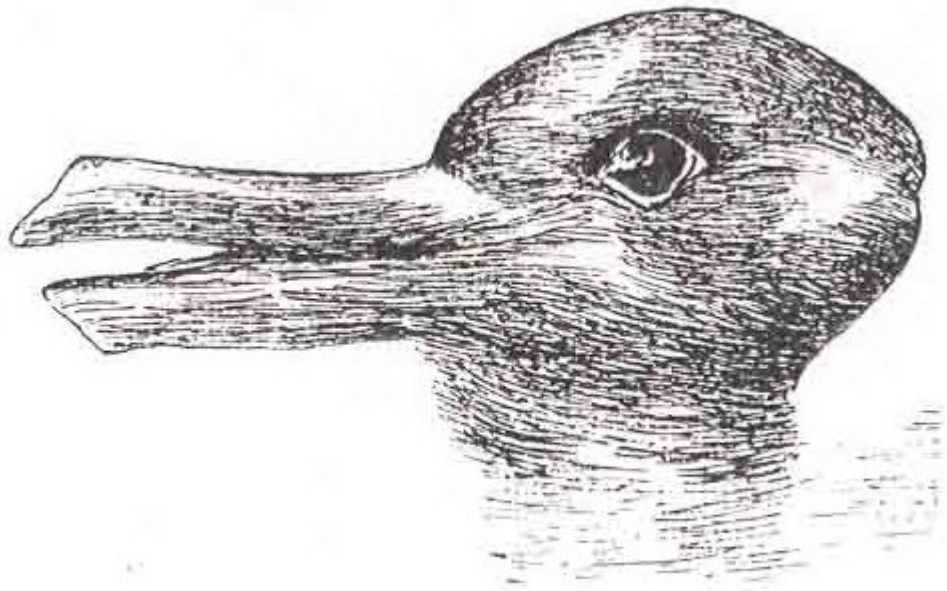
Interpretability Approaches
applied to Predictive Models in Clinical Healthcare

Trends in Bioinformatics

Intermediate Presentation

Tom Martensen, Axel Stebner



Interpretability is the degree to which a human can understand the cause of a decision.

- Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” arXiv Preprint arXiv:1706.07269.

1. **Technique**
2. **Use Case**
3. **Problem Statement**
4. **Envisioned Solution**
5. **Methods: LIME, Decision Rules**
6. **Contribution**
7. **Desired Outcome**

Interpretability is the degree to which a human can understand the cause of a decision.

- Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” arXiv Preprint arXiv:1706.07269.

Technique:
Interpretation of Machine Learning Models

Machine Learning

WHAT

Technique:
Interpretation of Machine Learning Models

Machine Learning

WHAT

Interpretability

WHY

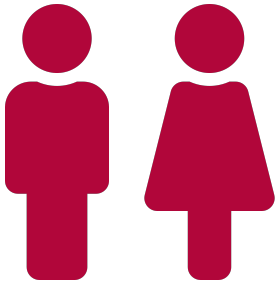
Technique:
Interpretation of Machine Learning Models

Machine Learning
WHAT

Interpretability
WHY

Social Acceptance

Safety and Testing



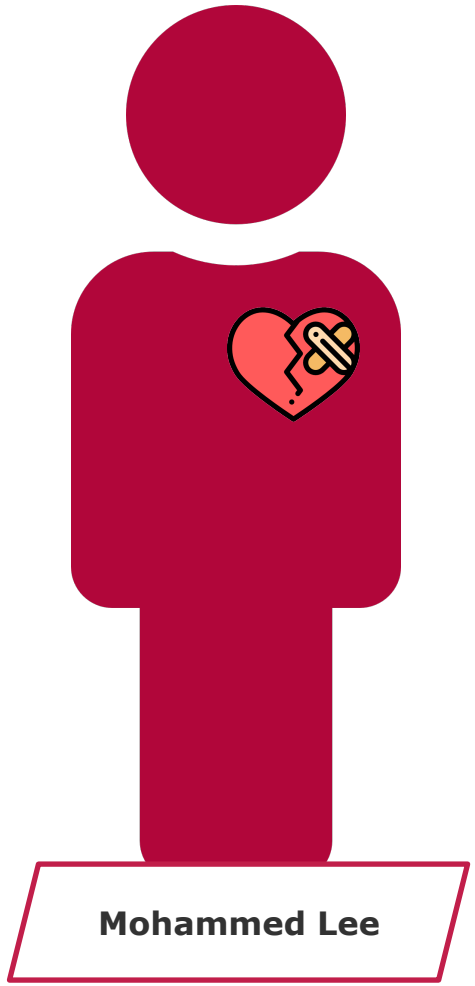
Bias Detection

Auditing

Use Case:
Therapy of Acute Kidney Injury

UniversitätsKlinikum Heidelberg

HPI Hasso Plattner Institut



Cardiopulmonary Bypass

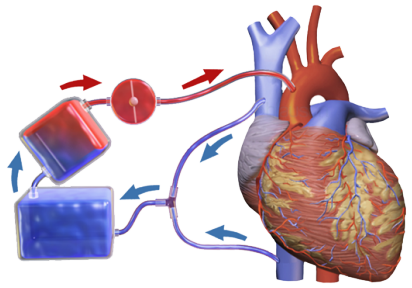
Decreased Blood Flow

Decreased Kidney Perfusion

Acute Kidney Injury (AKI)

Intermittent Hemodialysis (IHD)

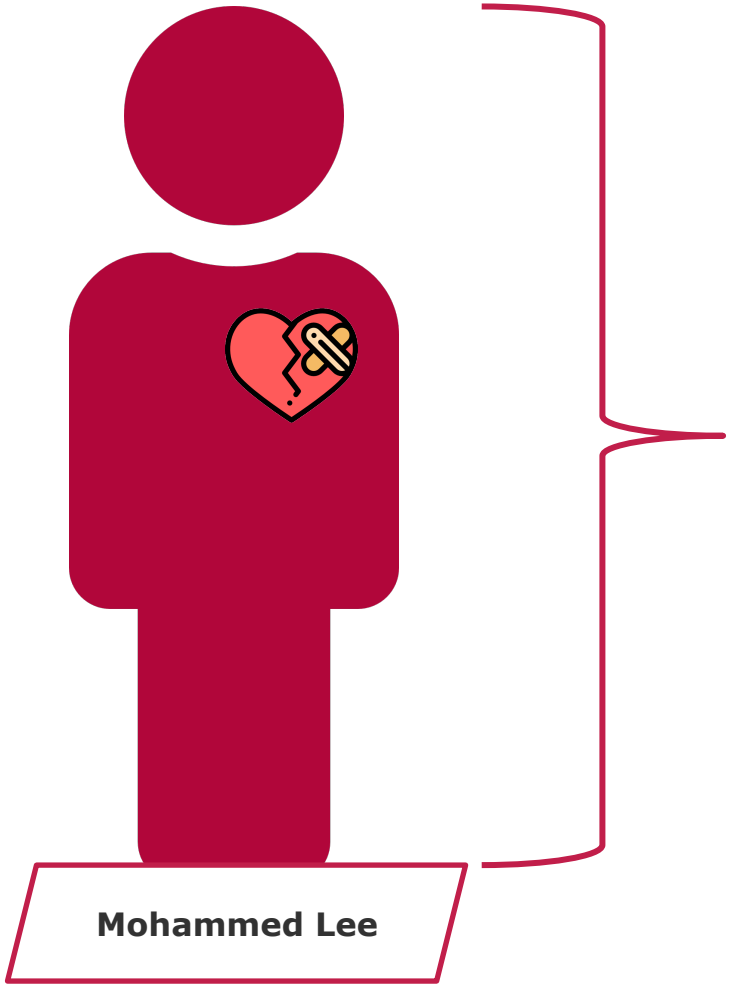
Continuous Renal Replacement Therapy (CRRT)



Use Case:
Therapy of Acute Kidney Injury

UniversitätsKlinikum Heidelberg

HPI Hasso Plattner Institut



98% Accuracy!

