Explainable ML

CORRELAID

GOOD CAUSES. BETTER EFFECTS.

Agenda

Introduction

- Why interpretable ML?
- Business context

Theory

- A methodology for interpretable ML
- Methods for interpretable ML

Practice

- Demo SHAP

Some further links

- Hands-On Interpretable ML: https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608
- Interpretable ML Book https://christophm.github.io/interpretable-ml-book/taxonomy-of-interpretability-methods.html
- Why and how (no formulas!): https://towardsdatascience.com/why-how-interpretable-ml-7288c5aa55e4
- NIPS 2017 Lundberg: wrote the python package for SHAP https://www.facebook.com/nipsfoundation/videos/1553537531404147/
- Inspiration: https://hci.iwr.uni-heidelberg.de/teaching/seminar_explainable_ML_2018
- Video on LIME https://www.youtube.com/watch?v=KP7-JtFMLo4
- LIME founding paper: https://arxiv.org/abs/1602.04938
- SHAP values: https://towardsdatascience.com/one-feature-attribution-method-to-supposedly-rule-them-all-shapley-values-f3e04534983d
- SHAP values https://github.com/slundberg/shap

Why does ML need to be interpretable?

Transparency is necessary for law enforcement

- Black boxes take decisions, but who is responsible for these decisions?
- Still highly debated question and growing need to address it
- Especially in ethically complex situations, we want to avoid our models to have the same biases that humans have

Why does ML need to be interpretable?

No understanding - no trust – (no usage)

- *“Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model.”* (Ribeiro et al. 2016)

- We tend to trust things we understand or at least can follow

- What's the best model worth if no one uses it?

Why does ML need to be interpretable?

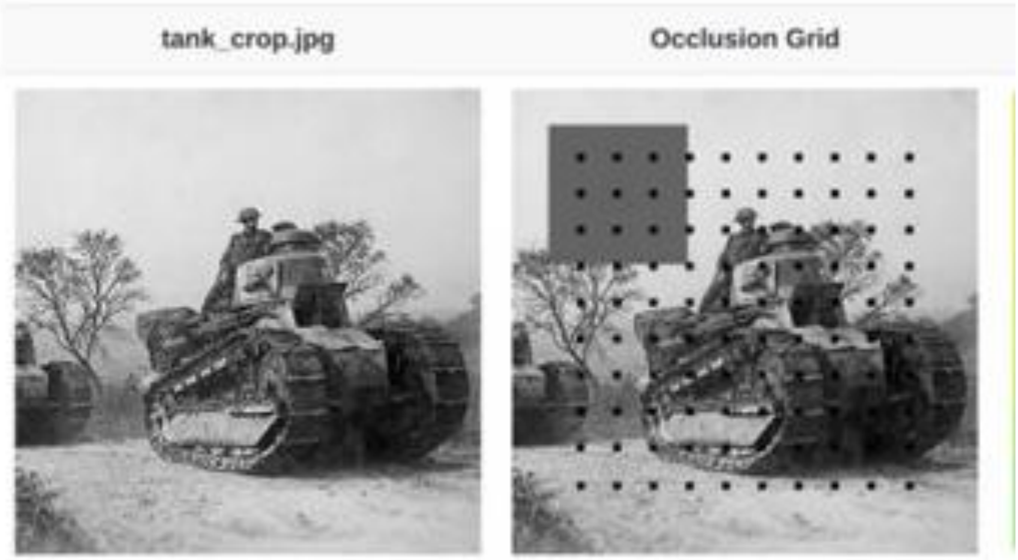
Best results in our field can be coined by human/machine interaction projects

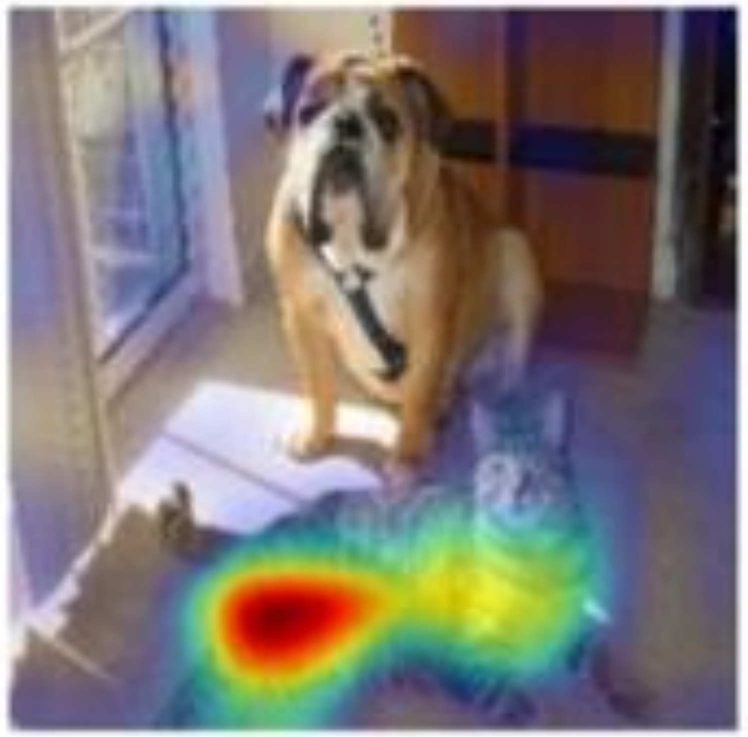
- A model is a *representation* of reality. Mapping is human-made.
- Incomplete data can be completed by human knowledge.
- No person will be convinced by a human not convinced by the model results.

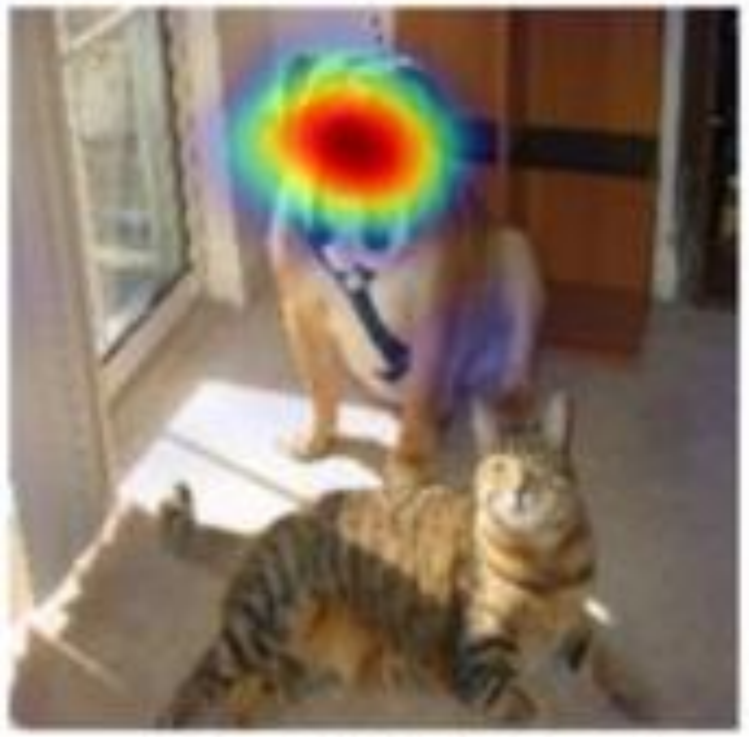
Why does ML need to be interpretable?

