



# EXPLAINABLE AI OPENING THE MACHINE LEARNING BLACK BOX

---

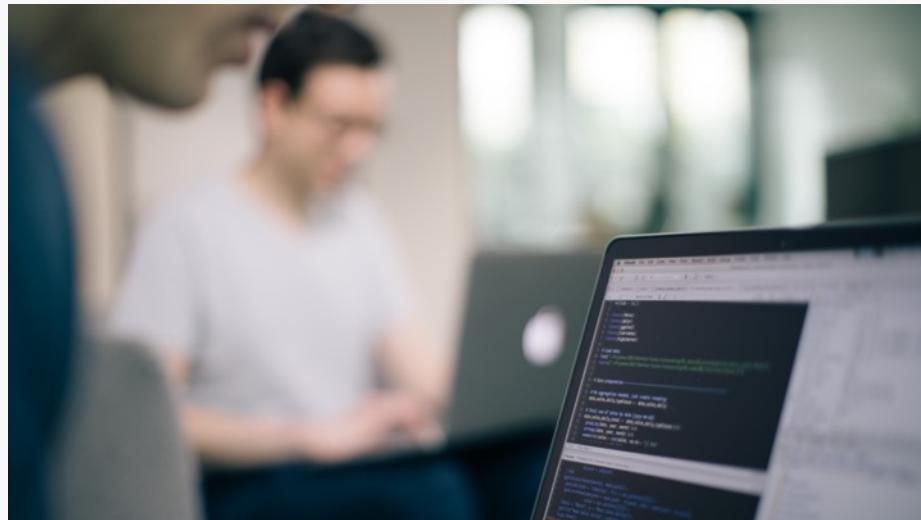
**STATWORX GmbH**  
Fabian Müller, Head of Data Science

Frankfurt, 25.04.2019

# AGENDA

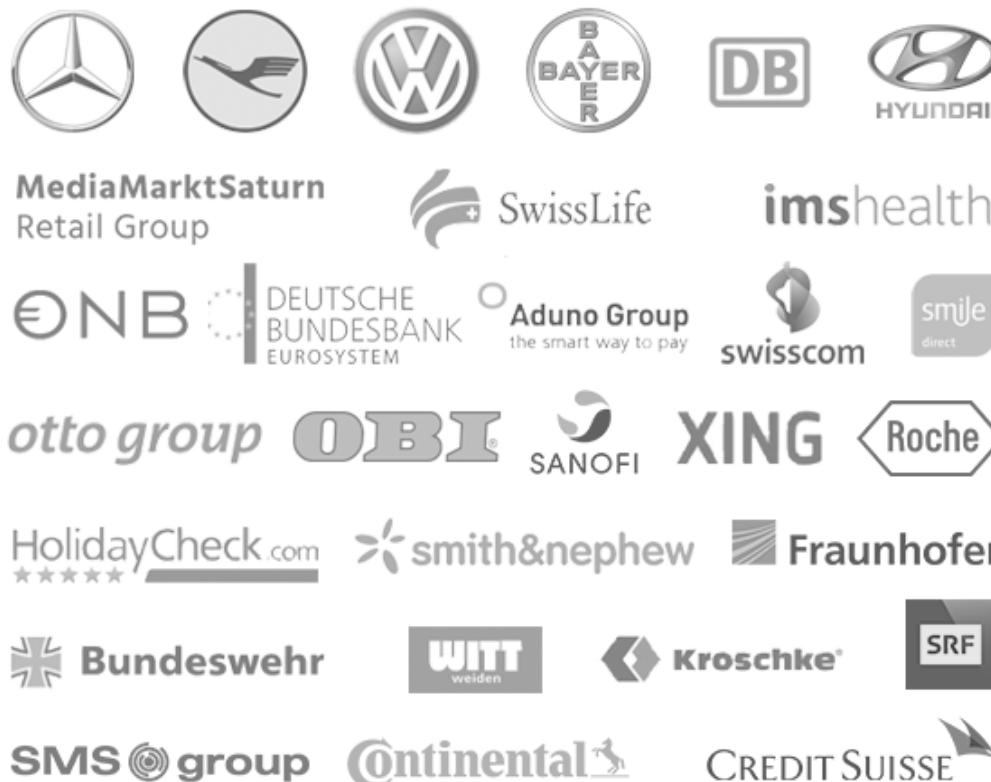
Key questions I am going to answer in this talk

STATWORX



- 1 WHO IS STATWORX? 😊
- 2 WHAT IS MODEL INTERPRETABILITY?
- 3 WHY DO WE NEED MODEL INTERPRETABILITY?
- 4 WHAT METHODS ARE AVAILABLE?
- 5 EXAMPLE IMPLEMENTATIONS USING CUTE PETS!

### SOME OF OUR CUSTOMERS

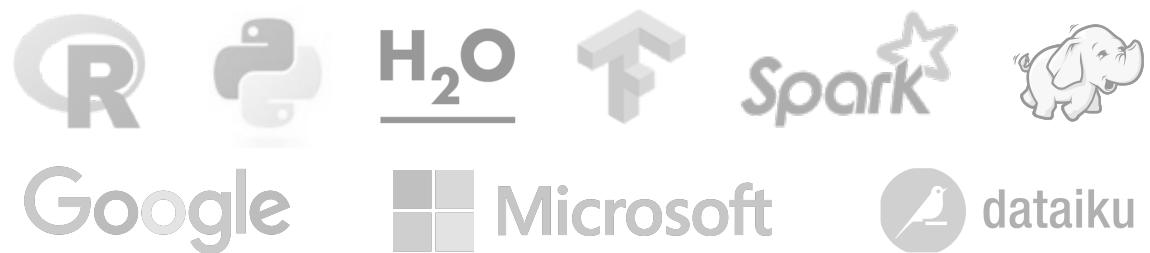


### COMPANY PROFILE

STATWORX is a consulting company for data science, machine learning, and AI located in Frankfurt, Vienna and Zurich. We support our customers in the development and implementation of data science and machine learning projects as well as data driven products.

