

Text Clustering

Sameer Singh and Conal Sathi

BANA 290: ADVANCED DATA ANALYTICS

MACHINE LEARNING FOR TEXT

SPRING 2018

May 8, 2018

Upcoming...

Homework

- Homework 2 has been out
- Due this week: May 11, 2017
- Not much activity on Piazza

Project

- Instructions for proposal are out
- Due: May 15th

Outline

Unsupervised Machine Learning

K-Means Clustering

DBSCAN Algorithm

Hierarchical Clustering

Outline

Unsupervised Machine Learning

K-Means Clustering

DBSCAN Algorithm

Hierarchical Clustering

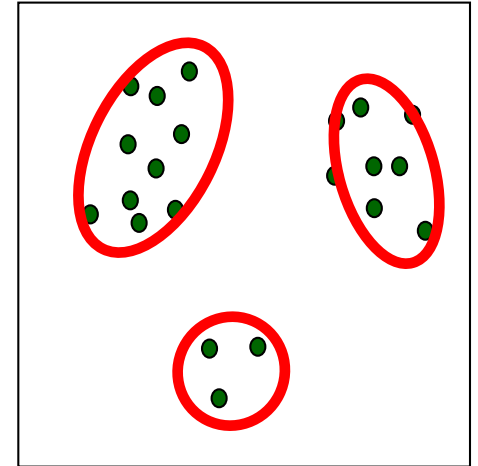
Unsupervised learning

Supervised learning

- Predict target value (“y”) given features (“x”)

Unsupervised learning

- Understand patterns of data (just “x”)
- Useful for many reasons
 - Data mining (“explain”)
 - Missing data values (“impute”)
 - Representation (feature generation or selection)



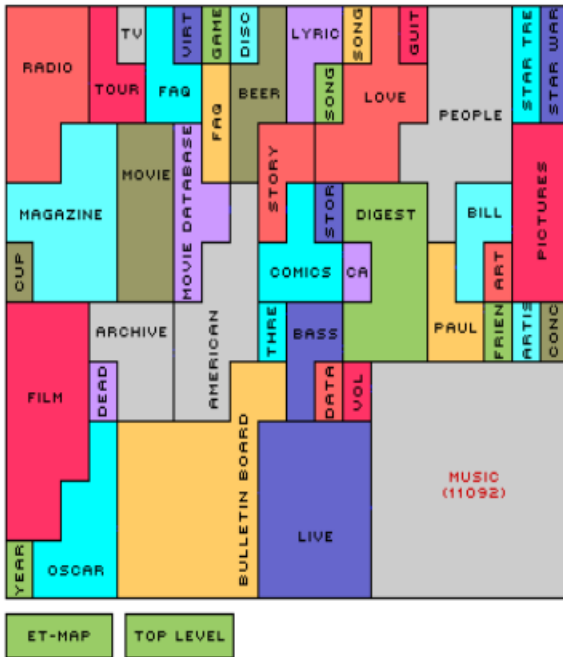
One example: *clustering*

- Describe data by discrete “groups” with some characteristics

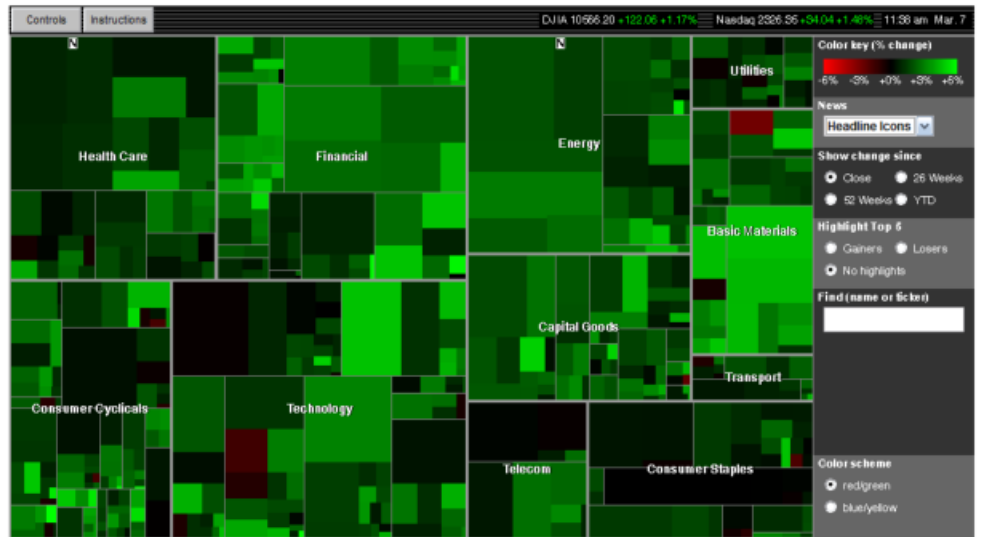
Text Clustering

- Whole corpus analysis/navigation
 - Better user interface
 - partition it into groups of related docs
- For improving recall in search applications
 - Better search results
- For better navigation of search results
- For speeding up vector space retrieval
 - Faster search

Clustering for Corpus



A Map of Yahoo!, Mappa.Mundi
Magazine, February 2000.



Map of the Market with Headlines

Clustering for Search



Results 1-20 of about 891,953 | [Details](#)

[Sources](#) [Sites](#) [Time](#) [Topics](#)

Top 354 Results [remix](#)

- + [Google-analytics.com](#) (90)
- + [Mining](#) (23)
- + [Marketing](#) (42)
- + [Customer](#) (29)
- + [Image](#) (29)
- + [Business intelligence](#) (17)
- + [Visualization](#) (11)
- + [Text Analytics API](#) (5)
- + [University](#) (10)
- + [Introduction to Text Analysis](#) (5)
- + [Cookies, Google Analytics](#) (13)
- + [Reviews](#) (6)
- + [Science](#) (8)
- + [Surveys](#) (10)
 - [Field](#) (4)
- + [Real Estate](#) (7)
- + [Book, E-Books](#) (5)
 - [Fraud](#) (4)

[Text Analytics API | Microsoft Azure](#) [new window](#) [preview](#)

Turn unstructured **text** into meaningful insights with the **Text Analytics** API from Microsoft Azure. Extract information with sentiment analysis and more.

<https://azure.microsoft.com/.../cognitive-services/text-analytics> - - Yippy Index V

[Text mining - Wikipedia](#) [new window](#) [preview](#)

Text mining, also referred to as **text** data mining, roughly equivalent to **text analytics**, is the process of deriving high-quality information from **text**.

https://en.wikipedia.org/wiki/Text_analytics - - Yippy Index V

[Text Analytics API overview \(Microsoft Cognitive Services ...\)](#) [new window](#) [preview](#)

Text Analytics API in Azure Cognitive Services for sentiment analysis, key phrase extraction, and language detection.

<https://docs.microsoft.com/.../Text-Analytics/overview> - - Yippy Index V


[Text Analytics | What is Text Analytics?](#) [new window](#) [preview](#)

Clarabridge pioneered **text analytics** and sentiment analysis and our **text** mining software remains a core component of all CEM services.



<https://www.clarabridge.com/text-analytics> - - Yippy Index V

What is **Text Analytics** ? - Editor Review User Reviews [new window](#) [preview](#)

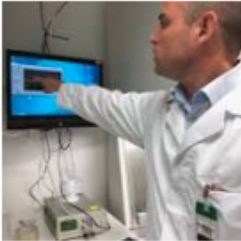
Clustering News



[Headlines](#) [Local](#) [For You](#) [U.S.](#)

 [Sign in](#) 


Top Stories



Muzzled watchdog: How killing the nuclear deal could make it easier for Iran to pursue the bomb in secret
Washington Post · 2h ago

RELATED COVERAGE

More progress likely in absence of nuclear deal
From Iran · Mehr News Agency - English Version · May 7, 2018



Eric Schneiderman, Accused by 4 Women, Quits as New York Attorney General
New York Times · 9h ago

RELATED COVERAGE

Four Women Accuse New York's Attorney General of Physical Abuse
Highly Cited · The New Yorker · 13h ago

In the News

Iran

Donald Trump

Met Gala

Israel

Eric Schneiderman

Melania Trump

North Korea

Don Blankenship

West Virginia

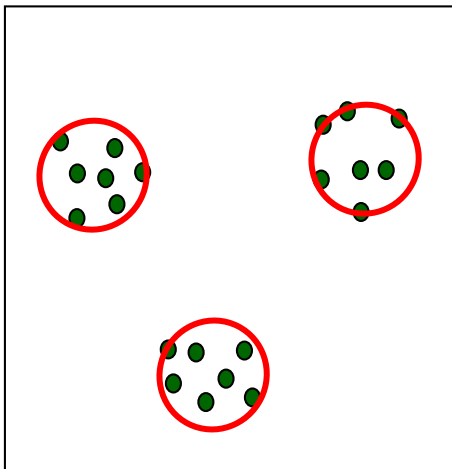
Kim Jong-un

Clustering

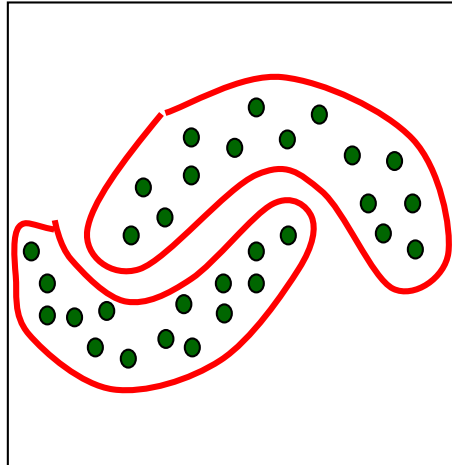
Clustering describes data by “groups”

The meaning of “groups” may vary by data!

