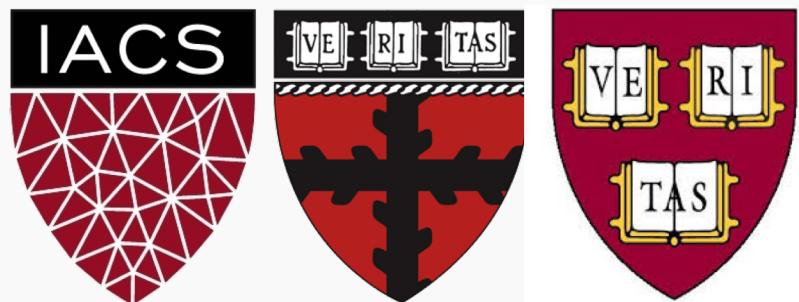


Lecture #12: Responsible Data Science and Wrap-Up

CS-S109A: Introduction to Data Science
Kevin Rader



HARVARD
Summer School

ANNOUNCEMENTS

- The **Final Exam** (individual work) is available and due Monday, Aug. 3.
 - There is a 1% penalty if handed in up to 24 hours late (max = 99/100)
- A few clarifications:
1. Unless specifically mentioned as departure delay, and reference to a delay should be assumed to be an arrival delay.
 2. Determining whether a flight is delayed or not it always whether the actual time of departure/arrival is 15 minutes or more after its scheduled time.
- **Friday's lab** will include general review: both conceptual and technical (partially based on today's **quiz**)



Outline

- Ethical Data Science
- Case Study: Bias in the COMPAS algorithm
- Wrap-Up



Ethical Data Science

*slides modified by CS 109's guest speaker in Fall 2018:

Julia Stoyanovich, NYU

<https://engineering.nyu.edu/faculty/julia-stoyanovich>



Online price discrimination

THE WALL STREET JOURNAL.

WHAT THEY KNOW

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.

WHAT PRICE WOULD YOU SEE?



lower prices offered to buyers who live in more affluent neighborhoods

<https://www.wsj.com/articles/SB10001424127887323777204578189391813881534>



CS-S109A: RADER

Online job ads



Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



One experiment showed that Google displayed adverts for a career coaching service for “\$200k+” executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

<https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study>

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for “\$200k+” executive jobs 1,852 times to the male group and only 318 times to the female group. Another experiment, in July 2014, showed a similar trend but was not statistically significant.



CS-S109A: RADER

Racial bias in criminal sentencing

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



A commercial tool COMPAS automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

The tool correctly predicts recidivism 61% of the time.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Blacks are almost twice as likely as whites to be labeled



Is data science impartial?

Data science is algorithmic, therefore it cannot be biased! And yet...

- All traditional evils of discrimination, and many new ones, exhibit themselves in the data science eco system
- Bias that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- Transparency helps prevent discrimination, enable public debate, establish trust
- Technology alone won't do: also need policy, user involvement and education



Data, responsibly

Because of its power, data science must be used responsibly



fairness



transparency

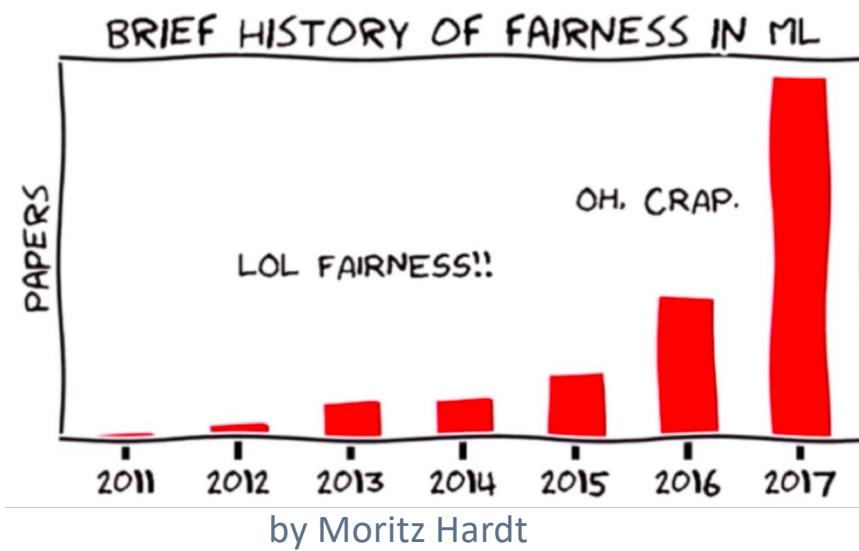


diversity



data protection

Fairness in ML



Fairness: a lack of “bias”

What are the tasks we are interested in?

