Data Science

Lecture 10: Clustering

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut’s University of Technology Thonburi

Topics

- Similarity
- K-Means Clustering

https://colab.research.google.com/drive/1Ts03orFSjtPhbCABcsAbBNWjV3Ncf2Zo

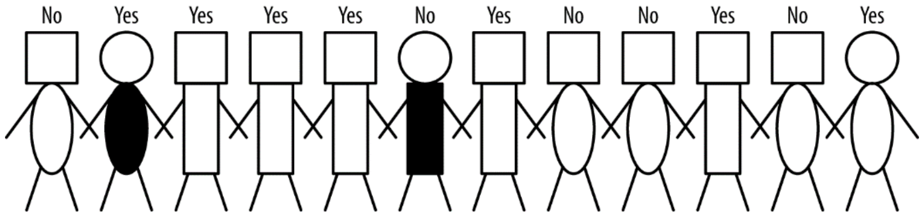
Similarity

Part 1

Similarity

- Similarity underlies many data science methods and solutions

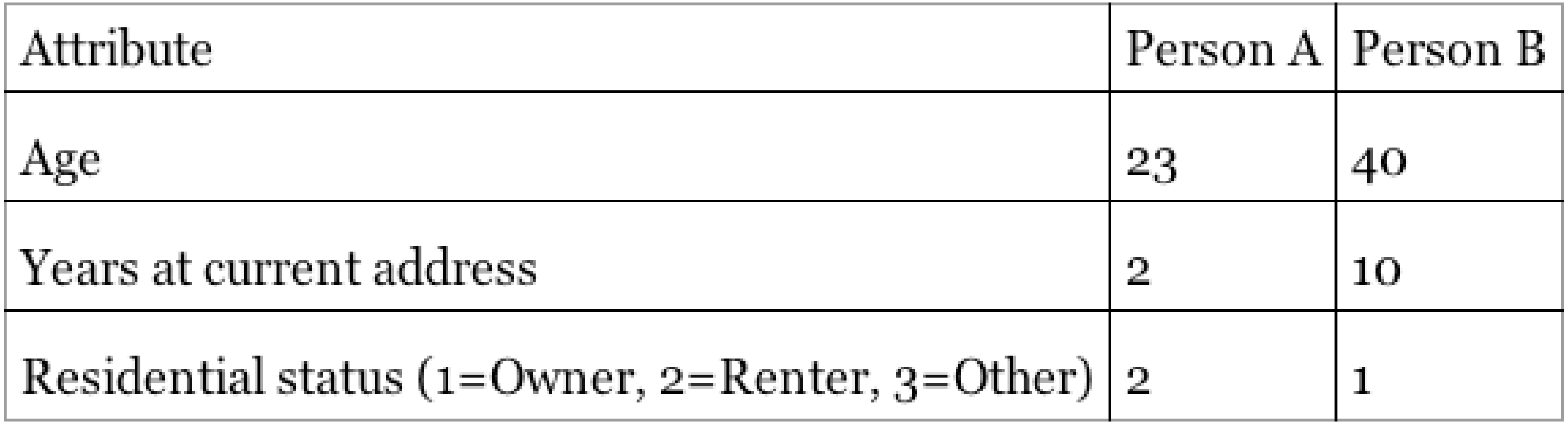
- If two things (people, companies, products) are similar in some ways, they often share other characteristics as well.



Tasks involving with similarity

- Retrieve similar things directly
  - IBM wants to find companies that are similar to their best business customers
- Use in classification or regression
- Group similar items together into clusters
  - To see whether customer base contains group of similar customers and what they have in common
- Provide recommendations of similar products
  - Amazon, Netflix
- Provide reasoning from similar cases
  - Knowledge management, troubleshooting

