Welcome to CME 250 Introduction to Machine Learning!

Spring 2020 – Online version

April 14th 2020

Stanford | Institute for Computational & Mathematical Engineering

Office Hours

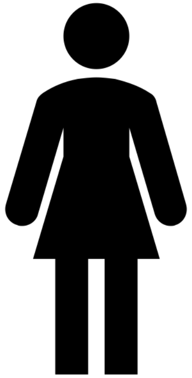
- Tuesdays: 10:30 am – 11:30 am
- Fridays: 12 pm – 1 pm (Starting this Friday)

Today’s schedule

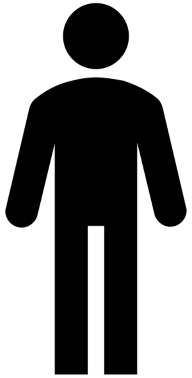
- Unsupervised Learning: Goals and Challenges
- Clustering
- Similarity / Dissimilarity Matrix
- K-means
- Hierarchical Clustering
- Gaussian Mixture Model

Let's get to know each other…

Breakout room



You



Another student

Name

Location

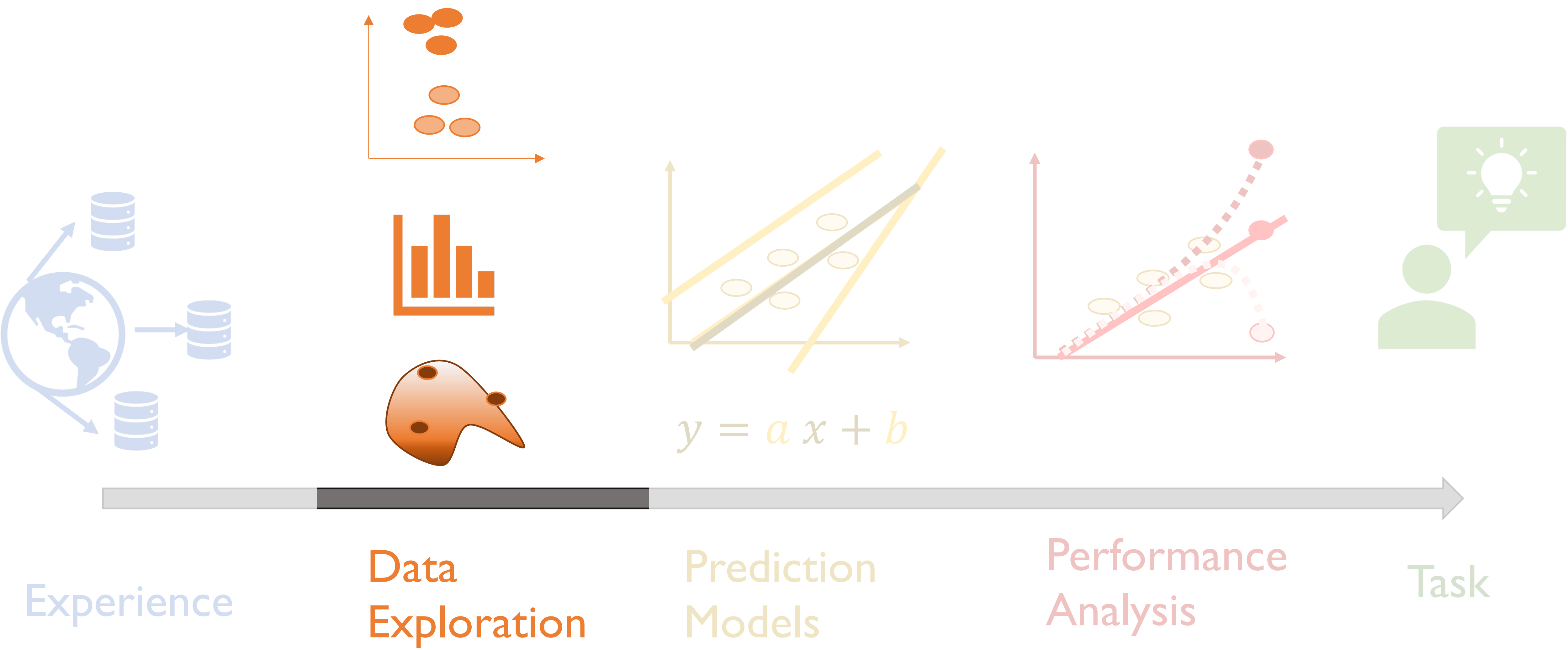
Department

Year

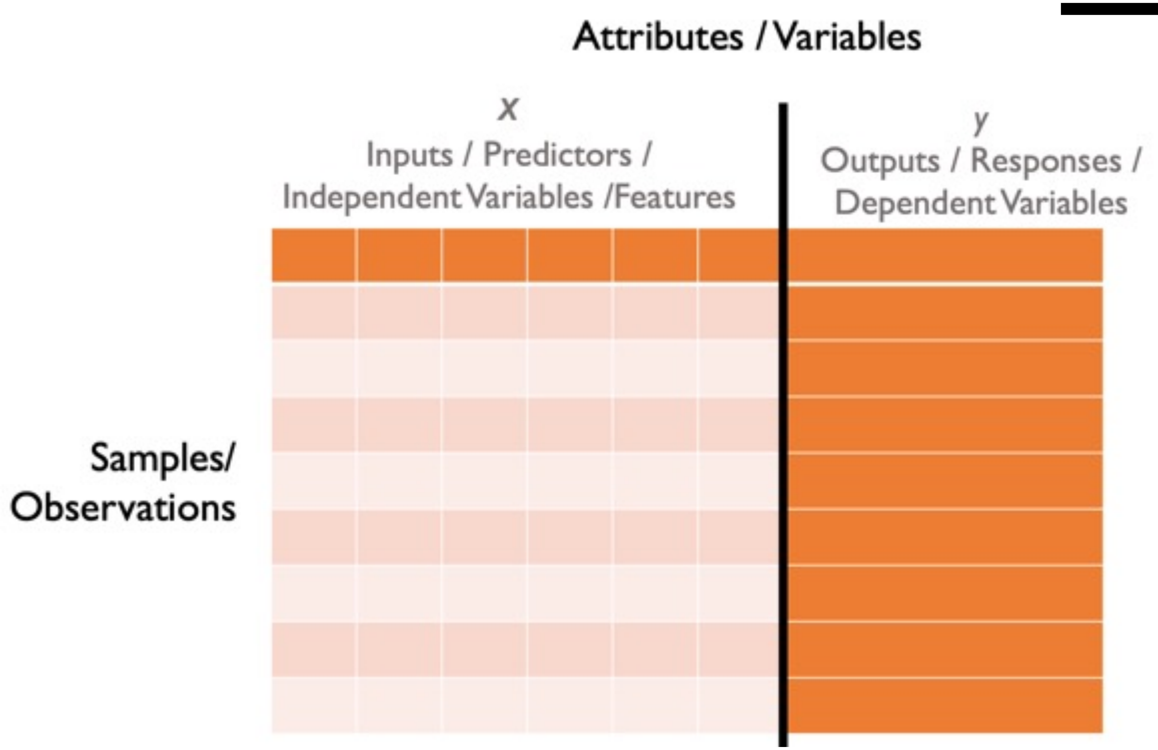
Interest in applying ML to …

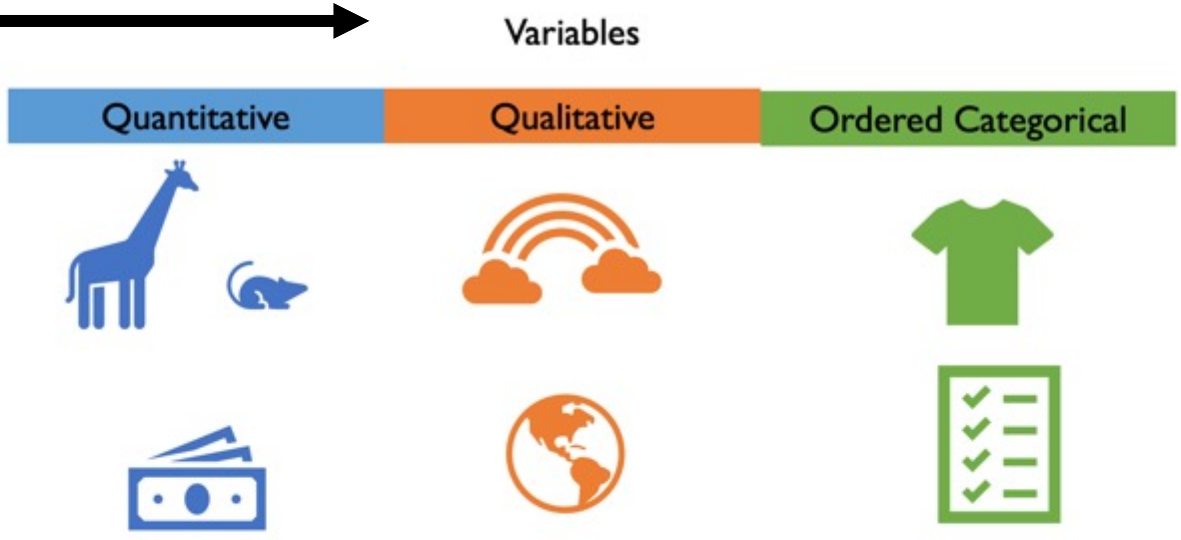
3 mins

Chat/Audio/Video


$y = a\,x + b$
Experience
Data Exploration
Prediction Models
Performance Analysis
Task

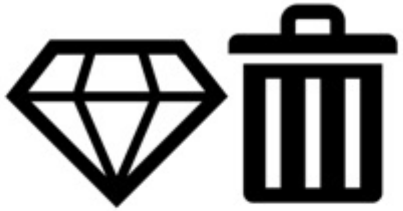
Last Class: Variable types





Order	✓	✗	✓
Distance	✓	✗	✗

Last Class: Exploratory Data Analysis



Data Quality

Quick Insights

Summaries

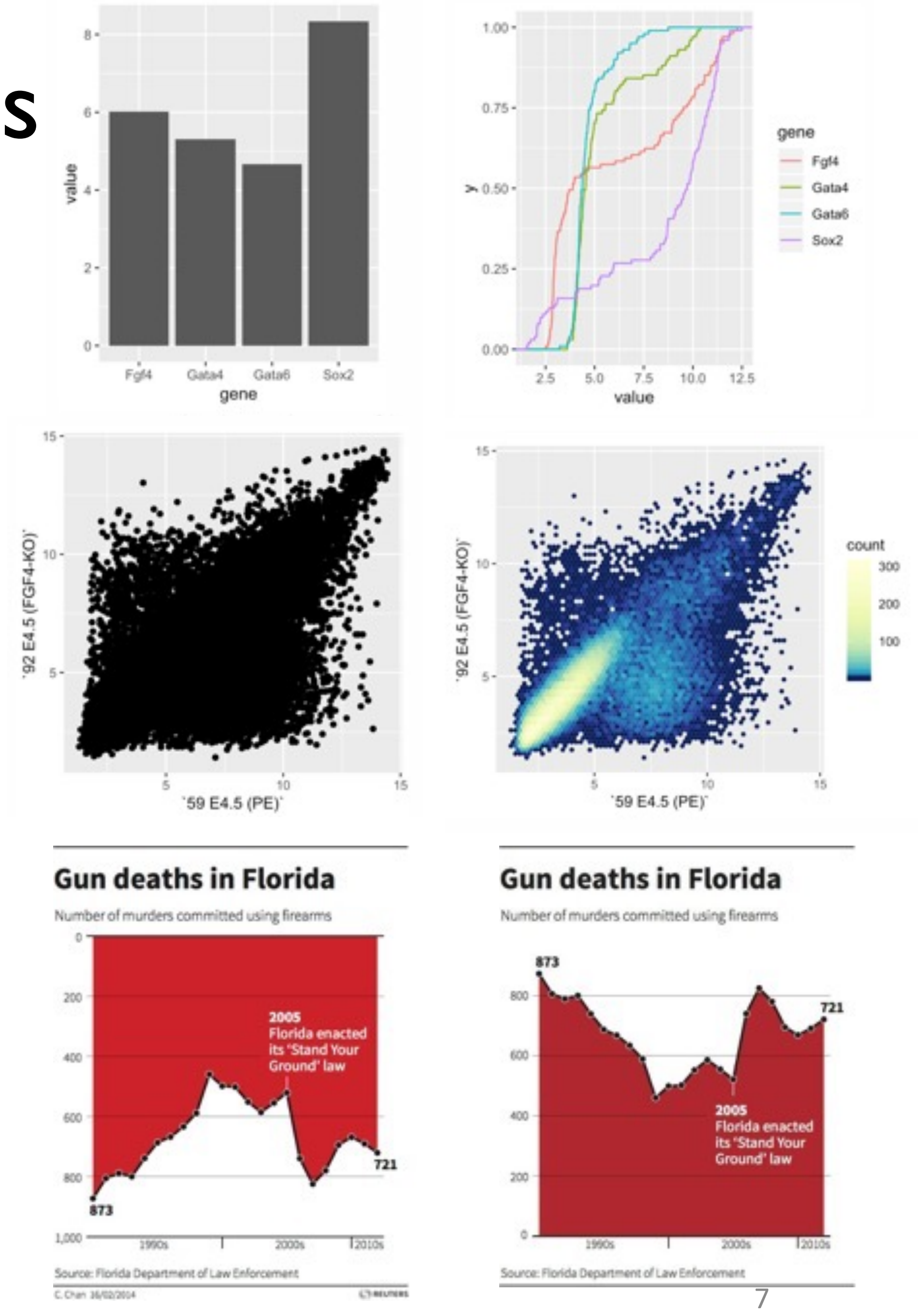
Pandas - Python

tidyverse -R

Visualization

Seaborn, Plotly … - Python

ggplot -R



Unsupervised Learning Part I: Clustering

Data Exploration

Introduction to Statistical Learning

Chapter 10.1: Intro to Unsupervised Learning,

10.3: Clustering

10.5: Practical Lab in R

Elements Statistical Learning

Chapter 13.2: K-means vs. Gaussian Mixture Models

14.3: Similarity Matrix and Clustering

8.5: Gaussian Mixture Model and EM

What is Unsupervised Learning?

Patterns + Properties
in Data

Clustering

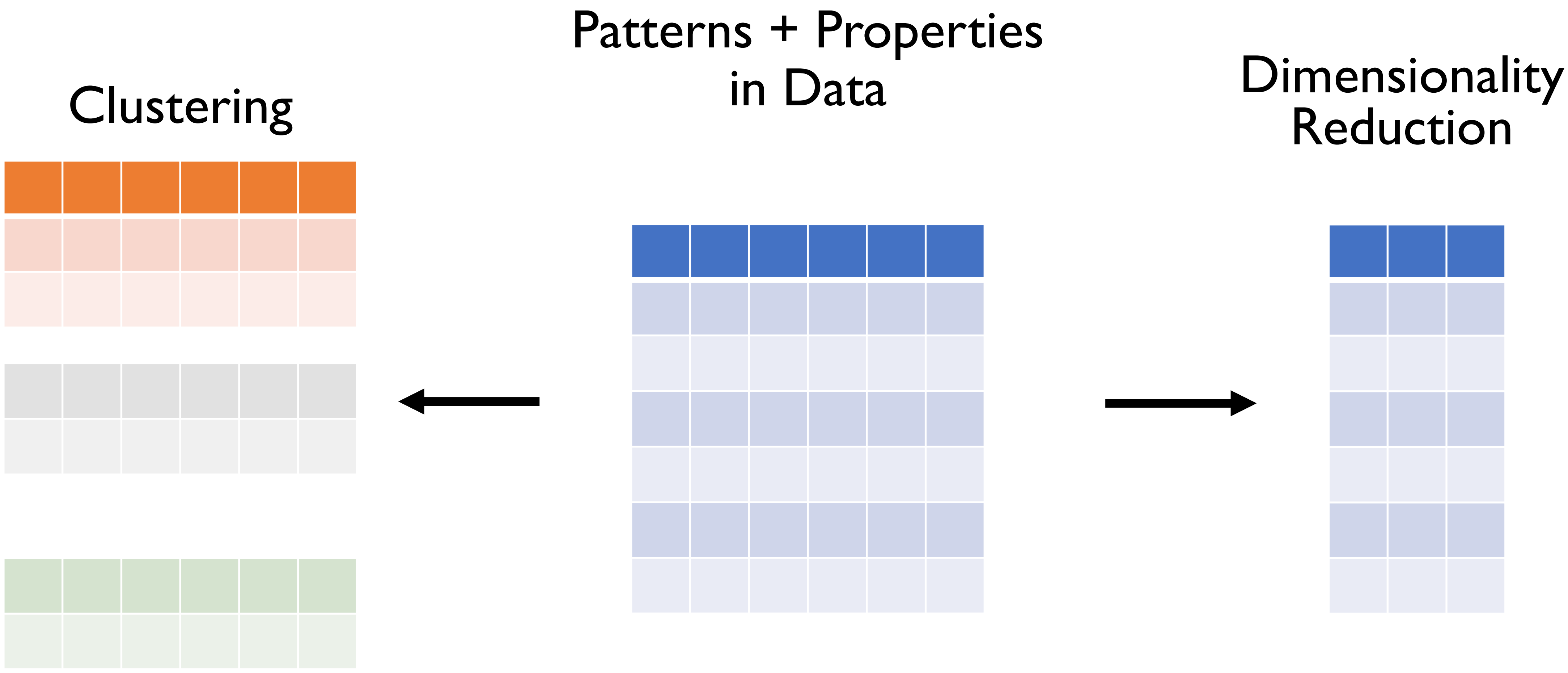
Subgroups that
explain variations
in data

Dimensionality
Reduction

Approx. same
information with
less variables

Challenge: What does it mean to be close?

What is Unsupervised Learning?

