Transparency & Interpretability

Prof. Julia Stoyanovich
New York University

stoyanovich@nyu.edu
@stoyanoj

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

Supplementary reading

The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

Box 1 | Research questions

DOI:10.1038/s42256-020-0171-8

To trust the behavior of complex AI algorithms, especially in mission-critical settings, they must be made intelligible.

BY DANIEL S. WELD AND GAGAN BANSAL

The Challenge of Crafting Intelligible Intelligence

key insights

DOI:10.1145/3282486

Automated Decision Systems (ADS)

Automated Decision Systems (ADS)

process data about people

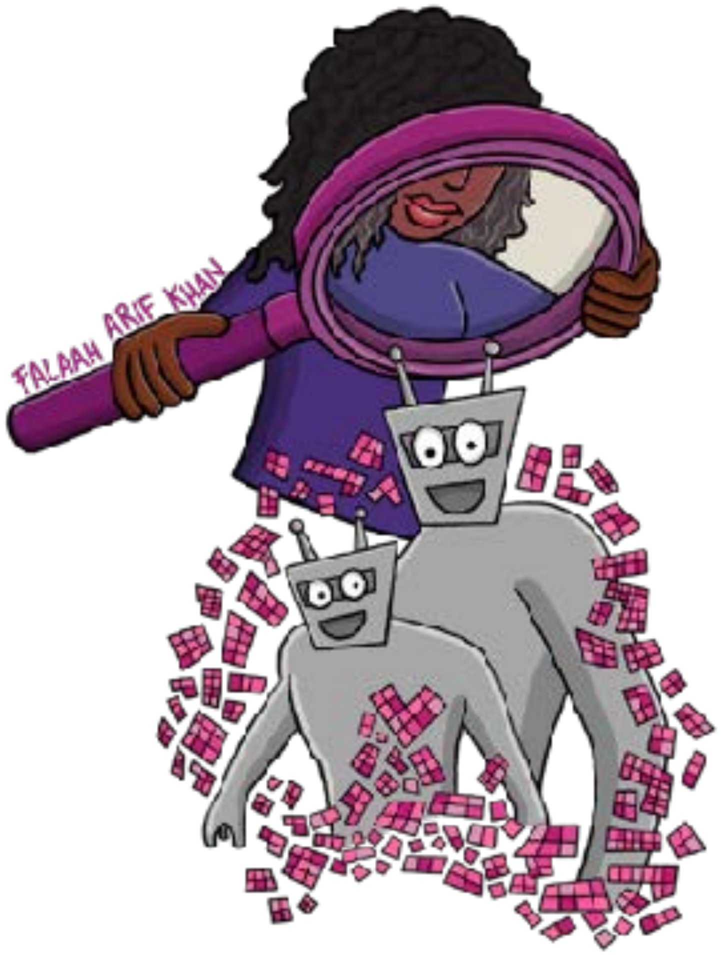
help make consequential decisions

combine human & automated decision making

aim to improve **efficiency** and promote **equity**

are subject to **auditing** and **public disclosure**

Terminology & vision



transparency, interpretability,
explainability, intelligibility

responsible AI

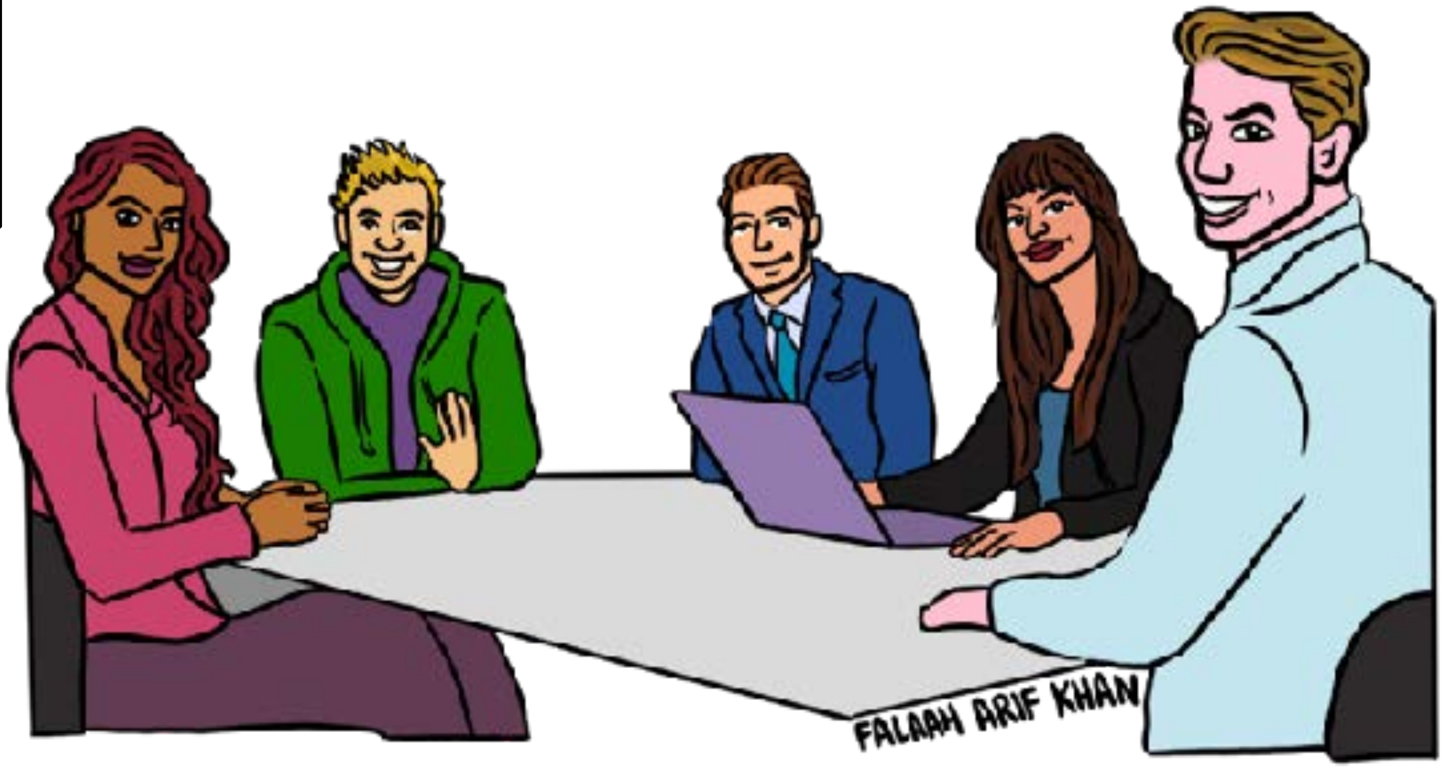
agency, responsibility

Interpretability for different stakeholders

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?



r/ai

examples

ADS in medical imaging

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

FACEBOOK AI NYU Langone Health

fastMRI

Accelerating MR Imaging with AI

What is fastMRI?

fastMRI is a collaborative research project between Facebook AI Research (FAIR) and NYU Langone Health. The aim is to investigate the use of AI to make MRI scans up to 10 times faster.

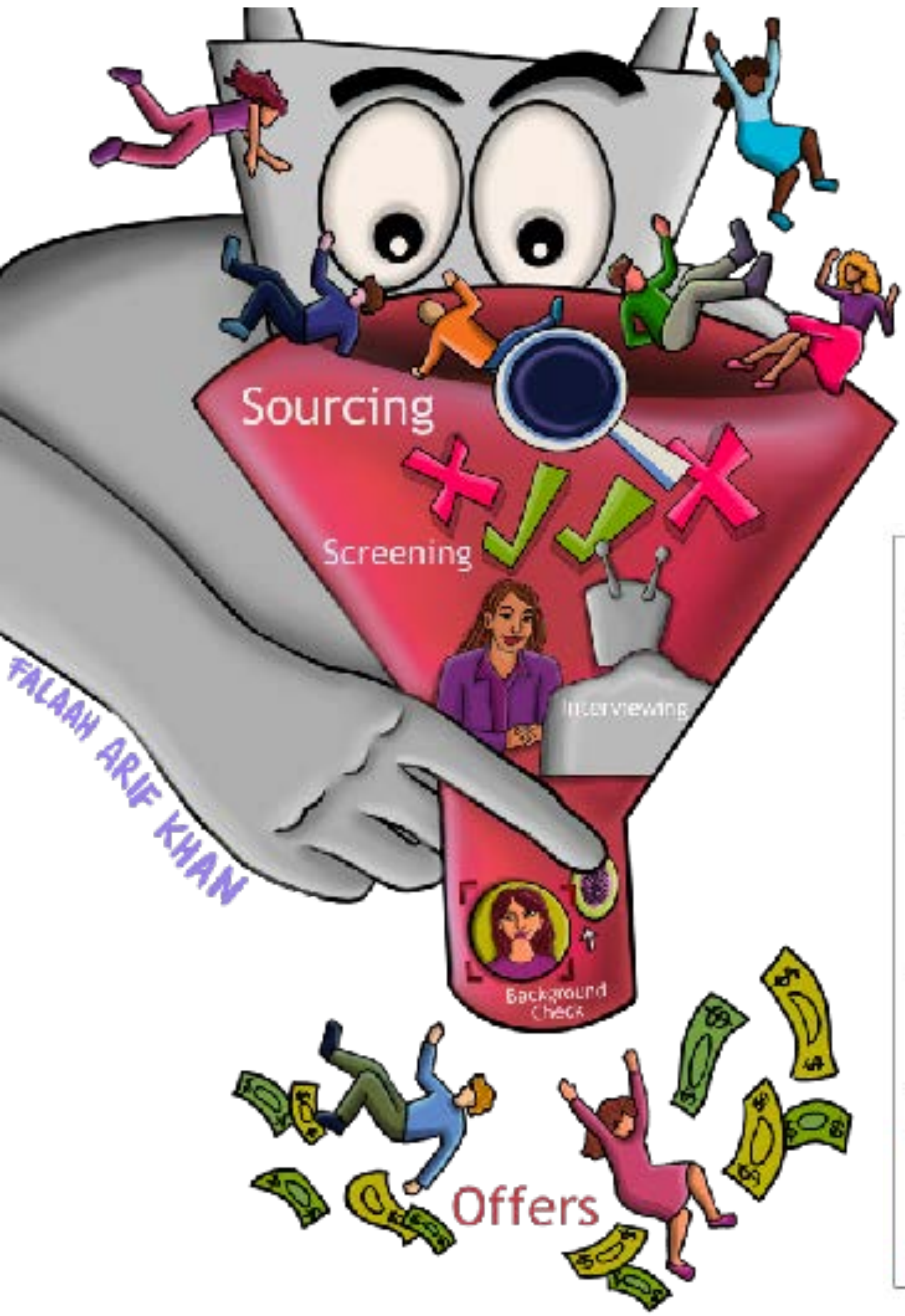
By producing accurate images from under-sampled data, AI image reconstruction has the potential to improve the patient's experience and to make MRIs accessible for more people.

