Winter 12-2019

# Understanding the Impact of Machine Learning on Enterprise Data Management: A Taxonomic Approach

Martin Fadler

Christine Legner

# Understanding the Impact of Machine Learning on Enterprise Data Management: A Taxonomic Approach

*Completed Research Paper*

**Martin Fadler**
Faculty of Business and Economics (HEC)
University of Lausanne
1015 Lausanne, Switzerland
martin.fadler@unil.ch

**Christine Legner**
Faculty of Business and Economics (HEC)
University of Lausanne
1015 Lausanne, Switzerland
christine.legner@unil.ch

## Abstract

Enterprise data management (EDM) is a critical success factor for leveraging the increasing data volume and variety but relies mostly on manual efforts and human intervention. While first studies show the significant potential of ML techniques, they have a strong technical focus and only address isolated problems. Against this backdrop, our study sheds light on how ML techniques can advance EDM. Based on the analysis of 43 ML cases, our study makes two main contributions: First, we suggest a taxonomy that links ML applications to concepts from EDM and data curation. Second, we identify nine archetypes that provide an overview of typical application areas of ML in EDM. We find that ML techniques induce a shift from manual data maintenance in a reactive mode to data creation in a proactive mode. Our analysis also reveals that some archetypes build on a rich body of research from the database community.

### *Keywords*

Enterprise Data Management, Machine Learning, Data Curation, Data Quality, Taxonomy

## Motivation and Research Objectives

As the digitalization of business advances, companies face an ever-increasing amount and variety of data generated by users and sensors. This creates new challenges for enterprise data management (EDM) (Abbasi et al. 2016; Sivarajah et al. 2017). A major concern of EDM has always been data quality (DQ), defined as data that is *"fit for purpose"* (Otto 2011; Wang 1998). While the simple principle of "garbage in garbage out" holds true for any form of data processing, it becomes even more critical as business will increasingly employ intelligent systems that learn from data (Redman 2018). Not surprisingly, top management considers DQ as a major concern when they think about implementing artificial intelligence (Pyle and José 2015), business analytics or self-service business intelligence (BARC 2018). However, the traditional EDM approaches are very manual and lack scalability to cope with increasing data volumes and variety (Stonebraker and Ilyas 2018). This is where machine learning (ML), and data-driven approaches come into play. Researchers (Zhu et al. 2014) and practitioners (Bean 2017) have proposed various usage scenarios of ML to improve DQ and to automate certain EDM tasks. The most prominent example is *Data Tamer*, a data curation system developed at MIT that entails ML has been able to reduce data curation costs in three real world examples by about 90% (Stonebraker et al 2013). While there is some evidence of the significant potential of ML techniques for EDM, the existing studies develop solutions that address very specific data problems. So far, there is neither an analysis of ML application areas in this domain, nor have the ML techniques been linked to the foundational concepts of EDM. Hence, we lack a systematic understanding of this dynamic and rapidly evolving field.

To address this gap, our overall research study objective is to understand *how machine learning techniques can support enterprise data management*. We therefore address two sub-questions:

***RQI:*** *What elements describe machine learning techniques for enterprise data management?*

***RQII:*** *Which archetypes of machine learning for enterprise data management can be distinguished?*

## Method

We follow the taxonomy development process suggested by Nickerson, Varshney, and Muntermann (2013), as it provides a concise method and is frequently applied for structuring emerging fields in the IS domain (Beinke et al. 2018; Püschel et al. 2016). As the first step, we determined the purpose of our taxonomy as describing and classifying ML techniques for EDM. This translates into two meta-characteristics: the EDM context, i.e. the specific situation in which the ML technique is applied, and the ML application, i.e. the concrete way of applying ML to support EDM. In the second step, we defined objective and subjective criteria as ending conditions to terminate the development process. In the third step, we conducted two iterations to develop our taxonomy using a combined deductive-inductive approach. In the first iteration (conceptual-to-empirical approach), we analyzed the literature on EDM, data curation and ML to collect the first set of dimensions and characteristics and base our taxonomy on categories which show a scientific rigor. In the second iteration (empirical-to-conceptual approach), we collected cases describing ML techniques for EDM from three different sources to ensure that all dimensions and characteristics are expressive enough to answer our research question. From the academic literature, we identified cases that applied ML techniques for data curation (17 cases). In parallel, we conducted five expert interviews and two focus groups with EDM experts (11 cases) and screened the market for innovative tools that offer ML techniques for EDM (15 cases). After review, our set consisted of 43 cases which we used for the taxonomy development. To answer our second research question, we analyzed the complete set of cases based on the taxonomy to derive archetype.

## Research Outcomes and Contributions

The outcome of this research is twofold: first, the taxonomy to classify ML techniques for EDM and second, the archetypes to describe typical application scenarios of ML in EDM.

