Decision Making by Machine Learning Algorithms

Sampath Kannan

Computer and Information Sciences

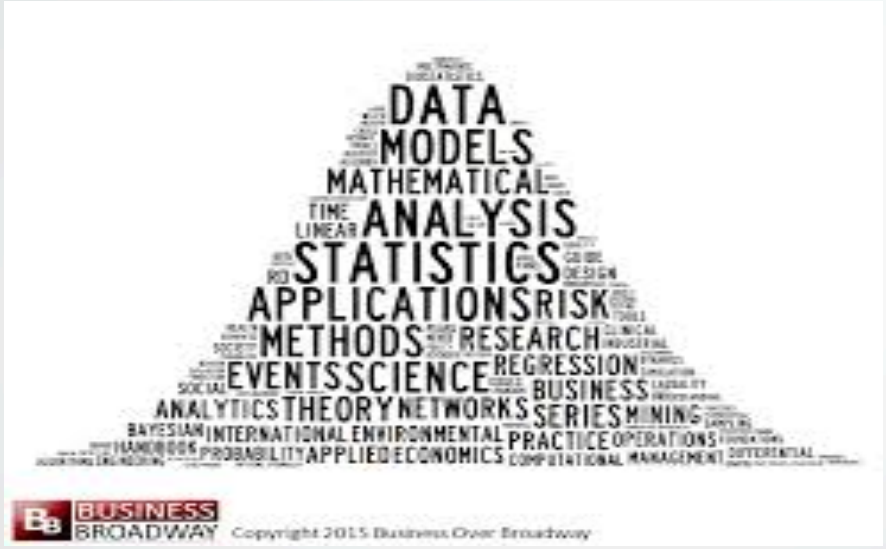
University of Pennsylvania

In the Information Age

- Data drives decisions …
  … What car should I buy? Where should I buy it?
- But I might be the `object' of some decisions:
  - What medicine will I be given?
  - What movies will be recommended to me?
- Such decisions classify people
  - loan-worthy or not
  - recidivism risk or not
  - college-ready or not
- These decisions/recommendations are made by machine learning algorithms.

What is Machine Learning ?

- Just statistics



- With a focus on ***prediction***
- With a focus on ***engineering***

Machine Learning - Illustration

Just one kind of learning – supervised learning. … also
Unsupervised learning
Reinforcement learning
Active learning … and others

Supervised Classification

Name	Income	Debt	Assets	Loan Amt	Credit Score	Paid back loan?
Alice	60,000	20,000	50,000	100,000	570	Yes
Bilal	45,000	0	30,000	40,000	550	Yes
Celia	25,000	5,000	90,000	20,000	650	No
Ding	75,000	10,000	10,000	30,000	630	No
Elvira	90,000	15,000	20,000	50,000	720	Yes
Franz	50,000	10,000	60,000	80,000	680	No
Gautam	70,000	25,000	25,000	30,000	600	Yes

X Y

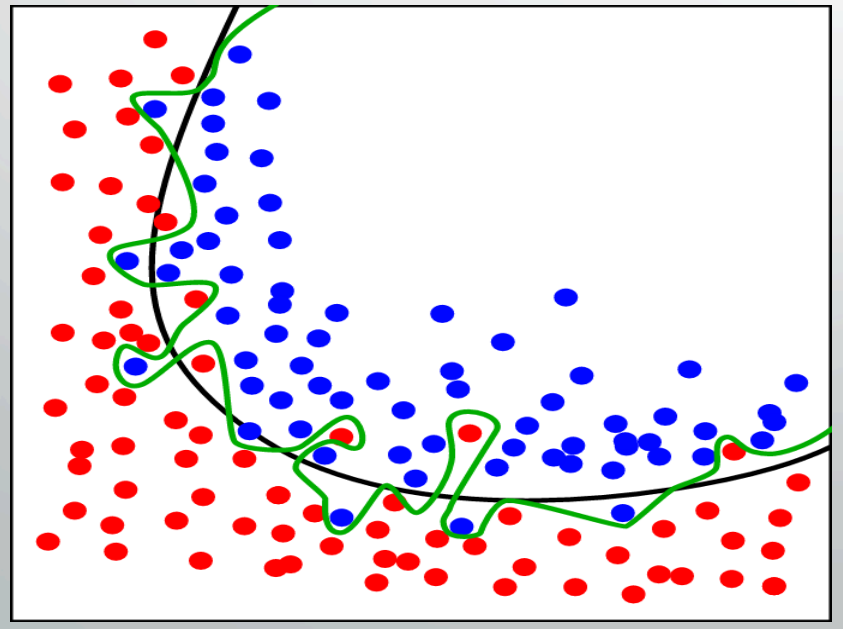
- Training data: Points with d *features* X, and a *label* Y
- Goal: Find a rule to predict *label* Y from *features* X for new data.

