# CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

# Admin

- Office hours today: **12:30-1:30** (in my office)
- Email me ASAP if you want to present on the last day of classes (photographer visiting)
- In lab Thurs: **project meetings** with all groups
- Project presentations: **Wed, Dec 18, 1-4pm**

# Outline for December 10

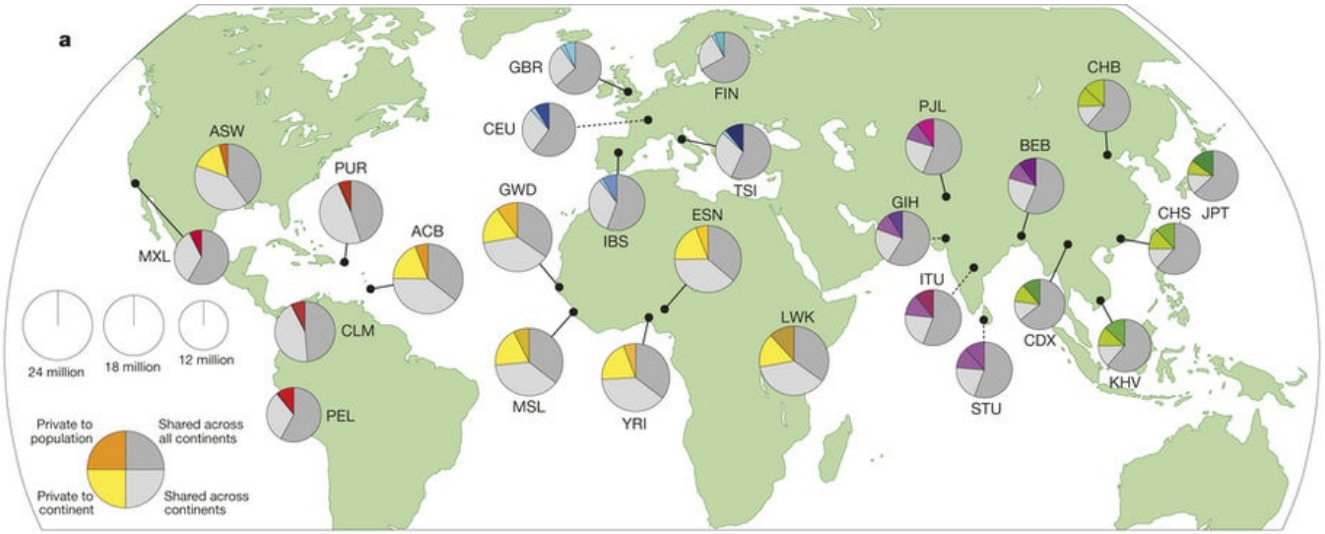
- Finish: Principal Component Analysis (PCA)
- Introduction to bias in ML
- Hand back Midterm 2, common issues

# Outline for December 10

- Finish: Principal Component Analysis (PCA)
- Introduction to bias in ML
- Hand back Midterm 2, common issues

# The 1000 Genomes project

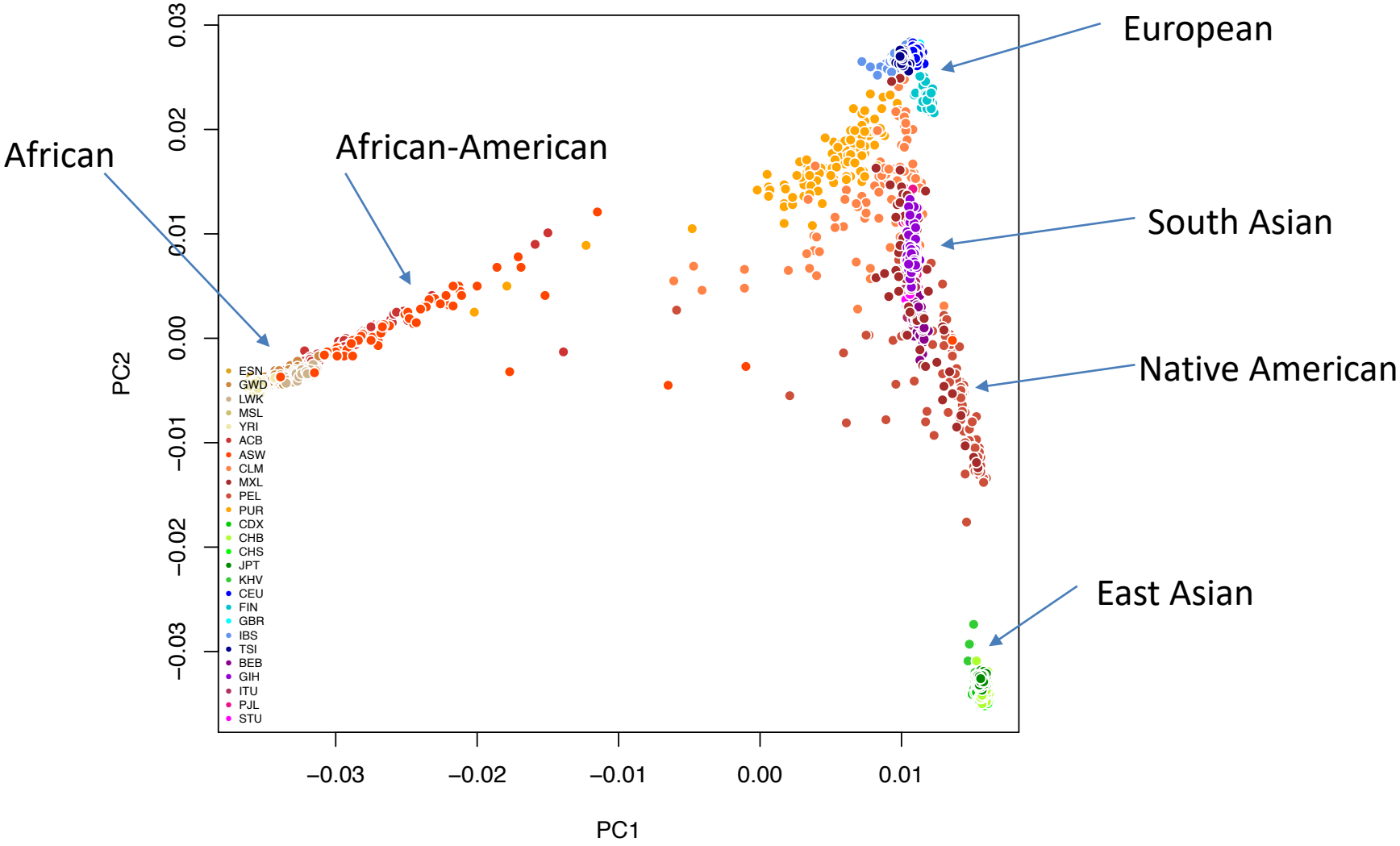
- Whole-genome sequence data from 2504 individuals from 26 populations
- A catalog of human genetic variation, useful as a reference or imputation panel
- Completely public. Download from ftp://ftp-trace.ncbi.nih.gov/1000genomes/




