A Survey on Data Quality Dimensions and Tools for Machine Learning

Yuhan Zhou, Fengjiao Tu, Kewei Sha, Junhua Ding, & Haihua Chen
haihua.chen@unt.edu

Department of Information Science
University of North Texas

Sep 18, 2024

Outline

- Introduction
- Data Quality Dimensions
- Data Quality Tools
- Illustrations of DQ Tools

Introduction

Research Summary

- An overview of 4 DQ dimensions, and 12 metrics in ML, with the definitions, and examples
- A summary and comparative analysis of 17 DQ evaluation and improvement tools in the last 5 years
- A development workflow for the framework and function designs of DQ tools

Data Quality Dimensions

Definitions of DQ dimensions, metrics

- DQ dimensions define aspects of data quality that can be measured and provide the reasons for measurement.
- DQ metrics answer what and how to measure.

Data Quality Dimensions

(1) Intrinsic dimension

Intrinsic dimension can be assessed by measuring internal attributes or characteristics of data based on given references.

Corresponding Metrics

- Correctness:
  - A record in a dataset is free of errors.
  - Data is correctly labeled if it is a labeled record.
- Duplication:
  It measures if the same instances repeat in the dataset, especially in both the training and test datasets.
- Trustworthiness:
  It defines how factual the source that provides the information is. It can be subjectively evaluated, such as indicating the level on a scale, or the data can go through fact-check algorithms.

Data Quality Dimensions

(2) Contextual dimension

It ensures the data aligns with the goals of ML projects.

Corresponding Metrics

- Class imbalance:
  It evaluates if the distribution across the known classes is biased or skewed.
- Completeness:
  A complete dataset should include as few missing values as possible.
- Comprehensiveness:
  A dataset contains all representative samples from the population.
- Unbiasedness:
  It refers to whether the training data has a distribution bias or historical bias.
- Variety:
  Each validation dataset and test dataset should contain a significant amount of new data compared to the corresponding training dataset.

Data Quality Dimensions

(3) Representational dimension

Representational dimension assesses the formats and structures of data, such as if the data is concisely and consistently represented, but also interpretable.

Corresponding Metrics

- Conformity:
  It measures how much the data conforms to the conventions for capturing information in a certain manner, including machine-readable data structures and formats for capturing specific attributes.
- Consistency:
  It requires data to be presented in the same format and to be compatible with previous data.

Data Quality Dimensions

(4) Accessibility dimension

Accessibility dimension evaluates the extent of obtaining either the entire or some portion of the data. Availability allows users to use and share the data with safety controls.

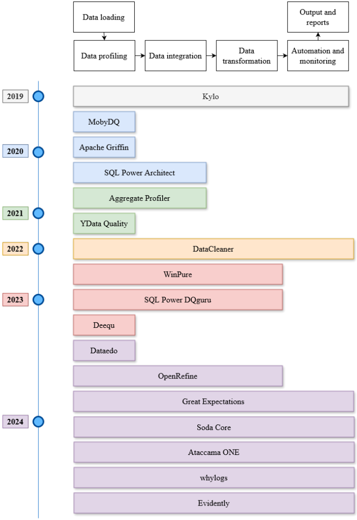
Corresponding Metrics

- Availability:
  High data availability ensures that data is readily accessible with defined user permissions for access and modifications.

Data Quality Tools

17 DQ tools, including applications and Python libraries

- Data profiling
- Data integration
- Data transforming
- Automation and monitoring



Data Quality Tools

Functions of DQ tools

- Data profiling:
  - To have an overview of the data, know DQ issues, and decide on corresponding fixing strategies
  - Frequency statistics, duplication analysis, data pattern discovery, drill-through analysis, etc.
- Data integration:
  - To maintain consistency when the customer wants to merge data from a different source
  - Ataccama ONE supports large-load data integration with seamless performance and continuous DQ checks during the integration

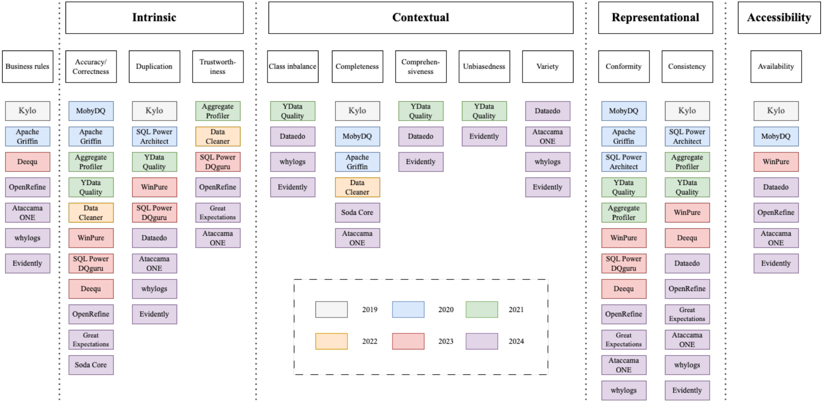
Data Quality Tools

Functions of DQ tools

- Data transforming:
  - To take action to fix the issues presented in the data profiling or discovery stage
  - Data cleaning, matching, merging, and de-duplicating tasks, etc.
- Automation and monitoring:
  - After users confirm the effectiveness of the DQ evaluation and transformation results
  - To automatically re-activate the workflow when new data is coming and generate up-to-date reports
  - Time-sensitive tasks