Models can be improved through understanding

- When the classifying mechanism itself is important, explainable ML can help debug
- Does my classifyier really capture what I want it to capture? Then I can make inference of the data I need!




(c) Grad-CAM 'Cat'


(i) Grad-CAM 'Dog'

Side Note – What's interpretable?

- "Interpretability" or "Explainability" has been around for quite some time, but in 2016, Zachary Lipton points out:

> *"From this, we might conclude that either:*
>
> *(i) the definition of interpretability is universally agreed upon, but no one has managed to set it in writing, or*
>
> *(ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character.*
>
> *Our investigation of the literature suggests the latter to be the case."*

Side Note – What's interpretable?

So what's interpretability? (Lipton 2016)

> *"The demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real-world costs in a deployment setting."*

- ➢ What does that mean?
  - ➢ **We need interpretability, when just a score and good model metrics are not enough**
  - ➢ **We do not only care, how often a model is right, but in which examples it is right**
    - ➢ **This is where human/machine interaction is possible**

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

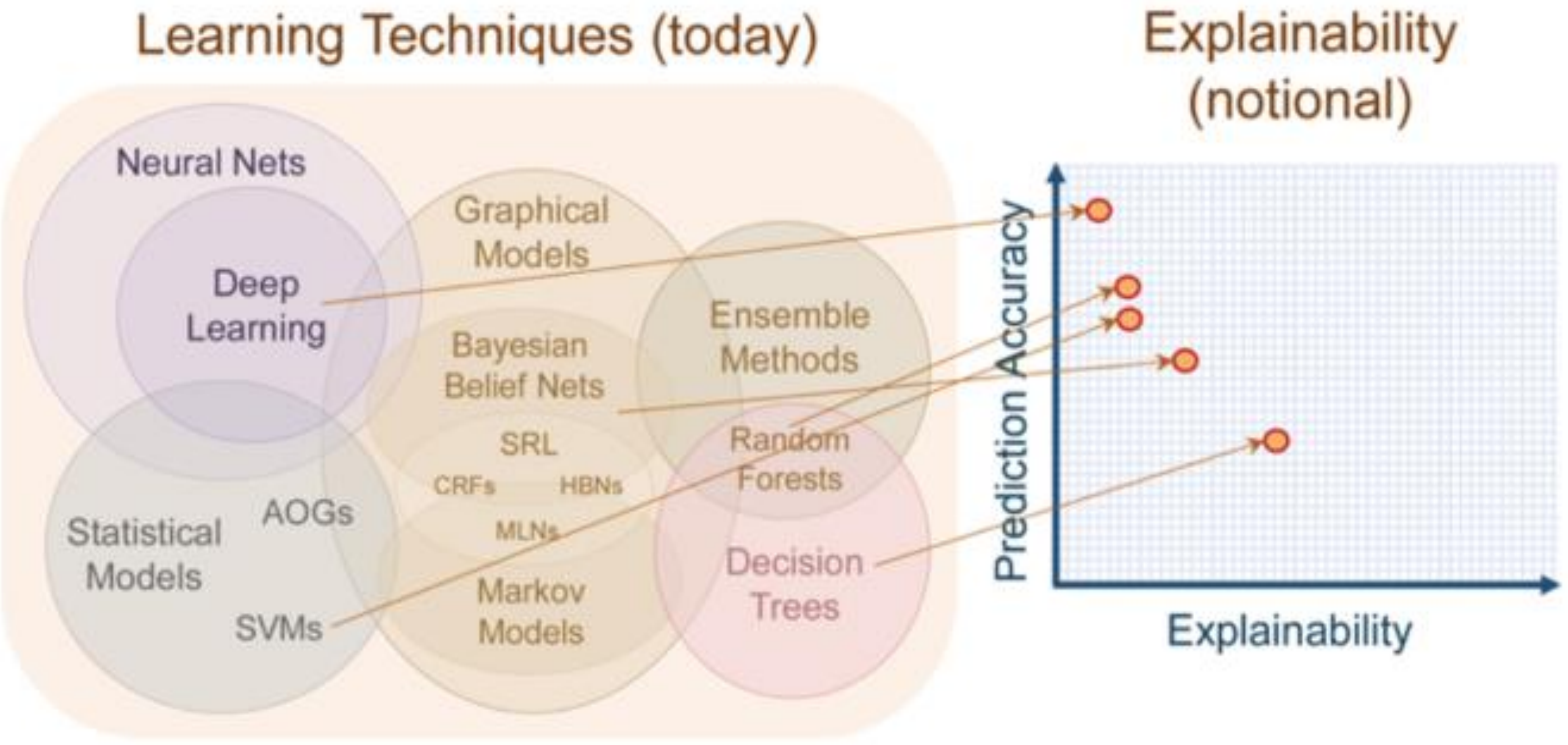
Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017))

"Interpretability is the degree to which a human can consistently predict the model's result"

(Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016))

Why does ML need to be interpretable?

We have a trade off right now between models that yield highest results in their predictions and the explainability of these models.



