

CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021



HVERFORD
COLLEGE

Admin

- **Midterm 1** due Thursday at the beginning of class (take in a 2.5 hour block)
- **Thursday**: begin in-person classes with Prof. Farias
- **Lab on Thursday**: project meetings with all groups
 - okay if you've been focusing on the midterm instead – use the time to make progress on the project
- **Note-taker**: Rahul

Outline for November 16

- Biases in data collection
- Biases in data usage
- Issues that arise with algorithm choice
- Open time for midterm review Q&A

What does it mean to claim that algorithms are biased (or racist or political...)?

```
3 model = initialization(...)
4 n_epochs = ...
5 train_data = ...
6 for i in n_epochs:
7     train_data = shuffle(train_data)
8     X, y = split(train_data)
9     predictions = predict(X, model)
    error = calculate_error(y, predictions)
    model = update_model(model, error)
```

Pseudocode from [A Gentle Introduction to Mini-Batch Gradient Descent and How to Configure Batch Size](#)

Is machine learning fair by default?

“After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. ‘This program had absolutely nothing to do with race... but multi-variable equations,’ argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound.”

-Gilian Tett

Issues

- Target variable/intended use
- Subjective labels
- Proxy variables
- Feature Selection/Engineering
- Source of training data
- Transparency
- Validation

Challenges

Algorithms do not exist in a bubble

- Inherit the prejudices of their designers
- Reflect cultural biases
- Difficult to identify - can entrench/enhance issues
- Deny historically disadvantaged groups full participation

Outline for November 16

- Biases in data collection
- Biases in data usage
- Issues that arise with algorithm choice
- Open time for midterm review Q&A

“The Missing Diversity in Human Genetic Studies” (*Cell*, 2019)

