Explain!
Or I will sue you!

Przemysław Biecek

Warsaw R Users Group
20 . 02 . 2019

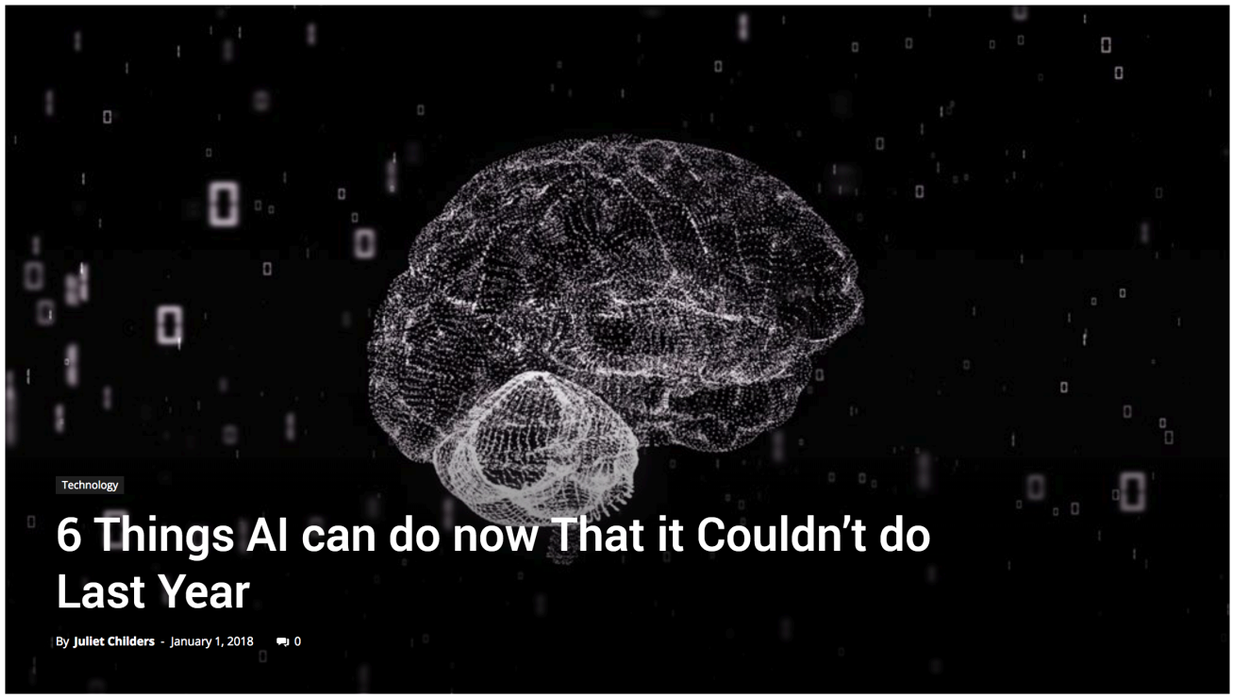
59 impressive things artificial intelligence can do today

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

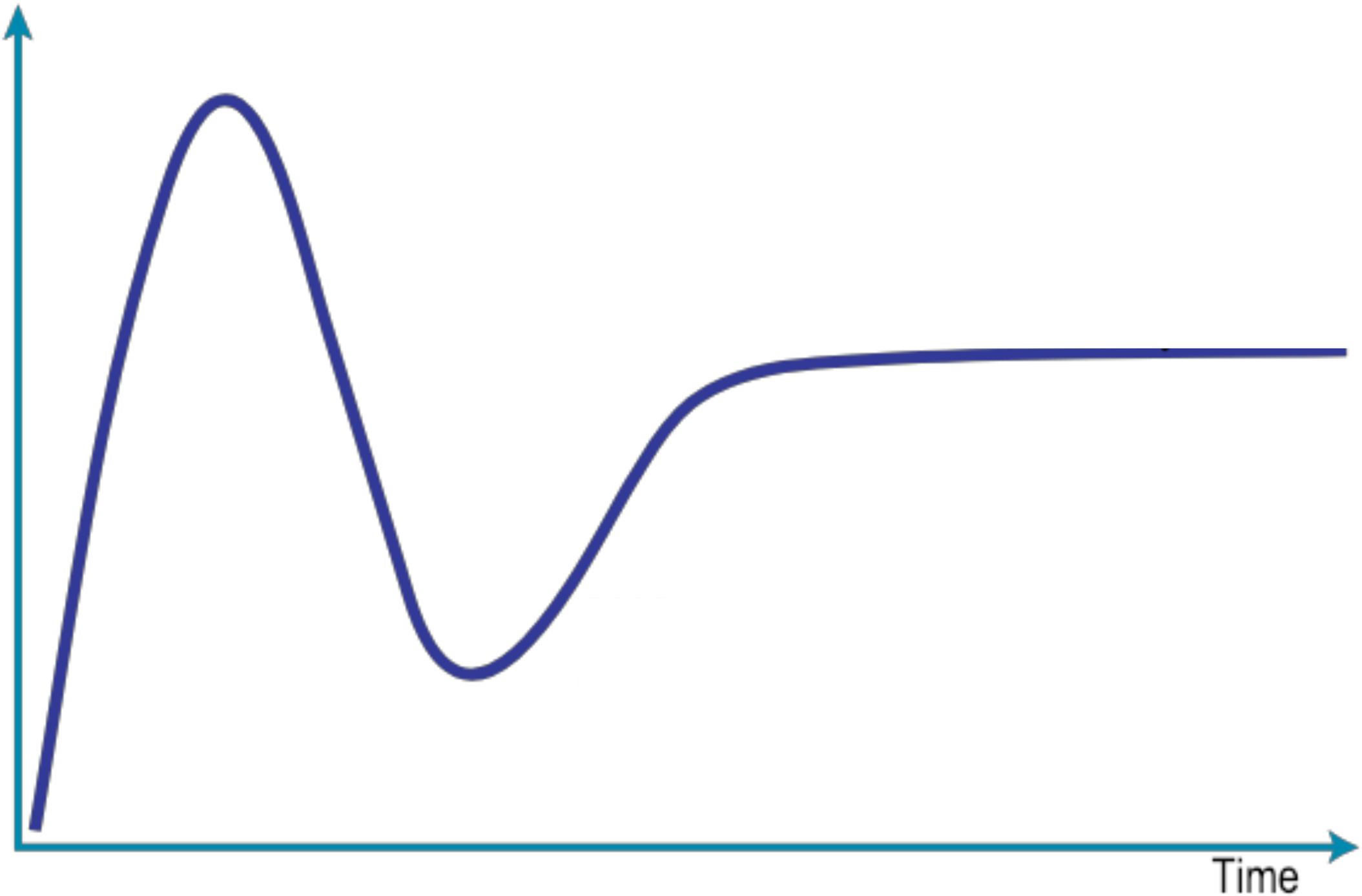
Major goals

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

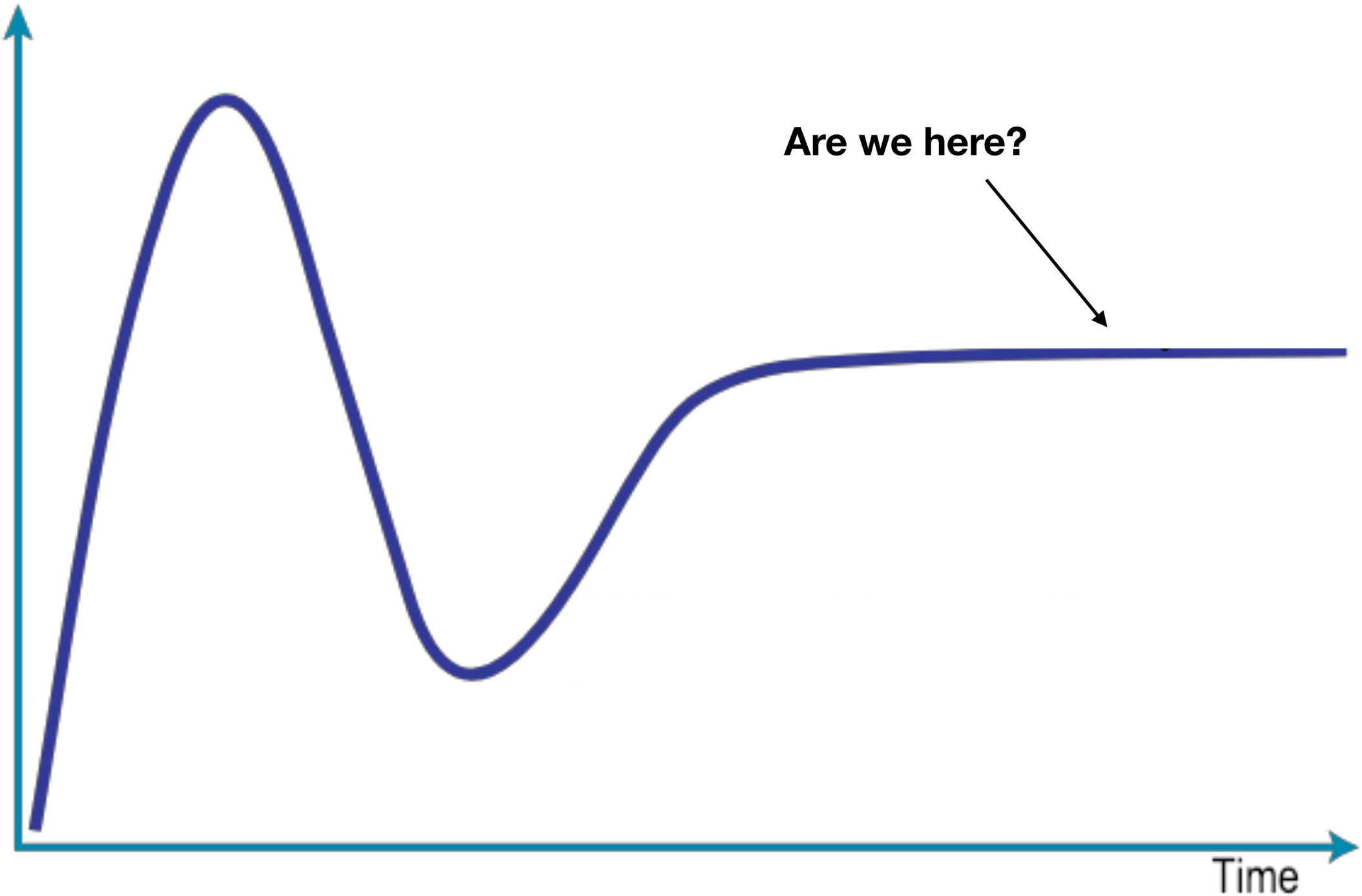
Approaches

Symbolic

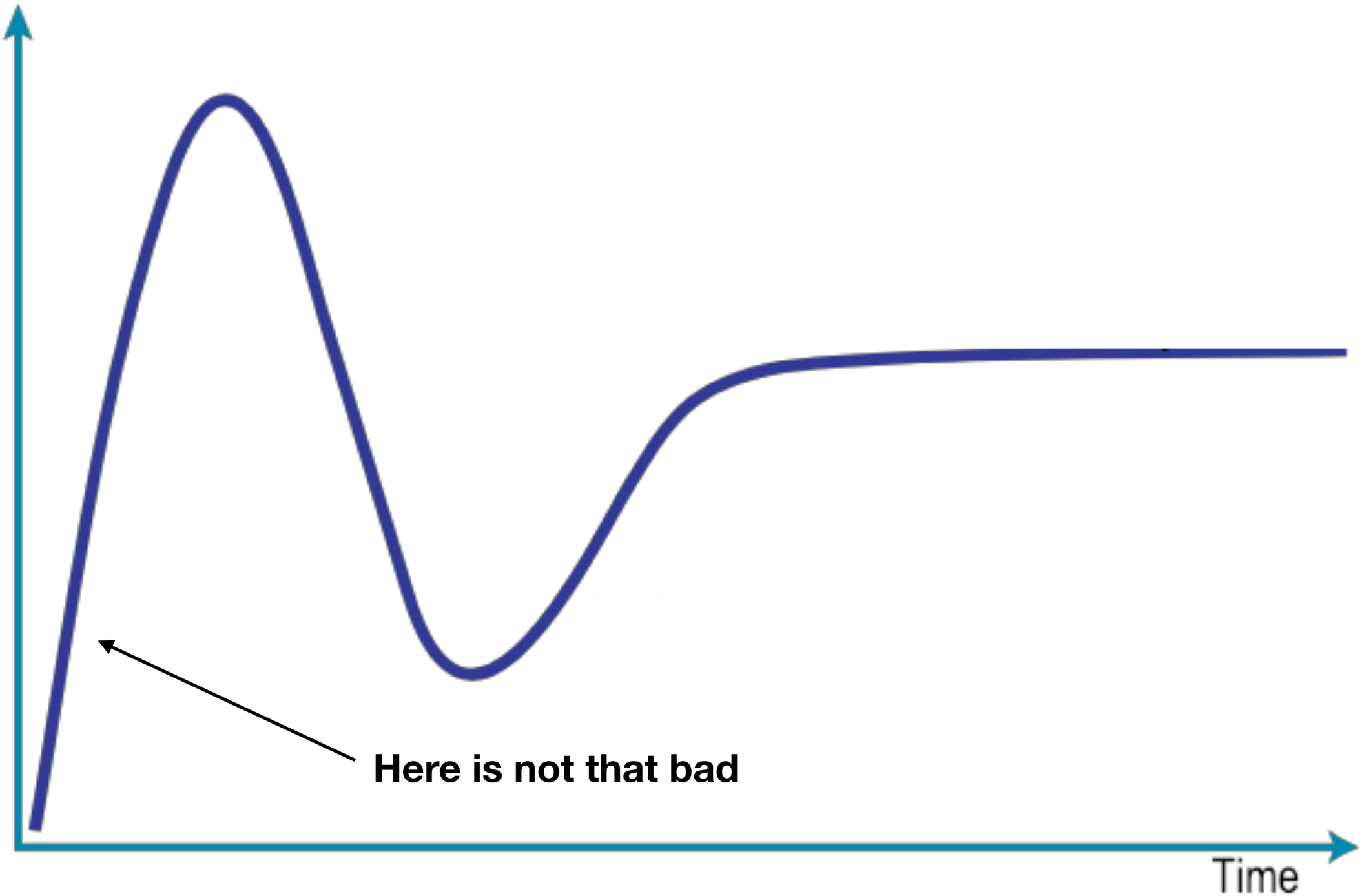
Hype Cycle for Predictive Models



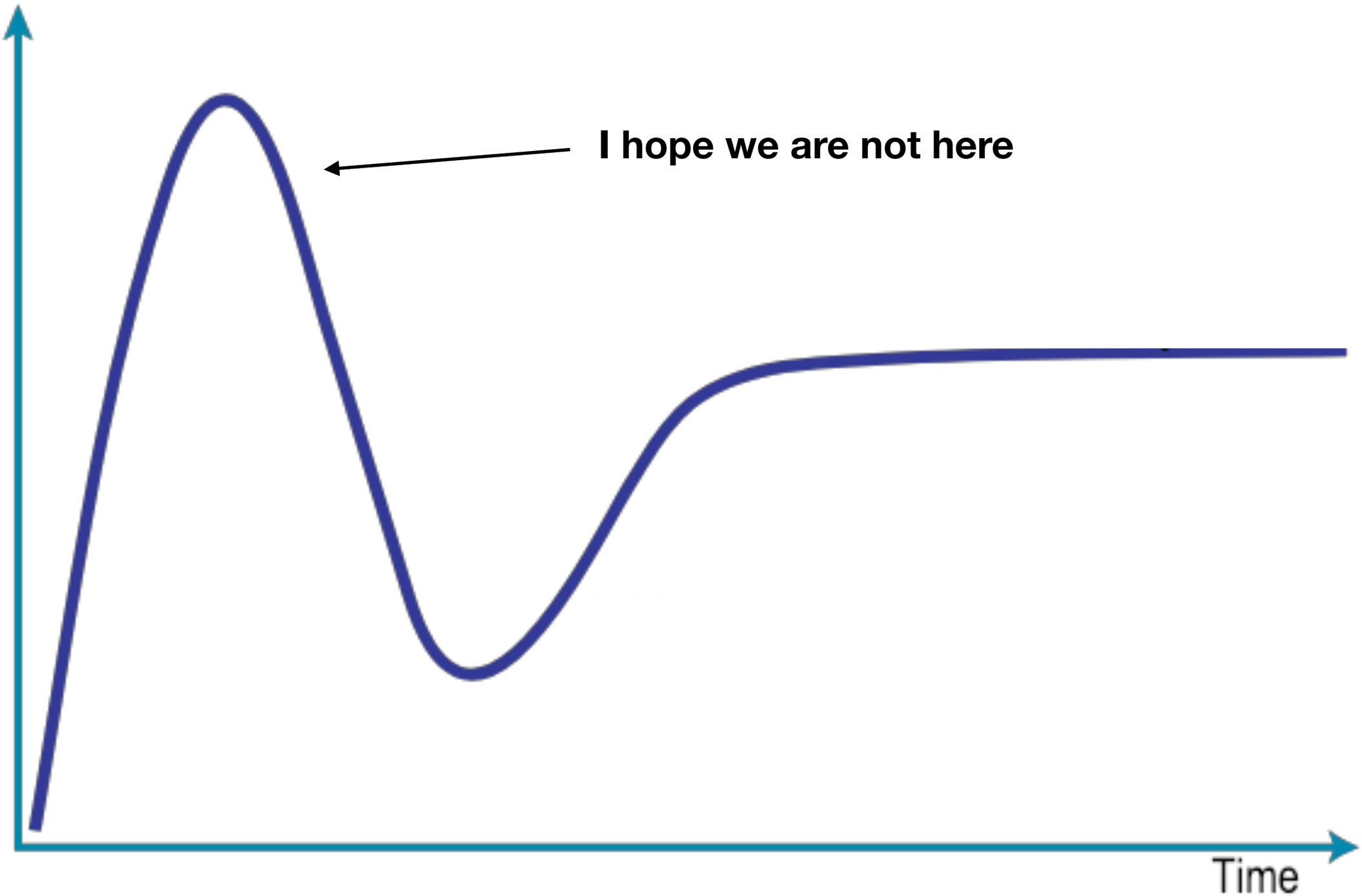
Hype Cycle for Predictive Models



Hype Cycle for Predictive Models



Hype Cycle for Predictive Models



Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offered for declared research purposes.[2]

