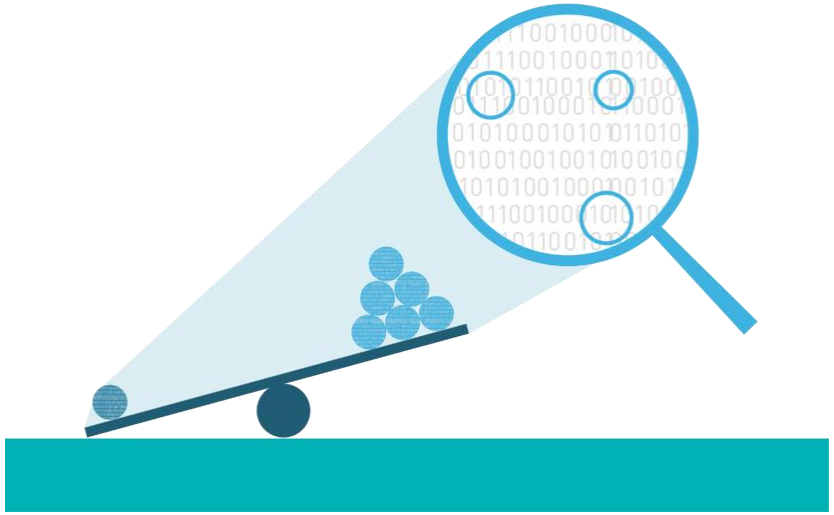
Explainable Anomaly Detection (xAD)

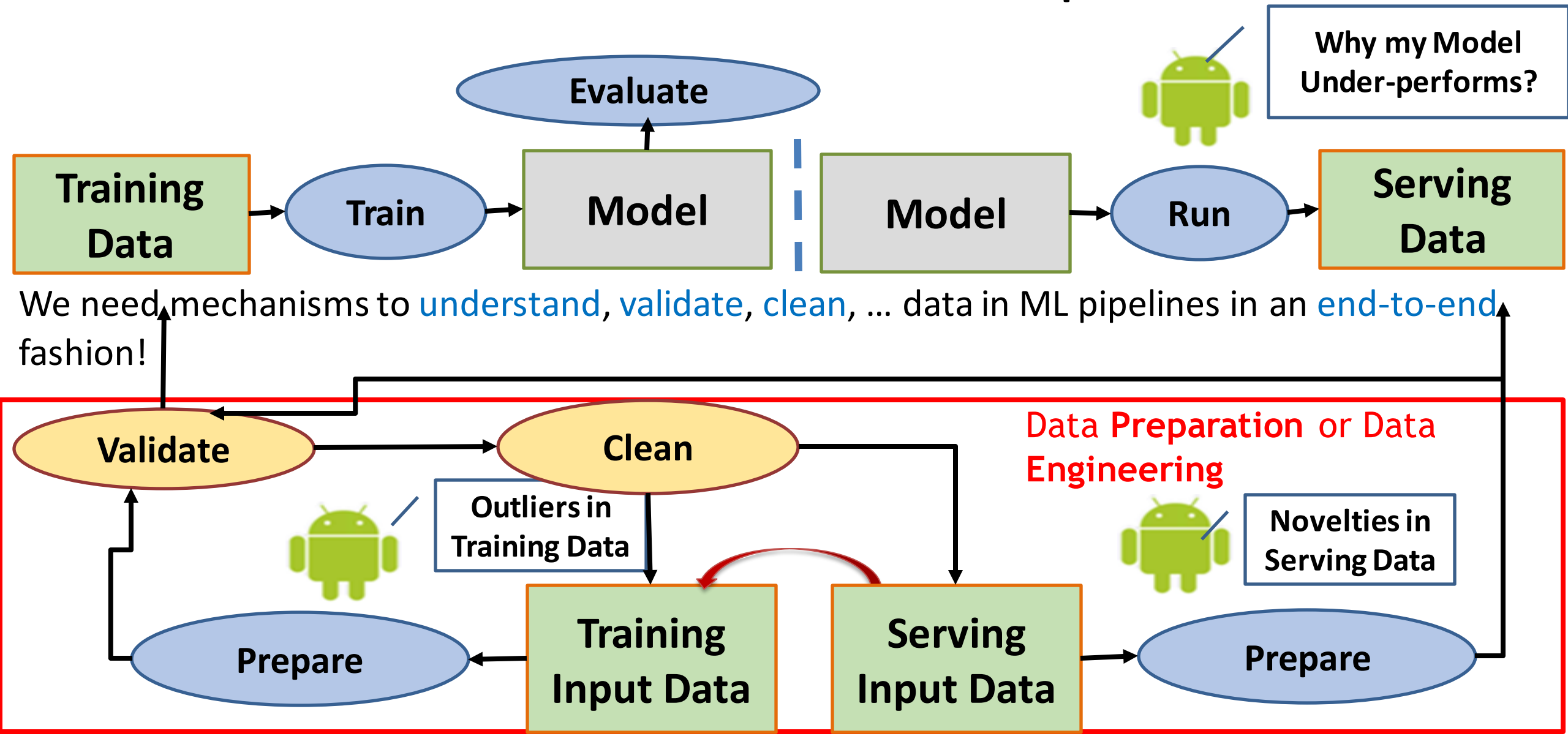


Vassilis Christophides

N. Myrtakis I. Tsamardinos E. Simon

University of Crete,
Heraklion Greece

Data-centric View of ML Pipelines



N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. 2018. Data Lifecycle Challenges in Production Machine Learning: A Survey. SIGMOD Rec. 47, 2 (December 2018), 17-28.

Data Quality in the ML Era !

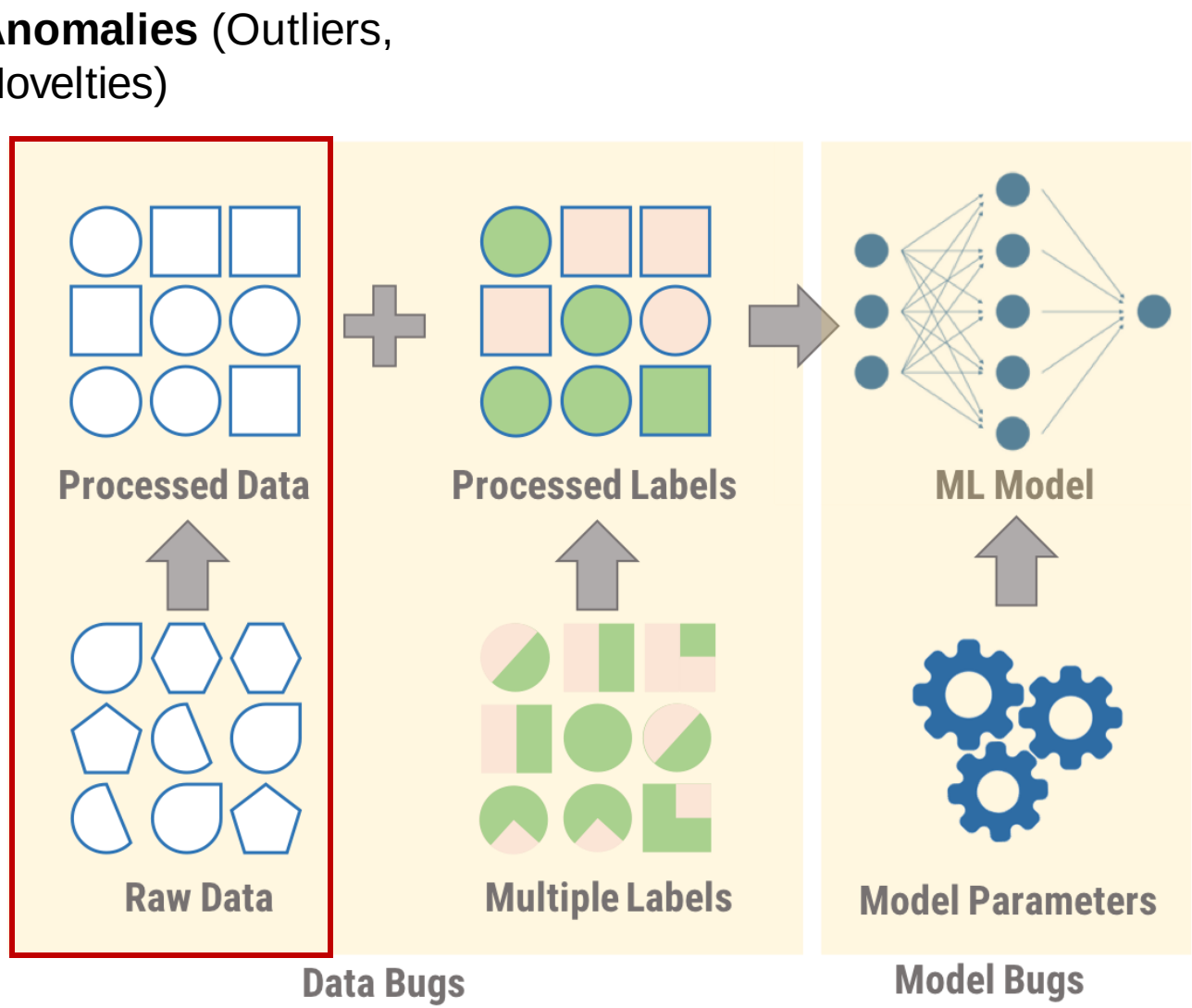
- An ML model is only as good as its data, and no matter how good a training algorithm is, the ultimate quality of automated decisions lie in the data itself! [European Union Agency for Fundamental Rights Data quality and artificial intelligence – Mitigating Bias and Error to Protect Fundamental Rights 2019]
- Only 3% of companies are making decisions based on data that meets basic quality standards [Harvard Business Review 2017]
- Most companies attempting to implement AI will fail and one of the primary reasons is the lack of enough clean training data [Techgenix 2019]

https://derivsource.com/2020/11/05/the-real-cost-of-poor-data-quality-400-million-or-much-much-more/

https://www.dataversity.net/impact-data-quality-machine-learning-era/
https://www.dataversity.net/challenges-for-data-governance-and-data-quality-in-a-machine-learning-ecosystem/

Data Quality for Building Production ML Systems

- Data quality management is a well-established area of database research and several measures to assess the quality of data in databases (*unsupervised*) can be used in commercial products
- There is a need to relook at this approach from the lens of building machine learning models (*supervised*)!
  - Need algorithms and tools that assess the quality of training/serving datasets and take remediate actions on labeled data errors
- Implement data quality management as a task of AutoML tools!



