14. Fair and Transparent Machine Learning

Blase Ur and Mainack Mondal

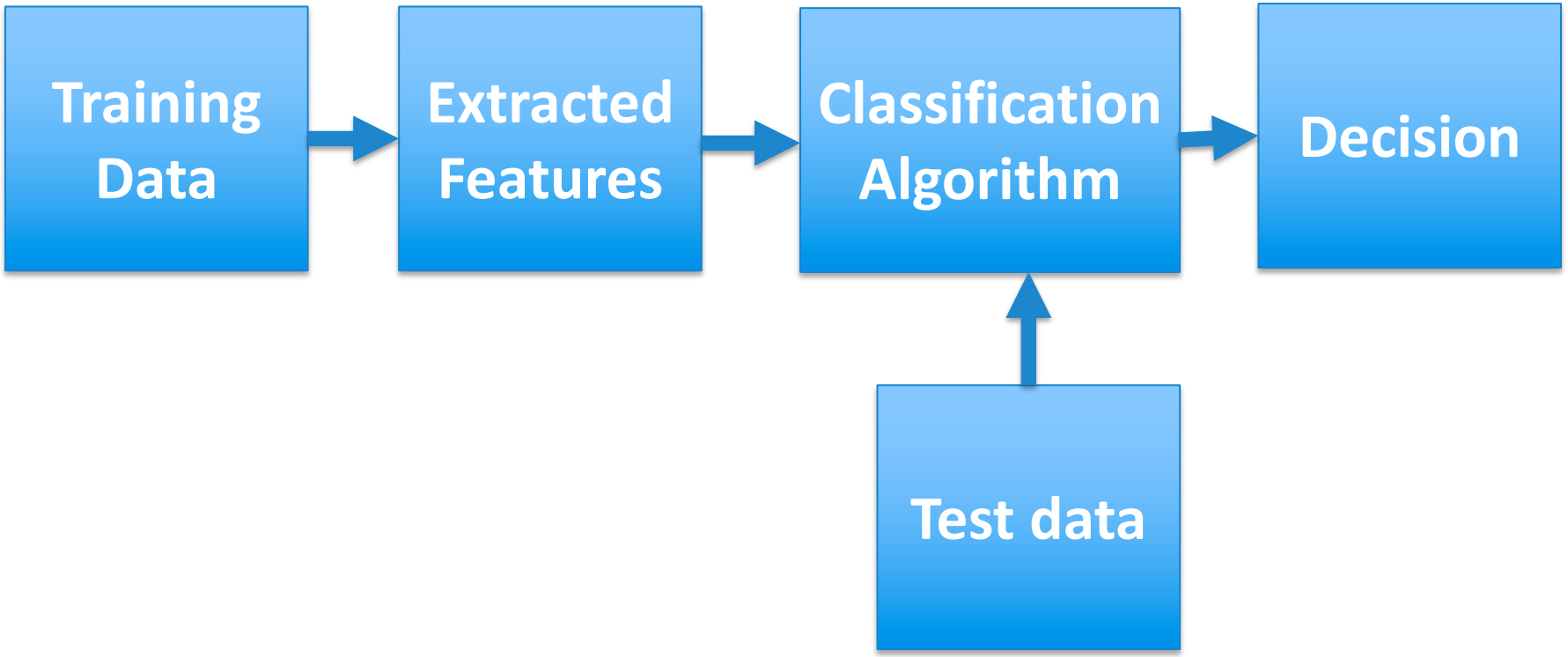
May 9th, 2018

CMSC 23210 / 33210

THE UNIVERSITY OF CHICAGO

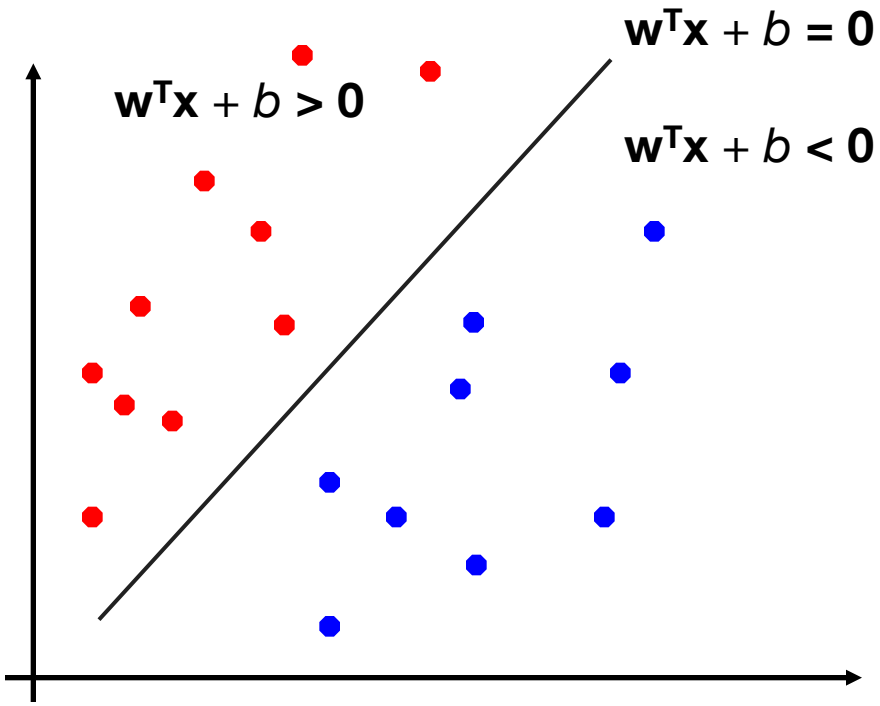
Security, Usability, & Privacy Education & Research

What is machine learning (ML)?