- predictive analytics and inferential interpretations

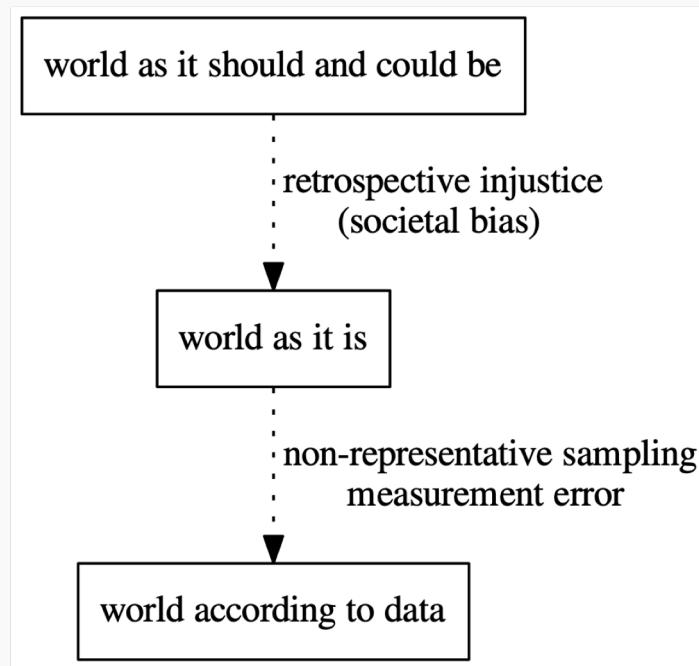


What do we mean by bias?

- **statistical bias**: a model is biased if it doesn't summarize the data correctly
- **societal bias**: a dataset or a model is biased if it does not represent the world “correctly”, e.g., data is not representative, there is measurement error, or the world is “incorrect”

Depends on your perspective: the world as it is or as it should be?

“Biased data”



from “Prediction-Based Decisions and Fairness” by Mitchell, Potash and Barocas, 2018

when data is about people, bias can lead to discrimination



The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employee, differently based on a protected characteristic such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact is the result of systematic disparate treatment, where disproportionate adverse impact is observed on members of a protected class.

- In the US, disparate impact measured by comparing the rate of positive outcomes for the protected class divided by rate of positive outcomes for the population as a whole.
- If that ratio is less than 80%, it is considered discriminatory



Vendors and outcomes

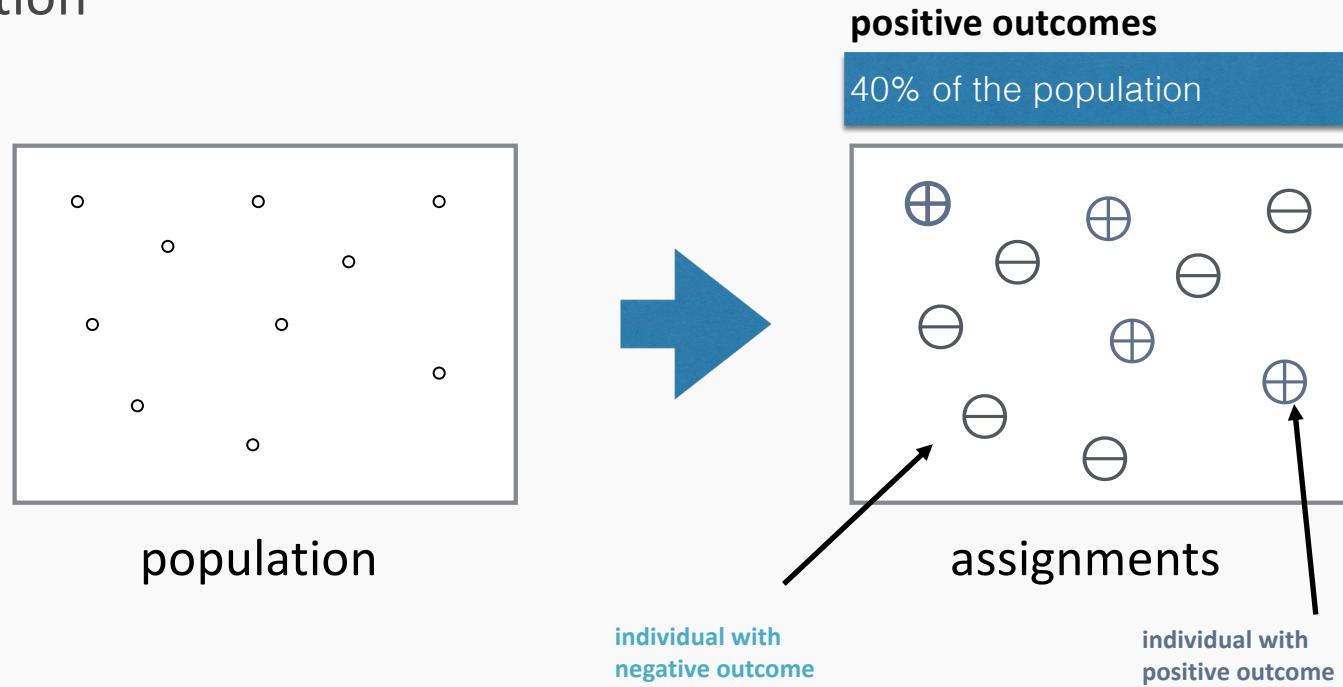
Consider a vendor assigning positive or negative outcomes to individuals.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount



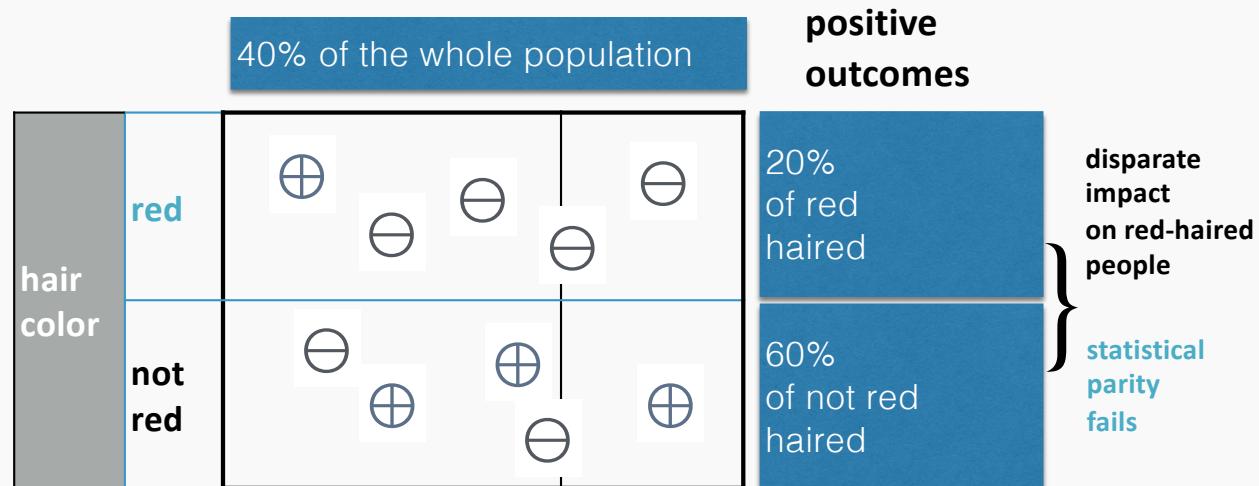
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



Sub-populations may be treated differently

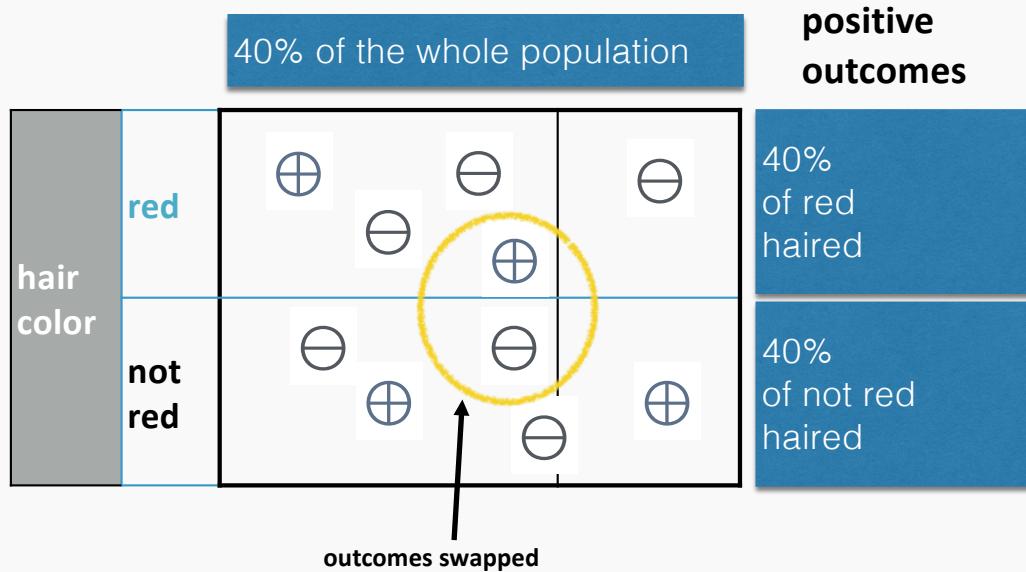
Sub-population: those with red hair
(under the same assignment of outcomes)



Statistical parity

Statistical parity (a popular group fairness measure):

demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

Now consider the assignments under both
hair color (protected) and **hair length** (innocuous)

		hair length		positive outcomes
		long	not long	
hair color	red	\oplus	\ominus \ominus \ominus \ominus	
	not red	\oplus \oplus \oplus	\ominus \ominus	

Deniability
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

20% of red haired
60% of not red haired



Blinding is not an excuse

19

Removing **hair color** from the vendor's assignment process does not prevent discrimination! This is upheld in the legal system.



