Just Machine Learning

Tina Eliassi-Rad

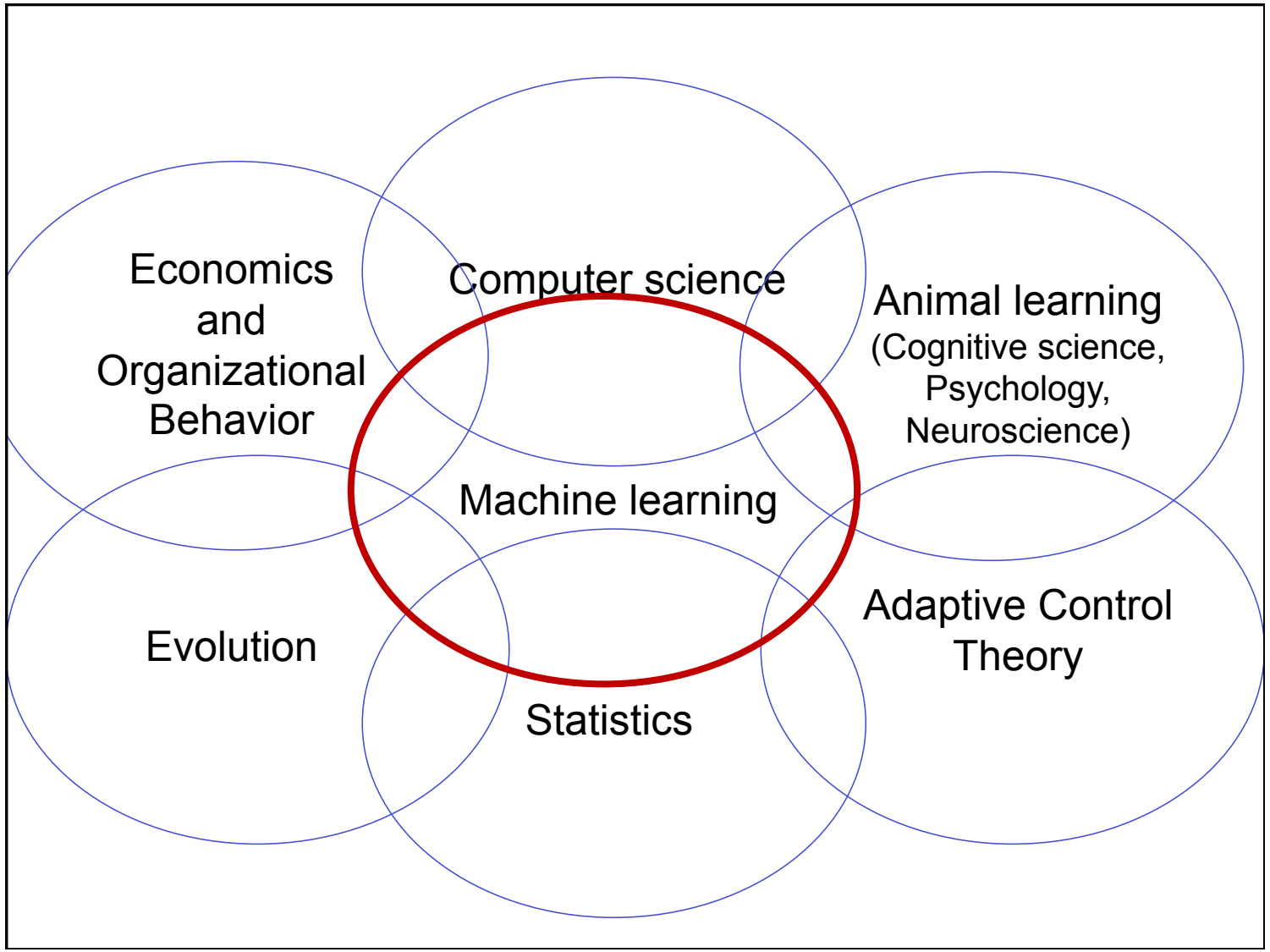
tina@eliassi.org @tinaeliassi

http://eliassi.org/safra17.pdf

What is machine learning?

https://xkcd.com/1838/

Machine learning emerged from AI



CS, Statistics, Machine Learning

- Computer Science: How can we build machines that solve problems, and which problems are inherently tractable/intractable?
- Statistics: What can be inferred from data plus a set of modeling assumptions, with what reliability?
- Machine Learning: How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

Place of ML within computer science

- The application is too complex for people to manually design algorithms – e.g., computer vision
- The application requires that the software customize to its operational environment after it is fielded – e.g., speech recognition

(Tom Mitchell, The Discipline of Machine Learning, July 2006)

- Arthur Samuel coined the term machine learning (1959)
  - *Field of study that gives computers the ability to learn without being explicitly programmed*
  - The Samuel Checkers-playing Program

What does “learning” mean?

