Interpretability Approaches
applied to Predictive Models in Clinical Healthcare

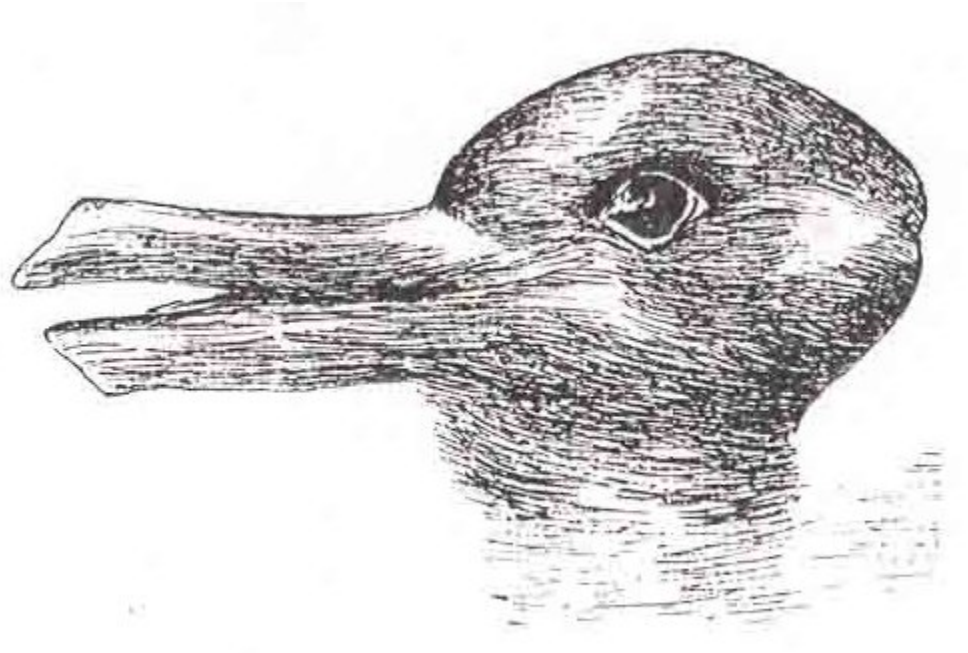
Trends in Bioinformatics

Final Presentation

Tom Martensen, Axel Stebner

Agenda

1. Recap
2. Methods
    1. Building a Clinical Prediction Model
    2. Applying Interpretability Methods in Detail
    3. Making Interpretability Available for Domain Experts
3. Results
4. Outlook



Recap:
Visions & Objectives

Hasso Plattner Institut

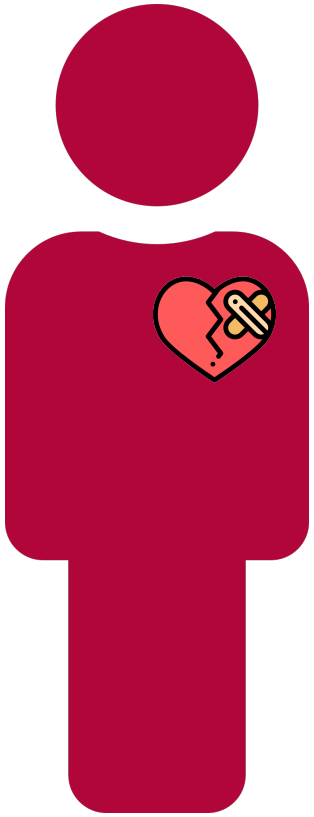
VISION 1
Find and validate medical hypotheses regarding mortality and recovery of AKI

- Train CPM
- Predict patient outcomes
- Gather interpretations
- Derive and evaluate clinical hypotheses

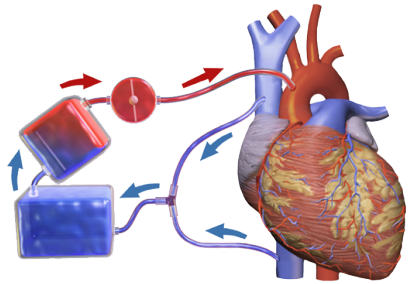
VISION 2
Make interpretations of CPMs available to physicians

- Interpret any CPM
- Make interpretations comparable side-by-side
- Show complexity-faithfulness tradeoff

Recap:
Use Case – Acute Kidney Injury



Mohammed Lee



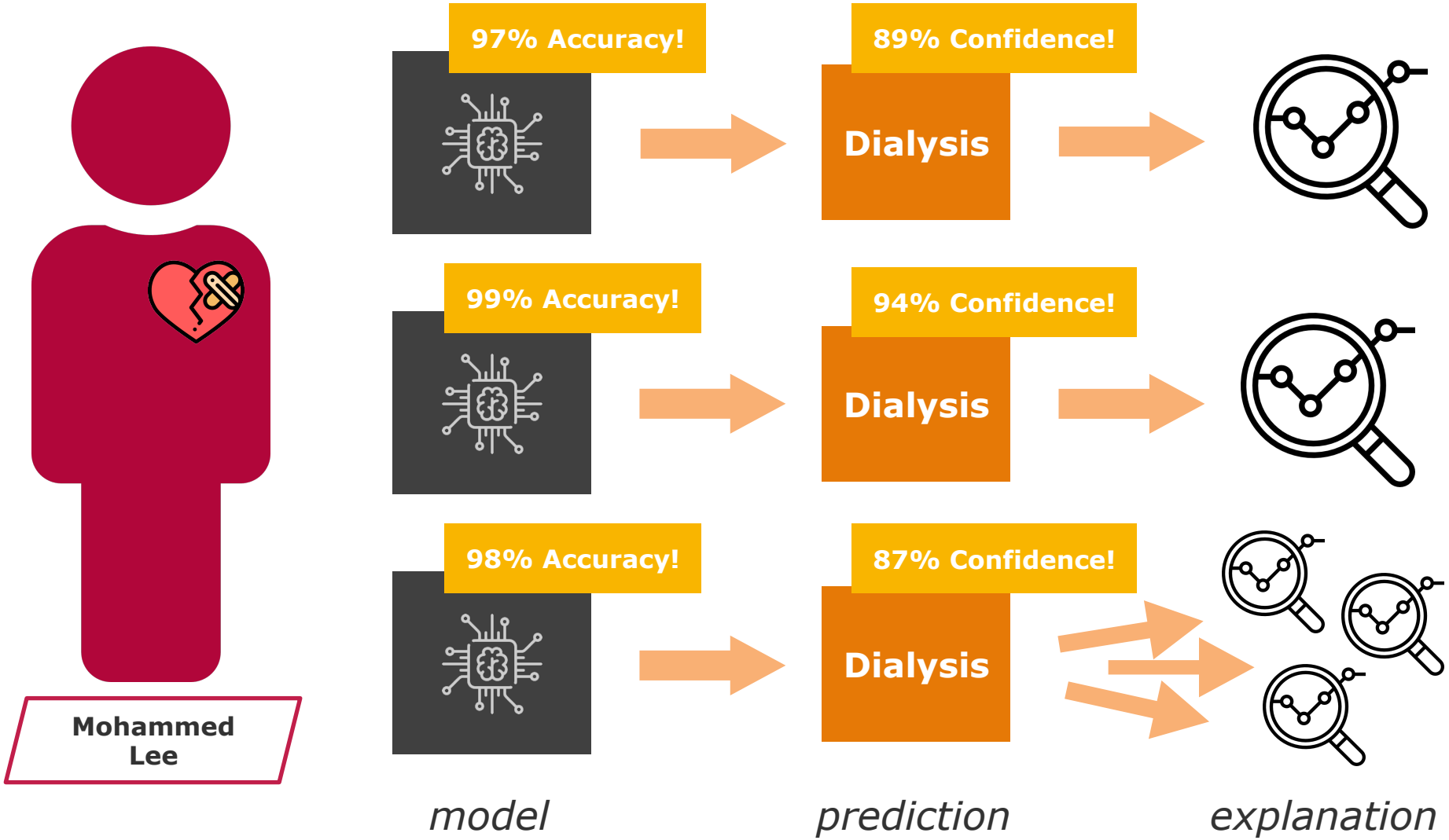
Cardiopulmonary Bypass

- Decreased Blood Flow
- Decreased Kidney Perfusion

Acute Kidney Injury (AKI)

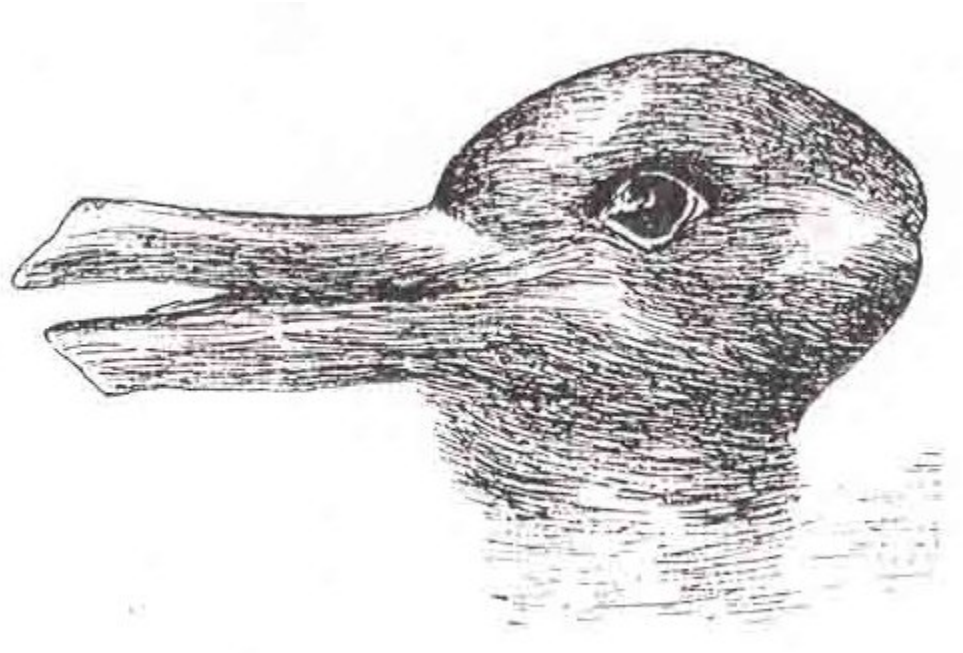
- Intermittent Hemodialysis (IHD)
- Continuous Renal Replacement Therapy (CRRT)

Recap:
Use Case – Therapy of Acute Kidney Injury



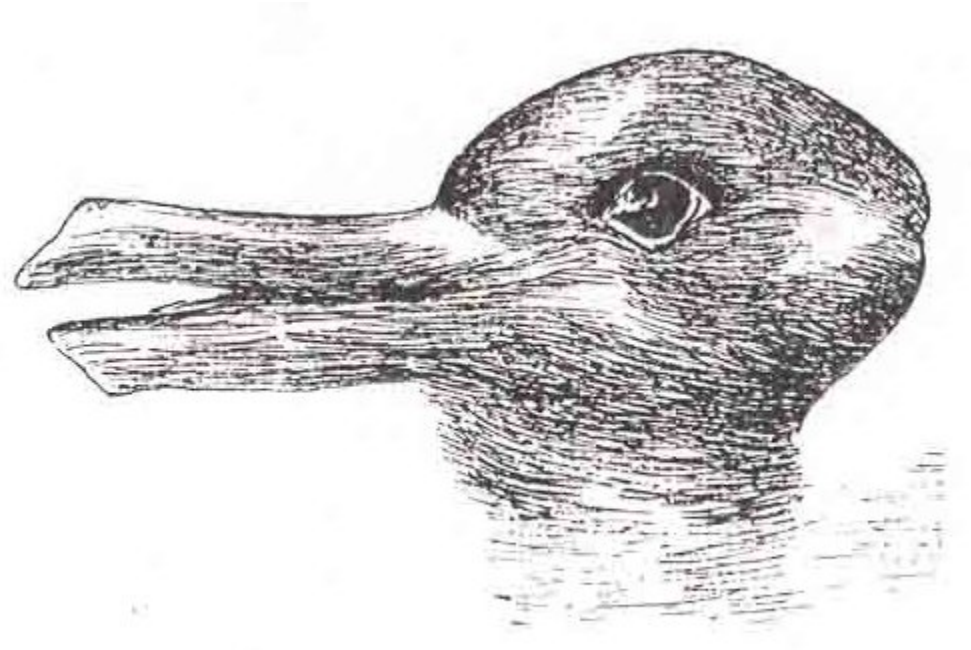
Methods

- Building a Clinical Prediction Model
- Applying Interpretability Methods in Detail
- Making Interpretability Available for Domain Experts