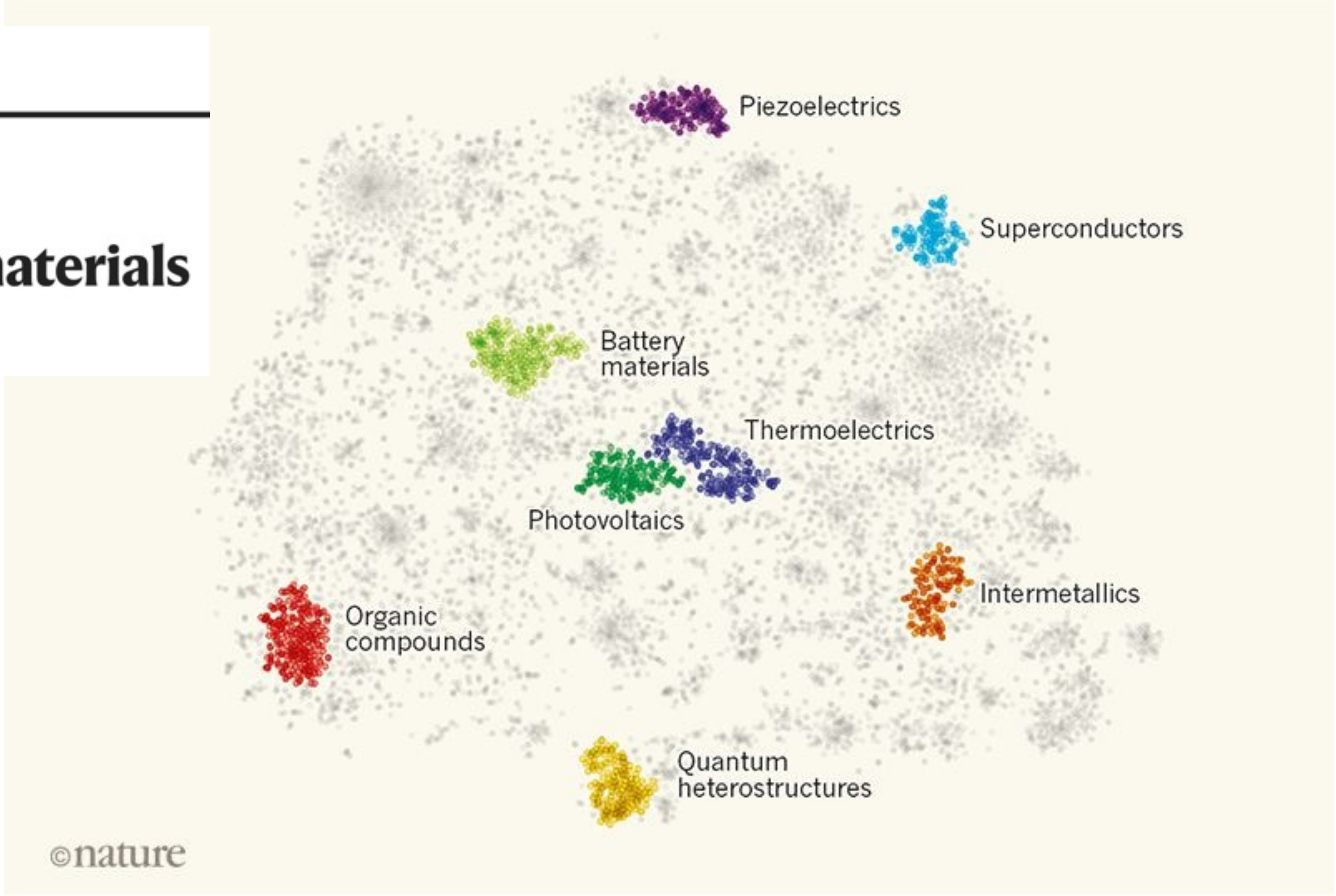


Clustering

nature

NEWS AND VIEWS · 03 JULY 2019

Text mining facilitates materials discovery



https://www.nature.com/articles/d41586-019-01978-x

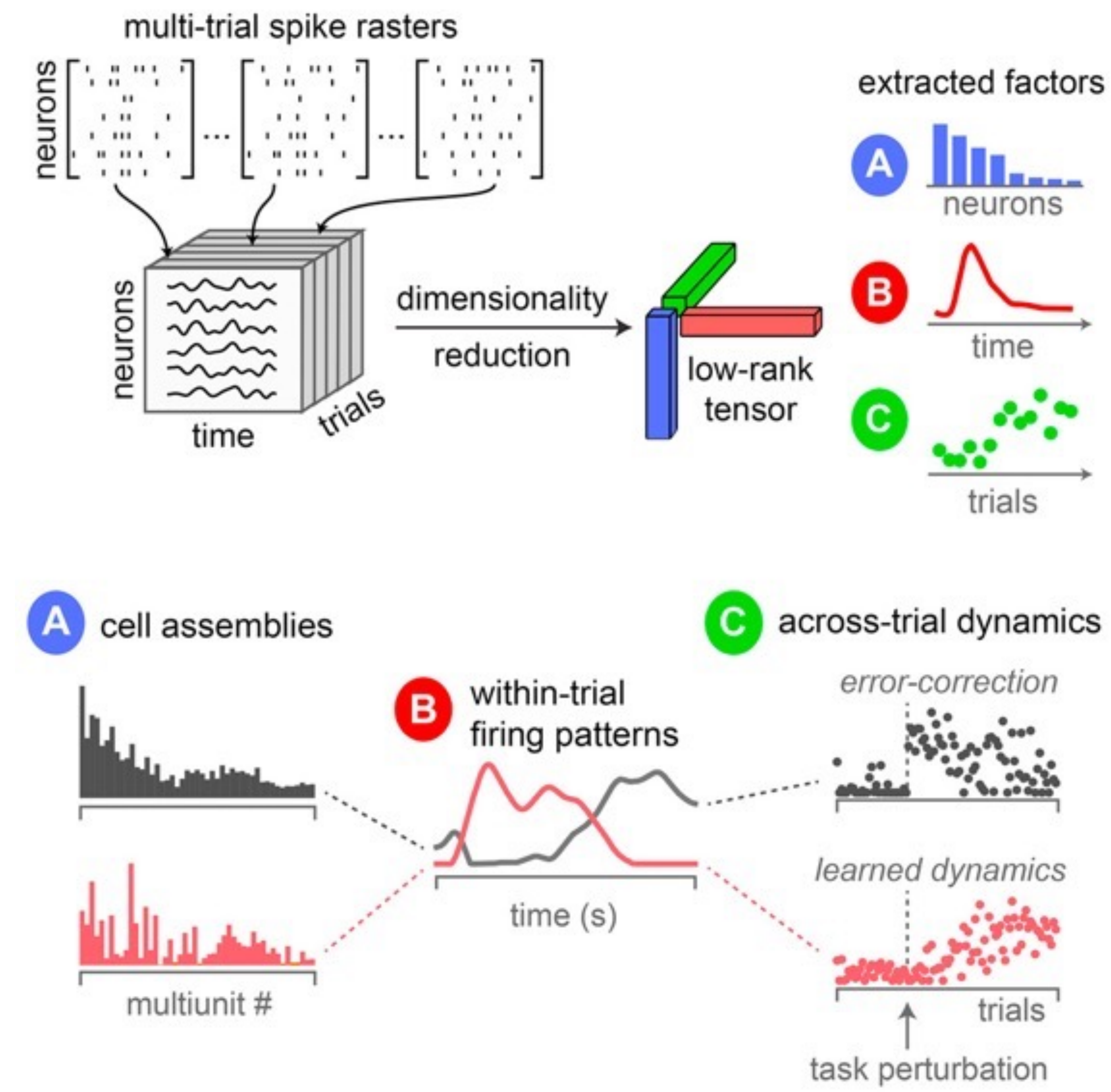
Dimensionality Reduction

Neuron
Volume 98, Issue 6, 27 June 2018, Pages 1099-1115.e8

NeuroResource

Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis

Alex H. Williams [1,13], Tony Hyun Kim [2], Forea Wang [1], Saurabh Vyas [2,3], Stephen I. Ryu [2,11], Krishna V. Shenoy [2,3,6,7,8,9], Mark Schnitzer [4,5,7,9,10], Tamara G. Kolda [12], Surya Ganguli [4,6,7,8]

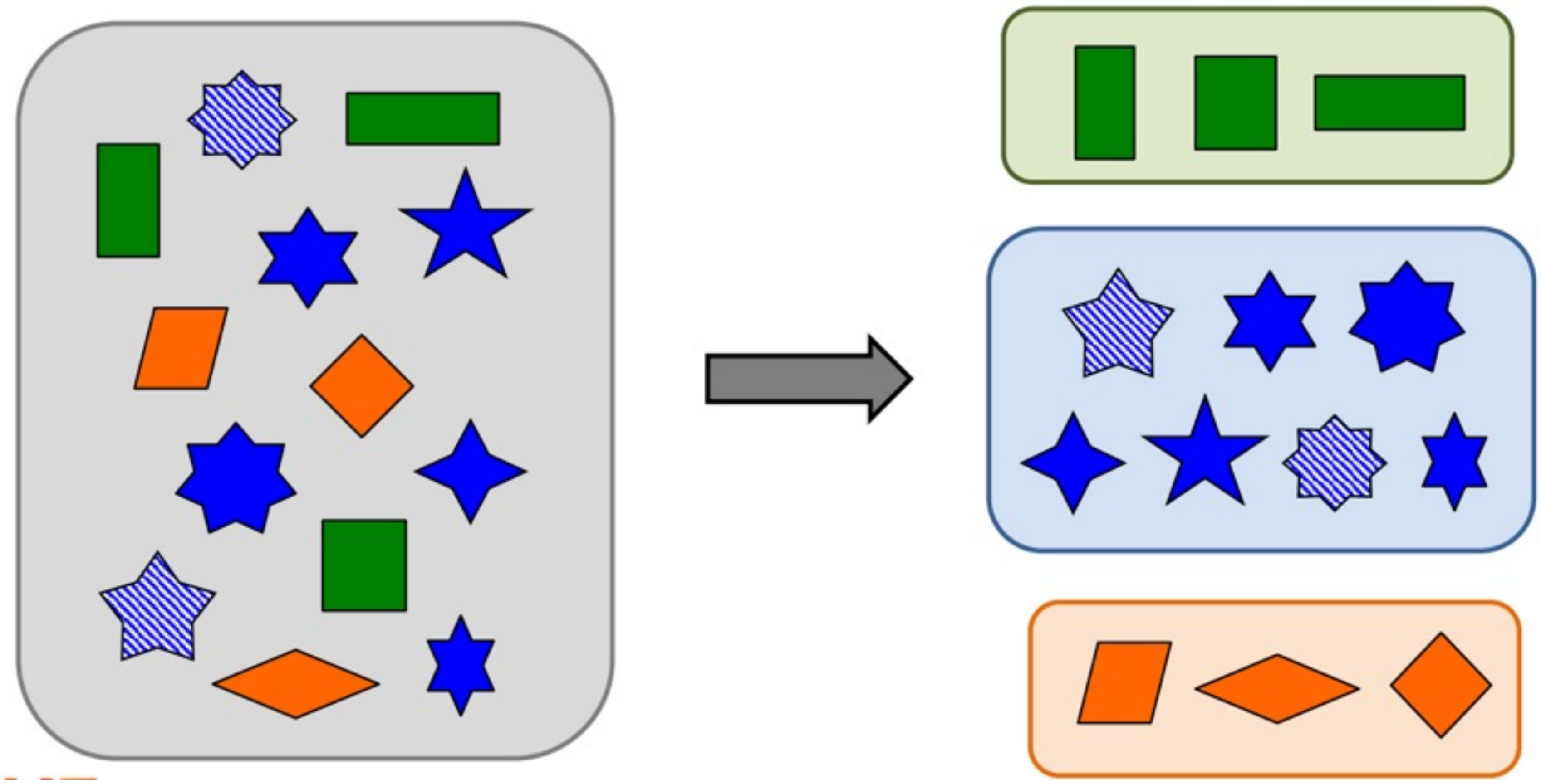


https://doi.org/10.1016/j.neuron.2018.05.015

What is clustering?

observations inside each group are alike,

observations between groups are different



What does it mean to be close? = Dissimilarity

Measure how close two samples are

$$d(x^{(1)}, x^{(2)})$$

Quantitative	Qualitative	Ordered Categorical

Quantitative

Squared error

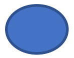
$$d(x^{(1)}, x^{(2)}) = \left(x^{(1)} - x^{(2)}\right)^2$$

Absolute error

$$d(x^{(1)}, x^{(2)}) = |x^{(1)} - x^{(2)}|$$

Qualitative

Dummy variable:
Is a …? 0/1

				
	1	0	0	0
	0	0	1	0

Ordered Categorical

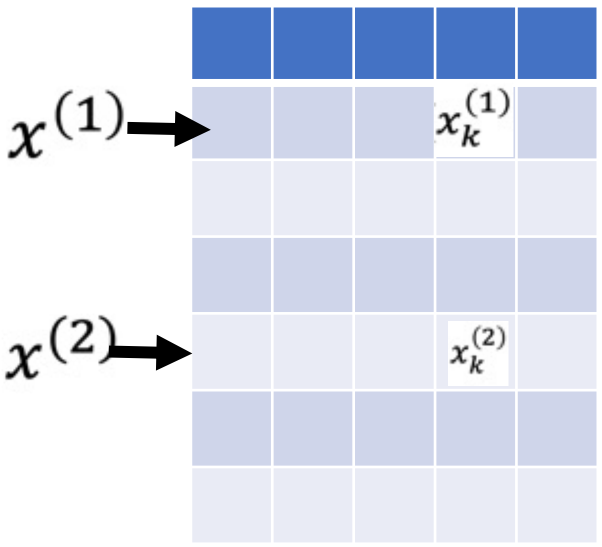
$$\tilde{x} = \frac{index \;- 0.5}{\#\, classes}$$

	Index	0-1 scale
	1	0.125
	2	0.375
	3	0.625
	4	0.875

How to combine dissimilarities

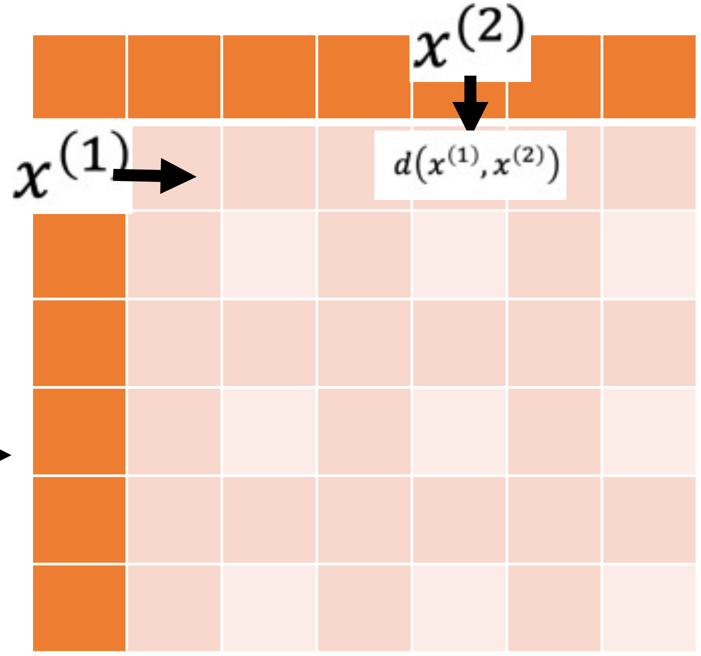
Measure how close two samples are

$$d(x^{(1)}, x^{(2)})$$



$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{\#\,vars} \boxed{w_k} * d_k(x_k^{(1)}, x_k^{(2)})$$

Variable Influence



(Similarity)
Dissimilarity Matrix

Dissimilarity may be more important than choosing clustering algorithm

What if all of the variables are related?

