Thesis Proposal:

The Role of Embeddings in Data-Driven Augmentation

Federico Dassereto

October 2, 2020

Abstract

Maecenas ipsum velit, consectetuer eu, lobortis ut, dictum at, dui. In rutrum. Sed ac dolor sit amet purus malesuada congue. In laoreet, magna id viverra tincidunt, sem odio bibendum justo, vel imperdiet sapien wisi sed libero. Suspendisse sagittis ultrices augue. Mauris metus. Nunc dapibus tortor vel mi dapibus sollicitudin. Etiam posuere lacus quis dolor. Praesent id justo in neque elementum ultrices. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. In convallis. Fusce suscipit libero eget elit. Praesent vitae arcu tempor neque lacinia pretium. Morbi imperdiet, mauris ac auctor dictum, nisl ligula egestas nulla, et sollicitudin sem purus in lacus. Praesent in mauris eu tortor porttitor accumsan. Mauris suscipit, ligula sit amet pharetra semper, nibh ante cursus purus, vel sagittis velit mauris vel metus. Aenean fermentum risus id tortor. Integer imperdiet lectus quis justo. Integer tempor. Vivamus ac urna vel leo pretium faucibus. Mauris elementum mauris vitae tortor. In dapibus augue non sapien. Aliquam ante. Curabitur bibendum justo non orci.

Morbi a metus. Phasellus enim erat, vestibulum vel, aliquam a, posuere eu, velit. Nullam sapien sem, ornare ac, nonummy non, lobortis a, enim. Nunc tincidunt ante vitae massa. Duis ante orci, molestie vitae, vehicula venenatis, tincidunt ac, pede. Nulla accumsan, elit sit amet varius semper, nulla mauris mollis quam, tempor suscipit diam nulla vel leo. Etiam commodo dui eget wisi. Donec iaculis gravida nulla. Donec quis nibh at felis congue commodo. Etiam bibendum elit eget erat.

Maecenas ipsum velit, consectetuer eu, lobortis ut, dictum at, dui. In rutrum. Sed ac dolor sit amet purus malesuada congue. In laoreet, magna id viverra tincidunt, sem odio bibendum justo, vel imperdiet sapien wisi sed libero. Suspendisse sagittis ultrices augue. Mauris metus. Nunc dapibus tortor vel mi dapibus sollicitudin. Etiam posuere lacus quis dolor. Praesent id justo in neque elementum

1 Motivation and Description

Since the rise of the World Wide Web, we have experienced an 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 [25], where the key properties are **V**olume, **V**elocity, **V**ariety, **V**eracity and **V**alue.

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. The uncontrolled nature of data manifests in large repositories of data (data lakes), in which both structured and unstructured data are stored. The peculiarity of data lakes lies in the fact that there is uncertainty about the presence of metadata describing the data themself. Furthermore, 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. 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.

Following the path traced by Tim Berners-Lee et al. in [2] regarding Open Data publication and maintenance, many organizations are publishing data, augmentating 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. At a processing level, the main difficult 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 [27] and graphs [31]. In the last year, an approach exploting embedding to solve data integration task [5] has been published. This work can be seen as forerunner for the integration of embeddings in data management tasks, even if it works on tables from the same context.

A data scientist who is trying to solve machine learning tasks, frequently found herself in the 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).

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 augmentation problem itself nor with the usage of embeddings. Few approaches that try to augment tables with respect to 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 [23], 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 [8].

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 and index for horizontal augmentation on the data lakes scenario, as well as the integration of embeddings in the augmentation of KBs. Eventually, an embedding algorithm for KBs will summarize unified framework the two previous results. Our claim is that a key ingredient to reach 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 [15], 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). We remand to Section 3.4 for further discussion on functional dependencies.

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 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 [9] and then its impact as query engine into a data-driven microblogs geolocation algorithm [13]. We further analyze the tuning of parameters to obtain higher quality embedding and summarized our results in [10]; 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 [6] 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. We remand to Section 4 for the discussion on the preliminary results obtained regarding the augmentation problem.

