Fairness in ML

These slides were assembled by Eric Eaton, with grateful acknowledgement of the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution. Please send comments and corrections to Eric.

Fairness

- Widespread algorithms with many small interactions
  - e.g. search, recommendations, social media
- Specialized algorithms with fewer but higher-stakes interactions
  - e.g. medicine, criminal justice, finance
- At this level of impact, algorithms can have unintended consequences
- Low classification error is not enough, need **fairness**

Based on materials by David Madras

Regulated domains

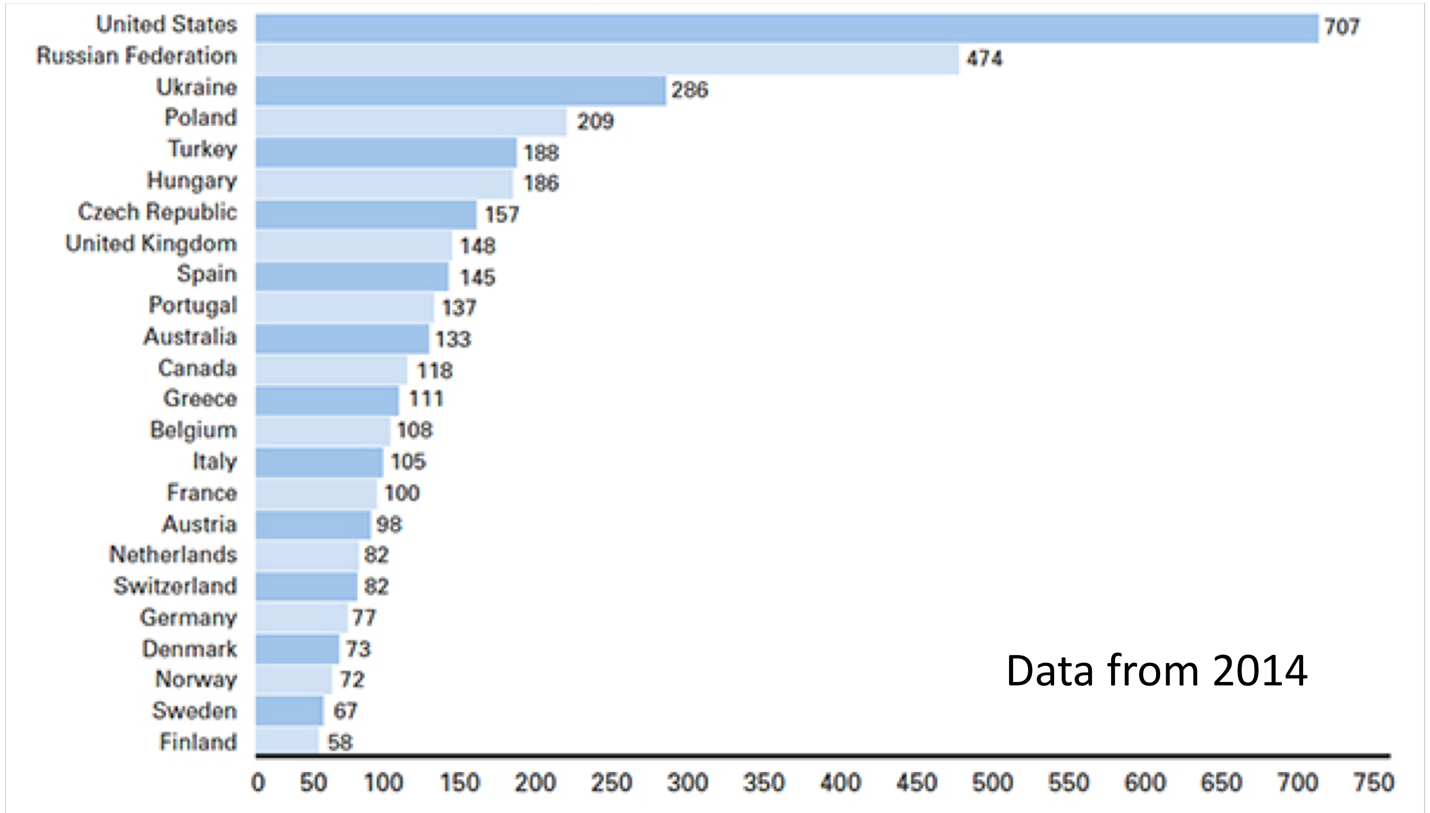
- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **Public Accommodation** (Civil Rights Act of 1964)

