Putting Artificial Intelligence Back into People's Hands

Toward an Accessible, Transparent and Fair AI

16 November 2019 · SFScon, Bozen, Italy
Vincent Lequertier · FSFE volunteer · https://vl8r.eu

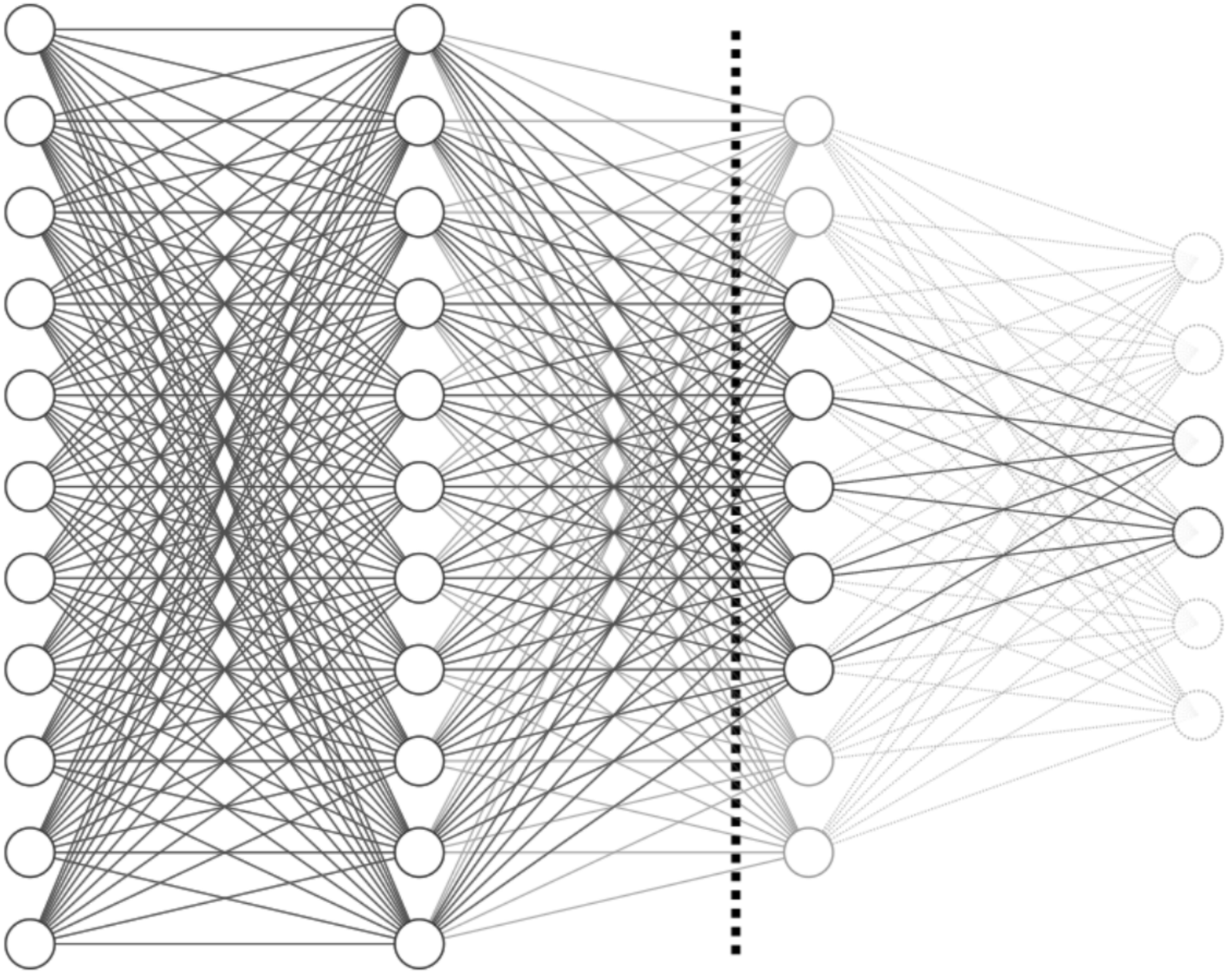
fsfe

Agenda

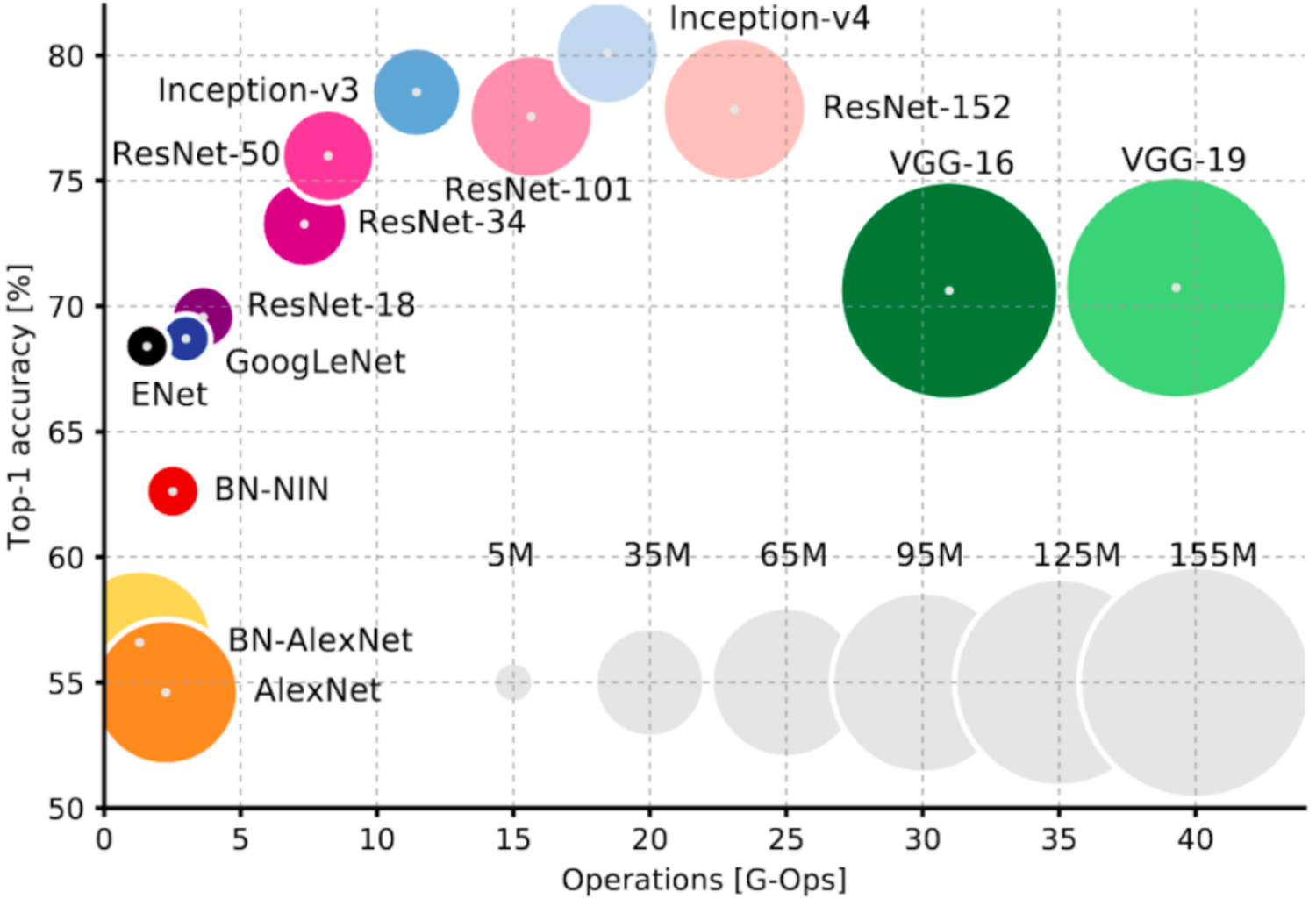
- How to create accessible Artificial Intelligence?
- Can AI be transparent and accurate?
- How to build fairness into AI?