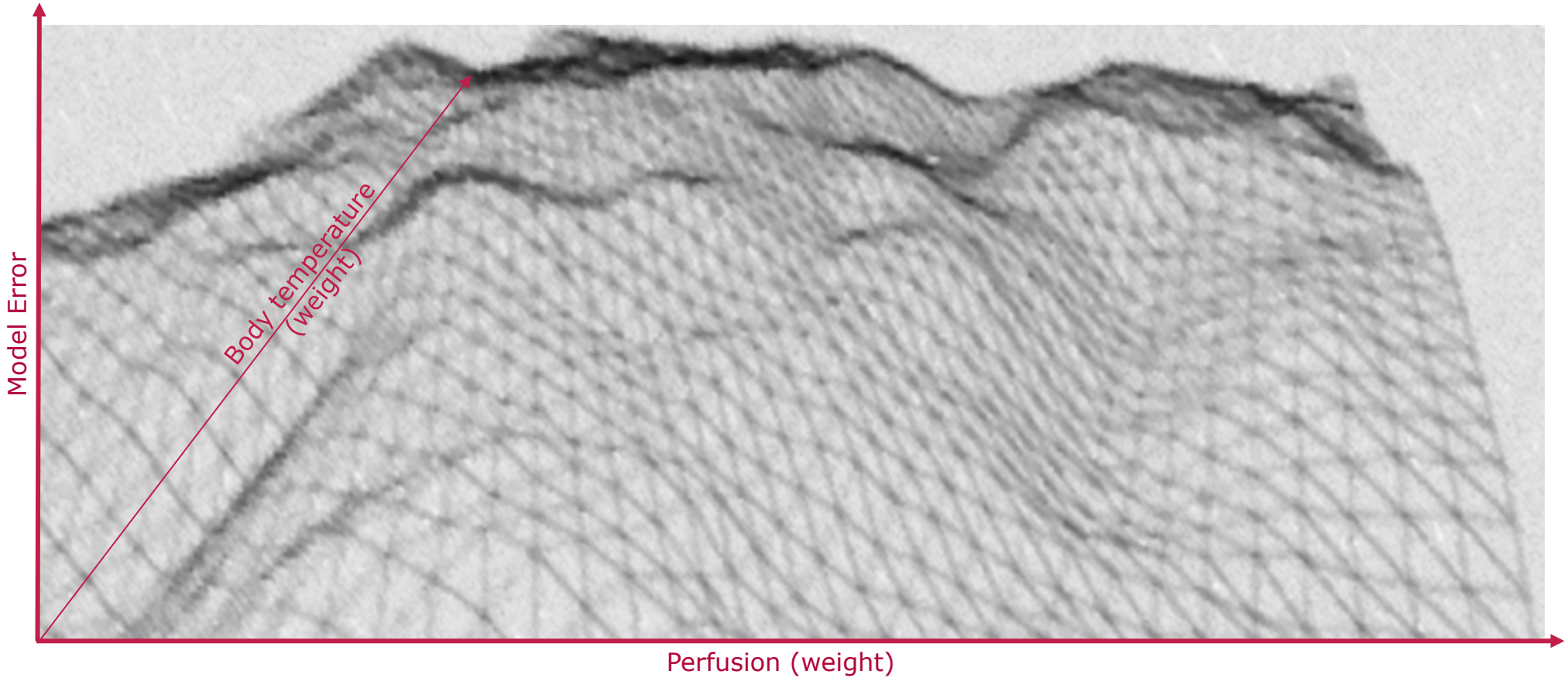
89% Confidence!

Continuous Renal Replacement Therapy

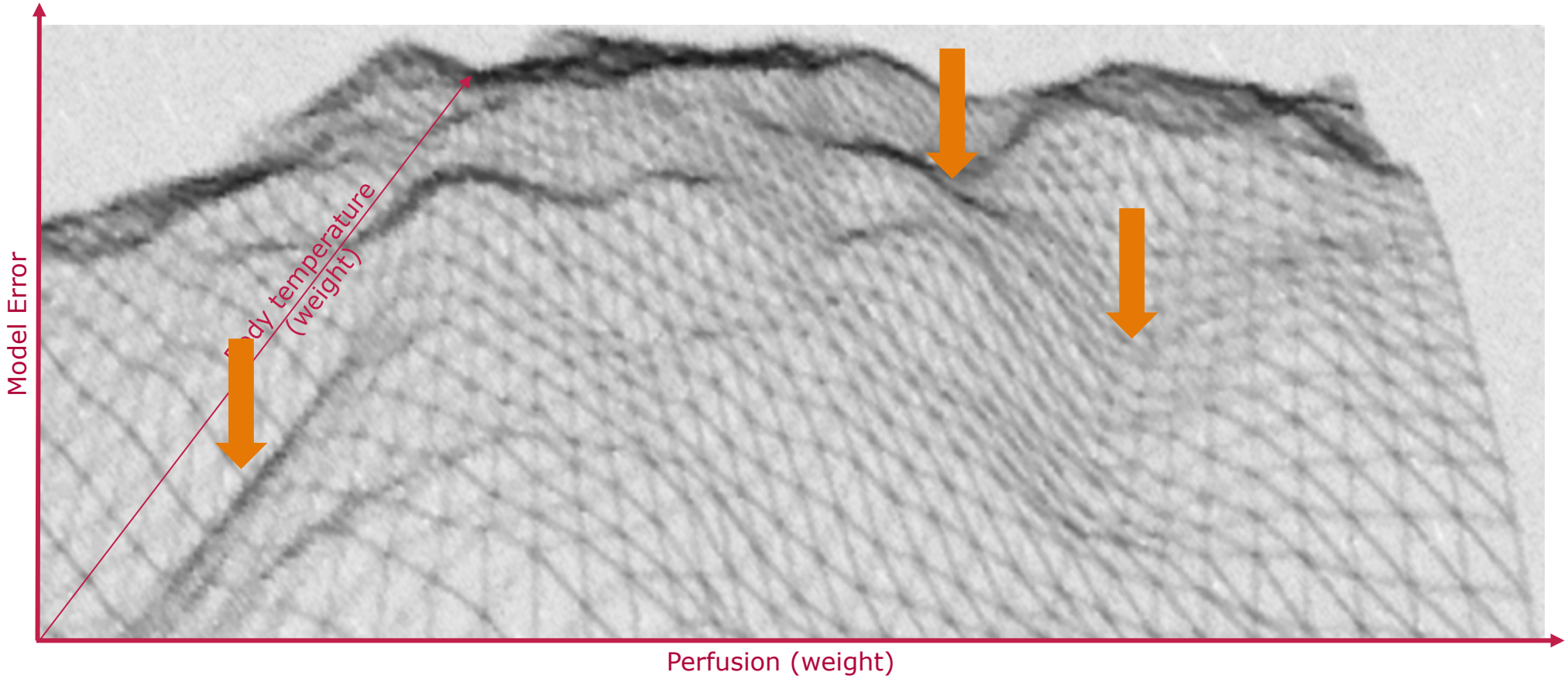
Interpretability Approaches

Stebner
Martensen
27.11.2018
Chart **8**

Complication: Multiplicity of Good Models



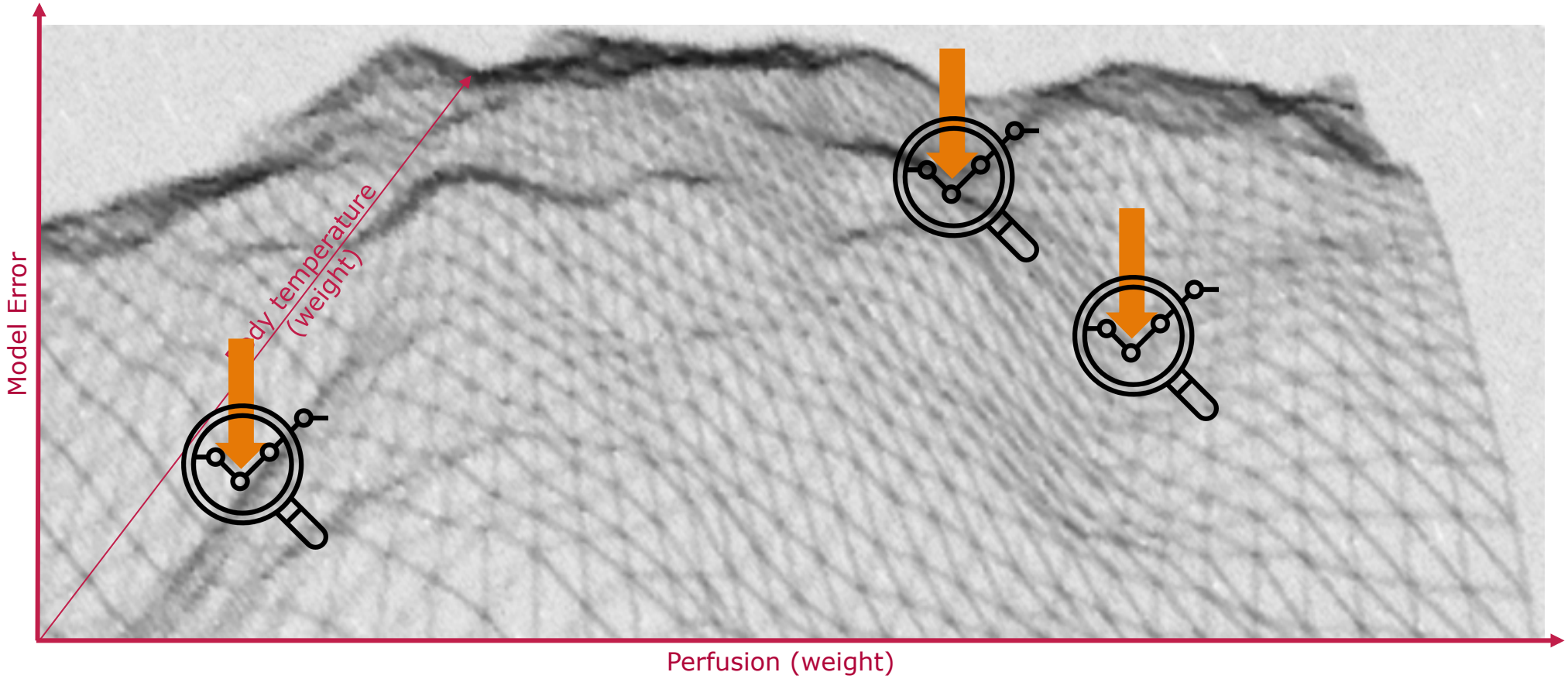
Complication: Multiplicity of Good Models



Which of these weightings should be chosen?

Which explanation is useful and valuable?