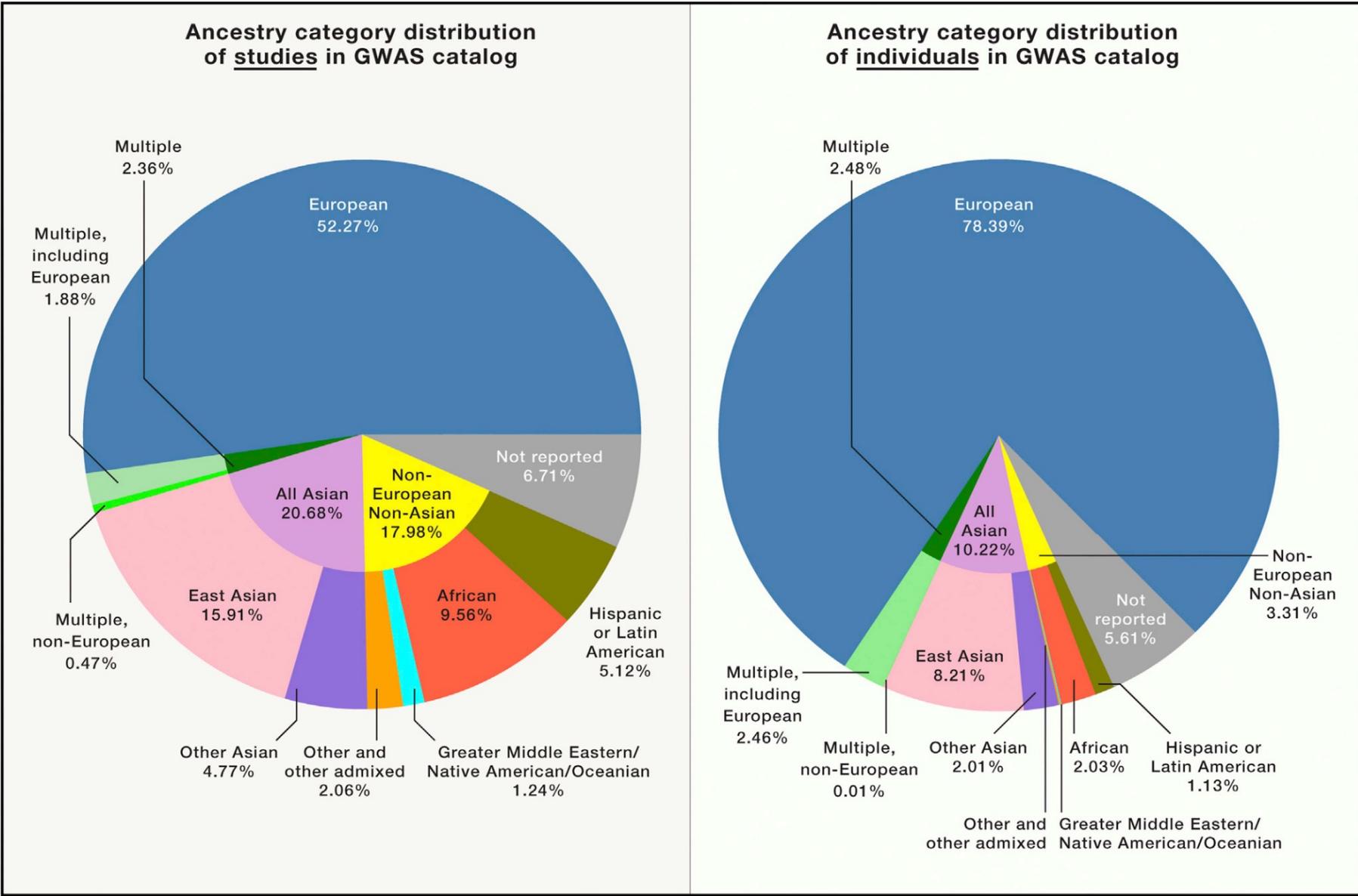
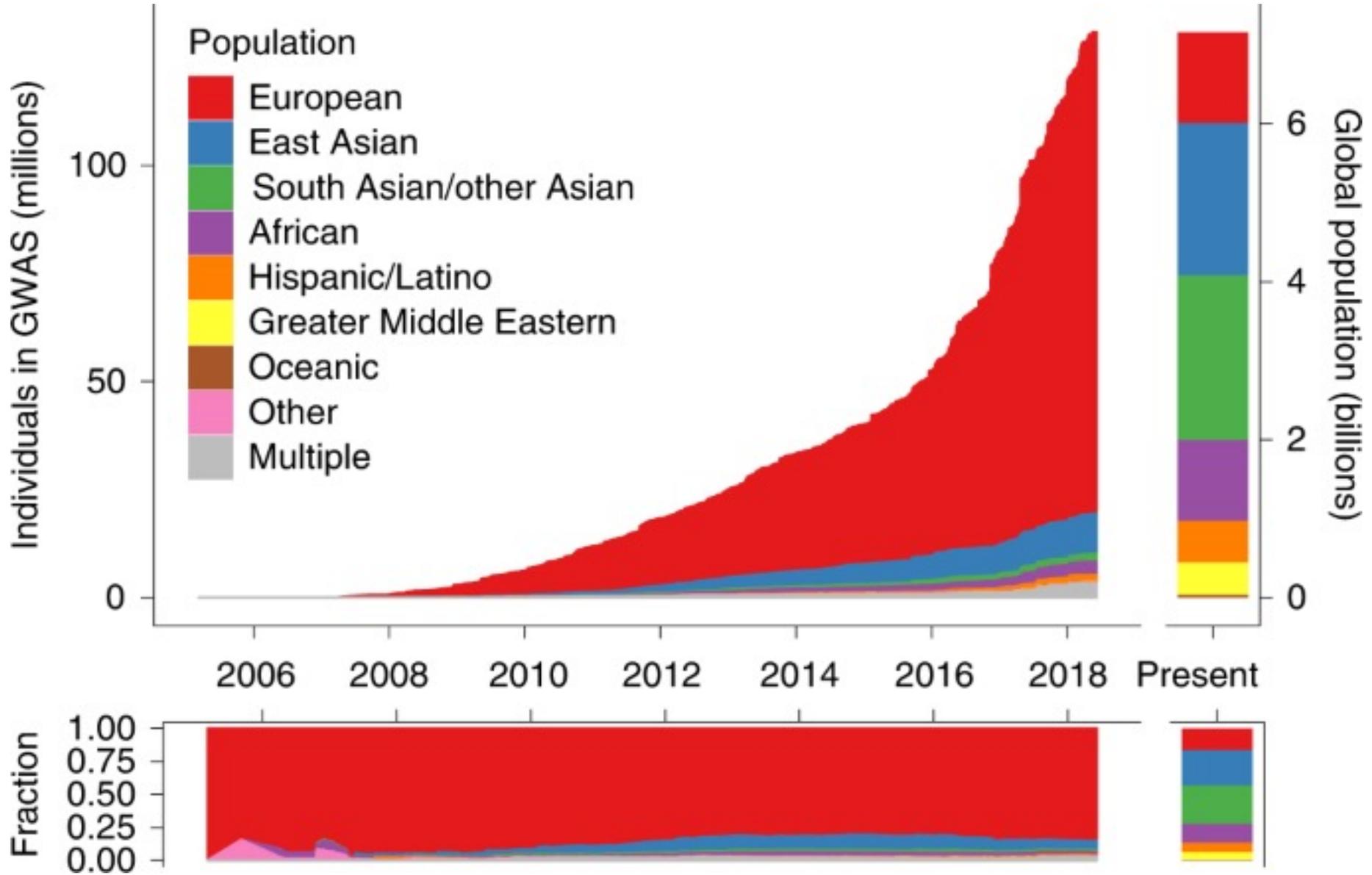


