# Visualize, Explore and Explain Predictive ML Models

PyData Warsaw 2018

Przemysław Biecek

# 59 impressive things artificial intelligence can do today

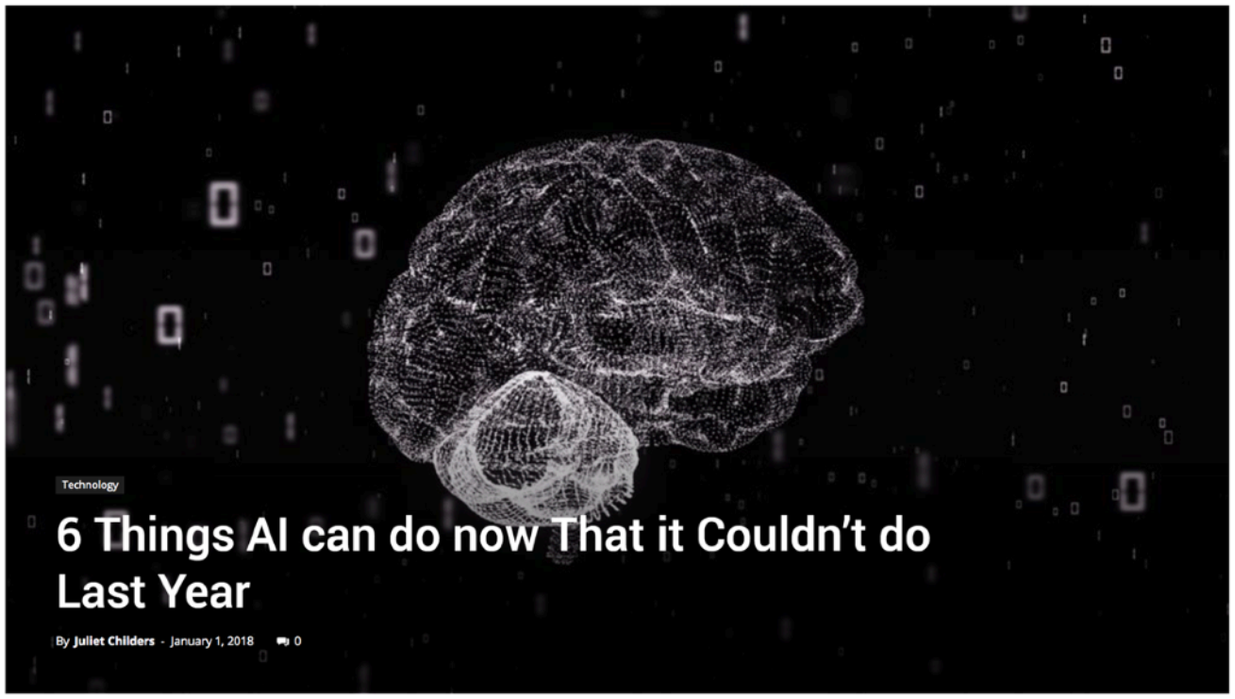
Ed Newton-Rex, Medium Mar. 7, 2017, 9:48 AM

**2050.**

That's the year in which artificial intelligence will be able to perform **any intellectual task a human can perform**, according to one survey of experts at a recent AI conference. Anything and everything any person has ever done in all of history—all of it doable, by 2050, by intelligent machines.

Streeter Lecka/Getty Images






# Applications of artificial intelligence

From Wikipedia, the free encyclopedia

Artificial intelligence, defined as intelligence exhibited by machines, has many applications in today's society. More specifically, it is Weak AI, the form of A.I. where programs are developed to perform specific tasks, that is being utilized for a wide range of activities including medical diagnosis, electronic trading, robot control, and remote sensing. AI has been used to develop and advance numerous fields and industries, including finance, healthcare, education, transportation, and more.

**Contents** [hide]



**Artificial intelligence**

**Major goals**

Knowledge reasoning
Planning
Machine learning
Natural language processing
Computer vision
Robotics
Artificial general intelligence

**Approaches**

Symbolic

//dzone.com/articles/cv-r-cvs-retrieval-system-based-on-job-description

DZone AI Zone

Over a million developers have joined DZone

REFCARDZ GUIDES ZONES JOBS | Agile AI Big Data Cloud Database DevOps Integration IoT Java Microservices Open Source

# Using Machine Learning to Retrieve Relevant CVs Based on Job Description

If you've ever tried to hire anyone, you know how difficult it can be to pour through hundreds of resumes and find the right one. AI can take the pain out of the process!

by Francis C. Fernandez-Reyes · Oct. 16, 17 · AI Zone · Tutorial

**How hard it can be?**

We use the average word embeddings (AWE) model for retrieving relevant CVs based on a job description. We present a step-by-step guide in order to combine domain-trained word embeddings with pre-trained embeddings for Spanish documents (CVs). We also use Principal Component Analysis (PCA) as a reduction technique used to put similar dimensions to word embeddings results.

## Architecture Description

Information retrieval (IR) models are composed of an indexed corpus and a scoring or ranking function. The main goal of an IR system is to retrieve relevant documents or web pages based on a user request. During the retrieval, the scoring function is used to sort the retrieved documents according to their relevance to the user query. The classic IR models such as BM25 and language models are based on the bag-of-words (BOW) indexing scheme. BOW models have two major weaknesses: they lose the context where a word appears and they also ignore its semantics. Latent semantic indexing (LSI) is a technique used to handle this problem but when the number of documents increases, the process of indexing becomes computationally expensive. The standard technique used to overcome this is to train word or paragraph embeddings over a corpus or use pre-trained embeddings.

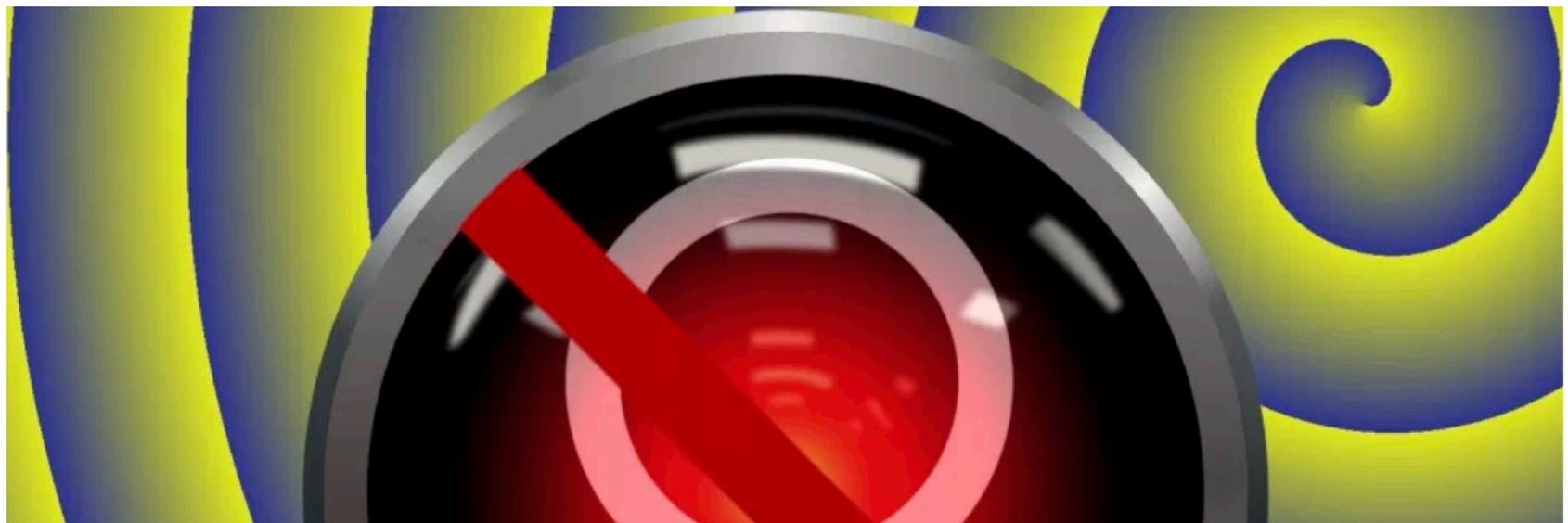
# Amazon Fired Its Resume-Reading AI for Sexism

Now abandoned, the project show

It began to penalize resumes that included the word "women," meaning phrases like "volunteered with Women Who Code" would be marked against the applicant. It specifically targeted two all-women's colleges, although sources would not tell Reuters which ones.

Amazon built the system in 2014 and scrapped it in 2017, after concluding that it was unsalvagable -- sources told Reuters that it rejected applicants from all-woman colleges, and downranked resume's that included the word "women's" as in "women's chess club captain."

**Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist**



> The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.
>
> Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

VICE News

AMAZON

# Amazon's resumé robot taught itself to discriminate against female applicants, report finds

By **Rex Santus** Oct 10, 2018

THE VERGE TECH SCIENCE CULTURE CARS

TECH | AMAZON | ARTIFICIAL INTELLIGENCE

21

# Amazon reportedly scraps internal AI recruiting tool that was biased against women

*The secret program penalized applications that contained the word "women's"*

By James Vincent | @jjvincent | Oct 10, 2018, 7:09am EDT

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

# Amazon scraps secret AI recruiting tool that showed bias against women

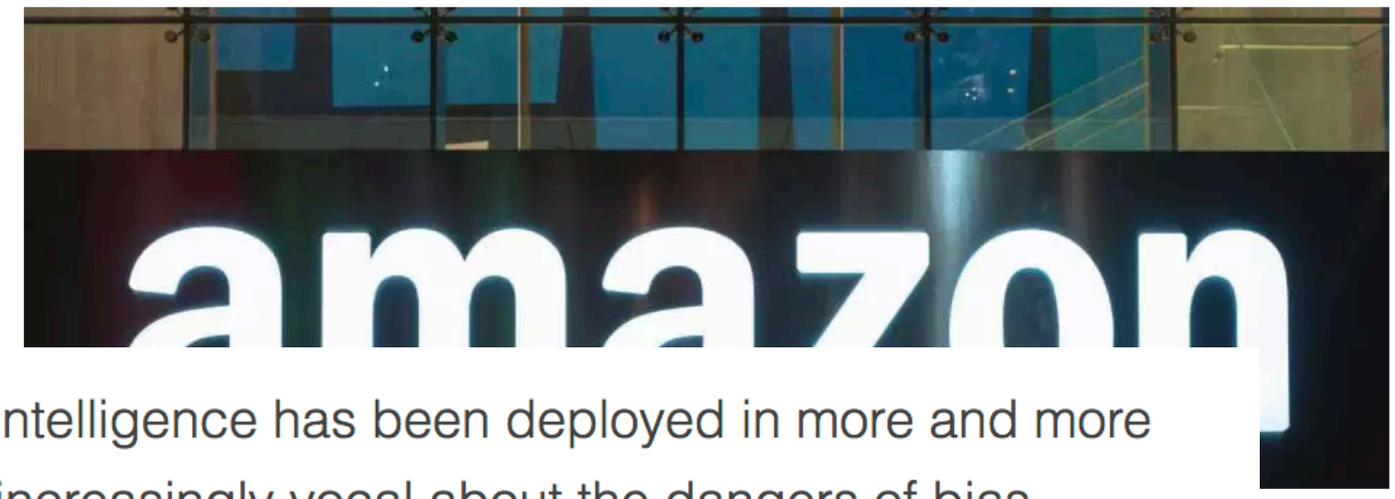
Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon
specialists uncovered a big problem: the



# Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

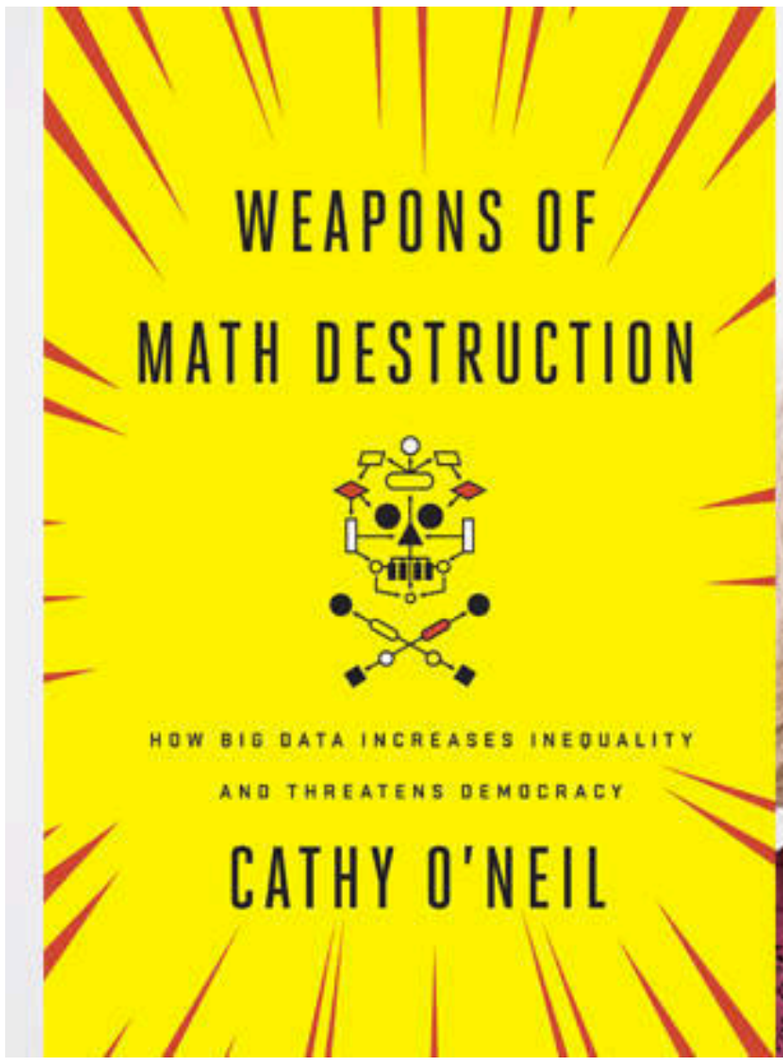



Over the past few years, as artificial intelligence has been deployed in more and more contexts, researchers have become increasingly vocal about the dangers of bias. Prejudices about gender and race can easily creep into a range of AI programs — everything from facial recognition algorithms to those used by the courts and hospitals.

