



ENTITY/RELATIONSHIP PROFILING

Whitepaper

SUMMARY

We introduce Entity/Relationship Profiling as the discovery-oriented, data-driven counterpart to classical Entity/Relationship Modeling. The latter offers a top-down enterprise view how data should be organized and managed, with support for alignment, compliance, and governance. E/R Profiling offers a bottom-up, data-centric view for the discovery of business entities and their relationships from data, with insight into data drifts, stories and tendencies. For the ultimate aim of data-driven decision making, data professionals ought to strive for an equilibrium between enterprise models and data assets. In this whitepaper, we provide a brief survey on academic and commercial data profilers, explain the technical features required for E/R Profiling in practice, and the benefits of E/R Profiling for tasks that data professionals need to complete routinely.

1. BACKGROUND

Entity/Relationship (E/R) Modeling captures business requirements, available in natural language, as Entity/Relationship models that are used to roll out data models in support of data-driven decision making. Indeed, E/R models provide an intuitive framework that resembles elements of natural language in the form of modeling features, yet is formal and tailored enough for translation into logical and physical data models. In the same way stories are assembled from entities and relationships between them, insight that is derived from data are stories about business entities and exciting relationships between them. Gaining a competitive advantage from such insight is what organizations often describe as finding diamonds in the dirt, a needle in haystacks, hidden treasure, or spinning straw into gold.

Classical E/R Modeling is a top-down approach that sets up a data infrastructure supporting business goals. However, data is now available in volume, variety, velocity, etc. Organizations need to discover what stories data assets hold for them and their customers. These bring up challenges like detecting the relevance of data sets, connecting data silos, integrating new data sources, or simply getting to know data (KYD). What insight stories does your data hold? Do your data assets resemble the goals of your enterprise? How is drift in data and business goals detected? In embracing these challenges and seeking answers to these questions, enterprises need to know how business entities, such as products or customers, and relationships between them, such as sales or acquisition, can be identified uniquely. This calls for Entity/Relationship Profiling, which we define as the task of computing all ways of uniquely identifying all business entities and their relationships from given data sources. Interestingly, there is a big divide between academic knowledge and commercial tools that address E/R Profiling.

Technically speaking, E/R Profiling has two computational problems at its core: 1) The discovery of uniqueness constraints from data sets, and 2) The discovery of inclusion dependencies from data sets. The decision variants of these problems are $W[2]$ - and $W[3]$ -complete, respectively, both in the number of columns. Hence, column scalability has strong boundaries (that is, the efficiency of discovery is very unlikely even when the number of columns is fixed). Despite these computational boundaries, excellent progress on each of these problems has been made over the last 35 years of research. Many algorithms have been brought forward that return rich results within reasonable time on rather challenging data sets.

In commercial tools, discovery is reduced to the analysis of given keys, functional dependencies, or foreign keys. That is, SQL queries simply return the degree by which a given constraint from these classes holds on the given data set. Some other tools restrict the key discovery problem to syntactic searches of single fields that have the letters "ID" in them, or validate for each single field its ratio of unique values. Similarly, some tools can discover unary inclusion dependencies (those with a single field). Hence, academic knowledge on the discovery problem is not accessible to public.

We introduce E/R Profiling as the discovery-oriented, data-driven complement to classical E/R Modeling, and highlight requirements for enabling E/R Profiling in practice, as well as the benefits of E/R Profiling for fundamental data tasks. Indeed, E/R Profiling maximizes automation and human-computer engagement in finding business keys and foreign keys, and therefore minimizes resources required while solving tasks critical for various data professionals, such as (i) amplify the accuracy and transparency of reports that analysts communicate, (ii) optimize the fit of logical data models that architects create, (iii) boost the effectiveness of data pipelines that engineers build, (iv) surge the effectiveness of feature stores that scientists manage, and (v) magnify data linkage and insight from catalogs that stewards maintain.

2. ACADEMIC AND COMMERCIAL DATA PROFILERS

The need for E/R Profiling and collaboration between researchers and practitioners becomes apparent when we chart the current landscape of academic prototypes and commercial profilers.

