Microsoft

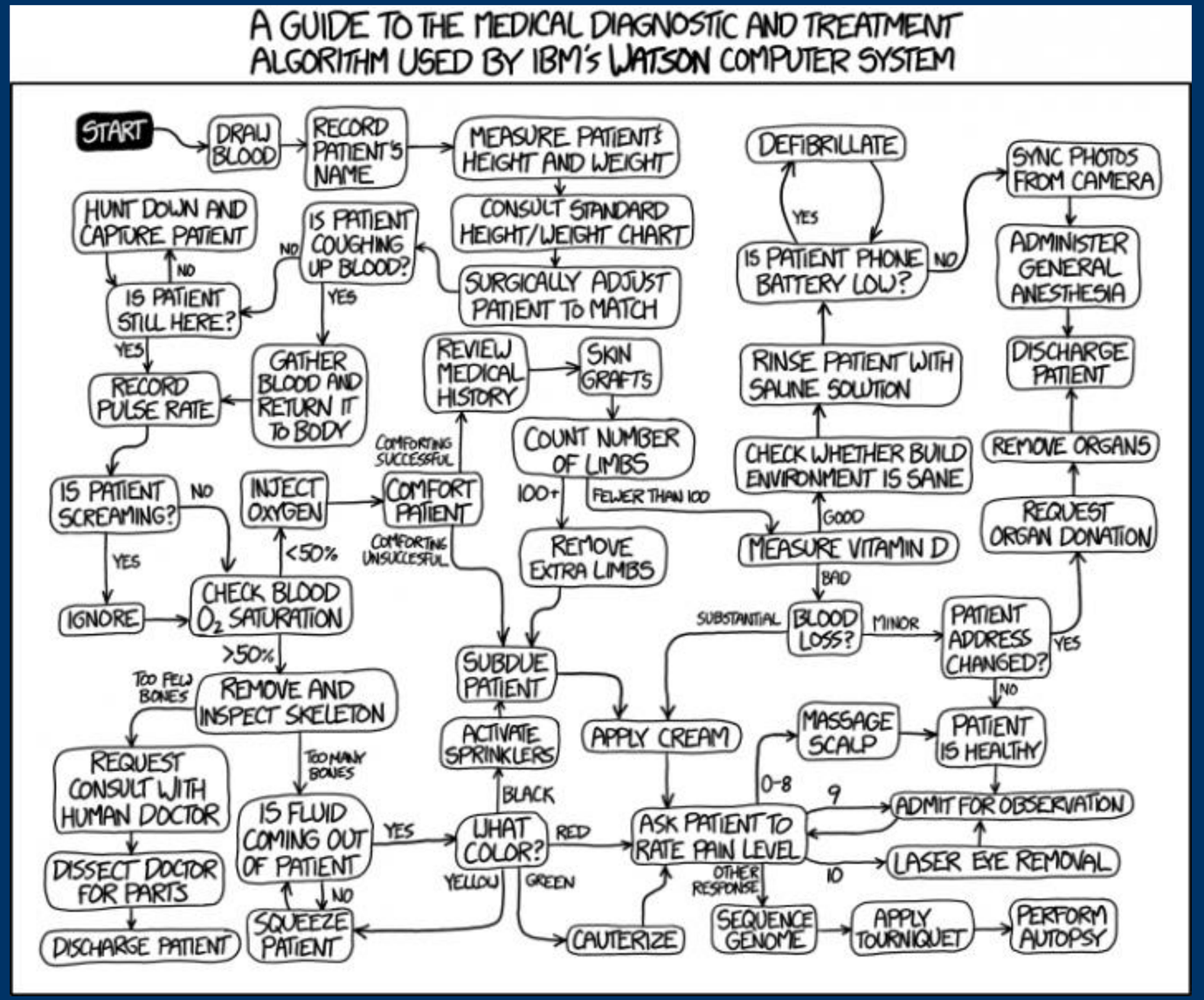
Excuse Me, Why Has Your AI Denied My Loan?

A Look at Machine Learning Model Explainability

Sorin Pește

Cloud Solution Architect, Data & AI @ Microsoft

source: xkcd.com



CODECAMP_BUCHAREST

/ 9 November 2019

/ JW Marriott Grand Hotel Bucharest

Global partners

Cognizant Softvision

EDUROM

Diamond partners

METRO SYSTEMS

Ness Digital Engineering

HARMAN A SAMSUNG COMPANY

iQST make it work

tremend

pipedrive

solarwinds

COMPUTARIS

EY Building a better working world

UBISOFT

Atos

NN

PSS PROSOFT SOLUTIONS

orange

Platinum partners

stefanini

SYSTEMATIC

MindGeek

REGINA MARIA REȚEAUA PRIVATĂ DE SĂNĂTATE

Gold partners

JET BRAINS

Entrepreneurship

rubikhub mix to match

HOW TO WEB

TBR

SPINLAB THE HHL ACCELERATOR

Welness Partners

NESCAFÉ

Berlin's vitamin aqua

Media partners

CANOPY Taking digital to new heights

.exposé SURSA TA DE STIRI

IT MANIA

Tech partners

LOOQ VIRTUAL REALITY SOLUTIONS

Liked By

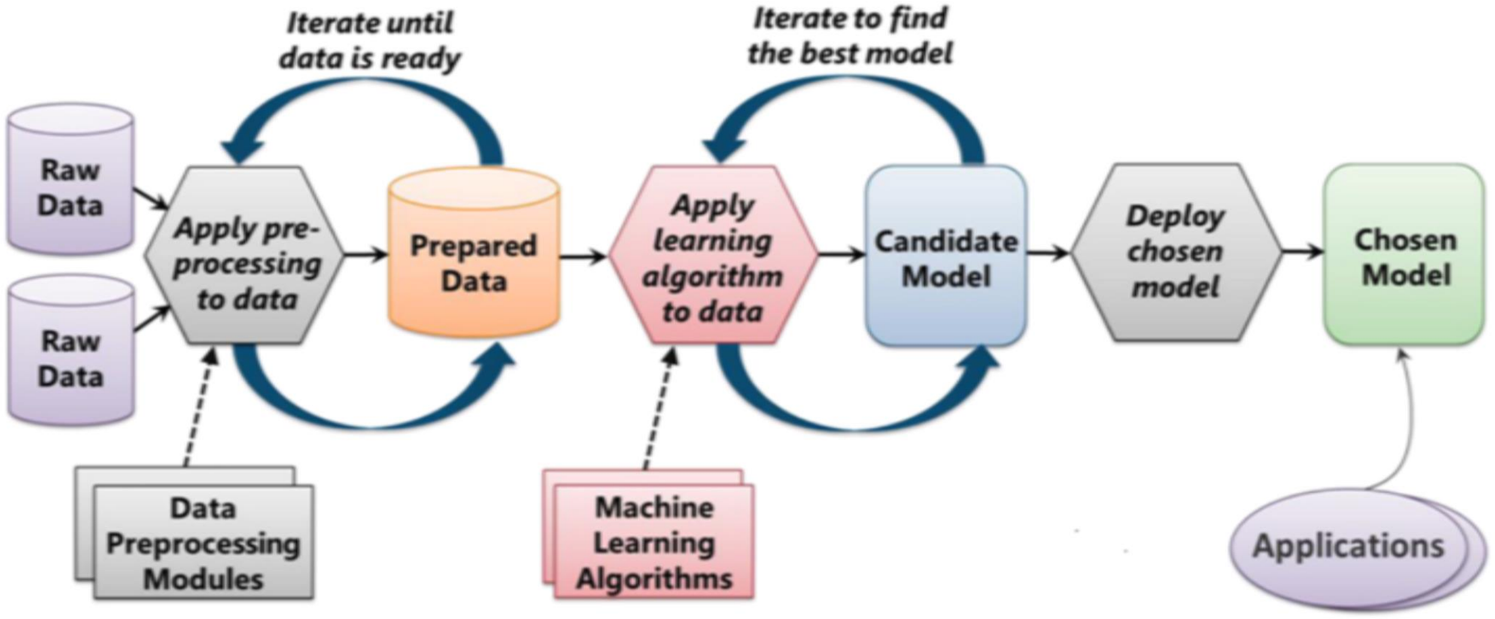
RADIO GUERRILLA DESCHIDEM MINȚI DIN 2004

Demo Code

https://github.com/neaorin/Explaining-ML-Models/

Why Should You Care About Explainability In Machine Learning?