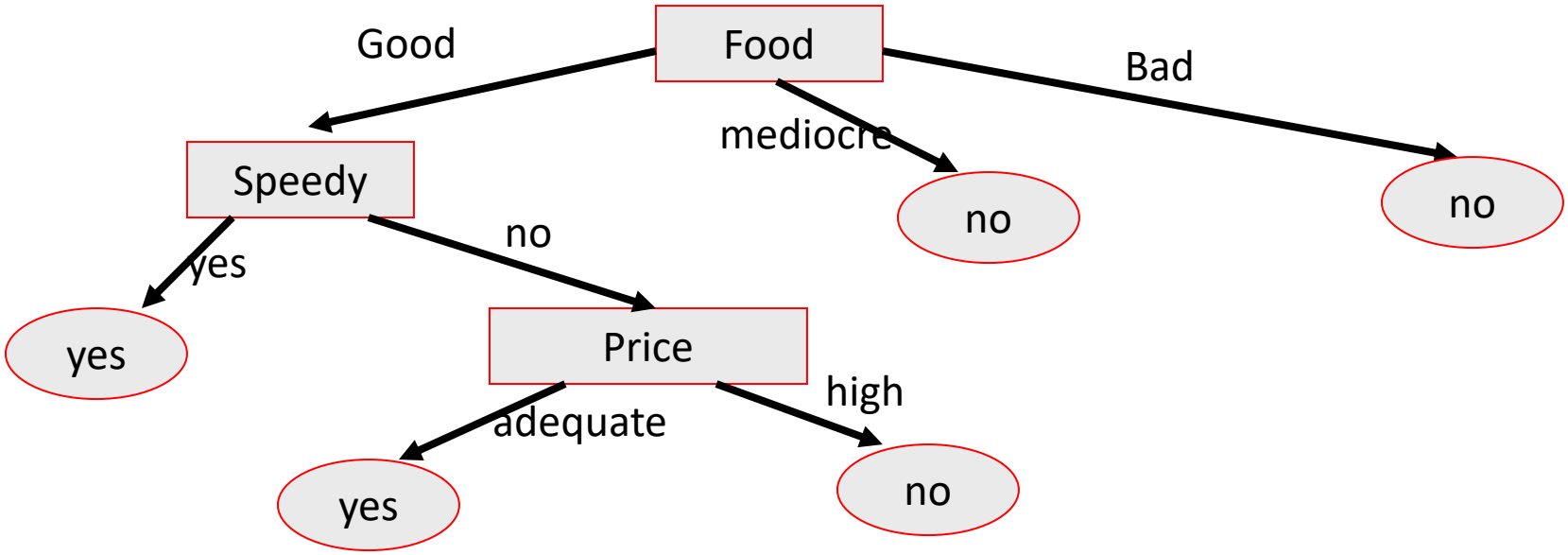
ML Example: Linear classifier

Binary classification: the task of separating classes in feature space



$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$

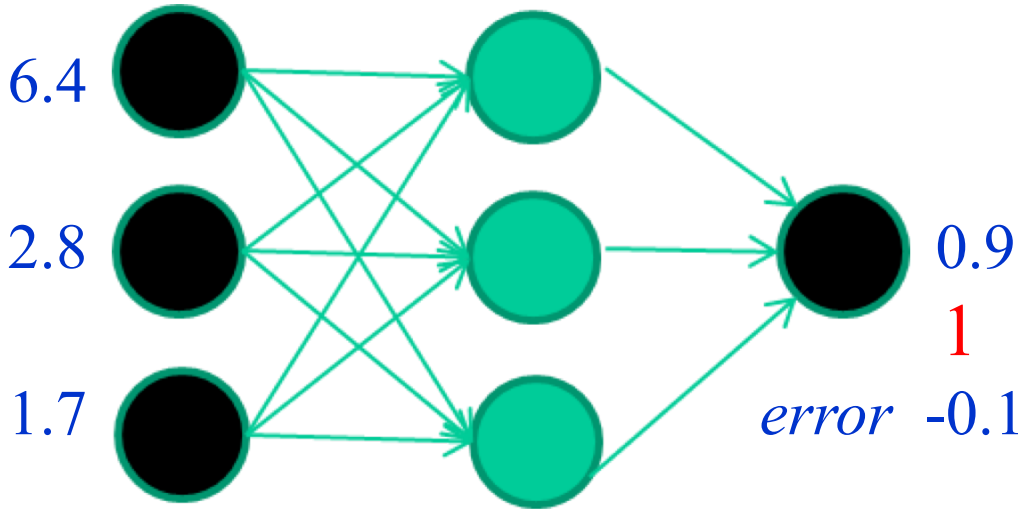
ML Example: Decision trees



ML example: Deep learning

Training data

Fields			***class***
1.4	2.7	1.9	0
3.8	3.4	3.2	0
6.4	2.8	1.7	1
4.1	0.1	0.2	0
etc …			

