Interpretable Machine Learning

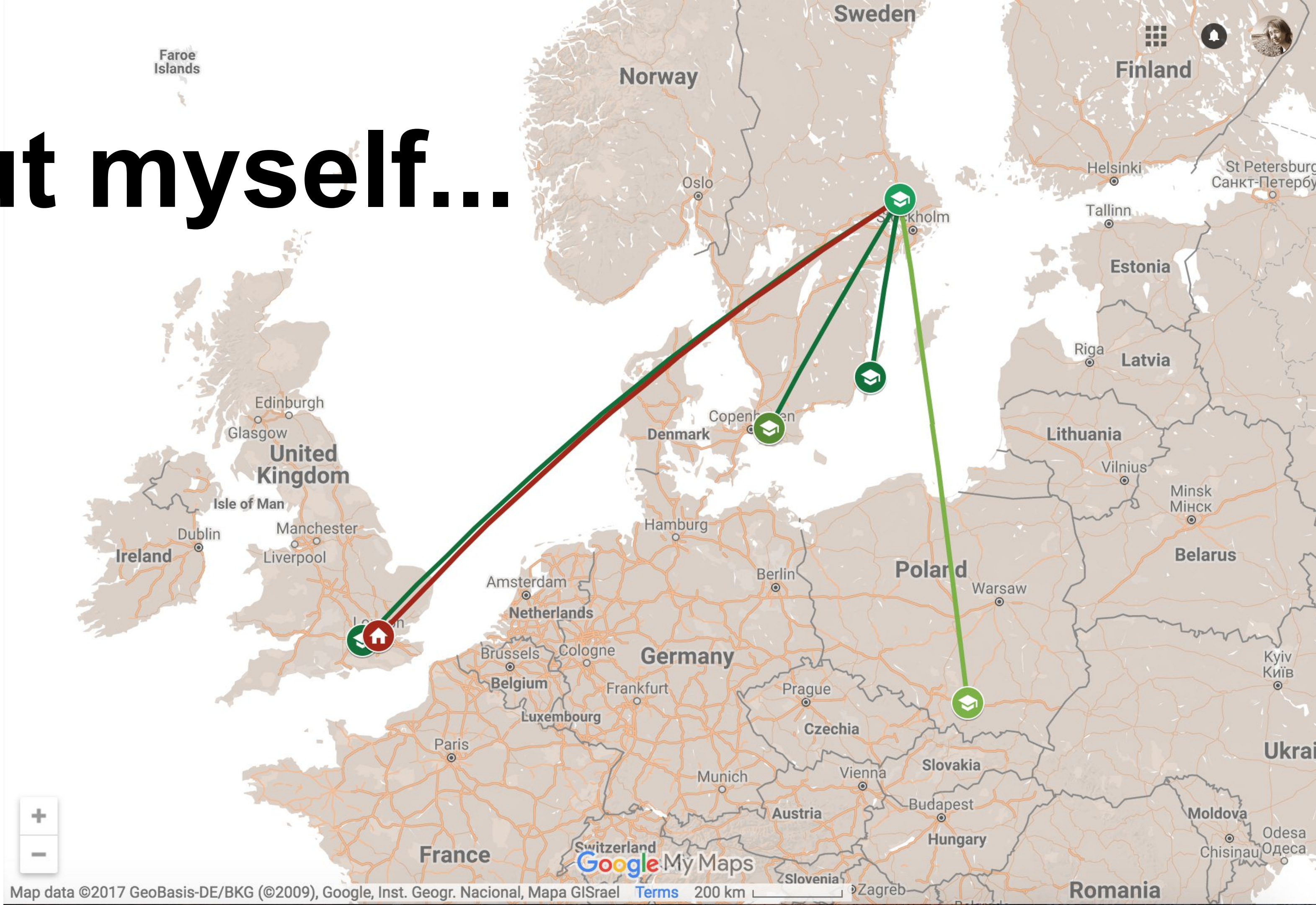
Using LIME Framework

kasia.kulma@aviva.com

Kasia Kulma (PhD), Data Scientist

@KKulma

About myself...



