Data Science in the Business World

BUSI/COMP 488

Daniel M. Ringel

UNC Kenan-Flagler Business School
Spring 2023

March 23, 2023

Class 19: Algorithmic Bias

Identification and Prevention

UNC KENAN-FLAGLER BUSINESS SCHOOL

NOTICE

This lecture is about sources and examples of algorithmic bias. The objective of this lecture is to stimulate a discussion on how to recognize and overcome biases in machine learning, artificial intelligence, and in us humans.

Discussion topics and examples can affect individuals differently. They can be painful, hurtful, shocking and may even feel offensive. Topics of this lecture include race, gender and sexual orientation.

For these reasons, this lecture is not compulsory for BUSI/COMP 488 in 2023.
Students may skip this lecture or leave at any point in time without having to provide any justification.

At the same time, I hope to receive constructive feedback from anyone who attended or watched this lecture. Let me know how I can improve this lecture, make it easier to digest, tune topics or language. Should I unintentionally communicated anything in a way that is either offensive or not conducive to students' learnings, please let me know how I can correct myself in the future.

If you have concerns about this lecture that you are hesitant to share with me directly, please contact your respective program office, or UNC's Equal Opportunity and Compliance Office on the internet at https://eoc.unc.edu/, by phone 919.966.3576, or by e-mail at eoc@unc.edu

Credit: Steven Ray Miller

What do you see?

What do you see?

What do you see?

What do you see?

Prototype Theory

One purpose of categorization is to ***reduce the infinite differences*** among stimuli ***to*** behaviorally and ***cognitively usable proportions***

There may be some central, ***prototypical notions*** of items that arise from stored typical properties for an object category (Rosch, 1975)

Fruit

Bananas
"basic level"

Unripe Bananas

Rosch, E. and Mervis, C.B., 1975. Family resemblances: Studies in the internal structure of categories. Cognitive psychology, 7(4), pp.573-605.

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and Exclaims:

"I can't operate on this boy, he is my son!"

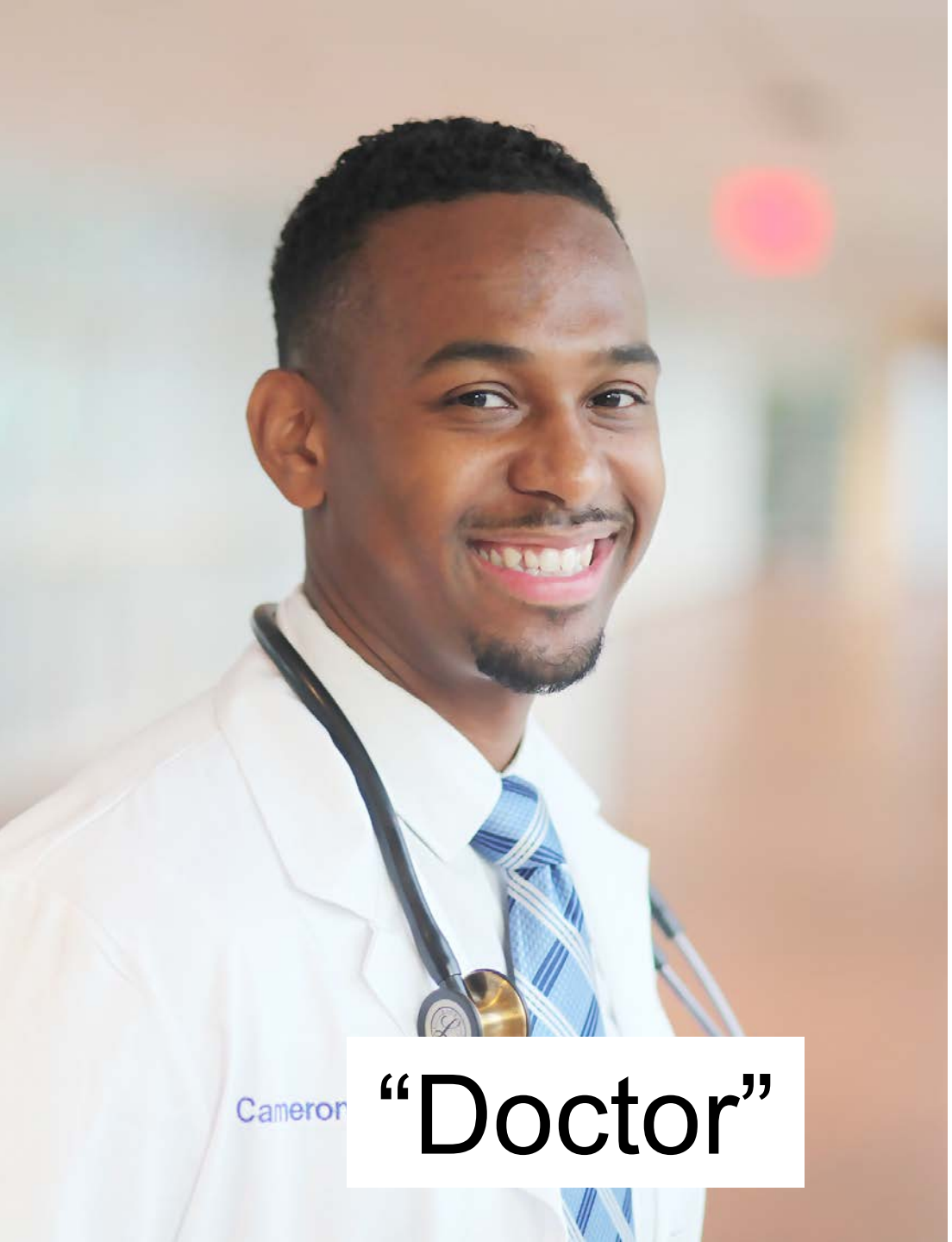
How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and Exclaims:

"I can't operate on this boy, he is my son!"

How could this be?!



“Doctor”

“Female Doctor”

The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

https://www.bu.edu/today/2014/bu-research-riddle-reveals-the-depth-of-gender-bias/

Basic Text Analysis

World learning from Text

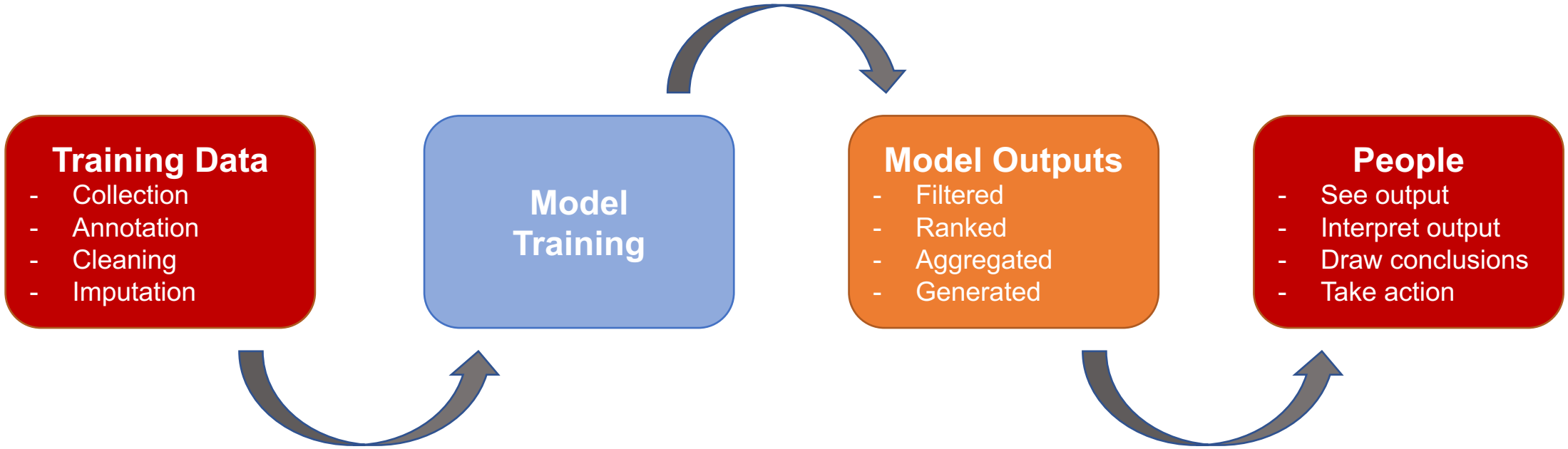
Gordon and Van Durme, 2013

Word	Frequency in Test Corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhaled”	168,985

Gordon, J. and Van Durme, B., 2013, October. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction* (pp. 25-30).