The Well-posed Learning Problem:

A computer program is said to **learn** from experience E w.r.t. some task T and some performance measure P, if its performance on T, as measured by P, **improves** with experience E. -- Tom Mitchell (1997)

Example

- Suppose your email program observes which emails you mark as spam and which you do not, and based on that information learns how to better filter spam.

Example

- Suppose your email program observes which emails you mark as spam and which you do not, and based on that information learns how to better filter spam.
- Task T: Classifying emails as spam or not spam

Example

- Suppose your email program observes which emails you mark as spam and which you do not, and based on that information learns how to better filter spam.
- Task T: Classifying emails as spam or not spam
- Experience E: Observing you label emails as spam or not spam

Example

- Suppose your email program observes which emails you mark as spam and which you do not, and based on that information learns how to better filter spam.
- Task T: Classifying emails as spam or not spam
- Experience E: Observing you label emails as spam or not spam
- Performance P: The number (or fraction) of emails correctly classified as spam/not spam

Some “success” stories

- IBM Watson defeats the best human competitors in Jeopardy!
- Google AlphaGo Model defeats Euro Go Campaign
- Speech recognition: Amazon Alexa, Apple Siri, Google Go, …
- Image recognition
- Translation
- Fraud detection
- Self-driving cars
- Recommendation systems: Amazon, NetFlix, …

Ethical issues in AI

Top 9 (h/t Julia Bossman)

Unemployment	Artificial stupidity	Security
Wealth inequality	Evil genies	Robot rights
Humanity	Singularity	Racist/sexist robots

http://bit.ly/2etvH3X

Ethical issues in AI

Top 9 (h/t Julia Bossman)

Unemployment	Artificial stupidity	Security
Wealth inequality	Evil genies	Robot rights
Humanity	Singularity	Racist/sexist robots

http://bit.ly/2etvH3X

Unemployment

- The end of work?
- The threat of automation
  - Manufacturing
  - Trucking
  - Loan underwriting
- AI and the Future of Work: http://futureofwork.mit.edu

Wealth inequality

- Increased wealth inequality is a likely consequence of the end of work
- The owners of AI tech/IP will be advantaged
- “Automated Inequality” by Nicolas Yan (Harvard Political Review, 2016)
  - http://harvardpolitics.com/world/automation/
- “The outcome—shared prosperity or increasing inequality—will be determined not by technologies but by the choices we make as individuals, organizations, and societies. If we fumble that future—if we build economies and societies that exclude many people from the cycle of prosperity—shame on us.” -- Erik Brynjolfsson (HBR 2015)

Humanity

- Altering our behaviors and interactions?
- Vying for your attention (for ad $$$)
  - Click-baiting, fake news, polarization, tech addiction
- “You'll Be Outraged at How Easy It Was to Get You to Click on This Headline” by Bryan Gardiner (Wired, 2015)
  - https://www.wired.com/2015/12/psychology-of-clickbait/
- MIT Conference on Digital Experimentation: http://bit.ly/2i56mgl

Artificial stupidity

- Adversarial machine learning
  - Wikipedia entry: http://bit.ly/2gW908q
- Exploitation of stupidity
  - Poisoning vs. evasion attacks
  - Black box vs. white box attack
- Fooling Google's Cloud Vision API

Evil genies

- Unintended consequences due to poorly defined tasks or faulty experience/data (garbage in, garbage out)
- “Customer Experience, Opaque AI And The Risk Of Unintended Consequences” by Adrian Swinscoe (Forbes, 2017)
  - http://bit.ly/2zq4nP6