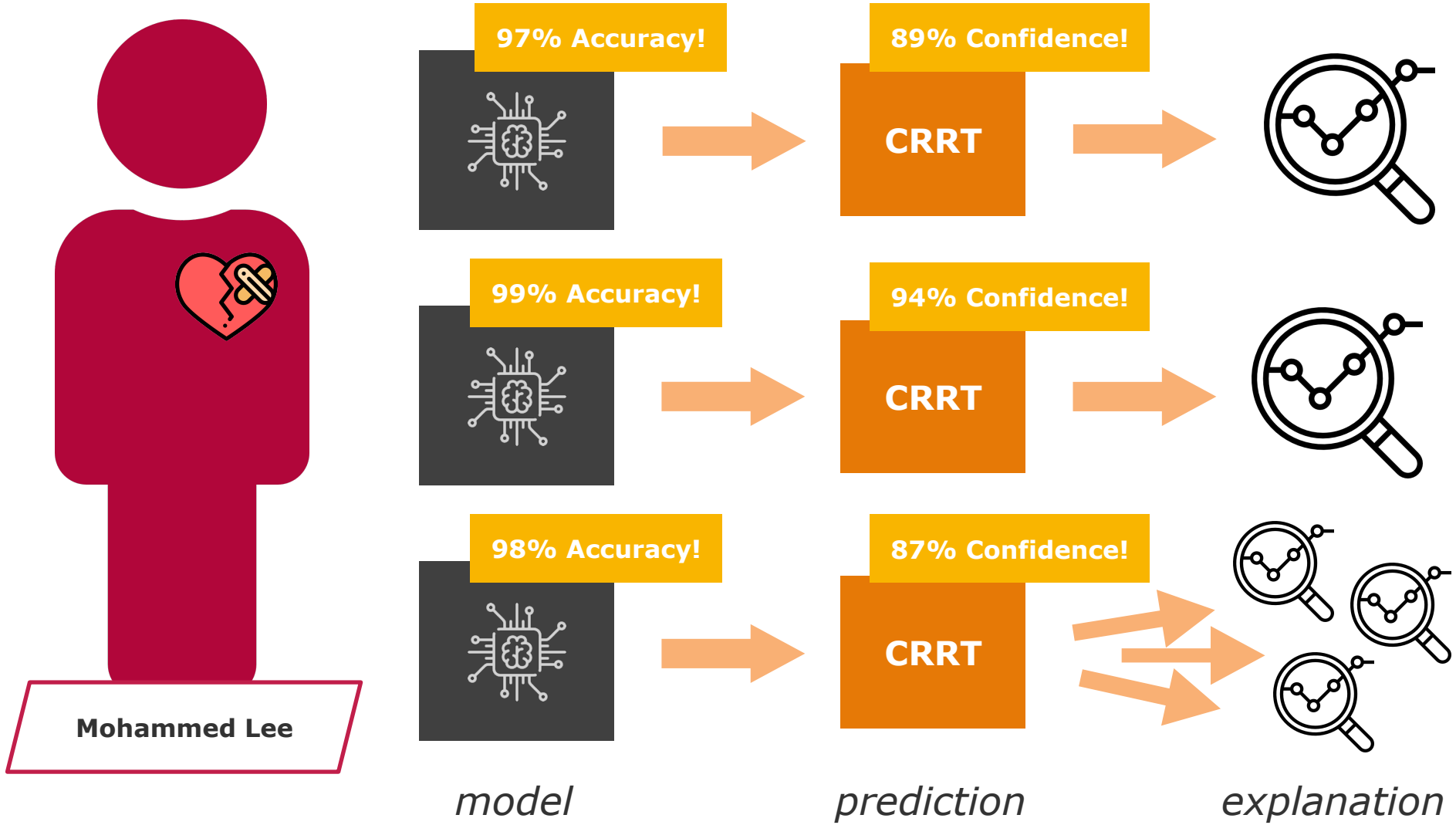
Complication: Multiplicity of Good Models



Which of these weightings should be chosen?

Which explanation is useful and valuable?

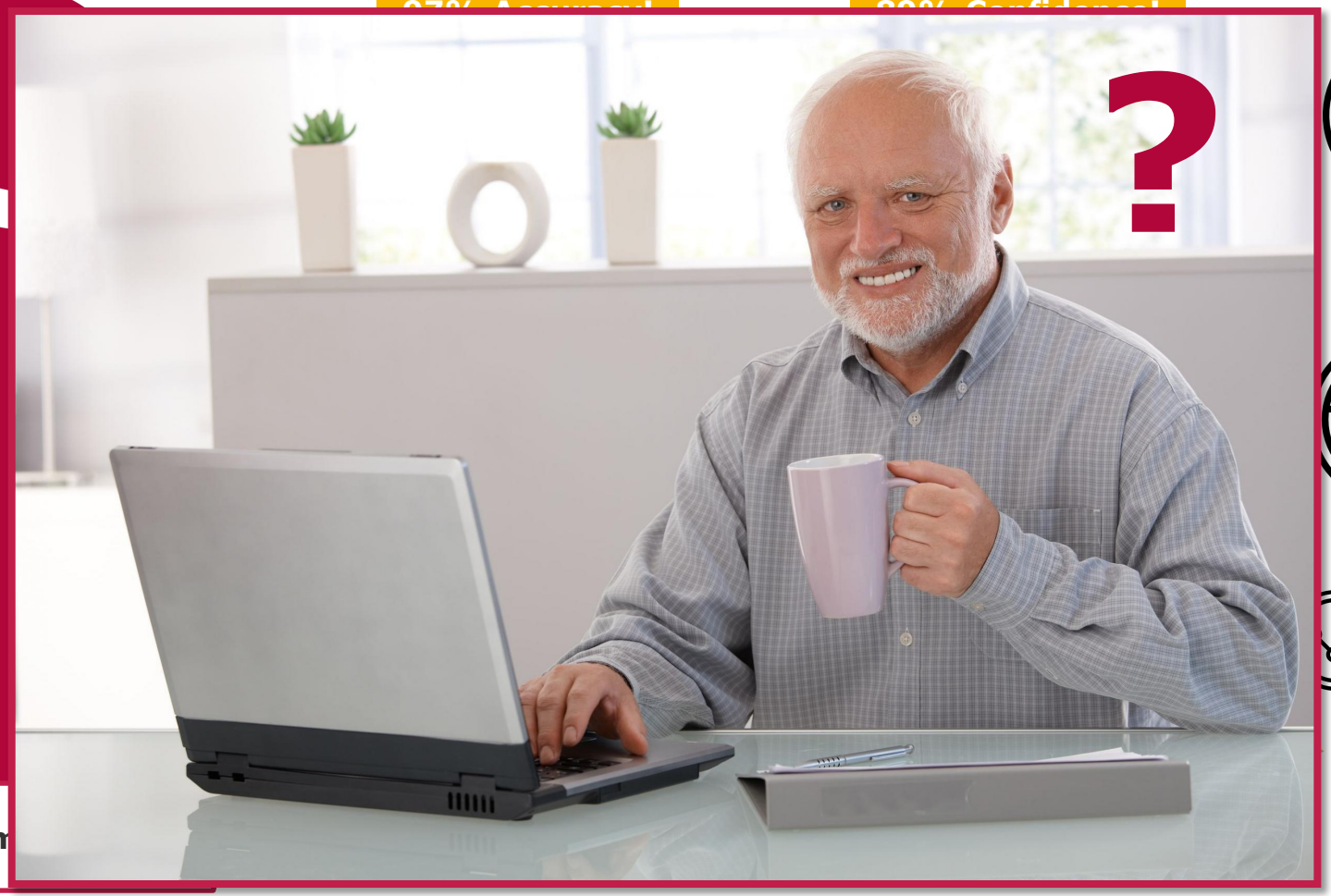
Use Case:
Therapy of Acute Kidney Injury



