Data Trading Similarity Signature
An Extended Data Trading Framework for Human and Non-Human Actors

Sebastian Lawrenz
Clausthal University of Technology
sebastian.lawrenz@tu-clausthal.de

Hendrik Poschmann
Ostfalia University of Applied Science
he.poschmann@ostfalia.de

Vera Stein
Clausthal University of Technology
vera.stein@tu-clausthal.de

Andreas Rausch
Clausthal University of Technology
andreas.rausch@tu-clausthal.de

Abstract

Fair and secure data trading is one of the most prominent challenges of the 21st century. This paper presents a second iteration of an approach to develop a data marketplace concept by checking consumer requirements. The main problem we identified is data quality and the question: Would a dataset fulfill the consumer requirements? Starting from an approach that uses a binary response set to answer the question of whether requirements are met, we concluded that a description of consumer requirements needs to be quantitatively comparable. The novel approach presented here identifies similarities between datasets and consumer requirements. It forms a unique, fingerprint-like similarity signature for each dataset, which can be interpreted by both human and non-human actors. The approach is deducted and designed by using the Design Science Research Methodology and discussed critically in the end.

1. Motivation and introduction

You can have data without information, but you cannot have information without data – Daniel Keys Moran

With the shift to a data-driven society, data is becoming a commodity. Data is a commodity, which can be traded as well as every physical commodity. But since we cannot touch or see data before buying it, the process in data trading is different from other trading processes. Purchase decisions are not based on product reviews or trying out a product yourself. Accordingly, new purchase decision processes must be researched and established.

1.1. Data and information

The definition of data and information is different depending on the scientific field and literature. A couple of scientific publications describe data as the basis of information and, therefore, the basis for knowledge and wisdom, as shown in the data information knowledge and wisdom pyramid [1]. In the rest of this paper, data refers to both data and information as a commodity subject to be traded.

Data is a kind of digital goods distinguished from other goods by characteristics like non-rivalry, infinite expansibility, and combinability [2]. Furthermore, the value of data is quantified by the news content (also called the level of surprise) [3], which leads to one of the key challenges in data trading – the product (the dataset itself) cannot be shown before it is sold [4].

1.2. Problem statement

The previously mentioned challenge is already tackled by a work of [4], where the authors present a framework for data trading by designing a data marketplace. This work mainly focuses on the question of how a data consumer can check his specific requirements by checking the data quality:

Data quality describes the value of a dataset from the data consumer's point of view with regard to the requirements. High data quality means a good match of the requirements and the characteristics of the dataset [4].

The approach presented a framework that translates informal requirements into a formal logic (based on the predicate logic) and verifies these requirements in a secure runtime. One main limitation of this approach was the binary set of answers *{yes, no}*. This way, it could only be checked if all requirements were met or not. However, the mentioned approach does not take the partial fulfillment of requirements into account.

Furthermore, non-human stakeholders, such as autonomous systems, were not considered as possible data providers or consumers. Nevertheless, non-human actors (autonomous systems), such as machines or robots, are essential stakeholders for data ecosystems, such as a circular economy ecosystem [5]. While

URI: https://hdl.handle.net/10125/79937
978-0-9981331-5-7

technically providing data transfer for a non-human actor is already easily viable via REST interfaces, the question of how a non-human actor can describe his requirements for a data set has yet to be answered.

1.3. Research methodology

The research of this paper has been conducted by using the *Design Science Research Methodology*. The *Design Science* paradigm concentrates on the development of an artifact in a specific context [6]. The preceding work was already developed by using *Design Science Research* [4]. Respectively, this paper uses the already existing data marketplace as a starting point. Furthermore, we use the *three-cycle view* from *Design Science* as proposed by Hevner [7]. Hevner introduces three closely related cycles of activities. The *Relevance Cycle* analyzes the environment and the domain, the *Rigor Cycle* investigates specific domain knowledge, and the *Design Cycle* finally supports the design of a new or extended artifact [7]. Section 1.4. presents a Scenario aligned to the *Design Science Research*, which shows our framework.

1.4. Scenario

Baseline Situation: A data marketplace is already existing. This data marketplace provides services for a data provider to create new offers. The marketplace itself does not contain any datasets, just the metadata about the datasets and a description of the offers. A data consumer can find and buy the current offers via the data marketplace and verify its requirements in a limited/binary way (further details see Section 2.2.).

Challenge: Prospective data consumers are not limited to human beings but also include autonomous systems. These autonomous systems have limited autonomy over their purchase decisions (e.g., budgets). In our case, the autonomous system is a disassembly robot that requires data about traction batteries. In the marketplace, several potential datasets are offered. These datasets vary in price, metadata, description, and content itself. Some datasets would provide a high value for the robot, and others a limited or nearly zero. The robot now has to decide which datasets would fulfill his requirements and have the best price-service ratio. However, the robot has knowledge about previous datasets, where the process performed well.