The 1000 Genomes Project Consortium, 2015; https://www.nature.com/articles/nature15393

*Slide: Iain Mathieson*

# Global population structure

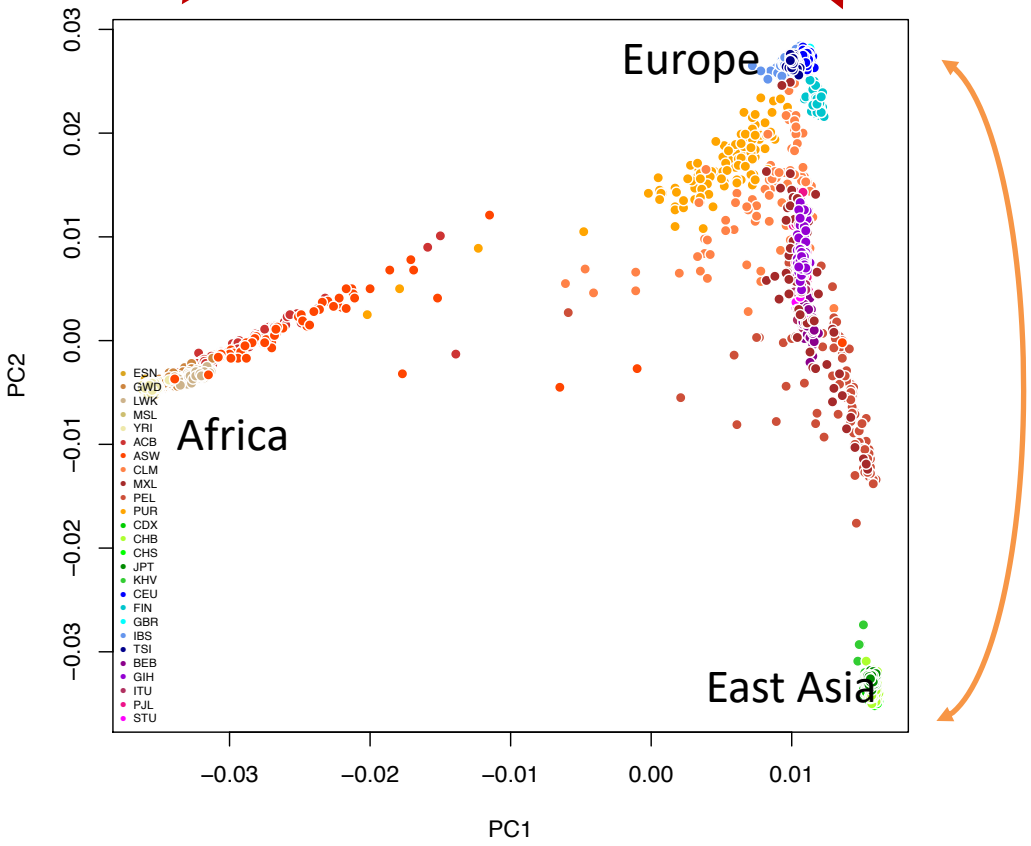


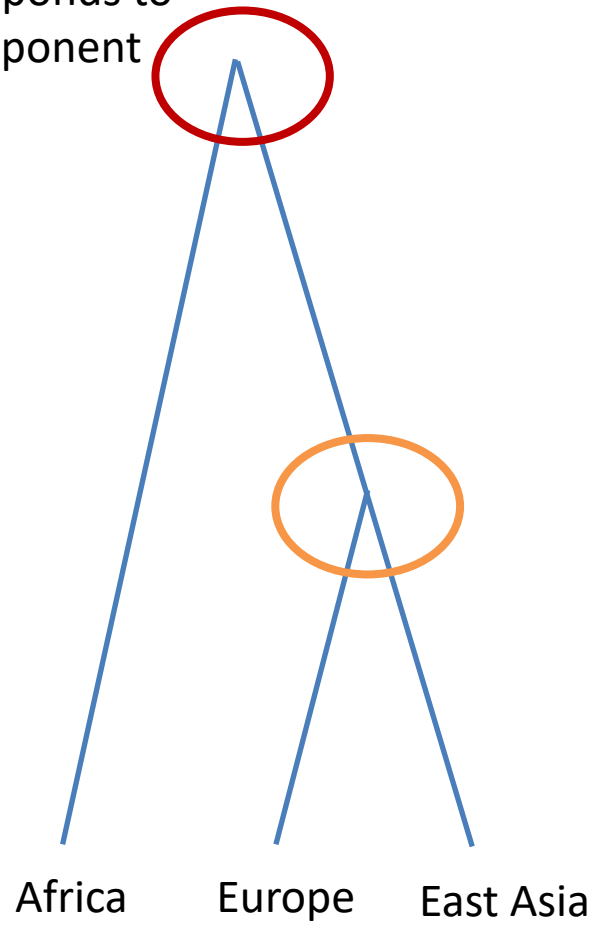

*Slide: Iain Mathieson*

