cii Student Papers 2021

Karlsruhe Institute of Technology
aifb
CRITICAL INFORMATION INFRASTRUCTURES RESEARCH GROUP

cii Student Papers - 2021

Research Group Critical Information Infrastructures (cii)
Karlsruhe Institute of Technology
Department of Economics and Management
Institute of Applied Informatics and Formal Description Methods

Web: cii.aifb.kit.edu

Corresponding Editor:
Prof. Dr. Ali Sunyaev
Kaiserstr. 89
76133 Karlsruhe, Germany
Phone: +49 721 608-43679
Email: sunyaev@kit.edu

DOI: 10.5445/IR/1000138902

Editorial

Critical information infrastructures are sociotechnical systems comprising essential software components and information systems with pivotal impact on individuals, organizations, governments, economies, and society. Here at the Karlsruhe Institute of Technology (KIT), our research group works on practice- and research-driven challenges concerned with the design, development, and evaluation of reliable and secure software and information systems. Our research features a strong focus on Internet and health care industries as well as on the industry-specific application of secure and trustworthy AI models. The principal goal of our research is theorizing on and designing the applications and methods required for the creation and innovation of sociotechnical systems with promising value propositions. Our work accounts for the multifaceted use contexts of information and communication technologies with research on human behavior affecting critical information infrastructures and vice versa. This enables us to rigorously generate strong theoretical insights while simultaneously producing research outputs of relevance to practical audiences.

Every year, our research group supervises more than 150 bachelor and master students during their studies at the KIT. Because it is of our utmost interest to offer a compelling, engaging, and fruitful teaching experience for our students, we actively introduce our research topics to the students in various seminars and lectures and give the students not only the chance to work on emerging topics in practice and research but also provide them a first insight into our daily work as researchers. During our courses, students mostly work in groups and deal with problems and issues related to sociotechnical challenges in the realm of (critical) information systems. Usually, topics align with our current research endeavors. In addition, students also propose their own research topics or even perform their studies in cooperation with small, medium, or large-sized organizations. Topics vary every semester and range from privacy risks when using disruptive health information systems (Gojka et al. 2021; Yari et al. 2021); securing data in cloud, fog and edge services (Lins et al. 2021); designing and implementing distributed ledgers (Kannengießer et al. 2020); understanding how to achieve trustworthy AI in autonomous vehicles (Renner et al. 2021); to the specification and training of sophisticated AI models to solve prevalent issues in economies and our society, such as the COVID-19 pandemic (Pandl et al. 2021b). Students are supervised by our research group throughout the complete research process, including an assistance on to how to identify and structure problems, apply appropriate research methods rigorously, develop and communicate a potential solution, and finally write a scientific report.

Bringing the scientific work closer to students and actively engaging students in our everyday work offers many opportunities, not only for the students but also for our research group and the research community as well as practice in general. Students get involved in timely practice problems that research is trying to solve. Knowledge and experience gained throughout the seminars and lectures help students to better understand and apply the theoretical foundations learned during prior lectures. Offering such scientific courses also helps students to gain first experiences or deepen already existing knowledge on how to write upcoming bachelor and master theses. Indeed, most students continue their research after finishing the seminars or lectures either in form of a thesis, as part of their novel student job at our research group, or even voluntary in their free time. Students also commonly get in touch with organizations, enabling first cooperation and even paving the way for upcoming employments.

Our seminars and lectures often result in excellent student works. Unfortunately, brilliant student works have far too often disappeared into drawers in the past, although disruptive and game-changing insights have been proposed by students. While we as a research group always highly acknowledge brilliant students works, incorporate their findings into our own research projects, and have often published exceptional findings on scientific conferences and in journals (e.g., Gräbe et al. 2020; Kannengießer et al. 2020; Lins et al. 2021; Petry et al. 2021; Renner et al. 2020; Schmidt-Kraepelin et al. 2020b), we believe that students' huge efforts and promising findings often get little appreciation. Therefore, we conceived the idea for this book, offering students the possibility to publish their excellent works in this dedicated miscellany. We are very pleased to present such a collection for the first time, summarizing the best student works of 2020 and the first half of 2021. The works included in this miscellany come from four different courses, which offer students a broad selection of topics related to (critical) information systems:

Emerging Trends in Digital Health:

The seminar *Emerging Trends in Digital Health* aims at providing insights into current topics in the field of information systems with a focus on innovative digital healthcare systems. Kicking off with a short introduction and corresponding topics, students can choose to work on many different topics around the lectures and research topics of the research group, including genomics, distributed ledger technology, artificial intelligence, and gamification in healthcare. An example of a publication in this area is the paper by Thiebes et al. (2020c), which address the lack of knowledge concerning business models of DTC genetic testing services by systematically identifying the salient properties of various DTC genetic testing service business models as well as discerning dominant business models in the market.

Emerging Trends in Internet Technologies:

Similiar, the seminar *Emerging Trends in Internet Technologies* aims at providing students insights into current topics in the field of information systems while mainly focusing on fundamental and innovative Internet technologies. Students are offered a selection of topics around the lectures and present research of our group including distributed ledger technology (Kannengießer et al. 2020), cloud, fog and edge computing (Lins et al. 2018; Renner et al. 2020), artificial intelligence (Renner et al. 2022; Thiebes et al. 2020a), security, and privacy (Gojka et al. 2021; Yari et al. 2021). For example, our research group has recently provided a thorough conceptualization of the phenomena *Artificial Intelliegence as a Service* given a lack of conceptual clarity in academia and practice (Lins et al. 2021).

Critical Information Infrastructures:

The course *Critical Information Infrastructures* introduces students to the world of complex sociotechnical systems that permeate societies on a global scale. Being offered every winter term, master students learn to handle the complexities involved in the design, development, operation, and evaluation of critical information infrastructures. In the beginning of the course, critical information infrastructures are introduced on a general level (Dehling et al. 2019). The following sessions focus on an in-depth exploration of selected cases that represent current challenges in research and practice. Students work in groups of four on specific topics and must write a course paper. The research group has also published a book chapter providing a discussion on the characteristics and challenges of critical information infrastructures (Dehling et al. 2019).

Digital Health:

The course *Digital Health* introduces master students to the subject of digitization in healthcare. Students learn about the theoretical foundations and practical implications of various topics surrounding digitization in healthcare, including health information systems, telematics, big healthcare data, and patient-centered healthcare (e.g., Pandl et al. 2021a; Rädsch et al. 2021; Thiebes et al. 2020b; Warsinsky et al. 2021). After an introduction to the challenge of digitization in healthcare, the following sessions focus on an in-depth exploration of selected cases that represent current challenges in research and practice. Students work in groups of three to four on specific topics and must write a course paper. At the moment, the research group deals, among others, with the topic of a lack of knowledge concerning best practices in the design and implementation of gamification for health-related mobile apps by identifying archetypes of gamification approaches that have emerged in pertinent health-related mobile apps and analyzing to what extent those gamification approaches are influenced by the underlying desired health-related outcomes (Schmidt-Kraepelin et al. 2020a).

Out of these courses, we selected the student works that represent the best and most interesting studies. The student works in this book cover a wide range of research problems, including a deep dive of the promising GAIA-X project, direct-to-consumer genetic testing, machine learning in digital health, learning from IT security catastrophes, the overlap of user preferences with search engine performances, and a review of the application of social comparison theory in mobile health research.

- Deckers et al. investigate how direct-to-consumer business models have changed and evolved due to novel advances in genetic technology, consumer demand, and legal regulations. Looking at archived service provides websites, they compared the services' business models along the taxonomy developed by Thiebes et al. (2020c).
- The study by Kruse and Kramsakov provides deeper insights into the historical business models of the direct-to-consumer genetic testing companies 23andMe and Ancestry by conducting a systematic literature review.
- Enderle, Remmele and Stöcker performed a literature research on the current application of machine learning for the discovery of antibiotics. Their results show, that neural networks, support vector machines, and decision trees are widely used in the generation and discovery of novel drugs.
- Similar to the previous study, Seitter et al. also performed a literature review on machine learning methods for antibiotic discovery. Their findings show that currently machine learning is mainly used for virtual screening and end-to-end approaches.
- Budig et al. explored trade-offs between privacy-preserving and explainable machine learning in healthcare by performing a literature review and subsequently evaluating mentioned explainability methods in terms of their ability to preserve privacy.
- Traub et al. developed a post IT catastrophe checklist by studying select literature on past incidents. In doing so they provide a valuable tool for practitioners and researches to analyse IT security catastrophes in a structured manner.
- Bänfer et al. conducted interviews with members of the GAIA-X project, an initiative to create a federated, open data infrastructure based on common European values, to gain insights on challenges and opportunities the venture faces.
- The study by Abt et al. undertook a four-step approach, including an online survey and expert assessment, with the aim of evaluating the overlap of user preference with search engine performance. Their findings show that search results quality, ease of use, and privacy protection are the main drivers of search engine performance.
- Last, Binder, Rauh and Schröter shed light onto the use of social comparison techniques such as leaderboards for mobile health applications, by conducting a systematic literature search. Their results indicate a high potential for social comparison features, while practical implementation is still scarce.

Ending this brief overview, we are thankful and glad that students once again took their free time to revise and further improve their submitted and graded papers to ensure the high quality of this book. Next to the students who authored the articles for this miscellany, this book would not have been possible without the dedicated research associates of our research group, who supervised the students during the courses. Therefore, we would like to take the opportunity to express our sincere thanks for their active support, motivation, and commitment in the supervision of all student works from the cii research group.

We are looking forward to continuing in our mission of excellent teaching and want to publish the best student works annually within a miscellany, thus bringing scientific work closer to students.

Sincerely,

Ali Sunyaev, Maximilian Renner, Philipp A. Toussaint, Scott Thiebes, Sebastian Lins

Miscellany Team 2021

Supervising Research Associates

References

Dehling, T., Lins, S., and Sunyaev, A. 2019. "Security of Critical Information Infrastructures," in *Information Technology for Peace and Security : IT Applications and Infrastructures in Conflicts, Crises, War, and* Peace, C. Reuter (eds.), Wiesbaden, Germany: Springer Vieweg, pp. 319-339.

Gojka, E.-E., Kannengießer, N., Sturm, B., Bartsch, J., and Sunyaev, A. 2021. "Security in Distributed Ledger Technology: An Analysis of Vulnerabilities and Attack Vectors," in *Advances in Intelligent Systems and Computing*, K. Arai (eds.), London, GB: Springer.

Gräbe, F., Kannengießer, N., Lins, S., and Sunyaev, A. 2020. "Do Not Be Fooled: Toward a Holistic Comparison of Distributed Ledger Technology Designs," in in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS 2020),* Maui, HI, USA.

Kannengießer, N., Pfister, M., Greulich, M., Lins, S., and Sunyaev, A. 2020. "Bridges between Islands: Cross-Chain Technology for Distributed Ledger Technology," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS 2020),* Maui, HI, USA.

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., and Sunyaev, A. 2021. "Artificial Intelligence as a Service – Classification and Research Directions," *Business & Information Systems Engineering* (63), pp. 441-456.

Lins, S., Schneider, S., and Sunyaev, A. 2018. "Trust is Good, Control is Better: Creating Secure Clouds by Continuous Auditing," *IEEE Transactions on Cloud Computing (6:3),* pp. 890-903.

Pandl, K. D., Feiland, F., Thiebes, S., and Sunyaev, A. 2021a. "Trustworthy Machine Learning for Health Care: Scalable Data Valuation with the Shapley Value," in *CHIL '21: Proceedings of the Conference on Health, Inference, and Learning, M. Ghassemi* (eds.), Association for Computing Machinery (ACM), pp. 47-57.

Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., and Sunyaev, A. 2021b. "How Detection Ranges and Usage Stops Impact Digital Contact Tracing Effectiveness for Covid-19," *Scientific Reports* (11:1), 9414.

Petry, L., Lins, S., Thiebes, S., and Sunyaev, A. 2021. "Technologieauswahl Im Digitalpakt: Wie Werden Entscheidungen Im Bildungssektor Getroffen?," *HMD Praxis der Wirtschaftsinformatik*.

Rädsch, T., Eckhardt, S., Leiser, F., Pandl, K. D., Thiebes, S., and Sunyaev, A. 2021. "What Your Radiologist Might Be Missing: Using Machine Learning to Identify Mislabeled Instances of X-Ray Images," in *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*.

Renner, M., Münzenberger, N., von Hammerstein, J., Lins, S., and Sunyaev, A. 2020. "Challenges of Vehicle-to-Everything Communication. Interviews among Industry Experts," in *15th International Conference on Wirtschaftsinformatik (WI2020),* Potsdam, Germany: GITO Verlag, pp. 1831-1843.

Renner, M., Lins, S., Söllner, M., Thiebes, S., and Sunyaev, A. 2021. "Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial Intelligence-capable Technology," in *Proceedings of the 42nd International Conference on Information Systems (ICIS21)*, Austin, TX, USA.

Renner, M., Lins, S., Söllner, M., Thiebes, S., and Sunyaev, A. 2022. "Understanding the Necessary Conditions of Multi-Source Trust Transfer in Artificial Intelligence," in *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS2022),* Maui, HI, USA.

Schmidt-Kraepelin, M., Toussaint, P. A., Thiebes, S., Hamari, J., and Sunyaev, A. 2020a. "Archetypes of Gamification: Analysis of Mhealth Apps," *JMIR mHealth uHealth* (8:10), e19280.

Schmidt-Kraepelin, M., Warsinsky, S., Thiebes, S., and Sunyaev, A. 2020b. "The Role of Gamification in Health Behavior Change: A Review of Theory-Driven Studies," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS 2020),* Maui, HI, USA.

Thiebes, S., Lins, S., and Sunyaev, A. 2020a. "Trustworthy Artificial Intelligence," *Electronic Markets (31)*, pp. 447-464.

Thiebes, S., Schlesner, M., Brors, B., and Sunyaev, A. 2020b. "Distributed Ledger Technology in Genomics: A Call for Europe," *European Journal of Human Genetics* (28), pp. 139-140.

Thiebes, S., Toussaint, P. A., Ju, J., Ahn, J. H., Lyytinen, K., and Sunyaev, A. 2020c. "Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing," *Journal of Medical Internet Research* (22:1), e14890.

Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., and Sunyaev, A. 2021. "Are Gamification Projects Different? An Exploratory Study on Software Project Risks for Gamified Health Behavior Change Support Systems," in *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS2021)*, Kauai, HI, USA.

Yari, I. A., Dehling, T., Kluge, F., Eskofier, B., and Sunyaev, A. 2021. "Online at Will: A Novel Protocol for Mutual Authentication in Peer-to-Peer Networks for Patient-Centered Health Care Information Systems," in *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS 2021)*, Kauai, HI, USA.

Table of Contents

Direct-to-Consumer Genetic Testing: A History of Business Models

Emerging Trends in Digital Health, Summer Term 2020

Felix Deckers
Master Student
Karlsruhe Institute of Technology
felix.deckers@t-online.de

Valon Gjonaj
Master Student
Karlsruhe Institute of Technology
valongjonaj@gmail.com

Eugenia Pawlenko
Master Student
Karlsruhe Institute of Technology
eugn991@gmail.com

Ali Yüksel
Master Student
Karlsruhe Institute of Technology
ali.yuksel1995@web.de

Abstract

Background: Due to the intervention of the American Food and Drug Administration (FDA) in the previously unregulated market of direct-to-consumer (DTC) genetic testing in 2013, providers were forced to adapt certain business practices to the new requirements. This has brought changes in the business models, which will be researched further.

Objective: With this thesis we look at the changes in the business models of DTC genetic testing providers in the past.

Methods: For this purpose, the first systematically developed taxonomy by Thiebes et. al. (2020) is used, which is a snapshot of the year 2018, but does not consider the past. The referenced taxonomy is applied to the years 2008/2009 as well as 2013/2014 to investigate the temporal course of the changes. 55 services are selected and examined in the different time periods.

Results: Our paper shows that regulatory measures can have an impact on the business models of DTC genetic test providers. To understand the extent of the changes, further research on the background and interrelationships of the DTC genetic test market is required. This will allow us to better assess the response of DTC genetic test providers to warnings and bans by regulators. In turn, this will allow legislators to be more specific about the alignment of competitive conditions and consumer protection.

Conclusion: The changes in the business models between the periods 2013/2014 (after the FDA intervention) and 2018 are particularly evident.

Keywords: direct-to-consumer, genetic testing, business models, taxonomy

Introduction

Problem Definition

In April 2003 the Human Genome Project published the full sequence of the human genome to the public (What is the Human Genome Project? n. d.). As early as 2005, the first genetic test providers entered the private market and offered tests that were sold directly to the private end consumer, also known as direct-

to-consumer (DTC) genetic testing (What is direct-to-consumer genetic testing? 2020). These companies offered services for sequencing and interpretation of the human genome for a variety of purposes like ancestry, predictions on health and paternity testing. Driven by decreasing costs of sequencing and analysis of gene data (DNA Sequencing Costs: Data 2019), genetic testing became more and more popular (Bowen 2018, para. 2; Regalado 2018, 2018). In 2013, the unregulated market led the Food and Drug Administration (FDA) to protect consumers by issuing cease and desist letters to suppliers for certain business activities. As a result, genetic testing providers gradually adapted their tests to the new regulatory requirements, which Allyse et. Al (2018) defines as "DTC 2.0". The FDA approved the first tests under new conditions subsequently in 2015, which leads us to the current form of DTC genetic testing known today. This was followed in 2020 by a scientific study of business models by Thiebes et. al. (2020). Within that work a taxonomy of business models was developed and then all services were divided into archetypes (cluster). This was the first systematic taxonomy development in the field of DTC genetic testing. The results of scientific study represent a snapshot of the period under research (2018) and call for further research in this area. Specifically, they invoke for an examination of past time periods and the change of the business models over time. If we now also consider the timeline of Allyse et al. (Figure 2), with its important events in the world of DTC genetic testing, we see that an overview of the development of business models over time makes an important contribution (knowledge gain) to the current state of research. This knowledge gain is valuable beyond the scope of research. It would be particularly interesting for established and future market participants in the DTC genetic industry and players in the health care segment. Furthermore, seeing how the business models have developed over the last two decades would be useful for consumers, to better assess the decision to use such DTC genetic testing services.

This leads us to the question:
What changes have occurred for business models of DTC genetic testing providers in the periods 2008/2009 and 2013/2014?

Objective of the Work

The aim of this work is to contribute two further periods from the past to the taxonomy of Thiebes et al., which reflects the year 2018. The results of an applied taxonomy provide an understanding of the business model changes in the DTC genetic testing environment due to more data and information. Once the changes have been identified, an attempt is made to analyze why the changes have taken place. The gained knowledge makes contribution to research by identifying the full-time span in the emerging competitive landscape of DTC genetic testing business models. It can help end consumers as well as health professional's decision-making process in selecting the right service for their own needs and interests. Moreover, our findings can serve the genetic testing industry to differentiate its business models from those of other competitors. Finally, the results will enable regulators to design laws that allow competition without compromising consumer protection.

Structure of the Work

This work is structured in five chapters and can be summarized as follows. The previous chapter defines the problem statement and objective by outlining the research gap in the changes of business models of DTC genetic testing services before 2018. Chapter two describes the theoretical background of DTC genetic testing and highlights the difference between sequencing and genotyping. The developed taxonomy of business models in DTC genetic testing by Thiebes et al. (2020), will be introduced in chapter 2.3. Chapter three presents the research approach, which consists of four steps. Chapter 3.1 describes how to select the relevant time periods that optimally reflect the changes in business models. In chapter 3.2, the selection of the DTC genetic testing services to be examined, will be introduced. This includes the reduction of services, supporting tools to be applied and the identification of services to be provided. Following, in chapter 3.3, the data collection will be explained. Chapter 3.4 focuses on the data comparison of the collected data. Chapter four will present the results of the services selection within the relevant time periods. Finally, the data collection and the data comparison will be stated. After presenting the research results, they will be discussed in chapter five. Principal findings will be stated, and theoretical and practical implications will be derived. Lastly the limitations of the research approach will be reflected and an outlook for further research will be given.

Theoretical Background of Genetic Testing and the Taxonomy

Direct to Consumer Genetic Testing

In the following, the term "Direct-to-Consumer Genetic Testing" (DTC genetic testing) will be considered. DTC genetic tests "are marketed directly to customers via television, print advertisements, or the Internet (What is direct-to-consumer genetic testing? 2020)." These tests can be bought online or in stores. In order to analyze the customers' DNA, the "customers send the company a DNA sample and receive their results directly from a secure website or in a written report (What is direct-to-consumer genetic testing? 2020)." Genetic tests allow people to get an "access to their genetic information without necessarily involving a health care provider or health insurance company in the process (What is direct-to-consumer genetic testing? 2020)." There are different methods of genome data collection which we will discuss in more detail in the next chapter.

Sequencing vs. Genotyping

In general, there are two ways of obtaining genome data. A distinction is made between sequencing and genotyping. In genotyping, there are two different procedures. Whole genome sequencing (WGS) is a laboratory process to "analyze the entire genomic DNA sequence of a cell at a single time, providing the most comprehensive characterization of the genome (Morgensztern 2018)." After the publication of the Human Genome Project in 2003, WGS became available which generated the reference for human genome sequences (Morgensztern 2018). One level below there is a laboratory method used "to focus on targeted sequencing of the protein coding regions of the genomic DNA (Gleason und Juul 2018)", called whole exome sequencing. It is a tool for "gene discovery for complex diseases and for facilitating the accurate diagnosis of individuals (Gleason und Juul 2018)."

On the other hand, genotyping can be used to determine "which genetic variants an individual possesses (Difference Between DNA Genotyping & Sequencing 2020)." Depending on the variants of interest and the resources available, "genotyping can be performed through a variety of different methods. For looking at many different variants at once, especially common variants, genotyping chips are an efficient and accurate method. They do, however, require prior identification of the variants of interest (Difference Between DNA Genotyping & Sequencing 2020)." DTC genetic testing can be offered due to the different methods of genome data collection.

Overview of the Taxonomy of Business Models in DTC Genetic Testing

In business, there is a wide range of definitions of business models. Since our work is based on the paper of Thiebes et al. (2018), we use their definition of business models. Business models are understood as "a representation of a firm's underlying core logic and strategic choices for creating and capturing value within a value network (Shafer et al. 2005)."
Based on this, in the paper "Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing" (Thiebes et al. 2020) a taxonomy was developed in which 277 services consisting of six clusters could be compared and analyzed. The taxonomy was developed using the methodology of Nickerson et al. (2018) and consists of seven iterative steps and provides guidelines for each step in the taxonomy development, as shown in Figure 1 (Nickerson et al. 2013).

The final taxonomy arose from the application of this methodology. 15 dimensions with two to four characteristics resulted in a total of 41 characteristics which are mutually exclusive. The dimensions were divided into the following four categories (Thiebes et al. 2020):

- Strategic Choices
- Value Network
- Create Value
- Capture value

Furthermore, six clusters were classified so that each cluster addresses a different target group. Thereby, each cluster is focusing a different target group. In cluster four for example, providers of simple health tests

are considered (Thiebes et al. 2020). The exact classification and description of the clusters is described in the paper (Thiebes et al. 2020).

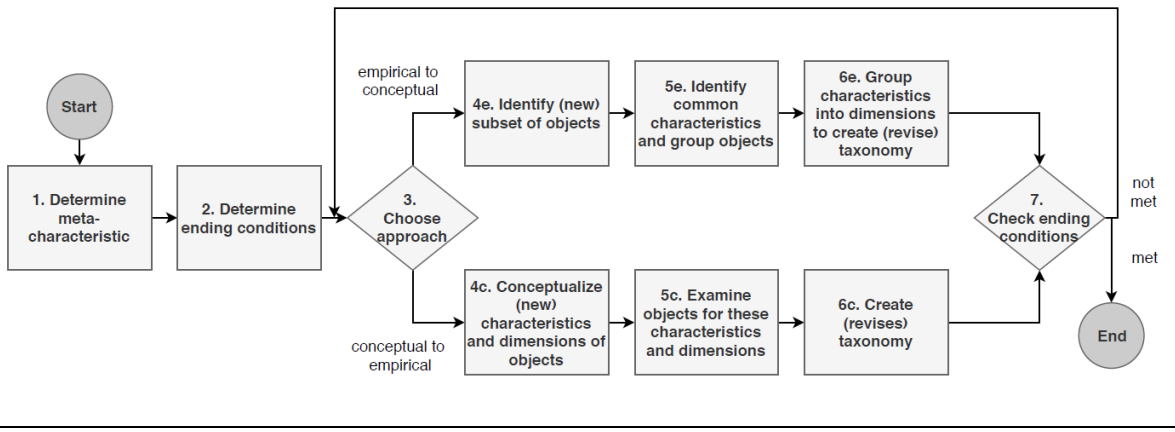

Figure 1. Taxonomy Development (Nickerson et al. 2013)

Research Method

Selection of Relevant Time Periods

As already described in chapter one, despite the establishment of DTC genetic testing providers since 2005, full regulatory intervention by the FDA did not occur until 2013 (Allyse et al. 2018,). In response to a previously published Report of the US Government Accountability Office (GAO) about investigations into business practices in 2010 (Kutz 2010), the FDA first sent warning letters to the largest DTC genetic testing companies. Later, in 2013, cease and desist letters were sent out to several DTC genetic providers (Green und Farahany 2014). The choice of periods is made depending on those crucial events. There, the greatest changes are assumed to occur. Accordingly, business models before and after the FDA intervention in 2010 and 2013 will be considered. Figure 2 illustrates the events and selected time periods in a timeline, which will be explained in the following.

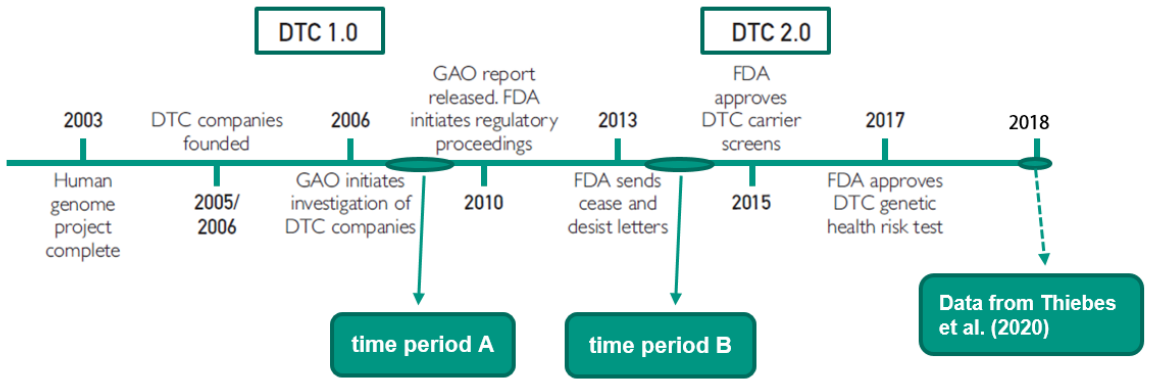

Figure 2. Timeline of Events Related to DTC Genetic Testing in Style of (Allyse et al. 2018)

The first period is set to 2008/2009 (time period A). During this period, there was no clear regulatory framework for evaluating the analytical and clinical validity and clinical utility of DTC genetic tests (Andrew S. Robertson 2009). Although some of the DTC genetic products contained very sensitive health data, the procedures for securing consent were poor (Skirton et al. 2012). Especially as DTC genetic tests were handled as commercial transactions and not as medical devices.

Furthermore, there was no restriction on panel content and no validation of user understanding was required (Allyse et al. 2018 .

For the second period the choice is 2013/2014 (time-period B). It is chosen in such a way that a regulatory procedure has already been initiated and cease and desist letters have been sent by the FDA, but DTC Carrier Screens have not yet been approved (Allyse et al. 2018). The aim is to assess whether and how DTC genetic providers were reacting to the circumstances. According to the regulations, DTC genetic tests were increasingly medically integrated and panel content was classified according to risk level. In addition, analytical validation became necessary, as well as validation of user understanding. Furthermore, there was a stronger separation between health and entertainment in terms of results. Due to the lack of data and redirects on Archive.org, both periods must be expanded to two years to have a bigger source of data.

Selection of DTC Genetic Testing Services

Reduction of Services to be Examined

Since the evaluation of 277 DTC services for two periods would go beyond the scope of a seminar paper, the target is 60 services. This seems reasonable in view of the time and the group size of four persons. The six clusters defined in the paper of Thiebes et al. (2020) are to remain in place. Accordingly, an attempt is made to evaluate ten providers per cluster for each period.

Supporting Tools for the Selecting Process

Two tools are used to support the process for reducing the number of providers from 277 to at most 60. On the one hand, the data availability for the periods A and B is checked via archive.org. Second, the providers are ranked according to their degree of popularity with the help of Alexa.com. The two websites will be described subsequently.

Alexa.com offers various analysis tools such as Search Engine Optimization (SEO) analysis or keyword research (Alexa - Alexa Internet - About Us n. d.). Regarding to the popularity ranking, however, the statistics on website traffic are in the foreground. Alexa.com was chosen because it contains information on most DTC genetic testing providers. After entering the web address of the respective DTC genetic service, a list appears which includes the site rank. The rank is calculated using a combination of the average number of daily visitors and page views of the website in the last three months (Alexa - Alexa Internet - About Us n. d.).

Archive.org is a digital non-profit archive for websites and other cultural artefacts as well as the largest existing database for this kind of information in the world (Hoeren 2006). Using the Wayback Machine to check archived Websites, there are two capabilities. On one hand there is the option to save only certain screenshots of the site, these are marked with a blue dot in the calendar view. If, on the other hand, the entire website is saved, but in certain circumstances at various times, a green dot appears as Figure 3 shows.

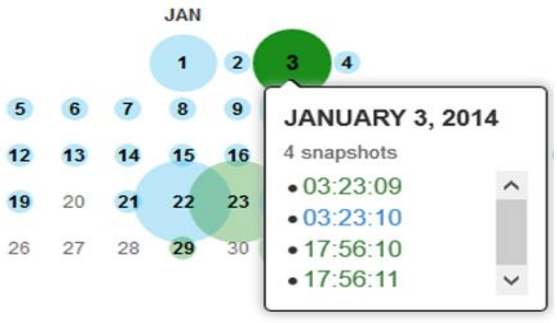

Figure 3. Calendar View in Archive.org. Following the Source (Wayback Machine n. d.)

For the selection of the services, first the two time periods A and B are searched for the availability of green dots on archive.org. Afterwards, providers for which data is available in both periods are included in the selection. When the taxonomy is used later, information is taken from the fully archived websites.

Identification of Services to be Examined

According to chapter 3.2.1, we want to determine ten services per cluster. To reach this, data must be created which indicates whether green dots are present for period A and period B. If this is the case, an "x" will be placed in the cell "available (green dot?)" for the examined service according to the respective period. Table 1 shows an excerpt of the excel table used and illustrates the procedure.

Cluster 1		**Period A (2008/2009)**		**Period B (2013/2014)**	
Name	**Website**	**available (green dot?)**	**not available**	**available (green dot?)**	**not available**
23andMe	https://www.23andme.com	x		x	
ActX	https://www.actx.com/	x		x	
ANABOLICGenes	https://anabolicgenes.com		x		x
Ancestry DNA	http://ldna.ancestry.com		x	x	
Arivale	https://www.arivale.com/		x		x
AthGene	http://www.athgene.com/		x		x
Athletigen	https://athletigen.com/		x	x	

Table 1. Excerpt of the Table Used to Identify Existing Data from of Service Websites

After the database has been created, it must be determined if more than ten services have archived data (green dots) for the respective periods. If this is the case, the filtered services are classified in descending order using the website ranking of Alexa.com. This is done to choose the potentially most popular services. For this purpose, another table is created in which the corresponding ranking for each service is listed. Table 2 shows an excerpt of this table.

Cluster 1	**Website**	**Alexa Ranking**
23andMe	https://www.23andme.com/health/	7.480
ActX	https://www.actx.com/	4.039.673
ANABOLICGenes	https://anabolicgenes.com/en/index.html	6.711.453
Ancestry DNA	http://ldna.ancestry.com/offers/buyKit.aspx	1.330
Arivale	https://www.arivale.com/	2.210.718
AthGene	http://www.athgene.com/	10.626.398
Athletigen	https://athletigen.com/	1.202.240

Table 2. Excerpt of the Table Used to Rank Services

The selection is carried out for all six clusters for the defined time periods and is explained step by step in the following sections. The detailed excel tables regarding the data identification and website ranking are available by the authors upon request.

Approach for Selection of Time Period A:

For the beginning of the analysis, period A is examined first, as it is prior to period B. We assume that less datasets are available for period A than for period B because many of the services were still under development in period A (Allyse et al. 2018).
For the selection of the services we prioritize the explicitly mentioned services in the referenced paper (Thiebes et al. 2020). These can be found in Table 3.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
23andMe	African Ancestry	DNA Diagnostics Center	SkinDNA	Genetic Genie	Dante labs
FamilyTreeDNA	EasyDNA	Alpha Biolabs	International Bioscience	Promethease	Full Genomes Corporation
MyHeritageDNA	FitGenes	Dadchecksilver	Pillcheck	Roots for Real	Helix
Genographic Project	The Making of Me	Who'z the daddy		My Genetic Health	Genes for Good

Table 3. Explicitly Mentioned Services in Referenced Paper

We give priority to these services because of their representativeness and, if possible, examine them for changes within the 15 dimensions.

The selection process was carried out according to the following procedure:

1. START.

At this point we have a list of services which websites have been stored for the period on archive.org with the mentioned "green dots". So far, we have not yet selected exactly which services we want to use in the analysis later.

2. Select services out of list that are archived with a "green dot" and proceed procedure.
3. Are more than ten Services archived?
    - If not applicable (less than ten Services), stop the iterative process and proceed with the analysis of the services.
    - If applicable perform step four.
4. Selection of services explicitly mentioned in the paper (see table 3).
5. Filling up to ten services by comparing the services with Alexa Ranking (possibility that not all websites have been ranked).
6. Are exactly ten services selected?
    - If not applicable randomly fill up to exactly ten services.
    - If applicable perform step seven.
7. STOP the iterative process and proceed with the analysis of the development.

For a better understanding the process is shown in a flow chart (see Figure 4).

Approach for Selection of Time Period B:

The identification of the services for period B is based on the results from period A. The services from period A are transferred to period B if archived data exists for period B. This is done to be able to compare the same services through different periods afterwards.

The approach of period A is therefore extended with step two. For a better understanding of the entire workflow, the entire workflow is shown again below. Apart from step two, the workflows in period A and B are identical.

1. START.
2. Transfer of the selected services from period A and check, if we also have archived data for these services for period B.

At this point we have a list of services which websites have been stored for the period on Archive.org with the mentioned "green dots". So far, we have not yet selected exactly which service we want to use in the analysis later.

3. Select service out of list that are archived with a "green dot" and proceed procedure with the select services and the checked services of period A.

4. Are more than ten Services archived?
    - If not applicable (less than ten Services) Stop the iterative process and proceed with the analysis of the services.
    - If applicable perform step five.
5. Selection of services explicitly mentioned in the paper (see table 3).
6. Filling up to ten services by comparing the services with Alexa Ranking (possibility that not all websites have been ranked).
7. Are exactly ten services selected?
    - If not applicable randomly fill up to exactly ten services.
    - If applicable perform step eight.
8. STOP the iterative process and proceed with the analysis of the development.

For a better understanding the process is shown in a flow chart (see Figure 5).

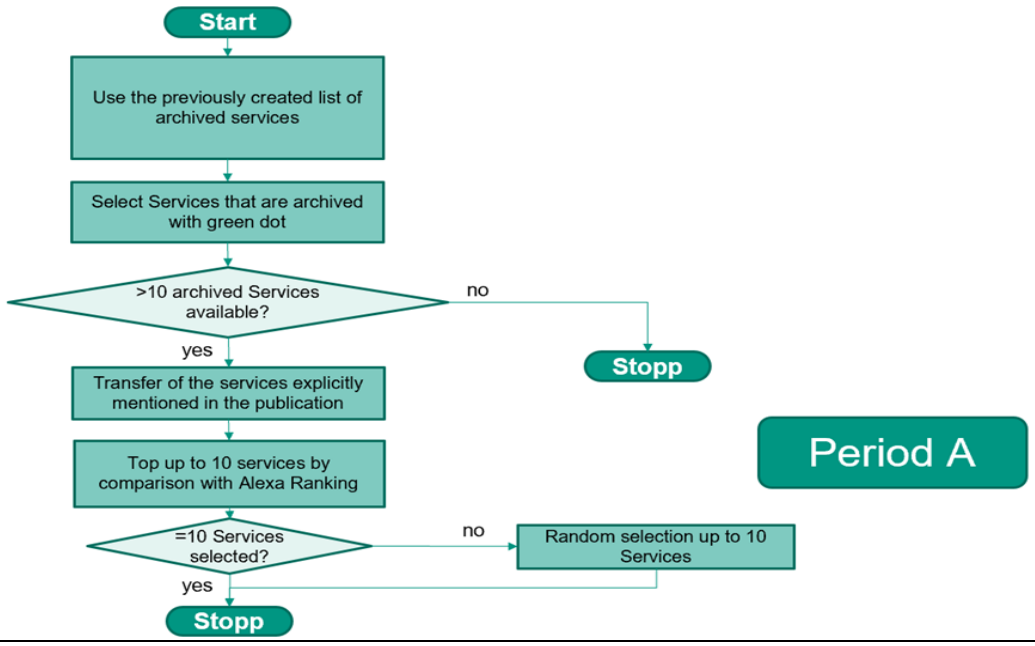

Figure 4. Process Flow Service Selection Period A

Data Collection

The following paragraph explains how the already identified services are examined for the 15 dimensions from the paper by Thiebes et al. (2020). The services are classified per cluster and time period by the same group member to ensure a consistent analysis within the clusters. This is done because the classification of the individual dimensions can differ depending on the individual's perception.

Additionally, the first cluster was examined by all the group members together to ensure a common understanding of each dimension of the taxonomy. To classify the services, the already identified website addresses of the respective services from the referenced paper are used to access the archived websites at Archive.org. Then the websites for the periods A and B are successively examined for the 15 different dimensions and corresponding characteristic.

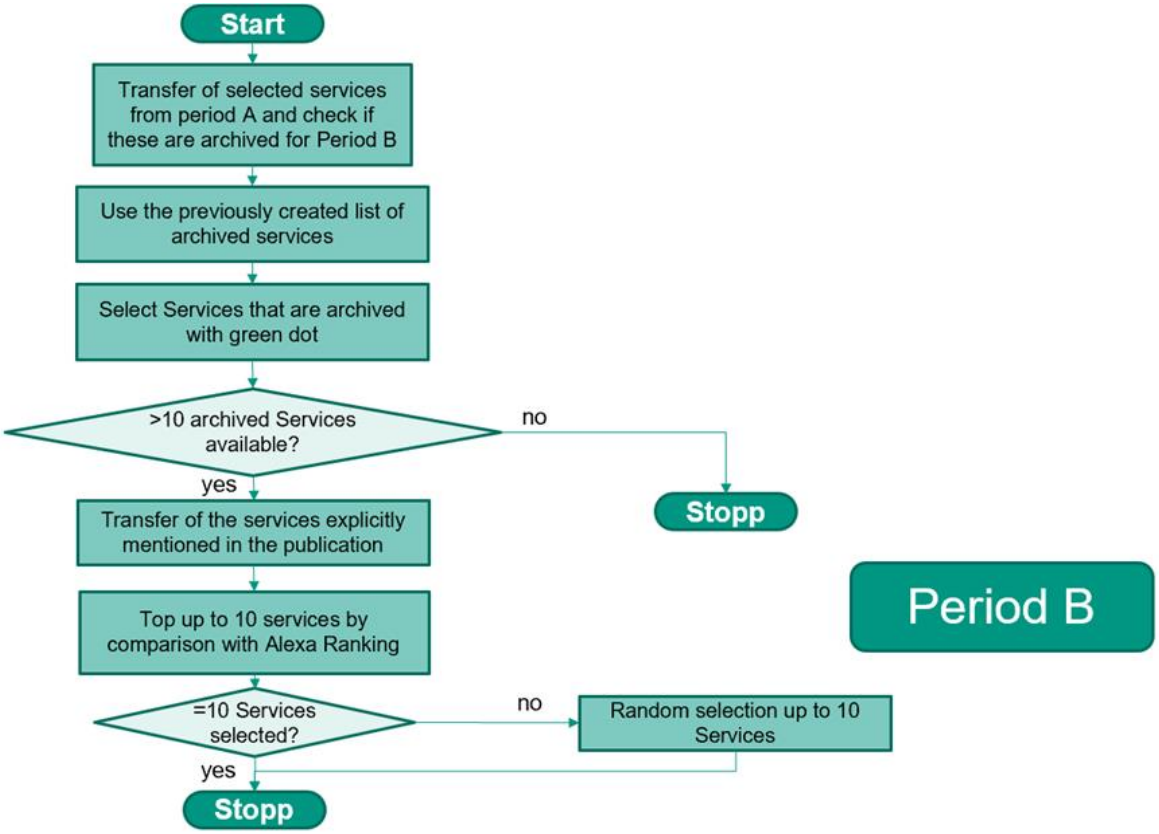

Figure 5. Process Flow Service Selection Period B

The results are notated according to the following methodology:

- If a dimension can be characterized, an "x" is placed in the corresponding characteristic.
- If it is not possible to specify the characteristics of a dimension exactly, all characteristics of the dimension under investigation are marked with a question mark "?"
- In case the service is not existing for period A or B, it is marked as "Company not existing".
- If the service does not provide DTC genetic tests for the period A and B, it will be entered as "no DTC test offered".

Company	**Cluster**	**Data available?**			**Line**	**Business purpose A-->B**	**Business purpose B-->C**	**Region of operation A-->B**
		A	**B**	**C**				
23andMe	1	data available	data available	data available	S1	No Change	No Change	Change
ActX	1	data available	data available	data available	S2	Change	No Change	Change
DNA Ancestry Project (Genebase)	1	data available	data available	data available	S3	No Change	No Change	No Change
Myriad	1	data available	data available	data available	S6	No Change	No Change	Change
Ancestry DNA	1	no data	data available	data available	S7	no data	No Change	no data

Table 4. Overview of the Changes Template (Exemplary Representation)

Data Comparison

After data collection and filling the taxonomy with both periods, a method is needed to reflect the changes in a summarized form. For this purpose, we have created an Excel file that is based on the taxonomy data and visualizes the changes between period A to B and B to C.

Number of Changes						
Dimension / Cluster	**1**	**2**	**3**	**4**	**5**	**6**
Business purpose A-->B	1	1	0	1	0	0
Business purpose B-->C	0	1	0	0	1	1
Region of operation A-->B	3	1	0	1	0	0
Region of operation B-->C	1	1	0	2	0	3
Consumer target group A-->B	2	1	0	1	0	0
Consumer target group B-->C	3	6	3	4	6	8
Consumer research consent A-->B	2	1	0	1	0	0
Consumer research consent B-->C	5	2	3	1	3	5
Distribution channel A-->B	1	1	0	1	0	0
Distribution channel B-->C	1	3	4	0	0	5
Sample site A-->B	1	1	0	1	0	1
Sample site B-->C	1	3	4	0	0	4
Sampling kit provider A-->B	2	1	0	1	0	1
Sampling kit provider B-->C	5	2	1	0	7	5
Sample storage A-->B	2	0	1	1	0	0
Sample storage B-->C	4	4	3	0	3	5
Genome test type A-->B	1	1	0	1	0	1
Genome test type B-->C	0	1	0	1	0	6
Data storage A-->B	1	0	0	1	0	0
Data storage B-->C	1	2	4	0	4	3
Data ownership A-->B	4	0	0	1	0	0
Data ownership B-->C	3	1	2	0	2	2
Data processing A-->B	1	1	0	1	0	0
Data processing B-->C	1	2	0	1	2	4
Fee type A-->B	1	1	0	1	0	0
Fee type B-->C	2	2	0	0	0	1
Fee payer A-->B	1	1	0	1	0	0
Fee payer B-->C	1	2	1	0	1	2
Reselling of genome data A-->B	1	1	0	1	0	0
Reselling of genome data B-->C	2	2	0	0	0	2
Average no. of changes (w/o 0)	**x = 2,125 ≈ ℤ =3**					

Table 5. Overview of Changes per Dimension and Cluster

In Table 4 the first step is to check whether the corresponding data for all periods A, B and C (C-period data if delivered by the Thiebes et al. (2020) taxonomy) is available for each service. Then one of the options "No Information in A or B, No Data, Change, No Change is displayed.

The meaning is as follows:

- "No Data" – if there was no data available for the respective service in that period.
- "No information in A (or B)" – if there was a "?" in the characteristic of a dimension in period A or B.
- "Change" – if the information/data changed in the characteristic of the dimensions in the time periods AB or BC.
- "No Change" – if no data has changed in the characteristics of a dimension.

Using the table in Table 4, the changes per dimension and cluster are now to be counted. This should produce a table with the number of changes, which can be seen in Table 5.

To show above-average amounts of changes, we calculate the mean value without zeros for the entire table and round up the result to the next larger integer ($\mathbb{Z}$). The rounding up indicates that we only focus on above-average changes in the context of this work. This number serves as an indicator for the conditional formatting in Table 5, which follows the rule: Mark all cells yellow, which are greater or equal to $\mathbb{Z}$. The detailed excel tables of the data comparison are available by the authors upon request.

Results

Results of Selection of Services

The services of Period A and B were examined for each cluster and selected according to the iterative process in chapter 3.2. Using the tables mentioned above, which have been shown in extracts in Tables 1 and 2, in total 30 services for period A and 55 services for period B were selected. In Table 6, the selected services are presented in tabular format. For Cluster one, six services were selected for period A and ten services for period B. The selection for Cluster two also resulted in six services for period A and ten services for period B.

	Cluster 1		**Cluster 2**	
No.	**Period A**	**Period B**	**Period A**	**Period B**
1	23andMe	23andMe	Genebase	Genebase
2	ActX	ActX	GenePlanet	GenePlanet
3	DNA Ancestry Project (Genebase)	DNA Ancestry Project (Genebase)	Genosense (and DNA Plus)	Genosense (and DNA Plus)
4	Myriad	Myriad	Holistic Health International	Holistic Health International
5	Family Tree DNA	Family Tree DNA	Home DNA Inc.	Home DNA Inc.
6	Genographic Project	Genographic Project	The Carlson Company LLC	The Carlson Company LLC
7		Ancestry DNA		African Ancestry
8		MyHeritage		FitGenes
9		Map My Genome		The Makings of Me
10		DNAFit		DexaFit LLC

Table 6. Overview of Selection of Services to be Examined Cluster One and Two

For Cluster three, eight services were selected for period A and ten services for period B. The selection for Cluster four resulted in nine services for period A and ten services for period B.

For Cluster five, no services were selected for period A and seven services for period B. The selection for cluster six resulted in one service for period A and nine services for period B.

Results of Data Collection

After the services for both periods were selected, the business models were examined for their characteristics. This was done using the taxonomy. The services divided into six clusters were examined for their characteristics in the mentioned 15 dimensions. Among other things, the type of the distribution

channel and the sample kit provider were identified. The other characteristics can be taken from Table 2 of the paper by Thiebes et al. (2020).

	Cluster 3		**Cluster 4**	
No.	**Period A**	**Period B**	**Period A**	**Period B**
1	American Paternity	American Paternity	Athena Diagnostics Inc.	Athena Diagnostics Inc.
2	DNA Solutions USA	DNA Solutions USA	Consumer Genetics	Consumer Genetics
3	Genetic Profiles	Genetic Profiles	DNALYSIS Biotechnology	DNALYSIS Biotechnology
4	Genetic Testing Laboratories Inc. (GTL)	Genetic Testing Laboratories Inc. (GTL)	EnteroLab	EnteroLab
5	Genetica DNA Laboratories Inc.	Genetica DNA Laboratories Inc.	GeneLex	GeneLex
6	LabCorp	LabCorp	International Biosciences	International Biosciences
7	Molecular Diagnostic Services (PTY) LTD	Molecular Diagnostic Services (PTY) LTD	Natera	Pink or Blue
8	Who´z the daddy?	Who´z the daddy?	Pink or Blue	Smart DNA
9		Alpha Biolabs	Smart DNA	QUEST Diagnostics
10		DNA Diagnostics Center (DDC)		Counsyl

Table 7. Overview of Selection of Services Cluster to be Examined Three and Four

First, period A was considered. Of a total of 30 services in period A, two service providers did not offer DTC genetic tests. Furthermore, it could be determined that one service provider did not exist at that time.

	Cluster 5		**Cluster 6**	
No.	**Period A**	**Period B**	**Period A**	**Period B**
1		Enlis Genomics	Medichecks	Medichecks
2		Geneknot		Full Genomes Corporation
3		Genetic Genie		YSEQ DNA Origins Project
4		NutraHacker		Foundation Medicine Inc.
5		Promethease		WeGene
6		Sports Gene LCC		Caligenix
7		Xcode		Asper Biogene
8				Bio Logis
9				Kailos Genetics

Table 8. Overview of Selection of Services to be Examined Cluster Five and Six

In addition, in 36 of 450 cases the exact characteristics could not be determined. In period B, 55 services were considered. One of the services did not offer DTC genetic tests. In our analysis, the exact characteristics could not be determined for this period B in 73 of 825 cases. In summary, we were nevertheless able to determine 91% of the examined characteristics. In over 90% of the cases, service providers were already profit-oriented and operated globally more often than locally. Furthermore, most undetectable characteristics were in the dimension sample storage. (As the list of all services examined would go beyond the scope of this work, the detailed results can be found in Multimedia Appendices which are available by the authors upon request.)

Results of Data Comparison

The following section illustrates the results of changes within the dimensions from period A to B and from period B to C broken down into six clusters.

Number of Changes							
Dimension / Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**∑ of Changes / Dimension**
Business purpose A-->B	1	1	0	1	0	0	3
Business purpose B-->C	0	1	0	0	1	1	3
Region of operation A-->B	3	1	0	1	0	0	5
Region of operation B-->C	1	1	0	2	0	3	7
Consumer target group A-->B	2	1	0	1	0	0	4
Consumer target group B-->C	3	6	3	4	6	8	30
Consumer research consent A-->B	2	1	0	1	0	0	4
Consumer research consent B-->C	5	2	3	1	3	5	19
Distribution channel A-->B	1	1	0	1	0	0	3
Distribution channel B-->C	1	3	4	0	0	5	13
Sample site A-->B	1	1	0	1	0	1	4
Sample site B-->C	1	3	4	0	0	4	12
Sampling kit provider A-->B	2	1	0	1	0	1	5
Sampling kit provider B-->C	5	2	1	0	7	5	20
Sample storage A-->B	2	0	1	1	0	0	4
Sample storage B-->C	4	4	3	0	3	5	19
Genome test type A-->B	1	1	0	1	0	1	4
Genome test type B-->C	0	1	0	1	0	6	8
Data storage A-->B	1	0	0	1	0	0	2
Data storage B-->C	1	2	4	0	4	3	14
Data ownership A-->B	4	0	0	1	0	0	5
Data ownership B-->C	3	1	2	0	2	2	10
Data processing A-->B	1	1	0	1	0	0	3
Data processing B-->C	1	2	0	1	2	4	10
Fee type A-->B	1	1	0	1	0	0	3
Fee type B-->C	2	2	0	0	0	1	5
Fee payer A-->B	1	1	0	1	0	0	3
Fee payer B-->C	1	2	1	0	1	2	7
Reselling of genome data A-->B	1	1	0	1	0	0	3
Reselling of genome data B-->C	2	2	0	0	0	2	6
Average no. of changes (w/o 0)	**x = 2,125**		**≈**	**𝕫 = 3**			
∑ of Changes / Cluster	54	46	26	24	29	59	∑ 241
∑ of Changes / Cluster for A to B	24	12	1	15	0	3	∑ 58
∑ of Changes / Cluster for B to C	30	34	25	9	29	56	∑ 183

Table 9. Results of Changes per Dimension and Cluster

As shown in Table 9, calculating the mean value without zeros provides a value of 2.152. This is rounded up to three due to the discreteness stipulation. Therefore, 33 changes are highlighted in yellow and will be investigated further. From all dimensions and time periods examined, 13.7% (33/241) of potentially relevant changes are identified. A further comparison of the individual intervals AB and BC reveals that the number of changes in interval BC is larger.
Looking at the table horizontally, it is evident that some dimensions contain lots of potentially relevant changes, while others have none. The *Consumer target group* dimension contains relevant changes for all clusters. Furthermore, the dimensions *Consumer research consent* and *Sample storage* also show a lot of changes. In contrast, the dimensions *Fee type, Fee payer* and *Reselling of genome data* show below average changes. Vertically, and thereby cluster-dependent, Cluster six contains the most changes, Cluster four the least.

Discussion

Principle Findings

By applying the taxonomy and adding the two periods A and B, changes in the business models for DTC genetic test providers were identified. As was highlighted in chapter 4.3, there were significantly more changes between periods B and C. This could be caused by the reaction of DTC genetic testing providers to cease and desist letters from the FDA. After one of the largest providers was given a warning, it can be assumed that other smaller services adapted their business models as well (Sharkey 2019). DTC genetic testing was declared as a medical product and therefore approval by the FDA was required (Sharkey 2019). Due to the classification as a medical product, higher requirements regarding clinical and analytical validity are applied (Sharkey 2019). As a result, DTC genetic testing providers have changed the way they interpret the tests and involve physicians in the evaluation of the analysis. This development is also reflected in our results in the *Sample site* dimension.

With reference to the selection and investigation of services in chapter 4.1, it is noticeable that no companies in cluster five could be selected in period A. DTC genetic testing providers in cluster five mainly focus on the interpretation of DNA analyzes, which they mainly obtain from third party providers. The establishment of the companies in cluster five could be based on the fact that the interpretation by large companies such as 23andme was prohibited by the FDA and these services were taken over by other providers such as *Promethease* (Regalado, 2018. Another reason could be the FDA's warning letters to the major suppliers in 2010, following which numerous companies focusing on the interpretation of health-related information were founded in the following years (Allyse et al. 2018).

Furthermore, the results show that, on the one hand, changes in the dimension *Consumer target group* are present across all clusters and, on the other hand, these are also the most changes within a dimension. In 2017, the FDA authorized DTC tests for ten hereditary diseases (Sharkey 2019. As a result, it can be assumed that the DTC genetic test providers may have expanded their product portfolio and therefore reached new target groups in response.

Implications for Researchers and Practitioners

For Researchers:

For research purposes, the results provide an overview of general changes in the DTC market. They also provide a suitable basis for further questions and more detailed work in the future. From the overview of changes in certain time intervals, correlations with regulatory events can be drawn, and further research into the background of the DTC genetic test market can be facilitated. Following the paper of Thiebes et al. (2020), this paper presents deeper insights into the application of business models in the health care segment.

For Practitioners:

The results of this elaboration can be used to track the performance of the competitors from industry perspective. In addition, the findings can be applied by established DTC genetic testing providers, potential newcomers, and other healthcare players to make better decisions for the future. Regulatory authorities

benefit from a better assessment of the reaction of the DTC genetic test market to regulatory interventions and can use this knowledge for future proceedings. Furthermore, politicians and legislators could be made more sensitive to the handling of sensitive health data. On one hand this could ensure the secure handling of end-consumer genetic data. On the other hand, it could help to create a level playing field. The results also help to raise the awareness of consumers who are considering using such DTC services.

Limitations

The target number of services was set at maximum of 60 services and could not be achieved. Using the methodology in chapter 3.2.3 only 55 services met the selection criteria, which were further analyzed.

At the beginning, two clusters were classified collectively by all group members to create a general understanding. During the data collection, each cluster was classified by the same group member to ensure consistent analysis within the clusters. Accordingly, different persons were responsible for different clusters. Since the classification of each dimension may vary depending on individual perceptions, the data collected are not collected with a completely unambiguous decision making. A collection from a single individual could produce a more accurate and consistent data collection.

The referenced green dots on the Archive.org website correspond to a large network of redirects, which attempt to represent the web address with individual stored pages for the past. As a result, it is possible that the various subpages were saved at different dates. This is important for us because we have fixed periods in which we carry out the analysis. With constantly changing points in time for the storage of subpages, the analysis of the individual dimensions of the taxonomy is more difficult, since often the stored subpages are no longer in the period and are therefore unusable. Larger intervals or a different data source could provide a solution to this problem.

For a better comparability of the data, the already known results from the paper by Thiebes et al. (2020) were adapted. All "informed guesses (x)" have been converted to "applicable characteristics" which decreases the quality of the data but allows comparison.

Future Work

Since most of the non-detectable characteristics were in the dimension *Sample storage,* it could be interesting to research why the regulations for sample storage were so non-transparent in the past. Furthermore, there were found just a few changes in dimensions such as *Fee payer* and *Reselling genome data*. An additional research could provide information on why only minor changes in these dimensions have taken place. In addition, there is a field of research on the providers in Cluster four, as here scarcely any changes can be observed over time. It might be interesting to know why there so few changes in this cluster have been. To make more precise statements about the reasons of the changes of the business models, the examination of the remaining services from the work of Thiebes et. al. (2020) could give a bigger overall picture of the history of the DTC genetic testing industry.

Conclusion

The DTC genetic testing market has only developed in the last 15 years and offers several advantages such as fast and affordable genetic testing for consumers. Despite the rapid technical and economic development, many regulatory measures and consumer protection laws have fallen by the wayside, especially in the early days (Allyse et al. 2018). After Thiebes et al. (2020) gave a first insight into the business models of DTC genetic test providers with the help of a specially developed taxonomy, our paper provides further in-depth findings based on past time periods. The taxonomy is applied to business models of up to 55 providers in the periods 2008/2009 and 2013/2014, and the changes between these periods and 2018 are shown. This provides new information about the development of the business models over time and can serve as a basis for further research. The causes of the changes and the exact relationships between the actions of actors in the health care segment and political decision-makers and consumers could be investigated.

References

Alexa - Alexa Internet - About Us 1996 - 2021. (retrieved from: https://www.alexa.com/about; last accessed: July 14, 2020).

Allyse, M. A., Robinson, D. H., Ferber, M. J., and Sharp, R. R. 2018. "Direct-to-Consumer Testing 2.0: Emerging Models of Direct-to-Consumer Genetic Testing," *Mayo Clinic proceedings* (93:1), pp. 113-120

Andrew S. Robertson 2009. "Taking Responsibility: Regulations and Protections in Direct-to-Consumer Genetic Testing," *Berkeley Technology Law Journal* (Vol. 24, No. 1,), pp. 213-243.

Bowen, S. 2018. *Consumer Genetic Testing Is Booming: But What are the Benefits and Harms to Individuals and Populations?* (retrieved from: https://blogs.cdc.gov/genomics/2018/06/12/consumer-genetic-testing/; last accessed: July 12, 2020).

Difference Between DNA Genotyping & Sequencing 2020. https://customercare.23andme.com/hc/en-us/articles/202904600-Difference-Between-DNA-Genotyping-Sequencing; last accessed July 22, 2020.

DNA Sequencing Costs: Data 2019. (retrieved from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data; last accessed: July 21, 2020).

Gleason, C. A., and Juul, S. E. (eds.) 2018. *Avery's diseases of the newborn*, Philadelphia, PA: Elsevier.

Green, R. C., and Farahany, N. A. 2014. "Regulation: The FDA is overcautious on consumer genomics," *Nature* (505:7483), pp. 286-287.

Hoeren, T. 2006. *Die «Wayback-Machine» - Was bringen Internetarchive für die- Recht-Steuern-Wirtschaft*. (retrieved from: https://rsw.beck.de/cms/main?docid=164816; last accessed: July 22, 2020).

Kutz, G. 2010. "GAO-10-847T Direct-To-Consumer Genetic Tests: Misleading Test Results Are Further Complicated by Deceptive Marketing and Other Questionable Practices,"

Morgensztern, D. (ed.) 2018. *IASLC thoracic oncology*, Philadelphia, PA: Elsevier.

Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A method for taxonomy development and its application in information systems," *European Journal of Information Systems* (22:3), pp. 336-359.

Regalado, A. 2018. *2017 was the year consumer DNA testing blew up*. (retrieved from: https://www.technologyreview.com/2018/02/12/145676/2017-was-the-year-consumer-dna-testing-blew-up/; last accessed: July 21, 2020).

Shafer, S. M., Smith, H. J., and Linder, J. C. 2005. "The power of business models," *Business Horizons* (48:3), pp. 199-207.

Sharkey, C. 2019. "Direct-to-Consumer Genetic T o-Consumer Genetic Testing: The FDA's Dual Role as s Dual Role as Safety and Health Information Regulator," *DePaul Law Review*.

Skirton, H., Goldsmith, L., Jackson, L., and O'Connor, A. 2012. "Direct to consumer genetic testing: a systematic review of position statements, policies and recommendations," *Clinical genetics* (82:3), pp. 210-218.

Thiebes, S., Toussaint, P. A., Ju, J., Ahn, J.-H., Lyytinen, K., and Sunyaev, A. 2020. "Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing," *Journal of medical Internet research* (22:1), e14890.

Wayback Machine 2014. (retrieved from: https://web.archive.org/web/20140415000000*/http://23andme.com/; last accessed 20 July 2020).

What is direct-to-consumer genetic testing? 2020. (retrieved from: https://ghr.nlm.nih.gov/primer/dtcgenetictesting/directtoconsumer; last accessed: July 12, 2020)

What is the Human Genome Project? 2018. (retrieved from: https://www.genome.gov/human-genome-project/What; last accessed: July 22, 2020).

History of Business Models: The Case of 23andMe and Ancestry

Emerging Trends in Digital Health, Summer Term 2020

Hepke Kruse
Bachelor Student
Karlsruhe Institute of Technology
hepkekruse@yahoo.de

Maxim Kramsakov
Bachelor Student
Karlsruhe Institute of Technology
maxim.kramsakov@student.kit.edu

Abstract

Background: Initial research results show that the market of Direct-to-Consumer (DtC) genetic testing today is dominated by heterogeneous business models. Thus, a valuable contribution to the lack of knowledge and understanding about business models in this market is already given. Previous research has been limited to three different research areas: First, the medical and ethical value of DTC genetic tests is expatiated in most of the scientific work. Only little research is found about current snapshots of the business models of various DTC genetic test providers as well as the general market development since the beginning of the sector as second and third research areas.

Objective: This scientific work follows on from this and closes the gap between the current snapshots of the business models and the general market development by examining and comparing the concrete evolution of the business models of two selected DTC genetic test providers.

Methods: Employing a systematic literature review, we reconstruct and analyze the development of the business models of 23andMe and Ancestry that represent probably the two greatest players among this market segment.

Results: In the past, 23andMe acted very innovatively with huge objectives and visions and benefited greatly from its role as a pioneer while accepting risks. In addition, 23andMe uses the market of direct-to-consumer genetic testing as a base for two additional pillars, namely drug research and -development as well as research services for pharmaceutical-, biological- and research companies. Ancestry, on the other hand, is more cautious. Major modifications only will be made if Ancestry observes the success of a new strategy at another market participant. Otherwise, Ancestry prefers to stick with efficient and well-functioning business models. In addition, Ancestry remains focused on direct-to-consumer genetic testing as its primary market and does not rely on other pillars in its business model.

Conclusion: The development and modification of business models may be strongly influenced by the corporate philosophy and the self-image of the company. Comparing the two providers shows that the development of the business models over time can proceed very differently despite this small niche market.

Keywords: direct-to-consumer genetic testing, business model, 23AndMe, Ancestry, history, literature review, development

Einleitung

Hintergrund und Zielsetzung

Während der entscheidende Wert von Gentests in der Vergangenheit lediglich auf Forschung und Medizin bei der Nutzung von medizinischen Gentests lag, zieht dieser Sektor mit der Zeit auch zunehmend die privaten Kunden an. Zu Beginn der Zweitausender entwickelte sich rasch ein Markt speziell für Gentests an Privatkunden, die sich ein Test-Set bei einem Anbieter im Internet bestellen und die Resultate anschließend als Bericht zurückerhalten. Zum einen werden damit Stammbaumtests und nicht medizinische Lifestyletests angeboten (ammar 2019; Hudson et al. 2007). Andererseits können aber auch Gentests als Medizintests zum Trägerstatus genetisch bedingter Krankheiten oder zur genetischen Gesundheit ganz ohne ärztliche Anordnung in Auftrag gegeben werden (Hudson et al. 2007). Die wissenschaftliche Validierung und klinischen Nützlichkeit der Testergebnisse (Gurwitz & Bregman-Eschet 2009) solcher Direct-to-Consumer-Gentests (DtC-Gentests) werden oftmals angezweifelt, da sie keine Validierung der Methodik und keine Interpretation der kundenspezifischen Ergebnisse durch einen Mediziner erfuhren (Hudson et al. 2007). Ebenso ist es beim Datenschutz, da die Daten der Kunden für weitere Forschung verwendet werden (ammar 2019). Die DNA-Sequenz jedes Menschen ist einzigartig und somit eine sehr sensible Information, die ausreichend geschützt werden muss. Auf dem deutschen Markt sind DtC-Gentests auf Grund dieses Risikos nicht zugelassen (Borry et al. 2010). Eine Gentestanordnung und -interpretation ist nur durch einen Mediziner gestattet. Trotz kontroverser Diskussionen, vielerlei Kritik in der Vergangenheit und dem zeitweisen Verbot zur Ausführung von medizinischen DtC-Gentests (Green & Farahany 2014) gibt es heute in den USA eine Vielzahl an DtC-Gentest-Anbietern, deren Geschäfts-modelle auf verschiedene Schwerpunkte spezialisiert sind. So fokussiert sich der Anbieter Ancestry auf Ahnenforschung, 23andMe auf charakteristische Eigenschaften, die auf bestimmte Gene zurückzuführen sind wie beispielsweise das Alzheimerrisiko oder das Risiko für Höhenangst. Auch bei den Zahlungs- und Datenschutzmodellen sowie den Zielgruppen gibt es einige Unterschiede (Thiebes et al. 2020). Insbesondere hinsichtlich des Datenschutzes hat sich durch den Weiterverkauf kundenspezifischer genetischer Daten eine Vielzahl neuer gesetzlicher Regularien entwickelt, die einen erheblichen Einfluss auf den Markt für DtC-Gentests haben.

Die bisherige Forschung beschränkte sich entweder auf den medizinischen und ethischen Wert solcher DtC-Gentests oder auf aktuelle Momentaufnahmen der Geschäftsmodelle verschiedener DtC-Gentest Anbieter und auf die grundsätzliche Marktentwicklung seit Anbeginn des Sektors. Diese wissenschaftliche Arbeit schließt daran an und untersucht und vergleicht die konkrete Entwicklung der Geschäftsmodelle zwei ausgewählter DtC-Gentest Anbieter. Das erlangte Wissen vermag hoffentlich an vielen Einsatzstellen einen Mehrwert liefern. Zum einen besitzt die Betrachtung der historischen Geschäftsmodelle zweier Anbieter das Potential, der allgemeinen Bevölkerung ein differenzierteres Bild für die Chancen und Risiken sowie die Vor- und Nachteile und auch Erfolge- und Misserfolge vergangener Geschäftsmodelle von DtC-Gentest Anbietern zu vermitteln. Selbiges gilt ebenso für aktuelle DtC-Gentest Anbieter und potenzielle zukünftige Anbieter, die einen Markteintritt in Erwägung ziehen und ihre Produktstrategie festlegen. Schließlich soll diese Arbeit als Hilfestellung für gesetzliche Regularien zu dienen, um vergangene gesetzliche Einschränkungen in Bezug auf die Geschäftsmodelle besser zu verstehen und auch zukünftig adäquat an neue Trends anpassen zu können. Neben den praxisorientierten Zielen möchten wir ebenfalls einen Betrag für die Wissenschaft und weitere Forschungen leisten, da es grundsätzlich wenig Forschung in diesem Gebiet gibt. Zumal wir dem Aufruf einer wissenschaftlichen Abhandlung mehrerer Forschungsexperten dieses Fachgebiets folgen (Thiebes et al. 2020), die Veränderung der Geschäftsmodelle verschiedener DtC-Gentest Anbieter zu untersuchen, nachdem sie selbst eine Momentaufnahme der aktuellen Geschäftsmodelle erstellt haben. Unsere wissenschaftliche Arbeit soll dieser Momentaufnahme mehr Verständnis und Hintergrundinformationen aufzeigen, auf welchen Schritten und Meilensteinen die Anbieter sich dorthin entwickelt haben. Dazu ist die Vorgehensweise unserer Arbeit, anhand eines systematischen Literaturreviews zunächst die historische Entwicklung der Geschäftsmodelle zweier Anbieter zu rekonstruieren und zu analysieren und diese als zweiten Schritt im Vergleich einander gegenüberzustellen. Folgende Ergebnisse resultieren aus unserer Arbeit: 23andMe handelt als agiler, wandelnder Innovator mit mehreren Geschäftsbereichen und weist eine andere Entwicklung seines Geschäfts auf als Ancestry, welches sich als zurückhaltendes, auf Sicherheit basierendes Unternehmen präsentiert, das sich allein auf den Direct-to-Consumer-Gentest Markt fokussiert. Wir wollen verdeutlichen, dass selbst in diesem kleinen Marktsektor die Entwicklungen der Geschäftsmodelle bis zum

heutigen Stand ganz unterschiedlich verlaufen können und daher auch die heutigen Cluster ihre Begründung erhalten.

Definition Direct-to-Consumer Gentests

Als Direct-to-Consumer (DtC) Gentests werden meist im Internet käufliche Gentests bezeichnet, die ohne Zutun eines Arztes erfolgen. Anders als bei ärztlichen Gentests ist weder bei der Durchführung noch bei der Interpretation der Ergebnisse ein medizinischer Fachmann anwesend (Thiebes et al 2020). Ein Interessent kann bei einem Anbieter, wie beispielsweise „Ancestry" oder „23AndMe", einen Gentest erwerben und erhält daraufhin einen Behälter zugesendet, den er mit einer Speichelprobe gefüllt zurücksenden muss. Nach mehreren Wochen erhält der Interessent die Ergebnisse des Tests, darunter medizinische Fakten, wie beispielsweise genetisch bedingte Risiken an bestimmten Krankheiten zu erkranken, aber auch rein unterhaltende Ergebnisse, wie „Mückenstichhäufigkeit" oder „Abneigung zum Geschmack von Koriander". Je nach Anbieter erhält man überwiegend unterhaltende oder überwiegend medizinische Ergebnisse, es gibt aber auch solche die sich primär auf Ahnenforschung fokussieren. Den Startschuss für den mittlerweile internationalen Trend DtC-Gentests gab das Unternehmen „23AndMe" mit der Eröffnung ihres Online-Shops im Jahr 2007 (Allyse, Robinson, Ferber, und Sharp 2018). Doch ebenso, wie sich DtC-Gentests im Laufe der Zeit unter Interessenten an Beliebtheit erfreuten, warfen sie bei Verbraucherschützern und der Regierung große Unsicherheiten auf (Hogarth, Javitt, und Melzer 2008). Kritiker haben früh erkannt, dass die Ergebnisse der Tests ungenau sind. Durch die Abwesenheit medizinischer Aufsicht kann es zu einer Fehlinterpretation der Ergebnisse kommen, wodurch Interessenten möglicherweise teure und nicht notwendige medizinische Untersuchungen durchführen ließen (Covolo, Rubinelli, Ceretti, und Gelatti 2015). Nicht nur diese mögliche Fehlinterpretation, sondern auch der Schutz der Genomdaten im Rahnem von DtC-Gentests mussten unter behördliche Aufsicht gestellt werden. Infolgedessen mussten sich viele Anbieter auf behördliche Regulationen einrichten, welche direkte Auswirkungen auf ihre Geschäftsmodelle hatten.