Google Flu Trends data, South Africa

Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offerec

Support The Guardian

Contribute → Subscribe →

Search jobs | Sign in | Search

News | Opinion | Sport | Culture | Lifestyle | More

World UK Science Cities Global development Football **Tech** Business Environment Obituaries

Google

Google Flu Trends is no longer good at predicting flu, scientists find

Researchers warn of 'big data hubris' and the importance of updating analytical models, claiming Google has made inaccurate forecasts for 100 of 108 weeks

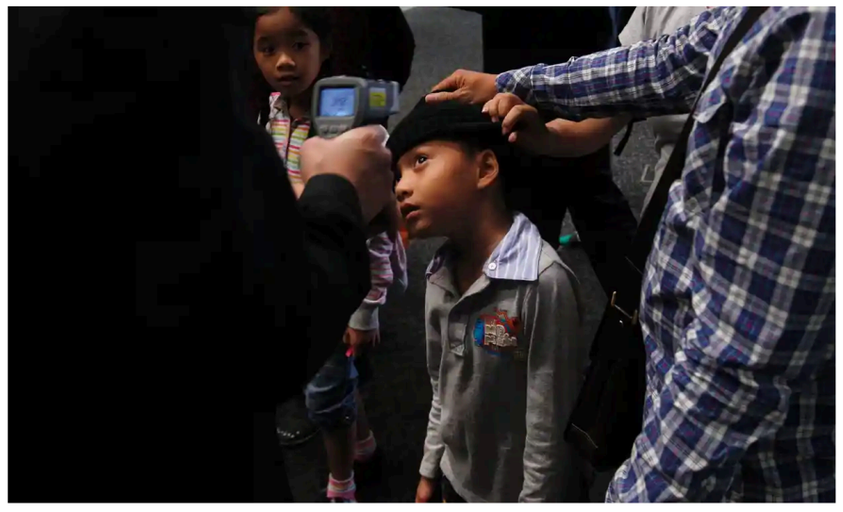
Charles Arthur

@charlesarthur

Thu 27 Mar 2014 10.27 GMT

200 7

This article is over **4 years old**



▲ Airport security personnel take a body temperature reading of a boy as he arrives at Hong Kong International Airport April 9 2013, following concerns over a deadly strain of bird flu. Photograph: Tyrone Siu/Reuters

Forbes | Billionaires | Innovation | Leadership | Money | Con

61,215 views | Mar 23, 2014, 09:00am

Why Google Flu Is A Failu

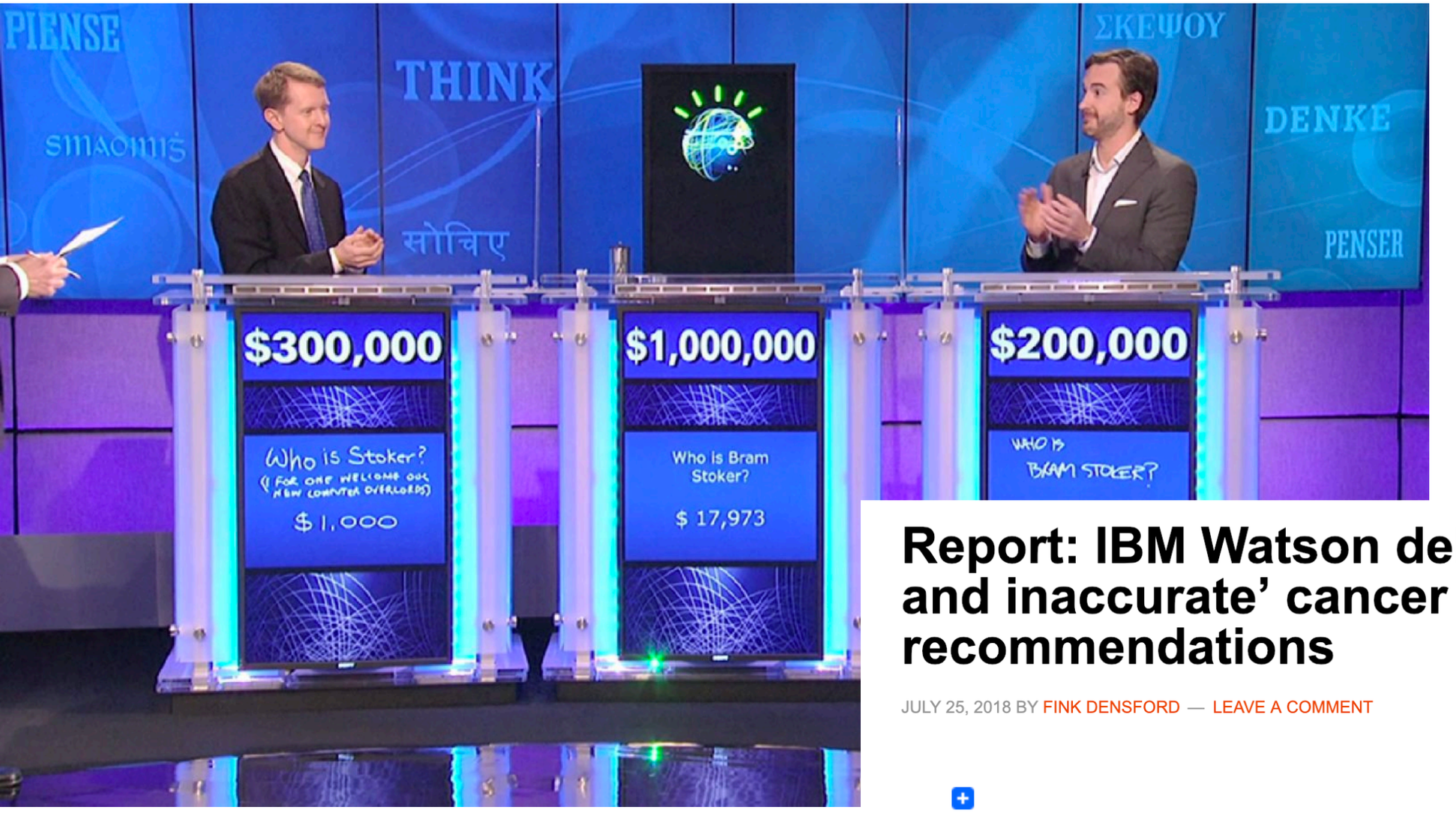
Steven Salzberg Contributor

Pharma & Healthcare

It seemed like such a good idea at the time.

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
सोचिए
PENSER
$300,000
Who is Stoker?
((FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS))
$1,000
$1,000,000
Who is Bram Stoker?
$ 17,973
$200,000
WHO IS BRAM STOKER?
$5600



Report: IBM Watson delivered ‘unsafe and inaccurate’ cancer recommendations

JULY 25, 2018 BY FINK DENSFORD — LEAVE A COMMENT

Internal documents from IBM Watson Health (NYSE:IBM) indicate that the company’s Watson for Oncology product often returns “multiple examples of unsafe and incorrect treatment recommendations,” according to a new report from *STAT News*.

The documents come from slides presented last year by IBM Watson Health’s deputy chief health officer, according to the report, and include feedback from customers that indicated the product is “often inaccurate” and that its recommendations bring to light “serious questions about the process for building content and the underlying technology.”

The issues were blamed on training the Watson product received by IBM engineers and physicians at the Memorial Sloan Kettering Cancer Center, which included “synthetic,” or hypothetical patients and cases, instead of real patient data, *STAT* reports.

https://www.massdevice.com/report-ibm-watson-delivered-unsafe-and-inaccurate-cancer-recommendations/

COMPAS (software)

From Wikipedia, the free encyclopedia

COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool developed by Northpointe (now equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

Contents [hide]

Risk Assessment [edit]

The COMPAS software uses an algorithm to assess potential recidivism risk. Northpointe created risk scales for general and violent recidivism, and for pretrial misconduct. According to the COMPAS Practitioner's Guide, the scales were designed using behavioral and psychological constructs "of very high relevance to recidivism and criminal careers."[3]

COMPAS (software)

From Wikipedia, the free encyclopedia

COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool developed by Northpointe (now equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

Contents [hide]

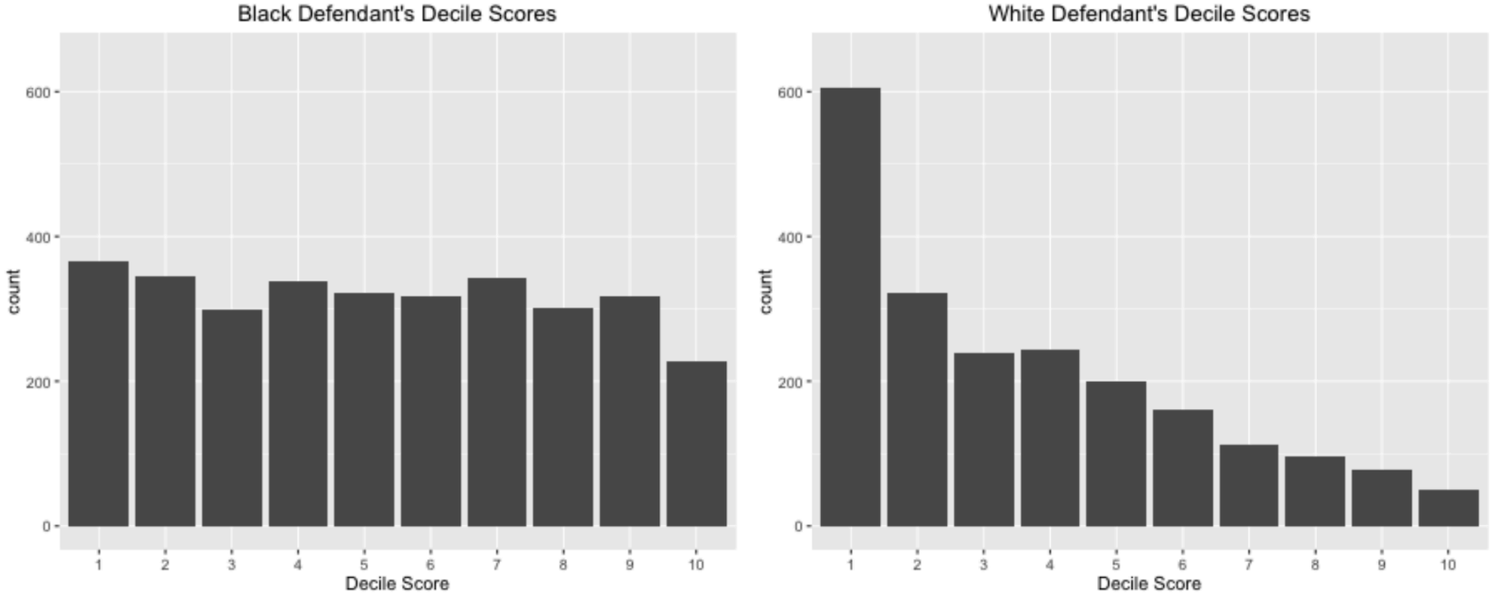
Risk Assessment [edit]

The COMPAS software uses an algorithm to assess poten
scales for general and violent recidivism, and for pretrial r
Practitioner's Guide, the scales were designed using beha
high relevance to recidivism and criminal careers."[3]

Analysis

We analyzed the COMPAS scores for "Risk of Recidivism" and "Risk of Violent Recidivism." We did not analyze the COMPAS score for "Risk of Failure to Appear."

We began by looking at the risk of recidivism score. Our initial analysis looked at the simple distribution of the COMPAS decile scores among whites and blacks. We plotted the distribution of these scores for 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.



These histograms show that scores for white defendants were skewed toward lower-risk categories, while black defendants were evenly distributed across scores. In our two-year sample, there were 3,175 black defendants and 2,103 white defendants, with 1,175 female defendants and 4,997 male defendants. There were 2,809 defendants who recidivated within two years in this sample.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

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

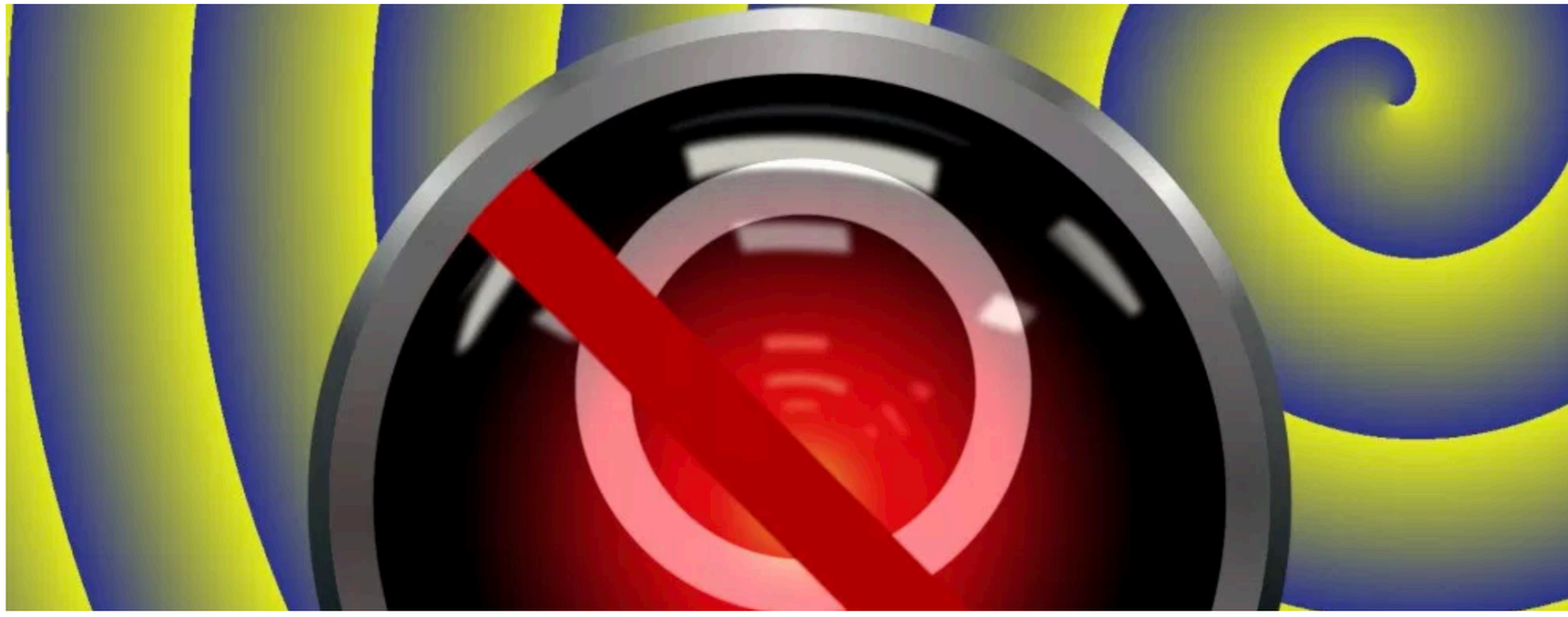
Amazon scraps secret AI showed bias against wom

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

> 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.

