Clustering

Big data on the web

With 2.41 billion monthly active users as of the second quarter of 2019.

300 million photo uploads per day

Every minute on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded.

Twitter has 330 million monthly active users (as of 2019 Q1)

Half a billion tweets are sent out each day (Mention, 2018).

That equates to 5,787 tweets per second.

Big data on the web

Over 50 billion pages indexed and more than 2 million queries/min

Articles from over 10,000 sources in real time

~4.5 million photos uploaded/day

48 hours of video uploaded/min; more than 1 trillion video views

What to do with such Big Data?

- Extract information to make decisions

- Evidence-based decision:
  - data-driven vs. analysis based on intuition & experience

- Analytics, business intelligence, data mining, machine learning, pattern recognition

Decision Making

- Data Representation
  - Features and similarity

- Learning
  - Classification (labeled data)
  - Clustering (unlabeled data)

Most big data problems have unlabeled objects!

Clustering

Given a collection of (unlabeled) objects, find meaningful groups

What is a cluster?

A group of the same or similar elements
gathered or occurring closely together

Galaxy clusters

Birdhouse clusters

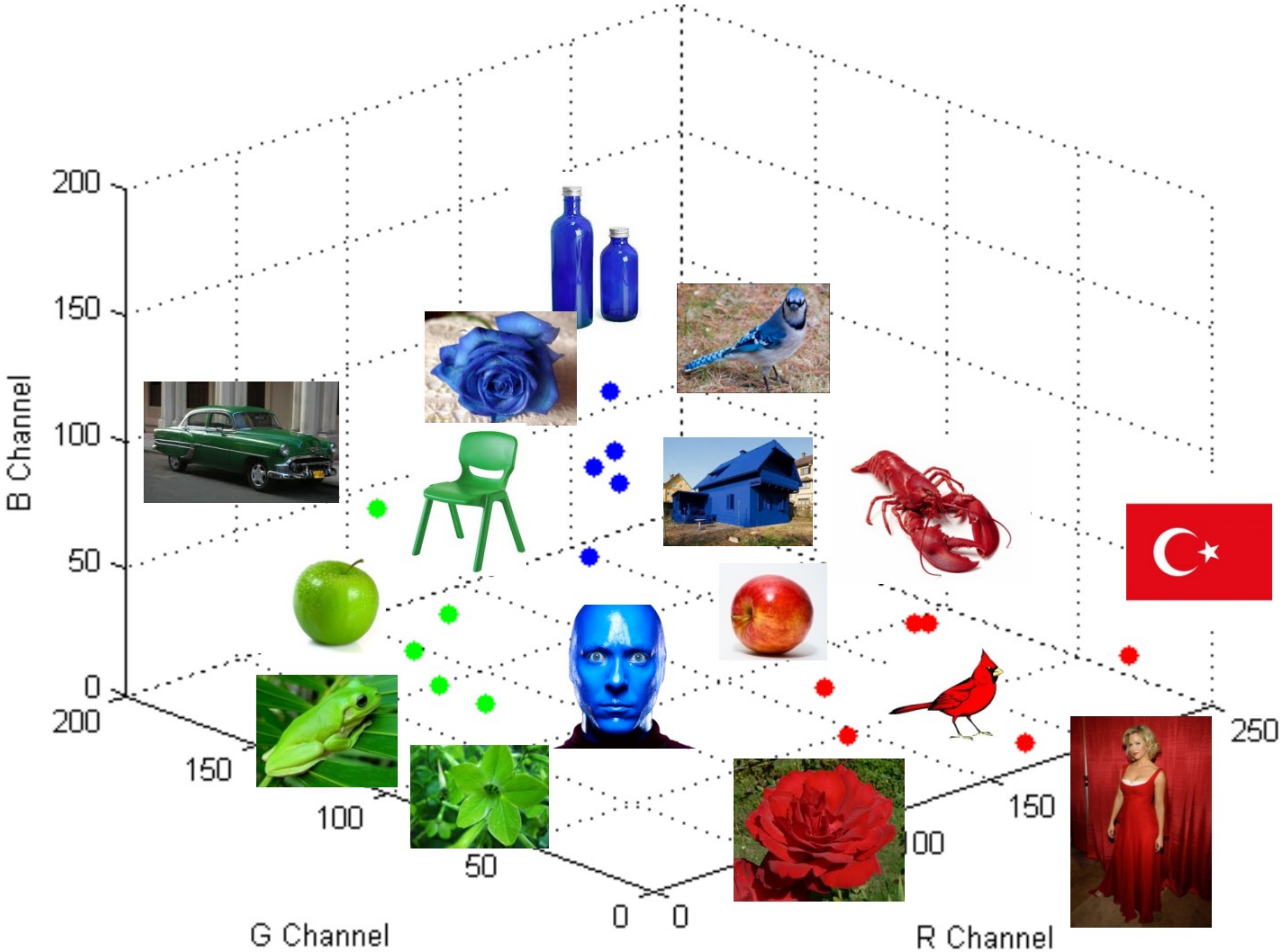
Cluster munition

Cluster computing

Cluster lights

Hongkeng Tulou cluster

Pattern Matrix



n x d pattern matrix

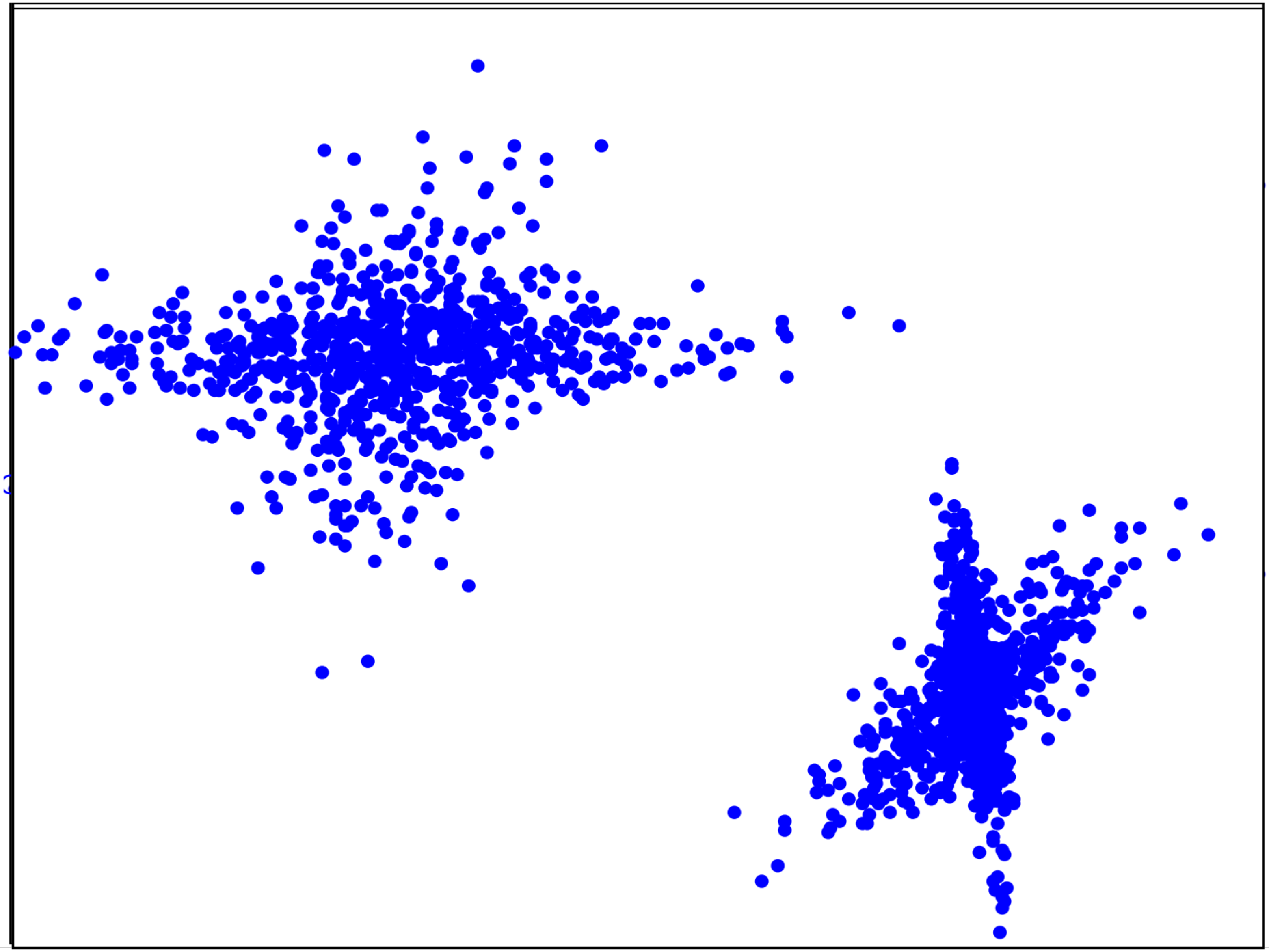
Similarity matrix

Polynomial kernel: $K(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^T\mathbf{y}+1\right)^4$

