

CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025



HAVERFORD
COLLEGE

- **Project proposal** feedback was given over email
- **Lab 8** due ~~Wednesday~~ Thursday
- **Final project** instructions/rubric posted
- Plan for next two weeks:
 - Today: finish statistics + unsupervised learning (end midterm 2 material)
 - Thursday and next Tuesday: midterm review
 - **Midterm in-class Thursday April 17**

Outline for today

- Bootstrapping
- Bagging (bootstrap aggregation)
- Revisit data visualization
- Unsupervised learning

Outline for today

- Bootstrapping
- Bagging (bootstrap aggregation)
- Revisit data visualization
- Unsupervised learning

The Bootstrap



In an 18th century story by Rudolph Erich Raspe, Baron Munchausen falls to the bottom of a deep lake.

About to drown, he has the idea to lift himself up by pulling on his bootstraps

(In the original German version, he pulls himself up by his hair, left).

Obviously impossible, this story gave its name to a statistical technique (Efron, 1975) that seems magical, in the sense that you can get something (estimates of uncertainty) for nothing!

In general, the bootstrap is an incredibly useful statistical technique – perhaps one of the most useful in all of modern statistics. You should use it everywhere.

Example: estimating the mean

Data, $\mathbf{X} = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]$

From some distribution with mean μ - we want to learn about μ

Estimate of the mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 4$

How good is this estimate?

Sample standard deviation $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = 3.16$

By the central limit theorem, we know that \bar{X} is approximately normally distributed with variance $\frac{s^2}{n}$ so we can construct confidence intervals and p-values for μ etc... “95% of the time, the 95% CI will contain the true value”.

The bootstrap: Resampling

Data, $\mathbf{X} = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]$

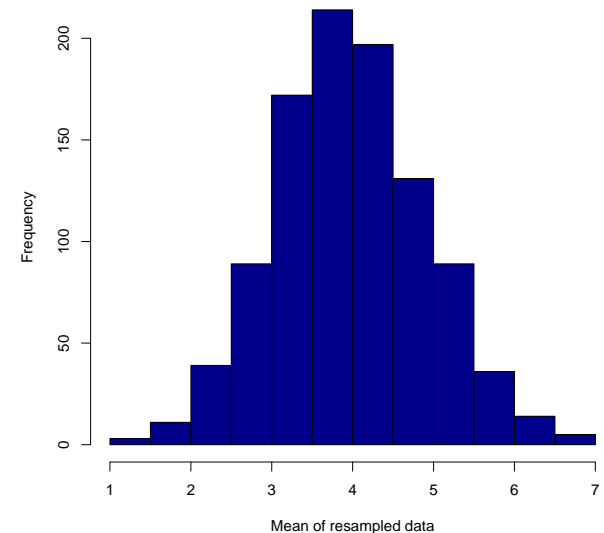
Compute Mean

Resample, with
replacement, T
times

1 8 2 4 6 10 1 1 1 8	→	4.2
1 0 1 6 4 1 4 2 1 2	→	2.2
8 1 6 2 6 4 2 4 10 2	→	4.5
8 3 4 2 10 8 10 8 8 1	→	6.2
6 4 6 4 6 4 2 4 3 4 0	→	4.3
...	→	...
...	→	...

Use the means from the
resampled data to estimate
the distribution!

95% of the means are
between 2.3 and 5.9 (T=1000)



The bootstrap: Resampling

“Estimate the range (Max—Min)”

Data, $\mathbf{X} = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]$

Compute Range

Resample, with
replacement, T
times

1 8 2 4 6 10 1 1 1 8 → 9

1 0 1 6 4 1 4 2 1 2 → 6

8 1 6 2 6 4 2 4 10 2 → 9

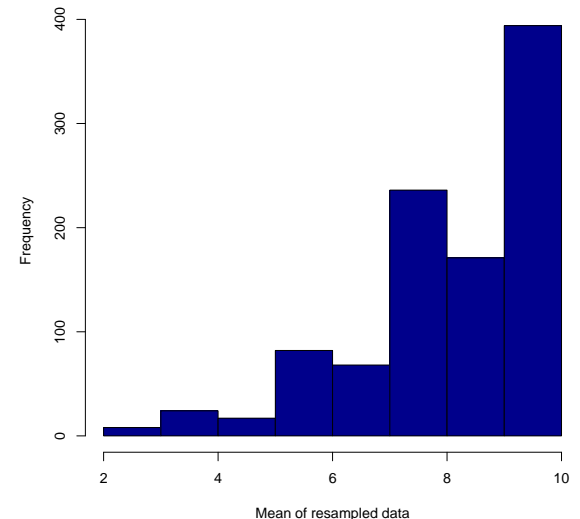
8 3 4 2 10 8 10 8 8 1 → 8

6 4 6 4 6 4 2 4 3 4 0 → 6

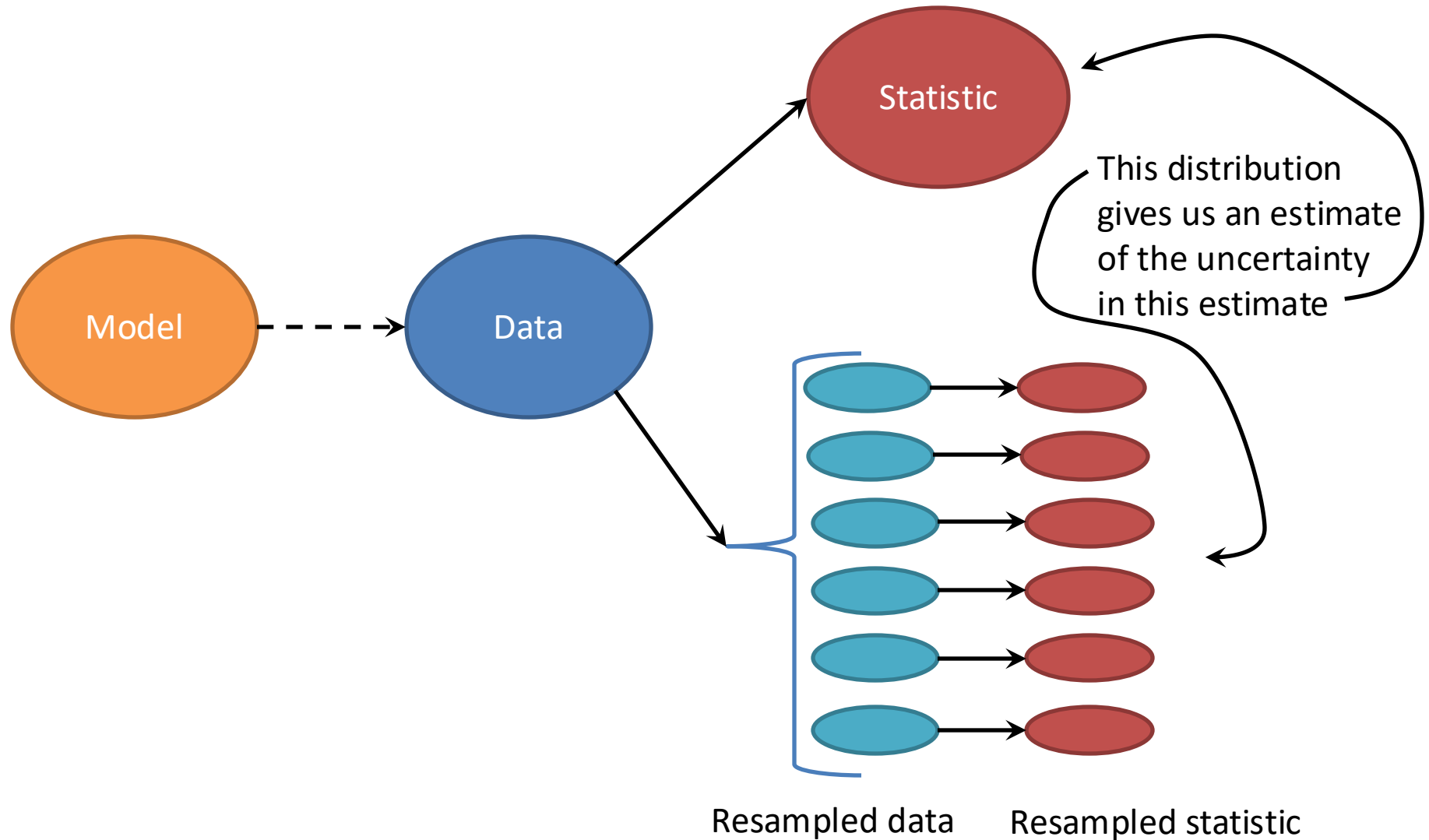
... → ...

... → ...

Use the ranges from the
resampled data to estimate
the distribution!



The bootstrap: Resampling



The bootstrap: Resampling

- The key point is that as long as we can resample our data (which we can always do).
- And calculate the thing we want to estimate (which we can almost always do).
- We can bootstrap anything, and get a sense of how good our estimate is.
- We do not need to make any assumptions about the underlying distribution. For example, to apply the central limit theorem.

The bootstrap: Resampling

- In general resampling or permutation method can answer most of the statistical questions that we are interested in (is the mean zero? are these distributions the same?)
- Why then in intro stats did we learn about t-tests, z-scores, and the central limit theorem instead of permutation tests and bootstrapping?
- Because when statistics was invented in the 1920s, people didn't have computers!

Bootstrap example

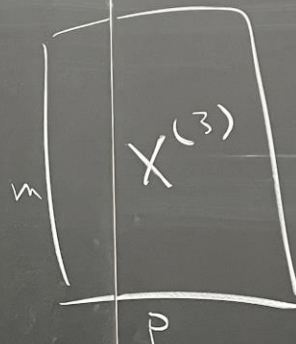
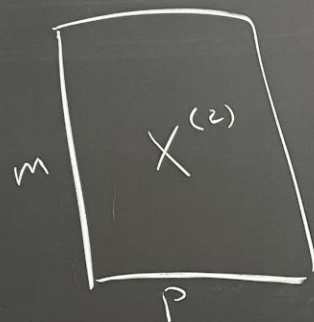
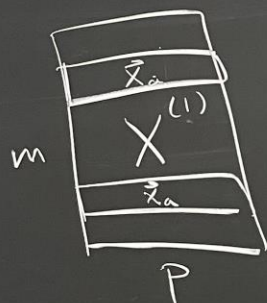
Setup: you obtain 0.87 accuracy on a test dataset using a new algorithm

Goal: find a 95% confidence interval for your estimate

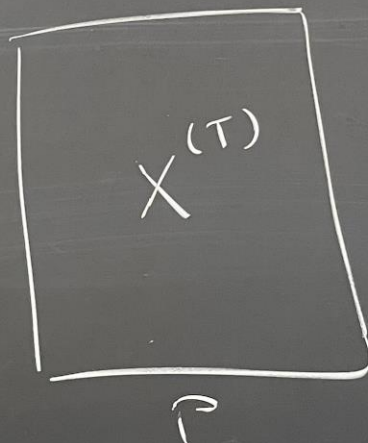
Bootstrap

① bootstrap T times, run our method on each new dataset

$T=1000$
(example)



...



