

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

We are in the midst of a global trend to regulate the use of algorithms, artificial intelligence (AI) and automated decision systems (ADS). As reported by the *One Hundred Year Study on Artificial Intelligence*: “AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy.” Major cities, states and national governments are establishing task forces, passing laws and issuing guidelines about responsible development and use of technology, often starting with its use in government itself, where there is, at least in theory, less friction between organizational goals and societal values.

In the United States, New York City has made a public commitment to opening the black box of the government’s use of technology: in 2018, an ADS task force was convened, the first of such in the nation, and charged with providing recommendations to New York City’s government agencies for how to become transparent and accountable in their use of ADS. In a 2019 report, the task force recommended using ADS where they are beneficial, reduce potential harm and promote fairness, equity, accountability and transparency. Can these principles become policy in the face of the apparent lack of trust in the government’s ability to manage AI in the interest of the public? We argue that overcoming this mistrust hinges on our ability to engage in substantive multi-stakeholder conversations around ADS, bringing with it the imperative of interpretability — allowing humans to understand and, if necessary, contest the computational process and its outcomes.

Remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm, and how we can empower humans to exercise agency — to adopt or challenge an algorithmic decision. Consider, for example, scoring and ranking — data-driven algorithms that prioritize entities such as individuals, schools, or products and services. These algorithms may be used to determine credit worthiness,

Box 1 | Research questions

- **What are we explaining?** Do people trust algorithms more or less than they would trust an individual making the same decisions? What are the perceived trade-offs between data disclosure and the privacy of individuals whose data are being analysed in the context of interpretability? Which potential sources of bias are most likely to trigger distrust in algorithms? What is the relationship between the perceptions about a dataset’s fitness for use and the overall trust in the algorithmic system?
- **To whom are we explaining and why?** How do group identities shape perceptions about algorithms? Do people lose trust in algorithmic decisions when they learn that outcomes produce disparities? Is this only the case when these disparities harm their in-group? Are people more likely to see algorithms as biased if members of their own group were not involved in

algorithm construction? What kinds of transparency will promote trust, and when will transparency decrease trust? Do people trust the moral cognition embedded within algorithms? Does this apply to some domains (for example, pragmatic decisions, such as clothes shopping) more than others (for example, moral domains, such as criminal sentencing)? Are certain decisions taboo to delegate to algorithms (for example, religious advice)?

- **Are explanations effective?** Do people understand the label? What kinds of explanations allow individuals to exercise agency: make informed decisions, modify their behaviour in light of the information, or challenge the results of the algorithmic process? Does the nutrition label help create trust? Can the creation of nutrition labels lead programmers to alter the algorithm?

and desirability for college admissions or employment. Scoring and ranking are as ubiquitous and powerful as they are opaque. Despite their importance, members of the public often know little about why one person is ranked higher than another by a résumé screening or a credit scoring tool, how the ranking process is designed and whether its results can be trusted.

As an interdisciplinary team of scientists in computer science and social psychology, we propose a framework that forms connections between interpretability and trust, and develops actionable explanations for a diversity of stakeholders, recognizing their unique perspectives and needs. We focus on three questions (Box 1) about making machines interpretable: (1) what are we explaining, (2) to whom are we explaining and for what purpose, and (3) how do we know that an explanation is effective? By asking — and charting the path towards answering — these questions, we can promote greater trust in algorithms,

and improve fairness and efficiency of algorithm-assisted decision making.

What are we explaining?

Existing legal and regulatory frameworks, such as the USA Fair Credit Reporting Act and the EU’s General Data Protection Regulation, differentiate between two kinds of explanations. The first concerns the outcome: what are the results for an individual, a demographic group or the population as a whole? The second concerns the logic behind the decision-making process: what features help an individual or group get a higher score, or, more generally, what are the rules by which the score is computed? Selbst and Barocas argue for an additional kind of an explanation that considers the justification: why are the rules what they are? Much has been written about explaining outcomes, so we focus on explaining and justifying the process.

Procedural justice aims to ensure that algorithms are perceived as fair and

DOI:10.1145/3282486

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

BY DANIEL S. WELD AND CACAN BANSAL

The Challenge of Crafting Intelligible Intelligence

ARTIFICIAL INTELLIGENCE (AI) SYSTEMS have reached or exceeded human performance for many circumscribed tasks. As a result, they are increasingly deployed in mission-critical roles, such as credit scoring, predicting if a bail candidate will commit another crime, selecting the news we read on social networks, and self-driving cars. Unlike other mission-critical software, extraordinarily complex AI systems are difficult to test: AI decisions are context specific and often based on thousands or millions of factors. Typically, AI behaviors are generated by searching vast action spaces or learned by the opaque optimization of mammoth neural networks operating over prodigious amounts of training data. Almost by definition, no clear-cut method can accomplish these AI tasks.

Unfortunately, much AI-produced behavior is alien, that is, it can fall in unexpected ways. This lesson is

most clearly seen in the performance of the latest deep neural network image analysis systems. While their accuracy at object recognition on naturally occurring pictures is extraordinary, imperceptible changes to input images can lead to erratic predictions, as shown in Figure 1. Why are these recognition systems so brittle, making different predictions for apparently identical images? Unintelligible behavior is not limited to machine learning: many AI programs, such as automated planning algorithms, perform search-based local search and inference whose complexity exceeds human abilities to verify. While some search and planning algorithms are possible complete and optimal, intelligibility is still important, because the underlying primitives (for example, search operators or action descriptions) are usually approximations. The only trust a proof that is based on (possibly) incorrect premises.

Despite intelligibility’s apparent value, it remains remarkably difficult to specify what makes a system “intelligible.” We discuss desiderata for intelligible behavior later in this article. In brief, we seek AI systems where it is clear what factors caused the system’s action, allowing the users to predict how changes to the situation would have led to alternative behaviors, and permits effective contest of

Key insights

- There are important technical and social reasons to prefer inherently intelligible AI models (such as linear models or QM) over deep neural networks. Furthermore, intelligible models often have comparable accuracy.
- When an AI system is based on an inscrutable model, it may explain its decisions by making those decisions into a discrete, explanatory world using techniques such as local approximations and vocabulary transformations.
- Results from psychology show that explanation is a process, but thought of as a representation between explainer and listener. We advocate for users to be able to interact with explanation systems that can respond to a wide range of follow-up questions.

DOI:10.1038/s42256-020-0171-8

DOI:10.1145/3282486

Automated Decision Systems (ADS)

Automated Decision Systems (ADS)

process data about people

help make consequential decisions

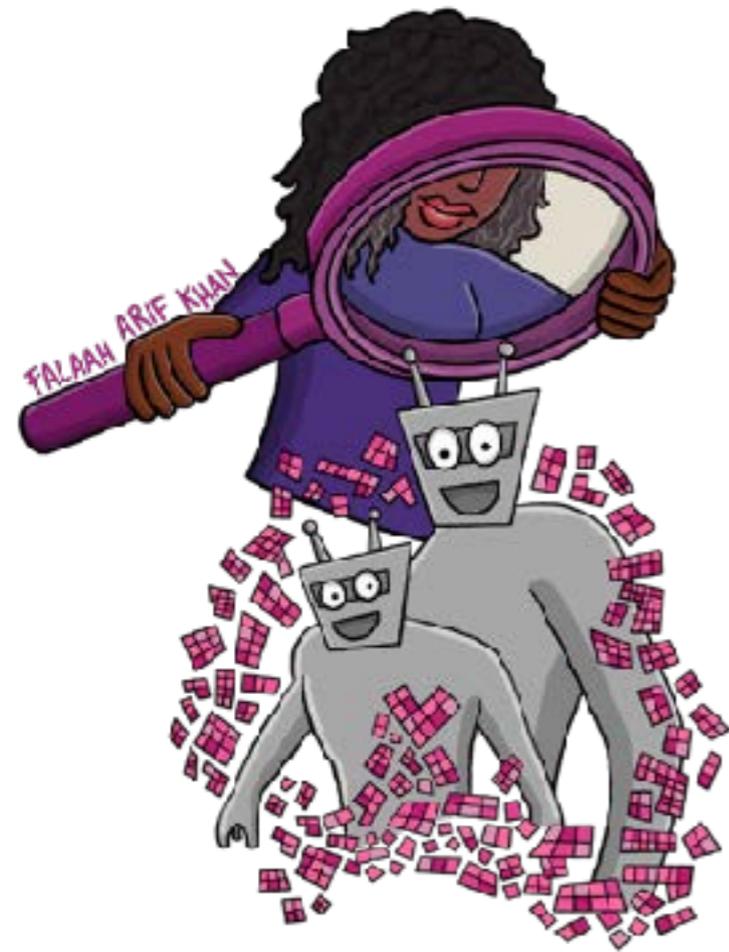
combine human & automated decision making

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?



examples

ADS in medical imaging

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?



