



DSCI 531

Prof. Kristina Lerman

Spring 2025

Outline

- Course description
 - Why Fairness and Bias in AI?
 - Examples of bias in AI
 - What is ethics?
- Course goals
 - What you will get out of the course
 - Description of topics
- Course details
 - Contact
 - Workload
 - Grading, etc.

AI is Everywhere

Democracy



News ranking algorithms

- Does the algorithm create filter bubbles?
- Does the algorithm disproportionately censor content?



Algorithmic justice

- Does the algorithm discriminate against a racial group in granting parole?
- Does a predictive policing system increase the false conviction rate?

Markets



Algorithmic trading

- Do algorithms manipulate markets?
- Does the behaviour of the algorithm increase systemic risk of market crash?



Algorithmic pricing

- Do algorithms of competitors collude to fix prices?
- Does the algorithm exhibit price discrimination?

Kinetics



Autonomous vehicles

- How aggressively does the car overtake other vehicles?
- How does the car distribute risk between passengers and pedestrians?



Autonomous weapons

- Does the weapon respect necessity and proportionality in its use of force?
- Does the weapon distinguish between combatants and civilians?

Society



Online dating

- Does the matching algorithm use facial features?
- Does the matching algorithm amplify or reduce homophily?



Conversational robots

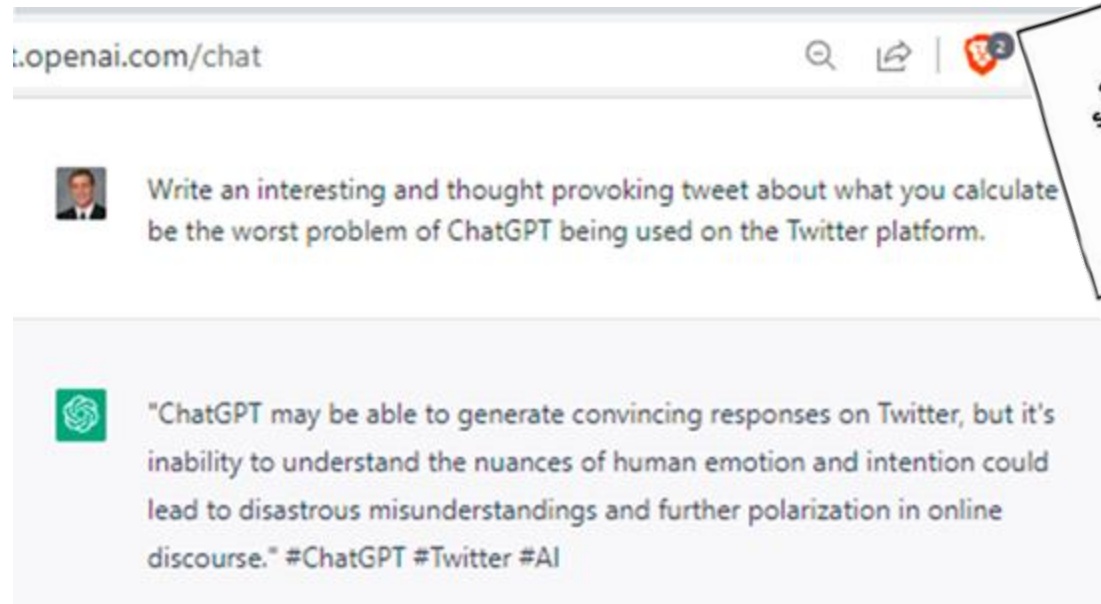
- Does the robot promote products to children?
- Does the algorithm affect collective behaviours?



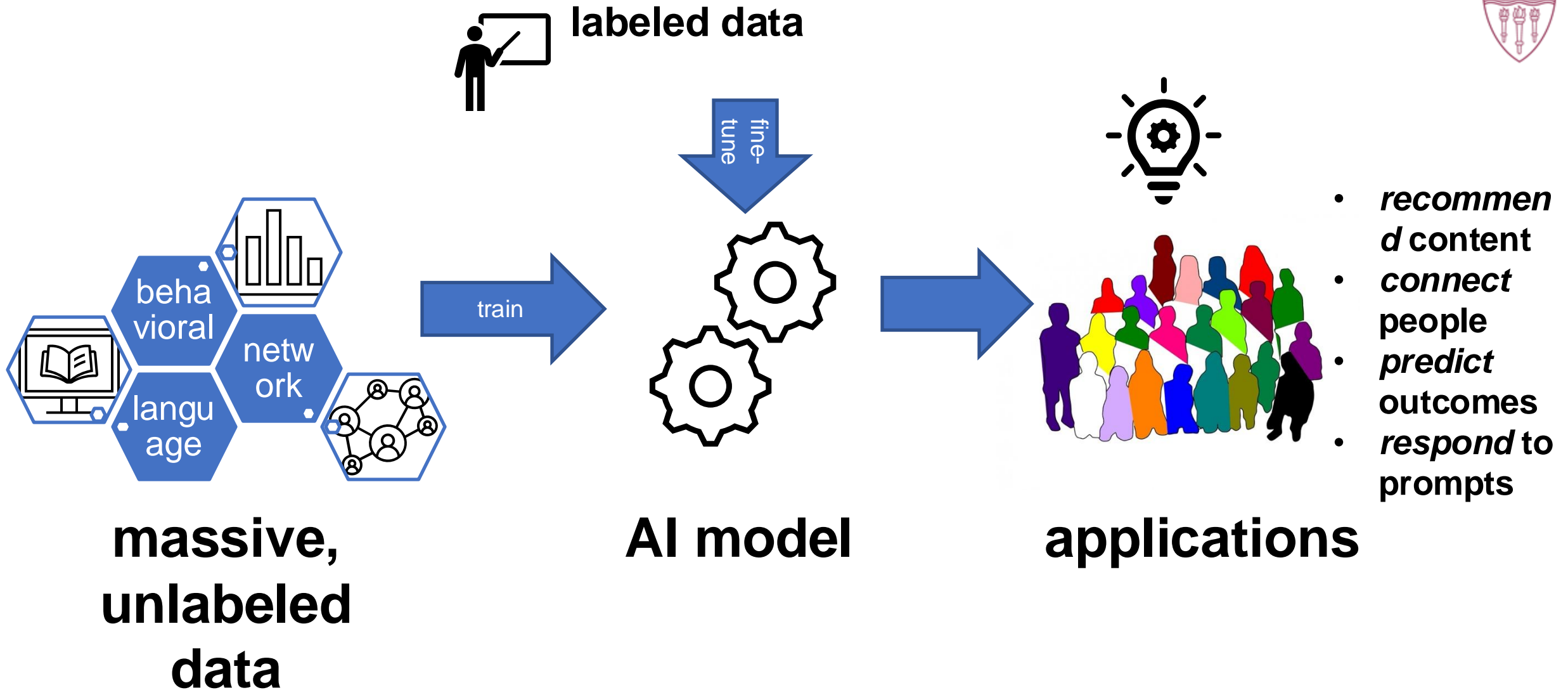
AI is Everywhere: Generative AI

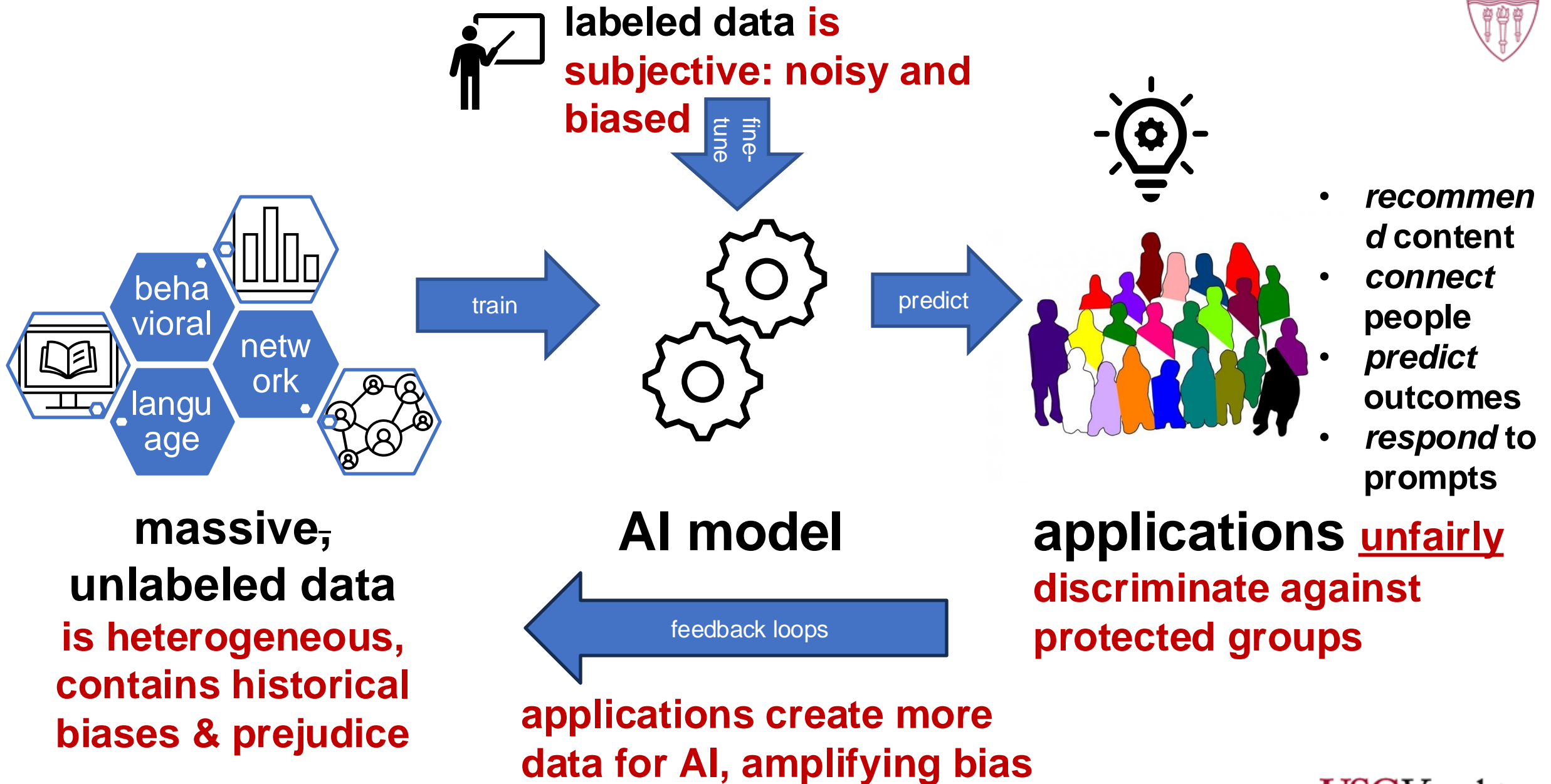
Text generation: GPT3, ChatGPT can generate essays and responses in any style