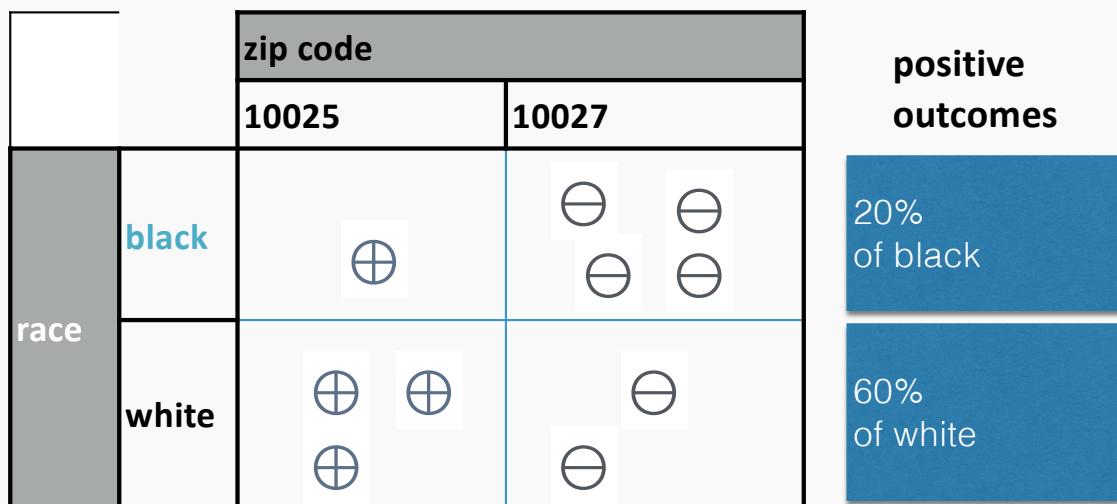
Assessing disparate impact

Discrimination is assessed by the effect on the protected sub-population, not by the input or by the process that lead to the effect.



Redundant encoding

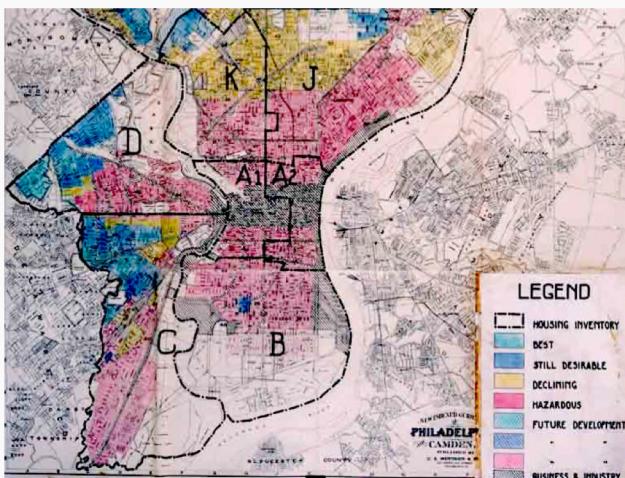
Let's replace hair color with race (protected), and hair length with zip code (innocuous)



Redlining

Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



wikipedia

Households and businesses in the red zones could not get mortgages or business loans.



Imposing statistical parity

May be contrary to the goals of the vendor

positive outcome: offered a loan



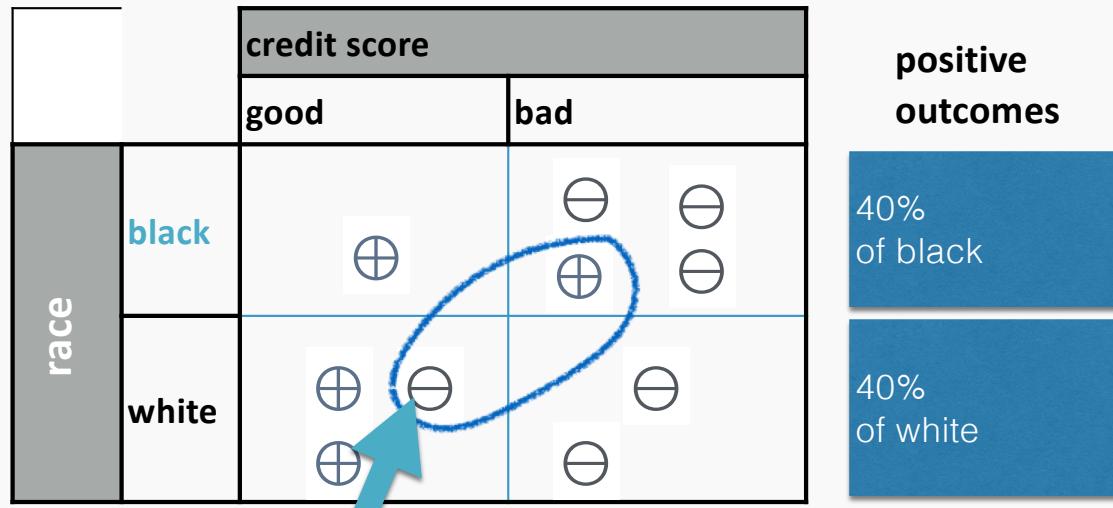
Impossible to predict loan payback accurately.
Use past information, which may itself be biased.