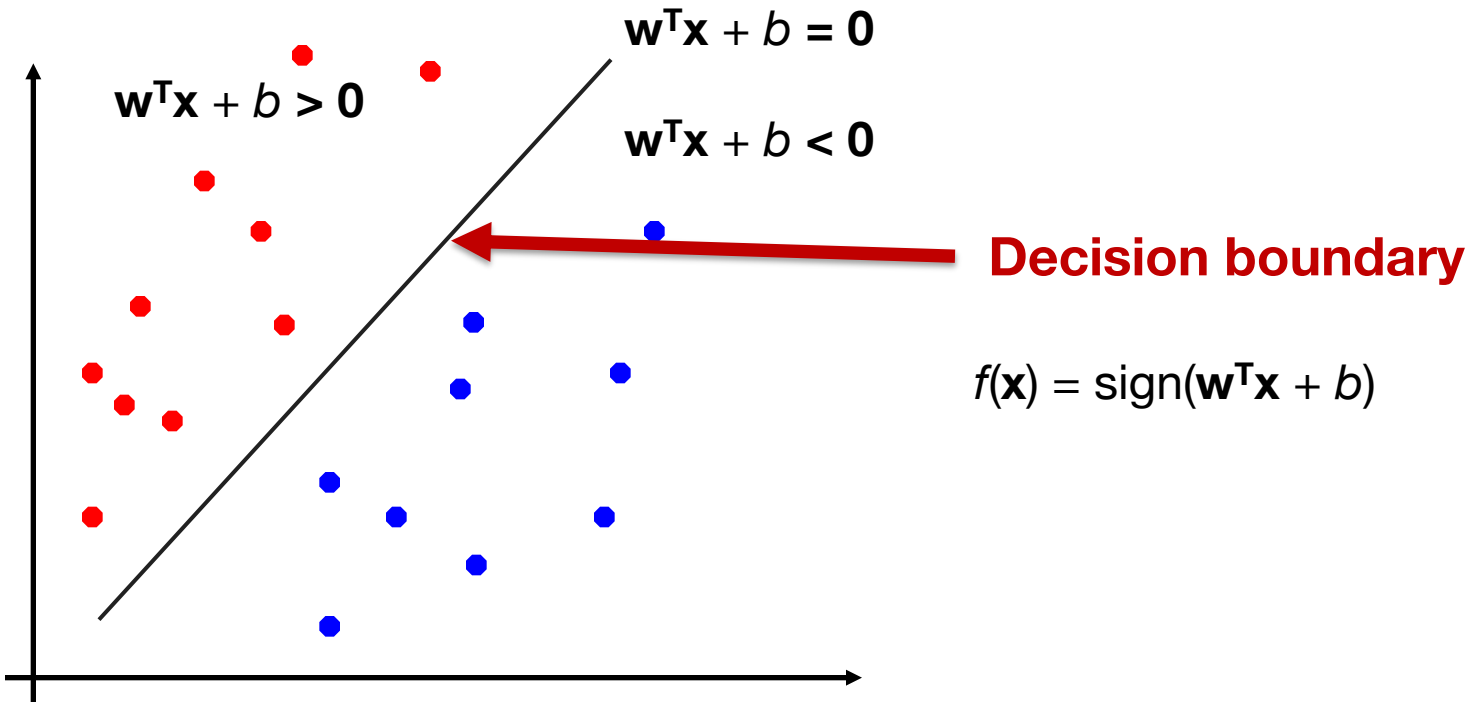


Repeat thousands, maybe millions of times
each time taking a random training instance
make slight weight adjustments

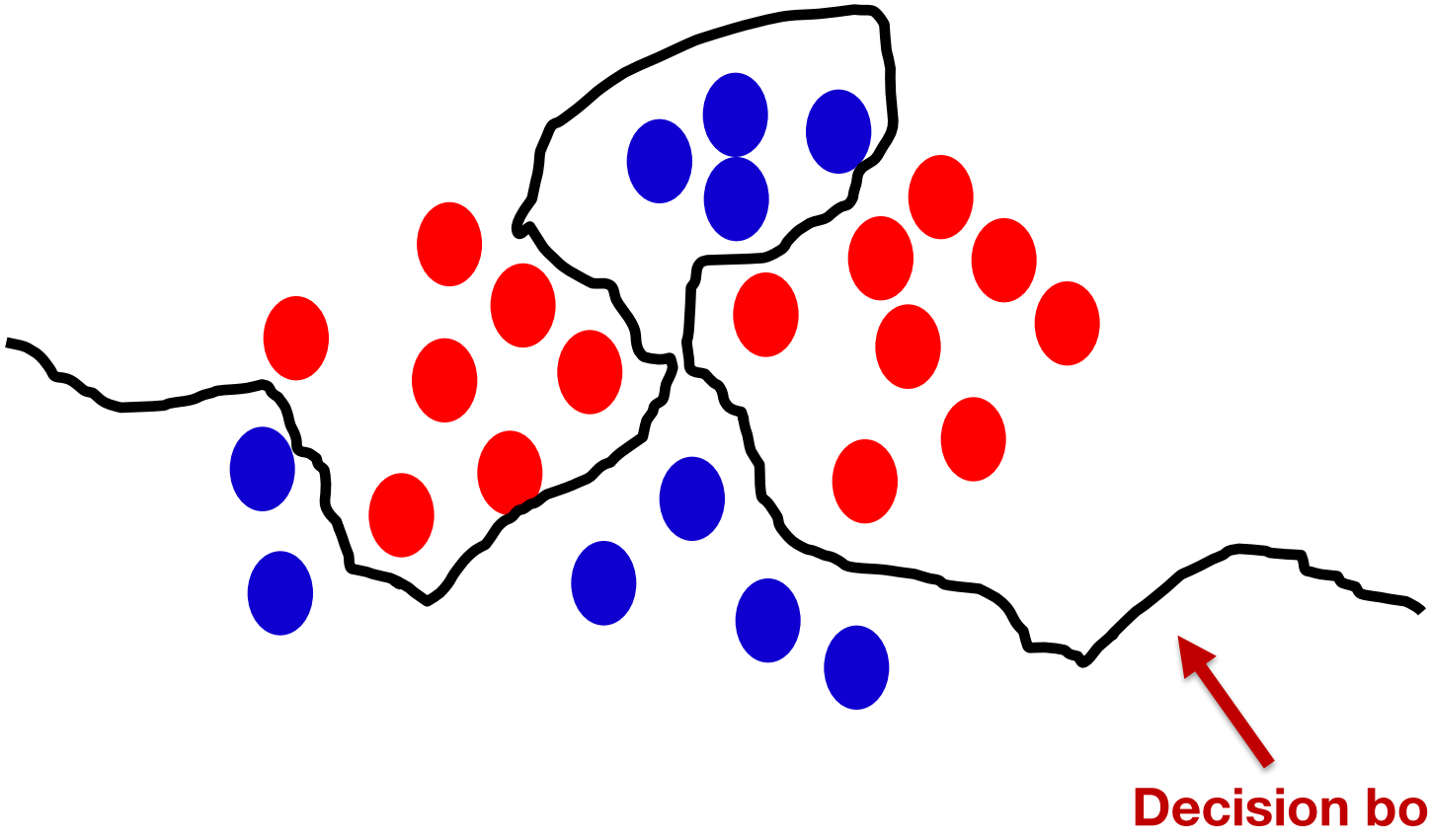
Algorithms for weight adjustment are designed to reduce error

ML Example: Linear classifier revisited

Binary classification: the task of separating classes in feature space



Decision boundaries for other (complex) algorithms



Why do we bother about fair machine learning algorithms?

