# CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD COLLEGE

# Admin

- **Midterm 1** due today!
- **No lab today**

- After Thanksgiving break
  - 3 classes on advanced Data Science topics
  - 3 classes for project presentations
  - Final project check-ins during lab

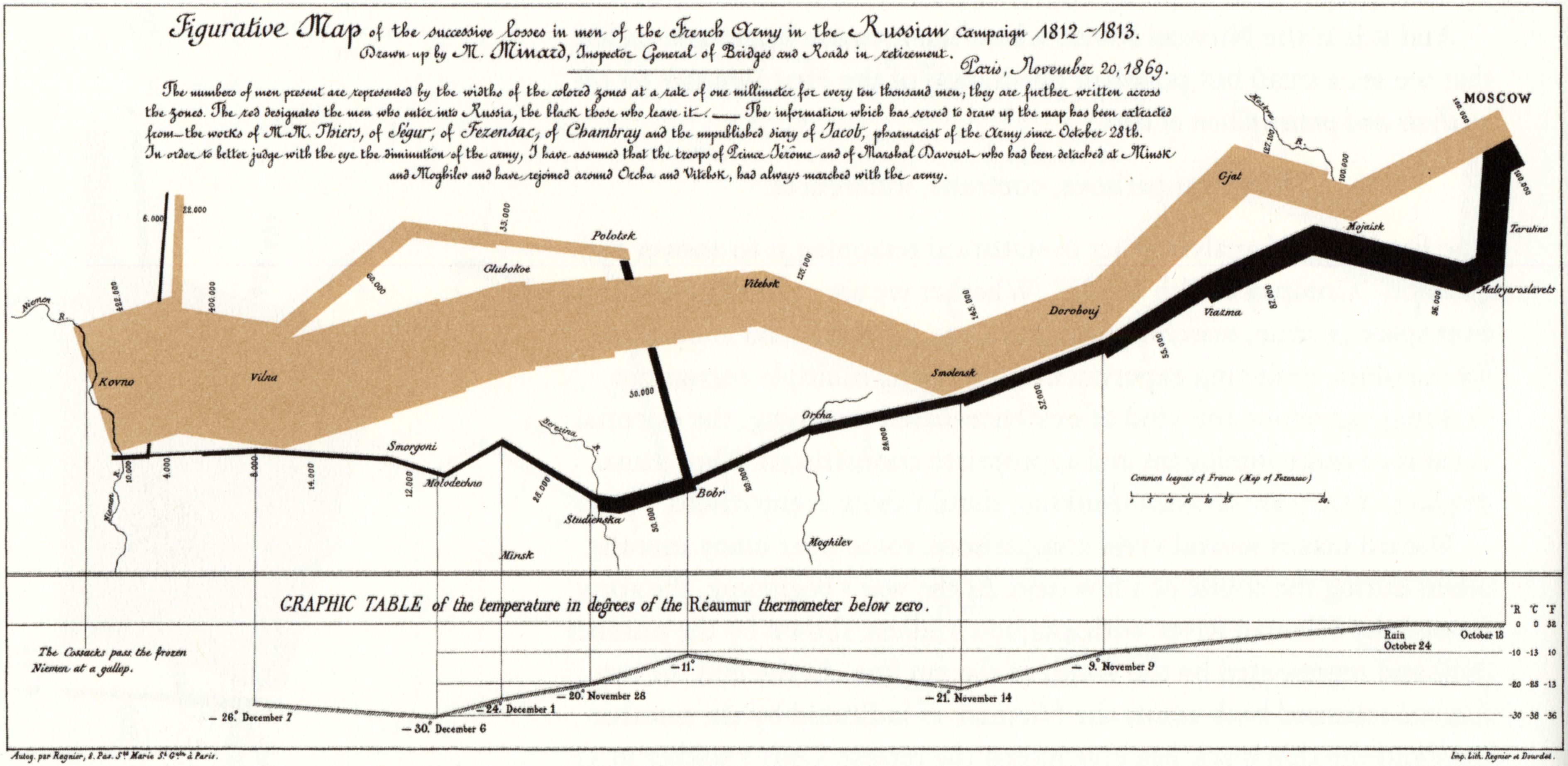
# Outline for November 21

- Revisit data visualization
- Real-world data science exercise
- Begin: clustering (K-means)

# Outline for November 21

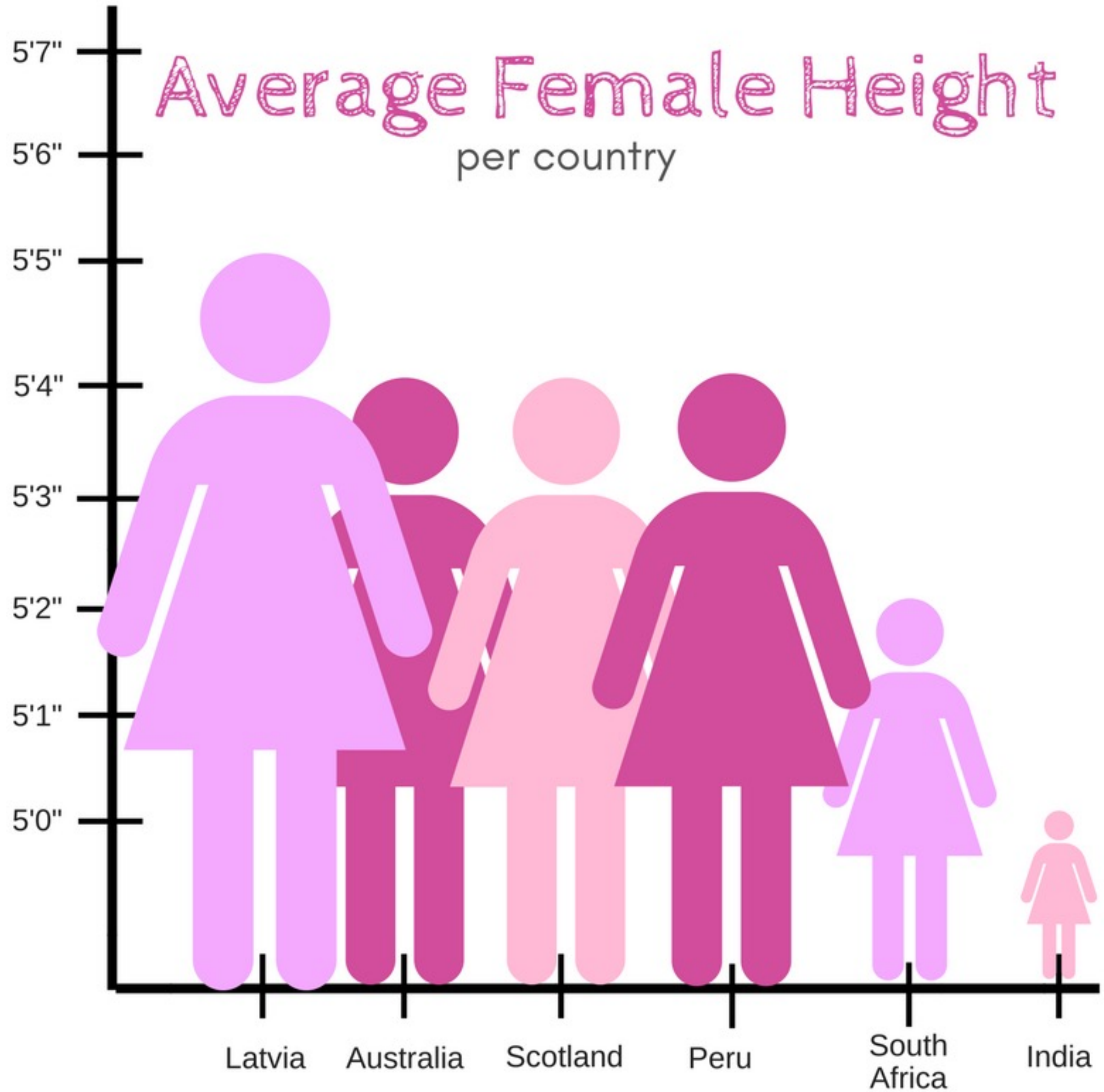
- Revisit data visualization
- Real-world data science exercise
- Begin: clustering (K-means)