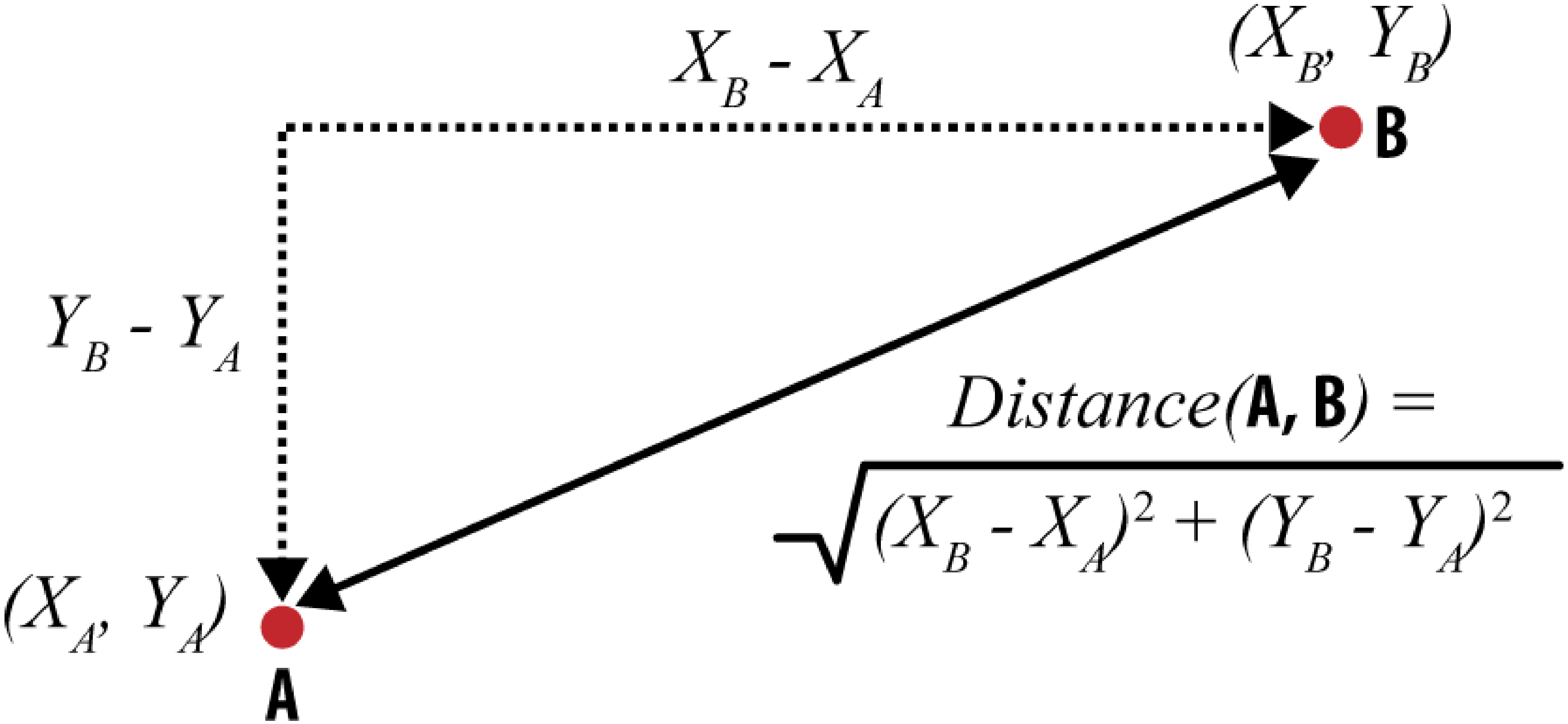
Example: credit card application



Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1

- Multiple attributes
- No single best method for reducing them to a single distance measurement