- Machine learning algorithms are widely used
  - Search, recommendations, social media
- Some algorithms make high-stake decisions
  - Medicine, criminal justice, finance
- Machine learning can have unintended consequences
  - Low classification error is not enough, need fairness

How is ML used in decision making?

By learning from past human decision data

Features

Users

	A_1	A_2		A_m
U_1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,m}$
U_2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,m}$
U_n	$X_{n,1}$	$X_{n,2}$		$X_{n,m}$

Classification Algorithm

Accept
Reject
Accept

	w_1	w_2	...	w_m

Weight vector

Predicted weight (U_i) = $(w_1 \times X_{i,1}) + (w_2 \times X_{i,2}) + \ ... + (w_m \times X_{i,m})$

Accept if predicted weight is positive

Case study: COMPAS

- COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
  - Predicting if a defendant should receive bail
  - Unbalanced false positive rates: more likely to wrongly deny a black person bail

	White	Black
Wrongly labeled high-risk	23.5%	44.9%
Wrongly labeled low-risk	47.7%	28.0%

Types of algorithmic fairness

Distributive Fairness

Fairness of **decision making** outcomes

Procedural Fairness

Fairness of the **decision making** process

Distributive fairness: Motivation

- Doctrine of disparate impact
- US law concerning employment, housing, etc.
  - "*practices [...] considered discriminatory and illegal if they have a disproportionate adverse impact on persons in a protected class*"
  - *Griggs vs. Duke Power co.*, 1971

Distributive fairness: application

- Equal Employment Opportunity Commission

- The 80% rule
  - If 50% of male applicants get selected for the job, at least 40% of females should also get selected

Do not use sensitive features

Sensitive features: gender, race, etc.

	Race	A_2		A_m
U_1	$X_{1,1}$	$X_{1,2}$	…	$X_{1,m}$
U_2	$X_{2,1}$	$X_{2,2}$	…	$X_{2,m}$
U_n	$X_{n,1}$	$X_{n,2}$		$X_{n,m}$

	w_1	w_2	…	w_m

Weight vector

Accept
Reject
Accept

Predicted weight (U_i) = ($w_1 \times X_{i,1}$) + ($w_2 \times X_{i,2}$) + … + ($w_m \times X_{i,m}$)

Do not use sensitive features

Sensitive features: gender, race, etc.

	~~Race~~	A_2		A_m
U_1	~~$X_{1,1}$~~	$X_{1,2}$	…	$X_{1,m}$
U_2	~~$X_{2,1}$~~	$X_{2,2}$	…	$X_{2,m}$
U_n	~~$X_{n,1}$~~	$X_{n,2}$		$X_{n,m}$

	0	w_2	…	w_m

Weight vector

Classification Algorithm

Accept
Reject
Accept

Predicted weight $(U_i) = (0 \times X_{i,1}) + (w_2 \times X_{i,2}) + \ldots + (w_m \times X_{i,m})$

Indirect unfairness

- Correlations between sensitive and non-sensitive features

- Example: What is the gender of a user visiting https://www.iamalpham.com/ ?

Controlling indirect discrimination

- Introduce fairness constraints

- Idea: Optimize the given function *under the constraints*

Key Insight: Limit the covariance between sensitive feature value and distance from decision boundary [Zafar et al. WWW'17]

Distributive fairness is not enough

Distributive fairness

Fairness of **decision making outcomes**

Example

- **Equal misclassification rates**
  - *Grant bail to high risk white defendants*
  - *Deny bail to low risk black defendants*

Procedural fairness

Fairness of the **decision making process**

Example

- **Fairness of using features**
  - *Grant bail based on the criminal history of the defendant's father*

Is it fair to use a feature?

Normative approach

Prescribe how fair decisions ought to be made

Anti-discrimination laws

- Sensitive (race, gender) vs non-sensitive features

Descriptive approach

Describe human perceptions of fairness

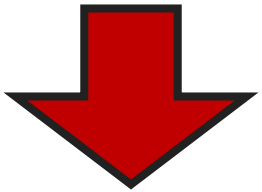
Beyond discrimination?

- Volitional? - father's history
- Relevant? - education
- …

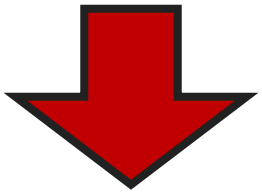
This and some of the following sides/numbers taken from [Grgic-Hlaca et al., WWW'18]

COMPAS revisited

- Helps judges decide if a person should be granted bail



Input: Defendant's answers to the COMPAS questionnaire



Output: Prediction of the defendant's criminal risk

COMPAS questionnaire

137 questions, 10 topics

Current criminal charges	Criminal attitudes
Criminal history	Neighborhood safety
Substance abuse	Criminal history of friends & family
Stability of employment	Quality of social life
Personality	Education & behavior in school

No questions about sensitive features!

Is it fair to use these features to make bail decisions?