The Explainability — Accuracy tradeoff (source)

“As exciting as their performance gains have been, though, there’s a troubling fact about modern neural networks: Nobody knows quite how they work. And that means no one can predict when they might fail.” (http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable)

“The result is that modern machine learning offers a choice among oracles: Would we like to know *what* will happen with high accuracy, or *why* something will happen, at the expense of accuracy? The “why” helps us strategize, adapt, and know when our model is about to break. The “what” helps us act appropriately in the immediate future.” (http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable)

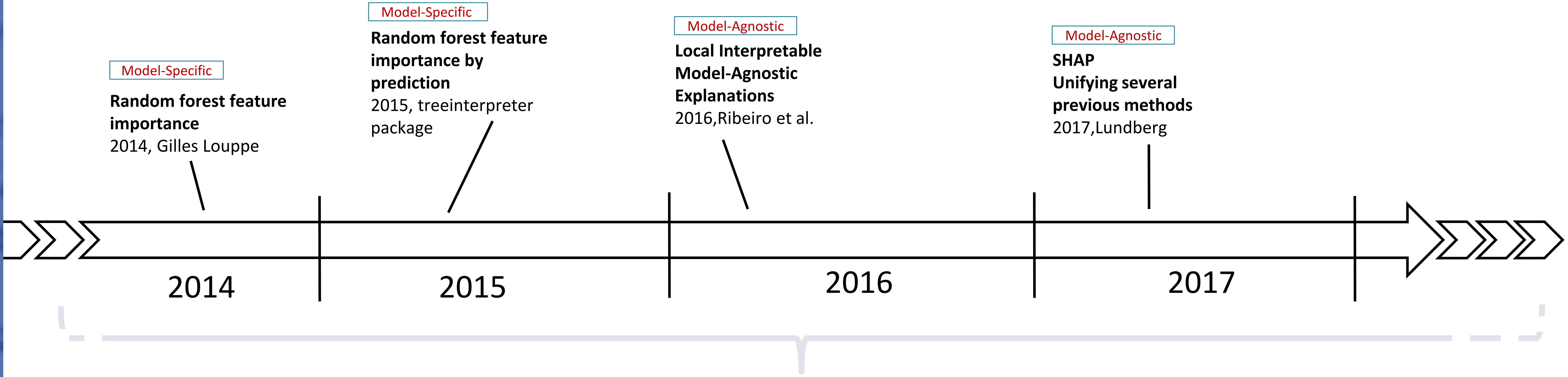
In an ideal world, we need both: Interpretability and accuracy (or whatever measure you want). Still the tradeoff exists. But we start dealing with it!

Much talk to arrive here: THIS is the reason of this workshop, show you some of these approaches.

https://medium.com/@Zelros/a-brief-history-of-machine-learning-models-explainability-f1c3301be9dc

Model Interpretability

Recent development: Taking black boxes apart



Model-Specific

Random forest feature importance
2014, Gilles Louppe

Model-Specific

Random forest feature importance by prediction
2015, treeinterpreter package

Model-Agnostic

Local Interpretable Model-Agnostic Explanations
2016,Ribeiro et al.

Model-Agnostic

SHAP
Unifying several previous methods
2017,Lundberg

Fast evolution in the field of neural networks
and deep learning, all in need for
interpretability

Quelle: https://medium.com/@Zelros/a-brief-history-of-machine-learning-models-explainability-f1c3301be9dc

Agenda

Introduction

- Why interpretable ML?
- Business context

Theory

- A taxonomy for interpretable ML
- Methods for interpretable ML

Practice

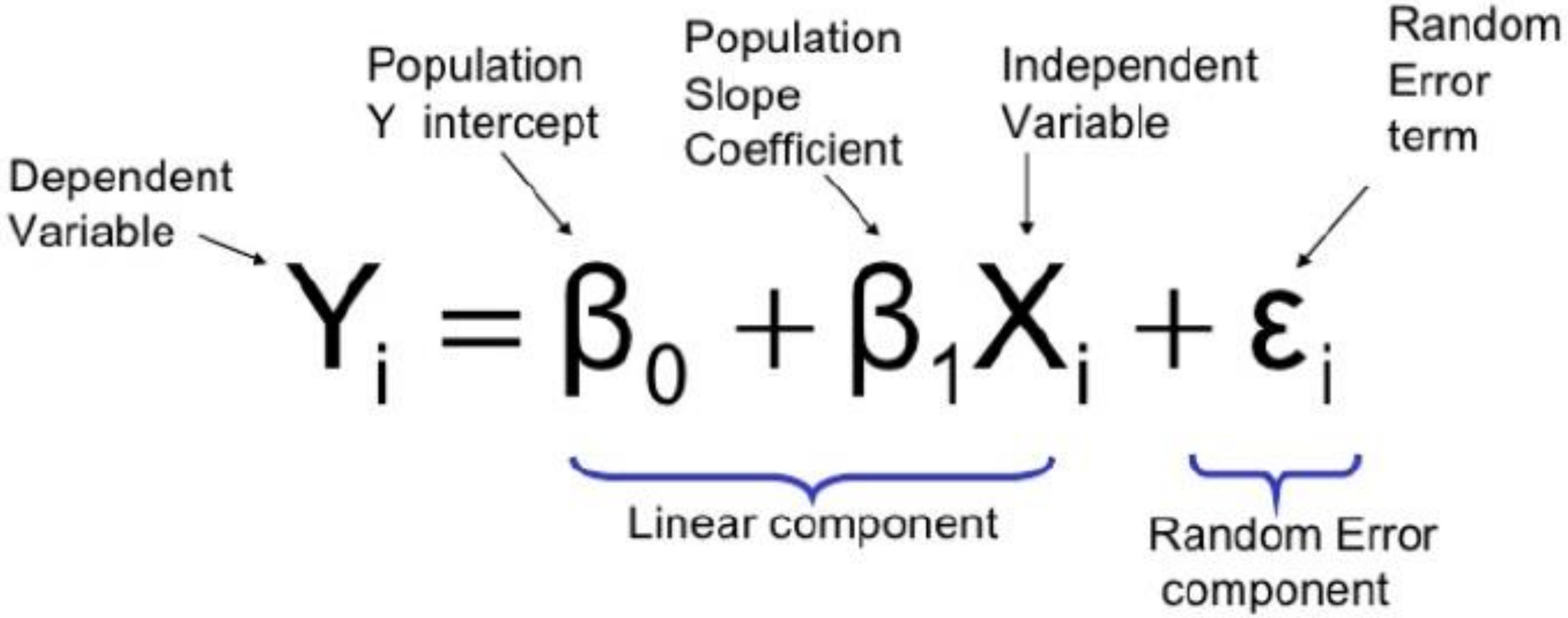
- Demo SHAP

How to make models better?