2011	3	35+
FOUNDED	OFFICES	EMPLOYEES
200+	50+	1000+
DATA SCIENCE PROJECTS	INDUSTRY CUSTOMERS	DATA ACADEMY PARTICIPANTS

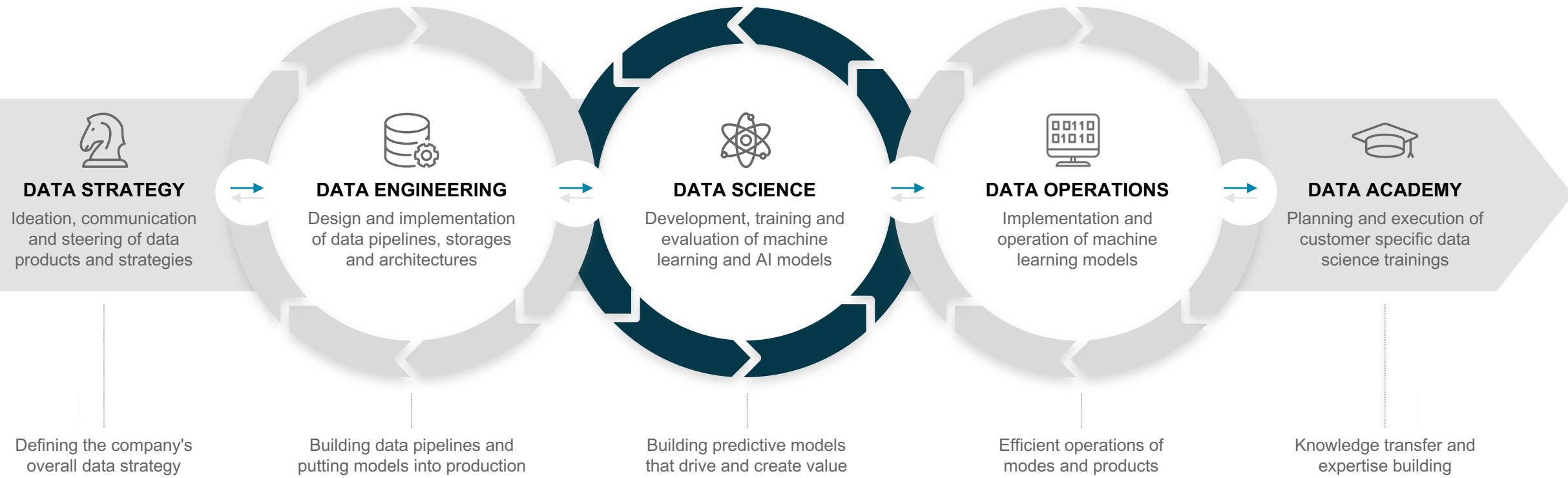
### TOOL STACK & PARTNER



# END-2-END DATA CONSULTING

STATWORX

We support our customers along the whole process of data driven decision making



# MOTIVATION

What makes Danny more likely to get adopted compared to Beauty and Kat?

Danny



$p = 73\%$

Kat



$p = 20\%$

Beauty



$p = 57\%$

STATWORX

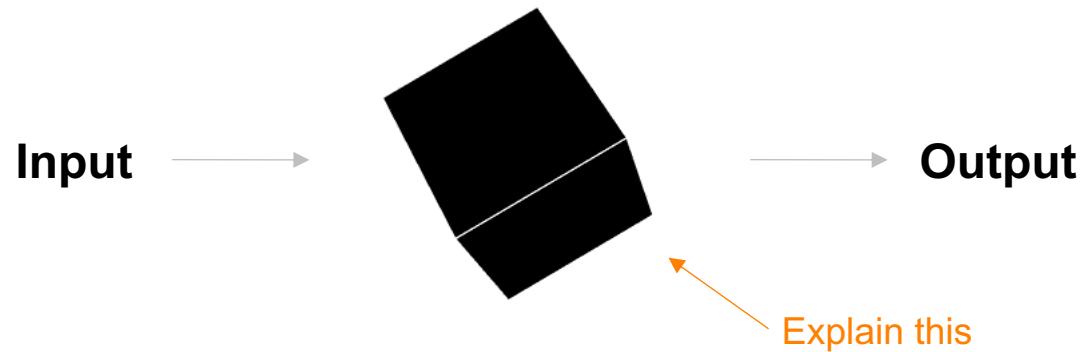
The screenshot shows the Kaggle competition page for 'PetFinder.my Adoption Prediction'. The title 'PetFinder.my Adoption Prediction' is displayed, along with the subtext 'How cute is that doggy in the shelter?'. The page includes navigation links for Overview, Data, Kernels, Discussion, Leaderboard, Rules, and Team. The Overview section contains a description of the competition, stating that millions of stray animals suffer on the streets or in shelters every day, and that adopting them can save lives and create families. It also mentions that PetFinder.my has been Malaysia's leading animal welfare organization since 2008. The page highlights that adoption rates are correlated with photo characteristics, and that PetFinder is working on AI tools to improve these qualities. The competition has 1,775 teams and 12 days left until merger day.

# DEFINITION

What model interpretability is about

***“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 NOT about understanding all bits and bytes of the model for all data points (we cannot).

**It's about knowing enough for your downstream tasks.**

# IMPORTANCE OF INTERPRETABILITY

STATWORX

Some (certainly unrepresentative) thoughts from Google about model interpretability

***“People can’t explain how they work, for most of the things they do.”***

Geoffrey Hinton, Google Fellow

***“Don’t limit your solutions to what the simple human mind can wrap itself around.”***

Cassie Kozyrkov, Chief Decision Intelligence Engineer, Google

# IMPORTANCE OF INTERPRETABILITY

STATWORX

Why we should care about model interpretability

## 1. Mismatched objectives and multi-objective trade-offs:

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (simply minimizing the loss does not work).

## 2. Human curiosity

Gain knowledge from data and learn from models (e.g. improve feature engineering by understanding the model).

## 3. Detect bias

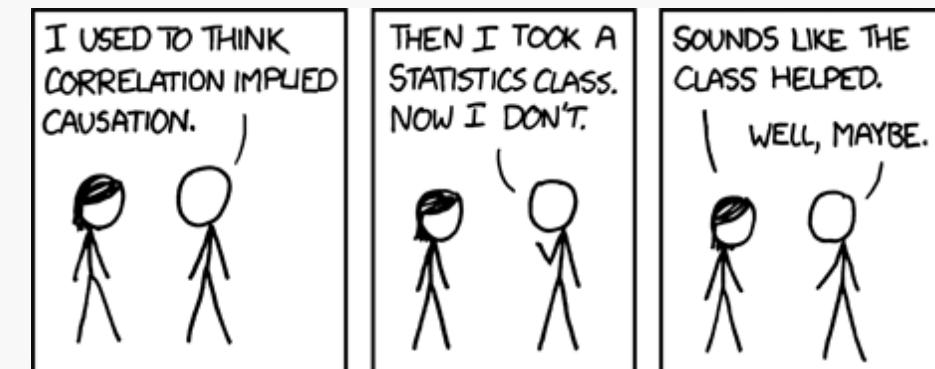
Discrimination against a minority when only minimizing arbitrary loss.

## 4. Social acceptance and trust

It is easier for humans to trust a system that explains its decisions compared to a black box and sometimes insights are more valuable than predictions.

## 5. Debugging

You can not test everything, if anything goes wrong, interpretability can help to find problems quickly, also humans tend to start asking questions when things go wrong.



# IMPORTANCE OF INTERPRETABILITY

STATWORX

Why we should care about model interpretability

## 1. Mismatched objectives and multi-objective trade-offs:

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (simply minimizing the loss does not work).

## 2. Human curiosity

Gain knowledge from data and learn from models (e.g. improve feature engineering by understanding the model).

## 3. Detect bias

Discrimination against a minority when only minimizing arbitrary loss.

## 4. Social acceptance and trust

It is easier for humans to trust a system that explains its decisions compared to a black box and sometimes insights are more valuable than predictions.

## 5. Debugging

You can not test everything, if anything goes wrong, interpretability can help to find problems quickly, also humans tend to start asking questions when things go wrong.



# IMPORTANCE OF INTERPRETABILITY

STATWORX

Why we should care about model interpretability

## 1. Mismatched objectives and multi-objective trade-offs:

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (simply minimizing the loss does not work).

## 2. Human curiosity

Gain knowledge from data and learn from models (e.g. improve feature engineering by understanding the model).

## 3. Detect bias

Discrimination against a minority when only minimizing arbitrary loss.

## 4. Social acceptance and trust

It is easier for humans to trust a system that explains its decisions compared to a black box and sometimes insights are more valuable than predictions.

## 5. Debugging

You can not test everything, if anything goes wrong, interpretability can help to find problems quickly, also humans tend to start asking questions when things go wrong.



# IMPORTANCE OF INTERPRETABILITY

STATWORX

Why we should care about model interpretability

## 1. Mismatched objectives and multi-objective trade-offs:

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (simply minimizing the loss does not work).

## 2. Human curiosity

Gain knowledge from data and learn from models (e.g. improve feature engineering by understanding the model).

## 3. Detect bias

Discrimination against a minority when only minimizing arbitrary loss.

## 4. Social acceptance and trust

It is easier for humans to trust a system that explains its decisions compared to a black box and sometimes insights are more valuable than predictions.

## 5. Debugging

You can not test everything, if anything goes wrong, interpretability can help to find problems quickly, also humans tend to start asking questions when things go wrong.