Gathering human moral judgments
[Grgic-Hlaca et al., WWW'18]

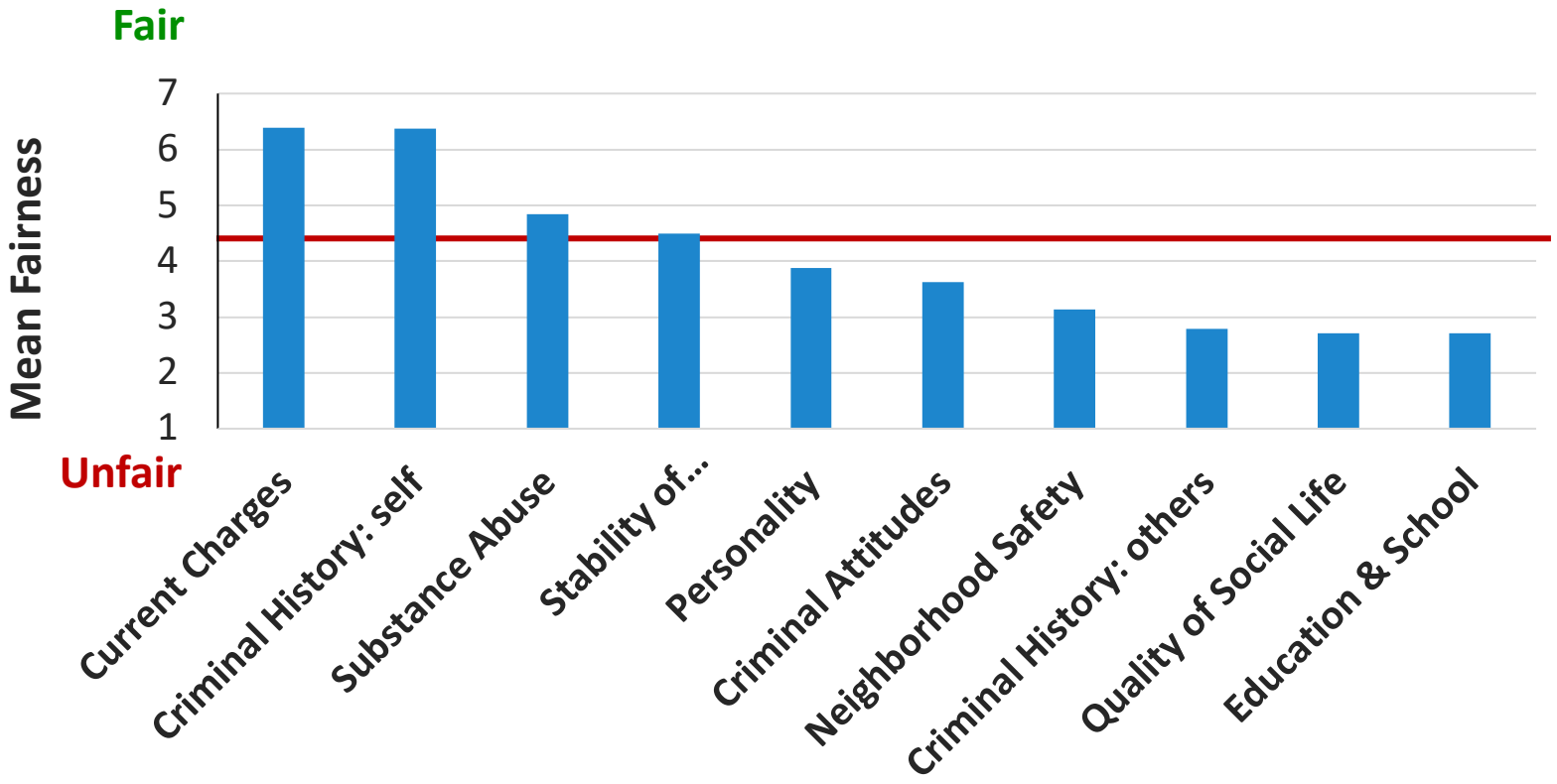
- Fairness of using features for making bail decisions

- US criminal justice system – US respondents
  - 196 Amazon Mechanical Turk master workers
  - 380 SSI survey panel respondents, census representative