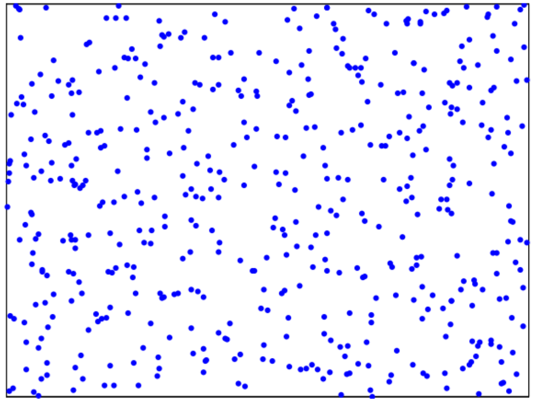
16	15	14	4	6	6	4	3	1
15	16	14	4	5	5	6	4	3
14	14	16	9	9	9	8	7	4
4	4	9	16	15	15	9	10	6
6	5	9	15	16	16	7	8	4
6	5	9	15	16	16	7	8	4
4	6	8	9	7	7	16	16	14
3	4	7	10	8	8	16	16	14
1	3	4	6	4	4	14	14	16

n x n similarity matrix

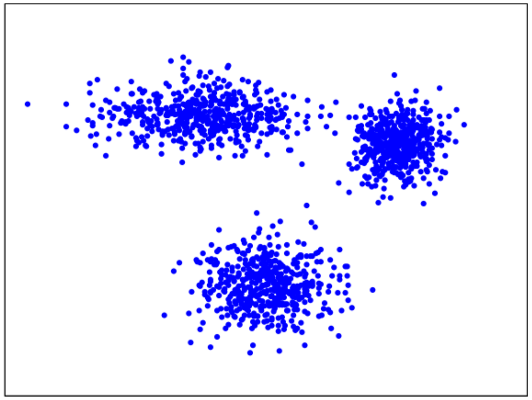
Data clusters in 2D space

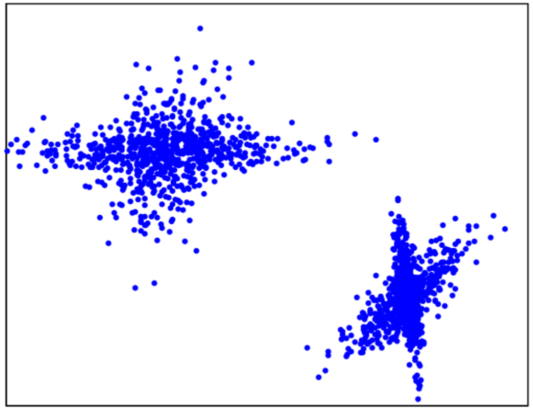


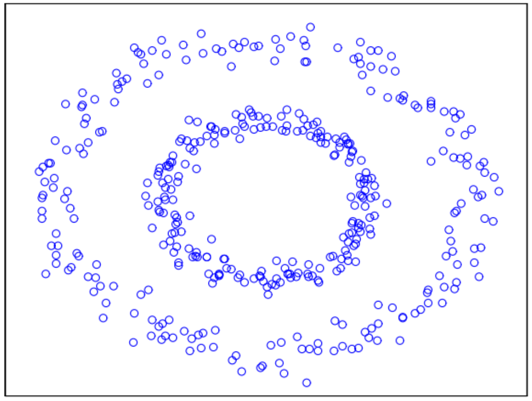
Challenges of clustering

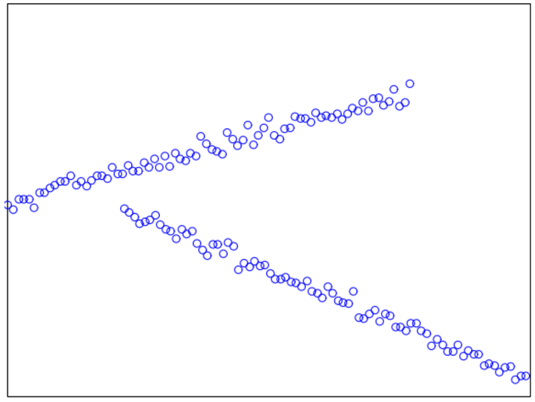


- Measure of similarity
- No. of clusters
- Cluster validity
- Outliers







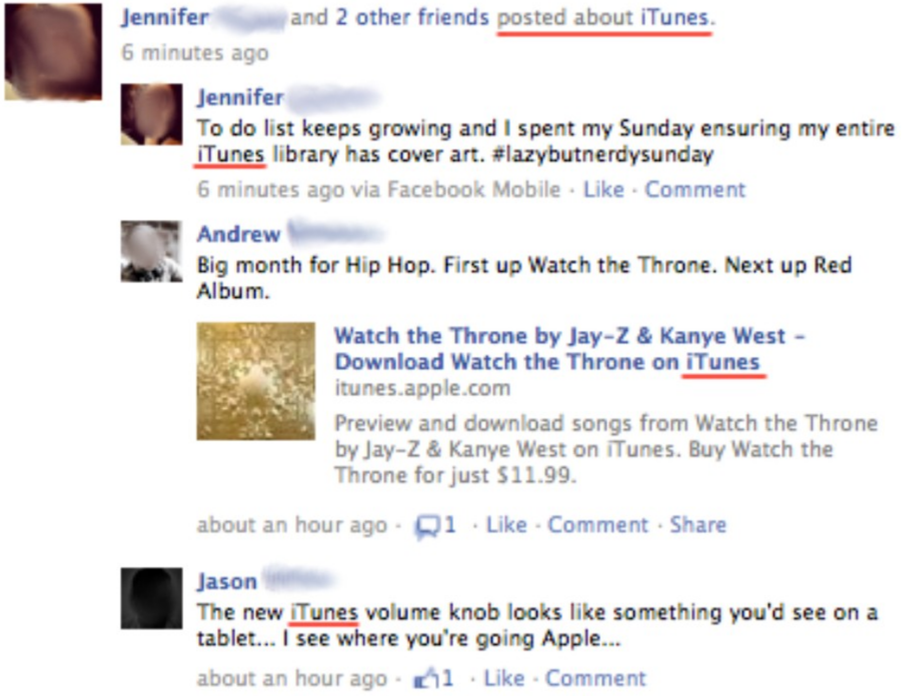


Clustering plays a key role in data analytic

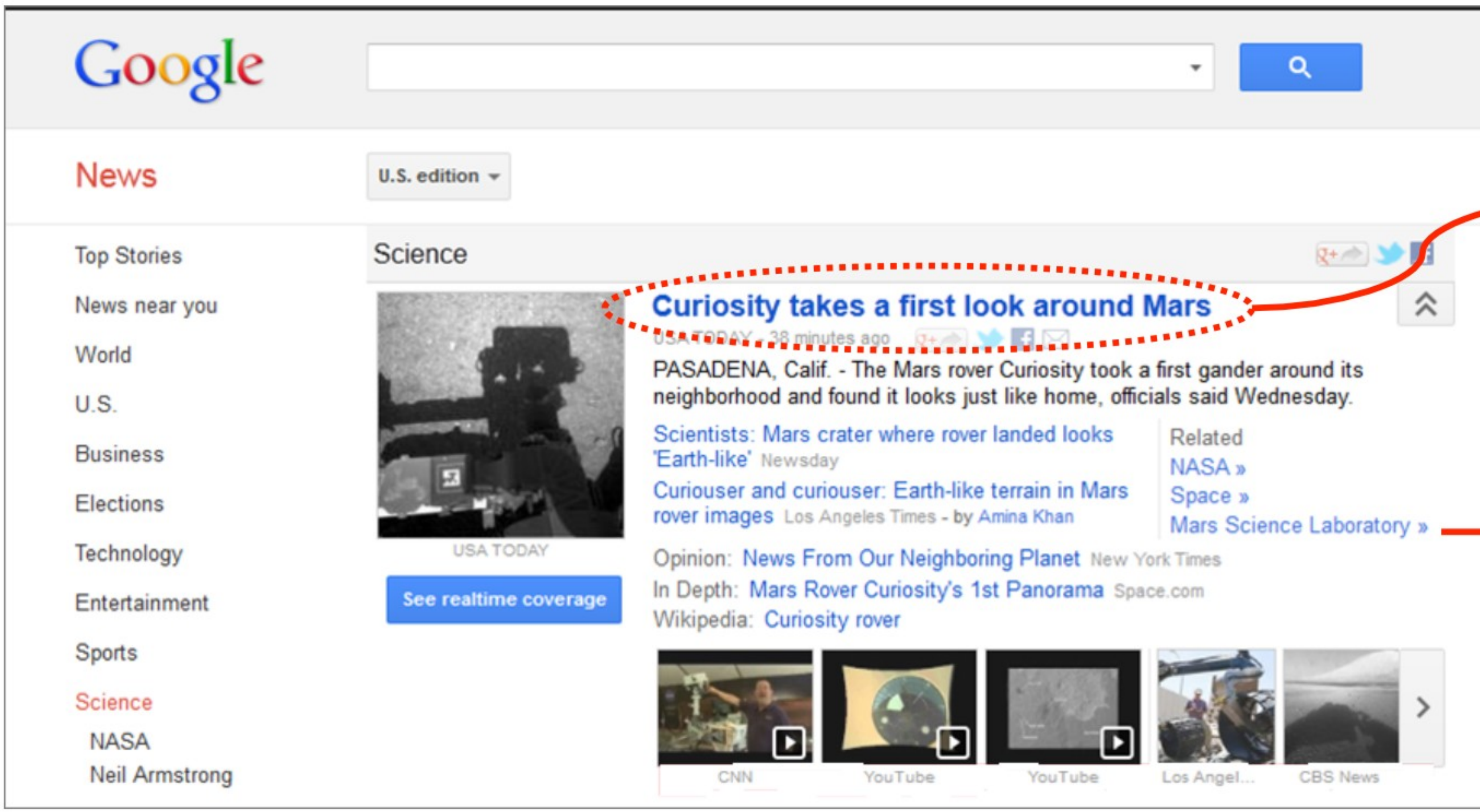
- Not feasible to “label” large collection of objects
- No prior knowledge of the number and nature of groups (clusters) in data
- Clusters may evolve over time
- Clustering provides efficient browsing, search, recommendation and organization of data

