

Toward more responsible and ethical data scientists

Bernease Herman
CSE 583: Software Engineering for Data Scientists
University of Washington
November 18, 2019

Agenda

Sources of bias in algorithms

Machine learning pipeline

ML-specific sources of bias

Other efforts toward ethical data sci

Questions

Concern / Source of Bias

Purposeful harm or manipulation

At times we do see algorithms that are
purposely intended to be unfair and biased.

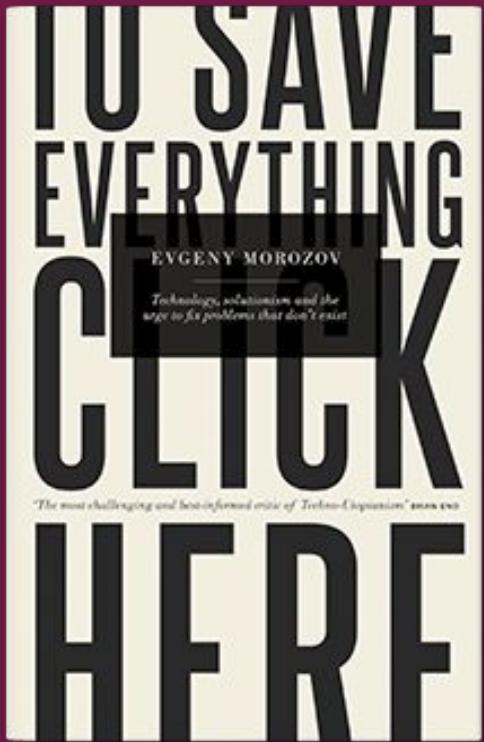
These are an obvious concern.

DRUGS!

**'The Code
I'm Still
Ashamed
Of'**

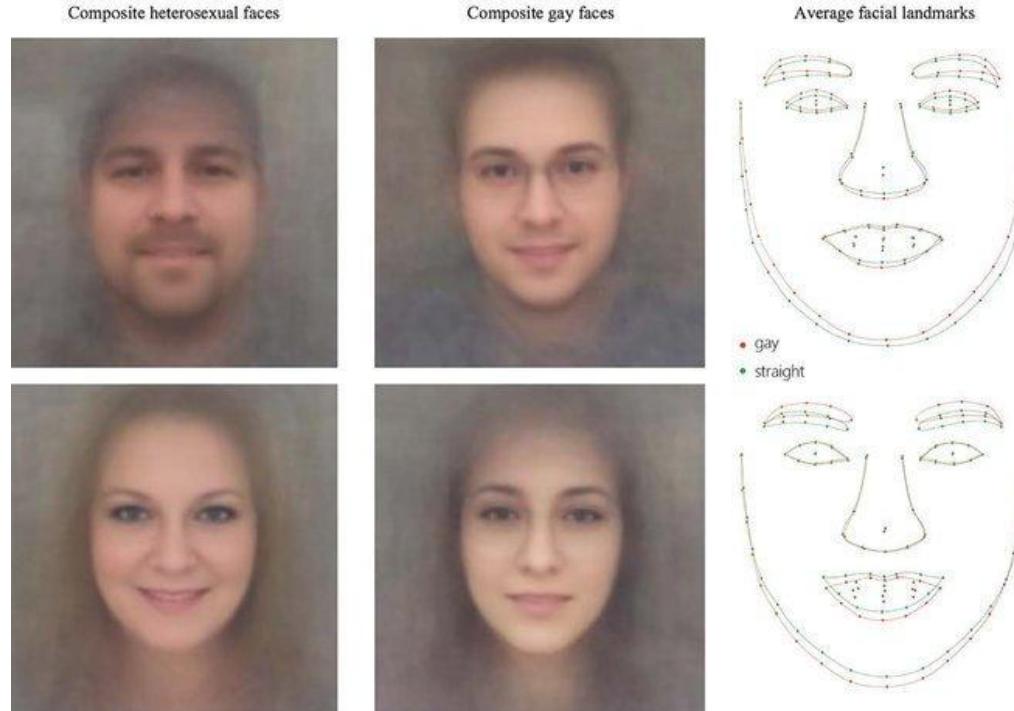


Concern / Source of Bias



Technological solutionism
Making tech products
because we're *able*
regardless of whether or
not we *should*.

Algorithmic “Gaydar” Research in 2017



Source: Wang and Kosinski (2017)

Concern / Source of Bias

Ubiquity, consolidation, and trust

While human decision makers can be arguably more biased, there are often many (somewhat independent) humans with different biases. Not the case with tech.



Concern / Source of Bias

Limited perspective of algorithm writer

Algorithms meant to handle a large, diverse range of data are subject to rules from the algorithm writer's perspective.

Concern / Source of Bias

Differences or errors in the algorithm implementation

Errors can be embedded into the system unknowingly.

Concern / Source of Bias

Measurement error, or mismeasured inputs

When gathering input data, we may have recorded an incorrect value.

Human-specified algorithms

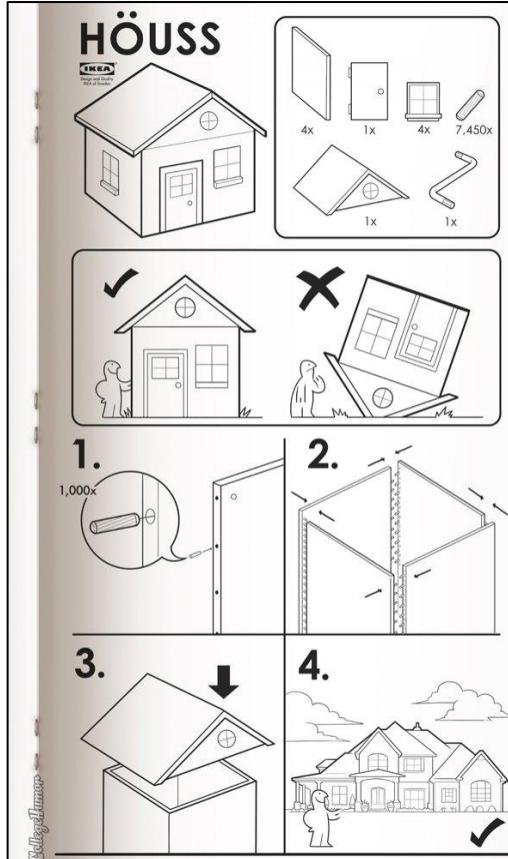
What's meant by human-specified?

Cela fut le 4^e Oct.

101

By dinner time,

I knew we were going to get this far, so I was prepared, and I knew it would be a job for me. But I am with and happy, ~~if I don't forget you, an sister.~~ The great part now is I got their opinion delivered, so that leaves my fifth & further engagement. Long ride to go to Mr. Secretary, and oh then there's another, like to come to him in another sort of form, and with another kind of finish to a hand. It is an obvious his purpose to fully implement its fiscal and separate surrounding their purpose, a state of circumstances that will have produced the terrible economical and financial trouble in the Europe and around India. But I burst in open this, and his measures set as oppose with all the team of consequences not has contact of the long duration. So with love, ~~by you~~



```
string sInput;
int iLength, iN;
double dblTemp;
bool again = true;

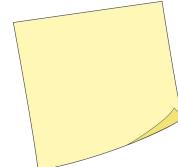
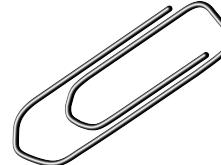
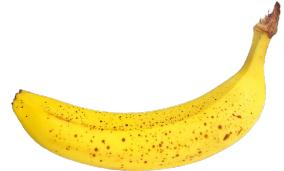
while (again) {
    iN = -1;
    again = false;
    getline(cin, sInput);
    system("cls");
    stringstream(sInput) >> dblTemp;
    iLength = sInput.length();
    if (iLength < 4) {
        again = true;
        continue;
    } else if (sInput[iLength - 3] != '.') {
        again = true;
        continue;
    } while (++iN < iLength) {
        if (isdigit(sInput[iN])) {
            continue;
        } else if (iN == (iLength - 3)) {
```

