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

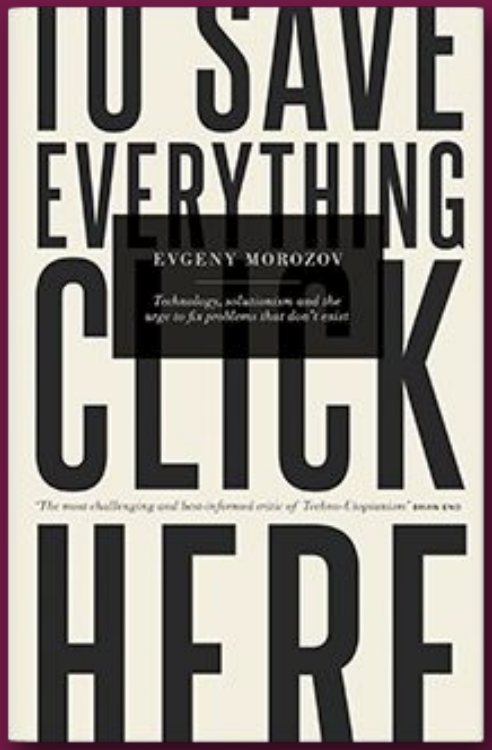
Purposeful harm or manipulation

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’

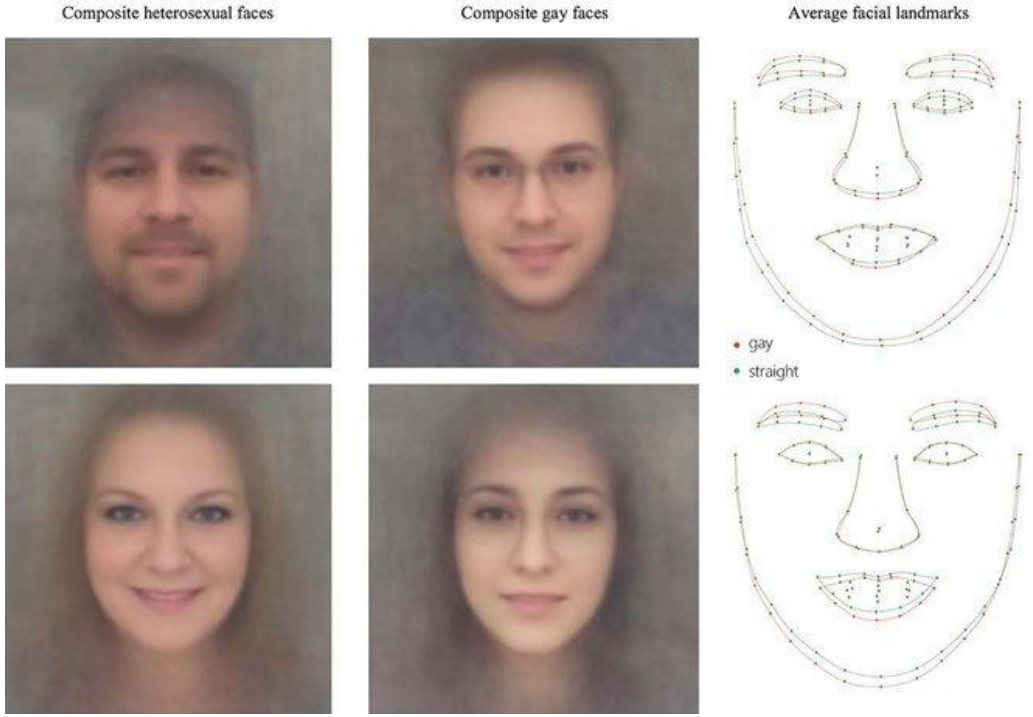
Concern / Source of Bias



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.


NEWS

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.

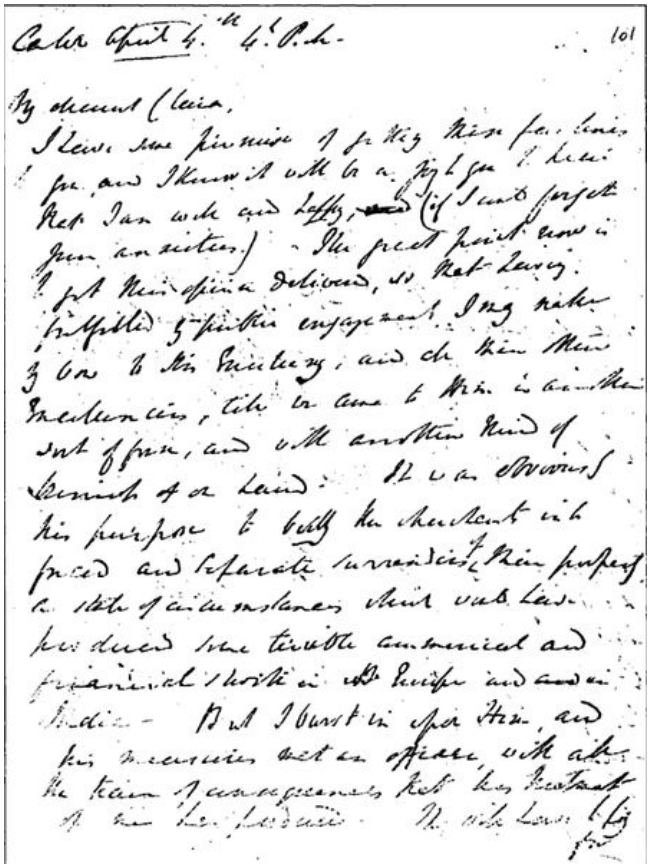
Concern / Source of Bias

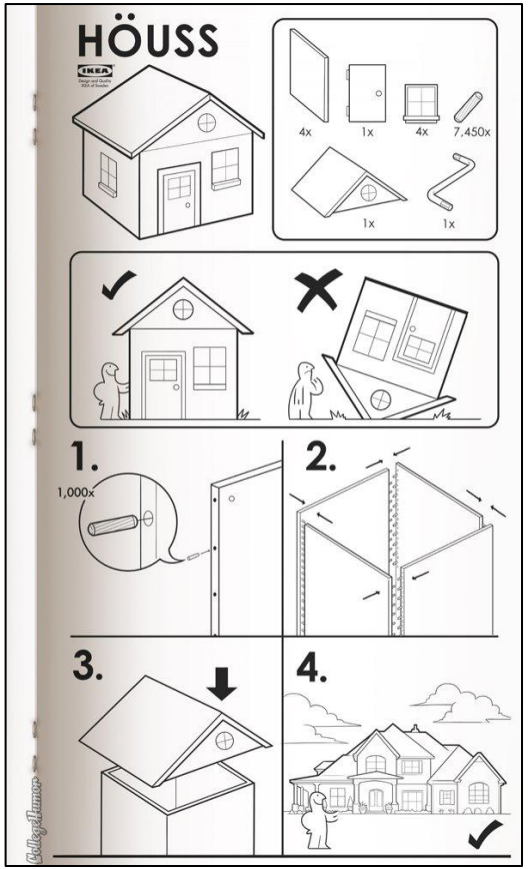
Measurement error, or mismeasured inputs

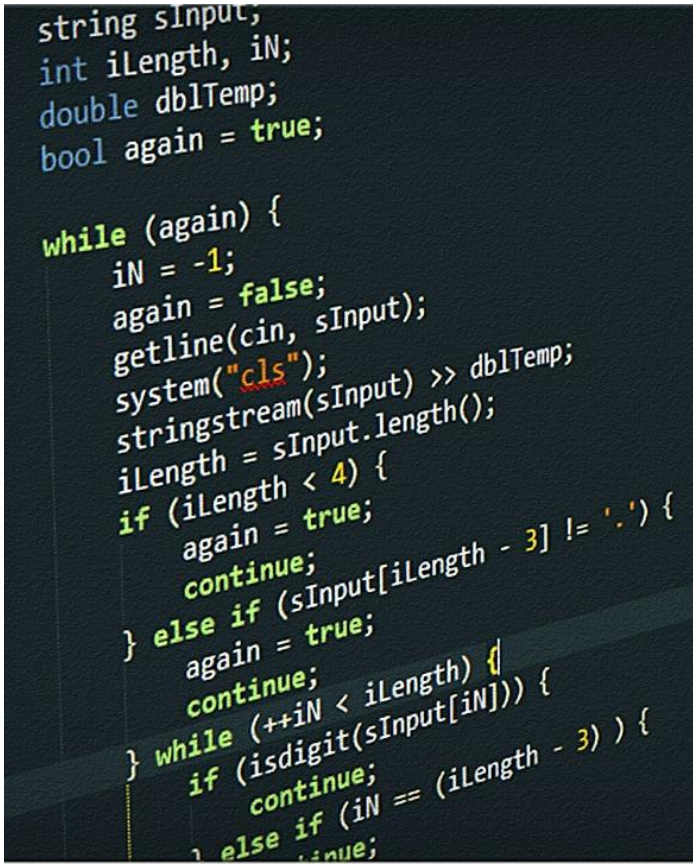
When gathering input data, we may have recorded an incorrect value.

Human-specified algorithms

What's meant by human-specified?