Definition Geschäftsmodelle

Obwohl das Forschungsfeld „Geschäftsmodell" fortwährend intensiv bearbeitetet wird, hat sich noch keine allgemein akzeptierte Definition für Geschäftsmodelle ergeben (Fielt 2013). Fakt ist je-doch, dass jedes Unternehmen ein Geschäftsmodell hat, ungeachtet davon, ob es klargestellt wurde oder nicht (Fielt, 2013). Oft wird dazu geneigt, das Geschäftsmodell als eine Art Strategie zu sehen, nach welcher ein Unternehmen agiert; Vielmehr ist die Strategie Teil des Geschäftsmodells (Onetti et al. 2012). Heute existiert eine Vielzahl von Definitionen, welche oft ähnliche Eigenschaften aufweisen bzw. die gleichen Aspekte eines Modells beschreiben. Durch die Synthese einiger Definitionen nach Shafer et. al (2005) konnten aus anerkannten wissenschaftlichen Arbeiten vier zentrale Komponenten von Geschäftsmodellen herausgearbeitet werden:

- Strategische Entscheidungen (strategic choices)
- Wertschöpfung (creating value)
- Wertnutzung bzw. -erfassung (capturing value)
- Wertschöpfungsnetzwerk (Value network)

Des Weiteren konnten Shafer et. al (2005) durch das Zusammenführen der vier Eigenschaften und unter Einbeziehung aller vorheriger Definitionen eine allgemeine Definition für Geschäftsmodelle auf-stellen:

Ein Geschäftsmodell ist die Repräsentation der der Firma zugrundeliegenden Kern-Logiken und strategischen Entscheidungen zur Wertschöpfung und -Nutzung innerhalb eines Wertschöpfungsnetz-werks. Hierbei bezeichnet Kern-Logik Unternehmenswerte und Vorgehensschemata, die die generelle Richtung aller Unternehmensentscheidungen angibt. Ein Beispiel hierfür wäre eine Firma, die sich ausschließlich auf Bio-Produkte fokussiert und deshalb auch bei allen Entscheidungen einen „grünen" Weg wählt. Die Kern-Logik bildet also die Grundlage für alle strategischen Entscheidungen. Das Wertschöpfungsnetzwerk ist hierbei das Feld, in welchem das Unternehmen agiert. Bei DtC-Gentest Anbietern handelt es sich meistens um das Internet. Hier ist der Wettbewerb größer und die Wichtigkeit eines innovativen Geschäftsmodells nimmt zu (Thiebes et al. 2020). Es ist also gerade im Internet wichtig, sich von der Konkurrenz abzuheben (Shafer et al. 2005), um nicht im Meer der DtC-Anbieter zu versinken. Im Rahmen dieser Arbeit betrachten wir Geschäftsmodelle nach der oben vorgeschlagenen Definition

durch Shafer et. al (2005) und zeigen, wie wichtig das Geschäftsmodell für den Erfolg eines Anbieters ist und welche Auswirkungen externe Faktoren darauf haben.

Methodik

Vorgehen zur Auswahl Geeigneter Anbieter

Bevor wir unsere Arbeit zur Analyse der historischen Geschäftsmodellentwicklung verschiedener Anbieter beginnen konnten, musste eine zentrale und entscheidende Fragestellung beantwortet werden. Zunächst gilt es zu klären, welche Anbieter grundsätzlich in den Fokus unserer Arbeit gerückt werden. Wir entschieden uns aus persönlichem Ermessen für die zwei Anbieter Ancestry und 23andMe. Beide Anbieter sind erfolgreich auf dem Markt, sind große Anbieter und besitzen insbesondere ganz verschiedene Marktschwerpunkte. Ancestry fokussiert sich auf die Ahnenforschung, 23andMe neben der Ahnenforschung mehr auf gesundheitsbasierende Gentests. Diese Tatsache ermöglicht uns eine gute Vergleichbarkeit.

Diese wissenschaftliche Arbeit soll in Form eines systematischen Literaturreviews durchgeführt werden. Sowohl für den Anbieter Ancestry als auch 23andMe lief die Analyse der Historie der Geschäftsmodelle parallel ab. Dazu nutzten wir den Ansatz zur effektiven Literaturrecherche von Levy & Ellis (2006), welcher Aufschluss über das Sammeln (Input) und Verarbeiten der Literatur sowie über das Verfassen der Arbeit (Output) gibt.

Methodik bei Anbieter 23andMe: Vorgehen des Literaturreviews

Input:

Der Suchstring für 23andMe lautete „abstract (Direct-to-Consumer genetic testing OR 23andMe) AND abstract (business model*)“. Bei der Auswahl der vier Datenbanksysteme PubMed, Proquest, ScienceDirect und Ebscohost wurde auf eine Mischung von medizinisch-technischen und wirtschaftlichen Datenbanken wert gelegt. Damit erwarten wir uns eine vielfältigere Menge an Quellen, da unsere fokussierten Begriffe bzw. Hauptthemen Geschäftsmodelle und DtC-Gentests aus jenen Wissenschaften hervorgehen.

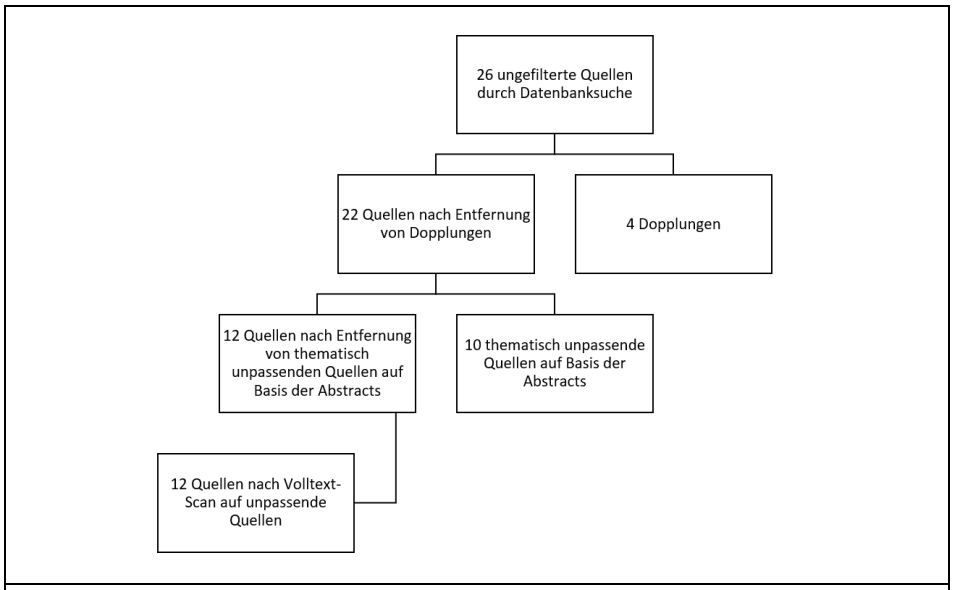

Abbildung 1. Filtern der Literatur zu 23andMe

Für 23andMe erhielt man insgesamt 26 Quellen (Vgl. Anhang A) in den vier Datenbanken zusammengenommen. Nach Entfernung von Dopplungen blieben 22 Quellen. Beim Lesen der Abstracts konnten weitere 10 Quellen ausgeschlossen werden. Größtenteils sind sie von unseren thematischen Schwerpunkten abgewichen, indem der Fokus nicht auf den Geschäftsmodellen oder den von uns

untersuchten Anbietern lag. In diesem Sinn wurden unter anderem zwei Quellen ausgeschlossen, da sie vor dem Gründungsjahr von 23andMe veröffentlicht wurden. Somit können diese Quellen zwar den DtC-Gentest Markt beleuchten, aber keine Aussage über das Geschäftsmodell von 23andMe treffen und sind daher für uns nicht von Relevanz. Im Anschluss wurden die Volltexte der übrigen 12 Quellen auf Relevanz und Eignung überprüft. Es konnten allerdings keine weiteren Quellen ausgeschlossen werden. Stattdessen wurden über eine Rückwärtssuche die in unseren Quellen relevanten zitierten Quellen in unsere Betrachtung einbezogen, auch wenn dies den Einbezug von nicht-wissenschaftlicher Literatur bedeutet.

Verarbeitung:

Die wichtigen, aussagekräftigen Informationen der Quellen über die Geschäftsmodelle des DtC-Gentest Anbieters 23andMe wurden kleinschrittig in den Verlauf einer zeitlichen Abfolge eingefügt.

Output:

Als Ergebnis erhalten wir eine Darstellung des historischen Verlaufs des Geschäftsmodells von 23andMe, welcher nach der Darstellung von Ancestry einen Vergleich der zwei Anbieter ermöglicht.

Methodik bei Anbieter Ancestry: Vorgehen des Literaturreviews

Für Ancestry wurden die gleichen Datenbankensysteme mit einer Vielzahl verschiedener Suchstrings durchsucht, die aber nicht die erwünschten Ergebnisse geliefert haben. Die gezielte Suche nach dem Unternehmen „Ancestry.com“ oder „Ancestry, Inc.“ konnte zwar verhindern, dass englische Artikel zur Ahnenforschung auftauchten, lieferte jedoch nur die Einführungsliteratur und wenige andere Arbeiten als brauchbare wissenschaftliche Quellen. Durch „Backward-“ und „Forward-Search“ jedoch konnten vielversprechende und überwiegend nicht-wissenschaftliche Quellen gefunden werden, die viel Inhalt zur Historie des Geschäftsmodells hergeben. Dazu haben wir uns die Quellen aus der Einführungsliteratur und aus der wenigen gefundenen Literatur angeschaut und nach der Qualität bewertet. Nach langer Qualitätsanalyse konnten wir die Historie aus 8 Quellen zusammensetzen.

Ergebnisse

Historie des Geschäftsmodells von 23andMe – ein Dreiteiliges Geschäftsmodell

Im Jahr 2007 öffnete das Startup 23andMe seinen Webstore (Turrini, 2018) und bot ihrer Zielgruppe, den Enthusiasten (Borry et al. 2010), die aus reiner Neugierde Interesse an den Gentests hegen, persönliche Genom-Dienstleistungen an mit dem Ziel, die Forschung des Gesundheitswesens zu revolutionieren (Hogarth 2017). Zu den Dienstleistungen von 23andMe gehören neben Berichten über die genetische Abstammung auch persönliche Informationen über den Trägerstatus, das Wohlbefinden und Gesundheitseigenschaften (Whaley & McGuire 2018). Dazu führt 23andMe einen ganzheitlichen Genomscan durch (Gurwitz & Bregman-Eschet 2009). Viele Wettbewerber auf dem Markt verwenden monogenetische Tests, d.h. sie testen auf gesundheitliche Risiken, die nur auf ein einziges Gen als Verursacher zurückzuführen sind. Ein ganzheitlicher Genomscan hingegen ermöglicht durch die sogenannte Einzel-Nukleotid-Polymorphie-Technologie (Whaley & McGuire 2018) multigenetische Tests. Das Auftrittsrisiko einer zu untersuchenden gesundheitsbasierenden Eigenschaft wird nicht nur durch ein Gen, sondern durch das Zusammenspiel mehrerer Gene hervorgerufen (Tutton & Prainsack 2011), welches mit dieser Technologie auf komplexe Art getestet wird. Damit hebt sich 23andMe direkt von Beginn an von dem Großteil der Wettbewerber ab. Die bisherige Anzahl an enthusiastischen Kunden war jedoch für eine fortlaufende Sicherung des Unternehmensbestands nicht ausreichend (Borry et al. 2010). Daher teilte 23andMe seinen Service Ende 2009 in zwei einzelne Dienstleistungen auf (Tutton & Prainsack 2011), genetische Abstammung und die Gesundheitstests. Man erwartete, dass nur an Abstammung interessierte Menschen 23andMe als Anbieter auswählen, wenn sie eine darauf zugeschnittene Dienstleistung erhalten. Mit dieser Strategie wurden bis Juni 2010 insgesamt 50.000 Kun-den genetisch untersucht. Im November 2010 führte man die beiden Dienstleistungen wieder zusammen und begann mit einem monatlichen Abonnement-Preis.

Dennoch sorgte der gesamte DtC-Gentest Markt bei der allgemeinen Bevölkerung sowie Experten in der Genforschung für eine kontroverse Diskussionen und viel Kritik (Hudson et al. 2007). Man war besorgt in Bezug auf die Glaubhaftigkeit der Tests, die Sicherheit der DNA-Verwendung, die Privatsphäre genetisch

riskanter Informationen und ein mangelndes Vertrauen in nicht-persönliche Beratung (Wilde et al. 2010). Die Experten forderten neben analytisch-wissenschaftlicher Validierung auch die klinische Validierung und Verwendbarkeit der Tests (Borry et al. 2010) und befürchteten Fehlinterpretationen der Tests bei Kunden (Allyse et al. 2018). Um diese Punkte zu analysieren und entsprechende Regularien zu verschärfen, begann im Jahr 2010 die FDA ihre Nachforschungen. Noch im gleichen Jahr erhielt 23andMe eine Verwarnung der FDA, dass der Verkauf von gesundheitsbasierten Gentests als medizinische Hilfsmittel ohne Genehmigung der FDA vor Markteintritt rechtlich nicht gestattet ist (Spector-Bagdady 2016). 23andMe hingegen sah seine Tests zu gewissen Gesundheitsrisiken und zu Trägerinformationen von Beginn an nicht als medizinischen Diagnosebericht, sondern nur als Informationen zu Bildungszwecken, welche rechtlich keine Genehmigung der FDA benötigten (Tutton & Prainsack 2011). Dennoch verkündete 23andMe im Jahr 2012 das Ziel, eine Million Nutzer zu sammeln (Wojcicki 2012) und reichte als erster DtC-Gentest Anbieter die geforderten, vor Markteinführung abzugebenden Papiere ein. Gleichzeitig startete 23andMe eine Werbekampagne eines noch nicht autorisierten Produktes, was die FDA im Jahr 2013 mit einer formalen Unterlassungsanordnung der gesundheitsbasierten Tests von 23andMe quittierte (Spector-Bagdady & Pike 2014). Folglich musste 23andMe sein Dienstleistungsangebot für Kunden dann auf die Abstammungsanalyse beschränken. Die Anzahl der Verkäufe sank stark. Schon seit 2011 gab es starke Preissenkungen von anfangs 999$ über Zwischenschritte auf 99-199 im Jahr 2012, was die Datenbank auf 40.000 Kundenprofile verdoppeln ließ (Brady 2013).

Als Chance nutze 23andMe parallel die Zusammenarbeit mit der FDA zur Bearbeitung der vormarktlichen Freigabebestimmungen (Spector-Bagdady 2016). Es scheint, als würde 23andMe die Autorität der FDA zu akzeptieren und die Sichtweise über ihre gesundheitsbasierten Gentests geändert zu haben. Zuvor wurde beharrlich daran festgehalten, dass ihre gesundheitsbasierten Gentests ausschließlich zu Informationszwecken und nicht zur medizinischen Diagnostik dienen (Tutton & Prainsack 2011). Stattdessen stieg nun die medizinische Integration in die Dienstleistung von 23andMe, denn im Jahr 2015 bekam 23andMe nach einigen Validierungsstudien die offizielle Genehmigung der FDA für den Trägertest des autosomal rezessiven Bloom-Syndroms (23andMe 2015). Somit war 23andMe der erste DtC-Gentest Anbieter, dessen Berichte die Standards der FDA zur klinischen und wissenschaftlichen Validierung einhielten (Spector-Bagdady 2016). Das Marketing weiterer autosomaler Trägertests wurde von der FDA ohne vormarktliche Autorisierung toleriert, sodass 23andMe wieder einen Teil ihrer gesundheitsbasierenden Tests anbieten durfte. Die Anforderungen der FDA, die zur Genehmigung erfüllt werden mussten, inkludieren sowohl eine analytische Validierung als auch die Sicherstellung des Nutzerverständnisses, dass dieser die Ergebnisse korrekt interpretieren kann. Im April 2017 erteilte die FDA schließlich die Genehmigung, einen Test für 10 polygenetische Risikofaktoren auf den Markt zu bringen (Saukko 2017), den sogenannten „23andMe's Personal Genome Service" (Allyse et al. 2018). Zu den zehn zu testenden Risiken gehörten unter anderem Alzheimer und Parkinson. Damit erlaubte die FDA die ersten DtC-Gentests als zugelassene Medizinprodukte. Dafür musste zu jedem Testbericht allerdings eine Reihe an Informationen zum Verständnis des Kunden beigelegt sein (Janssens 2019). Weitere Gegebenheiten waren von der vormarktlichen Untersuchung befreit (Allyse et al. 2018), wenn sie sich an die Anforderungen der neu entwickelten Genetic-health-risk-Kategorie hielten, und schloss medizinische Diagnosetests für größere, schwerwiegendere Fälle eindeutig aus, ebenso wie Tests auf Krebs oder Pränatal-Tests. Die Erfahrungen mit 23andMe als Vorreiter sollten der FDA für klarere Regularien in der Zukunft dienen (Allyse M 2018).

Genau diese Rolle des innovativen Vorreiters möchte 23andMe auch auf dem Markt vertreten. Bereits zu Gründungszeiten äußerte 23andMe die Vision, die Forschung des Gesundheitswesens zu revolutionieren (Hogarth 2017). Es besteht kein starres Konzept, an dem sich festgehalten werden muss. Stattdessen reagiert 23andMe flexibel und agil auf Änderungen äußerer Gegebenheiten und passt sich schnell an. Der ideale Beweis zur Bereitschaft des agilen Wechsels der Geschäftsstrategie ist die bereits erwähnte Zusammenarbeit mit der FDA. Zuvor wurde aus Sicht von 23andMe ausschließlich mit nicht medizinischen Tests gearbeitet. Durch die Kollaboration mit der FDA steigt 23andMe in den Markt der medizinischen Tests ein und ist damit Pionier auf dem Markt. Die Marktstrategie lautet, vor anderen Wettbewerbern neue Dienstleistungen anzubieten, inklusive der Genehmigungen (Whaley & McGuire 2018) und die Kunden mit andersartigen, neuen Dienstleistungen für sich zu gewinnen.

Ein weiterer wichtiger Aspekt in dem Geschäftsmodell von 23andMe ist der Umgang mit Kundendaten. Neben dem DtC-Gentests Angebot an Kunden, hatte das Geschäftsmodell von Beginn an auch das langfristige Ziel, mit Hilfe der Kunden eine riesige Bio- und Datenbank zu erstellen und diese für

Forschungs- und Entwicklungszwecke sowohl intern in dem Unternehmen als auch extern zu nutzen. Dazu sichert sich 23andMe die Rechte an den Speichelproben (Tutton & Prainsack 2011). Die Rechte an grundsätzlichen Kundendaten bleiben weiterhin beim Kunden, dürfen jedoch für Forschungszwecke genutzt werden (Spector-Bagdady 2016). Gleichzeitig werden Kunden oft dazu aufgefordert, freiwillig an Umfragen teilzunehmen und weitere Daten über sich der Forschung von 23andMe, aber auch von dritten Parteien zur Verfügung zu stellen. Dieser Fakt ermöglicht 23andMe noch viel mehr Geschäftsmöglichkeiten, als der alleinige Verkauf von DtC-Gentest mit strengeren Datenschutzregelungen es könnte. Zum einen ermöglicht es den Verkauf von Zugang zu den Daten für klinische Forschungsorganisationen und biopharmazeutischen Unternehmen (ammar 2019). So hat 23andMe Zugriffsvereinbarungen mit 30 pharmazeutischen und biotechnologischen Unternehmen, neben Forschungs- und non-profit-Organisationen (Servick 2015). Dazu gehört eine Kooperation mit dem Pharmakonzern Pfizer ab 2015 (Ratner 2014). 23andMe analysiert Kundendaten auf gewisse Immunerkrankungen wie Morbus Crohn und gibt diese dann an Pfizer weiter. Neben der Vertragsforschung mit Pfizer existiert unter anderem auch ein 60Mio$ Deal mit dem Unternehmen Genentech über die vollständige Genomsequenzierung (Spector-Bagdady 2016). Hinzu kommen die Zusammenarbeit mit Forschungsinstitutionen (Spector-Bagdady 2016) wie der Universität Chicago im Jahr 2013 und staatliche Forschungsgelder unter anderem im Jahr 2014 von den National Institutes of Health. 23andMe verkauft folglich also nicht nur Zugriff auf die Kundendaten, sondern nutzt sie auch für eigene Forschungsbeiträge. Der Zugriff auf solch eine große Datenbank wie diejenige von 23andMe kann die Art verändern, wie Pharmaunternehmen Arzneien entwickeln und Versicherungsunternehmen Lebensrisiken bewerten (Fisk 2014). Diese Chance sieht auch 23andMe und richtet sich eine eigene Abteilung zur Arzneimittelentwicklung ein, um in diesen Markt einzutreten (Hogarth 2017). Tatsächlich ist die unvermutete Haupteinnahmequelle von 23andMe sogar die Vertragsforschung und -entwicklung für die Pharmaunternehmen (Hogarth 2017).

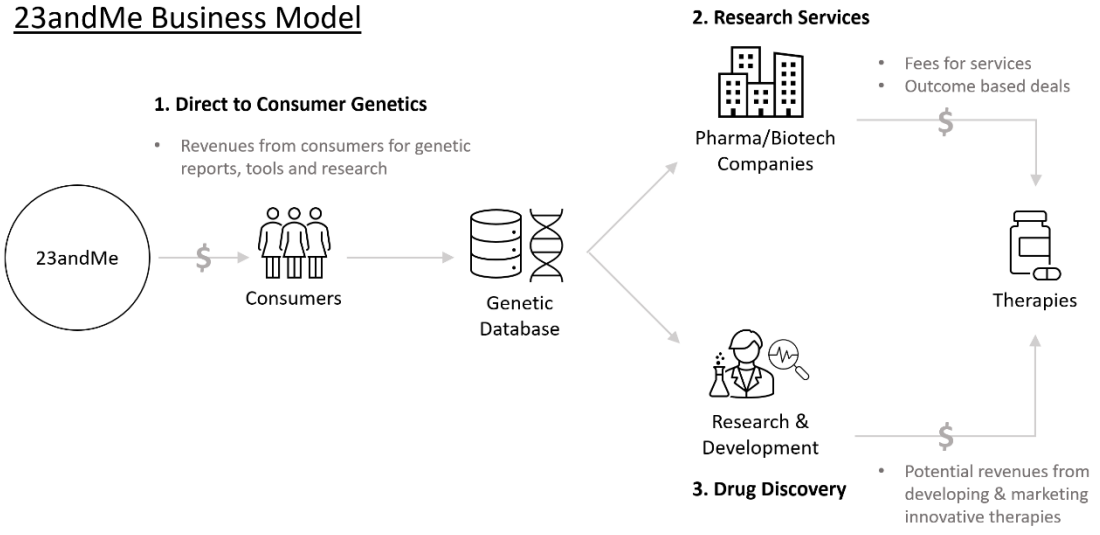

Abbildung 2. 23andMe Business Model (Whaley & McGuire, 2018 Zitiert von www.revenuesandprofits.com, 2017)

In jedem Fall hat 23andMe es erfolgreich geschafft, sich eine Vielzahl von Einkommenssträngen aufzubauen, seien es Kundenerträge, staatliche Forschungsgelder, Vertragsforschung für Pharmaunternehmen oder auch die Arzneimittelentwicklung. Dies lässt Rückschlüsse auf das Geschäftsmodell ziehen. Das Anbieten von DtC-Gentests ist nicht der alleinige Geschäftszweig im Businessplan von 23andMe, sondern besteht zusammen mit Vertragsforschung für Pharma- und Biounternehmen und der Arzneimittelentwicklung aus einem dreiteiligen Geschäftsmodell.

Es wird auch von der Unterscheidung zwischen einem Frondend-Business, den DtC-Gentests an Kunden, und einem Backend-Business als größte genetische Bio- und Datenbank der Welt gesprochen (Spector-

Bagdady 2016). Diese Aufteilung des Geschäftsmodells in drei Bereiche hat einen prägnanten Vorteil für 23andMe, denn Multi-Produkt-Strategien reduzieren häufig Marktrisiken (Whaley & McGuire 2018). Trotz getrennter Betrachtung der drei Geschäftsstandbeine, sind diese dennoch miteinander verwoben und üben gegenseitige Einflüsse aufeinander aus, sind also abhängig voneinander (Whaley & McGuire 2018). Alle drei Standbeine leben von dem Hinzukommen von neuen genetischen Daten und der Kunde bleibt zentraler Bestandteil des Geschäftsmodells von 23andMe. Auch für die Forschung ist das Erhalten von immer weiteren Datensätzen von Bedeutung, sodass eine Vielzahl an Kunden generiert werden muss. Das Vorantreiben des Forschungsbereichs könnte auch einen potenziellen Einfluss auf die stark gefallenen Preise für DtC-Gentests gehabt haben. Bei den preiswerten Tests von 23andMe können viele Wettbewerber nicht mithalten, sodass die Kundschaft sich für 23andMe entscheidet. Dies trägt direkt zur erhöhten Generierung von Kundendaten bei (Spector-Bagdady 2016). Die niedrigen Preise würden die Kosten wahrscheinlich nicht decken. Stattdessen besitzt 23andMe ein externes Venture-Capital-Fond-Netzwerk. Zu den Sponsoren gehört unter anderem Google (Whaley & McGuire 2018).

Seit 2017 konnten keine größeren Veränderungen des Geschäftsmodells von 23andMe festgestellt werden. Das aktuelle Geschäftsmodell entspricht demnach weiterhin dem beschriebenen dreigeteilten Modell. Dennoch existiert kein klares, eindeutiges Geschäftsmodell. (Hogarth 2017). Einerseits agiert 23andMe im Business-to-Business-Markt mit den Pharmaunternehmen und anderseits aber auch im Business-to-Consumer-Markt mit den DtC-Gentests an Kunden. Bezüglich dieses Aspekts eines Geschäftsmodells ist 23andMe nicht eindeutig. Dennoch basiert ihr gesamtes Geschäftsmodell auf der Basis zuwachsender Kundenzahlen und deren Bereitschaft, ihre genetischen Daten zur Verfügung zu stellen. Im Prinzip entspricht das Verfolgen diverser Geschäftsstränge genau der zuvor beschriebenen Flexibilität und Agilität, die 23andMe als innovativen Vorreiter ausmachen.

Historie des Geschäftsmodells von Ancestry

1983 wurde „Ancestry Publishing“, als Herausgeber für Bücher, Zeitschriften und Newsletter zu Familiengeschichte und Genealogie gegründet. Diese waren nicht zwingend an Genealogen oder andere Fachmänner gerichtet, sondern an Enthusiasten und Amateure der Familienforschung. Der heutige „Kern“-Wert des Unternehmens, dass das Erforschen der eigenen Geschichte in eigener Hand liegt, lassen sich also bis zur Gründung zurückdatieren.

1996 wurde Ancestry dann durch den Tech-Startup „Infobases“ übernommen und die Internetseite „Ancestry.com“ wurde ins Leben gerufen (Groo 2020). Ancestry.com entwickelte sich zu einer Komposition familienhistorischer Datenbanken, bei der Kunden gegen Abonnement-Zahlungen Zugriff auf diese erhalten und damit Ahnenforschung betreiben können. Bis 2000 hatte Ancestry.com noch viele kleinere Datenbanken aufgekauft, zusammengeführt und mehrere Webseiten zur Familienforschung online gestellt, die nicht mehr nur in der U.S.A verfügbar waren. Zu diesen Webseiten gehörten beispielsweise MyFamily.com oder Familyhistory.com, welche 2009 aber wieder zu Ancestry.com zusammengeführt worden sind.

2012 wurde Ancestry.com für 2.6 Mrd. USD von mehreren privaten Investoren, einer sogenannten Private-Equity-Gesellschaft, aufgekauft. Im gleichen Jahr wurde auch AncestryDNA veröffentlicht.

Mit AncestryDNA können Interessenten einen Speicheltest durchführen und erhalten nach der Auswertung Informationen über ihre ethnische Herkunft. Dann besteht die Möglichkeit sich ggf. mit weit entfernten Verwandten in Verbindung zu setzen. 2012 scheint dafür spät zu sein, wenn man bedenkt, dass der wahrscheinlich größte Konkurrent, 23AndMe, ähnliche DNA-Test-Sets bereits vor 5 Jahren veröffentlicht hat. Ancestry hat aber einen entscheidenden Vorteil, welcher das Unternehmen zum größten und mächtigsten Anbieter von DNA-Tests zur Ahnenforschung macht; seit der Gründung hat Ancestry massenweise Daten über Familiengeschichte gesammelt, ganze Dekaden lang konnten Besucher der Website selbst Stammbäume zusammenstellen und Familienangehörige suchen. Darüber hinaus hat sich Ancestry über die Zeit auch Zugriff zu einer massiven Zahl an Archiven und Datenbanken zur Familienforschung verschafft und digitalisiert.

Nach nur drei Jahren hatte der AncestryDNA-Test über 1 Million Kunden (Phillips 2016). Der entwickelte Ahnenforschungs-DNA-Test war also das perfekte Komplement zur historischen Familiendatenbank von Ancestry. Der Ahnenforschungs-Test hat dank der Datenbank nicht nur umfassendere Ergebnisse, als die der Konkurrenz liefern können, sondern die Datenbank auch erweitert und somit für künftig noch genauere

Ergebnisse gesorgt. Ancestry hat mit AncestryDNA ein sich selbst stabilisierendes Produktmodell erschaffen, welches sich auf Langfristigkeit fokussiert. Mit Werbesprüchen wie „Entdecke deine Familiengeschichte“ oder „eine Familie die darauf wartet entdeckt zu werden“ und dem vergleichsweise niedrigen Preis, reduziert Ancestry die Distanz zur eigenen Geschichte und richtet sich somit primär an Hobby-Forscher, die ohne viel Aufwand möglichst viel über ihre Vergangenheit in Erfahrung bringen möchten (Thiebes et al. 2020). Für Ancestry ist es wichtig, mit ihrem Marketing möglichst viele Kunden für sich zu gewinnen, da der Kunde fester Bestandteil ihres Geschäftsmodells wurde. Dabei ist es egal, ob die Kunden den DNA-Test nutzen oder nur Zugriff auf die Datenbank abonnieren, um nach ihren Ahnen zu Forschen; Wichtig ist, dass jeder Kunde Daten liefert, um die Datenbank von Ancestry zu erweitern. Nur dann kann für die Zukunft garantiert werden, dass die Ergebnisse ihres Tests aktuell bleiben und Ancestry ihre Marketing-Versprechen halten kann. Die Einordnung ihres Tests im Billig-Preis-Segment resultierte zwar in weniger umfangreichen und nicht absolut verlässlichen Tests, lockte natürlich aber viel mehr Kunden an als ein teures Produkt.

Im Juli 2015 wagte sich Ancestry an die Veröffentlichung des gesundheitlichen DNA-Tests „AncestryHealth“ und versuchte dabei nicht auf dieselben Hürden wie 23AndMe zu stoßen. Zu seinen Anfängen war AncestryHealth nur ein Umfrage-Tool zur eigenen Gesundheit (Phillips 2016), entwickelte sich aber schnell zu einem vollwertigen medizinischem Gen-Test. Ancestry wählte für ihren Test jedoch einen völlig anderen Ansatz als die Konkurrenz. Die AncestryHealth-Tests werden durch einen Arzt bestellt, welcher für PWNHealth arbeitet, ein Netzwerk aus Ärzten, gegründet im selben Jahr, in dem die Tests veröffentlicht wurden. Sicherlich kein Zufall, denn Ancestry hatte damit einen entscheidenden Trumpf. Die Tatsache, dass ein Arzt den Test verordnet, lässt keine Regulierung der FDA ähnlich wie bei 23AndMe zu, da es sich in diesem Sinne nicht mehr um einen DtC-Test handelt. Obwohl der Test qualitativ keine besseren Ergebnisse liefert als die Tests der Konkurrenz, konnte Ancestry mit der Entscheidung, einen Arzt einzubeziehen, viel Kritik und Rechtsstreite umgehen, welchem derartige Tests seit 2007 zum Opfer fallen. Es scheint fast so, als wäre es Teil ihres Geschäftsmodells die Konkurrenz zu beobachten, ihre Fehler zu erkennen und beim Aufstellen ähnlicher Produkte diese Fehler zu umgehen. Ihr Gesundheitstest hat sich jedoch nie der gleichen Beliebtheit erfreut wie der Ahnentest. Gründe dafür sind der massive Erfolg im Ahnenforschungswesen, aber auch der Name des Unternehmens den viele Personen mit Ahnenforschung assoziieren.

Heute im Jahr 2020 hat sich bei Ancestrys Geschäftsmodell kaum etwas verändert. Das ist auch keineswegs nötig – Ancestry hat sich damit mittlerweile zum Marktführer für Ahnenforschungs-Tests erhoben. Sie werben immer noch damit, den „Forscher“ im Kunden zu wecken und die Ahnenforschung als Reise in die Vergangenheit zu bezeichnen. Auf ihrer Website trifft man auf Sprüche wie „Flucht, Abschied, Neuanfang, Liebe - Deine Vorfahren machen dich zu der Person, die du heute bist“, welche die eigene Vergangenheit aussehen lässt wie einen Roman, der darauf wartet, gelesen zu werden. Die Geschichte über die Vorfahren vieler Personen ist natürlich nicht derart mit Spannung gefüllt, wie Ancestry es anpreist, aber schon in der Vergangenheit hat sich gezeigt, dass viele Enthusiasten sich gerade wegen des Marketings für Ancestry entscheiden. Ancestry hat heute die größte Sammlung an Aufzeichnungen zur Familiengeschichte mit über 27 Billionen Einträgen, die bis ins 13. Jahrhundert zurückreichen. Jeden Tag kommen, laut eigener Aussage, zwei Millionen weitere Datensätze hinzu; Dadurch nimmt auch der AncestryDNA-Test stetig an Informationsgehalt zu. Heute gibt der Test weit mehr her als die bloße Information über die Abstammung. Er zeigt die mögliche Wanderbewegung der Vorfahren an und zeigt mögliche historische Ursachen für diese.

Ancestry hat über 1700 Mitarbeiter weltweit und über 18 Millionen Kunden von AncestryDNA. In gewisser Weise sind die Kunden auch als Mitarbeiter zu sehen, wenn man bedenkt, dass durch sie die Datenbanken mit DNA-Daten und Stammbäumen befüllt werden. Ancestry hat DtC-Gentests an die Genealogie geknüpft und die Auffassung von Ahnenforschung für viele Menschen von Grund auf verändert (Groot 2020). Ancestry hat Ahnenforschung zu einer Tätigkeit für jedermann gemacht, und nicht mehr nur für Historiker.

Vergleich der Anbieter 23andMe und Ancestry

Aus den Historien geht klar hervor, dass 23AndMe Spitzenreiter für medizinische DtC-Gentests ist und Ancestry Spitzenreiter für genealogische Tests. Beide Unternehmen profitieren von dem wachsenden allgemeinen Interesse an DtC-Gentests. Eine große Anzahl an Kunden stellt die Grundlage für die Geschäftsmodelle beider Unternehmen dar.

Beide Geschäftsmodelle folgen dem ähnlichen Ziel, eine möglichst mächtige Sammlung an Daten zu etablieren. 23AndMe hat durch den großen Bestand an DNA-Daten den Vorteil, Gentests durch Abgleiche schneller durchzuführen und einen großen Teil zur Forschung beitragen zu können. Ancestry kann mit ihrer Genealogie-Datenbank ebenfalls zur Forschung beitragen und kann dank der Breite immer umfangreichere Ergebnisse der Ahnenforschungs-Tests liefern. Im Werdegang der beiden Unternehmen lassen sich aber große Verschiedenheiten erkennen. 23AndMe konnte es an die Spitze des Markts bringen, indem sie andere Herangehensweisen bei der DNA-Analyse als die Konkurrenz wählten und sich somit unter allen erheben konnten. Sie gelten als der Vorreiter für DtC-Gentest-Unternehmen und sind in ihrer Laufbahn als erste auf Probleme gestoßen, die für diesen Markt noch unbekannt waren. Diese Probleme konnten sie aber gewandt bewältigen und so die Agilität ihres Geschäftsmodells unter Beweis stellen. Durch ihre visionäre Denkweise und Bereitschaft mit der FDA zu kooperieren, konnten sie den DtC-Gentest-Markt revolutionieren und das Vorzeigebeispiel für weitere solche Unternehmen werden. Das Handeln des Unternehmens Ancestry könnte man fast als gegenteilig betrachten. Sie waren bei allen strategischen Entscheidungen stets zurückhaltend und geduldig, und haben versucht, aus den vergangenen Fehlern anderer Unternehmen zu lernen. Ancestry hat nur intensiv durchdachte Entscheidungen getroffen, was sie stets zu Erfolg geführt hat. Sie haben an ihren Grundwerten festgehalten und haben durch das Sammeln und Digitalisieren von Daten bedeutende Erfolge in der Ahnenforschung leisten können.

Auch bei der Verarbeitung von personenbezogenen Daten gehen beide Unternehmen unterschiedlich vor. Ancestry verwendet die DNA-Daten nur zu internen Zwecken wie zur Erweiterung ihrer Services, zur Untersuchung aggregierter genetischer Daten und zur Verbindung mit möglichen Verwandten, sog. „DNA-Matches“. Sie bieten ebenfalls die Option, Daten freiwillig zur statistischen und historischen Forschung freizugeben. 23AndMe auf der anderen Hand behält die Rechte an den Ergebnissen des DNA-Tests und verkauft diese weiter an Pharmahersteller und Institute zur Forschung. Durch diese zusätzlichen Einnahmen jedoch kann 23AndMe zu einem wesentlich niedrigeren Preis einen ähnlich umfangreichen Service wie die teurere Konkurrenz bieten.

Wie man hier deutlich erkennen kann, können beide Geschäftsmodelle zu großem Erfolg führen. Aus diesen zwei Historien geht hervor, dass der Markt für DtC-Gentests vielfältige Geschäftsmodelle hervorbringt, die stark von äußeren Faktoren beeinflusst werden.

Diskussion und Fazit

Grundsätzliche Resultate

Bei 23andMe und Ancestry handelt es sich um zwei unterschiedliche DtC-Gentest Anbieter, die auf ihre eigenen Weisen erfolgreich sind. 23andMe offenbart sich als junges, flexibles, agiles, innovatives Unternehmen mit dem Ziel Revolutionäres zu leisten. Es agiert häufig als Vorreiter und ist seinem Markt voraus, kann dadurch aber auch leichter Fehler begehen. Ancestry hingegen besinnt sich gerne auf funktionierende Marktstrategien und wagt erst Neues, wenn ein gewisser Erfolg bereits bei anderen DtC-Gentest Anbietern vermerkt wurde. In diesem Sinne existiert nicht nur heute, sondern auch in der vergangenen Entwicklung ein heterogenes Geschäftsumfeld.

Grenzen

Das gesteckte Ziel, die Historien der Geschäftsmodelle zweier DtC-Gentest Anbieter zu rekonstruieren und zu analysieren, ist eindeutig erreicht worden. Auch der Vergleich der zwei Anbieter ist auf Grund unserer Anbieterwahl gut durchführbar. Wir wählten 23andMe und Ancestry deswegen aus, weil wir uns sehr unterschiedliche Entwicklungen und somit eine kontrastreiche Gegenüberstellung erwarten. Bei der Auswahl anderer Anbieter wären genauso andere Ergebnisse möglich gewesen. Im Zweifelsfall könnte man unsere Auswahl in Frage stellen. Insbesondere hätte man an Stelle von Ancestry eventuell einen Anbieter finden können, der mehr wissenschaftliche Literatur bietet. Hinzu wären die Historie und somit ein Vergleich von noch mehr Anbietern eine gute Ergänzung gewesen, um ein noch facettenreicheres Bild zu erhalten. Dies wäre ein Anlass für weitere, zukünftige Forschung, weitere Anbieter zu untersuchen. Vielleicht wäre auch der Einbezug von noch mehr Datenbanken zu einem anderen oder nur detaillierteren Ergebnis durch mehr Quellenangaben gekommen. Außerdem konnten wir nicht alleinig mit wissenschaftlicher Literatur arbeiten, da grundsätzlich nur sehr wenig wissenschaftliche Literatur zu dem

Forschungsthema bekannt ist und diese größtenteils ebenfalls auf nicht-wissenschaftliche Literatur zurückgreift. Zumindest bei 23andMe waren wir bestrebt, die Verwendung von jedweder nicht wissenschaftlichen Literatur einzuschränken. Dazu bezogen wir nur nicht wissenschaftliche Literatur ein, die von unseren gefundenen Quellen an wissenschaftlicher Literatur zitiert wurde, um einen Anhaltspunkt für eine dennoch akzeptable Qualität der nicht-wissenschaftlichen Quellen zu haben. In dem Sinne vertrauten wir auf das Urteilsvermögen von Wissenschaftlern, dass die von ihnen verwendeten Quellen gewisse Qualitätsstandards erreichen. Diese Annahme ist anfechtbar. Es könnten dennoch unseriöse Quellen vorhanden sein, die unsere Aussagen verfälschen und uns zu falschen Entschlüssen führen konnten.

Auswirkungen und Zukünftige Forschung

Unsere Arbeit wird hoffentlich sowohl ihren Beitrag zur Forschung als auch zur Praxis leisten und zu einem umfassenderen Verständnis der zwei DtC-Gentest Anbieter in Bezug auf die historische Entwicklung ihrer Geschäftsmodelle beitragen. In Verbindung mit der Momentaufnahme der heutigen Geschäftsmodelle von Thiebes et al. (2019), an die wir anknüpften, manifestiert sich ein grundlegendes Wissen über die Geschäftsmodelle. Als Grundlage für weitere Forschung könnte unsere Arbeit, wie im letzten Kapitel angerissen, ein Anreiz zur Untersuchung der historischen Geschäftsmodelle weiterer DtC-Gentest Anbieter sein, um ein facettenreicheres Bild einerseits im Detail für jeden interessanteren Anbieter, anderseits aber auch als Big Picture ganzheitlich zu erhalten. Ein weiterer Forschungsansatz könnte auch die Untersuchung der Vernetzung des DtC-Gentest Markts mit weiteren Märkten sein.

In der Praxis könnte unsere Arbeit hilfreich für staatliche Regulatoren, Kunden, die Anbieter selbst, aber auch für Wissenschaftler im Gesundheitsbereich sein. Die erstellte Historie gibt Aufschluss über verschiedenste Änderungen des Geschäftsmodells und könnte zum besseren Verständnis bei der Gesetzgebung und der geeigneten Regulierung dienen. Gleichzeitig kann es auch auf neue Aspekte hinweisen, die zukünftig einer potenziellen Regulierung unterzogen und daher genauer betrachtet werden sollten, wie beispielsweise eine potenzielle Regulierung der privaten Datenbanken der Anbieter in Bezug auf den Datenschutz. Dies wäre auch für Kunden von Interesse. Außerdem könnten auch DtC-Gentest Anbieter profitieren, um unter anderem auch ihr eigenes Geschäftsmodell zu begutachten und Schlüsse von den Erfahrungen von 23andMe und Ancestry für sich zu nutzen.

Fazit

Unsere Arbeit zeigt die unterschiedliche Handhabung der Anbieter im DtC-Gentest Markt eindeutig ebenso wie die Heterogenität selbst bei einem kleinen Nieschenmarkt wie diesem. Der eine Anbieter fokussiert sich in seinem Geschäftsmodell vollständig auf seinen DtC-Gentest Markt, der andere nutzt ihn für ein noch größeres Geschäft und besitzt noch weitere Standbeine im Geschäftsmodell. Es wurde ein Mehrwert für Forschung und Entwicklung generiert, der hoffentlich als Basis für weitere Forschung genutzt werden kann.

Weiterführende Informationen

Auf Anfrage kann eine genaue Übersicht über die Vorgehensweise und die Ergebnisse des systematischen Literaturreviews zu 23andMe mitgeteilt werden.

References

Ancestry. 2019. “Ancestry Launches Consumer Genetics Tests for Health,” *STAT*, , October 15. (retrieved from: https://www.statnews.com/2019/10/15/ancestry-health-launch/; last accessed: July 22, 2020).

Ancestry. 2020a. “Ancestry Company Facts.” (retrieved from: https://www.ancestry.com/corporate/about-ancestry/company-facts; last accessed: July 18, 2020).

Ancestry. 2020b. “Ancestry Moves Further into Consumer Genetics | MIT Technology Review.” (retrieved from: https://www.technologyreview.com/2015/07/16/110196/ancestry-moves-further-into-consumer-genetics/; last accessed July 19, 2020).

Ancestry. 2020c. “Datenschutzerklärung - Ancestry.De.” (retrieved from: https://www.ancestry.de/cs/legal/privacystatement; last accessed: July 24, 2020).

Ancestry. 2020d. "Our Story | Ancestry Corporate." (retrieved from: https://www.ancestry.com/corporate/about-ancestry/our-story; last accessed July 18, 2020).
AncestryDNA. 2020. *AncestryDNA | Introducing AncestryDNA | Ancestry*. (retrieved from: https://www.youtube.com/watch?time_continue=71&v=SgSpSmdjr7s&feature=emb_title; last accessed July 18, 2020).
Bartol, J. 2013. "Re-Examining the Gene in Personalized Genomics," *Science & Education* (22:10), pp. 2529–2546.
Boell, S., and Cecez-Kecmanovic, D. 2014. "A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches," *Communications of the Association for Information Systems* (32), pp. 257–286.
Booth, A. 2008. "Unpacking Your Literature Search Toolbox: On Search Styles and Tactics," *Health Information & Libraries Journal* (25:4), pp. 313–317.
Borry, P., Cornel, M. C., and Howard, H. C. 2010. "Where Are You Going, Where Have You Been: A Recent History of the Direct-to-Consumer Genetic Testing Market.," *Journal of Community Genetics* (1:3), pp. 101–106.
Borry, P., Howard, H. C., Sénécal, K., and Avard, D. 2010. "Health-Related Direct-to-Consumer Genetic Testing: A Review of Companies' Policies with Regard to Genetic Testing in Minors.," *Familial Cancer* (9:1), pp. 51–59.
Brady, D. 2013. "23andMe Wants to Take Its DNA Tests Mass-Market.," *Bloomberg.Com*, 64.4.
Brocke, J. vom, Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., and Cleven, A. 2009. *Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process*, presented at the (retrieved from: http://www.alexandria.unisg.ch/Publikationen/67910; last accessed: June 10, 2020).
Castle, D., and Ries, N. M. 2007. "Ethical, Legal and Social Issues in Nutrigenomics: The Challenges of Regulating Service Delivery and Building Health Professional Capacity," *Nutrigenomics* (622:1), pp. 138–143.
Caulfield, T., and McGuire, A. L. 2012. "Direct-to-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses.," *Annual Review of Medicine* (63), pp. 23–33.
Covolo, L., Rubinelli, S., Ceretti, E., and Gelatti, U. 2015. "Internet-Based Direct-to-Consumer Genetic Testing: A Systematic Review," *Journal of Medical Internet Research* (17:12), p. e279.
Du, L., and Wang, M. 2020. "Genetic Privacy and Data Protection: A Review of Chinese Direct-to-Consumer Genetic Test Services.," *Frontiers in Genetics* (11), p. 416.
Ducournau, P., Gourraud, P.-A., Rial-Sebbag, E., Cambon-Thomsen, A., and Bulle, A. 2013. "Direct-to-Consumer Health Genetic Testing Services: What Commercial Strategies for Which Socio-Ethical Issues?" *Health Sociology Review* (22:1), pp. 75–87.
Duhaime-Ross, A. 2015. "Ancestry Is Talking to the FDA about Using DNA to Estimate People's Risk of Disease," *The Verge*, October 12. (retrieved from: https://www.theverge.com/2015/10/12/9487685/ancestry-com-dna-test-kit-disease-risk-fda; last accessed: July 20, 2020).
Fielt, E. 2013. "Conceptualising Business Models: Definitions, Frameworks and Classifications," *Journal of Business Models* (1:1), pp. 85–105.
Fink, A. 2019. *Conducting Research Literature Reviews: From the Internet to Paper*, SAGE Publications.
Fisk, P. 2014. "ARE YOU READY TO CHANGE THE WORLD IN 2014?," *Business & Economy* (9:1), pp. 50–57.
Green, R. C., and Farahany, N. A. 2014. "Regulation: The FDA Is Overcautious on Consumer Genomics.," *Nature* (505:7483), pp. 286–287.
de Groot, J. 2015. "International Federation for Public History Plenary Address: On Genealogy," *The Public Historian* (37:3), pp. 102–127.
Groot, J. de. 2020. "Ancestry.Com and the Evolving Nature of Historical Information Companies," *The Public Historian* (42:1), pp. 8–28.
Gurwitz, D., and Bregman-Eschet, Y. 2009. "Personal Genomics Services: Whose Genomes?," *European Journal of Human Genetics* (17:7), pp. 883–889.
Hamzelou, J. 2020. "The Business of DNA Analysis," *New Scientist* (245:3269), p. 15.
Harper, J. C., Kennett, D., and Reisel, D. 2016. "The End of Donor Anonymity: How Genetic Testing Is Likely to Drive Anonymous Gamete Donation out of Business.," *Human Reproduction (Oxford, England)* (31:6), pp. 1135–1140.

Hogarth, S. 2017. "Valley of the Unicorns: Consumer Genomics, Venture Capital and Digital Disruption," *New Genetics and Society* (36:3), pp. 250–272.
Hogarth, S., Javitt, G., and Melzer, D. 2008. "The Current Landscape for Direct-to-Consumer Genetic Testing: Legal, Ethical, and Policy Issues," *Annual Review of Genomics and Human Genetics* (9:1), pp. 161–182.
Hudson, K., Javitt, G., Burke, W., and Byers, P. 2007. "ASHG Statement* on Direct-to-Consumer Genetic Testing in the United States," *The American Journal of Human Genetics* (81:3), pp. 635–637.
Hwang, J., and Christensen, C. M. 2008. "Disruptive Innovation in Health Care Delivery: A Framework for Business-Model Innovation.," *Health Affairs (Project Hope)* (27:5), pp. 1329–1335.
Janssens, A. C. J. W. 2019a. "Chapter 6 - Direct-to-Consumer Genetic Testing," in *Clinical Genome Sequencing*, A. Tibben and B. B. Biesecker (eds.), Academic Press, pp. 89–101. (retrieved from: http://www.sciencedirect.com/science/article/pii/B9780128133354000064).
Janssens, A. C. J. W. 2019b. "Proprietary Algorithms for Polygenic Risk: Protecting Scientific Innovation or Hiding the Lack of It?" *Genes* (10:6), p. 448.
Kaufman, D. J., Bollinger, J. M., Dvoskin, R. L., and Scott, J. A. 2012. "Risky Business: Risk Perception and the Use of Medical Services among Customers of DTC Personal Genetic Testing.," *Journal of Genetic Counseling* (21:3), pp. 413–422.
Lancet. 2010. "New Guidelines for Genetic Tests Are Welcome but Insufficient.," *Lancet (London, England)* (376:9740), p. 488.
Lee, S. S.-J., and Crawley, L. 2009. "Research 2.0: Social Networking and Direct-To-Consumer (DTC) Genomics," *The American Journal of Bioethics* (9:6–7), pp. 35–44.
Levy, Y., and Ellis, T. J. (n.d.). "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research," *Informing Science: The International Journal of an Emerging Transdiscipline* (9), pp. 181–212.
Manchester Research. 2020. "The Genealogy Boom | Research Explorer | The University of Manchester." (retrieved from: https://www.research.manchester.ac.uk/portal/en/publications/the-genealogy-boom(2fcda487-c5fc-44d4-8aaf-8461ee15837d).html).
Nightingale, P., and Martin, P. 2004. "The Myth of the Biotech Revolution.," *Trends in Biotechnology* (22:11), pp. 564–569.
Onetti, A., Zucchella, A., Jones, M. V., and McDougall-Covin, P. P. 2012. "Internationalization, Innovation and Entrepreneurship: Business Models for New Technology-Based Firms," *Journal of Management & Governance* (16:3), pp. 337–368.
Petty, R., and Guthrie, J. 2000. "Intellectual Capitalliterature Review: Measurement, Reporting and Management," *Journal of Intellectual Capital* (1:2), pp. 155–176.
Phillips, A. M. 2016. "'Only a Click Away — DTC Genetics for Ancestry, Health, Love…and More: A View of the Business and Regulatory Landscape,'" *Applied & Translational Genomics* (8), Personal Genomics: Complications and Aspirations, pp. 16–22.
Phillips, K. A., Trosman, J. R., and Douglas, M. P. 2019. "Emergence of Hybrid Models of Genetic Testing Beyond Direct-to-Consumer or Traditional Labs.," *JAMA: Journal of the American Medical Association* (321:24), pp. 2403–2404.
Ratner, M. 2014. "Pharma Partners with Efforts to Pool Patient Genotype and Phenotype Data," *Nature Biotechnology* (32:10), pp. 967–967.
Rowley, J., and Slack, F. 2004. "Conducting a Literature Review," *Management Research News* (27:6), pp. 31–39.
Saukko, P. 2017. "Shifting Metaphors in Direct-to-Consumer Genetic Testing: From Genes as Information to Genes as Big Data," *New Genetics and Society* (36:3), pp. 296–313.
Schwartz, L. M., and Woloshin, S. 2019. "Medical Marketing in the United States, 1997-2016: The Journal of the American Medical Association," *JAMA* (321:1), p. 80.
Scudellari, M. 2018. "Get Paid for Your Genetic Data [Spectral Lines]," *IEEE Spectrum* (55), pp. 6–6.
Servick, K. 2015. "Can 23andMe Have It All?" *Science (New York, N.Y.)* (349:6255), pp. 1472–4, 1476–477.
Shafer, S. M., Smith, H. J., and Linder, J. C. 2005. "The Power of Business Models," *Business Horizons* (48:3), pp. 199–207.
Spector-Bagdady, K. 2016. "'The Google of Healthcare': Enabling the Privatization of Genetic Bio/Databanking.," *Annals of Epidemiology* (26:7), pp. 515–519.
Spector-Bagdady, K., and Pike, E. R. 2014. "Consuming Genomics: Regulating Direct-to-Consumer Genetic and Genomic Information," *NEBRASKA LAW REVIEW* (92), p. 70.

Templier, M., and Paré, G. 2015. "A Framework for Guiding and Evaluating Literature Reviews," *Communications of the Association for Information Systems* (37:1).

The Drum. 2020. "As the DNA Market Slows, Ancestry's Marketing Leans into Its Family History Business | The Drum." (retrieved from: https://www.thedrum.com/news/2020/03/13/the-dna-market-slows-ancestry-s-marketing-leans-its-family-history-business; last accessed: July 19, 2020).

Thiebes, S., Toussaint, P., Ju, J., Ahn, J.-H., Lyytinen, K., and Sunyaev, A. 2020. "Valuable Genomes: A Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing," *Journal of Medical Internet Research*.

Thrush, S. A., and McCaffrey, R. 2010. "Direct-to-Consumer Genetic Testing: What the Nurse Practitioner Should Know," *The Journal for Nurse Practitioners* (6:4), pp. 269–273.

Turrini, M. 2018. "Online Genomes: Problematizing the Disruptiveness of Direct-to-Consumer Genetic Tests," *Sociology Compass* (12:11), p. e12633.

Tutton, R., and Prainsack, B. 2011. "Enterprising or Altruistic Selves? Making up Research Subjects in Genetics Research.," *Sociology of Health & Illness* (33:7), pp. 1081–1095.

Wadman, M. 2008. "Gene-Testing Firms Face Legal Battle.," *Nature* (453:7199), pp. 1148–1149.

Wessel, M. 2016. *How Big Data Is Changing Disruptive Innovation.*, p. 1. (retrieved last: http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dbsu%26AN%3d118431853%26site%3dehost-live; last accessed: July 31, 2020)).

Whaley, G., and McGuire, S. 2018. "23andMe: Future of Personal Genomics Services Business?" *Journal of Case Studies* (36:1), pp. 78–96.

Wilde, A., Meiser, B., Mitchell, P. B., and Schofield, P. R. 2010. "Public Interest in Predictive Genetic Testing, Including Direct-to-Consumer Testing, for Susceptibility to Major Depression: Preliminary Findings," *European Journal of Human Genetics* (18:1), pp. 47–51.

Willever-Farr, H., Zach, L., and Forte, A. 2012. "Tell Me about My Family: A Study of Cooperative Research on Ancestry.Com," in *Proceedings of the 2012 IConference*, IConference '12, Toronto, Ontario, Canada: Association for Computing Machinery, February 7, pp. 303–310.

Williams-Jones, B. 2003. "Where There's a Web, There's a Way: Commercial Genetic Testing and the Internet.," *Community Genetics* (6:1), pp. 46–57.

Wojcicki, A. 2012. "One Million Strong," *23andMe Blog*, December 11. (retrieved from: https://blog.23andme.com/news/one-million-strong/; last accessed: July 24, 2020).

Wolfberg, A. J. 2006. "Genes on the Web" Direct-to-Consumer Marketing of Genetic Testing.," *New England Journal of Medicine* (355:6), pp. 543–545.

Wright, C. F., and Gregory-Jones, S. 2010. "Size of the Direct-to-Consumer Genomic Testing Market.," *Genetics in Medicine: Official Journal of the American College of Medical Genetics* (12:9), p. 594.

Machine Learning Techniques in Antibiotic Discovery

Emerging Trends in Digital Health, Summer Term 2020

Tilman Enderle
Master Student
Karlsruhe Institute of Technology
ufpxj@student.kit.edu

Nina Remmele
Master Student
Karlsruhe Institute of Technology
uorvg@student.kit.edu

Jakob Stöcker
Master Student
Karlsruhe Institute of Technology
uqecb@student.kit.edu

Abstract

Background: Artificial intelligence as well as machine learning are now widely used, including in medicine. One potential use case is an otherwise lengthy and yet not necessarily successful process, the discovery of new antibiotics. This process, marked by setbacks and costly, is becoming less attractive to pharmaceutical companies, resulting in fewer antibiotics being developed. In addition, there is the increased incidence of multidrug-resistant germs that cannot be controlled by conventional antibiotics, creating additional demand.

Objective: This paper addresses the question of which machine learning methods are applied in which steps of the discovery of a new antibiotic, in which steps the application of the methods makes sense, which advantages and disadvantages they entail, and an outlook on further potentials.

Method: For the literature review, a forward search was performed on three different databases, starting from a search string containing keywords relevant to the topic. The initial 850 hits could be narrowed down to 25 relevant publications after further screening.

Results: The literature found shows that neural networks, support vector machines as well as decision trees have been used so far in the generation and discovery of structures of new potential drugs, but also in the assessment of potential efficacy.

Conclusion: In the future, ML methods are likely to be used more frequently, as an algorithm can work much faster in this use case, and the increasing computational capacity will likely accelerate over time. The focus is likely to be on KNN, as this algorithm is very easy to train, requires little preprocessing, and does not require kernels such as SVM.

Keywords: machine learning, antibiotics, artificial neural networks, support vector machine, decision tree, random forest

Einleitung

Motivation und Hintergrund

Der Themenbereich der künstlichen Intelligenz und die damit verbundenen Techniken des maschinellen Lernens (ML) haben mittlerweile in vielen unterschiedlichen Gebieten Anwendung gefunden - digitale Sprachassistenten auf Mobiltelefonen, die Auswertung von großen Datenmengen, aber auch in der Medizin. So ist es unter anderem bereits möglich, Krebszellen durch Algorithmen zu erkennen, wofür sonst langjährig ausgebildete Experten benötigt werden. Ein weiterer Anwendungsfall ist die Unterstützung bei

der Entdeckung neuer Medikamente. Dieser Prozess ist gerade bezüglich neuer Antibiotika besonders langwierig und geprägt von Rückschlägen, da nicht jedes potentielle Präparat die gewünschte Wirkung erzielt oder zu viele Nebenwirkungen mit sich zieht. Infolgedessen sinkt die Motivation von Pharmaunternehmen, neue Antibiotika zu entwickeln und stattdessen nach profitableren Produkten zu forschen.

Diesem Problem steht zusätzlich der hohe Bedarf neuer Antibiotika aufgrund des vermehrten Vorkommens multiresistenter Keime gegenüber. Immer schneller werden Bakterien resistent gegen die herkömmlichen Antibiotika und so wird aus einem normalerweise unbedenklichen Kontakt mit einem Erreger plötzlich ein lebensgefährlicher, wenn das Immunsystem des Patienten durch eine andere Erkrankung bereits geschwächt ist und sich die Infektion ungehindert ausbreiten kann. Daher wird die Dringlichkeit neuer Wirkstoffe in Antibiotika immer höher. Diese multiresistenten Keime werfen die Medizin nahezu in ein Zeitalter vor antibiotischer Behandlungspraktiken zurück und waren laut einer Studie des Europäischen Zentrums für Prävention und Kontrolle von Krankheiten (ECDC) allein in Deutschland für 2.400 und in Europa 33.000 Todesfälle verantwortlich (Cassini et al. 2019).

Ziele der Arbeit

In der vorliegenden Arbeit befassen wir uns mit der Frage, welche Verfahren des maschinellen Lernens in welchen Schritten der Entdeckung eines neuen Antibiotikums angewendet werden. Zusätzlich gehen wir darauf ein, bei welchen Schritten die Anwendung der beschriebenen Verfahren überhaupt Sinn ergibt und welche Vor- und Nachteile sie mit sich bringen. Des Weiteren geben wir Ausblick über noch folgende Potentiale.

Grundlagen

Diese Arbeit handelt von zwei Themen, zum einen von ML, zum anderen von Antibiotika. In diesem Kapitel wird eine Einführung in beide Themen gegeben.

Maschinelles Lernen

ML behandelt das automatische Lösen von komplexen Problemen durch Algorithmen, die sehr schwer mit konventionellen Programmiermethoden zu lösen sind (Rebala et al. 2019). Beim ML gibt es meist eine große Datengrundlage, mit der ein Problem gelöst werden soll. Ein Beispiel hierzu ist die Bilderkennung. Mit klassischen Programmiermethoden ist es sehr schwer, wenn nicht unmöglich, einen Algorithmus zu schreiben, der erkennen kann, was auf Bildern abgebildet wird. Hierfür müsste man eine Vielzahl an Regeln programmieren und versuchen, alle Eventualitäten bei Bildern abzudecken. Da Bilder aber nie genau das Gleiche zeigen, ist dies mit konventionellen Methoden unmöglich. ML verwendet hier einen anderen Ansatz, die notwendigen Strukturen und Regeln werden selbstständig erlernt (Rebala et al. 2019). Am Beispiel der Bilder würde man hierfür einem ML-Algorithmus einen Trainingsdatensatz geben. In diesem Datensatz sind Bilder und die Bedeutung der Bilder gespeichert. Der Algorithmus lernt dann selbstständig Regeln, die diese Bilder mit den Bedeutungen verknüpfen. Diese Regeln können dann auf andere Bilder übertragen werden. ML-Algorithmen lösen Probleme meistens besser als Menschen, da sie alle Datenpunkte eines Datensatzes betrachten (Rebala et al. 2019).

Eine Schwachstelle dieser Methode ist jedoch, dass nicht klar ist, wie genau das Problem vom Algorithmus gelöst wurde (Rebala et al. 2019).

ML hat in der Mitte der 2000er einen Aufschwung erfahren. Immer mehr Daten sind vorhanden, mithilfe derer man die Algorithmen trainieren kann. Auch werden Computer immer leistungsfähiger, um mit den Daten gut umgehen zu können und die Algorithmen werden immer weiter verbessert (Rebala et al. 2019). So sind zum Beispiel, bedingt durch die Rechnerleistung, künstliche neuronale Netze (KNN) immer relevanter geworden, da Computer nun leistungsstark genug sind, um diese zu trainieren. KNN sind der Versuch, ein menschliches Gehirn nachzubilden, indem künstliche Neuronen automatisch so verknüpft werden, dass sie ein bestimmtes Problem lösen können. Durch diese Technik wurden Technologien wie Sprach- und Bilderkennung erst möglich (Rebala et al. 2019).

Biologische Datenbanken

ML-Methoden benötigen eine große Datenmenge. Für Antibiotika gibt es einige medizinische bzw. biologische Datenbanken, in denen antibakteriell wirkende Substanzen mit ihren biologischen Attributen aufgelistet sind. So gibt es zum Beispiel viele Datenbanken, die Peptide enthalten. Zu erwähnen ist die Antimicrobial Peptide Database (APD), welche antimikrobielle Peptide beinhaltet (Wu et al. 2019). Diese Datenbank kann von ML-Algorithmen dazu benutzt werden, die Strukturen von antimikrobiellen Peptiden zu lernen und dann mithilfe von anderen Peptid-Datenbanken neue antimikrobielle Peptide zu entdecken. Dies funktioniert, da in den Datenbanken die biologische Sequenz der Peptide vermerkt ist. Eine Sequenz beschreibt die atomaren Zusammensetzungen der Stoffe und der chemischen Bindungen, welche diese Atome verbinden. So kann man durch ML versuchen herauszufinden, welcher Teil der Peptide für die antimikrobielle Wirkung verantwortlich ist. Einige Datenbanken für Peptide finden sich in (Wu et al. 2019).

Antibiotika

„Antibiotika sind kleine Moleküle mit antimikrobieller Wirkung, die zur Behandlung von Infektionen – im engeren Sinne bakteriellen Infektionen – genutzt werden." (Fritsche 2016). Sie sind eine bedeutende Entwicklung der modernen Medizin, da es erst mit ihrer Hilfe möglich war, viele Infektionskrankheiten zu bekämpfen. Hierbei wirken sie entweder bakteriostatisch, verhindern also die Vermehrung von Bakterien, oder bakterizid, das heißt sie töten die Bakterien (Fritsche 2016). Antibiotika können aus vielen unterschiedlichen Stoffen bestehen. So gibt es zum Beispiel die Aminoglykosid-Antibiotika „eine Kombination von Aminozucker und Cyclohexanen" (Fritsche 2016), Chinolone, „synthetische Antibiotika auf Basis eines stickstoffhaltigen doppelten Sechsrings mit einer Carbonylgruppe […] und einer Carbonsäuregruppe" (Fritsche 2016). β-Lactame, Antibiotika mit einem „Lactamring von drei Kohlenstoff- und einem Stickstoffatom" (Fritsche 2016) oder auch (Poly-)Peptid-Antibiotika.

„Polypeptid-Antibiotika bestehen aus linearen, ringförmig geschlossenen oder verzweigten kurzen Ketten von Aminosäuren" (Fritsche 2016). Auch sie können entweder bakteriostatisch oder bakterizid wirken. Hierbei wirken sie an der Zellmembran der Bakterien und stören die Transportmechanismen, weswegen die Bakterien nicht mehr schädlich sind (Fritsche 2016).

Polypeptide Antibiotika spielen eine wichtige Rolle im Zusammenhang mit der Entdeckung durch ML und werden in vielen Quellen behandelt (vergleiche z.B.: (Grafskaia et al. 2018), (Giguère et al. 2015).

Methoden

Unsere Arbeit ist das Ergebnis eines Literaturreviews. Um ein grundsätzliches Verständnis zu erlangen, lasen wir zunächst die anfänglich vorgegebene Literatur. Auf dieser Basis konnten wir unseren Forschungskreislauf und auch unsere darin enthaltene Forschungsfrage erstellen. Die Forschungsfrage lautet: „Welche Techniken des maschinellen Lernens eignen sich zur Entdeckung von Antibiotika?". Als Antwort auf diese Frage erhoffen wir uns die Identifikation und Einordnung der verwendeten Techniken bei der Entdeckung von Antibiotika. Des Weiteren möchten wir die Einsatzgebiete identifizieren.

Für die eigentliche Literaturrecherche benutzten wir drei verschiedene Datenbanken:

1. bioRxiv: Die medizinische Datenbank passt gut aufgrund der medizinischen Seite unseres Themas. Hier erhalten wir viele Informationen über die medizinische Sicht, also über Antibiotika beziehungsweise Peptide. Außerdem hat uns diese Datenbank sehr viele, auf den ersten Blick treffend erscheinende Quellen geliefert.
2. IEEE Xplore: Diese Datenbank wählten wir aus, da sie die Seite der Informatik abdeckt. Wir fanden es wichtig, auch Literatur zu finden, die nicht aus rein medizinischer Sicht, sondern mit Informatik-Verständnis verfasst wurde. Der Themenbereich des maschinellen Lernens beziehungsweise von Deep Learning ist hiermit gut abgedeckt.
3. Scopus: Um einen allgemeinen Überblick über beide Themen zu erhalten, entschieden wir uns für Scopus, eine gemischte Datenbank über verschiedenste Themenbereiche (Naturwissenschaften, Technik, Medizin, Sozialwissenschaften und Kunst- und Geisteswissenschaften). Sie half uns, die allgemeine Sicht zu überprüfen.

Abbildung 1 zeigt beispielhaft unser Vorgehen für die Datenbank bioRxiv, da wir bei dieser am meisten Paper gefunden haben.

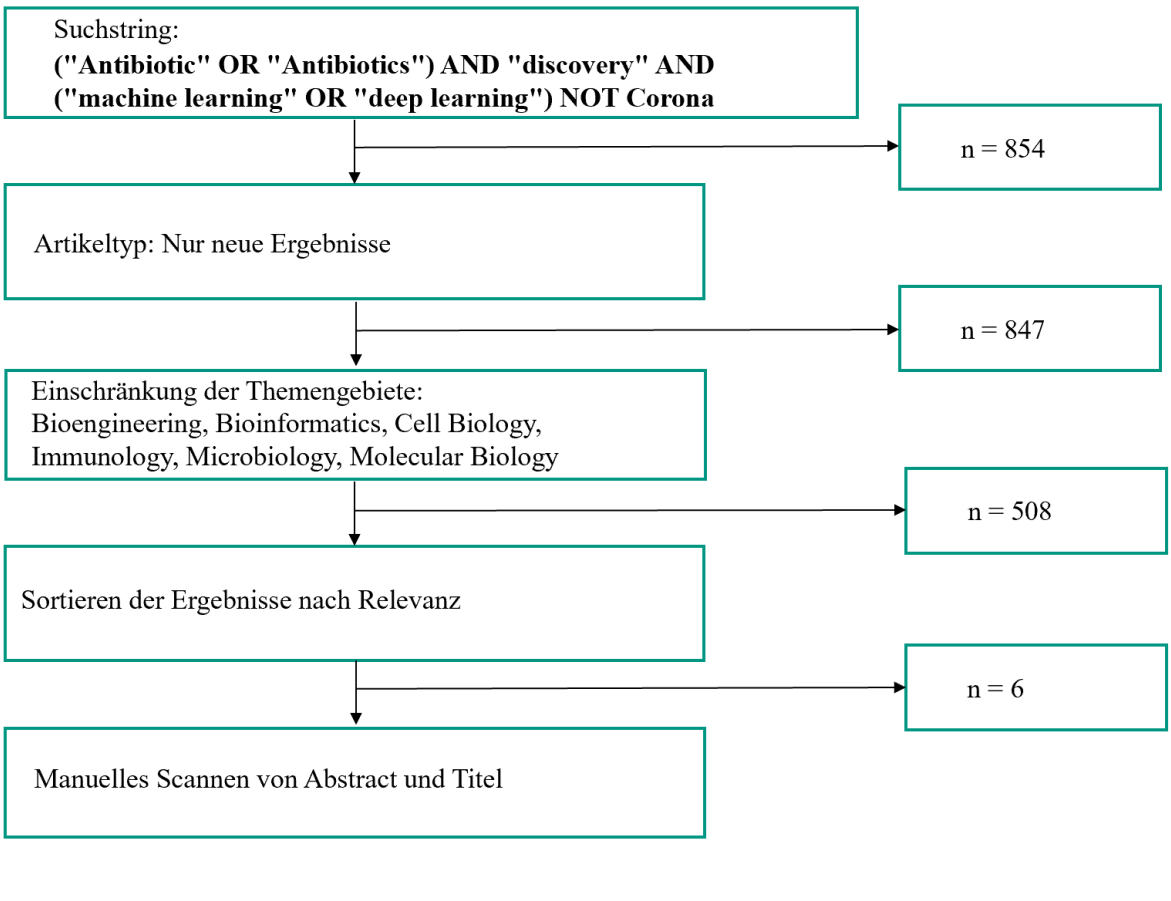

Abbildung 1. Ablauf der Suche

Für alle Datenbanken verwendeten wir den gleichen Suchstring, den wir nur an die jeweilige Syntax anpassten.

