數據科學分析的土壤 ——
Trust AI & Data Quality Pipeline

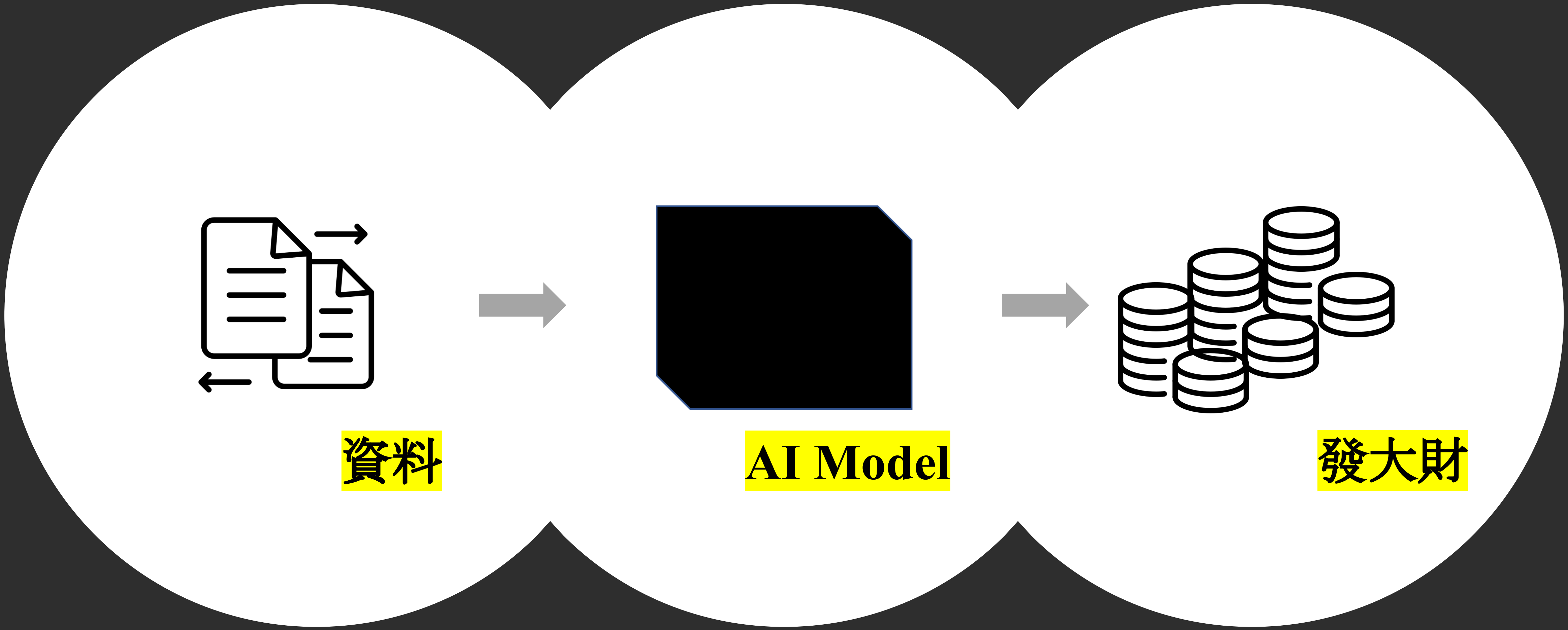
| 周成康. 陳瑾叡. 吳驊祐. 史康宇. 黃戎僔. 林孟璇
| 指導教授:李家岩
| 指導助教:陳彥彣. 方鈺學

TAIDQ

Table of Contents

BACKGROUND

The Imagined AI



The Real AI

- 資料爬蟲
- 資料庫
- 資料整併

資料

- 資料探勘
- 視覺化
- 特徵工程
- 訓練、驗證

資料科學

- 模型部屬
- MLOps

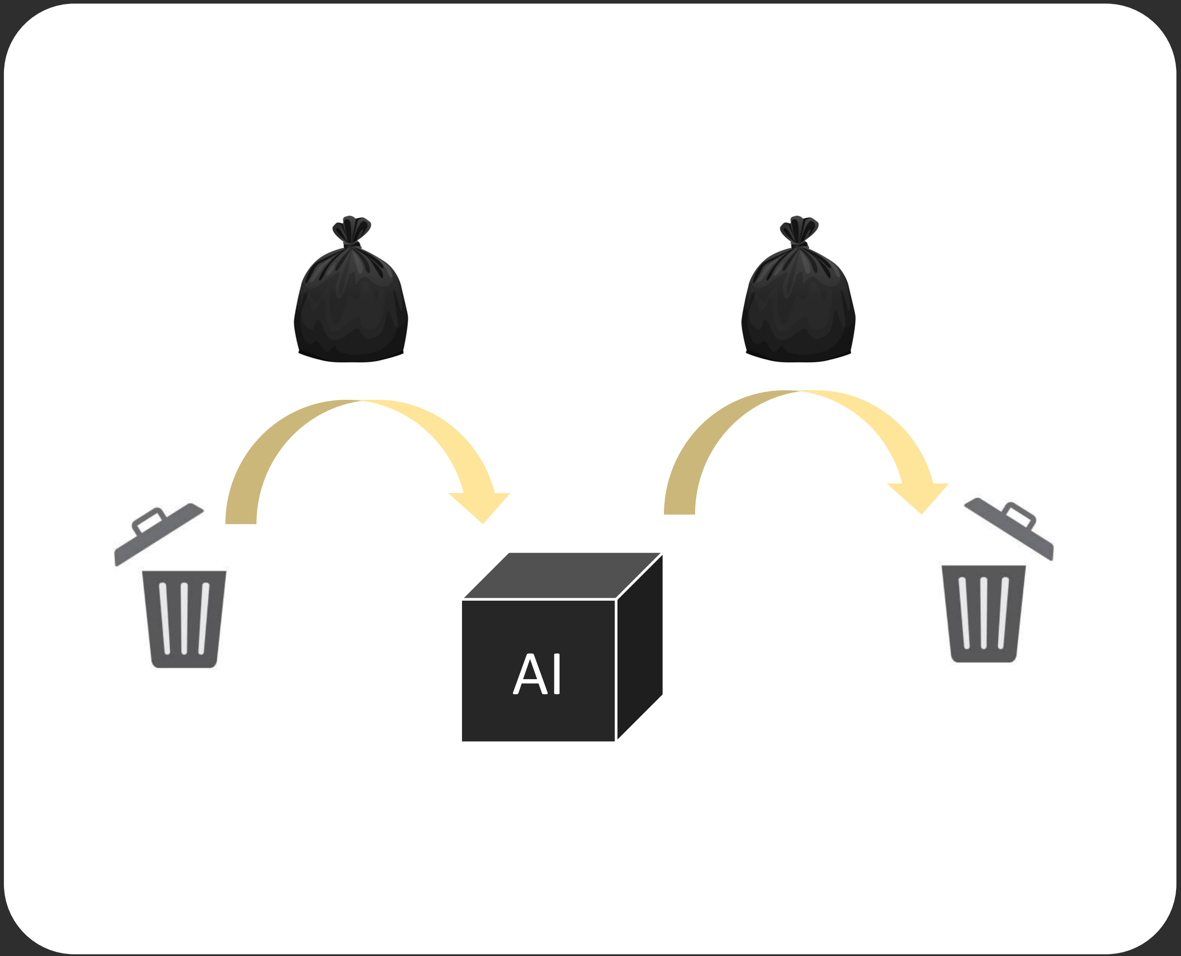
架構整理

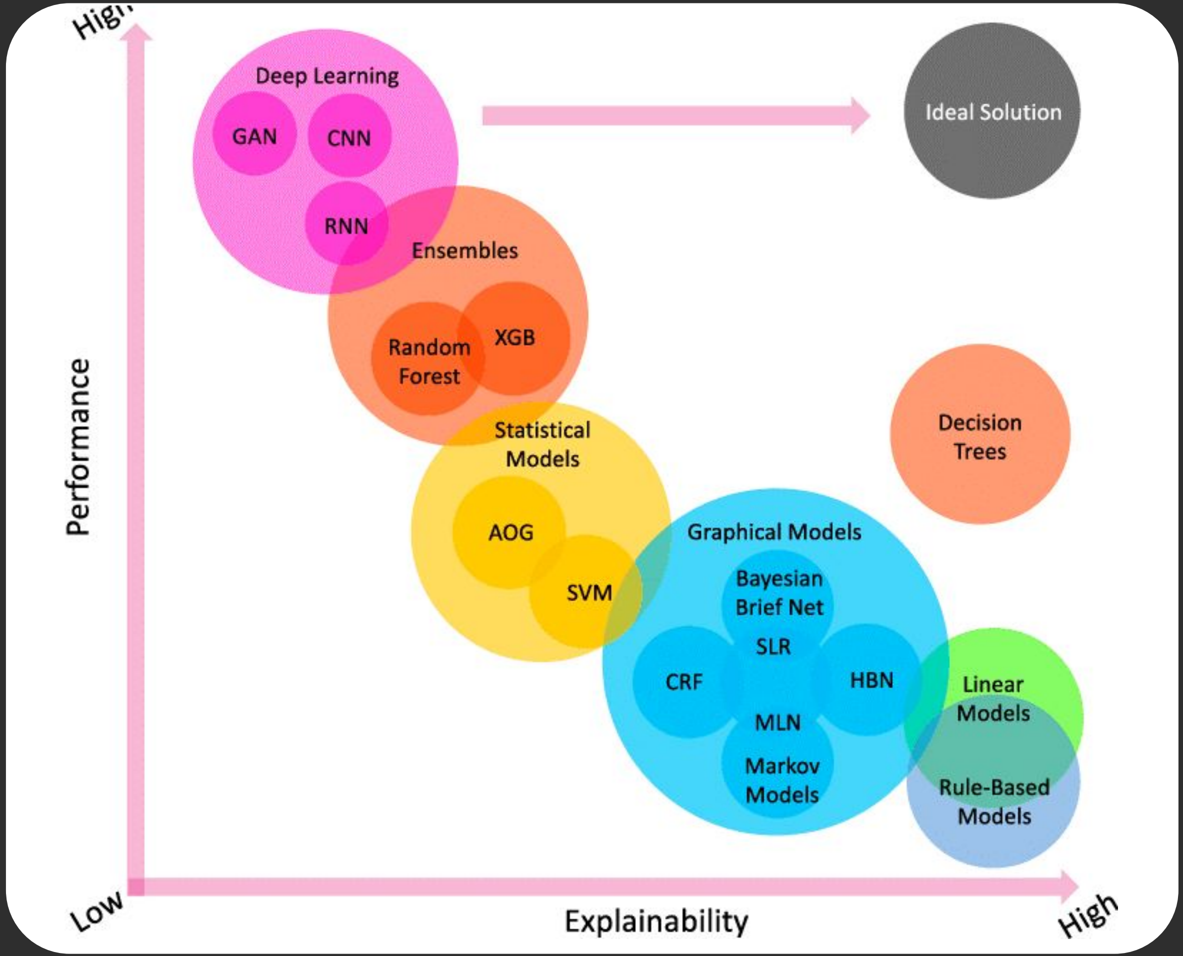
收入取決於…

?