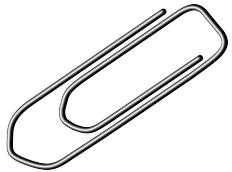


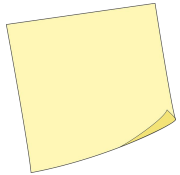
More difficult: Object classification

Object classification algorithm
Input: image
Output: predicted object class

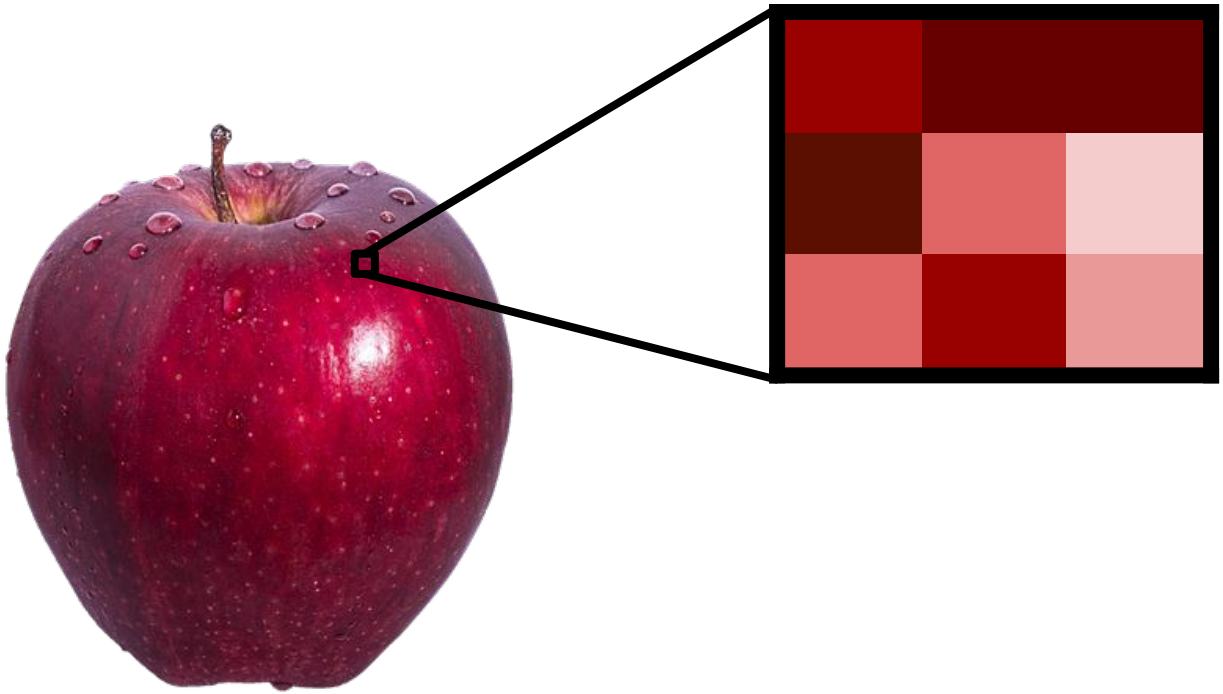








Representation of digital images

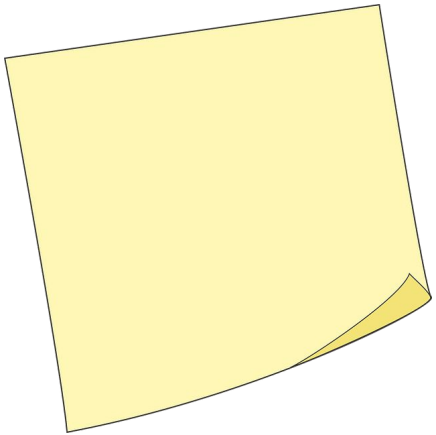


Writing rules for human-specified algo

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.

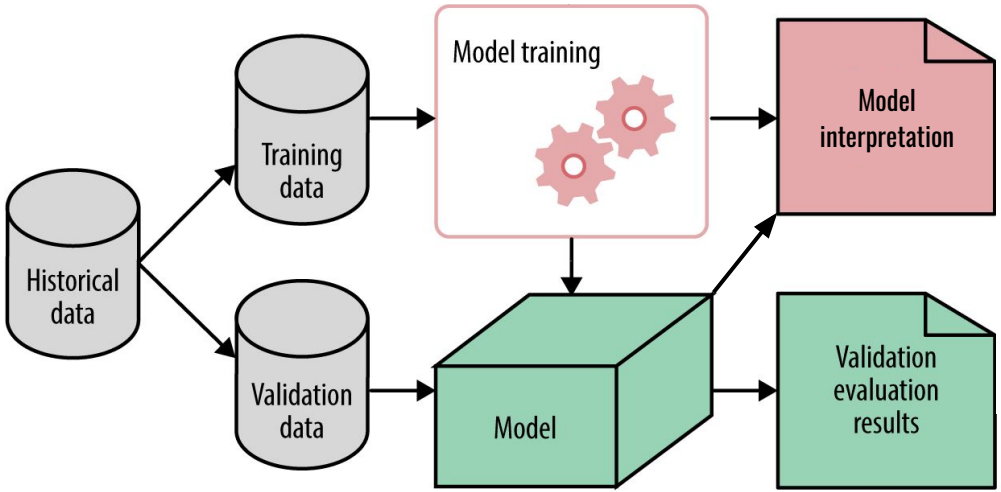
EXAMPLES: Generating features for object classification algorithm

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

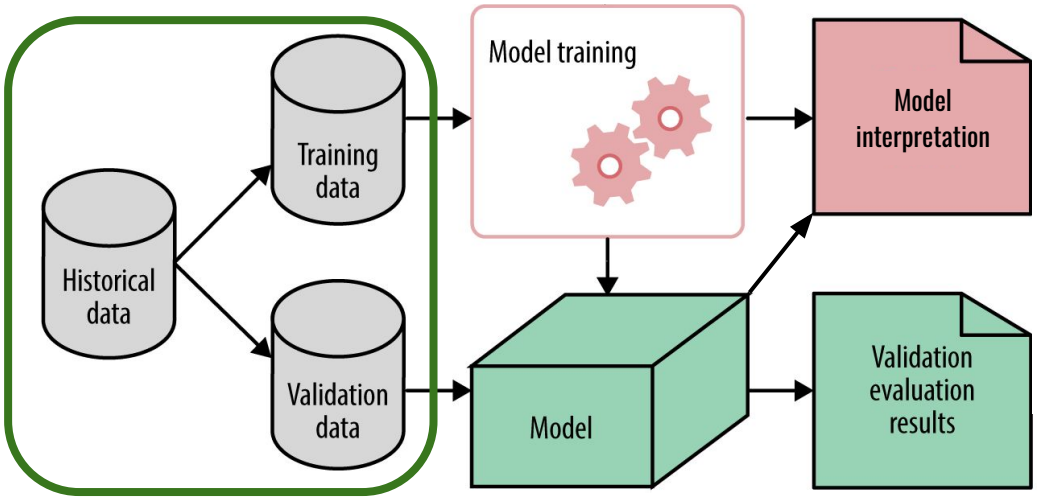
Machine learning pipeline



Modified graphic. Original from Alice Zheng

How do we choose and collect dataset?

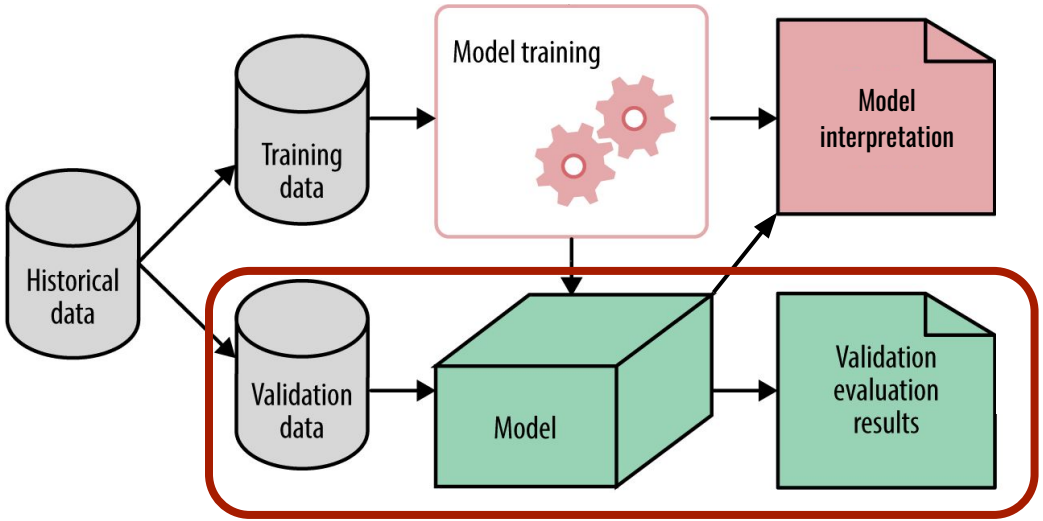
Data Curation · Evaluation · Interpretation



Modified graphic. Original from Alice Zheng

How to check performance of ML?

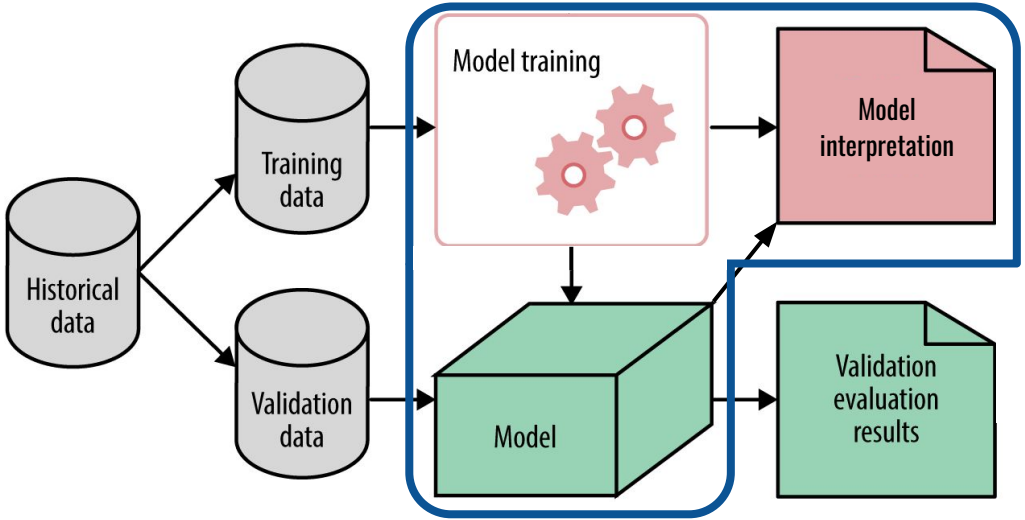
Data Curation · **Evaluation** · Interpretation



Modified graphic. Original from Alice Zheng

What's happening on the inside?

