

Современные методы анализа данных

Ансамбли

Analytics

- customer segmentation
 - preferences
 - behavior
 - personalization
- comparison of options
 - differences between groups in behavior
 - A/B tests
- relationships between parameters
 - what influences behavior
 - what does not work
 - interventions (what to change)

Features

- AI services and applications
 - (main idea)
 - recommendations
 - predictions generating
 - ...
- smart features
 - (additional functions)
 - autocomplete
 - ordering
 - options recognition
 - vote
 - text
 - Images

Why we use data

Analytics

with their help to answer some questions of interest to us

How justified are the conclusions?

- adequate data
- correct methods
- assumptions
- method limitations

Inference

Features

to build models that will perform the desired function

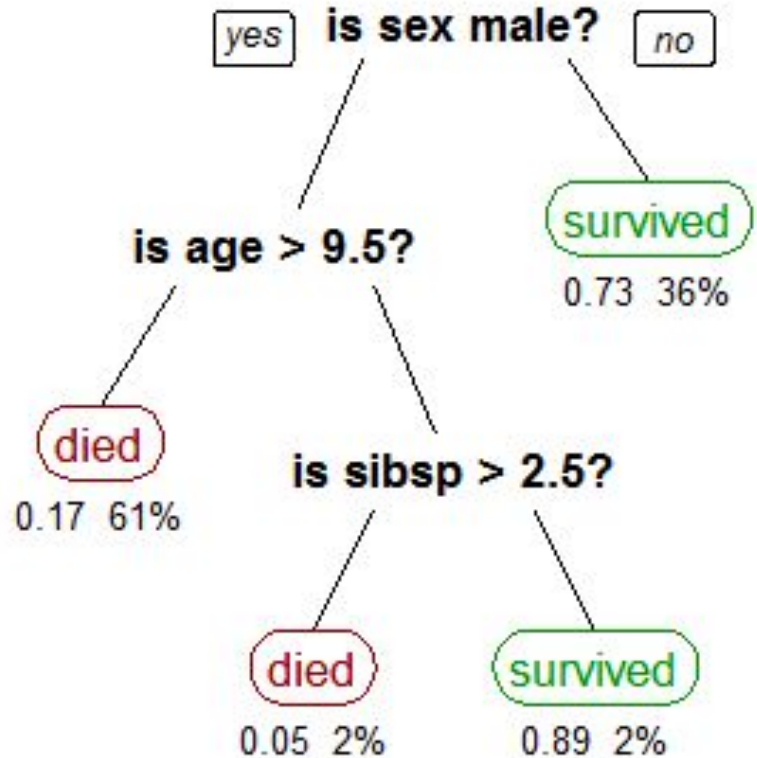
working or not:

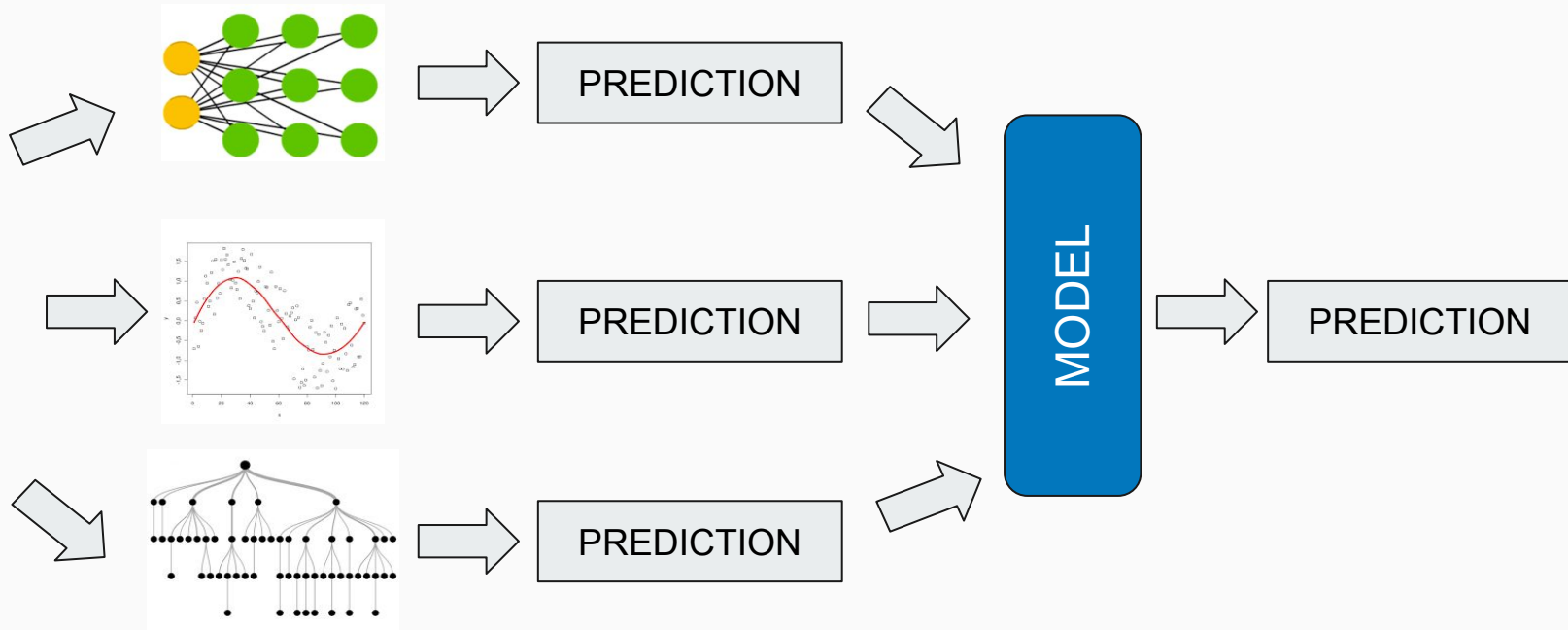
- not wrong (almost)
- implementation restrictions
 - speed
 - resource
- intensity adequate data

Prediction

Models: trees

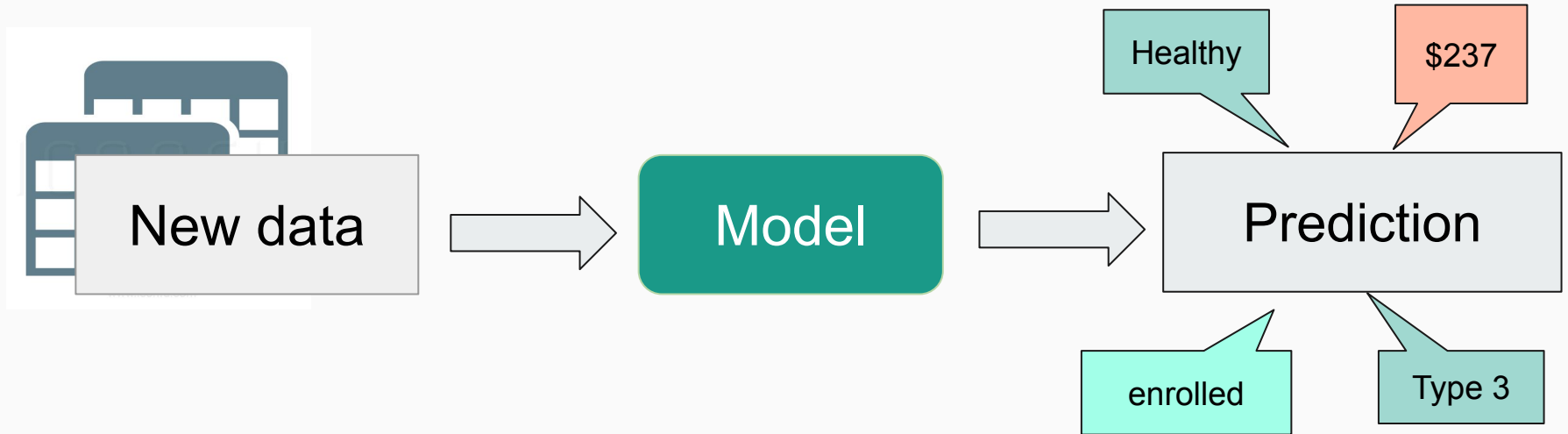
Classification tree for prediction of survivability on the Titanic





Пример: stacking

"Black box"



WHY

USE AI

- Quality improvement
- Opportunities for automation

Complex AI

- Limiting the ability to understand
- User mistrust

Interpretation instruments

- Decision Factors
- The logic of the models
- Difficulty in application



Husky

VS



Wolf

Ribeiro M. T., Singh S., Guestrin C. Why should i trust you?: Explaining the predictions of any classifier
(<https://arxiv.org/pdf/1602.04938.pdf>)