Interpretability Approaches

Stebner
Martensen
27.11.2018
Chart **12**

Use Case:
Therapy of Acute Kidney Injury

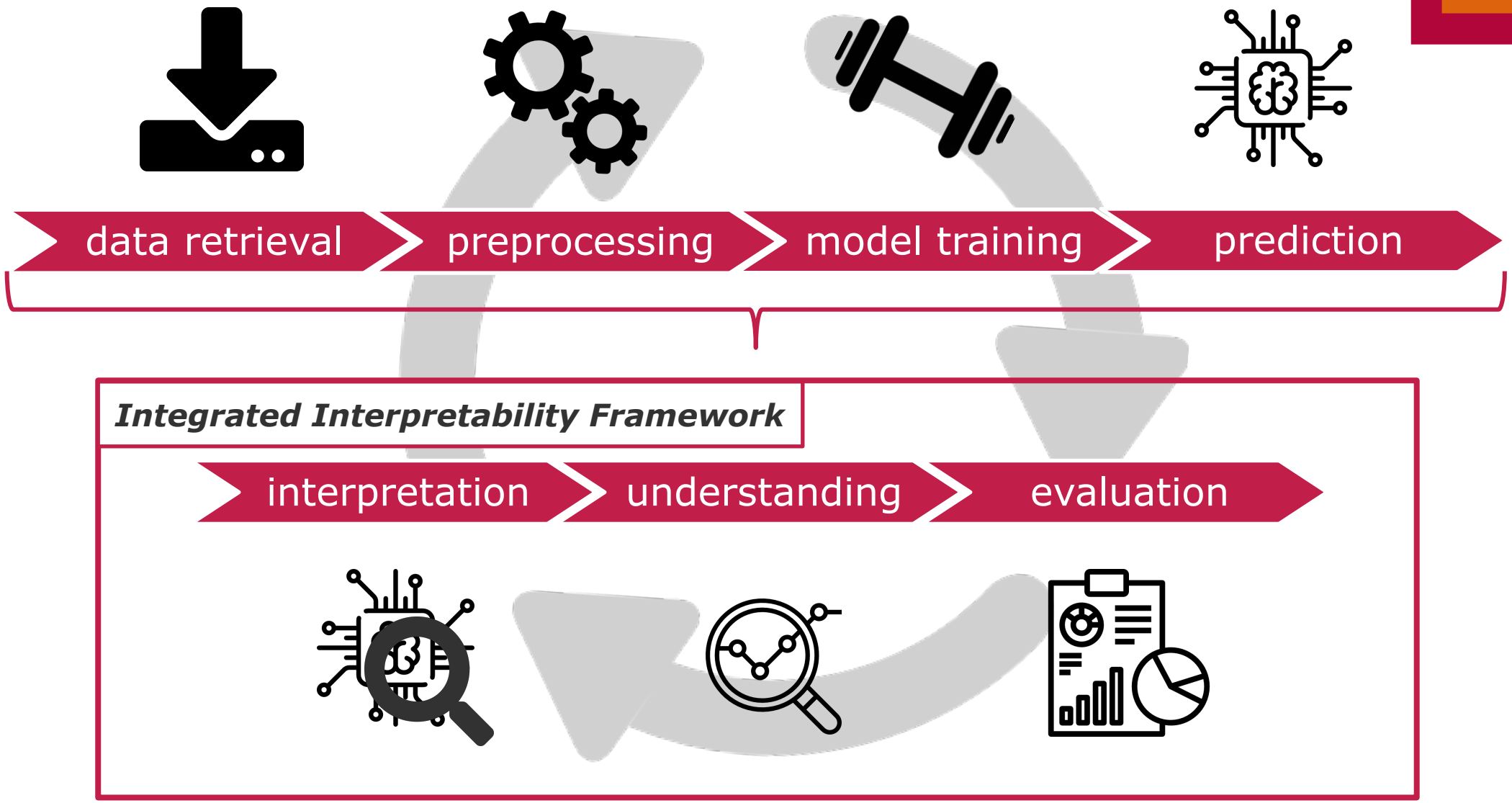


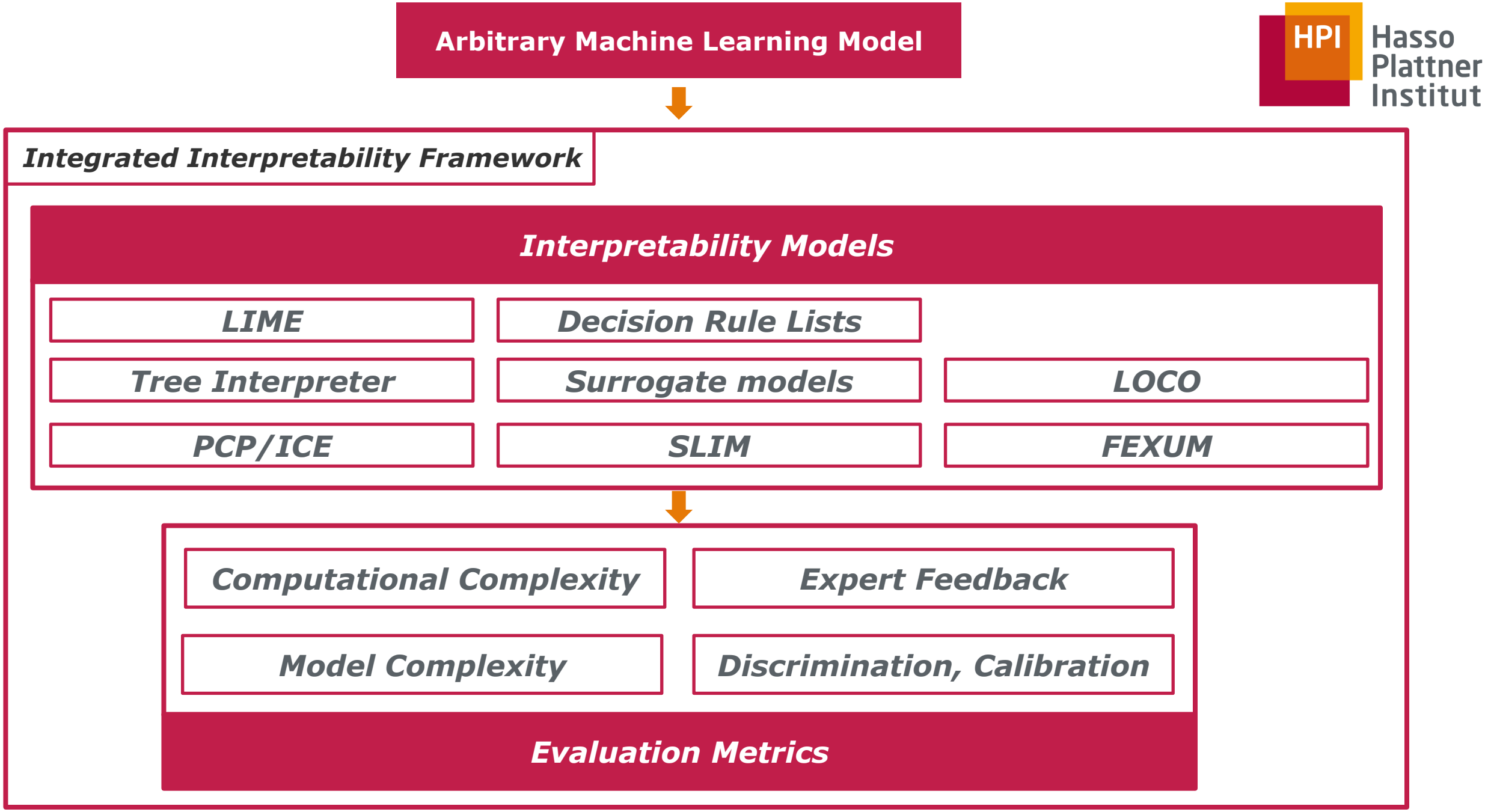
model *prediction* *explanation*

Interpretability Approaches

Stebner
Martensen
27.11.2018
Chart **13**

Integrated Interpretability Framework


data retrieval
preprocessing
model training
prediction
Integrated Interpretability Framework
interpretation
understanding
evaluation


Arbitrary Machine Learning Model
Integrated Interpretability Framework
Interpretability Models
LIME
Decision Rule Lists
Tree Interpreter
Surrogate models
LOCO
PCP/ICE
SLIM
FEXUM
Computational Complexity
Expert Feedback
Model Complexity
Discrimination, Calibration
Evaluation Metrics

Interpretability Models: Dimensions of Interpretability

INTRINSIC MODELS

vs

POST-HOC MODELS

METHOD OUTCOME

latent representations

feature ranking

surrogate model

MODEL-SPECIFIC

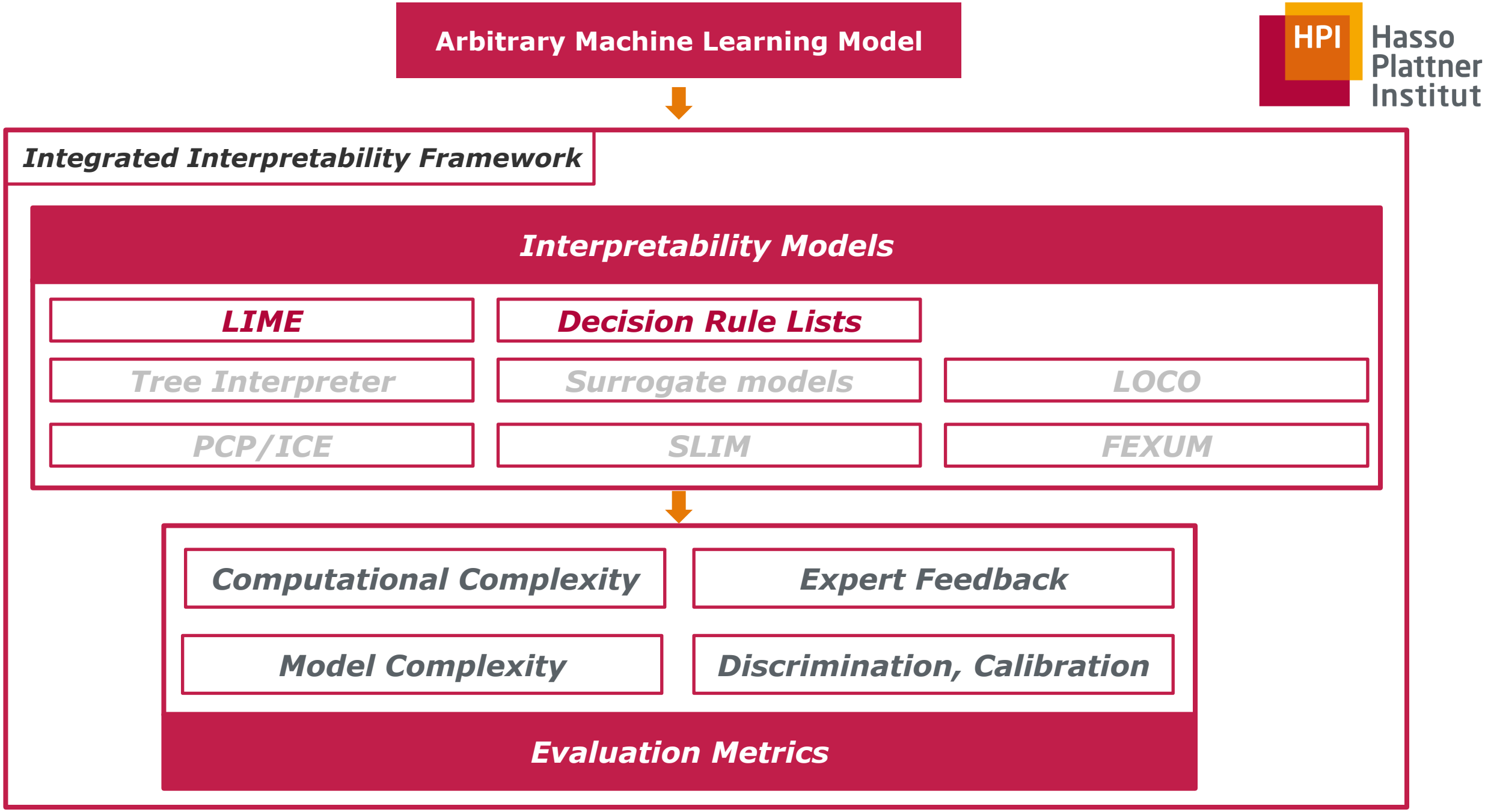
vs

MODEL-AGNOSTIC

LOCAL SCOPE

vs

GLOBAL SCOPE


Arbitrary Machine Learning Model
Integrated Interpretability Framework
Interpretability Models
LIME
Decision Rule Lists
Tree Interpreter
Surrogate models
LOCO
PCP/ICE
SLIM
FEXUM
Computational Complexity
Expert Feedback
Model Complexity
Discrimination, Calibration
Evaluation Metrics