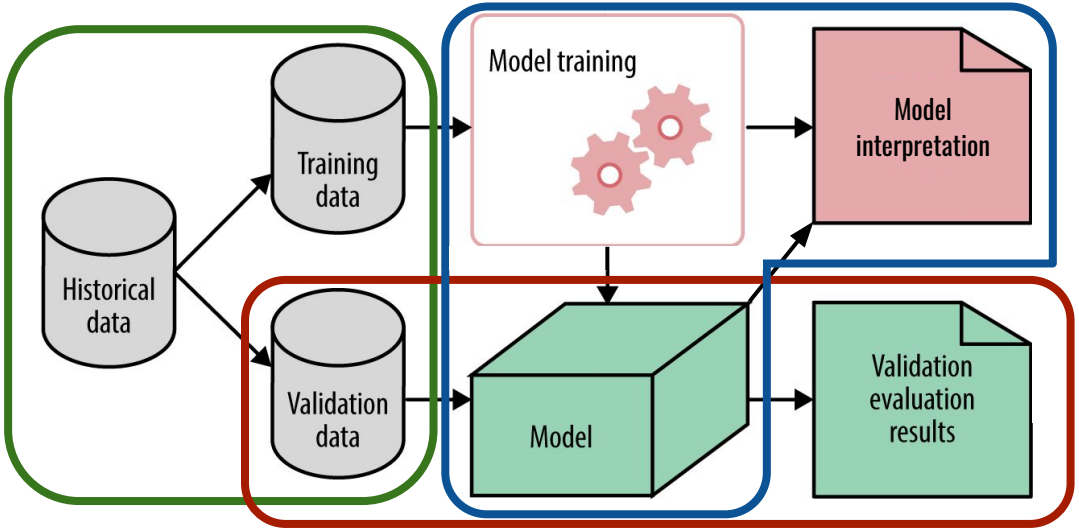
Data Curation · Evaluation · **Interpretation**



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

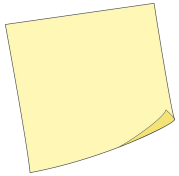
Labeling error

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

Source of bias: Labeling error

scissors ✓

apple ✓



window ✗

shopping bag ✓

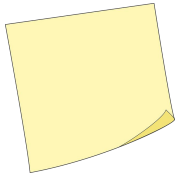
Source of bias: Labeling error

scissors
scissors
scissors

✓

apple
poison
apple

✓



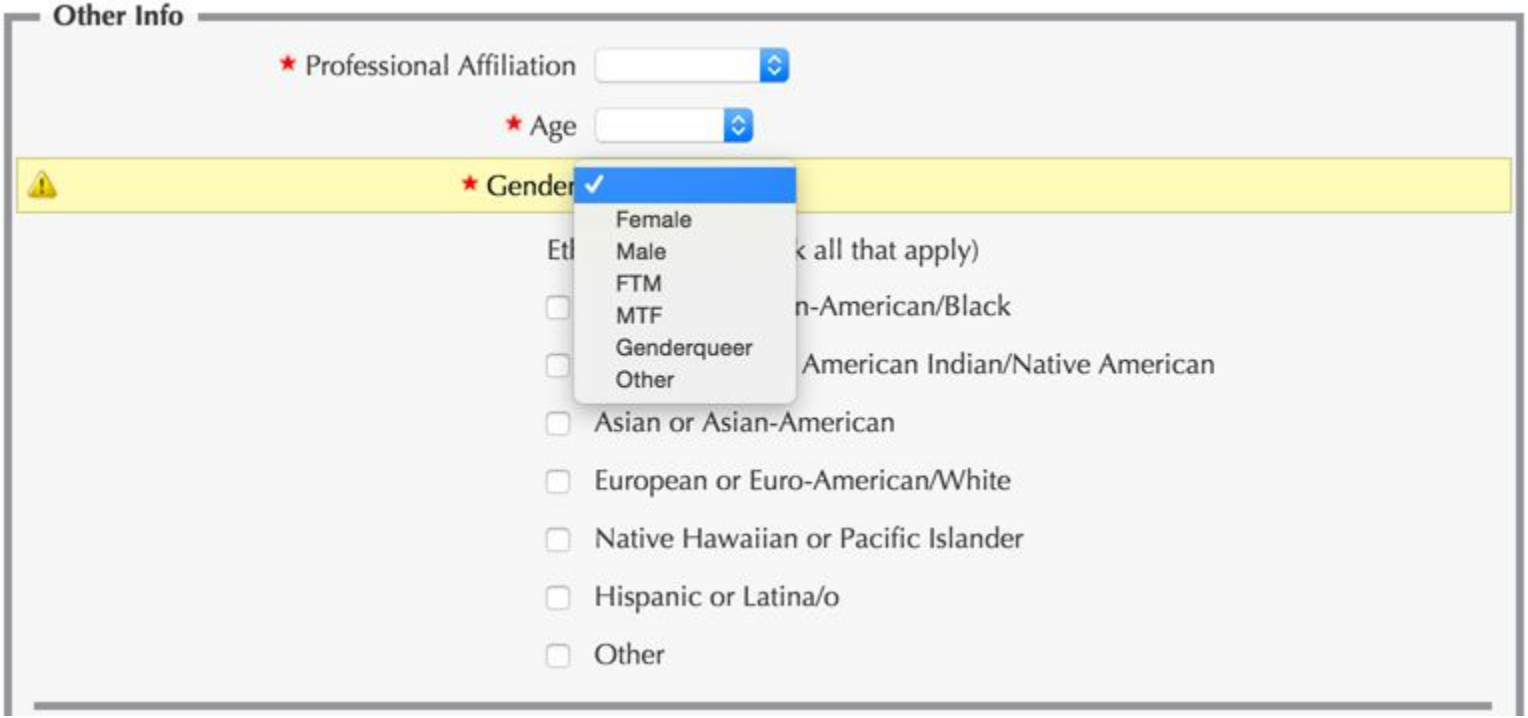
window
stickies
paper

✗

shopping bag
lock
lock

✓

Source of bias:
Self-reported labeling error



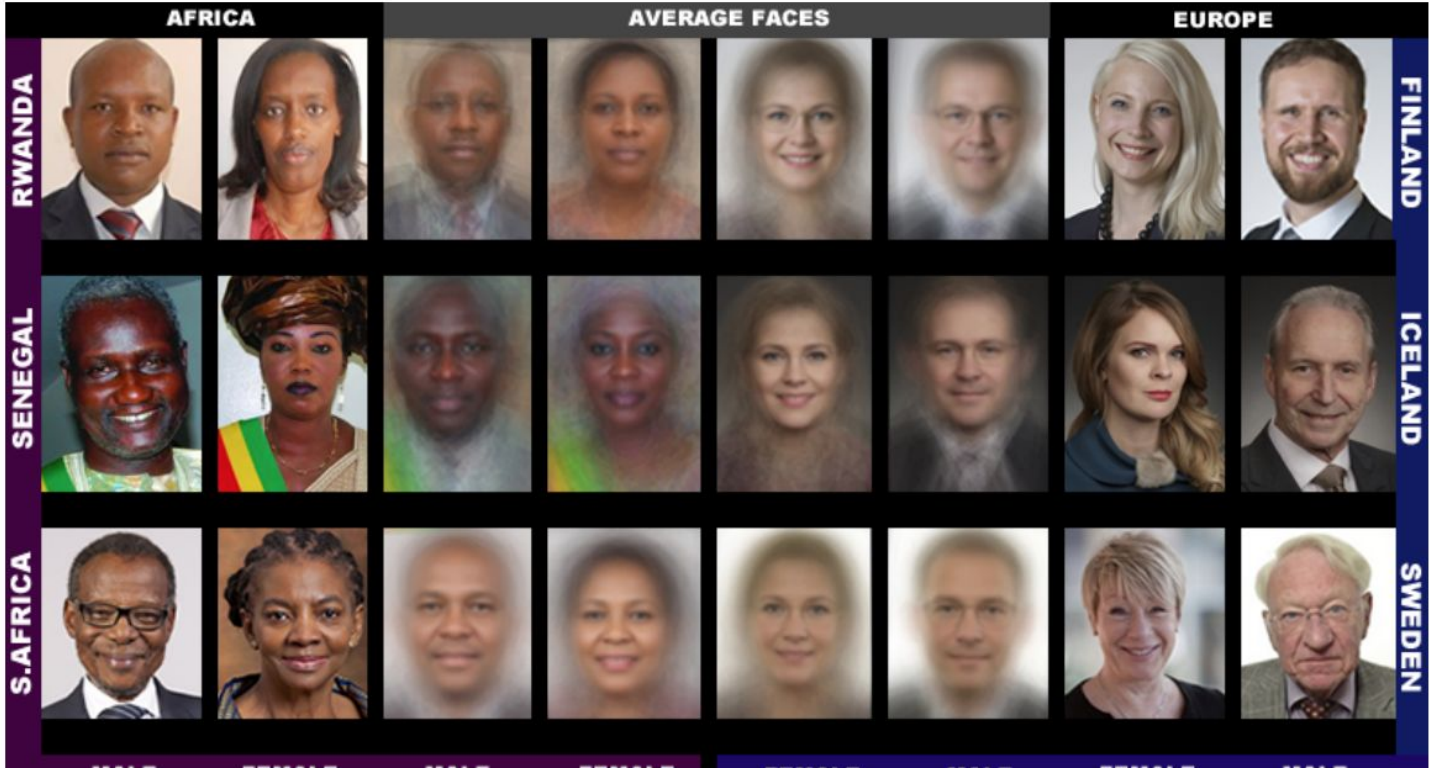
Concern / Source of Bias

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



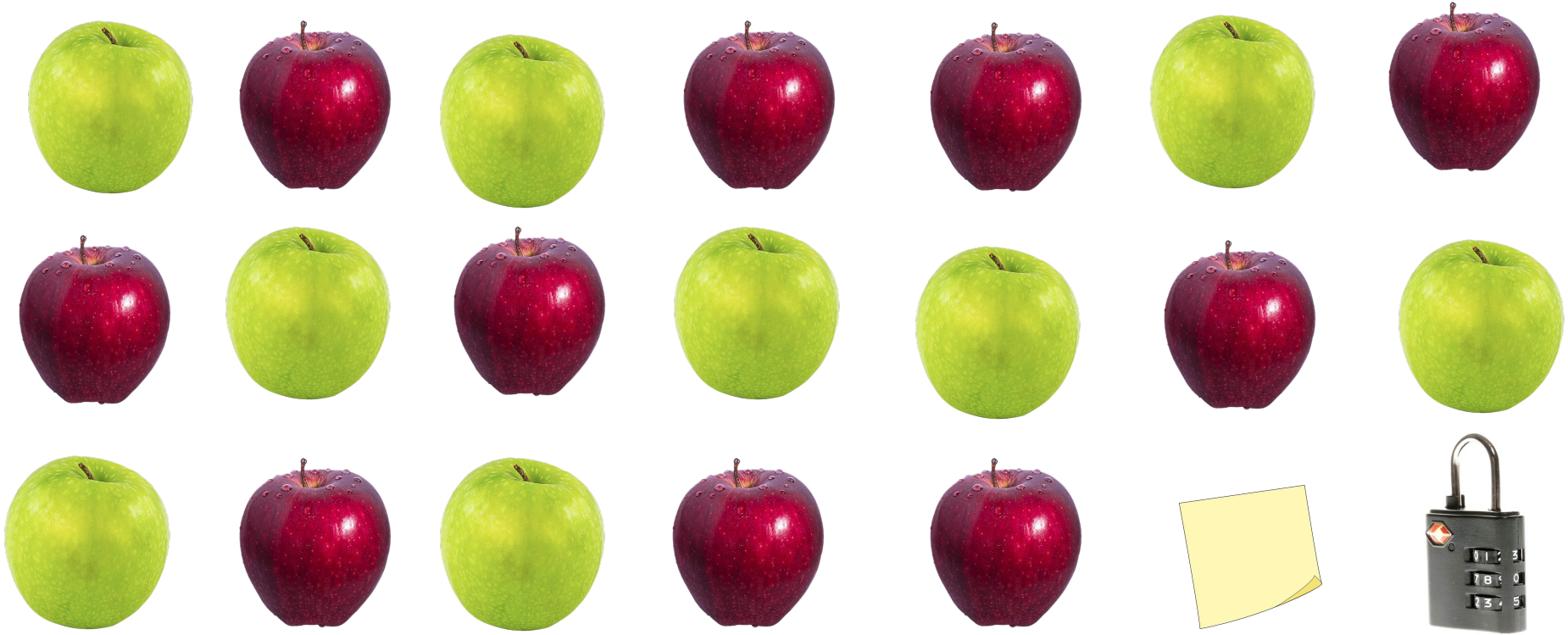
Dataset is not balanced across gender or skin tone