Terminology

- The labeled data the algorithm is given is called training data
- Data that the algorithm will label is called test data
- Training data is drawn from some unknown distribution on people
- Assume test data also comes from the same distribution (!!)
- After training, algorithm produces a decision rule or hypothesis
- This rule is used to classify the new people (test data)
- Decision rule has training error and generalization (test) error.

What doesn't work

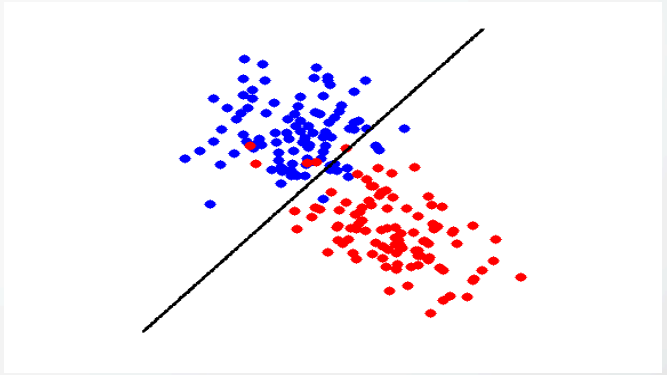
- Decision rule that exactly fits training data.
- This *overfits…*
  - We haven't **learnt;** only **memorized**
  - We wouldn't know how to generalize, i.e., classify a person whose data is different
  - To learn, we must **summarize / simplify**



This inevitably introduces errors

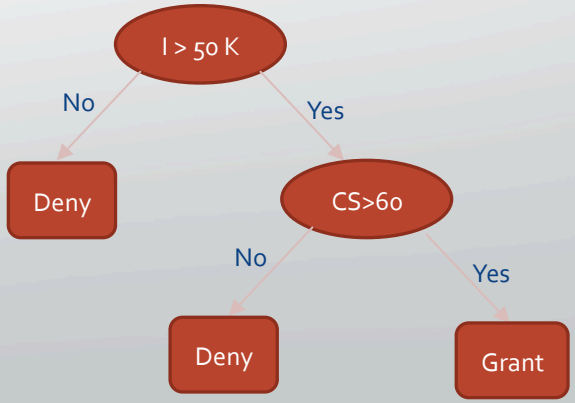
What do we mean by `simplify'?

- Limit decision rules to come from a "simple" class $\mathcal{C}$… for example
  - *linear functions*



  - *small decision trees*

y



Why might machine learning be “unfair”?

- Many reasons:
  - Data might encode existing biases.
    - E.g. labels are not “Committed a crime?” but “Was arrested.”
  - Data collection feedback loops.
    - E.g. only observe “Paid back loan?” if the loan was granted.
  - Different populations with different properties.
    - E.g. “SAT score” might correlate with label differently in populations that employ SAT tutors.
  - Less data (by definition) about minority populations.

Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program
OCTOBER 12, 2018 | 1:00 PM

TAGS: Women's Rights in the Workplace, Women's Rights, Privacy & Technology

In 2014, a team of engineers at Amazon began working on a project to automate hiring at their company. Their task was to build an algorithm that could review resumes and determine which applicants Amazon should bring on board. But, according to a Reuters report this week, the project was canned just a year later, when it became clear that the tool systematically discriminated against women applying for technical jobs, such as software engineer positions.

