Thesis Proposal:

The Role of Embeddings in Data-Driven Augmentation

Federico Dassereto

October 7, 2020

Abstract

Querying data lakes containing a large amount of unstructured data is a difficult task, accentuated by the facility of producing data due to the rise of the World Wide Web. A traditional operation on data lakes is to search for joinable and unionable tables, to find tables related to specific user requests. Existing approaches search for columns with the best overlap possible, returning columns from tables in data lakes with high joining or unionability rating. These approaches fail to capture the situation in which a data scientist wants to find related tables related to a particular query table when she is trying to build a machine learning model to predict a column of her query table. The reason for which existing approaches fail is that they do not consider the impact that a joining table has on the predicting tasks, due to the lack of semantic surrounding the data lake.

The goal of the thesis is to define an automatic table augmentation framework, to improve the performance of machine learning models in predicting target columns of a query table. The idea is to start by studying the augmentation problem as a problem on data lakes only, with measures and concepts derived from information theory and functional dependencies, and then integration embeddings in the process. Embeddings will be integrated by mapping tables to knowledge bases and exploiting their semantics to automatically produce the augmentation. More precisely, the thesis will first propose a framework for automatic horizontal table augmentation in data lakes, allowing a data scientist to add relevant features to her query table from a ranking of joining tables. Furthermore, an extension comprehending a mapping of tables to knowledge bases and exploiting knowledge bases embeddings will be developed, to study the role of embeddings in automatic tables augmentation. Finally, to sum up, the previous results, an ad-hoc embedding algorithm for knowledge bases will be developed to emphasize the relevant properties a knowledge base should have to produce a relevant augmentation.

1 Motivation and Description

Since the rise of the World Wide Web, we have experienced exponential growth in publicly available data. Data has become a crucial element in everyday life and many devices continuously produce and store them in a forest of different formats. This exponential growth has posed many challenges to scientists (even going so far as to create the figure of the *data scientist*) in every step in which data are traditionally used: from data collection to

integration, from processing to visualization. The impact of *Big Data* has been summarized in [40], where the key properties are **V**olume, **V**elocity, **V**ariety, **V**eracity and **V**alue. Following the path traced by Tim Berners-Lee et al. in [4] regarding Open Data publication and maintenance, many organizations are publishing data, augmenting tremendously the **V**olume and the **V**alue of such data. On the other hand, the explosion of sources providing data increases their **V**ariety, requiring new techniques to integrate data.

In data science, it is increasingly the case that the main challenge is not in integrating known data, rather it is in finding the right data to solve a given data science problem. Today, data is a mass (uncountable) like dust, and data surrounds us like dust, even lovely structured data. Data is so cheap and easy to obtain that it is no longer important to always get the integration right and integrations are not static things (**V**elocity component). Data integration research has embraced and prospered by using approximation and machine learning.

A data scientist who is trying to solve machine learning tasks, frequently found herself in a situation in which there are not enough data to build a good model. Given the amount of data previously discussed, it would be very useful a framework which enables her to find automatically new features or samples to improve the model. Broadly, two main classes of augmentation can be distinguished: *Horizontal Augmentation*, in which new significant features are added to the dataset and *Vertical Augmentation*, in which new samples (tuples) are added to the dataset. Ideally, the two classes can be seen as the search for joinable (Horizontal) or unionable (Vertical) tables in the data lakes situation, and Knowledge Bases (KB) completion or extensions (vertical).

The uncontrolled nature of data manifests in large repositories of data (*data lakes*), in which both structured and unstructured data are stored. In data lakes, it is common the situation in which there is a lack of schemas, making traditional database approaches to integrating or querying data difficult to pursuit or even infeasible. In the absence of a data schema, an effective metadata system becomes essential to make data queryable and thus prevent the lake from turning into a data swamp, i.e., an unusable data lake [60, 27]. Nevertheless, it is frequent the case in which metadata are not provided. To overcome this issue, approaches mining metadata from tables have been proposed [2, 58]. Along with the uncertainty regarding the quality of the data, the amount of such data makes it infeasible also the traditional human-in-the-loop framework, since hand-labeling or manual rating of very large amounts of data is extremely expensive. It is easy to see that searching in data lakes is a complicated task, both in terms of time and methodology.

At a processing level, the main difficulty arisen by the multiple sources and possibly unstructured data is given by the interpretation, i.e., the semantic layer. To extract semantics from unstructured data, in the last years the concept of *embedding* has been proposed. Embeddings are predominantly used in learning representations of texts [42] and graphs [46]. In the last year, an approach exploiting embeddings to solve data integration tasks [9] has been published. This work can be seen as a forerunner for the integration of embeddings in data management tasks, even if it works on tables from the same context. A key contribution to the integration of embeddings by Martin Grohe in [25], where many research questions have been posed, with a crucial point on *"Can we answer queries on the embedded data?"*, instead of current approaches, that exploit embeddings to answer queries. The main point arisen is that there is a lack of a unified framework, or theory, that clearly indicates the path through embeddings complete integration.

In this thesis, we focus on the problem of *using embeddings for automatic augmentation of data to improve machine learning tasks*. We believe that it is a crucial integration task that has not yet been explored, either as an augmentation problem itself nor with the usage

of embeddings. Few approaches that try to augment tables concerning a repository have been proposed; the two closest approaches proposed until now show different problems: (i) a set of rules to decide if it is safe to avoid a join or not [35], restricted to a *pure relational* settings, i.e., the schemas of each table is known; and (ii) a feature selection algorithm working on a matrix in which the number of attributes is much larger than the number of tuples [12].