# Visualization can illuminate…

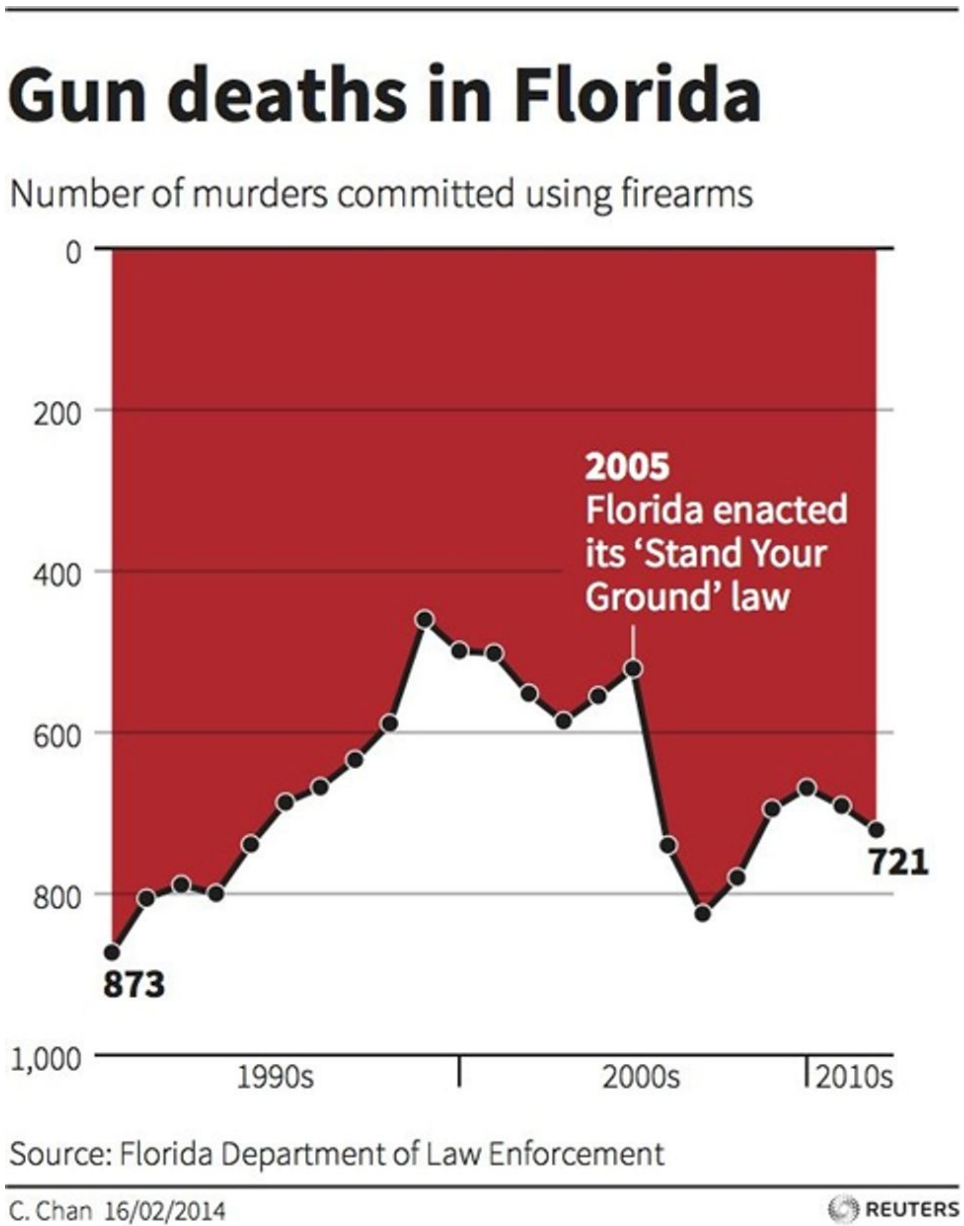



Size of Napoleon's army on the advance (in tan) and retreat (in black) from Moscow in 1812

