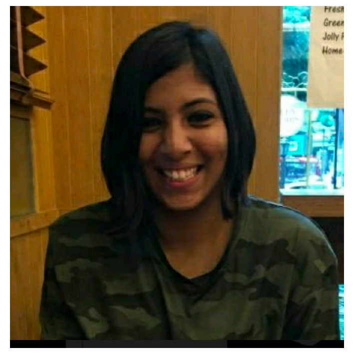
Automated Data Cleaning Can Hurt Fairness in ML-based Decision Making

Shubha Guha
UvA

Joint work with



Falaah Arif Khan
NYU

Julia Stoyanovich
NYU

Sebastian Schelter
UvA

ML in the Real World

- Used in critical decision-making processes.
  - Can reproduce or amplify pre-existing bias.
  - Bias can lead to unlawful discrimination. [1]
- Most ML applications in production are data-intensive, and require data cleaning. [2]
  - Large data size and short redeployment intervals mean that data quality issues are often addressed with automated cleaning techniques.

Data Quality and Fairness

- Evidence that data from historically disadvantaged groups may have poorer data quality. [7]
  - Systematic differences in data quality can potentially have negative impact on ML model fairness. [8]
- Evidence that data quality issues hurt predictive accuracy of ML models. [5]

Impact of Automated Data Cleaning on Fair Decision-Making

- RQ1: Does the incidence of data errors track demographic group membership in ML fairness datasets?
- RQ2: Do common automated data cleaning techniques impact the fairness of ML models trained on the cleaned datasets?

Sensitive Attributes

- Identified from occurrence [1] of unlawful discrimination according to US labor law [19] or European non-discrimination law [20].
- All datasets partitioned into *privileged group* and *disadvantaged group*.
  - Depends on the ML task which group is considered privileged vs. disadvantaged.