More difficult: Object classification

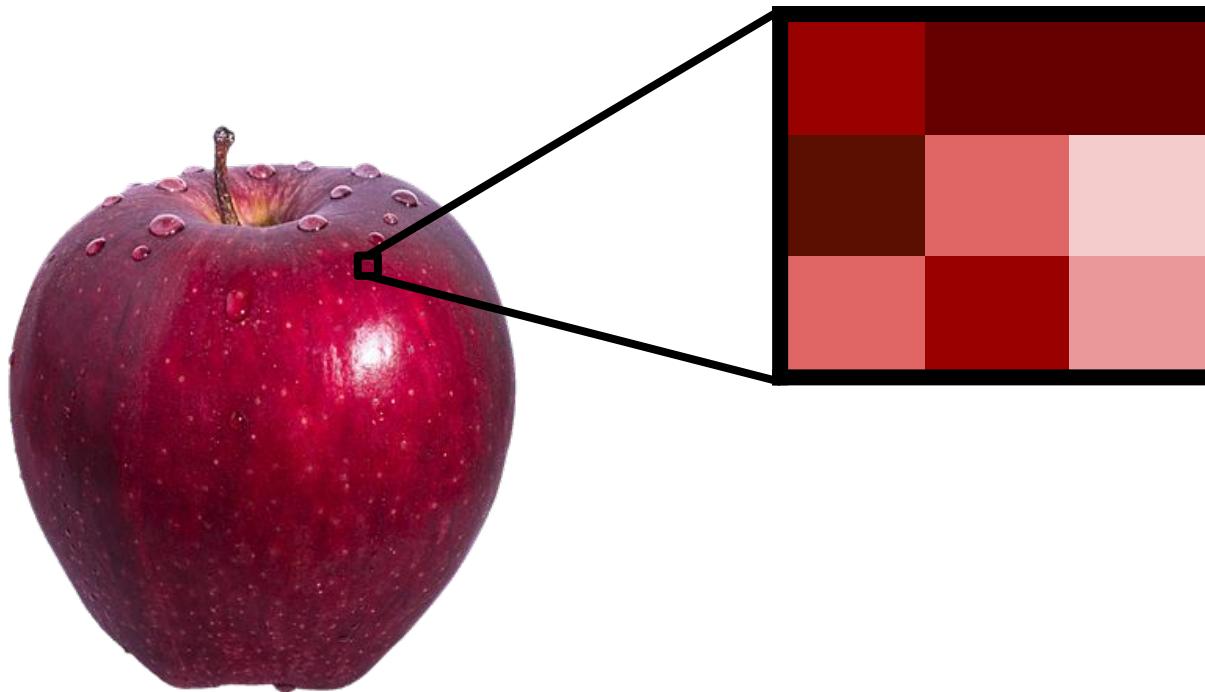
Object classification algorithm

Input: image

Output: predicted object class



Representation of digital images



Writing rules for human-specified algo

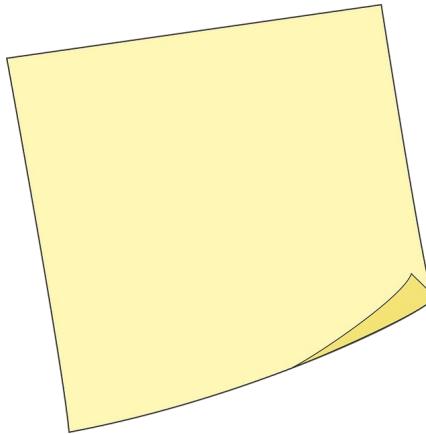
Object classification algorithm

Input: image

Output: predicted object class

```
if img.center_pixel == 'red':  
    object = 'apple'  
elif img.center_pixel == 'yellow':  
    object = 'banana'  
...
```

We'll need more features than that



ACTIVITY: Generating features for object classification algorithm

In groups of 2-4 people sitting nearby, think of 1-2 new features for our image algo.

Original input: image (pixels)

Features:

1. color of the center pixel
2. % of pixels are background (white)
- 3.



EXAMPLES: Generating features for object classification algorithm

Original input: image (pixels)

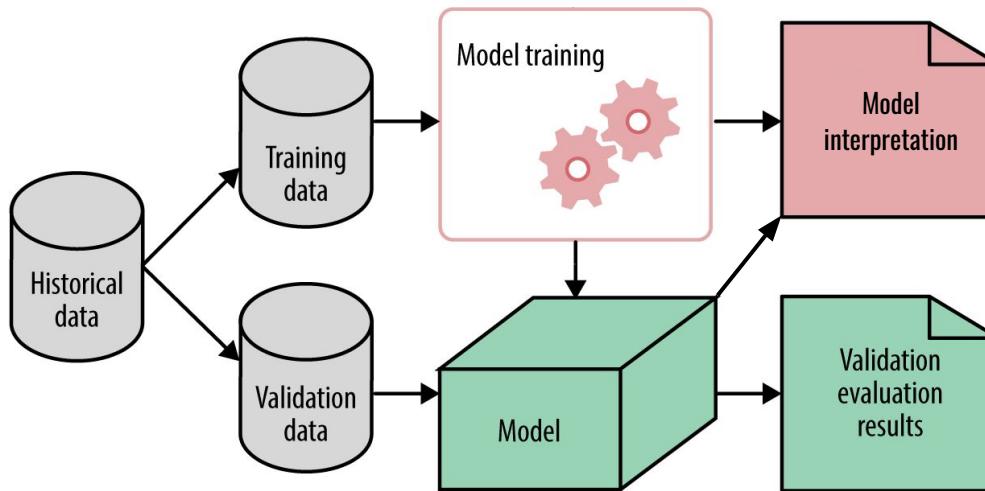
Features:

1. color of the center pixel
2. % of pixels are background (white)
3. how rectangular / narrow is shape
4. # of distinct colors
5. background holes



Machine learned algorithms

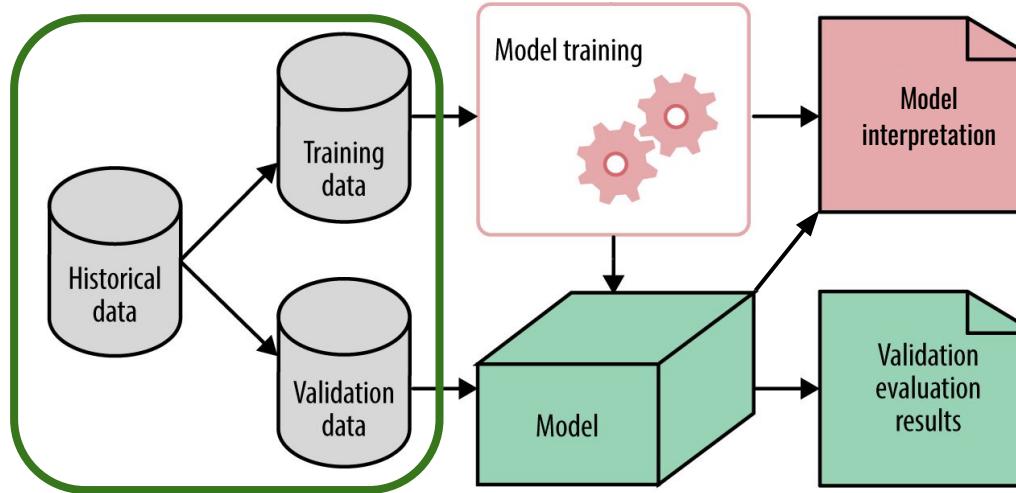
Machine learning pipeline



Modified graphic. Original from Alice Zheng

How do we choose and collect dataset?

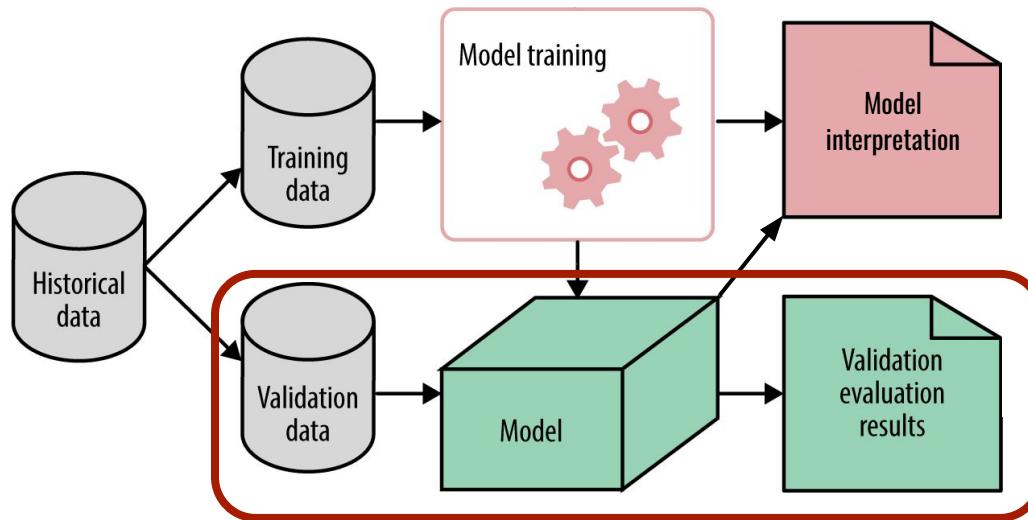
Data Curation • Evaluation • Interpretation



Modified graphic. Original from Alice Zheng

How to check performance of ML?

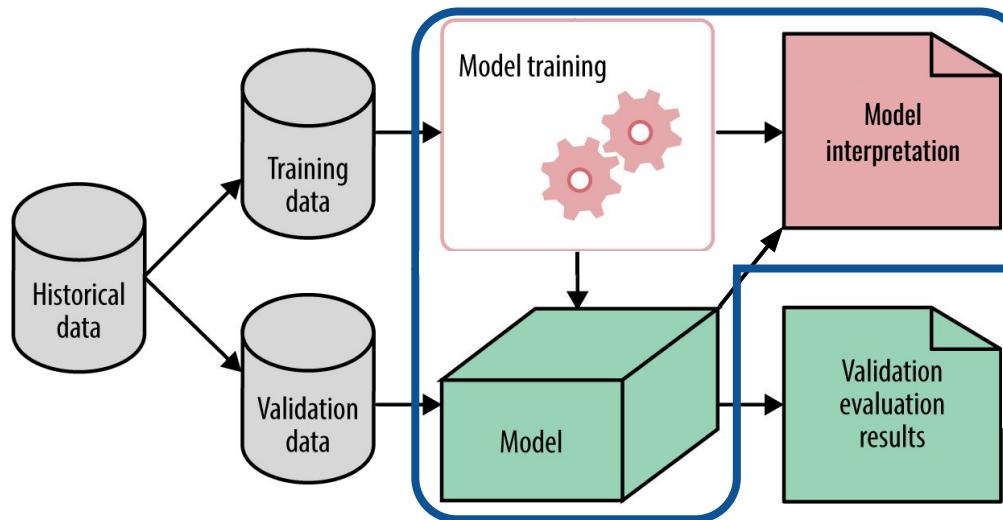
Data Curation • **Evaluation** • Interpretation



Modified graphic. Original from Alice Zheng

What's happening on the inside?