Human Reporting Bias

Common Workflow in Machine Learning



The Root of Algorithmic Biases are often the Training Data

Training Data
- Collection
- Annotation
- Cleaning
- Imputation

Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity

Common Biases in Machine Learning Applications

Training Data

People:
Interpretation!

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

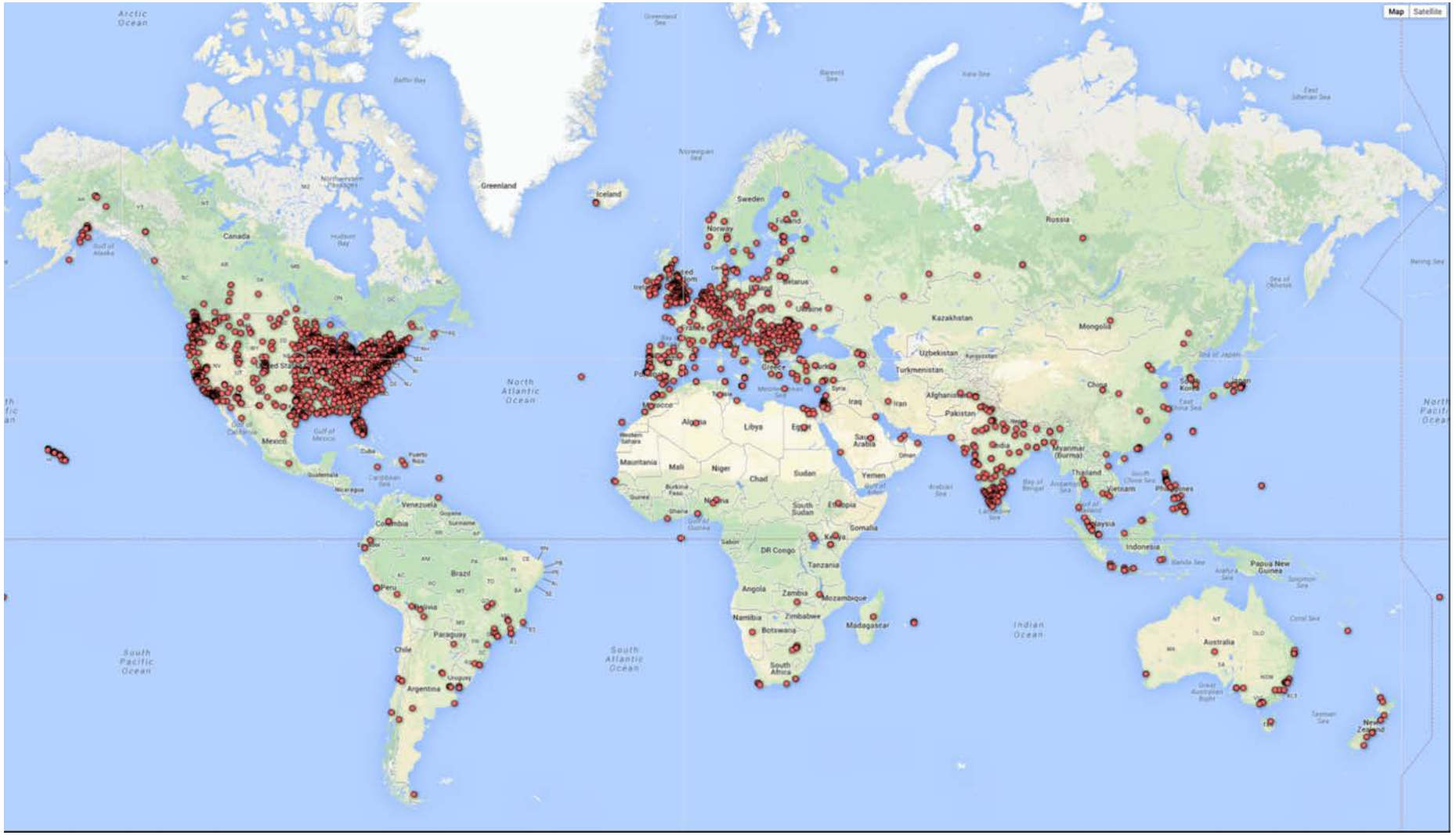
Correlation fallacy: Confusing correlation with causation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

Learn more at: https://developers.google.com/machine-learning/glossary/

Biases in Data

Selection Bias: Selection does not reflect a random sample



© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek

Distribution of Amazon Mechanical Turk Workers across the world.

Much research and model training relies on crowd-sourced annotation and labeling.

Why might this be a problem?

Biases in Data

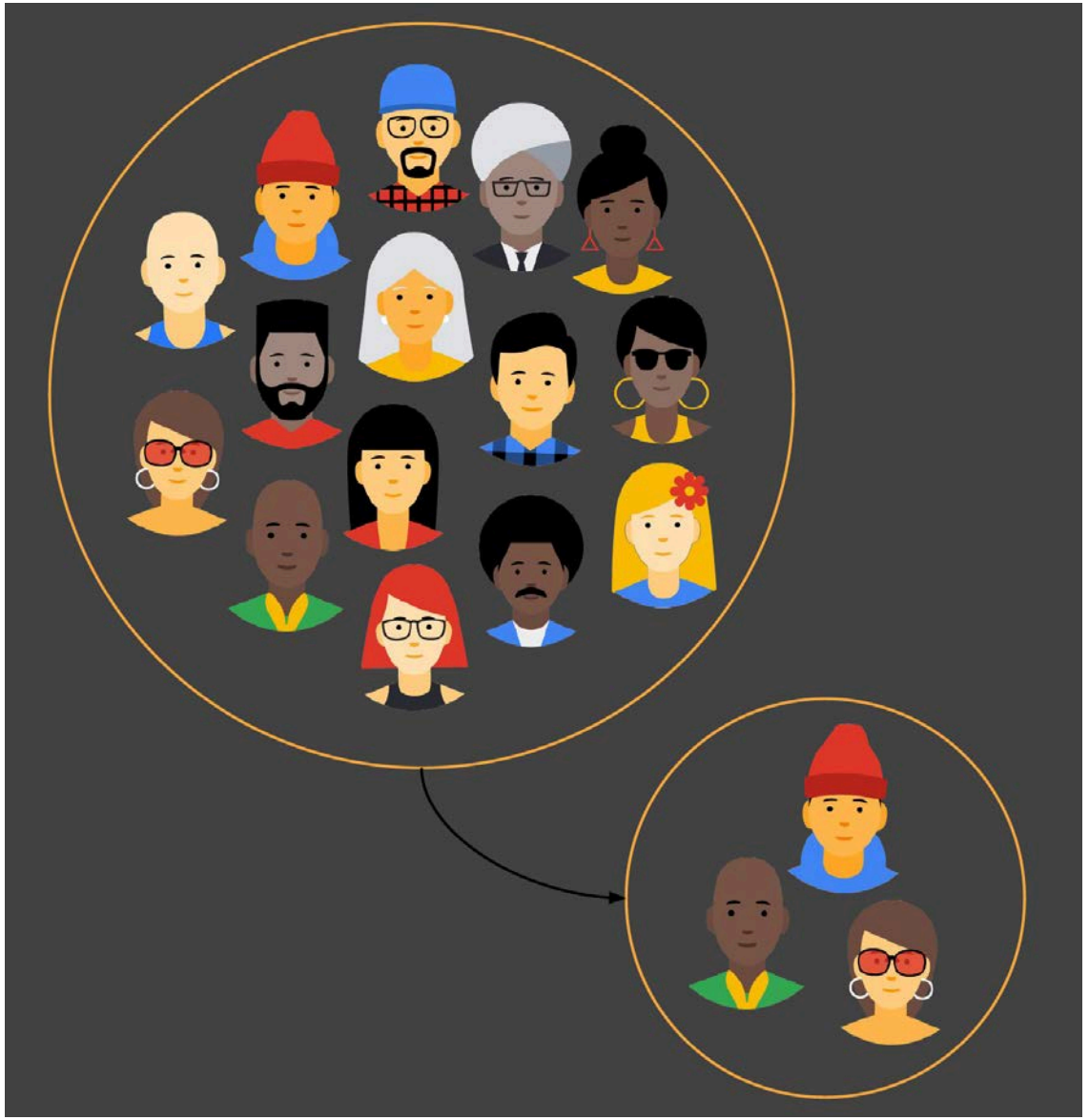
Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

