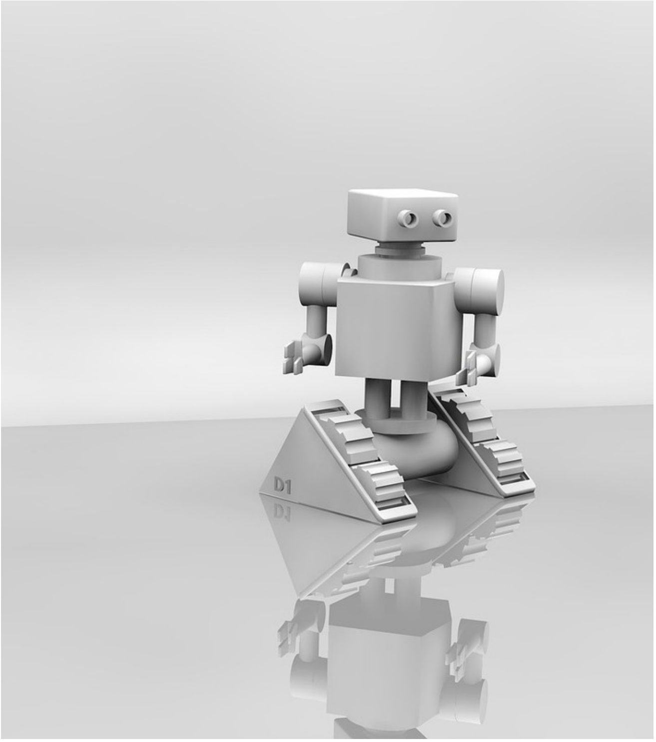


CPSC 481 Artificial Intelligence

Dr. Mira Kim

Mira.kim@fullerton.edu

What we will cover this week

- The Ethics of AI
  - Fairness and Bias
  - Privacy

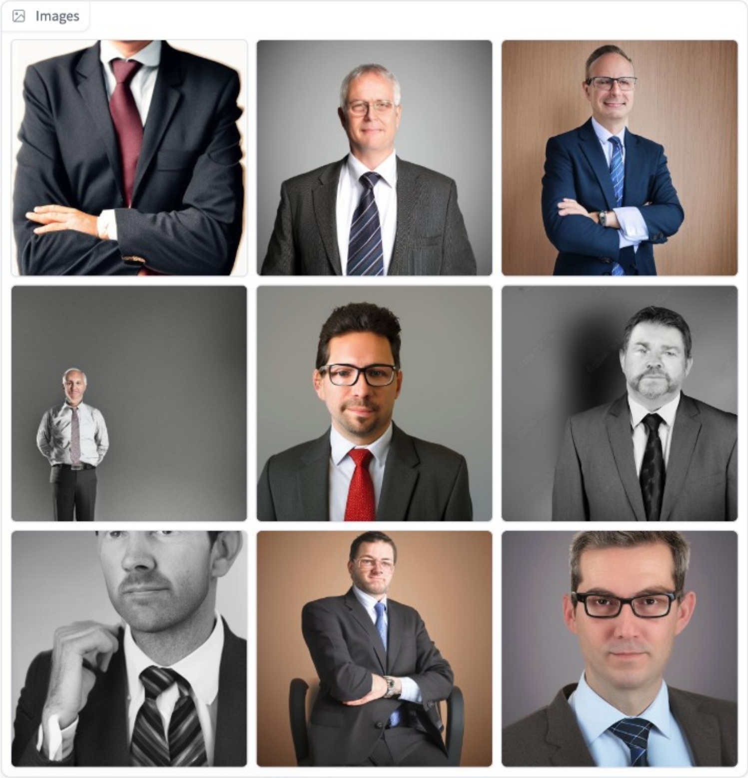
The Ethics of AI

- Every organization that creates AI technology, and everyone in the organization, has a responsibility to make sure the technology contributes to good, not harm.