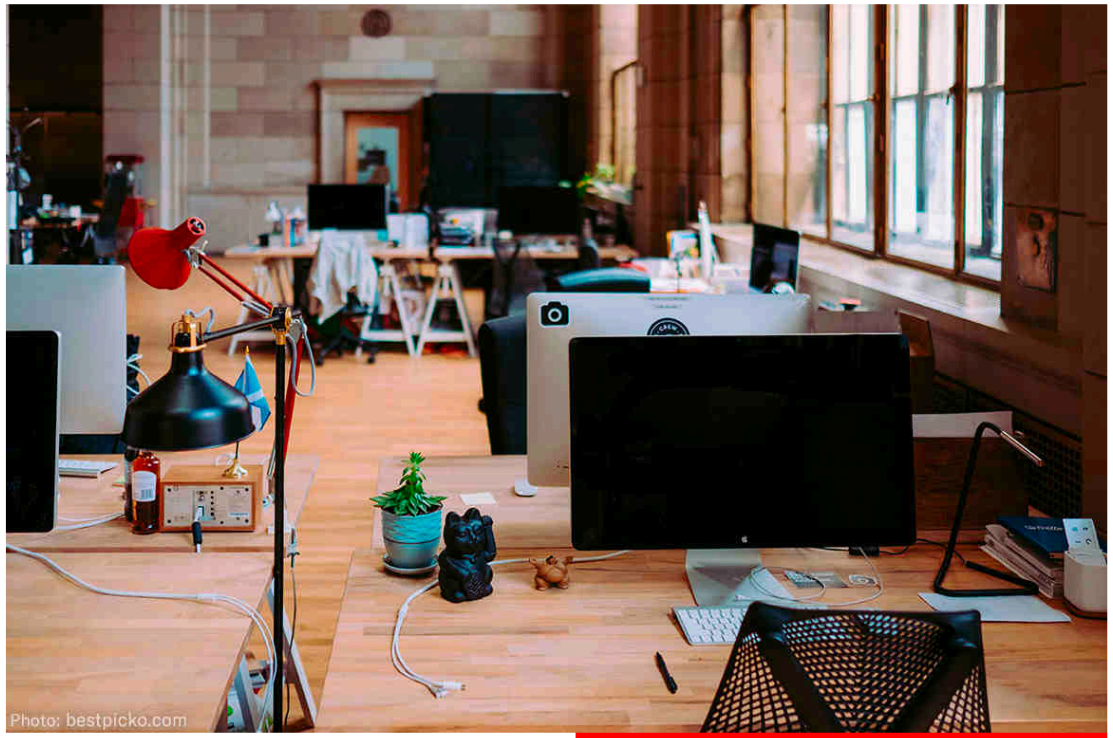


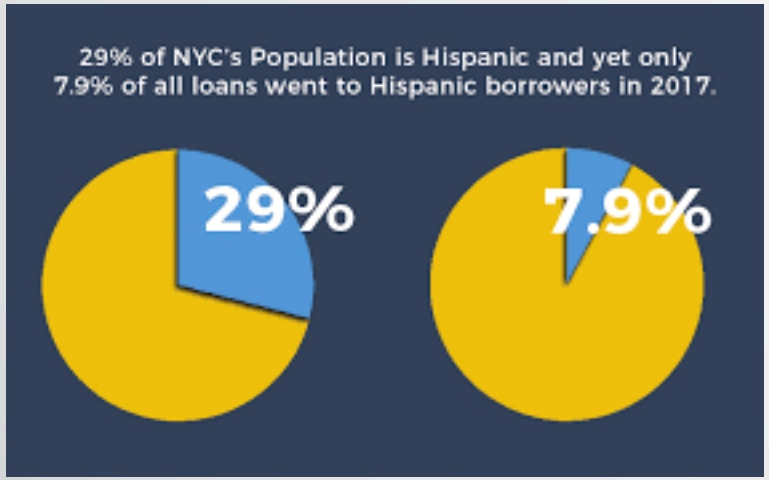
Photo: bestpicko.com

It shouldn't surprise us at all that the tool developed this kind of bias. The existing pool of Amazon software engineers is overwhelmingly male, and the new software was fed data about those engineers' resumes. If you simply ask software to discover other resumes that look like the resumes in a "training" data set, reproducing the demographics of the existing workforce is virtually guaranteed.