Findings consistent across both samples

Human judgments of fairness

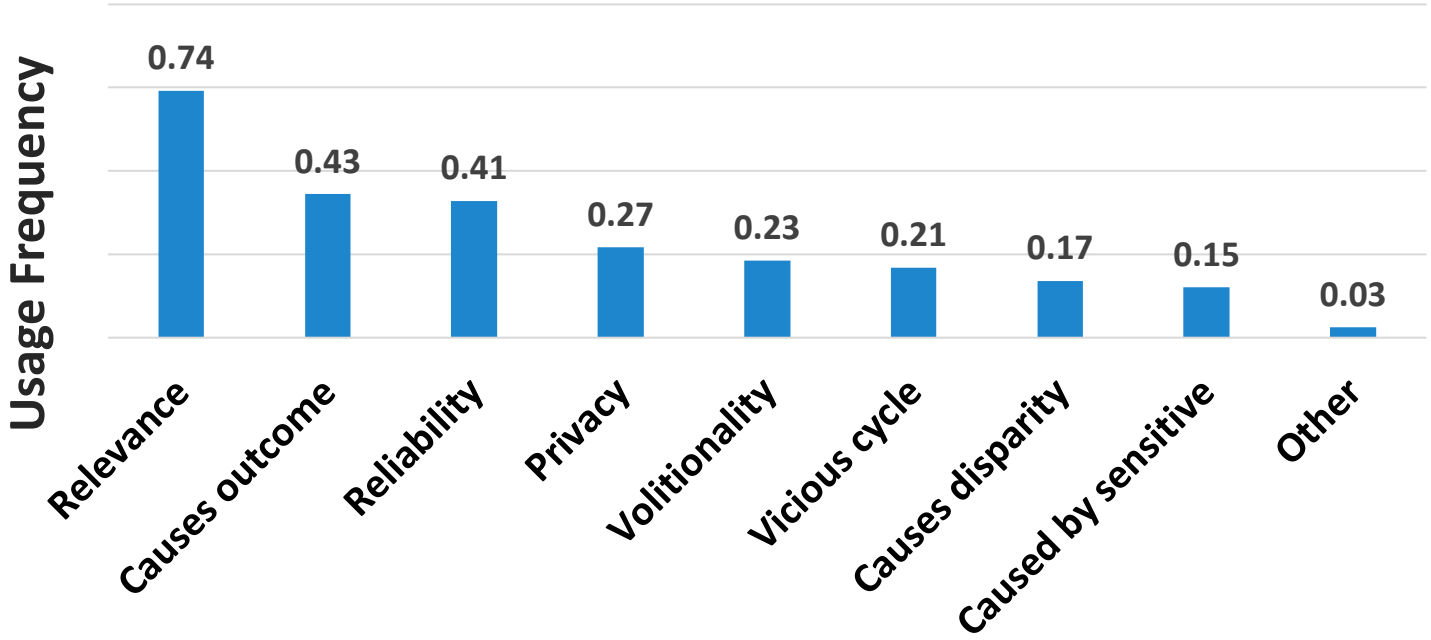
Rating the fairness of using a feature



People consider most of the features unfair

Reasons behind fairness judgments

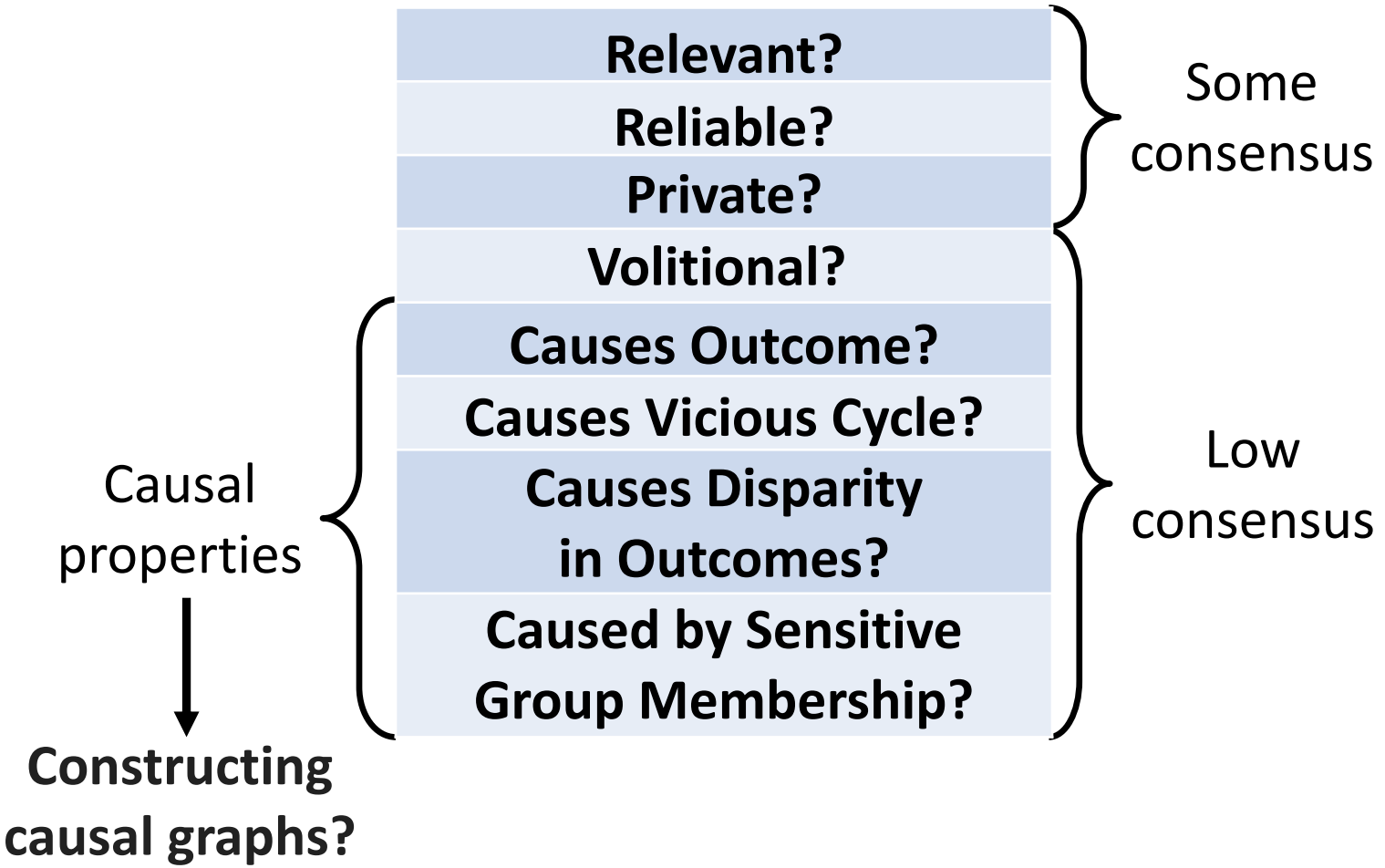
Why is it (un)fair to use a certain feature?



There is more to fairness than discrimination!

They can predict fairness judgement with 88% accuracy

Latent properties for fairness judgement



More notions of fairness

- Group fairness
  - Equalize two groups S and T at the level of outcomes
  - Prob[Getting Credit | S] = Prob[Getting Credit | T]

- Can be abused
- Select barkrupt persons in S and random persons in T
  - Should they get the same credit?