"(Antibiotic OR Antibiotics) AND discovery AND (machine learning OR deep learning) NOT Corona"

Unseren Suchstring entwickelten wir iterativ. Wir suchten erst nach einzelnen Wörtern, anschließend setzen wir beide Themengebiete zusammen. Da Peptide als Antibiotika betrachtet werden können, führten wir diesen Begriff nicht in unserem Suchstring auf. Auf der Informatik-Seite entschieden wir uns nur für die zwei Begriffe „machine learning" und „deep learning", da mit weiteren Begriffen des maschinellen Lernens keine zusätzlichen relevanten Quellen gefunden wurden. Den Zusatz „NOT Corona" verwendeten wir, da sich viele der aktuellen Paper mit diesem Thema befassten, der Corona-Virus aber nicht bakteriellen Ursprungs ist und damit für unsere Arbeit nicht relevant.

Um von den ungefähr 800 Papern, die wir mit diesem Suchstring insgesamt fanden, einige ausschließen zu können, filterten wir auch nach Themengebieten. Wir entschieden uns beispielsweise für die Bereiche Bioinformatics, Immunology, Microbiology, Computer Science, Medicine und Pharmacology.

Anschließend scannten wir Titel und Abstract der Paper manuell. Dabei achteten wir darauf, dass bereits der Titel beide Themengebiete enthielt und nicht nur auf maschinelles Lernen oder Antibiotika einging. Auch im Abstract suchten wir nach beiden Bereichen. Sobald wir über das Paper die folgenden zwei Fragen bejahen konnten, kam es in die engere Auswahl.

1. Behandelt das Paper ML, Deep Learning oder generell eine Methode aus dem Bereich der künstlichen Intelligenz in Kombination mit Antibiotika oder Peptiden?

2. Spielt das maschinelle Lernen eine unterstützende Rolle bei der Entdeckung neuer Antibiotika beziehungsweise neuer, geeigneter Wirkstoffe?

Nach dem Scan hatten wir 30 Paper, die wir ausführlich lasen. Danach konnten wir weitere fünf aussortieren, die doch nicht zu unserem Thema passten. Die übrigen 25 Paper analysierten wir, um herauszufinden, welche ML-Methoden in welchen Schritten der Antibiotika-Entdeckung verwendet wurden. Drei Methoden fanden wir besonders auffallend, weshalb wir in unserem Literaturreview auf genau diese drei näher eingehen wollen. Dabei handelt es sich um die Methoden Support Vector Machine (SVM), KNN und Random Forests (RF). Diese wurden häufig genannt und in den Ergebnissen als gut funktionierend beschrieben. Besonderes Augenmerk werden wir auf KNN legen, da diese in den Papern am häufigsten verwendet wurden. Außerdem sind wir der Meinung, dass diese Methode am stärksten in die Zukunft weist und bald immer öfter angewendet werden wird, da sich die Fähigkeit, den KNN-Algorithmus zu trainieren, weiter verbessern wird. Am zweithäufigsten fanden sich SVM, darauf schlussendlich gefolgt RF, wenn auch schon seltener und daher von uns weniger ausführlich behandelt.

In unserer Ergebnisdiskussion erläutern wir zunächst die Anwendungsgebiete. In welchen Schritten wurden die verschiedenen Methoden des MLs wirklich angewendet? Anschließend erstellten wir eine Übersicht über Vor- und Nachteile der Methoden. Um unsere Arbeit abzuschließen, fassen wir im Fazit unser Paper kurz zusammen und geben eine persönliche Einschätzung zur weiteren Entwicklung von ML-Methoden im Bereich der Antibiotika-Entwicklung.

Ergebnisse

In der Literatur stößt man immer wieder auf dieselben Methoden, die verwendet werden, um mithilfe von ML Antibiotika zu entdecken. Drei der am häufigsten vorkommenden werden im Folgenden vorgestellt und die Inhalte der Paper kurz erläutert.

Künstliche Neuronale Netze

In den letzten Jahren wurde verstärkt auf die Nutzung von KNN bzw. tiefe KNN bei der Entdeckung von neuen Antibiotika gesetzt. Dies liegt daran, dass Rechner nun leistungsstark genug sind, um in angemessener Zeit solche Netze berechnen zu können. Ein KNN ist der Versuch des Menschen, ein menschliches Gehirn mit Maschinen nachzuahmen.

Das Gehirn ist ein „hochkomplexer, nichtlinearer und paralleler Computer“ (Haykin 1999). Es kann seine Bestandteile, Neuronen, so anordnen, dass es Aufgaben wie Bilderkennung schneller und flexibler erledigen kann als digitale Computer (Haykin 1999). Dies ist möglich, da der Mensch über die Zeit durch Erfahrung „Regeln“ gelernt hat, aus denen er diese Fähigkeiten ableiten kann. In seiner grundlegenden Form ist ein KNN eine Maschine, die designt ist, um die Arbeitsweise des Gehirns nachzuahmen (Haykin 1999). Das heißt am Anfang ist es „blank“ und kennt keine Regeln. Durch den Prozess des Trainierens lernt das KNN Regeln, sodass es für einen gegebenen Input einen gewünschten Output liefert.

Im Buch „Neural networks, a comprehensive foundation“ von Haykin wird ein KNN folgendermaßen definiert (Haykin 1999):

Definition 1 (Künstliche neuronale Netze)

„Ein neuronales Netzwerk ist ein massiver parallel verteilter Prozessor, der aus einfachen Verarbeitungseinheiten besteht, die eine natürliche Tendenz haben, Erfahrungswissen zu speichern und zur Verfügung zu stellen. Es ähnelt dem Gehirn in zweierlei Hinsicht:

1. *Wissen wird vom Netzwerk von seiner Umgebung durch einen Lernprozess gewonnen.*
2. *Interneuron-Verbindungsstärken, so genannte synaptische Gewichte, werden verwendet, um das erworbene Wissen zu speichern.“*

Hierbei kann man KNN für zwei Klassen von Problemen verwenden: supervised- und unsupervised learning.

Definition 2 (Supervised learning)

„Supervised learning ist das Lernen auf Basis von Input-Output-Paaren. Ein KNN lernt für einen Input die Gewichte so anzupassen, dass der gegebene Output approximiert wird. Es leitet also eine Funktion von gelabelten Trainingsdaten ab."

Definition 3 (Unsupervised learning)

„Unsupervised learning ist ein selbstorganisiertes Lernen, das dazu führt, dass unbekannte Strukturen in Daten gefunden werden. Es findet hierbei ohne Zielwerte statt und erfordert nur Inputwerte."

Bei den meisten der Verfahren zur Entdeckung von Antibiotika handelt es sich um supervised learning. Ein KNN besteht im Wesentlichen aus Folgenden Elementen:

1. Input-Layer: Jedes KNN besitzt eine Input-Layer. Diese ist lediglich eine Fan-Out-Schicht und führt keine Berechnungen durch (Abraham 2005, S. 907). Eine Fan-Out-Schicht verteilt die Inputdaten auf die nächste Schicht.
2. Hidden-Layer: die meisten Netze besitzen eine oder mehrere Hidden-Layer. Diese Schichten bestehen aus mehreren Neuronen (Hidden-Units), wovon jedes Einzelne im Grunde eine (log) lineare Funktion ist. Eine Hidden-Layer führt ein nichtlineares Mapping vom Eingaberaum in einen (meist) höherdimensionierten Raum durch, dessen Aktivierungsfunktion aus einer Klasse von Funktionen ausgewählt ist, die Basisfunktionen genannt werden (Abraham 2005). Aktivierungsfunktionen sind Funktionen, die den Zusammenhang zwischen dem Netzinput und dem Aktivitätslevel eines Neurons darstellen. Dieses Aktivitätslevel wird dann durch eine Ausgabefunktion in einen Output transformiert, der anschließend von dem Neuron an andere Neuronen weitergesendet wird.
3. Output-Layer: Die Output-Layer nimmt die Werte der letzten Hidden-Layer entgegen und wandelt diese in den Output des KNN um.

Bei einem tiefen KNN handelt es sich um ein KNN, welches mehrere Hidden-Layers besitzt. Dadurch ist es in der Lage, komplexere Probleme zu lösen.

Verwendung

Im Jahr 2009 wurde im Paper „Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs" (Cherkasov et al. 2009) ein KNN darauf trainiert, quantitative Modelle der antibiotischen Aktivität von Peptiden aufzustellen. Die besten Peptide, die das Modell entdeckte, waren effektiv gegen eine breite Anzahl an multiresistenten „Superbugs" und damit genauso gut oder besser als vier konventionelle Antibiotika (Cherkasov et al. 2009). Hierbei wurde festgestellt, dass die Nutzung von KNNs die Entdeckung von antimikrobiellen Peptiden stark beschleunigen kann, selbst wenn man als Startpunkt für das Modell einen halb zufälligen Punkt wählt (Cherkasov et al. 2009). Ein ähnlicher Ansatz wurde auch von (Fjell et al. (2007) verfolgt. Hier wurde ebenfalls ein KNN trainiert, um die antibakterielle Aktivität von Peptiden zu modellieren (Fjell et al. 2007). Auch hierbei wurden Peptide identifiziert, die Wirkung gegen „Superbugs" zeigen (Fjell et al. 2007).

Müller et al. (2018) haben ein Recurrent Neural Network (RNN), genauer gesagt ein sogenanntes long short-term memory (LSTM) RNN, für ein kombinatorisches de novo-Peptid Design verwendet (Müller et al. 2018). RNN sind KNN, die Muster in sequentiellen Daten entdecken und neue Daten aus dem gelernten Kontext generieren (Müller et al. 2018). LSTM sind spezielle RNN, die sich Zustände merken können. Hierbei wird über Input-Gates kontrolliert, welche Informationen aufgenommen werden sollen, über Output-Gates, welche Informationen ausgegeben werden sollen und über Forget-Gates können einmal aufgenommene Informationen wieder vergessen werden. (Müller et al. 2018) haben „ein zweischichtiges unidirektionales LSTM RNN mit 256 Speichereinheiten pro Schicht trainiert. Der Ausgang der zweiten LSTM-Schicht wurde in eine dicht verbundene Feedforward-Schicht mit 22 Ausgangsneuronen eingespeist, wobei die Ausgangssignale mit einer Softmax-Funktion kombiniert wurden" (Müller et al. 2018). Mithilfe dieses Netzes wurde die Struktur von Peptiden generiert, die möglichst eine antibakterielle Wirkung haben sollten. Hierbei wurde anhand von Daten aus „A database for antimicrobial peptides (ADAM)", „APD" und „Database of Anuran defense peptide (DADP)" (Müller et al. 2018). Alle drei Datenbanken sind öffentlich zugänglich.

Nachdem das LSTM die Strukturen generiert hatte, wurde mittels eines antimikrobielle-Peptide Vorhersage Tools vom CAMP Server (Waghu et al. 2016) die antimikrobielle Aktivität hervorgesagt mit dem Ergebnis, dass 82% der generierten Sequenzen als aktiv bestimmt wurden (Müller et al. 2018).

Su et al. (2019) haben eine weitere KNN-Architektur benutzt, um antimikrobielle Peptide zu identifizieren, ein Convolutional Neural Network. CNNs besitzen Convolutional-Layers, die eine Matrix entgegennimmt, und einzelnen Teilen der Matrix Bedeutung zuordnen kann. Dadurch sind sie in der Lage, Muster in Daten zu erkennen. CNN sind recht komplex zu berechnen und können daher erst in der Praxis benutzt werden, seitdem man auf GPUs trainieren kann. Jedoch liefern sie meistens gute Ergebnisse und gelten als State-of-the-Art für Klassifikation (Lee et al. 2016). So war es Su et al. (2019) möglich, anhand von CNN ein Modell zu erstellen, das die existierenden Modelle übertrifft (Su et al. 2019).

Auch von Monteiro et al. (2020) wurde ein CNN benutzt. In dem Paper werden CNN genutzt, um aus „1D-Rohdaten, proteinhaltigen Aminosäuresequenzen und SMILES-Strings zur Darstellung der chemischen Struktur des Medikaments“ (Monteiro et al. 2020) einen Feature-Vektor zu extrahieren, der dann als Input für ein Fully Connected Neural Network (FCNN) dient (Monteiro et al. 2020). Dieses FCNN klassifizierte den Input als antibakteriell oder nicht antibakteriell (Monteiro et al. 2020).

Im Jahr 2020 nutzten Stokes et al (2020) ein tiefes KNN, um Moleküle mit antibakterieller Aktivität vorherzusagen. Dabei fokussierten sie sich nicht nur auf Peptide. Hierbei entdeckte das Modell acht antibakterielle Stoffe, die sich strukturell von bisherigen Antibiotika unterscheiden (Stokes et al. 2020). Unter anderem identifizierten sie Halicin als potenzielles Antibiotikum. Das Netz wurde auf Moleküle, die das Wachstum von E. coli Bakterien hemmen, trainiert und um eine Reihe von molekularen Merkmalen erweitert. Anschließend wurde Hyperparamter-Optimierung und Ensembling durchgeführt. Das so trainierte Netz wurde dann auf mehrere chemische Bibliotheken angewendet und die vielversprechendsten Kandidaten ausgewählt (Stokes et al. 2020). Das Netz wurde als binärer Klassifikator aufgebaut. Es wurde also nur unterschieden, ob ein Stoff das Wachstum von E. coli hemmt oder nicht (Stokes et al. 2020).

Support Vector Machine

Nach Lee et al (2016) ist eine SVM ein Algorithmus für maschinelles Lernen. SVNs sind binäre lineare Klassifikatoren. Das bedeutet, sie verfügen über eine Hyperebene, um Datenpunkte durch Maximierung des Abstandes vom nächstgelegenen Punkt in jeder Klasse zur trennenden Hyperebene in zwei Klassen zu trennen. Sie können nur Datenpunkte bearbeiten, die als linear trennbar bezeichnet werden, d.h. für die es eine Hyperebenen-Grenze gibt. Die Grenze wird anhand von Beispielen im Trainingssatz positioniert, die als Stützvektoren bezeichnet werden. Die optimale Hyperebene wird Maximum-Margin-Hyperebene genannt. Sie wird durch Training der SVM mit einem Trainingsdatensatz festgelegt, in dem die Merkmale und Klassifikationen der Daten bekannt sind (Pereira et al. 2015). Sie maximiert also die Trennungsspanne zwischen verschiedenen Klassen (Monteiro et al. 2020).

Damit SVMs auch nichtlineare Probleme lösen können - die meisten Probleme in der Realität sind nichtlinear - werden sogenannte Kernels benutzt. Kernels sind Funktionen mit speziellen Eigenschaften, die Datenpunkte in einen höherdimensionalen Raum abbilden (Rebala et al. 2019). Daher wird der sogenannte Kernel-Trick angewendet, sobald ein nichtlineares Problem auftritt. Durch diesen Trick wird eine lineare Trennung eines nichtlinearen Klassifikationsproblems ermöglicht (Spänig and Heider 2019).

Verwendung

In ihrer Studie fokussieren sich Pereira et al. (2015) auf die Anwendung von ML-Techniken, um bleiähnliche Moleküle auf dem Weg zu neuen Antitumor-Medikamenten und Antibiotika zu nutzen. Die Techniken, wie unter anderem SVM, wurden angewendet, um zwei verschiedene Klassen von Verbindungen vorherzusagen - aktive und nicht-aktive. Dabei fokussieren sich die Wissenschaftler auf drei verschiedene Klassifizierungsaufgaben: die komplette biologische Aktivität, die Antitumor Aktivität und die antibiotische Aktivität. Die Ergebnisse legen nahe, dass der implementierte computergestützte Ansatz mit quantenchemischen Deskriptoren verwendet werden kann, um die Aktivität neuer oder bestehender Naturstoffe ohne Aufzeichnung derer Bioaktivität vorherzusagen (Pereira et al. 2015).

Nicht nur zur Entdeckung von herkömmlichen Antibiotika wurden Techniken des ML benutzt. Im Paper “Sequence-based analysis and prediction of lantibitoics: a machine learning approach” behandeln

Poorinmohammad et al. (2018) Lantibiotika. Lantibiotika - eine Gruppe von ribosomal synthetisierten Peptiden - stellen eine wichtige Gruppe neuartiger und vielversprechender antimikrobieller Mittel dar, die eine hohe Wirksamkeit im Kampf gegen die Resistenz gegen Antibiotika aufzeigen könnten. Um genaueres Wissen über Lantibiotika zu erlangen und passende Sequenzen vorhersagen zu können, testen Poorinmohammad et al. (2018) vier verschiedene Techniken des maschinellen Lernens, darunter auch SVM. Die Klassifikation mit SVM wurde hier mit MATLAB Code durchgeführt. Die Algorithmen wurden trainiert, um ein Vorhersagemodell für Lantibiotika zu erstellen, welches auf Aminosäuresequenzen basiert, d. h. ihr Ziel ist es, mittels einer Methode des maschinellen Lernens ein prädiktives Modell zu entwickeln, um die Aktivität von Lantibiotika allein anhand ihrer Aminosäuresequenz genau und präzise vorherzusagen. In ihrem Versuch fanden Poorinmohammad et al. (2018) heraus, dass SVM vor allem mit kleineren Datensätzen gut angewendet werden kann. SVM-Algorithmen eignen sich gut für die Klassifizierung biologischer Daten und vor allem für die Klassifizierung der Protein- und Peptid-Informationen.

Die Studie von Porto et al. (2017) konzentriert sich auf die Prognosefähigkeit von ML-Methoden, wie auch SVM, um antimikrobielle Sequenzaktivitäten zu bestimmen. Die Autoren sehen antimikrobielle Peptide als eine vielversprechende Alternative zu herkömmlichen Antibiotika an und fokussieren sich daher auf diese. Das von ihnen genutzte Trainingsset beinhaltet Sequenzen von Peptiden und Nicht-Peptiden. Porto et al. (2017) nutzt verschiedene antimikrobielle Vorhersage-Systeme, die je unterschiedliche ML-Methoden unterstützen. Alle von ihnen entwickelten Systeme, die SVM unterstützen, haben eine Genauigkeit von über 90%. Im Moment funktionieren die Prädiktoren gut, um die Sequenzen vorherzusagen, auf die sie trainiert wurden. In einem Szenario des echten Lebens werden die Systeme allerdings genutzt, um Sequenzen vorherzusagen, die nicht im Trainingsset vorhanden sind. Die Studie kommt zu der Schlussfolgerung, dass die Entwicklung eines Prädiktors mit hoher Spezifität, der die bestehenden Systeme gemeinsam nutzt, die beste Alternative darstellt.

Entscheidungsbäume

Decision Trees

Decision Trees, zu Deutsch auch Entscheidungsbäume, sind Werkzeuge aus dem Bereich des Data Minings und maschinellen Lernens, die dazu verwendet werden, Klassifizierungsaufgaben durchzuführen (Olson 2020). Grundsätzlich ist ein Entscheidungsbaum ein spezieller Graphentyp, er besteht also aus einer Menge von Knoten und Kanten. Die Besonderheit liegt hierbei darin, dass der Graph eine geordnete Struktur und keine geschlossenen Kanten hat, es sind also keine Schleifen vorhanden (Pavlov 2019).

Ein Entscheidungsbaum unterteilt, ausgehend von einem Wurzelknoten, über die Kanten den Wertebereich eines Attributes in zwei disjunkte Teilmengen, was als eine Entscheidung anhand dieses Attributes zu verstehen ist. Die Entscheidung kann dabei binär und nicht-binär getroffen werden, aus dem Knoten führen also genau zwei oder mehr als zwei Kanten heraus. In der einfachsten Form folgen auf den Wurzelknoten direkt zwei Blätter, sogenannte Klassifikationen. Normalerweise beinhaltet ein Entscheidungsbaum aber beliebig viele weitere innere Knoten, welche ebenfalls wieder Entscheidungen darstellen.

Der Begriff des Lernens oder Trainierens kann bei einem Entscheidungsbaum als Prozess des Aufbaus der Baumstruktur, also der Wahl der abzufragenden Attribute, verstanden werden. Hierfür existieren bereits zahlreiche Algorithmen wie beispielsweise der Gini Index, CART oder C4.5, auf diese kann hier aufgrund des Umfangs aber nicht detaillierter eingegangen werden. Alle diese Algorithmen haben jedoch gemeinsam, dass sie den Entscheidungsbaum auf Basis von bereits vorhandenen Datensätzen trainieren (Rutkowski et al. 2020).

Random Forests

Random Forests sind eine Erweiterung des oben beschriebenen Klassifikators. Sie bestehen aus einer Sammlung von Entscheidungsbäumen, welche jeweils unabhängig eine Klassifikation eines Objektes vornehmen. Anschließend werden alle diese Klassifikationen zusammengefasst und durch eine Wahl die populäre Klasse, die letztendliche Outputklasse, bestimmt (Olson 2020).

Verwendung

In der oben bereits erwähnten Studie von Pereira et al. (2015) wurden neben der SVM auch Entscheidungsbäume und Random Forests verwendet, um aktive und nicht-aktive Klassen von Verbindungen in bleiähnlichen Molekülen vorherzusagen. Zwar lieferte der Ansatz der Random Forests keine Verbesserung der Vorhersagewahrscheinlichkeit, allerdings konnte die Performance im Vergleich zur SVM erhöht werden.

Auch in der Arbeit von Lira et al. (2013) fanden Entscheidungsbäume Anwendung bei der Entdeckung von bisher unbekannten antimikrobiellen Peptiden. Lira et al. (2013) trainierten dabei einen Entscheidungsbaum mit dem J48-Algorithmus auf Basis von 60 bereits bekannten Verbindungen. Der Entscheidungsbaum kategorisierte dabei auf fünf Ebenen mit neun Knoten anhand der physiochemischen Eigenschaften einer Verbindung die antimikrobielle Aktivität von neuen Verbindungen in die Kategorien „keine“, „niedrig“, „mittel“ und „hoch“. Dabei konnten eine Gesamtgenauigkeit sowie eine Präzision von nahezu 70% erreicht werden, gerade bei den hoch aktiven und daher besonders relevanten Verbindungen konnte sogar eine Präzision von fast 90% erreicht werden. Lira et al. (2013) zeigten somit, dass die Verwendung von Entscheidungsbäumen für die Bewertung der antimikrobischen Aktivität synthetischer Petptide gerade bei der Findung neuer Modelle für die Anwendung in der Entwicklung von neuen Medikamenten hilfreich sein kann, wenn bereits bekannte Peptide als Designgrundlage verwendet werden. Besonders der Zeit- und Kostenaufwand in der Entwicklung kann dabei reduziert werden.

Ähnliche Erfolge konnten (Debeljak et al. (2007) in ihrer Arbeit mit Random Forests erzielen. Diese fanden neben SVM-Anwendung bei der Untersuchung von neuen Klasse antimikrobieller Wirkstoffe. Dafür wurden quantitativen Struktur-Wirkungs-Beziehungen (QSAR) erstellt, um die antimikrobiellen Aktivitäten zu bestimmen. Diese QSARs wurden wiederum durch multivariate Modelle wie Random Forest erstellt. Letztendlich konnte gezeigt werden, dass die Random Forest Methode zwar nur ähnliche Ergebnisse bezüglich der durchschnittlichen Genauigkeit erzielte, mit manuell definierten Parametern aber deutlich stabilere Ergebnisse erreichte als die verglichenen Methoden.

Ergebnisdiskussion

Im Folgenden werden die im vorherigen Kapitel näher erläuterten Methoden des maschinellen Lernens gegenübergestellt. Alle vorgestellten Methoden wurden verwendet, um vielversprechende Bestandteile für Antibiotika zu entdecken. Es handelt sich hierbei um neue, aber auch schon zuvor für andere Verwendungszwecke bekannte Bestandteile. Alle drei Methoden behandeln leicht unterschiedliche Anwendungsgebiete. So werden KNNs dafür benutzt, bereits bestehende Sequenzen in aktive und nicht-aktive Sequenzen zu klassifizieren. Ein weiterer Anwendungsbereich von KNNs ist die Entdeckung neuer Strukturen von Stoffen, die aktiv sein könnten, bisher aber nicht als solche bekannt sind.

SVM hingegen dient der Vorhersage passender Sequenzen, die als Bestandteile für Antibiotika geeignet sind. Es handelt sich hierbei sowohl um bereits bekannte als auch unbekannte Sequenzen. Des Weiteren können hiermit aber auch aktive und nicht-aktive Verbindungen vorhergesagt werden.

Die dritte Methode, Entscheidungsbäume und RF finden dagegen lediglich Anwendung bei der Bestimmung einer voraussichtlichen Wirksamkeit von möglichen Kandidaten. Neue Strukturen können nicht generiert werden.

Bei Pereira et al. (2015) erreichte nicht SVM, sondern RF das beste Ergebnis in Bezug auf die Performance. Das AntiMarin-Testset zeigte hier, dass die 161 True Positives eine durchschnittliche Wahrscheinlichkeit von ca. 83% haben, antibiotische Aktivität aufzuweisen. Die Vorhersagekraft beider Techniken des maschinellen Lernens war hier in etwa gleich. Bei Poorinmohammad et al. (2018) erreichte SVM bei drei von sieben Modi bei der Klassifizierung eine Genauigkeit von über 90%, bei anderen drei über 80 Prozent. Auch die anderen Performance Parameter der Methode waren höher als 80%. Da RF hier zu ungenau war, wurde es in diesem Paper nicht weiter betrachtet. SVM erreicht im Paper von Monteiro et al. (2020) eine sehr hohe Genauigkeit, also eine hohe Rate der korrekt klassifizierten Vorhersagen, von etwa 90%. Allerdings wies der Algorithmus damit die niedrigste Genauigkeit auf. Andere Modelle, wie z. B. RF und eine Kombination aus CNN und FCNN zeichneten sich durch eine Genauigkeit über 92% aus.

Die vorgestellten Methoden haben alle unterschiedliche Vor- und Nachteile. Für KNN zählt Haykin in seinem Buch „Neural networks: A comprehensive foundation“ mehrere Vorteile auf (Haykin 1999):

- Nichtlinearität: Ein KNN ist nichtlinear. Dies führt dazu, dass ein KNN auch zum Lösen von sehr komplizierten Problemen geeignet ist.
- Input-Output Mapping: Dadurch, dass das Netz einen Input auf einen gegebenen Output abbildet, sind zum Trainieren eines KNN keine zusätzlichen Daten notwendig, weswegen es einfach zu trainieren ist.
- Kontextuelle Informationen: Da jedes Neuron im Netzwerk potenziell von der globalen Aktivität aller anderen Neuronen im Netzwerk betroffen ist, verarbeitet ein KNN kontextuelle Informationen auf natürliche Weise.

Des Weiteren ist die Vorhersage schnell, wenn das KNN erst einmal die Strukturen der Daten gelernt hat. Natürlich gibt es auch Nachteile bei dieser Methode. Zum einem handelt es sich um einen Blackbox Ansatz. Das heißt, es können zwar die Struktur der KNN angepasst und einige Parameter geändert werden, aber es kann nicht genau nachvollzogen werden, auf welchen Grundlagen ein KNN zu einem Ergebnis kommt. Auf der anderen Seite dauert das Trainieren eines KNN lang, vor allem, wenn man eine komplizierte Struktur wie ein CNN oder LSTM wählt. Dieses Problem wird aber durch die zunehmend steigende Rechenkraft und die Möglichkeit auf GPUs trainieren zu können immer kleiner. Zudem ist das Ergebnis eines KNN von den Trainingsdaten abhängig. Sind diese schlecht gewählt, so leidet die Ergebnisqualität dementsprechend darunter.

SVM haben als Vorteil, dass sie schon auf geringen Datenmengen ein gutes Ergebnis liefern. Andere Methoden, wie z. B. KNN benötigen mehr Daten zum Trainieren. Des Weiteren generalisieren sie gut, was es einfacher macht, neue Daten zu klassifizieren. Da die Klassifizierung durch SVM mithilfe der Stützvektoren funktioniert, ist sie besonders schnell. Nachteile von SVM sind, dass sie sehr lange zum Trainieren von großen Datensätzen brauchen und dass sie recht anfällig für Rauschen, also falsche Daten, sind.

Entscheidungsbäume bieten den Vorteil, dass sie sehr gut zu interpretieren sind. Man kann gut nachvollziehen, warum eine Instanz einer bestimmten Klasse zugeordnet wurde. Außerdem sind sie sehr effizient in der Auswertung. Jedoch bringen sie die Gefahr des Overfittings mit sich. Außerdem sind optimale Entscheidungsbäume exponentiell zur Größe schwieriger zu bestimmen, da sich Knoten immer zu zwei oder mehreren neuen Knoten verzweigen.

Tabelle 1 stellt eine Übersicht aller erklärten Vor- und Nachteile dar.

Methode	Vorteile	Nachteile
KNN	• Nichtlinearität • Input-Output Mapping • Kontextuelle Informationen • Schnelle Vorhersage, wenn erstmals trainiert	• Blackbox Ansatz • Benötigt viel Rechenkapazität zum Trainieren • Stark trainingsdaten-abhängig
SVM	• Große Generalisierungs-fähigkeit zur Lösung realer Probleme • Modell kann mit relativ kleinen Trainingssets generiert werden • Sehr effiziente Methode zur Klassifizierung von Objekten	• Trainiert lange auf großen Datensätzen • Sehr anfällig für Rauschen in den Daten
Entscheidungs-bäume	• Einfache Interpretation des Baumes • Effiziente Auswertung	• Gefahr Overfitting • Optimaler Entscheidungsbaum exponentiell schwierig zu bestimmen

Tabelle 1. Vor- und Nachteile der Vorgestellten Methoden

Fazit

In unserer Arbeit behandelten wir die Frage „Welche Techniken des maschinellen Lernens eignen sich zur Entdeckung von Antibiotika?“. Nach der genauen Erläuterung der Grundlagen - Antibiotika und maschinelles Lernen - untersuchten wir zur Beantwortung unserer Forschungsfrage drei ML-Methoden. Für die Methoden künstliches neuronales Netz, Support Vector Machine und Entscheidungsbäume bzw. Random Forests fanden wir heraus, für welche Anwendungszwecke diese verwendet werden. Unser Ergebnis lautet, dass sich jeder dieser drei Algorithmen eignet, in Zukunft bei der Entedckung neuer Antibiotika zu unterstützen. Interessant war, dass die ML-Methoden nicht nur verwendet wurden, um neue, aktive Sequenzen zu entdecken, sondern auch schon bekannte Sequenzen untersucht wurden. Die bekannten Sequenzen wurden zuvor in einem anderen medizinischen Kontext eingesetzt, fielen aber trotzdem durch ihre antibakterielle Aktivität auf.

Unserer Einschätzung nach werden zukünftig wohl häufiger ML-Methoden zur Entdeckung neuer Antibiotika bzw. neuer passender Sequenzen eingesetzt. Da ein Algorithmus deutlich mehr Daten in einer kürzeren Zeit erfassen kann, ist die Wahrscheinlichkeit höher, eine aktive Sequenz auch aus einem anderen Kontext viel schneller zu entdecken, als es Menschen durch gewöhnliche Forschung schaffen würden. Dies bringt einen großen Vorteil gegen die schnelle Resistenz vieler Bakterien gegen herkömmliche Antibiotika.

Ein weiterer Grund warum in Zukunft vermutlich häufiger Techniken des maschinellen Lernens angewendet werden, ist die steigende Rechenkapazität von Computern. Die wachsende Rechenleistung ermöglicht die Benutzung komplizierter Strukturen, die sehr flexibel und vielseitig sind. Es besteht die Möglichkeit, diese auch auf Grafikeinheiten zu trainieren, was den kompletten Prozess erheblich beschleunigt.

Wir denken, dass der weitere Fokus vor allem auf KNN liegen wird, da dieser Algorithmus sehr gut zu trainieren ist, nicht viel Preprocessing durchlaufen muss und zum reibungslosen Funktionieren auch keine Kernels wie SVM benötigt.

Referenzen

Cassini, A., Högberg, L. D., Plachouras, D., … , and Hopkins, S. 2019. “Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis,” *The Lancet Infectious Diseases* (19:1), pp. 56-66.

Cherkasov, A., Hilpert, K., Jenssen, H., Fjell, C. D., Waldbrook, M., Mullaly, S. C., Volkmer, R., and Hancock, R. E. W. 2009. “Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs,” *ACS Chemical Biology* (4:1), pp. 65-74.

Debeljak, Z., Skrbo, A., Jasprica, I., Mornar, A., Plecko, V., Banjanac, M., and Medić-Sarić, M. 2007. “QSAR study of antimicrobial activity of some 3-nitrocoumarins and related compounds,” J*ournal of chemical information and modeling* (47:3), pp. 918-926.

Fjell, C. D., Hancock, R. E. W., and Cherkasov, A. 2007. “AMPer: A database and an automated discovery tool for antimicrobial peptides,” *Bioinformatics* (23:9), pp. 1148-1155.

Fritsche, O. 2016. Mikrobiologie, Berlin, Heidelberg: Springer Spektrum.

Giguère, S., Laviolette, F., Marchand, M., Tremblay, D., Moineau, S., Liang, X., Biron, É., and Corbeil, J. 2015. “Machine Learning Assisted Design of Highly Active Peptides for Drug Discovery,” *PLoS Computational Biology* (11:4).

Grafskaia, E. N., Polina, N. F., Babenko, V. V., Kharlampieva, D. D., Bobrovsky, P. A., Manuvera, V. A., Farafonova, T. E., Anikanov, N. A., and Lazarev, V. N. 2018. “Discovery of novel antimicrobial peptides: A transcriptomic study of the sea anemone Cnidopus japonicus,” *Journal of Bioinformatics and Computational Biology* (16:2).

Haykin, S. S. 1999. Neural networks: A comprehensive foundation, Upper Saddle River, NJ: Prentice Hall.

Lee, E. Y., Fulan, B. M., Wong, G. C. L., and Ferguson, A. L. 2016. “Mapping membrane activity in undiscovered peptide sequence space using machine learning,” *Proceedings of the National Academy of Sciences of the United States of America* (113:48), pp. 13588-13593.

Lira, F., Perez, P. S., Baranauskas, J. A., and Nozawa, S. R. 2013. "Prediction of antimicrobial activity of synthetic peptides by a decision tree model," *Applied and environmental microbiology* (79:10), pp. 3156-3159.

Monteiro, N. R. C., Ribeiro, B., and Arrais, J. 2020. "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," *IEEE/ACM transactions on computational biology and bioinformatics*.

Müller, A. T., Hiss, J. A., and Schneider, G. 2018. "Recurrent Neural Network Model for Constructive Peptide Design," *Journal of chemical information and modeling* (58:2), pp. 472-479.

Olson 2020. Predictive Data Mining Models, Springer Singapore.

Pavlov, Y. L. 2019. Random Forests, Berlin/Boston: De Gruyter, Inc.

Pereira, F., Latino, D. A. R. S., and Gaudêncio, S. P. 2015. "QSAR-assisted virtual screening of lead-like molecules from marine and microbial natural sources for antitumor and antibiotic drug discovery," *Molecules* (20:3), pp. 4848-4873.

Poorinmohammad, N., Hamedi, J., and Moghaddam, M. H. A. M. 2018. "Sequence-based analysis and prediction of lantibiotics: A machine learning approach," *Computational biology and chemistry* (77), pp. 199-206.

Porto, W. F., Pires, Á. S., and Franco, O. L. 2017. "Antimicrobial activity predictors benchmarking analysis using shuffled and designed synthetic peptides," *Journal of Theoretical Biology* (426), pp. 96-103.

Rebala, G., Ravi, A., and Churiwala, S. 2019. An Introduction to Machine Learning.

Rutkowski, L., Jaworski, M., and Duda, P. 2020. Stream Data Mining: Algorithms and Their Probabilistic Properties.

Spänig, S., and Heider, D. 2019. "Encodings and models for antimicrobial peptide classification for multi-resistant pathogens," *BioData mining* (12), p. 7.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. 2020. "A Deep Learning Approach to Antibiotic Discovery," *Cell* (180:4), 688-702.

Su, X., Xu, J., Yin, Y., Quan, X., and Zhang, H. 2019. "Antimicrobial peptide identification using multi-scale convolutional network," *BMC bioinformatics* (20:1), p. 730.

Waghu, F. H., Barai, R. S., Gurung, P., and Idicula-Thomas, S. 2016. "CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides," *Nucleic acids research* (44:D1).

Wu, Q., Ke, H., Li, D., Wang, Q., Fang, J., and Zhou, J. 2019. "Recent Progress in Machine Learning-based Prediction of Peptide Activity for Drug Discovery," *Current topics in medicinal chemistry* (19:1), pp. 4-16.

Review: Machine Learning Methods in Antibiotic Discovery

Emerging Trends in Digital Health, Summer Term 2020

Eileen Seitter
Bachelor Student
Karlsruhe Institute of Technology
ueiam@student.kit.edu

Tessa Buttenberg
Bachelor Student
Karlsruhe Institute of Technology
uorbd@student.kit.edu

Eda Akgöz
Bachelor Student
Karlsruhe Institute of Technology
uopeo@student.kit.edu

Emre Karyagdi
Bachelor Student
Karlsruhe Institute of Technology
ureom@student.kit.edu

Abstract

Background: The World Health Organization considers antibiotic resistance as one of the greatest economic and public health challenges of our time, which is why there is an urgent need for the discovery of new antibiotics. To improve the research process regarding the time and economic resources, machine learning techniques are used. These can be applied at various points in the development process and can be realized with different concepts. An overview of the most current and promising approaches is therefore of significant importance.

Objective: We aimed to provide an overview of the use of machine learning techniques in antibiotic discovery. The objective was to identify the most important methods that are useful and promising in antibiotic research. We sought to classify existing approaches by area of application.

Methods: The research paper was based on a systematic literature search including the databases ACM Digital Library, AIS EBSCOhost, IEEE Xplore Digital Library and PubMed. After defining central core terms and a resulting search string, 489 results were obtained. These were filtered and grouped by temporal currency (from december 2019 to June 2020) and relevance. Marginal topics and under-researched methods were omitted, while promising approaches were researched in more detail through an additional backwards search. Finally, the work is based on 30 currently relevant research publications.

Results: The review identified two main areas in which machine learning techniques can be usefully applied in antibiotic discovery: virtual screening and end-to-end approaches. In the area of virtual screening, the quantitative structure-activity relationship (QSAR) methods, successful extensions of classical methods and virtual screening of molecular fragments emerge as promising. QSAR methods use classification methods (e.g., decision trees). For the extensions of classical screening methods, random forest or direct-message passing deep neural network (D-MPNN) provide useful support. In the fragment-based drug discovery approach the hunting FOX algorithm is applied. In terms of end-to-end concepts, research is currently being conducted on the two holistic concepts for prototype-based drug discovery and drug interaction, which are built on conditional diversity networks or convolutional neuronal networks.

Conclusion: The actual success of the novel models presented can only be assessed in a few years. Furthermore, the discovery of antimicrobial peptides, which offer an alternative to antibiotics in the fight against bacterial infections, was not included in this work.

Keywords: antibiotic discovery, machine learning, deep learning, neuronal network, virtual screening, quantitative-structure-activity-relationship, fragment-based drug discovery, phenotypic drug discovery, prototype-based drug discovery, drug interaction

Einleitung

Die Entdeckung des Penicillins von Alexander Fleming im Jahr 1928 war nicht nur einer der bedeutendsten medizinischen und gesellschaftlichen Fortschritte im 20. Jahrhundert, sondern gilt auch als Auslöser vieler weiterer antimikrobiell wirkender Medikamente (Vogelmeier 2018). Antibiotika sind heutzutage in der modernen Medizin kaum noch wegzudenken. Durch sie werden Behandlungen von bakteriellen Infektionen möglich, womit große Operationen, wie beispielsweise Organtransplantation, Behandlung von Frühgeborenen, sowie Chemotherapien für Tumore ein geringeres Risiko aufweisen, da sie postoperative Komplikationen durch Entzündungen minimieren (O'Neill 2016). Durch neue medizinische Eingriffe wurden seither Millionen von Menschenleben gerettet (Laxminarayan et al. 2013).

In Anbetracht der immer häufigeren Verwendung von Antibiotika für medizinische und nicht medizinische Einsätze, wie z.B. in der Agrarindustrie oder Krankheitsprävention, droht die „Wunderwaffe" zu verschwinden und stellt die Menschheit vor eines der schwerwiegendsten Probleme weltweit. Aufgrund der übermäßigen und unkontrollierten Verwendung von Antibiotika für den therapeutischen und paratherapeutischen Einsatz kommt es zu einem Selektionsdruck und bakterielle Infektionserreger entwickeln Resistenzeigenschaften (Witte & Klare 1999). Dadurch entstehen mehrfachresistente Bakterienstämme. Unter Resistenz versteht man im Allgemeinen, dass die Konzentration eines Chemotherapeutikums am Infektionsort nicht mehr ausreicht, um eine bakterielle Infektion zu bekämpfen, sprich diese zu töten oder deren Wachstum zu hemmen. Zu dieser Definition gehören verschiedene Aspekte; die über einem Schwellwert liegende minimale Hemmkonzentration (MIC) des Chemotherapeutikums für den Erreger, die Pharmakokinetik der Substanz und das klinische Resultat. Dies führt dazu, dass immer mehr durch Bakterien ausgelöste Krankheiten nicht angemessen behandelt werden können, da diese Resistenzen gegen die zurzeit existierenden Antibiotika aufweisen. Durch diese schwer behandelbaren Infektionen sterben jedes Jahr weltweit rund 700.000 Menschen und ohne Intervention wird die jährliche Zahl der Toten im Zusammenhang mit der antimikrobiellen Resistenz (AMR) in den nächsten 35 Jahren schätzungsweise auf 10 Millionen ansteigen (O'Neill 2014). Im Zusammenhang mit den ökonomischen Kosten würde es eine Verringerung des Bruttoinlandsprodukts um 2% bis 3,5% bedeuten, was schätzungsweise einem Verlust von 100 Billionen USD entspricht. Aus diesem Grund betrachtet die Weltgesundheitsorganisation (WHO) Antibiotikaresistenzen als eine der größten wirtschaftlichen und gesundheitspolitischen Herausforderungen unserer Zeit. Die Zunahme lebensbedrohlicher Infektionen durch AMR und die in Zukunft damit verbundenen Kosten zeigen die Dringlichkeit für die Entdeckung neuer Antibiotika. Das Problem ist jedoch, dass viele Pharmaunternehmen die Entwicklung neuartiger Antibiotika aufgrund einer Reihe von rationalen Gründen und wegen des hohen Risikos eines Versagens meiden (Harel & Radinsky 2018). Die Herstellung eines neuen Medikaments ist ein teurer und langwieriger Prozess, der über 500 Millionen Dollar kostet und länger als 10 - 15 Jahre dauern kann. Die erste Phase ist die Arzneimittelentdeckung, in der potenzielle Arzneimittel identifiziert werden, bevor ein Arzneimittelkandidat für klinische Studien ausgewählt wird. Dies erfolgt meistens in vitro. Obwohl in der Vergangenheit einige Medikamente zufällig entdeckt wurden (z.B. Minoxidil und Penicillin), sind heute systematischere Ansätze gefragt, die im Hinblick auf die zeitliche und wirtschaftliche Komponente effektiver sind.

Der technologische Wandel hat in der Vergangenheit bahnbrechende Möglichkeiten geschaffen. Wie kann nun neue Technologie bei der Problemstellung von AMR und der Notwendigkeit neuer Antibiotika eingesetzt werden? Eine Möglichkeit ist die Verwendung von Techniken des Maschinellen Lernens (ML), welche in den letzten Jahren verstärkt verfolgt wurde und traditionelle in vitro Methoden mit in silico Verfahren ablösen bzw. unterstützen. In der Forschung existieren verschiedene Ansätze mit Hilfe unterschiedlicher ML Methoden die Effektivität zu steigern, die Kosten zu senken, sowie AMR zu bekämpfen. Diese Techniken des ML können an verschiedensten Stellen des Entwicklungsprozesses

ansetzten. Einige Möglichkeiten des Einsatzes von ML-Methoden liegen bei der Entdeckung neuer antimikrobieller Verbindungen, im Verstehen der Angriffspunkte und Wirkweisen von Antibiotika, bei Analysen von Resistenzausbreitungen, sowie in der Entwicklung neuer Modelle zur sinnvollen Verwendung von antimikrobiellen Chemotherapeutika. Besonders im Hinblick auf die Entdeckung von neuen Antibiotika kann in unterschiedlichen Richtungen geforscht werden. Hier stehen vor allem das Screening von Datenbanken, Fragmentanalysen und -zusammensetzungen, der Einfluss des Zellmetabolismus und phänotypische Wirkweisen im Vordergrund. Auch die Techniken des ML weisen eine ausgedehnte Bandbreite auf. Methoden des Supervised Learning (dt.: überwachtes Lernen), sowie des Unsupervised Learning (dt.: unüberwachtes Lernen) kommen als Forschungshilfen in Frage. Insbesondere Deep Learning (dt.: tiefes Lernen) und probabilistische Modelle gelten als aussichtsreiche Kandidaten zur Prozessunterstützung in der Entdeckung neuer Antibiotika. Aufgrund der Vielfältigkeit der unterschiedlichen Herangehensweisen ist es bisher unklar, welche Ansätze in Zukunft verfolgt und weiter erforscht werden sollen.

Mit der vorliegenden Arbeit wird ein Überblick über existierende Ansätze für den Einsatz des ML in der Entdeckung von neuen Antibiotika gegeben. Diese werden hinsichtlich des Einsatzgebiets und der Methodik voneinander abgegrenzt und verglichen. Möglicherweise lassen sich hier Aussagen über den Erfolg der einzelnen Vorgehensweisen ableiten und Zukunftstendenzen erkennen.

Die zentrale Forschungsfrage lautet dementsprechend wie folgt: Welche Techniken des Maschinellen Lernens zur Entdeckung von Antibiotika existieren in Forschung und Praxis? Diese Arbeit zeigt die Ergebnisse eines systematischen Literaturreviews auf Basis einer biologischen, einer technischen und zwei allgemeinwissenschaftlichen Datenbanken. Somit gibt sie einen Überblick über die derzeit wichtigsten Einsatzgebiete von ML in der Antibiotikaentdeckung, welcher zur Orientierung für zukünftige Arbeiten zu diesem oder ähnlichen Themen dienen kann.

Der Rest dieser Arbeit ist wie folgt strukturiert. Der zweite Abschnitt stellt den theoretischen Hintergrund von Antibiotika und die Techniken des ML vor. Im Teilbereich Methodik wird detailliert auf den Forschungsansatz, einschließlich der verwendeten Recherchekriterien, eingegangen. Abschnitt vier präsentiert unsere Ergebnisse, worauf deren Diskussion folgt. Abschließend wird die Arbeit mit einem Zukunftsausblick abgeschlossen.

Theoretischer Hintergrund

Biologische Grundlage

Antibiotika leitet sich vom griechischen Wort anti bios (gegen das Leben) ab. Im Allgemeinen versteht man unter Antibiotika Chemotherapeutika mit antimikrobieller Wirkung, welche selektiv toxisch auf bakterielle Strukturen wirken (Fritsche 2016).

Die Stärke dieser Wirkung wird durch die MIC gemessen, welche die geringste wachstumshemmende Konzentration abbildet.

Aufgrund der Vielfältigkeit in ihrer chemischen Struktur lassen sich Antibiotika unterschiedlich klassifizieren. Zum einen unterscheidet man zwischen Schmalspektrumantibiotika und Breitbandantibiotika, letztere wirken gegen eine ganze Reihe von Mikroben. Zum anderen lassen sich Antibiotika nach ihrem Effekt auf die Erreger einteilen. Hier betrachtet man bakteriostatische Antibiotika, welche das Wachstum hemmen, indem sie die Vermehrung unterbinden, und bakterizide Antibiotika, welche die Bakterien abtöten. Des Weiteren ist die Vielfältigkeit der Angriffspunkte von Antibiotika in dieser Arbeit von großer Bedeutung. Folgende Angriffsziele werden im Allgemeinen definiert:

Angriffsziel	*Beschreibung*
Zellmembran	Veränderung der Permeabilität der Plasmamembran
Zellwandsynthese	Beeinflussung des Zellinnendrucks durch Behinderung des Aufbaus einer Peptidoglykanschicht
Proteinsynthese	Störung von Transkription und Translation
Replikation des Erbguts	Keine Zellteilung durch Störung der DNA-Replikation
Stoffwechselwege	Verhinderung der Synthese wichtiger Stoffe
Tabelle 1. Übersicht über die Gängigsten Angriffspunkte von Antibiotika (Fritsche 2016)	

Konzepte des Maschinellen Lernens

Maschinelles Lernen

Maschinelles Lernen ist ein Zweig der künstlichen Intelligenz, wodurch Lösungsvorschläge für komplexe Aufgaben gefunden werden (Nilsson 1998). Mit dem Computing werden Systeme entworfen, die durch Daten und Erfahrungen künstliches Wissen generieren. Durch das Erkennen von Mustern in vorliegenden Datenbeständen ist es möglich, genauere Modelle auf Basis dieser Daten zu erstellen. Diese spezifizierten Modelle ermöglichen es, Lösungen für Probleme zu finden.

Deep Learning

Ein Ansatz für das ML ist das Deep Learning (Goodfellow et al. 2015). Deep Learning ist eine besondere Art des ML, welcher große Kraft und Flexibilität erreicht, indem er lernt, die Welt als verschachtelte Hierarchie von Konzepten zu repräsentieren. Dieser Ansatz nutzt große Datenmengen und analysiert diese. Ihre Funktionsweise beruht auf das Lernen des menschlichen Gehirns. Mit Hilfe des neuronalen Netzes kann das Erlernte immer wieder mit neuen Inhalten verknüpft werden. Dies führt dazu, dass durch dieses Prinzip Prognosen und Entscheidungen getroffen werden können ohne, dass diese hinterfragt werden.

Neural Networks

Convolutional Neural Networks

Convolutional Neural Networks (CNN) sind eine spezielle Form von künstlichen neuronalen Netzen zur Verarbeitung von Daten, die eine bekannte gitterartige Topologie haben (Goodfellow et al. 2016). In mindestens einer ihrer Schichten wird die Faltung anstelle der allgemeinen Matrixmultiplikation verwendet. Die Faltung ist eine Art der linearen Operation.

Eine typische Schicht eines Convolutional Networks besteht aus drei Stufen (Goodfellow et al. 2016). In der ersten Stufe führt die Schicht mehrere Faltungen parallel durch, um eine Reihe von linearen Aktivierungen zu erzeugen. In der zweiten Stufe (oft auch Detektorstufe genannt) durchläuft jede lineare Aktivierung eine nichtlineare Aktivierungsfunktion. In der dritten Stufe wird eine Pooling-Funktion verwendet, um die Ausgabe der Schicht weiter zu modifizieren. Eine Pooling-Funktion ersetzt die Ausgabe des Netzes an einem bestimmten Ort durch eine zusammenfassende Statistik der nahegelegenen Ausgaben. Beispielsweise meldet die maximale Pooling-Funktion die maximale Ausgabe innerhalb einer rechteckigen Nachbarschaft. Wenn es mehrere Einheiten von Convolutional Layers, gefolgt von einem Pooling Layer gibt, bezeichnet man es als Deep Convolutional Neural Networks.

Fully Connected Neural Network

Diese Art von neuronalen Netzen ähnelt den traditionellen neuronalen Netzen, bei denen alle Neuronen miteinander verbunden sind. Der Output ist das Ergebnis der gewichteten Summe aller Outputs, die durch die zuvor verbundenen Neuronen gegeben sind und auf die eine Aktivierungsfunktion angewendet wird (Monteiro et al. 2020).

Autoencoder

Autoencoder sind künstliche neuronale Netzwerke, welche eine grundlegende Rolle beim unsupervised Learning und in tiefen Architekturen für Transfer-Lernen und andere Aufgaben spielen (Baldi 2012). Sie sind einfache Lernkreise, welche darauf abzielen, Eingänge in Ausgänge mit möglichst geringer Verzerrung umzuwandeln. Dabei werden verschiedene Merkmale von hoch-dimensionalen Daten extrahiert, um eine komprimierte Repräsentation zu erlangen. Der Ablauf besteht aus zwei Schritten; dem Encoder und dem Decoder. Der Encoder komprimiert den Input bis zu einem latenten Eigenvektor und der Decoder versucht den Input wieder zu konstruieren.

Variational Autoencoders (VAE) sind spezielle Encoder-Decoder-Architekturen, die versuchen, die Datenverteilung auf eine Weise zu lernen, die später zur Generierung neuer Beispiele abgetastet werden kann (Harel & Radinsky 2018).

Random Forest

Random Forest (RF) wird als eine Art Klassifikations- und Vorhersagemodell betrachtet (Lan & Pan 2019). Es besteht aus mehreren Entscheidungsbäumen, die sich jeweils leicht unterscheiden und die gemeinsam einen „Wald“ bilden. Die Methode nutzt die Ergebnisse der verschiedenen Classification Trees (CT), um eine bestmögliche Vorhersage treffen zu können. Die Klassifizierung neuer Daten erfolgt anschließend durch Betrachtung aller angepassten CTs sowie entsprechender Mehrheitswahl, d.h. die Klasse, welche am meisten vorhergesagt wurde, wird auch insgesamt prognostiziert (Pereira et al. 2015).

Methoden

Die Seminararbeit wird sich mit der Problemstellung in Form von einer systematischen Literaturrecherche beschäftigen. Dazu werden zunächst die relevanten Publikationen ermittelt und ausgewertet.

Zur Ermittlung der Quellen wurde eine stichwortbasierte Suche auf unterschiedlichen Online-Datenbanken durchgeführt. Diesbezüglich wurden biologische, allgemeine und informatikbasierte Datenbanken mithilfe einer Stichwortkombination durchsucht. Die Datenbanken ACM Digital Library (ACM), AIS EBSCOhost (EBSCO), IEEE Xplore Digital Library (IEEE) und PubMed wurden bei der Suche verwendet. Für die Suche wurden englische Begriffe festgelegt, da die zentralen Publikationen zum Thema des maschinellen Lernens bei der Antibiotikaentdeckung in englischer Sprache veröffentlicht wurden. Die Herleitung der Stichwortkombination erfolgte mithilfe einer Excel Tabelle, bei der die einzelnen Stickwortkombinationen der jeweiligen Kommilitonen ausgewertet und zu einem gemeinsamen Ergebnis führten. Zu Beginn wurden die drei Kernbegriffe „antibiotic“, „machine learning“ und „deep learning“ festgelegt. Diese stellen die zentralen Begriffe aus der Zielsetzung dar. Anschließend wurden die jeweiligen verwandten Begriffe für diese Kernbegriffe festgelegt und mit einem OR miteinander verknüpft, damit diese bei der Suche abgefangen werden. Die Verknüpfung der Kernbegriffe erfolgt durch jeweils ein AND. Durch dies wird gewährleistet, dass die ermittelten Publikationen einen Bezug zur Zielsetzung der Arbeit haben werden. Durch das Zusammensetzen der Begriffe und der Operatoren ergibt sich folgende Stichwortkombination zur Ermittlung der Quellen*: antibiotic*AND (machine learning OR deep learning OR neuronal network) AND discovery*.

Die initiale Suche ergab insgesamt 489 Ergebnisse. Bei den Datenbanken ACM und Ebsco musste bei der Suche eine zeitliche Beschränkung vorgenommen werden, da die Anzahl der Publikationen zu groß war. Da dieses Forschungsgebiet sehr neu ist und ein Großteil der für die Publikation relevanten Forschung nach 2019 publiziert wurde, lässt sich die Annahme treffen, dass die Menge an wissenschaftlichen Arbeiten sehr begrenzt ist. Es ist davon auszugehen, dass die aktuellsten Publikationen zum Thema in Summe alle relevanten, vorher publizierten Arbeiten aufgreifen. Auf Basis dieser Annahme und unter Betrachtung des zeitlichen Rahmens dieser Arbeit, wurde die Suche der Artikel zwischen Dezember 2019 und Juni 2020

eingeschränkt. Ältere Publikationen werden durch die Suche in PubMed und IEEE, sowie die weiterführende, detailreichere Recherche in den später identifizierten Kategorien abgedeckt. Nach dieser zeitlichen Beschränkung ist die Anzahl an Publikationen bei der Datenbank EBSCO auf 281 gesunken und bei ACM auf 121. Zudem kam es zu 5 Ergebnissen bei IEEE und zu 81 Ergebnissen bei PubMed. Die übrig geblieben Ergebnisse wurden nach ihrer Relevanz beurteilt. Während des Schreibens wurde zudem eine Rückwärtsrecherche durchgeführt, bei der anhand des Literaturverzeichnisses vorhandener Publikationen die Literaturanzahl vergrößert wurde. Die zuvor identifizierten Artikel präsentierten Techniken des ML, bei denen jedoch nicht genügend Informationen über Quantitative-Structure-Activity-Relationship (QSAR) und das Screenen von Datenbanken gefunden wurden, durch die Rückwärtsrecherche wurden diese Informationslücken gedeckt.

Zunächst wurden die wissenschaftlichen Arbeiten anhand Titel und Abstract überprüft, bevor im weiteren Verlauf die Relevanz der Volltexte betrachtet wurde. Es wurden nur Arbeiten in die Recherche miteinbezogen, die sich eindeutig auf die Methoden des ML bei der Entdeckung neuer Antibiotika beziehen. Aus zeitlichen Gründen wurden deshalb hauptsächlich Publikationen eliminiert, die sich mit ähnlichen und umfangreichen Konzepten befassen, wie z.B. Antibiotikaresistenz und Peptide. Die Suche auf unterschiedlichen Datenbanken führte dazu, dass Publikationen mehrmals ermittelt wurden. Mithilfe von dem Literaturverwaltungsprogramm Zotero, wo alle ermittelten wissenschaftlichen Arbeiten hochgeladen wurden, wurden alle Duplikate entfernt. Nach dem Entfernen reduzierte sich die Anzahl der zu betrachtenden Arbeiten auf 30.

Anschließend wurden die Publikationen nach Relevanz der Arbeit, ausreichender Hintergrund über die eingesetzten Techniken des ML, Aktualität und Erfolg der Anwendung des Verfahrens analysiert. Mit einer weiteren Excel Tabelle wurde geprüft, welche Methoden in den jeweiligen wissenschaftlichen Arbeiten behandelt wurden und dementsprechend kategorisiert.

Um Gemeinsamkeiten und Diskrepanzen zwischen den Methoden des ML bei der Entdeckung neuer Antibiotika zu erkennen, wurde eine qualitative Inhaltsanalyse durchgeführt. Zu den ermittelten Daten bei der Inhaltsanalyse gehören zum einen die Methoden des ML, die Prozessschritte bzw. Ziele und zum anderen die Kategorie der Publikationen.

Die Vorgehensweisen, die gleiche oder ähnliche Prozesstritte bzw. Einsatzgebiete aufweisen, wurden gruppiert, in dem eine einheitliche Bezeichnung festgelegt wurde. Zudem wurden auch Methoden ermittelt, die nur einzeln vorkamen und zu keiner Gruppe zugeordnet werden konnten. Diese wurden aufgrund geringer Information nicht in den Ergebnissen aufgezeigt und bearbeitet. Die ermittelten Gruppen stellen dabei Techniken des ML für unterschiedliche Einsatzbereiche in der Antibiotikaentdeckung dar und werden im folgenden Kapitel beschrieben.

Methoden des Maschinellen Lernens in der Entdeckung von Antibiotika

Virtuelles Screening: Prädiktive Methoden

Quantitative-Struktur-Wirkungs-Beziehung Methoden

QSAR ist ein Spezialfall des virtuellen Screenings, welcher einen Forschungsansatz bei der Entwicklung neuer Antibiotika darstellt. Target ist ein Biomolekül, an welches sich ein Wirkstoff binden kann. Die Wirkung des Wirkstoffs kann sich dadurch entfaltet (Hanzlik et al. 2007). Dieser Ansatz gehört zu den Methoden des Wirkstoffdesigns, der einen Wirkstoff über eine Datenbankstruktur identifiziert oder aufgrund der Kenntnis der Struktur des Targets herstellt und anschließend synthetisiert (Batool et al. 2019). Bei der Vorhersage von biologischen Profilen, der Generierung von Wirkstoffen und der Identifizierung von Treffern nimmt die Methode einen besonderen Stellenwert ein. QSAR beschreibt das quantitative Verhältnis zwischen chemischen Substanzen und ihren biologischen Aktivitäten. Ziel ist es, Vorhersagen zu chemischen und biologischen Eigenschaften neuartiger Moleküle zu treffen, um diese anschließend strukturell darstellen zu können (Patel et al. 2014). Dieses Verfahren ist insbesondere für die Arzneimittelentdeckung essenziell, da gleichartige Strukturen in der Regel ähnliche Wirkweisen zeigen. Die Besonderheit gegenüber anderen Molekülstrukturen ist, dass durch QSAR sowohl in 2D als auch in 3D

modelliert werden kann (Danishuddin & Khan 2016). Die 2D-Methode beruht auf der Berechnung und der Gegenüberstellung molekularer Eigenschaften, wohingegen die 3D-Methode strukturbasiert ist.

Im Rahmen der QSAR-Studien haben sich quantenchemische Deskriptoren bei der Vorhersage antibiotischer und biologischer Aktivitäten als nützlich erwiesen (Pereira et al. 2015).

Das Erstellen eines effektiven Modells für die Entdeckung von Antibiotika mithilfe von QSAR-Deskriptoren erfordert zunächst das Extrahieren von Trainings- und Testsätzen aus Datenbanken (Pereira et al. 2015). Bei diesem Vorgang wird auf bestimmte Arten von biologischen Aktivitäten, wie z.B. die antibiotische Wirkung, geachtet. Die durch die Datenbank ermittelten chemischen Strukturen werden als SMILES Strings gespeichert. SMILES ist eine chemische buchstabenbasierte Notationssprache, die speziell für den Computergebrauch von Chemikern entwickelt wurde (Weininger 1988). Die Aufteilung der Datensätze in Trainings- und Testdatensätze erfolgt zufällig (Pereira et al. 2015). Um Ähnlichkeiten zwischen Datensätzen von unterschiedlichen Verbindungen zu zeigen, wird die Principal Component Analysis (PCA) verwendet (Pereira et al. 2015). PCA dient dazu, große Datensätze zu vereinfachen und zu strukturieren, indem die statistischen Variablen ihren Hauptkomponenten genähert wird (Tibaduiza et al. 2016). Diese weitverbreitete Technik soll einen allgemeinen Eindruck von allen Datensätzen liefern, die von der QSAR verwendet wird.

Da die biologischen und antibiotischen Aktivitäten von Verbindungen nicht einfach quantifiziert werden können, wird mithilfe von molekularen Deskriptoren dieser Vorgang vereinfacht. (Cherkasov 2005).

Ein Deskriptor stellt die chemischen Eigenschaften eines Moleküls dar (Danishuddin & Khan 2016). Diese Darstellung erfolgt in einer numerischen Form. Deskriptoren spiegeln eine Vielzahl an Aspekten intra- und intermolekularer Wechselwirkungen wider. Die in den Deskriptoren enthaltenen Informationen hängen von der molekularen Darstellung und dem definierten Algorithmus ab. Damit ein Modell mit statistischen Methoden erstellt werden kann, müssen die Deskriptoren invariant zu der Größe des Moleküls und der Anzahl der enthaltenen Atome sein.

Für den Einsatz der Deskriptoren im QSAR-Modell werden entsprechende Auswahlkriterien zu Grunde gelegt (Danishuddin & Khan 2016). Diese Auswahlkriterien können Techniken wie euklidischer Abstand, gegenseitige Informationen oder Korrelationsmethoden sein. Eine Möglichkeit, so eine Auswahl zu treffen, sind Correlation-Based Feature Selection (CFS) - Filter (Pereira et al. 2015). CFS ist ein Algorithmus, der die Merkmalsauswahl für ML durch einen korrelationsbasierten Ansatz bearbeitet (Hall 1999). Dieser evaluiert die Nützlichkeit einzelner Deskriptoren für die Prognose der gegebenen biologischen Aktivitäten. Der korrelationsbasierte Algorithmus, wird durch Experimente ausgewertet, bei denen künstliche und natürliche Datensätze verwendet werden. Mithilfe der künstlichen Datensätze werden irrelevante und redundante Merkmale schnell identifiziert und selektiert. Relevante Merkmale kennzeichnen sich dadurch aus, dass ihre Relevanz nicht von anderen Merkmalen abhängt.

Um antibiotische Aktivitäten vorhersagen zu können, wird ein Modell erstellt, das unter Verwendung von internen und externen Bewertungen das QSAR- Modell validiert (Pereira et al. 2015). Vor diesem Hintergrund werden ML-Methoden wie z.B. CT oder RF gegenübergestellt. Dabei werden die antibiotischen Wirkungen der aktiven und nicht aktiven Verbindungen prognostiziert.

Durch den Einsatz des k-Nearest Neighbors (kNN) - Algorithmus, können Vorhersagen über aktive Verbindungen durch den Mehrheitsbeschluss der k ähnlichen Verbindungen im Trainingssatz getroffen werden (Pereira et al. 2015).

Ein CT ist eine grafische Darstellung möglicher Ergebnisse bzw. Auswirkungen zusammenhängender Entscheidungen. Diese ermöglichen es Lösungsansätze, Entscheidungsmöglichkeiten und Vorhersagen bester Entscheidungen zu treffen (Ganvankar & Sawarkar 2017). CTs werden durch einen iterativen Prozess erstellt und sequenziell aufgebaut. In der Regel beginnt ein CT mit einem einzelnen Knoten, von dem Verzweigungen ausgehen, welche die jeweiligen Entscheidungen darstellen. Diese Verzweigungen werden durch entsprechende Regeln erzeugt, wodurch jeder Knoten durch einen einzelnen Deskriptor definiert wird. Es werden ca. 200 - 250 Deskriptoren verwendet, um die Vorhersage der antibiotischen Wirkungen zu treffen (Pereira et al. 2015). Wenn die Werte an den Verzweigungen unterhalb eines bestimmten Deskriptorwertes sind, werden diese einem untergeordneten Knoten zugeordnet, liegen diese jedoch oberhalb, werden sie einem anderen Zweig zugeordnet. Damit ein CT Vorhersagen bezüglich antibiotischer

Wirkungen treffen kann, müssen gewisse Eigenschaften und Entscheidungskriterien mit einer bestimmten Anzahl an Deskriptoren ermittelt werden.

Neben den CTs wird auch der RF-Algorithmus verwendet, der zu Beginn die Trainingssätze auswählt und anschließend die geteilten Knoten ermittelt (Pereira et al. 2015). Die RF- Modelle werden mit 500 Bäumen und unterschiedlichen Ansätzen von Deskriptoren durchgeführt, wobei die Anzahl der Deskriptoren durch die Aufteilung der Knoten definiert wird. Je nach gesuchter Aktivität, z.B. antibiotischer Aktivität variiert die Anzahl der Deskriptoren im Modell. Zudem bestimmt die Methode die Wichtigkeit eines Deskriptors durch die Zunahme der Fehlklassifizierung, die dann auftritt, wenn die Werte des Deskriptors ständig verändert werden. Wie auch bei den CTs hängt die Vorhersage beim RF bezüglich der antibiotischen Wirkungen von den Eigenschaften und Entscheidungskriterien der einzelnen CTs ab, welche eine unterschiedliche Anzahl an Deskriptoren benötigen.

Durch den Vergleich der Ergebnisse der jeweiligen ML- Methoden kann gezeigt werden, dass der implementierte rechnergestützte Ansatz unter Verwendung von Deskriptoren zur Vorhersage biologischer und antibiotischer Aktivitäten bereits vorhandener oder neuer Molekülstrukturen erfolgreich verwendet werden kann (Pereira et al. 2015).

Die QSAR bildet die Basis des Virtuellen Screenings. Im Folgenden Kapitel wird die Bedeutung des Virtuellen Screenings im Rahmen des QSAR dargestellt.

Erfolgreiche Erweiterungen Klassischer Methoden

Um neue Antibiotika zu entdecken, reicht einfaches Screening von existierenden Datenbanken kaum aus, weswegen die Methoden ergänzt werden müssen.

Virtuelles Screening kann nicht nur anhand von molekularen Deskriptoren durchgeführt werden, sondern auch anhand des bakteriellen phänotypischen Fingerabdrucks (BPF) (Zoffmann et al. 2019). Der BPF beschreibt die morphologischen Änderungen, die Verbindungen ab der kleinsten effektiven Dosis (LOED) bei Bakterien hervorrufen. Die LOED ist in der Regel kleiner als die MIC, wodurch unter anderem Verbindungen entdeckt werden können, die zu typischen Screeningkonzentrationen zwar nicht antimikrobiell wirken, aber dennoch morphologische Veränderungen hervorrufen, sodass einerseits mit höheren Konzentrationen getestet werden kann oder andererseits sie durch medizinische Chemie zu Antibiotika weiterentwickelt werden könnten. Des Weiteren können durch virtuelles Screening auf Basis des BPF nicht nur Aussagen darüber getroffen werden, ob die Verbindungen wirken, sondern auch auf welche Art und Weise, womit Informationen für die Analyse der Struktur-Wirkungsbeziehung gewonnen werden. Ein RF - Modell kann mit dem Ziel genutzt werden, die zu testenden Verbindungen und ein Set an Referenzantibiotika in Gruppen gleicher Wirkungsweisen zu Clustern. Beim Trainieren des Modells wurde die Anzahl der Bäume bei n = 1000 konstant gehalten und mtry wurde zwischen drei Werten variiert. Mit einer Out-of-bag Validierung, welche den erwarteten Prädikationsfehler eines Modells schätzt (Cho et al. 2019), wurde gezeigt, dass das Modell korrekt klassifiziert. Ergebnis der RF Klassifizierung ist eine 3D-Projektion der Distanzmatrix, in der klare Cluster mit klaren Abgrenzungen zu erkennen sind. Beim Screening mit A. baumannii – einem von zwei getesteten Bakterienstämmen – konnte sogar ein Cluster identifiziert werden, was eine unterschiedliche Wirkungsweise zu bekannten Antibiotika besitzt.

