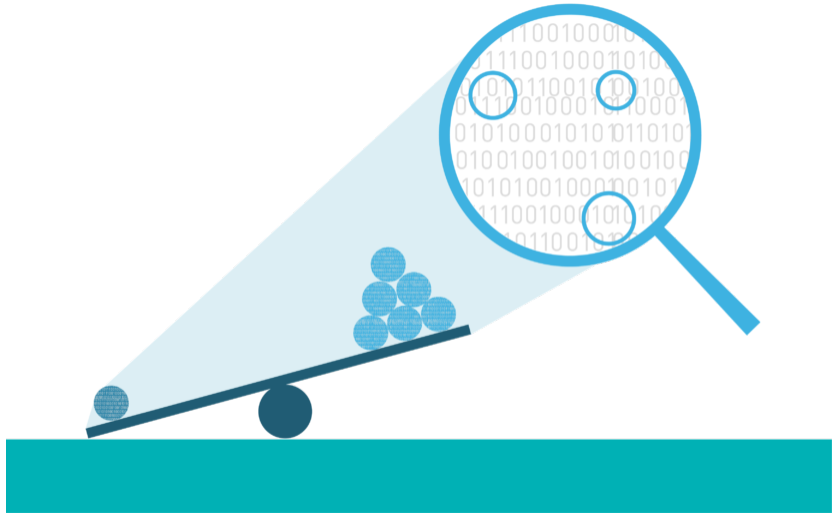
Explainable Anomaly Detection (xAD)

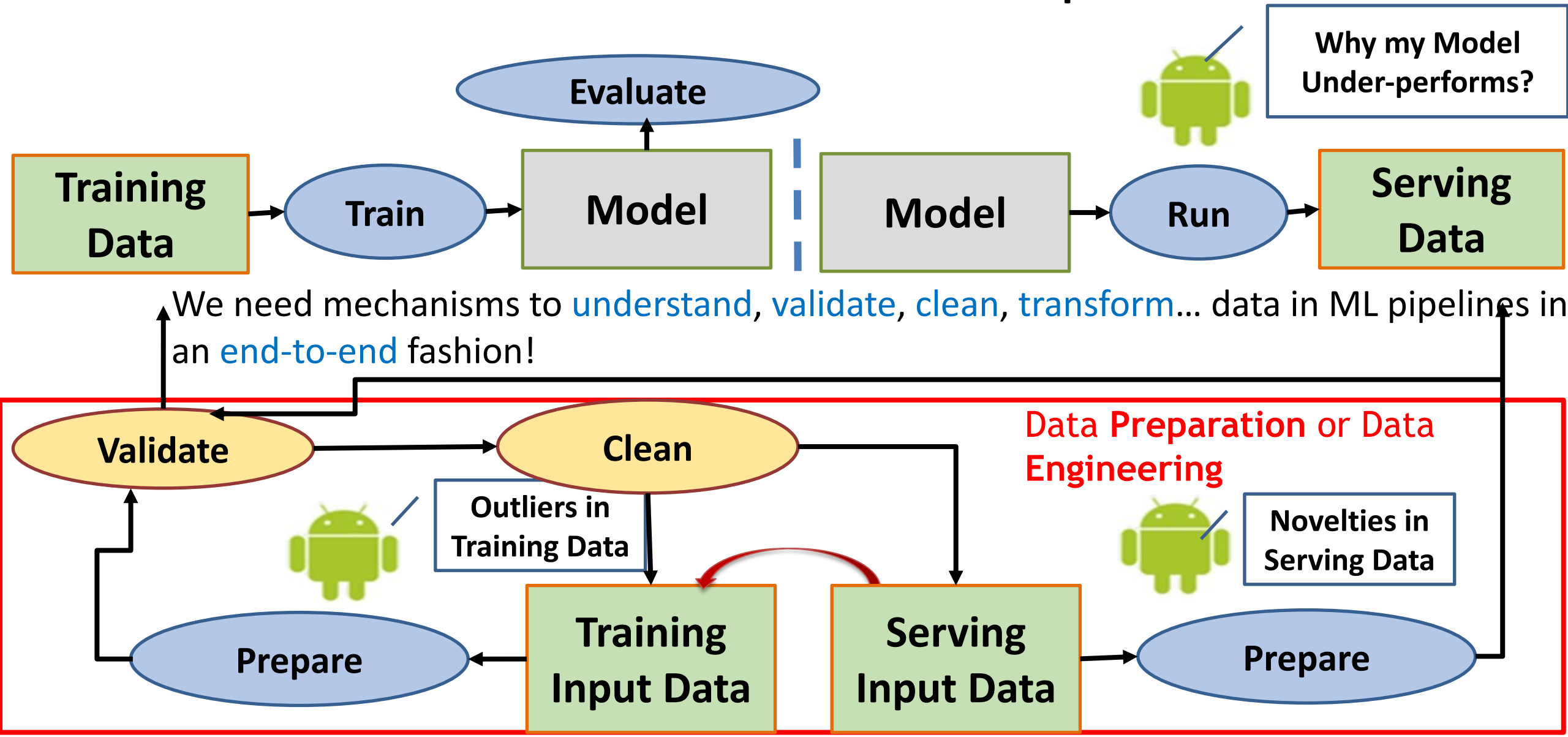


Vassilis Christophides ENSEA

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.

Point of Failure of Data-Intensive Systems: Data Quality!

- 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]
- We should also ensure that used data and algorithms, are making decisions based on individual merits and not on systematic bias that runs through our society!