Biases in Data

Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of
but
that some groups are represented less positively than others.



https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

Biases in Data

Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators

read more: https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Biases in Interpretation

Confirmation Bias: The tendency to search for, interpret, favor, and/or recall information in a way that confirms preexisting beliefs

© kris straub - Chainsawsuit.com

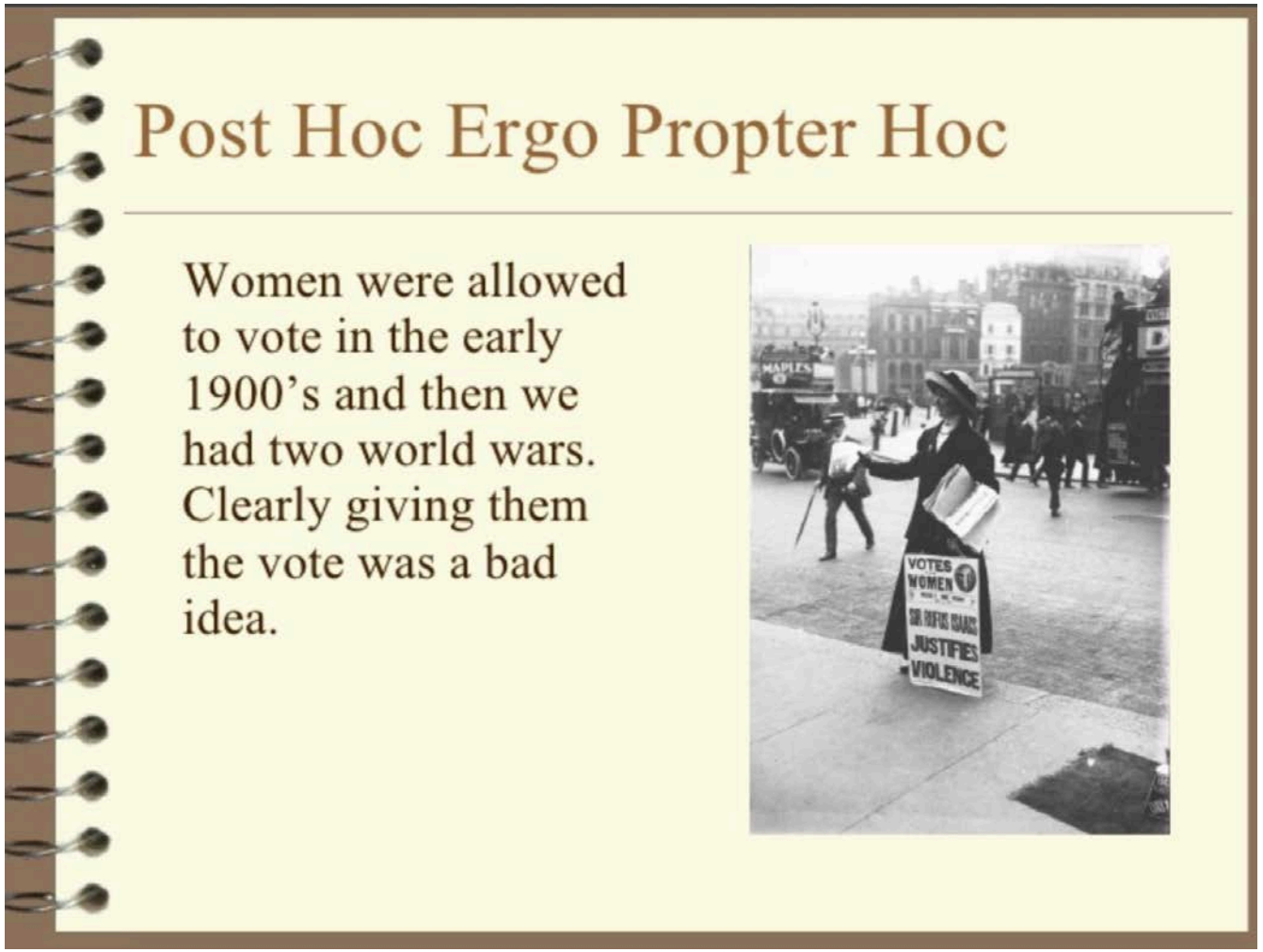
Biases in Interpretation

Overgeneralization: Coming to a conclusion based on information that is too general and/or not specific enough (related: overfitting)

http://www.sciencecartoonsplus.com/index.php

Biases in Interpretation

Correlation fallacy: Confusing correlation with causation



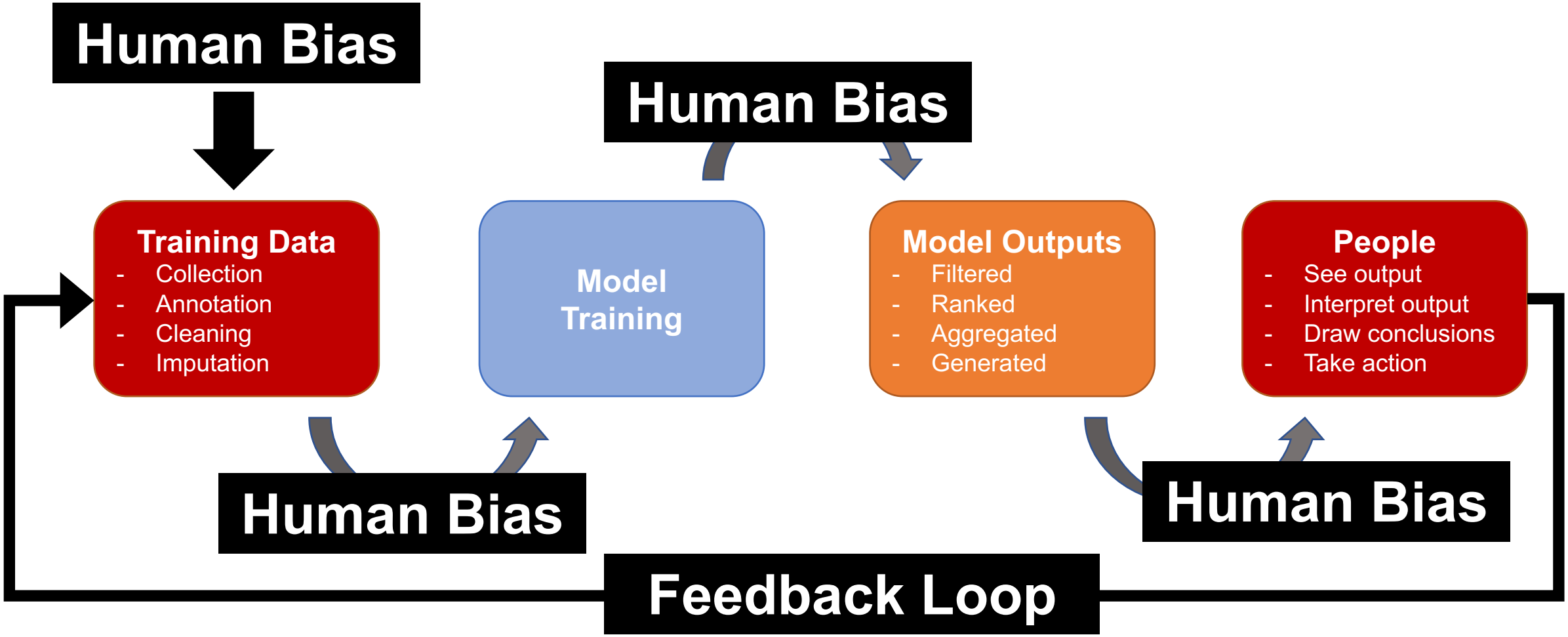
© mollysdad - Slideshare - Introduction to Logical Fallacies

Biases in Interpretation

Automation Bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

thedailyenglishshow.com | CC BY 2.0

Human Bias heavily impacts Machine Learning and AI



- **Bias Network Effect**
- **Bias "Laundering"**

Key Take-Aways

Human data ***perpetuates*** human biases

As Algorithms learn from human data,
the result is a ***bias network effect***

Bias Laundering happens when we ignore the biases of automated decision systems because of the illusion that all algorithms must be objective since they are *driven by math* or *run by a computer*

Bias comes in different “flavors”

Bias in Statistics and ML

- Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
- The "bias" term b (e.g., y = mx + b)