Methods

- **Building a Clinical Prediction Model**
- Applying Interpretability Methods in Detail
- Making Interpretability Available for Domain Experts



Methods:
Building a Clinical Prediction Model

Building a Clinical Prediction Model: Data Retrieval

HPI Hasso Plattner Institut

MIMIC-III Database

LAB EVENTS	ICU STAYS	ICU STAY VITALS (FIRST DAY)
▪ Different lab values ▪ Flagged ▪ Timestamp	▪ Start ▪ End	Aggregated lab values of first day of ICU stay

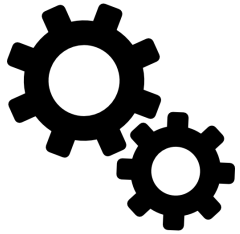
Procedure Events	Labels	AKI Patients

▪ All procedures in hospital
▪ Timestamp
	▪ Labels for classification:
	  ▪ Dosage
	  ▪ Therapy type
		▪ Patient master data
		▪ Only patients with AKI

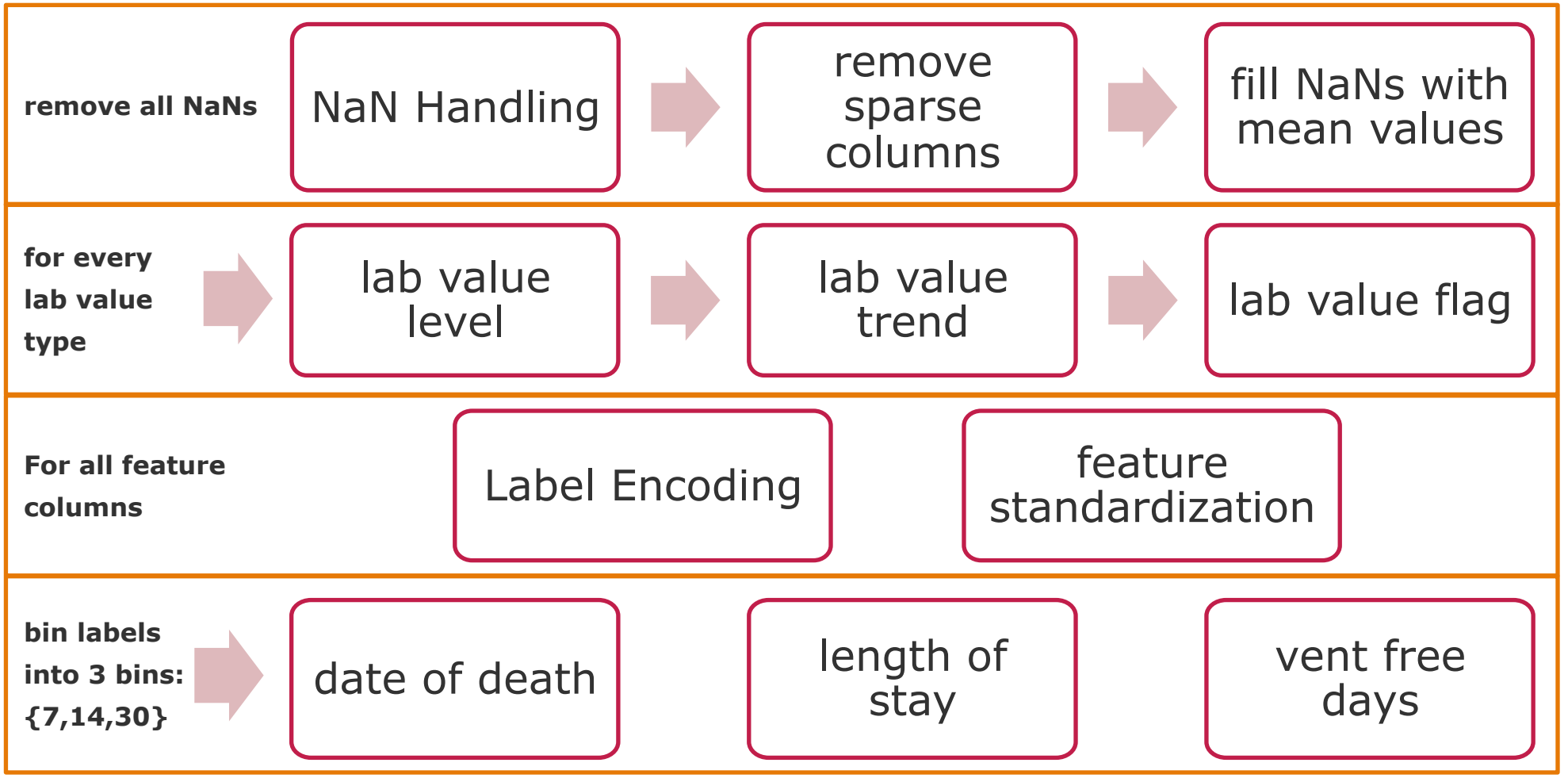
Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **13**

Building a Clinical Prediction Model: Data Preprocessing



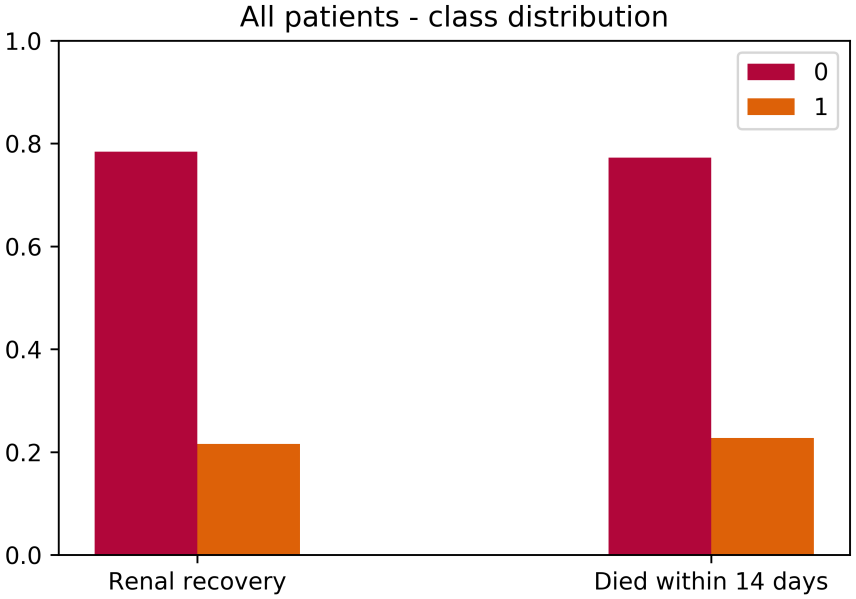
Hasso Plattner Institut



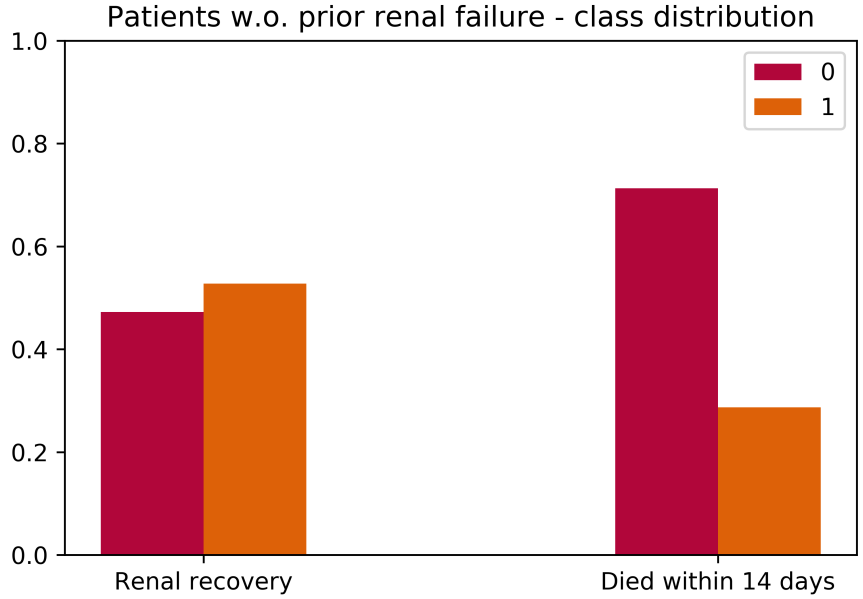
Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **14**

Building a Clinical Prediction Model: Data Preprocessing – Dataset Characteristics



→ 2945 instances



→ 944 instances

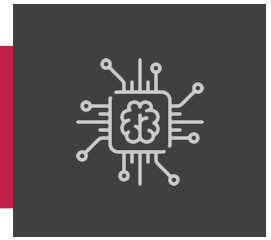
Building a Clinical Prediction Model: Model Training

2000

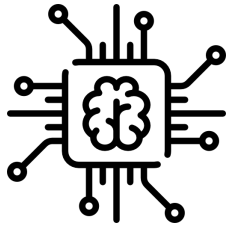
Random Parameter Search:

- Randomly pick parameters from specified range
- Create classifier
- 5-fold cross validation
- Evaluate with AUROC score

Trained model with optimal parameter setting



Building a Clinical Prediction Model: Prediction Patient Outcomes



Hasso Plattner Institut

results_gb_all_0_DIED_14_DAYS.dat
{'criterion': 'friedman_mse', 'loss': 'exponential', 'max_depth': 160, 'max_leaf_nodes': 653, 'min_samples_leaf': 38, 'n_estimators': 740}

results_gb_all_0_RENAL_RECOVERY.dat
{'criterion': 'mse', 'loss': 'exponential', 'max_depth': 77, 'max_leaf_nodes': 202, 'min_samples_leaf': 68, 'n_estimators': 841}

results_gb_not_all_0_RENAL_RECOVERY.dat
{'criterion': 'friedman_mse', 'loss': 'deviance', 'max_depth': 5, 'max_leaf_nodes': 569, 'min_samples_leaf': 15, 'n_estimators': 903}

results_gb_not_all_0_DIED_14_DAYS.dat
{'criterion': 'mse', 'loss': 'exponential', 'max_depth': 120, 'max_leaf_nodes': 362, 'min_samples_leaf': 14, 'n_estimators': 165}

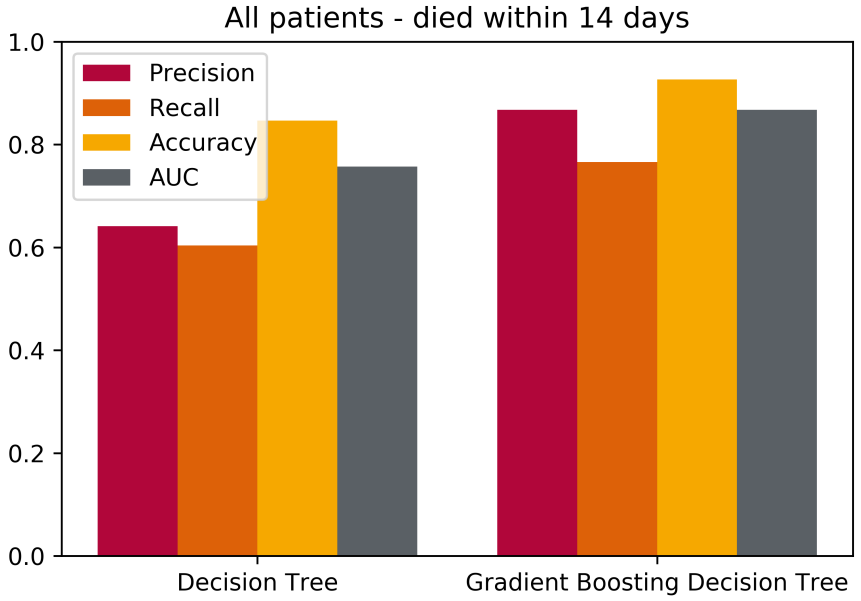
results_dt_not_all_0_RENAL_RECOVERY.dat
{'criterion': 'gini', 'max_depth': 34, 'max_leaf_nodes': 941, 'min_samples_leaf': 6}

