UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Tools and techniques for security and privacy of big data

Healthcare system as a case study

MASTER OF SCIENCE THESIS

Author: Shiva Toutounchiavval

Date: 22/06/2017

Supervisor

PhD. Josep Solé Pareta

FACULTAT D'INFORMÀTICA DE BARCELONA

This thesis is presented by Shiva Toutounchiavval

In fulfilment of the requirements for the degree of

Master in Innovation and Research in Informatics

With specialization in

Service Engineering

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Professor Josep Solé Pareta for his patience, motivation, enthusiasm, guidance, and immense knowledge that helped me in all the time of research and writing of this thesis.

I will forever be thankful to my parents and brother, who always have supported me by their unconditional kindness and family love.

My very profound gratitude is to my love Carlo who his presencs has been a turning point in my life. This accomplishment would not have been possible without him. Thank you

Thanks to my dear friends Rodrigo, Isabel, Raquel with whom I expeienced very nice moments. Thanks for your support.

CONTENTS

CHAPTER 1

1 Introduction

This section gives brief introduction about big data and secury problems.

Big data and the processing related to this concept has become important part of advanced science and business. Huge amount of information are produced daily not only by the complexity of new science (think for example elemental particles generation and detection in collider's physics) but also in everyday life by millions of people just by means of online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions. Moreover a new market has been open by mobile phones and their applications. These data are stored in databases, which efficiency has to exponentially increase because of the need of catching, framing, storing, and investigating this information in real time or in the shorted possible time. Another consequence of this new life style is the need of sharing this amount of information between million of database and programming devices, i.e. the creation of standards in order to increase as much as possible the level of compatibility.

Other important point is that the amount of this information generated and stored is quickly expanding. This means that the big data generation and analysis has become an hot topics for science of data management, mainly because of the increasing request of company (it doesn't matter the levels) to convert this huge resources into data and information that can be useful in order to accomplished with their different addresses.

All this is responsible for the creation of the terminology “big data”. Big data are found in astronomy[1] (eg, the Sloan Digital Sky Survey of telescopic information), retail sales (eg, Walmart's expansive number of transactions), search engines[2] (eg, Google's customization of individual searches based on previous web data), and politics[3] (eg, a campaign's focus of political advertisements on people most likely to support their candidate based on web searches). This tendency of digitizing information has involved of course also medicinal records, and has been rapidly promoted by healthcare industry as busyness paradigm.

As a consequence, healthcare industry abruptly extended the used information in terms of complexity, diversity and timeliness. Clinical trends have as significant role as health-care costs in big data’s growth. For example instead of traditional treatment, evidence-based medicine is replaced which is based on the precise available information. In this case strong and improved algorithms can provide clear evidence and accurate information from individual data sets and also to make insurers to reassess their prescient models.

By increasing cost of healthcare services and health insurance premiums, sense of having better healthcare management and high quality of care is expected more. This need make health care industry to be proactive rather than reactive in order to enhance the quality of care and to reduce the costs. Innovative achievements have a special place in enabling healthcare to be proactive and also it is costly. It means this cost can cause a serious threat for the security of patient’s information. In order to reduce the cost privacy of information may be neglected by Healthcare managements. Privacy-preserving data mining (PPDM) [4] has been studied effective data mining algorithms in order to reduce the privacy risk. Moreover Internet attacks are other concern of systems administrators.

1.1 Objective of this Master Thesist

This provided report aims to deal with tecqniques, security and privacy of big data. Research goal is referring to this point that Persistent and Sophisticated targeted network attacks are continuously challenging today's enterprise security teams. By exploring different aspects of high performance networks, the major objective of this Master Thesis is to present fundamental tools and techniques with deep focus on possibilities, impediments and challenges for providing network security and privacy in Big Data. Since the healthcare industry harnesses the power of big data, security and privacy issues are at the focal point as emerging threats and vulnerabilities continue to grow. As a case study, this Master thesis will also review the state-of-the-art of security and privacy issues in big data as applied to healthcare industry.

The whole document is organized as follows:

Chapter 1 gives brief introduction about big data and secury problems.

Chapter 2 provide more information about big data and its characteristics, their diffusion, size of big data, related problems with privacy and other possible vulnerabilities of security and privacy in big data such as web tracking is discussed.

Chapter 3 provides a clear insight of privacy and security in healthcare system.

Chapter 4 demonstrates fundamental techniques with deep focus on possibilities, impediments and challenges for providing network security and privacy in Big Data. Chapter.

Chapter 5 presents various companies which they seem as pioneer in introducing tools and techniques in order to provide accurate information while security of privacy is guaranteed. At the end, after discussing status of art of the problem and compliance tools that is used now a day, we present the open issues, related with subject.

Chapter 6 refers to the research direction, related subjects and issues which are under investigation.

Chapter 7 demonstrates potential PhD project related to the issues that can be improved.

Finally, Brief conclusions at the end of report remind reader that, while data protection can be challenging in a big data context, the benefits will not be achieved at the expense of data privacy rights; and meeting data protection requirements will benefit both system and individuals.

CHAPTER 2

2 Theoretical framework: Big data and Privacy issues

This chapter starts by big data definition, internet security and web tracking phenomena. More particular characteristics are defined that differentiate big data from more traditional forms of data processing. Next discussion in this chapter is analysing internet security issues and the main implications for data protection.

2.1 Big data (definition)

Big data is an evolving term for massive data sets which are large, complex and consist of various types. These variety and complexity has aroused many difficulties for storing, dealing with traditional data processing application software , analyzing and visualizing them for providing accurate information. Big Data' is considered as similar as 'small data', but bigger in size. It means that different approaches such as Techniques, tools and architecture are required to deal with it. This amount of data cannot be analyzed by traditional computing techniques. Many organizations, industries and social Medias are involved with big data. Statistic shows that 500+terabytes of new data gets ingested into the databases of social media site **Facebook**, -- see glossary -- every day[5]. This data is mainly produced in terms of photo and video uploads, message exchanges, putting comments etc. In is interesting to know that 10+terabytes of data can be produced by Single Jet engine in 30 minutes of a flight time. This number may arrive to Petabytes[6] by thousand flights per day.

2.1.1 Characteristic of Big Data

In the following few characteristic of big data are enumerated[7].

- **Volume*:*** Volume or size refers to the amount of dat (1000 terabytes). Volumn has important role and influence on scalable storage and adequate support in various organization and industries. Many organization faced problem in the case of large volume of data.

- **Variety:** other aspect of 'Big Data' is its variety. Variety refers to aggregation of many kinds of data, both structured and unstructured, including emails, photos, videos, monitoring devices, PDFs, audio, web server logs, etc. Dealing with this variety of data needs high level of techniques and technologies. But should be considered that effective analytics guarantee better decision making and provide more accurate information and results.

- **Velocity:** The term 'velocity' mentions to how fast the data is coming in. In traditional databases, data loads and update periodically whereas big data is processed and analyzed in real- or near-real-time. This aspect of big data is very crucial for many organizations and industries such as hospitals and clinics in order to access to up-to-date information

- **Variability:** variability refers to the inconsistency which can impact on data homogenization. It is resulting from several disparate data types and sources, or when

collecting data based on language processing. In this case data depends on its meaning and each word may have several meanings.

2.1.2 Storing, selecting and processing of Big Data [8]

Big data storage primarily supports storage and input/output operations on storage with a very large number of data files and objects. Consider that in this case data storage needs to be fast, intuitive, effective, safe and cost-effective. A typical big data storage architecture is made up of a redundant and scalable supply of direct attached storage (DAS) pools, scale-out or clustered network attached storage (NAS) or an infrastructure based on object storage format. The storage infrastructure is connected to computing server nodes that enable quick processing and retrieval of big quantities of data. Moreover, most big data storage architectures/infrastructures have native support for big data analytics solutions such as Hadoop, Cassandra and NoSQL.selection – see glossary -- will be based on data characteristics.

2.1.3 The structure of big data [9]

Three types of data should be considered, structured, unstructured, and semi-structured.

- **Structured Data:** structured data can be stored properly in database with all rows and columns. It is organized and has a fixed size. For example current data warehouse contains only structured data. Data may be text or numerical but the structure is pre-defined and could be easily stored and managed in relational databases, hence structured.

- **Unstructured Data:** In general unstructured data consists of data which is raw, unorganized and can be stored without the known format by the system and can't be fit into a database. Usually these kinds of data would be converted into structured data. The list of unstructured data includes Emails, Word Processing Files, PDF files, Spreadsheets, Digital Images, Video, Audio, and Social Media Posts. The Unstructured data is divided into machine generated or human generated data. Here are some examples of machine-generated unstructured data:

    - **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth – see glossary -- and you get the picture.

    - **Scientific data:** This includes seismic imagery, atmospheric data, and high energy physics.

    - **Photographs and video:** This includes security, surveillance, and traffic video.

    - **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic seismic profiles.

    - The following list shows a few examples of human-generated unstructured data:

    - **Text internal to your company:** Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.

- **Social media data:** This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr – see glossary --.

- **Mobile data:** This includes data such as text messages and location information.

- **Website content:** This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

- **Semi-Structured Data:** Semi-structured data can't be fitted in relational databases. They need some special processed in order to be analyzed and stored. Examples of semi-structured : CSV – see glossary -- XML – see glossary -- and JSON – see glossary -- documents are semi structured documents, NoSQL databases – see glossary -- are considered as semi structured.

2.2 Internet security

Internet security is related to the concept of computer security which deals with threats of the Internet, often consists of browser security additionally network security used to different applications or operating systems. Its goal is protecting internet from attacks by building up rules and measures[10]. Recent researches have presented Internet as unsafe and insecure channel for transferring data which is caused a high risk of intrusion or fraud, such as phishing[11]. Different techniques have been utilized to secure the exchange of information, including encryption and from-the-ground-up engineering[12].

2.2.1 Threats

Information security threats have been appeared in wide range of structures and ways. Some of significant and common threats are software attack, theft of intellectual property, identity theft, theft of equipment or information, sabotage, and information extortion. Every day many users of internet are affected by software attacks such as Viruses[13], worms, phishing attacks, and Trojan horses.
Some types of attacks are explained briefly as following:

- **Malware** is a kind of software with unique characteristics to disrupt computer operation, damage information, or gain authorized access to private computer systems. It can infect many programs by executing itself. Malware is characterized by its pernicious goal, acting against necessities of the PC client, and does not contain software that causes unexpected damage because of some lacks. Badware is other word which is used for both malicious concepts malware and unintentionally harmful software. There are different types of malware listed below:[14]

  - **Viruses*:*** Computer ***Viruses*** are programs that can replicate their structures or effects by infecting other files or structures on a computer. The common use of a virus is to take over a computer to steal data.

  - **Worms:** Computer **worms** are programs that can replicate themselves in order to spread to other computer network, executing malicious function throughout.

  - **Ransomware:** It is considered as a kind of malware software that limit user by blocking access to a computer system and makes unsuspecting user to pay in order to remove

restriction, hence it is considered as "scareware". In this case system's screen or users' files entirely will be blocked. These kinds of malwares usually infect system or encrypt certain file types. User's system can be infected by these types of threats in various ways, for example visiting malicious websites, malicious payloads, and attachments from spammed email. The best way to dealing with ransomware is installing antivirus and keep it updated.

- **Spyware** refers to programs that are installed on a computer system in order to monitor activities and collect information without user's consent. For example it is used in order to monitor employees' or children browsing activities or to target advertising in a marketing campaign. Moreover, it can impact on privacy by abusing the software.

- **A Trojan horse,** commonly known as a destructive program that, unlike viruses, is not replicated itself but it can be harmful, usually user is unaware that he has installed a Trojan horse, because it is installed in hidden way by attachment or free program, such as a game and continue its activities without introducing itself as a malware.

- **A denial-of-service attack** (DoS attack) or distributed denial-of-service attack (DDoS attack) try to overwhelm computers resources in order to make them unavailable or difficult for users to use them. Targets of this kind of attack are crashing server in a way that make recovery so hard, failure to access a particular website and increasing volume of spam email.

- **Phishing** is a form of fraud in order to steal confidential information such as login credentials or account information, via email or web page. In this way, fake email or link, is sent to the person by known contact or famous organization. In this way, attachment of email can be a malware which be installed on user's devise or can be a inserting link to encourage user by clicking on it, reveal his personal information such as account IDs or credit card details.[15], [16]. Insurance group RSA[17] said that phishing accounted for worldwide losses of $1.5 billion in 2012.

Unsafely environment in IT field such as stealing intellectual property[18]has provoked many concerns among businesses. One of the thriving and prevalent businesses for thieves in virtual world especially Internet, is Identity theft. The goal and target of Identity theft is obtaining personal and vital information in order to Profit-seeking purposes. Today Information Technology is improving very fast and most of the electronic devices are being summarized in one small cell phone. Moreover, high data capacity and performance, and variety of applications tempt people to store more public or confidential information in their mobile which can be stolen easily. Also, many interesting sites such as Facebook and LinkedIn are persuading people to register and share their information in order to receive social advantages. It means everyday more information is being shared through network which needs security, protection and support against different sophisticated network attacks. Competitive market may provoke sabotage. For example many black hat hackers are focusing on organization′s website in order to damage company's reputation by loosing data. Information extortion is other kind of way that attacker steals information for exchanging them by specific amount of money. Apart of using security software, user consciousness and carefulness is essential to protect user's system of any future network attacks. Mass information is not just related to individuals. Governments, military, corporations, financial institutions, hospitals and private businesses are also collecting a great amount of confidential information about their employees, customers, products, research and financial status. The greater part of this data is currently gathered, handled and saved on electronic computers and transmitted to other computers via network. From a business point of view, information security must be adjusted against cost; a mathematical economic way for to deal with this concern is given by the Gordon-Loeb Model[19].

2.3 Web tracking

We tracking depend on its type; can be considered as a serious threat for privacy and security in internet. For over 10 years, tracking systems have been observing where you go and what you look for on the Web, without user authorization. In General, Web tracking affects all the systems and platforms and all related operation such as extraction, storage, analyzing, selling, buying and auctioning of personal online data, is concealed and unauthorized. Today many individual's information that voluntarily are revealed on well-known web sites, are being tracked and analyzed. Target of this tracking for all famous and well known company such as Facebook, Apple and Google is very clear. All individual's information are gathered for advertising purpose. Moreover, these companies need to know more about their clients and customer's interests in order to design better product and increase their revenues. In the business, all the activities of online user through network will be recorded and reported, without a little attention is paid to her and how she is affected by web tracking. Although, sometime users can benefit from this tracking since they receive useful information fitting their interests. Other synonyms of web tracking which are used in this concept are 'Internet tracking'[20] or 'online tracking'[21]. Big data and web tracking are complicatedly associated and their development shows up commonly dependent. Storage capacity play a significant role in efficiency of web tracking. Without enough and appropriate space, web tracking may lose its reputation as an interesting application[22]. All activities of online user in social media or web search are gathered and stored in Big Data warehouses for different purpose, some are used in advertising and some for real-time-bidding. Gathered information consists of all type of data such as online blogs, pictures, texts, Tweets, emails, videos and technical details attached. Apart of personal information, emotional expressions or affective exchanges can be recorded. For example when a user search specific hotel with specific characteristic, this information can be saved by web tracker in order to find similar hotels which are closed to user interest. In general, purpose of collecting data in this way is trading with customer information. Problem arise when this information are leaked to third parties in order to abuse them or accessed by government agencies and identity thieves. Web tracking, based on the type of use, can be divided to different groups as following.[23]

- **User-Oriented Search**

  A well-known use of tracking is collecting data based on user's interests. The term Filter Bubble[24] is introduced by Pariser, resulted from information when a website algorithm selectively guesses what information a user would like to see based on data about the user (such as location, past click-behavior and search history). As a result, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles. (Definition is taken from Wikipedia).