To enable the broader research community to participate in this important project, NYU Langone Health has released fully anonymized raw data and image datasets. Visit our github repository, which contains baseline reconstruction models and PyTorch data loaders for the fastMRI dataset.

https://fastmri.org/

r/ai

ADS in hiring



What are we explaining?

To **Whom** are we explaining?

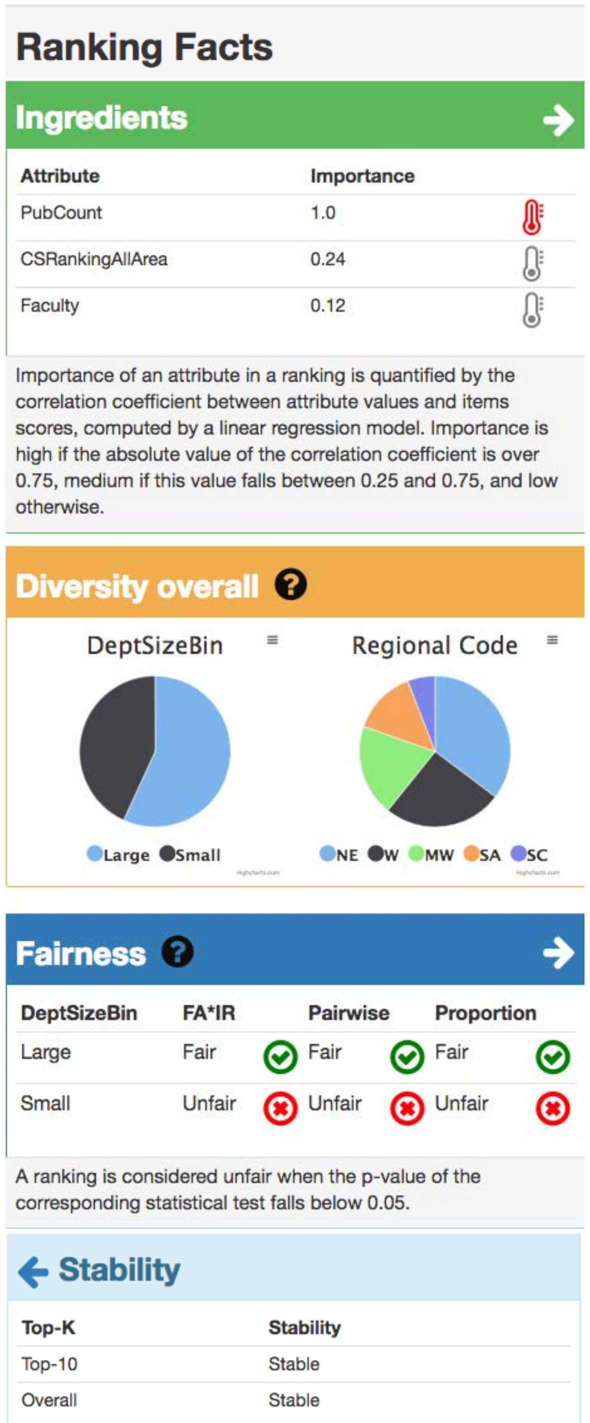
Why are we explaining?

ACCOUNTANT
Acme Partners

Qualifications:	BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications
Personal data to be analyzed:	An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.
Additional assessment:	AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to **selectively decline use of AI analysis for any of their personal** data or to review and challenge the results of such analysis.

Nutritional labels for ADS?



Ranking Facts

Ingredients

Attribute	Importance
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Diversity overall

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

comprehensible: short, simple, clear

consultative: provide actionable info

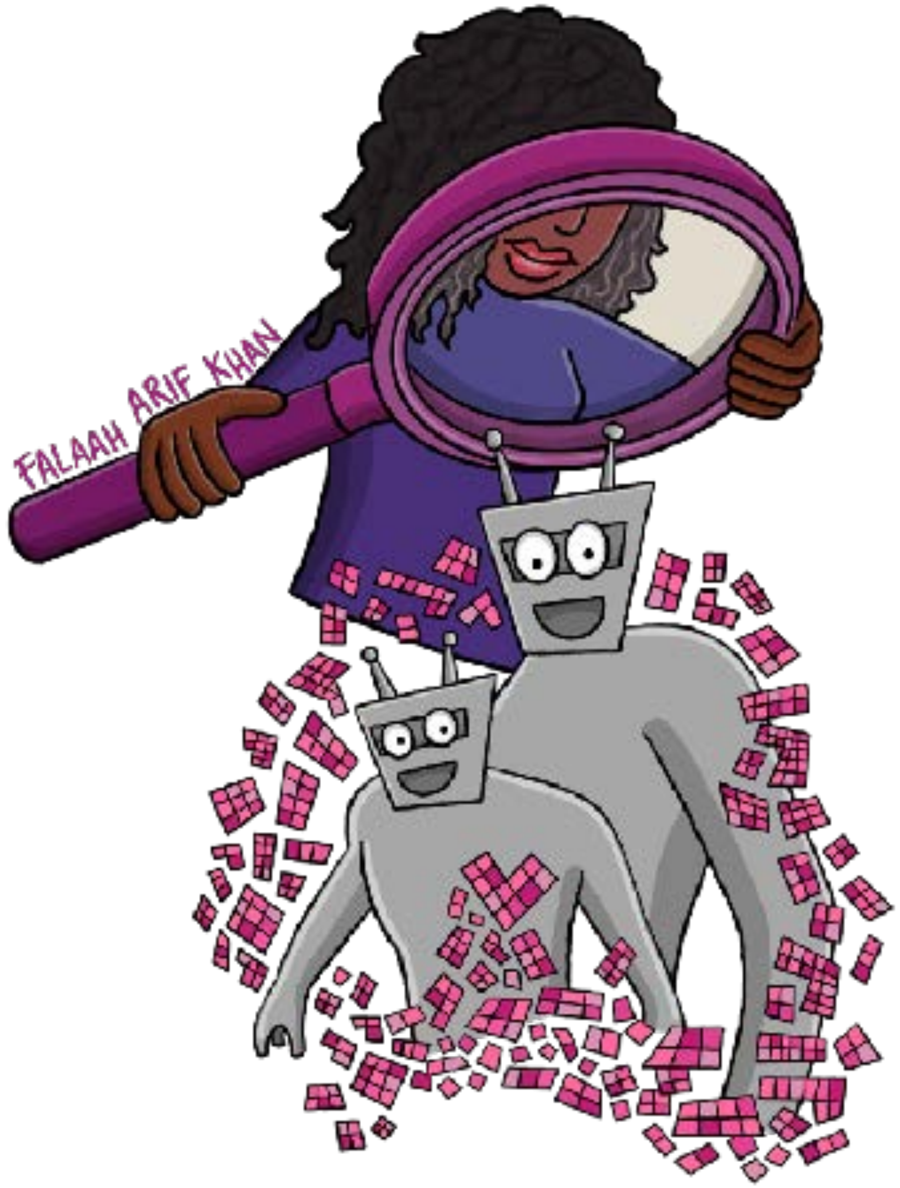
comparable: implying a standard

computable: incrementally constructed

[Yang, Stoyanovich, Asudeh, Howe, Jagadish, Miklau (2018)]
[Stoyanovich, Howe (2019)]

explaining black box models

What are we explaining?



How does a system work?

How **well** does a system work?

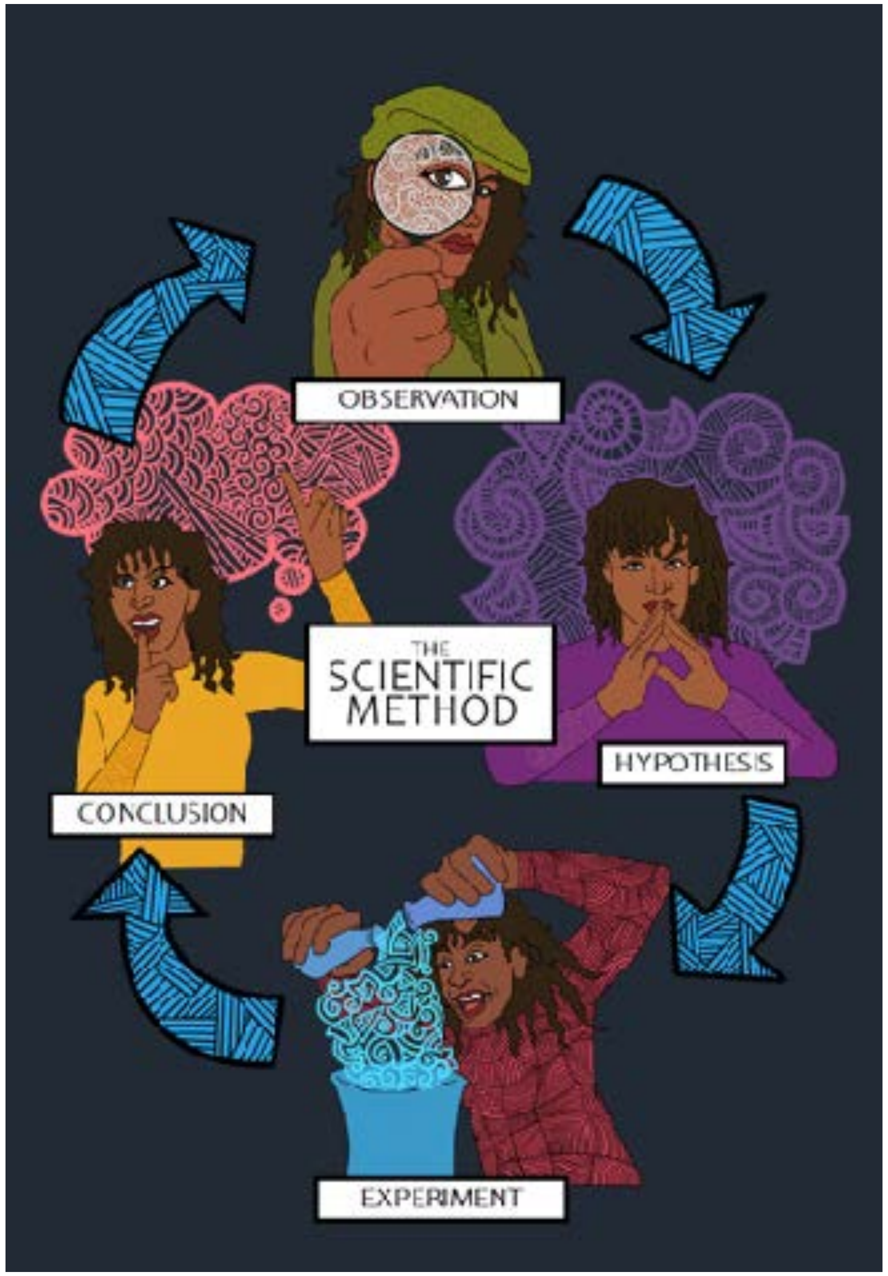
What does a system do?