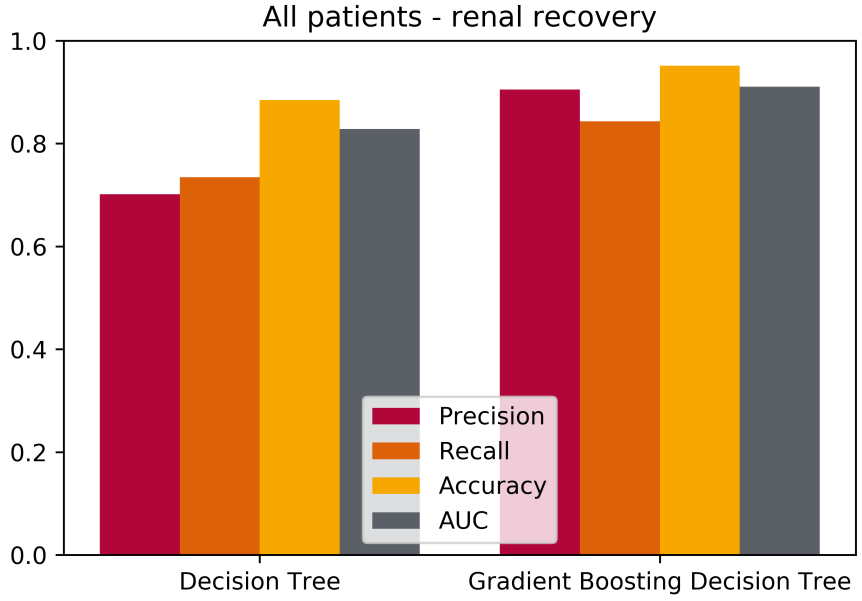
results_dt_all_0_RENAL_RECOVERY.dat
{'criterion': 'gini', 'max_depth': 50, 'max_leaf_nodes': 965, 'min_samples_leaf': 9}

results_dt_all_0_DIED_14_DAYS.dat
{'criterion': 'gini', 'max_depth': 142, 'max_leaf_nodes': 522, 'min_samples_leaf': 4}

results_dt_not_all_0_DIED_14_DAYS.dat
{'criterion': 'gini', 'max_depth': 127, 'max_leaf_nodes': 315, 'min_samples_leaf': 14}

Building a Clinical Prediction Model: Prediction Patient Outcomes



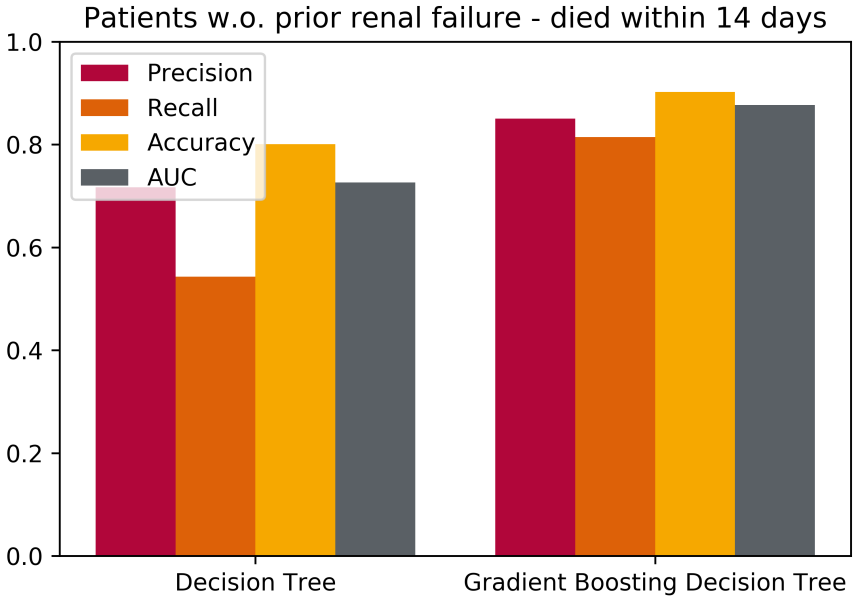


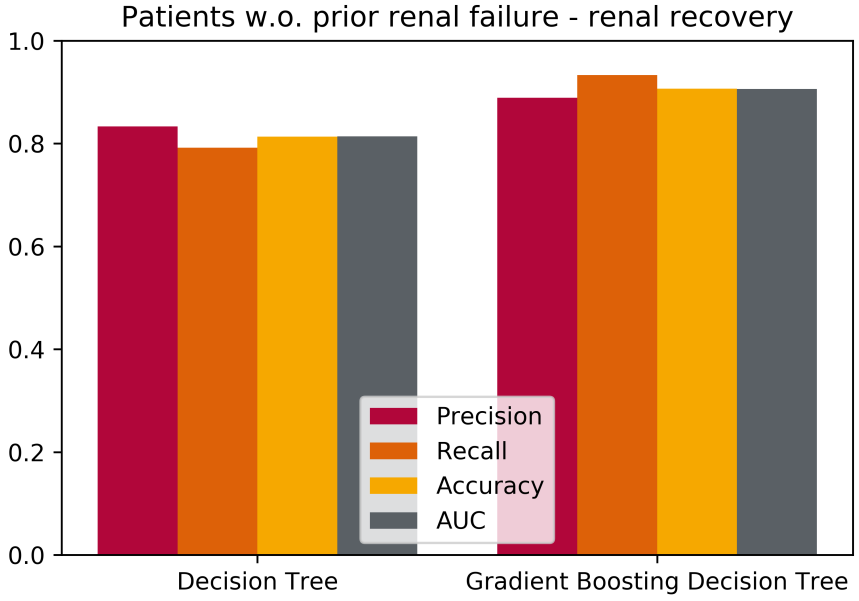
Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **21**

data retrieval > preprocessing > model training > prediction

Building a Clinical Prediction Model: Prediction Patient Outcomes



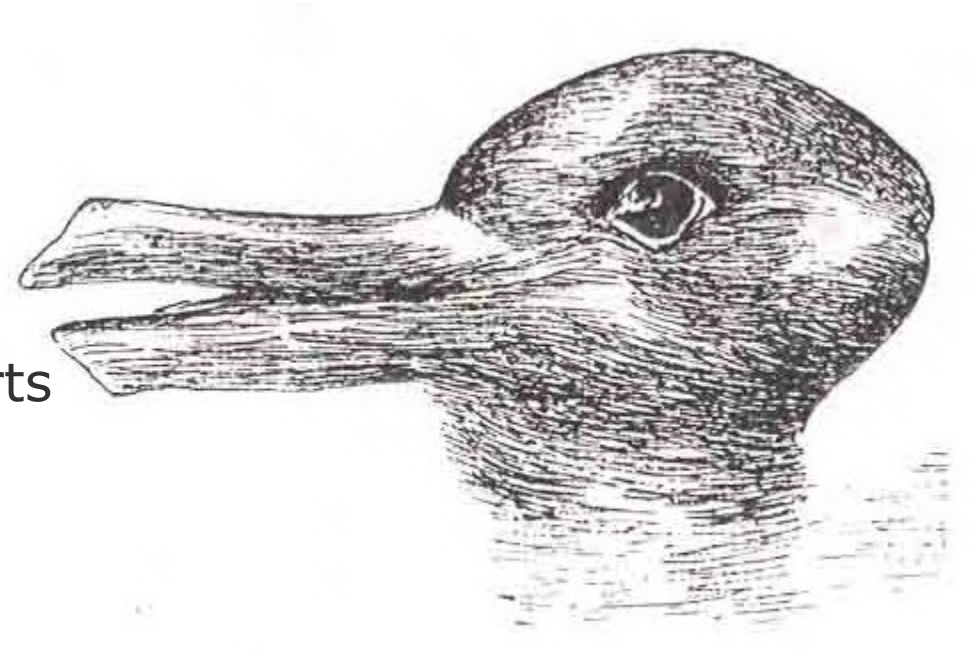


Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **22**

Methods

- Building a Clinical Prediction Model
- **Applying Interpretability Methods in Detail**
- Making Interpretability Available for Domain Experts



Methods:
Applying Interpretability Methods in Detail

- Model-based feature importances
- Global Surrogate
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

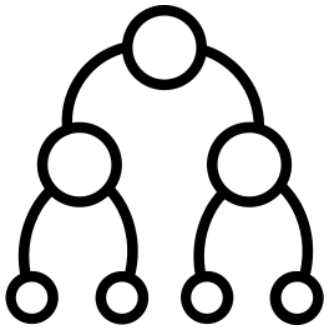
Methods:
Applying Interpretability Methods in Detail

- **Model-based feature importances**
- Global Surrogate
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

Applying Interpretability Methods in Detail: Model-based Feature Importances

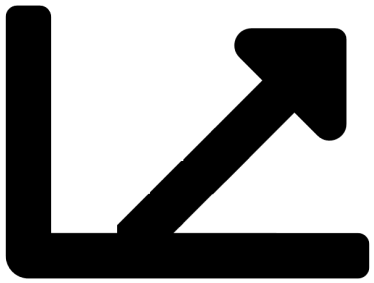
Decision Tree:

= Gini importance



Linear Regression:

Coefficients of linear function



Applying Interpretability Methods in Detail: Model-based Feature Importances

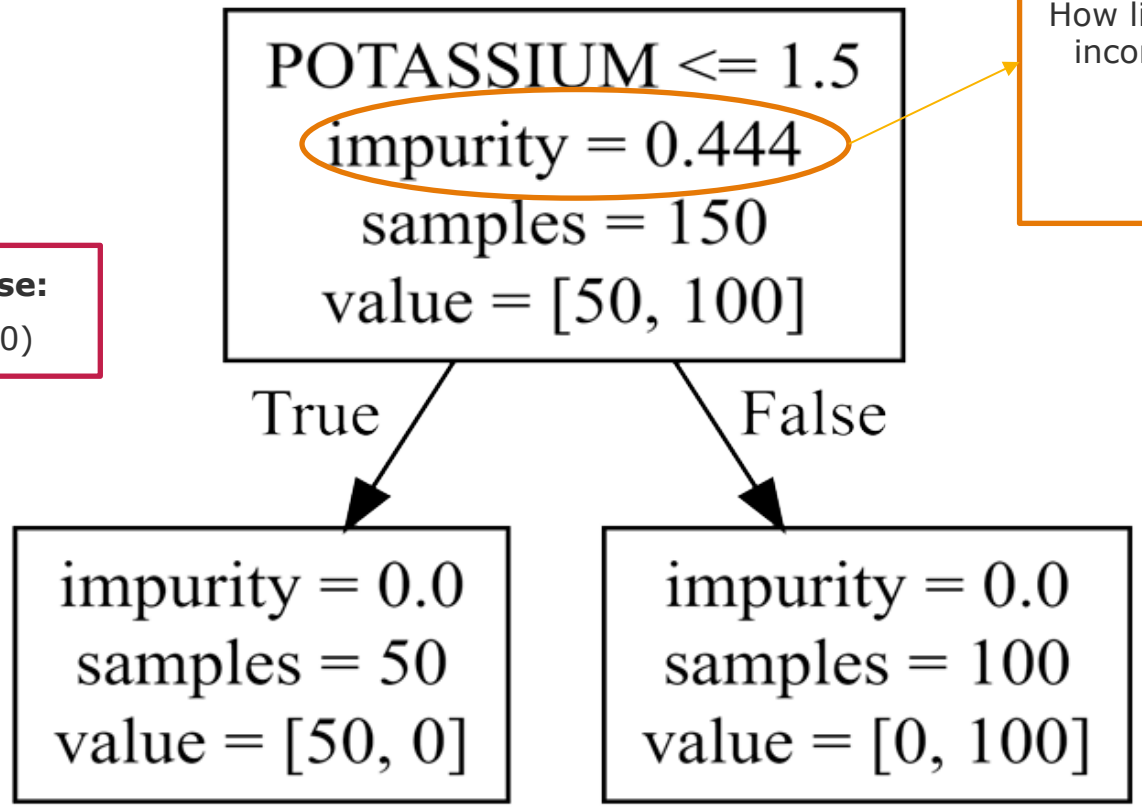
Decision Tree:

Gini importance

= impurity decrease to descendent nodes

Impurity Decrease:

0.444 – (0.0 + 0.0)



Gini Impurity:

How likely is it to randomly label an instance incorrect, based on the distribution of the label?