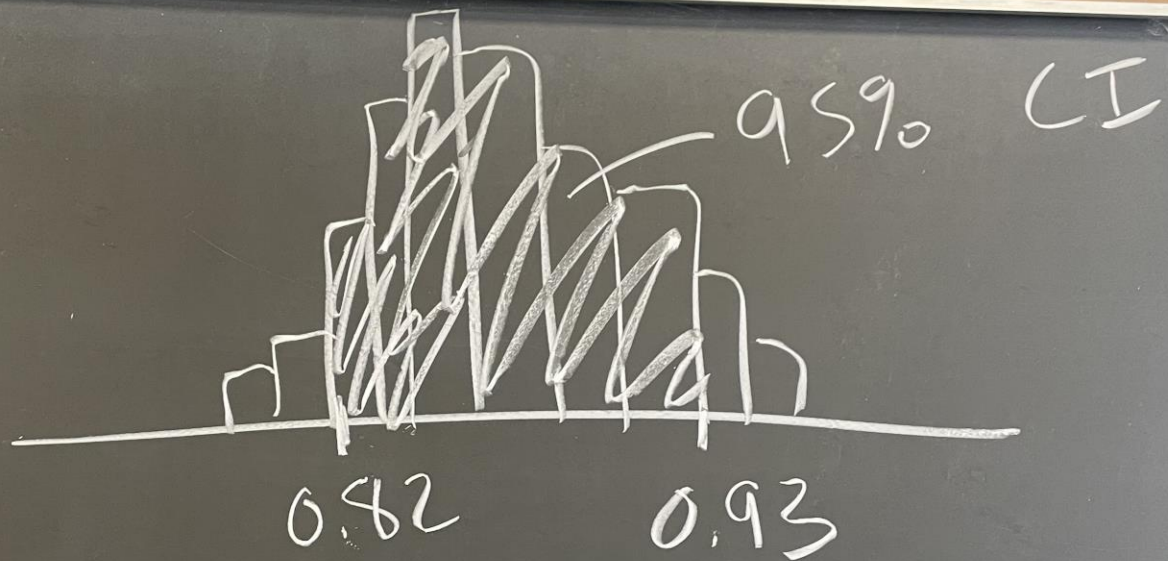
record
results.
accuracy

[0.82, 0.91, 0.86, ..., 0.95]

② Sort results

③ take the middle 95% (ex: 950)

$\rightarrow I = (0.82, 0.93)$

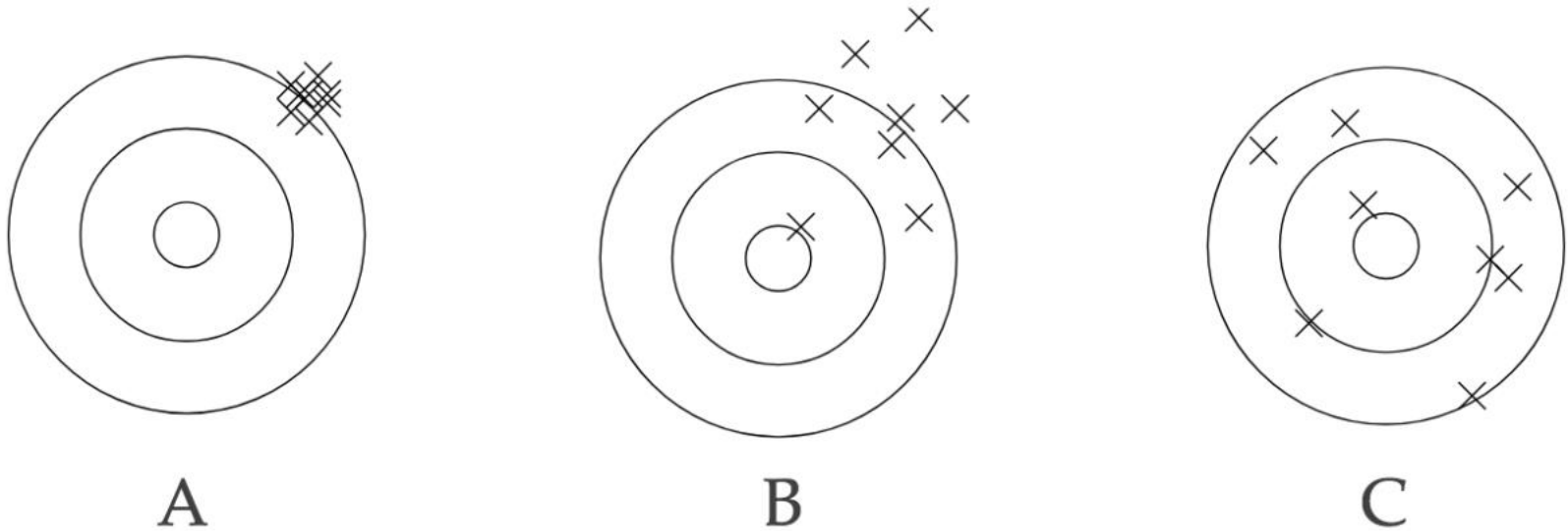


Outline for today

- Bootstrapping
- Bagging (bootstrap aggregation)
- Revisit data visualization
- Unsupervised learning

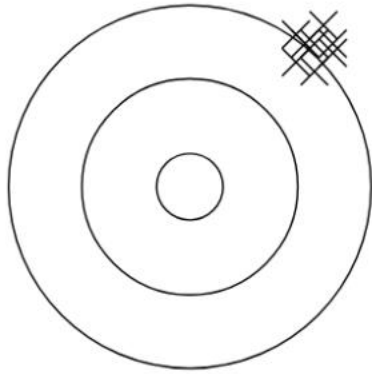
Bagging (Bootstrap Aggregation)

Motivation: bias and variance

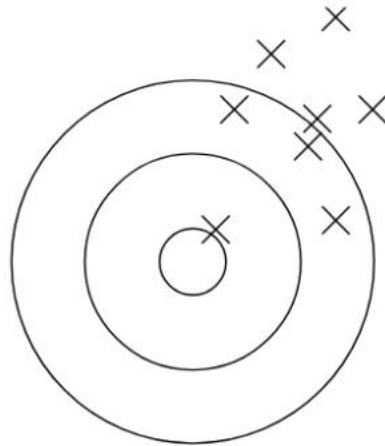


Label each picture with variance (high or low) and bias (high or low)

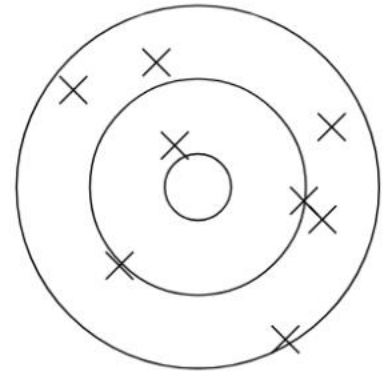
Motivation: bias and variance



A



B

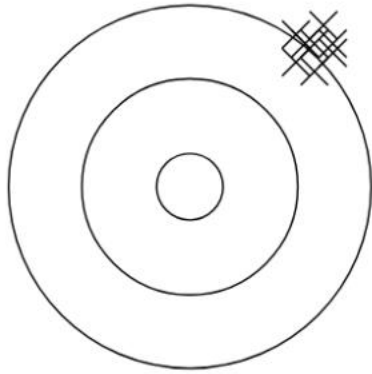


C

Variance: low
Bias: high

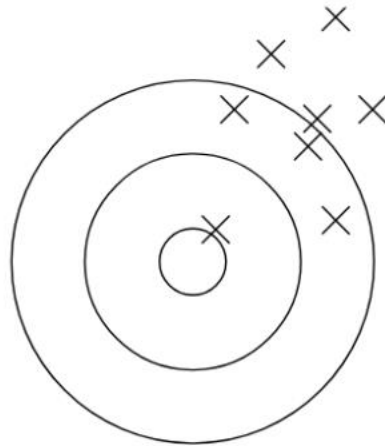
Label each picture with variance (high or low) and bias (high or low)

Motivation: bias and variance



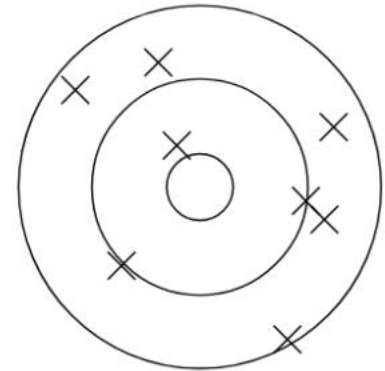
A

Variance: low
Bias: high



B

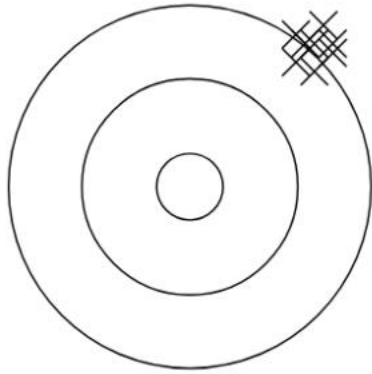
Variance: high
Bias: high



C

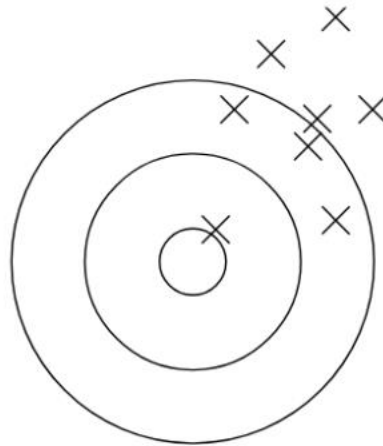
Label each picture with variance (high or low) and bias (high or low)

Motivation: bias and variance



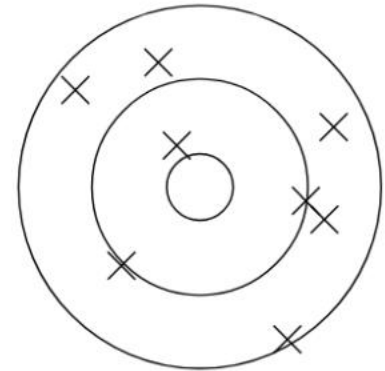
A

Variance: low
Bias: high



B

Variance: high
Bias: high

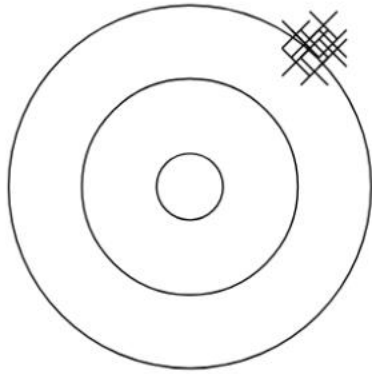


C

Variance: high
Bias: low

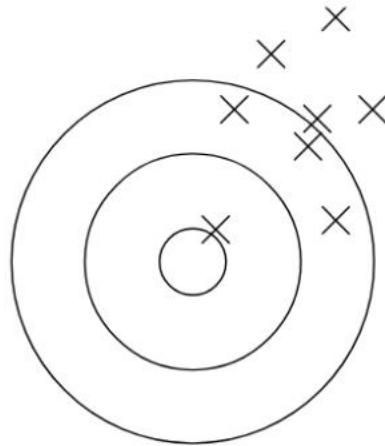
Label each picture with variance (high or low) and bias (high or low)

Motivation: bias and variance



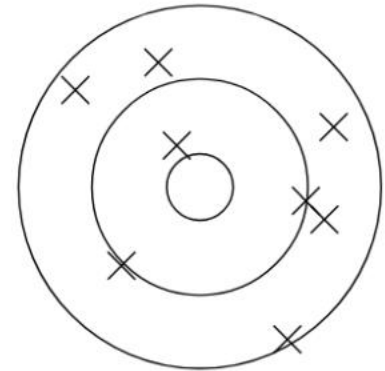
A

Variance: low
Bias: high



B

Variance: high
Bias: high



C

Variance: high
Bias: low

This is the type of classifier
we want to average!