The goal of this thesis is to study how to stabilize the automatic increase and make it scalable to possibly huge tables or data lakes. The idea is to define indexing structures, based on information-theoretic measures, to make the search for joining tables fast and adapt to the variety of existing joining possibilities (one-to-one, one-to-many, many-to-one, many-to-many). More precisely, the thesis will proposed an index for horizontal augmentation on the data lakes scenario, as well as the integration of embeddings in the augmentation of tables via KBs. Eventually, an embedding algorithm for KBs will summarize in a unified framework the two previous results. Augmenting via KBs can be done with the usage of *Knowledge Graphs*, exploiting the graph model to represent semantics information, e.g., links between entities, objects or even abstract concepts. Knowledge Graphs (KGs) are usually associated with linked open data, and allow a user to perform queries directly on the graph and moving among different sources of open data, i.e., they allow for retrieval of explicit knowledge. Furthermore, it is possible to add an ontology as a schema layer and exploit its logical rules inference system to retrieve also implicit knowledge. We claim that a key ingredient to reaching automatic augmentation is the concept of functional dependency, thanks to its ability to catch (possibly fuzzy) relations between attributes in tables. Functional dependencies have been extended even to graphs [20], so this concept seems suitable for both the scenarios we consider, i.e., tables augmentation and knowledge bases augmentation (as long as a KB can be seen as a graph).

Previous Works. From the above discussion, we would like to underline the three main topics of the proposal: *open data*, *data-driven augmentation* and *embeddings*. In order to familiarize with these three main concepts, we deepened them in the first year and published some works. Thanks to a long term collaboration with Prof. Michela Bertolotto (University College Dublin) and Dr. Laura Di Rocco (Northeastern University, Boston) started during my Master thesis entitled *Embeddings for Geospatial Ontologies Representation*, we were able to embed geographical ontologies onto a suitable space, on which we first evaluated its quality [14] and then its impact as query engine into a data-driven microblogs geolocation algorithm [18]. We further analyze the tuning of parameters to obtain higher quality embedding and summarized our results in [15]; the key idea is that a fine-grained tuning of the parameters could drastically improve the quality of the embedding in capturing semantic similarities.

On the other hand, we also worked in the Open Data domain, trying to expand an existing approach on Linked Open Data source selection [10] by introducing the concept of embedding to better associate context information. The work is in its conclusive part as the experimentation is going through and will be ready soon for submission. A discussion on the preliminary results obtained regarding the augmentation problem can be found in Section 4.

The remainder of this proposal is organized as follows: Section 2 places this document in a specific reference area and highlights the goals of the proposal; Section 3 discusses the state of the Art of the main topics of the proposal, i.e., embeddings, open data, augmentation and functional dependencies discovery; Section 4 presents in detail the goals of the thesis, the methodology we plan to follow to reach them and the preliminary results obtained; finally, Section 5 shows the plan of work for the remaining two years.

2 Reference Area and Relevance of Goals

To the best of our knowledge, no automatic table augmentation framework adapt to work at Internet Scale exists. We believe that such a framework would be very relevant to exploit the large amount of unstructured data available on the web and on data lakes, allowing not only to improve machine learning models but also to retrieve semantic information from a variety of schemaless tables. The proposal lies in the reference area "Data Augmentation to improve Machine Learning Tasks", intending to provide to a data scientist an augmented set of data which enables her to improve model performances. The approach we propose differ from other existing solutions in the following aspects:

- Existing approaches mainly focus on predicting the usefulness of a join under the relational hypothesis, i.e., the schema information are known. To this end, we plan to develop an approach that is agnostic to the knowledge of the schemas;
- Existing approaches that try to augment data against a repository perform very expensive joins, and are evaluated concerning repositories of limited size; other existing approaches on repositories run features selection algorithms on a very large matrix derived by the join of all the joinable tables. To this end, we plan to develop an approach that does not materialize any kind of join and is scalable to hundreds of thousands of tables in a repository;
- Existing works on single tables identify the most relevant features in the table, according to a specific target. We plan to overcome the single table assumption by indexing all the tables in a repository and identifying the most relevant features, and the relative tables, that improve the machine learning model performances;
- Other approaches simply return the set of joinable tables, usually with a threshold to allow for relaxed (imperfect) join, without any ranking. We plan to rank the returned tables according to the improvement that each of them will guarantee to the model.

3 State of the Art

In this section, we discuss related work that we consider relevant for the proposed research project. In order to facilitate reading, the discussion has been organized into three main sections, corresponding to the main concepts introduced in Section 1 and 2. Each of the three sections is in turn divided into sub-parts to further facilitate reading. The first one discusses Open Data and their common operations, as well as existing augmentation approaches, the second one introduces embeddings approaches and their relevance to the proposal, finally the third discusses knowledge graphs.

3.1 Open Data

The increased spreading of Open Data has been possible thanks to fact that many organizations publish their data following the Open Data principles [4]. As discussed in Section 1, these data are continuously produce and stored; their dynamic nature makes it impossible to apply previous management techniques, such as the creation of a global schema [3] and to keeping track of joins path known or mined from the data[19, 16]. All of the above-mentioned techniques requires tables to have schema information or meaningful attribute names, as well as managing pre-computed join-paths. Both these practices are impracticable for an open platform at Internet scale, since the number of tables can easily reach

millions or more. The most relevant task while dealing with Open Data is the discovery part, which include searching for tables containing specific value(s) based on keywords [8] or containment. Reconnecting to what we discussed in Section 1, we further identify two macro-areas in Open Data discovery, namely Join Search and Union Search.