E/R Profiling belongs to data profiling, for which academic surveys are available such as [1,2], including latest trends [3].

Data profiling is divided into data, structure and relationship discovery. Hence, E/R Profiling is focused on the latter task. Academic research on the discovery problem has started in the

1980s, and continued until now. In particular, a recent paper has established that decision variants for the discovery of keys and functional dependencies is $W[2]$ -complete in the arity of the given constraints, hence very likely to be intractable even when the focus is on discovering any key or functional dependency with a given number of columns [4]. For inclusion dependencies, the decision variant of discovery is even $W[3]$ -complete in the arity [4]. Despite these natural boundaries on column efficiency, algorithms have been brought forward that solve the discovery problem on large data sets [5,6,7,8,9,10].

Other directions in relationship discovery include incremental, distributed and relaxed profiling [11,12,13,14,15,16, 17,18,19,20,21]. As one may expect, academic research has had a strong focus on improving the efficiency of discovery. However, analyses of the results, their impact and use cases are usually not discussed, and algorithms are not publicly accessible in most cases. The [Metanome](#) tool represents a comprehensive prototype system that gives uniform access to discovery algorithms developed in academia for various classes of constraints [6]. Compared to data profiling tools from industry, academic tools like Metanome are far ahead in terms of addressing the discovery problem, but at the expense of engaging users with the interplay of constraints and data. Perhaps an exception is the prototype system [DataProf](#) [21] which provides concise data summaries to enable the iterative discovery of business rules and dirty data. However, it is only implemented for a particular class of uniqueness constraints.

There is an entire industry of commercial data profilers available. Most have a strong focus on data and structure discovery, and strong limitations on relationship discovery. In contrast to academic profilers, they engage users with the data but at the expense of discovering relationships. Examples of open-source and commercial profilers include the [Aggregate Profiler](#), [Atlas](#), the [Melissa Data Profiler](#), [SAS DataFlux](#) or [TIBCO Clarity](#). Other data profilers are part of software packages and address the *analysis problem* for keys, functional dependencies, and foreign keys. Analysis simply means that the user selects a candidate constraint, and the tool computes to which degree the candidate holds on the given data source. Compared to discovery, analysis is a simple problem. For instance, [Microsoft DOCS](#) applies the analysis to user-specified keys or foreign keys, and [SAP BODS](#) requires candidate columns for key analysis, and pairs of matching columns across tables for foreign key analysis. In [Talend Open Studio](#), users need to select key and foreign key columns for analysis.

The [IBM InfoSphere Information Analyzer](#) enables the discovery of keys and foreign keys, restricted to those with a single column (that is, unary keys and unary foreign keys). This is a strong limitation. The other tool is the [Informatica Data Explorer](#). Users can discover primary key candidates by specifying the maximum arity of any candidate key that will be returned, a threshold by which any candidate key shall hold, and the maximum number of records used for mining, with a default value of 1000 only. The latter is a strong limitation. There is no distinction between candidates for primary keys and unique constraints, and no support for validation on the entire data set. Foreign key discovery is limited to a syntactic search of primary key fields in other tables that have matching domains. Hence, the discovery is not based on the underlying data or any inclusion thresholds as defined by the SQL standard.

In summary, academic and commercial tools have different focus points: the former has strong capabilities in discovering constraints at the expense of user engagement, while the latter facilitates engagement of users with the data at the expense of discovery.

The purpose of this whitepaper is to show the benefits of combining the best of both worlds, which can be illustrated by our [DataViadotto Entity/Relationship Profiler](#). At the same time, we call both researchers and practitioners to action on more blue sky research and knowledge transfer into tools.

3. FEATURES OF ENTITY/RELATIONSHIP PROFILING

The combination of features that make E/R Profiling possible in practice can be summarized under the acronym DESMIDS: **D**ata-driven, **E**xpressive, **S**ample-based, **M**etric, **I**ndustry-compliant **D**iscovery at **S**cale. Desmids are green algae, see Fig.2. Their presence is an indication of high-quality water. Similarly, the presence of the DataViadotto Profiler facilitates high entity and high referential integrity across all your data assets. These form the foundation for all data processing and analytical tasks.



