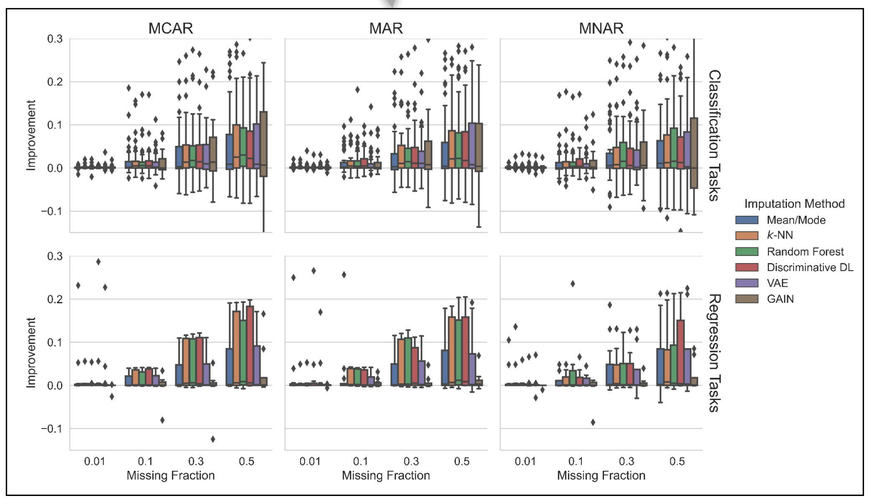
Automating Tabular Data Quality

Sebastian Jäger
Calgo Lab
sebastian.jaeger@bht-berlin.de

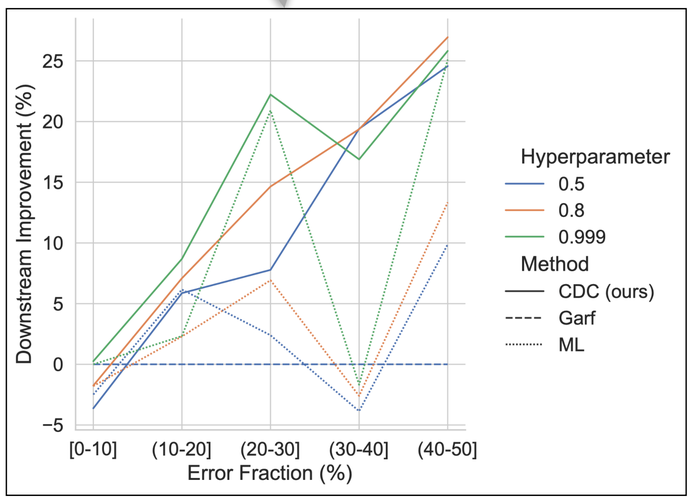
BHT
Berliner Hochschule für Technik

Patient	Age	ENAR Insurance	EAR Incident	Physician	ECAR Blood Oxygenation
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	-	Fall	Alice	84%
Tariq Lee	87	Uninsured	Hart attack	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	87%
Finn Cruz	63	-	Heart atrack	Bob	91%
Liam Wong	42	Health Corp.	Traffiv accidemt	Bob	86%