Data

R Ladies

DataCamp

meetup

R-tastic

https://kkulma.github.io

Data

DataCamp

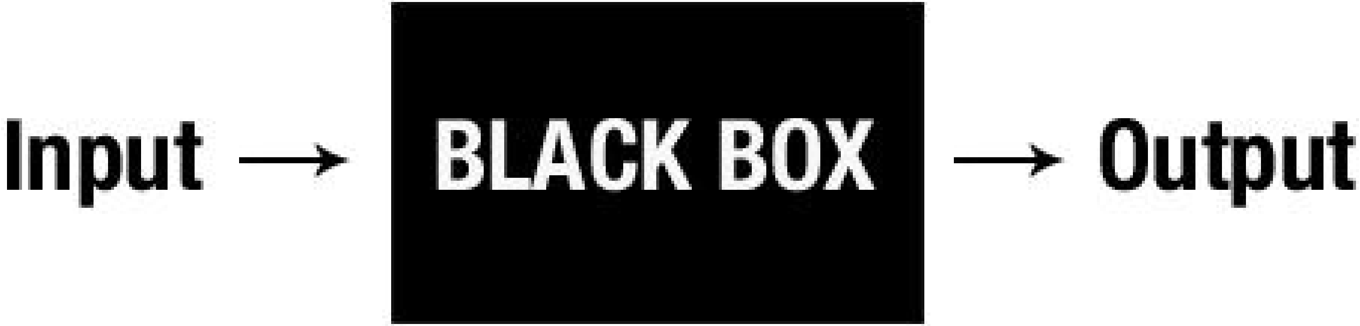
R Ladies

meetup

R-tastic

https://kkulma.github.io

Non-Data



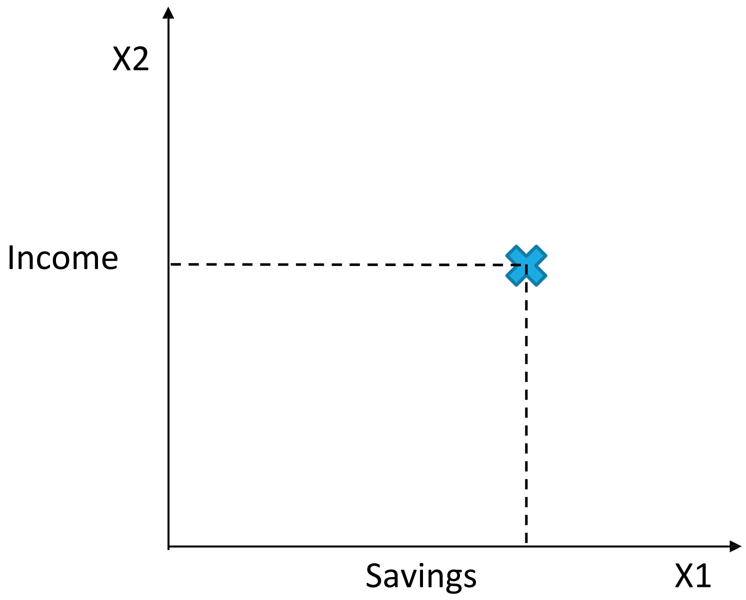
System that performs behaviour but you don’t know how it works

Will the loan default?

Will the loan default?

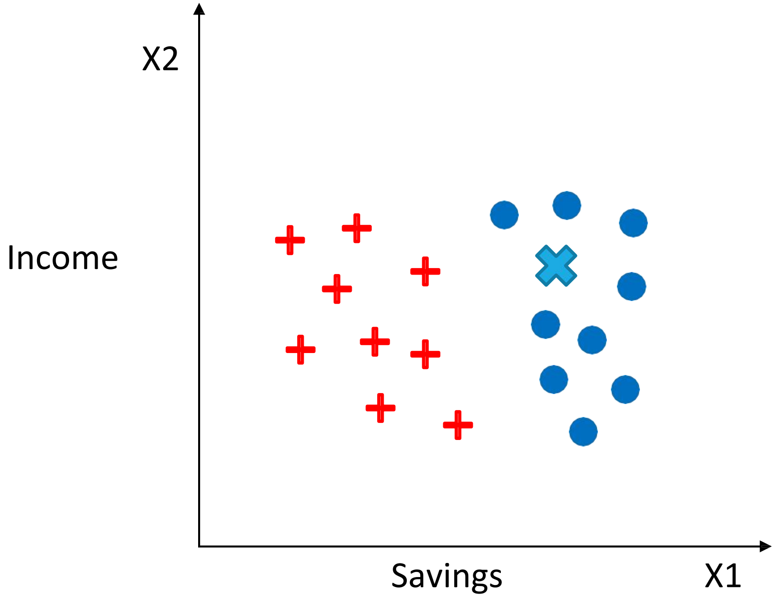
Is it a cancer cell? Will this prisoner commit a crime? Will this machine break down?

Will the loan default?



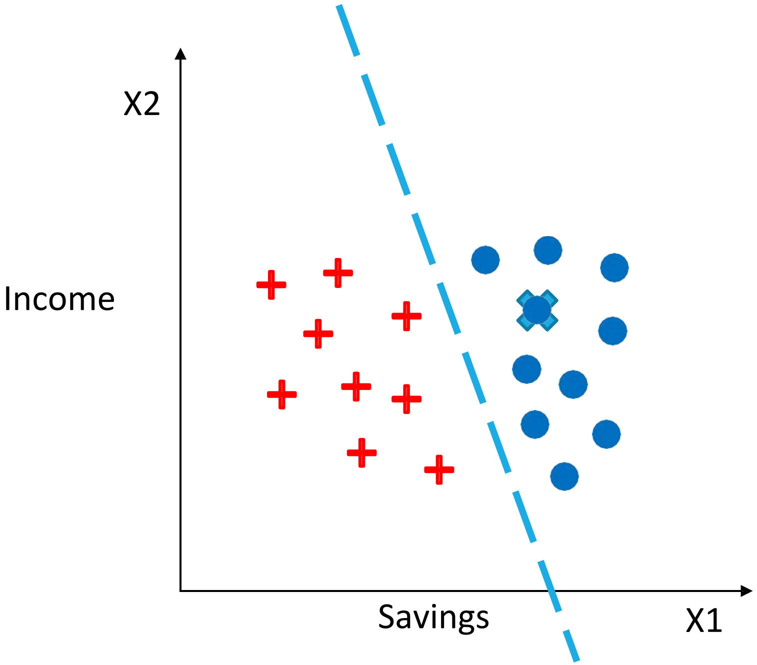
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Get Historical Data



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

Linear Classifiers



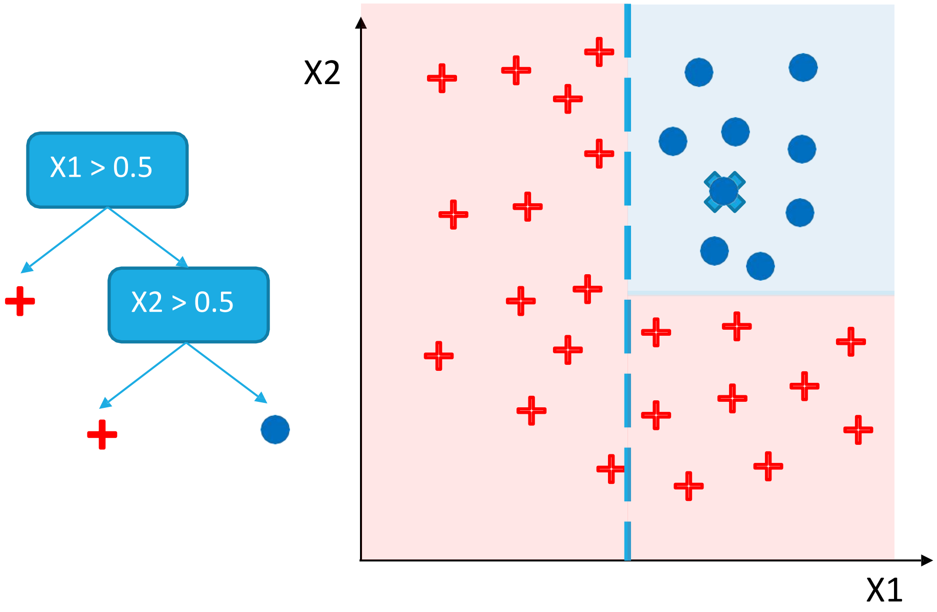
YOU CAN INTERPRET IT!