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



TECH AMAZON ARTIFICIAL INTELLIGENCE

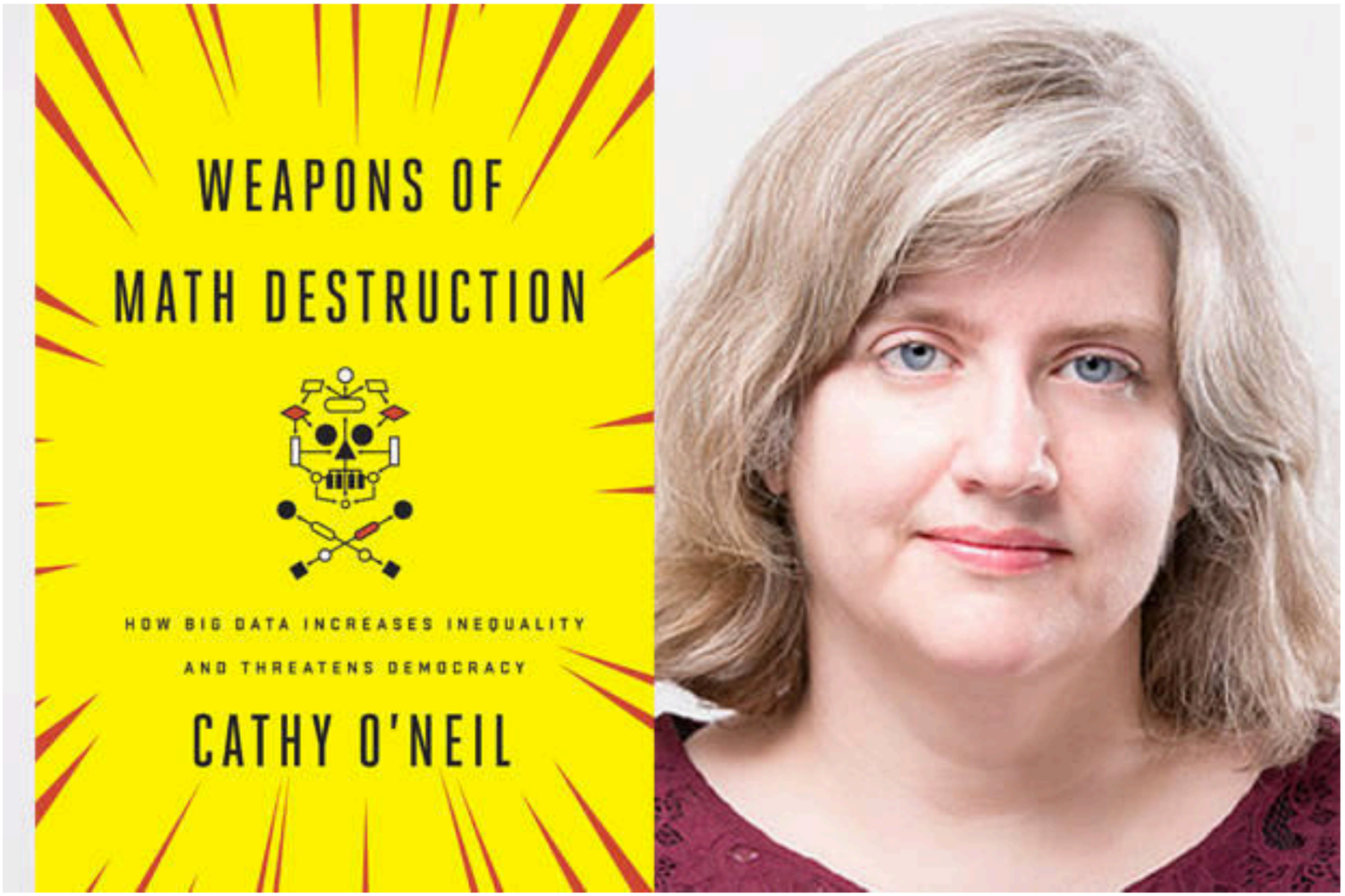
21

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

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

Cathy O'Neil:

The era of blind faith in ~~big data~~ machine learning must end

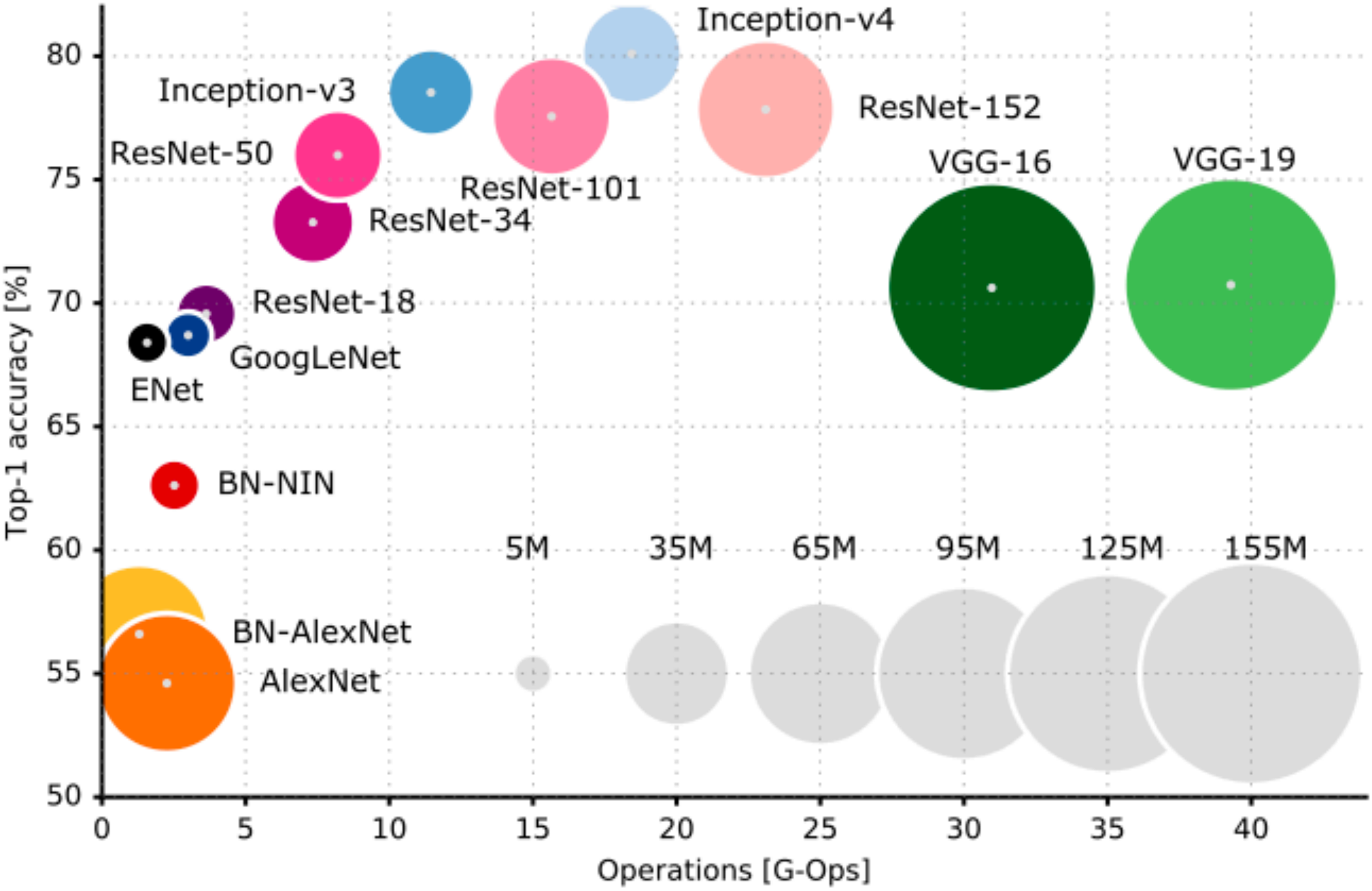


- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of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

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

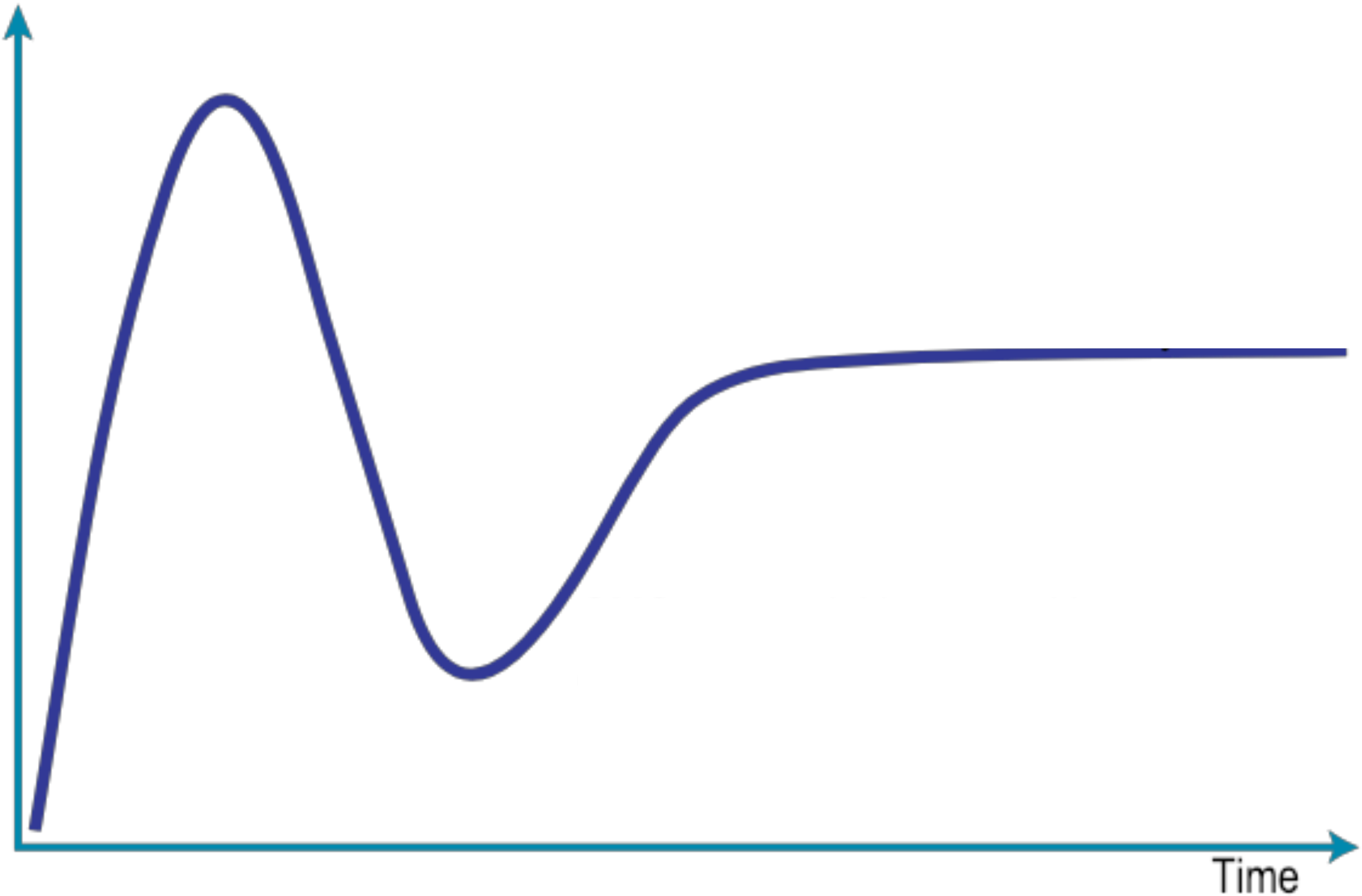
https://en.wikipedia.org/wiki/Social_Credit_System

Black boxes are even more opaque

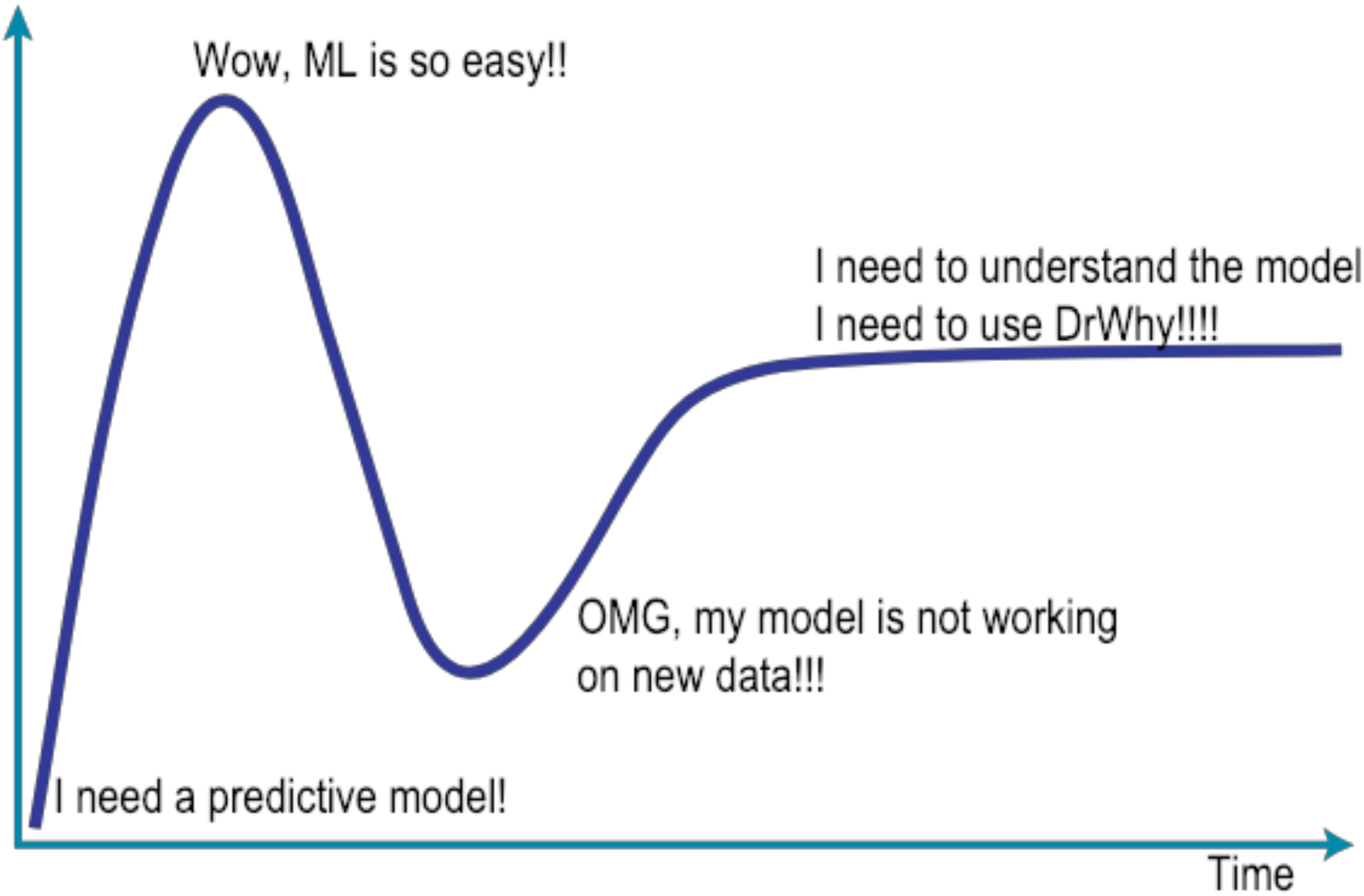


https://www.topbots.com/a-brief-history-of-neural-network-architectures/

Hype Cycle for Predictive Models



Hype Cycle for Predictive Models



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 explanation** (**explanation**) is a right to be given an explanation for an output of the algorithm. Such rights primarily refer to individual right explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a person who a is denied may ask for an explanation, which could be "Credit bureau X reports that you declared bankruptcy last year; this is 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 debate.

Contents [hide]

Prawo do wyjaśnienia decyzji kredytowej dla każdego! Sukces Panoptykonu!

