

# Explaining Explainable AI

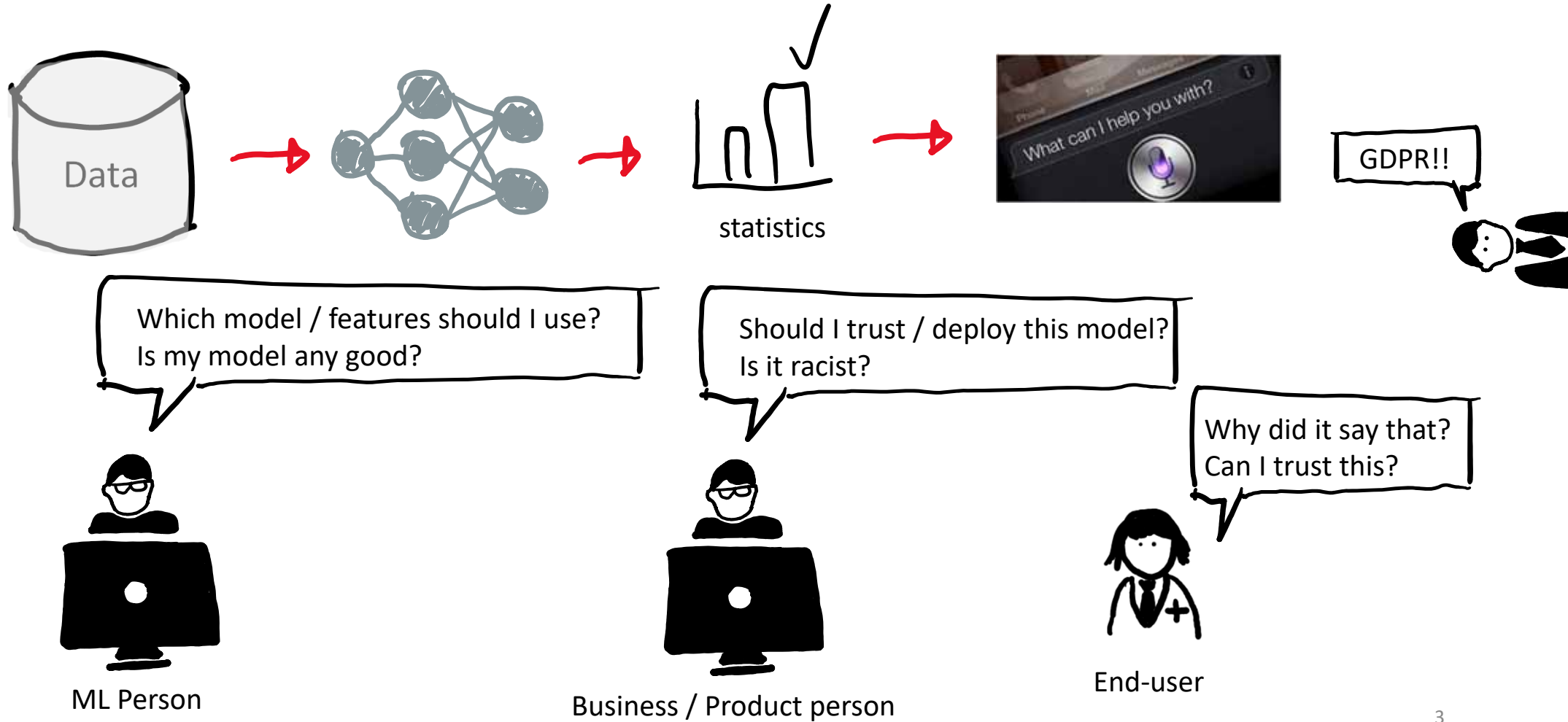
Marco Tulio Ribeiro (Microsoft Research)

# ML according to ML101



What's missing?

# ML in real life: humans matter



# explain

[ik'splān] 

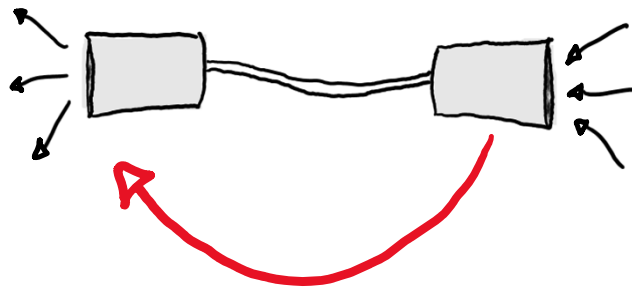
VERB

make (an idea, situation, or problem) clear to someone by describing it in more detail or revealing relevant facts or ideas.

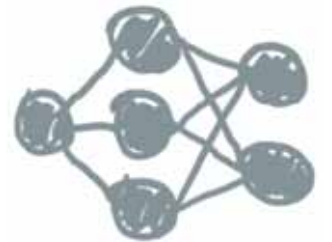
Purpose



Explanation  
Type



Technique



# Cheat sheet

① Purpose

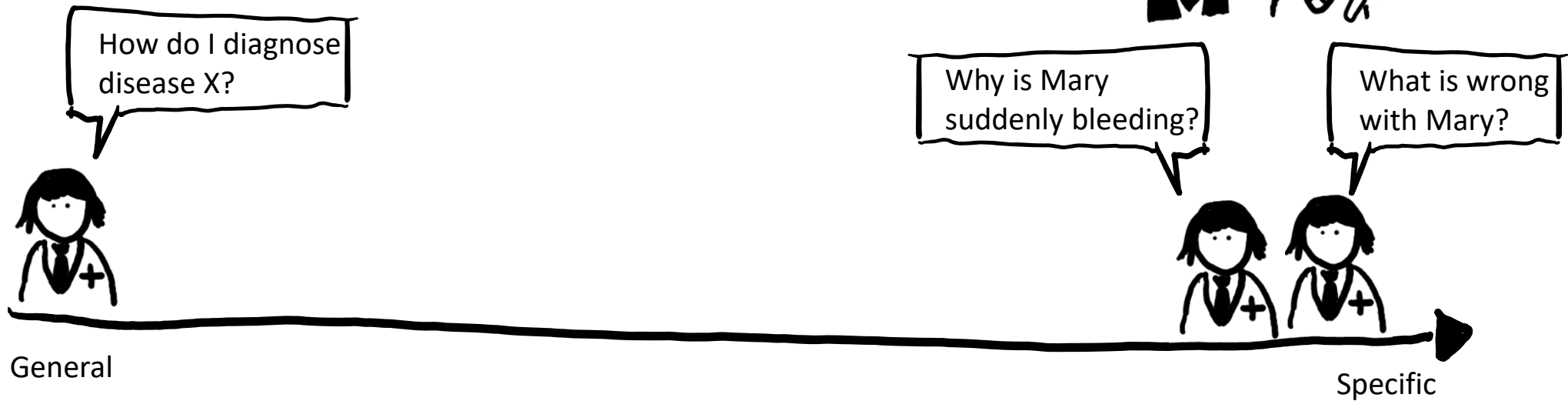
② Explanation Type

③ Technique

- What are explanations for?

# Purpose of explanation

## 1. Augment humans

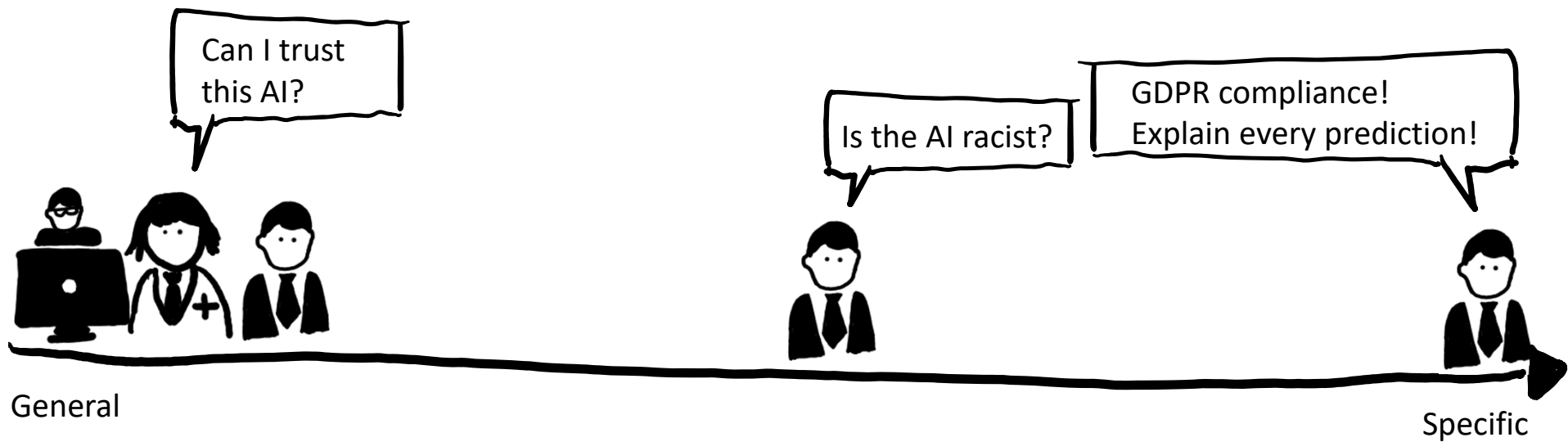


# Purpose of explanation



# Purpose of explanation

## 2. Help humans evaluate the AI





# Purpose of explanation



# Purpose of explanation

## 3. Help humans improve the AI

Ad related to latanya sweeney ⓘ

[Latanya Sweeney Truth](#)

[www.instantcheckmate.com/](http://www.instantcheckmate.com/)

Looking for **Latanya Sweeney**? Check **Latanya Sweeney's Arrests**.

Ads by Google

[Latanya Sweeney, Arrested?](#)

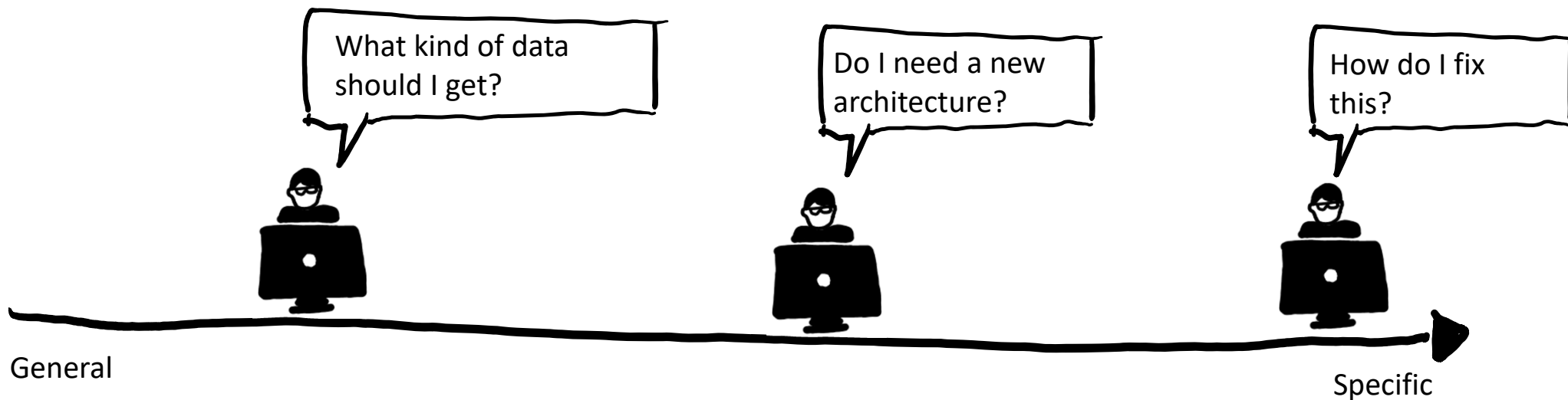
1) Enter Name and State. 2) Access Full Background Checks Instantly.