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

https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect

Example: Anomalies in Healthcare Data

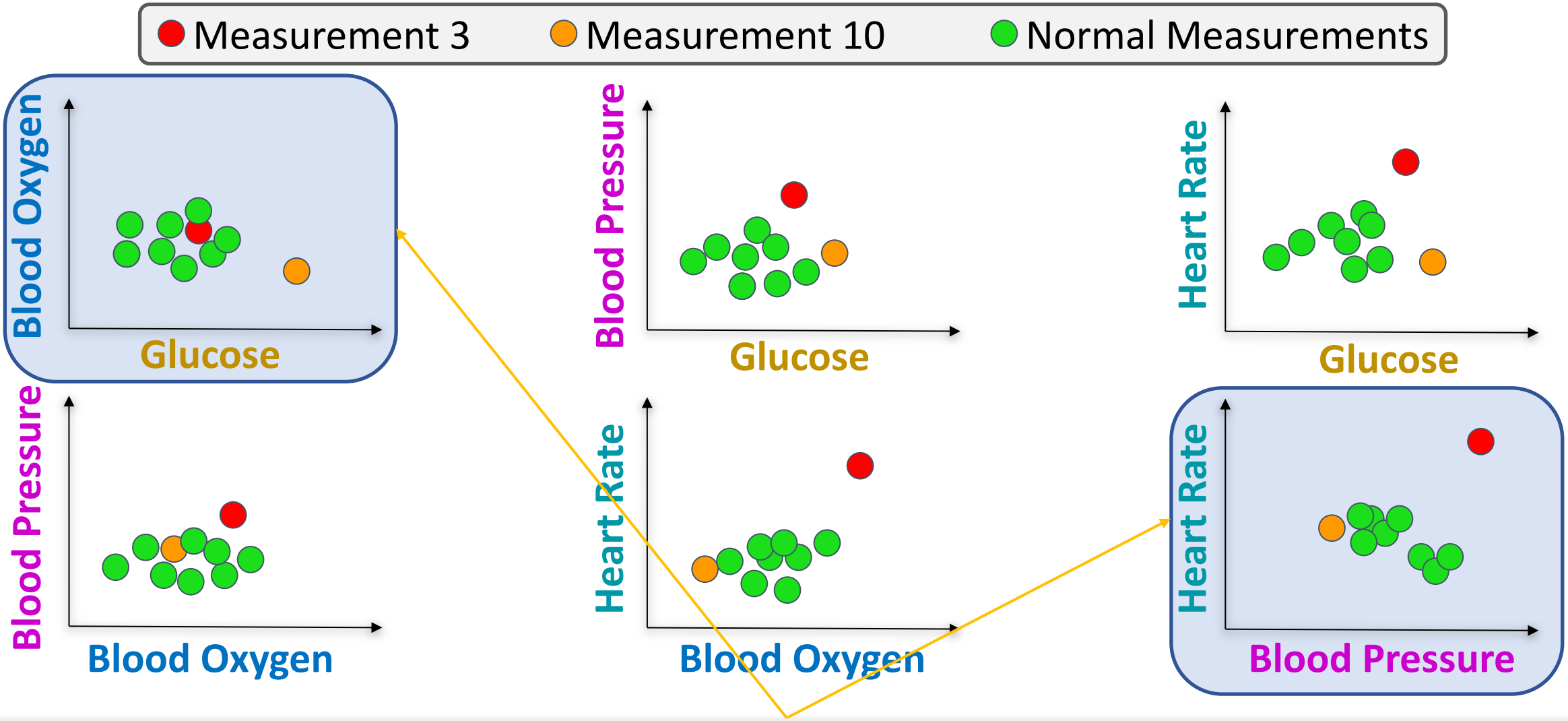
Analysis Task: Detect possible abnormal measurements for a patient

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

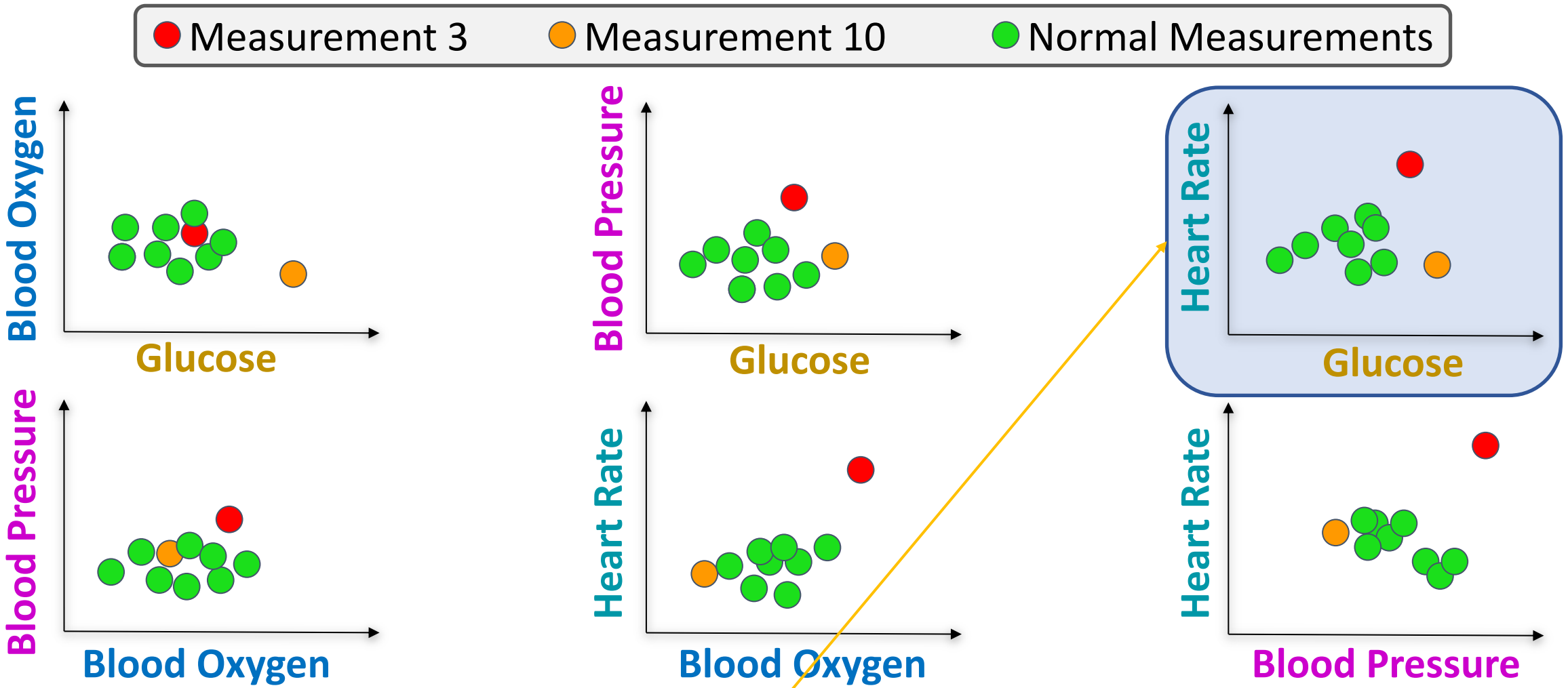
- Why did M3 and M10 get a high score?
- Should I alert patients for health problems?

2D Subspaces Explaining Anomalies: Local



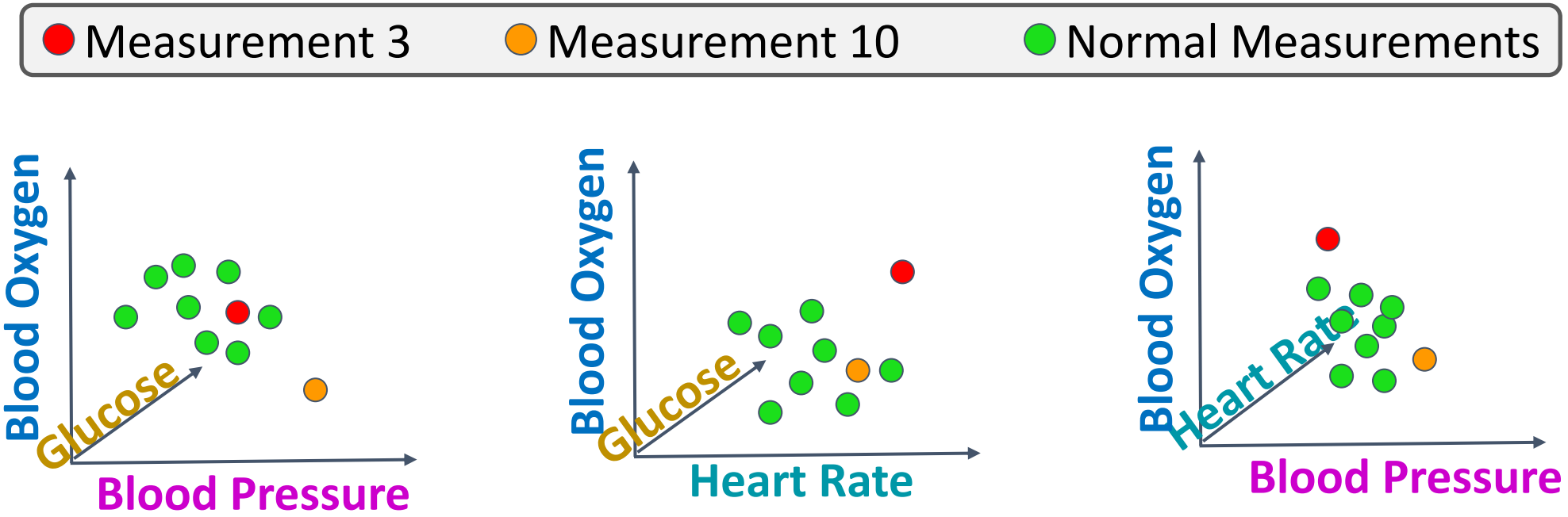
Local Explanation: Find subspaces that maximize anomalousness of individual samples