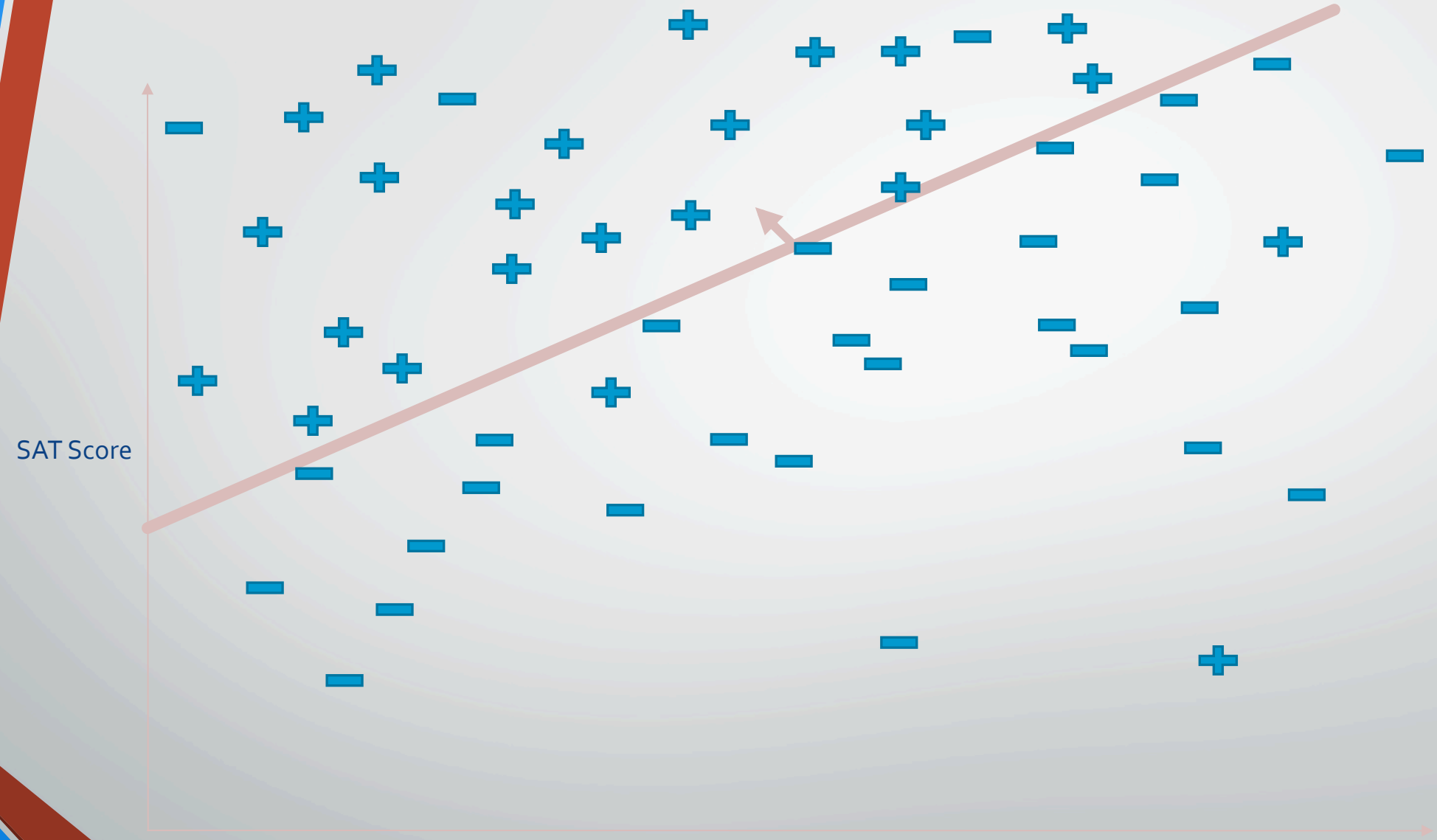
In the case of the Amazon project, there were a few ways this happened. For example, the tool disadvantaged candidates who went to certain women's colleges presumably not attended by many existing Amazon engineers. It similarly downgraded resumes that included the word "women's" — as in "women's rugby team." And it privileged resumes with the kinds of verbs that men tend to use, like "executed" and "captured."