13.02.2019

Po dzisiejszym posiedzeniu połączonych komisji sejmowych mamy dobre wiadomości: udało się przekonać rząd i posłów do wprowadzenia zmian w prawie bankowym, które zmienią proces przyznawania kredytów. W miejsce „czarnej skrzynki", która wypluwa niezrozumiałe decyzje bez uzasadnienia, ma się pojawić przejrzysta procedura, wzmacniająca prawa osób ubiegających się o kredyt. To duży przełom w relacjach bank–klient i odpowiedź ustawodawcy na postulaty, które zgłaszamy od dawna.

Krótka historia długiej walki

Od momentu, w którym pojawił się rządowy projekt tzw. ustawy sektorowej (wdrażającej RODO w 160 ustawach) i propozycje zmian w prawie bankowym, walczyliśmy o to, żeby każdy konsument starający się o kredyt miał prawo zażądać od banku wyjaśnienia oceny zdolności kredytowej i decyzji, którą w jego sprawie podjął bank. W pierwszej wersji projektu było przewidziane jedynie szczątkowe i nieprecyzyjnie sformułowane prawo do wyjaśnienia, dotyczące wyłącznie decyzji podejmowanych w sposób automatyczny, czyli bez udziału człowieka. Uznaliśmy, że to za mało, i walczyliśmy o więcej. Tym bardziej, że od kilku lat prawo do wyjaśnienia oceny zdolności kredytowej (w każdej sytuacji) mają przedsiębiorcy.

Na pierwszym posiedzeniu specjalnej podkomisji powołanej do prac nad tym pakietem zaapelowaliśmy do rządu i posłów o wprowadzanie dalej idących gwarancji dla konsumentów. Ten pomysł poparli posłowie wszystkich partii, a przewodniczący podkomisji Edward Siarka wezwał Ministerstwo Cyfryzacji do wypracowania kompromisu ze „stroną społeczną" i przedstawicielami banków. Negocjacje i praca nad brzmieniem przepisów zajęły trzy tygodnie, ale warto było poczekać.

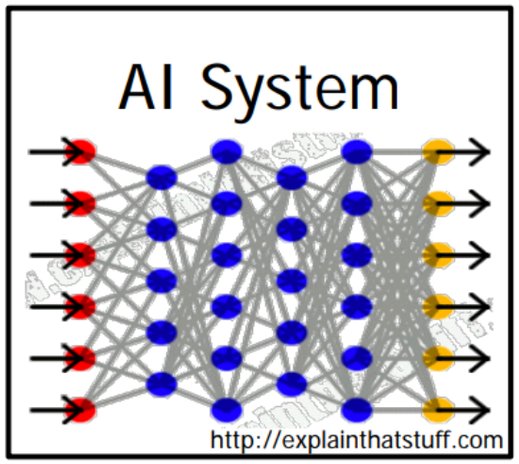
Koniec z „czarną skrzynką" przy udzielaniu kredytów

Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe,

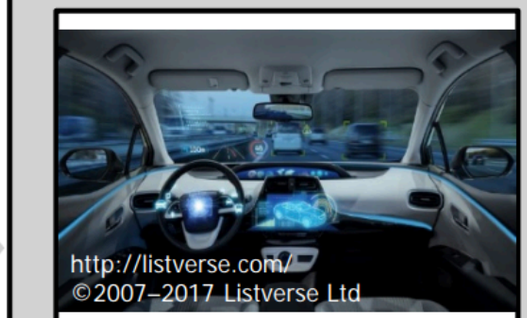
Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe, które miały wpływ na ocenę zdolności kredytowej. To kwintesencja zgłoszonej dzisiaj przez rząd i popartej przez posłów ze wszystkich ugrupowań zmiany w prawie bankowym.

https://panoptykon.org/wiadomosc/prawo-do-wyjasnienia-decyzji-kredytowej-dla-kazdego-sukces-panoptykonu

The Need for Explainable AI



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



Transportation

Finance

Security

Legal

Medicine

Military

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

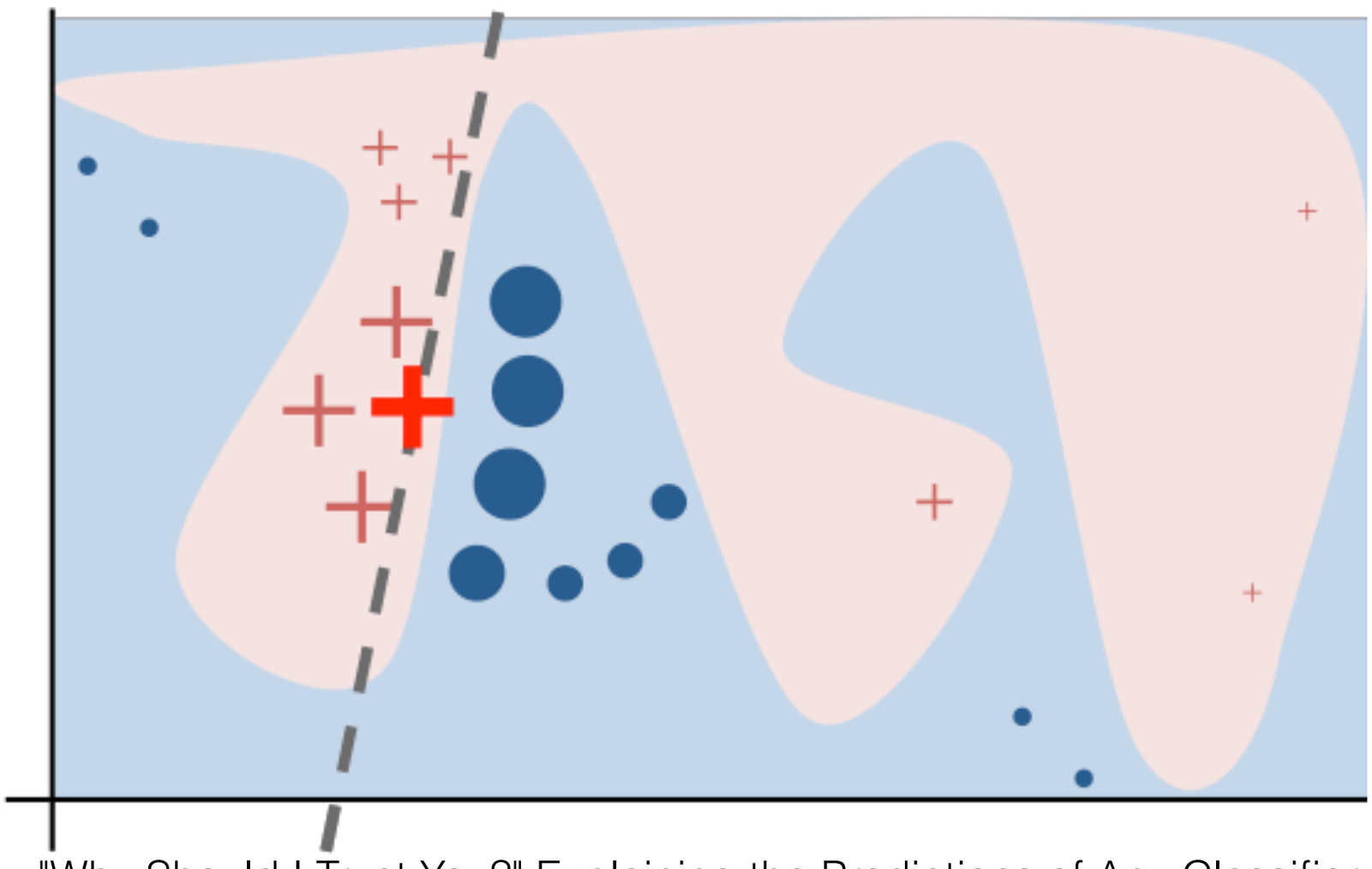
Agenda

- LIME, what it is and why it does not work
- SHAP, what it is and why it does not work
- Break Down, what it is ……..

Local Model approximations

LIME - Local Interpretable Model-Agnostic Explanations

Locally approximate the complex black-box model with an easier to interpret white-box model constructed on interpretable features.



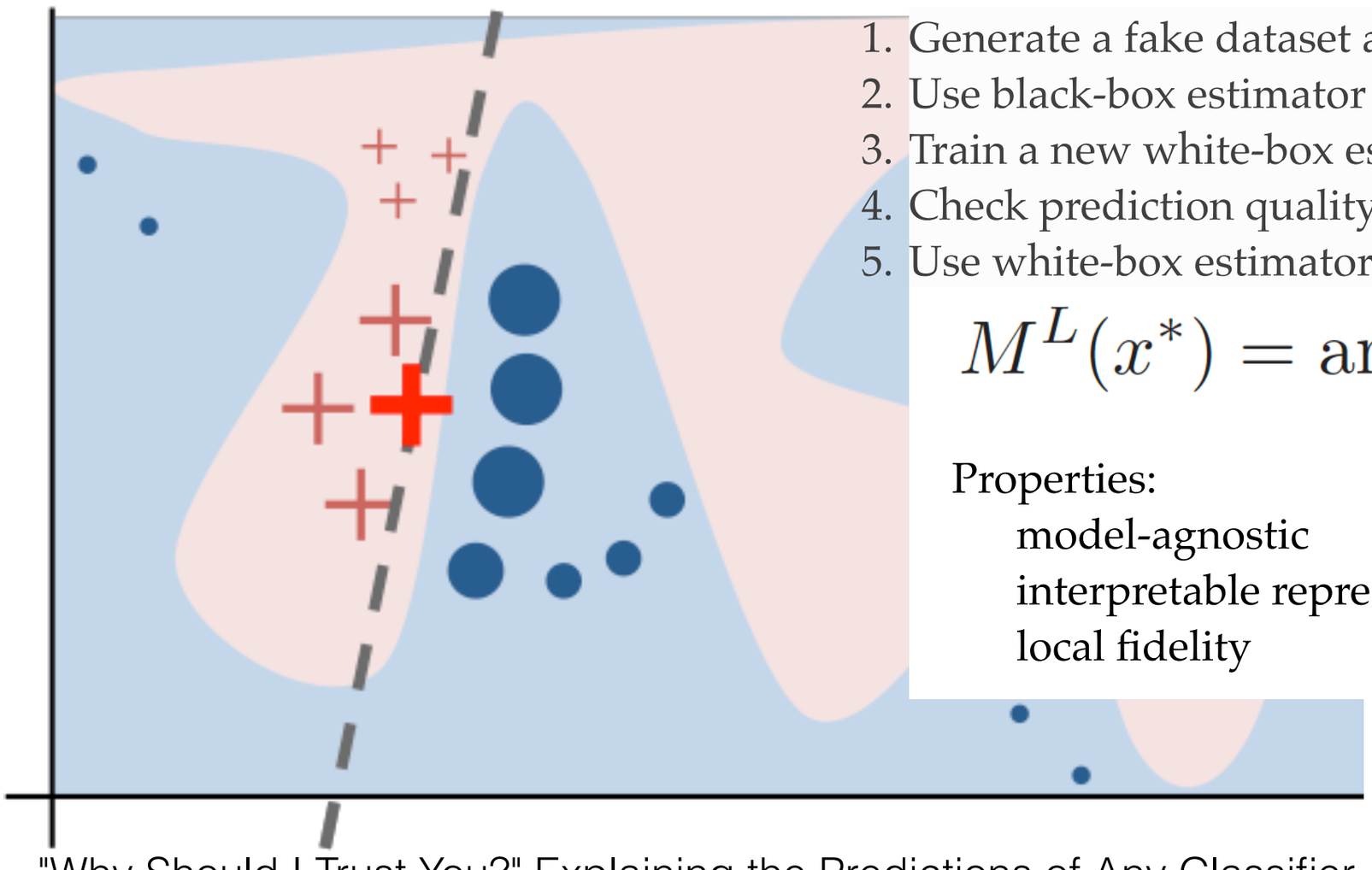
"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

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

Local Model approximations

LIME - Local Interpretable Model-Agnostic Explanations

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.

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

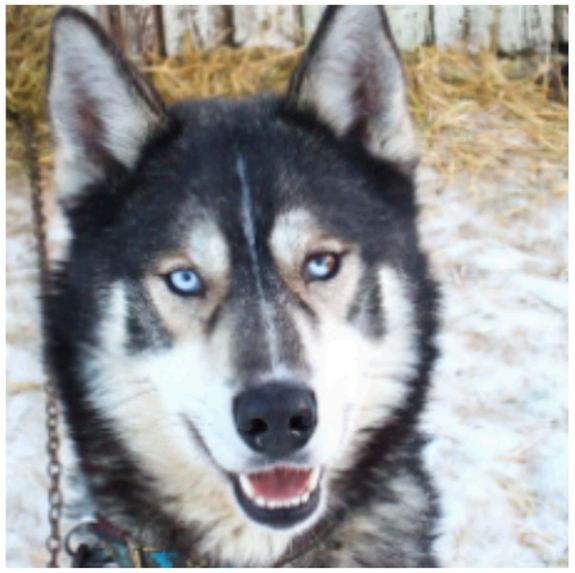
Properties:

- 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)

LIME
Pros and Cons



(a) Husky classified as wolf

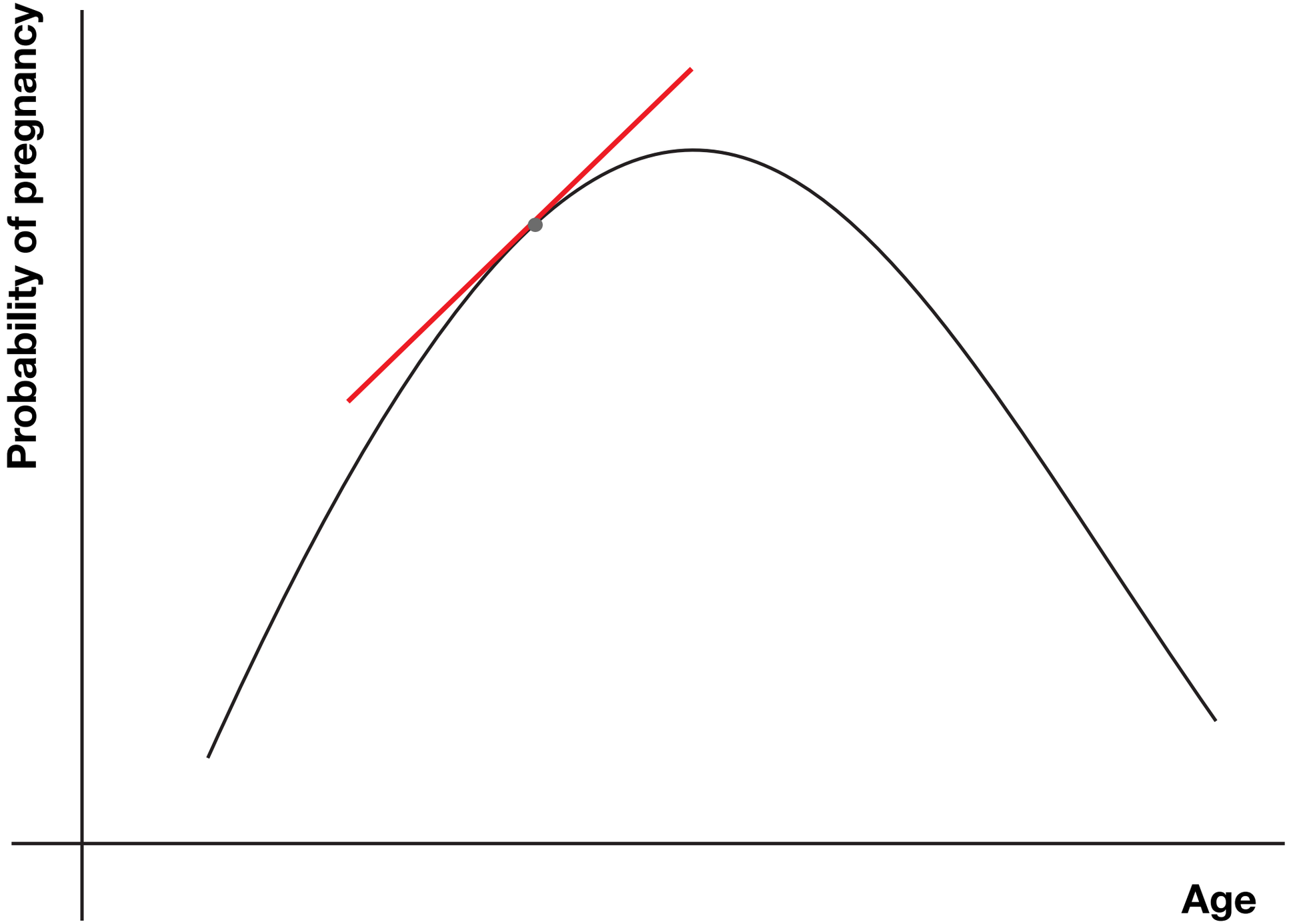
(b) Explanation

Pros

- Compelling idea
- Spectacular applications for images
- Good performance for text (but anchors are better)