Join Search. Given a query table $T_q(A_1, A_2, ...)$ with join attribute A_j, a joinable table is a relation $T(B_1, B_2, ...)$ such that T has at least one attribute B_i that is equi-joinable with A_j. To be more precise, the values in the two attributes must have significant overlap. The problem has been largely posed as set similarity search, where the values of the attributes are sets and a similarity function determines the relatedness. A frequent similarity function is the Jaccard Index, which suffers to the problem of unfairly advantages small domains [65]. Many solutions have been proposed exploiting this idea, but they all work under the assumption of having tables with average sets size ranging from few columns to a maximum of few hundreds. This is not the case of open data lakes, where there hundreds of thousands of tables possibly with millions of rows. To solve this issue, LSH Ensemble [65], an index based on locality-sensitive hashing [23] and MinHash [31] that uses containment, poses the problem as a domain search, where each attribute of a table is a domain. LSH Ensemble requires on a threshold value, making it an approximate algorithm for joining tables discovery. The family of approximate algorithms are scalable in performance, however they tend to suffer from false positive and negative errors, especially when the distribution of set sizes is skewed (as frequent in data lakes). Furthermore, using a threshold may confuse users, who have no knowledge of which data exists in the lake and therefore do not know how to set a reasonable threshold that will retrieve some, but not too many answers. A possible alternative to threshold search is to retrieve the top-k tables with higher containment. Another effective algorithm for join search is JOSIE [64], a scalable exact top-k overlap set similarity search algorithm. JOSIE minimizes the cost of set reads and exploits an inverted index to find the top-k sets. Due to its inverted index structure and the prefix filter, a fast trick to solve the threshold version of set similarity search problem [11], JOSIE outperforms its approximate counterparts for small k.

Union Search. Given a query table $T_q(A_1, A_2, ..., A_n)$ with n attributes, a unionable table is a table $T(B_1, B_2, ..., B_k)$ such that T has at least one attribute B_i that is unionable with some attribute A_j from the query table. Unionability means that, given attribute A_j and its domain (set of values), and attribute B_i and its domain, it is likely to exist a third domain D from which both A_j and B_i are sampled. An approach to finding unionable tables is through schema matching, where the problem is to match the attributes of two or more tables (or schemas) [28, 53]. Two tables that match on i attributes can presumably be unioned on those attributes. Matching is done largely heuristically using similarity functions over schema (attribute names) and sometimes values (for example, using a set similarity measure) or value distributions [33]. Although scalable schema matching and ontology alignment have been studied extensively, the best solutions are drawn when considering the unionability as a search problem. In particular, in [44] they find the k tables that have the highest likelihood of being unionable with a search table T on some subset of attributes. In few words, they determine if a table S that can be unioned with T on c attributes is more or less unionable than a table R that can only be unioned on $d < c$ attributes.

Data Integration & Augmentation. There has been extensive prior work on data mining, data augmentation, knowledge discovery, and feature selection, but only a few approaches to automatic augmentation have been proposed. Kumar et al. [35] proposed a set of rules to determine if avoiding to perform a join would be safe in a relational context,

while in [56] the rules of [35] are applied to high-capacity classifiers to test their validity. Other approaches to augmentation are devoted return a set of related tables, such as [6], where a relatedness search in data lakes is performed, and [21] that automatically identify joins between tables representing similar entities. While these systems help users to discover new data and explore relationships between datasets, they do not automatically determine whether or not such new information is useful for a predictive model. To overcome this lack, recently a framework called ARDA [12] has been proposed. ARDA works as a two steps algorithm, firstly it searches for join and then prunes out irrelevant features with features selection algorithms. ARDA materializes the joins between the input table and the tables in data lake, making it difficult to ensure scalability (in fact, their experiments are run on data lakes of less than a hundred tables).

3.2 Embeddings

Embeddings allow systems to capture non-geometric data in a mathematical structure, useful for easier comparison of data. A lot of attention has been devoted to word embeddings, i.e., embeddings of documents in which every word is represented by a vector.

Word embeddings have been proposed in a NLP context thanks to their ability to capture semantic relations among words in a text. Word2Vec [43] is an example of a pretrained embedding that embeds into Euclidean Space. Later, other embedding techniques like GloVe [51], BERT [17] and ELMo [52] have been proposed, only to mention the most famous. All these algorithms belong to the family of Euclidean embeddings, i.e., project texts onto a Euclidean space. Although these methods are very effective on texts, they fail to accurately capture a different kind of sources such as graphical structures [47].

In contrast, hyperbolic embeddings are particularly suitable to embed hierarchical data. As described in [34], hierarchical structures show a hidden hyperbolic geometry. On such observation, many hyperbolic embeddings have been proposed: the Poincaré Disk Model [47], the Lorentz Model [45] and a convex entailment cones approach [22]. It is worth to notice that Poincaré and Lorentz models produce the same projection in different spaces; the only difference is in the way they are computed, since the Lorentz model leads to more stable computations. Finally, a mixed approach was proposed in [26] to take advantage of the properties of both hyperbolic and Euclidean spaces.