Eine weitere Möglichkeit virtuelles Screening zu verbessern ist der Einsatz von Deep Learning. Durch ein direct-message passing deep Neural Network (D-MPNN) können Moleküle akkurater dargestellt werden als durch manuell erstellte Deskriptoren oder Fingerprints (Yang et al. 2019b). Die Neuronen dieses D-MPNN stellen einfach zu berechnende Molekülmerkmale dar. Das Modell beschränkt sich nicht auf die Darstellung der Moleküle mit Faltungen oder Deskriptoren, sondern nutzt Faltungen, wodurch die Flexibilität beim Lernprozess hoch ist, und Deskriptoren, so dass das Modell auf eine starke Grundlage aufbaut. Da das Modell die Daten direkt von der Graphendarstellung der Moleküle erhebt, lässt es sich in die Kategorie Graph Convolutional Neural Networks (GCNN) einordnen. Der Unterschied zu klassischen message passing Neural Networks besteht darin, dass die Nachrichten nicht von Neuron zu Neuron weitergeben werden, sondern von Bindung zu Bindung, um unnötige Iterationen zu vermeiden und so die Performance zu verbessern. Beim Trainieren des D-MPNN entstehen durch Iterationen Verbindungen höherer Ebenen, die Informationen über die Verbindungen und somit auch Neuronen in ihrer Nähe enthalten. Diese können dann in einen kontinuierlichen Vektor übersetzt werden, der ein komplettes Molekül darstellt. Der Vektor wird anschließend noch mit Molekülmerkmalen ergänzt, die mit dem RDKit

– eine Open Source Software für Chemoinformatik mit der unter anderem Deskriptoren für ML generiert werden können (Landrum 2006) – berechnet wurden. Das so entwickelte D-MPNN wurde mit 2335 strukturell diversen Molekülen trainiert und auf drei unterschiedliche Datenbanken angewandt (Stokes et al. 2020). Nach jeder Anwendung wurde das D-MPNN mit den daraus gewonnenen Daten neu trainiert. Das Modell berechnet eine Punktzahl zwischen 0 und 1 (im folgenden prediction Score), wobei die 0 keine antibakterielle Aktivität und die 1 antibakterielle Aktivität repräsentiert. Ein Vergleich mit fünf weiteren Methoden zeigt, dass das Modell gut geeignet ist, um strukturell diverse, antibakterielle und zu existierenden Antibiotika unähnliche Verbindungen zu finden. Das erste Screening fand mit dem Drug Repurposing Hub, eine Bibliothek, bestehend 6111 Molekülen in unterschiedlichen Stadien der Medikamentenentwicklung (Stokes et al. 2020), statt, wobei aus den 6111 Molekülen 99 Verbindungen mit dem höchsten prediction Score ausgewählt wurden und auf antibakterielle Aktivität gegen E. coli getestet wurden. Von den 51 aktiven Molekülen wurden prediction Scores, Forschungsstadium, strukturelle Unähnlichkeit zum Trainingsset und Toxizität betrachtet und Halicin wurde in allen Bereichen als gut eingeschätzt. Mit Halicin wurden in vitro und in vivo Tests durchgeführt, die eine Breitbandaktivität gegen mehrere Bakterienarten und Resistenzgene in E. coli suggerieren. Das Anwenden auf die zweite Datenbank, welche Moleküle enthält, die sich strukturell deutlich vom Trainingsset unterschieden, zeigt, dass das Modell generalisieren kann. Hier wurden keine aussichtsreichen Kandidaten identifiziert. Mit der dritten Datenbank ZINC15, welche eine virtuelle Sammlung von ca. 1,5 Milliarden Molekülen für in silico Screening ist, wurde die Anwendungen auf große Maßstäbe getestet. Von den 1,5 Milliarden wurden ca. 107 Millionen Moleküle mit antibiotikaähnlichen physikalisch-chemischen Eigenschaften für das Screening ausgewählt. Aus den 3260 Verbindungen mit prediction Scores > 0,8 wurden die 23 mit einer Tanimoto Ähnlichkeit < 0,4 für empirisches Testen ausgewählt. Davon zeigen acht eine Wachstumsinhibition von mindestens einem der fünf getesteten Bakterienarten und zwei zeigen eine Breitbandaktivität.

Virtuelles Screening von Molekülfragmenten

In der wissenschaftlichen Literatur wird neben der Antibiotika-Entdeckung der Fokus auch auf die Molekularstrukturerkennung gelegt. In diesem Themenbereich zeichnet sich der Ansatz Fragment-Based-Drug-Discovery (FBDD) mit seiner Effizienz und seinen aussichtsreichen Ergebnissen aus. FBDD ist ein Ansatz, welcher zunehmend in der pharmazeutischen Industrie eingesetzt wird. Insbesondere zur Verringerung der Attrition und Lieferung von Leitfaden für zuvor schwer zu handhabenden biologischen Zielen (Murray & Rees 2009). FBDD identifiziert Liganden mit niedrigem molekularem Gewicht (~150 Da), die sich an biologisch wichtigen Makromolekülen anbinden (Murray & Rees 2009). Der dreidimensionale experimentelle Bindungsmodus dieser Fragmente wird bestimmt unter Verwendung von Röntgenkristallographie oder Kernspinresonanzspektroskopie. Dieser wird für die Optimierung in potente Moleküle mit arzneimittelähnlichen Eigenschaften verwendet (Murray & Rees 2009). Abbildung 1 verschafft einen Überblick über die Effizienz des FBDDs Ansatz.

Im Vergleich zum Hochdurchsatz-Screening erfordert der Fragment-Ansatz weniger Screening der Verbindungen und bietet trotz der geringeren Anfangspotenz der Screening-Treffer, effizientere und fruchtbarere Optimierungsaktionen (Murray & Rees 2009). Zudem decken Fragment basierte Bibliotheken einen größeren chemischen Raum ab (Tam et al. 2019).

Der Algorithmus Hunting FOX, „Hunting FOX“ für „Hunting Fragments of X“, stellt ein gutes Beispiel für die Methodik des Fragment-Ansatzes dar und untermauert die Wichtigkeit der Neuentdeckung von niedermolekularen Wirkstoffen mit erwünschten Eigenschaften, die später als Baustein für die Medikament Entwicklung dienen.

Ein wichtiger Schritt bei der Suche nach niedermolekularen Wirkstoffen ist die Identifizierung neuer chemischer Zuleitungen (Murray & Rees 2009). Der Hunting FOX Ansatz entspricht einem konzeptionellen Beweis, indem auf rationale Weise ein chemisches Vokabular identifiziert wird (Mansbach et al. 2020). Dieses „chemische Vokabular“ steht in Zusammenhang mit einer bestimmten Arzneimittelaktivität von Interesse, ohne bekannte Regeln anzuwenden (Mansbach et al. 2020). Der Hunting FOX Algorithmus identifiziert automatisch eine Reihe von relevanten Fragmenten, welche für hybrid Fragment-basierten Entwicklung von Molekülen in Frage kommen. Diese sind in der Lage, P. aeruginosa (Stäbchenbakterium) zu durchdringen (Mansbach et al. 2020). Im Gegensatz zu dem herkömmlichen FBDD Ansatz, berücksichtigt der Hunting FOX Algorithmus alle möglichen Fragmente innerhalb einer Reihe von

Verbindungen, von einer Einfachbindungslänge im Radius um ein Zentralatom bis zu 10 Bindungslängen (Mansbach et al. 2020).

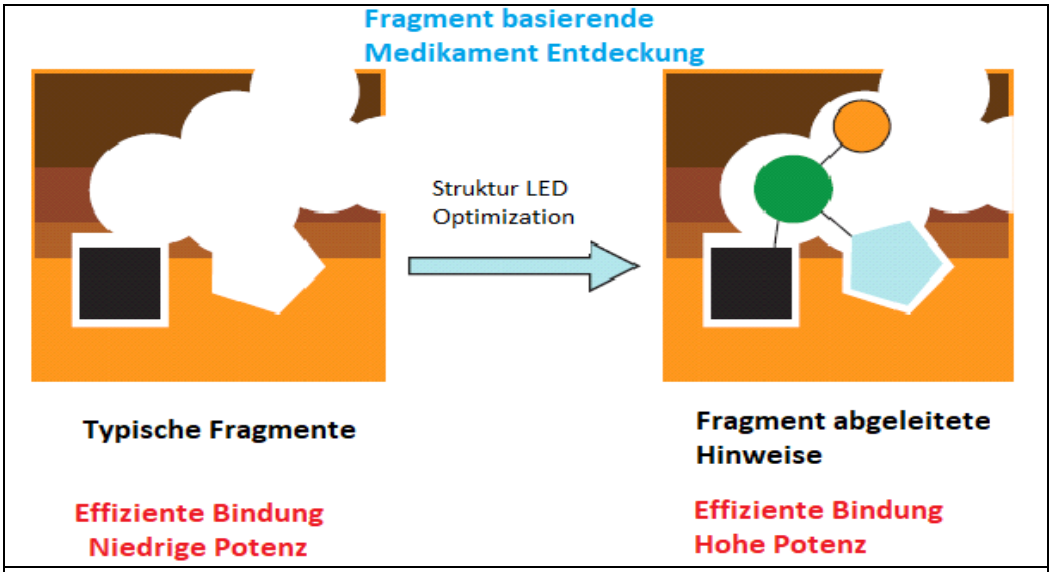

Abbildung 1. Übersicht über FBDD (Quelle: Murray & Rees, 2009)

Der Gesamtworkflow des Algorithmus besteht aus vier Teilen: (i) Definition einer Repräsentation für die Zusammensetzung; (ii) Experiementelle Messung und Datenbereinigung für eine ausgewählte Teilmenge des kuratierten Datensatzes, um die Eingabe der Arzneimittelaktivität in den Algorithmus zu ermöglichen; (iii) Durchführung einer Feature Selektion zur Identifizierung eines Vokabulars mit relevanten submolekularen Fragmenten; (iv) Anpassung eines Vorhersagemodells basierend auf dem identifizierten Vokabular (Mansbach et al. 2020). Abbildung 2 stellt diesen Aufbau Graphisch dar.

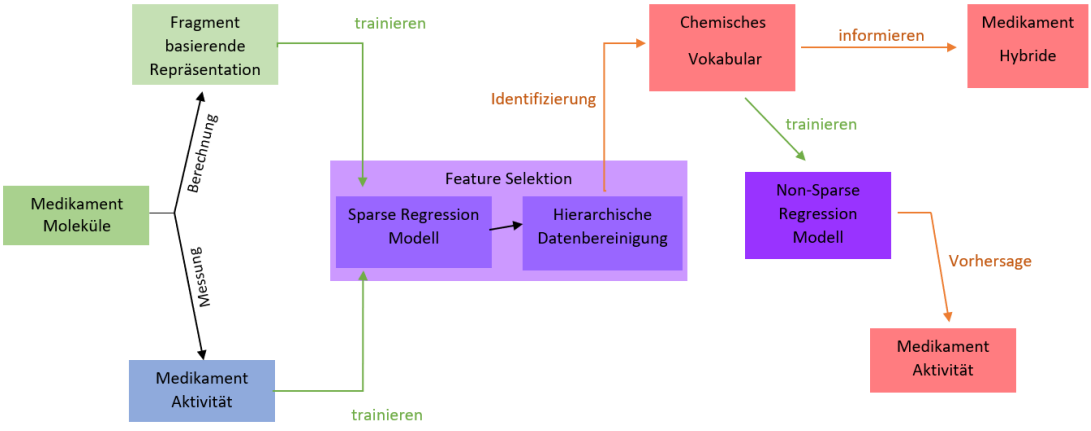

Abbildung 2. Aufbau des Hunting FOX Algorithmus (Quelle: Mansbach et al., 2020)

Die Definition einer Repräsentation für die Zusammensetzung erfolgt dadurch, dass eine 2-dimensionale Darstellung der Moleküle aufgestellt wird. Unter Betrachtung jedes einzelnen Atoms, werden die Fragmente bestimmt, die aus dem jeweiligen Atom plus jener Atome, welche sich im Radius von k ($1 \leq k \leq 10$) Bindungen befinden, bestehen (Mansbach et al. 2020). Insgesamt ergeben sich 22.139 verschiedene Fragmente, die den Trainingssatz von 595 Molekülen bilden (Mansbach et al. 2020). Jedes Molekül M wird als ein $N_f = 22{,}139$ langer Häufigkeitsvektor repräsentiert, bei dem jeder Eintrag der Summe des Eintretens

eins bestimmten Fragments entspricht (Mansbach et al. 2020). Schließlich werden diese mit der Anzahl an Atomen in dem Molekül normiert (Mansbach et al. 2020).

Die Basis für die hoch intrinsische Resistenz bei P. aeruginosa ist die niedrige äußere Membrane Permeabilität, gekoppelt mit sekundär Resistenz Mechanismen wie Antibiotische Efflux pumpen (Hancock 1998). Efflux pumpen sind Transportproteine, die an der Extrusion von toxischen Substraten, einschließlich nahezu alle Klassen von klinisch relevanten Antibiotika, vom Zellinneren in die äußere Umgebung beteiligt sind (Webber & Piddock 2003). Der Hunting FOX Algorithmus verwendet zur Trennung der Auswirkungen von Efflux pumpen und äußere Membranpermabilität, kürzlich erstellte verschiedene Mutanten-Stämme von gramnegativ Bakterien (Mansbach et al. 2020). Nach dieser Kategorisierung werden die vorhandenen Daten einer Feature Selektion unterzogen. Diese setzt sich aus 2 Schritten zusammen: (i) permissive LASSO, Regulierungen, um nicht prädiktive Variablen zu eliminieren und (ii) hierarchische Bereinigung, um verbleibende Redundanzen zu beseitigen (Mansbach et al. 2020). LASSO (least absolute shrinkage and selection operator) ergänzt, die Defizite des klassischen Linearen Model Selektionen, indem es die Auswahl der Variablen und die Schätzung der Parameter gleichzeitig durchführt (Wang et al. 2007). Die resultierenden Daten liefern wahrscheinlich aktive Fragmente, die anschließend nach möglichen molekularen Beiträgen überprüft werden. Dafür wird eine voreingenommene grobkörnige Simulation der Molekulardynamik ausgeführt, bei der ein MARTINI-Model Kraftfeld eingesetzt wird (Mansbach et al. 2020), welches mehrere Atome zu einer „virtuellen" Perle zusammenfasst, die durch ein effektives Potenzial interagieren (Monticelli et al. 2008).

End-to-End Konzepte

Prototypbasierte Arzneimittelforschung

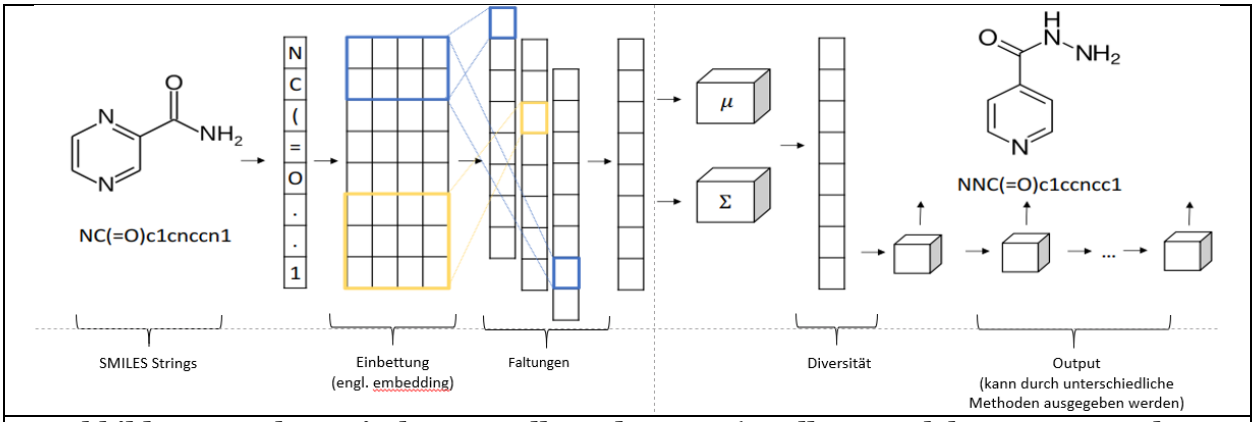

Abbildung 3. Schematische Darstellung des CDN (Quelle: In Anlehnung an Harel & Radinsky, 2018)

Ein weiterer Forschungsansatz in der Entdeckung von Antibiotika sind die Verwendung von sog. Komplettlösungen (engl. End-to-end) für neuronale Netze. Diese Modelle können beispielsweise in der prototypbasierten Medikamentenentdeckung verwendet werden, insbesondere für die Antibiotikaforschung. Bei einem Prototyp-basierten Verfahren startet man von einem Molekül (dem Prototyp), dass bereits gewünschte Eigenschaften hat und versucht eine chemische und strukturell ähnliche Verbindung ausgehend von dem Prototyp zu bekommen (Harel & Radinsky 2018). Da der chemische Raum von medikamentenähnlichen Molekülen schätzungsweise zwischen 10^{23}- 10^{60} Verbindungen liegt (Polishchuk et al. 2013), sind Algorithmen gefragt, welche automatisch potenzielle Moleküle ausgehend von dem Prototyp generieren. Eine Möglichkeit dies zu realisieren ist die Verwendung von Conditional Diversity Networks (CDN) (Harel & Radinsky 2018). CDN sind ein unüberwachte ganzheitliche Methode des maschinellen Lernens, welche speziell für die Thematik von prototypbasierten Verfahren entwickelt wurden. Abbildung 3 zeigt die schematische Darstellung der Architektur. Die Hypothese dieses Konzeptes ist, dass eine Verzerrung der Molekülbildung in Richtung bekannter Medikamente zu validen Molekülen führt. Im Wesentlichen beruht der Ansatz auf VAE, welche mit einer Diversitätskomponente erweitert wird. VAE ermöglichen die Probenahme von Molekülen, die näher an einem Arzneimittelprototyp liegen, und erhöht somit die Wahrscheinlichkeit, ein validiertes Arzneimittel mit ähnlichen Eigenschaften zu erzeugen.

Damit die Stichprobe sich ebenfalls vom Input (den Prototypen) unterscheidet, wird eine Diversitätskomponente eingeführt. Die Prototyp-getriebenen Hypothesengenerierung wird als ein Problem eines bedingten Datenerzeugungsprozess definiert. Das Modell arbeitet mit einem gegebenen Molekül-Prototyp und erzeugt verschiedene Moleküle als Kandidaten. Um medikamentenähnliche Moleküle zu identifizieren, wird nach dem Lipinski Kriterium vorgegangen, welches ein gebräuchliches qualitatives Maß für das Design chemischer Medikamente ist (Lipinski 2000).

Die CDN-Architektur beginnt mit der Codierung der Moleküle in SMILES Strings über die Encoder-Funktion (Harel & Radinsky 2018). Daraufhin wird jedes Zeichen in der SMILES-Darstellung in seiner Dimension d eingebunden und Faltungen über verschiedene Teilstring-(Filter-)Größen (z. B. chemischer Substrukturen) werden angewendet. Die extrahierten Merkmale werden verkettet und verbundene Schichten werden dem Encoder übergeben, der als Ausgabe einen gemittelten Vektor und einen Vektor der Standardabweichung berechnet. Diese stellen die Verteilung der Merkmale für den Prototyp da. Im VAE werden die Vektoren in einen Decoder eingespeist, welcher versucht, die Rekonstruktion der ursprünglichen Eingabe zu optimieren und eine der zuvor bekannten Darstellung zu erhalten. Während der Generierung werden die Merkmalsvektoren aus der früheren Verteilung abgetastet, und ihre Ausgabe wird an den Decoder weitergeleitet, der eine neue Darstellung erzeugt. Der VAE Generationsprozess wird zudem erweitert durch eine Diversitätsschicht, welche verrauschte Daten einer Stichprobenverteilung generiert. Dies sind somit die Daten des Encoders, jedoch mit größerer Varianz, wodurch der Molekülraum um das Ursprungsmolekül mit veränderbarer Diversität des Raumes untersucht werden kann. Das entspricht der Variabilität im chemischen Raum. Die Ausgabe der Diversitätsschicht ist eine Probe aus einer bedingten Diversitätsverteilung, die im Folgenden beschrieben wird:

Diverse z = (n x $\hat{\sigma}_i$) + $\hat{\mu}_i \sim$ N ($\hat{\mu}_i$, $\hat{\sigma}_i{}^2$ x D) mit n $\sim$ N (O, D)

Dabei ist z die latente Darstellung des Moleküls (hier mit Einberechnung der Diversität), n der verrauschte Datensatz der Stichprobe, verteilt mit der Gauß-Verteilung und dem Diversitätsparameter D und $\hat{\sigma}_i$, $\hat{\mu}_i$ die Standardabweichung bzw. Mittelwert der i-ten Stichprobe (Harel & Radinsky 2018).

Diese Diversitätsdarstellung (Diverse z) ist nun der Input für den Decoder zu Moleküldarstellung (Harel & Radinsky 2018). Der Decoder ist ein rekursives, generatives neuronales Netz (LSTM), welches das Ausgangssignal des Encoders als Input nimmt. Unter der Long short-term memory (LSTM) versteht man eine Technik, welche mit dem Gradientenverfahren zum Trainieren von neuronalen Netzen verwendet wird und dabei Kurzstreckenabhängigkeiten zwischen den Proben der Teilsequenzen der Trainingsdaten lernen (Sherstinsky 2020). Den ersten Zustand des Decoders bildet die kodierte Darstellung. Die Verbindungen setzt sich dann sequenziell zusammen, indem er die Verteilung über die Zeichen in jedem Zeitschritt auf der Grundlage seines aktualisierten Zustands und des Eingangszeichens aus dem vorherigen Schritt bearbeitet.

Beim Trainieren des Decoders werden richtige nächste Symbole (der Molekülstruktur) dem System gegeben, auch wenn es falsch vorhergesagt wurde (Harel & Radinsky 2018). Während der Generierung des nächsten Symbols wird entweder nach der Auswahl des am besten bewerteten Zeichen (argmax) oder durch Sampling aus derselben Verteilung vorgegangen. Zum Erhalten von gültigen syntaktischen SMILES Strings werden die Rekonstruktionsfehler minimiert, die beim Darstellen des Prototyps auftreten, damit so das Model die Koordinaten der wichtigsten Daten für die Variation der chemischen Umgebung erlernt. Die Darstellung der unteren Dimension, die in einen gültigen Prototyp rekonstruiert werden kann, gibt dem System Informationen über eine (valide) Darstellung, welche zum Trainieren von gültigen SMIILES Strings genutzt werden kann. Mit dieser Vorgehensweise wird umgangen, dass das Model Wissen über die Zusammensetzung von Atomen, Ringen oder Ähnlichem hat. Es funktioniert somit ohne das Verständnis von Medikamenten, sondern nur durch das Trainieren von Datensätzen mit medikamentenähnlichen Molekülen, wo keine bekannten Medikamente enthalten sind.

Wechselwirkung mit dem Wirkstoff (DTI)

Der Ansatz verfolgt eine ganzheitliche Deep Learning Model Lösung, welche aus einer Kombination von Deep Learning Architekturen, dem CNN und einem FCNN besteht (Monteiro et al. 2020). Abbildung 4 zeigt die Architektur dieses Systems. Ziel ist das Prognostizieren der Interaktion zwischen dem Medikament und dem Zielort, wobei das Ergebnis eine positive oder negative Wechselwirkung in binärer Darstellung (0/1) aufzeigen soll. Die Rohdaten, bestehend aus Protein-Aminosäuresequenzen (Zielort) und SMILES Strings

(Medikament), werden aussortiert, wenn ihre Wortlänge einen zuvor definierten Schwellwert überschreitet. Damit wird erzielt, dass jeder Input durch die gleiche Anzahl charakterisiert ist und der entsprechenden Art des Merkmales. Daraufhin können beide Inputs in eine ganzzahlige Zahl kodiert werden. Dabei werden die Proteine durch eine Protein-Substitutionstabelle (Yu et al. 2010) in Gruppen nach physikalisch-chemischen Eigenschaften kodiert und die SMILES Strings, beruhend auf einem Wörterbuch mit 32 Kategorien (Anzahl der verschiedenen Zeichen), in ganzheitliche Zahlen umgewandelt (Monteiro et al. 2020). Die Werte entsprechen nun verschiedenen Kategorien. Damit größeren Zahlen im Modell keinen höheren Einfluss zugewiesen wird, wird eine ein One-Hot Kodierungsschicht (auch 1-aus-n-Code genannt) angewendet, welcher die Zahlen jeweils als einen binären Vektor darstellt.

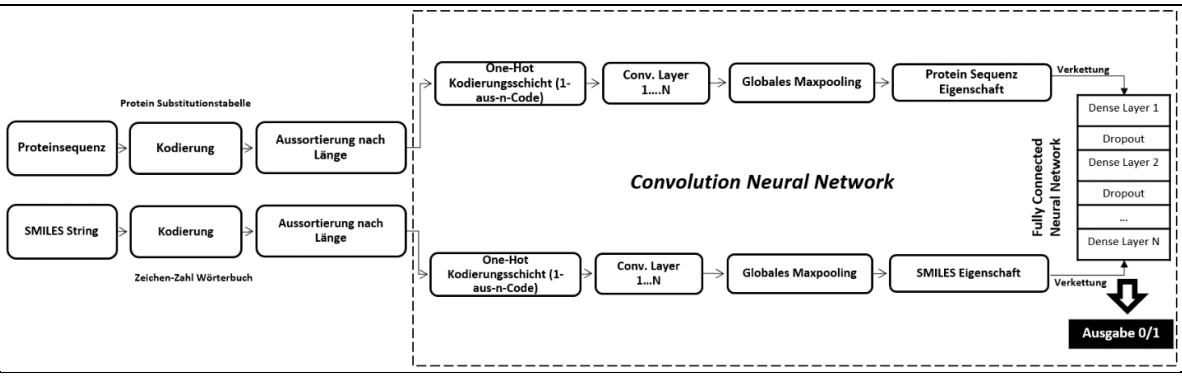

Abbildung 4. Darstellung des End-to-end Ansatzes für DTI (Quelle: In Anlehnung an Monteiro et al., 2020)

Für das CNN werden zwei parallele Serien von 1D-Faltungsschichten verwendet, eine für die Proteinsequenzen und eine weitere für die SMILES-Strings, um tiefe Muster (Repräsentationen oder lokale Abhängigkeiten) aufzudecken (Monteiro et al. 2020). Daraufhin wird nach jeder Reihe von Convolutional Layer ein Maxpooling Layer angewendet, um die räumliche Größe jeder Merkmalskarte auf ihr maximal repräsentatives Merkmal zu reduzieren. Die erhaltenden Strukturen werden zu einzelnen Merkmalsvektoren verkettet, welche nun DTI Paare darstellen. Die resultierenden Vektoren sind nun der Input für die FCNN Architektur. Um die Überanpassung zu reduzieren, wird zwischen jeder vollständig verbundenen Schicht ein Dropout verwendet. Dies ist eine Regulierungsmethode, bei der eine zuvor spezifisch festgelegte prozentuale Anzahl an Neuronen ausgeschaltet wird, welche während des Trainings Korrelationen bilden. Nach dieser Architektur folgt eine Ausgabeschicht, dass aus einem Neuron besteht und die Art der Interaktion (0/1), da es sich um ein binäres Klassifikationsproblem handelt, zurückgibt.

Diskussion

Die aktuelle klinische Pipeline für antibakterielle Wirkstoffen ist unzureichend, um das Resistenzproblem zu lösen, da vor allem die geringe Spezifität der Screening-Methoden und der Mangel an geeigneten Verbindungen in den chemischen Bibliotheken der Pharmaunternehmen große Hürden für die Medikamentenentwicklung darstellen (WHO 2019). ML bietet eine Möglichkeit die Screening-Methoden zu verbessern, da im Vergleich zu klassischem Screening mehr Daten in kürzerer Zeit verarbeitet werden können.

QSAR Methoden zeigen, dass ML in Verbindung mit molekularen Deskriptoren antibiotische Eigenschaften von Verbindungen erfolgreich vorhersagen kann (Pereira et al. 2015). Diese Modelle kommen in der Entwicklung von Medikamenten schon regelmäßig zum Einsatz, allerdings sind die klassischen QSAR Methoden nicht für den Einsatz mit den zunehmend großen Datenmengen geeignet (Zhang et al. 2017). Dies und auch die Beschränkung auf die Erfassung der Eigenschaften mittels Deskriptoren limitieren QSAR Modelle, weshalb die Erweiterung des Konzepts verfolgt wird.

Der Einsatz von Deep Learning in Form von CNN kann die Charakterisierung der Moleküle mittels Deskriptoren flexibilisieren (Ching et al. 2018; Yang et al. 2019b). Dadurch können im Vergleich zu klassischen Methoden wie z.B. SVM bessere Ergebnisse erzielt werden (Monteiro et al. 2020; Stokes et al. 2020). Wie die beste Performance der CNN Modelle erreicht werden kann, indem z.B. der generierte Input mit weiteren Informationen ergänzt wird, ist noch herauszufinden, denn die hier vorgestellten Modelle

treffen gegensätzliche Aussagen. Dies verdeutlicht, dass auch das Ergebnis von Deep Learning Modellen stark vom Studiendesign abhängt und damit als eine Art des ML den gleichen Restriktionen unterliegt wie andere ML Methoden (Ching et al. 2018).

Deep Learning Modelle dieser Art können allerdings nur Aussagen über Korrelation und nicht Kausalität treffen (Ching et al. 2018) Obwohl Deep Learning Screening Methoden also gute Vorhersagekraft besitzen, können aufgrund ihrer Black Box Eigenschaft keine Aussagen über biologische Vorgänge getroffen werden (Zhang et al. 2017). Allerdings kann auch in diesem Forschungsfeld Deep Learning erfolgreich integriert werden, sodass Wirkungsweisen und Vorgänge besser verstanden werden können, was wiederum die Identifizierung neuer Wirkungsweisen ermöglicht und somit die Antibiotikaentwicklung bereichern kann (Camacho et al. 2018; Yang et al. 2019a).

Auch das Problem der Limitierung der chemischen Datenbanken lässt sich mithilfe von ML in Angriff nehmen. Sowohl prototypenbasierten Konzepte, welche neue Zusammensetzungen im chemischen Raum der Datenbanken generieren (Harel & Radinsky 2018), als auch das Screening von Molekülfragmenten (Mansbach et al. 2020), was den chemischen Raum sogar erweitert, bieten Möglichkeiten von den Verbindungen.

Aufgrund der zeitlichen Limitation dieser Arbeit, wurden Veröffentlichungen bezüglich des Einsatzes von ML in der Entdeckung von antimikrobiellen Peptiden (AMP) nicht aufgenommen. AMP bieten eine Alternative zu Antibiotika in der Bekämpfung bakterieller Infektionen, können diese aber nicht gänzlich ersetzten (WHO 2019). Dennoch gibt es Parallelen beim Einsatz von ML, denn unter anderem findet auch virtuelles Screening Anwendung (Lee et al. 2018). Es ist zu ergründen, ob und wie sich die hier angewendeten Konzepte auf die Antibiotikaforschung übertragen lassen.

Weiterhin lässt sich der tatsächliche Erfolg der vorgestellten, neuartigen Modelle in Form von wirksamen Medikamenten erst in einigen Jahren einschätzen, denn die erfolgreiche Generierung von Kandidaten für präklinische und später klinische Studien ist ein wichtiger Schritt, verspricht aber keinesfalls einen erfolgreichen Durchlauf des Medikamentenzulassungsprozesses. Es lässt sich aber dennoch annehmen, dass mithilfe von weiterer Forschung auf Basis der vorgestellten ML Konzepte insbesondere Deep Learning die Erweiterung des chemischen Raums und Verbesserung von Screening-Methoden zum Zwecke der Antibiotikaentdeckung in Angriff genommen werden können. Eine der größten Herausforderungen scheint eine gute Auswahl der der Modelle zugrundeliegenden biologischen und chemischen Daten zu sein.

Fazit

In der vorliegenden Arbeit wurden die unterschiedlichen Ansätze des ML und der Antibiotika Entdeckung sowie der daraus resultierenden Effizienz vorgestellt und erläutert. Dabei wurde eine systematische Literaturrecherche zu Grunde gelegt, bei welchem die die ML Ansätze identifiziert und damit verbunden den Einsatz in die Antibiotika Entdeckung klarstellt. Neben der begrifflich systematischen Grundlegung wurden die relevanten ML Ansätze erläutert. Es wurden zwei Konzepte ausgewählt, die bei der Entdeckung neuer Antibiotika von großer Bedeutung sind. Vor diesem Hintergrund wurden die Bereiche in Virtuelles Screening und End-to-End Konzept gegliedert. Zu jedem Bereich wurde eine Beschreibung angegeben und die Konzepte entsprechend vorgestellt, welche für die Entdeckung der Arzneimittel essenziell.

Referenzen

Baldi, P. 2012. “Autoencoders, Unsupervised Learning, and Deep Architectures,” in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, JMLR, pp. 37–49.

Batool, M., Ahmad, B., and Choi, S. 2019. “A Structure-Based Drug Discovery Paradigm,” *International Journal of Molecular Sciences* (20:11), p. 18.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. 2018. “Next-Generation Machine Learning for Biological Networks,” *Cell* (173:7), pp. 1581–1592.

Cherkasov, A. 2005. “Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks,” *International Journal of Molecular Sciences* (6:1), pp. 63–86.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C.,

Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H. S., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., and Greene, C. S. 2018. "Opportunities and Obstacles for Deep Learning in Biology and Medicine," *Journal of The Royal Society Interface* (15:141), p. 47.

Cho, G., Jung, K., and Hwang, H. 2019. "Out-of-Bag Prediction Error: A Cross Validation Index for Generalized Structured Component Analysis," *Multivariate Behavioral Research* (54:4), pp. 505–513.

Danishuddin, and Khan, A. U. 2016. "Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design," *Drug Discovery Today* (21:8), pp. 1291–1302.

Fritsche, O. 2016. *Mikrobiologie*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Gavankar, S. S., and Sawarkar, S. D. 2017. *Eager Dicision Tree*, presented at the 2017 2nd International Conference for Convergence in Technology (I2CT), IEEE, April, pp. 837–840.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, Adaptive Computation and Machine Learning Series, Cambridge, MA: MIT Press.

Hall, M. A. 1999. "Correlation-Based Feature Selection for Machine Learning," Hamilton, New Zealand: University of Waikato.

Hancock, R. E. W. 1998. "Resistance Mechanisms in Pseudomonas Aeruginosa and Other Nonfermentative Gram-Negative Bacteria," *Clinical Infectious Diseases* (27:1), pp. 93–99.

Hanzlik, R. P., Koen, Y. M., Theertham, B., Dong, Y., and Fang, J. 2007. "The Reactive Metabolite Target Protein Database (TPDB) – a Web-Accessible Resource," *BMC Bioinformatics* (8:1), p. 95.

Harel, S., and Radinsky, K. 2018. "Accelerating Prototype-Based Drug Discovery Using Conditional Diversity Networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom: ACM, July 19, pp. 331–339.

Lan, H., and Pan, Y. 2019. "A Crowdsourcing Quality Prediction Model Based on Random Forests," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Beijing, China: IEEE, June, pp. 315–319.

Landrum, G. 2006. RDKit: Open-Source Cheminformatics.

Lee, E. Y., Wong, G. C. L., and Ferguson, A. L. 2018. "Machine Learning-Enabled Discovery and Design of Membrane-Active Peptides," *Bioorganic & Medicinal Chemistry* (26:10), pp. 2708–2718.

Lipinski, C. A. 2000. "Drug-like Properties and the Causes of Poor Solubility and Poor Permeability," *Journal of Pharmacological and Toxicological Methods* (44:1), pp. 235–249.

Mansbach, R. A., Leus, I. V., Mehla, J., Lopez, C. A., Walker, J. K., Rybenkov, V. V., Hengartner, N., Zgurskaya, H. I., and Gnanakaran, S. 2020. "Machine Learning Algorithm Identifies an Antibiotic Vocabulary for Permeating Gram-Negative Bacteria," *Journal of Chemical Information and Modeling*, pp. 2838–2847.

Monteiro, N. R. C., Ribeiro, B., and Arrais, J. 2020. "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 12.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. 2008. "The MARTINI Coarse-Grained Force Field: Extension to Proteins," *Journal of Chemical Theory and Computation* (4:5), pp. 819–834.

Murray, C. W., and Rees, D. C. 2009. "The Rise of Fragment-Based Drug Discovery," *Nature Chemistry* (1:3), pp. 187–192.

O'Neill, J. 2014. "Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations," Review on Antimicrobial Resistance: Tackling Drug-Resistant Infections Globally, London United Kingdom: Wellcome Trust, December.

O'Neill, J. 2016. "Tackling Drug-Resistant Infections Globally: Final Report and Recommendations," Review on Antimicrobial Resistance: Tackling Drug-Resistant Infections Globally, London United Kingdom: Wellcome Trust, May.

Patel, H. M., Noolvi, M. N., Sharma, P., Jaiswal, V., Bansal, S., Lohan, S., Kumar, S. S., Abbot, V., Dhiman, S., and Bhardwaj, V. 2014. "Quantitative Structure–Activity Relationship (QSAR) Studies as Strategic Approach in Drug Discovery," *Medicinal Chemistry Research* (23:12), pp. 4991–5007.

Pereira, F., Latino, D., and Gaudêncio, S. 2015. "QSAR-Assisted Virtual Screening of Lead-Like Molecules from Marine and Microbial Natural Sources for Antitumor and Antibiotic Drug Discovery," *Molecules* (20:3), pp. 4848–4873.

Polishchuk, P. G., Madzhidov, T. I., and Varnek, A. 2013. "Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data," *Journal of Computer-Aided Molecular Design* (27:8), pp. 675–679.

Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A. K. M., Wertheim, H. F. L., Sumpradit, N., Vlieghe, E., Hara, G. L., Gould, I. M., Goossens, H., Greko, C., So, A. D., Bigdeli, M., Tomson, G., Woodhouse, W., Ombaka, E., Peralta, A. Q., Qamar, F. N., Mir, F., Kariuki, S., Bhutta, Z. A., Coates, A., Bergstrom, R., Wright, G. D., Brown, E. D., and Cars, O. 2013. "Antibiotic Resistance—the Need for Global Solutions," *The Lancet Infectious Diseases* (13:12), pp. 1057–1098.

Sherstinsky, A. 2020. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *Physica D: Nonlinear Phenomena* (404), p. 43.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackerman, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. 2020. "A Deep Learning Approach to Antibiotic Discovery," *Cell* (180:4), pp. 688-702.e13.

Tam, B., Sherf, D., Cohen, S., Eisdorfer, S. A., Perez, M., Soffer, A., Vilenchik, D., Akabayov, S. R., Wagner, G., and Akabayov, B. 2019. "Discovery of Small-Molecule Inhibitors Targeting the Ribosomal Peptidyl Transferase Center (PTC) of M. Tuberculosis," *Chemical Science* (10:38), pp. 8764–8767.

Tibaduiza, D. A., Mujica, L. E., Rodellar, J., and Güemes, A. 2016. "Structural Damage Detection Using Principal Component Analysis and Damage Indices," *Journal of Intelligent Material Systems and Structures* (27:2), pp. 233–248.

Vogelmeier, C. 2018. "Antibiotika: die wichtigste medizinische Entdeckung des 20. Jahrhunderts," *DMW - Deutsche Medizinische Wochenschrift* (143:15), pp. 1057–1057.

Wang, H., Li, G., and Tsai, C.-L. 2007. "Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (69:1), pp. 63–78.

Webber, M. A., Piddock, L. J. V. 2003. "The Importance of Efflux Pumps in Bacterial Antibiotic Resistance," *Journal of Antimicrobial Chemotherapy* (51:1), pp. 9–11.

Weininger, D. 1988. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Modeling* (28:1), pp. 31–36.

Witte, W., Klare, I., and Robert-Koch-Institut 1999. "Antibiotikaresistenz bei bakteriellen Infektionserregern," *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* (42), pp. 8–16.

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C., and Collins, J. J. 2019a. "A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action," *Cell* (177:6), pp. 1649-1661.e9.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. 2019b. "Analyzing Learned Molecular Representations for Property Prediction," *Journal of Chemical Information and Modeling* (59:8), pp. 3370–3388.

Yu, C.-Y., Chou, L.-C., and Chang, D. T.-H. 2010. "Predicting Protein-Protein Interactions in Unbalanced Data Using the Primary Structure of Proteins," *BMC Bioinformatics* (11), p. 10.

Zhang, L., Tan, J., Han, D., and Zhu, H. 2017. "From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery," *Drug Discovery Today* (22:11), pp. 1680–1685.

Zoffmann, S., Vercruysse, M., Benmansour, F., Maunz, A., Wolf, L., Blum Marti, R., Heckel, T., Ding, H., Truong, H. H., Prummer, M., Schmucki, R., Mason, C. S., Bradley, K., Jacob, A. I., Lerner, C., Araujo del Rosario, A., Burcin, M., Amrein, K. E., and Prunotto, M. 2019. "Machine Learning-Powered Antibiotics Phenotypic Drug Discovery," *Scientific Reports* (9:1), p. 5013.

"Antibacterial Agents in Clinical Development: An Analysis of the Antibacterial Clinical Development Pipeline." 2019. Geneva: World Health Organization, p. 48.

Trade-offs Between Privacy-Preserving and Explainable Machine Learning in Healthcare

Emerging Trends in Digital Health, Summer Term 2020

Tobias Budig
Bachelor Student
Karlsruhe Institute of Technology
tobias.budig@student.kit.edu

Alexander Dietz
Bachelor Student
Karlsruhe Institute of Technology
alexander.dietz@student.kit.edu

Selina Herrmann
Bachelor Student
Karlsruhe Institute of Technology
uppch@student.kit.edu

Abstract

Background: *M*achine Learning has enormous potential for applications in various fields. Explainability and privacy are two key questions when training a Machine Learning model especially in critical information infrastructure such as the healthcare sector.

Objective: The goal of this paper is to identify the current state of research and possible trade-offs between explainability and privacy of Machine Learning models. Furthermore, the aim was to identify possible ways of implementing explainability methods in Federated Learning, a privacy-preserving setting.

Methods: First, we have conducted a systematic literature review to identify possible trade-offs. Second, we evaluated and selected methods that one can theoretically implement without risking privacy in a Federated Learning application with a focus on medical image analysis.

Results: Our results show that only a few researchers have so far been discussing possible trade-offs between explainable and privacy-preserving Machine Learning. The three relevant papers show that there is a natural trade-off, and a higher level of explainability can make a model more vulnerable to attacks and therefore have a higher risk of privacy leakage. For our implementation example of explainable Machine Learning methods in Federated Learning we came to the result that it seems to be theoretically possible. Our selected methods are SHapley Additive exPlanations, Gradient-weighted Class Activation Mapping, and Local Interpretable Model-Agnostic Explanations. However, experiments would be necessary to confirm these ideas.

Conclusion: To summarize, we showed that possible trade-offs between explainable and privacy-preserving Machine Learning methods exist, though, is not yet fully discussed in the literature.

Keywords: machine learning, privacy-preserving, XAI, federated learning, trade-offs

Introduction

Motivation

Machine learning (ML) has enormous potential in healthcare and has been used in medicine since the very beginning of the field (Ahmad et al. 2018). However, only in recent years, the importance of ML-based solutions in healthcare has been recognized, and in a few years, it will become indispensable.

For example, ML helps immensely in the detection of chronic diseases. The careful analysis of medical data ensures that maladies are detected earlier, and patients receive better care (Chen et al. 2017). As the University of Pennsylvania was able to co-develop a technology, which can train an artificial intelligence (AI) so that it can identify brain tumors in x-ray images using the privacy-preserving method federated learning (Intel 2020).

The increasing computing capacity, the use of electronic health records in hospitals, and the availability of data cause this area to evolve rapidly (McCradden et al. 2020). Nevertheless, big data also entails ethical concerns about responsibility, trust, and accountability, among others. In the implementation of ML, public views are fundamental. On the one hand, to encourage companies to invest in AI, on the other hand, to support educational initiatives that encourage trust and support among the population. These ethnic controversies and the speed with which this technology is advancing and developing can affect public confidence.

An important issue is to protect the privacy of an individual. Therefore, privacy-preserving machine learning (PPML) tries to ensure that one cannot exploit the used training data. Especially in the healthcare sector as (McCradden et al. 2020) showed in her study about "Ethical concerns around use of artificial intelligence in health care research [...]", this is essential, as sensitive data is involved, hence trust is inevitable. While some people are open to new technologies, many cautious users are skeptical about them and wait to see how they develop (HSBC 2017). Advertisers and designers make every effort that their product inspires confidence and spend vast amounts of money to ensure this. The survey by HSBC also shows the distrust towards technology in medicine. Only 14% state that they would trust a humanoid robot programmed by leading surgeons to perform open-heart surgery on them, compared to already 9% who would trust a family member led by a surgeon.

An explanation is essential to build trust. However, there is no mathematical definition of explainability (Molnar 2019). A non-mathematical definition is that explainability "is the degree to which a human can understand the cause of a decision." It is easier for one to understand why predictions or decisions have been made, the better the explanation of the ML model is. However, there is the problem that many ML methods are hard to explain, as they work like a black box (Ahmad et al. 2018). Explainable machine learning (EXML) allows the user to understand, question, and even improve the ML system.

Furthermore, the explainability and privacy of ML models is also part of ongoing political discussions, as we have seen with regulations such as European General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA). These regulations include statements about how data can be used for automated decision and some even argue for a "right for explanation" (Goodman and Flaxman 2017) in the GDPR. Hence, the topic is not only part of current research but also essential for practical implementations. For the given practice problem, it is of interest if there are trade-offs between EXML and PPML and how these are relevant to the field of healthcare.

Objectives

Privacy and explainability are crucial when using ML in healthcare. There is currently only limited literature discussing both topics and especially their trade-offs. Therefore, our main objective with this work is to examine possible trade-offs between EXML and PPML.

An overview would benefit future research on these topics. To answer this question, we want to discuss the findings of the relevant literature and present the effects of EXML on privacy exploitation risk, how PPML can hamper this risk and what are overall outcomes of PPML and EXML on a model's performance. Thus, giving an overview of the latest research on these topics. As healthcare is one of the fields, we want to assess the relevance of EXML and PPML for healthcare.

To do so, we examine if different EXML methods can be implemented in a PPML setting. In this context, we have chosen the example of federated learning for Image Analysis as the implementation of EXML would be highly beneficial in this application.

Background

Privacy-Preserving Machine Learning

As mentioned in our introduction, one of the key challenges of ML in healthcare is keeping the patient's data private and secure. For medical research purposes, it is highly beneficial to share data and collect larger data sets to train better ML models. Therefore, helping medical personnel make more precise and correct decisions. However, medical records are some of the most sensitive data, and keeping this data private is inevitable. Due to this need, researchers have developed privacy-preserving methods.

We will introduce the most crucial ones that are or will be relevant to healthcare applications, such as differential privacy (DP), trusted execution environment (TEE), cryptographic approaches, and federated learning.

Differential Privacy

DP is a privacy-preserving method that is based on a probabilistic model (Dwork and Roth 2014). Given the situation of a database where each row represents the information about an individual, the goal is to ensure that a single change to the database does not affect the results by much. Therefore, the input of an individual is privacy secured. Adding random noise, mostly LaPlace or Gaussian distributed, to the data accomplishes this.

Definition 1. (Differential Privacy) A randomized mechanism M: $D \rightarrow R$ with domain D and range R satisfies (ε,δ)-differential privacy if for any two adjacent inputs $x,y \in D$ and for any subset of outputs of $S \subseteq R$ it holds that

$$\Pr[M(x) \in S] \leq \exp(\varepsilon) \Pr[M(y) \in S] + \delta,$$

where δ represents the privacy budget (Dwork and Roth 2014).

In general, a smaller (ε,δ)-value provides higher privacy. This definition provides multiple advantages such as "composability, group privacy, and robustness to auxiliary information" (Abadi et al. 2016).

Trusted Execution Environment

A TEE is defined as "a secure, integrity-protected processing environment, consisting of processing, memory, and storage capabilities" (Asokan et al. 2014). TEEs are already in use in many of our mobile devices as a way to secure two-factor authentication or IoT applications (Ekberg et al. 2013). In the context of ML TEEs can provide a solution for protecting the privacy of training data as the training of a model can be run within the TEE. A user sends encrypted data to the enclave, which then decrypts the data and trains the model (Narra et al. 2019).

This approach is also interesting for healthcare applications as a clinic could send sensitive encrypted data to a research site to support the training of a ML model, for example, on image classifications of brain tumors. By doing so, a large number of clinics can train the model as long as the clinic trusts the TEE. Current work also proposes the use of TEE in a federated learning setting (Mo and Haddadi 2019). This approach should hamper the information leakage as the model is trained within a TEE on the client-side, and one only encrypted data is sent to the central server.

Cryptographic Approaches

Two of the most relevant cryptographic approaches are secure multiparty computation (SMC) and homomorphic encryption (HE). SMC is a method that "enables computation on sensitive data from multiple sources while maintaining privacy" (Chen et al. 2019). It guarantees that only each computation reveals the outcome to other users. The theoretical approach was introduced in 1982 by A. C. Yao (Yao 1982), but due

to the high need for computation power, it took until 2004 for a first general notable implementation (Malkhi et al. 2004). One of the most used methods is Shamir's secret sharing (Shamir 1979).

HE requires algorithms to allow certain operations to be carried out on ciphertexts (Aslett et al. 2015). By doing so, there is no difference between operating on the ciphertext or the original message. The user encrypts the data with a public key, and only certain individuals with a private secret key can decrypt the data at any point in time.

In healthcare, HE can be used to share data between research institutes or store sensitive data securely in a cloud (Raisaro et al. 2018). Furthermore, there are also first implementations of HE in other PPML methods, like federated learning, to ensure a higher level of privacy (Xu et al. 2019). Current restraints in HE includes the high computational cost (Aslett et al. 2015), the inability to perform division operations, and the large data size of the ciphertext.

Federated Learning

By definition, federated learning is the collective training of one model by multiple clients orchestrated by one central server. It provides the advantage of decentralized training data and a high level of privacy by design. Especially in the context of the GDPR federated learning is currently a promising technique to ensure PPML.

Since its introduction in 2016 (Konečný et al. 2016), a wide range of applications use federated learning. One of the most prominent is the use in the Gboard mobile keyboard by Google. Each user downloads the current model, trains it with their local data, and then sends the updated model back to the server, which averages all inputs to compute the latest model. By doing so, the user can get recommendations for any input on their smartphone based on the data of millions of other users without ever revealing their private data to any other user or server (Yang et al. 2018).

Federated learning currently faces also challenges in the implementation and security (Kairouz et al. 2019), which are also relevant to the field of healthcare. First, the central server can be an exploitable weakness as it has to handle all the clients' inputs. Furthermore, it requires a certain level of trust by the clients as some centralized authority has to decide on what to train and how to train the model. There is a risk of information leakage from the server to the client or at the server. An adversary could use this information to reconstruct a client's data based on the model and the gradient update.

Therefore, it is advisable to implement further privacy-preserving techniques such as DP, TEE, or cryptographic approaches in the design of a federated learning system to hamper such attacks. So far, it seems that DP and cryptographic approaches are the only efficient techniques (Huang et al. 2020). Due to the promising results concerning privacy, researchers also try to implement applications in various fields of healthcare, such as dentistry (Schwendicke et al. 2020), wearables (Chen et al. 2020), and image analysis (Li et al. 2019).

Explainable Machine Learning

In this section, we want to give a brief overview of EXML techniques with a focus on the explanation of images predictions.

EXML refers to an ML system that "produces details or reasons to make its functioning clear and easy to understand" (Arrieta et al. 2020). To reach this, different approaches are developed depending on the class of ML model. Here we differentiate between transparent and complex ML models, which consider the inherent complexity. For complex models, we need to generate an explanation after training. These so-called "post-hoc" techniques can be classified into two dimensions.

First, the scope of the explanation, global or local, describes whether the aim of an explanation is only for one prediction or the overall behavior. Second, explainable algorithms can be, on the one hand, model-specific if they only work with some kind of model architecture. On the other hand, model-agnostic algorithms are independent of the model properties and can, therefore, be applied to all kinds of ML models (Molnar 2019).

Transparent and Complex Machine Learning Models

Simple ML models, like linear regression or decision trees, can be interpreted without further techniques. Here, humans can interpret the weights of the regression to understand the model's decisions. Their advantage is providing out-of-the-box explainability, but shallow models are not as powerful as complex ones in nonlinear domains.

On the other hand, we need to use post-hoc methods for complex ML models like neural networks to achieve explainability. In this case, we cannot interpret the neural network's weights directly due to the huge number of parameters and the inner complexity (Molnar 2019). Therefore, there is a trade-off between the intrinsic model explainability and the model accuracy (Arrieta et al. 2020).

Local and Global Explanations of Machine Learning Models

ML explanations can have a different scope. On the one hand, local explainability describe which input features have a big impact on one specific predicted result. Common methods are Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). They are useful to justify an end-user (Doshi-Velez and Kim 2017). On the other hand, global explanations describe the behavior of the whole model. Methods, such as ProtoDash (Gurumoorthy et al. 2017), try to investigate and describe the patterns learned by the model. Incorrect learnings (e.g. biases) can be discovered here (Doshi-Velez and Kim 2017). However, local explanations can also help to identify biases in ML predictions (Selvaraju et al. 2017).

Model-Agnostic and Model-Specific Explainability

Model-agnostic post-hoc algorithms only need the input object and the result of an ML model to provide explainability for this model. Therefore, they do not depend on specific model structure or architecture. There are two main classes of model-agnostic post-hoc methods.

First, explanation by simplification, where we try to approximate parts of the complex ML model with simpler ones. Here, LIME is the most known approach. It builds local linear models trained by the original model input-output pairs to explain individual predictions. One can apply it to tabular, image, or text data. LIME for image explanations will be discussed later in section 5.3.2.

The second class, feature relevance explanations, tries to rank or measure the impact of features on the predicted result. SHAP is currently the most popular approach in this field. It uses each possible combination of input features to get a marginal contribution to the result (Lundberg and Lee 2017). SHAP can be used for every neural network like a convolutional neural network (CNN).

In contrast, model-specific techniques can use internal parameters like architecture or optimization algorithm. They can benefit from direct access to the weights of a model. For Instance, Grad-CAM, the most common model-specific technique for CNNs, utilizes the gradient of the last convolutional layer (Selvaraju et al. 2017).

Methodology

Data Collection

Our approach for this paper can be separated into two different methods. For section 2 and section 5.2 we used a non-systematic literature review. The trade-offs between PPML and EXML were analyzed by a systematic literature review. To do so, we focused our search on the following scientific databases: IEEE Xplore, ArXiv, ACM Digital Library, AIS Electronic Library, Science Direct, and Scopus.

We believe that these databases are the most relevant for the discussed topics and cover a wide range of journals and conference publications. We used the following search string: TIKEAB((explain* OR interpret*) AND "machine learning" AND (privacy OR federated OR trust*)).

This search string requires the publication to have at least one of the search terms in the fields of explainability and privacy, respectively. Furthermore, it requires the papers to discuss ML. For the database

ArXiv, we included only papers published since 2018 as papers on ArXiv are not peer-reviewed. We conducted the search on June 3rd, 2020 and had 526 results shown in Table 1 grouped by their database.

Database	Number of Results
IEEE Xplore	100
ArXiv	150
ACM Digital Library	189
AIS Electronic Library	5
Science Direct	21
Scopus	61

Table 1. Results of Literature Review

Data Analysis

As the following step, we analyzed our results and looked for papers that discuss PPML and EXML. The majority of the found papers only discussed one of the topics and were, therefore, not applicable for a literature review on the trade-offs. We were able to identify three papers that discussed trade-offs. Section 4.1 examines these papers. It seems like the topic is part of ongoing research and can be a topic for further papers. Shokri et al. (Shokri et al. 2019) and Harder et al. (Harder et al. 2020) have stated to be the first ones to discuss possible trade-offs. Besides, these papers were recently published Backward search from our results has shown that ML models are vulnerable to Membership Inference Attacks. In other words, the attacker can exploit these black-box models only by accessing input and output queries to reconstruct members of the training set of the model (Shokri et al. 2017; Song et al. 2019; Truex et al. 2019). Even worse, (Oh et al. 2019) showed, that it is possible to reverse-engineer some parts of Neural Networks like architecture or hyperparameters.

These results and the mentioned points in the papers raise three questions

1. Does Explainable Machine Learning increase privacy exploit probability?
2. Which Privacy-Preserving Machine Learning techniques save Machine Learning models from exploitation?
3. What impact do Explainable and Privacy-Preserving Machine Learning have on the model's accuracy?

which will be discussed in the Results section.

Trade-Offs

After conducting the systematic literature review as described in 3.2, we got three relevant papers. Two of it – (Harder et al. 2020) and (Shokri et al. 2019) - discuss the privacy explainable trade-off directly. Furthermore, (Arrieta et al. 2020) focus on EXML in general but also investigate the privacy-explainable question from data-fusion perspective. To start, we first present the relevant key findings of the three relevant papers for further discussion.

Relevant Papers

Privacy Risks of Explaining Machine Learning Models

Shokri et al. (2019) discuss the possibilities for an adversary to use a model explanation to infer sensitive data of the training's set. In their work, they focus on Membership Inference Attacks and Reconstruction Attacks. The authors conducted experiments using gradient-based attribution methods or record-based influence measures. The used data sets included two sets with binary features and up to circa 200,000 records. Another two sets with mixed Features and up to circa 100,000 records. Furthermore, the authors used CIFAR-100, a benchmark data set for image classification. They used "fully connected multi-layer networks with tanh activations" for training the datasets with binary and mixed features and a convolutional neural network for the CIFAR-100 image dataset.

When running a Membership Inference Attack, the adversary tries to determine the training set's data point. On the other hand, a Reconstruction Attack tries to get as many data of the training set as possible, basically reconstructing the training set.

They show that an adversary can exploit record and feature-based explanations. Furthermore, she can get the training set membership information of a data point. The only way to reduce information leakage is by adding noise to the data, using DP. In addition, they were able to conduct Reconstruction Attacks to extract parts of a training set given a record-based explanation. In this context, they also state that minorities in the training data have a high risk of being revealed by this attack.

The authors believe to be the first to discuss the trade-offs between explainability and privacy of a ML model.

Interpretable and Differential Private Predictions

Harder et al. (2020) main question is if it is possible to have an explainable model without data lost cause of privacy protection. They claim to be the first ones to enquire about this question. Their approach is using locally linear maps (LLM) on a family of simple models, which should ensure privacy and do not expose the whole model, only the LLM.

LLM act as an approximation of differentiable functions, i.e., a collection of piecewise linear functions. If sufficient linear maps are available, local models compared to complex model counterparts have a relatively low loss inaccuracy. The more complex the data, the higher the loss of prediction accuracy. For using LLM, explainability and privacy are necessary.

The authors explain two ways to gain explainability. One relies on inherently explainable models. The other is post-processing schemes, for example, using gradient-based attributions. To provide privacy, they use DP, which lead them to the conflict that adding a large among noise ensures a high level of privacy but is disadvantageous for predicting accuracy. They also state that high dimensional parameters imply high security lost. That is why they propose using a relatively small network, which can be partly trained and guarantees privacy using LLM. Their contributions are proposing a novel family of explainable models, providing explanations for 'local' and 'global' DP on classification and suggesting using random projections to deal better with privacy and accuracy trade-offs.

After explaining their method using LLM, they eventuate the trade-off between privacy, explainability, and accuracy with experiments. They realize that it is tough to assure all three assertions. For example, reducing the level of privacy and removing random projections implies better explainability results. Meanwhile, private training benefits form increasing the dimensionality of random projections. Also, raising the number of LLM, the accuracy of private models decreases because the privacy budget is distributed over more parameters.

In conclusion, it is possible to use LLM to provide explainability and privacy. However, the data set, which they use, is simple and relatively small, so there is still the question: what is the limit of complexity for the LLM models? They also state that "several open questions for future research remain".

Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI

The literature review "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI" by (Arrieta et al. 2020) gives an overview of the current EXML research. Moreover, the authors provide a global taxonomy of EXML by classify current explainable techniques. In the beginning, they state that responsible AI needs explainable AI as well as privacy-preserving AI. Therefore, the authors investigated not only current EXML techniques but also their privacy impact.

Despite the explainability of ML models enable third parties to investigate the model, many papers about EXML do not cover the privacy topic.

A general privacy concern of ML models are attacks like membership inference attacks or approaches to reverse engineer the model's parameter, even with black-box models where third persons only have access to the input features and output predictions. EXML can, therefore, be used to increase the attacker's success probability.

Furthermore, the authors explore the field of responsible ML for data fusion. They state that EXML can compromise privacy for data fusion on different levels (data, model, Knowledge) (Arrieta et al. 2020).

As a result, the authors conclude that further research is needed to ensure explainability as well as privacy for ML in general and data fusion.

Results

In the next subsections, we describe the contributions of the reviewed literature to the questions motivated in the Data Analysis section.

Does Explainable Machine Learning Increase Privacy Exploit Probability?

In general, ML models are vulnerable to privacy leaks through membership inference attacks (Shokri et al. 2017) or reverse-engineering attacks (Oh et al. 2019; Shokri et al. 2019) showed that EXML techniques can increase the privacy exploit probability because some EXML techniques like "gradient-based methods can leak a significant amount of information, much beyond what is leaked through the predicted labels." Moreover, they showed that record-based influence measures - a technique that explains the result by outputting the most critical point from the training set - is even worse. An adversary could reconstruct nearly 99% (Shokri et al. 2019) of the used example. Experiments by (Shokri et al. 2019) show, that especially minorities are vulnerable to these attacks. Here, the authors demonstrated for a diabetic hospital dataset that data rows of minorities like children or African-American people could be regenerated. Outliers are more likely to memorize by the model during training is the explanation for this behavior by (Shokri et al. 2019).

In contrast, new, designed explainability techniques can be privacy-preserving. Harder et al. (2020) describes a novel approach with LLM to guarantee privacy by making the gradient differential private. Nevertheless, in this paper, they have specific constraints like a relatively simple dataset and the lack of interaction with a complex counterpart (Harder et al. 2020). Therefore, we cannot interpolate these results for other cases.

Furthermore, Arrieta et al. (2020) states that common EXML techniques like LIME or SHAP are not investigated towards privacy concerns yet.

Which Privacy-Preserving Machine Learning Techniques Save Machine Learning Models from Exploitation?

To tackle the privacy question raised by model explanations, both - Shokri et al. (2019) and Harder et al. (2020) - propose to use noise towards the gradients to guarantee data protection. There are two techniques offered. First, differential private training should be immune to gradient-based attacks due to "gradient-based explanations only interact with the model, and not with the underlying training set." (Shokri et al. 2019). Especially smoothed gradients are resistant to the attacks performed by Shokri et al. (2019). The

average of the gradients in the surrounding area to the original point, together with adding Gaussian noise, seem to be a privacy-preserving approach, as the authors state.

Second, Harder et al. (2020) developed an explainable method to provide privacy-preserving local and global explanations. The aim is to provide a set of LLMs to approximate the neural network's predictions. The authors realize this by adding weighted linear functions where the differential private stochastic gradient descent computes the linear coefficients. That means, adding noise to the gradient to make it differential private (Harder et al. 2020).

In addition, the authors used random projection (Johnson-Lindenstrauss transformation) to reduce the dimensions of the privatized parameters to increase accuracy (Harder et al. 2020).

What Impact do Explainable and Privacy-Preserving Machine Learning Have on the Model's Accuracy?

Harder et al. (2020) states that there is a natural trade-off between privacy and accuracy. He also describes a triangle trade-off between privacy, explainability, and accuracy of an ML model. All three goals cannot be achieved simultaneously. If one focuses on two of the three trades, the third one will be decreasing.

Also, DP comes with the cost of accuracy lost, which one could tackle by using a bigger dataset as Shokri et al. (2019) states. Arrieta et al. (2020) summarizes that there is a "need for further research toward the development of XAI Tools capable of explaining ML models while keeping the model's confidentiality in mind."

Implementation of Explainable Machine Learning in Federated Learning for Image Analysis

In the following section, we like to present a highly relevant example of the use of PPML in healthcare and how one can implement EXML methods in this scenario.

Convolutional Neural Networks for Image Analysis

CNNs are a special class of neural networks that use convolutions in place of general matrix multiplication (Goodfellow et al. 2016). Forms of CNN are widely used in the field of visual computing and have shown great performance in object recognition and image classification (Ciregan et al. 2012; Cireşan et al. 2011; Lawrence et al. 1997). Therefore, CNN is the standard for medical imaging applications.

Federated Learning for Medical Image Analysis

Medical images, taken by techniques like Magnet Resonance Imaging, X-ray, and ultrasound, present one of the greatest opportunities for using AI in healthcare. Clinics have millions of medical information. This data would be a good training set for a ML model (Dash et al. 2019). However, especially in the field of diagnosis, which most image analysis is part of it is crucial to have PPML and EXML models. Images like a brain scan are highly sensitive data, and it is in the patient's interest to have them maximally secured and not open to any other non-authorized third party. If one uses this data to train a ML model, it is even more important to ensure the privacy of the training data as multiple hospitals might use the final model. Therefore, it should not be possible to extract any of the original training data from the model.

For these reasons, we discuss the use of one of the most promising approaches in the field, federal learning for the application of medical image analysis. First implementations showed that it is possible to train a model on brain tumor segmentation data with federated learning that had a 99% accuracy of a model with shared data (Sheller et al. 2019).

Furthermore, as we have seen in our previous sections, it is also crucial in the healthcare sector to offer the explainability of ML models to the end-user like a doctor. A decision based on images like a brain scan can have far-reaching consequences, and therefore it is inevitable to explain to the decision-maker if she is supposed to rely on the results of a ML model. Due to this fact, we present possible ways to incorporate explainability methods in a model that one trained with the federal learning approach.

Explainability Methods for Image Classification

Gradient-Weighted Class Activation Mapping

Selvaraju et al. (2017) introduced Grad-CAM in 2017. The method is based on Class Activation Maps (Zhou et al. 2016). Class Activation Maps highlight discriminative areas of an image. However, it is only applicable to a particular kind of CNN architecture. For this reason, the researchers developed Grad-CAM as a method that did not require any particular kind of CNN architecture. One can apply it to any already trained CNN. It uses gradient information in the last convolutional layer to highlight discriminative areas such as certain objects or classes. Due to the reason that this layer has "the best compromise between high-level semantics and detailed spatial information" (Selvaraju et al. 2017).

Local Interpretable Model-Agnostic Explanations

As a model-agnostic explainable method, LIME can be applied to all ML models and therefore for CNNs, too. This method aims to approximate the decisions of the model at a specific point in a linear way. To achieve this, one generates a set of random data-points. In the case of image data, this means to divide the image into sections called super-pixels. As the second step, the model performs predictions of every sample point. Due to the local aim of the explanation, all samples are weighted in descending order of distance from the original point. In the last step, the model trains a linear classifier by the weighted random sample set and its predicted values. As a result of image data, super-pixels are turned on or off to identify the areas that contribute to the classification (Ribeiro et al. 2016).

SHapley Additive exPlanations

SHAP introduced in section 2.2.2 is a method to explain individual predictions. This method uses optimal Shapley Values from coalitional game theory. Shapely Values indicate how the prediction can be fairly distributed among the features. To do so, one calculates each feature's contribution to an individual prediction (Molnar 2019).

To achieve this, the model generates a super-set of all possible combinations ("coalitions") of input features for every prediction. After that, for every of these 2^F coalitions, where F is the number of features, a model is trained. The difference between the prediction of two elements of the super-set computes the marginal contribution. Finally, weighing the marginal contributions results in a feature-wise explanation. (Lundberg and Lee 2017) describes a technique to approximate the process by reducing the number of coalitions.

DeepSHAP is a model-specific sister algorithm of SHAP and researchers use it for CNNs. It leverages knowledge from internal weights and is optimized for Neural Networks (Lundberg and Lee 2017).

Result

Model-agnostic (e.g. SHAP, LIME) and model-specific (e.g. Grad-CAM) explainability approaches combined with federated learning should be possible because they only need local data even if methods require access to the model's parameters. Hence, one can perform the methods on a local and global level of federated learning. For a Shapley values explainability approach (e.g. SHAP), Wang (2019) proposes a method to balance model explainability and privacy in a federated learning setting. They showed that this approach enables explanations at the local guest system, without getting access to detailed information on guest data. Furthermore, explainability techniques can help detect bias in the training data as it gives a visual explanation of a decision. This is especially useful in healthcare, as there is a high risk of biased data and decisions (Gianfrancesco et al. 2018).

SHAP provides a complete explanation of all features as it is based on a well-founded theory (Molnar 2019). The unified federated features of SHAP give useful information about the contribution of the federated features from the guest party, although the guest data does not need to be released (Wang 2019). Nevertheless, the Shapley Value method has a bad performance. However, SHAP provides approximations that lead to an increase in computational performance and are, therefore, more applicable.

LIME is a model-agnostic explainability method, meaning it is possible to change the underlying model. The short and human-friendly explanations are easy to understand even for a layperson or in situations

with little time and where a full explanation is not required (Molnar 2019). It provides a correct definition of distance in general, however in this case, especially for the distance between super-pixels, Molnar recommends trying to use different kernels settings. A disadvantage of LIME is that it is not stable as Alvarez-Melis and Jaakkola (2018) state, "On the robustness of interpretability methods showed that even close points can result into different explanations."

In federated learning, the number of members' privacy increases because the local model's parameters are merged at the central server. However, there is the problem of unequal distributed data, which could decrease the model's performance and increase the bias (Bonawitz et al. 2019).

Discussion

In the previous sections, we have shown the current state of research on the trade-offs between PPML and EXML, as well as the possible implementation of EXML methods in a federated learning environment for healthcare focusing on the particular example of image analysis. We have seen that both topics are part of ongoing research and highly relevant in healthcare.

Principal Findings

In general, most current ML models are vulnerable to membership inference or reverse engineering attacks, leading to a loss of privacy or intellectual property. These results are independent from EXML methods. A key inside by our literature review is that EXML could increase the exploit probability of ML models. To tackle this issue, new EXML techniques, based on differential private gradients, are developed. The approach seems to be promising due to the properties of DP. Another approach to guarantee privacy is federated learning. It can help protect the privacy of learning data between clients and the central server and between two clients. EXML methods do not require access to the model itself (model-agnostic) or if the only access to the local model of the client (model-specific). Therefore, we expect EXML techniques together with federated learning to protect privacy as Wang (2019) showed for Shapley Values. Furthermore, concerning our example in Section 5.2, we see EXML for image data as one of the most relevant and promising applications in healthcare. An explanation of an image is an intuitive approach that can be understood by all kinds of users.

Implications for Research and Practice

These results lead to the conclusion that users must handle ML models carefully. Even black-box models, which often make up application programming interfaces, can be partially reverse-engineered even without EXML (Dash et al. 2019). Researchers should focus more on EXML methods that are privacy-preserving to provide more real-world value. By doing so, they enable medical institutions to safely use ML methods. Responsible ML needs to be privacy-preserving and explainable.