Clustering Users on Facebook

- ~300,000 status updates per minute on tens of thousands of topics
- Cluster users based on topic of status messages



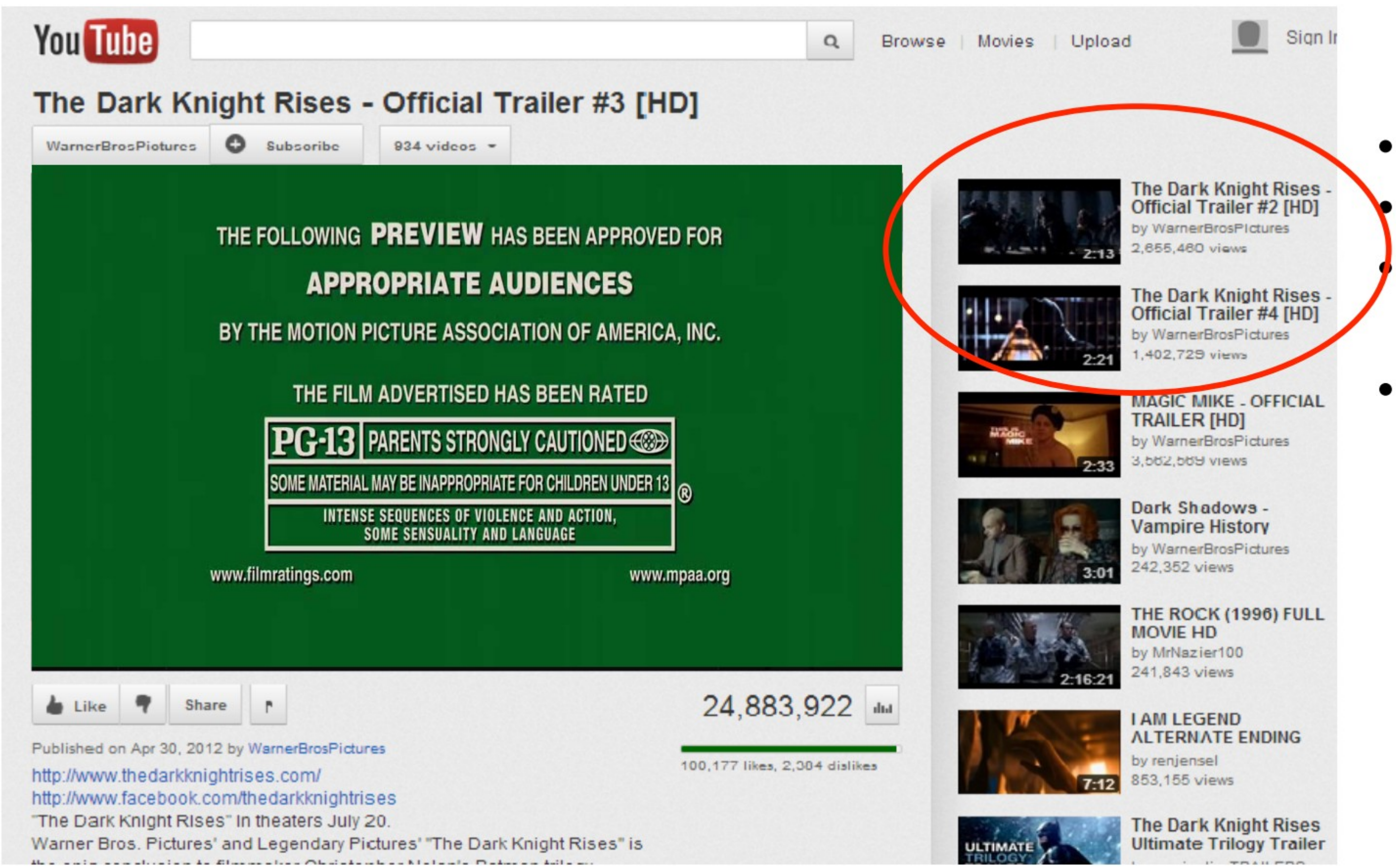
Clustering Articles on Google News



Topic cluster

Article Listings

Clustering Videos on Youtube



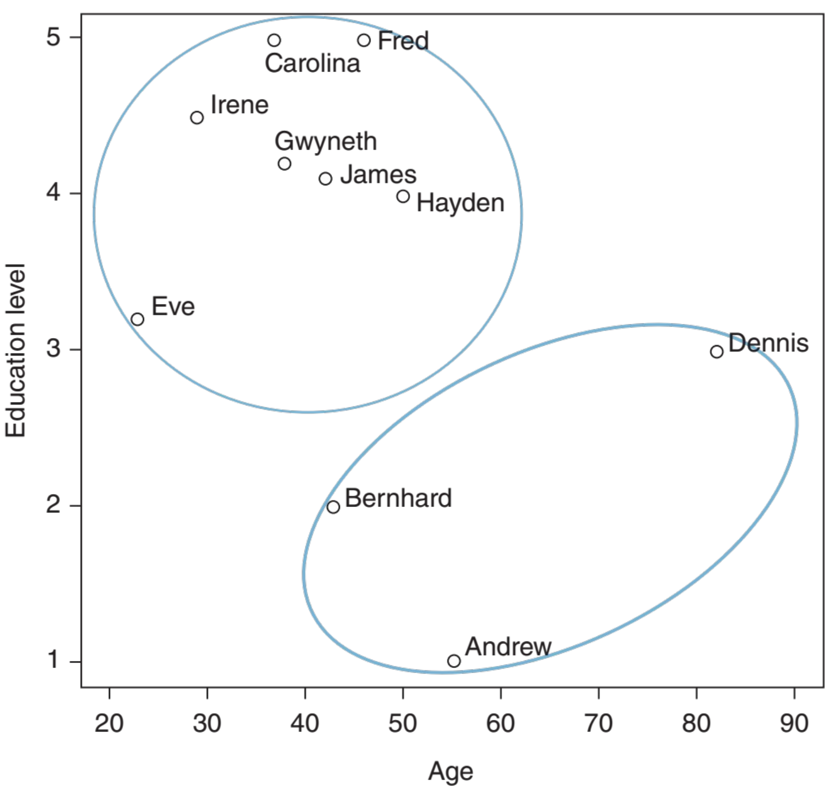
- Keywords
- Popularity
- Viewer engagement
- User browsing history

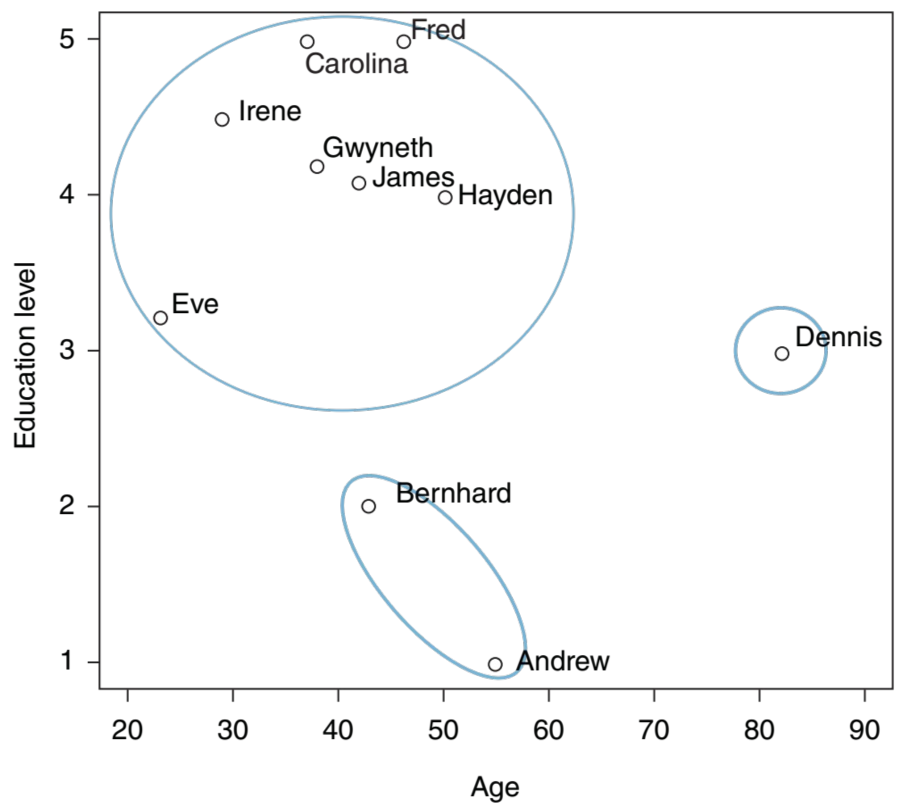
Distance Measures
e.g., Organizing dinners

Table 5.1 Simple social network data set.

Name	Age	Educational level
Andrew (A)	55	1
Bernhard (B)	43	2
Carolina (C)	37	5
Dennis (D)	82	3
Eve (E)	23	3.2
Fred (F)	46	5
Gwyneth (G)	38	4.2
Hayden (H)	50	4
Irene (I)	29	4.5
James (J)	42	4.1

Distance measure tells us which other objects in the same data set are more similar and which are more dissimilar.