Cognitive biases

- Confirmation bias
- Recency bias
- Optimism bias

Algorithmic Bias

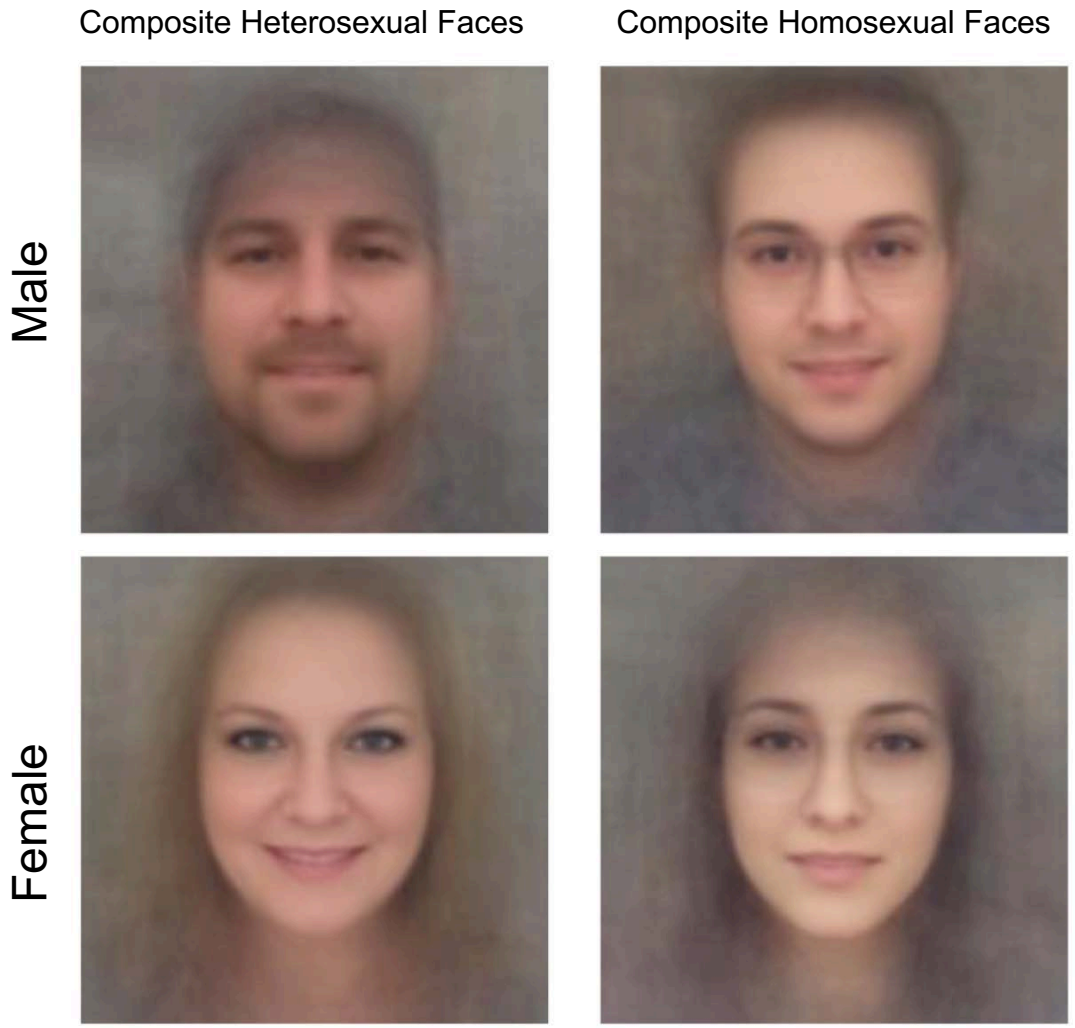
Unjust, unfair, or prejudicial treatment of people
related to race, income, sexual orientation, religion, gender,
and other characteristics historically associated with discrimination and marginalization,
when and where they manifest in algorithmic systems or algorithmically aided decision-making

"Although neural networks might be said to write their own programs, they do so towards ***goals set by humans, using data collected for human purposes****. If the data is skewed, even by accident, the computers will* ***amplify injustice****."*

— The Guardian

https://www.theguardian.com/commentisfree/2016/oct/23/the-guardian-view-on-machine-learning-people-must-decide

Predicting Homosexuality



https://psyarxiv.com/hv28a/

Wang and Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, 2017.

"**Sexual orientation detector**" using 35,326 images from public profiles on a US dating website.

quote from paper:

"*Consistent with the prenatal hormone theory [PHT] of sexual orientation, gay men and women tended to have gender-atypical facial morphology*."

???
!!!

Wang, Y. and Kosinski, M., 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2), p.246.

Predicting Homosexuality

Differences between homosexual and and heterosexual faces in selfies relate to grooming, presentation, and lifestyle — that is, **differences in culture, not in facial structure**

Selection Bias
+
Experimenter's Bias
+
Correlation Fallacy

Check out this more elaborate discussion on **Medium**
"*Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?*"

https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Google Vision Cloud: Image Classification

Faces | Objects | **Labels** | Logos | Web | Properties | Safe Search

Screenshot from 2020-03-31 11-27-22.png

Technology	68%
Electronic Device	66%
Photography	62%
Mobile Phone	54%

https://twitter.com/nicolaskb/status/1244921742486917120

Google Vision Cloud: Image Classification

Try the API

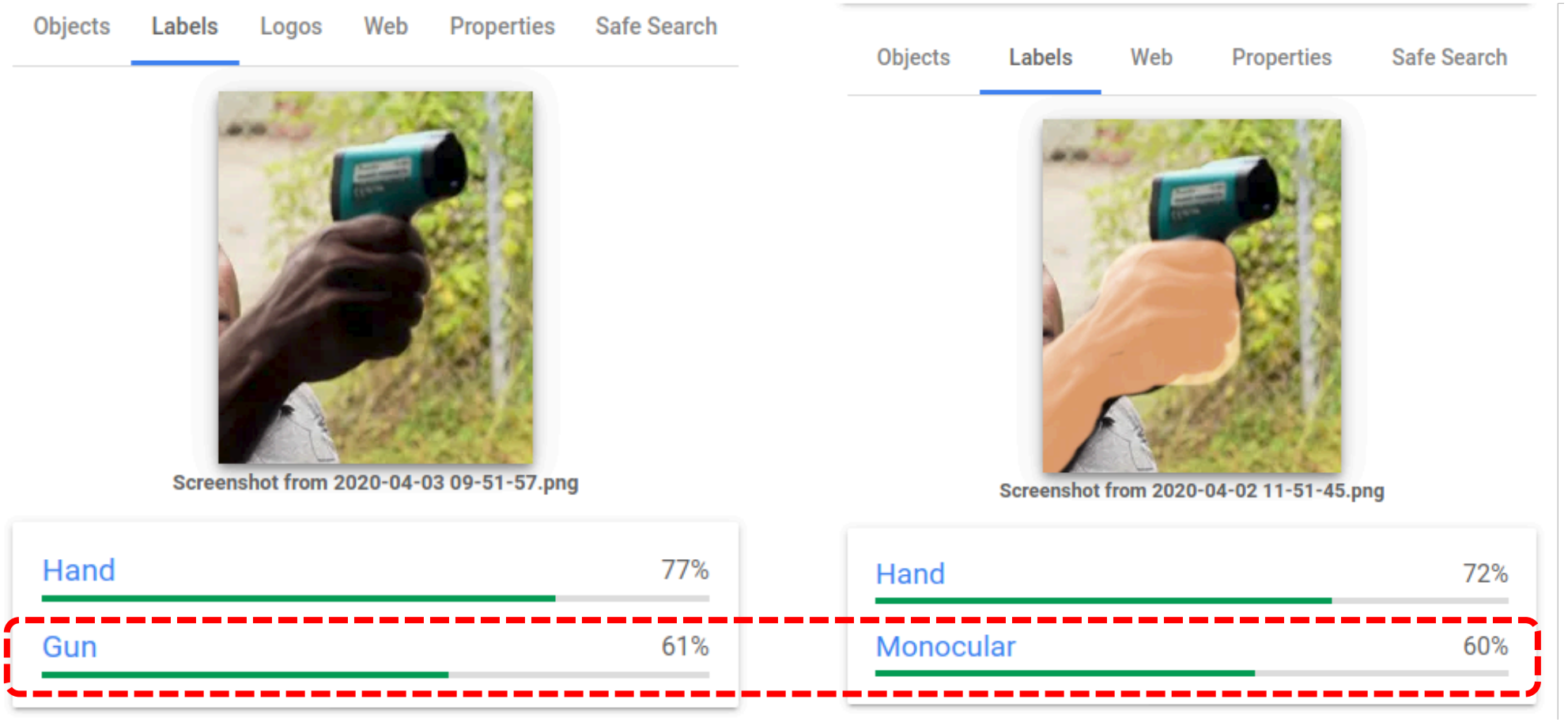
Faces | Objects | **Labels** | Web | Properties | Safe Search