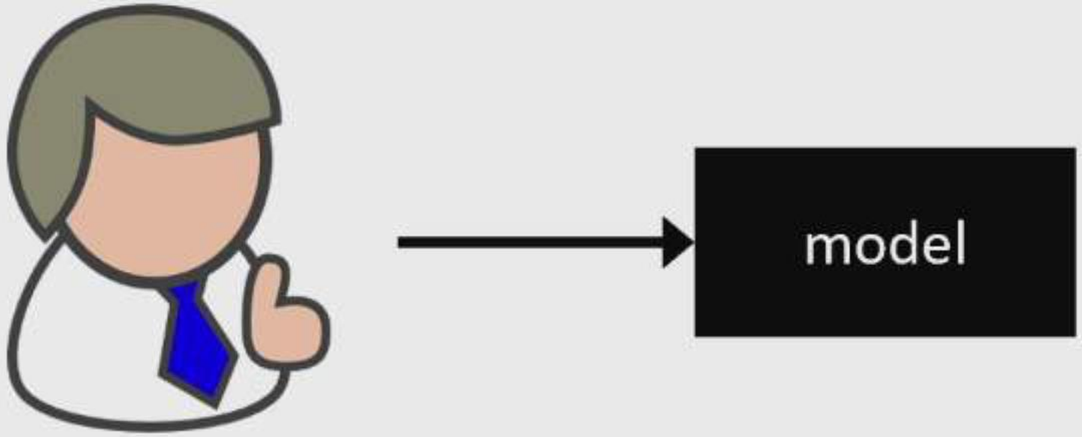
The Machine Learning Process



From "Introduction to Microsoft Azure" by David Chappell

https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa

John, a bank customer



John, a bank customer

John, a bank customer

chance John will have repayment problems

55%

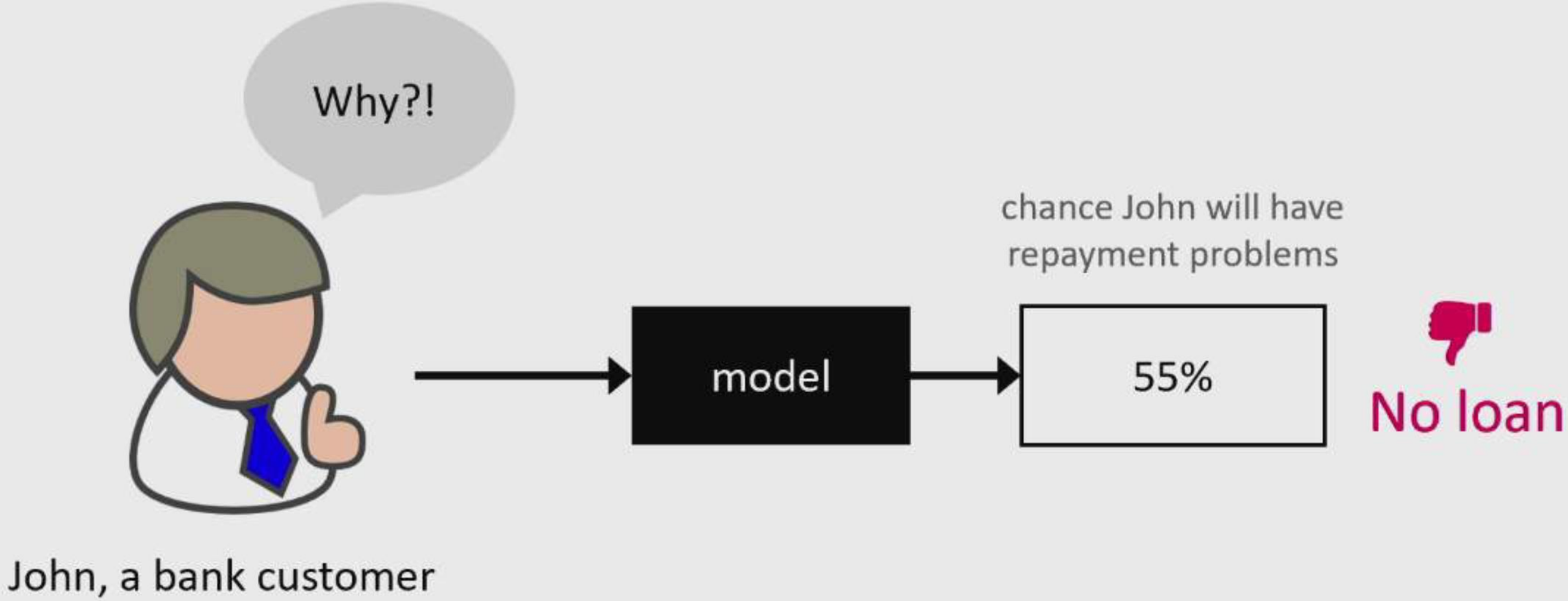
John, a bank customer

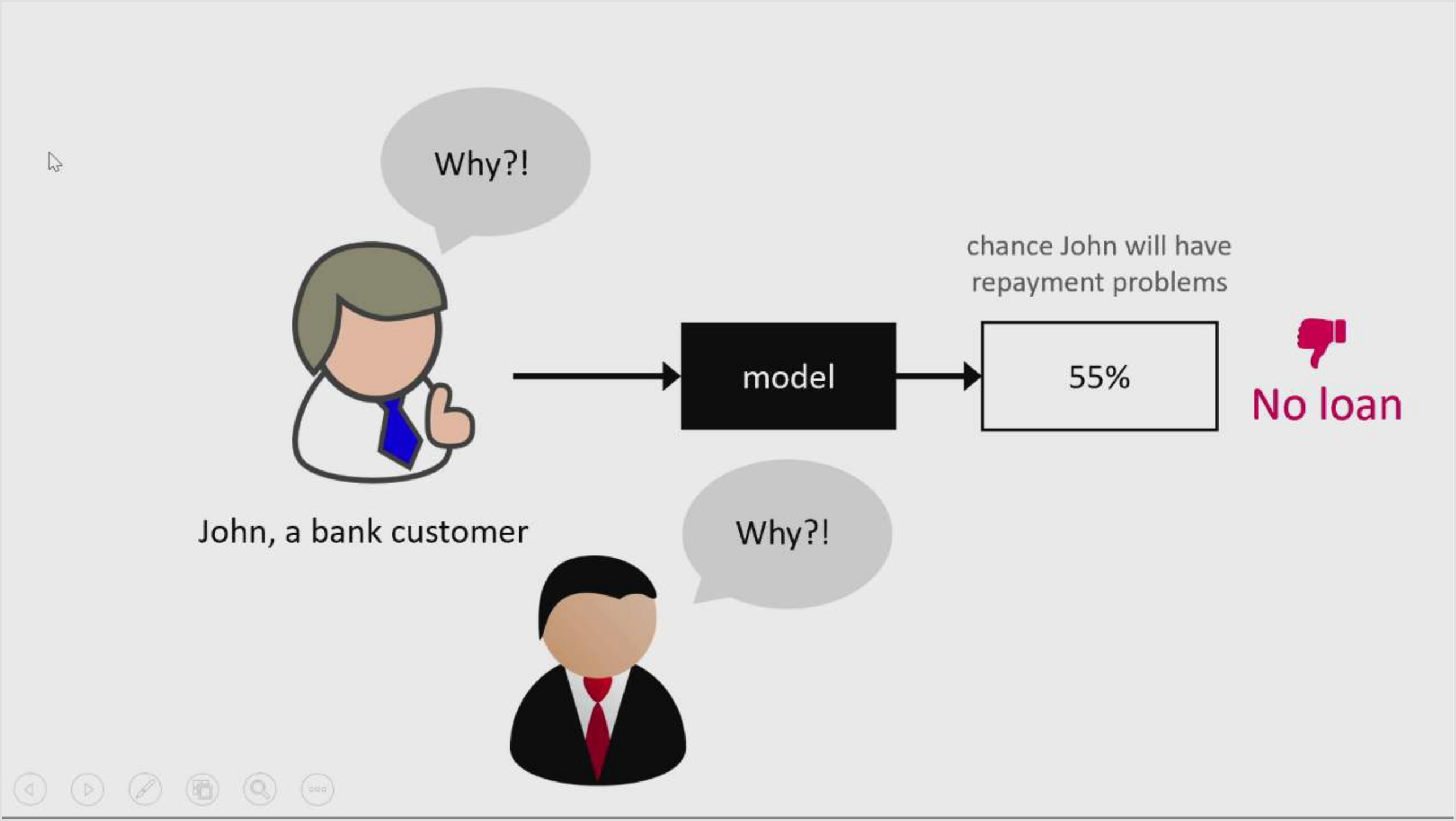
model

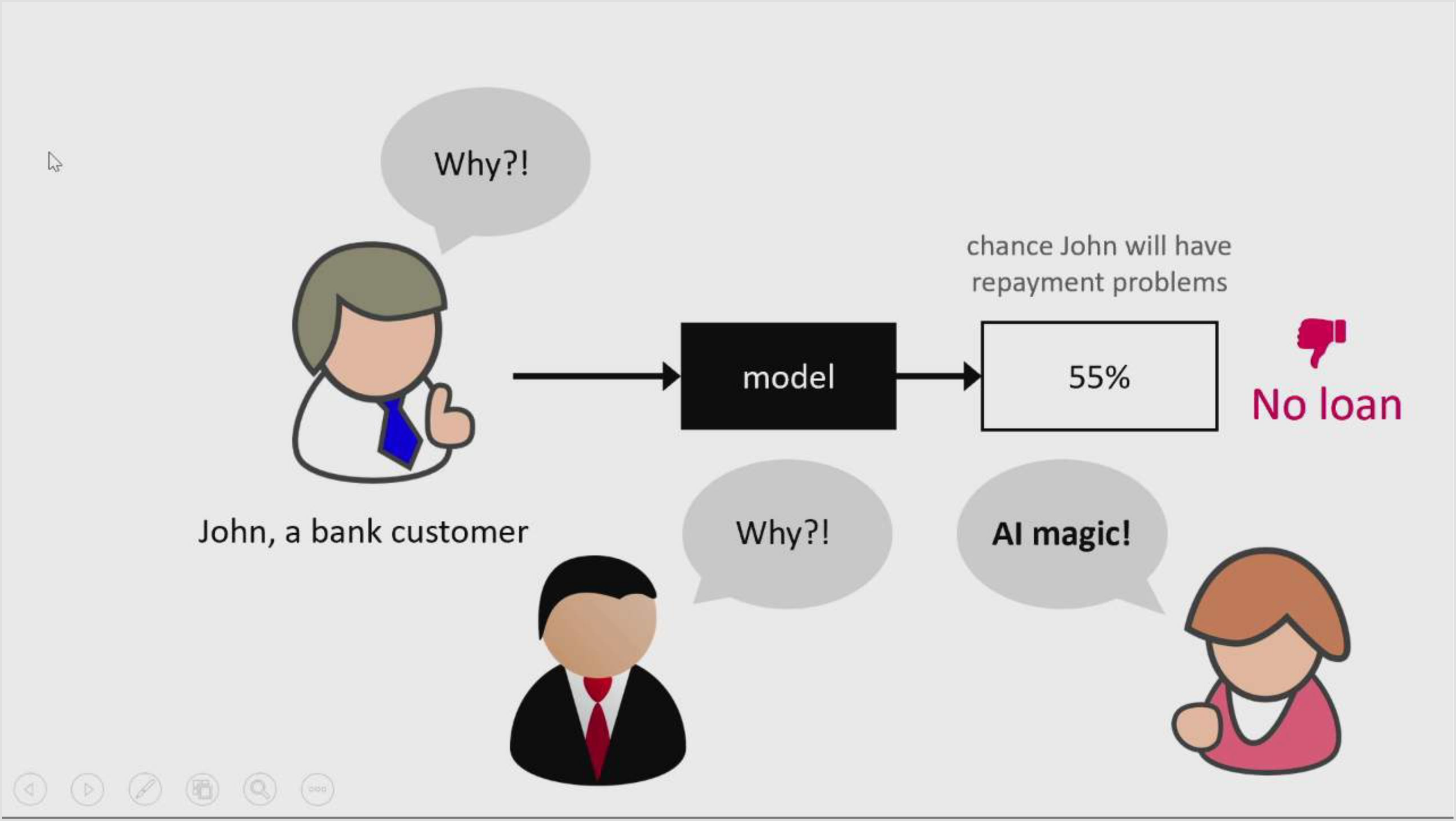
chance John will have repayment problems

55%

No loan


Why?!
chance John will have repayment problems
model
55%
No loan
John, a bank customer


Why?!
chance John will have repayment problems
model
55%
No loan
John, a bank customer
Why?!


Why?!
chance John will have repayment problems
model
55%
No loan
John, a bank customer
Why?!
AI magic!

BUSINESS NEWS OCTOBER 10, 2018 / 6:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Artificial Intelligence

Military artificial intelligence can be easily and dangerously fooled

AI warfare is beginning to dominate military strategy in the US and China, but is the technology ready?

by **Will Knight** Oct 21, 2019

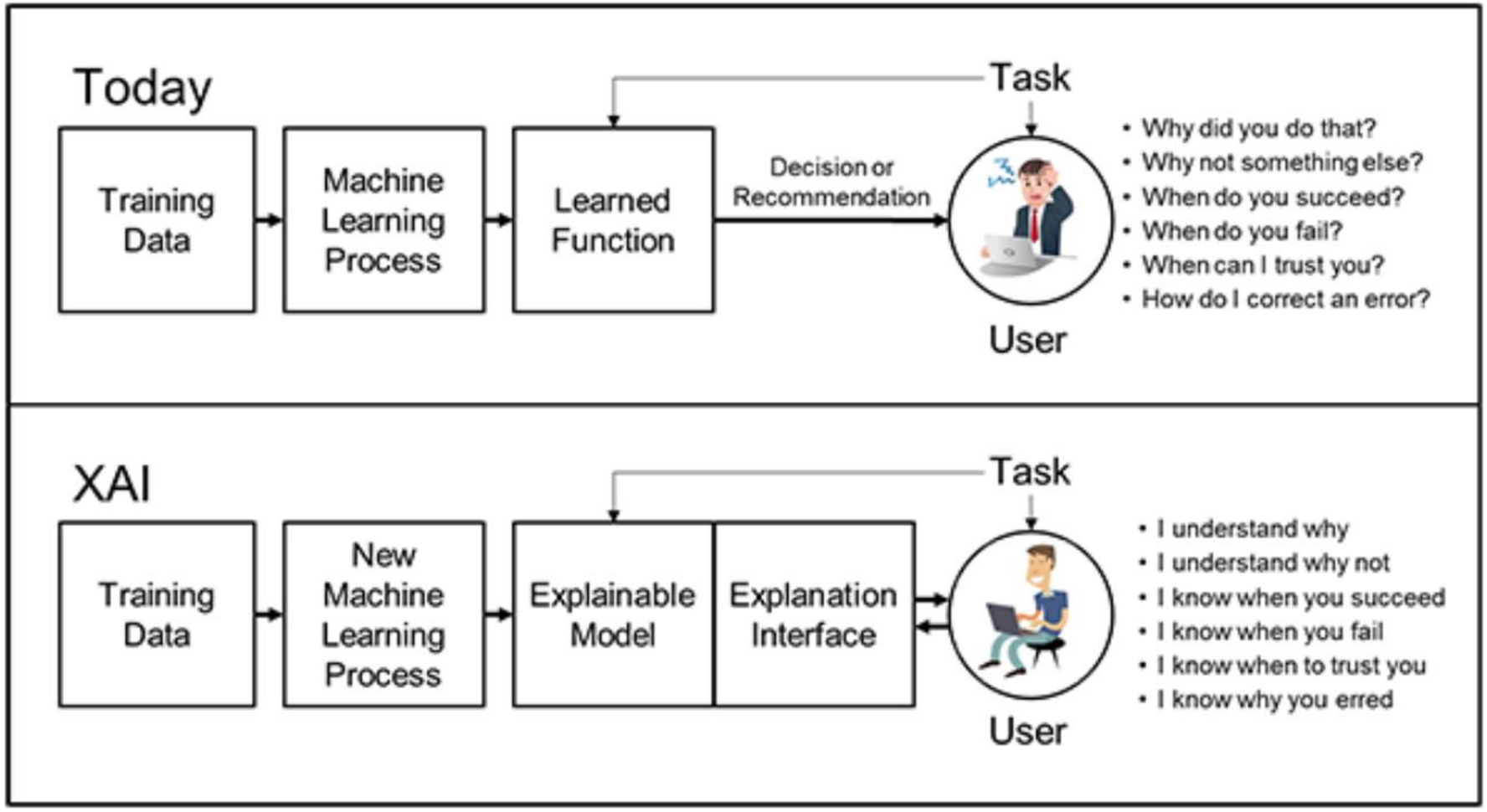
Last March, Chinese researchers announced an ingenious and potentially devastating attack against one of America's most prized technological assets —a Tesla electric car.

The team, from the security lab of the Chinese tech giant Tencent, demonstrated several ways to fool the AI algorithms on Tesla's car. By subtly altering the data fed to the car's sensors, the researchers were able to bamboozle and bewilder the artificial intelligence that runs the vehicle.