個人造化

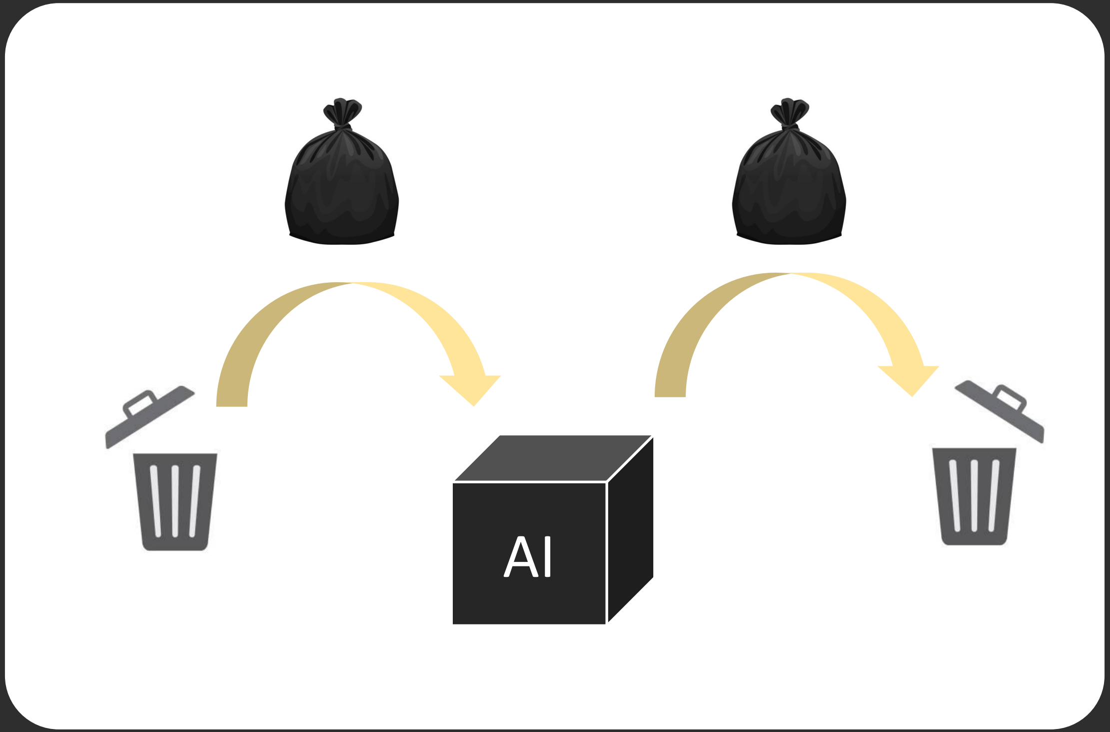
Dive into Details





GIGO: Garbage-in-garbage-out

- Wrong & Inappropriate Records
- Redundant Data
- Missing Records
- Insufficient Domain Knowledge

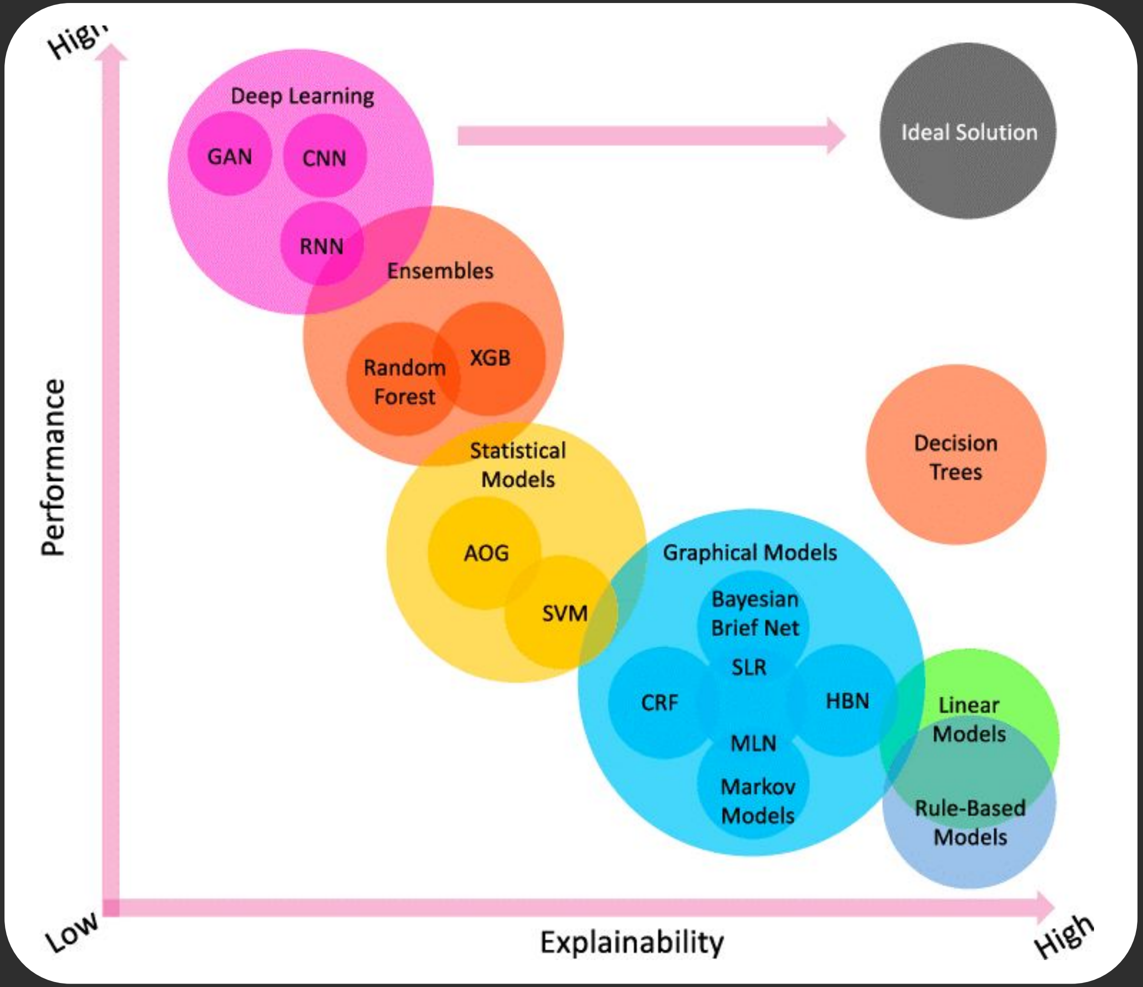


Dilemma: Performance vs Explainability

With optimal value, but lack optimal solution

Not Trustworthy Enough

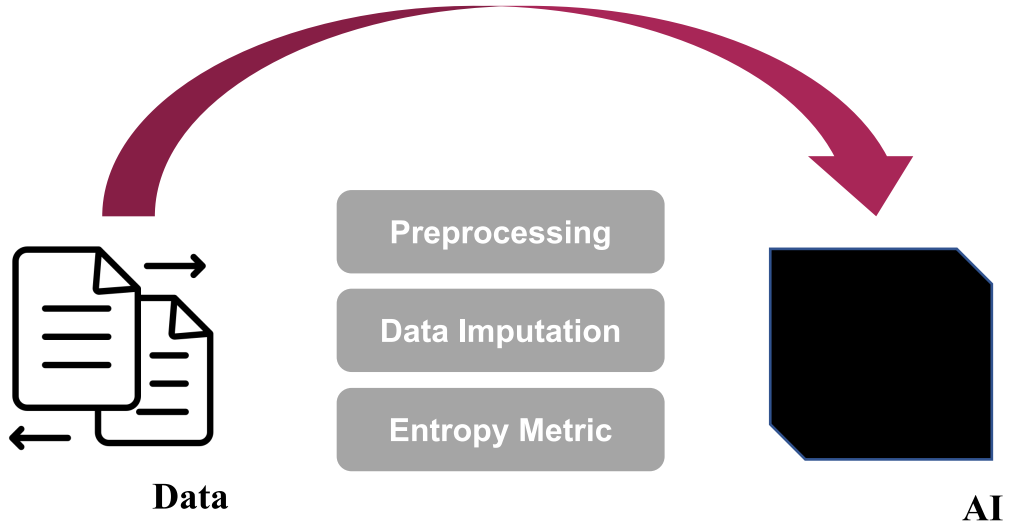
Ambiguous & Unfairness Mechanism



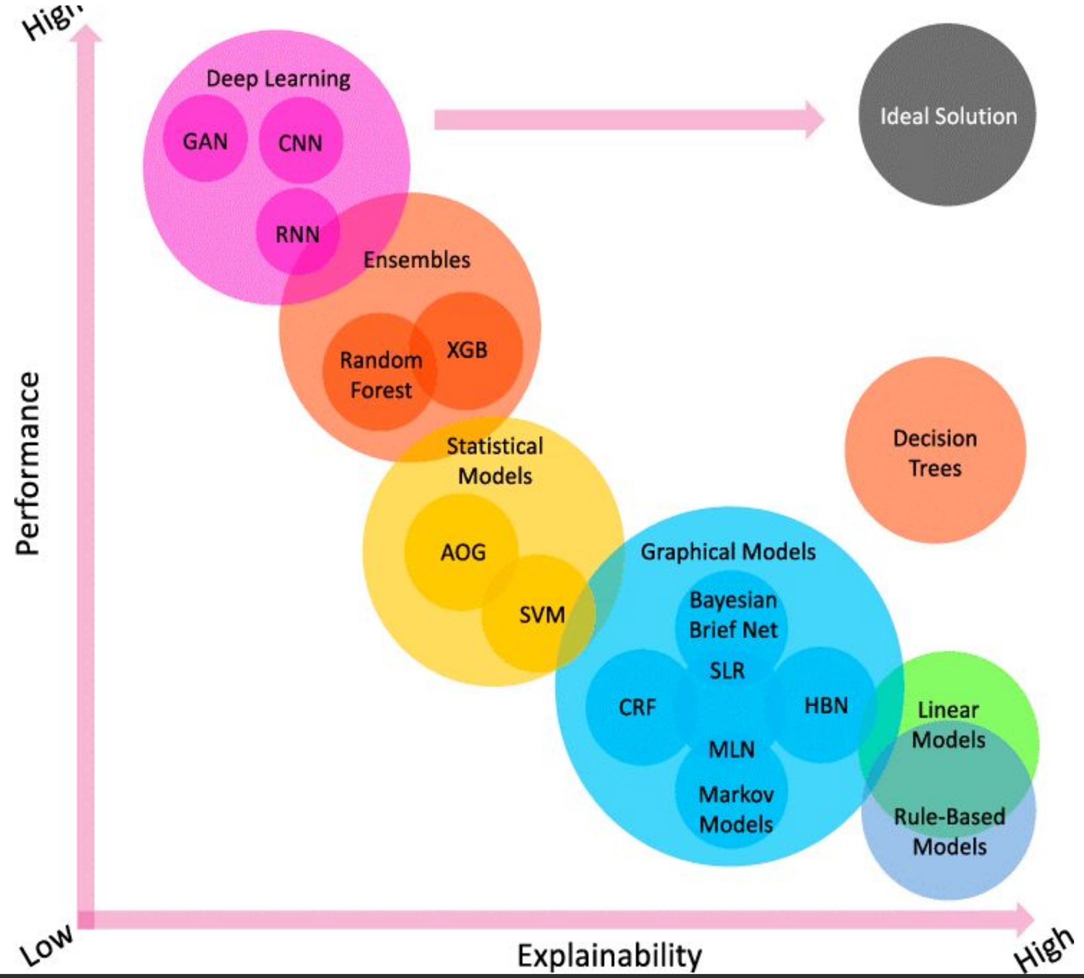
Arrieta et al., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI,"