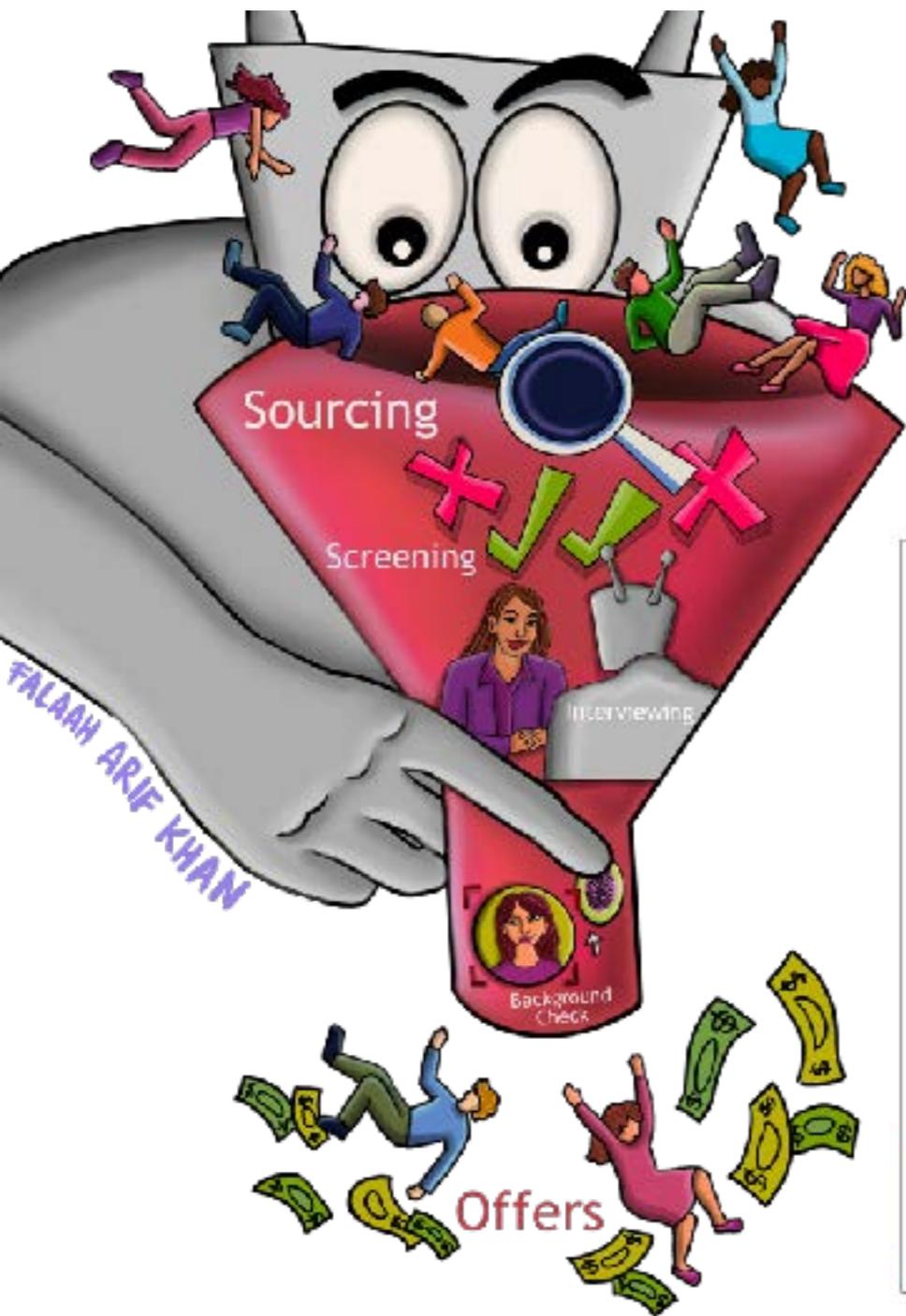
What is fastMRI?

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.

ADS in hiring



What are we explaining?
To **Whom** are we explaining?
Why are we explaining?

ACCOUNTANT

Acme Partners

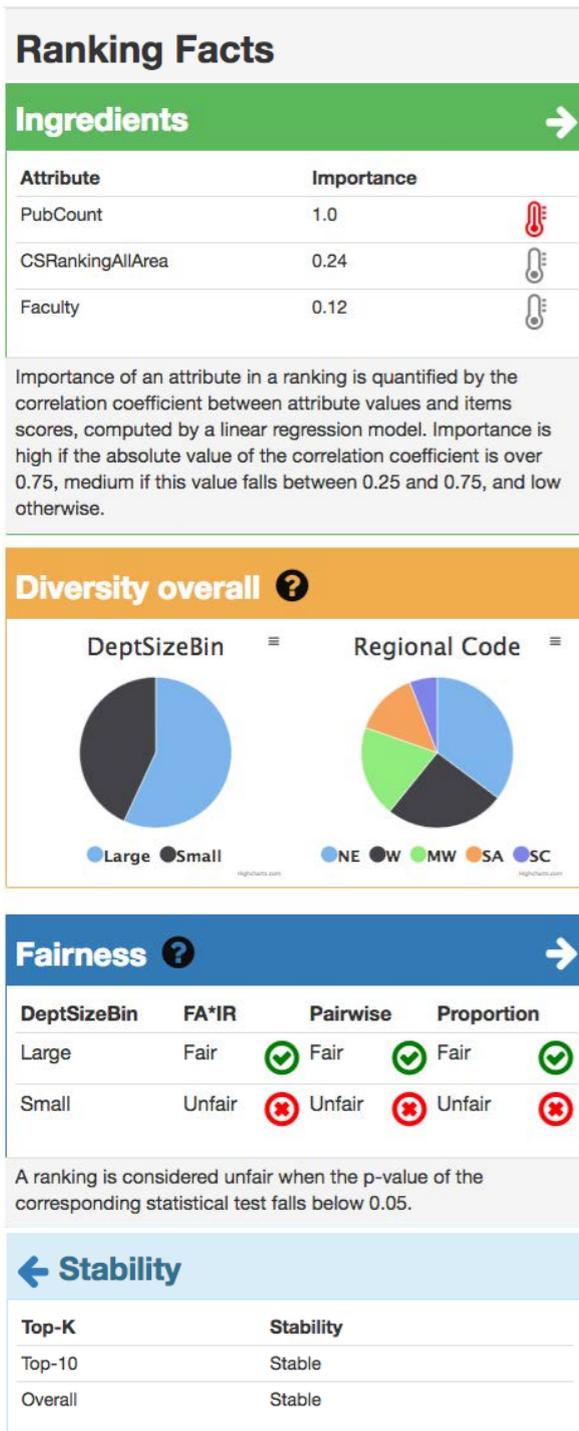
Qualifications: BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications

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.

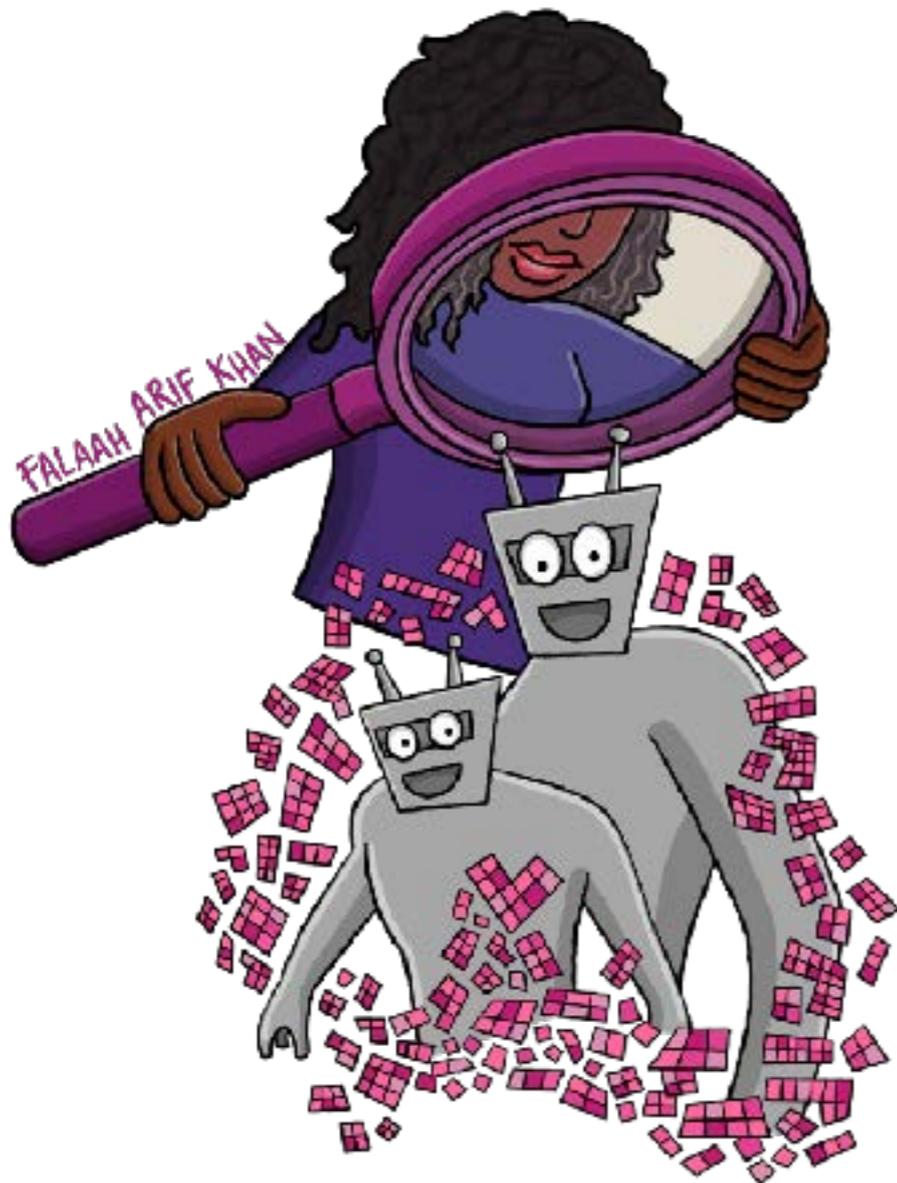
Nutritional labels for ADS?



comprehensible: short, simple, clear
consultative: provide actionable info
comparable: implying a standard
computable: incrementally constructed

explaining black box
models

What are we explaining?



How does a system work?

How **well** does a system work?