In one case, a TV screen contained a hidden pattern that

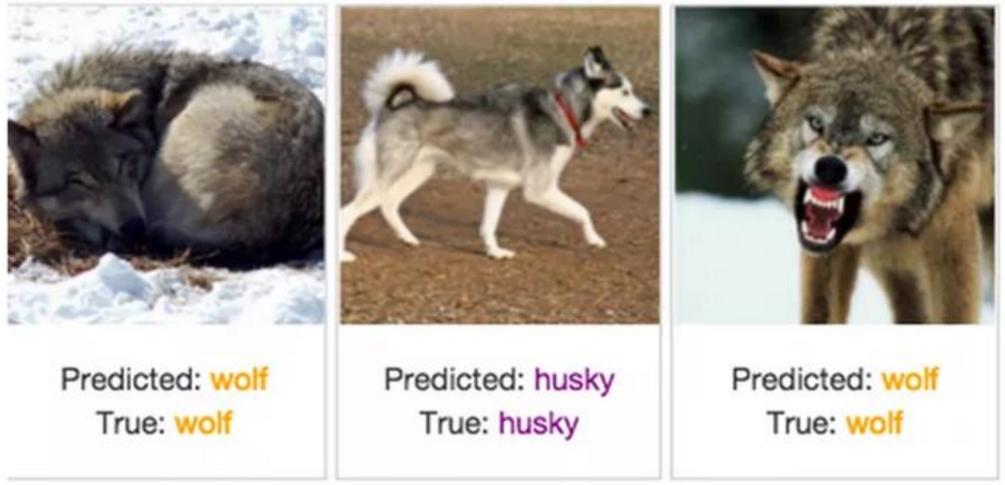
https://www.technologyreview.com/s/614497/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/

We Need AI to be Explainable

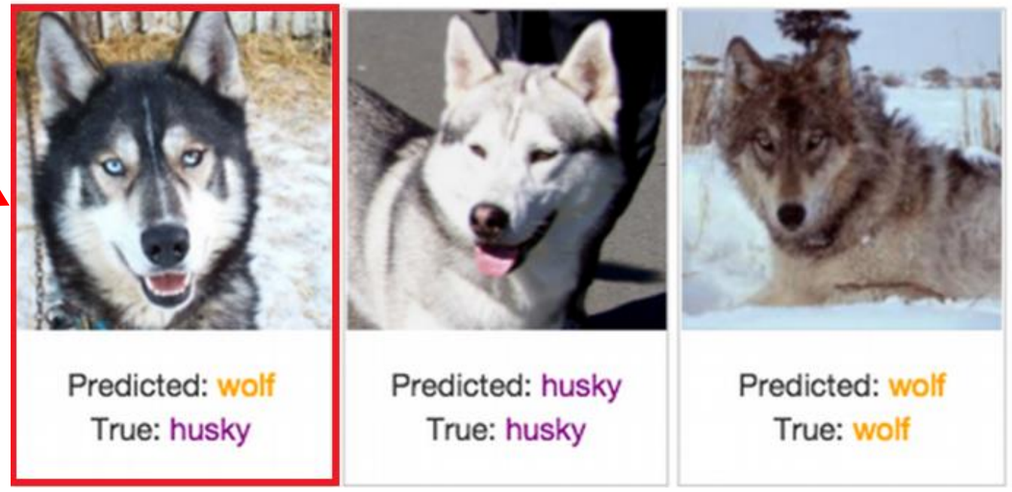


https://www.darpa.mil/program/explainable-artificial-intelligence

Don’t Mistake Model Performance for Trustworthiness

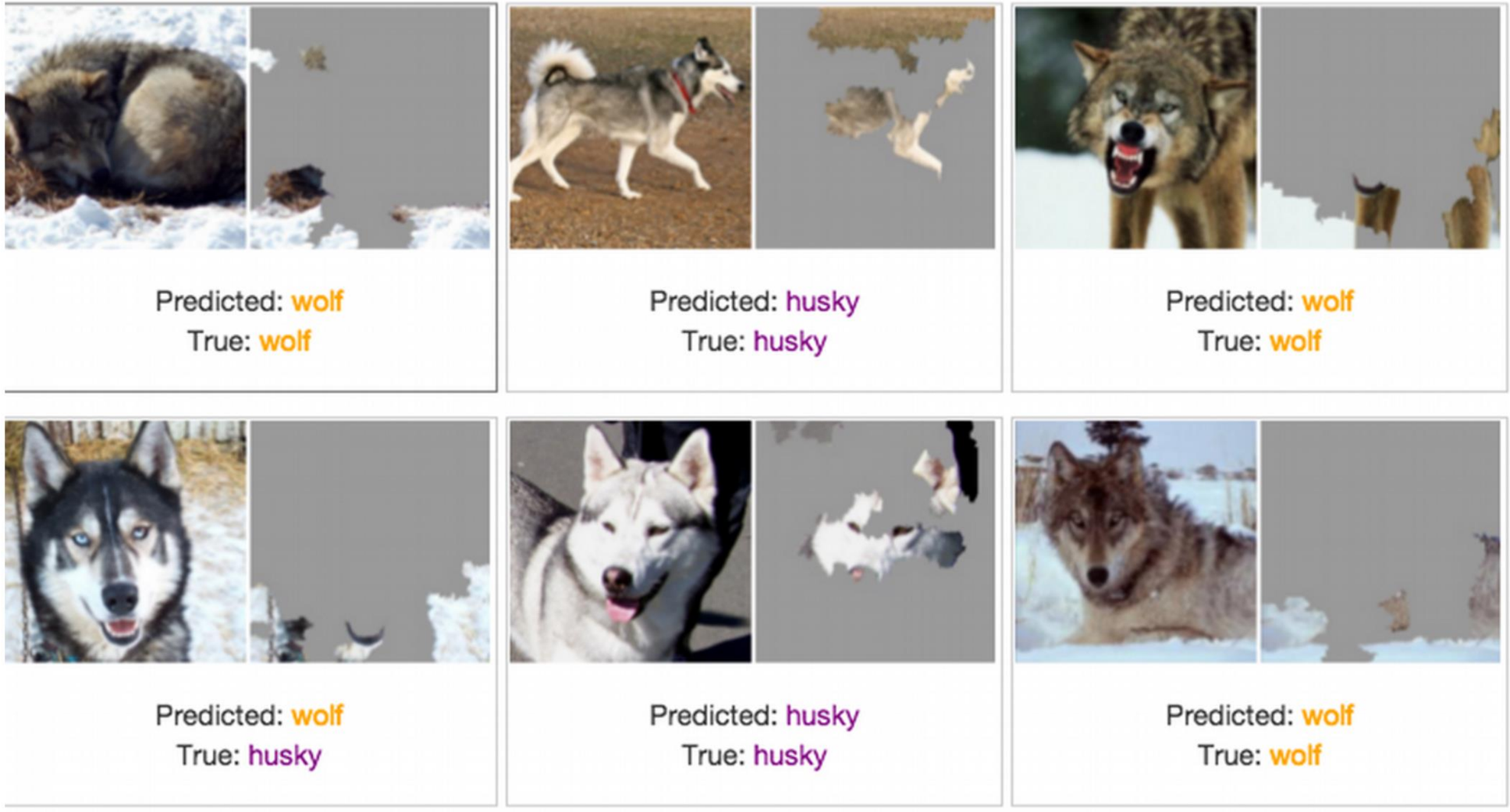


Only One Mistake!



https://theblue.ai/blog/lime-models-explanation/

Don’t Mistake Model Performance for Trustworthiness



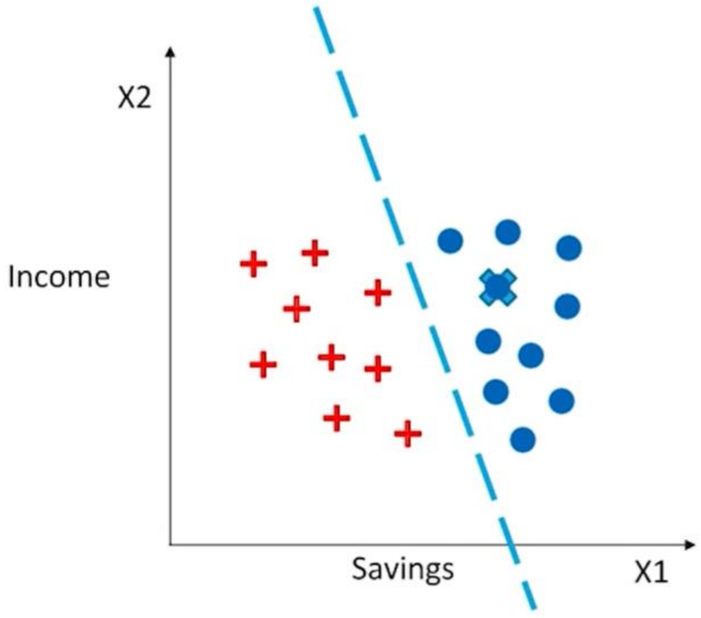
https://theblue.ai/blog/lime-models-explanation/

	Interpretable	Accurate
Complex model	✗	✓
Simple model	✓	✗

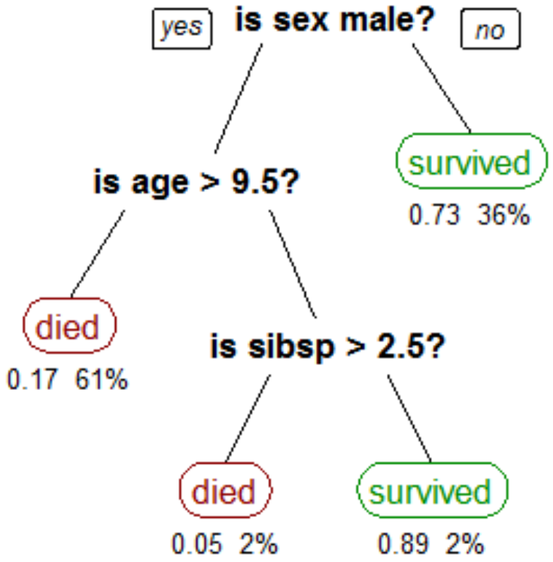
Interpretable or accurate: **choose one**.

Some Models Are Easy to Interpret

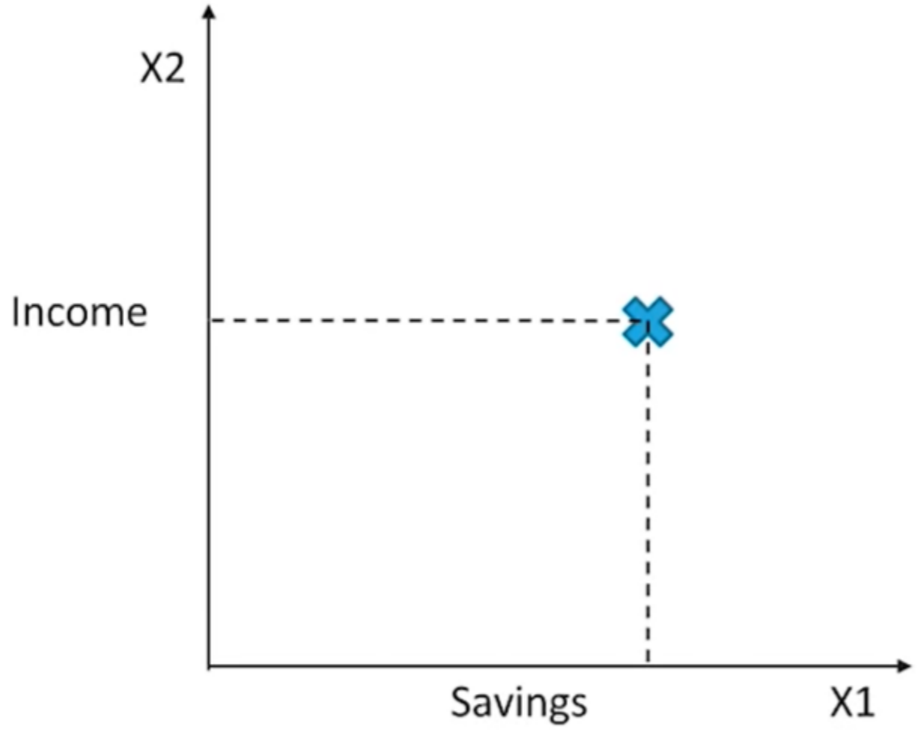
Linear / Logistic Regression



Single Decision Tree

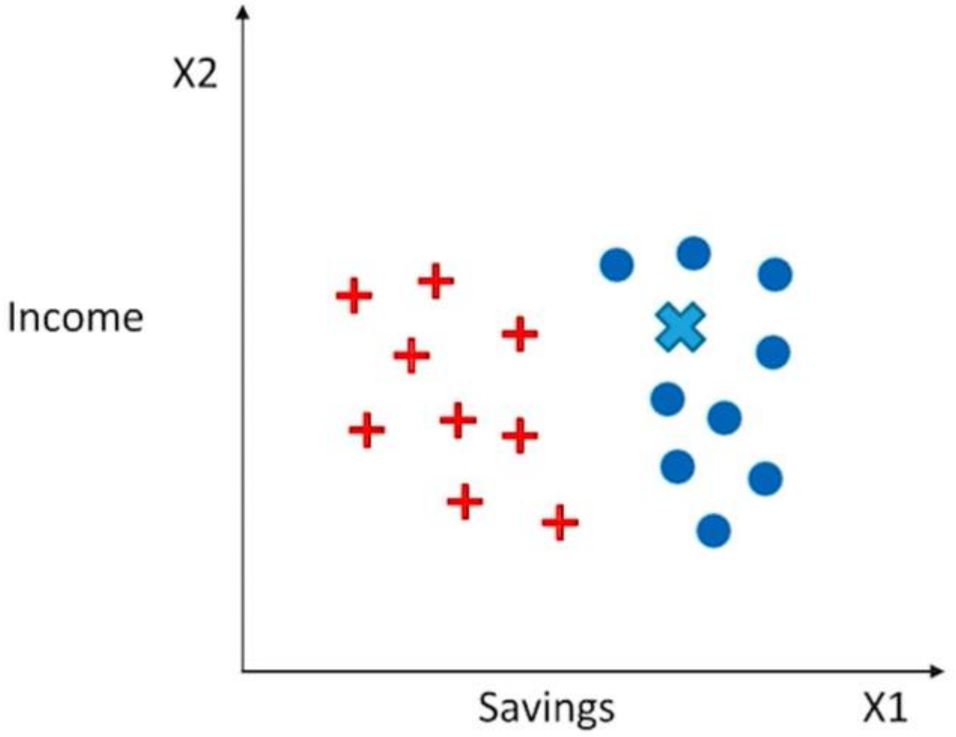


Will the loan default?