Buyers

Total amount of sales per product

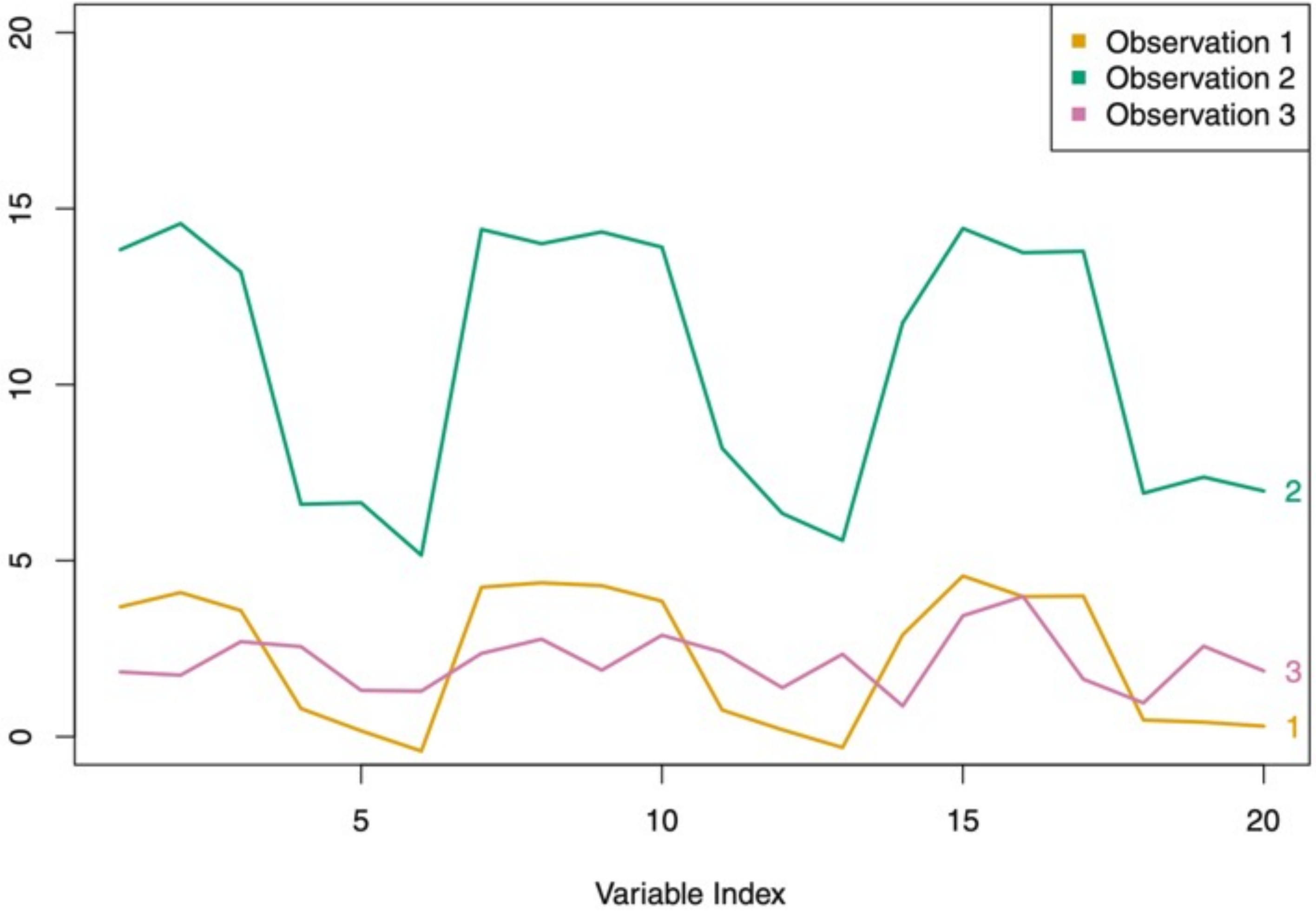
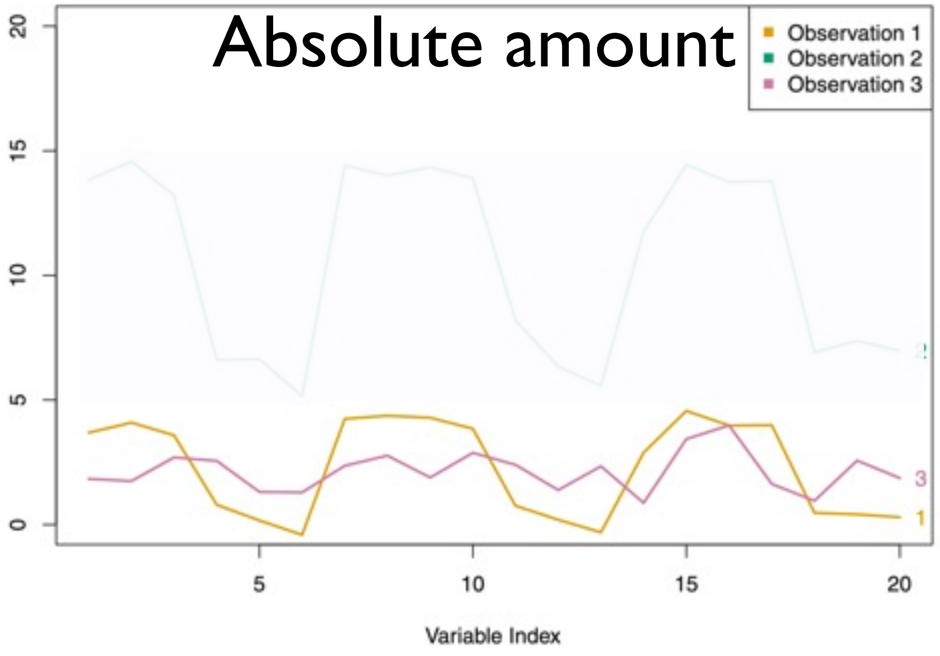
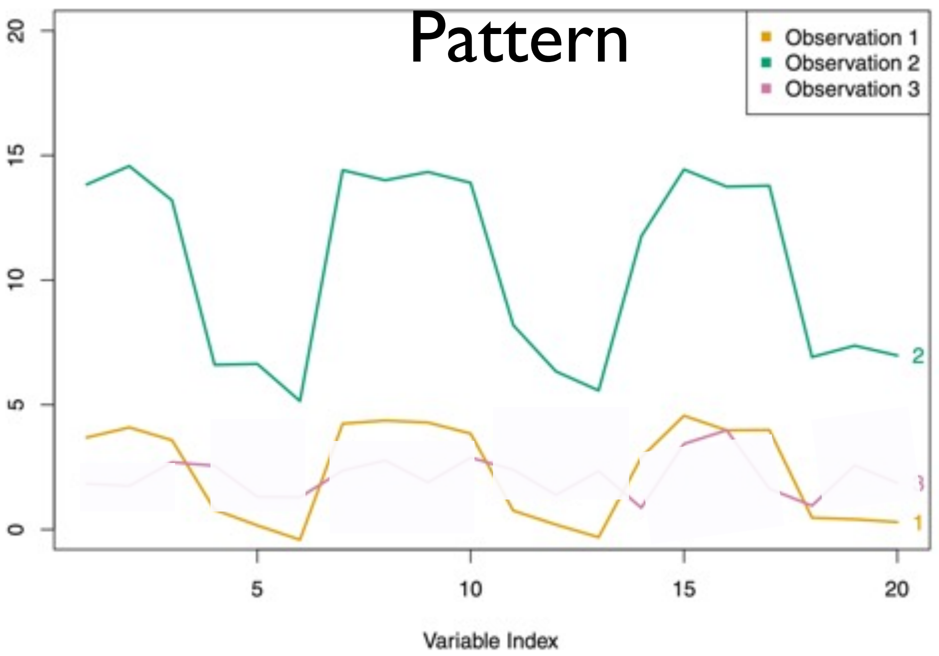


Figure 10.13 ISL (8th printing 2017)

Products in a supermarket

What if all of the variables are related?





Euclidean distance / Squared error

$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \left(x_k^{(1)} - x_k^{(2)}\right)^2$$

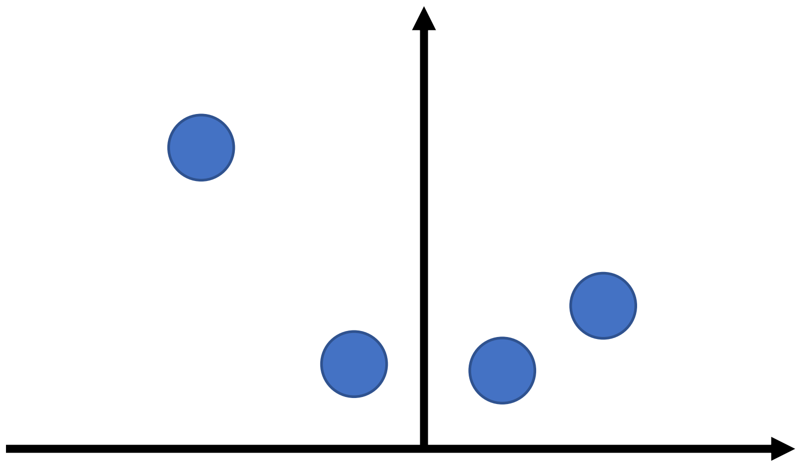
Correlation

$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \frac{\left(x_k^{(1)} - \bar{x}^{(1)}\right)\left(x_k^{(2)} - \bar{x}^{(2)}\right)}{\sqrt{\sum_{k=1}^{P}\left(x_k^{(1)} - \bar{x}^{(1)}\right)^2}\sqrt{\sum_{k=1}^{P}\left(x_k^{(2)} - \bar{x}^{(2)}\right)^2}}$$

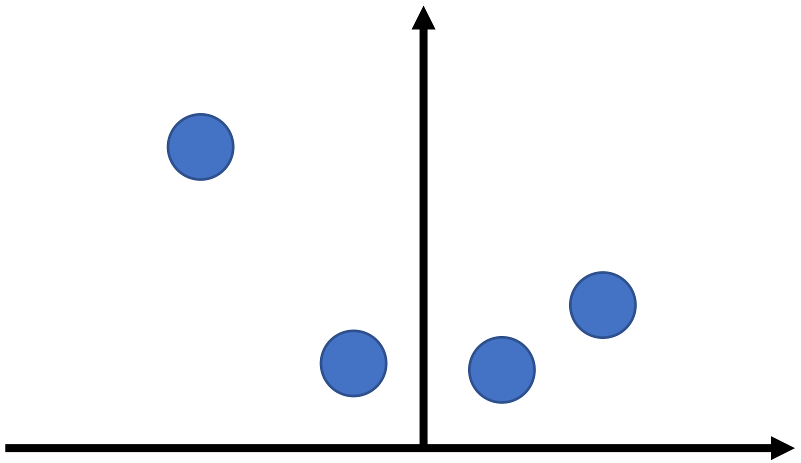
*If means are zero = cosine similarity

What if all of the variables are related?

Absolute amount



Pattern



Euclidean distance / Squared error

$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \left(x_k^{(1)} - x_k^{(2)}\right)^2$$

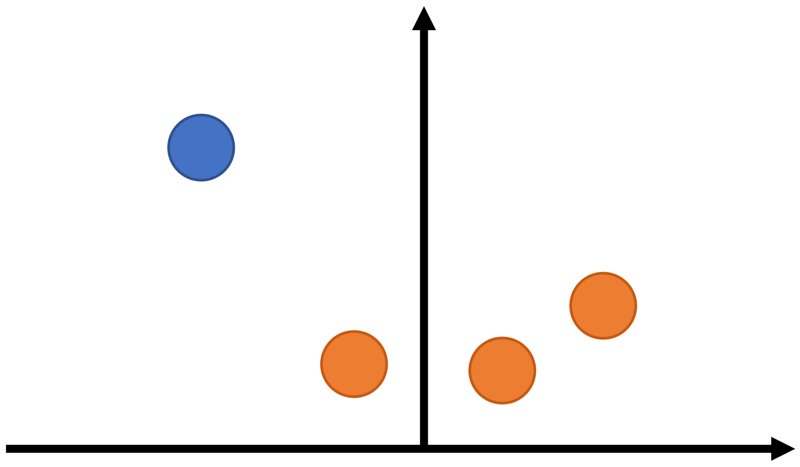
Correlation

$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \frac{\left(x_k^{(1)} - \bar{x}^{(1)}\right)\left(x_k^{(2)} - \bar{x}^{(2)}\right)}{\sqrt{\sum_{k=1}^{P}\left(x_k^{(1)} - \bar{x}^{(1)}\right)^2}\sqrt{\sum_{k=1}^{P}\left(x_k^{(2)} - \bar{x}^{(2)}\right)^2}}$$

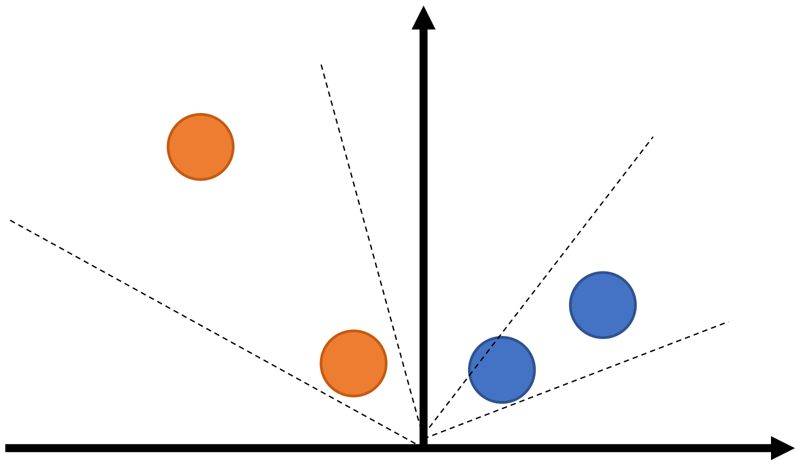
*If means are zero = cosine similarity

What if all of the variables are related?

Absolute amount



Pattern



Euclidean distance / Squared error

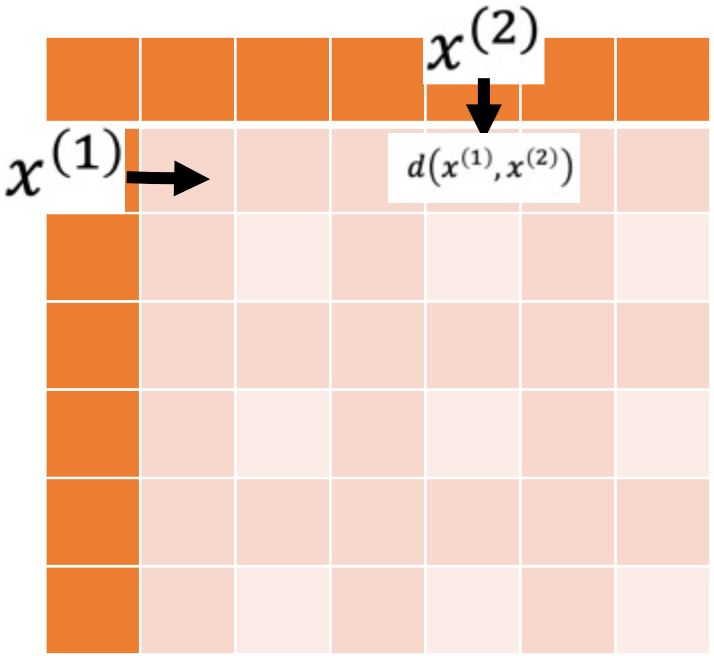
$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \left(x_k^{(1)} - x_k^{(2)}\right)^2$$

Correlation

$$d\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{P} \frac{\left(x_k^{(1)} - \bar{x}^{(1)}\right)\left(x_k^{(2)} - \bar{x}^{(2)}\right)}{\sqrt{\sum_{k=1}^{P}\left(x_k^{(1)} - \bar{x}^{(1)}\right)^2}\sqrt{\sum_{k=1}^{P}\left(x_k^{(2)} - \bar{x}^{(2)}\right)^2}}$$

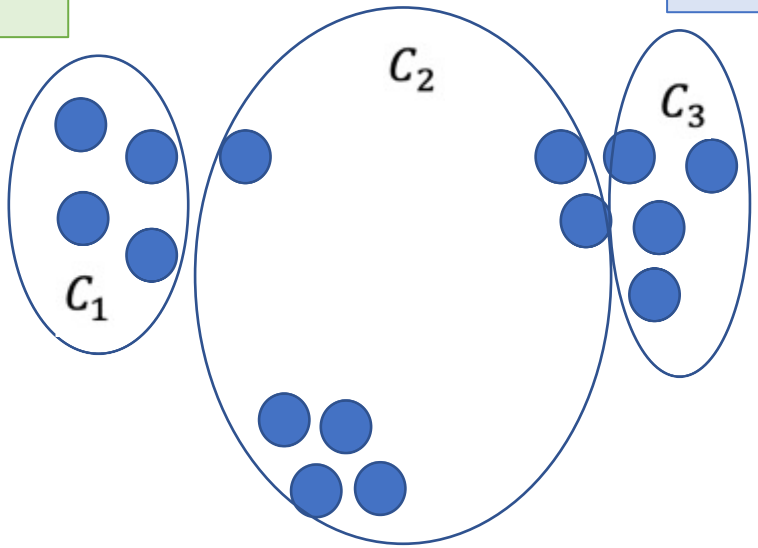
*If means are zero = cosine similarity

Clustering in terms of dissimilarity



Dissimilarity Matrix

$$\min_{C_1 \dots C_L} \underbrace{\sum_{l=1}^{L}}_{\text{Sum over clusters}} \underbrace{\sum_{i,j\in C_l}}_{\text{Sum over samples in cluster}} d(x^{(i)}, x^{(j)})$$



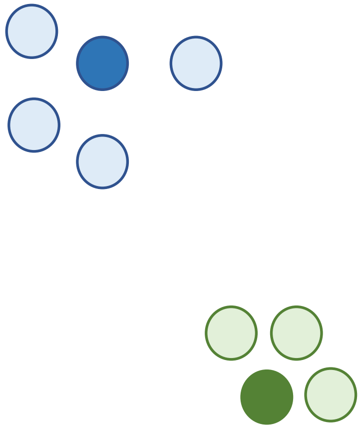
Combinatorial problem: complete enumeration is non-feasible

Types of clustering algorithms

Subgroups that explain variation

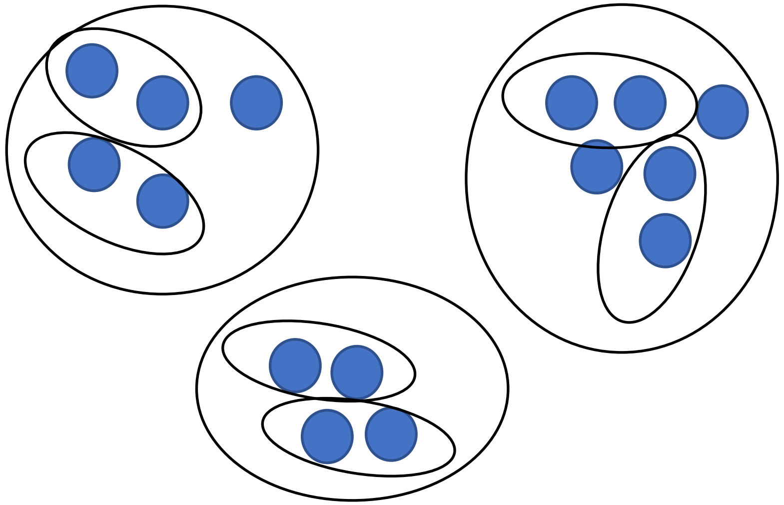
K-means

Prototypes

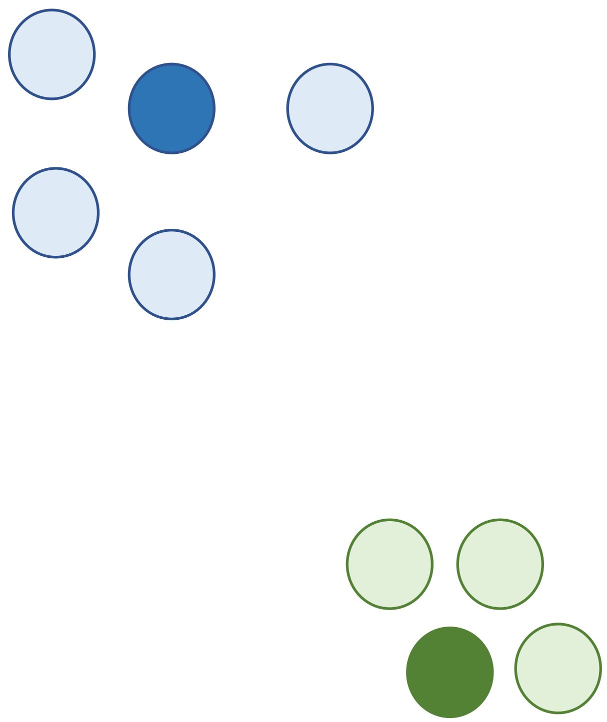


Hierarchical

Agglomerative



K-means



Find prototypes and clusters
that minimize

$$J = \sum_{l=1}^{L} \sum_{i \in C_l} d(x^{(i)}, \tilde{x}_l)$$

K-means algorithm (Lloyd's Algorithm)

(0) Initialize clusters (At random, far apart points, domain knowledge, another clustering)