Why was I ___ (mis-diagnosed / not offered a discount / denied credit) ?

Are a system's decisions discriminatory?

Are a system's decisions illegal?

But isn't accuracy sufficient?



How is accuracy measured? FPR / FNR / …

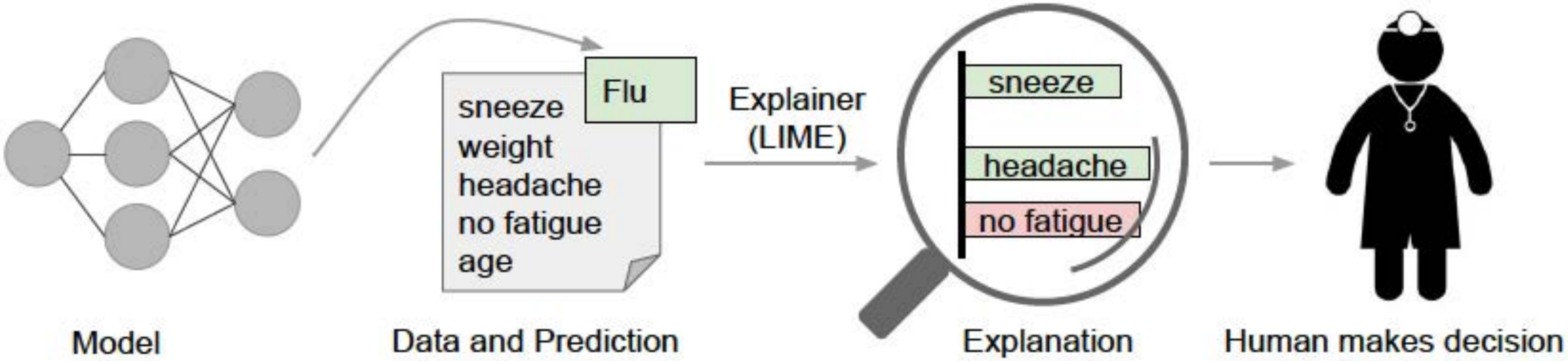
Accuracy for whom: over-all or in sub-populations?

Accuracy over which data?

There is never 100% accuracy. Mistakes for what reason?

Explanations based on features

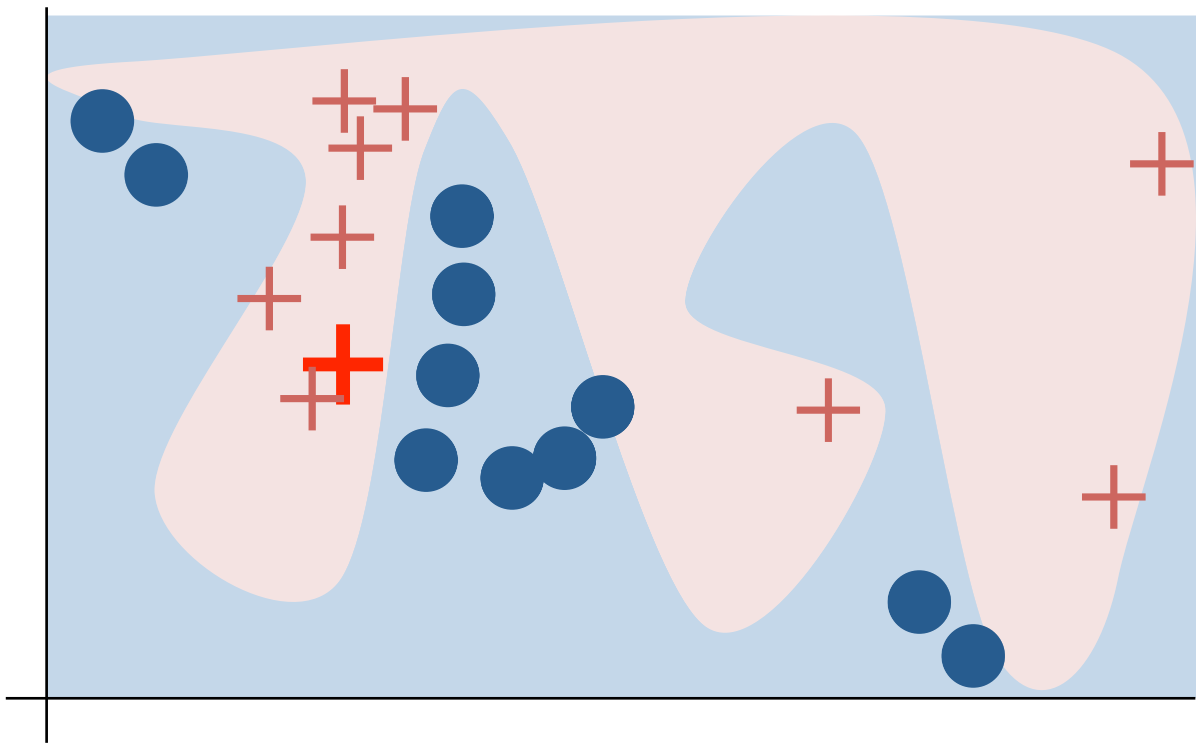
features in **green** (“sneeze”, “headache”) support the prediction (“Flu”), while features in **red** (“no fatigue”) are evidence against the prediction



what if patient id appears in green in the list?

LIME: Locally Interpretable Model-Agnostic Explanations

1. sample points around +
2. use original model to assign class labels



Key ideas

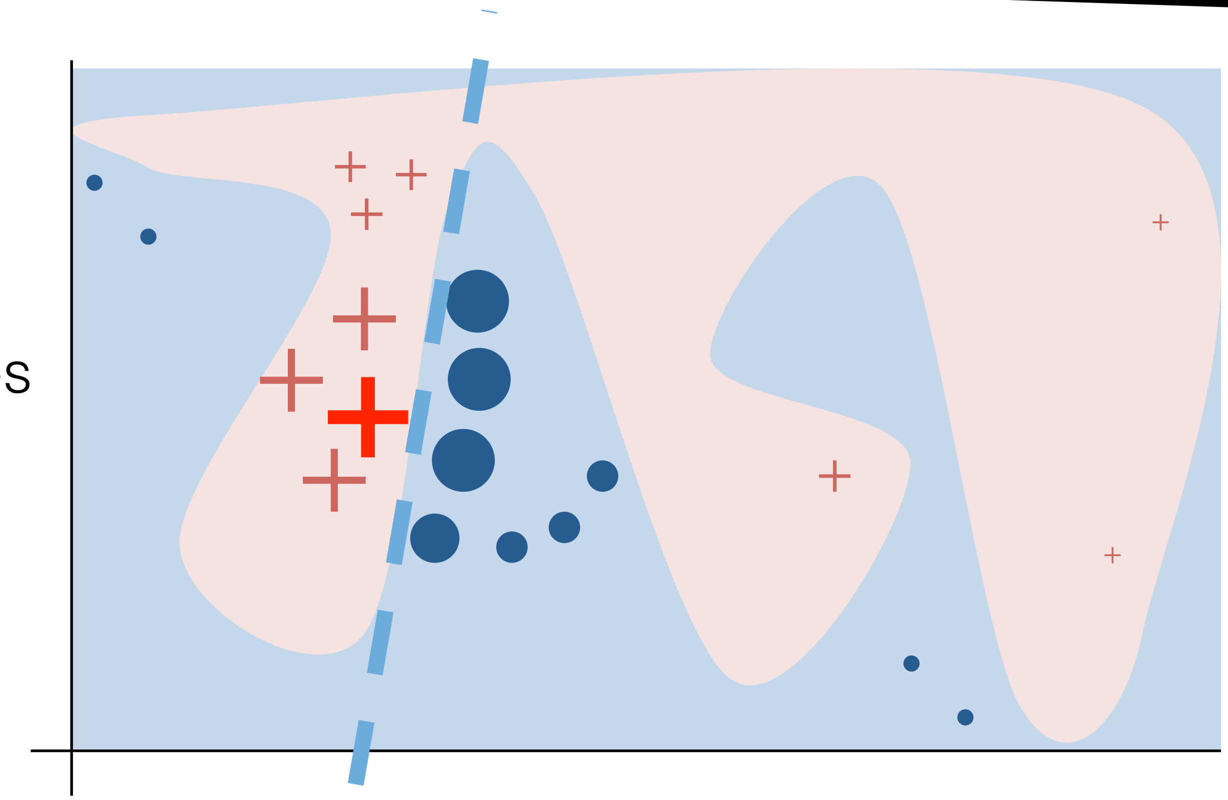
interpretable features

interpretable models

locally faithful explanations

LIME: Locally Interpretable Model-Agnostic Explanations

1. sample points around +
2. use original model to assign class labels
3. weigh points according to distance from +
4. learn interpretable model according to samples



