Putting Artificial Intelligence Back into People's Hands

Toward an Accessible, Transparent and Fair AI

2 February 2020 · FOSDEM, Brussels, Belgium

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

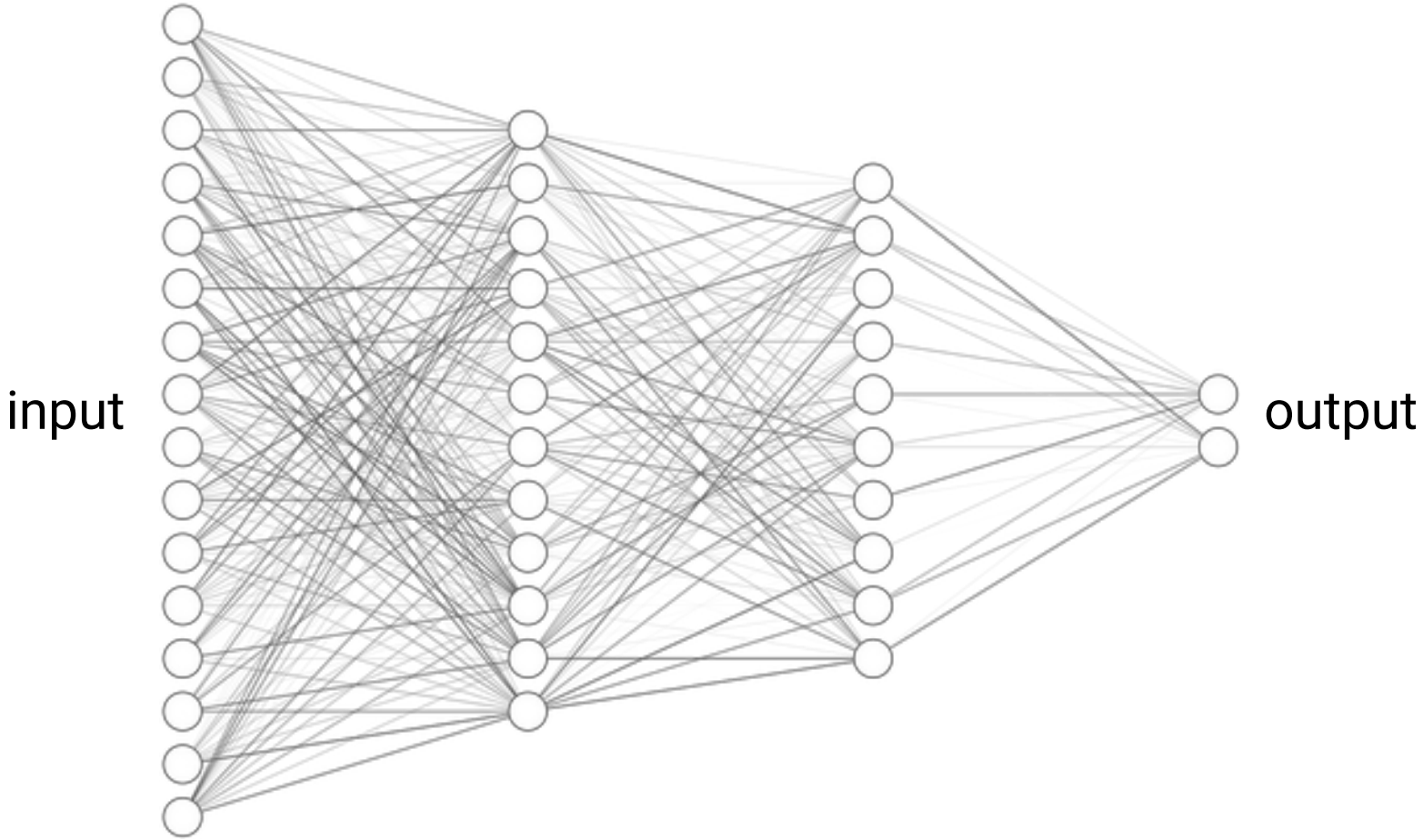
fsfe

Agenda

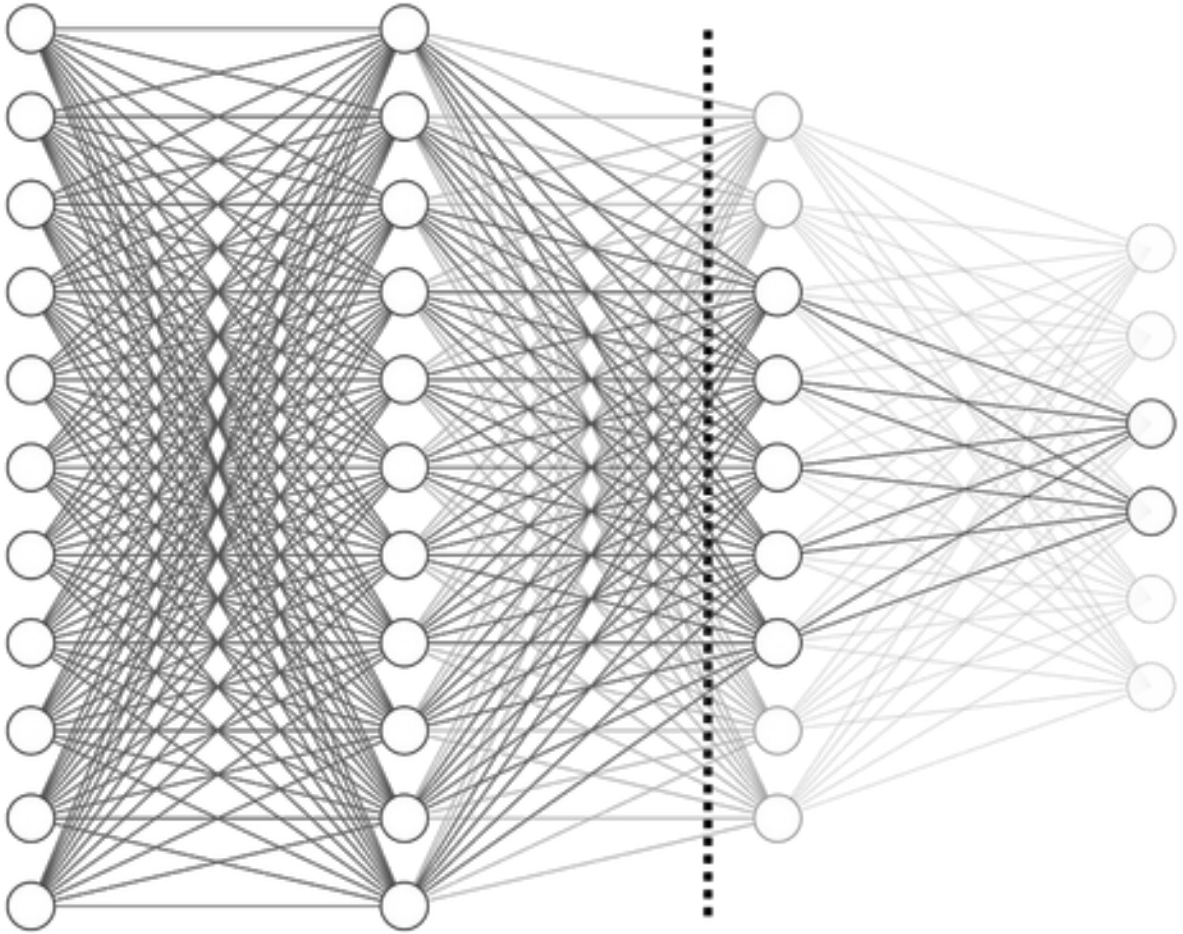
- How to create accessible Artificial Intelligence?
- Can AI be transparent and accurate?
- How to build fairness into AI?