- **Online Advertising**

  Initial important goal of developing web tracking was having better marketing and high sale profit. By technology Improvement and innovation of complex applications and web sites, web tracking decided to improve its operation and add more sophisticated features such as controlling user's behavior, audience segmentation, and targeting. For these reason, tracking is considered so practical and efficient in most of websites. According to Wills and Tatar [25] survey, Google is one of well known web sites that collects information and use them for online ads.
  AdStack[26] is a technology that provides service related to advertisements in marketing. In this case when the user open the spesific email, based on user information the advertisement is downloaded.

- **Web Analytics and Usability Tests**
  Starting critical objective of creating web tracking was having better advertising and high deal benefit. Other objective of web tracking is increasing the usability and quality of website, especially with websites becoming more and more interactive and advanced by latest technology and various features. According to [24], web analytics and usability tests are two applications which are used by website in order to collect, report, and analyze user's data in order to identifying measures to improve website in the future.

- **Assessing Financial Credibility**
  Many companies benefit from information of user in social connection in order to assess the creditworthiness of their clients. For example for the company Lenddo – see glossary – is important to know the user's friends in Facebook. If one of friends has delay in her/his payment related to the company, may have bad impact on user. Many companies such as Kreditech – see glossary --, eBay – see glossary --, or Amazon accounts – see glossary – are using personal data from Face Book, user's location or other social medias which are gaied through internet[27].

- **Price Discrimination**
  Tracking also can influence the price of products by assessing economic circumstance related of the potential clients. Companies modify the price by analyzing collected information such as ZIP code, the date of birth, the region that user lives, income, habits and interest[28].

- **Determining the Insurance Coverage**
  Web tracking can help insurance company to reduce their risk by reviewing information related to the client lifestyle. In this case, all activities of customer can be recorded, such as kind of product that is bought, subject of survey in internet, magazine subscriptions. Analyzing this information provides the risk of cancer, accident that customer may be faced. Allfinanz[29] and TCP LifeSystems[30] develop software, which has this ability to determine the risk in mentioned way.

- **Impact on the Job Market**
  One of the most important effects can be perceived in the job market. Many employees are judged based on their online social activities. This judgment in many cases can be unfair because of receiving incorrect or bad quality of information[31]. One case is related to a person who was refused by employer because of attending to the mental health conference. After this case, Finland's Data Protection Ombudsman is prohibited employers from googling job candidates [32].

- **Identity Theft**
  Collected information can be used in dangerous way. Stealing social security number is one of challenges that can have destructive effect on person. It can pave the path to have access on sensitive websites (e.g., banking or loan services)[33].

2.4 Privacy issues

Big Data itself is an important and complex subject that directly affects to security. In this case, analysts have an obligation or better to say responsibility about clients to be straightforward about data gathering and usage. Some important concerns are explained as following[34] :

- All individual data which are gathered in databases or social networking website can be consolidated with many external extensive informational, reveal some information of person which is secretive and unauthorized to be kept or known by other people.

- More information which are collected for using in different business of the organization, is unauthorized and user doesn't know that his/her privacy is not secure.

- Another important issue is related to Social classification. In this case collected data being used by educated person for science research or improving business while restricted people will be treated worse.

- Big Data utilized by law requirement will build the odds of certain labelled individuals to experience the bad effects of unfavourable results without the capacity to claim or notwithstanding having information that they are being separated.

How big data is used is important in justice point of view. Individual's data may be analysed investigate purposes, eg to identify general patterns and connections, or it might utilize individual information to settle on choices influencing people. This collected data and its analytics may influence people in different way. Some of them seem fair and interesting for user when receiving advert related to their interest. Profiling is other concept that has a more intrusive influence upon individuals. In this case people can be affected by their information which is collected from different organization. Therefor GDPR (General Data Protection Regulation) [35] devise some rules that personal data must be "*processed fairly, lawfully and in a transparent manner in relation to the data subject*". By contrast big data analytics seems deceptive. They retrieve data and analyze them and provide unexpected result. By increasing different kind of social media, much information may be shared by certain organizations in order to know too much about individuals. There is concern about this amount of information which can be shared to security organization by social Medias such as Facebook and Yahoo or even sometime companies make decision to hire employee based on their information which is shared in Face book or link dine. This trend can put individual`s life in danger. For example, data gathered in electronic health record systems in accordance with HIPAA/H ITECH provisions (see following chapter) is already raising concerns about violations of one's privacy. One important issue in research field is providing algorithms which are using individual information randomly in order to guarantee privacy [36].

CHAPTER 3

3 The healthcare system

This chapter discusses the importance of privacy, security and, web tracking in health care and the challenges which are associated with them. Health Insurance Portability and Accountability Act (HIPAA) and the Security Rules are also explained. We examine the storing and transmission of sensitive patient data in the modern health care system and discuss current security practices that healthcare system providers institute to comply with HIPAA Security Rule regulations.

Healthcare sciences and big data are inseparably connected. This link becomes tighter by the day and guides the change of healthcare systems in terms of knowledge and customized medical solutions. The price to pay for healthcare companies is of course the volume of exchanged information, the compatibility between different sources, and the need of fast analytics of collected data. Each person indeed can generate an enormous information trail in everyday life just by using its own mobile or other network connected devices. This data can yield critical connections between's factors, for example, our genome, eating diet, way of life, and family history and lead to better and personalised diagnosis, treatments, medical solutions. Moreover, genetic investigation is now day making possible the almost complete sequencing of personal multiple genes, i.e. it is progressively driving towards the use of the personalised medicine, i.e. big data are becoming a clinical tool. The healthcare industry has largely investigated big data methods for the diagnosis, the treatment, and the management of the population health, as well as, from economical point of view, defining strategy, planning and administration opportunities. On the other hand, since healthcare companies are moving from a volume-based (quantity) to a value-based models/technologies. This has led to the association between quality of care and costs, determining the key role of data quality in this change. The first challenge of the healthcare industry is managing this extended volumes of data in order to access and link this information in real time: the personal continuous request of information from the global network translates in healthcare systems with large amounts of clinical, financial, genomic, social and environmental data necessary for scientific investigation and patient care in order to understand and check population health and eventually prevent diseases and developing optimized treatments. The analysis of the multiple collected information increases knowledge of patients about their health condition and prevents harmful health occurrence such as diabetes. In this case, the definition of the patient condition became more complex, by defining a new health scale for measuring in which needs clinical and physical data are coherently complemented by the psychological and social environment data related to the patients[36]. These changes has had an inevitable counterpart, in terms of individuals' rights to privacy since this kind of data could have different market and different purposes. The developing and spreading personal medicinal data (generation, electronic storing and analysis) has been garanteed by traditional policies which claims that sharing recorded informatin in helthcare industry is impossible and lead to countless advantages for the business companies and heath care industries in particular, but comporting an unacceptable cost for individual protection. This new framework has forced the government authorities to request new protection standards, determining the developing and implementation new adequate measures to ensure the security of private health information. In this context the introduction in 1996 of the Privacy and Security Rules of the Health Insurance Portability and Accountability Act (HIPAA) – see section 3.3 -- is the most significant document that the US Federal Government has produced to ensure the protection of patient privacy in the medicinal healthcare system. HIPAA Security and Privacy Rules has permitted the identification of obstacles actively helps in their removal, also if their implementation to safeguard sensitive medical

data has met different needs, complains and frustrations between health care workers and administrative and Information Systems staff[37].

3.1 Use, Benefits and drawback of Big Data in health care

In each business strategy the capability to understand customers' needs at different scales and as consequence the possibility to orient productions and markets in a focused and effective way represents a primary challenge. In 2012 the **Center for Economics and Business Research** estimated a combined benefit for the UK economy in adopting big data related technologies in £216 billion just in the period 2012-17. In this estimation, £149 billion is the fraction originated just by business effectiveness[38]. On the other hand, also customers could profit from more focused advertisements and tailored offers and at the same time improve services and products. For instance, the process of applying for insurance can be made less demanding, with fewer questions to reply, if the insurance company can get other relevant information by means of big data methods. This apply of course also to public and government administrations, where Big data analytic helps to plan and deliver more efficient services and create constructive outcomes that improve quality of individuals' lives. Like the majority of the companies, also the heath care enterprises are involved in this process, receiving new information, which are generated in many different forms. Big data have thus the potentiality to change and orient also healthcare systems.
More in details, for the healthcare system, the advantages coming from big data are[39]

- Understanding and Targeting Customers
- Understanding and Optimizing Medical Business Processes
- Personal Quantification and Performance Optimization in medical area
- Improving Sanitary structure Performance
- Improving Medical Science and Research
- Optimizing Machine and Device Performance
- Improving Security and Law statements in medical applications
- Improving and Optimizing services at City and Country scales
- Financial Trading

This corresponds to the definition of a new pathway for healthcare system, which can be summarized in this way: [4]

- **Right living:** Playing an active role in medical treatment, including disease prevention can build value by patients. In this case, they are encouraged to make sane choice of life-style, for example, an equilibrated diet and exercise, no alcohol or smoking, in other words, actively taking care of themselves.

- **Right care:** This pathway requires the guarantee that each patient get the most convenient, suitable treatment available. Moreover depending to medical protocols, right care foresees that, across different healthcare providers/suppliers, all the operators should have the same information and work with the same target, i.e. avoiding multiple efforts and un-optimized treatments.

- **Right supplier:** This pathway ensures that all patients, independently from their status, will be attended by high-performing experts that are best matching the duties and the tasks to be achieved, in order to provide the best possible result. The "Right supplier" in this manner has two implications: the correct match of supplier and expertises guarantee the capability to face the task proposed but only after a selection of the providers.

- **Right value:** target of this pathway is providers and customers put effort to increase healthcare reputation and value while attending or advancing its quality. In this case some measures that have influence on cost-effectiveness of care should be considered such as removing fraud, waste, or abuse in the system.

- **Right innovation**: This pathway includes the identification of new treatments and therapies to deal with patient care, and the enhancement of innovation engines — like for example, by advancing medicine and boosting medical research. To fulfil this pathway, companies are requested to improve the use of their big data, for example, by searching new potential targets, like in pharmaceutics research enhancing towards new optimized molecules. Companies could use the collected information also for predictive personalised models in clinical treatment.

On the other hand several drawbacks can be identified to the introduction of big data in healthcare system, like:

- **Quality of Data:**[40,41] Decision making and predictive models are strictly connected to data availability, and roughly speaking more data availability is leading to more reliable models, i.e. more successful outcomes. However collecting data and take care of their storing and analysis is not costless. Companies managers and leaders will face the always increase need of analysing and storing this information in the most possible efficient way. For this reason data collection must be fundamentally focused on quality of the data, more than quantity, in order to eliminate redundant information which could slow down the data processing. This guideline opens to new inquiries like how it can be guaranteed which information is important, how much information would be sufficient for decision making and which level of accuracy to attribute to predictive models.

- **Repurposing data:** [36] In the use of big data technologies a further issue is represented by the use of collected information for a reason different respect with the original aim, and the possible selling of this information to an alternate company. This is the power and at the same time is a drawback of big data analytic since predictive models are able to "fish" new insights and establishing new connections between different datasets. Companies, like **DataSift**[42] provide the collection and analysis of online data produced by social media services (Twitter's GNIP for instance) to make their predictions more effective. Estimating potential number of visitors in tourist websites has been improved by including Geotagged photo on Flickr or the profiles of supporters on social Medias[43]. In the same way data which is obtained from mobile phone is successfully used for counting and analysing of to the footfall in retail centres[44]. Advertising campaigns are now influenced by customer's nationality in order to have better advertisement according to customer interest. Movements and shopping habits in airport terminals are tracked in order to set rents and prices for the costumers.[45]

3.2 Practical cases

The healthcare system can benefit enormously from the use of advanced analytics and big data technologies. However critics of data security worry about patient records which are a prime target for cyber thieves, because they yield personal information (like Social Security numbers, Medicare information and prescription information…). Some examples are provided[46]:

Evidence based medicine:[47] The paradigm of "Cookbook medical solutions" (i.e the use of a battery of medical tests to analyze and determine the nature of illness, by progressively eliminating not compatible causes) is progressively substituted by "evidence-based medicine" (i.e. the identification of symptoms to narrow the possible diagnoses). Beth Israel Deaconess Medical Center in Boston, is utilizing sensitive medical data from two million patients collecting through mobile phone application in order to reduce the possible diagnosing causes. In the same framework, doctor notes are automatically encoded into standards so that "hypertension" and "high blood pressure" are simultaneously found when searching for this topic.

Cancer Prevention[48]**:** Public Health England (PHE) suspected that tumour survival rates in the UK were low compared with equivalent ones in Europe, because it is enough advanced to be cured. For this PHE inaugurate the "Routes to Diagnosis"[49] in order to evaluate how and when cancer was diagnosed in people. By using complex algorithms, PHE deals about a big data project that was actively involving 118 million records on 2 million patients from different information sources. The project has revealed that patient diseases were increased from 2006 to 2013. A key issue emerging from results of the analysis was that in 2006 almost 25% of cancer cases were only diagnosed in an emergency when the patient came to hospitals. Patients diagnosed in this way have lower survival prospective compared to other situations. To overcome this problem, PHE put in place several initiatives to promote preventive diagnosis. The results were then published in 2015, showing that by 2013 just 20% of cancers were diagnosed as an emergency. The main achievements of the project was that the increase of population attention towards cancer diagnosis, which raise awareness of the symptoms of lung cancer and help people to spot the symptoms early.

Valence Health, Improving Outcomes and Reimbursements:[50] The elevate cost of Medical services are one fundamental reason why efficiency promised by new big data technology is so attracting healthcare industry. Strategy and decision making machines in healthcare systems tightly connected with financial matters. Digitizing and sharing of healthcare data together with the promise of decreasing costs in storage and parallel processing are driving the big data healthcare revolution and the MapR Converged Data Platform was finally identified as the best solution[51]. This platform is proving the first primary pool for the different health care companies, storing a huge amount of big data: daily new 3,000 inbound data feeds are added, with 45 different types of datasets. These entries include not only lab test results, health records, prescriptions, immunizations, patient medications and drug taking, but also claims and payments together with identifiant financial data from doctors and hospitals. The data collected are used to improve the specialists' and clinics' choice and to make faster and effective reimbursement process. According to collected data which is growing fast, made companies to change strategies and corresponding infrastructures. Following example demonstrates the efficiency enhancement permitted by the creation of MapR Platform. In the past the processing time due to the collection of 20 million lab records was estimated around 22 hours, mainly because of the limitation of single machine CPU hardware. MapR parallel processing significantly reduce the duration of data treatment (from 22 hours to 20 minutes). Big data technologies also makes possible to apply client requests as soon as possible which before were exceptionally hard such as modification or erasing request of a specific

entry. While in an ordinary database this operation can be considered successfully concluded only after ¾ weeks (in order to remove all possible links or redundancy), MapR Platform provides point-in-time recuperation of the entries that permit to change or remove that particular record.

Liason Technologies: Streaming System of Record for Healthcare:[52]
Integration, management, and data protection between different data sources has stimulated the generation of Liaison Technologies based on cloud-solutions, whose constraints should meet HIPAA requirements – section 3.3 – and the creation of standard for data sharing. With MapR Platform, the portion of data accessed by a healthcare operator is modified and now inserted directly in the existing record, by simultaneous saving a log for each change of data. As an example, the patient record may be accessed by different operators (companies, doctors, and scientists). Changes can be operated in real time into the MapR-DB, HBase, MapR-DB JSON document, graph, and search databases. The entire operator will keep the most update as possible data view in their most adequate format. To summarize, by means of MapR platform and Liaison technologies healthcare system can secure the whole assembly of data elements in single shot, avoiding the spreading out of the establishing of priority access right and alternate solutions.