Along with the rise of word embeddings, the need of quality indicators for these structures raised too. Initially, no common way of assessing the quality of embeddings was developed. In fact, each embedding algorithm was evaluated in a task-dependent way, which means that if the embedding performs well on a particular task it is considered "good". Recently, the concept of distortion was introduced in [55] and slightly modified in [14]. To the best of our knowledge, [14] is the first attempt to evaluate the use of embeddings in a geographical context. The distortion essentially measures how well the distances in original metric space are preserved in the embedding. Thus, it can be seen as a task-independent evaluation. In hyperbolic embeddings, all the links between entities in the structure are treated equally, but this could not be always the case; in such a situation, Knowledge Graphs Embeddings (KGE) can be applied.

Finally, hybrid approaches mixing hyperbolic and Euclidean embeddings have been proposed, based on the observation that in a text there are many asymmetric word relations. The main approaches in this direction are an adaption of GloVe in a Cartesian product of spaces [59] and a learning model in a spherical space [41].

As mentioned in Section 1, a preliminary approach exploiting embeddings for data integration tasks have been proposed [9]. In such an approach, tables are represented as graphs, by adding nodes representing columns and rows identifiers and then embedding

in Euclidean spaces following the framework of the random walk. The drawback of such an approach is that is assumed to be known context from which each table comes from, making it unsuitable for our scenario.

3.3 Knowledge Graphs

A knowledge graph is a set of triples, in the form $\langle$ **s,p,o** $\rangle$, where **s** and **o**, subject and object respectively, are entities and **p** is the predicate between **s** and **o**. KGs are traditionally store as RDF (Resource Description Framework) [13]. Since a rigorous definition of KGs does not exists, according to [49] any RDF graph can be considered a KG, as long as they (i) mainly describes real-world entities and their interrelations, organized in a graph; (ii) defines possible classes and relations of entities in a schema; (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains. This general definition allows us to consider KGs structures like DBpedia [36] and its ontology, a knowledge base extracted from Wikipedia and YAGO, [57] which is linked as well as the DBpedia ontology. A complete survey on how to build, model and query KGs can be found in [32]. Recently, the idea of embedding KGs emerged, to include them in machine learning tasks such as links prediction [54] and question answering [29]. A very effective algorithm called TransE was proposed in [7], in which embeddings are learned in a translational way; this work generated many variants such as TransH [62] and TransR [37]. A thorough survey on embedding techniques for KGs can be found in [24].

4 Goals, Methodology and Preliminaries

In this section, we start with an highlight of the macro goals of the thesis, then we expand the content of Section 1 by detailing our specific goals and the feasibility of each of them. Finally, we summarize the work done in the first year concerning the proposed goals.

4.1 Goals

The high-level goals of this thesis are:

- To show the relevant role that embeddings can play in searching tasks;
- To show that embeddings can be used directly to answer queries instead of being an orthogonal part of the answering process.

The overall goal of this research proposal is the study of the role of embeddings in data augmentation, from tables augmentation on data lakes to an embedding algorithm for knowledge bases, passing through the integration of embeddings in existing tables augmentation framework exploiting knowledge bases. We organize the work into four main objectives, each leading to specific results. The aim of Objective 1 (indexing data lakes to augment) is to analyze the state of the art on (i) open data search, and (ii) sketching and approximation of information-theoretic measures, and to define an index for tables augmentation in data lakes based on information-theoretic measures. The aim of Objective 2 (augmenting via knowledge bases) is to analyze the state of the art approaches in knowledge bases representation and augmentation, along with their role in augmenting tables and integrating of embeddings in such approaches. Objective 3 (embedding knowledge bases) aims at proposing an embedding algorithm for knowledge bases, by creating a mathematical structure that encapsulates relevant features for catching the bast augmentation possible, and an evaluation of the proposed algorithm. Finally, Objective 4 (stay

up-to-date on embeddings) is an objective which is equally distributed over the three years, in order to stay up-to-date with the newest embedding technologies and algorithms. It is worth to be over the three years since it is a very dynamic topic, and the knowledge gained in recent years on the subject would not be lost.

Objective 1: Indexing data lakes to augment.

1. *Analysis of the state of the art approaches* related to (i) open data searching and indexing, and (ii) information-theoretic measures approximation and sketching. There exist many frameworks aiming at searching on open data, for tasks such as joinability and unionability, implementing different techniques. Information-theoretic measures are very useful and, for large datasets, can require a lot of time to be computed. We plan to study how to approximate such measures as well as sketching them on a subset of the available data.

2. *Definition of an index for tables augmentation in data lakes* based on information-theoretic measures, taking advantage of existing indexing structures (e.g., inverted index) by storing the information about a table and its columns, along with information theory measures regarding columns. In such a way, exploiting the fast search offered by the index structures, it would be possible to retrieve joining columns. The key idea is that we will exploit information-theoretic measures to simulate the behavior of functional dependencies.

Objective 2: Augmenting via knowledge bases.

1. *Analysis of the state of the art approaches* for knowledge bases construction, representation and augmentation, along with their role in augmenting tables and integrating of embeddings in such approaches. Since the variety of knowledge bases available, often built available as knowledge graphs, this analysis will be deeply and sharply isolate the fundamental features that a KB is required to have.

2. *Definition of a framework that exploits knowledge bases and possibility of including embedding to augment*, by understanding how embeddings algorithms works on knowledge bases and find the best possible representation to maximize the augmentation. Knowledge bases semantics can be very useful in extending tables, by adding not only features to maximize the augmentation but also to add new relevant tuples. Once the augmentation through embeddings will be completed, a comparison between with and without embedding approaches will be conducted, in order to highlight the key role of embeddings.