Figure 2. Summary of GWAS Studies by Ancestry for Studies in the GWAS Catalog through January 2019
 We show the distribution of ancestry categories in percentages included in GWAS (<https://www.ebi.ac.uk/gwas/home>) based on the study (left) and based on the total number of individuals (right).

“Clinical use of current polygenic risk scores may exacerbate health disparities”
(*Nature Genetics* 2019)



Example: job ads based on historical data

- Prestigious job ads automatically shown to men but not women
- Screenshot of Google image search for “CEO”



Example: Facial Recognition and Dataset Bias

1. HP Webcam

<https://www.youtube.com/watch?v=t4DT3tQqgRM>

2. Gender Shades

<https://www.youtube.com/watch?v=-ydGhdYd0M>



Example: cameras and webcams

- Many cameras and webcams have not been trained with racial and ancestral diversity in mind



Are Face-Detection Cameras Racist?

By Adam Rose | Friday, Jan. 22, 2010

<http://content.time.com/time/business/article/0,8599,1954643,00.html>

Example: loans and credit

- Housing loans (mortgages) given/denied automatically; correlate with neighborhoods and race
- Features that correlate with whether or not a user will pay back a loan:

- Borrower type of computer (Mac or PC).
- Type of device (phone, tablet, PC).
- Time of day you applied for credit (borrowing at 3am is not a good sign).
- Your email domain (Gmail is a better risk than Hotmail).
- Is your name part of your email (names are a good sign).

REPORT

Credit denial in the age of AI

Aaron Klein · Thursday, April 11, 2019

<https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/>

Example from: Suresh Venkatasubramanian

Outline for November 16

- Biases in data collection
- **Biases in data usage**
- Issues that arise with algorithm choice
- Open time for midterm review Q&A

LEFT OUT

UK biobank recruitment reflected diversity (in 2001; ref. 11). Analyses do not.

United Kingdom population (2001)

5.5%
Minority proportion



UK Biobank participants

5.4%
Minority proportion
(recruitment)



0.06%
Minority proportion
(analyses)

Outline for November 16

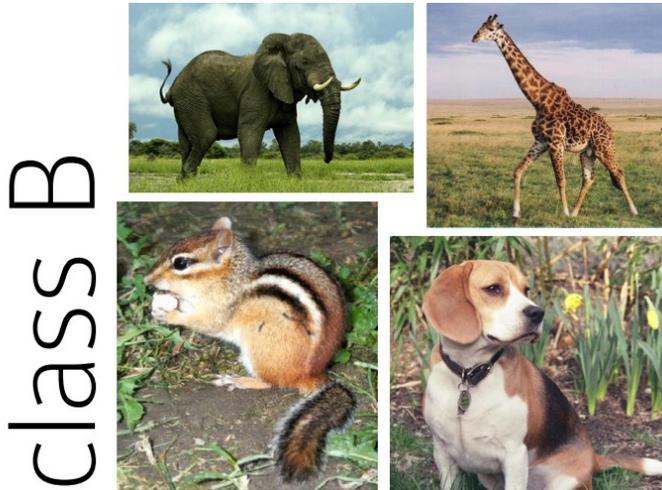
- Biases in data collection
- Biases in data usage
- **Issues that arise with algorithm choice**
- Open time for midterm review Q&A