Screenshot from 2020-03-31 11-23-45.png

Gun	88%
Photography	68%
Firearm	65%
Plant	59%

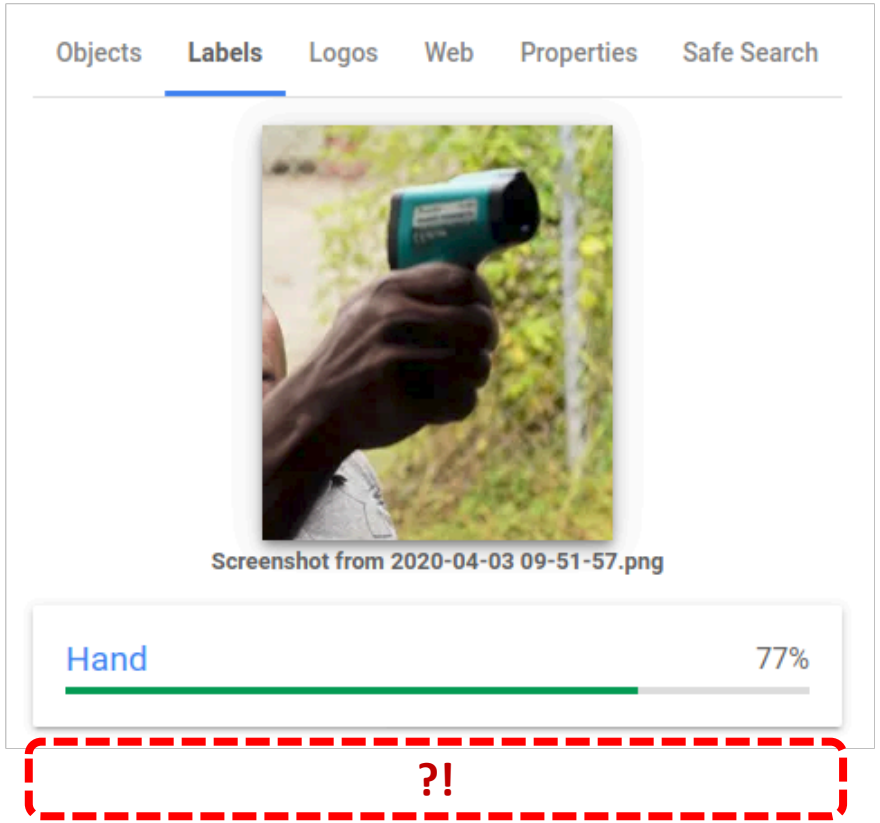
https://twitter.com/nicolaskb/status/1244921742486917120

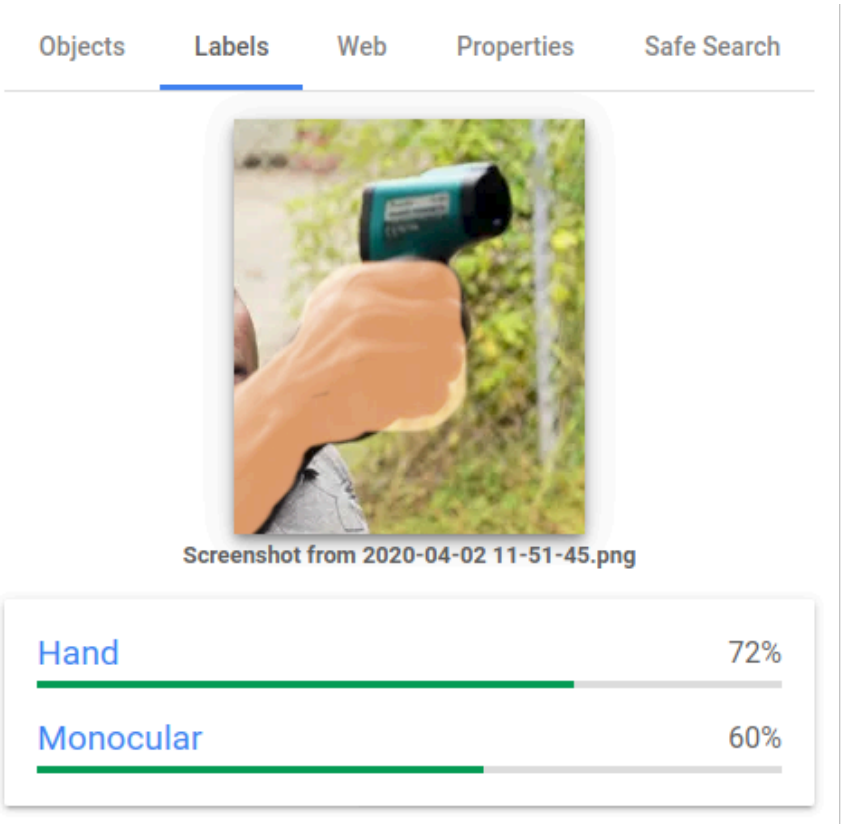
Google Vision Cloud: Isolating the Problem



https://algorithmwatch.org/en/google-vision-racism/

Google Vision Cloud: Correcting the Problem





https://algorithmwatch.org/en/google-vision-racism/

Discussion Point: Does this SOLVE the problem?

Joy Buolamwini – Founder of the Algorithmic Justice League

https://time.com/5520558/artificial-intelligence-racial-gender-bias/ https://www.youtube.com/watch?v=QxuyfWoVV98 https://www.youtube.com/watch?v=eRUEVYndh9c

A ***seemingly*** straight-forward Machine Learning task
such as Binary Classification
can be ***Treacherous***

Identifying Algorithmic Bias: A Starting Point for Everyone

Testing for Fairness & Inclusion in Binary Classification

Disaggregated Evaluation

- Create for each (subgroup, prediction) pair
- Compare across subgroups

Example:

- women, job eligibility
- men, job eligibility

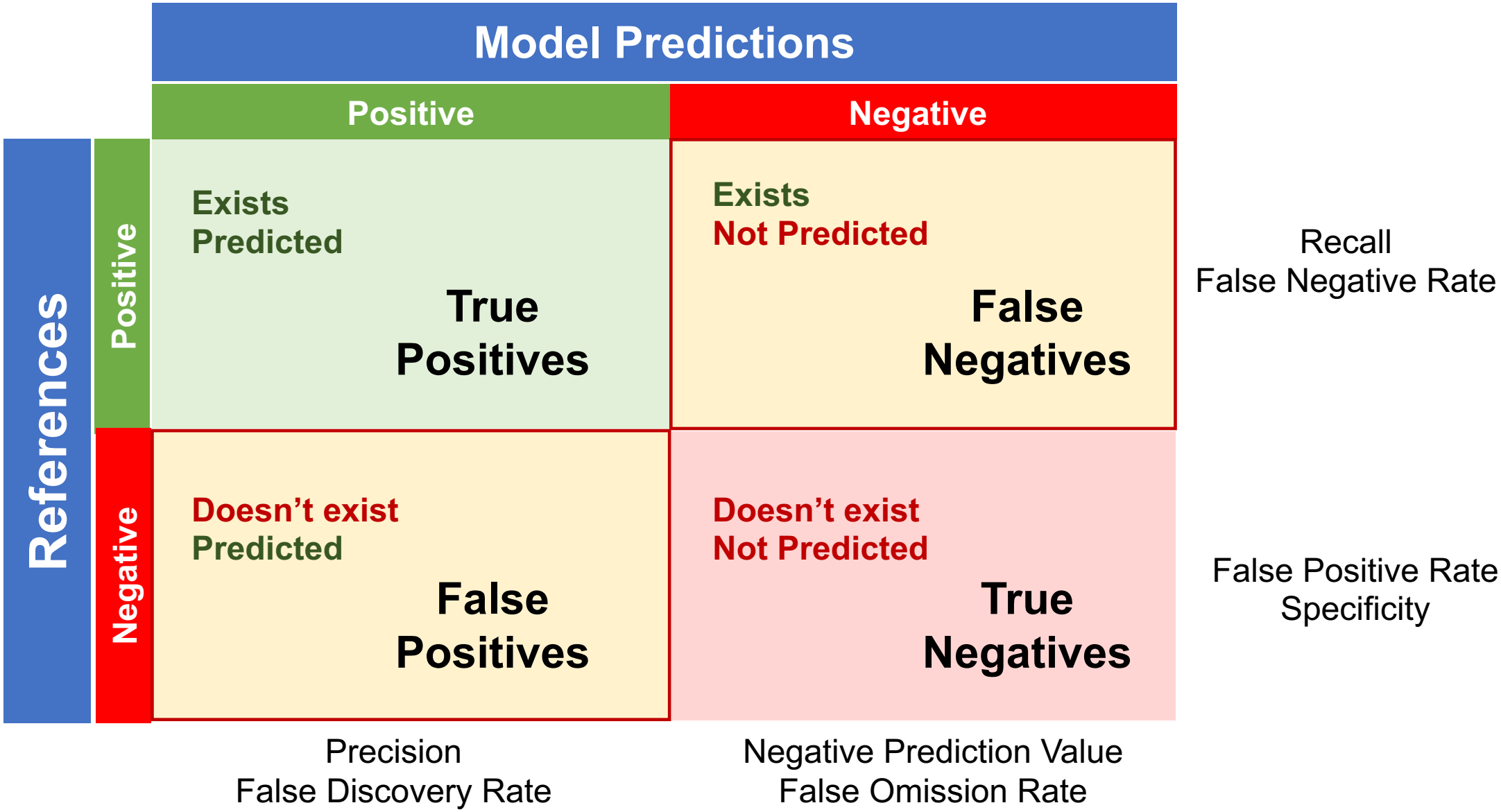
Intersectional Evaluation

- Create for each (subgroup1, subgroup2, prediction) pair
- Compare across subgroups

Example:

- Native American women, job eligibility
- Asian men, job eligibility

The Confusion Matrix



Evaluating Fairness & Inclusion

*Numerical Example**

Male Job Eligibility

True Positives (TP) =	72	**False Positives (FP) =**	2
False Negatives (FN) =	2	**True Negatives (TN) =**	404

Precision = TP / (TP + FP) = **0.973**

Recall = TP / (TP + FN) = **0.973**

Female Job Eligibility

True Positives (TP) =	8	**False Positives (FP) =**	8
False Negatives (FN) =	2	**True Negatives (TN) =**	222

Precision = TP / (TP + FP) = **0.500**

Recall = TP / (TP + FN) = **0.800**

Precision is intuitively the ability of the classifier not to label as positive a sample that is negative.
Recall is intuitively the ability of the classifier to find all the positive samples.

* this numerical example is available as an excel on the course website on CANVAS

Equality of Opportunity

*Numerical Example**

Male Job Eligibility

True Positives (TP) =	72	**False Positives (FP) =**	2
False Negatives (FN) =	2	**True Negatives (TN) =**	404

Precision = TP / (TP + FP) = **0.973**

Recall = TP / (TP + FN) = **0.973**

Female Job Eligibility

True Positives (TP) =	8	**False Positives (FP) =**	8
False Negatives (FN) =	2	**True Negatives (TN) =**	222

Precision = TP / (TP + FP) = **0.500**

Recall = TP / (TP + FN) = **0.800**

Equality of Opportunity fairness criterion: *Recall* is equal across subgroups

We should hire equal proportion of individuals from the qualified fraction of each group

Precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

Recall is intuitively the ability of the classifier to find all the positive samples.

* this numerical example is available as an excel on the course website on CANVAS

Predictive Parity

*Numerical Example**

Male Job Eligibility			
True Positives (TP) =	72	**False Positives (FP) =**	2
False Negatives (FN) =	2	**True Negatives (TN) =**	404

Female Job Eligibility			
True Positives (TP) =	8	**False Positives (FP) =**	8
False Negatives (FN) =	2	**True Negatives (TN) =**	222

Precision = TP / (TP + FP) =	**0.973**	**Precision** = TP / (TP + FP) =	**0.500**
Recall = TP / (TP + FN) =	**0.973**	**Recall** = TP / (TP + FN) =	**0.800**

Predictive Parity fairness criterion: *Precision* is equal across subgroups

A prediction for a candidate should reflect the candidate's real capability of doing the job

Precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

Recall is intuitively the ability of the classifier to find all the positive samples.

* this numerical example is available as an excel on the course website on CANVAS

A Lot depends on the Application at Hand: Privacy in Images

False Positives Might be Better than False Negatives

False Positive: Something that *doesn't* need to be blurred gets blurred.

Bummer…

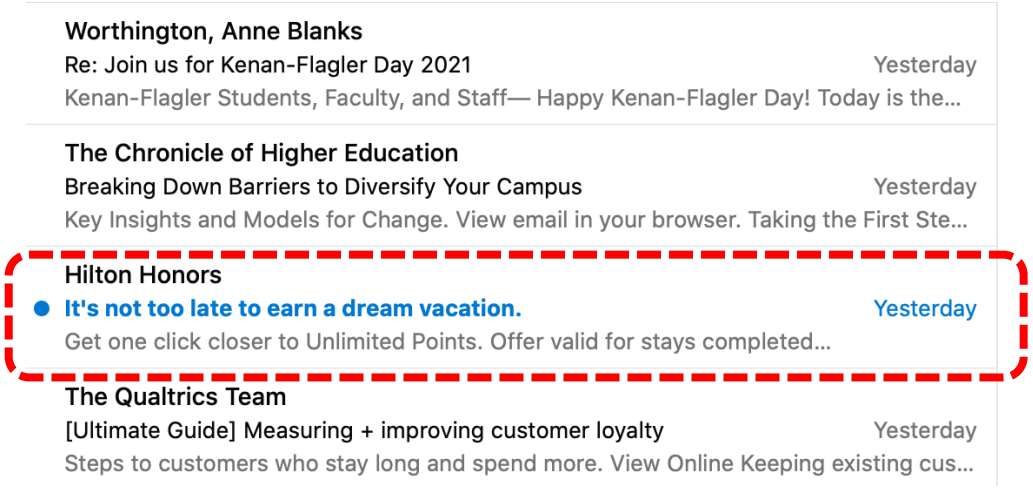
False Negative: Something that *needs* to be blurred is not blurred.

Fraud and identity theft!

A Lot depends on the Application at Hand: Spam Detection

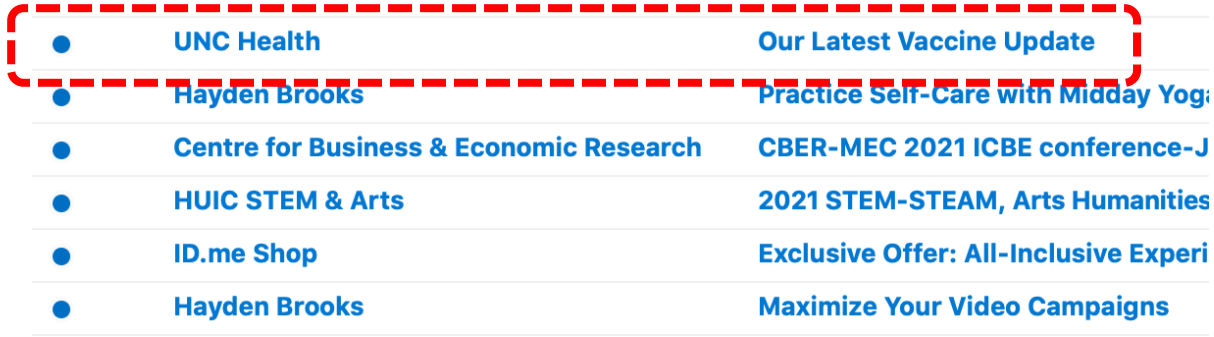
False Negatives Might be Better than False Positives

False Negative: Email that is SPAM is not caught → appears in someone's inbox



can be annoying …

False Positive: Email identified as SPAM → is removed from someone's inbox



When it is an important message, it's a loss!