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

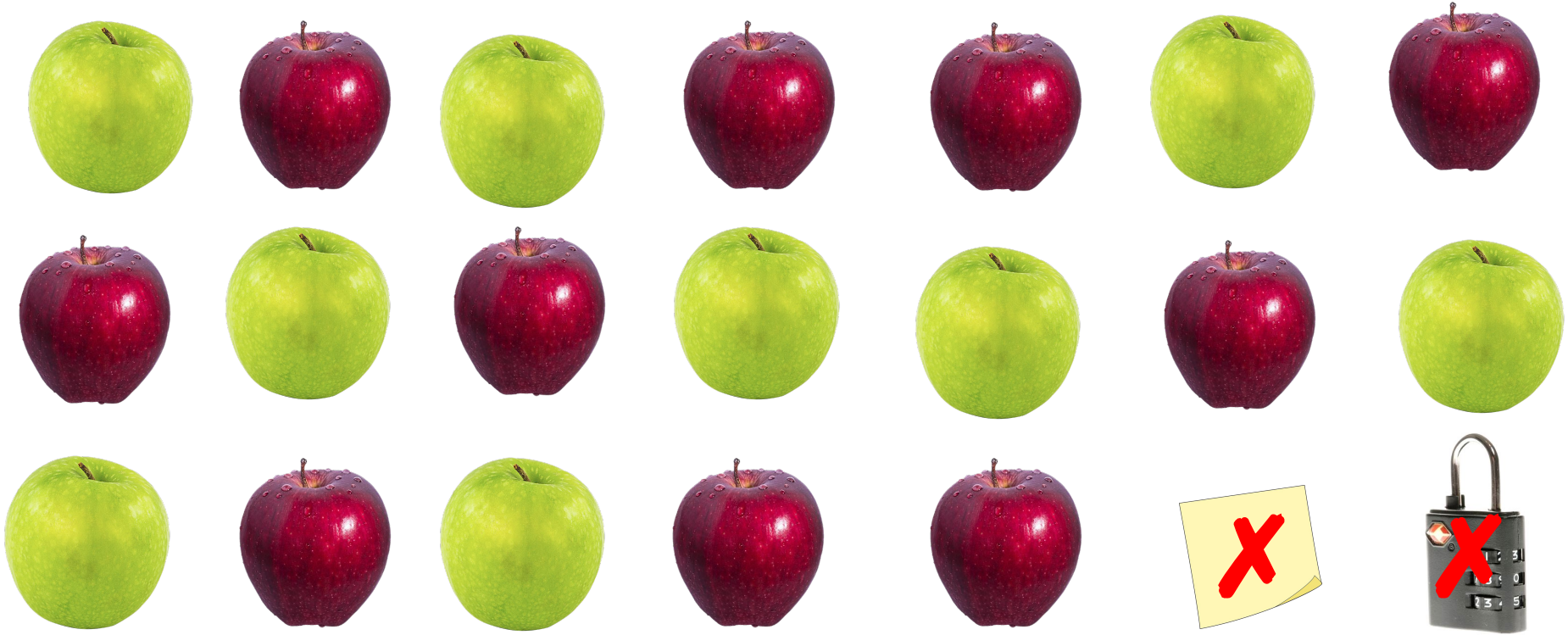


Writing rules like human-specified algo

Object classification algorithm
Input: image
Output: predicted object class
Instructions / Rules:

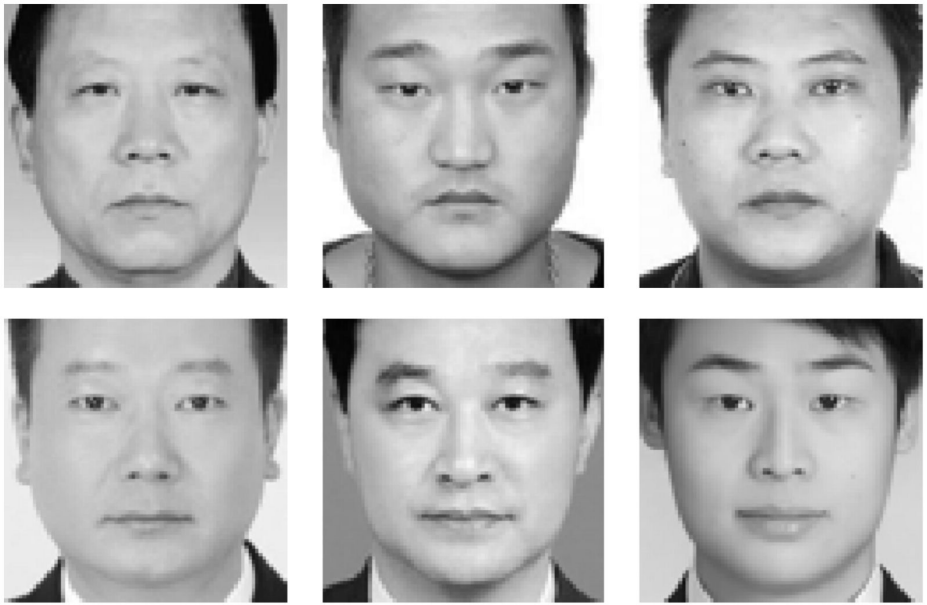
```
# 100% of the time, no matter what input
object = 'apple'
```

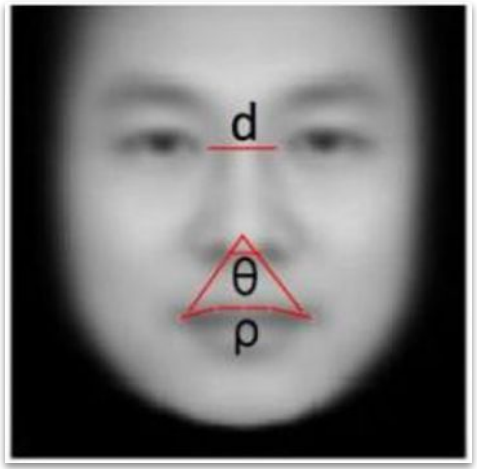
Source of bias: Very unrepresentative dataset & evaluation metrics



Dataset bias and flawed evaluation

Wu and Zhang 2016, arxiv: 1611.04135v2



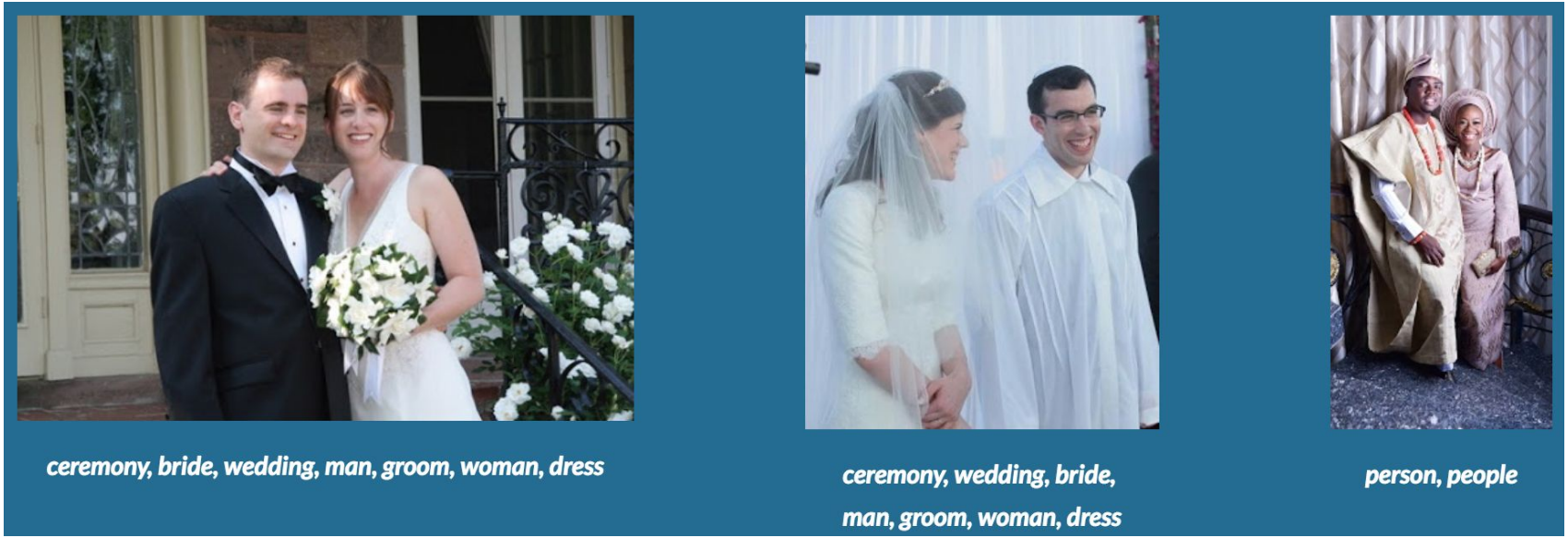


New approaches to evaluation

Evaluating geographic robustness

NeurIPS 2018 InclusiveImages Competition

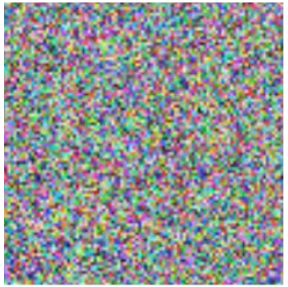
Challenging cultural bias in images



NeurIPS 2018 InclusiveImages Competition

Evaluating generative models

Noise ~ N(0,1)



