Machine Learning to Data Management: A Round Trip

Laure Berti-Équille, Angela Bonifati, Tova Milo

Machine Learning to Data Management: A Round Trip

Laure Berti-Equille [1], Angela Bonifati [2], Tova Milo [3]

[1] *Aix-Marseille Université, LIS (CNRS), France*
laure.berti-equille@univ-amu.fr
[2] *Université Claude Bernard Lyon 1, LIRIS (CNRS), France*
angela.bonifati@univ-lyon1.fr
[3] *Tel Aviv University, Tel Aviv, Israel*
milo@post.tau.ac.il

***Abstract*—With the emergence of machine learning (ML) techniques in database research, ML has already proved a tremendous potential to dramatically impact the foundations, algorithms, and models of several data management tasks, such as error detection, data cleaning, data integration, and query inference. Part of the data preparation, standardization, and cleaning processes, such as data matching and deduplication for instance, could be automated by making a ML model "learn" and predict the matches routinely. Data integration can also benefit from ML as the data to be integrated can be sampled and used to design the data integration algorithms. After the initial manual work to setup the labels, ML models can start learning from the new incoming data that are being submitted for standardization, integration, and cleaning. The more data supplied to the model, the better the ML algorithm can perform and deliver accurate results. Therefore, ML is more scalable compared to traditional and time-consuming approaches. Nevertheless, many ML algorithms require an out-of-the-box tuning and their parameters and scope are often not adapted to the problem at hand. To make an example, in cleaning and integration processes, the window sizes of values used for the ML models cannot be arbitrarily chosen and require an adaptation of the learning parameters. This tutorial will survey the recent trend of applying machine learning solutions to improve data management tasks and establish new paradigms to sharpen data error detection, cleaning, and integration at the data instance level, as well as at schema, system, and user levels.**

I. Detailed Outline of the Tutorial

The tutorial is organized into an introductory SWOT analysis (Section I-A), four main parts (described in detail from Section I-B to Section I-E), and final conclusions (Section I-F) as follows.

A. ML for Data Management: SWOT Analysis

The tutorial start with an introduction to the relevant concepts in Machine Learning from the data management perspective. We explore the use of ML techniques as a tool to express and quantify data patterns and knowledge transfer for representing and analyzing data, using examples from literature in database cleaning/repair, data integration, and query inference. We first provide an overview of the opportunities and limitations, alongside with the computational challenges associated with ML techniques applied to data management. We articulate the tutorial into four main parts related to the presented levels of data management: (1) instance-based; (2) schema-based; (3) system-based, and (4) user interaction-based data management.

B. ML in Instance-Based Data Transformation Tasks

In this part, we illustrate the role played by ML techniques in data processing at the data instance level, through ML applications to data cleaning, database repairing, and data fusion. This list of topics is by no means exhaustive albeit representative of the diversity of problems where ML tools have proved useful. We will review recent DB/ML research leveraging:

- Clustering applied to anomaly detection and data cleaning [16], [41], [49], detection of patterns of glitches [7], [8], replacement of erroneous or missing values, and deduplication [12];
- Classification applied to database repairing [44], [47], regression classification used in record linkage [27], and kNN for data fusion;
- Semi- and supervised learning models for similarity and blocking functions used in deduplication and record linkage [10], [11], and active learning in entity resolution [46];

- Bayesian analysis applied to data cleaning [17], probabilistic inference in data repairing [39], disambiguation (conflict resolution), and data fusion [20], [21], [35]; and
- Model optimization and statistical model training with guarantees [31] used in learning from samples for progressive data cleaning.

Finally, we will discuss the pros and cons of ML applications to instance-based data management tasks and we will also review some work in ML to address noisy data labels and model robustness and discuss its applicability to data quality and integration.