Artificial Intelligence accessibility

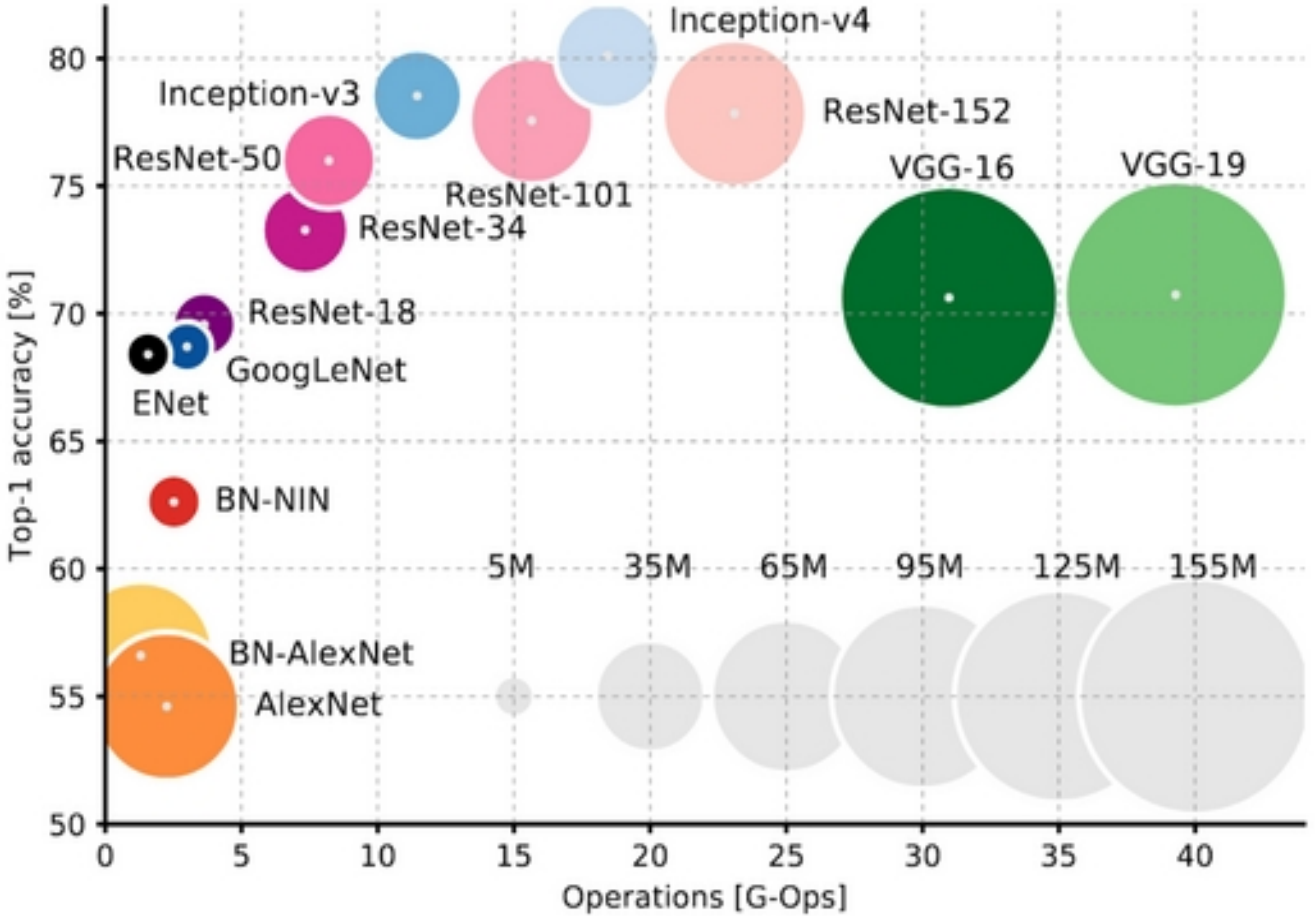
What is a neural network?



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

How to make AI accessible?

- Make it easy to reuse the model (ONNX format)
- Release the training code and the dataset under a Free licence
- Consider the number of FLOP when designing the model

Artificial Intelligence transparency

fsfe

AI is used for critical matters

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

Why do we want transparency?

