CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD COLLEGE

Outline

- Midterm 2 Review
  - Logistic regression and cross entropy
  - Naive Bayes
  - Disparate impact
- Less focus
  - tSNE
  - t-tests

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

“Certifying and Removing Disparate Impact” Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

“Certifying and Removing Disparate Impact” Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Direct discrimination: $C = f(X)$

* Female instrumentalist not hired for orchestra
* Some ethnic groups not allowed to eat at a restaurant

“Certifying and Removing Disparate Impact” Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Indirect discrimination: $C = f(Y)$

* but strong correlation between X and Y
* Ex: housing loans
* Ex: programming experience

"Certifying and Removing Disparate Impact" Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

Notation

features $\begin{cases} X: \text{protected feature} \\ Y: \text{other features} \end{cases}$

$\left.\right\}$ $X=0$ minority group; $X=1$ majority group

9, 10, 6

label $\{\, C$: binary outcome $\underbrace{(\text{hired, admitted})}_{C=1}$

if not: $C=0$

Disparate Impact

$$P(C=1 \mid X=0) \leq 0.8\, P(C=1 \mid X=1)$$

Idea: if we can predict X from Y

there could be disparate impact

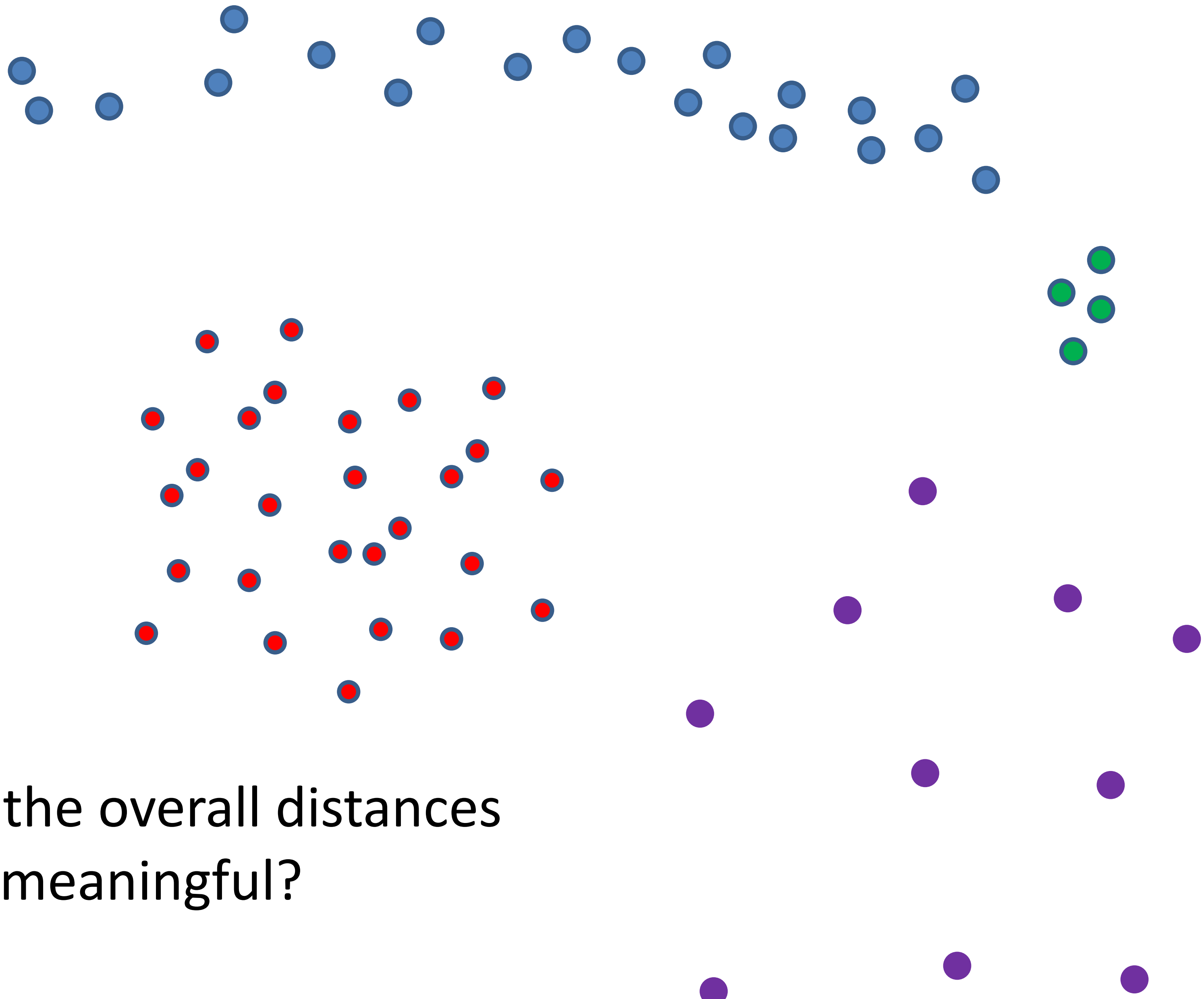
① train $f : \underbrace{Y}_{\substack{\text{input} \\ \text{(features)}}} \to \underbrace{X}_{\substack{\text{output} \\ \text{(label)}}}$ (f is a classifier)

② use to predict X → get $\underbrace{\text{BER}}_{\text{balanced error rate}}$

③ if BER is too low, could be disparate impact.