Extends to marketing and advertising; not limited to final decision

This list sets aside complex web of laws that regulates the government

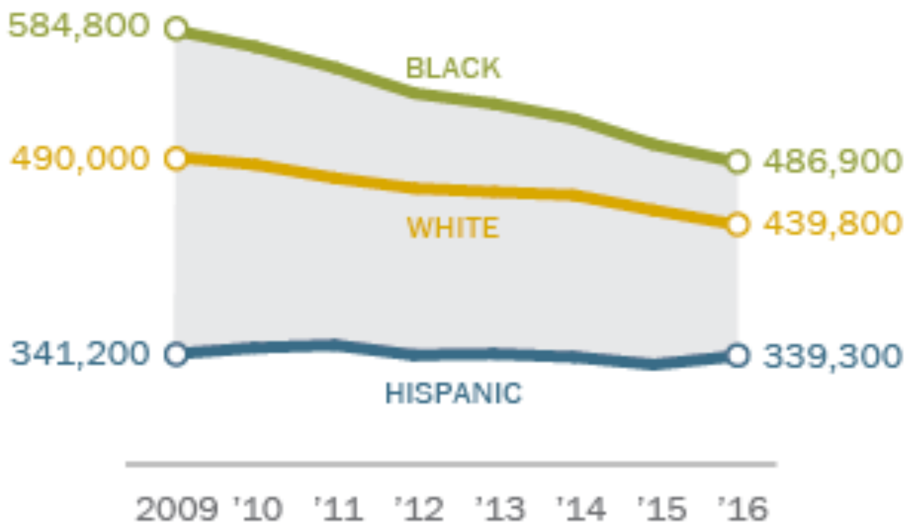
Background on US Prison Population

Incarceration Rates per 100,000



Source: https://www.apa.org/monitor/2014/10/incarceration

Sentenced Federal and State Prisoners by Race and Hispanic Origin, 2009-2016



Note: Whites and blacks include only those who are single-race, not Hispanic. Hispanics are of any race. Prison population is defined as inmates sentenced to more than a year in federal or state prison.
Source: Bureau of Justice Statistics.

PEW RESEARCH CENTER

Source: https://www.pewresearch.org/fact-tank/2018/01/12/shrinking-gap-between-number-of-blacks-and-whites-in-prison/

Case Study: COMPAS

- Software by Northpointe that predicts recidivism
- Used by judges in determining sentencing and bail
- Scores derived from 137 questions answered by defendants or pulled from criminal records:
    - “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 in fights while at school?”
    - Agree or disagree? “A hungry person has a right to steal”
    - Agree or disagree? “If people make me angry or lose my temper, I can be dangerous.”
    - Race is **not** one of the questions
- The exact method of determining the score is kept as a trade secret