IF 10X1 + X2 - 5 > 0

OTHERWISE

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

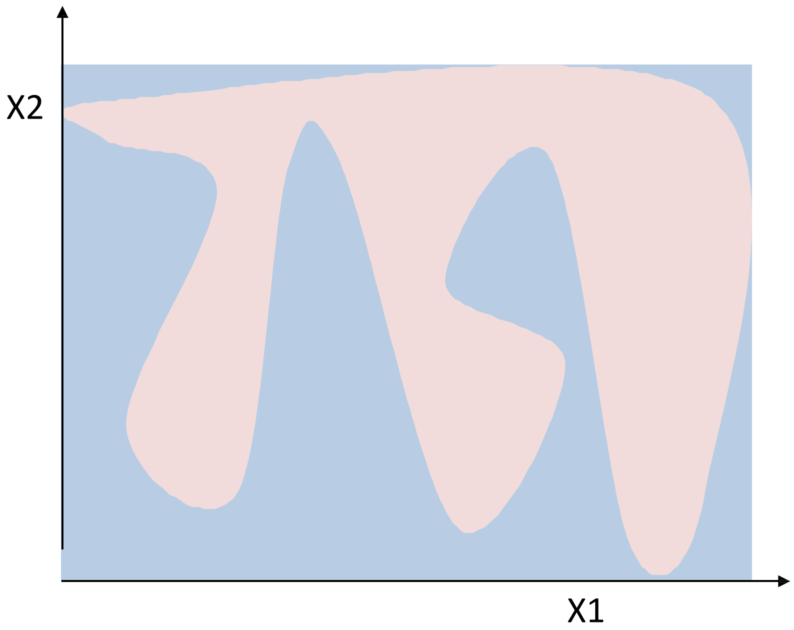
Decision trees



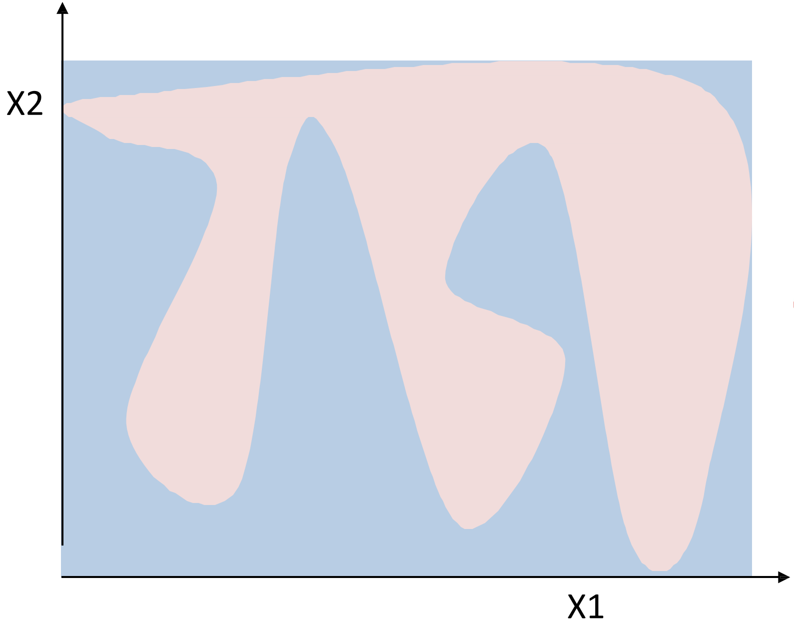
YOU CAN STILL INTERPRET IT!

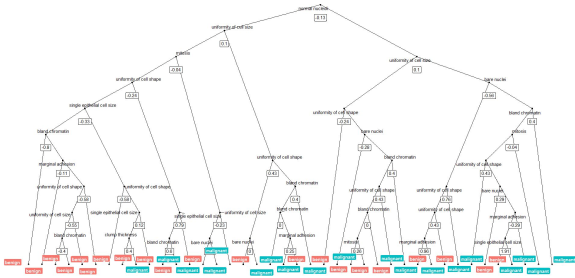
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

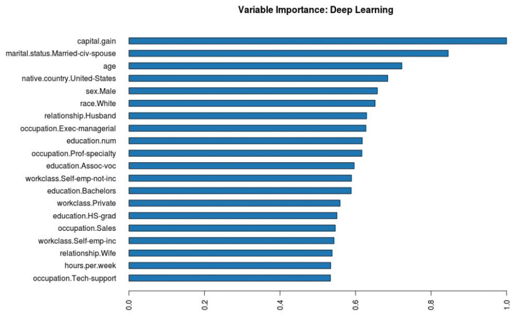
Big Data: More Complexity & More Dimensions



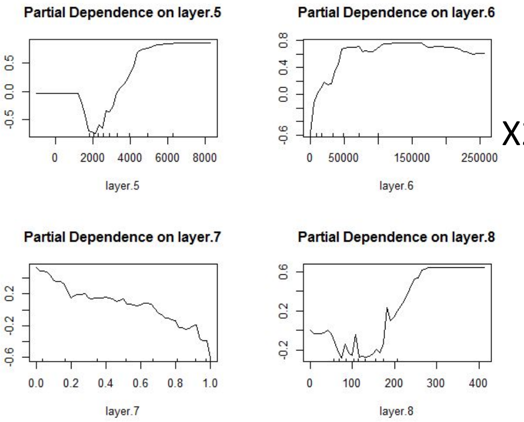
Big Data: More Complexity & More Dimensions