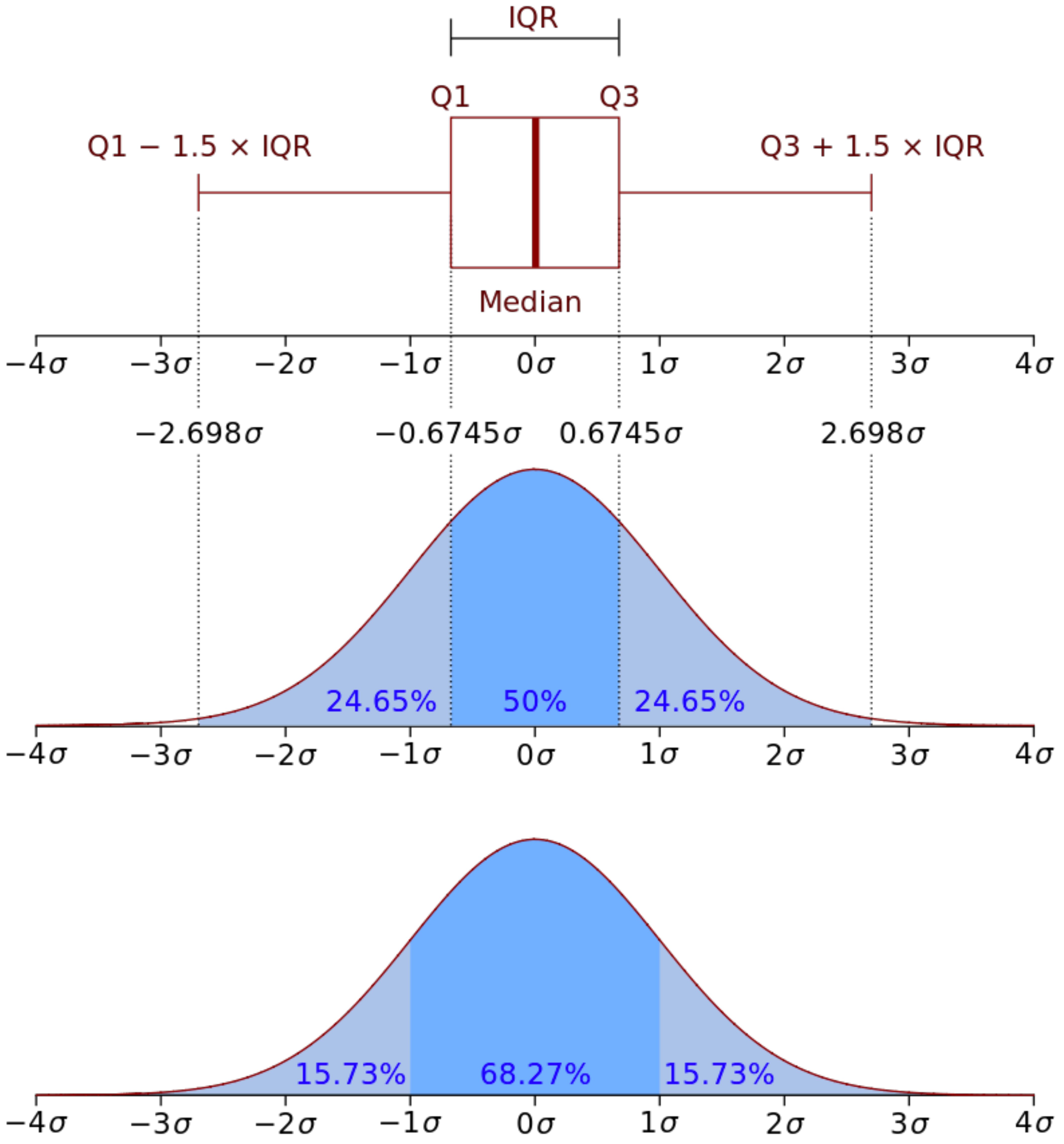
Benchmark Datasets

name	**source**	**number of tuples**	**number of attributes**	**sensitive attribute(s)**
`adult`	census	48,844	12	sex, race
`folk`	census	378,817	10	sex, race
`credit`	finance	150,000	8	age
`german`	finance	1,000	18	age
`heart`	healthcare	70,000	11	sex

TABLE I
BENCHMARK DATASETS USED IN ML FAIRNESS RESEARCH.

Error Detection Strategies

- Missing values
- Outliers
  - Standard deviation
  - Interquartile range
  - Isolation Forest
- Label errors