1.4. Contribution and outline

In this paper, we propose a data trading framework for human and non-human actors. This paper makes three contributions to the research area of data marketplaces and data trading:

1. The analysis of the application domain of data marketplaces in a second iteration and the deduction of requirements in a structured way using Design Science.
2. An overview of the current trends and related work in data marketplaces and scientific theories to overcome the identified main problem of checking the data quality.
3. A novel machine learning approach to check data quality against the data consumers' requirements. The process is introduced in a structured way and based on autoencoding and k-means. This approach allows human and non-human data consumers to compare datasets to find the best match for their requirements.

The remainder of this paper is structured as follows: Section 2 introduces the relevant background of the existing data marketplace and the current limitations. Furthermore, we discuss the role of non-human actors in data ecosystems and collect their requirements. Section 3 presents the current state of research and provides the specific domain knowledge. Section 4 presents our overall approach: An extended data marketplace framework. Section 5 discusses the approach critically, and Section 6 concludes the paper and outlines our next steps.

2. Background

In this section, we introduce the necessary background for our paper. This section is oriented on the *Design Science Research Methodology* and represents the *Relevance Cycle* and the *Rigor Cycle*. In the first part, we introduce the current state of data marketplaces, show the relevance of our work (*Relevance Cycle*), and later the necessary research background and related work (*Rigor Cycle*).

2.1. Data marketplace requirements

A data marketplace describes a legal framework for data trading. In general, these terms tackle all topics related to making data profitable and around the new emerging business models for data exchange, such as data collection, aggregation, processing, enrichment, and buying and selling processes [8]. Instead of tackling the whole frame of this research area and emerging market, we only focus on trading *datasets*. A dataset, as a commodity, is a completed collection of different data, including metadata, for example, an SQL table with instance values and column names. Trading with data streams, such as IoT data streams, is out of scope in this research.

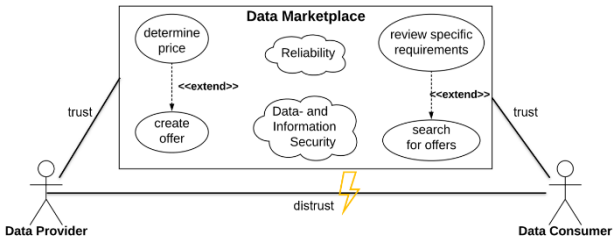

Figure 1: Functional and non functional requirements for a data marketplace

In the preceding work of [4], the main requirements and use cases for a data marketplace were conducted, namely:

Data Privacy: The value of a dataset itself is characterized by the news content or level of surprise [3], [9]. Since this is the fact, showing the content of a dataset will reduce the value of the same dataset. Accordingly, a data provider wants to ensure the privacy of the dataset.

Requirements Privacy: A similar case applies to the data consumer. A data consumer does not want to share the specific requirements for a dataset since a data provider could derive the business model (or idea) or the specific value of a dataset.

Both requirements lead to the main challenge in a data marketplace, based on the lack of trust between data providers and data consumers: *If no one is willing to share information about the datasets, how can they be traded*?

A first step to overcome this lack of trust is by providing a secure and reliable platform, such as a data marketplace. Moreover, the authors deducted some prominent use cases for data marketplaces, as shown in figure 1:

The primary use case for the *data provider* is to create an offer with some optional assistance like a price suggestion. For a *data consumer*, the primary use is to search for offers, including a decision support assistance based on a review of his specific requirements (against his data quality criteria). The two optional use cases are discussed further in the following subsection 2.2. Since this paper is focused on the assistance processes for the data provider and the data consumer, trivial use cases like the closing buy/sell are excluded from the scope of this research.

2.2. (Previous) data marketplace architecture and limitations

Figure 2 shows a high-level approach for the previous data marketplace architecture, as described in detail in the preceding work [4]. This approach is designed to tackle the requirements and use cases introduced in Section 2.1. The *Offer Creation* package supports the data provider by creating an offer, and the *Meta Data Client (A)* analyzes the dataset to be sold automatically. The datasets are not stored locally and not in the marketplace itself to avoid security issues. Every offer in the *Offer List (C)* contains only the meta data and the description of the dataset. A data consumer can find offers via the *Search (D)* component and later check his specific requirements using the *Requirements Adapter* (E). The *Secure Runtime (G)*, a container-based sandbox, finally evaluates the specific requirements for the dataset in a closed environment and provides the