Image generation: DALL-E, Stable diffusion can generate art in any style

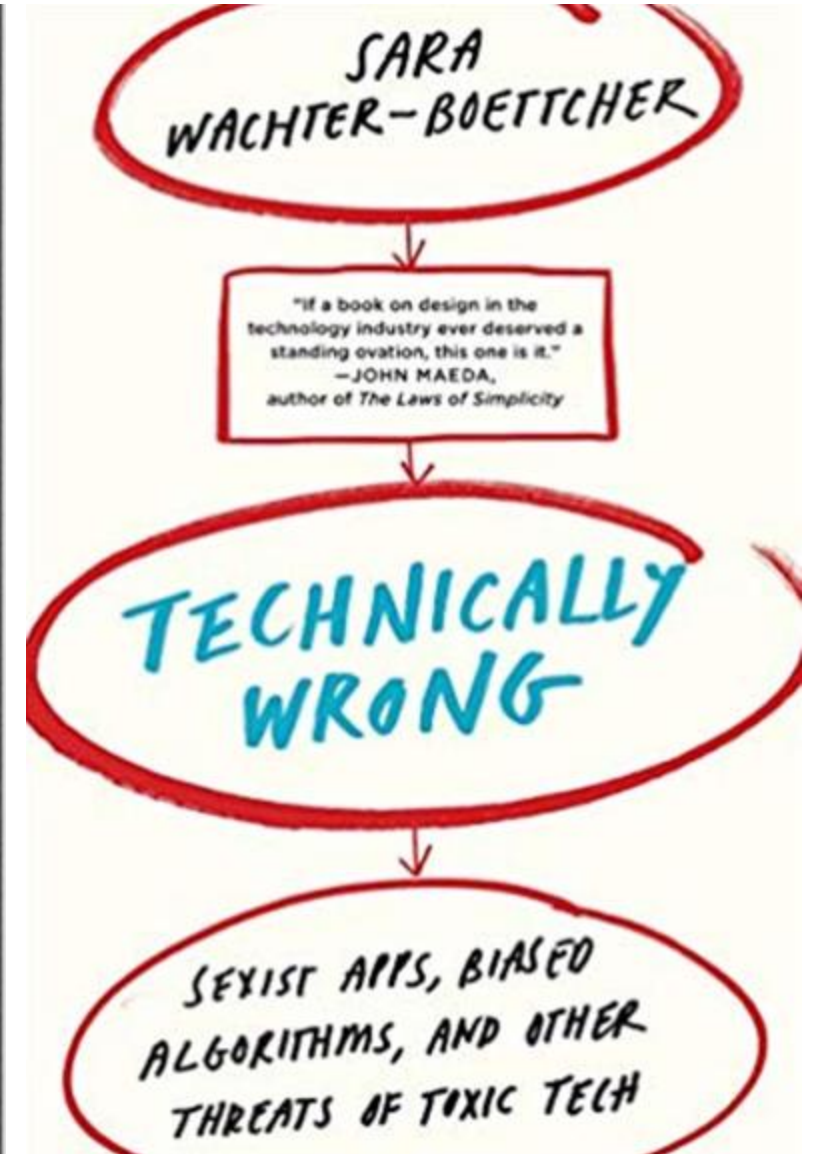
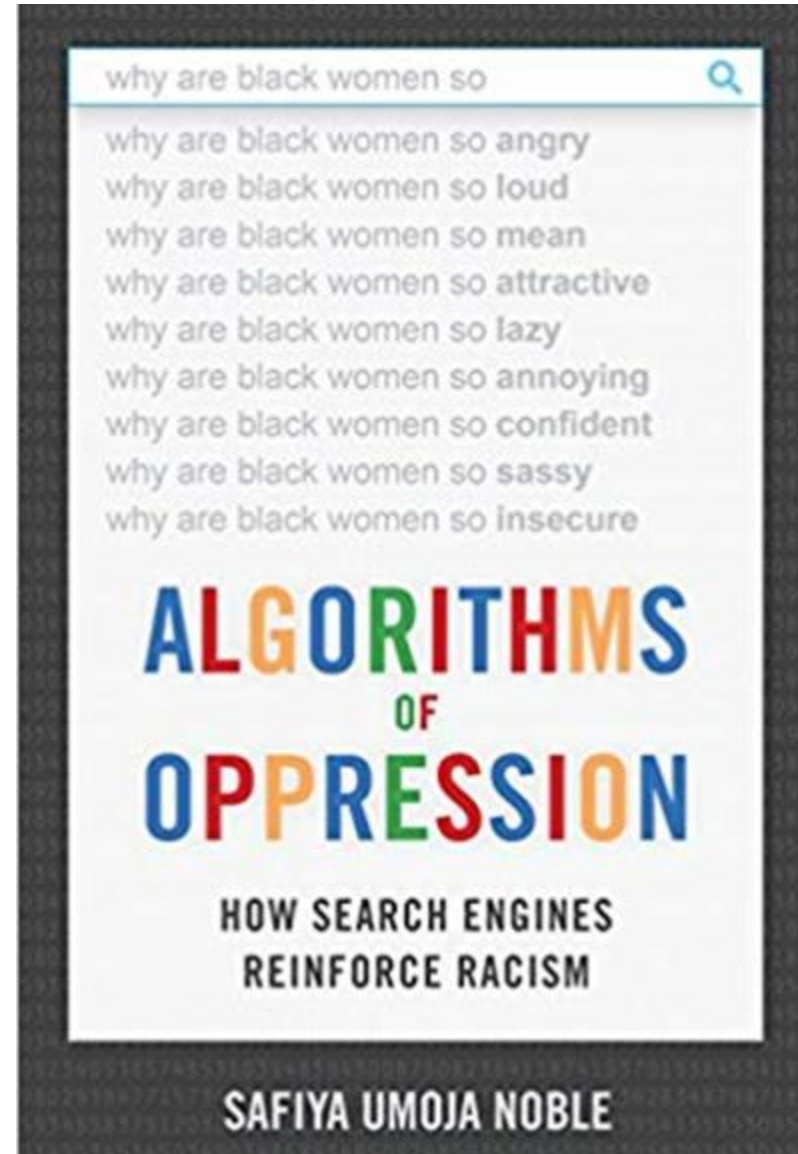
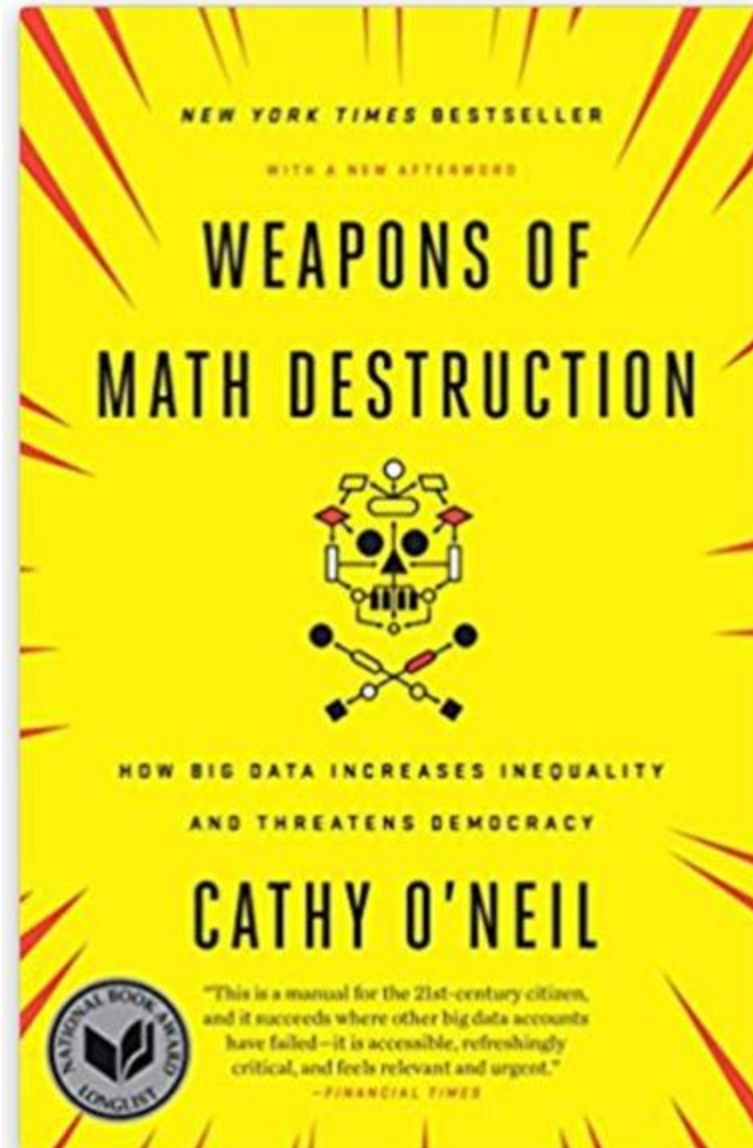


<https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-an-algorithm-took-my-work.html>





What could go wrong?





Discriminating via AI - some examples



unprofessional hairstyles



professional hairstyles



AI and Bias



“Unprofessional hairstyles” vs “Professional hairstyles” in Google image search (circa 2018)

AI and Bias

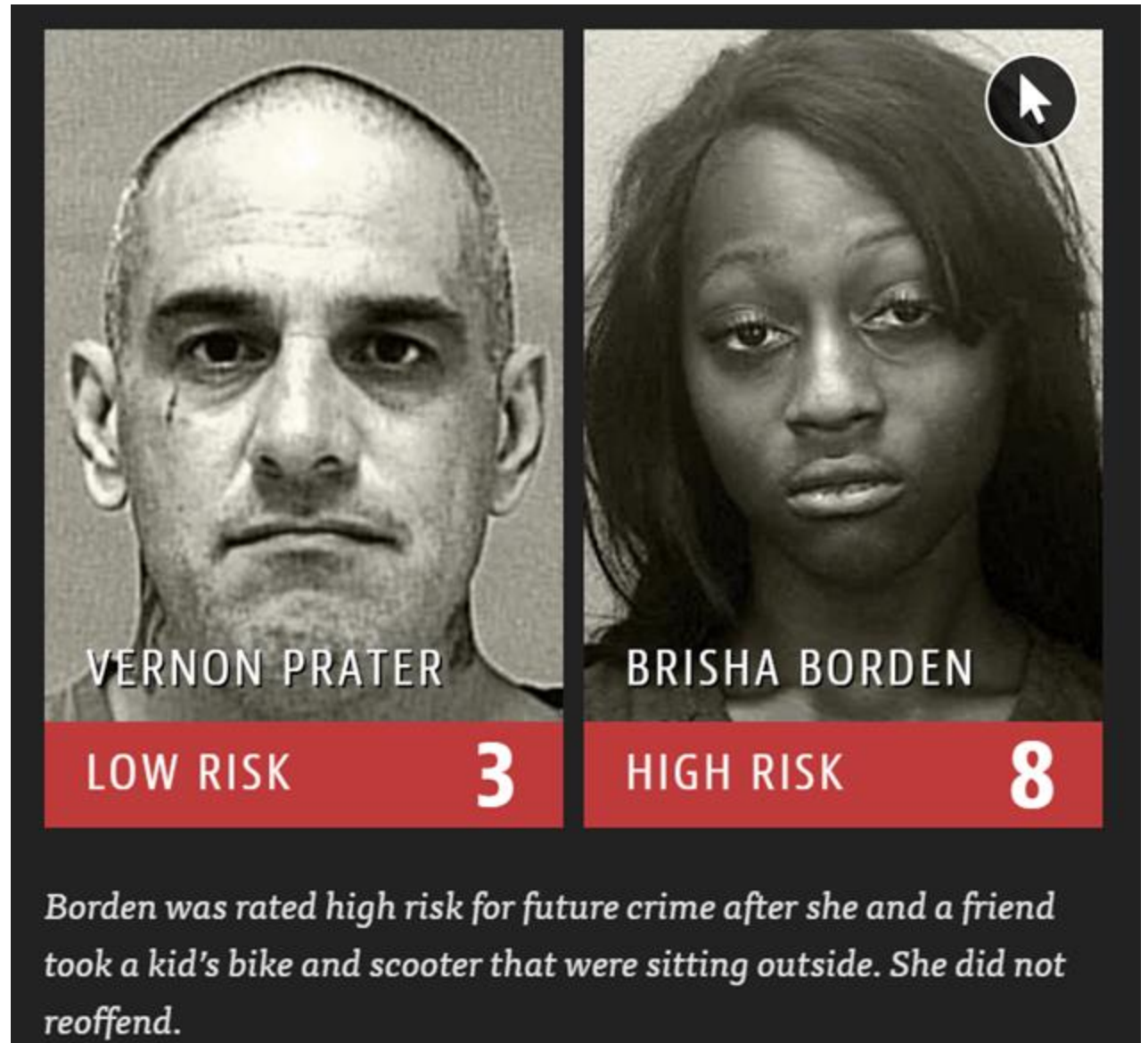


Joy Buolamwini

Facial recognition

Bias in automated criminal risk assessment

COMPAS tool systematically
gives black defendants higher
risk scores for future
recidivism



The image displays two mugshot-style photographs side-by-side. The left photograph is of a white male, Vernon Prater, with a low risk score of 3. The right photograph is of a black female, Brisha Borden, with a high risk score of 8. Below the photographs, a red bar contains the risk assessment results. A mouse cursor icon is visible in the top right corner of the right photograph.

Name	Risk Level	Score
VERNON PRATER	LOW RISK	3
BRISHA BORDEN	HIGH RISK	8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

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

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.



DHH ✓ @dhh · Nov 8, 2019



Replying to @dhh

So nobody understands THE ALGORITHM. Nobody has the power to examine or check THE ALGORITHM. Yet everyone we've talked to from both Apple and GS are SO SURE that THE ALGORITHM isn't biased and discriminating in any way. That's some grade-A management of cognitive dissonance.