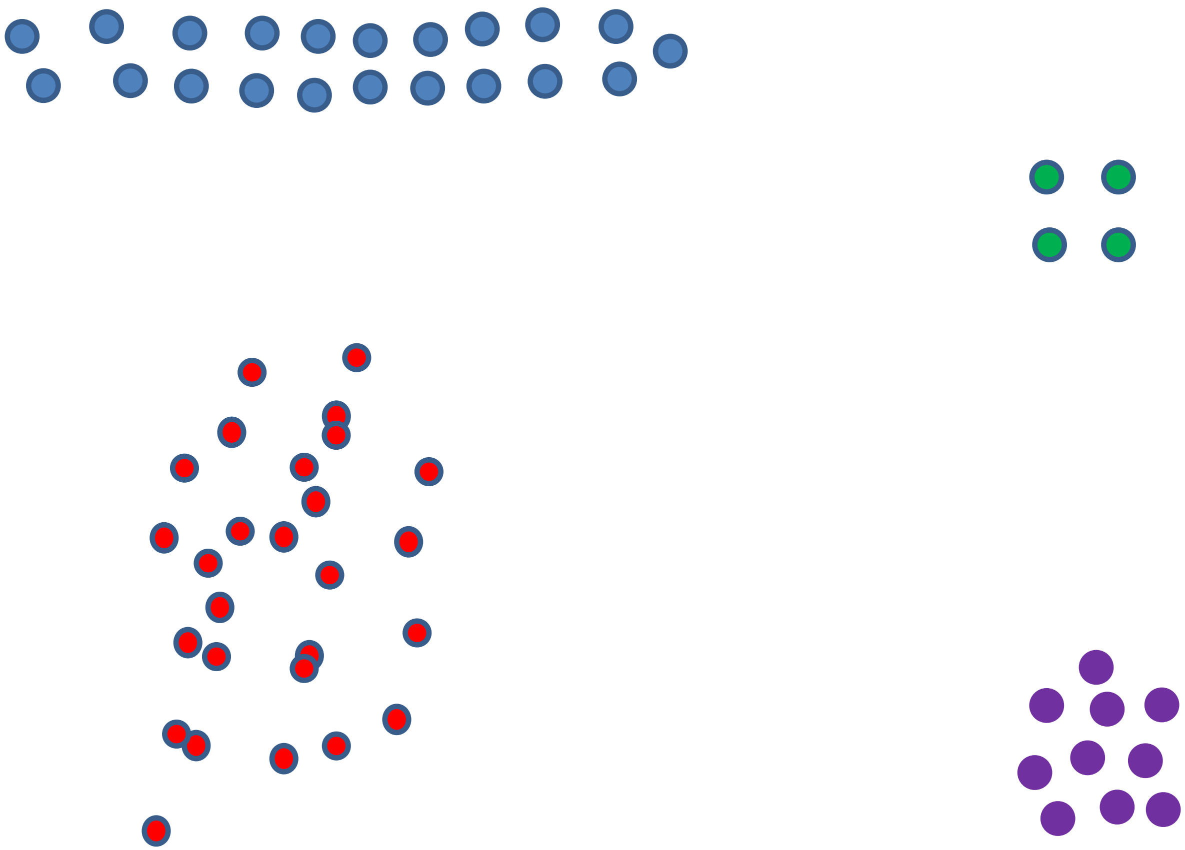
Bootstrap demo

See video tutorial on Piazza!



What if the overall distances are not meaningful?

Focus on your neighbors

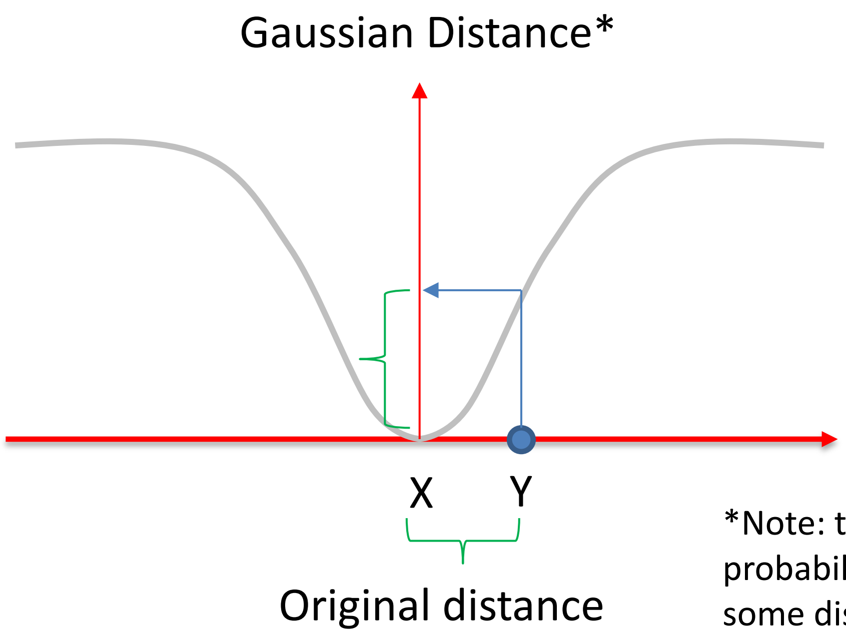


What if the overall distances
are not meaningful?