(1) Iterate until clusters do not change

(a) Find best cluster for each point $x^{(i)}$

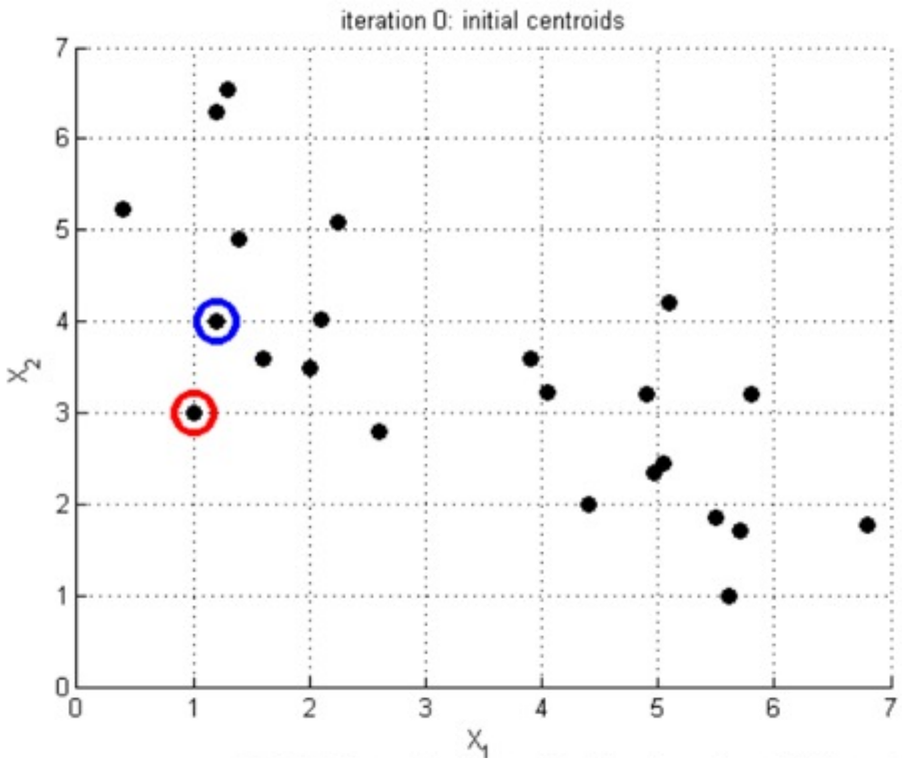
$$\min_{1,\ldots,l} d(x^{(i)}, \tilde{x}_l)$$

(b) Compute prototype $\tilde{x}_l$ for each cluster C_l

Centroid: d = squared error	Center: any dissimilarity
$\tilde{x}_l = \frac{1}{\lvert C_l \rvert} \sum_{i \in C_l} x^{(i)}$	$\tilde{x}_l = x^{(\tilde{j})}: \tilde{j} = \underset{j \in C_l}{\operatorname{argmin}} \sum_{i \in C_l} d(x^{(i)}, x^{(j)})$

* Also known as k-medoids

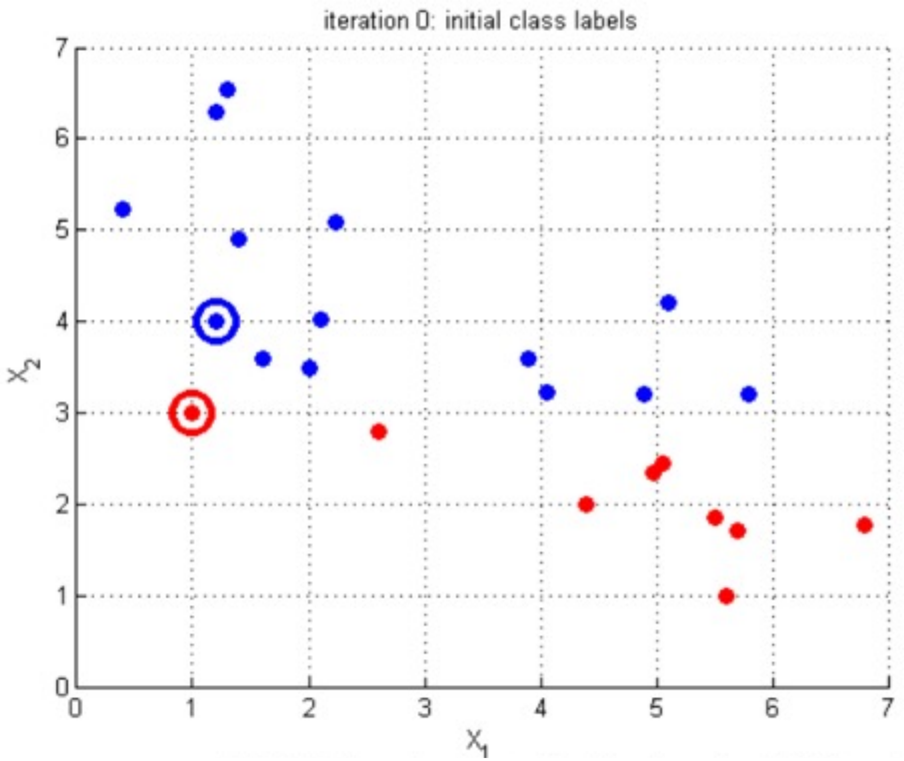
K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)

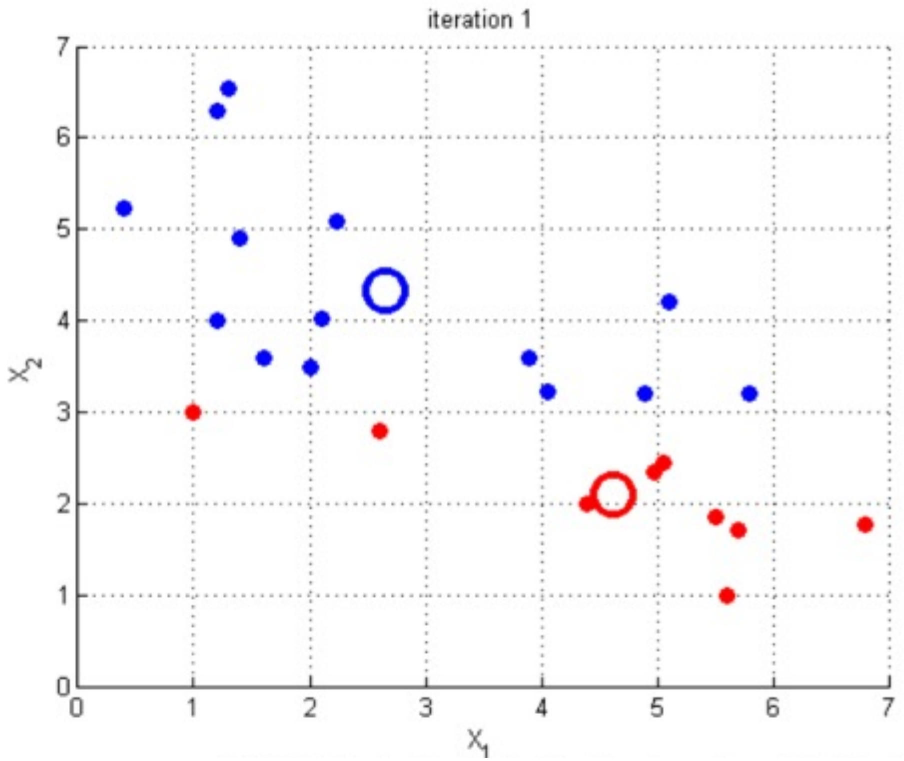


Pick initial centroids

Assign initial clusters

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



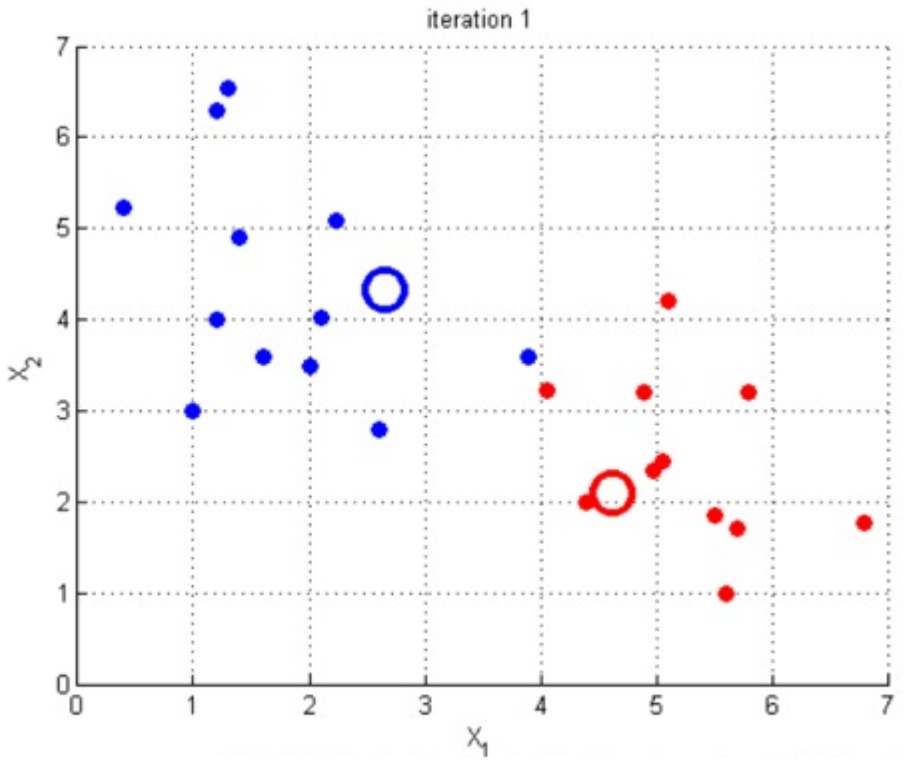
Pick initial centroids

Assign initial clusters

Update centroids

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

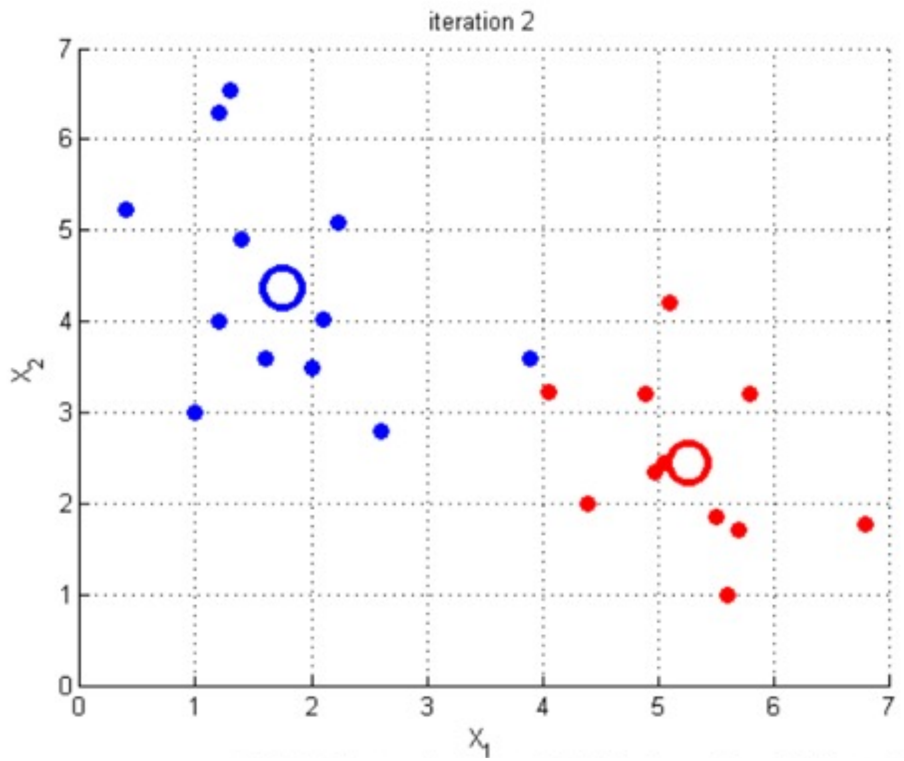
Assign initial clusters

Update centroids

Reassign clusters

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

Assign initial clusters

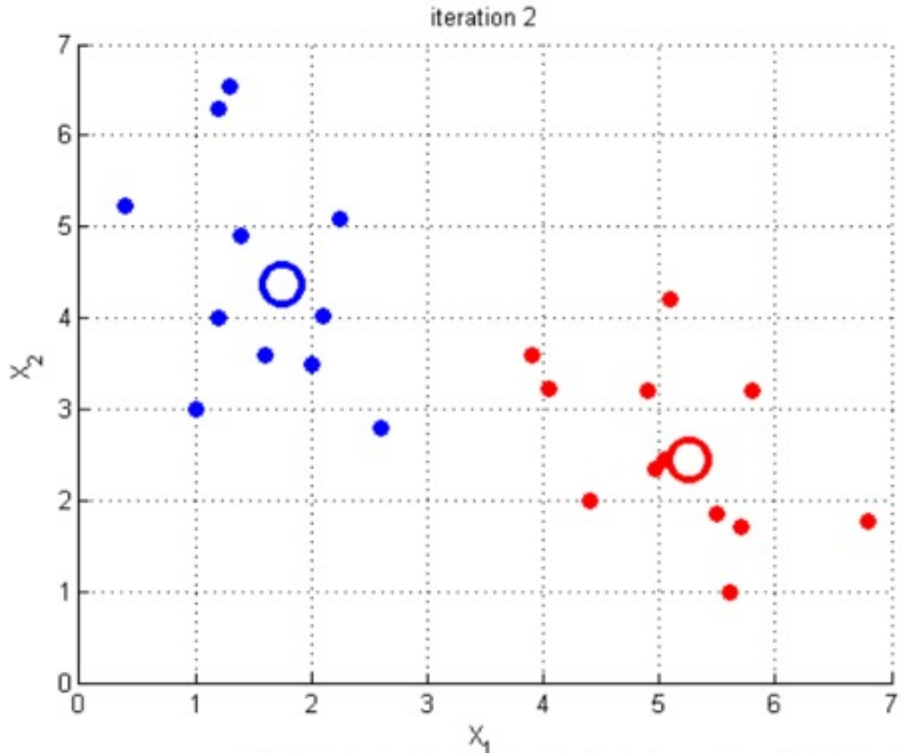
Update centroids

Reassign clusters

Update centroids

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

Assign initial clusters

Update centroids

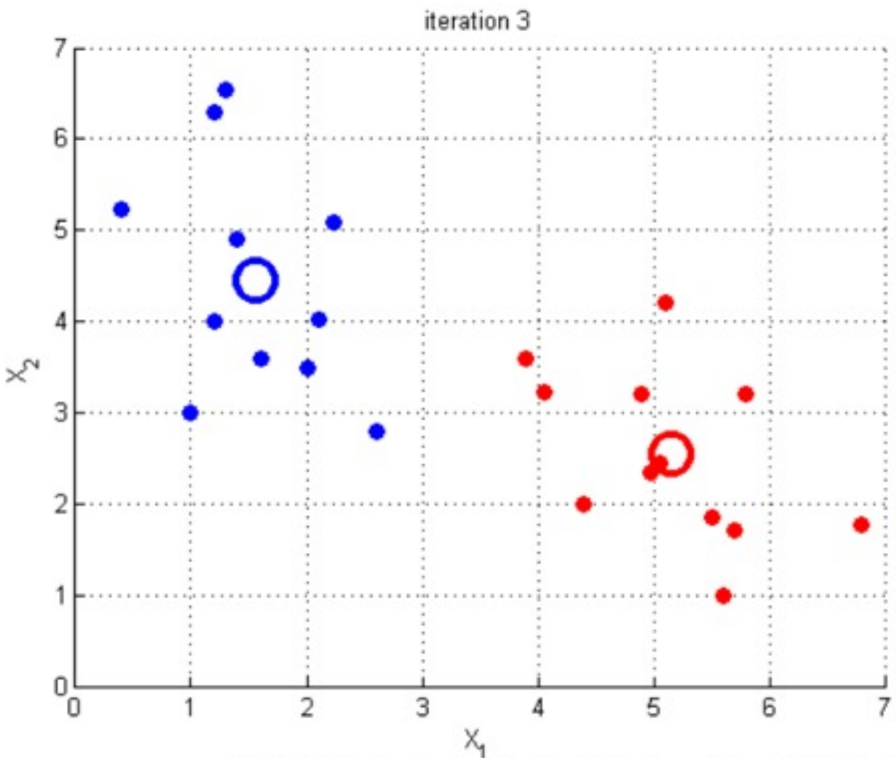
Reassign clusters

Update centroids

Reassign clusters

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

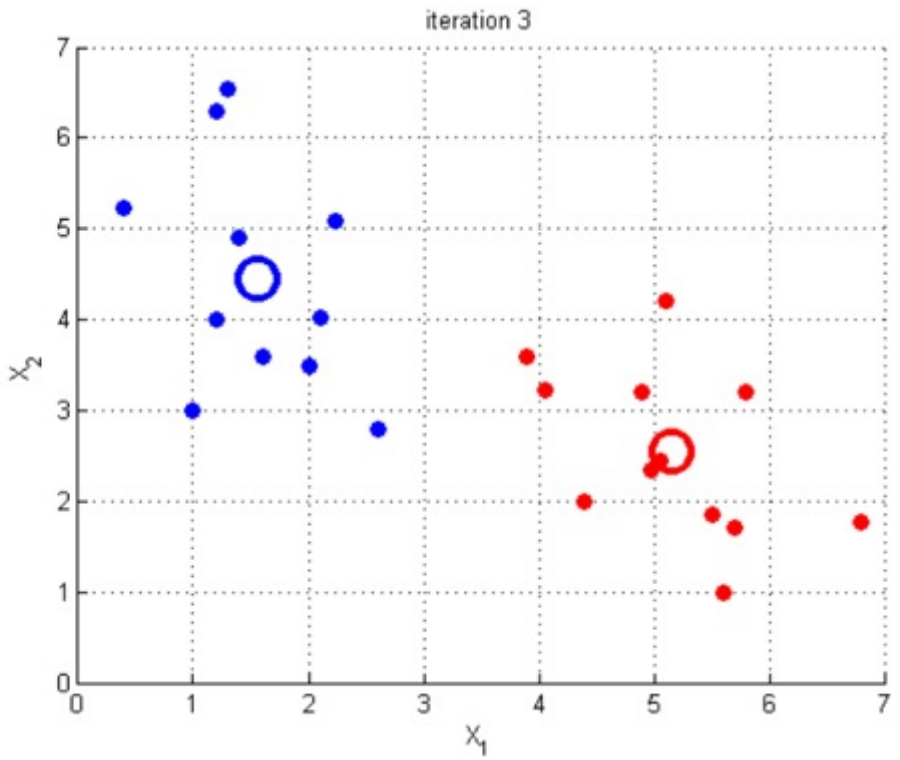
Update centroids

Reassign clusters

Update centroids

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

Update centroids

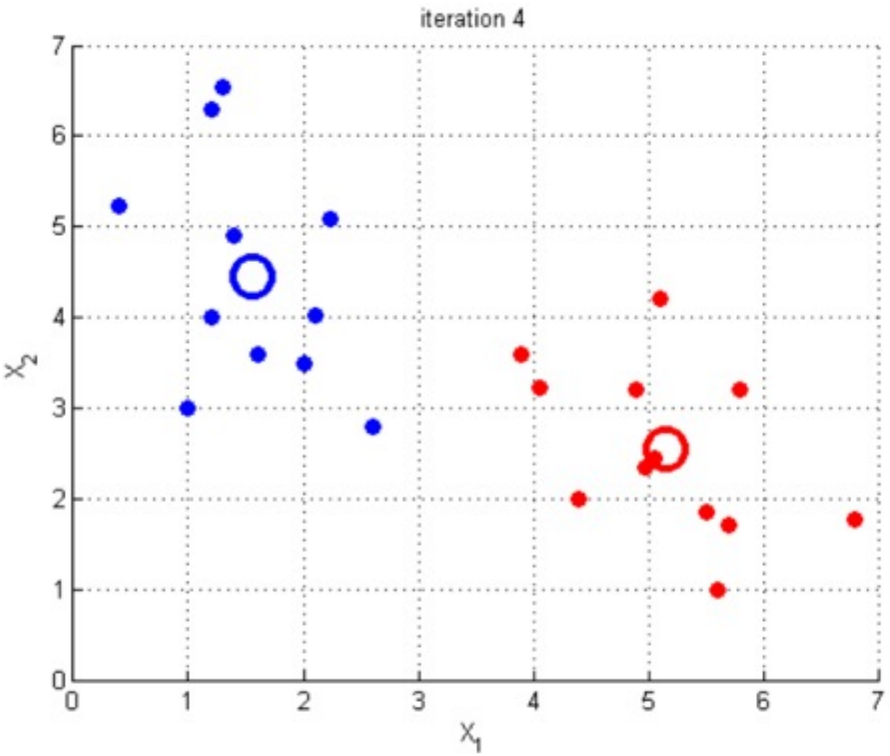
Reassign clusters

Update centroids

Reassign clusters

* Simulation done by Karianne Bergen

K-means algorithm (Lloyd's Algorithm)