Geometric approach
Euclidean distance

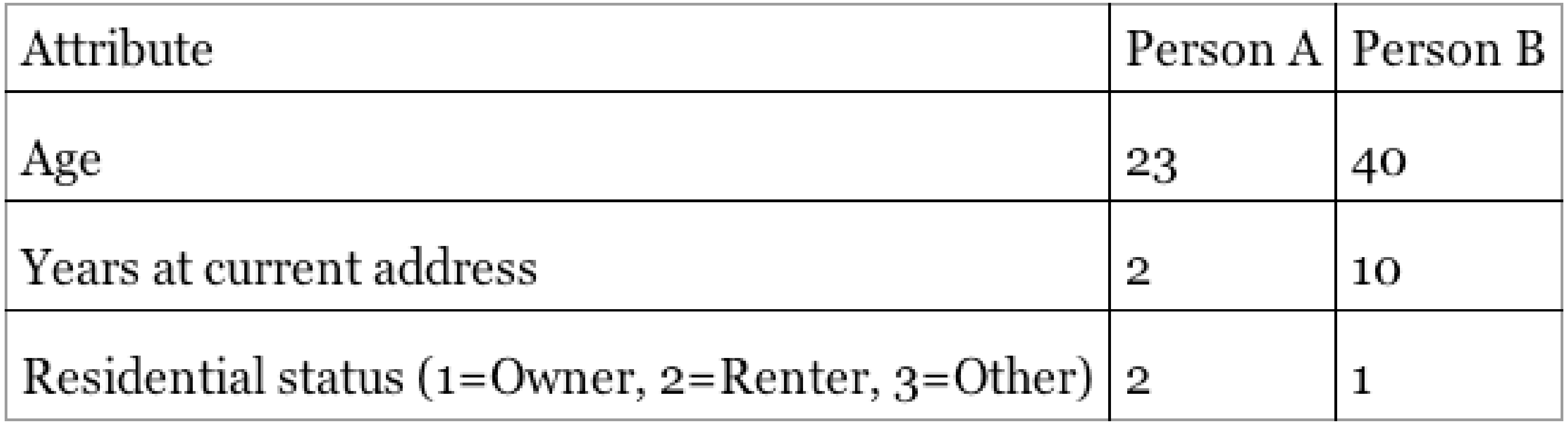


$$\textit{Distance}(\mathbf{A}, \mathbf{B}) = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}$$

Equation 6-1. General Euclidean distance

$$\sqrt{(d_{1,A} - d_{1,B})^2 + (d_{2,A} - d_{2,B})^2 + \ldots + (d_{n,A} - d_{n,B}}$$

Example: credit card application



Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1

$$d\left(A,B\right)=\sqrt{\left(23-40\right)^2+\left(2-10\right)^2+\left(2-1\right)^2}$$

$$\approx 18.8$$

- No unit, no meaningful interpretation
- Useful for comparing the similarity of one pairs to others

Distance Functions

Equation 6-2. Euclidean distance (L2 norm)

$$d_{\text{Euclidean}}(\mathbf{X},\mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdot}$$

Equation 6-3. Manhattan distance (L1 norm)

$$d_{\text{Manhattan}}(\mathbf{X},\mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_1 = | x_1 - y_1 | + | x_2 - y_2 | + \cdot$$

Equation 6-4. Jaccard distance

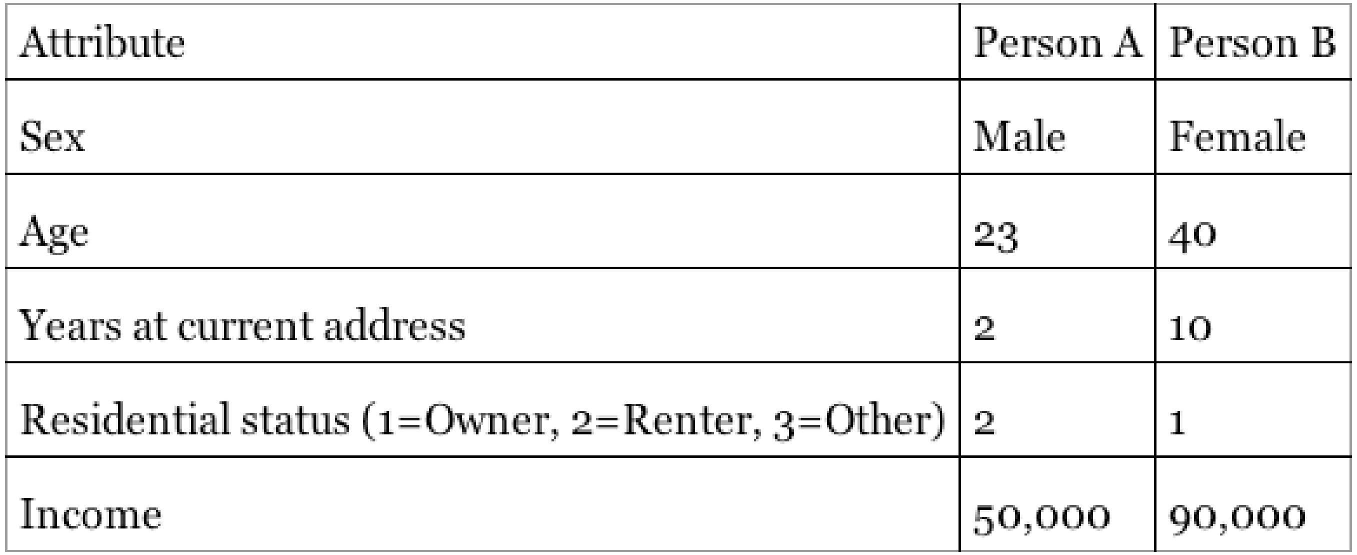
$$d_{\text{Jaccard}}(X,\ Y) = 1 - \frac{| X \cap Y |}{| X \cup Y |}$$

