Ethics in Data

Kristina Lerman

Principal Scientist, USC Information Sciences Institute
Research Professor, USC Computer Science Dept.

USC Viterbi
School of Engineering

What is Ethics?

- Ethics refers to the concern that humans have always had for figuring out how to **live their best life**
- How do we identify a **good life**, one worth choosing from among all the different ways of living that lay open to us?
- **Moral principles** that govern behavior
  - E.g., “do unto others as you have them do unto you”

AI shapes how people live a good life: makes it easier for some but harder for others

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?

USC Viterbi
School of Engineering

AI 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
  - Who **benefits** from AI? e.g., AI automation is profitable for companies.
  - Who is **harmed**? e.g., AI automation increases income inequality.
- *AI fairness*: How do we know who is harmed by AI, who benefits?
- *AI justice*: How do we ensure that access to the benefits of AI, and exposure to its harms, are distributed in the right way?

Ethically significant benefits of AI

- **Human understanding** – AI can make sense of data to help us better understand the world
  - E.g., drug discovery
- **Social and economic efficiency** - once we understand the world, we can design or intervene in its systems to improve their functioning
- **Predictive accuracy & personalization** - more precisely tailor actions to be effective in achieving good outcomes for specific individuals, groups, and circumstances.

Ethically significant harms of AI

- **Harms to privacy & security** - Even anonymized datasets can be linked & merged to reveal intimate facts, including medical and mental health history, private conversations at work and at home, genetic makeup and predispositions, reading and Internet search habits, political and religious views, which can be re-constructed and stored without your knowledge or informed consent.
- **Harms to transparency & autonomy** – transparency refers to the ability to see how a given system works. Autonomy is the respect for the capacity of individuals to make decisions and is expressed in research through informed consent
- **Harms to fairness & justice** – from errors and inaccuracies and unjust, hidden biases in data that rest on falsehoods, sampling errors, and unjustifiable discriminatory practices. These are especially concerning when affecting legally protected subgroups.

Other dimensions of harm

- **Explainability** – do we understand how the system arrived at its decision/outcome, e.g., why a loan was denied? Do we know what can be done to change the decision, e.g., what needs to change to get the loan approved?
- **Stability** – when deployed, will AI lead to similar outcomes in each deployment? Or will there be large variance in outcomes (chaos)? What are the unintended consequences?
- **Economic inequality** – AI systems designed to replace workers rather than enhance their performance. Income inequality has grown as AI automation replaced routine work.

→ Trustworthy AI

How is it possible for AI to be unethical?