# Cathy O'Neil:

# The era of blind faith in ~~big data~~ must end

black boxes




- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of mathematics to prevent them from asking questions”.

https://www.theguardian.com/books/2016/oct/27/cathy-oneil-weapons-of-math-destruction-algorithms-big-data

# Right to explanation

From Wikipedia, the free encyclopedia

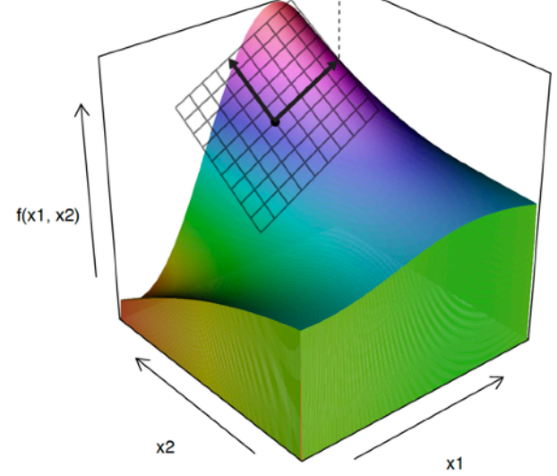
In the regulation of algorithms, particularly artificial intelligence and its subfield of machine learning, a **right to expla explanation**) is a right to be given an explanation for an output of the algorithm. Such rights primarily refer to individ explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a perso is denied may ask for an explanation, which could be "Credit bureau X reports that you declared bankruptcy last yea considering you too likely to default, and thus we will not give you the loan you applied for."

Some such legal rights already exist, while the scope of a general "right to explanation" is a matter of ongoing debat

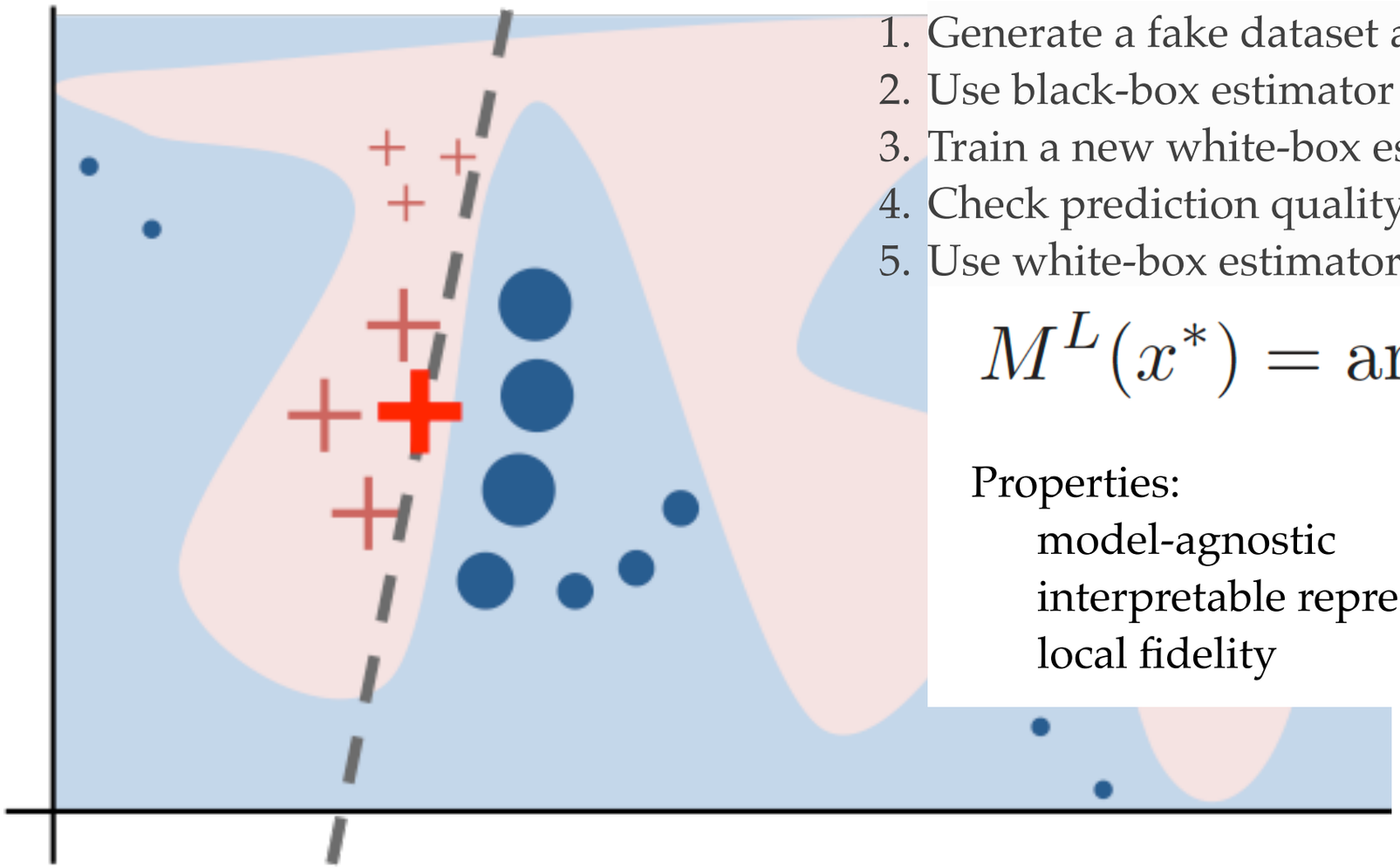
**Contents** [hide]

# We need model explainers!

# Local Model approximations




A different approach to model explanation is to locally approximate the complex black-box model with an easier to interpret white-box model constructed on interpretable features.



1. Generate a fake dataset around x.
2. Use black-box estimator to get target values y.
3. Train a new white-box estimator for (y,x).
4. Check prediction quality of a white-box classifier.
5. Use white-box estimator as an explanation of black-box model.

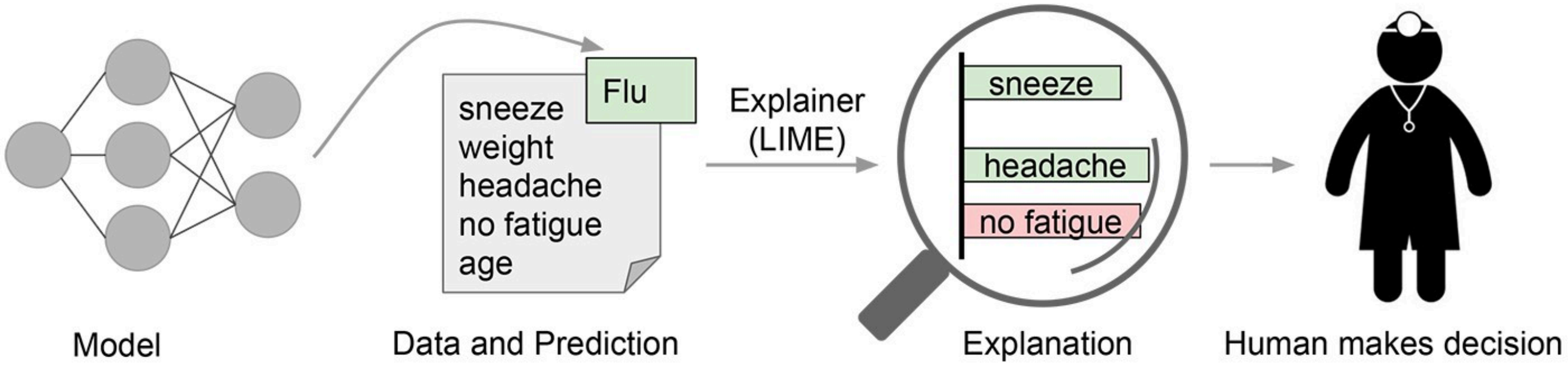
$$M^L(x^*) = \arg\min_{g \in G} L(f, g, \Pi_{x^*}) + \Omega(g$$

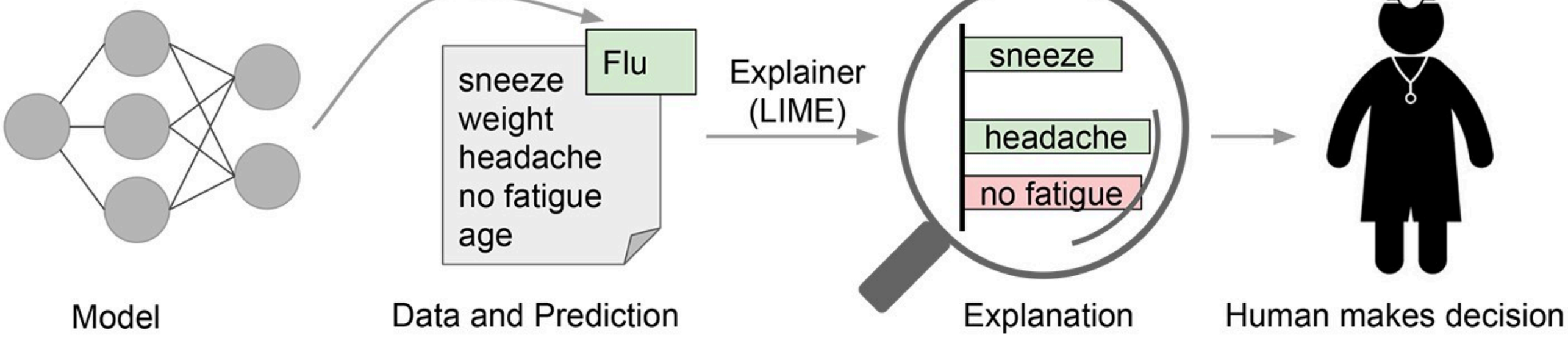
Properties:

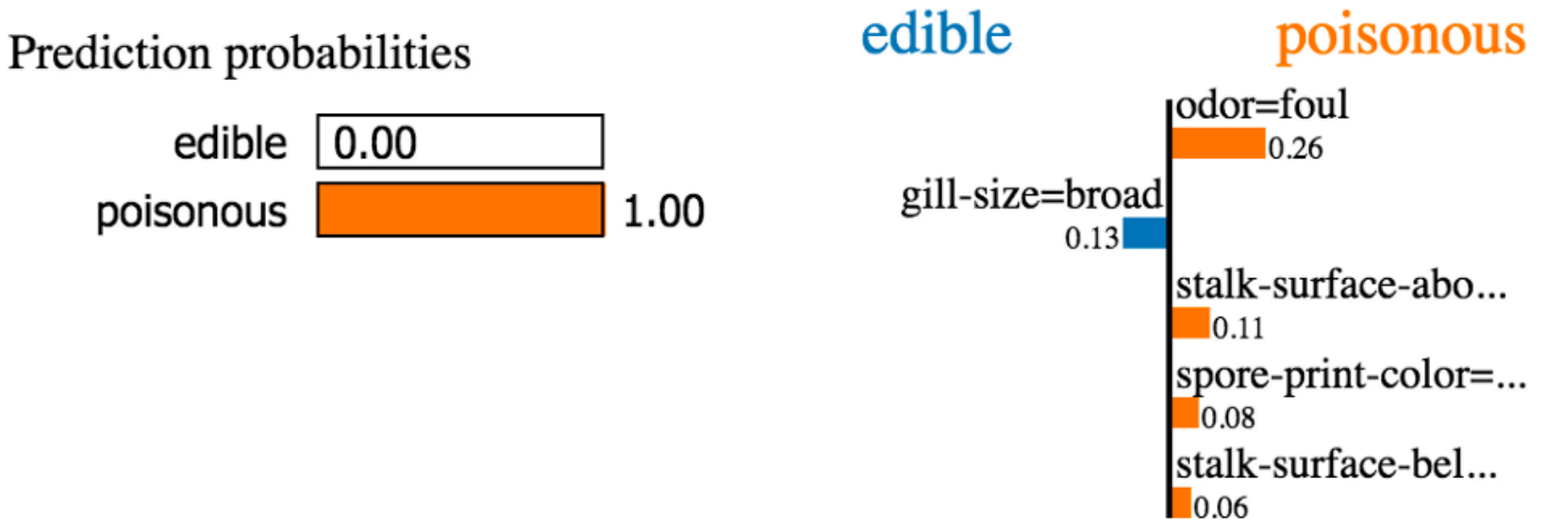
- model-agnostic
- interpretable representation
- local fidelity