Key ideas

interpretable features

interpretable models

locally faithful explanations

[Ribeiro, Singh & Guestrin, 2016]

r/ai

When accuracy is not enough

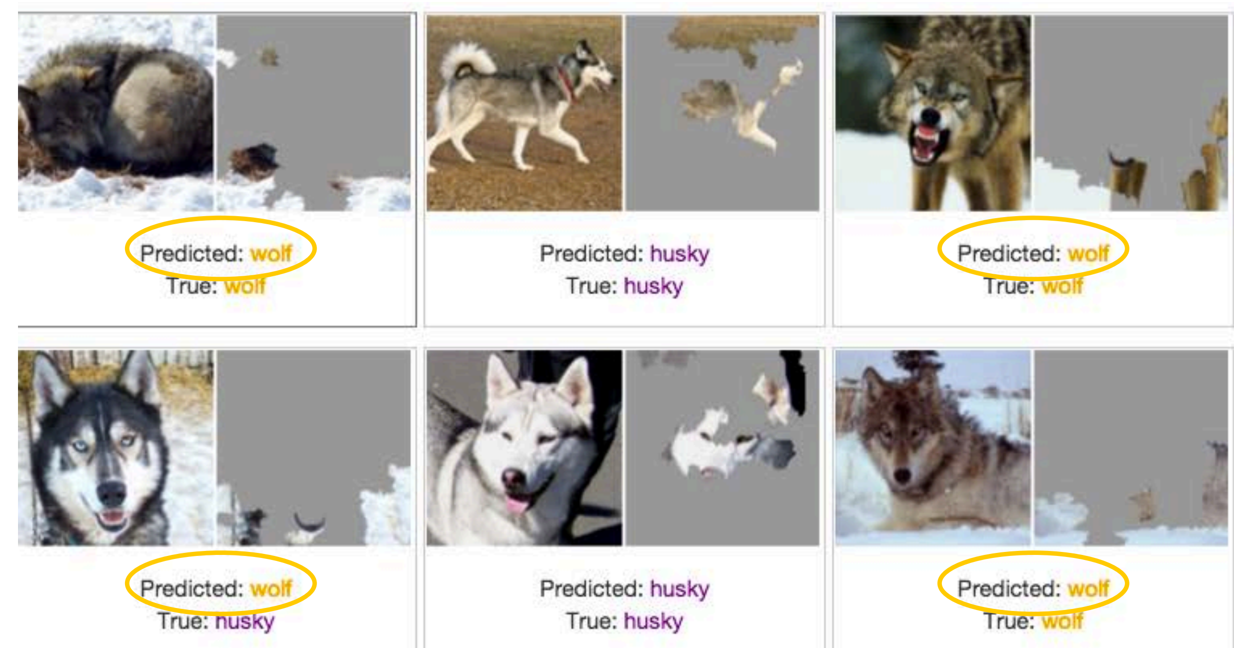
Train a neural network to predict wolf v. husky

Only 1 mistake!!!

Do you trust this model?
How does it distinguish between huskies and wolves?

A snow detector!

Explanations for neural network prediction



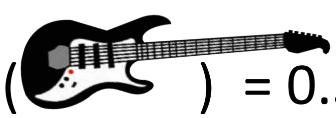
We've built a great snow detector… ☹

[Ribeiro, Singh & Guestrin, 2016]

When accuracy is not enough

Explaining Google's Inception NN

probabilities of the top-3 classes
and the super-pixels predicting each

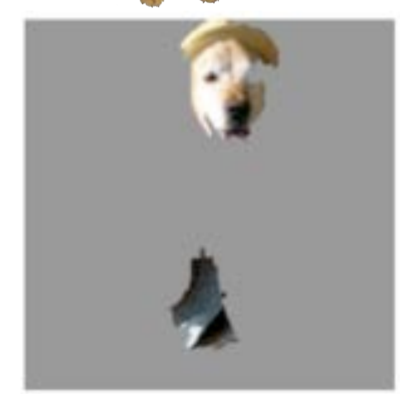
P() = 0.32

Electric guitar - incorrect but
reasonable, similar fretboard

P() = 0.24

Acoustic guitar

P() = 0.21



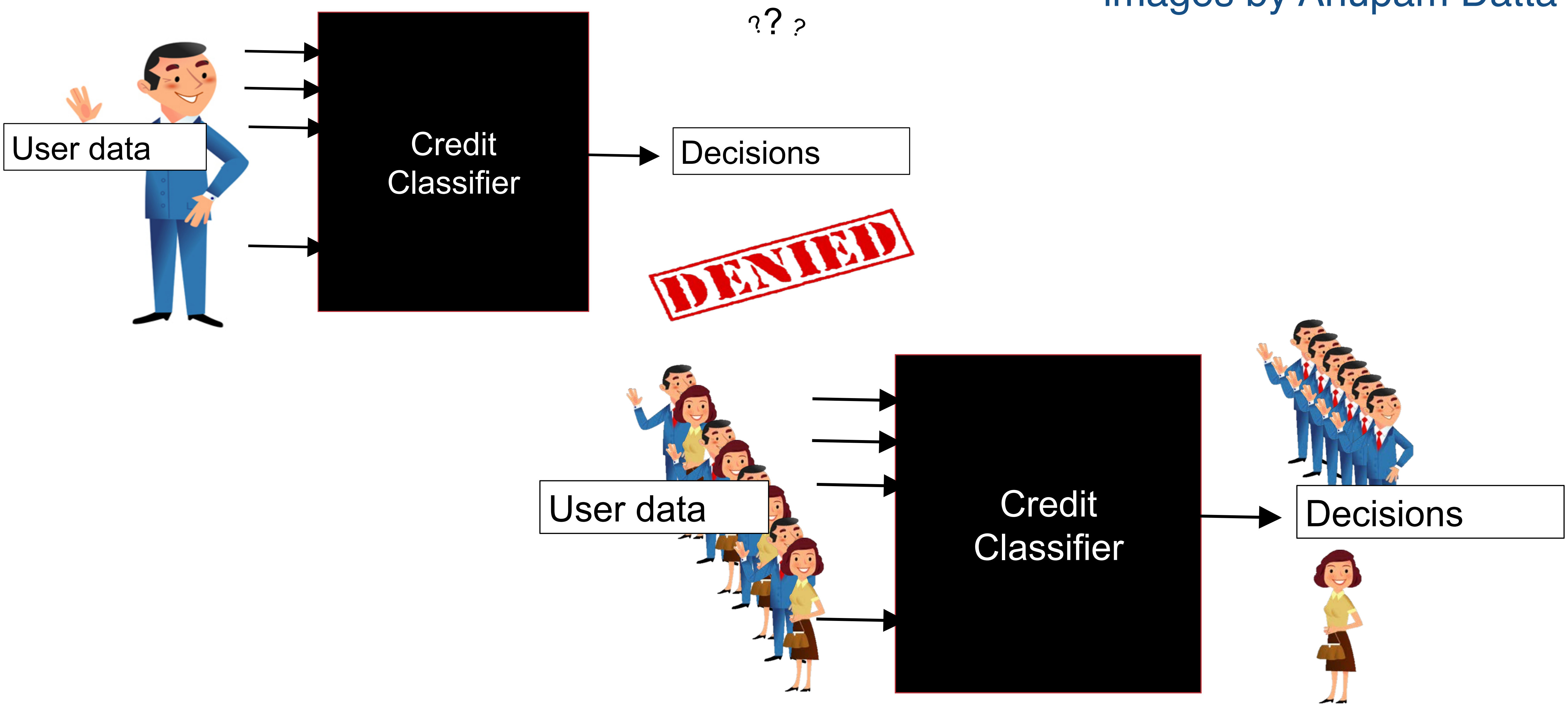
Labrador

[Ribeiro, Singh & Guestrin, 2016]

quick discussion

Auditing black-box models

images by Anupam Datta



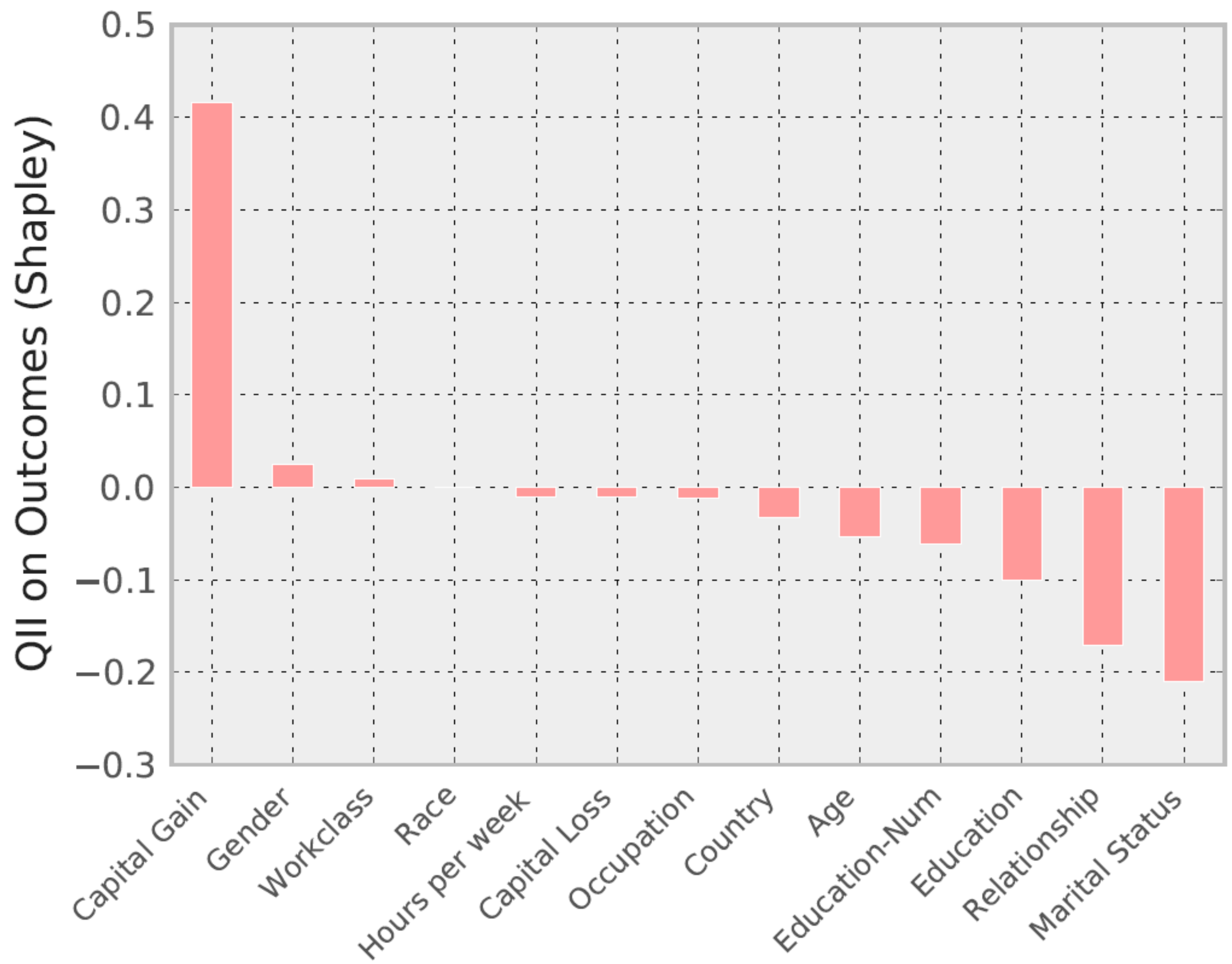
[Datta, Sen & Zick, 2016]

r/ai

Transparency report: Mr. X

How much influence do individual features have a given classifier's decision about an individual?

