INFO 290T

Human-Centered Data Management

Profiler

Thoughts on Paper?

- Interface?
- Evaluation?
- Writing?

What I found neat about the paper…

- A focus on visualization of anomalies of various types
- Ways to correlate/explain column anomalies based on other columns
- A neat distance-based objective to both explain anomalies and value distributions
  - D(column, anomaly class) or D (column, column')

Performance Evaluation

- The paper evaluates performance for brushing and linking interactions, and concludes they are interactive for up to 1M rows. Is this sufficient?

Performance Evaluation

- The paper evaluates performance for brushing and linking interactions, and concludes they are interactive for up to 1M rows. Is this sufficient?

- Not quite! Showing "related" columns to a given (anomalous) column can be quite time-consuming.
  - Why might this not be a problem?

Performance Evaluation

- The paper evaluates performance for brushing and linking interactions, and concludes they are interactive for up to 1M rows. Is this sufficient?

- Not quite! Showing "related" columns to a given (anomalous) column can be quite time-consuming.
  - Why might this not be a problem?
    - Can be done offline

Issues in data

The paper enumerates four (five, including FD violations) types of data quality issues:

- Missing values
- Errors (e.g., during data entry)
- Extreme values
- Inconsistencies (outliers that may be erroneous)

What are other types of data quality issues *not* caught by Profiler?

Issues in data

The paper enumerates four (five, including FD violations) types of data quality issues:

- Missing values
- Errors (e.g., during data entry)
- Extreme values
- Inconsistencies (outliers that may be erroneous)

What are other types of data quality issues *not* caught by Profiler?

Duplicate entities, values "in range" that are "incorrect", numerical discrepancies (date of birth & current date)

Fitting into the workflow

At what stage would Profiler fit into a data scientist's workflow? At the start, in the middle, or towards the end?

If you were to use Profiler, how would you use it?

Once you discovered some anomalies, how would you go about fixing it?