[www.instantcheckmate.com/](http://www.instantcheckmate.com/)

[Latanya Sweeney](#)

Public Records Found For: **Latanya Sweeney**. View Now.

[www.publicrecords.com/](http://www.publicrecords.com/)



# Cheat sheet

## ① Purpose

- What are explanations for?

- Augment humans
- Evaluate the AI
- Improve the AI

- What questions do you have?



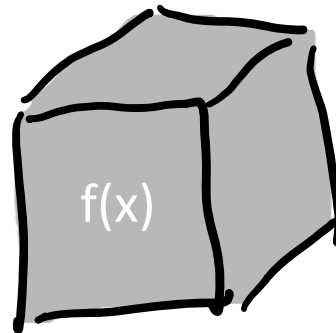
## ② Explanation Type

## ③ Technique

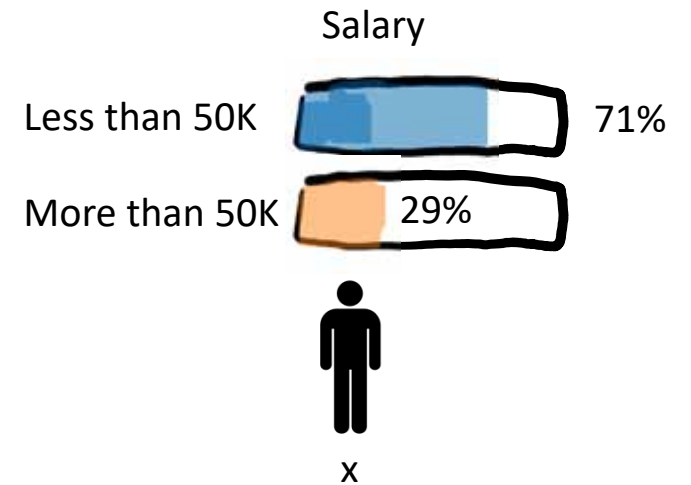
# Example: predicting income

Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	$\leq$ High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	$\leq 40$
Country	United States

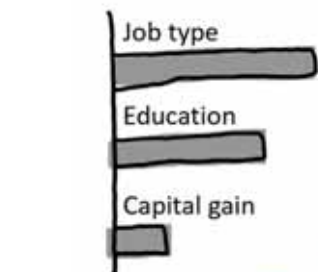
Census Data



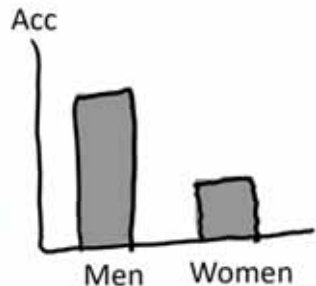
AI



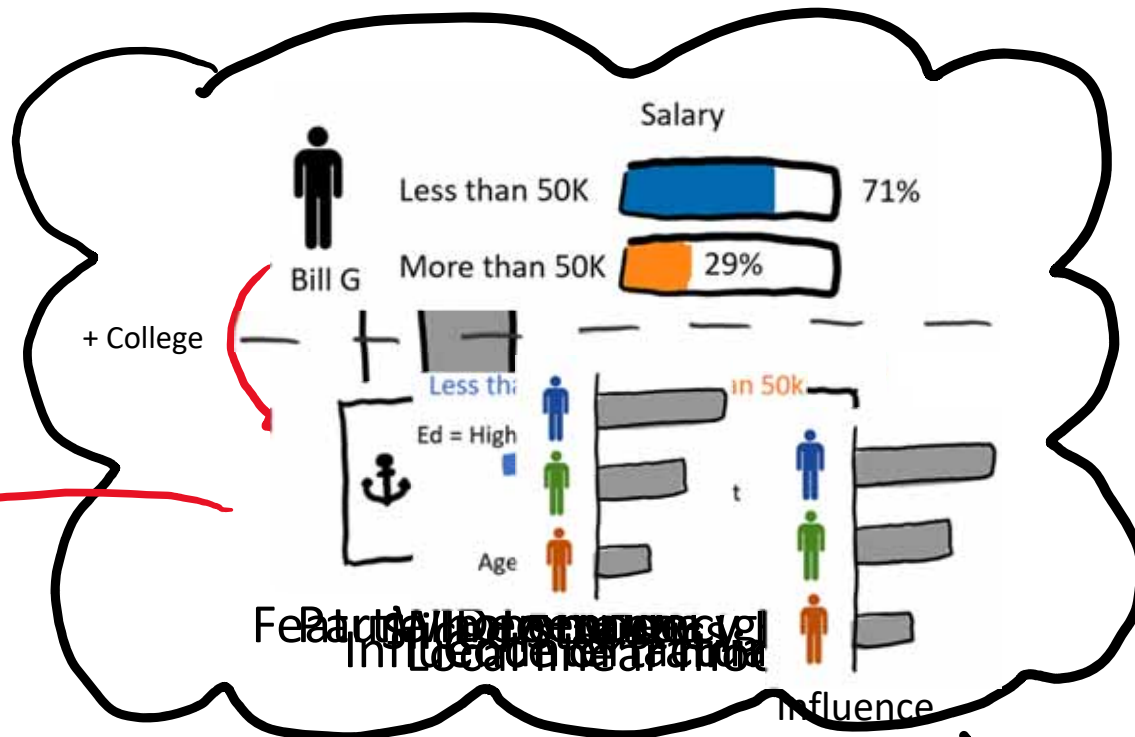
# Explanation(s)



Feature importance



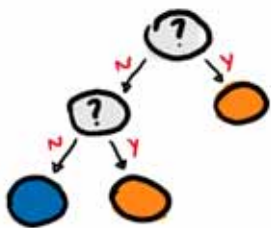
Sliced Statistics



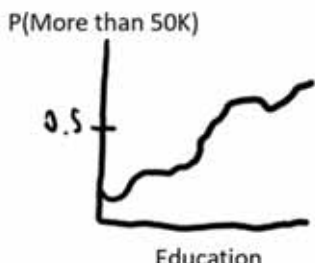
Feature importance  
Influence

Influence

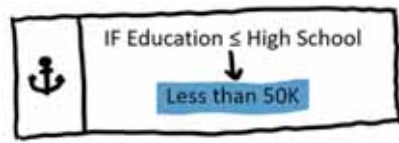
General ▶ Specific



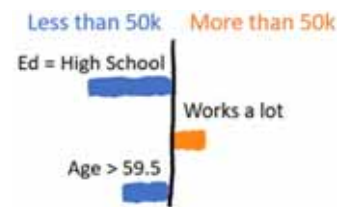
Whole model



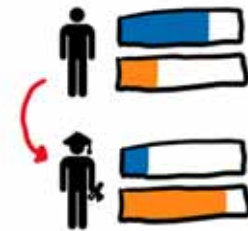
Partial Dependency Plot



Anchor



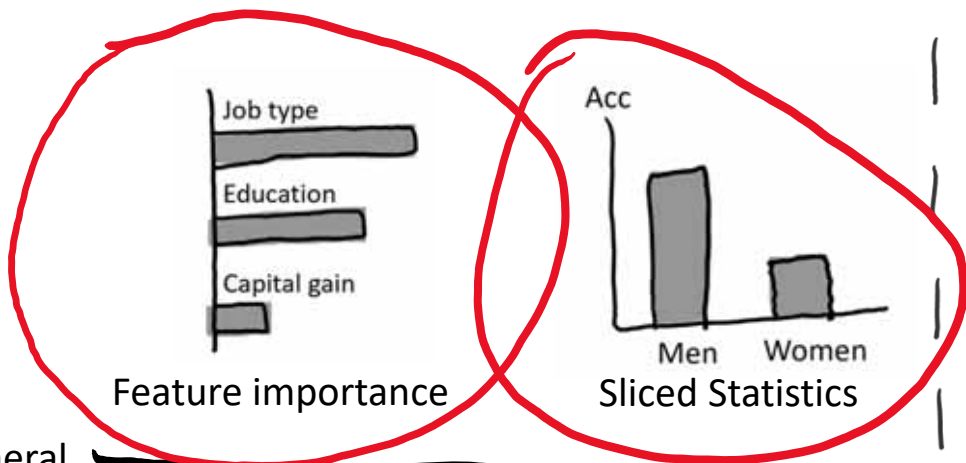
Linear model



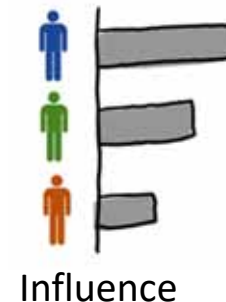
Counterfactual

# Explanation(s)

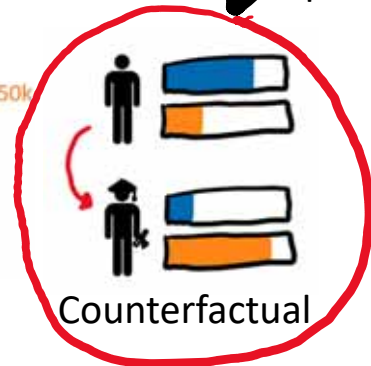
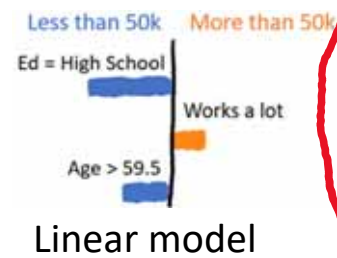
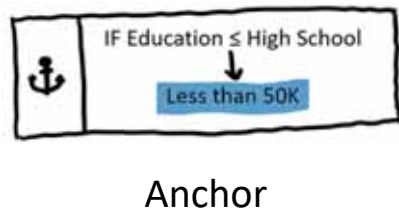
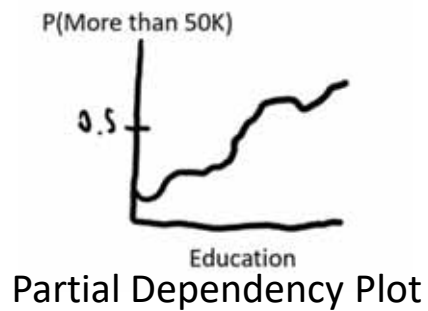
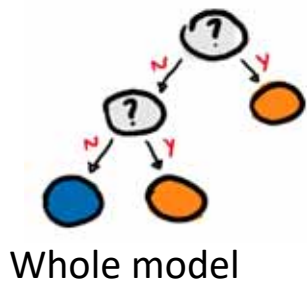
Which one is the best?