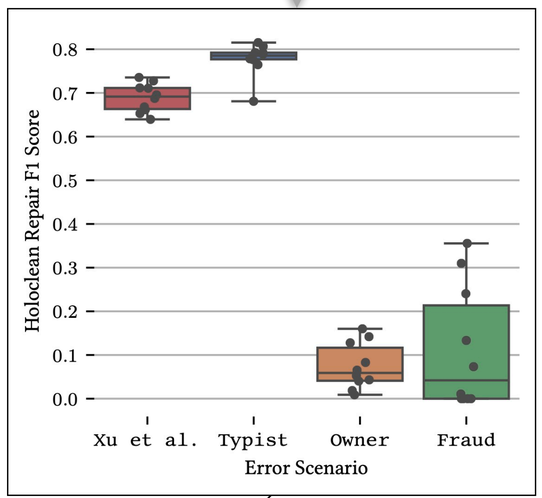
Missing Value Imputation

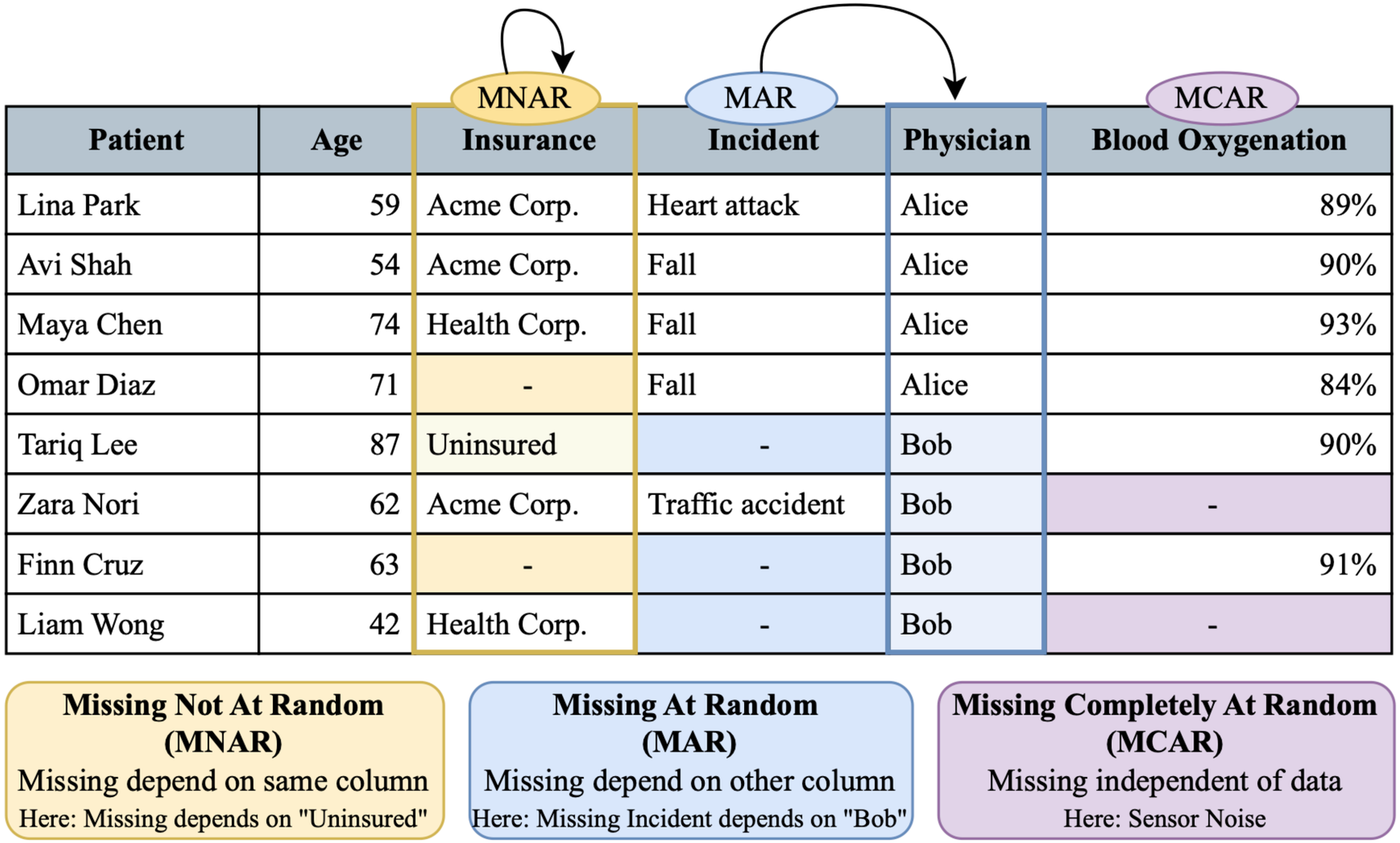


Conformal Data Cleaning



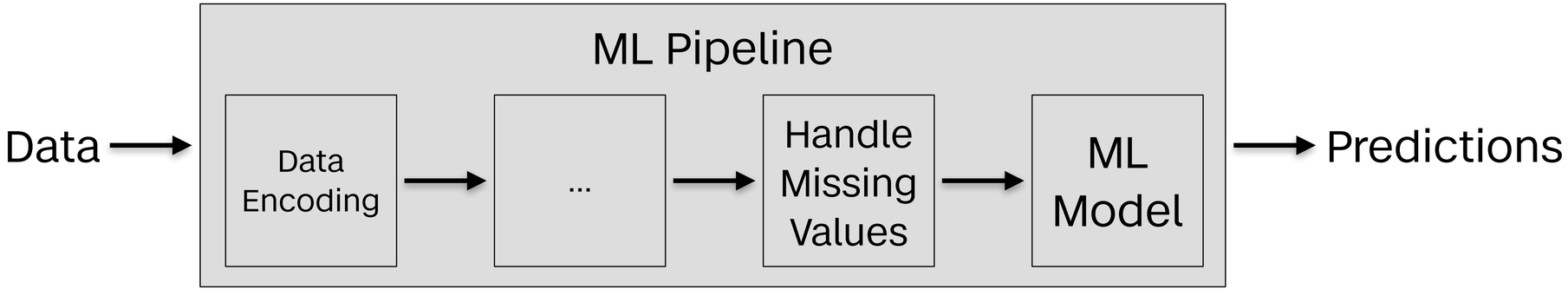
Generating Realistic Errors





Stef van Buuren. 2018. Flexible imputation of missing data (2 ed.). CRC Press, Taylor & Francis Group

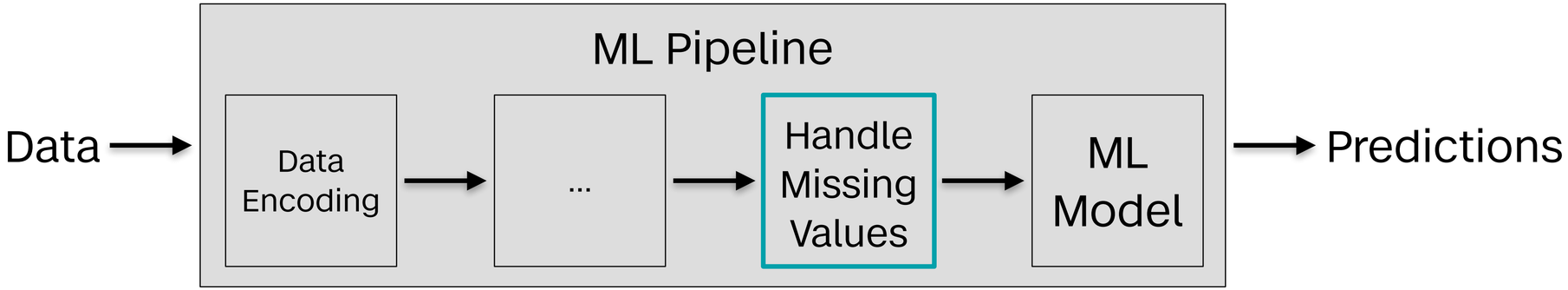
A Benchmark for Data Imputation Methods



Jäger et al., Frontiers in Big Data, 2021
https://github.com/se-jaeger/data-imputation-paper

A Benchmark for Data Imputation Methods

Patient	Age	Insurance	Incident	Physician	Blood Oxygenation
		MNAR	MAR		MCAR
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	-	Fall	Alice	84%
Tariq Lee	87	Uninsured	-	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	-
Finn Cruz	63	-	-	Bob	91%
Liam Wong	42	Health Corp.	-	Bob	-



$$X_c = \hat{f}_c(X_{\{1,\ldots,d\}\setminus\{c\}}),\ \forall c \in \{1,\ldots,d\}$$

```
def fit(self, data: pd.DataFrame) → Imputer:
    for column in data.columns:
        self._imputer[column] = RandomForest.fit(data.drop(columns=column), data.loc[column])
```

Jäger et al., Frontiers in Big Data, 2021
https://github.com/se-jaeger/data-imputation-paper

A Benchmark for Data Imputation Methods

- 69 heterogeneous datasets from OpenML
  - MCAR, MAR, MNAR
  - 1%, 10%, 30%, 50% missing values
- 6 imputation methods
  - mean/mode
  - RF, KNN (classic ML)
  - discriminative DL
  - GAIN, VAE (generative ML)
- Different scenarios and experiments
  - Here, we assume ML pipeline is trained on clean data

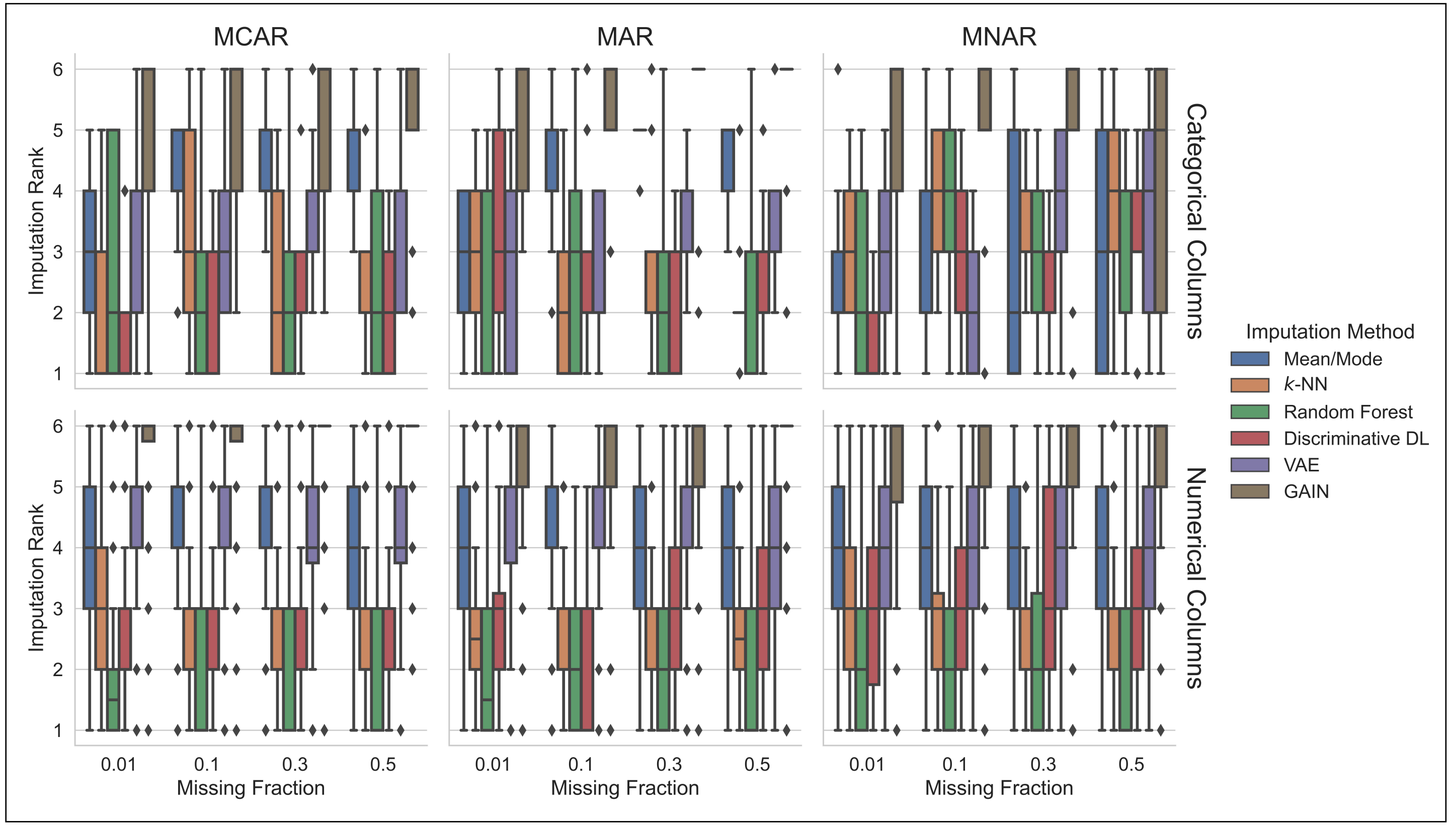
Jäger et al., Frontiers in Big Data, 2021
https://github.com/se-jaeger/data-imputation-paper