More notions of fairness

- Individual fairness
  - Treat similar persons similarly

Similar for the classification task

Similar outcomes

Transparency in Machine Learning

Most of the ML systems are black boxes

- We don’t know about
  - Input features
  - Classification algorithm

Most of the ML systems are black boxes

- We don't know about
  - Input features
  - Classification algorithm

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

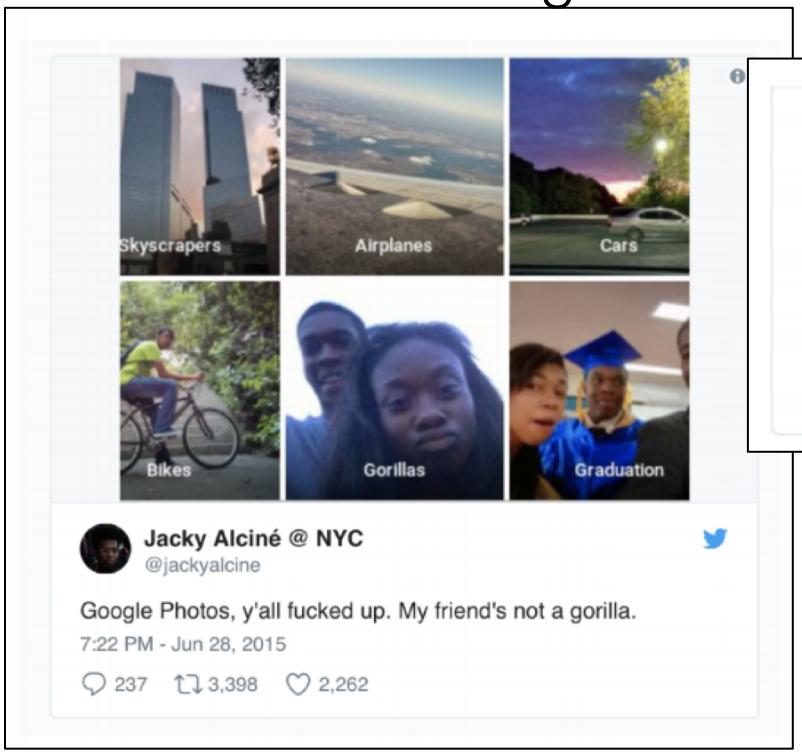
It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

In what appears to be an unintended side effect of Staples' pricing methods—likely a function of retail competition with its rivals—the Journal's testing also showed that areas that tended to see the discounted prices had a higher average income than areas that tended to see higher prices.

Most of the ML systems are black boxes

- We don’t know about
  - Input features
  - Classification algorithm



Most of the ML systems are black boxes

- We don’t know about
  - Input features
  - Classification algorithm

Google self-driving car hits a bus

Dave Lee
North America technology reporter

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky

Most of the ML systems are black boxes

- We don’t know about
  - Input features
  - Classification algorithm

- Gives rise to software bugs
  - Inadvertent unfairness, discrimination and even death
  - How to bring transparency to ML systems?

How to make ML systems transparent?

- Some big data systems might inadvertently make use to sensitive attributes

- Even the developers might not be aware of such data associations

Big Data Systems should not use sensitive attributes/features

- Race, gender, health information
  - Only allowed in specific contexts

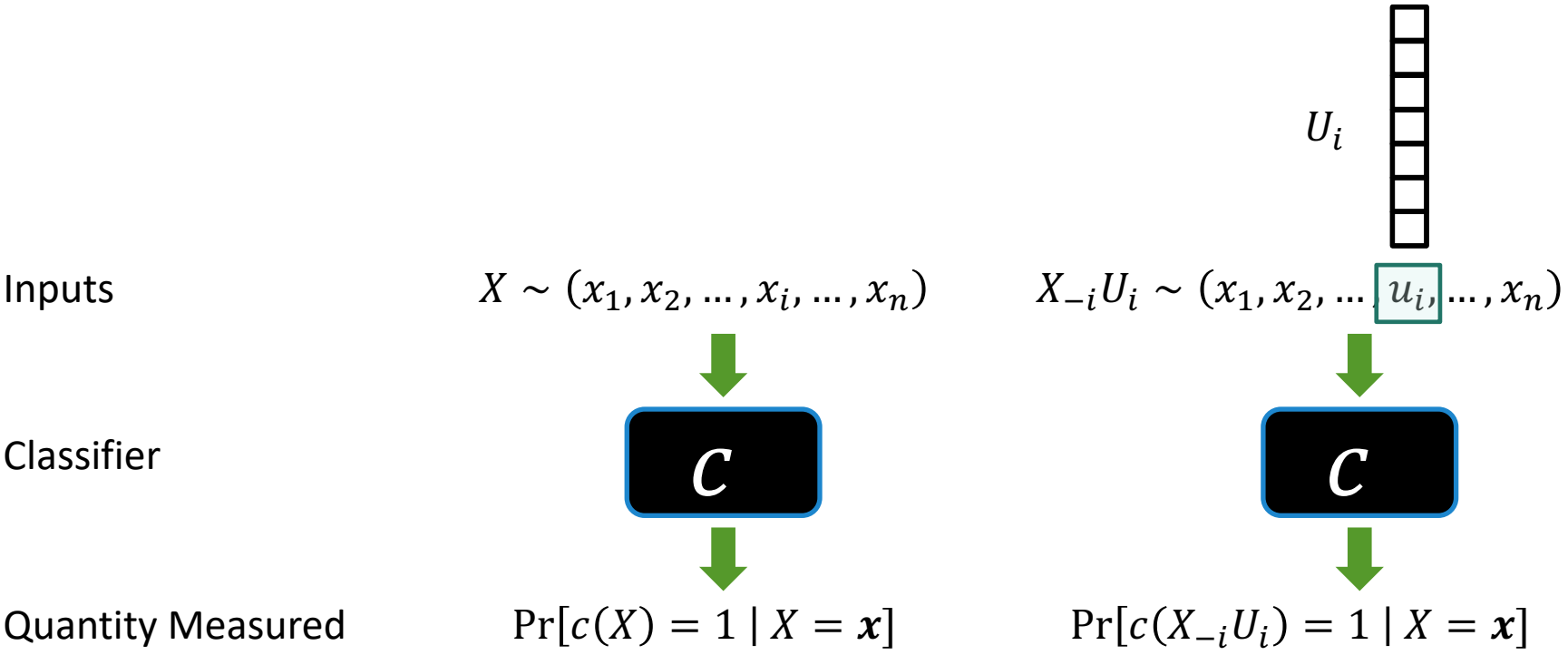
- Alice is denied credit by algorithm from a bank
  - Q1: Which input has most influence in my credit denial?
  - Q2: Which input has most influence in women's credit decision?
  - Q3: Which input influences men getting more positive outcomes than women?