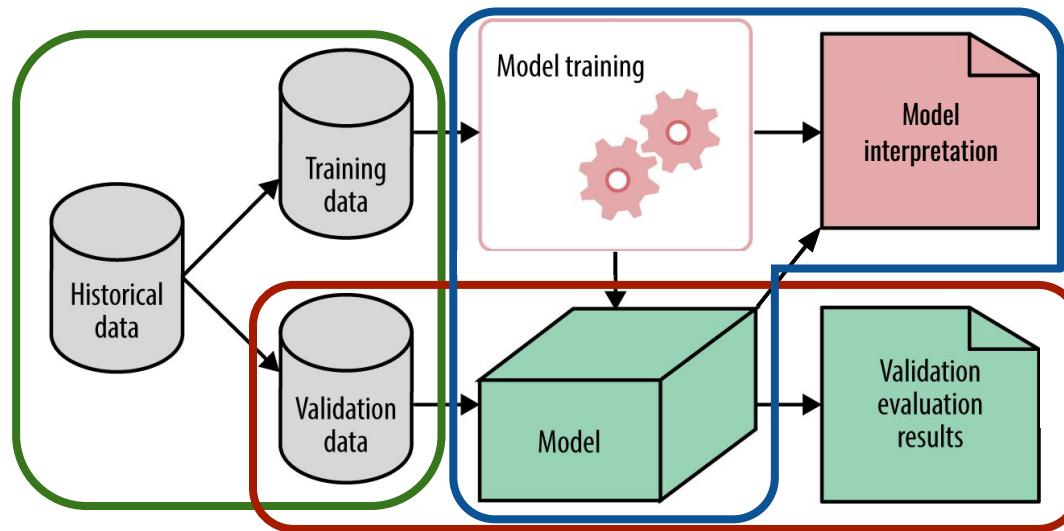
Data Curation · Evaluation · Interpretation



Modified graphic. Original from Alice Zheng

My broad research interests in ML

Data Curation · Evaluation · Interpretation



Modified graphic. Original from Alice Zheng

Concern / Source of Bias

Labeling error

When gathering labels, we may have recorded an incorrect value.

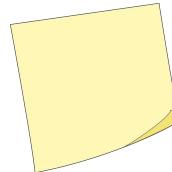
Source of bias: Labeling error



scissors



apple



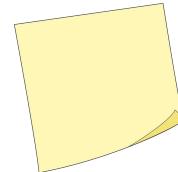
window



shopping bag



Source of bias: Labeling error



scissors

scissors

scissors

apple

poison

apple

window
stickies

paper

shopping bag

lock

lock



Source of bias: Self-reported labeling error

Other Info

★ Professional Affiliation

★ Age

★ Gender ⚠

Female (all that apply)

Male Asian-American/Black

FTM American Indian/Native American

MTF

Genderqueer

Other

Asian or Asian-American

European or Euro-American/White

Native Hawaiian or Pacific Islander

Hispanic or Latina/o

Other

Concern / Source of Bias

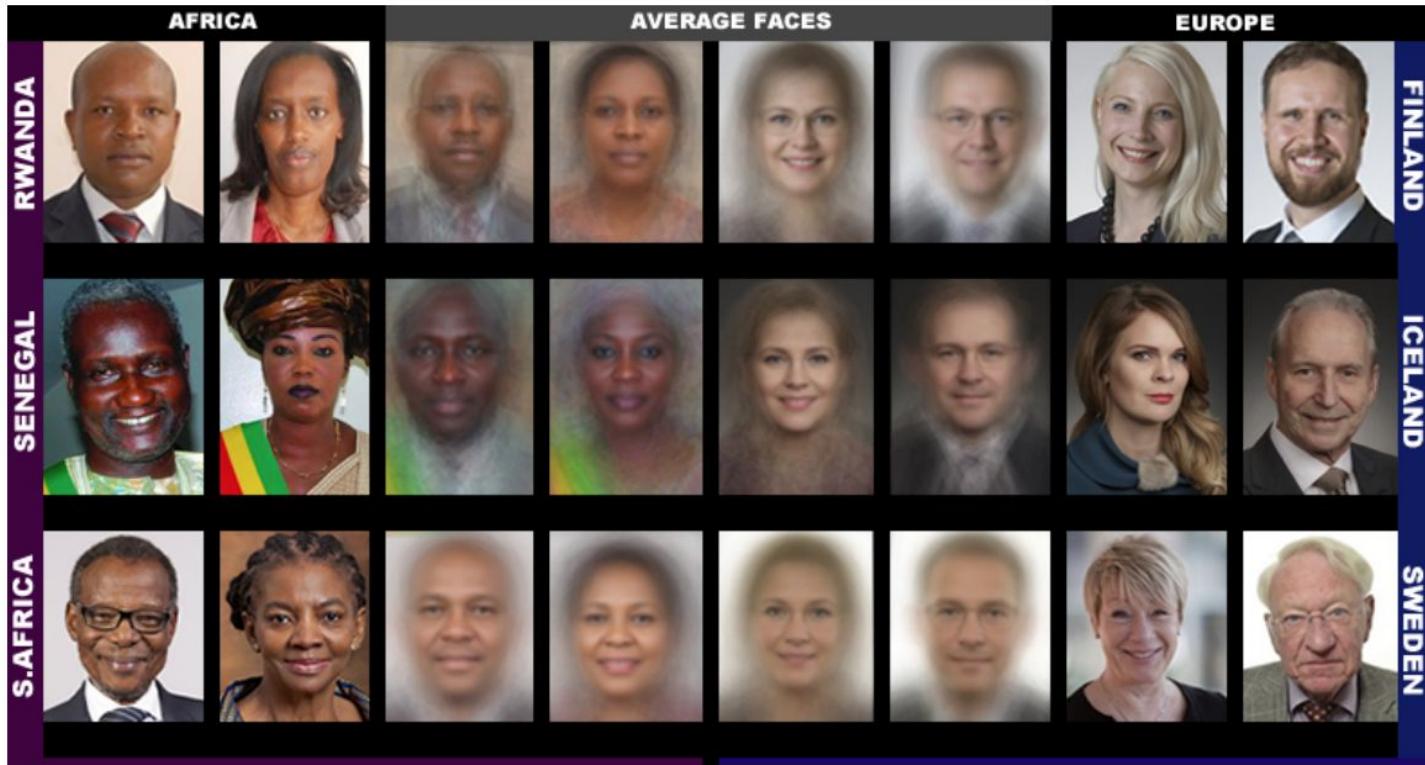
Unrepresentative dataset

The contents (or their proportion) of the dataset is not representative of the real world.

Source of bias: Unrepresentative dataset



Gender Shades project



Dataset is not balanced across gender or skin tone

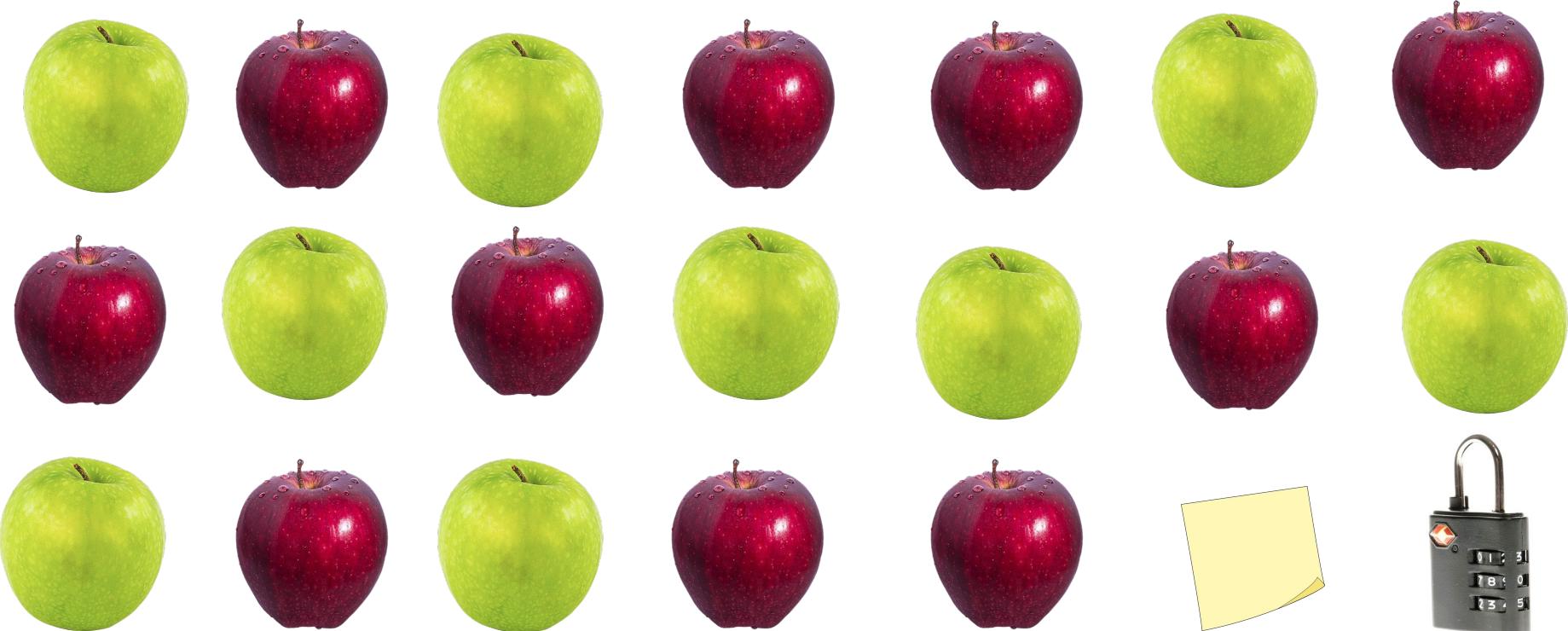


Concern / Source of Bias

Insufficient evaluation metrics

The metrics we use to evaluate the fit of our model do not capture all relevant aspects of the task.