Focus on your neighbors

tSNE (t-distributed Stochastic Neighborhood Embedding)

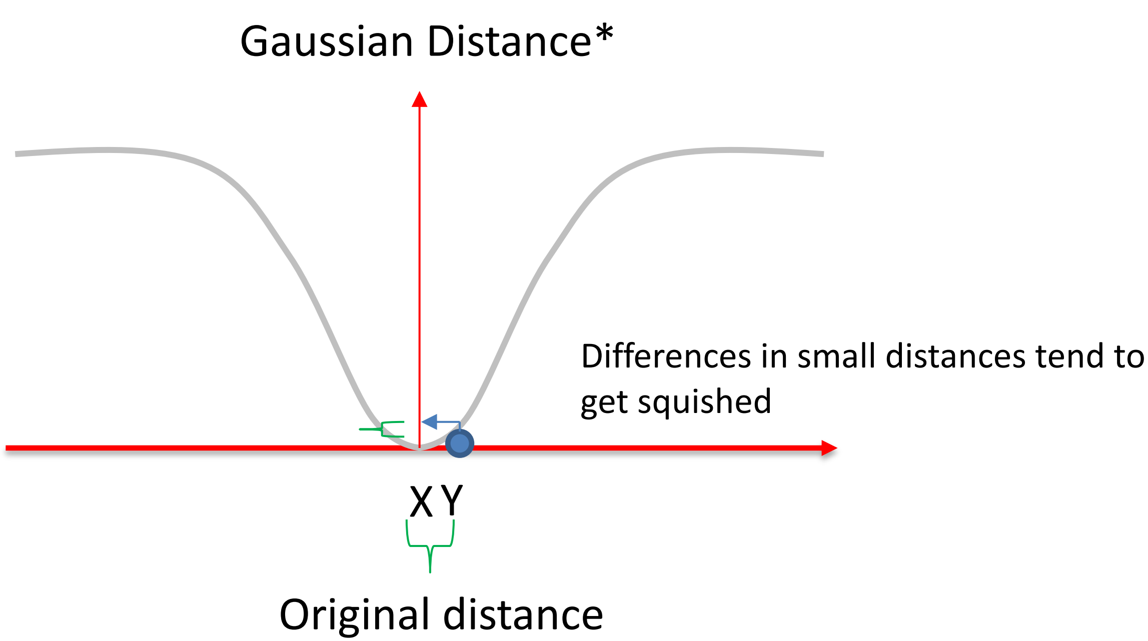
- Define distances between a point X to a point Y by a Gaussian function centered at X



*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)

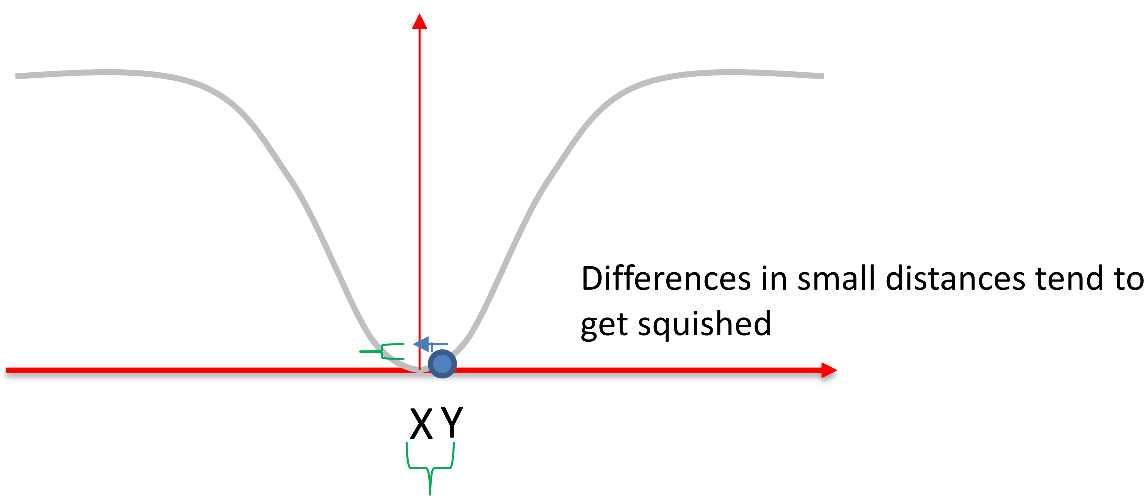
- Define distances between a point X to a point Y by a Gaussian function centered at X



tSNE (t-distributed Stochastic Neighborhood Embedding)