images by Anupam Datta



Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

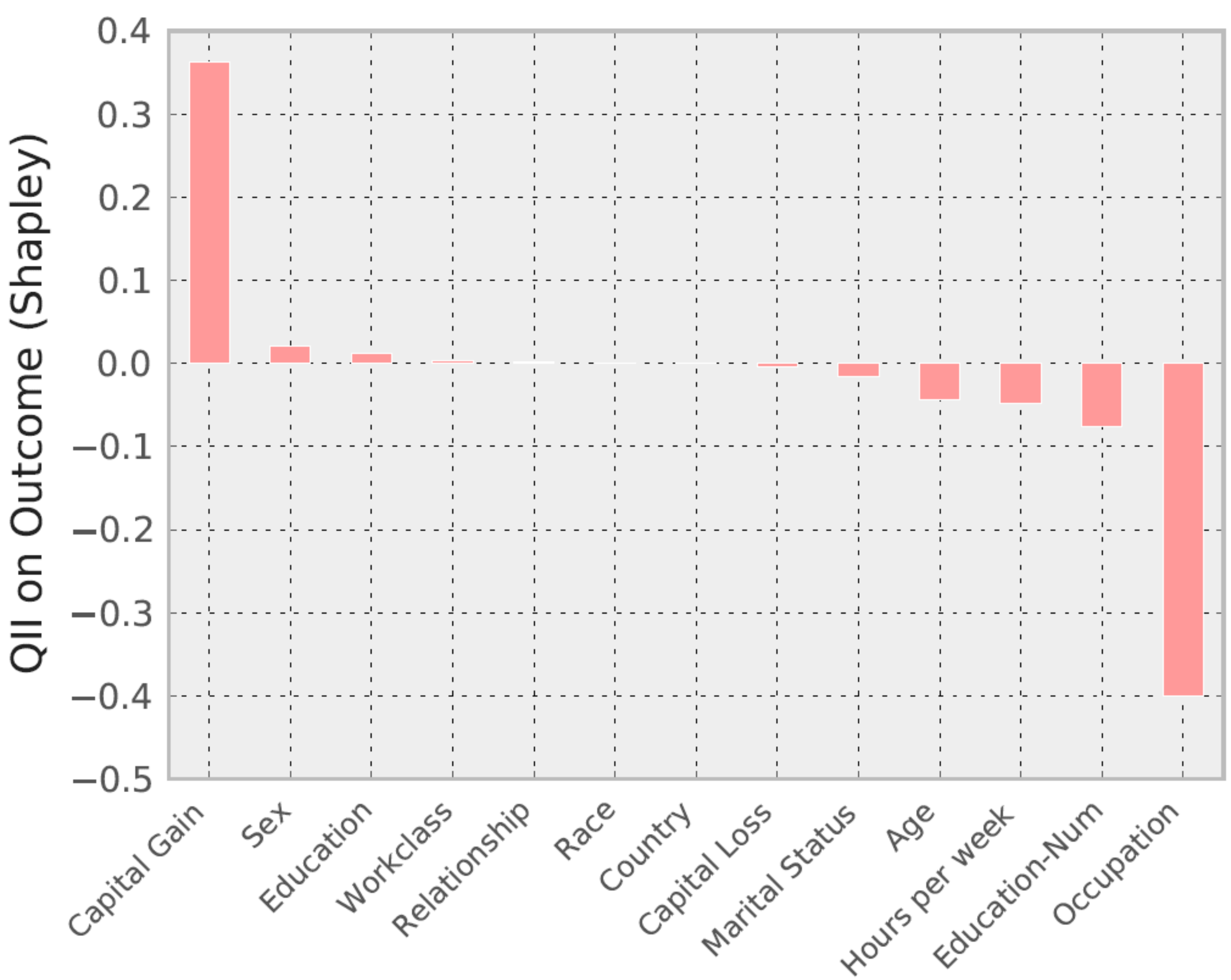
income

[Datta, Sen & Zick, 2016]

Transparency report: Mr. Y

Explanations for superficially similar individuals can be different

images by Anupam Datta



Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	$41310
Capital loss	$0
Work hours per week	24
Country	Mexico

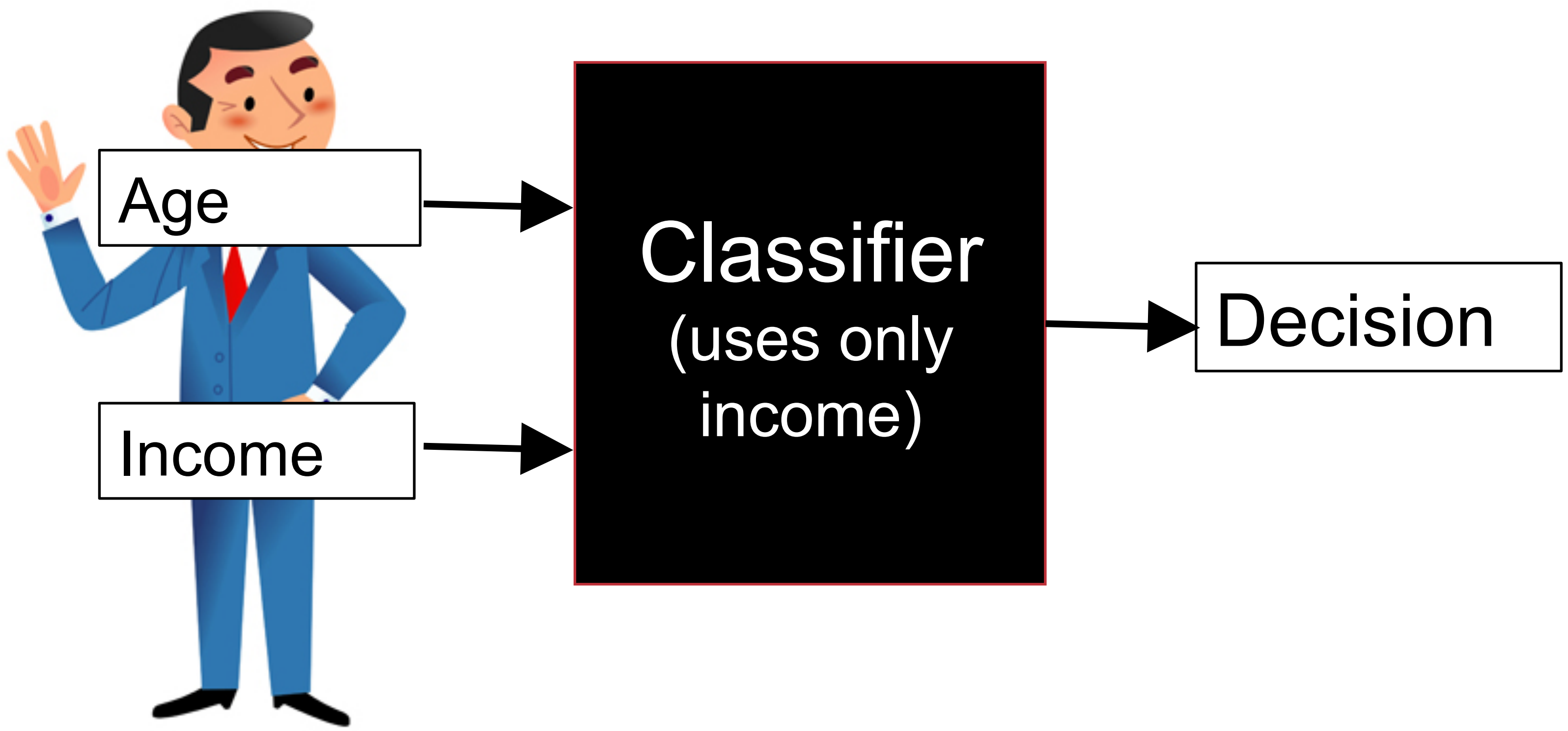
income

[Datta, Sen & Zick, 2016]

QII: Quantitative Input Influence

images by Anupam Datta

For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.



replace features with random values from the population, examine the distribution over outcomes