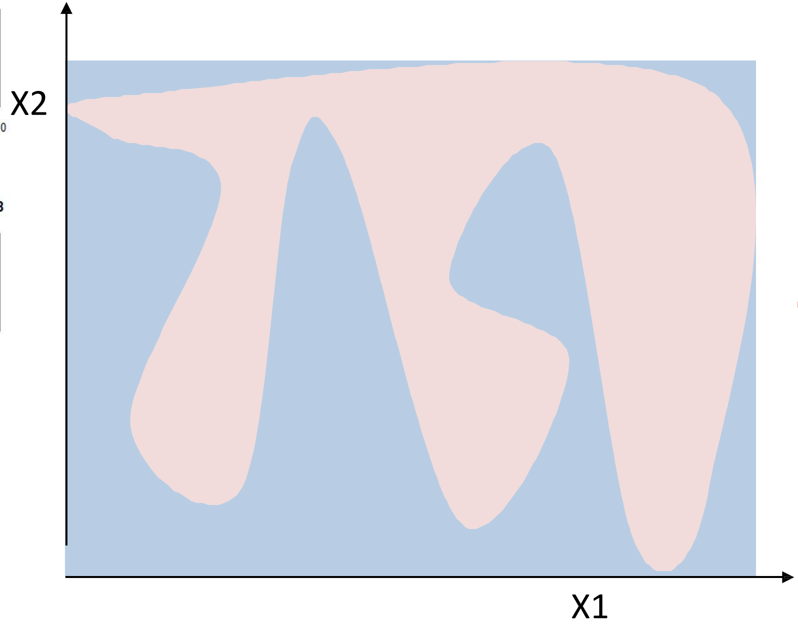


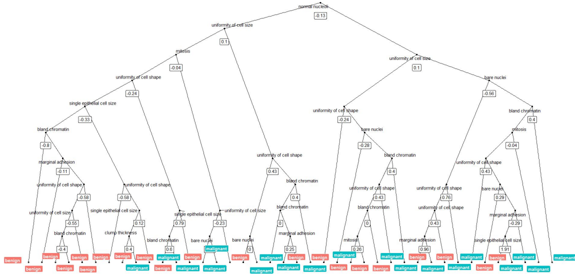


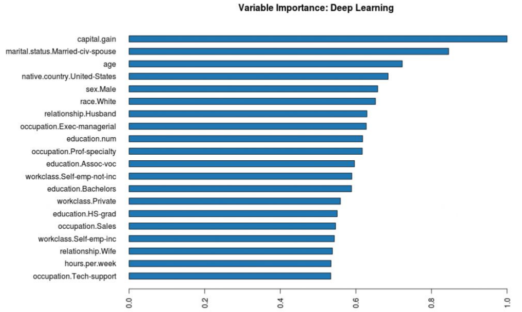


Big Data: More Complexity & More Dimensions

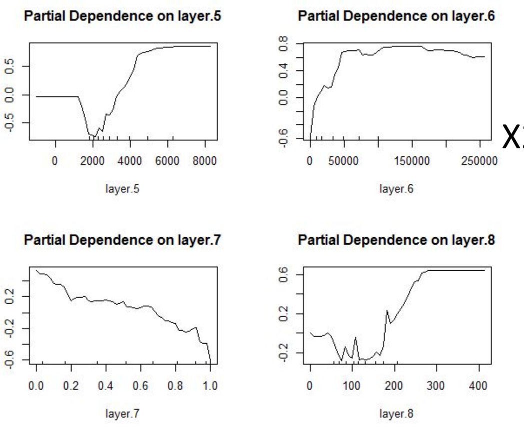


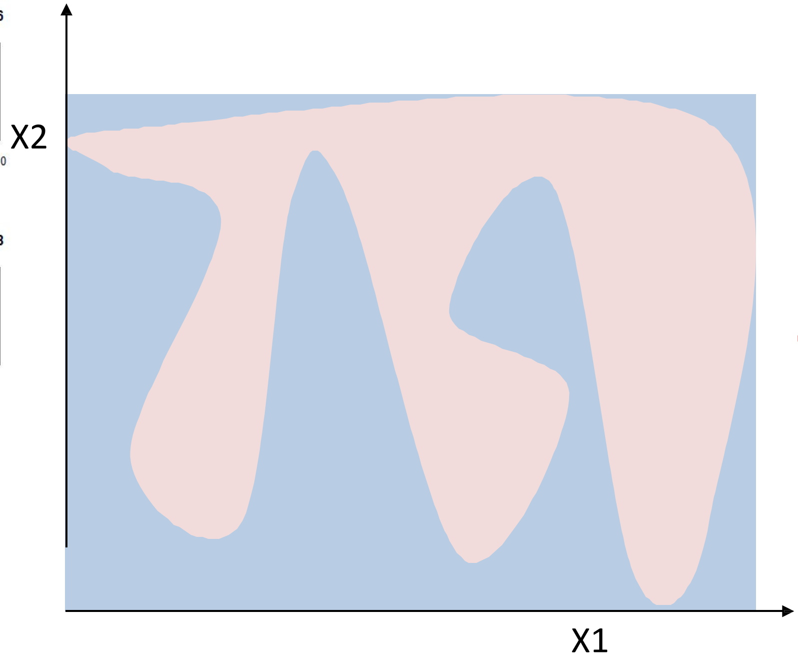


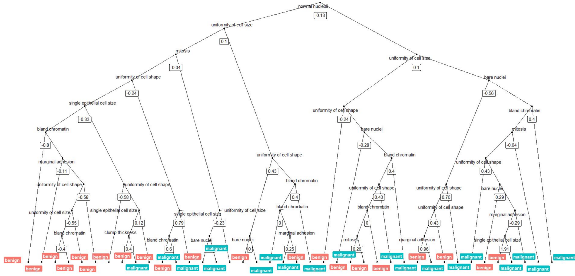


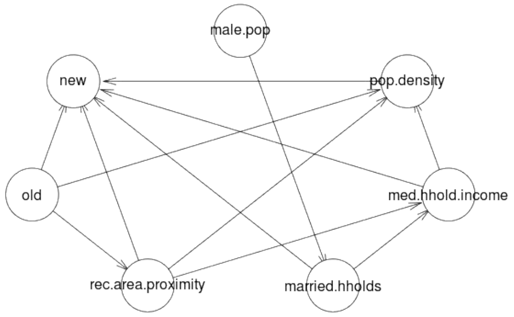


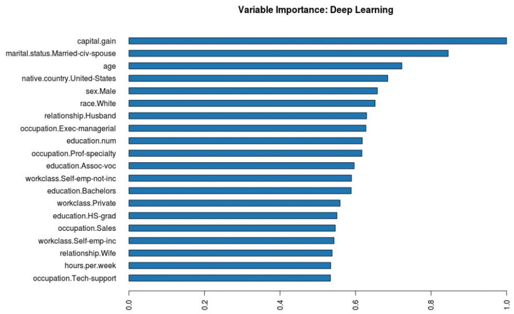
Big Data: More Complexity & More Dimensions