Objective 3: Embedding knowledge bases.

1. *Definition of an embedding algorithm for knowledge bases*, by representing in the best way the isolated relevant and general properties that KBs should have for being suitably embedded, along with an extensive evaluation on its efficiency compared to existing embedding techniques. To the best of our knowledge, in the state of the art of embedding algorithms, no algorithm exists that exploits the semantics of its entities, rather they exploit the structural properties of the knowledge base to project it onto a new semantic space. This objective will be founded on the results obtained in Objectives 1 and 2, in order to understand if embeddings effectively play a role. Since this objective is the farthest from the time of writing, not many details are reported.

Objective 4: Stay up-to-date on embedding techniques.

1. *Analysis of the state of the art approaches*, of continuously published embeddings works. Many embeddings approaches are yearly proposed, in a variety of formats: from completely new approaches to slightly changes on existing ones, from pre-trained embedding on large structures to new benchmarks. This objective includes the tracking of theoretical new ideas and implementations, in order to able to integrate new embeddings ideas in this work.

4.2 Methodology

In the following, we describe the methodology we envision for achieving each objective described in Section 4.1.

Objective 1: Indexing data lakes to augment.

1. *Analysis of the state of the art approaches* related to (i) open data searching and indexing, and (ii) information-theoretic measures approximation and sketching. We will review the most relevant solutions to open data problems, like table joinability and unionability, as well as the relevant results obtained in approximating and sketching information-theoretic measures. To this aim, we will take into account both the scientific literature related to data management and to data analysis and learning.

2. *Definition of an index for tables augmentation in data lakes* based on information-theoretic measures. We plan to propose and index for automatic table augmentation in data lakes. In particular, given a table T, a join column t_j of T, a target column t_y of T and a data lake D composed of n tables $(D_1, ..., D_n)$, the index will return a ranking of tables in D, joining on t_j that provides the best augmentation for T, i.e., mostly improves the performances of a machine learning model that has t_y as target.

Objective 2: Augmenting via knowledge bases.

1. *Analysis of the state of the art approaches* for knowledge bases construction, representations and augmentations. We will conduct a deep and complete analysis of the state of the art approaches and algorithms to manipulate and represent KBs, as well as embedding algorithms for KBs. The main distinction in the embedding approaches will be among the ones with a geometric interpretation, i.e., transitional operations between vectors and the ones based on matrix factorization and co-occurrences, i.e., the family of embedding closer to word embeddings.

2. *Definition of a framework that exploits knowledge bases and possibility of including embedding to augment and Comparison.* In the case of semantic data lakes, i.e., data lakes with known (or derivable) mappings on KGs, embeddings can be exploited to retrieve relations between entities in the KG semantic space (for example through a nearest neighbors search or even more complex ways) and produce the augmentation. Once the framework is designed and implemented, an evaluation of the two scenarios (augmenting via KGs only and KGs embedding) will be conducted to show the advantages of using embeddings. In such a situation, an augmentation using only KGs will be taken as baseline; it could work, in its simpler version, with a mapping from tables values to entities in the KG on which a basic join operation can be performed.

Objective 3: Embedding knowledge bases.

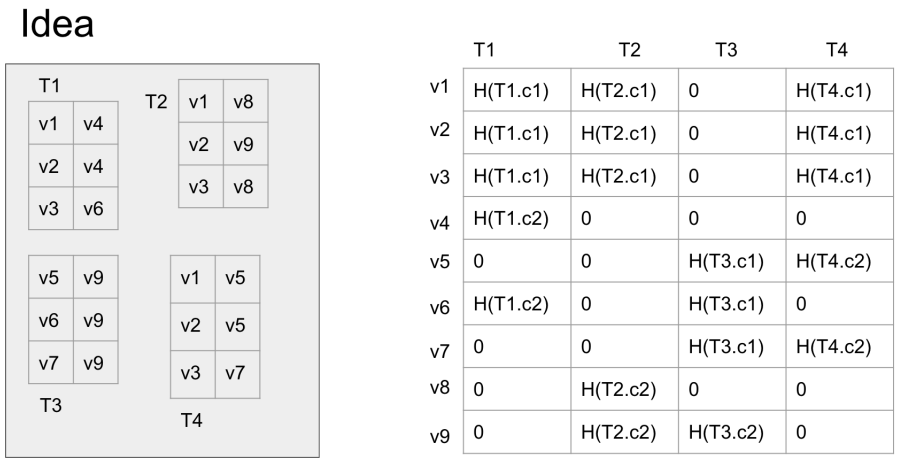

Figure 1: Sketch of entropy-based index for open data. When a value v is present in column j of table i, in position (v, i) it is stored the entropy of the column of j of table i.

1. *Definition of an embedding algorithm for knowledge bases.* Once Objective 2 is completed, we will have a picture of what features of existing KGs embedding have a positive impact on the augmentation and which one could be discarded on modified. To this aim, extensive knowledge of embedding algorithms will result in the definition of ad-hoc algorithms to embed KGs to provide table augmentation. The key concept for such an objective is that we do not want to propose a complete novel algorithm for embedding KBs, rather we plan to isolate and represent, in an appropriate way, the most important features of the KB to improve existing algorithms for KBs embeddings. In a sense, we want to integrate the semantic lying in the KB in the embedding process of existing algorithms, that up to now only consider structural properties. At such a point, a comparison between all of the approaches will be done: tables augmentation in data lakes only, table augmentation via KGs, table augmentation via KGs embeddings, and table augmentation via our KGs embedding algorithm.