Limitations and Future Research

Due to the few papers, we have found and the simple assumptions in the research methods, a general statement on possible trade-offs is so far not possible. A next step could be to test the common EXML algorithms as Grad-CAM, SHAP and LIME, for their influence on specific attack scenarios. DP for the gradient or training data seems to be a promising approach that researchers should continue to test. For example, a possible research question could be to focus on EXML for DP and evaluate the privacy loss. In addition, federated learning seems to be another useful approach, since the data remains local, and the risk is, therefore, more in the data transfer than in the ML model itself. A significant challenge here is to carry out a meaningful fusion of the data in the central server because the populations of the nodes can be different and of different quality.

Also, we see advantages concerning the common challenges of federated learning. First, federated learning does not eliminate the risk of biased training data. However, using EXML methods can identify such bias and therefore improve the overall results and accuracy. For example, the authors of Grad-CAM show that they were able to identify gender bias in a training data set with their visual explanation technique (Selvaraju et al. 2017). Moreover, the fairness of the federated setting is highly relevant and also linked to biased training data. Visual explanations would be able to identify the misrepresentation of minority

groups. Hence, it is beneficial to implement EXML methods in federated learning for image classification in the field of healthcare.

Overall, the reviewed literature speaks not with the same voice. One group was able to reverse-engineer a black box model and could, therefore, attack privacy. The other group described a novel approach to prevent such attacks. Now it is not clear how EXML techniques influence privacy attacks. Hence, further research is needed.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. 2016. "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, *CCS '16, New York, NY, USA*: Association for Computing Machinery, pp. 308–318.

Ahmad, M. A., Eckert, C., and Teredesai, A. 2018. "Interpretable Machine Learning in Healthcare," *The IEEE Intelligent Informatics Bulletin*, pp. 1–7.

Alvarez-Melis, D., and Jaakkola, T. S. 2018. "On the Robustness of Interpretability Methods," *ArXiv Preprint ArXiv*:1806.08049.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., and others. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), Elsevier, pp. 82–115.

Aslett, L. J., Esperança, P. M., and Holmes, C. C. 2015. "A Review of Homomorphic Encryption and Software Tools for Encrypted Statistical Machine Learning," *ArXiv Preprint ArXiv:*1508.06574.

Asokan, N., Ekberg, J., Kostiainen, K., Rajan, A., Rozas, C., Sadeghi, A., Schulz, S., and Wachsmann, C. 2014. "Mobile Trusted Computing," *Proceedings of the IEEE* (102:8), pp. 1189–1206.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, H. B., and others. 2019. "Towards Federated Learning at Scale: System Design," *ArXiv Preprint ArXiv*:1902.01046.

Chen, M., Hao, Y., Hwang, K., Wang, L., and Wang, L. 2017. "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access* (5), pp. 8869–8879.

Chen, V., Pastro, V., and Raykova, M. 2019. "Secure Computation for Machine Learning with SPDZ," *ArXiv Preprint ArXiv:*1901.00329.

Chen, Y., Qin, X., Wang, J., Yu, C., and Gao, W. 2020. "FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare," *IEEE Intelligent Systems (*35:4), pp. 83–93.

Ciregan, D., Meier, U., and Schmidhuber, J. 2012. "Multi-Column Deep Neural Networks for Image Classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3642–3649.

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. 2011. "Flexible, High Performance Convolutional Neural Networks for Image Classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11*, AAAI Press, pp. 1237–1242.

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. 2019. "Big Data in Healthcare: Management, Analysis and Future Prospects," *Journal of Big Data* (6:1), p. 54.

Doshi-Velez, F., and Kim, B. 2017. "Towards a Rigorous Science of Interpretable Machine Learning," *ArXiv Preprint ArXiv*:1702.08608.

Dwork, C., and Roth, A. 2014. "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science* (9:3–4), pp. 211–407.

Ekberg, J.-E., Kostiainen, K., and Asokan, N. 2013. "Trusted Execution Environments on Mobile Devices," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13, New York, NY, USA:* Association for Computing Machinery, pp. 1497–1498.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. 2018. "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine (*178:11), American Medical Association, pp. 1544–1547.

Goodman, B., and Flaxman, S. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *AI Magazine (*38:3), Association for the Advancement of Artificial Intelligence (AAAI), pp. 50–57.

Gurumoorthy, K. S., Dhurandhar, A., and Cecchi, G. 2017. "Protodash: Fast Interpretable Prototype Selection," *ArXiv Preprint ArXiv*:1707.01212.

Harder, F., Bauer, M., and Park, M. 2020. "Interpretable and Differentially Private Predictions.," in *AAAI*, pp. 4083–4090.

HSBC. 2017. Trust in Technology, pp. 3–5. (retreived from: https://www.hsbc.com/-/media/hsbc-com/newsroomassets/2017/pdfs/170609-updated-trust-in-technology-final-report.pdf; last accessed: July 31, 2020).

Huang, X., Ding, Y., Jiang, Z. L., Qi, S., Wang, X., and Liao, Q. 2020. "DP-FL: A Novel Differentially Private Federated Learning Framework for the Unbalanced Data," *World Wide Web*.

Intel. 2020. Intel Works with University of Pennsylvania in Using Privacy-Preserving AI to Identify Brain Tumors, Intel. (retrieved from: https://newsroom.intel.com/news/intel-works-university-pennsylvania-using-privacy-preserving-ai-identify-brain-tumors/#gs.9pds0t; last accessed: July 31, 2020).

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., and others. 2019. "Advances and Open Problems in Federated Learning," *ArXiv Preprint ArXiv:*1912.04977.

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. 2016. "Federated Learning: Strategies for Improving Communication Efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*.

Lawrence, S., Giles, C. L., Ah Chung Tsoi, and Back, A. D. 1997. "Face Recognition: A Convolutional Neural-Network Approach," *IEEE Transactions on Neural Networks (*8:1), pp. 98–113.

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., and Feng, A. 2019. Privacy-Preserving Federated Brain Tumour Segmentation.

Lundberg, S. M., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Malkhi, D., Nisan, N., Pinkas, B., and Sella, Y. 2004. "Fairplay - Secure Two-Party Computation System," in *USENIX Security Symposium*.

McCradden, M., Baba, A., and A. Saha A, et al. 2020. "Ethical Concerns around Use of Artificial Intelligence in Health Care Research from the Perspective of Patients with Meningioma, Caregivers and Health Care Providers: A Qualitative Study," *CAMJ Open*, pp. 90–95.

Mo, F., and Haddadi, H. 2019. Efficient and Private Federated Learning Using TEE, p. 1. (retireved from: https://eurosys2019.org/wp-content/uploads/2019/03/eurosys19posters-abstract66.pdf; last accessed: July 31, 2020).

Molnar, C. 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (retrieved from: https://christophm.github.io/interpretable-ml-book/; last accessed: July 31, 2020).

Narra, K. G., Lin, Z., Wang, Y., Balasubramaniam, K., and Annavaram, M. 2019. Privacy-Preserving Inference in Machine Learning Services Using Trusted Execution Environments.

Oh, S. J., Schiele, B., and Fritz, M. 2019. "Towards Reverse-Engineering Black-Box Neural Networks," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer*, pp. 121–144.

Raisaro, J. L., Klann, J. G., Wagholikar, K. B., Estiri, H., Hubaux, J.-P., and Murphy, S. N. 2018. "Feasibility of Homomorphic Encryption for Sharing I2B2 Aggregate-Level Data in the Cloud," *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science (2017), American Medical Informatics Association*, pp. 176–185.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Schwendicke, F., Samek, W., and Krois, J. 2020. "Artificial Intelligence in Dentistry: Chances and Challenges," *Journal of Dental Research* (99:7), pp. 769–774.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626.

Shamir, A. 1979. "How to Share a Secret," *Commun. ACM* (22:11), New York, NY, USA: Association for Computing Machinery, pp. 612–613.

Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., and Bakas, S. 2019. "Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi, S. Bakas, H.*

Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (eds.), Cham: Springer International Publishing, pp. 92–104.
Shokri, R., Strobel, M., and Zick, Y. 2019. "Privacy Risks of Explaining Machine Learning Models," *ArXiv Preprint ArXiv*:1907.00164.
Shokri, R., Stronati, M., Song, C., and Shmatikov, V. 2017. "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18.
Song, L., Shokri, R., and Mittal, P. 2019. "Membership Inference Attacks against Adversarially Robust Deep Learning Models," in *2019 IEEE Security and Privacy Workshops (SPW)*, IEEE, pp. 50–56.
Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. 2019. "Demystifying Membership Inference Attacks in Machine Learning as a Service," *IEEE Transactions on Services Computing*, IEEE.
Wang, G. 2019. "Interpret Federated Learning with Shapley Values," *ArXiv Preprint ArXiv*:1905.04519.
Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., and Ludwig, H. 2019. "HybridAlpha," P*roceedings of the 12th ACM Workshop on Artificial Intelligence and Security - AISec'19*, ACM Press.
Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. 2018. "Applied Federated Learning: Improving Google Keyboard Query Suggestions," *ArXiv Preprint ArXiv:*1812.02903.
Yao, A. C. 1982. "Protocols for Secure Computations," in *23rd Annual Symposium on Foundations of Computer Science (Sfcs 1982),* pp. 160–164.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. 2016. "Learning Deep Features for Discriminative Localization," i*n 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929.

Learning from IT Security Catastrophes: A Post Catastrophe Analysing Checklist

Emerging Trends in Internet Technologies, Summer Term 2020

Felix Traup
Master Student
Karlsruhe Institute of Technology
felix.traup@kit.edu

Christina Speck
Master Student
Karlsruhe Institute of Technology
spe.christina@gmail.com

Felix Deckers
Master Student
Karlsruhe Institute of Technology
felix.deckers@t-online.de

Peter Lorenz
Master Student
Karlsruhe Institute of Technology
peter-lorenz@outlook.com

Abstract

Background: Our reliance on IT increases the impact of each IT security incident, often making them a costly catastrophe. IT security guidelines and standards aim at recommending security measures, that should prevent IT security catastrophes. However, guidelines and standards tend to be very brought and are not always up to date on security recommendations, protecting from the latest threats.

Objective: This work builds up a checklist on how to learn from the latest IT security catastrophes after they have happened. After immediate firefighting has ended, our checklist can be used by both practitioners and researchers and offers suggestions on how IT security catastrophes can be analysed.

Methods: We iteratively analysed literature on past IT security catastrophes to build a checklist that considers practice as well as research, to help learning from IT security catastrophes. We chose IT security catastrophes, that best reflect our defined IT security catastrophe spectrum. Finally, we provided an exemplary instantiation of the developed checklist on the IT security catastrophe caused by the ransomware WannaCry.

Results: As a result, we developed a checklist that enables both practitioners and researchers to analyse IT security catastrophes in a structured manner after immediate firefighting has ended. Also, we provided a definition, that helps classify an IT security issue as catastrophe.

Conclusion: Further iterations with the most recent IT security catastrophes are advised to continually improve our checklist. Our work thereby contributes to the awareness, on how important it is to build methods in order to learn from past catastrophes.

Keywords: IT security catastrophe, IT security, IT security guidelines, IT security standards, cyber security, security incident

Introduction

With the fast development and increasing complexity of IT (Information Technology) (Alenezi & Zarour 2020), it is argued that IT security becomes more and more difficult (Shin & Williams 2008). Consequently, IT security catastrophes (ITSC) are more likely to occur. In addition, our increasing reliance on IT (Attwood 2020) increases the number of affected market participants (e.g., IT provider, IT consumer), in the occurrence of an ITSC. The occurrence of an ITSC can therefore become very costly. Heartbleed, as an example of ITSC, has made data of many companies vulnerable to attack, due to a simple coding error in an open-source cryptographic library, OpenSSL (Wheeler 2014). A simple coding error sufficed for an estimate of "24–55% of HTTPS servers in the Alexa Top 1 Million" (Durumeric et al. 2014) to be vulnerable. In the case of the Yahoo data breach, three billion user accounts were exposed in the result of an attack (Ormerod 2019). Due to the extent of damage ITSCs can trigger, IT security becomes a crucial requirement for the success of market participants, such as IT providers.

Market participants usually have measures in place to shield themselves from an IT security threat and protect their data (Kersten 2020). Many guidelines are proposed from reliable institutions (e.g., BSI) to support the creation of those measures. However, these guidelines tend to be very broad, applicable to most market participants and covering as many IT security threats as possible (Fliehe 2014). Even though broad guidelines intend to help as many market participants as possible, generic language might not lead to practical IT security measures. Moreover, most guidelines are renewed in fixed time intervals (mostly yearly), not necessarily covering the most recent IT security threats (Schönbohm 2020). With standards such as BSI being updated only yearly, market participants are sometimes left with outdated IT security measures in place over the course of the remaining year.

Furthermore, existing work and especially guidelines often focus on the ex-ante prevention of ITSCs and include ex-post measures only as an ad hoc approach. After immediate 'firefighting' has ended, available ex-post analysis often focuses solely on their impact in one area (e.g., social aspects, technical aspects, legal responsibility) (Trautman & Ormerod 2017b) or in one specific user group (Jackson 2014). Although ex-post analysis offers potential for learning from past ITSCs, not all market participants feel compelled to implement learned IT security measures, such as in the case of Heartbleed, where many companies failed to patch their servers (Durumeric et al. 2014). Often, conclusive studies are missing, that show what a past ITSC made us learn, considering IT security measures. A habit of deducing practical security measures from past ITSCs could be beneficial with regard to the increasing complexity of IT, mentioned above. In the search for a structured approach to deduce practical security measures out of ITSCs, we want to answer the following research question:

How can security measures be deduced from use cases of IT security catastrophes?

To answer the research question, we develop a checklist of analytic considerations for practitioners as well as researchers, to investigate ITSCs ex-post for practical security measures. We focus on a predefined set of ITSC cases, that aims at reflecting the majority of the ITSC spectrum (as defined in "The IT Security Catastrophe Spectrum"). On each of our predefined ITSC cases, we want to map out the considerations that were made after the ITSC, to avoid similar mistakes in the future. The resulting list of considerations serves as checklist of analytic considerations after the occurrence of an ITSC. Finally, the checklist is subjected to an exemplary instantiation with a recent ITSC.

Our work is structured as follows:

- Literature identification: The first part focuses on an identification of literature on our preselected set of ITSC cases. Only literature on the ex-post analysis of the ITSC cases are relevant.
- Deduction: The identified literature is then evaluated for usable analytic considerations, that can be made after the occurrence of an ITSC.
- Synthesis: Iteratively, the hereby found considerations are merged to a checklist.
- Instantiation: At last, the checklist is instantized on a current ITSC.

We expect the checklist to be applicable to an ITSC, regardless of its kind, in order to investigate an ITSC, after its occurrence, for applicable security measures. This work contributes to the development process of approaches on how to learn from ITSCs in three ways. First, we offer potential for the further development of an ITSC ex-post analysing checklist. Second, we provide a possibility for quick and systematic generation

of applicable security measures, in both practice and research. Third, the checklist builds an overview of key security issues, triggering ITSCs.

Foundations of IT Security Catastrophes and their Handling

Foundational Concepts of IT Security

From an IT Security Issue to an IT Security Catastrophe

To understand what triggers ITSCs, the term ITSC needs to be clearly defined, starting with the term 'security'. The terms 'security' and 'safety' are often used interchangeably in daily life, some dictionaries even define them as synonyms (e.g. Merriam-Webster 2020). However, in the context of information technology it is important to distinguish between the two concepts. We follow the definition, provided by Nas (2015). Hence, the following definitions of 'safety' and 'security' will be considered in our work:

> Safety: The state of being away from harm caused by random natural forces.
>
> Security: The state of being away from harm caused by active attackers with malicious intentions.

A non-IT example of both measures of safety and security would be the act of closing all windows before leaving your home. This prevents both rain (random natural force) and thieves (malicious attacker) from entering a house or an apartment.

Not all IT issues harming the IT security can be classified as ITSC. Even further, not all IT issues harm the IT security. We consider an IT issue to be harmful for the IT security, when the IT issue prevents a system (a) to be ready for correct service, (b) to lack improper system alterations, and (c) to prevent unauthorized disclosures. This triad of attributes are called (a) Availability, (b) Integrity, and (c) Confidentiality (Avizienis et al. 2004). Therefore, IT issues compromising at least one attribute can be classified as IT security issue.

We distinguish between three major stages of IT security issues. This definition builds up a taxonomy provided by the NIS Cooperation Group, 2018:

- IT security incidents: "Incidents affecting the security of network and information systems, in any sector of society" (ib.).
- ITSC: Incidents which must be notified to national authorities under the EU-level requirements of Article 14&16 of the NIS directive (Directive (EU) 2016/1148, 2016), Article 13a of the Framework directive (Directive 2009/140/EC, 2009), and Article 19 of the EIDAS regulation (Regulation (EU) No 910/2014, 2014) or under any further national legislation.
- Large-scale ITSC: IT security catastrophes requiring EU cooperation (NIS Cooperation Group 2018).

The respective EU directives, frameworks and regulations build the foundation for national legislation. With focus on different types of IT systems, services, and industries, these legal texts establish that incidents having a 'significant' or 'substantial' impact on continuity, provision, operation, and/or integrity of a service provided need to be notified to national authorities and the European Commission (Directive 2009/140/EC 2009; Regulation (EU) No 910/2014 2014; DIRECTIVE (EU) 2016/1148 2016). Thus, an IT security incident, that has been notified to national authorities and the European Commission, is categorized as ITSC.

The IT Security Catastrophe Spectrum

After defining ITSCs, we need to understand the main triggers for ITSCs. The main triggers mentioned in this work are inspired by the EU Commission's 'Cybersecurity Incident Taxonomy' (NIS Cooperation Group 2018). "The scope of this taxonomy is cybersecurity incidents in general, for the sake of completeness." (ib.). Hence, the taxonomy provides us with a thorough categorization of IT incidents and thereby ITSCs into their trigger categories. Five trigger categories can be identified *(see Figure 1)*, according to the taxonomy (all cited from ib.):

- System failures: "The incident is due to a failure of a system, i.e. without external causes."
- Natural phenomena: "The incident is due to a natural phenomenon."

- Human errors: "The incident is due to a human error, i.e. system worked correctly, but was used wrong [sic.]."
- Malicious actions: "The incident is due to a malicious action."
- Third party failures: "The incident is due to a disruption of a third-party service, like a utility."

As described above *(see Chapter "From an IT Security Issue to an IT Security Catastrophe")*, natural phenomena are the trigger of safety issues. Therefore, we mention natural phenomena in our pentagon for the sake of completeness, but as we focused on security, we did not include ITSCs triggered by natural phenomena into our deductions.

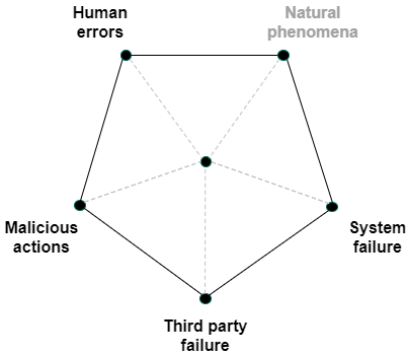

Figure 1. Triggers for IT Security Catastrophes

IT Security Guidelines/Standards

Once an IT security incident counts as ITSC and the further aggravation of the impact on market participants has been stopped, affected market participants might reevaluate the IT security in place. IT security guidelines and standards provide help for the reevaluation of IT security. Although, in order to be able to observe IT security guidelines correctly, a demarcation to the IT security standards must be made. Standards and guidelines provide guidance for improving cyber security, "but guidelines usually lack the level of consensus and formality associated with standards" (Scarfone 2009). Guidelines are applied when standards need to be extended by a technical specification or technical reports and the immediate release of a standard is not possible (International Organization for Standardization 2016). This is the case when the topic in question is still in development (Scarfone 2009).

The large number and diversity of today's security standards and guidelines is a result of the different needs developed by companies (e.g. different industries) but also from the roles and responsibilities of people in the company (Fliehe 2014). This is the reason why numerous committees worldwide are working on the development of IT security standards and guidelines. Due to the large number of different standards and guidelines we will only briefly explain the most common ones.

ISO 27001: The ISO 27001 consists of general requirements for the management of the introduction, implementation, and continuous improvement of information security (Kersten 2020). The standard does not describe explicit methods and criteria for establishing information security. Instead, it provides guidelines that can be selected and applied by the respective user. Therefore, the standard can be used in organizations of different type and size and can be applied in any country. It was first published in 2005 and was revised in 2013 and 2015 due to technical corrections (Kersten 2020).

BSI-Standard 200-1: The BSI-Standard 200-1 is published by the German 'Bundesamt für Sicherheit in der Informationstechnik' (BSI) and was first published in 2008 (BSI-Standard 100-1) and completely revised in 2017 (Bundesamt für Sicherheit in der Informationstechnik 2018). The BSI designs information security in digitisation through prevention, detection and reaction for the state, economy and society and publishes standards or guidelines that are updated regularly (Schönbohm 2020).

The goal of the standard is to find an efficient procedure to manage information security in different organizations, no matter if they are individuals or international organizations (Bundesamt für Sicherheit in der Informationstechnik 2018). In addition, the user is continuously informed about how the recommendations of the standard can be adapted to suit the individual/organization. (Bundesamt für Sicherheit in der Informationstechnik 2018). This standard is primarily intended for persons in charge of information security, security officers, security experts, security consultants, and all parties interested in or charged with information security management.

NIST Cybersecurity Framework: The National Institute of Standards and Technology (NIST) published the NIST Cybersecurity Framework 2013 and revises it regularly, on average bi-annually (Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1 2018; "Evolution of the Framework" 2019).

The framework was developed to improve cybersecurity in any organization or community. It uses already existing and effectively working standards, guidelines, and practices to offer a flexible way to address and manage cybersecurity risk in a cost-effective way (Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1 2018).

Related Work

As well as preventing ITSCs through complying to IT security standards and guidelines, learning from ITSC ex-post is no new concept. It is occasionally found in academia under specific subject areas (e.g. computer science, legal science) (Jackson 2014; Trautman & Ormerod 2017a; Wheeler 2014). Unlike common belief, the ex-post analysis of ITSC is not reserved to computer scientists. An ITSC can be analysed on different levels. For instance, in the context of ITSC, the legal community is debating on security obligations IT provider have towards their users and on penalties for cyber attackers. After an ITSC, one question will always arise: did the IT provider act within their legal obligations? If so, should laws be adapted? If not, are the consequences of the breach of law not heavy enough? (Trautman & Ormerod 2017b).

Another remarkable example for the development of an approach to ex-post analyse an ITSC comes from the universe of law librarians (Jackson 2014). Jackson suggests a set of recommendations that law librarians should assess after 'firefighting' after the catastrophe has ended. This contribution is very similar to our work but due to the limited user focus only applicable to law librarians. Nevertheless, this example serves as an inspiration for building a post-catastrophe analysing checklist.

Research Methods

Criteria for Selection of IT Security Catastrophes

In order to build the checklist, we select ITSC cases, that we analyse for considerations to make after the occurrence of an ITSC. Our selection of ITSC is based on the following three criteria: (1) be current (recent catastrophes), (2) comply with our ITSC definition, and (3) cover the selection scheme.

First, the reason for the importance of actuality is that recent catastrophes are more valuable to our checklist. This assumes that older vulnerabilities are already captured in the regularly updated guidelines and could therefore be prevented today using software or guidelines. Secondly, to use only catastrophes that are relevant for a broad crowd we decided that the analysed ITSC should comply with our definition of ITSCs presented in Chapter "From an IT Security Issue to an IT Security Catastrophe". Third, the last criteria described above is the coverage of the pentagon 'Cybersecurity Incident Taxonomy' *(see Chapter "The IT Security Catastrophe Spectrum")* through the analysed catastrophes. The target is therefore to cover all trigger categories with the selection of ITSC cases, except 'Natural phenomena'.

Literature Identification on Selected IT Security Catastrophes and Analysis

We conduct a qualitative literature identification on the selected ITSC cases, which represent our ITSC set. From the identified literature, we deduce a list of analytic considerations that authors made for developing practical recommendations after the occurrence of each ITSC in the ITSC set. The resulting analytic considerations are then merged into a structured approach in the form of a checklist that practitioners and researchers can use as a starting point to investigate recent ITSCs after they have occurred.

The literature identification with its deduction of analytic considerations is conducted in the following iterative manner *(see Figure 2)*:

- We choose an ITSC case *i* out of our ITSC set, composed of *N* ITSC cases.
- We deduct analytic considerations by conducting a literature identification on the respective case *i*, answering the questions:
- How did authors analyse the ITSC case *i* ex-post in our identified literature? The answer of this question provides us with an unstructured list of analytic considerations.
- From which of the analytic considerations could practical measures be deduced from? With this step we have a more refined list of analytic considerations with their respective practical recommendations.
- We add the deduced list of analytic considerations to our checklist. Similar considerations are merged into one consideration.
- If we haven't conducted our literature identification on each ITSC case from our ITSC set *(i≤N)*, we go into the next iteration *(i=i+1)*, with the next ITSC case of our ITSC set.

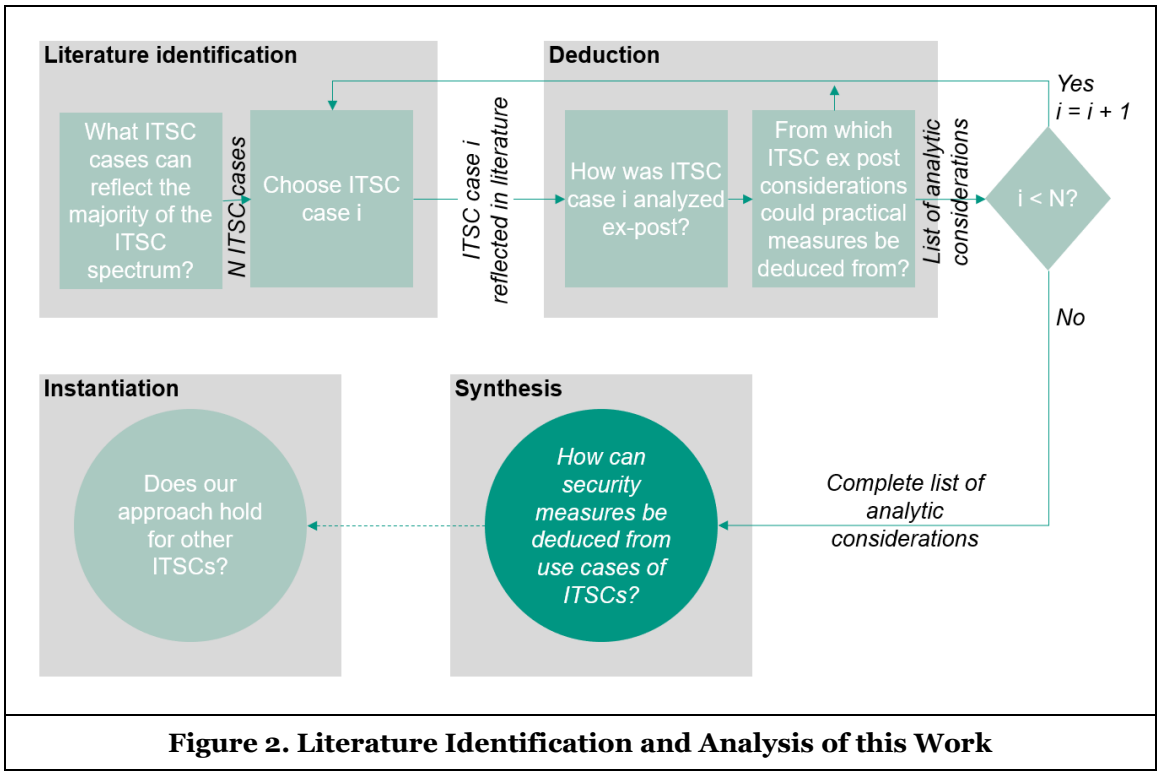

Figure 2. Literature Identification and Analysis of this Work

Finally, after *N* iterations, our complete list of analytic considerations builds our checklist. To provide a verification of the resulting checklist, we conduct an exemplary instantiation of a recent ITSC, that has not been used for the building of the checklist.

The literature identification itself needs to be adapted to the ITSC case. Not every ITSC case is covered by academia. To avoid relying exclusively on grey literature in some ITSC cases, we conduct the literature identification as follows. First, the name of the ITSC case represents the search string in an EBSCOhost search. If the EBSCOhost search either doesn't provide any result or the results are only constituted by grey literature, a search on Publish or Perish (Harzing 2007) is conducted with the same search string as in the EBSCOhost search. The goal of the Publish or Perish search is to find a research paper that covers the ITSC case by suggesting practical measures that could have prevented the ITSC case. Further, that research paper can be used as a starting point of a forward reference search, with citing literature indicated by Google Scholar.

A forward reference search focuses on the publications created after an article's date of publication, so called secondary studies (Felizardo et al. 2016). The forward reference search is usually carried out in several

iterations. The starting point reveals a certain number of citing articles in the first iteration. Based on these citing articles, relevant articles can be identified and investigated for their citing articles – resulting in another iteration. Since our work focuses on the development of a checklist, our forward reference search continues if the previous iteration delivers a novel analytic consideration for our checklist, otherwise the search is aborted.

Whether the resulting list of literature is based on the EBSCOhost or the forward reference search, the list is then categorized into literature that is included into or excluded from the deduction phase. Literature is excluded if it is non-English, inaccessible, off-topic, out of scope or of low relevance. Low relevance includes literature, where no central argument is based on the respective ITSC. Moreover, grey literature is treated as the respective author's opinion.

Deducing a Structured Checklist for Analysing IT Security Catastrophes Ex-Post

Selected IT Security Catastrophes

Before identifying the literature, according to Chapter "Literature Identification on Selected IT Security Catastrophes and Analysis", we select the ITSCs to be searched for in literature, according to Chapter "Criteria for Selection of IT Security Catastrophes". Our selection of ITSCs is composed of five different ITSCs *(N=5)*, which are described below.

Heartbleed

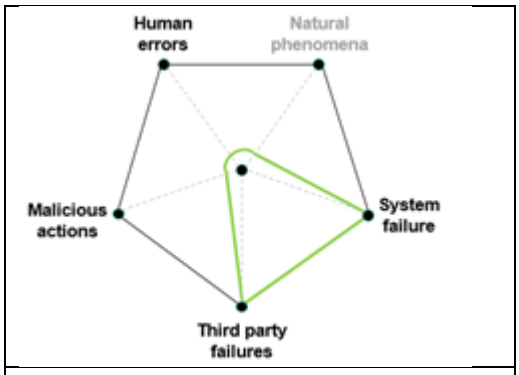

Figure 3. Heartbleed Classification

Heartbleed is a buffer over-read failure in the OpenSSL cryptography library which affected about 500.000 websites since 2012 (Mutton 2014) and was caused by an update on March 3, 2012 which implemented 'Heartbeat'. The update improves the monitoring of the connection status and reduces the administrative effort for long-lasting client/ server connections. (Carvalho et al. 2014).

The beat is composed of a short character string and a number which indicates the length of the string (Durumeric et al. 2014). The character string is sent to the server and answered with the same string, but the replying software does not check whether the specified string length and the actual length match. Malicious attackers could use the missing query to send a short string and a large number which makes the server reply with the random parts of his main memory located right next to the short string until the given number of chars is reached. Using the described method, attackers can capture valuable information such as private keys, passwords or usernames which can be used to steal data or eavesdrop communication from the accounts (Lewis 2014a). Although OpenSSL is widely used, the failure was not found until April 7, 2014 (Durumeric et al. 2014). One week later the failure was published simultaneously with the patch, but to fix the vulnerability updating the system is not enough, it is also necessary to revoke old certificates and issue new ones and to change the private key.

The classification of Heartbleed in the Cybersecurity Incident Taxonomy is the following *(see Figure 3)*:

- System Failure: The failure is caused by a lack of one specific test in the software.
- Third Party Failure: Programming error was caused by the use of a trusted third-party software (Open SSL protocol).

Xerox

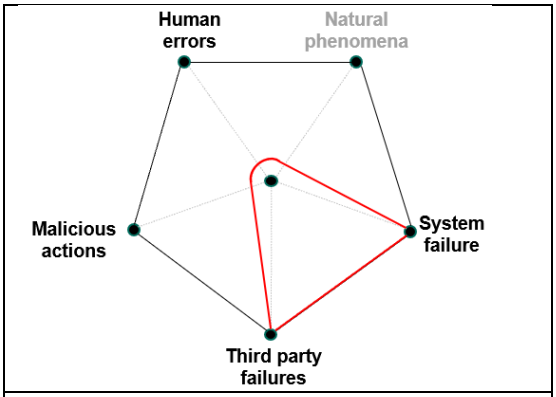

Figure 4. Xerox Classification

In 2013 the computer scientist David Kriesel discovered in computer scans that the printing workstations of the Xerox company happened to randomly replace certain characters in the scans with other similar-looking ones (Jehle 2018). Xerox stated that they have been aware of the bug for 8 years and pointed out that there is a warning for this issue in the user manual. Nevertheless, the customers used the wrong settings for scanning and created unintentionally manipulated documents (Kriesel 2013b) *(later further investigated, please refer to Chapter "Instantiation of Checklist with the WannaCry Ransomware Attack").*

The bug was caused by incorrect parameterization when encoding in the JBIG2 image format (Kriesel 2013b), which led to breaches of integrity in the form of information loss and incorrect changes in PDF-documents. In response to the Xerox bug, the BSI has published a guideline that prohibits the use of JBIG2 in tandem with pattern matching due to the loss of data integrity (BSI Technische Richtlinie 03138 Ersetzendes Scannen Bundesamt für Sicherheit in der Informationstechnik 2015; Kriesel 2015).

The classification of Xerox in the Cybersecurity Incident Taxonomy is the following *(see Figure 4)*:

- System Failure: Due to faulty parameterization in image format JBIG2.
- Third Party Failure: Error occurs at the retail customer and was caused by Xerox (third party).

Yahoo Data Breach

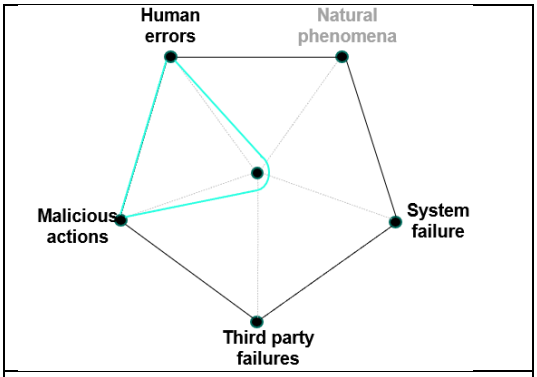

Figure 5. Yahoo Data Breach Classification

In 2013 an unnamed third party hacked into the Yahoo's Data Network and stole up to 3 billion user accounts with sensitive information like names, telephone numbers, passwords, and unencrypted security questions (R. McMillan 2017; Pearce 2018). This was accomplished by sending a phishing mail to a Yahoo employee with access to the data network and account management tool (Williams 2017). The hackers then gained access to user accounts by creating faked session cookies and searched for sensitive user information.

The classification of the Yahoo Data Breach in the Cybersecurity Incident Taxonomy *(see Figure 5)*:

- Human Error: First access to system through phishing mail by incautious employee.
- Malicious Action: Use of data to access user accounts and steal user data.

Amazon Server Shutdown

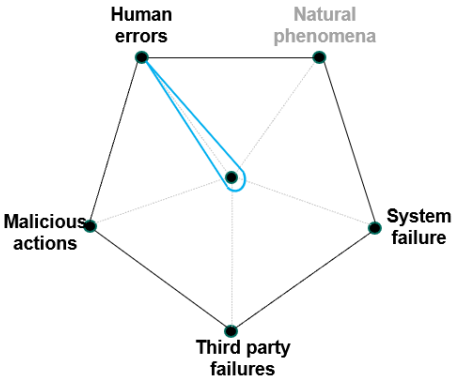

Figure 6. Amazon Server Shutdown Classification

On February 28, 2017, a four-hour disruption occurred in the Amazon S3 Service in the Northern Virginia (US-EAST-1) region (Amazon Web Service 2017, Summary of the Amazon S3 Service Disruption). Due to an incorrect command from a software engineer, a larger than intended number of servers was removed from the network during a routine server check-up (Guy 2017). The accidentally removed servers were part of other server subsystems, causing a chain reaction resulting in further server shutdowns. Customers of the server provider had to accept increased access times to their websites or failures of parts of their billing system during the downtime (How E-Commerce Sites Were Affected by the Amazon S3 Outage Apica Inc. 2017). Affected customers were among others Victoria's Secret, Nike, and Target.

The classification of the Amazon Server Shutdown in the Cybersecurity Incident Taxonomy *(see Figure 6)*:

- Human Error: The reason for the outage was an incorrect input of a software engineer.

Equifax Data Breach

Equifax Inc. is one of the largest American consumer credit reporting agencies and was the target of a 76 days long-term data breach started on May 13, 2017 (Bottum 2018). The vulnerability exploited for the malicious action is caused by an open-source web application framework called Apache Struts 2. Although the failure of the Software was known months in advance and the patch was available since March 7, 2017 it hasn't been installed on the Equifax system until July 29, 2017 (Rasalam & Elson 2019). Among the possibly gathered data are the four main personal security identifiers: name, address, birth date and social security number (Primoff & Kess 2017), which can be used for many crimes. Since consumer credit reporting is widespread, about 145 million individuals have been affected (Bottum 2018).

The classification of the Equifax Data Breach in the Cybersecurity Incident Taxonomy *(see Figure 7)*:

- Human Error: Administrators failed to install the patch after the release.
- Third Party Failure: The error was caused by a vulnerability of an open source software.

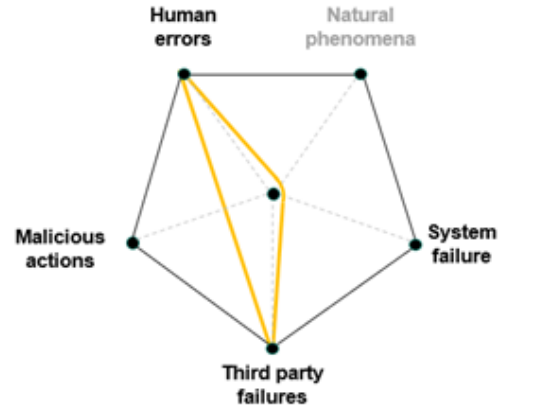

Figure 7. Equifax Classification

Deducing from Heartbleed

With *i=1*, we start our literature identification and analysis with Heartbleed. The following analysis was conducted on a set of literature found on EBSCOhost using the search term 'Heartbleed'. The search was limited to the attribute 'Scholarly (Peer Reviewed) Journals' and, after removing duplicates, provided us with 32 individual items of literature. Eleven search results were off topic, mainly because they were out of scope. Six papers were of low relevance to our research question, mainly because they did not contain any recommendations. After limiting the result list to publications in English or German and therefor excluding one paper, we conducted our deduction on a set of 14 pieces of literature. Three of these were articles in the non-peer reviewed industry magazines 'Computer' (Garber 2014; Wheeler 2014), and 'IEEE Security & Privacy' (Carvalho et al. 2014) that were published in the direct aftermath of the publication of the Heartbleed issue. The remaining eleven were published between 2014 and 2018. A brief overview can be found in *Table 1*.

Search Term		'Heartbleed' limited to attribute 'Scholarly (Peer Reviewed) Journals'	
Database		EBSCOHOST	
Total search results		32	
Excluded	Off topic	11	
	Of low relevance	6	
	Non-English	1	
Included	Peer reviewed journal papers	11	(Allen et al. 2017; Beurdouche et al. 2017; T. Chen et al. 2016; Durumeric et al. 2014; Jackson 2014; Kamp 2014; Kang & Park 2017; Kupsch & Miller 2014; M. McMillan 2015; Vassilev & Celi 2014; Zhang et al. 2018)
	Grey literature	3	(Carvalho et al. 2014; Garber 2014; Wheeler 2014)

Table 1. Overview of Literature Search Results for 'Heartbleed'

From these 14 pieces of literature, we deduced three steps of our structured checklist *(see Table 2)*:

- Testing and verification (Beurdouche et al. 2017; Carvalho et al. 2014; T. Chen et al. 2016; Kang & Park 2017; Kupsch & Miller 2014; Vassilev & Celi 2014; Wheeler 2014)
- User interaction (Jackson 2014)
- Importance of cybersecurity in companies and/or society (Garber 2014; Kamp 2014; M. McMillan 2015)

Testing and Verification

As the Heartbleed vulnerability has been active for more than two years *(see Chapter "Selected IT Security Catastrophes")*, the question of why it has not been discovered earlier became part of several research papers. A potential research hypothesis is that "Heartbleed created a significant challenge for current software assurance tools, and we do not know of any such tools that were able to discover the Heartbleed vulnerability at the time of announcement." (Kupsch & Miller 2014). The reasons for the technical difficulty of detecting the issue lie in (a) the use of pointers, (b) the complexity of the execution path from buffer allocation to misuse, (c) validity bytes of the TLS message being a subset of the allocated buffer, and (d) contents of the buffer not coming directly from the attacker (Kupsch & Miller 2014). In recent years, there have been new approaches to software testing that enhance our abilities to find issues like Heartbleed. Both metamorphic testing (T. Chen et al. 2016) and fuzzy mining (Kang & Park 2017) were proven to – with little alteration – be able to redetect the Heartbleed malfunction. Formal verification of its composite state machine showed that OpenSSL not only did not check the content of the client's heartbeat message (Heartbleed), but also allowed for skipping the 'verify' message in TLS exchange protocols and thus opened doors for client impersonation (Beurdouche et al. 2017). Furthermore, negative testing methods (Wheeler 2014), context configured source code weakness analysers (Carvalho et al. 2014) and Memcheck (Vassilev & Celi 2014) have been assumed, but not proven to be able to detect Heartbleed.

Unit of analysis	Steps of approach						Unassigned		
Articles/papers	**Testing & verification**		**User interaction**		**Importance of IT security**		**Undefined category**		
Citation	New methods for software testing	Formal verification of software	Access rights	User training	Strategic importance of IT security in organizations	Funding of open source software	Complexity of IT systems	Server certificates and passwords	New sources of information about security threads
Allen et al. 2017									x
Beurdouche et al. 2017			x						
Carvalho et al. 2014	x						x		
Chen et al. 2014	x								
Durumeric et al. 2014								x	
Garber 2014						x			
Jackson 2014			x	x					
Kamp 2014a						x			
Kang & Park 2017	x								
Kupsch & Miller 2014	x								
Liang Zhang et al. 2018								x	
McMillan 2015					x				
Vassilev & Celi 2014	x								
Wheeler 2014	x						x		

Table 2. Concept Matrix of Heartbleed Literature

The cited literature shows that Heartbleed was an inspiration for researchers to fruitfully work on testing and verification. These altered methods could then be applied to different settings, including commercial IT systems relevant for practitioners. We therefor include in our structured checklist the following two recommendations for researchers and practitioners:

- For practitioners: Check, whether some researchers have introduced a new testing method to find issues like the recent catastrophe.
- For researchers: Check, why existing testing methods did not find the issue and if existing testing patterns could be modified or altered to find similar issues.

Commercial Focus and Importance

In the early aftermath of Heartbleed, The Wallstreet Journal stated the total annual funding of the OpenSSL foundation was less than $2,000 USD (Yadron 2014). Although this number was put in perspective by Steve Marquess of OpenSSL foundation (Marquess 2014), it generated a public discussion about funding of open source software and cybersecurity in general. “Security experts say Heartbleed’s existence for two years without being discovered illustrates that a lot of important open source Internet [sic.] software is not funded, developed, or reviewed carefully enough.” (Garber 2014). Kamp states that the most productive way to assure decent software quality was to pay the developers (Kamp 2014). In the context of commercial IT in hospitals, it has been shown that “[o]rganizations that do not invest proactively in IT security face a significant risk of incurring much greater costs from incidents involving compromised data security.” (M. McMillan 2015).

From the cited research we include in our structured checklist the following recommendations for practitioners and researchers:

- For practitioners: Check the funding of your organization’s IT security department.
- For researchers: Check, whether the case could have been prevented by better funding of the software developers or IT security department.

User Interaction

The Heartbleed flaw was also used as inspiration for researchers to advocate on increased user training. A remarkable contribution was made by Jackson (Jackson 2014), researching on the learnings that law librarians could take from IT security catastrophes. They suggest that law librarians should for example encourage users to use safe internet connections and “[…] assist [the IT department ed.] by evaluating security products and services and participating in the educational process.” (Jackson 2014). Furthermore, it is arguable that the philosophy of freedom of law libraries materialized in non-restrictive access rights might impose a security risk (Jackson 2014).

Although it is possible to argue that Heartbleed could not necessarily have been prevented by users, we consider the issue of user training and access rights as relevant step of our structured checklist. We deduct the following recommendation for practitioners:

- For practitioners: Rethink access rights and strengthen user training.

Further Areas Mentioned but not yet Included in the Structured Checklist

Further research areas mentioned in or applied in the literature on the matter of Heartbleed were the complexity of IT systems (Carvalho et al. 2014; Wheeler 2014), server certificates or password changes (Durumeric et al. 2014; Zhang et al. 2018) and the use of new sources of information to find out about IT security incidents earlier (Allen et al. 2017). These areas were not covered thoroughly enough to allow for generalization. Only together with deductions from other IT security catastrophes, these areas could form further steps of the structured checklist developed in this paper.

Deducing from Xerox

For the analysis of the Xerox catastrophe *(i=2)* the following literature analysis was conducted. In the beginning 3 results (Aldemir et al. 2020; Rodríguez-Díaz & Sánchez-Cruz 2014; Shiah & Yen 2013) were

found on EBSCOHOST with the search string 'JBIG2' for the period 2013-2020. None of the three search results have any connection to the topic. For that reason, the search was extended to Publish or Perish (Harzing 2007) using the search term 'Xerox JBIG2'. The search provided 73 pieces of literature, of which 2 were included in the deductions. An additional forward reference search based on Kriesel (2013b) provided 13 pieces of literature of which 13 are used in the deduction. The second iteration of the literature search yielded only grey literature. The grey literature is included in the deduction, because it supports the containment of the Xerox catastrophe. A brief overview can be found in *Table 3*.

Search Term		'Xerox JBIG2' and 'Published between 2013 and today'	
Database		Google Scholar	
Total search results in 1st iteration		73	
Excluded	Off-topic	66	
	Low relevance	0	
	Non-English	5	
	Double	0	
	No access	0	
Included	Grey Literature	2	(Jehle 2018; Kriesel 2013b)
	Journal Paper	0	
Forward research based on		(Kriesel 2013b)	
Total search results in 2nd iteration		13	
Excluded	Off-topic	0	
	Non-English	0	
	Double	0	
	No access	0	
Included	Grey Literature	13	(Chang 2013; Coldewey 2013; Collis 2013; Coy 2013; Jehle 2018; Kelion 2013; Kriesel 2013a; "Xerox schaltet Zahlendreher-Funktion ab" 2013b, Kriesel 2013b, 2013c, 2013d; "Xerox Scan-Problem: Was Sie wissen müssen" 2013a; Kriesel 2015)
	Journal Papers	0	

Table 3. Overview of Literature Search Results for 'Xerox'

From these 13 pieces of literature, we included 13 and deduced one new step for our structured checklist (see Table 4):

- Communication of bug/failure (Chang 2013; Coldewey 2013; Collis 2013; Coy 2013; Jehle 2018; Kelion 2013; Kriesel 2013a, 2013b; "Xerox schaltet Zahlendreher-Funktion ab" 2013b, Kriesel 2013c, 2013d; "Xerox Scan-Problem: Was Sie wissen müssen" 2013a; Kriesel 2015)

User Interaction

In addition to the discovery of the category "Communication of bug", the category "User interaction" is examined again, since many referenced sources criticized a poor user training. As a result, the category is examined from a different perspective.

During the uncovering of the problem, it was discovered that Xerox had been aware of the error for eight years and that recommendations for users were included in the manual as well as in the printer's associated software ("Xerox schaltet Zahlendreher-Funktion ab" 2013b; Kriesel 2013d). Despite these recommendations it is not clear to the user which settings lead to a possible change in the document during scanning ("Xerox Scan-Problem: Was Sie wissen müssen" 2013a; Kelion 2013; Kriesel 2013d). As a result of the misunderstanding that arose during the application of the machines, Xerox has published a statement

regarding the correct application of the scan settings. This led to the rectification of the situation ("Xerox Scan-Problem: Was Sie wissen müssen" 2013a). Furthermore, the scanner settings have been revised about unambiguous comprehensibility in order to exclude further errors. We therefore include the following step for practitioners into our method:

- For practitioners: Introduce an unambiguous user training.

Unit of analysis	Assigned									
Articles/papers	**User interaction**		**Importance of IT security**			**Communication of bug/failure**				
Citation	Access rights	User training	Strategic importance of IT security in organizations	Explicit roles responsible for cybersecurity	Funding of open source software	Transparent communication with media	How to reach all affected users?	Transparent communication with public	Publication of press release after careful investigation	If third-party software is used, customers must be informed of any errors/problems of the third-party software
Kriesel D. 2013a						x	x	x	x	
Xerox Inc. 2013		x				x	x	x		
BSI 2015						x	x	x		
Jehle C. 2018						x	x	x		
Coldeway D. 2013										x
Colins H. 2013						x		x	x	
Chang J.M. 2013						x		x		
Kelion L. 2013		x				x		x		
Coy P. 2013						x		x	x	
Spiegel online 2013						x		x	x	
Kriesel D 2013b		x				x	x	x	x	
Kriesel D 2013c		x				x	x	x	x	
Kriesel D 2013d		x								

Table 4. Concept Matrix of Xerox Literature

Communication of Bug/Failure

After the discovery of the error and the explanation of the consequences - "models dating back to 2005 offer Jbig2 compression, representing hundreds of thousands of individual units"- (Kelion 2013), Xerox initially did not take the error seriously (Coldewey 2013; Kriesel 2013b, 2013d). As a result, the discovery got published by Kriesel and then was critically discussed in international press shortly after (Chang 2013; Collis 2013; Coy 2013). Even after its publication by the press, Xerox's still considered the bug as minor, because the company assumed that only a minority of users could experience the bug (Kelion 2013).

Communicating the bug more quickly and transparently would have probably meant less attention and less damage to the company (Kriesel 2013a).

Furthermore, it must be considered how all users affected by the error can be contacted as quickly as possible ("Xerox schaltet Zahlendreher-Funktion ab" 2013b; Jehle 2018; Kriesel 2013d). In the Xerox case, contacting affected users is difficult because the scanners are distributed through subcontractors and, as a result, complete user information is not available. Whether all affected systems have been patched to date cannot be verified.

From the cited research we include in our structured checklist the following recommendations for practitioners and researchers:

- For practitioners: Check, whether the current catastrophe requires a change in the communication strategy of your organisation.
- For researchers: Check, whether the error and the resulting consequences would not have occurred or would have been less if the information had been diffused more rapidly.

Deducing from Yahoo Data Breach

For the analysis of the Yahoo data breach *(i=3)* the following literature analysis was conducted. In the beginning 2 peer reviewed journal paper were found with the search string 'yahoo data breach', using the database EBSCOhost. None of the two search results were included in the deduction, due to non-English content and no relation to the topic. Further, the search was extended to Publish or Perish (Harzing 2007). With the search string 'yahoo data breach', 9 pieces of literature were found, of which 1 was included into the deduction. An additional forward reference search provided 46 pieces of literature of which 15 are used in the deduction. Of those 15, the second iteration was based on 5 pieces of literature, that deal with the Yahoo data breach as source for new learning potential as well as provide new steps to our approach. Through the second iteration 2 new pieces of literature were included into the deduction. However, no new step has been found for our checklist, so the forward reference search has been aborted after 2 iterations. A brief overview can be found in *Table 5*, for more detail, please refer to attached multimedia appendix.

From these 17 pieces of literature, we used 8 pieces of literature to deduce two new steps for our structured checklist *(see Table 6)*:

- Applicability of prevention methods (Blackwell 2018; Blue et al. 2017; Jalkanen 2019), taking up 2 sub-steps of the former category 'Unassigned'
- Law enforcement (Brill & Jones 2016; Edwards 2018; Ogle 2019; Ormerod 2019; Trautman & Ormerod 2017a)

Applicability of Prevention Methods

Since the Yahoo data breach has affected users' email accounts, possible data affected by the breach include "bank and family details as well as passwords that users share between systems or have received in their email accounts" (Thielman 2016). Although Yahoo has the responsibility to protect user data, the breach has revealed that users could have protected themselves with simple actions. For instance, saving emails and sharing passwords via email are aspects that favored the accessibility of sensitive data for intruders. Moreover, one password was often used for different websites (Weiss 2016), revealing once more that users need a different password for each website. Therefore, on top of changing user passwords regularly, possible practicable security instructions can be: (1) Emails should be deleted when the user has no use for them in the future, (2) one password should provide authentication for one website.

On the company side, it has been argued that Yahoo could have reset passwords automatically in order to protect their users' accounts (Jalkanen 2019). Further, Yahoo has been known to not attach enough importance to data security (Blackwell 2018), since their security measures have been questioned by own employees. For one, Yahoo did not use end-to-end encryption, but instead used "depreciated one-way encryption functions such as MD5 and SHA-128" (Blue et al. 2017). Moreover, Yahoo could have provided a reward program for hackers – a common practice for IT companies (e.g., Google Vulnerability Reward Program) (Pabrai et al. 2020) - in order to reveal existing IT security vulnerabilities (Jalkanen 2019).

Therefore, practicable security instructions for organizations, that can be learned from the Yahoo data breach are: (1) favour end-to-end encryption, (2) provide a reward program for hackers.

Starting point for the forward reference search		(Trautman 2017)	
Database		Publish or Perish	
Total search results in 1st iteration		46	
Excluded	Off-topic	9	
	Low relevance	8	
	Non-English	1	
	Double	2	
	No access	11	
Included	Grey Literature	5	(Jalkanen 2019; King 2017; Rife 2019; Trautman et al. 2020; Wynne 2019)
	Journal Paper	10	(Beckmann et al. 2018; Blue et al. 2017; Brill & Jones 2016; Edwards 2018; Ogle 2019; Ormerod 2019; Trautman 2018; Trautman & Ford 2018; Trautman & Ormerod 2017a, 2018)
Total search results in 2nd iteration		11	
Excluded	Off-topic	5	
	Non-English	1	
	Double	3	
	No access	0	
Included	Grey Literature	1	(Blackwell 2018)
	Journal Paper	1	(Furey & Blue 2019)

Table 5. Overview of Literature Search Results for 'Yahoo Data Breach'

Since the ability to formulate security measures, that could have prevented the ITSC, express applicability of the resulting prevention methods, we include the following steps into our checklist:

For practitioners: Check, whether user security instructions can be formulated and check, whether security instructions can be formulated for the members of your organization.

For researchers: Check whether the given instructions could have prevented the ITSC.

Law Enforcement

After being considered the biggest data breach in history, the question of accountability has been raised in academia (Brill & Jones 2016; Ormerod 2019). Trautman even argues in his law review, whether Yahoo's conduct before and after the breach "constitutes a breach of the [legal ed.] duty to provide security, the duty to monitor, the duty to disclose, or some combination thereof" (Trautman 2017).

For instance, Yahoo's conduct after the breach has been criticized for its way of disclosing the data breach. In fact, user information was already compromised in 2014. Only in 2016, the data breach had been disclosed for the first time, leaving many users compromised for two years (Trautman 2018). Allegedly, Yahoo knew about the breach long before the public was informed (Edwards 2018; Ogle 2019). Yahoo had even the chance to protect their users after the breach has been noticed by the company, by simply resetting all user passwords, as mentioned above. Yahoo's then-CEO Marissa Mayer, however, feared that a password change would drive email users away to other services (Ormerod 2019).

Moreover, Yahoo's conduct before the breach revealed how deliberately Yahoo's management chose to take a risk, by not providing the necessary security measures for their users. In fact, Yahoo employees advised

to increase security measures (e.g., by adopting end-to-end encryption). Yahoo's management responded, that "end-to-end encryption would block the company from indexing and searching the contents of users' message data" (Ormerod 2019) and thereby preventing the company to use that data for tailored advertisement purposes in the future.

Unit of analysis	Steps of checklist														
Articles/papers	**User interaction**		**Importance of IT security**			**Communication of bug/failure**					**Applicability of prevention methods**			**Law enforcement**	
Citation	Access rights	User training	Strategic importance of IT security in organizations	Explicit roles responsible for cybersecurity	Funding of open source software	Transparent communication with media	How to reach all affected users?	Transparent communication with public	Publication of press release after careful investigation	If third-party software is used, customers must be informed of any errors/problems of the third-party software	Assess complexity of own IT system	Assessing problem for instructions concerning provider protection	Assessing problem for instructions concerning user protection	Assess if laws are heavy enough to discourage unlawful behavior	Assess if laws are enforced to punish breach, miscommunication
Beckmann et al. 2018								x							
Blackwell 2018												x			
Blue et al. 2017												x			
Brill & Jones 2016														x	x
Edwards 2018			x											x	x
Furey & Blue 2019			x												
Jalkanen 2019			x					x				x			
King 2017			x												
Ogle 2019														x	x
Ormerod 2019			x											x	x
Rife 2019			x												
Trautmann 2018		x	x												
Trautmann & Ford 2018			x												
Trautmann & Ormerod 2017														x	x
Trautmann & Ormerod 2018			x												
Trautmann et al. 2020						x	x	x							
Wynne 2019			x			x	x	x							

Table 6. Concept Matrix of Yahoo Data Breach Literature

With Yahoo's decisions on security, the company chose profitability over the investment in security. Yahoo as well as other companies tend to balance profitability and security, in order to maximise profits (Trautman 2018). Ormerod argues that "underinvesting in security and retaining excessive data are often perfectly rational decisions" (Ormerod 2019). However, in the case of Yahoo, the costs of not protecting users' information has not been assessed correctly. Legislators therefore felt the need to put more pressure on companies.

Only after the Yahoo data breach, the Federal Trade Commission (FTC) Policy Statement on Unfairness has been broadened to "pursue companies that fail to adequately protect users' and customers' information" (Ormerod 2019). Until then, "the FTC has only alleged unfairness in instances involving the unauthorized disclosure of (1) directly-identifiable personal information (2) that is clearly 'sensitive.'" (Brill & Jones 2016). Consequently, the cases, that have been mostly brought forward, were cases involving provable monetary harm (Brill & Jones 2016).

Yahoo's "settlement of $35 million penalty to settle charges that it misled investors by failing to disclose one of the world's largest data breaches" (Ormerod 2019) seems minuscule, considering that three billion user accounts have been exposed (Ormerod 2019). We therefor include the following steps to our checklist:

- For practitioners: Check, whether laws are enforced to punish the occurred ITSC and its possible miscommunication.
- For researchers: Assess, whether the current case law complies with the meaning of the law.

Deducing from the Amazon AWS Outage

Search Methodology		Two searches with the term 'Amazon AWS outage', the second one including 'SmartText Searching' and results limited to the attribute 'Scholarly (Peer Reviewed) Journals'	
Database		EBSCOHOST	
Total search results		64	
Excluded	Off topic	38	
	Of low relevance	7	
	Non-English	5	
Included	Peer reviewed journal papers	4	(Araujo et al. 2019; Arora et al. 2018; B. Chen & Curtmola 2017; Ivanova et al. 2018)
	Grey literature	10	("Amazon AWS outage impacts dozens of online retailers" 2017; Frank 2017; Gaudin 2017; Jaeyoung et al. 2019; Li et al. 2020; Rash 2017; Saran 2017; Scott 2018; Weise 2017a, 2017b)

Table 7. Overview of Literature Search Results for 'Amazon AWS outage'

The following analysis was conducted on a set of literature found on EBSCOHOST using the search term 'Amazon AWS outage'. The search provided 18 results. A second repetition using the 'SmartText Searching'-feature and limiting the results to the attribute 'Scholarly (Peer Reviewed) Journals' provided further 46 search results. Of these 64 pieces of literature, 14 were included in the deductions *(see Table 7)*. A brief overview can be found in *Table 7*. For a detailed view on the concepts covered in literature on AWS, see *Table 8*.

From nine of these 14 pieces of literature, we deduced one further step of our structured checklist. The remaining five pieces did not add any further category or did not deviate significantly from the recommendations deduced from the other ITSC. The deduced further step of our checklist is:

- Integrity (Araujo et al. 2019; Arora et al. 2018; B. Chen & Curtmola 2017; Ivanova et al. 2018; Jaeyoung et al. 2019; Li et al. 2020; Rash 2017; Saran 2017; Scott 2018)

Units of analysis	**Data loss prevention**		**Testing**
Citation	New software for assuring integrity	Rethink system architecture	New testing methods for software
Araujo et al. 2019	x	x	
Arora et al. 2018	x	x	
Chen & Curtmola 2017	x		x
Ivanova et al. 2018		x	
Jaeyoung et al. 2019		x	
Li et al. 2020	x		
Scott 2018		x	
Amazon AWS outage impacts dozens of online retailers 2017			
Weise 2017a			
Weise 2017b			
Frank 2017			
Gaudin 2017			
Rash 2017		x	
Saran 2017		x	

Table 8. Concept Matrix of Amazon AWS Outage Literature

Integrity

When a significant proportion of AWS-servers was shut down by a human error in 2017 (Amazon Web Services Inc. 2017; Frank 2017; Gaudin 2017; Weise 2017b), according to the blog Retailcustomerecperience.com, more than 50 online retailers experienced downtimes in their system ("Amazon AWS outage impacts dozens of online retailers" 2017; Weise 2017a). This downtime showed that even in the age of seemingly all-available cloud computing, availability might become an issue. In the direct aftermath the 2017 security catastrophe, Chen and Curtmola proposed a method for server-side testing in a distributed cloud setting where nodes might act maliciously. In this setting the proposed server-side repair mechanism will help to assure integrity (B. Chen & Curtmola 2017). The AWS outage was also an inspiration for researchers to develop and suggest new and safer cloud architectures (Araujo et al. 2019; Arora et al. 2018; Ivanova et al. 2018). In media and specialized magazines the need for new and safer cloud architectures was stretched (Jaeyoung et al. 2019; Rash 2017; Saran 2017; Scott 2018; Wang et al. 2018).

From our literature identification we therefor deduce the following suggestions for practitioners and researchers:

- For practitioners: Check, if there are new methods available for assuring data integrity.
- For researchers: Check, if existing system architecture models could be changed or altered or if the system could be enhanced by software to better cope with the risk of data loss or improper alteration.

Deduction from the Equifax Data Breach

The following analysis is conducted on a set of literature found on EBSCOHOST using the search term 'Equifax data breach'. In prevention of receiving many off-topic results, only papers between the data breach in 2017 and today are considered. The search provided 26 pieces of peer-reviewed literature of which 9 are used for the deduction. A brief overview can be found in *Table 9* and *Table 10*.

Search Term		'Equifax data breach' limited to attributes 'Scholarly (Peer Reviewed) Journals' and 'Published between 2017 and today'	
Database		EBSCOHOST	
Total search results		26	
Excluded	Off topic	11	
	Low relevance	3	
	No access	3	
Included	Relevant peer reviewed journal papers	9	(Ahmad & Barbacki 2019; Berghel 2017; Bottum 2018; Lin 2018; Luszcz 2018; Rasalam & Elson 2019; Skedsvold 2017; Tantleff 2017; Trope 2018)

Table 9. Overview of Literature Search Results for 'Equifax Data Breach'

From these 9 pieces of literature, we deduced three steps for our structured checklist:

- Commercial focus and importance: (Ahmad & Barbacki 2019; Berghel 2017; Lin 2018; Luszcz 2018; Trope 2018)
- Testing and verification: (Ahmad & Barbacki 2019; Berghel 2017; Luszcz 2018; Rasalam & Elson 2019; Trope 2018)
- Communication of bug/failure: (Bottum 2018; Skedsvold 2017; Tantleff 2017; Trope 2018)

Commercial Focus and Importance

After the 'Apache Struts 2' vulnerability was disclosed on March 8, 2017 (Bottum 2018) the IT department was assigned to install the published security patch within 48 hours, according to the testimony of the former Equifax CEO (Smith 2017). About six months and 143 million stolen datasets later it is clear that the IT department for unknown reasons has not followed their orders. The failure implies that the importance of cybersecurity in the specific firm is insufficient. Four of nine analysed paper recommend increasing the importance of cybersecurity in the company to avoid this kind of human failure (Lin 2018; Berghel 2017; Ahmad & Barbacki 2019; Trope 2018). The rise of importance can be designed in different ways, but it should sensitize the employees and implement a system which prevents security updates from being missed. Additionally, the donation amount for used open source code should be reconsidered (Luszcz 2018) to support them during testing to avoid errors in advance.

- For practitioners: Ensure that all employees are aware of the monetary risk that goes along with a weak cybersecurity and search for tools which coordinate the software actuality.
- For researchers: Compare existing software actuality coordination systems.

Testing and Verification

During the attacking period from March 13, 2017 until July 29, 2017 the Equifax IT security department ran scans to identify vulnerable system on a regular basis (Rasalam & Elson 2019; Smith 2017). But because of unacceptable deficits in their security systems, they were not able to detect their vulnerabilities (Ahmad & Barbacki 2019) and also failed to identify the unauthorized accesses until July 29, 2017 (Trope 2018). According to Jeff Luszcz "Development teams need to move beyond applying patches, and address vulnerabilities in processes. Taking preventive action keeps code safer, saving time and money while protecting reputations" (Luszcz 2018). Even if a firm will not follow Luszcz's attitude, the firm should at

least use verified or validated test methods to ensure exploring simple vulnerabilities like missing patches. Also, the employees must be trained significantly better, because that could prevent a data breach of this scope (Berghel 2017).

- For practitioners: Search for new test methods and software structured checklists to detect vulnerabilities. In addition, sensitize your employees to the relevance of installing necessary security software.
- For researchers: Check, how to improve existing testing models to detect security vulnerabilities.

Units of analysis	Steps of checklist									
Article/Papers	**Communication of bug failures**				**Testing and verification**			**Commercial focus and importance**		
Citation	How to reach all affected users?	Transparent communication with public	Publication of press release after careful investigation	If third-party software is used, customers must be informed of any errors/problems of the third-party software	New testing methods for software	Formal verification of software	Administrator training	Increase strategic importance of Cybersecurity in companies	Introduce explicit roles responsible for cybersecurity	Change funding of open source software
Ahmad and Barbacki 2019							x	x		
Berghel 2017							x	x		
Bottun 2018		x								
Lin 2018								x		
Luszcz 2018					x					x
Rasalam & Elson 2019					x					
Skedsvold 2017	x	x								
Tantleff 2017		x								
Trope 2018		x			x		x	x		

Table 10. Concept Matrix of Equifax Data Breach Literature

Communication of Bug/Failure

Equifax disclosure the data breach on September 7, 2017, five weeks after it was found (Skedsvold 2017) without an explanation for this delay and excluding their own negligence on the situation (Trope 2018). Although they had weeks to prepare, there was no communication strategy for the affected people which lead to a complete overstrain of telephone and website (Skedsvold 2017). Due to the delay, there had been a fifty-state class action lawsuit, because people were deprived of the opportunity to compensate for the impact as quickly as possible (Bottum 2018). To avoid such a situation in the future, the EU adopted the 'General Data Protection Regulation' in 2018, which stipulates that the injured party must be informed within 72 hours (Tantleff 2017).

- For practitioners: Check, whether the communication strategy fits the General Data Protection Regulation and reaches all affected user after a data breach.
- For researchers: Check, if another communication strategy might have lowered the commercial impact of the data breach.

Resulting Checklist

Using the deductions of the five ITSCs, we build a structured checklist which can be used to ex-post-analyse future ITSCs. It is necessary to specify the structure the approach should have. However, the order of steps by nature remains arbitrary and carries no meaning. Within the steps, the recommendations are sorted by the criterion 'expected ease of application'. This criterion consists of the expected time and resources necessary to follow the recommendation. However, these measures can differ between users, as different users have different abilities. Therefor the ordering of recommendations within the steps is to be seen as arbitrary.

We developed our approach according to the following requirements:

- General usability for ITSC with different root causes
- Not every step is relevant for every ITSC
- Ability to consider the user's time / resource restrictions
- Easily extendable for further ITSC
- Applicable for both practitioners and researchers

Structured Checklist for Analysing IT Security Catastrophes

In accordance with the requirements mentioned above the model of choice is a structured checklist divided in seven parts, one for each category of deduction. To address both target groups, all parts are divided into a practitioner and a researcher section. Within the sections for practitioners the deductions are arranged in ascending order according to the time and resources required to implement the recommendations. Within the researcher section the ordering of the checklist items is not time critical, but elective. The advantage for users is that the required parts of the checklist for the current ITSC can be selected and the rest skipped. Which section of the checklist is important depends on the origin of the case under consideration. To give a rough idea of how it might work, we will use the taxonomy introduced in "The IT Security Catastrophe Spectrum". to show which parts of the checklist might be the key sections for the four considered corners of the pentagon.

The consideration of an ITSC should focus on the origin of the ITSC and how to prevent a similar happening in the future. In the case of a human error the checklist parts 'applicability of prevention methods', 'commercial focus and importance' and 'user action' might be the most important checklist parts to consider. A raise of awareness or using a simple warning signal may have prevented an ITSC like the 'Amazon AWS Outage'. To review a malicious action the main attention should be with 'integrity' and 'testing and verification' so that after assuring the data integrity new tests and verification methods can be used to prevent further catastrophes. Is the source a 'third party failure' then 'testing and verification', 'communication of bug failure' and 'commercial focus and importance' should be considered as most important. Apart from prevention thorough testing or increasing the funding of open source software, it is necessary to communicate bugs to the users so that they can initiate measures. For 'System failure' we recommend a focus on 'integrity' and 'testing and verification' to assure the data security for further incidents and prevention through better gap detection.

Applicability of Prevention Methods

Applicability of prevention methods contains among others the question on why existing guidelines, prevention methodologies or others have not prevented the ITSC. Practitioners and researchers can apply this guideline item as follows:

- For practitioners:

1. Assure that the current user security instructions of your company cover the behavior that caused or enhanced the recent ITSC. In addition, review the instructions to ensure that they are understandable and complete.
2. If the regarded vulnerability was caused by human error or a misunderstanding of security instructions, you should consider whether the lessons learned can and should be formulated into security instructions for members of your organization.

- For researchers:
  1. With regards to the variety of applied IT security standards (see "IT Security Guidelines/ Standards") the question arises of what caused these standards to not prevent the most recent ITSC. Potential research hypothesis could include that instructions were either not sufficiently specific or applied incorrectly.

Commercial Focus and Importance

This item of the checklist focuses on the commercial dimension of an ITSC and can be applied by researchers and practitioners as follows:

- For practitioners:
  1. As demonstrated by Equifax a simple neglected security patch can force a CEO to resign and plummet the share price. Ensure that all employees are aware of the monetary risk that goes along with a weak cybersecurity and search for tools which coordinate the software actuality.
  2. Check the funding of your organization's IT security department. Although it is often stated that there is no glory in prevention, the proactive cost for a strong cybersecurity is lower than potential damage caused by an ITSC. To avoid this, also invest money or resources for testing of used open source code.
- For researchers:
  1. Especially in situations where the potential thread remains undetected for a long period of time, the question arises of why the developer community did not detect the issue earlier. A potential research hypothesis could be that a better funding of the software developers could have detected the issue earlier or even prevented it in first place. Possible influencing factors include tight software deadlines with insufficient staff capacity or a low budget for pricy security software licenses.
  2. From the ITSC cases analysed in this work the question arises if and why the IT security issues remain open for a longer than necessary time period. Potential areas of research include a comparison of existing software actuality coordination systems and conducting a quality appraisal.