Why do I make more money?  
What are important predictors of income?  
How do I make my model fair?



General ▶ Specific



# Cheat sheet

## ① Purpose

### • What are explanations for?

- Augment humans
- Evaluate the AI
- Improve the AI

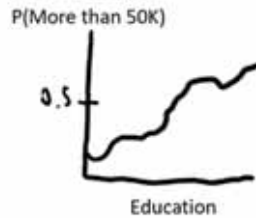
### • What questions do you have?



## ② Explanation Type



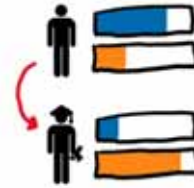
Linear model



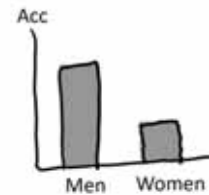
Partial Dependency Plot



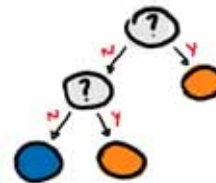
Anchor



Counterfactual



Sliced Statistics



Whole model


⋮

## ③ Technique

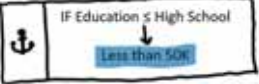
# Technique: how to get explanations

Can I please have Interpretable ML?

I'm comfortable with




Changed my mind, I want



Interpretable model


Ok, let's train a different model  
Sure, let's train a linear model!



- Exact
- Fast

Black box explanations

Sure, bring your model and we'll explain it



- More accurate
- Flexible



# Technique: how to get explanations

Can I trust the AI to diagnose patients alone?



Interpretable model

Sure! Give me a few more years : )



- Exact
- Fast

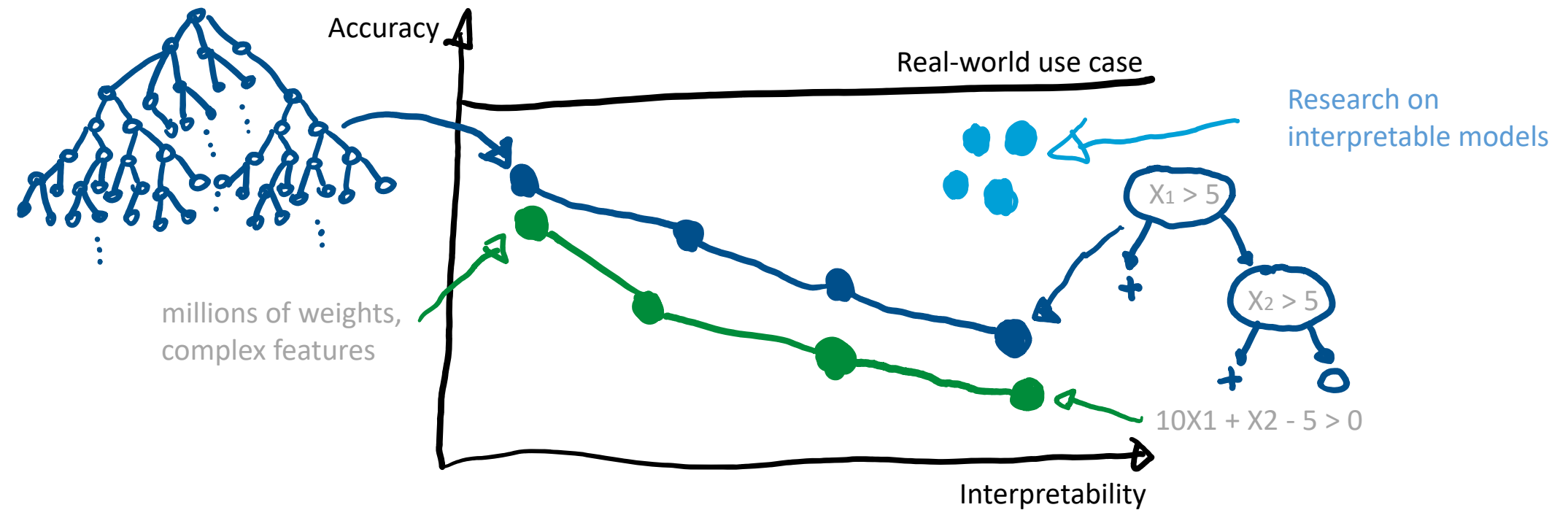
Black box explanations

Let me give you Rich's contact info...