Local Approach LIME: Dimensions of Interpretability

INTRINSIC MODELS

vs

POST-HOC MODELS

METHOD OUTCOME

latent representations

feature ranking

surrogate model

MODEL-SPECIFIC

vs

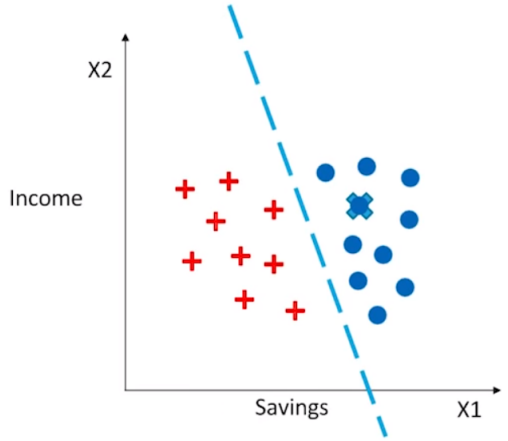
MODEL-AGNOSTIC

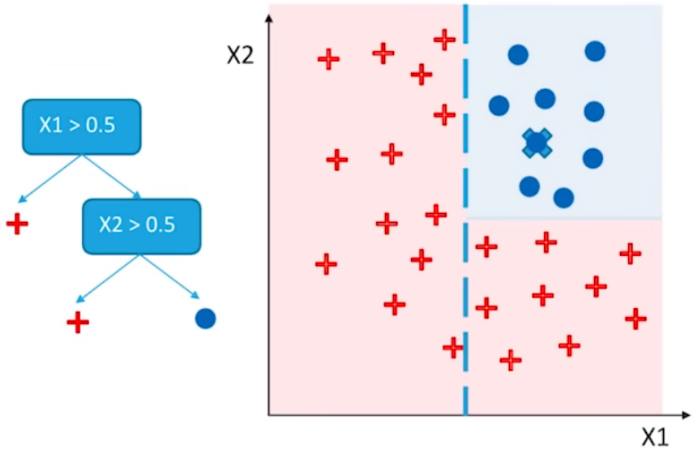
LOCAL SCOPE

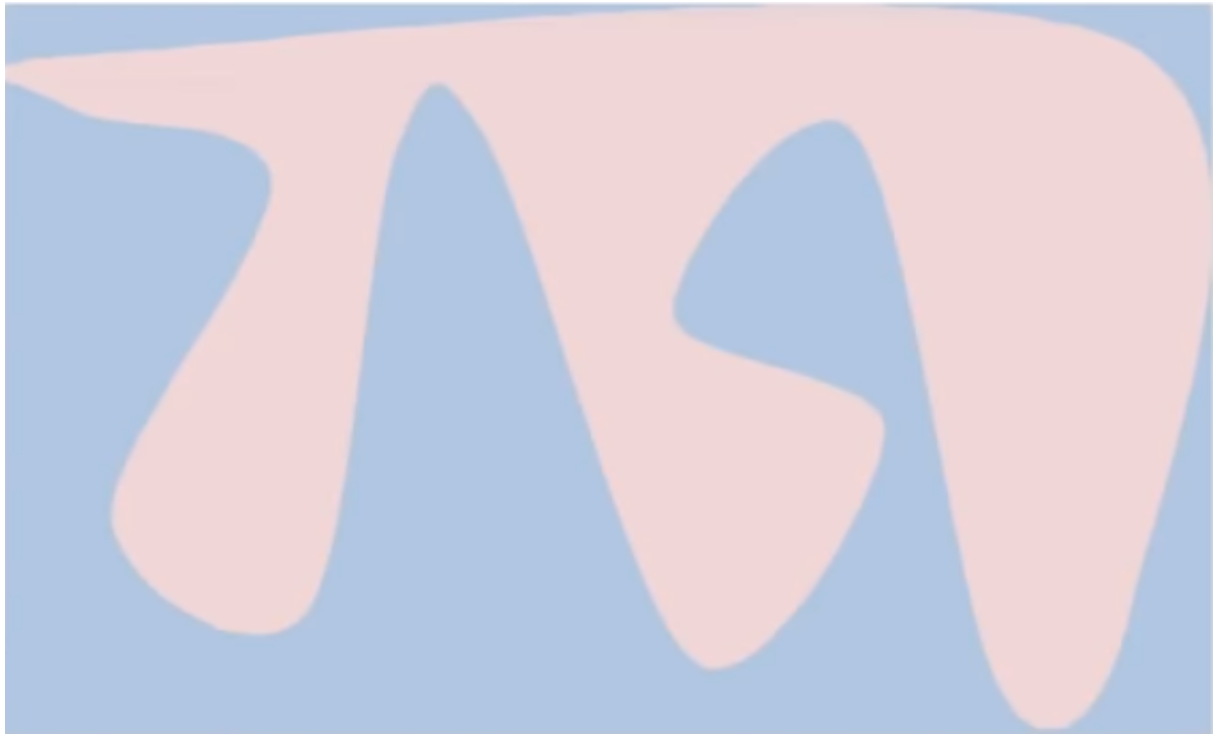
vs

GLOBAL SCOPE

Local Approach LIME: What does it do?