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

Port to R: Thomas Lin Pedersen (2017) https://github.com/thomasp85/lime
Other implementations: live (Staniak, Biecek 2018) and iml (Molnar 2018)


Flu
sneeze
weight
headache
no fatigue
age
Explainer
(LIME)
sneeze
headache
no fatigue
Model
Data and Prediction
Explanation
Human makes decision


sneeze
weight
headache
no fatigue
age
Flu
Explainer
(LIME)
sneeze
headache
no fatigue
Model
Data and Prediction
Explanation
Human makes decision


Prediction probabilities
edible
0.00
poisonous
1.00
edible
poisonous
odor=foul
0.26
gill-size=broad
0.13
stalk-surface-abo...
0.11
spore-print-color=...
0.08
stalk-surface-bel...
0.06
Feature
Value
odor=foul
True
gill-size=broad
True
stalk-surface-above-ring=silky
True
spore-print-color=chocolate
True
stalk-surface-below-ring=silky
True

# Anuvada: Interpretable Models for NLP using PyTorch

https://github.com/EdGENetworks/anuvada

**Positive review**

a chemist develops a fabric that never gets dirty or wears out , but it is seen as a threat to the survival of various industries . in this delightful ealing studios comedy , guinness is marvelous as the mild - mannered but persistent chemist . greenwood , with her sensual voice , plays the love interest ; parker is her harried father . thesiger is amusing as a patriarch of the fabric industry . while telling an engaging story , the film also raises some intriguing questions about science , the economy , and politics . it is adeptly directed by mackendrick , who would go on to make \the _UNK " and the sublime \"sweet smell of _UNK " later in the 1950s . "

**Negative review**

worst movie i have seen since _UNK afternoon . i suppose that this is a horror / comedy . i pretty much predicted every scene in this movie . the special - effects were not so special . i believe that i could come up with as good of effects from what i have lying around the house . i wish i could have something good to say about this movie , but i am afraid that i do n't . even coolio should be ashamed of appearing in such a turkey . i do , after a little thought , have one thing good to say about this movie - it ended .

# “What is Relevant in a Text Document?”

https://arxiv.org/pdf/1612.07843.pdf

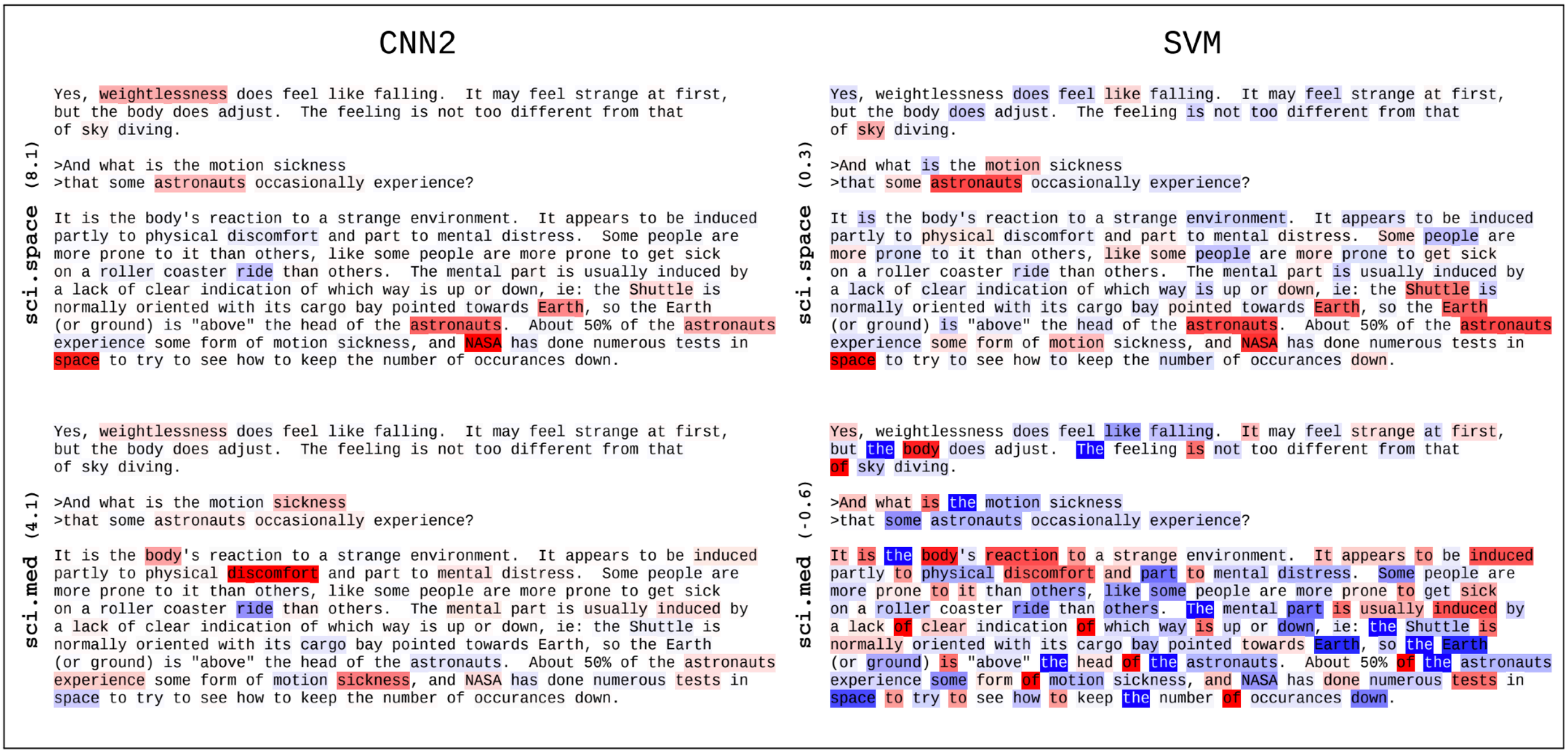
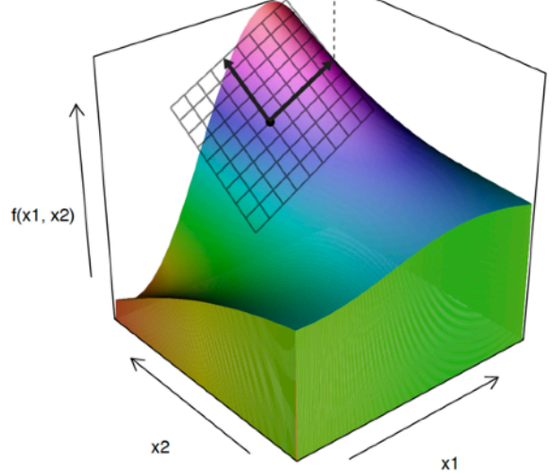



**Figure 2.** LRP heatmaps of the document `sci.space 61393` for the CNN2 and SVM model. Positive relevance is mapped to red, negative to blue. The color opacity is normalized to the maximum absolute relevance per document. The LRP target class and corresponding classification prediction score is

# Local Model approximations




Spectacular use-cases for image data and text data.

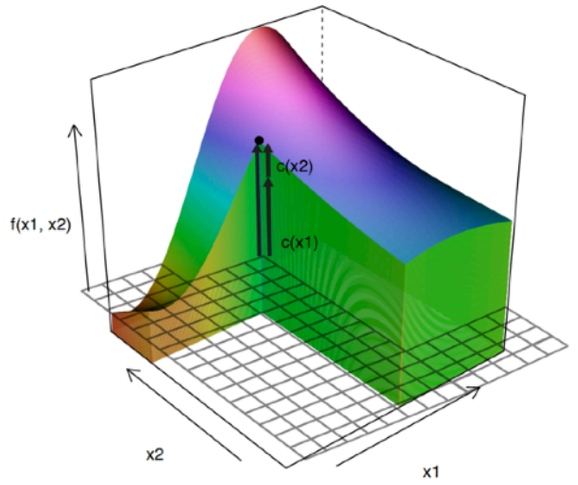
Not that developed for tabular data (what are interpretable features?
do we have to reduce continuous variables to sets of binary features).

(a) Husky classified as wolf

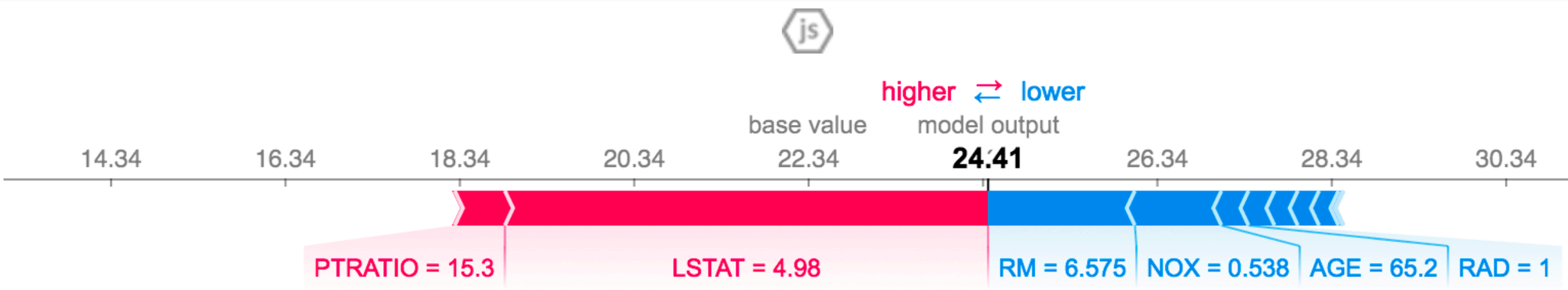
(b) Explanation

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

# Variable attributions




SHAP (SHapley Additive exPlanations) is a unified framework for interpretation of model predictions. It has desired properties (Local accuracy, Missingness, Consistency) and may be seen as unification of other approaches like DeepLIFT, Layer-Wise Relevance Propagation, LIME.




**A Unified Approach to Interpreting Model Predictions**

**Scott M. Lundberg**
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

**Su-In Lee**
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

# Visualizing and Understanding Atari Agents

http://www.interpretable-ml.org/nips2017workshop/papers/06.pdf

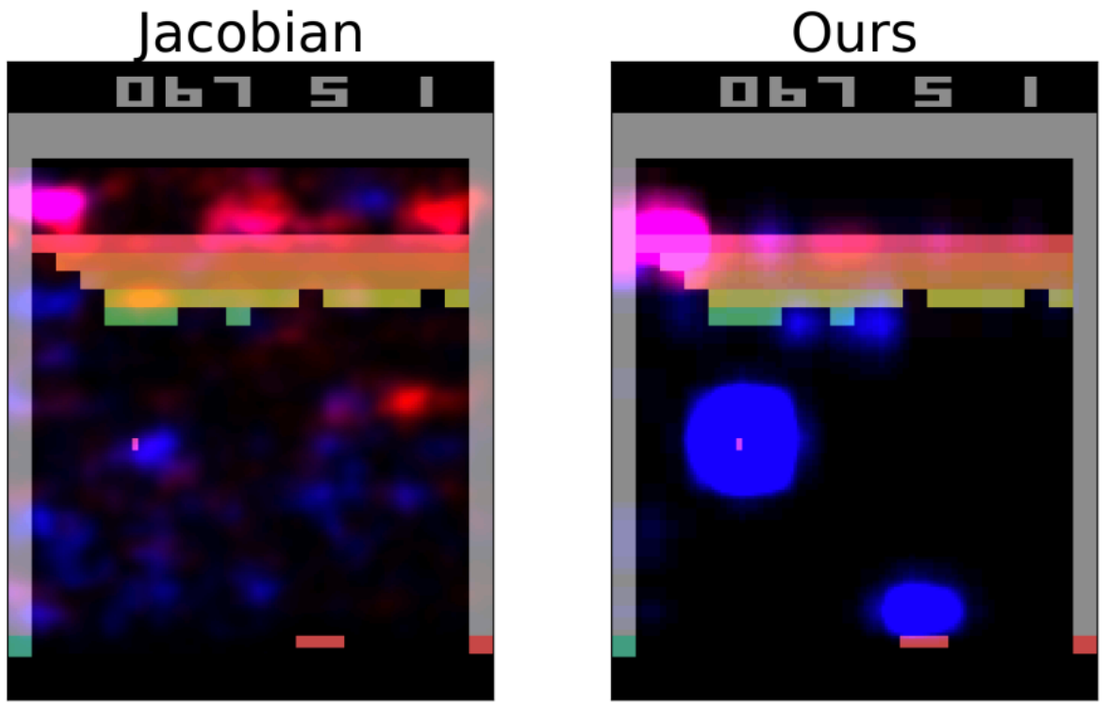
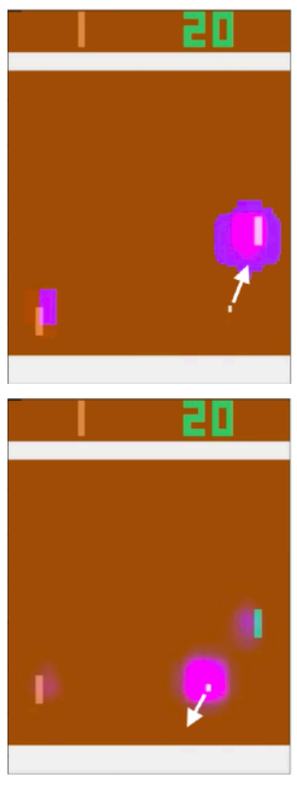
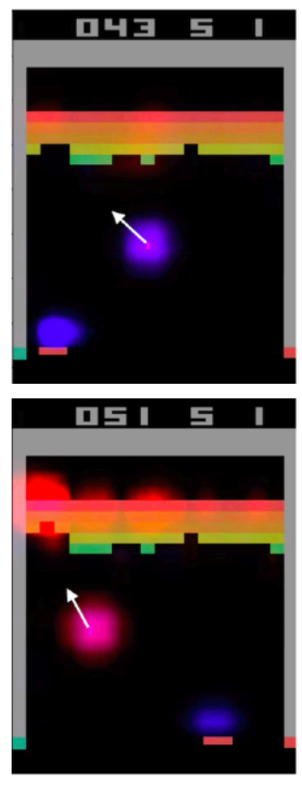



Figure 1: Comparison of Jacobian saliency to our perturbation-based approach. We are visualizing an actor-critic model [16]. Red indicates saliency for the critic; blue is saliency for the actor.