Artificial Intelligence accessibility

Leveraging other models: fine-tuning



Bigger models are not more accurate



Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications

fsfe

How to make AI accessible?

- Make it easy to reuse the model's parameters
- Release the training code and datasets under a Free licence
- Consider computational complexity when designing the model

Artificial Intelligence transparency

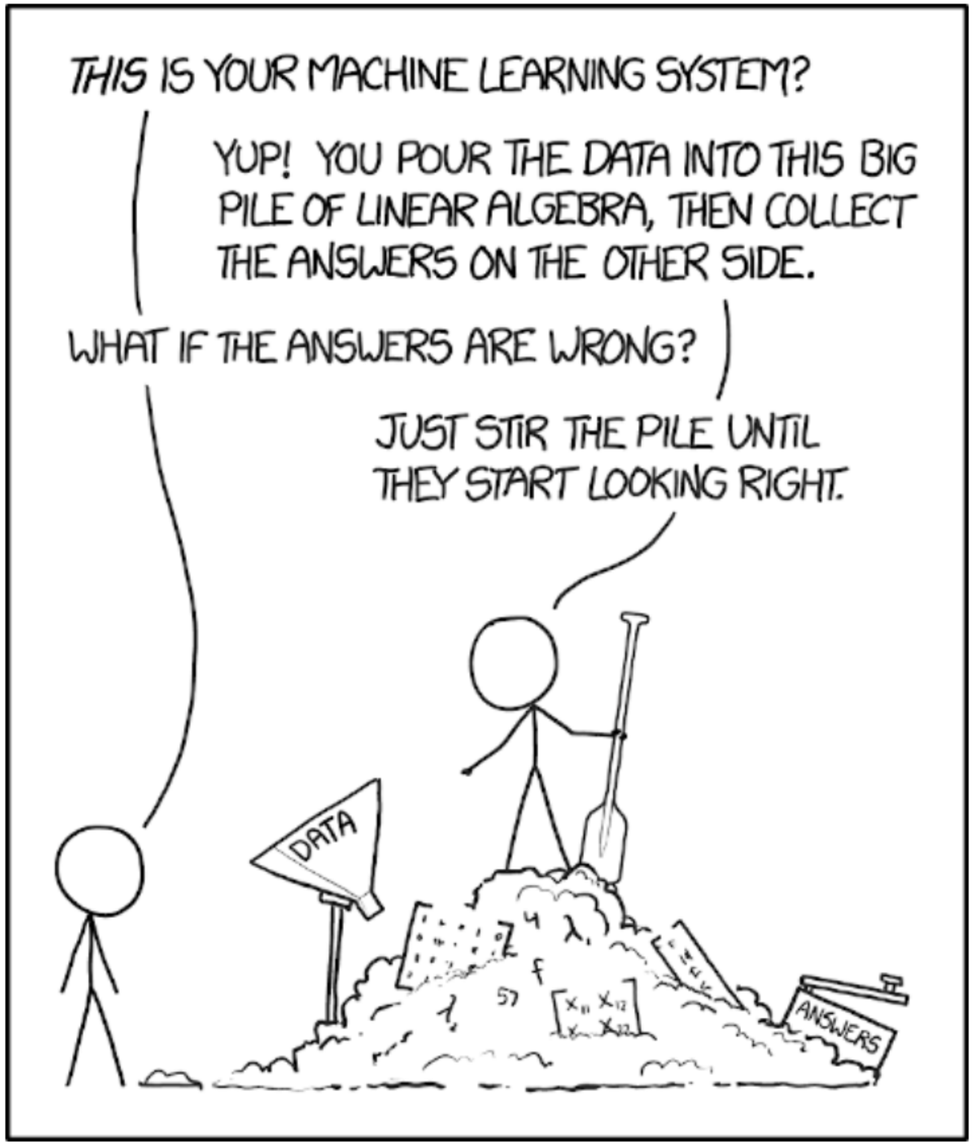
AI is used for critical matters

- Loan approval
- Justice
- Healthcare
- Self-driving cars

Why do we want AI transparency?