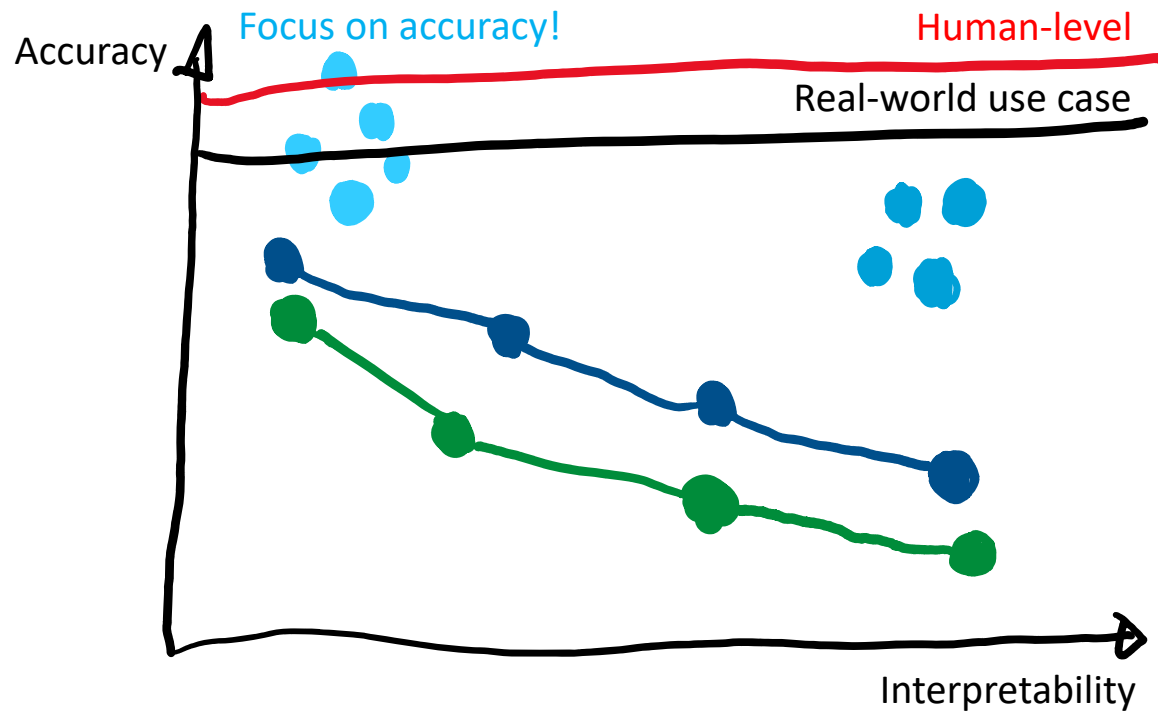


- More accurate
- Flexible

# Accuracy vs Interpretability

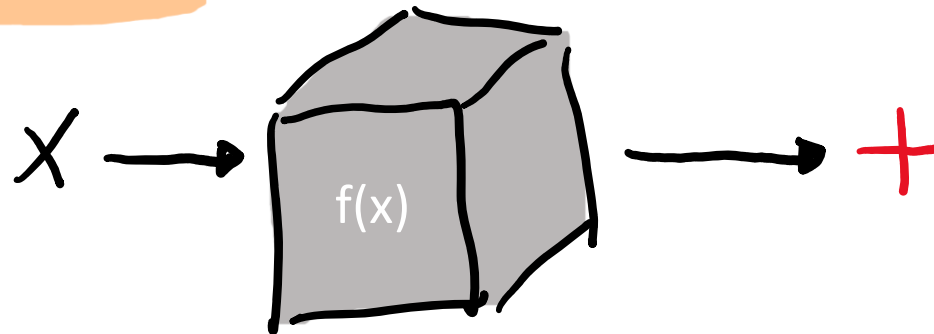


# Deep Learning / Ensembles



# Being Model-Agnostic...

Ignore the internal structure

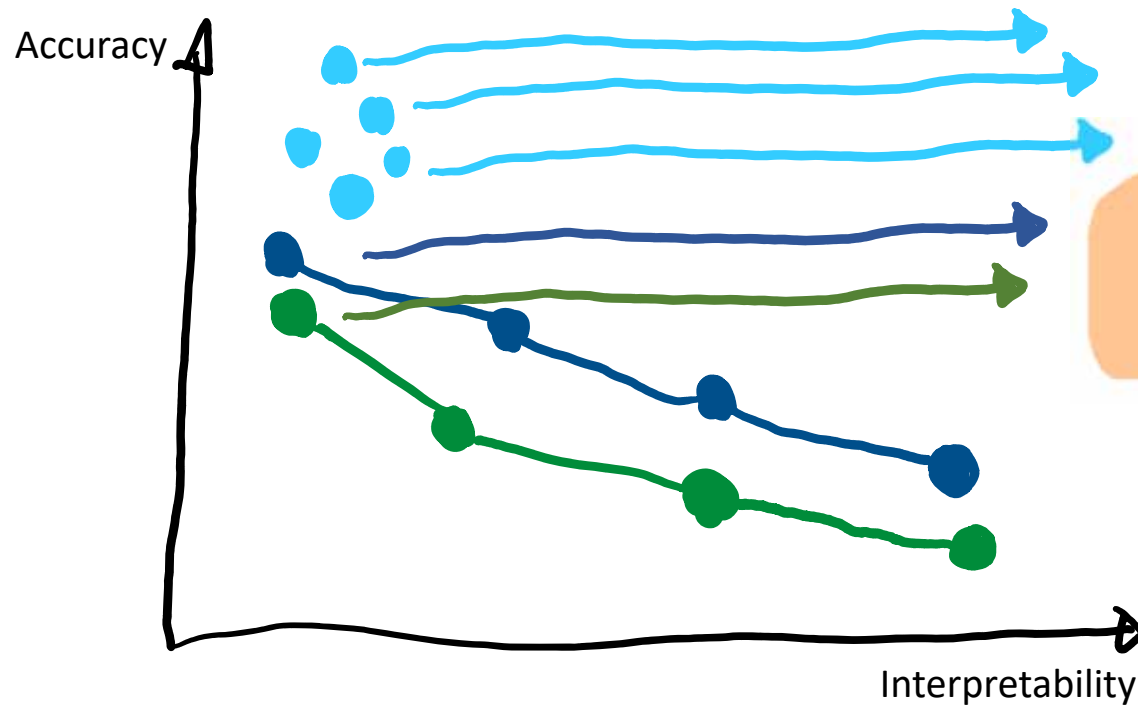


Explain any existing, *or future* model

Compare any 2 models to each other

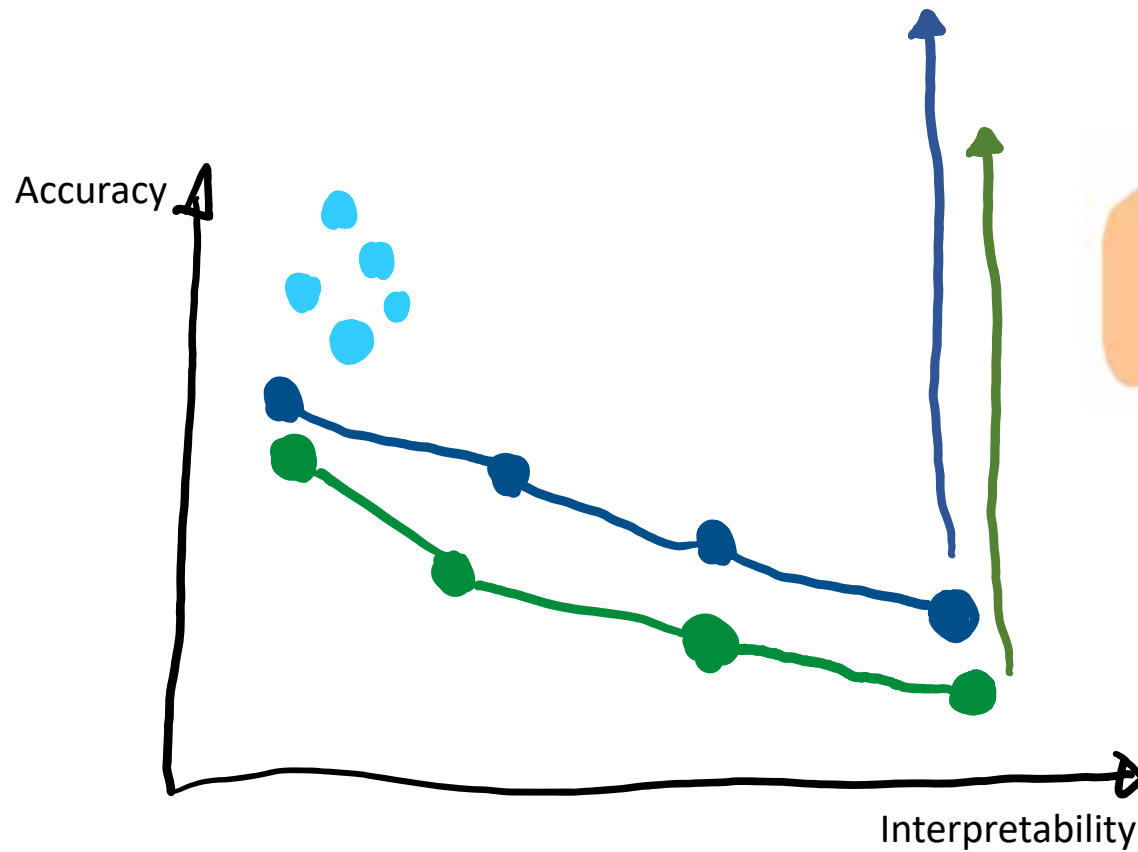
Adapt explanation to target user

# Model-Agnostic: Explain Any Classifier!



Make everything interpretable!

# Interpretable model



Avoid shortcuts,  
'real' accuracy

# Cheat sheet

## ① Purpose

• What are explanations for?

- Augment humans
- Evaluate the AI
- Improve the AI

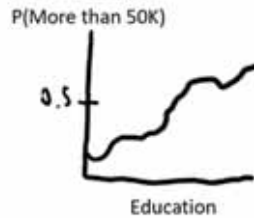
• What questions do you have?



## ② Explanation Type



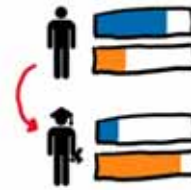
Linear model



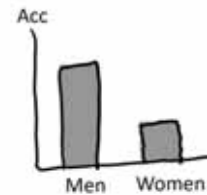
Partial Dependency Plot



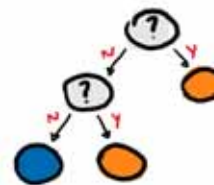
Anchor



Counterfactual



Sliced Statistics



Whole model

⋮

## ③ Technique

Interpretable models

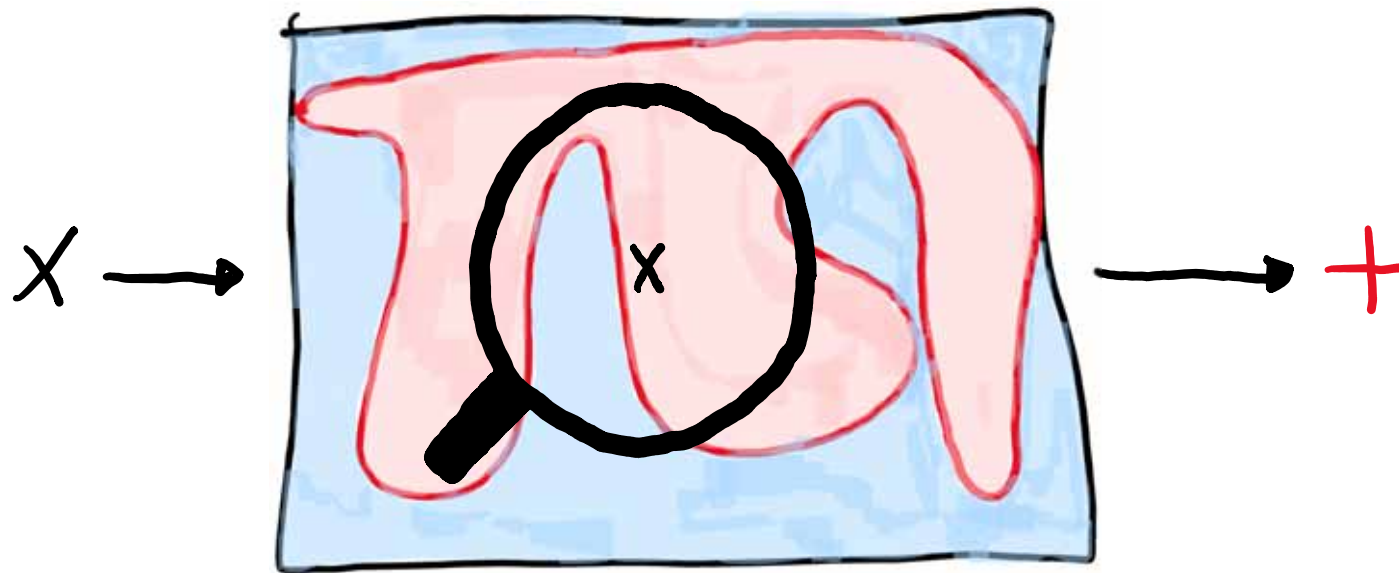
Black box

# LIME

"Why Should I Trust You?": Explaining the Predictions of Any Classifier  
Ribeiro et al, KDD 2016

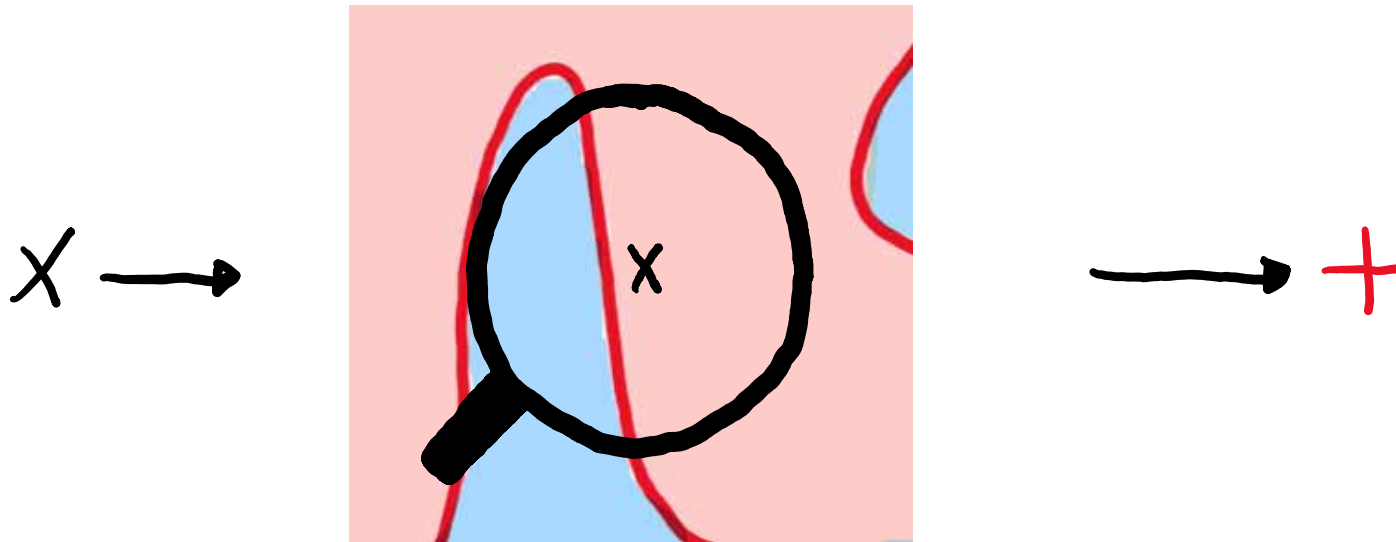


# Explaining individual predictions



“Global” explanation is too complicated

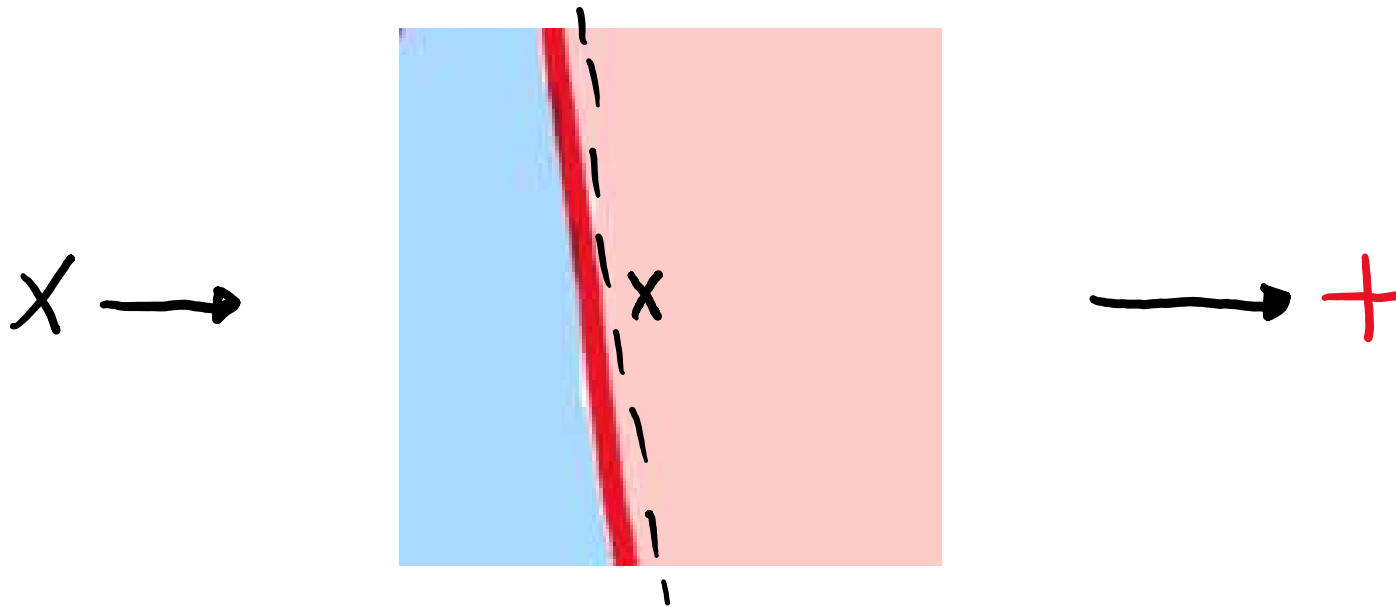
# Being Local...