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

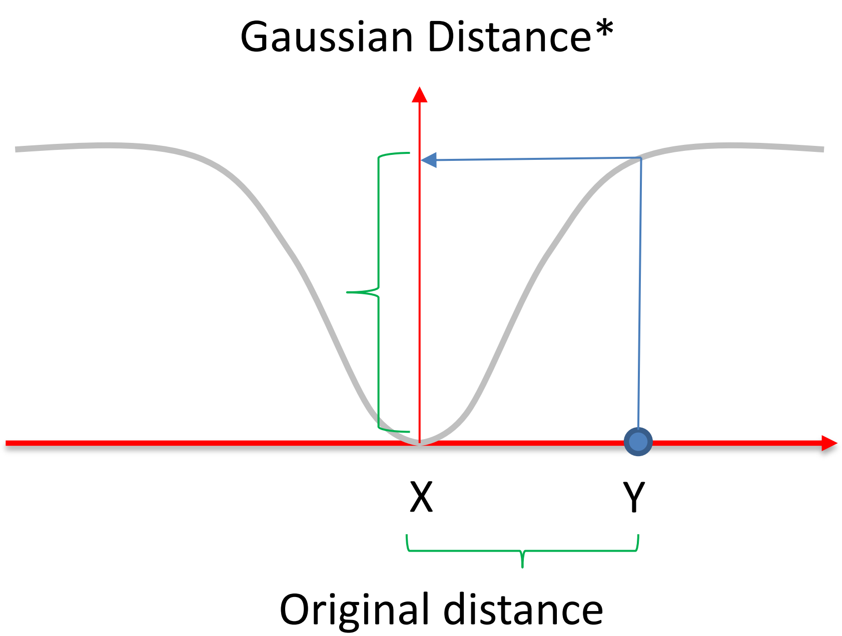
Gaussian Distance*



Original distance

tSNE (t-distributed Stochastic Neighborhood Embedding)

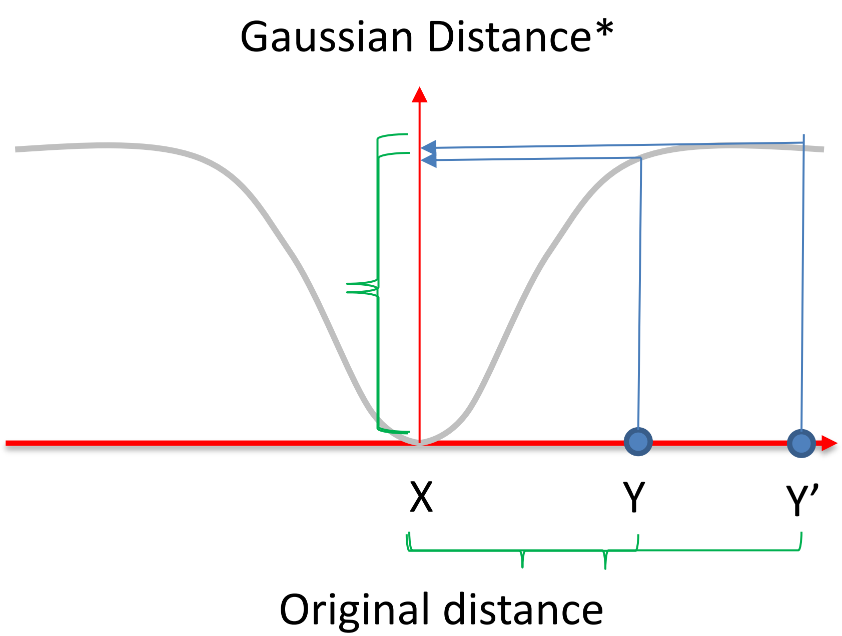
- Define distances between a point X to a point Y by a Gaussian function centered at X



Difference in large distances tend to get REALLY squished

tSNE (t-distributed Stochastic Neighborhood Embedding)