Inductive Bias

Training Data



class A



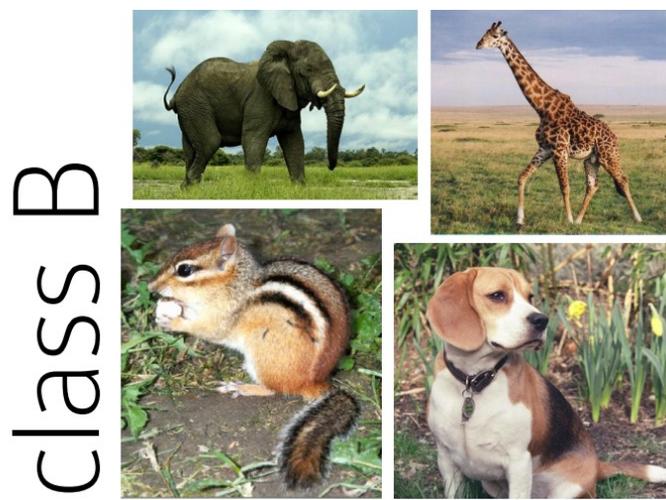
class B

Testing Data



Inductive Bias

Training Data



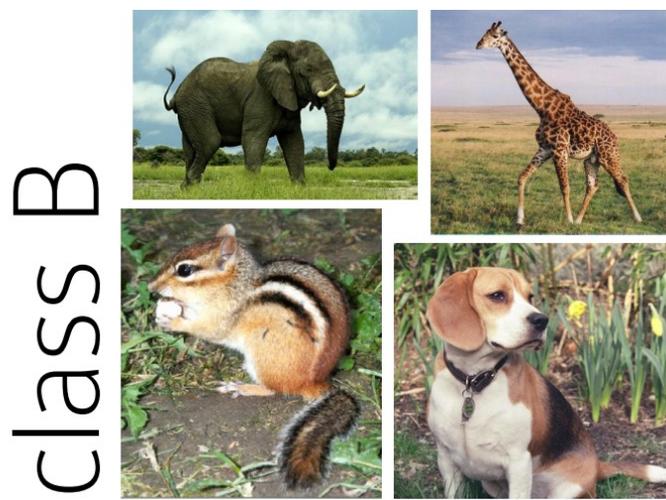
Testing Data



A: "fly"
B: "no fly"