https://dawn.cs.stanford.edu/2018/08/30/debugging2

Example: Anomalies in Healthcare Data

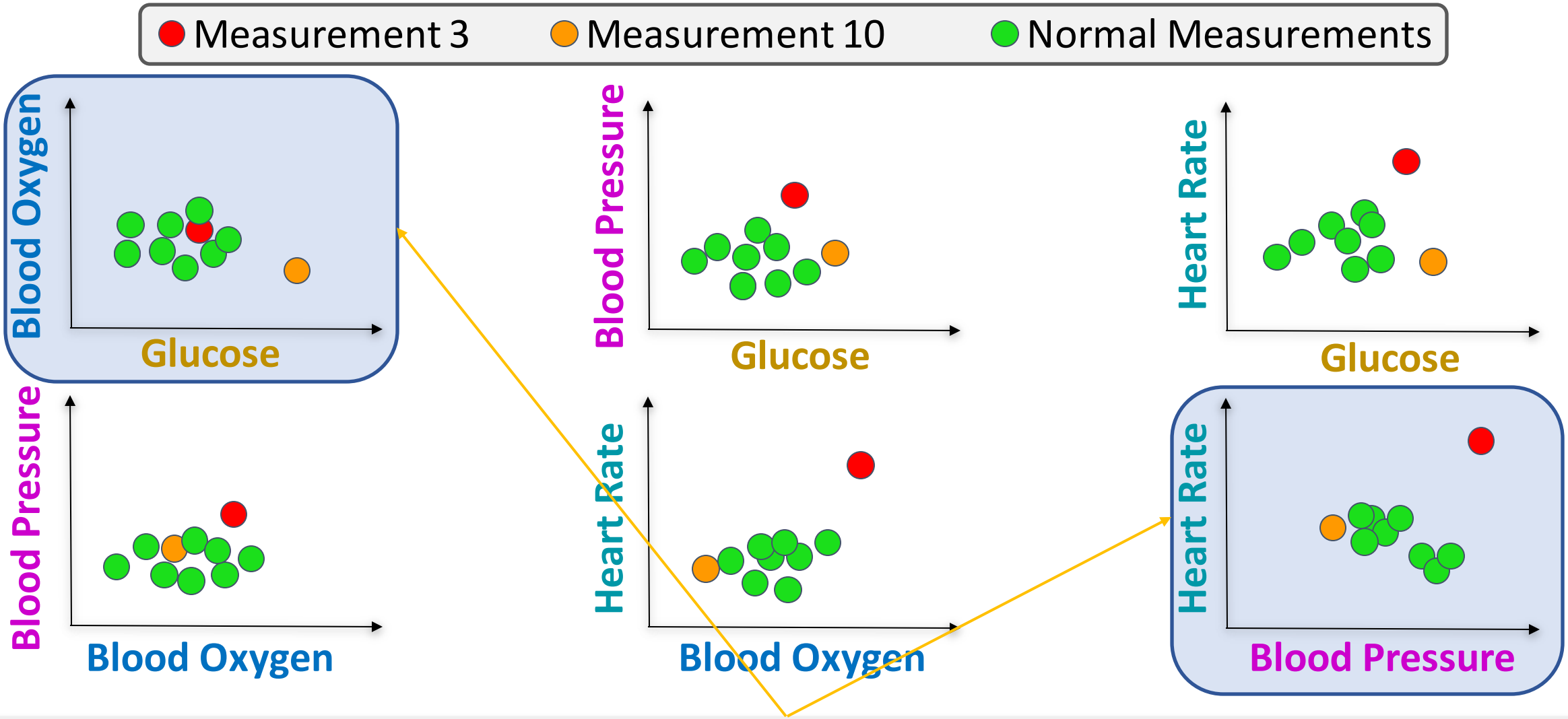
Analysis Task: Detect possible abnormal measurements for a patient

Scores close to **0** for normal samples and close to **1** for anomalies

Measurements	Glucose (mg/dL)	Blood Oxygen (%)	Blood Pressure (systolic)	Heart Rate (beats/m)	Anomaly Score
M1	100	95	100	100	0.3
M2	130	94	125	95	0.35
M3	140	97	150	160	0.9
⋮	⋮	⋮	⋮	⋮	⋮
M10	150	90	120	105	0.95

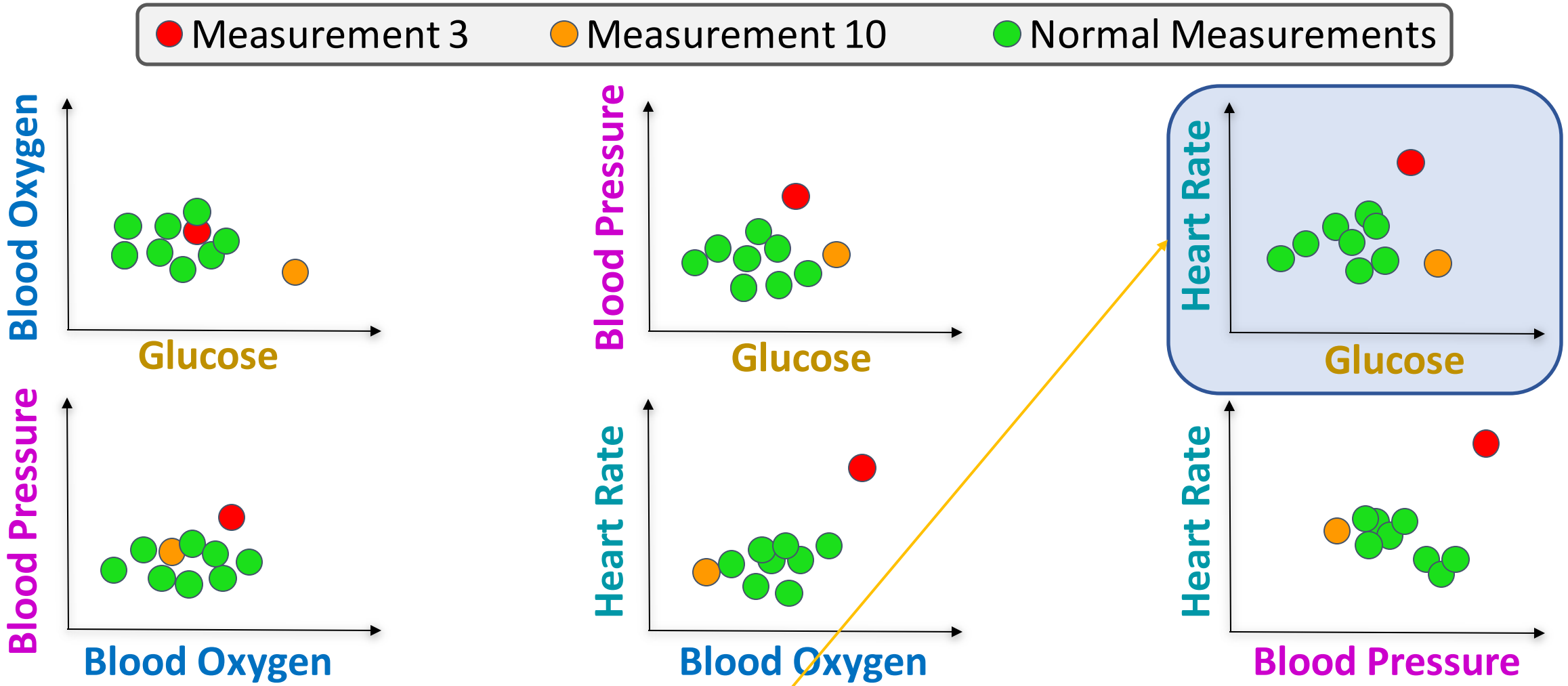
- Why did M3 and M10 get a high score?
- Should I alert patients for health problems?

2D Subspaces Explaining Anomalies: Local



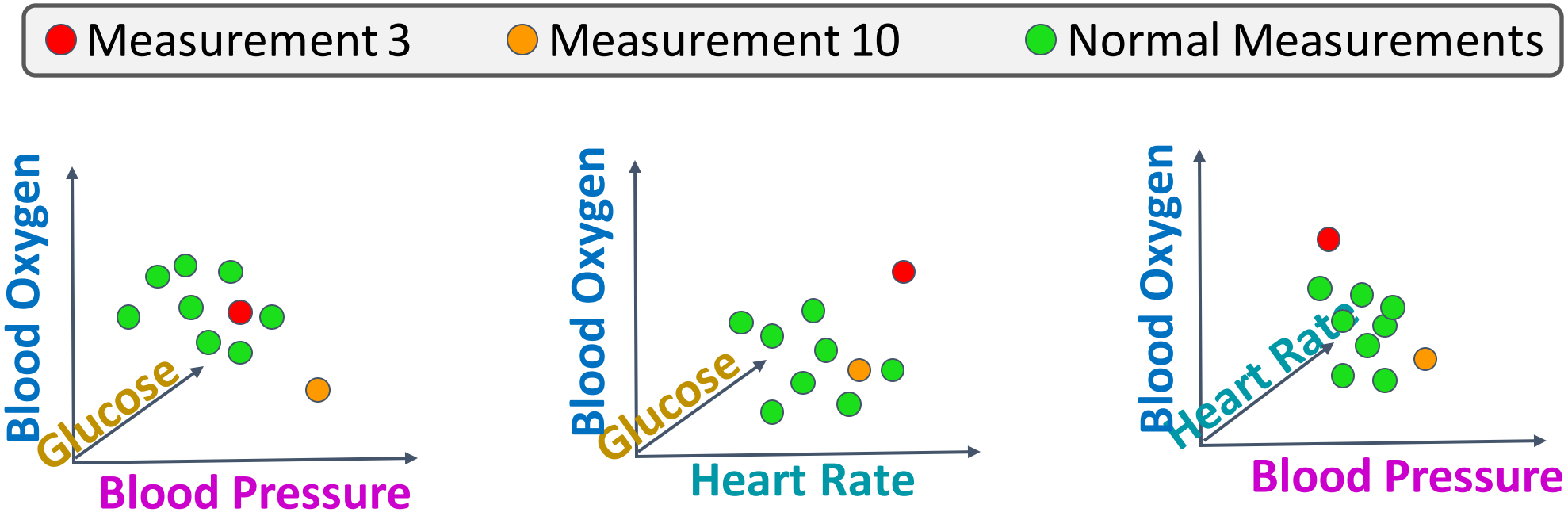
Local Explanation: Find subspaces that maximize anomalousness of individual samples

2D Subspaces Explaining Anomalies: Global



Global Explanation: Find subspaces that summarize the anomalousness of as many samples as possible

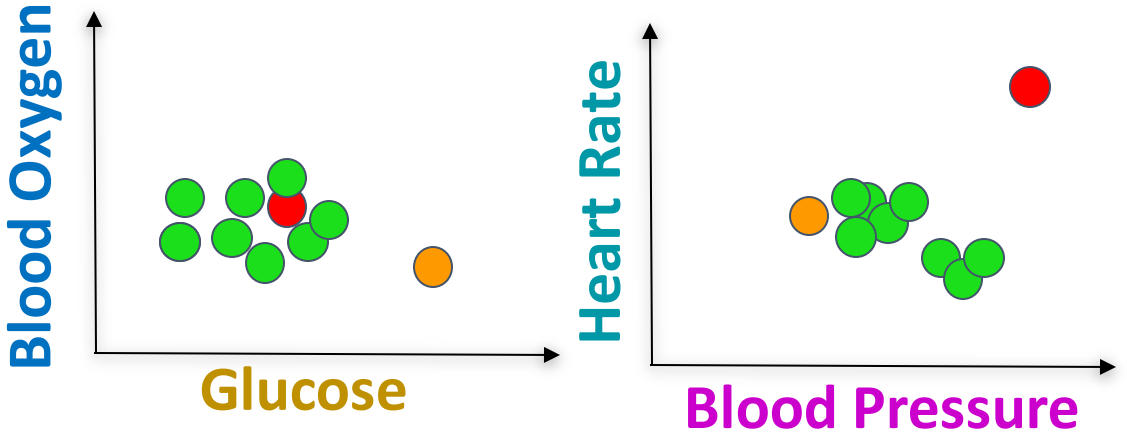
3D Subspaces Explaining Anomalies: Higher Dim.



We don't know in advance the **dimensionality of 'best explanations'**!