C. ML in Schema-Based Data Transformation Tasks

In this part, we focus on the application of ML techniques to schema-driven tasks in data management, such as schema and constraint inference, schema mapping and query specification. Again, our list is not meant to be exhaustive but aims at fostering the discussion on the usage of learning in all its facets in the above tasks. The presentation will include the following topics:

- Schema and schema mapping discovery techniques [2], [9], [29], [33] to fight database decay and facilitate data integration;
- Usage of input data examples to help the user specify complex data transformation tasks [15], [28];
- Usage of machine learning in data source reconciliation [18];
- Learning in rule discovery and information extraction [23], [26], [32], [40]; and
- Query specification paradigms leveraging grammar induction techniques [3], [13], [14].

We will review recent approaches for data transformation, schema, and constraint discovery that can benefit from learning based on input data examples. We will discuss the advantages and limitations of these techniques and pinpoint a few extensions.

D. ML in System-Oriented Data Management Tasks

In this part, we will discuss about a recent trend in our community to use ML techniques for obtaining "trained database systems", i.e., databases that can learn from past query workloads (or past query executions and optimized plans) the behavior to be adopted in upcoming querying and tuning tasks. We will put under lenses the many ML approaches for database query optimization and bulk data processing systems by highlighting their advantages and their possible impact on full-fledged database systems. We will (not exhaustively) discuss about the following trends in the DB community:

- Predicting query answering based on the past history of queries [34];
- Predicting the decisions of a query optimizer [38] and performing database tuning [43] by leveraging ML techniques;
- Feature engineering and labeling are bottlenecks in ML techniques that hinder their adoption in database-oriented tasks; we will review works to facilitate those tasks [1], [5], [36];
- Finally, we will focus on the connections between ML and Databases and the unsolved challenges in this area [24], [37].

E. ML in User-Guided Data Management Tasks

In this part, we will discuss the limitations of pure ML approaches and how users can help to complement the efforts. As illustrative examples, we will examine common data management tasks such as entity resolution and data cleaning [4], [6], [19], [25], [45], [48], [50]. We will consider the interplay between ML-based algorithms and crowd-sourcing, and highlight where users' input is essential. Specifically, we will discuss three dimensions of the problem:

- How users can help in improving the data itself, e.g., by detecting errors [42], gathering missing data and choosing among possible data repairs [4], [48];
- How they can assist in gathering meta-data that facilitates improved data processing [6], [45];
- How can we find and identify the most relevant crowd to complement the ML efforts in a given data management task [19], [50].

F. Lessons Learned and Perspectives

We foresee two major outcomes from this tutorial. In the short term, we expect that this tutorial will lead to a more effective use of ML techniques in data management applications. In the long term, we hope that understanding the benefits and limits of the application of ML to the modeling, representation, and analysis of data will lead to a better interaction between data management and ML when designing the next-generation database management systems.

II. Presenters' Biography

Laure Berti-Equille received her Ph.D. degree in Computer Science from University of Toulon in France in 1999. From 2000-2010, she was a tenured Associate Professor at University of Rennes 1, and a 2-years visiting researcher at AT&T Labs Research in New Jersey, USA, as a recipient of the prestigious European Marie

Curie Outgoing Fellowship (2007-2009). From 2011-2017, she joined IRD, the French Institute of Research for Development, as a Research Director. From 2014-2017, she was a Senior Scientist at Qatar Computing Research Institute (Hamad Bin Khalifa University). She is now is a full Professor at Aix-Marseille University (AMU) in France. Her interests are at the intersection of large-scale data science, data analytics, and machine learning with a focus on data quality and truth discovery research. She initiated the very first workshop editions on information and data quality in information systems (IQIS 2005) and in databases (QDB 2009 and 2016) in conjunction with SIGMOD and VLDB respectively, and co-organized the first French workshops on Data and Knowledge Quality in conjunction with EGC (Extraction et Gestion de Connaissances) in 2005, 2006, 2010, and 2011. Laure is serving as an associated editor of the ACM Journal on Data and Information Quality and served as a Program Chair of the International Conferences on Information Quality (ICIQ) in 2012 and 2016. She has received various grants from the French Agency for National Research (ANR), the French National Research Council (CNRS), and the European Union.