Answer: wolf

Model: wolf







Answer: husky

Model: husky



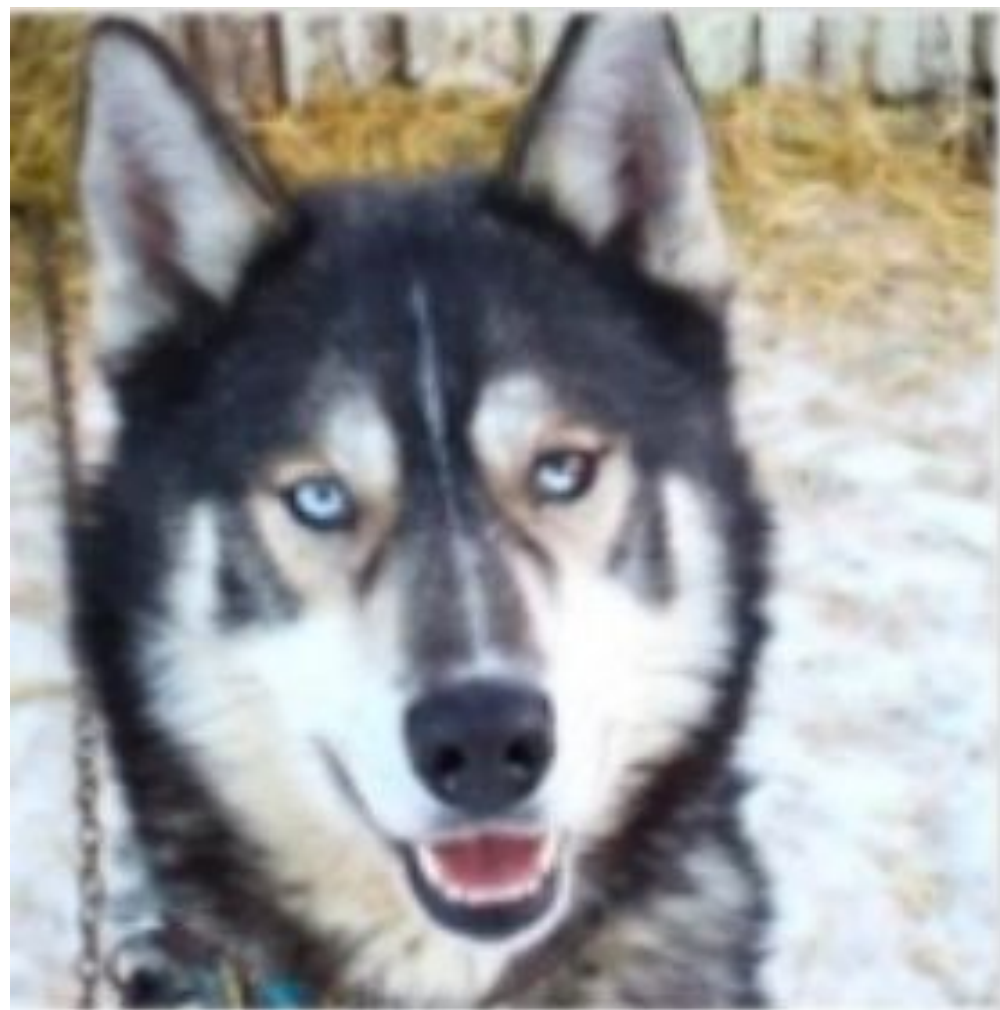


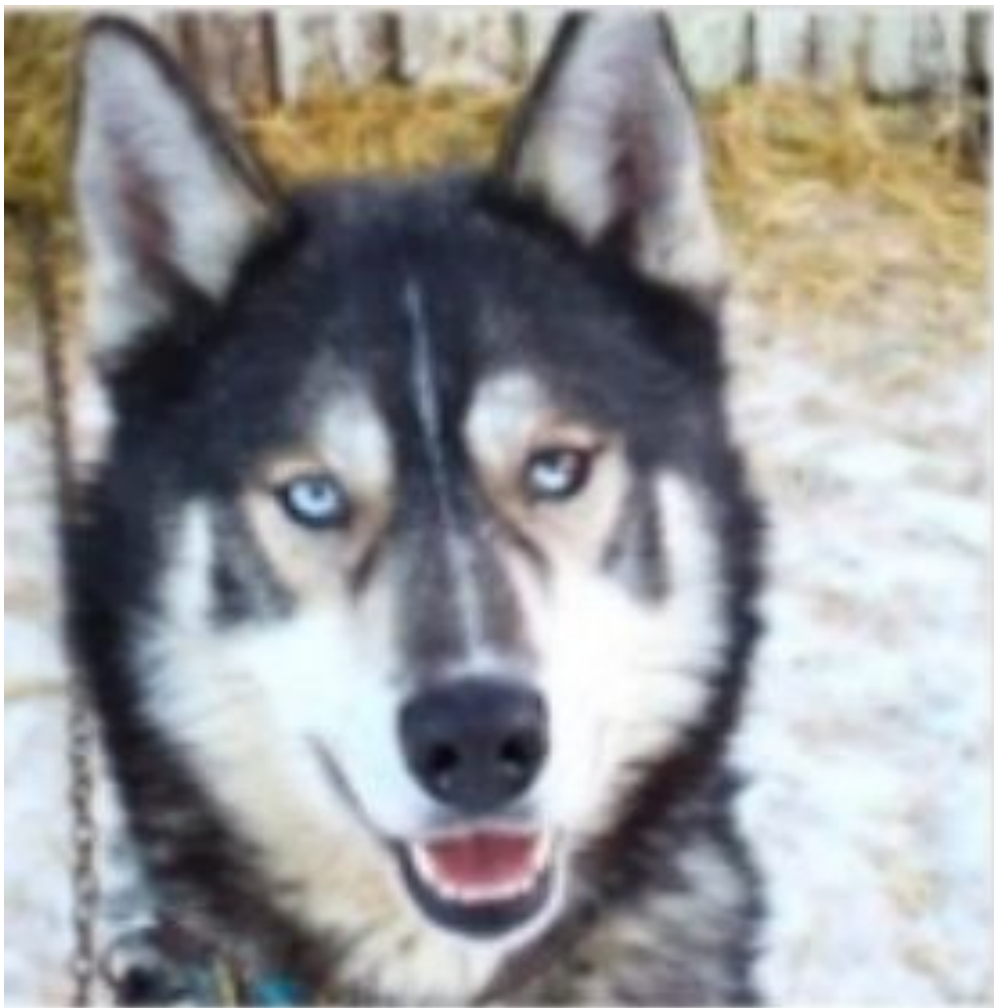


Answer: wolf

Model: wolf



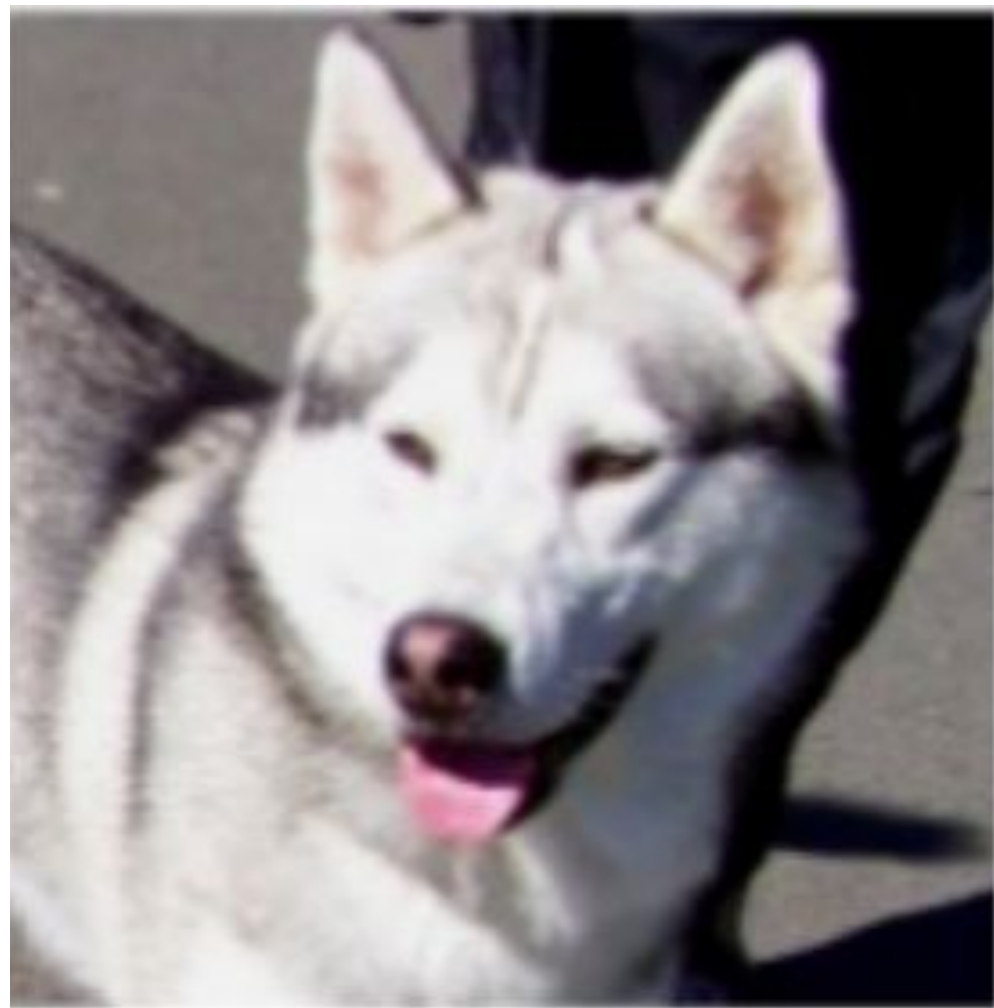


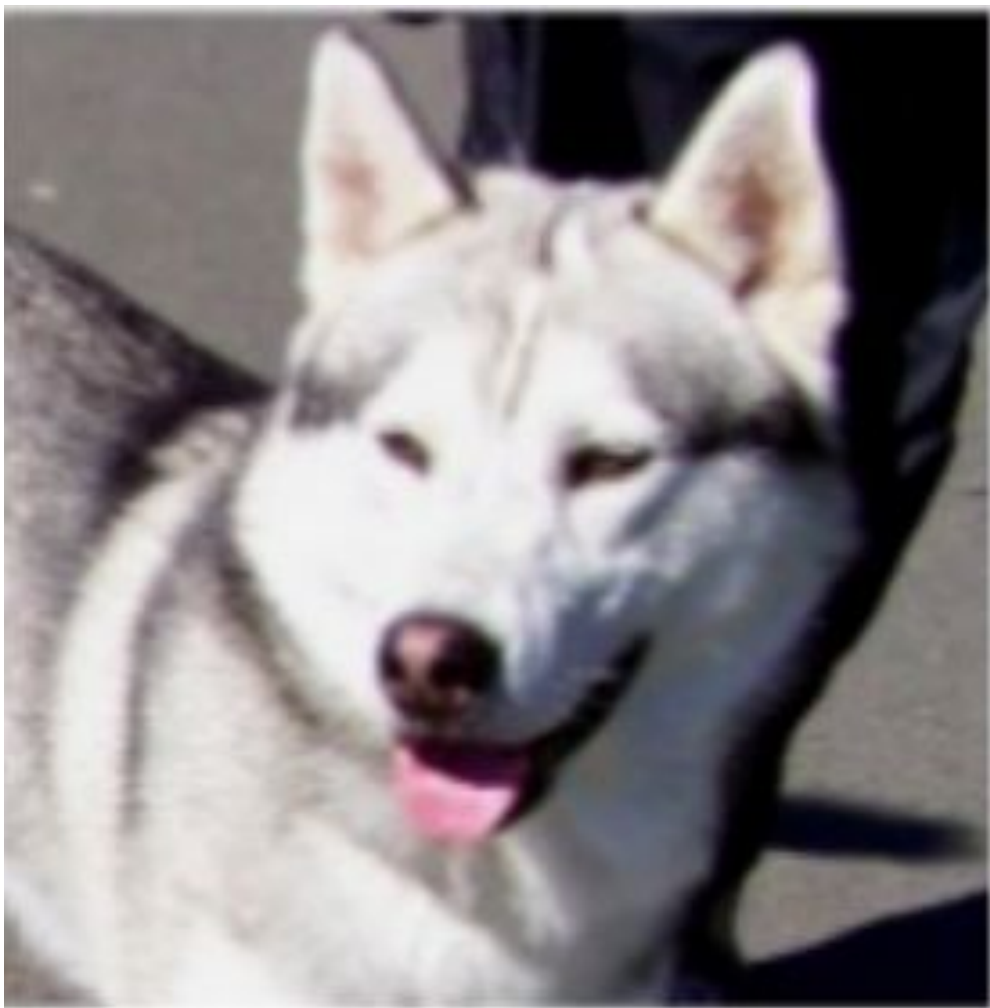


Answer: husky

Model: wolf