Objective 4: Stay up-to-date on embedding techniques.

1. *Analysis of the state of the art approaches*, of continuously published embeddings works. In order to review and keep trace of the most relevant embedding approaches, the surveyed techniques we will then be organized into a taxonomy, according to their hypothesis, usage and domain of application.

4.3 Preliminaries

In this section we will present the activities done and the results achieved in the first year.

1. We have analyzed the state of the art approaches related to (i) open data searching and indexing, (ii) information-theoretic measures sketching and approximations and (iii) functional dependencies discovery. We found out that existing searching frameworks on open data only cares about finding the best, or the best *top-k* set, of *joining* tables or columns. This has many drawbacks, first of which the fact that the best join does not say anything about the best augmentation. For instance, the best join that a table can have is with itself, meaning a void augmentation.

2. We also noticed that there are information-theoretic measures that are particularly suitable in representing, or at least approximating, the behavior of functional dependencies. In particular, the conditional entropy $H(Y|X) = H(X) - I(X,Y)$, where $H(X)$ in the Shannon entropy of variable X and $I(X,Y)$ is the mutual information between variables X and Y, has lower bound 0 and upper bound $H(X)$. These two cases coincides with full functional dependency[1] when $H(Y|X) = 0$ (meaning that the value of Y is totally determined by X) and complete independence, when $H(Y|X) = H(Y)$ (meaning that the knowledge of variable X has no impact on the knowledge of Y). This observation, mixed with the intuition that if a functional dependency (in its descriptive interpretation) between a set of attributes and a specific target holds these attributes are interesting for the augmentation, means that discovering FDs that minimize the conditional entropy[2] produce the best augmentation. However, the approach of identifying a target attribute and discover FDs in such a way requires the materialization of the join, that, as discussed in Section 3, is unfeasible on data lakes.

3. We also analyzed the role of Functional Dependencies (FDs). FDs express relationships between attributes of a database relation. An FD $X \rightarrow Y$ states that the values of attribute set X *uniquely* determine the values of attribute Y. Despite their rich semantic information, functional dependencies are extremely hard to compute, as they belong to the W[2] complexity class [5]. Particularly interesting are the non-trivial (FDs that cannot be derived from others) and minimal functional dependencies. Many works have been proposed to mine functional dependencies from tables: lattice-based approaches such as TANE [30] and DFD [1] that performs search and pruning on the lattice structure; row-based methods derive candidate FDs from two attribute subsets, namely agree sets and difference-sets, which are built by comparing the values of attributes for all possible combinations of tuples pairs, such as DepMiner [38] and FastFD [63]; incremental approaches exploiting the concepts of tuple partitions and monotonicity of FDs to avoid the re-scanning of the database [61]. Of course, all of these approaches fail to scale to very large tables both in terms of time and memory usage.

   We analyzed many of above-mentioned FDs discovery algorithms, both exact and approximate, implemented in a popular framework named Metanome [48]. We found out, as expected, that no existing algorithm for FDs discovery is scalable to data lakes of thousands of tables and many Gigabytes of space. All of the experiments we run on single tables with thousands of rows and 30-40 attributes result in memory exceeding error or timeout. In sight of this, we chose to not follow the direction of pure FDs discovery, but to limit their usage to theoretical studies, such as upper and lower bounds studies and pruning techniques to improve our information-theoretic based index.

   The idea of inferring FDs from data has attracted a different view of such concept: the point of view change from *uniquely implies* (roughly speaking, each left side of FDs is key with respect to the right side) to *better identify* (each left side of FDs is a "substitute" of the right side). These kinds of FDs quantifies how much the left side approximates the right side; it very useful in the situation in which correlations

[1] Note that, despite the differentiation that we made in Section 3 about functional dependency, up to now the interpretation is the same on classical database FDs and descriptive FDs.

[2] Actually, our intuition derived by the fraction of information, i.e., a normalized variant of the mutual information defined as $\frac{I(X;Y)}{H(X)}$. However, it is possible to show that minimizing the conditional entropy is equal to maximize the fraction of information.

discovery and features elimination is needed. The mining of such an idea of FDs has been proposed by using information-theoretic measures like the reliable fraction of information [39] and the smoothed mutual information [50]. The (partial) drawbacks of these last approaches are that (i) they require to know which variable the user wants to "describe", i.e., the right side of the functional dependency and (ii) the right side must be composed of a single attribute, while the left side is generally made by several attributes. Furthermore, these approaches does not scale well with large tables with many attributes, while are able to manage table with many rows.

4. To overcome the issue of the join materialization, we decided to exploit the idea of an inverted index, in which the information of each table is stored, along with each column entropy and the column-table association. The peculiarity of this index is that it looks like a sparse matrix, since all the unique values in all the tables are indexed, and the entropy information for each column is stored only when i-th values belongs to j-th column. Figure 4.3 shows a sketch of the designed index.

5. We are currently working on the implementation of such an index, along with the research for understanding if mixing up the entropies is enough to simulate the conditional entropy behavior or a bit more elaborated metrics is required.

5 Research Plan

We distinguished our objectives in different classes, i.e., the objectives that are supposed to be completed in a short time and objectives which require more work to be completed. This distinction is reported in Table 1.