OVERVIEW

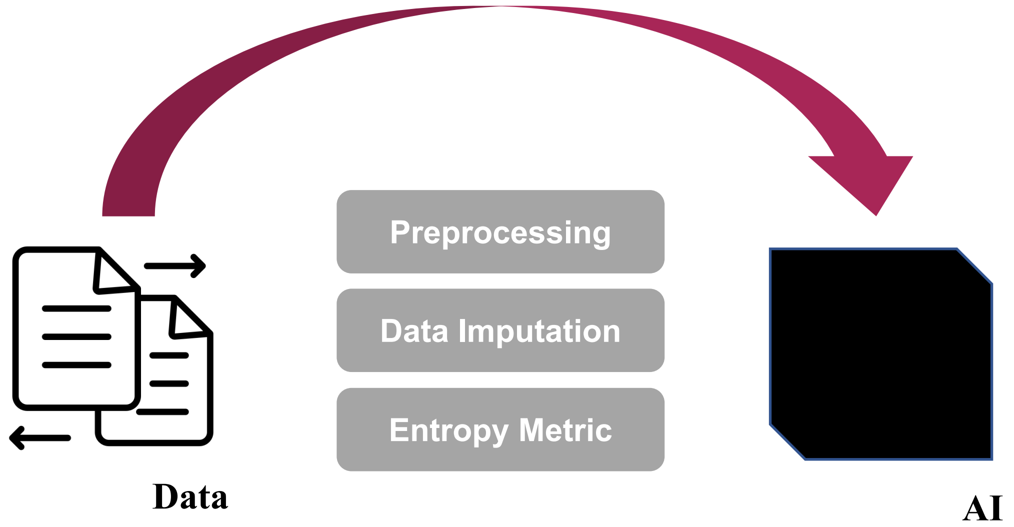
Solution: TAI&DQ Pipeline



Data Quality

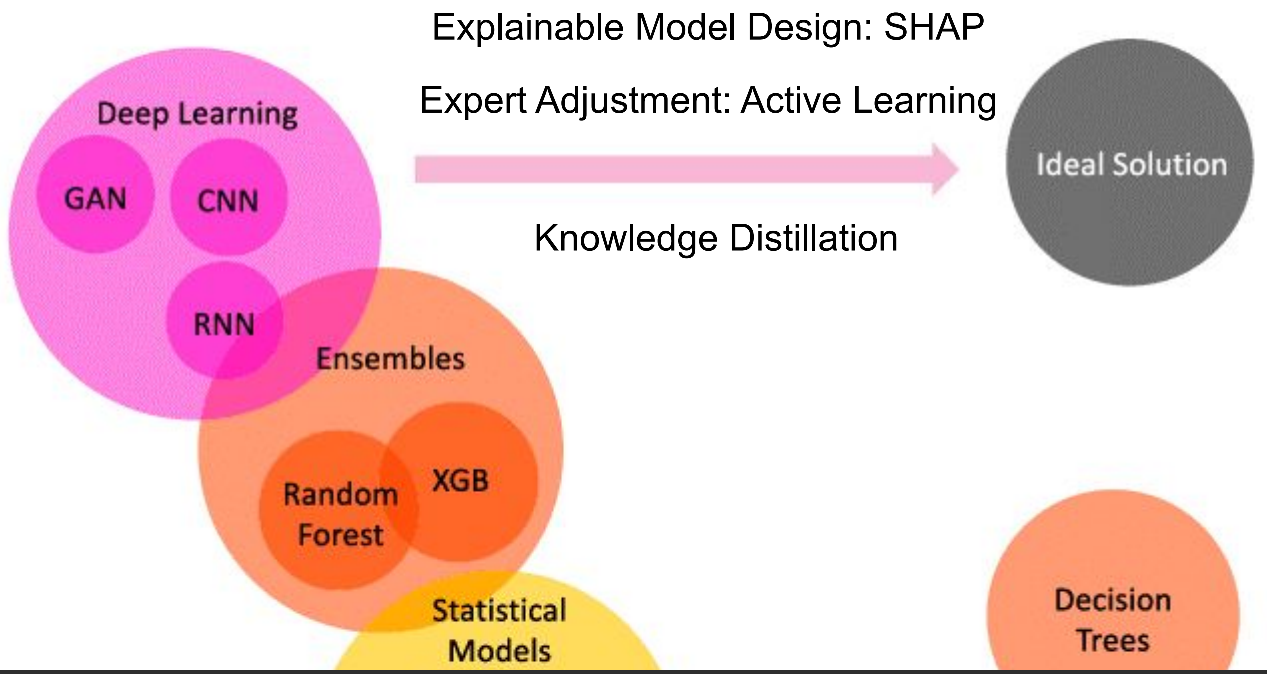


Solution: TAI&DQ Pipeline

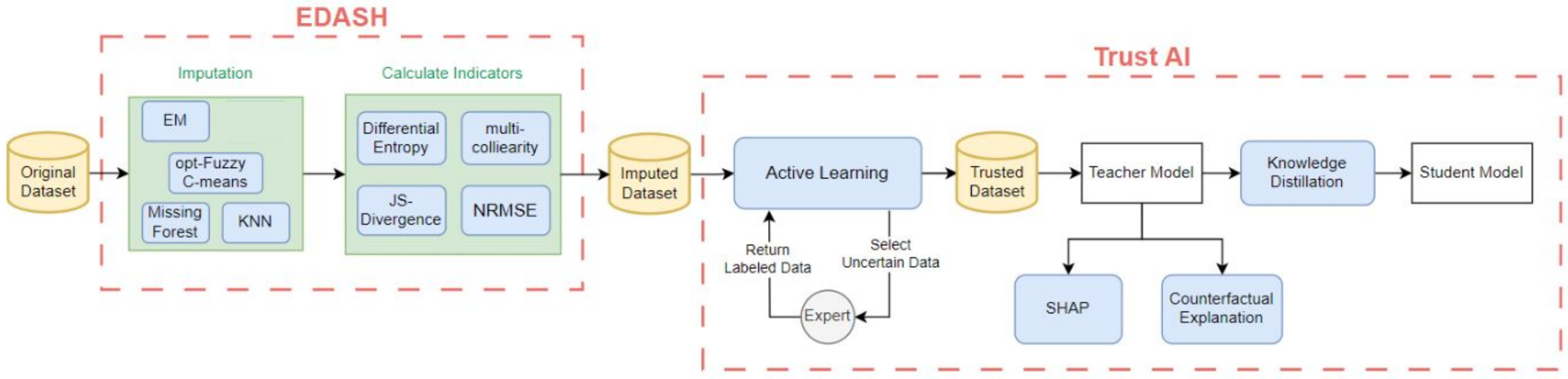


Data Quality

Trust AI



Solution: TAI&DQ Pipeline



Make the Black Box Transparent & Better Data

Trust AI & Data Quality

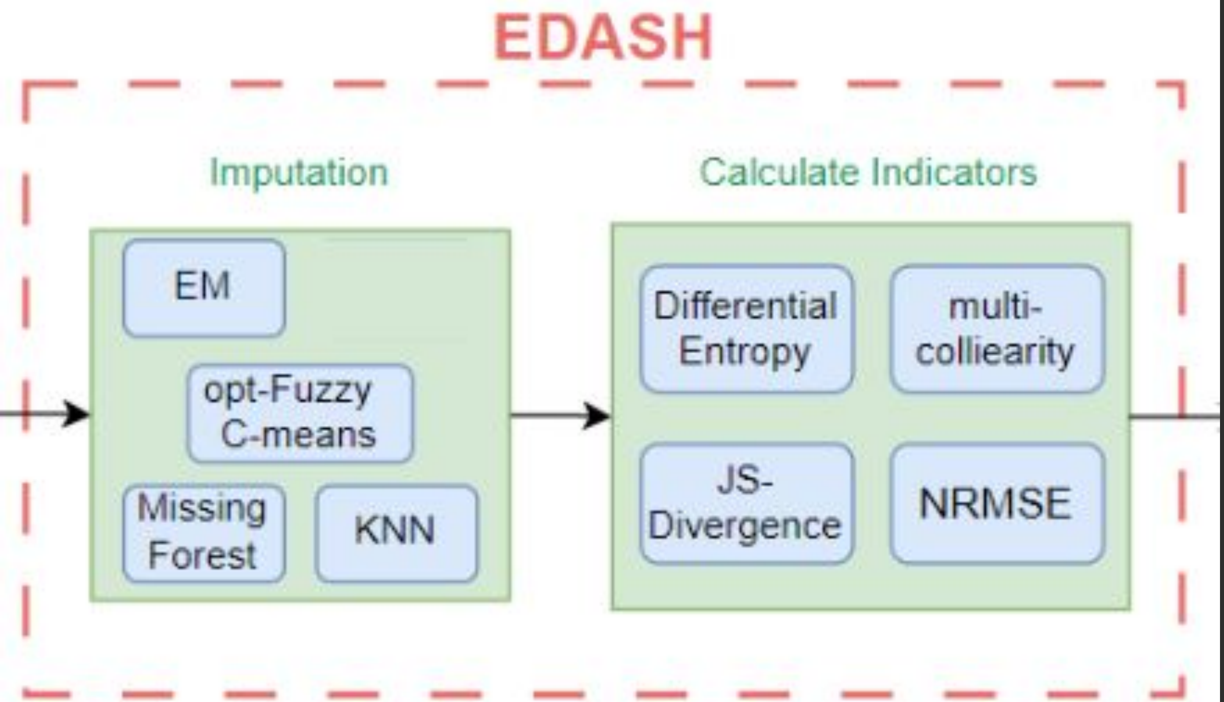
Dataset and Task Description

Gas Sensor Array Drift Dataset from University of California San Diego

Target: use chemical sensor signal to predict the class of gas

- 128 sensor signal(numerical) & Gas Class(categorical)
- 13910 records
- Classification Task
- Randomly simulating NaN values for imputation simulation

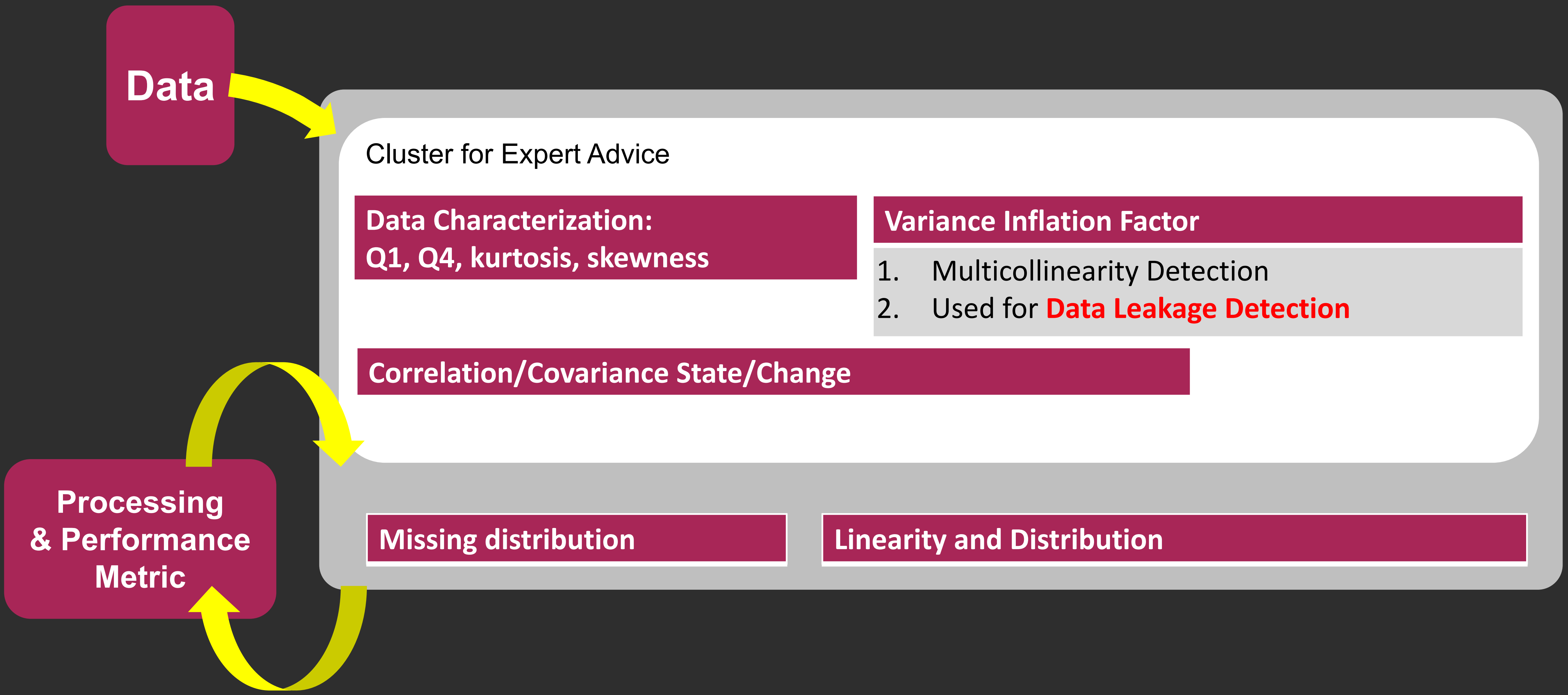
Gas Class	1	2	3	4	5	6
# of each class	2565	2926	1641	1936	3009	1833