Fig. 2: Desmids

So what do these features mean and why are they unique across profiling tools?

3.1. Data-driven

Our tool derives its output exclusively from underlying data, in contrast to other profilers who simply use constraints specified from the database schema or the names of fields. As mentioned before, current data profilers focus on the analysis of given keys and foreign keys, rather than their discovery. Hence, there is little support for data professionals in linking existing data sets, integrating new data sets, or understanding the relevance of data assets for projects. Because of that inability, the relevance of data sets remains hidden, data sets remain isolated, and insight remains undiscovered. In turn, current profiling tools fail to inform data professionals about the power data sets hold for the project at hand. In contrast, data-driven profilers pro-actively discover how business entities and their relationships can be identified uniquely across data sets, such as the key/foreign key relationship illustrated in Fig. 3.

Team's Team	Team's Team	Team's Team	Team's Team	Team's Team	Team's Team	Team's Team	Team's Team	Team's Team
Year	Lg ID	Tm ID	Opp ID	W	T	L	OTL	
1909	NHA	COB	HAI	1	1	0	null	
1909	NHA	COB	LES					
1909	NHA	COB	MOS					
1909	NHA	COB	MOW					
1909	NHA	COB	OT1					
1909	NHA	COB	REN					

Year	Lg ID	Tm ID	Franch ID	Co	H ID	Dw ID	Rank
1909	NHA	COB	BKN	null	null	null	4
1909	NHA	HAI	MTL	null	null	null	5
1909	NHA	LES	TBS	null	null	null	7
1909	NHA	MOS	MOS	null	null	null	6
1909	NHA	MOW	MTW	null	null	null	1
1909	NHA	OT1	STE	null	null	null	2
1909	NHA	REN	REN	null	null	null	3
1910	NHA	MOC	MTL	null	null	null	2
1910	NHA	MOW	MTW	null	null	null	4

Fig. 3: A key/foreign key relationship for a Hockey data set

3.2. Expressive

Our tool is **expressive**, meaning that it discovers all single-column and multi-column unique constraints and inclusion dependencies, including foreign keys, that hold on given data sets. In contrast, other tools are single-column profilers which can simply say how many values in a single column are unique and how many missing values occur in the column. This prevents current profilers from recommending different ways to uniquely identify business objects and their relationships. Our tool does not have that restriction. In fact, we even profile the class of *certain keys* [14], which are multi-column candidate keys that even permit missing data without losing the ability to uniquely identify all records.

3.3. Samples

Our tool provides smart data **samples**, meaning that users can inspect carefully chosen samples of data that show why a uniqueness constraint or relationship holds or does not hold, as illustrated in Fig. 4. Such

samples help users decide whether a constraint represents a rule important to their business. Even more, when a business rule is violated, then the offending data is dirty by definition (since the data does not conform to a business rule). As a consequence, our tool helps users understand by which rules their data is truly governed and identify data that violate entity or referential integrity. Furthermore, the engagement with samples allows users realize which different artificial identifiers refer to the same business object (entity resolution) and which data elements are duplicated.

Hockey.TeamVsTeam [year,oppID] -> Hockey.Teams [year,tmID]

Hockey.TeamVsTeam								Hockey.Teams						
year	lgID	tmID	oppID	W	L	T	OTL	year	lgID	tmID	franchID	confID	divID	rank
1969	NHL	CHI	STL	4	2	0	null	1969	NHL	STL	STL	null	WD	1
1969	NHL	MNS	STL	2	4	2	null	1977	WHA	CIN	CIN	null	null	7
1977	WHA	NEW	CIN	8	3	0	null	1980	NHL	COR	NJD	CC	SM	5
1980	NHL	CAL	COR	1	3	0	null							

Fig. 4: Data samples supporting the validity of a mined relationship

3.4. Metric

Our tool is driven by various **metrics**, enabling users to set thresholds by which keys and relationships must hold for them to be found. For example, a meaningful key or foreign key may be violated due to dirty data. Tools that only discover constraints that hold on the data would not be able to find them. In contrast, with the *DataViadotto Profiler* users can discover business keys and foreign keys even when they do not hold on the data set. The metrics also make it possible to rank all discovered keys and relationships accordingly. For example, uniqueness constraints may rank higher than others whenever they have fewer missing data in them. Likewise, we may rank foreign keys based on the ratio of records that have existing references. In fact, users can choose how to rank the output of our discovery algorithms, making it easier for them to identify those relationships that matter more to them.