Source: https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/tree/tree.py

Applying Interpretability Methods in Detail: Model-based Feature Importances

Advantages:

+ Highly compressed, global insight
+ Availability

Disadvantages:

– Faithfulness linked to the error of the model
– Understandability for lay person
– Definition differs per model type

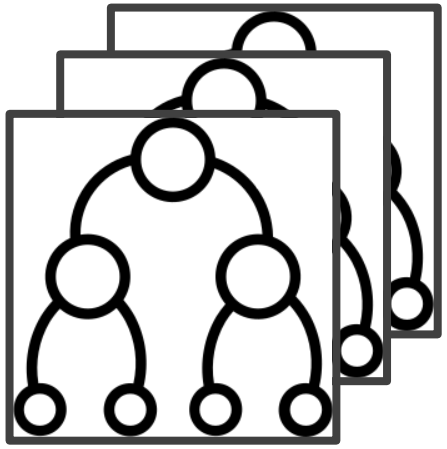
Methods:
Applying Interpretability Methods in Detail

- Model-based feature importances
- **Global Surrogate**
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

Applying Interpretability Methods in Detail: Global Surrogate

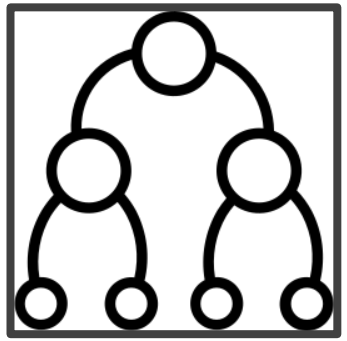
IDEA:

Approximate complicated model output with simpler model



Random forest classifier

Predictions: [0, 1, 0, 1, 1, 0]

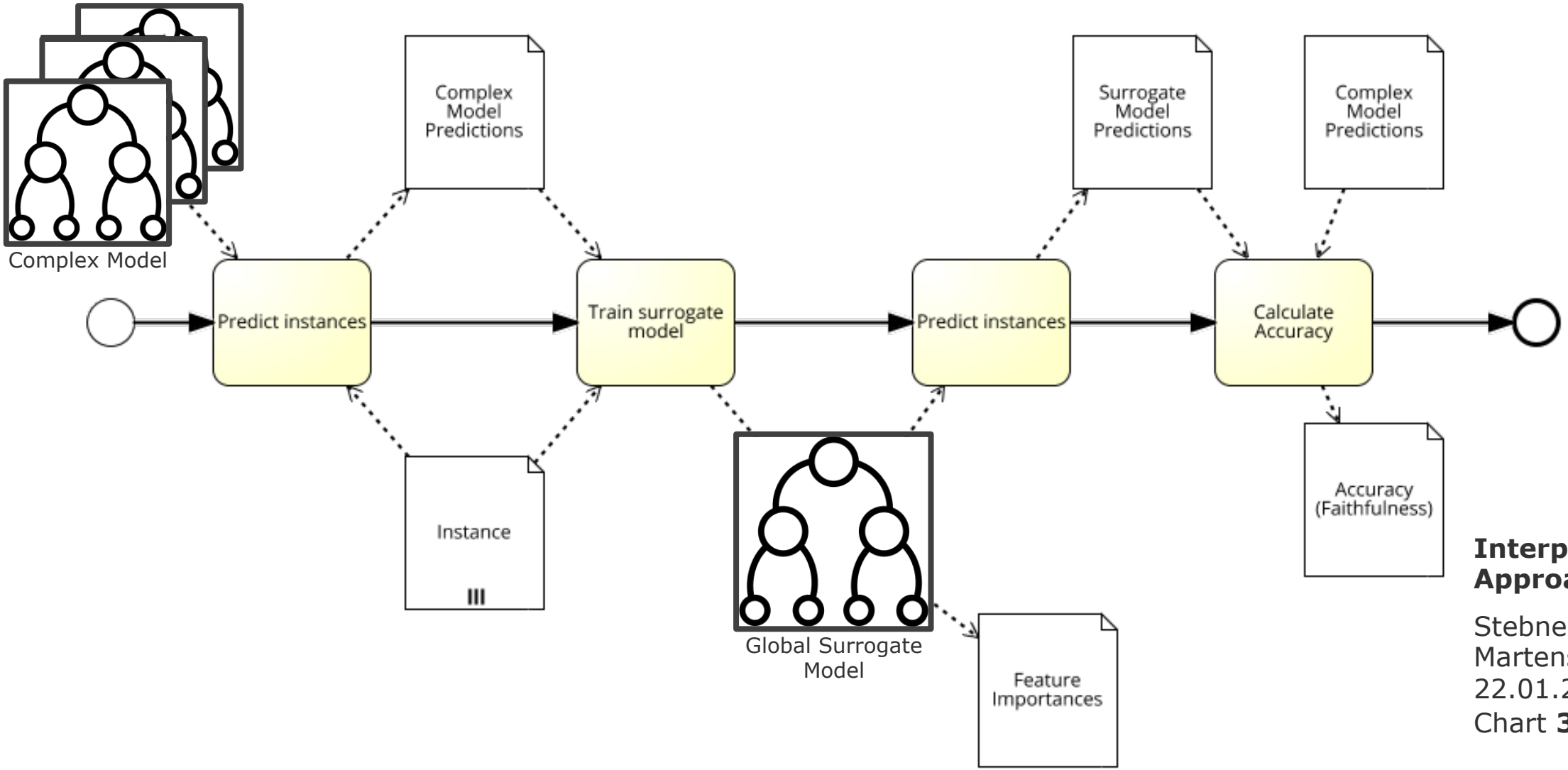


Decision Tree (Surrogate)

Predictions: [0, 0, 0, 1, 1, 0]

→ 83.33 % accuracy

Applying Interpretability Methods in Detail: Global Surrogate



Applying Interpretability Methods in Detail: Global Surrogate

Advantages:

+ Applicable to any original model (model-agnostic)
+ Surrogate models are “arguably” intuitive
+ Approximation easily measurable

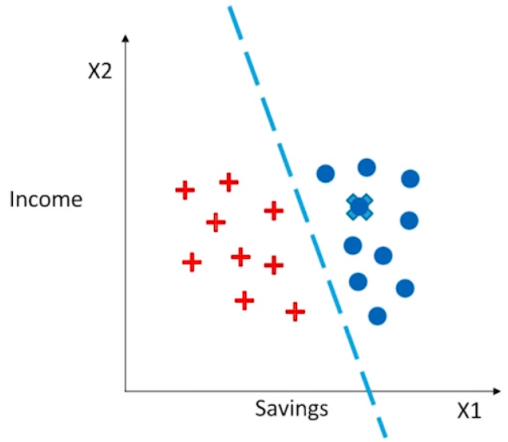
Disadvantages:

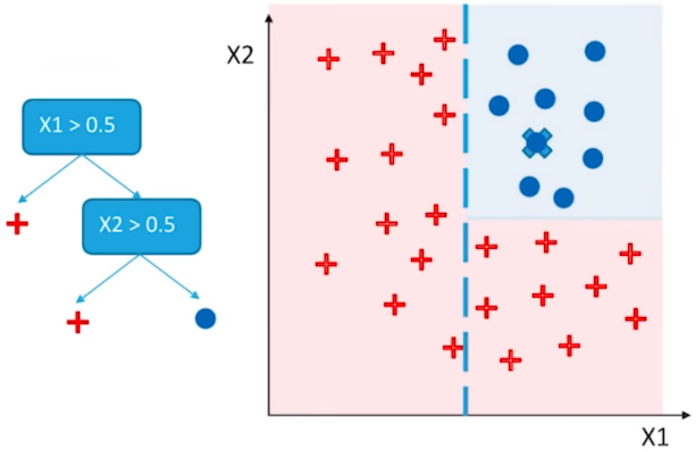
– Conclusions about model and not data
– Close for one subset of data, divergent for another?
– Intrinsically interpretable models?

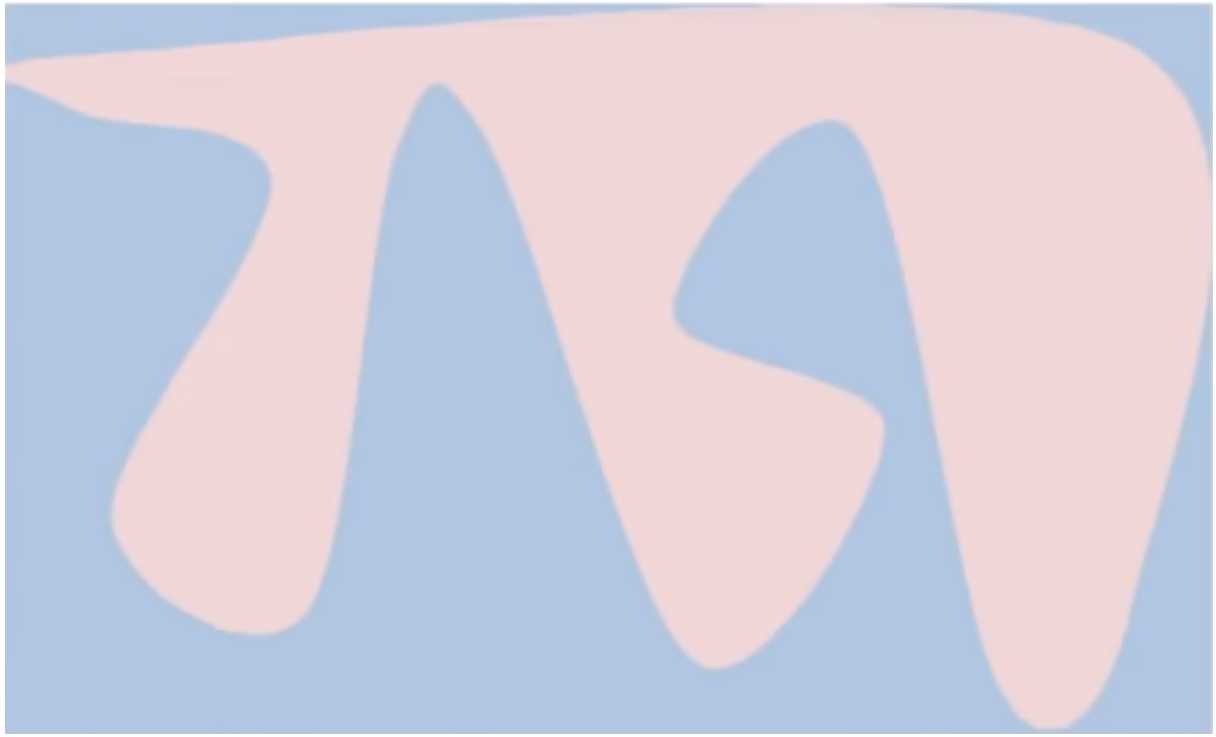
Methods:
Applying Interpretability Methods in Detail

- Model-based feature importances
- Global Surrogate
- **Local Interpretable Model-Agnostic Explanations (LIME)**
- Shapley values