(a) Pong: kill shot

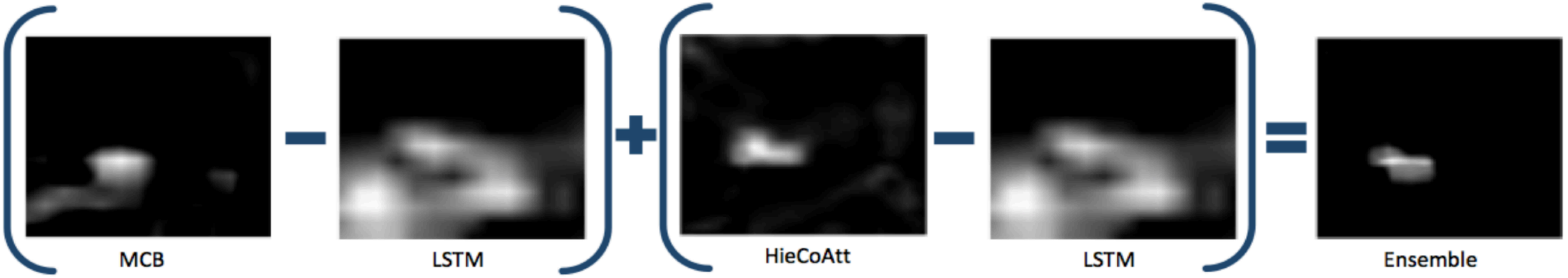
(b) SpaceInvaders: aiming



(c) Breakout: tunneling

# Ensembling Visual Explanations for VQA

http://www.cs.utexas.edu/~ml/papers/rajani.vigil17.pdf




**Q:** The car at the front of the train is what color? **A:** red

Ensemble

# On the Robustness of Interpretability Methods

https://arxiv.org/pdf/1806.08049.pdf

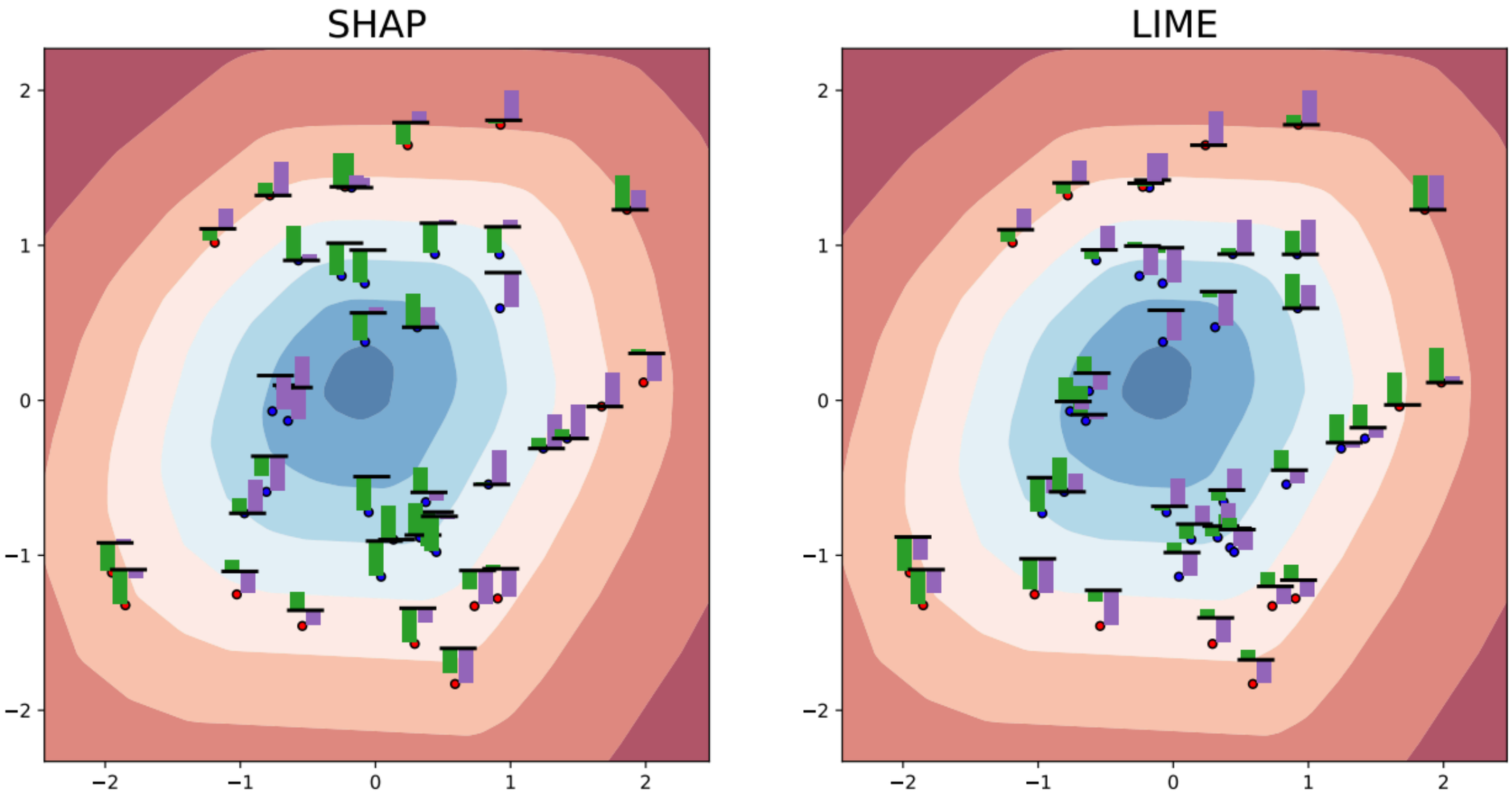
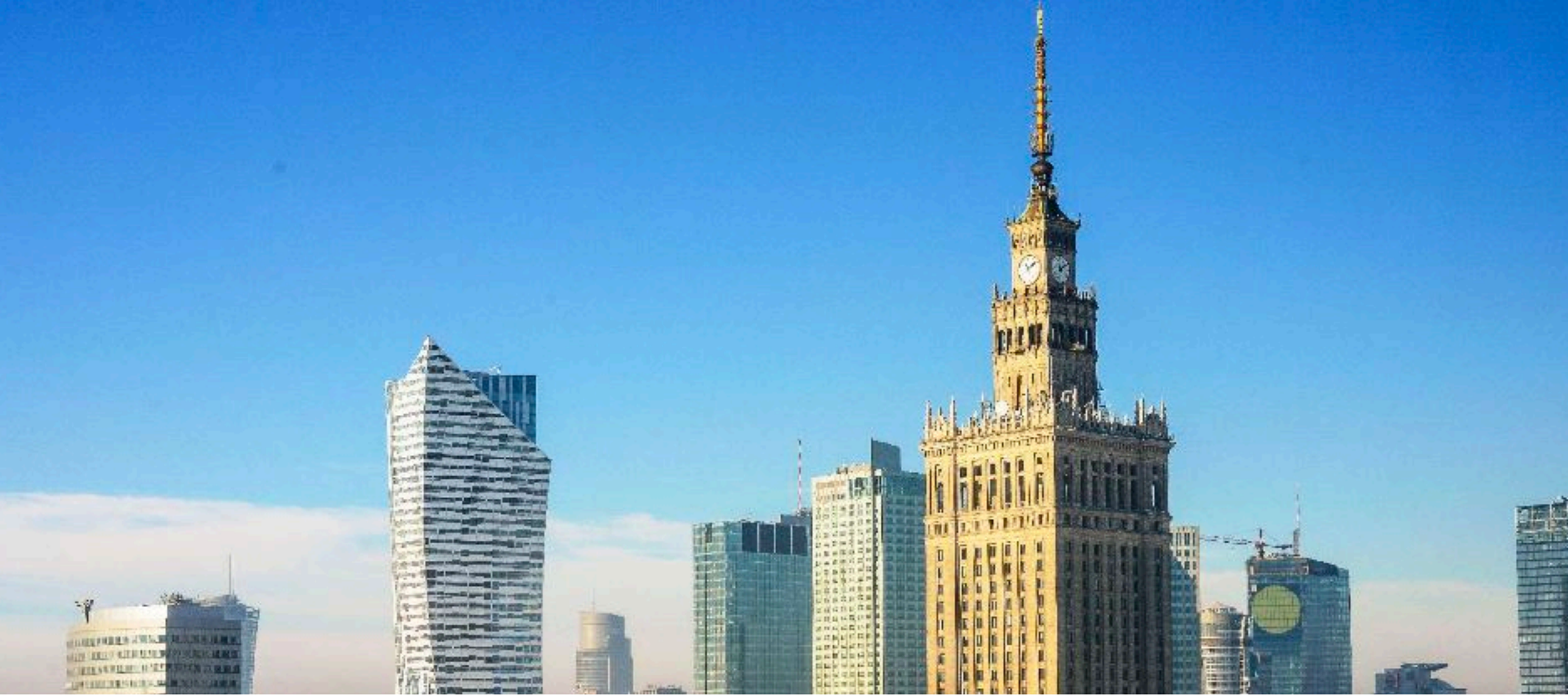



Figure 1: LIME and SHAP explanations for two simple bi-
nary classifiers: a linear SVM (top row) and a two-layer

# We can do even better!



```
library("DALEX")
head(apartments)
```

| m2.price | construction.year | surface | floor | no.rooms | district |
|---|---|---|---|---|---|
| 5897 | 1953 | 25 | 3 | 1 | Srodmiescie |
| 1818 | 1992 | 143 | 9 | 5 | Bielany |
| 3643 | 1937 | 56 | 1 | 2 | Praga |
| 3517 | 1995 | 93 | 7 | 3 | Ochota |
| 3013 | 1992 | 144 | 6 | 5 | Mokotow |

# Let's create two competing models

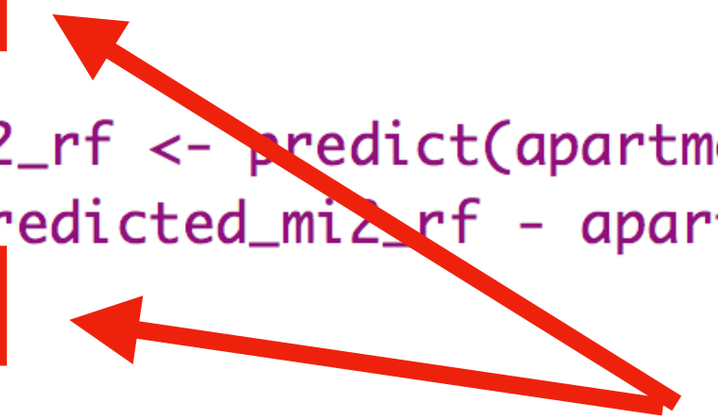
```
> library("DALEX")
> apartments_lm_model <- lm(m2.price ~ construction.year + surface + floor +
+                             no.rooms + district, data = apartments)
>
> library("randomForest")
> set.seed(3)
> apartments_rf_model <- randomForest(m2.price ~ construction.year + surfac
+                                     no.rooms + district, data = apartme
```