USC Viterbi
School of Engineering

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

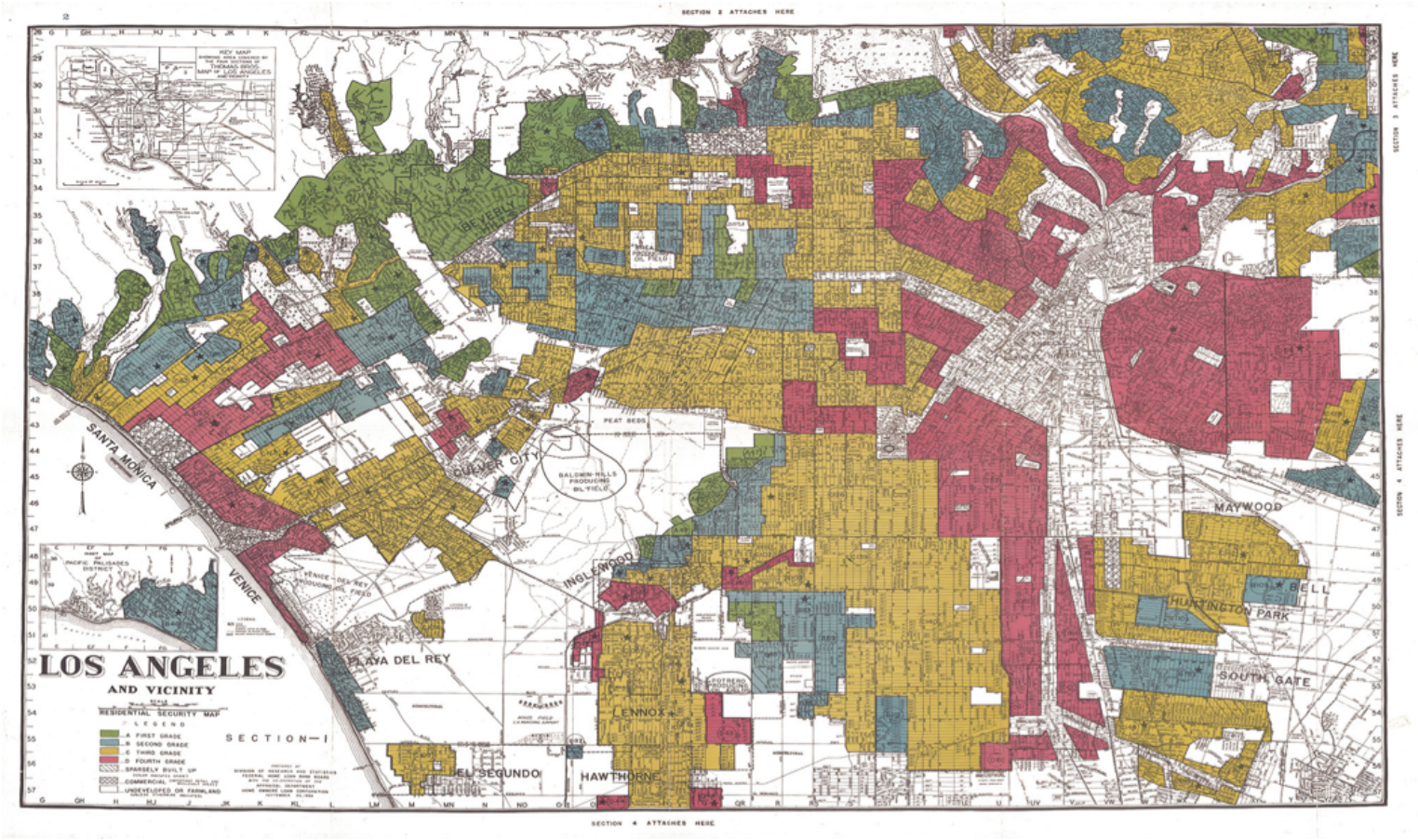
USC Viterbi
School of Engineering

Legally problematic when harm protected classes

US Federal laws define protected classes to include:

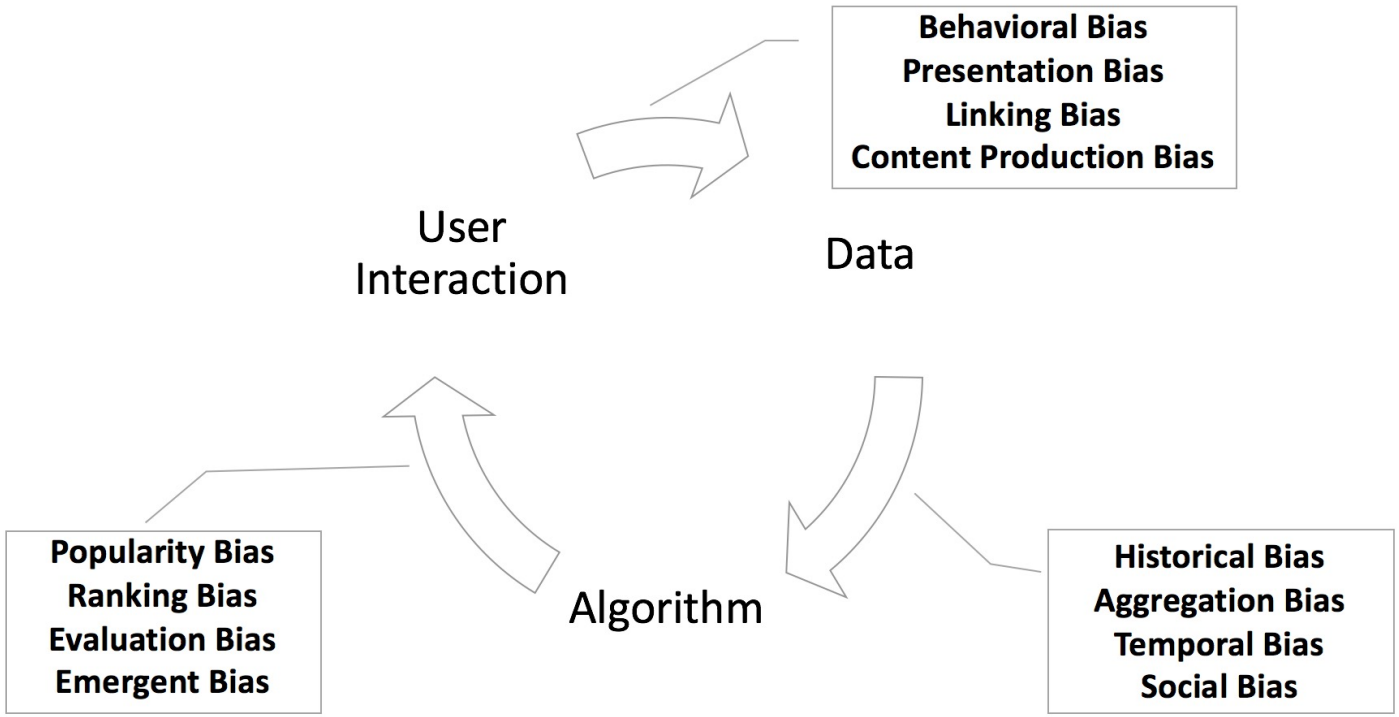
- Race.
- Color.
- Religion.
- National origin or ancestry.
- Sex/Gender.
- Age.
- Physical or mental disability.
- Veteran status.
- Genetic information.
- Citizenship.