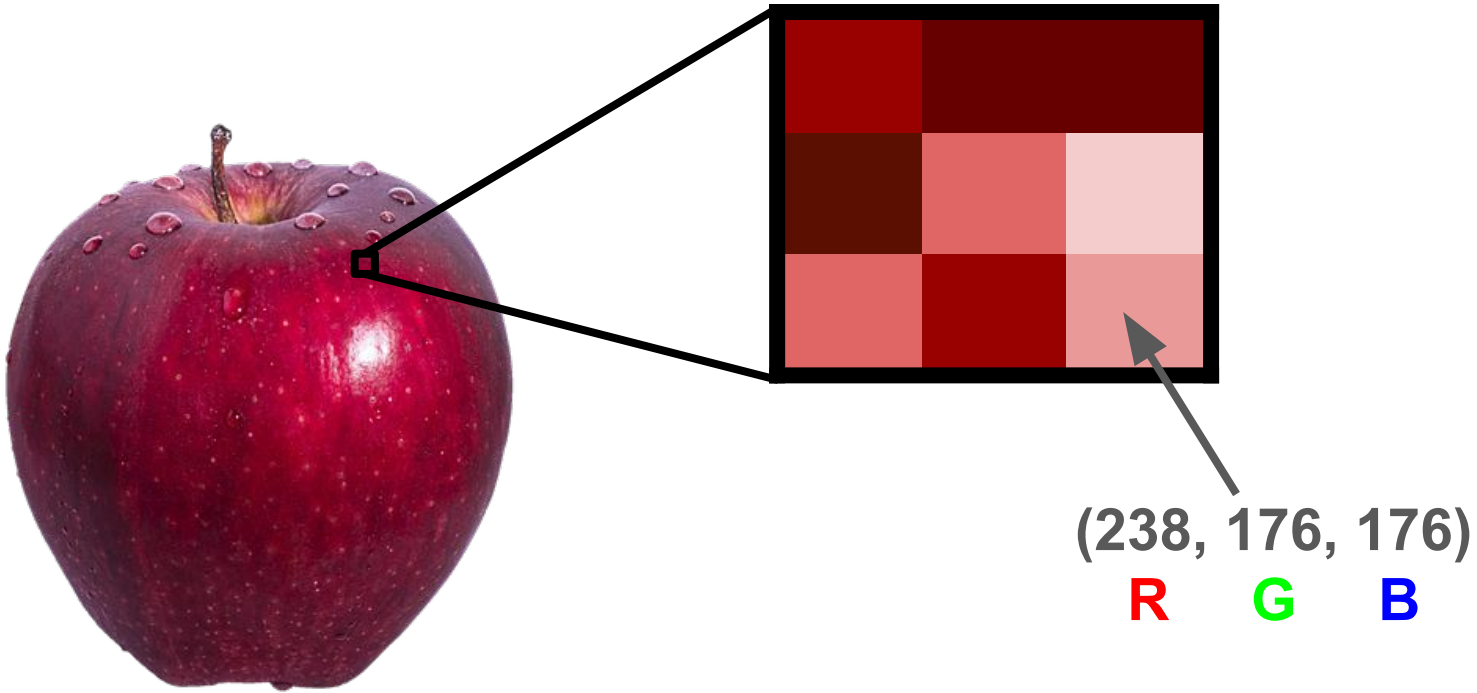
Generative Model

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

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

Representation of objects

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

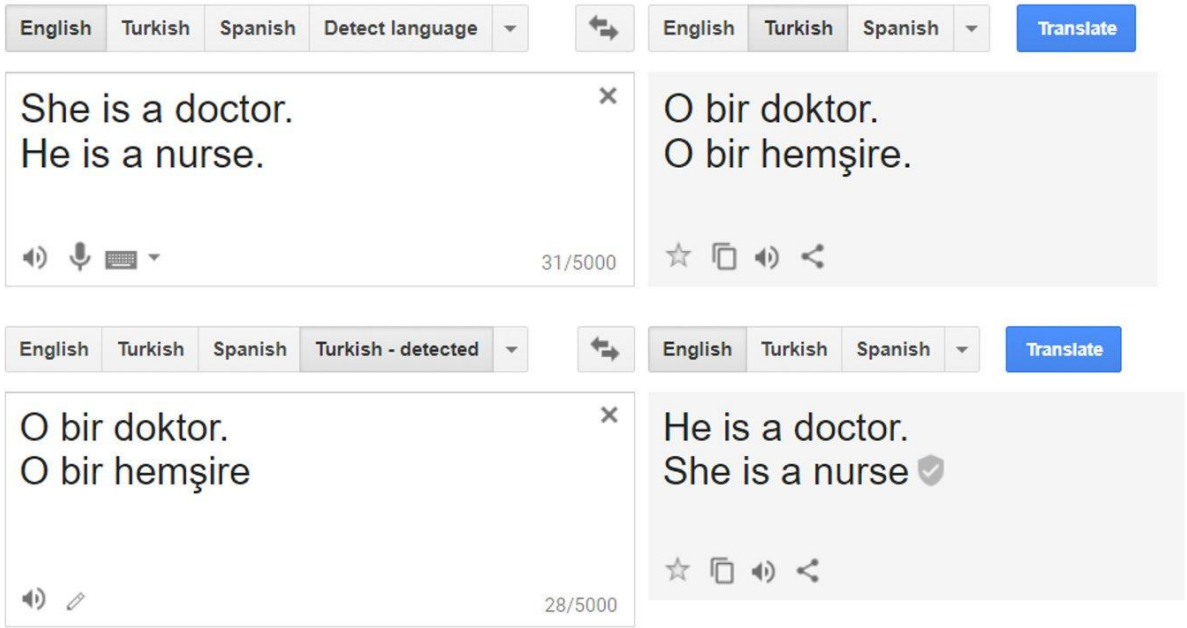
$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \text{paper clip}$$

Representation of words

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{“man”} \qquad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{“welder”} \qquad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \text{“macho”}$$

Bias shown in Google Translate

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

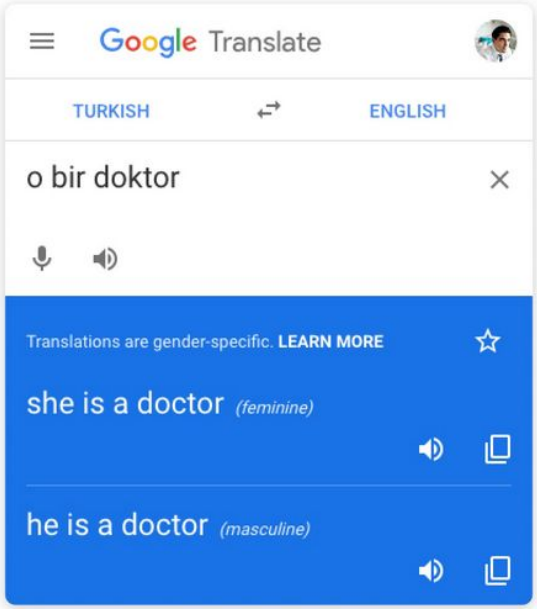


Bias shown in Google Translate

Google responds in December 2018 with one solution.

Before

After



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 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

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 *she* 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 *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher 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 *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.

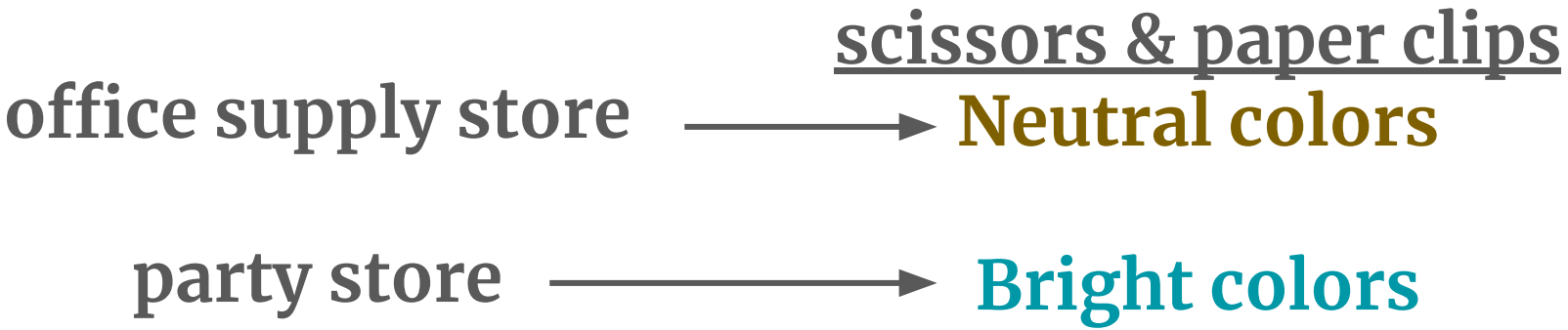
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Concern / Source of Bias