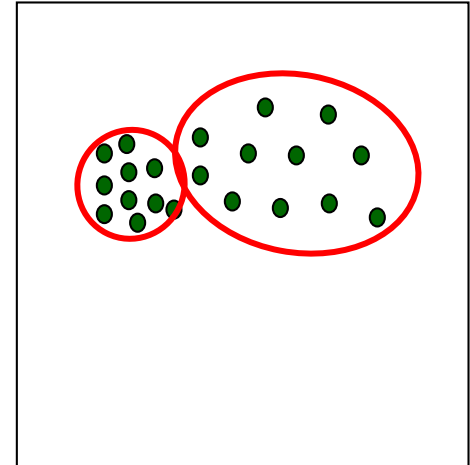
Examples



Location



Shape

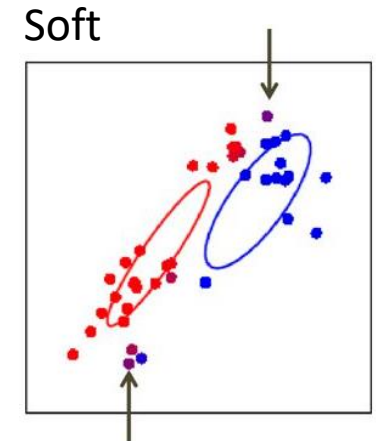
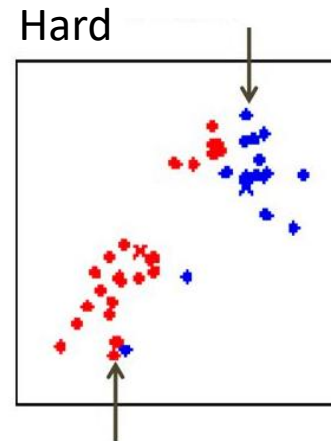


Density

Types of Clustering

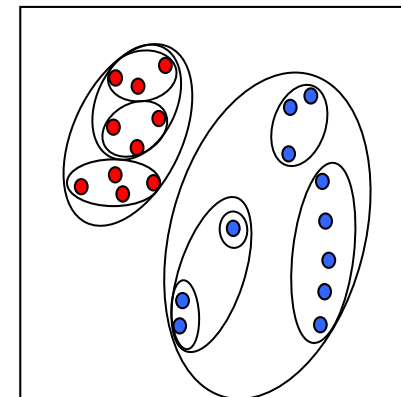
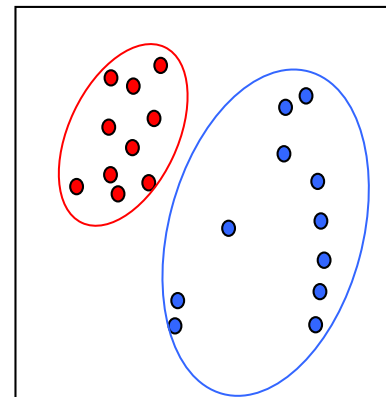
Hard vs Soft

- **Hard:** One cluster per point
- **Soft:** Weighted membership to all



Flat vs Hierarchical

- **Flat:** Fixed number of clusters
- **Hierarchical:** Tree of clusters



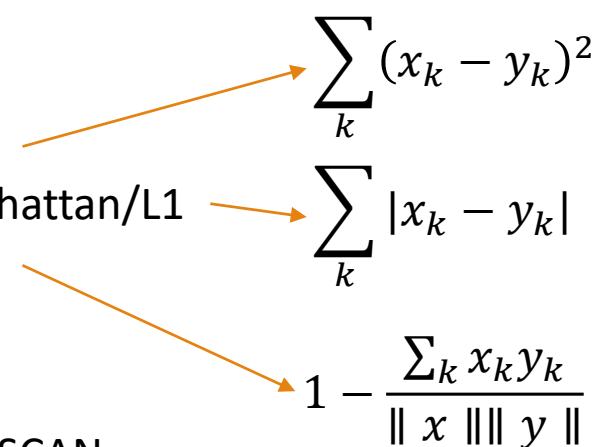
What do I need for Clustering?

What's a point? A vector!

- Bag of words, Ngrams
- TFIDF weighted, sometimes features
- ... next time, we'll learn vectors

Distance between vectors

- Euclidean/L2
- City-block/Manhattan/L1
- Cosine Distance


$$\sum_k (x_k - y_k)^2$$

$$\sum_k |x_k - y_k|$$

$$1 - \frac{\sum_k x_k y_k}{\|x\| \|y\|}$$

Algorithm

- Kmeans, DBSCAN
- Hierarchical Clustering
- ... many others, active area of research

In-Class Activity 1

Outline

Unsupervised Machine Learning

K-Means Clustering

DBSCAN Algorithm

Hierarchical Clustering

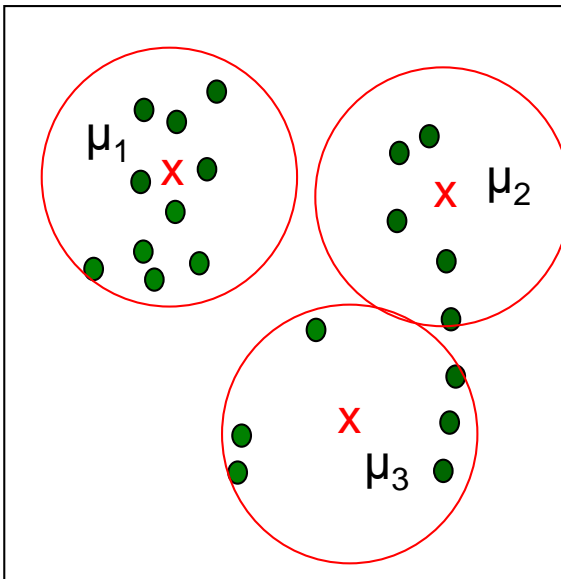
K-Means Clustering

A simple clustering algorithm

Hard and Flat

Iterate between

- Updating the assignment of data to clusters
- Updating the cluster's summarization



Notation:

Data example i has features x_i

Assume K clusters

Each cluster c “described” by a center μ_c

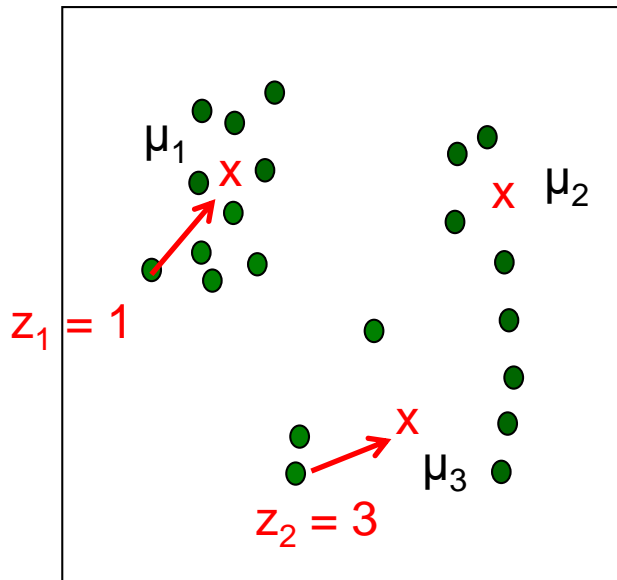
Each cluster will “claim” a set of nearby points

K-Means Clustering

A simple clustering algorithm

Iterate between

- Updating the assignment of data to clusters
- Updating the cluster's summarization



Notation:

- Data example i has features x_i
- Assume K clusters
- Each cluster c “described” by a center μ_c
- Each cluster will “claim” a set of nearby points
- “Assignment” of i^{th} example: $z_i = 1..K$

K-Means Clustering

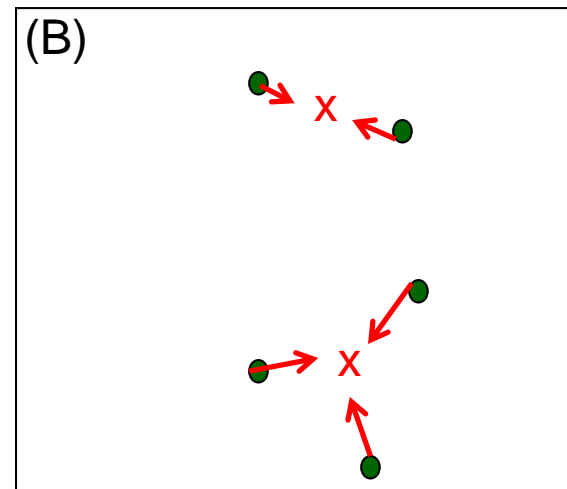
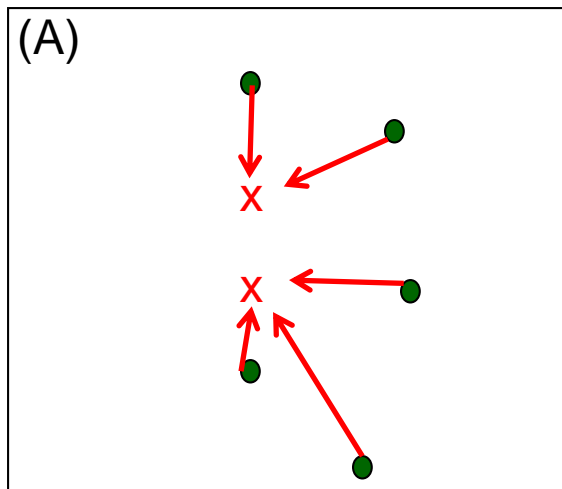
Iterate until convergence:

- (A) For each datum, find the closest cluster

$$z_i = \arg \min_c \|x_i - \mu_c\|^2 \quad \forall i$$

- (B) Set each cluster to the mean of all assigned data:

$$\forall c, \quad \mu_c = \frac{1}{m_c} \sum_{i \in S_c} x_i \quad S_c = \{i : z_i = c\}, \quad m_c = |S_c|$$



Demo Time!