# Let's create two competing models

```
> library("DALEX")
> apartments_lm_model <- lm(m2.price ~ construction.year + surface + floor +
+                             no.rooms + district, data = apartments)
>
> library("randomForest")
> set.seed(3)
> apartments_rf_model <- randomForest(m2.price ~ construction.year + surfac
+                                     no.rooms + district, data = apartmen
>
> predicted_mi2_lm <- predict(apartments_lm_model, apartmentsTest)
> sqrt(mean((predicted_mi2_lm - apartmentsTest$m2.price)^2))
[1] 283.0865
>
> predicted_mi2_rf <- predict(apartments_rf_model, apartmentsTest)
> sqrt(mean((predicted_mi2_rf - apartmentsTest$m2.price)^2))
[1] 283.3479
```

# Let's create two competing models

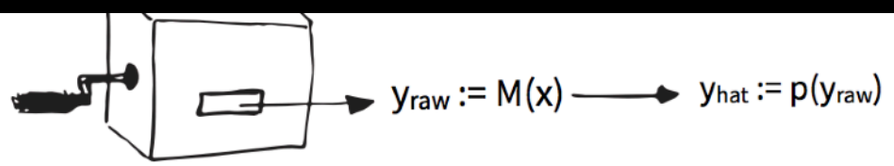
```
> library("DALEX")
> apartments_lm_model <- lm(m2.price ~ construction.year + surface + floor +
+                             no.rooms + district, data = apartments)
>
> library("randomForest")
> set.seed(3)
> apartments_rf_model <- randomForest(m2.price ~ construction.year + surfac
+                                     no.rooms + district, data = apartmen
>
> predicted_mi2_lm <- predict(apartments_lm_model, apartmentsTest)
> sqrt(mean((predicted_mi2_lm - apartmentsTest$m2.price)^2))
[1] 283.0865
>
> predicted_mi2_rf <- predict(apartments_rf_model, apartmentsTest)
> sqrt(mean((predicted_mi2_rf - apartmentsTest$m2.price)^2))
[1] 283.3479
```