- Ensure fairness
- Provide transparency
- Respect privacy
- Ensure safety
- Limit harmful uses of AI
- Promote collaboration

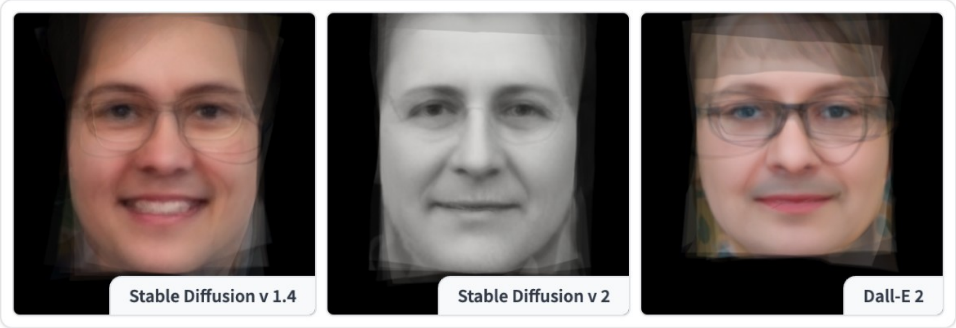
- Establish accountability
- Uphold human rights and values
- Reflect diversity/inclusion
- Avoid concentration of power
- Acknowledge legal/policy implications
- Contemplate implications for employment

Bias in image models



"Manager" by Stable Diffusion.

Model generated white men 97% of the time when given prompts like “CEO” or “director.”

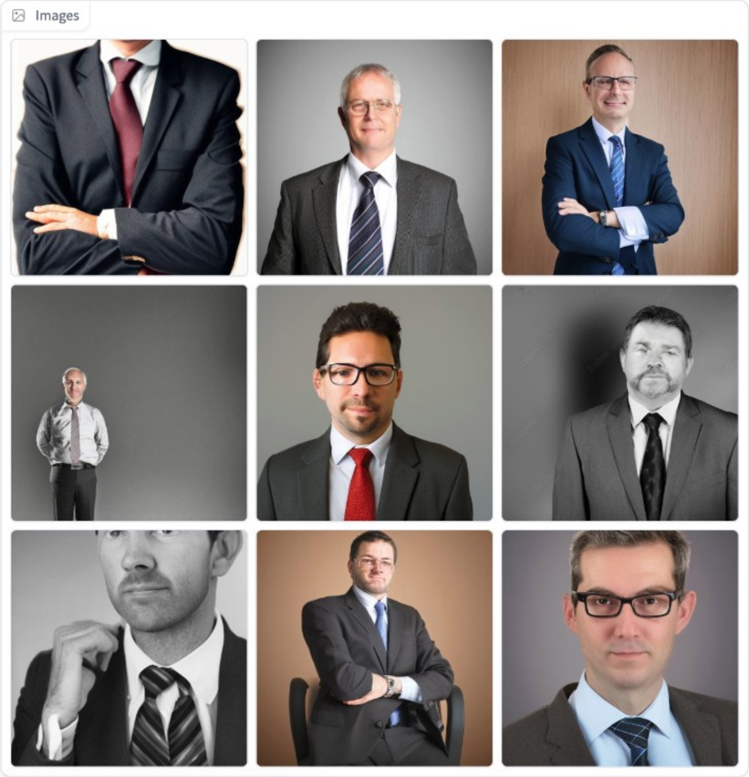


The average face of a teacher generated by Stable Diffusion and DALL-E 2.

Average face of a teacher

Source - MIT Technology Review
https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

Bias in image models



"Manager" by Stable Diffusion.

"Compassionate manager" by Stable Diffusion.

- Attaching different adjectives changes images
- -> Stereotypical gender biases

Adding “compassionate,” “emotional,” or “sensitive”

-> generate a woman

Adding “stubborn,” “intellectual,” or “unreasonable”