Cons

- Hard to find interpretable features for tabular data
  (look for localModel and Mateusz Staniak)
- Local approximation for raw features may be misleading

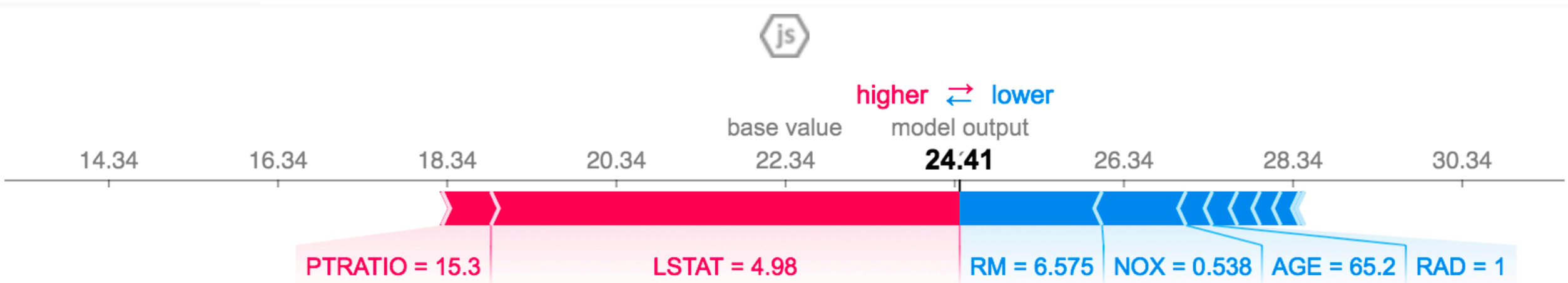

Probability of pregnancy
Age

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

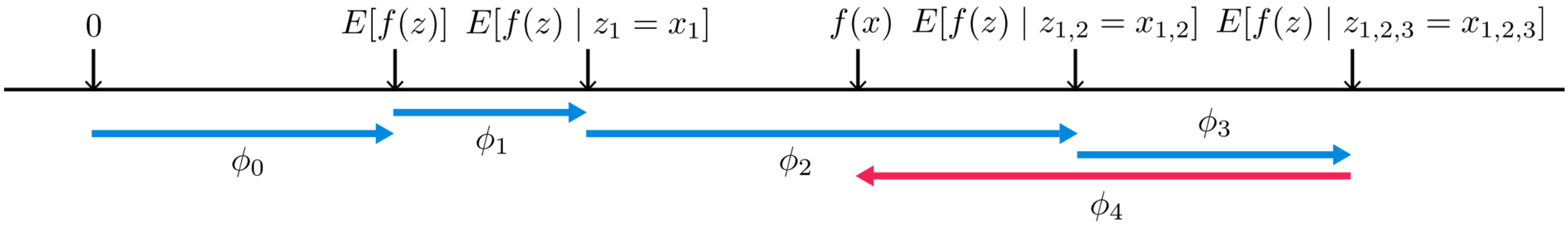
Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

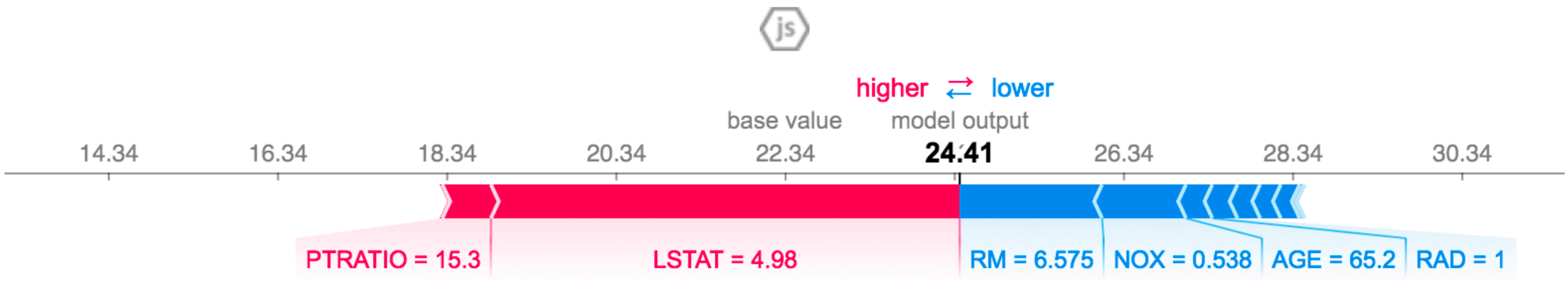


http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

Features are players
Models are coalitions
Shapley values will attribute values to features

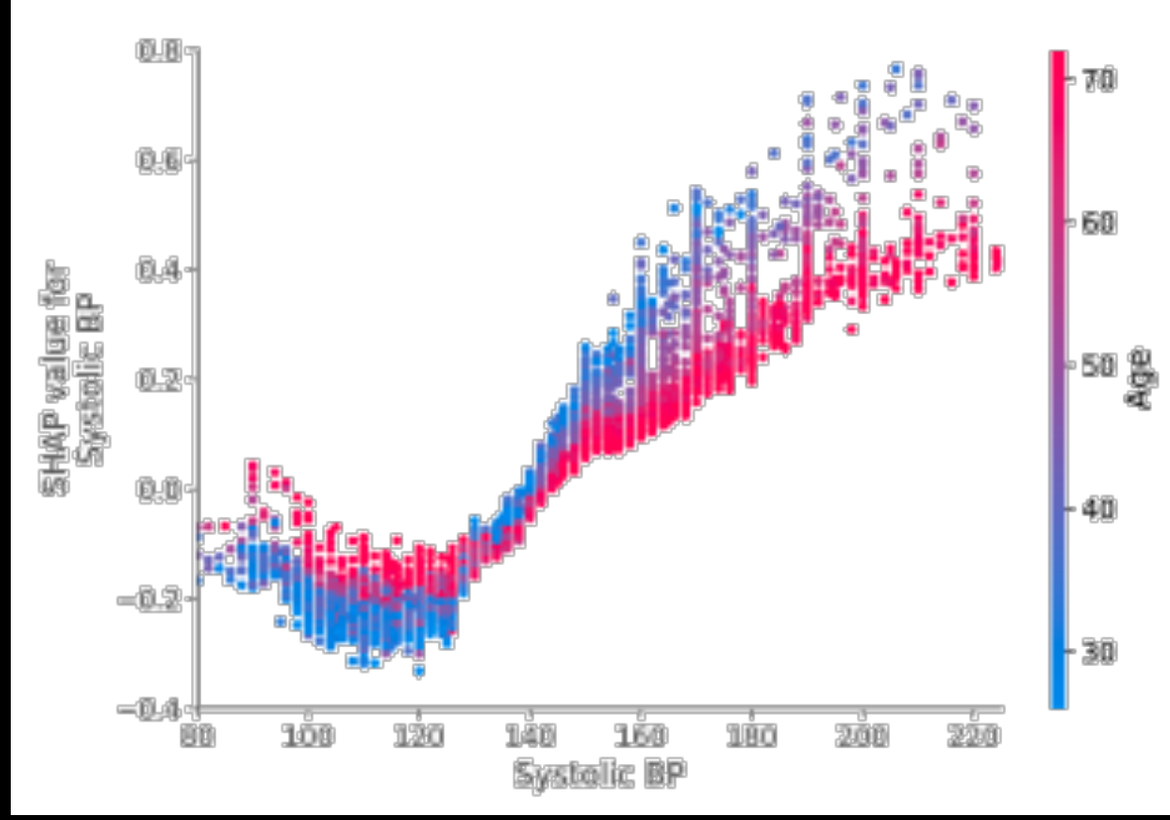
$$\phi_i(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N|-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$





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

SHAP
Pros and Cons



Pros

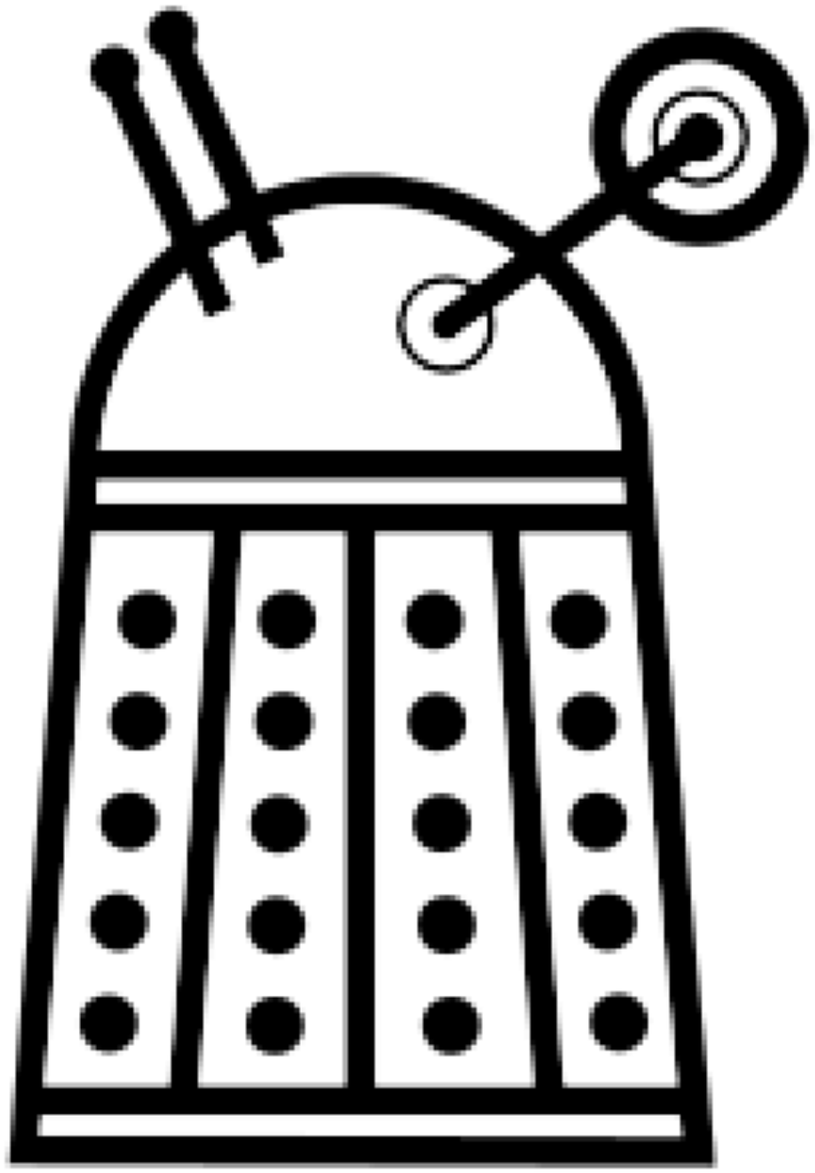
- Based on game theory - cool!
- Easy to use and visually appealing implementation.
- Unification of some other approaches.

Cons

- Additive (there are some diagnostic plots, but still…)
- Additive (additive explanations for non additive models?)
- Additive

explain!

explain!



explain!

README.md

DALEX

CRAN 0.2.6 | downloads 1359/month | downloads 14K | build passing | coverage 89%

DALEX: Descriptive mAchine Learning EXplanations

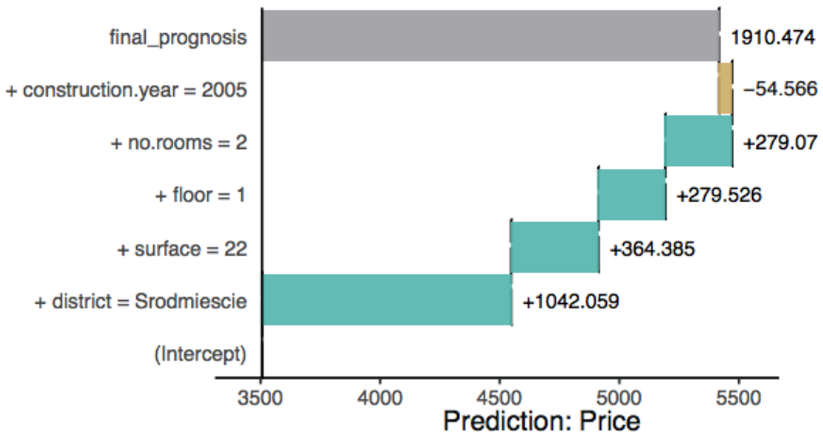
Machine Learning models are widely used and have various applications in classification or regression tasks. Due to increasing computational power, availability of new data sources and new methods, ML models are more and more complex. Models created with techniques like boosting, bagging of neural networks 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.

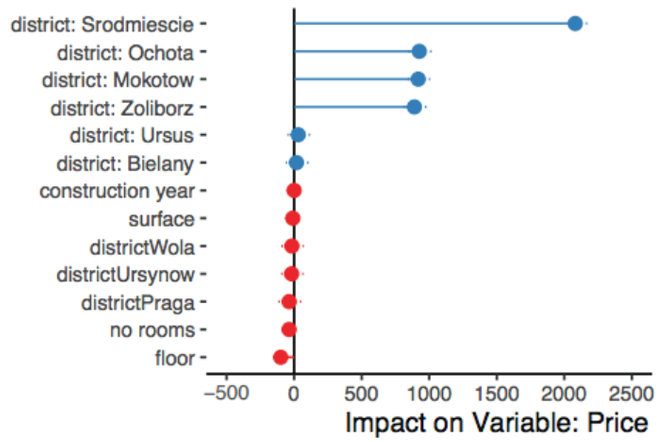
Find more about DALEX in this Gentle introduction to DALEX with examples.

Experimental Python version pyDALEX.