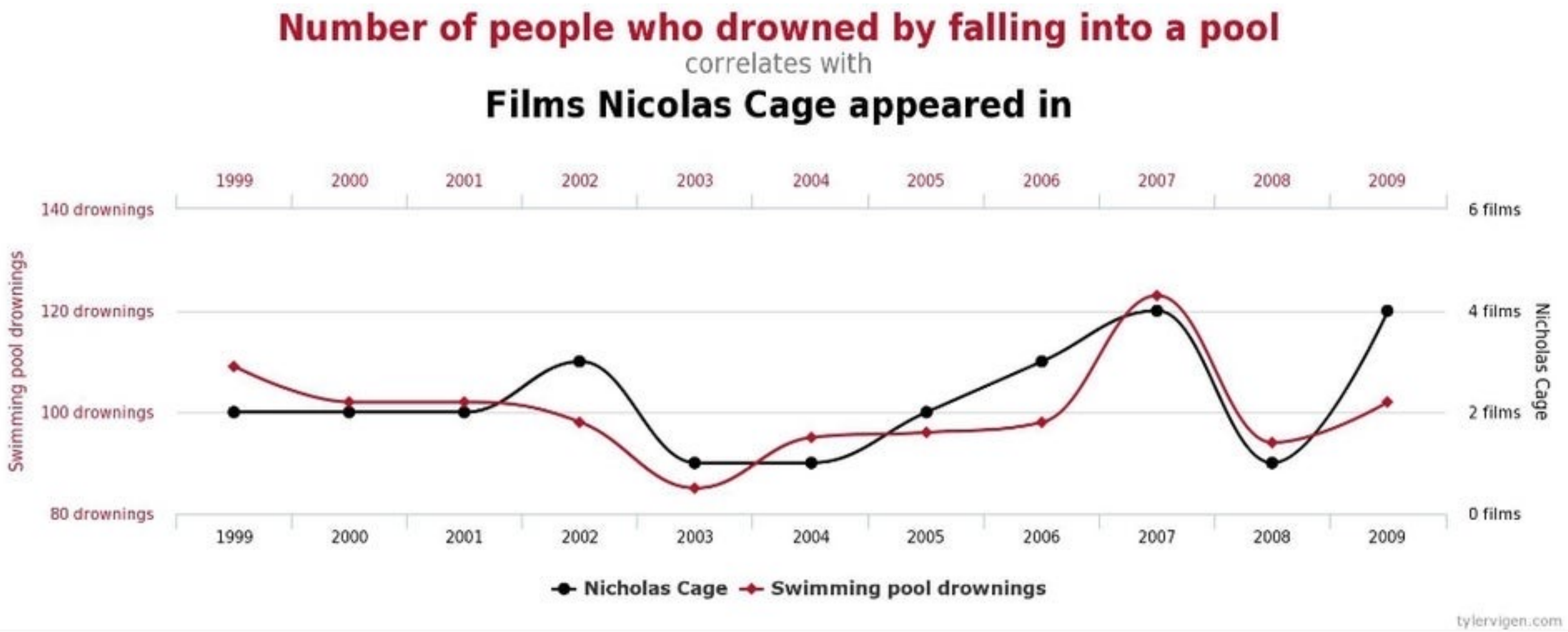
# … but also mislead



# … but also mislead



# … but also mislead



# Visualizing amounts


Figure from "Fundamentals of Data Visualization" by Claus Wilke

# Visualizing amounts


Figure from “Fundamentals of Data Visualization” by Claus Wilke

# Visualizing distributions

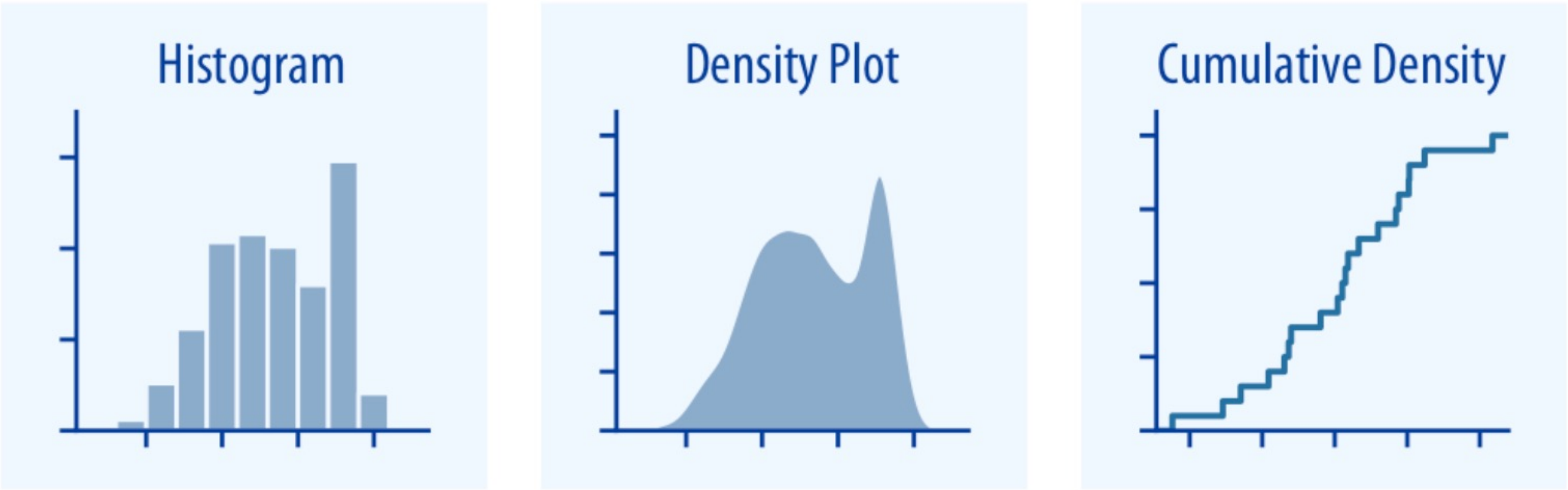



Figure from “Fundamentals of Data Visualization” by Claus Wilke

# Visualizing distributions

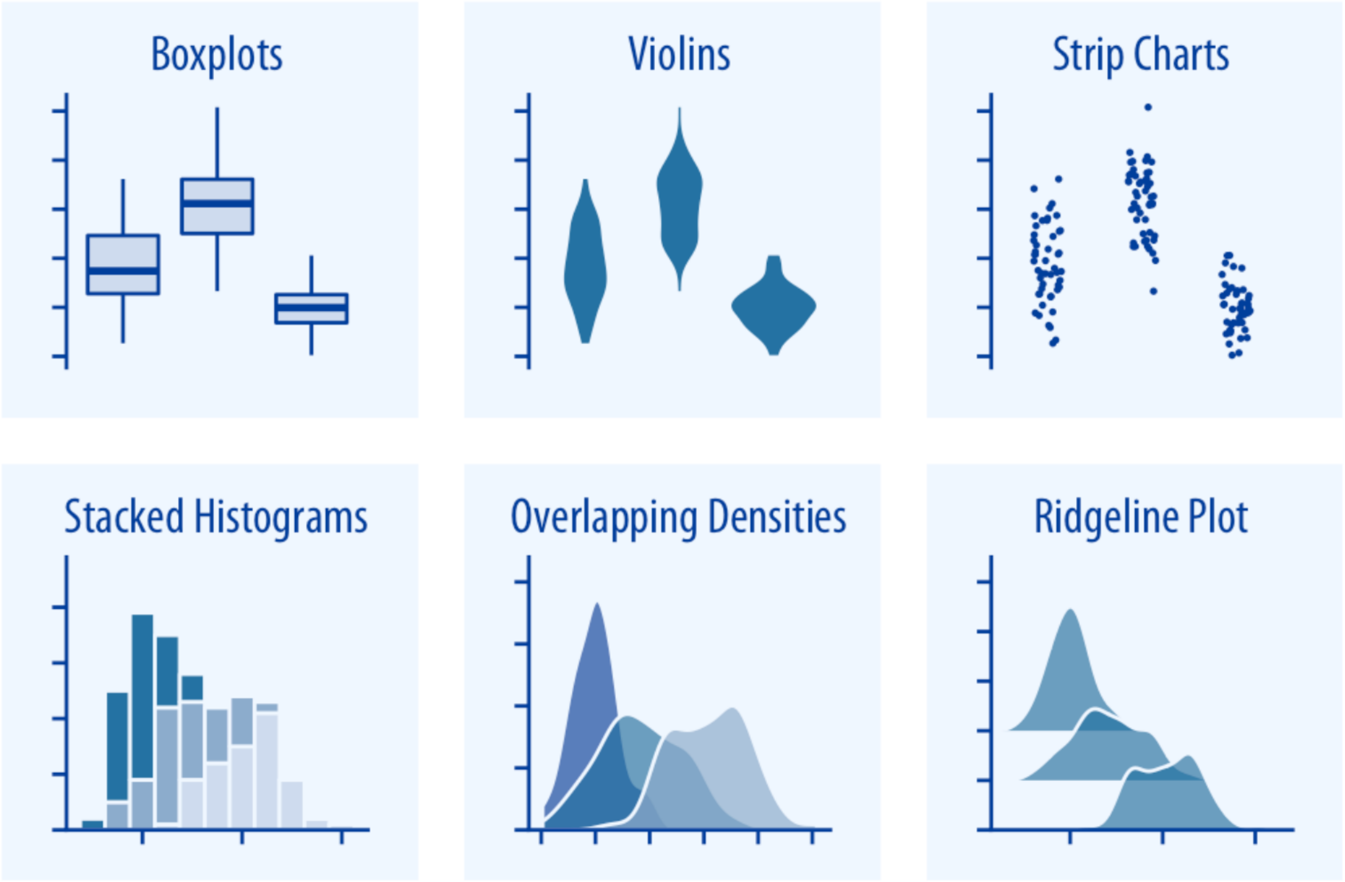
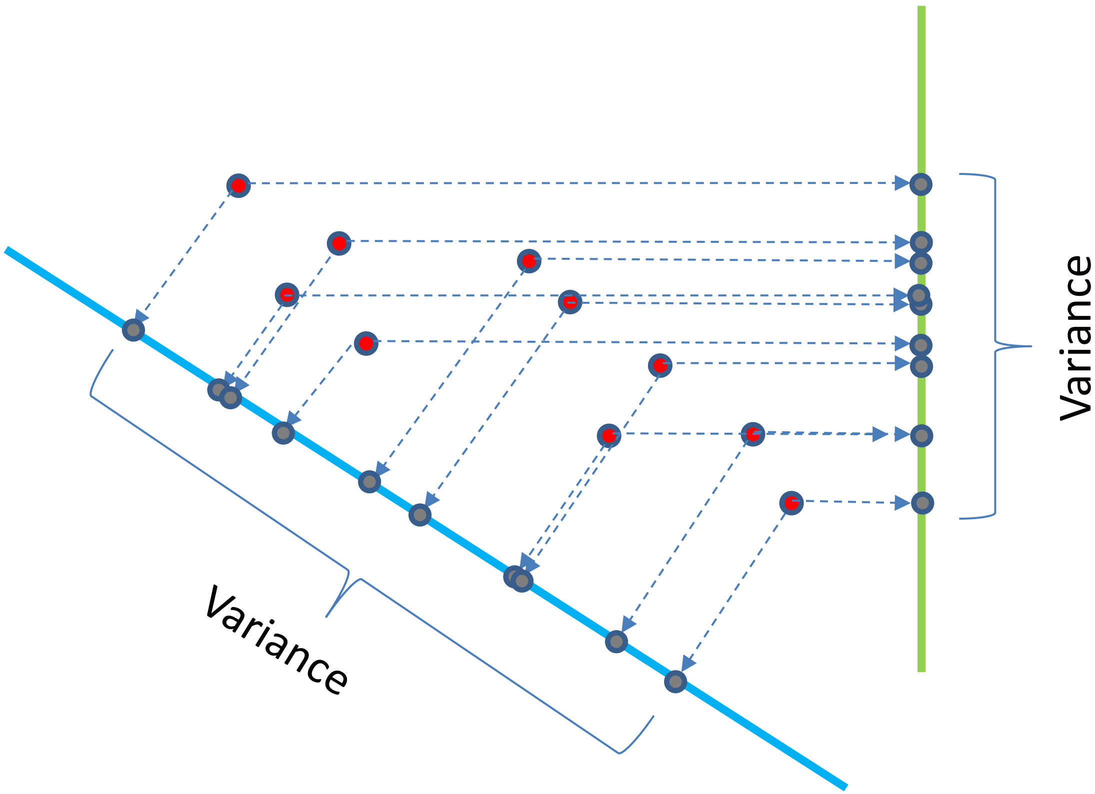



Figure from “Fundamentals of Data Visualization” by Claus Wilke

# Alternative to PCA

# Reducing dimensions

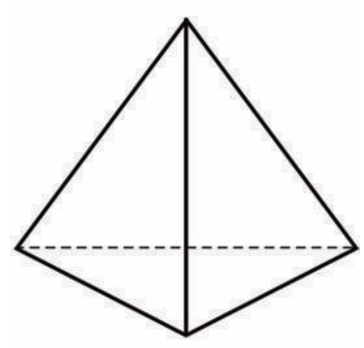
- How?
  - Project the points from high-dimensions to low dimensions



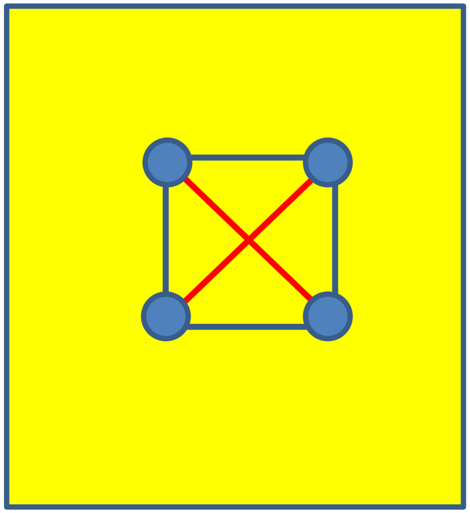

Prefer the blue line because more spread of the original data is represented→ Principal Component Analysis (**PCA**)

# Reducing dimensions

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



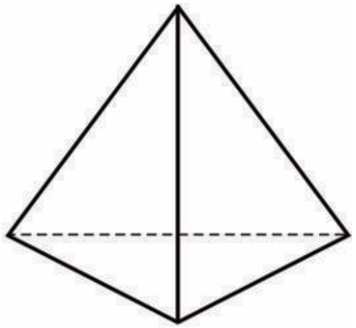
Tetrahedron with length 1 sides.
All pairwise distances between the four points = 1



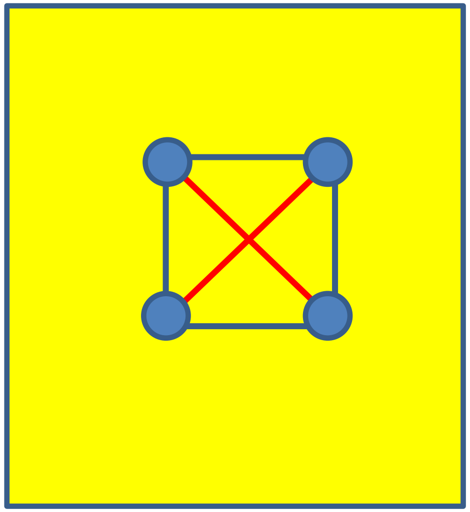
Try to arrange four points in 2D such that pairwise distances are as closest to the original pairwise distances