We aim at organizing our research in the three years according to eight activities, listed in Table 2. Each activity corresponds to a thesis objective or to the thesis writing. In particular, Table 2, for each activity, points out each sub-objective and the time units dedicated to it, while Table 3 presents their schedule in the three years.

Objective	**Long/Short Term Task**
1.1	Short Term Task
1.2	Short Term Task
2.1	Short Term Task
2.2	Long Term Task
3	Long Term Task
4	Always in Progress
5	Long Term Task

Table 1: Long/Short Term feasibility of each objectives presented in Section 4.

References

[1] Ziawasch Abedjan, Patrick Schulze, and Felix Naumann. "DFD: Efficient functional dependency discovery". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 2014, pp. 949–958.

[2] Patricia C Arocena, Boris Glavic, Radu Ciucanu, and Renée J Miller. "The iBench integration metadata generator". In: *Proceedings of the VLDB Endowment* 9.3 (2015), pp. 108–119.

Table 2: Activities

Activity	Description	Months	Colour
1	Analysis of the state of the art on open data, sketching and approximation of information theory measures	1st Year	
2	Definition of an index for tables augmentation in data lakes based on information-theoretic measures	1st Year	
3	Analysis of the state of the art approaches for knowledge bases construction, representation and augmentation	8	
4	Definition of a framework that exploits knowledge bases and possibility of including embedding to augment and Comparison	8	
5	Definition of an embedding algorithm for knowledge bases	6	
6	Continuative analysis of the state of the art approaches embeddings works	16	
7	Experimental Evaluation	10	
8	PhD thesis writing	4	

Table 3: Workplan

Activity	Year1	Year2	Year3
1			
2			
3			
4			
5			
6			
7			
8			

[3] C. Batini, M. Lenzerini, and S. B. Navathe. "A Comparative Analysis of Methodologies for Database Schema Integration". In: *ACM Comput. Surv.* 18.4 (1986), 323–364.

[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far". In: *Semantic services, interoperability and web applications: emerging concepts*. 2009, pp. 205–227.

[5] Thomas Bläsius, Tobias Friedrich, and Martin Schirneck. "The parameterized complexity of dependency detection in relational databases". In: *11th International Symposium on Parameterized and Exact Computation (IPEC 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017.

[6] Alex Bogatu, Alvaro AA Fernandes, Norman W Paton, and Nikolaos Konstantinou. "Dataset Discovery in Data Lakes". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020, pp. 709–720.

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating embeddings for modeling multi-relational data". In: *Advances in neural information processing systems*. 2013, pp. 2787–2795.

[8] Dan Brickley, Matthew Burgess, and Natasha Noy. "Google Dataset Search: Building a search engine for datasets in an open Web ecosystem". In: *The World Wide Web Conference*. 2019, pp. 1365–1375.

[9] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. "Creating embeddings of heterogeneous relational datasets for data integration tasks". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020, pp. 1335–1349.

[10] Barbara Catania, Giovanna Guerrini, and Beyza Yaman. "Exploiting Context and Quality for Linked Data Source Selection". In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019, 2251–2258.

[11] Surajit Chaudhuri, Venkatesh Ganti, and Raghav Kaushik. "A primitive operator for similarity joins in data cleaning". In: *22nd International Conference on Data Engineering (ICDE'06)*. IEEE. 2006.

[12] Nadiia Chepurko, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. "ARDA: Automatic Relational Data Augmentation for Machine Learning". In: *arXiv preprint arXiv:2003.09758* (2020).

[13] Dario Colazzo, François Goasdoué, Ioana Manolescu, and Alexandra Roatiş. "RDF analytics: lenses over semantic graphs". In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 467–478.

[14] Federico Dassereto, Laura Di Rocco, Giovanna Guerrini, and Michela Bertolotto. "Evaluating the effectiveness of embeddings in representing the structure of geospatial ontologies". In: *The Annual International Conference on Geographic Information Science*. 2019, pp. 41–57.

[15] Federico Dassereto, Laura Di Rocco, Shanley Shaw, Giovanna Guerrini, and Michela Bertolotto. "How to Tune Embedding Parameters in Geographical Ontologies". In: *Under Submission.*. 2020.

[16] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed Elmagarmid, Ihab F Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. "The data civilizer system". In: *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017*. 2017.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[18] Laura Di Rocco, Federico Dassereto, Michela Bertolotto, Davide Buscaldi, Barbara Catania, and Giovanna Guerrini. "Sherloc: a knowledge-driven algorithm for geolocating microblog messages at sub-city level". In: *International Journal of Geographical Information Science* (2020), pp. 1–32.

[19] Ronald Fagin, Laura M Haas, Mauricio Hernández, Renée J Miller, Lucian Popa, and Yannis Velegrakis. "Clio: Schema mapping creation and data exchange". In: *Conceptual modeling: foundations and applications*. 2009, pp. 198–236.

[20] Wenfei Fan, Yinghui Wu, and Jingbo Xu. "Functional dependencies for graphs". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 1843–1857.

[21] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. "Aurum: A data discovery system". In: *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE. 2018, pp. 1001–1012.

[22] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. "Hyperbolic entailment cones for learning hierarchical embeddings". In: *arXiv preprint arXiv:1804.01882* (2018).

[23] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. "Similarity search in high dimensions via hashing". In: *Vldb*. Vol. 99. 6. 1999, pp. 518–529.

[24] Palash Goyal and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey". In: *Knowledge-Based Systems* 151 (2018), pp. 78–94.