Is statistical parity sufficient?

Statistical parity (a popular group fairness measure)

demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Individual fairness

any two individuals who are similar w.r.t. a particular task should receive similar outcomes



Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By ADAM LIPTAK JUNE 29, 2009

The New York Times



Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

Case opinions

- Majority** Kennedy, joined by Roberts, Scalia, Thomas, Alito
Concurrence Scalia
Concurrence Alito, joined by Scalia, Thomas
Dissent Ginsburg, joined by Stevens, Souter, Breyer

Laws applied

Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e^{et seq.}



CS-S109A: RADER

Two notions of fairness

individual fairness



equality

group fairness



equity

two intrinsically different world views



On the (im)possibility of fairness

2 intrinsically different worlds views

- What you see is what you get (WYSIWYG) - individual fairness
- We are all equal (WAE) - group fairness

Construct Space (CS)	Observed Space (OS)	Decision Space (DS)
intelligence	SAT score	
grit	high-school GPA	performance in college
propensity to commit crime	family history	recidivism
risk-averseness	age	

Goal: tease out the difference between beliefs about fairness and mechanisms that logically follow from those beliefs.



Racially identifying names

[Latanya Sweeney; CACM 2013]

The screenshot shows a Google AdSense page. On the left, there's a large Google AdSense logo. In the center, there are two ads. The top ad is for "Latanya Sweeney, Arrested?" from instantcheckmate.com, which includes a link to www.instantcheckmate.com/. The bottom ad is for "Latanya Sweeney" from publicrecords.com, which includes a link to www.publicrecords.com/. To the right of these ads is a screenshot of the instantcheckmate website for "LATANYA SWEENEY". It shows her profile picture, address (1420 Centre Ave, Pittsburgh, PA 15219), DOB (Oct 27, 1959), and a "CERTIFIED" badge. Below this is a "Criminal History" section stating she has never been arrested. At the bottom is a table titled "Possible Matching Arrest Records" with one row: "No matching arrest records found."

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

racially identifying names trigger ads suggestive of a criminal record

<https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>



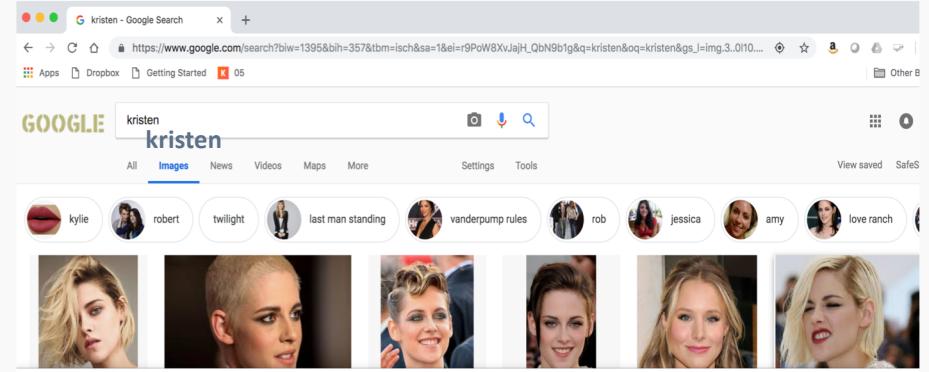
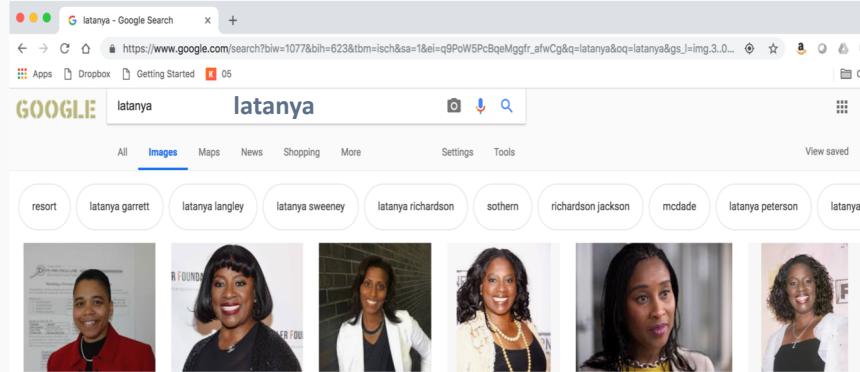
Observations