Pick initial centroids

Assign initial clusters

Update centroids

Reassign clusters

Update centroids

Reassign clusters

Update centroids

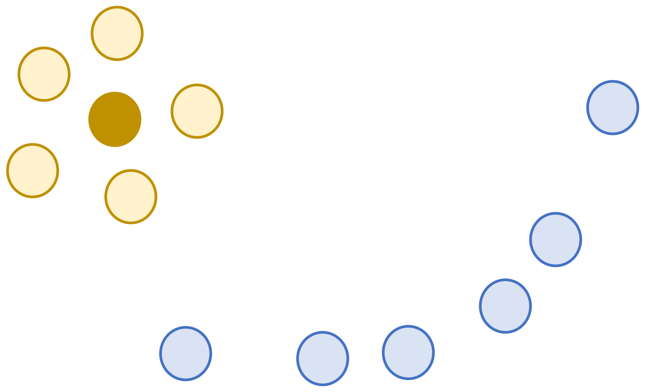
Reassign clusters

Converged

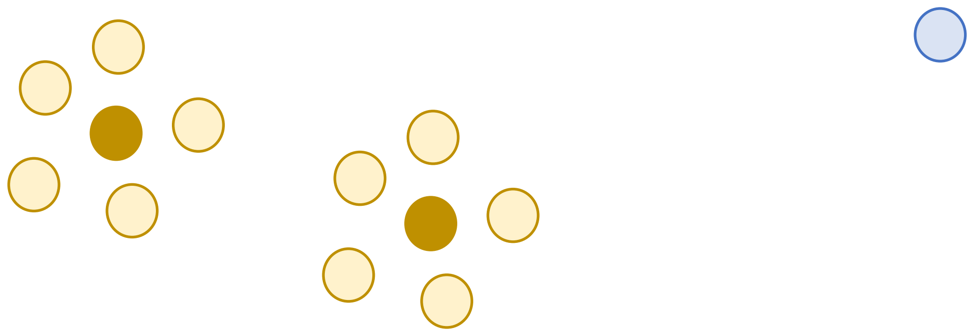
* Simulation done by Karianne Bergen

Challenges of K-means

All clusters are spherical
and of the same size.



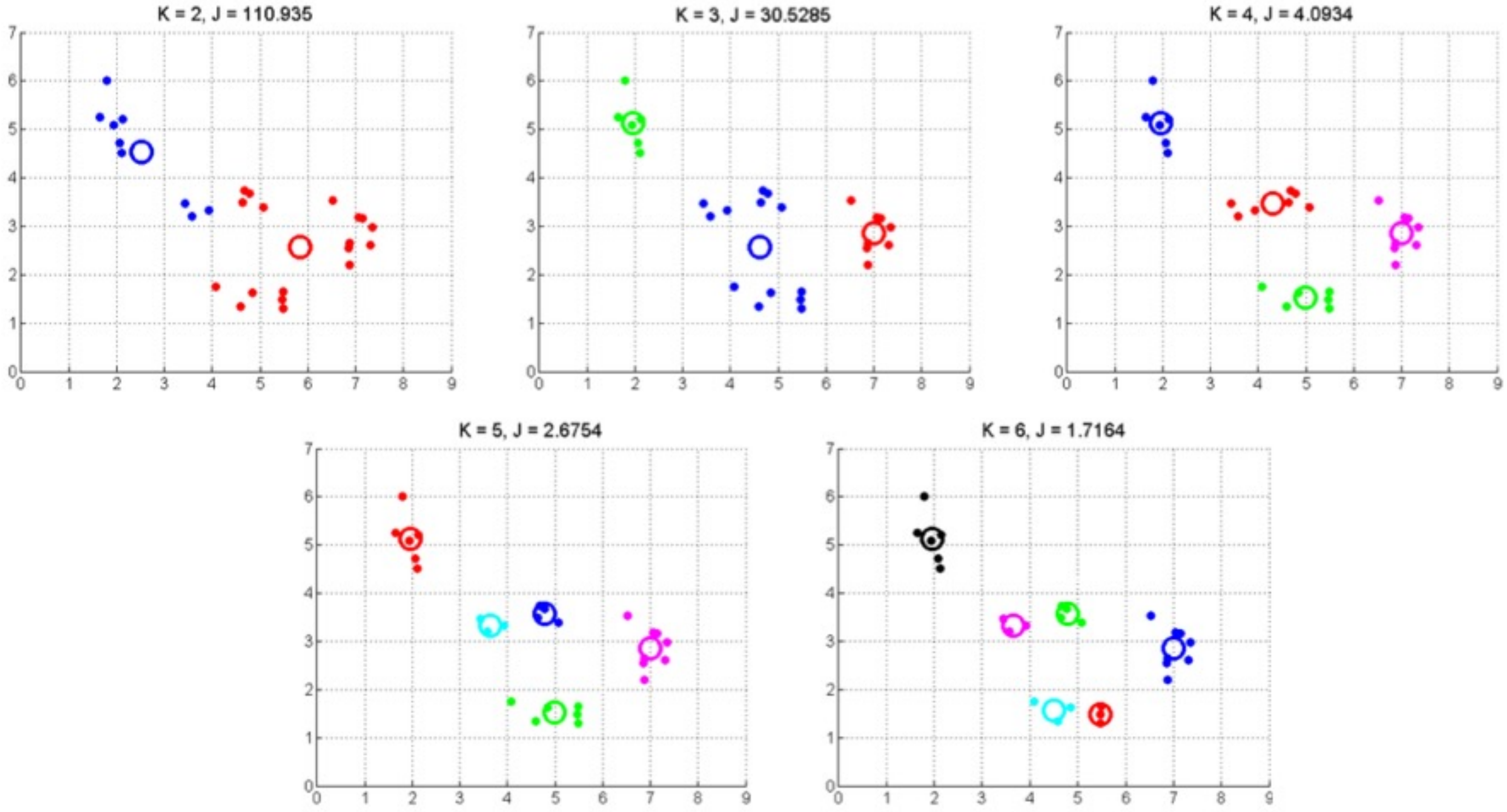
Sensibility to outliers
depending on
dissimilarity measure



Fixed number
of clusters

Choose number of clusters

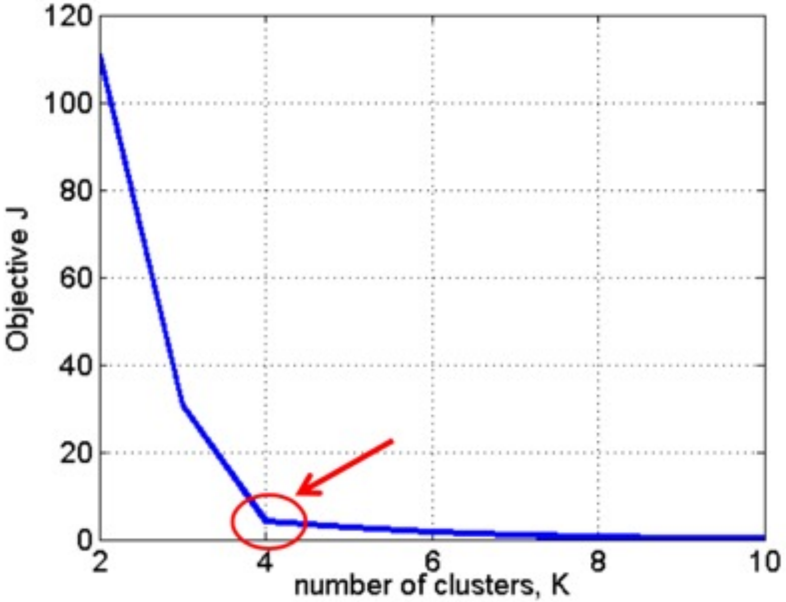
$$J = \sum_{l=1}^{L} \sum_{i \in C_l} d(x^{(i)}, \tilde{x}_l)$$

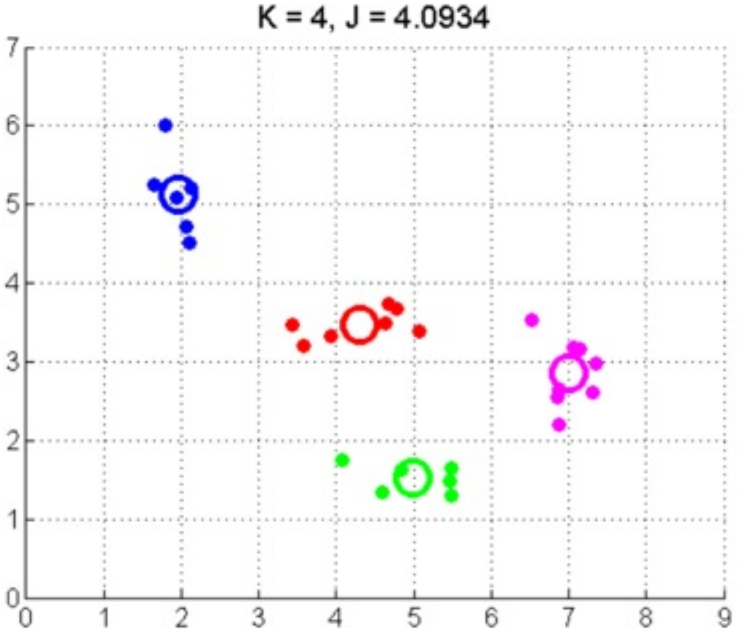


Choose number of clusters

$$J = \sum_{l=1}^{L} \sum_{i \in C_l} d(x^{(i)}, \tilde{x}_l)$$

Some heuristics:
- ✓ For each k repeat multiple times and select best J
- ✓ Find "elbow" in K vs J