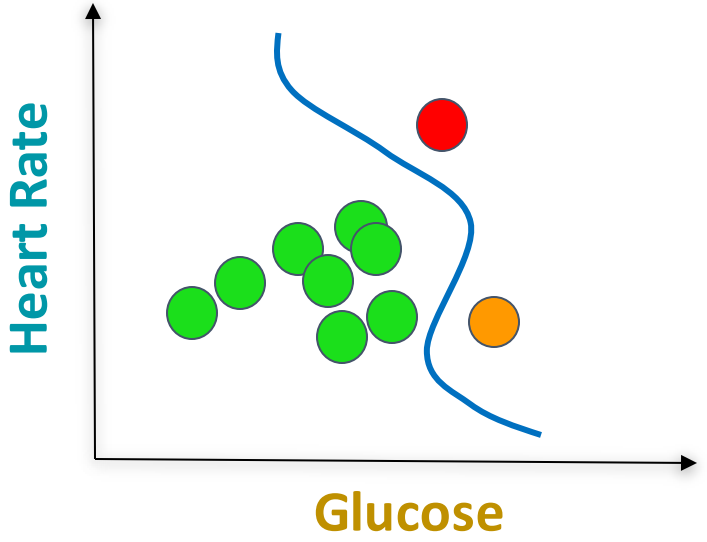
Descriptive vs Predictive Anomaly Explanations

Descriptive Explanation



A feature subset that can maximize the anomalousness score of samples as seen by a detector

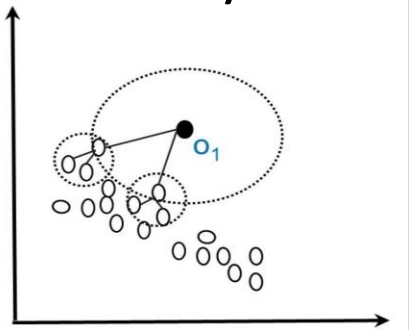
Predictive Explanation



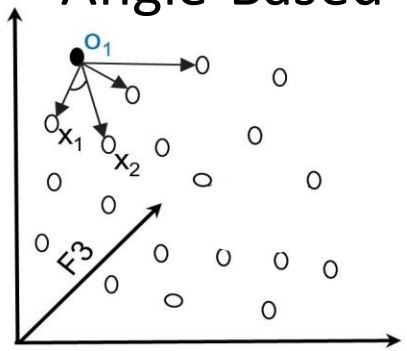
A minimal subset of features leading to a predictive model that best approximates the decision boundary of a detector

How Can We Produce Predictive Explanations ?

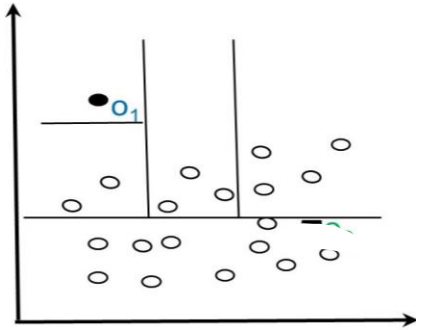
Density-Based



Angle-Based

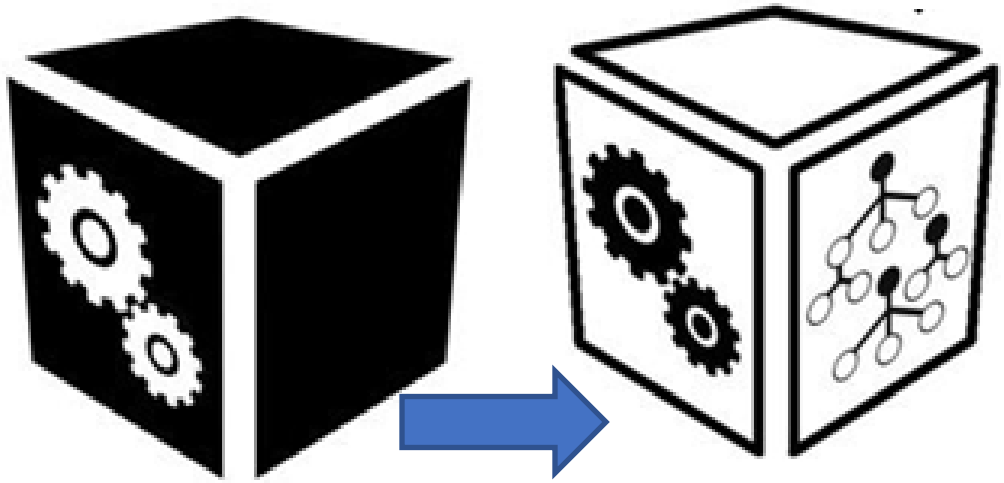


Isolation-Based



Unsupervised Anomaly Detector

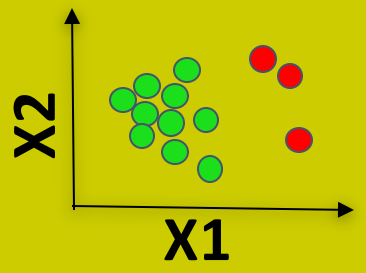
Surrogate Supervised Model



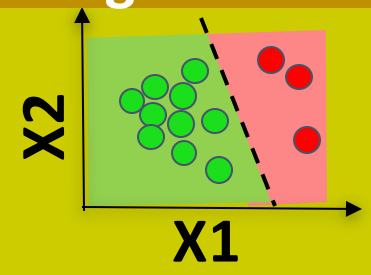
Samples **Labelled** by the **Detector**

Generalization Performance Estimation

Feature Selection Algorithms

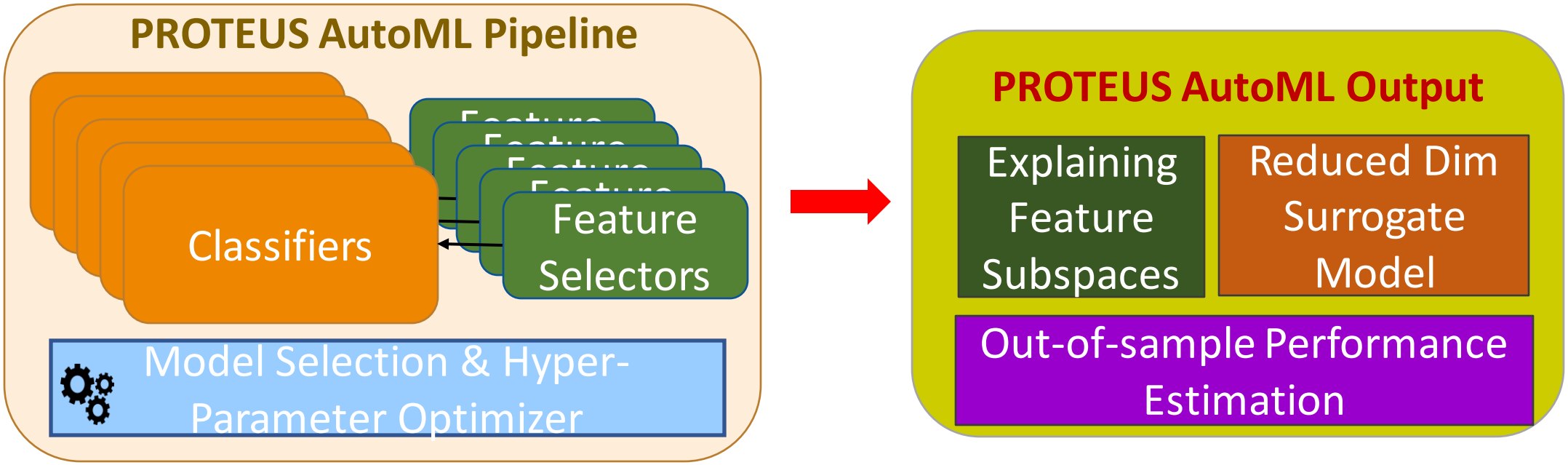


Classification Algorithms



AutoML

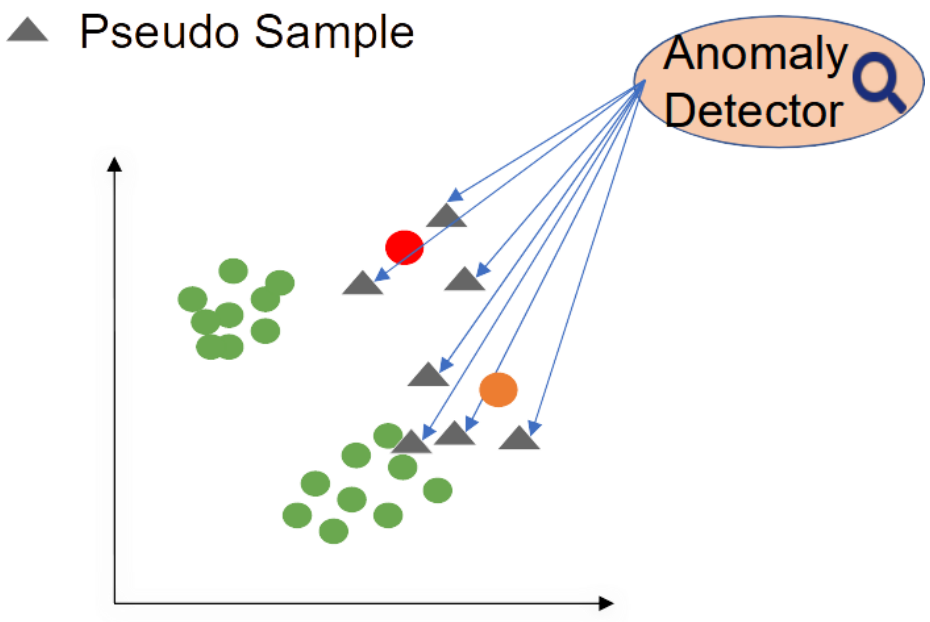
PROTEUS Outcome and Design Choices



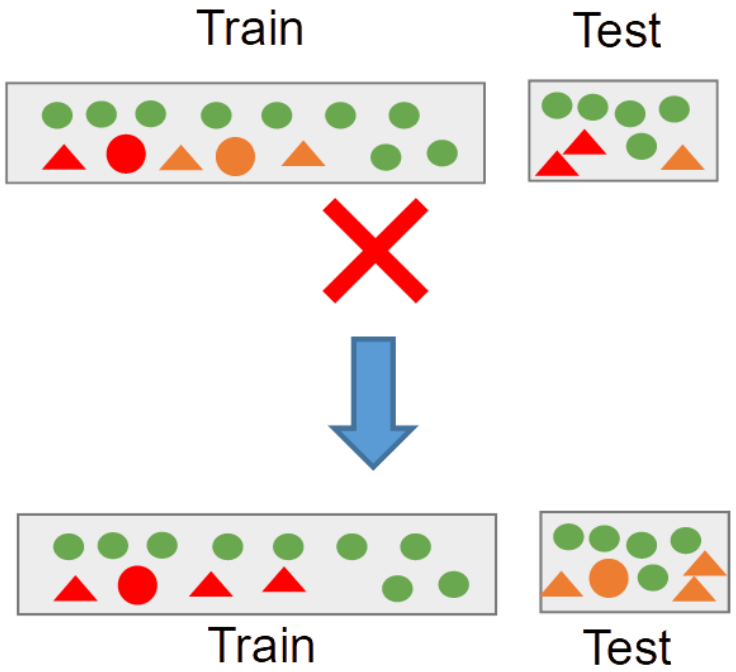
1. How to treat the inherent imbalance nature of the anomaly class ?
2. How we avoid information leakage between train and test sets ?
3. How provide reliable performance estimates ?