2D Subspaces Explaining Anomalies: Global



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

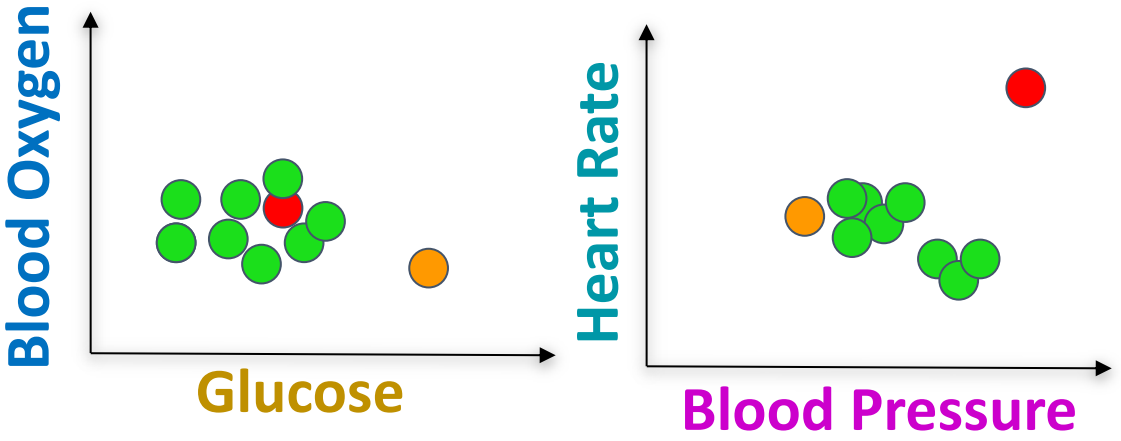
3D Subspaces Explaining Anomalies: Higher Dim.



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

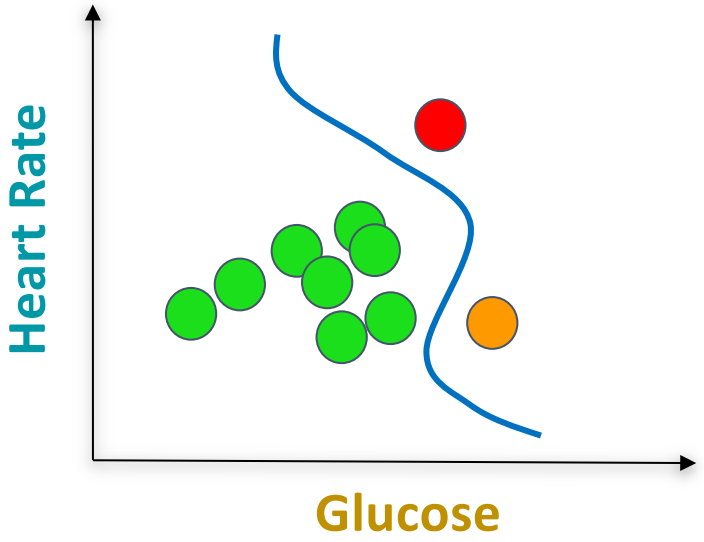
Descriptive vs Predictive Anomaly Explanations

Descriptive Explanation



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

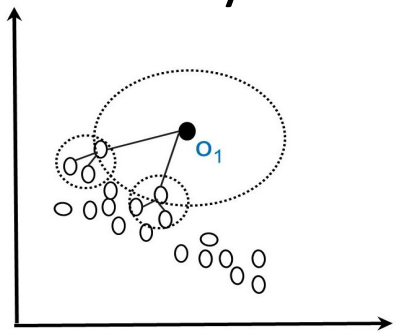
Predictive Explanation



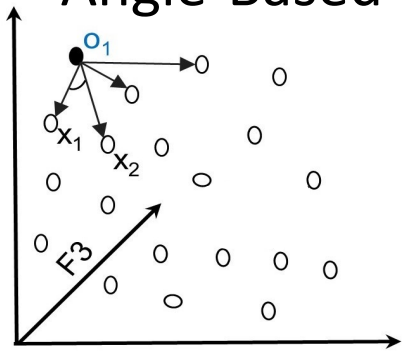
A minimal subset of features leading to a predictive model that best approximates the decision boundary of a detector

How Can We Produce Predictive Explanations ?