- Allows to interpret the result
- Builds trust in the model
- Helpful for debugging
- We require people to justify themselves

Parameters are not meant to be transparent



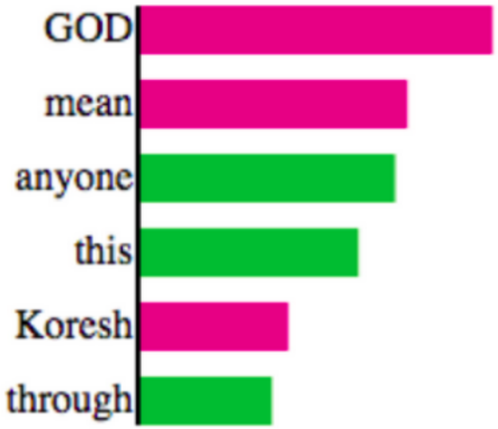
xkcd.com

LIME: Debugging and selecting models

Local Interpretable Model-Agnostic Explanations

Algorithm 1

Words that A1 considers important:



Predicted:

Atheism

Prediction correct:

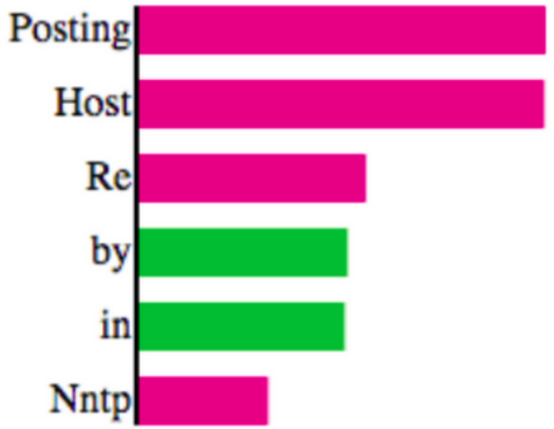
✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD**!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:



Predicted:

Atheism

Prediction correct:

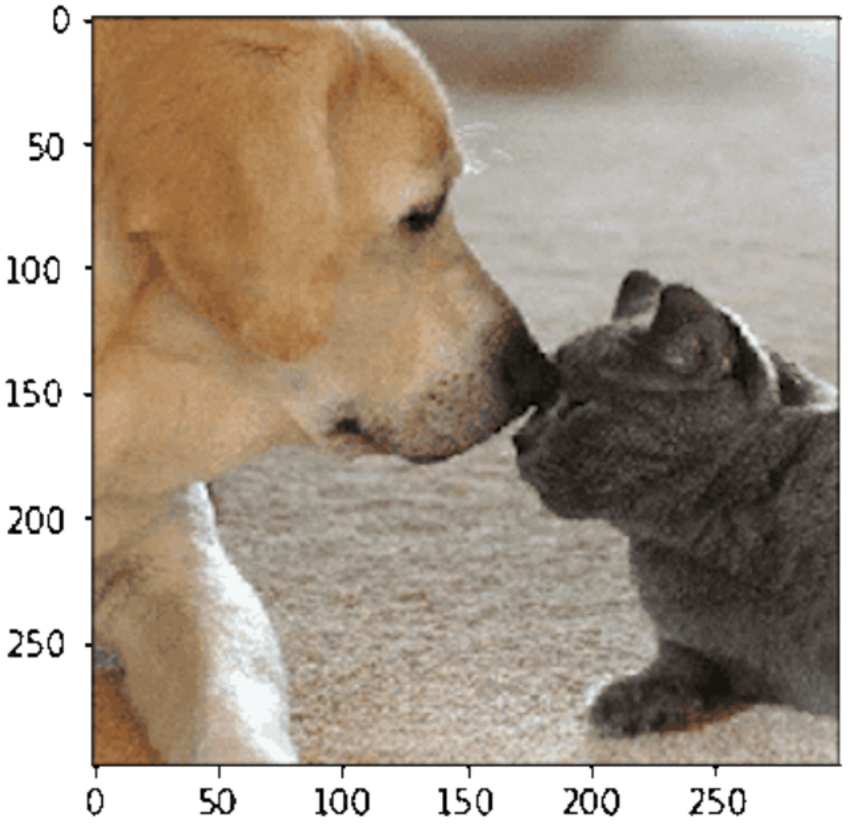
✓

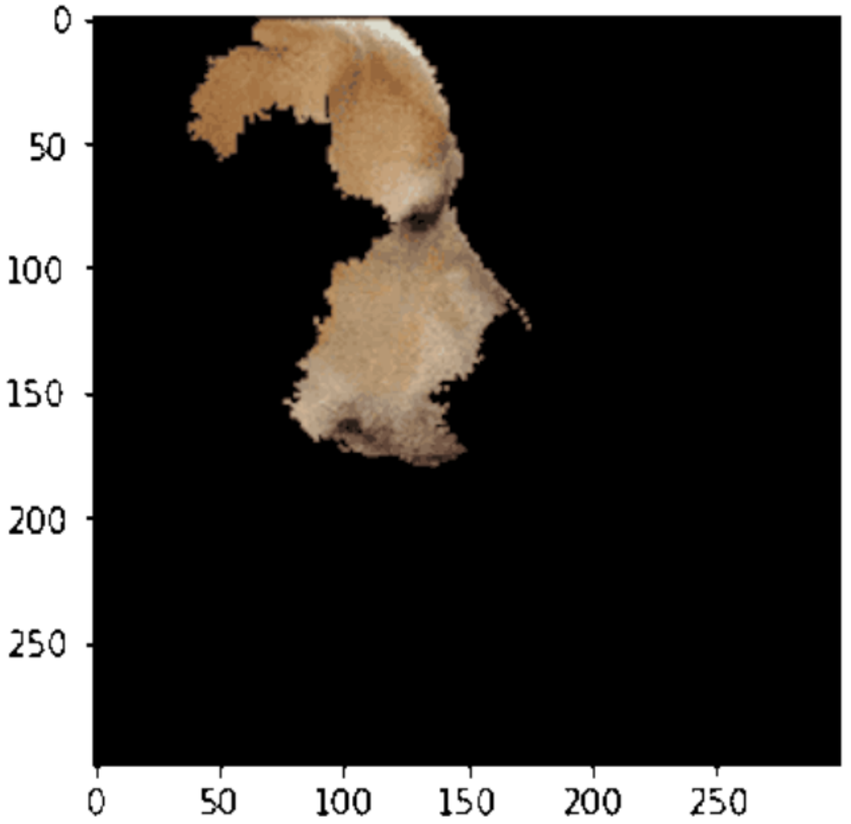
Document

From: pauld@verdix.com (Paul Durbin)
Subject: **Re**: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Tulio Ribeiro, M., Singh, S., & Guestrin, C. (2016). " Why Should I Trust You?": Explaining the Predictions of Any Classifier

Making sense of images classification

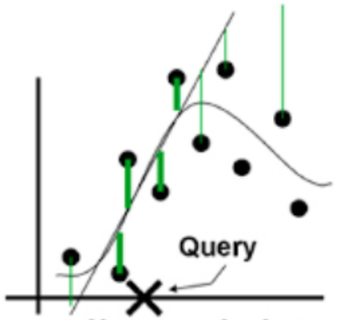




How it works?