Singularity

- What if super-intelligence emerges?
- Wikipedia entry: http://bit.ly/1TMfHXl

Security

- Weaponization of AI in both physical and cyber space
  - Drones, robot soldiers
- Wikipedia entry: http://bit.ly/2z7B4At

Robot rights

- Can robots have moral status?
- When is a Robot a Moral Agent?” by John P. Sullins (International Review of Information Ethics, 2006)
  - http://bit.ly/2z5KAUF
- Georgia Tech Robot Ethics
  - https://ethics.gatech.edu/robot-ethics

Ethical issues in AI

Top 9 (h/t Julia Bossman)

Unemployment	Artificial stupidity	Security
Inequality	Evil genies	Robot rights
Humanity	Singularity	Racist/sexist robots

http://bit.ly/2etvH3X

Racist/sexist robots

- “Bias in Computer Systems” by Batya Friedman and Helen Nissenbaum (ACM Transactions on Information Systems, 1996)
  - https://dl.acm.org/citation.cfm?id=230561
- “Algorithmic Bias in Autonomous Systems” by David Danks and Alex John London (IJACI 2017)
  - http://bit.ly/2zrdbnX
- UC Berkeley Course on Fairness in Machine Learning
  - https://fairmlclass.github.io
- Fairness, accountability, and transparency
  - FatML Conferences: https://www.fatml.org

Racist Robots in the News

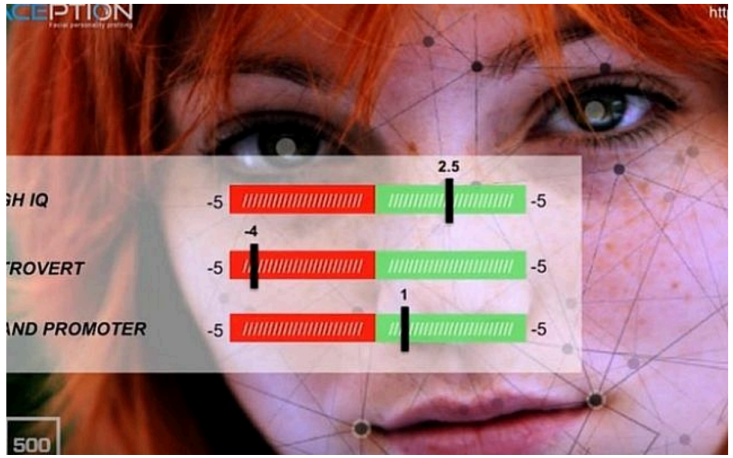
FACEPTION 'CAN MATCH AN INDIVIDUAL WITH VARIOUS PERSONALITY TRAITS AND TYPES WITH A HIGH LEVEL OF ACCURACY'

New Israeli facial imaging claims to identify terrorists and pedophiles

Tel Aviv start-up Faception says its face 'classifiers' can spot criminals and even great poker players in a split second, but the experts are not convinced

By SUE SURKES
24 May 2016, 10:52 pm | 9

3K shares



An image taken from a May 2016 presentation by Faception co-founder Shai Gilboa (screen capture: YouTube)

A Tel-Aviv based start-up company says it has developed a program to identify personality types such as terrorists, pedophiles, white collar offenders and even great poker players from facial analysis that takes just a fraction of a second.

OPINION | TECH

'Gaydar' Shows How Creepy Algorithms Can Get

Imagine what an oppressive government could do with it.

By Cathy O'Neil
409 September 25, 2017, 6:30 AM EDT

Watch out. *Photographer: Jin Lee/Bloomberg*

Artificial intelligence keeps getting creepier. In one controversial study, researchers at Stanford University have demonstrated that facial recognition technology can identify gay people with surprising precision, although many caveats apply. Imagine how that could be used in the many countries where homosexuality is a criminal offense.

Nikon S630

GOOGLE

Google Photos Mistakenly Labels Black People 'Gorillas'

BY CONOR DOUGHERTY JULY 1, 2015 7:01 PM 41

Email
Share
Tweet
Save
More

Google continued to apologize Wednesday for a flaw in Google Photos, which was released to great fanfare in May, that led the new application to mistakenly label photos of black people as "gorillas."

The company said it had fixed the problem and was working to figure out exactly how it happened.

"We're appalled and genuinely sorry that this happened," said a Google representative in an emailed statement. "We are taking immediate action to prevent this type of result from appearing."