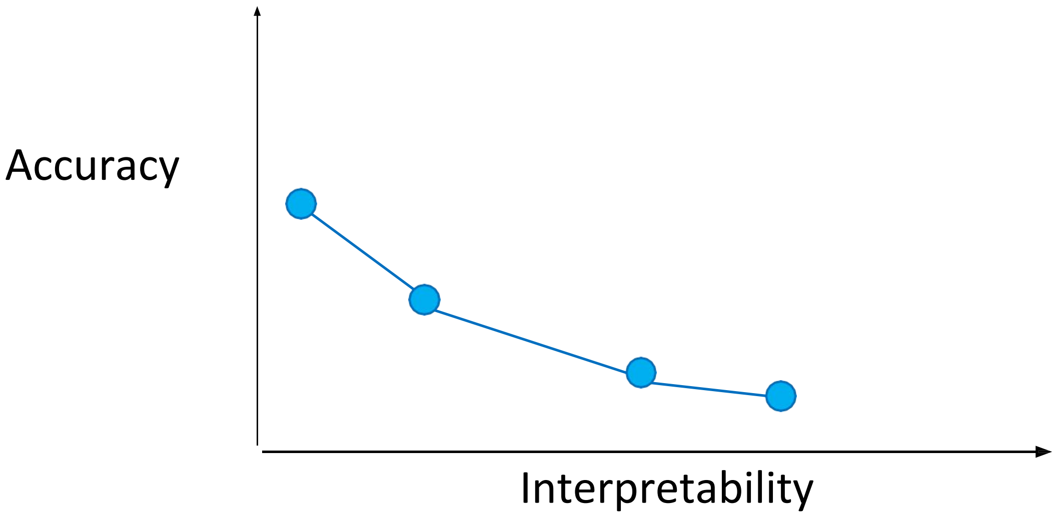






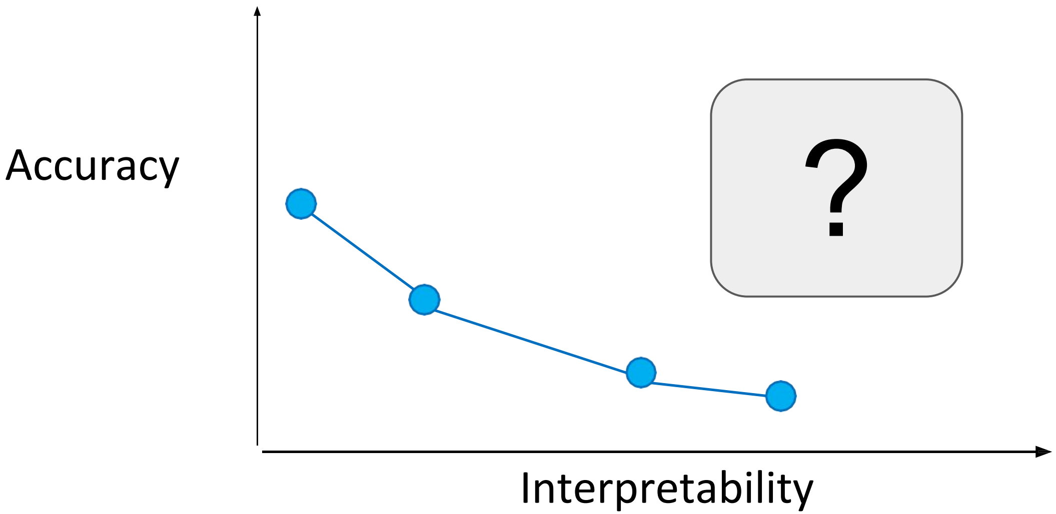


Accuracy VS Interpretability



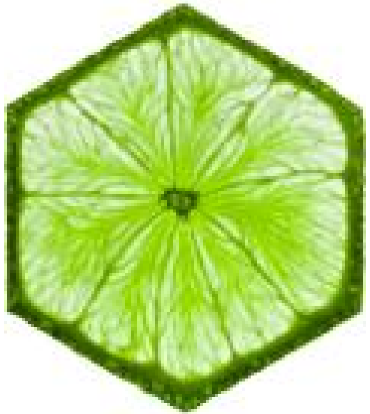
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Accuracy VS Interpretability

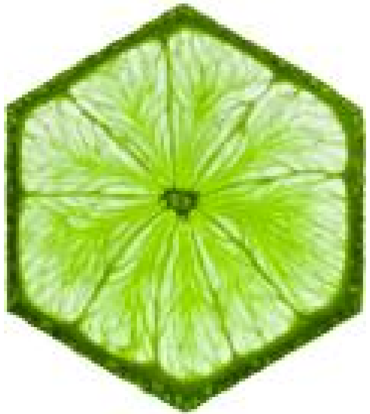


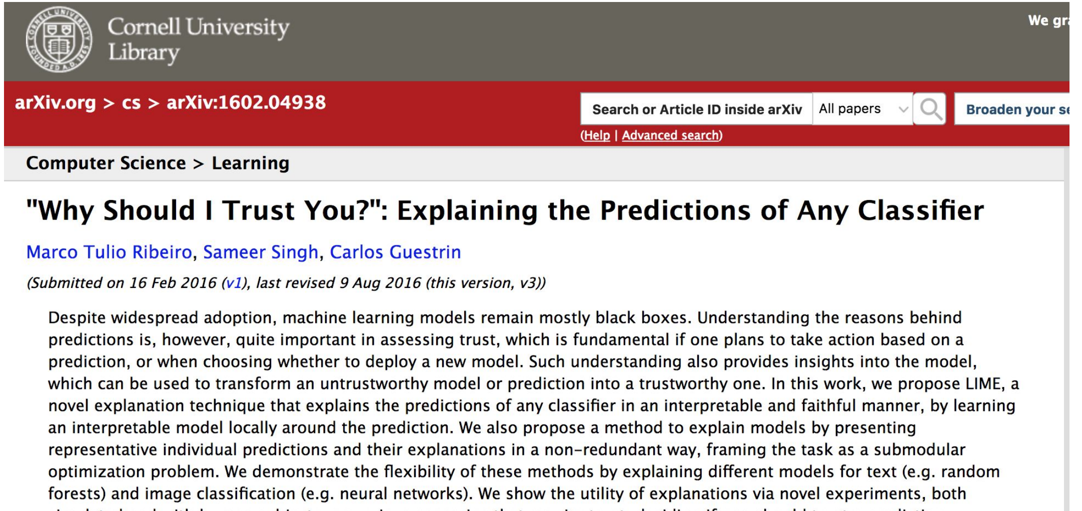
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Local
Interpretable
Model-agnostic
Explanations