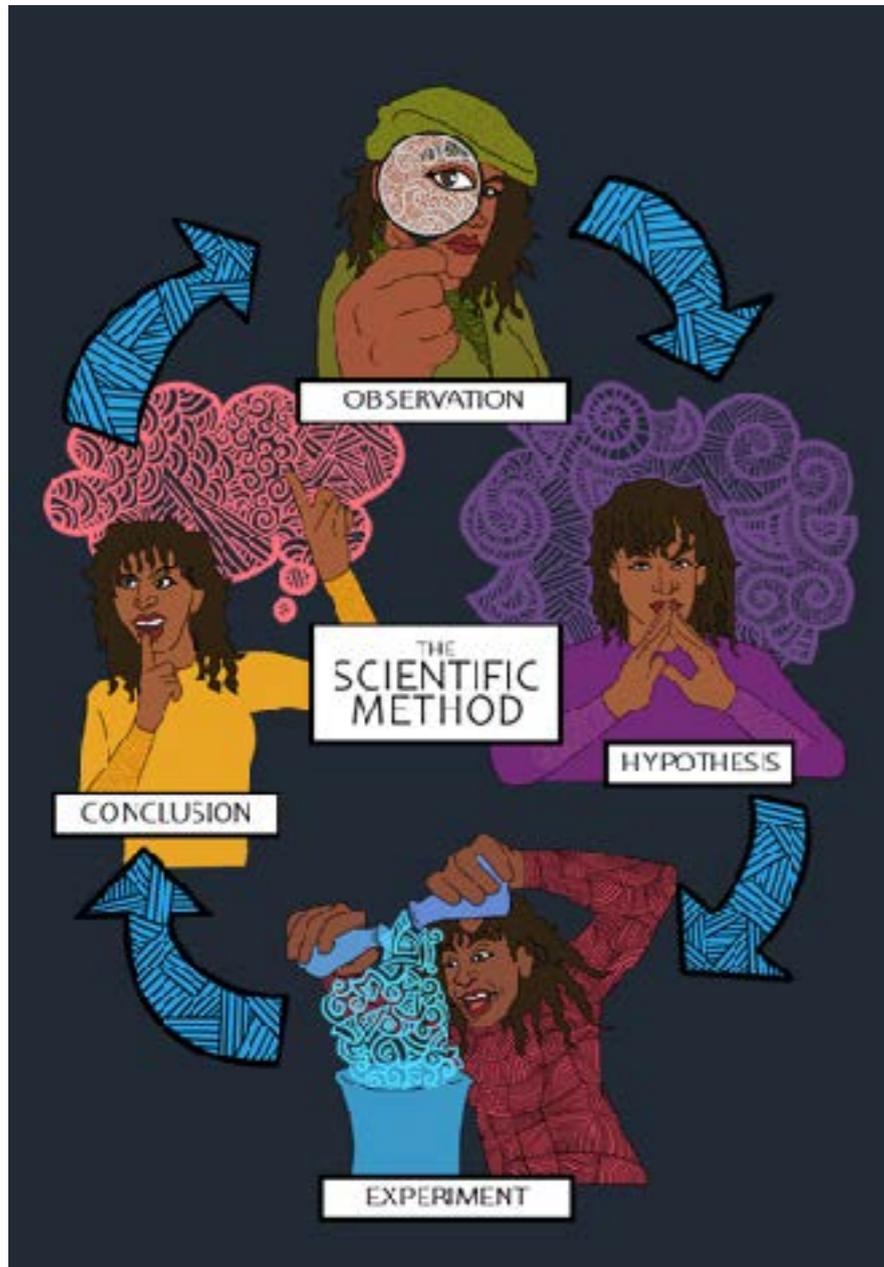
What does a system do?

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 / ...

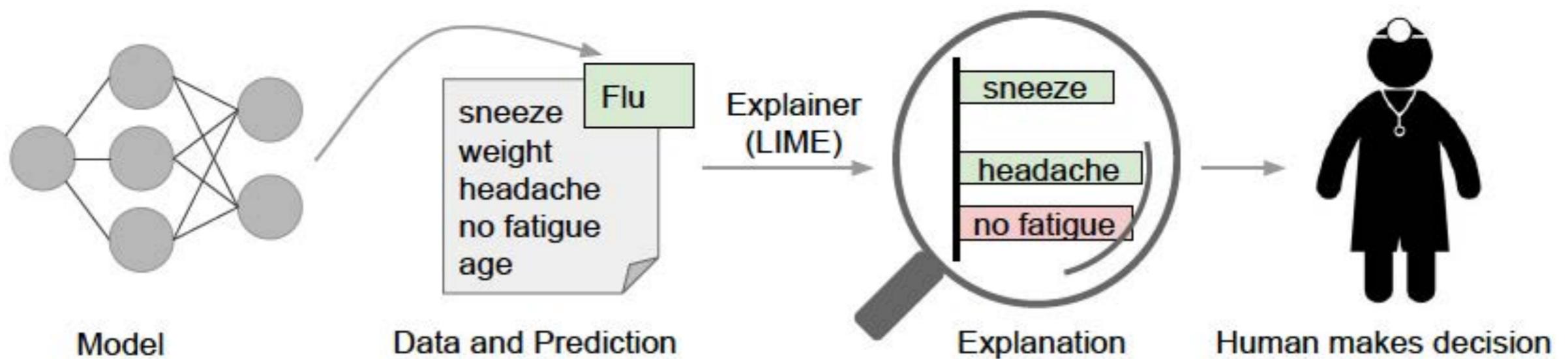
Accuracy for whom: over-all or in sub-populations?

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



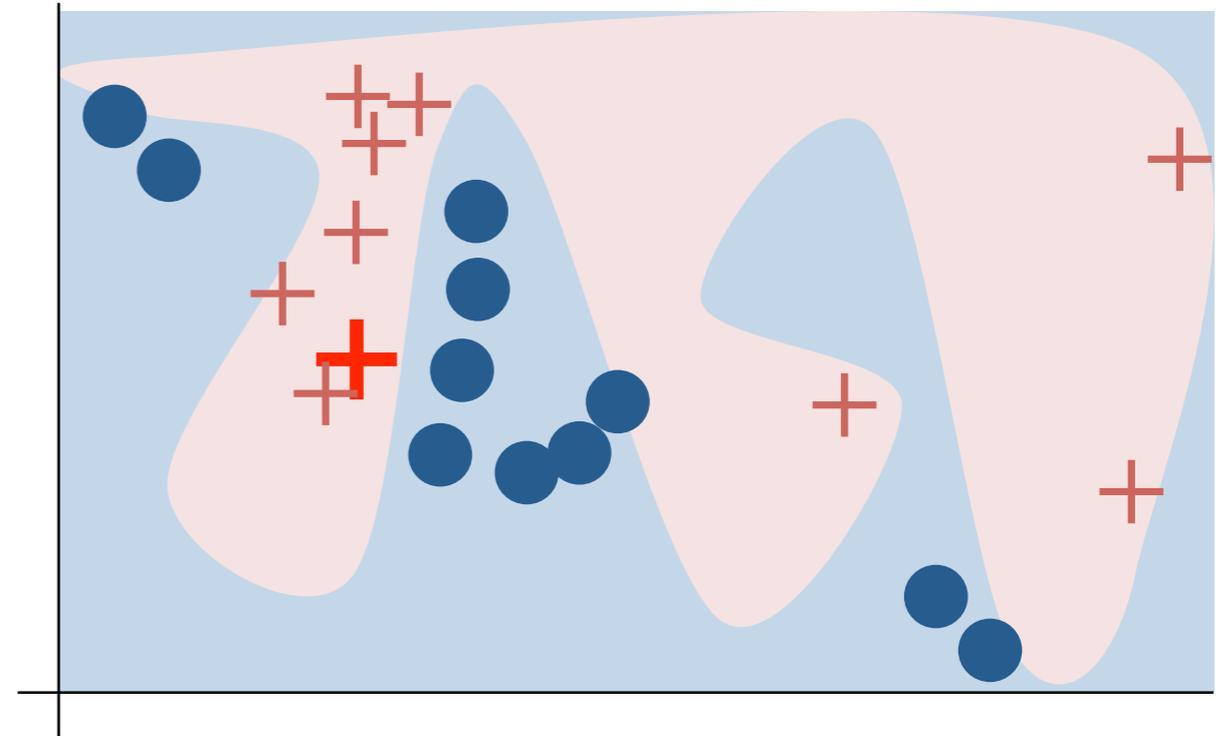
what if patient id appears in green in the list?