[Latanya Sweeney; CACM 2013]

Ads suggestive of a criminal record, linking to Instant Checkmate, appear on google.com and reuters.com in response to searches for “Latanya Sweeney”, “Latanya Farrell” and “Latanya Lockett”*

No Instant Checkmate ads when searching for “Kristen Haring”, “Kristen Sparrow”* and “Kristen Lindquist”*

* next to a name associated with an actual arrest record



Why is this happening?

[Latanya Sweeney; CACM 2013]

Possible explanations (from Latanya Sweeney):

- Does Instant Checkmate serve ads specifically for black-identifying names?
- Is Google's Adsense explicitly biased in this way?
- Does Google's Adsense learn racial bias based on click-through rates?

How do we know which explanation is right?

We need transparency!



Codes of ethics

The screenshot shows the ACM (Association for Computing Machinery) website. At the top, there's a navigation bar with links to Digital Library, CACM, Queue, TechNews, Learning Center, and Career Center. Below that is a secondary navigation bar with links to Join, Volunteer, myACM, and Search. The main menu includes About ACM, Membership, Publications, Special Interest Groups, Conferences, Chapters, Awards, Education, Public Policy, and Governance. A breadcrumb trail indicates the user is at Home > Code Of Ethics. The main content area features a large banner with the text "ACM Code of Ethics and Professional Conduct". Below the banner, the title "ACM Code of Ethics and Professional Conduct" is displayed, followed by a "Preamble" section. The Preamble states: "Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession." It goes on to explain that the Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. The Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle. Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, and principles involving compliance with the Code are given in Section 4. The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making. When thinking through a particular issue, a computing professional may find that multiple principles should be taken into account, and that different principles will have different relevance to the issue. Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration. The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.

[PDF of the ACM Code of Ethics](#)

On This Page

- Preamble
- 1. GENERAL ETHICAL PRINCIPLES.
 - 1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
 - 1.2 Avoid harm.
 - 1.3 Be honest and trustworthy.
 - 1.4 Be fair and take action not to discriminate.
 - 1.5 Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
 - 1.6 Respect privacy.
 - 1.7 Honor confidentiality.
- 2. PROFESSIONAL RESPONSIBILITIES.
 - 2.1 Strive to achieve high quality in both the processes and products of professional work.
 - 2.2 Maintain high standards of



Codes of ethics

The screenshot shows a mobile application interface for the "Community Principles on Ethical Data Practices". The background is dark blue. At the top left is a "BACK" button with a circular arrow icon. In the center is a white circle containing a blue scales of justice icon. Below the icon, the title "Community Principles on Ethical Data Practices" is displayed in white. To the right of the title is a "SUBSCRIBE" button with a blue outline. A text block explains the purpose of the code of ethics: "This code of ethics for data sharing is created and proposed for adoption by the data science community to reflect the behaviors and principles for the responsible and ethical use and sharing of data by data scientists." Below this, another text block encourages community participation: "As a community-driven crowdsourced effort, you can join the discussion and contribute to the next version of the Community Principles on Ethical Data Sharing." A small box contains a link: "NSF contacts - Google Docs docs.google.com/document/d/.../edit". The word "OVERVIEW" is centered below the text blocks. To the right of the overview text is a large block of explanatory text about the development of the principles. At the bottom right are social media sharing icons for LinkedIn, Facebook, and Twitter. On the left side of the screen, there is a vertical navigation menu with links: "OVERVIEW", "BACKGROUND", "VALUES", "PRINCIPLES", "AUTHORS", and "SIGNATORIES". Below this menu are two blue rounded rectangular buttons labeled "SIGN" and "JOIN".

This code of ethics for data sharing is created and proposed for adoption by the data science community to reflect the behaviors and principles for the responsible and ethical use and sharing of data by data scientists.

As a community-driven crowdsourced effort, you can join the discussion and contribute to the next version of the Community Principles on Ethical Data Sharing.

NSF contacts - Google Docs
docs.google.com/document/d/.../edit

OVERVIEW

The Community Principles on Ethical Data Practices are being developed by people from the data science community in conjunction with data science organizations. These principles focus on defining ethical and responsible behaviors for sourcing, sharing and implementing data in a manner that will cause no harm and maximize positive impact. The goal of this initiative is to develop a community-driven code of ethics for data collection, sharing and utilization that provides people in the data science community a standard set of easily digestible, recognizable principles for guiding their behaviors.

