AI for data quality management

Mattia Ferrini
is a director at KPMG Switzerland.

Francesco Guidi
is a manager at KPMG Switzerland.

The importance of data quality in analytics and AI projects is well recognized. But what can AI do for data quality management?

Every day seems to deliver new solutions and promises in data analytics, data science and artificial intelligence. Before investing in a shiny new emerging technology, however, ask yourself one question: is my data quality fit for purpose? Unfortunately, many organizations have low trust in their own data quality, which paralyzes the investment in data-driven solutions. How can organizations efficiently achieve their data quality goals? How can AI improve data quality management processes?

What is data quality? In simple terms, data has the desired quality if it is fit for use and meets the requirements set by its users. Data quality can be classified along seven different dimensions: completeness, consistency, accuracy, timeliness, validity, currency and integrity.

We know from experience that it is not always easy to rigorously define data quality requirements. At the same time, data quality requirements are based on trade-offs: how does the impact of an incorrect data point compare with the cost of ensuring its quality? Can a business confidently move ahead with the deployment of a solution with the data at hand and the existing data quality management processes? The existing definition of data quality relies on our ability to estimate the repercussions of poor data quality on business activities and processes. For instance, incorrect information on product weights can lead to an inefficient estimation of storage needs, or the overestimation of the profitability of a new distribution center, or even hinder the possibility of exporting goods into a certain country. The consequences can be severe as data quality issues can trigger cascades: "compounding events causing negative downstream effects" ([Samb21]). The importance of data quality can never be understated, especially in AI projects. To the point that there is an increasing number of AI experts advocating a transition from model-centric AI, where the focus is on model improvement, to data-centric AI, which claims that the focus should be on curating data.

MODERN DATA QUALITY PROCESSES

The starting point of data quality is a rigorous definition of all the entities and fields in your data model. From there, the building blocks of a data quality system are manually defined rules, sets of constraints that limit the value that data is expected to assume. Can price be negative? Can the field value *country* be NULL? The need for precise definitions and data quality rules propagates downstream as the data is transformed from raw to informative dashboards and data visualizations.

Traditionally, the identification of data quality issues has been performed through two complementary approaches: one based on a 24/7 monitoring of the data quality rules; a second one based on the setup and design of a workflow that allows the consumers of data to easily report quality issues. This approach is still necessary as it is possible that rules might not offer complete coverage. At the same time, it is possible that rules become obsolete. It is therefore necessary to offer stewards a platform to not only report data quality issues but also contribute to the maintenance and improvement of rules.

Alongside workflows for the identification of data quality issues, mature organizations have workflows for their remediation. Remediation workflows should facilitate the collection of all the necessary information on the records identified as faulty, offer transparency on their root causes and provide input on how issues should be remediated.