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



USC Viterbi
School of Engineering

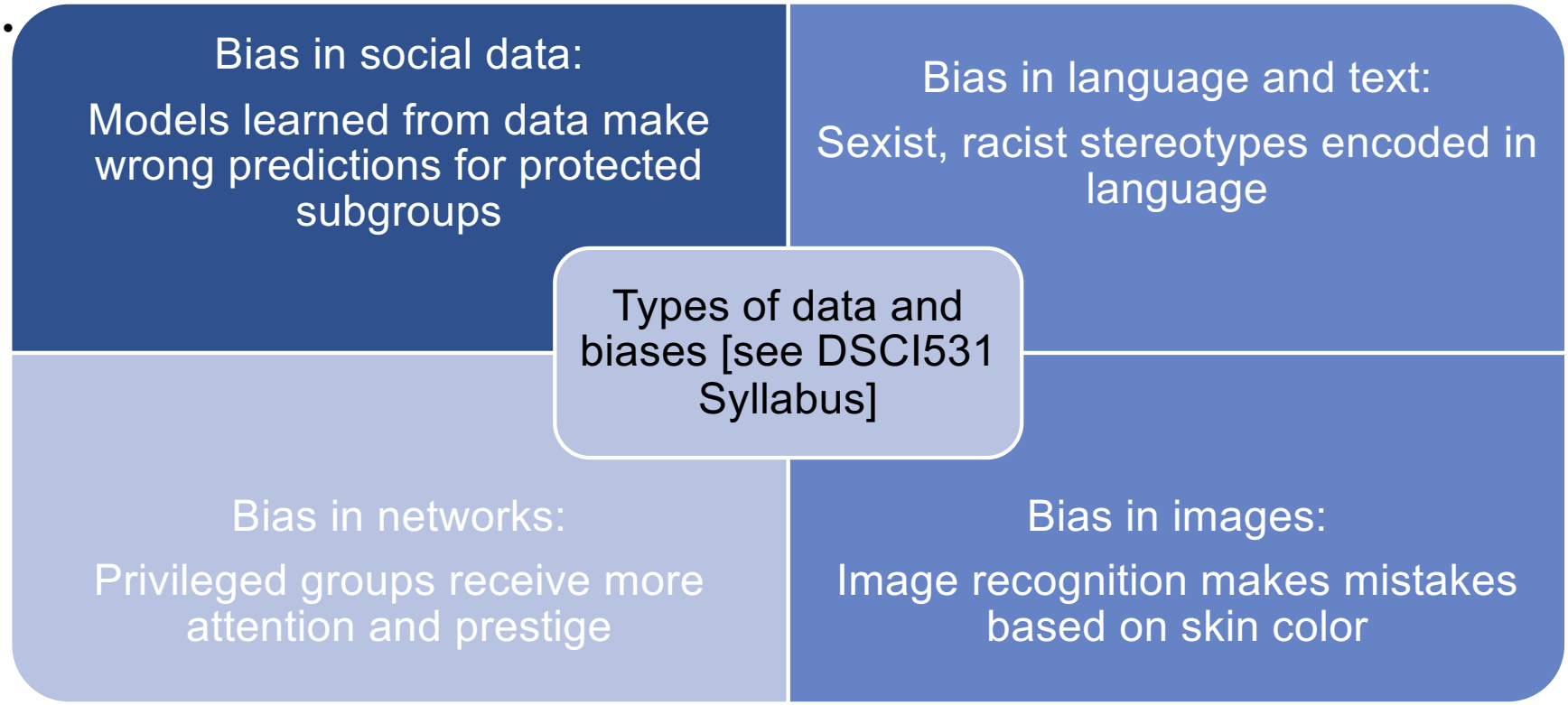
Biases in data are propagated by algorithms



Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

USC Viterbi
School of Engineering

Different types of data results in different biases



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

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

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

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

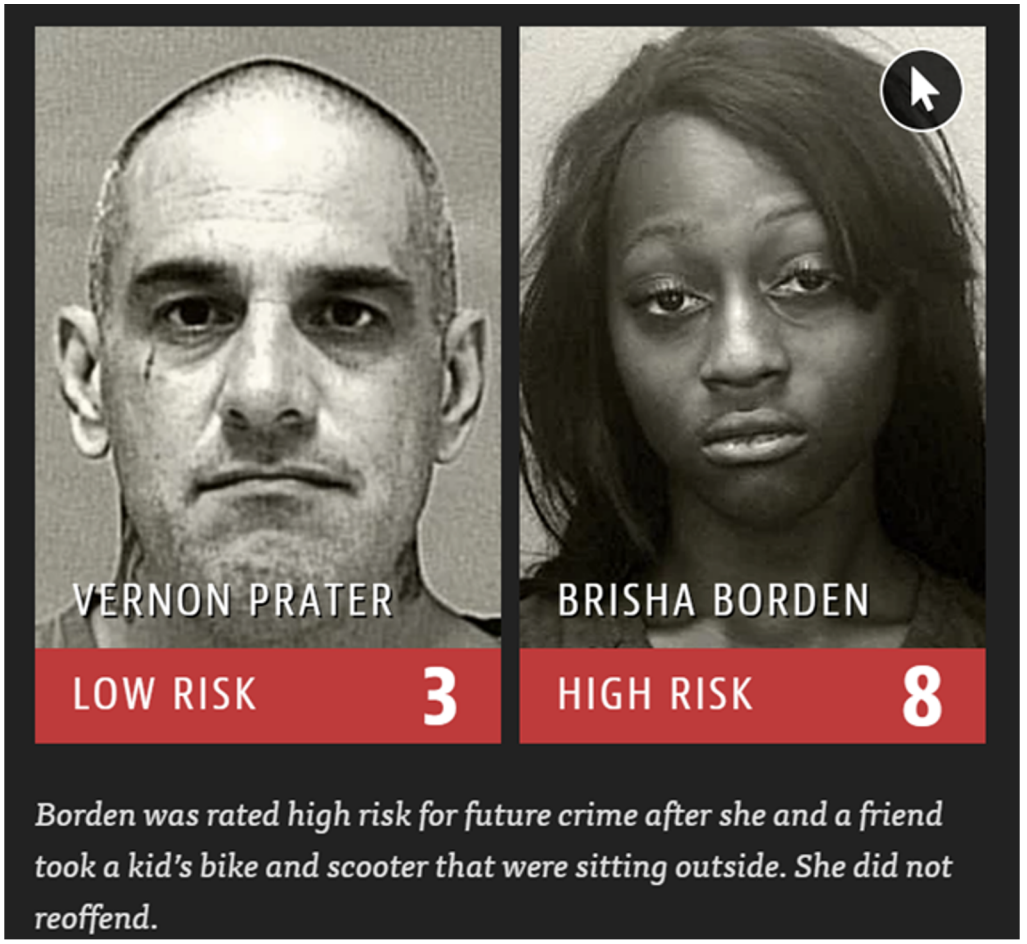
Bias in images: Gender classification errors depend on gender and skin color