K-means for compression

Pixels

R	G	B

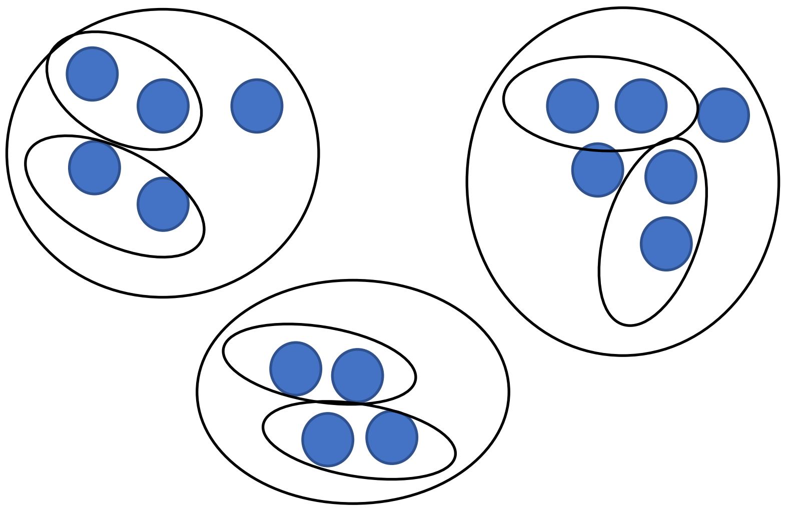
K-means
+
Replace by centroid

K=3

K=5

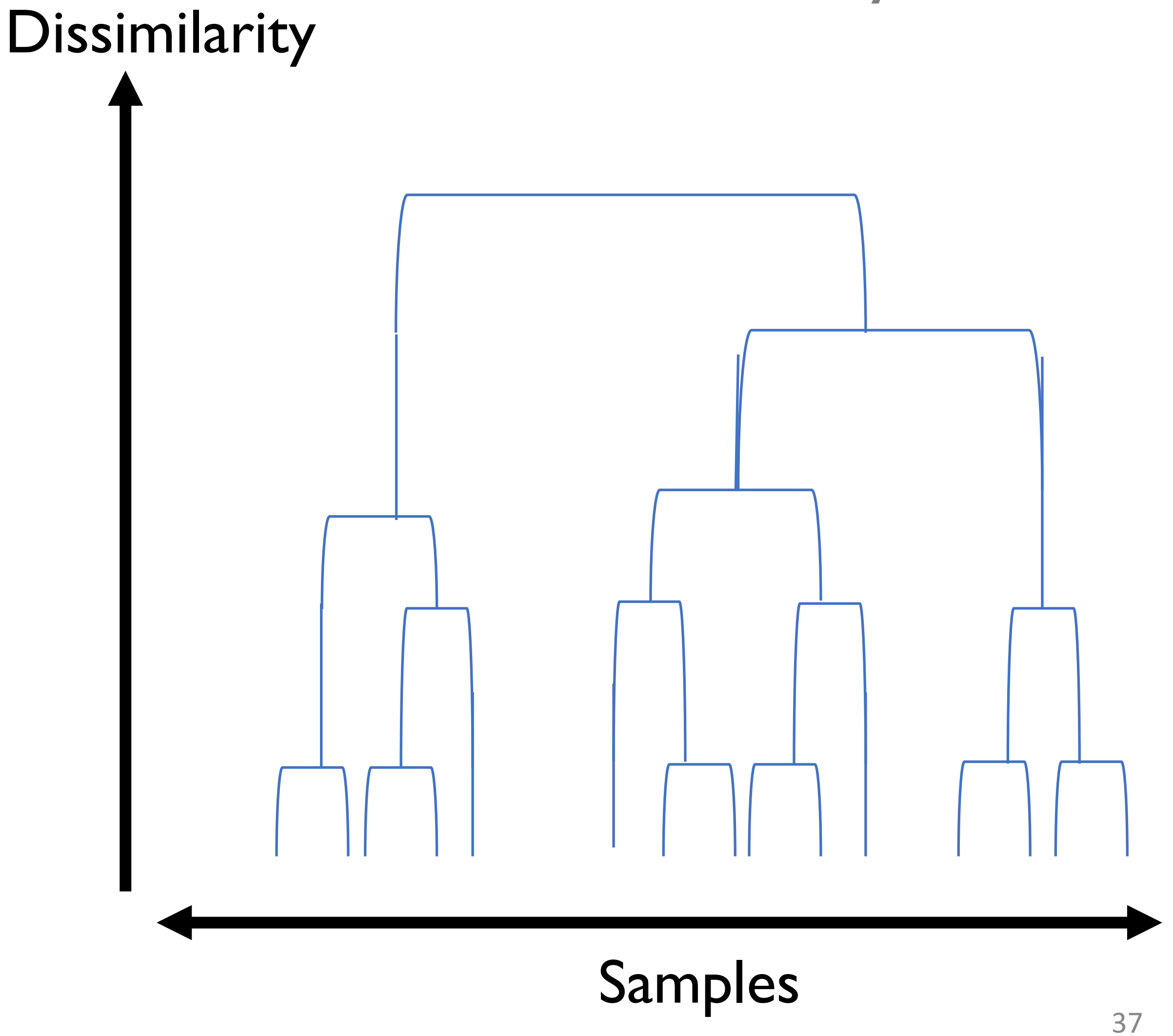
K=20

Hierarchical Clustering

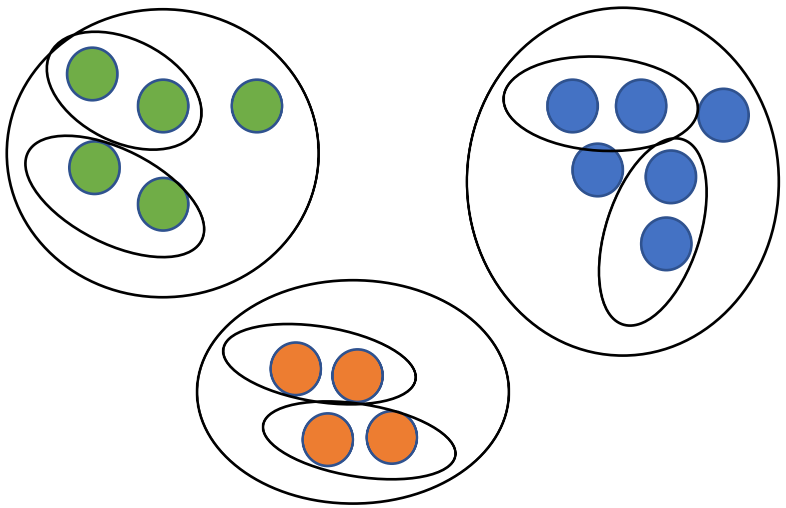


Create dendrogram

Tree hierarchy

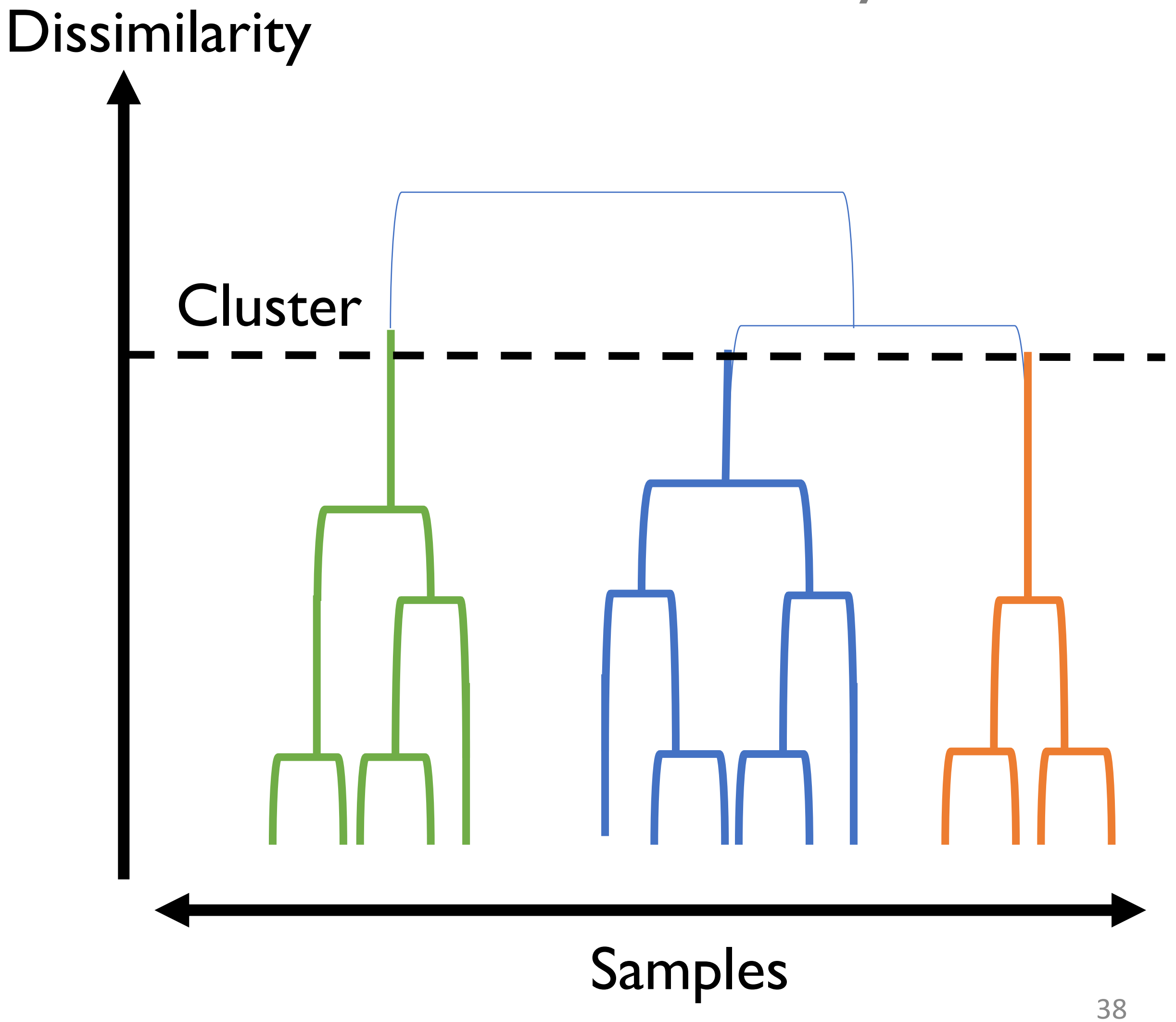


Hierarchical Clustering



Create dendrogram

Tree hierarchy



Hierarchical clustering algorithm (agglomerative)

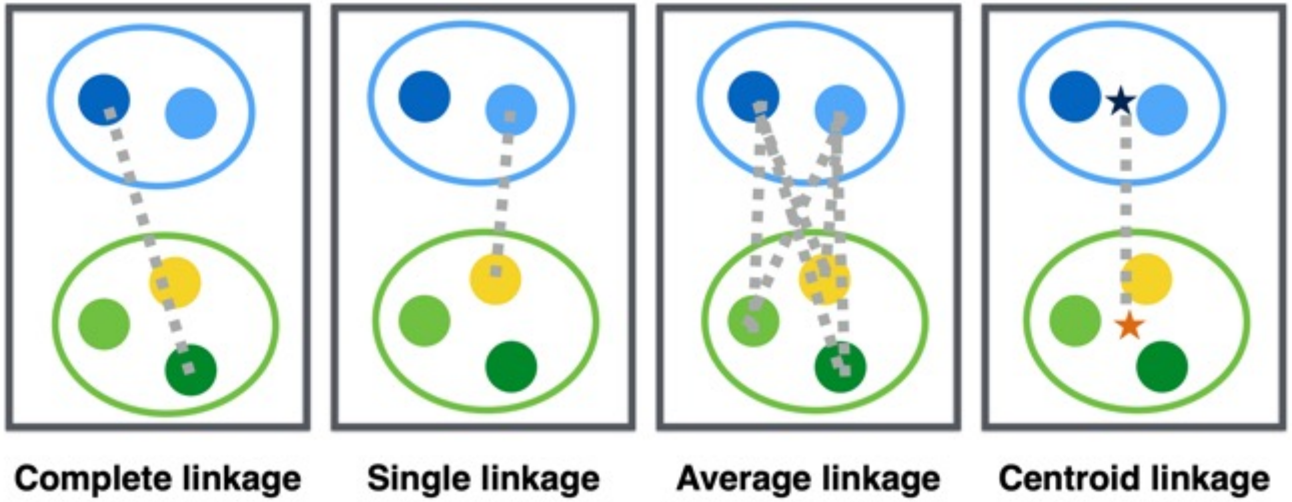
(0) Start with N clusters (each observation is a cluster)

(1) Repeat until 1 cluster left

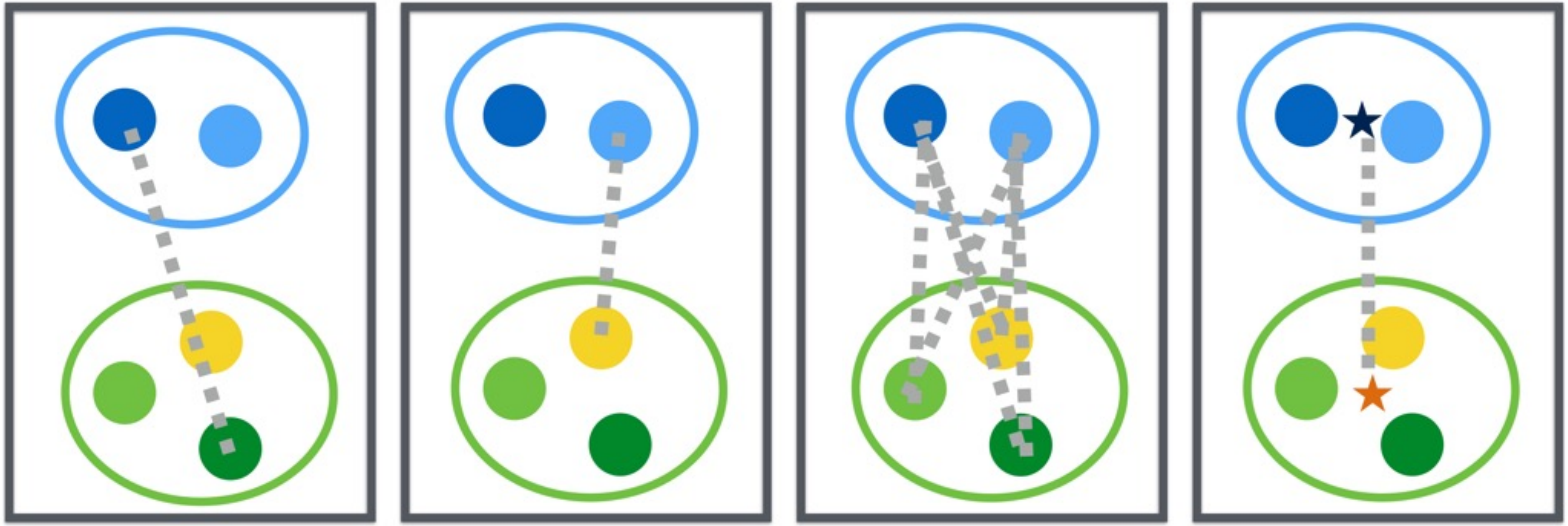
(a) Merge clusters with the least dissimilarity

(dissimilarity = height in dendrogram)

(b) Compute dissimilarity between **clusters = Linkage**



Hierarchical clustering algorithm (agglomerative)



Complete linkage	Single linkage	Average linkage	Centroid linkage
$\max_{\substack{x^{(i)}\in C_1,\\ x^{(j)}\in C_2}} d(x^{(i)}, x^{(j)})$	$\min_{\substack{x^{(i)}\in C_1,\\ x^{(j)}\in C_2}} d(x^{(i)}, x^{(j)})$	$\sum_{\substack{x^{(i)}\in C_1,\\ x^{(j)}\in C_2}} \frac{d\left(x^{(i)}, x^{(j)}\right)}{\lvert C_1\rvert\,\lvert C_2\rvert}$	$d\left(\sum_{x^{(i)}\in C_1} \frac{x^{(i)}}{\lvert C_1\rvert}, \sum_{x^{(j)}\in C_2} \frac{x^{(j)}}{\lvert C_2\rvert}\right)$

Hierarchical clustering algorithm (agglomerative)

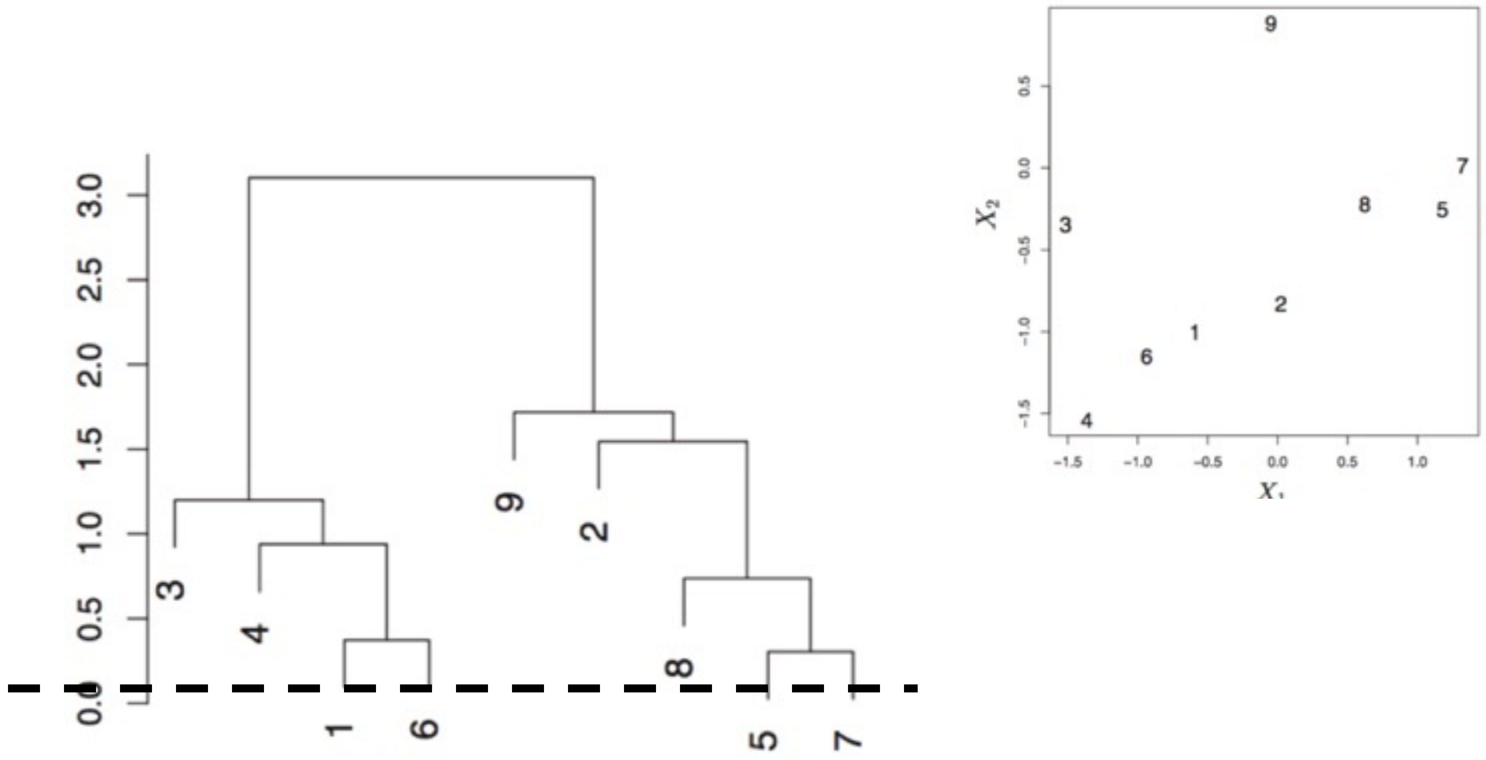
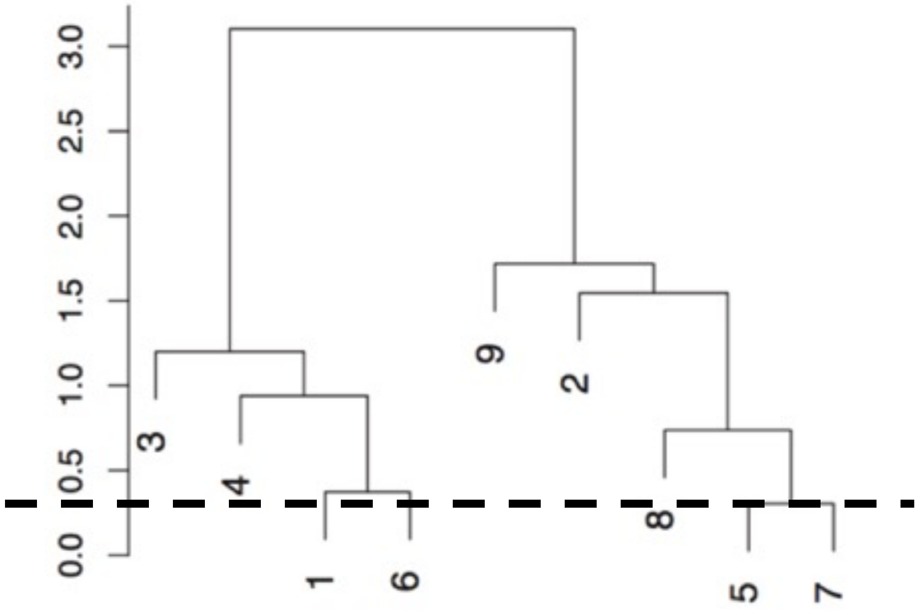


FIGURE 10.11, ISL (8th printing 2017)

Hierarchical clustering algorithm (agglomerative)



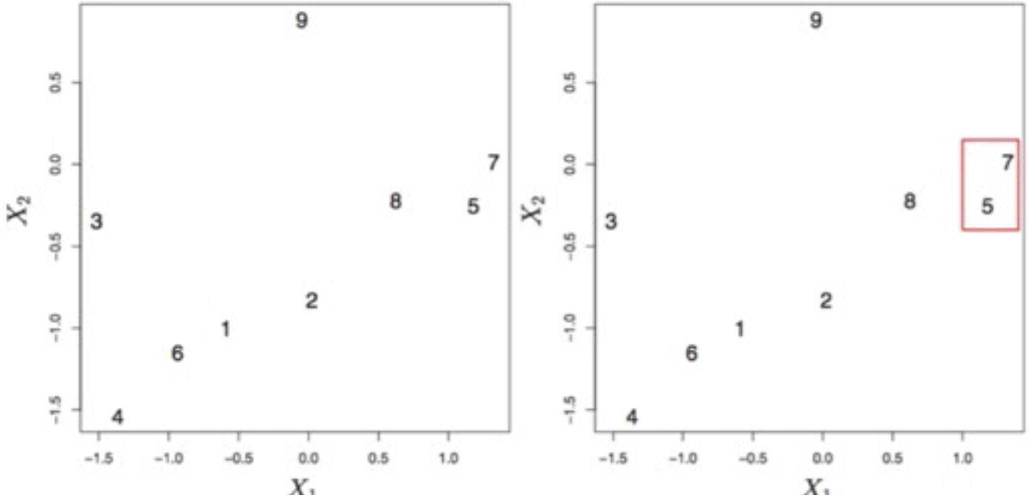
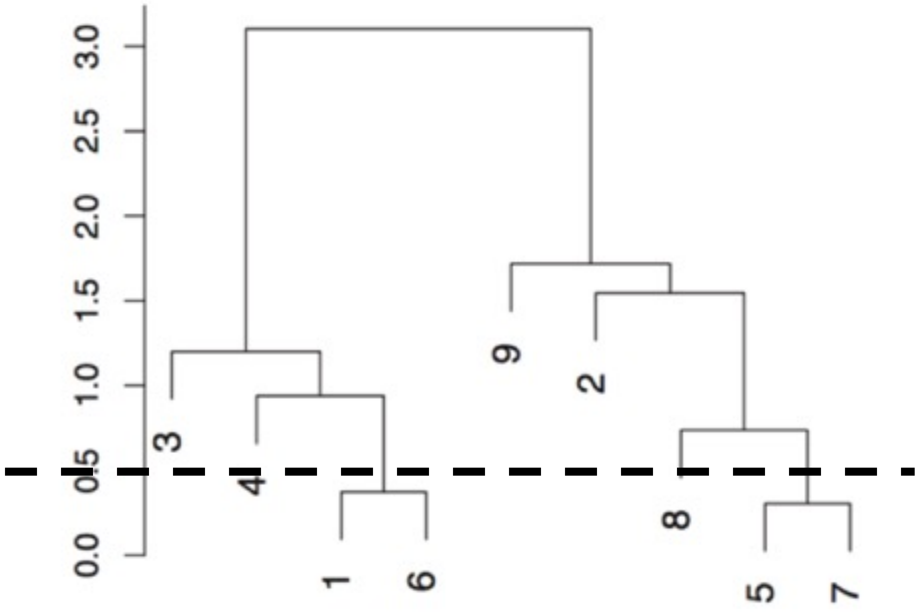