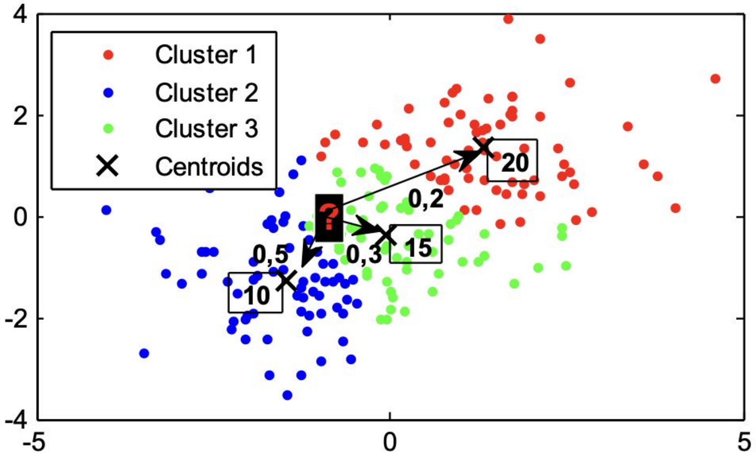
Module 1
EDASH
(EDA Dashboard)

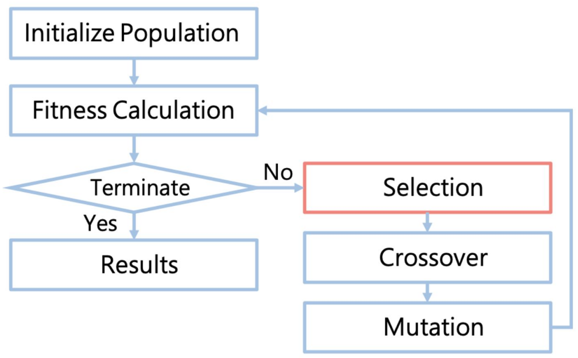
- Data Profiling
- Imputer Methods
- Data Quality

Data Profiling for Data Characterization



Imputer Methods

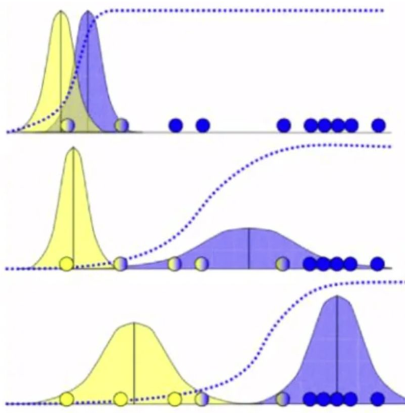


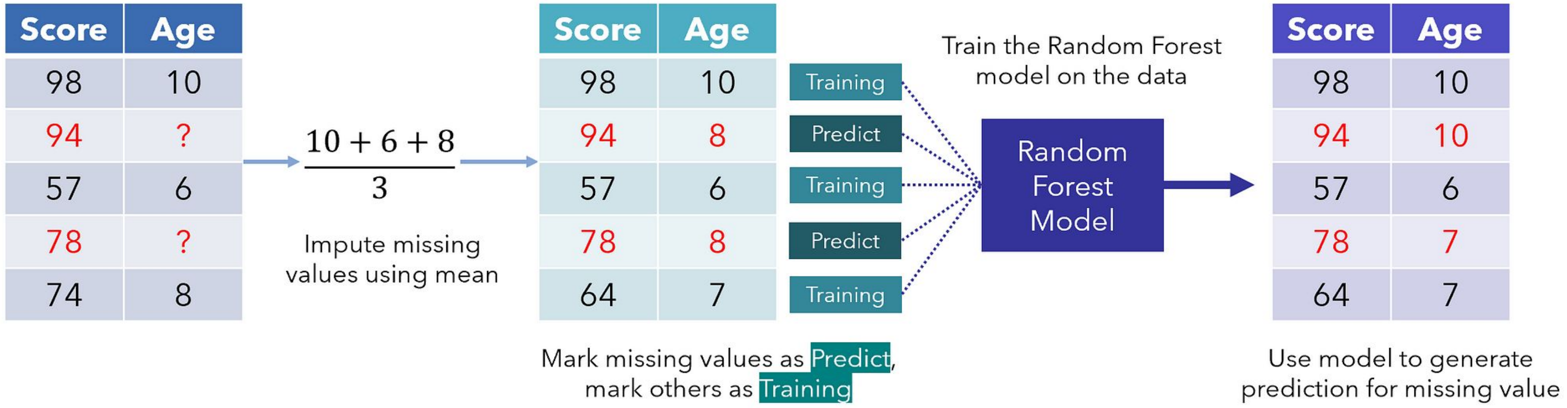


$$\{\theta^0, \alpha^0, \beta^0\} \rightarrow \text{E step: } \{M^0, N^0\} \rightarrow \text{M step: } \{\theta^1, \alpha^1, \beta^1\} \rightarrow \text{E step: } \{M^1, N^1\} \rightarrow \ldots$$

From the basis of **Statistical Learning & ML,** derives different imputation

Suitable for different situations.





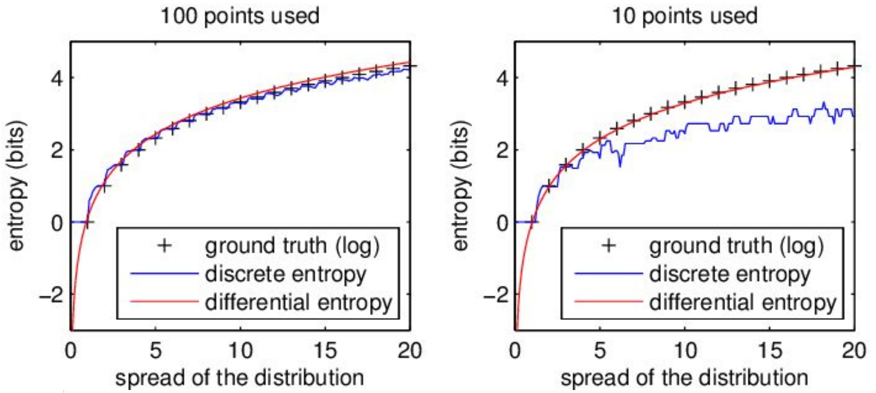
Data Quality

This step is for data or imputer selection.

$$H(X) = -\sum_{x \in X} P(x) \cdot \log P(x)$$

$$h(X) = -\int_{\mathcal{X}} f(x) \log f(x) dx$$

$$\frac{1}{2} \log \left(2\pi e \sigma^2\right)$$

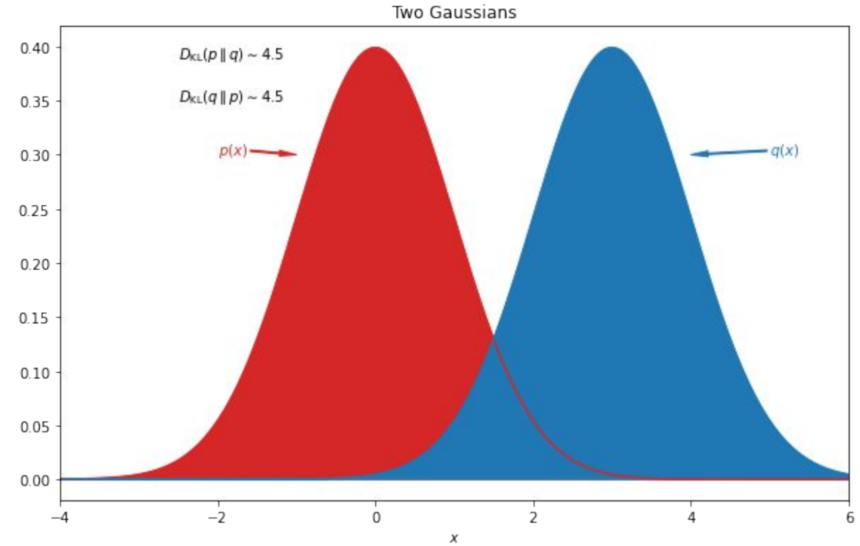


- Stable even with few data points
- Can be used for both numerical and categorical data type.
- Run faster in canonical form, without data points simulation

Profiling module | Processing

$$D_{KL}(P\|Q) = \frac{1}{2}\left(\ln\left(\frac{\sigma_Q^2}{\sigma_P^2}\right) + \frac{\sigma_P^2}{\sigma_Q^2} + \frac{(\mu_P - \mu_Q)^2}{\sigma_Q^2} - 1\right)$$

$$D_{JS}(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M) \quad \text{,where } M = \frac{1}{2}(P+Q)$$