<http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html>

Choosing Number of Clusters

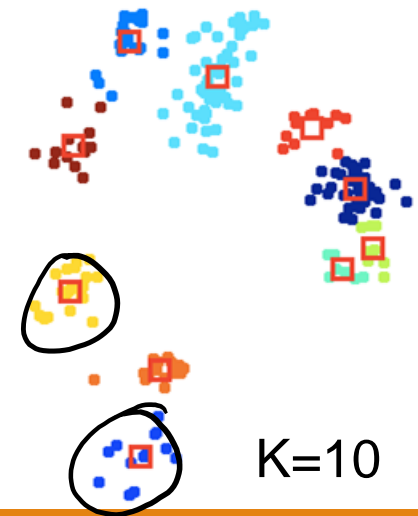
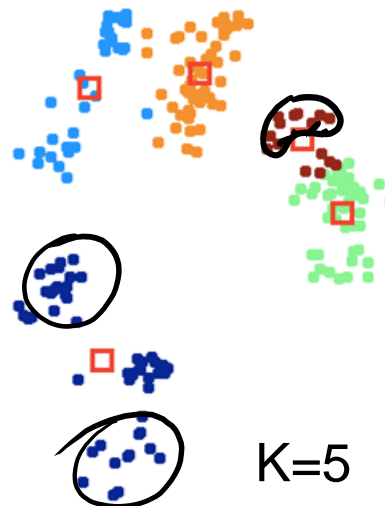
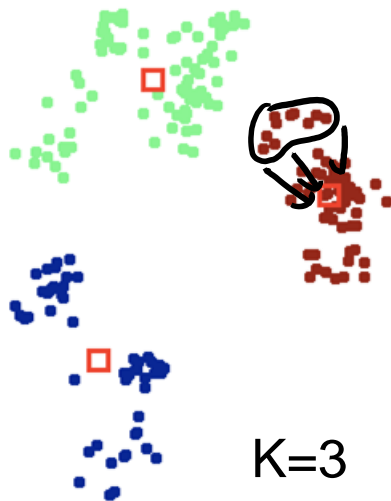
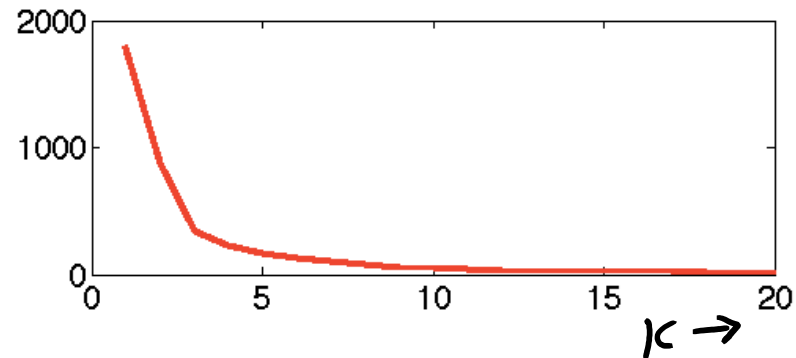
With cost function

$$C(\underline{z}, \underline{\mu}) = \sum_i \|x_i - \mu_{z_i}\|^2$$

what is the optimal value of k?

Cost always decreases with k!

A model complexity issue...



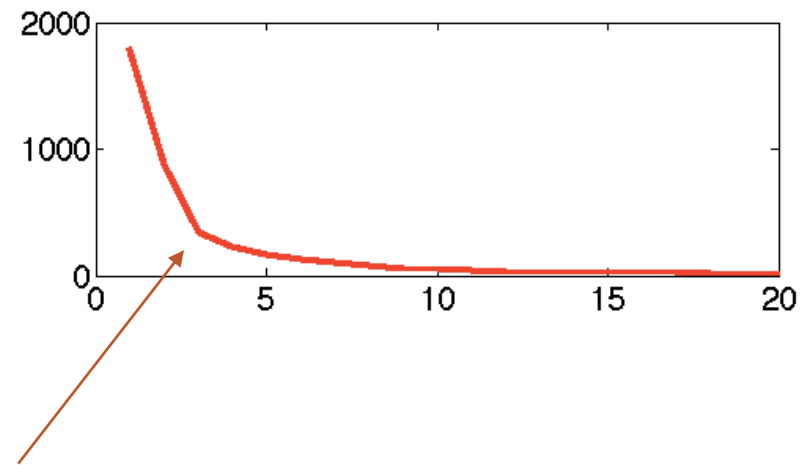
Choosing Number of clusters

With cost function $C(\underline{z}, \underline{\mu}) = \sum_i \|x_i - \mu_{z_i}\|^2$

what is the optimal value of k?

Cost always decreases with k!

A model complexity issue...



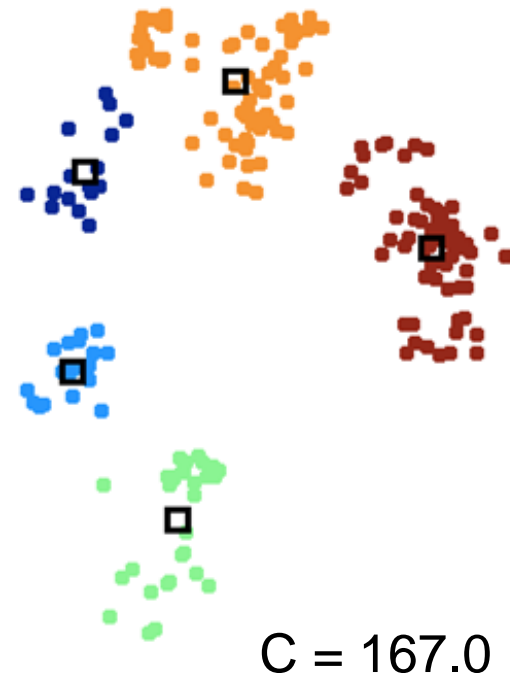
Pick the point of diminishing returns
i.e. the “elbow”

Initialization

Multiple local optima, depending on initialization

Try different (randomized) initializations

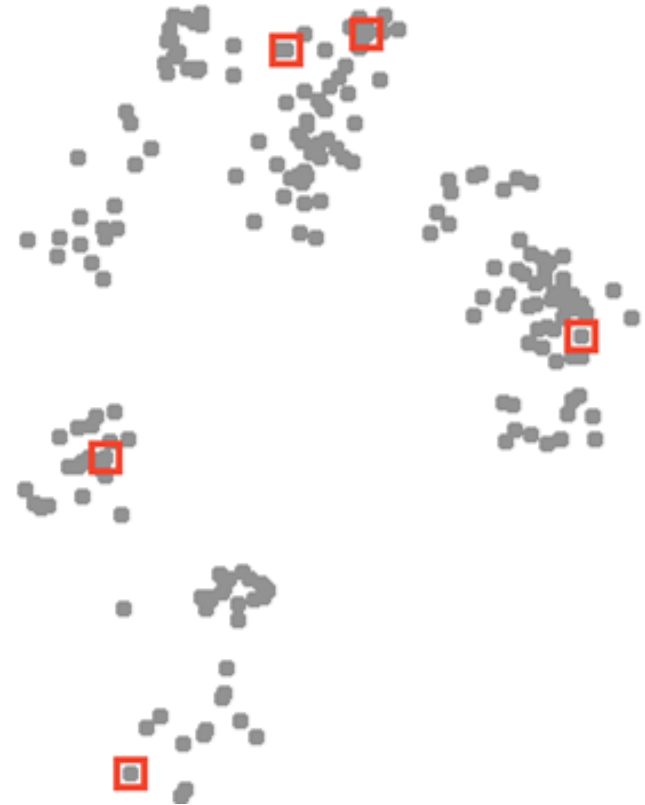
Can use cost C to decide which we prefer



Initialization methods

Random

- Usually, choose random data index
- Ensures centers are near some data
- Issue: may choose nearby points