Angela Bonifati received her Ph.D. degree in Computer Science from Politecnico di Milano in 2002. After graduating she worked as a postdoctoral researcher at the INRIA research institute in Paris. She then obtained a permanent position as a researcher at the Italian National Research Council in 2003. She is now a full Professor in France (since 2011), currently at University of Lyon 1. Her research focuses on advanced database applications such as data integration and exchange, web and graph databases, query inference by considering both structured and semi-structured data models. She has been visiting professor in several foreign universities, such as Stanford University, UBC and Saarland University. Angela served as the Program Chair of several international conferences, including ICDE 2011 (Semi-structured data Track) and ICDE 2018 (Information Extraction and Data Cleaning and Curation Track), WebDB 2013, and XSym 2009. She is currently associate editor of the VLDB Journal, ACM Transactions on Database Systems (TODS) and Distributed and Parallel Databases. She has been the recipient of the prestigious Palse Impulsion Starting Grant at the University of Lyon (IDEX) in 2016. She has received grants from the French and Italian Ministry of Science and the French National Research Council (CNRS).

Tova Milo received her Ph.D. degree in Computer Science from the Hebrew University, Jerusalem, in 1992. After graduating she worked at the INRIA research institute in Paris and at University of Toronto and returned to Israel in 1995, joining the School of Computer Science at Tel Aviv university, where she is now a full Professor. She is the head of the Database research group and holds the Chair of Information Management. She served as the Head of the Computer Science Department from 2011-2014. Her research focuses on large-scale data management applications such as data integration, semi-structured information, Data-centered Business Processes and Crowd-sourcing, studying both theoretical and practical aspects. Tova served as the Program Chair of several international conferences, including PODS, VLDB, ICDT, XSym, and WebDB, and as the chair of the PODS Executive Committee. She served as a member of the VLDB Endowment and the PODS and ICDT executive boards and as an editor of TODS and the Logical Methods in Computer Science Journal. Tova has received grants from the Israel Science Foundation, the US-Israel Binational Science Foundation, the Israeli and French Ministry of Science and the European Union. She is an ACM Fellow, a member of Academia Europaea, a recipient of the 2010 ACM PODS Alberto O. Mendelzon Test-of-Time Award, the 2017 VLDB Women in Database Research award, the 2017 Weizmann award for Exact Sciences Research, and of the prestigious EU ERC Advanced Investigators grant.

REFERENCES

[1] M.R. Anderson, M.J. Cafarella. Input Selection for Fast Feature Engineering. ICDE 577-588, 2016.

[2] P. Andritsos, R.J. Miller, P. Tsaparas, Information-Theoretic Tools for Mining Database Structure from Large Data Sets. SIGMOD:731-742, 2004.

[3] T. Antonopoulos, F.k Neven, F. Servais. Definability Problems for Graph Query Languages. ICDT:141-152, 2013.

[4] A. Assadi, T. Milo, S. Novgorodov. DANCE: Data Cleaning with Constraints and Experts. ICDE:1409-1410, 2017.

[5] S.H. Bach, B. Dawei He, A. Ratner, C. R. Learning the Structure of Generative Models without Labeled Data. ICML:273-282, 2017.

[6] M. Bergman, T. Milo, S. Novgorodov, W. Chiew Tan. Query-Oriented Data Cleaning with Oracles. SIGMOD:1199-1214, 2015.

[7] L. Berti-Equille, T. Dasu, D. Srivastava. Discovery of complex glitch patterns: A novel approach to Quantitative Data Cleaning. ICDE 2011: 733-744.

[8] L. Berti-Equille, J.M. Loh, T. Dasu. A masking index for quantifying hidden glitches. Knowl. Inf. Syst. 44(2): 253-277, 2015.