“Global” explanation is too complicated

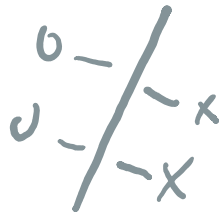
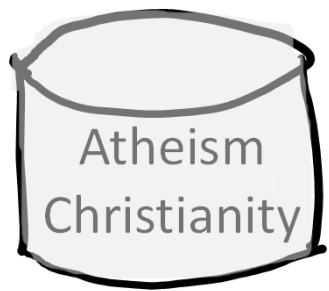
# Being Local...

Explanation is locally faithful

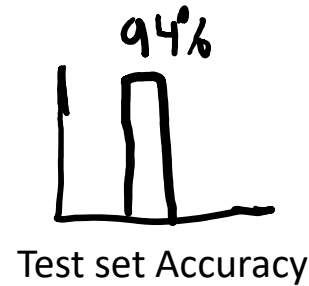


“Global” explanation is too complicated

# What an explanation looks like



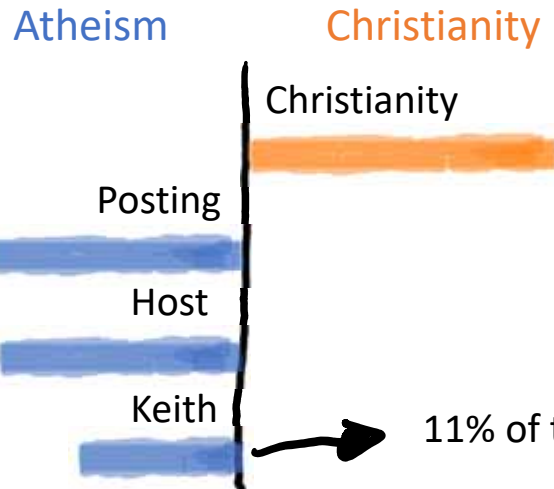
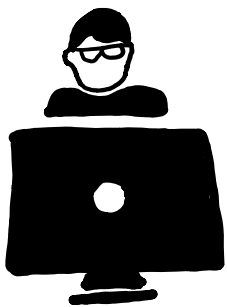
RBF SVM



From: Keith Richards  
Subject: Christianity is the answer  
NTTP-Posting-Host: x.x.com

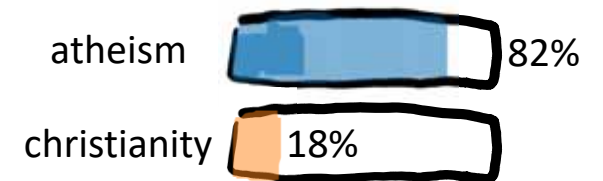
I think Christianity is the one true religion.  
If you'd like to know more, send me a note

Model seems good,  
let's look at some explanations  
on second thought...



11% of training, **always** in atheism

Prediction Probabilities



# Local Interpretable Model-Agnostic Explanations

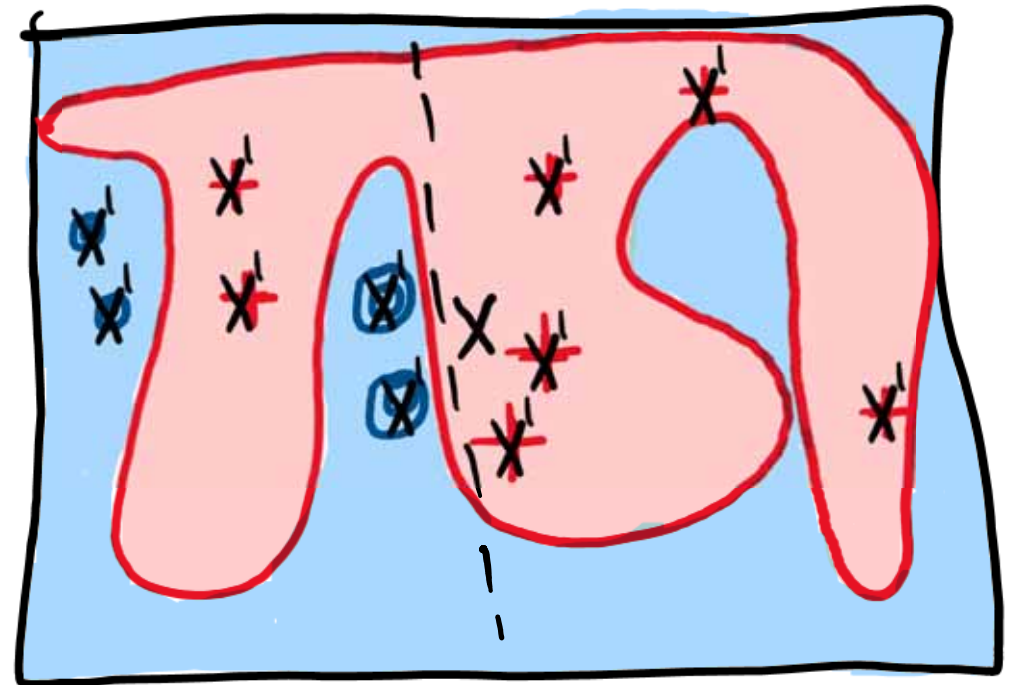
1. Sample points around  $x$

2. Get predictions from complex model

3. Weight samples according to distance to  $x$

4. Learn simple model on weighted samples

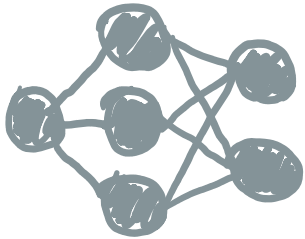
5. Use simple model to explain  $x$



# Interpretable representations

$x$  (embeddings)

3.1 | 1.7 | 8.4 | 0.1 | ...



$x'$  (words)

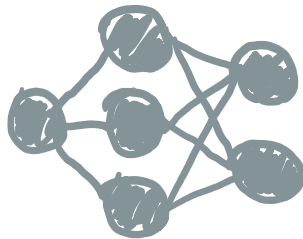
This is a horrible movie.



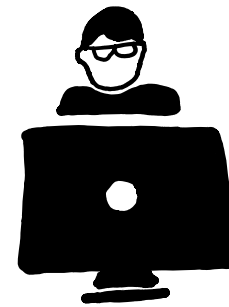
We use  $x'$  to perturb and explain

# Interpretable representation: images

$x$  (3 color channels / pixel)



$x'$  (contiguous superpixels)



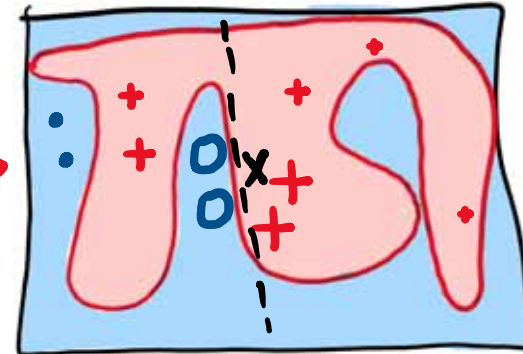
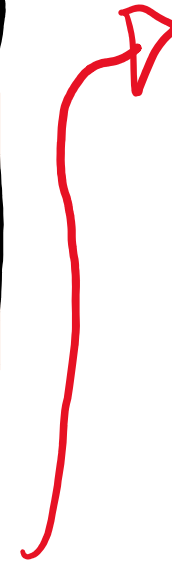
# Sampling example - images



Original Image  
 $P(\text{labrador}) = 0.21$



Perturbed Instances	$P(\text{Labrador})$
	0.92
	0.001
	0.34



Locally weighted regression



Explanation



# Explaining Global behavior

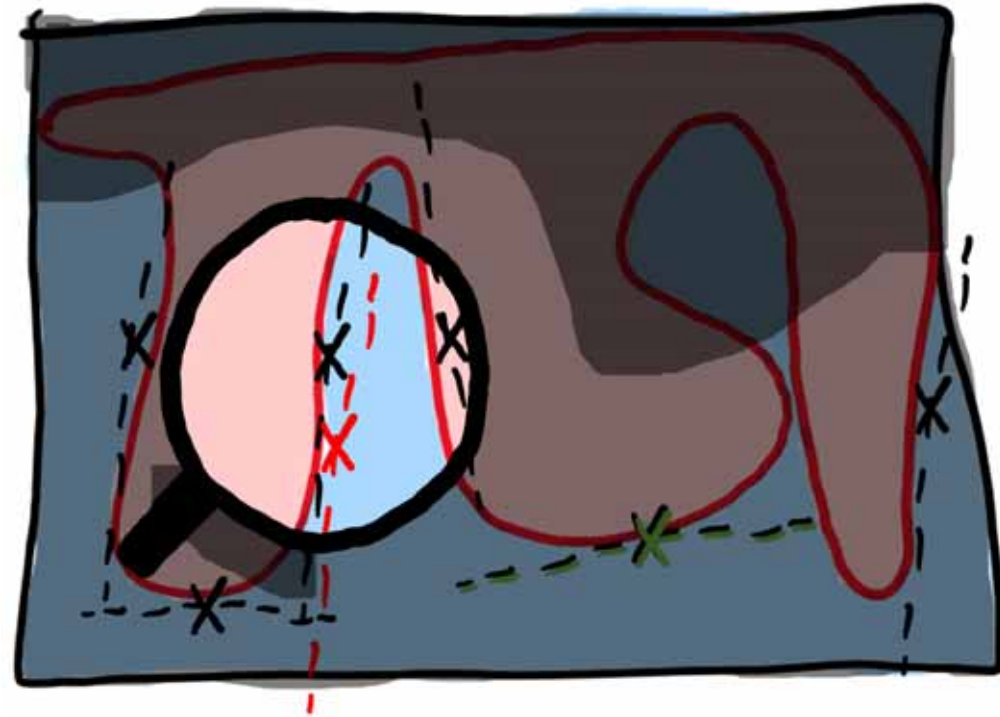
LIME explains a single prediction

Can't examine all explanations,  
pick  $k$  to show the user

Chosen set must be representative...