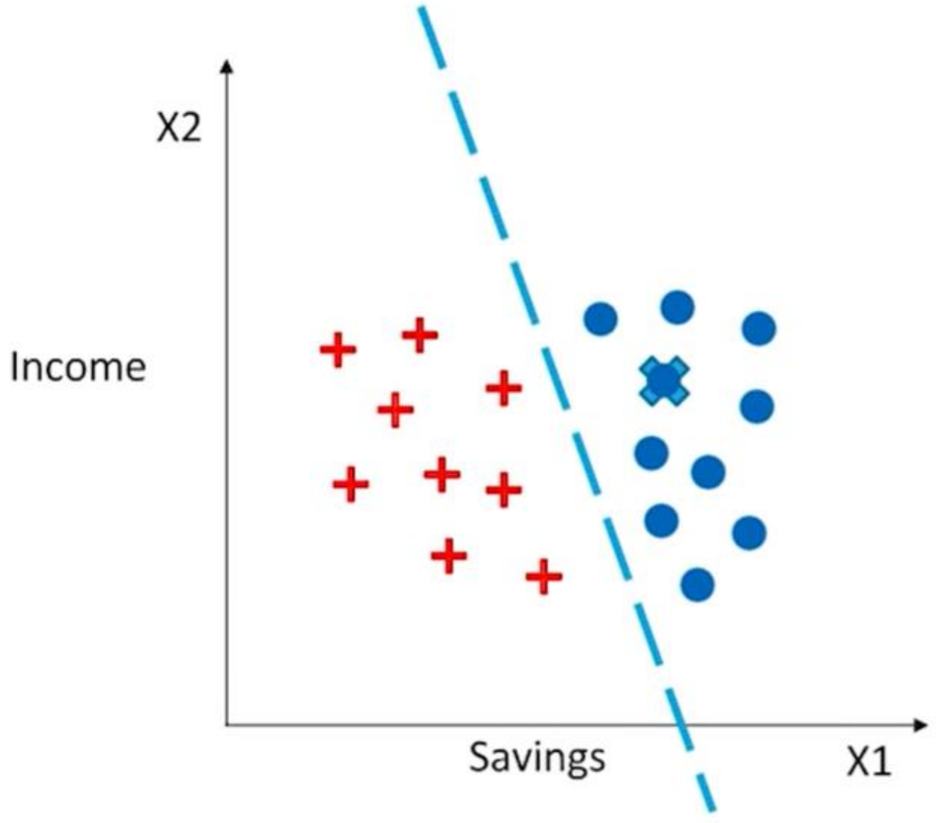
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Get Historical Data



Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Linear Classifiers



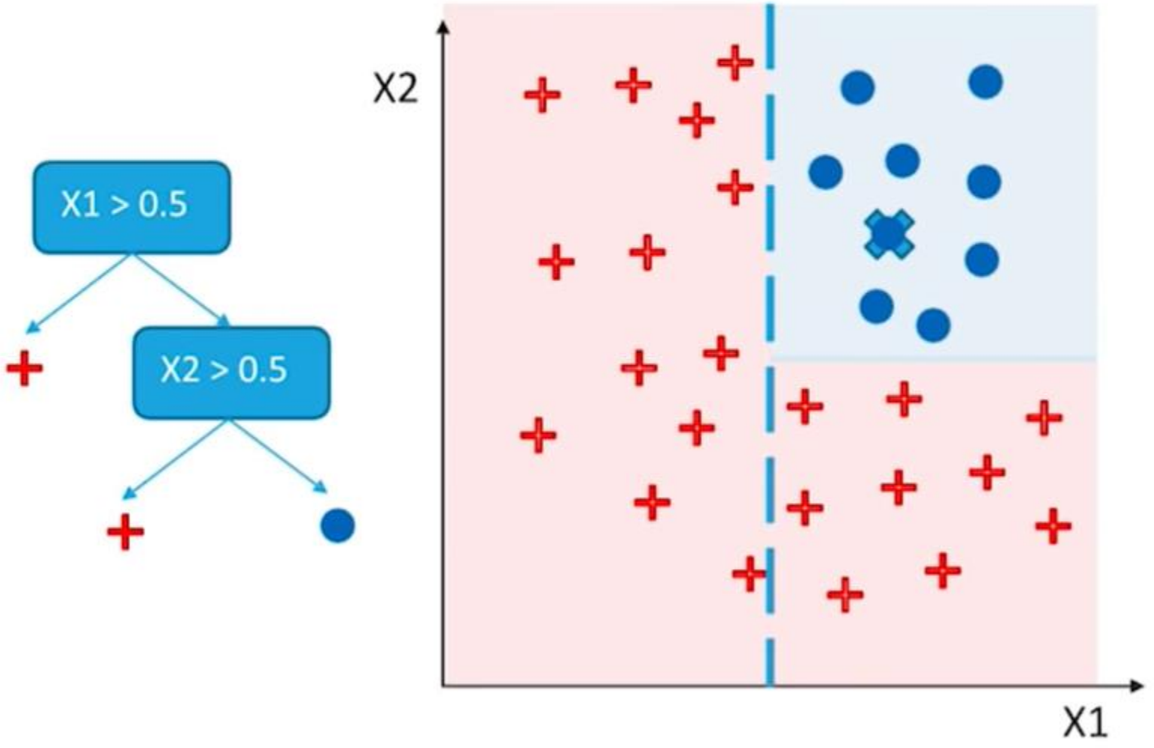
YOU CAN INTERPRET IT!

IF 10X1 + X2 - 5 > 0

OTHERWISE

Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Decision trees

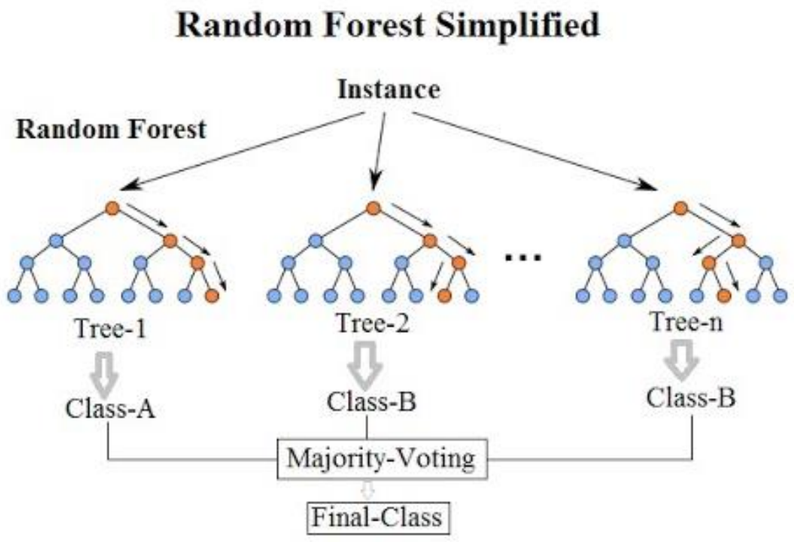


YOU CAN STILL INTERPRET IT!

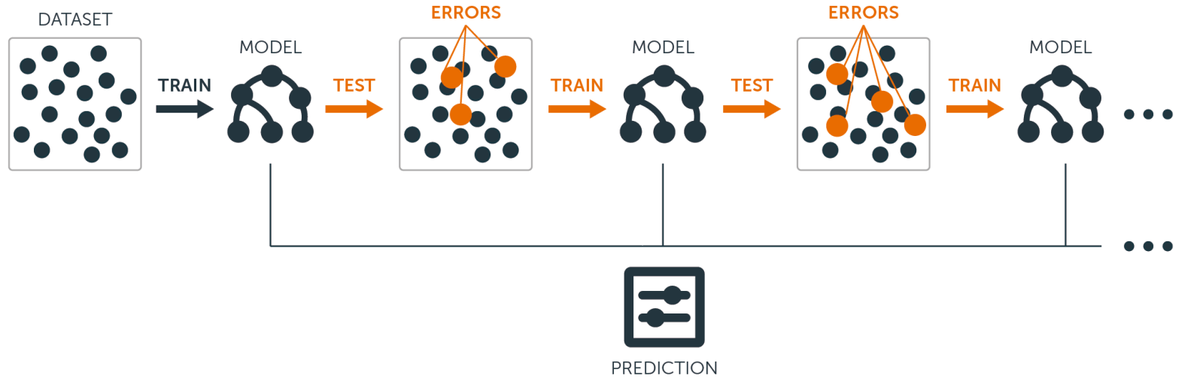
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Some Models Are Harder to Interpret

Random Forest



Boosted Trees



Some Models are REALLY HARD to Interpret

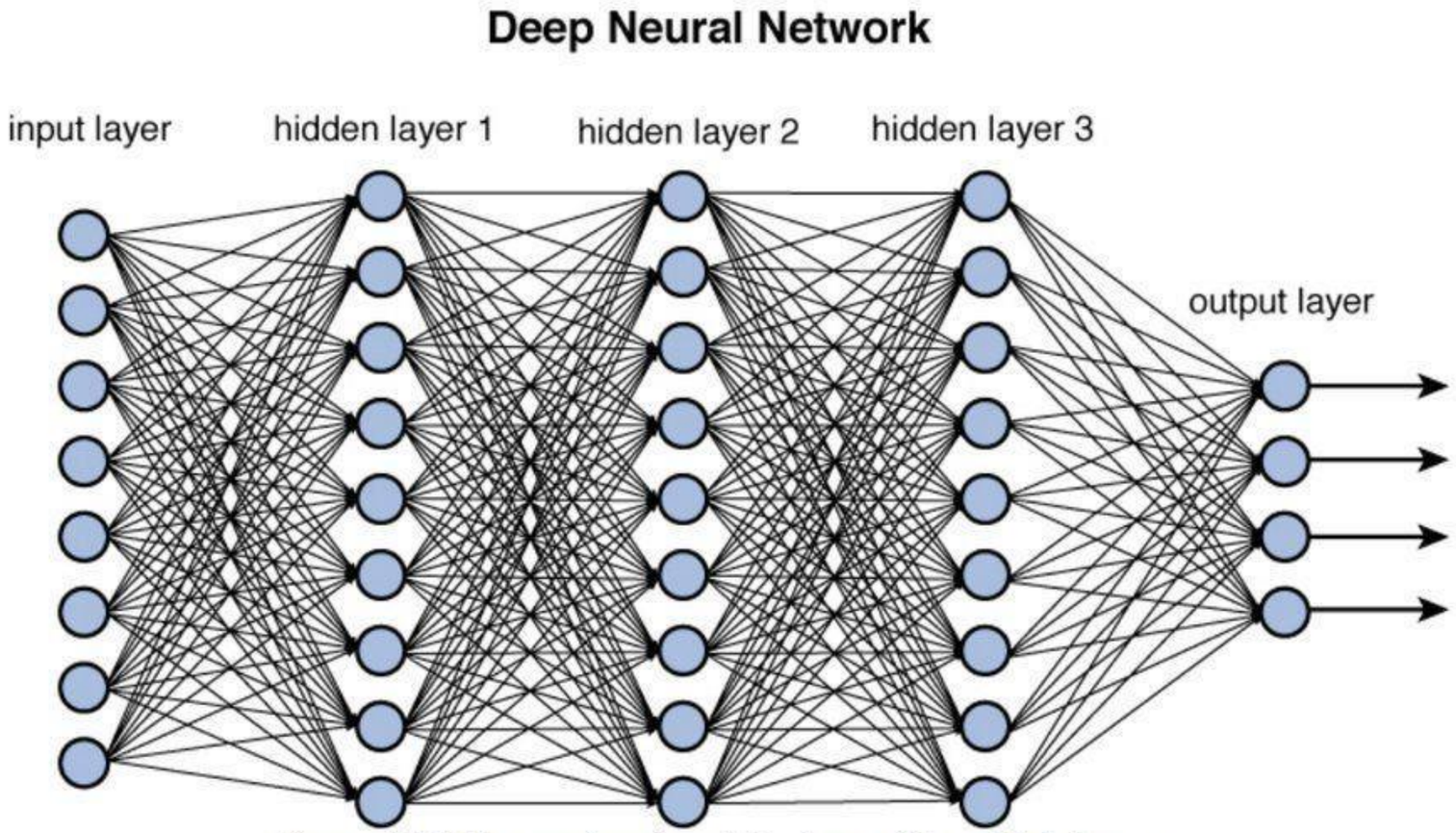


Figure 12.2 Deep network architecture with multiple layers.

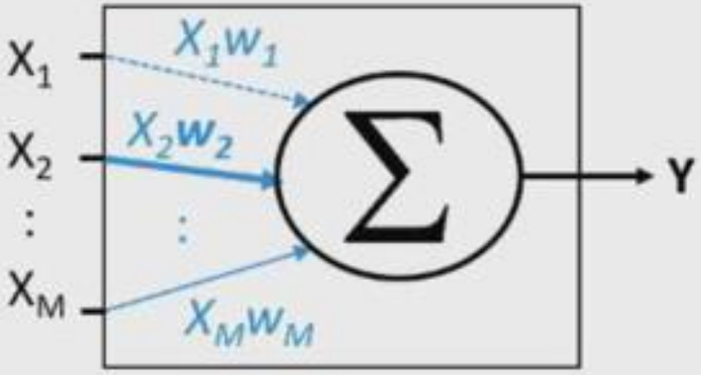
Explainable to Whom?