# What causes these patterns?

## Populations splits separate populations

Oldest split corresponds to first principal component

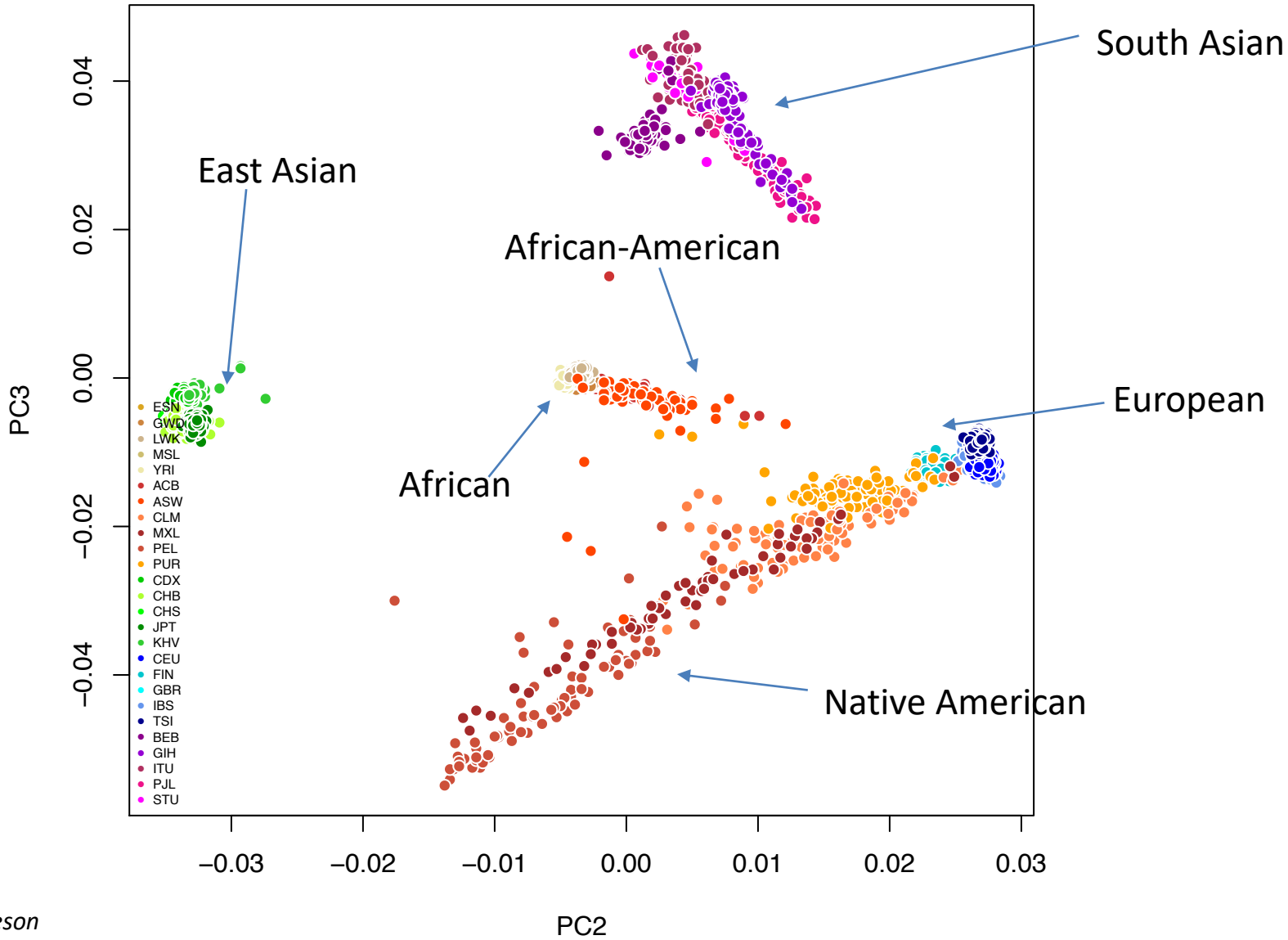



*Slide: Iain Mathieson*

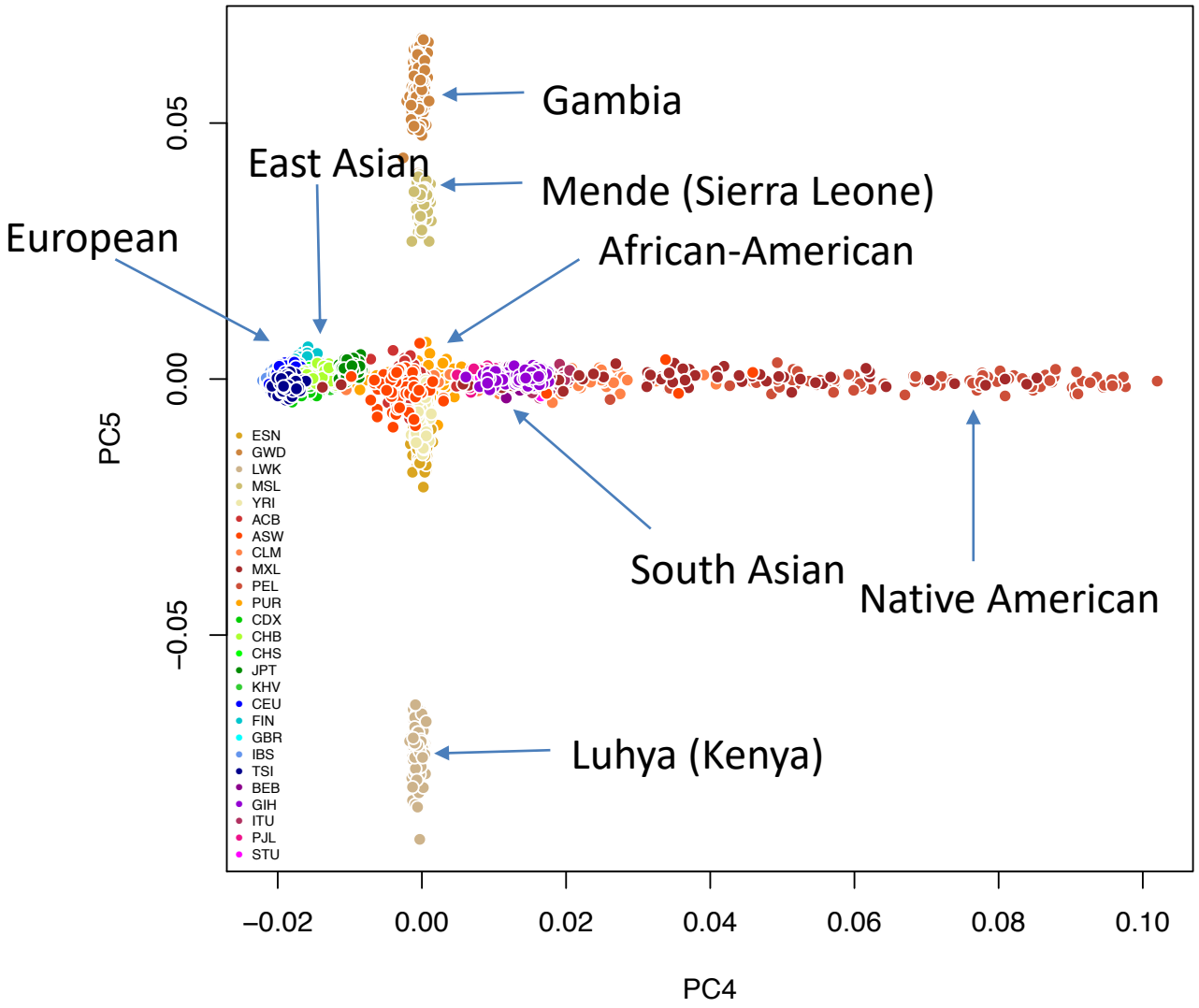
McVean 2009: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686