Inductive Bias

Training Data

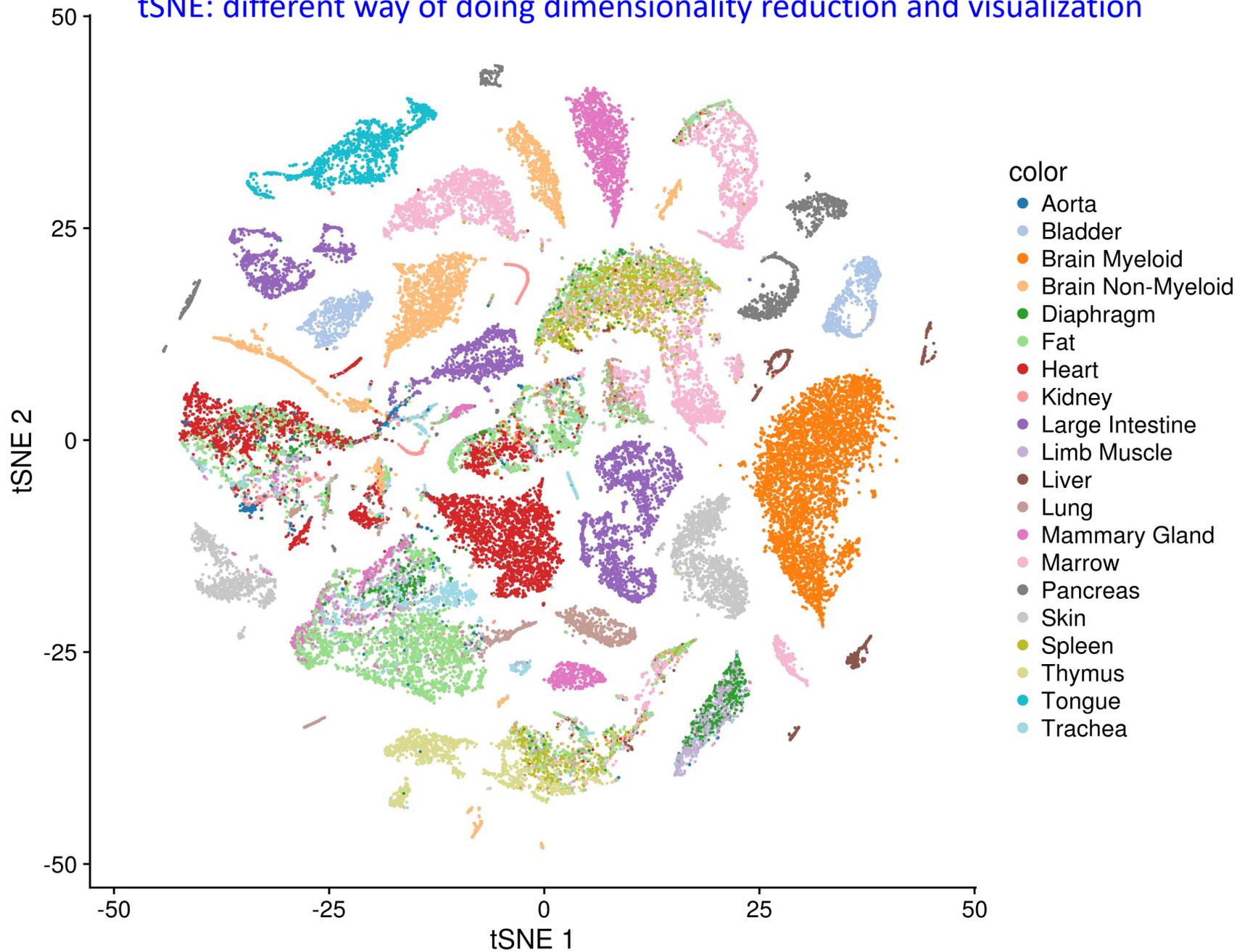


Testing Data

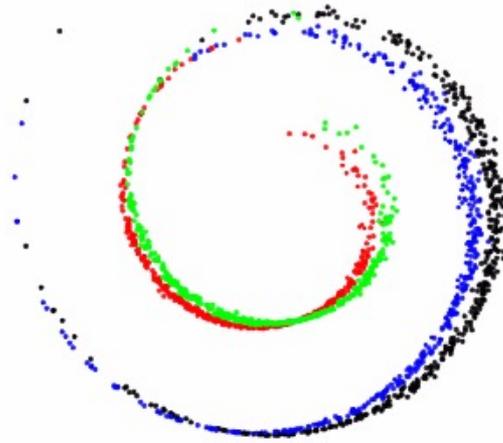


A: "bird"
B: "mammal"

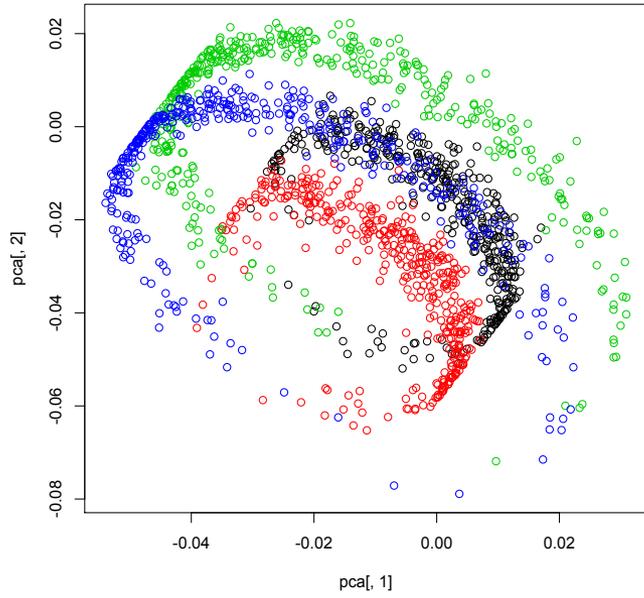
tSNE: different way of doing dimensionality reduction and visualization