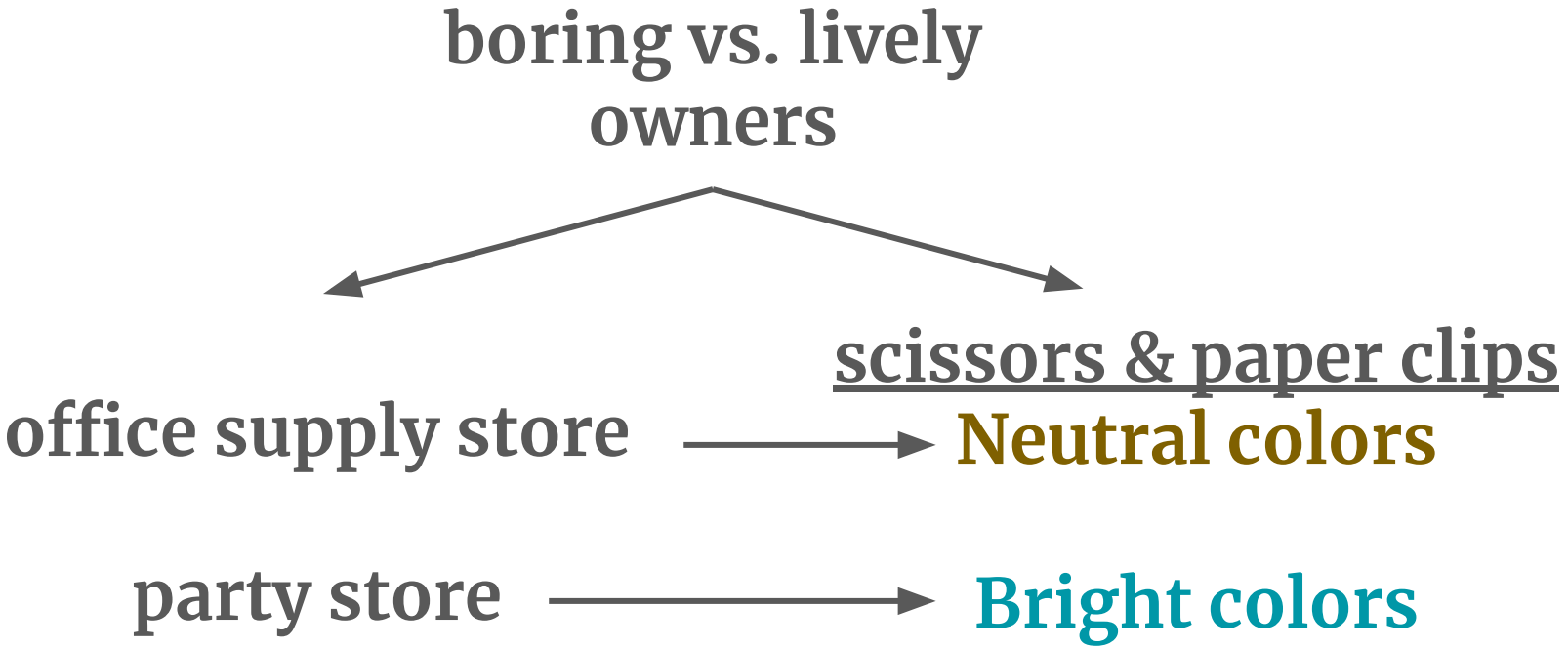
Confounding variables

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



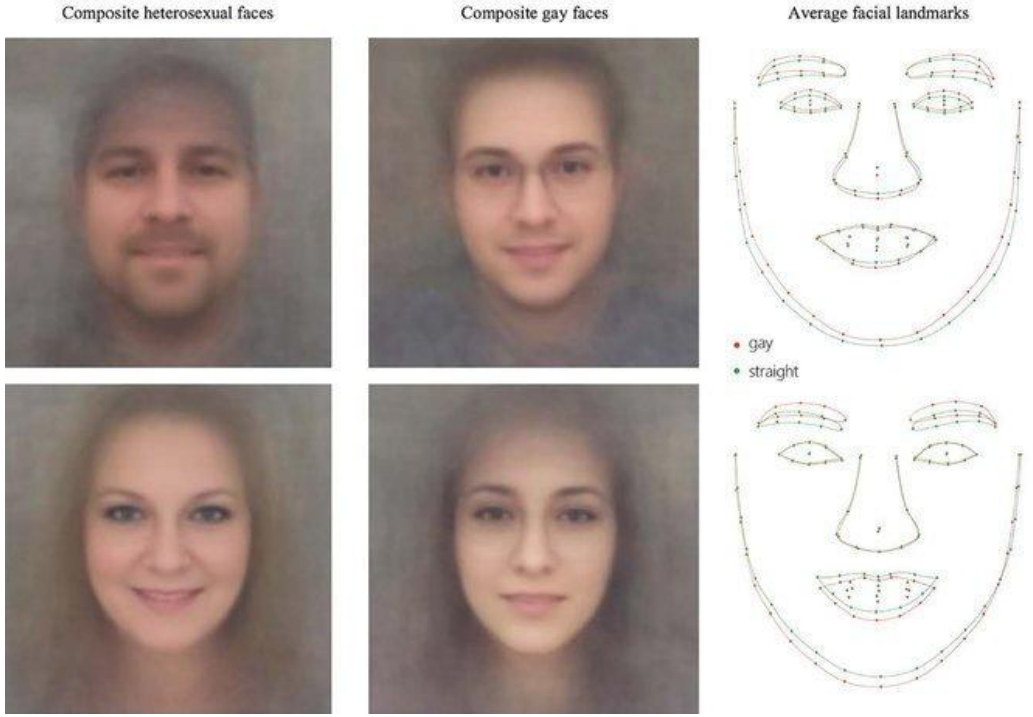
Constrained to what we can measure or observe

Decision space	Construct space	Observed space
Performance in college	Intelligence	IQ
Performance in college	Success in High School	GPA
Recividism	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

Friedler et al. 2016 “On the (Im)possibility of fairness”

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

Dwork et al. 2016 “Fairness Through Awareness”

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?

Dwork et al. 2016 “Fairness Through Awareness”

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

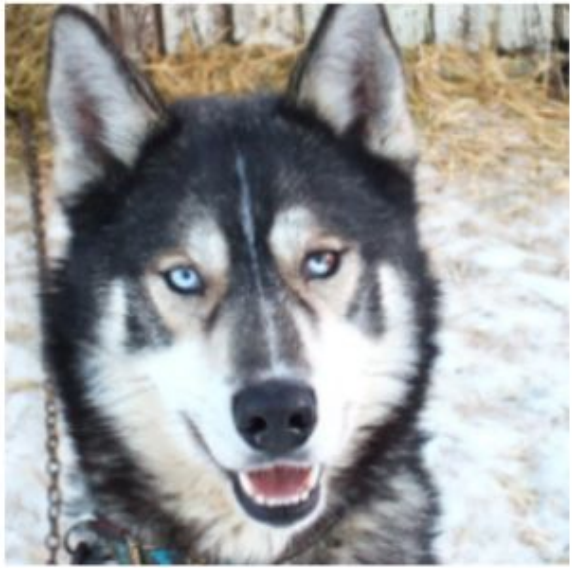
What do we do when our models are right for the wrong reasons?

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)

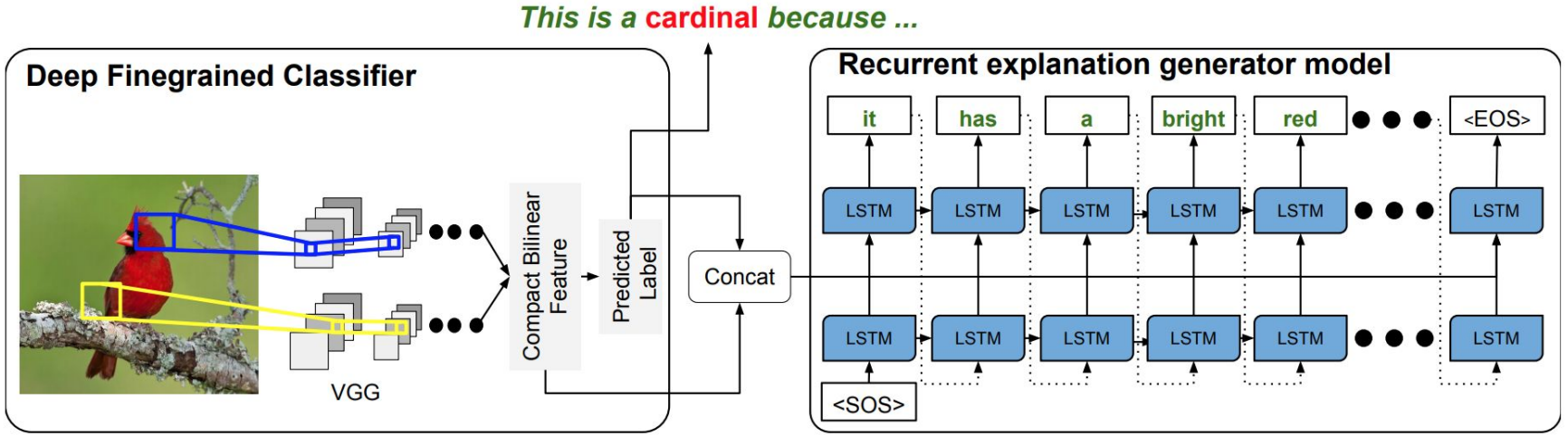
Text explanations

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 artic tern because this is a white bird with a black head and orange feet.

Explanations can be persuasive

Herman 2017

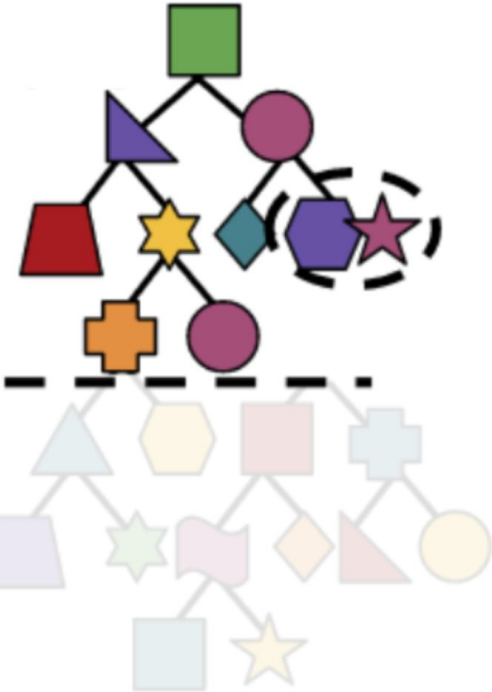
When tailoring our model explanations to human preferences and judges, our models may learn to prioritize persuasive explanations over introspective ones.

Splitting model form from simplicity

Herman 2017

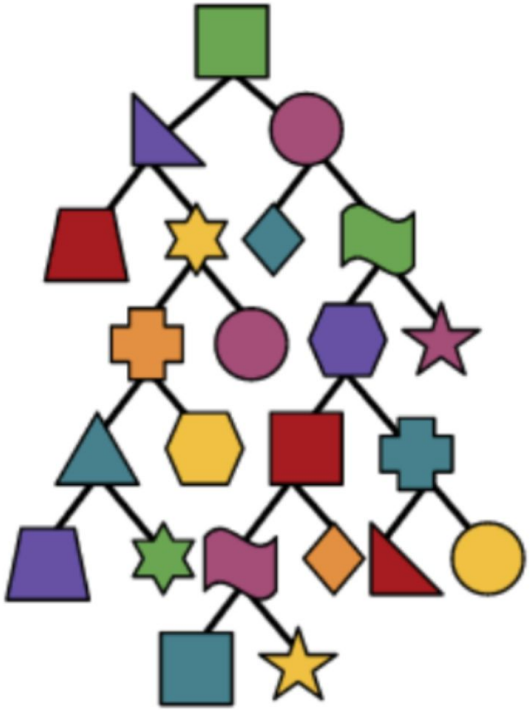
Simultaneously coerced into suitable **model form** (e.g., decision tree) and reduced in **complexity** (e.g., model size).

Difficult to evaluate across complexity preferences.