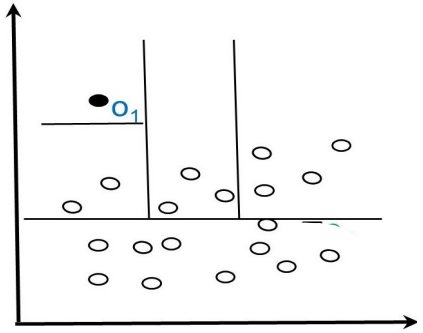
Density-Based



Angle-Based



Isolation-Based

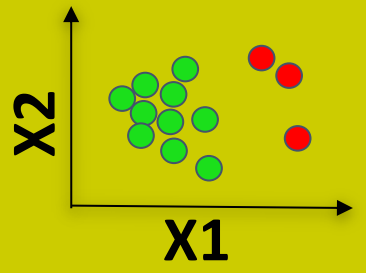


Unsupervised Anomaly Detector

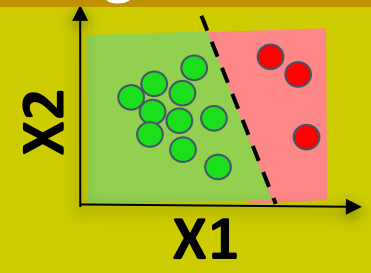
Surrogate Supervised Model

Generalization Performance Estimation

Feature Selection Algorithms

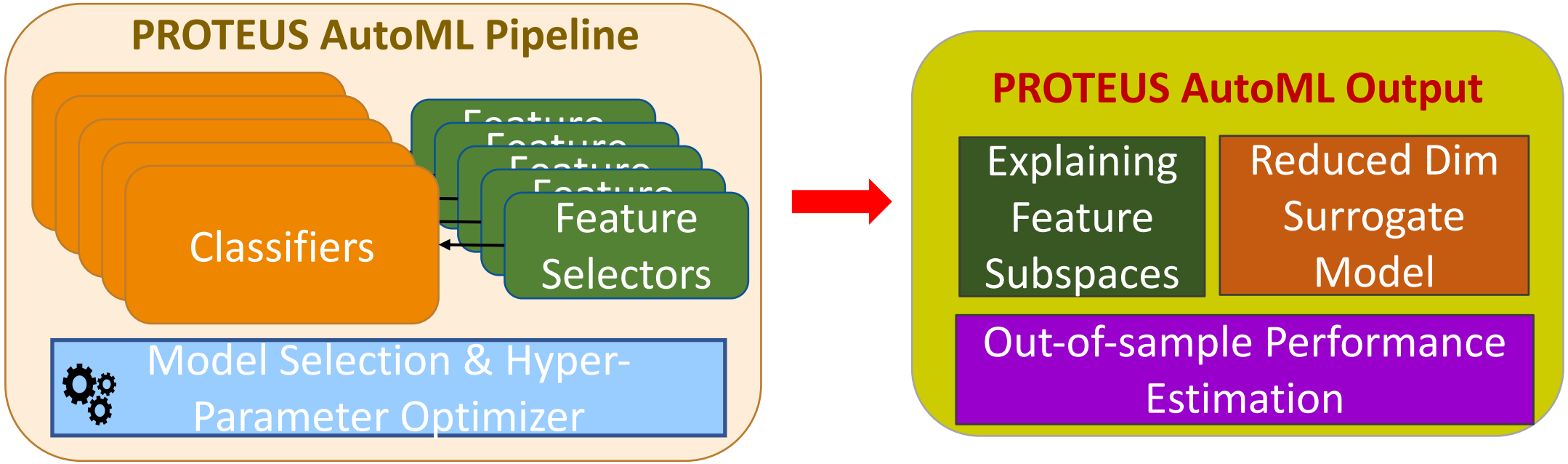


Classification Algorithms



AutoML

PROTEUS Outcome and Design Choices



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

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	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**

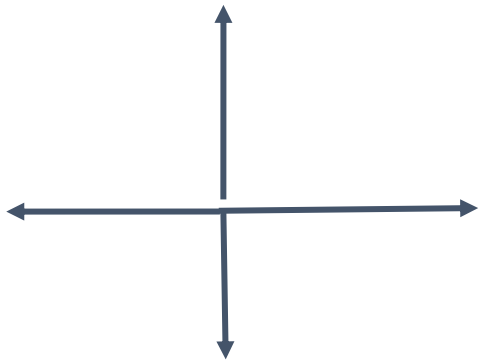
Adding irrelevant features to the **synthetic** dataset: 77%, 88%, 92%, 94%, 96%
Adding irrelevant features to every **real** dataset: 30%, 60%, 90%

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

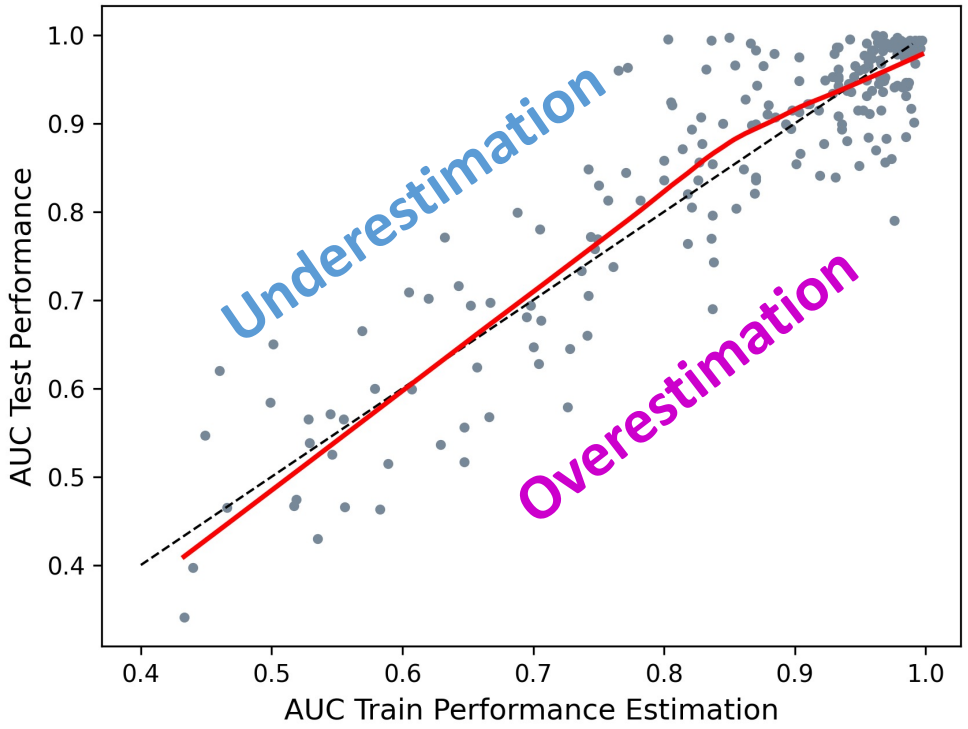
5 for the synthetic and 3 per real dataset

255 analyses

1 synthetic dataset and 3 real datasets

PROTEUS Performance Estimation

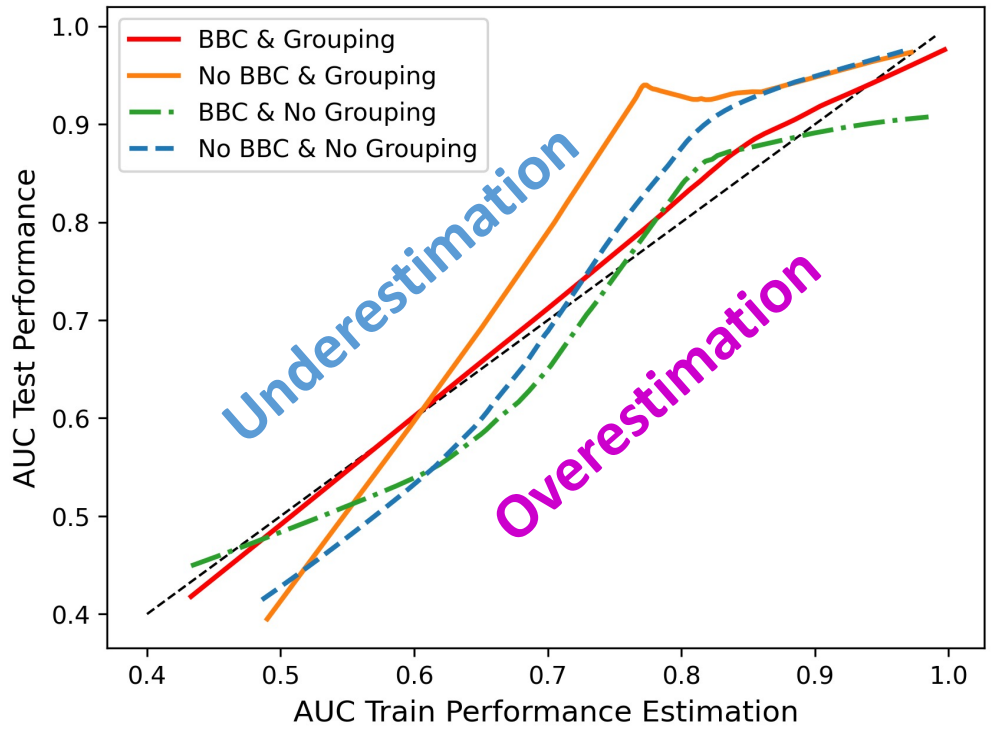
Q: Do the design choices of PROTEUS contribute to provide an accurate performance estimation ?