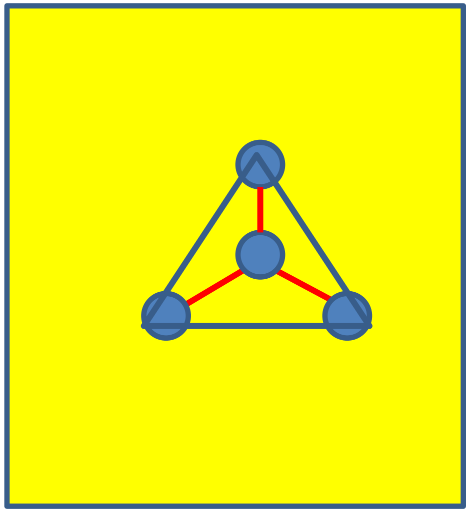
# Reducing dimensions

- 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



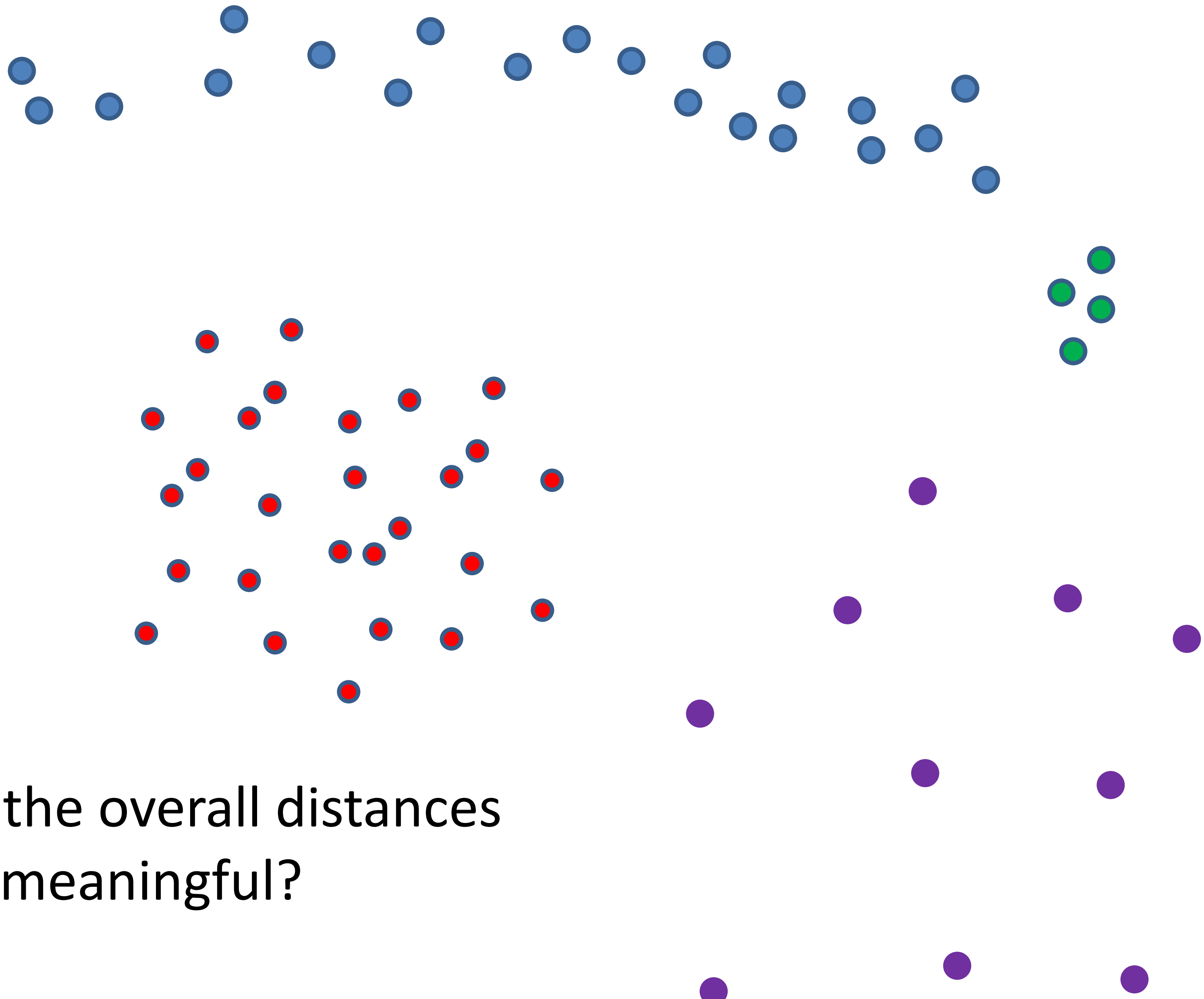


A lot of the time we want to create clusters.

Distances in the original data may not be meaningful

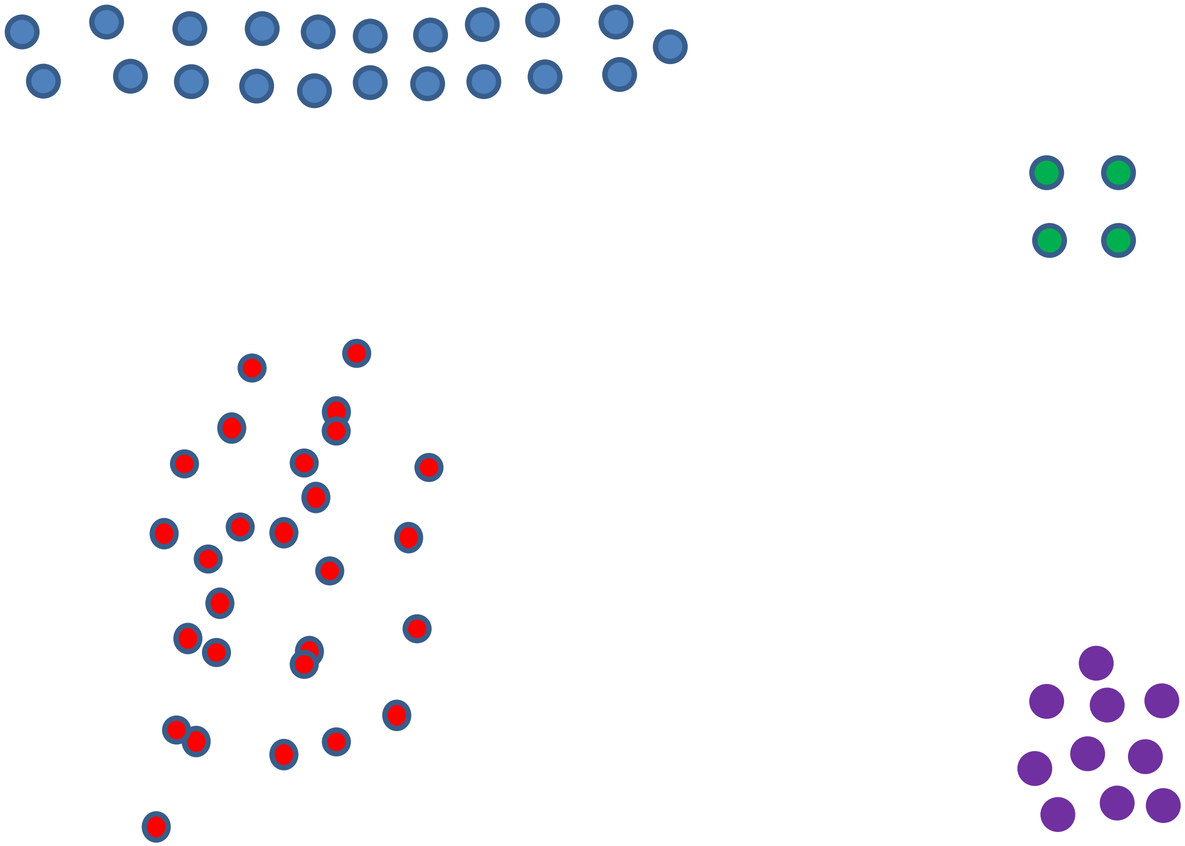
So we want some kind of embedding that preserves clustering

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



What if the overall distances are not meaningful?

Focus on your neighbors

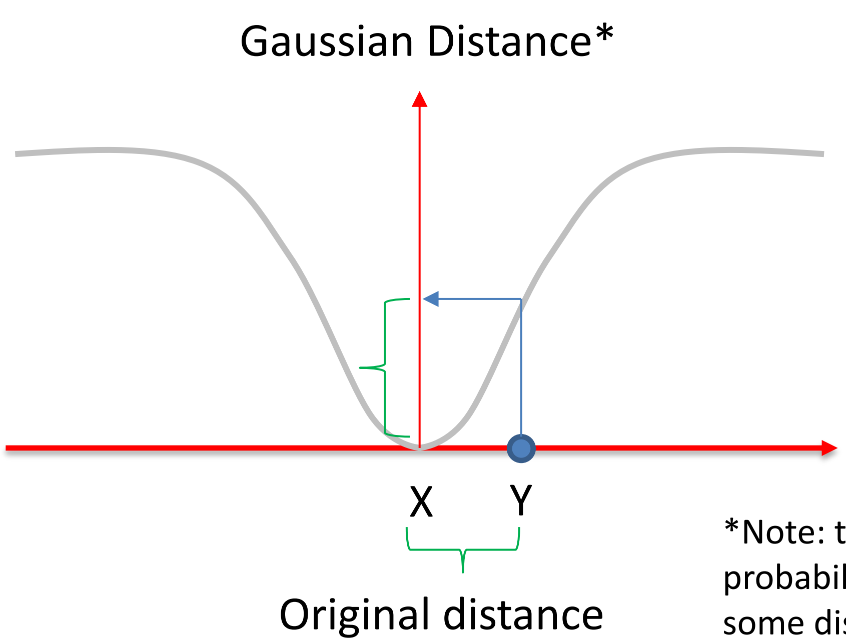


What if the overall distances are not meaningful?