Differences between Values of Common Attribute Types

- For quantitative attributes:

$$d(a.b) = |a - b|$$

- For qualitative (categorical) attributes:
  - Ordinal: $$d(a,b) = (|pos_a - pos_b|)/(n-1)$$
  - Nominal: $$d(a,b) = \begin{cases} 1\ , & \text{if } a \neq b \\ 0\ , & \text{if } a = b \end{cases}$$

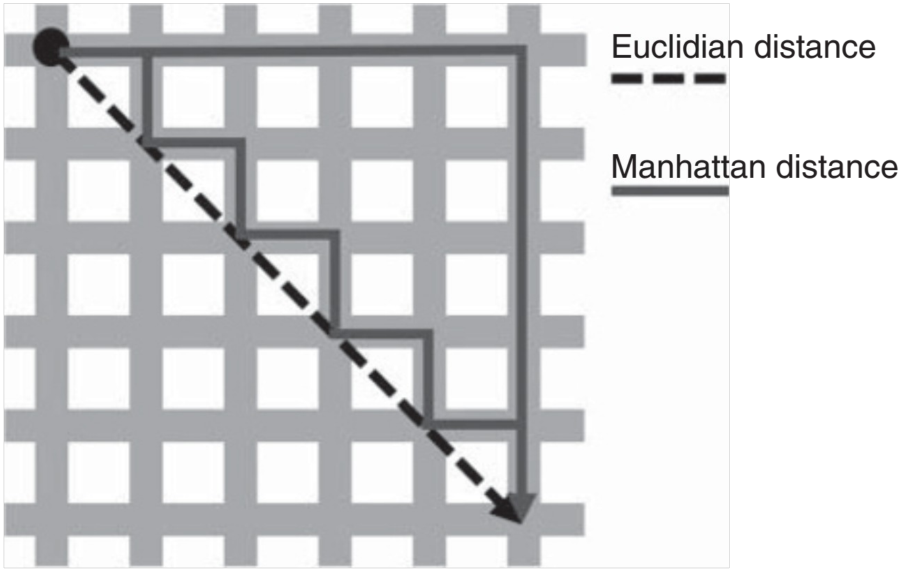
Usually computing the distance between two objects consists of aggregating the distances, usually differences, between their corresponding attributes.

Distance Measures for Objects with Quantitative Attributes

- An object represented by a vector of m quantitative attributes can be mapped to an m-dimensional space.
- Several distance measures are particular cases of the Minkowski distance.

$$d(p,q) = \sqrt[r]{\sum_{k=1}^{m} |p_k - q_k|^r}$$

  - For the Manhattan distance, r = 1
  - For the Euclidean distance, r = 2



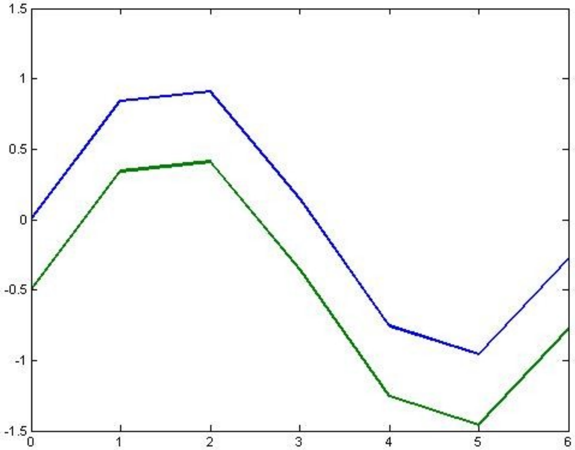
Distance Measures for Objects with Quantitative Attributes

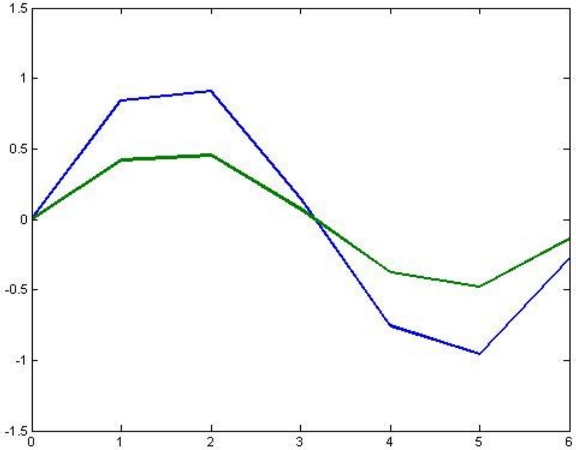
- Correlation distance

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

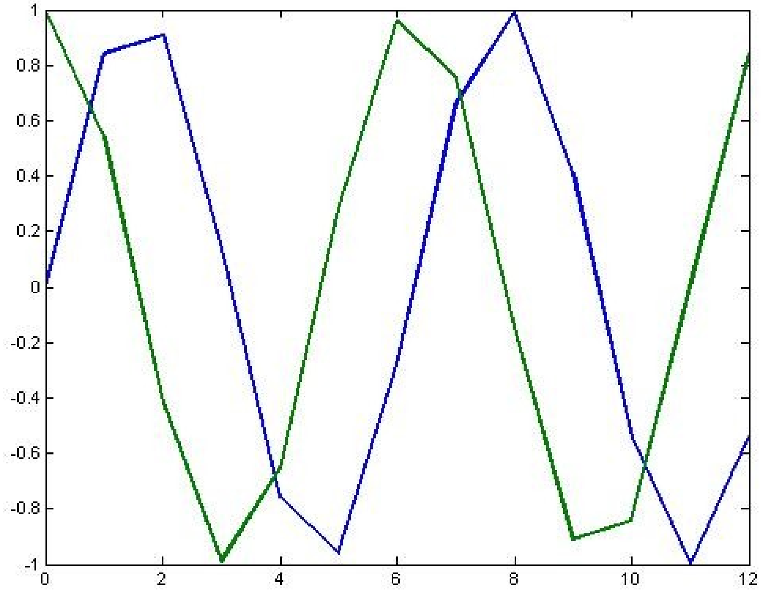
$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

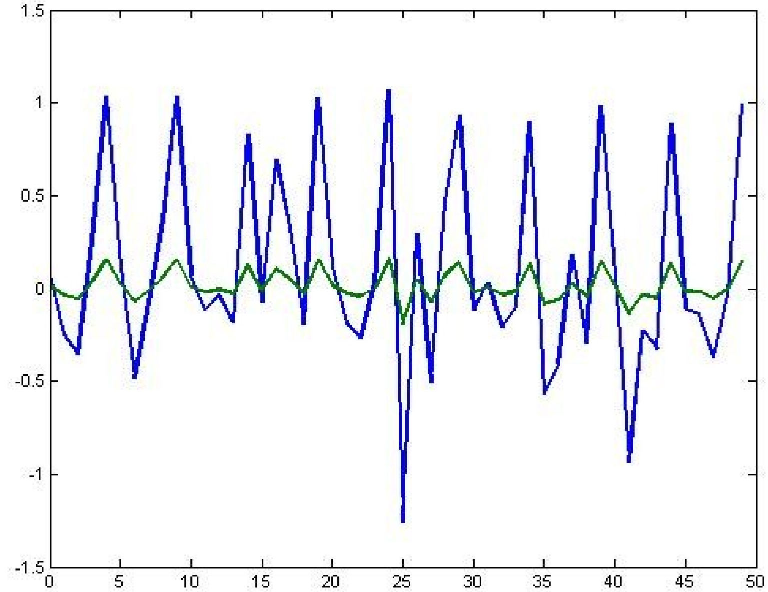
$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$





How about these?





Distance Measures for Non-conventional Attributes

Non-conventional Attributes

- biological sequences
- time series
- images
- sound
- Video

- All these non-conventional attribute types can be converted into quantitative or qualitative types

Distance Measures for short sequences (text)

- The Hamming distance can be used for sequences of values and these values are usually characters or binary values.
- The Hamming distance is the number of positions at which the corresponding characters or symbols in the two strings are different.
  - distance between the strings “James” and “Jimmy” is 3
  - and between “Tom” and “Tim” is 1

Distance Measures for short sequences (text)

- For short sequences that can have different sizes we use edit distance.
- The edit distance measures the minimum number of operations necessary to transform one sequence into another.
- The possible operations are: insertion (of a character), removal (of a character) and substitution (of a character by another).
  - The edit distance between the strings “Johnny” and “Jonston” is 5, since it is necessary to substitute the characters h, n, n, y with n, s, t, o (four operations), and to add a character n to the end (a fifth operation).
- A similar idea is used in bioinformatics to compare DNA, RNA and amino acid sequences.

Distance Measures for long sequences (texts)

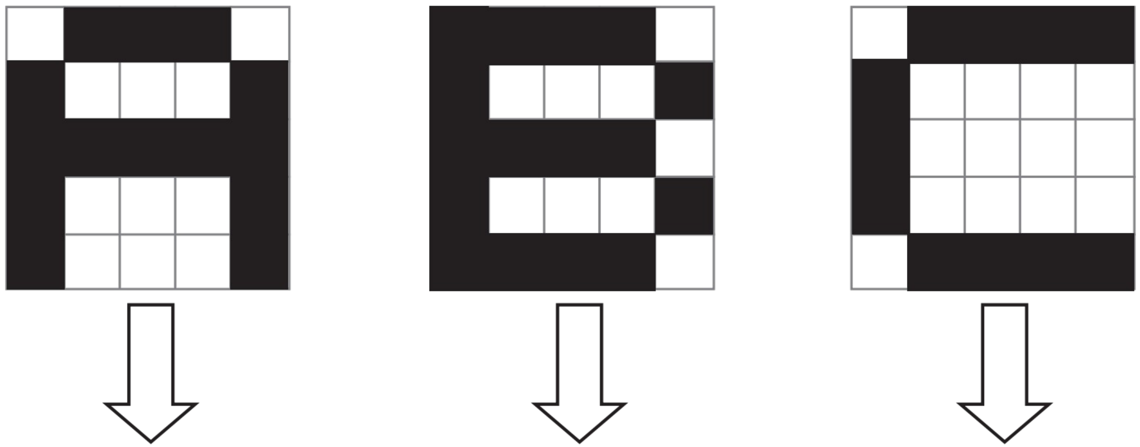
For long texts we can use “bag of words”:

- For example, for the two texts:
  - A = “I will go to the party. But first, I will have to work.”
  - B = “They have to go to the work by bus.”