Source of bias: Very unrepresentative dataset & evaluation metrics



Writing rules like human-specified algo

Object classification algorithm

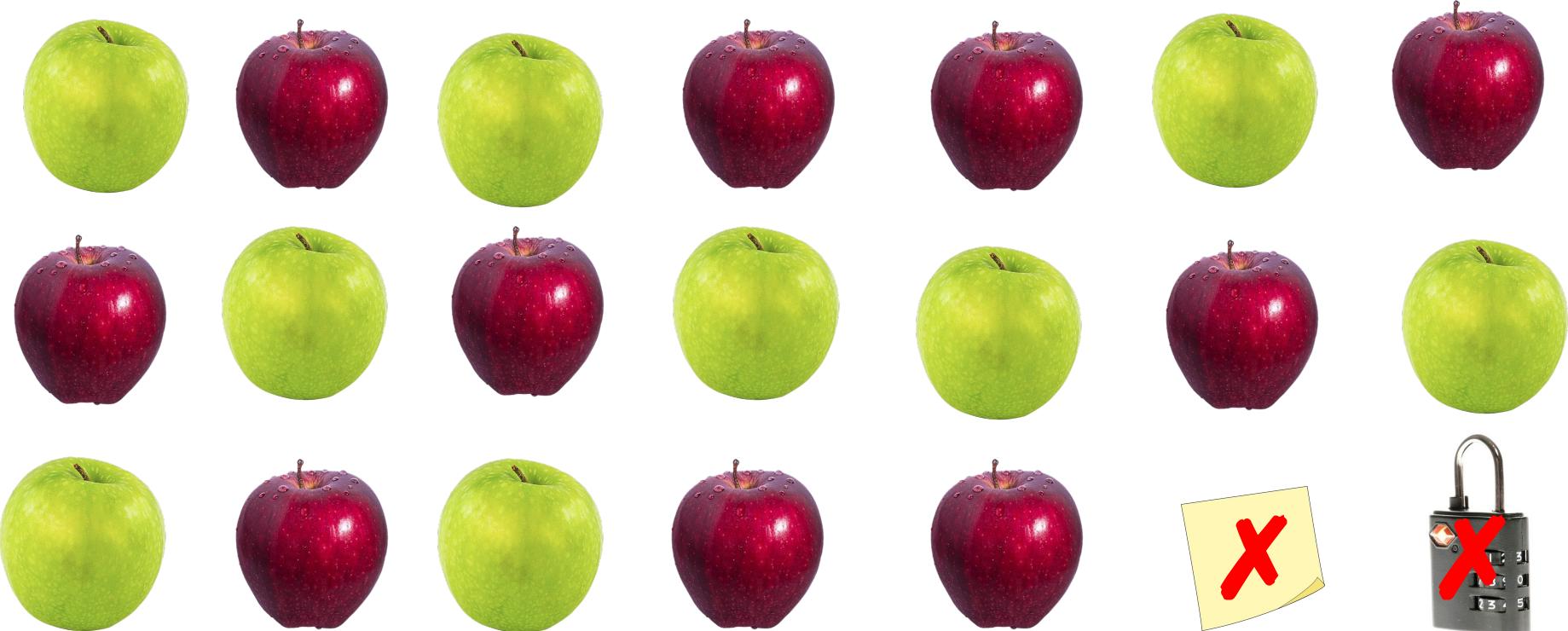
Input: image

Output: predicted object class

Instructions / Rules:

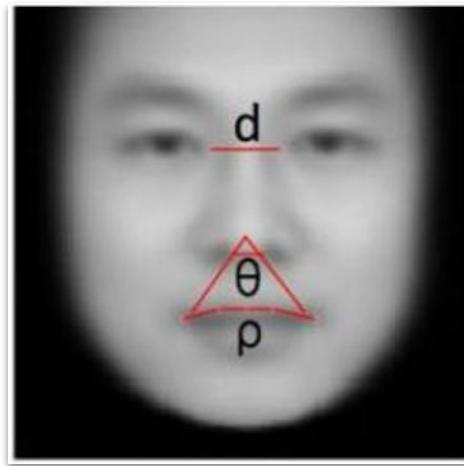
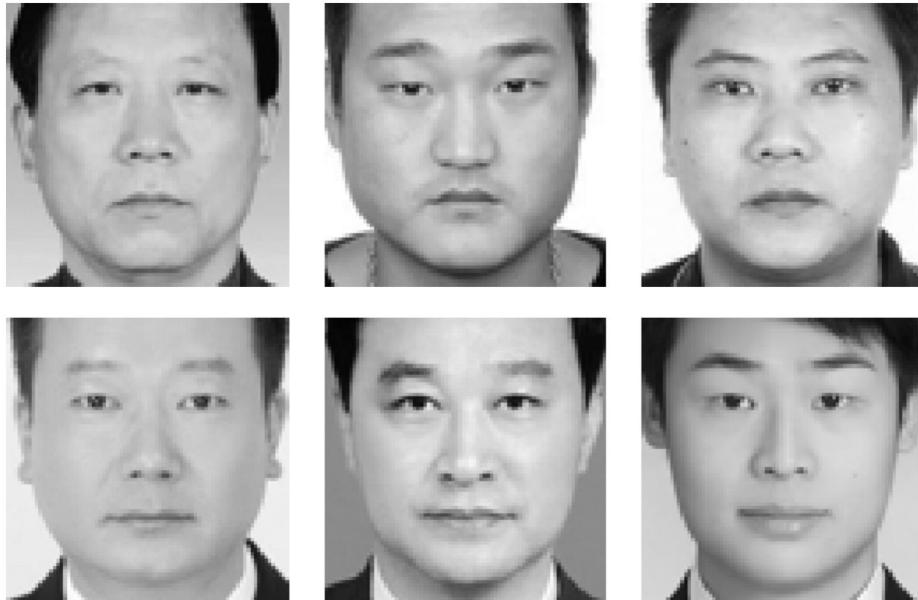
```
# 100% of the time, no matter what input  
object = 'apple'
```