Data Scientist

End User

Types Of Explainers

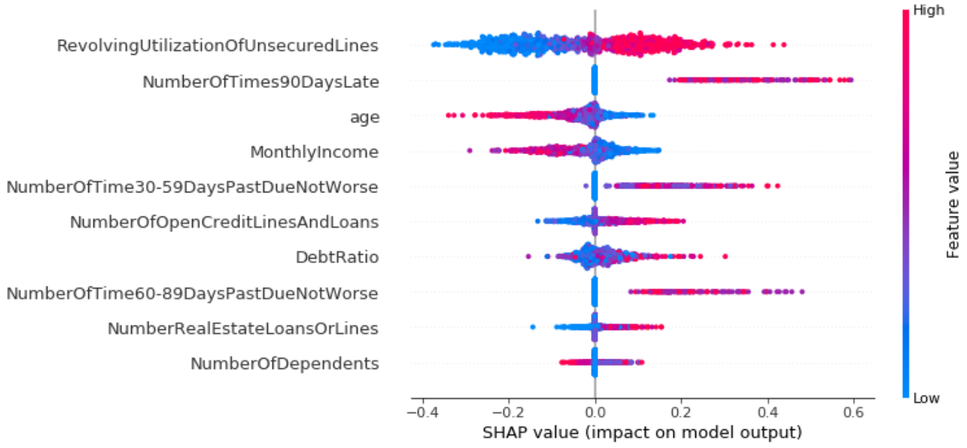
Glass Box



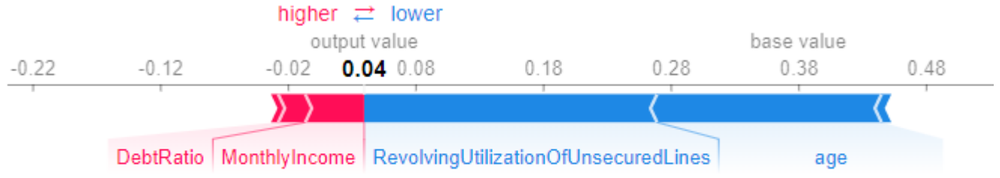
Black Box

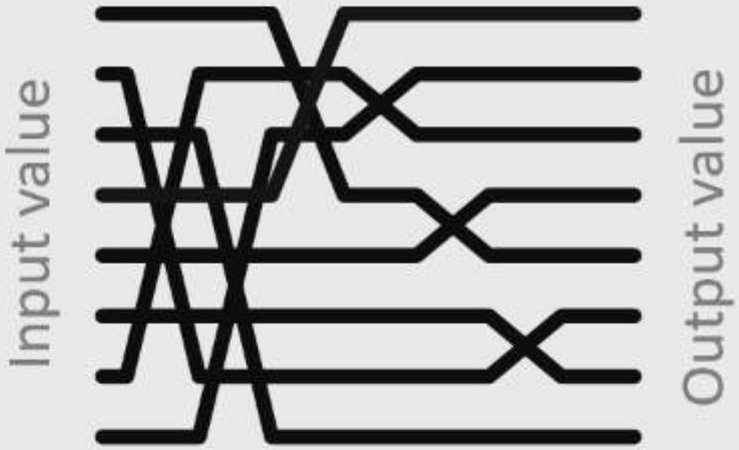
Types Of Explanations

Global
(Model-wide)

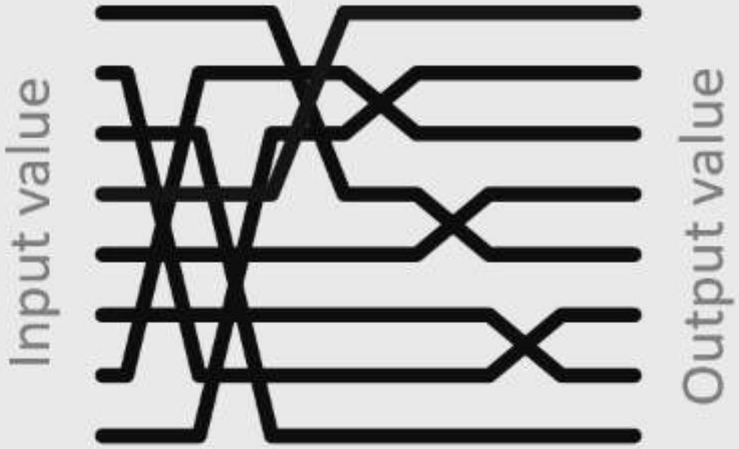


Local
(Single Prediction)





Complex models are inherently complex!



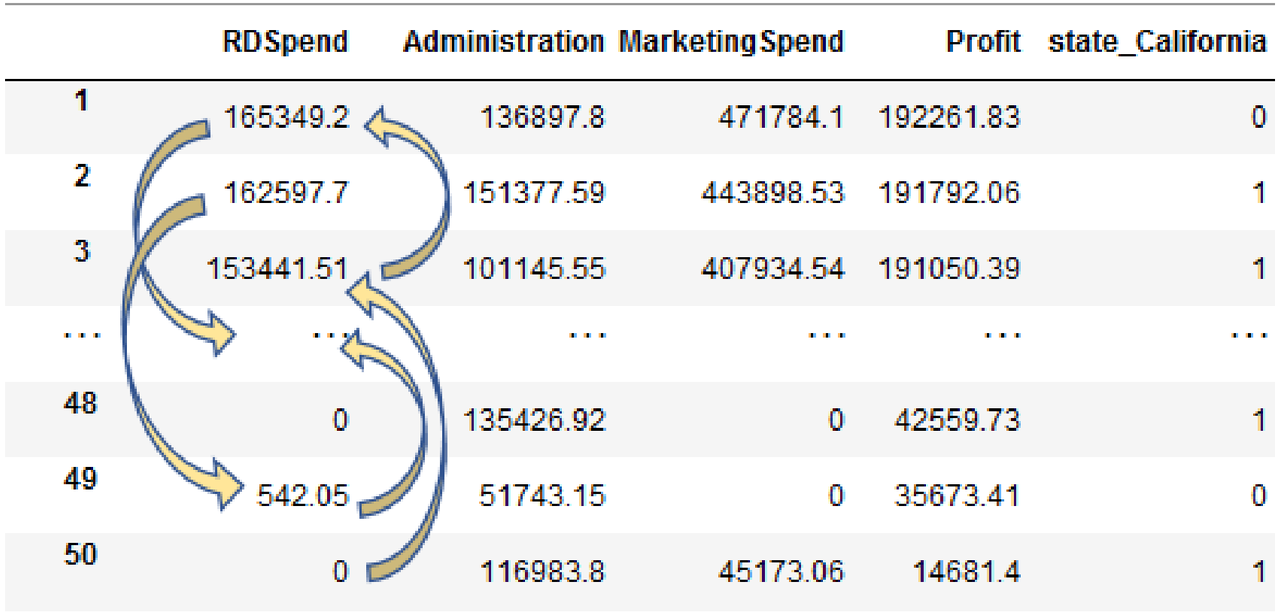
Complex models are inherently complex!

But a single prediction involves only a small piece of that complexity.

Permutation Importance

Permutation Importance

- Global, Black Box
- Measure model performance decreases when feature is not available (permuted)



	RDSpend	Administration	MarketingSpend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
…	…	…	…	…	…
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001 (available online at https://www.stat.berkeley.edu/%7Ebreiman/randomforest2001.pdf).

eli5 0.10.1

```
pip install eli5
```

Latest version

Last released: Aug 29, 2019

Debug machine learning classifiers and explain their predictions

Navigation

Project description

Release history

Download files

Project links

Homepage

Statistics

Project description

pypi v0.10.1 build passing codecov 97% docs failing

ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions.

It provides support for the following machine learning frameworks and packages:

Demo

Permutation Importance with `eli5`

Permutation Importance - Advantages

- Global insight into the model
- Useful for feature selection
- Very fast

Permutation Importance - Disadvantages

- Correlated features are a problem
- Not clear whether to use training or testing data

https://christophm.github.io/interpretable-ml-book/feature-importance.html#feature-importance-data