This code is not intended to be all encompassing. Rather, these principles will provide academia, industry, and individual data scientists a common set of guidelines for driving the development of standards, curriculums, and best practices for the ethical use and sharing of data, ultimately advancing the responsible and ethical use of data as a collective force for good.



Case Study: Bias in the COMPAS algorithm



Racial bias in criminal sentencing

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



Bernard Parker, left, was rated high risk; Dylan Fuggett was rated low risk. (Josh Ritchie for ProPublica)

ProPublica is “an independent, non-profit newsroom that produces investigative journalism in the public interest.”

ProPublica investigated COMPAS, a commercial tool that automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



Automated risk assessment

Goal: predict the likelihood of some category of future crime

Used by judges to assist in (1) assigning bail amounts and (2) sentencing

Intended use of the algorithm was **different** - to provide supportive interventions!



Algorithm

Input: attributes of individuals, based on a 137-question questionnaire answered by defendants. Questions are from the following categories:

- current charges, criminal history, non-compliance history, family criminality, peers, substance abuse, residence/stability, social environment, education, vocation, leisure/recreation, social isolation, criminal personality, anger, criminal attitudes
- race is not a feature!



Example questions

- “Was one of your parents ever sent to jail or prison?”
- “How many of your friends/acquaintances are taking drugs illegally?”
- “How often did you get into fights while at school?”
- Agree or disagree with statements such as:
 - “A hungry person has a right to steal.” and
 - “If people make me angry or lose my temper, I can be dangerous.”



Training Data

The original training data were based on a subset of Michigan and New York inmates and parolees. Even though it was trained on mostly male observations, the score is said to be “gender-neutral.”

The algorithm/score was revised and tuned to a new training group: “normative data were sampled from over 30,000 COMPAS Core assessments conducted between January 2004 and November 2005 at prison, parole, jail and probation sites across the United States.” This is the normative group that the risk assessments are based on today.

Output: risk score [1,10]



- risk for general recidivism

Output Score

Output: risk score [1,10]

- risk for general recidivism
- risk for violent recidivism

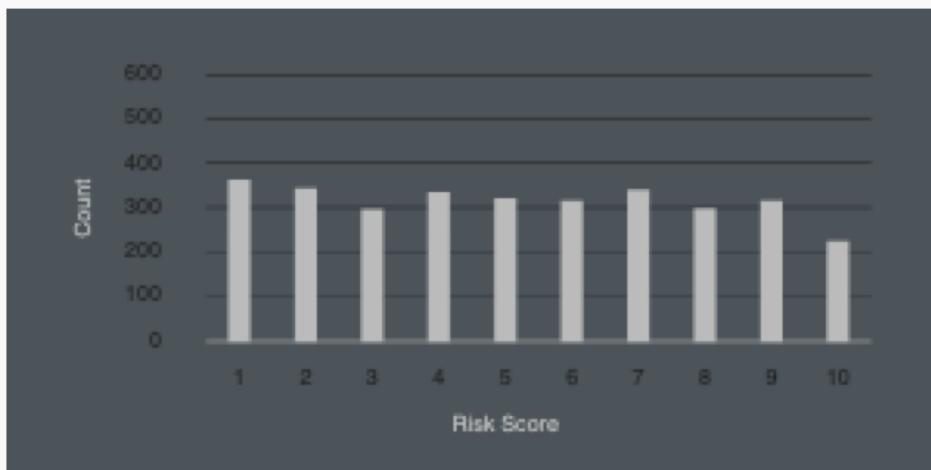
These scores are often put into 2 or 3 categories of low-risk, medium-risk, and high-risk of recidivism.

The predicted risk groups are then assessed via the standard false positive and false negative rates, and the AUC score is used to tune the output and compare to other predictive models.

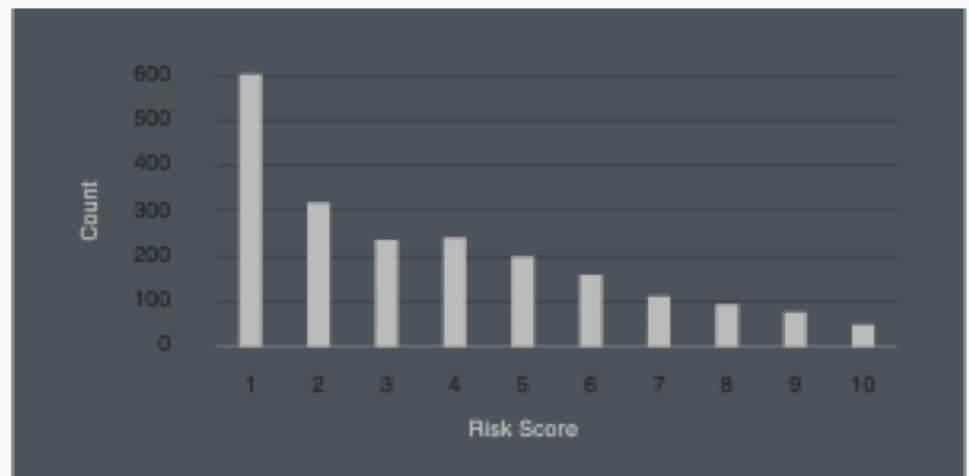


Risk Assessment Results (in Broward County, Fla.)