Data: The central Ingredient

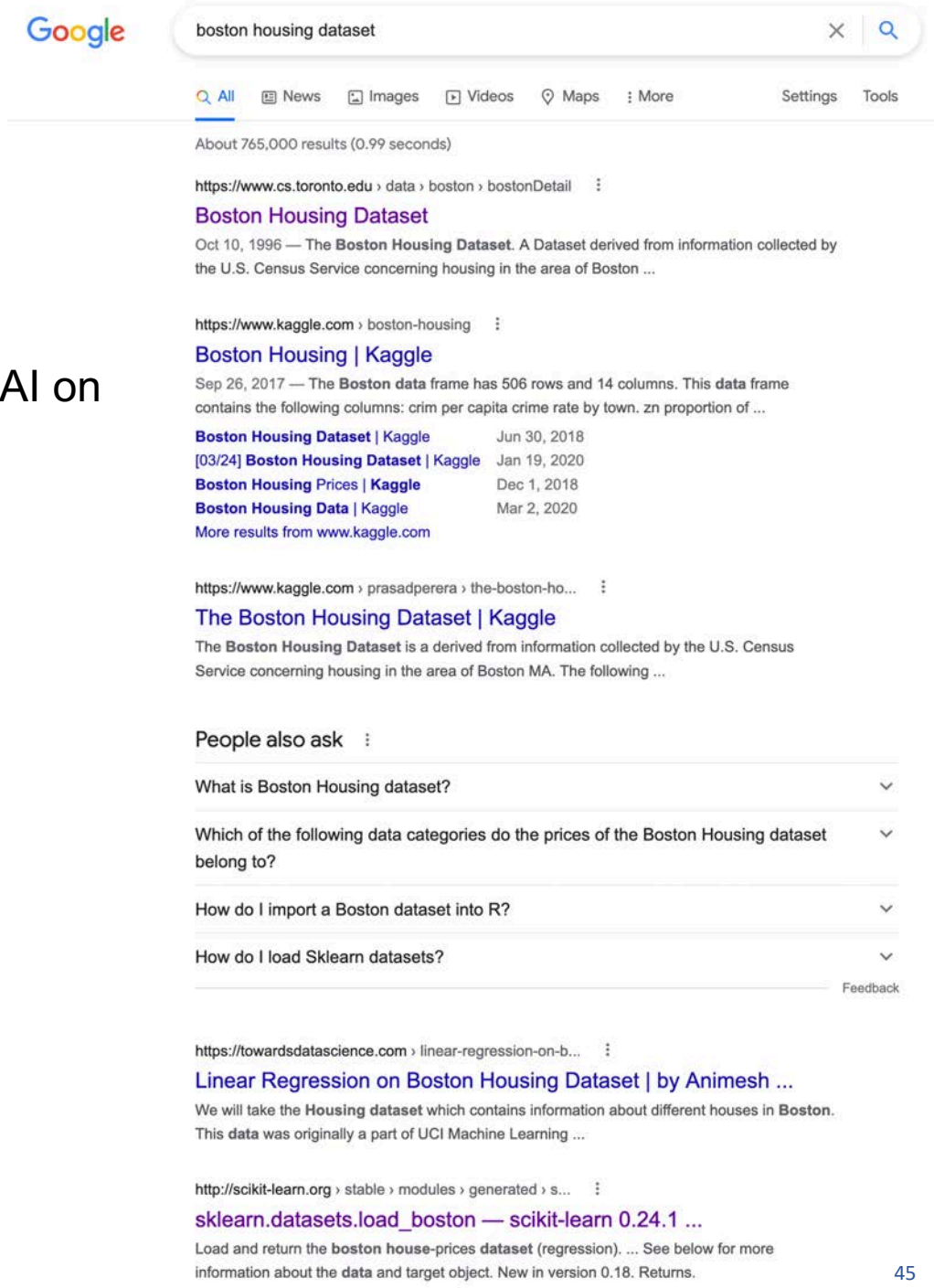
Using data blindly can lead to adverse outcomes

- There are many historic datasets that we can apply ML to, or train AI on
- These data can contain hidden biases
- Before you use data (e.g., specific features), ask yourself:

    1. What do individual **features really capture**?
    2. **Who collected** these data?
    3. **How** were these data **collected**?
    4. What was the **purpose or intent** for collecting these data?
    5. Were these data previously **preprocessed** or **transformed**?
    6. What are the **implications** for my specific applications?

Are seemingly innocent data reall so innocent?

Ex: Boston Housing Dataset



The Boston Housing Dataset

Python code to conveniently load it

```
from sklearn.datasets import load_boston
boston = load_boston()
print(boston.DESCR)
```

```
**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```

Language!

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980. N.B. Various transformations are used in the table on pages 244-261 of the latter.

The Boston Housing Dataset, revisited

In an effort to replace the term "blacks" with the more appropriate term "black people", I raised it with the scikit-learn community: Issue 19657

I am glad that the community, after short debate, responded swiftly: Pull Request 19661

```
!pip uninstall scikit-learn -y
!pip install git+git://github.com/scikit-learn/scikit-learn.git
from sklearn.datasets import load_boston
boston = load_boston()
print(boston.DESCR)
```

```
  **Data Set Characteristics:**

      :Number of Instances: 506

      :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

      :Attribute Information (in order):
          - CRIM     per capita crime rate by town
          - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
          - INDUS    proportion of non-retail business acres per town
          - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
          - NOX      nitric oxides concentration (parts per 10 million)
          - RM       average number of rooms per dwelling
          - AGE      proportion of owner-occupied units built prior to 1940
          - DIS      weighted distances to five Boston employment centres
          - RAD      index of accessibility to radial highways
          - TAX      full-value property-tax rate per $10,000
          - PTRATIO  pupil-teacher ratio by town
          - B        1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
          - LSTAT    % lower status of the population
          - MEDV     Median value of owner-occupied homes in $1000's
```

The Boston Housing Dataset: Problem Solved?

The change in language does not resolve the fact that variable B is problematic