	I	will	go	to	the	party	but	first	have	work	they	by	bus
A	2	2	1	2	1	1	1	1	1	1	0	0	0
B	0	0	1	2	1	0	0	0	1	1	1	1	1

- Each text is converted into a quantitative vector, where each position is associated with one of the words

Distance Measures for Images



0	1	1	1	0
1	0	0	0	1
1	1	1	1	1
1	0	0	0	1
1	0	0	0	1

1	1	1	1	0
1	0	0	0	1
1	1	1	1	0
1	0	0	0	1
1	1	1	1	0

0	1	1	1	1
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	1	1	1	1

	1^{st} row					2^{nd} row					3^{rd} row					4^{th} row					5^{th} row				
A	0	1	1	1	0	1	0	0	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1
B	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0
C	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1

Hundreds of clustering algorithms are available; many are “admissible”, but no algorithm is “optimal”

- K-means
- Gaussian mixture models
- Kernel K-means
- Fuzzy k-means
- DBSCAN
- Nearest neighbor
- Hierarchical clustering

Clustering Validation

- The automatic validation measures are divided into three categories:
  - **External indices:** The external criteria uses external information, such as **class label**, if available, to define the quality of the clusters in a given partition. Two of the most common external measures are the **correct-RAND** and **Jaccard**.
  - **Internal indices:** The internal criteria looks for **compactness** inside each cluster and/or separation between different clusters. Two of the most common internal measures are the **silhouette index**, which measures both compactness and separation, and the **within-groups sum of squares**, which only measures compactness.

silhouette index

It measures:

- How close to each other the objects inside a cluster are.
- The separation of different clusters

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i) & , \quad \text{if } a(x_i) < b(x_i) \\ 0 & , \quad \text{if } a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1 & , \quad \text{if } a(x_i) > b(x_i) \end{cases}$$

- $a(x_i)$ is the average distance between x_i and all other objects in its cluster
- $b(x_i)$ is the minimum average distance between x_i and all other objects from other clusters.
- The average of all s(xi) gives the partition silhouette measure value

Within-groups sum of squares

The within groups sum of squares is given by:

$$s = \sum_{i=1}^{K} \sum_{j=1}^{J_i} sed(p_j, C_i)$$

where K is the number of clusters and J_i is the number of instances of cluster i, and C_i is the centroid of cluster i.

Jaccard external measure

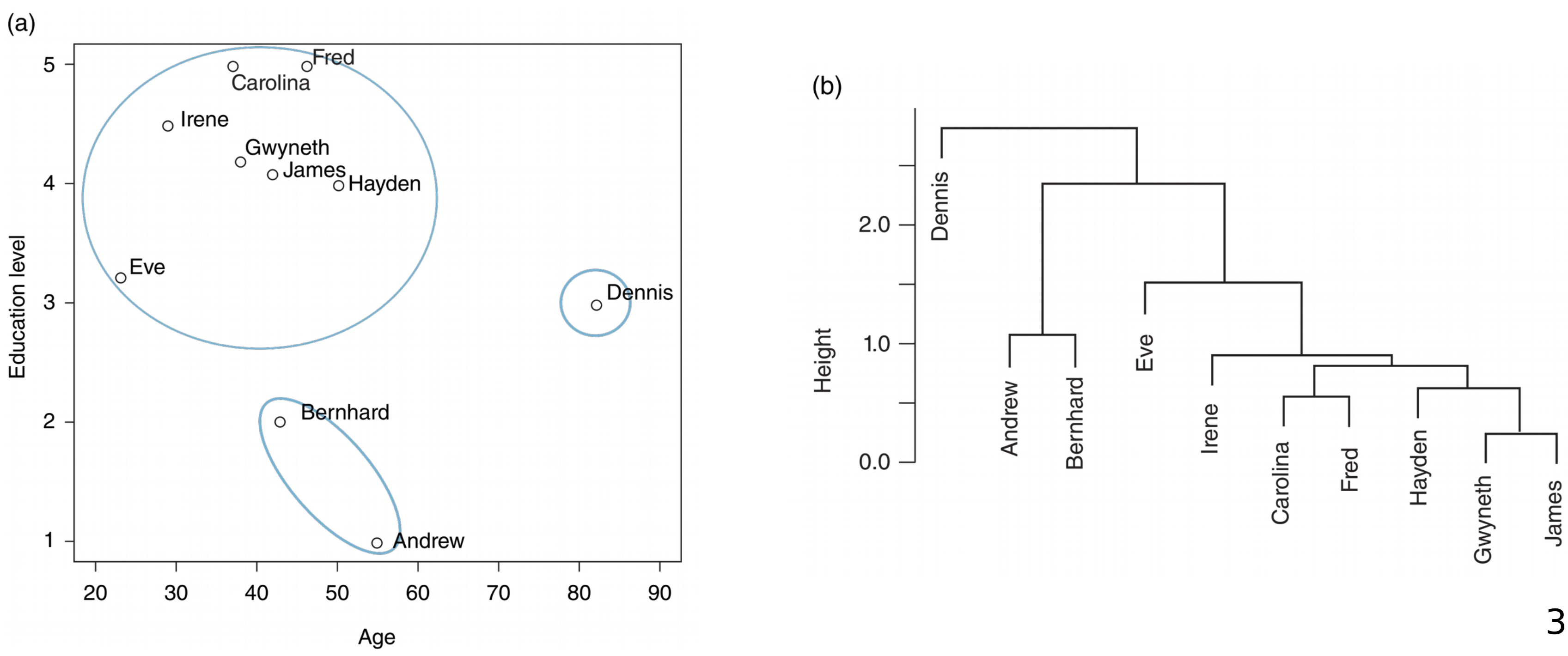
- It needs data labels.
- It evaluates how uniform the distribution of the objects in each cluster is with respect to the class label.

$$J = (M_{11})/(M_{01} + M_{10} + M_{11})$$

- M_{01} is the number of objects in other clusters but with the same label
- M_{10} is the number of objects in the same cluster, but with different labels
- M_{00} is the number of objects in other clusters with different labels
- M_{11} is the number of objects in the same cluster with the same label.

Categories of Clustering algorithms

- a) Most techniques define partitions in one step (partitional clustering),
- b) While others progressively define partitions, either increasing or decreasing the number of clusters (hierarchical clustering).



Categories of Clustering algorithms

- Another criteria is the approach used to define what a cluster is:
  - **Separation-based:** each object in the cluster is closer to every other object in the cluster than to any object outside the cluster
  - **Prototype-based:** each object in the cluster is closer to a prototype representing the cluster than to a prototype representing any other cluster
  - **Graph-based:** represents the data set by a graph structure associating each node with an object and connecting objects that belong to the same cluster with an edge
  - **Density-based:** a cluster is a region where the objects have a high number of close neighbors (i.e. a dense region), surrounded by a region of low density
  - **Shared-property:** a cluster is a group of objects that share a property

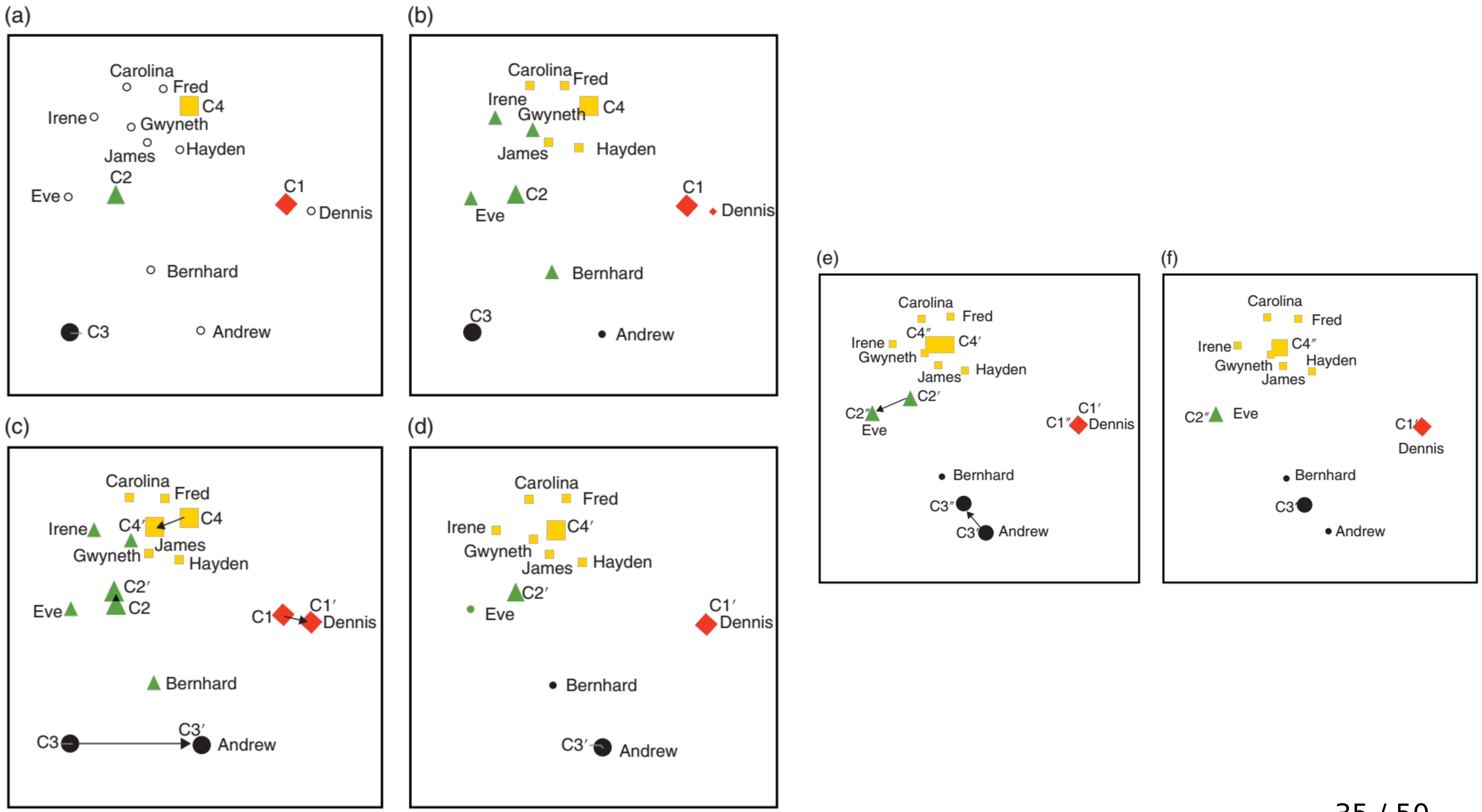
- K-means:
  - The most popular clustering algorithm and a representative of partitional and prototype-based clustering methods
- DBSCAN:
  - Another partitional clustering method, but in this case density-based
- Agglomerative hierarchical clustering:
  - A representative of hierarchical and graph-based clustering methods.