# IMPORTANCE OF INTERPRETABILITY

STATWORX

Why we should care about model interpretability

## 1. Mismatched objectives and multi-objective trade-offs:

The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (simply minimizing the loss does not work).

## 2. Human curiosity

Gain knowledge from data and learn from models (e.g. improve feature engineering by understanding the model).

## 3. Detect bias

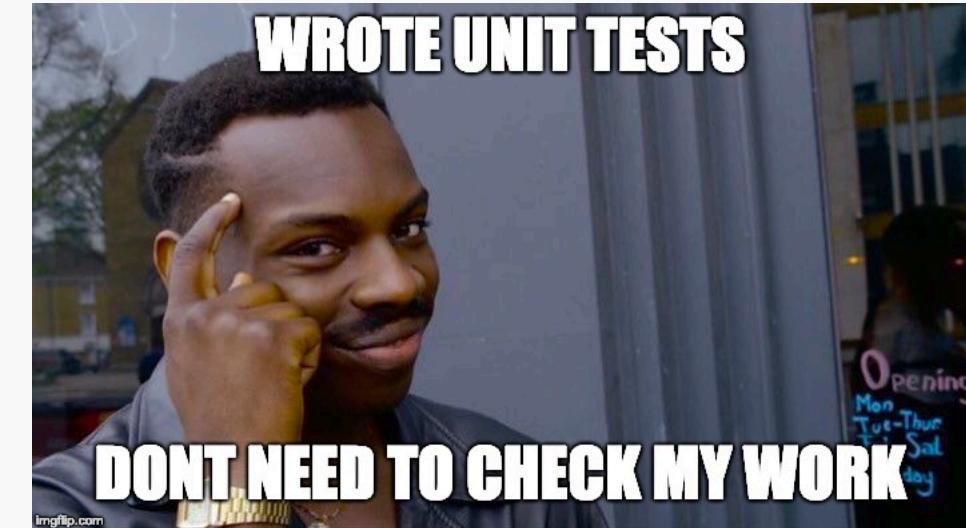
Discrimination against a minority when only minimizing arbitrary loss.

## 4. Social acceptance and trust

It is easier for humans to trust a system that explains its decisions compared to a black box and sometimes insights are more valuable than predictions.

## 5. Debugging

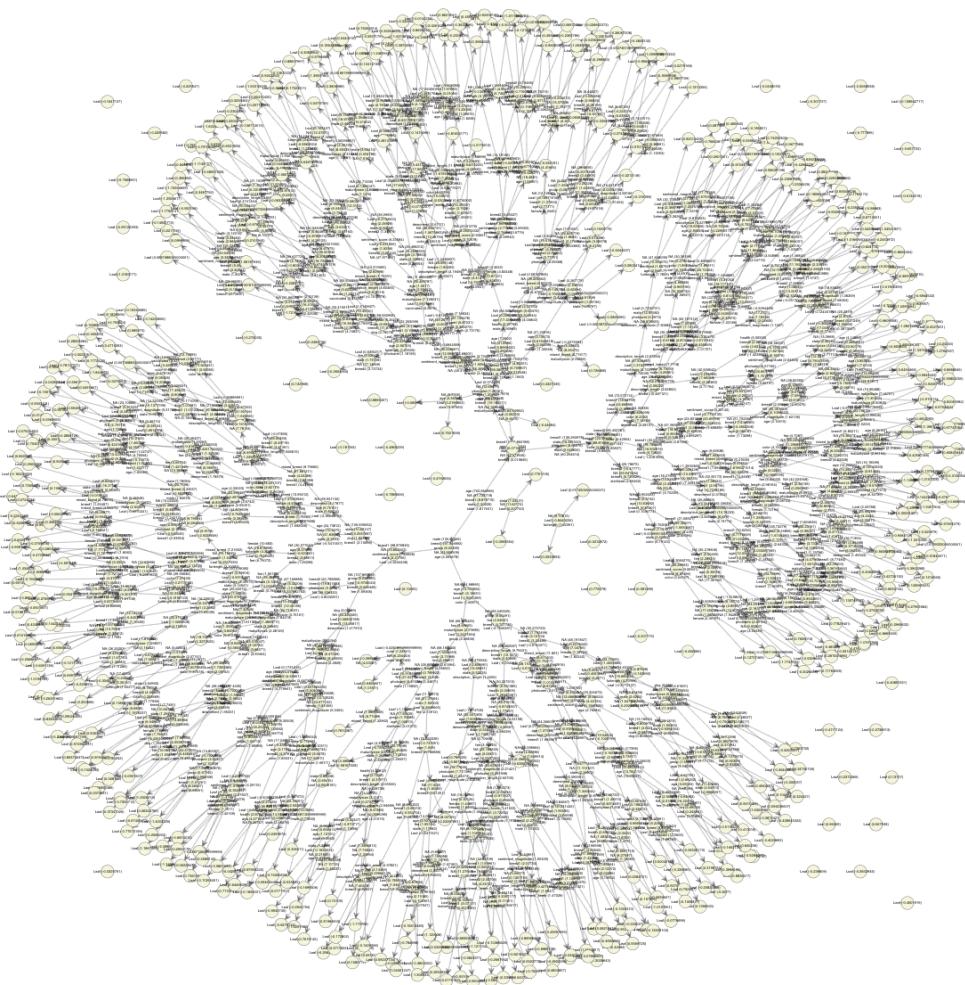
You can not test everything, if anything goes wrong, interpretability can help to find problems quickly, also humans tend to start asking questions when things go wrong.



# INTERPRETATION IS NO FREE LUNCH

STATWORX

Since interpretation is important, let's take a look at the model... ok shit!



dmlc  
**XGBoost**

# INTERPRETATION IS NO FREE LUNCH

STATWORX

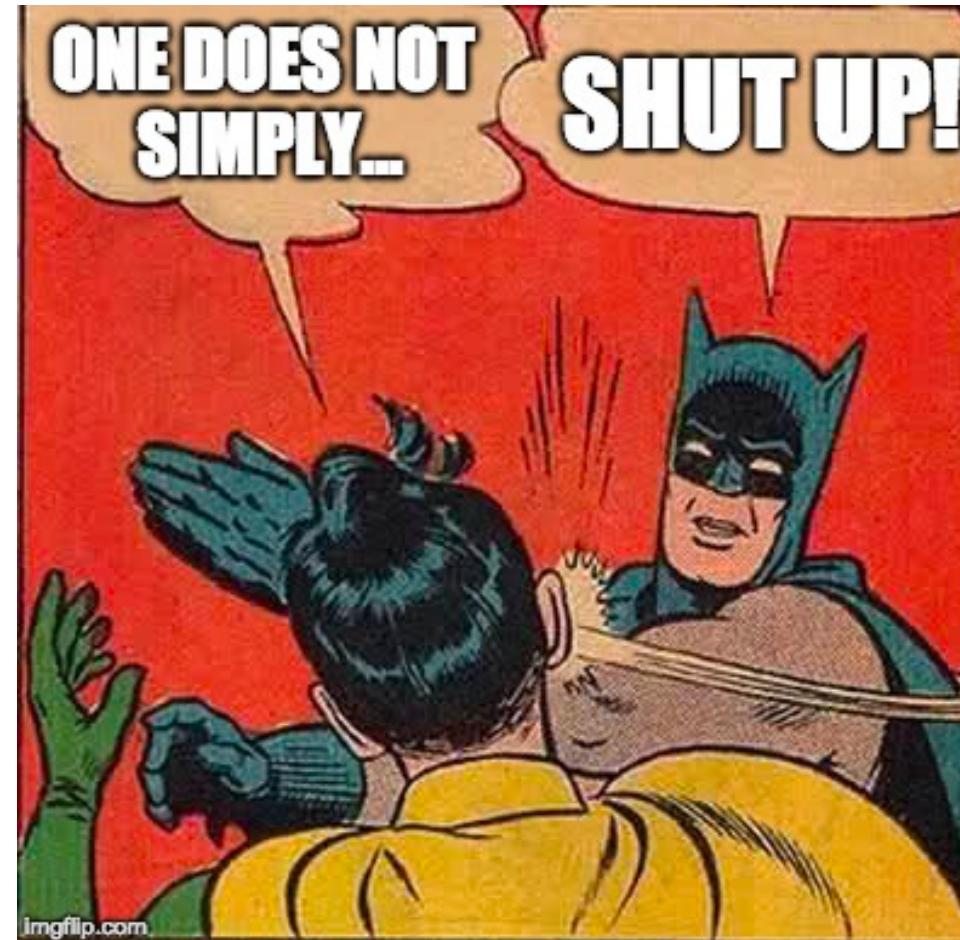
One does not simply interpret ML models, or does one?