Quantitative input influence (QII)
[Datta et al, IEEE S&P 2016]

- Causal intervention

- Quantity of interest

QII: Causal intervention



Causal intervention: replace x_i with value from an independent random sample u_i. In other words, replace feature with random values from the population

QII: Quantity of Interest

Outcome of an individual	• $\Pr[c(X) = c(\boldsymbol{x}_0) \mid X = \boldsymbol{x}_0]$
Outcomes of a group of individuals	• $\Pr[c(X) = 1 \mid X \text{ is female}]$
Disparity between group outcomes	• $\Pr[c(X) = 1 \mid X \text{ is male}] - \Pr[c(X) = 1 \mid X \text{ is female}]$

QII: Marginal importance

- Individual interventions often have very low effect
- The marginal importance of feature i
  - How much impact does i have given that we have already intervened on feature set S?
  - Then calculate the average of those impacts
  - Aggregate based on economics, game theory etc.

How to help developers debug their ML systems?

- Fairtest [Tramer et al., Euro S&P'17]

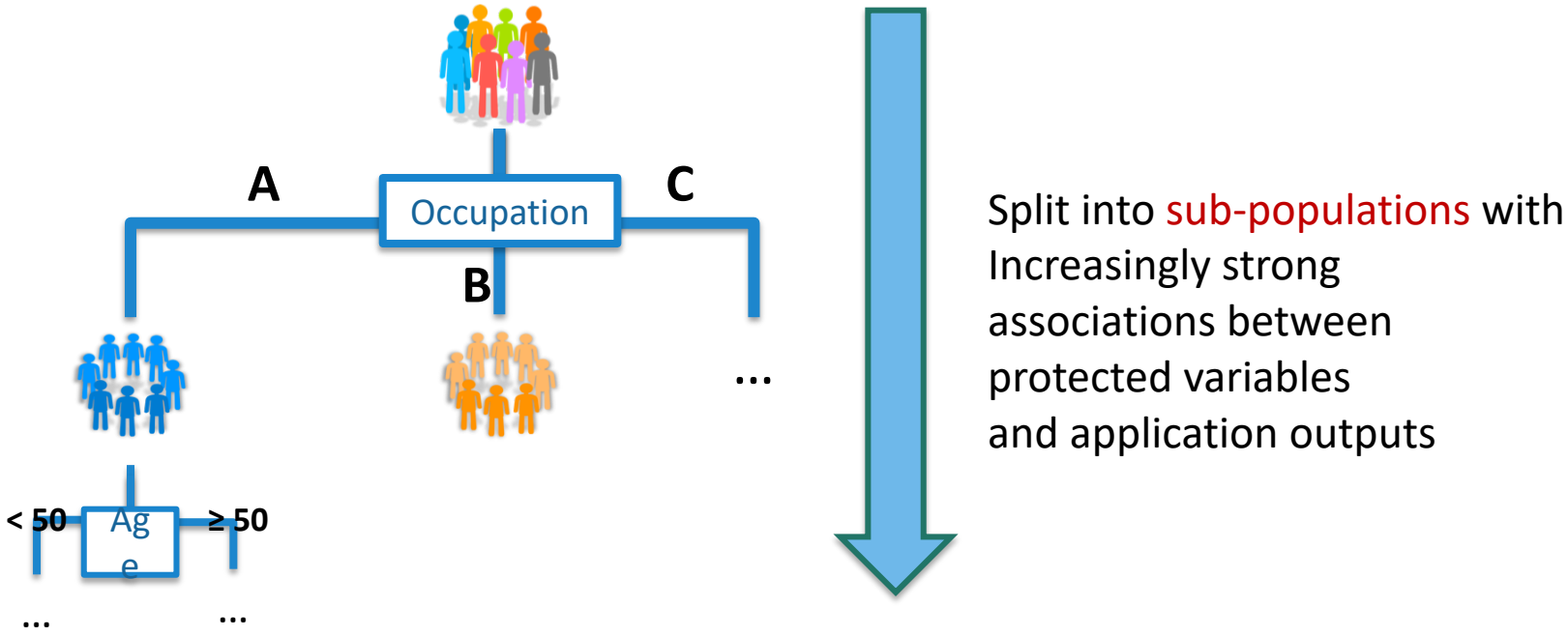
- DeepXplore [Pei et al., SOSP'17]

Fairtest: framework for detecting unwarranted data associations

- Reading for today
  - Key idea: find and examine the subgroups with highest adverse effects
  - Lets just check an example

Detecting disparate impact of location based pricing on low income population

Goal: find most strongly affected user sub-populations



Split into sub-populations with Increasingly strong associations between protected variables and application outputs

Association guided decision trees

DeepXplore: Systematic testing of deep neural networks

- Developers leverage test set to measure accuracy of their deep learning systems
  - How good is the coverage of test set?
  - Expanding test set means more manual labeling

DeepXplore: Systematic testing of deep neural networks

- Coverage metric: What fraction of neurons are activated by a test suite?

- Generate test inputs to activate more and more neurons
  - Utilized output of multiple deep neural networks
  - Formulate neuron coverage as optimization problem