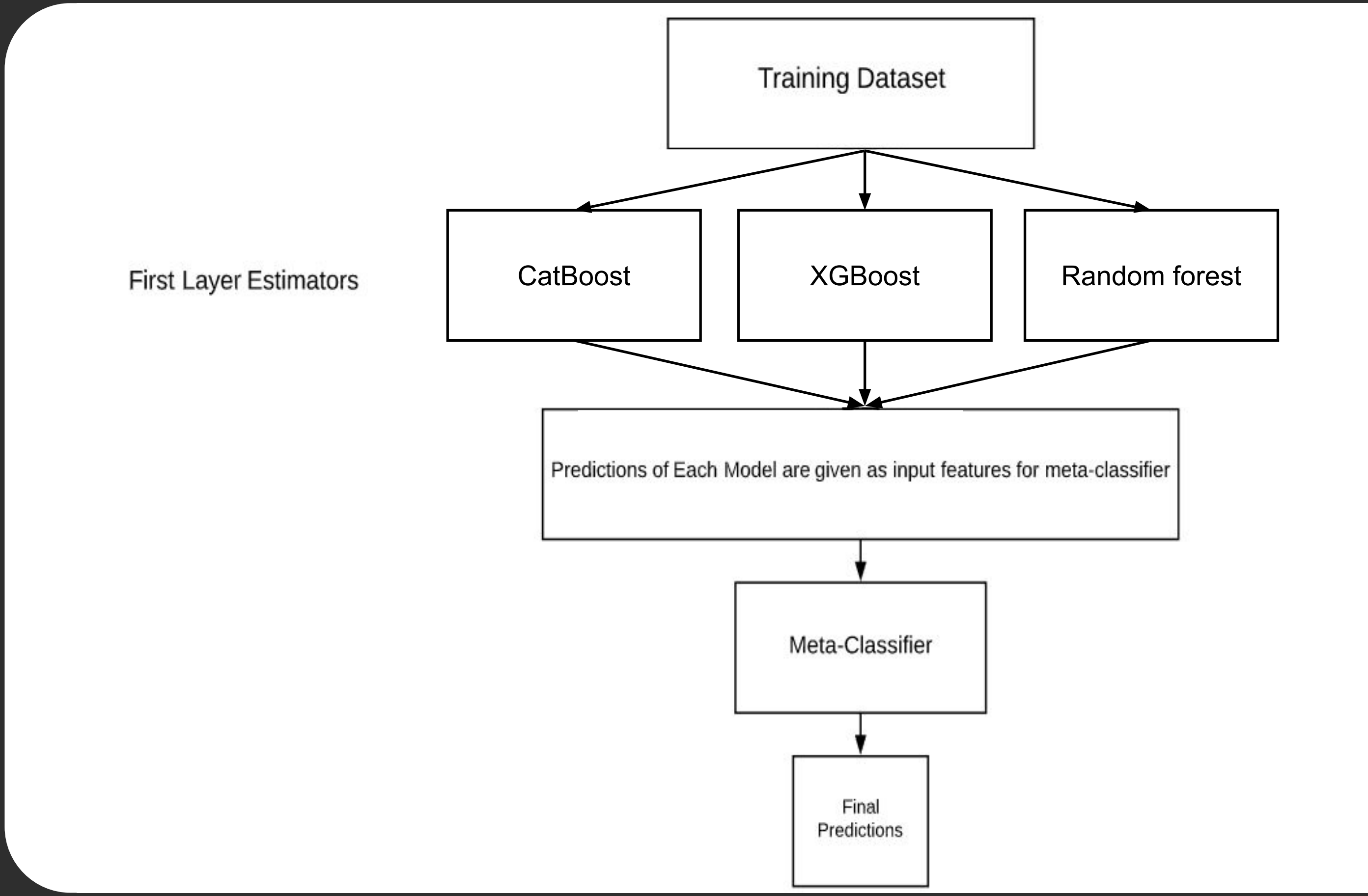
Stability & Performance Evaluation

Profiling module | Processing

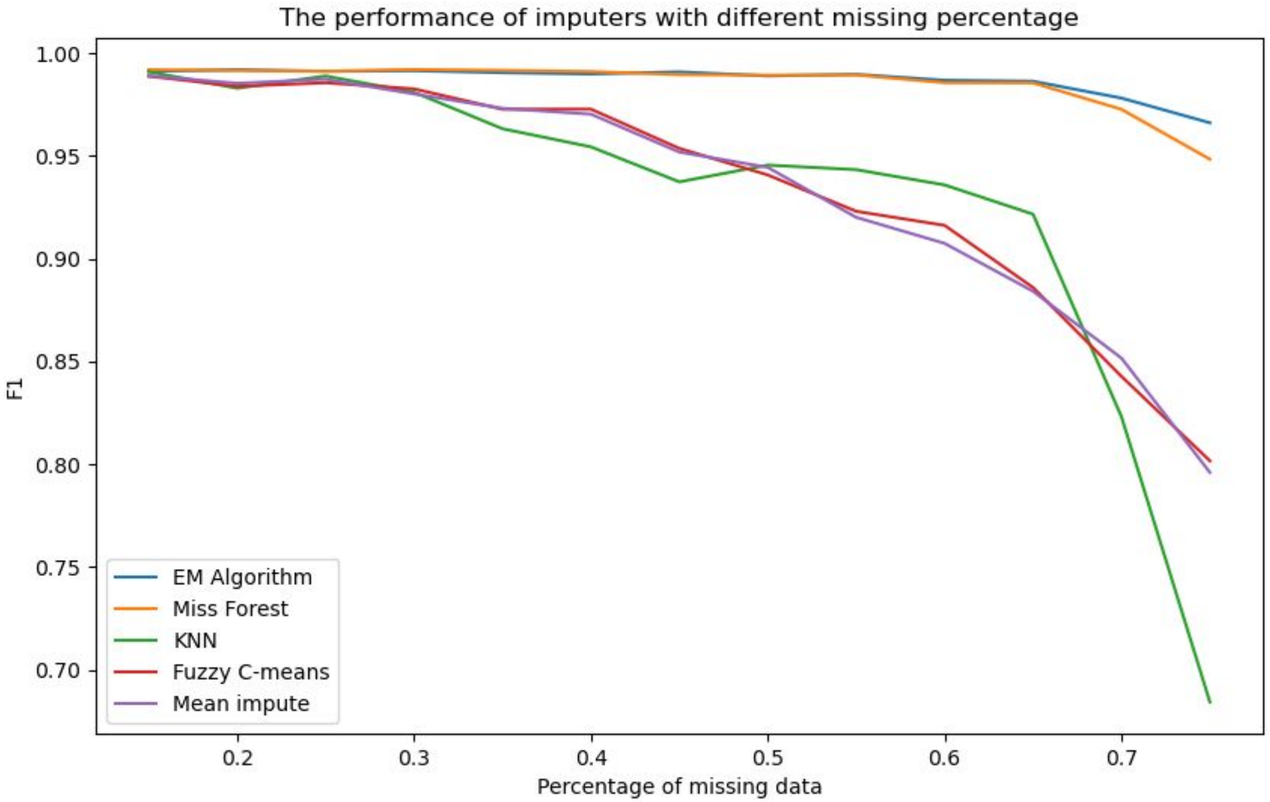
Stability Evaluation

Performance Evaluation

- Able to hold the nonlinearity
- Generalizability by Stacking ensemble
- Explainability to some degree

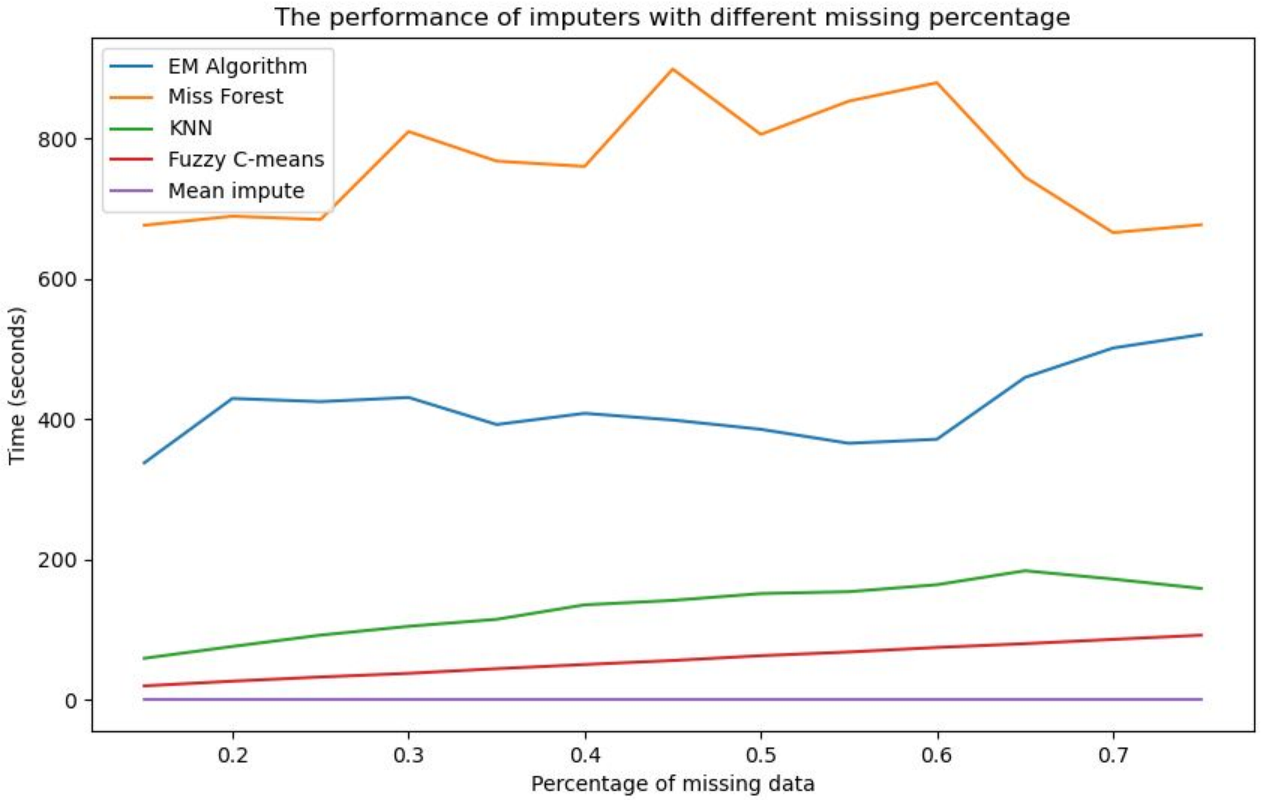


Experiments - Performance



- Best Results : EM & MissForest
- KNN performance drops when missing rate is high

Experiments - Time



- EM & MissForest are time consuming
- It’s better to use them when large missing rate.

Conclusion

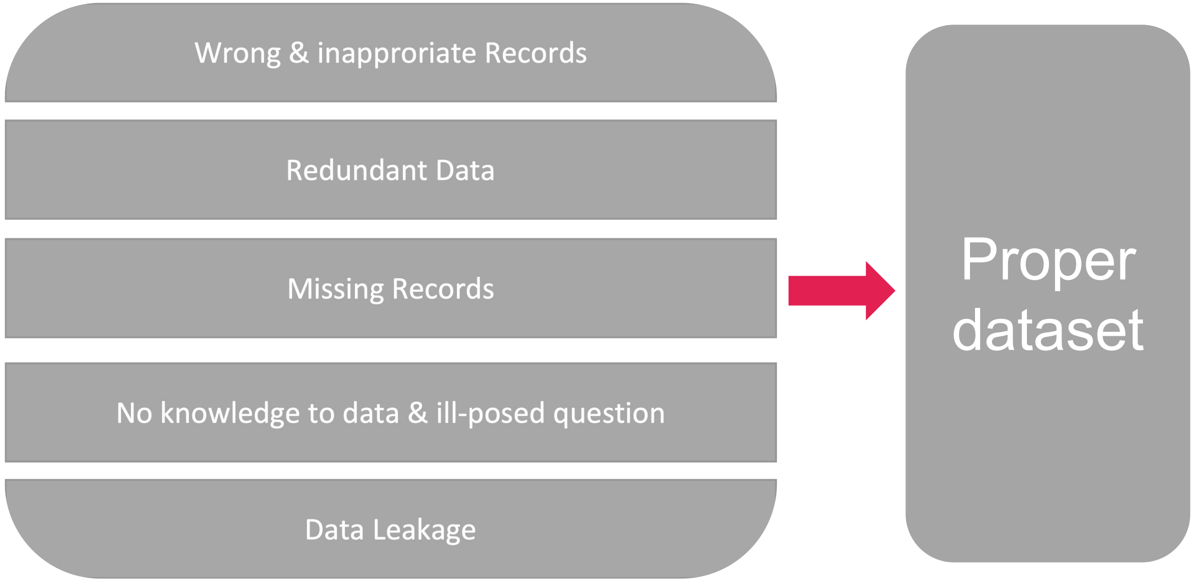
What do we get

- Raw dataset → Knowledge beforehand of dataset characterization and detection
- Missing dataset → complete dataset with information and characterization preserved.
- Provide stability/accuracy detection of 2 datasets.