QII: Quantitative Input Influence

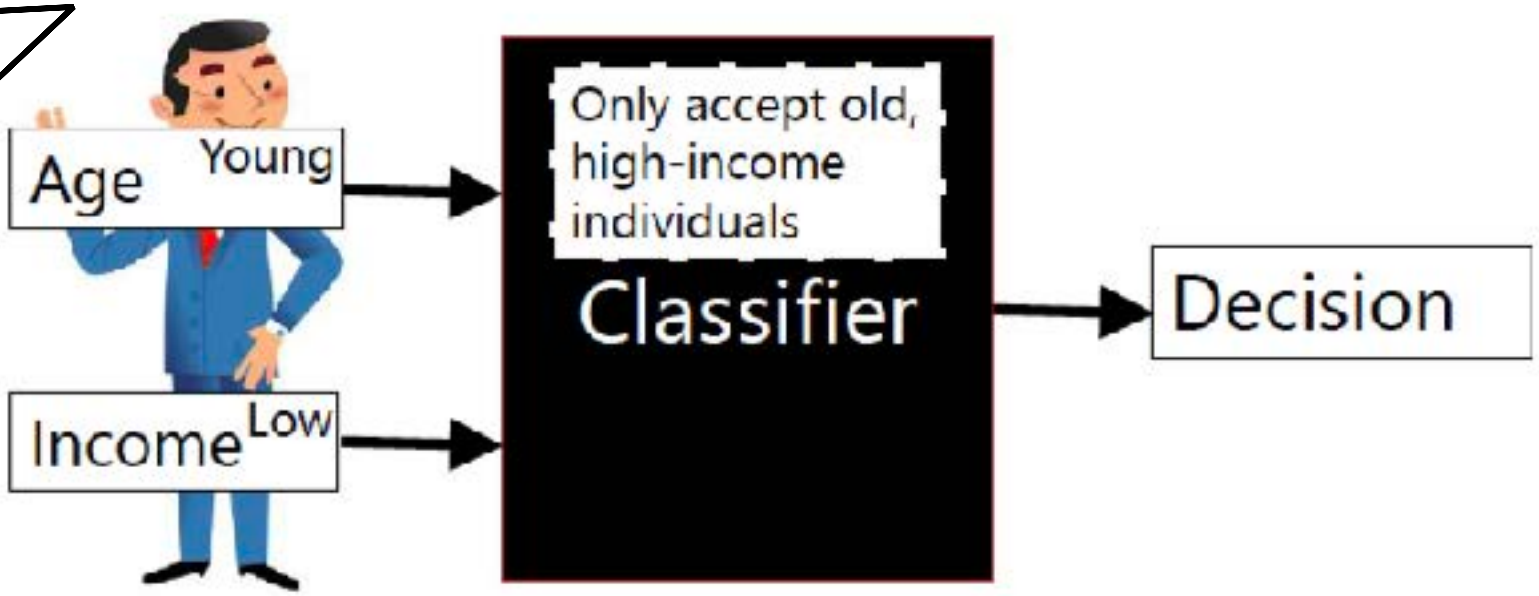
images by Anupam Datta

For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.

Key ideas

intervene on an input feature, measure its **importance**

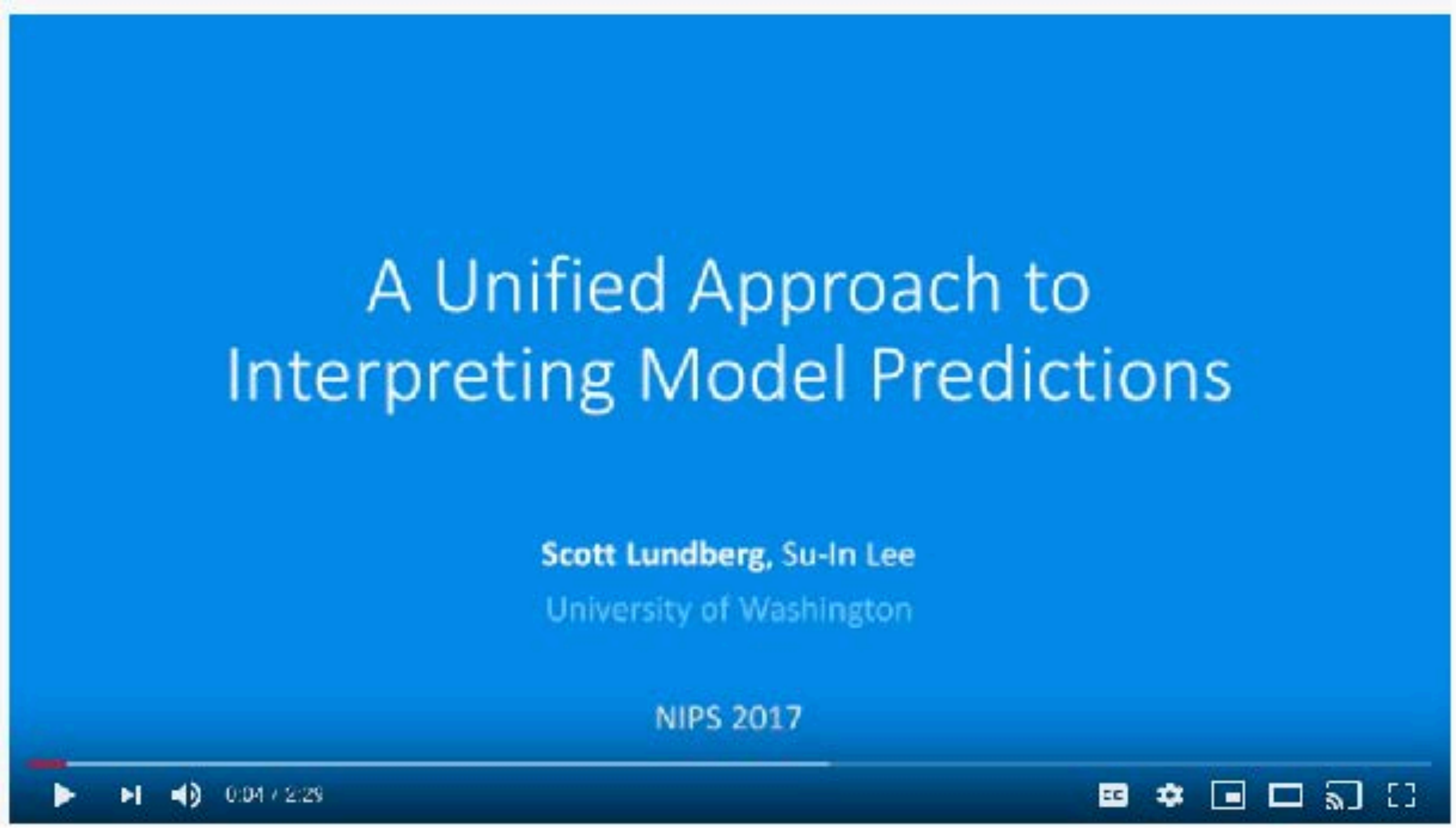
aggregate feature importance using its **Shapley value**



in this case, intervening on one feature at a time will have no effect

SHAP: Shapley Additive Explanations

A unifying framework for interpreting predictions with “additive feature attribution methods”, including LIME and QII, for **local explanations**