A Benchmark for Data Imputation Methods

Training on Complete Data - Imputation Quality

- Rank imputation tools for given group:
  - Missing Mechanism
  - Missing Fraction
  - Column Type

- RF, KNN, discriminative DL perform best
- Generative deep learning perform worst



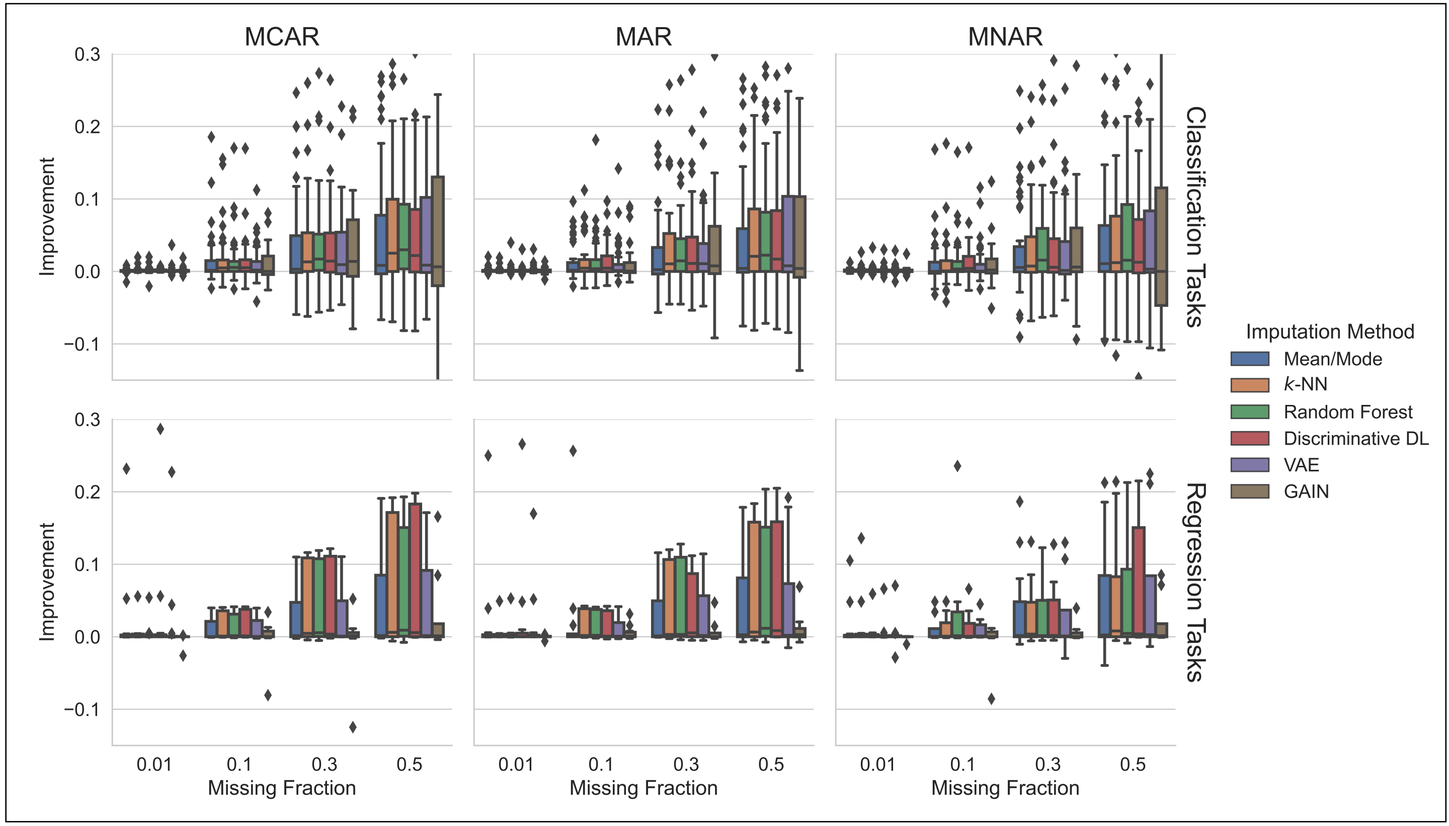
Jäger et al., Frontiers in Big Data, 2021
https://github.com/se-jaeger/data-imputation-paper

A Benchmark for Data Imputation Methods

Training on Complete Data - Impact on Downstream Task

- Downstream improvement* for given group:
  - Missing Mechanism
  - Missing Fraction
  - Downstream Type

- More missing values yield better improvement
- Classic ML yield better improvements + typically no degradations



$$* \text{ improvement} = \frac{\text{imputed} - \text{incomplete}}{\text{baseline}}$$