Answer: husky

Model: husky



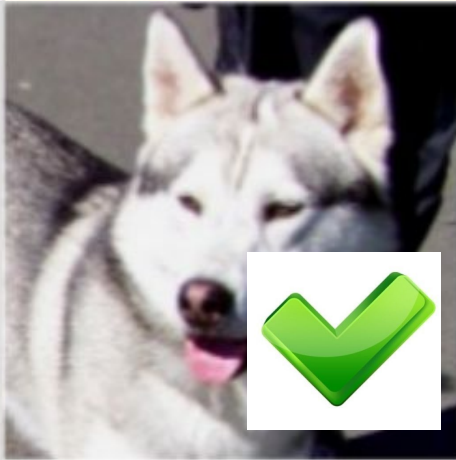
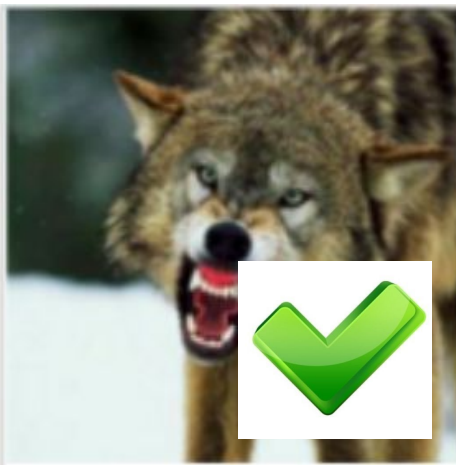


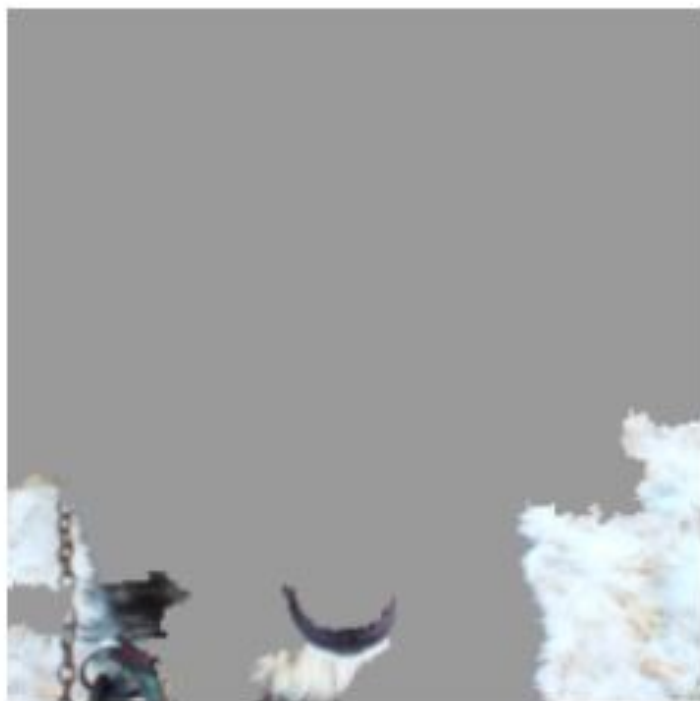
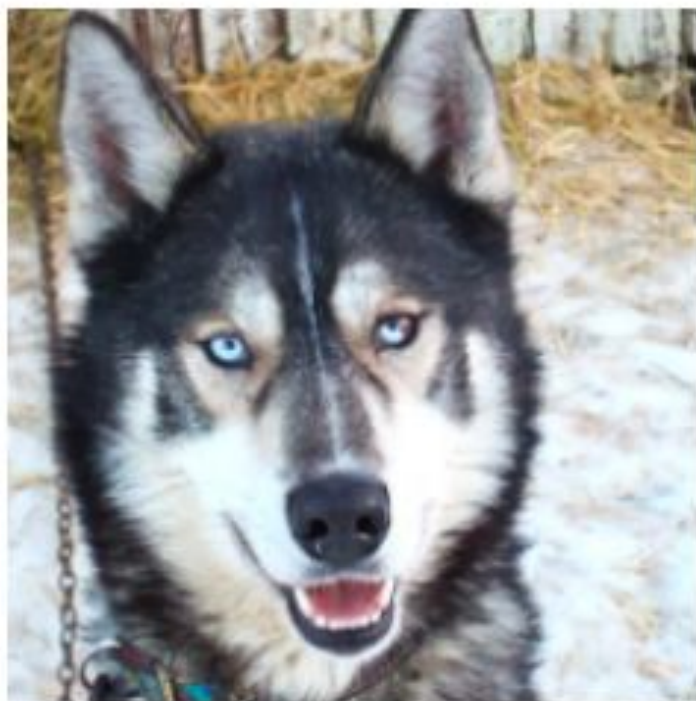


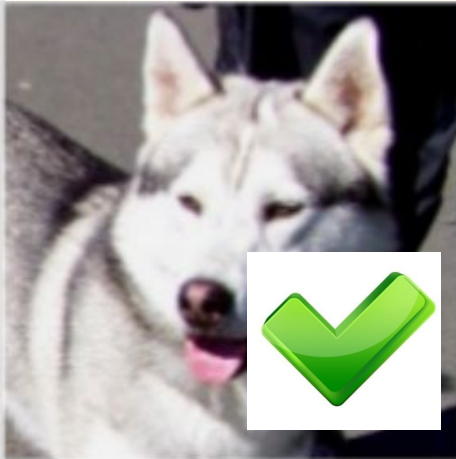
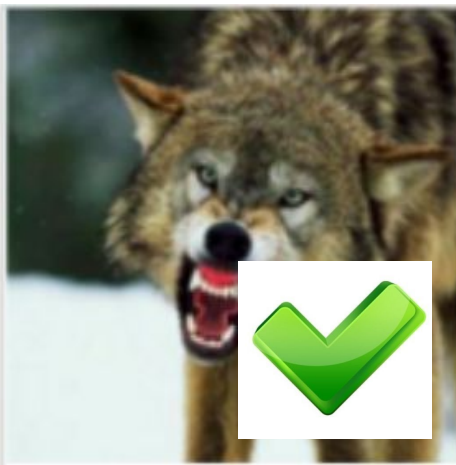
Model: **wolf**