# INTERPRETATION IS NO FREE LUNCH

STATWORX

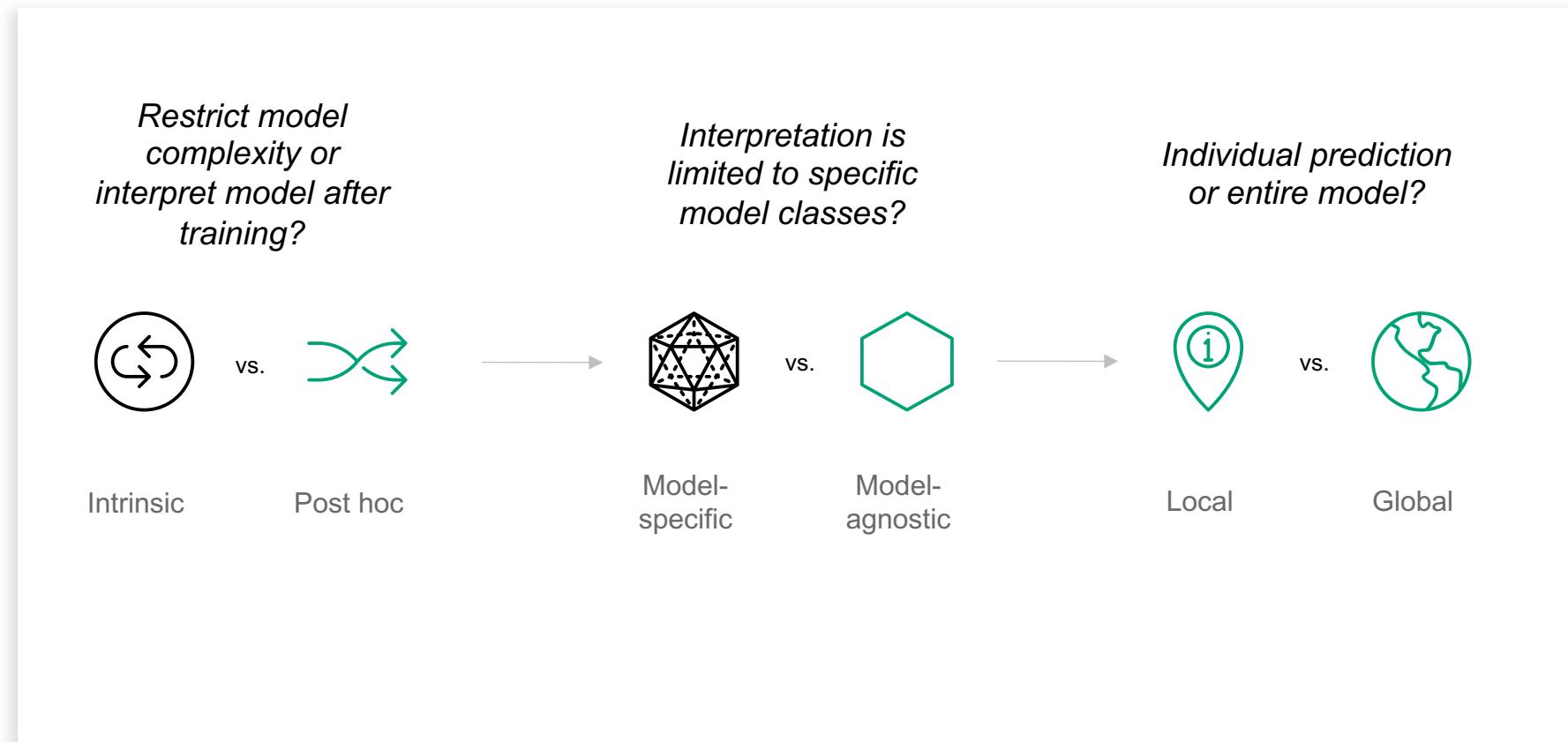
One does not simply interpret ML models, or does one?



# TAXONOMY OF INTERPRETABILITY

STATWORX

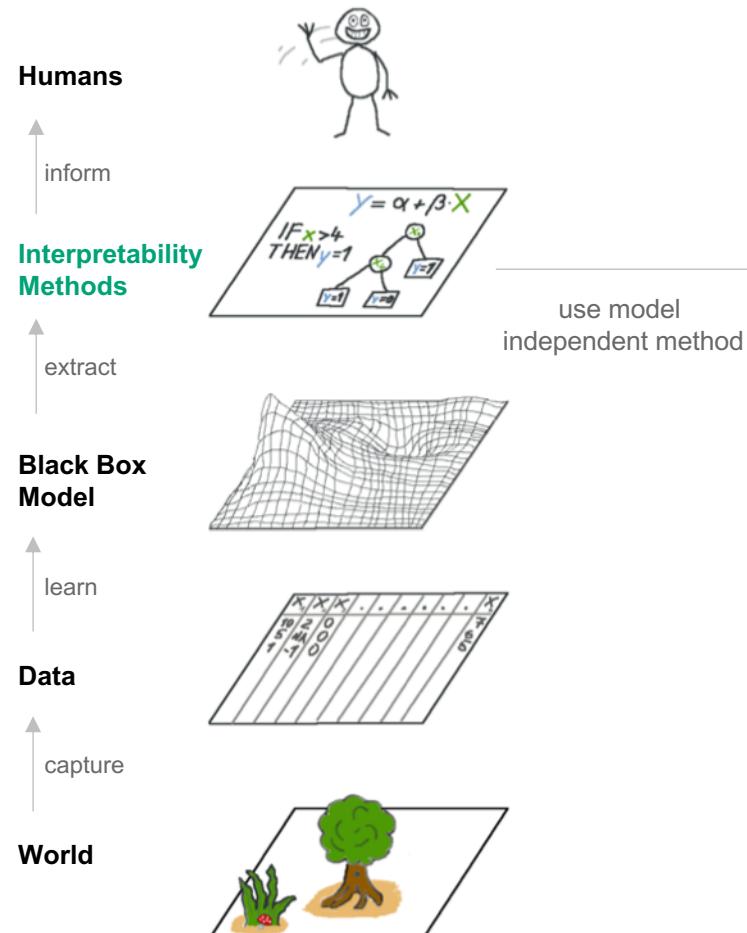
There is more than one method for model interpretability