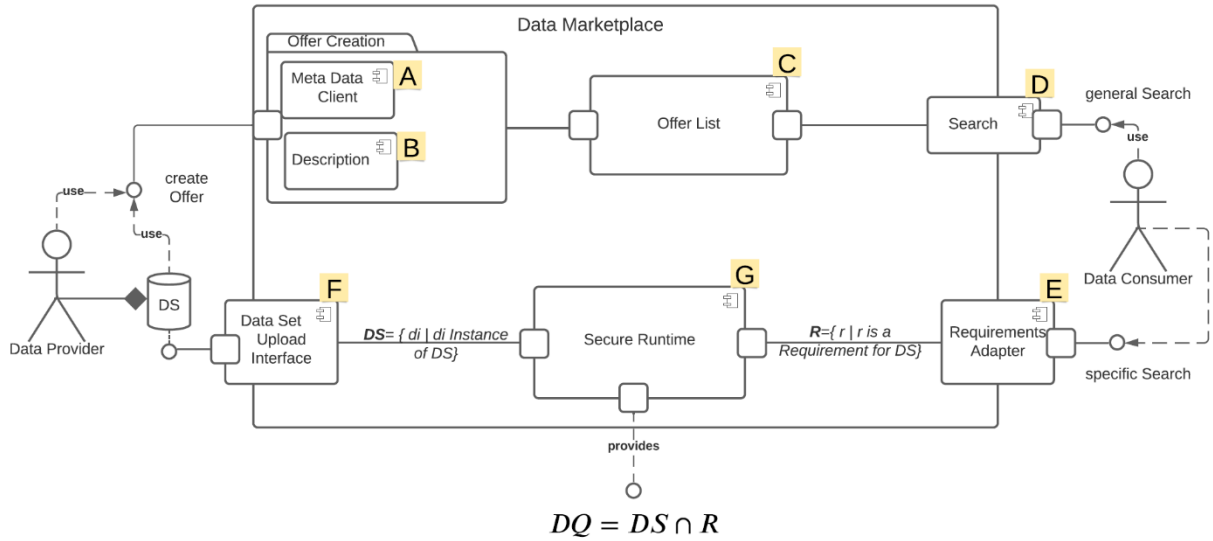

Figure 2: High level data marketplace architecture (based on [4])

data consumer a binary answer, if its requirements are fulfilled {yes} or not fulfilled {no}.

Limitations: The current approach was limited to two values {0, 1}: The requirements are fulfilled or not fulfilled. Cases like partially fulfilled requirements (e.g., for 90%) are not possible with this approach. Furthermore, the approach does not support the purchase decision in case of multiple datasets fulfilling the requirements (or even not). Another limitation we identified during the evaluation with some testers in an unstructured case study was that the current approach is not designed to compare different datasets easily. Since this requirement came from several participants during the evaluation study, we consider it in the scope of this paper as a goal.

Furthermore, most testers were not able to describe their requirements *R* with our approach. They considered the process of using Prolog to describe requirements to be too complicated and not very purposeful. *Accordingly, a more straightforward approach is necessary to describe the requirements.*

2.3. Autonomous systems

In the last decade, intelligent and almost autonomous systems made their next step in development. The Internet of Things (IoT), for example, aims to connect every device over the internet and thereby provide new communication paths for autonomous systems and smart devices [10]. In parallel, this also boosts the research stream of autonomous and self-adaptive systems. A self-adaptive system can evaluate and change its own behavior in response to the environment [11]. The growing IoT infrastructure creates new applications and opportunities for autonomous devices and enables devices to utilize data from many different sources [12].

In the scope of the research project *Recycling 4.0*, an autonomous robot system to disassemble traction batteries was already presented [13]. During its process lifecycle, the robot system needs different kinds of data: For example, different training datasets for the AI-based functions and data about the current product, like the state of health. In this Recycling 4.0 framework, a data marketplace was already presented, as an option to gather data, via REST [14]. Accordingly, in the current implementation, the robot system can buy and download data via the marketplace. Still, the search and the decision process are limited and need the support of human actors. Obviously, this is a limitation in the process and contrast to the vision of autonomous systems. *Respectively, a framework is necessary that allows autonomous systems to describe their requirements for data as well.*

2.4. Problem statement and design goal

The implementation of the data marketplace is currently limited in its functionality. On this basis, we made a further iteration of collecting requirements for the data marketplace, according to the *Relevance Cycle* of the *Design Science Research Methodology*. Within the scope of our evaluations, the main point of criticism is the restricted purchase decision support.

Vision: In an ideal data marketplace, the consumer can compare all available datasets against each other and choose the most suitable dataset for its requirements. Furthermore, a consumer can be either a human or a non-human actor.

Problem: Since complete product transparency of the datasets is not possible because the disclosure of the dataset will reveal the information content, methods are necessary to support the data consumers. One way was therefore already presented but limited because the set of answers was limited to {yes, no} and it was relatively inflexible in comparing different datasets against each other. Moreover, initial case studies revealed that the use of the methodology is too complex and not suitable for most of the users.

Method: To overcome these limitations, as described above, an extended data trading framework is necessary to compare the data consumer requirements against different datasets on offer and support their purchase decision in an extended way.

In summary, *the Design Goal of this paper is to provide an extended data trading framework that allows data consumers to compare datasets against each other in terms of their requirements in a simple way and includes non-human actors.*

3. Related work

While the last Section 2. was representing the *Relevance Cycle* of the *Design Science Research Methodology*, this section represents the *Rigor Cycle* and provides domain knowledge. We first present some related research work in the area of data marketplaces, followed by the basics of understanding and interpreting data, in order to find similarities.

3.1. Data marketplaces - quo Vadis?

Current trends in data trading are already explored by Stahl et al. in different surveys [8], [15], [16]. In three iterations, Stahl et al. reported and presented the current trends related to data trading. Especially in their third iteration [17], one main result was the information paradox, as already discussed in a previous paper [4]. The information paradox directly leads to the challenges

of data quality and requirements to a dataset. Further trends Stahl et al. identified are related to the pricing models of datasets and the integrity of the same datasets (data origin).

Azcoita et al. identified similar challenges in their work, like: *Data buyers need to have a way to estimate the value of a coalition of datasets*, and *data buyers need to protect against strategic data sellers* [18].