Answer: **wolf**







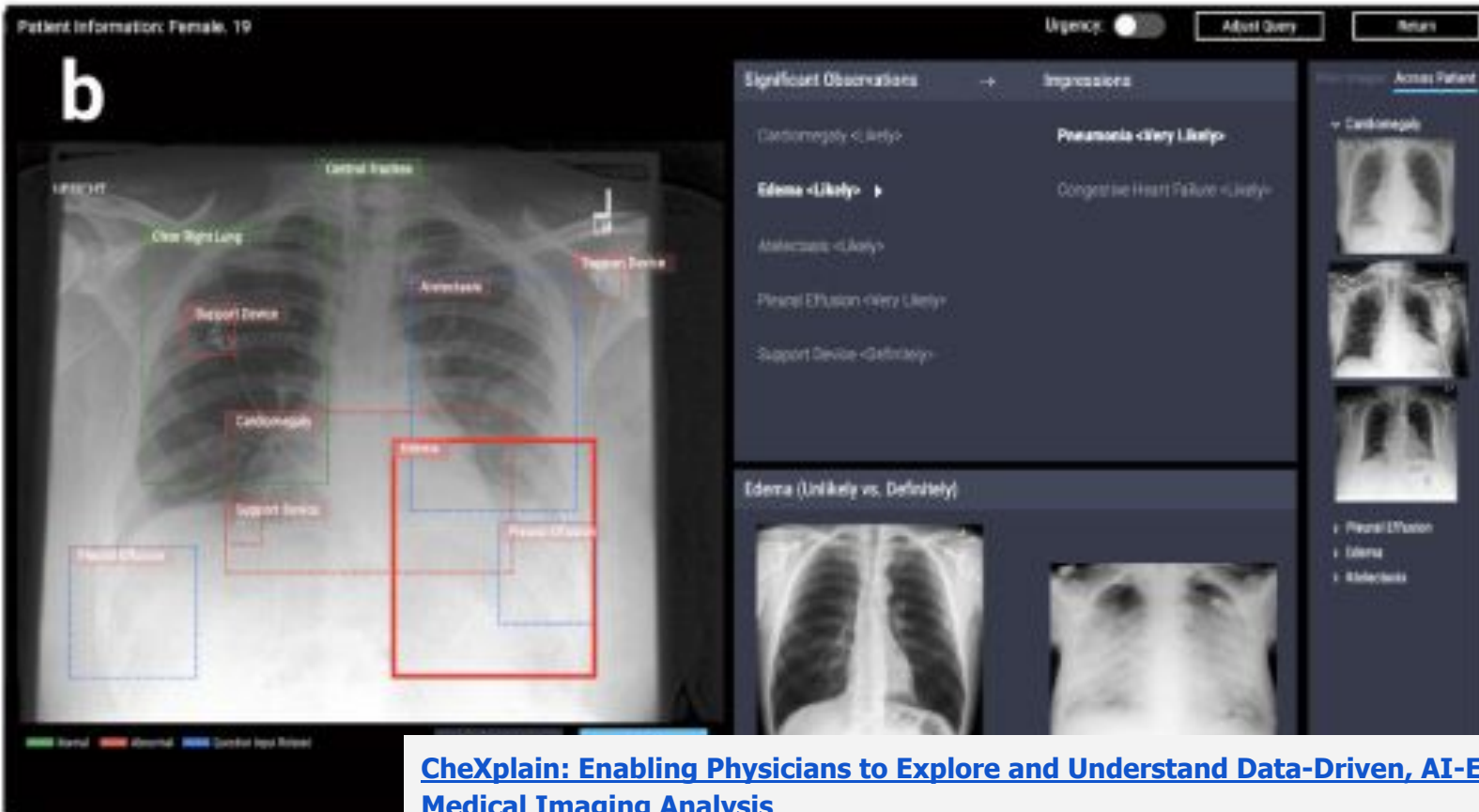


CheXplain: Image Predictions Explained (CHI 2020)

Patient Information: Female, 19

Urgency: ☐ ☒ Adjust Query Return

b



The interface displays a chest X-ray with several annotations: 'Clear Right Lung' (green), 'Central Trachea' (green), 'Support Device' (red), 'Aortic Arch' (blue), 'Edema' (red), 'Pneumothorax' (red), 'Pleural Effusion' (blue), and 'Support Device' (red). A legend at the bottom indicates: Green = Normal, Red = Abnormal, Blue = Question Input Related.


Significant Observations

- Cardiomegaly <Likely>
- Edema <Likely>
- Atelectasis <Likely>
- Pleural Effusion <Very Likely>
- Support Device <Definitely>

Impressions

- Pneumonia <Very Likely>
- Congestive Heart Failure <Likely>

Edema (Unlikely vs. Definitely)

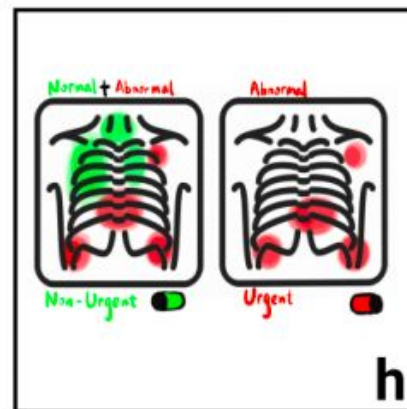
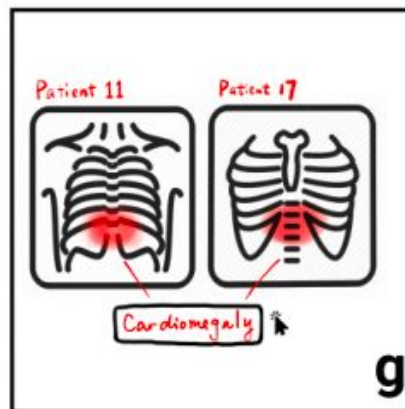
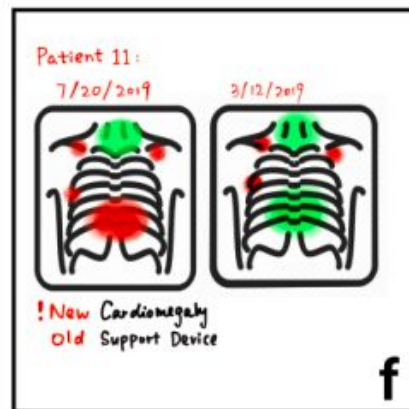
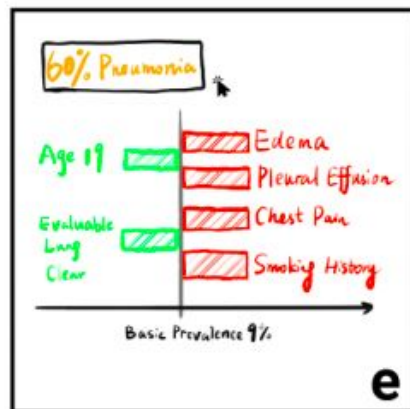
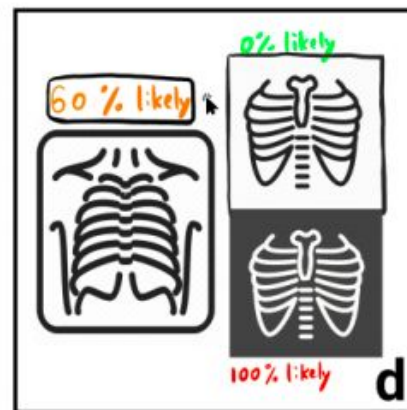
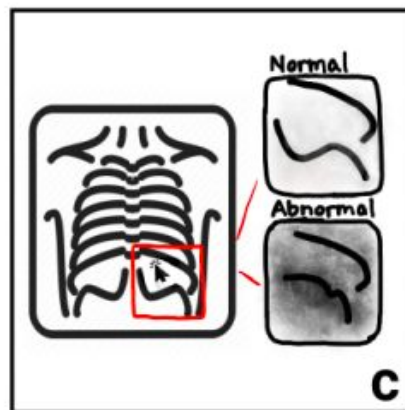
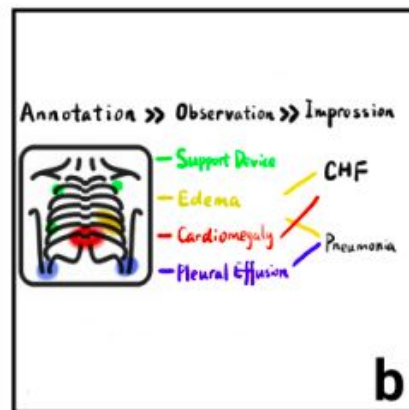
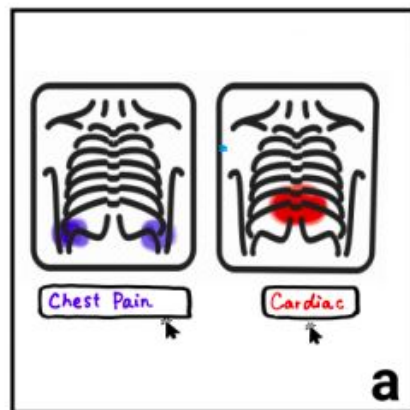


Across Patient

- Cardiomegaly
- Pleural Effusion
- Edema
- Atelectasis

CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis

CheXplain: Image Predictions Explained (CHI 2020)



Interpret the uninterpretable

Interpreted Machine Learning

What are interpretations?

Interpretability - the degree to which a person is able to understand the reasons for a decision

The purpose of interpretation is a description of the internal logic of the system

Why are interpretations needed?

- Rationale for Decision Making
- Detecting Bias in Models
- Fulfillment of requirements for “transparency” (GDPR)

Who needs interpretations?