# TAXONOMY OF INTERPRETABILITY

STATWORX

Post hoc, model-agnostic methods for interpretation



**Model-agnostic methods build solely upon the model's predictions and thus separate learning from explanation.**

**Advantages of model-agnostic methods:**

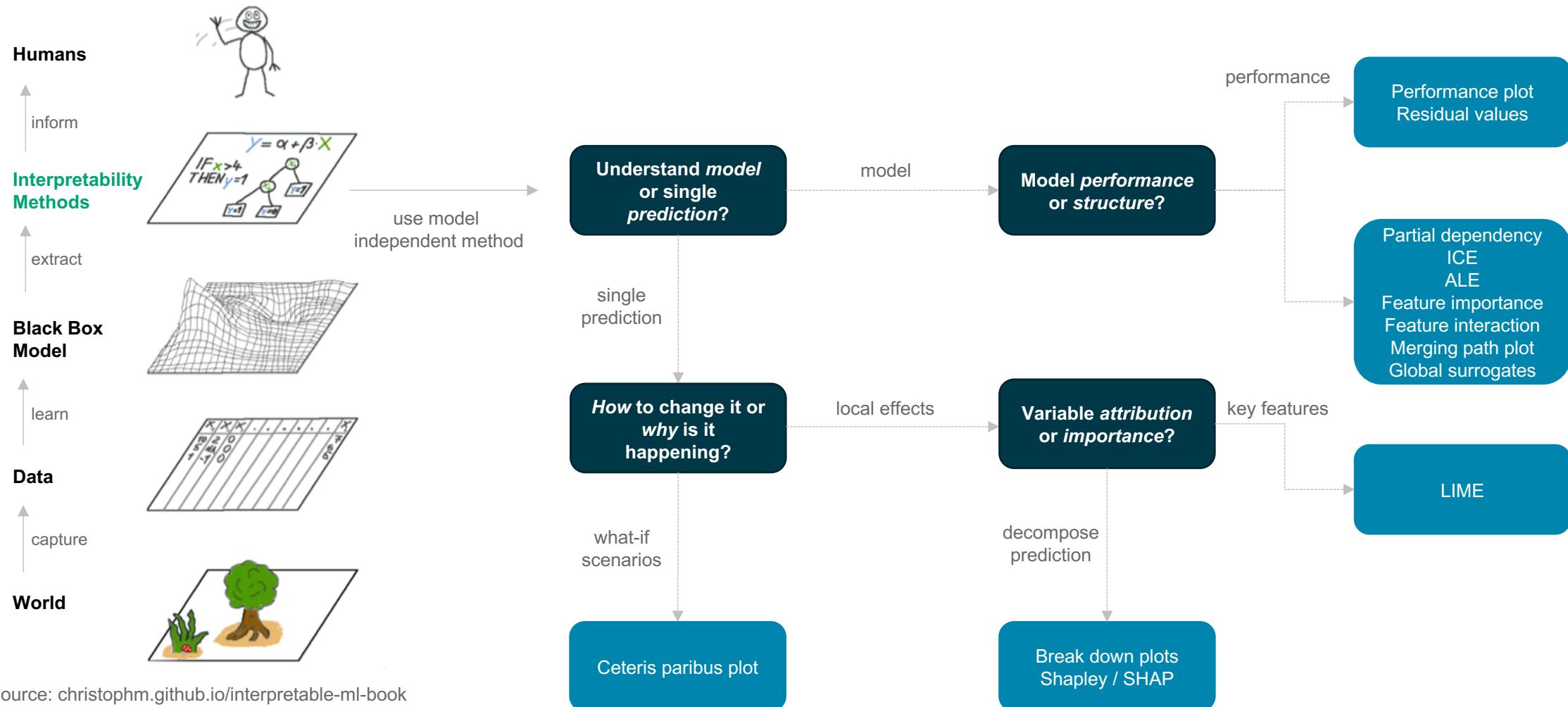
- ✓ High flexibility
- ✓ Model-independent explanation
- ✓ Explanations are comparable between different model classes

Source: christophm.github.io/interpretable-ml-book

# TAXONOMY OF INTERPRETABILITY

STATWORX

What method to use when?



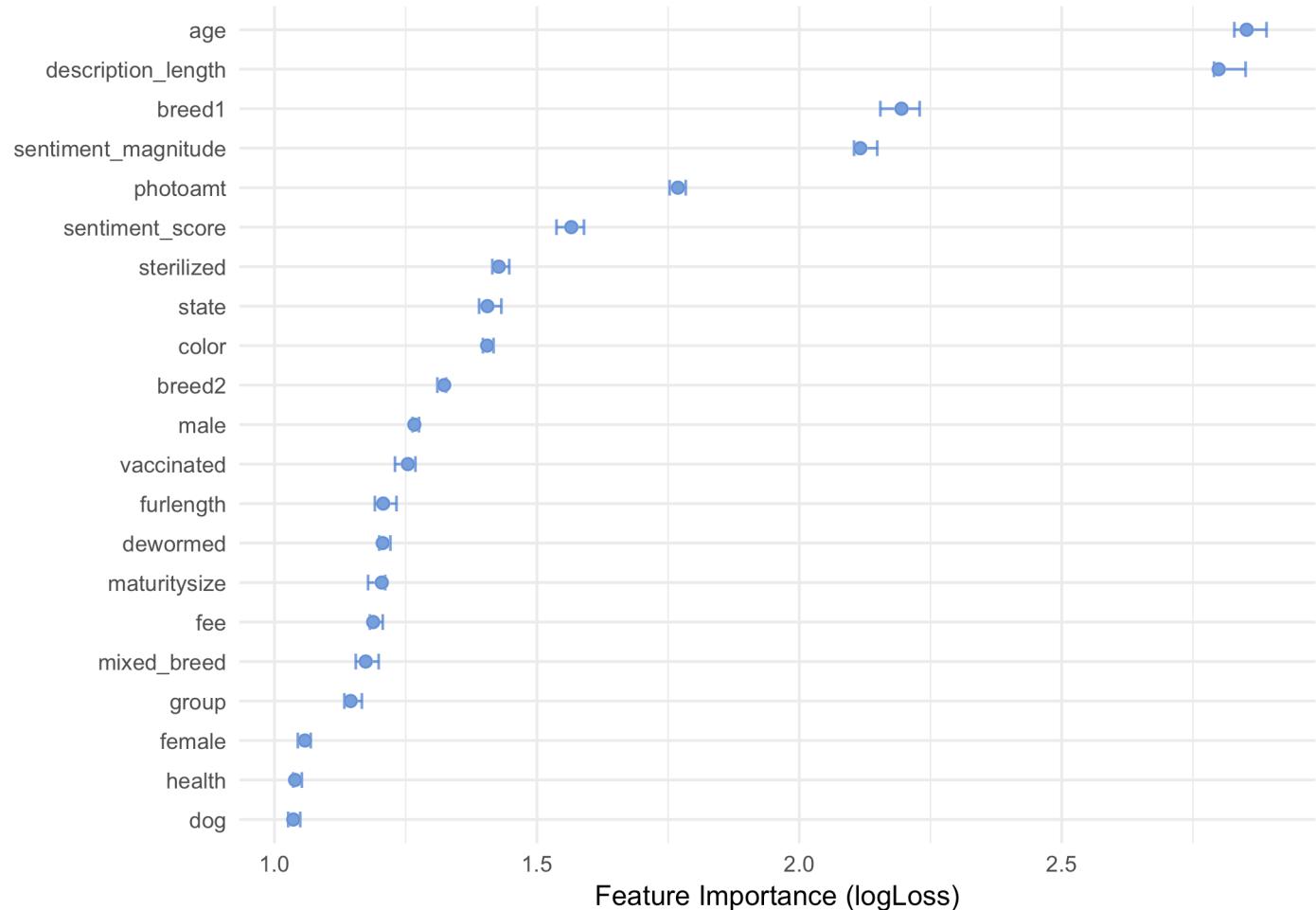
# ADOPTION PREDICTION

STATWORX

Explaining the overall model structure with visualized feature importance

Feature importance explained by:  
**Permutation feature importance**

- Measure feature importance by increase in model error after feature permutation
- ✓ Intuitive calculation and easy interpretation
- ✓ Comparable between algorithms
- ✗ Feature importance on training or testing data?
- ✗ Importance based on prediction error (no variance component)