From self-driving cars to photos, Google, like every technology company, is constantly releasing cutting-edge technologies with the understanding that problems will arise and that it will have to fix them as it goes. The idea is that you never know what problems might arise until you get the technologies in the hands of real-world users.

In the case of the Google Photos app — which uses a combination of advanced computer vision and machine learning techniques to help users collect, search and categorize photos — errors are easy to spot. When the app was unveiled at the company's annual developer show, executives went through carefully staged demonstrations to show how it can recognize landmarks like the Eiffel Tower and give users the ability to search their photos for people, places or things — even things as specific as a particular dog breed.

Google's Speech Recognition Has a Gender Bias

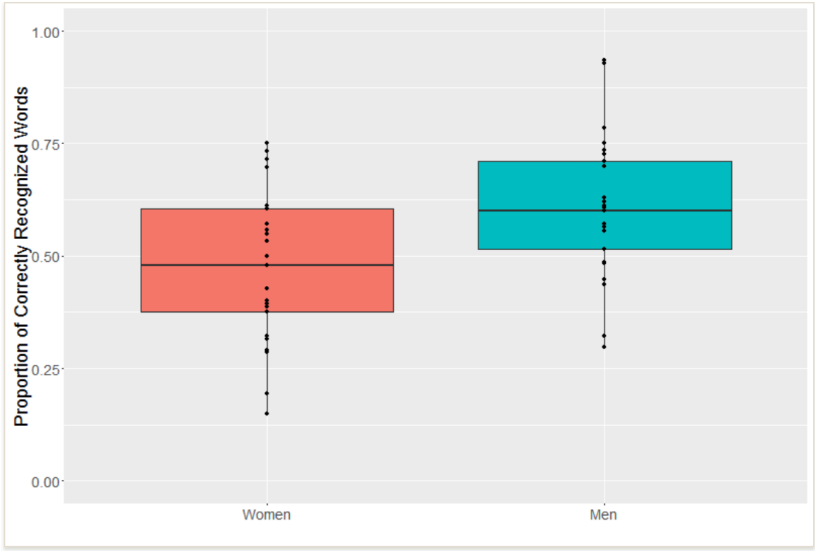
July 12, 2016

GOOGLE'S SPEECH RECOGNITION HAS A GENDER BIAS

Posted by **Rachael Tatman** in **Uncategorized** and tagged with **computational linguistics**, **gender**, **linguistics**, **sociolinguistics**, **speech recognition**, **speech signal**, **speech technology**

In my last post, I looked at how Google's automatic speech recognition worked with different dialects. To get this data, I hand-checked annotations more than 1500 words from fifty different accent tag videos .

Now, because I'm a sociolinguist and I know that it's important to stratify your samples, I made sure I had an equal number of male and female speakers for each dialect. And when I compared performance on male and female talkers, I found something deeply disturbing: YouTube's auto captions consistently performed better on male voices than female voice ($t(47) = -2.7$, $p < 0.01$.) . (You can see my data and analysis here.)



On average, for each female speaker less than half (47%) her words were captioned correctly. The average male speaker, on the other hand, was captioned correctly 60% of the time.

It's not that there's a consistent but small effect size, either, 13% is a pretty big effect. The Cohen's d was 0.7 which means, in non-math-speak, that if you pick a random

ProPublica's Study of NorthPointe Software

PRO PUBLICA

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

TayTweets: Microsoft's Twitter Bot

SECTIONS HOME SEARCH

The New York Times

TECHNOLOGY

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

By DANIEL VICTOR MARCH 24, 2016

Tay's Twitter account. The bot was developed by Microsoft's technology and research and Bing teams.