# Global population structure




*Slide: Iain Mathieson*

# Global population structure



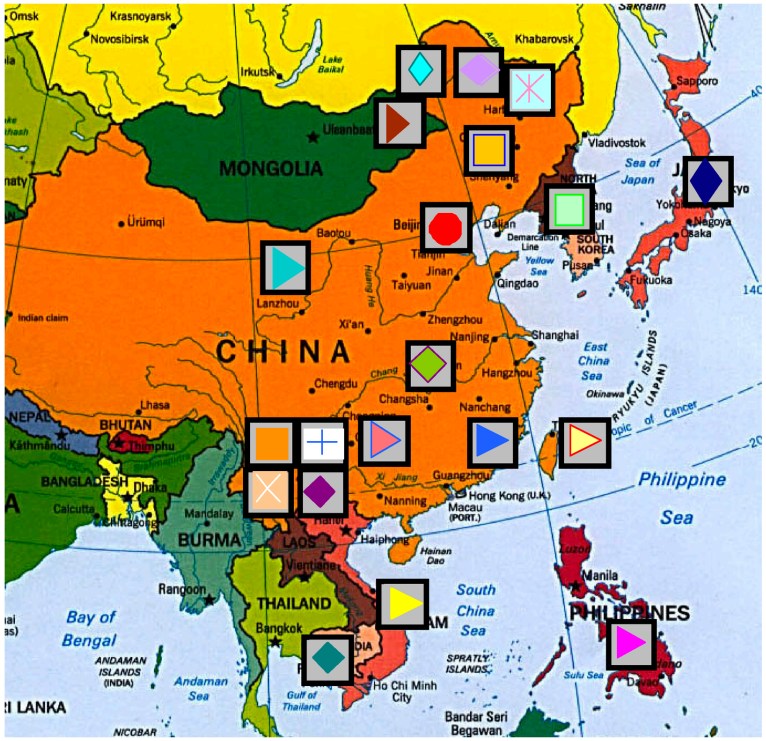

*Slide: Iain Mathieson*

# Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays

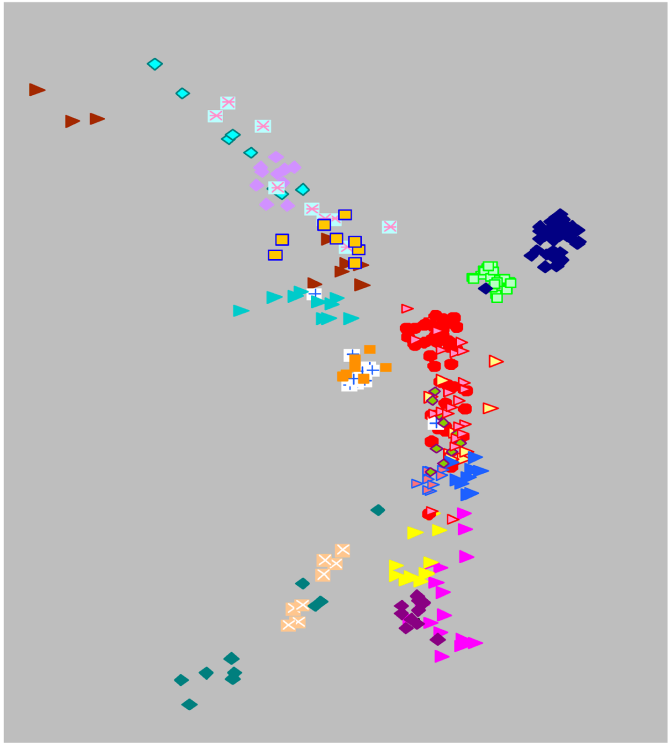
Chao Tian, Roman Kosoy, Annette Lee, Michael Ransom, John W. Belmont, Peter K. Gregersen, Michael F. Seldin

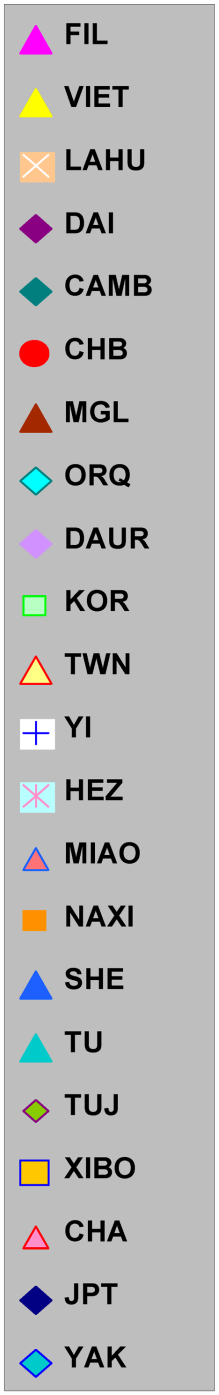
Published: December 5, 2008 • https://doi.org/10.1371/journal.pone.0003862