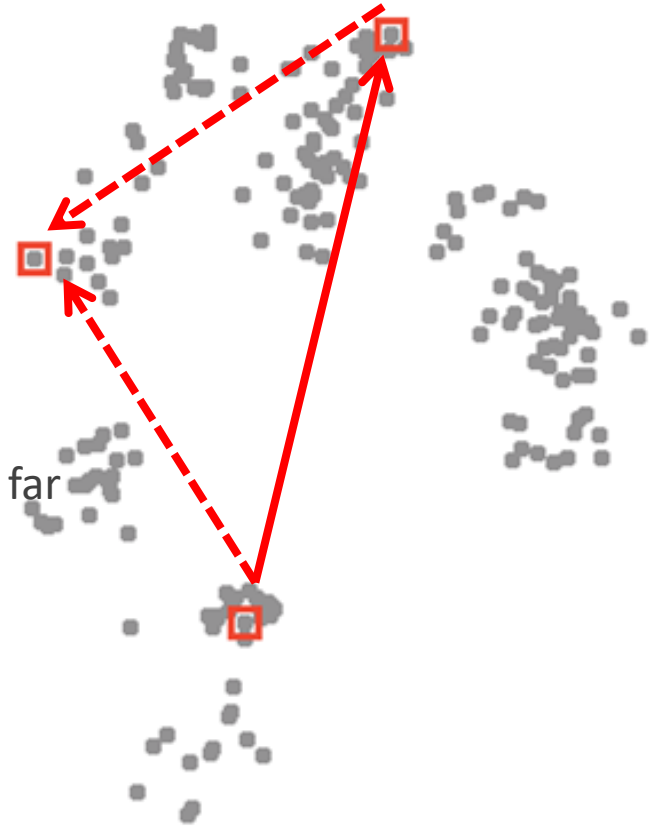
Initialization methods

Random

- Usually, choose random data index
- Ensures centers are near some data
- Issue: may choose nearby points

Distance-based

- Start with one random data point
- Find the point farthest from the clusters chosen so far
- Issue: may choose outliers



Initialization methods

Random

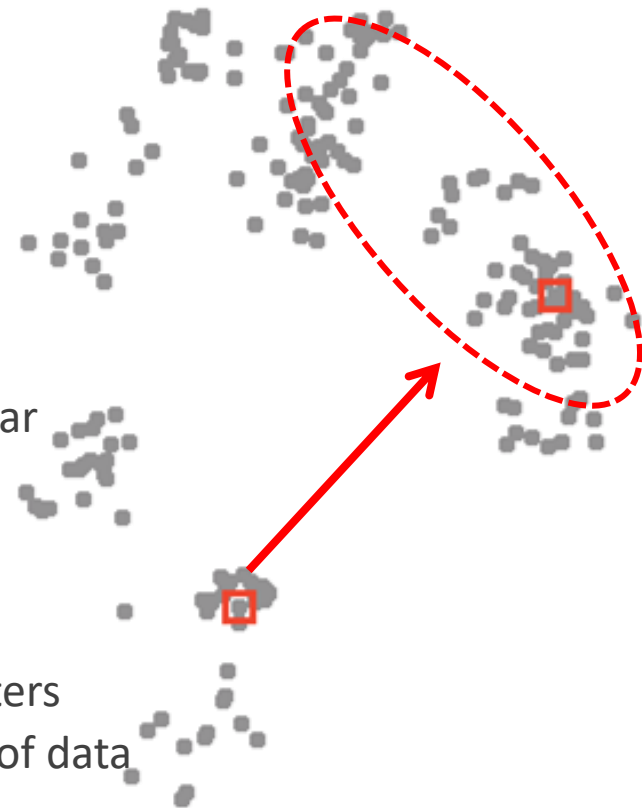
- Usually, choose random data index
- Ensures centers are near some data
- Issue: may choose nearby points

Distance-based

- Start with one random data point
- Find the point farthest from the clusters chosen so far
- Issue: may choose outliers

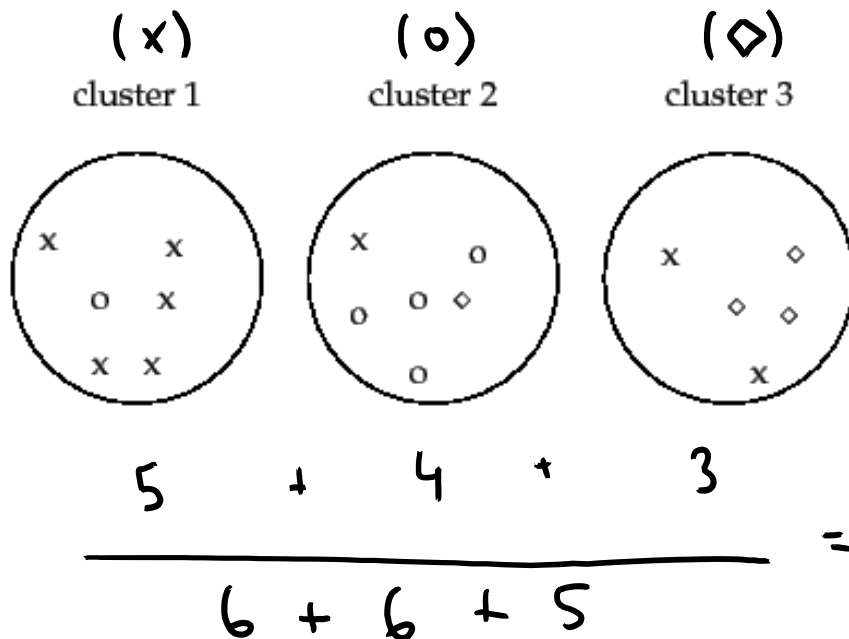
Random + distance (“k-means++”)

- Choose next points “far but randomly”
- $p(x) \propto \text{squared distance from } x \text{ to current centers}$
- Likely to put a cluster far away, in a region with lots of data



With Labels: Is Clustering Good?

What if we also have labeled data?

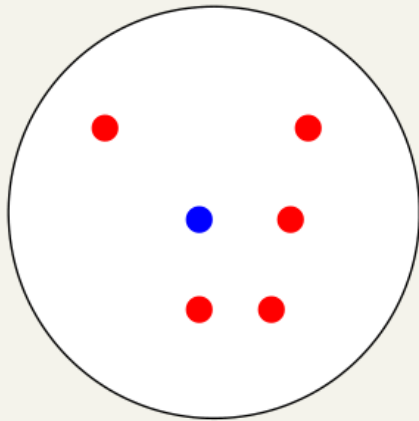


- For each cluster, get majority label
- Count points in that cluster that have that label
- Add them up, divide by total points

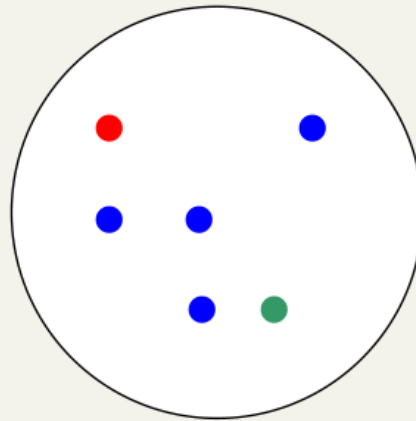
$$\text{Purity}(C, Y) = \frac{1}{N} \sum_k \max_j c_k \cap y_j$$

Points in cluster k Points with label j

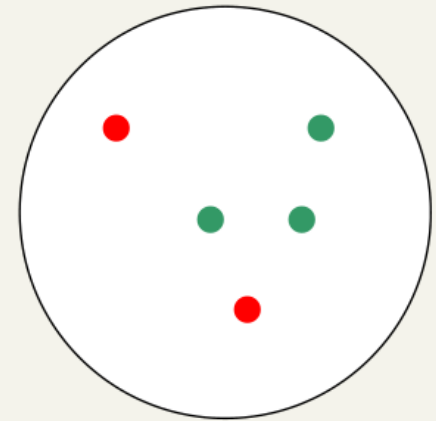
Another Example



Cluster I



Cluster II



Cluster III

$$\frac{5 + 4 + 3}{6 + 6 + 5}$$

Summary

K-Means clustering

- Clusters described as locations (“centers”) in feature space

Procedure

- Initialize cluster centers
- Iterate: assign each data point to its closest cluster center
- : move cluster centers to minimize mean squared error

Properties

- Always converges
- initialization important

Choosing the # of clusters, K

- The “elbow” method

Outline

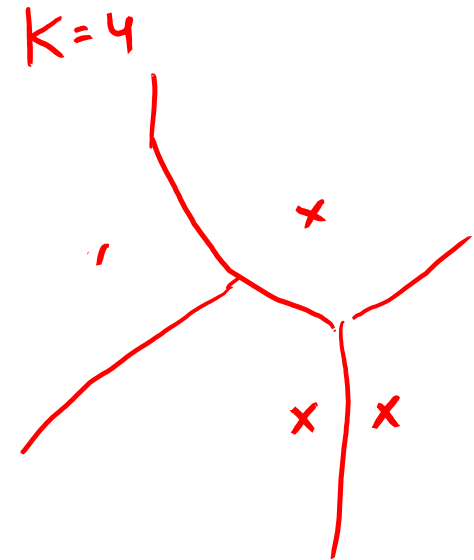
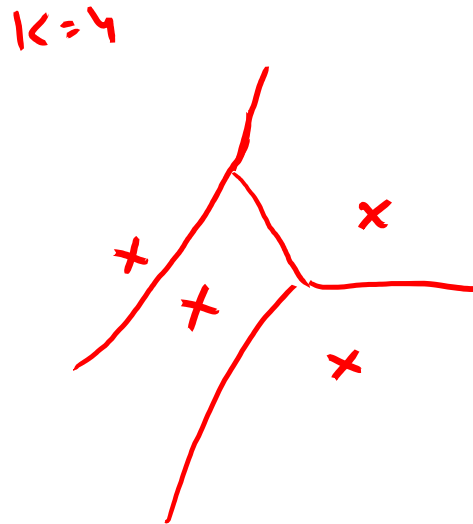
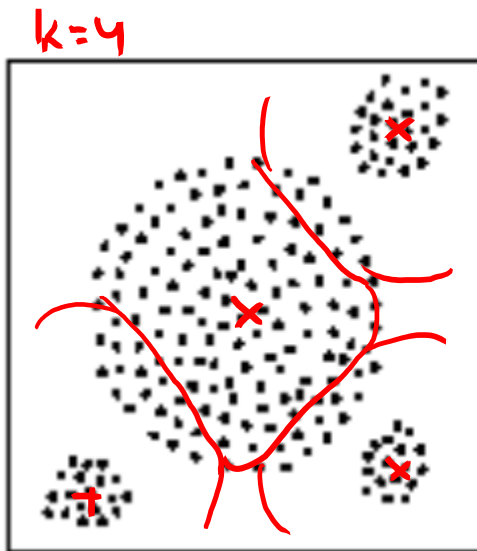
Unsupervised Machine Learning

K-Means Clustering

DBSCAN Algorithm

Hierarchical Clustering

Does K-Means always work?