**Which one is better?**

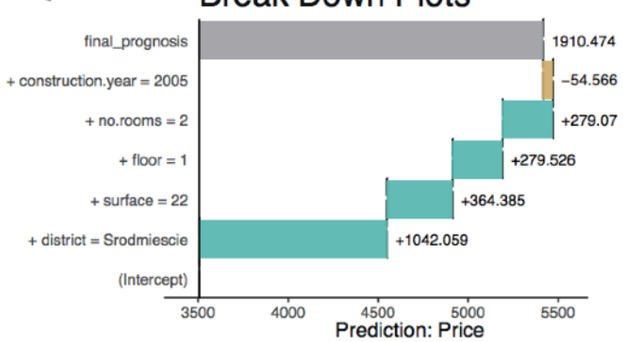
# DALEX architecture

**A)** model

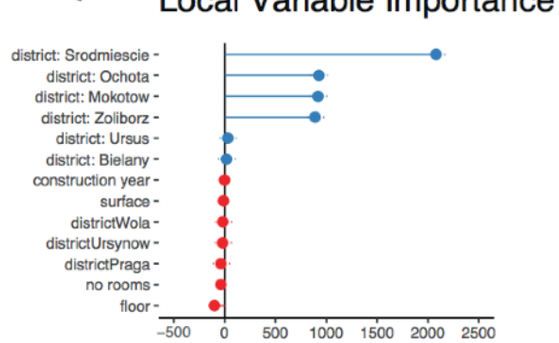



**B)** explain(model; data; y; predict_function)

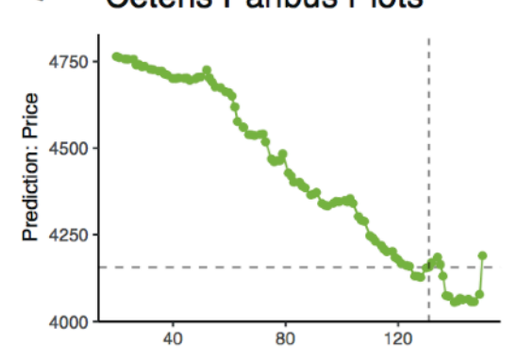
**C)**



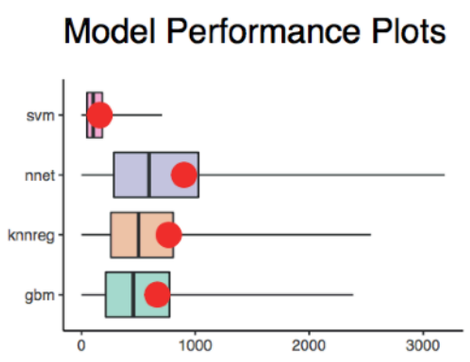

**D)**



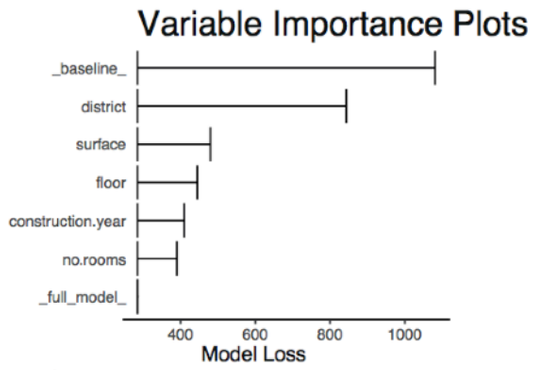

**E)**



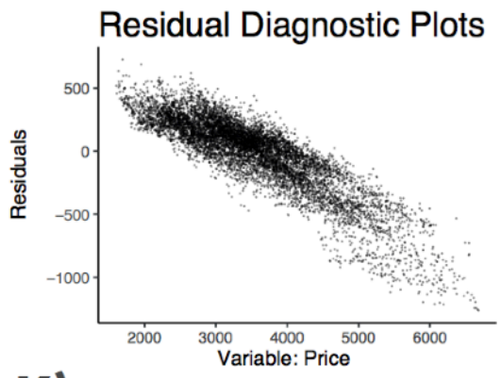

**F)**



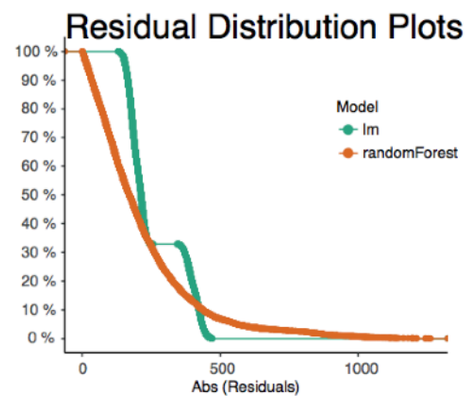

**G)**



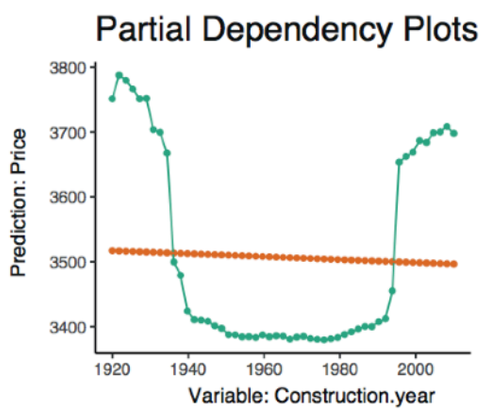

**H)**



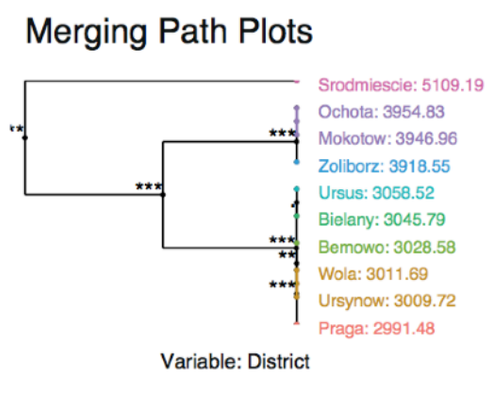

**I)**




**J)**




**K)**




**Wrap a model**

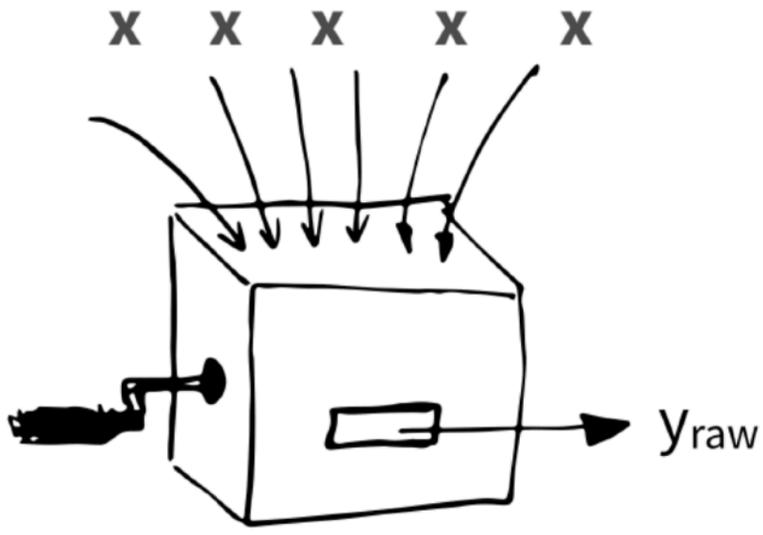
↓

**Create an explainer**

↓

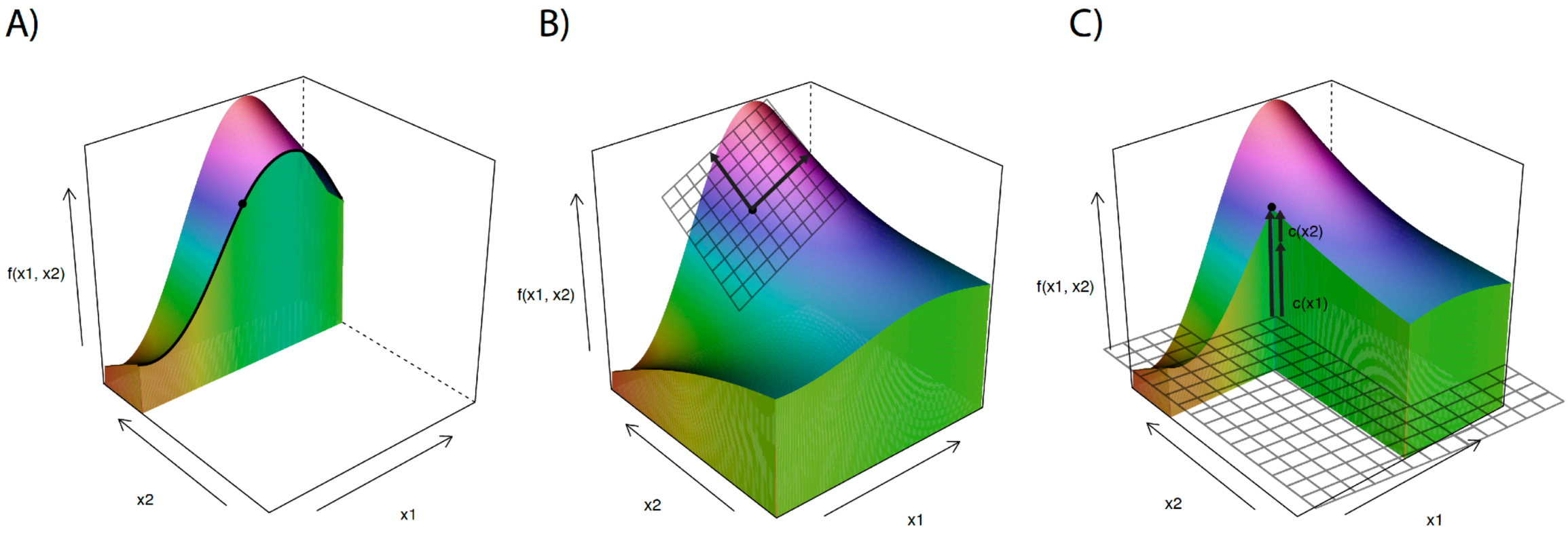
**Plot the explainer**

# How can we explain model predictions?




Black box model is a function

$$f : R^p \rightarrow R$$



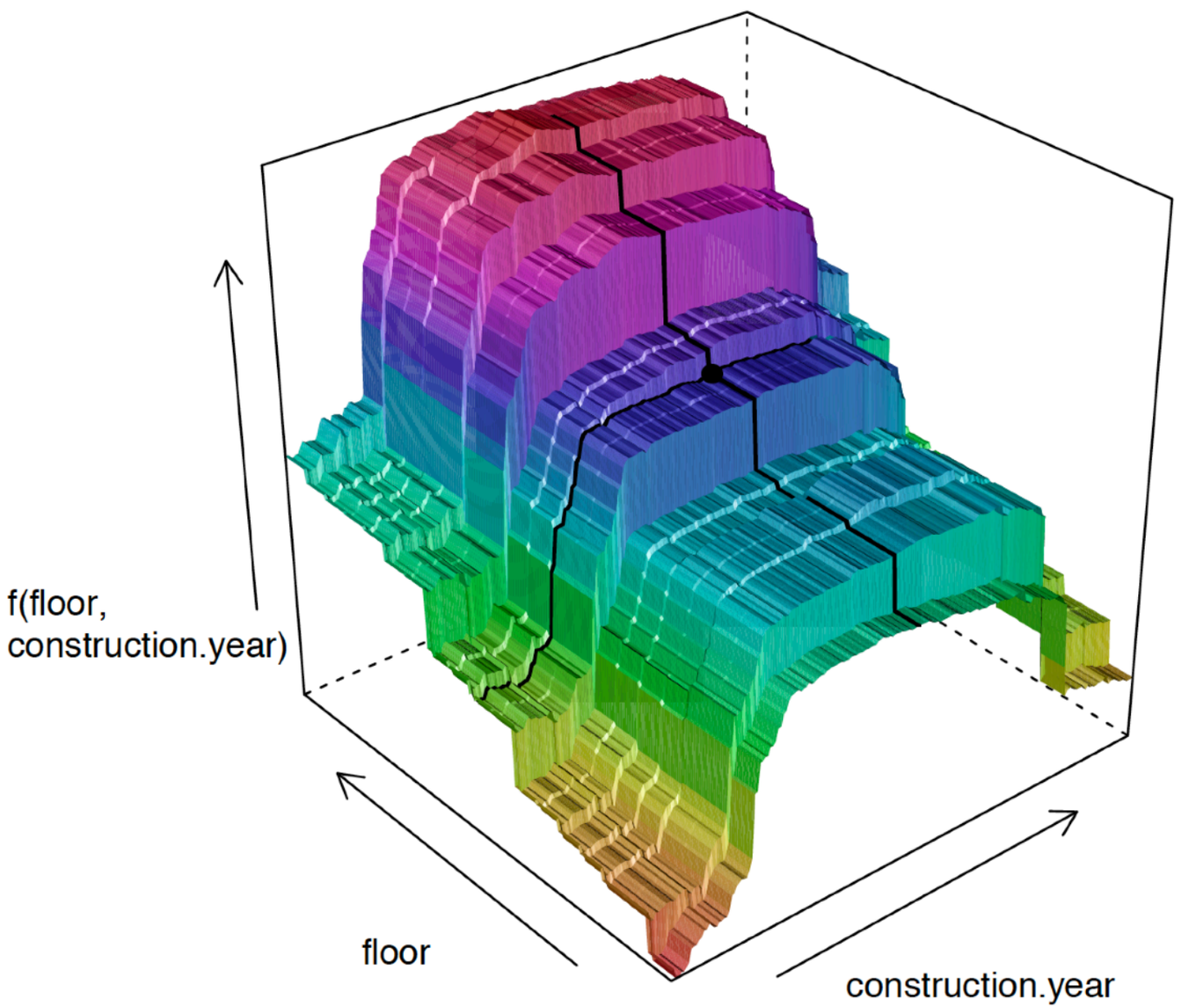
# Ceteris Paribus Profiles

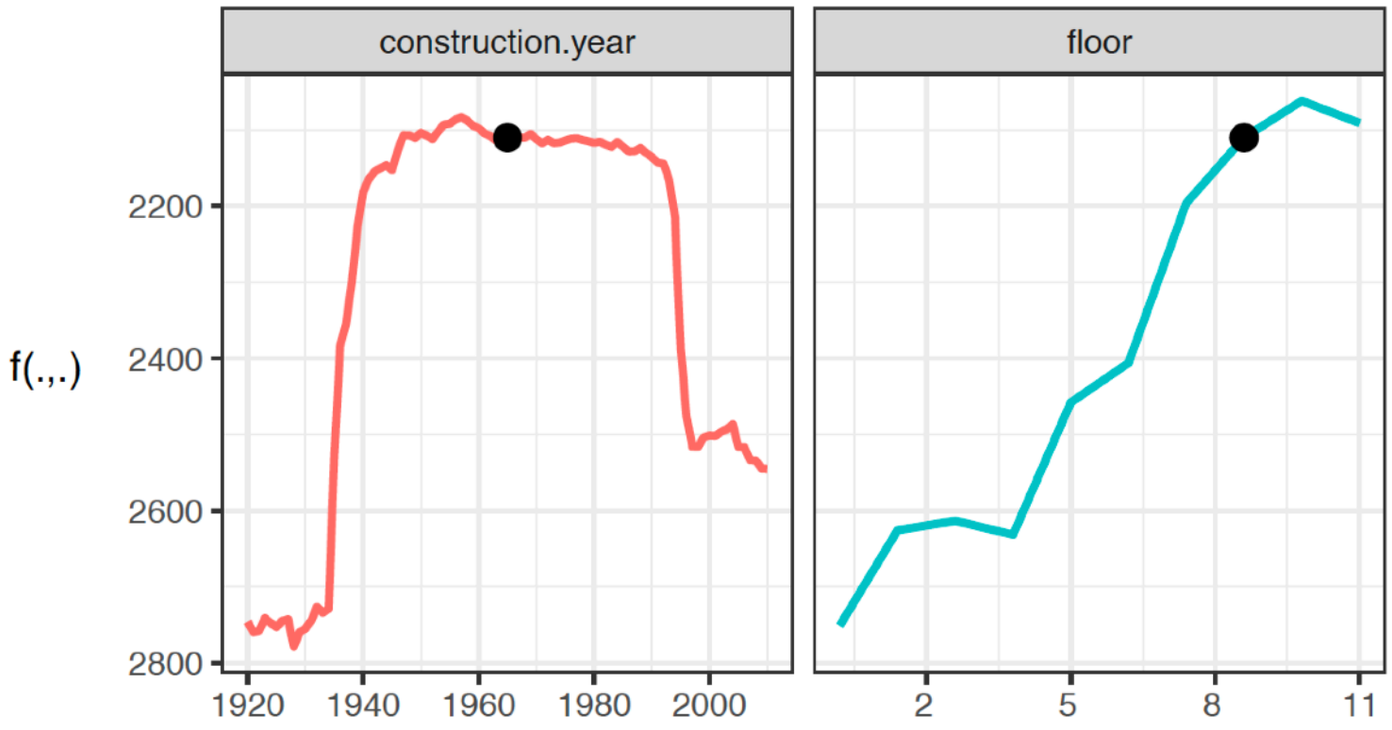
Ceteris paribus is a Latin phrase meaning “other things held constant” or “all else unchanged”.

It is a method for exploration of model responses given only a single variable is changed.

More formally

$$CP^{f,j,x}(z) := f(x|^j = z)$$



# Ceteris Paribus Profiles

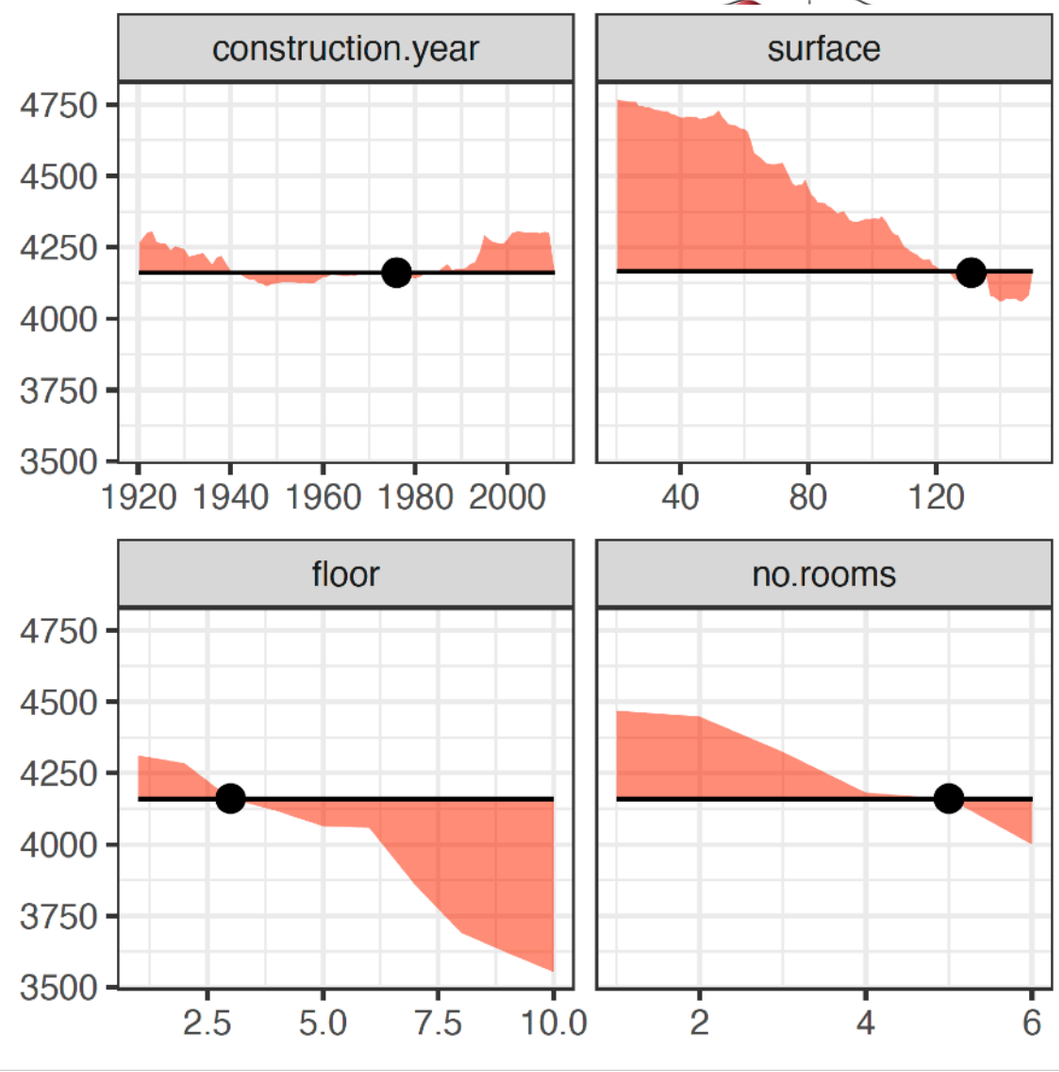
Based on Ceteris Paribus Profiles you may calculate local variable importance.

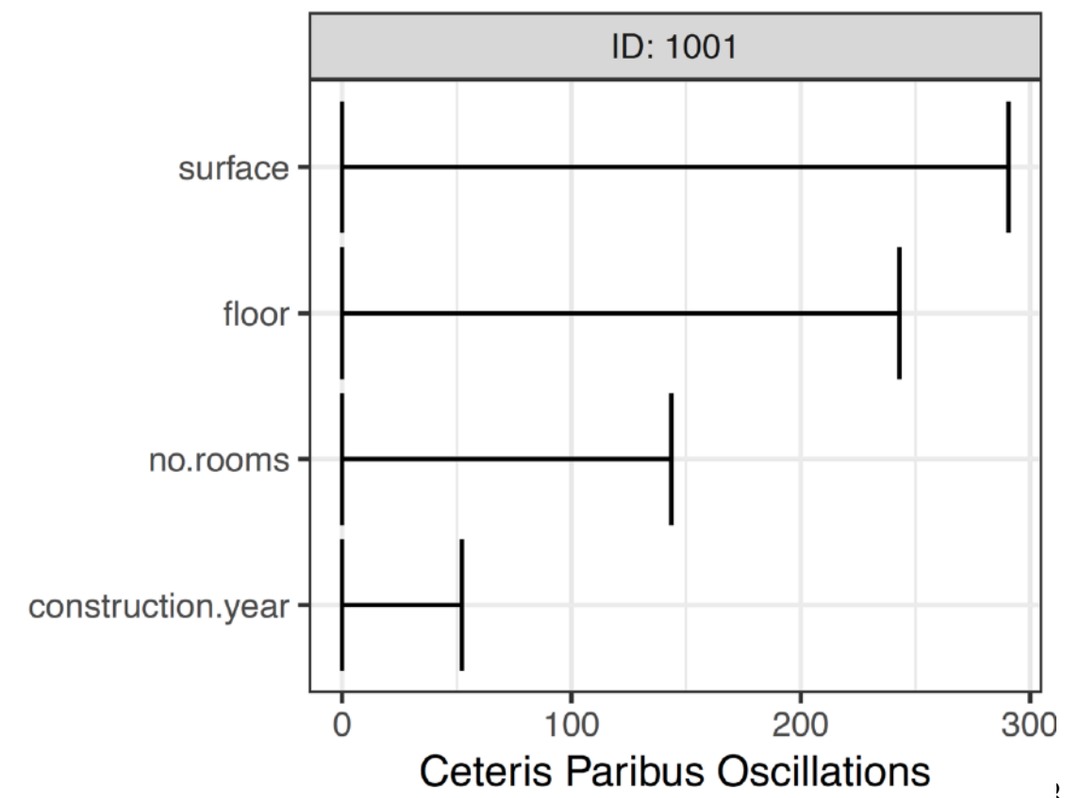
One way to do so is to integrate CP oscillations over model predictions.