3.5. Industry-compliant

Our tool is **industry-compliant**, meaning that our results conform fully to the definition of SQL semantics of missing data. Unique constraints enable the unique identification of all records with no missing data on any columns of the constraint, while foreign keys may apply simple, partial, or full SQL semantics for missing data. Our tool enables the discovery for all of these semantics. This ensures that entity and referential integrity is measured at the level data specialists deem appropriate. Hence, the impact of missing data can be analysed and controlled as required by the application an organisation targets. Other profilers do not take into account missing data other than reporting how many values in a column are missing. This may lead to poor entity and referential integrity, and to incorrect analysis and reporting. In contrast, our tool even allows users to specify which values they want our tool to regard as missing. This is important as missing data is often disguised as actual data when entered, such as the numerical value 0 or a string such as 01-01-1900.

3.6. Discovery

Our tool solves the **discovery** problem, as already described in the leading paragraphs. It is not surprising that current profilers cannot solve this problem due to its computational complexity. Our tool is built on the latest academic research in this area, including the state-of-the-art algorithms we continue to develop ourselves. For instance, Fig. 5 shows two relationships discovered by our tool, including various measures that help assess the relevance of each relationship. In this case, both relationships are many-to-one key-foreign key relationships that can be materialized with an inner join. This measures are automatically discovered from the underlying data.



Source table ↑	Target table	Source columns	Target columns	① Inclusion (simple)	① Inclusion (partial)	① Inclusion (full)	① Coverage	① Max cardinality	① Uniqueness	① Join type
Hockey.TeamVsTeam	Hockey.Teams	year, tmlID	year, tmlID	100%	100%	100%	100%	29	100%	*  1
Hockey.TeamVsTeam	Hockey.Teams	year, oppID	year, tmlID	100%	100%	100%	100%	29	100%	*  1

Fig. 5: Discovered Relationships, Measures for their Degrees of Validity, and Best Available Join Type

3.7. At Scale

Our tool is **scalable** within the boundaries of the underlying computational problem. For example, we employ smart data sampling strategies that return results as accurately as possible and scale to any number of records. While few unique constraints and relationships use more than three columns, our tool enables users to specify the maximum number of columns in unique constraints and relationships that our algorithms should mine. Our

algorithms only return results without redundancy in them, in order to reduce the time required to inspect them. For example, all unique constraints returned are minimal. Hence, removing any field from a unique constraint that has been discovered means that records exist with duplicate values on all remaining fields, and this is presented in our samples. In addition, our tool promotes the discovery of unique constraints before the discovery of relationships between them. Indeed, making the results of key discovery available as input for relationship discovery will tremendously cut down on the time required to discover all relationships that reference these unique constraints. This is not just a facilitator for efficiency but represents the actual process of discovery as it should happen: first we find all ways in which business entities can be uniquely identified, and then we find all relationships between them, that is Entity/Relationship Profiling. However, the DataViadotto Profiler can also find relationships not restricted to foreign keys alone, such as inclusion dependencies.

4. Benefits of Entity/Relationship Profiling

It is deliberating to realize that any actionable insight derived from data ought to be a narrative assembled from the business entities (keys) and relationships (foreign keys) that govern the data. For data professionals with a background in data modeling, this is no surprise: Peter Chen made this explicit in his Entity-Relationship Model from the mid 1970s. However, the rise of the data-era now demands for tools that facilitate *Entity-Relationship Profiling* to find an equilibrium between the data gathered and the models built to enable their analysis.