Overview of Proteous Design Choices

1 Supervised Oversampling



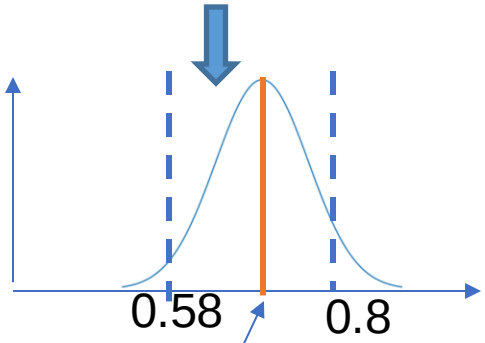
2 Anomaly Grouping



3 Bootstrap Bias Correction (BBC)

Out-of-sample Predictions

	$Model_1$	…	$Model_n$
s_1	0.8	…	0.6
…	…	…	…
s_n	0.4	…	0.55



Perf. Est. = $\frac{\sum_{i=1}^{B} Performance_i}{B}$

PROTEUS Search Space of Surrogate Models

Classification Algorithms

Random Forest	**# Trees**: {100, 300, 500}, **MinLeafSize**: {1, 2, 3} **Split Criterion**: {Entropy}
Support Vector Machines	**Kernel**: {Linear, RBF, Poly}, **C**: {1, 5, 10}, **γ**: {1, 2, 5}, **degree**: {2, 3}
K-Nearest Neighbors	**K**: {5, 10 ,15}

Cartesian Product

Feature Selection Algorithms

SES	**max_k**: {2,3}, **α**: {0.01, 0.05, 0.1}
FBED	**threshold**.: {0.01, 0.05, 0.1} , **iters**: {0, 1, 3}
Lasso	**λ**: {0.001, 0.01, 0.1, 0.2}

Pseudo Samples Per Anomaly (ps): 0, 3, 10

Exhaustive Grid Search

Configuration 1: **SES**: max_k=2 , α=0.01 & **Random Forest**: #Trees=100, MinleafSize:1, Split Crit.: Entropy & ps=0

Configuration 2: **SES**: max_k=3 , α=0.01 & **Random Forest**: #Trees=100, MinleafSize:1, Split Crit.: Entropy & ps=0

⋮ ⋮

Configuration 1800: **Lasso**: λ=0.2 & **KNN**: K=15 & ps=10

Related Work & Baselines

Method	Category	Detector Agnostic	Global Explanation	Predictive Explanation
SHAP [Lundberg et al. 2017]	Black-box model explainer	✓	✗	✗
CA-Lasso [Micenková et al. 2013]	Post-hoc anomaly explainer	✓	✗	✗
LODA [Penvy T. 2015]	Explainable anomaly detector	✗	✗	✗
PROTEUS *[Myrtakis* et al. 2021]	AutoML anomaly explainer	✓	✓	✓

Real and Synthetic Datasets

Dataset Name	# Features	# Samples	Anomaly Ratio	IF	LOF	LODA
Synthetic	5	867	1%	0.96	1.0	0.92
Wisconsin Breast Cancer	30	377	5%	0.95	0.94	0.96
Ionosphere	33	358	36%	0.85	0.93	0.87
Arrhythmia	257	452	15%	0.80	0.74	0.75

Characteristics **Detectors AUC in Train**

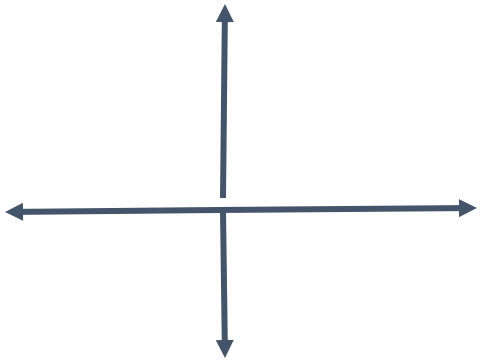
Adding irrelevant features to the **synthetic** dataset: 77%, 88%, 92%, 94%, 95%
Adding irrelevant features to every **real** dataset: 30%, 60%, 90%

Experimental Setting

- Each dataset was stratified and split to 70% training - 30% held out
- Up to 10 features were selected as explanation based on their scores
- Experimental Dimensions

PROTEUS (fs, full, ca-lasso, shap, loda)

Detectors (IF, LOF, LODA)



Datasets with irrelevant features

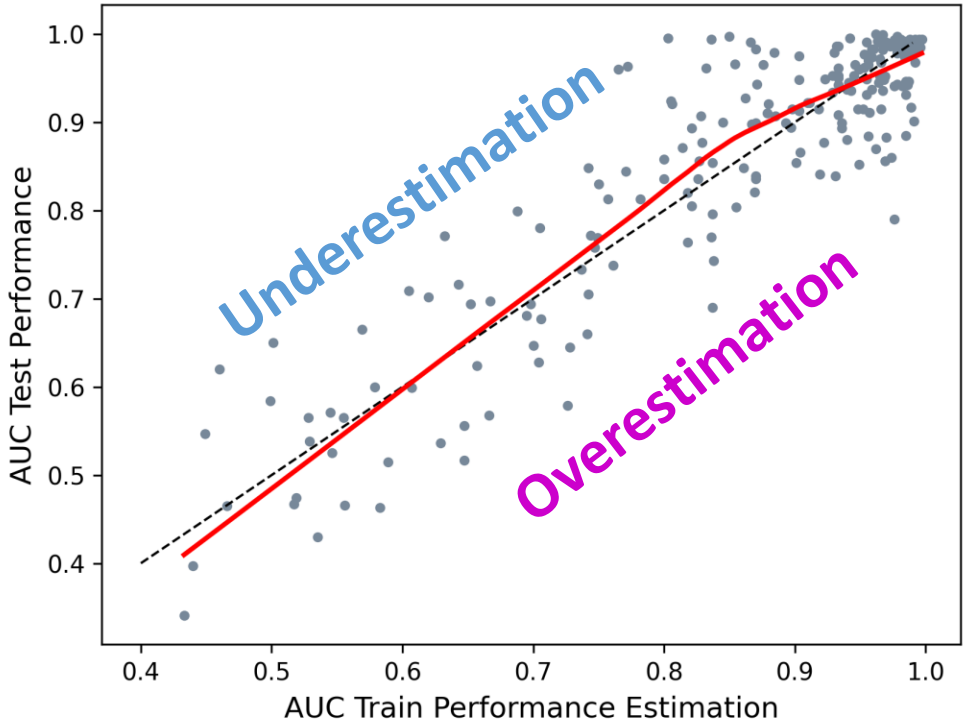
5 for the synthetic and 3 per real dataset

255 analyses

1 synthetic dataset and 3 real datasets

PROTEUS Performance Estimation

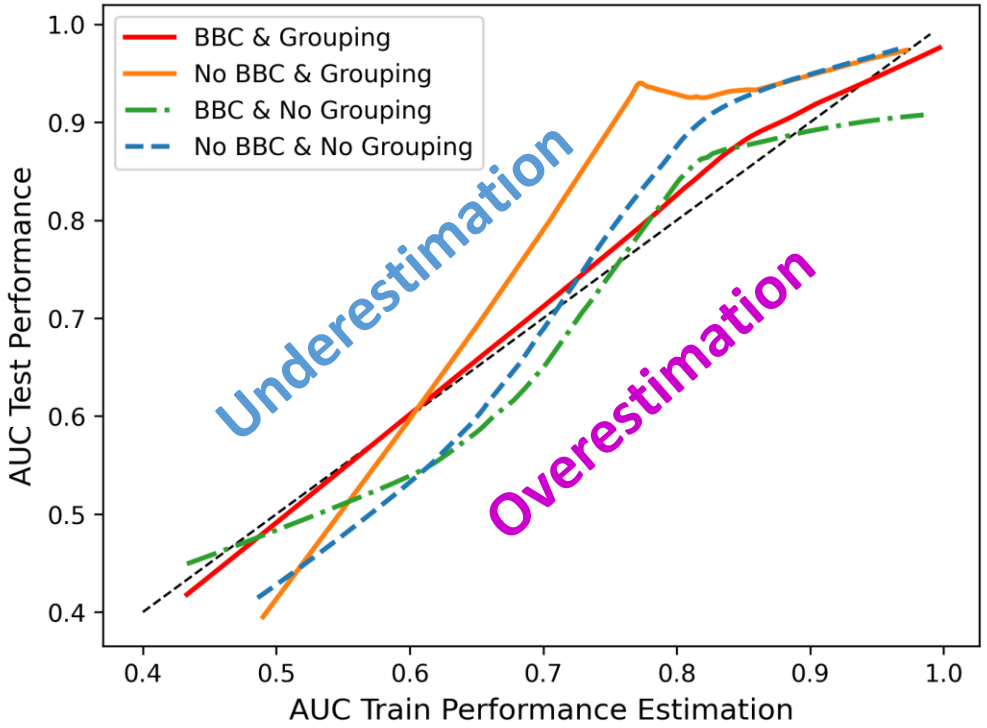
Q: Do the design choices of PROTEUS contribute to provide an accurate performance estimation ?



- Each point represents the train and test performance for a particular analysis
- The **dashed black diagonal** line indicates the zero bias
- The red line is the loess smoothing curve

Ablation Analysis

Q: How is the accuracy of performance estimation affected by different design choices ?



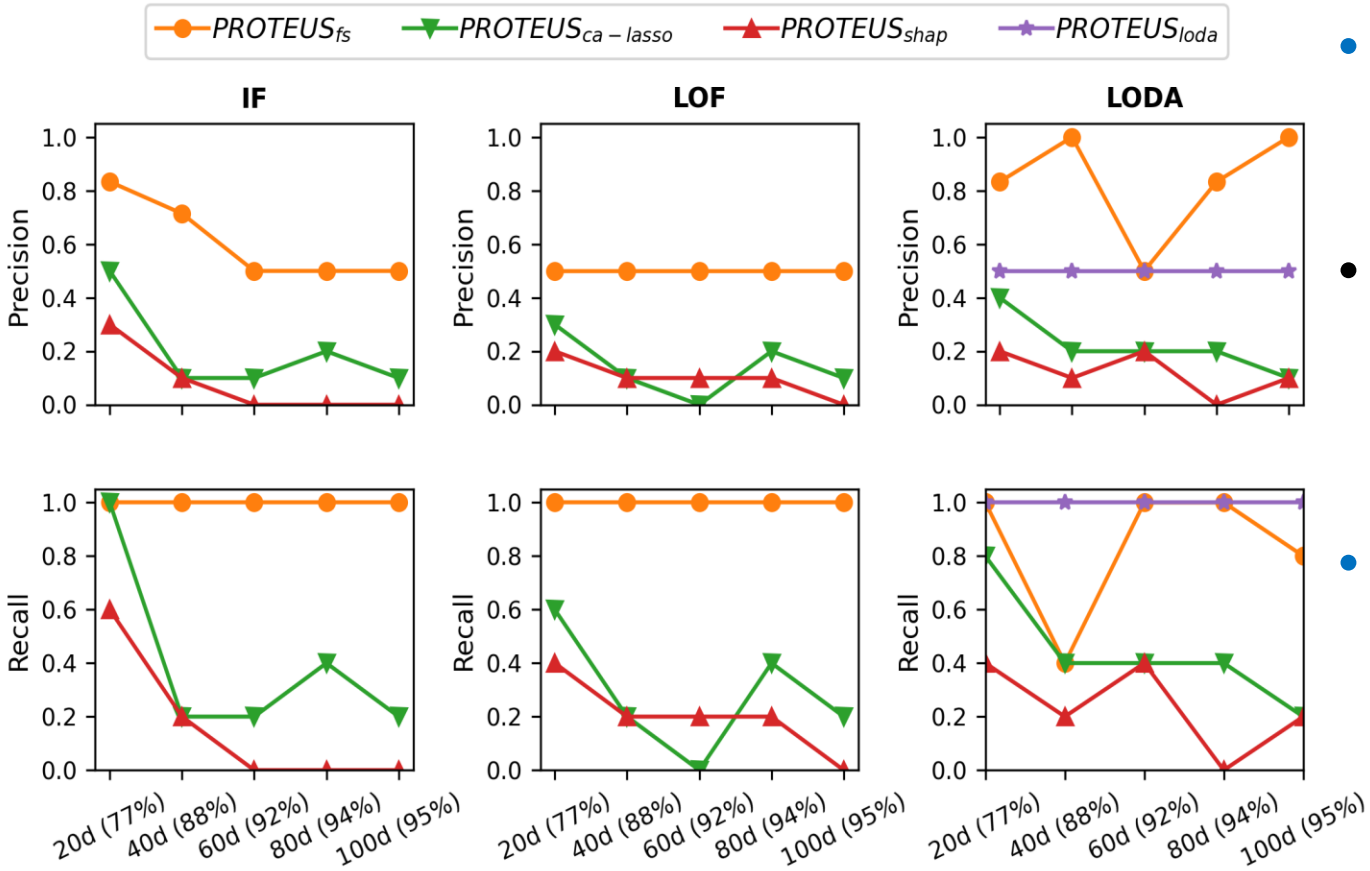
BBC & Grouping	No BBC & Grouping	BBC & No Grouping	No BBC & No Grouping
0.05	0.88	0.11	0.25