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

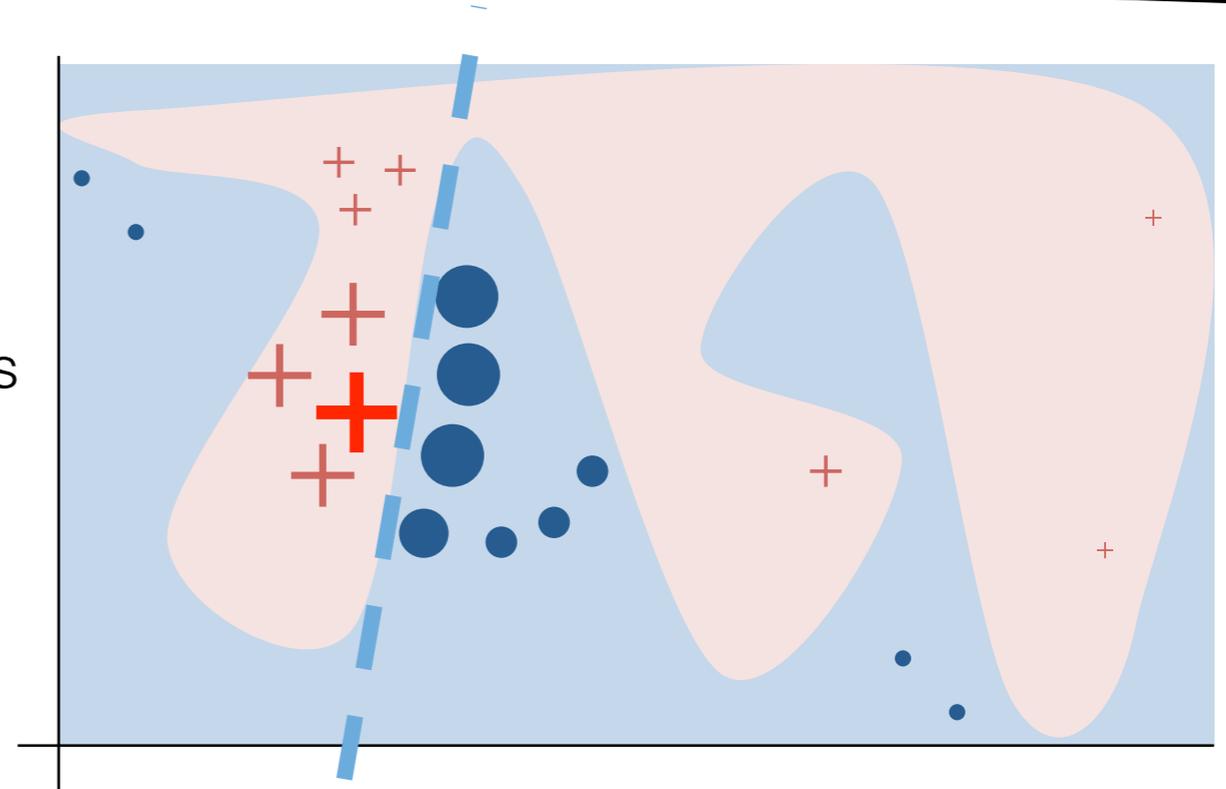
1. sample points around +
2. use original model to assign class labels
3. weigh points according to distance from +
4. learn interpretable model according to samples

Key ideas

interpretable features

interpretable models

locally faithful explanations



When accuracy is not enough

Train a neural network to predict **wolf** v. **husky**



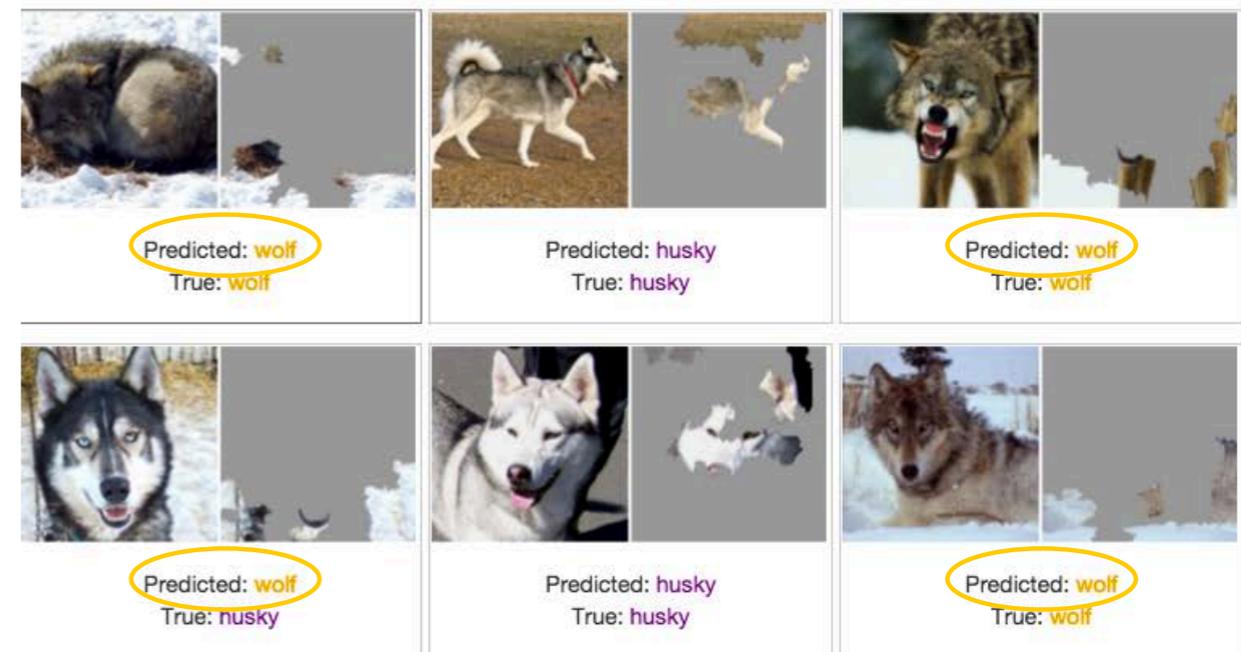
Only 1 mistake!!!

Do you trust this model?

How does it distinguish between huskies and wolves?