Agarwal et al. propose in their paper *A Marketplace for Data: An Algorithmic Solution* a mathematical model for a 2-sided market [19]. T numerical features present m Sellers, and N buyers each having T labels which they want to predict well. The marketplace should match the buyers one after another to the sellers setting the prize depending on how many buyers arrived before. Buyers can test their own ML algorithms. The central market sets individual prizes. The biggest challenges regarding the data aquisation are truthfulness, revenue maximization, revenue division, and computational efficiency.

In *Big Data Market Optimization Pricing Model Based on Data Quality*, the authors propose to describe quality by accuracy, completeness redundancy, data volume, latency, response time, timeliness [20]. The quality score is a linear function of all of those factorss, each one specifically weighted. Moreover, they consider the use of quality by an exponential-based utility function. The parameters of this function are determined by minimizing the Sum-Squared-Error of the assumed classification utility to the real accuracy. Furthermore, a profit function depending on quality and price is proposed, and the profit is optimized with Karush-Kuhn-Tucker optimization.

Nevertheless, different authors propose machine-learning-based approaches to calculate the price of datasets, like, for example, Jia et al. based on k-nearest-neighbors [21].

Fernandez, Subramaniam, and Franklin discussed in *Data Market Platforms: Trading Data Assets to Solve Data Problems* problems and challenges for data trading in general [22]. Such as conducted, in previous work [4], they identified *pricing*, *degree of trust*, and *data quality* (fulfillment of requirements). Moreover, they concluded that the understanding of data mainly hurdles all these problems. We currently lack theories of how to understand (big) data, combine different data, or define a fair for datasets.

3.2. Understanding data and data similarities

The field of research that tries to bridge these lacks mentioned above is *Data Science*. *Data Science* is the study of the generalizable extraction of knowledge from data [23]. Even if it is hard to pin down what exactly data science is, at a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data [24]. Moreover, methods of *Machine Learning* (ML) and *Artificial Intelligence* (AI) are closely related to data science. AI is divided into three main parts: *Supervised learning*, *unsupervised learning*, and reinforcement learning. While supervised learning needs *labeled data*, unsupervised learning techniques are working with *unlabeled data* [25]. Labeled data describes a set of data that has been tagged with one or more labels to describe their characteristics (a label for images could be cat or dog). Unlabeled data instead are not previously classified or characterized. In the field of unsupervised learning, the machine tries to recognize patterns in the input data that deviate from the structureless noise.

3.2.1. Autoencoder. One type of neural network that learns in an unsupervised manner is an *autoencoder*. An autoencoder aims for learning a representation, so-called *embedding*, for a dataset, by reducing the dimensions of the dataset. The schema of a basic autoencoder is shown in figure 3.

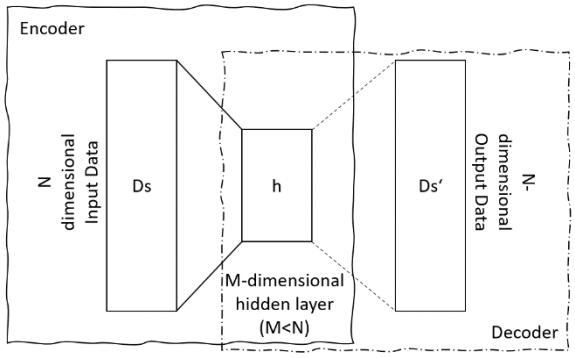

Figure 3. Autoencoder overview (own figure based on [25])

An example of an autoencoder field of application is the compression of images. There, a large tensor of data should be compressed to one with a much smaller size and dimensionality. This compression phase is called encoding. The ML algorithm uses a decoding phase for evaluation, in which it tries to rebuild the original data from the smaller size tensor. Depending on how close those rebuilt data is to the original one, it is evaluated positive or negative in a learning phase.

3.2.2. Cluster analysis. Another concept of unsupervised learning is the Cluster analysis. Cluster analysis refers to methods for discovering similarities and features in (normally large) unlabeled datasets. Cluster analysis is a key technique in order to find structures in immense amounts of data, which (after standardized preconditioning) may form individual,

fingerprint-like patterns to compare different requests and offers in trading processes. Cluster analysis is done by using some metric or distance. For example, one can use Euclidian or Manhattan norm to compare a set of numerical vectors to calculate distances between the data points.

Wagner and Wagner explored a variety of clustering algorithms in the literature, discussed their advantages and disadvantages, and came up with the first step towards formalization [26].

Huang et al. propose a parameter-free spectral clustering method that promises to overcome many open issues of spectral clusterings, such as dealing with noise [27].

One special kind is the k-means-clustering, which is rather old in the history of machine learning [28].

The value of k, the number of clusters, must be given as input to the algorithm. In each iteration of the algorithm, we have k-means of clusters. Based on the chosen metric or distance, the means are moved to a better position step by step. To find that better position, all data points are assigned to the cluster with the closest mean, and afterward, the new means are calculated by finding the average of all data points of that cluster. This process can be stopped as soon as there are no relevant changes in the positions of the means anymore.