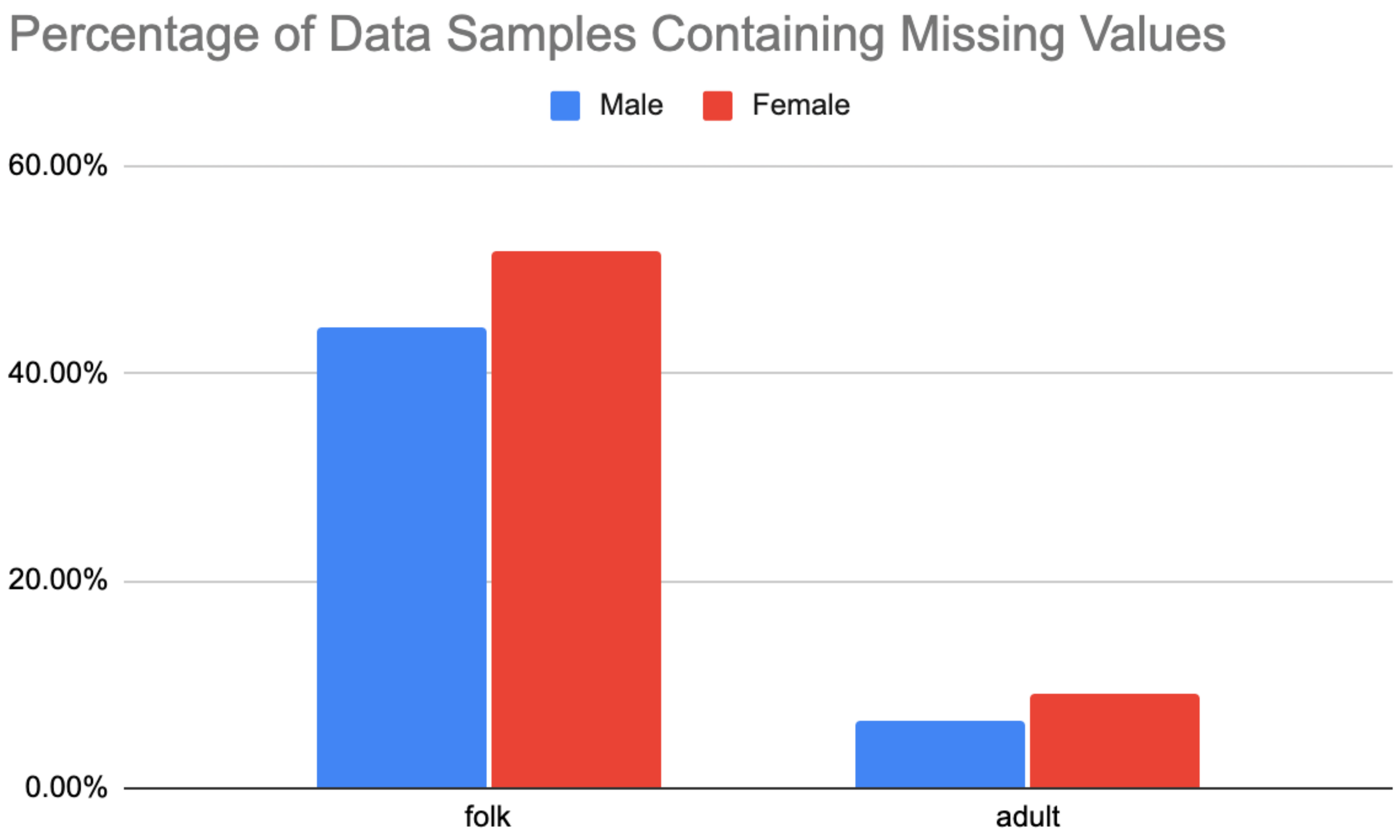


Data Cleaning Methods

- Missing value imputation
  - Column mean or mode (numerical)
  - Column mode or constant “dummy” value (categorical)
- Outlier repair
  - Replace detected outliers with mean or mode of column (numerical)
- Label error repair
  - Flip labels of flagged tuples

RQ1: Demographically Disparate Data Quality Issues

- Counted corrupt exemplars from privileged vs. disadvantaged groups.
- Reported only cases that pass significance test.



RQ1: Higher Rates of Missing Values for Disadvantaged Groups

RQ2: Impact of Automated Data Cleaning on Fairness

- Adapted existing CleanML benchmark for joint data cleaning and model training. [5]
- For each configuration:
  - One of the 5 datasets: adult, folk, credit, german, heart.
  - One of 3 ML model types: logistic regression, nearest neighbors, gradient-boosted decision trees.
  - One error detection strategy and one repair method.
  - 20 different train/test splits, 5 random seeds for hyperparameter search.
- In total, 26,400 models trained and evaluated.