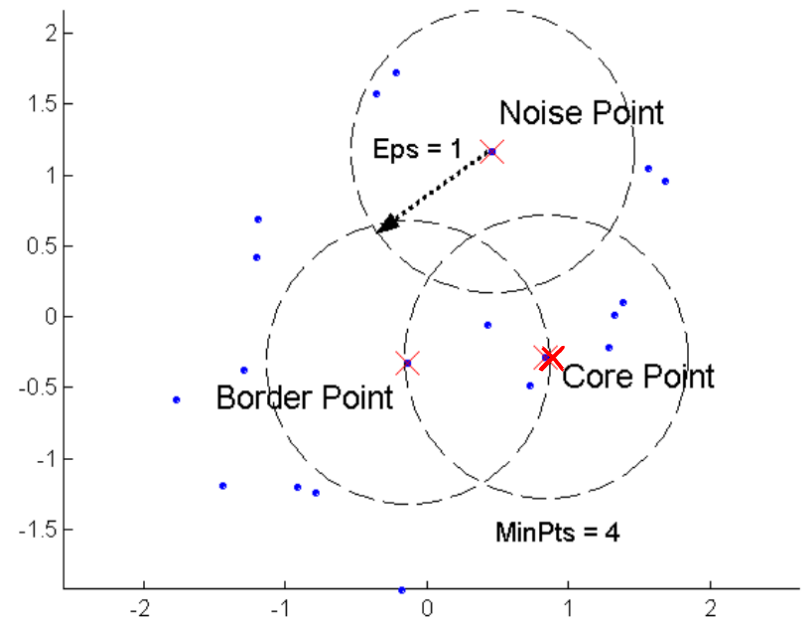
DBSCAN

Radius, ϵ

Min samples

No need for number of clusters

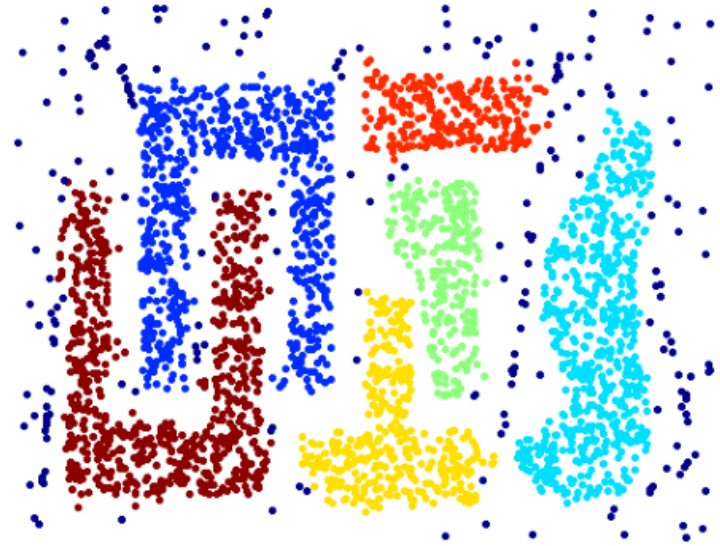
- Three types of points
- **Core points:** that have $>\text{min_samples}$ neighbors in radius ϵ
 - These are “interior” points
- **Border points:** $<\text{min_samples}$ neighbors (in radius ϵ), but has a core point
 - Defines the edges of the clusters
- **Noise points:** Any other points
 - These are outliers to the clusters



DBSCAN Example: Ideal Eps



Original Points

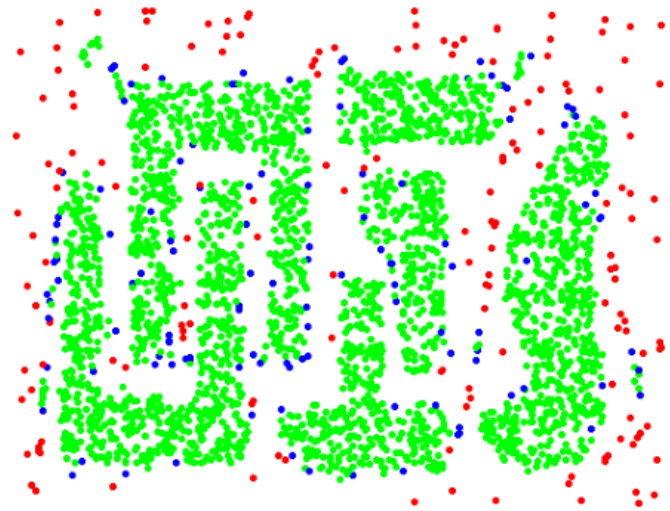


Clusters

DBSCAN Example: Eps too big



Original Points



Point types: **core**,
border and **noise**

In-Class Activity 2

Outline

Unsupervised Machine Learning

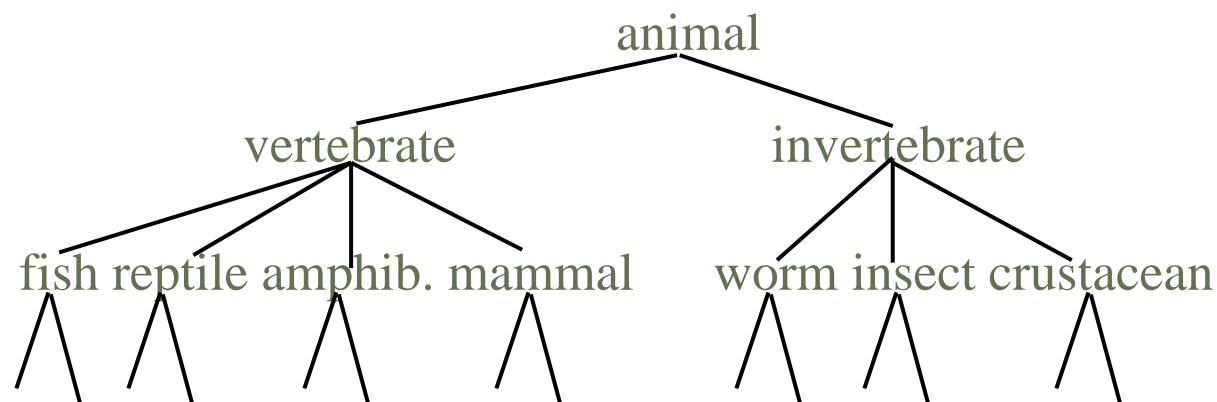
K-Means Clustering

DBSCAN Algorithm

Hierarchical Clustering

Hierarchical Clustering

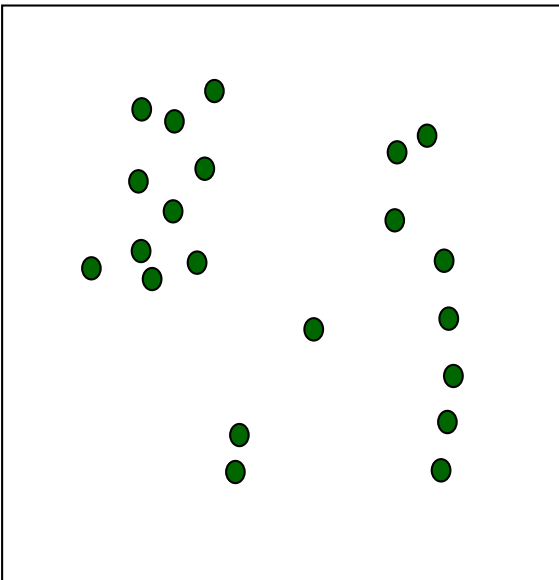
Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.