C




D





Handout 22

eigenvalues $\begin{cases} \lambda_1 = \frac{3}{5} \\ \lambda_2 = 0 \end{cases}$

high $\Rightarrow$ PC1

low $\Rightarrow$ PC2

$\Rightarrow W = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}$

$p \times r$ ($r$: new dim)

**eigenvectors:**

$$(A - \lambda \cdot I)\vec{v} = \vec{0}$$

$$\left( \begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix} - \begin{bmatrix} 3/5 & 0 \\ 0 & 3/5 \end{bmatrix} \right) \begin{bmatrix} v_{1a} \\ v_{1b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

($\vec{v}_1$)

$\det(\cdots) = 0$

$T = XW$

$n \times r$ $\quad$ $n \times p$ $\quad$ $p \times r$

$$\begin{bmatrix} -3/10 & -3/10 \\ -3/10 & -3/10 \end{bmatrix} \begin{bmatrix} v_{1a} \\ v_{1b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-\frac{3}{10} v_{1a} = \frac{3}{10} v_{1b}$$

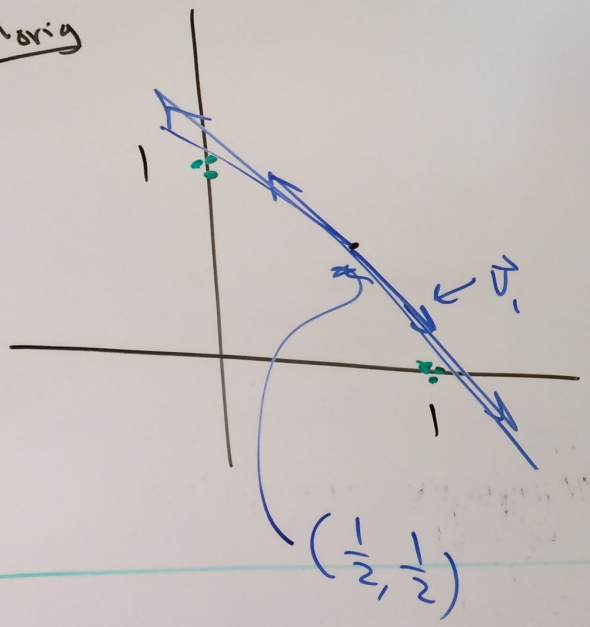
$$\Rightarrow \vec{v}_1 = \underbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}_{P(1}$$

$$\Rightarrow \vec{v}_2 = \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{P(2}$$

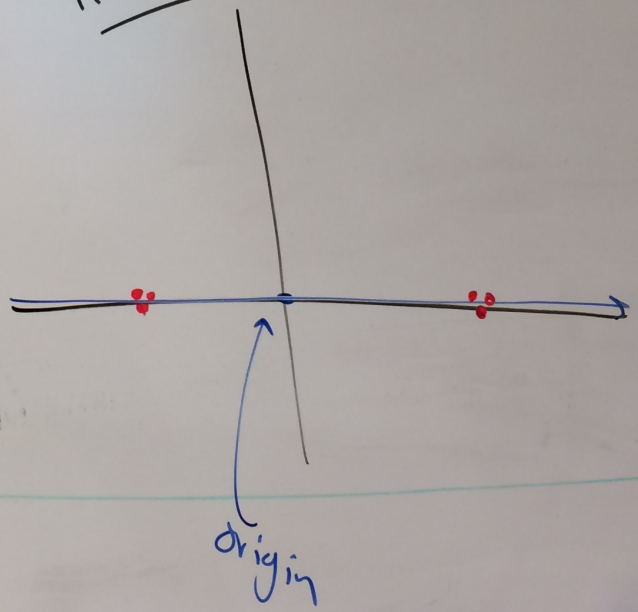
$$T = XW = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \begin{matrix} \vec{v}_1 & \vec{v}_2 \end{matrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

no more variation.

$X_{orig}$




transformed



# Outline for December 10

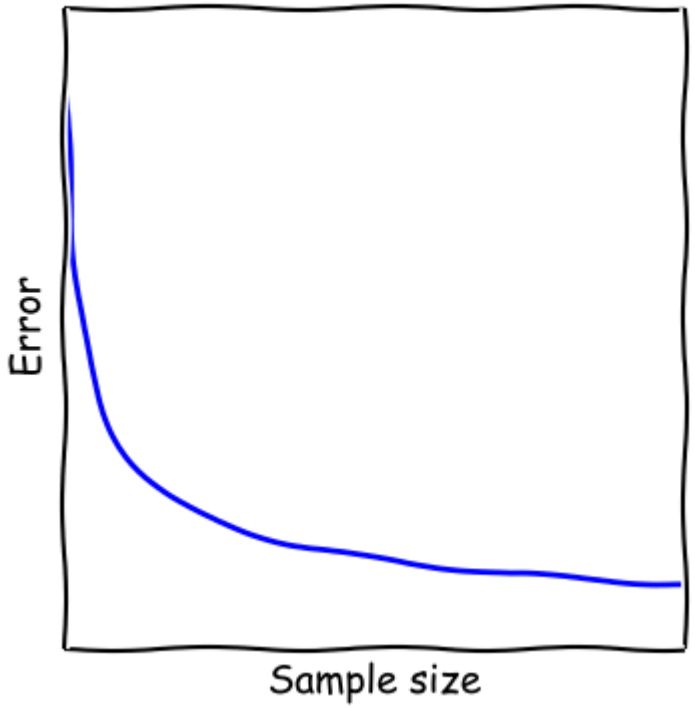
- Finish: Principal Component Analysis (PCA)
- Introduction to bias in ML
- Hand back Midterm 2, common issues

# “How big data is unfair” (takeaways)


- ML is not fair by default, even though it relies on “neutral” multi-variable equations
- If training data reflects social biases, algorithm will likely incorporate them
- “Protected” attributes (race, gender, religion, sexual orientation, etc) often redundantly encoded