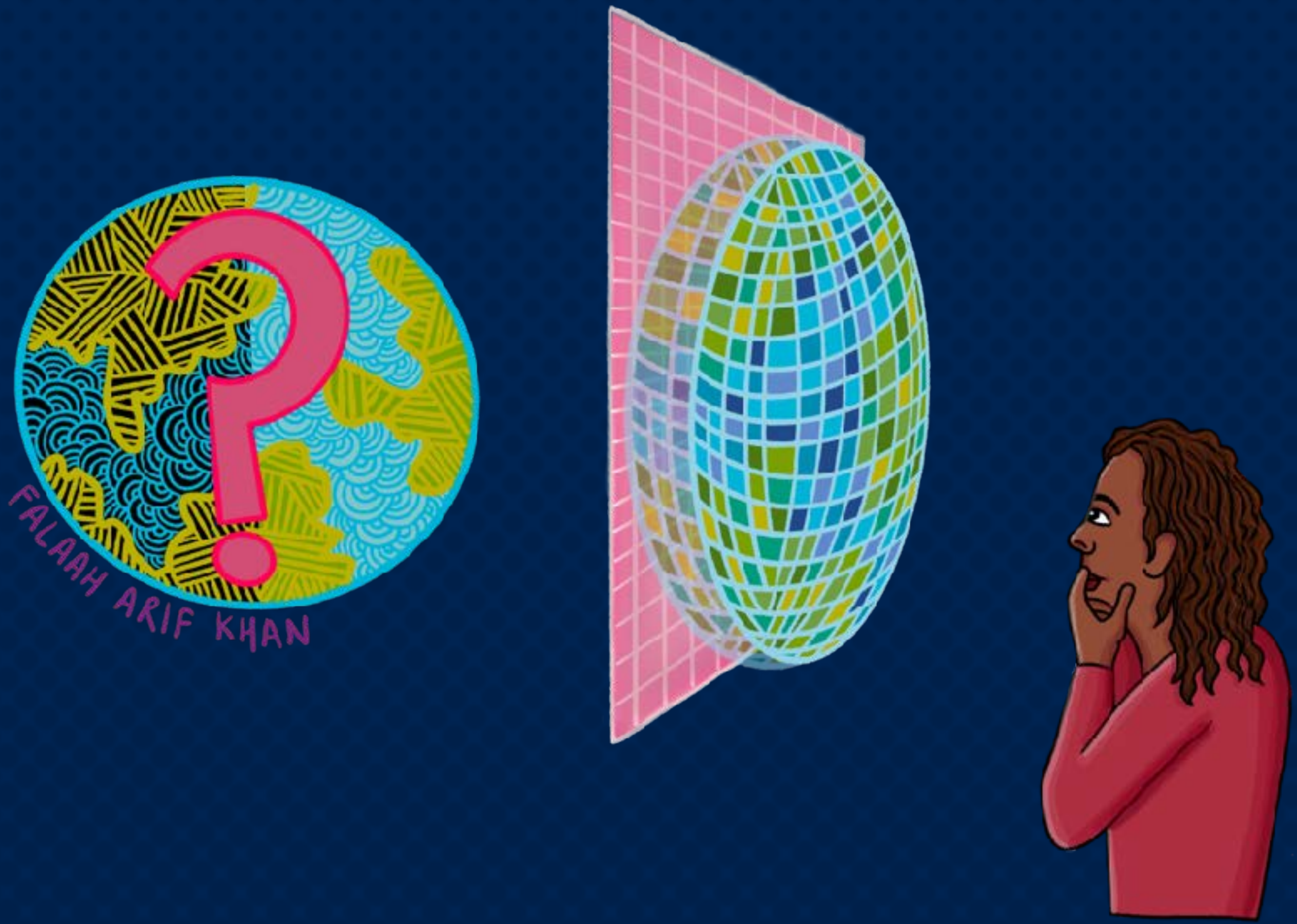


https://www.youtube.com/watch?v=wjd1G5bu_TY

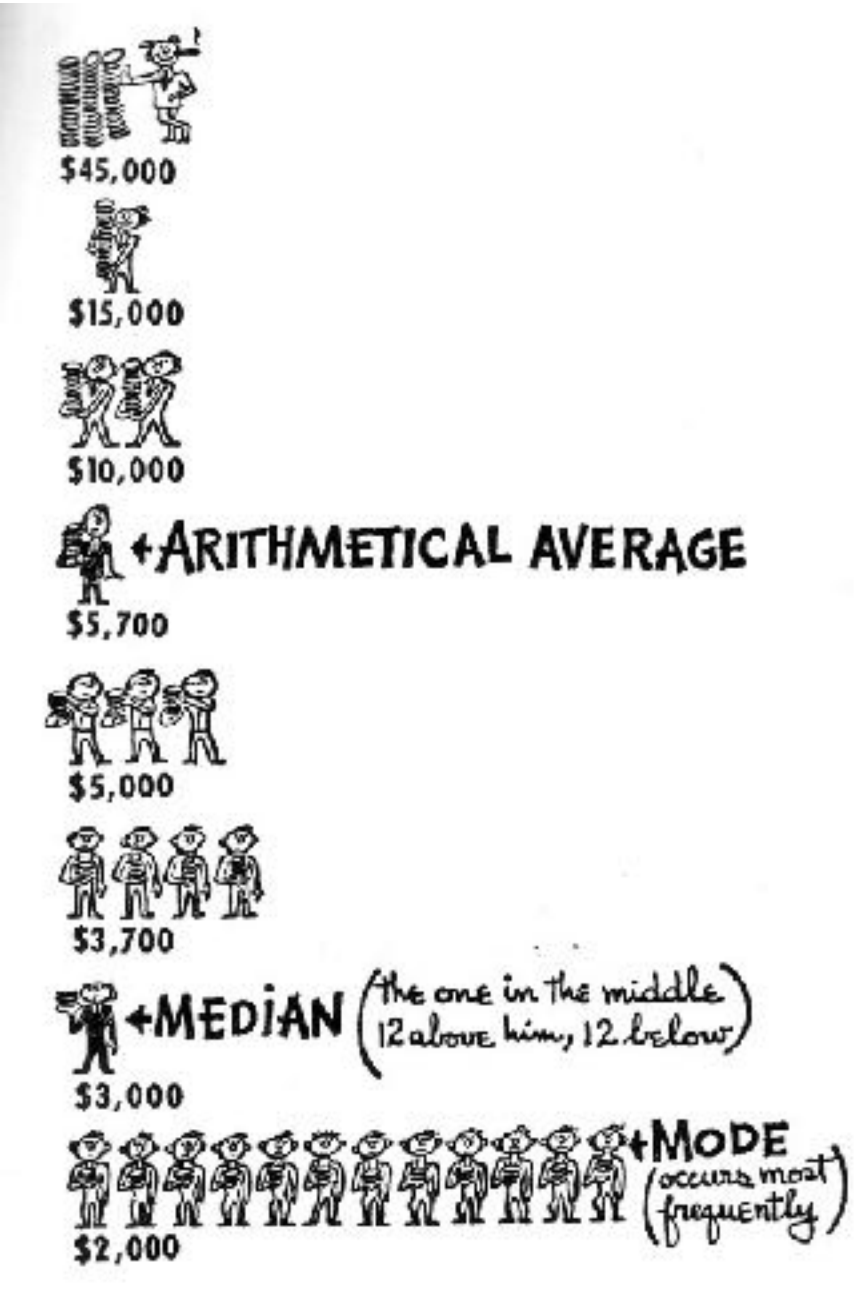
[Lundberg & Lee, 2017]

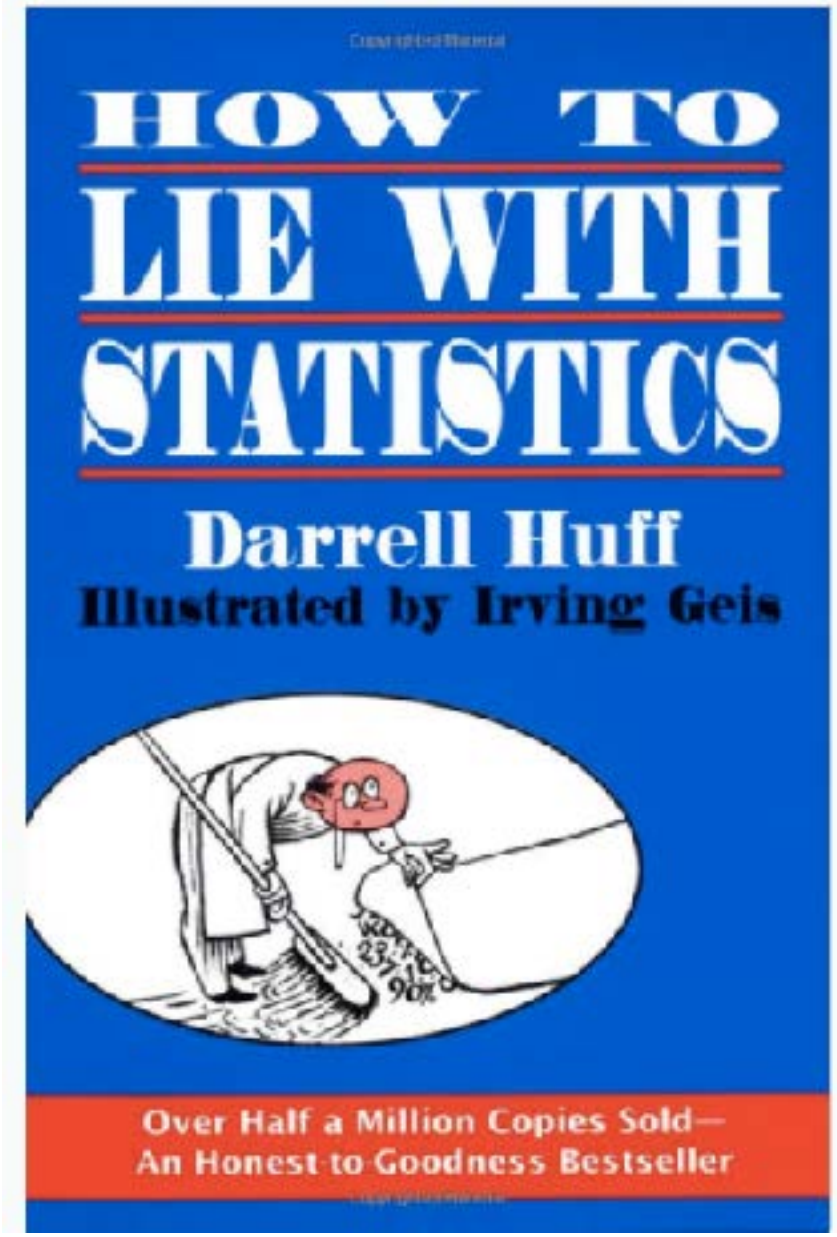
explaining ADS

Explaining the data



The well-chosen average





Explaining bias

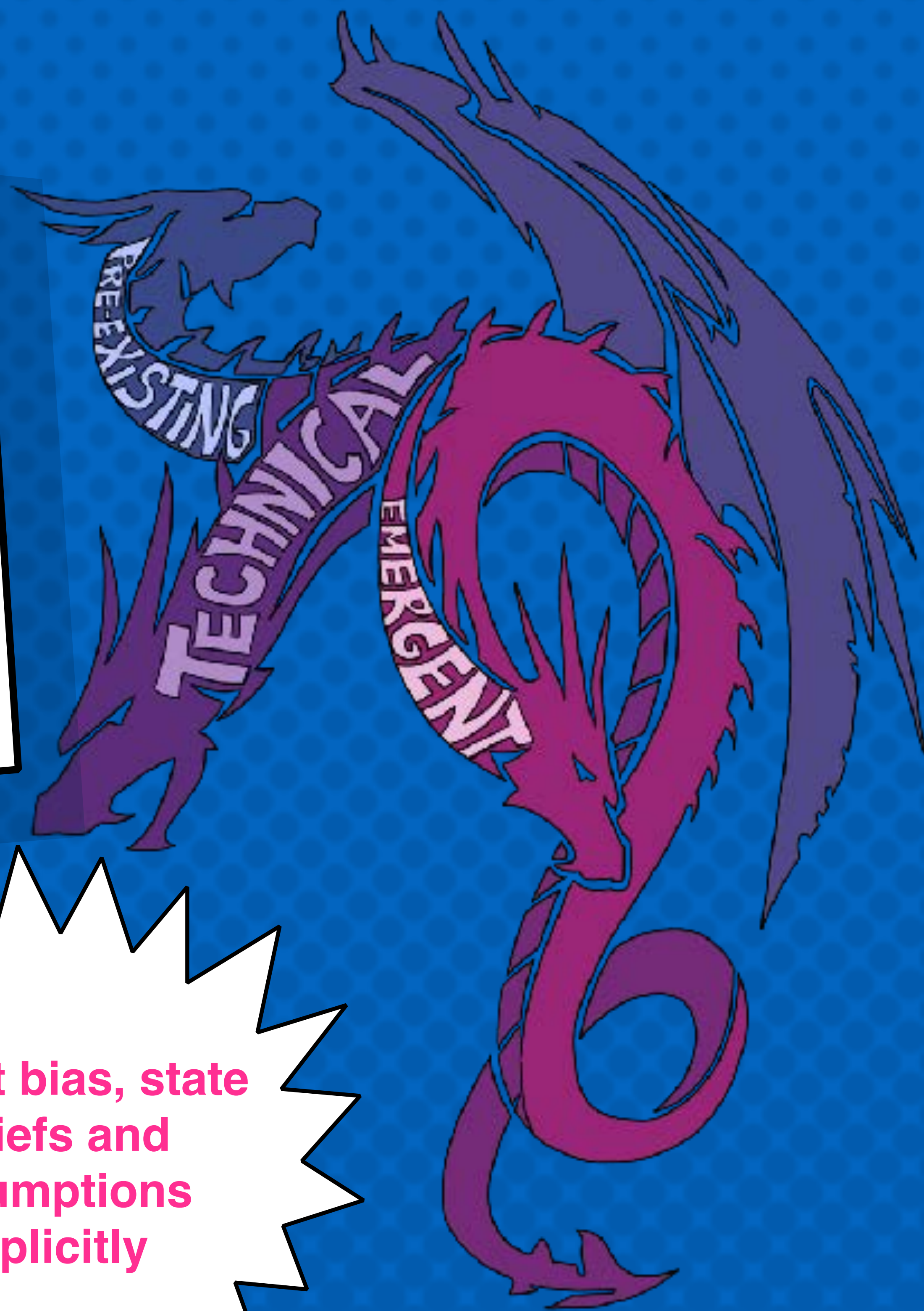
Pre-existing is independent of an algorithm and has origins in society

Technical is introduced or exacerbated by the technical properties of an ADS

Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

to fight bias, state beliefs and assumptions explicitly



Explaining the models

Explaining the decisions

everyone is at the table!