Explanations for neural network prediction



We've built a great snow detector... ☹️

When accuracy is not enough

Explaining Google's Inception NN

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



$$P(\text{Electric guitar}) = 0.32$$



Electric guitar - incorrect but
reasonable, similar fretboard

$$P(\text{Acoustic guitar}) = 0.24$$



Acoustic guitar

$$P(\text{Labrador}) = 0.21$$

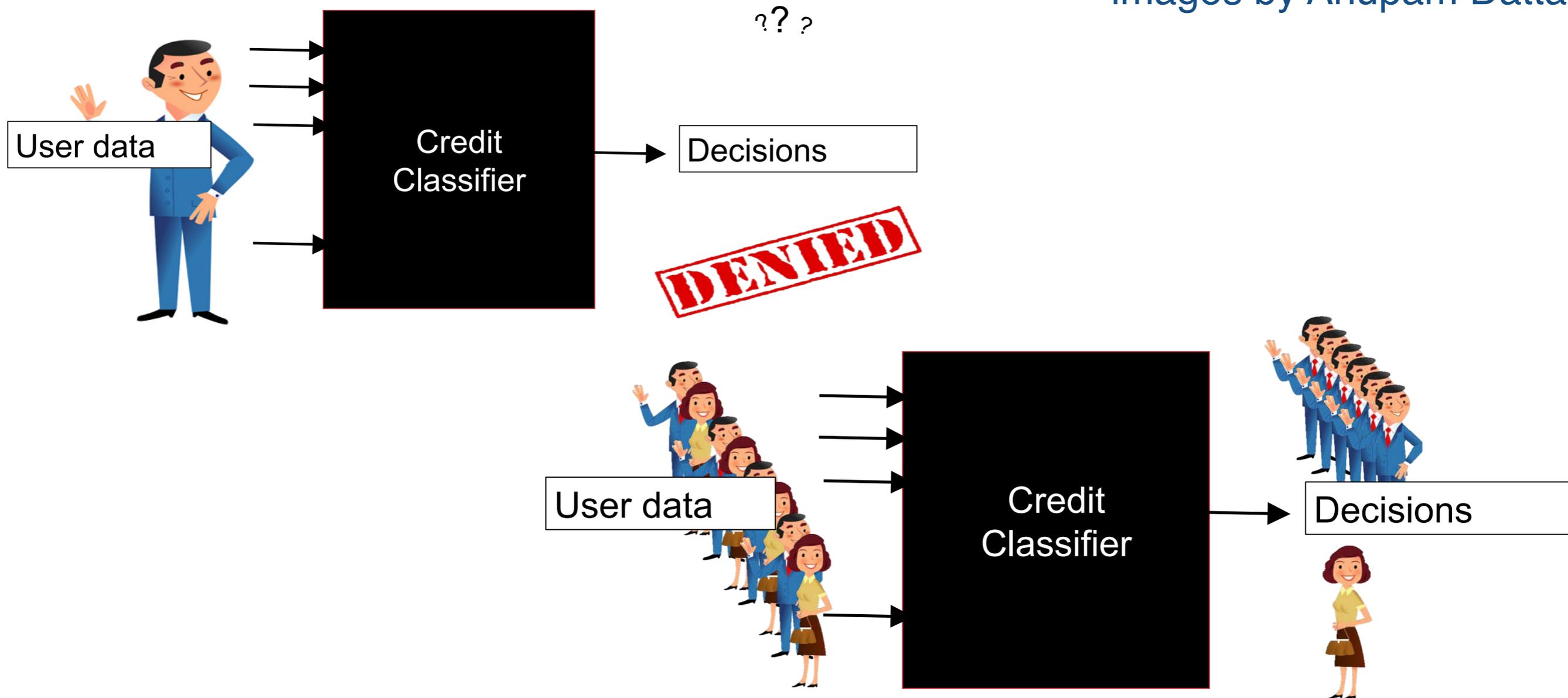


Labrador

quick discussion

Auditing black-box models

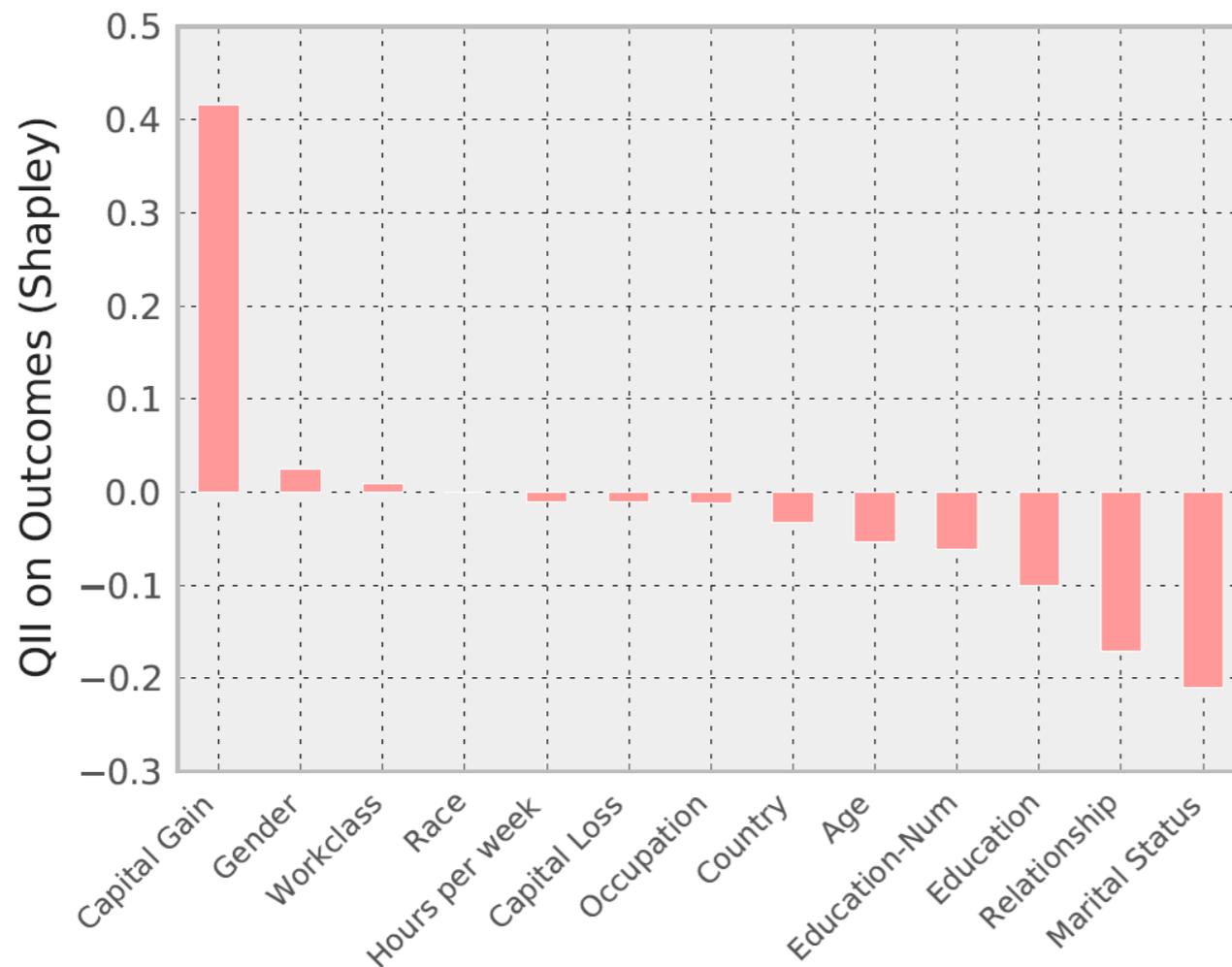
images by Anupam Datta



Transparency report: Mr. X

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

images by Anupam Datta



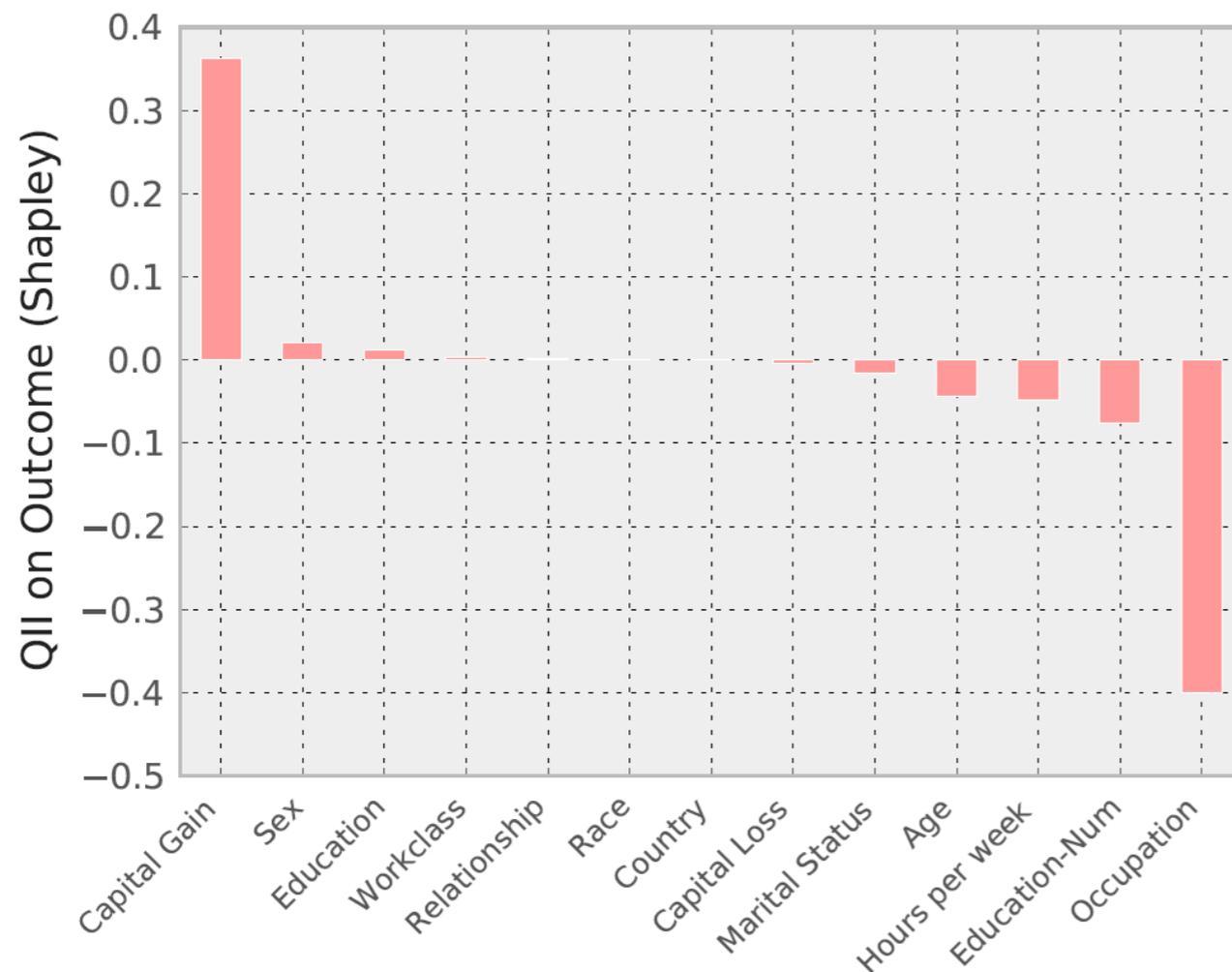
Age	23
Workclass	Private
Education	11 th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	\$14344
Capital loss	\$0
Work hours per week	40
Country	Vietnam

income

Transparency report: Mr. Y

images by Anupam Datta

Explanations for superficially similar individuals can be different



DENIED

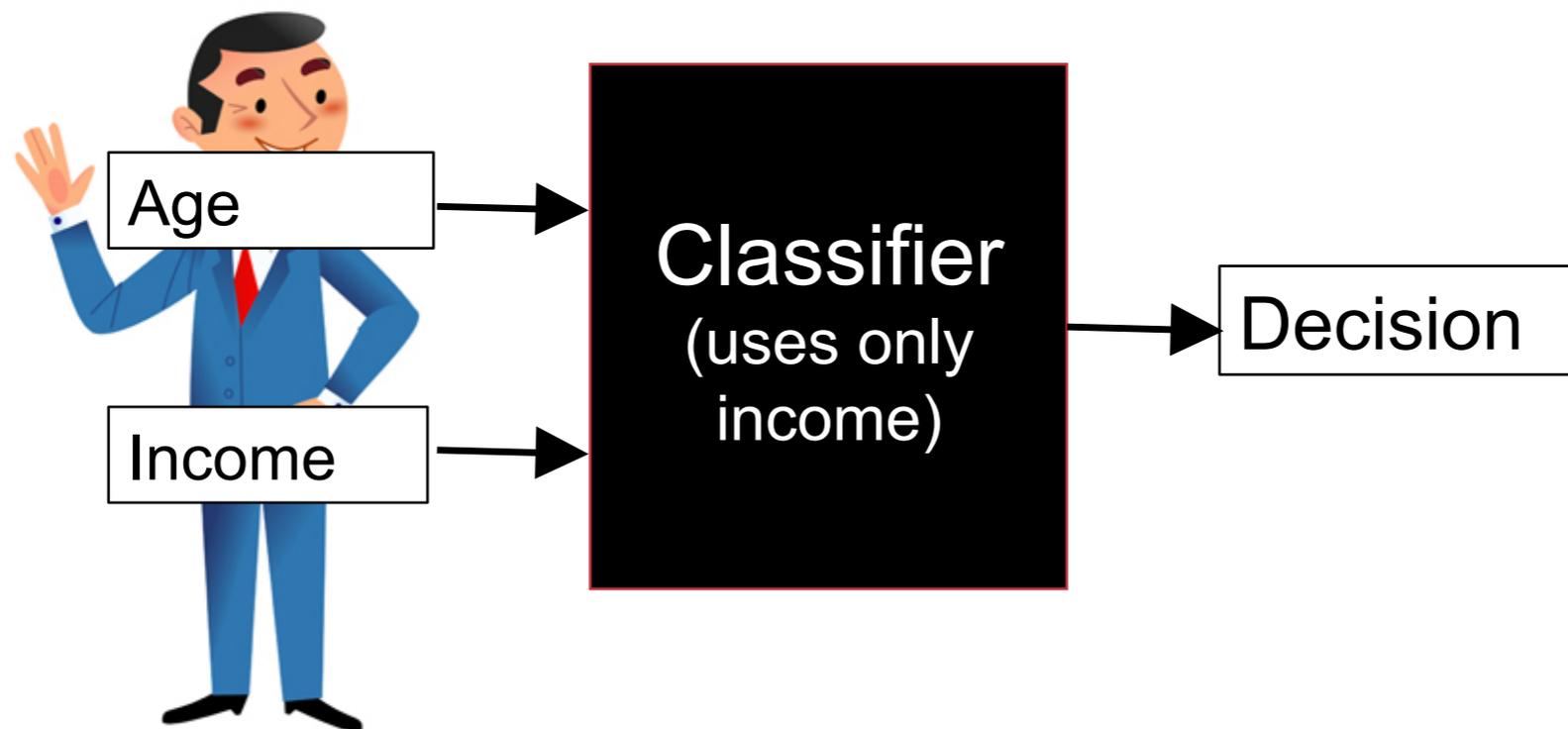
Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	\$41310
Capital loss	\$0
Work hours per week	24
Country	Mexico