Novartis Genomics: Next Generation Sequencing (NGS):[53] is a cutting edge big data application that permits the connection of immense pool of raw heterogeneous data (originated by different organizations) which are continuously subjected to movements and/or modifications. All this requires workflow tools that are sufficiently strong to handle such huge amounts of data and at the same time versatile enough to update rapidly according to the change in the exploration methods. This project also represented a milestone in the compatibility terms, since it makes possible the integration of from Novartis[54] respect with information from different external organizations, as 1000 Genomes, NIH's GTEx (Genotype-Tissue Expression), and TCGA (The Cancer Genome Atlas). The Novartis group selected Hadoop and Apache Spark to make effectively its own work process framework.

Healthcare IOT Start-up: Working to Classify Heart Conditions Faster:[55] This is another example that shows the efficiency enhancement due to the introduction of MapR-FS Platform in the analysis of cardiovascular diseases. One basic test in this kind of diagnosis is the identification of anomalies in heart rhythm. This identification was first done manually by the sanitary operator. The process was taking over 24 hours, a long lag before doctors can access the patient data, increasing the risk of medical emergencies. With MapR-FS, Telemed[56] was first able to investigate data from various medical devices directly via NFS – see glossary -- into a computer cluster for real-time patient diagnosis. Respecting HIPAA constraints, Telemed model is now exported in several hospitals to treat patient data and as consequence it increases the pool of medical device company data, making more reliable its predictions.

Unitedhealthcare[57](Fraud, Waste, and Abuse): UnitedHealthcare gives medical benefits and services to about 51 million of people. The organization counts with more than 850,000 doctors and care experts and around 6,100 doctor's facilities across the USA. Their Payment Integrity group (i.e. the end-to-end claims processing) is successfully processing without delay, more than one million claims each day (i.e. a minimum of 10 TB of information every day) basing the efficiency of such procedure to capillary management and constrain of big data type, storing and sharing. United Healthcare came up with a unique dual model strategy, which meant focusing on operationalizing savings, while at the same time pursuing innovation to constantly leverage the latest technologies. By using Hadoop for reviewing claims, prescriptions, medical plan, subject, contracted care providers, including over 36 data assets, UnitedHealthCare was thus able to building a predictive analytic models (like PCR, True Fraud, Ayasdi...) that can identify inaccurate or suspicious

declaration based on the systematization and repeatability. The results are the availability of a rank-ordered list of potentially fraudulent providers they can pursue in a targeted.

Possible Adverse Effects of Data Transparency:[58] On the other hand, it was clearly also possible to take advantage of big data revolution (and in particular about the data transparency) in healthcare by pursuing objectives that create value only for some specific clients. In healthcare, some operators could try to take advantage of big data, without regard to clinically proven outcomes. For example, owners of Magnetic Resonance Imaging (MRI) machines, in order to cut fixed costs across more patients, could be more proactive in identifying and selecting unwanted patients and disease areas. The data can thus be used to manipulate the market and the healthcare services, regardless of clinical need, and patients could end up pursuing and receiving unnecessary MRIs. Exporting this model on bigger scale, the big data analytic can in principle kill the healthcare value, since operators will invest more in beneficial treatment area and patient health and life condition will not improve. This ethical issues, involving patients, providers, and payers, demands for the definition of an "right care" values and protection respect to this kind of possible abuses and demand to see the suitable evidence presenting that certain services are essential.

3.3 How big data affect the health care system in the case of privacy and security [59]

The last years have demonstrated that Big data technologies have sensible enhanced quality of care given to patients, and, at the same time, they have enhanced the quality of life for patients in terms of life expectancy; but as already shown in previous paragraphs, there are some associated dangers, i.e. to the spreading out of healthcare information together with sensitive data (that could permit the identification of the patient) or the use of these information for prejudicial purposes. The potentiality of big data in improving the quality of patient care has as counterpart the risks associated to Privacy. Privacy needs to be respected, no matter about the advancement of medical science. This difficult doesn't involve only company's managements but also government institutions.

President Obama requested the assistance from the Department of Health and Human Services (HHS) and asked for input from healthcare stakeholders on the use of big data to benefit the health industry. In 2014 Obama asked the HHS to "consult with stakeholders to assess how federal laws and regulations can best accommodate big data analyses[60]." The HHS was also asked to "develop recommendations for ways to promote and facilitate research through access to data while safeguarding patient privacy and autonomy."[61] In the two sessions of December 2014 and February 2015, The Health IT Policy Committee's Privacy and Security Workgroup enumerate the advantages of the use of big data and the sensitive problem that big data could cause.

The preliminary results from these two sessions are well summarized in a virtual workgroup meeting in which co-chair of the Privacy and Security Workgroup, Stanley Crosley, believes patients' right to privacy must be preserved: Crosley said[62], "Patients should not be surprised or harmed by collections, uses, or disclosures of their information. Nowhere is this more difficult than with big data". In this framework, one of the major problems related to the use of big data in healthcare system is the potential use of the collected information against individuals. Healthcare data could be used by health insurance companies to decide about approval or deny of prime, setting the level of financial risk represented by each individual.

Responsibility and correct use of big data technologies minimizing the privacy risks associated is involved significant challenges. Currently the laws covering the use of data are inconsistent and incoherent, with some states prohibiting the use of health data to discriminate. On the other hand there are situations in which discrimination is "expressly permitted." For example, the Health Insurance Portability and Accountability Act prohibit the use of health information for

discriminatory purposes; however other entities holding the exact same data may be permitted to use it. The health industry may be heavily regulated, but not all industries have the same protections[63]

3.4 Health Insurance Portability and Accountability Act (HIPAA)

In 1996 Health Insurance Portability and Accountability Act (HIPAA) was published in Public Law 104-19. "HIPAA specifies the privacy, security and electronic transaction standards with regard to patient information for all health care providers"[64]. HIPAA set two noteworthy principles: insurance reform (pre-existing medical conditions should be kept independently from any possible changes in job position), administrative simplification (reduction of health care costs by standardizing information transactions)[65]. Electronic Data Interchange (EDI) has turned into a crucial issue in helping health care organizations meet the requests of high volume patient loads, as well as maintaining efficient business partner relationships and expand their costumer targets. This was leading to put a huge effort in the standardization of electronic data (think to the creation of National Standard Format, or NSF done by American National Standards Institute ANSI) in order to facilitate ease the sharing of electronic information within and across different platforms. In this sense HIPAA validates and assists electronic data transactions, but never losing the issue of privacy and security that may stem from converting to the use of vulnerable electronic transactions.
HIPAA includes Health Plans (managed care organizations), Health Care Clearing houses (billing companies, community health management information systems), and Health Care Providers (doctors, nurses, therapists). HIPAA monitors the standard code sets (for diagnoses and procedures) and privacy standards (mandating the use of Protected Health Information, or PHI), by the introduction of unique identifiers (such as a National Provider Identifier and Tax ID number), and security standards (physical, technical, and administrative procedures to secure PHI). HIPAA normally apply to all USA healthcare companies and it foresees for the one that are not in compliance with its standards fines up to $250,000 and up to 10 years in prison for managers[78]. Other HIPAA guarantees and benchmarks involved protection portability, medical plan freedom (i.e. the possibility to change healthcare operators in order to have the best treatments), and the fraud enforcement against Medical and insurance fraud.
According to HIPAA regulations, a calendar for the individual compliance of healthcare companies exists. The electronic transactions must be subject to standardization with a maximum date of October 16, 2003[66]. The most generic privacy regulations (referred to the most covered entities) required compliance by April 14, 2003 while it has been fixed to the April 14, 2004 the compliances for small health plans. Security standards for electronic PHI have an impending compliance deadline of April 21, 2005, except for small health plans, which have until April 21, 2006 to comply[67].
Oral, written, and electronic forms of PHI are detailed regulated by HIPAA. The motivation behind the Privacy Rule is to "meet the pressing need for national standards to control the flow of sensitive health information and to establish real penalties for the misuse or improper disclosure of this information"[68]. General principles built up inside the Privacy Rule are the utilization and exposure of PHI for treatment, payment, and health care operations,, the base important utilize and sensitive of data, formation of creation of de-identified information or a restricted informational collection, use of gauges to business accomplices through contract, application to data about the deceased, adherence to the notice of security practices, and application as secured elements parts of associations that are not secured elements[69].
As a rule, the Privacy Rule protects people's PHI by managing how and when a person's PHI might be revealed and for what reasons. It stipends people greater inclusion by permitting them particular rights to get to their medicinal records and to demand revisions, to approve or limit the exposure of

their data in specific conditions, to be informed of the route in which their data is imparted to others, and to be informed of their rights identifying with privacy.[70]

3.5 Security rule

The Security Rule of HIPAA has introduced specific format for the electronic form of PHI, specific standards of storing, transmission, confidentiality and protection against unauthorized users access and threats to security or integrity[71]. Also if the minimum security level covering healthcare big data is guaranteed by HIPAA, no standards for computer applications is identified, since HIPAA establishes behavioural standards and it requires interested operators to develop their own practices that accomplish to these behavioural standards.
More in details, the criteria established by HIPAA Security rules can be enumerated as it follows:
(1) Administrative safeguards: i.e formal practices to manage security and personnel;
(2) Physical safeguards: i.e protection of computer systems and the facilities within which they reside;
(3) Technical safeguards: i.e. it means to control and monitor information access, including technology to secure data-in-transit;
(4) Organizational requirement—business associate contracts;
(5) Policies and procedures and documentation requirements—similar to those within the final Privacy Rule.[71]
In this sense the compliance with HIPAA rules is obtained by first identification of risky areas that need be changed in order to satisfy the security standards, and then by design and implementation of adequate solutions to the problems. The most common solution is the creation of a privacy and security officer which is entitled to establish and to review the security and privacy administrative controls. Simultaneously companies provide a plan promoting the compliance and determining the internal policies and procedures to its realization. Examples of steps in the plan include establishing security certification processes for employees and contractors, updating employee records to indicate the level of security appropriate to each job position, assessing the compliance of the Management Information Systems (MIS) department with the Security Rule standards, and explicitly stating consequences for non-compliance with new security rules. [72]

3.6 Electronic information in health care

The "Information Age"[73] has prompted exceptional open doors for financial development in the medicinal services industry. The need to be competitive on the market was forcing all the healthcare operators to adopt the big data technologies, but keeping in mind the end goal from both financial and ethical point of view[74]. The main consequence of this revolutionized is the "software robotization"[75] This automation has revolutionized the industry in the way of time effectiveness and increased capabilities. Tracey Banks, from North Central Women's Health Partners in Texas, estimates that converting from paper records to CompuMed Systems Electronic Medical Records (EMR) system saves 10 min per patient[76]. In fact, Hooda, Dogdu, and Sunderraman state, "the efficiency of health care operations is directly linked with the amount of automation of the health care information system"[77]. However, this efficiency and cost reduction should not have a counterpart in the privacy of individual in terms of spreading out of private medical information. In this framework is important to understand HIPAA's Privacy and Security Rules and to recognize how privacy may be compromised by the new technology in term of how health care facilities use computers and other technology.
Prior to the massive introduction of computer networks in health care organizations, patient information was less accessible and paper records made copying, transferring, and storing

information difficult. Now in the *information age* an hospital using computers can record information about patient admission, discharge, transfer, laboratory results, radiology reports, and billing with an extremely reduce cost and, more important, this information is readily accessible to other operators[78]. A part from reducing cost and time in paper data processing, the big data technologies improved cash flow, management decision-making, and clinical ancillary services[85]. In this sense Computer revolution has the definition of Health care in "commodity-driven and improving coordination of services will bring the focus more toward the two billion patient encounters per year"[79].
However computer networks technologies are now necessary also to protect the accuracy of patient information. The healthcare system infect is now extremely structured, with operators specialized in different areas or services (think for instance to physicians with different expertises: generic doctors for check-up, orthopaedic doctors, cardiologists and so on). Each operators produces information which must be as much detailed and meticulously as possible in order to be read and exchanged with accuracy passed from one health care provider to another.[87]
To overcome this problem different solutions have been implemented. One of this is Record Link[80] (automation software) is highly used in hospitals, and according to Judy Ferraro, Director of Medical Records at Elmhurst Memorial Healthcare in Illinois, it drastically reduced the number of incomplete medical records by at least 75%[[81]]. In Massachusetts, Blue Cross Blue Shield has financed $50 million in creating a state-wide EMR system, which Governor Mitt Romney estimates could save the state millions in addition to improving the accuracy of patient records. Massachusetts may in fact be the first state to implement state-wide electronic medical records[82].
Universal Patient Record (UPR) is a nationwide project aimed to electronically store all patient health data in one single file, thus eliminating the problems related with collecting the medical information contained in separated medical records originated by different medical facilities or companies and permitting the transfer of this information between international healthcare operators[84]. On the other hand a deterrent to the implementation of such a system is the staggering financial burden associated with such an undertaking. Brewin relates various estimates by specific nations of the incredible cost of implementing even a nationwide EMR system. Despite obvious obstacles, i.e. the cost and required collaboration of such project, the prospective of a future in which doctors around the world could access to online records in order to better understand and collaborate, sharing information and intuitions, may hope to a significant improvement in the quality of health care. To compete effectively in the health care market and play a key role in it automation software and EMR look the only possible choice that guarantees the survival to companies. Indeed there is no an alternative option. It has been already demonstrate for the small hospital structure that refused to pass to robotization in favour of keeping the paper records the price to pay was incorporation in bigger structure or the extinction[84].

3.7 Privacy of electronic information in health care

If it is clear that big data methods must be mandatory included in the framework of healthcare system, it is also true that their appearance look responsible for a renewed and sudden preoccupation about privacy and security issues. Eddy described a medical record as consisting of "information divulged by a uniquely vulnerable human being, worried in some manner about the core of her very existence, to a trusted person with superior knowledge."[83,84] The information included in a medical record is extremely personal: it contains dietary habits, sexual orientation, sexual activities, employment status, income, eligibility for public assistance, history of diseases, treatments rendered, medications taken, diagnostic information, psychological profiles, genetic testing, family history, doctor's and nurse's subjective notes about patient personality and mental state, and much more [[77,86]]. In this sense a medical record is "an entire work-up of your being."[85]

This spreading out of this information could be devastating to an individual's social, religious life if used in the uncorrected way. Big data technology indirectly permits to this information (despite the intense private nature) the possibility to be sent across word, in a matter of seconds, making them available with just a click of a mouse. This is a radical change in the relationship between doctor–patient and it is as a consequence affecting the whole quality of health care system[86]. Private information that could be useful in diagnosis of specific diseases could be omitted by the patient if this latter one could be scared by the possibility of this information could be spread out and can be than accessible to other people. These threats to privacy can affect efficiency of treatment and in general the quality of health care has prompted the government to include privacy and security provisions under HIPAA regulation[86].

CHAPTER 4

4 Good Practices in healthcare system in a big data context

In the previous chapter we discussed several key data protection implications that can arise from the use of big data analytics in healthcare system. Now we turn to some of the compliance tools adopted by the systems in order to meet their data protection obligations and protect people's privacy rights.