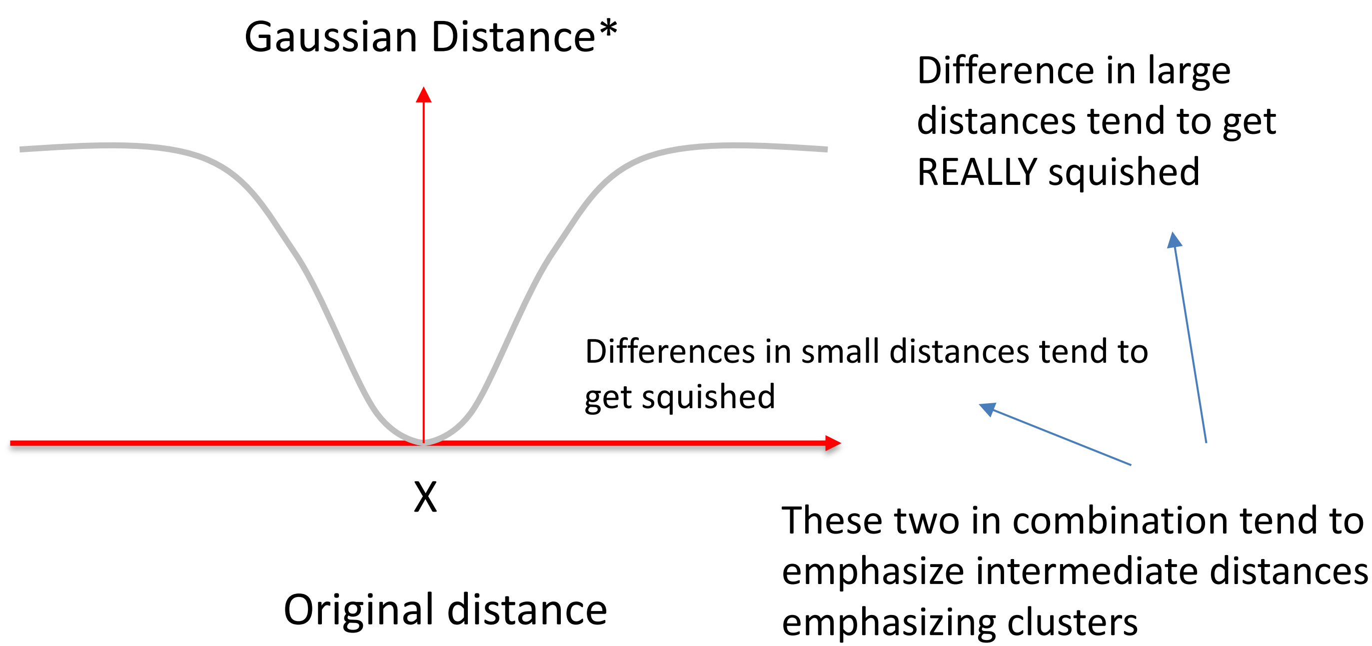
- Define distances between a point X to a point Y by a Gaussian function centered at X



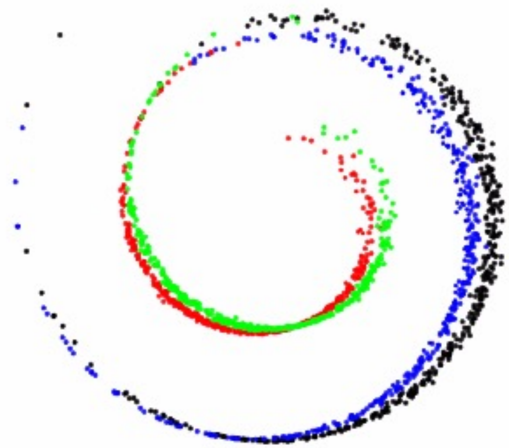
Difference in large distances tend to get REALLY squished

tSNE (t-distributed Stochastic Neighborhood Embedding)

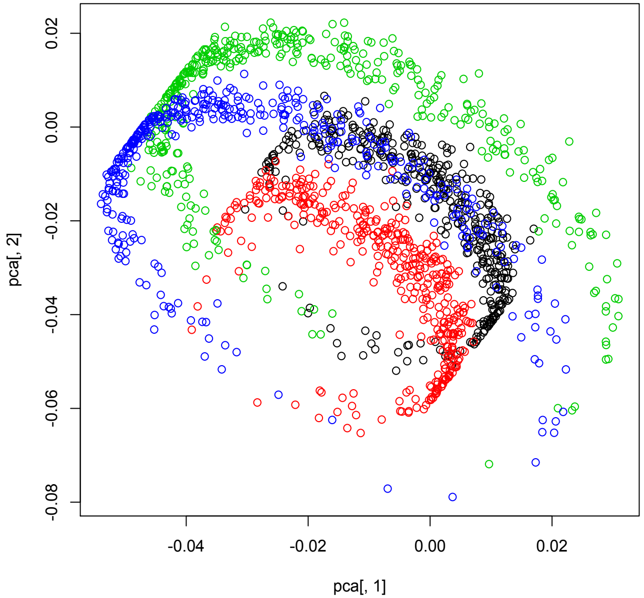
- Define distances between a point X to a point Y by a Gaussian function centered at X