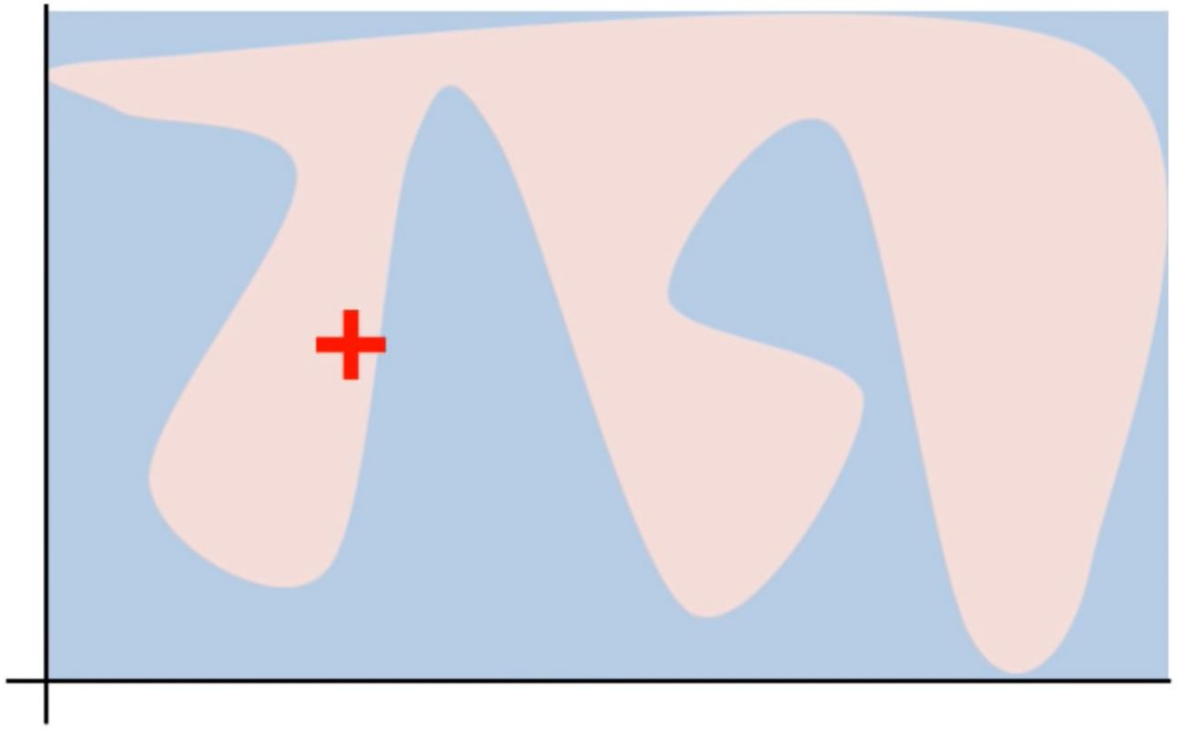
LIME

LOCAL

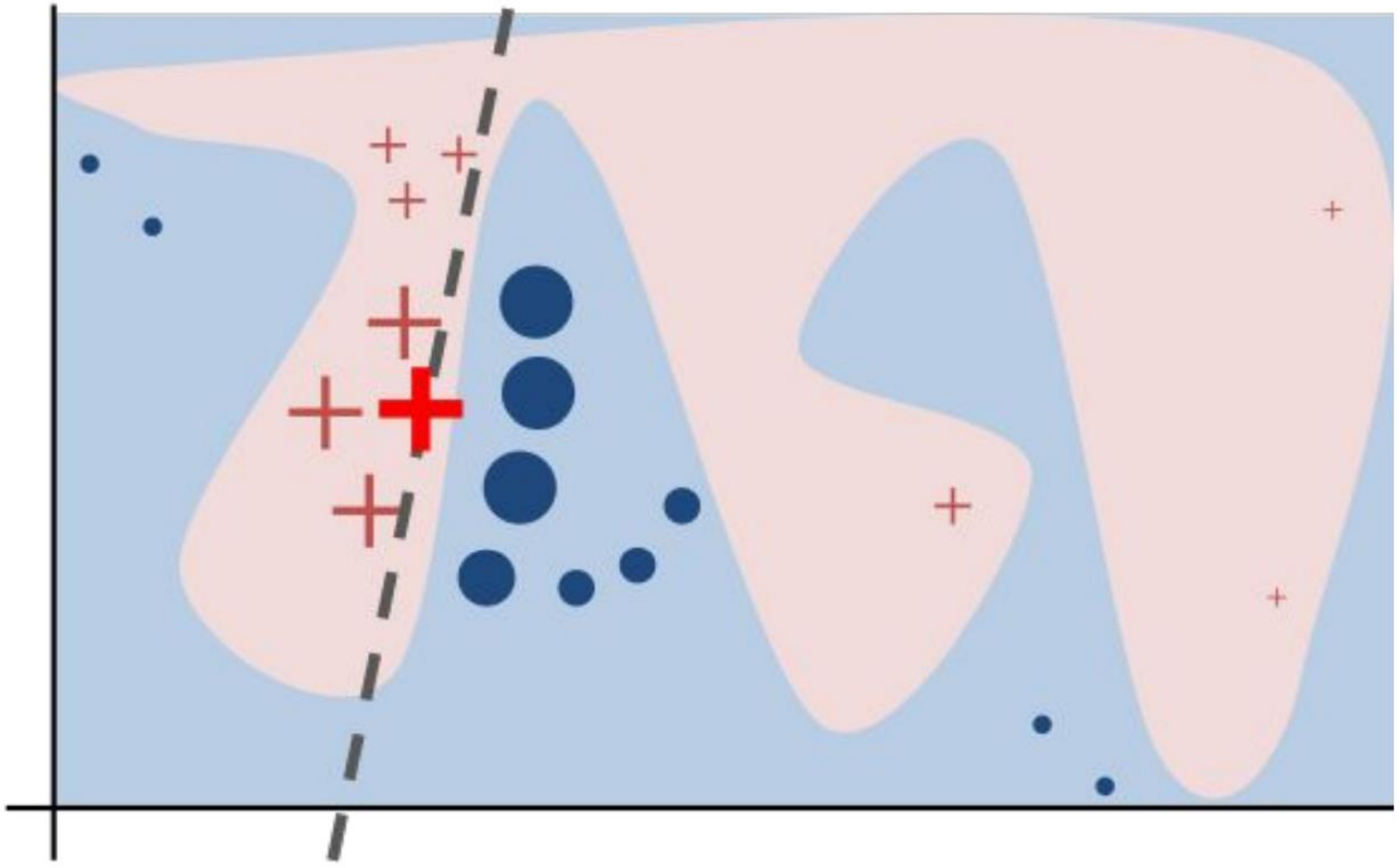
INTERPRETABLE

MODEL-AGNOSTIC

EXPLANATIONS

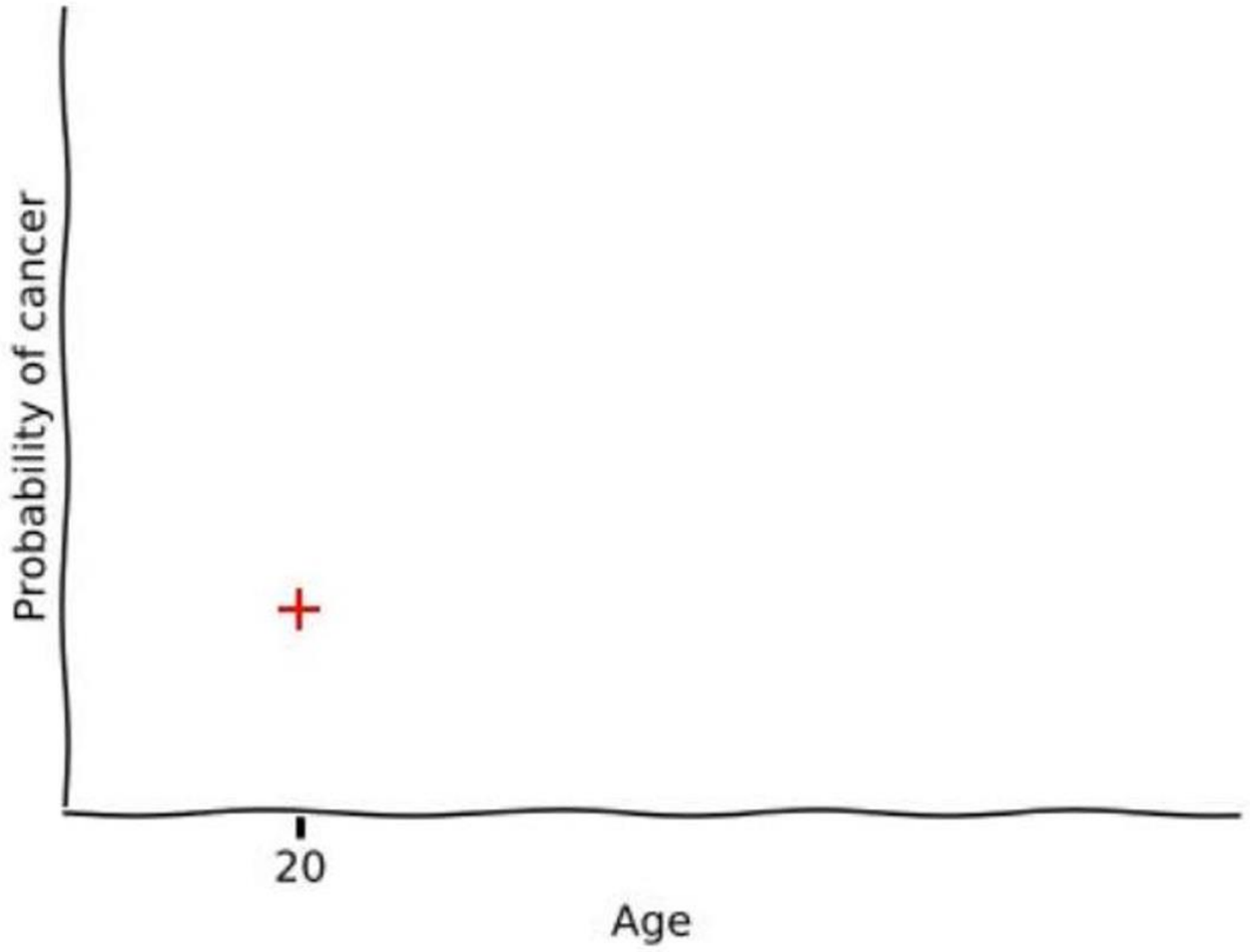


How LIME Works



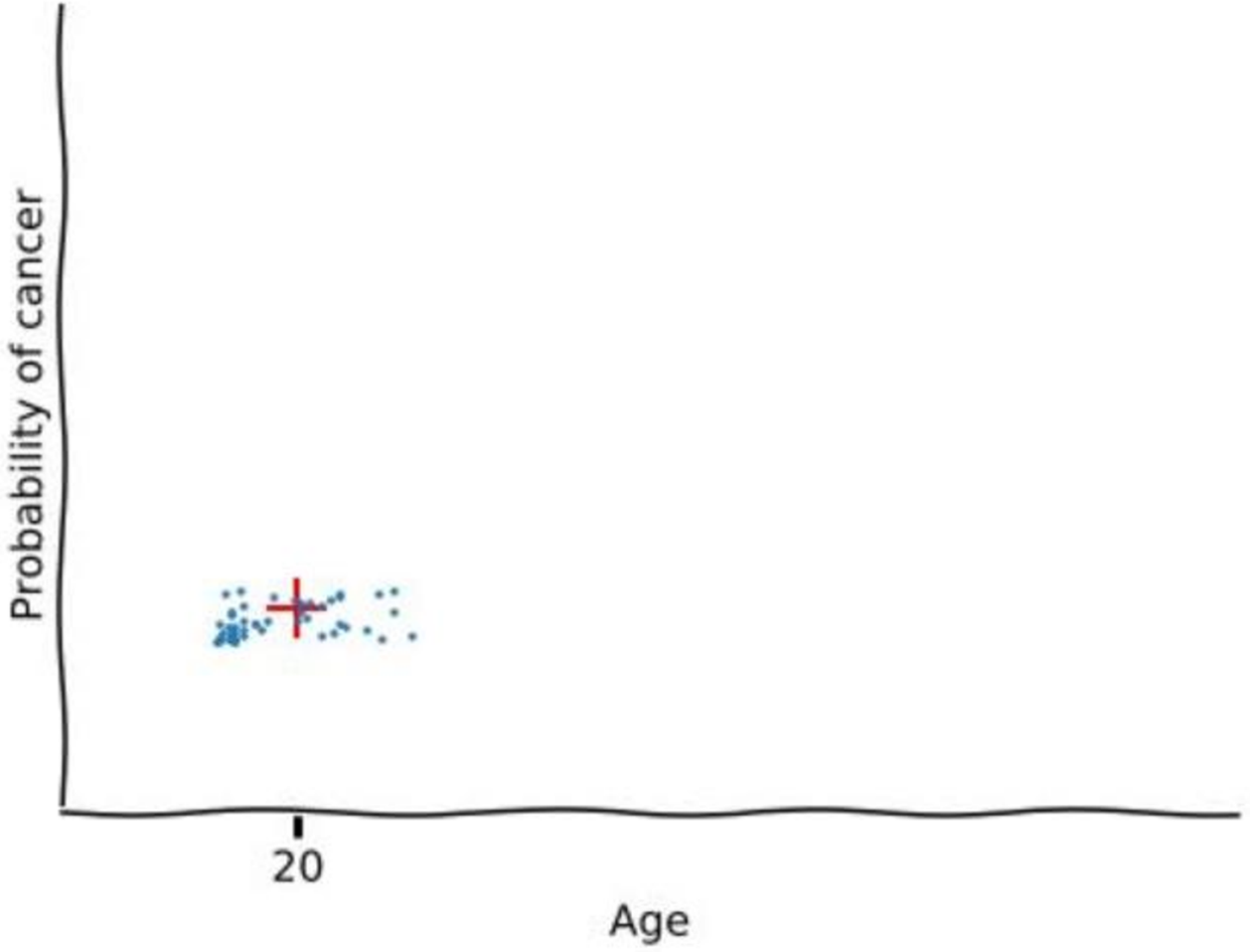
"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

How LIME Works

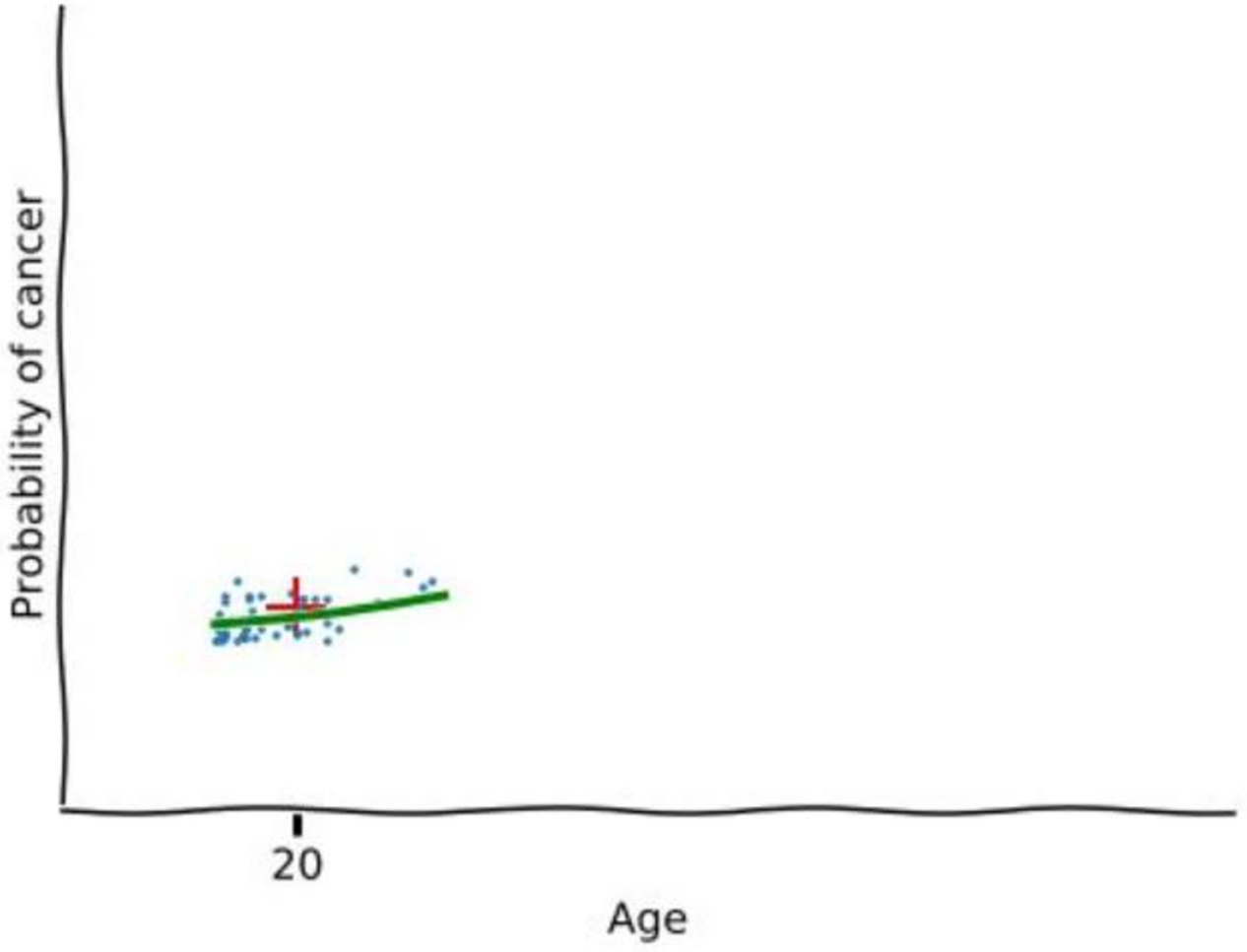


Open the Black Box: an Introduction to Model Interpretability with LIME and SHAP - Kevin Lemagnen https://www.youtube.com/watch?v=C80SQe16Rao