Communication of Bug/Failure

This section collects analytic measures in the field of communication around the ITSC.

- For practitioners: Xerox and Equifax have led to a public discussion about the adequateness of the communication strategy of the respective inflicted parties, as described above. We suggest practitioners to assess whether the current catastrophe requires a change in the communication strategy of an organisation. For organizations operating within the European Union, we recommend aligning the communication strategy with the General Data Protection Regulation of the European Commission.
- For researchers: In the aftermath of Xerox and Equifax the abovementioned public discourse concluded that the respective company's communication was sub-optimal. The question arises if and to what extent a different communication strategy could lead to a more rapid spread of information causing a shorter duration and therewith a less negative impact of an ITSC.

Integrity

This checklist item covers the dimension of integrity, as defined above. Practitioners and researchers can apply this section as follows:

- For practitioners: In the aftermath of the AWS outage questions arose on whether the integrity of data was always assured during this downtime. As mentioned above, the research community used this incident as motivation to develop a variety of new methods to assure data integrity in complex IT systems. Since future ITSC might lead to similar topics research, we recommend practitioners to check, if there are new methods available for assuring data integrity.
- For researchers: After the AWS shutdown happened, the public questioned the integrity of Amazon's servers. Especially in the age of large, distributed IT systems and cloud architectures, this question gains importance. A potential topic of research could be to check, if existing system architecture models could be changed or altered or if the system could be enhanced by software to better cope with the risk of data loss or improper alteration.

Testing and Verification

This section includes analytic measures in the field of software testing and verification.

For practitioners: Check, whether researchers have introduced a new testing method to find issues like the recent catastrophe. In addition, we recommend sensitizing your employees to the relevance of installing necessary security software.

- For researchers:
  1. Many ITSC reveal that the applied testing methods have failed to detect an IT system's vulnerability. A short-term field of research could be to improve existing testing models to expose this specific weakness.
  2. In the longer run it should be considered why existing testing methods did not find the current issue and assessed if existing testing patterns could be modified or altered to find similar issues in the future.

User Interaction

User interaction collects recommendations on how to interact with users. It is especially relevant for practitioners.

- For practitioners:
  1. Some IT security catastrophes are caused by human error, as described above. We recommend practitioners to assess whether their current policy of access rights aligns with minimizing the risk of potential damage.
  2. Xerox and Heartbleed have revealed inconsistencies or ambiguities in user training material. We therefor recommend practitioners in the aftermath of a future ITSC to assess whether the respective catastrophe is unambiguously mapped in user training.

Law Enforcement

This section covers the legal dimension of an ITSC and can be applied by practitioners and researchers as follows:

- For practitioners: In the aftermath of the Yahoo data breach, a discourse about accountability arose in the US legal community. For future ITSC we therefor recommend practitioners to build an opinion on whether laws are enforced to punish certain behavior with regards to an ITSC (e.g., potential miscommunication).
- For researchers: If a future ITSC reveals behavior of one market participant, that is not in the spirit of the law, the question might arise, whether current laws are properly formulated to discourage unlawful behavior. A potential research topic could also be to assess whether the current case law complies with the meaning of the law.

Instantiation of Checklist with the WannaCry Ransomware Attack

Before the checklist can be instantiated on the example of WannaCry, we need to describe how WannaCry identifies as ITSC. WannaCry is commonly regarded as "the most successful ransomware to date in terms of its spread to the market." (Bundesamt für Sicherheit in der Informationstechnik 2019, translated by

author). Ransomware is a type of software that maliciously encrypts files or even whole computer systems and blocks user access until a specified ransom is paid. "Wannacry [sic.] attacked many hospitals, companies, universities and government organization across at least 150 universities, having more than 2,00,000 victims" (Mohurle & Patil 2017). The ransomware used a vulnerability in Microsoft's Server Message Block (SMB) protocol to encrypt user data into a .WCRY format and demands between $300-$600 USD of ransom for the decryption, to be paid in Bitcoin. (Kumar et al. 2018). In the ITSC spectrum, WannaCry covers both malicious actions and third-party failures. The first category emerges from the fact, that the United Stated, Canada and Australia hold North Korea as accountable for the ransomware attack (Bossert 2017) – therefor assuming a malicious intention. The categorization as third-party failure is due to the fact, that infected organizations relied on Microsoft Windows – which then turned out to be vulnerable.

The classification of WannaCry in the Cybersecurity Incident Taxonomy *(see Figure 8)*:

- Malicious Action: North Korea blamed for the attack.
- Third party failure: Vulnerability in Microsoft Windows that made other organizations hostage of the ransomware.

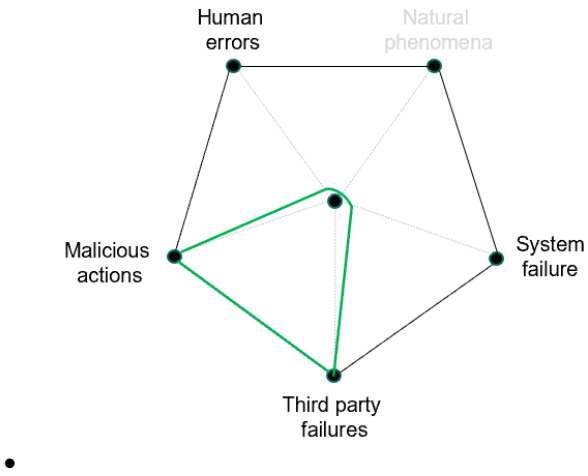

Figure 8. WannaCry Classification

The case of WannaCry therefor clearly defines as ITSC. Further we collect examples in the fields of research and practice that could have been inspired by our checklist. The focus thereby lies in visualizing the instantiation step required for seamless application of our approach; this means to decide which checklist item is relevant for the recent catastrophe. With the aim of exemplifying the necessary de-generalization step, the cited work is to be seen as incomplete and arbitrary list of examples with the focus on illustrating what an application of our checklist could look like.

Applicability of Prevention Methods

In the aftermath of WannaCry the UK department of health conducted a survey on the infected NHS instances and concluded "organisations [sic.] infected by WannaCry shared the same vulnerability and could have taken relatively simple action to protect themselves." (Department of Health 2018)
With regards to the variety of applied IT security standards (see "IT Security Guidelines/Standards") the question arises of what caused these standards to not prevent the most recent ITSC. Potential research hypothesis could include that instructions were either not sufficiently specific or applied incorrectly; to our knowledge there is no such work currently focusing on this matter in the context of WannaCry.

Commercial Focus and Importance

We consider the item of 'commercial focus and importance' as highly relevant for the analysis of the ITSC WannaCry since it is commonly regarded as "the most successful ransomware to date in terms of its spread

to the market." (Bundesamt für Sicherheit in der Informationstechnik 2019, translated by author). Both researchers and practitioners used the incident as inspiration for working on the topic of commercial focus and importance.

- What practitioners did: Castillo & Falzon (2018) used the catastrophe of WannaCry to investigate the impact of cyberattacks on stock market returns. They showed that "WannaCry had a positive effect on the equity returns of cybersecurity companies and cybersecurity investment vehicles" (Castillo & Falzon 2018). This again undermines the importance of cybersecurity.
- What researchers did:
  1. Hockey (2020) stated that "The research revealed that only 24% of those surveyed felt that cyber security budgets were at an adequate level – this is particularly surprising considering the publicity surrounding the WannaCry attack and the calls to action that occurred as a result." This media report we see as indication that researchers have used WannaCry as motivation to investigate funding and budgeting of IT security.
  2. The second part of our checklist item to our knowledge remains un-researched. Potential areas of further research in the context of WannaCry include a comparison of existing software actuality coordination systems and the identification of potential gaps.

Communication of Bug/Failure

- What practitioners did: WannaCry was successfully used as inspiration to study the effect of crisis-communication on company reputation (Mañas-Viniegra et al. 2019). During the WannaCry catastrophe, Telefonica transferred the role of the "traditional official spokesperson to the social networks of one of its directors." (Mañas-Viniegra et al. 2019).
- What researchers did: To our best knowledge there is no piece of research investigating the sensitivity of impact of catastrophes with regards to the style of communication; WannaCry could be a motivation for such research.

Integrity

- What researchers did: In the years after the WannaCry-attack, researchers have used this catastrophe as a motivation to "propose a more fundamental approach to protecting valuable files by applying the Moving Target Defense (MTD) concept." (Lee et al. 2019) Others have used the incident as inspiration to develop new data integrity verification systems for cloud storages (Apolinário et al. 2018). Due to the nature of ransomware, we consider the topic of integrity as important for the analysis of WannaCry.

Testing and Verification

We consider the checklist item of 'testing and verification' as highly important to the catastrophe of WannaCry. Since the ransomware used "a zero-day vulnerability that exist in Windows Server Message Block (SMB)" (Lee et al. 2019), the incident was used as a motivation for researchers to work on the topic of testing and verification. This can be seen an example of a potential application of our checklist.

- What researchers did:
  1. "[The] emerge of new ransomware families, such as WannaCry, showed that ransomware keeps evolving and cyber criminals are upgrading the ransomware code with more sophisticated features" (Akbanov et al. 2019). Researchers have used this continuous evolvement as motivation to improve existing testing models to expose or further investigate potential gateways for ransomware: For example, the methods of software defined networking (SDN) (Akbanov et al. 2019) and static/dynamic malware analysis (Jones & Shashidhar 2017) were slightly modified and used for detection and analysis of the WannaCry ransomware.
  2. In the years after WannaCry, an enhanced form of the control flow graph (CGF) representation was suggested to better understand the dynamic behaviour of ransomware (Nguyen et al. 2018).

User Interaction

To our best knowledge, WannaCry has not been used as an inspiration to work on the topic of user interaction.

Law Enforcement

To our best knowledge, WannaCry has not been used as an inspiration to work on the topic of law enforcement.

Discussion

Principle Findings

The central point of this work is to answer the original research question: How can security measures be deduced from use cases of IT security catastrophes? Therefore, we identified the most recent ITSCs until the Pentagon described in "The IT Security Catastrophe Spectrum" was covered apart from the excluded section. The abstraction of the measures identified to remedy the ITSC were combined to learn how to use the ITSC. After sorting the measures, they have been instantized and compiled in a checklist which enables us to give the most specific instructions possible for prevention. An essential part of our checklist is the usage in the time before necessary measures are implemented in official guideline documentations. It is therefore intended to provide practitioners with a list of potential gaps that should be closed immediately and to provide researchers with ideas to create new pointers for future guidelines and adjustments. Of course, the checklist is a dynamic element that will never be completed, as each new ITSC can add a new perspective. In line with the research question, the resulting checklist provides an approach that can already help practitioners and researchers to derive security measures for future ITSCs.

Contributions

To help learning from future ITSCs, the results indicate that a checklist providing an overview of past ITSC ex-post analytic considerations could help practitioners and researchers in developing practical recommendations in IT security. Issues in the implementation, management, or regulation of IT security, that favored the occurrence of an ITSC, are more likely to be reconsidered in future decisions, when provided with a checklist of considerations. Therefore, we can assume that our work helps learning lessons from past ITSCs. However, our contributions for academia and practice are very different.

Academia

For researchers, we contribute a potential for further development of an ITSC ex-post analysing checklist. With an ITSC ex-post analysing checklist, the complex concept of learning from an ITSC is broken down to understandable pieces of information and is easier to grasp. Thereby the checklist provides a foundation of what information is needed to learn from ITSCs.

First, the development of the checklist showed, that ITSC needs to be well defined. With a clear definition of ITSCs, an ITSC can be identified as such. The identification of an ITSC then helps to assess how pressing the topic of learning from this ITSC is. ITSCs are rather rare events, but often end in cost intensive effects. Further, by offering a view on what makes an IT security issue an ITSC, we contribute to the definition of ITSCs.

Second, our work shows, that in order to learn from mistakes in IT security, transparency is necessary. Transparency is essential, not only in practice but also in academia, in order to make sense of the complexity of IT security. The more an ITSC was covered by literature, the easier practical recommendations could be formulated for the improvement of the overall IT security. Heartbleed, that had been covered more thoroughly by literature, than Xerox, for instance, lead to a quicker deduction of practical recommendations.

Third, the checklist builds an overview of security issues, triggering ITSCs, reflected in academia. Based on this overview, it can be assessed how well academia covers the analysis of ITSCs. Some of our chosen ITSCs

have not been covered by academia at all (e.g., Xerox). Therefore, our work contributes to the awareness of ITSC research and of the need for more contributions in this topic area.

Practitioners

For practitioners, our results provide a systematic possibility to generate applicable IT security measures, after the occurrence of an ITSC. Companies, that have been affected by the ITSC may be in a hurry to learn from the ITSC, in order to quickly prevent similar happenings.

Further, our checklist requires the expertise of different areas of expertise (e.g., computer science, corporate management, legal department). The collaborative work in different areas of expertise can be administered to different departments in a company, which is simplified by the checklist. Learning from a recent ITSC then happens in parallel. In order to keep track of the progress of learning from a recent ITSC, in these different departments, the checklist provides the possibility to create an overview of the progress, by simply checking the points, that have been considered.

Limitations

Although our structured checklist provides many advantages, it is based on the analysis of 'catastrophes' which were categorized as such by their impact on different areas of interest, e.g., society, technology, industries and/or individual companies. By nature, this categorization is binary: An IT security incident either does or does not belong to the category 'catastrophe'. We do not further distinguish the magnitude and dimension of impact. Our results depend on the definition of 'catastrophe' and might slightly differ if IT incidents with a lower magnitude of impact were included in the deductions.

However, this binary categorization considers that even with different root triggers, incidents might have similar outcomes allowing for a generalizable ex-post treatment featured in our structured checklist. In the category 'active attacker' we consequentially do not distinguish between different motifs, intentions, knowledge backgrounds, group memberships or other attributes of the attacker or attack vectors. This reflects the implicit assumption that although these attributes might differ, the attacks can have a similar impact at this work's level of generalization.

Like all methods based on the analysis of historic information, our structured checklist works better, if input information remains within the historically manifested range. However, if for any reason, one or several input attributes change to values far outside the historically manifested range, our structured checklist might provide inconclusive or implausible results.

We selected ITSC according to their root trigger and reduced the spectrum to aspects covered by the definition of security. This led to a de-facto exclusion of the safety-relevant category of natural phenomena. However, there is no standardized definition of 'safety' and 'security' and interpretation differs among authors. It remains undetermined if and to what extent the addition of 'natural phenomena' to the definition of 'security' changes aspects of our structured checklist.

We conducted the selection process of ITSC until each of the potential root causes in our spectrum was covered at least once. The inclusion of additional ITSCs leaves room for future research, potentially broadening, updating, or enhancing our suggested approach.

The goal of our work was to provide researchers and practitioners with a hands-on approach to analysing ITSC after they have happened. The abovementioned selection of a limited number of ITSCs from the potentially unlimited set of catastrophes that have happened in the past, together with the methodology of literature identification (in contrast to a thorough literature review across all existing literature) and the grey-literature nature of this work distinguish our hands-on approach from a scientific method. Future contributions could be made by iteratively applying and re-enhancing the structured checklist, by including future literature, and by conducting a scientific review process.

The exemplary instantiation of the proposed structured checklist aims to exemplify the necessary re-generalization step, as described above. It is conducted with a single ITSC by the authors of this paper. Together with the selection of a historic ITSC, this introduces laboratory conditions. Practicality, generalizability, verifiability, and applicability – and therewith also validity – of our checklist remain to be proven and leave room for future research.

We conducted our deduction on literature published in English and German. We excluded all papers in other languages. It remains undetermined whether or to what extent the inclusion of further languages changes aspects of our structured checklist.

The literature identification is focused on the scientific database EBSCOhost. Therefor a piece of literature that EBSCOhost's search engine does not reveal, will not be used in our structured checklist. In the case of the Yahoo data breach and the Xerox bug this shortcoming of the database led to the necessity of using alternative sources of literature. In the process of further enhancing the structured checklist until it eventually becomes a method, future research can increase the number of databases used for the identification of literature.

Generalization of analytic steps such as we did in this work require a rather general formulation. This assures applicability on different types of IT security issues but makes it necessary for users to select which parts of the structured checklist are relevant for the most recent ITSC. Generic wording introduces a potential source of ambiguity and disputability since the application of our approach remains in the hands of the user.

Future Work

In the process of determining our checklist three unresolved questions respectively research fields have emerged which could possibly be resolved by future research. First, the checklist could be extended by examining ITSCs caused by natural catastrophes. This would lead to a complete coverage of the spectrum and would provide another possibility for the classification of ITSCs, thus potentially extending the checklist with new components. Second, the checklist can be further enhanced by an iterative application in which future literature is examined and a review process of the previous checklist is conducted.

Third, to cover the spectrum, only as many catastrophes were evaluated until the spectrum of the pentagon was completely covered. An analysis of further catastrophes can update, enhance, and extend the checklist. Furthermore, the checklist has been instantized with only one disaster, requiring a review with other catastrophes to create a valid scientific method. By further improving the checklist, the checklist may become a method that allows researchers to examine catastrophes from all perspectives and could possibly allow a faster and more effective recovery of future IT catastrophes.

Conclusion

As the last conclusion, after which we will leave the interested reader to his or her own impression, we state that our checklist by nature can never be complete. It is designed to grow, change, and evolve with every further application and therefor continuously adapts to changing environmental conditions. This allows for flexibility and enables a wide spectrum of potential applications. We are looking forward to seeing our work evolve and adapt further!

References

Ahmad, I., and Barbacki, K. 2019. "Recent Developments in Canadian Privacy Law and the Digital Charter," *Business Lawyer,* (75), pp. 1647–1654.

Akbanov, M., Vassilakis, V. G., and Logothetis, M. D. 2019. "Ransomware detection and mitigation using software-defined networking: The case of WannaCry," *Computers & Electrical Engineering*, (76:), pp. 111–121.

Aldemir, E., Gezer, N. S., Tohumoglu, G., Barış, M., Kavur, A. E., Dicle, O., and Selver, M. A. 2020. "Reversible 3D Compression of Segmented Medical Volumes: Usability Analysis for Teleradiology and Storage," *Medical Physics*, (47:4), pp. 1727–1737.

Alenezi, M., and Zarour, M. 2020. "On the Relationship between Software Complexity and Security," *International Journal of Software Engineering & Applications*, (11:1).

Allen, T. T., Sui, Z., and Parker, N. L. 2017. "Timely Decision Analysis Enabled by Efficient Social Media Modeling," *Decision Analysis*, (14:4), 250–260.

"Amazon AWS outage impacts dozens of online retailers" 2017. RetailCustomerExperience.Com, 1. (retrieved from: http://www.redibw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3 dbsu%26AN%3d122283161%26site%3dehost-live; last accessed: July 31, 2020)

Amazon Web Service 2017. "Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1)" (retrieved from: https://aws.amazon.com/de/message/41926/; last accessed: July 31, 2020).

Apica Inc. 2017. "How E-Commerce Sites Were Affected by the Amazon S3 Outage" (retrieved from: https://www.apicasystems.com/blog/top-100-e-commerce-sites-effected-amazons-s3-out-age/#:~:text=Newer%20websites%20are%20pulling%20data,servers%20when%20the%20outage%20happened; last accessed: July 31, 2020).

Apolinário, F., Pardal, M., and Correia, M. 2018. "S-Audit: Efficient Data Integrity Verification for-Cloud Storage," *IEEE International Conference On Trust, Security And Privacy In Computing And Communications,* No. 17.

Araujo, N. J. P., Pianto, D. M., and Ralha, C. G. 2019. "MULTS: A Multi-cloud Fault-tolerant Architecture to Manage Transient Servers in Cloud Computing," *Journal of Systems Architecture*, (101:101651).

Arora, V., Nawab, F., Agrawal, D., and Abbadi, A. E. 2018. "Janus: A Hybrid Scalable Multi-Representation Cloud Datastore," *IEEE Transactions on Knowledge & Data Engineering*, (30:4), pp. 689–702.

Attwood, A. I. 2020. "Changing Social Learning Theory Through Reliance on the Internet of Things and Artificial Intelligence," *Journal of Social Change*, (12:1).

Avizienis, A., Laprie, J., Randell, B., and Landwehr, C. 2004. "Basic Concepts and Taxonomy of-Dependable and Secure Computing," *IEEE Transactions on Dependable and Secure Computing*. (:1), pp. 1–14.

Beckmann, M., Scheiner, C. W., and Zeyen, A 2018. "Moral Disengagement in Social Media Generated Big Data," in *Lecture Notes in Computer Science*, G. Meiselwitz (Ed.), Vol. 10913, Social Computing and Social Media. User Experience and Behavior: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I (pp. 417–430), Cham: Springer International Publishing.

Berghel, H. 2017. "Equifax and the Latest Round of Identity Theft Roulette," *Computer*, (50:12), pp. 72–76.

Beurdouche, B., Bhargavan, K., Delignat-Lavaud, A., Fournet, C., Kohlweiss, M., Pironti, A., and Zinzindohoue, J. K. 2017. "A Messy State of the Union: Taming the Composite State Machines of TLS," *Communications of the ACM*, (60:2), pp. 99–107.

Blackwell, J. 2018. "Best Practices to Obtain and Maintain PCI Compliance," University of Oregon, Oregon, USA (retrieved from: https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/24352/Blackwell_2018.pdf?sequence=1&isAllowed=y; last accessed: July 31, 2020).

Blue, J., Furey, E., and Condell, J. 2017. "A Novel Approach for Secure Identity Authentication in Legacy Database Systems," *2017 28th Irish Signals and Systems Conference (ISSC),* pp. 1–6.

Bossert, T. P. 2017. "It's Official: North Korea Is Behind WannaCry," *The Wall Street Journal* (retrieved from: https://www.wsj.com/articles/its-official-north-korea-is-behind-wannacry-1513642537; last accessed: July 31, 2020).

Bottum, T. 2018. "Material Breach, Material Disclosure," *Minn. L. Rev.* (: 103), pp. 2095–2134.

Brill, H., and Jones, S. 2016. "Little Things and Big Challenges: Information Privacy and the Internet of Things," *American University Law Review,* (66:1183).

Bundesamt für Sicherheit in der Informationstechnik 2015. "BSI Technische Richtlinie 03138 Ersetzendes Scannen," *Bundesamt für Sicherheit in der Informationstechnik*.

Bundesamt für Sicherheit in der Informationstechnik 2018. "BSI Standard 200-1. (BSI-Standard)," *Bonn: Bundesamt für Sicherheit in der Informationstechnik*.

Bundesamt für Sicherheit in der Informationstechnik 2019. "Ransomware Bedrohungslage, Prävention & Reaktion 2019" (retrieved from: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Cyber-Sicherheit/Themen/Ransomware.pdf?__blob=publicationFile&v=6; last accessed: July 31, 2020).

Carvalho, M., DeMott, J., Ford, R., and Wheeler, D. A. 2014. "Heartbleed 101," *IEEE Security & Privacy*, (12:4), pp. 63–67.

Castillo, D., and Falzon, J. 2018. "An Analysis of the Impact of WannaCry Cyberattack on Cybersecurity Stock Returns," *Review of Economics & Finance*, pp. 93–100 (retrieved from: http://www.bapress.ca/ref/ref-article/1923-7529-2018-03-93-08.pdf; last accessed: July 31, 2020).

Chang, J. M. 2013. "Xerox Machines Change Documents After Scanning: Some of the company's Workcentre machines are altering numbers in documents" (retrieved from: https://abcnews.go.com/Technology/xerox-machines-change-documents-scanning/story?id=19895331; last accessed: July 31, 2020).

Chen, B., and Curtmola, R. 2017. "Remote Data Integrity Checking with Server-side Repair," *Journal of Computer Security*, (25:6), pp. 537–584.

Chen, T., Kuo, F., Ma, W., Susilo, W., Towey, D., Voas, J., and Zhou, Z. 2016. "Metamorphic Testing for Cybersecurity," *Computer*, (49:6), pp. 48–55.
Coldewey, D. 2013. "Copier Conundrum: Xerox Machines Swap Numbers During Scans" (available at https://www.cnbc.com/id/100945451; retrieved July 31, 2020).
Collis, H. 2013. "When a Copy is NOT a Copy: University Researcher Notices that Xerox Machine CHANGES Documents after scanning" (retrieved from: https://www.dailymail.co.uk/news/article-2386900/When-copy-NOT-copy-University-researcher-notices-Xerox-machine-changes-documents-scanning.html; last accessed: July 31, 2020).
Coy, P. 2013. "Some Xerox Scanners Can Alter Documents by Accident" (retrieved from: https://www.bloomberg.com/news/articles/2013-08-09/some-xerox-scanners-can-alter-documents-by-accident; last accessed July 31, 2020).
Department of Health 2018. "Investigation: WannaCry Cyber Attack and the NHS" (retrieved from: https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf; last accessed: July 31, 2020).
Durumeric, Z., Payer, M., Paxson, V., Kasten, J., Adrian, D., Halderman, J. A., and Beekman, J. 2014. "The Matter of Heartbleed," in Proceedings of the 2014 Conference on Internet Measurement Conference, Williamson, A. Akella, and N. Taft (Eds.), USA: ACM Press, pp. 475–488.
Edwards, B. P. 2018. "Cybersecurity Oversight Liability," *Georgia State University Law Review*, (35:663).
Directive 2009/140/EC, 2009 Official Journal of the European Union. 2009.
Regulation (EU) No 910/2014, Official Journal of the European Union. 2014.
Directive (EU) 2016/1148, Official Journal of the European Union. 2016.
Felizardo, K. R., Mendes, E., Kalinowski, M., Souza, É. F., and Vijaykumar, N. L. 2016. "Using Forward Snowballing to update Systematic Reviews in Software Engineering," in Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, Piscataway, NJ: IEEE, pp. 1–6.
Fliehe, M. 2014. "Kompass der IT-Sicherheitsstandards: Auszüge zum Thema Elektronische Identitäten" (retrieved from: https://www.bitkom.org/sites/default/files/file/import/140311-Kompass-der-IT-Sicherheitsstandards.pdf; last accessed: July 31, 2020).
Frank, B. H. 2017. "AWS Says a Typo Caused the Massive S3 Failure This Week," *CIO* (13284045:3).
Furey, E., and Blue, J. 2019. "Can I Trust Her? Intelligent Personal Assistants and GDPR," *2019 International Symposium on Networks, Computers and Communications (ISNCC)*, (2019), pp. 1–6.
Garber, L. 2014. "News Briefs," *Computer*, (47:6), pp. 12–17.
Gaudin, S. 2017. "AWS Blames a Typo for Tuesday's Outage," *CIO* (13284045:2).
Guy, B. 2017. "Analysis: Rethinking Cloud Architecture After the Outage of Amazon Web Services," (retrieved from: https://www.geekwire.com/2017/analysis-rethinking-cloud-architecture-outage-amazon-web-services/; last accessed: July 31, 2020).
Harzing, A. W. 2007. "Publish or Perish" (retrieved from: https://harzing.com/resources/publish-or-perish; last accessed: July 31, 2020).
Hockey, A. 2020. "Uncovering the Cyber Security Challenges in Healthcare," *Network Security*, (2020:4), pp. 18–19.
International Organization for Standardization 2016. ISO/IEC Directives Part 1. (SO/IEC DIR 1:2016-05(en)). Switzerland: IEC Central Office.
Ivanova, D., Borovska, P., and Zahov, S. 2018. "Development of PaaS Using AWS and Terraform for Medical Imaging Analytics," *AIP Conference Proceedings*, (2048:1).
Jackson, D. W. 2014. "Cybersecurity: Breaches and Heartbleed to BYOD-- Are Bankers, Entertainment Company Executives, Celebrities, Postal Workers, Ice Cream Lovers, Home Builders, and ClOs the Only Ones Who Should Be Concerned?" *Law Library Journal*, (106:4), pp. 633–643.
Jaeyoung, D., Sengupta, S., and Swanson, S. 2019. "Programmable Solid-State Storage in Future Cloud Datacenters," *Communications of the ACM*, (62:6), pp. 54–62.
Jalkanen, J. 2019. "Is Human the Weakest Link in Information Security?: Systematic Literature Review", *University of Jyväskylä* (retrieved from: https://jyx.jyu.fi/handle/123456789/64186; last accessed: July 31, 2020).
Jehle, C. 2018. "Xerox-Software Verändert Eingescannte Zahlen" (retrieved from: https://www.heise.de/tp/features/Xerox-Software-veraendert-eingescannte-Zahlen-3961586.html; last accessed: July 31, 2020).

Jones, J., and Shashidhar, N. 2017. "Ransomware Analysis and Defense: WannaCry and the Win32 environment," *International Journal of Information Security Science*, (6:4), pp. 57–69.
Kamp, P. 2014. "Quality of Software Costs Money--Heartbleed Was Free," *Communications of the ACM*, (57:8), pp. 49–51.
Kang, J., and Park, J. H. 2017. "A Secure-Coding and Vulnerability Check System Based on Smart-Fuzzing and Exploit," *Neurocomputing*, (256:), pp. 23–34.
Kelion, L. 2013. "Xerox to Update Scan Software after Switched Number Outcry," (retrieved from: https://www.bbc.com/news/technology-23610405; last accessed: July 31, 2020).
Kersten, H. 2020. "IT-Sicherheitsmanagement nach der neuen ISO 27001: ISMS, Risiken, Kennziffern, Controls (2., aktualisierte Auflage). Edition <kes>.
King, D. 2017. "Yahoo! Inc.: Corporate Profile Part 1: Buyer Profile" (retrieved from: http://www.infosecgirl.net/uploads/3/4/9/6/34965713/yahoo_profile.pdf; last accessed: July 31, 2020).
Kriesel, D. 2013a. "Möglicher Workaround für Zeichenersetzungen in Xerox Scankopierern" (available at http://www.dkriesel.com/blog/2013/0806_work_around_for_character_substitutions_in_xerox_machines; retrieved July 31, 2020).
Kriesel, D. 2013b. "Xerox Scanners/Photocopiers Randomly Alter Numbers in Scanned Documents" (retrieved from: http://www.dkriesel.com/en/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning; last accessed: July 31, 2020).
Kriesel, D. 2013c. "Xerox-Scankopierer Verändern Geschriebene Zahlen" (retrieved from: http://www.dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning?redirect=1; last accessed: July 31, 2020).
Kriesel, D. 2013d. "Telefonkonferenz mit Xerox" (retrieved from: http://www.dkriesel.com/blog/2013/0806_conference_call_with_xerox; last accessed: July 31, 2020).
Kriesel, D. 2015. "BSI überarbeitet Richtlinie RESISCAN, verbietet JBIG2" (retrieved from: http://www.dkriesel.com/blog/2015/0317_bsi_verbietet_jbig2?s=bsi; last accessed: July 31, 2020).
Kupsch, J. A., and Miller, B. P. 2014. "Why Do Software Assurance Tools Have Problems Finding Bugs Like Heartbleed?"
Lee, S., Kim, H. K., and Kim, K. 2019. "Ransomware Protection Using the Moving Target Defense Perspective," *Computers & Electrical Engineering*, (78:), pp. 288–299.
Li, Z., Romano, P., and van Roy, P. 2020. "Transparent Speculation in Geo-Eeplicated Transactional Data Stores," *Journal of Parallel & Distributed Computing*, (143:), pp. 129–147.
Lin, H. 2018. "Once More unto the Breach: Recent thefts of credit data show how little power consumers have over their own information. This has to change," *Hoover Digest: Research & Opinion on Public Policy*, pp. 127–130.
Luszcz, J. 2018. "Apache Struts 2: How Technical and Development Gaps Caused the Equifax Breach," *Network Security*, (2018:1), pp. 5–8.
Mañas-Viniegra, L., Niño González, J. I., and Martínez Martínez, L. 2019. "Transparency as a Reputational Variable of the Crisis Communication in the Media Context of Wannacry Cyberattack," *la Transparencia Como Variable Reputacional de la Comunicación de Crisis en el Contexto Mediático del Ciberataque Wannacry*, (:48), pp. 149–171.
Marquess, S. 2014. "Speeds and Feeds of Money, Responsibility, and Pride" (retrieved from: https://veridicalsystems.com/blog/of-money-responsibility-and-pride/; last accessed: July 31, 2020).
Mathew, A. R. 2019. "Cyber Security through Blockchain Technology," *International Journal of Engineering and Advanced Technology*, (9:1), pp. 3821–3824.
McMillan, M. 2015. "The Cost of IT security," *Healthcare Financial Management*, (69:4), pp. 44–47.
McMillan, R. 2017. "Yahoo Triples Estimate of Breached Accounts to 3 Billion" (retrieved from: https://www.wsj.com/articles/yahoo-triples-estimate-of-breached-accounts-to-3-billion-1507062804; last accessed: July 31, 2020).
Merriam-Webster. 2020. "Definition of Security" (retrieved from: https://www.merriam-webster.com/dictionary/security; last accessed: July 31, 2020).
Mohurle, S., and Patil, M. 2017. "A Brief Study of Wannacry Threat: Ransomware Attack 2017," *International Journal of Advanced Research in Computer Science*, (8:), pp. 1938–1940.
Nas, S. 2015. "The Definitions of Safety and Security," *Journal of ETA Maritime Science*, (3:2), pp. 53–54.
National Institute of Standards and Technology 2018. "Framework for Improving Critical Infrastructure Cybersecurity," Version 1.1.

National Institute of Standards and Technology 2019. "Evolution of the Framework" (retrieved from: https://www.nist.gov/cyberframework/evolution; last accessed: July 31, 2020)

Nguyen, M. H., Le Nguyen, D., Nguyen, X. M., and Quan, T. T. 2018. "Auto-Detection of Sophisticated Malware Using Lazy-Binding Control Flow Graph and Deep Learning," *Computers & Security*, (76:), pp. 128–155.

NIS Cooperation Group. 2018. "Cybersecurity Incident Taxonomy" (retrieved from: https://ec.europa.eu/information_society/newsroom/image/document/2018-30/cybersecurity_incident_taxonomy_00CD828C-F851-AFC4-0B1B416696B5F710_53646.pdf; last accessed: July 31, 2020).

Ogle, J. 2019. "Identities Lost: Enacting Federal Law Mandating Disclosure & Notice after a Data Security Breach," *Arkansas Law Review*, (72:), pp. 221.

Ormerod, P. C. 2019. "A Private Enforcement Remedy for Information Misuse," *Boston College Law Review*, (60:), pp. 1893.

Pabrai, N., Keller, J., Lin, J., Hupa, A., and Bacchus, A. 2020. "Vulnerability Reward Program: 2019 Year in Review" (retrieved from: https://security.googleblog.com/2020/01/vulnerability-reward-program-2019-year.html; last accessed: July 31, 2020).

Pearce, A. 2018. "Can Companies Disclaim and Limit Liability for Data Breaches in Online Terms of Service?" *Journal of Internet Law*, pp. 3–6.

Primoff, W., and Kess, S. 2017. "The Equifax Data Breach: What CPAs and Firms Need to Know Now," *The CPA Journal*, pp. 14–17.

Rasalam, J., and Elson, R. J. 2019. "Cybersecurity And Management's Ethical Responsibilities: The Case Of Equifax And Uber," *Global Journal of Business Pedagogy Volume*, (3:3), pp. 8–15.

Rash, W. 2017. "AWS Outage Demonstrates Need for Redundancy Even in the Cloud," *EWeek*, pp. 1.

Rife, R. H.B.L. 2019. "Improving Information Security Awareness Training Through Real-Time Simulation Augmentation," *Northcentral University, La Jolla, California, USA* (retrieved from: https://search.proquest.com/openview/f50c6ff238c7b134b8b4be517ca02b5b/1?pq-origsite=gscholar&cbl=18750&diss=y).

Rodríguez-Díaz, M. A., and Sánchez-Cruz, H. 2014. "Refined Fixed Double Pass Binary Object Classification for Document Image Compression," *Digital Signal Processing*, (30:), pp. 114–130.

Saran, C. 2017. "Amazon Web Services Outage Shows Vulnerability of Cloud Disaster Recovery," *Computer Weekly*, pp. 4–6.

Satheesh Kumar, M., Ben-Othman, J., and Srinivasagan, K. G. 2018. "An Investigation on Wannacry Ransomware and its Detection," *in 2018 IEEE Symposium on Computers and Communications (ISCC),* Piscataway, NJ: IEEE*,* pp. 1–6.

Scarfone, K. 2009. "Cyber Security Standards" (retrieved from: https://tsapps.nist.gov/publicatison/get_pdf.cfm?pub_id=152153; last accessed: July 31, 2020).

Schönbohm, A. 2020. "Kurzprofil BSI: Deutschland * Digital * Sicher * BSI," (retrieved from: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Presse/BSI-Kurzprofil.pdf?__blob=publicationFile&v=8; last accessed: July 31, 2020).

Scott, B. 2018. "How a Zero Trust Approach can Help to Secure your AWS Environment," *Network Security*, (2018:3), pp. 5–8.

Shiah, C.-Y., and Yen, Y.-S. 2013. "Compression of Chinese Document Images by Complex Shape Matching," *The Computer Journal*, (56:11), pp. 1292–1304.

Shin, Y., and Williams, L. 2008. "Is complexity really the enemy of software security?" *in Proceedings of the 4th ACM workshop on Quality of protection,* A. Ozment & K. Stølen (Eds.), New York, NY: ACM, pp. 47.

Skedsvold, M. C. 2017. "A Duty to Safeguard: Data Breach Litigation Through a Quasi-Bailment Lens," *J. Intell. Prop. L.* (:25), pp. 201–226.

Smith, R. F. 2017. "Prepared testimony of Richard F. Smith before the US house committee on energy and commerce subcommittee on digital commerce and consumer protection," *US House of Representatives*, pp. 1–8.

Tantleff, A. K. 2017. "Equifax Breach Affects 143M," *Journal of Health Care Compliance,* (Sep/Oct2017, Vol. 19), pp. 45–46.

Thielman, S. 2016. "Yahoo Hack: 1bn Accounts Compromised by Biggest Data Breach in History," *The Guardian* (retrieved from: https://www.theguardian.com/technology/2016/dec/14/yahoo-hack-security-of-one-billion-accounts-breached; last accessed: July 31, 2020).

Trautman, L. J. 2017. "Corporate Directors' and Officers' Cybersecurity Standard of Care: The Yahoo Data Breach," *SSRN Electronic Journal*, (66:1231).
Trautman, L. J. 2018. "How Google Perceives Customer Privacy, Cyber, E-Commerce, Political and Regulatory Compliance Risks," *William and Mary Business Law Review*, (10:), pp. 1.
Trautman, L. J., and Ford, J. 2018. "Nonprofit Governance: The Basics," *Akron Law Review*, (52:), pp. 971.
Trautman, L. J., Hussein, M. T., Ngamassi, L., & Molesky, M. J. 2020. "Governance of the Internet of Things (IoT)" (retrieved from: https://arxiv.org/pdf/2004.03765; last accessed: July 31, 2020).
Trautman, L. J., and Ormerod, P. C. 2017a. "Industrial Cyber Vulnerabilities: Lessons from Stuxnet and the Internet of Things," *University of Miami Law Review*, (72:), pp. 761.
Trautman, L. J., and Ormerod, P. C. 2018. "Wannacry, Ransomware, and the Emerging Threat to Corporations," *Tennessee Law Review*, (86:), pp. 503.
Trope, R. L. 2018. "When Incident Response Goes Awry: Cybersecurity Developments," *Business Lawyer*, (:74), pp. 229–241.
Vassilev, A., and Celi, C. 2014. "Avoiding Cyberspace Catastrophes through Smarter Testing," *Computer*, (47:10), pp. 102–106.
Wang, Y., Li, G., Ma, M., He, F., Song, Z., Zhang, W., and Wu, C. 2018. "GT-WGS: an Efficient and Economic Tool for Large-Scale WGS Analyses Based on the AWS Cloud Service," *BMC Genomics*, (19:), pp. 89-N.PAG.
Weise, E. 2017a. "Amazon Server Blackout Wreaks Havoc on Websites," *USA Today* (retrieved from: http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dasn%26AN%3dJ0E066392195117%26site%3dehost-live; last accessed: July 31, 2020).
Weise, E. 2017b. "Ugh! Typo to Blame for Server Blackout," *USA Today* (retrieved from: http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dasn%26AN%3dJ0E309800732017%26site%3dehost-live; last accessed: July 31, 2020).
Weiss, N. E. 2016. "The Yahoo! Data Breach—Issues for Congress," *CRS Insight* (retrieved from: https://fas.org/sgp/crs/misc/IN10586.pdf; last accessed: July 31, 2020).
Wheeler, D. A. 2014. "Preventing Heartbleed," *Computer*, (47:8), pp. 80–83.
Williams, M. 2017. "Inside the Russian Hack of Yahoo: How They Did It" (retrieved from: https://www.csoonline.com/article/3180762/inside-the-russian-hack-of-yahoo-how-they-did-it.html; last accessed: July 31, 2020).
Wynne, J. 2019. "Impact of Data Breach and Privacy Concerns on Organisation's Performance," *University of Liverpool, Liverpool, England*.
Xerox 2013a. „Xerox Scan-Problem: Was Sie wissen müssen: Fragen und Antworten" [Press release], (retrieved from: https://www.xerox.com/assets/pdf/ScanningQAincludingAppendixA-de.pdf; last accessed: July 31, 2020).
Xerox 2013b. „Xerox schaltet Zahlendreher-Funktion ab" (retrieved from: https://www.spiegel.de/netzwelt/gadgets/xerox-zahlendreher-funktion-bei-multifunktionsdruckern-abgestellt-a-918238.html; last accessed: July 31, 2020).
Yadron, D. 2014. "Massive OpenSSL Bug 'Heartbleed' Threatens Sensitive Data," *The Wall Street Journal* (retrieved from: https://www.wsj.com/articles/web-encryption-tool-is-flawed-researchers-say-1396986692; last accessed: July 31, 2020).
Zhang, L., Choffnes, D., Dumitras, T., Levin, D., Mislove, A., Schulman, A., and Wilson, C. 2018. "Analysis of SSL Certificate Reissues and Revocations in the Wake of Heartbleed," *Communications of the ACM*, (61:3), pp. 109–116.

A Deep Dive of the GAIA-X Project: Analysis of the Major Opportunities & Challenges

Critical Information Infrastructures, Winter Term 20/21

Miguel Andre Bänfer
Master Student
Karlsruhe Institute of Technology
uwgcg@student.kit.edu

Lauritz Bühler
Master Student
Karlsruhe Institute of Technology
uhewf@student.kit.edu

Luise Möller
Master Student
Karlsruhe Institute of Technology
uqebf@student.kit.edu

Amélie Svensson
Master Student
Karlsruhe Institute of Technology
udesh@student.kit.edu

Abstract

Background: In recent years, data has become an increasing relevant resource and source of competition for companies. However, the data market is currently dominated by US companies, limiting the independence of European companies. Therefore, the project GAIA-X was introduced, to create a federated, open data infrastructure based on common European values. As this constitutes a large project, involving multiple businesses, research institutes and governments, difficulties can arise which could possibly hinder the realization of the project.

Objective: To aid in the project's success, the following paper analyzes key opportunities and challenges of GAIA-X, to help maintain clarity over its goals and understand challenges faced.

Methods: Several interviews with project members from different backgrounds were conducted. Open Coding was then used to interpret the given insights and identify the opportunities and challenges mentioned.

Results: The control, sovereignty and protection of data, as well as interoperability were seen as vital opportunities of the project. These would aid in the creation of data markets, foster innovation and lead to more independence of large service providers. On the other hand, the project still faces significant challenges and risks. Especially the coordination of the various project members poses a challenge, but also cyber security concerns and potential adoption issues by the market are mentioned.

Conclusion: GAIA-X bears many opportunities that have the potential to create major impact for companies regarding their data initiatives, but also faces some significant challenges which have to be overcome before the project can reach its full potential. To help aid tracking and overcoming these challenges as well as prevent them from becoming risks, a risk management team should be created as a central authority on all subject related matters.

Keywords: critical information infrastructure, data sovereignty, european cloud infrastructure, federated data infrastructure, federated services, GAIA-X, open-source initiative

Introduction

In the recent years, the use of IT services has increased significantly, especially in the corporate context, thereby creating a new area, in which data has become a commodity for organizations, governments and individuals. The data, which has become ever more present in everyday life, has become a central resource and is handled within data ecosystems using so called data infrastructures (Oliveira & Lóscio, 2018; Kitchin, 2014). As data becomes an increasing relevant parameter of competition, helping with the development of Artificial Intelligence (AI), Machine Learning and furthering digitalization, making it available plays an essential role (European Parliament, 2020). However, the market leaders in cloud computing, software development, search engines and social networks are mainly non-European companies, restricting the independence of European companies (European Parliament, 2020). Here, the EU sees a risk regarding the stored data of European companies and their competitiveness (BMWi, 2020b). In addition, currently high lock-in effects exist, limiting data portability and interoperability, which are seen as essential for data sharing (European Parliament, 2020).

Due to the current situation of the market and data being an essential part of competitiveness, there is a strong incentive to build a "high performance, competitive, secure and trustworthy data infrastructure for Europe" (BMWi & BMBF, 2019). To this end, the project GAIA-X was introduced as a "federated, open data infrastructure based on European values" (BMWi & BMBF, 2019). Its core objective is to increase industrial competitiveness and to reduce European dependence from non-European companies by creating an "open, federated, secure and trustworthy data and cloud infrastructure for Europa as the basis for a digital ecosystem" (BMWi, 2020b). GAIA-X, which will be structured into regional hubs, is building a network of an increasing number of member states across Europe (BMWi, 2020a), with many larger and smaller corporations already participating or interested in joining (BMWi, n.d.).

The unprecedented number of companies and member states joining the initiative in the early stages sent a clear positive signal regarding the importance of the project and its potential impact. But this could also pose a threat, as multiple participants could potentially mean diversified interests in the project goals and therefore the focus it should have. In addition, a larger project can face multiple challenges ranging from strategic over technical to user acceptance. This work thus is trying to answer the research question: *What are the main opportunities and challenges for GAIA-X?* This allows to understand and avoid (or at least mitigate) the faced challenges, thereby helping with a smoother realization of the project. In addition, understanding the main opportunities that members see in the project, can help maintain clarity of the goals that are supposed to be achieved. Following this line of thought, this paper therefore aims to analyze and deliver a structured overview, as well as a comparison of the main challenges and opportunities, by conducting qualitative research in the form of semi-structured interviews with project members and industry experts. In this paper, opportunities are referred to outcomes or properties that potentially add value for users, industry and other market players resulting from GAIA-X. Challenges on the other hand describe potential obstacles, problems and highly complex tasks. Furthermore, risks describe challenges that could lead to the failure of the project under certain circumstances.

To thoroughly answer our research question, the paper is structured into the following chapters *Background*, *Method*, *Results* and *Conclusion*. In *Background,* the GAIA-X project will be presented and defined, highlighting its objectives as well as its structure. Following this, in *Method* our research approach will be introduced within three separate sub-chapters. Firstly, we present how we proceeded for our interview acquisition, secondly, we describe the interview preparation and execution process and lastly, we provide an insight on how we analyzed the interviews using coding methods. In the chapter *Results* the main findings of the interviews are presented. This three-parted chapter introduces and describes the main opportunities and challenges, which are mentioned by the interview partners. In addition, it further analyzes our findings and highlights the importance and interrelationships between the mentioned opportunities and challenges. The final chapter *Conclusion* summarizes the principal findings and introduces the implications for practice & research. As GAIA-X is in the middle of the development, the implications focus mainly on suggestions to aid in the successful development of the project. In addition, limitations of the research will be discussed and an outlook on future research is given.

Background

Definition of GAIA-X

GAIA-X is a European project that started in Autumn 2019. To foster its establishment the international, non-profit "GAIA-X foundation AISBL" has been founded in Belgium (BMWi 2020a). The project aims at building a "high-performance, competitive, secure and trustworthy data infrastructure based on European values" (BMWi & BMBF 2019). These European values are set as seven guiding principles (BMWi 2020b).:

1. European data protection
2. Openness and transparency
3. Authenticity and trust
4. Digital sovereignty and self-determination
5. Free market access and European value creation
6. Modularity and interoperability
7. User-friendliness

GAIA-X is seen as a step towards a federated data infrastructure that creates a digital ecosystem for Europe. In that digital ecosystem data can be stored and shared between participants to spur innovation while digital sovereignty is ensured (BMWi & BMBF 2019). To implement the data infrastructure and to create the digital ecosystem an architecture of standards will be developed. The architecture of standards defines regulatory, industry-specific and technical standards (BMWi 2020c).

Objectives of GAIA-X

Europe's current digital infrastructure is based on a few major non-European companies that also offer a larger variety of digital services than European companies do. The objective of GAIA-X is to reduce European dependence from non-European companies and to increase industrial competitiveness at the same time. GAIA-X aims to achieve "an open, federated, secure and trustworthy data and cloud infrastructure for Europe as the basis for a digital ecosystem" (BMWi 2020b). Because a digital ecosystem is based on shared data, the objective is to create a data sharing architecture, an EU federation of cloud infrastructure, related infrastructure and data services (BMWi, 2020d). This objective goes along with the European Data Strategy that aims to create one single market for data, which is still open to international data (European Commission, 2020). To make cloud services more attractive for European companies and to enhance trust in these services, GAIA-X is based on the seven principles mentioned beforehand. Additionally, the GAIA-X network can be extended by edge-components. By sharing and using data together data pools are created, which can be used to develop new ideas and business models by science and companies (BMWi & BMBF 2019). Through an open data approach (BMWi & BMBF 2019) and common standards, the data infrastructure improves the exchange in and between domains. Inter- and intra-domain data exchange might facilitate the development of advanced smart services, like artificial intelligence as a service (BMWi 2020e). Several domain-specific use cases tailored to their market players have been compiled. Users might choose services corresponding to their individual needs (BMWi 2020e).

Structure of GAIA-X

GAIA-X is organized in hubs that build a network between an increasing number of member states. The widespread distribution of members across Europe corresponds to the diverse regional economic structures (BMWi, 2020a). The digital infrastructure is structured in three architecture levels. It includes the network level (hardware and data-transfer networks), the data level (data storages incl. software) and the service level (data processors, services, functions, applications). The data and service level together form the data infrastructure (BMWi & BMBF 2019).

Figure 1 shows a high-level overview of the architecture of GAIA-X. The digital ecosystem delivered by GAIA-X can be differentiated in an infrastructure ecosystem and a data ecosystem. In the center of both ecosystems are the four domains of federation services. They create the basis for an interoperable and legally secure connection between providers and users. These federation services include "federated identity and access management, a federated catalogue with a directory of all providers and services, specifications and solutions for the sovereign exchange of data, as well as certification and compliance" (BMWi 2020a).

Thus, they ensure that participants comply with the guiding principles (BMWi 2020d). The infrastructure ecosystem provides infrastructure components to store, transfer and process data as well as services (BMWi 2020d).

The main assets of the infrastructure ecosystem are called "Nodes". These Nodes build a network and, further, can be structured in hierarchies. Each one represents a computational resource, e. g., hardware, a data center, or infrastructure operation services (BMWi 2020d). In the data ecosystem, the main asset constitutes the data, which is made available by a data provider, owned by a data owner, and received by a data consumer. The data owner decides to make data available for other participants or to keep it private. Data can be exchanged or traded by checking self-description regarding protection and other restrictions. The self-description includes information about the owner, usage policies, provenance details, technical descriptions, content related descriptions, accompanied by a statement of proof. Further details, like legal aspects, can also be included (BMWi 2020d). Any participant who registers for GAIA-X needs to accept its terms and conditions to ensure each participant's self-determination over their data (BMWi 2020d).

GAIA-X declared commitment to neutrality. For this reason, any new and any established provider that satisfies GAIA-X terms, conditions and guiding principles can become a Node in the network (BMWi & BMBF 2019). GAIA-X will be an open ecosystem based on existing solutions. Therefore, non-European companies that provide essential data and analysis infrastructure can also collaborate (BMWi 2020b).

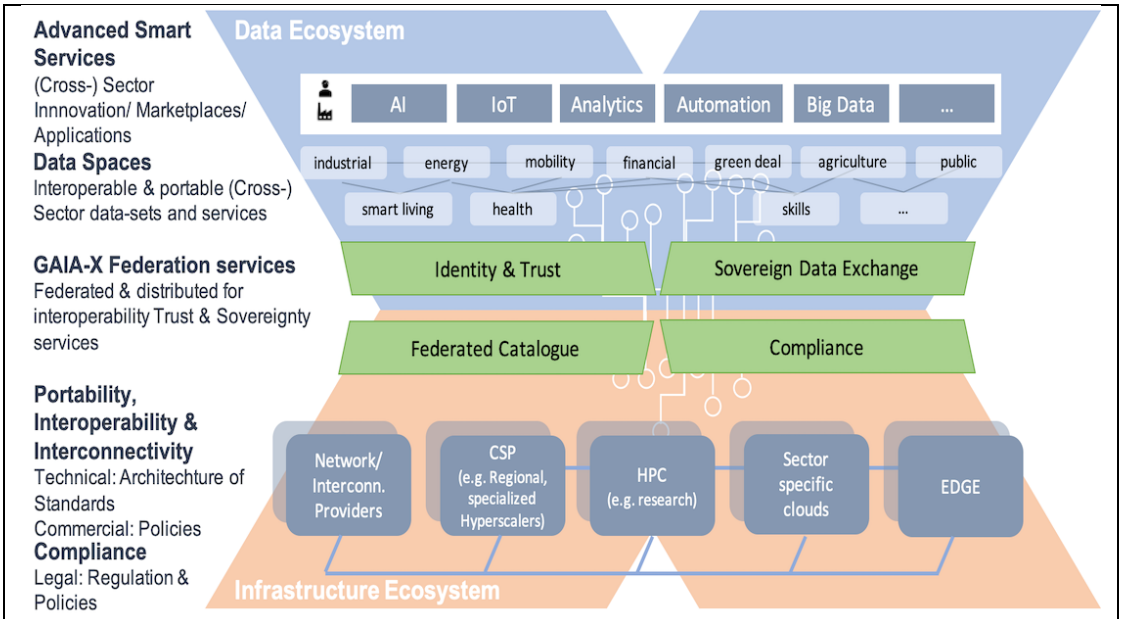

Figure 1. High-Level Overview of the GAIA-X Architecture (BMWi, 2020a)

Methods

Interview Acquisition

For the acquisition of our interview partners, we first reflected on the areas from which we wanted to interview people. It was important to us that we put together a balanced pool of partners from the different areas of the GAIA-X development (industry, research, politics) as well as future users of GAIA-X.

Using the webpage of the GAIA-X summit on November 18th and 19th (GAIA-X Summit 2020), we identified possible experts for our interviews. We received an overwhelming response from the contacted parties, conducting a total of seven interviews with nine interview partners.

In the following Table 1 we have summarized the pseudonymized demographic data of our interview partners. As one can see, we interviewed people from both the development and user side, who work at medium-sized or large organizations and are involved in the project in various ways.

Interview Execution

To prepare the expert interviews, we developed an interview guideline which consists of general questions complemented by more specific queries adapted to the experts' sectoral background and respective role in the project.

Speaker	**Background**	**Size of the Organization**
No. 1	German Goverment	1750 employees
No. 2	Cloud Service Provider	> 200.000 employees
No. 3 and No. 4	Association	200 employees
No. 5 and No. 6	Research Insititute	532 members
No. 7	Applied Research Insititute	239 employees
No. 8 (written interview)	Applied Research Insititute	95 employees
No. 9	French Association	1 million members

Table 1. Demographic Data

This guideline therefore is divided into two parts: The first set of questions investigates usual aspects such as the interviewee's professional experience and other demographic data. After briefly introducing our research aim, the interview continues with a second set of questions depending on the experts' sectoral background. Differentiated according to the interviewees' involvement in GAIA-X ("computer scientists" vs. "corporate user" perspective), we prepared two different question pools containing more detailed queries. This second part of our interview guideline is structured into several key words, for instance IT-security, interoperability, dependance and user expertise. Each keyword comprises various questions examining a specific area of interest. Our approach therefore follows the concept of a semi-structured interview. This offers the opportunity to dynamically include new questions which might arise while conducting the interviews (Myers, 2009). By structuring our research interests with the help of this guideline, we ensure a continuous flow of the interview. Furthermore, utilizing a prepared question pool ensures covering our main areas of interest and facilitates our coding analysis by making the experts' insights comparable.

To guarantee an academic approach and enquire all relevant aspects, we decided to conduct the interviews in pairs of two. One person was responsible for the introduction and formalities whereas the other one conducted the interview and asked the main questions. Furthermore, due to scheduling conflicts we had to change the format of one interview (No. 8) from video call to a written questioning.

Interview Analysis

After transcribing the interviews, we analyzed them using the Open Coding method. The interviews had an average length of 1 hour with some of them slightly longer (80min) and others shorter (33min). The Open Coding method allowed us to identify the most important concepts and thus the main issues of opportunities and challenges. Each section was considered one by one and summarized with the help of a generic term (a code). The constant comparison of similarities and differences between existing codes and the respective section helped to find connections (Myers, 2009). In order to avoid keeping the codes from becoming too general, we have refined codes by adding more detailed names in some cases, e.g., *Challenge_Interoperabilitaet* and *Challenge_Interoperabilitaet_Entwicklungsumgebung*.

A total of 96 codes were used, which can be divided into 6 clusters:

1. Demographic data (5 codes)
2. Project GAIA-X (3 codes)
3. Motivation and definition of GAIA-X (8 codes)
4. Opportunities and biggest potential (20 codes)
5. Challenges and potential risks (28 and 10 codes)

6. Miscellaneous (22 codes)

Based on these clusters as well as the individual codes, we were able to formulate the most important concepts and connections for our further analysis of challenges and opportunities. Our code book has 20 codes for opportunities of the GAIA-X project with a total of 145 hits. Further, it contains 48 codes referring to challenges with 124 hits in total. Out of these challenges we identified 10 codes as potential risks with 23 hits.

For a detailed analysis, we examined our different clusters: The first cluster, demographic data, comprises generic information regarding the experts and their related organization. It is subdivided into 5 categories such as personal and corporate information or the interviewee's connection to GAIA-X. To further categorize the acquired knowledge related to GAIA-X, we introduced the clusters project GAIA-X and motivation and definition of GAIA-X. These two categories consist of various individual codes which contain project-related information as well as visions, its planned roadmap and industry-specific use cases. The next cluster named opportunities and biggest potential displays subcodes regarding data related topics, interoperability of the GAIA-X approach, potential innovations, especially for Small and Medium Enterprises (SME's), and positive responses driving the projects progress. The fifth cluster challenges and potential risks is two-parted. The codes for challenges include e.g., project specification, hyperscalers, market acceptance, interoperability or data availability. Closely connected to this, we summarized other codes depicting risks. The identified subcategories for risks describe cyber security, data protection, costs, a slow development and dependency on cloud service providers. The last cluster called miscellaneous includes all codes which could not be assigned clearly or would represent a very small cluster representing too many details. Some exemplary codes attributed to this category are legal foundations or user expertise.

Results

Opportunities

The opportunities largely coincide with the seven guiding principles listed in the project's position papers (see Background). Some of the opportunities were mentioned frequently and, in all interviews, while others only occurred in some interviews. The codes within the cluster of opportunities were grouped thematically and result in nine key opportunities as listed in *Table 2*.

The assurance of *data control, sovereignty and protection* is of great importance. Currently, the market leaders in IT services are mainly non-European, which means that European standards for data security and ownership are not guaranteed. GAIA-X is intended to counteract this situation by introducing integrity with regard to data: Services offered on GAIA-X will comply with the EU requirements for data protection, data sovereignty and data control. Thus, the immense amount of data that is produced nowadays, can be stored securely. The EU is striving for data sovereignty. This implies that users of GAIA-X services can be confident that their data is secure and data misuse is prevented. Users also have control over their data, which means "as the provider of a data set, [they] […] always see who, how, when and where the data is being processed" (Speaker 1). Thus, it could enable consuming IT services as "cloud services in areas where it is not possible today from a regulatory point of view", e.g., in the financial sector or in health care management (Speaker 1).

One of the biggest benefits GAIA-X will bring is the development of a *data market*. The large mass of data produced only becomes of added value when it is shared and can be used properly (Speaker 9). Through an organized data market, SMEs that produce data can "attach a kind of price tag to it, so that you can trade data, so that you can provide companies that need data" (Speaker 1) with targeted data such as developers of AI. Through the aspect of data control, companies can decide exactly what data they offer and to whom. Access to data from a wide range of sectors and the cross-sectoral use of this data will give rise to new business models, innovations and services that we cannot yet imagine (Speaker 4).

Another central advantage of GAIA-X is the *interoperability* function. Interoperability in this case means that information can be exchanged easily via interfaces and common standards. It will have two major benefits. The interface function will enable a data market. Thanks to the data security, protection, and control features, the market will be organized and secure. The transfer of information to other services and ecosystems will allow users to be more flexible. Thus, users will have the flexibility to switch service providers when needed, scale services, or distribute their workload across multiple services. This gives users

the advantage of gaining independence, flexibility and avoiding a lock-in effect (Speaker 5, Speaker 9). Furthermore, the interoperability feature and the open market strategy in Europe mean that service providers are competing with each other to a greater extent than before, and that the importance of single service providers is decreasing. This should lead to a dynamic market in which services develop and a market that "massively cushions the shockwaves that such a thing [the loss of a service provider] could cause, because it is possible to migrate capacities that are available and booked there without any problems" (Speaker 5).

Opportunities	**Description**
Data control, sovereignty, protection	In contrast to the services provided by non-European hyperscalers, data protection, sovereignty and control are to be assured in GAIA-X based on European standards.
Data market	Users should be able to offer data on an organized data market and receive financial compensation in return or be able to procure data.
Interoperability	Common standards and technical interfaces will enable an interoperability function allowing the exchange of data.
Innovation and flexibility	New business models such as advanced smart services can be developed, and users achieve greater flexibility and independence.
Transparency	GAIA-X services aim to be more transparent than existing IT services, so that processes can be retraced.
Cyber security	GAIA-X is expected to be secure against cyber-attacks despite its status as a future digital center.
Promotion of SMEs	European SMEs are to be supported in the areas of digitalization, independence and data exchange which will foster innovation.
Momentum	GAIA-X is currently receiving a lot of positive reactions from politicians, developers and future users, and this should be seized as an opportunity.
Project management approach	The large number of project members with different backgrounds provides the necessary diversity of expertise for an IT project of this size.

Table 11. Overview of Opportunities

Another central advantage of GAIA-X is the *interoperability* function. Interoperability in this case means that information can be exchanged easily via interfaces and common standards. It will have two major benefits. The interface function will enable a data market. Thanks to the data security, protection, and control features, the market will be organized and secure. The transfer of information to other services and ecosystems will allow users to be more flexible. Thus, users will have the flexibility to switch service providers when needed, scale services, or distribute their workload across multiple services. This gives users the advantage of gaining independence, flexibility and avoiding a lock-in effect (Speaker 5, Speaker 9). Furthermore, the interoperability feature and the open market strategy in Europe mean that service providers are competing with each other to a greater extent than before, and that the importance of single service providers is decreasing. This should lead to a dynamic market in which services develop and a market that "massively cushions the shockwaves that such a thing [the loss of a service provider] could cause, because it is possible to migrate capacities that are available and booked there without any problems" (Speaker 5).

The interoperability function paired with the data market could promote creativity, *innovation and flexibility* in the European market. The new gained flexibility could help develop new businesses and business models. For example, companies with limited resources such as start-ups have the possibility to pick exactly the needed services from the GAIA-X catalogue in the beginning and "can migrate directly to any new server environments without the whole fuss, because I [they] know immediately who can do it following the same technical requirements" (Speaker 3) while still being able to scale or change their services later on in case of growth and development. In addition, the advantage of common standards and interfaces could also be used in advancing digitalization in more conservative sectors (Speaker 5). Speaker 5 sees that "especially in the energy and healthcare sectors […] the potential of GAIA-X is infinite, because these are primarily perceived as conservative industries where nothing has happened for a long, long time". The ultimate vision is the development of advanced smart services through connecting data or adapting

algorithms across sectors. One example occurred in the analysis of data from agricultural and environmental field, which could then be coupled with health data to find more efficient correlations (Speaker 5).

Another important advantage of GAIA-X is the *transparency* of processes for its users. The transparency refers on the one hand to the services offered and on the other hand to the service providers themselves. It is intended to lead to an increase in confidence in GAIA-X on the side of the users. They ought to be informed about the different ways of data usage and to be able to comprehend it. The open-source approach should make it possible to verify how internal processes work, so that users are not facing a kind of "black box" when they decide to use a service (Speaker 2). Speaker 7 suggests a certification mechanism comparable to eco-labels for food, which could increase the comprehensibility of the level of compliance to GAIA-X standards. This would allow users to decide which data should be protected through higher standards and for which - e.g., public data - existing standards are sufficient (Speaker 1).

Cyber security plays an important role in the development of GAIA-X which aims to be designed to be secure from cyber-attacks. Its users can trust the data infrastructure and be confident that their data will be safe from attacks. Presumably, GAIA-X will be located at the digital center, making it attractive to cyber-attacks. However, a "security by design" approach is being applied, meaning that the security aspect is being incorporated into the concept phase from the "very beginning of the development" to make the infrastructure as secure as possible (Speaker 3). The interviewed project members were confident that the experts involved will develop the appropriate security mechanisms to ensure that future users will be able to use GAIA-X without any concerns.

"The focus in the strategies of GAIA-X is on *small and medium-sized enterprises*" *(SMEs)* and their promotion (Speaker 2). Regarding the user side, GAIA-X brings three major advantages to SMEs. Firstly, digitalization in companies is to be simplified and accelerated, as currently many smaller companies do not yet use clouds (Speaker 1). Secondly, the interoperability function gives them "much greater power, also over the cloud providers" and makes them more independent (Speaker 1). The third major advantage for SMEs is the data market, through which they can generate additional profit as data providers or procure data from a wider area (Speaker 5). From the service provider perspective, GAIA-X represents for "smaller special cloud providers […] an interface, […] a technology […] so that they can interconnect" there and offer their services (Speaker 1).

In general, GAIA-X is perceived very positively by the industry (both involved partner companies and future users) and by policy makers, "within a year there has been a *momentum* for GAIA-X and also an encouragement […] [like no other] in the last 10 years for any IT topic or even digital topic" (Speaker 3). Politicians, future users, and "the industry really have a need for […] a data infrastructure" like GAIA-X to keep the European market competitive (Speaker 1). On the one hand, GAIA-X aims to fully utilize the potential of produced and existing data masses and to obtain a higher data protection standard. On the other hand, it will also help users to gain more independence from service providers and, as a result, to strengthen the European economy. The political support is reflected for example in the fast creation of the GAIA-X association (AISBL). The potential for the industry is reflected in the constantly growing number of interested companies to become part of the project. This momentum should be used to introduce the first parts of GAIA-X to future users in a timely manner and to maintain trust and acceptance.

Despite *project management* challenges, which are discussed more detailed in the next section, the way the project is set up and driven holds opportunities. The involvement of industry partners, developers, policy makers and other experts with a wide range of backgrounds and from different countries brings a great deal of expertise in many areas to the project (Speaker 9). Furthermore, "future users can already have a say in how the open data infrastructure should look and be used" (Speaker 8) as they are involved from the beginning on. Furthermore, they are helping to identify needs as well as requirements. Use cases are introduced and discussed to model and test as realistically as possible. In contrast to the hyperscalers, GAIA-X is driven more by a top-down approach, which makes it possible to agree on the standard requirements followed by incorporating and structuring them from the beginning on. Thus, a greater diversity of service providers can be achieved than from the bottom-up approach as seen in the US where "individual players who, simply because they are the first, then absolutely dominate the market" (Speaker 7).

Challenges

As in the opportunity section, there are frequently and rarely occurring challenges. Challenges arise in all stages of the project.

Challenges	**Description**
Coordination of project members	Only a few working groups, but a lot of project members with different interests might lead to underemployment and disagreements.
Stagnating development	Stagnation can be caused by "coordination of the members" and might lead to decreasing trust and acceptance, eventually to participants leaving the project.
Specification of GAIA-X	Difficulties in finding unanimous specifications of GAIA-X can be experienced, due to disagreements or particular-interests among the project members themselves.
Creating interoperability	To create interoperability interfaces, standards and functions that ensure inter- and intra-domain exchange of data need to be defined.
Data security	The data security concept is in the initial development phase. Market players see difficulties in sharing data because of competitive advantages.
Demand and fulfilment of expectations	Market acceptance is required. Expectations cannot be defined ultimately due to GAIA-X facing a dynamic market environment.
Dealing with hyperscalers	Hyperscalers are vital to GAIA-X, but could also turn into main competitors.

Table 3. Overview of Challenges

The challenge mentioned the most refers to the *coordination of project members* due to a high number of participating parties. The project started with a few project members and has grown very fast. Currently, there are 181 participating companies and institutions in several countries, mainly in Germany and France. The high number of participants make project management and decision-making processes more complex (Speaker 5). "If every enterprise brings a handful of employees, then we talk about 500 to 900 workers. If they all want to be included in the [six] working groups, it will be organizationally difficult to keep such a group working and to bring results. There would be too much friction" (Speaker 2). Due to the high number of participants, project members that are willing to work on GAIA-X, might not be actively engaged within the six existing working groups in the technical committee. A 'first come, first served' approach might not satisfy all participants and lead them to think "Thanks, but no thanks. I will leave" (Speaker 2).

Coordinating the high number of participants is a challenge on the higher management levels leading to inconsistent expectations. For example, France and Germany are both developing individual federation services on their own (Speaker 4). Although all members aim at building an open, transparent data infrastructure, there are various particular-interests because individual participants have different expectations how they might benefit from GAIA-X. Additionally, for participating companies the voluntary contribution to GAIA-X might conflict with their annual financial statement (Speaker 5). Further, there are disagreements considering the content of the frequently published position papers. Some co-authors want it as marketing material, others prefer a strategic paper and yet others aim to include more technological details. The task is to steer a middle course between defining terms specifically and setting a framework that everyone agrees with (Speaker 2). Finding the middle ground between different positions might lead to uncertainties, thus for some project members the meaning of the federation service "sovereign data exchange" is still ambiguous (Speaker 4).

The challenge of a *slow development* is another representation of the coordination challenge of project members caused by time-consuming disagreements or contradicting particular-interests among the participants themselves. A slowing development of the project might lead to the risk of a decreasing momentum, as well as the *decrease of trust and acceptance in GAIA-X*. This could either lead to participants reducing their contribution to GAIA-X or to participants leaving the project (Speaker 2).