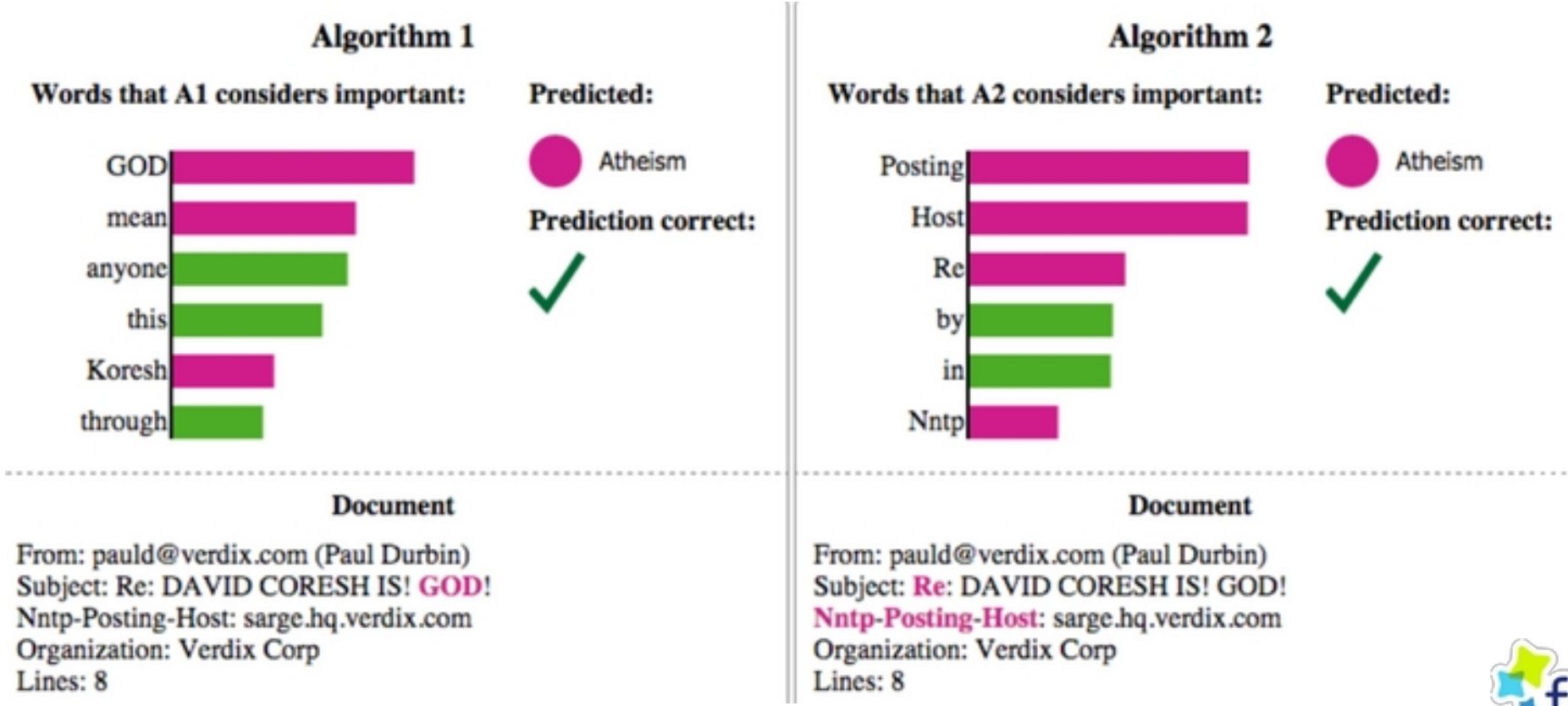
- Allows to interpret the result
- Builds trust in the model
- Makes debugging easier