FIGURE 10.11, ISL (8th printing 2017)

Hierarchical clustering algorithm (agglomerative)



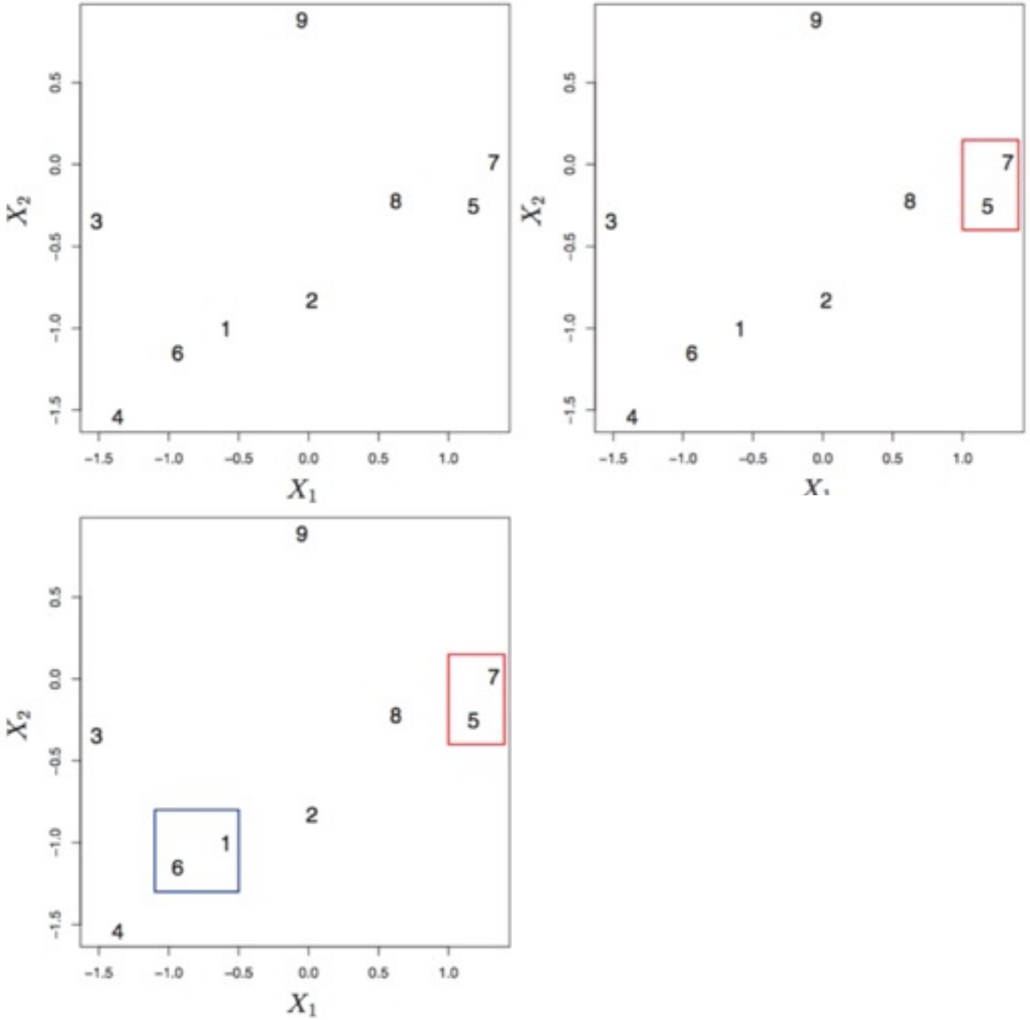
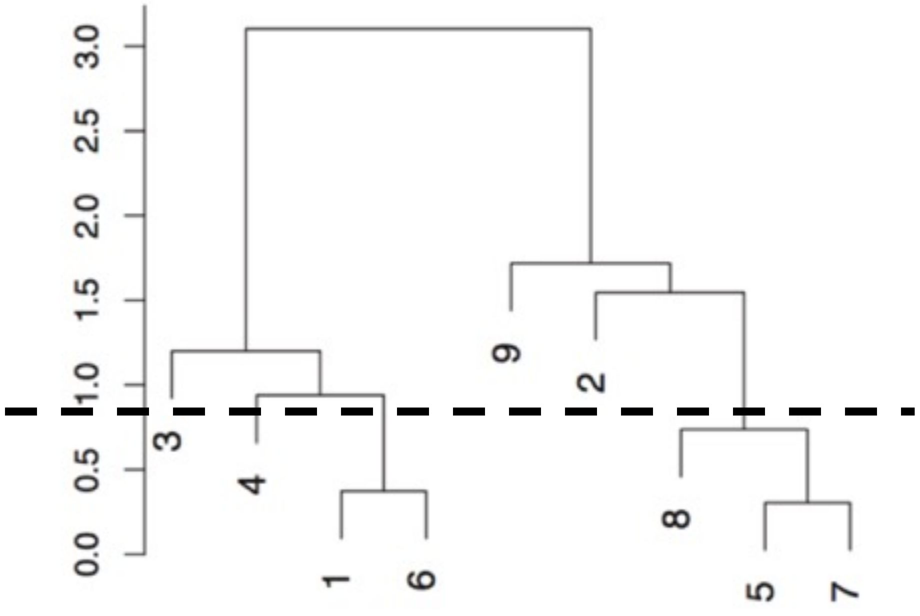


FIGURE 10.11, ISL (8th printing 2017)

Hierarchical clustering algorithm (agglomerative)



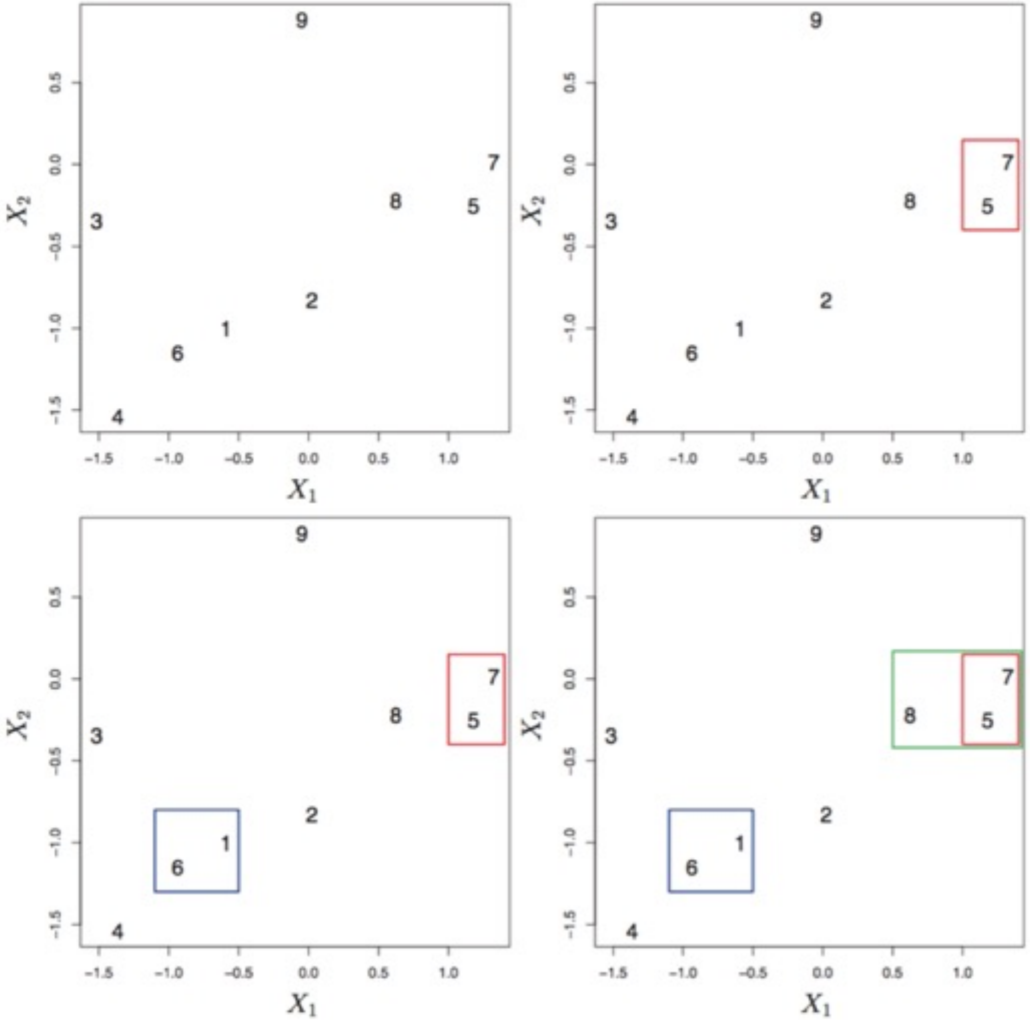


FIGURE 10.11, ISL (8th printing 2017)

Challenges of Hierarchical Clustering

Sensibility to dissimilarity & linkage

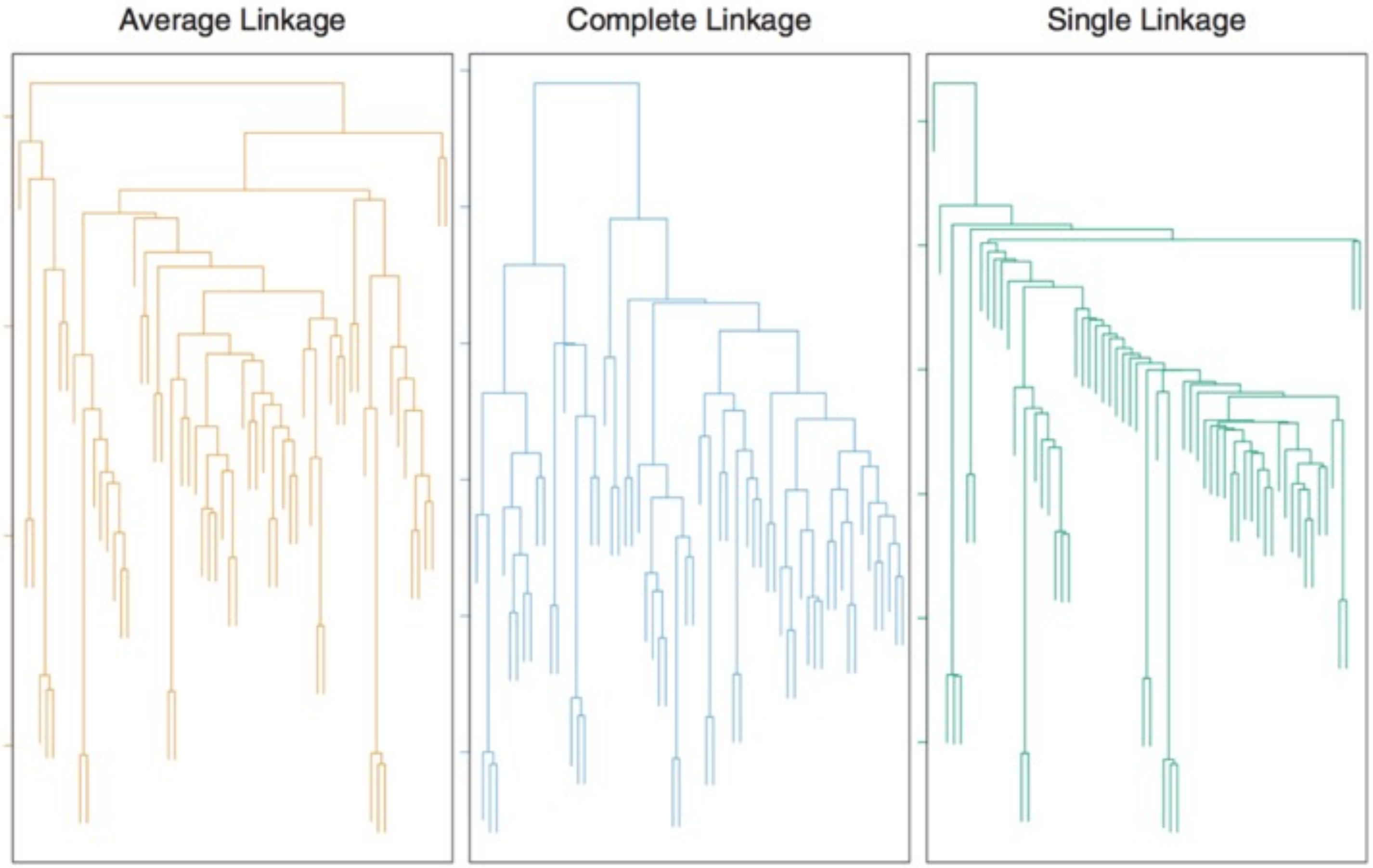


FIGURE 10.12, ISL (8th printing 2017)

Recompute linkage at each step.

Types of clustering algorithms

Subgroups that explain variation

K-means

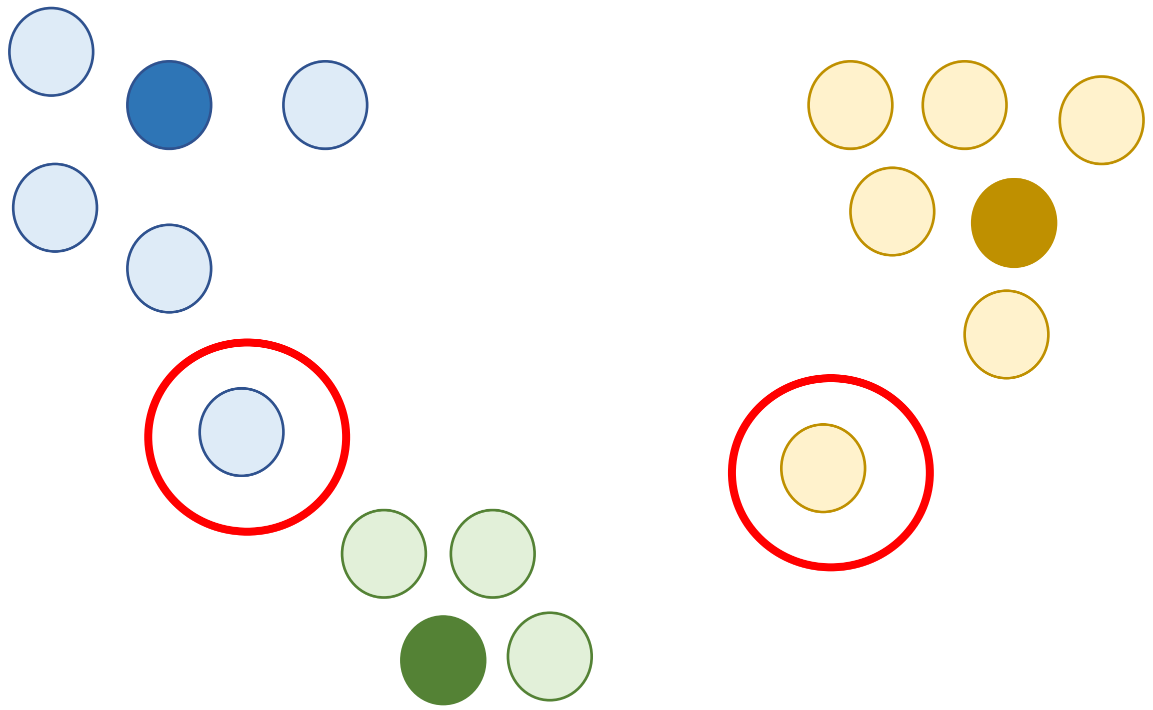
Hierarchical

How to check robustness?