income

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

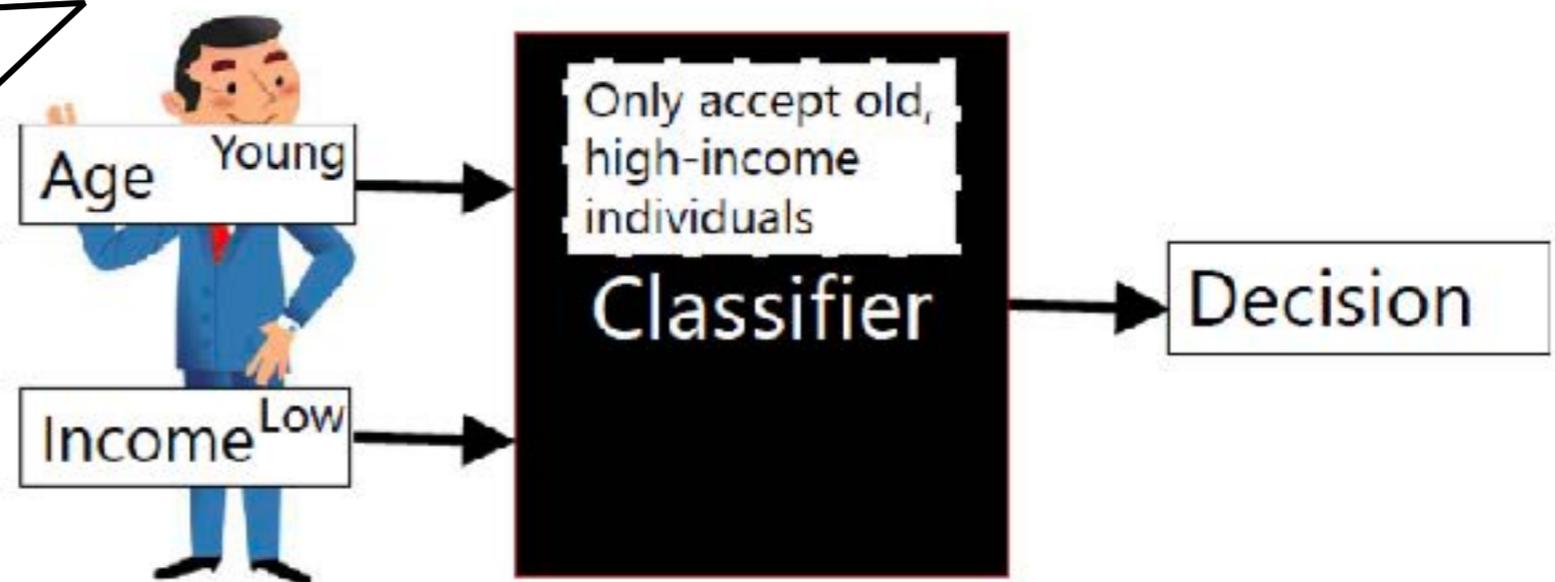
images by Anupam Datta

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**

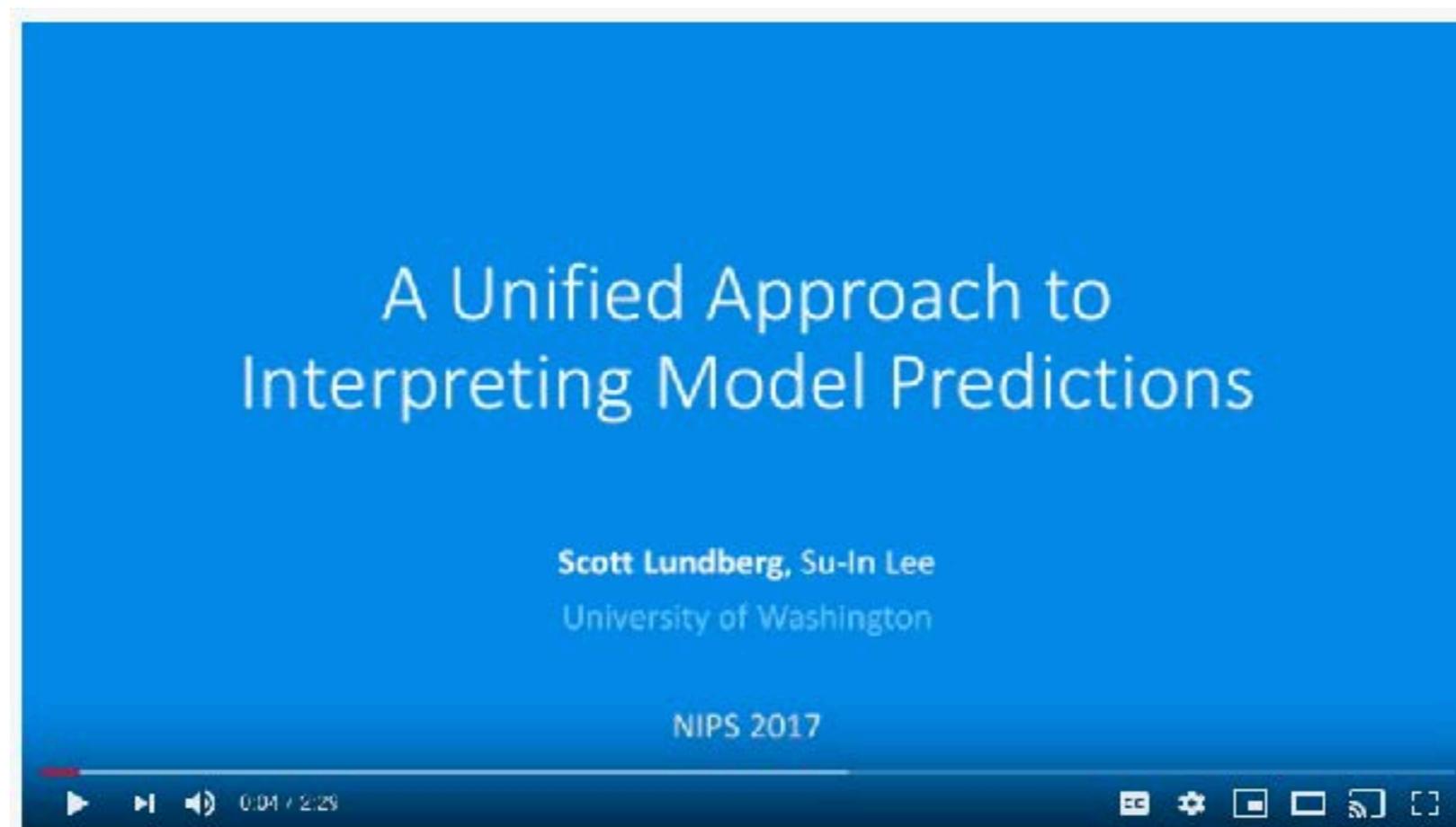
aggregate feature importance using its **Shapley value**