- 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 ?



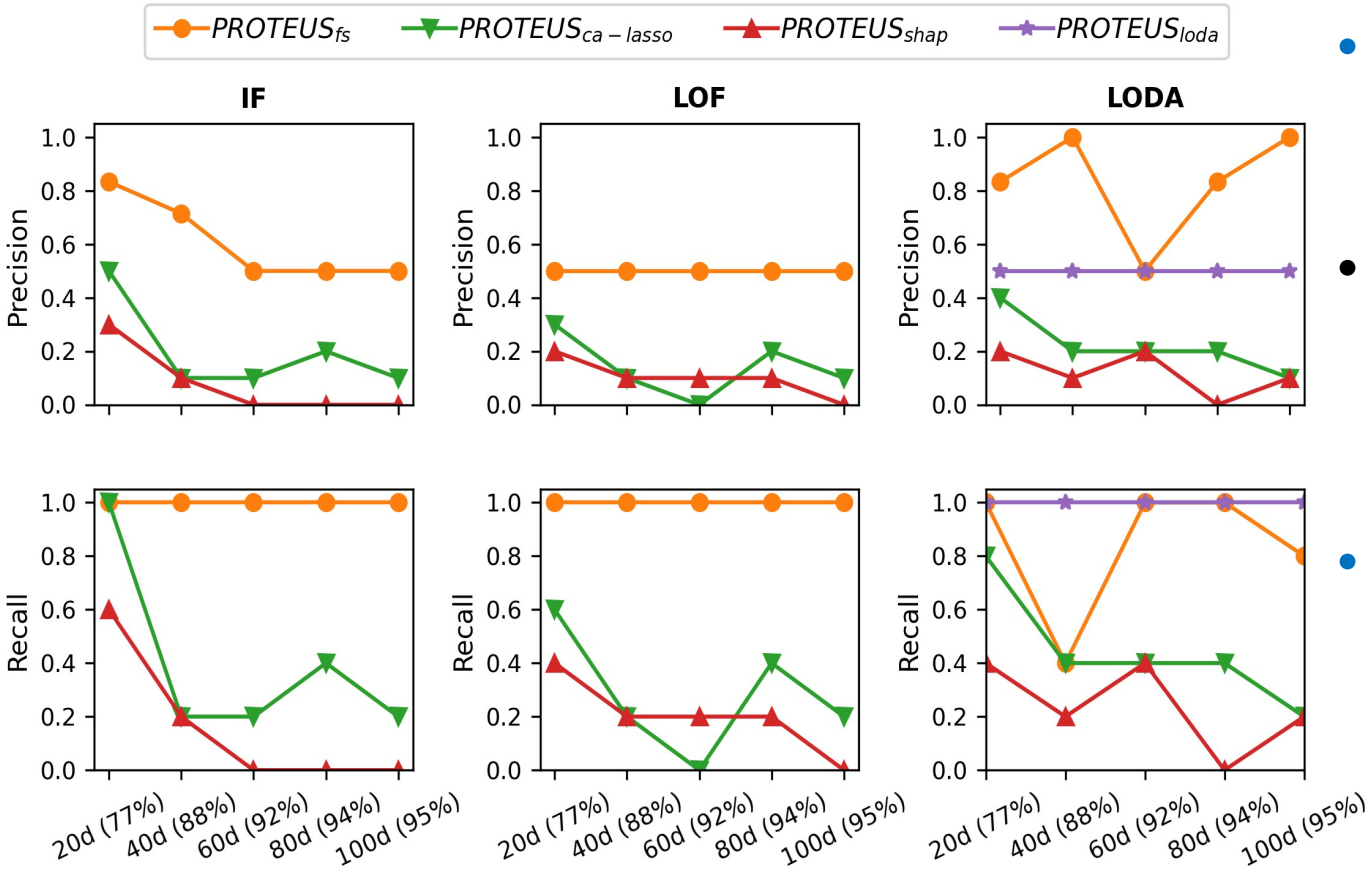
BBC & Grouping	No BBC & Grouping	BBC & No Grouping	No BBC & No Grouping
0.05	0.88	0.11	0.25

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)?



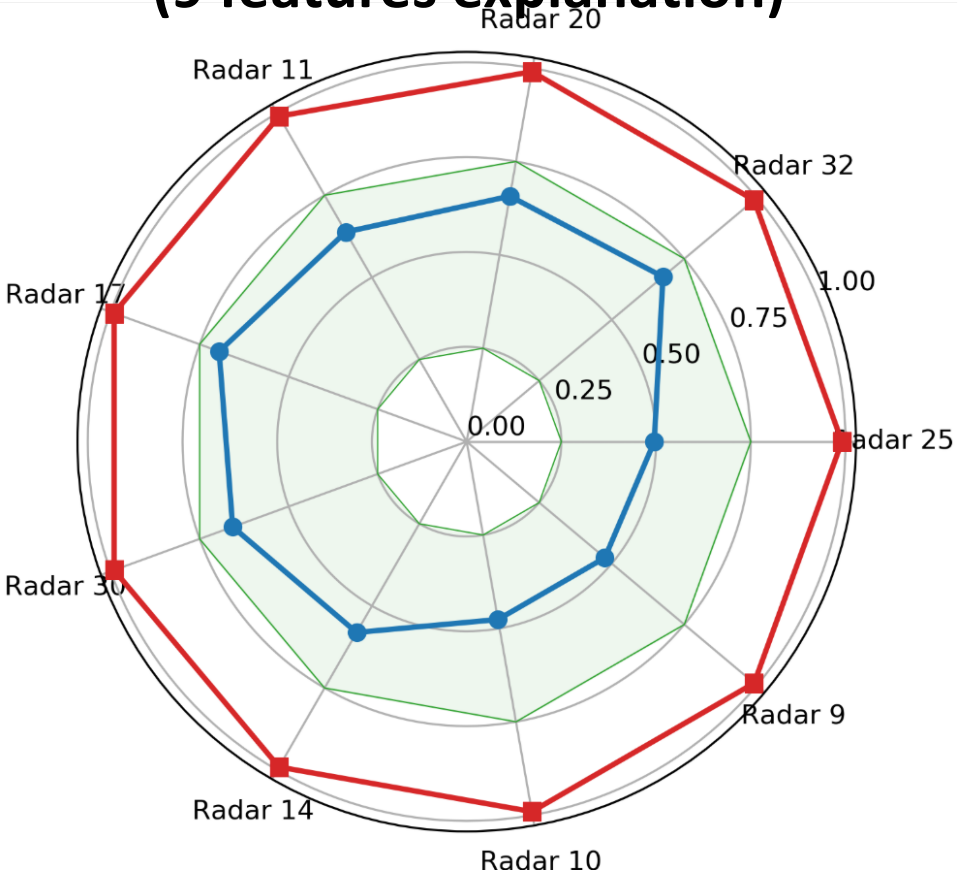
- Feature selection algorithms of PROTEUS$_{fs}$ exhibit the highest overall precision suboptimal recall
- Unlike SHAP and CA-Lasso, PROTEUS$_{fs}$ exhibits a robust performance when varying data dimensionality, regardless of the employed detector
- PROTEUS$_{fs}$ approximates well the recall of the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