Science, Oct 2017

REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1]* Joanna J. Bryson,[1,2]* Arvind Narayanan[1]*

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture. The general idea that text corpora capture semantics, including cultural stereotypes and empirical associations, has long been known in corpus linguistics (*1*, *2*), but our findings add to this knowledge in three ways. First, we used word embeddings (*3*), a powerful tool to extract associations captured in text corpora; this method substantially amplifies the signal found in raw statistics. Second, our replication of documented human biases may yield tools and insights for studying prejudicial attitudes and behavior in humans. Third, since we performed our experiments on off-the-shelf machine learning components [primarily the Global Vectors for

response times when subjects are asked to pair two concepts they find similar, in contrast to two concepts they find different. We developed our first method, the Word-Embedding Association Test (WEAT), a statistical test analogous to the IAT, and applied it to a widely used semantic representation of words in AI, termed word embeddings. Word embeddings represent each word as a vector in a vector space of about 300 dimensions, based on the textual context in which the word is found. We used the distance between a pair of vectors (more precisely, their cosine similarity score, a measure of correlation) as analogous to reaction time in the IAT. The WEAT compares these vectors for the same set of words used by the IAT. We describe the WEAT in more detail below.

Most closely related to this paper is concurrent

the reaction latencies of four pairings (flowers + pleasant, insects + unpleasant, flowers + unpleasant, and insects + pleasant). Greenwald *et al.* measured effect size in terms of Cohen's *d*, which is the difference between two means of log-transformed latencies in milliseconds, divided by the standard deviation. Conventional small, medium, and large values of *d* are 0.2, 0.5, and 0.8, respectively. With 32 participants, the IAT comparing flowers and insects resulted in an effect size of 1.35 ($P < 10^{-8}$). Applying our method, we observed the same expected association with an effect size of 1.50 ($P < 10^{-7}$). Similarly, we replicated Greenwald *et al.*'s finding (*5*) that musical instruments are significantly more pleasant than weapons (see Table 1).

Notice that the word embeddings "know" these properties of flowers, insects, musical instruments, and weapons with no direct experience of the world and no representation of semantics other than the implicit metrics of words' co-occurrence statistics with other nearby words.

We then used the same technique to demonstrate that machine learning absorbs stereotyped biases as easily as any other. Greenwald *et al.* (*5*) found extreme effects of race as indicated simply by name. A bundle of names associated with being European American was found to be significantly more easily associated with pleasant than unpleasant terms, compared with a bundle of African-American names.

In replicating this result, we were forced to slightly alter the stimuli because some of the original African-American names did not occur in the corpus with sufficient frequency to be included. We therefore also deleted the same number of European-American names, chosen at random, to balance the number of elements in the sets of two concepts. Omissions and deletions are indicated in our list of keywords (see the supplementary materials).

In another widely publicized study, Bertrand and Mullainathan (*7*) sent nearly 5000 identical résumés in response to 1300 job advertisements, varying only the names of the candidates. They

NIPS, Dec 2016

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Bias in computer systems
(Friedman & Nissenbaum, 1996)

- Identified three sources of bias
    1. Preexisting bias
    2. Technical bias
    3. Emergent bias
- “We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.”

How do computer scientists define fairness?

- Probabilistically
- Lots of parity (i.e., “fairness”) definitions
  - Decisions should be in some sense probabilistically independent of sensitive features values (such as gender, race)
  - There are many possible senses

Confusion matrix

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

- **Accuracy**: How often is the classifier correct? *(TP+TN)/total*
- **Misclassification (**a.k.a. **Error) Rate**: How often is it wrong? *(FP+FN)/total*
- **True Positive Rate** (**TPR**, a.k.a. **Sensitivity** or **Recall**): When it's actually yes, how often does it predict yes? *TP/actual yes*
- **False Positive Rate** (**FPR**) : When it's actually no, how often does it predict yes? *FP/actual no*
- **Specificity** (**1 – FPR**) : When it's actually no, how often does it predict no? *TN/actual no*
- **Precision (**a.k.a. **Positive Predictive Value)**: When it predicts yes, how often is it correct? TP/predicted yes
- **Negative Predictive Value:** When it predicts no, how often is it correct? TN/predicted no
- **Prevalence**: How often does the yes condition actually occur in our sample? actual yes/total

http://bit.ly/2xApsRz

Lots of parity definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity
- Accuracy parity
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Positive predictive value parity
- Negative predictive value parity
- Predictive value parity
- ...

See https://fairmlclass.github.io
for definitions.