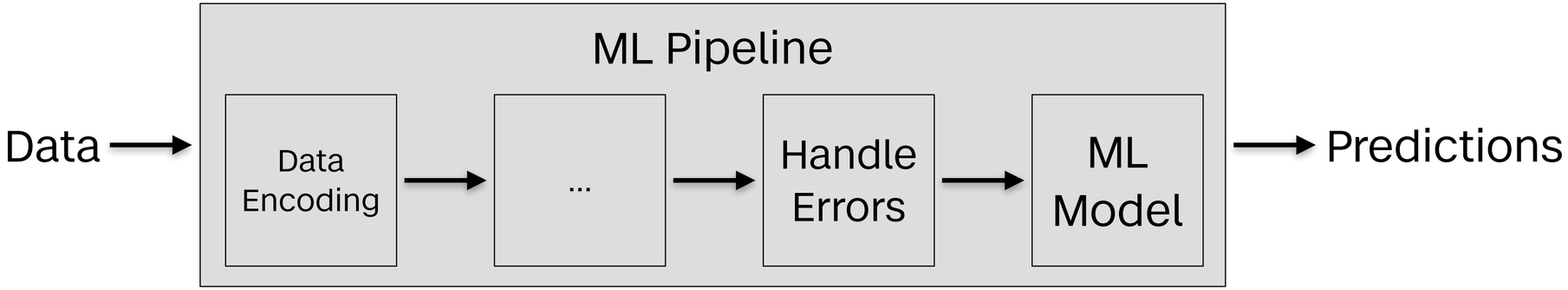
imputed: our imputation step

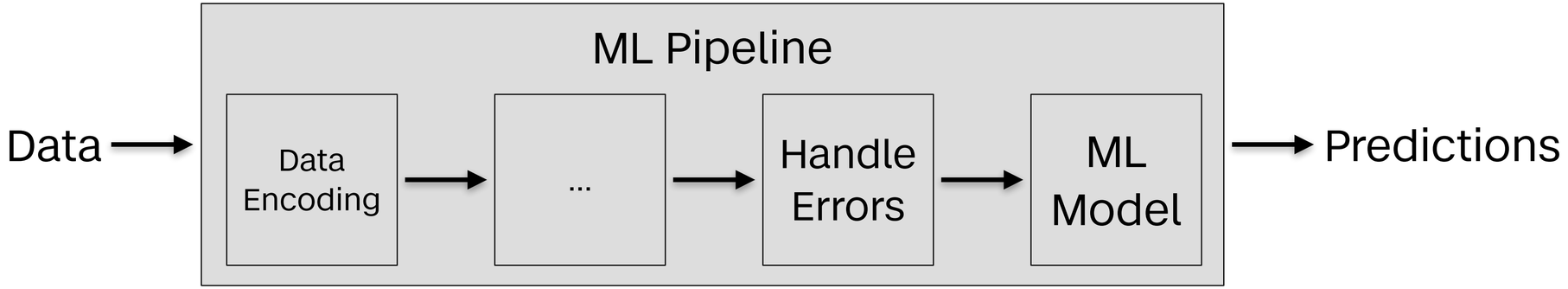
incomplete: naive imputation step

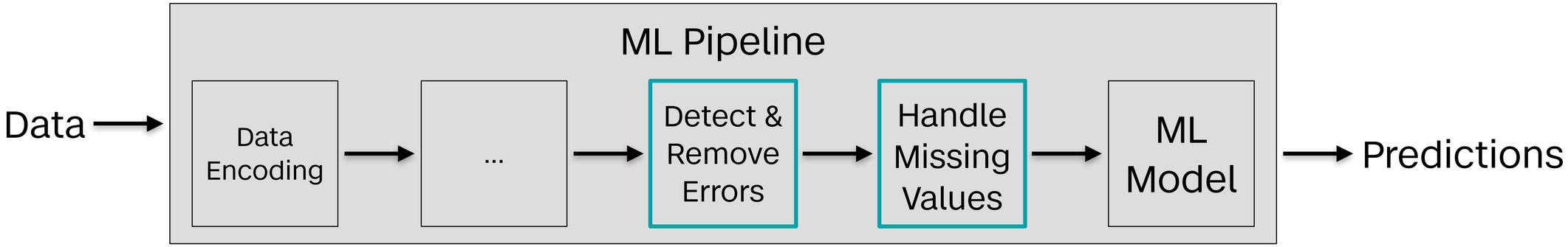
baseline: clean test data

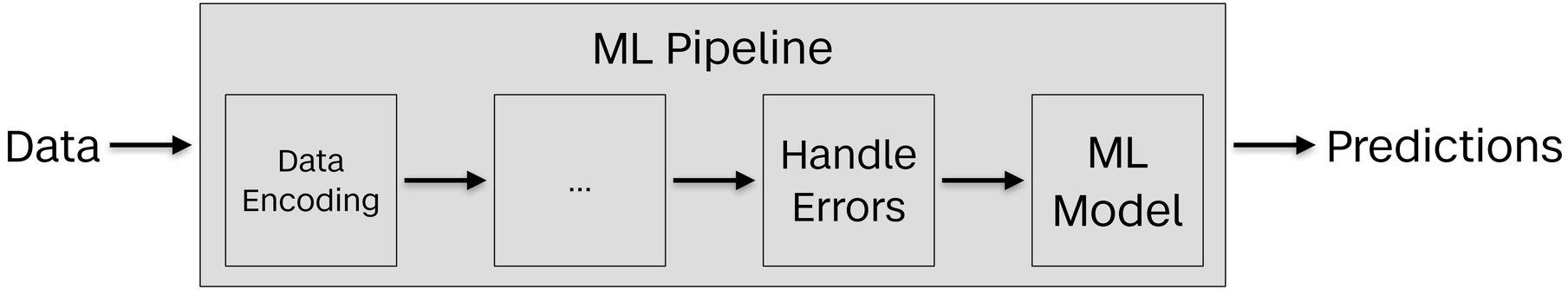
Jäger et al., Frontiers in Big Data, 2021
https://github.com/se-jaeger/data-imputation-paper

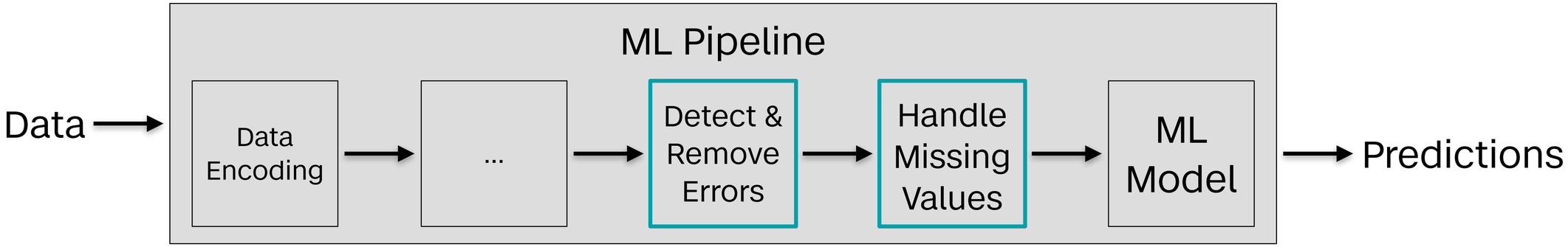
Patient	Age	Insurance	Incident	Physician	Blood Oxygenation
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	XX	Fall	Alice	84%
Tariq Lee	87	Uninsured	Strlke	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	87%
Finn Cruz	63	AAA	Fsll	Bob	91%
Liam Wong	42	Health Corp.	Accidemt	Bob	86%

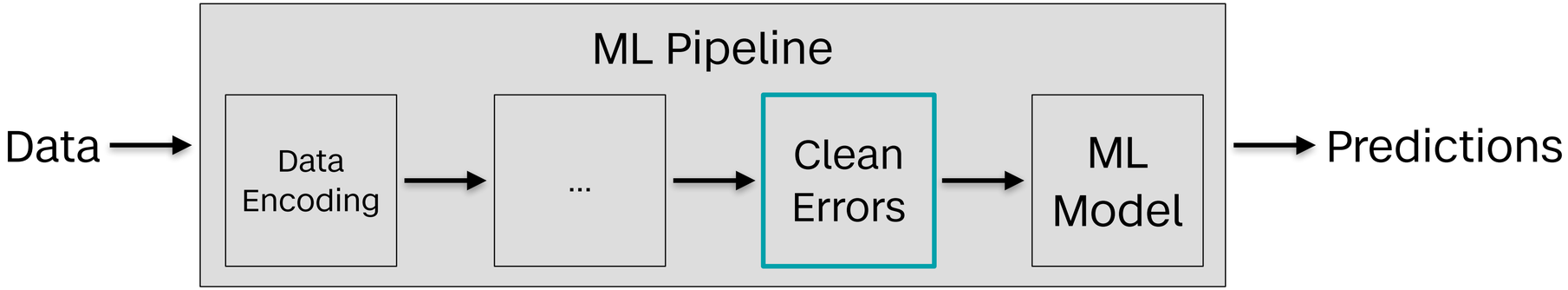

ML Pipeline
Data
Data Encoding
...
Handle Errors
ML Model
Predictions


ML Pipeline
Data
Data Encoding
...
Handle Errors
ML Model
Predictions


ML Pipeline
Data
Data Encoding
...
Detect & Remove Errors
Handle Missing Values
ML Model
Predictions