Use „white-box“ models	Often, simple linear relationships or single decision trees do not capture enough of the truth
Look for model performance metrics	Can be misleading → model inherent problems stay uncovered (data leaking, poor data…)
AB testing	Very expensive and often not possible
Make models interpretable	Computationally more expensive and also not easy to interpret – but a start

Linear/Logistic Regression

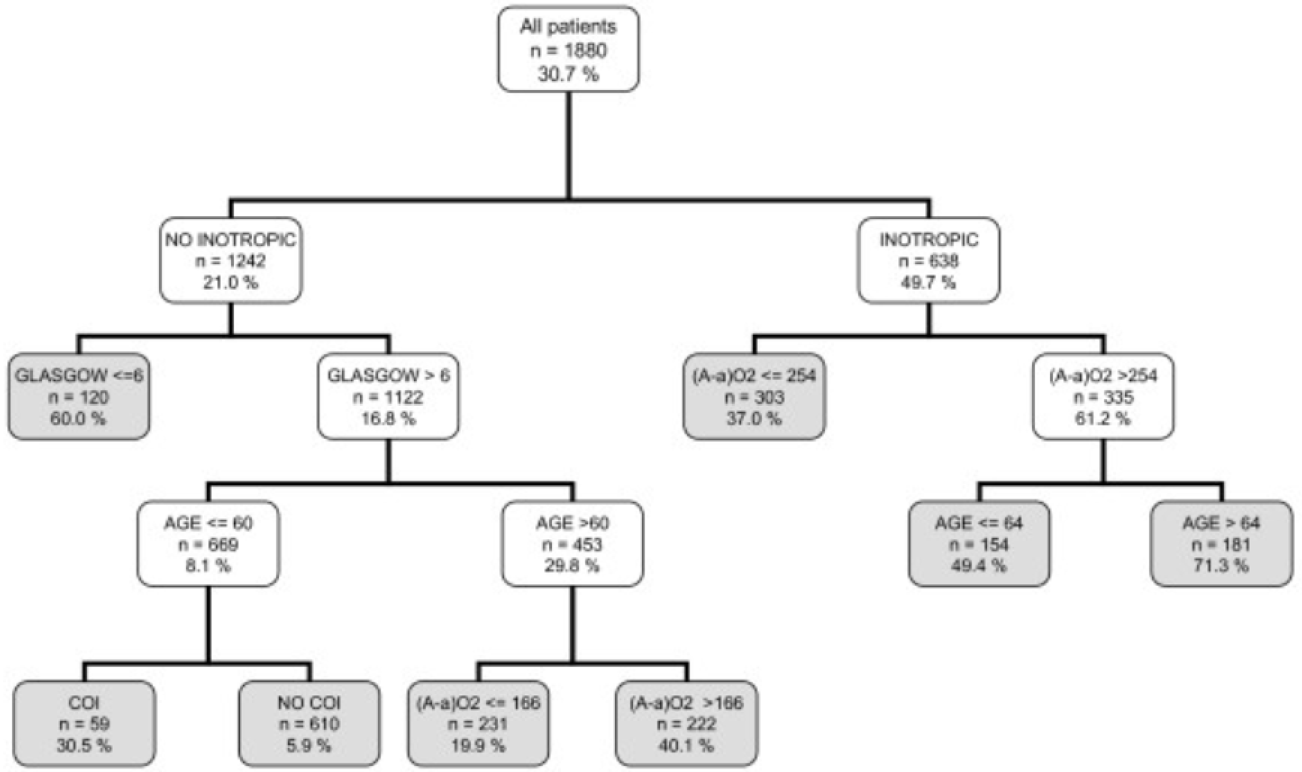
- Point Estimates
- P-Values
- Odds Ratios



$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Decision Trees

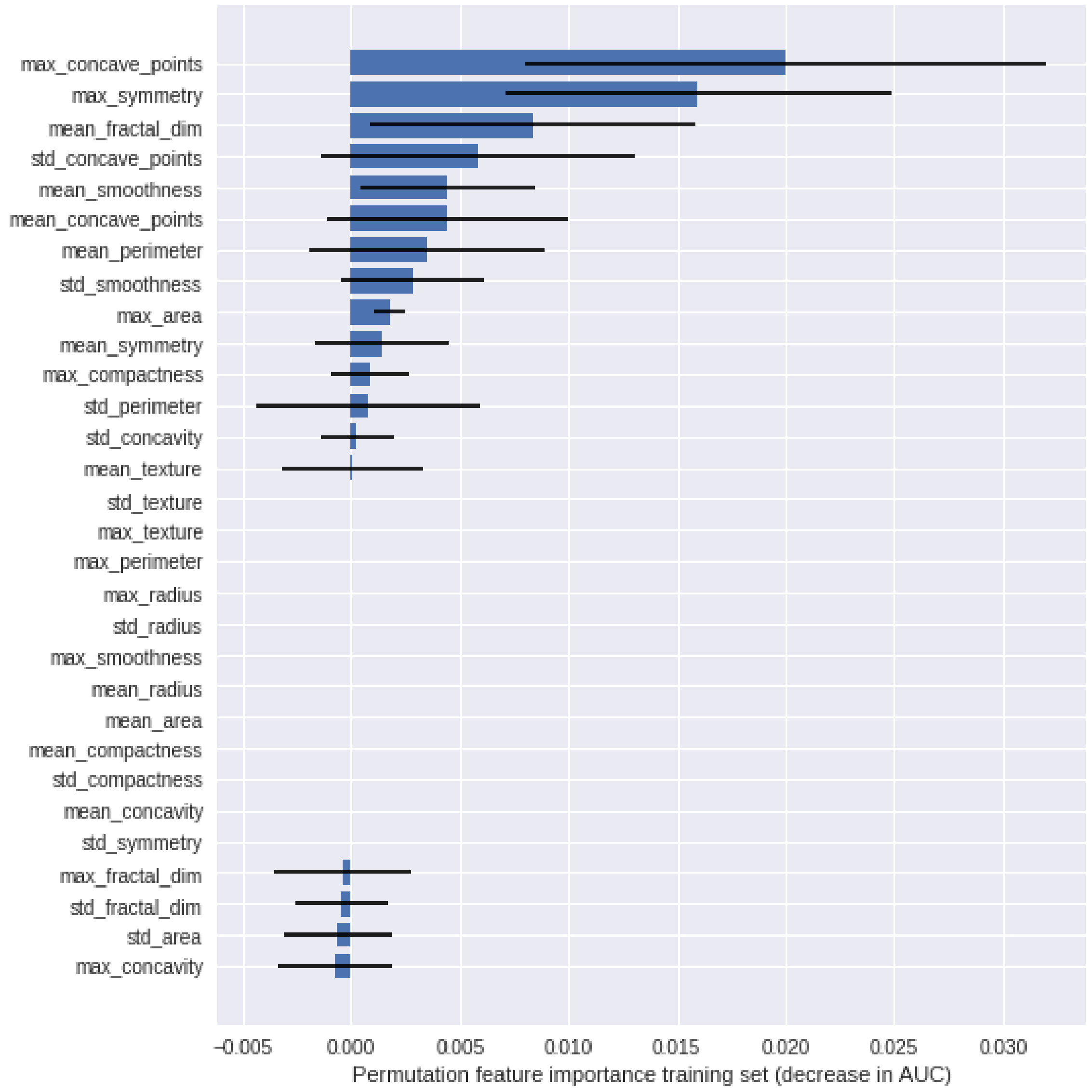
- Variables have easy to follow split-points that segment outcomes
- Shorter path length is more interpretable


All patients n = 1880 30.7 %
NO INOTROPIC n = 1242 21.0 %
INOTROPIC n = 638 49.7 %
GLASGOW <=6 n = 120 60.0 %
GLASGOW > 6 n = 1122 16.8 %
(A-a)O2 <= 254 n = 303 37.0 %
(A-a)O2 >254 n = 335 61.2 %
AGE <= 60 n = 669 8.1 %
AGE >60 n = 453 29.8 %
AGE <= 64 n = 154 49.4 %
AGE > 64 n = 181 71.3 %
COI n = 59 30.5 %
NO COI n = 610 5.9 %
(A-a)O2 <= 166 n = 231 19.9 %
(A-a)O2 >166 n = 222 40.1 %

→ Though even “simple models” are not automatically interpretable. Depends on the dimensionality!