Local
Interpretable
Model-agnostic
Explanations





Computer Science > Learning

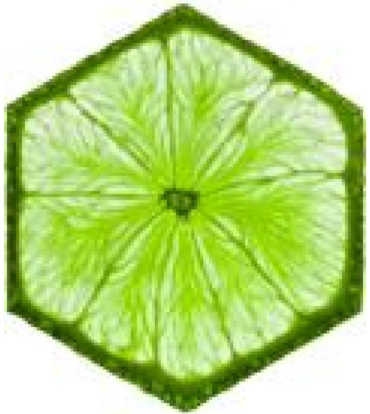
"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(Submitted on 16 Feb 2016 (v1), last revised 9 Aug 2016 (this version, v3))

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one. In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both

Local
Interpretable
Model-agnostic
Explanations



Cornell University Library

arXiv.org > cs > arXiv:1602.04938

Search or Article ID inside arXiv | All papers | Broaden your s

(Help | Advanced search)

Computer Science > Learning

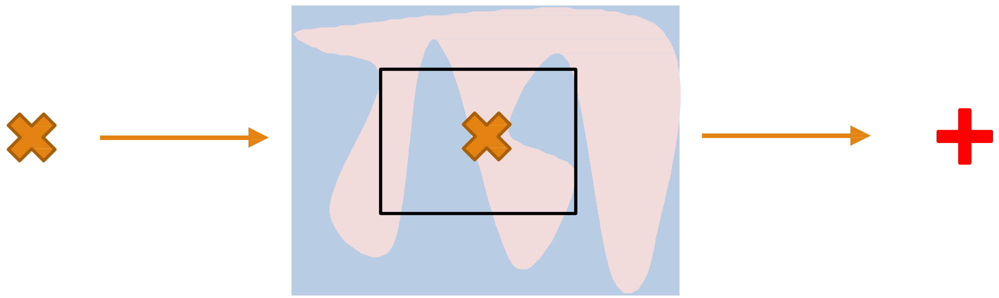
"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(Submitted on 16 Feb 2016 (v1), last revised 9 Aug 2016 (this version, v3))

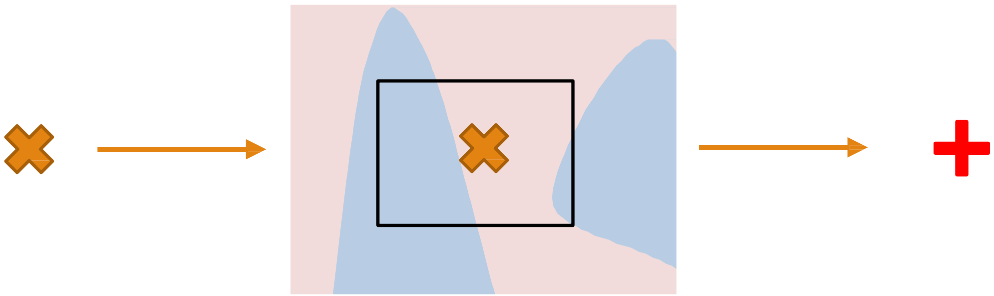
Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one. In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both

Being Local and Model-Agnostic…



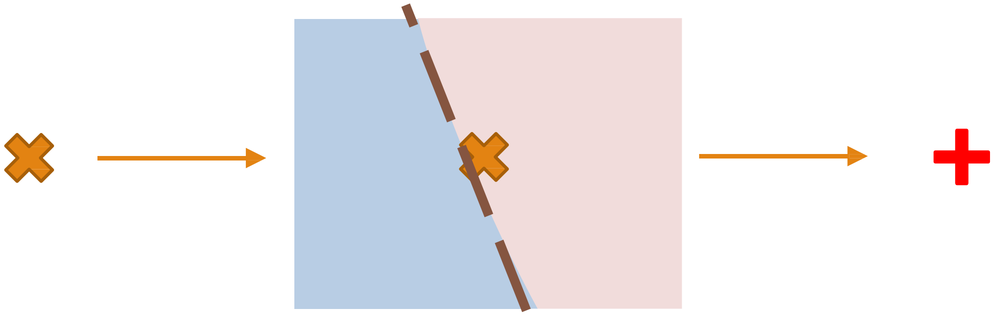
Source: https://www.youtube.com/watch?v=LAm4QmVaf0E&t=3658s

Being Local and Model-Agnostic…



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

Being Local and Model-Agnostic…



Explanation is an interpretable model, that is locally accurate

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

HOW LIME WORKS

1. Permute data*

2. Calculate distance between permutations and original observations*

3. Make predictions on new data using complex model

4. Pick m features best describing the complex model outcome from the permuted data.*

5. Fit a simple model to the permuted data with m features and similarity scores as weights *

6. Feature weights from the simple model make explanations for the complex models local behaviour

HOW LIME WORKS

1. Permute data*
2. Calculate distance between permutations and original observations*
3. Make predictions on new data using complex model
4. Pick m features best describing the complex model outcome from the permuted data.*
5. Fit a simple model to the permuted data with m features and similarity scores as weights *
6. Feature weights from the simple model make explanations for the complex models local behaviour

HOW LIME WORKS

1. Permute data*
2. Calculate distance between permutations and original observations*
3. Make predictions on new data using complex model
4. Pick m features best describing the complex model outcome from the permuted data.*
5. Fit a simple model to the permuted data with m features and similarity scores as weights *
6. Feature weights from the simple model make explanations for the complex models local behaviour

HOW LIME WORKS

1. Permute data*

2. Calculate distance between permutations and original observations*

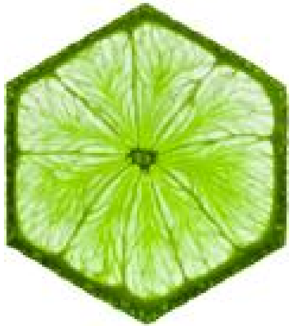
3. Make predictions on new data using complex model