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

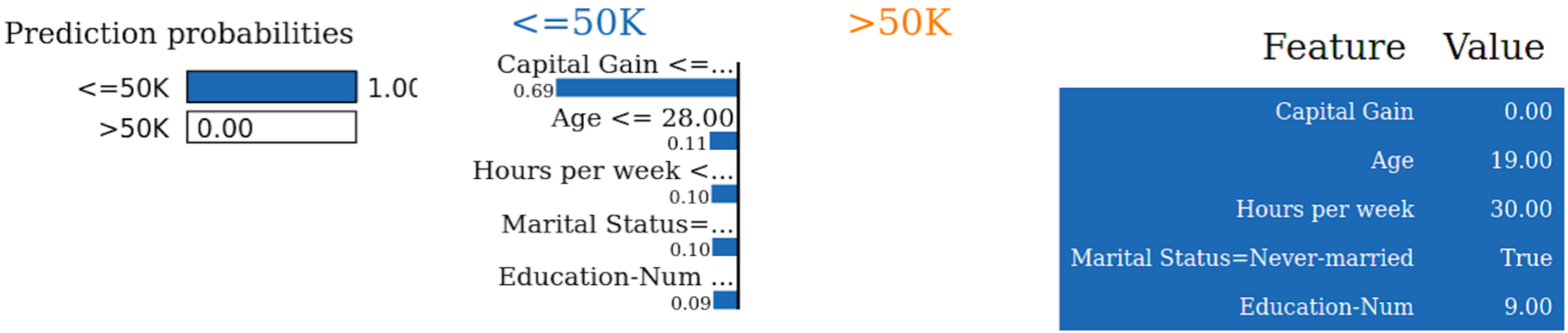


Locally weighted regression

Explanation

oreilly.com, Local Interpretable Model-Agnostic Explanations (LIME): An Introduction

Also for tabular data



Artificial Intelligence fairness

Protecting "car color" is easy

Brand	Seats	Year	Color	Speed (km/h)
A	5	2011	blue	150
B	2	2012	black	200
C	5	2010	red	250

Protecting gender is not easy

gender	hobby	education	salary
male	jogging	CS degree	35k
female	artistic swimming	self-taught	37k
female	women's volleyball team	PhD	35k
male	scuba-diving	CS degree	37k

Think about correlation before removing an attribute

Why an algorithm can be unfair?

- Bias in the data itself
- Trained with the wrong metrics (bias by proxy)
- Bad prediction model
- Bias is hard to notice
- *“With great power comes great responsibility”* (Peter Parker)

Vocabulary

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
- Positive Predicted Values (PPV)
- Negative Predicted Values (NPV)