Source of bias: Very unrepresentative dataset & evaluation metrics



Dataset bias and flawed evaluation

Wu and Zhang 2016, arxiv: 1611.04135v2



New approaches to evaluation

Evaluating geographic robustness



OpenImages Training Set



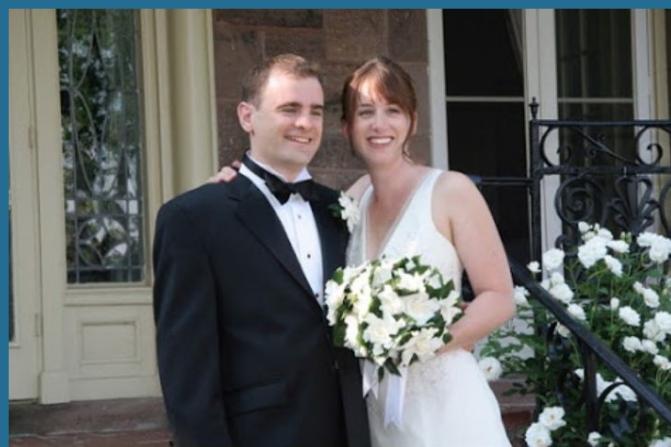
InclusiveImages
Public Validation



InclusiveImages
Private Final Test

NeurIPS 2018 InclusiveImages Competition

Challenging cultural bias in images



ceremony, bride, wedding, man, groom, woman, dress



*ceremony, wedding, bride,
man, groom, woman, dress*

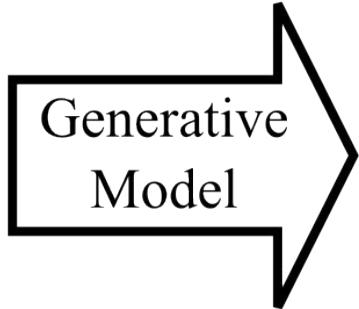


person, people

NeurIPS 2018 InclusiveImages Competition

Evaluating generative models

Noise $\sim N(0,1)$

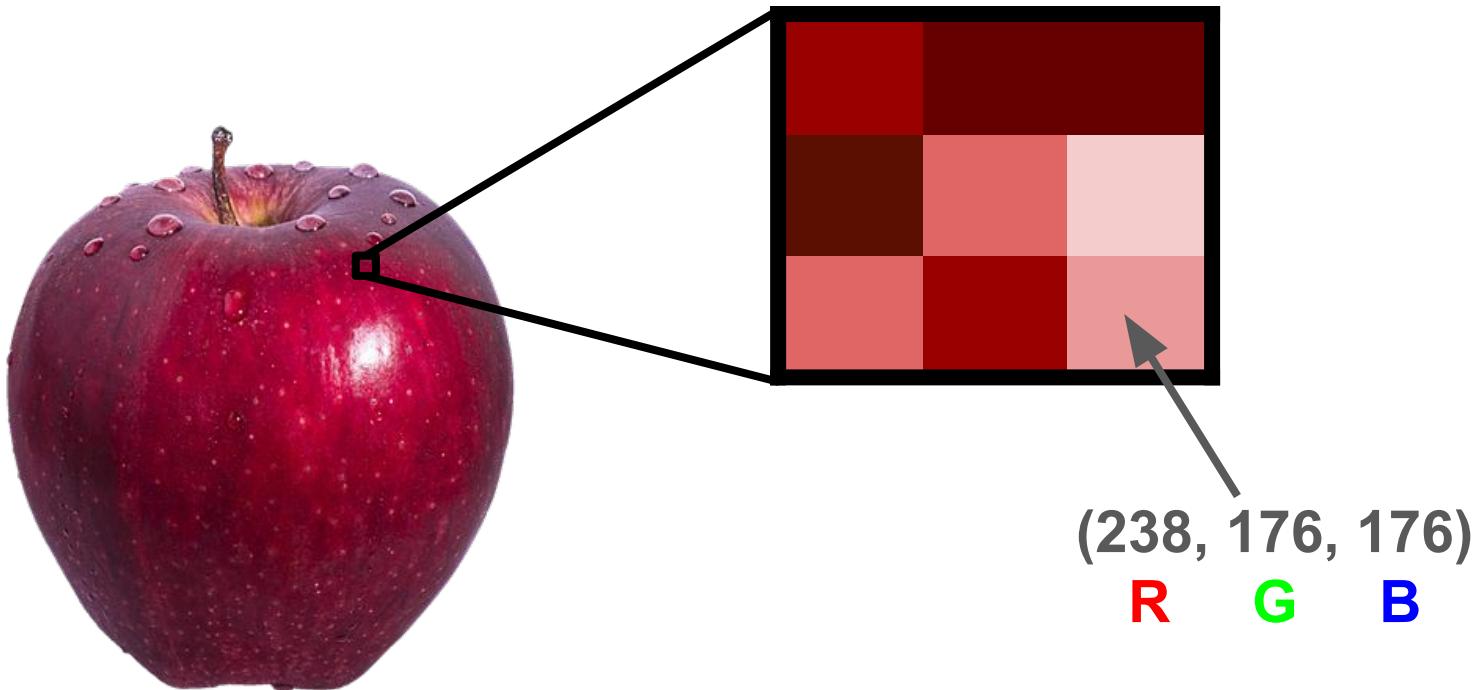


Concern / Source of Bias

Faulty representations

The structures to represent objects and concepts in the system leads to bias.

Representation of digital images



Representation of objects

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

= apple =



=



Representation of objects

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \text{apple}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \text{paper clip}$$

Representation of words

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{"man"}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{"welder"}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{"macho"}$$

Bias shown in Google Translate

English to Turkish to English shows bias in some job roles.

The screenshot displays two separate Google Translate sessions. Each session has language selection bars at the top: English, Turkish, Spanish, Detect language, and a 'Translate' button. The first session shows the input 'She is a doctor.
He is a nurse.' being translated to 'O bir doktor.
O bir hemşire.' and then back to 'He is a doctor.
She is a nurse.' with a checkmark. The second session shows the input 'He is a doctor.
She is a nurse.' being translated to 'O bir doktor.
O bir hemşire' and then back to 'She is a nurse.' with a shield icon.

Bias shown in Google Translate

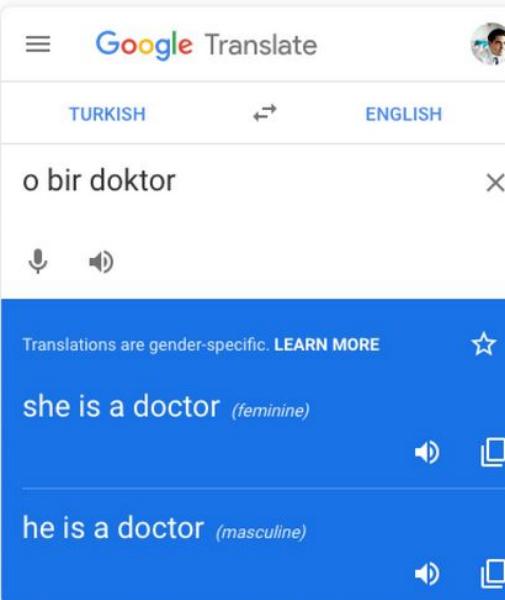
Google responds in December 2018 with one solution.

Before



The screenshot shows the Google Translate mobile app interface. At the top, it says "Google Translate". Below that, "TURKISH" is selected as the source language and "ENGLISH" as the target language. A blue text box contains the Turkish phrase "o bir doktor". Below the text box, there are two buttons: a microphone icon for voice input and a speaker icon for audio output. The English translation "he is a doctor" is displayed in a blue box at the bottom, accompanied by a small checkmark icon and a star icon for favoriting. The entire interface is white with blue highlights.

After



The screenshot shows the same Google Translate interface after a update. The English translation now reads "she is a doctor (feminine)". The rest of the interface remains the same, with the blue text box containing "o bir doktor", the microphone and speaker icons, and the blue box at the bottom containing the corrected translation.

Debiasing word embeddings

The popular representations for words considered the equivalent job to male programmers for women to be homemakers

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

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding.

Debiasing word embeddings

The popular representations for words considered the equivalent job to male programmers for women to be homemakers

Extreme <i>she</i> occupations		
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme <i>he</i> occupations		
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she-he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Gender stereotype <i>she-he</i> 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

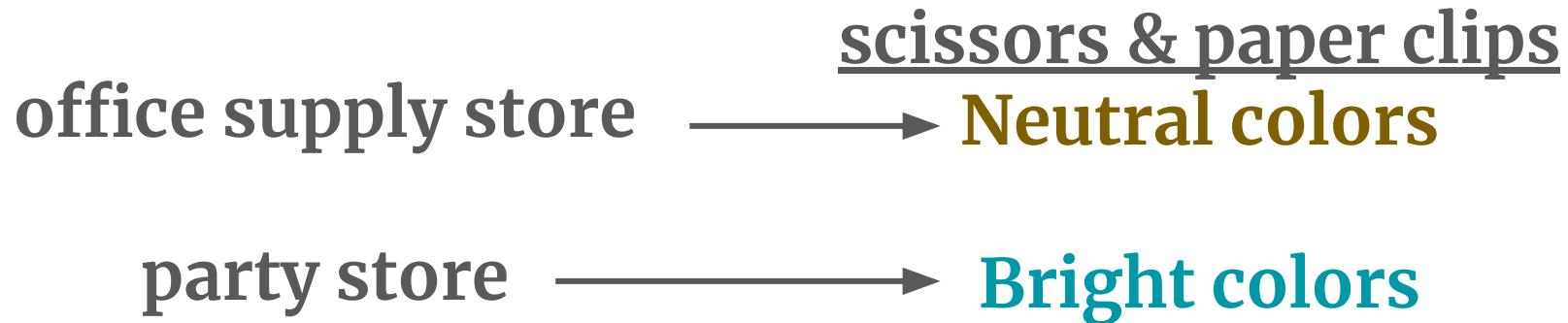
Gender appropriate <i>she-he</i> analogies.		
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Concern / Source of Bias