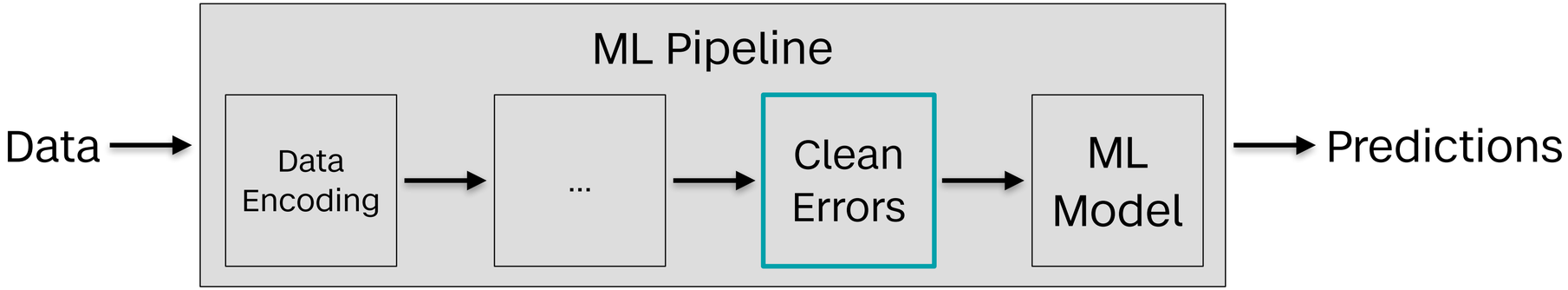

ML Pipeline
Data
Data Encoding
...
Handle Errors
ML Model
Predictions


ML Pipeline
Data
Data Encoding
...
Detect & Remove Errors
Handle Missing Values
ML Model
Predictions


ML Pipeline
Data
Data Encoding
...
Clean Errors
ML Model
Predictions

From Data Imputation to Data Cleaning — Automated Cleaning of Tabular Data Improves Downstream Predictive Performance

Patient	Age	Insurance	Incident	Physician	Blood Oxygenation
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	XX	Fall	Alice	84%
Tariq Lee	87	Uninsured	Strlke	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	87%
Finn Cruz	63	AAA	Fsll	Bob	91%
Liam Wong	42	Health Corp.	Accidemt	Bob	86%



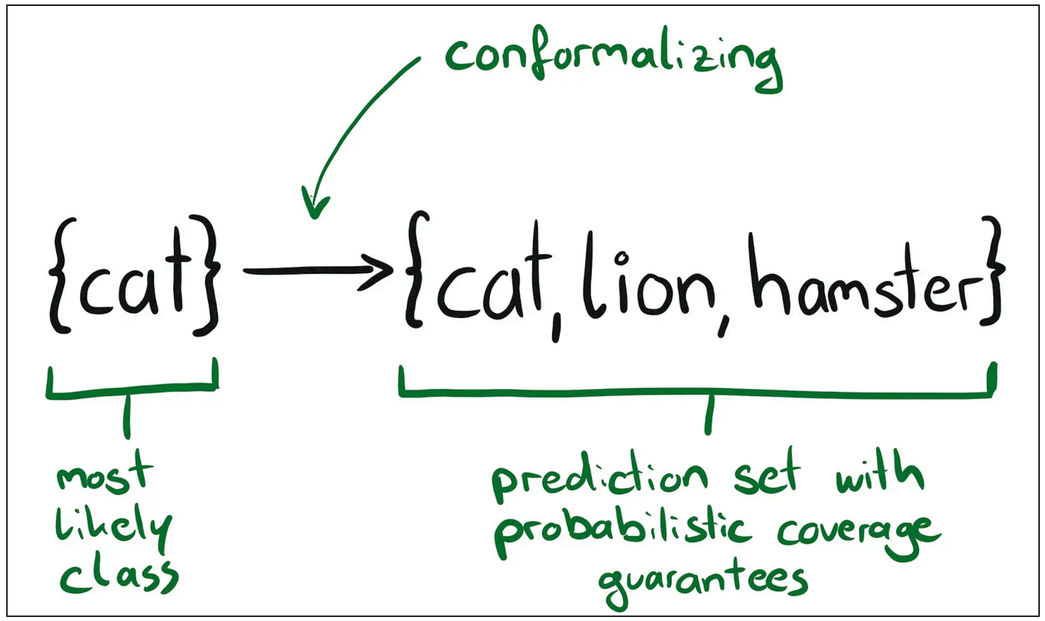
$$X_c = \widehat{cleaner}_c(X_{\{1,\ldots,d\}\setminus\{c\}}),\ \forall c \in \{1,\ldots,d\}$$

Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning

Conformal Prediction

- Distribution-free and model-agnostic uncertainty calibration method
- Only assumes data points are exchangeable*
- Turns black-box point predictor into set predictor
- Prediction sets have statistical ‘coverage’ guarantees
- User chooses ‘confidence level’
- Prediction set size represents ‘amount of uncertainty’

- Coverage:
  - $\mathbb{P}(y_{test} \in C(X_{test})) \geq$ confidence level
  - ‘for 99 out of 100 test samples, the prediction set contains the true label’



* ML typically assumes i.i.d. data, which implies exchangeable data points

Week #1: Getting Started With Conformal Prediction For Classification
Week #3: Conformal Prediction For Regression

Conformal Data Cleaning

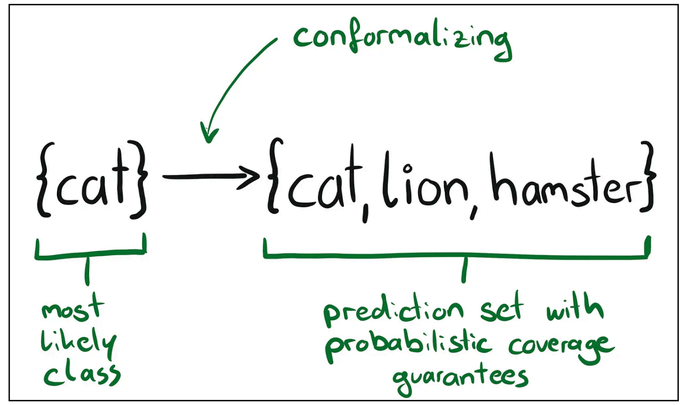
$$X_{i,c}^{test} = D_{i,\{1,\ldots,d\}\setminus\{c\}}^{test}$$

$$C_{i,c}, \hat{y}_{i,c} = \widehat{cleaner}_c(X_{i,c}^{test})$$

$$\tilde{D}_{i,c}^{test} = \begin{cases} D_{i,c}^{test}, \text{if } D_{i,c}^{test} \in C_{i,c} \\ \hat{y}_{i,c}, \quad \text{otherwise} \end{cases},$$

Patient	Age	Insurance	Incident	Physician	Blood Oxygenation
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	XX	Fall	Alice	84%
Tariq Lee	87	Uninsured	Strlke	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	87%
Finn Cruz	63	AAA	Fsll	Bob	91%
Liam Wong	42	Health Corp.	Accidemt	Bob	86%