Equation 6-5. Cosine distance

$$d_{\text{cosine}}(\mathbf{X},\mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\| \mathbf{X} \|_2 \cdot \| \mathbf{Y} \|}$$

Similarity calculation
Heterogeneous attributes

- Sometime, attributes may be numeric and categorical
- Also, values may not have the same range
- The variable thus need to be scaled/normalized by its range before calculating similarity

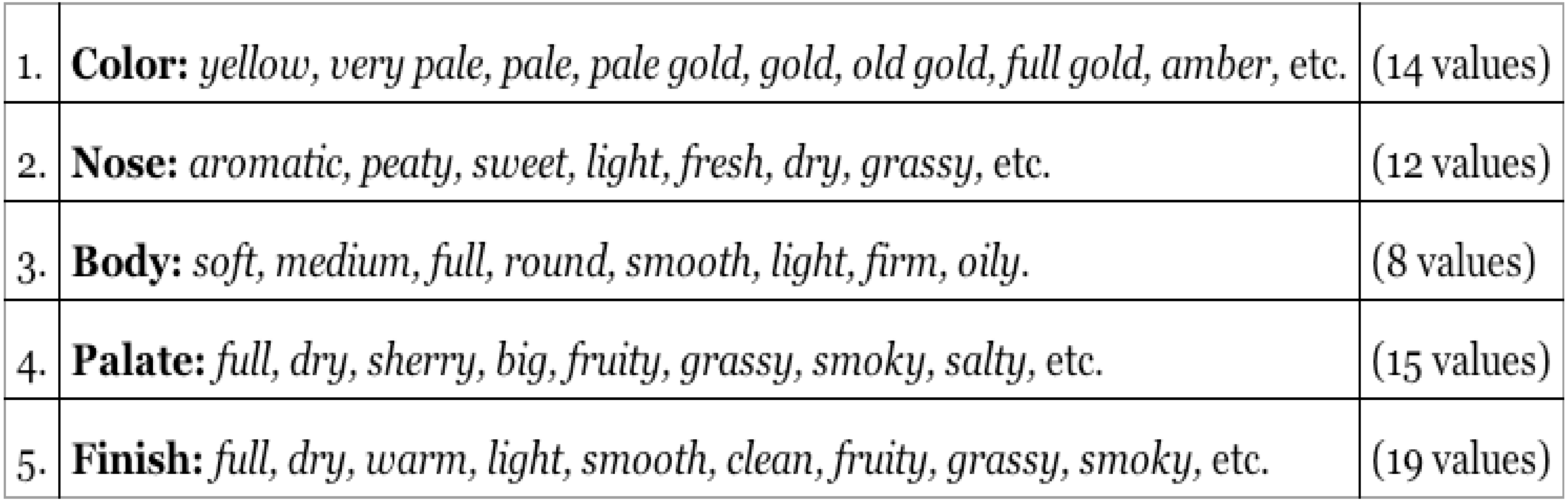


Attribute	Person A	Person B
Sex	Male	Female
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1
Income	50,000	90,000

Nearest neighbor reasoning

- We could use distance to
  - find the companies most similar to our best corporate customers
  - find the online consumers most similar to our best retail customers
- Once we have found these, we can take whatever action is appropriate in the business context
- The most similar instance is called nearest neighbor