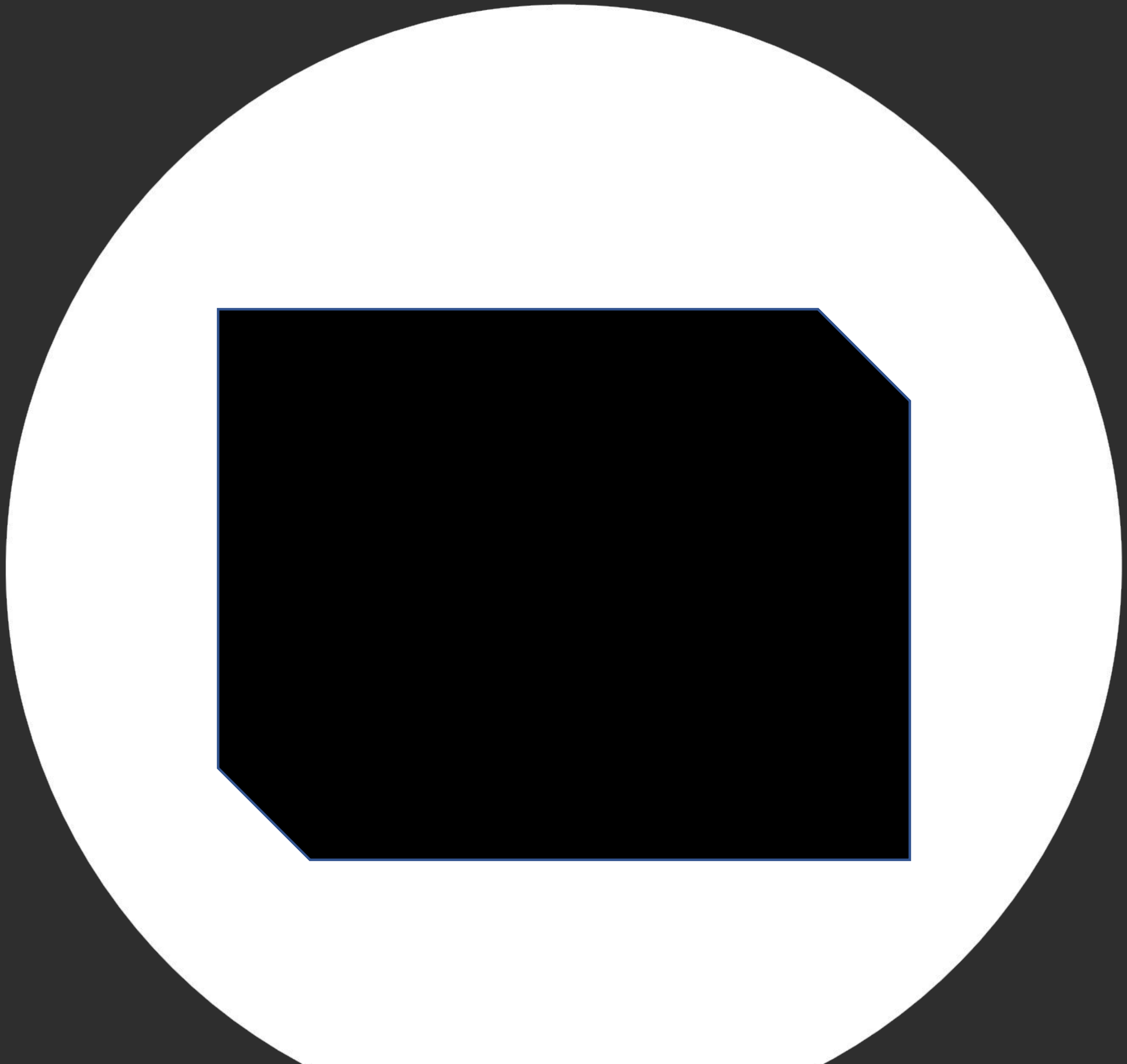
⇒ SOP for data quality & proper dataset

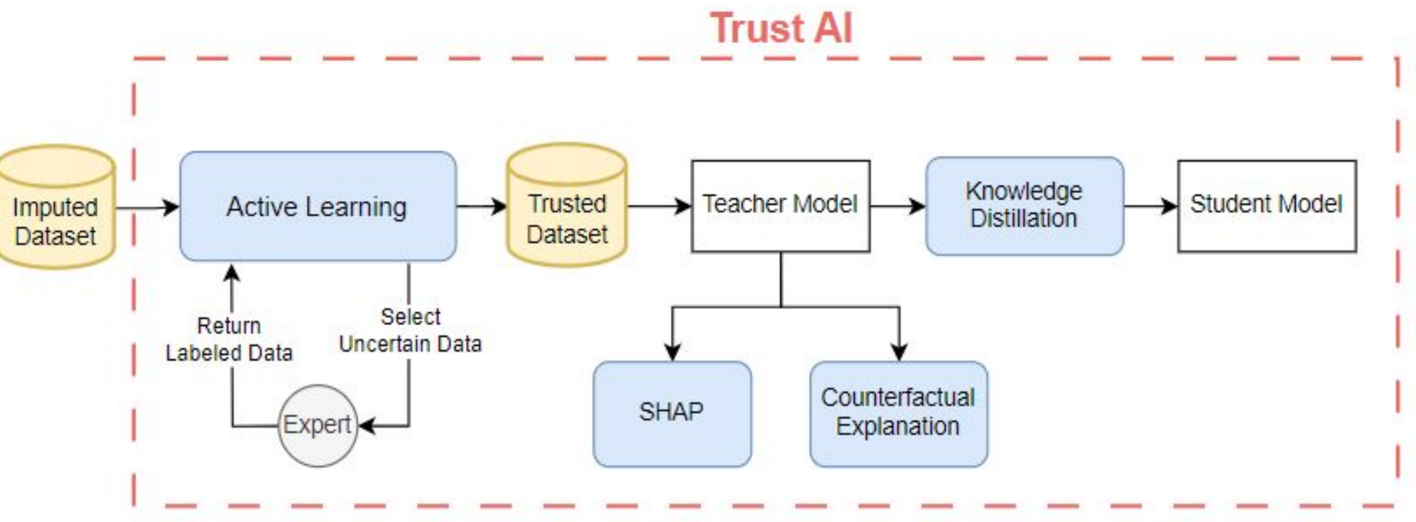
What does it mean

- Reduce economical & time cost
- Enhance performance of model
- Enhance understanding of dataset and task design



Now, let’s open the black box





Module 2
Trust AI

- Active Learning
- Knowledge Distillation
- SHAP Explanation

Active Learning

Query the Oracle → Uncertainty Elimination

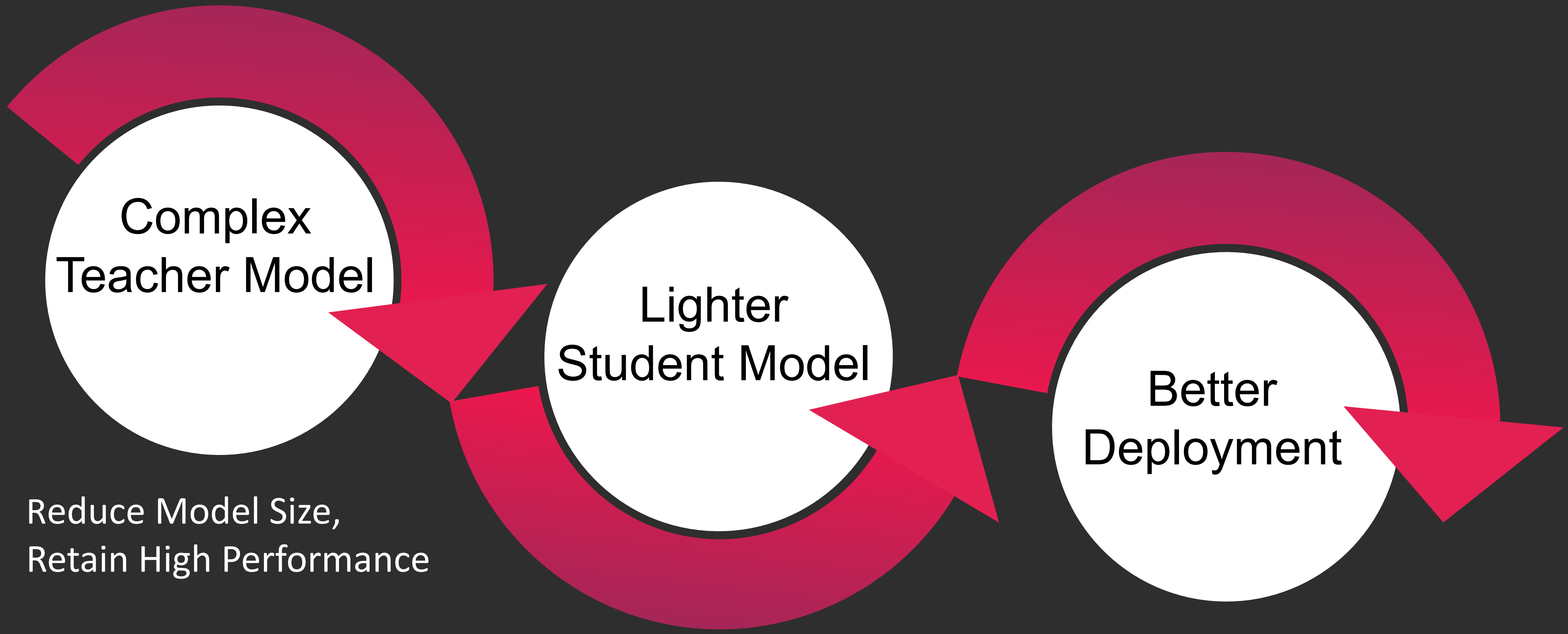
Pool-Based Sampling → Avoid Selection Bias; Faster Training Process

Knowledge Distillation – Trust AI

1. Previous Model with Most Info
2. Correctness-Consistency
3. Efficient Subsequent Data Acquisitions
4. Low Cost, High Valued Training Set
5. Better Scalability and Maintainability

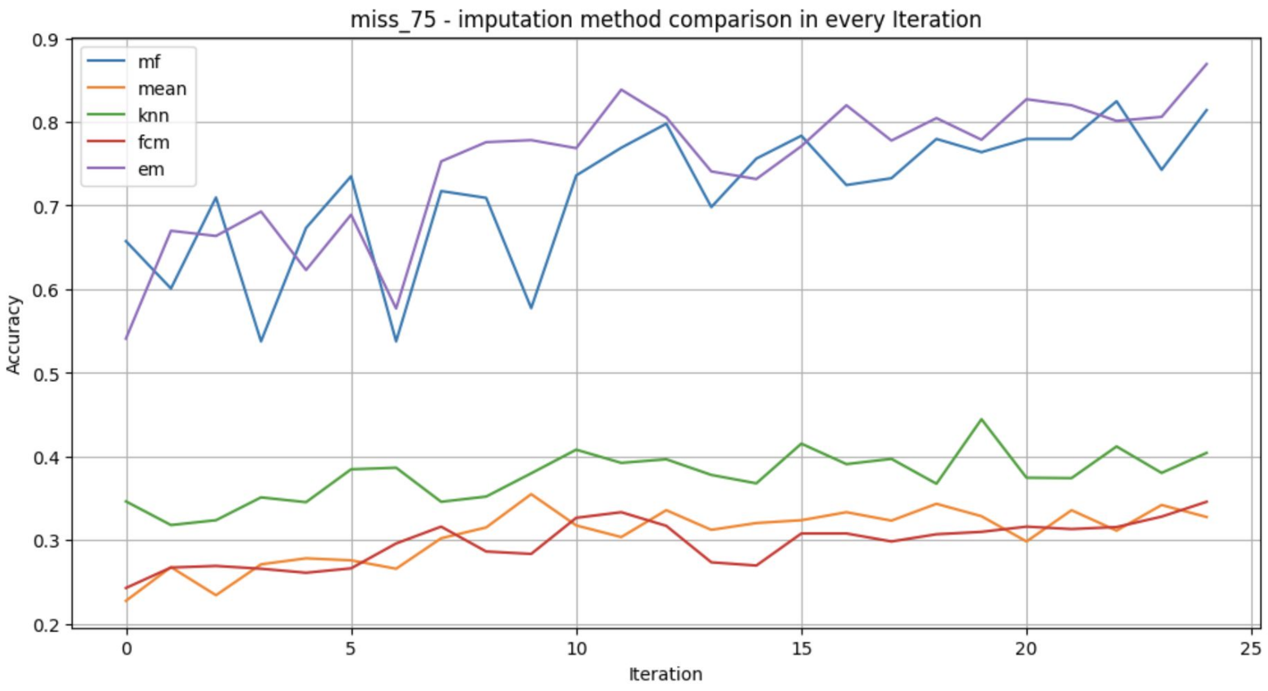
Kwak et al., “TrustAL: Trustworthy Active Learning using Knowledge Distillation”, 2022

Knowledge Distillation - Compression



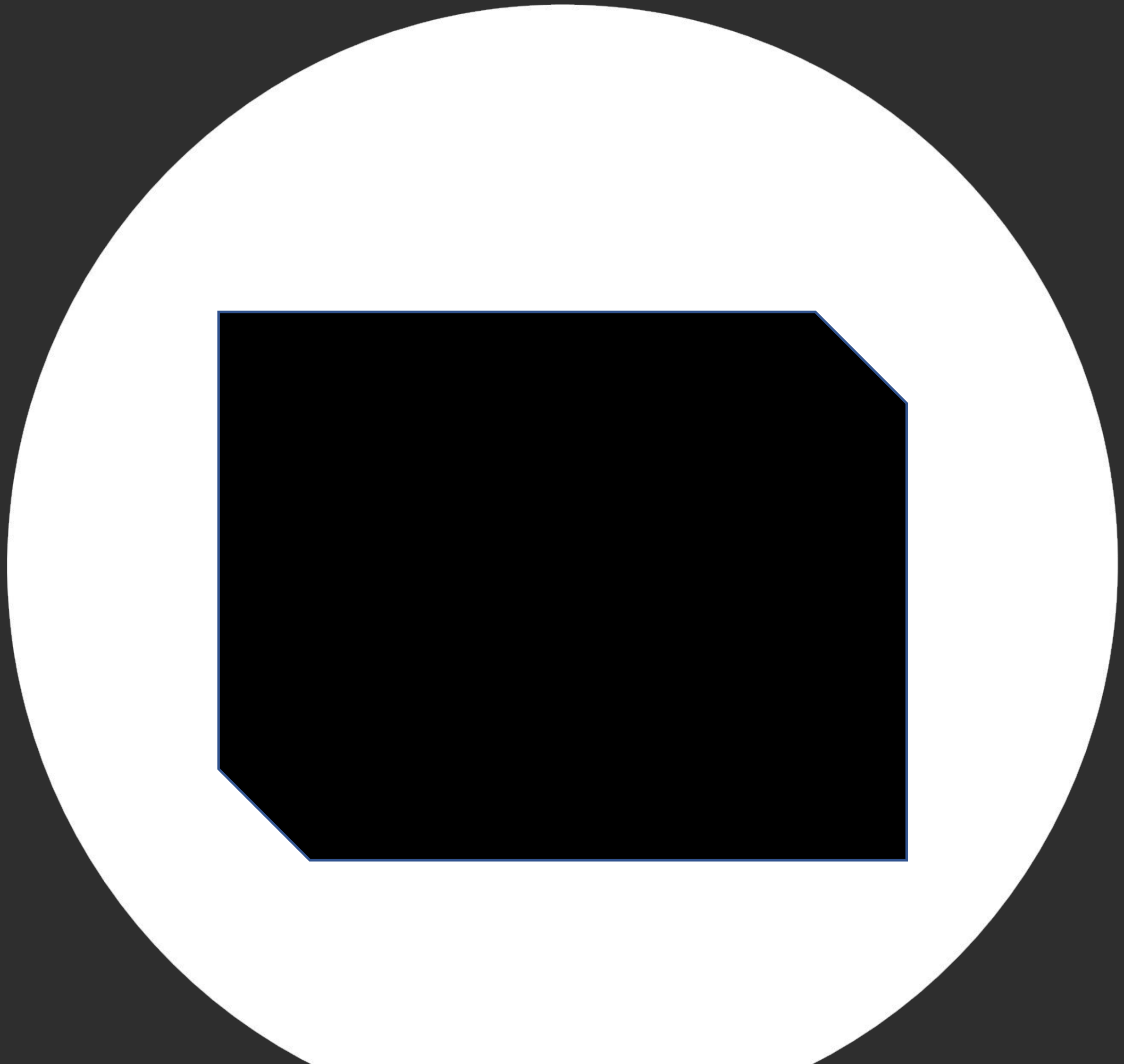
Reduce Model Size,
Retain High Performance

Experiments



- KNN drops with higher missing rate
- EM have most benefit with increasing iterations.