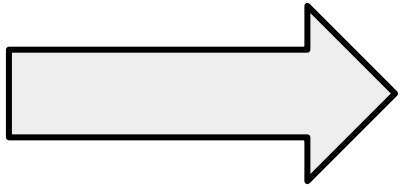


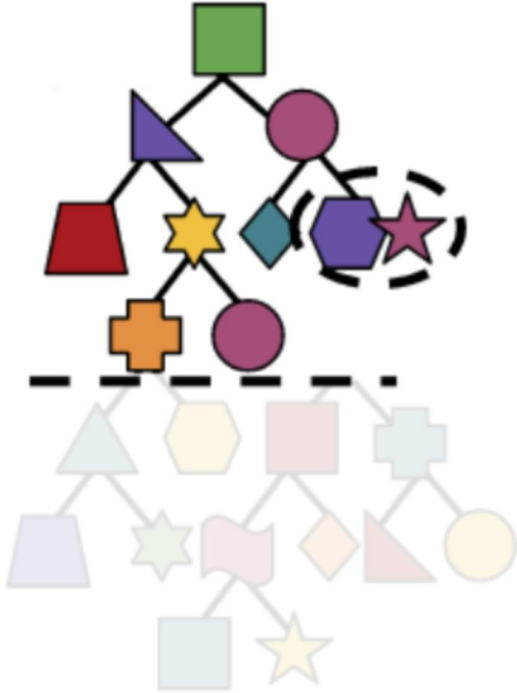
Splitting model form from simplicity

Herman 2017

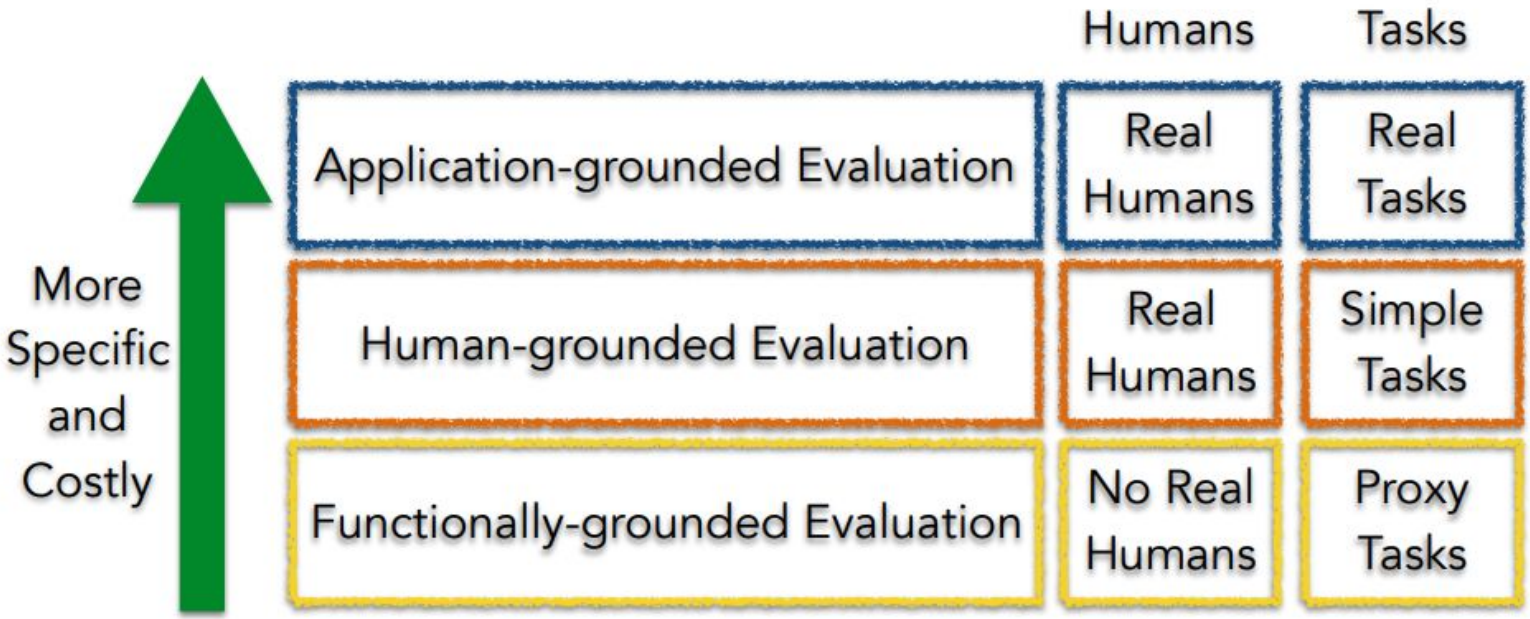


Keeps model form and reduction of complexity separate. Improves evaluation and adaptability.





Taxonomy of evaluation for interpretability applies broadly



Doshi-Valez and Kim 2017

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?

Social Pressure and Voter Turnout

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

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?

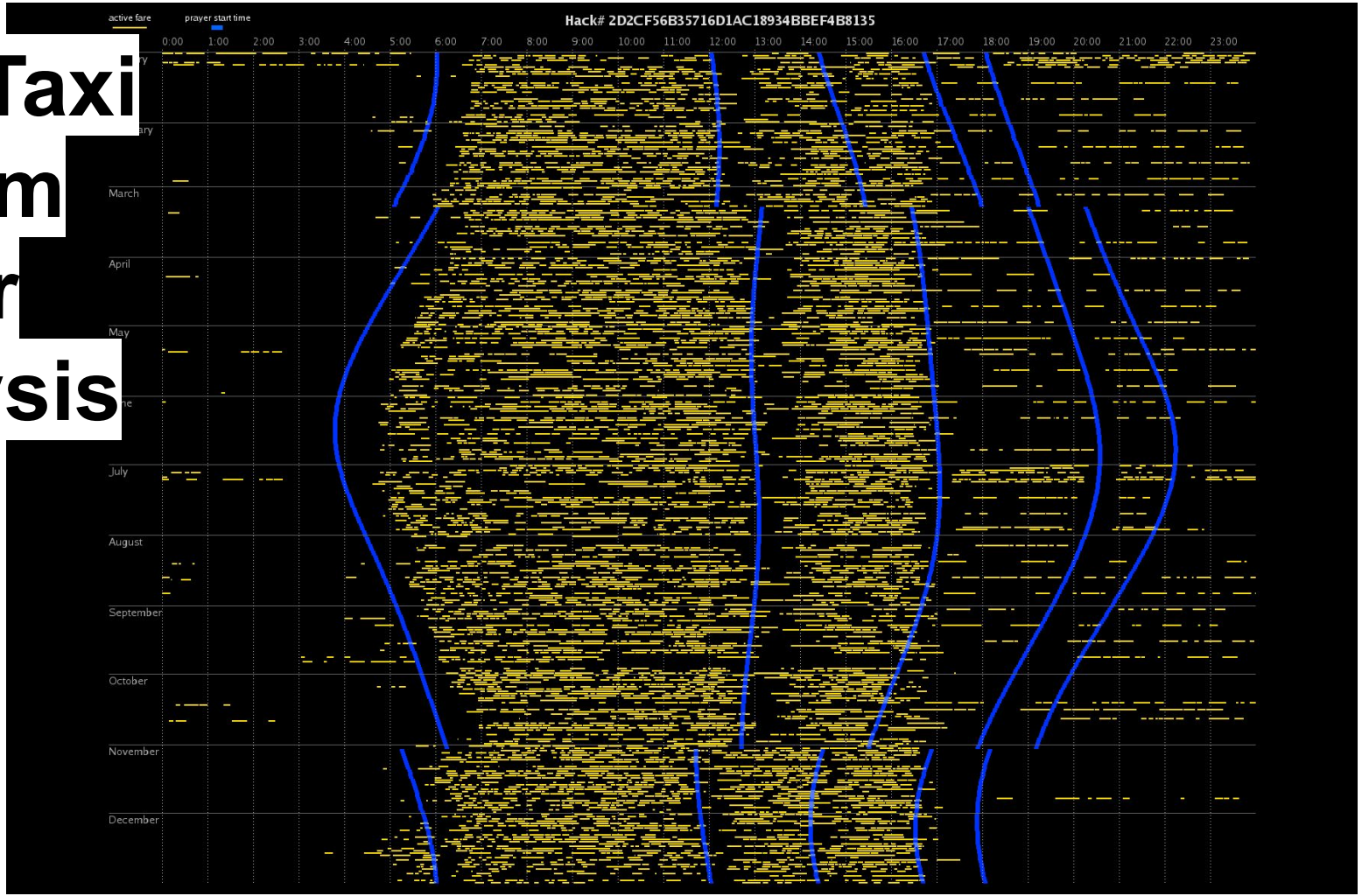
Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

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.

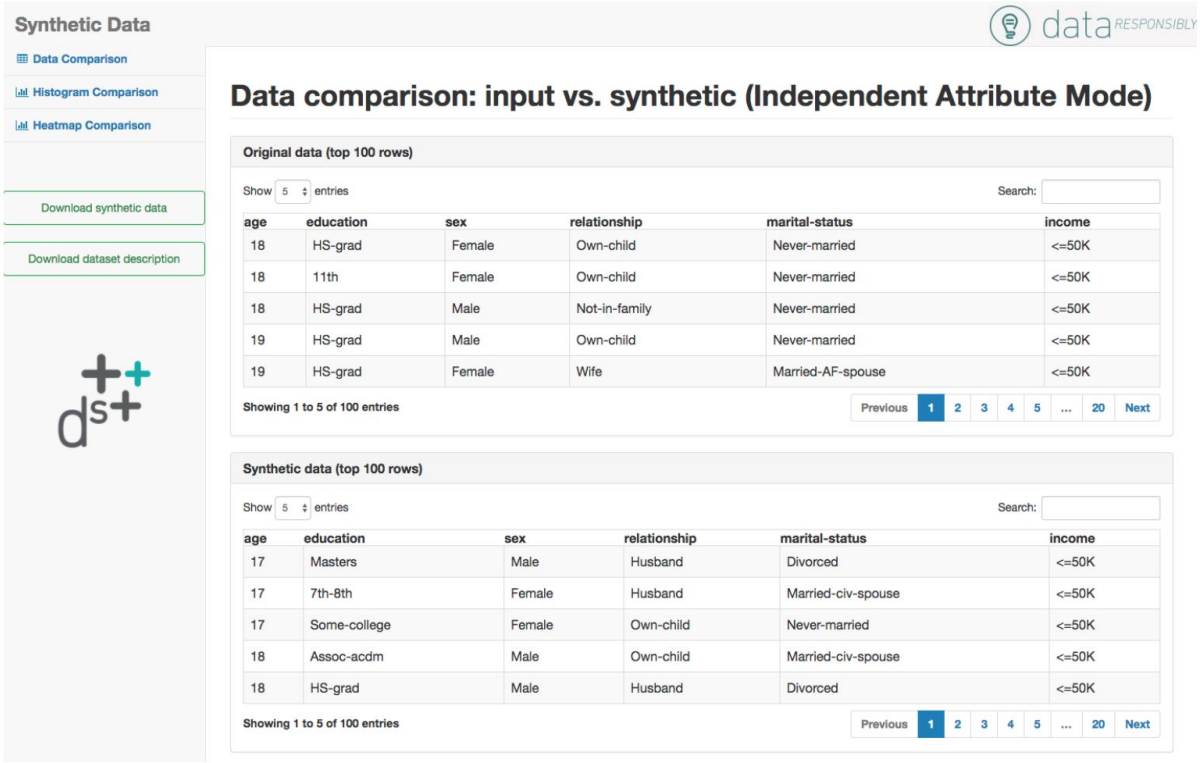
DO YOUR CIVIC DUTY — VOTE!