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

USC Viterbi
School of Engineering

Bias in automated criminal risk assessment

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



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

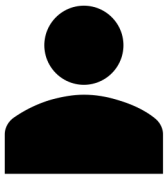
USC Viterbi School of Engineering

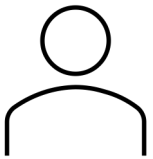
ProPublica study

- Data: 7000 people arrested in Broward County, FL (2013-2014)
  - COMPAS algorithm predicts whether a defendant will commit a crime within next 2 years

	Did commit a crime	**Did not commit a crime**
Will commit crime	True Positives (TP)	*False Positives* (FP)
Will not commit a crime	*False Negatives* (FN)	True Negatives (TN)

Prior Offense	1 attempted burglary
LOW RISK	**3**
Subsequent Offenses	3 drug possessions





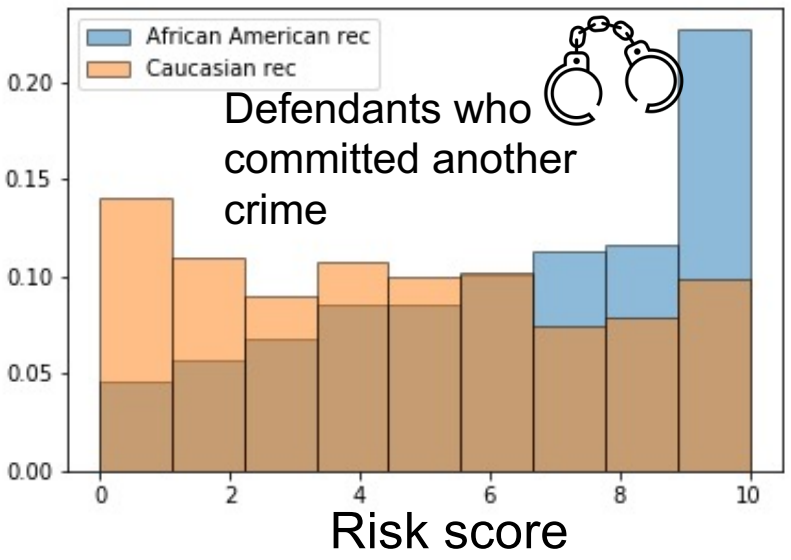
Prior Offense	1 resisting arrest without violence
HIGH RISK	**10**
Subsequent Offenses	None

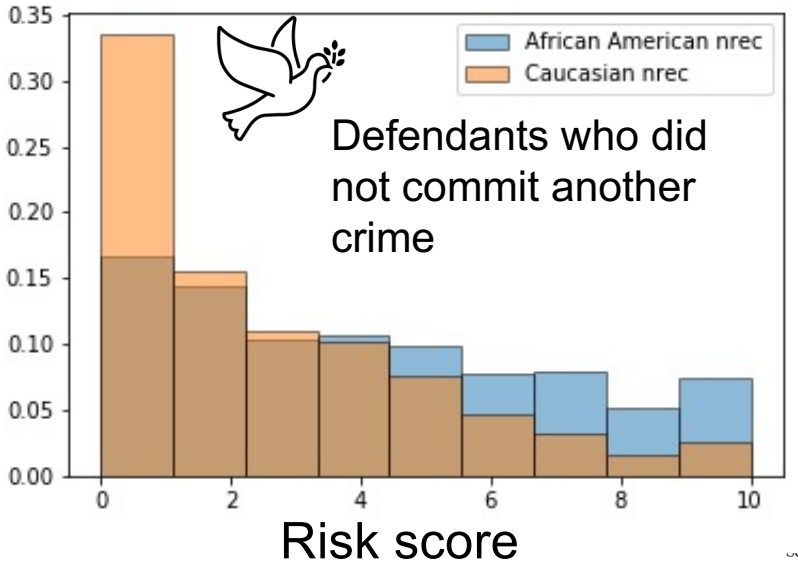
Racial disparities in COMPAS risk scores

Significant racial disparities:

- Algorithm falsely flagged black defendants as future criminals at 2X the rate for white defendants
- Algorithm falsely flagged White defendants as low risk more often than black defendants.

