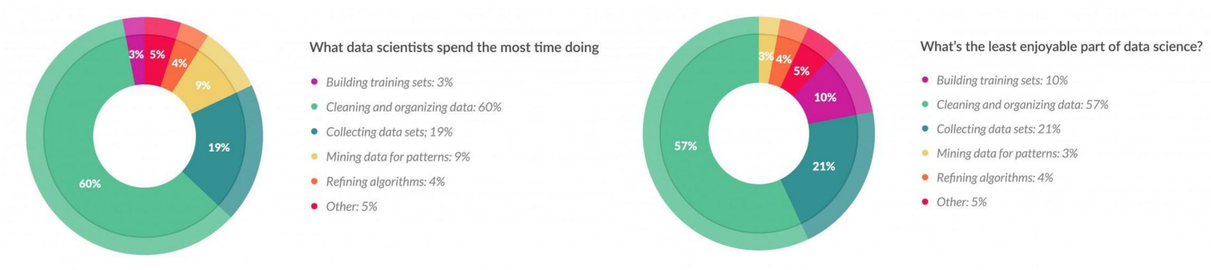
PIClean: a Probabilistic and Interactive Data Cleaning System

Zhuoran Yu, Xu Chu

1) Data Quality

- Real life data is often dirty
  - Data error rates in industry: 1% - 30% [Redman, 1998]
  - It is estimated that data quality problems cost U.S. businesses more than $600 billion a year [The Data Warehousing Institute 2002]
  - One in three business leaders does not trust the data they use to make decisions [IBM white paper 2011]
- Data scientists view data cleaning as the most time-consuming and least enjoyable work [Press, 2016]



Pictures credit to Gil Press, 2016

2) Limitations of Current Methodologies

- Usability of Rule-based Cleaning:
  - Rule-based cleaning is an important class of tools to capture data inconsistencies [Fan, 2012]
  - Data engineers mostly still need to write cumbersome ad-hoc scripts that encode rules
  - A large amount of money and human labor are spent on designing and confirming rules
- Coverage of Long-tail Errors:
  - Even with combination of all available tools, detection efficiency on real-world datasets is far from satisfactory [Abedjan, 2016]
  - Up to 40% in real datasets can remain undetected

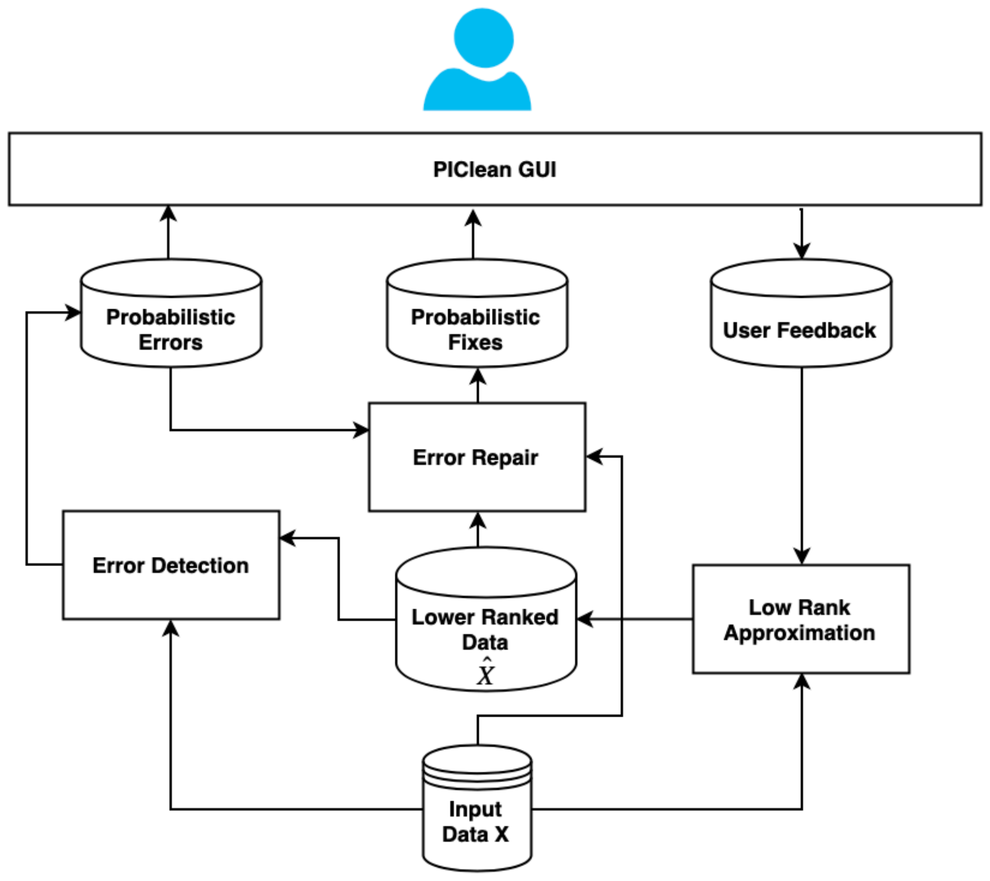
3) Our Approach

We propose PIClean, a probabilistic and interactive data cleaning system.

- Backend:
  - Our backend is built with low-rank approximation
  - Low-rank approximation has its power to capture dependencies between columns of tables
  - This approach saves effort of manually specifying cleaning rules
  - Our algorithm computes probability of each cell being erroneous and provide candidate fixes
- Frontend:
  - Probabilities associated with cells are presented to users
  - Users are able to accept, reject or provide fixes for next iteration of computation
  - Users are also provided with candidate fixes ranked in order for references

Our system limits extra input from users to save human efforts.

4) System Architecture



6) Categorical Data

Real-world data tables may contain categorical values, our system provide different encoding strategies for categorical values with different features:

- One-hot Encoding: For independent values with small domain such as gender, color, etc.
- Character-level Encoding: For values with specific structures such as email address, phone numbers.
- Multi-dimensional Scaling: For values such that distance between different values can be properly defined by distance functions

Backend computation is associated with encoded data and probabilities for categorical cells are calculated by probabilities of all fields in encoded vectors.