Case Study: COMPAS

Table 1: ProPublica Analysis of COMPAS Algorithm

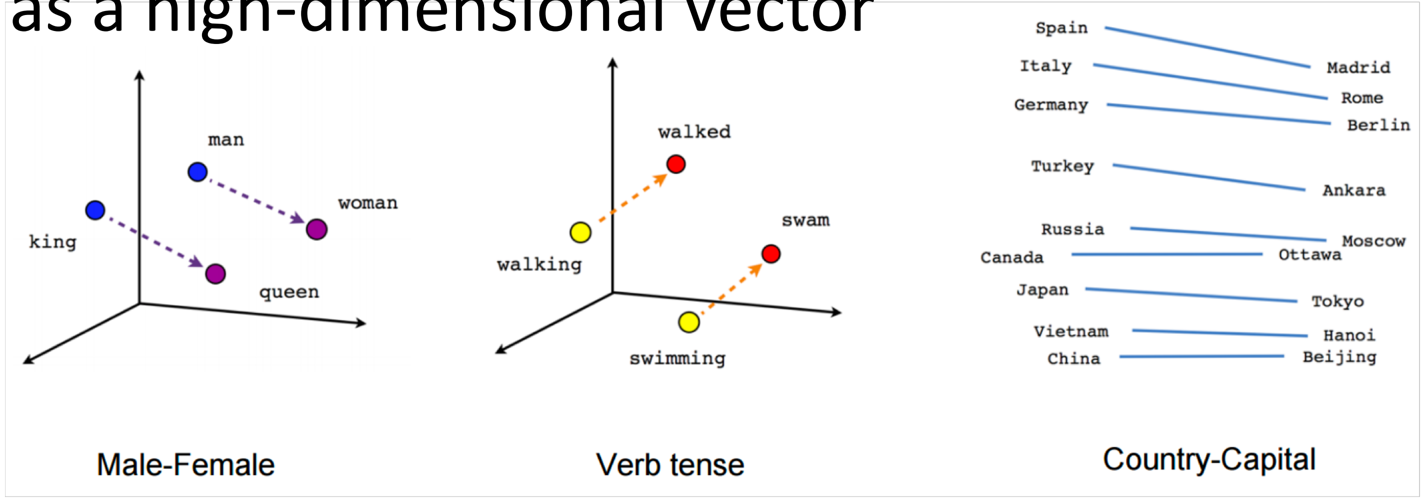
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

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

- African Americans are almost twice as likely as Caucasians to be incorrectly labeled as high risk
- Software predictions can have <u>real</u> consequences

Example: Bias in Word Embeddings (Bolukbasi et al. 2016)

- Studied word2vec word embeddings trained on Google News
- word2vec represents each word as a high-dimensional vector
- Vector arithmetic can be used to answer analogies like:
  - Paris : France ≅ London : England



- Other analogies with stereotyped answers:
  - man : woman ≅ programmer : homemaker
  - man : woman ≅ surgeon : nurse

Gender stereotype *she*-*he* analogies.

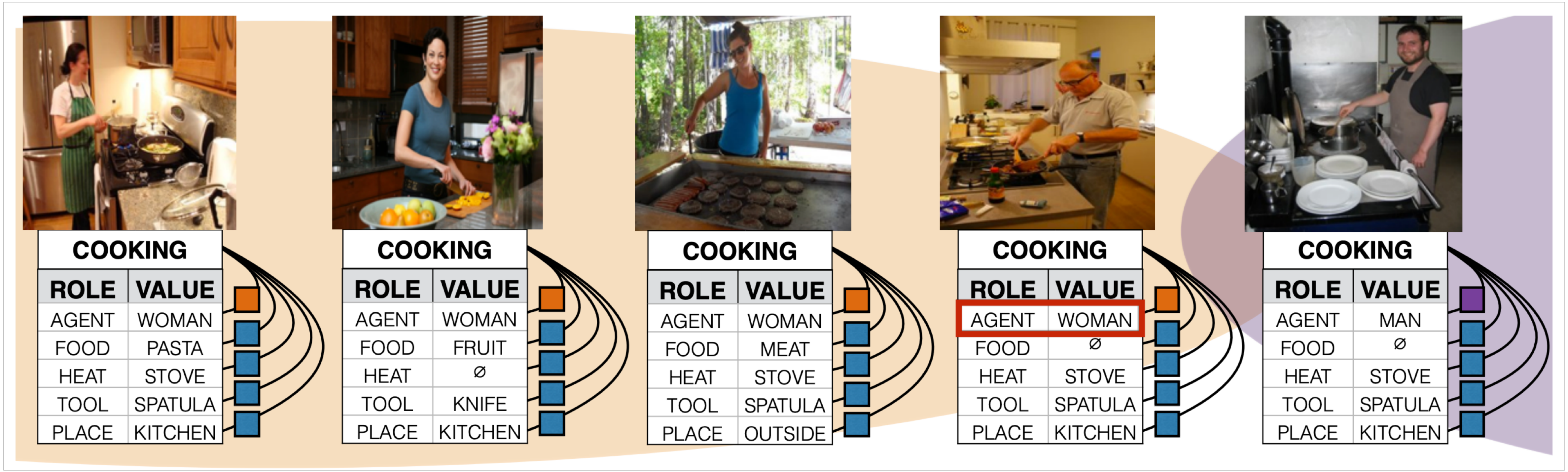
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she*-*he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Example: Bias in Image Classification