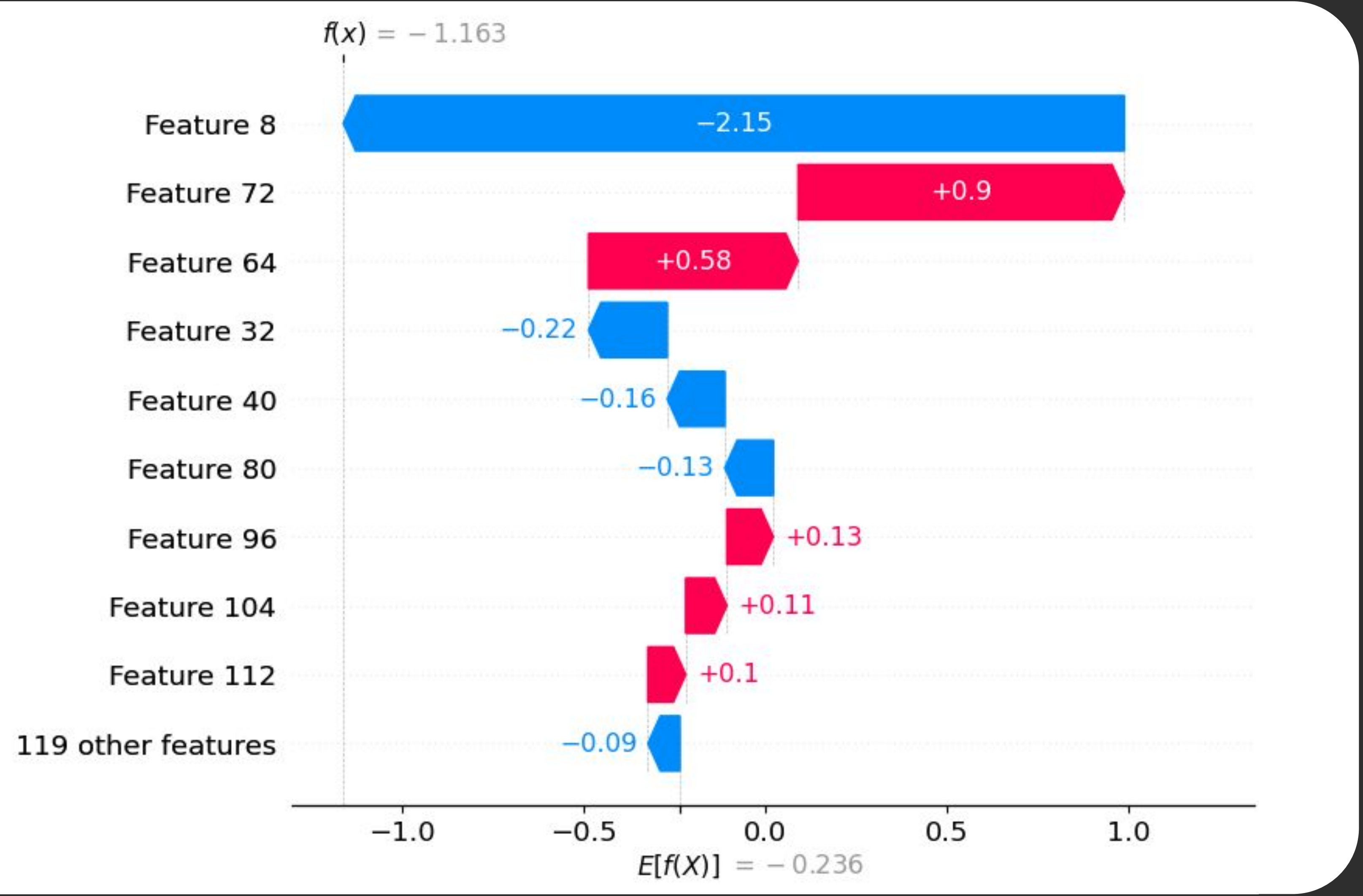
Now, let’s REALLY open the black box



SHAP

- Complicated models are difficult to understand intuitively.
- SHAP opens the black box.

SHAP



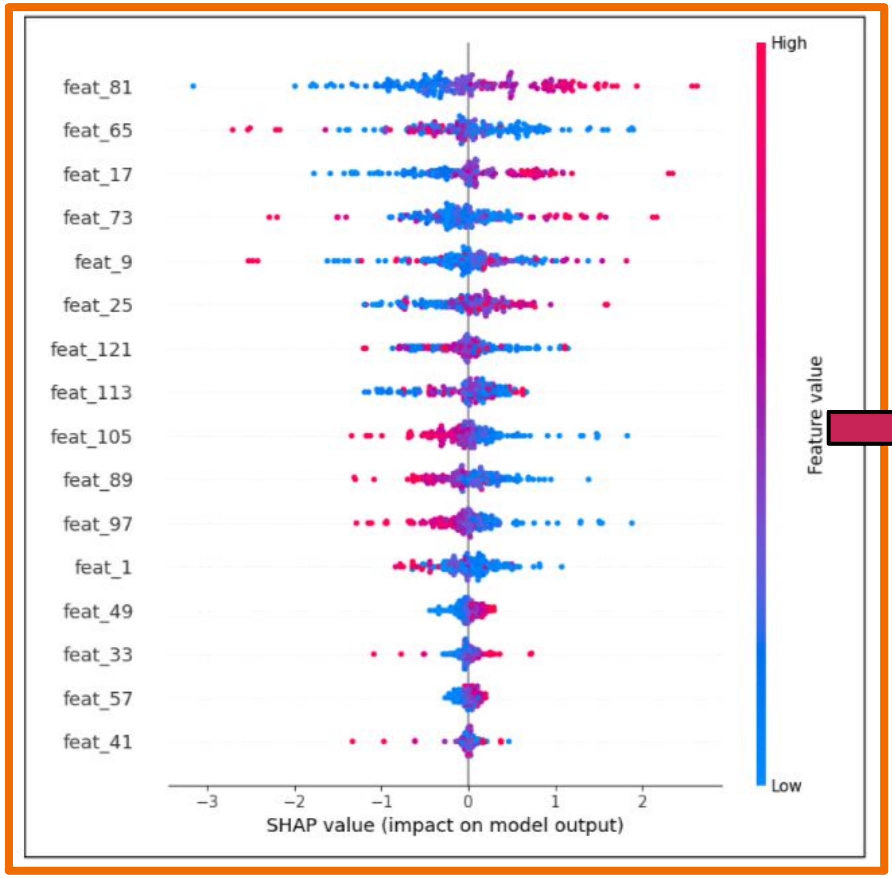
SHAP is like evaluating the contribution of each member in a project.

Predicted value: calculated by the sum of contributions of features.

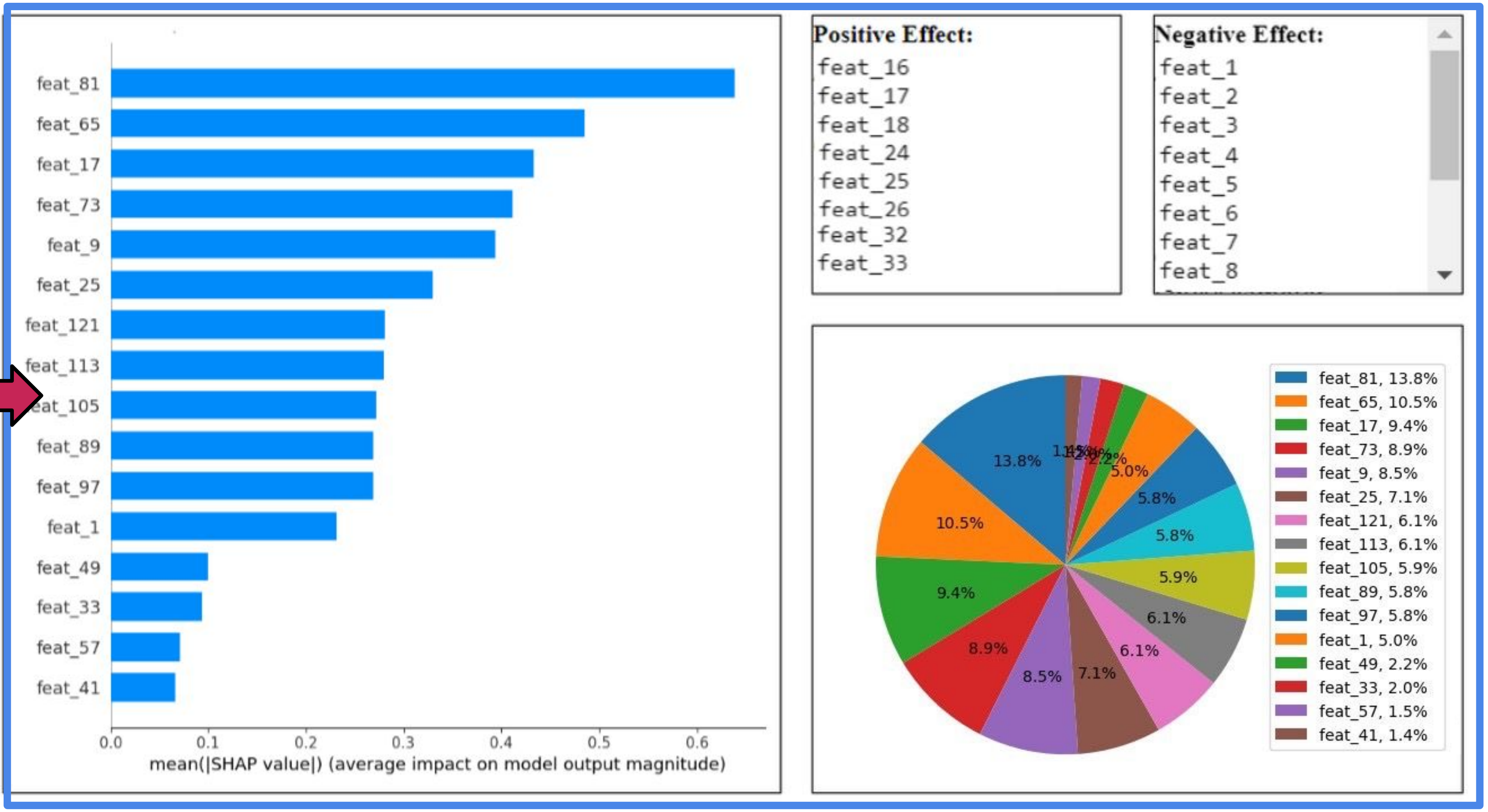
The sum of contributions of features results in the predicted value of target label.