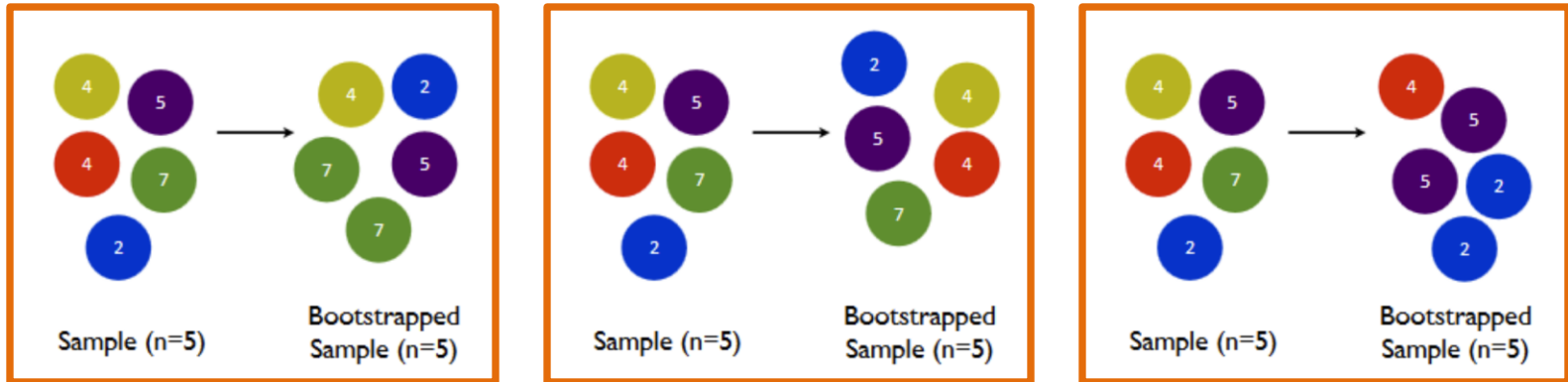
Label each picture with variance (high or low) and bias (high or low)

Ensemble Idea

- Average the results from several models with **high variance** and **low bias**
 - Important that models be diverse (don't want them to be wrong in the same ways)
- If n observations each have variance s^2 , then the mean of the observations has variance s^2/n (reduce variance by averaging!)

Bagging Algorithm

- ❖ Bagging = Bootstrap Aggregation [Brieman, 1996]
- ❖ *Bootstrap* (randomly sample with replacement) original data to create many different training sets
- ❖ Run base learning algorithm on each new data set independently



Desmond Ong, Stanford

Bagging (Bootstrap Aggregation)

Train:

for t in range(T):

- * create bootstrap sample $X^{(t)}$ of size n
from training data
- * train on $X^{(t)}$ to get model $h^{(t)}$

Test:

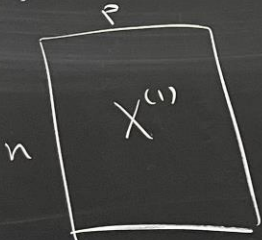
for each test example, the T classifiers **vote**
on the label

Random Forests

Random Forests
train

$T=3$

bootstrap



model

Hennis

refit classifier



$h^{(2)}$



$h^{(3)}$



test
 $\vec{x} = \begin{bmatrix} \text{outlook} & \text{temp} & \text{wind} & \text{hum} \\ \text{rain} & \text{high} & \text{low} & \text{high} \end{bmatrix}$

$$h^{(1)}(\vec{x}) = +1$$

$$h^{(2)}(\vec{x}) = -1$$

$$h^{(3)}(\vec{x}) = -1$$

Vote!
(average)

$$\boxed{h(\vec{x}) = -1}$$

★ ★

Outline for today

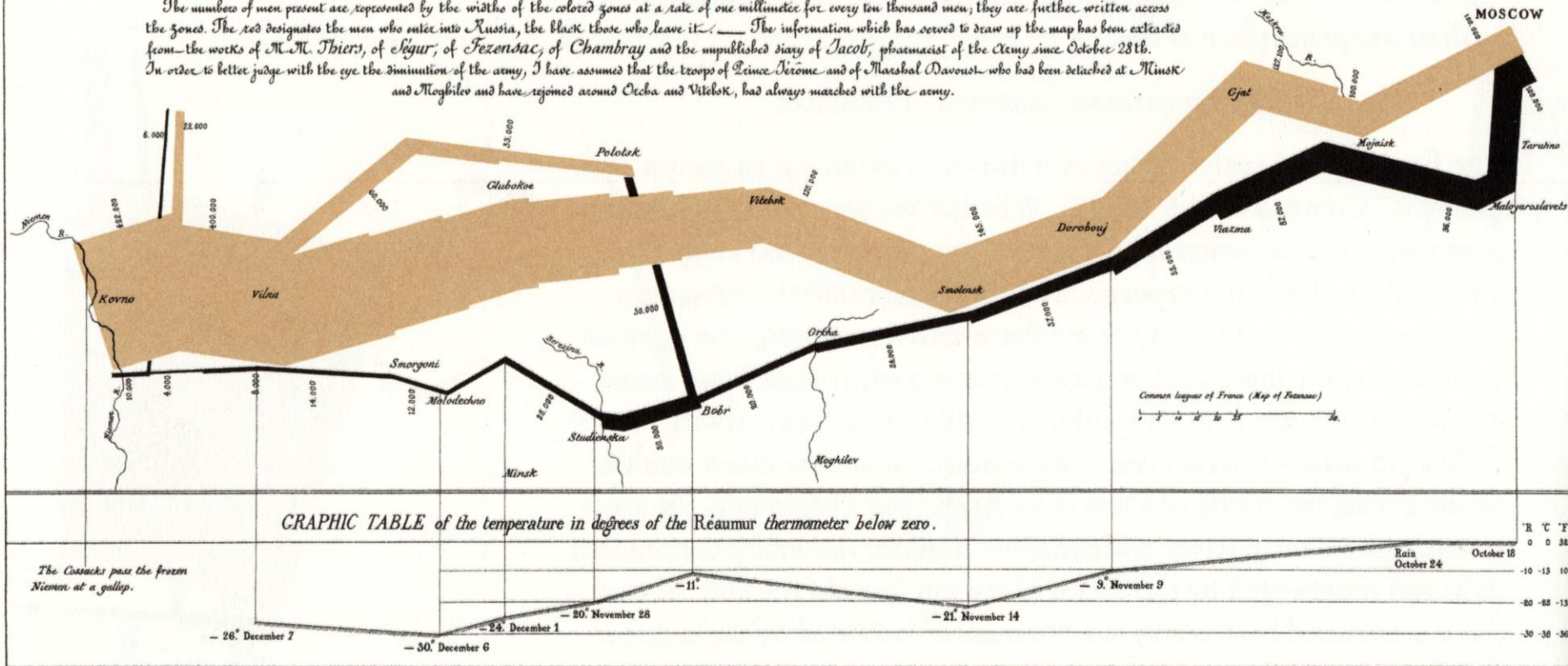
- Bootstrapping
- Bagging (bootstrap aggregation)
- **Revisit data visualization**
- Unsupervised learning

Visualization can illuminate...

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~1813.

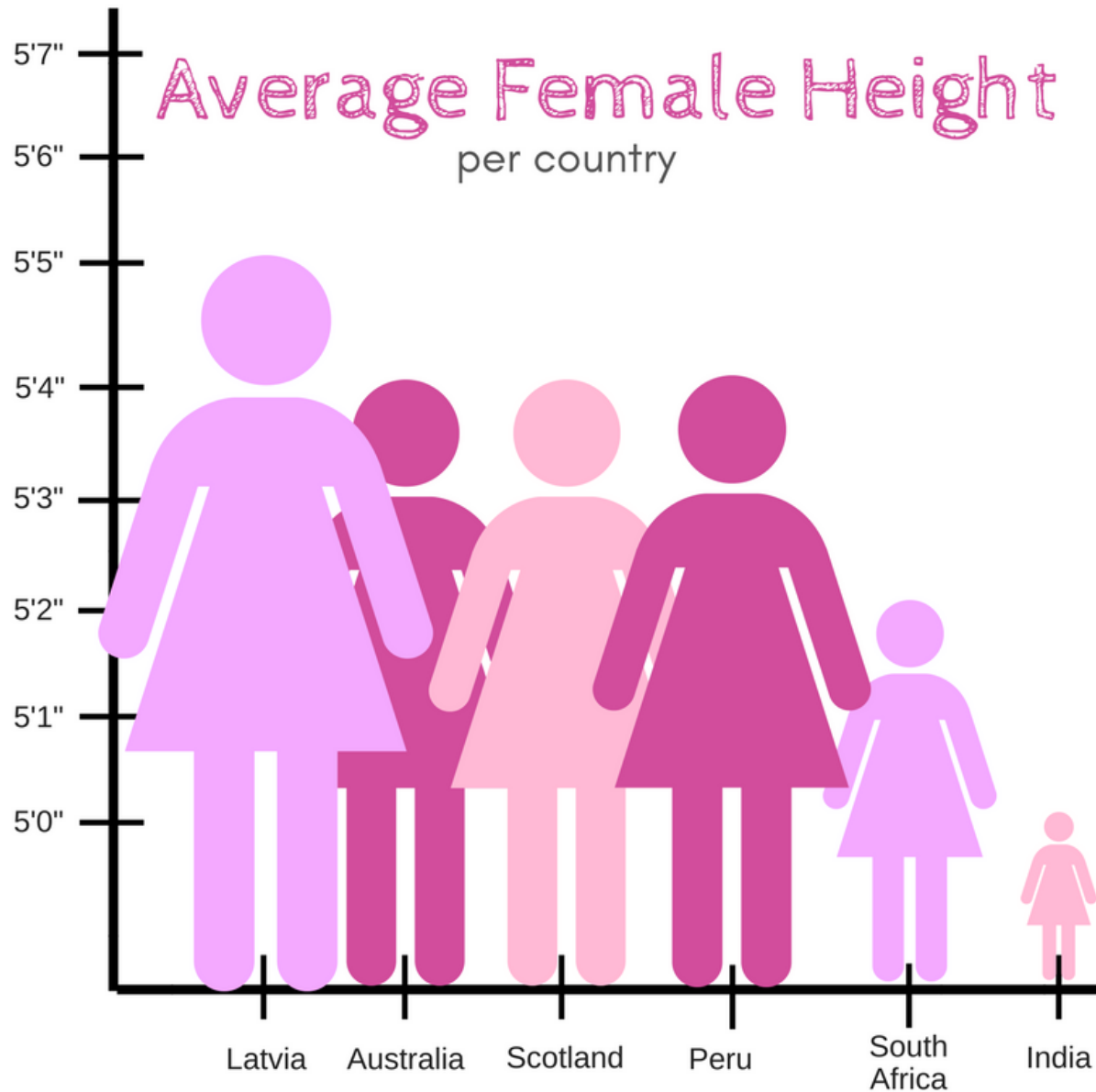
Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement. Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter into Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M. Thiers, of Fozensac, of Chambray and the unpublished diary of Jacob, pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davoust who had been detached at Minsk and Moghilev and have rejoined around Orcha and Vitebsk, had always marched with the army.