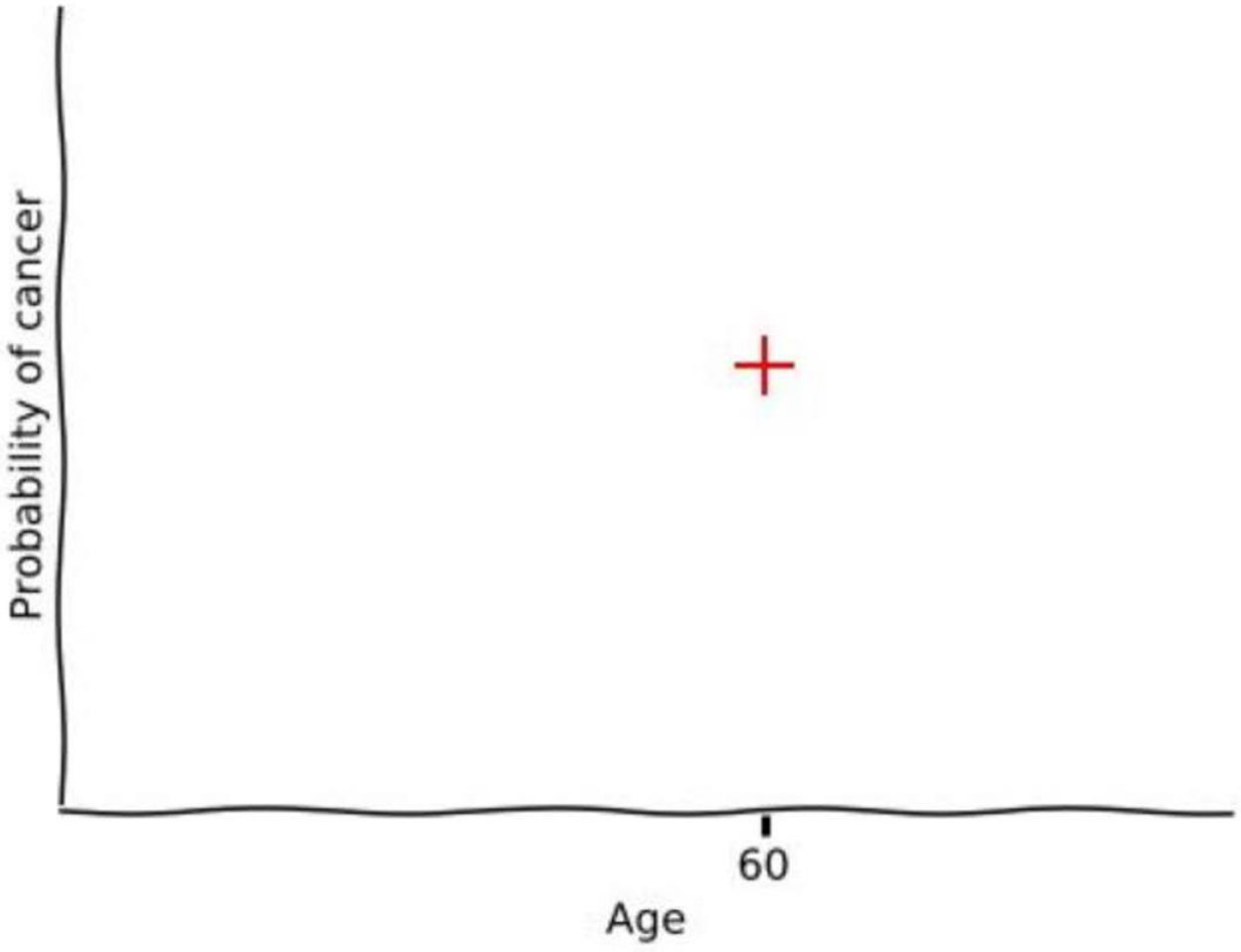
How LIME Works



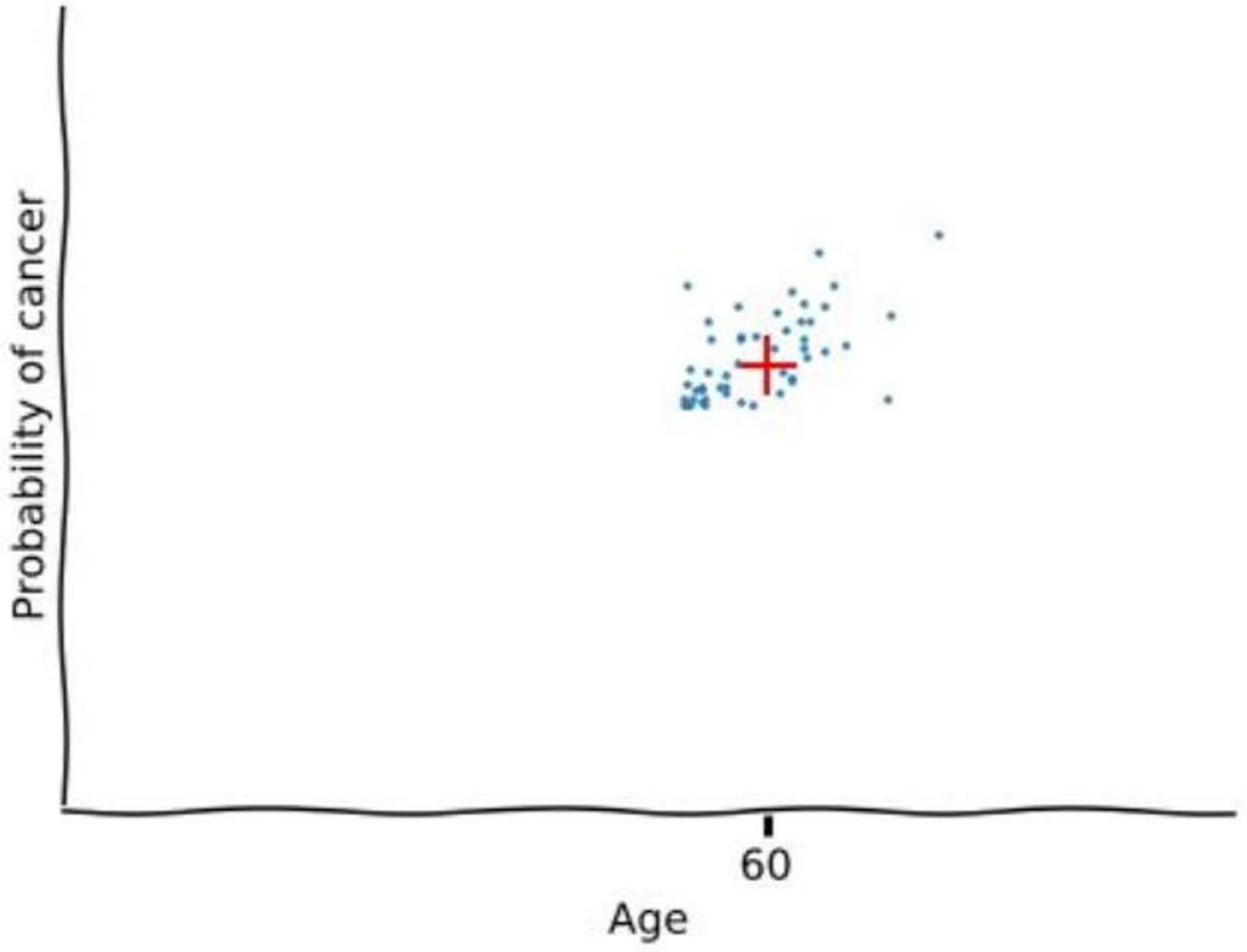
How LIME Works



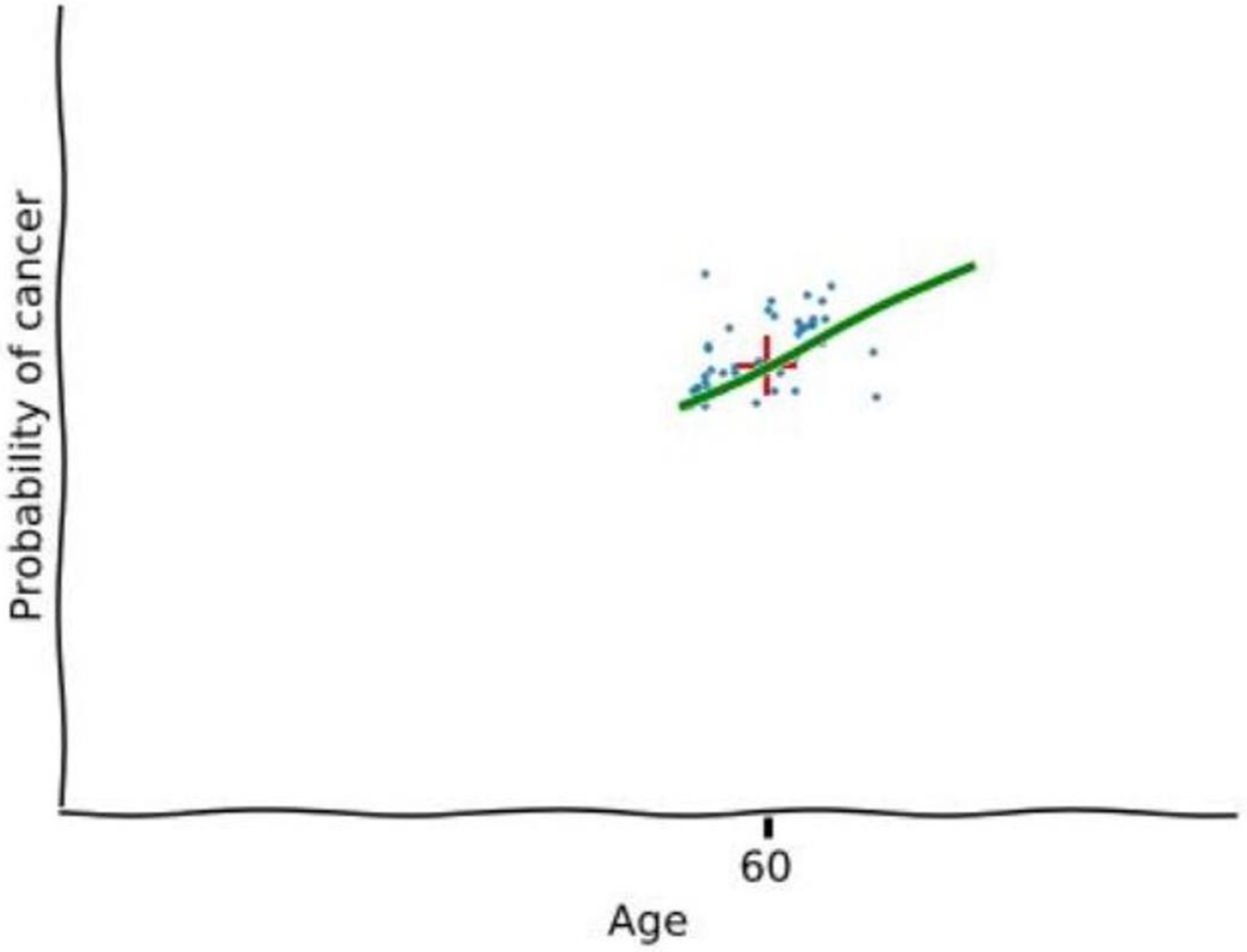
How LIME Works



How LIME Works



How LIME Works



How LIME Works

1. Choose an **observation** to explain
2. Create new dataset around observation by sampling from training data distribution
3. Calculate distances between new points and observation, that's our measure of **similarity**
4. Use model to predict class of the new points
5. Find the subset of **m** features that has the strongest relationship with our target class
6. Fit a linear model on fake data in **m** dimensions weighted by similarity
7. Weights of linear model are used as explanation of decision

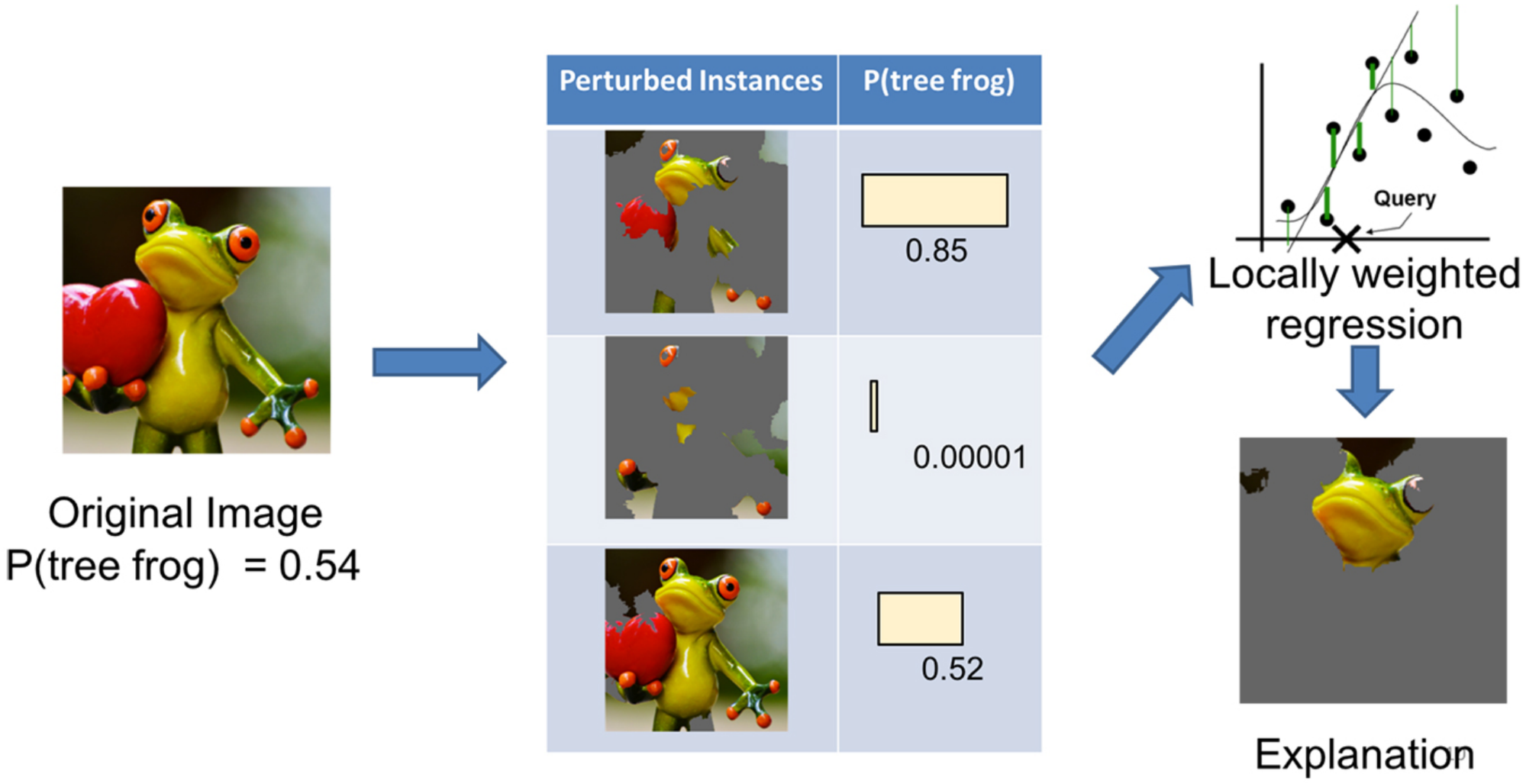
Open the Black Box: an Introduction to Model Interpretability with LIME and SHAP - Kevin Lemagnen https://www.youtube.com/watch?v=C80SQe16Rao

LIME Also Works With Image Data

Original Image

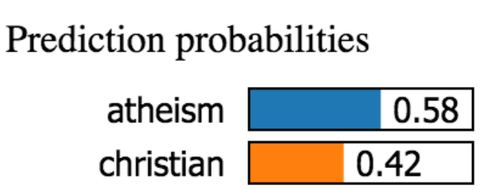
Interpretable Components

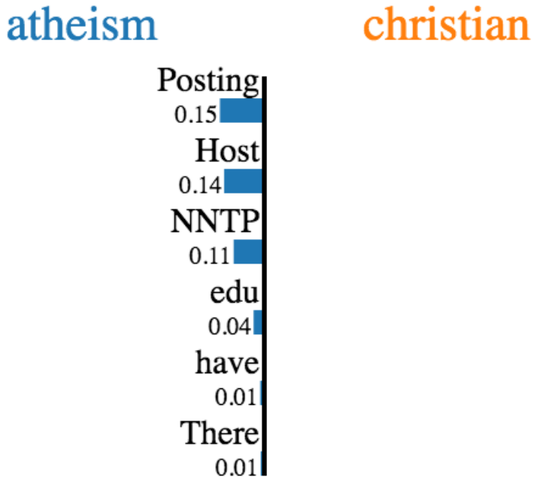
LIME Also Works With Image Data



https://github.com/marcotcr/lime/

LIME Also Works With Text Data





Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

https://github.com/marcotcr/lime/

Demo

Local Explanations with `LIME`

LIME - Advantages

- Local explanations, useful for end users
- Works with tabular data, images and text
- Black box

LIME - Disadvantages

- Depends on random sampling, can be unstable
- Linear model surrogate can be inaccurate
  - This can be tested with R-squared score
- Can be slow for complex models

https://christophm.github.io/interpretable-ml-book/feature-importance.html#feature-importance-data

SHAP

SHAP

Lundberg and Lee. A unified approach to interpreting model predictions
NeurIPS 2017 *(oral presentation)*

Lundberg and Lee. An unexpected unity among methods for interpreting model predictions
NeurIPS Workshop on Interpretable Machine Learning in Complex Systems 2016 *(best paper award)*

0

Base rate

20%

0

$E[f(x)]$

Base rate

20%

$E[f(x)]$

Prediction for John

55%

$f(x)$

0

Base rate

20%

$E[f(x)]$

Prediction for John

55%

$f(x)$

0

How did we get here?