Classification



Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning

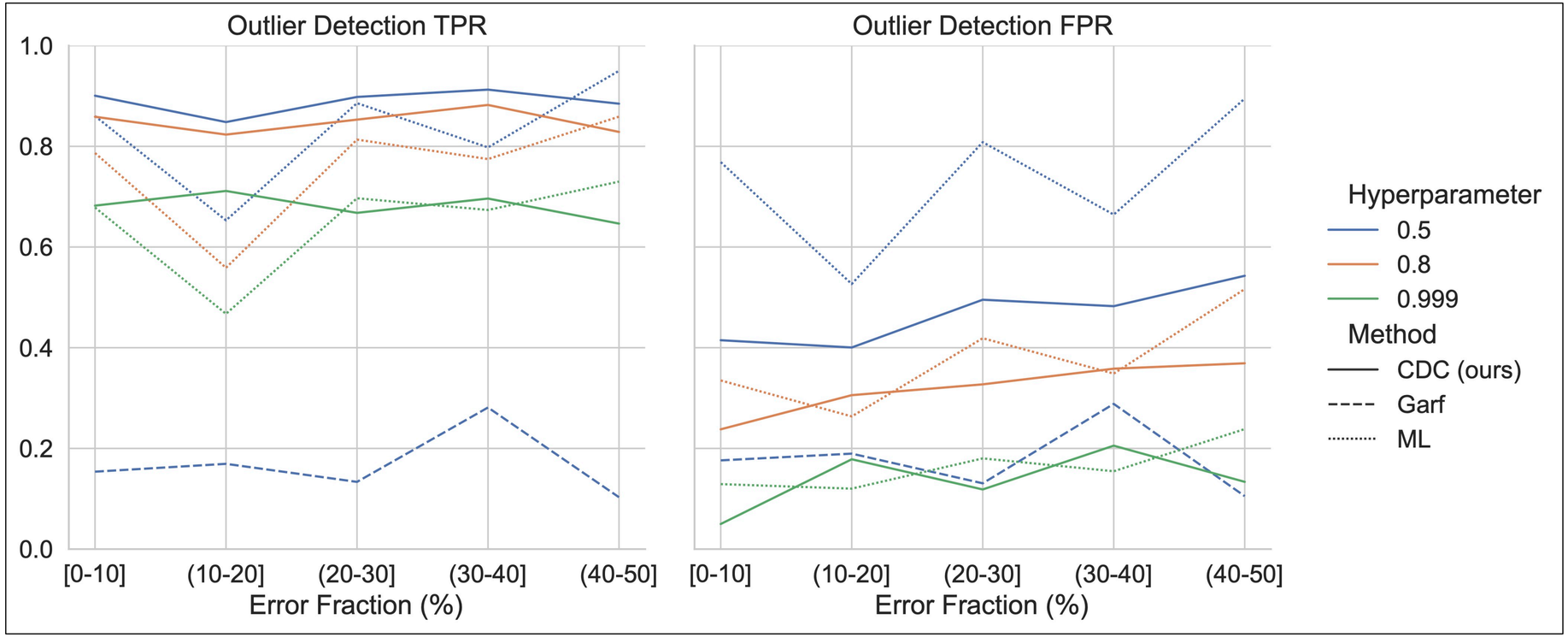
From Data Imputation to Data Cleaning —
Automated Cleaning of Tabular Data Improves Downstream Predictive Performance

- 16 heterogeneous datasets from OpenML
  - 4 error types
  - 1%, 5%, 10%, 30%, 50% erroneous cells
  - => for each dataset 20 corrupted versions
  - On average 11% ± 14, min 0%, max 41% erroneous cells

Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning

From Data Imputation to Data Cleaning —
Automated Cleaning of Tabular Data Improves Downstream Predictive Performance

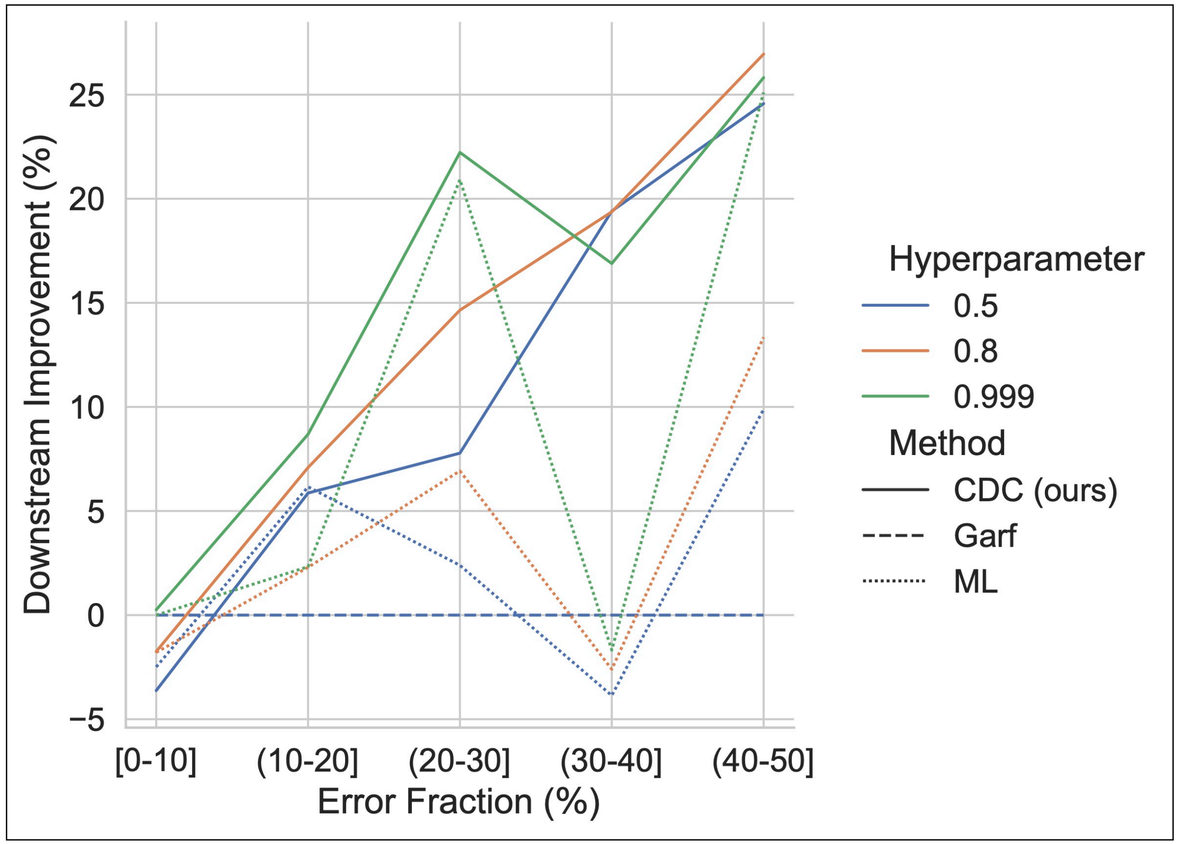
- High confidence level …
  - detects less errors
  - introduces fewer errors



Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning

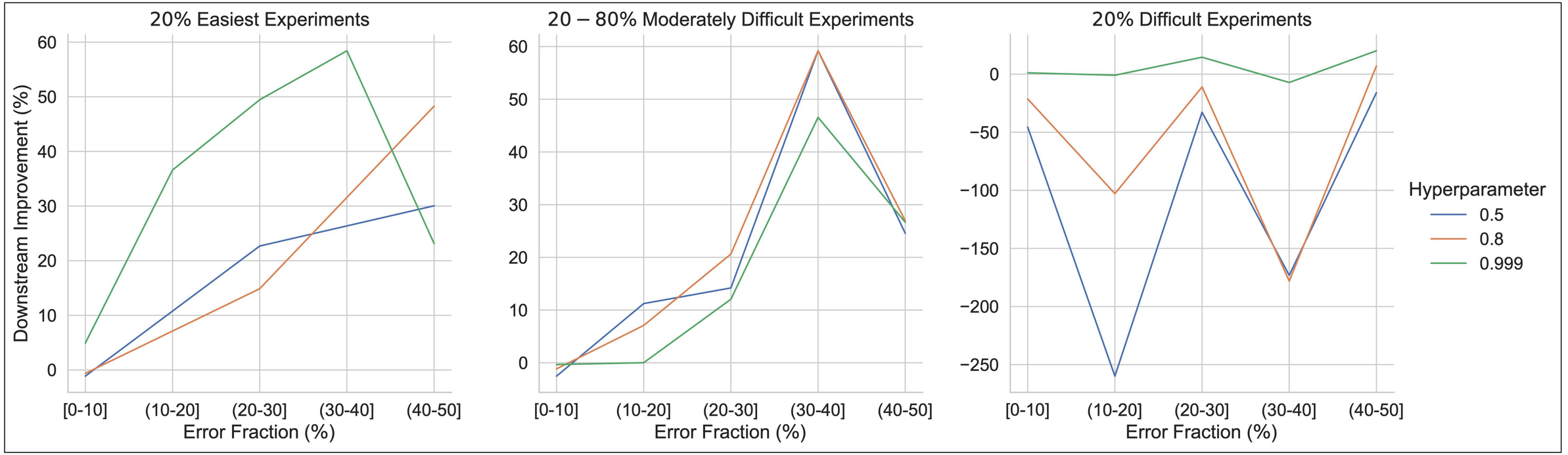
From Data Imputation to Data Cleaning —
Automated Cleaning of Tabular Data Improves Downstream Predictive Performance