4.1 Security practices to ensure protection of private health information under HIPAA

Guarantee the security of private health information by HIPAA-friendly environment is a complex procedure. This security can be achievable through high cooperation between all individuals of healthcare service organization, from the most elevated positioned chairmen to the nurses and clinicians who provide daily care. According to Schneider and Mercuri, many elements affect guaranteeing security, such as evaluation and alleviation of security and privacy risks, approach and systems advancement, appropriate reaction and recuperation, assessment of business partner contracts, employing and end effects, consistency and mindfulness preparing and training, accessibility control and validation, auditing, and intermittent revision and modification. Organizations with innovative ways try to adapt themselves with the Security Rule. The ISO Common Criteria (CC) of the National Institute of Standards and Technologies is a good sample which is engaging with a project related to national defense and it is well-known in health care industry in order to validation, design and records-keeping. By appearing each new challenges to the security of PHI, new methodology to address these difficulties are presented. Consider that these changes and applies can be time-consuming which needs phased-in approach to give workers and MIS division both need sufficient time to control , investigate and troubleshoot as issues arise. In general, security is tended to in accordance with to administrative, physical and technical protection.[86], [77]

4.2 Administrative security safeguards

Administration has significant roles and influence in building up consistence with HIPAA security measures. In order to have high level of security, it is vital to invite professional officer in security and privacy for cooperation and the foundation of a security and security committee. All the strategies with the group cooperation can be devised to guarantee consistence with HIPAA, make approaches that reinforce protection and security, set up a representative/temporary worker certification on HIPAA standards, refresh worker records to show every worker's level of trusted status, and survey the adequacy of the MIS division in following new HIPAA principles. Moreover, results for inability and failure must be obviously expressed in order to agree to new privacy and security strategies. Administration is a principal decision maker to set the case for the rest of the organization and make them aware from changes. Importance of HIPAA policies depends on attention and effort of management's staff. They can influence on employees in a way that HIPAA policies being followed seriously or to be neglected.[87]

4.3 Physical security safeguards

Physical Safeguards are an arrangement of standards and rules sketched out in the HIPAA Security Rule that attention on the physical access to Protected Health Information (PHI). It is responsible for guaranteeing consistency with HIPAA's Security Rule may be the most straightforward approach to start the consistence process. Interestingly, Administrative Safeguards concentrate on strategy and policy, while Technical Safeguards concentrate on data protection. For instance, physical access controls may be utilized by an organization in order to confine the availability of information. There are many way to describe security safeguards. For example, using electronic swipe card or simple locks door to get to PCs or rooms that contain sensitive patient data, or have double-secure workstations such as recording storage areas, in order to increase protection and prevent unauthorized use of the computers[70]. Eliminate unnecessary data is other way to protect sensitive data. All information being saved by Technical security safeguards Electronic Data Interchange (EDI) for use by staff of healthcare. However, using network and computer may facilitate the process of gathering data, can lead to security risks. Raymond Panko, a leader in telecommunication networks and security, suggests, ". . . security compromises (successful attacks) are both widespread and varied" [87]. Virus, worms and their combination can be considered as important security threats which have destructive effect on computers. A Trojan horse is a good sample of threats which influence computers in hidden way. In order to confine these kinds of attacks, high performance and alternative Security safeguards are crucial. Firewall protection, host security, and cryptography elements, are simple kind of Security safeguards. The objective of security and HIPAA is to ensure patients' rights and benefits. Accordingly, many sorts of attacks can be neglected because of financial shortage which leads to decrease support and enough attention to the computer network and privacy. Attacks can be stopped by sophisticated firewall, by passing various levels, before arriving to the computer host. Firewall performed like filtering the packets which coming through a router. Next step is controlling packet by firewall which is provide by ISP and then it will checked in the application layer which could be an example of an anti-virus software (common examples are Norton AntiVirus TM and McAfeer Virus-Scan) before it be executed on the host computer. One of the most important cryptology that is recognized as HIPAA-friendly is fundamentals of cryptology. Classification, confirmation, message integrity, and anti-replay protection are four kinds of crypto- graphic systems which are defined by Panko. Fulfillment these techniques would support parties, guarantee of right messages, and assurance if a message is hacked by attacker, never arrive to the destination in order to protect receiver from attack. Link encryption or link level or link layer encryption is the data security process which data may be encrypted at the data link level as it is transferred between two points through of network. This method is prevalent in the lowest protocol layers and related systems are using Virtual Private Network (VPN) with many standards such as Secure Sockets Layer/Transport Layer Security (SSL/TLS), Point-to-Point Tunneling Protocol (PPTP), and IP security (IPsec). Link encryption has been used successfully within organizations, including the military, where the security of each link can be assured and in the near future will be implemented with wireless technology. Two-factor confirmation security method consists of username and password is used in VPN which guarantee privacy of user. Moreover, magnetic strips as third factor can boosts security. Biometrics options such as Token key generators also can increase security for organizations like hospitals or insurance company which are keeping sensitive information. Positive side of this setup is that confined communications, gain permission through options such as packet filtering, routers and switches, firewalls, authentication and encryption[88].

4.4 Solving web tracking in health care system

To control and limit tracking techniques, many strategies, methods, and tools are used. Some tools are available which are using defence strategies. In following section, some solutions are compared and discussed.

- **Blocking Advertisement Services:** Using blacklist in order to defence host computers, is considered good solution. As an example, Microsoft tracking protection list (TPL) is added to Internet Explorer 9 recently. According to [89], this list is characterized by precision of 96.3% and recall of 72.2%. It is possible to avoid third-party tracking by blocking connections to all third-party websites, however, only 37% of the requests to third-party websites are considered directly to a tracker.

- **Hiding the IP Address*:*** Other well-known solution is hiding IP address. Most of simple tracking techniques are using IP address, although most tracking strategies are designed base on application layer. In order to hide the IP address famous method from the remote site is to use anonymous proxy servers, virtual private networks (VPNs), or Tor. Proxy servers act as a mediator in all communication between two parties. They provide anonymization services which its responsibility is hiding IP address of the user or/and removing cookies from HTTP requests. Bypass internet censorship as the user connects, is the other service that is provided by proxy server. This service blocks the IP address of the target website and fetchs the web siteand displqyes it on host computer. Zend2.com and KPROXY are considered as a web-based anonymous proxy services that are provided free. In order to guarantee security in remote access application VPN protocol has been introduced. VPN nodes are located in singile networt secure the tunnel which is connection to particular VPN, by many protocls such as Point-to-Point Tunnelling Protocol (PPTP), Layer 2 Tunnelling Protocol (L2TP), Internet Protocol Security (IPsec), Secure Sockets Layer (SSL), or OpenVPN. SecurityKISS[90], CyberGhost[91], USA IP PPTP/L2TP/OpenVPN VPN service, and VPNReactor. L2TP/PPTP/OpenVPN VPN service. It avoids viewing user Internet connection; physical location from realizing what sites have been seen by user. Moreover, it gives you a chance to have access to the sites which are blocked. Tor Browser makes possible to use Tor on Windows, Mac OS X, or Linux without needing to install any software. It is self-contained (portable) and can be executed by USB flash device in order to secure anonymity.

.

- **Modification of Data Sent over the Network:** Privoxy[92] characteristic is a non-caching web proxy HTTP and SSL proxy server, with abilities of filtering for improving privacy, adjusting information and HTTP headers or controlling ace ss and expelling advertisements and different harmful Internet junk. Privoxy has a flexible design that can be figured easily according to user's interest.

.

- **Opt-Out Cookies*:*** Opt-out cookies are cookies created by a website on user's browser in order to prevent installing future cookies from that same website. They stop third party, installing cookies and tracking user's web page. In fact provide ads according to tracking instead of stopping them. The disadvantage of using opt-out cookies is that they can block cookies from a particular server, so they should be renewed manually and be managed via browser settings[93]. Moreover, cookies can be lost by cleaning the history. However, according to [94], less than 1% of web browsers are reported to use them.

- **Do Not Track. (DNT):**[95] it is a relatively new service in a web browser setting that informs service in order to disable its tracking of user. It is a new type of HTTP header field. By activating DNT in the browser, special signal will be sent to websites in order to stop tracing user's activity. Consider that this signal is a header, and not a cookie. Then by cleaning cookies from browsers, the performance and functionlity of the Do Not Track flag can not be changed. Target of DNT is working as a simple and persistent opt-out fron tracking96.

- **Using Privacy-Focused Search Engines*:*** This method is a Privacy-focused search engines (e.g., DuckDuckGo[97], Startpage Search Engine[98], and Ixquick Search Engine[99]) promise that your privacy will never be threatened. It means no data will be stored by website and each time user considered as new user that visit the web site. In this method which is used in privacy-focused search engines, there is no aim to forwarded user's information to the visited website.

- **Private Browsing Mode*:*** By using Private Browsing, user can visit websites without creating a search history[100]. Most of the important browsers provide this feature in order to cover user tracks and keep user anonymity. This name for private browsing is different in each browser. For example some main browsers (Firefox, Safari, Chrome, and Internet Explorer) they secure user computer and hide the visited site from local attacker. Private browsing is also recommended for checking emails on somebody else's computer that prevent any cookies saves the account information. Another threat is referred to the fingerprinting techniques, for example, used by fingerprintJS[101], which provide the same information in normal and confidential modes.

- **Clearing the Browser Cache and History:** This method, on the condition that the user properly cleans all storage in which the browser and its plugging can keep information, behaves almost in the same way as private browsing. The only exception is Safari, which in the private browsing mode continuously serves cookies and history from the earlier browsing activities.

- **Execution Blocking:** There are several browser extensions, which are able to block the execution of JavaScript (e.g., NoScript[102]), Flash (e.g., Flashblock[103]), and other script content. However, other tracking methods are not impacted by using these extensions.

- **J. E-Mail Aliases:** Email address is the most important information which should be protected seriously. Email address can be accessible easily by websites and enable them to create several account to the similar client or lead to increase spam email with unwanted advertising. In this method an email address is provided with another destination email address. Instead of using the original email address, email alias is used to send email to the receiver. The benefits of this method is preventing spam
  Free e-mail aliases are offered from many provider such (e.g., Inbox Alias[104], 33Mail[105], Jetable[106], Mailexpire[107], TrashMail[108], and Guerrilla Mail[109]), which can be configured in various lifetime, from 5 min, through one month, to being infinite. Advantages of this method is counteracts spam and ensures the clients' protection by providing safe situation in order to use their email address without any concern and make it public easily. Other reason to use this method is avoiding receiving Self-Destroying File Systems which are sent to host computer by email address in order to share sensitive data, althouh they have limit lifetime.

4.5 A philosophical prospective: humanizing Big Data

According to Strong[110], marketing researcher and author of ""Humanizing Big Data: Marketing at the Meeting of Data, Social Science, and Consumer Insight," any size business can use Big Data for better business decisions if they focus on the often neglected source of all this information, individual human beings. Humans and human behaviour are the engine behind Big Data. And, of course, understanding humans and human behaviour has always been a necessary part of succeeding in small business. "Humanizing Big Data" claims that value and power of data for all businesses is clear. Data itself is not dangerous when it is used for improving technology without considering the customer in the process. Problem appears when this information impact on human behaviour. In "Humanizing Big Data", Strong suggests that companies utilizing (or considering) Big Data should consider the arranging and ramifications of doing as such. In other to choose right approach, businesses should focus on metrics (sales, volume, etc.) that influence their bottom line. Correct metrics accompanied by context and understanding to the numbers that create Big Data can help business's progress. Strong also emphasize that customers should be considered in Big Data solutions by companies and negative impact of this information should be eliminated. He says that "Humanizing Big Data" in productive for all businesses, small and big. After all, Big Data is found everywhere: online, on social media, in email newsletter services, etc. important conclusion is gaining enough Knowledge in order to use data in a safe way instead of the focusing just on technology.

4.6 Exascale computing

Big data technologies have created new specialization skills. New companies have been created just to deal with data (management). This is the first step of towards a new framework in which, in order to minimize cost and educational training, companies are organized/communicate in infrastructures (hence the new IaaS; Infrastructure-as-a-Service). This is a main revolution in the operational procedure: before, companies were keeping data for themselves and all resources were found in "in house" and internet connection was not mandatory for many tasks. Today it is not realistic and competitive to develop applications, databases decision making machine and analytic methods without steadily get new input and compare solution in the network. Despite this extremely competitive environment, companies are still hesitating to outsource data management, i.e. the direct possibility to store and consult their own data when they more need. However there are specific cases, like for instance predictive medicine, where the big data analytic could guarantee remarkable improvements only if computational performance and processing capacity will be upgraded not only in terms of software, tools, and algorithms, but also in terms of computer facilities. This has led infrastructures to adopt Exascale computers—machines that perform one billion calculations per second and are over 100 times more powerful than today's fastest systems. Only these powerful processors indeed could break down the now-day limitations in storing and analysts of heavy clinical data (for example the genomic map or the easy access and visualization of 3D medical tomography). [111]

4.7 Infrastructure

The continuous change of volume, variety, veracity, and acquisition speed of big data require a continuous evolution of infrastructures in order to address and exploit this new flood of information. For this reason the most successful and reliable infrastructures have aIaaS level, and on the top of it a bases in which individual platforms and services can be provided. Infrastructures thus choose the virtualization and cloud computing[112] as more efficient and user-friendly platforms to develop, capture/manipulate, and store the huge amount of healthcare data[113]. The now-day infrastructure framework have already developed the skills in order to face the Big Data tsunami:

some companies like Hadoop, MapReduce, MongoDB, Cassandra, Lucene,- - see glossary – are already competitive on the market. In any case, despite these huge improvements, several questions remain still open. Dynamics, scalability, parallelization of code and CPU processing, and intercommunication within infrastructures are issues that still need to be addressed. Numerous applications and platforms, infect provide directly service from the cloud framework without guarantee about versatility, process speed and capability to work with non-traditional databases. For this reason, the existence and functionality of cloud infrastructures require a massive invest in solutions oriented in solving these issues, in order to effectively enable reliable execution of, for example, machine learning algorithms, pattern recognition of images, languages, media, artificial intelligence techniques, semantic interoperability, and 3D visualization and other services.
Moreover, the delicate information contained in Healthcare data fixes specific requirements on infrastructures (like regulations, security and privacy compliance, reliability). Several cloud based platforms, like Philips HealthSuite Digital Platform 4950[114] are actually in use in the Healthcare system.

This latter one is an example of multi-client cloud platform where clinical and medical systems and devices data are acquired, combined and analysed. Several access are entitle: different healthcare operators (doctors, physicians, researchers…) as well as individuals have access to personal health data, patient conditions on large scale populations, thus, offering for instance the possibility to doctors to personalize and make more accurate medical treatments on the base of more robust clinical diagnosis. The importance of cloud computing was recently highlighted by the European Commission through its European Cloud Initiative. They proposed a European "Open Science Cloud; a trusted, open environment for the scientific community for storing, sharing, and re-using scientific data and results; and a European Data Infrastructure targeting the build-up of the European super-computing capacity".[115]
Big Data for the healthcare community will be the most active partner of the European Open Cloud, since the data able to collect are a unique source of information for any kind of analytic developments.

4.8 Data integration

Big Data Repositories or Data Warehouses have daily to deal with information produced by various sources having specific format or sometime data come in unstructured forms.
The integration of different way of information require thus three operations, specifically, Extract, Transform and Load (ETL). ETL processes are thus responsible for identifying in the heterogeneous medical data records and results that can be considered structural, syntactic, and semantic. The syntactic heterogeneity appears in forms of different data access interfaces that need to be interpolated and mediated. Structural heterogeneity refers to different data models and different data schema models that require integration on schema level. Finally, the process of integration can result in duplication of data that requires consolidation. Structural heterogeneity deals with different data models and different data schema models that in order to be integration on entry level needs to be organized in categories. Finally, the process of integration can result in duplication of data that requires re-organization. Information extraction, machine learning, and semantic web technologies are the keys points in the process of data integration, in order to get rid out of un-necessary context data and make free relevant data. Information extraction is thus the mean to discriminate valuable from un-valuable data improving the accuracy of data integration routines, avoiding duplication and assuring data alignment. On the other hand active learning machine and algorithms ensure the efficiency and reliability of the automatic data integration routines in order to meet the desired level

of data quality. Finally, the semantic web technology gives the opportunity of a rapid visualization of data by selecting adequate categories trough the knowledge based graph and ontology: data can be represented by important concepts in a common framework of categories. The use of standardized categories facilitates collaboration, sharing, modelling, and reuse across applications[116].