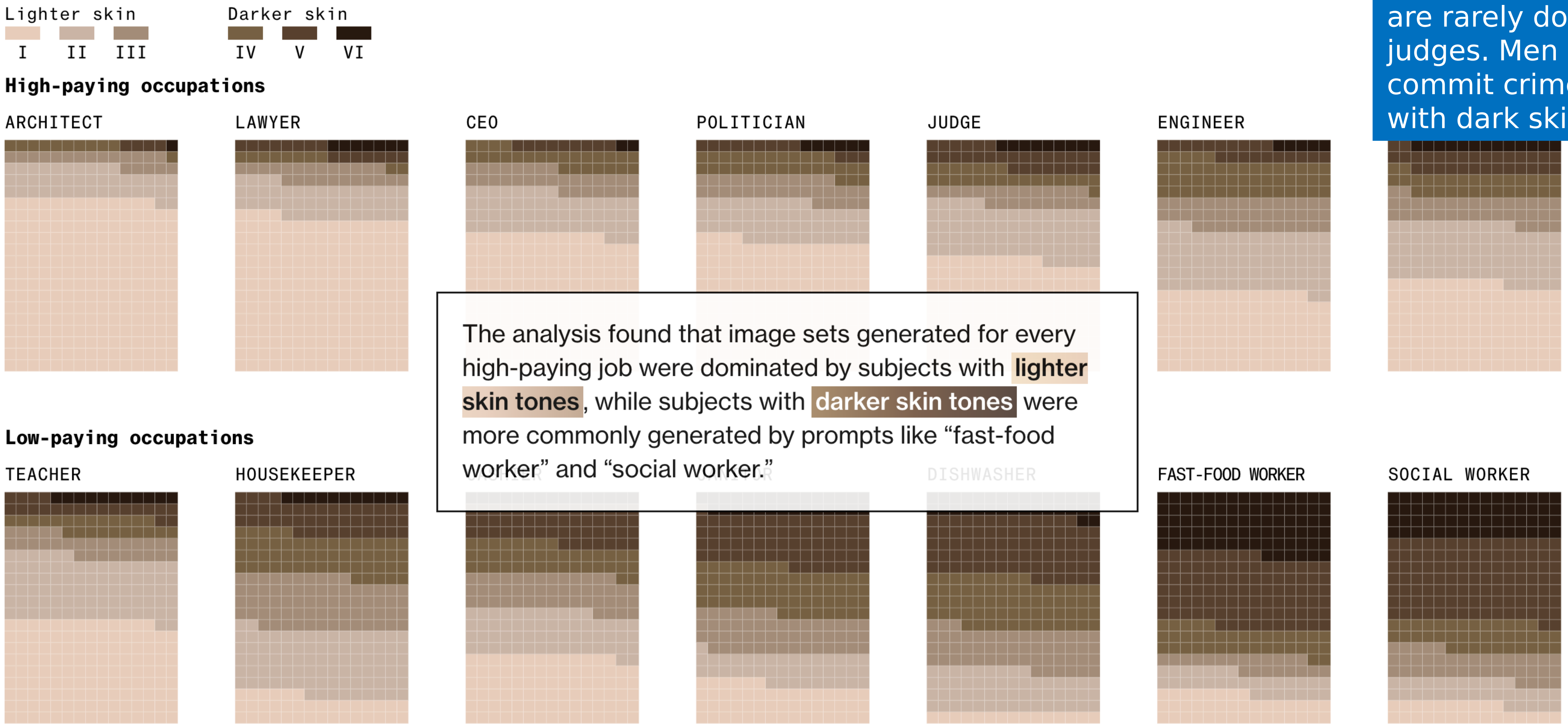
-> generate a man

Source - MIT Technology Review
https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