Data Quality Tools

DQ dimensions, metrics, and corresponding tools



	Intrinsic			Contextual					Representational		Accessibility
Business rules	Accuracy/ Correctness	Duplication	Trustworth-iness	Class inbalance	Completeness	Comprehen-siveness	Unbiasedness	Variety	Conformity	Consistency	Availability
⬜ Kylo	🟦 MobyDQ	⬜ Kylo	🟩 Aggregate Profiler	🟩 YData Quality	⬜ Kylo	🟩 YData Quality	🟩 YData Quality	🟪 Dataedo	🟦 MobyDQ	⬜ Kylo	⬜ Kylo
🟦 Apache Griffin	🟦 Apache Griffin	🟦 SQL Power Architect	🟧 Data Cleaner	🟪 Dataedo	🟦 MobyDQ	🟪 Dataedo	🟪 Evidently	🟪 Ataccama ONE	🟦 Apache Griffin	🟦 SQL Power Architect	🟦 MobyDQ
🟥 Deequ	🟩 Aggregate Profiler	🟩 YData Quality	🟥 SQL Power DQguru	🟪 whylogs	🟦 Apache Griffin	🟪 Evidently		🟪 whylogs	🟦 SQL Power Architect	🟩 Aggregate Profiler	🟥 WinPure
🟪 OpenRefine	🟩 YData Quality	🟥 WinPure	🟪 OpenRefine	🟪 Evidently	🟧 Data Cleaner			🟪 Evidently	🟩 YData Quality	🟩 YData Quality	🟪 Dataedo
🟪 Ataccama ONE	🟧 Data Cleaner	🟥 SQL Power DQguru	🟪 Great Expectations		🟪 Soda Core				🟩 Aggregate Profiler	🟥 WinPure	🟪 OpenRefine
🟪 whylogs	🟥 WinPure	🟪 Dataedo	🟪 Ataccama ONE		🟪 Ataccama ONE				🟥 WinPure	🟥 Deequ	🟪 Ataccama ONE
🟪 Evidently	🟥 SQL Power DQguru	🟪 Ataccama ONE							🟥 SQL Power DQguru	🟪 Dataedo	🟪 Evidently
	🟥 Deequ	🟪 whylogs							🟥 Deequ	🟪 OpenRefine	
	🟪 OpenRefine	🟪 Evidently							🟪 OpenRefine	🟪 Great Expectations	
	🟪 Great Expectations								🟪 Great Expectations	🟪 Ataccama ONE	
	🟪 Soda Core								🟪 Ataccama ONE	🟪 whylogs	
									🟪 whylogs	🟪 Evidently	

⬜ 2019 🟦 2020 🟩 2021

🟧 2022 🟥 2023 🟪 2024

Data Quality Tools

Comparative analysis

- The user experience
  - For tools that have not been updated for a long time, like MobyDQ, Apache Griffin, and SQL Power Architect, their user interfaces are relatively simple, and lacking in design and interaction.
  - Ataccama ONE and Evidently demonstrate useful guides, clear descriptions of features, example cases, and easy-to-navigate websites.
  - YData Quality and Evidently support low-code commands and make the tool more user-friendly.
- Integrating AI and GPT
  - Winpure, Ataccama ONE, Soda Core, and Evidently – have already stepped out to integrate AI and GPT technology into the modeling, rules suggestion, and monitoring tasks.

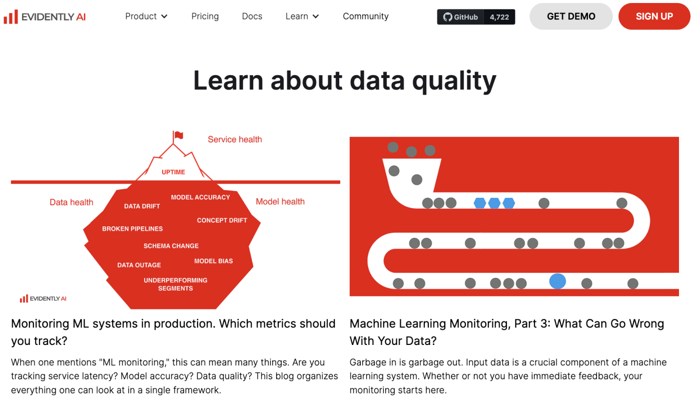
Data Quality Tools

Limitations of the tools

- Customized business rules
  - Not many tools enable customized DQ evaluation rules or revising the current rules.
- Dimensions and metrics
  - Not standardized.
  - Some adopted metrics are used generally for most data analysis, and only a few tools support evaluating DQ issues specific to ML tasks.
- The difficulty of automation and monitoring large-volume data

Illustration of DQ tools

Evidently AI



Learn about data quality

Monitoring ML systems in production. Which metrics should you track?

When one mentions "ML monitoring," this can mean many things. Are you tracking service latency? Model accuracy? Data quality? This blog organizes everything one can look at in a single framework.

Machine Learning Monitoring, Part 3: What Can Go Wrong With Your Data?

Garbage in is garbage out. Input data is a crucial component of a machine learning system. Whether or not you have immediate feedback, your monitoring starts here.

Figure: Evidently AI

Illustration of DQ tools

What is Evidently AI

Evidently helps evaluate, test, and monitor data and ML-powered systems. It is available both as an open-source Python library and a cloud platform.

- Predictive tasks: classification, regression, ranking, recommendations
- Generative tasks: chatbots, summarization
- Data monitoring: data quality and data drift for text, tabular data, and embeddings.

Next: Illustrations with Jupyter Notebooks