Residual Sum of Squares of the 4 design choices

- PROTEUS with BBC and CV with Grouping gives the most accurate estimation

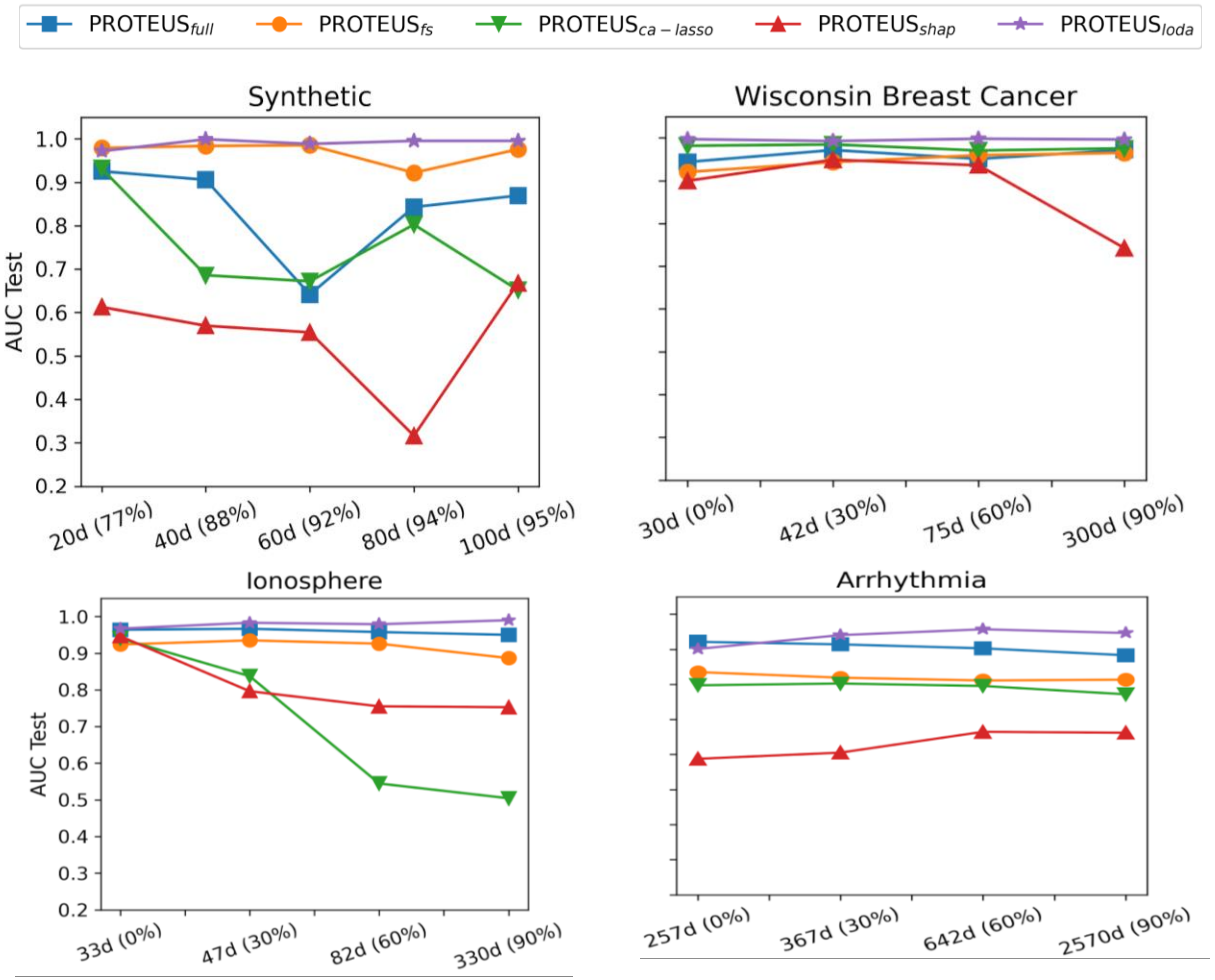
Relevant Features Identification Accuracy

Q: What is the precision and recall of discovered features w.r.t. synthetic gold-standard of anomaly explanations (with 5 relevant features)?



- Feature selection algorithms of PROTEUS$_{fs}$ exhibit the highest overall precision suboptimal recall
- Unlike SHAP and CA-Lasso, PROTEUS$_{fs}$ exhibits a robust performance when varying data dimensionality, regardless of the employed detector
- PROTEUS$_{fs}$ approximates well the recall of the explainable detector LODA which is the upper performance limit

Generalization Performance



- In synthetic dataset PROTEUS$_{fs}$ generalizes better than PROTEUS$_{full}$
- In real datasets PROTEUS$_{fs}$ is robust achieving high performance regardless of the employed detector
- PROTEUS$_{fs}$ approximates the optimal performance of LODA in a detector-agnostic manner

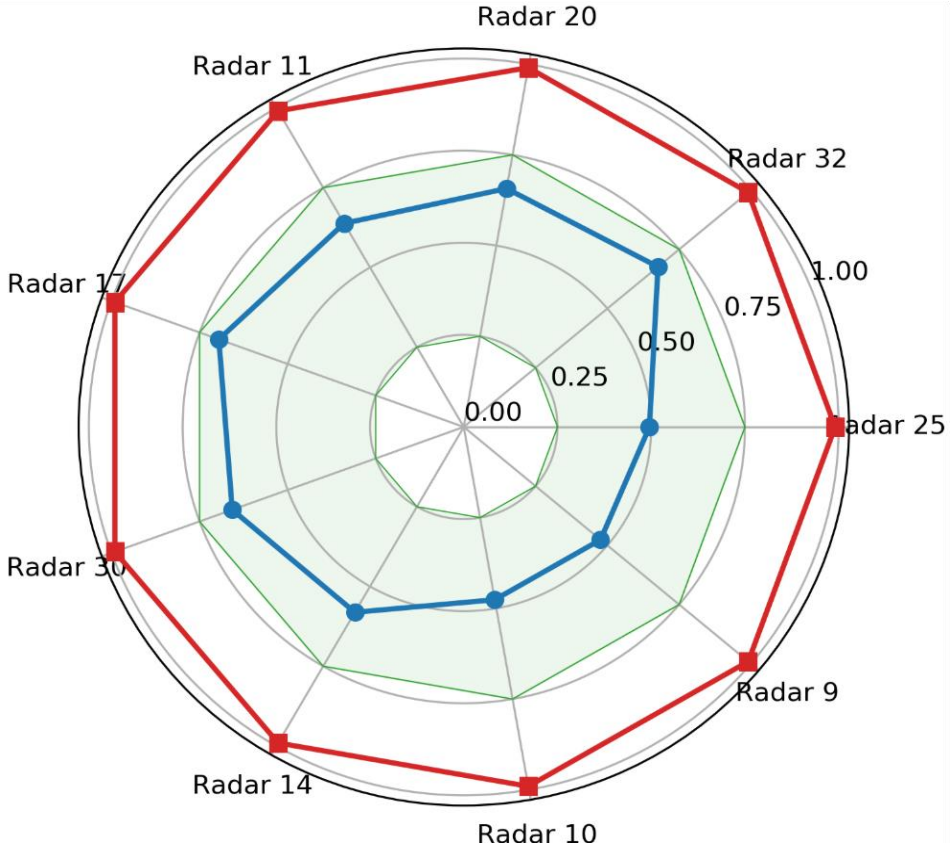
- AUC test performance averaged over the 3 detectors

Contrasting PROTEUS Surrogate Models With Unsupervised Anomaly Detectors

- Ionosphere Dataset (33 Features)

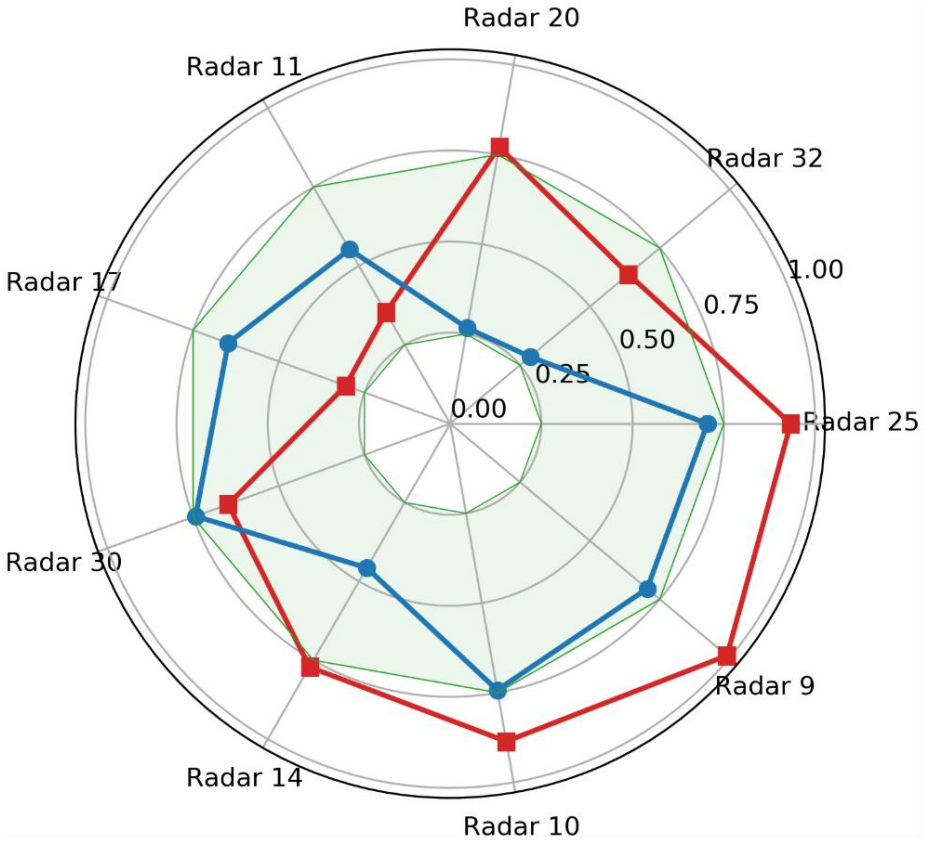
Proteus Agreement with LOF

(9 features explanation)



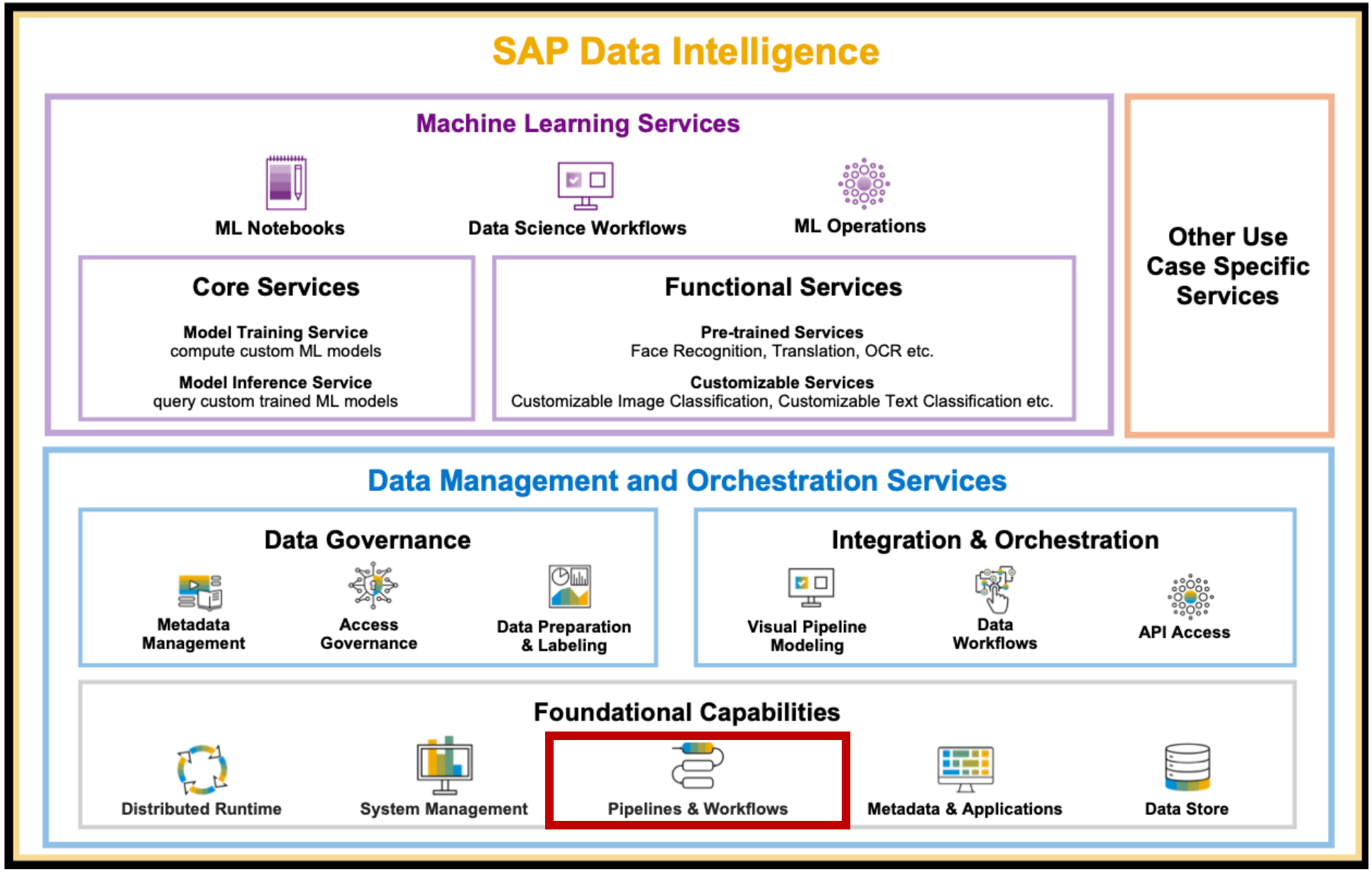
Proteus Disagreement with LOF

(9 features explanation)



