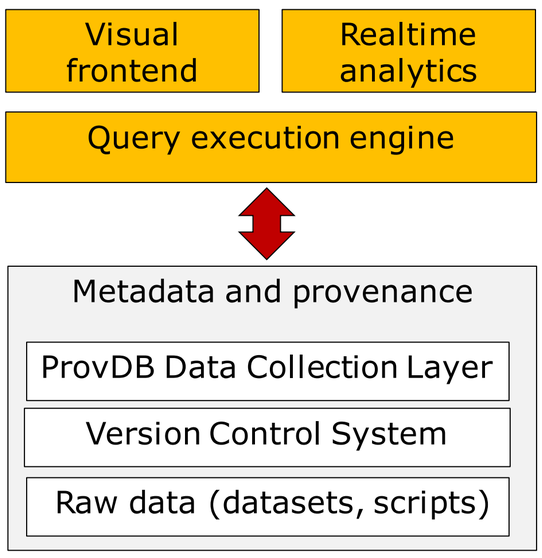
Approaches toward a Unified Framework for Data Management

Data-driven methods are becoming increasingly prevalent, while the data itself is becoming more dynamic. There is a clear need to move away from adhoc workflows toward a more scalable solution that integrates well with existing development processes. This poster aims to illustrate the current research effort toward such a framework.

Metadata and Provenance

Collaborative, adhoc data science workflows pose a challenge for traditional lifecycle management systems. Version control systems are based on a low-level file abstraction and lack capabilities to reason about data contained within versions and relationships between datasets.

ProvDB proposes a unified system to track version lineages of generated datasets, scripts and models to enable intospection and debugging of the workflows.

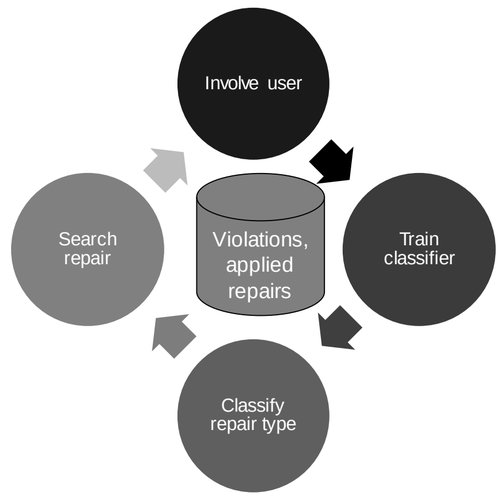


User challenge	Proposed solution
Define a schema for the provenance information a priori	A small base schema is fixed, but users can add arbitrary semi-structured metadata ("*schema-later*")
Capture metadata in a passive manner with minimal effort	Suite of provenance ingestors for popular frameworks (scikit-learn, caffe), default ingestor for UNIX shell commands
Generate useful insights from captured metadata	Logical data model is mapped into a Neo4J property graph to enable Cypher queries and visual exploration

Continuous Data Cleaning

Data-driven businesses need to ensure data quality not only for static data with fixed constraints, but also in dynamic environments where data changes frequently and constraints evolve over time. Inconsistencies can arise from incorrect data values or stale constraints.

The proposed system is able to operate on data streams by considering incremental changes instead of starting the repair process from scratch. It automatically suggests repairs and involves a human-in-the-loop to validate application semantics.



High-level overview of the proposed system

tid	FirstName	Surname	BranchID	Salary	Zip	City
t1	James	Brown	107	70K	50210	Miami
t2	Craig	Rosberg	107	50K	50210	Miami
t3	James	Brown	308	70K	21100	Atlanta
t4	Monica	Johnson	308	60K	21100	Houston
t5	James	Brown	401	80K	65300	NY
t6	Monica	Johnson	401	100K	65300	NY
t7	Mark	Douglas	401	130K	65300	Boston