This disparity cannot be explained by prior crimes, type of crimes, their age and gender

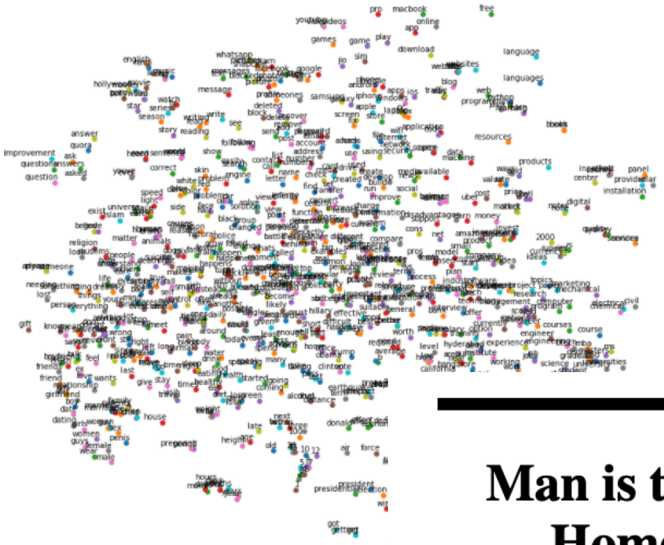


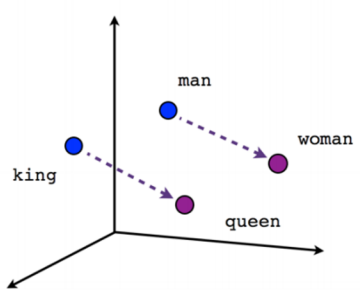


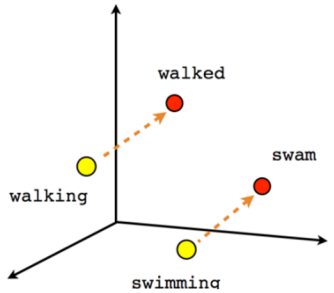
bi
School of Engineering

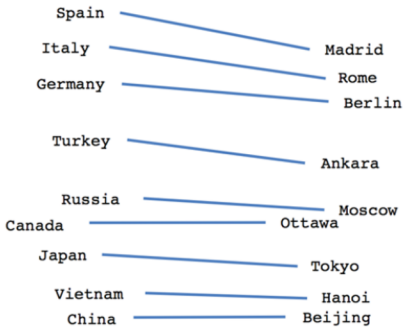
Bias in language

Word Embeddings









Country-Capital

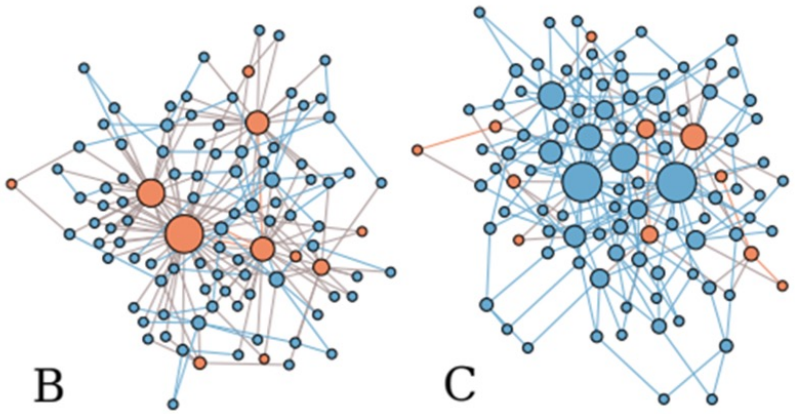
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

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

USC Viterbi
School of Engineering

Bias in networks

- Biased individual preferences for linking to similar/dissimilar individuals can skew how important these groups are in the network
  - B: preference to link to members of a different subgroup gives the minority more power
  - C: preference for linking to members of the same subgroup gives the majority more power



USC Viterbi
School of Engineering

Ethics in USC Data Science Curriculum

- **DSCI-531: Fairness in Artificial Intelligence**
- **Profs. Lerman & Morstatter**
- The course will explore topics in the intersection of data, language, networks and algorithms with fairness and bias through quantitative analysis and hands-on exploration. The course will introduce students to basic and advanced fairness concepts, including methods and apply them to societal data to study fairness and bias, and to understand their effects on learning algorithms. While there are no formal prerequisites for the class, students are expected to be proficient in programming, algorithm design and data structures, and to have taken college level or above courses in linear algebra and statistics. AI and machine learning courses are strongly recommended.

USC Viterbi
School of Engineering