K-means

Algorithm K-means

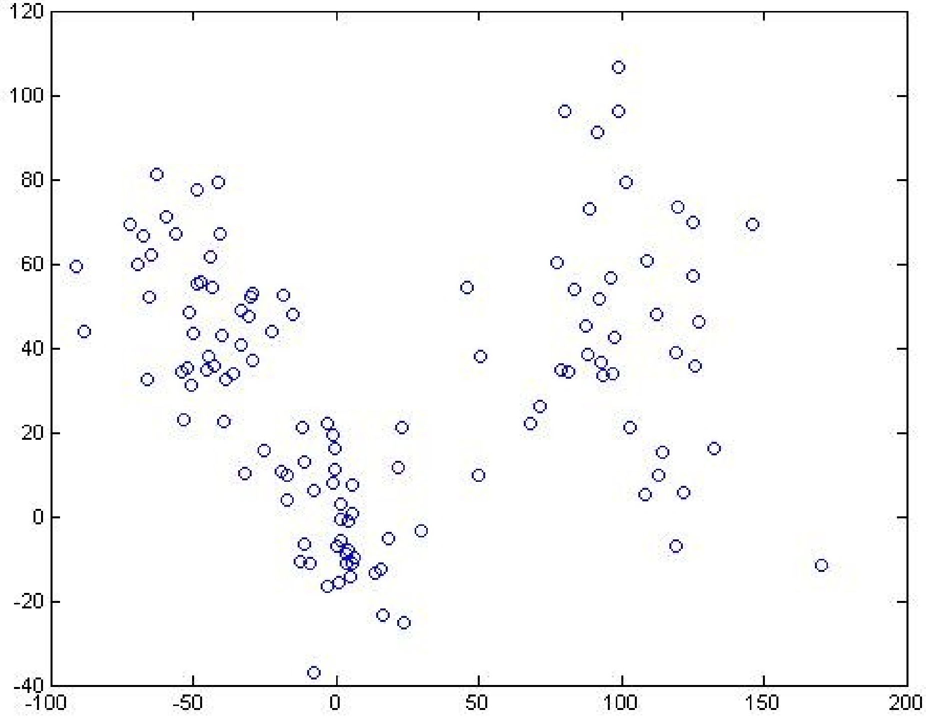
INPUT D the data set
INPUT d the distance measure
INPUT K the number of clusters
Define the initial K centroids (they are usually randomly defined, but can be defined explicitly in some software packages)
repeat
    Associate each instance in D with the closest centroid according to the chosen distance measure d
    Recalculate each centroid using all instances from D associated with it.
until No instances from D change of associated centroid.

Example



(a)
Carolina
Fred
C4
Irene
Gwyneth
James
Hayden
C2
Eve
C1
Dennis
Bernhard
C3
Andrew
(b)
Carolina
Fred
Irene
Gwyneth
C4
James
Hayden
C2
Eve
C1
Dennis
Bernhard
C3
Andrew
(c)
Carolina
Fred
C4
Irene
C4′
James
Gwyneth
Hayden
C2′
Eve
C2
C1
C1′
Dennis
Bernhard
C3
C3′
Andrew
(d)
Carolina
Fred
Irene
C4′
Gwyneth
James
Hayden
C2′
Eve
C1′
Dennis
Bernhard
C3′
Andrew
(e)
Carolina
Fred
C4″
Irene
C4′
Gwyneth
James
Hayden
C2′
C2″
Eve
C1′
C1″
Dennis
Bernhard
C3″
C3′
Andrew
(f)
Carolina
Fred
Irene
C4″
Gwyneth
Hayden
James
C2″
Eve
C1′
Dennis
Bernhard
C3′
Andrew

Example



How many clusters do you think there are in this data? How might it have been generated?