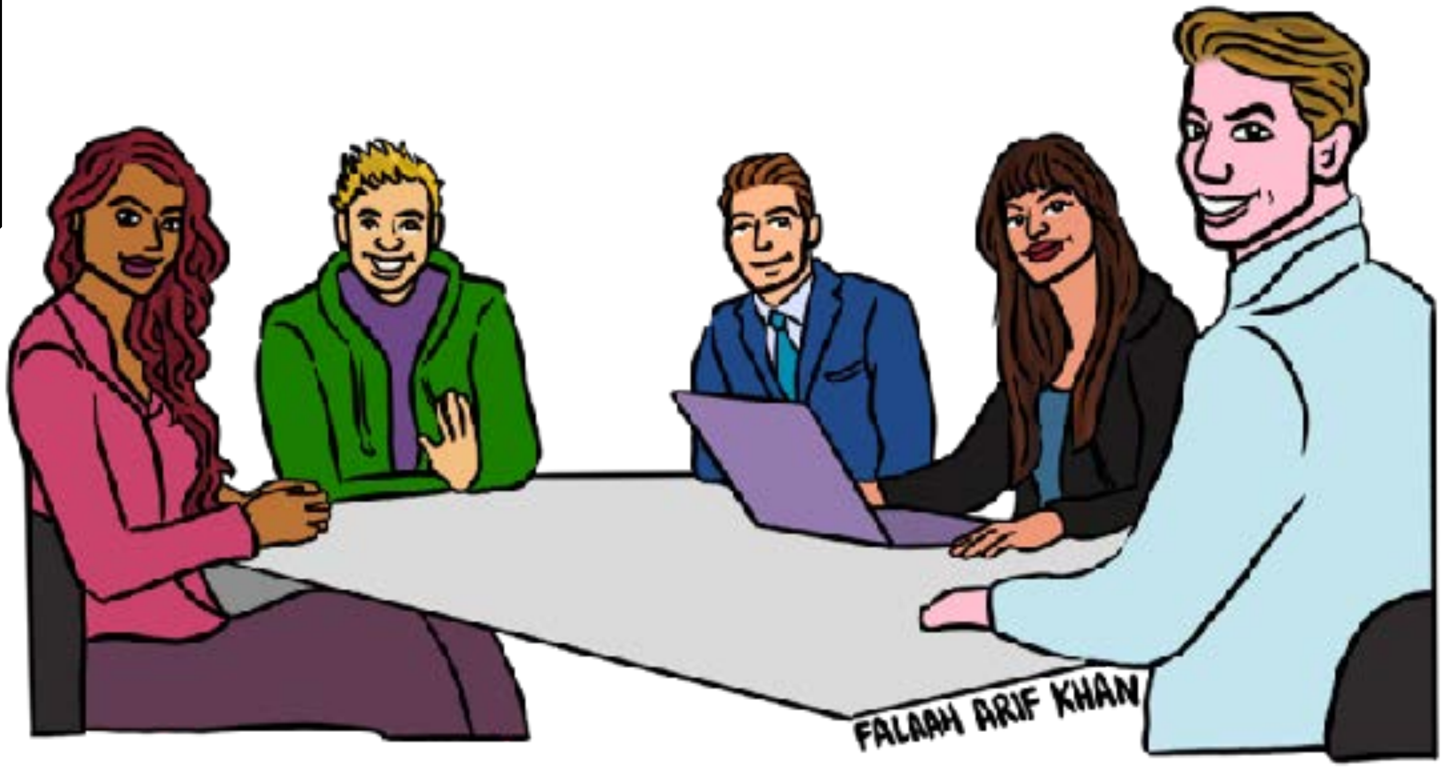
r/ai

Interpretability for different stakeholders

What are we explaining?

To **Whom** are we explaining?

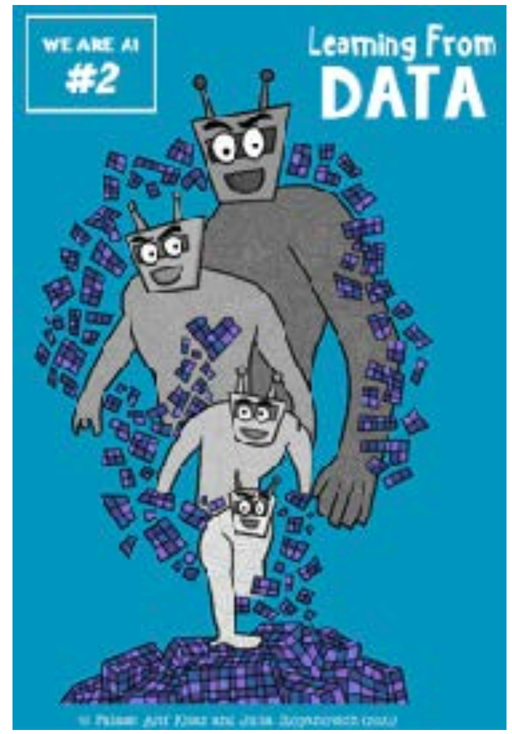
Why are we explaining?

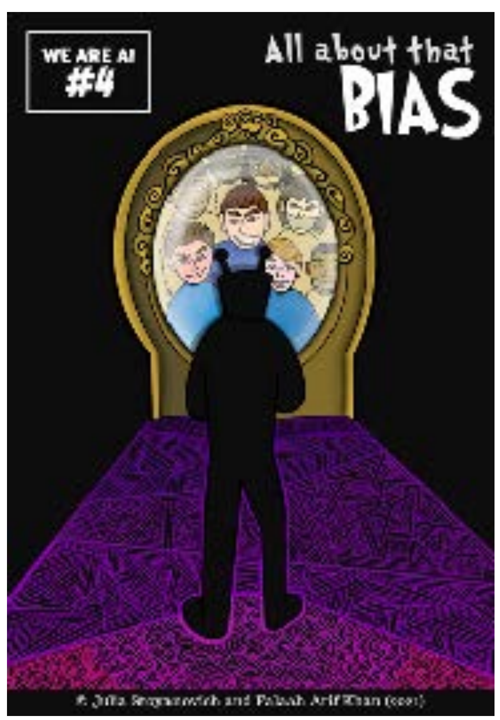


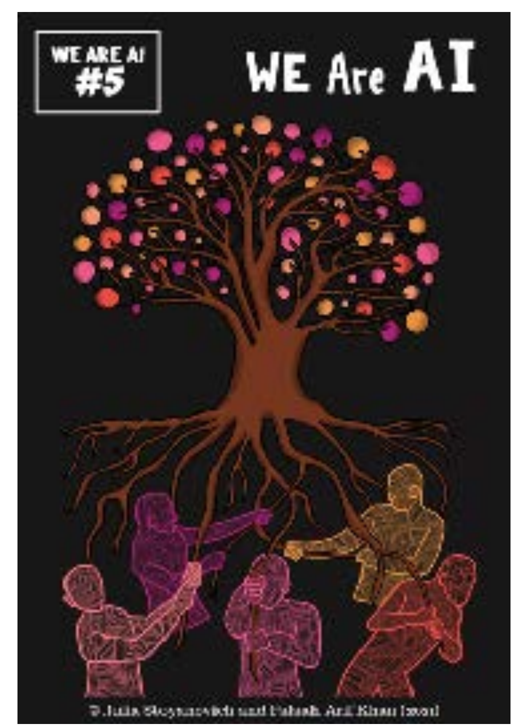
r/ai

dataresponsibly.github.io/we-are-ai

AI comics for the general public

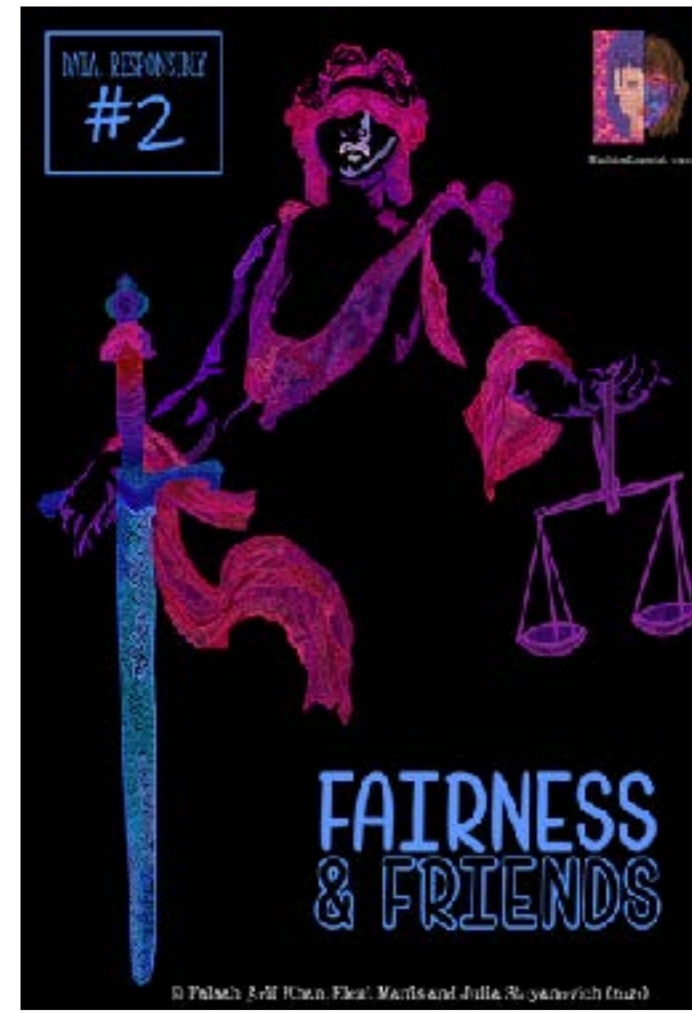






dataresponsibly.github.io/we-are-ai/comics

Scientific comics on AI



datarensponsibly.github.io/comics

Q&A
discussion

Thank you!

Julia Stoyanovich

Computer Science and Engineering
Center for Data Science
Visualization & Data Analytics Center
Center for Responsible AI
New York University