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**



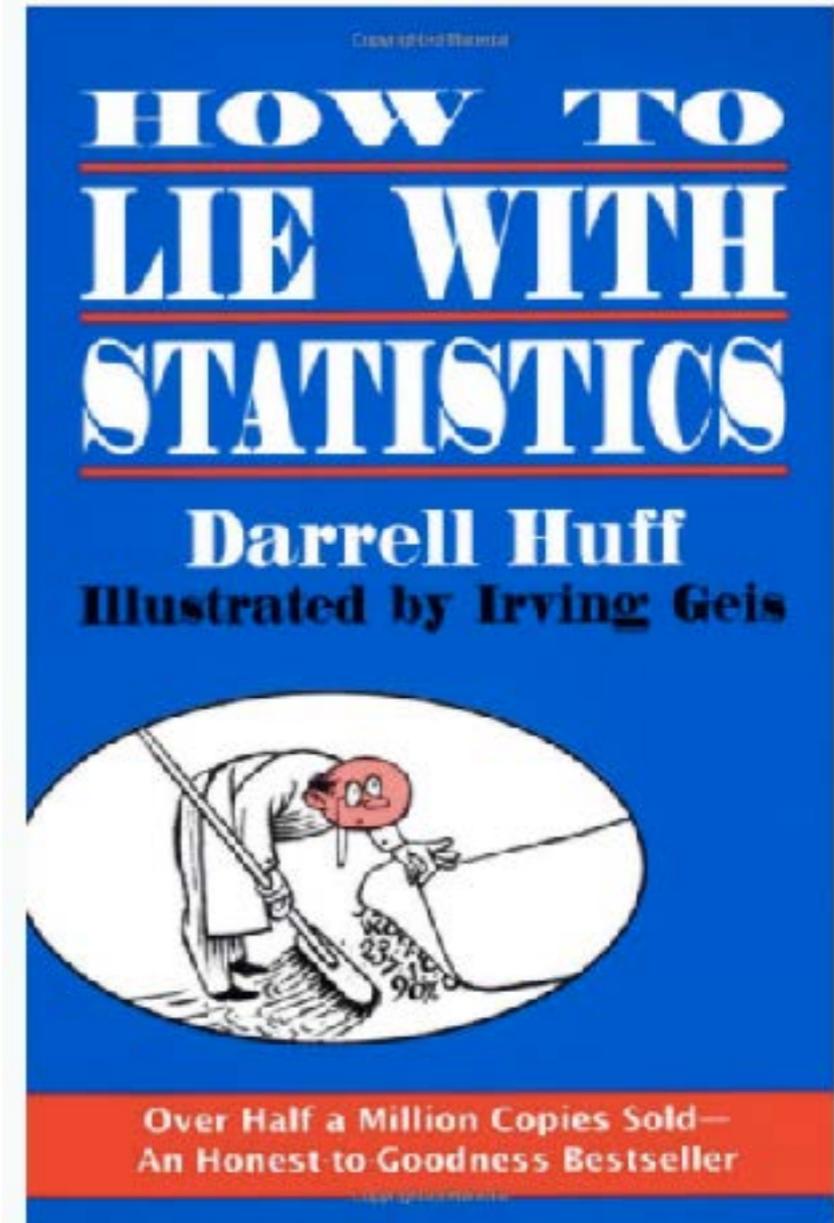
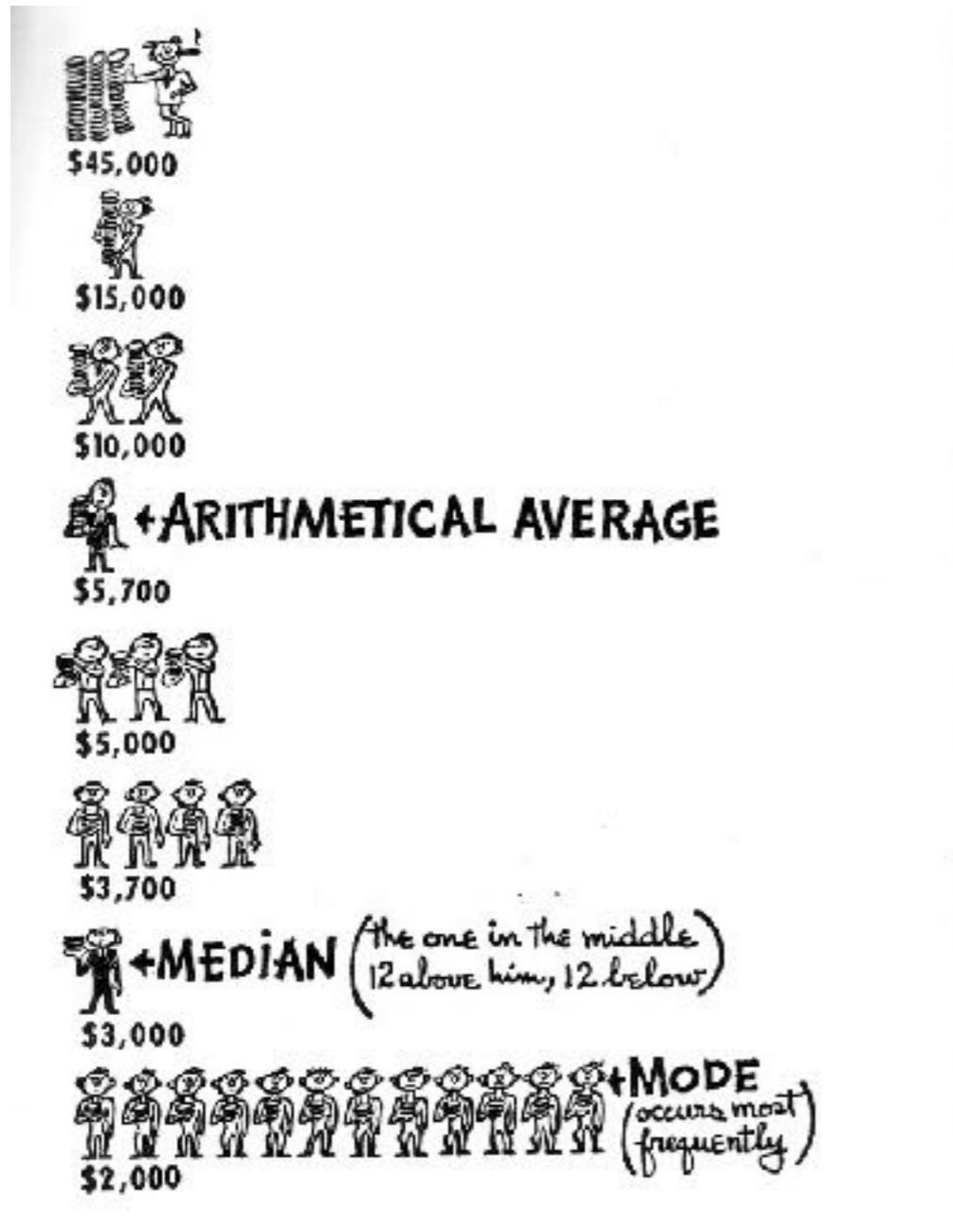
https://www.youtube.com/watch?v=wjd1G5bu_TY

explaining ADS

Explaining the data



The well-chosen average

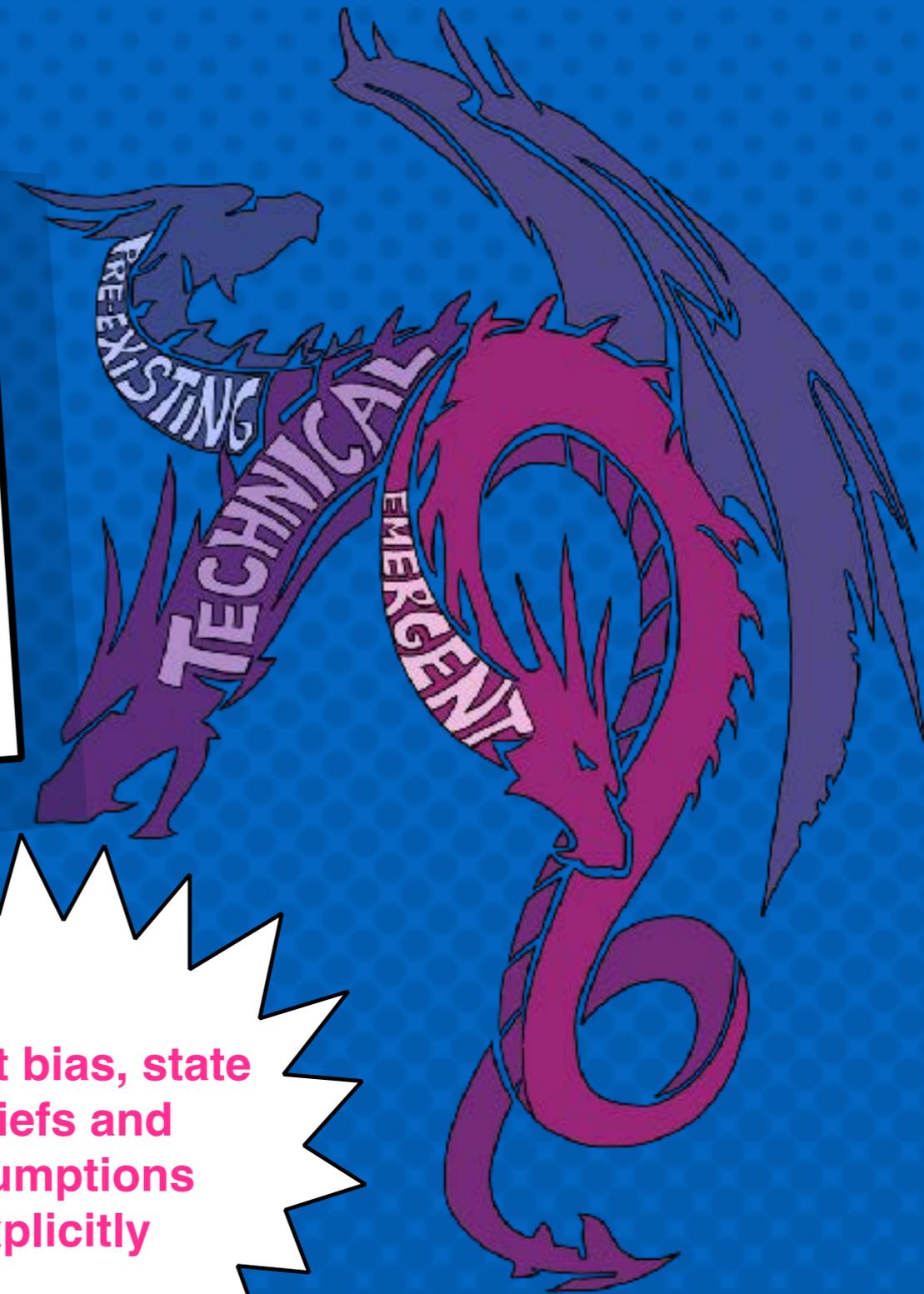


Explaining bias

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



FALAH ASIF KHAN

FALAH ASIF KHAN

to fight bias, state beliefs and assumptions explicitly

[Friedman & Nissenbaum (1996)]

Explaining the models



Explaining the decisions



*everyone is at
the table!*

Interpretability for different stakeholders

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?