The second most mentioned challenge is *creating interoperability* in the GAIA-X ecosystem. This challenge is multilayered. It begins by defining which interfaces, which standards and which functions need to be included to create yet non-existing, domain-overreaching interoperability (Speaker 1). Instead of creating new standards, it is planned to build on already existing standards. "At first, existing standards and new

offers have to be identified. Then, universal standards can be defined or a network that can connect multiple standards can be developed" (Speaker 5). Thereby, the groundwork done by the "International Data Spaces" incentive, that created standards for a secure and sovereign data exchange, might be used (Speaker 8). Additionally, to ensure interoperability, not only software but also hardware needs to be considered. Especially, the bandwidth in the GAIA-X network needs to be sufficiently large. It might be increased dynamically to handle the data streams within the technological framework (Speaker 3) and to keep latency as low as possible (Speaker 1). Furthermore, a certification mechanism that guarantees interoperability needs to be defined and established. One idea is to have different certificates describing the rank of interoperability for each service. They might be interpretated as "You have now achieved a Bronze interoperability level, or a Silver interoperability level, or a Gold interoperability level." (Speaker 1). But there are still unanswered questions: "Who will do that? Will a trusted third party commissioned by the GAIA-X AISBL or by the individual enterprises do it as a certification authority?" (Speaker 2).

Data security, one of the key aspects of GAIA-X, is considered a challenge on both the developer's and on the user's side. On the developer's side, "there are discussions about how something like this can really be implemented. […] For me the whole thing is still not so clear. We are still relatively at the beginning" (Speaker 4). On the user's side, "the fear of disclosing central data and business secrets when exchanging data across companies is a decisive reason for companies not to do so" (Speaker 8). Companies might disclose sensitive information including their competitive advantages or how to maintain them that competitors might be interested in. An enterprise might interpret that as "support one's own competitors in order to better compete against oneself" (Speaker 2). In the case of data security violations companies require rules for liabilities. "In particular, the complex requirements for data protection, such as those in terms of the General Data Protection Regulation, pose challenges for companies when exchanging data across companies and thus form an obstacle that must be reduced through uniform regulations with regard to the handling of data and with the help of controlled access to data ecosystems" (Speaker 8).

Another challenge is seen in the *demand* of GAIA-X once the project is rolled out on the market. To achieve significant demand market acceptance is required and "acceptance means performance" (Speaker 1). But GAIA-X will be a living construct that adapts to requirements of its users. Therefore, the services that need to be provided when GAIA-X is launched, cannot be identified definitively today (Speaker 5). Moreover, GAIA-X needs to be well known by potential users when it goes to market. Some group members perceive current marketing measures as insufficient and would prefer to have it organized more professionally. "I think marketing is also a bit of a step-motherly issue at GAIA-X, because it's assumed that government and corporate support will be enough to push the issue" (Speaker 5).

Further difficulties could arise from the *hyperscalers*, that are strongly incorporated and considered a vital part of GAIA-X "because without the services of Amazon, Google, Microsoft […] the whole thing won't work. We will still depend on them. […] They also have a very, very own interest in influencing GAIA-X and the question is whether they always play by the rules. We will see" (Speaker 5). That issue might be addressed in terms of governance and compliance. "One must assume that they will make two offers" (Speaker 5). Due to their market position, they might offer GAIA-X-compliant services but also cheaper equivalent services, without ensuring the user's data control, simultaneously (Speaker 5).

Discussion

Many of the aspects from the previous sub-chapters are closely connected and hold the potential for opportunities as well as challenges.

When it comes to *cyber security*, we observed that most interview partners actively involved in the project did not see cyber-attacks as a major challenge. However, our interview partner with strong technical background which was not part of the project seemed more concerned about potential attacks once GAIA-X is in the digital center of attention (Speaker 7). On the other hand, the involved interview partners argued that the matter of cyber security is incorporated from the beginning on in the concept phase to guarantee security (Speaker 3). On the technological side, the most modern standards and possibilities of IT-security will be implemented (Speaker 5). "I think a lot is being done to reduce the risks. But I would not presume that everything is always fully eliminated. That would be a bit too naïve." (Speaker 2).

Another point is the excitement and importance attached to the projects' outcome. One should take advantage of the current *momentum* and the positive response of project members, authorities, industry

and future users in order to progress quickly with the project and make use of the current market openness. If the project proceeds slowly, the enthusiasm might decrease. A slow development may entice participants to leave the project due to frustration. An option to keep the momentum high could be to develop and launch a test version of GAIA-X within the current year that provides certified services. Speaker 2 suggests to "take specific use cases and pre-implement them together with some companies in interaction as a mini-ecosystem". Considering that major investments are being made in cloud technology this year, the importance of a test version increases (Speaker 8). In order to meet users' expectations, users should be strongly incorporated within the development of the test version. As the project progresses, involving users ensures that their needs are actively addressed. If they are not involved, services that are no longer relevant to customers, as they are also in a dynamic market environment, might be developed.

Ultimately, GAIA-X must deliver its value proposition. If GAIA-X cannot ensure interoperability, data control and secure data exchange, it will not be established. Due to their value proposition, GAIA-X services might be more expensive than services outside the GAIA-X catalogue from established providers. For that reason, the *customer's willingness to pay* for secure digital services needs to be analyzed (Speaker 7). One interesting question for a market analysis might be "how many people would be willing to pay 5€ per month for an internet search engine that does not use one's data for other purposes?" (Speaker 7). Besides economic aspects, trust is considered as a key to the market acceptance of GAIA-X. In addition to technological measures regulatory measures can also be considered to make it as trustworthy as possible. On the legal side, trust can be increased by a legal framework that asserts certain rights (Speaker 7).

The fulfillment of the expected opportunities is challenged by its *technical realization*. The technical measures for data security, and the definition of common standards and interfaces for an interoperability function are crucial. This should enable data security, sovereignty, control, interoperability, independence for users, and establish a data market.

For the creation of such data markets the market players need to share their data. However, in the interviews the refusal to share data was discussed. Although the project members, which partly include potential users, know that for GAIA-X *willingness to share data* is vital, they see difficulties in sharing their data because they are afraid to disclose corporation trade secrets or other sensitive information (Speaker 8). However, competition might grow out of cooperation in the form of shared data (Speaker 3). The participants rather discuss about legal rights of use and licenses instead of comprehending GAIA-X as a big cooperation to spur innovation. If the opportunity of cooperation is not understood by corporations, and "if it's always about what's in it for me and not what's it in it for us, then the project might fail" (Speaker 3).

From a *project management* point of view, the involvement of a wide range of experts with different professional and cultural backgrounds brings the expertise needed for such a large IT-project (Speaker 9). It is crucial to get the input from developers, authorities, future service providers and potential users. However as mentioned earlier, the coordination of a large number of group members but also contradicting particular-interests might hinder the development of the project.

Besides, the correlations between opportunities and challenges, *interdependencies between the individual opportunities* can also be stated, meaning one opportunity makes the other possible. For example, to enable a functioning data market where data is traded based on European standards, data control must be with the data owners. They should be able to determine who has which access to which of their data. The data market serves as a basis for further opportunities, such as the promotion of innovation, of creativity or of existing small and medium-sized enterprises. Innovative processes and services such as advanced smart services can be developed mainly through access to a wide range of data. This is also due to the function of interoperability. It enables the use of data in different services and from different areas to create links. For example, by merging data from different sectors synergy effects can be used, such as applying analytical methods from medicine to quality management in industry (Speaker 1).

In general, all interview partners mentioned similar opportunities and challenges. It was noteworthy that the importance of challenges varied in the interviews. This can probably be traced back to two factors. On the one hand, many of our interview partners had different backgrounds (technical or business management). On the other hand, they are involved in different areas of the project, deal with other daily challenges and thus have a different perspective. The enthusiasm for GAIA-X was also reflected in the share of opportunities and challenges mentioned. Clearly challenges were less emphasized than opportunities during the interviews. This could be due to the project being in its starting phase.

Conclusion

Principal Findings

GAIA-X bears many opportunities that have the potential to create major impact for companies regarding their data initiatives, but also faces some significant challenges that have to be overcome before the project can reach its full potential.

Sharing data between companies can lead to the development of new business models and foster innovation within European companies. This is where GAIA-X promises significant opportunities over the current deployed solutions by focusing on interoperability as well as control of data and secure data exchange as key elements of the data infrastructure. This will also lead to higher independence of smaller enterprises from large service providers and foster competition. Given the potential of GAIA-X, it is unsurprising that the project is currently met by great enthusiasm from many different parties and therefore is experiencing a great momentum that pushes the project forward.

This seen enthusiasm might decrease quickly, due to stagnating development or differences in the vision of the future of GAIA-X, which could lead to serious project risks and companies even potentially leaving the project all together. Additional risks arise if the project cannot fulfill its promises, especially regarding interoperability and cyber security. It is also unclear if the project will be able to reach its full potential due to the fact that companies might have a mindset issue, hereby focusing more on the competition that could arise then on the potential positive impact newly formed cooperations could have. Further, the willingness of companies, as profit-oriented entities, to pay for these services are not yet assessed, leading potentially to a reduced market adoption if the benefits do not outweigh the costs. In general, it is important to keep in mind that challenges can potentially turn into risks if not addressed properly.

Implications for Practice and Research

To ensure GAIA-X' smooth development, challenges and risks should be monitored frequently. It would also be wise to create and implement concrete steps to mitigate challenges early on, to prevent them from becoming risks to the project in the future. For this, it would be helpful to create a risk analysis team that can function as a central authority for all related topics. As it is also key to keep the companies' focus on the positive side of the project, the team could aid in highlighting the key opportunities and help to clarify the vision of GAIA-X.

To overcome some of the more dominant challenges mentioned by the interview partners, we propose to create a clear Go-To-Market strategy that is communicated with all project members; especially testing functions early on with businesses could improve integration and acceptance later on. Furthermore, the customer's willingness to pay for services should be analyzed to understand the companies "cost-benefit" ratio and to facilitate the planning process.

Limitations

We have presented many interesting findings in the previous sections that have demonstrated a vast variety of challenges and opportunities regarding the GAIA-X project, but when reading and interpreting the given information, there are some limitations that need to be considered.

Even though we believe we were able to discover many of the given opportunities and challenges faced by the GAIA-X project, there could still be aspects of the project that we were unable to highlight and explore fully. This is mainly due to the limited number of *interview partners*. In addition, our interview partners were mainly German (with one French interview partner) as well as people who may be positively biased towards the project, as they represent stakeholders or project contributors. The limited number of interviews and the selection of the partners may have influenced our result to focus more on the opportunities, as well as to focus mainly on the opportunities and challenges faced by the German hub, companies and developers. Therefore, international insights might have been limited.

The research is conducted in a very *early phase of the GAIA-X project*, therefore scientific literature is limited. This limits the amount and variety of sources that could be consulted, meaning that the research is only based on a few key authors. In addition, due to the early phase of the project, the interview partners

may feel excitement that the project is moving forward and therefore focus on the positive aspects. Further, some challenges could be overlooked as they have not yet been materialized. This holds true especially for the user side of the story (regarding acceptance, necessary qualifications, and user experience), due to the fact that to this day no prototype exists and therefore the questions were discussed on a hypothetical basis.

As the research was conducted by students, the time was restricted to a semester and therefore limiting a more in-depth research of the presented topic area. Through this, an in-depth research with multiple interview partners and different approaches to the topic area was not possible. Due to the *time constraints*, we were only able to focus on qualitative research (interviews), limiting the insights regarding the impact and importance of the challenges and opportunities found.

Outlook Future Research

We believe our research serves as a basis for future investigations along the GAIA-X project and the opportunities and challenges faced. To gather more insights and complement the findings, further research is necessary. Regarding the limitations it would make sense to expand the interview base, by not only conducting more interviews but also trying to capture an even broader variety of members involved in the project; mainly interviewing partners from different participating countries and also people not directly involved in the project. This would not only give insights to differences in national challenges and opportunities but also display the differences between people working on the project and people who have a certain distance to GAIA-X. This could illuminate why certain companies are (not) willing to join the GAIA-X project and which challenges and opportunities they see as the dominant reasons. In addition, a quantitative study could be conducted, and therefore giving insights into the importance of the found opportunities and challenges, also eliminating biases.

References

BMWi, and BMBF. 2019. “Project GAIA-X: A Federated Data Infrastructure as the Cradle of a Vibrant European Ecosystem; Executive Summary,” *data-infrastructure.eu* (retrieved from: https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/das-projekt-gaia-x-executive-summary.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. 2020a. “GAIA-X - the European project kicks off the next phase,” *BMWi.de* (retrieved from: https://www.bmwi.de/Redaktion/EN/Publikationen/gaia-x-the-european-project-kicks-of-the-next-phase.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. 2020b. “GAIA-X: A Pitch Towards Europe,” *BMWi.de* (retrieved from: https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-a-pitch-towards-europe.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. 2020c. “GAIA-X: Policy Rules and Architecture of Standards,” *BMWi.de* (retrieved from: https://www.bmwi.de/Redaktion/EN/Publikationen/gaia-x-policy-rules-and-architecture-of-standards.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. 2020d. “GAIA-X: Technical Architecture,” *BMWi.de* (retrieved from: https://www.bmwi.de/Redaktion/EN/Publikationen/gaia-x-technical-architecture.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. 2020e. “GAIA-X: Driver of digital innovation in Europe,” *BMWi.de* (retrieved from: https://www.bmwi.de/Redaktion/EN/Publikationen/gaia-x-driver-of-digital-innovation-in-europe.html; last access January 6, 2021; last accessed: July 31, 2020).

BMWi. (n.d.). “FAQs on the GAIA-X project,” *BMWi.de* (retrieved from: https://www.bmwi.de/Redaktion/EN/FAQ/Data-Infrastructure/faq-projekt-gaia-x.html; last access January 6, 2021; last accessed: July 31, 2020).

Council of the European Union. 2008. “Council directive 2008/114/EC on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection,” *Official Journal of the European Union* (345), pp. 75–82.

European Commission. 2020. “Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions; A European strategy for data,” Brussels: EUR-Lex.

European Parliament. 2020. “Is data the new oil? Competition issues in the digital economy,” (retrieved from:

https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/646117/EPRS_BRI%282020%29; last access January 10, 2021; last accessed: July 31, 2020).
"GAIA-X Summit." 2020. (retrieved from: https://events.talque.com/gaia-x-summit/en/6iq6yI5LPSxaIRA6cmnq; last access December 20, 2020; last accessed: July 31, 2020).
Kitchin, R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE Publications Ltd.
Mell, P., and Grance, T. 2011. rep., *The NIST Definition of Cloud Computing*.
Myers, M. D. 2009. "*Qualitative Research in Business & Management,"* London: Sage.
Oliveira, M. I., and Lóscio, B. F. 2018. "What is a data ecosystem?" in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (doi: 10.1145/3209281.3209335).
Sunyaev, A. 2020. "*Internet Computing: Principles of Distributed Systems and Emerging Internet-Based Technologies"* (1st ed.), Cham, Switzerland: Springer.

Evaluating the Overlap of User Preferences with Search Engine Performances using UCISE

Critical Information Infrastructures, Winter Term 20/21

Raphael Abt
Master Student
Karlsruhe Institute of Technology
abtraphael@gmail.com

Jonathan Haigis
Master Student
Karlsruhe Institute of Technology
jonathan.haigis@student.kit.edu

Christina Stappen
Master Student
Karlsruhe Institute of Technology
christina.stappen@gmx.de

Abstract

Background: The assessment of information systems such as Search Engines (SEs) from the users' perspective needs to assure that it does not introduce a bias towards the systems' technical implementation. Instead, the socio-technical nature of SEs and a users' contextual evaluation based on previous experiences should be at the center of attention.

Objective: Therefore, we introduce a methodological approach for a User Centered Information System Evaluation (UCISE) and apply it to a SE evaluation. This approach allows for comparing selected SEs based on a derived set of user-based evaluation criteria.

Method: Our methodology consists of four steps, including a criteria synthesis, a criteria importance assessment by a large-scale survey and an expert assessment of selected SE performances regarding the user-based criteria. In the final step, we visualize the fit between the users' criteria preferences and the assessed SE performances.

Results: The first step of our methodological approach enabled the synthesis of a set of ten criteria for the SE differentiation from user-perspective. Of these ten criteria, only the three criteria 'search result quality', 'ease of use' and 'privacy protection' emerged as being perceived as 'extremely important' or 'important' by the user. The assessment of selected SEs with regard to all ten criteria shows significant differences between the providers, which lead to a very differentiated picture in the visualization of the fit of providers to user preferences.

Conclusion: Our research indicates that seven out of ten criteria are not decisive for a user's SE selection. The 'search result quality' emerges as the only criterion being perceived as 'extremely important' for a user's SE evaluation, followed by 'ease of use' and 'privacy protection'. When assessing the fits of three selected SE providers to the derived criteria, Google largely outperforms its competitors. This appears to be realized by overcompensating its major weakness in 'privacy protection' by superior results in 'search result quality' and 'ease of use' where the other assessed providers noticeably lack behind.

Keywords: user perspective, user-centered, search engine, information system, evaluation, assessment, Google, DuckDuckGo

Introduction

Within today's internet-based society, the World Wide Web has become one of the most important and highly valued services for information exchange and retrieval (Sunyaev 2020). Due to the web's crucial role for society and its vast amount of nearly two billion available websites (Statista 2019), quickly differentiating information of value from irrelevant information is of high importance. This is when SEs come into play. They enable users to filter the wealth of available information for content of interest. Consequently, it is not surprising that, together with the web's evolution, the global search engine market has evolved into a multi-billion-dollar market (Argenton and Prüfer 2012).

Upon closer inspection of this market, one can notice a significant imbalance of market shares. Handling a worldwide amount of more than 71% of all desktop and more than 91% of all mobile queries, *Google* is by far the globally leading player (Statista 2020). *Googles* strongest competitors are *Bing* and *Baidu*, each with approximately 13% of all desktop and with 1% or respectively 6% off all mobile queries (Statista 2020).

This concentration of market power among only a few providers may entail the risk of adverse consequences. One example is consumer dependency. With *Google* holding the lion's share of search queries, the majority of users depend on a single provider for being directed to the content of interest. Economists claim that such monopolistic markets can cause economic inefficiencies if providers abuse their market power for profit maximization (Varian 2016). In SE markets this is done by special relations with merchants as sponsoring third parties, where prices are raised when competition among merchants increases. Thereby big players like *Google* can circumvent natural market mechanisms and exploit their monopolistic position (Clemons and Wilson 2016). Another aspect is *Googles'* algorithmic editing of search results based on individual preferences which are known from previous web searches (Weinberg 2020). Considering that the visual appearance of search results heavily influences users' behavior (Pan et al. 2007), this distortion of displayed results again raises concerns about the highly concentrated market power (Commission 2010). Due to this level of concentrated power and the strong dependency both on the end of the consumer and third parties, *Google's* disruption or unintended consequences of the system could have detrimental effects on critical societal functions such as people's economic or social well-being (Sunyaev 2020). Therefore, it is evident that the major SEs can be considered Critical Information Infrastructure (CII).

Regardless of the known disadvantages of *Google's* market power, there still seem to be compensating factors that can explain the market success and the lack of focus on alternative providers. Possible reasons have already been in the focus of intense research. Findings range from economic explanations such as market failure, further to system specific factors like technically superior implementations, up to social or psychological influences like peer imitation or force of habit (Argenton and Prüfer 2012; Clemons 2019; Clemons and Wilson 2016; Dritsa et al. 2020). Nevertheless, single scientific metrics of professionals familiar with the technical details of SEs may lack meaning for the multidimensional evaluation of the market situation from the untrained users' perspective (Crudge and Johnson 2004). SEs, as CII, are information systems that incorporate a socio-technical nature (Sunyaev 2020). To avoid a possible bias towards a system view, an evaluation of SE market imbalance needs to focus on a socio-technical perspective and involve user preferences during contextual interaction with the technical subsystem. Additionally, from the user's point of view it might not be a decisive criterion how the SEs perform in terms of absolute measures but rather how they perform in relation to their competitors since this represents the possible range of alternatives. Accordingly, it appears that previous research does not fully reflect the evaluation scheme of the users. What are the distinctive features of SEs in the eyes of the users, and which of them dominate their decision for a certain SE? And how far off do users perceive the strengths and weaknesses of competitors such as *DuckDuckGo* or others from the market leader *Google*?

To understand the users' preferences for SE selection more profoundly, our research aims at facilitating a comparison from the users' perspective. Nevertheless, before being able to compare SEs from the users' position, it is first necessary to know the distinction criteria that reflect the user perspective. Therefore, this paper firstly focuses on identifying criteria that are decisive for the users' SE differentiation before

measuring how selected providers perform within these criteria in comparison to one another, leading to the following research questions:

- RQ 1: Which criteria are decisive for the users' search engine selection?
- RQ 2: How do selected providers perform within these criteria in relation to the market leader?

In order to answer these research questions, we identify four objectives that we examine in a four-step exploratory research approach that we call the User Centered Information System Evaluation (UCISE) and adapt it to fit a SE evaluation. First, to answer *RQ 1*, we derive search engine differentiation criteria from the user's perspective by combining previous research and current market implementations of SEs. Subsequently, the user perceived importance of the derived differentiation criteria is determined with the help of a large-scale survey. To answer *RQ 2*, we examine the performance of selected SEs with the help of an expert-assessment of selected SEs regarding the user-based criteria. Finally, the user-based criteria importance and the expert-based SE-performances regarding these criteria are combined in a two-dimensional framework for three different SEs. This allows to illustrate the overlap of their performances with user preferences (market fit) and to compare provider performances with one another.

Following the introduction, the Background section provides insights into relevant theories, concepts, terms, and related research, such as the System Usability Scale (SUS) and the repertory grid method. The Methodology section provides a detailed look into our research strategy consisting of the UCISE as a four-step methodological approach with a comprehensive look into the methods used for deriving user-based criteria, their importance ranking, the SE performance assessment and the result synthesis in a single framework. Afterwards, the Results section contains a description of our results while the Discussion lays an emphasis on discussing results and limitations of this work. Finally, the last section concludes our research and presents future research opportunities.

Background

The following section provides information on theoretical concepts and related research to ensure a common ground concerning the context of our work. A brief definition of essential terms and definitions regarding Information System (IS) and SE research is given in the first part. Subsequently, part two presents important findings of related research on SE assessment and user-based system evaluation as well as a basic concept of market fit analysis in business studies.

Terms and Definitions

A *Search Engine (SE)* is a program that examines a body of data for items that meet certain specified criteria and returns these items or their location (Butterfield and Ngondi 2016). Within the context of the World Wide Web, the term generally refers to a program that can be accessed via a website and allows users to search for web content. Therefore, in our work, we will use the term search engine or its abbreviation (SE) for web search engines that are publicly accessible via a web domain and focus on website-indexing of publicly accessible web domains.

Due to their set of interrelated components that serve the purpose of information retrieval, SEs classify as being *Information Systems (IS)*. An IS is a system consisting of several "interrelated components that work together to collect, process, store and disseminate information to support decision-making and control" (Jessup and Valacich 2008; Laudon and Laudon 1999). One of the fundamental characteristics of ISs is that they incorporate a socio-technical nature. ISs not only consist of software, hardware, and data components but also of organizational structures and people who use and interact with the system (Sunyaev 2020). As already stated in the Introduction, this is especially true for SEs. Their performance and results depend on both technical and social aspects. The technical aspects include architectural elements such as servers, databases and algorithms leading to the search results that match a particular search query. The dependency of the technical system's outcome on human-generated input stresses the role of the social subsystem (Orlikowski 2007). Due to this entanglement of social and technical components, SEs can be defined as so-called *Socio-Technical Systems*.

In turn, the strong entanglement of SEs with social components leads to another important term for this work, Critical Information Infrastructure (CII). Some ISs have become so essential for society that their disruption or malfunction will yield adverse consequences of critical magnitude, critical breadth, and

critical duration. Therefore, the term CII describes ISs, that in case of interruption can have disadvantageous effects on "vital essential societal functions or the health, safety, security, or economic and social well-being of people" (European Union 2008). As stated in the Introduction, the global SE market is dominated by only a few large providers. This resembles a largely monopolistic market structure of critical proportion. Since their position yields a competitive advantage through increasing data aggregation, large providers benefit of natural entry barriers for competitors. Therefore, it is not to be expected that a disturbance of the major providers could immediately be overcompensated by other providers. Additionally, due to the high market shares, the impact of a possible interruption or disturbance in one of the large providers would immediately affect a large user-base and therefore be of critical breadth. Thus, it is especially the level of concentrated power and the strong dependency, both on the end of the consumer and third parties, that account for the crucial role of SEs within society. Accordingly, SEs must be considered CII as they can have a severe impact on today's way of information dissemination and thus people's economic and social well-being (Sunyaev 2020).

Relevant Concepts and Related Research

As stated in the Introduction, SEs and their differentiation have already been in the focus of intense research. However, many existing SE comparisons on performance measures are based on isolated professional metrics from a system perspective. For example, Dritsa et al. (2020) propose an objective similarity score, using the content and ranking of search responses to measure SE performance. As a further example, Azzopardi et al. (2018) suggest a metric to measure the utility and costs of SE result pages. Whereas this focus may be necessary to evaluate technical details from a system perspective, it might not allow to holistically assess the socio-technical context of SEs, in a way that allows to explain their differentiation from a user's perspective. Crudge and Johnson (2004) state that when trying to understand user-based evaluation constructs, the derivation of evaluation constructs defined by the system objective arguably adopts a bias towards a system view. To avoid this bias, they find the repertory grid technique suitable for determining user-based SE evaluation constructs and capture the broad range of relevant evaluation criteria from a user engaged with a system. The repertory grid technique is based on semi structured interviews with the elicited data being recorded in a two-dimensional matrix. For their work, Crudge and Johnson (2004) used a sample of five SEs as elements which were presented to ten participants for dyadic elicitation to derive evaluation constructs that allowed discrimination of these SEs from user perspective. The suitability of this approach is suggested as the users' view on a system is based on a subconscious set of evaluation criteria derived by experiences, prior knowledge and the judgement of systems or objects to be alike or different from others (Kelly 2003). Therefore, it might not be a decisive criterion how the SEs perform in terms of absolute measures but rather how they perform in relation to their competitors, representing the possible range of alternatives.

It follows that user-based SE evaluation will always happen as a comparison to what the user is already familiar with and therefore needs a certain framing. Using the repertory grid technique, Crudge and Johnson (2004) elaborate a foundation of user-based differentiation dimensions in five thematic areas with detailed influencing factors. These thematic areas were named "Result issues", "Search mechanism", "Features", "Front page design" or "Advertisement". Since Crudge and Johnson (2004) do not elicit the user-perceived importance of the thematic areas, they suggest deploying a large-scale survey for future work. A more detailed description of the areas, including their influencing factors, will be provided in Part I of the Methodology section.

The relativity of evaluation constructs from the users' perspective is also stated by Brooke (1995) concerning the usability of ISs. Brooke (1995) argues that the usability of a system does not exist in any absolute sense and therefore is strongly context-dependent and subjective. To overcome this subjective nature of usability and allow a generalized system assessment that bears cross-system comparison, ten statements were introduced as items on a five-point Likert-scale that interviewees could agree or disagree with. The Likert-scale allows transforming individual opinions, perceptions or preferences into objective data with the help of an ordinal or interval scale. It therefore helps in the scientific assessment of attitudes towards a system by quantifying subjective preferential thinking in a validated and reliable manner (Joshi et al. 2015). To facilitate easy agree- or disagreement with the items, Brooke (1995) emphasizes that all statements should be formulated in first-person perspective and as extreme expressions regarding the topic being captured. By combining all ten items in a single score from 0-100, the SUS represents a composite overview of a

system's overall usability and therefore allows to globally quantify single subjective evaluations. Moreover, an adaption of this approach seems especially suitable for the purpose of SE evaluation since it allows to translate single subjective user evaluation into aggregated objective data while respecting the users' contextual nature of system evaluation. Additionally, when using multiple diversified statements, it permits to consider that from the users' position, not only single aspects of the technical subsystem of an IS but the whole socio-technical system is being evaluated.

Using relative concepts for system evaluation is not only utilized in IS research but is also a common approach in strategic management when assessing a firm's most efficient resource allocation. Grant (2005) suggests putting a firm's resources and capabilities into perspective by assessing them based on the two criteria of relative strength and strategic importance (see Figure 1). By comparing the firm's resources and capabilities in terms of relative strength to the aggregate of its competitors and considering their strategic importance, a firm can identify its key or superfluous strengths as well as key or irrelevant weaknesses. Whereas this framework was initially meant to identify the most efficient resource allocation for business purposes, it can be slightly adapted to assess the current fit of a firm to its market. Transferred to the context of SE evaluation from users' perspective, an adaption of this framework seems to be useful when comparing the strengths and weaknesses of SEs and identifying their importance to the user.

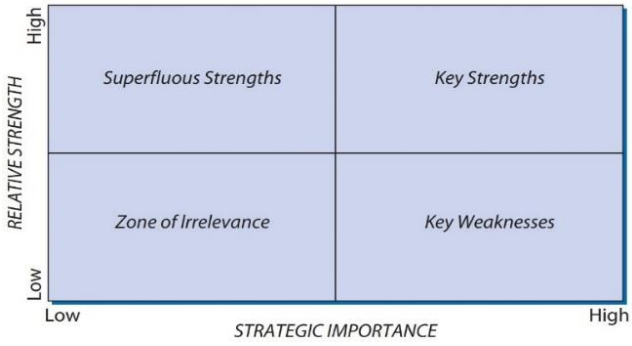

Figure 1. Framework for Appraising a Firm's Resources and Capabilities as Suggested by Grant (2005)

Methodology

According to the four identified objectives mentioned in the Introduction, we created the UCISE (see Figure 2) as a synthesized methodological research approach with four sequential parts to answer the two research questions. The UCISE is directly adapted to evaluate SEs but can easily be altered to fit other ISs. Part I focused on the derivation of SE differentiation criteria from the users' perspective. Existing criteria from related literature were gathered and regrouped to form the basis of our work. Additionally, to reflect the current market and consider the latest developments of SE-providers, a compilation of existing SEs was considered to finally form our suggested set of ten criteria for SE differentiation from the users' perspective. In Part II the ten criteria from users' perspective were then assessed regarding their user-perceived importance by an online survey. In parallel, Part III addressed the performance ranking of three selected SEs by an expert-assessment. A set of professionals assessed the SE-performance in regard to the user-based criteria from Part I. Finally, in Part IV previous results could be combined to illustrate the overlap of different providers with user preferences (and thus their market fit) and create comparability between different providers.

Part I: Derivation of Criteria for SE Differentiation

The derivation of criteria for SE differentiation from the users' perspective splits up into two components. The first component is based on former research of Crudge and Johnson (2004) using the repertory grid technique to derive five user-based evaluation thematic areas for SEs. Since establishing these five thematic areas originally aimed at unifying diverse user-derived statements from multiple semi-structured

interviews, the thematic areas and their descriptions had to be logically regrouped and renamed as shown by Table 1 to assure a consistent criteria description as the basis of our research.

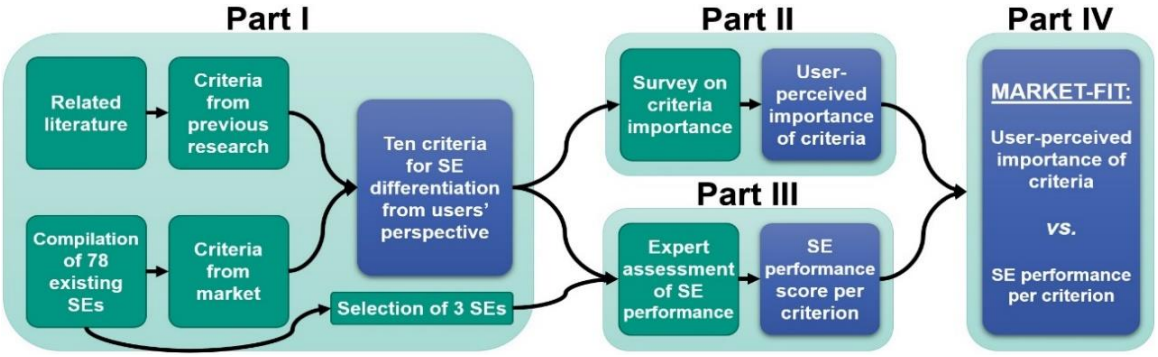

Figure 2. Synthesis of UCISE as a Four-Part Methodological Research Approach to Assess the Overlap of User Preferences with Search Engine Performances (Source: Own Visualization)

As it must be assumed that SEs have considerably evolved since the previous research of Crudge and Johnson (2004), the second component of Part I was introduced to ensure the criteria base's topicality. To ascertain that the final criteria also allow differentiating the current market, the criteria-set derived with the help of previous research was logically extended. As a users' evaluation construct is always based on experience, prior knowledge and the judgement of systems or objects to be alike or different from others (Kelly 2003), the logical extension of the criteria list was done by introducing additional criteria deemed necessary to compare and distinguish currently available SEs. For this cause, we compiled a set of over 500 SEs using various sources. Thereof, a set of 78 suitable SEs was derived by the exclusion of entries that were invalid (e.g. copyright update older than three years) or inappropriate for the context of our research (e.g. no website-indexing). The rules for exclusion and the complete list of 78 SEs can be taken from Appendix. Since a dyadic comparison of these 78 elements would have led to over 3000 direct comparisons, a random sample of 10 elements was used instead to reduce complexity. As this work aims to examine whether the current market imbalance reflects user preferences, this random sample was enriched by the six globally most-used SEs *Google*, *Bing*, *Yahoo!*, *Yandex*, *DuckDuckGo* and *Ask* (Statista 2020). This formed the final sample of 16 SEs used in Part I of this work, which can also be taken from the Appendix. The instances of this finalized SE sample were then compared in pairs, and new criteria were introduced whenever the derived criteria from literature (Table 1) were not deemed sufficient to state in which dimension the instances were alike or different.

These two components of Part I of the UCISE led to a final set of ten criteria as a basis for the evaluation of SEs from the users' perspective. It is important to understand, that the identification of these criteria is an important prerequisite for a user-centered system evaluation since the following steps of the methodological approach build on these user-based criteria. The final criteria set, including a detailed description of each criterion with exemplary features, is presented in the Results.

Part II: Determination of User-Perceived Criteria Importance

To answer our first research question and identify criteria that are decisive for the users' SE selection, the users' perceived importance of the derived SE criteria was captured in an online survey. Online surveys offer the possibility to collect self-reported data, and as such, are an ideal tool to gain insight into the user's perceived importance and preferences of the criteria (Bhattacherjee 2012). They are especially feasible by offering many advantages as a tool with efficient data collection, easy access to a large target group and low incurred costs (Selm and Jankowski 2006).

The survey design was chosen to contain three main parts: a general part about previous SE usage, the presentation and rating of the derived user-based SE criteria from Part III of the Methodology and a collection of demographic data. The detailed survey is available from the authors on request. By capturing

the users' previous SE experience and usage, the first part of the survey served to verify representation of the current market situation within our set of participants and thus assured the suitability of the survey results for generalization (Weimann and Brosig-Koch 2019). Part three served a similar purpose by capturing demographic features about our set of participants. In the second part the participants of the survey were first asked to arrange the ten criteria in descending order, according to their perceived importance, which yielded in a priority list from the users' position. Afterwards, the participants were asked to rate the ten criteria using a 7-step, symmetric Likert-scale (Joshi et al. 2015). This Likert-scale was chosen to range from *extremely unimportant* (weight 1) to *extremely important* (weight 7) with *neither unimportant nor important* (weight 4) as a neutral element. As mentioned in the Background section, this allowed to translate the subjectively perceived importance of SE criteria into an aggregate global mean. The 7-step Likert-scale was selected since it can increase the reliability of the participants' responses in a survey compared to a 5-step Likert-scale. This originates from the fact that it provides more varieties of options which in turn increase the probability of meeting the participants' subjective reality (Joshi et al. 2015).

Dimensions from Crudge and Johnson, 2004		**Regrouped/ renamed to**
Thematic area	***Influencing factors (excerpt)***	
Result issues	Content, quantity, balance	Search result quality
	Layout	Aesthetics
	Viewing sites (e.g. sneak-a-peek)	No. of search-related features
Search mechanism	For obtaining results, ease of use	Ease of use
	After obtaining results	No. of search-related features
Features	Directory, images, quantity	No. of search-related features
	Languages	Customization
	Presentation, location, visibility	Ease of use
Front page design	General, color, logo	Aesthetics
	Clarity and focus of search	Ease of use
Advertisements	Front page, results page, location	Disturbance of advertisement

Table 1. Re-Framed User-Based SE Evaluation Dimensions from Crudge & Johnson (2004)

To verify the completeness of the provided criteria base, by an additional question, participants were given the opportunity to mention criteria that were missing in the provided collection from their perspective. Additionally, to ensure the validity of the user preferences, the constructs of SE criteria were described to the user in detail, utilizing the definitions and descriptions as presented in Table 2. Thus, the survey ensured a common baseline about the evaluated constructs for all survey participants. To further increase the validity of the survey results, countermeasures were taken to avoid the common method variance (Podsakoff et al. 2003). The Common Method Variance (CMV) is a variance, especially present in self-report questionnaires, that is created by the measurement method rather than by the construct itself (Chang et al. 2010). To prevent the CMV, the survey was designed following guidelines of Chang et al. (2010), such as encouraging participants to truthfully state their own opinion or asking for the demographic data at the end of the survey.

In conclusion, based on the results of the user-based criteria importance ranking on the 7-step Likert-scale, it was possible to finally calculate an importance score for each criterion. This was done by calculating the mean of the Likert-scale rating from Question 5 of the survey for each criterion in the following way:

$$I_c = \frac{\sum_{i=1}^{7} (i * n_i)}{N}$$

I_c = Importance score for criterion $c \in$R
i = Position on Likert-scale $\in [1,7]$
n_i = Amount of survey participants selecting position i on Likert-scale
N = Total amount of survey participant

The priority list that was generated as a result from another survey question served as a control scheme to verify the plausibility of the calculated importance scores per criterion.

Part III: Assessment of SE Performance per Criterion

To answer the second research question about the relation of actual performances of SEs, it was first necessary to assess selected SE-performances in regard to the criteria from the users' perspective. As stated in the Background section, user-based measures are defined by the context in which the assessed system is used and thus do not exist in an absolute sense (Kelly 2003). Since many user-based criteria such as *ease of use* or *usability* cannot directly be quantified in an objective way, user-based SE performance assessment and its measures must consider their context and perception dependency (Brooke 1995). To take these difficulties into account when attempting to objectively assess the subjective topic of usability, Brooke (1995) developed the System Usability Scale (SUS), which was introduced in the Background section. For this work's purpose, the core principal of the SUS was adapted and modified to fit the UCISE approach for a user-based SE assessment. This was done by creating a questionnaire concerning SE performance in regard to the user-based criteria. This questionnaire contained between one and three statements for each of the ten user-based SE criteria. Within the final set of 21 statements, each item was created in such a way that it closely resembled one or multiple aspects of the respective user-based criterion-definition to allow the evaluation of SE performance from the users' point of view. A 7-step Likert-scale was then used to measure the degree of agreement or disagreement with the provided statements. The scale ranged from *strongly disagree* (weight 1) to *strongly agree* (weight 7) with *neither agree nor disagree* (weight 4) as the neutral element. To facilitate easy agree or disagreement with the provided items, all statements were formulated in the first-person perspective and as extreme formulations regarding the criterion being captured. A complete version of the questionnaire is available from the authors on request.

Subsequently, a set of three diverse SEs was selected to be assessed regarding their performances. This set consisted of the market leader *Google*, the most popular privacy-focused competitor *DuckDuckGo* and an independent provider *Gigablast* which maintains an own website index. By selecting SEs that differ vastly in regard to multiple criteria, such as aesthetics or popularity, it was considered that user-based system evaluation is contextual and always takes place in relation to existing experiences. Therefore, the selection of extreme exemplary SEs aimed at easier assessment of crucial differences in regard to the user-based criteria. Additionally, by selecting distinctive and unique SEs, this work also aims at examining a diverse range of findings to provide different cases of possible results and assess the feasibility of the UCISE.

After creating the questionnaire based on statements related to the user-based criteria and the composition of the set of three extreme SE examples, the actual assessment of SE performance was conducted with the help of an expert assessment. This method was chosen because relying on the experiences and the intuitions of experts in judging the presented situations allows an assessment of situations that otherwise are hard to measure or cannot be measured directly (Boiko 2018). To allow for a well-founded assessment of the degree to which the three examined SE performed, we chose to select experts with a background in Information Technology (IT), computer science or User Experience (UX) as they possess in-depth knowledge of the matter and it can be expected, that due to their profession, they have a larger base of experience than the average user. Therefore, it was assumed that they can judge on a sounder and more qualified context regarding the SE performance in the user-based criteria and not only based on their immediate perception.

During the assessment process, the participants were provided with the definition of the ten criteria, as presented in Table 2. By presenting the criteria, we were able to raise awareness of the different focus points

in the context of this work and direct the experts focus towards a more perception-based view of the average user. Afterwards, the participants were provided with the landing pages of the SEs one after the other. They could interact and use each SE for a maximum of 20 minutes to become familiar with it and develop an opinion. The order in which the three SEs were provided was randomized for each expert. After the assessment was completed for all 3 SEs of our set, the participants were asked to state their degree of agreement or disagreement to the questionnaire statements on the 7-step Likert-scale. The expert-assessment was led and documented by an interviewer to avoid uncontrolled data entry into the results. This removed the need to randomize the statement or criteria order, as the risk of wrong or faulty entries was minimized. While rating the statements, the participants were allowed to revisit the landing page of each SE if needed. Since SEs constantly evolve, it is important to note that the assessment was based on the state of the SEs as of January 2021. Each SE's performance was finally derived for each criterion by building the equally weighted mean of the expert agreement or disagreement to the statements assigned to the respective criterion.

As the questionnaire also contained negated statements, there was the need to negate their ratings to maintain score consistency. This was done by calculating 8 - $x_{h,i}$ for the actual score for these statements. Further details on this approach can be drawn from Brooke (1995).

$$S_c = \frac{\sum_{i=1}^{j} \left(\frac{\sum_{h=1}^{k} x_{h,i}}{k} \right)}{j}$$

S_c = Score for criterion $c \in$ R
h = Current statement in criteria c
k = Total amount of statements for $c \in [1,3]$
i = Current question
j = Total amount of questions
$x_{h,i}$ = Score for statement i from expert h , $x_{h,i} \in [1,7]$

Part IV: Illustration of Overlap of SE Performance with User Preference

As depicted in Figure 3, the final part of the UCISE combines the results of the three previous parts in an adaption of Grant's (2005) framework (see Background section). This allows to visualize the fit of a SE to users' preferences on the one hand and sets the ground for comparing performances of different SE providers on the other hand.

Based on the previous results, the ten user-derived criteria were considered as instances and arranged in a two-dimensional feature space for each examined SE. The two features of each criterion are the criterion's user-based importance and the expert-based performance of a SE regarding this criterion. The user-based criteria importance is known from the online survey and is assigned to the x-axis of the adapted framework. It is labelled "perceived importance". The higher the mean of the user-based importance ranking within the survey, the more important (further right) the respective criterion. Whereas the perceived importance of a criterion remains constant throughout all different frameworks, the second feature of each criterion varies depending on the performance of the examined SE within the user-based criterion. Since this performance is known by the expert assessment and based on comparing different providers with one another, this feature is labelled "relative strength". Within the framework, the relative strength is assigned to the y-axis.

After plotting the data, the strengths and weaknesses of each SE were revealed. For example, if a SE performs well in a criterion that is important to the users, this yields a crucial strength. If it performs weak in an important criterion, this yields a major weakness. The other way around, if a SE performs weak in a criterion that is unimportant to the users, this is a minor weakness. If it performs well in this unimportant criterion, this results in a useless strength. Therefore, to perfectly address user preferences, the relative strength of a SE in regard to a criterion should at least match the user-perceived importance of this criterion. Thus, SE should outperform the market (represented by the solid horizontal line) in more important criteria and underperform the market only in less relevant criteria. Regarding the user priorities it follows that a SE should at least perform as strong regarding a criterion as the users perceive the importance of this criterion.

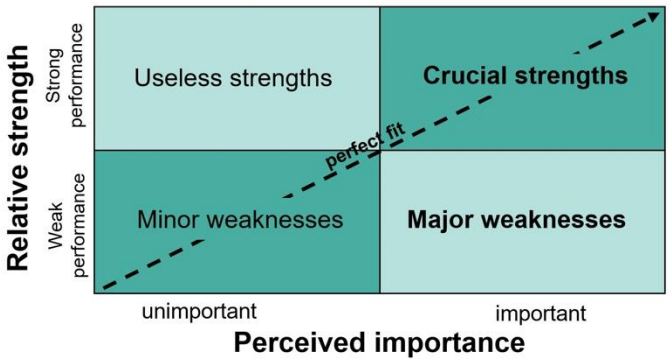

Figure 3. UCISE Framework Adapted for Evaluation of SE Market Fit Adapted from Grant (2005)

Thus a "perfect fit" to user preferences consists of criteria performances that match the frameworks diagonal line. The area over this curve marks an area when a SE outperforms the perceived importance in a certain criterion. Vice versa, the area under the curve indicates when a SE underperforms the perceived importance.

By changing the original x-axis of the framework that is suggested by Grant (2005) into user-perceived importance, the statement of the framework fundamentally changes. This change allows the evaluation of SE performances from the user's eye, permitting insights into a SE's adaption to user preferences. Due to the large-scaled survey of Part II, one can assume that (at least to a certain extent) a large enough user-base is captured to reflect the general preferences of users on the SE market. Therefore, the adaption of the framework can be used to illustrate the market fit of an IS and is deemed suitable for our UCISE research approach. It will in the following be referred to as the *UCISE framework* which in our case is adapted to evaluate SE market fit.

Results

The following section provides the representation of the results of our research. To allow the reader to draw immediate conclusions about the methodological origin of the results, the presentation is divided into the four parts of the UCISE as depicted in Methodology section. Hence, we first present the resulting ten final criteria for SE differentiation from the users' perspective. Subsequently, their user-perceived importance will be shown. Afterwards, the performance score of the three assessed SEs regarding the ten criteria is provided. Finally, the last part of this section puts all the previous results into context by arranging them in the dimensions of the UCISE framework adapted for SE evaluation (see Methodology Part IV).

Ten Criteria for SE Differentiation from Users' Perspective

Part I of the UCISE methodology led to the finding of ten final criteria for the differentiation of SEs from users' perspective as depicted in Table 2. Of these criteria, six were adapted from previous research by regrouping and renaming the five thematic areas from Crudge and Johnson (2004) and four criteria were derived from current market. The six criteria based on previous research were renamed to *aesthetics*, *customization*, *disturbance of advertisement*, *ease of use*, *number of search-related features* and *search result quality*. Additionally, the criteria *automated personalization*, *popularity*, *privacy protection* and *social or environmental contribution* were deemed necessary to differentiate the full range of current web search implementations. A detailed description of each criterion, including an exemplary specification, can be drawn from Table 2.

User-Perceived Importance of the Criteria

The conducted online survey of Part II of the methodological approach had 263 participants with a female to male ratio of 44.1% to 55.9%. The average age of the participants was 33.37 years, and 58% of the

Aesthetics	Describes whether the SE has a pleasant visual appearance. *E.g.: The chosen colors and proportional arrangements on the provided website are well balanced. The design is perceived as aesthetically.*
Automated personalization*	Describes the automated tailoring of the search results to the user based on the user profile. *E.g.: The SE automatically displays results based on my search history and personal information. It knows my individual preferences.*
Customization	Describes the possibility to control or alter the basic layout and default settings. *E.g.: The SE provides the possibility to modify and store presets which may alter search-behavior and/or visual default settings.*
Disturbance of advertisement	Describes the non-existence of paid links within the search results or their placement in a non-disturbing position. *E.g.: There is no placement of advertisement within the list of results or in areas that are interfering with the search process.*
Ease of use	Describes how effortless it is to use the SE. High ease of use describes the absence of difficulty or great effort when using the SE. *E.g.: The SE is intuitive to use and easy to navigate. The desired features are easy to find.*
Number of search-related features	Describes the scope and the variety of available functions that the SE provides in addition to the basic web search. *E.g.: There are plenty of available features like for example auto correction or proactive results (E.g., "do you mean XX"). Also refers to search filters like result restriction to images, videos or geographic regions.*
Popularity*	Describes that the SE is well known and used by many other users.
Privacy protection*	Describes that the SE does not accumulate search requests or meta-data for other reasons than delivering results to the current search process *E.g.: User data is not accumulated and post-processed for business purposes by the SE and/or is not passed on to third parties.*
Search result quality	Describes the delivery of qualitative results that enable the user to quickly find satisfactory information. *E.g.: A sufficient amount of relevant results is provided that can be found in appropriate time. They consist of the right amount of known and new links.*
Social/ environmental contribution*	Describes that the search engine focuses on charitable contributions that serve other than pure business purposes. *E.g.: The search engine is committed to be CO_2 neutral or donates money for charitable projects like social inequality or plastic waste.*

Table 2. Ten Criteria for SE Differentiation from Users' Perspective. The Four Criteria that Were Derived from Current Market are Marked with an Asterisk (*)

participants stated to hold a bachelor's degree or higher. When asked for their primary SE, 85% of the participants named *Google*, only 6% stated *Ecosia*, and 4% stated *DuckDuckGo* as their primary SE (5% stated other providers). The results concerning the importance score are depicted in Table 3. The ranking of the SE differentiation criteria according to their perceived importance on the 7-step Likert-scale revealed that *search result quality* is by far the most decisive criterion for SE differentiation from a users' perspective

with an importance score of 6.59 out of 7. The *ease of use* is ranked as the second most important criterion with an importance score of 5.3, followed by *privacy protection* with an importance score of 5.07. This importance ranking of the top three most important criteria is also confirmed by the survey control question that lead to the *Priority list* as displayed in Table 3.

The importance scores of the remaining seven criteria show that these criteria are all perceived as being between "neither important nor unimportant" and "rather unimportant", (please recall that 4 was defined as the neutral element). Additionally, their scores indicate only one point difference between the fourth most important criterion (*disturbance of advertisement* with 3.81) and the least important criterion (*customization* with 2.96). Furthermore, it can be drawn from the priority list that for these seven criteria the ranking that was derived from the control question differs from the ranking of the importance score.

The additional question verifying the completeness of the provided criteria base only led to 17 mentions in total. After unification and grouping, four main topics were revealed. Four mentions concerned the *interoperability* of SEs between different devices, three mentions concerned the *mobile friendliness* and two were aimed at the *transparency* in regard to open source approaches. The *speed* of the SE was mentioned eight times.

Rank	Criterion	Importance score (Q5)	Priority list (Q4 - control question)
1	Search result quality	6.59	1
2	Ease of use	5.3	2
3	Privacy protection	5.07	3
4	Disturbance of advertisement	3.81	5
5	Social/ Environmental contribution	3.79	6
6	Number of search-related features	3.71	4
7	Popularity	3.61	10
8	Aesthetics	3.25	7
9	Automated personalization	3.07	9
10	Customization	2.96	8

Table 3. User-Perceived Importance of the Ten Criteria for SE Differentiation from Users' Perspective Sorted by Descending Importance Score

SE Performance per Criterion

The SE performance score per user-based criterion resulted from the expert assessment in Part III of the UCISE approach. The detailed results of the assessment are presented in Table 4. As already stated, it was ensured that all experts had profound expertise concerning IT or UX. Therefore, the set of experts consisted of two IT-consultants, a software developer, a hardware developer, a design thinking expert and a data scientist. As presented in Table 4, within the sample of three selected providers, *Google* achieved the best ratings in *search result quality* (6.22), *ease of use* (6.67), *number of search related features* (6.22) as well as *aesthetics* (6.33), *automated personalization* (5.67) and *popularity* (7). This way, *Google* dominates the selected competitors in six of the ten evaluated criteria. *DuckDuckGo* was rated best in *privacy protection* (6.50) and *customization* (3.67). Additionally, it shares the best score for *social and environmental contribution* (3.67) together with *Gigablast*. Besides *social and environmental contribution*, *Gigablast* also achieved the best score for the criterion *disturbance of advertisement* (7). A remarkable low performance for *Google* is its score in terms of *privacy protection* (1.67) which is not only its lowest performing criterion but also a remarkable difference to its competitors. Both *Googles'* and *DuckDuckGos'* performance is above

the neutral element of 4 in eight of ten examined criteria. In comparison, *Gigablast* performs below the neutral element in eight of the ten criteria. The lowest overall score is achieved by *Gigablast* concerning the criterion *popularity*.

Criterion	**Google**	**DuckDuckGo**	**Gigablast**
Search result quality	6.22	5.28	1.94
Ease of use	6.67	6.42	3.75
Privacy protection	1.67	6.50	5.5
No. of search-related features	6.22	5.17	2.94
Disturbance of advertisement	5.25	5.33	7
Social/ Environmental contribution	3.5	3.67	3.67
Aesthetics	6.33	6.17	1.83
Customization	4.39	6.06	2.5
Automated personalization	5.67	2.33	2
Popularity	7	4	1

Table 4. Averaged SE Performance Scores per Criterion and SE. Bold Scores Represent the Best Score per Criterion Across the Three SEs

SE market fit

The last part of the results combines the findings of the three previous methodological Parts I-III. This is realized by the UCISE framework, which visualizes the fit of the three expert-assessed SEs to the importance of the user-based criteria. Figure 4 illustrates the market fit for the market leader *Google*. Starting from the top right, *Google* has a strong performance concerning the *search result quality* which is the most important criterion to SE users. Nevertheless, with 6.22 the performance in terms of *search result quality* still slightly lacks behind in relation to the user-perceived importance of this criterion which is 6.59. This is reversed when it comes to the second most important criterion *ease of use*. With 6.67 *Google's* performance is very strong whereas the user only perceives this criterion as rather important (5.3). Overall, *Google's* strong performance in terms of *search result quality* and *ease of use* regarding the two criteria perceived as most important lead to crucial strengths.

A crucial weakness however shows when looking at *Google's* weak to very weak performance in terms of *privacy protection* (Recall that this criterion is perceived third most important by the users). *Google's* performance in regard to almost all of the remaining seven criteria remains above the perfect fit. Nevertheless, the remaining seven criteria are all rated between neutral and rather unimportant in regard to their perceived importance which makes *Google's* performance in these criteria being classified as useless strengths. The only exception is *social/ environmental contribution* being a minor weakness.

The UCISE-frameworks of all examined SEs are depicted in Figure 5 with *Google's* framework being shown gain in part (a). *DuckDuckGo's* fit is illustrated in part (b) in the upper right of the figure. With 5.28 its performance in terms of *search result quality* is only scored as rather strong and therefore lower as *Google's* score. Nevertheless, when it comes to *ease of use DuckDuckGo* nearly performs as good as the market leader. In terms of the third most important criterion, namely *privacy protection*, *DuckDuckGo* outperforms *Google* by far. Similarly, to *Google's* performance, *DuckDuckGo* also performs above the line of the perfect fit for the seven criteria that are perceived less important. The criterion *automated personalization* is an additional minor weakness to *Google's* minor weakness in *social/ environmental contribution*.

Gigablast's market fit is shown in part (c) of Figure 5. In general, its performance is rated very contrary to that of *Google*. This can be drawn from the weak performance in the most important criterion *search result quality*. In this regard, *Gigablast* only scores 1.94 (Likert-scale for performance ranking started at 1).

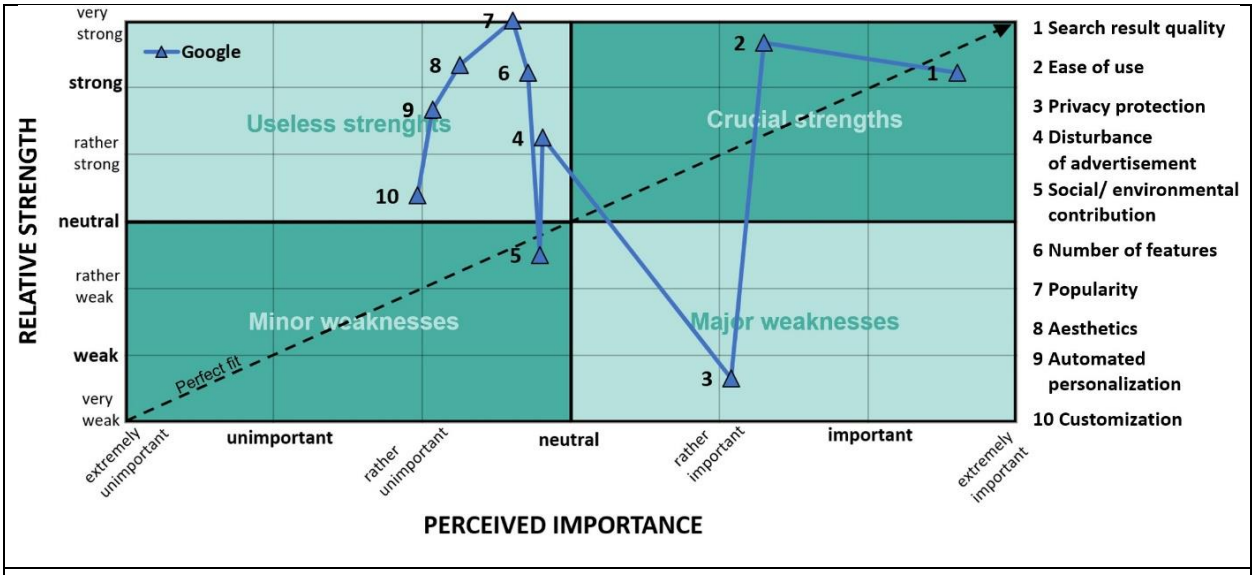

Figure 4. Market Fit of the Market Leader *Google*. (Source: Own Visualization)

Additionally, *Gigablast* only scores a 3.75 in regard to *ease of use*. This yields two major weaknesses concerning the two criteria that are perceived as most important by the users. *Gigablast's* only crucial strength is the *privacy protection* where it nevertheless falls short of *DuckDuckGo's* score. With a very strong performance regarding the *disturbance of advertisement*, *Gigablast* achieves a useless strength. In regard to all remaining and rather unimportant criteria, *Gigablast's* performance yields six minor weaknesses.

To summarize, *Google* dominates the examined SEs in regard to the two most important criteria *search result quality* and *ease of use*. *DuckDuckGo* takes the lead concerning the third most important criterion, *privacy protection* and otherwise largely resembles *Google's* performances. With *disturbance of advertisement* the third examined SE *Gigablast* only dominates a criterion that users perceive as neither important nor unimportant. A clear dominance regarding the remaining seven criteria that are perceived as rather unimportant cannot be observed. Nonetheless, *Gigablast's* performance mainly yields minor weaknesses concerning these criteria, whereas the other two providers mainly show useless strength.

Discussion

In order to better understand the overlap of user preferences with SE performance, we synthesized the UCISE approach for ISs with four sequential parts and directly applied it to SEs. Through this approach, it was possible to assess decisive criteria for the users' SE selection and differences between market leaders and competitors from the users' perspective. First, we derived ten criteria that allow a SE differentiation from the users' perspective (see Table 2) by building on differentiation schemes of previous research and introducing additional constructs with the help of currently available SEs on the market. Second, we examined the user-perceived importance of these ten criteria with the help of a large-scale survey with 263 participants that was deployed online. This led to the result of only three of the presented criteria being perceived as important in the eye of the users. The criterion with the by far highest importance is the *search result quality*. This criterion is the only one that is perceived as extremely important within our research. The second and third most important criteria are the *ease of use* and the *privacy protection* both being perceived as rather important by the users. The remaining seven criteria for SE differentiation from users' perspective are all perceived as being neutrally important or rather unimportant (see Table 3). Third, our methodological approach led to the performance assessment of three diverse SE providers (*Google, DuckDuckGo* and *Gigablast)* with the help of experts that have a professional relation to IS.

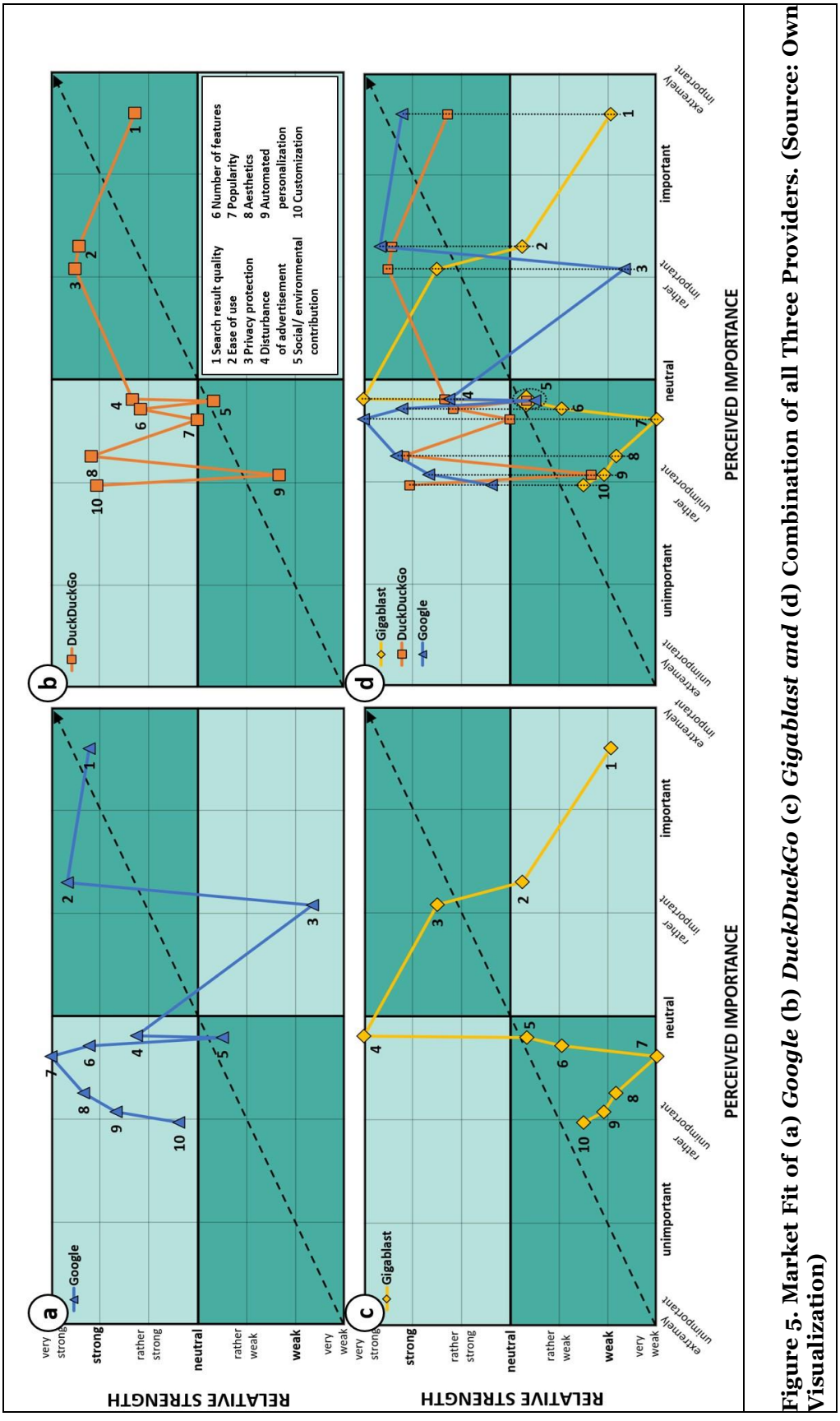

Figure 5. Market Fit of (a) *Google* (b) *DuckDuckGo* (c) *Gigablast and* (d) Combination of all Three Providers. (Source: Own Visualization)

Fourth, the results of the expert assessment as well as all previous findings were combined in a common framework. This UCISEs framework opposes an ISs (in our case SEs) performance regarding the user-based criteria to their user-perceived importance and thus leads to a visualization of a systems' market fit. It additionally allows to compare different providers to one another on an aggregated level. The combination of all findings in this manner resulted in the observation that *Google* outperforms its competitors in the first two most important criteria (*search result quality* and *ease of use*) but largely lacks behind in terms of *privacy protection* although this criterion is perceived as third most important to SE users. Therefore, *Google's* performance in regard to *search result quality* and *ease of use* yield crucial strengths and its weak *privacy protection* can be considered as a major weakness with both *DuckDuckGo* and *Gigablast* outperforming *Google* in this respect.

Our results indicate that it is worth noticing that user-based SE evaluation is done by a multidimensional construct of criteria that evolves over time and is not purely content related. In the previous work of Crudge and Johnson (2004), five thematic areas for SE differentiation from a users' perspective were identified. However, within our work these areas had to be regrouped to six criteria and appended with four additional criteria for allowing differentiation of the current market situation. One example is the recent tendency towards a focus on social and environmental contribution in all industry sectors, which has gained a lot of attention lately (Ruch et al. 2011). Due to this evolving behavior of evaluation constructs and the underlying exploratory nature of the UCISE approach, it has to be stated that it cannot fully be ensured that the elicited ten criteria reflect the topical users' perspective in a holistic manner. Nevertheless, our survey results indicate that with a total of 17 mentions, there were only a few suggestions for additional evaluation dimensions. Considering that the provided ten criteria descriptions already partially contained those additional suggestions, this finding is further put into perspective. For example, the suggestion of *speed* as an additional criterion is already partially contained in the description of *search result quality* by the wording "…in appropriate time…" (see Table 2). Additionally, it could be argued that *ease of use* is defined broad enough to also include the suggestion of *mobile friendliness*.

The findings about the perceived importance of the ten criteria illustrate that users mainly evaluate SE based on the three criteria *search result quality, ease of use* and *privacy protection*. This result is partially in line with the findings of a recently published paper of Prüfer et al.(2020), who assessed the data-drivenness of the current SE market by a discrete choice experiment. They find that both a reduction in the quality of search results and the degree of personalization (which is concluded as an equivalent to our definition of *privacy protection*) of a search engine have a significantly negative effect on user satisfaction. Additionally, Prüfer et al. (2020) also consider the number of ads to have a significant effect on user satisfaction, whereas in our work the *disturbance of advertisement* was perceived as neither important nor unimportant and thus did not show an equally significant effect on users.

Prüfer et al. (2020) also indicate that the criteria that they find to be significant do not have an equal weight, with the quality of search results being roughly twice as important as their definition of *privacy protection* (degree of personalization). This is also shown within our findings. It can be noted that there is already a significant margin within the averaged criteria importance within our research (*search result quality* with 6.59, *ease of use* with 5.3 and *privacy protection* with 5.07). This indicates that the *search result quality*, as the only criterion rated extremely important, largely dominates all other criteria. This should come as no surprise, given the fact that search engines are massive information retrieval systems (Seymour et al., 2011), tasked with finding and presenting relevant information to the user. An even larger gap shows between the third and fourth most important criteria with *disturbance of advertisement* being rated as neither important nor unimportant (3.81) on average. Hence, the large margins may suggest the classification of all ten criteria into three different levels of importance. Following this reasoning, the first importance level is suggested to consist of *search results quality* alone, the second level includes *ease of use* as well as *privacy protection* and the third level contains the remaining seven criteria.

Additionally, it is noteworthy that the seven criteria of the third level are only separated by 0.85 points in their importance score. This level is bounded by the fourth most important criterion *disturbance of advertisement* with an importance score of 3.81 and *customization* with the lowest overall importance of 2.96. Moreover, within the third importance level some criteria are only separated by 0.1 points in their importance scores. Keeping in mind that by these scores, the criteria are rated as neutrally important to rather unimportant, this suggests that users do not perceive them as decisive for the SE differentiation. This conclusion is supported by the results of the survey control question that yields the criteria priority list (see

Table 3). Within this list the same seven criteria are placed fourth to tenth rank but their order changes considerably. This supports the assumption that criteria of the third importance level are indeed perceived as less important and that their ranking could be interchangeable in contrast to the criteria of the first two importance levels that keep a consistent order in both the importance score and the priority list.

Another interesting aspect concerning the perceived importance of the criteria is that user preferences could appear to be slightly inconsistent or contradicting. As already stated, *search result quality* was ranked as the dominant criterion by a substantial margin. It therefore may seem surprising to find that *privacy protection* was rated as the third most important criterion. As previous research has shown, SE providers rely on personalization strategies to increase their perceived search result quality (Hannak et al. 2013; Teevan et al. 2010). Due to the necessity of aggregated user data in this regard, this comes at the cost of a users' privacy (Castellà-Roca et al. 2009). Hence, one can conclude that there is a trade-off between a high *search result quality* and *privacy protection* because a higher search result quality can be achieved by neglecting the users' data privacy. It follows that *automated personalization* and *search result quality* have an inverse effect on *privacy protection*. Nonetheless, by a high perceived importance of both *search result quality* and *privacy protection* and at the same time with *automated personalization* being perceived as rather unimportant (see Table 3), our findings of the survey suggest that users are not aware of this direct relation. One possible explanation for not perceiving this dependency could be the lack of knowledge and understanding of the general concept of personalization and how it is applied in the context of SEs. Untrained users without sufficient in-depth knowledge about technical details of search algorithms might not consider the logical connection between the criteria in their mental evaluation process. Another possible explanation for the high perceived importance of *privacy protection* could be a phenomenon known as the "privacy paradox". This phenomenon describes the discrepancy in a users' self-proclaimed preference to not disclose personal information and their actual behavior. While participants of surveys seem to place a high value on their privacy, they often disclose a lot of data for very little benefit in their actual behavior (Norberg et al. 2007). Hence, it is possible that the participants of our survey rated *privacy protection* according to their self-proclaimed intentions instead of their actual preference or behavior. Whereas the actual cause of this discrepancy is of great interest, a detailed examination of the underlying phenomenon was beyond the focus of this work. However, it must be kept in mind that concerning the importance score of *privacy protection* might overdraw the real influence of this criterion on a users' SE selection.

The findings derived by the UCISE framework concerning the market fit of our three examined SEs indicate that alternative SE providers like *Gigablast* and *DuckDuckGo* cannot keep up with *Google's* performance regarding its major strengths in *search result quality* and *ease of use*. Nevertheless, there are considerable differences between the two examined alternative competitors in terms of the performance gap to the market leader. Whereas the performance differences of *Gigablast* and *Google* are considerably large and therefore could explain their different market shares (Statista 2020), *DuckDuckGo* realizes a much more similar performance to *Google* (see Figure 5, d). The major differences of *DuckDuckGo* and *Google* arise from the three most important criteria. When considering *Google's* dominant market share (Statista, 2020), even *DuckDuckGo's* much stronger performance in terms of *privacy protection* does not seem to overcompensate its slight deficit of performance regarding the two most important criteria *search result quality* and *ease of use*.

This may be caused by two possible reasons. A first possibility is that, the already explained "privacy paradox" manifests in the results. Alternatively, this disproportional impact of criteria could lead to the assumption that the underlying Likert-scale for the ranking of the perceived criteria importance is not equidistant. In other words, it indicates that the difference between rather important and important is not the same as the difference between important and extremely important. This refers to a well-known discussion among researchers with two schools of thoughts where one claims the Likert-scale to be an ordinal scale and the other insisting on it to be an interval scale (Joshi et al., 2015). With the already mentioned findings of Prüfer et al. (2020), this once again supports the conclusion that performance deviations in more important criteria play a disproportionately large role for the users. This topic is taken up again in the Conclusion.

The findings about the user preferences yield three implications for SE providers in practice. First, in order for a SE provider to match the users' priorities, it is essential to focus on three distinctive criteria. Thereof, the main focus should be a providers' *search result quality*, as it is the only criterion that is perceived as extremely important by the users. Subsequently, an emphasis should be laid on the *ease of use* which is

perceived as rather important to important. *Privacy protection* is ranked with almost the same importance as *ease of use* and therefore should not be neglected by providers either. Nevertheless, when putting this result in the context of *Google's* major weakness in *privacy protection* and also taking into account the "privacy paradox", it follows that providers may overcompensate a weak performance in *privacy protection* with a strong to very strong performance in the two most important criteria. Still, especially for *Google's* competitors a credible focus on *privacy protection* should not be neglected and might be a differentiating criterion to the market leader due to its rather high importance.