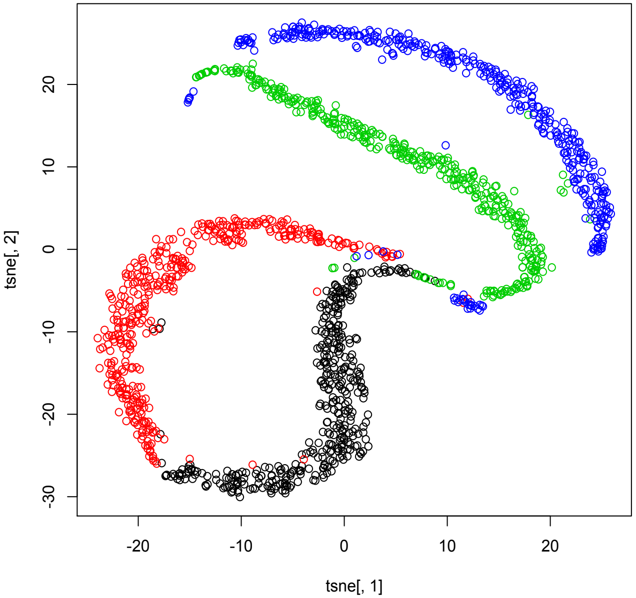
Original data



PCA

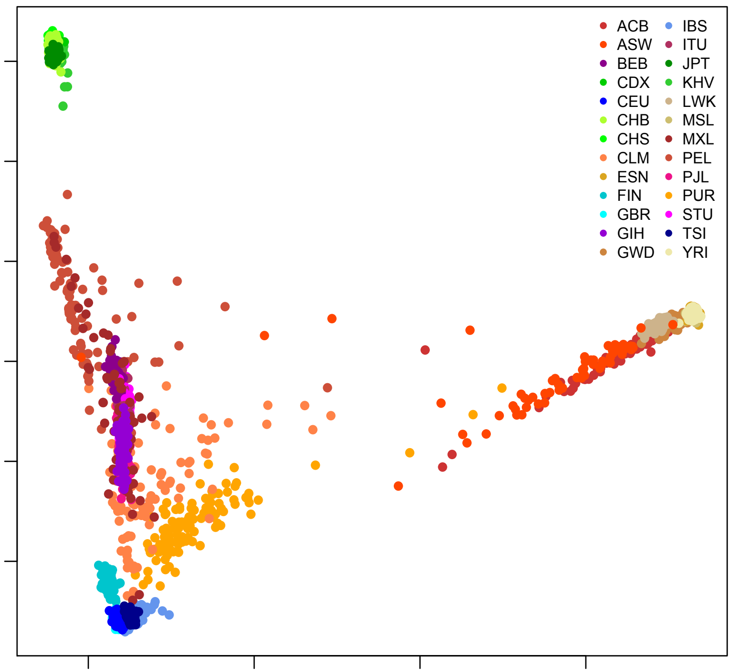


t-SNE

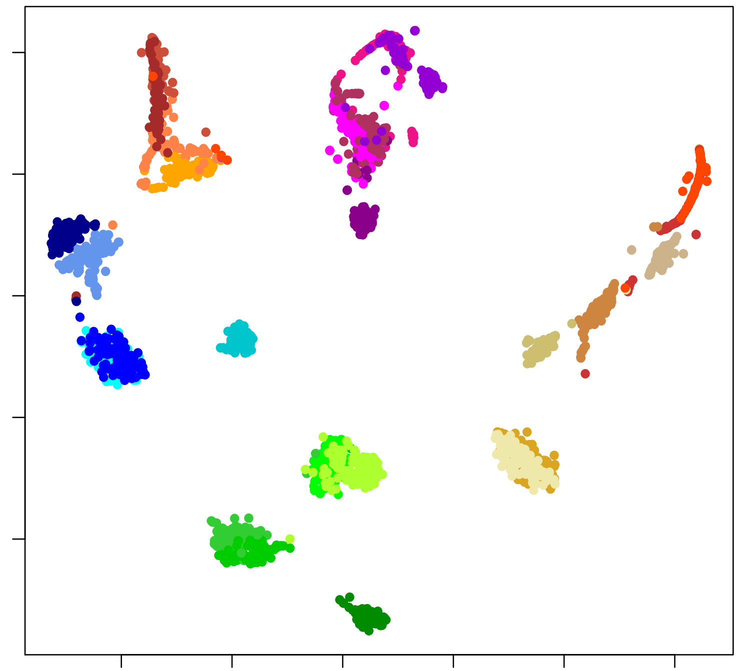


"swissroll data" **Dinoj Surendran**

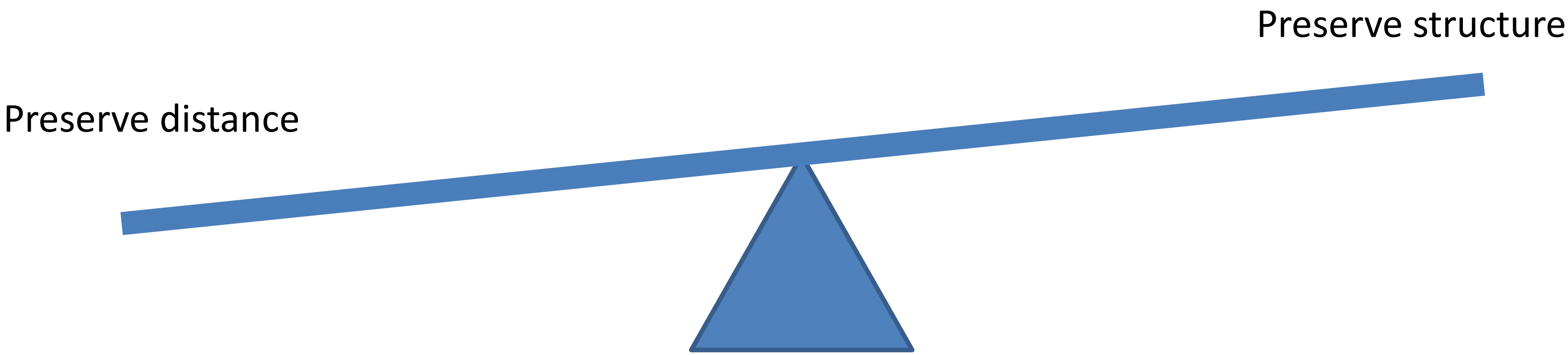
PCA



t-SNE



CHB	Han Chinese in Beijing, China	**MSL**	Mende in Sierra Leone
JPT	Japanese in Tokyo, Japan	**ESN**	Esan in Nigeria
CHS	Southern Han Chinese	**ASW**	Americans of African Ancestry in SW USA
CDX	Chinese Dai in Xishuangbanna, China	**ACB**	African Caribbeans in Barbados
KHV	Kinh in Ho Chi Minh City, Vietnam	**MXL**	Mexican Ancestry from Los Angeles USA
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	**PUR**	Puerto Ricans from Puerto Rico
TSI	Toscani in Italia	**CLM**	Colombians from Medellin, Colombia
FIN	Finnish in Finland	**PEL**	Peruvians from Lima, Peru
GBR	British in England and Scotland	**GIH**	Gujarati Indian from Houston, Texas
IBS	Iberian Population in Spain	**PJL**	Punjabi from Lahore, Pakistan
YRI	Yoruba in Ibadan, Nigeria	**BEB**	Bengali from Bangladesh
LWK	Luhya in Webuye, Kenya	**STU**	Sri Lankan Tamil from the UK
GWD	Gambian in Western Divisions in the Gambia	**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

-)