One Class Global SHAP

- One class global SHAP



- Divide to multiple figures



Counterfactual Explanation

Decision-Making Support

```
Query instance (original outcome : 1)
```

5	feat_116	feat_117	feat_118	feat_119	feat_120	feat_121	feat_122	feat_123	feat_124	feat_125	feat_126	feat_127	feat_128	Gas_Class
9	34.573879	39.702694	-7.137493	-12.199501	-15.406119	25665.841797	7.617512	11.160269	27.769402	31.185884	-6.301726	-9.816422	-13.670685	1

```
Diverse Counterfactual set (new outcome: 5)
```

115	feat_116	feat_117	feat_118	feat_119	feat_120	feat_121	feat_122	feat_123	feat_124	feat_125	feat_126	feat_127	feat_128	Gas_Class
169	34.573879	39.702693	-7.137493	-12.199501	-15.406119	25665.8416	7.617512	12.717600	27.769402	31.185885	-6.301726	-9.816422	-13.670685	5
169	34.573879	39.702693	-7.137493	-12.199501	-15.406119	25665.8416	7.617512	11.160269	27.769402	31.185885	-6.301726	-9.816422	-13.670685	5

DEMO

Business Model

關鍵合作夥伴

企業的數據分析團隊

價值主張

- 增進資料品質
- 維持高價值資料
- 提升模型效率
- 低成本高效訓練
- AI決策可解釋性

目標客群

- 有大量數據的產業, 如製造業
- 資料品質欠佳的公司
- 問卷調查發行者
- 渴望使用AI輔助決策的企業

關鍵活動

- 方法論實驗
- 平台建設
- 可解釋性實測

關鍵資源

- 開發人員
- 平台建設
- 方法論研究者
- 平台部署資源

通路

- 網頁平台
- 程式套件

顧客關係

- 提供企業方便的工具
- 開源成套件供開發者社群使用
- 建設成網站平台供使用者體驗

成本結構

平台維護

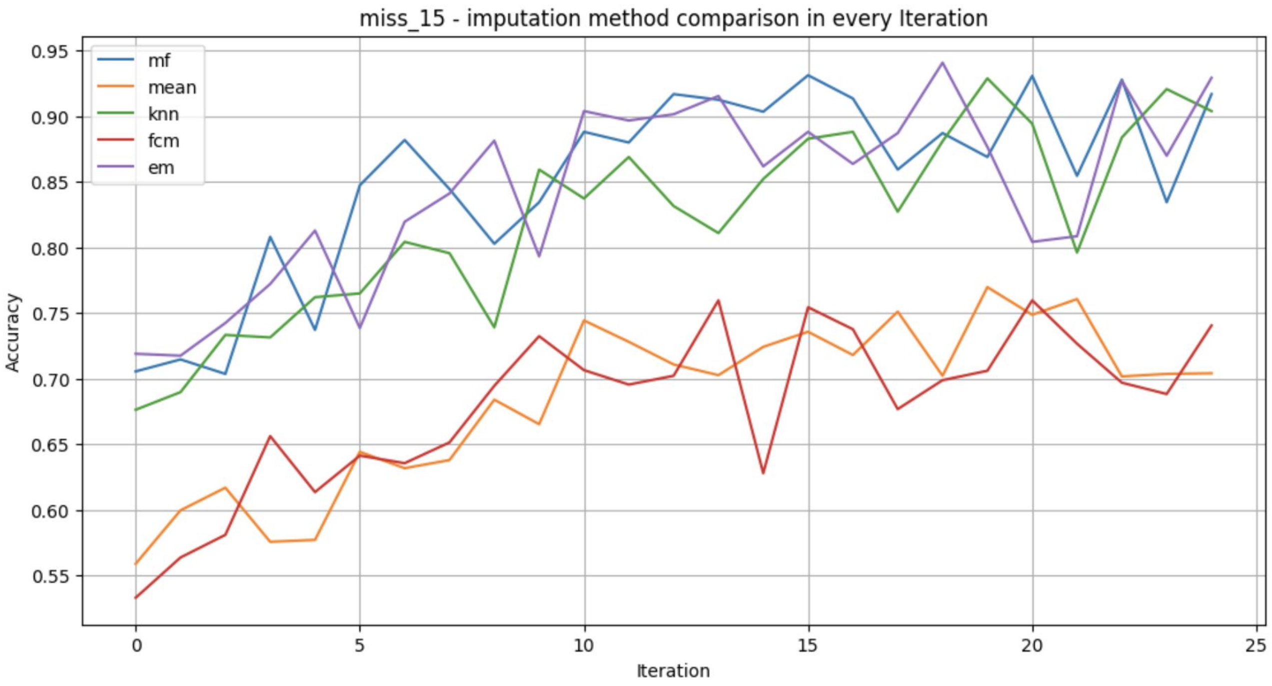
收益流

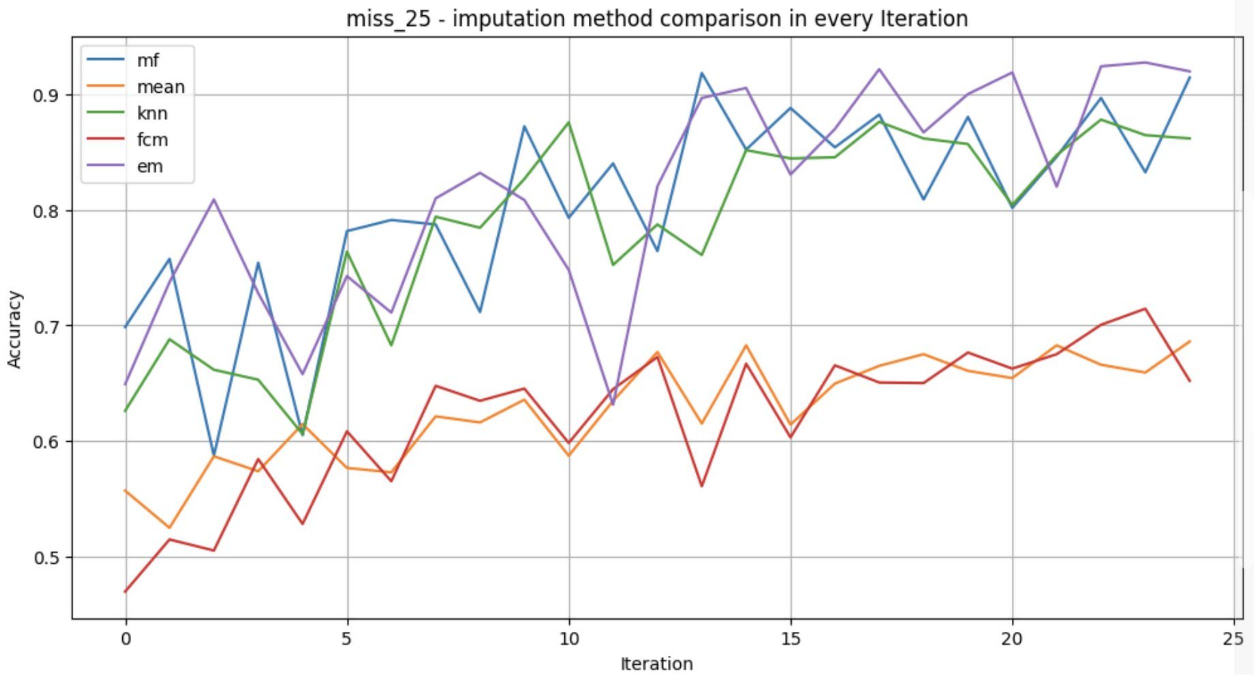
使用者費用

TAIDQ

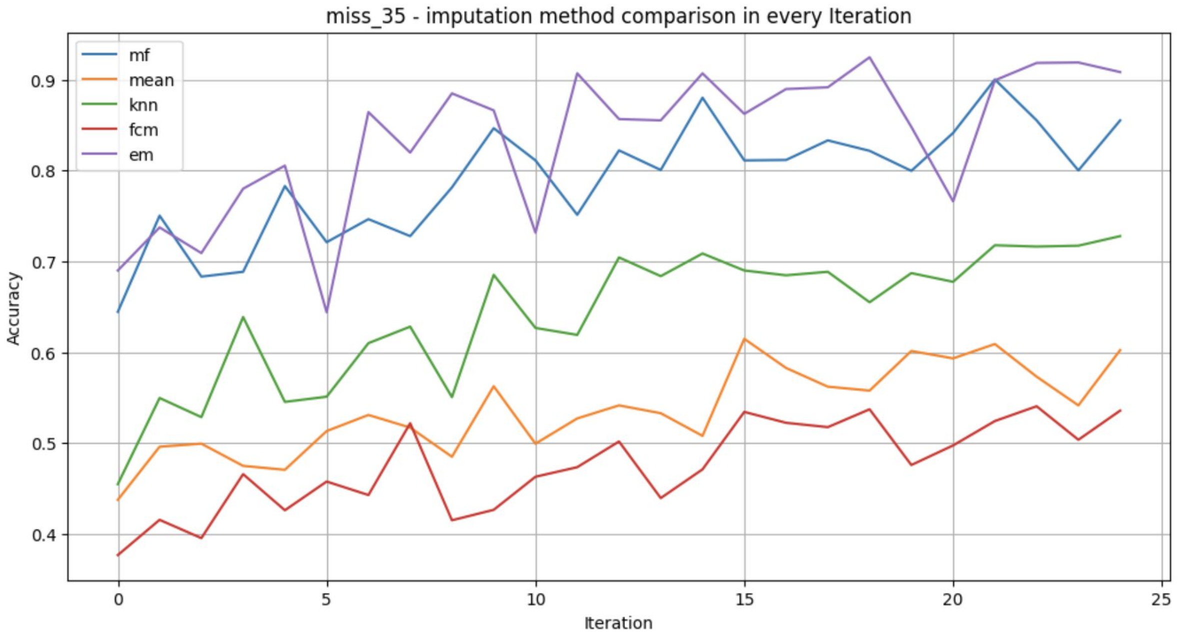
for listening

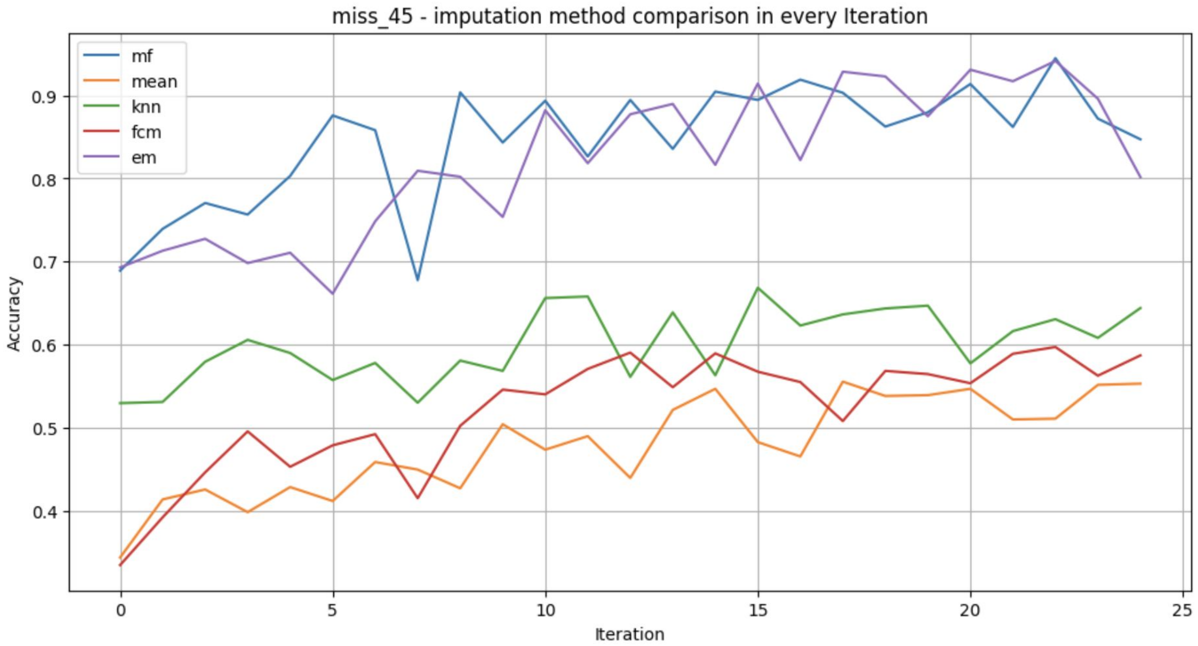
Appendix - Experiments





Appendix - Experiments





Appendix - Experiments