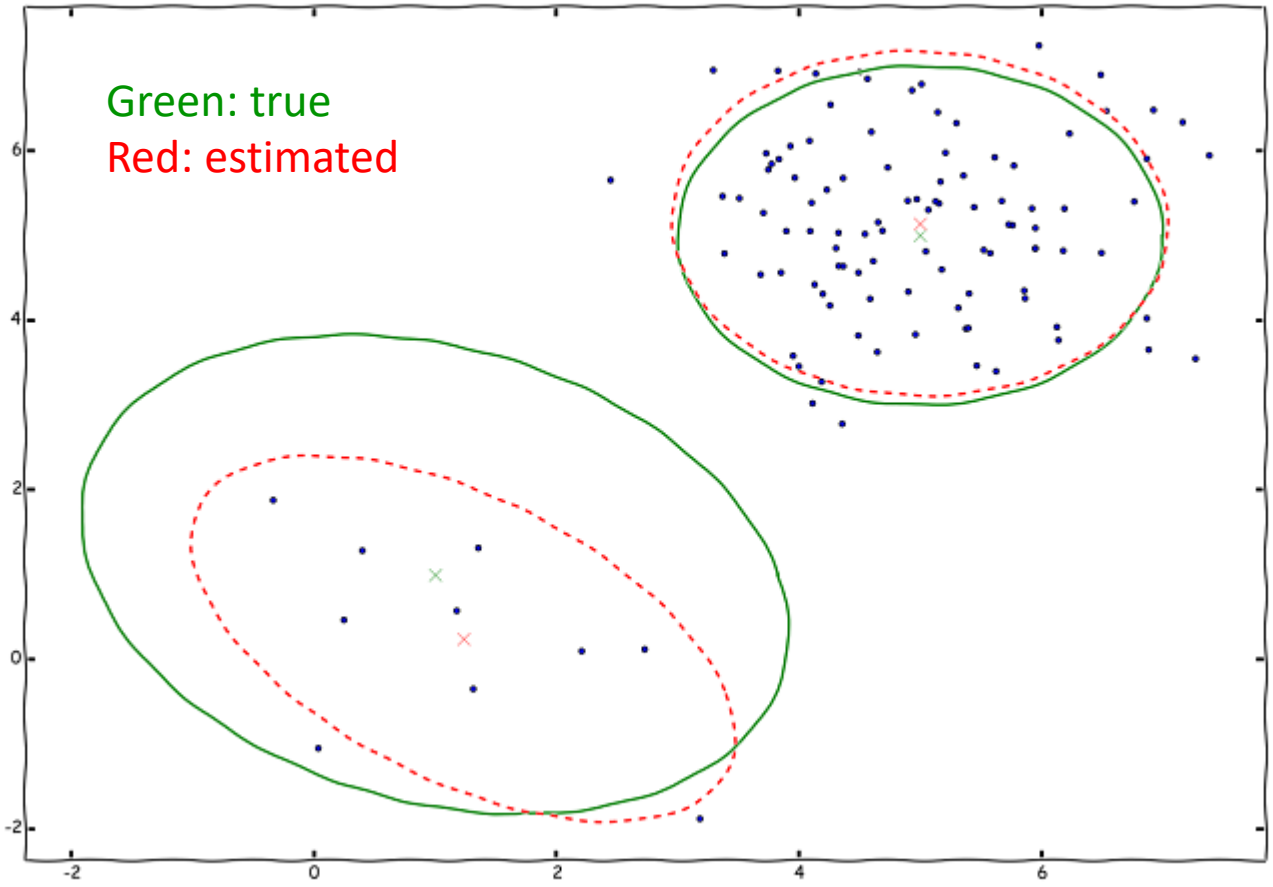
# Sample size disparity

- More data from majority will make results more accurate for that group
- Less accurate for the minority



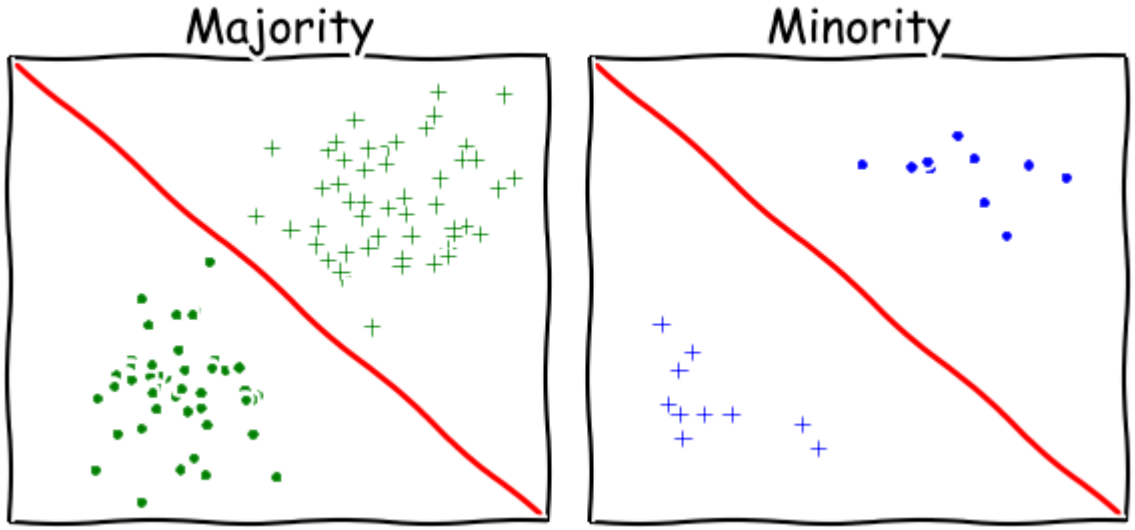

"The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate."
Image: Moritz Hardt

# Sample size disparity




"Modeling a heterogeneous population as a gaussian mixture and learning its parameters using the EM algorithm. As expected, the estimates for the smaller group are significantly worse than for the larger. Dashed red ellipsoids describe the estimated covariance matrices. Solid green defines the correct covariance matrices. The green and red crosses indicate correct and estimated means, respectively." Image: Moritz Hardt