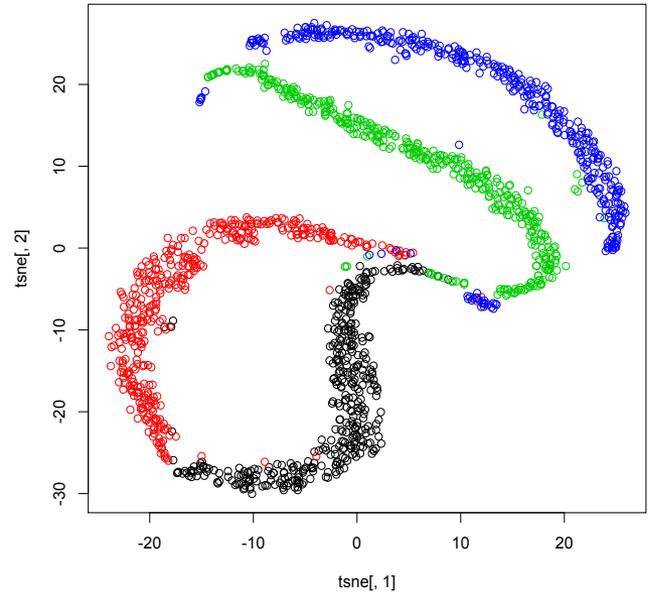
Original data



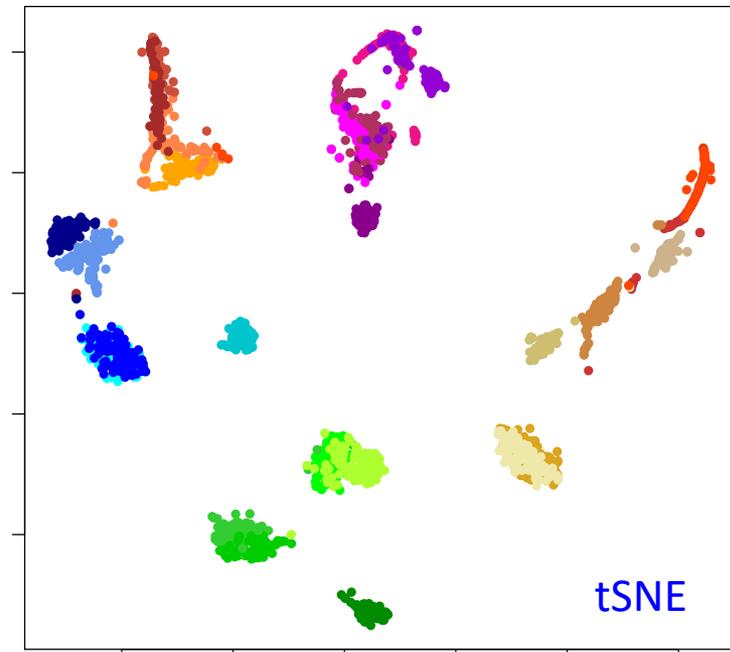
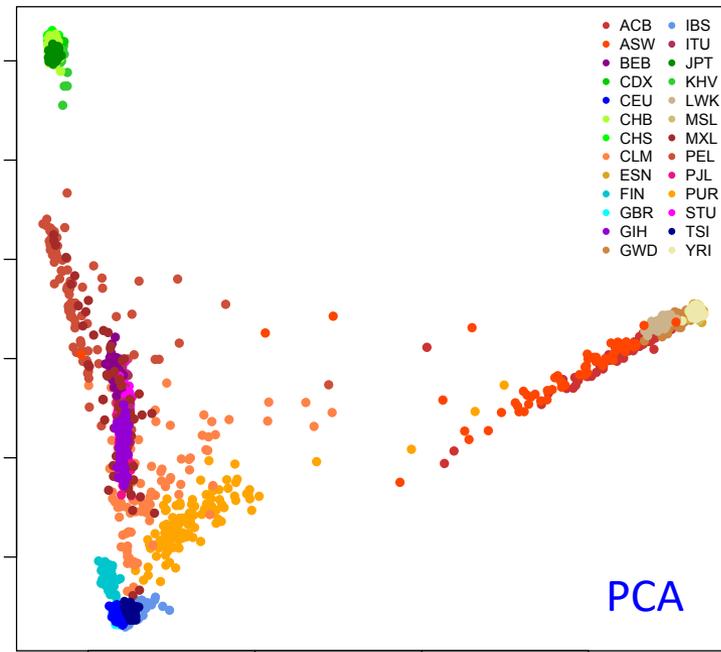
PCA