...and diverse

How: Submodular optimization



# LIME

## ① Purpose

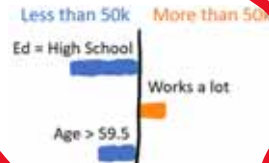
### • What are explanations for?

- ☐ Augment humans
- ☐ Evaluate the AI
- ☐ Improve the AI

### • What questions do you have?



## ② Explanation Type



Linear model

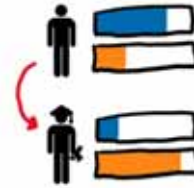
P(More than 50K)



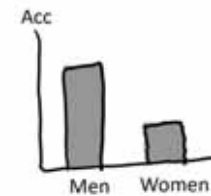
Partial Dependency Plot



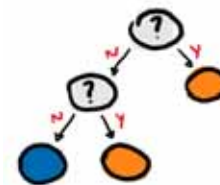
Anchor



Counterfactual



Sliced Statistics



⋮

## ③ Technique

Interpretable models

Black box

- LIME [Ribeiro et al 2016]

# Experiments

## Purpose

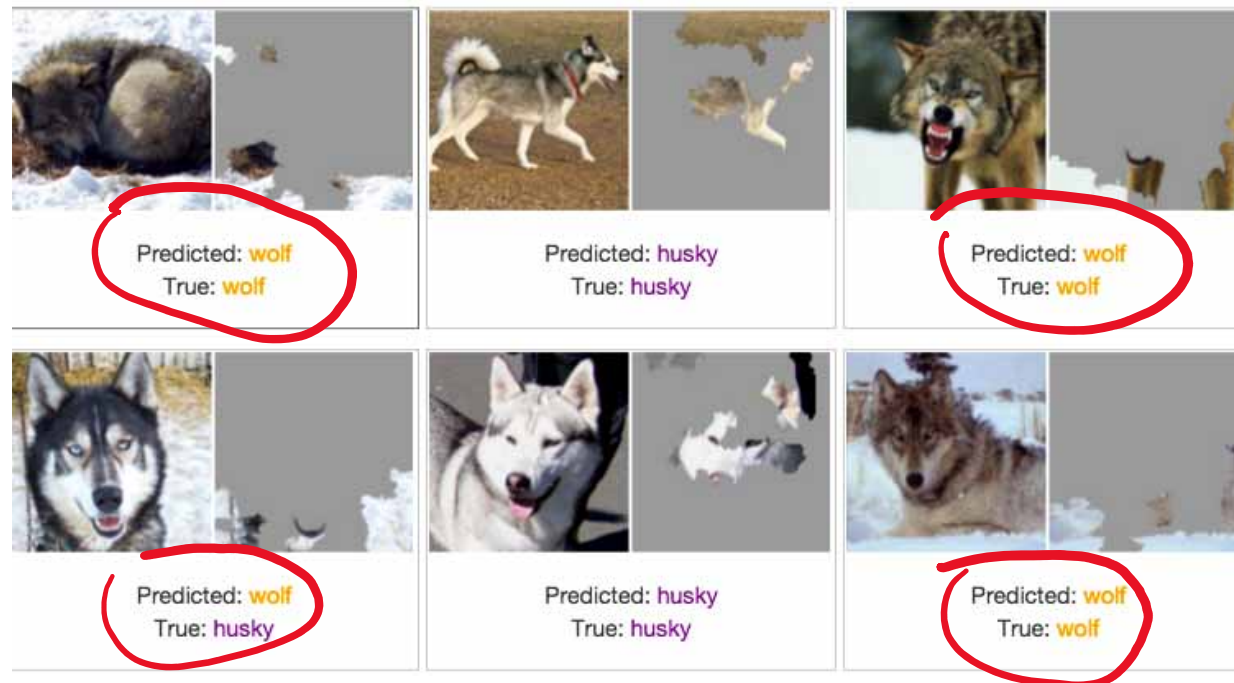
- Augment humans
- Evaluate the AI
- Improve the AI

# Experiment: Wolf or a Husky?



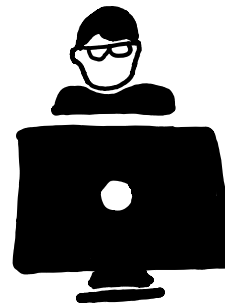
Only 1 mistake!