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	_____

NYC Taxi Muslim Driver Analysis



Synthetic datasets for privacy

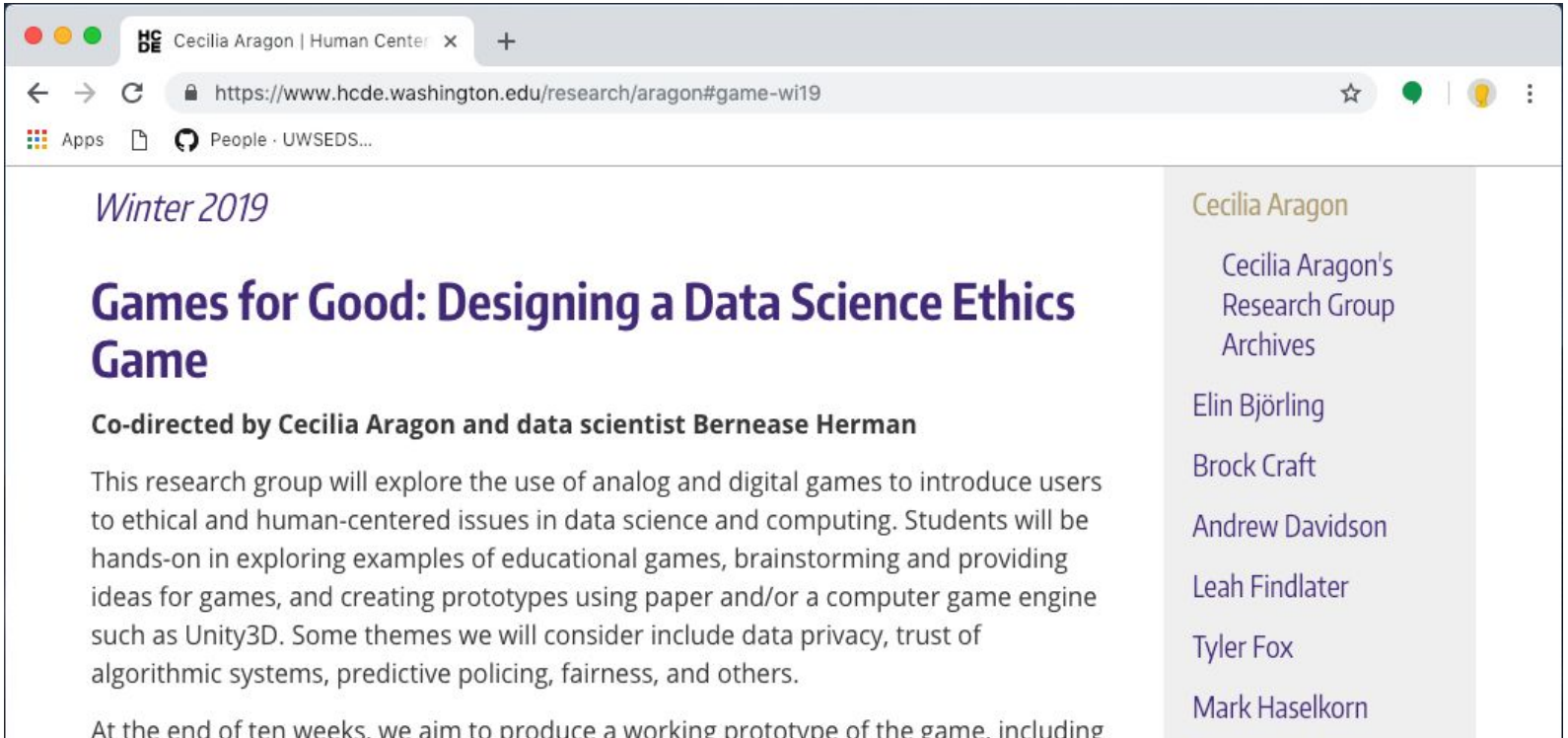


Bonus slides:

Other efforts toward ethical data science

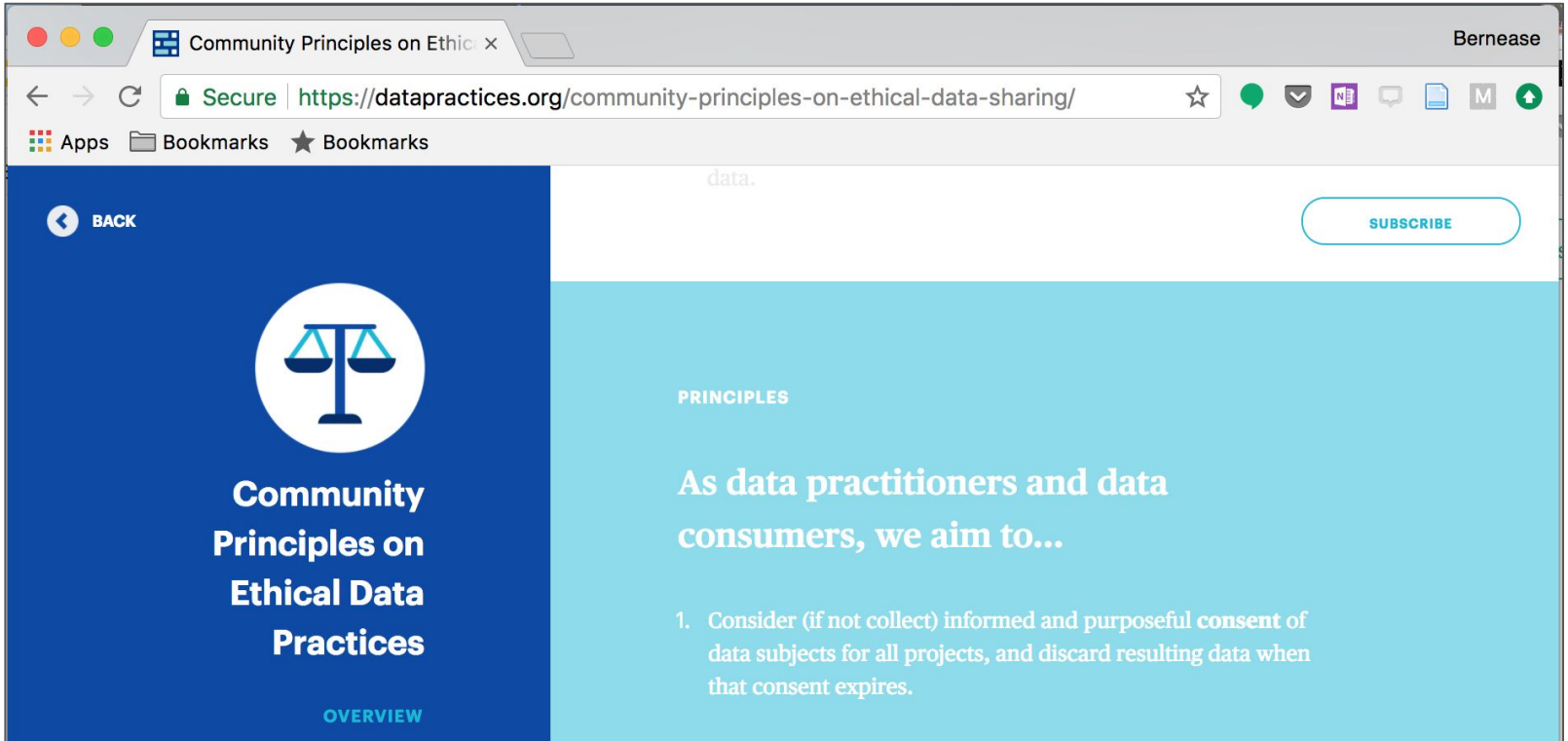
Educational Game on AI Ethics

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



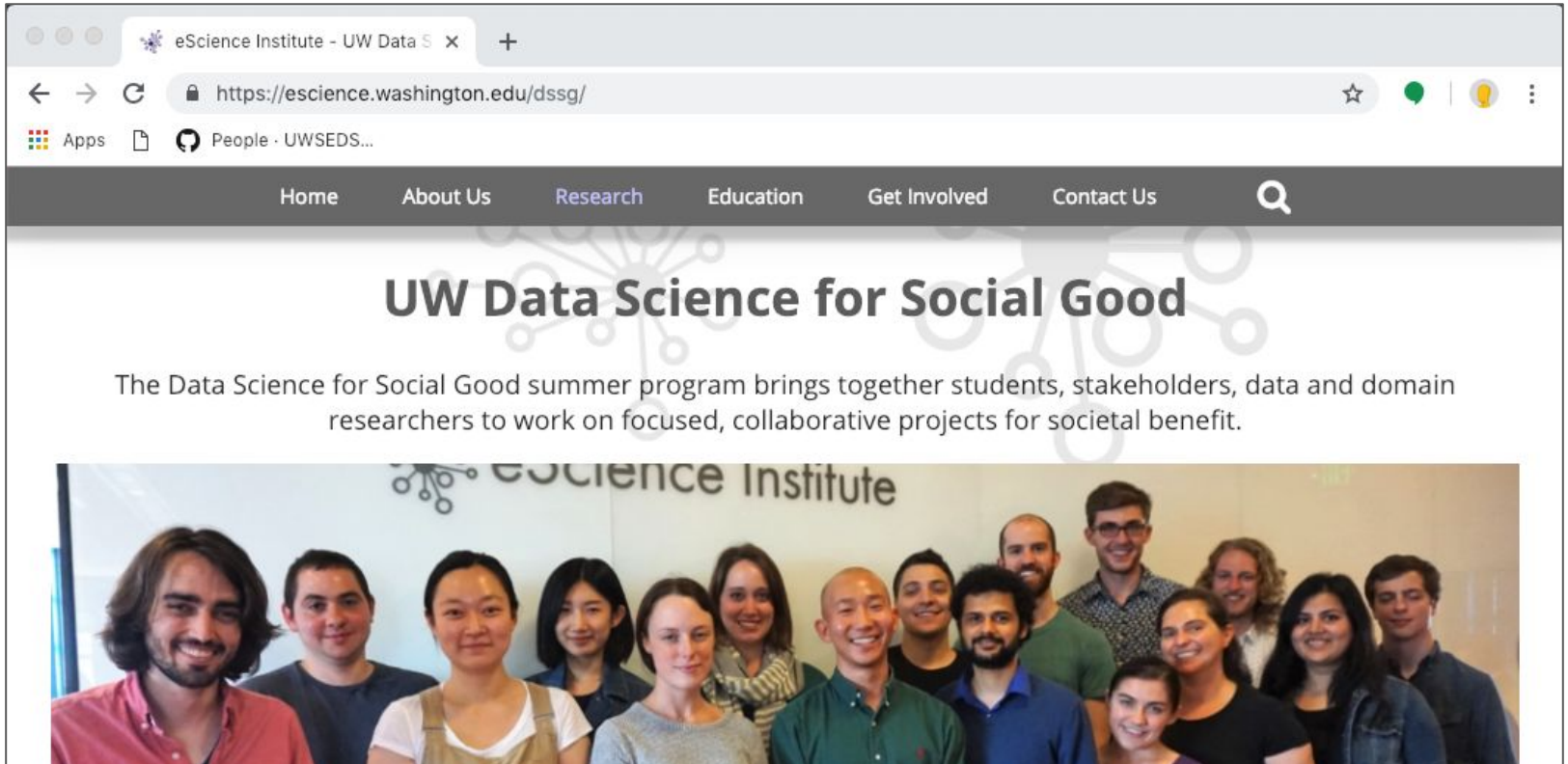
Codes of Ethics

Community Principles for Ethical Data Practices



Data Science for Social Good

Summer program at UW to help advance work in these areas



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