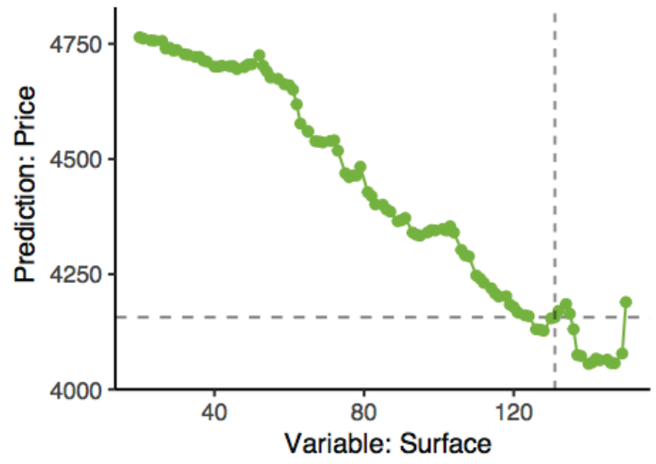
Break Down Plots



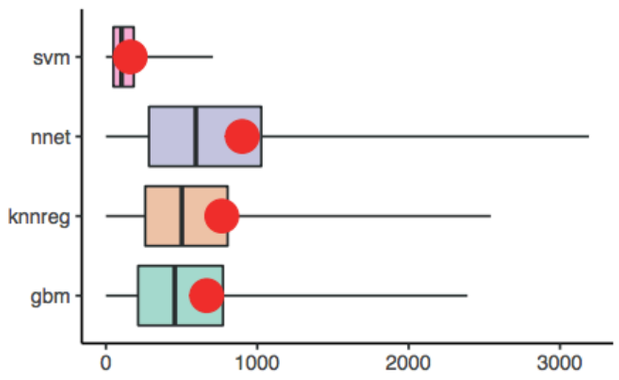
Local Variable Importance



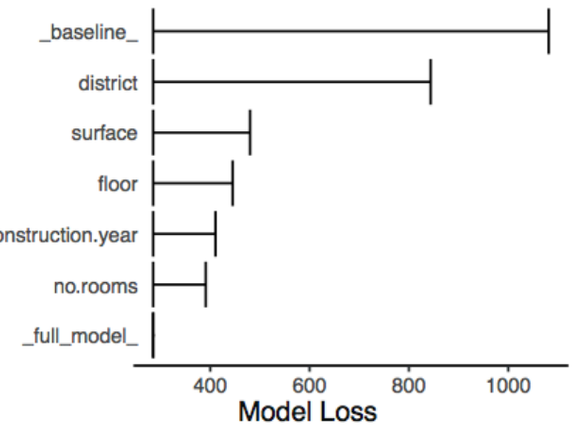
Ceteris Paribus Plots



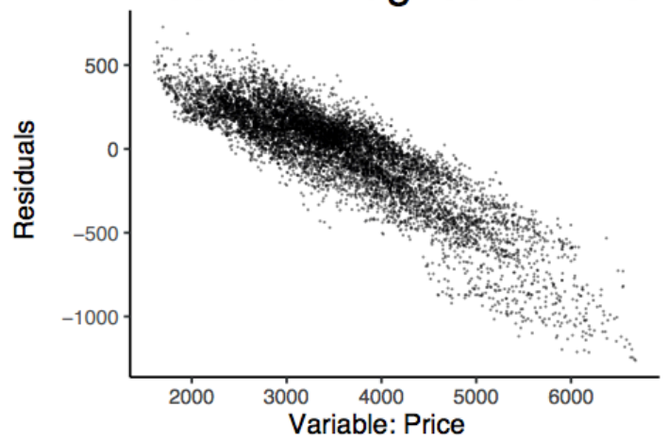
Model Performance Plots



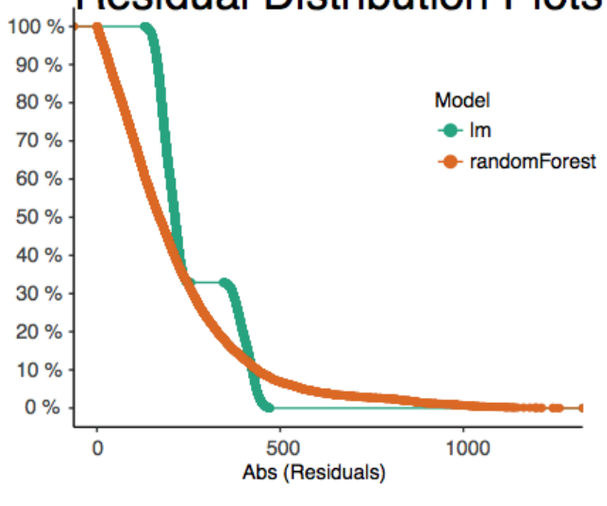
Variable Importance Plots



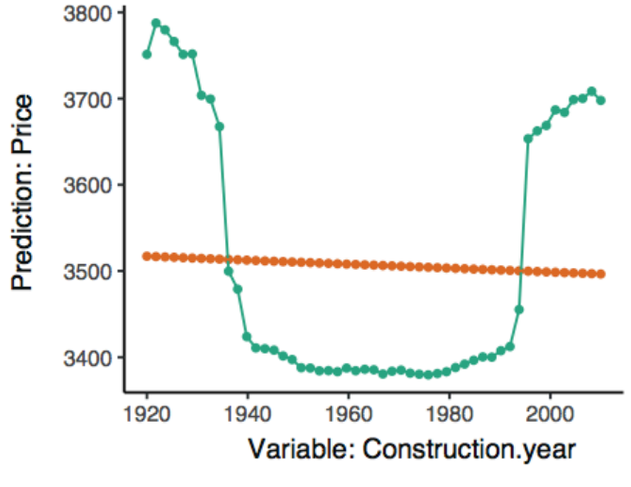
Residual Diagnostic Plots



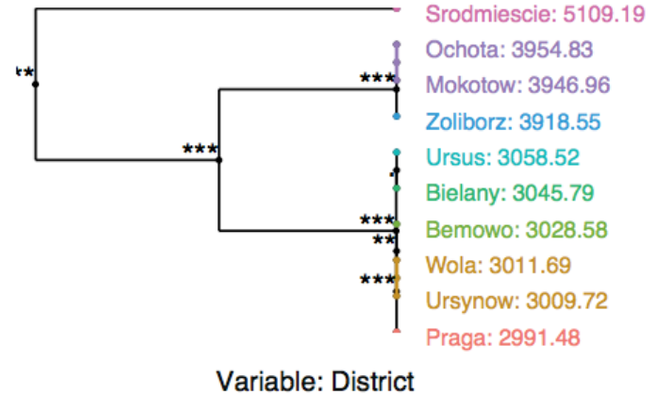
Residual Distribution Plots



Partial Dependency Plots

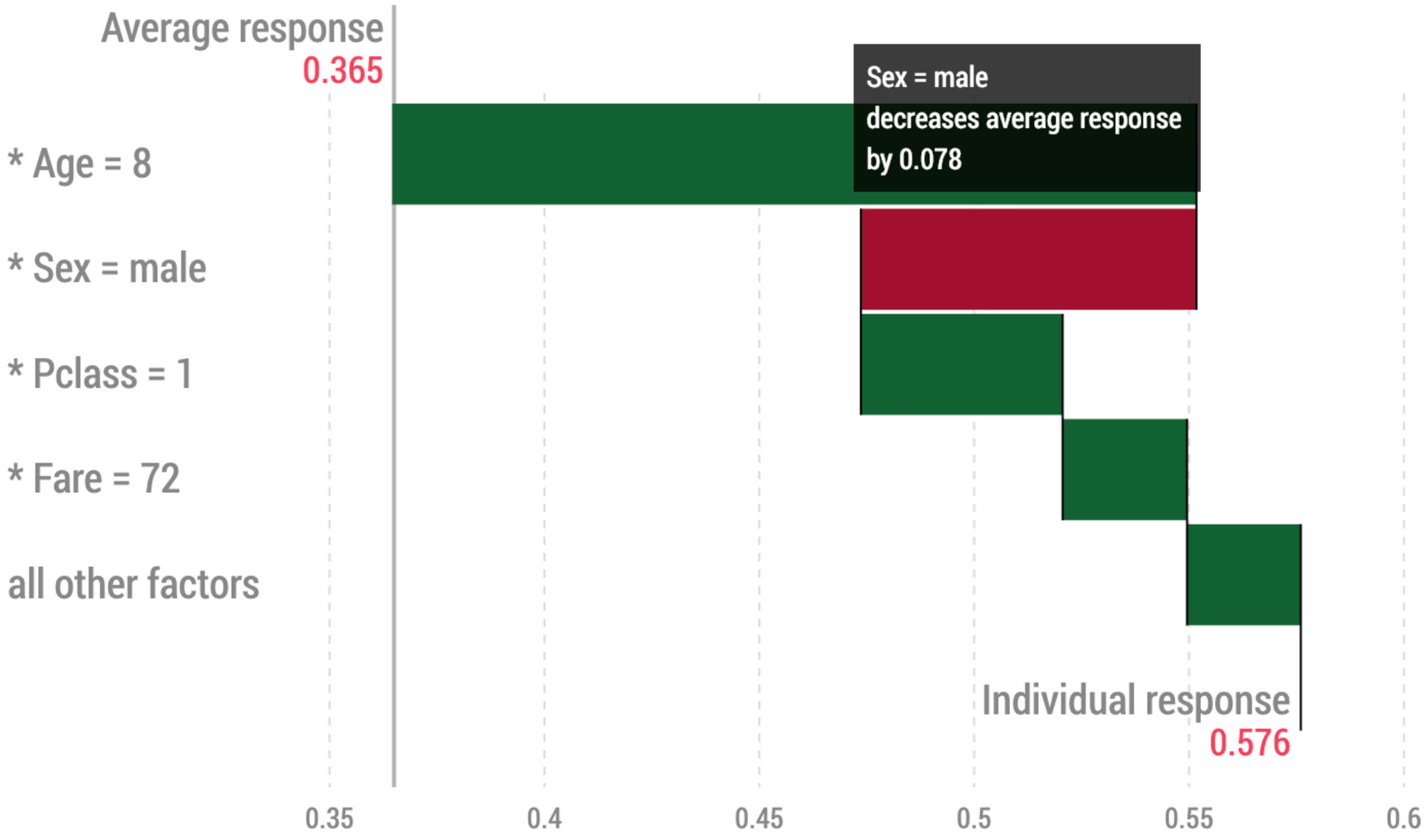


Merging Path Plots



Break Down 2

```
plotD3(rf_la)
```



What are the odds of surviving?

Master Philip Aks,[1] also known as Frank or Filly, was born in London, England on 7 June 1911.

He was the son of Polish immigrants, Sam Aks (1891-1970), a tailor, and **Leah Rosen** (1891-1967), natives of Łódź and Warsaw, respectively who had married the year before his birth before resettling in England.

In the months prior to Aks' birth his parents appeared on the 1911 census as residents of 198 St George Street, St George in the East, London and his father was described as a trouser machinist.

Contents

Titanic Passenger Summary

Name: Master Philip Aks (Frank)

Born: Wednesday 7th June 1911 in London, England
Age: 10 months and 8 days (Male)
Nationality: English
Last Residence: in London, England
3rd Class Passengers
First Embarked: Southampton on Wednesday 10th April 1912
Ticket No. 392091, £9 7s
Destination: Norfolk, Virginia, United States
Rescued (boat 11)
Disembarked Carpathia: New York City on Thursday 18th April 1912

Linked Biography

https://www.encyclopedia-titanica.org/

1. Get the data

```
library("titanic")
head(titanic_small)
```

```
  Survived Pclass     Sex Age SibSp Parch     Fare Embarked
1        0      3    male  22     1     0  7.2500        S
2        1      1 female  38     1     0 71.2833        C
3        1      3 female  26     0     0  7.9250        S
4        1      1 female  35     1     0 53.1000        S
5        0      3    male  35     0     0  8.0500        S
7        0      1    male  54     0     0 51.8625        S
```

2. Create a model

```
library("titanic")
head(titanic_small)

library("randomForest")
rf_model <- randomForest(Survived ~ Pclass + Sex + Age + SibSp +
                           Parch + Fare + Embarked,
                         data = titanic_small)
rf_model
```

```
Call:
 randomForest(formula = Survived ~ Pclass + Sex + Age + SibSp +
arked, data = titanic_small)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 19.05%
Confusion matrix:
    0   1 class.error
0 378  46   0.1084906
1  90 200   0.3103448
```

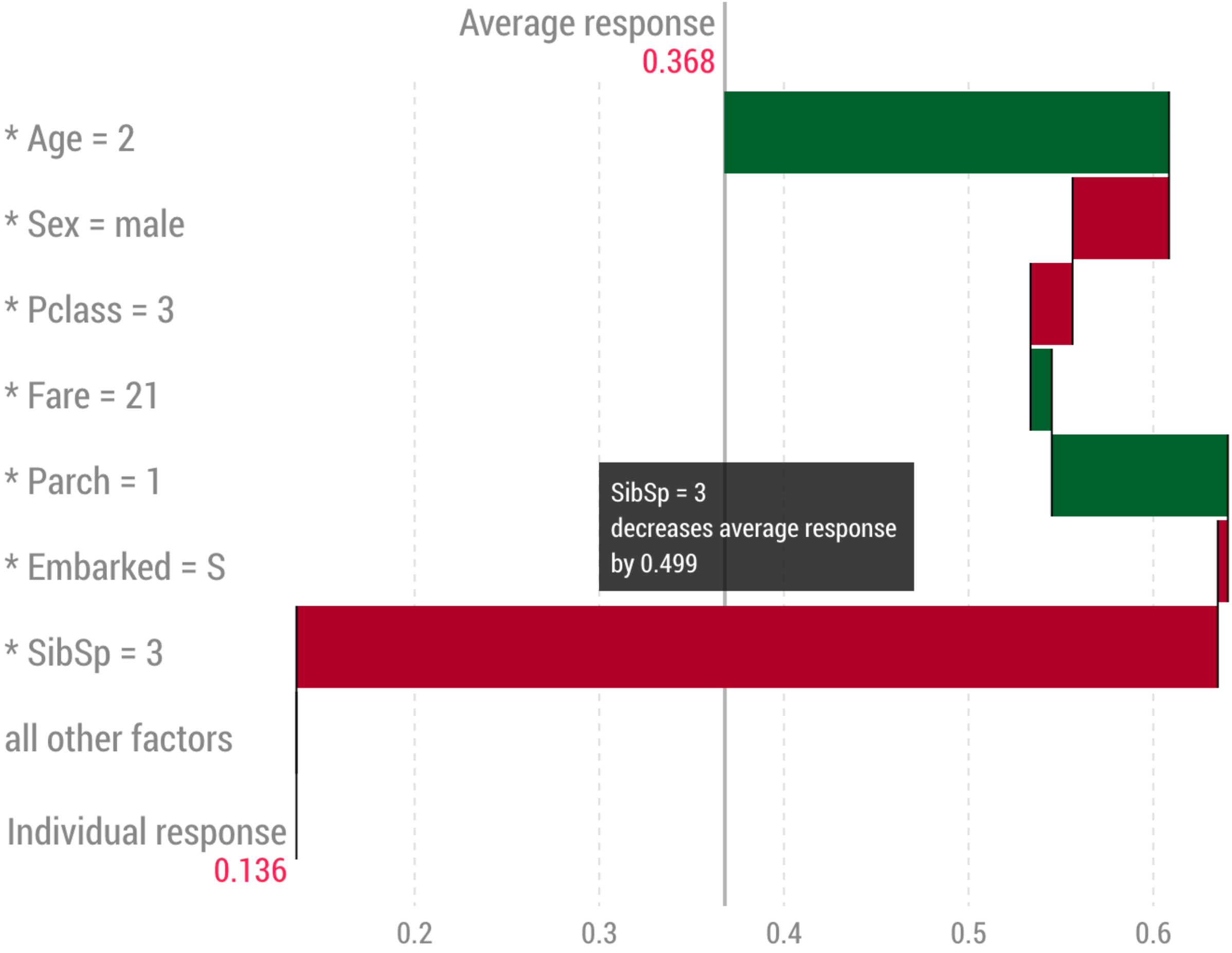
3. Prepare an explainer

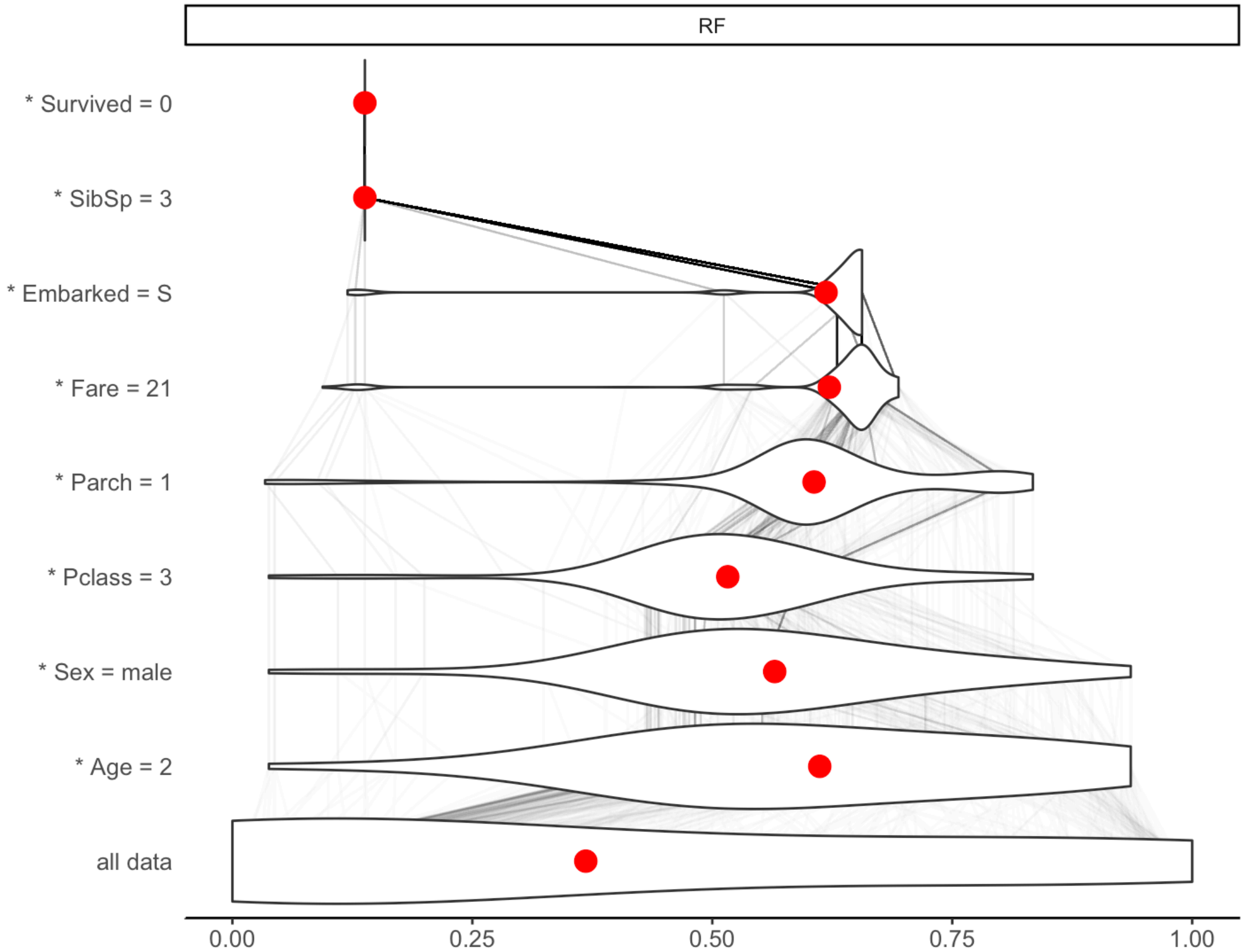
```r
library("titanic")
head(titanic_small)

library("randomForest")
rf_model <- randomForest(Survived ~ Pclass + Sex + Age + SibSp +
                           Parch + Fare + Embarked,
                         data = titanic_small)
rf_model
```