DHH ✓
@dhh

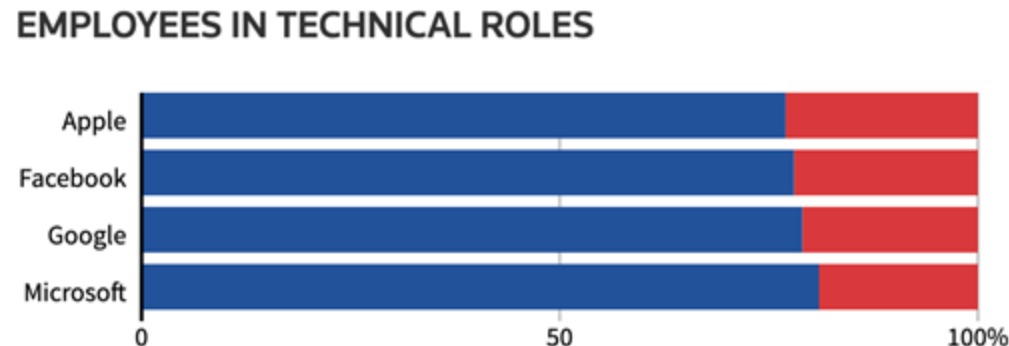
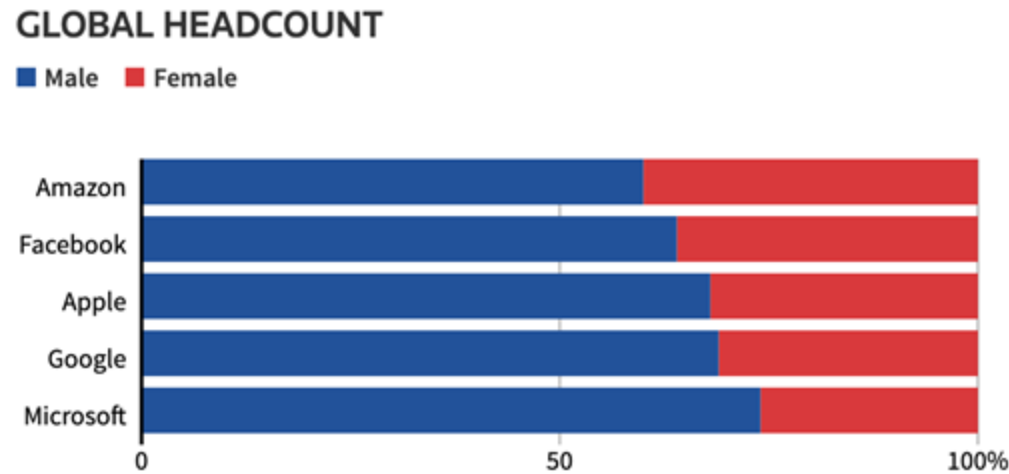
Apple has handed the customer experience and their reputation as an inclusive organization over to **a biased, sexist algorithm it does not understand, cannot reason with, and is unable to control**. When a trillion-dollar company simply accepts the algorithmic overlord like this...

♡ 3,361 3:29 PM - Nov 8, 2019



Amazon hiring tool

- 2014 - Amazon develops AI system to screen resumes.
- One year later, realized the screening tool contains gender bias.
- Why?





Ethics in AI

What do we mean by Ethics?

Ethics in the broadest sense refers to the concern that humans have always had for figuring out how best to live

- Ethics : how to **live the best life (for you, for society, etc.)**
- How do we identify a **good life** (one worth choosing)?
- **Moral principles** that govern behavior
 - E.g., “do unto others as you have them do unto you”

What does ethics have to do with technology?

- Technology shapes how human beings seek the good life.
 - Well-designed and well-used technologies can make it easier for people to live well (e.g., by allowing more efficient distribution of food, water, energy, or medical care). Poorly designed technologies can make it harder to live well (e.g., by toxifying our environment, or by reinforcing unhealthy or antisocial habits).
 - Technologies are not ethically 'neutral', for they reflect the values that we 'bake in' to them with our design. Technologies both reveal and shape what humans value, what we think is 'good' in life and worth seeking.
- Technologies are reshaping the global distribution of power, justice, and responsibility
 - Facebook, Google, Amazon, Apple, and Microsoft have more global political influence than many states and nations.
 - Damaging the fabric of society: democracy, mental health

Ethics and AI

- Individual benefits/harms
 - Well-designed and well-used AI can create individual **benefits**
 - **Reduced effort** - e.g., AI can plan for you
 - **Greater efficiency** - e.g., logistics; better distribution of goods
 - **Easier access to services** - e.g., chat bots
 - Poorly designed AI can create individual **harms**
 - Constraints on **autonomy**
 - **Toxic** effects - e.g., increasing hate speech, unsafe/unhealthy behaviors
 - Reduced efficiency - e.g., inability to account for low-probability high risk events
 - AI reflects the values that we 'bake in' to it with our design.
 - AI language models (e.g., ChatGPT) bake in stereotypes
 - AI trained on human decisions -> human ethics could play a role
- Societal benefits/harms:
 - Facebook, Google, Amazon, Apple, and Microsoft have more global political influence than many states and nations.

Technology ethics

- The positive and negative impacts of technology are distributed unevenly among individuals and groups.
- Technologies can create widely disparate impacts, creating 'winners' and 'losers' or magnifying existing inequalities, e.g., when benefits of a new technology are enjoyed only by wealthy nations while the burdens of environmental contamination produced by its manufacture fall upon citizens of poorer nations.
- How do we ensure that access to the benefits of new technologies, and exposure to their risks, are distributed in the right way? This is a question about technology justice, a matter of ethics.



Ethical Challenges in AI

Ethical issues in data

- Ethical issues are everywhere in the world of data, because data's collection, analysis, transmission and use can and often does profoundly impact the ability of individuals and groups to live well.

“big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”

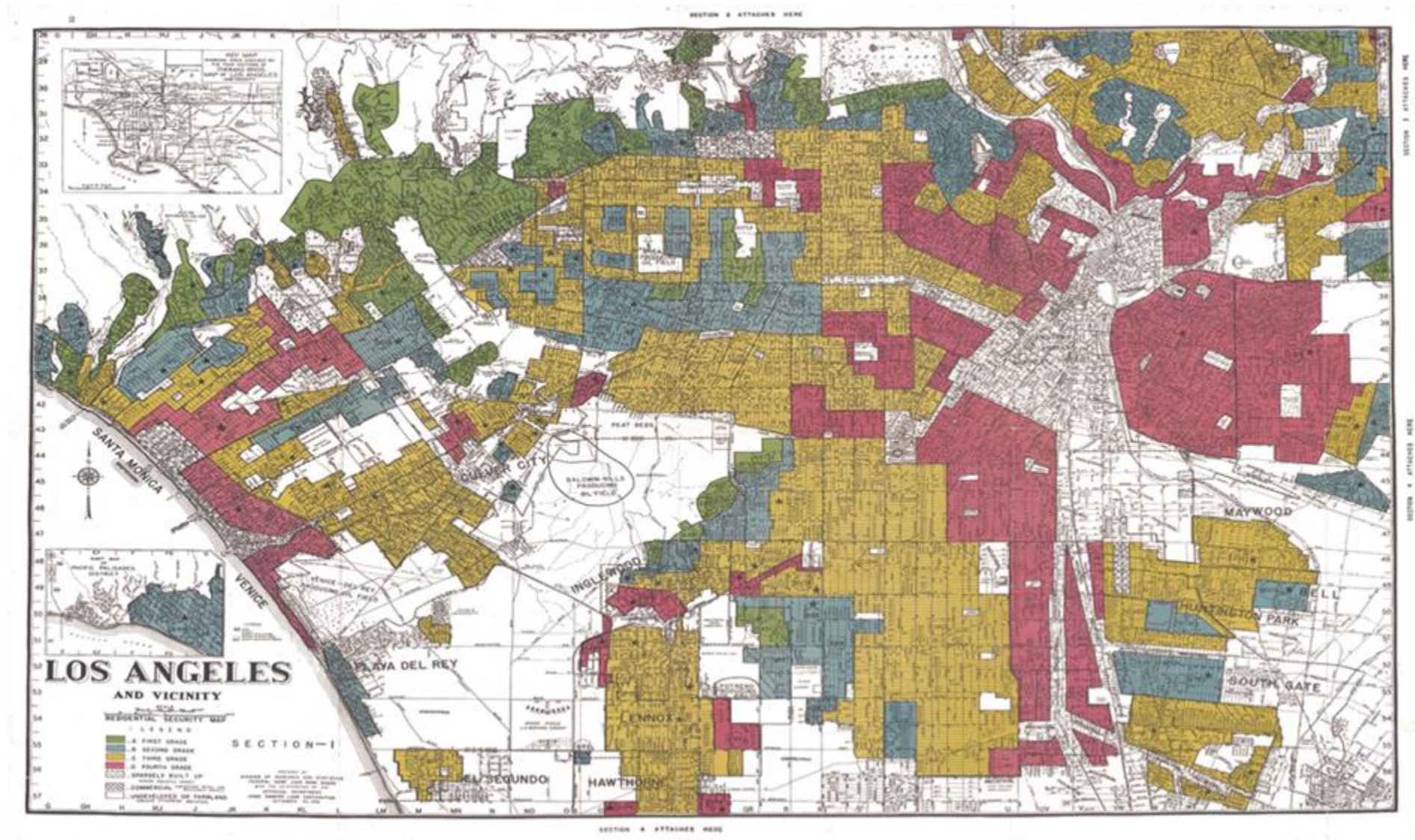
White House report *Big Data: Seizing Opportunities, Preserving Values*, 2014

What is bias?

bi·as
/'bīəs/

1. prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.
2. a systematic distortion of a statistical result due to a factor not allowed for in its derivation.
3. (cognitive) a systematic error in thinking that occurs when people are processing and interpreting information in the world around them and affects the decisions and judgments that they make.

“Redlining” – discrimination by ~~race~~ zipcode



Legally problematic when harms protected classes

US Federal laws define protected classes to include:

- Race.
- Color.
- Religion.
- National origin or ancestry.
- Sex/Gender.
- Age.
- Physical or mental disability.
- Veteran status.
- Genetic information.
- Citizenship.