Bias in image models

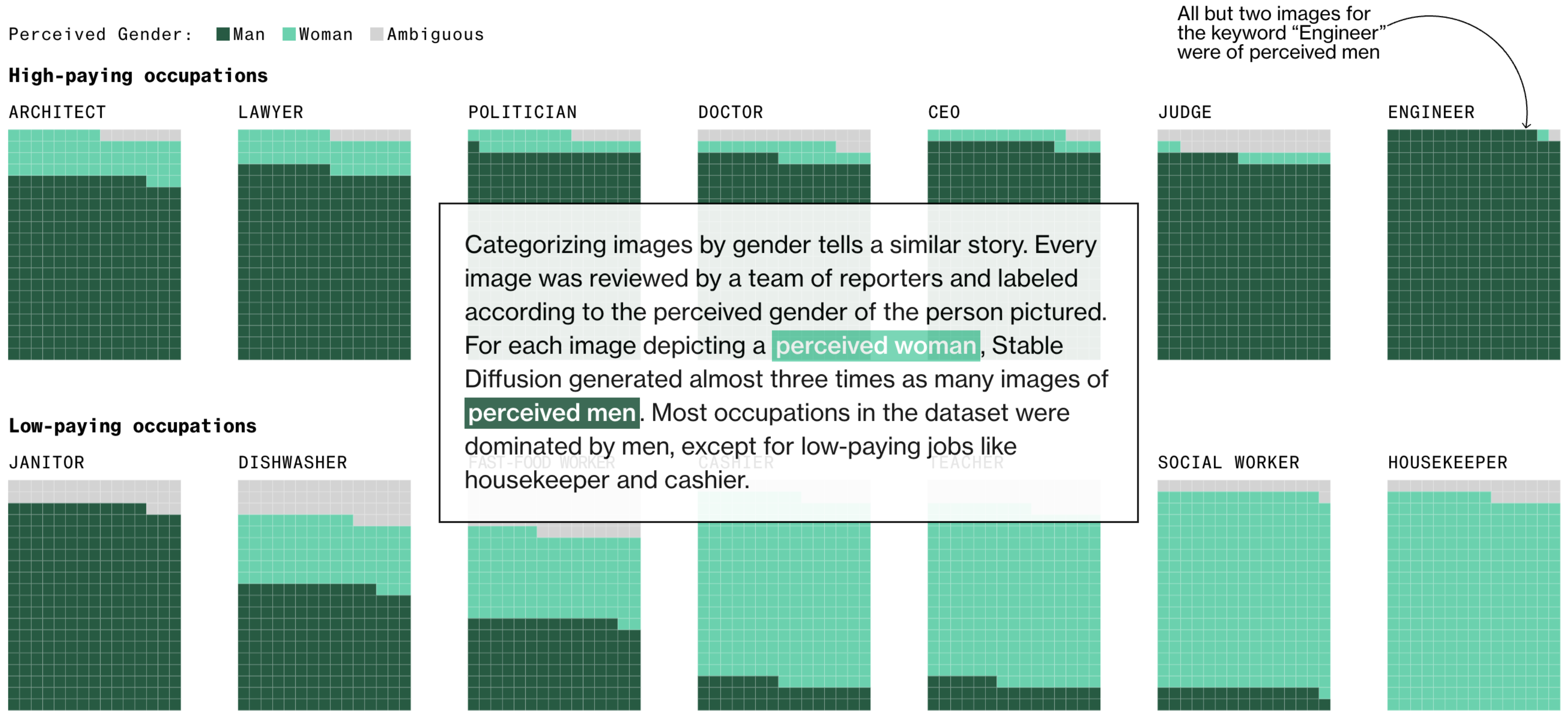
Bloomberg article - Humans are biased. Generative AI is even worse

The world according to Stable Diffusion is run by White male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes, while women with dark skin flip burgers.



Source: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bias in image models



Source: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

FAIRNESS AND BIAS

Fairness and bias

- AI/ML systems are being deployed in complex high-stakes settings
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
  - Non-discrimination (“fairness”)
  - Right to explanation
  - Safety