Contrasting PROTEUS Surrogate Models With Unsupervised Anomaly Detectors

- Ionosphere Dataset (33 Features)

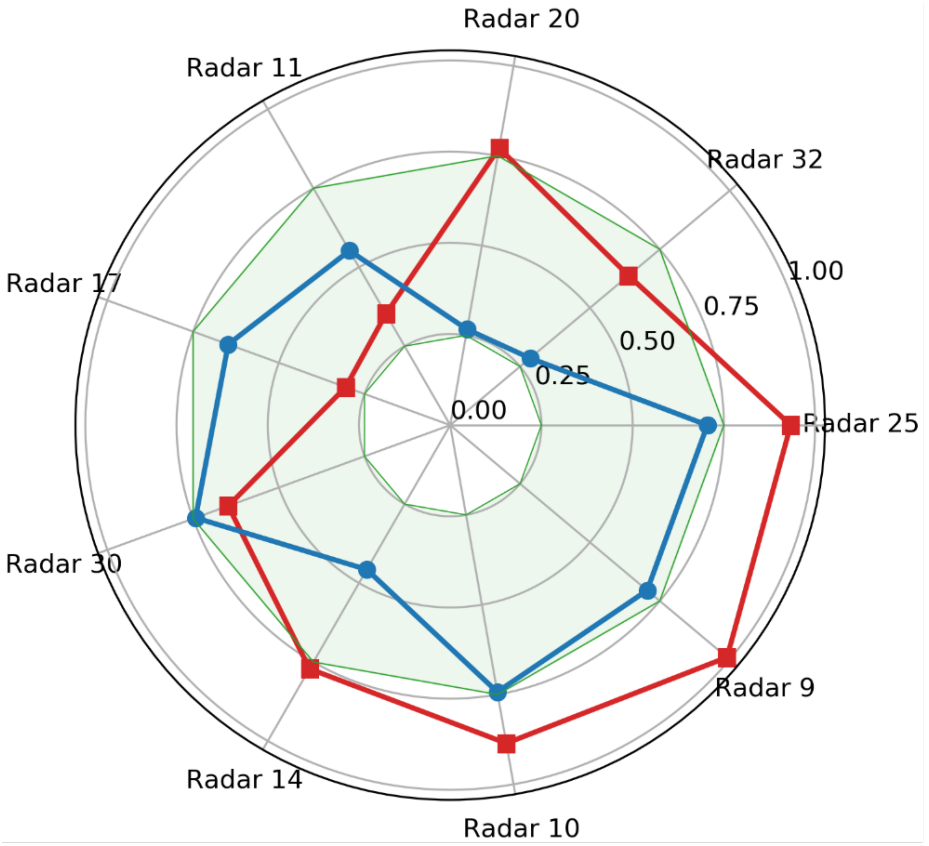
Proteus Agreement with LOF

(9 features explanation)



Proteus Disagreement with LOF

(9 features explanation)



Proteus Anomaly Proteus Normal

Conclusion

- Anomaly explanation ⟶ **a supervised classification problem with feature selection** ⟶ solved effectively as an AutoML problem
- **First** methodology for predictive, global, detector-agnostic anomaly explanations
- 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"

Acknowledgements

H.F.R.I.
Hellenic Foundation for Research & Innovation

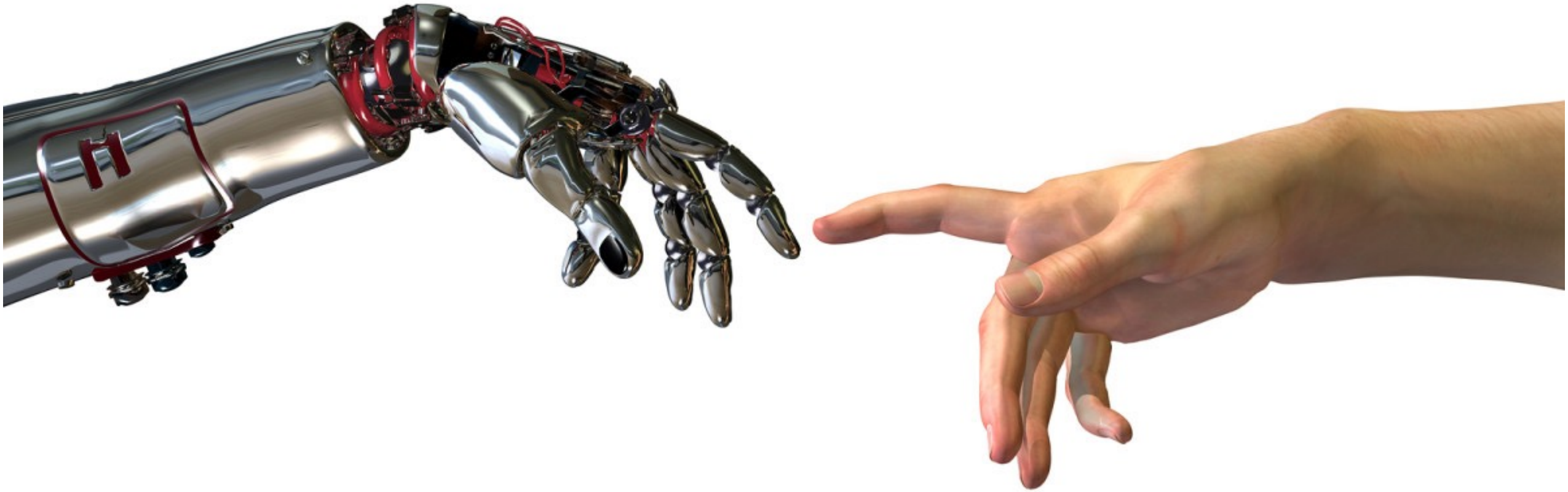
1st Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment

CERGY PARIS UNIVERSITÉ

institut d'études avancées

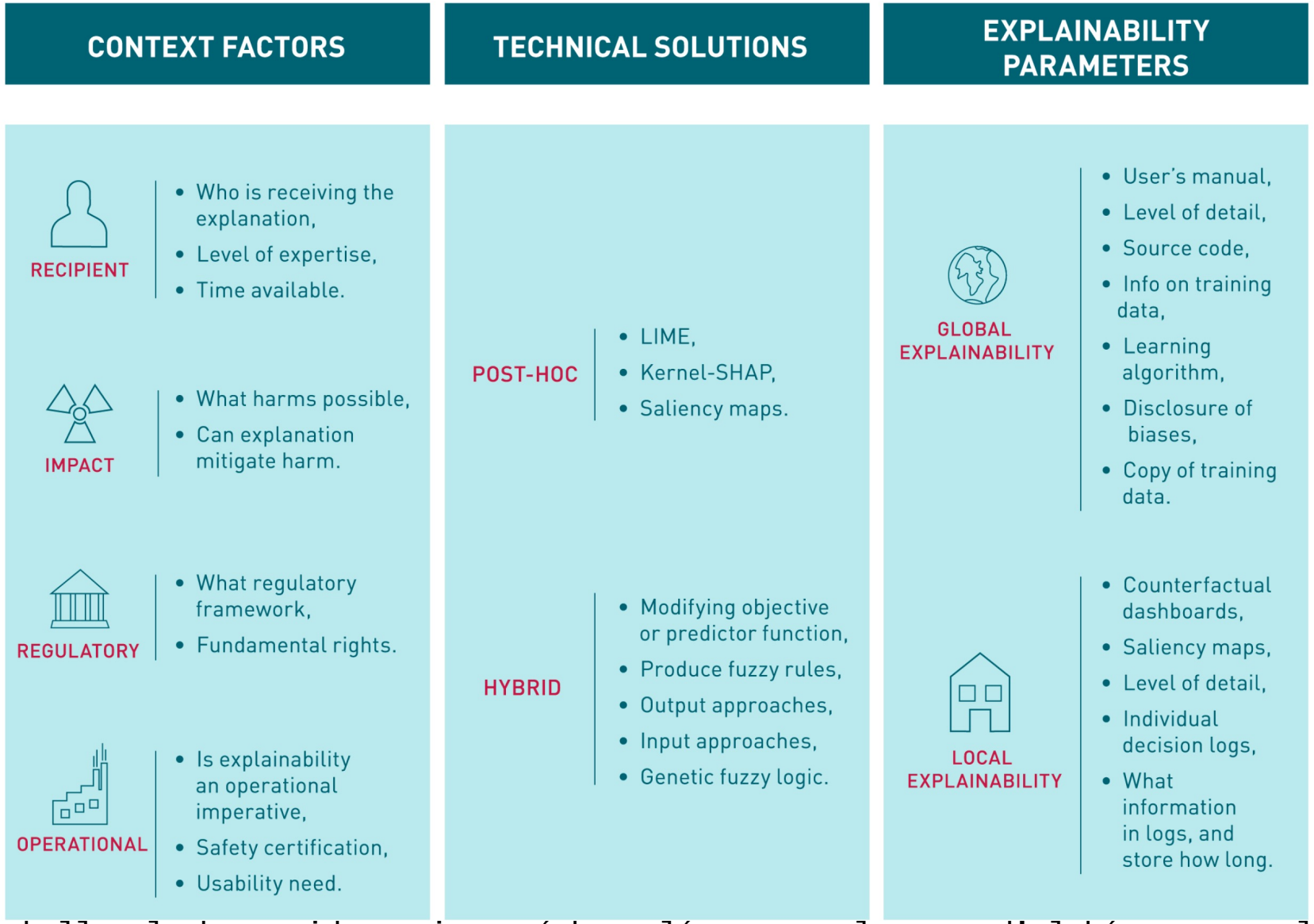
Fellows-in-Residence 2019-2020

Questions?



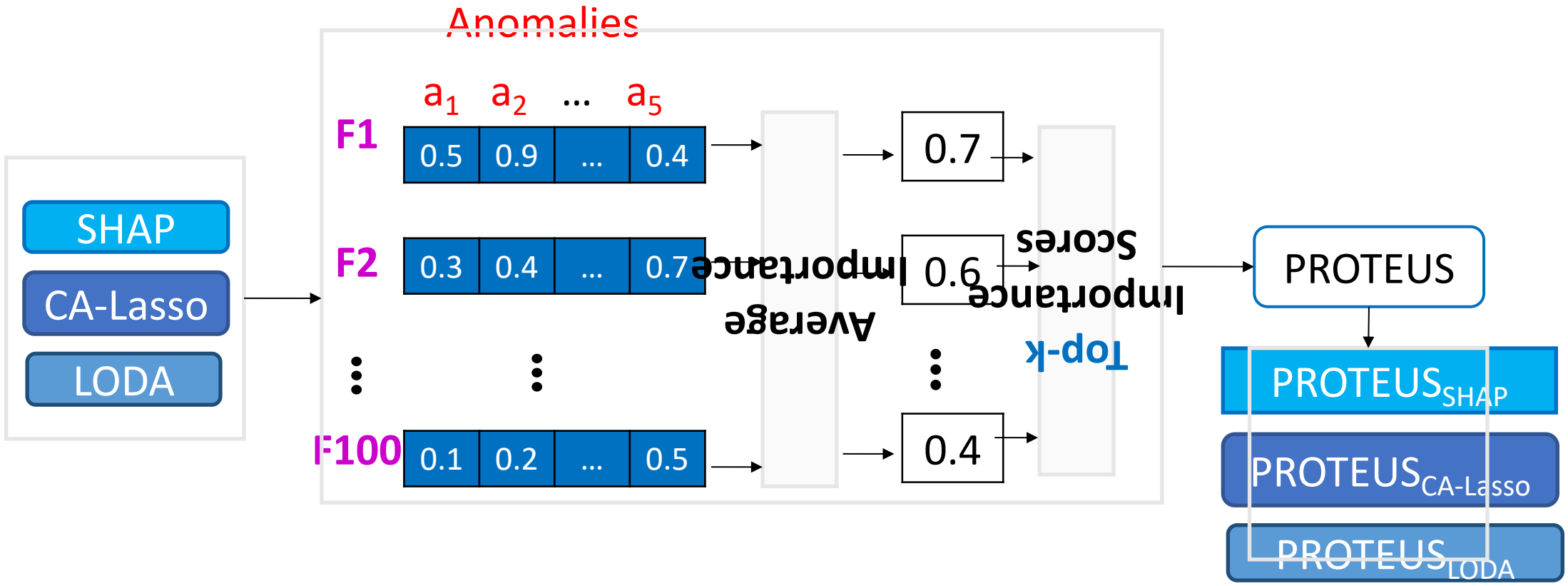
https://thenextweb.com/contributors/2018/10/06/we-need-to-build-ai-systems-we-can-trust/