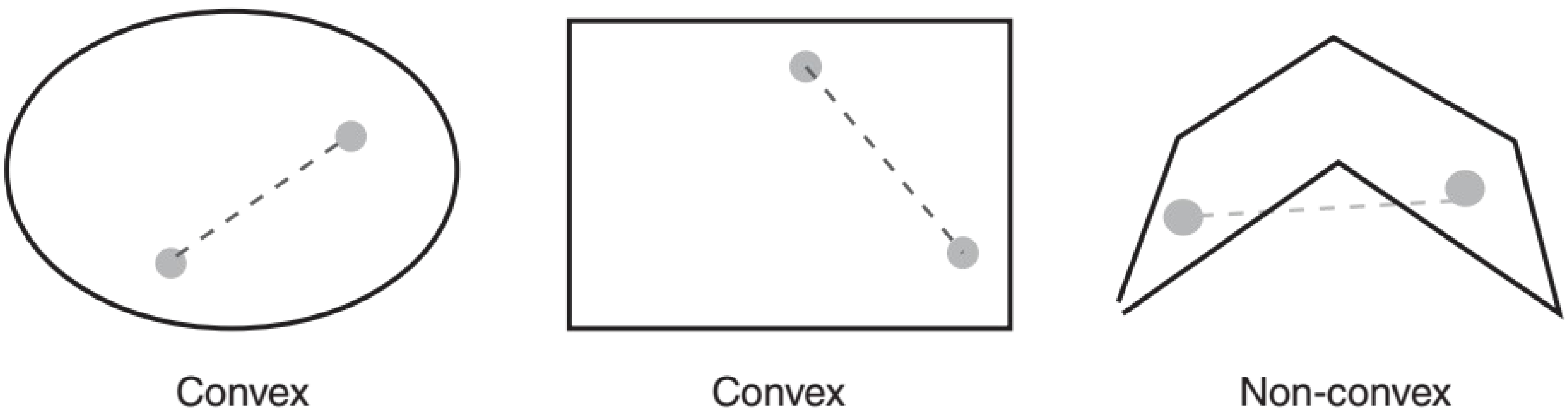
Example

$$k = 2$$

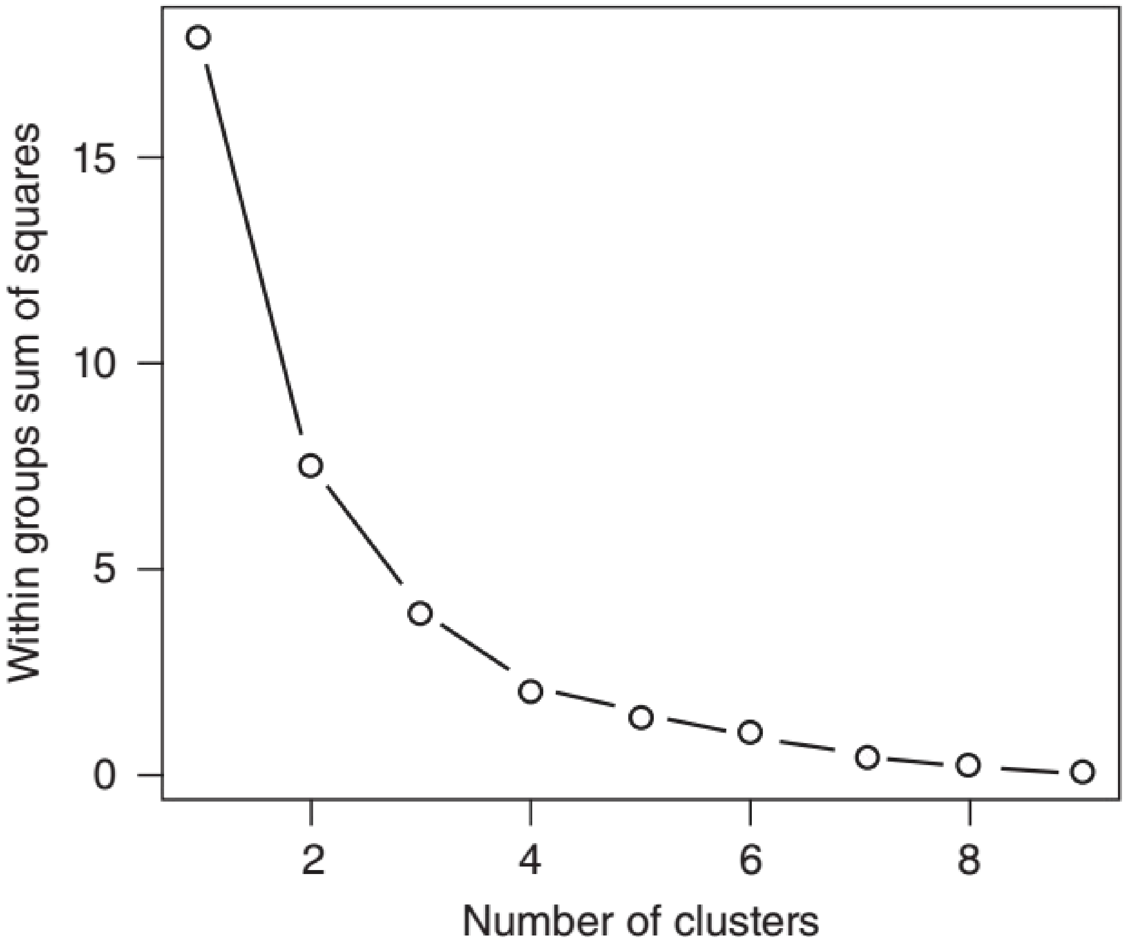
Weakpoints of K-means

- Random initialization means that you may get different clusters each time
- Data points are assigned to only one cluster (hard assignment)
- Implicit assumptions about the “shapes” of clusters
- You have to pick the number of clusters…

K-means clusters are convex



Choosing the right K



Jogota validation

- Jagota validation suggests a measure that emphasizes cluster tightness or homogeneity:

$$Q = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \mu_i)$$

- $|C_i|$ is the number of data points in cluster *I*

- Q will be small if (on average) the data points in each cluster are close

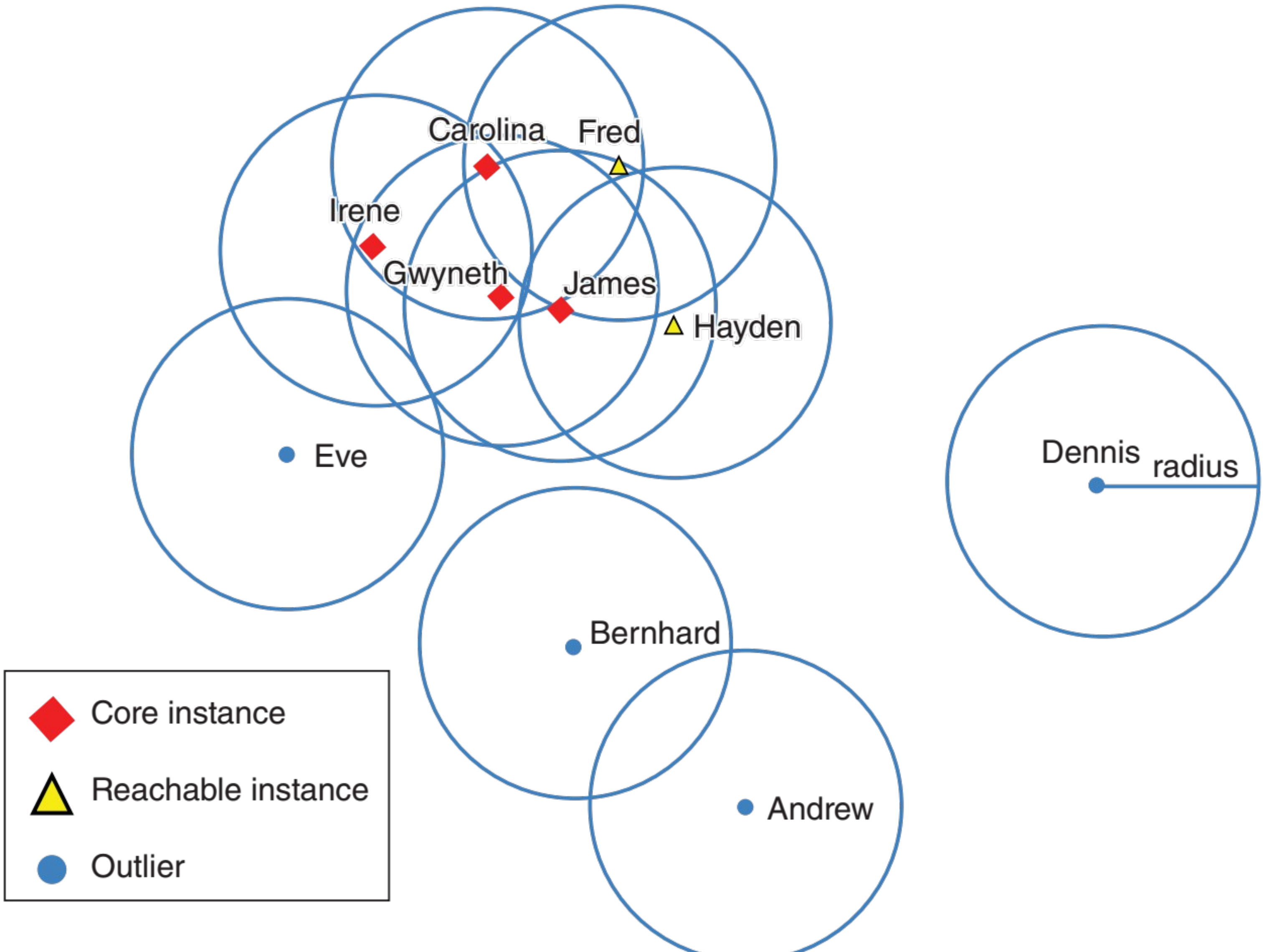
DBSCAN

density-based spatial clustering of applications with noise

- In contrast to k-means, DBSCAN automatically defines the number of clusters.
- In DBSCAN, objects forming a dense region belongs to the same cluster.
- Objects not belonging to dense regions are considered to be noise.

DBSCAN

density-based spatial clustering of applications with noise



DBSCAN

density-based spatial clustering of applications with noise

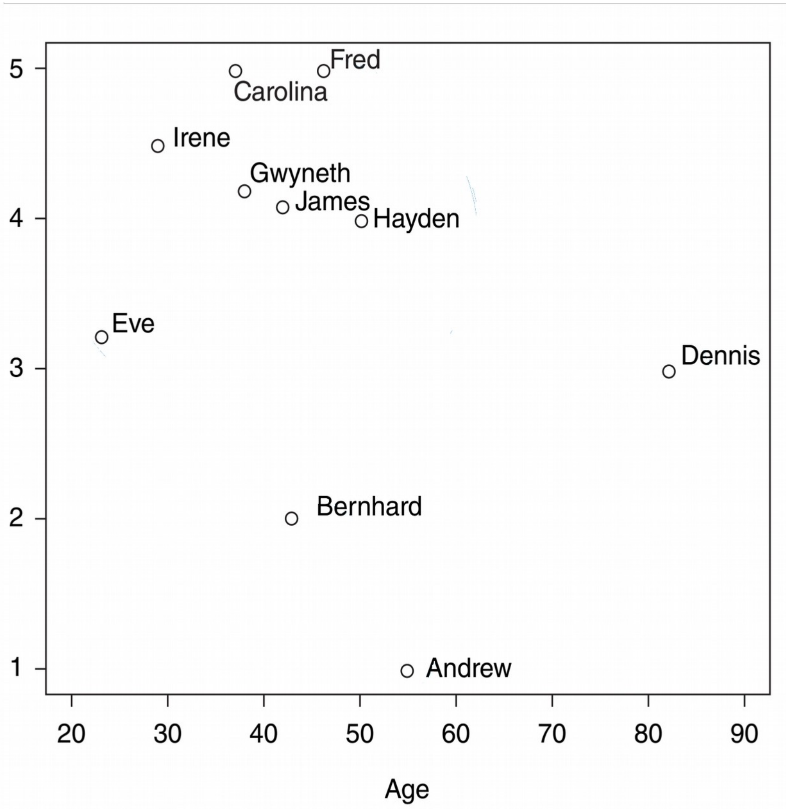
- A core instance p is an instance that directly reaches a minimum number of other instances.
- To be considered “directly reachable” an instance q must be at a lower distance from p than a predefined threshold.
- If p is a core instance, then it forms a cluster together with all instances that are reachable from it, directly or indirectly.
- Each cluster contains at least one core instance.
- DBSCAN also has some randomization on deciding to which core instance a given instance will be attached when there is more than one core instance that can reach it directly

DBSCAN
Pros & Cons

Advantages	Disadvantages
• It can detect clusters of an arbitrary shape • Robust to outliers	• Each time we run DBSCAN the results can be a bit different due to some randomness, but the results typically do not vary much between different runs • Computationally more complex than k-means • Difficulty in setting the hyper-parameter values

Agglomerative Hierarchical Clustering

- Hierarchical algorithms construct clusters progressively and based on pairwise distances.