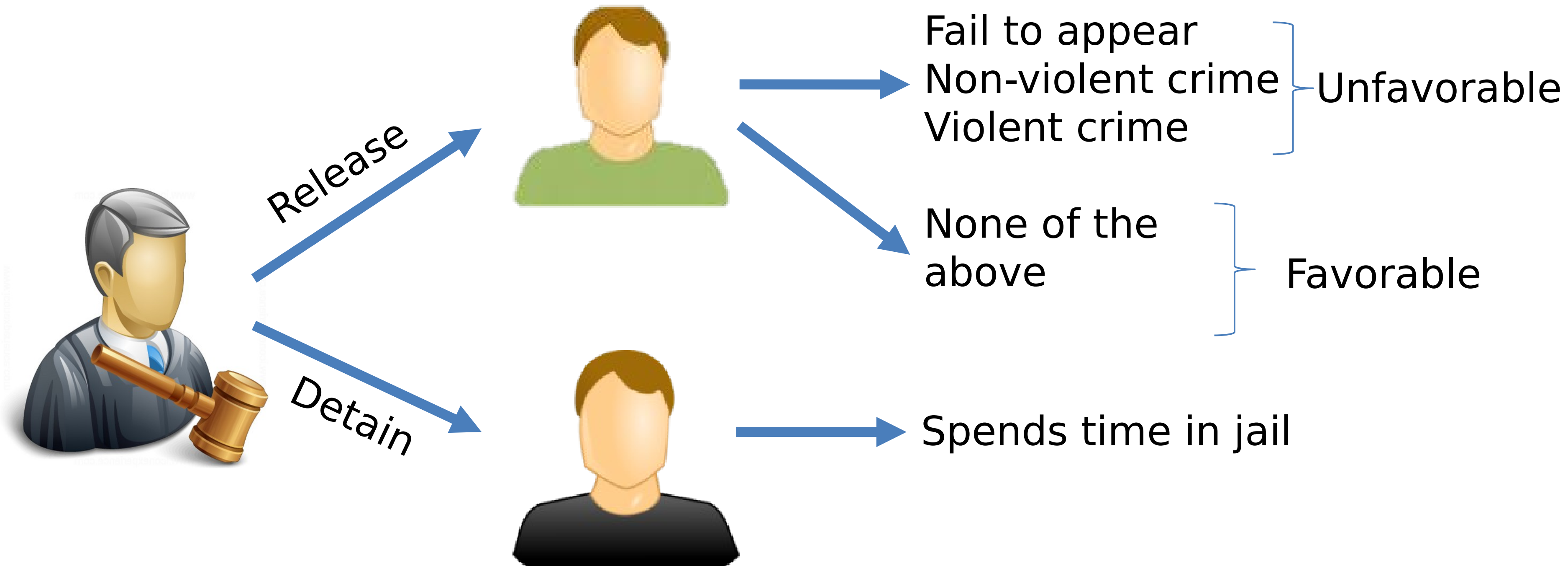
Real World Scenario: Bail Decision

- U.S. police make about 12M arrests each year

We consider the binary decision

- Release vs. Detain is a high-stakes decision
  - Pre-trial detention can go up to 9 to 12 months

Bail Decision



Judge is making a prediction:
Will the defendant commit ‘crime’ if released on bail?

Bail Decision-Making as a Prediction Problem

- Build a model that predicts defendant behavior if released based on their characteristics

Training examples ⊆ Set of All Released Defendants

Defendant Characteristics				Outcome
Age	Prev. Crimes	Level of Charge	...	
28	2	Felony	...	Crime
14	1	Misd.	...	No Crime
63	0	Misd.	...	No Crime
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.

Learning algorithm

Predictive Model

Test case

Defendant Characteristics				Outcome
35	3	Felony	.	?

Prediction:
Crime (0.83)

Accuracy and fairness

Where does bias come from?

1. Data reflects bias in the systems/people that create training data
   – "Bias in, bias out"
2. Unintended feedback loops
   – Send more police to one neighborhood → more arrests made → more crime! → send even more police → …
3. Data does not reflect true objectives
   – Have only arrest records, not who actually committed crimes
4. Different features have different correlations for different sub-groups
   – E.g., Do SAT verbal scores indicate college success for men/women equally?

Accuracy and fairness

- Model was optimized for accuracy
- But is it fair?
  - Is bail granted at equal rates for different groups?
  - Not necessarily – accuracy and fairness are different objectives
- What is fairness (for an algorithm)?
  - *Statistical parity*: fairness metrics in AI - the proportion of positive decisions made be the same for different sub-groups
  - *Equality of false negatives*: model mistakes must be made at the same rate for different sub-groups

Accuracy and fairness: solution 1

- Optimize accuracy and fairness objectives together
  - If only accuracy is maximized, algorithm will naturally optimize better for the majority population, as that has a greater impact on overall accuracy
  - Problem 1: not always possible for one model to maximize accuracy and maximize fairness
    - In fact, different fairness objectives can themselves conflict
      - Statistical parity vs. equality of false negatives
    - Humans/society will have to decide where in the accuracy-fairness trade-off curve we want to be
      - Could be different for different applications: bail decisions are more critical than targeted ads
  - Problem 2: models optimizing for different sub-groups can unintentionally discriminate people at the intersections of these same sub-groups
    - Model is fair with race and gender separately, but unfair to a particular race-gender class