Parameters are not meant to be transparent

xkcd.com

LIME: Debugging and selecting models

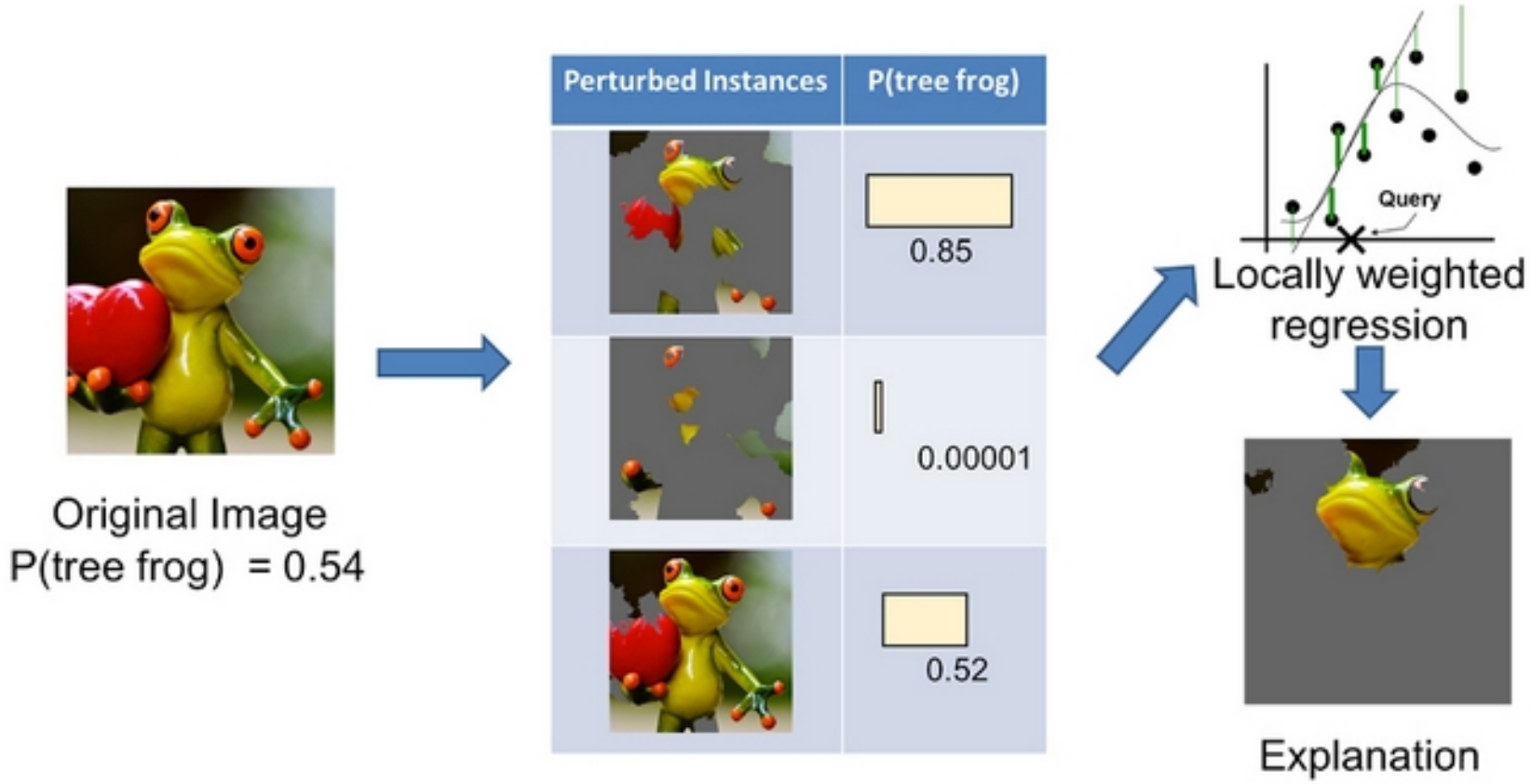
Local Interpretable Model-agnostic Explanations