Size of Napoleon's army on the advance (in tan) and retreat (in black) from Moscow in 1812

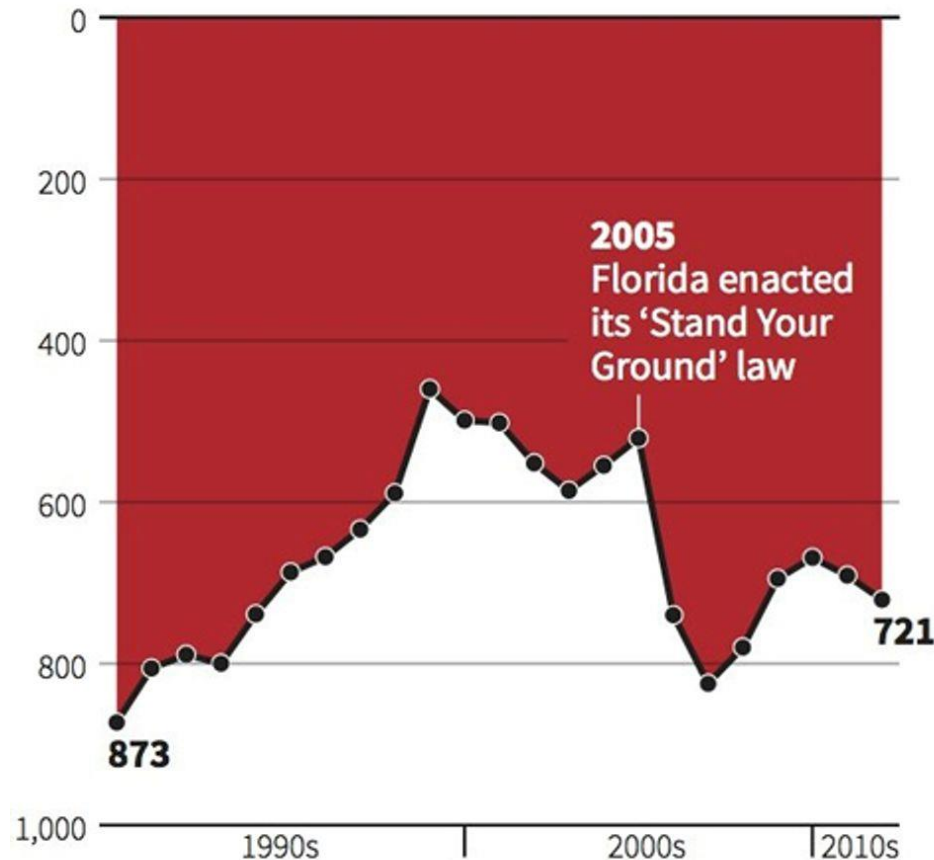
... but also mislead



... but also mislead

Gun deaths in Florida

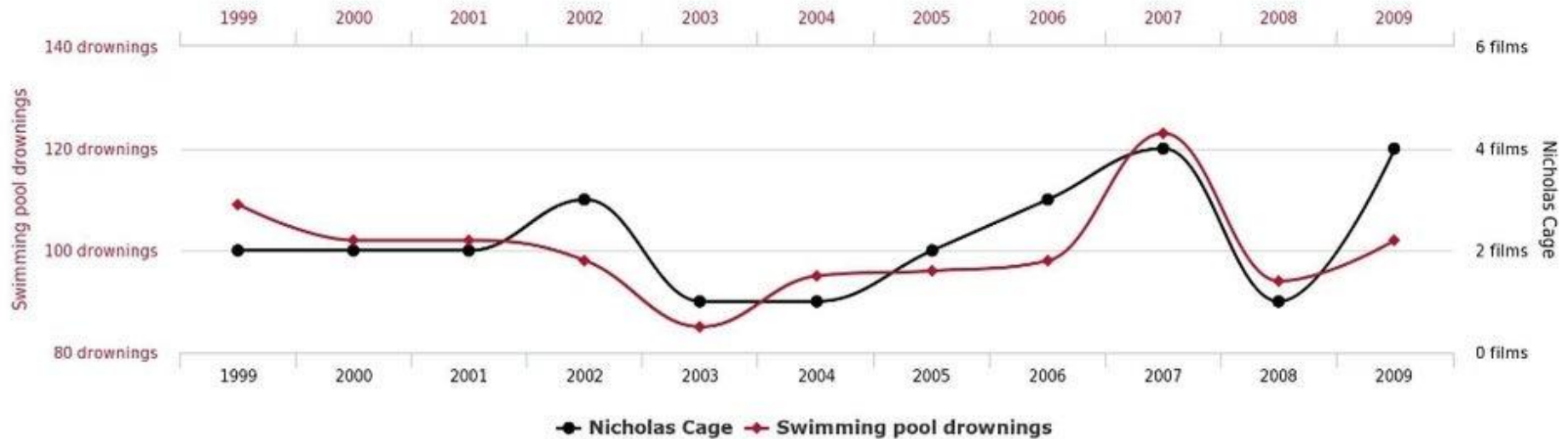
Number of murders committed using firearms



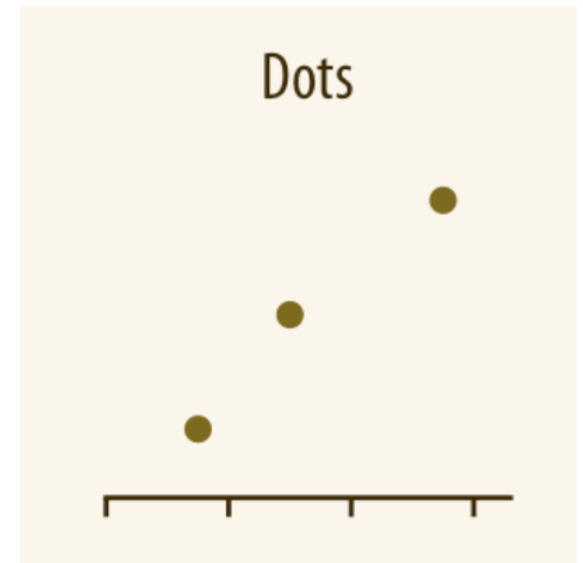
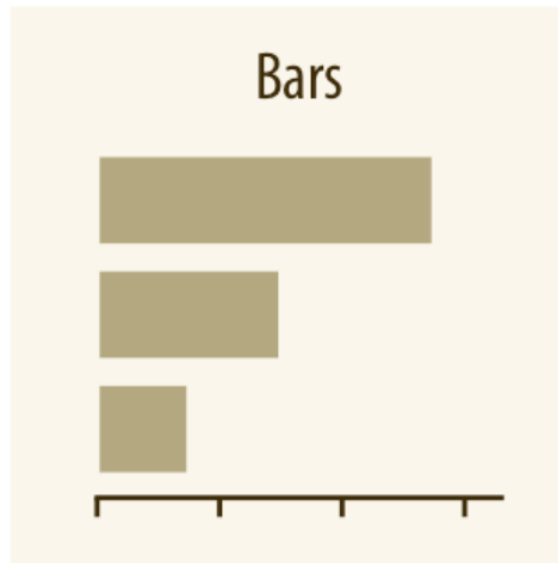
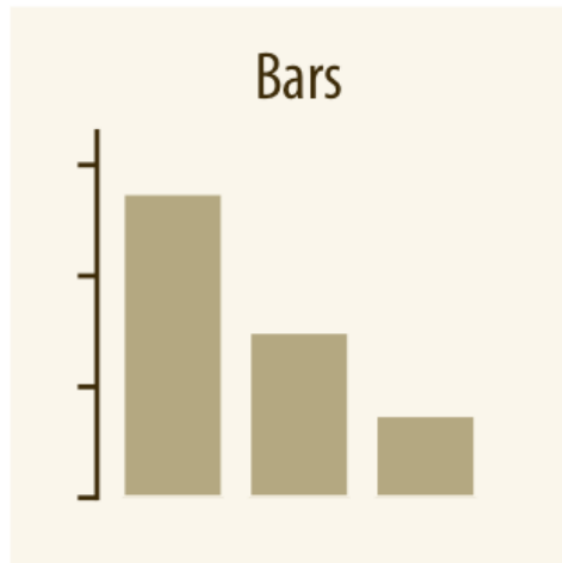
Source: Florida Department of Law Enforcement

... but also mislead

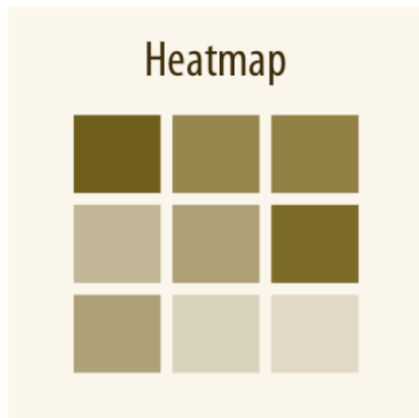
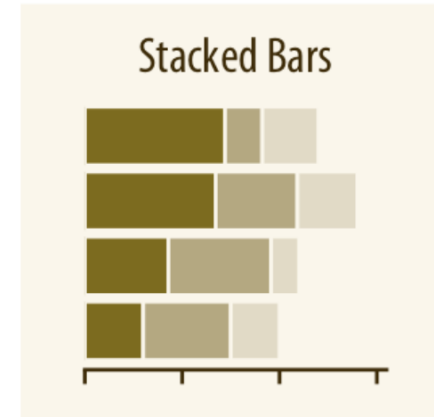
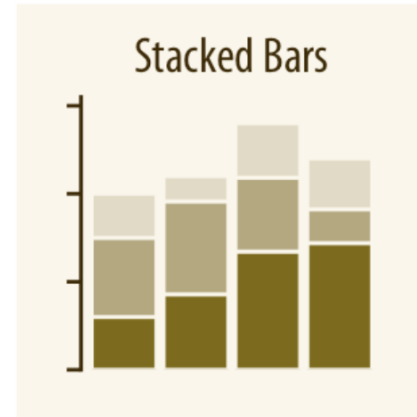
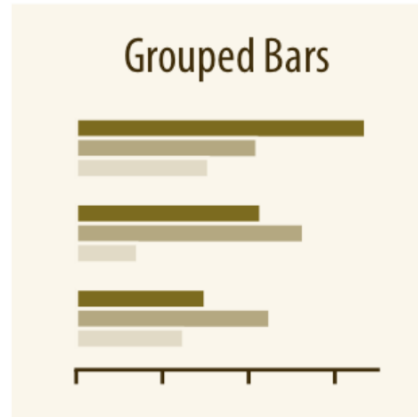
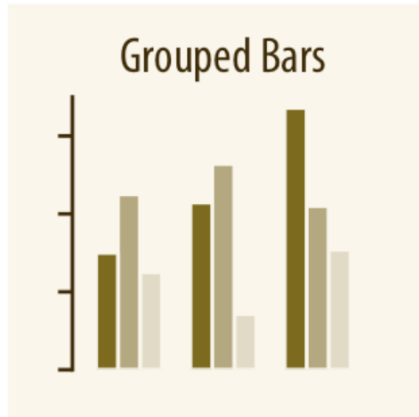
Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