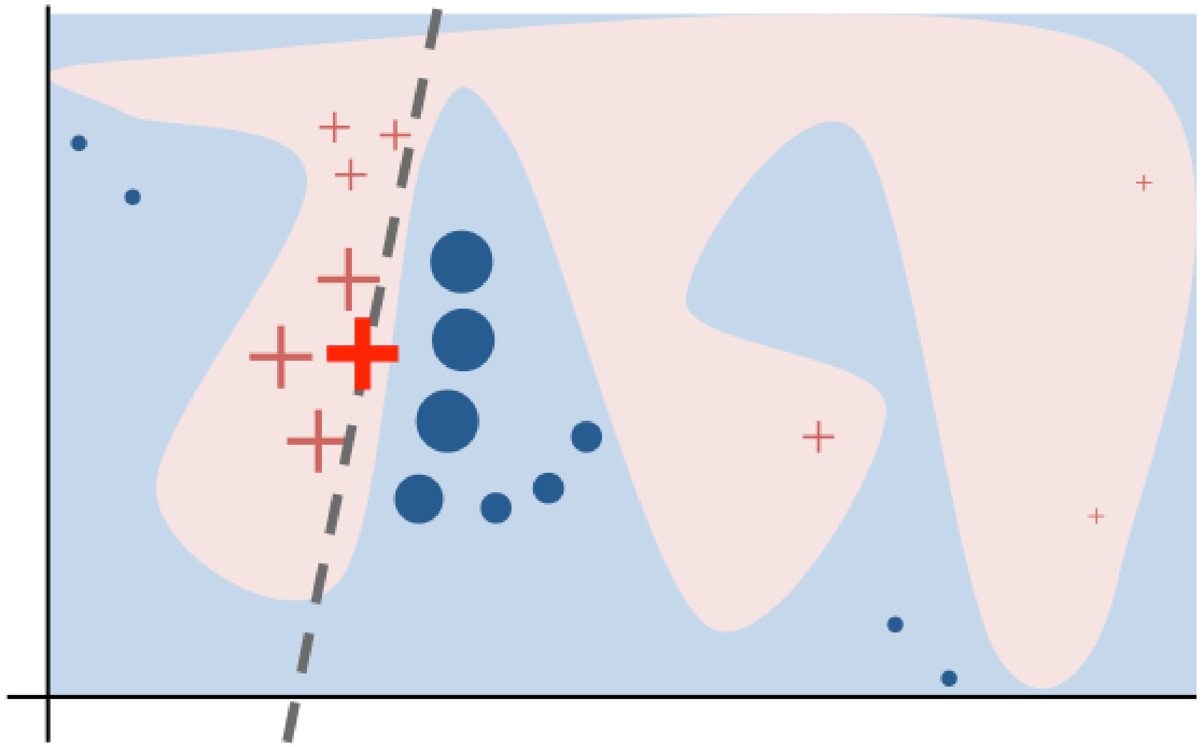
Random Forests & GBM

- **Feature ablation**: Variables are included / excluded in various model iterations
- Measure importance by decrease in accuracy or node purity
- **Feature importance**: Various ways to attribute feature importance (more on that later)



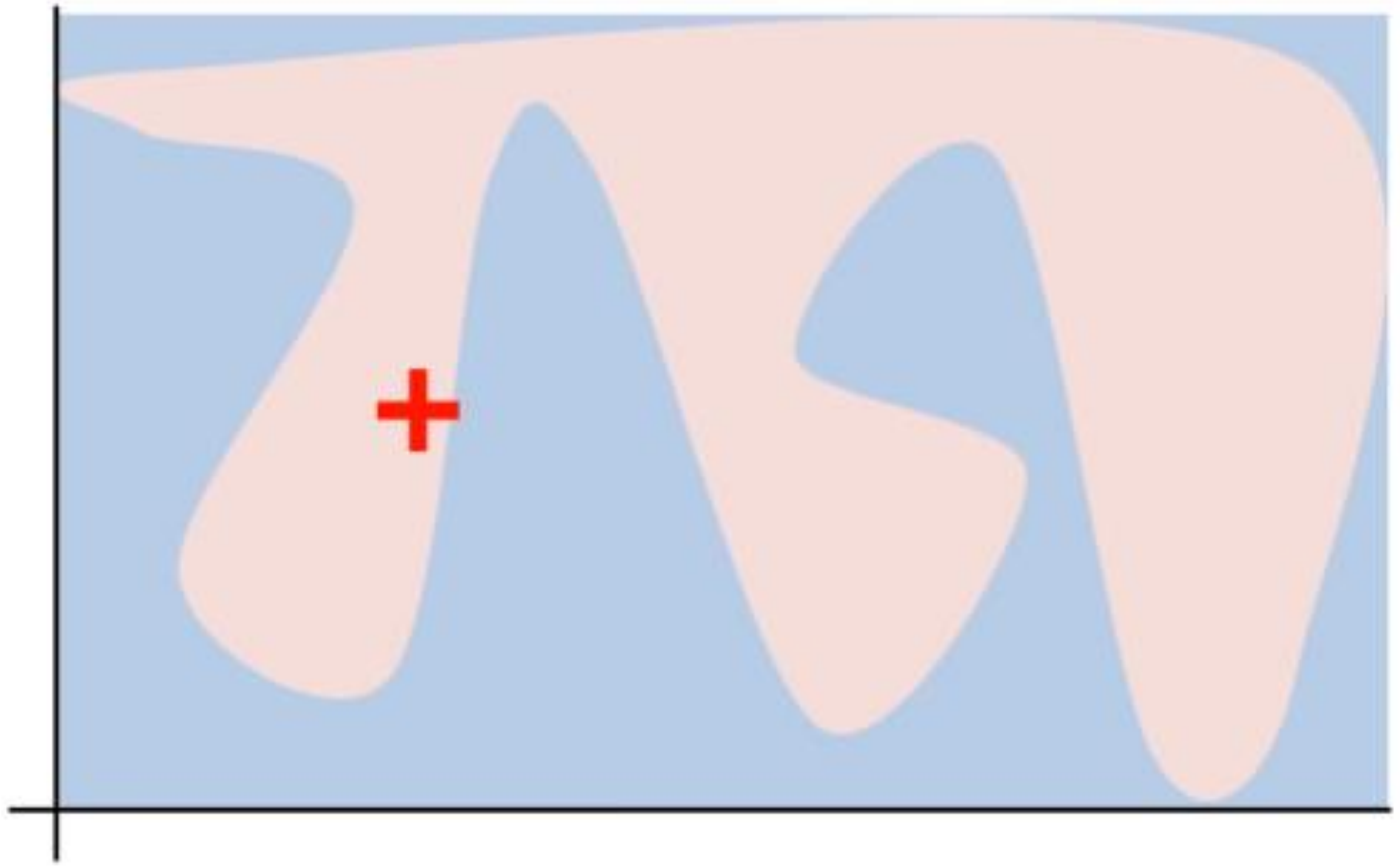
Local Interpretable Model-Agnostic Explanations (LIME)