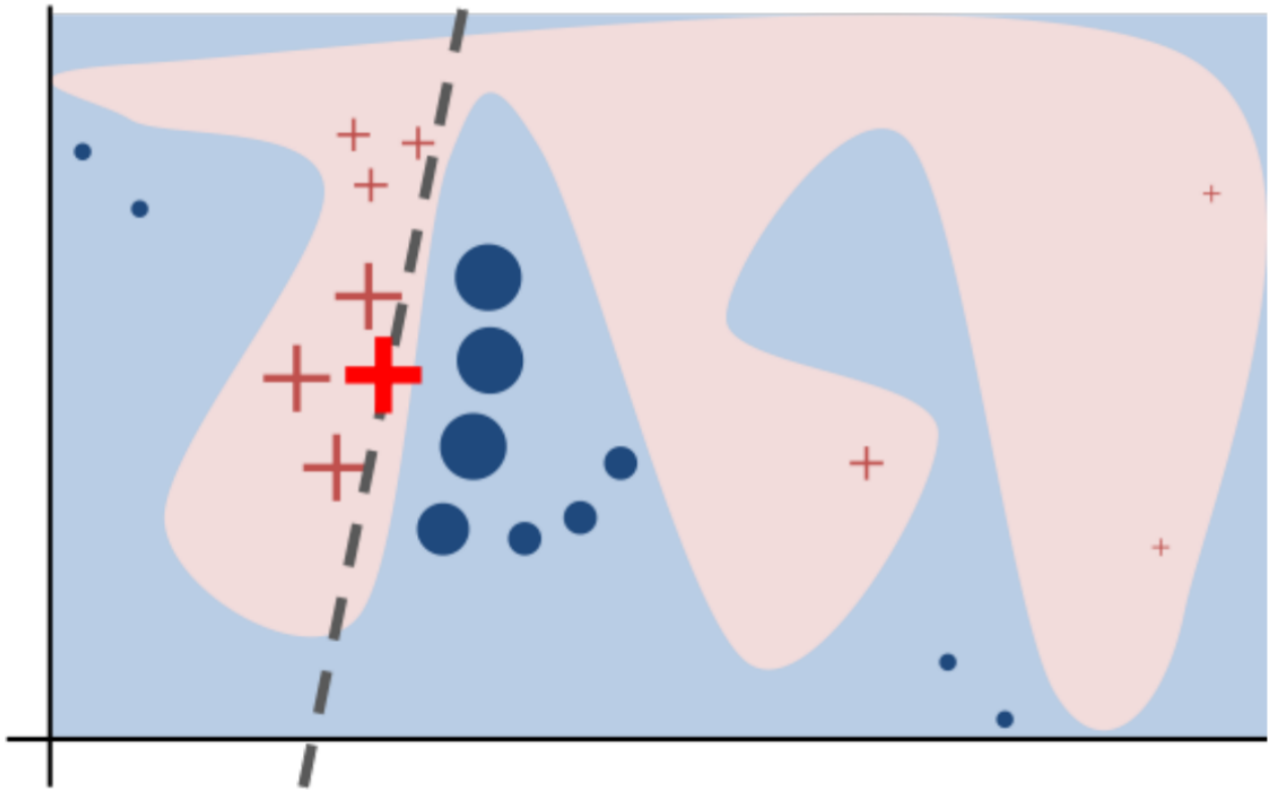






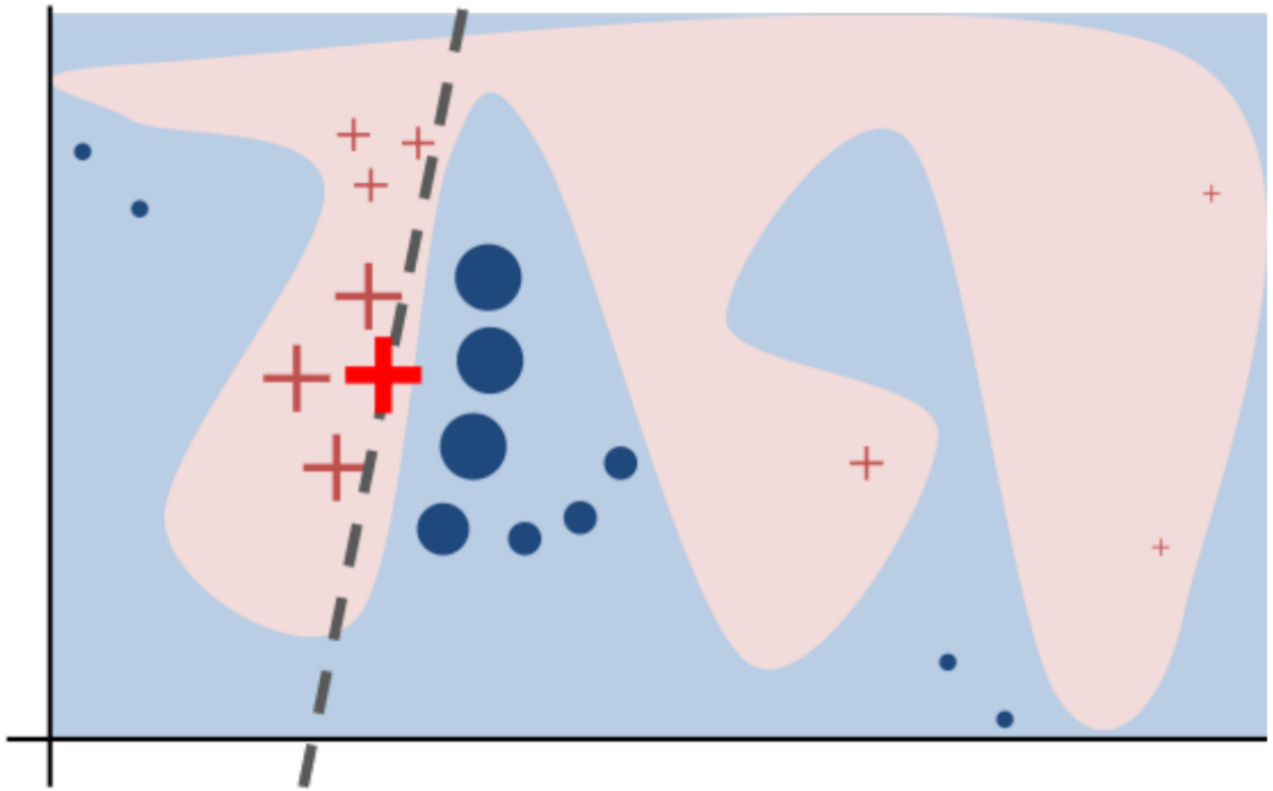
Local Approach LIME: What does it do?

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



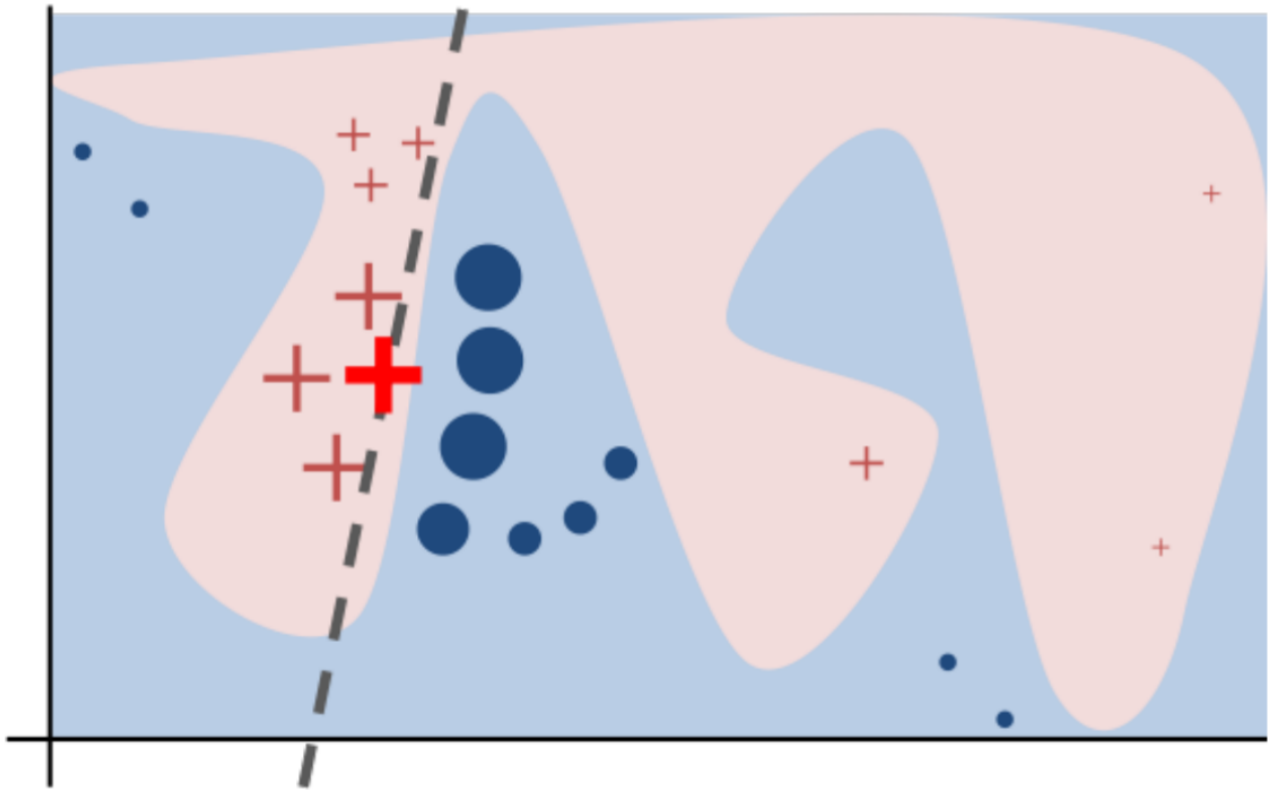
Local Approach LIME: What does it do?

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



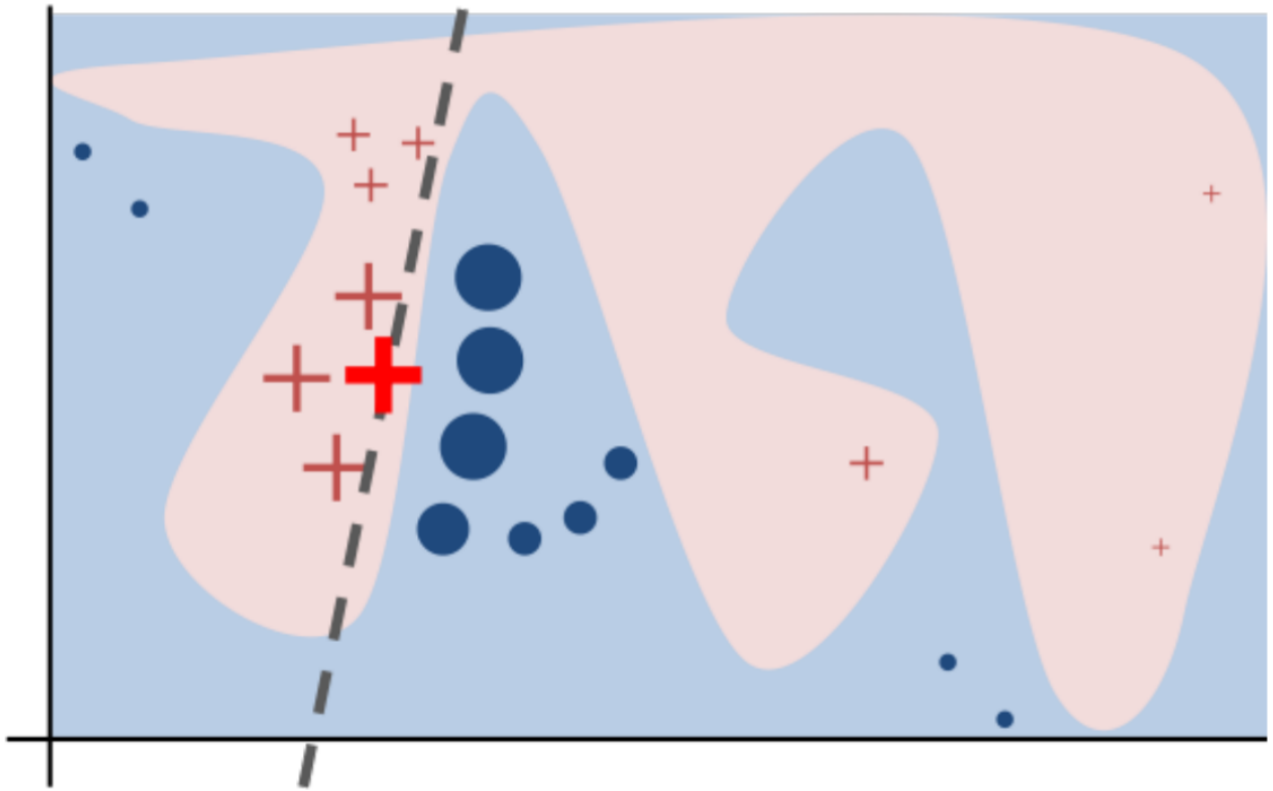
Local Approach LIME: What does it do?