# Neural Network Explanations

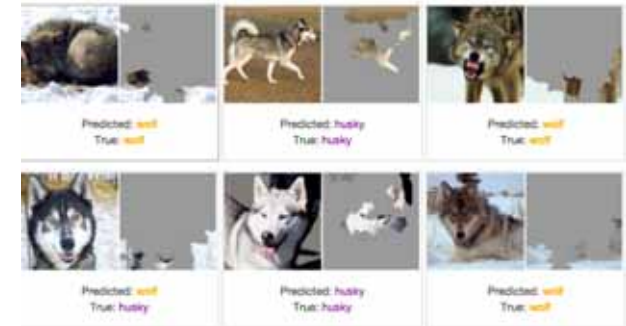


We built a deep snow detector

# Do ML people get this insight?



ML Person

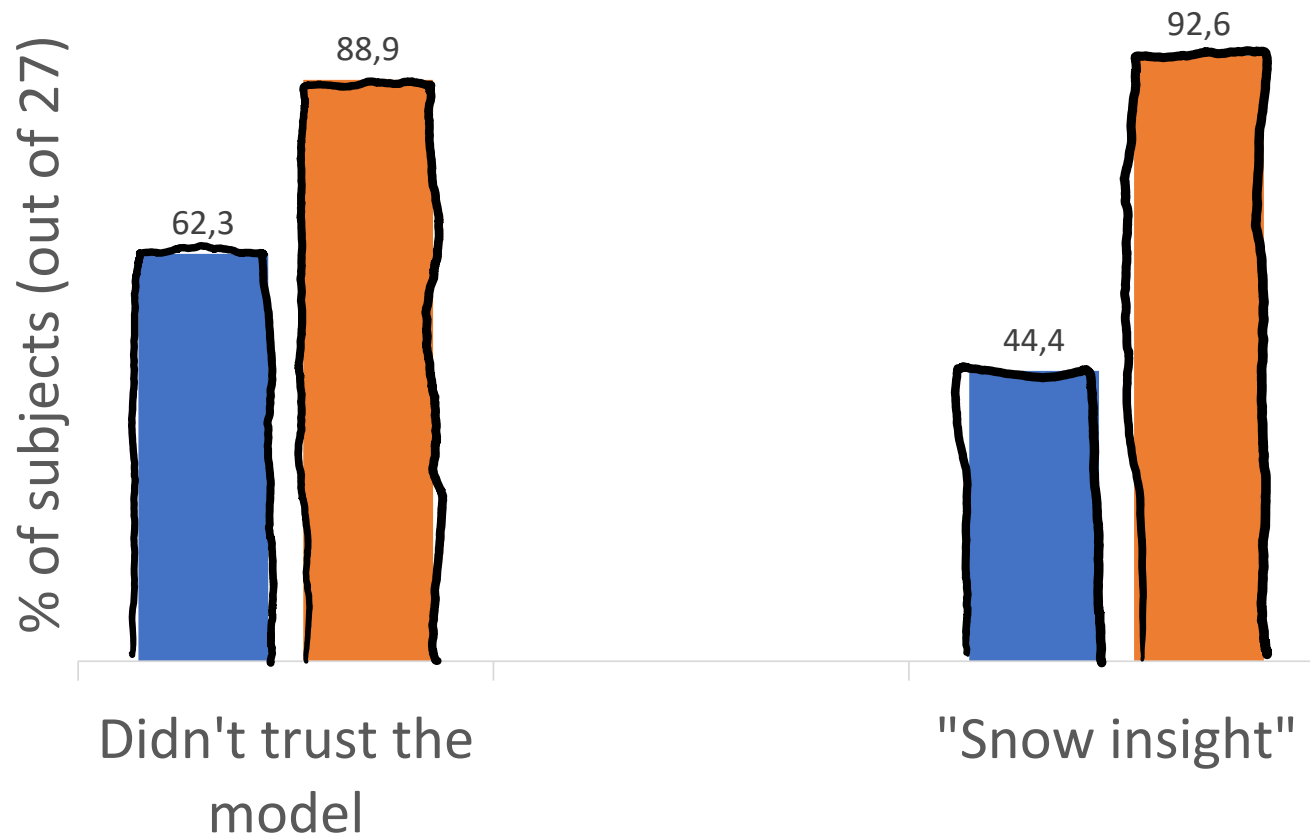


1. Would you trust this model?
2. What is the model doing?

# Explanations help experts get insights, avoid mistakes

■ Before explanations

■ After explanations



# Experiments

## Purpose

- Augment humans
- Evaluate the AI
- Improve the AI



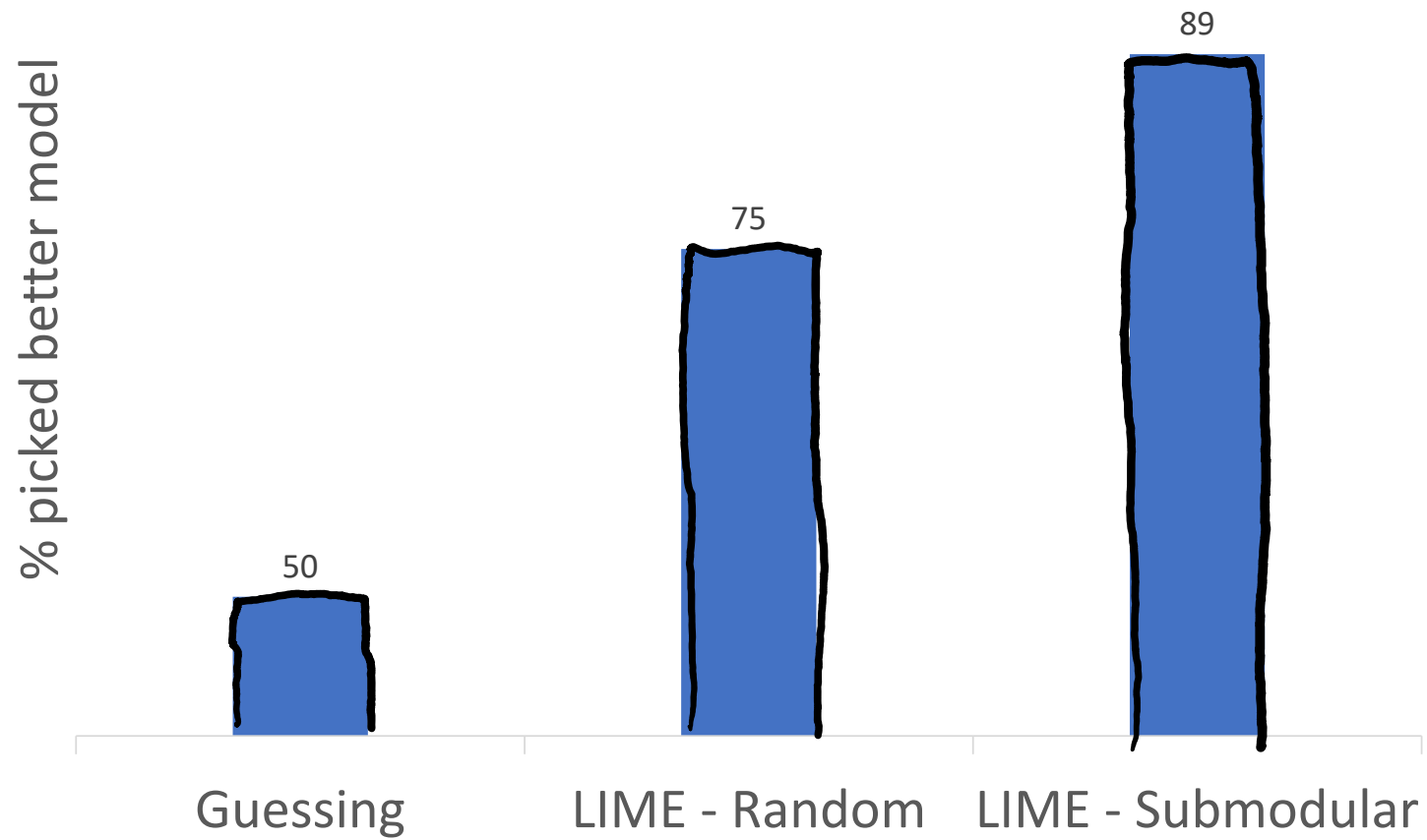
# Experiment: Model selection

Turkers asked to pick model that generalizes better

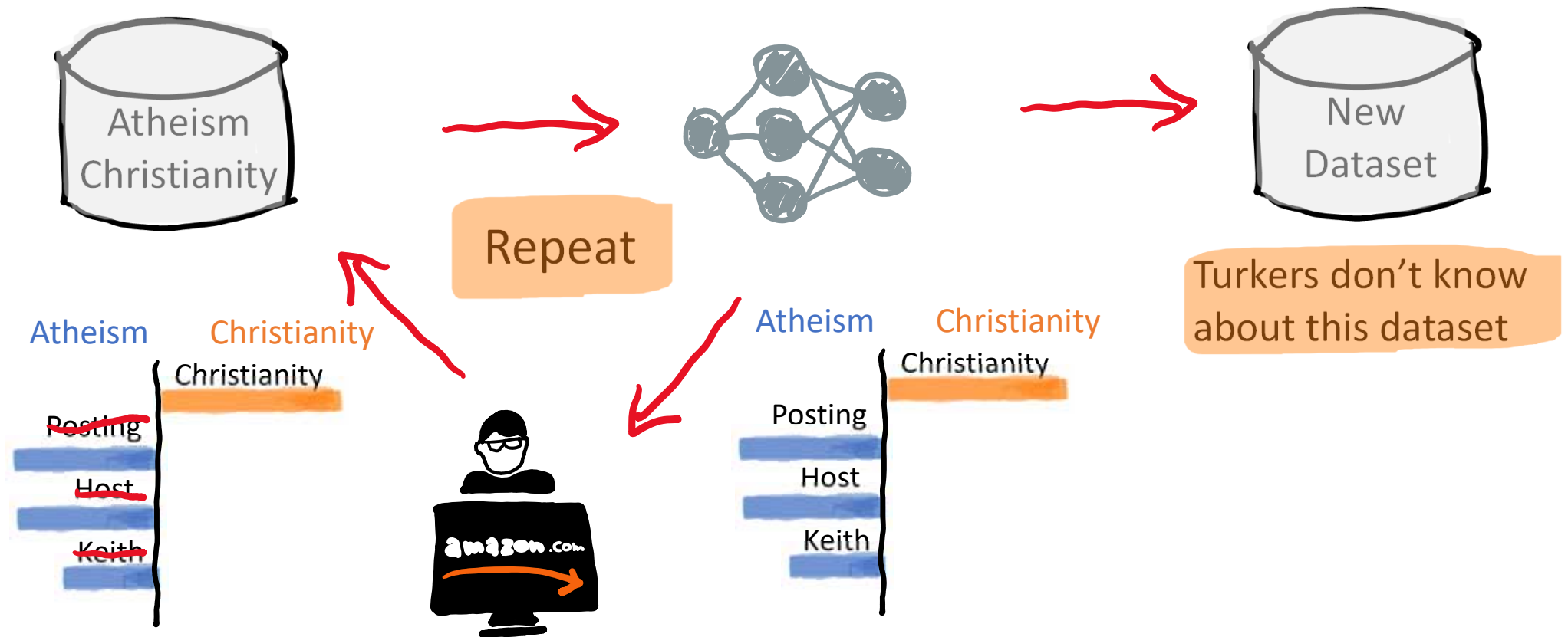
Example #3 of 6 True Class: ● Atheism Instructions Previous Next

Algorithm 1	Algorithm 2																								
<p><b>Words that A1 considers important:</b></p> <table style="width: 100%;"><tr><td>GOD</td><td style="text-align: right;">████████████████████</td></tr><tr><td>mean</td><td style="text-align: right;">██████████████████</td></tr><tr><td>anyone</td><td style="text-align: right;">████████████████</td></tr><tr><td>this</td><td style="text-align: right;">██████████████</td></tr><tr><td>Koresh</td><td style="text-align: right;">██████████</td></tr><tr><td>through</td><td style="text-align: right;">████████</td></tr></table> <p><b>Predicted:</b> <span style="color: magenta;">●</span> Atheism <b>Prediction correct:</b> <span style="color: green;">✓</span></p>	GOD	████████████████████	mean	██████████████████	anyone	████████████████	this	██████████████	Koresh	██████████	through	████████	<p><b>Words that A2 considers important:</b></p> <table style="width: 100%;"><tr><td>Posting</td><td style="text-align: right;">████████████████████</td></tr><tr><td>Host</td><td style="text-align: right;">██████████████████</td></tr><tr><td>Re</td><td style="text-align: right;">██████████</td></tr><tr><td>by</td><td style="text-align: right;">██████████</td></tr><tr><td>in</td><td style="text-align: right;">██████████</td></tr><tr><td>Nntp</td><td style="text-align: right;">████████</td></tr></table> <p><b>Predicted:</b> <span style="color: magenta;">●</span> Atheism <b>Prediction correct:</b> <span style="color: green;">✓</span></p>	Posting	████████████████████	Host	██████████████████	Re	██████████	by	██████████	in	██████████	Nntp	████████
GOD	████████████████████																								
mean	██████████████████																								
anyone	████████████████																								
this	██████████████																								
Koresh	██████████																								
through	████████																								
Posting	████████████████████																								
Host	██████████████████																								
Re	██████████																								
by	██████████																								
in	██████████																								
Nntp	████████																								
<p style="text-align: center;"><b>Document</b></p> <p>From: pauld@verdix.com (Paul Durbin) Subject: Re: DAVID CORESH IS! <span style="color: magenta;">GOD!</span> Nntp-Posting-Host: sarge.hq.verdix.com Organization: Verdix Corp Lines: 8</p>	<p style="text-align: center;"><b>Document</b></p> <p>From: pauld@verdix.com (Paul Durbin) Subject: <span style="color: magenta;">Re:</span> DAVID CORESH IS! <span style="color: magenta;">GOD!</span> <span style="color: magenta;">Nntp-Posting-Host:</span> sarge.hq.verdix.com Organization: Verdix Corp Lines: 8</p>																								

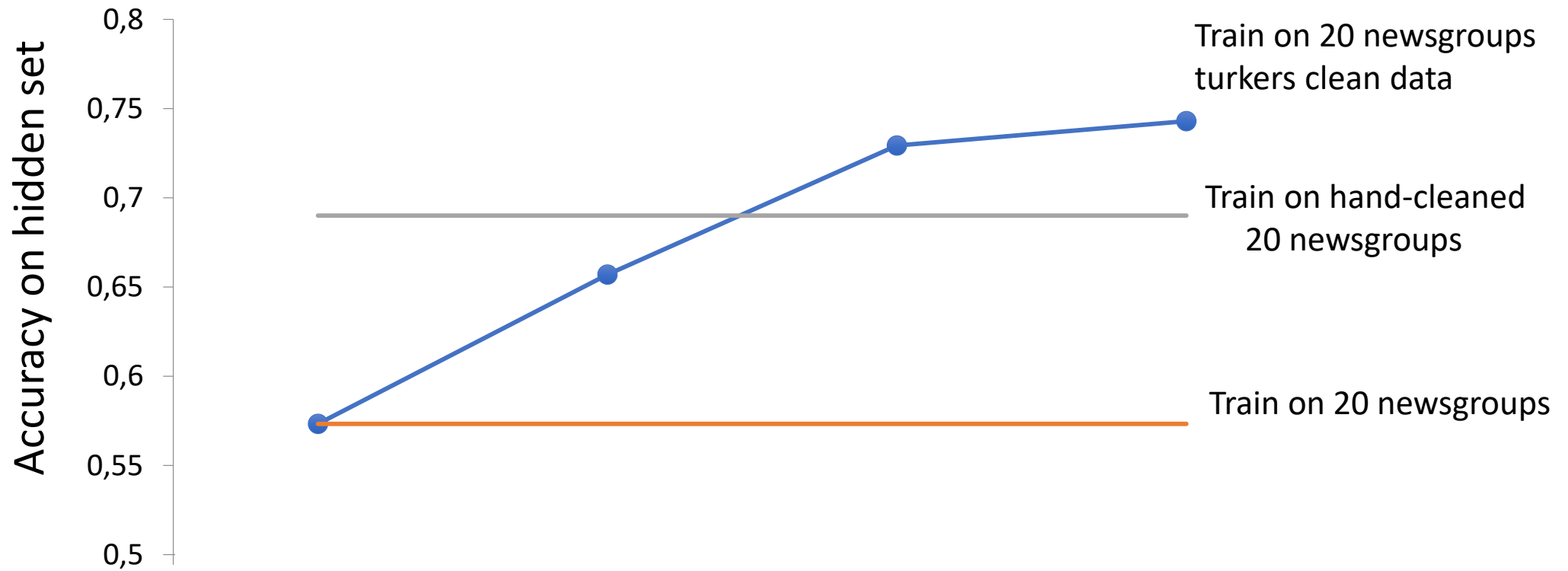
## Turkers can do model selection



# Fixing bad classifiers



## Turkers can do feature engineering



# Experiments

## Purpose

- Augment humans
- Evaluate the AI
- Improve the AI

# Cheat sheet

## ① Purpose

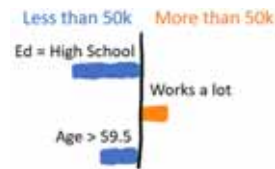
### • What are explanations for?