t-SNE

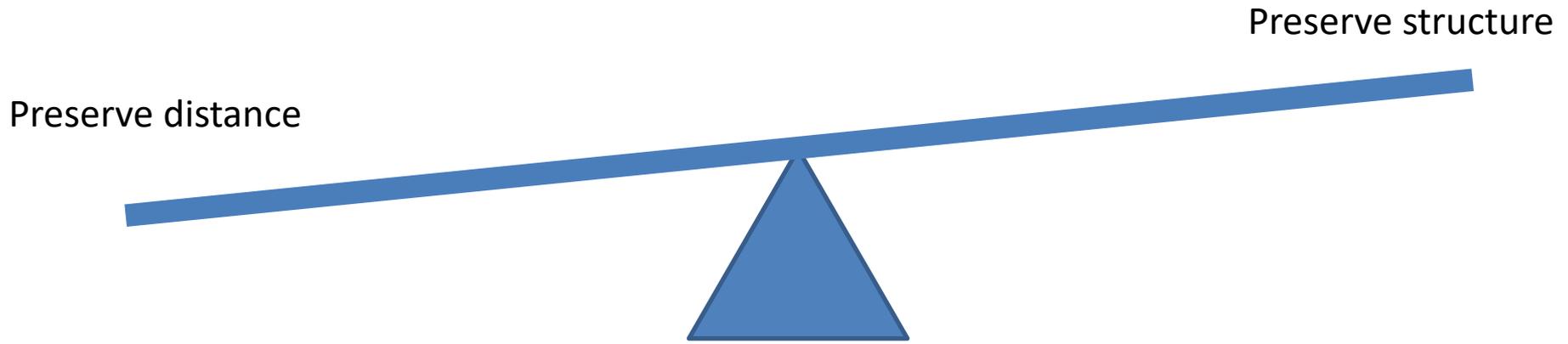


PCA vs. tSNE on genetic data



CHB	Han Chinese in Beijing, China
JPT	Japanese in Tokyo, Japan
CHS	Southern Han Chinese
CDX	Chinese Dai in Xishuangbanna, China
KHV	Kinh in Ho Chi Minh City, Vietnam
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
TSI	Toscani in Italia
FIN	Finnish in Finland
GBR	British in England and Scotland
IBS	Iberian Population in Spain
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
GWD	Gambian in Western Divisions in the Gambia

MSL	Mende in Sierra Leone
ESN	Esan in Nigeria
ASW	Americans of African Ancestry in SW USA
ACB	African Caribbeans in Barbados
MXL	Mexican Ancestry from Los Angeles USA
PUR	Puerto Ricans from Puerto Rico
CLM	Colombians from Medellin, Colombia
PEL	Peruvians from Lima, Peru
GIH	Gujarati Indian from Houston, Texas
PJL	Punjabi from Lahore, Pakistan
BEB	Bengali from Bangladesh
STU	Sri Lankan Tamil from the UK
ITU	Indian Telugu from the UK



How to visualize data always depends on the data, and the question

There is rarely if ever a single correct approach

Example: Propublica, *Machine Bias*

“Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision and case management of offenders. COMPAS was developed empirically with a focus on predictors known to affect recidivism. It includes dynamic risk factors, and it provides information on a **variety of well validated risk and needs factors designed to aid in correctional intervention to decrease the likelihood that offenders will reoffend.**”

- Northpointe COMPAS Practitioner's Guide

- An algorithm to assess potential recidivism risk. Risk scales for general recidivism, violent recidivism, and pretrial misconduct.
- Input: answers to 137 questions by defendant (or taken from records) uses such factors such as poverty, joblessness and other variables
- Northpointe reports that accuracy is high and equal across race (~68%)