$$vip_j^{CP}(x) = \int_{-\inf}^{inf} |CP^{f,j,x}(z) - f(x)|dz$$

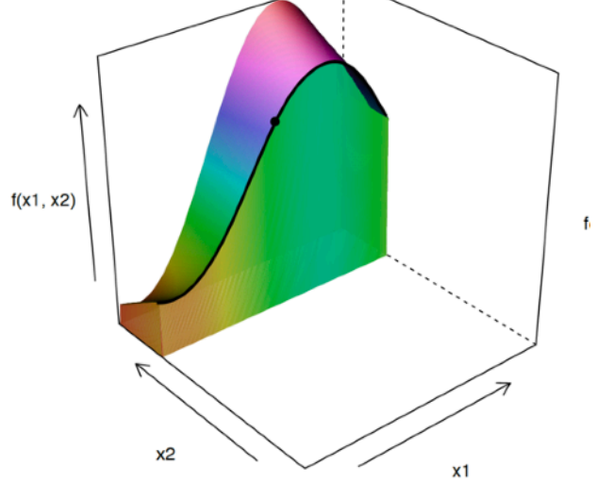
This leads to a straightforward estimator for variable importance

$$\widehat{vip_j^{CP}(x)} = \frac{1}{n}\sum_{i=1}^{n} |CP^{f,j,x}(x_i) - f(x)|$$



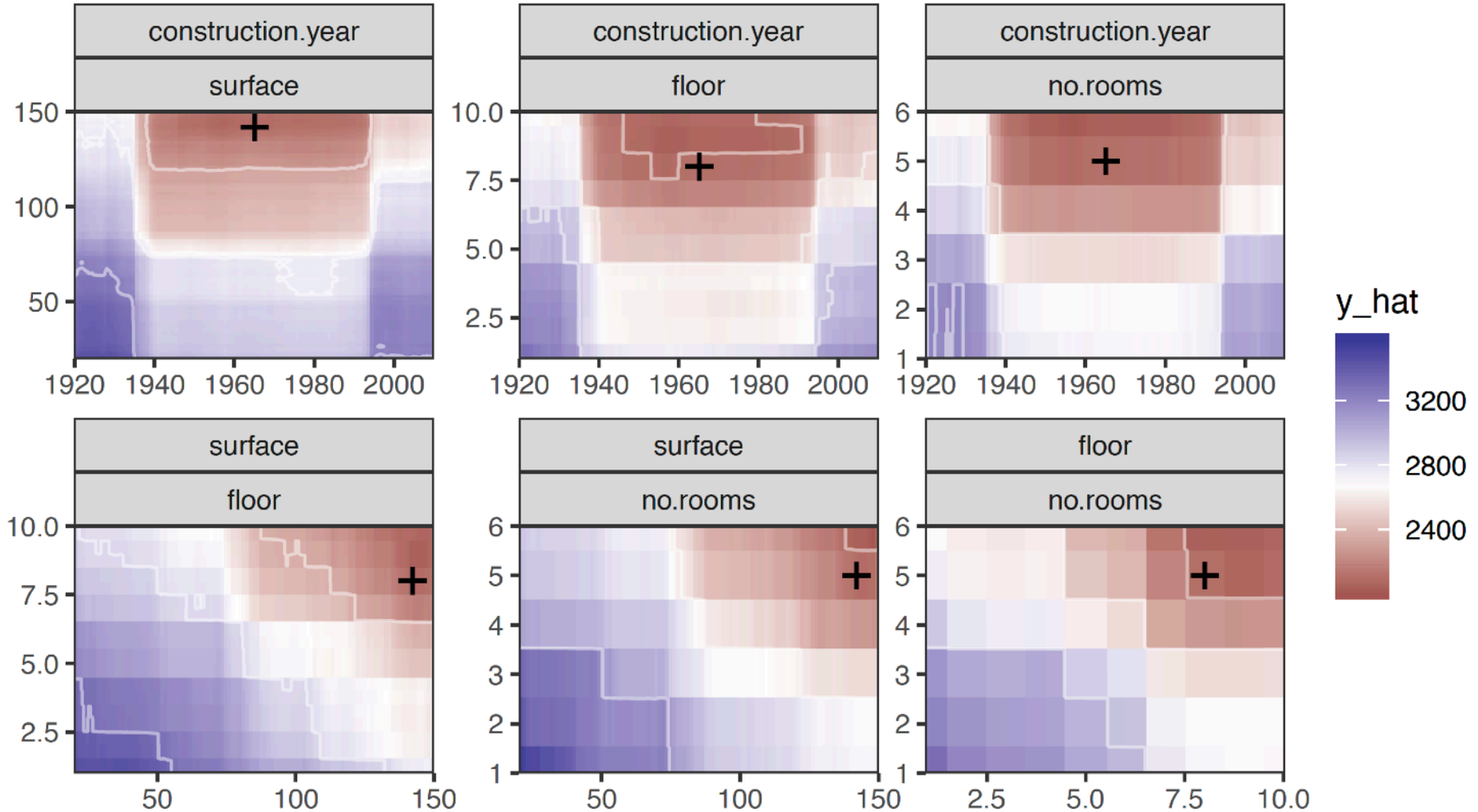
# Ceteris Paribus Profiles 2D



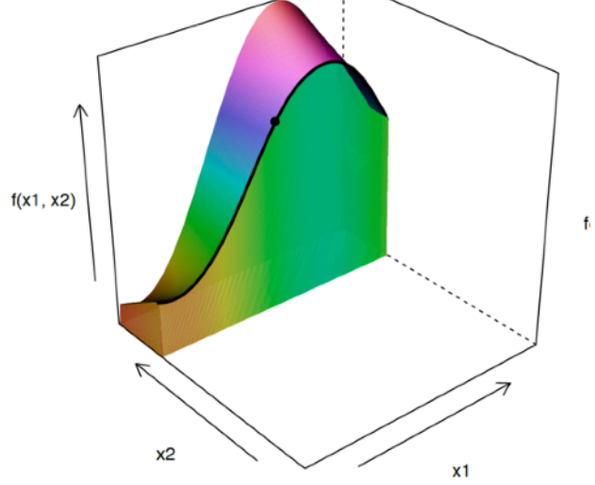

Ceteris Paribus Profiles may be easily calculated for larger number of variables.

Analysis of such profiles help to identify interaction between pairs of variables.

$$CP^{f,(j,k),x}(z_1, z_2) := f(x|^{(j,k)} = (z_1, z_2))$$

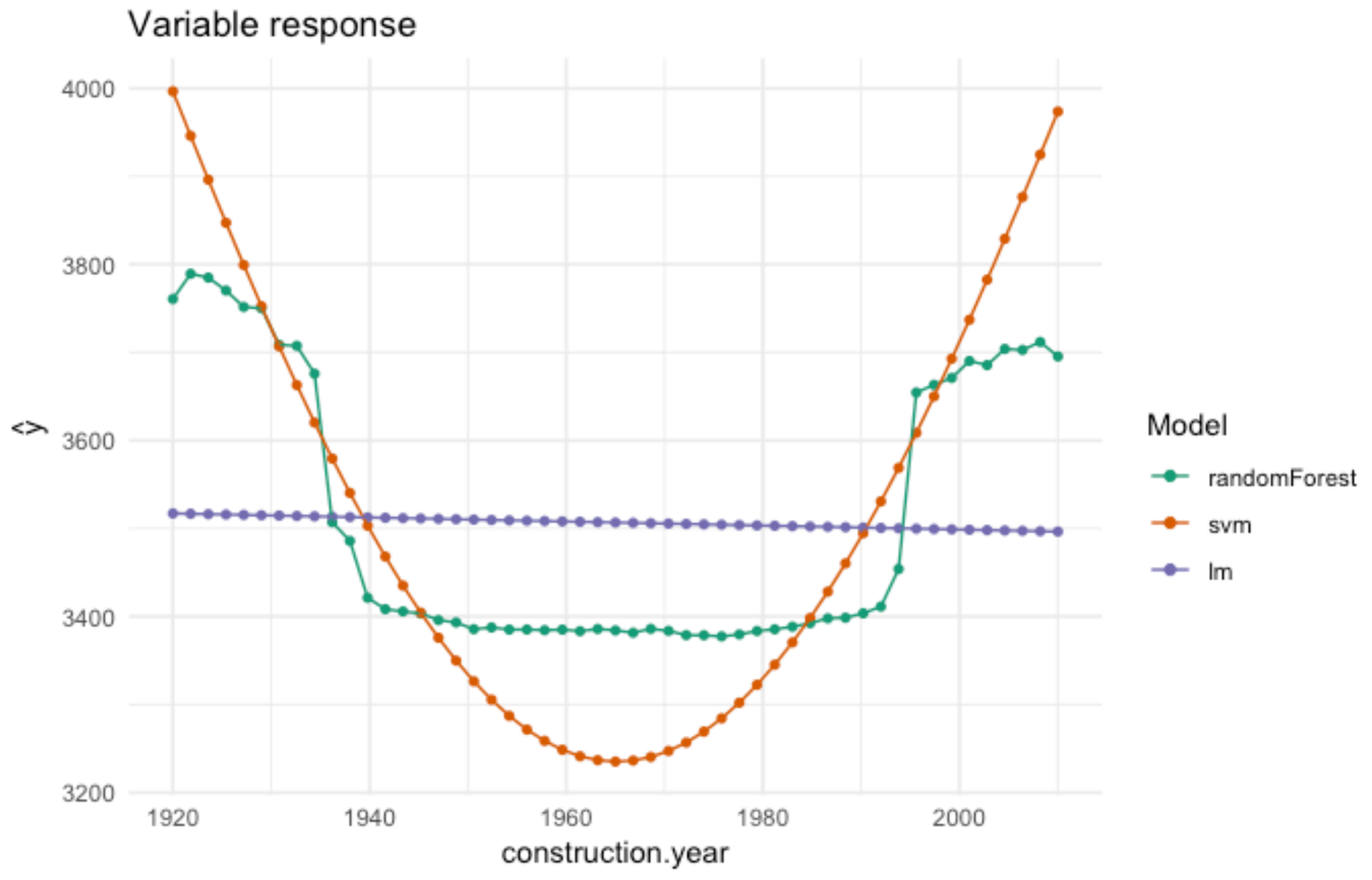


# Ceteris Paribus Profiles

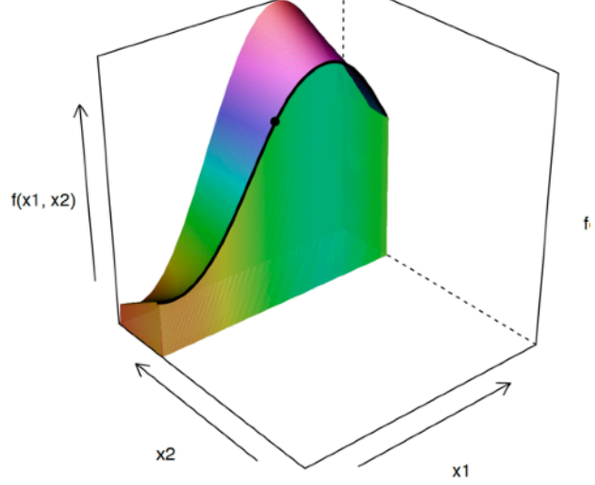



Ceteris Paribus Profiles may be used for comparisons of different models. Having few competing models one may compare their CP profiles.

Agreement (or lack of it) between competing models shows how stable are model responses.