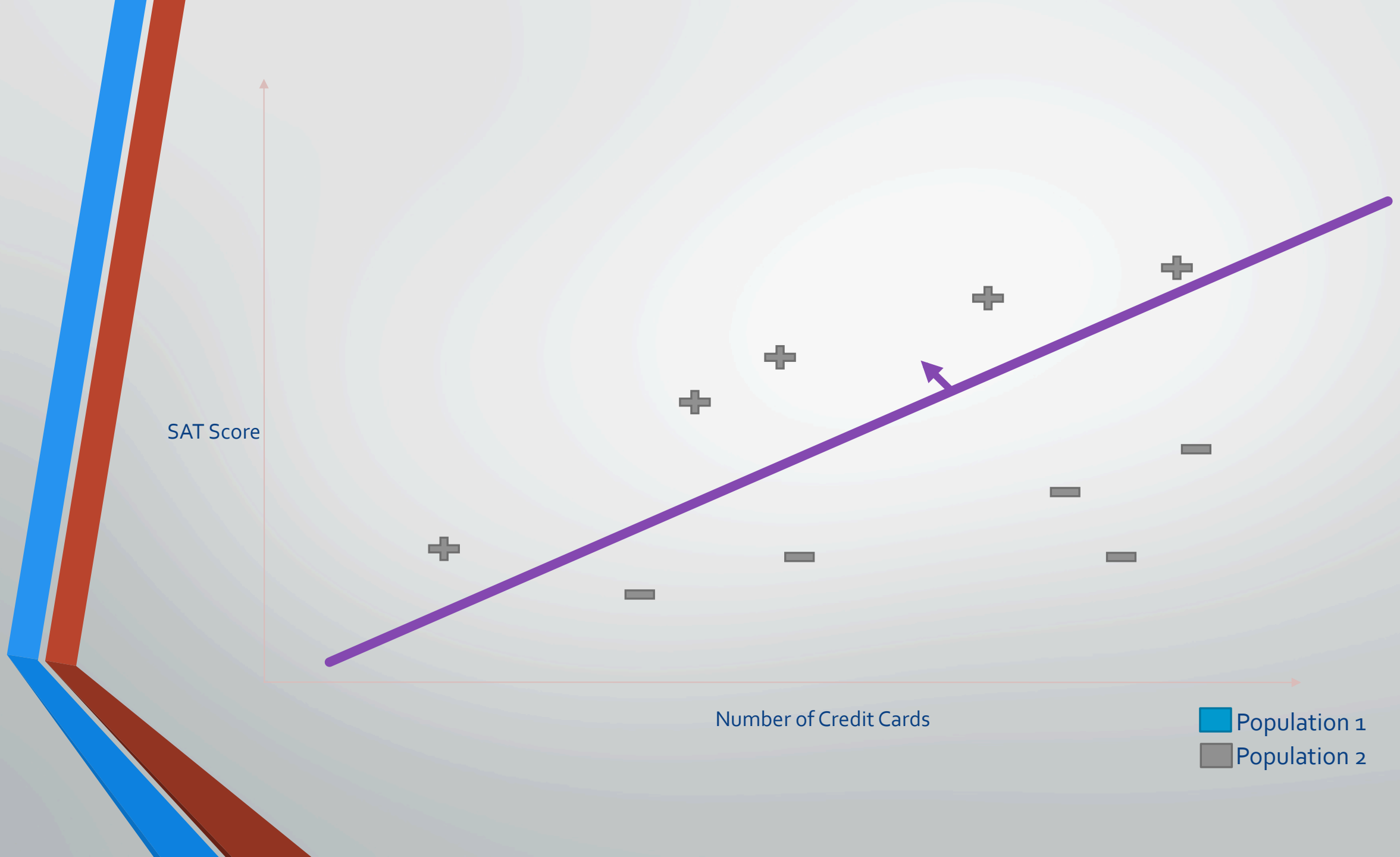
r"?

- Data collection feedback loops.
  - E.g. only observe “Paid back loan?” if the loan was granted.