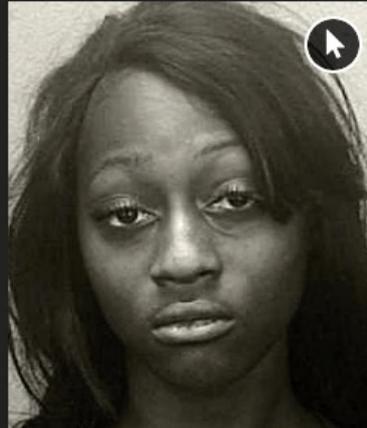
Example: Propublica, *Machine Bias*

Two Drug Possession Arrests

	
DYLAN FUGETT	BERNARD PARKER
LOW RISK 3	HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Example: Propublica, *Machine Bias*

Prediction Fails Differently for Black Defendants

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

White

African American

Example: Propublica, Machine Bias

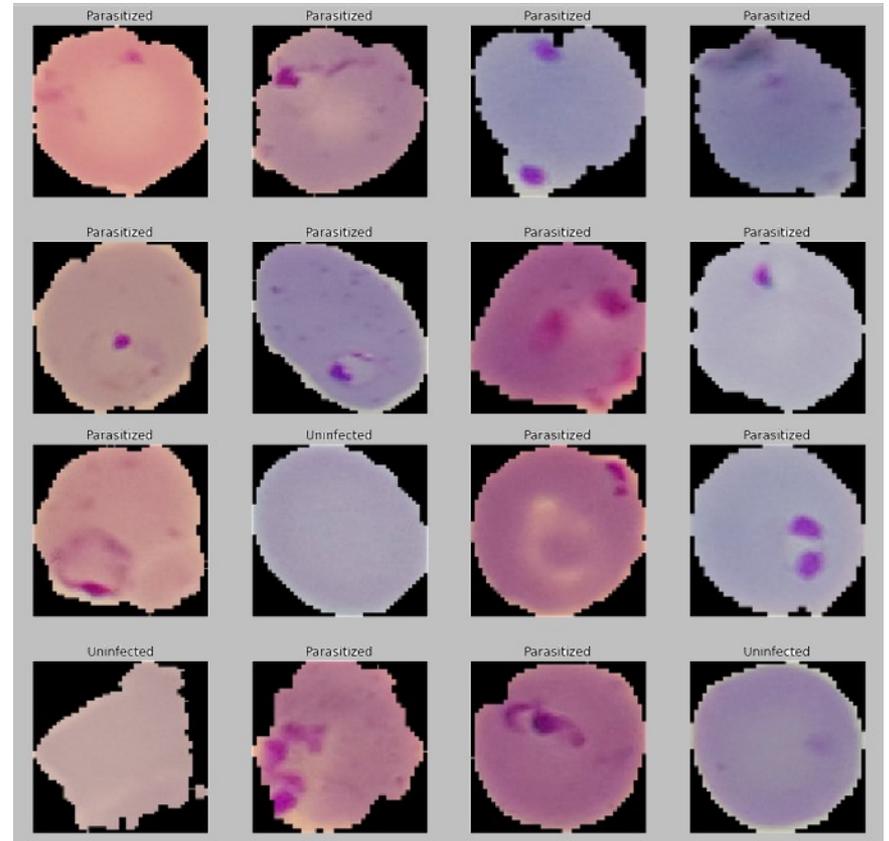
Prediction Fails Differently for Black Defendants

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

		White <u>pred</u>		African American <u>pred</u>	
		low	high	low	high
<u>true</u>	low		23.5 → 100%		44.9
	high	47.7			28.0

Are there any beneficial examples of Machine Learning?

- Medical scans
- Machine translation
- ML for accessibility
 - Speech to text
 - Image captions
 - Website navigation
- Neural networks in evolution



<https://lhncbc.nlm.nih.gov/publication/pub9932> (NIH malaria dataset)

Generative Approaches

- In population genetics we often need high-quality simulated data for validation (not many “labels” in evolution)
- Simulation parameters used for European populations were historically used for other groups
- Use of GANs (generative adversarial networks) can create accurate simulated data for any population

Admissions at Haverford

- Haverford has suddenly started receiving 10x more applications than usual
- You are tasked with creating an algorithm to determine whether or not an applicant should be admitted
- Questions:
 - How would you encode features?
 - How would you use past admission data to train?
 - What cost function are you trying to optimize?

Outline for November 16

- Biases in data collection
- Biases in data usage
- Issues that arise with algorithm choice
- **Open time for midterm review Q&A**