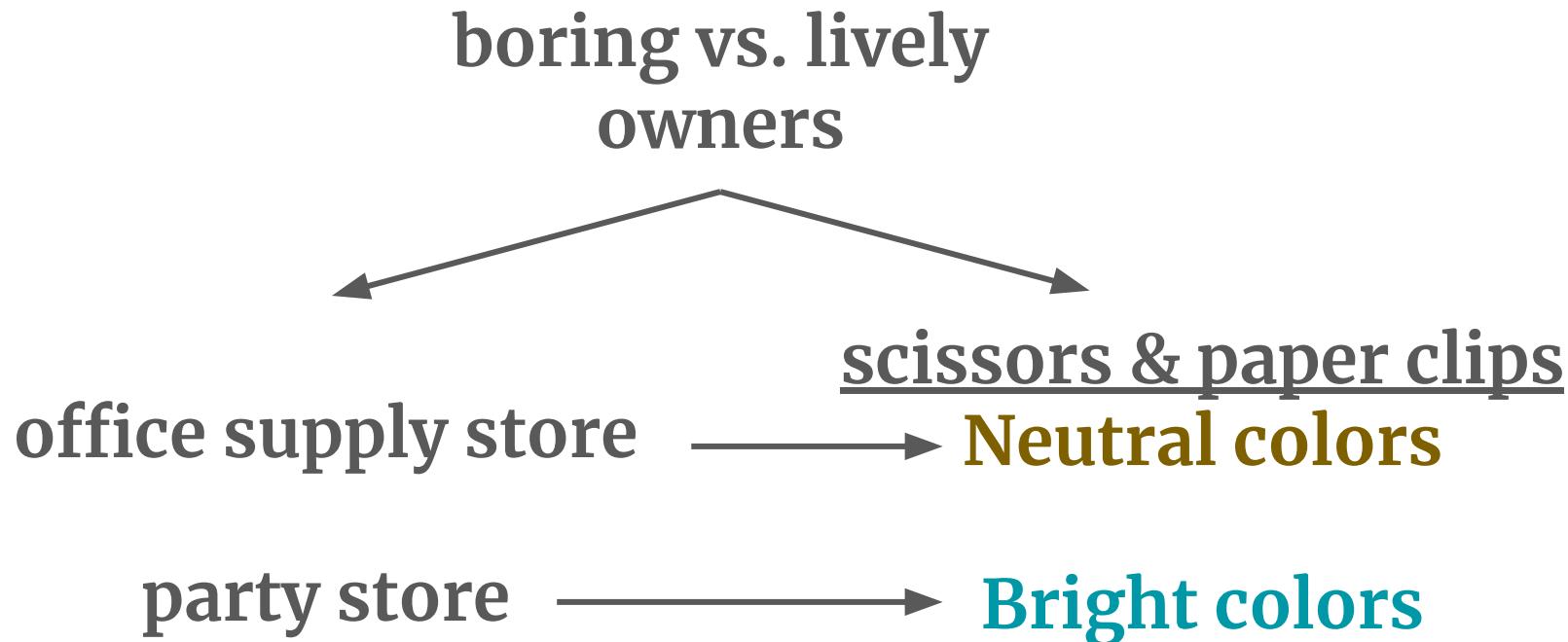
Confounding variables

We think we see a relationship between two things, but it turns out that there's a lurking factor causing the correlation.

Source of bias: Confounding variables



Source of bias: Confounding variables

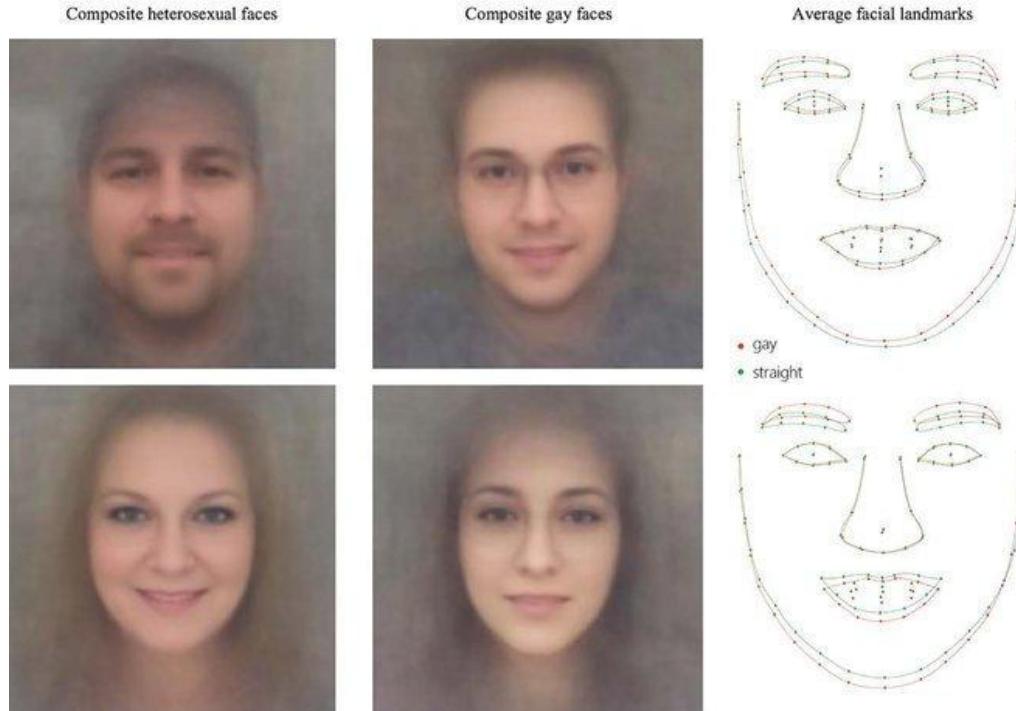


Constrained to what we can measure or observe

Decision space	Construct space	Observed space
Performance in college	Intelligence	IQ
Performance in college	Success in High School	GPA
Recidivism	Propensity to commit crime	Family history of crime
Recidivism	Risk-averseness	Age
Employee Productivity	Knowledge of job	Number of Years of Experience

Tab. 1: Examples of construct space attributes and their corresponding observed space attributes for different outcomes

Confounding in orientation detector



Source: Wang and Kosinski (2017)

Concern / Source of Bias

Fairness; Normative vs Positive Ethics

Even if our model works perfectly on historical data, that may not reflect societal views on how these tools should work. But how do we define fairness?

Amazon's hiring AI experiment

Media coverage of resume screening software, October 2018



Fairness of process (or “unaware” algorithms)

If we don't use protected information, we aren't being discriminatory!

But this isn't true: many protected classes are correlated with other features that can be included.

Fairness of process

Individual fairness

(or fairness through awareness)

Treat *similar** individuals as *similarly** as possible



*similar in distribution of outcomes

Fairness of process

Individual fairness

(or fairness through awareness)

Treat *similar** individuals as *similarly** as possible



*similar despite protected classes

How do we measure similarity?

Fairness of process

Individual fairness

Group fairness

(or statistical parity)

*There should be near equal outcomes across
protected and non-protected groups*

Concern / Source of Bias

Interpretability

What do we do when our models are right
for the wrong reasons?

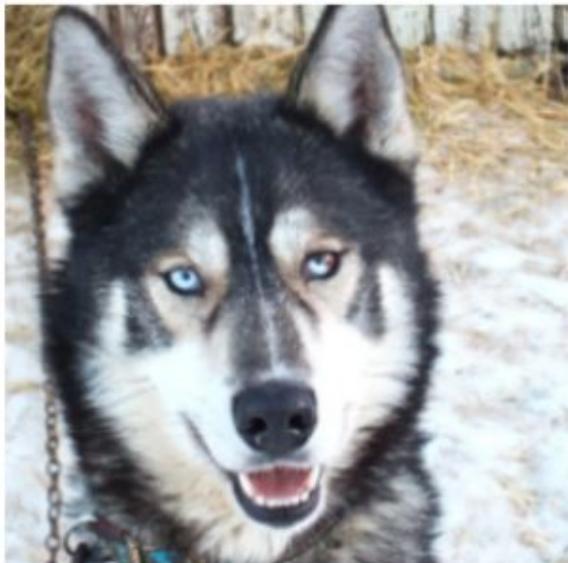
How would we know?

Large, complex models are rarely understood, even by experts

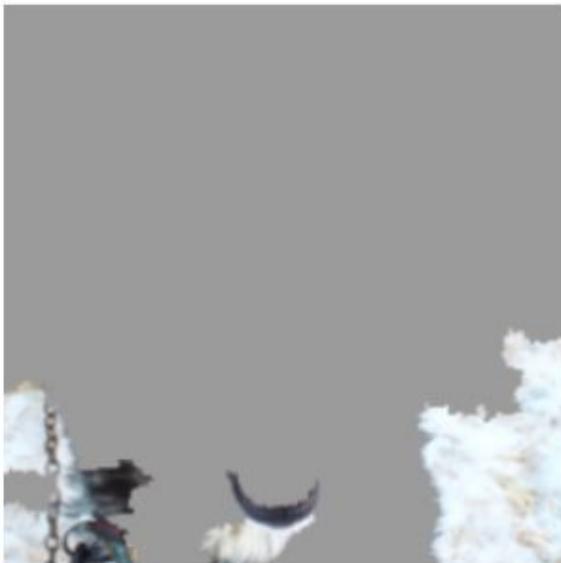


Models right for the wrong reasons

Ribeiro et al. 2015 (LIME)



(a) Husky classified as wolf



(b) Explanation

Multiple interpretability approaches

heavily borrowed from Kim & Doshi-Valez ICML 2017 tutorial

1. Fitting new models that are intrinsically interpretable
2. Post-hoc analysis of existing model (called an “explanation”)
3. Interpretable analysis of raw data

Explanations come in many forms

Linear regression models

(with a certain number of parameters)

Decision trees (or similar)

(with a certain depth/number of parameters)