(9) $P(C=1 \mid X=0) \leq 0.8\, P(C=1 \mid X=1)$

get loan

Black

get loan

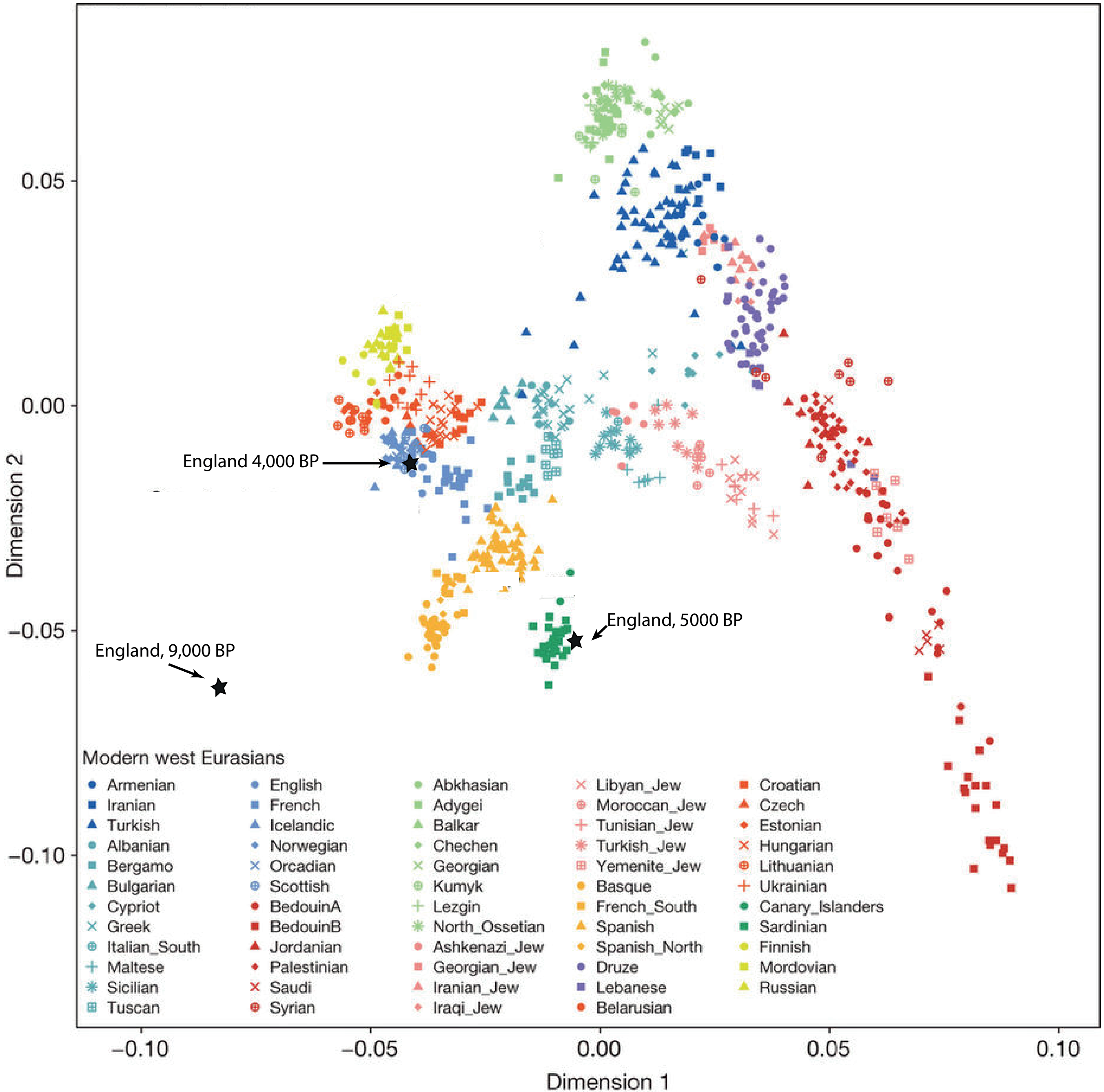
white

$0.275 \leq 0.8(0.35)$

$0.275 \leq 0.28$ ✓ $\Rightarrow$ disparate impact

...hted avg

... leaf entropies



Dimension 2
Dimension 1
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
England 4,000 BP
England, 5000 BP
England, 9,000 BP
Modern west Eurasians
Armenian
Iranian
Turkish
Albanian
Bergamo
Bulgarian
Cypriot
Greek
Italian_South
Maltese
Sicilian
Tuscan
English
French
Icelandic
Norwegian
Orcadian
Scottish
BedouinA
BedouinB
Jordanian
Palestinian
Saudi
Syrian
Abkhasian
Adygei
Balkar
Chechen
Georgian
Kumyk
Lezgin
North_Ossetian
Ashkenazi_Jew
Georgian_Jew
Iranian_Jew
Iraqi_Jew
Libyan_Jew
Moroccan_Jew
Tunisian_Jew
Turkish_Jew
Yemenite_Jew
Basque
French_South
Spanish
Spanish_North
Druze
Lebanese
Belarusian
Croatian
Czech
Estonian
Hungarian
Lithuanian
Ukrainian
Canary_Islanders
Sardinian
Finnish
Mordovian
Russian

P = pos

N = neg

H = healthy

D = disease