Applying Interpretability Methods in Detail: LIME





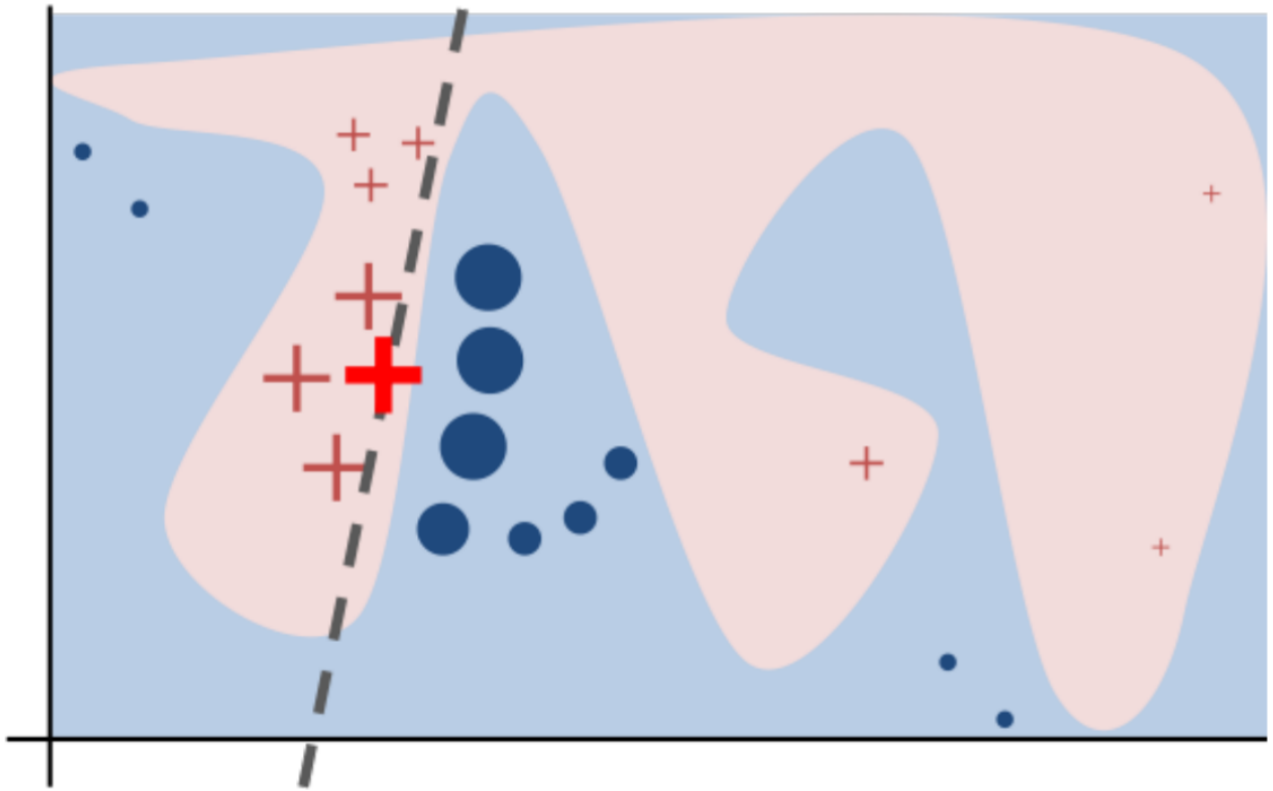


Interpretability Approaches

Stebner
Martensen
22.01.2019

Applying Interpretability Methods in Detail: LIME

1. *Perturbate data*
2. *Compute proximity*
3. *Make predictions*
4. *(Select features)*
5. *Fit a simple model*
6. *Extract explanations (feature weights)*



Applying Interpretability Methods in Detail: LIME

- Select a model family and train the model

Fidelity-Interpretability Trade-off

$\mathcal{L}(f, g, \pi_x)$

Unfaithfulness of the model

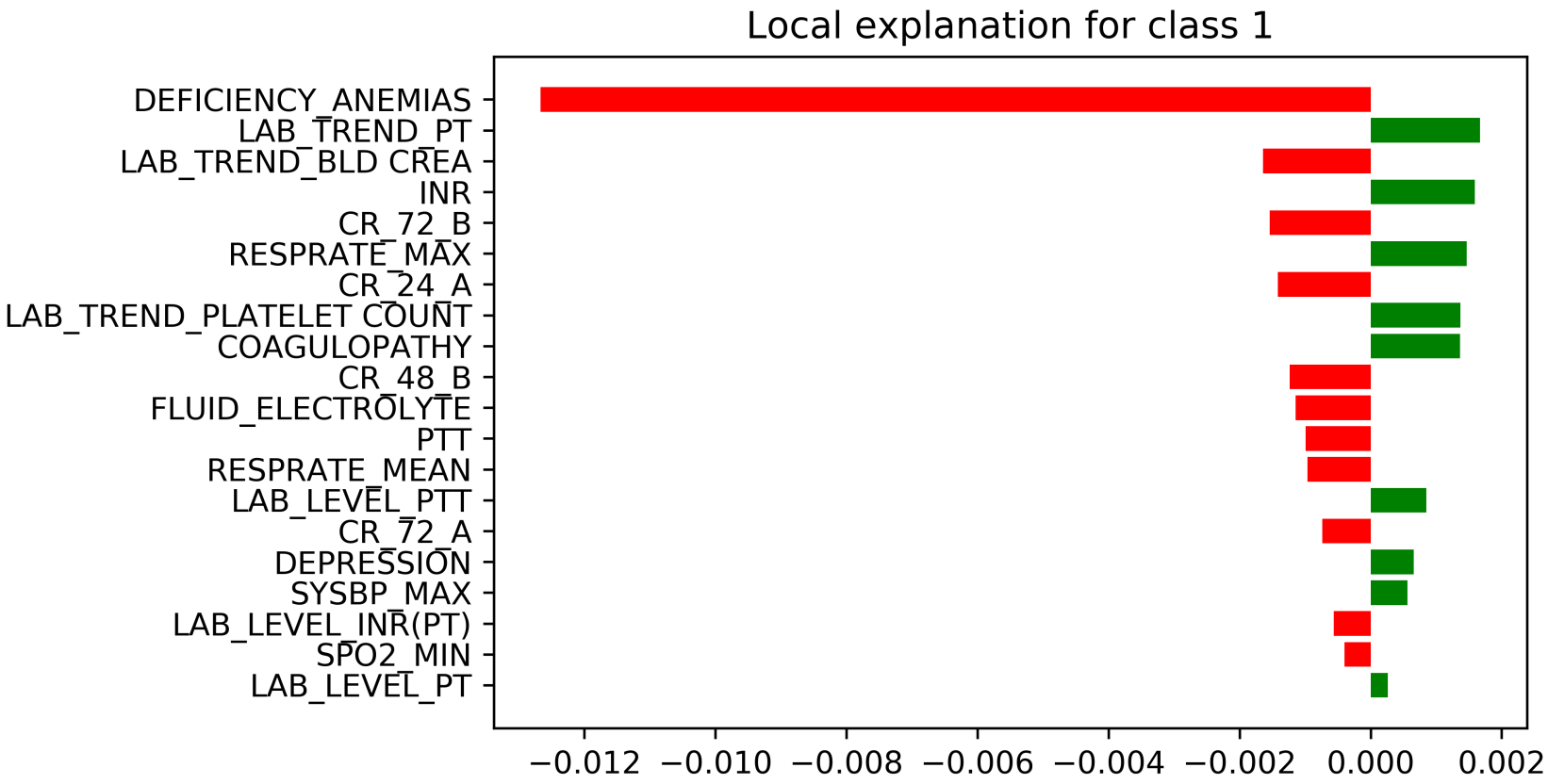
$\Omega(g)$

Complexity of the model

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- Extract explanations (e.g. model weights)

Applying Interpretability Methods in Detail: LIME



Applying Interpretability Methods in Detail: LIME Submodular Pick

Why Submodular Pick?

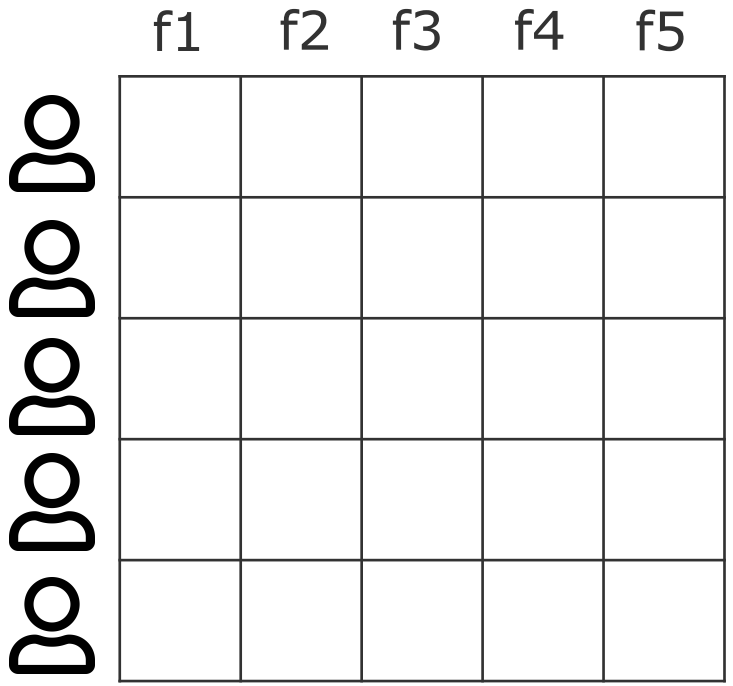
→ LIME is **Local** Interpretable Model Explanations

→ Submodular Pick explains model globally by combining local explanations

Parameters:

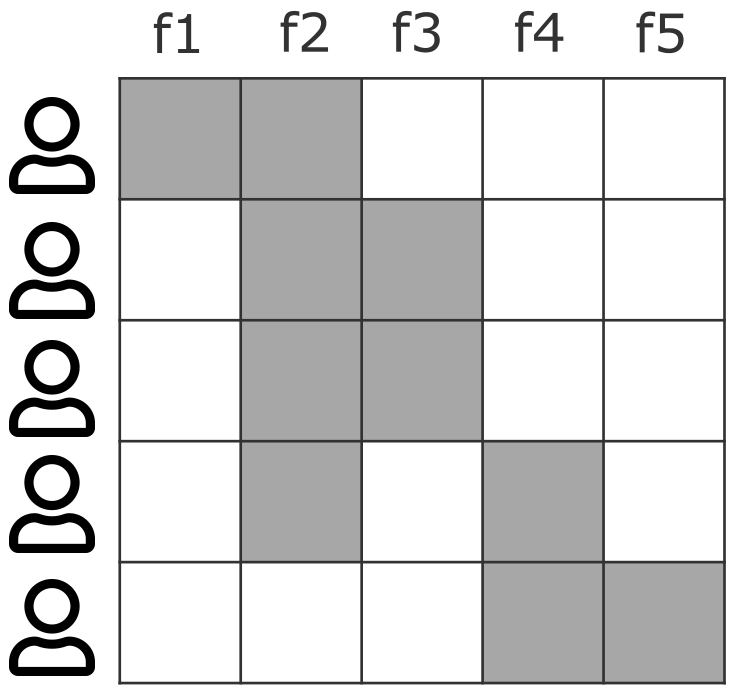
- # instances (10 percent of dataset)
- # explanations (1 percent of dataset)
- # features (complexity value)