Proteus Anomaly Proteus Normal

Anomaly Detection & Explanation Operators in SAP DI



https://developers.sap.com/topics/data-intelligence.html

Predictive Anomaly Explanation Pipeline in SAP DI

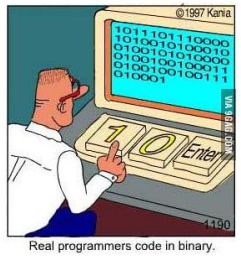
Data Stream

Multi-Dimensional Data Samples

Anomaly Detector

Scored Samples

High Low

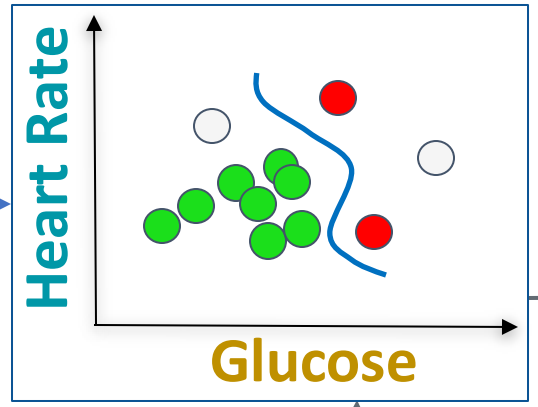


Anomaly Scores Binarization

Labelled Samples

New batch of data

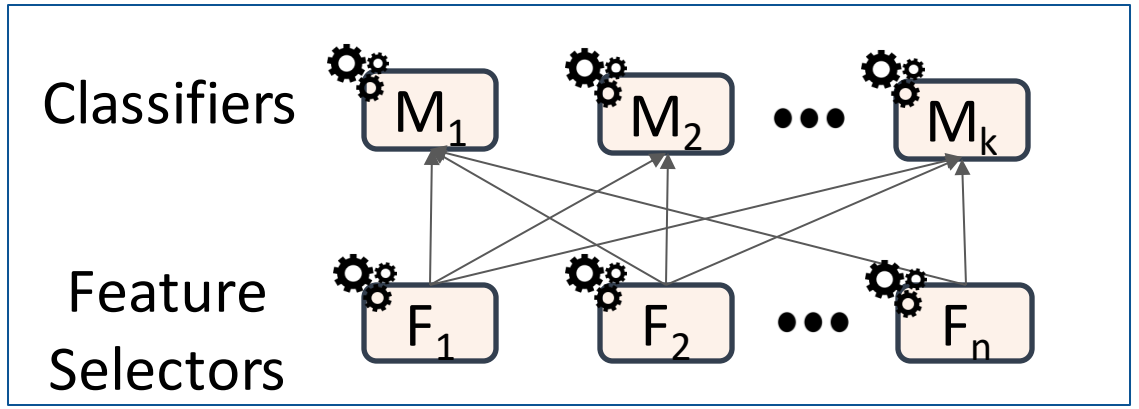
Reduced-dimensionality surrogate model



Alarms: Anomalies

Discarded: Normal Points

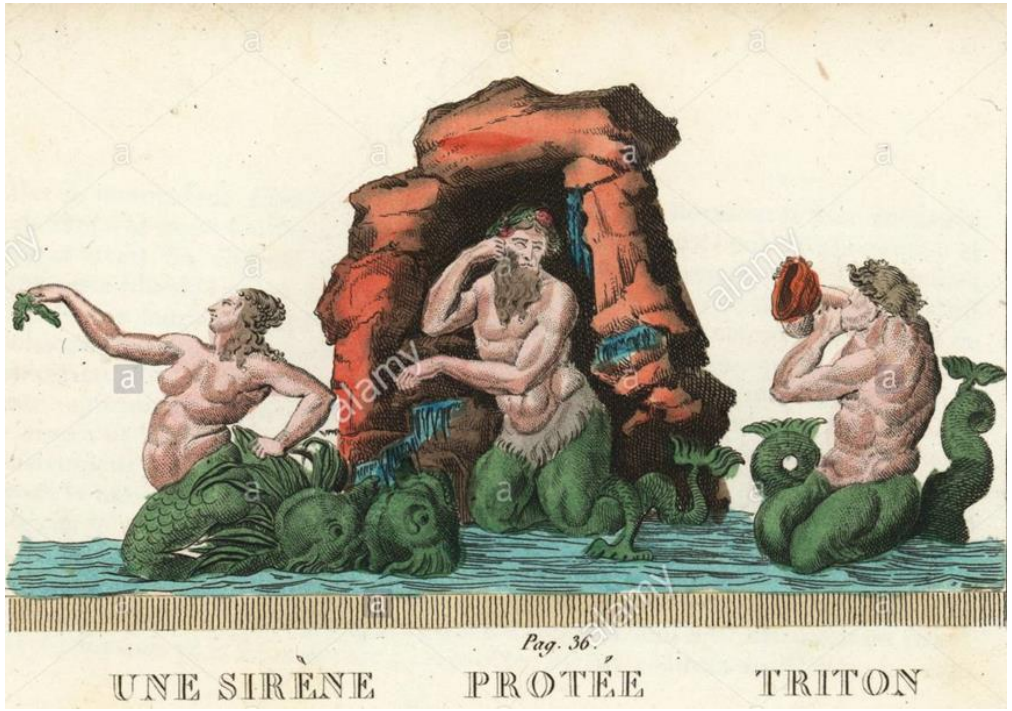
Human Analyst



PROTEUS AutoML Pipeline

Summary

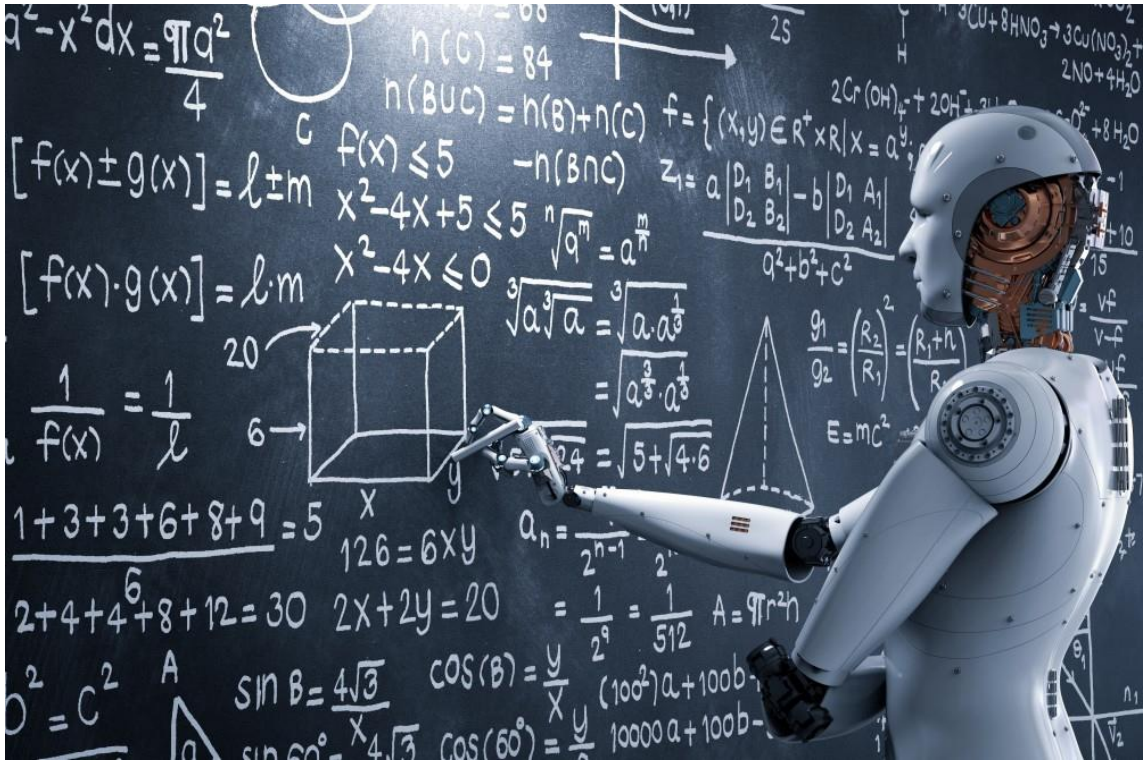
- Anomaly explanation ⟶ **a supervised classification problem with feature selection** ⟶ solved effectively as an AutoML problem
- **First** methodology for predictive, global, detector-agnostic anomaly explanations
  - Not all existing explanation formalisms can serve as a predictive model!
- PROTEUS is **robust** and **effective** discovering features relevant to anomalies
- Adequate design choices (Oversampling, BBC, CV with Grouping)⟶**accurate approximation of a detector's decision boundary**⟶accurate performance estimation



In Greek mythology, Proteus (Πρωτεύς) is an early prophetic sea-god or god of rivers and oceanic bodies of water, one of several deities whom Homer calls the "Old Man of the Sea"

Open Issues in xAD

- Explaining Anomalies in Data Streams
  - Online Anomaly Detection & Explanation
- Explainability of Time Series Models
  - Higher-level data abstractions to explain DL-based models for temporal patterns
- End-to-end explanations of data quality issues in ML pipelines
  - Coupling diagnosis of downstream ML models to upstream data preprocessing
  - Heterogenenous explanation models for different tasks: boolean expressions, decision trees, numerical score of feature importance (Linear Regression, SHAP)