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



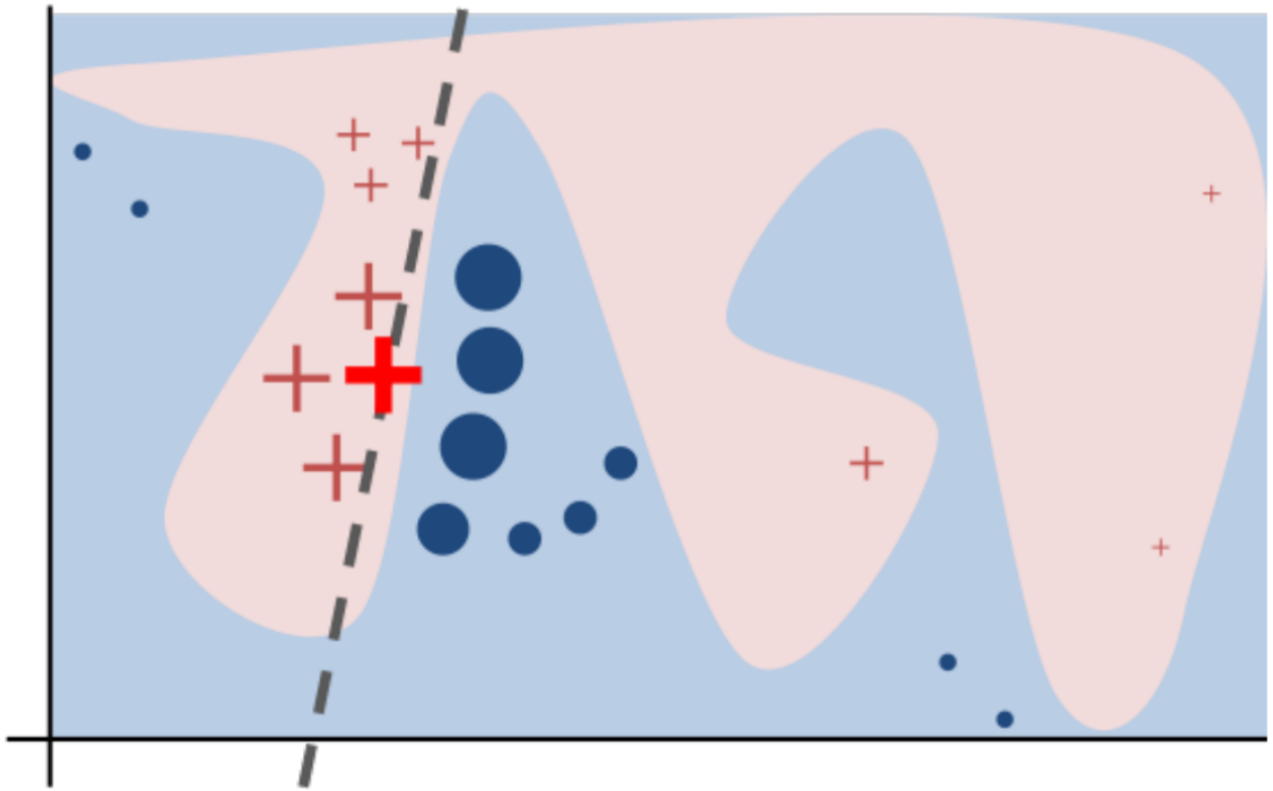
Local Approach LIME: What does it do?

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



Local Approach LIME: What does it do?

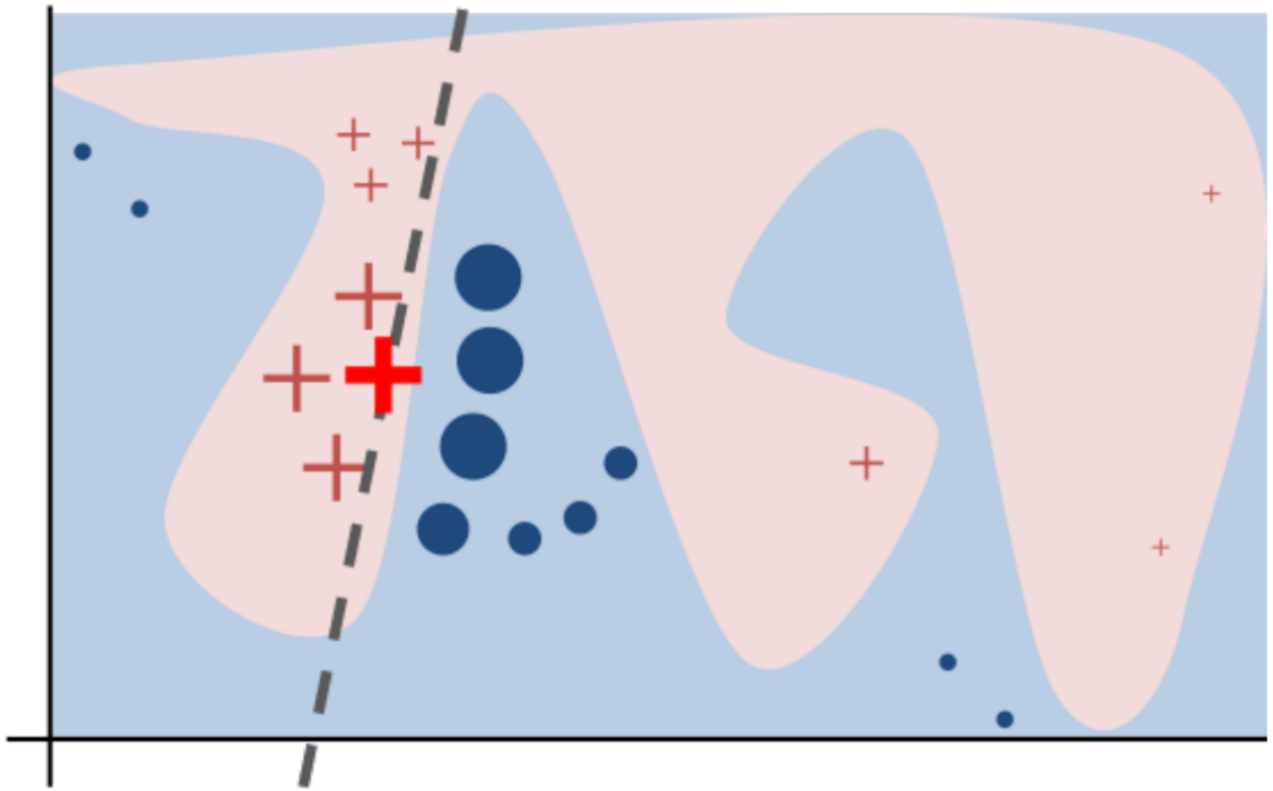
1. *Perturbate data*
2. *Compute proximity*
3. *Make predictions*
4. *(Select features)*
5. ***Fit a simple model***
6. *Extract explanations (feature weights)*



Local Approach LIME: What does it do?

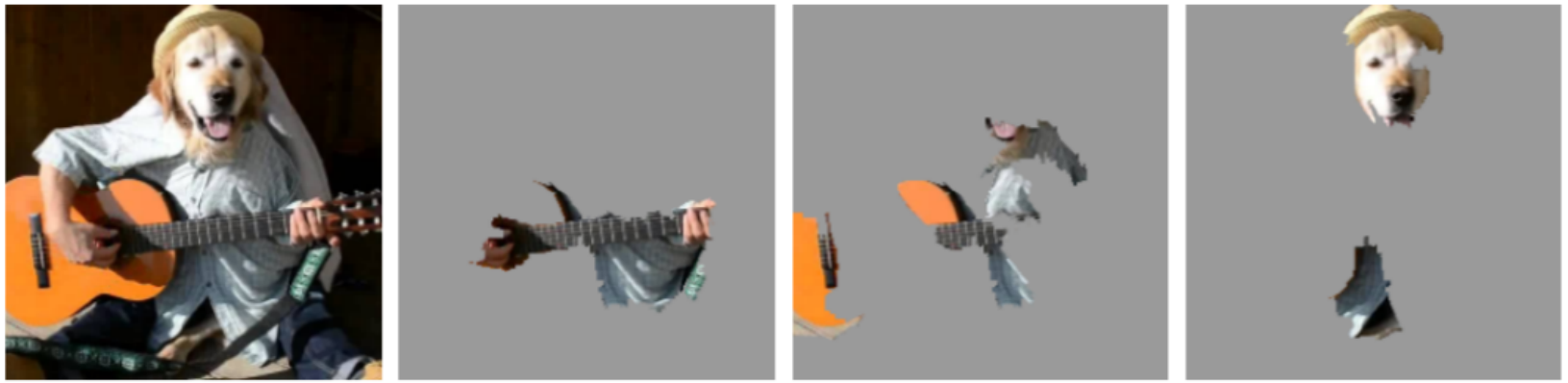
1. *Perturbate data*
2. *Compute proximity*
3. *Make predictions*
4. *(Select features)*
5. *Fit a simple model*
6. ***Extract explanations***

   (feature weights)