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

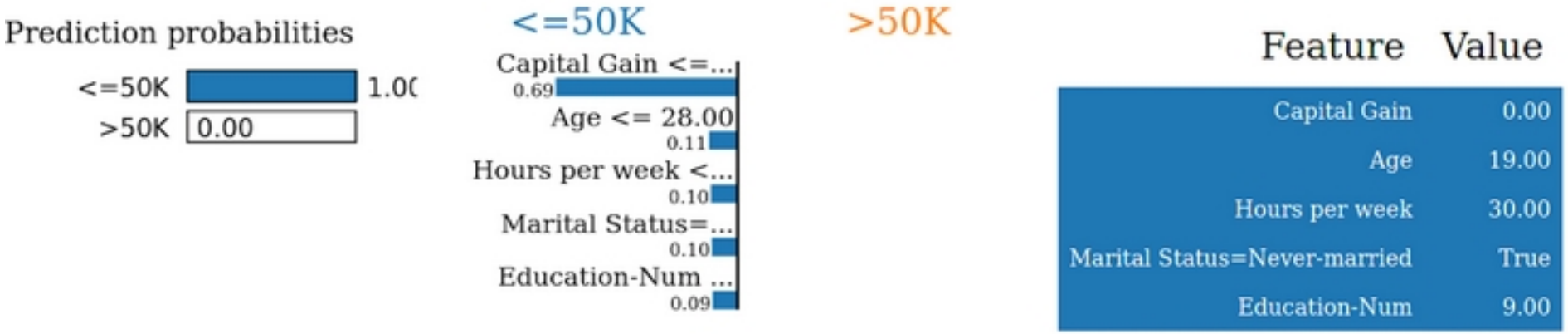
Making sense of images classification

How does it work?



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

Also for tabular data



Artificial Intelligence fairness

Protecting car colors 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
female	women’s volleyball team	CS degree	35k
male	football team captain	self-taught	37k
male	chess	CS degree	37k

❗ **Think about correlation before removing an attribute**

Vocabulary

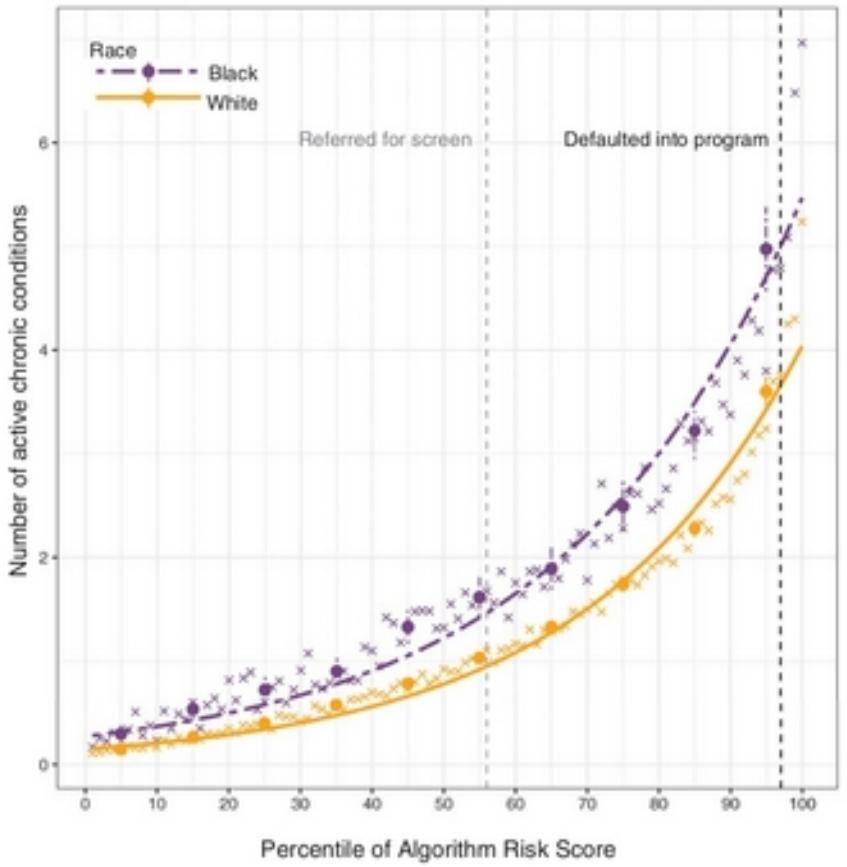
- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

COMPAS recidivism scoring

All defendants			Black defendants			White defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate	32.35		**FP rate**	44.85		**FP rate**	23.45	
FN rate	37.40		**FN rate**	27.99		**FN rate**	47.72	

propublica.org (2016)

Racial bias in healthcare



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

Why an algorithm can be unfair?

- Bias in the input data itself
- Training with the wrong metric (bias by proxy)
- Bad prediction model
- Bias is hard to notice
- "*With great power comes great responsibility*" (Peter Parker)

A 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

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

Thank you! Questions?

2 February 2020 · FOSDEM, Brussels, Belgium

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