COMPAS recidivism scoring

All Defendants

	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35

FN rate: 37.40

PPV: 0.61

NPV: 0.69

LR+: 1.94

LR-: 0.55

Black Defendants

	Low	High
Survived	990	805
Recidivated	532	1369

FP rate: 44.85

FN rate: 27.99

PPV: 0.63

NPV: 0.65

LR+: 1.61

LR-: 0.51

White Defendants

	Low	High
Survived	1139	349
Recidivated	461	505

FP rate: 23.45

FN rate: 47.72

PPV: 0.59

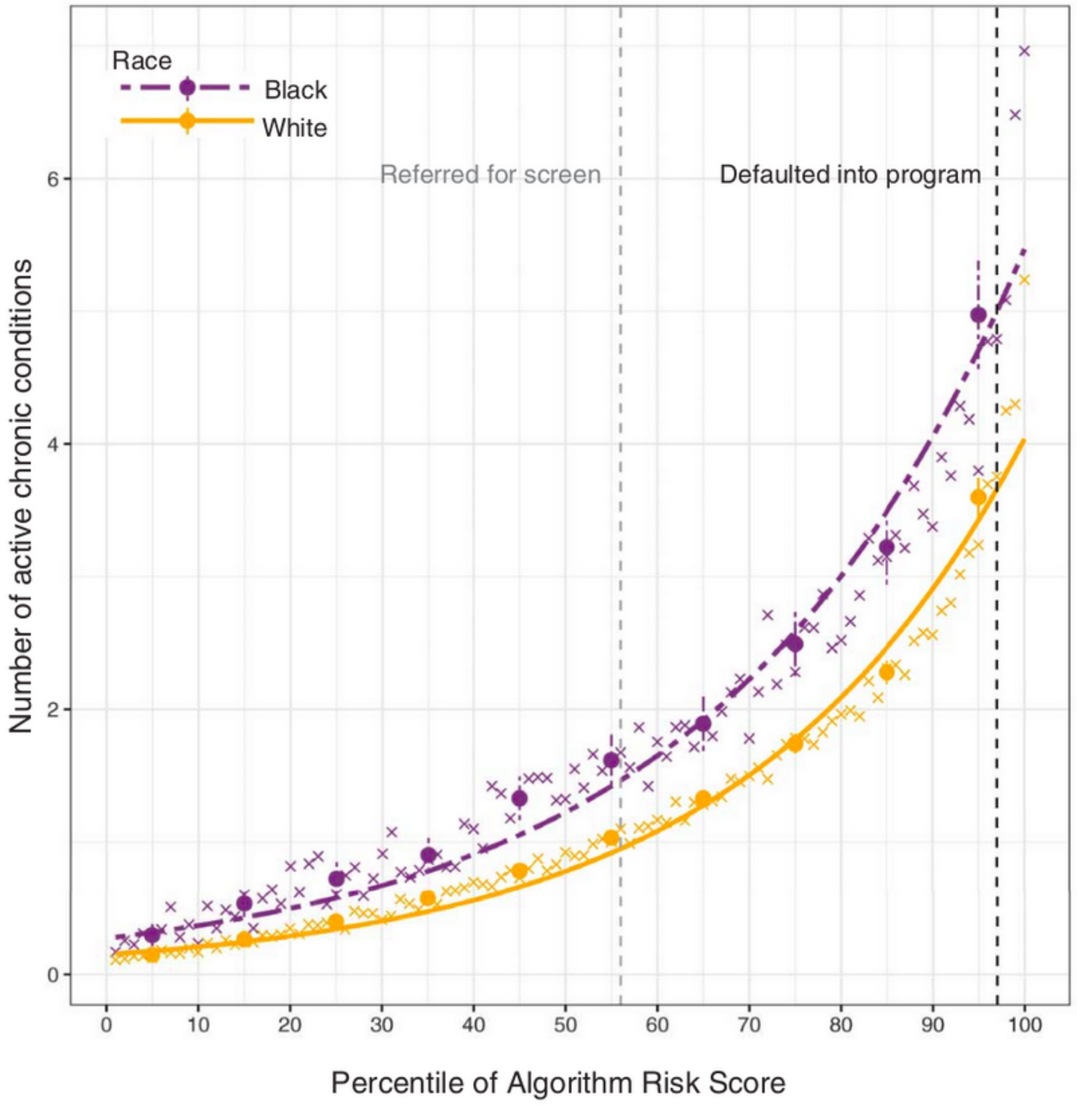
NPV: 0.71

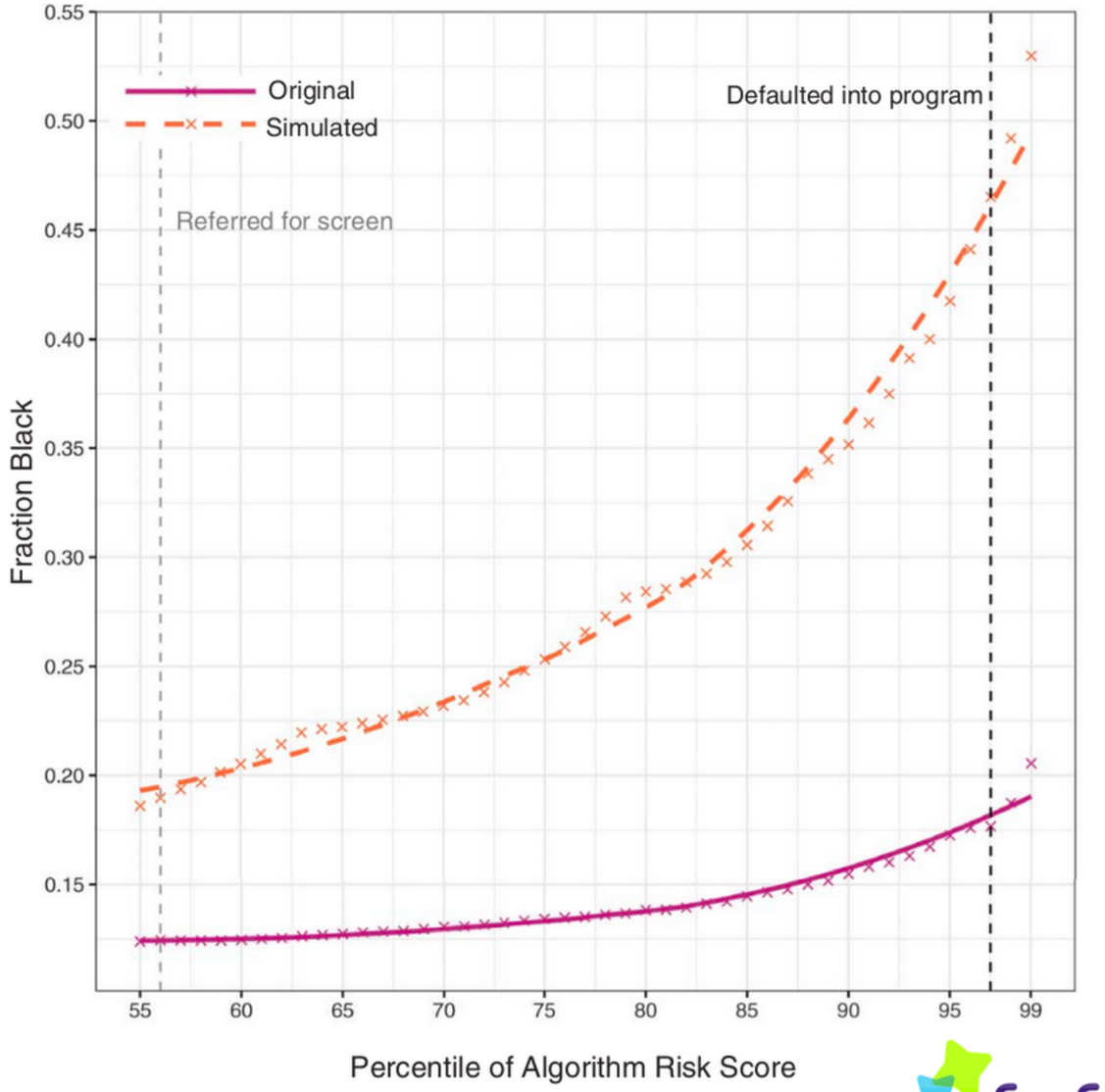
LR+: 2.23

LR-: 0.62

propublica.org, How We Analyzed the COMPAS Recidivism Algorithm (23 May 2016)

Racial bias in US healthcare





Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.

A plausible fair loss function

Let k be the number of values of a protected attribute
Let f: $y_{pred}, y_{true} \rightarrow s \in [0, 1]$ be a fairness function

$$loss = loss + \lambda \frac{\sum_{i=0}^{k} w_i f_i(y_{pred}, y_{true})}{\min_{\forall i \in [0,k[} f_i(y_{pred}, y_{true})}$$

Thank you! Questions?

Vincent Lequertier · FSFE volunteer · https://vl8r.eu