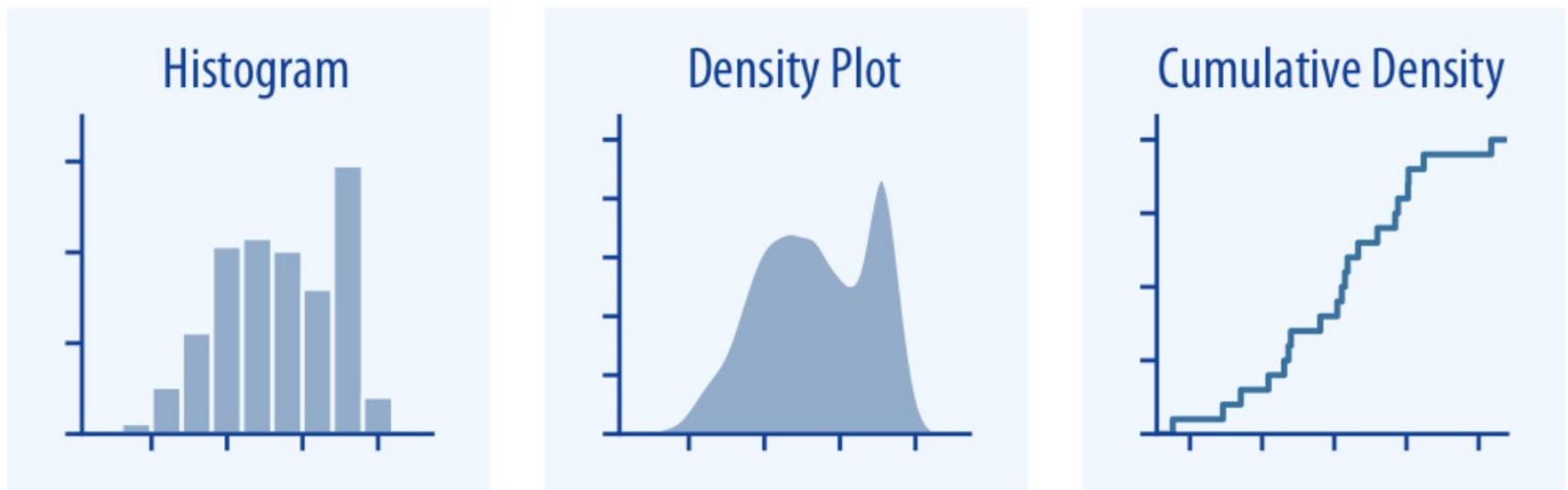
Visualizing amounts



Visualizing amounts

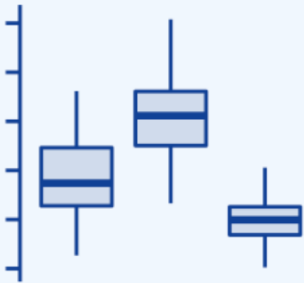


Visualizing distributions

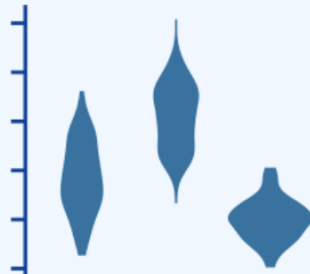


Visualizing distributions

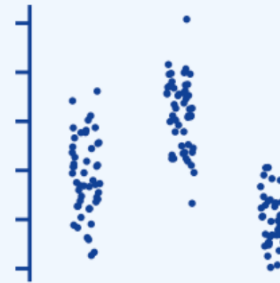
Boxplots



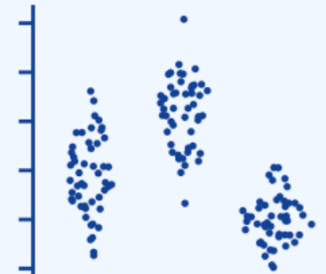
Violins



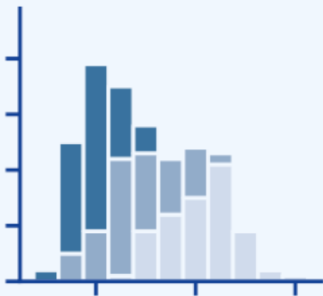
Strip Charts



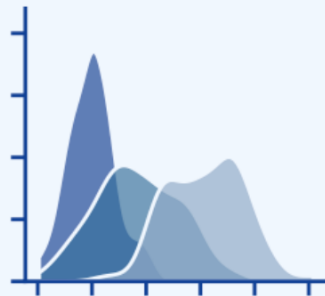
Sina Plots



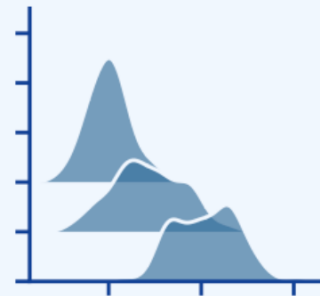
Stacked Histograms



Overlapping Densities



Ridgeline Plot



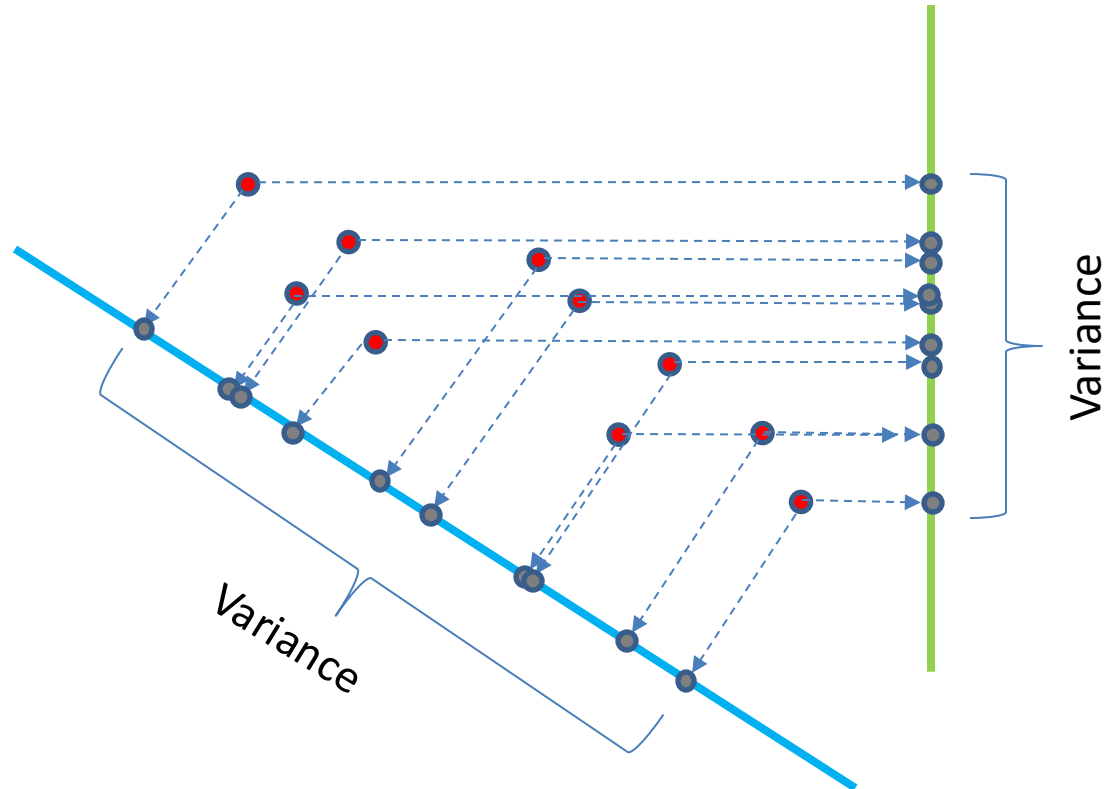
Quick break: write one midterm
topic/question on the notecard

Alternative to PCA

Reducing dimensions

- How?

- Project the points from high-dimensions to low dimensions

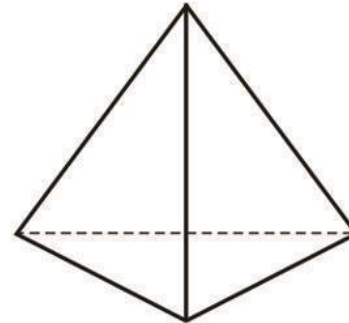


Prefer the blue line because more spread of the original data is represented → Principal Component Analysis (**PCA**)

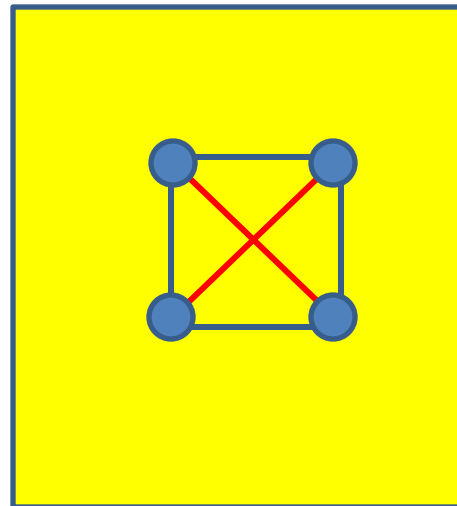
Reducing dimensions

- How?

- Project the points from high-dimensions to low dimensions
- Reconstruct high dimensional relationships in low dimensions



Tetrahedron with length 1 sides.
All pairwise distances between the four points = 1

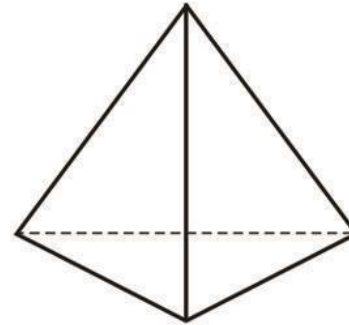


Try to arrange four points in 2D such that pairwise distances are as close to the original pairwise distances

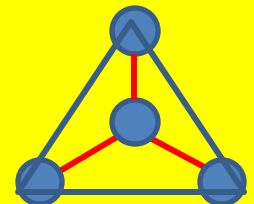
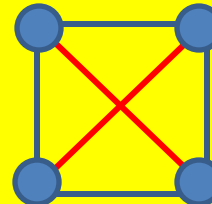
Reducing dimensions

- How?