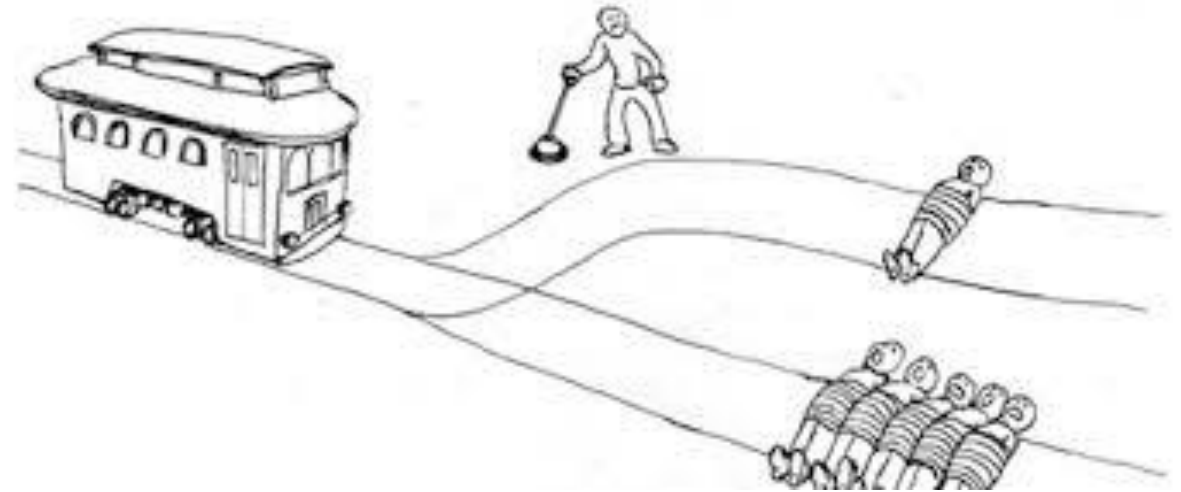
AI does not solve ethical dilemmas

The Trolley problem

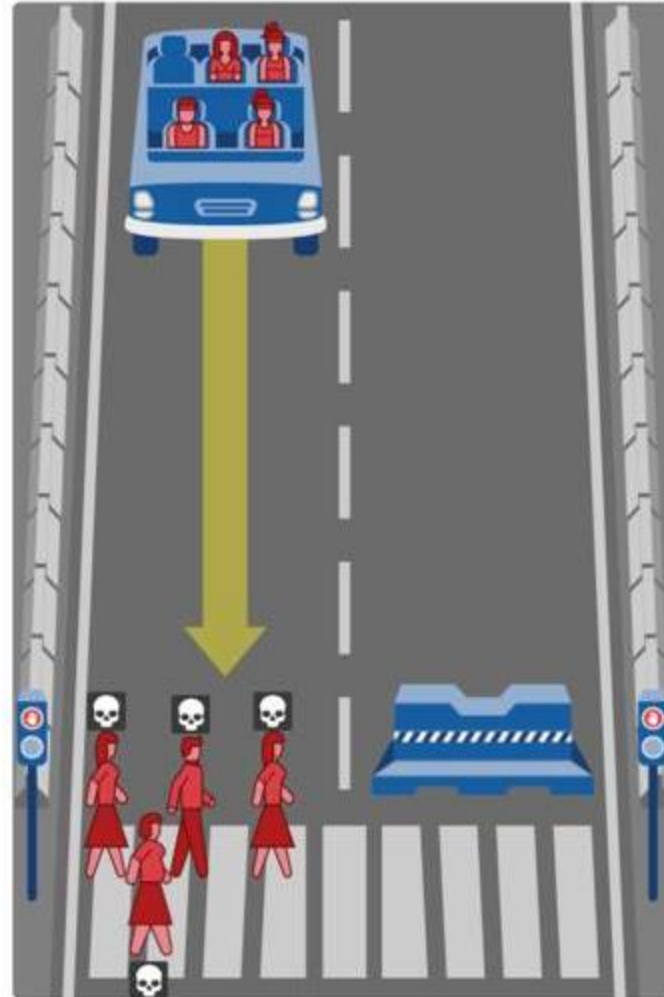
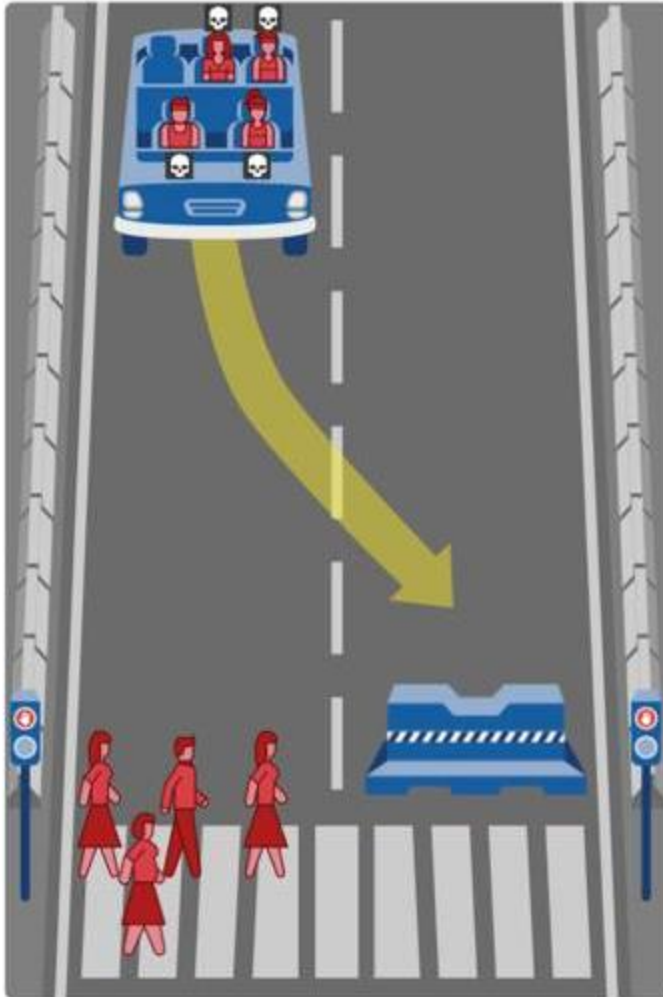
You see a runaway trolley moving toward five tied-up people lying on the tracks. You are standing next to a lever. If you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the more ethical option?



What should a self-driving car do?



Other ethical dilemmas

- AI infers gender
 - ads for high-income jobs are presented more often to men than to women
- AI infers race
 - ads for arrest records are significantly more likely to show up on searches for African-American names
- AI infers a person's income
 - Alternative 1: provides better product recommendations
 - Alternative 2: uses the information to set product prices
- AI infers a person's health state
 - Alternative 1: recommends lifestyle changes to improve health
 - Alternative 2: uses it to set health insurance rates
- AI directs our interactions
 - recommendations for friends, and news articles that further biases

Other dimensions of AI

- **Explainability** – How did the system arrived at its decision/outcome?
 - e.g., why a loan was denied?
- **Stability** – Will AI make consistent decisions?
 - E.g., is model sensitive to initial conditions? Will AI lead to similar outcomes in each deployment?
 - Are decisions inconsistent with small data changes?
- **Economic inequality** – Income inequality has grown as AI automation replaced routine work.



Trustworthy AI

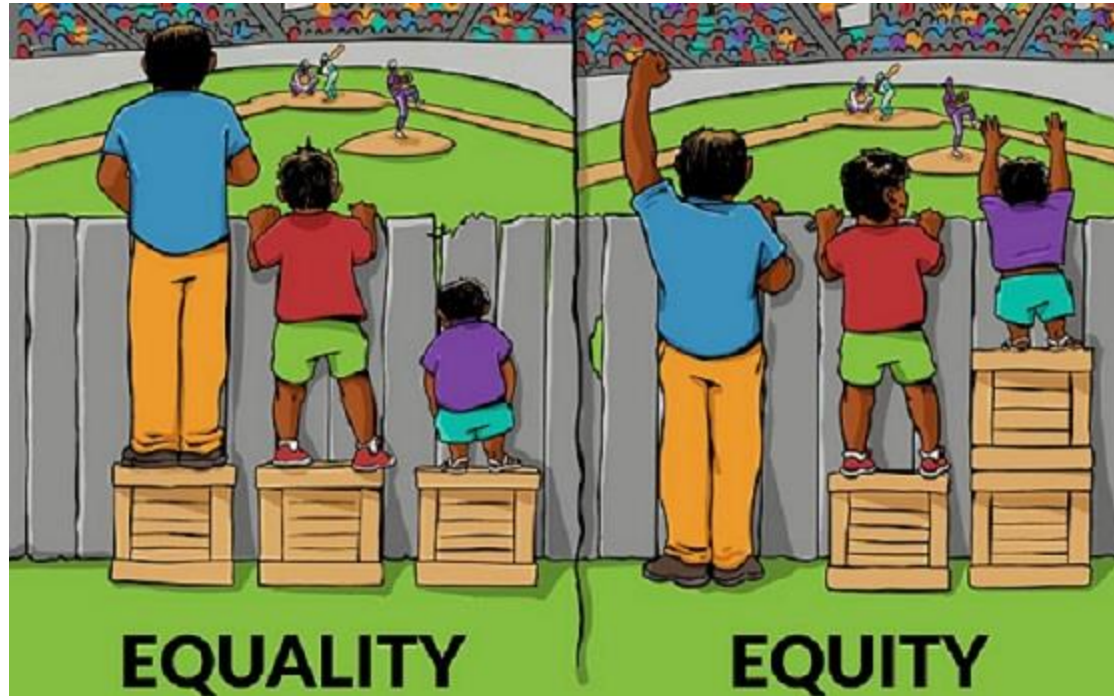


Overview of topics

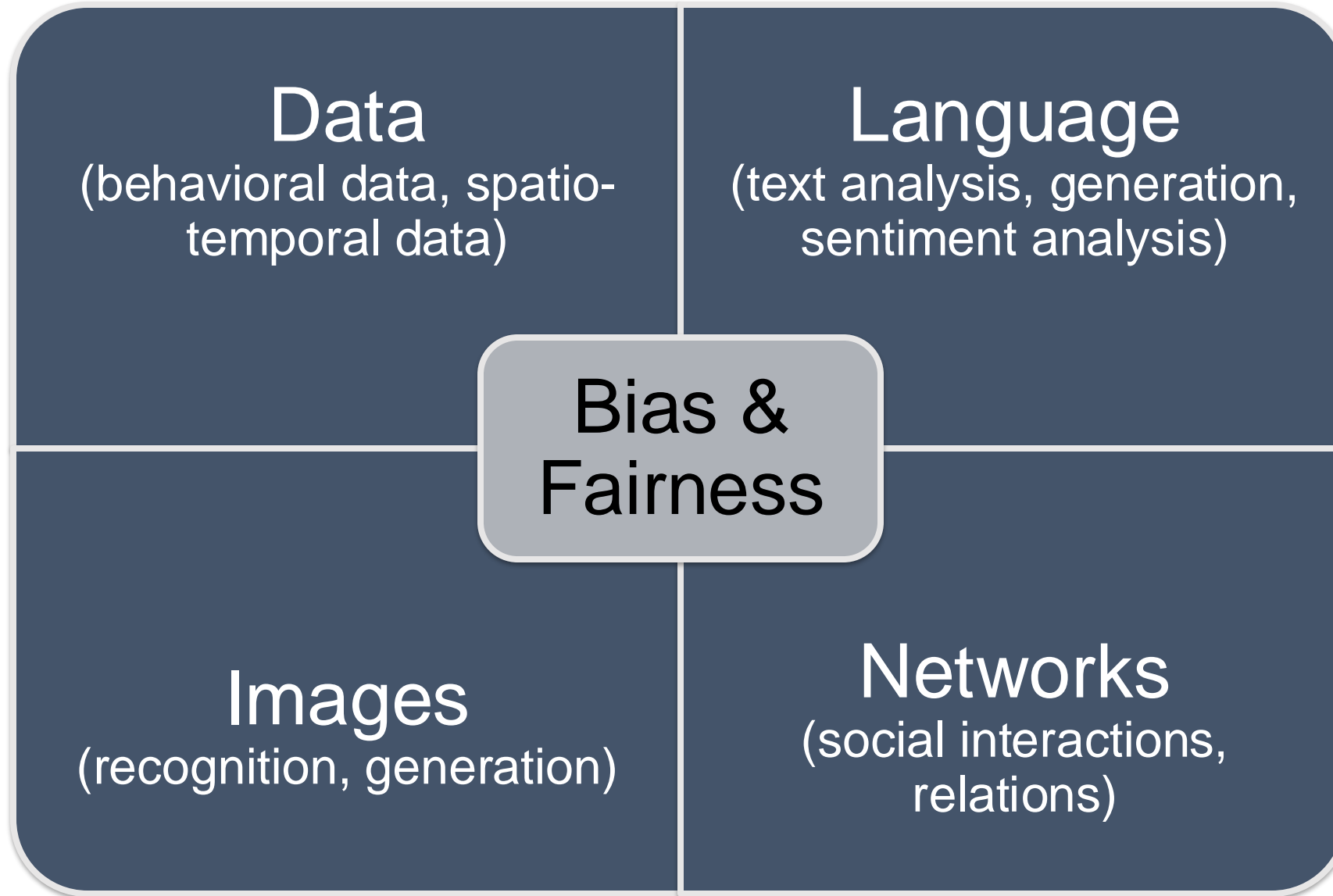
What is fairness?

- Intuitive property of fairness
 - Your private attributes (race, gender, ...) should not affect how you are treated by the algorithm (esp. when they are not relevant to the decision being made)
- In reality, dozens of measures of fairness,
- Many incompatible with each other and can only be simultaneously satisfied in a few special cases.

Fairness and equity

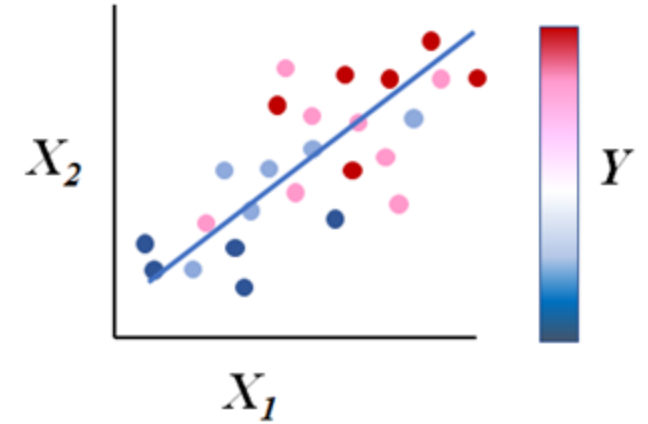
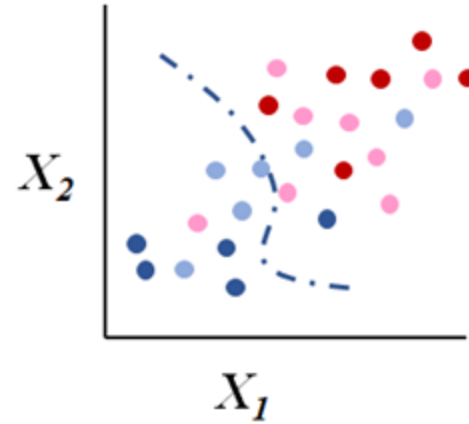
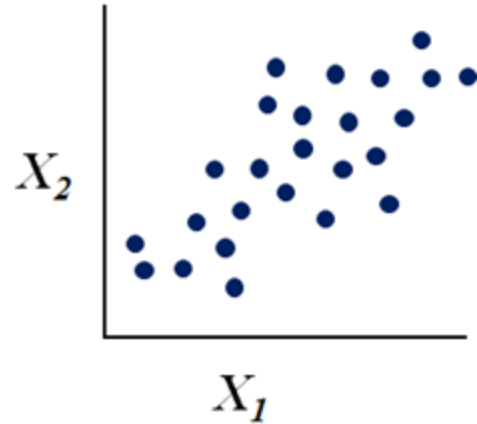