4. Pick m features best describing the complex model outcome from the permuted data.*

5. Fit a simple model to the permuted data with m features and similarity scores as weights *

6. Feature weights from the simple model make explanations for the complex models local behaviour

HOW LIME WORKS



1. Permute data*

2. Calculate distance between permutations and original observations*

3. Make predictions on new data using complex model

4. Pick m features best describing the complex model outcome from the permuted data.*

5. Fit a simple model to the permuted data with m features and similarity scores as weights *

6. Feature weights from the simple model make explanations for the complex models local behaviour

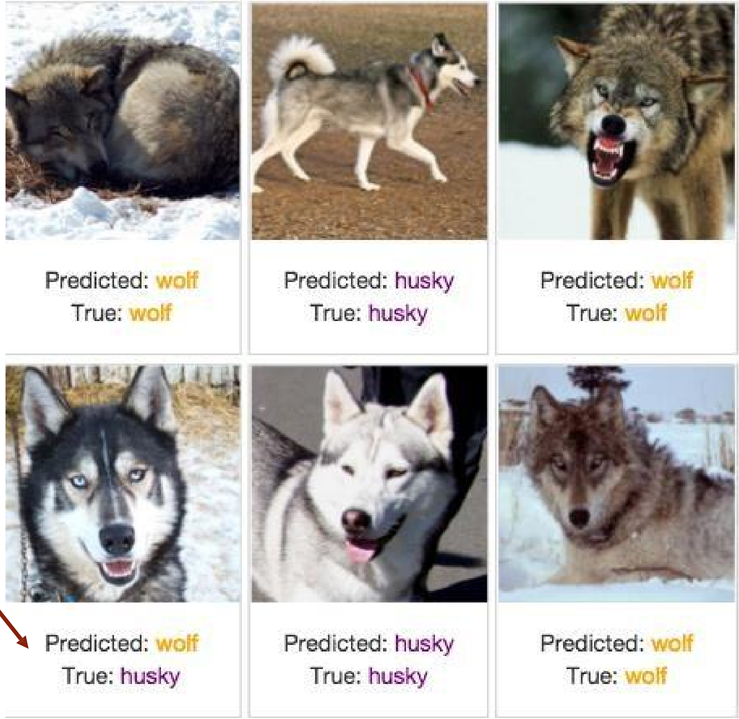
HOW LIME WORKS

1. Permute data*
2. Calculate distance between permutations and original observations*
3. Make predictions on new data using complex model
4. Pick m features best describing the complex model outcome from the permuted data.*
5. Fit a simple model to the permuted data with m features and similarity scores as weights *
6. Feature weights from the simple model make explanations for the complex models local behaviour

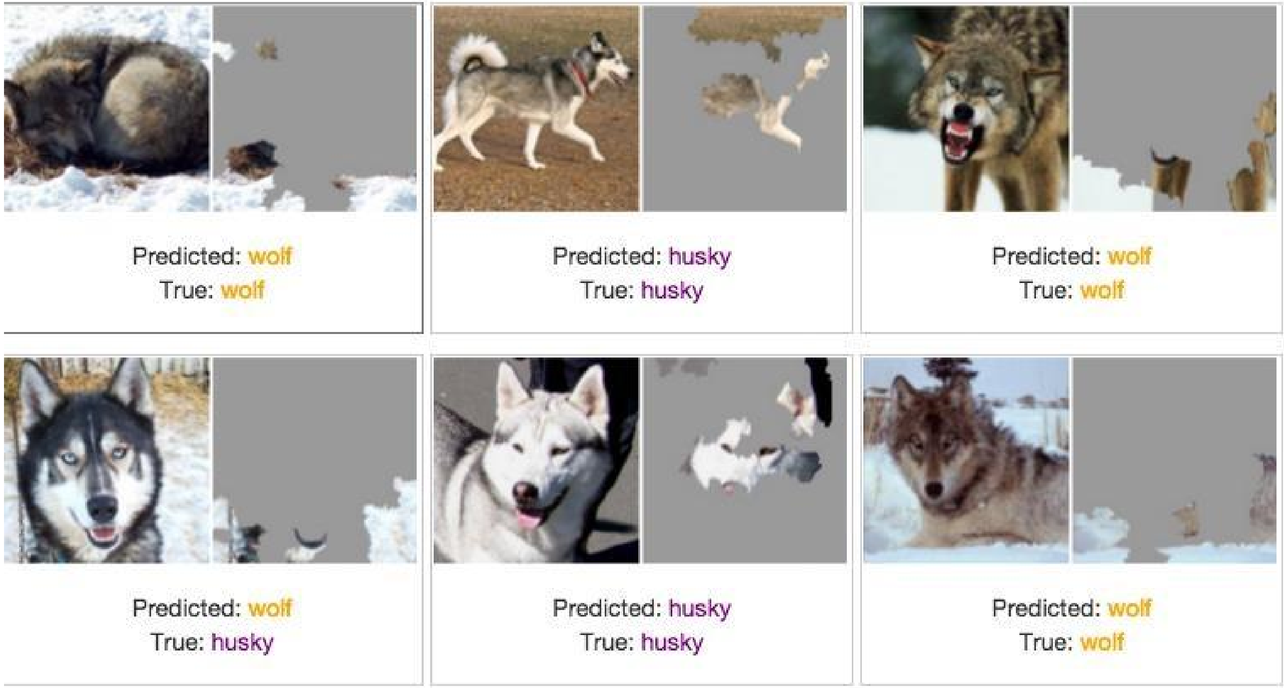
CAN YOU BUILD YOUR TRUST BASED ON ACCURACY?

CAN YOU BUILD YOUR TRUST BASED ON ACCURACY?

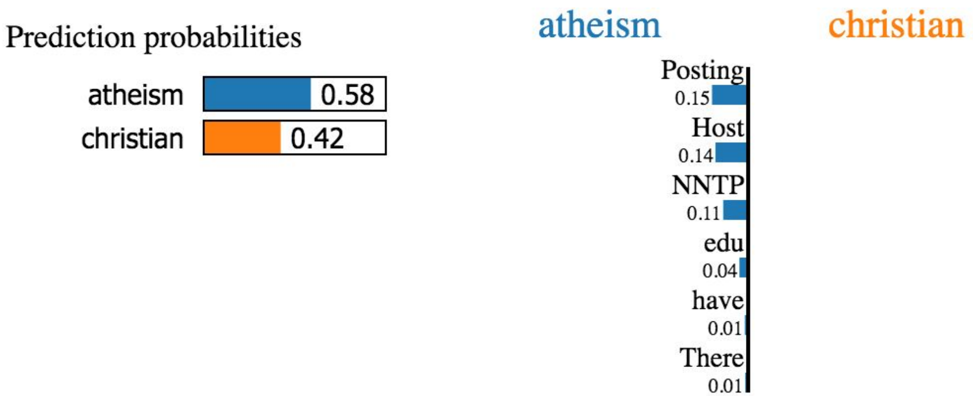
Only 1
mistake!