- Machine Learning Model Developers
- Decision makers (doctors, managers)
- Consumers of AI products

New direction

- Actively developing: XAI (Explainable AI), Interpretable ML
- Are parts of the major AI conferences:
 - Operationalizing Human-Centered Perspectives in Explainable AI (CHI 2021)
 - Human, ML & AI (CHI 2021)
 - Counterfactual Explanations in Explainable AI: A Tutorial (KDD 2021)
 - Privacy and Ethics -> Interpretability and Explainability (KDD 2021)
 - AI/ML & seeing through the black box (CHI 2020)
 - Coping with AI: not agAI! (CHI 2020)
 - Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities (NeurIPS 2020)
 - Algorithmic Fairness through the Lens of Causality and Interpretability (NeurIPS 2020)
 - Human-Centered Explainability for Healthcare (KDD 2020)
 - Interpretable Models (KDD 2020)
 - Explainable Models for Healthcare AI (KDD 2018),
 - Interpretable ML Symposium (NIPS 2017),
 - Explainable AI (IJCAI 2018),
 - Explainable artificial intelligence (XAI): Why, when, and how? (Strata Data Conference 2018)

About

- What problems arise (ethics, society)
- Solution Approaches: “Tweaked Algorithms”, Using Only Interpretable Models
- “Superstructure” of interpretation on the black box

Approaches

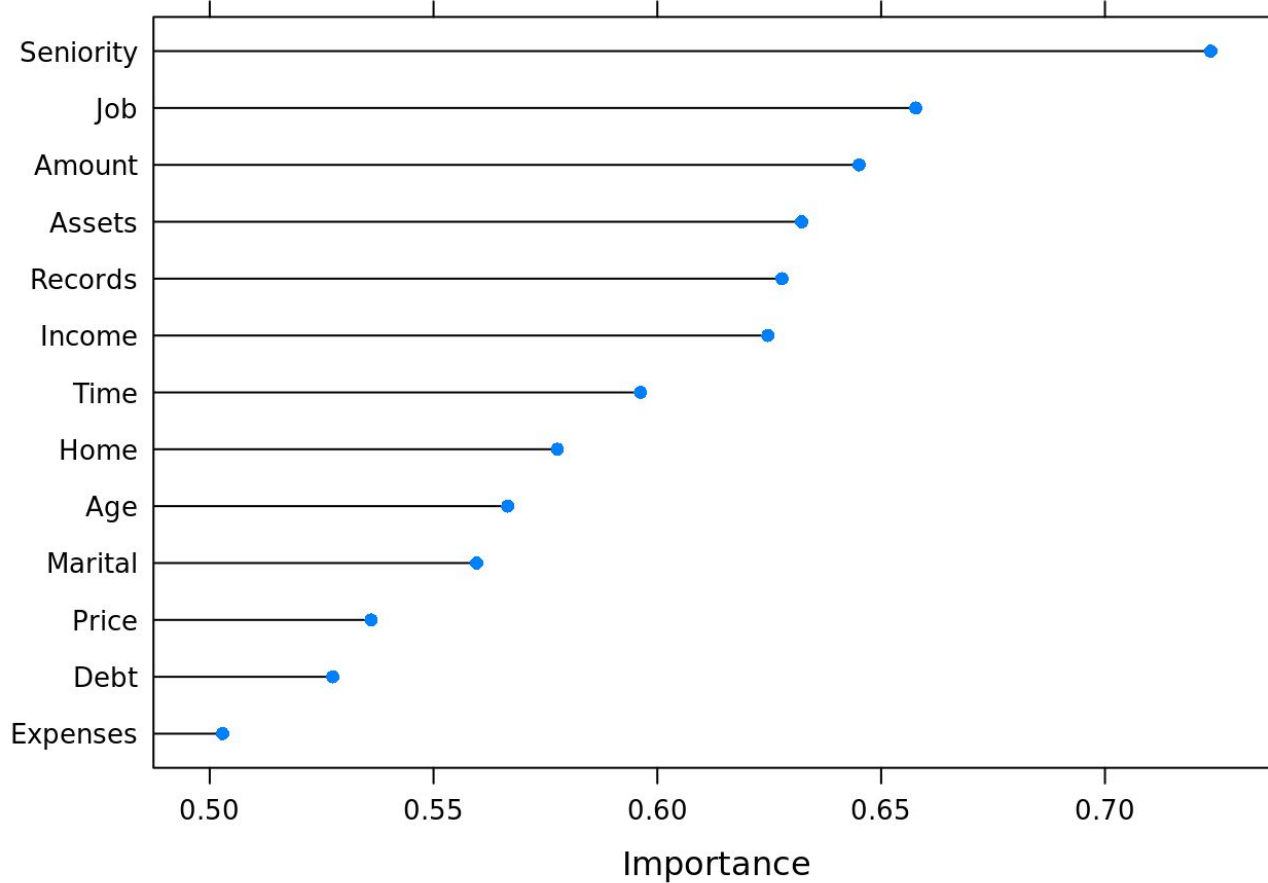
Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model (Global)	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	How
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	How , Why, Why not, What if
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	How , Why, Why not, What if
Explain a prediction (Local)	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	Why
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	Why , How to still be this
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	What if , How to be that, How to still be this
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	Why , Why not , How to be that
Example based	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	Why , How to still be this
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	Why , Why not , How to be that

Table 1. Taxonomy of XAI methods mapping to user question types. Questions in bold are the primary ones that the XAI method addresses. Questions in regular font are ones that only a subset of cases the XAI method can address. For example, while a global decision tree approximation can potentially answer *Why*, *Why not*, and *What if* questions for individual instances [58], the approximation may not cover certain instances.

Global interpretation

1. Assessment of the significance of features (importance), highlighting significant
 - model dependent
 - model-independent (changing the quality of the prediction)
2. Investigation of change in prediction when some variable changes (ICE plots)