# Cultural Differences



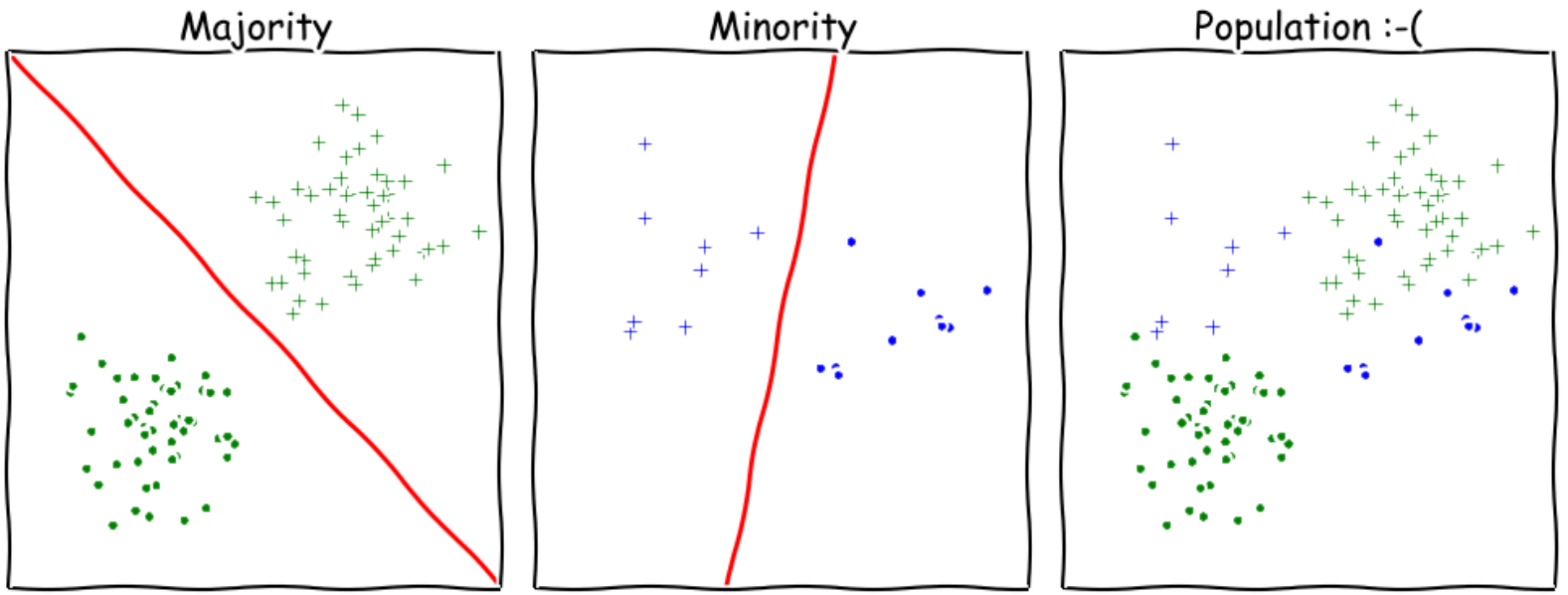

"Positively labeled examples are on opposite sides of the classifier for the two groups."
Image: Moritz Hardt

Goal: determine if a user profile (on Facebook, Twitter, etc) is genuine

* positive: real profile
* negative: fake profile

Feature: length of name

# Undesired complexity




“Even if two groups of the population admit simple classifiers, the whole population may not.” Image: Moritz Hardt

# Examples

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

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

- Prestigious job ads automatically shown to men but not women

  https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/

- Housing loans (mortgages) given/denied automatically; correlate with neighborhoods and race

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

- Predictive policing

  https://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

Examples modified from: Suresh Venkatasubramanian

# 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% |

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Example from: Ameet Soni

# Word-embedding examples

**Table 1. Summary of Word-Embedding Association Tests.** We replicated eight well-known IAT findings using word embeddings (rows 1 to 3 and 6 to 10); we also help explain prejudiced human behavior concerning hiring in the same way (rows 4 and 5). Each result compares two sets of words from target concepts about which we are attempting to learn with two sets of attribute words. In each case, the first target is found compatible with the first attribute, and the second target with the second attribute. Throughout, we use word lists from the studies we seek to replicate. $N$, number of subjects; $N_T$, number of target words; $N_A$, number of attribute words. We report the effect sizes ($d$) and $P$ values ($P$, rounded up) to emphasize that the statistical and substantive significance of both sets of results is uniformly high; we do not imply that our numbers are directly comparable with those of human studies. For the online IATs (rows 6, 7, and 10), $P$ values were not reported but are known to be below the significance threshold of $10^{-2}$. Rows 1 to 8 are discussed in the text; for completeness, this table also includes the two other IATs for which we were able to find suitable word lists (rows 9 and 10). We found similar results with word2vec, another algorithm for creating word embeddings, trained on a different corpus, Google News (see the supplementary materials).