Applying Interpretability Methods in Detail: LIME Submodular Pick



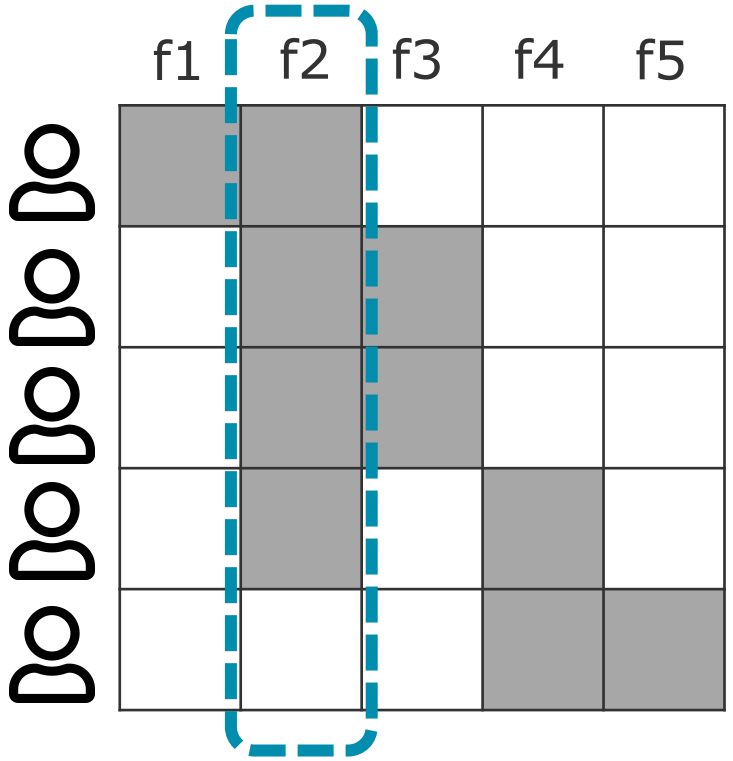
1. Select k instances

Applying Interpretability Methods in Detail: LIME Submodular Pick



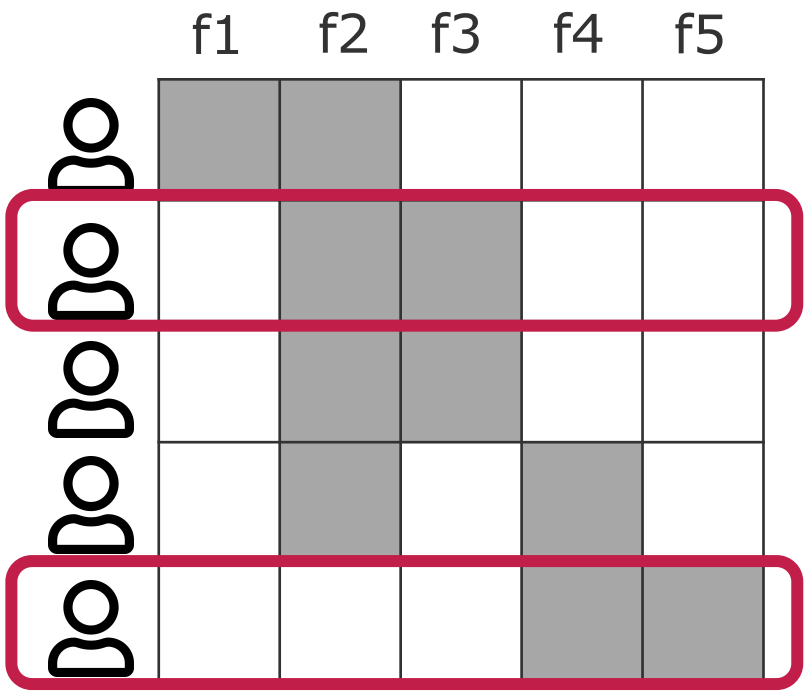
1. Select k instances
2. Get k local explanations and the important features

Applying Interpretability Methods in Detail: LIME Submodular Pick



1. Select k instances
2. Get k local explanations and the important features
3. (f2 has highest importance, because important in 4/5 explanations)

Applying Interpretability Methods in Detail: LIME Submodular Pick



1. Select *k* instances
2. Get *k* local explanations and the important features
3. (f2 has highest importance, because important in 4/5 explanations)
4. Pick *i* explanations with highest coverage

Applying Interpretability Methods in Detail: LIME Submodular Pick

Coverage of an explanation:

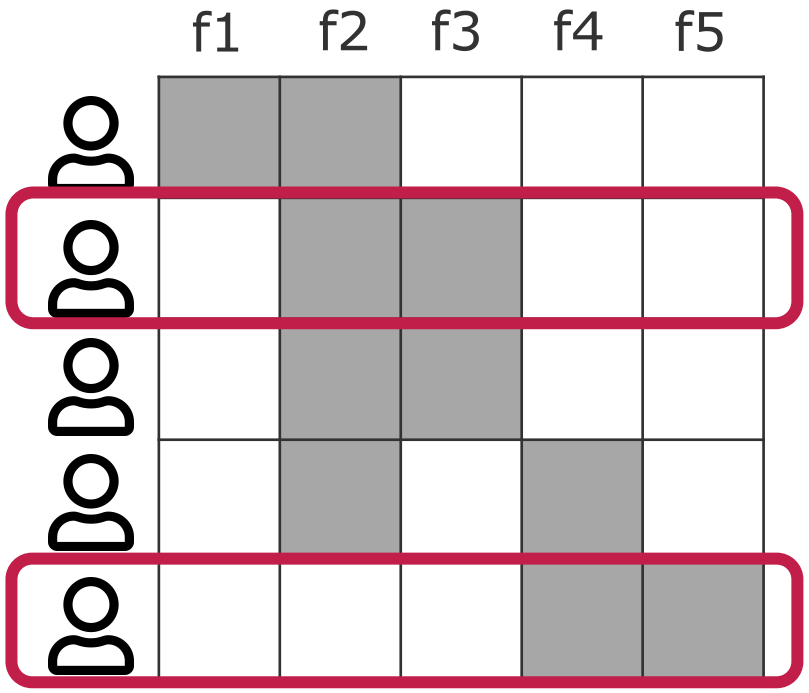
$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

for some set V.

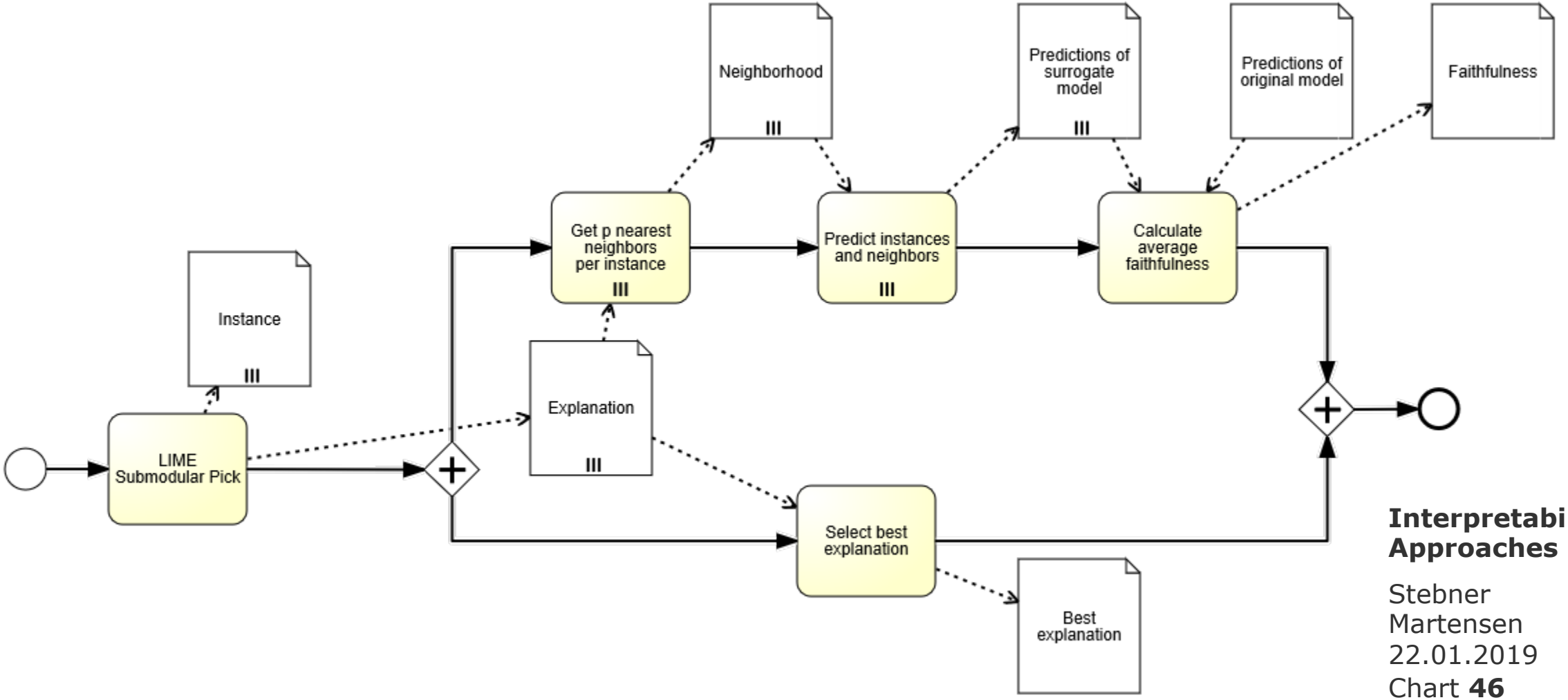
But which V?

Pick B explanations to maximize the coverage:

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}}\, c(V, \mathcal{W}, I)$$



Applying Interpretability Methods in Detail: LIME Submodular Pick - Evaluation



Applying Interpretability Methods in Detail: LIME Submodular Pick

Advantages:

+ Not model dependent, based on data!
+ Includes visualization
+ Local and global approach

Disadvantages:

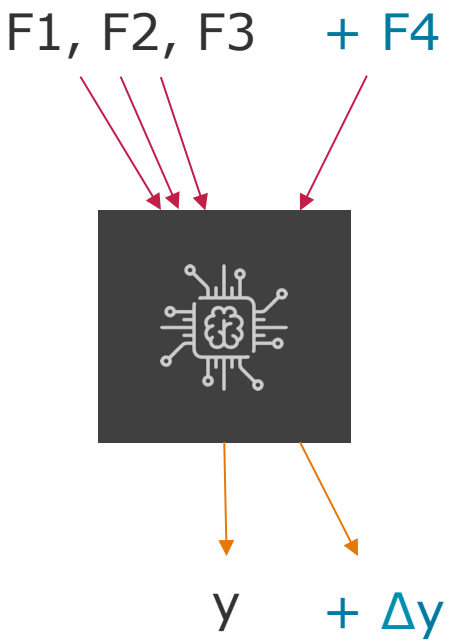
– Requires correct definition of neighborhood
– Submodular pick optimizes coverage, potentially disregards feature interactions
– Instability of model explanations (non-deterministic results)

Methods:
Applying Interpretability Methods in Detail

- Model-based feature importances
- Global Surrogate
- Local Interpretable Model-Agnostic Explanations (LIME)
- **Shapley values**

Applying Interpretability Methods in Detail: Shapley Values

How much did the feature **contribute** to the models prediction?



→ Figure out the **marginal contribution** of F4.

$$\varphi_i(x) = f(x_1, \ldots, x_n) - E[f(x_1, \ldots, X_i, \ldots, x_n)]$$

- Example for a simple linear model:

$$f(x_1, \ldots, x_n) \approx y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\varphi_i(x) = \beta_i x_i - \beta_i E[X_i]$$

→ care!: it's an additive model with no feature interactions

Applying Interpretability Methods in Detail: Shapley Values

$$\varphi_i(x) = \sum_{Q \subseteq S \setminus \{i\}} \frac{|Q|!(|S| - |Q| - 1)!}{|S|!} (\Delta_{Q \cup \{i\}}(x) - \Delta_Q(x)).$$

What? Some difference.

A Number. Maybe a scaling factor?

Summation over subsets?