2 Reference Area and Relevance of Goals

The current proposal lies in the reference area "Data Augmentation to improve Machine Learning Tasks", with the aim of providing to a data scientist an augmented set of data which enables her to improve her model performances. The approach we propose differ from other existing solutions for the following aspects:

- Existing approaches mainly focus on predicting the uselfulness of a join under the realtional hypothesis, i.e., the schema informations are known. To this end, we plan to develop an approach that is agnostic to the knowledge of the schemas;
- Existing approaches which try to augment data against a repository perform very expensive joins, and evaluate thier work with respect to 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 approaches that does not materialize any kind of join and is scalable to hundreds of thousands 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 to 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 treshold 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 present related work that we consider relevant for the proposed research project. In order to facilitate reading, the discussion has been organized into fuor main sections, corresponding to the main concepts introduced in Section 1 and 2. Each of the four sections is in turn divided into sub-parts to further facilitate reading. The first one discusses embeddings approaches and their relevance to the proposal, the second introduces Open Data and their common operations, the third provides an overview of existing augmentation approaches and finally the fourth reviews functional dependencies discovery approaches.

3.1 Embeddings

Embeddings allow 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., embedding 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 [28] is an example of a pre-trained embedding that embeds into Euclidean Space. Later, other embedding techniques like GloVe [33], BERT [12] and ELMo [34] 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 well capture a different kind of sources such as graphical structures [32].

In contrast, hyperbolic embeddings are particularly suitable to embed hierarchical data. As described in [22], hierarchical structures show a hidden hyperbolic geometry. On such observation, many hyperbolic embeddings have been proposed: the Poincaré Disk Model [32], the Lorentz Model [30] and a convex entailment cones approach [16]. 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 [18] 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, a common way of calculating the quality of embeddings was not 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 [36] and slightly modified in [9]. To the best of our knowledge, [9] 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.

A very effective algorithm called TransE was proposed in [3], in which embeddings are learned in translational way; this work generated many variants such as TransH [38] and TransR [24]. 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 [37] and a learning model in a spherical space [26].

As mentioned in Section 1, a preliminary approach exploting embeddings for data integration tasks have been proposed [5]. 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 random walks framework. The drawback of such an approach is that it is know thle context from which each table comes from, making it unsuitable for our scenario. AGGIUNGI Keynote PODS qua o sopra in motivations(meglio)

3.2 Open Data

The growth of Open Data diffusion has been possible thanks to fact that many organizations publish their data following the Open Data principles [2]. As discussed in Section 1, these data are continuosly produce and stored; their dynamic nature makes it impossible to apply previous managment techniques, such as the creation of a global schema [1] and to keeping track of joins path known or mined from the data[14, 11]. 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 the fact are intractable 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 definitely the discovery part, which include searching for tables containing specific value(s) based on keywords [4] or contaiment. Reconnecting to what we discussed in Section 1, we further indentify 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 attributes values are sets and a similarity function determines the relatedness. A frequent similarity function is the Jaccard Index, which suffers to problem of unfairly advantages small domains [40]. Many solutions have been proposed exploting 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 possibily with millions of rows. To solve this issue, a new similarity measure called containment have been proposed in LSH Ensemble [40], where an index based on locality-sensitive hashing [17] and MinHash [20] 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 often in data lakes). Furthermore, using a threshold may confuse users, who have no knowledge of what data exists in the lake and therefore do not know what is a good 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 contaiment. Another effective algorithm for join search is JOSIE [39], 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 [7], 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) [19, 35]. 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 [21]. 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 [29] 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.

3.3 Augmentation Approaches

3.4 Functional Dependencies

Functional dependencies (FDs) express relationships between attributes of a database relation. An FD $X \rightarrow A$ states that the values of attribute set X *uniquely* determine the values of attribute A.

4 Goals and Preliminaries

GOALS

- AGGIUNGI Embedding per ricerca in funzione di ML
- Studiare l'augmentation in termini di cosa esiste (entropia, FDs)
- Possibilità di augmenting verticale
- Introduzione delle KBs

5 Research Plan

Acknowledgements

So long, and thanks for all the fish.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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).

[9] 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.

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

[11] 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.

[12] 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).

[13] 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.

[14] 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.

[15] 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.

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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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).

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

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

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

[38] 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.

[39] 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.

[40] 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.