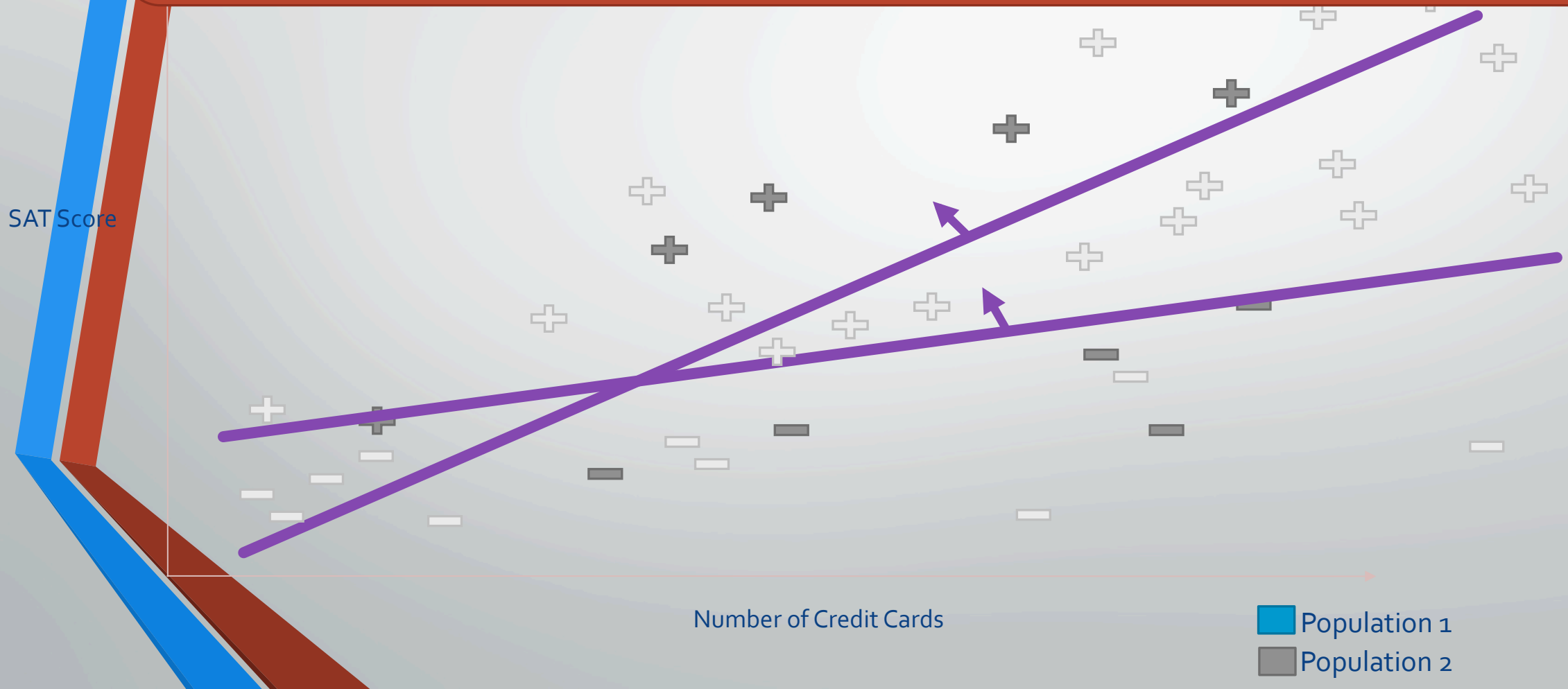
Different populations with different properties: Artificial example – College admissions