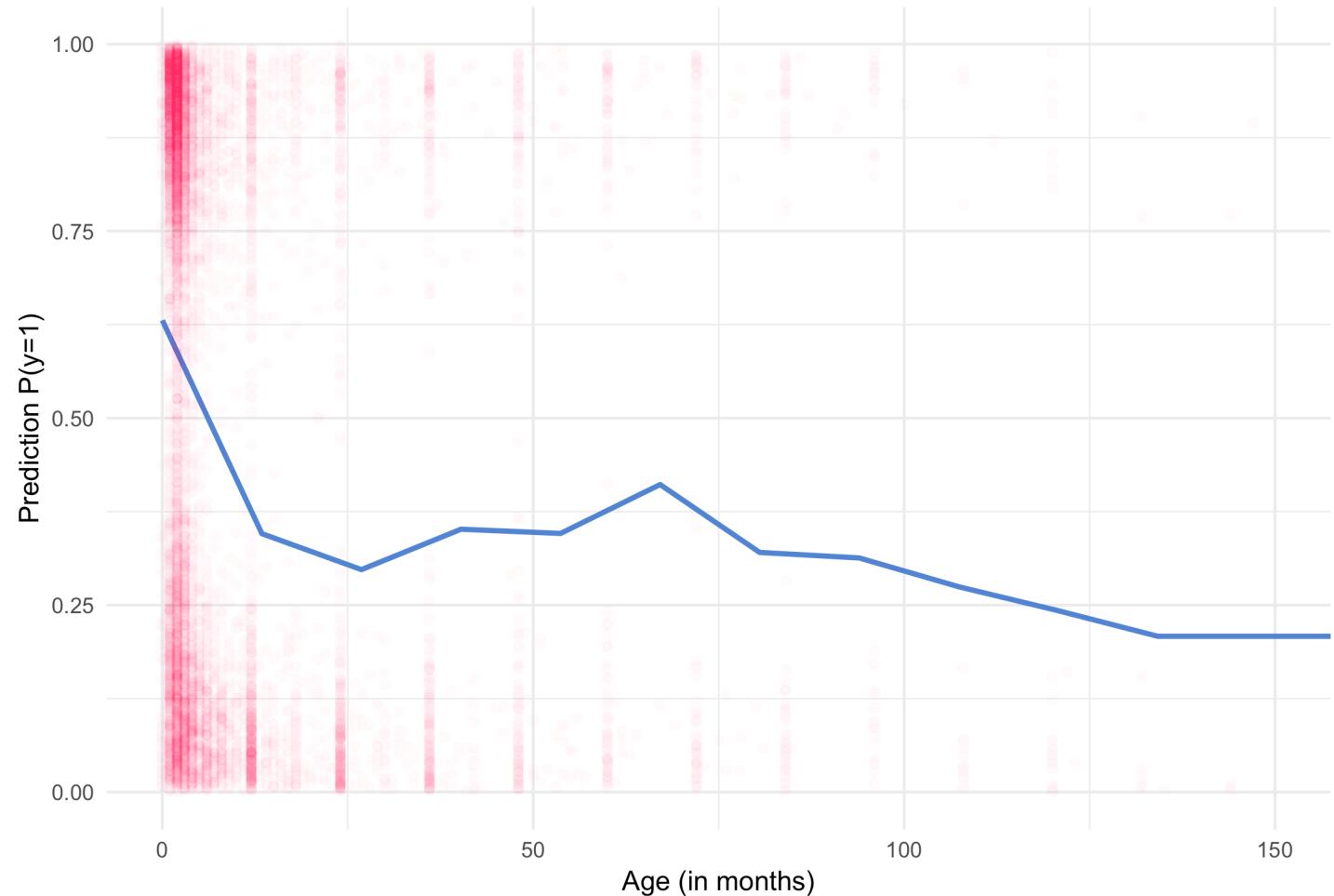
# ADOPTION PREDICTION

STATWORX

Explaining the effect of a single feature

Feature effect explained by:  
**Partial dependency plot (PDP)**

- PDP shows the marginal effect of one features on the predicted outcome
  - For a given feature value (e.g. age=2) take the average prediction over the entire dataset
- ✓ Intuitive calculation and easy interpretation
- ✗ Assumes independence between features



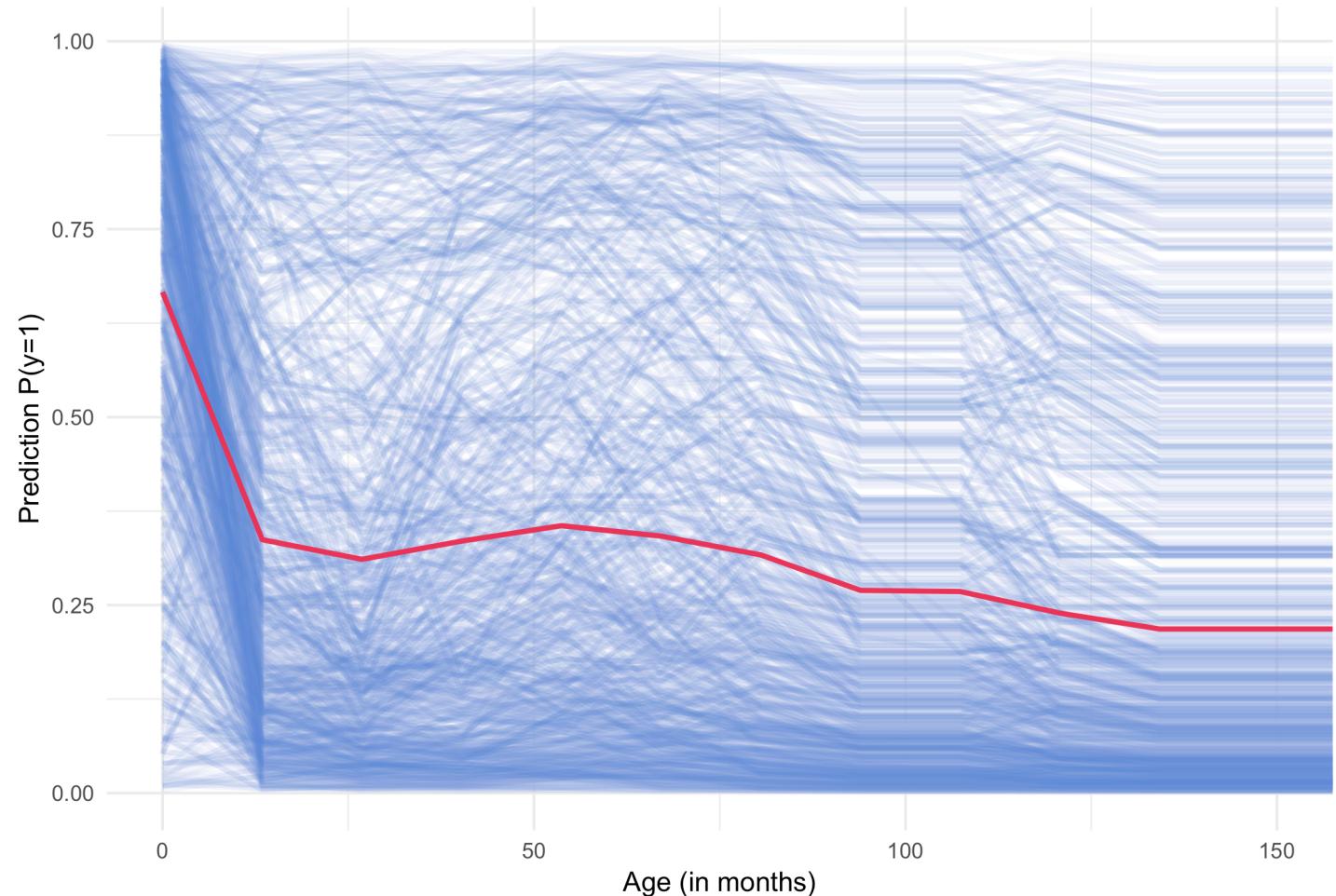
# ADOPTION PREDICTION

STATWORX

Explaining the effect of a single feature

Feature effect explained by:  
**Individual Conditional Expectation (ICE)**

- Shows dependence of the prediction on a feature for each instance separately
- ✓ Uncover obscure heterogeneous relationship
- ✗ Overplotting for medium/large data