CHAPTER 5

5 Current Available solutions

The distribution of big data inspires many companies to develop healthcare applications or similar innovations. To assess this trend, several company profiles and business models; according to the practices and method which were discussed in previous chapter; are reviewed in the following.

Big data impacts all the economical fields, proposing functional solutions to critical problem of the planet. Arguably, the most impressive change involves healthcare system and insurance companies, whose business area has demanded for long time changes and big data are currently creating new prospective. Medical services and healthcare companies always need to be updated respect with clients' requests and clients expect better engagement and better care conditions. In addition, the industry need to take care of government compliance and regulations to address specific issue (see HIPAA regulation in chapter 3), reducing the costs and enhance profits. They also need to improve access to healthcare solutions along with better risk management. Big Data has answered to these open questions: the huge amount of healthcare data are created daily by medical facilities (EMRs, labs, therapeutic treatments, CRM systems, medical prescription, etc, etc…). Processing this information by big data analytics has helped healthcare facilities and providers to improve care and delivery outcomes. Big Data has entered healthcare in the following ways:

- **Preventive healthcare**: Most of healthcare services are now cantered on curative healthcare. Nevertheless, the preventive healthcare value will be soon dominant. Information collected by applications and wearable can be shared between specialists who can address special aspects of diseases. This will likewise shield light of possible legislation modification in order to prevent risk or dangerous situations. Preventive healthcare will thus lower companies and government costs and enhance knowledge. Information assembled from medicinal records, protection records, applications, wearable, IOT sensors and online networking can all be used to create a more effective social systems.

- **Efficient Systems**: Big data analytics can transform the healthcare system into a social system in which operators have enough knowledge of the patient to suggest him best customized solutions. This can lead to improved resource utilization, and on long time scale to bring to expenses reduction followed by more fruitful outcomes. Big Data become thus a critical component in the designing of smooth efficient medical services frameworks.

- **Patient engagement:** Big Data helps specialists and other operators to get more insights about patient diseases. It is worthwhile to notice that in the framework of big data doctors have for the first time the possibility to prescribe unique healthcare solutions for patients instead of delivering generic care. This can change the relationship doctor-patient and the same time constitutes the basis for a culture of good heath life.

To encourage the exchange of information between healthcare operators, the Health authority in different nations has defined a standard format XML. They are further made available in various sub formats to provide transfer of: i) basic Patient Medical history, ii) complete Patient History including all changes done to the Patient records[117].

.

In the following some successful examples of commercial solutions are reported [59].

Opentracker®[118] distributes database solutions that acquire and store any conceivable sort of information. In order to simplify post-processing coding, **Opentracker®** comes with an effective application programming interface (API) that customers use with a specific aim to deliver distinctive question to the database. This api allows to easily to face the so called 'Data Tsunami' – i.e. the fact that a datatream for even a single user via social media such as Facebook, Instagram, and Twitter, networking via LinkedIn, or a consumer site such as Amazon contains innumerable pieces of information to be counted and put to use. In the following we illustrate one example, a sub-program to check of medical exercising. In this case the information pool will be made of individual records and separate datastreams (every step taken, start times, finish times, distances, average speed, calories burned, sessions, temperatures, weight, milestones, and so forth). Each information will be identified by a entry plus a passage/flag. **Opentracker®** takes care to store and chronological following these entries make them accessible and suitable for analytics methods (map reducing, hadoop…).

Asthmapolis®[119] has built a GPS-powered tracker that screens inhaler utilization by asthmatics. The data is ported to the main database and used to recognize individual, group, and population based patterns and is then merged with Centers for Disease Control and Prevention (CDC)[120] data about known asthma impetuses (for example, dust numbers in the Northeast and the impact of volcanic haze in Hawaii) to enable doctors to create customized treatment arranges and spot anticipation openings.

Ginger.io®[121] offers a portable application in which patients, (for example, those with diabetes) agree to be tracked by sanitary operators through their mobile phones and the doctors help them with behavioural health therapies. By checking the mobile sensors of smart-phones, the application records calling, messaging, places, and even the movement of data. The Ginger.io application coordinates this data with public research from the NIH and sources of behavioural health data. The data collected can thus be used to develop models. For example, the absence of a specific regular action could be a flag about patience health, as well as sporadic sleep can be interpreted as the signal of anxiety attack.

mHealthCoach®[122] bolsters patients on chronic care medication, giving training and advancing treatment adherence through an interactive system. The application use information from the Healthcare Cost and Utilization Project, supported by the Agency for Healthcare Research and Quality[123], and in addition results and notices from clinical trials (taken from the FDA's clinicaltrials.gov site). mHealthCoach can likewise be utilized by suppliers and payors to distinguish higher-chance patients and convey focused on messages and suggestions to them.

Another critical issue is the mix of clinical research systems conforming Big Data repositories. Integrating clinical research networks is a value broadly recognized by researchers and funding agencies, since connecting networks means clinical research more effectively and conforming communities to shared operational knowledge and data. **Li Ka Shing Centre for Health Information and Discovery of the University of Oxford[124]**, recently supported by a £90m initiative in Big Data and drug discovery; or the **NIH Big Data to Knowledge (BD2K)[125]** initiative

enabling biomedical scientists to capitalize on the Big Data being generated by the research communities.
Together with **Amazon Web Services[126] and Hitachi Data Systems 808**[127], the **Philips HealthSuite digital platform[128]** analyses and stores 15 PB of patient data that is collected from 390 million imaging studies, medical records and patient inputs. This empowers healthcare providers to efficiently impact patient care.

e-Zest®[129] has a dedicate healthcare services which is based on the straight collaboration of the major companies in this area. **e-Zest®** created particular solutions for healthcare analytics and healthcare management, among different geographical areas, enabling the use of big data technologies in provide cutting edge solutions and treatments.

e-Zest® gives business solutions for little and medium companies everywhere throughout the US. The real end clients of the application were HealthCare accreditation organization, by building up a framework that will help customers to provide maximum business. In this latter case, using a *.NET* coding, ***e-Zest®*** has developed a healthcare package where client in-house application (written often in Microsoft Access 2000) are still functional and compatible. ***e-Zest®*** engineering and development methodology spans from conceptualization, architecture, design, development, testing, deployment and enhancements to porting and ongoing support. The ***e-Zest®*** application serves as an electronic health record system for organizations that deliver healthcare as a service, such as hospitals and physicians. The system provides services to healthcare practitioners by providing them one stop solution for managing the patient information in a user friendly way. The application caters to the security concerns by displaying records through different levels such as therapeutic relationships, confidentiality and getting the latest information through government managed servers. The application is cloud-based software. It is written in Java, with a JAX-RS server-side REST API, and a JavaFX rich client that supports online/offline operation and secures data synchronisation. It also has a mobile client, written as a single-page HTML5 / JavaScript app that communicate directly with the REST API via Cross-Origin Resource Sharing (CORS).

Healthcare companies are facing the integration of new sources of information – whether it's payers payers accessing clinical and lab data when setting up Accountable Care Organizations or providers looking at claims data for driving Value Based Programs. Transparency can be only guaranteed if, both insurances and medical services suppliers are taking into account the social information coming out from physicians and doctors as an extra figure of merits. Supplementing this is the utilization personal medical devices, sensors and wearable by patients that provides additional diagnostic information.
Sagitec's platform has additionally an essential role in industry (specifically catalysis). The Sagitec Framework™ is object-oriented, database-agnostic software architecture based on top of Microsoft .NET. It has supporting and auxiliary programs, compilers, code libraries, an API and a well structured collection of high-value tools and components that take the capabilities of Microsoft's .NET language to the next level. The Framework is on incremental upgrade path – with each new release of .NET, the Framework flawlessly consolidates the upgrades without unfavourably affecting the business functionality that lies on top of it. As the Framework changes, so do singular client application – without paying attention to when the code is modified. Beyond the Sagitec Framework™, there are other management services software that can evolve and be used independently. This layered structure gives the flexibility to the code to include new components and utilities when customers require them. Moreover this design guarantees the customers less risks of "technical obsolescence".

HealHub™[130] helps biopharmaceutical companies and medicinal device producers to manage the process of ingesting any type of healthcare data, merging appropriate datasets, creating models and exposing insights through APIs. The pre-built package contains:

- Episodes of Care. HealHub™ uses 13 definitions for segmenting claims into appropriate cases.
- Patient Behavior Segments. Get a composite HUC (health risk, utilization, and cost) score to identify patient segments for target interventions
- Disease Cohorts
- Group patients into disease cohorts based on diagnosis codes and claims
- Social Media Integration

Acquiring and online analysing information, blogs, and websites for pre-determined measures such as physician ratings, product issues, and sentiment data on brand or health plan performance. Data information researchers can utilize pre-manufactured utilities inside HealHub™ as a starting point for models for transition in care, gaps in care, patient behavior, and provider performance. The operations group can then utilize the Ops Workbench[131] to interpret those models and to assemble lists, reports or integrate with downstream systems for further analysis. HealHubTM gives rapid access to significant knowledge – more than 100 data models including episode groupers, disease cohort builders, and segmentation by health behavior and cost drivers to assist operators and involve patients better. Implementation time for an organization's digital health stack is thus reduced by 70%, and the price of the service is less. HealHubTM utilizes Microsoft Azure widely for machine learning, analysis, and business intelligence. Moreover for what concerns enterprise-class security regulation, Azure is accomplishes the HIPAA requirements.

Specific standard language has been developed in the healthcare system. Here few examples:

MetaMap®[132] is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover national library of medicine Metathesaurus[133] concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques.

Apache cTAKES[134]: clinical Text Analysis and Knowledge Extraction System is an open-source natural language handling framework for free-text data extraction from clinical electronic healthcare records. It elaborates clinical notes, distinguishing them, sorting them, and classifying clinical structure names, drugs, diseases/disorders, symptoms effects, anatomical area of interest, and treatments.

cTAKES built using the UIMA Unstructured Information Management Architecture framework [135] and OpenNLP natural language processing toolkit. Its components are specifically trained for the clinical domain, and create rich linguistic and semantic annotations that can be utilized by clinical decision support systems and clinical research.

These components include:

- Named Section identifier
- Sentence boundary detector
- Rule-based tokenizer
- Formatted list identifier
- Normalizer
- Context dependent tokenizer
- Part-of-speech tagger
- Phrasal chunker
- Dictionary lookup annotator
- Context annotator
- Negation detector
- Uncertainty detector
- Subject detector
- Dependency parser
- patient smoking status identifier
- Drug mention annotator

openEHR[136] is an open standard format in healthcare informatics that use electronic healthcare records (EHRs) to issue management and storage, transferring and access of health information. In openEHR, all healthcare information for a man are stored as "one lifetime", vendor-independent, person-centred EHR. The openEHR specification includes an EHR Extraction format also if the generally speaking, there is not a uniformly accepted standard between the different healthcare operators (see for instance the two standards EN 13606 and HL7).

The openEHR specifications are maintained by the openEHR Foundation that is a not profit company supporting open research, development, and implementation of openEHR records. The standards of openEHR are the results of a mix of 15 years of and Australian research and development into EHRs and new models, including what has turned out to be known as the paradigm methodology [[137,138]] for specification of content.

The openEHR specifications[139] include information and service models for the EHR, demographics, clinical workflow and archetypes. They are designed to be the basis of a medico-legally sound, distributed, versioned EHR infrastructure.

Security on the treatment of Big Data in healthcare system required a robust architecture. The security architecture should assure fine-grained security, rich audit functionality, directory services, and segregation of duties into a unified security framework. Client systems (clinic, hospital, patients…) ask for full auditability, including mobility, differential content access, and continuous check about possible entries updates. Where necessary, security issues can be associated in identifying and limiting the access by IP or the more specific Media Access Control (MAC) address associated with LAN hardware.
Dynamic security Framework from ***Sagitec®***[140] provides the benefits of a package-based solution (including a rich set of functional designs, best practices, pre-built software modules, software architecture, and configuration tools) and inherent flexibility to adapt to the business environments and challenges of healthcare organization.

CHAPTER 6

6 Research direction

After the discussing status of art of the problem and compliance tools nowday in use, we present the open issues, related with the subject.

The health care industry is not the only one interest in protecting costumer's privacy and security. The Electronic Information Age has affected almost all aspects of human life. Actually, the healthcare services industry has profited by the new technologies offered by internet and big data analytics. Efficiency, reduction of costs, better outcomes are just few aspects of Information Age.
However, each internet service provider needs to deal with the continuously increasing request for protection and security concerning their transactions. E-Business clients would not agree to the trade or exchange of their private data to different organizations without their signed assent. For instance, if a business factory "A" sold machinery to a client C, the business company "B," who sells auxiliary equipments, may be greatly interested for advertising purposes to the data of C in possess by "A". On the other hand "A," would not have any interest to pitch client data to increase the volume of business of "B" without the assent of their clients also because of the danger of losing the confidence and the trust of C. Similarly, there are the individuals who might benefit from the buy of medical data, for example, a drug company that may benefit from acquiring the list of persons affected by a specific disease. The health care industry stands to benefit from the same security technology that e-businesses use. However, all the promise of big data revolution is in direct conflict with privacy regulation that greatly limits the use of electronic-PHI. Evidence-based medicine has great potential, but regulations of ePHI make it extremely difficult to aggregate all the evidence. Although search engine companies like Google collect health-related information all the time through search history, use of this type of information by covered entities under HIPAA is a difficult legal issue that has yet to be explored fully. One promising strategy adopted by some providers is the use of security Platform for privacy protection based on the (P3P) project. Basically, P3P is "a standard that is administered by the World Wide Web Consortium (W3C)" that "provides users the ability to control the use of personal information on the web site they visit"[141]. P3P has many useful features, such as enabling web sites to state their privacy policies in a standard format, which browser plug-ins receive in both machine and human readable formats. In other words, the decision about whether or not to accept a site's privacy conditions is automated. Clients can customize their preferences about the sharing of their data. A considerable number of the elements included in P3P are accomplished HIPAA requirements, like recognizing how individual data will be shared and providing to the user the possibility to review and modify if possible the data. P3P 1.0 standards were released in April 2002. Until now, roughly 20% of the main 500 web sites are using such specifications, and the number of new web site moving to P3P solution is increasing over time[151]. However still currently, a certain number of issues (like for example, P3P its experiencing technical errors and violations of P3P specifications by particular sites), have make weak the P3P project and the security and privacy are still an open problem, also if P3P demonstrates to be a valuable protocols in both electronic-business and medicinal services systems. On the other hand healthcare system need to face the web tracking phenomenon, that, keeping notes on individual's health, has been shown to be a tool for improving the marketing performance, but at the same time a danger in terms of privacy. The continuous improvements of tracking strategies make the specialists always work away on this subject. The primary task to be issue to by

information researchers is finding out how new tracking techniques are thought and developing identifying and neutralising solutions[142].

There are some approaches now under developments:

- Analysis of scripts originating from advertisement of known suppliers. JavaScript are at the basis of the most common tracking methods. Analysing the scripts, discriminating the domain names and the IP addresses is possible to identify the functions used in the browser and neutralize the code, by disenable the corresponding browser functionality. Sometime script is hidden inside the html code, making their identification harder. Defence mechanism is already installed as plug-in in common used browsers. This is the case of Firefox add-on showing which JavaScript code is running on a site page[143].