- Images from imSitu visual semantic role labeling (vSRL) dataset
  - Only 33% of cooking images are of men
  - Prediction with a (biased) conditional random field only predicts men in 16% of cooking images

Source: https://arxiv.org/pdf/1707.09457.pdf

Algorithmic Fairness

- How can we ensure that our algorithms act in ways that are fair?
  - This definition is vague and somewhat circular
  - Describes a broad set of problems, not a specific technical approach

- Related to ideas of :
  - **Accountability**: who is responsible for automated behavior? How do we supervise/audit machines that have large impact?
  - **Transparency/Explainability**: why does an algorithm behave in a certain way? Can we understand its decisions? Can it explain itself?
  - **AI safety**: how can we make AI without unintended negative consequences?
  - **Aligned AI:** How can AI make decisions that align with our values?

Why Fairness is Hard

- Suppose we are a bank trying to fairly decide who should get a loan
  i.e. Who is most likely to pay us back?
- Suppose we have two groups: A and B (the sensitive attribute)
  - This is where discrimination could occur
- The simplest approach is to remove the sensitive attribute from the data, so that our classifier doesn't know the sensitive attribute

Age	Gender	Employed?	Zip Code	Requested Amount	~~A or B?~~	Grant Loan?
37	F	Yes	24729	$50,000	~~A~~	Yes
23	M	Yes	11038	$30,000	~~B~~	Yes
72	F	No	10038	$90,000	~~A~~	Yes
39	F	Yes	30499	$70,000	~~A~~	No
45	M	No	20199	$60,000	~~B~~	No
68	M	Yes	30029	$50,000	~~B~~	No

Based on materials by David Madras

Legally Recognized “Protected classes”

- **Race** (Civil Rights Act of 1964)
- **Color** (Civil Rights Act of 1964)
- **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964)
- **Religion** (Civil Rights Act of 1964)
- **National origin** (Civil Rights Act of 1964)
- **Citizenship** (Immigration Reform and Control Act)
- **Age** (Age Discrimination in Employment Act of 1967)
- **Pregnancy** (Pregnancy Discrimination Act)
- **Familial status** (Civil Rights Act of 1968)
- **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
- **Genetic information** (Genetic Information Nondiscrimination Act)

Why Fairness is Hard

Age	Gender	Employed?	Zip Code	Requested Amount	A or B?	Grant Loan?
37	F	Yes	24729	$50,000	?	Yes
23	M	Yes	11038	$30,000	?	Yes
72	F	No	10038	$90,000	?	Yes
39	F	Yes	30499	$70,000	?	No
45	M	No	20199	$60,000	?	No
68	M	Yes	30029	$50,000	?	No

- However, this won’t work if the sensitive attribute is correlated with others
  - E.g., it is easy to predict race given other info (home address, financials, etc.)
- We need more sophisticated approaches

Group Fairness

- Key idea: “Treat different groups equally”
- Assess fairness based on **demographic parity**: require that the same percentage of A and B receive loans
    - What if 80% of A is likely to repay, but only 60% of B is?
    - Then demographic parity is too strong
- Could require equal false positive/negative rates
    - When we make an error, the direction of that error is equally likely for both groups
        - P(loan|no repay, A) = P(loan|no repay, B)
        - P(no loan|would repay, A) = P(no loan|would repay, B)

Individual Fairness

- Key idea: “Treat similar examples similarly”
- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaptation
  - Generative modelling/adversarial approaches

Looking Forward

- This is an open and active area of research
- Lots of progress, long way to go
- Law will catch up with ML technology eventually