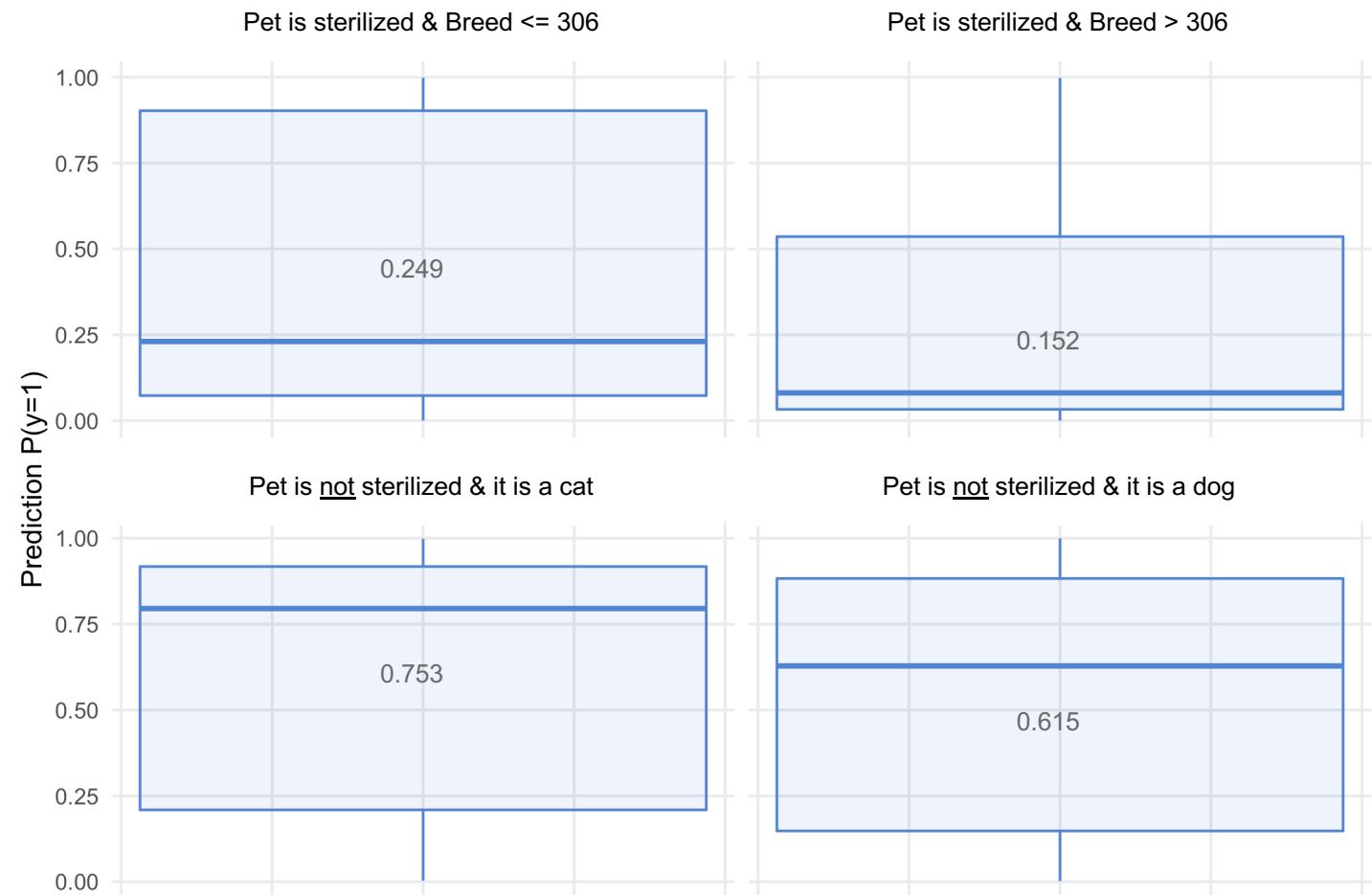
# ADOPTION PREDICTION

Explaining the model structure with a global surrogate model

Model structure explained by:

## Global Surrogate Tree

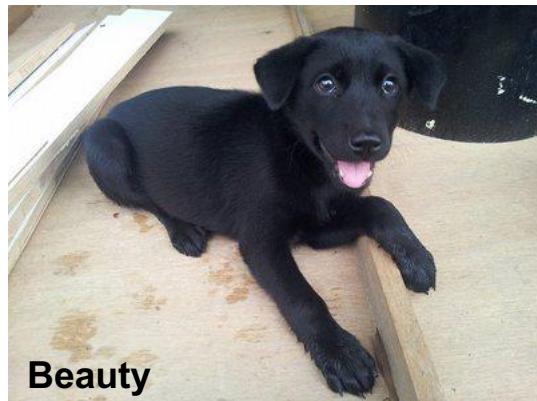
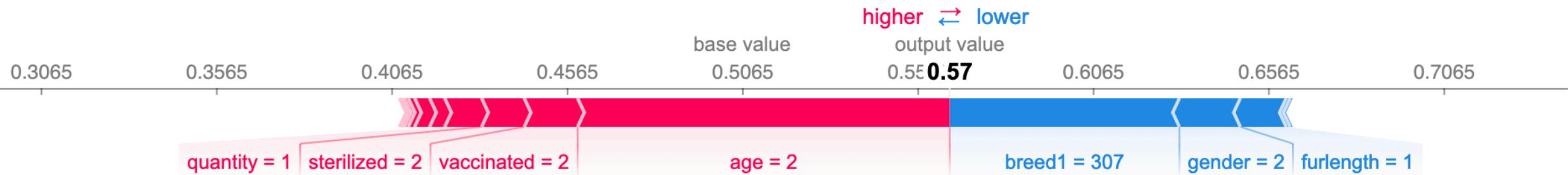
- Train interpretable model (e.g. tree) to approximate predictions of black box model
- ✓ Intuitive and straightforward
- ✓ Approximation quality can be measured
- ✓ Extendable to local surrogate models (e.g. LIME)
- ✗ Approximation can be (locally/globally) misleading