- Passive monitoring of network traffic helps to detect anomalies, which can be the results of the malicious active tracking methods. This concern traffic generated by the same IP number towards the IP numbers of advertisement providers or web tracking companies. [144]

- Depending on standards for internet pages (HTTP protocol and related versions), the tracking possibility can be enhanced or reduced. The analysis of the page extension could thus help in identify tracking phenomena. [145]

- It is possible to update protection system by checking continuously web program change logs and developer's technical blogs

- Analysis of behavior of popular web browser plug-in/ add-ons.

Putting aside the confusing overlaps of regulation, the value of more information to effective treatment and the pervasiveness of technology in everyday care, healthcare privacy professionals still face major challenges ahead. The value of stolen medical records is now higher than consumer credit information, making the healthcare industry a new target for cyber-attacks. Medical records are already some of the most regulated and protected data sets worldwide, but it seems that in the next few years, the necessity and difficulty of protecting them will only increase.

CHAPTER 7

7 Potential PhD project:

7.1 Methodology

Methodology Easy reading and protected access of EHR

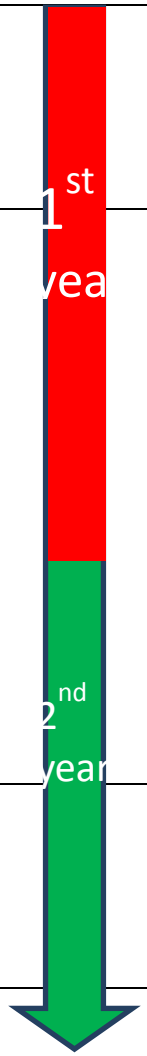
Electronic-PHI and multisource electronic health record (EHR) have become new valuable objects in the now-day healthcare system. Computer age and big data revolution have infect change dramatically the acquisition, the analysis, the storage and the sharing of medical information, before based on paper records. Changing medical records into electronic data has represented a big challenge for all healthcare operators but, at the same time, an improvement in term of efficiency, management costs, and quality of medical treatments.
It has been shown (in this thesis) that efficiency, cost reduction, and improve of life quality for the health companies and patients passes through the *creation of complex databases* (in which information, originated from different healthcare operators – generic doctors, hospital, service companies need to be integrated and complemented each other's) and the d*evelopment of algorithms of analysis* (which allow to use these information for predictive models).
On the other hand, security and protection of privacy are some of the fundamental factors in the development of high-quality tools in the healthcare sector. If no attention is paid to these aspects, there is substantial risk that individuals may come to harm in healthcare situations.
Several open questions remain (see the previous session of this manuscript).
Due to technological advances, addressing the growth in the size of data sets used by science, medicine and industry has become an increasingly central concern. Moreover different health information systems are not necessarily compatible with each other and thus information exchange of structured data is hampered, making thus extremely complicated the creation of "universal" pool of information. Finally the compliance of HIPAA regulations (i.e. privacy and respect of human rights) in some case looks still far from to be totally accomplished.
In this project, I propose to explore a new way of collecting, sharing, and privacy protecting healthcare related information by advancing data-mining algorithms for converting healthcare big data. The basic idea is thus addressing the following issue: handling of different types of data, sampling and bias them to remove "structural noise" (i.e. not relevant data). At the same time I propose a selective mechanism of access to the relevant information in order to assure the maximum privacy and discretion to people.
It has been proposed that the exchange of EHR data between different suppliers can be based on archetype methodology: in this sense EMR Data Extraction, Migration and Storage become an essential part of healthcare systems.
However the experience of using archetypes in deployed EHR systems is quite limited today. Currently deployed EHR systems with large user bases have their own proprietary way of representing clinical content using various models. Due to the complexity of medical terminology, the overall number of medical data models is very high, which makes the data sometime no accessible to the vast of the scientific community. My project is thus focused on the investigation of the main EHR content models and their flexibility of content representation. This could be a first fundamental step towards the migration of medical records to an archetype modelling, suitable for a data integration in the light of transparency and harmonization of the content.
Evidence suggests that if not carefully implemented, big data technologies could lead to potentially dangerous situations, in term of privacy, since big data sharing turns into the spreading out of

sensitive data. Human data must consider issues of data confidentiality, identifiability, extent of consent, and data usage agreements. All these ethical, social and legal aspects must be incorporated into a differential management of restricted access to sensitive data.

To take into account this aspect, in my project, I also propose to equip the archetype database of EHRs with differential privacy accesses: not all the digital documentations of a patient will be simultaneously available for all the operators and special authorization codes will prevent undesired accesses.
To realize such project, I propose to use open source based software (like Resource Entitlement Management System[1] or similar). This will allow the creation of a secured system without increasing companies' costs. The basic idea is that different users will have access to an uniformed electronic records database after providing their federated user identification number and after log into the system. Users will thus fill a data access application and agree to the dataset's terms of use.

7.2 Working Phase

In the following a possible time table of the project is presented.

Time	Activity	TOTAL TIME
Months 0-3	Preparing an initial research plan with supervisor. Preparing list of principal references on which the work will draw, and set objectives for the first year of study.	
Months 10-21	Analysis of EHR content structure structures. Analysis of differential privacy accesses database solutions available. Early Stage Assessment of PhD thesis to be submitted by the end of the first year. Developments of new solution (design of conversion EHR algorithms, definition and implementation of archetype models database with differential privacy access) Attendance to conference for ideas sharing and discussions. Presentation of results achieved.	
Month 22 – 24	Developments of new solution (design of conversion EHR algorithms, definition and implementation of archetype models database with differential privacy access)	

[1]Resource Entitlement Management System (REMS). https://confluence.csc.fi/display/REMS/Home. Accessed 14 Feb 2017.

Time	Activity	TOTAL TIME
Months 25-36	Developments of new solution (design of conversion EHR algorithms, definition and implementation of archetype models database with differential privacy access) Submit of third year report on the thesis. Thesis submission. Thesis defence	3rd year
	Submit of second year report on the thesis.	

Conclusions

Big data are the "new" business and scientific cutting edge frontier. The quantity of data and information that can be extracted from the computerized universe is exponentially growing up as the number of clients and as the number of communication media that are created in order to face always new and dynamic way to collect and process data. Data has changed the way we manage, analyze and leverage data in any industry. The big data revolution as thus affects of the aspects of the human life, and as a consequence also the health care systems. Healthcare analytics have the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general. Healthcare professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and look for best strategies to use these numbers.
However efficiency, reduction of costs and improvements of treatments have several drawbacks.
As technology evolves also the availability of data will progressively increase. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years we are indeed facing the challenge of "big data becoming really big data". This clear affects also healthcare systems, where new application or websites are collecting new health sensitive data. However it turned out the more data as been not necessary associated to better data and more important to new information.
In order to be efficiently analyzed a huge pool of data required the integration of the information provided by different and heterogeneous sources (hospitals, specialist doctors, physicians…). This process is only possible thanks to the creation of standards and the developing a new infrastructure where all data providers collaborate with each other, i.e the creation of common databases. Equally important is the implementing new data analysis tools and strategies. Several solutions have been proposed (see openEHR standards).
The other big obstacle standing in the way to use big data in healthcare is how medical data is spread across many sources governed by different states, hospitals, and administrative departments, i.e. the privacy and protections of sensitive information. As technologies evolve, also the experience and the expectations of privacy evolve. In the past, privacy was viewed as a personal good, rather than a societal one. As such, privacy was regarded as a matter of individual responsibility. Jurisdictions around the world adopted data protection laws that reflected Fair Information Practices (FIPs) — universal privacy principles for the handling of Personal data.
Legislation in different countries is not coherent. In 1996 USA promoted the HIPAA regulation in order to deal with the problem. Title I of HIPAA protects health insurance coverage for workers and their families when they change or lose their jobs, while title II, known as the Administrative Simplification (AS) provisions, requires the establishment of national standards for electronic health care transactions and national identifiers for providers, health insurance plans, and employers.
The effective compliance date of the HIPAA fixed for 2003 had guaranteed the patients and healthcare entities protection about the individuals PHI and their use. The enactment of the HIPAA has caused major changes in the way healthcare centers operate. Data protection compliance tools have been then developed in order to face incorrect and unauthorised used of PHI or undesired web-tracking. Companies and applications as Opentracker®, Asthmapolis®, Ginger.io®, mHealthCoach®, e-Zest®, MetaMap, Ctake are examples of healthcare related services respecting HIPAA regulations. However there are still several open questions. Promising strategy adopted by some providers is the use of security Platform for privacy protection based on the (P3P) project -- which permits to users to have the control on the use of personal information on the web site they visit and by new anti-tracking methods. However the rapid evolution of computer science always demands for the identification and neutralization of new treads.

Appendix Glossary

This appendix is dedicated to the explanation of some technical terms that has been inside the thesis.

Microsoft ACCESS
It is a database management system (DBMS) from Microsoft that combines the relational Microsoft Jet Database Engine with a graphical user interface and software-development tools. It is a member of the Microsoft Office suite of applications, included in the Professional and higher editions or sold separately. Microsoft Access stores data in its own format based on the Access Jet Database Engine. It can also import or link directly to data stored in other applications and databases.
Software developers, data architects and power users can use Microsoft Access to develop application software. Like other Microsoft Office applications, Access is supported by Visual Basic for Applications (VBA), an object-based programming language that can reference a variety of objects including DAO (Data Access Objects), ActiveX Data Objects, and many other ActiveX components. Visual objects used in forms and reports expose their methods and properties in the VBA programming environment, and VBA code modules may declare and call Windows operating system operations. (Definition is taken from Wikipedia)

Microsoft .NET or NET Framework (pronounced dot net)
It is a software framework developed by Microsoft that runs primarily on Microsoft Windows. It includes a large class library named Framework Class Library (FCL) and provides language interoperability (each language can use code written in other languages) across several programming languages. Programs written for .NET Framework execute in a software environment (in contrast to a hardware environment) named Common Language Runtime (CLR), an application virtual machine that provides services such as security, memory management, and exception handling. (As such, computer code written using .NET Framework is called "managed code".) FCL and CLR together constitute .NET Framework.
FCL provides user interface, data access, database connectivity, cryptography, web application development, numeric algorithms, and network communications. Programmers produce software by combining their source code with .NET Framework and other libraries. The framework is intended to be used by most new applications created for the Windows platform. Microsoft also produces an integrated development environment largely for .NET software called Visual Studio.
.NET Framework began as proprietary software, although the firm worked to standardize the software stack almost immediately, even before its first release. Despite the standardization efforts, developers, mainly those in the free and open-source software communities, expressed their unease with the selected terms and the prospects of any free and open-source implementation, especially regarding software patents. Since then, Microsoft has changed .NET development to more closely follow a contemporary model of a community-developed software project, including issuing an update to its patent promising to address the concerns.
.NET Framework led to a family of .NET platforms targeting mobile computing, embedded devices, alternative operating systems, and web browser plug-ins. A reduced version of the framework, .NET Compact Framework, is available on Windows CE platforms, including Windows Mobile devices such as smartphones. .NET Micro Framework is targeted at very resource-constrained embedded devices. Silverlight was available as a web browser plugin. Mono is available for many operating systems and is customized into popular smartphone operating systems

(Android and iOS) and game engines. NET Core targets the Universal Windows Platform (UWP), and cross-platform and cloud computing workloads. (Definition is taken from Wikipedia)

JAVA

It is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages in use, particularly for client-server web applications, with a reported 9 million developers. Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them.

The original and reference implementation Java compilers, virtual machines, and class libraries were originally released by Sun under proprietary licenses. As of May 2007, in compliance with the specifications of the Java Community Process, Sun relicensed most of its Java technologies under the GNU General Public License. Others have also developed alternative implementations of these Sun technologies, such as the GNU Compiler for Java (bytecode compiler), GNU Class path (standard libraries), and IcedTea-Web (browser plugin for applets).

The latest version is Java 8 which is the only version currently supported for free by Oracle, although earlier versions are supported both by Oracle and other companies on a commercial basis. (Definition is taken from Wikipedia)

API

In computer programming, an application programming interface (API) is a set of subroutine definitions, protocols, and tools for building application software. In general terms, it is a set of clearly defined methods of communication between various software components. A good API makes it easier to develop a computer program by providing all the building blocks, which are then put together by the programmer. An API may be for a web-based system, operating system, database system, and computer hardware or software library. An API specification can take many forms, but often includes specifications for routines, data structures, object classes, variables or remote calls. POSIX, Microsoft Windows API, the C++ Standard Template Library and Java APIs are examples of different forms of APIs. Documentation for the API is usually provided to facilitate usage. (Definition is taken from Wikipedia)

GUI

It is the graphical user interface, is a type of user interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, instead of text-based user interfaces, typed command labels or text navigation. GUIs were introduced in reaction to the perceived steep learning curve of command-line interfaces (CLIs), which require commands to be typed on a computer keyboard.

The actions in a GUI are usually performed through direct manipulation of the graphical elements. Beyond computers, GUIs are used in many handheld mobile devices such as MP3 players, portable media players, gaming devices, smartphones and smaller household, office and industrial controls. The term GUI tends not to be applied to other lower-display resolution types of interfaces, such as video games, or not including flat screens, like volumetric displays because the term is restricted to

the scope of two-dimensional display screens able to describe generic information, in the tradition of the computer science research at the Xerox Palo Alto Research Centre (PARC)
(Definition is taken from Wikipedia)

IP address
(Abbreviation of Internet Protocol address) is an identifier assigned to each computer and other device (e.g., printer, router, mobile device, etc.) connected to a TCP/IP network that is used to locate and identify the node in communications with other nodes on the network. IP addresses are usually written and displayed in human-readable notations, such as 172.16.254.1 in IPv4, and 2001:db8:0:1234:0:567:8:1 in IPv6.
Version 4 of the Internet Protocol (IPv4) defines an IP address as a 32-bit number.[1] However, because of the growth of the Internet and the depletion of available IPv4 addresses, a new version of IP (IPv6), using 128 bits for the IP address, was developed in 1995, and standardized as RFC 2460 in 1998. The IP address space is managed globally by the Internet Assigned Numbers Authority (IANA), and by five regional Internet registries (RIR) responsible in their designated territories for assignment to end users and local Internet registries, such as Internet service providers. Addresses have been distributed by IANA to the RIRs in blocks of approximately 16.8 million addresses each. Each ISP or private network administrator assigns an IP address to each device connected to its network. Such assignments may be on a static (fixed or permanent) or dynamic basis, depending on its software and practices.
(Definition is taken from Wikipedia)

Media Access Control (MAC)
In the IEEE 802 reference model of computer networking, the medium access control or media access control (MAC) layer is the lower sublayer of the data link layer (layer 2) of the seven-layer OSI model. The MAC sublayer provides addressing and channel access control mechanisms that make it possible for several terminals or network nodes to communicate within a multiple access network that incorporates a shared medium, e.g. an Ethernet network. The hardware that implements the MAC is referred to as a media access controller.
The MAC sublayer acts as an interface between the logical link control (LLC) sublayer and the network's physical layer. The MAC layer emulates a full-duplex logical communication channel in a multi-point network. This channel may provide unicast, multicast or broadcast communication service. (Definition is taken from Wikipedia)

Microsoft Azure
It is a cloud computing service created by Microsoft for building, deploying, and managing applications and services through a global network of Microsoft-managed data centers. It provides software as a service, platform as a service and infrastructure as a service and supports many different programming languages, tools and frameworks, including both Microsoft-specific and third-party software and systems.
Azure was announced in October 2008 and released on February 1, 2010 as Windows Azure, before being renamed to Microsoft Azure on March 25, 2014.
(Definition is taken from Wikipedia)

ATL
(ATL Transformation Language) is a model transformation language and toolkit developed and maintained by OBEO and AtlanMod. It was initiated by the AtlanMod team (previously called ATLAS Group). In the field of Model-Driven Engineering (MDE), ATL provides ways to produce

a set of target models from a set of source models. Released under the terms of the Eclipse Public License, ATL is an M2M (Eclipse) component, inside of the Eclipse Modeling Project (EMP). (Definition is taken from Wikipedia)

An electronic health record (EHR)

An electronic health record (EHR) or electronic medical record (EMR), refers to the systematized collection of patient and population electronically-stored health information in a digital format. These records can be shared across different health care settings. Records are shared through network-connected, enterprise-wide information systems or other information networks and exchanges. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information.