- Local interpretable model-agnostic explanations (LIME)
- Method for fitting local, interpretable models that can explain single predictions of any black-box machine learning model.
- Surrogate models are interpretable models that are learned on the predictions of the original black box model
- Instead of globally approximating a model, we approximate it locally



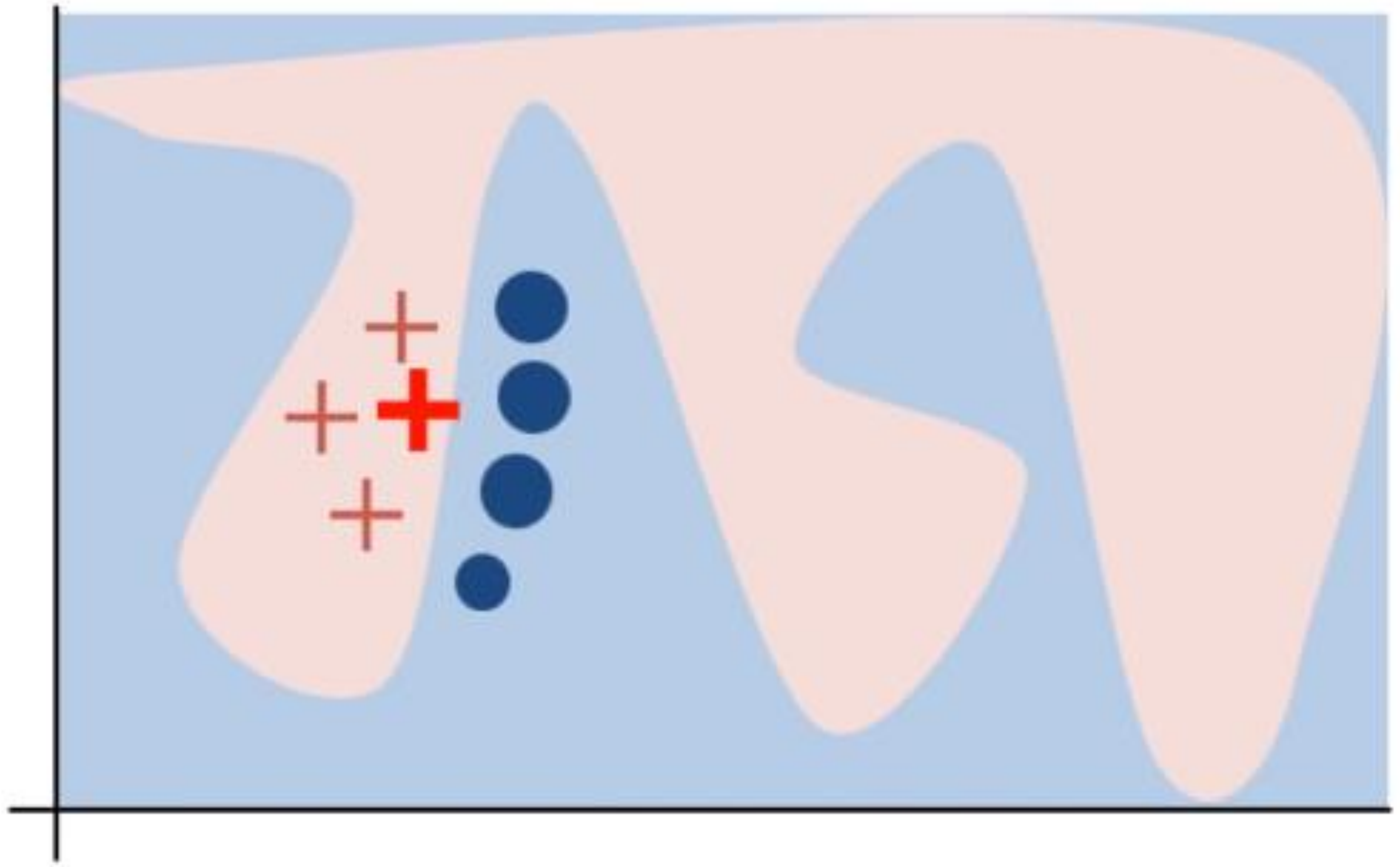
Local Interpretable Model-Agnostic Explanations (LIME)

- Pick data point
- Create “perturbed” data points
- Calculate distance between “perturbed” and original data point
- Make predictions on “perturbed” data points with complex model
- Pick m features best describing the complex model outcome from “perturbed” data
- Fit a simple model to the “perturbed” data with m and similarity scores as weights
- Feature weights from the simple model make explanations for the complex models local behavior