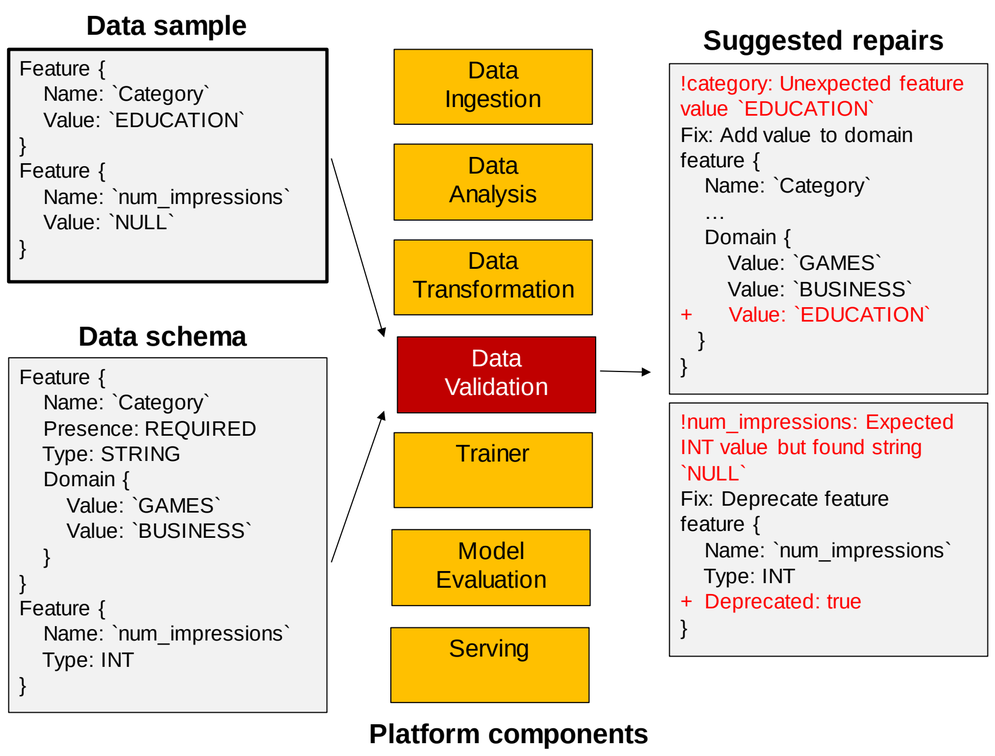
Constraint (FD)	Repair class
[FirstName, Surname] -> [Salary]	Constraint
[Zip] → [City]	Data

Data cleaning is reframed as a classification problem to repair the data, the constraint or both.

TFX: toward an end-to-end platform

TensorFlow Extended (TFX) is a machine learning platform implemented at *Google*. Driven by the need to keep track of experiment history in a centralized database and be resilient against disruptions from inconsistent data, TFX integrates aforementioned approaches into an end-to-end platform with shared configuration.

It allows continuous training and validation over evolving data, captures data transformation pipelines and exposes a simple interface for users of various levels of expertise to monitor and debug their workflows.



Conclusion

While TFX is built on top of *TensorFlow* for the machine learning use case, it provides a solid reference architecture for a general-purpose data management platform, including traditional ETL processes.

An open standard for metadata models and stricter formalization of data workflow activities through clear interfaces would not only enable researchers to contribute their work directly (e.g. classifiers for data cleaning), but also create opportunities for sharing metadata and generating new insights.

Collectively, this could be a step away from adhoc glue code, custom scripts and fragile systems with high technical debt toward more resilient, transparent and reproducable processes.

Janusch Jacoby

Data Engineering in der Praxis
WS 17/18 (Master)
Hasso Plattner Institute, Potsdam, Germany

References

H. Miao, A. Chavan, A. Deshpande. ProvDB: A System for Lifecycle Management of Collaborative Analysis Workflows.
M. Volkovs, F. Chiang, J. Szlichta, R. Miller. Continuous Data Cleaning.
Google Inc. TFX: A TensorFlow-Based Production-Scale Machine Learning Platform
N. Polyzotis, S. Roy, S. Whang, M. Zinkevich. Data Management Challenges in Production Machine Learning
S. Schelter. Data Management Challenges in Machine Learning Pipelines.

Hasso Plattner Institut