Impossibility results ☹

- Kleinberg, Mullainathan, Raghavan (2016)
- Chouldechova (2016)
- You can’t have your cake and eat it too

Some definitions

- X contains features of an individual (e.g., medical records)
    - X incorporates all sorts of measurement biases
- A is a sensitive attribute (e.g., race, gender, ...)
    - A is often unknown, ill-defined, misreported, or inferred
- Y is the true outcome (a.k.a. the ground truth; e.g., whether patient has cancer)
- C is the machine learning algorithm that uses X and A to predict the value of Y (e.g., predict whether the patient has cancer)

Some simplifying assumptions

- The sensitive attribute A divides the population into two groups a (e.g., whites) and b (e.g., non-whites)
- The machine learning algorithm C outputs 0 (e.g., predicts not cancer) or 1 (e.g., predicts cancer)
- The true outcome Y is 0 (e.g., not cancer) or 1 (e.g., cancer)

Impossibility results

- Kleinberg, Mullainathan, Raghavan (2016), Chouldechova (2016)
- Assume differing base rates – i.e., $Pr_a(Y=1) \neq Pr_b(Y=1)$ – **and** an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve
  a) Precision parity: $Pr_a(Y=1 \mid C=1) = Pr_b(Y=1 \mid C=1)$.
  b) True positive parity: $Pr_a(C=1 \mid Y=1) = Pr_b(C=1 \mid Y=1)$
  c) False positive parity: $Pr_a(C=1 \mid Y=0) = Pr_b(C=1 \mid Y=0)$

Impossibility results

- Kleinberg, Mullainathan, Raghavan (2016), Chouldechova (2016)
- Assume differing base rates – i.e., $Pr_a(Y=1) \neq Pr_b(Y=1)$ – **and** an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve

  a) Precision parity: $Pr_a(Y=1 \mid C=1) = Pr_b(Y=1 \mid C=1)$

  b) True positive parity: $Pr_a(C=1 \mid Y=1) = Pr_b(C=1 \mid Y=1)$

  c) False positive parity: $Pr_a(C=1 \mid Y=0) = Pr_b(C=1 \mid Y=0)$

"Equalized odds" -- Hardt, Price, Srebro (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."
-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

"Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups."

-- Kleinberg, Mullainathan, Raghavan (2016)

ProPublica and NorthPointe

- ProPublica's main charge was that black defendants experienced higher false positive rate
- Northpointe's main defense was that their risk assessment scores satisfy precision parity: $Pr_a(Y=1 \mid C=1) = Pr_b(Y=1 \mid C=1)$
- Due to the impossibility results, Northpointe's algorithm cannot satisfy "equalized odds"
  - Disproportionately high false positive rate for blacks
  - Disproportionately high false negative rate for whites

Group vs. individual fairness

- Fairness through awareness by Dwork, Hardt, Pitassi, Reingold, Zemel (2012)
- “People who are similar w.r.t. a specific (classification) task should be treated similarity.”
- Does not get around the impossibility results
- Assuming you have equal base rates, treating everyone equally is a good move

Solutions considered from the machine learning side

- Preprocessing or “massaging” the data to make it less biased
- Learning fair representations: encode data while obfuscating sensitive attributes
- Penalize the algorithm to encourage it to learn fairly
  - During training (e.g., through regularization or constraints) or as a post-processing step
- Allow the sensitive attributes to be used during training, but do not make them available to the model during inference time
- Causal inference
  - Caution: powerful but brittle

Solutions considered from the policy side

- Regulations 😱
- The EU has General Data Protections Regulation (GDPR) data laws going into effect in 2018
- These laws grant users the right to a logical explanation of how an algorithm uses our personal data
- Wikipedia entry: http://bit.ly/1lmrNJz

Just machine learning in an unjust world?

- Racist/sexist humans – e.g., biased judges
- Stupid algorithms are already in use – e.g., three-strikes laws, mandatory minimum sentencing
  - They don’t take enough empirical data into account
  - Machine learning can help here
    - Personalization, context-awareness, …

Where do we go from here?

“Computers may be intelligent, but they are not wise. Everything they know, we taught them, and we taught them our biases. They are not going to un-learn them without transparency and corrective action by humans.”
-- Ellora Thadaney Israni

https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

An interdisciplinary call for action

- You can’t have all the different kinds of fairness that you might want
  - Recall the impossibility results
- We need to work together across disciplines to reach agreement in terms of which kinds of “fairness” we want to optimize
  - Fairness based on explanation?
  - Fairness based on placement?

Thank you

Slides at http://eliassi.org/safra17.pdf