The taxonomy of ML techniques in EDM comprises seven dimensions that are structured according to the meta-characteristics EDM context and ML application. The EDM context characterizes the specific situation in which the ML technique is applied to solve a data-related issue. We suggest three dimensions: the Data production process (as the high-level process), the Data curation task (as the specific activity performed to curate data) and the DQ impact (as the benefit from data curation). The ML application meta-characteristic comprises four dimensions that characterize the Data level, the Learning strategy, the ML method, and Human – ML interaction. We then used the taxonomy to classify 43 collected cases that apply ML for EDM and analyzed the resulting patterns.

Based on the classification we derived 9 archetypes as a description of a group of ML techniques that follow a certain pattern in at least one of the taxonomy's dimensions and can be viewed as a homogenous group. We ordered the archetypes according to the corresponding Data production process: acquire & create (three archetypes), unify & maintain (four archetypes), protect & retire (one archetype) and discover & use (one archetype).

There are several interesting insights and implications from our research: We find that ML supports both reactive and proactive EDM. Of particular interest is that ML techniques shifts manual data maintenance efforts that were formerly executed by data custodians (in a reactive mode) to the data collector (in a proactive mode). This implies that, with ML, EDM is becoming an integrated part of each business function, rather than being delegated to specialized data management units. While ML has significant potential, we find that in most cases ML does support by proposing a solution, but does not completely substitute the human activities. Although some part of the data-related tasks can be automated, human effort is in all cases still required. This observation creates manifold research opportunities related to the socio-technical design of data production processes that integrate augmentation or assemblage with ML techniques.

Our study makes two main contributions: First, the suggested taxonomy provides a classification scheme that links ML techniques to concepts from EDM and data curation. It thereby connects research streams, that address complementary organizational and technical aspects, but are currently not connected. Second, the archetypes provide an overview of typical application areas of ML in EDM. Our analysis reveals that

some archetypes build on a rich body of research that has developed over the last decades, for instance in the case of archetype 5 (entity matching). In addition, we also observe archetypes that open interesting new fields, such as archetype 9 that applies ML to support the discovery of data by users. Our findings are relevant for both research and practice: The taxonomy and archetypes support practitioners in selecting and assessing suitable ML techniques for their data problems. For research, the taxonomy helps positioning work and derive opportunities for further research. Our findings thereby create the groundwork for future research on advancing EDM practices with ML. We also contribute to DQ research: Based on (Zhu et al. 2014)'s framework, we structure the field of database-related technical solutions for DQ, using the methods of ML and data mining.

## References


Abbasi, A., Sarker, S., and Chiang, R. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems* (17:2). (http://aisel.aisnet.org/jais/vol17/iss2/3).

BARC. 2018. "Top Business Intelligence Trends 2019 | What 2,700 BI Professionals Think," *BI Survey*. (https://bi-survey.com/top-business-intelligence-trends, accessed November 27, 2018).

Bean, R. 2017. "How Big Data Is Empowering AI and Machine Learning at Scale," *MIT Sloan Management Review*. (https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/, accessed November 12, 2018).

Beinke, J. H., Nguyen, D., and Teuteberg, F. 2018. "Towards a Business Model Taxonomy of Startups in the Finance Sector Using Blockchain," in *Proceedings of International Conference of Information Systems*, San Francisco, CA, p. 9.

Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* (22:3), pp. 336–359. (https://doi.org/10.1057/ejis.2012.26).

Otto, B. 2011. "Data Governance," *Business & Information Systems Engineering* (3:4), pp. 241–244. (https://doi.org/10.1007/s12599-011-0162-8).

Püschel, L., Röglinger, M., and Schlott, H. 2016. "What's in a Smart Thing? Development of a Multi-Layer Taxonomy," in *Proceedings of International Conference of Information Systems*, Dublin, Ireland, p. 19.

Pyle, D., and José, C. S. 2015. "An Executive's Guide to Machine Learning | McKinsey." (https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning, accessed November 27, 2018).

Redman, T. C. 2018. "If Your Data Is Bad, Your Machine Learning Tools Are Useless," *Harvard Business Review*, p. 5.

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. 2017. "Critical Analysis of Big Data Challenges and Analytical Methods," *Journal of Business Research* (70), pp. 263–286. (https://doi.org/10.1016/j.jbusres.2016.08.001).

Stonebraker, M., Bruckner, D., and Ilyas, I. F. 2013. "Data Curation at Scale: The Data Tamer System," in *CIDR Proceedings 2013*.

Stonebraker, M., and Ilyas, I. F. 2018. "Data Integration: The Current Status and the Way Forward," *IEEE Technical Committee on Data Engineering*, Data Engineering Bulletin, p. 7.

Wang, R. Y. 1998. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), pp. 58–65.

Zhu, H., Madnick, S., Lee, Y., and Wang, R. 2014. "Data and Information Quality Research: Its Evolution and Future," in *Computing Handbook, Third Edition*, H. Topi and A. Tucker (eds.), Chapman and Hall/CRC, pp. 16-1-16–20. (https://doi.org/10.1201/b16768-20).