[9] G.J. Bex, W. Gelade, F. Neven, S. Vansummeren. Learning Deterministic Regular Expressions for the Inference of Schemas from XML Data. WWW:825-834, 2008.

[10] I. Bhattacharya, L. Getoor. Entity Resolution. Encyclopedia of Machine Learning and Data Mining 2017:402-408, 2017.

[11] M. Bilenko, B. Kamath, R.J. Mooney. Adaptive Blocking: Learning to Scale Up Record Linkage. ICDM:87-96, 2006.

[12] M. Bilenko. Learnable Similarity Functions and their Applications to Clustering and Record Linkage. AAAI:981-982, 2004.

[13] A. Bonifati, R. Ciucanu, S. Staworko. Learning Join Queries from User Examples. ACM Trans. Database Syst.40(4): 24:1-24:38, 2016.

[14] A. Bonifati, R. Ciucanu, A. Lemay. Learning Path Queries on Graph Databases. EDBT:109-120, 2015.

[15] A. Bonifati, U. Comignani, E. Coquery, R. Thion. Interactive Schema Mapping Specification with Exemplar Tuples. SIGMOD:667-682, 2017.

[16] Y. Chung, S. Krishnan, T. Kraska. A Data Quality Metric (DQM): How to Estimate the Number of Undetected Errors in Data Sets. PVLDB 10(10):1094-1105, 2017.

[17] S. De, Y. Hu, V.V. Meduri, Y. Chen, S. Kambhampati. BayesWipe: A Scalable Probabilistic Framework for Improving Data Quality. ACM J. Data and Information Quality, 8(1), 5:1-5:30, 2016.

[18] A. Doan, P. Domingos, A.Y. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. SIGMOD:509-520, 2001.

[19] A. Doan, A. Ardalan, J. R. Ballard, S. Das, Y. Govind, P. Konda, H. Li, S. Mudgal, E. Paulson, P. Suganthan G. C., H. Zhang. Human-in-the-Loop Challenges for Entity Matching: A Midterm Report. HILDA@SIGMOD:12:1-12:6, 2017.

[20] X.L. Dong, L. Berti-Equille, D. Srivastava. Data fusion: resolving conflicts from multiple sources. Handbook of Data Quality, 293-318.

[21] X.L. Dong, L. Berti-Equille, D. Srivastava. Integrating Conflicting Data: The Role of Source Dependence. PVLDB:550-561, 2009.

[22] H. Fernau. Algorithms for Learning Regular Expressions from Positive Data. Inf. Comput. 207(4): 521-541, 2009.

[23] P.A. Flach, I. Savnik. Database Dependency Discovery: A Machine Learning Approach. AI Commun. 12(3):139-160, 1999.

[24] Frontiers in Massive Data Analysis. http://www.stat.berkeley.edu/~mmahoney/pubs/nrc-massive-data.pdf

[25] F. Geerts, G. Mecca, P. Papotti, D. Santoro. The LLUNATIC Data-Cleaning Framework. PVLDB, 6(9):625-636, 2013.

[26] J.M. Hellerstein, C. Ré, F. Schoppmann, D.Z. Wang, E. Fratkin, A. Gorajek, K.S. Ng, C. Welton, X. Feng,K. Li, A. Kumar. The MADlib Analytics Library or MAD Skills, the SQL. PVLDB, 5(12):1700-1711, 2012.

[27] Y. Hu, Q. Wang, D. Vatsalan, P. Christen. Improving Temporal Record Linkage Using Regression Classification. PAKDD:561-573, 2017.

[28] Z. Jin, M.R. Anderson, M.J. Cafarella, H. V. Jagadish. Foofah: Transforming Data By Example. SIGMOD:683-698, 2017.

[29] A. Kimmig, A. Memory, R.J. Miller, L. Getoor. A Collective Probabilistic Approach to Schema Mapping Discovery. ICDE:921-932, 2017.