3.3. Grounding

In summary, established solutions and best practices in the field of data marketplaces are still missing. Data marketplaces are still an emerging field of research. The use of data science techniques, such as machine learning algorithms, is rather typical for pricing datasets. Instead of finding fair pricing models, we hereby aim to verify datasets against data consumer requirements. Since the concrete consumer requirements are unknown, we identified some similarities of the concepts of unsupervised learning and provide grounding theories of these fields for our *Design Cycle*. Namely, autoencoding to extract features from datasets and clustering to find similarities between these features.

4. Overall approach

So far, we have introduced the baseline situation, especially the previous version of the data marketplace, which is described in [4] in more detail and summarized the requirements for our next iteration. Subsequently, we analyzed some related work to provide a knowledge base for our artifact. Now, our *Design Cycle* aims to reconstruct the data marketplace in further iterations until we reach a satisfactory design, such as proposed by Hevner [7].

4.1. Requirements and approach

As already deducted in this paper as well as in previous and related papers, one of the main challenges in data trading is verifying the *data quality*. Since the last approach was too complicated based on Prolog and

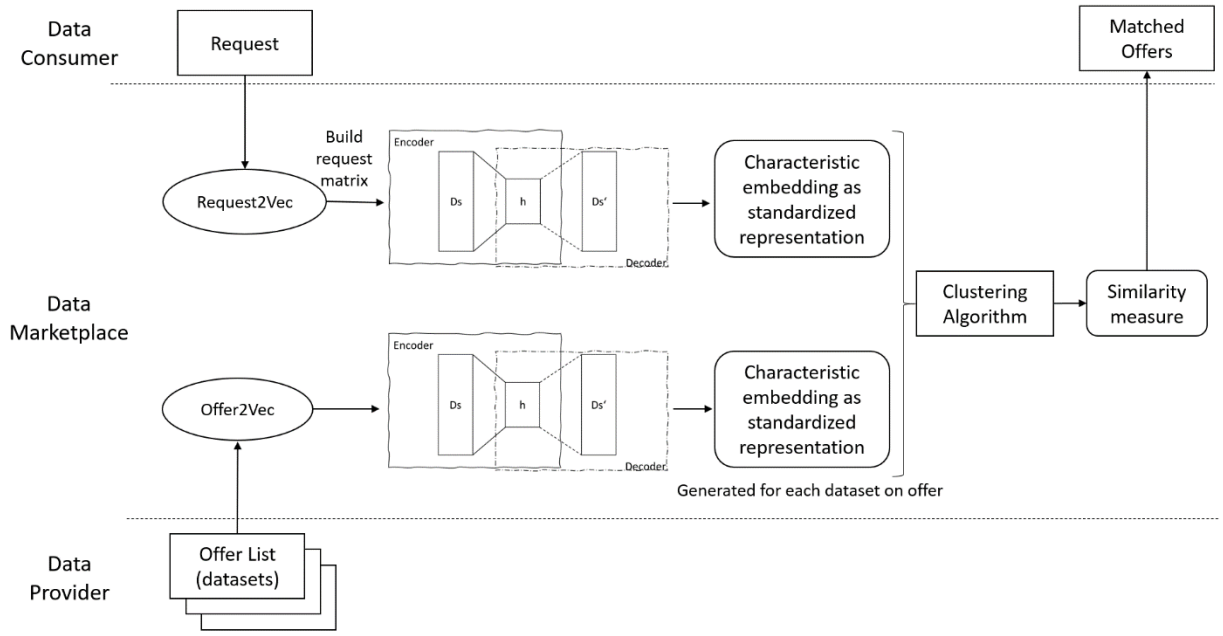

Figure 4: Process of finding data trading similarity signatures

first-order logic, a simplified approach is needed. Respectively, the following process and framework are designed to present a simple approach for data consumers to describe their data quality requirements and an approach suitable for autonomous systems.

Figure 4 shows the overall process. The feature values of all available datasets on the data marketplace are scaled to a normalized range before being encoded to characteristic embeddings of a specific form using a DNN-based autoencoder and stored in the *Offer List (see figure 2 and figure 5)*. The latter only stores the feature values and not the complete datasets. The request of the data consumer is translated into a machine-readable format (e.g., CSV), and a parser translates the requirements into a feature-oriented request matrix. This request can either be a dataset previously used with the needed characteristics, a formulated description of the required data, or direct input in matrix form. Afterward, an autoencoder creates a characteristic embedding of the request matrix analog to the encoding of the offers. Finally, these characteristics for requests and each offered dataset will be clustered by a clustering algorithm in order to determine the similarity between the request and each dataset. The highest matches of a cluster compared to the embedding of the request will be offered to the data consumer accordingly. The matched offer can be value-oriented or by means of potential utility.

4.2. Core concept

The main element of the method depicted in figure 5 is an autoencoder that can compress the data in the offered datasets and the request to an embedding of a standardized form and dimensionality. As we do not have the possibility to evaluate the datasets directly due to the privacy aspect, an autoencoder offers an unsupervised training process to learn representations containing the same amount of information in a much denser state. An efficient clustering can only be achieved due to this processing, as datasets are naturally different on diverse topics.