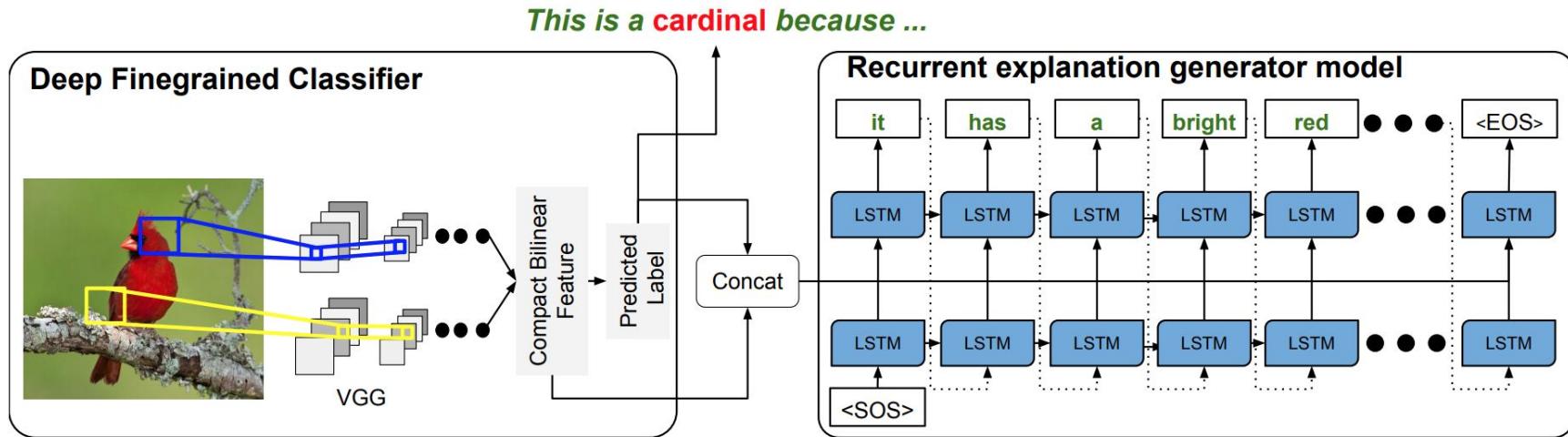
Text explanations

Visualizations (e.g., saliency maps)

more

Text explanations are one such form

Hendricks et al. 2016



This is a pine grosbeak because this bird has a red head and breast with a gray wing and white wing.



This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.



This is a pied-billed grebe because this is a brown bird with a long neck and a large beak.



This is an arctic tern because this is a white bird with a black head and orange feet.

Explanations can be persuasive

Herman 2017

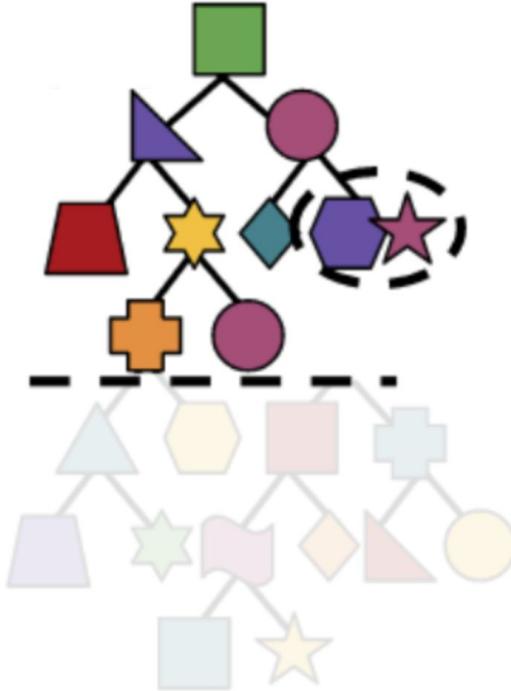
When tailoring our model explanations to human preferences and judges, our models may learn to prioritize persuasive explanations over introspective ones.

Splitting model form from simplicity

Herman 2017

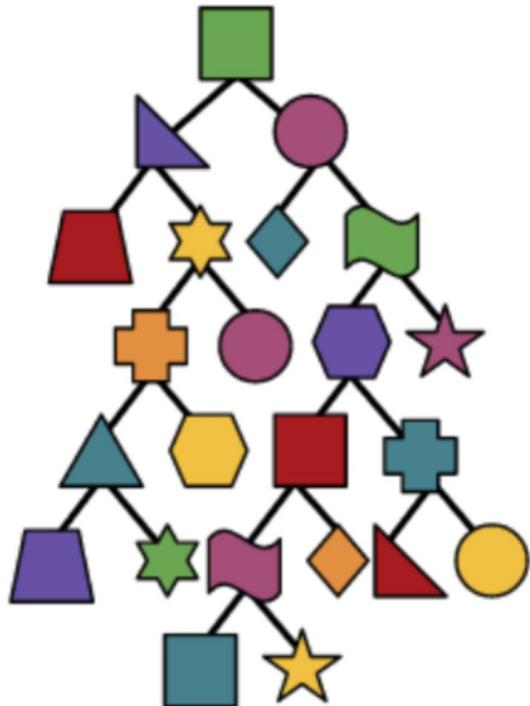
Simultaneously
coerced into suitable
model form
(e.g., decision tree)
and reduced in
complexity
(e.g., model size).

Difficult to evaluate
across complexity
preferences.

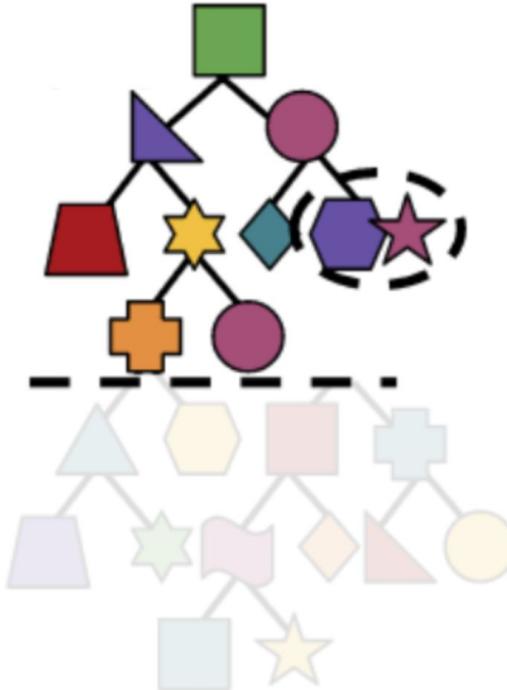
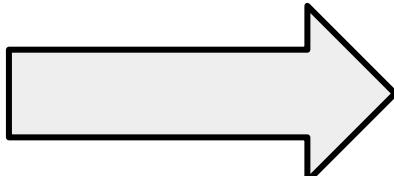


Splitting model form from simplicity

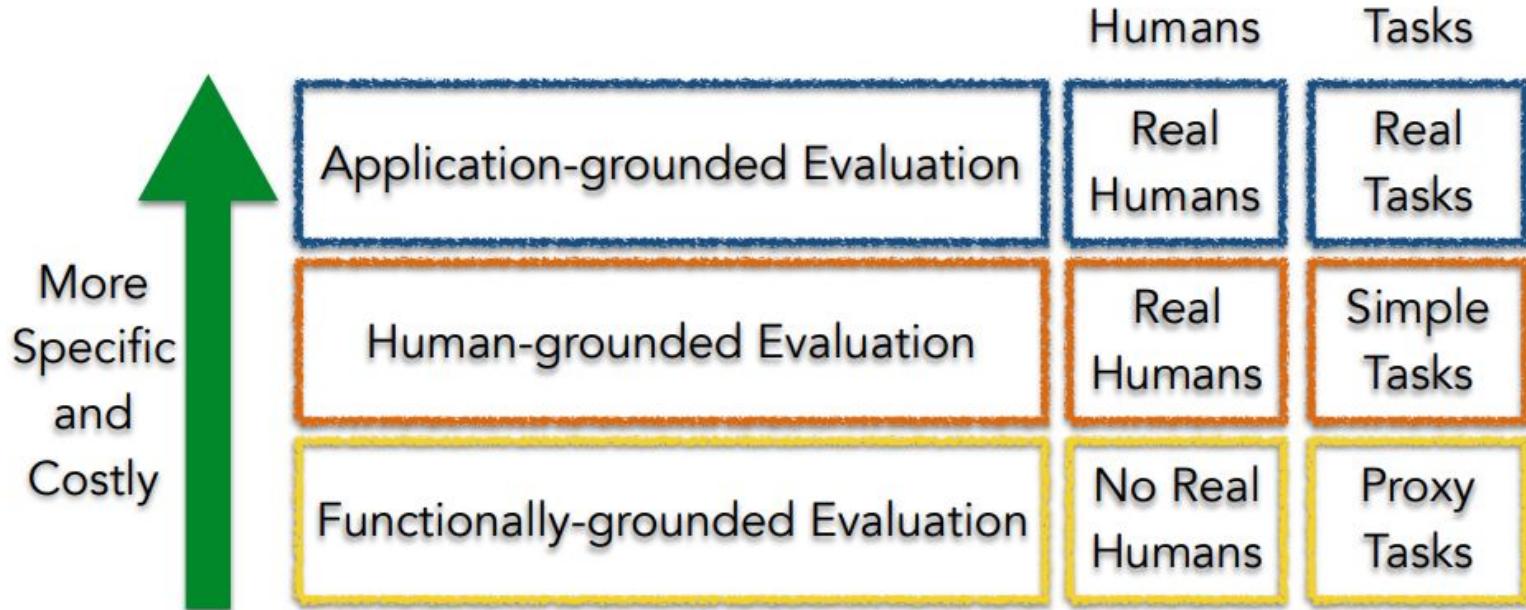
Herman 2017



Keeps model form
and reduction of
complexity separate.
Improves evaluation
and adaptability.