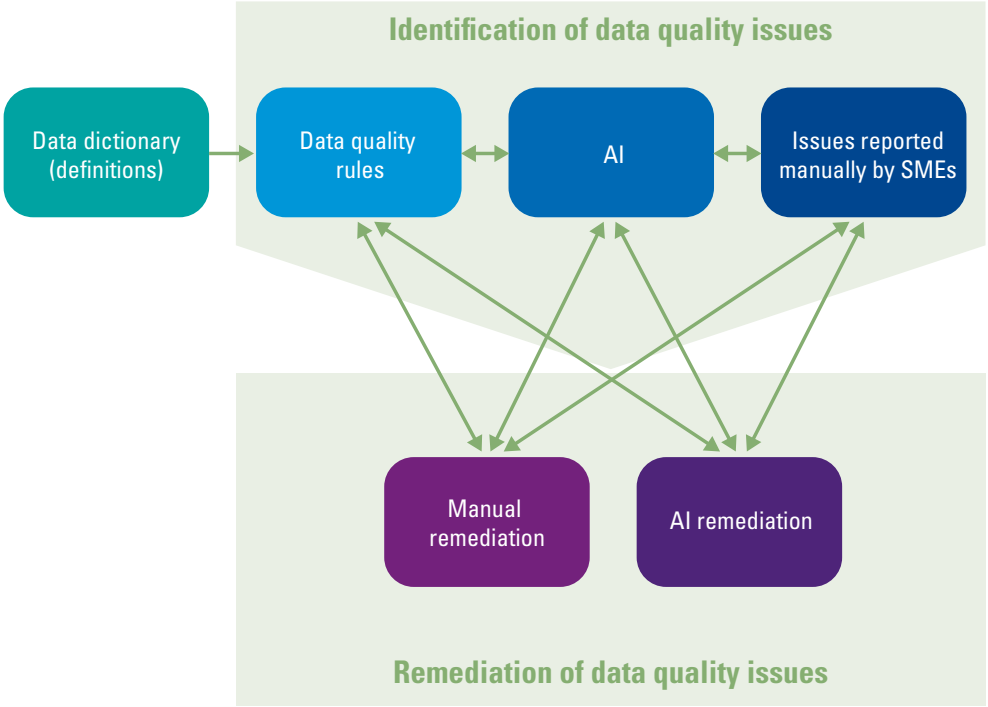

Figure 1. A modern process for identifying and remediating data quality issues.

Defining rules for identifying data quality issues as well as their manual remediation is very cumbersome. AI can boost the efficiency of both processes.

ARTIFICIAL INTELLIGENCE FOR THE IDENTIFICATION OF DATA QUALITY ISSUES

AI algorithms can automate the identification of issues. For example, the use of anomaly detection algorithms is particularly beneficial in the analysis of large, multi-dimensional datasets where the manual compilation of rules is complex and cumbersome. The use of anomaly detection algorithms also brings an additional benefit: anomaly detection algorithms can produce a score that quantifies the likelihood of data, whereas rules just result in a binary outcome (constraints are either satisfied or not).

Often, data is entered manually by human operators. Computer vision and natural language processing algorithms can also be leveraged to tap into the original data sources, such as PDF and word files, extract data and validate what has been typed in the system.

Furthermore, machine learning models can support the deduplication of data, estimating the probability that two records refer to the same entity (entity resolution).

ARTIFICIAL INTELLIGENCE FOR THE REMEDIATION OF DATA QUALITY ISSUES

AI models (regression, clustering) can support the imputation of missing data and the correction of anomalies. Alongside recommending a resolution for data quality issues, models can be designed to output a score that measures the degree of confidence that the AI system has in its recommendation. It is therefore possible to design human-in-the-loop remediation processes where operators are involved only when the confidence of the AI systems falls below a given threshold.

DATA QUALITY IS AN ONGOING EFFORT

Getting data quality right is not a one-off activity but an ongoing effort. Cleansing data should be a daily exercise. At the same time, improving the data quality management processes should also be a continuous improvement effort which entails reviewing data quality rules as well as improving the AI models involved in identifying and remediating data quality issues. Mature organizations are increasingly adopting a setup where data scientists are involved in data quality management efforts on an ongoing basis.

CONCLUSION

The importance of quality data in AI projects is well recognized in the industry: no data analytics or data science initiative can be successful if data quality requirements are not met. Garbage in – garbage out is an old and well-known truth that also applies to data analytics. However, the role that AI can play in improving data quality is often overlooked. Ask not what data quality can do for AI, ask what AI can do for data quality.

Reference

[Samb21] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L.M. (2021, May). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

About the authors

Mattia Ferrini is a director at KPMG AG. He has over 15 years of experience at the intersection of applied mathematics, data science, software engineering and business. Prior to joining KPMG in 2016, Mattia managed a high Sharpe-ratio quantitative hedge fund, pioneering the use of machine learning in the industry and he was responsible for EMEA pricing and revenue management at a large, international, e-marketplace where he led the development of algorithms to model marketplace dynamics and forecast user behavior. He has gained professional experience in various organizations worldwide.

Francesco Guidi is a manager at KPMG Switzerland. Overall, his work experience amounts to more than 20 years, of which the last 12 in data management from different perspectives. He worked in different companies with extensive participation across Europe, Africa, Middle East and Asia.