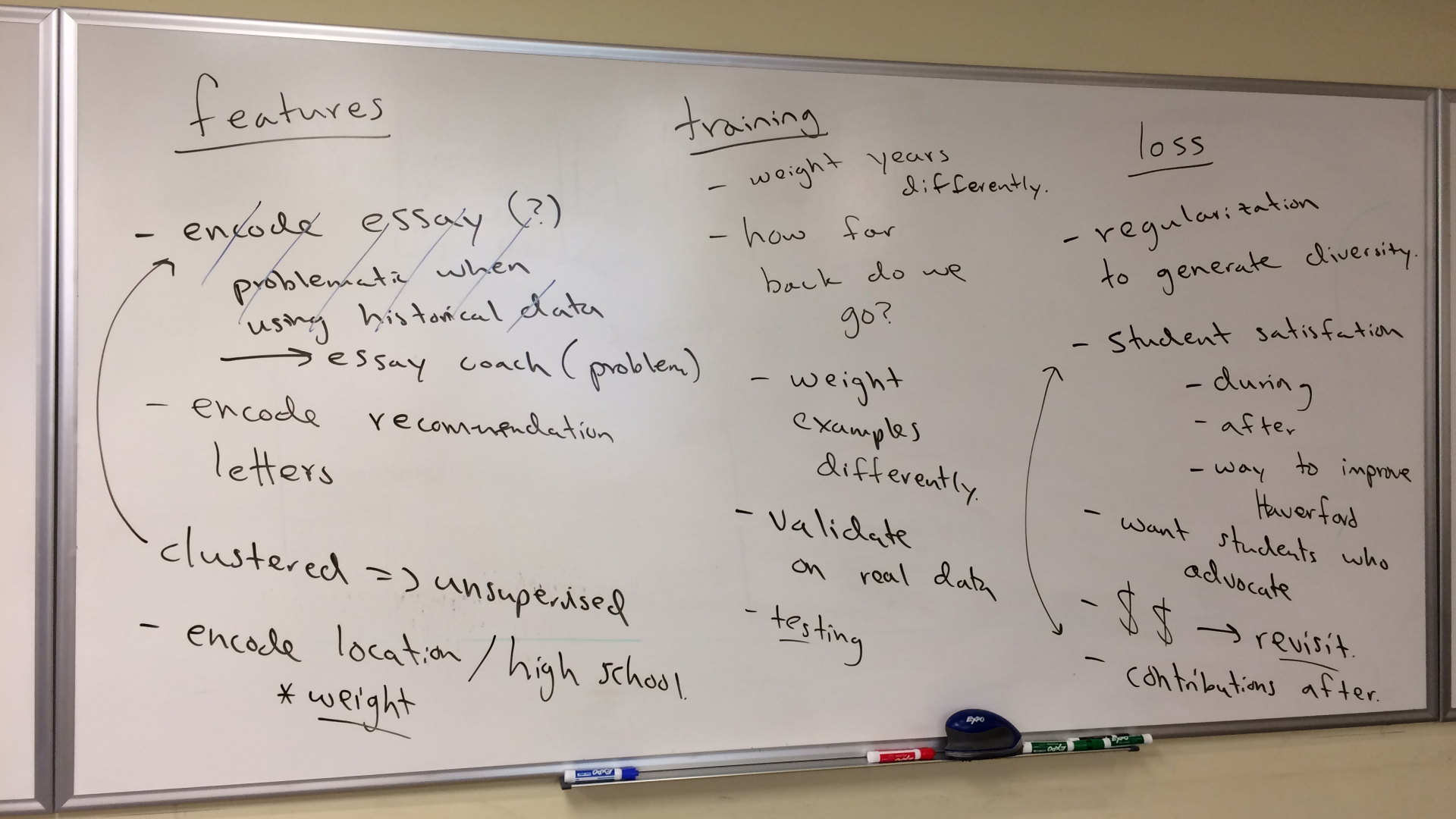
| Target words | Attribute words | Original finding | | | | Our finding | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ref. | $N$ | $d$ | $P$ | $N_T$ | $N_A$ | $d$ | $P$ |
| Flowers vs. insects | Pleasant vs. unpleasant | (5) | 32 | 1.35 | $10^{-8}$ | 25 × 2 | 25 × 2 | 1.50 | $10^{-7}$ |
| Instruments vs. weapons | Pleasant vs. unpleasant | (5) | 32 | 1.66 | $10^{-10}$ | 25 × 2 | 25 × 2 | 1.53 | $10^{-7}$ |
| European-American vs. African-American names | Pleasant vs. unpleasant | (5) | 26 | 1.17 | $10^{-5}$ | 32 × 2 | 25 × 2 | 1.41 | $10^{-8}$ |
| European-American vs. African-American names | Pleasant vs. unpleasant from (5) | (7) | Not applicable | | | 16 × 2 | 25 × 2 | 1.50 | $10^{-4}$ |
| European-American vs. African-American names | Pleasant vs. unpleasant from (9) | (7) | Not applicable | | | 16 × 2 | 8 × 2 | 1.28 | $10^{-3}$ |
| Male vs. female names | Career vs. family | (9) | 39k | 0.72 | $<10^{-2}$ | 8 × 2 | 8 × 2 | 1.81 | $10^{-3}$ |
| Math vs. arts | Male vs. female terms | (9) | 28k | 0.82 | $<10^{-2}$ | 8 × 2 | 8 × 2 | 1.06 | .018 |
| Science vs. arts | Male vs. female terms | (10) | 91 | 1.47 | $10^{-24}$ | 8 × 2 | 8 × 2 | 1.24 | $10^{-2}$ |
| Mental vs. physical disease | Temporary vs. permanent | (23) | 135 | 1.01 | $10^{-3}$ | 6 × 2 | 7 × 2 | 1.38 | $10^{-2}$ |
| Young vs. old people's names | Pleasant vs. unpleasant | (9) | 43k | 1.42 | $<10^{-2}$ | 8 × 2 | 8 × 2 | 1.21 | $10^{-2}$ |

## Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1]* Joanna J. Bryson,[1,2]* Arvind Narayanan[1]*

# Admissions at Haverford

- Haverford has suddenly started receiving 10x more applications than usual
- You are tasked with creating a Machine Learning 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 loss function are you trying to optimize?


features
- encode essay (?)
problematic when using historical data
→ essay coach (problem)
- encode recommendation letters
clustered => unsupervised
- encode location / high school
* weight
training
- weight years differently.
- how far back do we go?
- weight examples differently.
- validate on real data
- testing
loss
- regularization to generate diversity.
- student satisfaction
- during
- after
- way to improve Haverford
- want students who advocate
- $ $ → revisit.
- contributions after.

- use ML as a filter
- humans make borderline decisions

# Outline for December 10

- Finish: Principal Component Analysis (PCA)
- Introduction to bias in ML
- Hand back Midterm 2, common issues

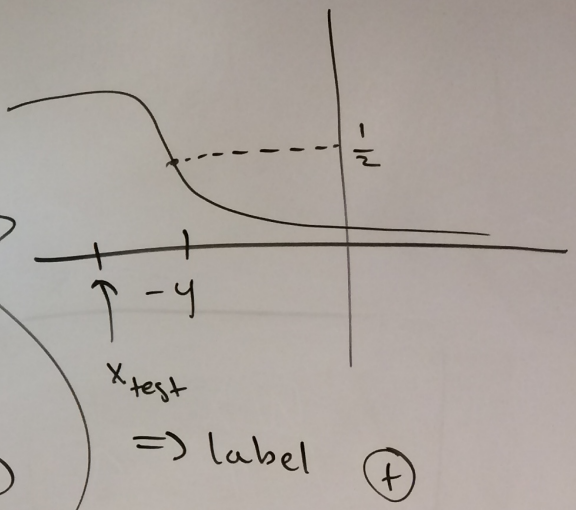
in-lab

(4) probably picking same feature every time

(7) mess up the weights

(9) $-8 - 2x > 0 \Rightarrow +$

$x < -4$




$\Rightarrow$ label +