- ☐ Augment humans
- ☐ Evaluate the AI
- ☐ Improve the AI

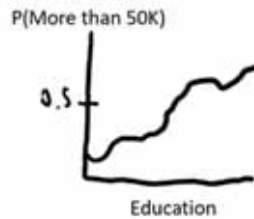
### • What questions do you have?



## ② Explanation Type



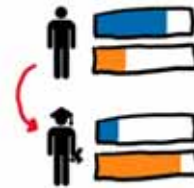
Linear model



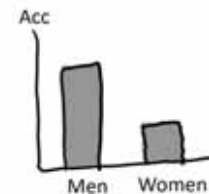
Partial Dependency Plot



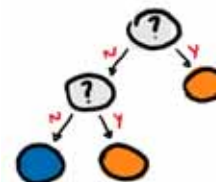
Anchor



Counterfactual



Sliced Statistics



Whole model

⋮

## ③ Technique

### Interpretable models

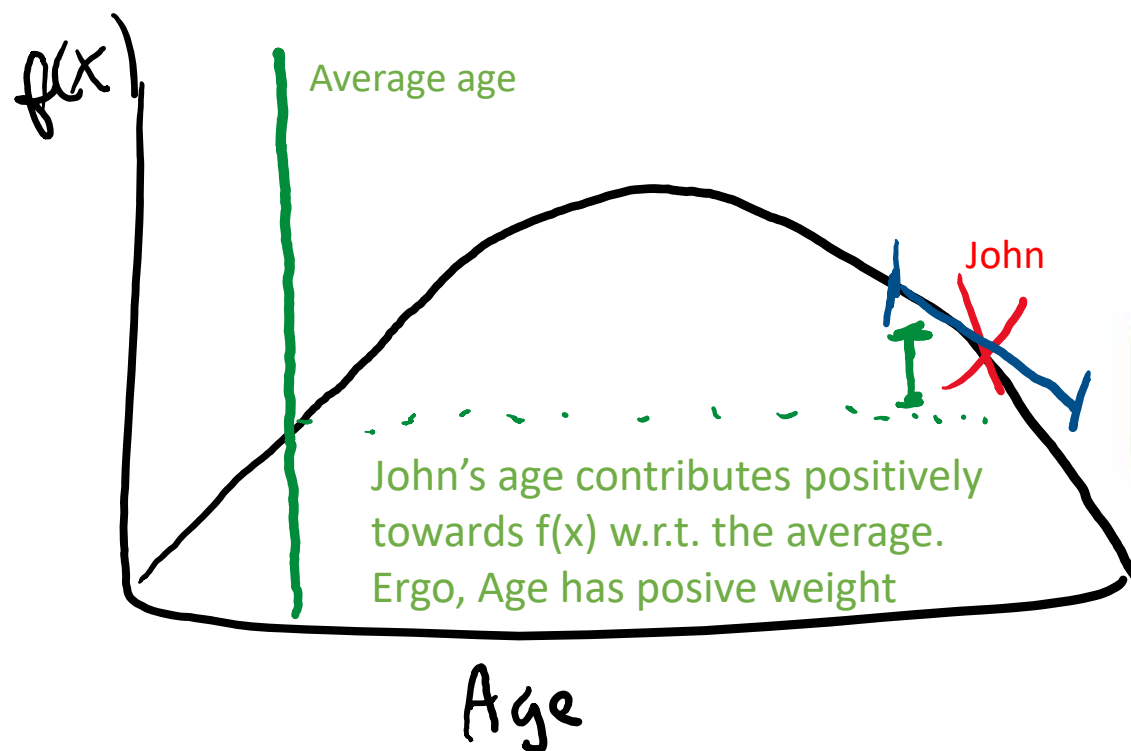
### Black box

- LIME [Ribeiro et al 2016]
- SHAP [Lundberg et al 2017]

# LIME vs SHAP

What's the difference?

# Different meanings for weights



LIME: weight is local approximation

SHAP: weight is contribution w.r.t baseline

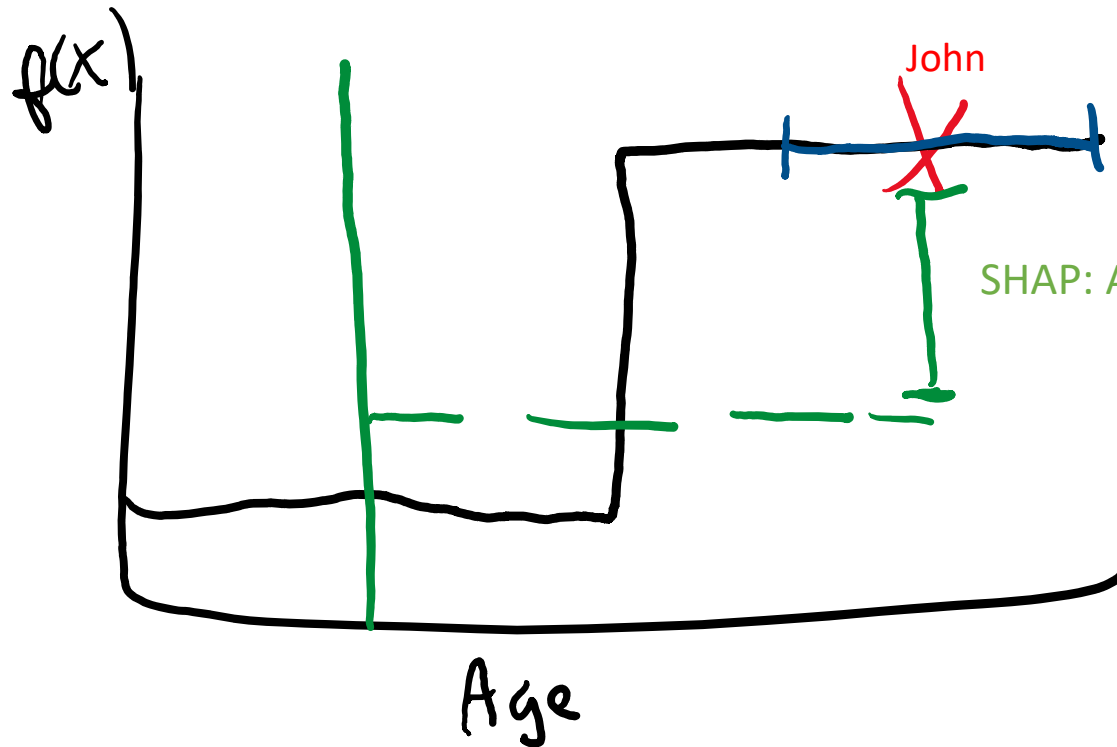
If you increase age,  $f(x)$  goes down  
if you decrease it,  $f(x)$  goes up  
Ergo, Age has negative weight

Which one is right?



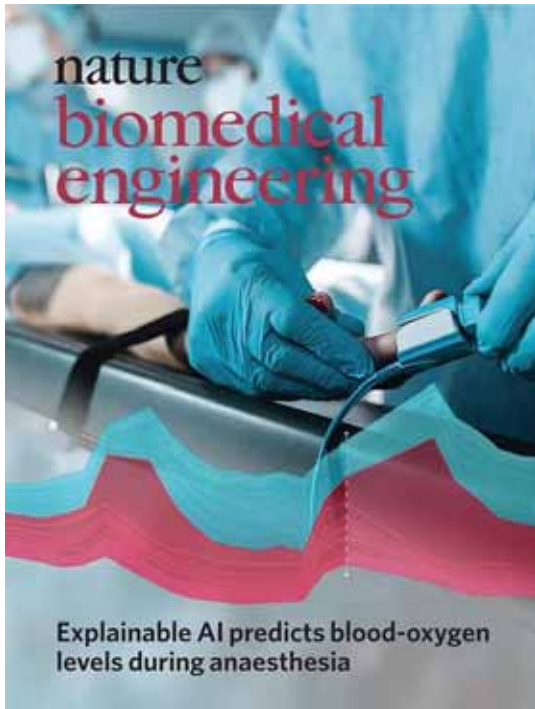
# Another example

LIME: Age doesn't matter

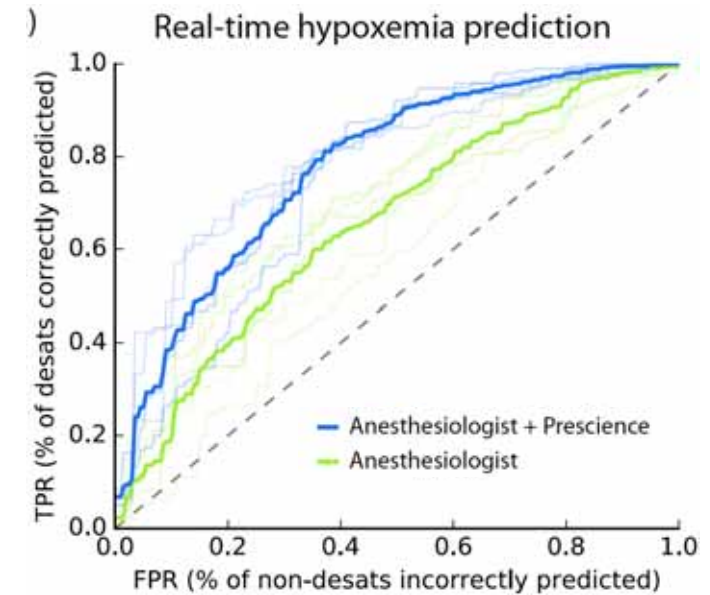


SHAP: Age has positive weight

# SHAP experiments Purpose



- Augment humans
- Evaluate the AI
- Improve the AI



Lundberg et al., Explainable machine-learning predictions for the prevention of hypoxemia during surgery, Nature Biomedical Engineering 2018 (*cover article*)

# Cheat sheet

## ① Purpose

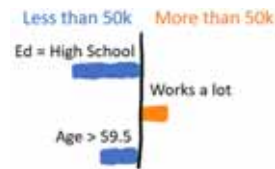
- What are explanations for?

- Augment humans
- Evaluate the AI
- Improve the AI

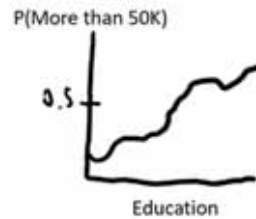
- What questions do you have?



## ② Explanation Type



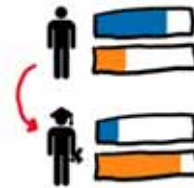
Linear model



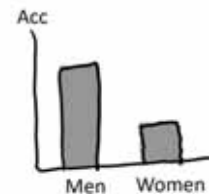
Partial Dependency Plot



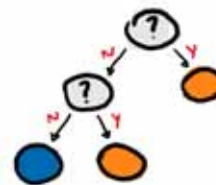
Anchor



Counterfactual



Sliced Statistics



Whole model

⋮

## ③ Technique

Interpretable models

Black box

- LIME [Ribeiro et al 2016]
- SHAP [Lundberg et al 2017]

# Explaining Explainable AI

Marco Tulio Ribeiro (Microsoft Research)