Local Approach LIME: Sample around the instance

1. Engineering of perturbation generator
   - Static data?
   - Pictures?
   - Time series?



2. Collect a set of permutations
3. Define a proximity measure

Local Approach LIME: Fit a simple model

- 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

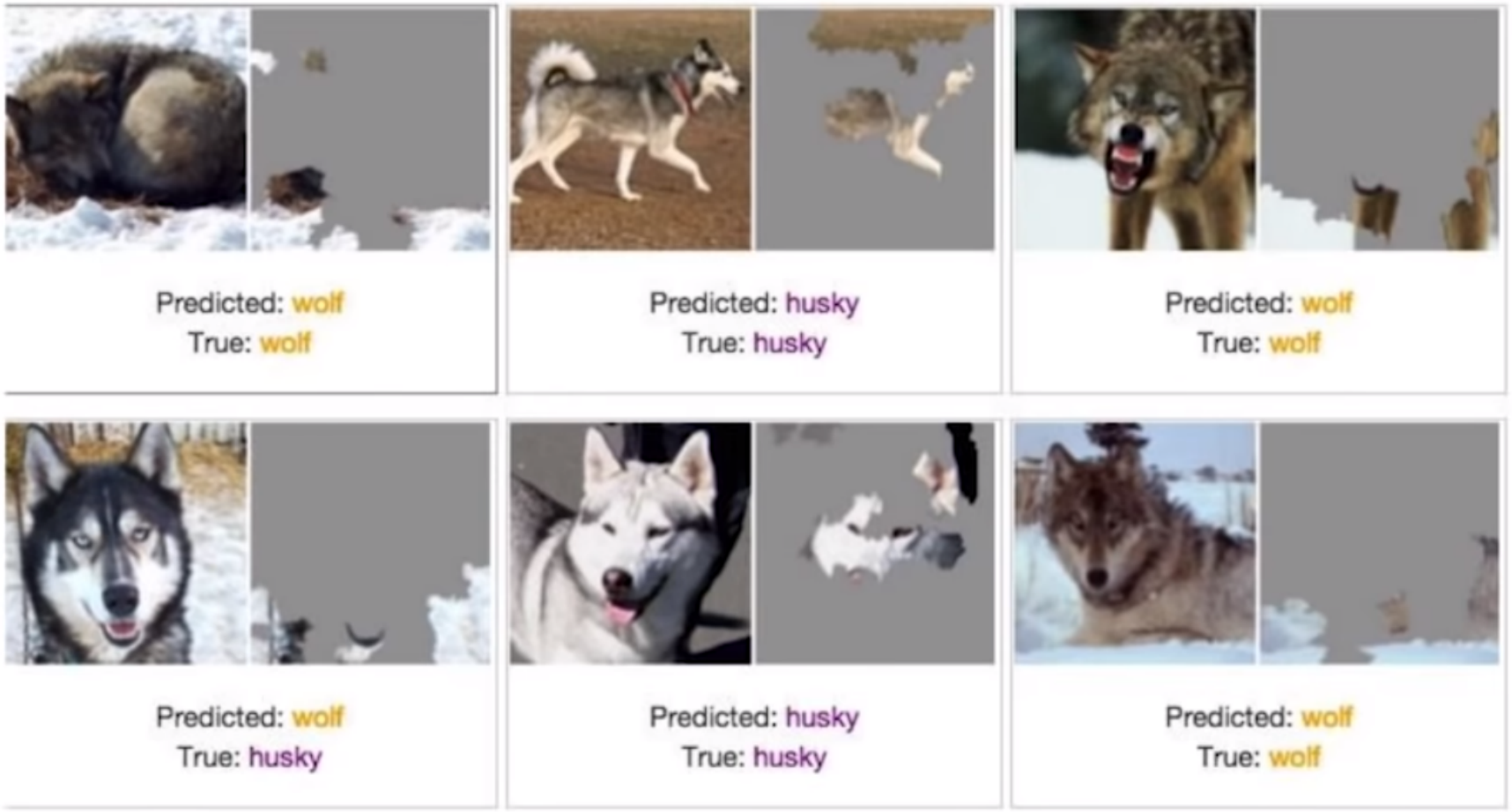
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- Extract explanations (e.g. model weights)

Local Approach LIME: Example

Only one mistake!

Local Approach LIME: Example



It's a great snow detector!

Global Approach: Decision Rule Lists (DRL)

```
IF perfusion = low && body_temp = high THEN risk = high
```

Support/Coverage:

Share of matched instances by rule

$$S(r) = \text{Matched instances} / \text{Total instances}$$

Accuracy/Confidence:

How accurate is the rule in predicting the correct class for the instances for which the rule applies?

$$A(r) = \text{Correctly classified matched instances} / \text{Total matched instances}$$

- Overlapping rules
- No matching rules

- Decision Lists
- Decision Sets

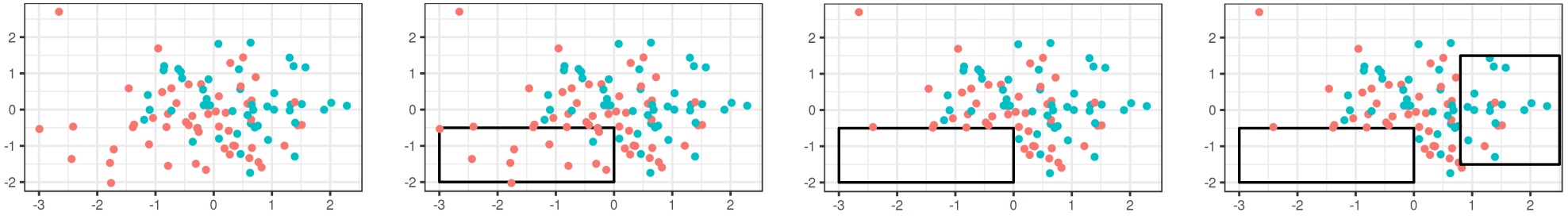
Inspired by: https://christophm.github.io/interpretable-ml-book/rules.html#sequential-covering

Global Approach DRL: Sequential Covering

Sequential Covering

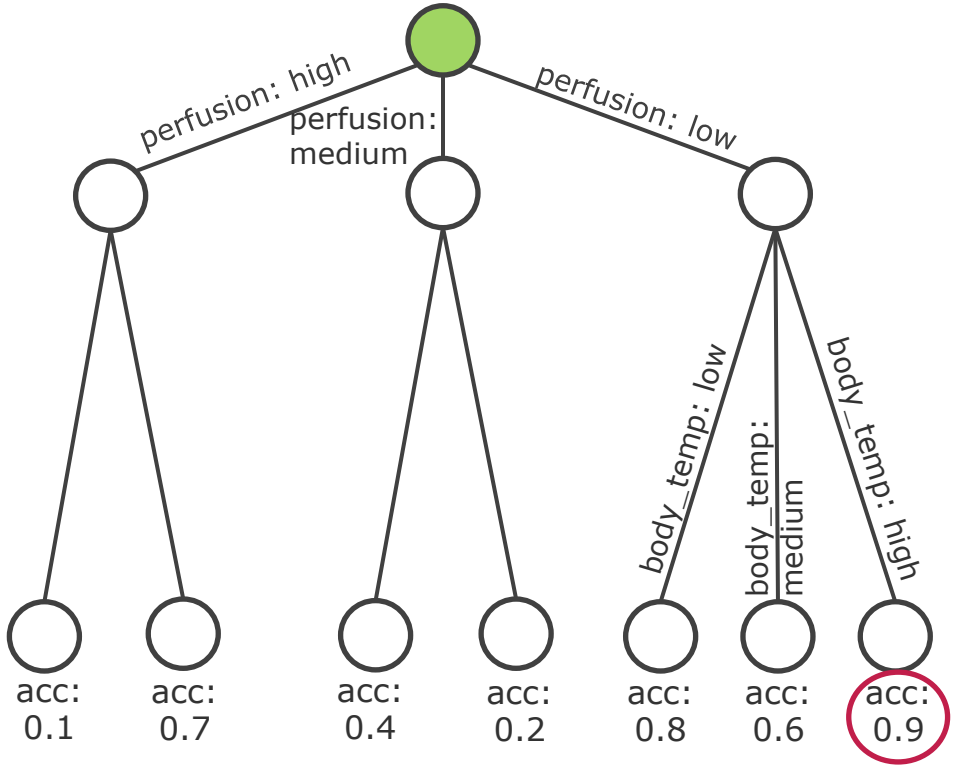
"Create rules until dataset is covered"

```
1   rule_list = []
2   WHILE rule_list.getCoverage() < threshold:
3         r = learn_rule(data)
4         rule_list.push(r)
5         data.removeInstancesCoveredBy(r)
6   return rule_list
```



Inspired by: https://christophm.github.io/interpretable-ml-book/rules.html#sequential-covering

Global Approach DRL: Learning a Rule



IF *perfusion = low && body_temp = high* THEN *risk = high*

Inspired by: https://christophm.github.io/interpretable-ml-book/rules.html#sequential-covering

Global Approach DRL: Dimensions of Interpretability

INTRINSIC MODEL

vs

POST-HOC MODELS

METHOD OUTCOME

interpretable decision rule list

feature ranking

surrogate model

MODEL-SPECIFIC

vs

MODEL-AGNOSTIC

LOCAL SCOPE

vs

GLOBAL SCOPE

Contribution

- Clinical Predictive Model for AKI Use Case
- Extend current Machine Learning Pipeline
- **Build Integrated Interpretability Framework**
  - Compare interpretability models
  - Select interpretability models based on evaluation metrics
    - Computational Complexity
    - Model Complexity
    - Expert Feedback
    - Discrimination & Calibration

Desired Outcome

Sources

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

Uni Heidelberg Logo
https://de.wikipedia.org/wiki/Universit%C3%A4tsklinikum_Heidelberg#/media/File:Universit%C3%A4tsklinikum_Heidelberg_logo.svg

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 on https://flaticon.com

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

Hide the Pain Harolds:

- https://static.independent.co.uk/s3fs-public/thumbnails/image/2017/07/11/11/harold-0.jpg
- https://ih0.redbubble.net/image.427352071.9413/ap,550x550,16x12,1,transparent,t.png

Interpretability Approaches applied to Clinical Predictive Modeling

Trends in Bioinformatics

Intermediate Presentation

Discussion

Possible Questions:

- 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?
- How could we engineer perturbations for time series in LIME?