Indeed, knowledge about keys and relationships is central for all data processing and analytical tasks. For example, direct benefits of using keys and foreign keys include the following.

1. Organize data effectively and efficiently by
 - a. Pathways to critical data elements
 - b. Points of reference for joins
 - c. Integrating data assets using matching fields across tables
 - d. Despite missing and inconsistent data
2. Better data quality by
 - a. Enforcing entity integrity with keys
 - b. Enforcing referential integrity with foreign keys
 - c. Finding all ids used for the same entity
 - d. Imputing missing data
 - e. Removing data redundancy with foreign keys
 - f. Avoiding sources of inconsistent data
3. Higher performance by
 - a. Faster access using unique indexes
 - b. Optimising join types
 - c. More trust in data with business keys and key/foreign keys
 - d. Actual insight from accurate reports
 - e. Less biased, more explainable, higher quality predictive analytics

Indeed, the *DataViadotto Profiler* maximizes automation in finding keys and foreign keys, and therefore minimizes resources required while solving tasks critical for various data professionals, such as:

- Amplify the accuracy and transparency of reports that data analysts communicate,
- Optimizing the fit of logical data models that data architects create,
- Boost the effectiveness of data pipelines that data engineers build,
- Surge the effectiveness of feature stores that data scientists manage, and
- Magnify data linkage and insight from data catalogues data stewards maintain.

5. CLOSING

In summary, the *DataViadotto Profiler* is the first commercial tool that makes Entity/Relationship Profiling possible in practice. Indeed, the unique combination of our features make data profitable and lift any data-related role to new levels of effectiveness and efficiency. As an employer you will enable staff to understand data better and faster, make the most of your data assets, and bring data-driven decision making to life.

References

- [1] Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock. 2018. Data Profiling. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00878ED1V01Y201810DTM052>
- [2] Jixue Liu, Jiuyong Li, Chengfei Liu, and Yongfeng Chen. 2012. Discover Dependencies from Data - A Review. *IEEE Trans. Knowl. Data Eng.* 24, 2 (2012), 251–264. <https://doi.org/10.1109/TKDE.2010.197>
- [3] Júlia Colleoni Couto, Juliana Damasio, Rafael Bordini, and Duncan Ruiz. 2022. New Trends in Big Data Profiling. In *Intelligent Computing*, Kohei Arai (Ed.). Springer International Publishing, Cham, 808–825. https://doi.org/10.1007/978-3-031-10461-9_55
- [4] Thomas Bläsius, Tobias Friedrich, and Martin Schirneck. 2022. The complexity of dependency detection and discovery in relational databases. *Theor. Comput. Sci.* 900 (2022), 79–96. <https://doi.org/10.1016/j.tcs.2021.11.020>
- [5] Heikki Mannila and Kari-Jouko Räihä. 1987. Dependency Inference. In *VLDB'87, Proceedings of 13th International Conference on Very Large Data Bases, September 1-4, 1987, Brighton, England*. 155–158. <http://www.vldb.org/conf/1987/P155.PDF>
- [6] Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. 2015. Data Profiling with Metanome. *Proc. VLDB Endow.* 8, 12 (2015), 1860–1863. <https://doi.org/10.14778/2824032.2824086>
- [7] Falco Dürsch, Axel Stebner, Fabian Windheuser, Maxi Fischer, Tim Friedrich, Nils Strelow, Tobias Bleifuß, Hazar Harmouch, Lan Jiang, Thorsten Papenbrock, and Felix Naumann. 2019. Inclusion Dependency Discovery: An Experimental Evaluation of Thirteen Algorithms. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. 219–228. <https://doi.org/10.1145/3357384.3357916>