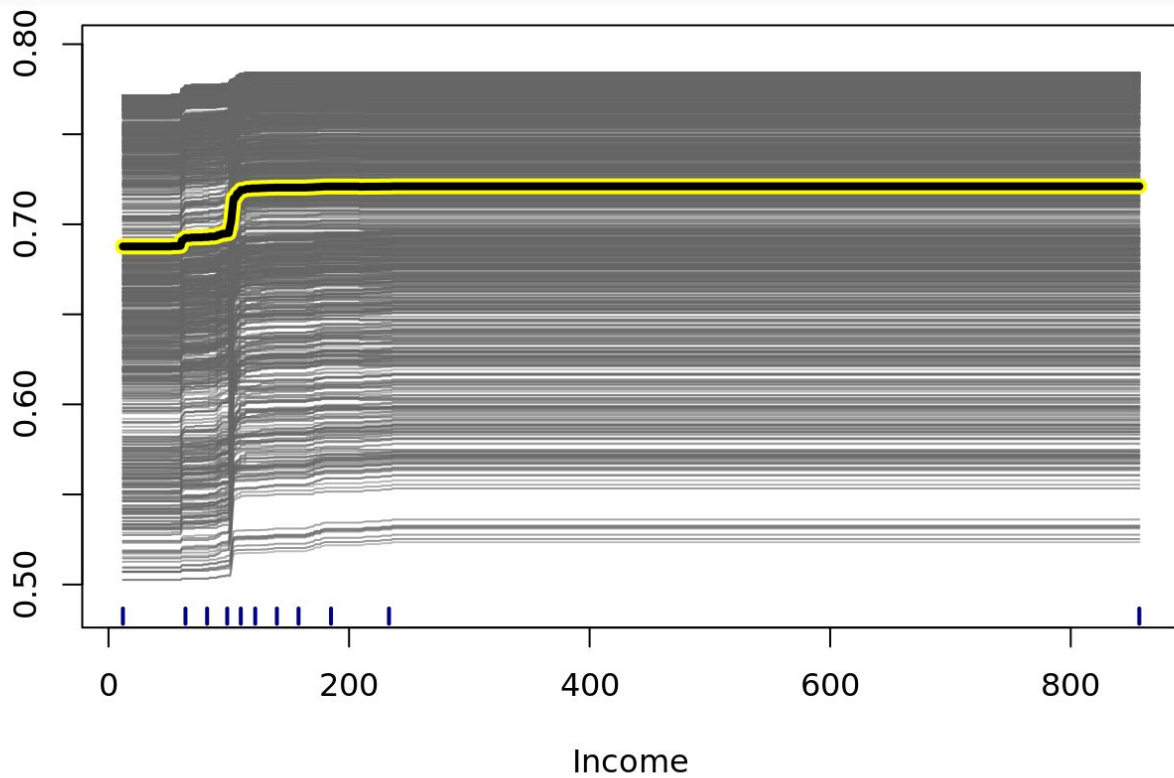
Feature Importance



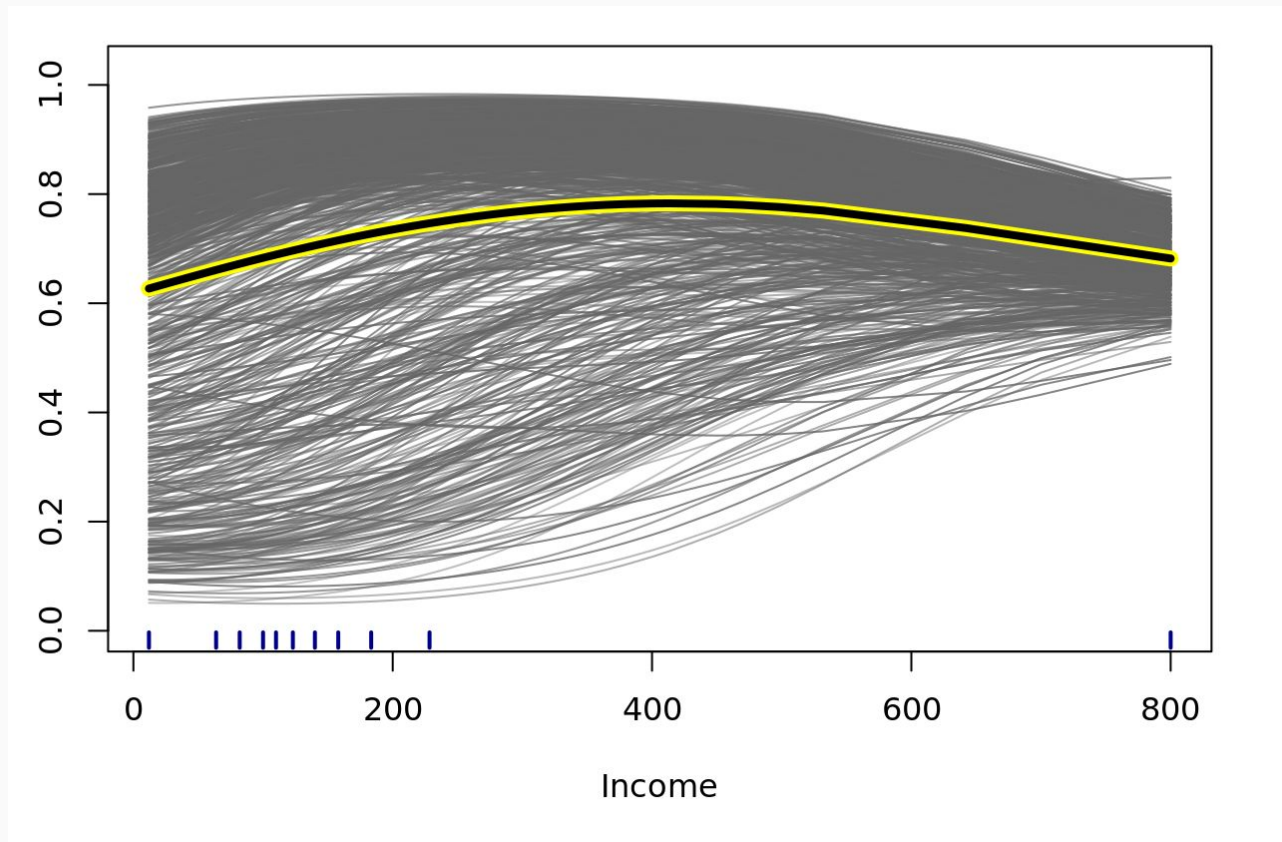
Visual exploration of variables

ICE (Individual Conditional Expectation) graphs - show the change in prediction when one of the variables changes

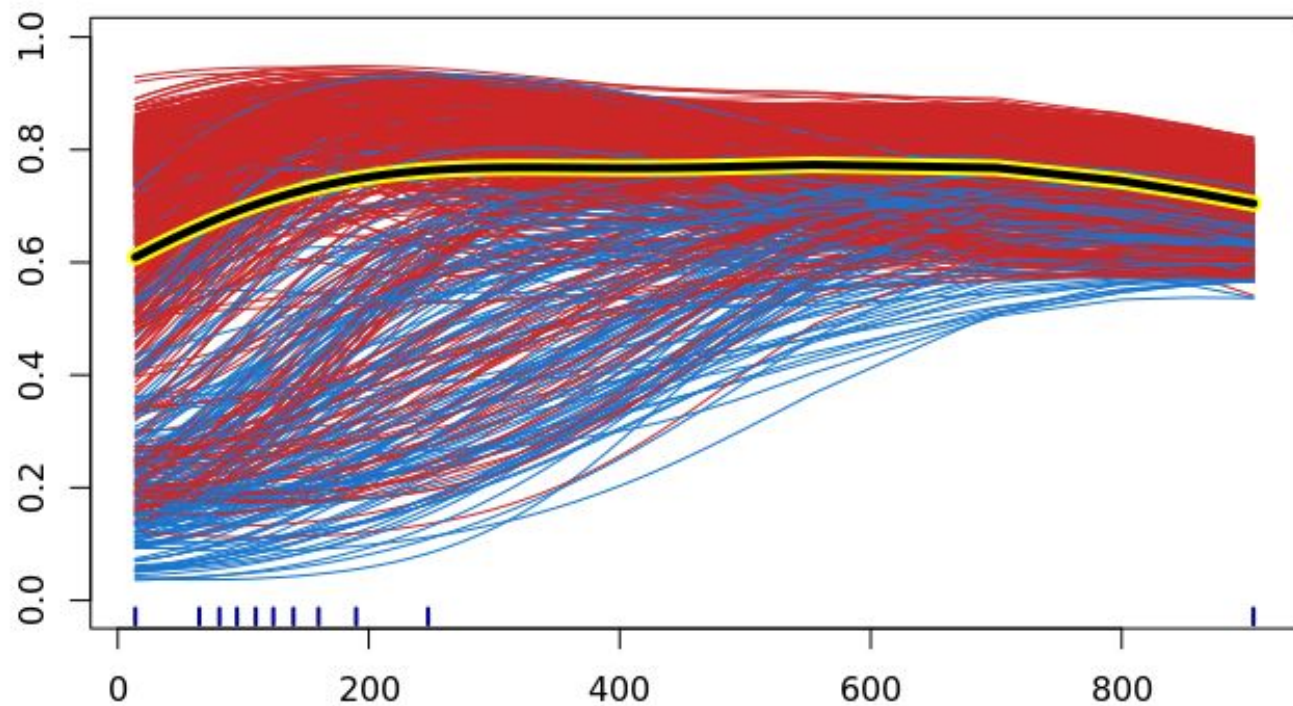
ICE graph: income (gradient boosting)



ICE Chart: Income (SVM)

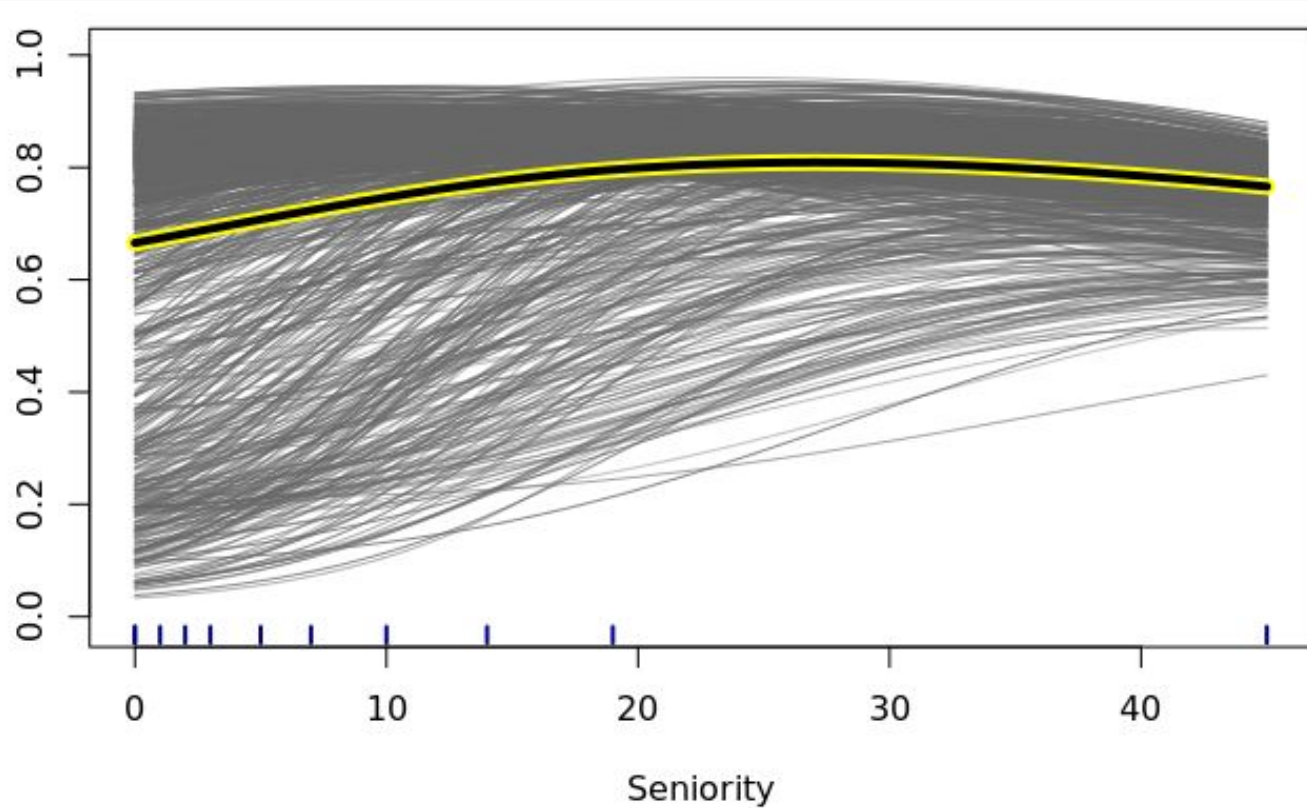


ICE Chart: Income and Debt (SVM)



Income colored by Records

ICE Graph: Work Experience (SVM)

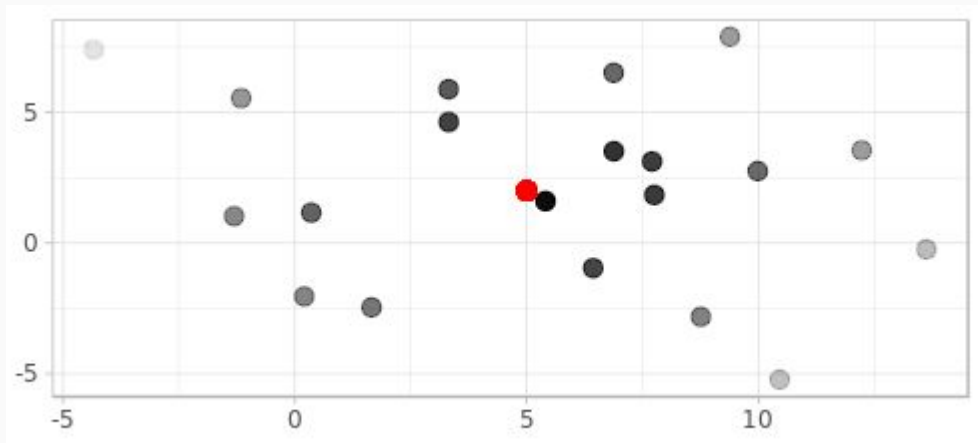


Local interpretation

- The global interpretation turns out to be a fairly generalized "averaged" over all data
- We want to investigate a specific example, to understand what factors led to the fact that the client has a bad credit status (and how to change it)
- To answer such questions, there are local interpretation algorithms.