- More errors yield higher improvements*
- High confidence level …
  - increases improvement
  - avoids performance degradation



$$* \text{ improvement} = \frac{\text{cleaned} - \text{dirty}}{\text{dirty}}$$

Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning



- “Difficulty” is based on the (relative) prediction set sizes
- Again, high confidence level prevents downstream degradation

Jäger and Bießmann, AISTATS 2024
https://github.com/se-jaeger/conformal-data-cleaning

Data quality is a hyperparmeter that can be optimized.

Towards Realistic Error Models for Tabular Data

Patient	Age	Insurance	Incident	Physician	Blood Oxygenation
Lina Park	59	Acme Corp.	Heart attack	Alice	89%
Avi Shah	54	Acme Corp.	Fall	Alice	90%
Maya Chen	74	Health Corp.	Fall	Alice	93%
Omar Diaz	71	- (was: Uninsured)	Fall	Alice	84%
Tariq Lee	87	Uninsured	Strlke (was: Stroke)	Bob	90%
Zara Nori	62	Acme Corp.	Traffic accident	Bob	87% (was: 85%)
Finn Cruz	63	- (was: Uninsured)	Fsll (was: Fall)	Bob	91%
Liam Wong	42	Health Corp.	Accidemt (was: Accident)	Bob	86% (was: 87%)

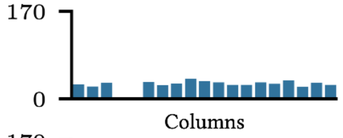
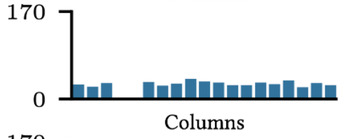
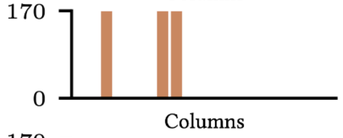
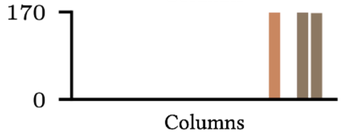
Errors Not At Random (ENAR)
Errors depend on same column
Here: Missing depends on "Uninsured"

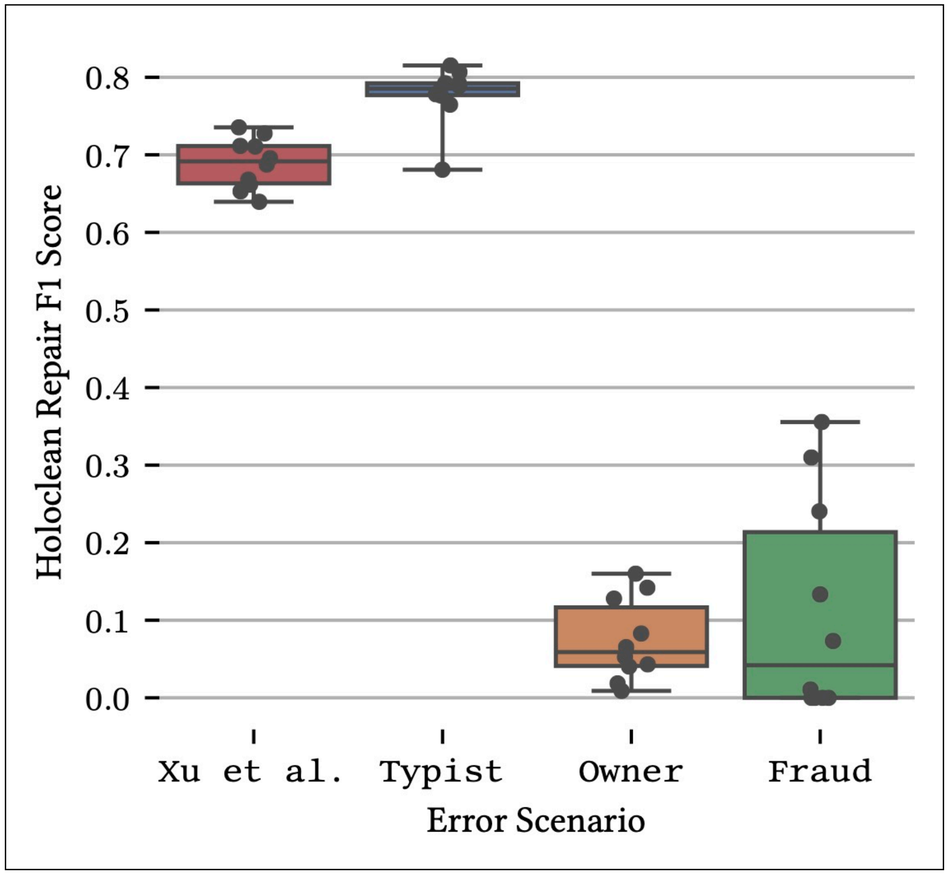
Errors At Random (EAR)
Errors depend on other column(s)
Here: Incident depends on "Bob"

Errors Completely At Random (ECAR)
Errors independent of data
Here: Sensor Noise

https://github.com/calgo-lab/tab_err

Towards Realistic Error Models for Tabular Data

Error Scenario	Errors	Error Mechanisms	Error Distribution (ECAR, EAR, ENAR)
Xu et al.	509	ECAR	
Typist	509	ECAR	
Owner	507	EAR	
Fraud	507	EAR, ENAR	



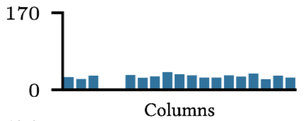
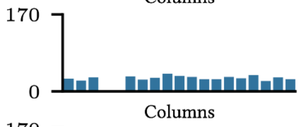
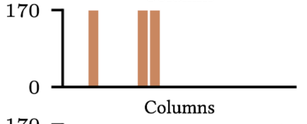
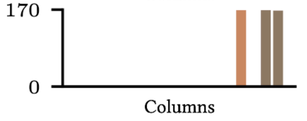
https://github.com/calgo-lab/tab_err

Towards Realistic Error Models for Tabular Data

- Disentangle **Error Mechanism** and **Error Type**
- Allows fine grained control of **Realistic Error Scenarios** ...
- ... or hands-free generation of perturbed datasets

- Possible applications:
  - Testing robustness of ML models or cleaning tools
  - Test data pipelines
  - Make data quality research reproducible by sharing error scenarios as code