Evaluation

- Predictive parity

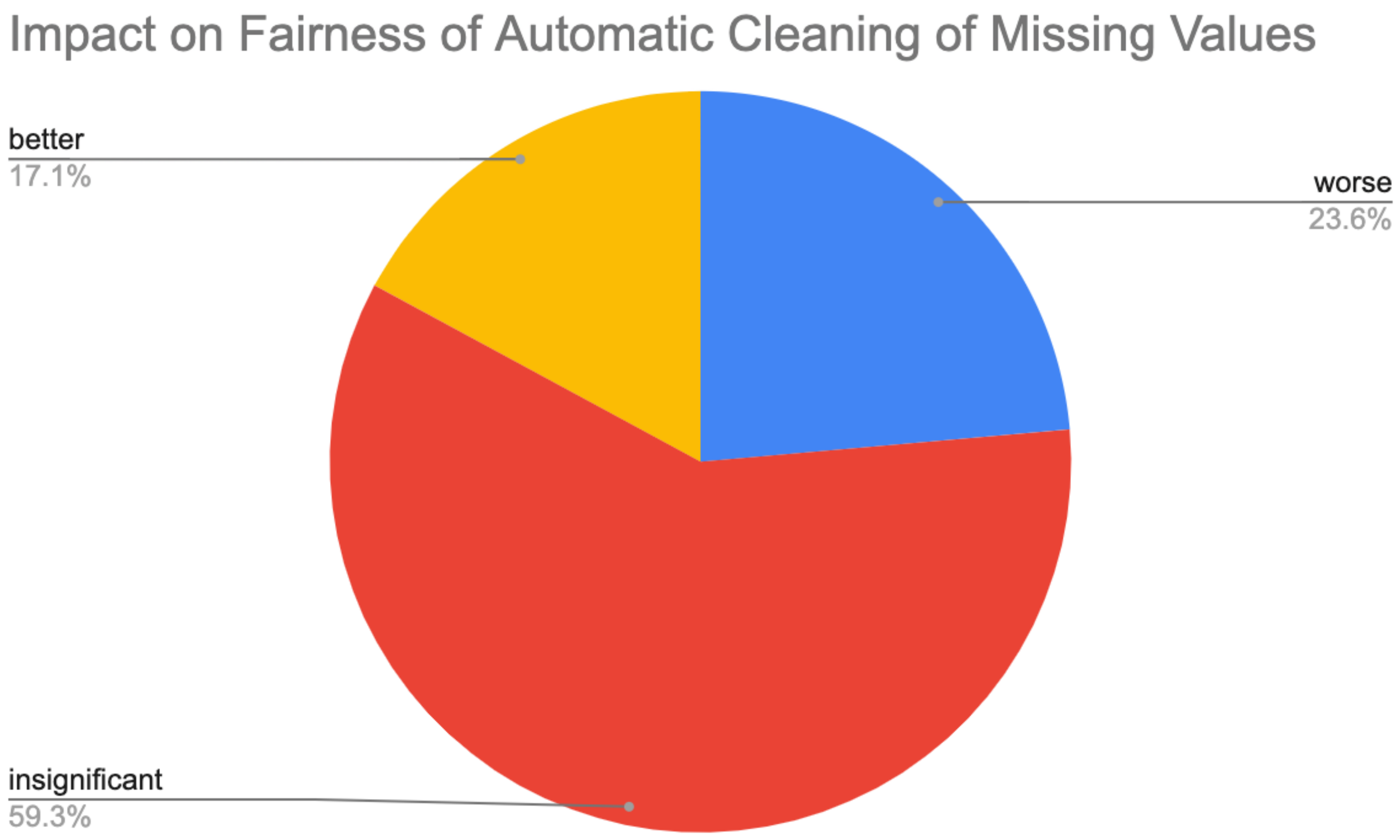
$$\frac{TP_{\text{priv}}}{TP_{\text{priv}} + FP_{\text{priv}}} - \frac{TP_{\text{dis}}}{TP_{\text{dis}} + FP_{\text{dis}}}$$

- Equal precision

- Equal opportunity

$$\frac{TP_{\text{priv}}}{TP_{\text{priv}} + FN_{\text{priv}}} - \frac{TP_{\text{dis}}}{TP_{\text{dis}} + FN_{\text{dis}}}$$

- Equal recall



RQ2: Negative Impact More Likely Than Positive Impact

Future Work

- Additional empirical evaluation with
  - Ground truth clean data
  - Data integrity constraints
  - More advanced data detection
  - More advanced data cleaning
    - Fairness-aware data cleaning methods
  - Intersectional formulations of demographic characteristics
  - Additional datasets from non-US sources

Thanks!

- Paper: https://ssc.io/pdf/demodq.pdf
- Code: https://github.com/amsterdata/demodq
- Find me on LinkedIn: https://www.linkedin.com/in/shubhaguha/