Applying Interpretability Methods in Detail: Shapley Values

$$\varphi_i(x) = \sum_{Q \subseteq S \setminus \{i\}} \frac{|Q|!(|S| - |Q| - 1)!}{|S|!} (\Delta_{Q \cup \{i\}}(x) - \Delta_Q(x)).$$

F1	F2	F3	F4
X			~
	X		~
		X	~
X	X		~
	X	X	~
X		X	~
X	X	X	~

- *S* is a set of all features
- *Q* a subset of *S* not including *i*

Applying Interpretability Methods in Detail: Shapley Values

$$\varphi_i(x) = \sum_{Q \subseteq S \setminus \{i\}} \frac{|Q|!(|S| - |Q| - 1)!}{|S|!} (\Delta_{Q \cup \{i\}}(x) - \Delta_Q(x)).$$

F1	F2	F3	F4
X	X		~

- *S* is a set of all features
- *Q* a subset of *S* not including *i*

f1	f2	E[F3]	f4	← Feature values with *i*
f1	f2	E[F3]	E[F4]	← Feature values without *i*

$$f_Q(x) = \mathbb{E}[f | X_i = x_i, \forall i \in Q]$$

$$\Delta_Q(x) = f_Q(x) - f_{\{\}}(x)$$

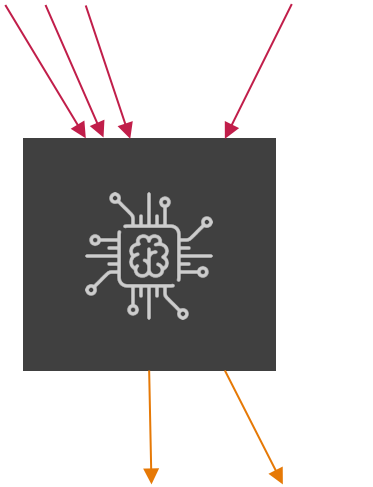
$$f_{\{\}}(x) = \mathbb{E}[f]$$

$$\Delta_{Q \cup \{i\}}(x) - \Delta_Q(x)$$

Applying Interpretability Methods in Detail: Shapley Values

$$\varphi_i(x) = \sum_{Q \subseteq S \setminus \{i\}} \frac{|Q|!(|S| - |Q| - 1)!}{|S|!} (\Delta_{Q \cup \{i\}}(x) - \Delta_Q(x)).$$

F1, F2, F3 + F4 + F6, F5, F7, …



y + Δy

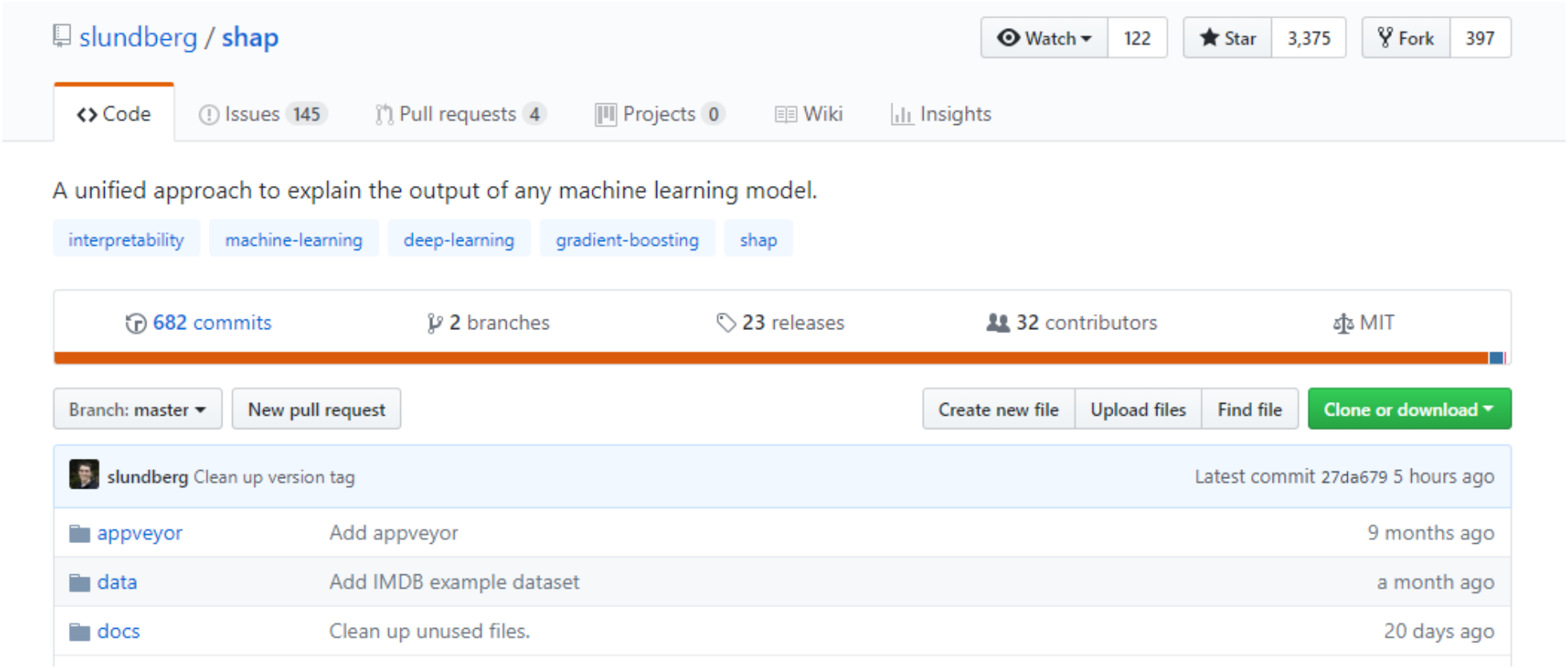
- $|Q|!$-many possible rearrangements
- $(|S| - |Q| - 1)!$-many possibilities to arrange features following *i*

Applying Interpretability Methods in Detail: Shapley Values

Some unique properties:

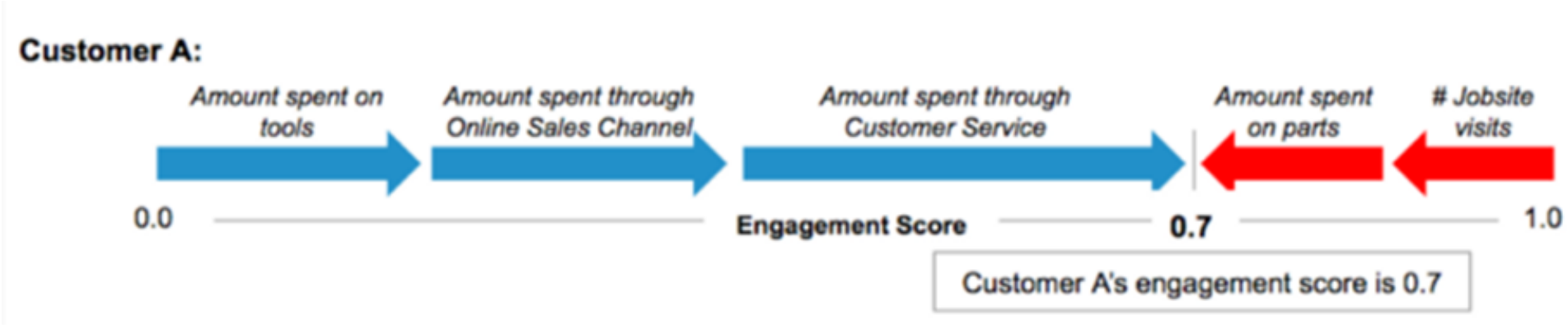
- **Efficiency**
  - Contributions add up to the difference of prediction and expectation
- **Symmetry**
  - Same value for same contributions
- **Dummy Feature**
  - Non-contributing features have value 0
- **Additivity**
  - Multi-model predictions (e.g. random forest) can be analyzed

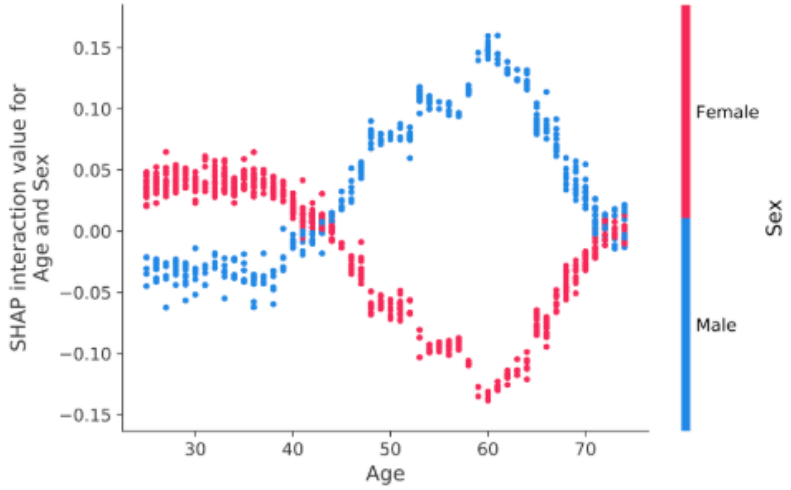
Applying Interpretability Methods in Detail: SHAP



Source: https://medium.com/civis-analytics/demystifying-black-box-models-with-shap-value-analysis-3e20b536fc80

Applying Interpretability Methods in Detail: SHAP





Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **56**

Source: https://medium.com/civis-analytics/demystifying-black-box-models-with-shap-value-analysis-3e20b536fc80

Applying Interpretability Methods in Detail: Shapley values and SHAP

Advantages:

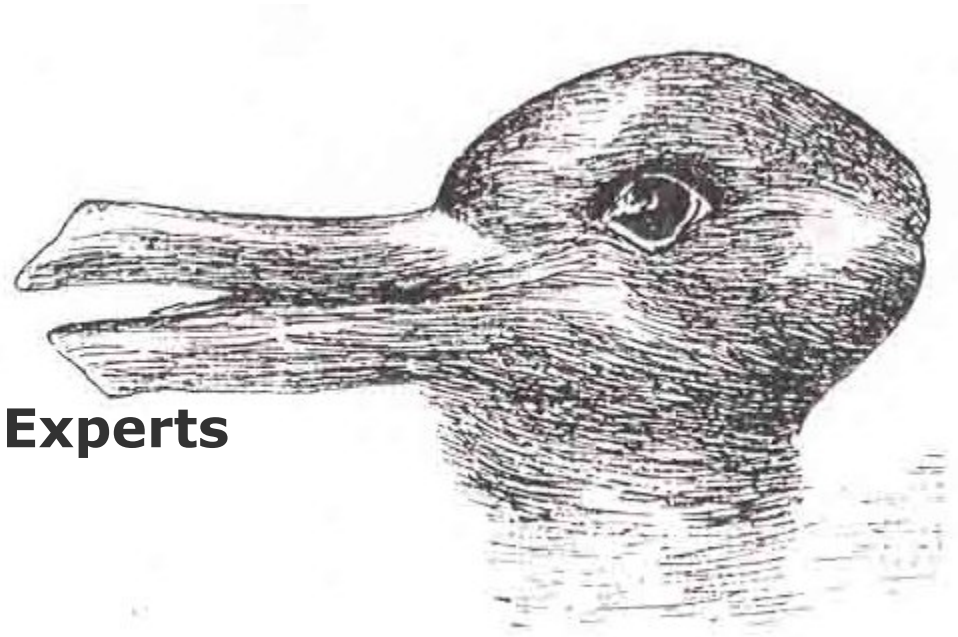
+ Contrastive explanations (with respect to the expectation)
+ Applicable for whole dataset, subset or single instance
+ Solid foundation from game theory

Disadvantages:

– Exponential computational complexity
– Always returns all features
– No prediction model

Methods

- Building a Clinical Prediction Model
- Applying Interpretability Methods in Detail
- **Making Interpretability Available for Domain Experts**



Methods:
Making Interpretability Available for Domain Experts

Requirements:

- Compare different Interpretability Method outputs for one CPM
- Rank interpretability models
- Faithfulness-Complexity tradeoff

Visualizations:

- Feature Importances
- Complexity-Faithfulness-Graph

Results

- Feature Importances
- Complexity-Faithfulness-Graph
- Clinical Hypotheses

Results:
Feature Importances

Feature	Model-based A.	LIME	Linear Surrogate Model	Tree Surrogate Model	SHAP
Age					
Platelets					
Blood Gas					
…					

- Comparing interpretability methods output for every feature
- Filter, sort, threshold, … operations
- (Weighted) average

Results:
Feature Importances

- Feature Importances ordered by maximal importance

Feature	Model-based importances	LIME	Linear Surrogate Model	Tree Surrogate Model	SHAP
Lab Flag PT			0.4024		
Lab Flag INR(PT)			-0.3983		
Deficiency Anemias			0.2606		
AIDS			-0.2547		
Lab Level Hematocrit (Calculated)			0.2515		
GFR_72	0.1618	0.0607		0.2127	0.1618
Lab Flag Bilirubin	0.0678			0.1610	
CR_72	0.0440	0.0397		0.1530	0.0440
Lactate	0.4508			0.1050	0.4508

Results:
Feature Importances

- Feature Importances ordered by occurrences (if occurred more than once)

Feature	Model-based importances	LIME	Linear Surrogate Model	Tree Surrogate Model	SHAP
GFR_72	0.1618	0.0607		0.2127	0.1618
CR_72	0.0440	0.0397		0.1529	0.0440
Lactate	0.0451			0.1050	0.0451
Lab Flag Bilirubin	0.0678			0.1610	
Bicarbonate	0.0295			0.0192	