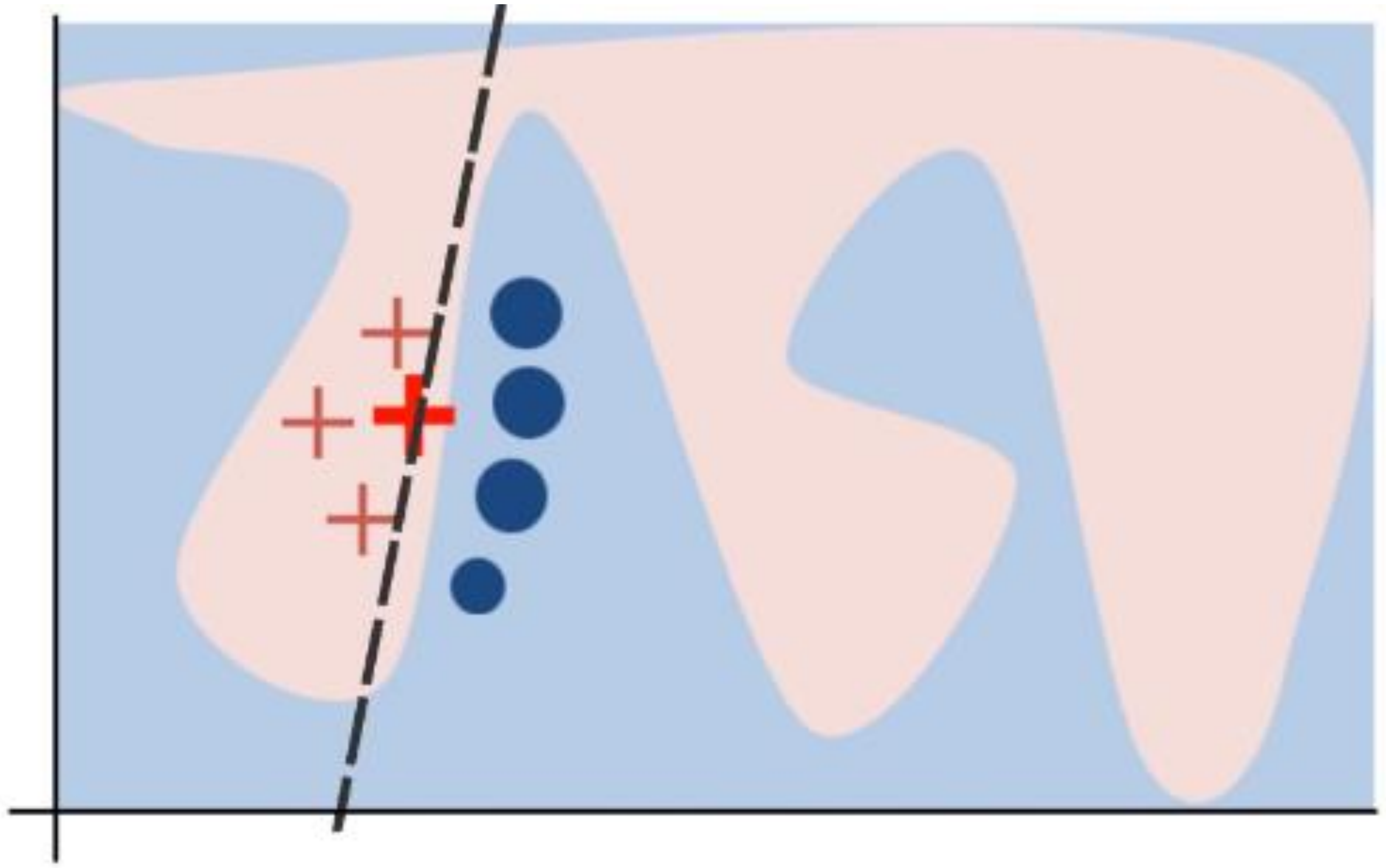
Local Interpretable Model-Agnostic Explanations (LIME)

- Pick data point
- Create “perturbed” data points
- Calculate distance between “perturbed” and original data point
- Make predictions on “perturbed” data points with complex model
- Pick m features best describing the complex model outcome from “perturbed” data
- Fit a simple model to the “perturbed” data with m and similarity scores as weights
- Feature weights from the simple model make explanations for the complex models local behavior



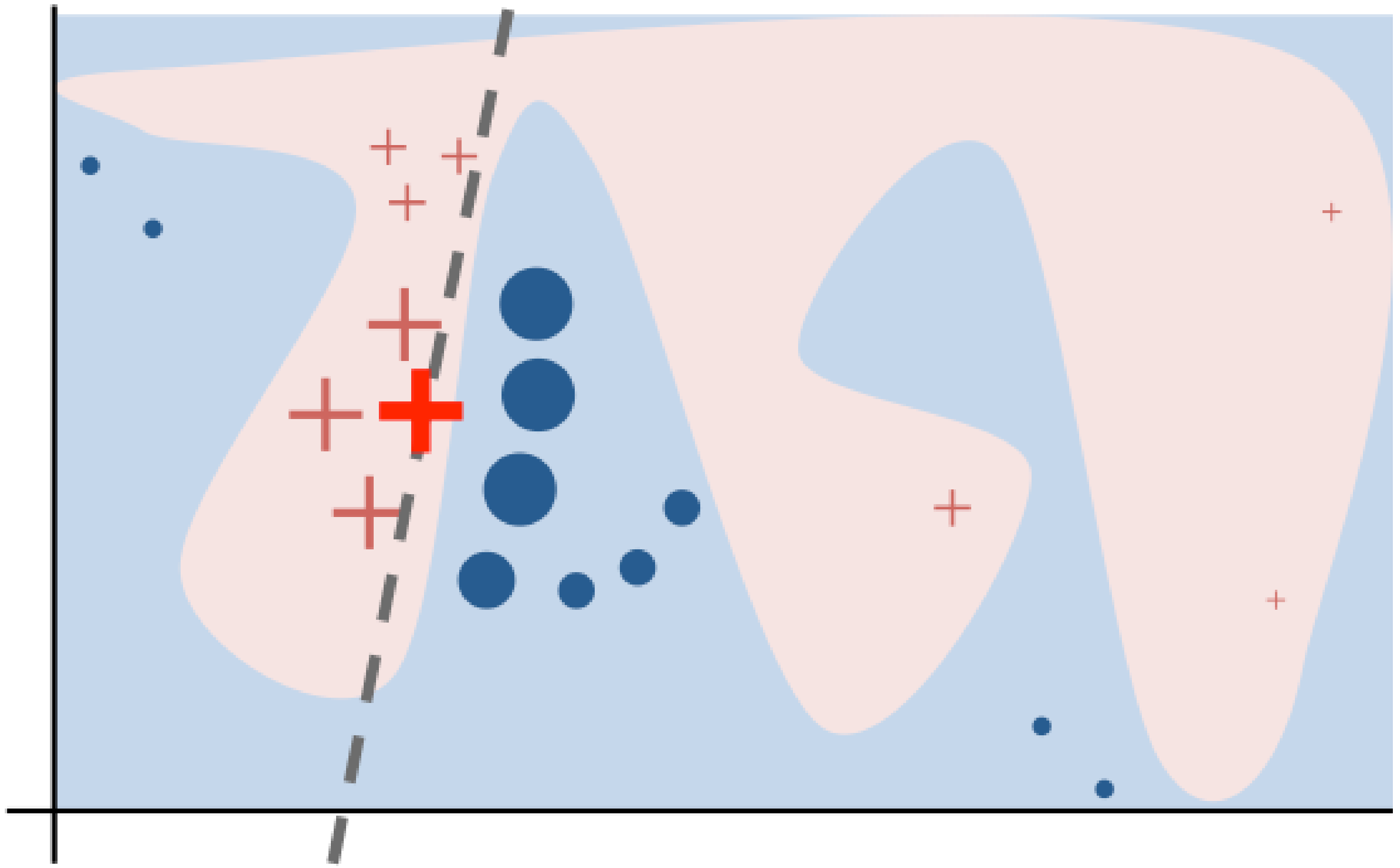
Local Interpretable Model-Agnostic Explanations (LIME)