Second, in regard to the market fit, it can be concluded that small providers may neglect strong performances in the seven criteria that are perceived as least important by the users, as this only yields in useless strengths. Especially when alternative providers face the problem of sparse resources the focus on more important criteria has a by far bigger impact on the market fit and user preferences.

Finally, when considering the dominating importance of the *search result quality* together with the findings of Prüfer et al. (2020) about the direct influence of data aggregation on the quality of search results, it can be concluded that there is a huge entry barrier for small competitors. Due to *Google's* high market share and thus large access to user data, it may be a considerable challenge for competitors to achieve an equally strong performance in terms of *search result quality*.

In addition to useful findings for practice, we find that this work also helps at expanding the existing, scientific knowledge base about user preferences in web search. In this regard, we see three main implications for research. By the derivation of ten criteria for SE differentiation from the users' perspective, this work provides an extensible foundation for user-based evaluation of SEs and insights into evaluation constructs from the users' point of view. Thereby, the findings emphasize that in order to understand a users' evaluation construct, it is important to holistically focus on the socio-technical nature of SEs instead of isolated technical aspects. Furthermore, by providing findings about user-perceived importance of the criteria base, this work may contribute to existing and future explanations of market imbalances, failures or other market-related phenomena when assessing the SE market. Finally, the synthesized UCISE approach, though being highly exploratory, may illustrate an initial starting point for the other evaluations of IS from user perspective. This approach consists of four steps containing a criteria derivation from the eye of the user, an assessment of their importance, an evaluation of the examined ISs in regard to the derived criteria and finally the comparison of the results in a visual UCISE framework that helps at comparing different systems and visualizing their market fit.

Nevertheless, the exploratory nature of the research approach leads to the fact that it naturally is not without limitations. First of all, the provided findings within this work are limited to the scope of SE research from users' perspective since the criteria are highly context-dependent and cannot serve as general criteria to evaluate IS. The UCISE therefore needs to be adapted to the respective research focus. Additionally, the results of our work only cover a first insight into the user-perceived importance of the criteria base that is limited to the description, as provided in Table 2. Therefore, the findings require careful abstraction if they are taken into account when explaining the current SE market imbalance. It should be clear to the reader that there are many additional (external) influencing factors that yield market success. These include among others, the pre-installation of SEs on the respective user devices, the users' force of habit, familiarity with other ISs in use or network effects and much more. Additionally, as stated by Prüfer et al. (2020), there is a vicious circle that leads to a self-enforcing effect of market dominance. This is caused by the better access of market leaders to user data which yields in stronger performances in regard to a SE's *search result quality*.

Further limitations of our work arise from possible distortions that were introduced by the exploratory nature of our methodological approach. Although it was tried to keep the influences to an absolute minimum, it needs to be mentioned that regardless of all efforts made, it must be assumed that there is a remaining bias especially in the first three parts of the UCISE approach. Firstly, the derivation process of the ten final criteria inevitably introduces a certain bias towards the authors perception and also remains to be dependent on the respective IS sample itself. Additionally, this step might introduce a bias that influences the results of the second part of the methodological approach. It is to be expected that the design of the descriptions of the ten criteria imposes a bias towards the survey participants' importance ranking. Furthermore, when taking a closer look at the demographic composition of the survey participants, it became evident that, with an average age of 33.37 years across all survey participants, our survey population is about 8 years below the average age in the European Union and consists of more academics (Eurostat

2019). This suggests the existence of a slight bias introduced by our selection of communication channels. Nevertheless, when viewing the survey results in regard to the participants' primarily used search engine, the shares of providers among the participants largely resembled the actual market shares of (Statista (2020), e.g. Google was stated primary SE by 85%). Therefore, we conclude that the results of this step should still be representative for a more general population. Additionally, the results in criteria importance of the introduced survey are also confirmed by the results of Prüfer et al. (2020), who managed to obtain a more diverse population and 821 participants. Subsequently, in the third part of the created UCISE approach it needs to be noted that the performance score per SE is strongly dependent on the statements that are provided to the assessing experts as well as on the experts themselves and their expertise.

Despite the limiting factors, each step of the chosen approach was selected carefully. The synthesis of the UCISE approach as well the findings derived by its application to SEs remain to be more of an initial step towards insights on the overlap of user preferences with search engine performances. Therefore, the reader should carefully reflect and abstract the provided insights as there remains the challenge of further objectification and bias minimization.

Conclusion and outlook

Since SEs are ISs incorporating both social and technical aspects, single scientific metrics with a technical focus lack meaning for their multidimensional evaluation from the untrained users' perspective. Therefore, to explain the current market imbalance and concentration of power among a few SE providers, it is essential to focus on the user perspective and preferences during contextual interaction with the whole socio-technical system. To approach the user-centered evaluation of SEs, the aim of this work was to carve out distinctive criteria that are decisive for a users' SE selection on the one hand *(RQ1)*, and to assess the performance of selected providers in relation to the current market leader *Google* on the other hand *(RQ2)*. In order to provide a methodological approach that allowed answering this question a User Centered Information System Evaluation UCISE was created and applied.

There are three key findings that follow from our research. First of all, SE evaluation from user perspective is multidimensional and we suggest an initial set of ten criteria to gain insights into the users' SE selection. Second, based on our deployed survey, we find only the three criteria *search result quality*, *ease of use* and *privacy protection* to be perceived as especially important in terms of user preferences. Thereof, users attach a particular importance to the *search result quality*. Third, according to our evaluated set of SEs, *Google* is able to score the best in the two criteria that are perceived as most important by the users and can thus overcompensate its major weakness in terms of *privacy protection*. With this performance, *Google* is able to dominate the other examined competitors *DuckDuckGo* and *Gigablast*. This could be one of many factors that helps at explaining, *Google's* market dominance and thus the large imbalance on the current SE market.

As our work is a first exploratory research approach for identifying possible reasons for *Google's* dominant market position, there are several possibilities of how our research could be continued. As presented in Methodology section, multiple precautions were taken to minimize the extent to which a bias could interfere with this work. Nevertheless, as stated in the Discussion, it must be assumed that there is a remaining influence of distortion due to the exploratory nature of the UCISE research approach. Thus, future research might be necessary to further minimize distorting influences.

It is essential to mention that, if a user perceives a SE along multiple dimensions, then all of these dimensions must be evaluated when holistically assessing the system. If this is not the case, then any conclusions drawn from the evaluation may themselves be subject to bias stemming from selection of certain measures (Crudge and Johnson 2004). As mentioned in within the Discussion, the online survey gave participants the possibility to state SE characteristics that were important to them. Although, this did not yield in identifying major new criteria, some suggestions were made by the survey participants that indicate the need for further refinement of the criteria-base in future research. A first suggestion in this direction would be the repetition of the research described in Crudge and Johnson (2004) with the input of a SE set that reflects the current SE market (like the one used within this work) to verify the completeness of our derived criteria list or identify additional ones.

Additionally, since deviations of a SE from the perfect fit to user preferences seem to be more relevant in more important criteria than in less important criteria, this raises the question, whether the underlying

importance scale of user preferences contains an exponential nature. Our findings in the last part of the Results indicate, that deviations from the perfect fit seem to be more relevant the further right the criterion is placed, which is in line with the findings of Prüfer et al. (2020). Therefore, it might be of interest to continue research concerning the nature of the gradation of user-perceived importance regarding the differentiation criteria.

The expert set presented in this study consisted of 6 experts only. The small sample size of expert opinions is susceptible to extreme individual opinions or misconceptions among the experts. By increasing the number of participants in the expert assessment, one could reassure robust and reliable ratings concerning the performance of examined SEs, leading to an increased reliability of the performance ratings. As a final suggestion for further refinement of the expert assessment we additionally suggest increasing the number of statements per criterion that lead to the experts' agree- or disagreement in order to increase the robustness of the SE performance score. Alternatively, there might generally be other suitable approaches to assess the performance of SEs in regard to the user-based criteria.

Finally, for further continuation of our work it could be of interest to enlarge the set of examined SEs as this will yield more insights concerning possible explanations for the current market imbalance. Therefore, we suggest the expansion of the examined SE sample, for example by considering larger competitors of *Google*, such as *Bing*, *Baidu* or *Yandex* to assess how far off user perceive their strengths and weaknesses in relation to the market leader. In doing so, one would be able to gain further insights towards the current SE market situation and *Googles* uncontested reign.

References

Argenton, C., and Prüfer, J. 2012. "Search Engine Competition with Network Externalities," *Journal of Competition Law \& Economics* (8:1), pp. 73–105.

Azzopardi, L., Thomas, P., and Craswell, N. 2018. "Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, New York, NY, USA: Association for Computing Machinery, pp. 605–614.

Bhattacherjee, A. 2012. *Social Science Research: Principles, Methods and Practices*, University of South Florida.

Boiko, Y. 2018. "Methods of Forming an Expert Assessment of the Criteria of an Information System for Managing Projects and Programs," *Technology Transfer: Fundamental Principles and Innovative Technical Solutions* (2), pp. 9–11.

Brooke, J. 1995. "SUS: A Quick and Dirty Usability Scale," *Usability Eval. Ind.* (189).

Butterfield, A., and Ngondi, G. E. 2016. *A Dictionary of Computer Science*, Oxford Quick Reference, Oxford University Press.

Castellà-Roca, J., Viejo, A., and Herrera-Joancomartí, J. 2009. "Preserving User's Privacy in Web Search Engines," *Computer Communications* (32:13), pp. 1541–1551.

Chang, S.-J., van Witteloostuijn, A., and Eden, L. 2010. "From the Editors: Common Method Cariance in International Business Research," *Journal of International Business Studies* (41), pp. 178–184.

Clemons, E. K. 2019. "Power and the Potential for the Abuse of Power in Online Gateway Systems: An Analysis of Google," in *New Patterns of Power and Profit*, Springer International Publishing, pp. 135–160.

Clemons, E. K., and Wilson, J. 2016. "Modeling Competition in Mandatory Participation Third Party Payer Business Models: The Complex Case of Sponsored Search," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 5210–5219.

Commission, E. 2010. *Antitrust: Commission Probes Allegations of Antitrust Violations by Google*. (retrieved from: https://ec.europa.eu/commission/presscorner/detail/en/IP_10_1624; last accessed: July 31, 2020).

Crudge, S. E., and Johnson, F. C. 2004. "Using the Information Seeker to Elicit Construct Models for Search Engine Evaluation," *Journal of the American Society for Information Science and Technology* (55:9), Wiley Online Library, pp. 794–806.

Dritsa, K., Sotiropoulos, T., Skarpetis, H., and Louridas, P. 2020. "Search Engine Similarity Analysis: A Combined Content and Rankings Approach," in *Web Information Systems Engineering -- WISE 2020*,

Z. Huang, W. Beek, H. Wang, R. Zhou, and Y. Zhang (eds.), Cham: Springer International Publishing, pp. 21–37.
European Union. 2008. "Council Directive 2008/114/EC on the Identification and Designation of European Critical Infrastructures and the Assessment of the Need to Improve Their Protection.," *Official Journal of the European Union* (L345), pp. 78–83.
Eurostat. 2019. *Median Age over 43 Years in the EU*. (retrieved from: https://ec.europa.eu/eurostat/de/web/products-eurostat-news/-/DDN-20191105-1; last accessed: July 31, 2020).
Grant, R. M. 2005. "Contemporary Strategy Analysis," *Notes* (7th ed.), Chichester: Wiley.
Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. 2013. *Measuring Personalization of Web Search*, WWW '13, New York, NY, USA: Association for Computing Machinery, pp. 527–538.
Jessup, L. M., and Valacich, J. S. 2008. *Information Systems Today: Managing in the Digital World*, (Vol. 3), Pearson Prentice Hall Upper Saddle River, NJ.
Joshi, A., Kale, S., Chandel, S., and Pal, D. 2015. "Likert Scale: Explored and Explained," *British Journal of Applied Science & Technology* (7), pp. 396–403.
Kelly, G. A. 2003. *The Psychology of Personal Constructs. Vol. 2*, London: Routledge.
Laudon, K. C., and Laudon, J. P. 1999. *Management Information Systems*, Prentice Hall.
Norberg, P., Horne, Dan, and Horne, David. 2007. "The Privacy Paradox: Personal Information Disclosure Intentions Versus Behaviors," *Journal of Consumer Affairs* (41), pp. 100–126.
Orlikowski, W. J. 2007. "Sociomaterial Practices: Exploring Technology at Work," *Organization Studies* (28:9), pp. 1435–1448.
Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L. 2007. "In Google We Trust: Users' Decisions on Rank, Position, and Relevance," *Journal of Computer-Mediated Communication* (12:3), pp. 801–823.
Podsakoff, P., MacKenzie, S., Lee, J.-Y., and Podsakoff, N. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* (88), pp. 879–903.
Prüfer, J., Inge, G., Klein, T., Kurmangaliyeva, M., and Prüfer, P. 2020. "Indikatorik Und Governance-Ansätze Zur Analyse Und Regulatorischen Gestaltung Datenbasierter Märkte in Deutschland. *Abschlussbericht Des Forschungsprojekts 11/19 Für Das Bundesministerium Der Finanzen*."
Ruch, T. J., Schmidt, N.-H., Decker, J., and Kolbe, L. M. 2011. "Ecosia - Who Cares about a Green Search Engine?," in *AMCIS*.
Selm, M., and Jankowski, N. 2006. "Conducting Online Surveys," *Quality and Quantity* (40), pp. 435–456.
Statista. 2019. *Chart: How Many Websites Are There?* (retrieved from: https://www.statista.com/chart/19058/how-many-websites-are-there/; last accessed: July 31, 2020).
Statista. 2020. "Dossier Zum Thema Suchmaschinen." (retrieved from: https://de.statista.com/statistik/studie/id/6997/dokument/online-suche/; last accessed: July 31, 2020).
Sunyaev, A. 2020. "Introduction to Internet Computing," in *Internet Computing*, Springer International Publishing.
Teevan, J., Dumais, S. T., and Horvitz, E. 2010. "Potential for Personalization," *ACM Trans. Comput.-Hum. Interact.* (17:1), New York, NY, USA: Association for Computing Machinery.
Varian, H. R. 2016. *Grundzüge Der Mikroökonomik*, Berlin, Boston: De Gruyter Oldenbourg. (retrieved from: https://www.degruyter.com/view/title/511799; last accessed: July 31, 2020).
Weimann, J., and Brosig-Koch, J. 2019. "Externe Validität," in *Einführung in Die Experimentelle Wirtschaftsforschung*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 29–41.
Weinberg, G. 2020. "Gabriel Weinberg's Answer to: What Does Google Know about Me? - Quora." (retrieved from: https://www.quora.com/What-does-Google-know-about-me/answer/Gabriel-Weinberg, accessed December 3, 2020; last accessed: July 31, 2020).

Appendix

Derivation of the set of 78 SEs

The final set of 78 SEs that was used to derive the 10 final SE-evaluation criteria was compiled in the following way. Using the databases of https://searchengine.party/, https://searchenginearchive.com/ and https://en.wikipedia.org/wiki/Search_engine a compilation of over 500 SEs could be derived. This compilation was then processed to a final set of 78 SEs using the following rules:

- Exclusion of SEs that were not available via a public domain
- Only included SEs that do website indexing (No people or product SEs)
- Exclusion of SEs that are limited to certain topics (e.g. vertical SEs), geographic regions or users (e.g. children)
- Exclusion of inactive SEs or SEs that are too old (e.g. copyright update older than 3 years)
- Exclusion of sister domains
- Exclusion of SEs that don't offer language support for neither English nor German
- Exclusion of SEs that only offer own start page and redirect to other sites for results
- Exclusion of SEs that were criticized to be malware or spyware
- Exclusion of sites that were permanently on maintenance during the research period

The final set of 78 SEs is available from the authors on request

The Theory of Social Comparison in Mobile Health Research

Digital Health, Winter Term 19/20

Kai Alexander Binder
Master Student
kai.binder@student.kit.edu

Maximilian Rauh
Master Student
maximilian.rauh@student.kit.edu

Marius Schröter
Master Student
urecq@student.kit.edu

Abstract

Background: Chronic diseases like type 2 diabetes that are often related to weight, and unhealthy lifestyles are becoming more common with some researchers even speaking of epidemics. mHealth technology combined with social comparison might help many people to become more active and healthier.

Objective: This literature review aims to research the current use of the social comparison theory in mHealth research.

Methods: A systematic literature review was conducted in six electronic databases (ProQuest, Scopus, PubMed, ScienceDirect, ACM Digital Library and IEEExplore). After applying the inclusion and exclusion criteria, 23 qualified papers were evaluated.

Results: The majority of research was conducted in the context of fitness apps with the purpose of fighting obesity and physical inactivity. The mHealth research is largely carried out on smartphones due to low cost and simplicity compared to activity trackers. Therefore, social comparison theory was mainly applied as a persuasive behaviour change technique (e.g. leaderboards).

Conclusion: We conclude that social comparison features have a high potential in mHealth applications but are still in a trial phase regarding the extent and the possibilities to give the user maximum enjoyment and satisfaction. However, the observed users felt more motivated by other behaviour change techniques such as goal-setting, feedback and self-control.

Keywords: social comparison, mHealth, literature review, persuasive behaviour

Introduction

Motivation

Nowadays, more and more people are seeing the importance of being active and maintaining a healthy lifestyle. Public health education contributed to the fact that the belief of one's lifestyle affecting the longevity is ingrained in Western societies (Newsom et al. 2005). In the beginning of every year, the internet, TV and magazines are flooded with advertisements for weight loss, fitness apps and health products. Furthermore, preventable chronic diseases like type 2 diabetes that are often related to weight and unhealthy lifestyle are becoming more common and some researchers are even speaking of epidemics (Wilding 2014). This shows that many people are actively looking for better ways to improve their health while others are overwhelmed or need individual advice to make their life and lifestyle healthier.

Many companies have realized the enormous potential of catering health services to these patients and customers. In particular, the proliferation of electronic health applications and smart devices (Statista 2019a) is facilitating their business endeavours. They make it easier than ever for consumers and patients to have access to health advice, fitness motivation or health tracking at all times (Meng et al. 2019). "The use of information and communication technology for health" is generally referred to as eHealth, with mobile health or mHealth specifically meaning "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants (PDAs), and other wireless devices" (WHO 2011). However, not only the WHO acknowledges the future relevance of this health segment. The mHealth market is projected to grow over 25% the next few years to reach a volume of more than USD 150 billion (Grand View Research 2018). Especially preventive healthcare and technical innovations are drivers for the increasing numbers. With the soaring amount of mHealth apps and their immediate availability over smartphones, patients and their providers are enabled to continuously monitor, discuss and improve the patient's health. Particularly, monitoring services, diagnosis services and healthcare system strengthening services are predicted to play major roles in the future (Grand View Research 2018). However, the mHealth market does face some challenges. Some of these challenges are more obvious such as the technological characteristics of the used devices like the computing power or screen sizes. In addition, many users are concerned about privacy and security of the devices and how the applications processes user data (Consolvo et al. 2006). The provided data is often highly sensitive and mHealth apps are attractive to getting hacked (S. Bhuyan et al. 2017).

All this potential and wide-spread usage begs one big question: How do you effectively design and develop mHealth devices and apps? Since mHealth involves both technology and the users' health, designers and developers need to keep technology acceptance and individual (health) behaviour in mind. Applications need to have a reasonable use case and be simple enough to be usable for even tech novices (Gurupur and Wan 2017). There are several factors that contribute to a successful usage. Health consciousness for example can moderate the credibility of mHealth applications (Meng et al. 2019). Gamification elements like leaderboards in fitness apps have a positive impact on users' attitude to use mHealth applications. Especially the competitive factor of such elements contributes to increased physical activities (Wu et al. 2015). Partially responsible for this notion is the social comparison between app users. The social comparison theory goes back to Festinger (Festinger 1954) who explained the competitive behaviour of humans as a socio-psychological process. This process drives people to self-evaluate by comparing themselves with others in order to get a feeling where they stand with their performances. The social comparison is not anything particular new and has been applied in a variety of setting over the last 60 years. However, since the mobile health market has only developed over the last 5-10 years, the application of social comparison in this research area is very sparsely. Nevertheless, the potential of using this comparison process to better cater mHealth devices and applications is undeniable. This leads to our research question:

RQ: How is the social comparison theory applied in mHealth research?

Objectives

The goal of this paper is to examine the current use of the social comparison theory in mHealth research to show research gaps and future research areas. In order to fulfil this goal, we perform a systematic literature review to get an overview of existing mHealth research. We then analyse relevant publications in order to highlight where the social comparison theory is already used in mHealth studies.

Structure

The next chapter describes the theoretical background of this paper. First, the social comparison theory is explained in greater depth. Additionally, mHealth and some mHealth research papers are outlined to show what typically is researched in this topic. Chapter three then presents the methodological approach of this paper. The ideal-typical approach for a systematic literature review is set out and we explain how we collected the papers and what inclusion and exclusion criteria we used for the selection process. Additionally, we disclose how we analysed the papers. In chapter four we present the selected papers that resulted from the literature review. These papers are then discussed in chapter five. Before we conclude our work by summarizing the findings in chapter seven, the limitations of our work are presented in chapter six.

Theoretical Foundation

Social Comparison Theory

In this segment, the theory of social comparison which Festinger established in 1954 will be presented in more detail. Social comparison is a pervasive social phenomenon (Suls et al. 2002) which can be defined as an individuals need to compare the status of the self with other individuals or groups in order to evaluate or improve the own performance (Hoorens and Damme 2012). In the words of Festinger (Festinger 1954), the drive for self-evaluation is based on comparison with other individuals.

Social comparison can be characterized by the direction in which it is performed which can be either upward or downward (Latané 1966). This means that individuals compare themselves to others whether they are better-off or worse-off (Buunk et al. 1990) in a variety of dimensions, such as status, capability and achievements (Shang et al. 2012). Although the target dimension to be compared can have both positive and negative effects, it is not intrinsically linked to either direction (Wu et al. 2015). “Rather, both upward and downward social comparisons are capable of generating positive or negative affective responses, depending on which aspect of the comparison is focused on” (Buunk et al. 1990). Indeed, a comparison results in either positive or negative consequences depending on whether the comparing individual represents themselves and the individual to be compared as the same or different, i.e., identification or contrast (Bailis and Chipperfield 2006).

Research has shown that individuals tend to compare themselves with others slightly better-off (Buunk and Gibbons 2007; Wheeler 1966), which supports Festinger’s hypothesis of an unidirectional upward drive in social comparison (Festinger 1954). Furthermore, the comparison to individuals or groups better-off provides two signals. On the one hand, individuals know that they are not performing as good as others but on the other hand it is possible for them to improve their current status (Wu et al. 2015). The exposure to upward targets provokes self-enhancement and can result in a motivational boost to become as good as the individual or group compared (Lockwood and Kunda 1997; Wheeler 1966). On the contrary, downward comparison gives comparers the information that they are not as bad off as others but at the same time they get the signal that there exists the possibility to get worse (Buunk et al. 1990). The positive outcome of a comparison to individuals or groups worse-off can be mood improvement by knowing you are better than others (Gerrard et al. 2005). “In addition, many correlational studies in populations facing some kind of threat have shown that well-being (…) is positively associated with perceiving oneself as better off than others (Buunk and Gibbons 2007).

As a conclusion from either upward or downward direction, it can be said that comparers who focus on the positive aspects of the same information may feel better, while those who focus on the negative aspects may feel worse (Wu et al. 2015). So how you feel in response to the information that another person is better or worse off than you may depend on how you interpret the information (Buunk et al. 1990). Regardless of either upward or downward comparison, Festinger (Festinger 1954) claims that when comparing a large number of individuals, one chooses the individual that is closest to his or her own performance. Particularly in the upward comparison of the self to individuals who are far from one's own performance, self-deflation may occur because the better one seems unattainable and one's own performance is perceived as fixed and unimprovable (Lockwood and Kunda 1997). Brown et al. (Brown et al. 2007) found a further characteristic feature of the theory of social comparison in their investigation of two different comparison environments. In their study, the focus is on competitive and cooperative environments. The researchers' conclusion is that competitive environments are likely to favour contrasting effects due to their focus and reward of individual performance, while cooperative environments may favour assimilation effects. In other words, competitive contexts promote a mindset in which the emphasis is on the self, leading to contrast effects, while cooperative contexts promote a mindset in which the emphasis is on the group, leading to assimilation effects.

Mobile Health (mHealth)

Due to the ongoing worldwide growth of mobile phone subscriptions and the penetration of smartphones, various mHealth initiatives are getting launched and established around the globe (ITU 2019). In 2011, the World Health Organization (WHO) elaborated a survey in order to determine the status of mHealth activity in the member states. The largest activity of mHealth initiatives were observed in the categories of health

call centres and emergency toll-free telephone services (WHO 2011), whereby the majority of activities was limited to small-scale projects mostly in pilot stage (WHO 2011). High-income countries reported more initiatives in general (WHO 2011) as well as more complex projects in the field of patient monitoring (WHO 2011).

Three major challenges can be derived from existing literature, namely: More research in order to provide policy makers a solid scientific foundation, cross-functional collaboration for the integration in existing health care systems and the fulfilling of high privacy and data security standards (WHO, 2011). In order to address the lack of formal evaluation in mHealth initiatives (WHO 2011) the US National Institute of Health organized the "mHealth Evidence Workshop" where researches came together and discussed the current evaluation standards (Kumar et al. 2013). This study showed methodologic issues in evaluating programs and endorses for further transdisciplinary scientific research in this filed (Kumar et al. 2013).

This is also supported by researchers from Columbia university suggesting a wide systematic approach with public and private partnerships that take all stakeholders into account in order to implement mHealth projects successfully (Mechael and Searle 2010). Regards to developing countries a study by Alam et al. (Alam et al. 2020) shows that the perceived reliability is a crucial factor for the adoption of mHealth services in Bangladesh. This is in line with the WHO's statement that public awareness campaigns must be conducted (WHO 2011). Considering these points, the WHO will support establishing global standards by providing a "National eHealth Roadmap Development Toolkit" to the member states (WHO 2011).

Data security and privacy is another concern that needs to be handled by policy makers within mHealth (WHO 2011). Although people are willing to share personal data for public good nowadays, a study by Raji et al. (2011) is showing that people were "most concerned about revealing conversation, commuting, and inherently private psychological states" (Raij et al. 2011). Thus, there is a requirement for policy intervention as well as secure hardware in mHealth technology. Therefore, researches from Dartmouth college were proposing and applying an adaptive security model for mHealth sensors in order to guarantee data security also for low-power mHealth sensors (Mare et al. 2011).

Methodology

In this chapter, the research method of a systematic literature review is explained in further detail. Since the mHealth research is an emerging area, it is of great importance to get a concise overview of existing literature, especially when looking for the application of a specific theory like the social comparison theory. The first section deals with how we performed the literature review while section two goes into more detail regarding the data analysis. This two-step process is illustrated in Figure 1.

Systematic Literature Review

In order to collect all relevant mHealth literature that incorporates the social comparison theory, we conducted a systematic literature review following the approach of Webster and Watson (2002) in order to get a comprehensive understanding of the research matter. Systematic reviews have the advantage of "evaluating and interpreting all available research relevant to a [...] phenomena of interest" (Kitchenham and Charters 2007). Concerning the data collection, we concentrated on online data bases with a focus on information systems. However, since mHealth research is also often concerned with health-related issues, medical data bases were also taken into consideration. This led to the following databases:

- ProQuest
- Scopus
- PubMed
- ScienceDirect
- Association for Computing Machinery (ACM) Digital Library
- IEEExplore

For finding relevant literature we tried to keep the search terms as generic as possible, since this literature review aims to get a very comprehensive overview of social comparison in mHealth. Mainly, our search strategy consisted of two parts: the social comparison theory and the mHealth area. For the search term mHealth, we opted for applying some variety and subtopics of mHealth on the search in order to hit as many relevant articles as possible.

This led to the search string:

> *Social comparison* AND *(mHealth* OR *mobile health* OR *gamification* OR *fitness apps OR fitness applications OR health apps OR health applications OR mobile phone)*

The search string was then applied on the database attributes title, abstract and key words which made sure that the application of the social comparison theory was actually a key part of the research study and not just a mere mention in the main body of the text. In addition, only peer reviewed articles and conference proceedings were taken into consideration for ensuring a certain qualitative standard and to reflect the latest findings respectively. The online database search resulted in a total of 151 articles. After sorting out 12 duplicates, we gathered 139 articles for further analysis. The titles and abstracts of these articles where then read by the three authors of this paper with at least two authors examining the same article. By examining all articles found and applying the following inclusion and exclusion criteria (Table 1), we selected 17 articles which were then analysed further in a second step. During analysis, 6 articles were added through backwards search. This gives a total of 23 articles.

Inclusion Criteria	**Exclusion Criteria**
Peer Reviewed or Conference Proceedings	Grey literature
Academic Research	Practical oriented
Social comparison as major theory	Social comparison minor theory
Focus on mHealth	About general health/eHealth topics
English language	Non-English papers
Accessible through university means	Not accessible with university license
	Literature Reviews

Table 1. Inclusion and Exclusion Criteria for the Article Selection

With at least two authors reading the same article we made sure that certain biases were kept to a minimum. However, different researchers can also result in a potentially different understandings of the inclusion and exclusion criteria. In order to assure an appropriate precision of the selection, we used inter-coder reliability score from Miles and Huberman (Miles and Huberman 1994). According to their proposal, the reliability score, calculated by the ratio of the number agreements and the total number of agreements and disagreements, should be approaching 0.90. With our coding, we achieved a final score of 0.87 which we deem as adequate. In case of a disagreement, the third researcher who has not read the article decided whether or not to include the paper into the analysis.

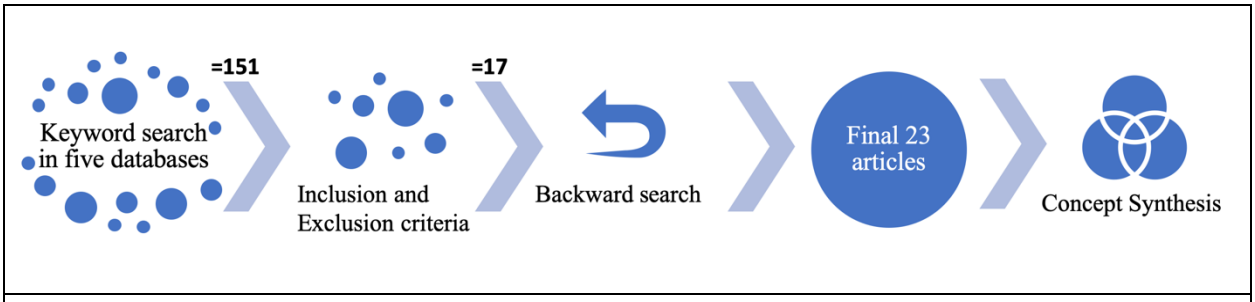

Figure 1. Systematic Literature Review with Data Collection and Analysis

Data analysis

With the 23 papers selected, we then proceeded to analysing the contents of each article. We followed the approach of Webster and Watson (2002). They propose a concept-centric analysis because in contrast to an author-centric approach, the literature review does not result in a mere summary of the articles but rather synthesises the contents and gives a better grasp of key topics that researchers deem as relevant (Webster and Watson 2002, p. xvi). For the reason of conciseness, we still opted for an author centric presentation in chapter four, *Results*. Within our overview, we used eight categories to give further analysis insights into the papers, namely

- study design (whether or not it is a qualitative or quantitative study with further information about the process),
- participants (information about the composition of the participants and their characteristics),
- setting (in which the mHealth aspect is measured),
- approach (how the participants where influence in their behaviours),
- aim (what was the goal that the mHealth application should achieve),
- outcomes (what was achieved),
- application of social comparison (which mechanism induced social comparison),
- type of mHealth technology (how did participants receive information about their health).

Results

The characteristics of the 23 studies are summarized in Table 2. For ease of reading, this section uses references to the indices used in Table 2. Seven studies were conducted in the USA[1, 3, 10, 11, 14, 15, 21], three in Canada[16, 17, 18] and in the Netherlands[2, 8, 13], two in Australia[6, 19] and in Singapore[20, 22], one in Belgium[5], China[7], Taiwan[9], and Israel[23], and two in multiple countries[4, 12]. The setting included daily routines[3, 9, 10, 12, 13, 15, 21, 23], physical and sport activities[4, 8, 11, 20, 22], eating habits[6, 14], leisure time[8] as well as smoking habits[12] and took place in either schools[2, 10, 21], sport clubs[13] or in the participants private life[3, 4, 5, 6, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23]. The number of participants ranged from 10 in a study on the influence of competition on physical activity[4] to 1743 middle school students, in which the study also examined the influence of competition on physical activity[21]. Of the 23 studies, 16 distinguished between male and female participants[2, 3, 4, 5, 6, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 23], and it is noticeable that these studies have a higher proportion of female (66%) than male participants. Two of the 23 papers are conceptual[1, 8], while seven use quantitative[2, 4, 5, 9, 13, 17, 18], six qualitative[7, 11, 12, 19, 20, 22] and eight papers use a combination of both quantitative and qualitative research methods[3, 6, 10, 14, 15, 16, 21, 23]. The promotion of physical health was the primary aim of eight studies[3, 4, 8, 10, 15, 19, 21, 23]. An increase in the consumption of vegetables was the aim of two studies[6, 14], and one aimed to reduce smoking[12]. Four studies focused on how mHealth applications should be designed[1, 2, 5, 9] and differed in their target groups. Another two studies compared smartphone apps and activity trackers[7, 11] and wanted to summarize these technologies in terms of techniques for behaviour change, such as social comparison. Six studies analysed the effects of social comparison methods on mHealth applications[13, 16, 17, 18, 20, 22], with three of them considering the susceptibility of participants to selected parameters[16, 17, 18]. The main mHealth tools used were smartphone or web apps[1, 2, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 20, 22, 23], wearable activity trackers[11, 19], a combination of the two[3, 10], or a fictitious fitness app[16, 17, 18]. Some studies had additional mHealth and technology intervention components, including motivational or persuasive text messages and dedicated social media groups[3, 6, 8, 9, 10, 13, 14, 16, 17, 19, 18, 23]. Six studies reported that the intervention was based on social comparison[1, 8, 13, 17, 20, 23]. A further seven studies alluded social comparison to behaviour change techniques[2, 6, 9, 10, 11, 16, 19], and nine studies had no clear theoretical basis[3, 4, 7, 12, 14, 15, 18, 21, 22]. The outcome of the studies is that participants considered social comparison in mHealth applications to be less important than change techniques in terms of goal-setting, feedback and self-control[2, 5, 17]. Nevertheless, the participants became more active, i.e. were more likely to achieve their activity goals[3, 23] and felt more motivated[12, 19]. However, this is mainly due to self-control rather than social comparison[10, 15]. Participants with a high social comparison orientation (SCO) benefit from such characteristics[1]. For example, participants placed high in a leaderboard compared themselves upward, thus increasing their physical activity[22]. This is an interesting finding, as 75% of the participants preferred to compare upwards[13]. However, the social comparison characteristics do not apply to the older generation, as they consider them indifferent.

Index	Author/Year	Method	Focus/Aim	Type of mHeatlh Technology
1	Arigo and Suls 2018	Conceptual	Determine how smartphone apps with social comparison features in a user centric way have to be designed	Apps on mobile phones
2	Belmon et al. 2015	Quantitative	Determine how young adults rate the use of certain behaviour change techniques in physical activity apps	Apps on mobile phones
3	Consolvo et al. 2006	Mixed Methods	Promote physical health and losing weight with more walking	Mobile phone with external pedometer
4	de Oliveira and Oliver 2008	Quantitative	Increase physical activity	App on mobile phones
5	DeSmet et al. 2019	Quantitative	Determine how young adults rate the use of certain behaviour change techniques in physical activity apps	Apps on mobile phones
6	Hendrie et al. 2019	Mixed Methods	Increase vegetable consumption	App on mobile phones
7	Huang and Zhou 2019	Qualitative	Clustering fitness-apps and determining the variety of dimensions regarding behaviour change techniques	Apps on mobile phones
8	Klein, Manzoora and Mollee 2017	Conceptual	Increase physical activity	App on mobile phones
9	Liao et al. 2018	Quantitative	Explore middle-aged adults' needs on functional features of mobile phone apps promoting physical activity	Apps on mobile phones
10	Lin et al. 2006	Mixed Methods	Increase physical activity	Pedometer and app on PC
11	Lyons et al. 2019	Qualitative	Increase physical activity	Activtiy Tracker
12	Maramis et al. 2019	Qualitative	Smoking behaviour modification with the help of social comparison	App on mobile phones
13	Mollee and Klein 2016	Quantitative	Determine the effects of both upward and downward social comparison	Web app
14	Mummah et al. 2016	Mixed Methods	Development of a theory-driven mobile app to increase vegetable consumption	App on mobile phones
15	Munson and Consolvo 2012	Mixed Methods	Increase physical activity	App on mobile phones

16	Oyibo, Adaji and Vaddileva 2019	Mixed Methods	Determine whether participants susceptibility to competition can be predicted based on their susceptibility to social comparison	Fitness app
17	Oyibo and Vassileva 2019a	Quantitative	Determine to which of six persuasive features in fitness apps participants are the most/least susceptible	Fitness app
18	Oyibo and Vassileva 2019b	Quantitative	Determine to which of six persuasive features in fitness apps participants are the most/least susceptible in respect to their culture	Fitness app
19	Tong, Colera and Laranjo 2018	Qualitative	Increase physical activity	Wearable
20	Wu, Kankanhalli and Huang 2015	Qualitative	Determine the impact of leaderboards on attitude and physical activity behaviour	App on mobile phones
21	Xu et al. 2012	Mixed Methods	Increase physical activity	Pedometer with web application
22	Zhou, Kankanhalli and Huang 2016	Qualitative	Determine the effects of social networking services on physical activity	App on mobile phones
23	Zuckerman and Gal-Oz 2014	Mixed Methods	Promote physical health with more walking	App on mobile phones

Table 2. Resulting Papers from the Literature Review

Discussion

Setting

The setting of the 23 qualified studies is distinguished in Table 2 and analysed systematically. As mentioned before, the majority of social comparison applications took place in a personal activity setup with regards to fitness activity. Two studies measured the influence of social comparison on personal activity in the most general way by analysing data from a mobile fitness app in combination with a wearable fitness tracker (Klein et al. 2017; Tong et al. 2018). Further papers targeted their research on a special group of people. Liao et al. (Liao et al. 2017) explored how the SC features will affect middle-aged adults, since a lot of studies have been conducted with younger people. In contrary one study by Xu et al. (Xu et al. 2012) was tailored to school students in order to increase their steps and physical activity comparison with other schools. Furthermore, a sedentary lifestyle as well as jobs are an influencing factor for chronic diseases Lin et al. (Lin et al. 2006) focused on promoting physical activity for people facing this lifestyle or job.

Two studies spotlighted the rise of opportunistic physical activity in their work. “These are where a person incorporates activities into her normal, everyday life to increase her overall level of physical activity” (Consolvo et al. 2006). Also, the research by a Media Innovation Lab distinguished that from *structured exercise* and investigated the impact of social comparison in consideration of both elements (Zuckerman and Gal-Oz 2014).

In addition, there are three papers dedicated to promoting running as a physical activity. Therefore a Singaporean study, targeted on fitness apps with social networking features, used objective data from

Runkeepers users' (Zhou et al. 2016). Besides this major app *Runkeepers*, an earlier study by this team used extracted running data from *Nike+* for the analysis (Wu et al. 2015). The persuasive features of social comparison were also applied in a running environment by an earlier study in 2008 (de Oliveira and Oliver 2008).

In regard to physical activity, three papers used a more systematic approach. One study provided the participants two fitness apps, one with sharing features and one without, whereby these focused on multiple sport activities in order to assess different behaviour change methods, such as social comparison (Munson and Consolvo 2012). The other two examined the design of a system for promoting physical activity. *Active2Gether* is one example for a systematic approach by creating an extensive app with multiple behaviour change techniques implemented (Klein et al. 2017). The second study assessed 13 electronic activity monitors with regards to different behaviour change techniques implemented (Lyons et al. 2014).

Besides all the research done to personal activity, there were also two studies focusing social comparison in the application of establishing a more vegetarian eating behaviour (Hendrie et al. 2019; Munson and Consolvo 2012). Both studies included an everyday tracking of vegetable consumption in a mobile app.

Only one study tackled a smoking behaviour change of the participants by implementing social comparison tools in a smartphone app (Maramis et al. 2019).

The research done by the University of Saskatchewan (Oyibo et al. 2019; Oyibo and Vassileva 2019a, 2019b) as well as the papers by Belom et al. (Belmon et al. 2015) and DeSmet et al. (DeSmet et al. 2019) investigated the users' susceptibility to certain behaviour change techniques (e.g. social comparison), thus no setting was applied.

Also the conceptual study by Arigo and Suls (Arigo and Suls 2018) and the comparison of fitness apps in Huang and Zhou's paper (Huang and Zhou 2019) did not apply a setting in accordance to the authors understanding.

Aim

A total of eight studies focused on physical health, precisely the promotion of physical activity. Three papers (Consolvo et al. 2006; Lin et al. 2006; Tong et al. 2018) justified their research explicit with the rise of overweight and obesity in modern societies. The authors citied WHO reports of global mortality in order to emphasise this issue. Today we can see an increase in the actuality of this reasoning, status 2016 the WHO reports "more than 1.9 billion adults, 18 years and older, were overweight" (WHO 2018). Further reasons for that phenomena can be found in an increased sedentary lifestyle and the availability of energy-dense foods (Lin et al. 2006). The study by Zuckerman and Gal-Oz is referring radically to western societies as "environments that promote physical inactivity" (Zuckerman and Gal-Oz 2014).

Others are just mentioning the sedentary lifestyle as a reason for promoting physical activity without being more detailed here (Klein et al. 2017; de Oliveira and Oliver 2008). One paper is targeting especially young adults' physical activity (Klein et al. 2017). Two papers are not carrying out further reasons for aiming on increasing objectives' physical health (Munson and Consolvo 2012; Xu et al. 2012).

Four studies are aiming on the design of mHealth applications with respect to the different behaviour change techniques implemented (Arigo and Suls 2018; Belmon et al. 2015; DeSmet et al. 2019; Liao et al. 2017). The behaviour change techniques have been structured in accordance with the taxonomy developed by Michie et al. (Michie et al. 2011). The effectiveness of the certain behaviour change technique is investigated as well as the question if they are targeted individually on the users' preference (Arigo and Suls 2018). This was conducted with adults in general (DeSmet et al. 2019) and with young Dutch adults (Belmon et al., 2015), whereby another aim here was to "explore whether these ratings [behaviour change technique preference rating] are associated with personality characteristics [...] and levels of physical activity" (Belmon et al. 2015). The enhancement of quality perception on behaviour change technique was part in Liao et al. (Liao et al. 2017) study on middle-aged users. They further aimed to investigate if the needs of middle-agers are offered (in app stores) yet and if physical activity or mobile phone self-efficacy influences the quality perception (Liao et al. 2017)

Six studies analysed the effects of social comparison methods on mHealth (Mollee and Klein 2016; Oyibo et al. 2019; Oyibo and Vassileva 2019a, 2019b; Wu et al. 2015; Zhou et al. 2016). One paper analysed the

effects on users' behaviour when they compare each other either downward or upward, while the participants indicated which method they prefer (Mollee and Klein 2016). Another study aimed at the impact of leaderboards since these are popular social comparison features in mHealth applications (Wu et al. 2015). The evaluation of effects occurring with social networking features like browsing other users' data and sharing own data was the aim of one paper (Zhou et al. 2016).

Oyibo et al. (Oyibo et al. 2019; Oyibo and Vassileva 2019a, 2019b) investigated in a total of three studies the effects of social comparison. First, the aim was to investigate if users' susceptibility to competition can be predicted based on their susceptibility for social comparison and if the results can be used for fitness app design (Oyibo and Vassileva 2019a). Second, they researched which is the most persuasive behaviour change feature (e.g. goal setting, reward, social comparison, …) and differentiated between users' currently exercising or intending to do so (Oyibo et al. 2019). Finally, in the third study they investigated the persuasiveness to certain features in regard to culture, whereby they observed a collectivist culture from Nigeria and an individualist culture from Canada and United States (Oyibo and Vassileva 2019b).

Two studies are aiming on increasing the individual's vegetable consumption in order to fight obesity (Hendrie et al. 2019; Mummah et al. 2016). In contrast to the beforementioned studies tackling physical activity, these papers are approaching the obesity by focusing on nutrition. While many commercial healthy eating apps are not developed scientifically, Hendrie et al. specifically want to tackle that problem by developing an app called *VegEze* that considers scientific and commercial validity (Hendrie et al. 2019). Also the research around the app *Vegethon* aims to integrate behaviour change theory as well as describing the iterative development of the app for further purposes (Mummah et al. 2016). Both of the papers are using the IDEAS framework for developing digital health interventions.

The last paper aimed to reduce smoking by developing an mHealth app with social comparison features. According to the authors this intervention contributes to the reduction of many diseases like the lack of physical activity (Maramis et al. 2019).

Types of mHealth Technology

The term mHealth means the health practice supported by mobile devices, such as smartphones, PDAs or other wireless devices. This of course enables continuous monitoring of patients while still giving them the freedom of normally pursuing their daily life. Nowadays, the most commonly used mobile device certainly is the smartphone. These little devices impact nearly half of the earth's population (Statista 2019a). One of the contributing reasons why smartphones became so popular is the ubiquity of mobile apps. Within our analysed journal articles, this prevalence can also be examined. Nearly all authors either used mobile applications to evoke healthier habits in their field studies or discussed their design. While the authors did not specify their reasoning why they wanted to use mobile apps for their studies, one reason very likely was the overall simplicity – both for developers/researchers and the users/participants. Even for older mobile phones, manufacturers provided development environments that made application development for developers/researchers much easier. For the modern systems Android and iOS, development kits are better supported than ever. This makes mobile apps an easy environment for researchers to implement design features and test out ideas.

Looking at the user side of applications, hardware and software improvements contributed to widespread usage and easier usability. Some studies from the late 2000s (Consolvo et al. 2006; Lin et al. 2006; de Oliveira and Oliver 2008) still needed to use pedometers or heartrate monitors to track participants' steps. These devices were either linked to the mobile phone (de Oliveira and Oliver 2008) or users even had to input the tracked steps from the pedometer into the application (Lin et al. 2006). For obvious reasons, this was very intrusive and unintuitive. Especially since the studies aimed at increased daily activity, wearing a pedometer-mobile phone combination in the office was perceived as "large and unattractive" (Consolvo et al. 2006). While additional hardware was necessary during the emergence of smartphones, modern mobile phones have received several sensors that can be used to draw conclusions about user activity. Zuckerman et al. (Zuckerman and Gal-Oz 2014) highlight the capabilities of more modern apps that have the potential to use sensor input and apply gamification features. Although more and better features definitely contribute to the prevalence of apps, one key aspect often emerges: ease of use. In their user centric design approach or their app *VegEze*, Hendrie et al. were told by their interviewees that their future app should be "quick and simple to use" (Hendrie et al. 2019). Generally, when it comes to developing the app, a user centric

design approach is highly recommended. While the outcomes of social comparison theory application will be discussed in more detail in the following segment, one example of the positive impact of social comparison on health behaviour shows the importance of user centric development. The creators of the anti-smoking app *QuitIt!* (Maramis et al. 2019) were able to successfully implement social comparison features to support smoking reduction of their users. Arigo and Suls (Arigo and Suls 2018) argue that one step towards more user centric design is for example the ability for the user to select his or her own preferred social comparison mechanism or data presentation method to elicit higher behaviour change.

As hinted before, a contributing factor to the dissemination of smartphones are their app eco-systems. Through app stores it is easier than ever to buy and download apps for any kind of use case. Every smartphone user has access to an enormous database of applications in order to find the one that caters the best to his or her needs. While mobile devices of any kind provide the necessary hardware, in the end it depends on the software to provide usability. This circumstance led to an vast business potential of a trillion US Dollars per year in the next five years (Statista 2019b). Due to these factors it is only logical for both researchers and practitioners to embrace apps for mobile health deployment.

Mechanisms to Elicit Social Comparison

As described earlier, social comparison theory can be found in everyday interactions of humans comparing themselves in order to position and evaluate their performances. Social comparison however does not only take place in the direct interaction between humans but can be evoked through different features or mechanisms. The following segment goes into further detail how and why the analysed studies apply these different mechanisms.

In the last chapter, a brief overview was given that showed that most studies focused on everyday activity or fitness activity in general. Very typical for these kind of scenarios in real-life competitions such as championships or the Olympics are leaderboards. They provide a good overview of the performances of the different participants, often with palpable metric statistics. In addition, they create a form of social pressure based on peers view on a person's image and social status (Wu et al. 2015). Many studies in our review use leaderboards to evoke social comparison (Hendrie et al. 2019; Mummah et al. 2016; Wu et al. 2015; Zuckerman and Gal-Oz 2014). *VegEze*, an app to promote vegetable consumption, opted for a deidentified leaderboard. It ranks different rewards which users can achieve and also shows how many active users are currently aiming to get the respective rewards (Hendrie et al. 2019). Even though their feedback in their user centric app design approach showed that only a third of users were hoping for some kind of social comparison feature, they still wanted to have rewards and leaderboards in their app to implement some kind of gamification. The similar app *Vegethon* also implemented a leaderboard that showed the first name of the user and compared them to other users with a similar baseline vegetable consumption to reduce potential discouragement (Mummah et al. 2016). In their user interviews however, some participants were considering anonymous leaderboards but also admitted that they would reduce accountability (Mummah et al. 2016). Wu et al. show that leaderboards do increase competitive climate and positively moderate the effect of social comparison (Wu et al. 2015). However, they also argue that for people with low self-efficacy, upward and downward comparisons do have negative influence on user's attitudes towards physical activity. In general, leaderboards might have issues regarding the motivation of underperforming users at the bottom of the ranking.

A possibility to solve this issue can be found in two of our selected papers. The app *Fish'n'Steps* (Lin et al. 2006) which aims at increased physical health, does also use rankings to elicit social comparison but disguises these rankings in form of depicted comic fish. Depending on the walked steps and whether participants reached their daily goals, their individual fish grows and shows happy or sad emotions. This has the benefit, that users receive higher motivation because on the one hand they want to see their fish grow and to be happy, on the other hand they want to have their fish looking as good or better than their fitness partners' fish (Lin et al. 2006). Whether or not this depiction was more effective in reaching increased physical activity will be discussed in more detail in the next segment. Another study that used a similar approach is the *American Horsepower Challenge* (Xu et al. 2012). Especially because the challenge is aimed at increasing physical activity of school children, it is of high importance to have proficient means to motivate the children. Xu et al. (2012) used a race between school buses to show the progress of each school in comparison to competing schools. As the study shows, pupils felt very competitive and

experienced high degrees of social comparison. A contributing factor might have been the ranking in form of school buses which is a more adequate depiction considering the context of young children in school.

Some studies even tried to mitigate the negative social pressure induced through social comparison by providing more supportive means. Munson and Consolvo (Munson and Consolvo 2012) explain their decision to include Facebook sharing due to the additional channel of possible social support from other close people that might not be participating in the fitness app but whose opinion still matters to the user. Tong et al. even implemented their own, app-specific "social forum" (Tong et al. 2018) to provide a network for support among users.

However, some studies that examine the preferred behaviour change techniques of young adults (Belmon et al. 2015) and more mature users (DeSmet et al. 2019; Liao et al. 2017) show that users of fitness or health apps do not necessarily expect or even want social support or social comparison. For young adults, goal setting features and self-monitoring is of higher importance than social support and social comparison (Belmon et al. 2015). For adults, even though some prefer social comparison features as long as they are not outperformed (Perski et al. 2018), the majority still values social comparison and social support less than the aforementioned self-monitoring and goal setting (DeSmet et al. 2019). Liao et al. (2017) confirm these findings for middle-aged users as well. In their study, they asked users about 52 design features for physical activity apps. Social comparison was categorized into the indifferent category (Liao et al. 2017). So even though many researchers and app developers use leaderboards, rankings or other forms of social comparison such as Facebook sharing or groups, many users and especially older users do not intend to conduct social comparison in their physical health. As mentioned in the last segment, user centric design is a key aspect to individually adapt features such as social comparison mechanisms to achieve the expected results of healthier behaviour.

Outcomes of Social Comparison

We have previously presented that most of the studies aimed to increase the physical health of their participants. One of the popular features was the implementation of social comparison features. Whether or not these mechanisms supported the achievement of goals will be discussed in the following segment.

In advance it can be said that in all selected articles in which the aim was improving health and fitness, participants reached their goal. When they used the app *TripleBeat,* users were much more effective during their training, though this was not significantly due to the competitive environment (de Oliveira and Oliver 2008). The user interface and goal-setting features were considered more relevant for increased activity. However, social comparison did increase the enjoyment of using the app. With the app *Fish'n'Steps* mostly users who were actively looking for a change in their activity habits experienced an additional boost to improve their health (Lin et al. 2006). The social comparison elicited through the depiction of the fish had an ambiguous influence on this enhancement (Lin et al. 2006). Some participants experienced higher degrees of motivation when their fish was growing. Nevertheless, a sad and smaller fish compared to competitors had a negative influence on the participants and some even stopped using the app all together. In addition, participants even felt overwhelmed by the competition and were exclaiming their concern that the app added unnecessary rivalry to the already competitive real life (Lin et al. 2006). The app *GoalPost* utilized Facebook sharing to provide support but also evoke comparison. The study of Munson and Consolvo (Munson and Consolvo 2012) confirms the ambiguity of social comparison. While one user saw benefits of sharing her activity data publicly due to higher accountability, many felt ashamed of openly publishing their potentially underperforming running times (Munson and Consolvo 2012). In their study with the app StepByStep, Zuckerman and Gal-Oz (Zuckerman and Gal-Oz 2014) tested both an baseline version as well as a leaderboard version. Results show that there is no significant difference between the two versions when it comes to increased physical activity (Zuckerman and Gal-Oz 2014). Furthermore, in the baseline version, a higher daily goal and more interaction with the app were correlated with more daily walking. This was not true for the competitive version. One explanation might be that with the baseline version, users open the app when interested in their walking and thus seeing the necessity to walk more. In the leaderboard version, users know that they are constantly compared to others. On the one hand, this means that users actively walk more without needing information about their current steps level. On the other hand, users might be adverted to seeing their potentially weaker performance compared to others when opening the app. This might even lead to decreased activity overall.

The negative influences can especially be examined in the *American Horsepower Challenge*. While most of the studies above focused on the physical health of young adults and adults, the study by Xu et al. (2012) aims at improving children's' health. The gamification features of the game improved motivation for more activity if not in the long run. The children participating in the challenge often checked their competitors step count to see where they stand and if they need to enhance their performance. However, in particular for schools in the same district, the rivalry became overcompetitive. When some students found a pedometer of a competing school, they kept it until the challenge was over to stop the rival student from collecting steps (Xu et al. 2012). While these actions are plausible, they might be counterproductive to the overall goal between all students to increase their activity. Even though negative impacts of social comparisons on physical activity might be neglectable, when it comes to more serious problems such as excessive drinking, consequences might be more serious. A more competitive environment makes some people more worried about failure and other users performing better than themselves, which may demoralise them and cause them to fall back into old patterns (Perski et al. 2018).

Lastly, it can be noted that the outcomes of social comparison features are ambiguous at best. There certainly are positive examples where social comparison mechanisms help to increase physical activity. Nevertheless, they are very rare and often mediated by goal-setting and self-monitoring features that have a greater influence on the users. Many participants have expressed their indifference or even dislike towards these social comparison features in health-related disciplines. As Xu et al. (2012) highlight, physical activities are often associated with sports that get broadcasted as tournaments or competitions (Xu et al. 2012). Health and fitness apps however are often targeted at *average joes* who want to improve their opportunistic activity. App designers are probably better off supporting this target group through social support and collaborative rather than competitive mechanisms. And again, there is not a one-size-fits-all approach when it comes to support vs. competition in this setting.

Limitations and Future Research

With regard to the research question, our approach was to examine how Festinger's theory of social comparison is applied in mHealth. However, there are only a limited number of studies that apply Festinger's theory in their research. The majority of the studies focused on social comparison as a factor of competition in their health applications. Social comparison was also used as a technique for behaviour change in order to guide the user to the desired behaviour (e.g. increasing physical activity). To collect enough qualified literature, we had to expand our search term by including the keywords *Gamification* and *Fitness app*. This biased search led to the next limitation that many of the collected papers only evaluated and approximated the design of fitness apps. The underlying reasons can be explained by the simple implementation of social comparison functions in mobile apps. Therefore, the contribution to mHealth as a general topic is limited.

Further research in this area is required to close the gap of implementing the theory of social comparison in mHealth. In particular, research needs to focus on the design of user-centric social comparison features to satisfy the individual preferences of users and to motivate them in the long run.

Conclusion

First of all, it can be said that due to the increasing number of obese and unhealthy people, it is very important to tackle this problem by all means. The fact that this is recognized and considered important by governments and organizations is a first step. mHealth in particular is predestined for this area, as most people own smartphones and can be encouraged to be physically active and lead a healthier lifestyle with the help of apps. It's no wonder, then, that growth projections for mHealth are exceeding expectations. The question this paper explored, however, is how social comparison theory is currently being implemented in mHealth applications. Our literature review adds an comprehensive overview to the small base of existing research papers regarding social comparison in mHealth. We conclude that social comparison features have high potential in mHealth applications but are still in a trial phase in terms of scope and ability to provide maximum enjoyment to users and satisfy their needs. In addition, it is becoming apparent that mHealth research will largely be conducted via smartphone apps. This can be explained by their low cost and simplicity. Participants do not need to be equipped with trackers or other wearables but can download an app to their own smartphone. If the future of mHealth is truly about smartphone apps, developers will need

to be careful to program in a user-centric way to allow customers to choose whether to use social comparison features, and if so, which ones. As our literature review showed, the majority of users do not yet see much added value in social comparison features, but rather feel motivated by behaviour change techniques such as goal setting, feedback, and self-monitoring. Some even feel that the world is already too competitive, and they have no need for such an environment in their free time. Others just want to compare upwards and don't want to be compared to peers who are doing worse. Here, too, it becomes clear how important it is to be user-centric.

References

Alam, M. Z., Hoque, M. R., Hu, W., and Barua, Z. 2020. "Factors Influencing the Adoption of MHealth Services in a Developing Country: A Patient-Centric Study," *International Journal of Information Management* (50:February), Pergamon, pp. 128–143.

Arigo, D., and Suls, J. M. 2018. "Smartphone Apps Providing Social Comparison for Health Behavior Change: A Need for Better Tailoring to Person and Context," *MHealth* (4), AME Publishing Company, p. 46.

Bailis, D. S., and Chipperfield, J. G. 2006. "Emotional and Self-Evaluative Effects of Social Comparison Information in Later Life: How Are They Moderated by Collective Self-Esteem?," *Psychology and Aging* (21:2), pp. 291–302.

Belmon, L. S., Middelweerd, A., Te Velde, S. J., and Brug, J. 2015. "Dutch Young Adults Ratings of Behavior Change Techniques Applied in Mobile Phone Apps to Promote Physical Activity: A Cross-Sectional Survey," *JMIR MHealth and UHealth* (3:4), JMIR Publications Inc., pp. e103–e103.

Brown, D. J., Ferris, D. L., Heller, D., and Keeping, L. M. 2007. "Antecedents and Consequences of the Frequency of Upward and Downward Social Comparisons at Work," *Organizational Behavior and Human Decision Processes* (102:1), pp. 59–75.

Buunk, A. P., and Gibbons, F. X. 2007. "Social Comparison: The End of a Theory and the Emergence of a Field," *Organizational Behavior and Human Decision Processes* (102:1), pp. 3–21.

Buunk, B. P., Collins, R. L., Taylor, S. E., VanYperen, N. W., and Dakof, G. A. 1990. "The Affective Consequences of Social Comparison: Either Direction Has Its Ups and Downs," *Journal of Personality and Social Psychology* (59:6), pp. 1238–1249.

Consolvo, S., Everitt, K., Smith, I., and Landay, J. A. 2006. "Design Requirements for Technologies That Encourage Physical Activity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, New York, NY, USA: Association for Computing Machinery, pp. 457–466.

DeSmet, A., De Bourdeaudhuij, I., Chastin, S., Crombez, G., Maddison, R., and Cardon, G. 2019. "Adults' Preferences for Behavior Change Techniques and Engagement Features in a Mobile App to Promote 24-Hour Movement Behaviors: Cross-Sectional Survey Study," *JMIR MHealth and UHealth* (7:12), JMIR Publications, pp. e15707–e15707.

Festinger, L. 1954. "A Theory of Social Comparison Processes," *Human Relation* (7:2), pp. 117–140.

Gerrard, M., Gibbons, F. X., Lane, D. J., and Stock, M. L. 2005. "Smoking Cessation: Social Comparison Level Predicts Success for Adult Smokers," *Health Psychology* (24:6), pp. 623–629.

Grand View Research. 2018. "MHealth Market Size Worth $151.57 Billion By 2025 | CAGR: 25.7%." (retrieved from: https://www.grandviewresearch.com/industry-analysis/mhealth-market; last accessed: July 31, 2020).

Gurupur, V., and Wan, T. 2017. "Challenges in Implementing MHealth Interventions: A Technical Perspective," *MHealth* (3), p. 32.

Hendrie, G. A., James-Martin, G., Williams, G., Brindal, E., Whyte, B., and Crook, A. 2019. "The Development of VegEze: Smartphone App to Increase Vegetable Consumption in Australian Adults," *Journal of Medical Internet Research* (3:1), pp. 1–16.

Hoorens, V., and Damme, C. Van. 2012. "What Do People Infer from Social Comparisons? Bridges between Social Comparison and Person Perception," *Social and Personality Psychology Compass* (6:8), pp. 607–618.

Huang, G., and Zhou, E. 2019. "Time to Work Out! Examining the Behavior Change Techniques and Relevant Theoretical Mechanisms That Predict the Popularity of Fitness Mobile Apps with Chinese-Language User Interfaces Guanxiong," *Health Communication* (34:12), pp. 1502–1512.

ITU. 2019. "Measuring Digital Development." (retrieved from:https://www.itu.int/en/ITU-

D/Statistics/Documents/facts/FactsFigures2019_r1.pdf; last accessed: July 31, 2020).
Kitchenham, B., and Charters, S. 2007. "Guidelines for Performing Systematic Literature Reviews in Software Engineering."
Klein, M. C. A., Manzoor, A., and Mollee, J. S. 2017. "Active2Gether: A Personalized m-Health Intervention to Encourage Physical Activity," *Sensors* (17:6), pp. 1–16.
Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., Riley, W. T., Shar, A., Spring, B., Spruijt-Metz, D., Hedeker, D., Honavar, V., Kravitz, R., Craig Lefebvre, R., Mohr, D. C., Murphy, S. A., Quinn, C., Shusterman, V., and Swendeman, D. 2013. "Mobile Health Technology Evaluation: The MHealth Evidence Workshop," *American Journal of Preventive Medicine* (45:2), Elsevier, pp. 228–236.
Latané, B. 1966. "Studies in Social Comparison - Introduction and Overview," *Journal of Experimental Social Psychology* (1:1), pp. 1–5.
Liao, G.-Y., Chien, Y.-T., Chen, Y.-J., Hsiung, H.-F., Chen, H.-J., Hsieh, M.-H., and Wu, W.-J. 2017. "What to Build for Middle-Agers to Come? Attractive and Necessary Functions of Exercise-Promotion Mobile Phone Apps: A Cross-Sectional Study," *JMIR MHealth and UHealth* (5:5), pp. 1–23.
Lin, J. J., Mamykina, L., Lindtner, S., Delajoux, G., and Strub, H. B. 2006. "Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game," in *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, Berlin, Heidelberg: Springer-Verlag, pp. 261–278.
Lockwood, P., and Kunda, Z. 1997. "Superstars and Me: Predicting the Impact of Role Models on the Self," *Journal of Personality and Social Psychology* (73:1), pp. 91–103.
Lyons, E. J., Lewis, Z. H., Mayrsohn, B. G., and Rowland, J. L. 2014. "Behavior Change Techniques Implemented in Electronic Lifestyle Activity Monitors: A Systematic Content Analysis," *Journal of Medical Internet Research* (16:8), pp. 1–15.
Maramis, C., Mylonopoulou, V., Stibe, A., Isomursu, M., and Chouvarda, I. 2019. "Developing a Smartphone Application to Support Smoking Behavior Change through Social Comparison," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, pp. 6922–6925. (https://doi.org/10.1109/embc.2019.8856672).
Mare, S., Sorber, J., Shin, M., Cornelius, C., and Kotz, D. 2011. "Adaptive Security and Privacy for MHealth Sensing," in *USENIX Workshop on Health Security (HealthSec)*.
Mechael, P., and Searle, S. 2010. "Barriers and Gaps Affecting MHealth in Low and Middle Income Countries : Policy White Paper," *Health San Francisco* (54:March), pp. 1–79.
Meng, F., Guo, X., Peng, Z., Zhang, X., and Vogel, D. 2019. "The Routine Use of Mobile Health Services in the Presence of Health Consciousness," *Electronic Commerce Research and Applications* (35), p. 100847.
Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., and French, D. P. 2011. "A Refined Taxonomy of Behaviour Change Techniques to Help People Change Their Physical Activity and Healthy Eating Behaviours: The CALO-RE Taxonomy," *Psychology and Health* (26:11), pp. 1479–1498.
Miles, M., and Huberman, M. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*, (2nd ed.), Sage Publication.
Mollee, J. S., and Klein, M. C. A. 2016. "The Effectiveness of Upward and Downward Social Comparison of Physical Activity in an Online Intervention," in *Proceedings - 2016 15th International Conference on Ubiquitous Computing and Communications and 2016 8th International Symposium on Cyberspace and Security, IUCC-CSS 2016*, pp. 109–115.
Mummah, S. A., King, A. C., Gardner, C. D., and Sutton, S. 2016. "Iterative Development of Vegethon: A Theory-Based Mobile App Intervention to Increase Vegetable Consumption," *International Journal of Behavioral Nutrition and Physical Activity* (13:1), pp. 1–12.
Munson, S. A., and Consolvo, S. 2012. "Exploring Goal-Setting, Rewards, Self-Monitoring, and Sharing to Motivate Physical Activity," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pp. 25–32.
Newsom, J. T., McFarland, B. H., Kaplan, M. S., Huguet, N., and Zani, B. 2005. "The Health Consciousness Myth: Implications of the near Independence of Major Health Behaviors in the North American Population," *Social Science & Medicine* (60:2), pp. 433–437.
de Oliveira, R., and Oliver, N. 2008. "TripleBeat: Enhancing Exercise Performance with Persuasion," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '08, New York, NY, USA: Association for Computing Machinery, pp. 255–264.

Oyibo, K., Adaji, I., and Vassileva, J. 2019. "Susceptibility to Fitness App's Persuasive Features: Differences between Acting and Non-Acting Users," in *ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 135–143.

Oyibo, K., and Vassileva, J. 2019a. "Investigation of Persuasive System Design Predictors of Competitive Behavior in Fitness Application: A Mixed-Method Approach," *Digital Health* (5), pp. 1–16.

Oyibo, K., and Vassileva, J. 2019b. "Investigation of the Moderating Effect of Culture on Users' Susceptibility to Persuasive Features in Fitness Applications," *Information* (10:11), pp. 1–21.

Perski, O., Baretta, D., Blandford, A., West, R., and Michie, S. 2018. "Engagement Features Judged by Excessive Drinkers as Most Important to Include in Smartphone Applications for Alcohol Reduction: A Mixed-Methods Study," *Digital Health* (4), SAGE Publications, pp. 1–15.

Raij, A., Ghosh, A., Kumar, S., and Srivastava, M. 2011. "Privacy Risks Emerging from the Adoption of Innocuous Wearable Sensors in the Mobile Environment," in *Conference on Human Factors in Computing Systems - Proceedings*, pp. 11–20.

S. Bhuyan, S., Kim, H., Isehunwa, O. O., Kumar, N., Bhatt, J., Wyant, D. K., Kedia, S., Chang, C. F., and Dasgupta, D. 2017. "Privacy and Security Issues in Mobile Health: Current Research and Future Directions," *Health Policy and Technology* (6:2), pp. 188–191.

Shang, R. A., Chen, Y. C., and Huang, C. C. 2012. "A Stage for Social Comparison - The Value of Information in Virtual Communities," in *Pacific Asia Conference on Information Systems (PACIS) 2012*, pp. 1–17.

Statista. 2019a. "Anzahl Der Smartphone-Nutzer Weltweit von 2016 Bis 2018 Und Prognose Für 2019." (https://de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/, accessed May 21, 2020).

Statista. 2019b. "Worldwide Mobile App Revenues in 2014 to 2023," *Statista.Com*. (https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/).

Suls, J., Martin, R., and Wheeler, L. 2002. "Social Comparison: Why, with Whom, and with What Effect?," *Current Directions in Psychological Science* (11:5), pp. 159–163.

Tong, H. L., Coiera, E., and Laranjo, L. 2018. "Using a Mobile Social Networking App to Promote Physical Activity: A Qualitative Study of Users' Perspectives," *Journal of Medical Internet Research* (20:12), pp. 1–13.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Q.* (26:2), USA: Society for Information Management and The Management Information Systems Research Center, xiii–xxiii.

Wheeler, L. 1966. "Motivation as a Determinant of Upward Comparison," *Journal of Experimental Social Psychology* (1:1), pp. 27–31.

WHO. 2011. "MHealth - New Horizons for Health through Mobile Technologies." (retrieved from:https://www.who.int/goe/publications/goe_mhealth_web.pdf; last accessed: July 31, 2020)).

WHO. 2018. "Obesity and Overweight."

Wilding, J. P. H. 2014. "The Importance of Weight Management in Type 2 Diabetes Mellitus," *International Journal of Clinical Practice* (68:6), BlackWell Publishing Ltd, pp. 682–691.

Wu, Y., Kankanhalli, A., and Huang, K. W. 2015. "Gamification in Fitness Apps: How Do Leaderboards Influence Exercise?," *2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015*, pp. 1–12.

Xu, Y., Poole, E. S., Miller, A. D., Eiriksdottir, E., Catrambone, R., and Mynatt, E. D. 2012. "Designing Pervasive Health Games for Sustainability, Adaptability and Sociability," in *Proceedings of the International Conference on the Foundations of Digital Games*, FDG '12, New York, NY, USA: Association for Computing Machinery, pp. 49–56.

Zhou, Y., Kankanhalli, A., and Huang, K. W. 2016. "Effects of Fitness Applications with SNS: How Do They Influence Physical Activity," in *2016 International Conference on Information Systems, ICIS 2016*.

Zuckerman, O., and Gal-Oz, A. 2014. "Deconstructing Gamification: Evaluating the Effectiveness of Continuous Measurement, Virtual Rewards, and Social Comparison for Promoting Physical Activity," *Personal and Ubiquitous Computing* (18:7), pp. 1705–1719.