Hierarchical Agglomerative Clustering

Initially, every datum is a cluster

Data:

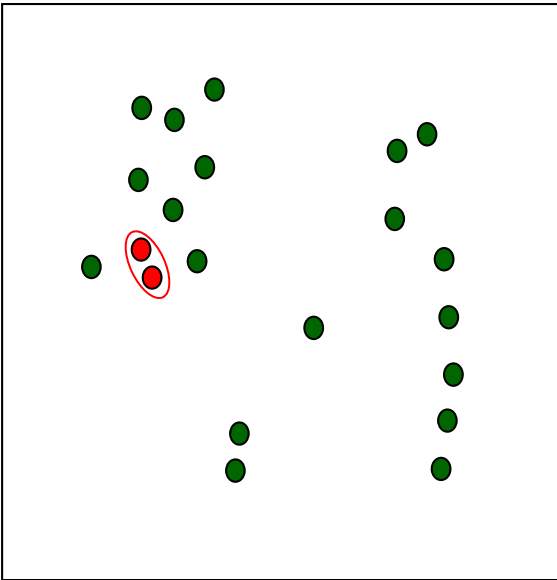


- A simple clustering algorithm
- Define a distance **between clusters**
- Initialize: every example is a cluster
- Iterate:
 - Compute distances between all clusters
 - Merge two closest clusters
- Save both clustering and sequence of cluster operations
- Result: “Dendrogram”

Iteration 1

Builds up a sequence of clusters (“hierarchical”)

Data:



Dendrogram:

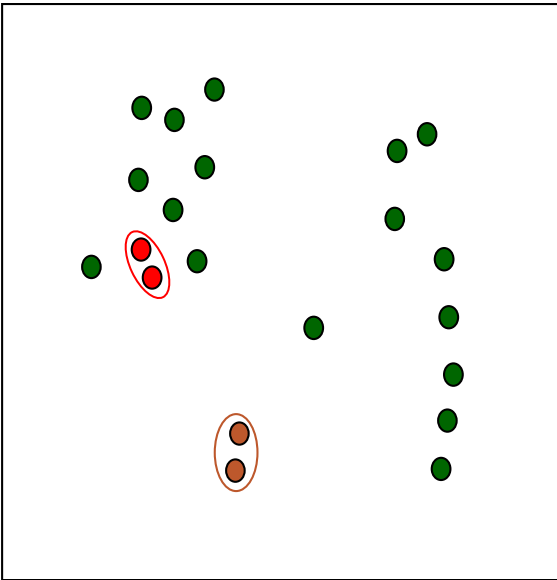


Height of the join
indicates dissimilarity

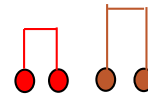
Iteration 2

Builds up a sequence of clusters (“hierarchical”)

Data:



Dendrogram:

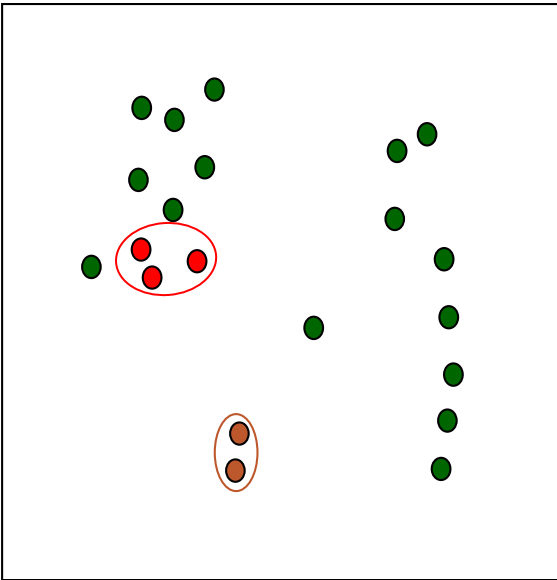


Height of the join
indicates dissimilarity

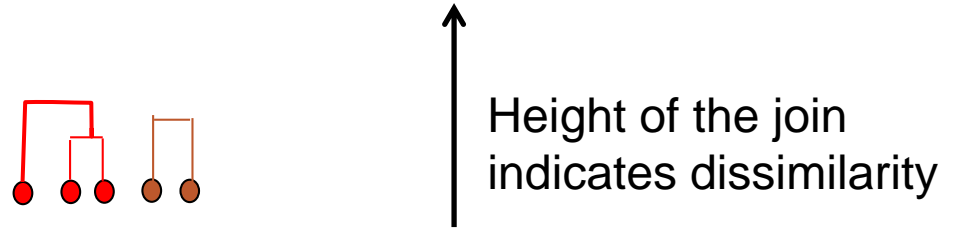
Iteration 3

Builds up a sequence of clusters (“hierarchical”)

Data:



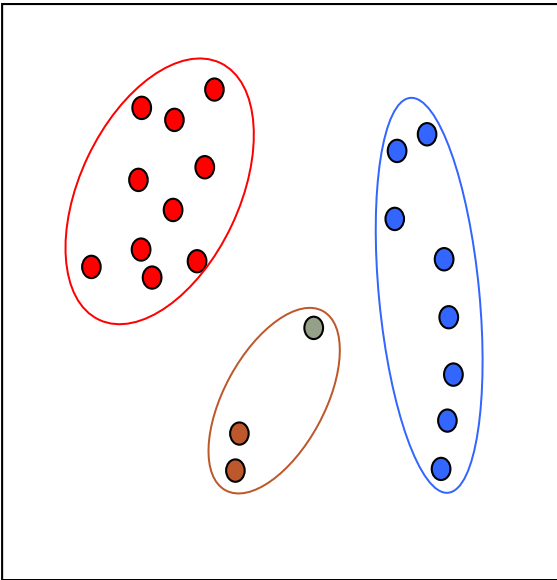
Dendrogram:



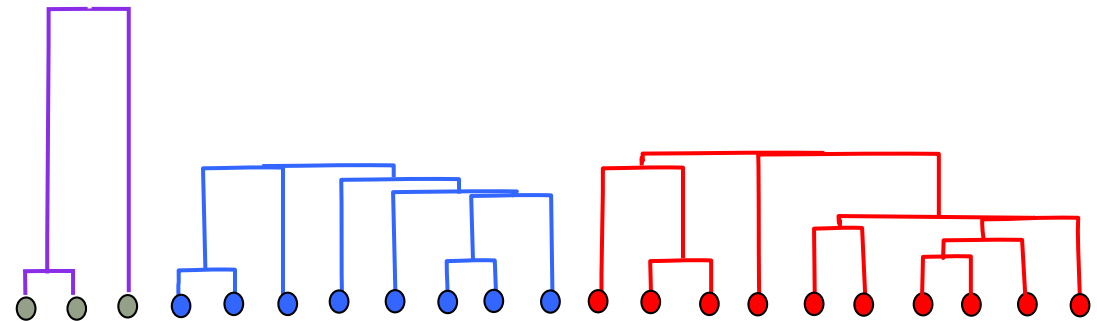
Iteration m-3

Builds up a sequence of clusters (“hierarchical”)

Data:



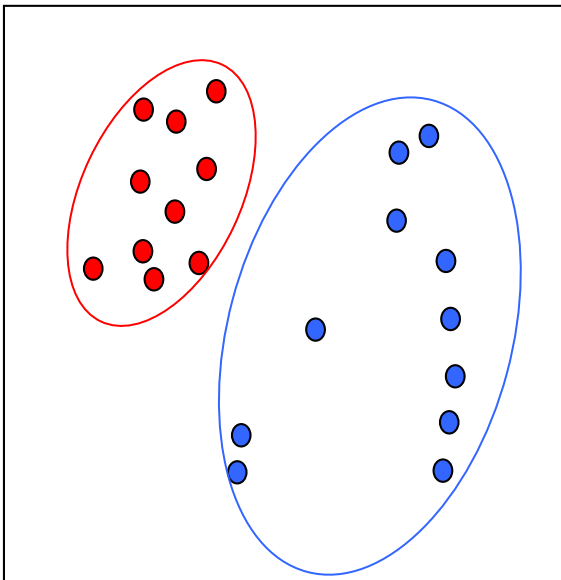
Dendrogram:



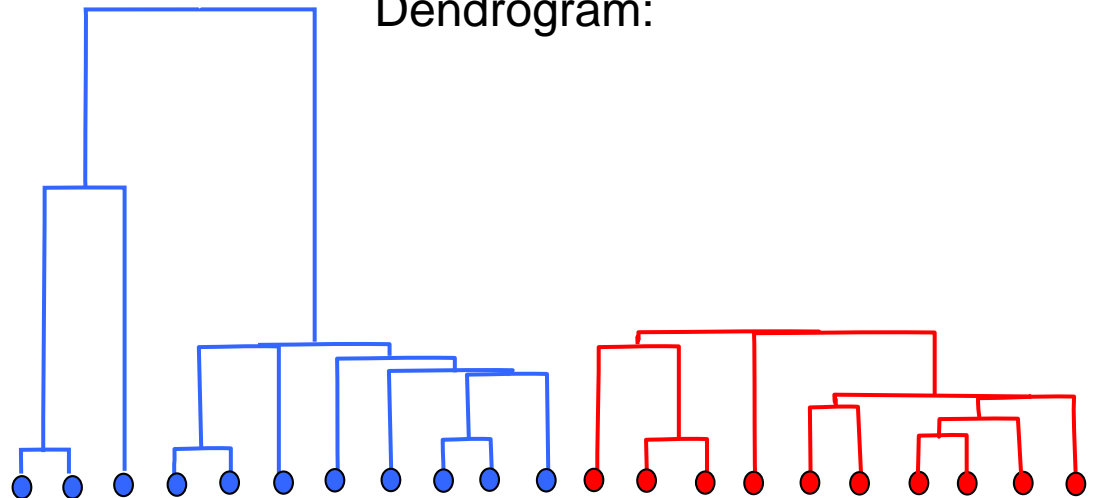
Iteration m-2

Builds up a sequence of clusters (“hierarchical”)

Data:



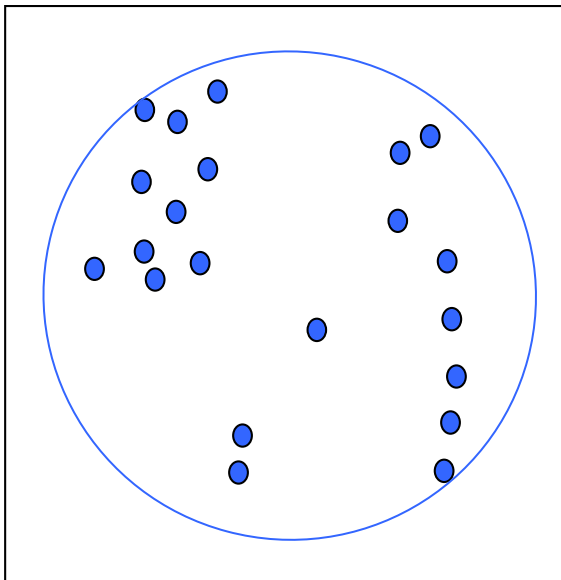
Dendrogram:



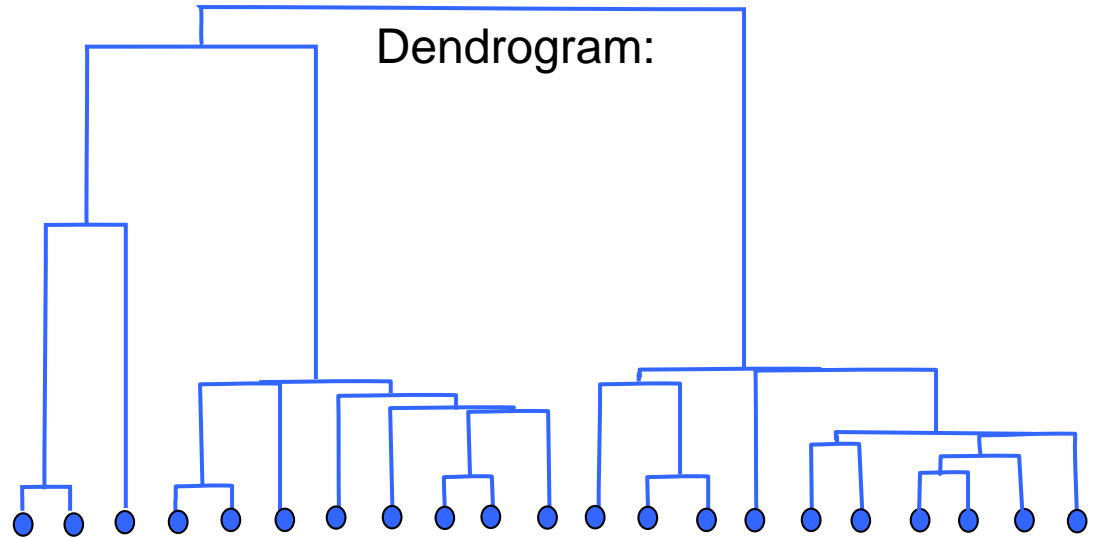
Iteration m-1

Builds up a sequence of clusters (“hierarchical”)

Data:



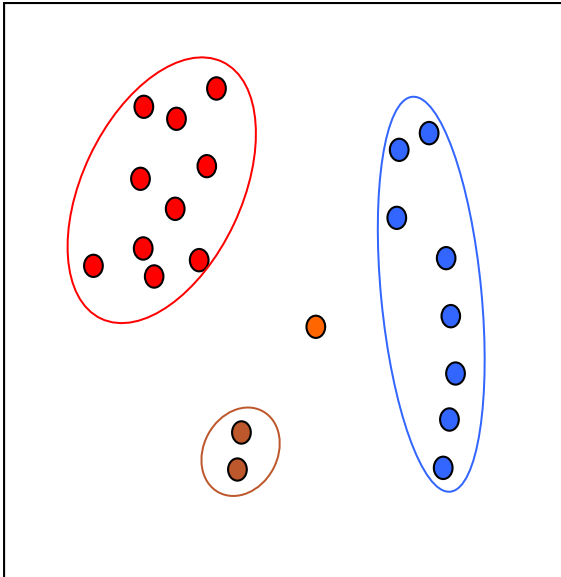
Dendrogram:



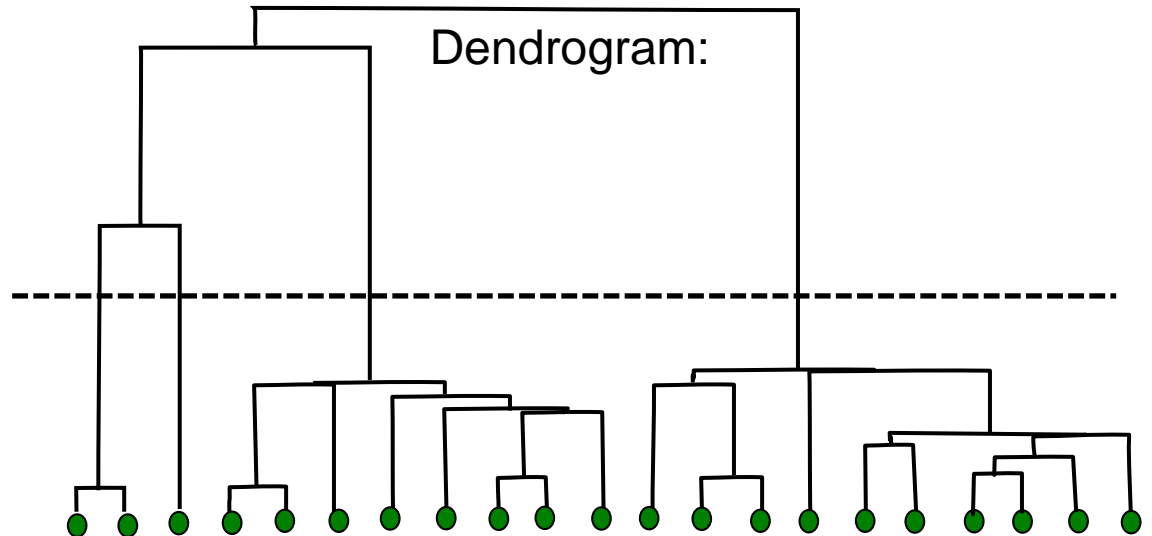
From dendrogram to clusters

Given the sequence, can select a number of clusters or a dissimilarity threshold:

Data:

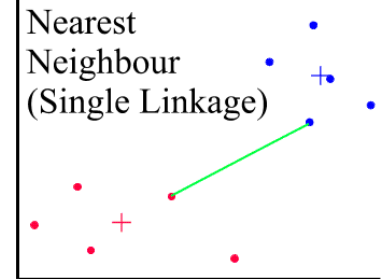


Dendrogram:



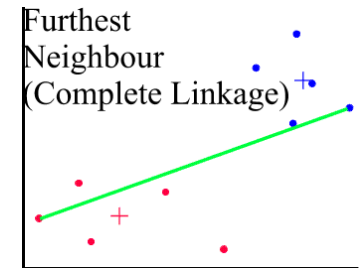
Cluster distances

$$D_{\min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \|x - y\|^2$$



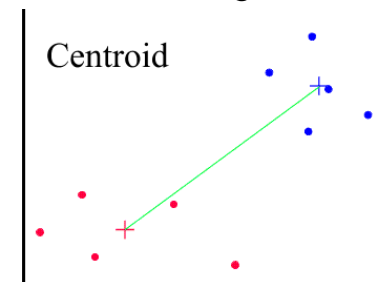
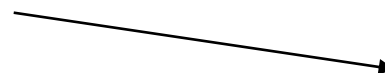
produces minimal spanning tree.

$$D_{\max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \|x - y\|^2$$

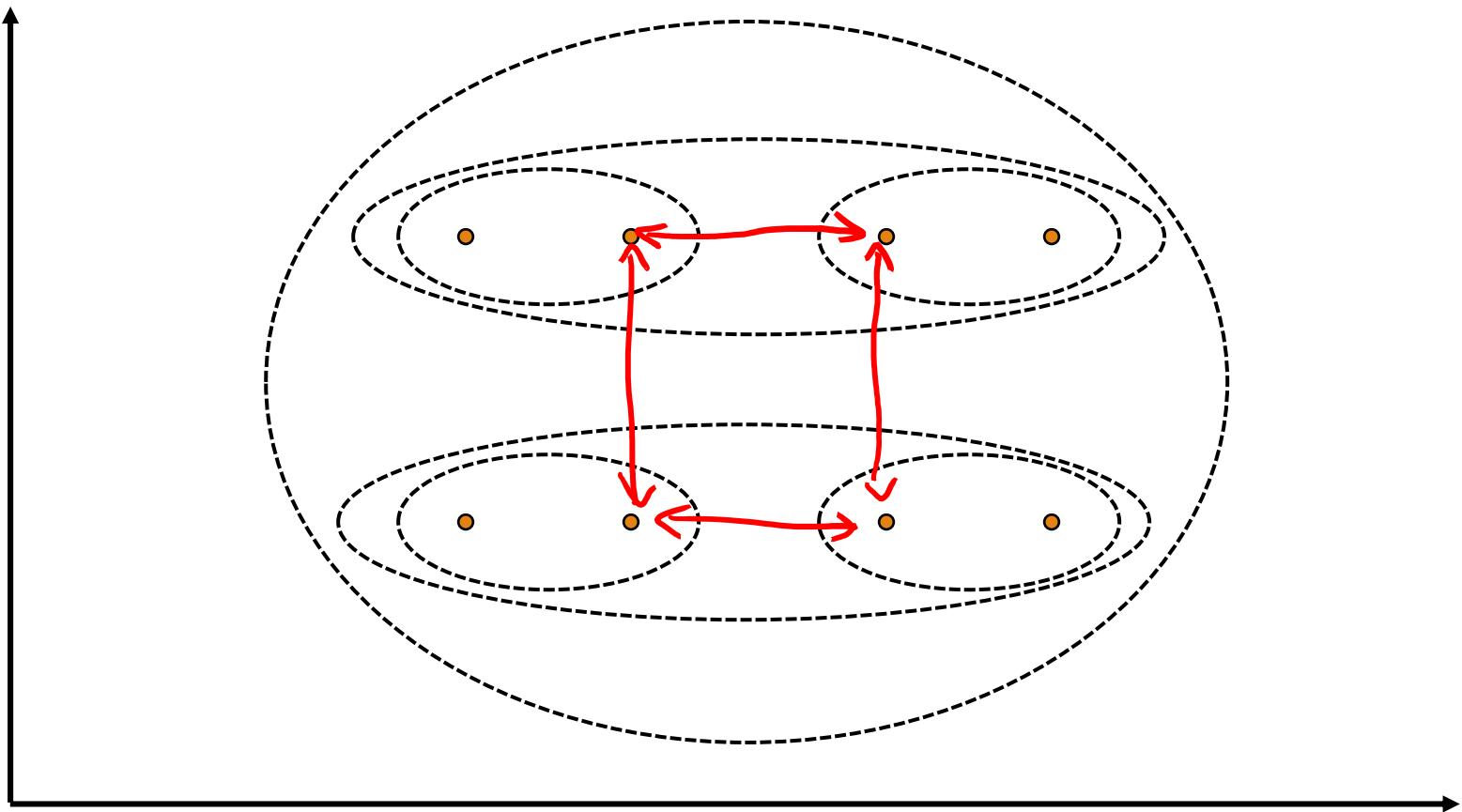


avoids elongated clusters.