- Project the points from high-dimensions to low dimensions
- Reconstruct high dimensional relationships in low dimensions



Tetrahedron with length 1 sides.
All pairwise distances between the four points = 1

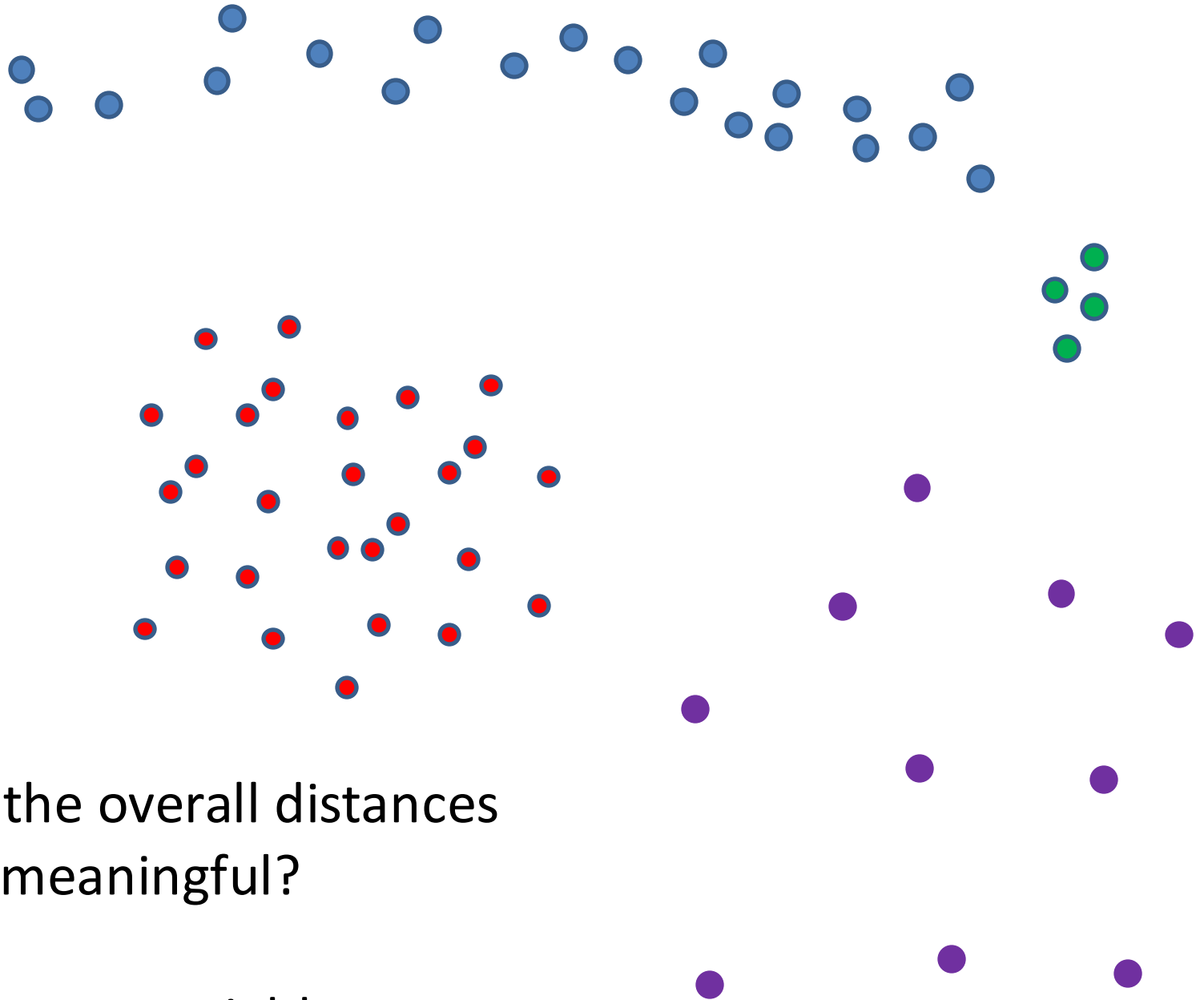


A lot of the time we want to create clusters.

Distances in the original data may not be meaningful

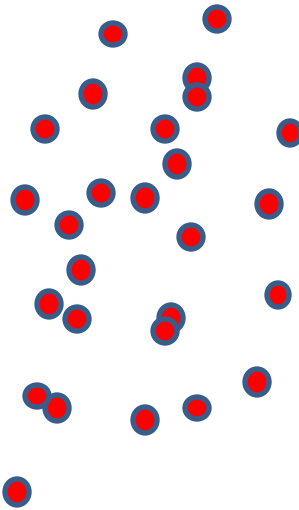
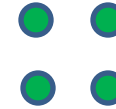
So we want some kind of embedding that preserves clustering

Linear projection (e.g. PCA) is only one type of embedding



What if the overall distances
are not meaningful?

Focus on your neighbors

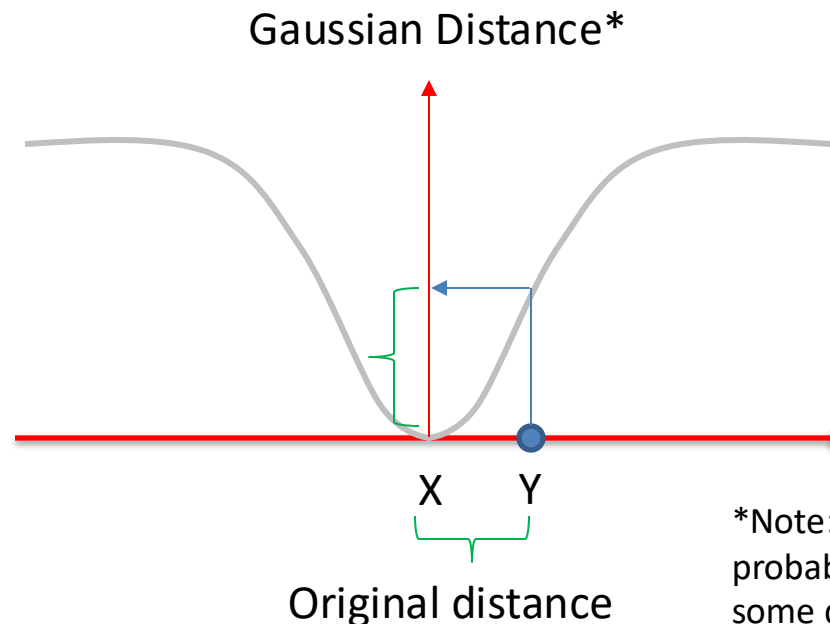


What if the overall distances
are not meaningful?

Focus on your neighbors

tSNE (t-distributed Stochastic Neighborhood Embedding)

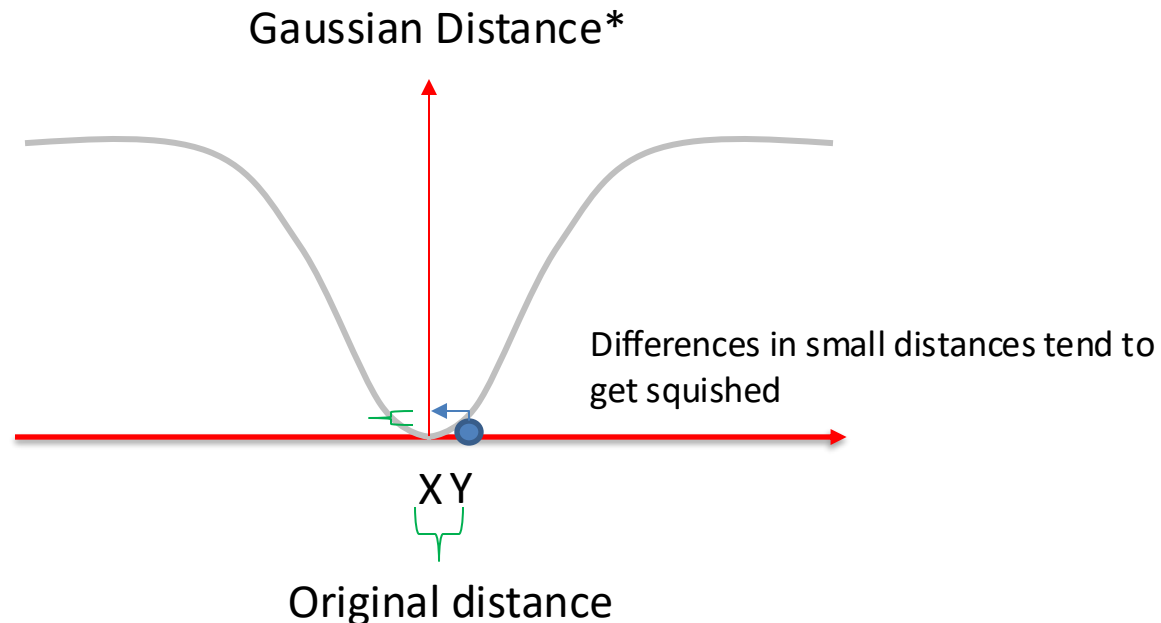
- Define distances between a point X to a point Y by a Gaussian function centered at X



*Note: the actual algorithm uses notions of probability (i.e., probability of finding Y at some distance from X). I use notion of distance as a proxy

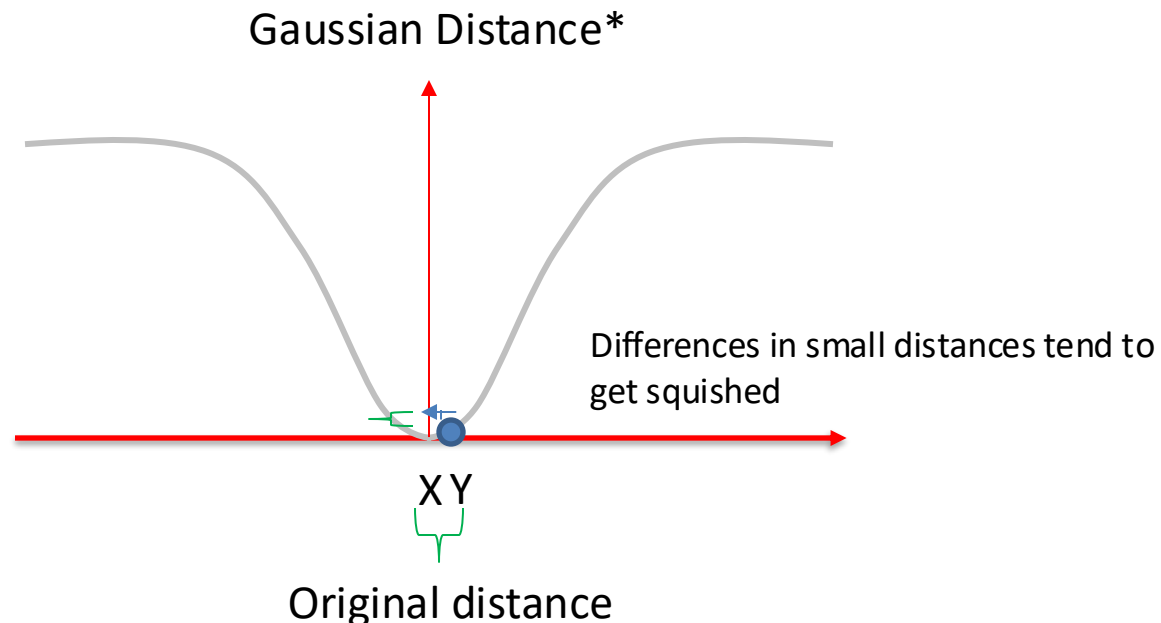
tSNE (t-distributed Stochastic Neighborhood Embedding)