Focus on your neighbors

# tSNE (t-distributed Stochastic Neighborhood Embedding)

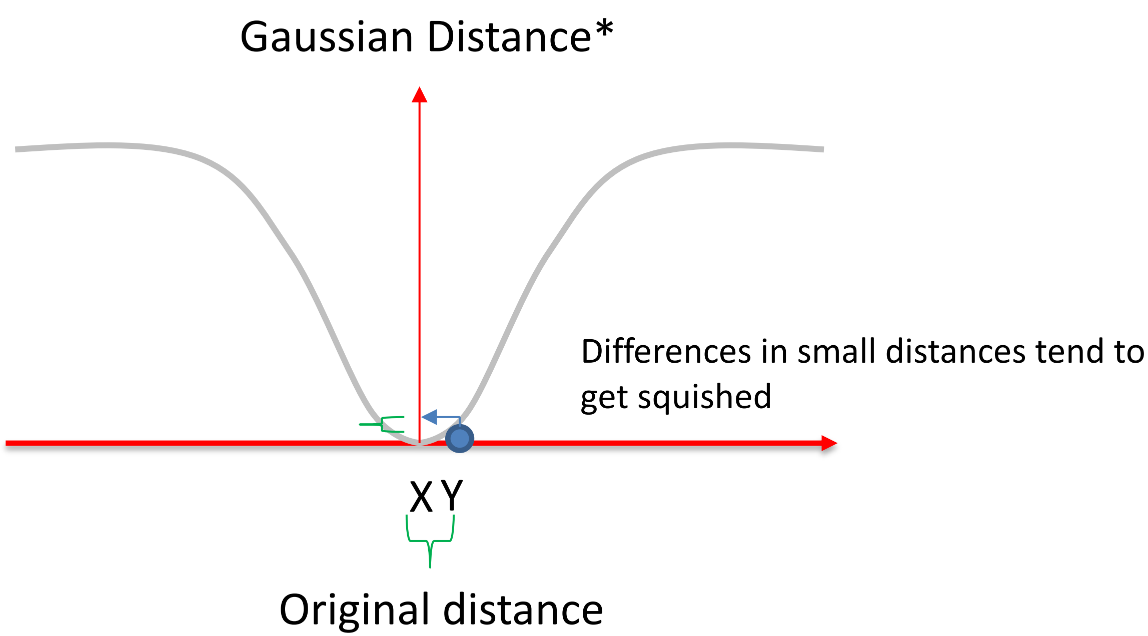
- Define distances between a point X to a point Y by a Gaussian function centered at X




*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

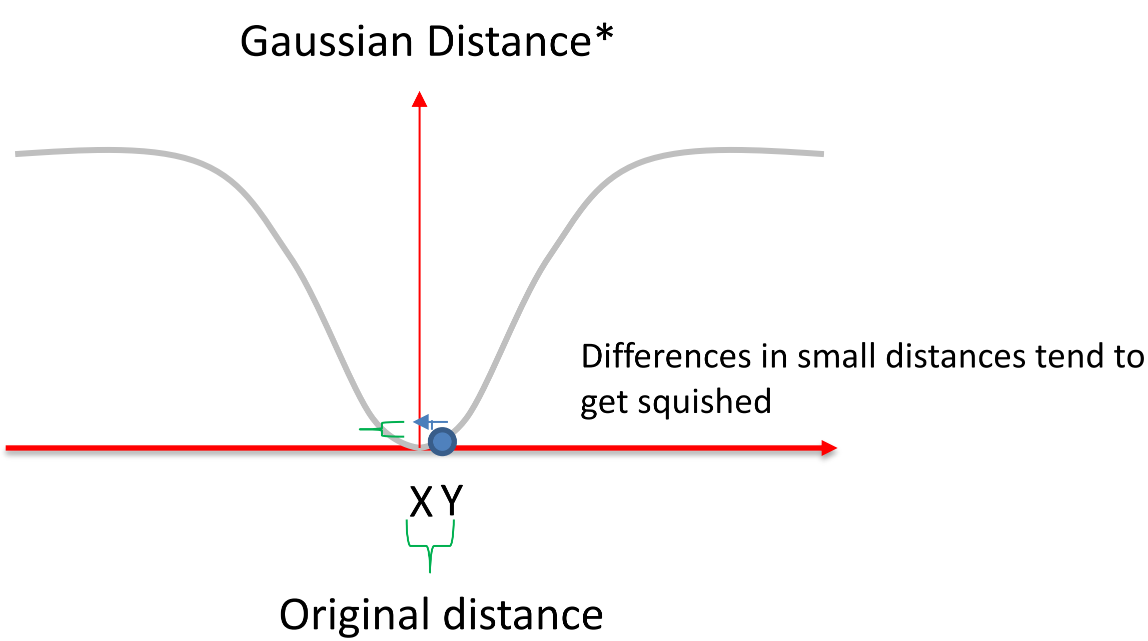
# tSNE (t-distributed Stochastic Neighborhood Embedding)

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



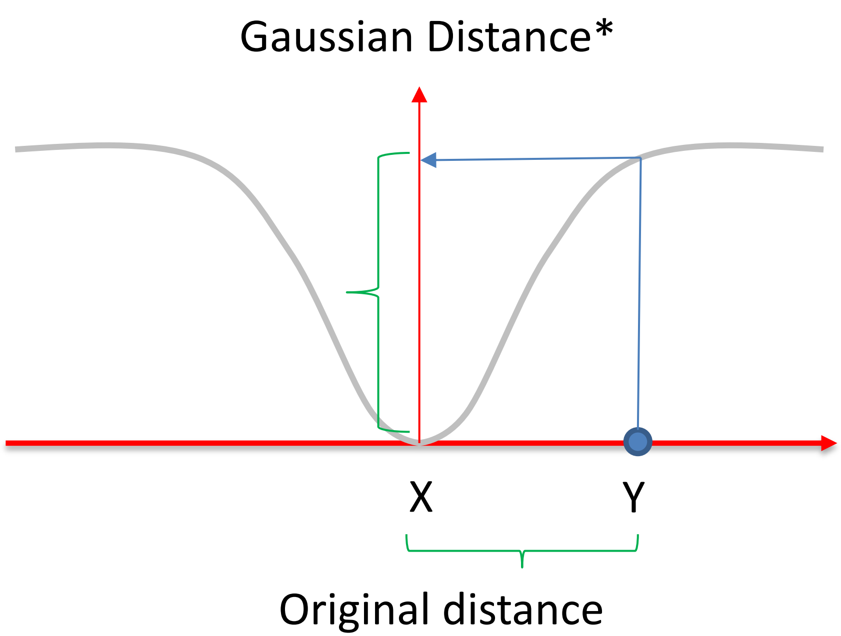
# tSNE (t-distributed Stochastic Neighborhood Embedding)