This course at a glance





Types of data analysis methods



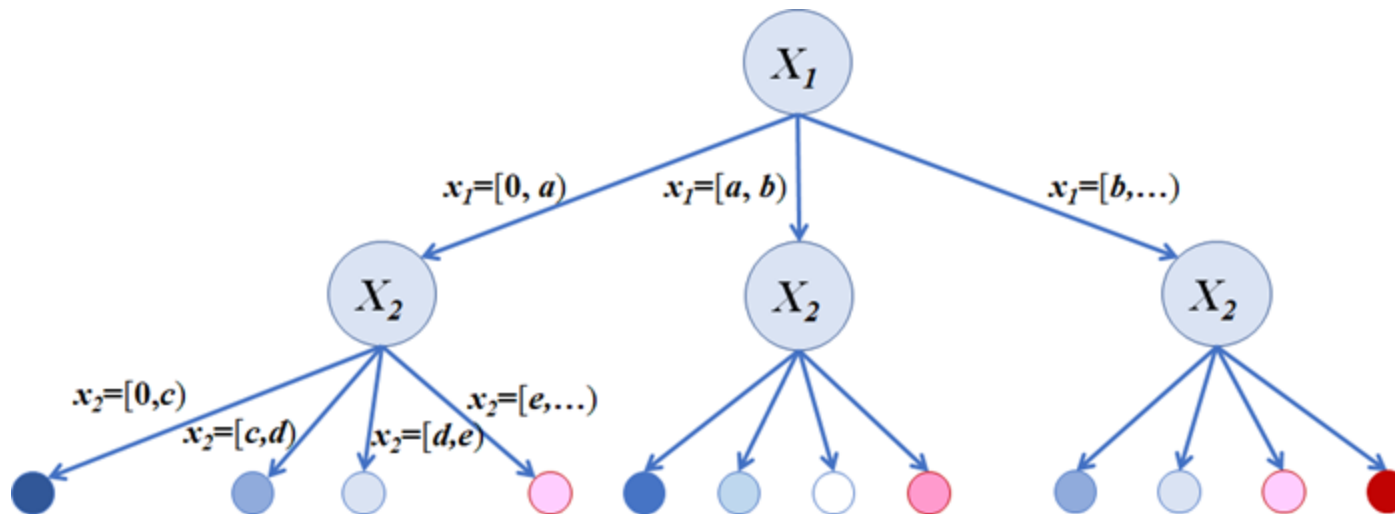
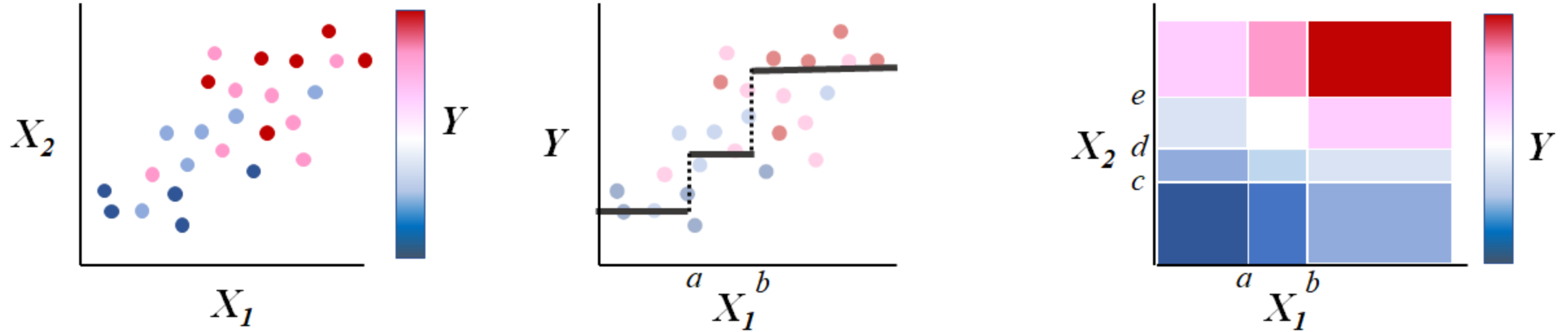
Unsupervised methods

- Input: a set of features
- Clustering
 - K-means, hierarchical, spectral,
- Dimensionality reduction
 - PCA, SVD, ...

Supervised methods

- Input: features & labels/outcomes
- Classification
 - Random forest, SVM, ...
- Regression
 - Linear regression, logistic ...

Data modeling and prediction



Sources of bias in data

Simpson's paradox

Subgroups within a population with different behaviors

Sampling bias

Subgroups are not equally represented

Heterogeneous data

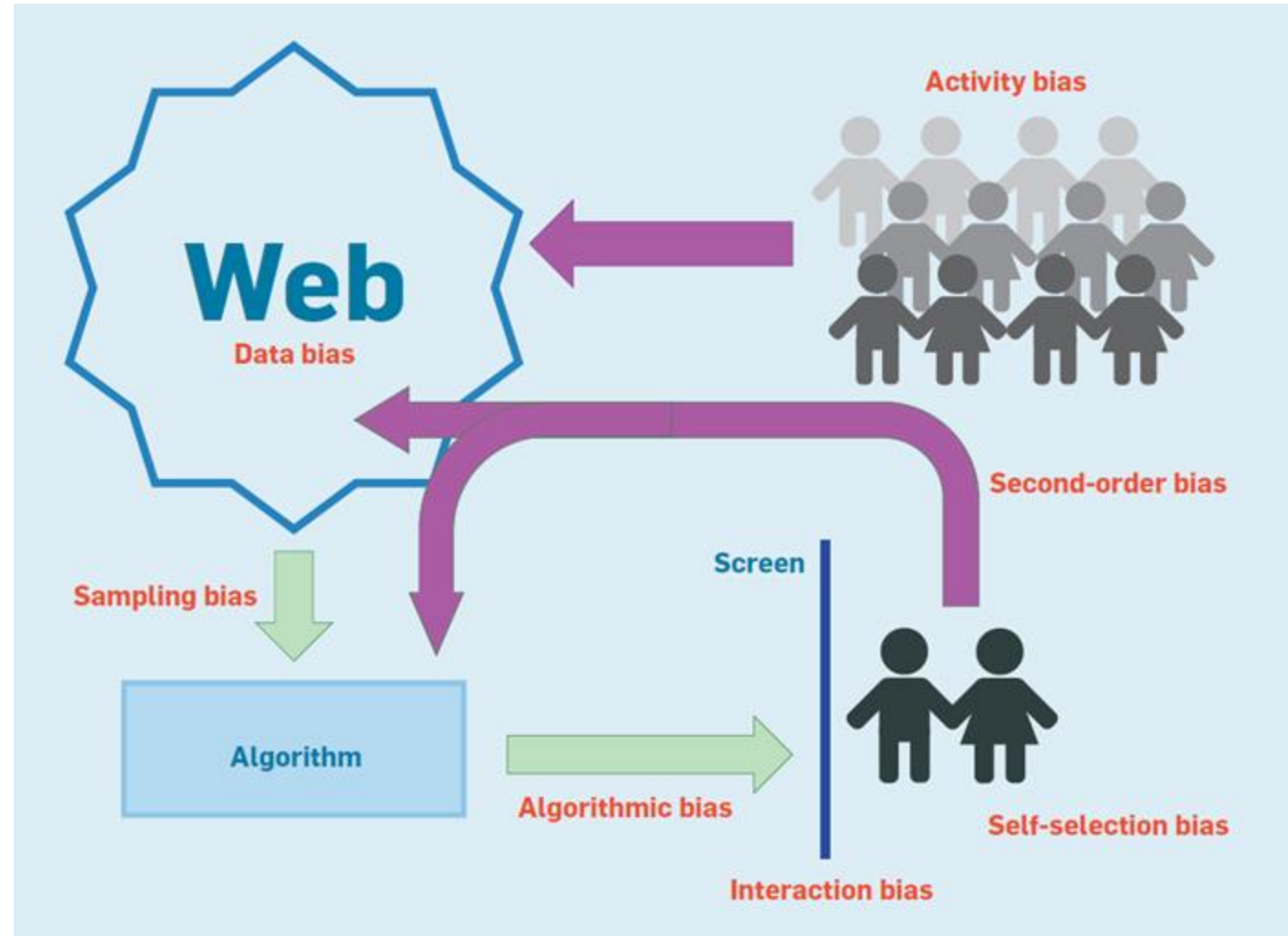
Survivor bias

Population changes when measuring trends

Cross-sectional data fallacy

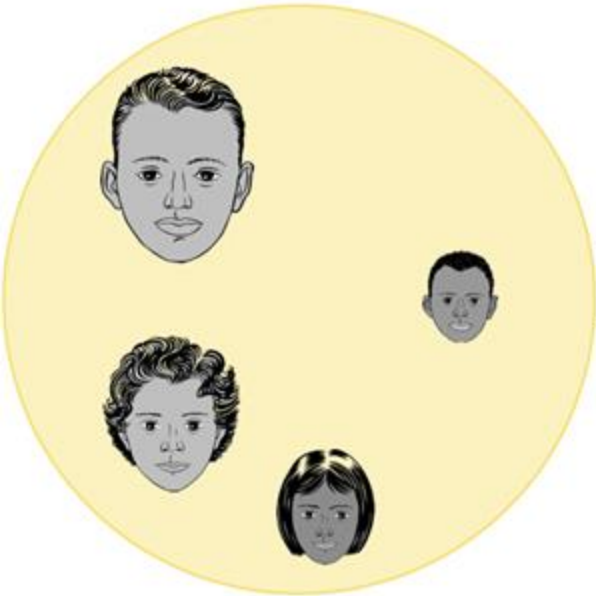
Population composed of different cohorts

Sources of bias in social data

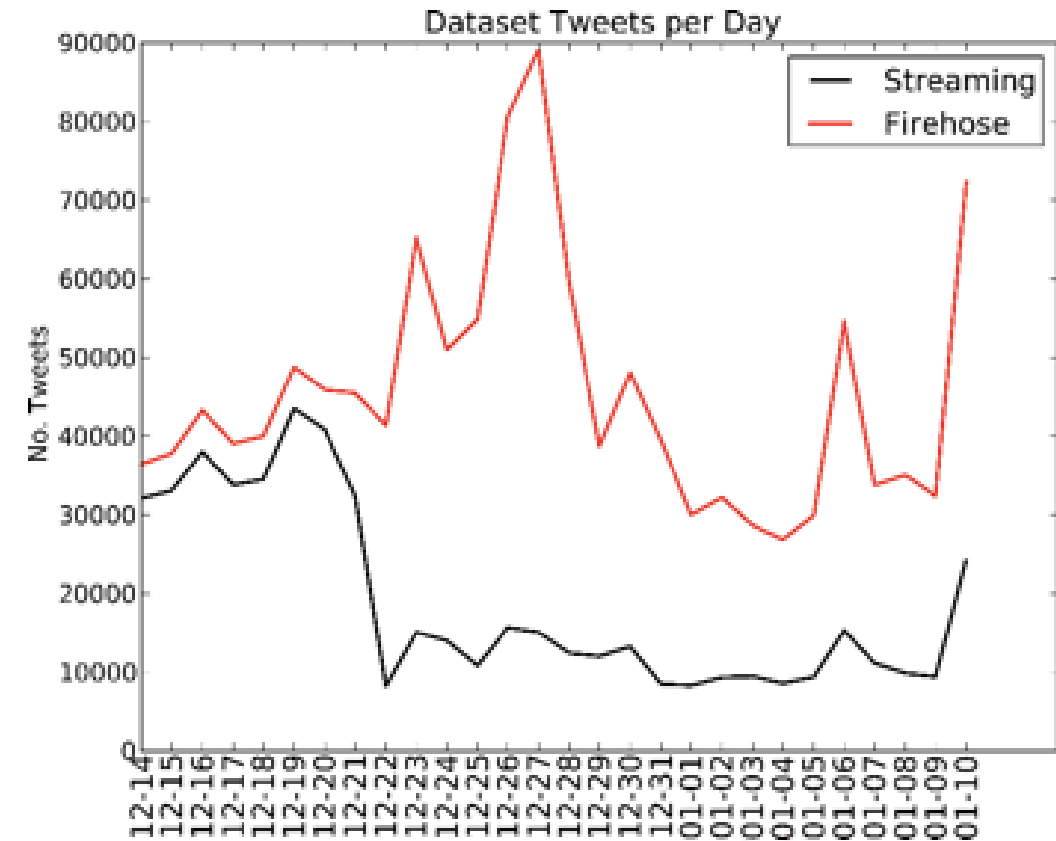
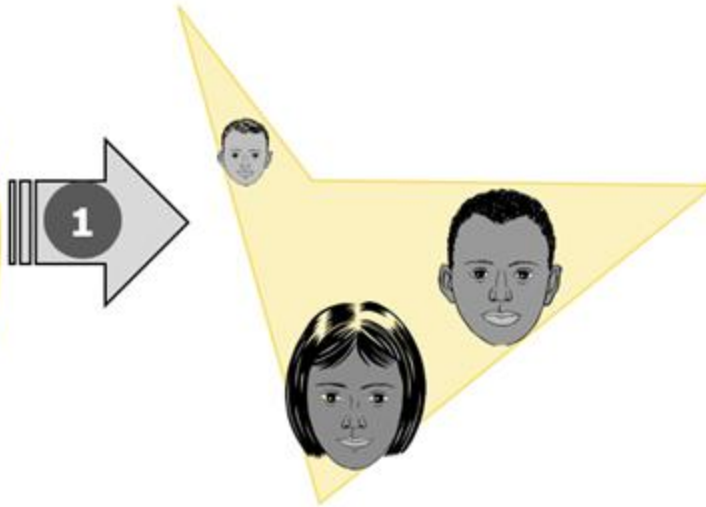