```
- B        1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
```

- From a data transformation perspective, for example, variable **B** is non-invertible
- Its parabolic nature does not allow us to reconstruct the original variable ***BK***, which is excluded from the dataset
- In other words, **BK** could be one of two different values (i.e., 2 X's for 1 Y), and we cannot reconstruct which one
- We are thus curtailed in our ability to investigate **BK** more deeply
- Furthermore, the value of ***0.63*** is simply set without further explanation as to why at this level
- There are additional aspects to consider regarding what variable **B** is meant to capture in the original 1978 publication (i.e., impact of self-segregation on house prices) that, from my perspective, could be related to systemic racism and the practice of redlining

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980. N.B. Various transformations are used in the table on pages 244-261 of the latter.

Machine Learning and AI can lead to Adverse Outcomes

The Human Factor

- Lack of insight into ***sources of bias*** in the data and model
- Lack of insight into the ***feedback loops***
- Lack of careful, ***disaggregated evaluation***
- Human ***biases in interpreting and accepting results***

→ **Intentional *or* Unintentional: Both can lead to the same adverse outcomes!**

Data Really Matter Really!

- Question ***where*** your data come from, ***who*** collected them, and for ***what purpose***
- ***Understand*** your Data: skews, correlations
- ***Abandon single training-set*** / testing-set from similar distribution
- Combine inputs from ***multiple sources***
- Use ***held-out test*** sets for hard use cases
- Talk to experts about ***additional signals***
- Be ***skeptical*** about data ***transformations*** (especially non-invertible ones)
- ***Document*** transformation, imputation and deletion of data

→ **Never just assume that your data are fine as they are!**

A new Era: Challenges with Large Language Models

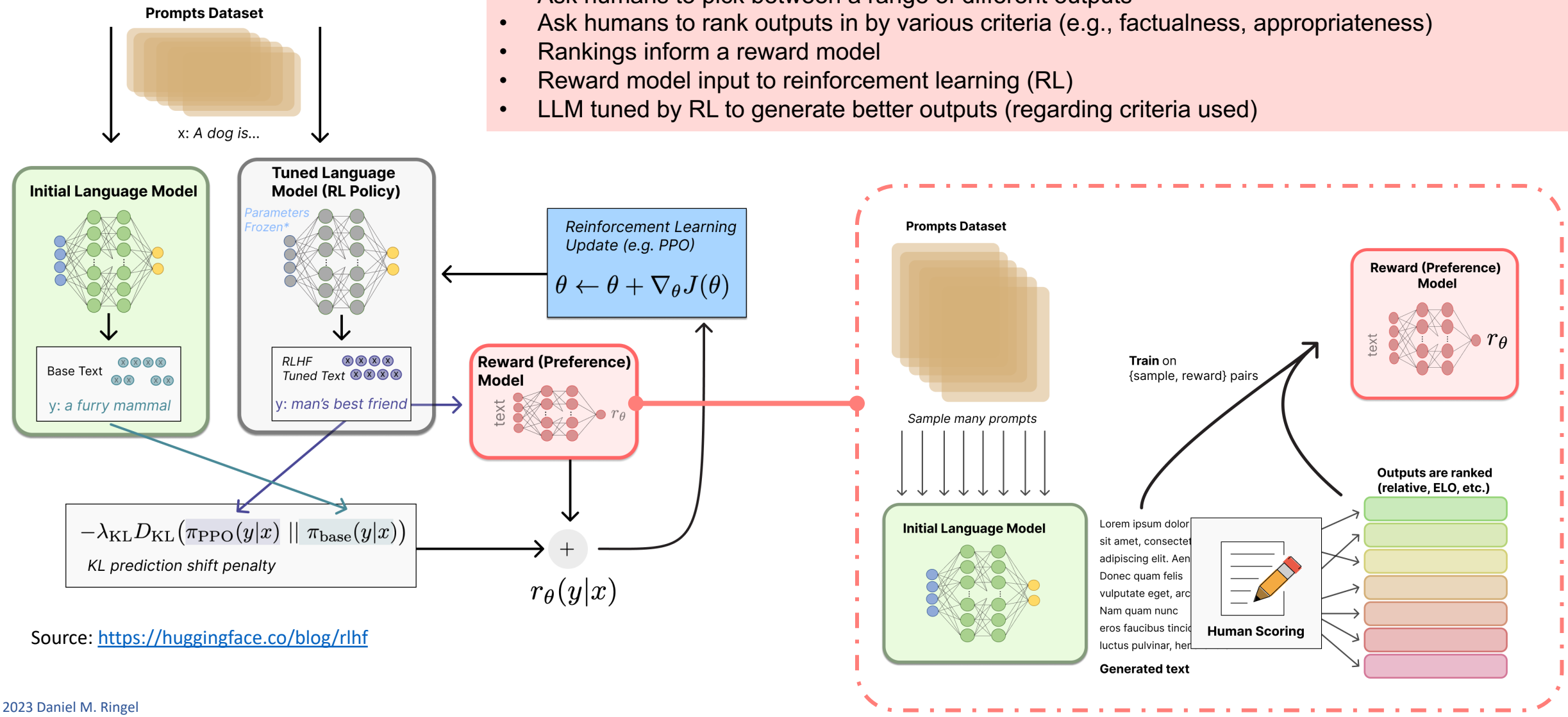
- Language models are **notorious bullshitters**
  - Excellent at predictions
  - Cannot reflect on the implication of their predictions
- **Hallucinating** is a polite term for making stuff up
- **Toxic training data exacerbate the problem**
  - Is social media (e.g., tweets, reddits) a good source to explain the world and give advice?
- **False sense of security from cited sources?**
  - Source make AI text look authoritative
  - People are less likely to question it

Mitigation strategies: Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF)

RLHF is a technique to improve LLM's answers based on feedback from humans

- Prompt LLM
- Ask humans to pick between a range of different outputs
- Ask humans to rank outputs in by various criteria (e.g., factualness, appropriateness)
- Rankings inform a reward model
- Reward model input to reinforcement learning (RL)
- LLM tuned by RL to generate better outputs (regarding criteria used)



Source: https://huggingface.co/blog/rlhf

Implications of Large Language Models

Firms
- Streamline operations
- Competitive advantage
- Reputation risks

Consumers
- Better customer experiences (CX)
- Empowerment at the cost of technology dependence
- Misleading information and false authority

Society
- Inequity and bias
- Balance between data-driven technology and human creativity
- Copyrights, patents, privacy

Other OpenAI models:
WebGPT can go and look up information on the web and give sources for its answers.
InstructGPT, a version of GPT-3 that OpenAI trained to produce text that was less toxic.

Call for Nominations: Recognize a Professor for their Teaching

Don't forget to put in YOUR nominations before Monday, March 27th, 2023!

The mediocre teacher tells.
The good teacher explains.
The superior teacher demonstrates.
The great teacher inspires.

William A. Ward

I cannot teach anybody anything;
I can only make them think.

Socrates

Tell me and I forget.
Teach me and I remember.
Involve me and I learn.

Benjamin Franklin

Your Vote Counts! Please Nominate here: https://tinyurl.com/weatherspoon2023

Looking Ahead

Next Class: Tuesday, March 28, 2023

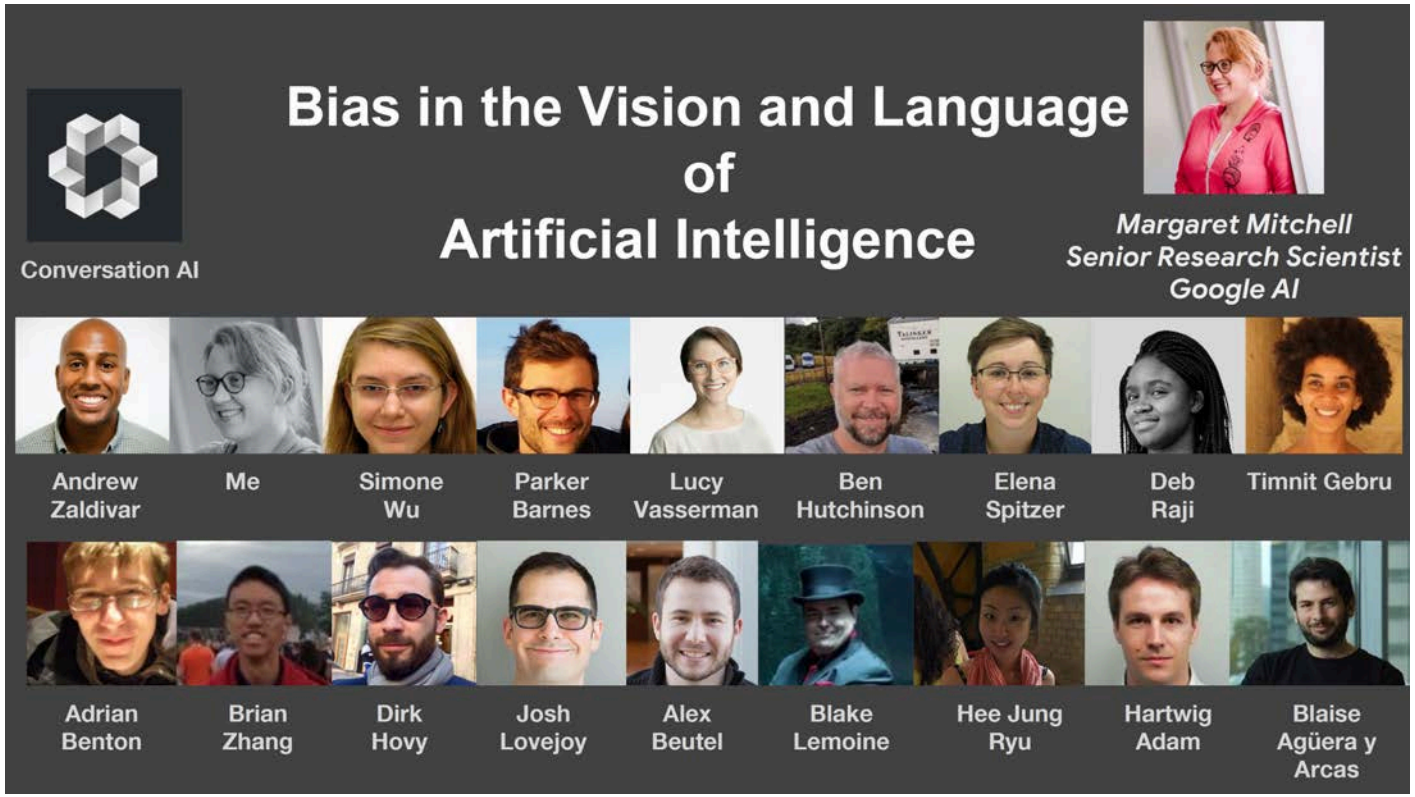
Team Assignment 3 due before 12 noon
Investing into Peer2Peer Lending
- Submit Presentation + Python Notebook but no data (if you use additional data, then provide a link to it).

Special Topics Tutorial: *P2P Loans*
- 03/22 (tonight!) from 5-6pm
- with Isaac (will be recorded)
- Zoom Link or ID: 985 4146 3939 Passcode: 844935

Quiz 3: *Predictive Modeling*
- Opens 9pm on 3/28
- Closes 4/2 at midnight
- Practice Quiz available by 3/28

This lecture is designed to stimulate discussion and builds heavily of the excellent work and materials of Margaret Mitchell, Senior Research Scientist at Google AI, and her team.

You can download their full presentation from Stanford's website



https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

Additional Reading on algorithmic bias in marketing:

Lambrecht, A. and Tucker, C., 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 65(7), pp.2966-2981.