Local interpretation: LIME

Let's look at the neighborhood of our example. In this way we will change the predictor values randomly and study how this affects the result



LIME: neighborhood approach

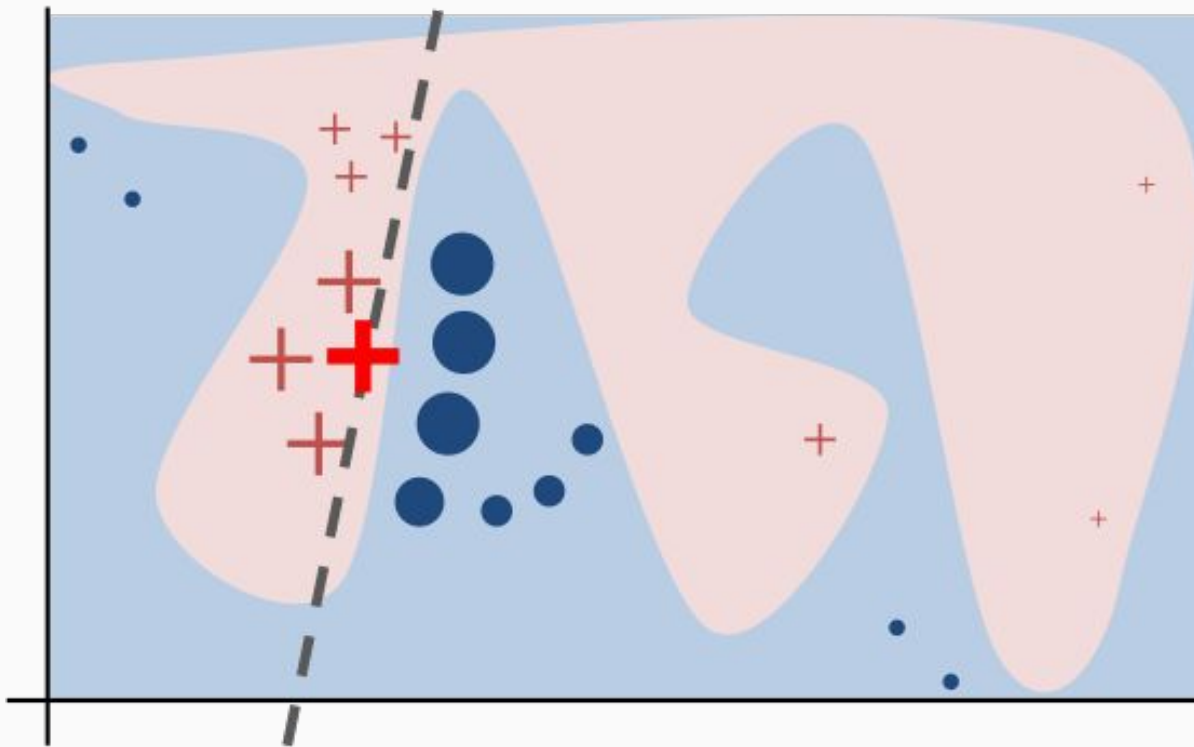
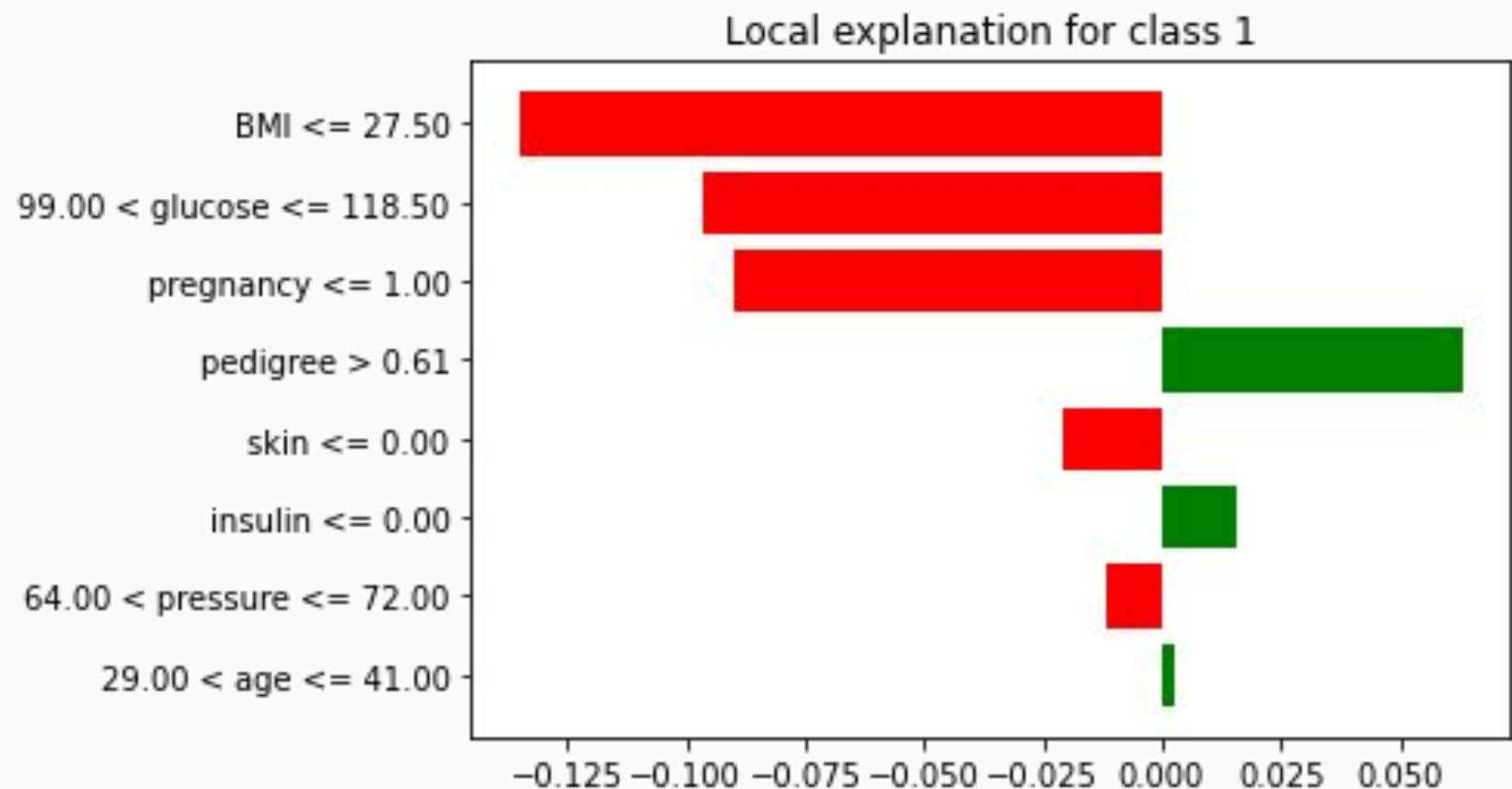


Figure from author's [Github](#)

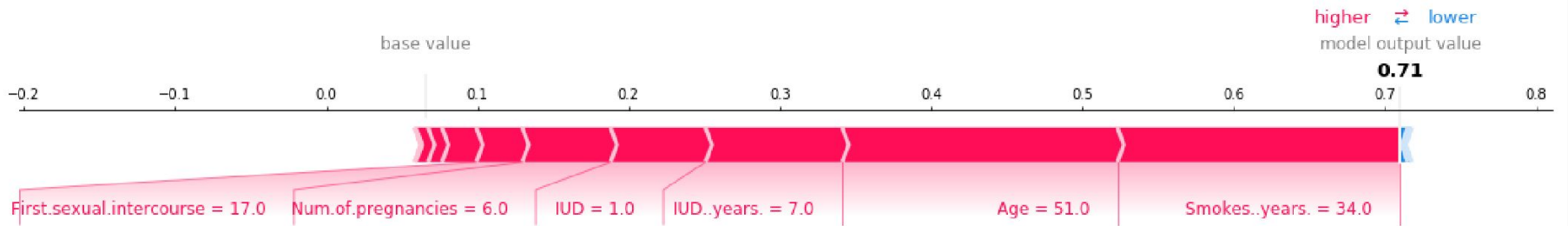
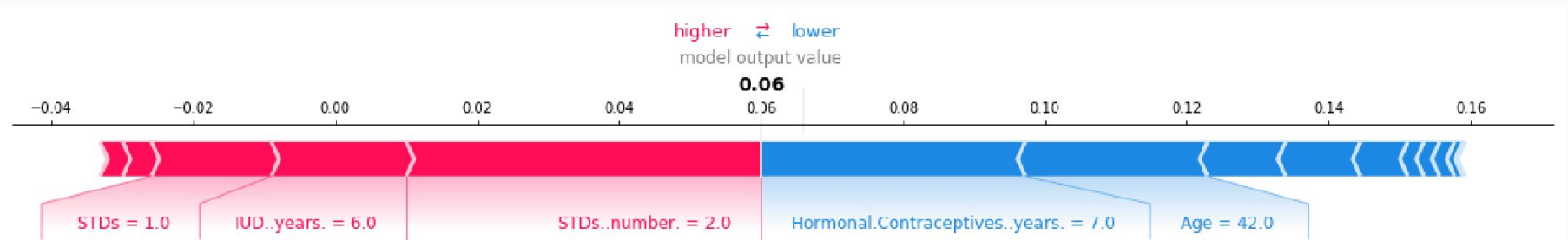
LIME: algorithm

- 1) generate artificial data around the example
- 2) we obtain a prediction for them according to our model
- 3) use some interpretable model (tree/regression) to relate 1 and 2.
Important: we weight the data - those that are closer to the original example (according to some proximity metric) weigh more
- 4) interpret the results (valid only for the neighborhood of the example)

LIME: example, prediction -- no diabetes



Not only LIME



Example from the book [Interpretable Machine Learning](#)

Not only LIME

- Accumulated Local Effects (ALE) Plot
- SHAP (SHapley Additive exPlanations)
- Anchors (by LIME's creators, but the results in form of IF-THEN rules)
- Counter-examples ("if I change the feature X, then the prediction will be reversed")
- Similar examples
- Influential Observations
- ...