… YES, IF YOU WANT TO BUILD A GREAT SNOW DETECTOR!



LIME IN TEXT ANALYTICS



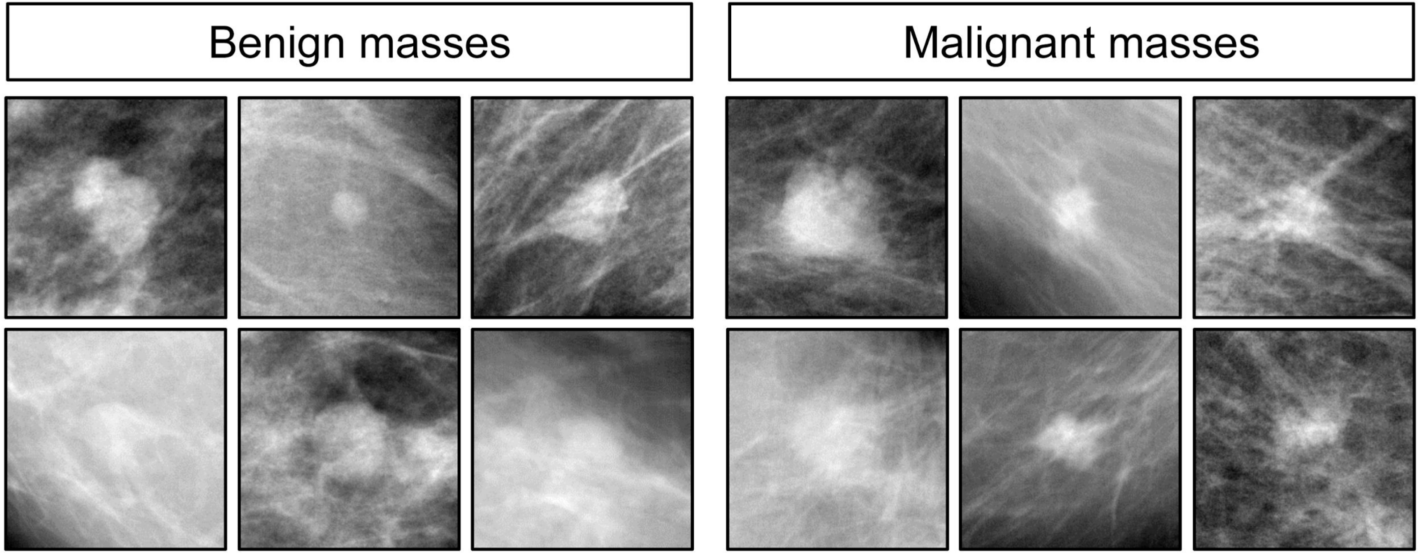
Text with highlighted words

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.

UNDERSTANDING CLASSIFICATION OF BENIGN AND MALIGNANT BREAST CANCER CELLS



LET'S SEE SOME CODE

WHY IS IT IMPORTANT?

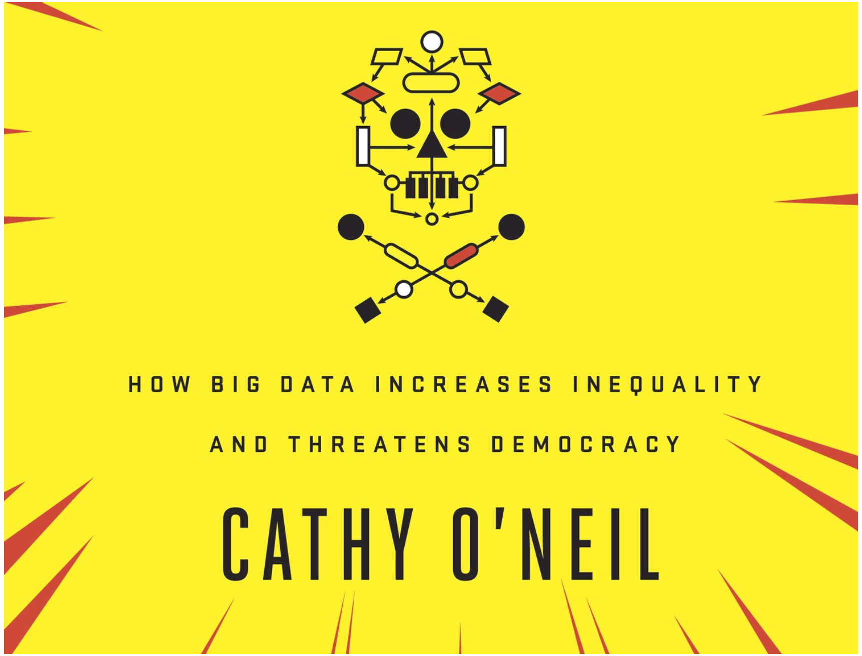
Trust	Predict	Improve
How can we trust the predictions are correct?	How can we understand and predict the behavior?	How do we improve it to prevent potential mistakes?

WHY IS IT IMPORTANT?

Trust	Predict	Improve
How can we trust the predictions are correct?	How can we understand and predict the behavior?	How do we improve it to prevent potential mistakes?
Being able to interpret the explanations and compare classifiers based on them	Improved prediction of model behavior and time to make that assessment when explanations were provided	Non-ML experts with explanations VS ML experts without explanations

KEEP
CALM
AND
COMPLY WITH
GDPR


KEEP
CALM
AND
COMPLY WITH
GDPR


HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL


KEEP
CALM
AND
COMPLY WITH
GDPR


HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL


TED Ideas worth spreading
WATCH
DISCOVER
Share
Add to list
Like
Zeynep Tufekci at TEDGlobal>NYC
We're building a dystopia just to make people click on ads
22:55