The Three Pillars of xAI



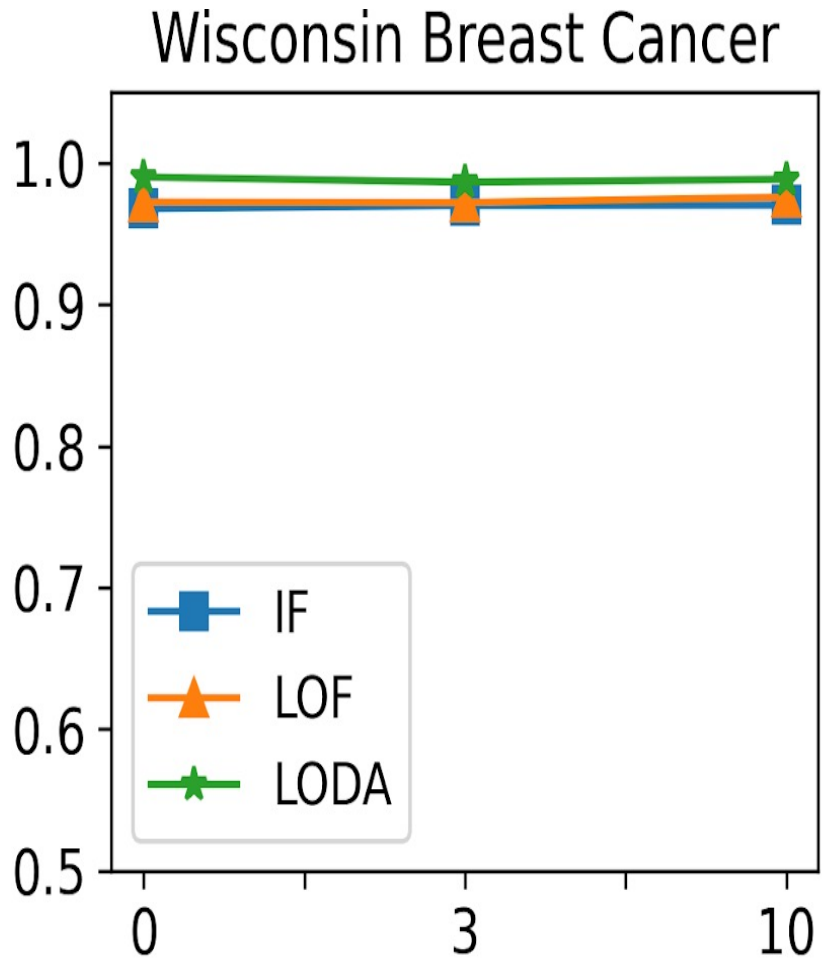
Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, et al. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. 2020. hal-02506409

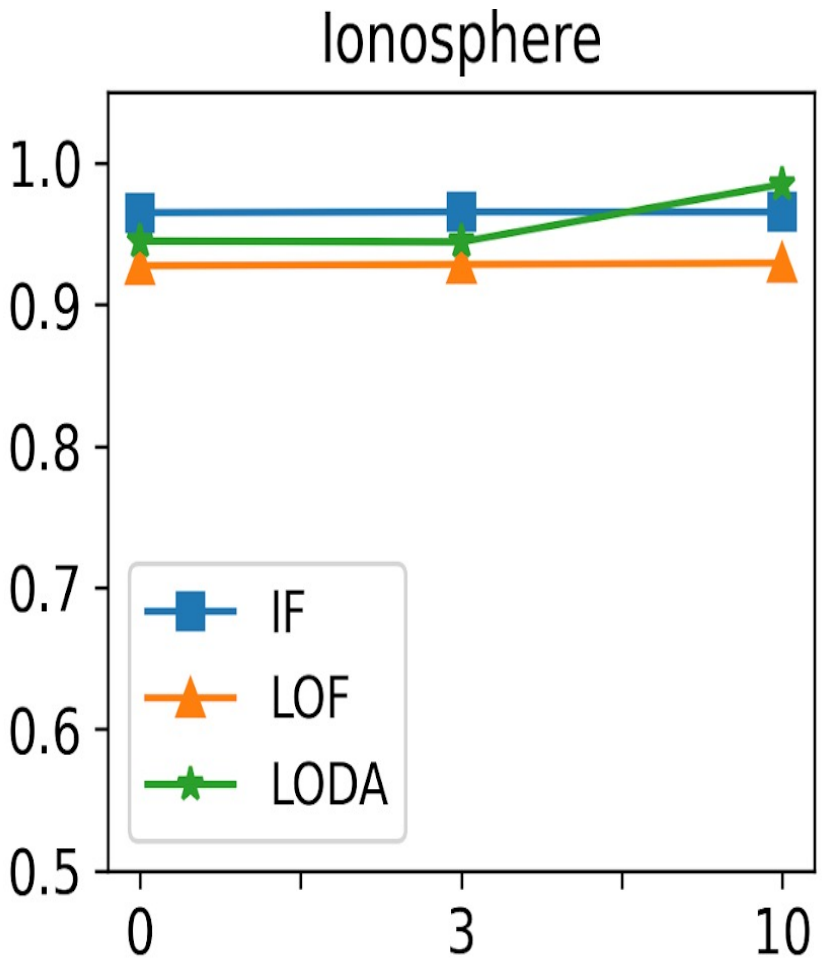
Creating a Predictive Explanation for Feature Importance Methods using PROTEUS

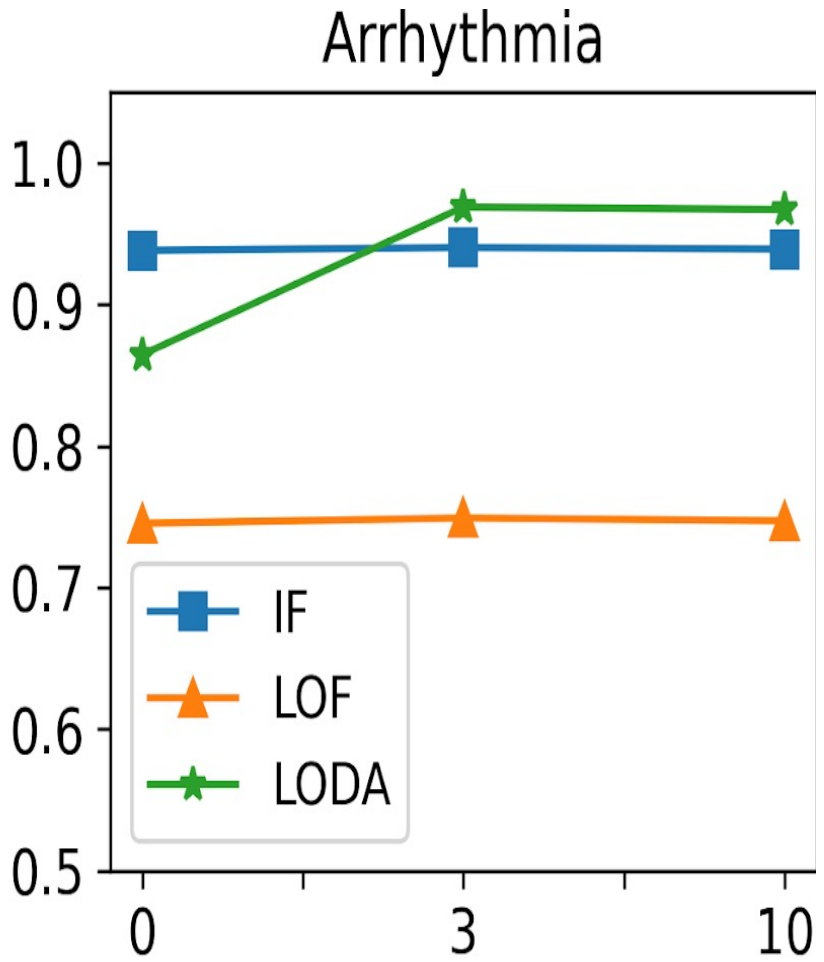


The Effect of Oversampling on Performance

- 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.*