EHR systems are designed to store data accurately and to capture the state of a patient across time. It eliminates the need to track down a patient's previous paper medical records and assists in ensuring data is accurate and legible. It can reduce risk of data replication as there is only one modifiable file, which means the file is more likely up to date, and decreases risk of lost paperwork. Due to the digital information being searchable and in a single file, EMR's are more effective when extracting medical data for the examination of possible trends and long term changes in a patient. Population-based studies of medical records may also be facilitated by the widespread adoption of EHR's and EMR's. (Definition is taken from Wikipedia)

A Uniform Resource Locator (URL)

A Uniform Resource Locator (URL), colloquially termed a web address, is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. A URL is a specific type of Uniform Resource Identifier (URI), although many people use the two terms interchangeably. A URL implies the means to access an indicated resource and is denoted by a protocol or an access mechanism, which is not true of every URI. Thus http://www.example.com is a URL, while www.example.com is not. URLs occur most commonly to reference web pages (http), but are also used for file transfer (ftp), email (mailto), database access (JDBC), and many other applications.

Most web browsers display the URL of a web page above the page in an address bar. A typical URL could have the form http://www.example.com/index.html, which indicates a protocol (http), a hostname (www.example.com), and a file name (index.html).

(Definition is taken from Wikipedia)

Transport Layer Security (TLS)

Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), both frequently referred to as "SSL", are cryptographic protocols that provide communications security over a computer network. Several versions of the protocols find widespread use in applications such as web browsing, email, Internet faxing, instant messaging, and voice-over-IP (VoIP). Websites use TLS to secure all communications between their servers and web browsers.

(Definition is taken from Wikipedia)

Pretty Good Privacy (PGP)

Pretty Good Privacy (PGP) encryption program provides cryptographic privacy and authentication for data communication. PGP is used for signing, encrypting, and decrypting texts, e-mails, files, directories, and whole disk partitions and to increase the security of e-mail communications. Phil Zimmermann developed PGP in 1991. PGP and similar software follow the OpenPGP standard (RFC 4880) for encrypting and decrypting data. (Definition is taken from Wikipedia)

Model-driven engineering (MDE)
It is a software development methodology that focuses on creating and exploiting domain models, which are conceptual models of all the topics related to a specific problem. Hence, it highlights and aims at abstract representations of the knowledge and activities that govern a particular application domain, rather than the computing (f.e. algorithmic) concepts. (Definition is taken from Wikipedia)

Lenddo
Lenddo is a Singapore-based software-as-a-service company which uses non-traditional data comprising social media and smartphone records in order to ascertain customers' financial stability. Its vision is "to improve financial inclusion for at least a billion people" in developing countries around the world.[1][2] In October 2013, Lenddo had over 350,000 members globally.
(Definition is taken from Wikipedia)

Facebook
Facebook may be accessed by a large range of desktops, laptops, tablet computers, and smartphones over the Internet and mobile networks. After registering to use the site, users can create a user profile indicating their name, occupation, schools attended and so on. Users can add other users as "friends", exchange messages, post status updates and digital photos, share digital videos and links, use various software applications ("apps"), and receive notifications when others update their profiles or make posts. Additionally, users may join common-interest user groups organized by workplace, school, hobbies or other topics, and categorize their friends into lists such as "People From Work" or "Close Friends". In groups, editors can pin posts to top. Additionally, users can complain about or block unpleasant people. Because of the large volume of data that users submit to the service, Facebook has come under scrutiny for its privacy policies. Facebook makes most of its revenue from advertisements which appear onscreen.
(Definition is taken from getwebcatalog.com)

Twitter
Twitter is an online news and social networking service where users post and interact with messages, "tweets", restricted to 140 characters. Registered users can post tweets, but those who are unregistered can only read them. Users access Twitter through its website interface, SMS or a mobile device app. Twitter Inc. is based in San Francisco, California, United States, and has more than 25 offices around the world. (Definition is taken from Wikipedia)

LinkedIn
LinkedIn is a business- and employment-oriented social networking service that operates via websites and mobile apps. Founded on December 28, 2002, and launched on May 5, 2003, it is mainly used for professional networking, including employers posting jobs and job seekers posting their CVs. As of 2015, most of the company's revenue came from selling access to information about its members to recruiters and sales professionals. As of September 2016, LinkedIn had more than 467 million accounts, out of which more than 106 million are active. As of April 2017, LinkedIn had 500 million members in 200 countries. LinkedIn allows members (both workers and employers) to create profiles and "connections" to each other in an online social network which may represent real-world professional relationships. Members can invite anyone (whether an existing member or not) to become a connection. The "gated-access approach" (where contact with any professional requires either an existing relationship or an introduction through a contact of theirs) is intended to build trust among the service's members. LinkedIn participated in the EU's International Safe Harbor Privacy Principles. (Definition is taken from Wikipedia)

Flick

Flick (pronounced "flicker") is an image hosting and video hosting website and web services suite that was created by Ludicorp in 2004 and acquired by Yahoo on March 20, 2005. In addition to being a popular website for users to share and embed personal photographs, and effectively an online community, the service is widely used by photo researchers and by bloggers to host images that they embed in blogs and social media. (Definition is taken from Wikipedia)

XML

In computing, Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The W3C's XML 1.0 Specification and several other related specifications—all of them free open standards—define XML. The design goals of XML emphasize simplicity, generality, and usability across the Internet. It is a textual data format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, the language is widely used for the representation of arbitrary data structures such as those used in web services. Several schema systems exist to aid in the definition of XML-based languages, while programmers have developed many application programming interfaces (APIs) to aid the processing of XML data.
(Definition is taken from Wikipedia)

Comma-separated values

In computing, a comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.
The CSV file format is not standardized. The basic idea of separating fields with a comma is clear, but that idea gets complicated when the field data may also contain commas or even embedded line-breaks. CSV implementations may not handle such field data, or they may use quotation marks to surround the field. Quotation does not solve everything: some fields may need embedded quotation marks, so a CSV implementation may include escape characters or escape sequences.
(Definition is taken from Wikipedia)

JSON

In computing, JavaScript Object Notation or JSON, is an open-standard file format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types (or any other serializable value). It is a very common data format used for asynchronous browser/server communication, including as a replacement for XML in some AJAX-style systems.
(Definition is taken from Wikipedia)

NoSQL

A NoSQL (originally referring to "non SQL", "non-relational" or "not only SQL") database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century, triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com. NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.
Motivations for this approach include: simplicity of design, simpler "horizontal" scaling to clusters of machines (which is a problem for relational databases), and finer control over availability. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) are different from those used by default in relational databases, making some operations faster in NoSQL. (Definition is taken from Wikipedia)

Google Earth
Google Earth is a computer program that renders a simulacrum of the Earth based on satellite imagery. It maps the Earth by the superimposition of images obtained from satellite imagery, aerial photography and geographic information system (GIS) onto a 3D globe.
It was originally an eponymous product sold by Keyhole, Inc. After Keyhole's acquisition by Google, Google Earth was released in June 2005, driving public interest in geospatial technologies and applications. (Definition is taken from Wikipedia)

Apache Hadoop
Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. (Definition is taken from Wikipedia)

Apache Cassandra
Apache Cassandra is a free and open-source distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacentres, with asynchronous master less replication allowing low latency operations for all clients. Cassandra also places a high value on performance. In 2012, University of Toronto researchers studying NoSQL systems concluded that "In terms of scalability, there is a clear winner throughout our experiments. Cassandra achieves the highest throughput for the maximum number of nodes in all experiments" although "this comes at the price of high write and read latencies.
(Definition is taken from Wikipedia)

Kreditech
Kreditech (a trading name of Kreditech Holding SSL GmbH) is a German online lender which offers loans to individuals based on their creditworthiness which is analyzed using their online data instead of using traditional credit rating information. Founded in 2012 by Sebastian Diemer and Alexander Graubner-Müller, Kreditech is headquartered in Hamburg, Germany; it particularly focuses its efforts on emerging market. (Definition is taken from Wikipedia)

eBay
eBay is a multinational e-commerce corporation, facilitating online consumer-to-consumer and business-to-consumer sales. It is headquartered in San Jose, California. eBay was founded by Pierre Omidyar in 1995, and became a notable success story of the dot-com bubble. Today it is a multibillion-dollar business with operations in about 30 countries. (www.ebay.com)
(Definition is taken from Wikipedia)

Amazon.com
Amazon.com also called Amazon, is an American electronic commerce and cloud computing company that was founded on July 5, 1994, by Jeff Bezos and is based in Seattle, Washington. It is the largest Internet-based retailer in the world by total sales and market capitalization. Amazon.com started as an online bookstore, later diversifying to sell DVDs, Blu-rays, CDs, video downloads/streaming, MP3 downloads/streaming, audiobook downloads/streaming, software, video games, electronics, apparel, furniture, food, toys, and jewelry. The company also produces consumer electronics—notably, Kindle e-readers, Fire tablets, Fire TV, and Echo—and is the

world's largest provider of cloud infrastructure services (IaaS and PaaS). Amazon also sells certain low-end products like USB cables under its in-house brand Amazon Basics. (www.amazon.com) (Definition is taken from Wikipedia)

Kabbage
Kabbage, Inc. is an online financial technology company based in Atlanta, Georgia. The company provides funding directly to small businesses and consumers through an automated lending platform. (Definition is taken from Wikipedia)

PayPal
PayPal Holdings, Inc. is an American company operating a worldwide online payments system that supports online money transfers and serves as an electronic alternative to traditional paper methods like checks and money orders. PayPal is one of the world's largest Internet payment companies. The company operates as a payment processor for online vendors, auction sites and other commercial users, for which it charges a fee. (Definition is taken from Wikipedia)

Walmart
Wal-Mart Stores, Inc. doing business as Walmart, is an American multinational retailing corporation that operates as a chain of hypermarkets, discount department stores, and grocery stores. Headquartered in Bentonville, Arkansas, the company was founded by Sam Walton in 1962 and incorporated on October 31, 1969. As of January 31, 2017, Walmart has 11,695 stores and clubs in 28 countries, under a total of 63 banners. The company operates under the name Walmart in the United States and Canada. It operates as Walmart de México y Centroamérica in Mexico and Central America, as Asda in the United Kingdom, as the Seiyu Group in Japan, and as Best Price in India. It has wholly owned operations in Argentina, Chile, Brazil, and Canada. It also owns and operates the Sam's Club retail warehouses. (Definition is taken from Wikipedia)

Apache Lucene
Apache LuceneTM is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Apache Lucene is an open source project available for free download. Please use the links on the right to access Lucene. (Definition is taken from Wikipedia)

Valence Health offers a complete suite of solutions that empower provider organizations as they transition to value-based care. Whether implemented in modules or as a whole, Valence has the expertise and technology platform that providers need to succeed in a rapidly evolving marketplace. Our three-in-one combination of focused consulting, innovative technology platform, and operational services sets us apart.
Since 1996, Valence has provided value-based care solutions that help hospitals, health systems and physicians more effectively manage patient populations to help them achieve clinical and financial rewards. Our integrated set of advisory services, population health technology and managed services support more than 85,000 physicians and 135 hospitals as they advance the health of 20 million patients. (Definition is taken from http://valencehealth.com/)

Infrastructure as a service (IaaS)
According to the Internet Engineering Task Force (IETF), the most basic cloud-service model is that of providers offering computing infrastructure – virtual machines and other resources – as a service to subscribers.

Infrastructure as a service (IaaS) refers to online services that abstract the user from the details of infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc. A hypervisor, such as Xen, Oracle VirtualBox, Oracle VM, KVM, VMware ESX/ESXi, or Hyper-V, runs the virtual machines as guests. Pools of hypervisors within the cloud operational system can support large numbers of virtual machines and the ability to scale services up and down according to customers' varying requirements. Linux containers run in isolated partitions of a single Linux kernel running directly on the physical hardware. Linux cgroups and namespaces are the underlying Linux kernel technologies used to isolate, secure and manage the containers. Containerisation offers higher performance than virtualization, because there is no hypervisor overhead. (Definition is taken from Wikipedia)

References

1 Zhang Y et al, “Astronomy in the big data era” Data Science Journal, 14, 11.

2 Anderson A and Semmelroth D, “BIG DATA and SEARCH ENGINES”

3 KATE MILTNER, MARY L. GRAY, “Critiquing Big Data: Politics, Ethics, International Journal of Communication, 8, 1633 (2014)

4 Agrawel R. amd Srikant R, “Privacy Preventing Data Meaning”, ACM SIGMOD 2000, 5/00 Dallas USA

5 Intel IT Centre, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012

6 MS KB article 275561 (2007-01-29). "Description of the new features that are included in Microsoft Jet 4.0". Microsoft. Retrieved 2008-06-19.

7 Seref SAGIROGLU and Duygu SINANC, “Big Data: A Review” 978-1-4673-6404-1/13/$31.00 ©2013 IEEE

8 http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs

9 https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/

10 Gralla, Preston (2007). “How the Internet Works”. Indianapolis: Que Pub. ISBN 0-7897-2132-5.

11 Rhee, M. Y. (2003). “Internet Security: Cryptographic Principles, Algorithms and Protocols”. Chichester: Wiley. ISBN 0-470-85285-2.

12 An example of a completely re-engineered computer is the Librem laptop (https://www.crowdsupply.com/purism/librem-13) which uses components certified by web-security experts. It was launched after a crowd funding campaign in 2015.

13 Stewart, James (2012). “CISSP Study Guide”. Canada: John Wiley & Sons, Inc. pp. 255–257. ISBN 978-1-118-31417-3 – via Online PSU course resource, EBL Reader.

14 https://en.wikibooks.org/wiki/Intellectual_Property_and_the_Internet/Internet_security

15 Ramzan, Zulfikar (2010). "Phishing attacks and countermeasures". In Stamp, Mark & Stavroulakis, Peter. Handbook of Information and Communication Security. Springer. ISBN 9783642041174.

16 Van der Merwe, A J, Loock, M, Dabrowski, M. (2005), “Characteristics and Responsibilities involved in a Phishing attack, Winter International Symposium on Information and Communication Technologies”, Cape Town, January 2005.

17 "2012 Global Losses From Phishing Estimated At $1.5 Bn". FirstPost. February 20, 2013. Retrieved December 21, 2014.

18 https://en.wikipedia.org/wiki/Information_security.

19 Gordon, Lawrence; Loeb, Martin (November 2002). "The Economics of Information Security Investment". ACM Transactions on Information and System Security. 5 (4): 438–457. doi:10.1145/581271.581274.

20 Cumbley R and Church P (2014) Is ‘‘Big Data’’ creepy? Computer Law and Security Review October 2013 29: 601–609. https://doi.org/10.1016/j.clsr.2013.07.007

21 Knight A and Saxby S (2014) "Identity crisis: Global challenges of identity protection in a networked world". Computer Law and Security Review 30: 617–632.