- ✓ How clusters change using subsets of data
- ✓ How clusters change changing parameters

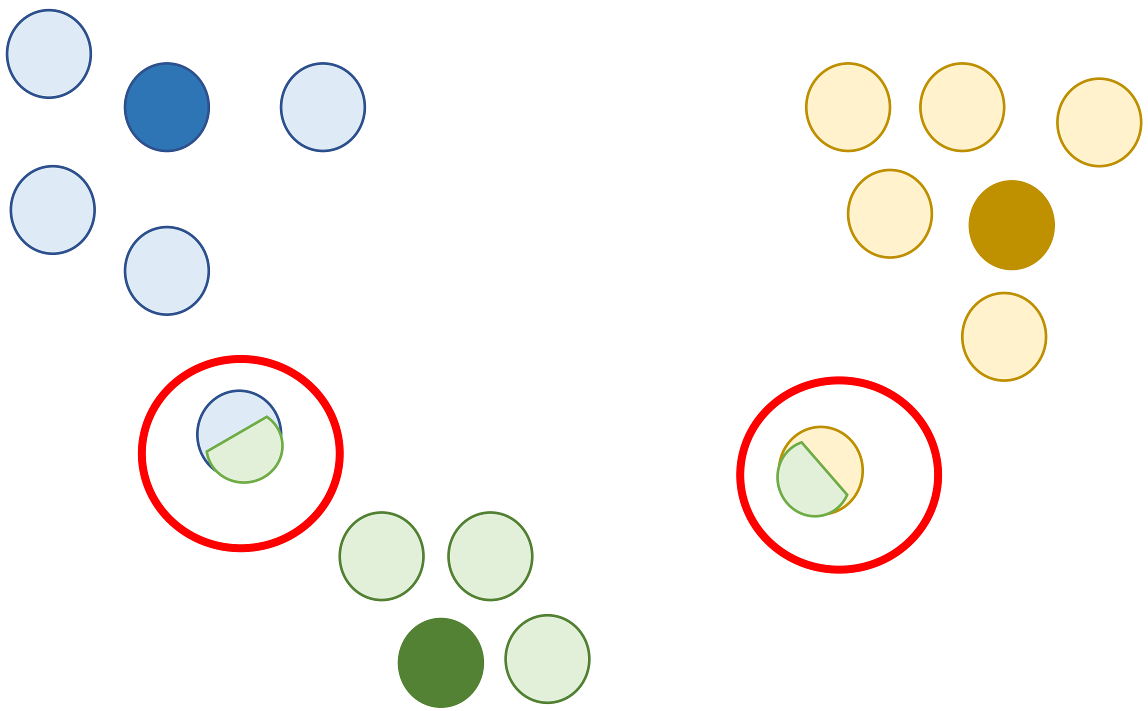
Hard clustering vs soft clustering

Hard Clustering



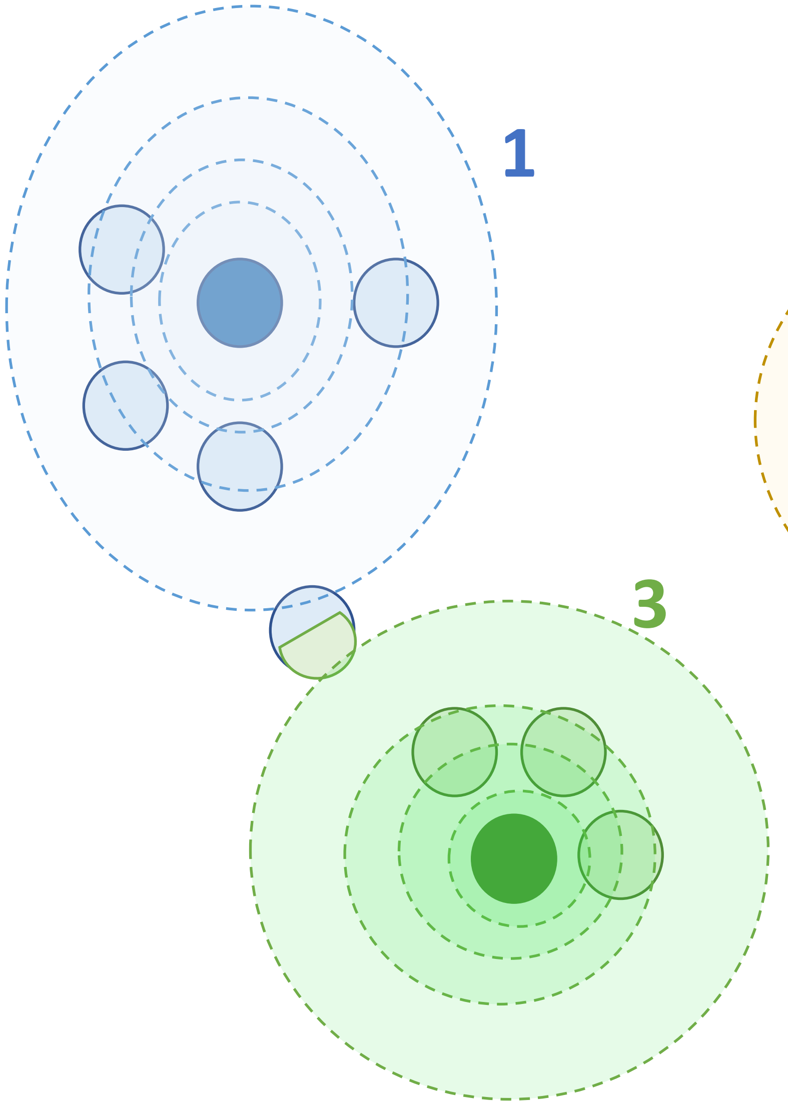
K-means, Hierarchical

Soft Clustering



Gaussian Mixture models

Gaussian Mixture Models (Informally)



Assumption: Mixture model

$$
\begin{aligned}
&P(X_i = x) \\
&= P(Z_i = 1)P(X_i = x|Z_i = 1) \\
&+ P(Z_i = 2)P(X_i = x|Z_i = 2) \\
&+ P(Z_i = 3)P(X_i = x|Z_i = 3)
\end{aligned}
$$

To do: Characterize

1. Marginal $P(X_i = x|Z_i = k)$: μ_k, Σ_k
2. Contribution $P(Z_i = k)$

https://stephens999.github.io/fiveMinuteStats/intro_to_em.html

GMM : Expectation-Maximization algorithm (Informally)

(0) Initialize marginals: μ_k, Σ_k (At random, another clustering)

(1) Iterate until convergence

(a) E-step: Responsibility/weight each observation i for each cluster j

$$\gamma_{Z_i}(k) = P(Z_i = k|X_i) = \frac{P(X_i|Z_i = k)P(Z_i=k)}{P(X_i)}$$ Bayes' rule

(b) M-step: Compute weighted mean and variance, using all observations

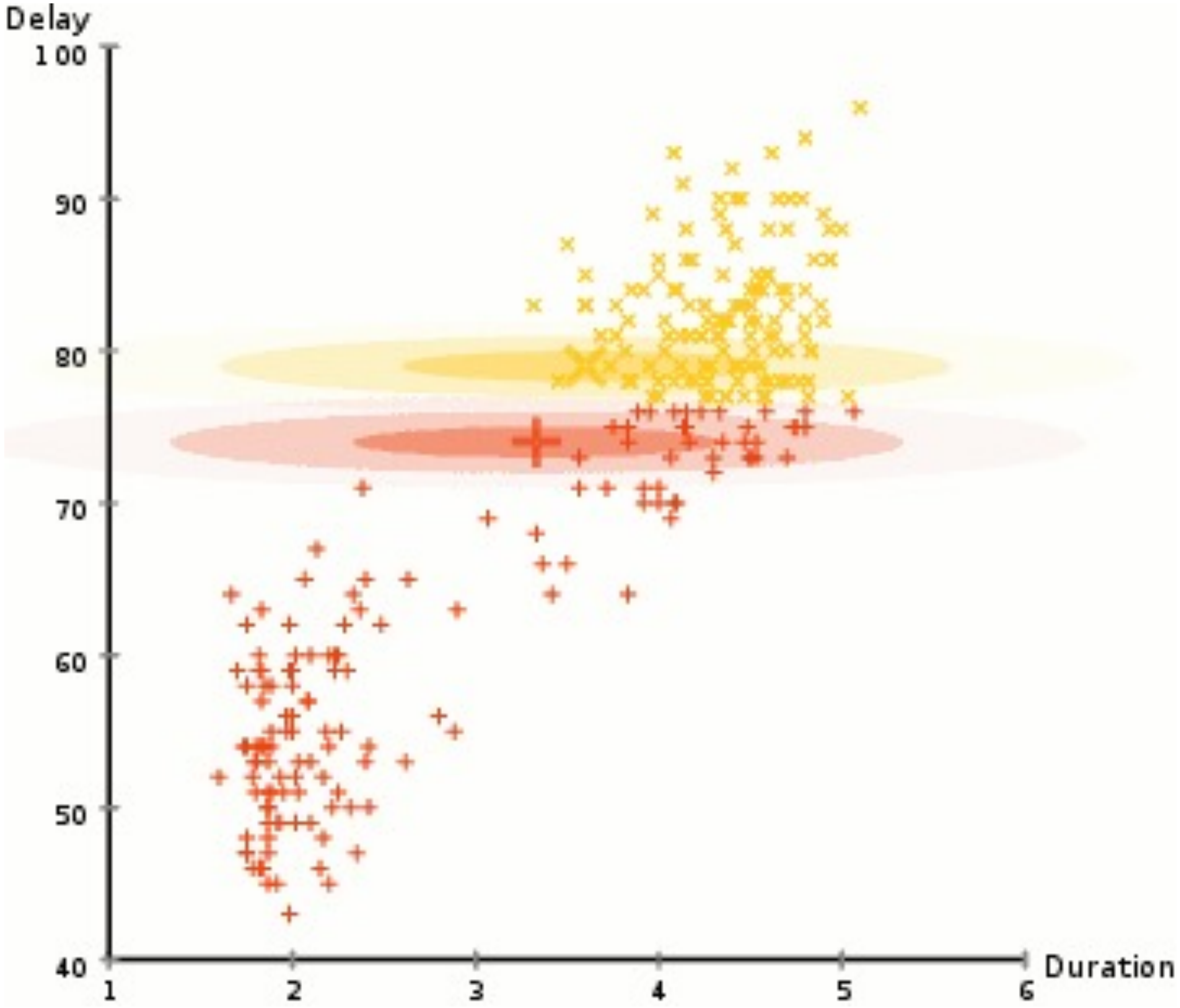
$$\hat{\mu_k} = \frac{\sum_{i=1}^{n} \gamma_{z_i}(k) x_i}{\sum_{i=1}^{n} \gamma_{z_i}(k)} = \frac{1}{N_k} \sum_{i=1}^{n} \boxed{\gamma_{z_i}(k)} x_i$$

$$\hat{\sigma_k^2} = \frac{1}{N_k} \sum_{i=1}^{n} \boxed{\gamma_{z_i}(k)} (x_i - \mu_k)^2$$

$$\hat{\pi_k} = \frac{N_k}{n} = P(Z_i = k)$$

https://stephens999.github.io/fiveMinuteStats/intro_to_em.html

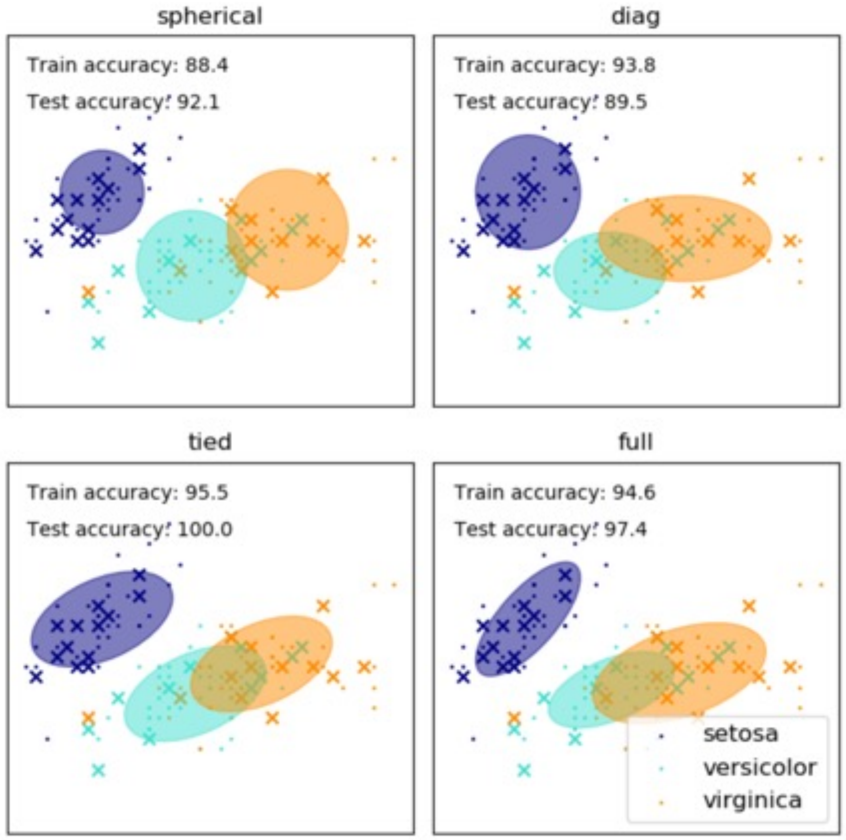
GMM : Expectation-Maximization algorithm (Informally)



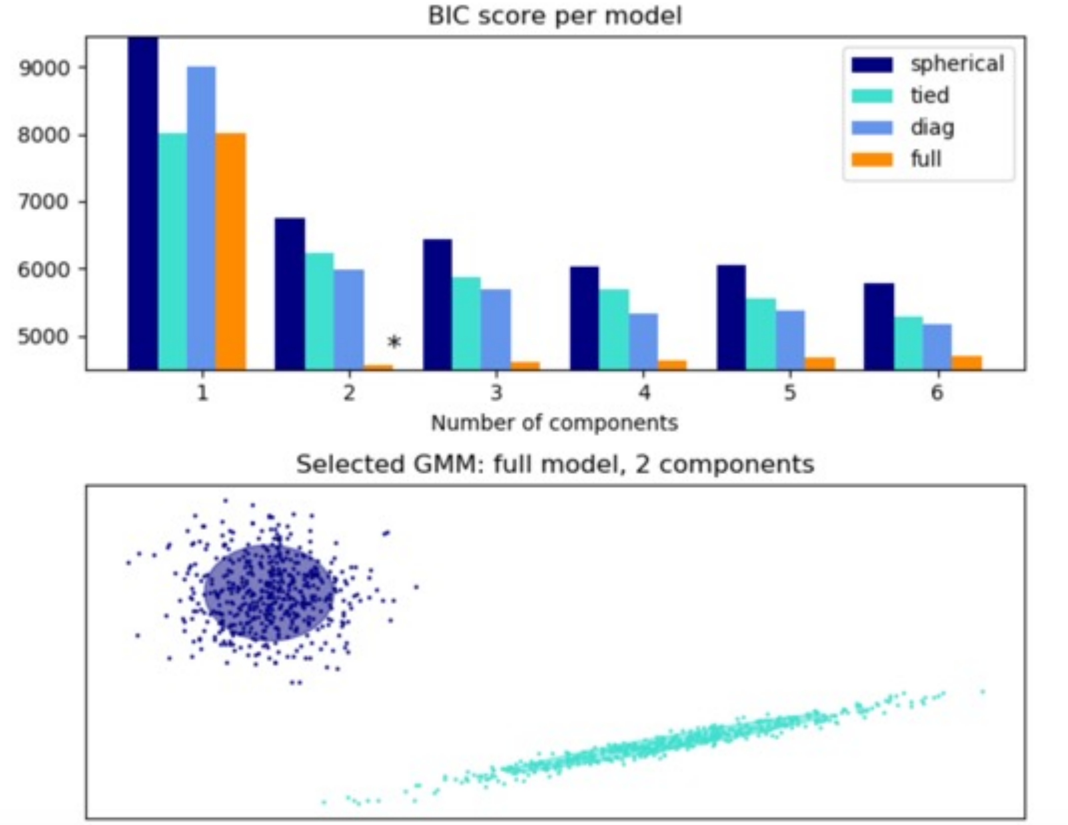
https://commons.wikimedia.org/wiki/File:EM_Clustering_of_Old_Faithful_data.gif

Challenges of Gaussian Mixture Models

Select form of covariance matrix.



Fixed number of components



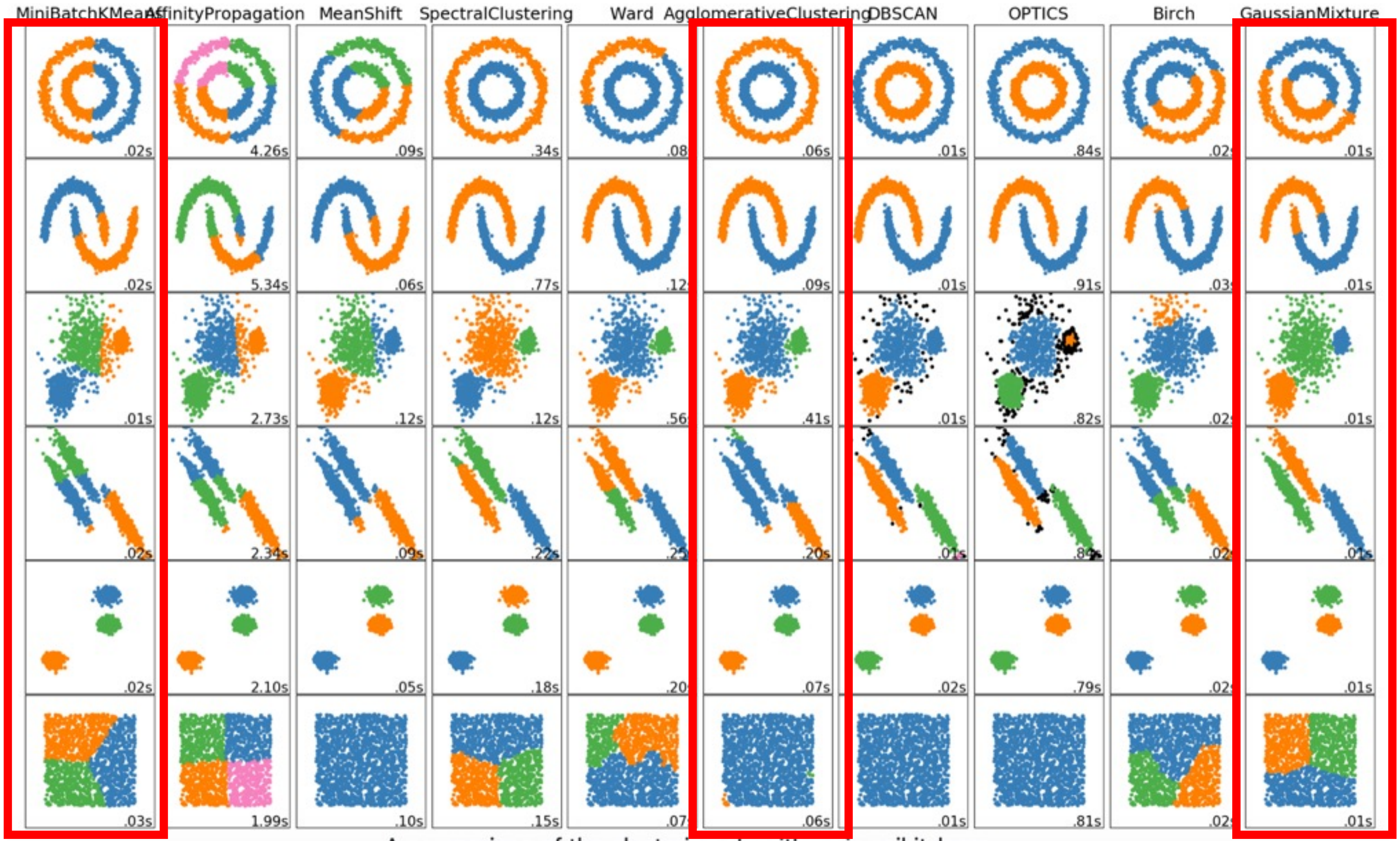
Code Available https://scikit-learn.org/stable/modules/mixture.html

Recap

Patterns + Properties in Data

- Clustering
  - Hard
    - K-means
      Prototypes
    - Hierarchical
      Dendrograms
  - Soft
    - GMMs
      Mixture Distribution
- Dimensionality Reduction

Dissimilarity/Similarity



A comparison of the clustering algorithms in scikit-learn

Code Available https://scikit-learn.org/stable/modules/clustering.html