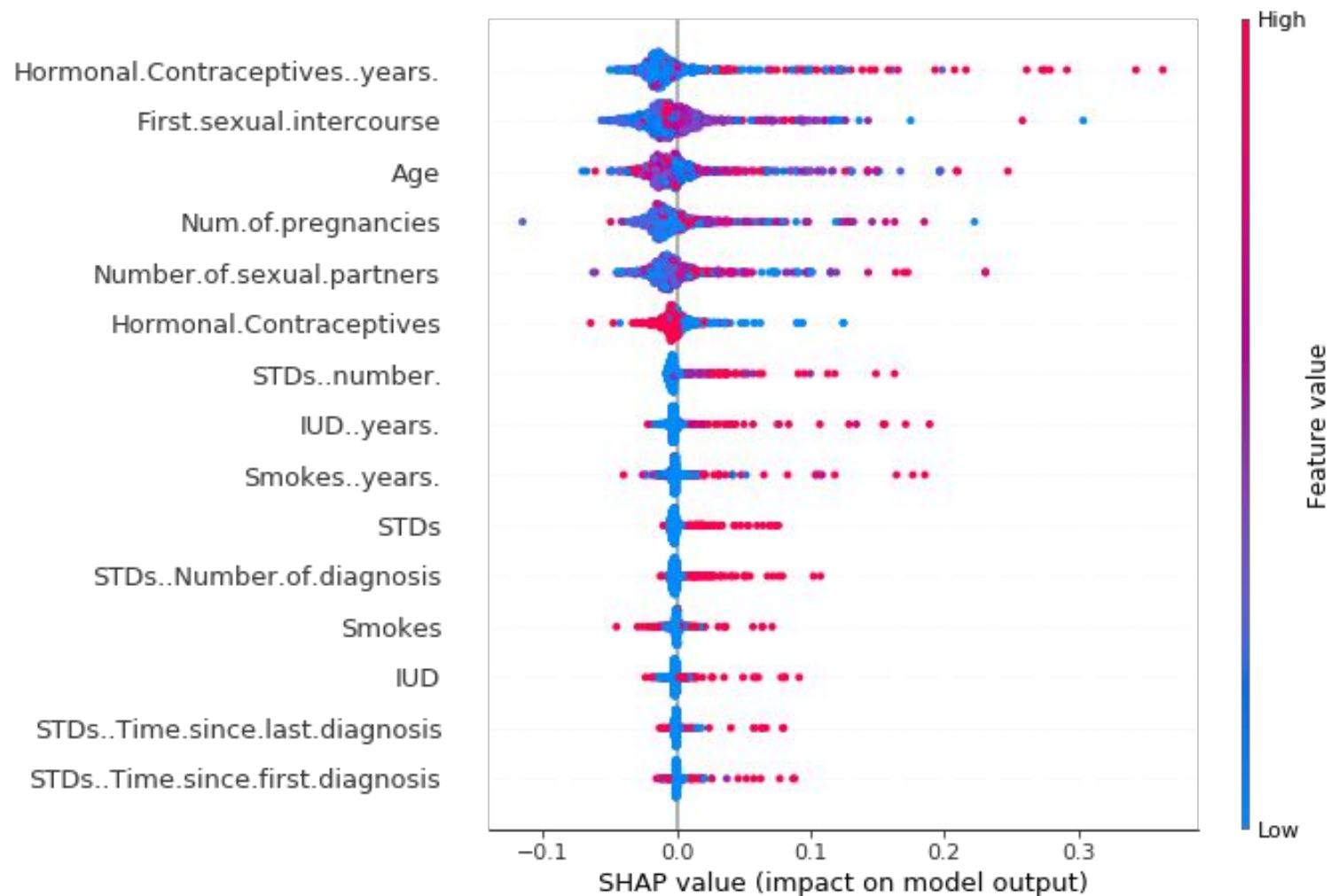
win-win: everything is fine

- build a complex multilevel ensemble with high accuracy (or other target metric)
- we build an interpretation on top to understand what affects what
- victory!

Why do we need simple models then?
And why do we need all sorts of diagnostics, etc.?

What can go wrong

- these methods interpret **the model, not the reality** (different model - different conclusions)
- approximation of approximation (“gartic phone”)
- models can be bad (low quality) and model interpretation models can be bad too
- incorrect inferences from the interpretation model
 - local methods give a local interpretation, one cannot draw general conclusions
 - there is no understanding of what exactly this or method / visualization shows



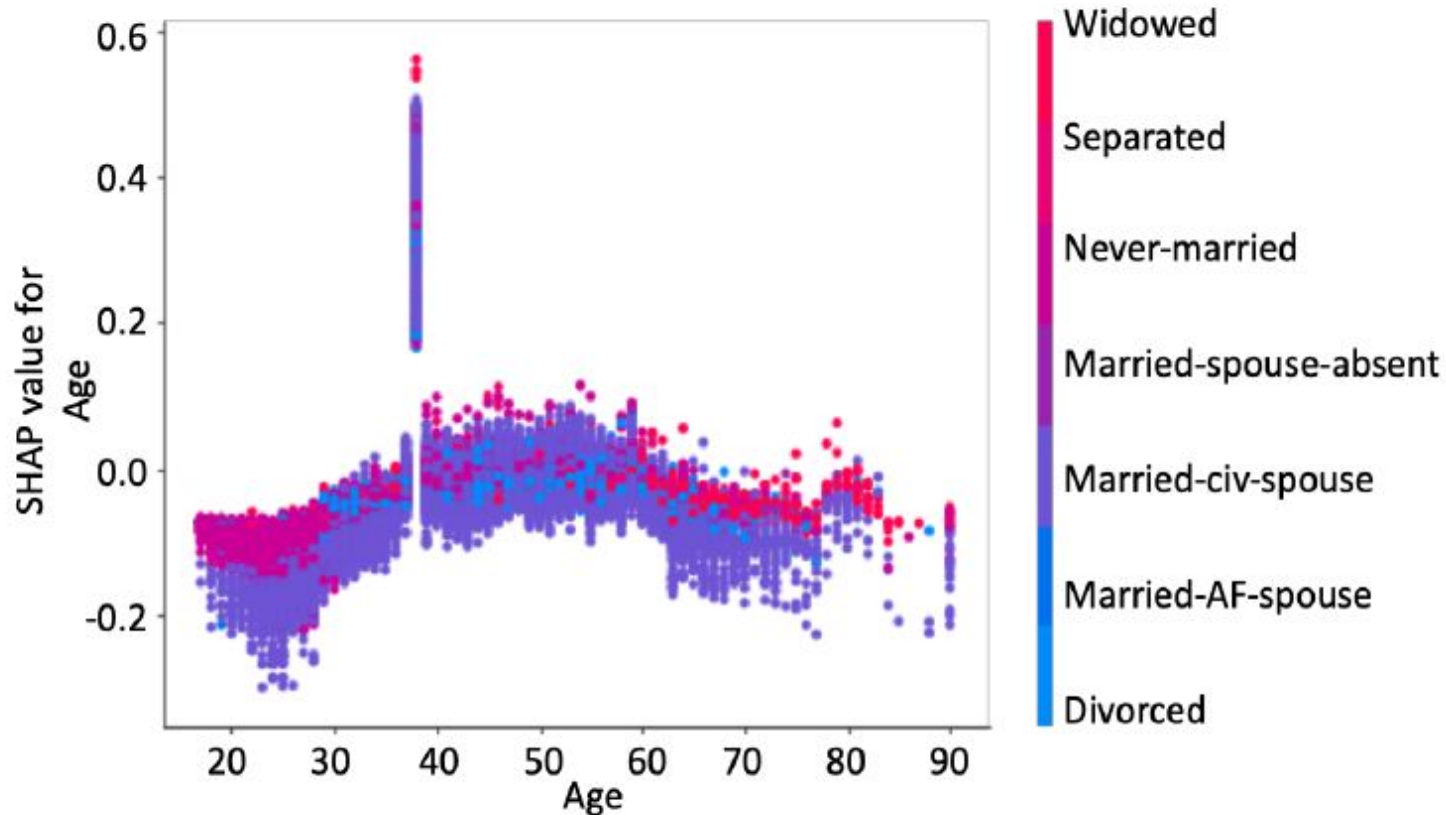
From [Interpretable Machine Learning](#)

Using iML

According to Kaur et al. ([Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning](#), CHI 2020), analysts use tools differently than intended by their developers

- tend to trust such models too much, without even understanding how they work, if the results are presented in a “scientific” format, with visualization and / or supported by links to publications
- tend to rationalize deviations - if there is a plausible explanation for an observation, then it is not considered a bias or an erroneous prediction
- people with more experience in ML interpret visualizations more correctly, but they are also more critical of them

An example visualization in a study



Useful links

- [Interpretable Machine Learning](#) by Christoph Molnar
- [Hands-on Machine Learning Model Interpretation](#) (examples в python)
- [lime](#)
- [Local Interpretable Model-Agnostic Explanations \(LIME\): An Introduction](#)
- <https://towardsdatascience.com/explainable-ai-xai-a-guide-to-7-packages-in-python-to-explain-your-models-932967f0634b> — somewhat complete overview of xAI with examples