- 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

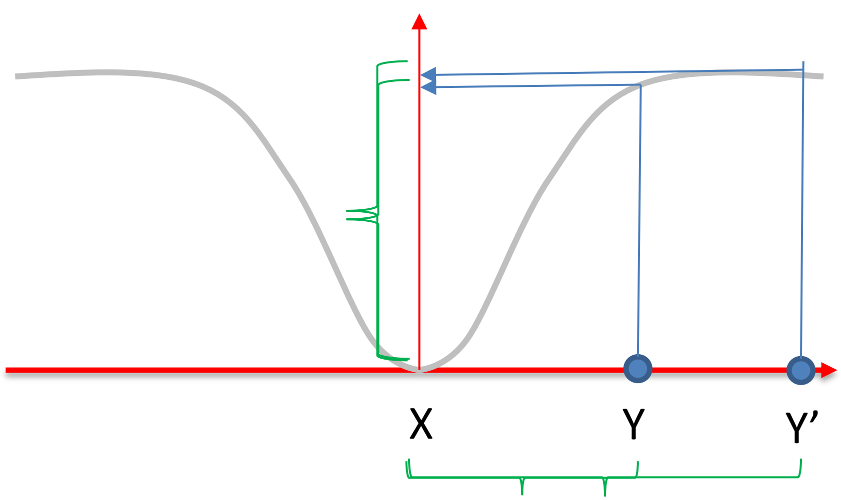



Difference in large distances tend to get REALLY squished

# tSNE (t-distributed Stochastic Neighborhood Embedding)

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

Gaussian Distance*

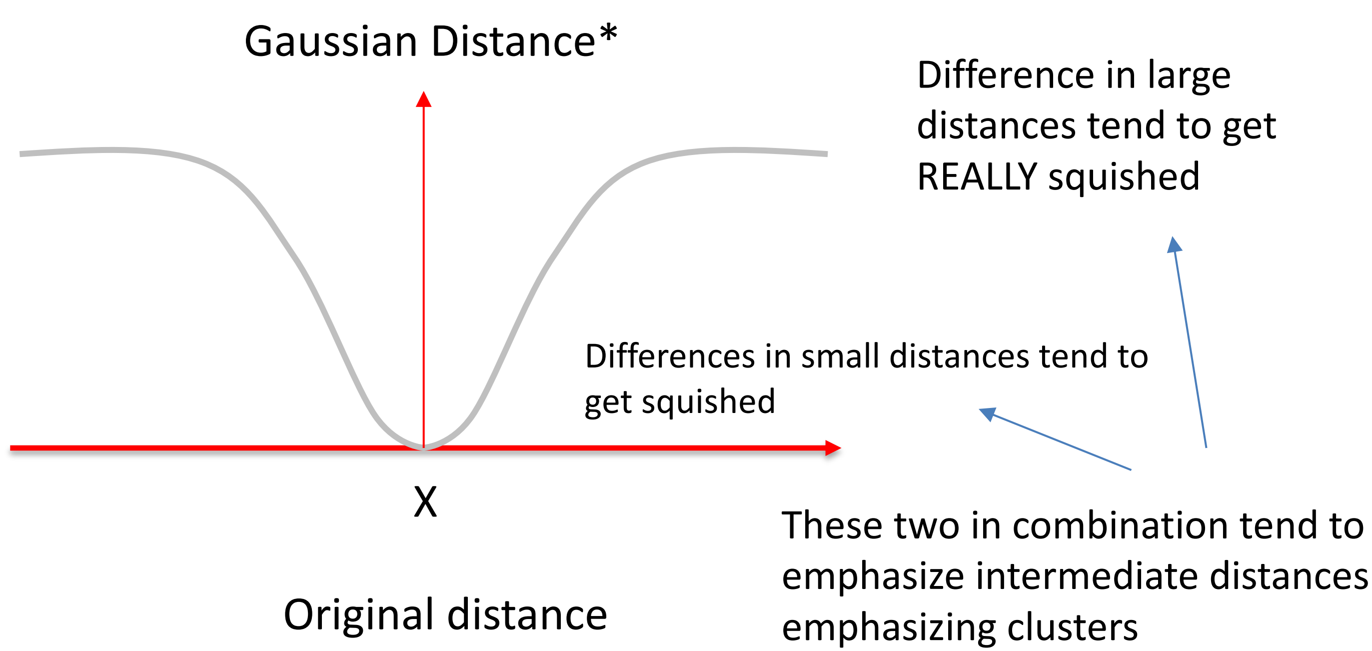



Difference in large distances tend to get REALLY squished

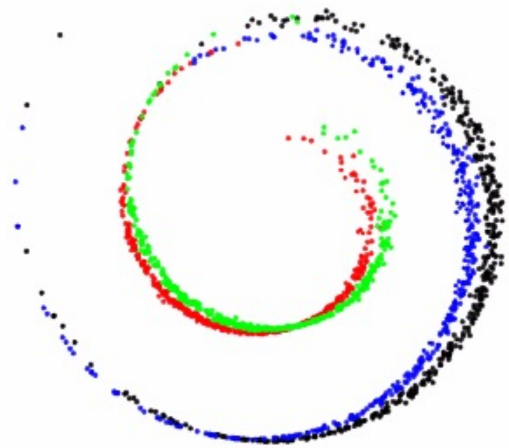
Original distance

# tSNE (t-distributed Stochastic Neighborhood Embedding)

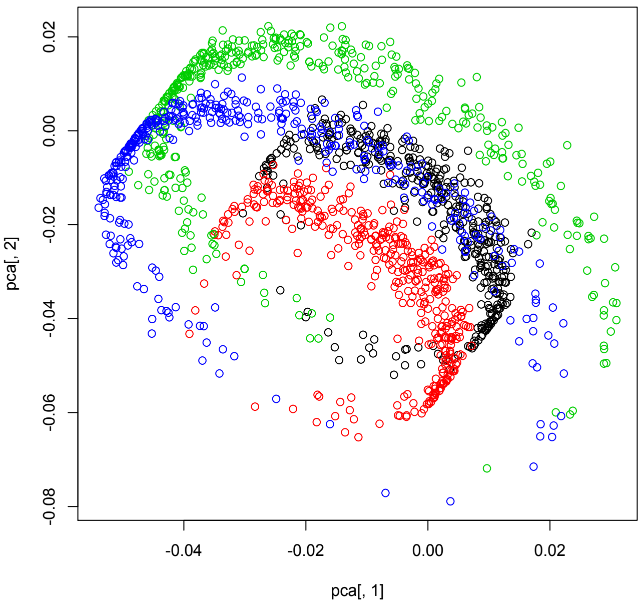
- Define distances between a point X to a point Y by a Gaussian function centered at X



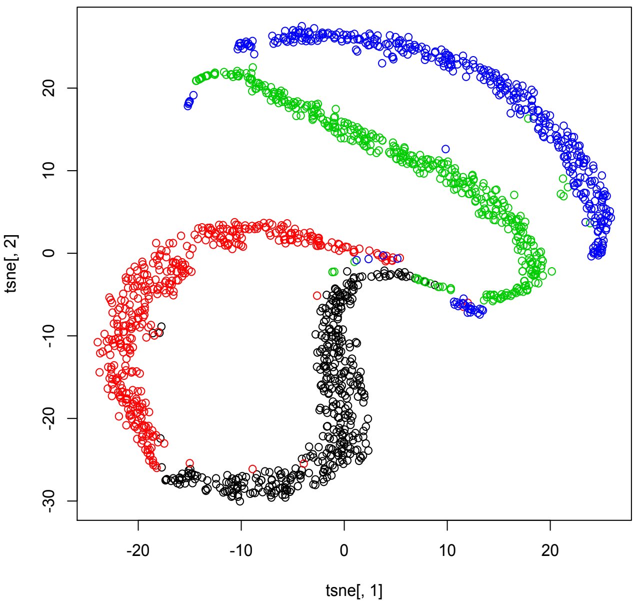
# Original data



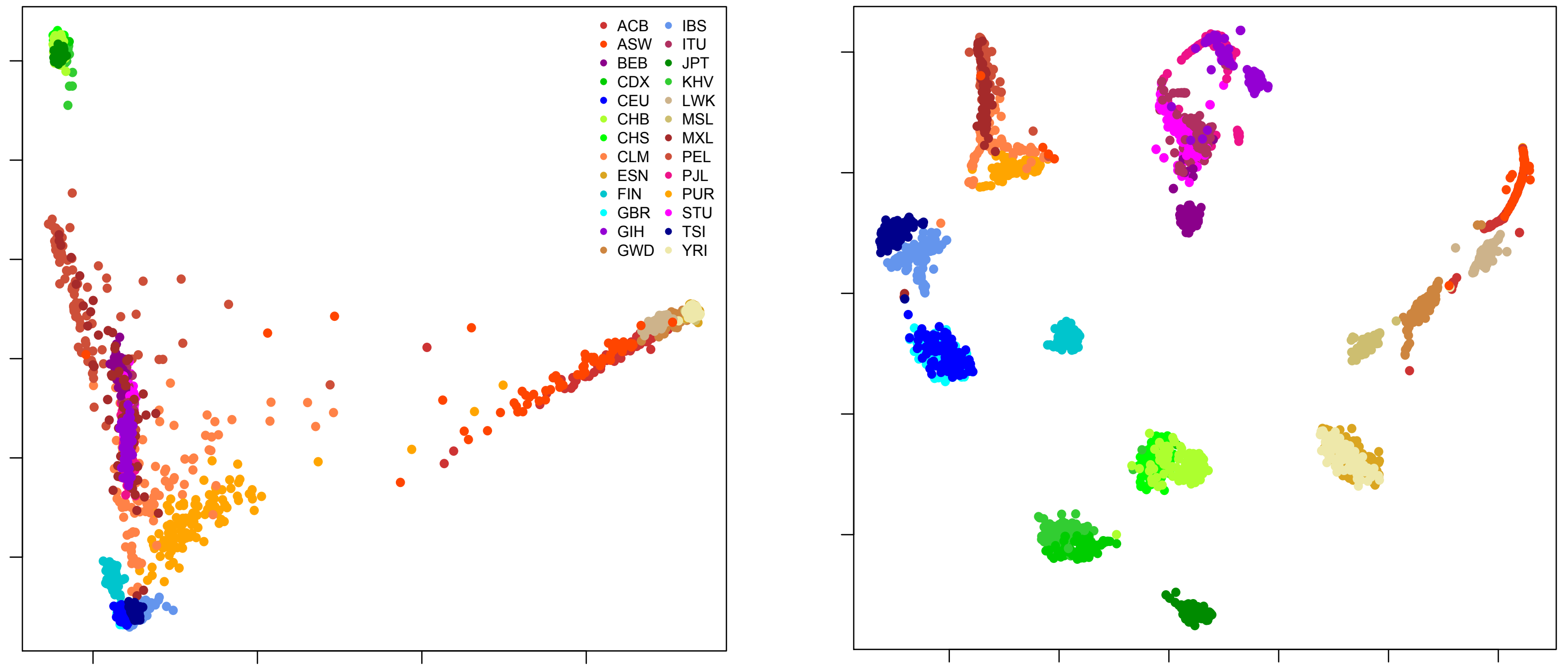
## PCA




## t-SNE

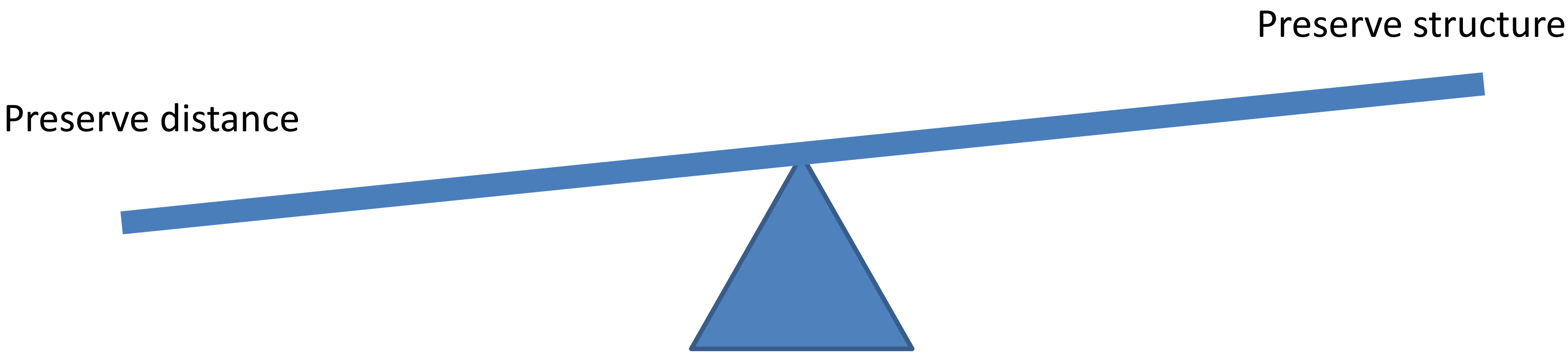



"swissroll data" **Dinoj Surendran**




| | |
|---|---|
| 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

# Outline for November 21

- Revisit data visualization
- Real-world data science exercise
- Begin: clustering (K-means)