- Define distances between a point X to a point Y by a Gaussian function centered at X



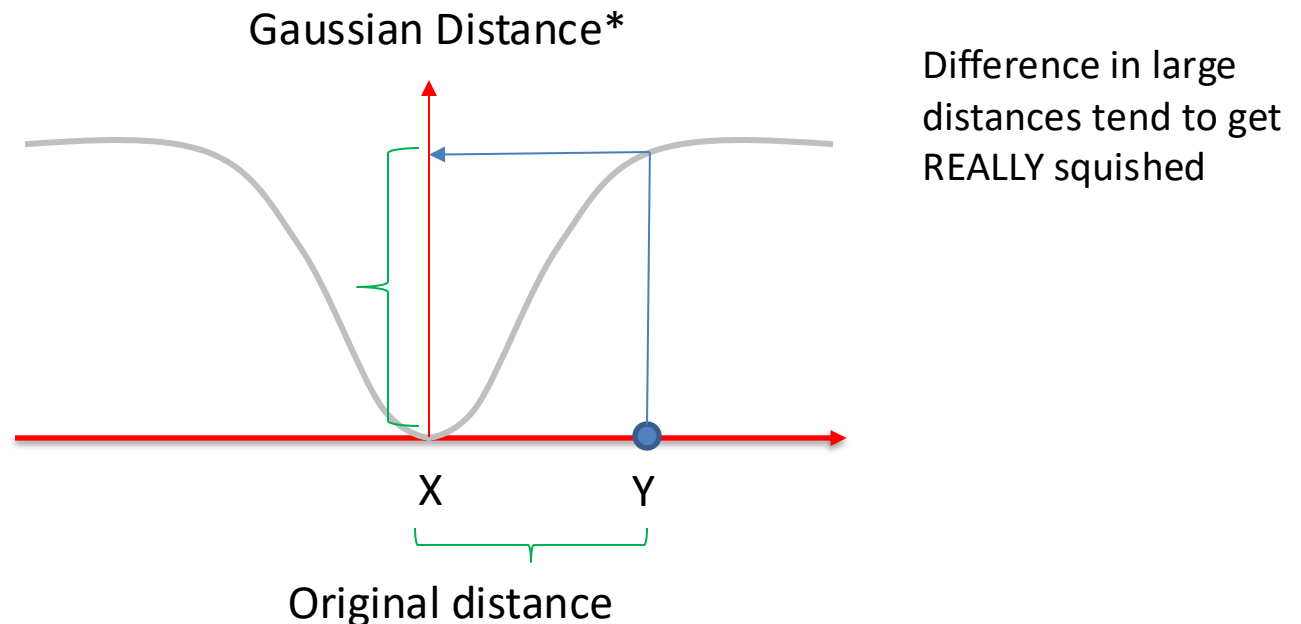
tSNE (t-distributed Stochastic Neighborhood Embedding)

- Define distances between a point X to a point Y by a Gaussian function centered at X



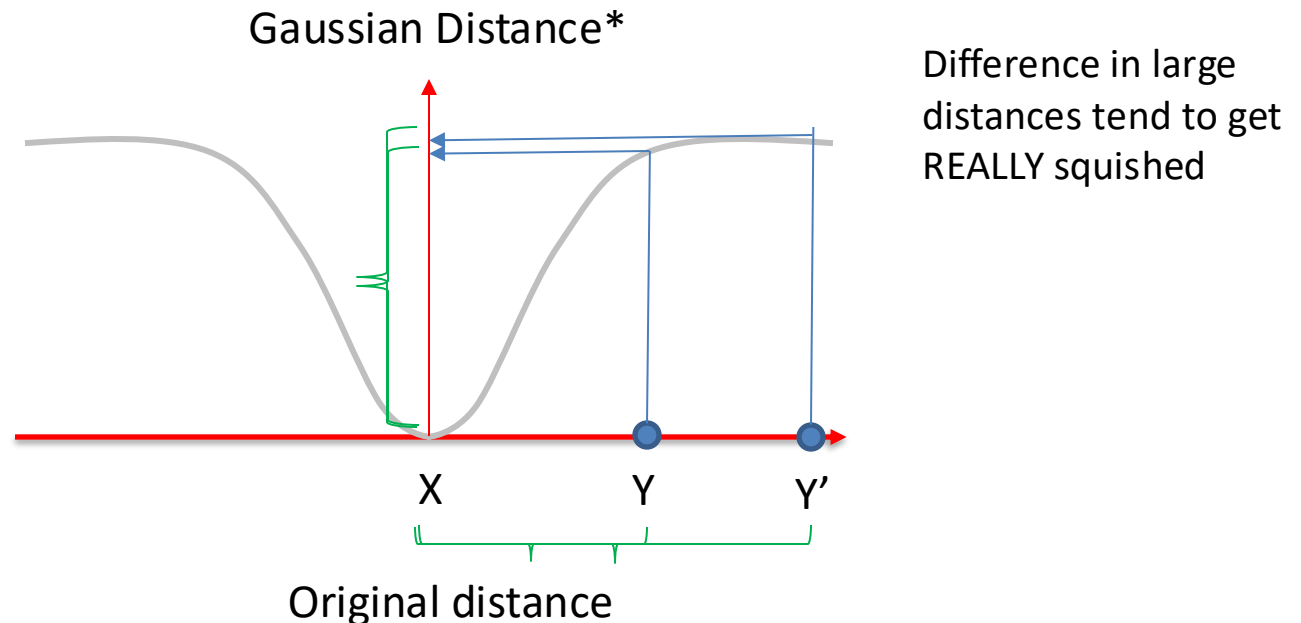
tSNE (t-distributed Stochastic Neighborhood Embedding)

- Define distances between a point X to a point Y by a Gaussian function centered at X



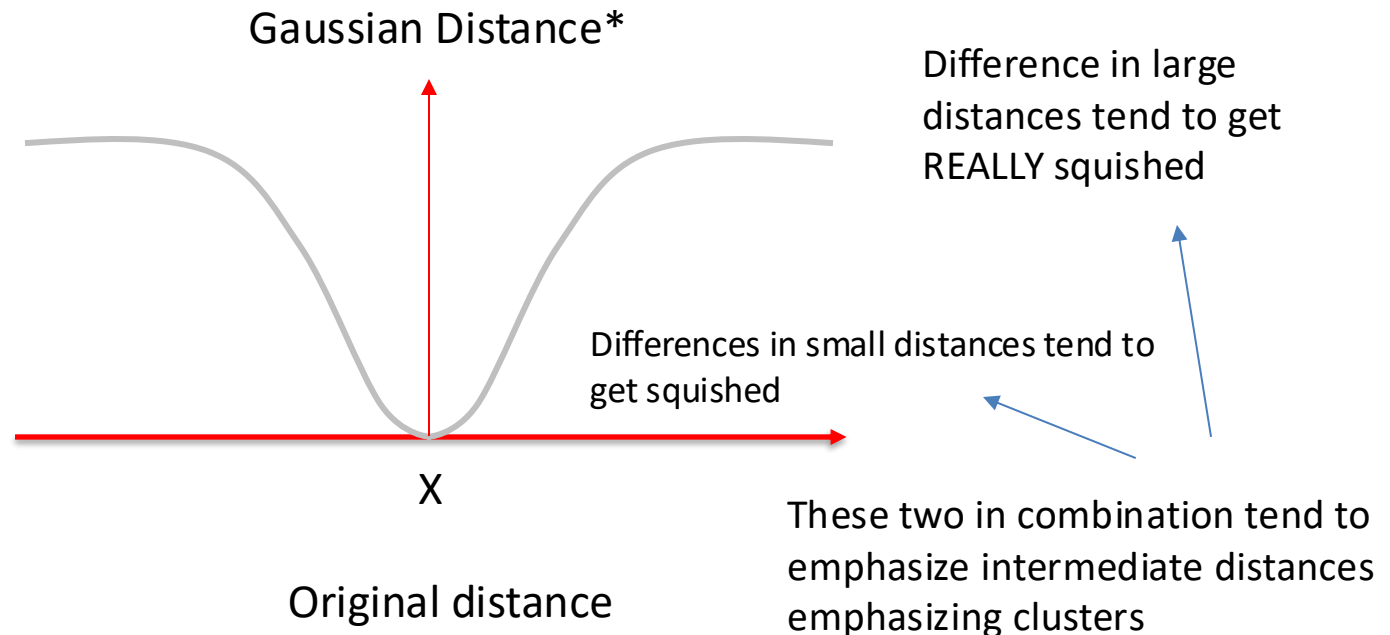
tSNE (t-distributed Stochastic Neighborhood Embedding)

- Define distances between a point X to a point Y by a Gaussian function centered at X

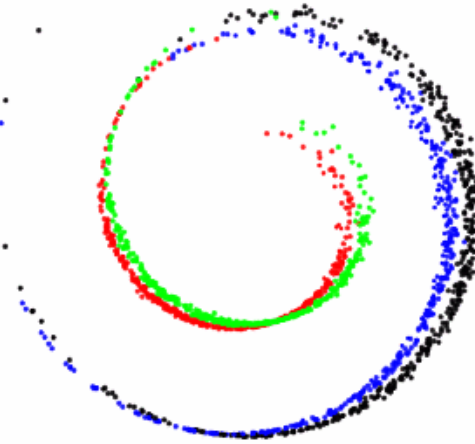


tSNE (t-distributed Stochastic Neighborhood Embedding)

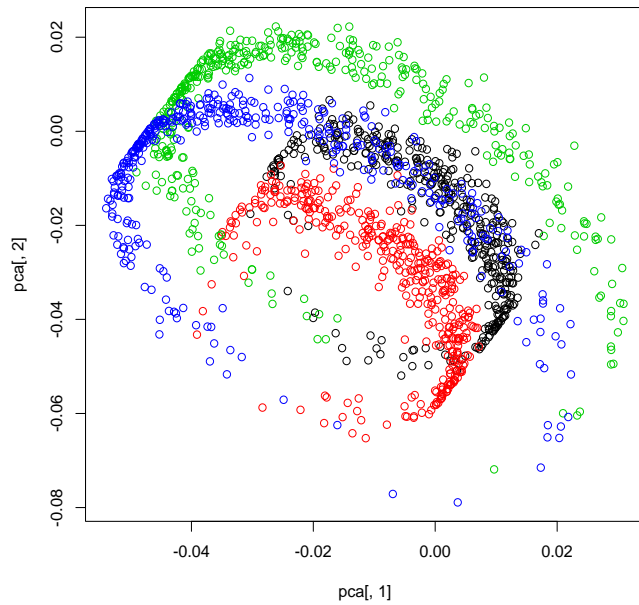
- Define distances between a point X to a point Y by a Gaussian function centered at X