22 Kambatla K, Kollias G, Kumar V, et al. (2014) “Trends in Big Data analytics”. Journal of Parallel and Distributed Computing 74: 2561–2573. Available online 2 February 2014

23 “Web Tracking: Mechanisms, Implications, and Defenses”
Tomasz Bujlow, Member, IEEE, Valentín Carela-Español, Josep Solé-Pareta, and Pere Barlet-Ros.

24 E. Pariser, “The Filter Bubble: What the Internet is hiding from you.” London, U.K.: Penguin, 2011.

25 C. E. Wills and C. Tatar, “Understanding what they do with what they know,” in Proc. ACM Workshop Privacy Electron. Soc., Oct. 2012, pp. 13–18.

26 AdStack |Email Optimization Platform, 2014.[Online]. Available at : http://adstack.com

27 “Facebook friends could change your credit score,” CNN, 2013. [Online]. Available: http://money.cnn.com/2013/08/26/technology/social/facebook-credit-score/index.html?hpt=hp_t2

28 J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, “Detecting price and search discrimination on the Internet,” in Proc. 11t ACM Workshop Hot Topics Netw., Seattle, WA, USA, Oct. 2012, pp. 79–84.

29 https://en.wikipedia.org/wiki/Bancassurance

30 www.tcplifesystems.com

31 “AP IMPACT: When your criminal past isn’t yours,” Yahoo Finance, 2011. [Online]. Available: https://www.yahoo.com/news/ap-impact-criminal-past-isnt-182335059.html

32 “Finnish employers cannot google applicants,” 2006. [Online]. Available: http://blogs.law.harvard.edu/infolaw/2006/11/15/finnish-employers-cannot-google-applicants

33 Face Recognition Study—FAQ—Heinz College—Carnegie Mellon University, 2011. [Online]. Available: http://www.heinz.cmu.edu/ãcquisti/face-recognition-study-FAQ

34 “Big data privacy: a technological perspective and review” , Priyank Jain, Manasi Gyanchandani, Nilay Khare, First Online: 26 November 2016 DOI: 10.1186/s40537-016-0059-y

35 “Big data, artificial intelligence, machine learning and data protection” ,20170301Version: 2.0. Information computer office, Elizabeth Denham.

36 Leroy Hood and Charles Auffray “Participatory medicine: a driving force for revolutionizing healthcare”, Genome Medicine20135:110 DOI: 10.1186/gm514

37 https://pharmaphorum.com/views-and-analysis/big-data-enables-the-promise-of-personalised-medicine/

38 Mervis, J. 2012. “Agencies Rally to Tackle Big Data”, Science, 336(4):22, June 6, 2012

39 “The Human Element of Big Data: Issues, Analytics, and Performance”. Geetam S. Tomar, Narendra S. Chaudhari, Robin Singh Bhadoria, Ganesh Chandra Deka September 30, 2016 by ChapmanandHall/CRCReference-351Pages-123B/WIllustrations
ISBN 9781498754156 - CAT# K27388

40 Terlink, Marc et al. The new hero of big data and analytics: The Chief Data Officer. IBM, June 2014. http://www-935.ibm.com/services/us/gbs/thoughtleadership/chiefdataofficer/ Accessed 15 June 2016

41 “Big data, artificial intelligence, machine learning and data protection” Data Protection Act and General Data Protection Regulation

42 http://datasift.com

43 Wood, Spencer A et al. “Using social media to quantify nature-based tourism and recreation.” Nature Scientific Reports, 17 October 2013 http://www.nature.com/articles/srep02976

44 “Smart Steps increase Morrisons new and return customers by 150%”.October 2013 http://dynamicinsights.telefonica.com/1158/a-smart-step-ahead-for-morrisons

45 www.databaseanswers.org/data_models/

46 https://mapr.com/blog/5-big-data-production-examples-healthcare/

47 Evidence-Based Medicine Working Group (November 1992). "Evidence-based medicine. A new approach to teaching the practice of medicine". JAMA. 268 (17): 2420–5. doi:10.1001/jama.268.17.2420. PMID 1404801

48 Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW Collaborative Biomedicine in the Age of Big Data: The Case of Cancer J Med Internet Res 2014;16(4):e101

49 http://www.ncin.org.uk/publications/routes_to_diagnosis

50 THE LANCET • Vol 363 • January 10, 2004 • www.thelancet.com. “Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies”

51 www. MapR.com

52 Mats Uddenfeldt “How Big Data is Improving Outcomes and Reducing Costs in Healthcare”, https://www.linkedin.com/pulse/how-big-data-reducing-costs-improving-outcomes-mats-uddenfeldt

53 http://genome.cshlp.org/content/20/9/1297.short

54 https://www.novartis.com/

55 https://mapr.com/blog/5-big-data-production-examples-healthcare/

56 http://www.ecografostelemed.es/

57 2013-2014 United Healthcare Annual Delegate Compliance Notice. Fraud, Waste and Abuse and General Compliance Requirements, https://www.unitedhealthcareonline.com/ccmcontent/ProviderII/UHC/en-US/Assets/ProviderStaticFiles/ProviderStaticFilesPdf/Tools%20and%20Resources/Training%20&%20Education/Fraud_Waste_Abuse_Training_FAQ.pdf

58 A big data revolution and healtcare accelerating value and innovation jnuary 2013 by Peter Groves,Basel Kayyali, David Knott, Steve Van Kuiken

59 http://www.hipaajournal.com/healthcare-big-data-privacy-and-security-workgroup-gives-preliminary-report-8029/

60 https://www.hrsa.gov/about/news/pressreleases/2010/100201a.html

61 Young B. Choi · Kathleen E. Capitan · Joshua S. Krause Meredith M. Streeper "Challenges Associated with Privacy in Health Care Industry: Implementation of HIPAA and the Security Rules"J Med Sys (2006) 30(1): 57–64 DOI 10.1007/s10916-006-7405-0

62 http://www.hipaajournal.com/healthcare-big-data-privacy-and-security-workgroup-gives-preliminary-report-8029/

63 T.-C. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos, "TrackAdvisor: Taking back browsing privacy from third-party trackers," in Proc. 16th Passive Active Meas. Conf., Mar. 2015, pp. 1–12.

64 Schneider J, Mercuri RT (2004). "The HIPAA-potamus in health care data security". Commun ACM 47(7)

65 Volonino L, Robinson SR (2004) "Principles and practice of information security: Protecting computers from Hackers and Lawyers", Prentice Hall, Inc., Upper Saddle River, NJ

66 Federal Register: "Rules and Regulations". 65(160), August 2000

67 Panko R (2004) "Corporate Computer and Network Security",Prentice Hall Inc., NJ

68 "Workgroup for Electronic Data Interchange" (WEDI) (2004) HIPAA Security White Papers

69 "Workgroup for Electronic Data Interchange" (WEDI) (2004) Security and Privacy Workgroup Introduction

70 T.-C. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos, "TrackAdvisor: Taking back browsing privacy from third-party trackers," in Proc. 16th Passive Active Meas. Conf., Mar. 2015, pp. 1–12.

71 Washington District of Colombia Department of Health. Retreived September 17, 2005 at http://dchealth.dc.gov/hipaa/hipaaps.shtm.

72 Internet Explorer 9 Tracking Protection Lists—Microsoft, 2014. [Online]. Available: http://ie.microsoft.com/testdrive/Browser/TrackingProtectionLists

73 Castells, M. (1999). The Information Age, Volumes 1-3: Economy, Society and Culture. Cambridge (Mass.); Oxford: Wiley-Blackwell

74 "The social contract core". Honolulu, Hawaii, USA — May 07 - 11, 2002 ISBN:1-58113-449-5 doi>10.1145/511446.511474

75http://www.cio.com/article/2390305/outsourcing/it-robots-may-mean-the-end-of-offshore-outsourcing.html

76 Hagland M. Customized automation: OB/GNY practices are finding EMR systems designed specifically for them. Healthcare Informatics Online, March 2004.

77 Hooda JS, Dogdu E, Sunderraman RJ (2004) Health level-7 compliant clinical patient records system. Commun ACM 47(7)

78 Goldberg IV (2000) Electronic medical records and patient privacy. Health Care Manager 18:3

79 Eddy AC (2000) Annals of health law, Annual 2000 v9 p1-72, A Critical Analysis of Health and Human Services' Proposed Health Privacy Regulations in Light of the Health Insurance Privacy and Accountability Act of 1996

80 http://www.record-link.com/

81 Essex D. The many layers of workflow automation. Health-care Informatics Online, June 2000. Available at http://www.healthcare-informatics.com/issues/2000/06 00/essex.htm

82 "Challenges Associated with Privacy in Health Care Industry: Implementation of HIPAA and the Security Rules". Choi, Y.B., Capitan, K.E., Krause, J.S. et al. J Med Syst (2006) 30: 57. Doi:10.1007/s10916-006-7405-0.

83 Eddy, A. Critical Analysis of Health and Human Services' Proposed Health Privacy Regulations in Light of the Health Insurance Privacy and Accountability Act of 1996. Annals of health law, 9, 1-72. (2000)

84 Flores Zuniga, A., Win, K. & Susilo, W. (2013). Secure exchange of electronic health records. User-Driven Healthcare: Concepts, Methodologies, Tools, and Applications (pp. 1403-1424). United States: Medical Information Science Reference

85 https://www.healthit.gov/buzz-blog/privacy-and-security-of-ehrs/privacy-security-electronic-health-records/

86 "Guide to Privacy and Security of Electronic Health Information" ,version 2,healthit.gov

87 Panko R (2005) "Business Data Networks and Telecommunications", 5th edn., Prentice Hall, Inc., NJ

88 Onam PW, Hanebutte N (2005) "Fundamentals—and Beyond-of Computer & Network Security." Powerpoint Slides from Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Waikola, Hawaii

89 T.-C. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos, “TrackAdvisor: Taking back browsing privacy from third-party trackers,” in Proc. 16th Passive Active Meas. Conf., Mar. 2015, pp. 1–12.

90 SecurityKISS\u2014Free VPN Service, 2014. [Online]. Available: http://www.securitykiss.com/index.php

91 “CyberGhost VPN: Surf Anonymously”, 2014. [Online].
Available: https://www.cyberghostvpn.com/en_us

92 Privoxy—Home Page, 2013. [Online]. Available: http://www.privoxy.org

93 J. R. Mayer and J. C. Mitchell, “Third-party web tracking: Policy and technology,” in Proc. IEEE Symp. Security Privacy, May 2012, pp. 413–427.

94 Web Tracking and User Privacy Workshop, 29 April 2011, 2011. [Online]. Available: http://w3.org/2011/04/29-w3cdnt-minutes.html

95 Tracking Preference Expression (DNT)—World Wide Web Consortium (W3C), 2014. [Online]. Available: http://www.w3.org/2011/tracking-protection/drafts/tracking-dnt.html

96 F. Roesner, T. Kohno, and D. Wetherall, “Detecting and defending against third-party tracking on the web,” in Proc. 9th USENIX Conf. Networked Syst. Design Implement., 2012, p. 12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2228298.2228315

97 “DuckDuckGo”, 2014. [Online]. Available: https://duckduckgo.com

98 “Startpage Search Engine”, 2014. [Online]. Available: https://startpage.com/eng

99 “Ixquick Search Engine”, 2014. [Online].Available: https://www.ixquick.com

100 G. Aggarwal, E. Bursztein, C. Jackson, and D. Boneh, “An analysis of private browsing modes in modern browsers,” in Proc. 19th USENIX Secur. Symp., 2010, p. 6.

101 Valve/Fingerprintjs · GitHub, 2014. [Online]. Available: https://github.com/Valve/fingerprintjs

102 NoScript Security Suite Add-Ons for Firefox— Mozilla Add-Ons, 2014. [Online]. Available: https://addons.mozilla.org/en-US/firefox/ addon/noscript

103 “Flashblock, Add-Ons for Firefox—Mozilla Add-Ons”, 2014. [Online].
Available: https://addons.mozilla.org/en-US/firefox/addon/flashblock

104 Inbox Alias—Protect Yourself, 2014 [Online].Available: http://inboxalias.com

105 33Mail.Com—Simple Free Disposable Email Address Service, 2014. [Online]. Available: http://www.33mail.com

106 Jetable.Org—Home, 2014. [Online]. Available: http://www.jetable.org

107 “Create a New Mailexpire Alias—Avoid Spam,” 2014. [Online]. Available: http://www.mailexpire.com

108 TrashMail—Disposable Email Addresses, 2014. [Online]. Available: https://trashmail.com

109 Guerrilla Mail—Disposable Temporary E-Mail Address, 2014. [Online].
Available: http://www.guerrillamail.com

110 “Humanizing Big Data: Marketing at the Meeting of Data, Social Science, and Consumer Insight” by Colin Strong, March 28, 2015. https://smallbiztrends.com/2015/03/humanizing-big-data-book-review.html

111 “Outsourcing data management” – IaaS, http://www.opentracker.net/article/understanding-big-data

112 M. Armbrust et al., “A view of cloud computing” Communication of the ACM, 53, 50, 2010.

113 “Big Data Technologies in Healthcare Needs, opportunities and challenges”. BDV Big Data Value Association.TF7 Healthcare subgroup. 12/21/2016

114 http://PhilipsHealthSuite.com and WullianallurRaghupathi and VijuRaghupathiQuelch, John A., and Margaret L. Rodriguez. "Philips Healthcare: Marketing the HealthSuite Digital Platform." Harvard Business School Case 515-052, May 2015. (Revised September 2015.)
115 http://europa.eu/rapid/press-release_IP-16-1408_en.htm

116 “Big Data Technology in Healthcare, Needs, opportunities and challenges”,TF7 Healthcare subgroup

117 http://www.e-zest.com/big-data-solutions-for-healthcare

118 www. opentracker.com

119 www.asthmapolis.com

120 https://data.cdc.gov/

121 www.Ginger.io.com

122 www.mHealthCoach.com

123 https://www.ahrq.gov/

124 http://www.ndmrb.ox.ac.uk/the-li-ka-shing-centre

125 https://commonfund.nih.gov/bd2k

126 https://aws.amazon.com/es/

127 https://www.hds.com/es-latam/home.html

128 http://www.usa.philips.com/healthcare/innovation/about-health-suite

129 http://www.e-zest.com/

130 http://www.sagitec.com/healhub

131 https://www.fluidops.com/en/products/information_workbench/

132 https://mapr.com/mapr-guide-big-data-healthcare/

133 https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

134 http://ctakes.apache.org/

135 https://uima.apache.org/

136 Specifications Editorial Committee. “openEHR EHR Extract IM”.openEHR Foundation. openEHR Foundation Retrieved 3 November 2015.

137 Beale, Thomas (2002). “Archetypes: Constraint-based Domain Models for Future-proof Information Systems” (PDF). Proceedings of the 11th OOPSLA Workshop on Behavioural Semantics. (PDF)

138 S. Heard & T. Beale. (eds.) (2007). “openEHR Architecture Overview” (PDF). openEHR Foundation. Retrieved 9 April 2013. (PDF)
139 openEHR Specification Program. “openEHR Specifications”. openEHR Foundation. Retrieved 3 November 2015.

140 www.Sagitec.com

141 W3C. «Guía Breve de Privacidad y P3P». http://www.w3c.es/Divulgacion/GuiasBreves/PrivacidadP3P

142 “Information Security and Privacy in Healthcare: Current State of Research1”. Ajit Appari (Ajit.Appari@Tuck.Dartmouth.edu) And M. Eric Johnson (M.Eric.Johnson@Tuck.Dartmouth.edu) Center for Digital Strategies Tuck School of Business Dartmouth College, Hanover NH

143 https://addons.mozilla.org/es/firefox/addon/javascript-deobfuscator/

144 http://www.cs.wustl.edu/~jain/cse567-06/ftp/net_monitoring/index.html

145 https://http2.github.io/