Accuracy and fairness: solution 2

- Make models interpretable:
  - Create a human-understandable explanation of model predictions
  - Versus “black box” model

Bail Decision-Making as a Prediction Problem

Build a model that predicts defendant behavior if released based on his/her characteristics

Training examp

Defenda Character				
Age	Prev. Crimes	Le Ch		
28	2	Fe		
14	1	M		
63	0	M		
.	.			
.	.	.	...	.
.	.	.	...	.

Does making the model more understandable/transparent to the judge improve decision-making performance?
If so, how to do it?

Test case

Defendant Characteristics				Outcome
35	3	Felony	.	?

Predictive Model

Prediction:
Crime
(0.83)

Give a human-readable explanation of the model

If Current-Offense = Felony:
If Prior-Felony = Yes and Prior-Arrests ≥ 1, then Crime
If Crime-Status = Active and Owns-House = No and Has-Kids = No, then Crime
If Prior-Convictions = 0 and College = Yes and Owns-House = Yes, then No Crime

If Current-Offense = Misdemeanor and Prior-Arrests > 1:
If Prior-Jail-Incarcerations = Yes, then Crime
If Has-Kids = Yes and Married = Yes and Owns-House = Yes, then No Crime
If Lives-with-Partner = Yes and College = Yes and Pays-Rent = Yes, then No Crime

If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
If Has-Kids = No and Owns-House = No and Moved_10times_5years = Yes, then Crime
If Age ≥ 50 and Has-Kids = Yes, then No Crime

Default: No Crime

Judges were able to make decisions 2.8 times faster and 38% more accurately (compared to no explanation and only prediction)

Motivation for Interpretability

- Auxiliary criteria are often hard to quantify (completely)
  - E.g.: “Safety”
    - Impossible to enumerate all scenarios violating safety of an autonomous car

- Fallback option: interpretability
  - *If the system can explain its reasoning, we can verify if that reasoning is sound with regards to auxiliary criteria like fairness, safety*

In-class exercise

- Discuss how you would design an interpretable AI system for the following scenarios. Write a human-readable explanation of the system. (Follow a similar format as slide 18)

1. Medical Diagnosis: A hospital uses an AI system to diagnose heart diseases. Based on certain criteria, the system concludes a low, medium or high risk of heart disease.
2. Loan Approvals in Banking: A bank uses an AI system to determine whether to approve or reject a loan application. The system examines different factors to make decisions.