$P(D \mid P) = 0.8$

$$P(N \mid H) = P(P \mid D) = \frac{P(P)\,P(D \mid P)}{P(D)} = \frac{P(P) \cdot 0.8}{\frac{1}{500}} = x$$

$$\boxed{\begin{aligned} P(P) &= P(P, D) + P(P, H) \\ &= P(D)P(P \mid D) + P(H)P(P \mid H) \\ &= \frac{1}{500}x + \frac{499}{500}\Big(1 - \underbrace{P(N \mid H)}_{x}\Big) \end{aligned}}$$

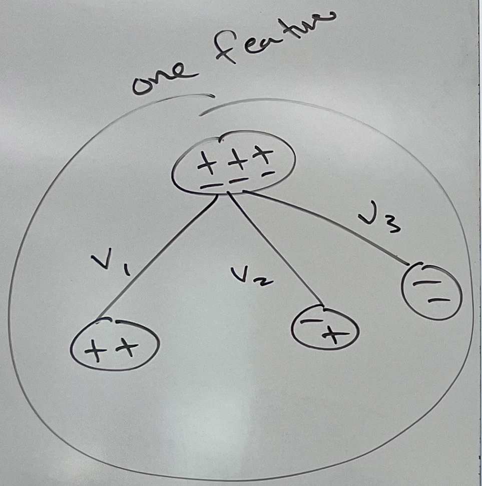
$$x = \frac{\left[\frac{1}{\cancel{500}}x + \frac{499}{\cancel{500}}(1-x)\right]0.8}{\frac{1}{\cancel{500}}}$$

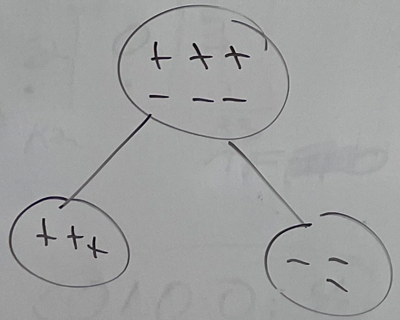
$$x = (x + 499(1-x))\,0.8$$

$$\frac{5}{4}x = x + 499 - 499x$$

$$\frac{5}{4}x = 499 - 498x$$

$$\boxed{x \approx 0.9995}$$





9

$$H(Y) = -\sum_{c \in vals(Y)} p(y=c) \log_2 p(y=c)$$

$$= - p(y=+) \log_2 p(y=+) - p(y=-) \log_2 p(y=-)$$

$$= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$$

$$= + \frac{1}{2} + \frac{1}{2} = 1$$

$$H(Y|X) = \sum_{v} p(X=v)\, H(Y|X=v)$$ weighted avg of leaf entropies

one feature

$$H(Y|\underbrace{X=v}) = -\sum_{c} p(y=c|x=v) \log_2 p(y=c|x=v)$$