[25] Martin Grohe. "word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data". In: *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 2020, pp. 1–16.

[26] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. "Learning mixed-curvature representations in product spaces". In: *International Conference on Learning Representations*. 2018.

[27] Rihan Hai, Sandra Geisler, and Christoph Quix. "Constance: An intelligent data lake system". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 2097–2100.

[28] Bin He and Kevin Chen-Chuan Chang. "Statistical schema matching across web query interfaces". In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 2003, pp. 217–228.

[29] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. "Knowledge graph embedding based question answering". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 105–113.

[30] Yka Huhtala, Juha Kärkkäinen, Pasi Porkka, and Hannu Toivonen. "TANE: An efficient algorithm for discovering functional and approximate dependencies". In: *The computer journal* 42.2 (1999), pp. 100–111.

[31] Piotr Indyk and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 1998, pp. 604–613.

[32] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. "A survey on knowledge graphs: Representation, acquisition and applications". In: *arXiv preprint arXiv:2002.00388* (2020).

[33] Jaewoo Kang and Jeffrey F Naughton. "On schema matching with opaque column names and data values". In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 2003, pp. 205–216.

[34] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. "Hyperbolic geometry of complex networks". In: *Physical Review E* 82.3 (2010), p. 036106.

[35] Arun Kumar, Jeffrey Naughton, Jignesh M Patel, and Xiaojin Zhu. "To join or not to join? thinking twice about joins before feature selection". In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 19–34.

[36] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic web* 6.2 (2015), pp. 167–195.

[37] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. "Learning entity and relation embeddings for knowledge graph completion". In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[38] Stéphane Lopes, Jean-Marc Petit, and Lotfi Lakhal. "Efficient discovery of functional dependencies and armstrong relations". In: *International Conference on Extending Database Technology*. Springer. 2000, pp. 350–364.

[39] Panagiotis Mandros, Mario Boley, and Jilles Vreeken. "Discovering reliable approximate functional dependencies". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 355–363.

[40] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. "Big data: the management revolution". In: *Harvard business review* 90.10 (2012), pp. 60–68.

[41] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. "Spherical text embedding". In: *Advances in Neural Information Processing Systems*. 2019.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[43] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013).

[44] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. "Table union search on open data". In: *Proceedings of the VLDB Endowment* 11.7 (2018), pp. 813–825.

[45] Maximilian Nickel and Douwe Kiela. "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". In: *CoRR* abs/1806.03417 (2018).

[46] Maximillian Nickel and Douwe Kiela. "Poincaré embeddings for learning hierarchical representations". In: *Advances in neural information processing systems*. 2017, pp. 6338–6347.

[47] Maximillian Nickel and Douwe Kiela. "Poincaré Embeddings for Learning Hierarchical Representations". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. 2017, pp. 6338–6347.

[48] Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. "Data profiling with metanome". In: *Proceedings of the VLDB Endowment* 8.12 (2015), pp. 1860–1863.

[49] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic web* 8.3 (2017), pp. 489–508.

[50] Frédéric Pennerath, Panagiotis Mandros, and Jilles Vreeken. "Discovering Approximate Functional Dependencies using Smoothed Mutual Information". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1254–1264.

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[52] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[53] Erhard Rahm. "Towards large-scale schema and ontology matching". In: *Schema matching and mapping*. 2011, pp. 3–27.

[54] Andrea Rossi, Donatella Firmani, Antonio Matinata, Paolo Merialdo, and Denilson Barbosa. "Knowledge Graph Embedding for Link Prediction: A Comparative Analysis". In: *arXiv preprint arXiv:2002.00819* (2020).

[55] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. "Representation Tradeoffs for Hyperbolic Embeddings". In: *International Conference on Machine Learning*. 2018.

[56] Vraj Shah, Arun Kumar, and Xiaojin Zhu. "Are Key-Foreign Key Joins Safe to Avoid when Learning High-Capacity Classifiers?" In: *Proceedings of the VLDB Endowment* 11.3 (2017).

[57] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge". In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 697–706.

[58] Isuru Suriarachchi and Beth Plale. "Crossing analytics systems: A case for integrated provenance in data lakes". In: *2016 IEEE 12th International Conference on e-Science (e-Science)*. 2016, pp. 349–354.

[59] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. "Poincare Glove: Hyperbolic Word Embeddings". In: *International Conference on Learning Representations*. 2018.

[60] Coral Walker and Hassan Alrehamy. "Personal data lake with data gravity pull". In: *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*. 2015, pp. 160–167.

[61] Shyue-Liang Wang, Ju-Wen Shen, and Tzung-Pei Hong. "Incremental discovery of functional dependencies using partitions". In: *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. Vol. 3. 2001, pp. 1322–1326.

[62] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. "Knowledge graph embedding by translating on hyperplanes." In: *Aaai*. Vol. 14. 2014. 2014, pp. 1112–1119.

[63] Catharine Wyss, Chris Giannella, and Edward Robertson. "Fastfds: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances extended abstract". In: *International Conference on Data Warehousing and Knowledge Discovery*. 2001, pp. 101–110.

[64] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. "JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes". In: *Proceedings of the 2019 International Conference on Management of Data*. 2019, pp. 847–864.

[65] Erkang Zhu, Fatemeh Nargesian, Ken Q Pu, and Renée J Miller. "LSH ensemble: internet-scale domain search". In: *Proceedings of the VLDB Endowment* 9.12 (2016), pp. 1185–1196.