# Discussion: admissions at Haverford

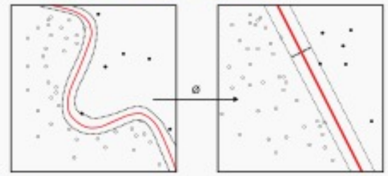
- Haverford has suddenly started receiving 10x more applications than usual
- You are tasked with creating an algorithm to determine whether or not an applicant should be admitted
- Questions:
  - How would you encode features?
  - How would you use past admission data to train?
  - What loss function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

# Outline for November 21

- Revisit data visualization
- Real-world data science exercise
- Begin: clustering (K-means)

**Supervised Learning:**
makes use of examples where we know the underlying "truth" (label/output)

**Machine learning and data mining**



**Problems** [show]

**Supervised learning** [hide]
(classification • regression)

Decision trees • Ensembles (Bagging, Boosting, Random forest) • *k*-NN • Linear regression • Naive Bayes • Neural networks • Logistic regression • Perceptron • Relevance vector machine (RVM) • Support vector machine (SVM)

**Clustering** [hide]

BIRCH • **Hierarchical** • *k*-means • Expectation-maximization (EM) • DBSCAN • OPTICS • Mean-shift

**Dimensionality reduction** [hide]

Factor analysis • CCA • ICA • LDA • NMF • PCA • t-SNE

**Structured prediction** [hide]

Graphical models (Bayes net, CRF, HMM)

**Anomaly detection** [hide]

*k*-NN • Local outlier factor

**Neural nets** [hide]

Autoencoder • Deep learning • Multilayer perceptron • RNN • Restricted Boltzmann machine • SOM • Convolutional neural network

**Reinforcement Learning** [hide]

Q-Learning • SARSA • Temporal Difference (TD)

**Theory** [show]

**Machine learning venues** [show]

**Machine learning portal**

V • T • E

**Unsupervised Learning:**
Learn underlying structure or features without labeled training data

Image: wikipedia

# Unsupervised learning: 3 main areas

1) Clustering: group data points into clusters based on features only

2) Dimensionality reduction: remove feature correlation, compress data, visualize data

3) Structured prediction: model latent variables (example: Hidden Markov Models)

# Unsupervised learning examples from biology: clustering

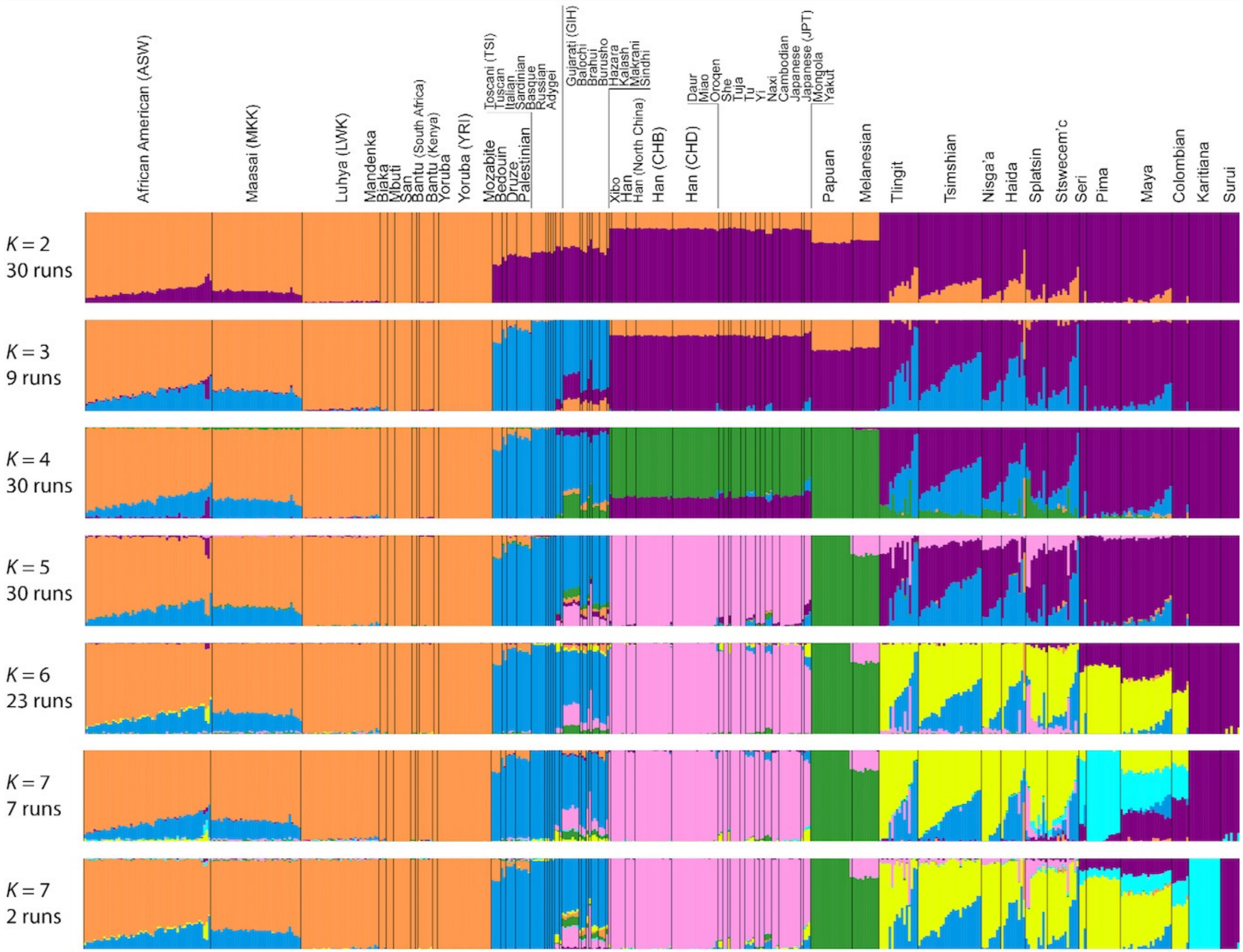
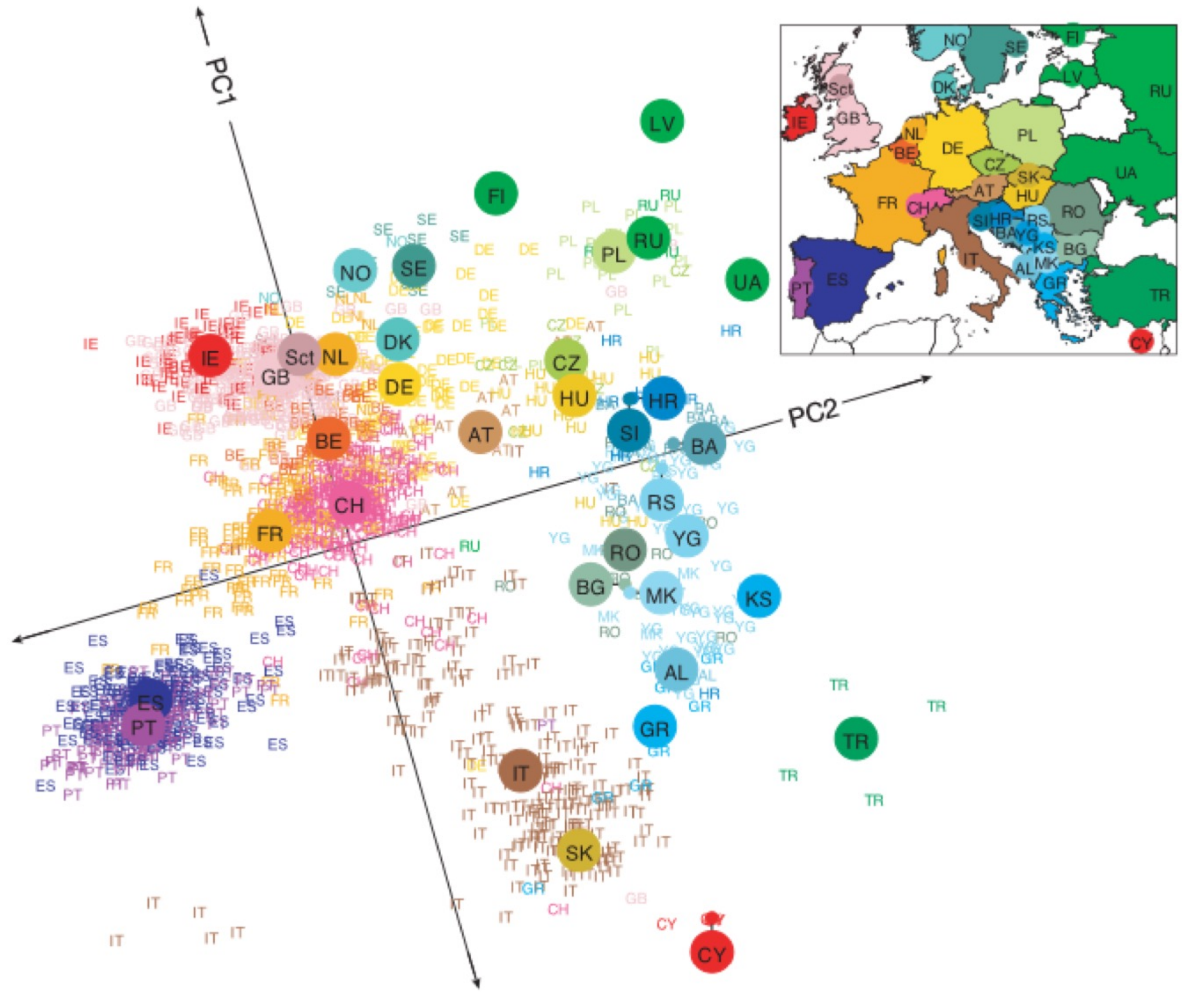