[30] S. Krishnan, J.n Wang, M.J. Franklin, K. Goldberg, T. Kraska, T. Milo, E. Wu. SampleClean: Fast and Reliable Analytics on Dirty Data. IEEE Data Eng. Bull., 38(3):59-75, 2015.

[31] S. Krishnan, J. Wang, E. Wu, M.J. Franklin, K. Goldberg. ActiveClean: Interactive Data Cleaning For Statistical Modeling. VLDB:948-959, 2016.

[32] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, H.V. Jagadish. Regular Expression Learning for Information Extraction. EMNLP:21-30, 2008.

[33] R.J. Miller, P. Andritsos. Schema Discovery. IEEE Data Engineering Bulletin, 26(3):40-45, 2003.

[34] Y. Park, A. Shahab Tajik, M. Cafarella, B. Mozafari. Database Learning: Toward a Database that Becomes Smarter Every Time, SIGMOD:745-758, 2017.

[35] R. Pradhan, S. Bykau, S. Prabhakar. Staging User Feedback toward Rapid Conflict Resolution in Data Fusion. SIGMOD 2017.

[36] A. Ratner, S.H. Bach, H.R. Ehrenberg, J.A. Fries, S. Wu, C. R. Snorkel: A System for Lightweight Extraction. CIDR 2017.

[37] C. Ré, D. Agrawal, M. Balazinska, M.J. Cafarella, M.I. Jordan, T. Kraska, R. Ramakrishnan. Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype? SIGMOD:283-284, 2015.

[38] M. Schleich, D. Olteanu, R. Ciucanu. Learning Linear Regression Models over Factorized Joins. SIGMOD:3-18, 2016.

[39] T. Rekatsinas, X. Chu, I.F. Ilyas, C. R. HoloClean: Holistic Data Repairs with Probabilistic Inference. PVLDB, 10(11): 1190-1201, 2017.

[40] A. Rostin, O. Albrecht, J. Bauckmann, F. Naumann, U. Leser. A Machine Learning Approach to Foreign Key Discovery. WebDB@SIGMOD, 2009.

[41] S. Song, C. Li, and X. Zhang. Turn Waste into Wealth: On Simultaneous Clustering and Cleaning over Dirty Data. KDD:1115-1124, 2015.

[42] S. Thirumuruganathan, L. Berti-Equille, M. Ouzzani, J.-A. Quian-Ruiz, N. Tang. UGuide: User-Guided Discovery of FD-Detectable Errors. SIGMOD Conference 2017: 1385-1397.

[43] D. Van Aken, A. Pavlo, G. J. Gordon, B. Zhang.Automatic Database Management System Tuning Through Large-scale Machine Learning. SIGMOD:1009-1024, 2017.

[44] M. Volkovs, F. Chiang, J. Szlichta, R.J. Miller. Continuous Data Cleaning. ICDE:244-255, 2014.

[45] J. Wang, S. Krishnan, M.J. Franklin, K. Goldberg, T. Kraska, and T. Milo. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data.SIGMOD:469-480, 2014.

[46] S.E. Whang, D. Marmaros, H. Garcia-Molina. Pay-as-you-go Entity Resolution. TKDE:1111-1124, 2013.

[47] M. Yakout, L. Berti-Equille, A.K. Elmagarmid. Don't be SCAREd: Use SCalable Automatic REpairing with Maximal Likelihood and Bounded Changes. SIGMOD:553-564, 2013.

[48] M. Yakout, A.K. Elmagarmid, J. Neville, M. Ouzzani, I.F. Ilyas. Guided Data Repair, PVLDB, 4(5):279-289, 2011.

[49] A. Zhang, S. Song, J. Wang, P.S. Yu. Time Series Data Cleaning: From Anomaly Detection to Anomaly Repairing. PVLDB, 10(10):1046-1057, 2017.

[50] C.J. Zhang, Z. Zhao, L. Chen, H. V. Jagadish, C. C. Cao. Crowdmatcher: Crowd-Assisted Schema Matching. SIGMOD:721-724, 2014.