20%

$E[f(x)]$

0

ϕ_0

Base rate

35%

0 $E[f(x)]$ $E[f(x) \mid x_1]$

ϕ_0 ϕ_1

Age = 20

70%

0 $E[f(x)]$ $E[f(x) \mid x_1, x_2$

ϕ_0 ϕ_1 ϕ_2

Day trader

90%

0 $E[f(x)]$ $E[f(x) \mid x_1, x_2, x_3]$

ϕ_0 ϕ_1 ϕ_2 ϕ_3

Open accounts = 1

55%

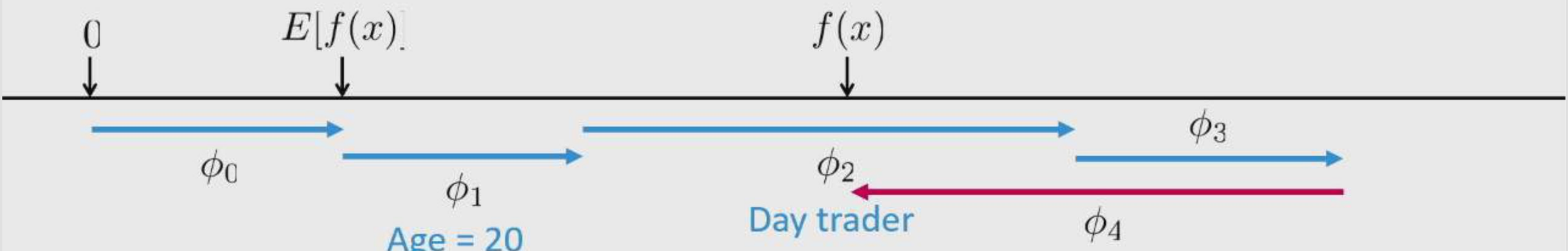
0 $E[f(x)]$ $f(x)$

ϕ_0 ϕ_1 ϕ_2 ϕ_3

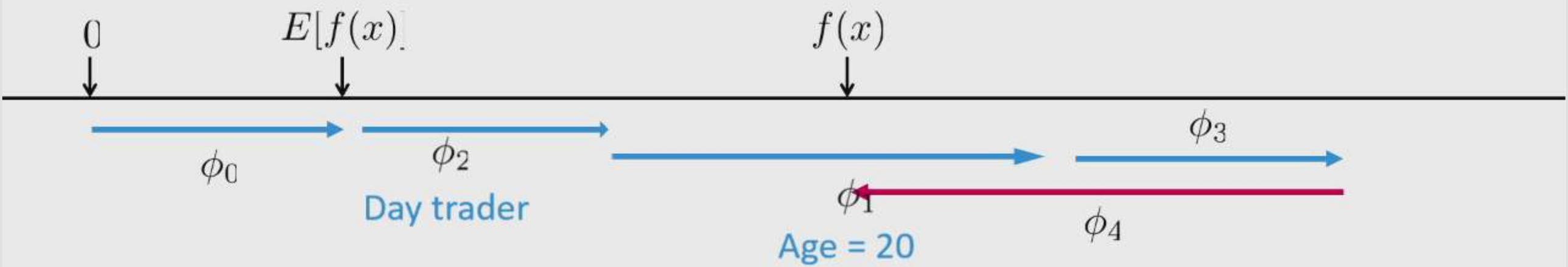
ϕ_4

Capital gains

The order matters!



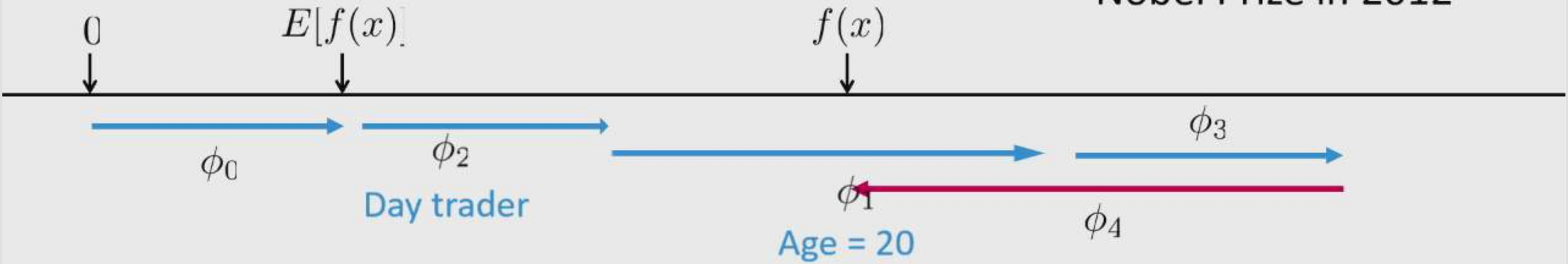
The order matters!



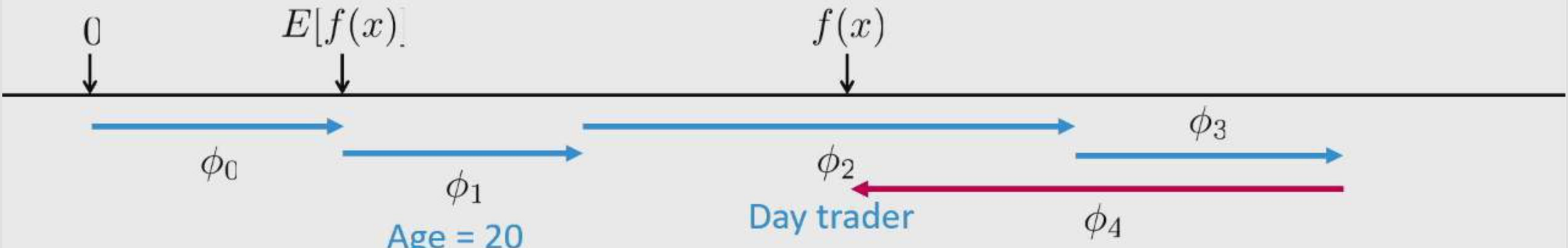
Lloyd Shapley

Nobel Prize in 2012

The order matters!

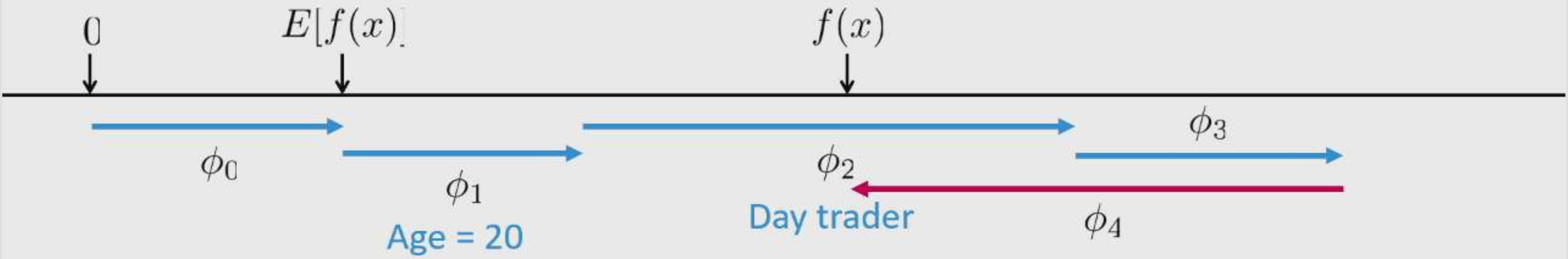


Shapley values result from **averaging over all N! possible orderings**.



SHapley Additive exPlanation (SHAP) values

Shapley values result from **averaging over all N! possible orderings**.



How SHAP Works

The weight of each feature is computed using the Shapley values method from game theory.

To get the importance of feature X{i}:

- Get all subsets of features S that do not contain X{i}
- Compute the effect on our predictions of adding X{i} to all those subsets

That can be computationally expensive, but SHAP has optimizations for different models (linear, trees, etc..)

Demo

Local And Global Explanations with `shap`

SHAP - Advantages

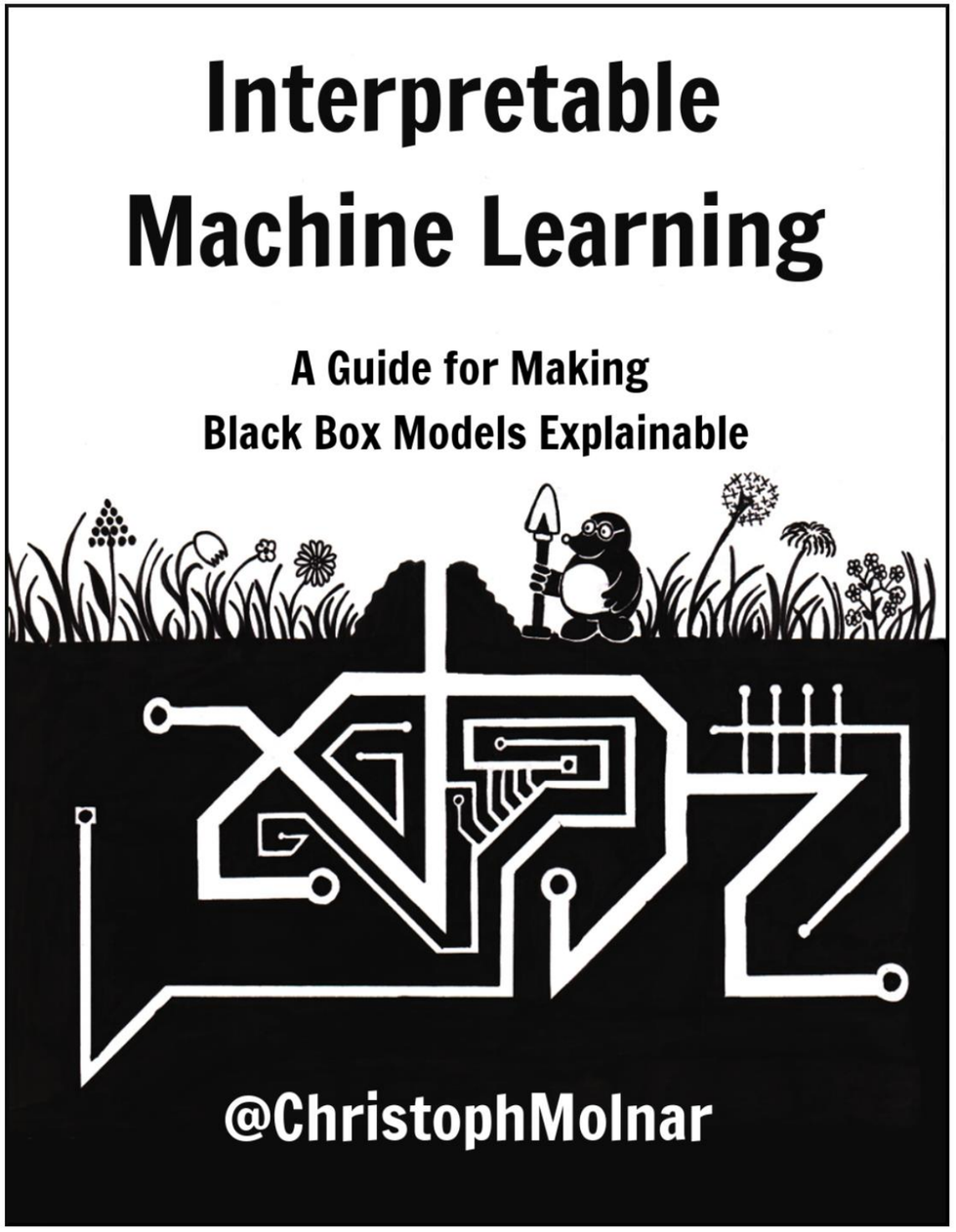
- Solid theoretical foundation in game theory
- Generates both local and global explanations
- Black box
  - optimizations are available for Glass box models

SHAP - Disadvantages

- Ignores feature dependence
- Black box (kernel) SHAP is slow

https://christophm.github.io/interpretable-ml-book/feature-importance.html#feature-importance-data


LIME
DeepLIFT
SHAP
Shapley reg. values
Relevance prop.
QII
Shapley sampling
Saabas



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

Questions?

CODECAMP_ ❤ FEEDBACK

codecamp.ro/feedback

Thank You!