Sampling and Representativity

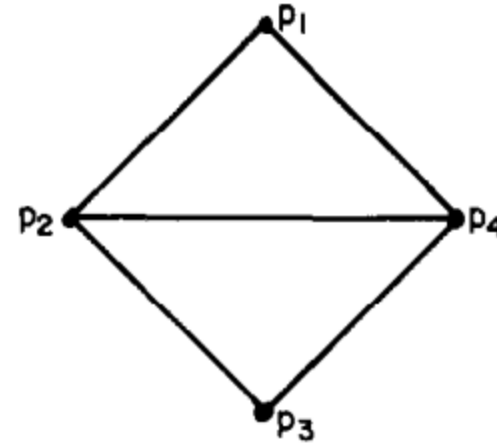
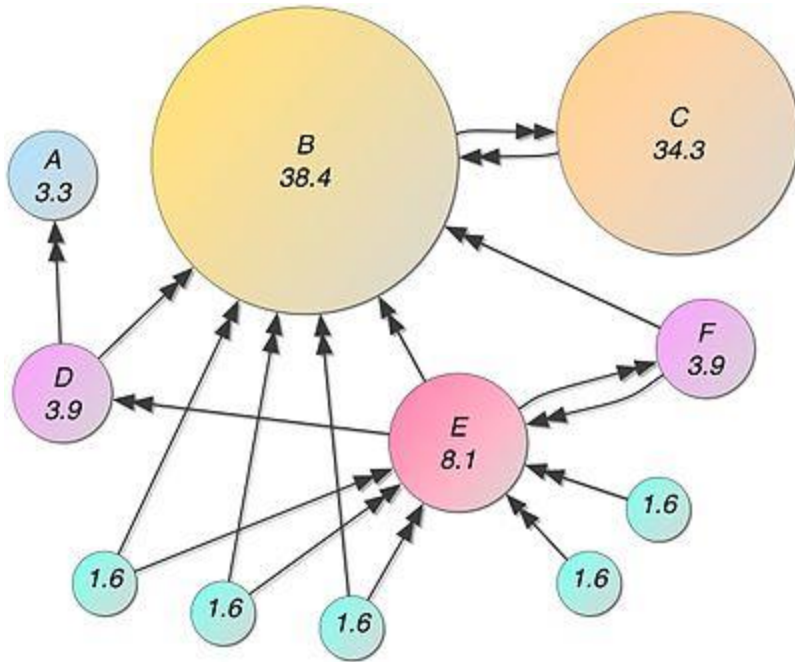
The World



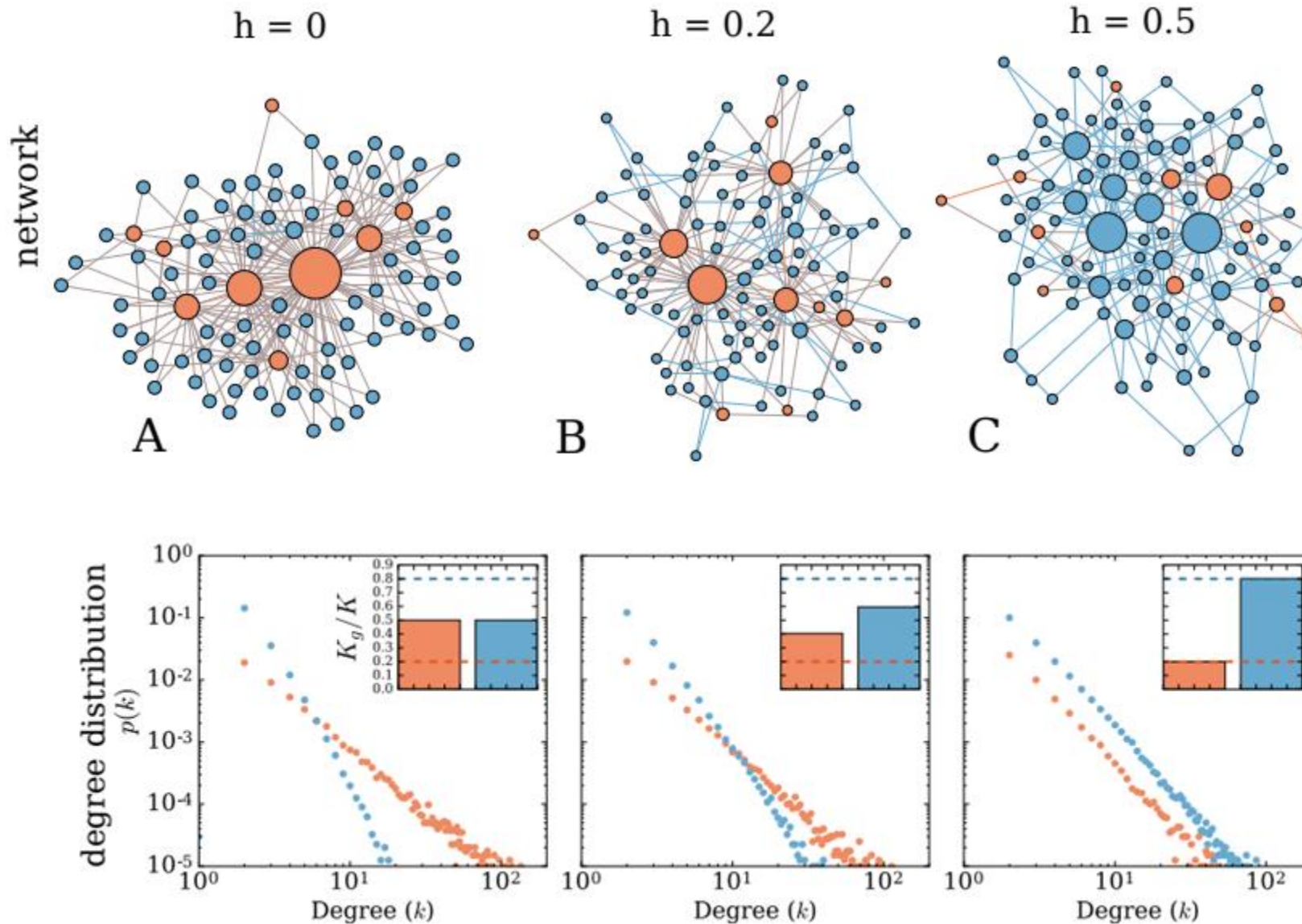
Social Media



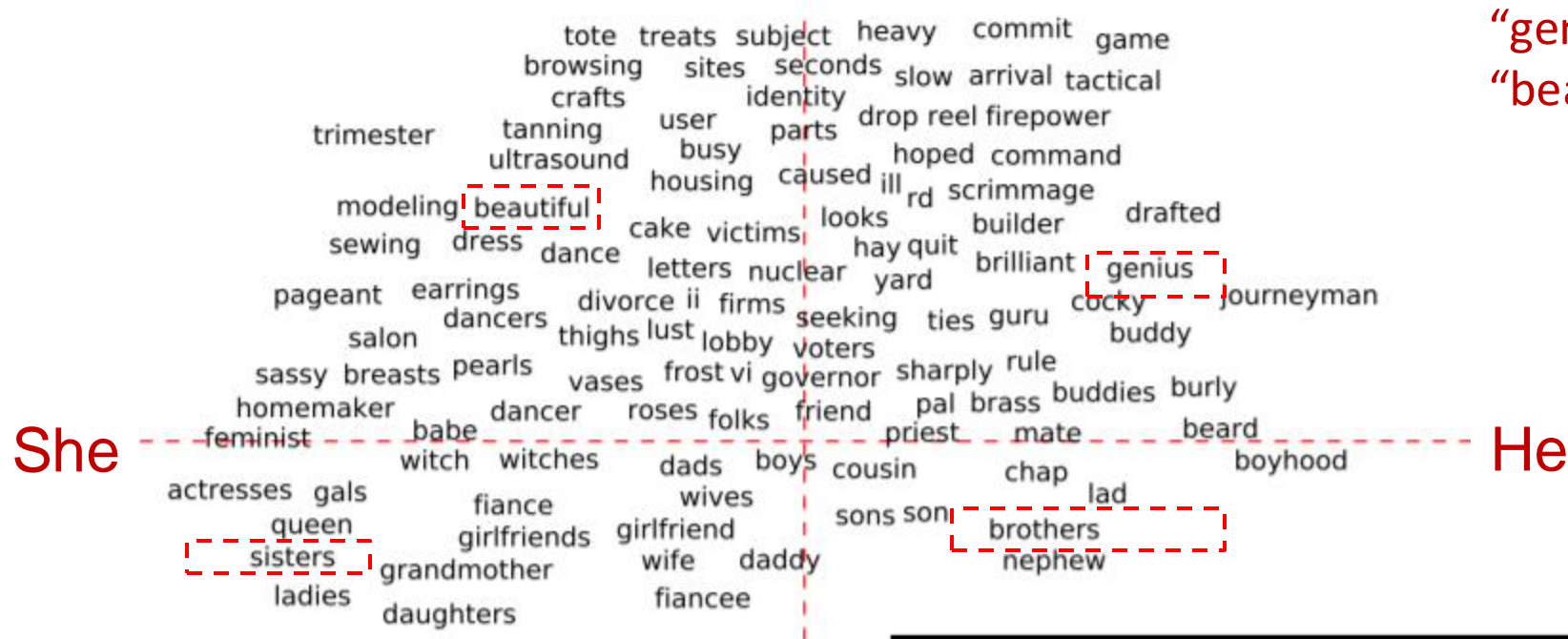
Networks: Representation and Metrics



Bias in networks



Bias in language



Source: <https://towardsdatascience.com/tackling-gender-bias-in-word-embeddings-c965f4076a10>

“genius” is male but
“beauty” is female?

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

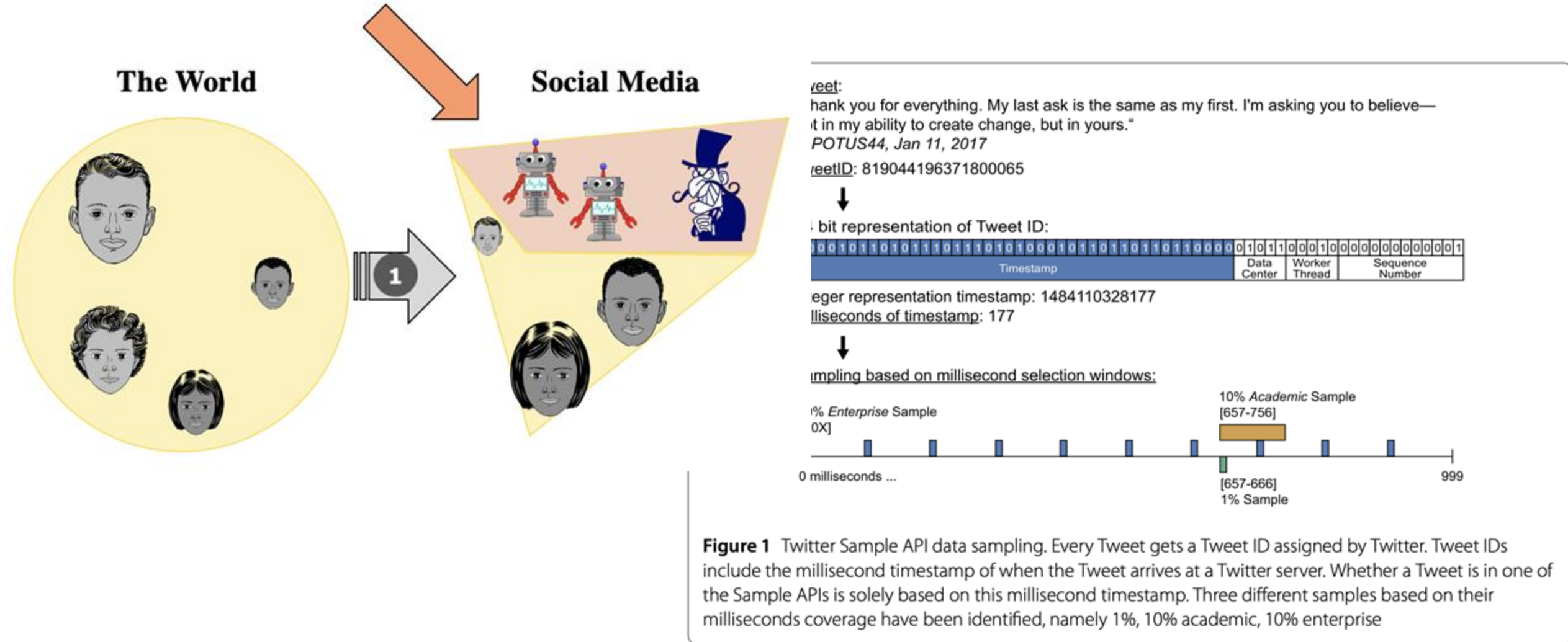
Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Manipulation and Deception





der was misidentified in **up to 7 percent** of lighter-skinned female photos.



identified in **up to 12 percent** of darker-skinned male



misidentified in **35 percent** of darker-skinned females



er was misidentified in **up to 1 percent** of lighter-skinned males photos.