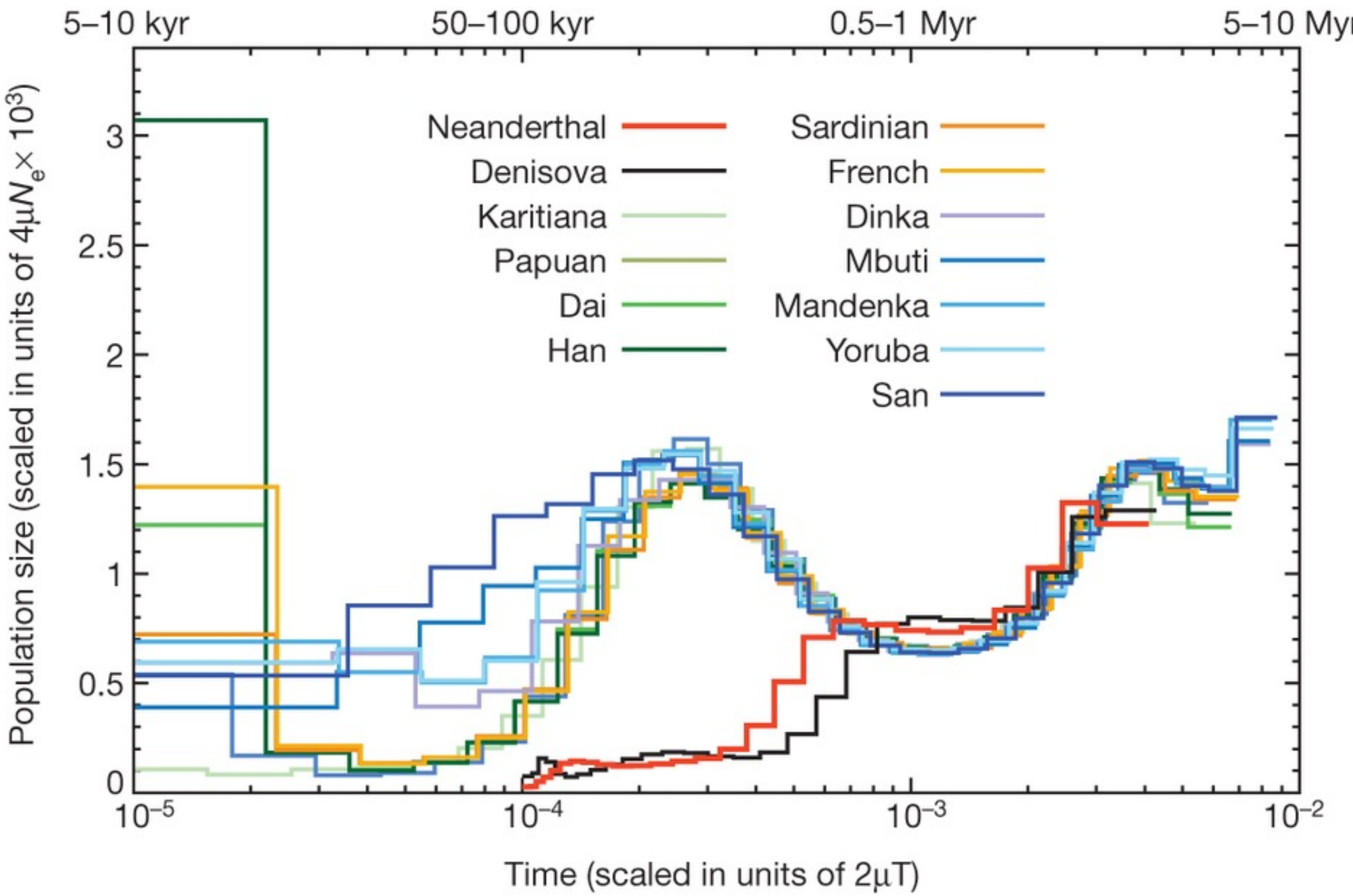

Figure: German Dziebel

# Unsupervised learning examples from biology: structured prediction




Genes mirror geography within Europe (2008)

# Unsupervised learning examples from biology: structured prediction




*The complete genome sequence of a Neanderthal from the Altai Mountains, Prufer et al (2014)*

# Clustering

* learn about the structure in our data

* cluster new data (prediction)

**GOAL** →

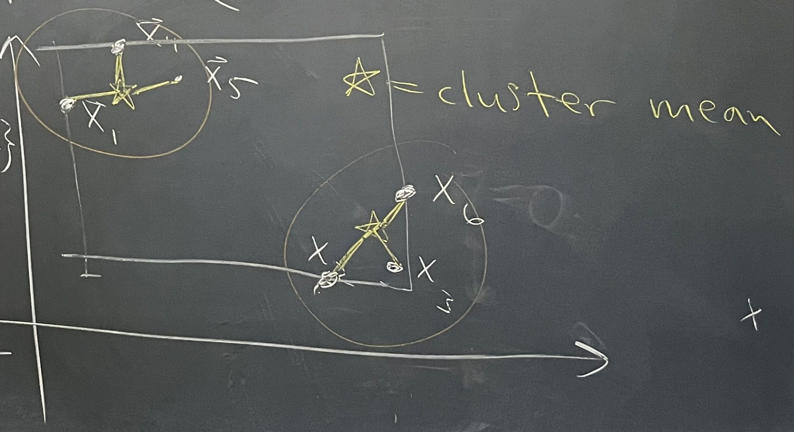
$$C_1 = \{\vec{x}_1, \vec{x}_4, \vec{x}_5\}$$

$$C_2 = \{\vec{x}_2, \vec{x}_3, \vec{x}_6\}$$

**Goal:** $\{C_1, C_2, \dots C_K\} = C$

**such that**

within cluster similarity is minimized



$$J(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \|\vec{x}_i - \vec{\mu}_k\|^2$$

minimizing this cost function

☆ cluster mean

# K-means algorithm

pick

0. initialization step

choose means (centers) randomly from the data

$$\vec{\mu}_1^{(1)}, \vec{\mu}_2^{(1)} \dots \vec{\mu}_K^{(1)}$$

iterate

1. E-step assign each datapoint to the closest mean
2. M-step recompute means as the cluster avg.

pick K

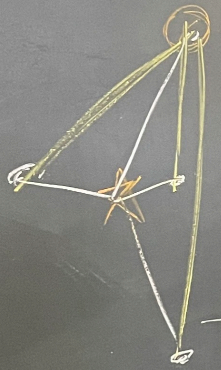
ers)

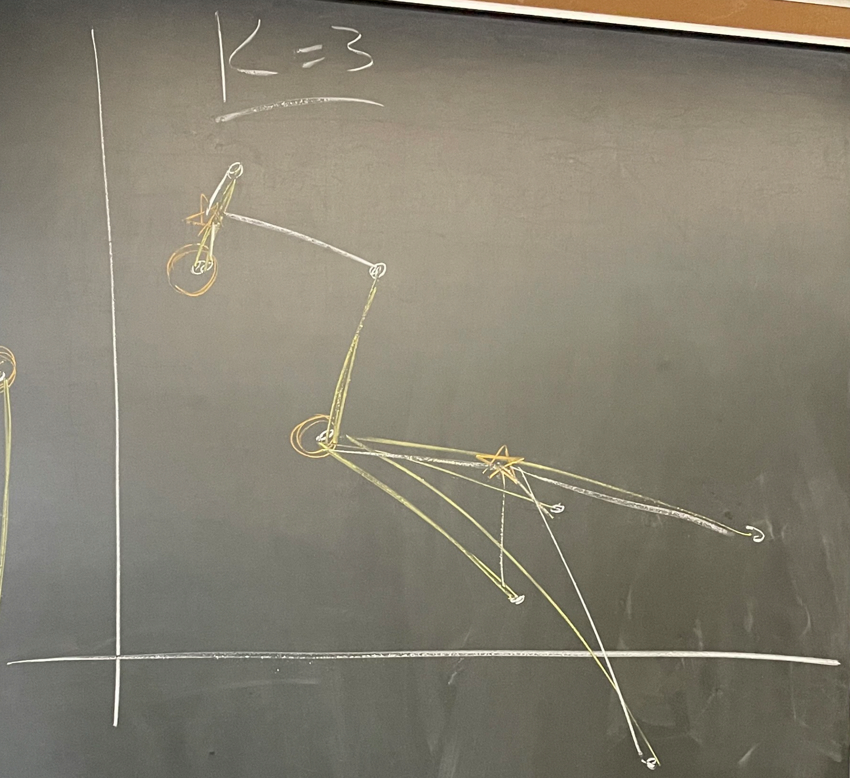
ata

datapoint

ns as

vg.






Stop: when cluster membership no longer changes

(or when you see a pattern you've seen before)