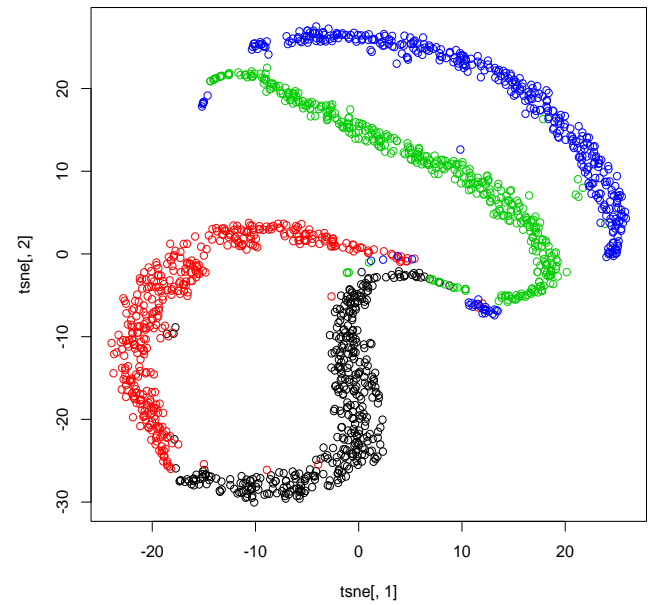
Original data



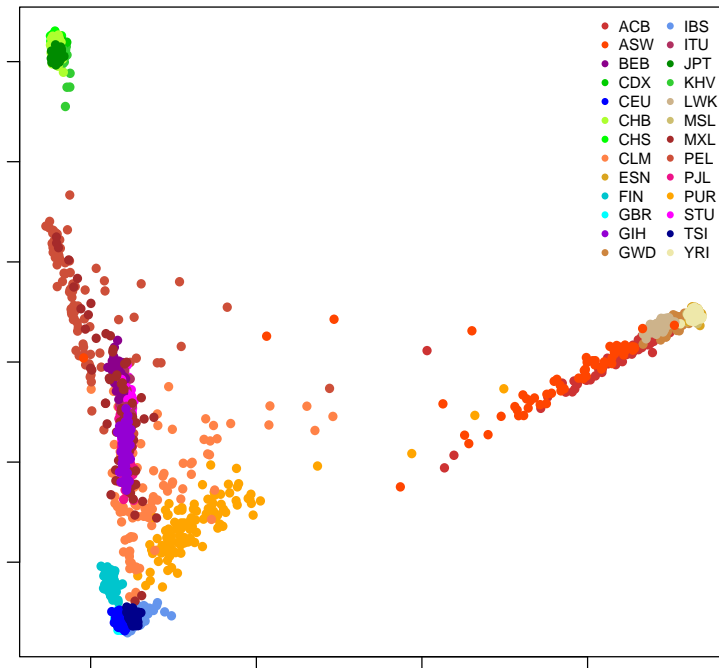
PCA



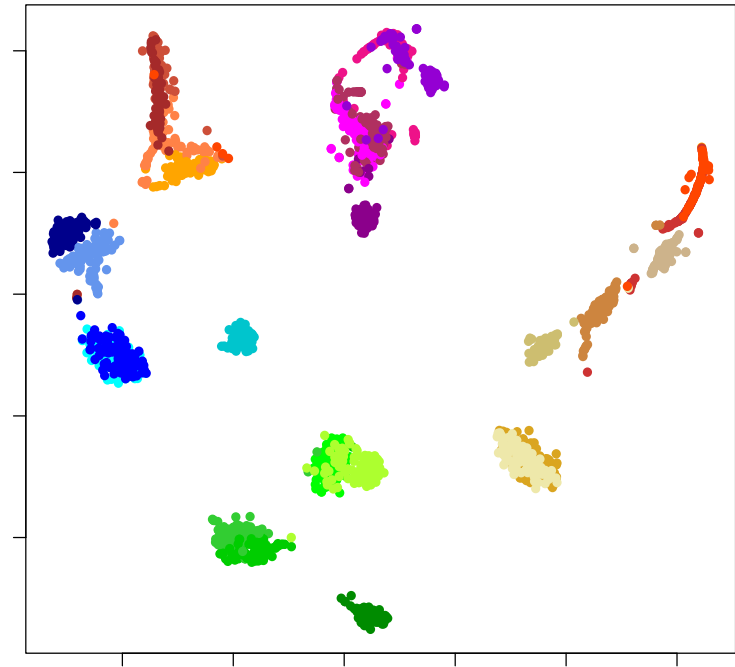
t-SNE



PCA

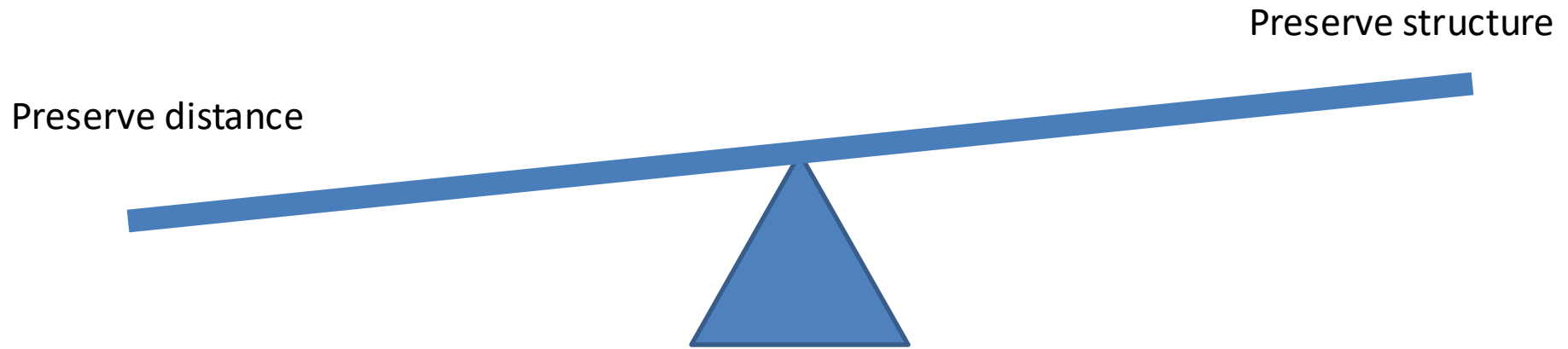


t-SNE



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

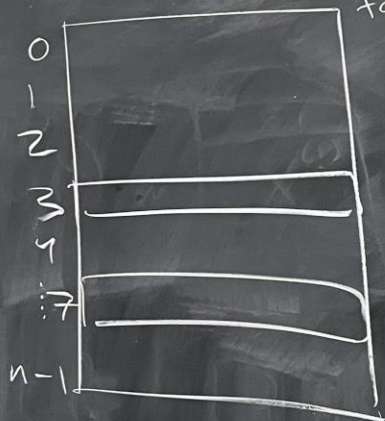


How to visualize data always depends on the data, and the question

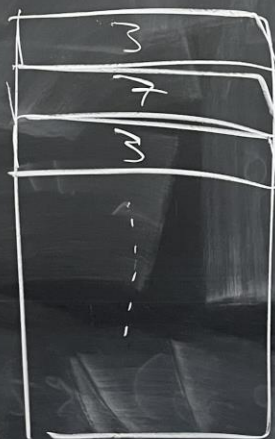
There is rarely if ever a single correct approach

Handout 19

① can assume $\text{rand_int } O(1)$
 for t in range (T) :
 for i in range (n) :
 $j = \text{rand_int}(n)$



bootstrap



Same size

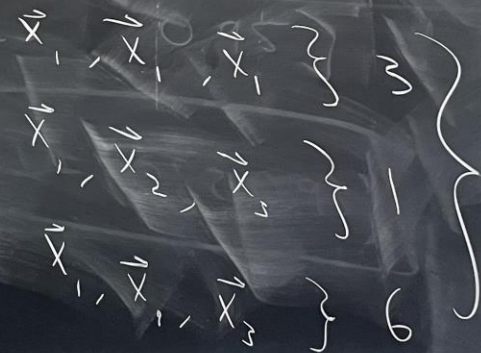
$O(nT)$

② original data

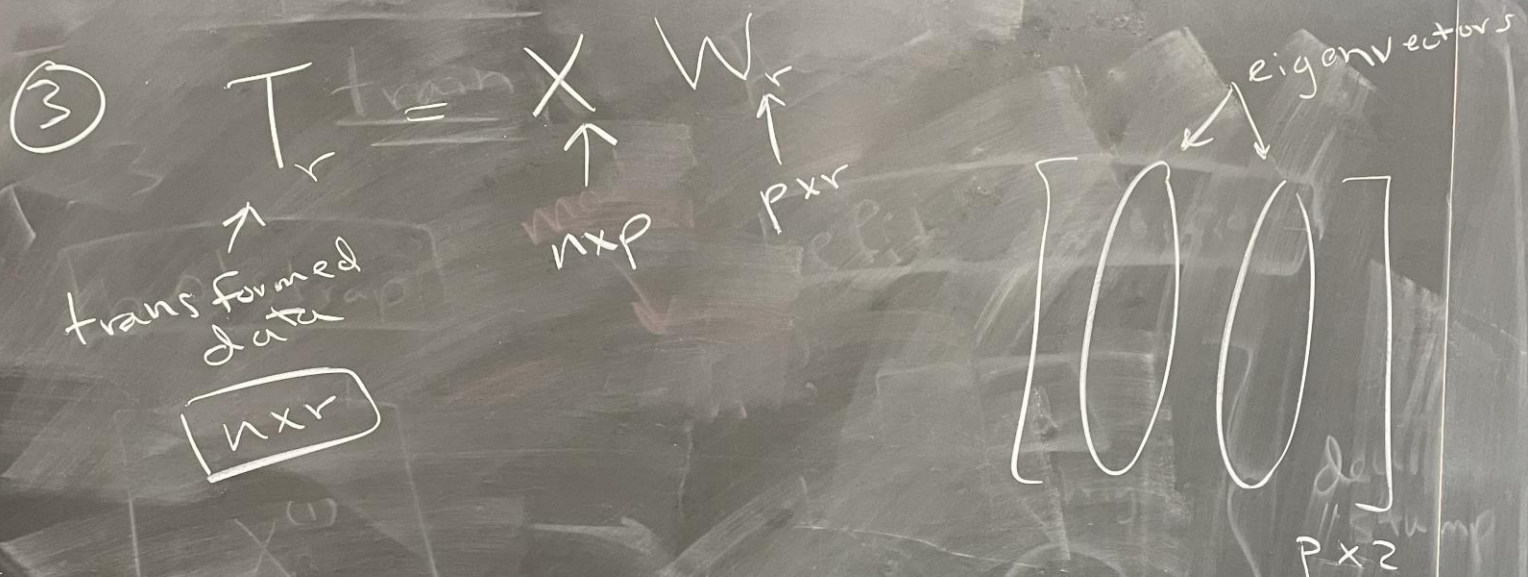


3 possible datasets

$n=3$



10 datasets



④ next time!

Outline for today

- Bootstrapping
- Bagging (bootstrap aggregation)
- Revisit data visualization
- Unsupervised learning

After midterm – will not be on the exam!

Quote of the week

“He who buys what he does not need, steals from himself.”

– Swedish Proverb