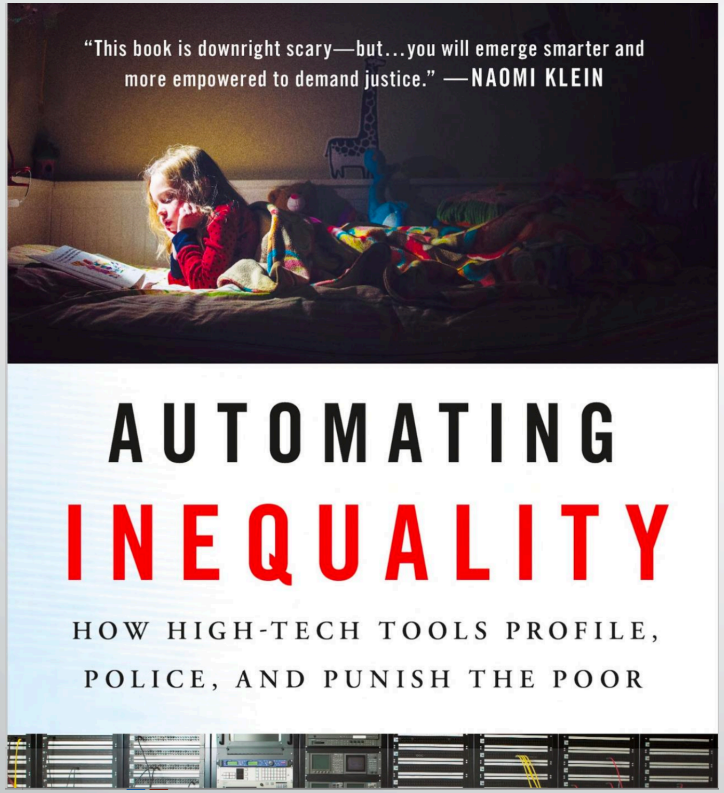

SAT Score
Number of Credit Cards
Population 1
Population 2

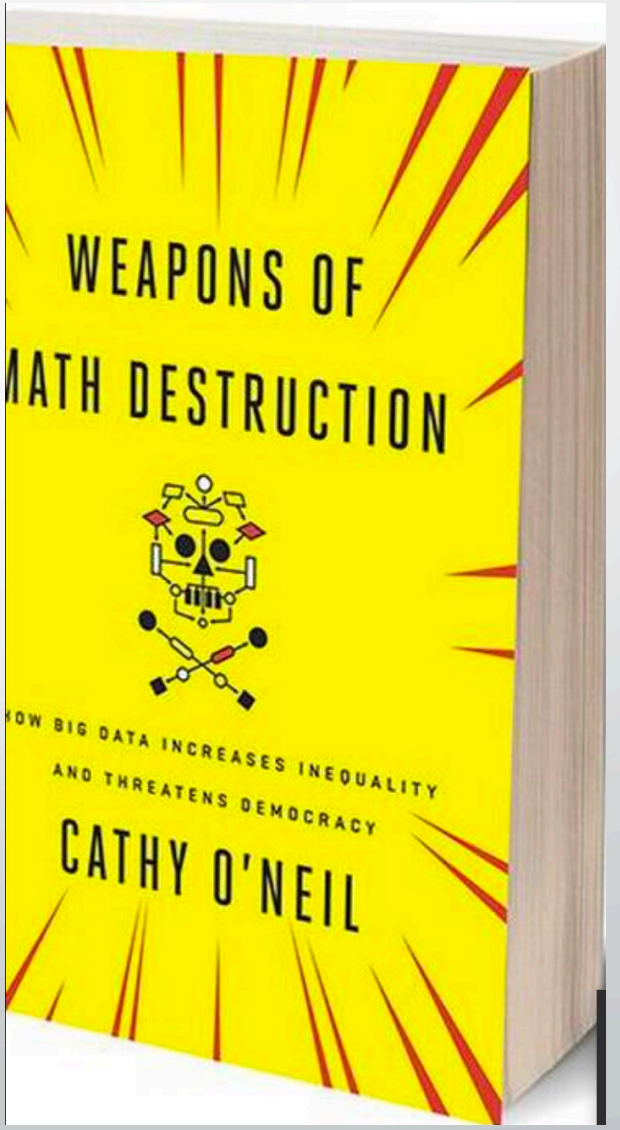
Upshot: To be "fair", the algorithm may need to *explicitly* take into account group membership. Else, optimizing accuracy fits the majority population.

Upshot: Algorithms trained on minority populations will be less accurate.
Qualified individuals will be denied at a higher rate.



Several excellent books describing what's wrong with state of affairs





This panel's perspective

- Theoretical Computer Science approach
- Precisely define what we want to solve:
  - Then worry about how to solve it efficiently
  - If possible exactly; if not approximately, but with provable guarantees
- So need precise definitions of fairness goals and other goals
  - Will hear more about this in the rest of the panel

How does this perspective help

- Precise formulations of classification accuracy and fairness goals allow us to
  - Have some measure of transparency in the decision process
  - Understand what goals can be achieved efficiently and what goals cannot
  - Prove theorems about the impossibility of simultaneously achieving several goals
  - Understand potential unintended consequences
  - Understand what goals align better with larger societal objectives such as minimizing crime rate or eliminating poverty

Final Thoughts

- Research on fairness in classification is in its nascent stages
- Definitions proliferate; decision algorithms are complex and opaque
- In the public sphere and in decisions impacting people's lives, need for more oversight and regulation of criteria and algorithms, with input from a range of experts including computer scientists
- Most of the work so far only addresses fairness in `one-shot classification' and does not address the compounding effects of discrimination through a pipeline of decisions.