Taxonomy of evaluation for interpretability applies broadly



Many applications and topics are particularly human-centered

Ethics in AI

Interpretable and explainable ML

Language and speech

Music information retrieval

Affective computing

Some applications of computer vision

Concern / Source of Bias

Privacy and data rights

Are people informed about and unlikely to
be harmed by the collection / use / release
of the data in question?

How can we mitigate? Compensate?

For more information: (517) 351-1975
email: etov@grebnner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRINCE STO
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C050
THE JACKSON FAMILY
9999 MAPLE DR
FLINT MI 48507

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

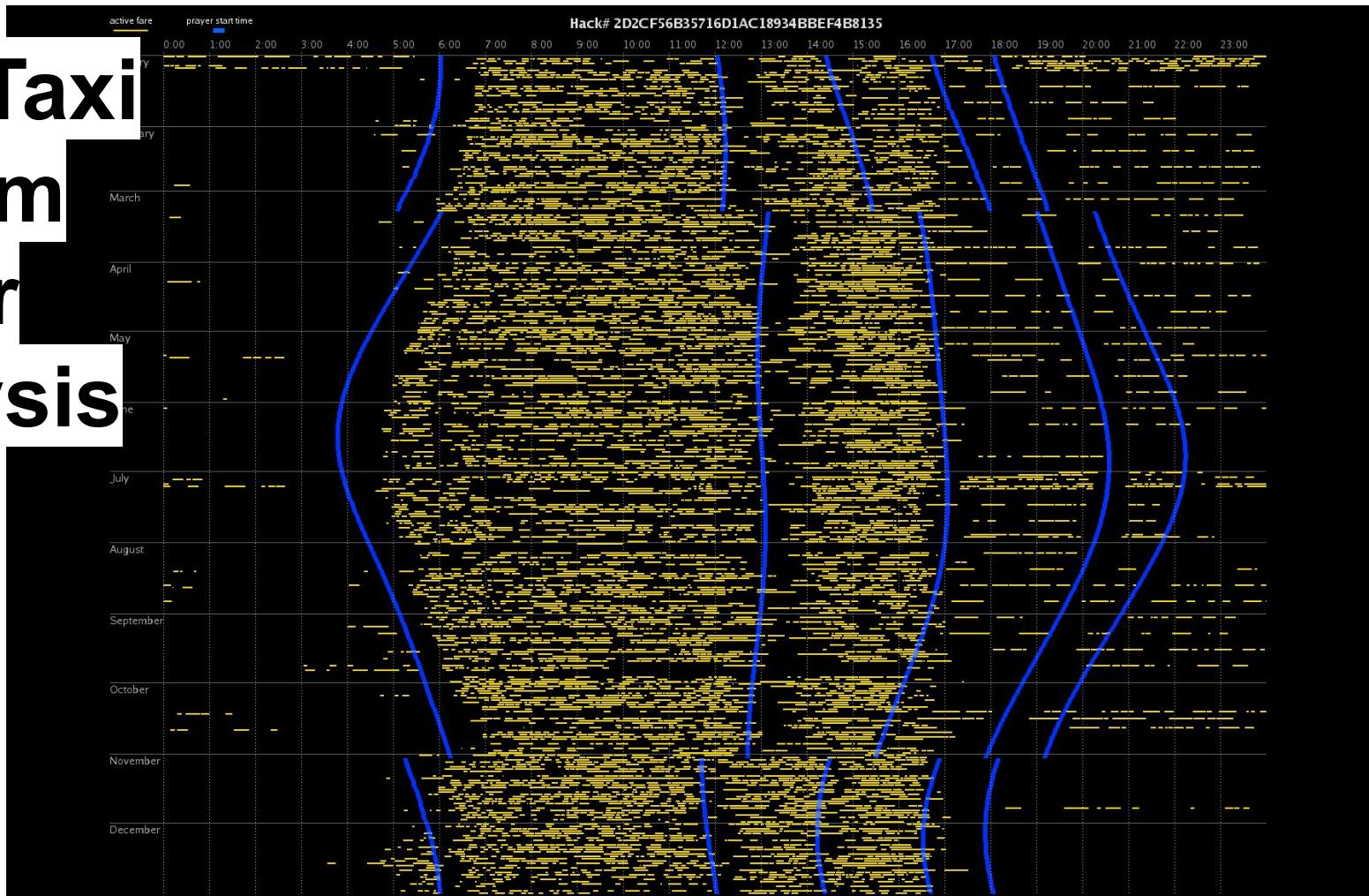
DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____
9999 BRIAN JOSEPH JACKSON		Voted	_____
9991 JENNIFER KAY THOMPSON		Voted	_____
9991 BOB R THOMPSON		Voted	_____
9993 BILL S SMITH			_____
9989 WILLIAM LUKE CASPER	Voted	_____	_____
9989 JENNIFER SUE CASPER	Voted	_____	_____

Social Pressure and Voter Turnout



NYC Taxi Muslim Driver Analysis



Synthetic datasets for privacy

Synthetic Data

- [Data Comparison](#)
- [Histogram Comparison](#)
- [Heatmap Comparison](#)

[Download synthetic data](#)

[Download dataset description](#)



Data comparison: input vs. synthetic (Independent Attribute Mode)

Original data (top 100 rows)

age	education	sex	relationship	marital-status	income
18	HS-grad	Female	Own-child	Never-married	<=50K
18	11th	Female	Own-child	Never-married	<=50K
18	HS-grad	Male	Not-in-family	Never-married	<=50K
19	HS-grad	Male	Own-child	Never-married	<=50K
19	HS-grad	Female	Wife	Married-AF-spouse	<=50K

Show 5 entries Search:

Showing 1 to 5 of 100 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [20](#) Next

Synthetic data (top 100 rows)

age	education	sex	relationship	marital-status	income
17	Masters	Male	Husband	Divorced	<=50K
17	7th-8th	Female	Husband	Married-civ-spouse	<=50K
17	Some-college	Female	Own-child	Never-married	<=50K
18	Assoc-acdm	Male	Own-child	Married-civ-spouse	<=50K
18	HS-grad	Male	Husband	Divorced	<=50K

Show 5 entries Search:

Showing 1 to 5 of 100 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [20](#) Next

Bonus slides:
**Other efforts toward
ethical data science**

Educational Game on AI Ethics

A light, educational game that addresses AI ethics (ongoing)

The screenshot shows a web browser window with the following details:

- Address Bar:** https://www.hcde.washington.edu/research/aragon#game-wi19
- Page Title:** Cecilia Aragon | Human Center
- Page Content (Left Column):**
 - Winter 2019*
 - ## Games for Good: Designing a Data Science Ethics Game
 - Co-directed by Cecilia Aragon and data scientist Bernease Herman**
 - This research group will explore the use of analog and digital games to introduce users to ethical and human-centered issues in data science and computing. Students will be hands-on in exploring examples of educational games, brainstorming and providing ideas for games, and creating prototypes using paper and/or a computer game engine such as Unity3D. Some themes we will consider include data privacy, trust of algorithmic systems, predictive policing, fairness, and others.
 - At the end of ten weeks, we aim to produce a working prototype of the game, including
- Page Content (Right Column):**
 - Cecilia Aragon
 - Cecilia Aragon's Research Group Archives
 - Elin Björling
 - Brock Craft
 - Andrew Davidson
 - Leah Findlater
 - Tyler Fox
 - Mark Haselkorn

Codes of Ethics

Community Principles for Ethical Data Practices

The screenshot shows a web browser window with the following details:

- Title Bar:** "Community Principles on Ethic x" and "Bernease".
- Address Bar:** "Secure | https://datapractices.org/community-principles-on-ethical-data-sharing/".
- Toolbar:** Includes icons for Apps, Bookmarks, and a star for bookmarks.
- Content Area:**
 - A blue sidebar on the left contains a white circle with a balance scale icon, a "BACK" button, and the text "Community Principles on Ethical Data Practices". Below this is an "OVERVIEW" button.
 - The main content area has a light blue background.
 - Text: "As data practitioners and data consumers, we aim to..." followed by a list of principles.
 - A "PRINCIPLES" section with a list of numbered items.
 - A "SUBSCRIBE" button at the bottom right.

Data Science for Social Good

Summer program at UW to help advance work in these areas

The screenshot shows a web browser window for the eScience Institute - UW Data Science for Social Good. The URL is https://escience.washington.edu/dssg/. The page features a dark header with navigation links for Home, About Us, Research, Education, Get Involved, Contact Us, and a search icon. Below the header, a large banner displays the text "UW Data Science for Social Good" over a background of network nodes. A descriptive paragraph explains the program's purpose: "The Data Science for Social Good summer program brings together students, stakeholders, data and domain researchers to work on focused, collaborative projects for societal benefit." At the bottom, there is a group photo of approximately 15 people, mostly young adults, standing in front of a wall with the "eScience Institute" logo.

eScience Institute - UW Data S X +

https://escience.washington.edu/dssg/

Apps People · UWSEDS...

Home About Us Research Education Get Involved Contact Us

UW Data Science for Social Good

The Data Science for Social Good summer program brings together students, stakeholders, data and domain researchers to work on focused, collaborative projects for societal benefit.

eScience Institute

Questions and Thank You

University of Washington eScience Institute
Gordon and Betty Moore Foundation
Alfred P. Sloan Foundation

Collaborators: Bill Howe, Julia Stoyanovich, Cecilia Aragon, Alina Arseniev-Koehler, Jay Rutherford, others

bernease@uw.edu