PRIVACY

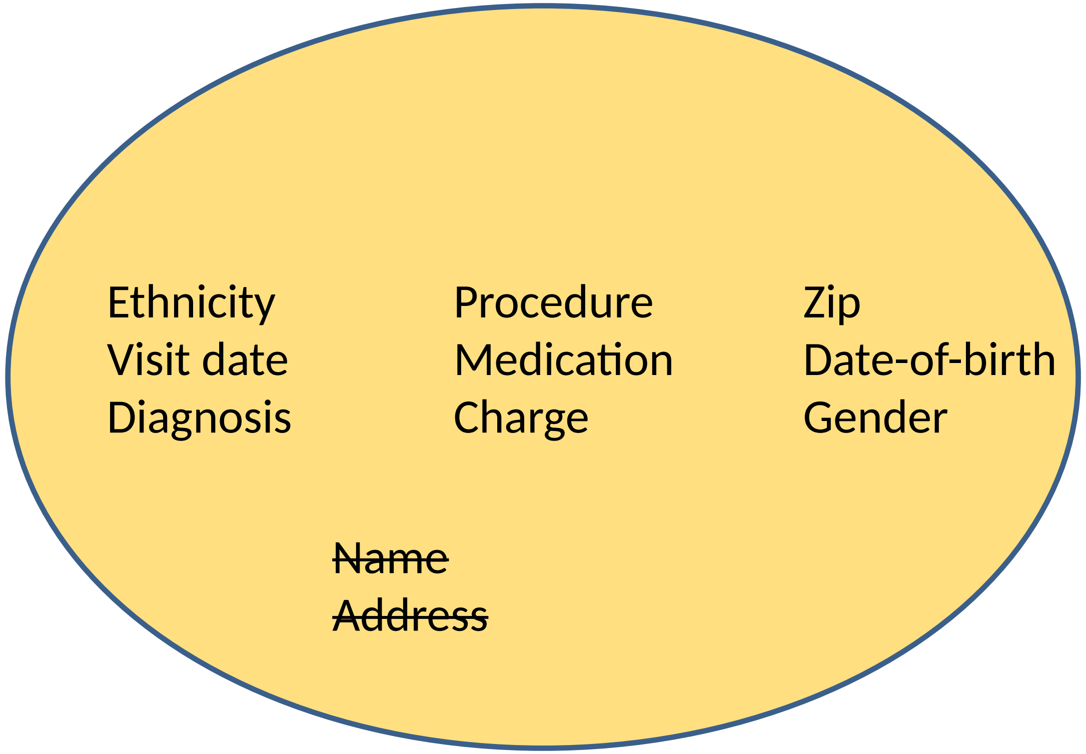
Privacy

- Data collectors have a moral and legal responsibility to be good stewards of the data they hold
- Individual’s right to privacy versus the value that society gains from sharing data
  - We want to cure diseases without compromising any individual’s right to keep their health history private

Privacy

- De-identification: eliminating personally identifying information (such as name and social security number) before releasing data to the public
  - E.g., so that medical researchers can use the data in analyses