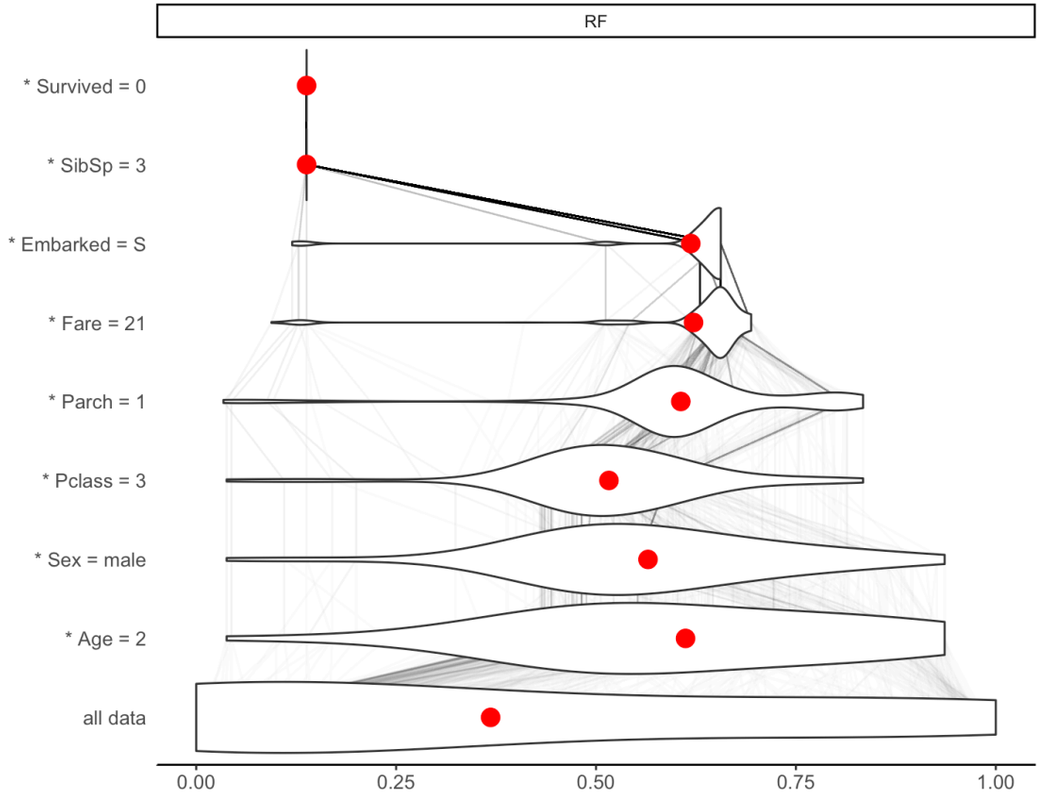
```r
library("DALEX2")
predict_fuction <- function(m,x) predict(m, x, type = "prob")[,2]
rf_explain <- explain(rf_model, data = titanic_small,
                      y = titanic_small$Survived == "1", label = "RF",
                      predict_function = predict_fuction)
```

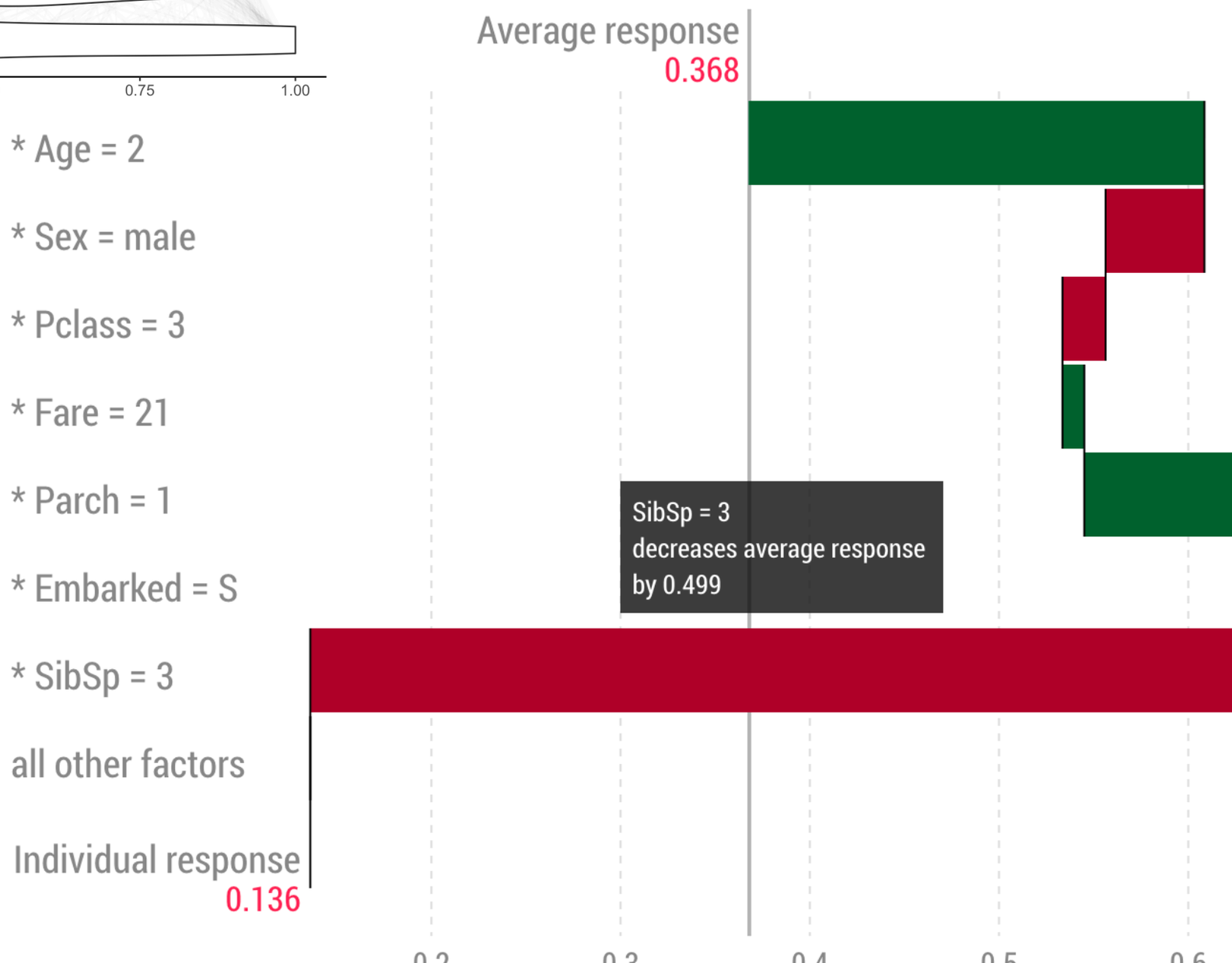
4. Plot explanations

```r
library("titanic")
head(titanic_small)

library("randomForest")
rf_model <- randomForest(Survived ~ Pclass + Sex + Age + SibSp +
                           Parch + Fare + Embarked,
                         data = titanic_small)
rf_model

library("DALEX2")
predict_fuction <- function(m,x) predict(m, x, type = "prob")[,2]
rf_explain <- explain(rf_model, data = titanic_small,
                      y = titanic_small$Survived == "1", label = "RF",
                      predict_function = predict_fuction)

library("breakDown2")
rf_la <- local_attributions(rf_explain, titanic_small[7,])
rf_la
plotD3(rf_la, max_features = 7)
```


Average response
0.368
* Age = 2
* Sex = male
* Pclass = 3
* Fare = 21
* Parch = 1
* Embarked = S
* SibSp = 3
all other factors
Individual response
0.136
SibSp = 3
decreases average response
by 0.499
0.2
0.3
0.4
0.5
0.6

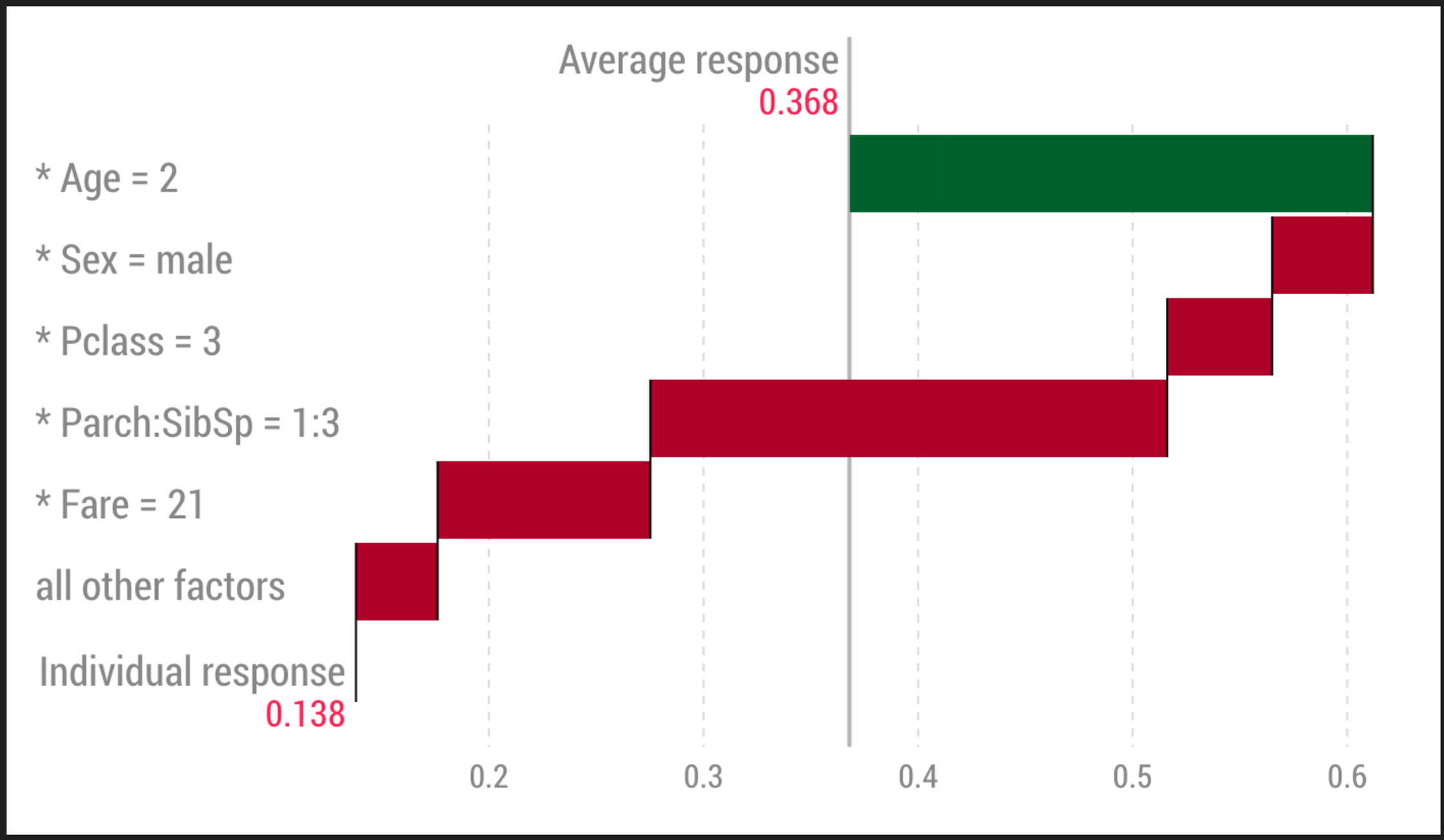

RF
* Survived = 0
* SibSp = 3
* Embarked = S
* Fare = 21
* Parch = 1
* Pclass = 3
* Sex = male
* Age = 2
all data
0.00
0.25
0.50
0.75
1.00


RF
* Survived = 0
* SibSp = 3
* Embarked = S
* Fare = 21
* Parch = 1
* Pclass = 3
* Sex = male
* Age = 2
all data
0.00
0.25
0.50
0.75
1.00


Average response
0.368
* Age = 2
* Sex = male
* Pclass = 3
* Fare = 21
* Parch = 1
* Embarked = S
* SibSp = 3
all other factors
Individual response
0.136
SibSp = 3
decreases average response
by 0.499
0.2
0.3
0.4
0.5
0.6

Break Down 2 with interactions

```
rf_la <- local_interactions(rf_explain, titanic_small[7,])
plotD3(rf_la)
```



Model Level Uncertainty

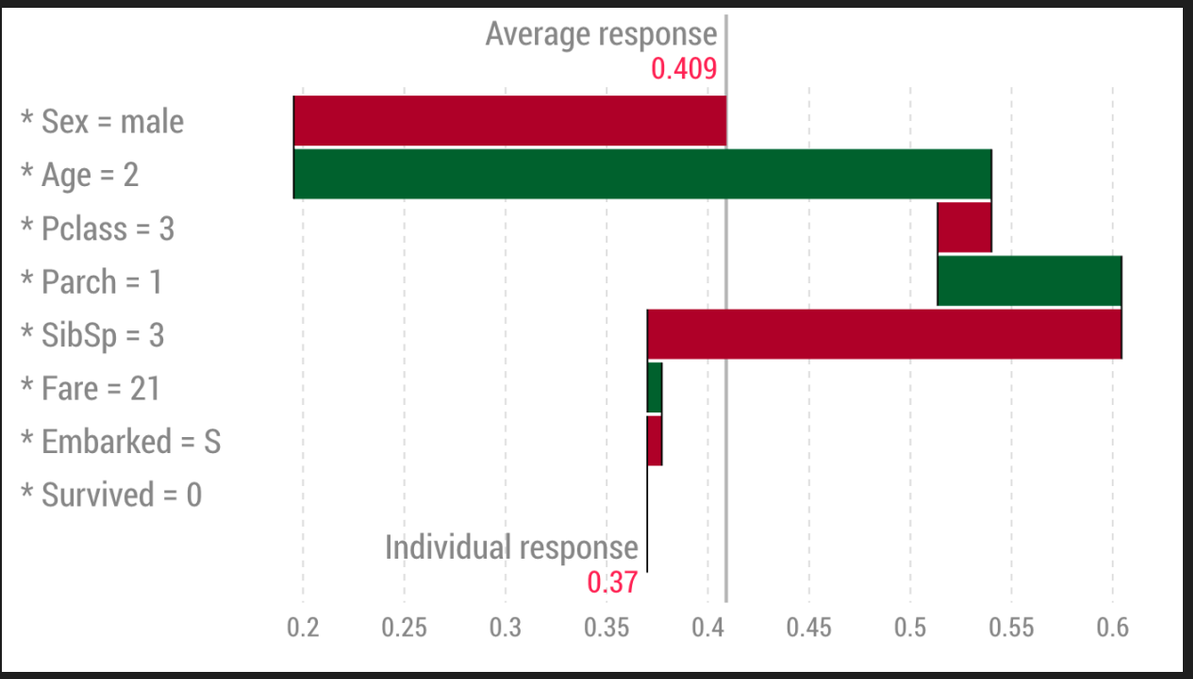
Some models are robust some are not.