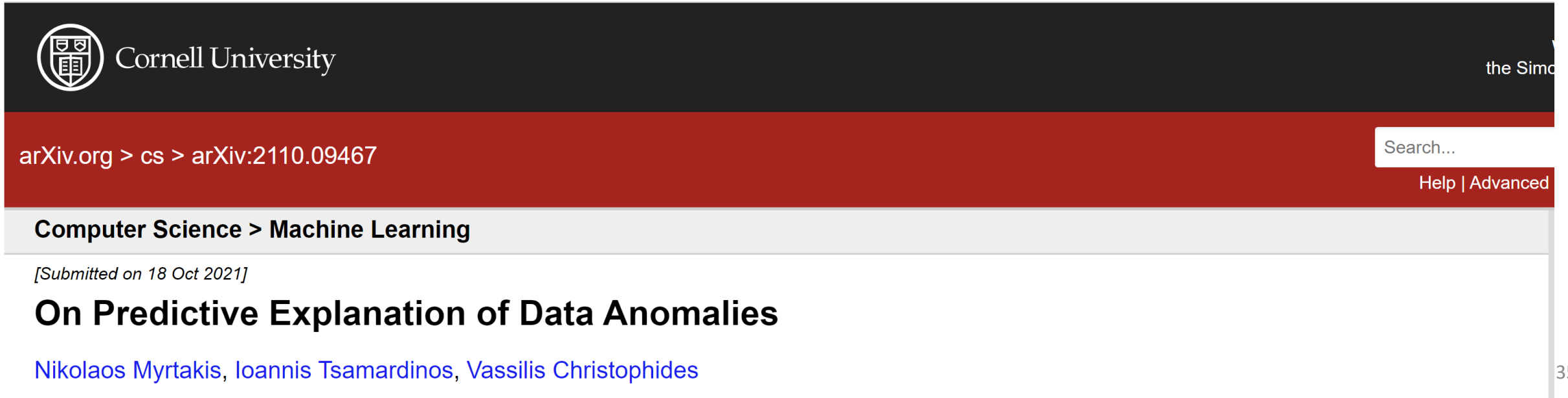


Acknowledgements

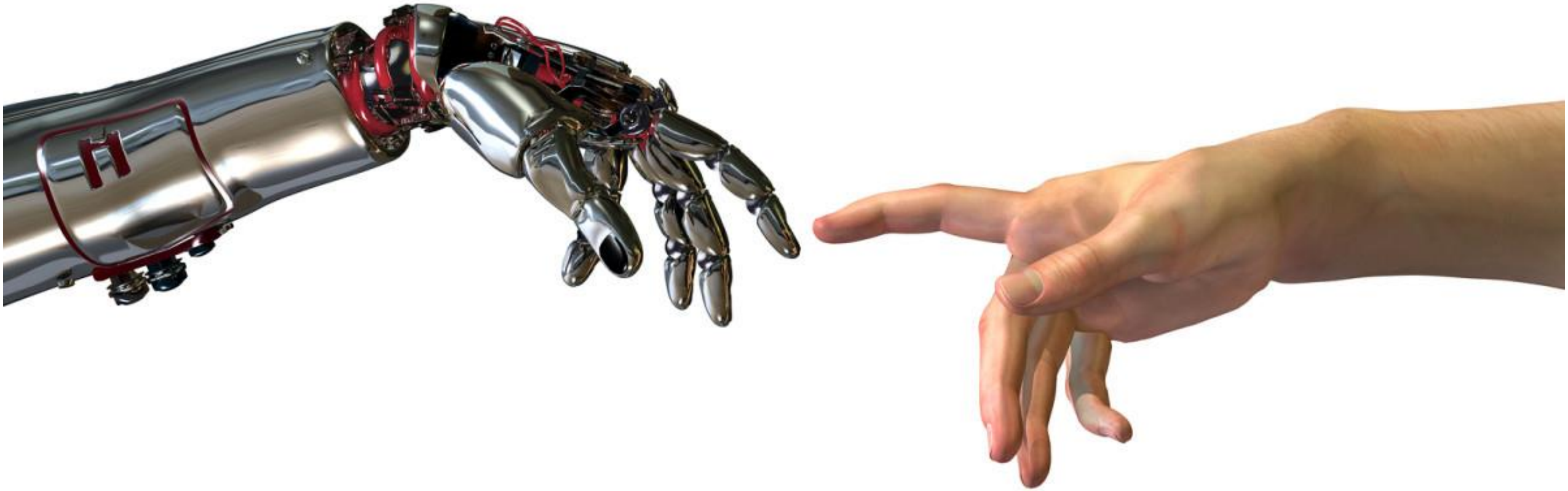
1st Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment

Fellows-in-Residence 2019-2020

https://arxiv.org/abs/2110.09467

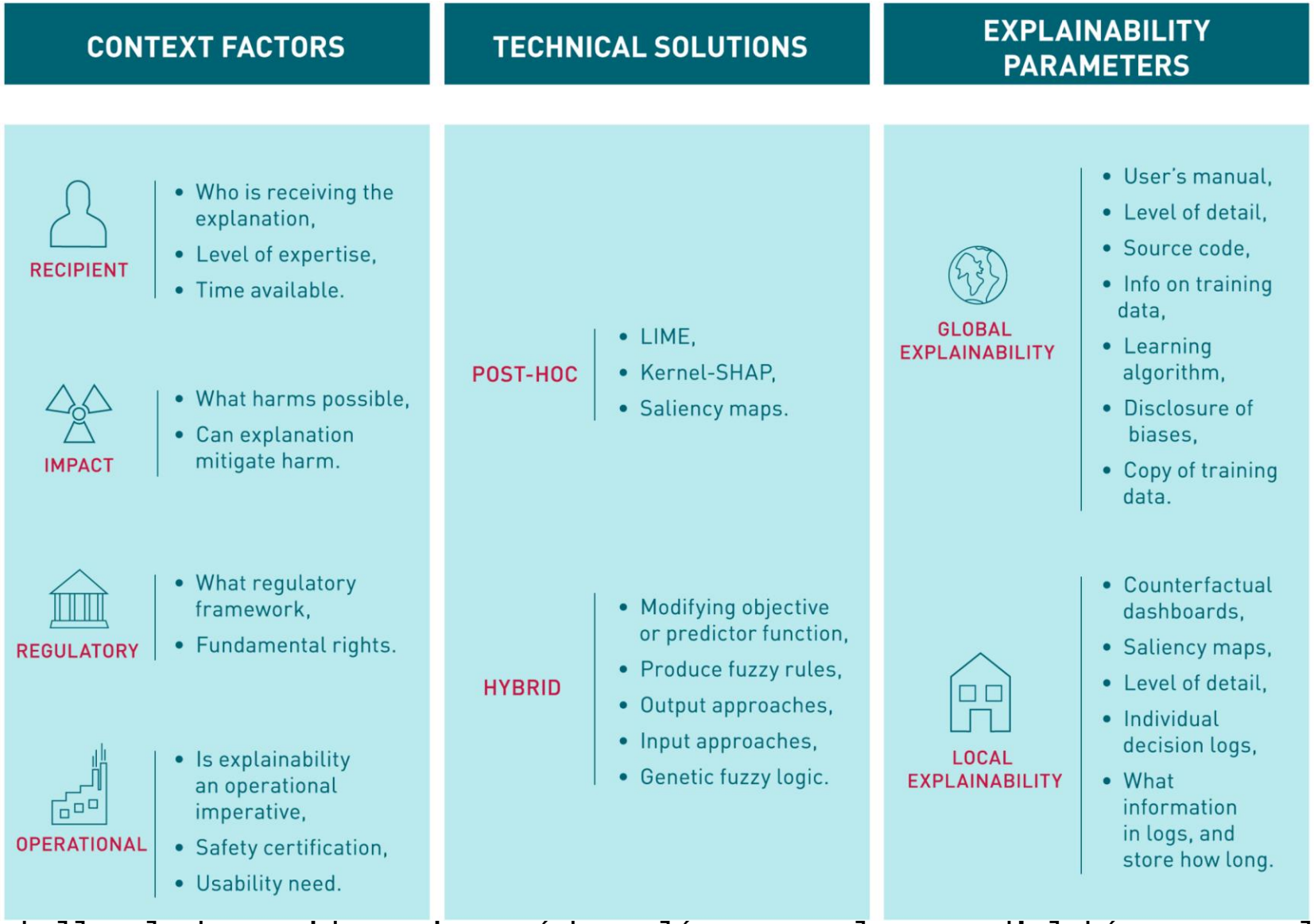


Questions?



https://thenextweb.com/contributors/2018/10/06/we-need-to-build-ai-systems-we-can-trust/

The Three Pillars of xAI

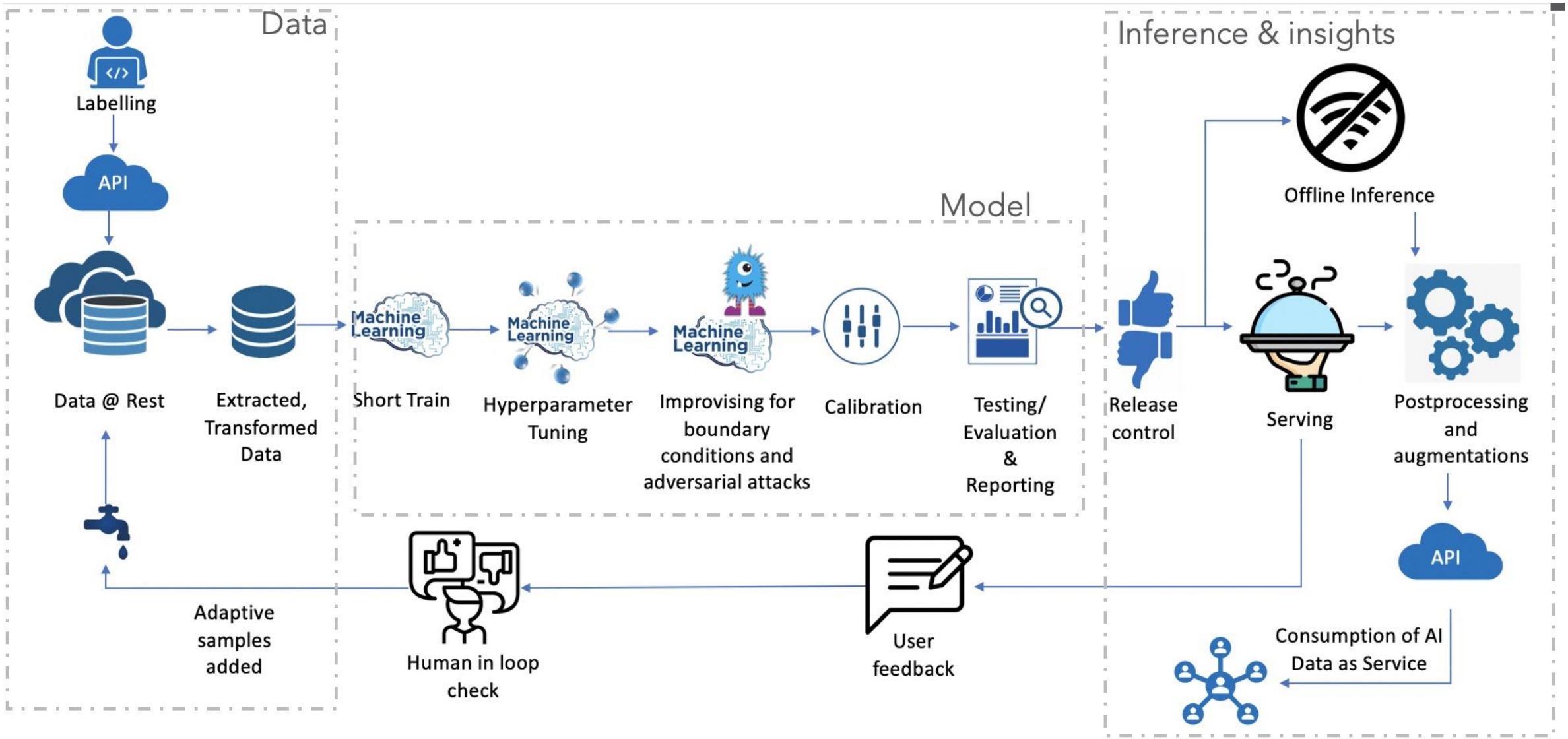


Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, et al. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. 2020. hal-02506409