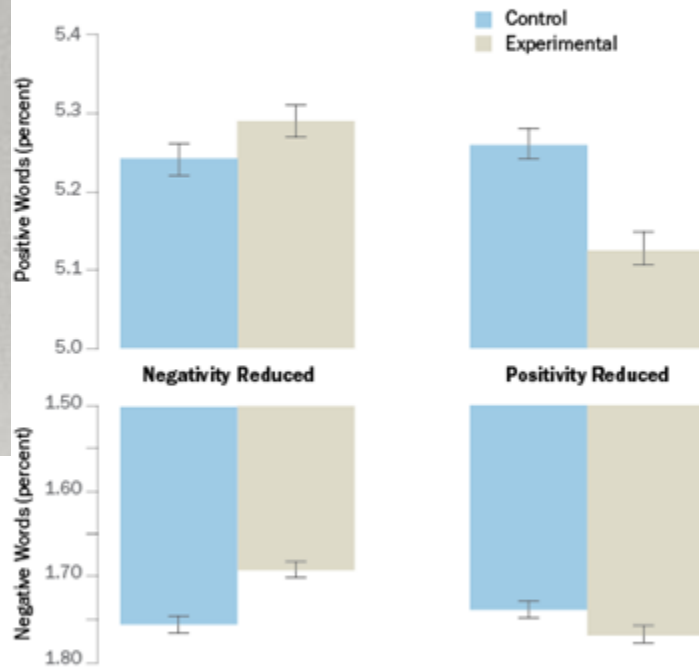
<https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>

Fairness in image analysis: Gender classification error rate

Ethics in Research

Facebook Test Finds Exposure to 'Emotional Content' Leads to Small Changes in Users' Expressed Emotions

Users exposed to different levels of positive or negative words in their 'news feed' provokes minimal changes in the emotional content of their own posts



Source: Proceedings of the National Academy of Sciences of the United States of America, June 17, 2014 vol. 111 no. 24

PEW RESEARCH CENTER



- History and Ethical Principles
- Defining Research with Human Subjects
- Internet-Based Research

**facebook,
don't run
tests on
me**

What you will get out of the course

- Learn the principles of data science, and how to use them responsibly
 - Social data
 - Networks
 - Language
- Learn how to create explainable and transparent models
 - not black box methods!
 - how to avoid common pitfalls of analysis
- Apply what you learn to an original research project
- Learn to present research findings, and learn from peers



Course details

Where to find me (Kristina Lerman)

- At USC
Immediately before or after class
- At ISI
ISI Rm. 932 (by appointment)
- Zoom (by appointment)
- Email: lerman@isi.edu



TA

Fiona Guo

- **Office:** ISI 921 **OR**
<https://usc.zoom.us/my/siyiguo>
- **Office Hours:** 10-11 Wednesday on Zoom
or by appointment
- **Contact Info:** siyiguo@isi.edu



TA

Ashwin Rao

- **Office:** ISI 911 **OR**
<https://usc.zoom.us/j/9026240639>
- **Office Hours:** 10AM - 11AM PT Monday on Zoom or by appointment
- **Contact Info:** mohanrao@usc.edu



Course Communication

- Brightspace – brightspace.usc.edu
 - Your USC login works on this account
 - Important announcements and logistics are posted here
 - Lecture slides will be posted before each lecture
 - All assignments will be submitted through Blackboard
- USC Slack - usc.enterprise.slack.com
 - #spring25-dsci-531-32427 channel (#spring25-dsci-531-32455 for DEN students)
 - course discussions
 - homework and project discussions
 - To ask questions, please mention(@) TAs and professor
- All questions should be posted (not emailed!)
 - If you know the answer to a posted question, please try to provide helpful suggestions

Readings

- Posted on the site each week
 - You can read it online or print them
 - Listed in syllabus (subject to change)
- Please read all required readings before the class they are covered

Slides

- Available online before the lecture
- NOT a replacement for the lecture
- You can print these out and make notes on them

Prerequisites & Recommendations

- Prerequisites
 - Working knowledge of Python
 - Recommended courses a plus
- Recommended Courses
 - CS561 – Introduction to AI
 - CS573 – Advanced AI
 - DSCI551 – Data Science
 - Probability and statistics
 - Networks or Graph Theory, NLP

Grading

Assignment	Points	% of Grade
Homeworks	20	20
Quizzes	30	30
Class participation	5	5
Project proposal	3	3
Literature review	5	5
Midterm report	8	8
Peer review	4	4
Project report	15	15
Project presentation	10	10
TOTAL	100	100

Workload - class participation

- Importance – 5% of the final score
- Easiest points to receive
 - Show up to class
 - Do not use social media in class ...
 - Answer questions
 - Ask questions

Workload - Homework

- Importance – 20% of the final score
- Submit on Blackboard
- Timeline
 - HW #1: Research ethics & Basics of Data Analysis
 - HW #2: Bias in Data and Prediction
 - HW #3: Bias in NLP
 - HW #4: Bias in Networks
- Due dates: see syllabus
- Grading rubric: see syllabus

Workload - Quizzes

- These are worth 30% of your final score
- Importance
 - 5 quizzes, roughly every other Monday **in class**
 - Basic ideas discussed in the class and/or the topics related to the recommended readings
 - Your lowest score will be dropped

Workload - Projects

- Importance – 50% of the final score
- Submit on Blackboard
- Grade breakdown
 - Proposal: 3%
 - Literature Review: 5%
 - Mid-term Report: 8%
 - Peer Review: 4%
 - Final Presentation: 10%
 - Apr 28 and 30, in class
 - Final Report: 15%
 - Due May 7

Course Project

- A research project based on what you have learned in class
- Be creative!
- An ideal project is one that you could publish a paper about
 - Empirical validation is important – show that your method beats state-of-the-art
- Teaming
 - 2-3 students
 - No individual projects

An Extensive Analysis of Exclusion Bias through Social Media

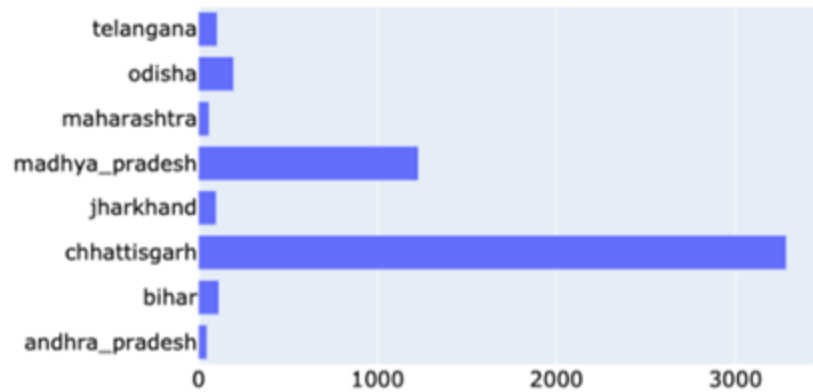


Figure 1: Number of posts from each states in CGNet

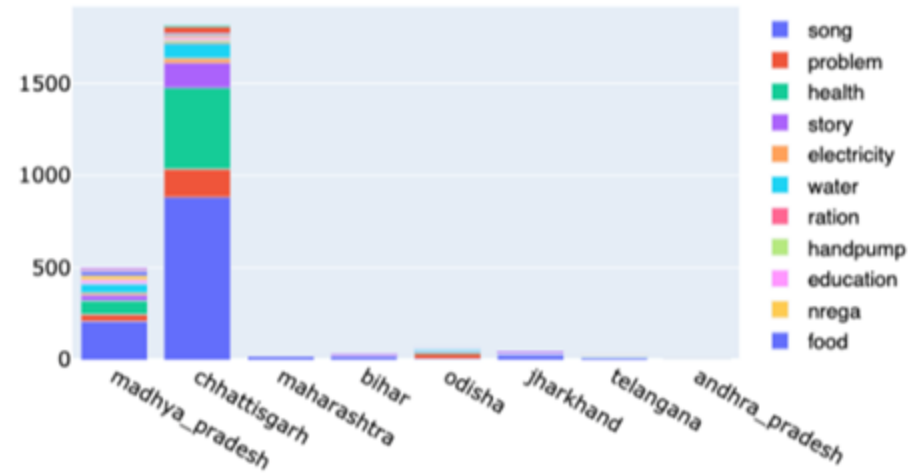
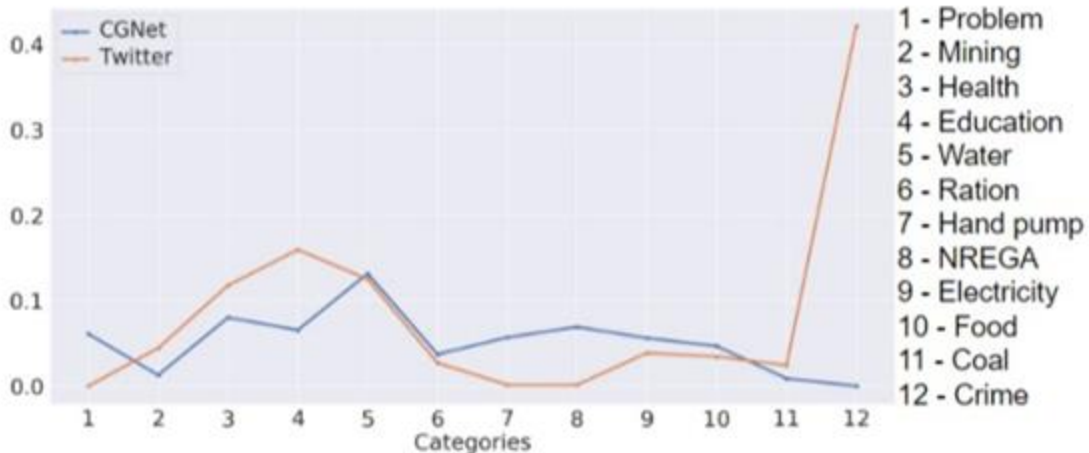
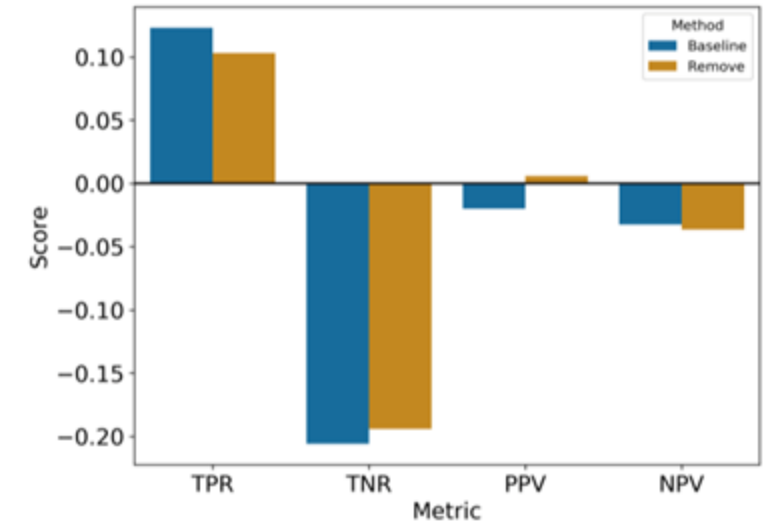
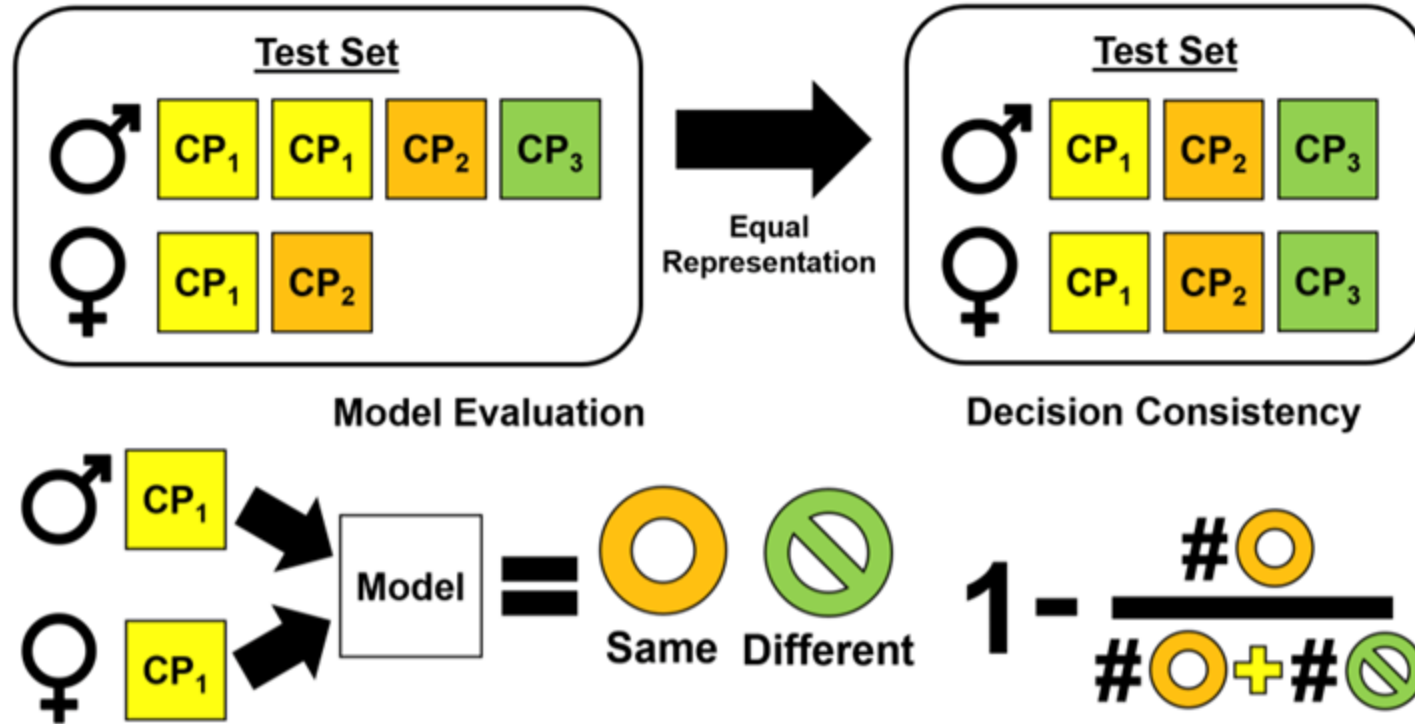