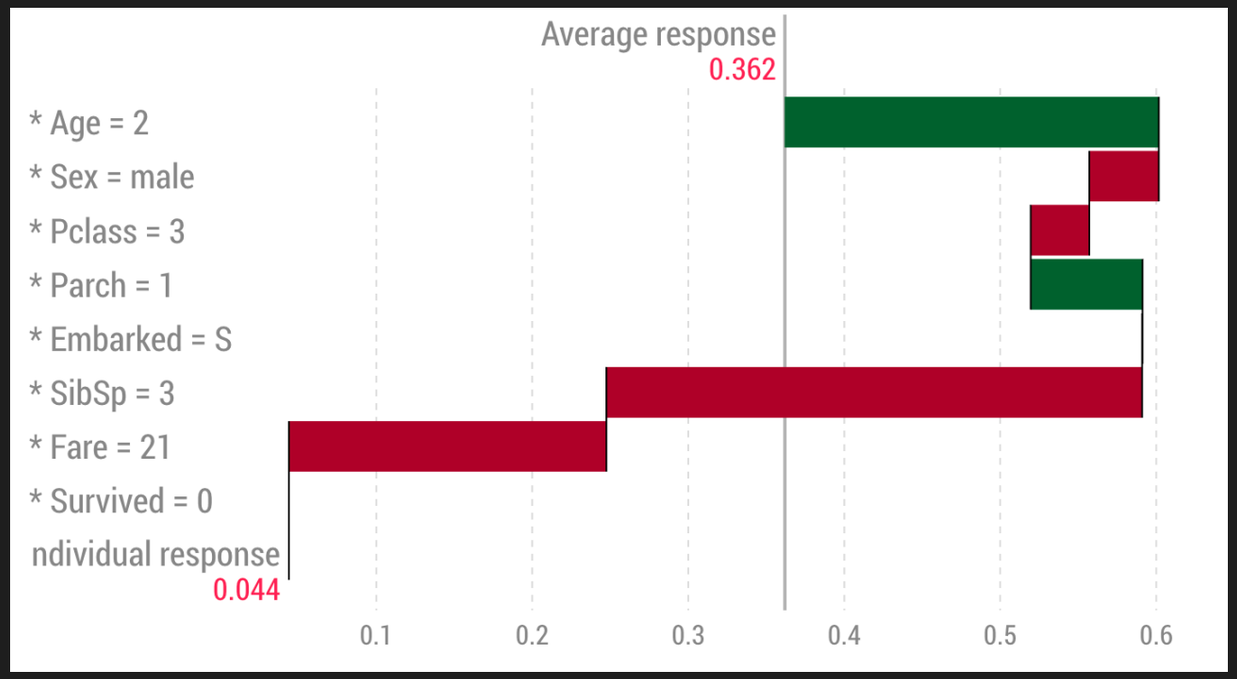
The more elastic is the model the less stable it is.

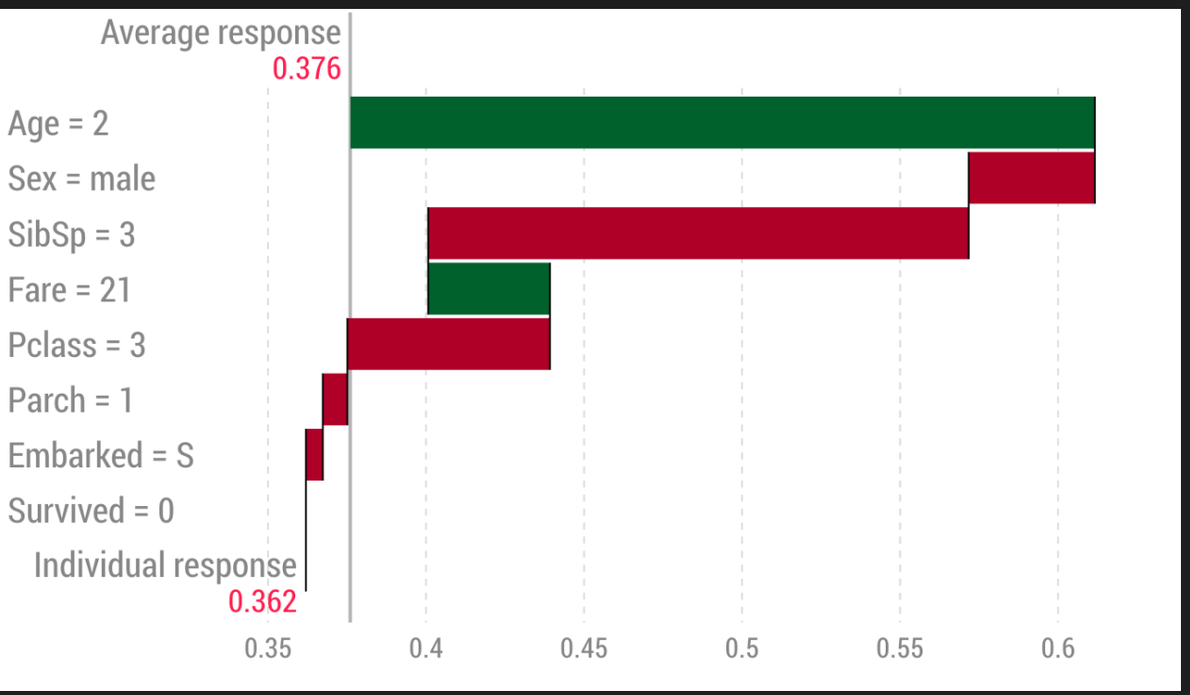
In explanations we should take into account model stability.

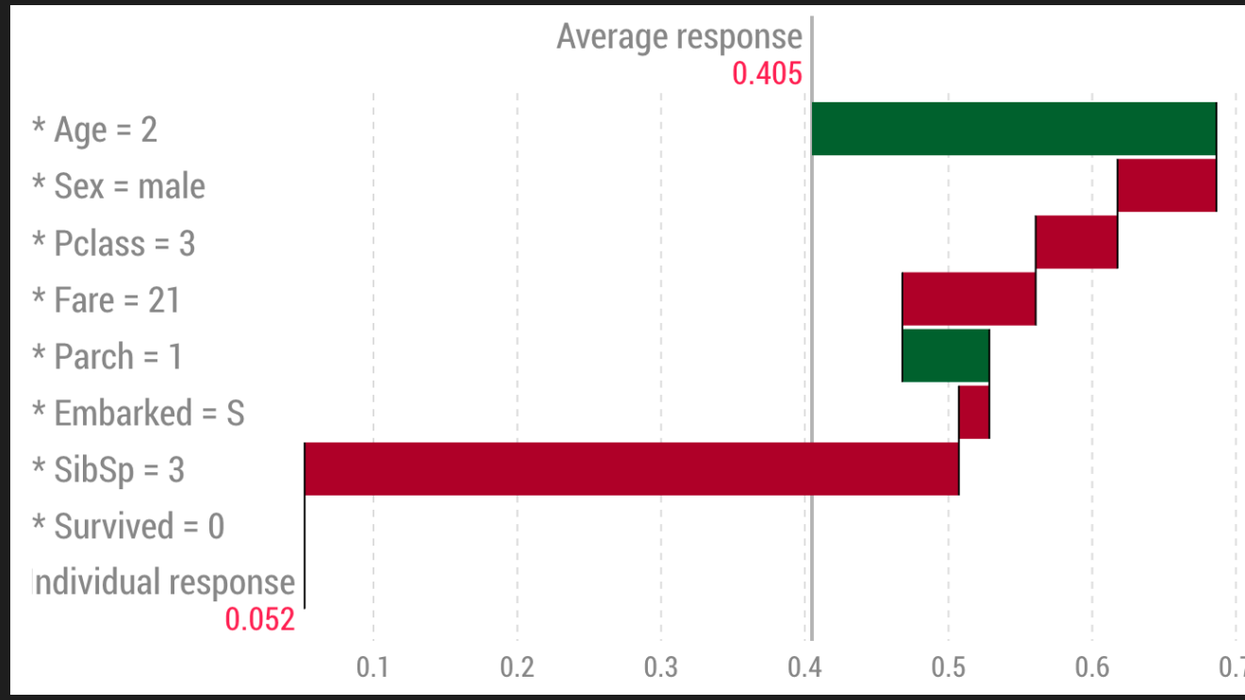
Bootstrap is a nice approach to assess the model stability.

```r
titanic_B <- titanic_small[sample(1:nrow(titanic_small), replace = TRUE),]

rf_model <- randomForest(Survived ~ Pclass + Sex + Age + SibSp +
                           Parch + Fare + Embarked,
                         data = titanic_B)
```









Explanation Level Uncertainty

Simple explanations for complex models?

How it is possible?

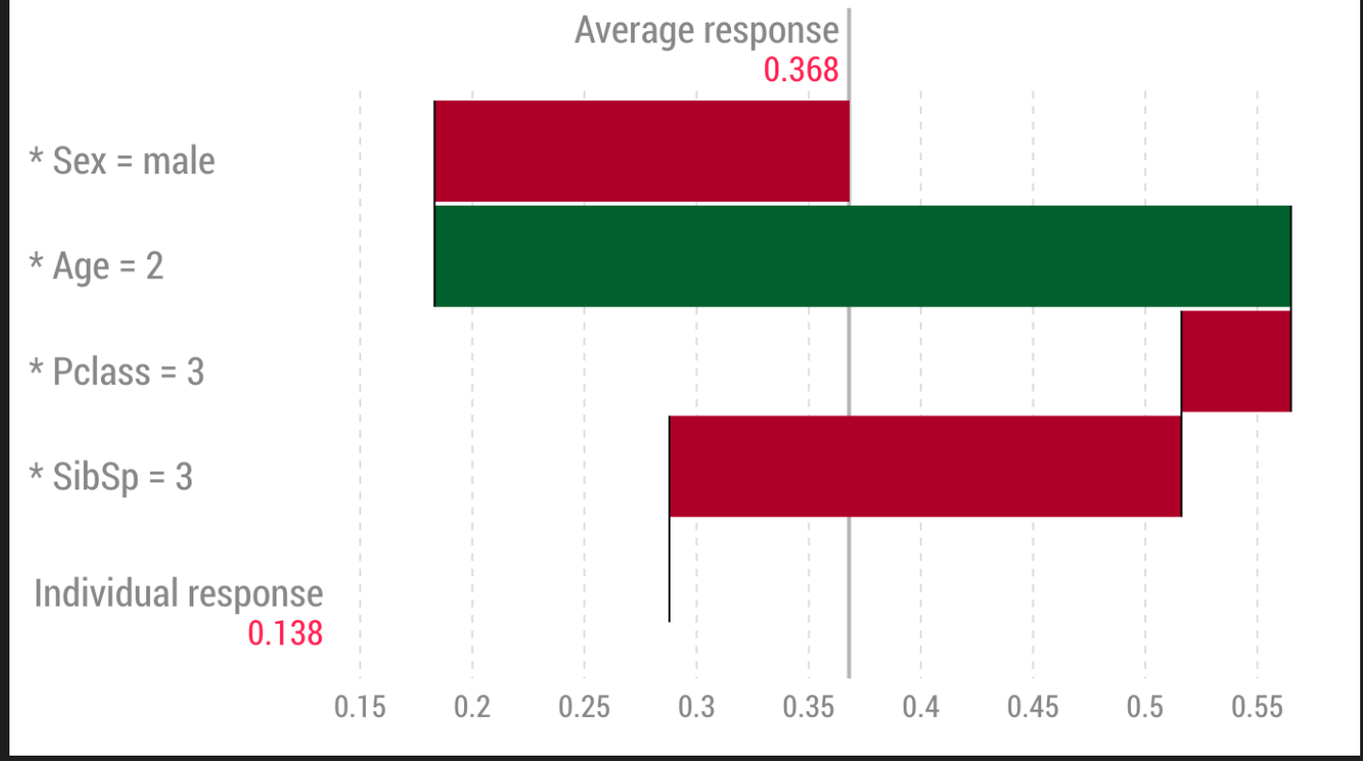
Explanations are simplified

-> the simpler the more we loose in the fidelity.

We need to understand how accurate are explanations!

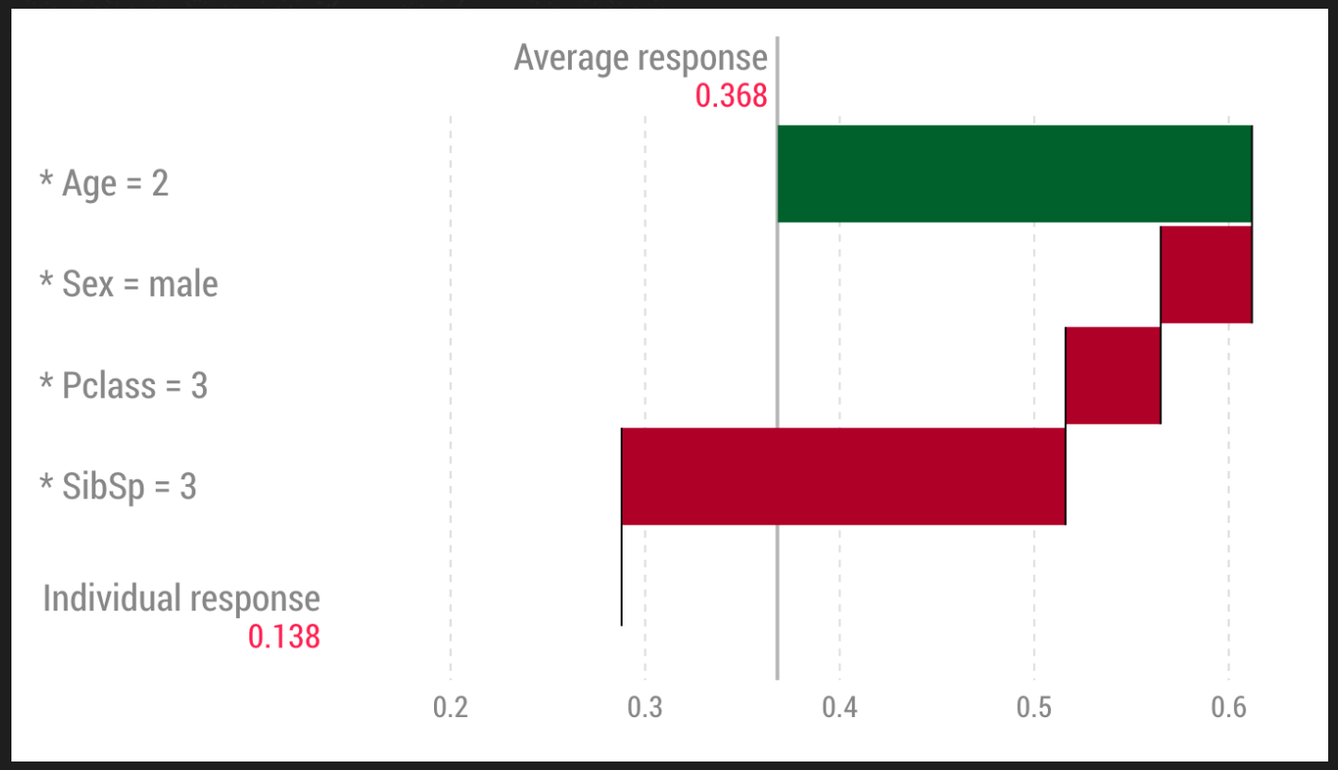
Order does matter

for non-additive models




```
rf_la <- local_attributions(rf_explain, titanic_small[7,],
                            order = c("Age", "Sex", "Pclass", "SibSp"))
plotD3(rf_la)
```

```
rf_la <- local_attributions(rf_explain, titanic_small[7,],
                            order = c("Sex", "Age", "Pclass", "SibSp"))
plotD3(rf_la)
```



Break Down 2 and r2d3

```r
# plot D3 object
r2d3::r2d3(
  data = x_as_list,
  script = system.file("breakDownD3.js", package = "breakDown2"),
  dependencies = system.file("tooltipD3.js", package = "breakDown2"),
  options = list(xmin =min_max[1], xmax = min_max[2],
                 model_avg = model_baseline, model_res = model_prediction),
  d3_version = "4"
)
```

```
// add small links between rectangles
svg
  .append("g")
  .selectAll("line")
  .data(data)
  .enter()
  .append("line")
  .attr("x1", function(d) { return xAxis(d.cummulative); })
  .attr("x2", function(d) { return xAxis(d.cummulative); })
  .attr("y1", function(d, i) { return yAxis(i) })
  .attr("y2", function(d, i) { return yAxis(i + 2)-2 })
  .style("stroke","black")
  .style("stroke-width","1")
```

Thank you!

Find more at DrWhy.AI