xAI Terminology and Definitions

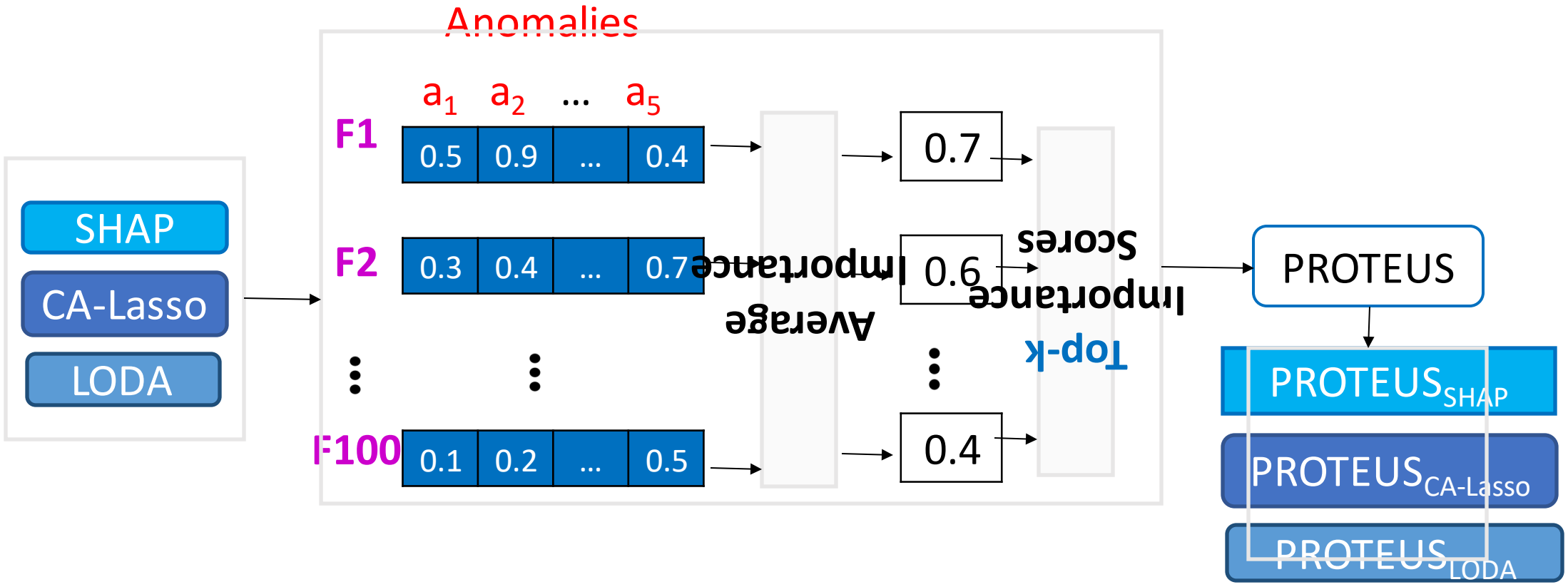
Thomas Rojat, Raphael Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Dıaz Rodrıguez. Explainable Artificial Intelligence (XAI) on Time Series Data: A Survey CoRR abs/2104.00950 (2021)

End-to-End Machine Learning System



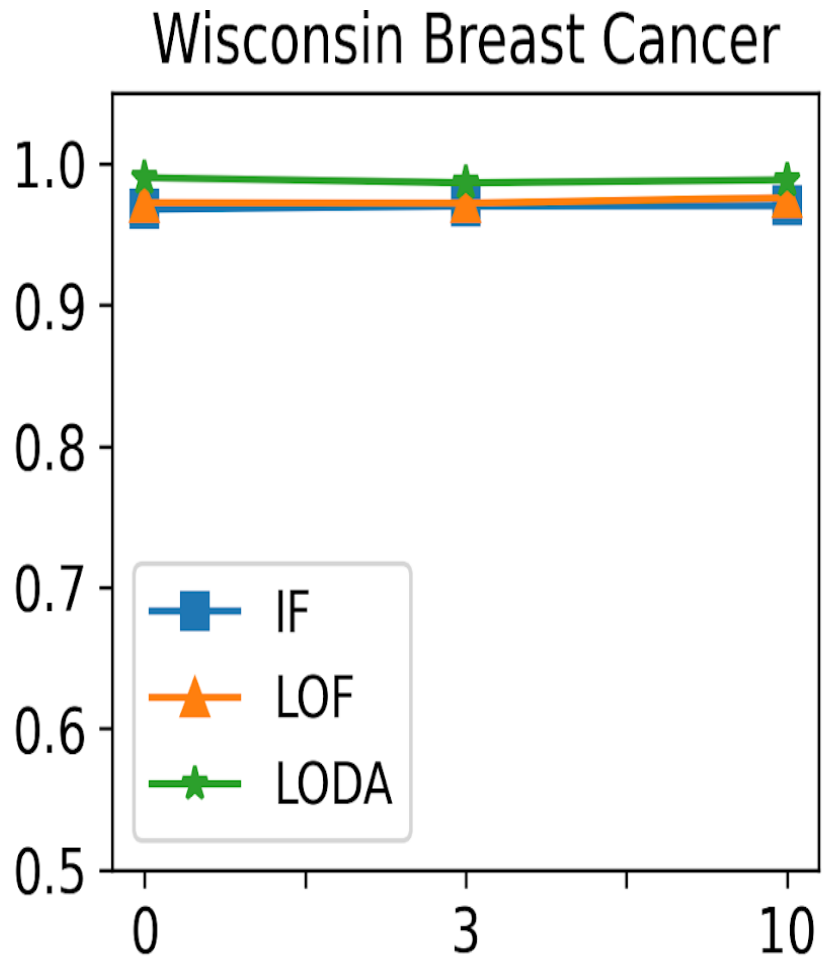
https://suneeta-mall.github.io/2019/12/23/Reproducible-ml-pipeline-k8s.html

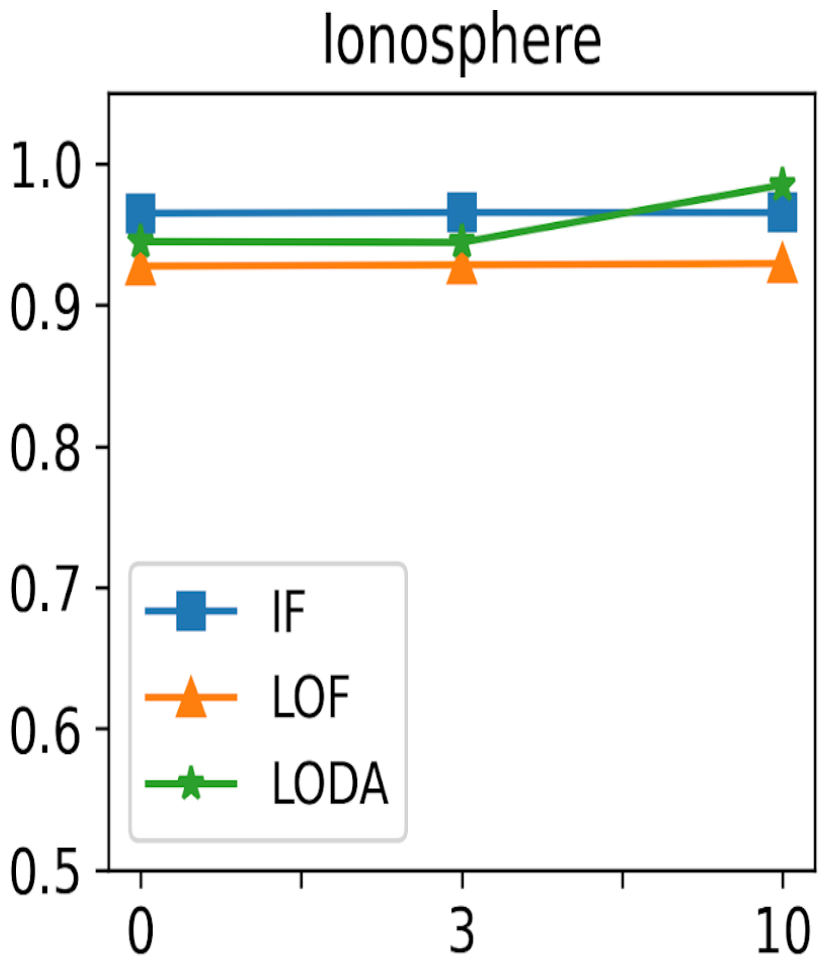
Creating a Predictive Explanation for Feature Importance Methods using PROTEUS

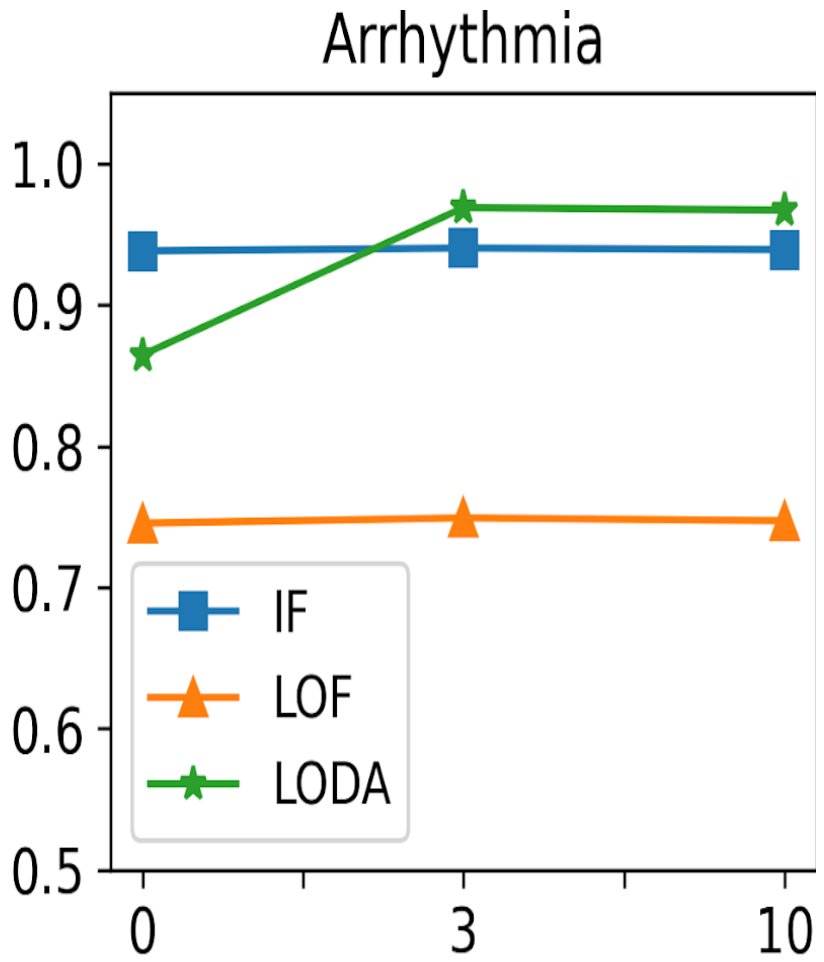


The Effect of Oversampling on Performance

- Effect of increasing pseudo-sample size per anomaly on AUC test performance







References

- **Breunig**, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. **2000**. LOF: identifying density-based local outliers. *SIGMOD*.
- **Jensen**, D. D.; and Cohen, P. R. 2000. Multiple Comparisons in Induction Algorithms. *do. Learn.*
- **Kriegel**, H.-P., Schubert, M., and Zimek, A. (**2008)**. Angle-based outlier detection. *KDD*.
- **Lagani**, V.; Athineou, G.; Farcomeni, A.; Tsagris, M.; Tsamardinos, I.; et al. **2017**. Feature Selection with the R Package MXM:Discovering Statistically Equivalent Feature Subsets. *Journal of Statistical Software.*
- **Liu**, F.T., Ting, K.M., Zhou, Z.H. (**2008**). Isolation forest. ICDM
- **Pevny**, T. **2015**. Loda: Lightweight on-line detector of anomalies. *Machine Learning.*
- **Tibshirani**, R. **1996**. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*
- **Tsamardinos**, I.; Greasidou, E.; and Borboudakis, G. **2018**. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning.*
- **Tsamardinos**, I.; Borboudakis, G.; Katsogridakis, P.; Pratikakis, P.;and Christophides, V. **2019**. A greedy feature selection algorithm for Big Data of high dimensionality. *Machine Learning.*