The autoencoder itself can be fairly easy in its composition, as simpler structures can be applied to a broader range of different datasets. Moreover, the encoding of the embedding *h* for each request or offer is the foundation for comparing the similarity measurement in the cluster analysis.

As the clustering algorithm needs to be adapted to just one form of embedding vector, the parameters (most notably, *k* in a *k*-means) can be determined once and then be applied in various process cycles. Continuing to run this principle in a real environment would lead to more data and improvement via batch-wise retraining for transfer learning.

In conclusion, data quality can be understood as an individual feature in each trading process defining the consensus of the offered dataset meeting the requirements of the prospective data consumer. For example, if two data consumers need different columns of a dataset, they would each only care about the percentage of null value in columns used by their algorithms. This requirement could be formalized to a query to check that. However, not all requirements are so clear and easy to formalize for humans and autonomous systems.

Therefore, we propose the idea that the data consumer can find datasets by looking for data that is similar to another dataset on which they know the algorithm to perform well. For example, suppose the data consumer already has the data of traction battery type A, from which a robot learned to optimize the recycling process of those batteries. In that case, it can give this previous dataset as an input. The marketplace will find data with similar content and features based on our process.

The similarity between offer and request has to be calculated without giving the data to one of the parties involved. The method must be able to cope with numeric and categorical data, ideally through numeric representation in a low dimensional space. The concept is based on the following fundamentals:

A significant concern in trading datasets is the amount of information and the percentage of noise or redundancies in a dataset. Furthermore, requested information may be included as a sub-set of a much larger set of data which can then be acquired partly or even in total (depending on the price). A modification to extract the required information is also possible.

As datasets come in various forms and dimensionalities, similarity clustering cannot be performed directly by using the initial data. In order to achieve a standardized and comparable input for the clustering, our approach uses autoencoding to create an embedding of the dataset. This step has two advantages: We can keep a maximum of the information content and therefore the *individuality*, both of the requests as of the offered dataset, and we get a standardized representation form for the comparison in the clustering step. Another surplus of this method is the practical encryption, as the DNN-encoding of the embeddings happens inside the marketplace infrastructure, and the neural network parameters are not publicly known.

In our proposed concept, we are interested in that cluster to which the datapoint of our request is assigned. The other data points in that cluster belong to datasets that are very similar to that one. Therefore, they can be proposed to the data consumer.

4.3. Marketplace integration

After introducing the fundamentals of the approach by using autoencoding and clustering, we present how we integrate these features in the already existing marketplace, as shown in figure 2. When a data provider creates a new offer, the autoencoding neural network learns the encoding transition. For that, he uploads the dataset, and the neural net chooses encoder Φ and decoder φ. Criteria for that is that the result of encoding a dataset Di to a vector hi (Equation 4.1) and afterward decoding of hi to Ďi (Equation 4.2) is as similar to Di as possible.

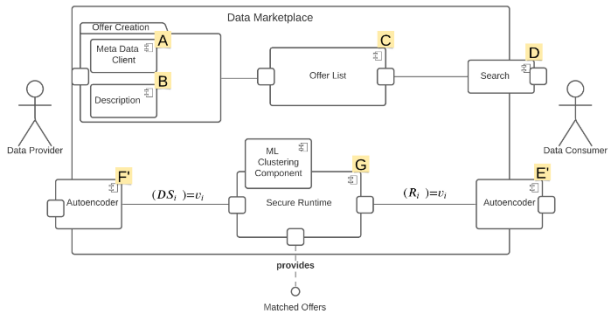

Figure 5: Integrated data marketplace structural view

$$\Phi(D_i) = h_i \quad (4.1)$$

$$\varphi(h_i) = \check{D}_i \quad (4.2)$$

The *autoencoder* component will replace the old *Meta Data Client (A)* from the *Offer Creation* package, which will still contain the *Description (B) (*see figure 2*)*.

If a data consumer is looking for some special dataset, a specific request is proposed. It consists of a dataset D_R on which he knows his algorithm to perform well or (/and) its requirements directly. The marketplace transforms those requirements to a feature matrix and executes autoencoding on the same matrix. If the autoencoder inputs the dataset D_R. Therefore, this autoencoder component will replace the Requirements Adapter (F) (see figure 2). Finally, the k-means cluster algorithm is running in the secure runtime. The produced outcome is the dataset with the minimal distance between the input datasets and the input requirements. The integrated structural view of our data marketplace is shown in figure 5.

4.4. Concept summary

In summary, the proposed concept compresses all datasets into smaller vector embeddings by autoencoding. This form of representation in a standardized structure can be used to compare the request to the datasets on offer in a way that only the matching of the information content matters. The input request set of the data consumer is chosen based on criteria of performance of similar datasets in former operational processes. After the encoding, all vector embeddings are clustered by k-means-clustering in order to determine the actual similarity. The closest offer is defined as a match to the request and forwarded to the data consumer as a recommendation.

5. Concept evaluation and discussions

Compared to the requirements and use cases that are shown in figure 1 and the results of our baseline data marketplace artifact, this section evaluates the new marketplace artifact against the requirements and the scenario firstly, and secondly, discusses the approach in a critical perspective.