Privacy: de-identification



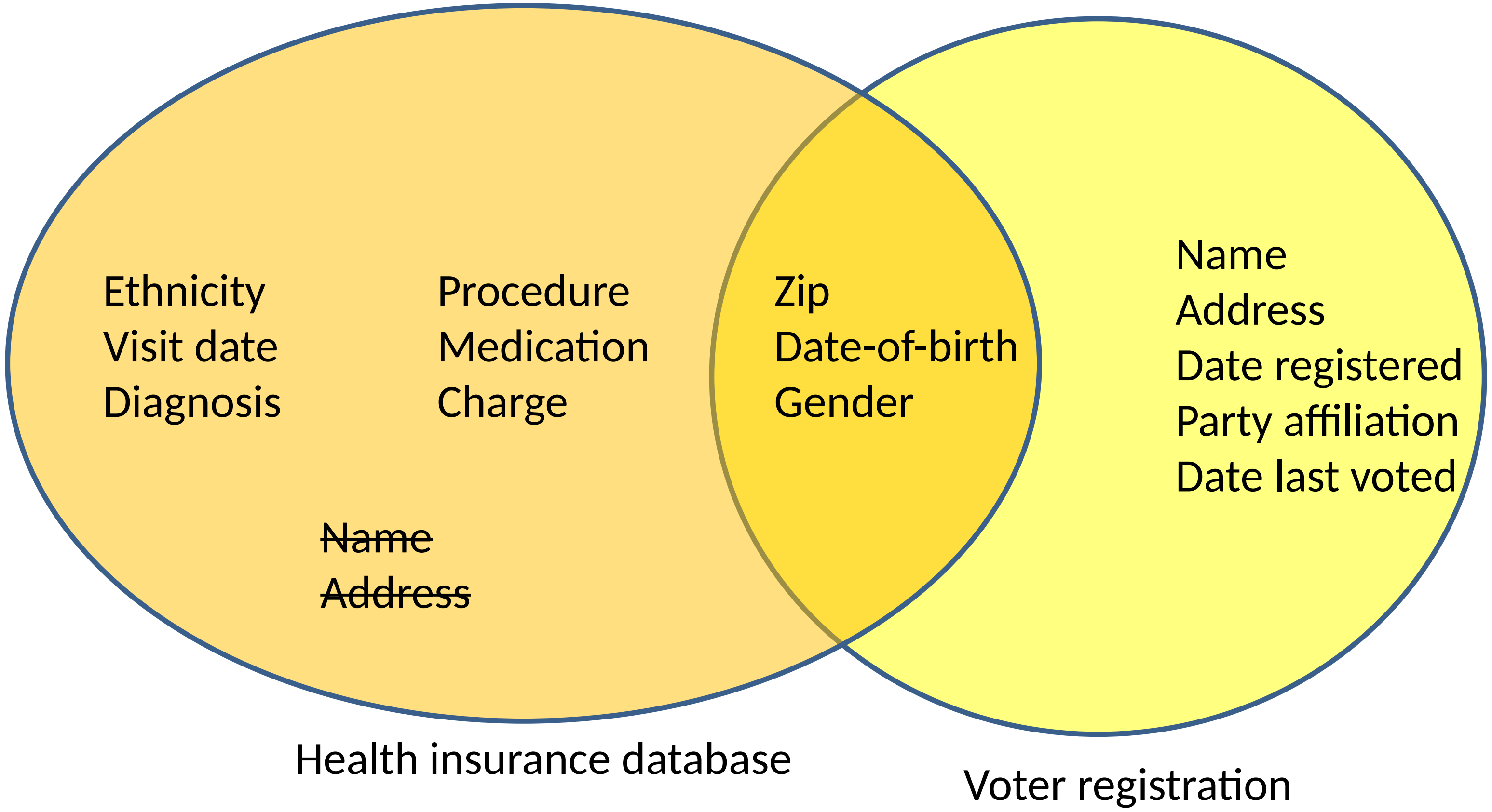
Health insurance database

Privacy: de-identification

- Are "de-identified" data more secure?
  - **Not necessarily!**
- 87% of the US population can be uniquely identified by ZIP code, gender, and date of birth (Sweeney, 2002).
- Sweeney identified William Weld, governor of Massachusetts, in a health insurance database for state employees by purchasing voter registration for Cambridge, Massachusetts, for $20 and linking ZIP code, gender, and date of birth to the "de-identified" medical database (Sweeney, 1997).

[Dr. Latanya Sweeney]

Privacy: de-identification



Need for Privacy

- The data contains:
  - Attribute values which can uniquely identify an individual
    - { zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals
    - { medical condition, salary, location }

	Identity-related Data			*Sensitive Data*	
#	***Zip***	***Age***	***Nationality***	***Name***	***Condition***
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Need for Privacy

	Identity-related Data			*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***Condition***
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Published Data

Data leak!

#	*Name*	*Zip*	*Age*	*Nationality*
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List

Source of Problem

- Even if we remove the direct uniquely identifying attributes
  - There are some fields that may still uniquely identify some individual!
  - The attacker can *join* them with other sources and identify individuals

	Identity-related Data			*Sensitive Data*
#	**Zip**	**Age**	**Nationality**	**Condition**
…	…	…	…	…

Quasi-Identifiers

Identifier: information that uniquely identifies a person
E.g., full name, or SSN
Quasi-Identifier: information that **can** identify a person **when combined with another dataset**

A solution: *K*-anonymity

- Proposed by Sweeney
- Change data in such a way that for each tuple in the resulting table there are at least (*k*-1) other tuples with the same value for the quasi-identifier – **K-anonymized table**

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Techniques for anonymization

1. Data Swapping – interchange values between records
2. Randomization – add random noise to the data
3. Generalization - replace the original value by a semantically consistent but *less* specific value
4. Suppression - data not released at all

Techniques for anonymization

#	*Zip*	*Age*	*Nationality*	*Condition*
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Generalization

Suppression (cell-level)

In-class Exercise

2-anonymize this dataset (for the Identity-related columns)

	Identity-related Data			*Sensitive Data*
#	**Zip**	**Age**	**Insurance**	**Condition**
1	13053	28	Kaiser	Heart Disease
2	13067	29	Kaiser	Heart Disease
3	13053	35	Aetna	Viral Infection
4	13067	36	Aetna	Cancer

References

- Russel and Norvig, Artificial Intelligence: A Modern Approach, 4th edition
  - Section 27.3 “The Ethics of AI”
- https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/
- https://www.bloomberg.com/graphics/2023-generative-ai-bias/
- Hima Lakkaraju, CS282BR: Topics in Machine Learning Interpretability and Explainability