- [8] Lan Jiang and Felix Naumann. 2020. Holistic primary key and foreign key detection. *J. Intell. Inf. Syst.* 54, 3 (2020), 439–461. <https://doi.org/10.1007/s10844-019-00562-z>
- [9] Martti Kantola, Heikki Mannila, Kari-Jouko Räihä, and Harri Siirtola. 1992. Discovering functional and inclusion dependencies in relational databases. *Int. J. Intell. Syst.* 7, 7 (1992), 591–607. <https://doi.org/10.1002/int.4550070703>
- [10] Ziheng Wei and Sebastian Link. 2019. Discovery and Ranking of Functional Dependencies. In 35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019. 1526–1537. <https://doi.org/10.1109/ICDE.2019.00137>
- [11] Loredana Caruccio and Stefano Cirillo. 2020. Incremental Discovery of Imprecise Functional Dependencies. *ACM J. Data Inf. Qual.* 12, 4 (2020), 19:1–19:25. <https://doi.org/10.1145/3397462>
- [12] Loredana Caruccio, Vincenzo Deufemia, Felix Naumann, and Giuseppe Polese. 2021. Discovering Relaxed Functional Dependencies Based on Multi-Attribute Dominance. *IEEE Trans. Knowl. Data Eng.* 33, 9 (2021), 3212–3228. <https://doi.org/10.1109/TKDE.2020.2967722>
- [13] Ihab F. Ilyas, Volker Markl, Peter J. Haas, Paul Brown, and Ashraf Aboulnaga. 2004. CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies. In SIGMOD. 647–658. <https://doi.org/10.1145/1007568.1007641>
- [14] Henning Köhler, Uwe Leck, Sebastian Link, and Xiaofang Zhou. 2016. Possible and certain keys for SQL. *VLDB J.* 25, 4 (2016), 571–596. <https://doi.org/10.1007/s00778-016-0430-9>
- [15] Henning Köhler, Sebastian Link, and Xiaofang Zhou. 2016. Discovering Meaningful Certain Keys from Incomplete and Inconsistent Relations. *IEEE Data Eng. Bull.* 39, 2 (2016), 21–37. <http://sites.computer.org/debull/A16june/p21.pdf>
- [16] Jyrki Kivinen and Heikki Mannila. 1992. Approximate Dependency Inference from Relations. In Database Theory - ICDT'92, 4th International Conference, Berlin, Germany, October 14-16, 1992, Proceedings. 86–98. https://doi.org/10.1007/3-540-56039-4_34
- [17] Ester Livshits, Alireza Heidari, Ihab F. Ilyas, and Benny Kimelfeld. 2020. Approximate Denial Constraints. *Proc. VLDB Endow.* 13, 10 (2020), 1682–1695. <https://doi.org/10.14778/3401960.3401966>
- [18] Eduardo H. M. Pena, Eduardo C. de Almeida, and Felix Naumann. 2019. Discovery of Approximate (and Exact) Denial Constraints. *Proc. VLDB Endow.* 13, 3 (2019), 266–278. <https://doi.org/10.14778/3368289.3368293>
- [19] Hemant Saxena, Lukasz Golab, and Ihab F. Ilyas. 2019. Distributed Implementations of Dependency Discovery Algorithms. *Proc. VLDB Endow.* 12, 11 (2019), 1624–1636. <https://doi.org/10.14778/3342263.3342638>
- [20] Hemant Saxena, Lukasz Golab, and Ihab F. Ilyas. 2019. Distributed Discovery of Functional Dependencies. In 35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019. 1590–1593. <https://doi.org/10.1109/ICDE.2019.00149>
- [21] Renjie Xiao, Yong'an Yuan, Zijing Tan, Shuai Ma, and Wei Wang. 2022. Dynamic Functional Dependency Discovery with Dynamic Hitting Set Enumeration. In 38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022. 286–298. <https://doi.org/10.1109/ICDE53745.2022.00026>
- [22] Ziheng Wei and Sebastian Link. 2018. DataProf: Semantic Profiling for Iterative Data Cleansing and Business Rule Acquisition. In Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018. 1793–1796. <https://doi.org/10.1145/3183713.3193544>

ABOUT DATAVIADOTTO

DataViadotto is the industry pioneer for Entity/Relationship profiling technology. The company draws on decades of academic research in the subject to make the process of discovering models from data more effective, efficient and intuitive. Ultimately, data becomes profitable.

FOR ADDITIONAL QUESTIONS, CONTACT DATAVIADOTTO

www.viadotto.tech