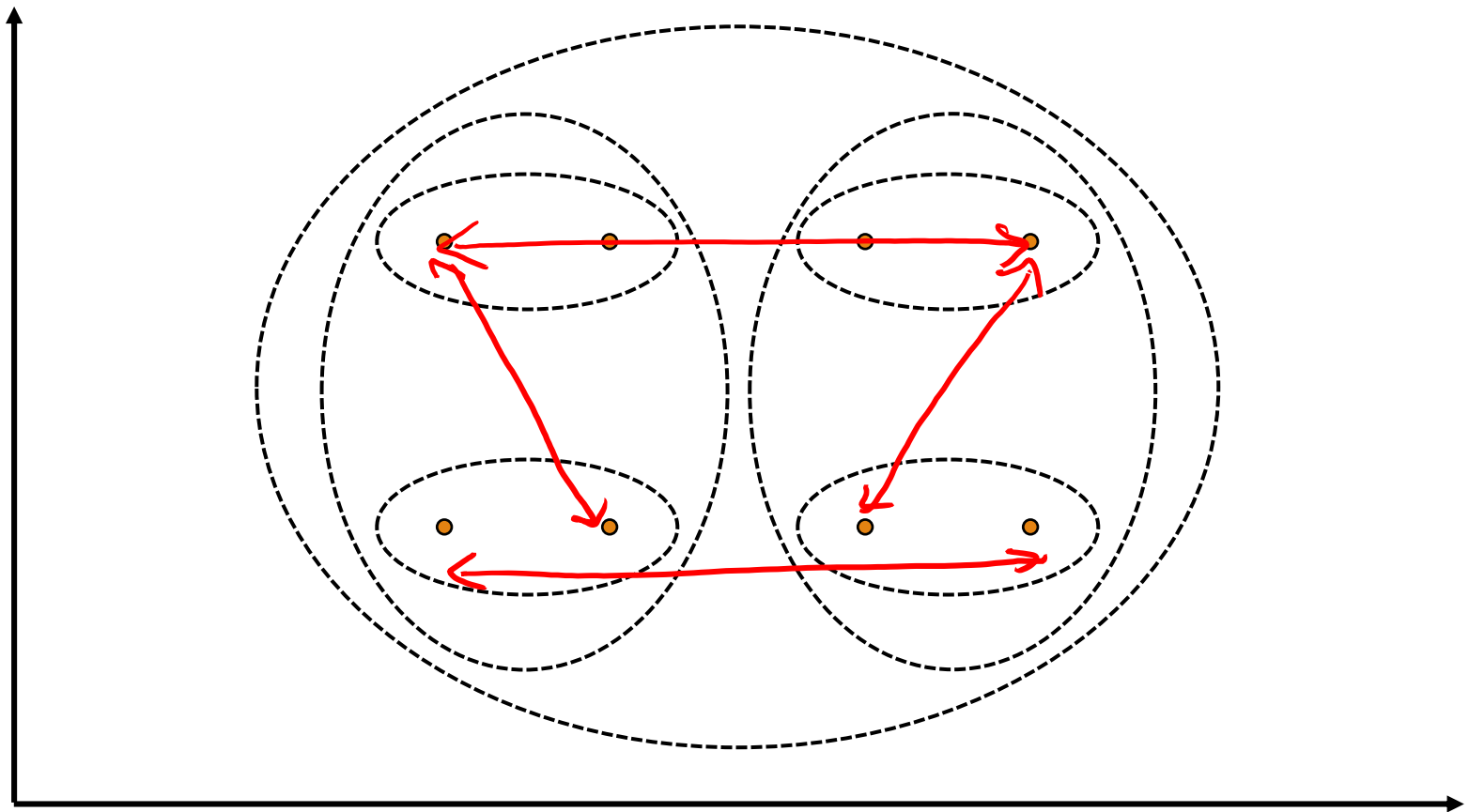
$$D_{\text{means}}(C_i, C_j) = \|\mu_i - \mu_j\|^2$$



Single Link Example



Complete Link Example



Summary

Agglomerative clustering

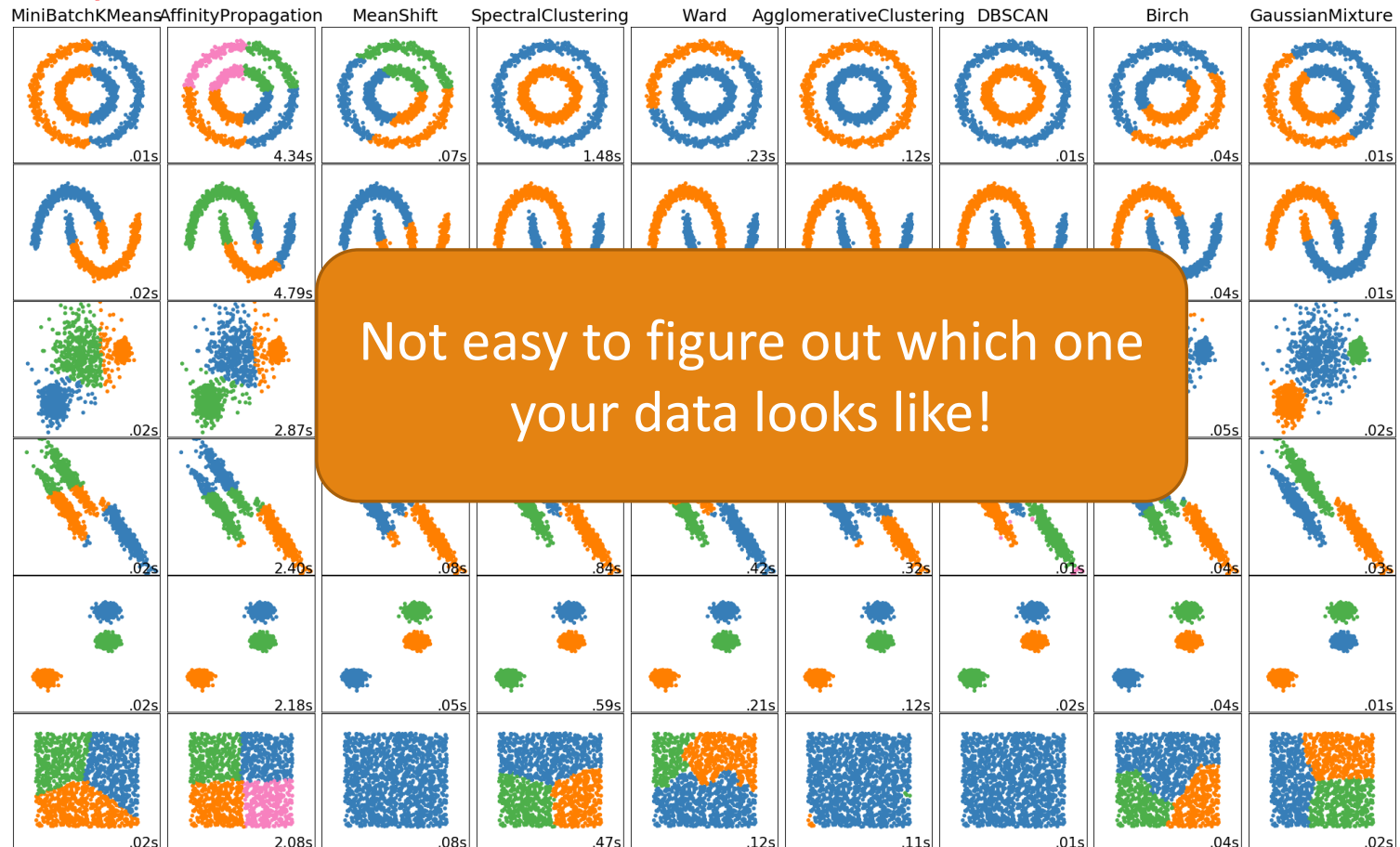
- Choose a cluster distance / dissimilarity scoring method
- Successively merge closest pair of clusters
- “Dendrogram” shows sequence of merges & distances

Agglomerative clusters depend critically on **dissimilarity**

- Choice determines characteristics of “found” clusters

K Means

Lots of Approaches!



In-Class Activity 3