5.1. Concept evaluation against the scenario

In brief, there are two important use cases for our proposed concept, the process when the data provider creates an offer instance at the marketplace and when a data consumer searches for offers that fulfill its requirements.

In the first case, a data provider has a dataset D_k which the data provider wants to sell. However, the data provider does not want to upload the data since the marketplace would store the dataset (in contrast to the data privacy requirements). Instead, the data provider creates an offer with attributes like name and description and generates a vector embedding v_k for the dataset. This vector can be calculated by the function Φ, which was learned by autoencoding in a training phase. (Equation 4.1) The generated vector is uploaded to the marketplace as it is needed later for finding similar datasets.

For the second case, the marketplace already contains a list of n offers, each created for a dataset D_i with $i \in \{1,2,..,n\}$. For each of these offers, a vector v_i with compressed data is uploaded to the secure runtime. In addition, we assume that the data consumer either has at least one dataset he knows the algorithm to perform on sufficient (such as described in the Scenario in Section 1.3.) or a description of the requirements. For this dataset D_0 the vector v_0 is generated, too, and this vector is given as an input to the algorithm. Now a k-means-clustering over the set of all vectors $\{v_0,v_1,v_2,..,v_n\}$ is executed, and it is determined which vectors are in the same cluster as v_0. For those, the offers they belong to are proposed to the data consumer.

In Section 2.1. and figure 1, we conclude four use cases and two non-functional requirements. The use cases for the *data provider* are planned to create an offer and to suggest fair pricing, especially in combination

with data security. Our extended approach fulfills these use cases since they were already fulfilled in the previous data marketplace architecture.

Furthermore, our new approach increases data security since the whole dataset is no longer stored in the secure runtime, just the vector v_i. Data consumers can verify their requirements by using already well-known datasets as inputs. A feature that bridges the complexity of describing requirements in a structured way by using Prolog. This feature was wished by many testers since data consumers can easily describe or create a good dataset but not describe it formally. We are currently planning an algorithm to create a dataset based on these features for these data consumers who prefer to write their requirements still structured using Prolog. All in all, all use cases and requirements are still fulfilled, and consumer requirements out of our case study are considered. Autonomous systems can now interact without human support with the marketplace.

5.2. Discussion and limitations

The main drawback of the extended approach is, as already mentioned, the need for an input dataset from the data consumer. However, conducting several interviews with data scientists is the preferred way to identify datasets. One huge advantage of using an autoencoder is the possibility of denoising [29]. The marketplace could learn "bad" or insufficient requirements but improve them on a higher level — another step towards an easy process. For the next steps, a better evaluation of our current concept is necessary. The current design phase focused on extending the previous concept to fulfill the new requirements but not on the evaluation.

In addition, further research in cluster algorithms is necessary, especially in the area of k-means algorithms and their performance and suitability in the domain of data marketplaces.

5.3. Alternatives

The current approach is in an early stage of development and was mainly focused on developing the architecture and concepts that fulfill the requirements, which we deducted in a second iteration. Even if the first small evaluation results are promising, there is still a high risk that the approach does not entirely fit. Regardless of the success of individual machine learning components currently selected, we can state that there is no getting about the use of machine learning. Machine learning is the perfect compromise between complex formal requirements and simplified requests.

An alternative solution could be based on labeled datasets. A label between requests and datasets that fulfills these requests. These data could trained by any kind of neural network.

6. Conclusion and outlook

This paper aimed to extend the previous data marketplace framework in a new iteration regarding the communication of consumer requirements, including autonomous machines. The research was conducted by using the Design Science Research Methodology. By field-testing the previous data marketplace and extending the scope from human to non-human actors as data consumers, we deduce old and new requirements in a structured way. Furthermore, we summarized the most important related works and current trends in the field of data trading. Finally, we presented our extended data trading requirements framework based on autoencoding and clustering algorithms. Therefore, the approach bridges the current limitations and increases data security. All in all, we were able to test our approach already with small use cases and limited datasets. However, a structured, statistically valid, and code-based analysis is still missing and subject to further investigation beyond the scope of this paper.

Another possible step of further improvement is to apply the developed approach on sub-sets and feature-specific components of datasets. This could lead to individual data signatures for each feature, which would be an opportunity to improve the formulating process of the requests, as a data consumer would be able to give discrete requirements on each feature without an example dataset into the process (generation of a comparable sequence).

In the next steps, we will provide this technical evaluation and discuss the accuracy of our algorithms. Moreover, further research in the area of clustering algorithms and the optimization of encoders are necessary to improve our accuracy of matching predictions. Evaluating different kinds of learning strategies and neuronal networks is also seen as a technical perspective for improvement. Re-evaluating the achievements of the approach on a second user-based case study will lead to a more substantial validity of the results regarding the initial premise of our research.

Conclusively, we see a high chance of establishing a new method that can serve as an informational fingerprint for individual datasets in the field of trading data. In perspective, we are also planning to extend the framework again to buy parts of datasets without manipulating the prices (arbitrage).

7. References

[1] Rowley, J., 2007, The wisdom hierarchy: representations of the DIKW hierarchy, Journal of Information Science, 33/2:163–180, DOI:10.1177/0165551506070706.