Results:
Complexity-Faithfulness-Graph

Complexity – Faithfulness – Tradeoff:

Complexity ~ Faithfulness

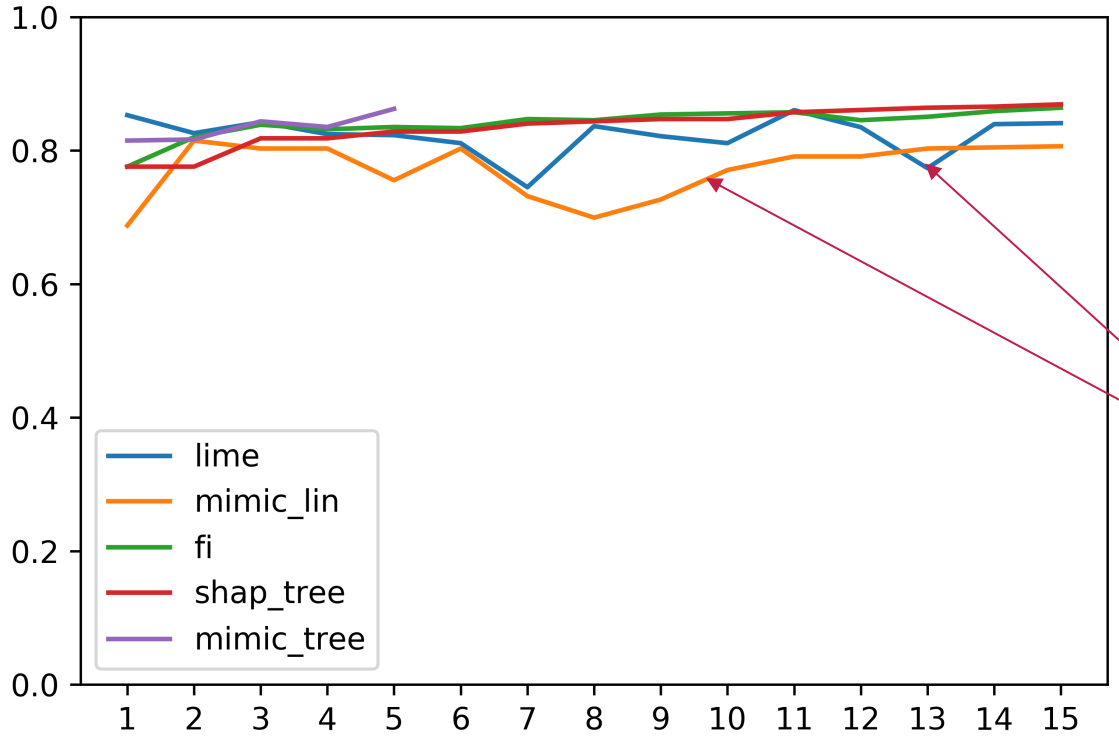
Complexity ~ 1 / Interpretability

Increased complexity -> increase in faithfulness
Increased complexity -> decrease in interpretability

Inspired by: Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. Leilani Gilpin et. al., Massachusetts Institute of Technology, 2018

Results:
Complexity-Faithfulness-Graph

Complexity – Faithfulness – Tradeoff:



Why does this not increase monotonically?

→ Maybe showing incompetence of linear models for complex relations

Results:
Tentative Clinical Hypotheses

Glomelural Filtration Rate 72h before procedure:

- Flow rate of filtered fluid through the kidney
- Known as indicator of kidney function

Creatinine Clearance Rate 72h before procedure:

- Volume of blood plasma cleared of creatinine per unit time
- AKI is defined as increase of CR over baseline

Do these patients have higher chances of survival/recovery because their AKI is detected earlier?

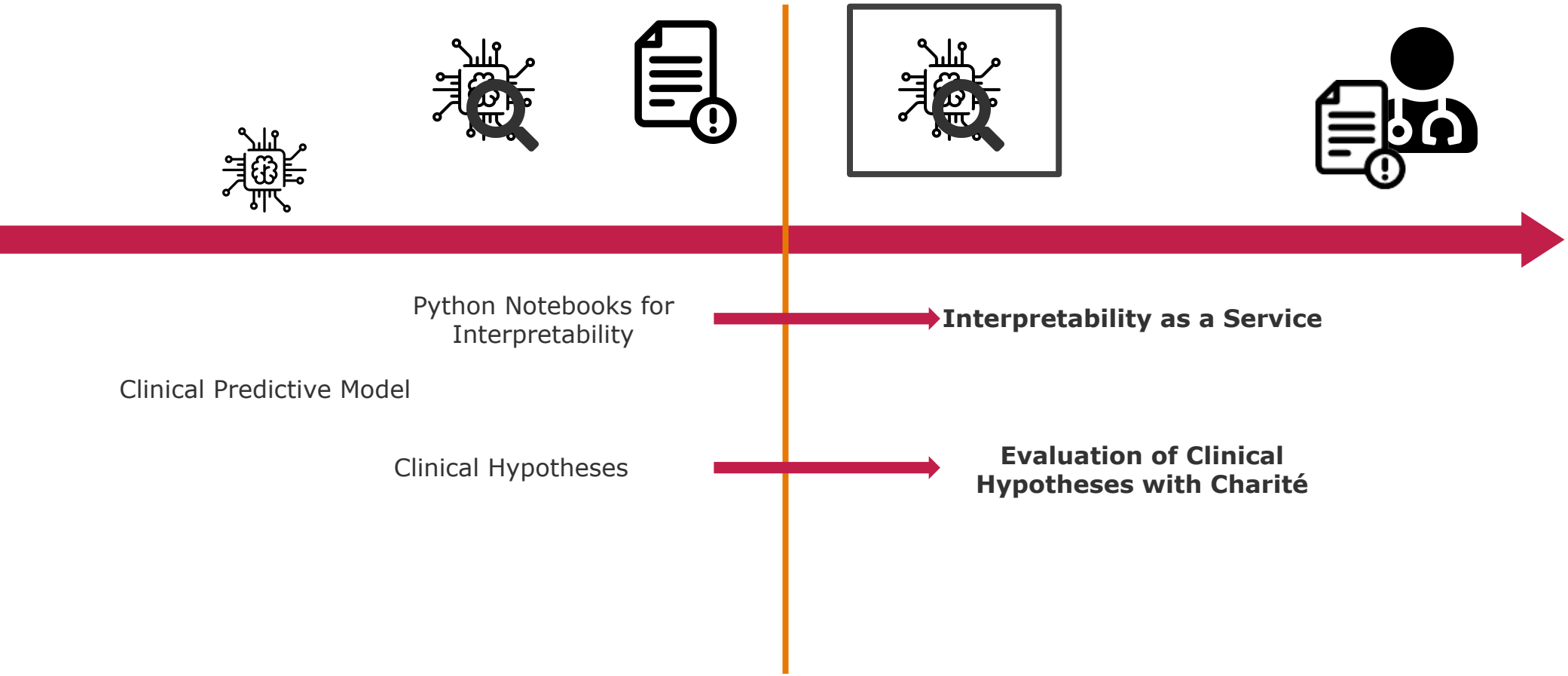
Bilirubin:

- Product of breakdown of red blood cells

Platelets / Thrombocytes:

- First responders to sites of damages in the body

Outlook:
Next Steps



Interpretability Approaches

Stebner
Martensen
22.01.2019
Chart **68**

Visions & Objectives: Contribution

VISION 1

Find and validate medical hypotheses regarding mortality and recovery of AKI

- Train CPM
- Predict patient outcomes
- Gather interpretations
- Derive and evaluate clinical hypotheses

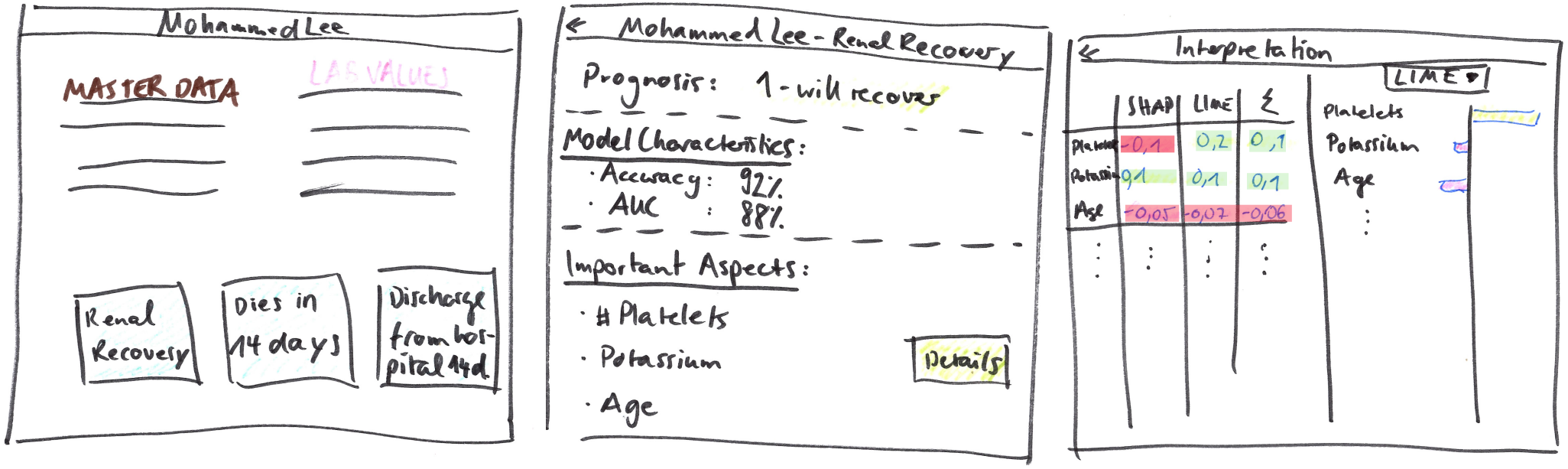
VISION 2

Make interpretations of CPMs available to physicians

- Interpret any CPM
- Make interpretations comparable side-by-side
- Show complexity-faithfulness tradeoff

Outlook: Future Work

- Evaluate Approach with different cohorts (Heidelberg database, different disease)
- Patient Predictor and Diagnosis Explainer (UI)



Sources

Duck-Rabbit-Illusion: https://en.wikipedia.org/wiki/Ambiguous_image#/media/File:Duck-Rabbit_illusion.jpg

Sherlock: https://images.fineartamerica.com/images-medium-large-5/sherlock-holmes-c1905-granger.jpg

Cardiopulmonary Bypass: https://upload.wikimedia.org/wikipedia/commons/thumb/2/24/Blausen_0468_Heart-Lung_Machine.png/300px-Blausen_0468_Heart-Lung_Machine.png

Injured Kidney: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQ4kVzdKHZ81KazmyE9YXLQvvqp9iF00PI56PfPI0MOV_Fxorw1aA

Error Plane: https://image.slidesharecdn.com/navdeepmlinov0117-171102184007/95/ideas-on-machine-learning-interpretability-9-638.jpg?cb=1509648095

Icons by Fontawesome (https://fontawesome.com/license) and by Freepik, Appzgear, Pixel perfect , phatplus & Eucalyp on https://flaticon.com

LIME: https://www.slideshare.net/0xdata/interpretable-machine-learning-using-lime-framework-kasia-kulma-phd-data-scientist

Feature Importances from sci-kit learn: https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/tree/tree.py

LIME Paper: Ribeiro et al. “Why Should I Trust You?” Explaining the Predictions of Any Classifier (ACL Proceedings 2016)

Interpretable Method (Dis-)Advantages: Molnar, C. (2018). Interpretable Machine Learning. Retrieved from https://christophm.github.io/interpretable-ml-book/

Evaluating Interpretability: Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. Leilani Gilpin et. al., Massachusetts Institute of Technology, 2018

Interpretability Approaches applied to Clinical Predictive Modeling

Trends in Bioinformatics

Intermediate Presentation

Discussion

Possible Questions:

- How should we normalize the importances, so that they are actually comparable?
- As a patient, in how much level of detail would you expect your doctor to explain Machine Learning results?
- As a physician, how do you want to be trained for interpretable models?