Figure 2: Distribution of each category wrt each state

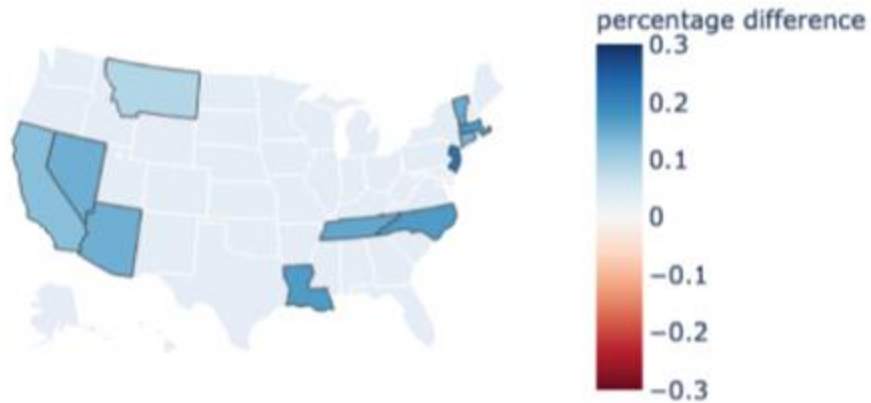


Unfair Fairness Evaluation



Regional Ageism in Traffic Policing Activity

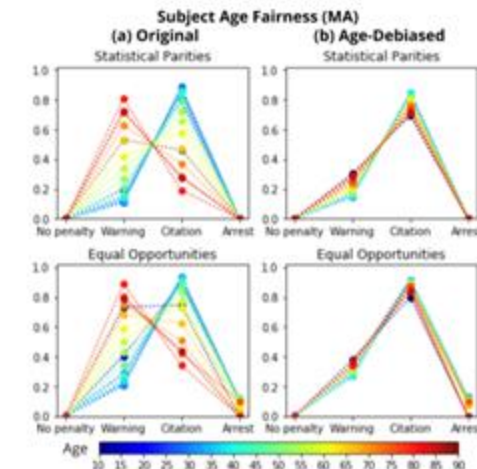
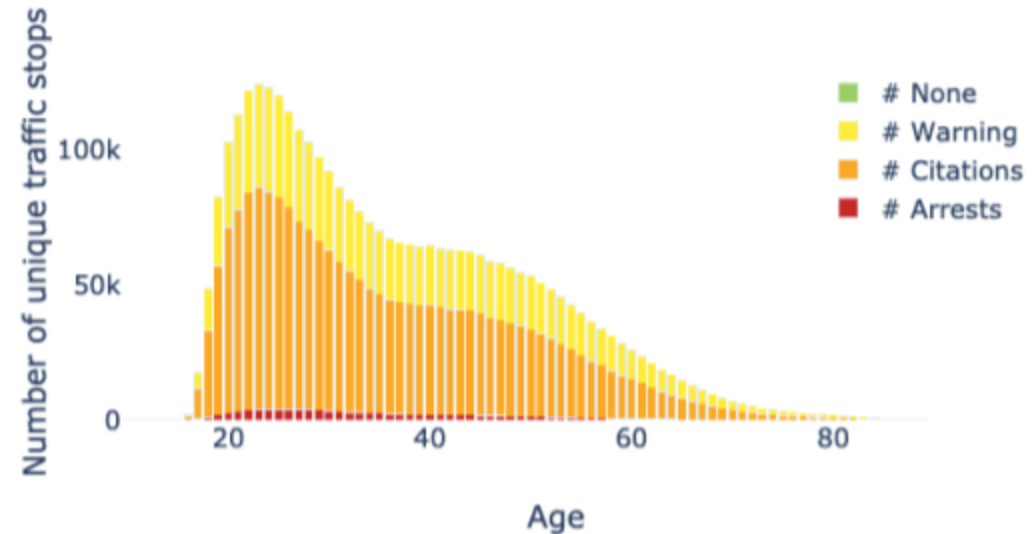
(a) Ages 20-29 Stop vs Census Makeup (stop % - census %)



(b) Ages 50+ Stop vs Census Makeup (stop % - census %)



(a) Stop Distribution Type by Age in MA



Project examples



Bias in COVID-19 Detection from Audio

“This project focuses on looking for biases that exist in COVID-19 detection based on coughing sounds. We hope to explore whether the cough sound screening can be a more cost-effective but reliable alternative to the PCR test in the post covid period. In particular, we will work on checking the accuracy and reliability of detection results for all age groups, genders, etc.”

Bias in Mate Selection for Speed Dating

“...The trained [dating app] models will, however, reflect the data’s biases towards particular demographic groupings, but few researchers have paid attention to them. In our work, we build regression models combined with pre-processing techniques to perform the dating successful rate prediction, and deploy a number of evaluation mechanisms

Questions to ask

What do you want to predict?

What are common datasets? Are some demographics over/underrepresented in these data?

What are common features gleaned from these data? Do they harm certain classes?

Example: hate detection model that flags people who use AAE

(Example!) Data sets

- Social data
 - Harvard Dataverse
 - <https://dataverse.harvard.edu/>
- Online networks/social media
 - SNAP <http://snap.stanford.edu/>
- Health
 - NHANES <https://www.cdc.gov/nchs/nhanes/index.html>
- Education
 - NCES <https://nces.ed.gov/>
 - <https://pslcdatashop.web.cmu.edu/>
- COMPAS Dataset: <https://github.com/propublica/compas-analysis/>
- UCI Adult Dataset: <https://archive.ics.uci.edu/ml/datasets/Adult>

Grading policy – Late days

- 7 days combined
 - 4 homework assignments and 5 course project assignments (proposal, lit review, midterm report, peer review, final report)
 - No late days allowed for quizzes
- Max 2 days for each submission
- Submissions after 2 late days or when 7 late days are used up will lose 20% points per day.

Project Proposal

- Proposal should include:
 - Title
 - This should be the title of your final paper
 - Authors
 - The people who will be involved in the project
 - A description of the project
 - Be sure to state what you think is new and innovative about your project
 - The use of pictures and screen shots is encouraged
 - The date you would like to present your project
 - Length – 2 pages (max)
 - Format: Latex

Project Paper

- Due at the beginning of the last class
- Length: 5 pages
- Format: Latex – same as project proposal
 - use overleaf.com
- Content:
 - Title block
 - Abstract
 - Body
 - Introduction
 - Related work
 - Methods
 - Results
 - Conclusion
 - References

Project paper: Title Block

- Title of the paper
 - Choose a title that highlights the contribution of your paper
 - Keep it short (<16 words)
 - Title should summarize the main outcome of the study
- Full names of the authors
 - Address includes affiliation:
 - University of Southern California
 - Computer Science Department
 - Los Angeles, CA 90089
 - Your email address

Project paper: Abstract

- An abstract is a 100-250 word summary of your paper
 - Identify the problem you are solving
 - Describe your solution
 - Summarize the results that support your solution
- An abstract is NOT the same as the introduction
 - NEVER repeat the abstract in the introduction
 - You might restate portions of it in the paper, but using different words

Project paper: Introduction

- Introduction is a Mini Paper
- The formula
 - The **problem** and why is it important
 - **State of the art** and its failing (~Related Work)
 - Your **contribution** (~Methods)
 - Its **benefits** (~Results)
 - End with the **big picture** (~Conclusion)

Project paper: Body of Paper

- Introduction
- Sections might include:
 - Motivating application or example
 - Approach or Methods
 - Results or **Evaluation**
 - **Evaluation is one of the most important components of the paper!**
 - Related work
- Conclusion/Discussion
 - Summarizes the contribution, including any conclusions and directions for future research
- Use pictures, screen shots, and diagrams

Project paper: References

- References to both related work and work that you build on
- Use the “named” bibliography style [Ambite, 2004].
- Bibliography
 - Ambite, Jose Luis, 2004. Planning by Rewriting, Journal of Artificial Intelligence, Kluwer Academic Publishers, 4(2), pg 27—34.

Cheating and academic integrity

- Not tolerated!
- No second chances – all infractions will be reported
 - First offense is automatic failure in the class
 - Second offense is suspension from the University
- Examples:
 - ChatGPT, or other LLMs
 - Turning in someone else's work
 - Copying from someone else during a quiz
 - Doing a project that uses someone else's work without giving them credit
- We will follow USC policies:
<https://sjacs.usc.edu/students/academic-integrity/>

Plagiarism is cheating

- What is plagiarism?
 - More than 7 words copy and pasted from another work
 - But what if I want to quote someone verbatim?
 - “put it in quotes and reference” (K. Lerman, *Course Introduction*, p. 72, 2023)
- Examples
 - Copying code for homework
 - Copying text or figures from someone’s paper
 - Copying text from a blog and passing it off as your own work

ChatGPT: Acceptable uses of GenAI

GenAI is a tool but not a substitute for writing. Tips on using it well:

- Brainstorm project ideas
 - Chat with it and ask to refine your project ideas
- Generate paper outline
- Review existing literature
 - But, make sure to check the references exist! You will be responsible for AI hallucinations
- Editing, grammar checking, etc.
 - Prompt: “Polish text: ...”
 - But, don’t plagiarize GPT: its prose is bland and won’t earn you full credit.
- Create illustrations for your presentation
 - Include source acknowledgement, e.g., “Open DALL-e ...”

Rules of behavior

- Attendance
 - Please be on time
- Phone use
 - If it makes noise, turn it off or to vibrate mode in class
 - No texting! No web surfing!
 - Pay attention! You are paying for the privilege of attending the course – make the most of it

Once the course is over

- Directed research (1-2 MS or Phd Students)
- Research Assistantships (Phd Students)
 - We can also recommend you for positions in other groups
- Teaching Assistantships (for PhD students)