Black Defendants' Risk Scores



White Defendants' Risk Scores



Recidivism Results (in Broward County, Fla.)

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%



Discussion Questions:

1. Is COMPAS a fair algorithm? Think about this from both the individual and group perspective.
2. How can the COMPAS algorithm predict different rates of recidivism for racial groups even though race was not an input into the prediction model? (How) can this be corrected?
3. What factors could be causing the observed differences



Course Wrap-Up



Modules

The semester has been organized into 4 major ‘modules’:

- Module 0: Intro to Data and Data Science (and Python)
- Module 1: Regression
- Module 2: Classification
- Module 3: Ensemble Methods

We have learned various approaches to perform both predictions and inferences within each of these frameworks.



Module 1: Regression Methods

When is it appropriate to perform a regression method? What regression models have we learned?

1. Linear Regression (simple, multiple, polynomial, interactions, model selection, Ridge & Lasso, etc...)
2. k -NN
3. Regression Trees

What is the main difference between these types of models?



Module 2: Classification Methods

When is it appropriate to perform a classification method? What classification models have we learned?

1. Logistic Regression: same details as linear regression apply
2. k -NN
3. Classification Trees

What is the main difference between these types of models (advantages and disadvantages)? When should you use each method?



Module 3: Ensemble Methods

What does it mean for a model to be an ensemble method?

1. Bagging Trees
2. Random Forests
3. Boosting Models
4. Stacking Models
5. Neural Networks

What approach does each model take to improve prediction accuracy?



Choosing between Models

How can we choose between our various methods/models to answer a question at hand? What approaches/measures can we use to make this determination?

1. In-sample: AIC, BIC (barely mentioned)
2. Out-of-sample: Cross-Validation

What measure(s) should we use when we perform cross-validation?



Dealing with Data Issues

What issues have arisen when dealing with real data? How have we handled them?

1. Categorical Predictors: might make sense to one-hot encode
2. Missing Data: might make sense to impute
3. High Dimensionality: might make sense to use a data reduction technique.
4. Too many observations: do preliminary analysis on a subset

How are predictions affected? How are inferences affected?



Dealing with High Dimensionality

What does ‘high dimensionality’ mean? What issues arise when this happens? How can we handle it?

1. Model Selection: subset variable selection
2. Regularization: LASSO and Ridge like approaches (penalize the loss function)
3. PCA: create new predictor variables that encapsulate the ‘essence’ of all your predictor data with a minimal number of variables.

How can we compare methods to determine which approach is best?



Other things we've learned

- Scraping, Data Gathering, Data Wrangling
- EDA: Visualization and Summary Statistics
- t -tests and p -values: probabilistic/ approaches to perform inferences
- Bootstrapping: empirical approach to perform inferences
- Misclassification Rates, Types of Errors, Confusion Matrices/Tables, and ROC Curves
- Bias-Variance Trade-off
- Train vs. Test vs. Validation
- Standardization vs. Normalization. When should we do it?
- Data Ethics

Anything lingering questions or thoughts?



Other things we haven't discussed

There are lots of topics we have not covered in one semester...some are covered in 109B in the Spring:

- AB testing and Causal Inference
- Unsupervised Classification/Clustering
- Smoothers
- Bayesian Data Analysis
- Reinforcement Learning
- Other versions of Neural Networks (and ‘Deep Learning’)
- Interactive Visualizations
- Database Management (SQL, etc.)
- Cloud Computing and Scaling (AWS)
- And much, much more...



Courses Related to Data Science

- CS 109B: Advanced Topics in Data Science
- CS 171: Visualizations
- CS 181/281: Machine Learning
- CS 182: Artificial Intelligence (AI)
- CS 205: Distributive Computing
- Stat 110/210: Probability Theory
- Stat 111/211: Statistical Inference
- Stat 139: Linear Models
- Stat 149: Generalized Linear Models
- Stat 195: Intro to Statistical Machine Learning

This list is not exhaustive!



What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results



Thanks for all your hard work!

It's been a long semester for everyone involved. Thank you for your patience, your hard work, and your commitment to data science!

It's sad to see you go...



CS-S109A: RADER