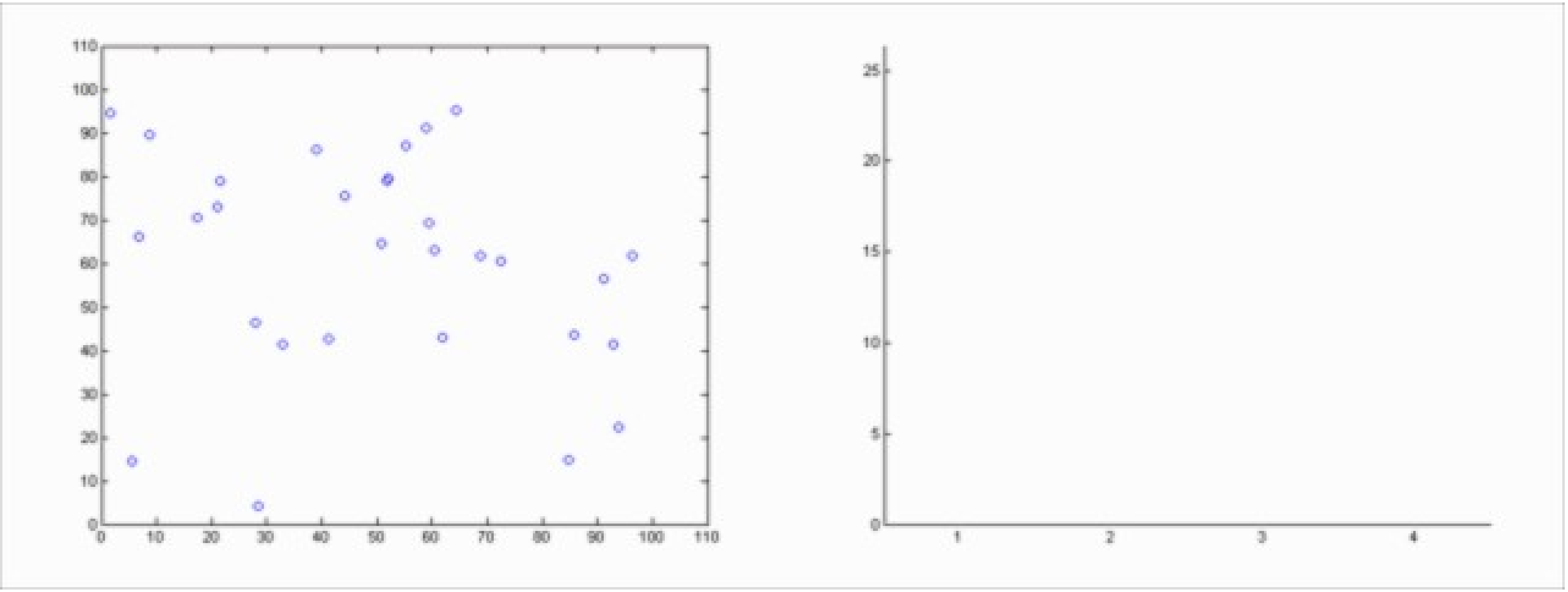


	A	B	C	D	E	F	G	H	I	J
A	0									
B	1.07	0								
C	3.26	2.33	0							
D	2.26	2.53	3.17	0						
E	2.6	1.54	1.63	3.65	0					
F	3.11	2.31	0.56	2.7	1.98	0				
G	2.67	1.71	0.62	2.87	1.2	0.79	0			
H	2.32	1.59	1.11	2.12	1.78	0.8	0.76	0		
I	3.13	2.1	0.63	3.47	1.06	1.12	0.6	1.35	0	
J	2.51	1.61	0.76	2.61	1.36	0.73	0.26	0.5	0.86	0

Agglomerative Hierarchical Clustering

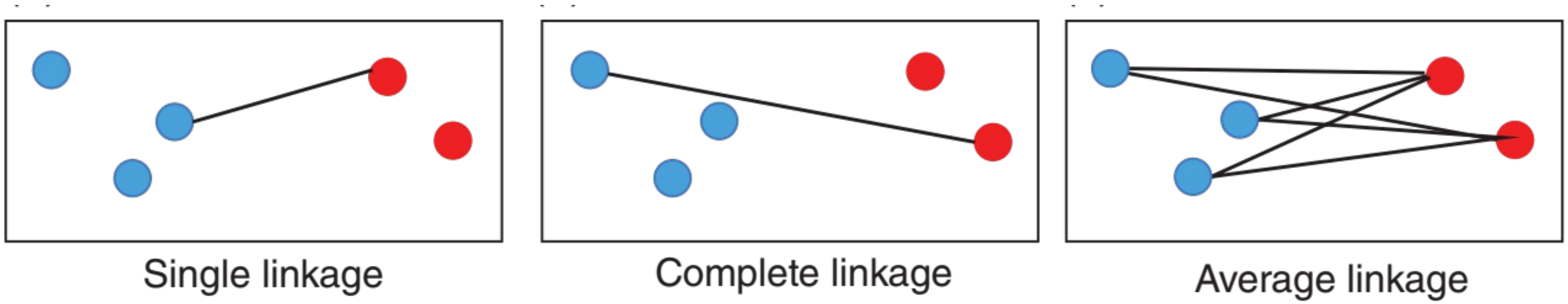
- We start with every data point in a separate cluster.
- We keep merging the most similar pairs of data points/clusters until we have one big cluster left.
- This is called a bottom-up or agglomerative method.

Agglomerative Hierarchical Clustering Demo

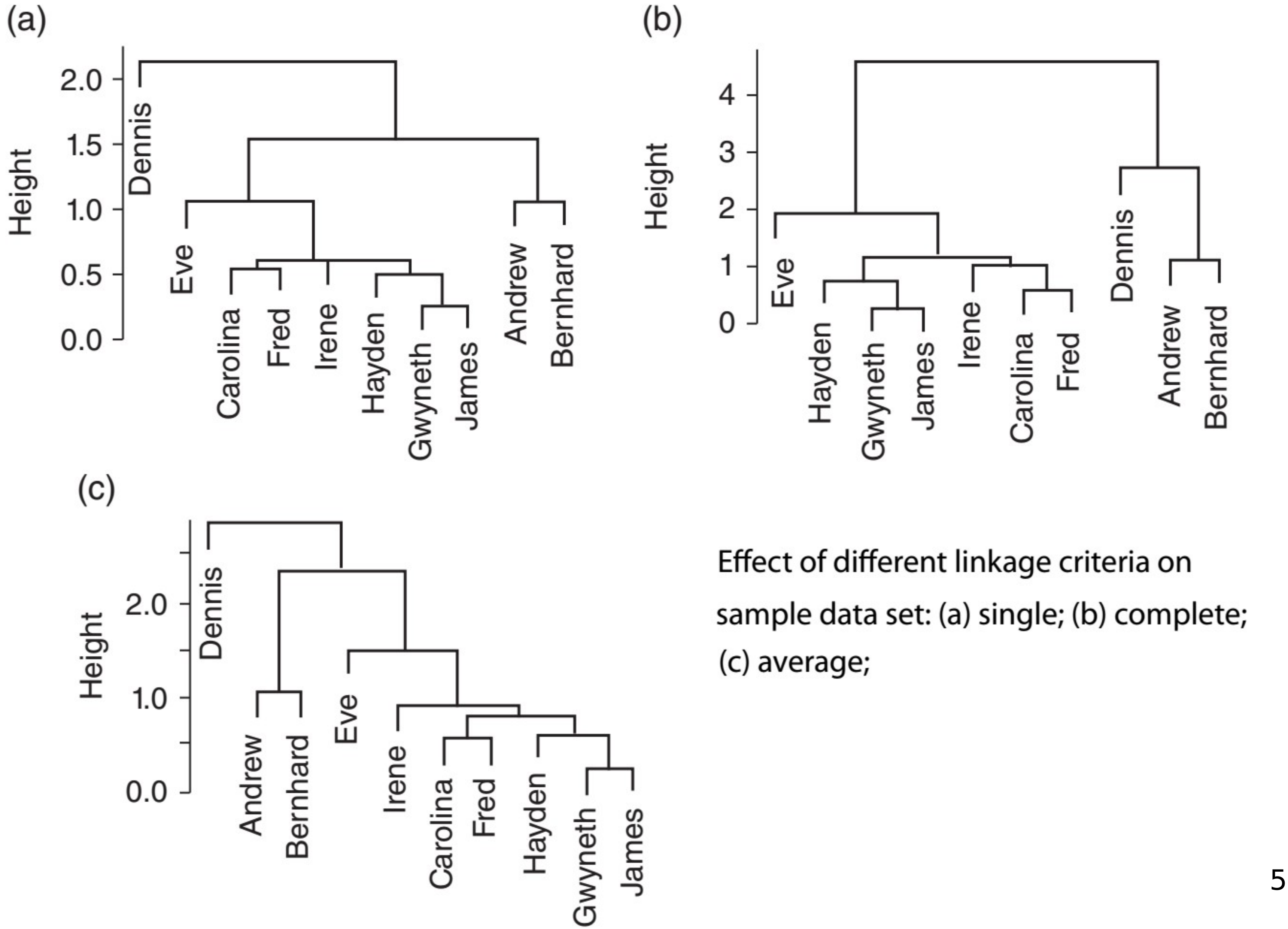


Cluster Linkage

- **The single linkage:** This measures the distance between the closest instances, one from each set. It favors the appearance of a dominant cluster.
- **The complete linkage:** This measures the distance between the most distant instances, one from each set. It favors similar clusters
- **The average linkage:** This measures the average distance of every pair of instances, each instance of a pair from each set. It is in between the two previous approaches.



Effect of linkage criteria



Effect of different linkage criteria on sample data set: (a) single; (b) complete; (c) average;