[2] Quah, D., 2003, Digital goods and the new economy.

[3] Shannon, C. E., Weaver, W., 1949, THE MATHEMATICAL THEORY OF COMMUNICATION.

[4] Lawrenz, S., Rausch, A., 2021, Don`t Buy A Pig In A Poke A Framework for Checking Consumer Requirements In A Data Marketplace, Proceedings of the 54th Hawaii International Conference on System Sciences, 0:4663–4672, DOI:10.24251/hicss.2021.566.

[5] Kintscher, L., Lawrenz, S., Poschmann, H., 2021, A Life Cycle Oriented Data-Driven Architecture for an Advanced Circular Economy, Procedia CIRP, 98:318–323, DOI:10.1016/J.PROCIR.2021.01.110.

[6] Wieringa, R. J., 2014, Design science methodology: For information systems and software engineering. .

[7] Hevner Alan, R., 2007, A Three Cycle View of Design Science Research, Scandinavian Journal of Information Systems, 19/2:87–92.

[8] Stahl, F., Schomm, F., Vossen, G., Vomfell, L., 2016, A classification framework for data marketplaces, Vietnam Journal of Computer Science, 3/3:137–143, DOI:10.1007/s40595-016-0064-2.

[9] Itti, L., Baldi, P., 2009, Bayesian surprise attracts human attention, Vision Research, 49/10:1295–1306, DOI:10.1016/j.visres.2008.09.007.

[10] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M., 2015, Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications, IEEE Communications Surveys and Tutorials, 17/4:2347–2376, DOI:10.1109/COMST.2015.2444095.

[11] Macías-Escrivá, F. D., Haber, R., Del Toro, R., Hernandez, V., 2013, Self-adaptive systems: A survey of current approaches, research challenges and applications, Expert Systems with Applications, 40/18:7267–7279, DOI:10.1016/j.eswa.2013.07.033.

[12] Kyriazis, D., Varvarigou, T., 2013, Smart, autonomous and reliable Internet of Things, Procedia Computer Science, 21:442–448, DOI:10.1016/j.procs.2013.09.059.

[13] Poschmann, H., Brüggemann, H., Goldmann, D., 2021, Fostering End-of-Life Utilization by Information-driven Robotic Disassembly, Procedia CIRP, 98:282–287, DOI:10.1016/j.procir.2021.01.104.

[14] Kintscher, L., Lawrenz, S., Poschmann, H., 2021, A Life Cycle Oriented Data-Driven Architecture for an Advanced Circular Economy, Procedia CIRP, 98:318–323, DOI:10.1016/j.procir.2021.01.110.

[15] Stahl, F., Schomm, F., Gottfriedm, V., 2014, The data marketplaces survey revisited, /March:25.

[16] Stahl, F., Schomm, F., Vomfell, L., Vossen, G., 2015, Marketplaces for digital data: Quo vadis?, Westfälische Wilhelms-Universität Münster, European Research Center for Information Systems (ERCIS), Münster.

[17] Stahl, F., Schomm, F., Vomfell, L., Vossen, G., 2015, Marketplaces for digital data: Quo vadis?, Westfälische Wilhelms-Universität Münster, European Research Center for Information Systems (ERCIS), Münster.

[18] Azcoitia, S. A., Paraschiv, M., Laoutaris, N., 2020, Computing the Relative Value of Spatio-Temporal Data in Wholesale and Retail Data Marketplaces, /February 2020.

[19] Agarwal, A., Dahleh, M., Sarkar, T., 2019, A marketplace for data: An algorithmic solution, ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation, pp. 701–726, DOI:10.1145/3328526.3329589.

[20] Yang, J., Zhao, C., Xing, C., 2019, Big Data Market Optimization Pricing Model Based on Data Quality, Complexity, 2019:5964068, DOI:10.1155/2019/5964068.

[21] Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., et al., 2018, Efficient task specific data valuation for nearest neighbor algorithms, Proceedings of the VLDB Endowment, 12/11:1610–1623, DOI:10.14778/3342263.3342637.

[22] Fernandez, R. C., Subramaniam, P., Franklin, M. J., 2020, Data market platforms: Trading data assets to solve data problems, Proceedings of the VLDB Endowment, 13/11:1933–1947, DOI:10.14778/3407790.3407800.

[23] Dhar, B. V., Dhar, V., 2013, Data Science and Prediction, Commun. ACM, 56/12:64–73, DOI:10.1145/2500499.

[24] Provost, F., Fawcett, T., 2013, Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data, 1/1:51–59, DOI:10.1089/big.2013.1508.

[25] Joshi, A. V, 2020, Machine Learning and Artificial Intelligence. Springer.

[26] Wagner, S., Wagner, D., 2007, Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

[27] Huang, J., Nie, F., Huang, H., 2015, A new simplex sparse learning model to measure data similarity for clustering, IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua/Ijcai:3569–3575.

[28] Shi, N., Liu, X., Guan, Y., 2010, Research on k-means clustering algorithm: An improved k-means clustering algorithm, 3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010, pp. 63–67, DOI:10.1109/IITSI.2010.74.

[29] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., 2010, Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, Journal of Machine Learning Research, 11:3371–3408.