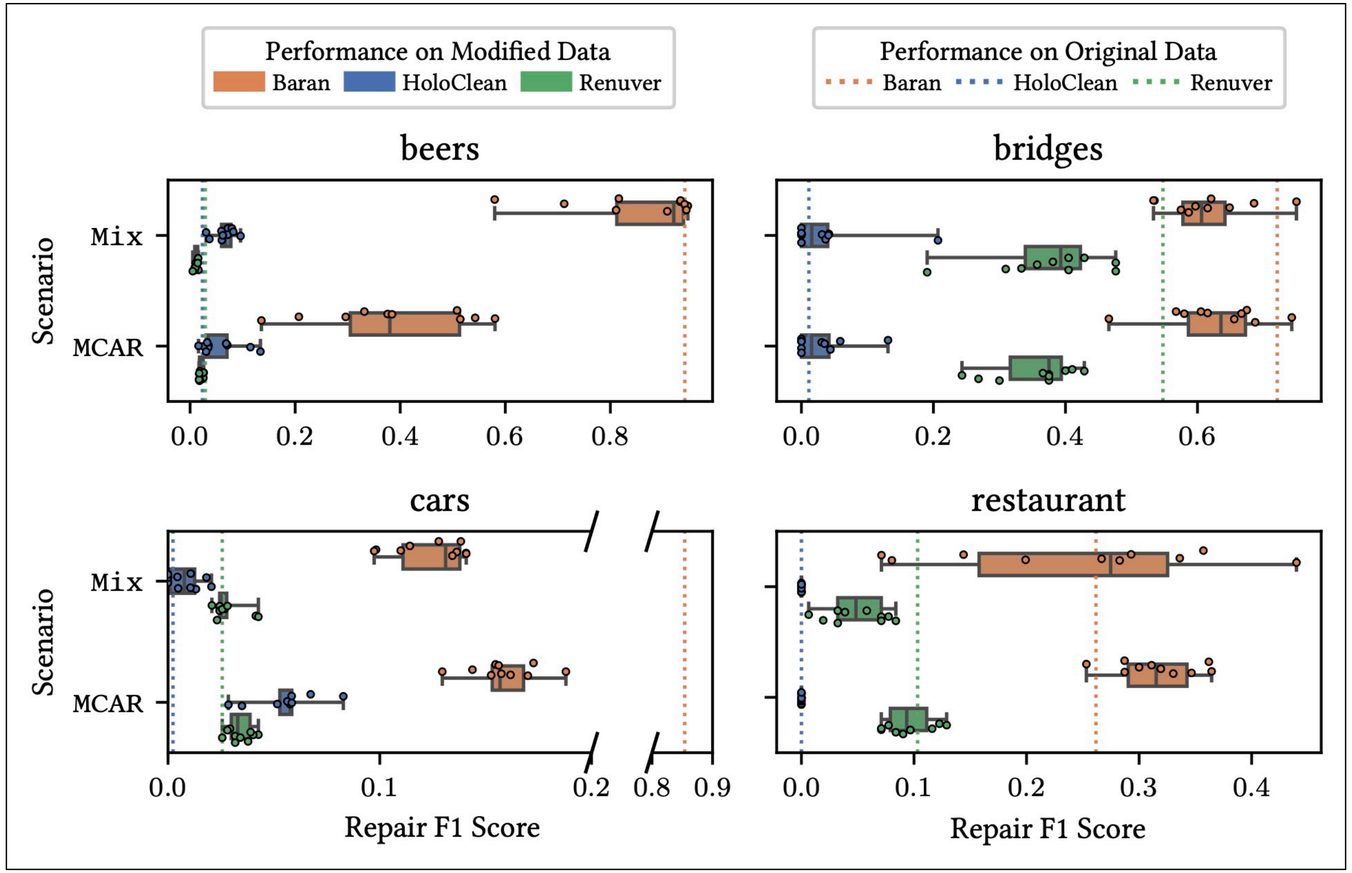
- pip install tab-err
- https://github.com/calgo-lab/tab_err/blob/main/examples/1-Getting-Started.ipynb

Error Scenario	Errors	Error Mechanisms	Error Distribution
			ECAR EAR ENAR
Xu et al.	509	ECAR	
Typist	509	ECAR	
Owner	507	EAR	
Fraud	507	EAR, ENAR	

https://github.com/calgo-lab/tab_err

Towards Realistic Error Models for Tabular Data

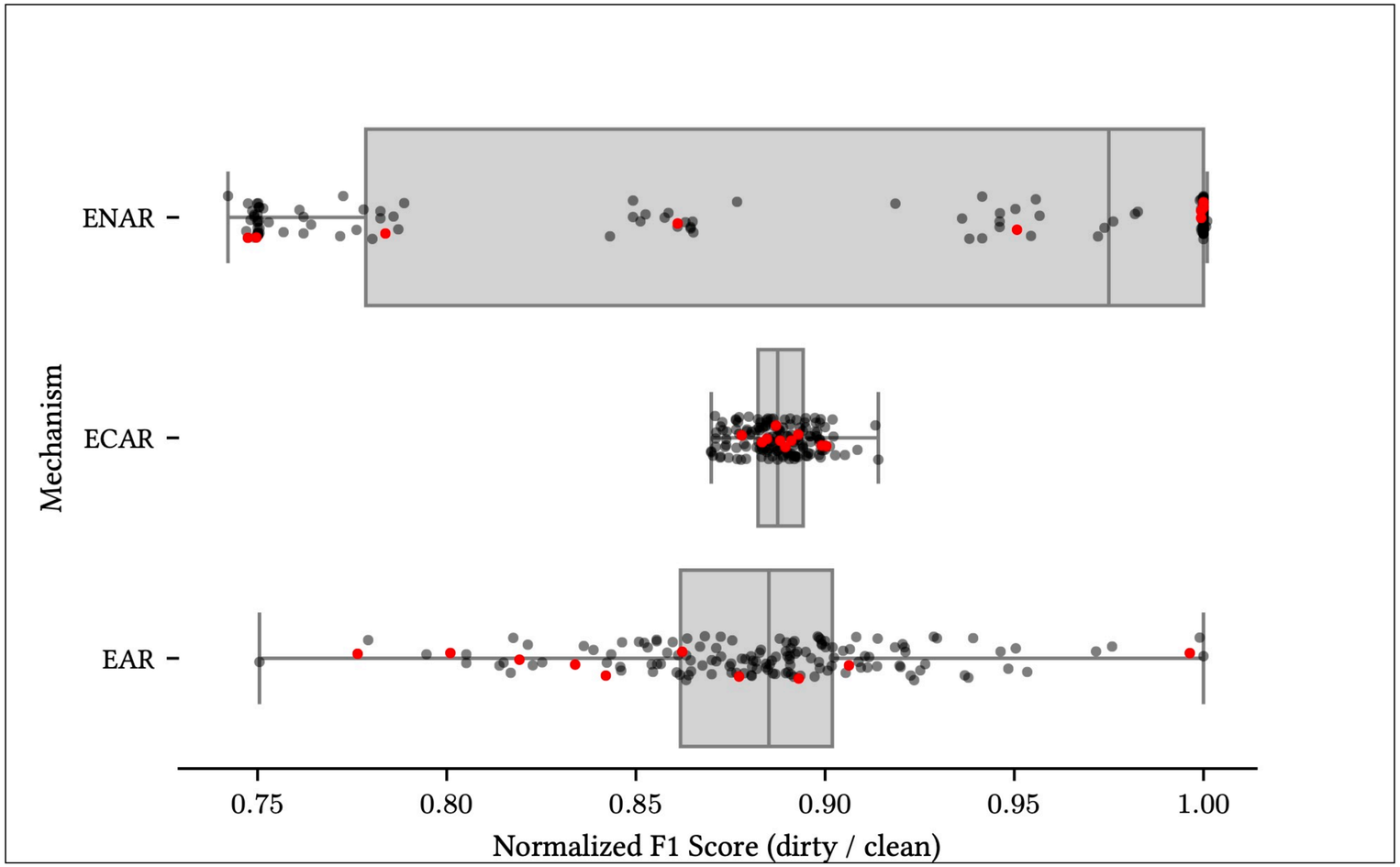
Impact of Realistic Errors on Data Cleaning Tools



https://github.com/calgo-lab/tab_err

Towards Realistic Error Models for Tabular Data

Impact of Realistic Errors on Downstream ML Performance



https://github.com/calgo-lab/tab_err

Thanks!