- Pick data point
- Create “perturbed” data points
- Calculate distance between “perturbed” and original data point
- Make predictions on “perturbed” data points with complex model
- Pick m features best describing the complex model outcome from “perturbed” data
- Fit a simple model to the “perturbed” data with m and similarity scores as weights
- Feature weights from the simple model make explanations for the complex models local behavior

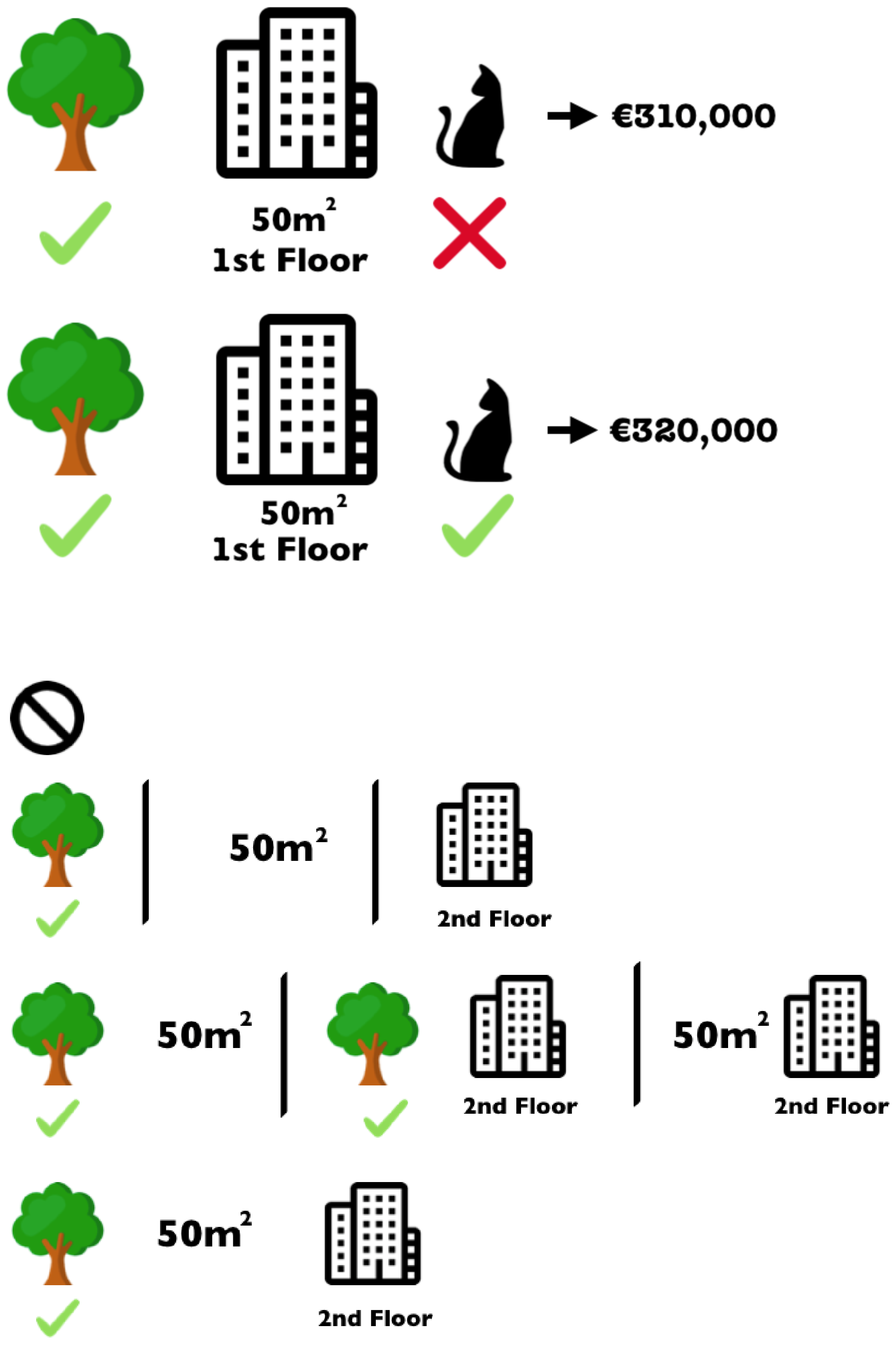


Local Interpretable Model-Agnostic Explanations (LIME)

- Pick data point
- Create “perturbed” data points
- Calculate distance between “perturbed” and original data point
- Make predictions on “perturbed” data points with complex model
- Pick m features best describing the complex model outcome from “perturbed” data
- Fit a simple model to the “perturbed” data with m and similarity scores as weights
- Feature weights from the simple model make explanations for the complex models local behavior



Shapley Values



Example: ML Model to predict housing prices

- For each coalition, compute the prediction w/ and w/o the selected feature value and take the difference to get the marginal contribution.
- The Shapley value is the average marginal contribution of a feature value across all possible coalitions

Interpretation:

- How much has each feature value contributed to the prediction compared to the average prediction?
- Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

Shapley Values

- Features ~ Players
- Finding each player’s marginal contribution, averaged over every possible sequence in which the players could have been added to the group
- E.g. for three features: ABC, ACB, BCA, BAC, CAB, and CBA
- For each sequence, capture the marginal payoff that accrued to each player. Then, average all of these payoffs together, and you have the Shapley values for each player.
- Next step: Instead of working directly with sequences, we work on different subsets of players, and weights those subsets based on what portion of all sequences they represent.
- SHAP: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] .$$

Notation: |F| is the size of the full coalition. S represents any subset of the coalition that doesn’t include player i, and |S| is the size of that subset. The bit at the end is just ”how much bigger is the payoff when we add player i to this particular subset S”

CORRELAID

For instant updates on our work follow us on Twitter

@CorrelAid

Receive the latest infos on our projects via Facebook

@WeAreCorrelAid

Sign up for our mailing list on

www.correlaid.org