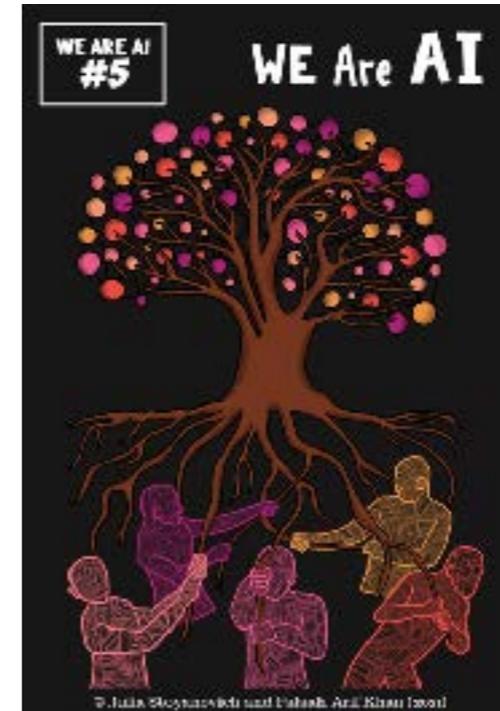
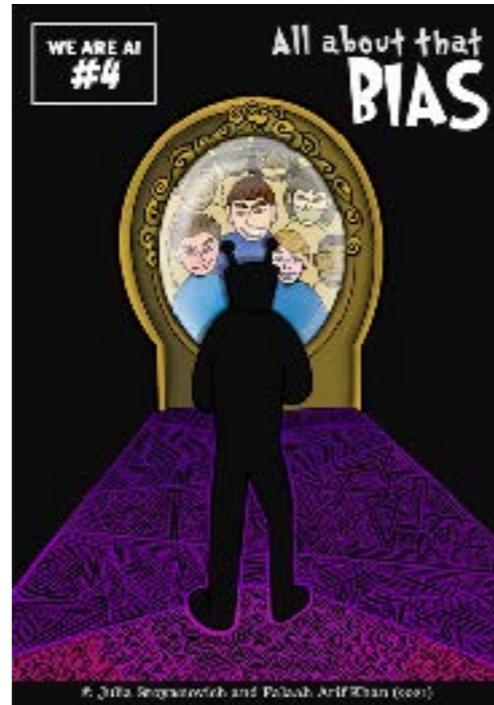
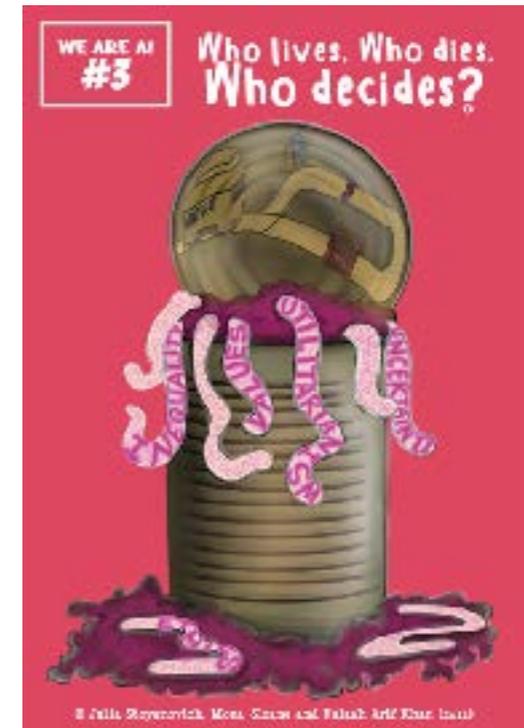
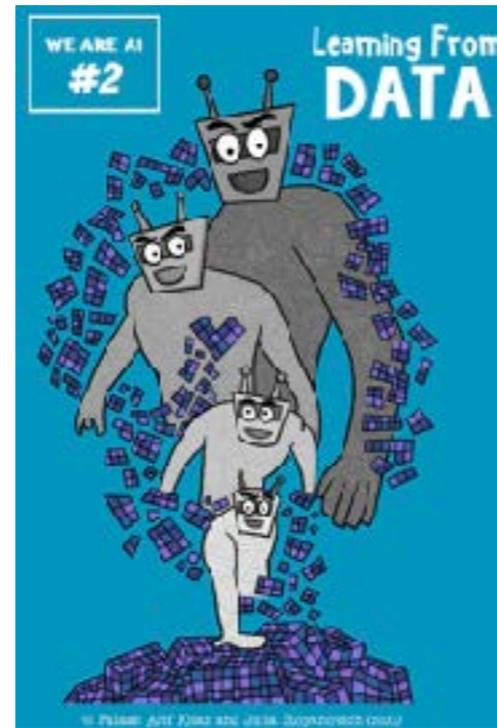
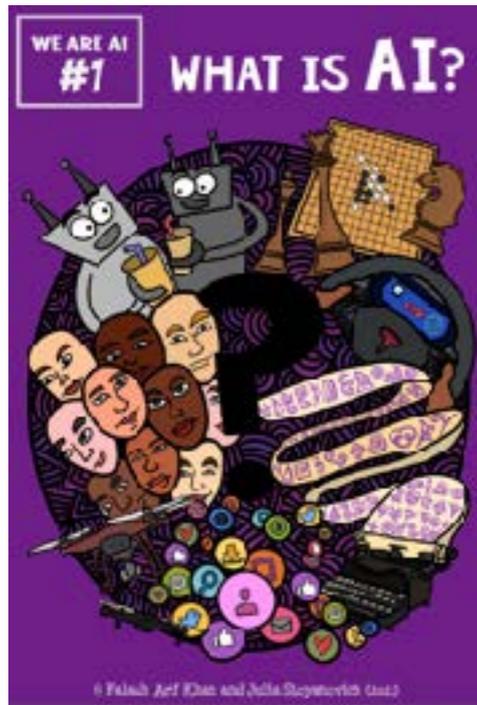
We are AI

taking control of technology
powered by NYU Center for Responsible AI

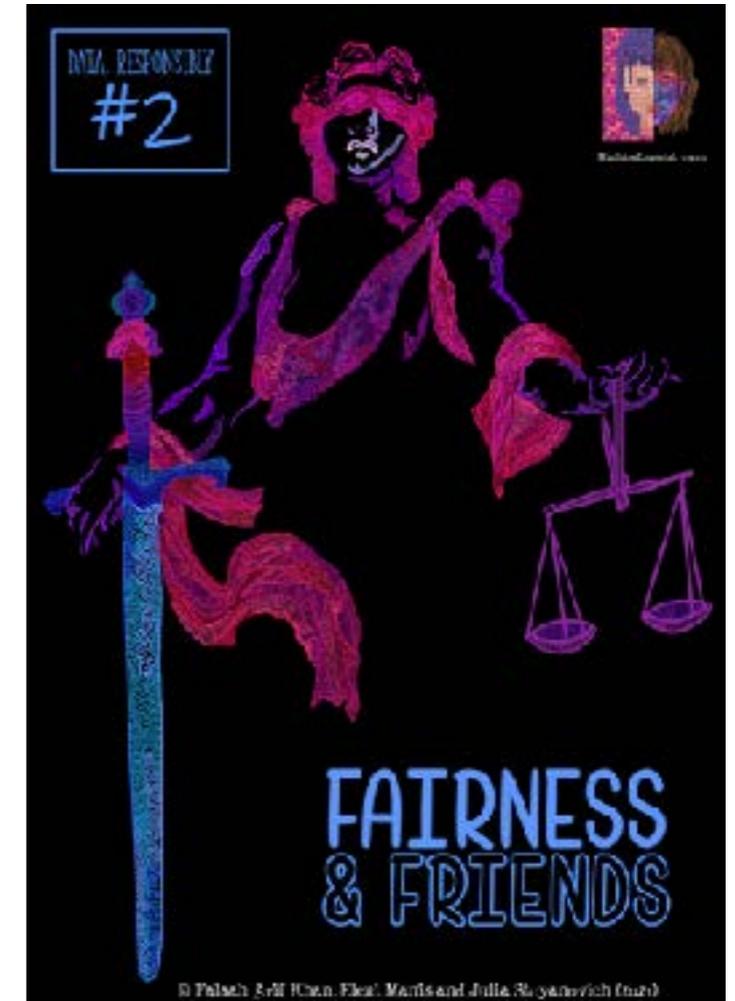
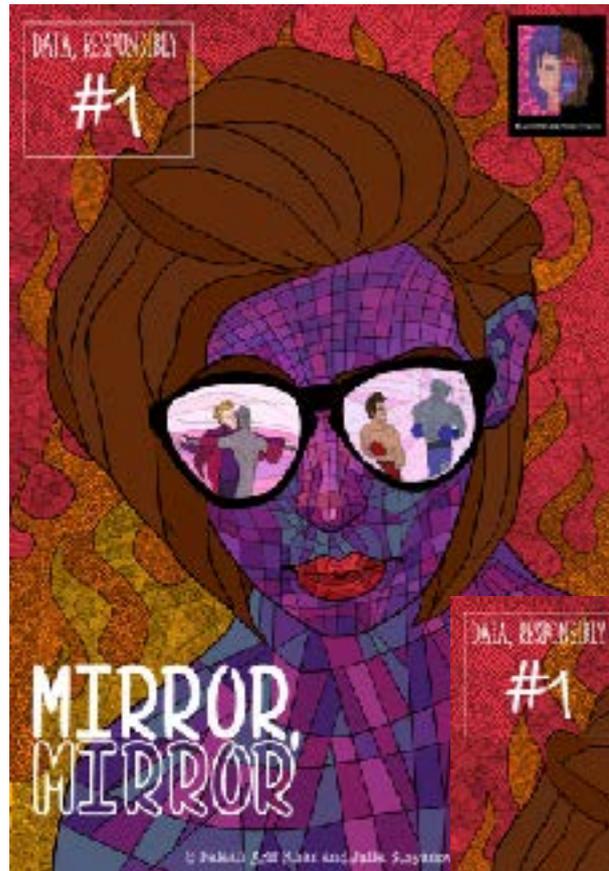
r/ai center
for
responsible
ai



AI comics for the general public



Scientific comics on AI





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