# ADOPTION PREDICTION

STATWORX

Beauty's probability for adoption explained by Shapley Values



Beauty

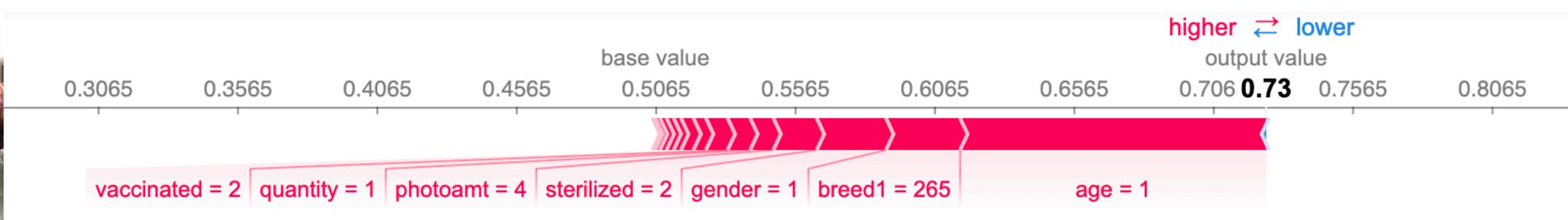
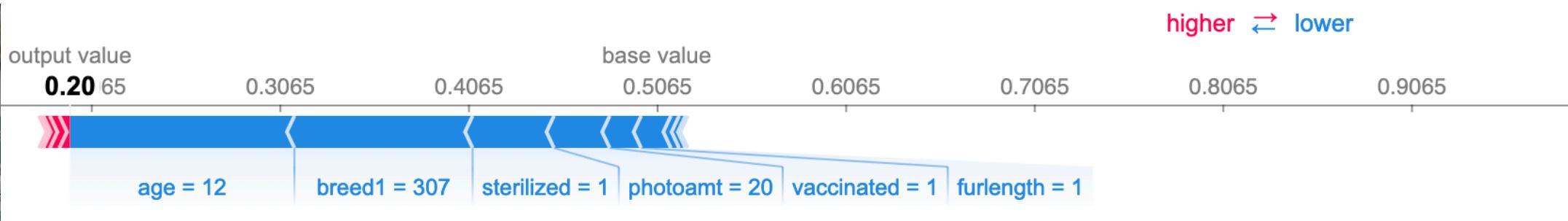
Explain Beauty's probability of adoption:

- + Beauty's age of two years makes her more likely to get adopted (by ~10%)
- + Being neither sterilized nor vaccinated makes her more likely to get adopted (by ~3%)
- Beauty is of mixed breed, which makes her less likely to get adopted (by ~7%)
- Beauty is a female dog, which makes her less likely to get adopted (by ~3%)
- Beauty has short hair, which makes her less likely to get adopted (by ~2%)

# ADOPTION PREDICTION

STATWORX

What about the Shapley Values for Danny and Kat?



# ADOPTION PREDICTION

STATWORX

One does not simply interpret deep learning models, or does one?



# ADOPTION PREDICTION

One does not simply interpret deep learning models, or does one?

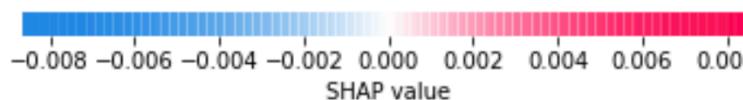
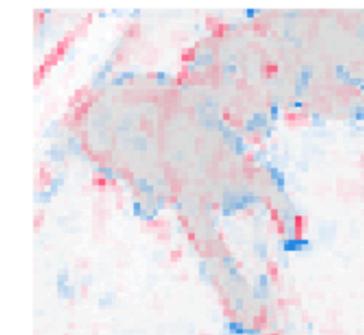
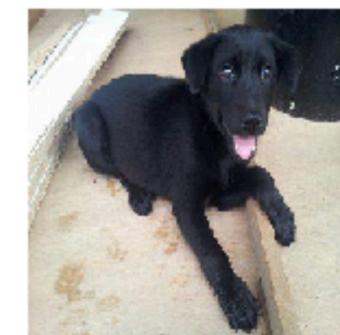
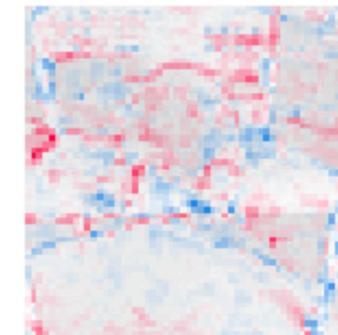
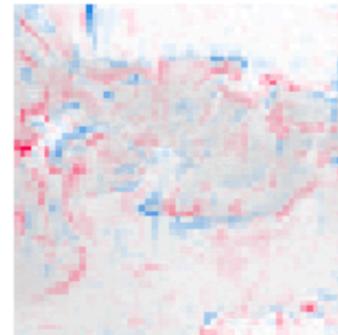
STATWORX



# ADOPTION PREDICTION

STATWORX

Explaining fine-tuned VGG16 predictions with Shapley Values



# CONCLUSION

STATWORX

What to take home

## Interpretability is important

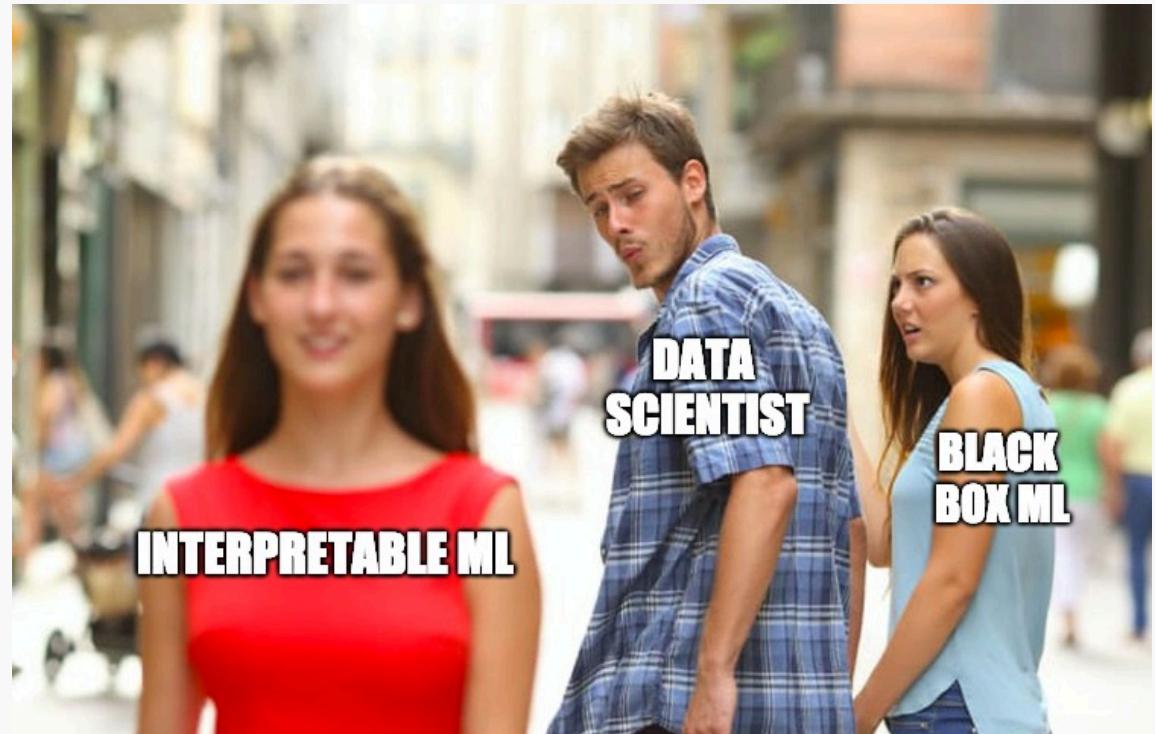
- ... to accurately describe real-world tasks
- ... to satisfy human curiosity
- ... to detect bias in your models
- ... to generate social acceptance and trust
- ... to allow for easy model debugging

## Interpretability is “easy” to implement

- Use model-agnostic methods
- Global methods: PDP, ICE & surrogate models
- Local methods: Shapley Values (and Lime)
- and many more

## Resources

- [github.com/fabianmax/pet-finder](https://github.com/fabianmax/pet-finder)
- [www.statworx.com/blog](http://www.statworx.com/blog)



# WE ARE HIRING! JOIN OUR TEAM



(Senior) Data Scientist (m/w/d)  
(Senior) Data Engineer (m/w/d)

VIELEN DANK  
FÜR IHRE AUFMERKSAMKEIT.

---

**KONTAKT**

Fabian Müller, Head of Data Science  
[fabian.mueller@statworx.com](mailto:fabian.mueller@statworx.com)  
[www.statworx.com](http://www.statworx.com)

**STATWORX**

Wöhlerstr. 8-10  
60323 Frankfurt am Main