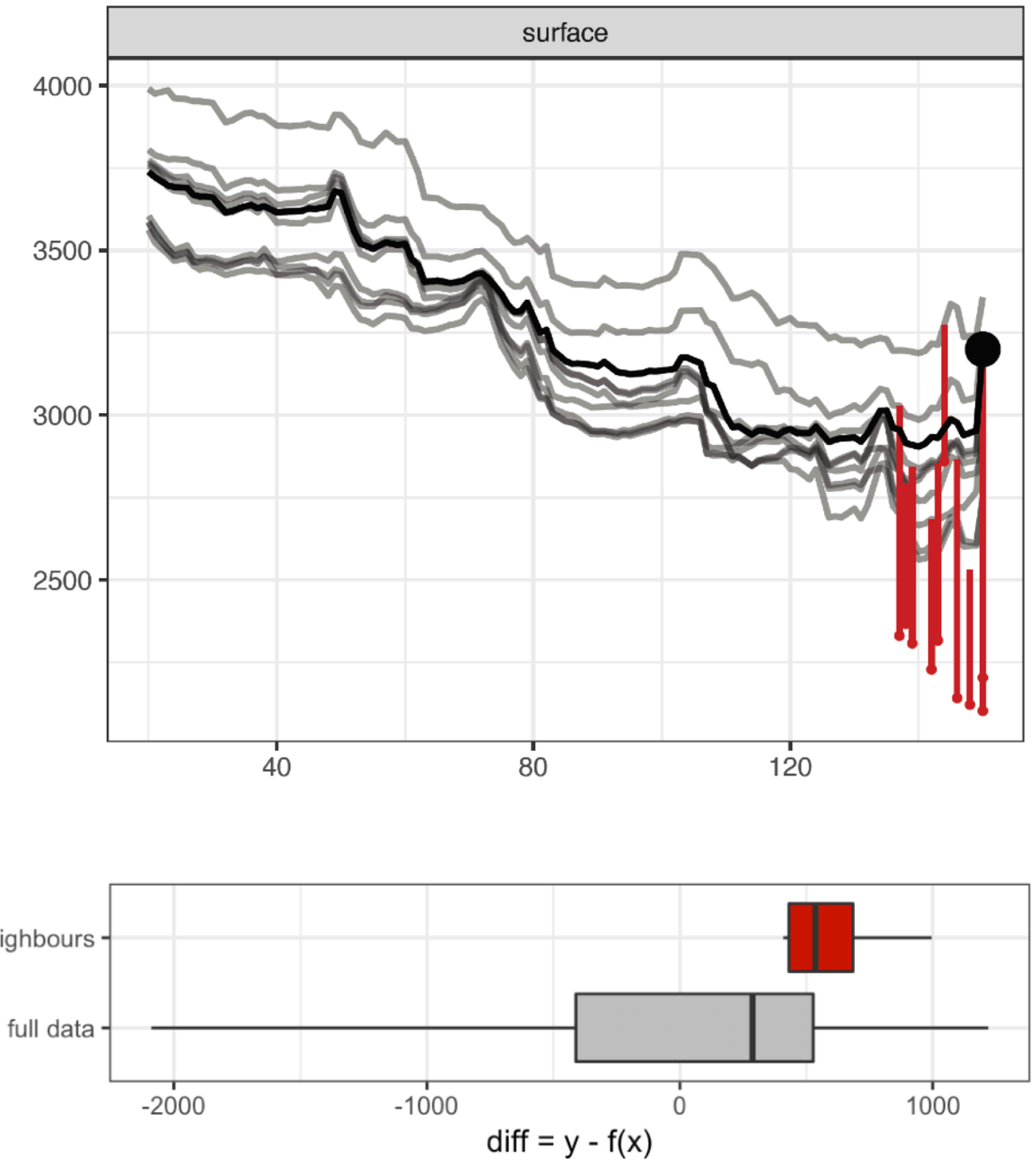
# Ceteris Paribus Profiles




Ceteris Paribus Profiles may be used for validation of local fidelity of a model.

If model predictions are biased (e.g. for Random Forest) then we may additionally perform some checks.

Analysis of profiles for labeled neighbours help to assess model smoothness and stability around the point of interest.



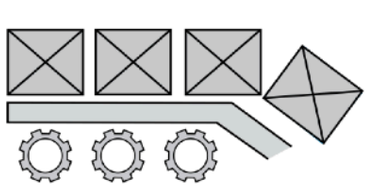
https://pbiecek.github.io/DALEX_docs/

1 Introduction | 2 Architecture of DALEX | 3 Model understanding | 4 Prediction understanding | 5 Ceteris Paribus Profiles | 6 Epilogue | 7 Exercises

# DALEX: Descriptive mAchine Learning EXplanations

## DrWhy

**Model Factory**



mlr
tidymodels
caret

**Data**

training data
test data
validation data
next release
next application

**DALEX2**

model preprocessing
explain()
predict()
loss_function()

Local explainers

**ceterisParibus2**

What If profiles
1d, 2d profiles
local importance with oscillations

**breakDown2**

Sequential variable attribution, interactions, Shapley Values

**localModel**

interpretable features for tabular data

Global explainers

**VarImp**

variable importance
model performance
model comparisons

**SAFE**

Surrogate feature extraction for white box models

**response**

PDP
AlePlot

Auditors

**auditor**

model diagnostic
validation of model assumptions
performance plots

**localAuditor**

audit for local explainers

# modelDown: pkgdown for models

https://github.com/MI2DataLab/modelDown

## modelDown

build passing

`modelDown` generates a website with HTML summaries for predictive models. Is uses DALEX explainers to compute and plot summaries of how given models behave. We can see how exactly scores for predictions were calculated (Prediction BreakDown), how much each variable contributes to predictions (Variable Response), which variables are the most important for a given model (Variable Importance) and how well out models behave (Model Performance).

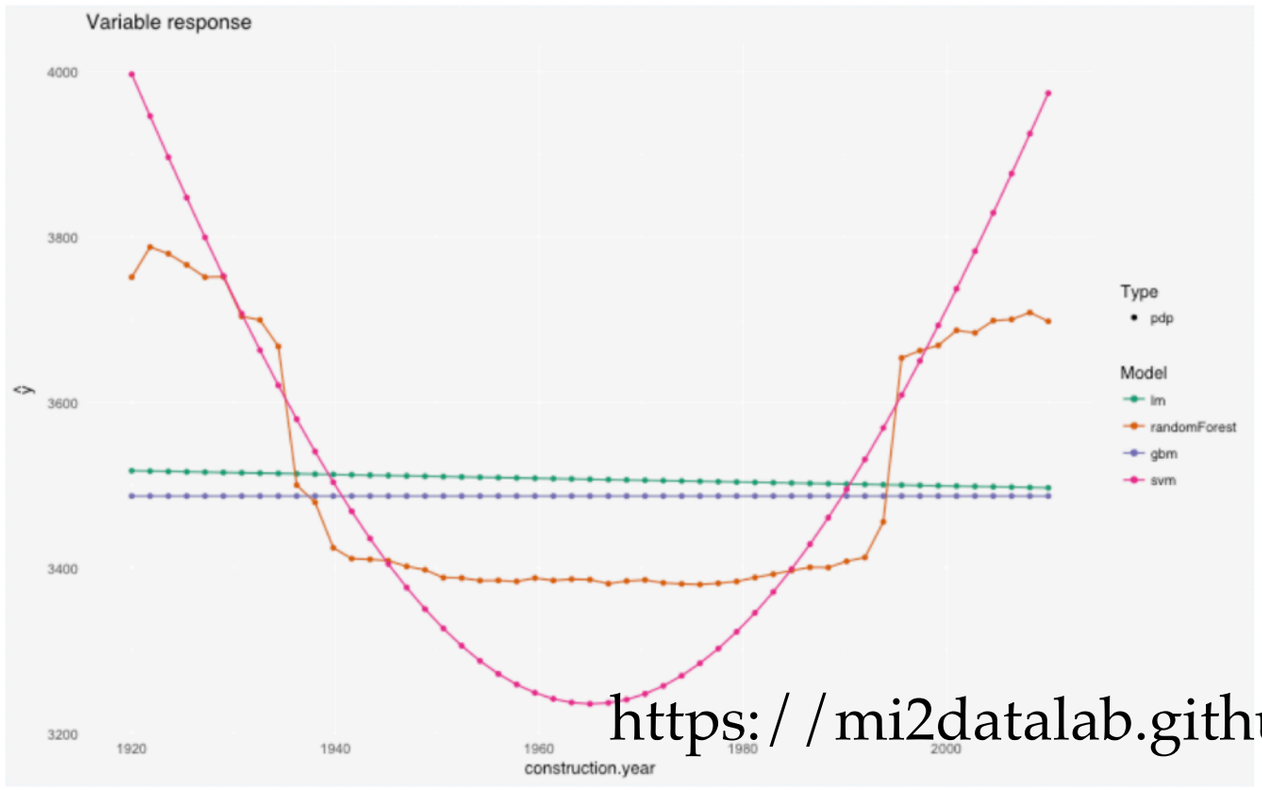
`pkgdown` documentation: https://mi2datalab.github.io/modelDown/

An example website for regression models: https://mi2datalab.github.io/modelDown_example/

modelDown | Model Performance | Variable Importance | Variable Response | Prediction BreakDown

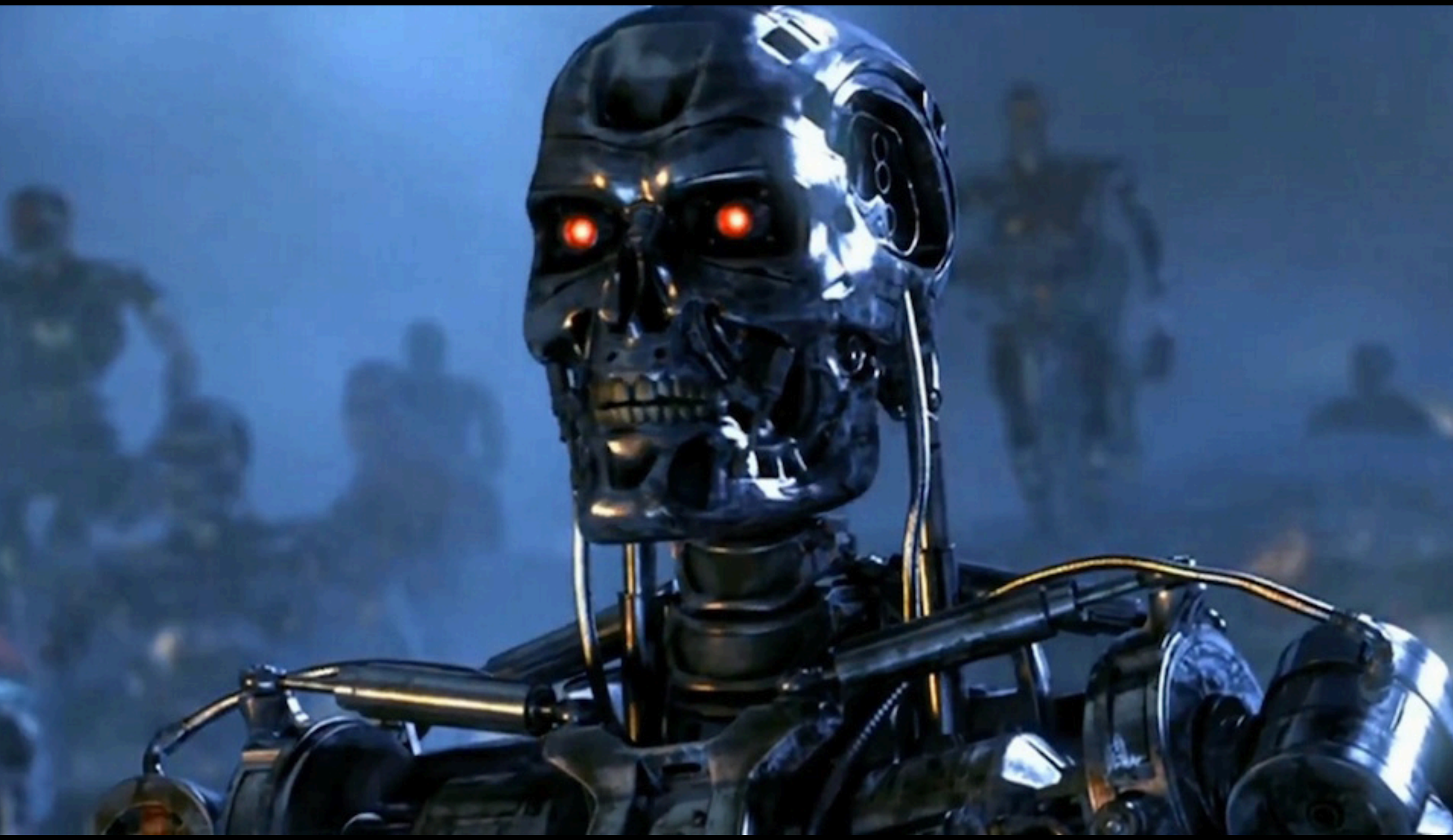
- construction.year
- district
- floor
- no.rooms
- surface

### construction.year




https://mi2datalab.github.io/modelDown_example/

# Machine Learning Models will replace humans



# Machine Learning Models will empower humans

# Find more at https://github.com/pbiecek/DALEX

] | https://github.com/pbiecek/DALEX

# DALEX


CRAN 0.2.3 | downloads 1841/month | downloads 4050 | build passing | coverage 92%

## DALEX: Descriptive mAchine Learning EXplanations

Machine Learning models are widely used and have various applications in classification or regression tasks. Due to increasing computational power, availability of new data sources and new methods, ML models are more and more complex. Models created with techniques like boosting, bagging of neural networks are true black boxes. It is hard to trace the link between input variables and model outcomes. They are use because of high performance, but lack of interpretability is one of their weakest sides.

In many applications we need to know, understand or prove how input variables are used in the model and what impact do they have on final model prediction. `DALEX` is a set of tools that help to understand how complex models are working.

Find more about DALEX in this Gentle introduction to DALEX with examples.

## DALEX Stories

- An interactive notebook with examples: launch binder

### How to use DALEX

- How to use DALEX with caret
- How to use DALEX with mlr
- How to use DALEX with H2O
- How to use DALEX with xgboost package
- How to use DALEX for teaching. Part 1
- How to use DALEX for teaching. Part 2
- breakDown vs lime vs shapleyR

# Take aways

- Modelling is not only about performance, it’s also about interpretability,
- Understand what you are doing (this may require math), do not just glue random snippets of code,
- Do no harm.