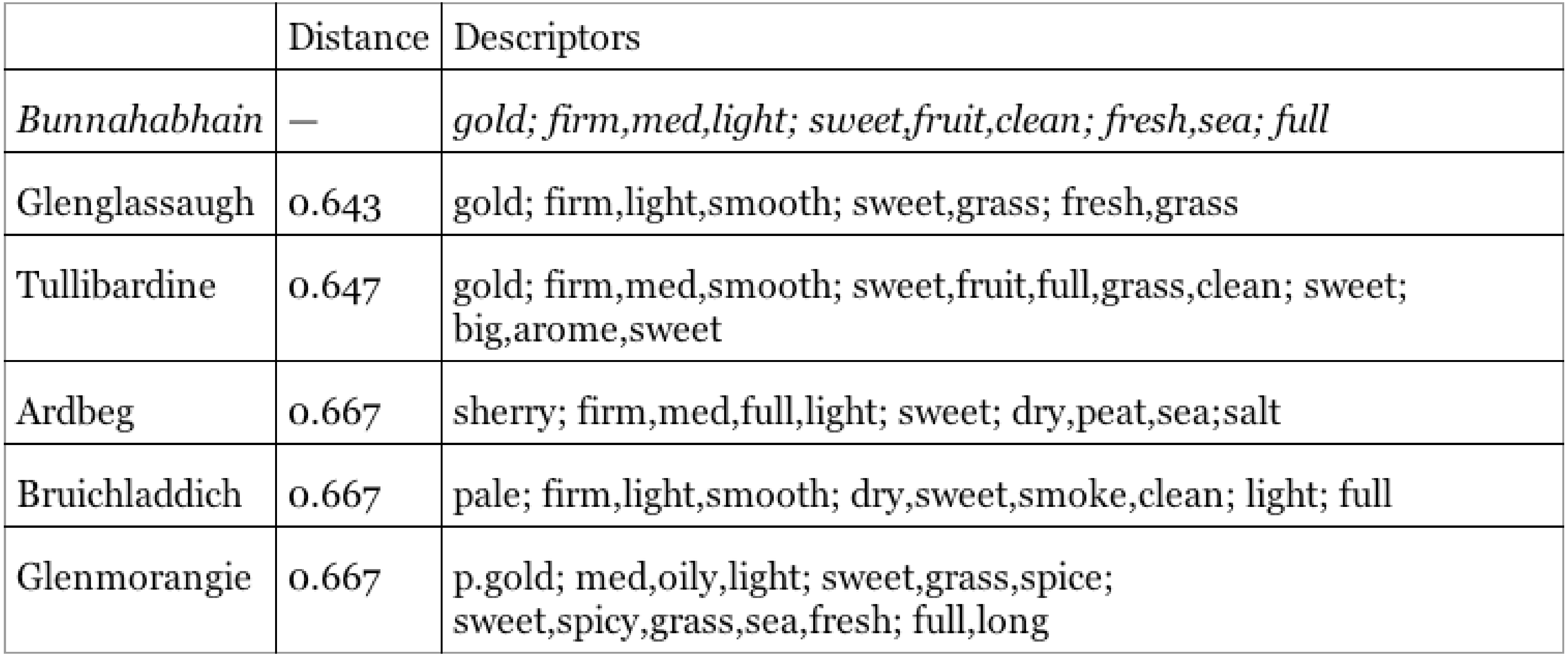
Example: Wine comparison Attribute list



1.	**Color:** *yellow, very pale, pale, pale gold, gold, old gold, full gold, amber,* etc.	(14 values)
2.	**Nose:** *aromatic, peaty, sweet, light, fresh, dry, grassy,* etc.	(12 values)
3.	**Body:** *soft, medium, full, round, smooth, light, firm, oily.*	(8 values)
4.	**Palate:** *full, dry, sherry, big, fruity, grassy, smoky, salty,* etc.	(15 values)
5.	**Finish:** *full, dry, warm, light, smooth, clean, fruity, grassy, smoky,* etc.	(19 values)

68 binary attributes

Example: Drink Comparison Most similar to *Bunnahabhain*



	Distance	Descriptors
Bunnahabhain	—	*gold; firm,med,light; sweet,fruit,clean; fresh,sea; full*
Glenglassaugh	0.643	gold; firm,light,smooth; sweet,grass; fresh,grass
Tullibardine	0.647	gold; firm,med,smooth; sweet,fruit,full,grass,clean; sweet; big,arome,sweet
Ardbeg	0.667	sherry; firm,med,full,light; sweet; dry,peat,sea;salt
Bruichladdich	0.667	pale; firm,light,smooth; dry,sweet,smoke,clean; light; full
Glenmorangie	0.667	p.gold; med,oily,light; sweet,grass,spice; sweet,spicy,grass,sea,fresh; full,long

Dataset are available at http://adn.biol.umontreal.ca/~numericalecology/data/scotch.html

Look-alike customers

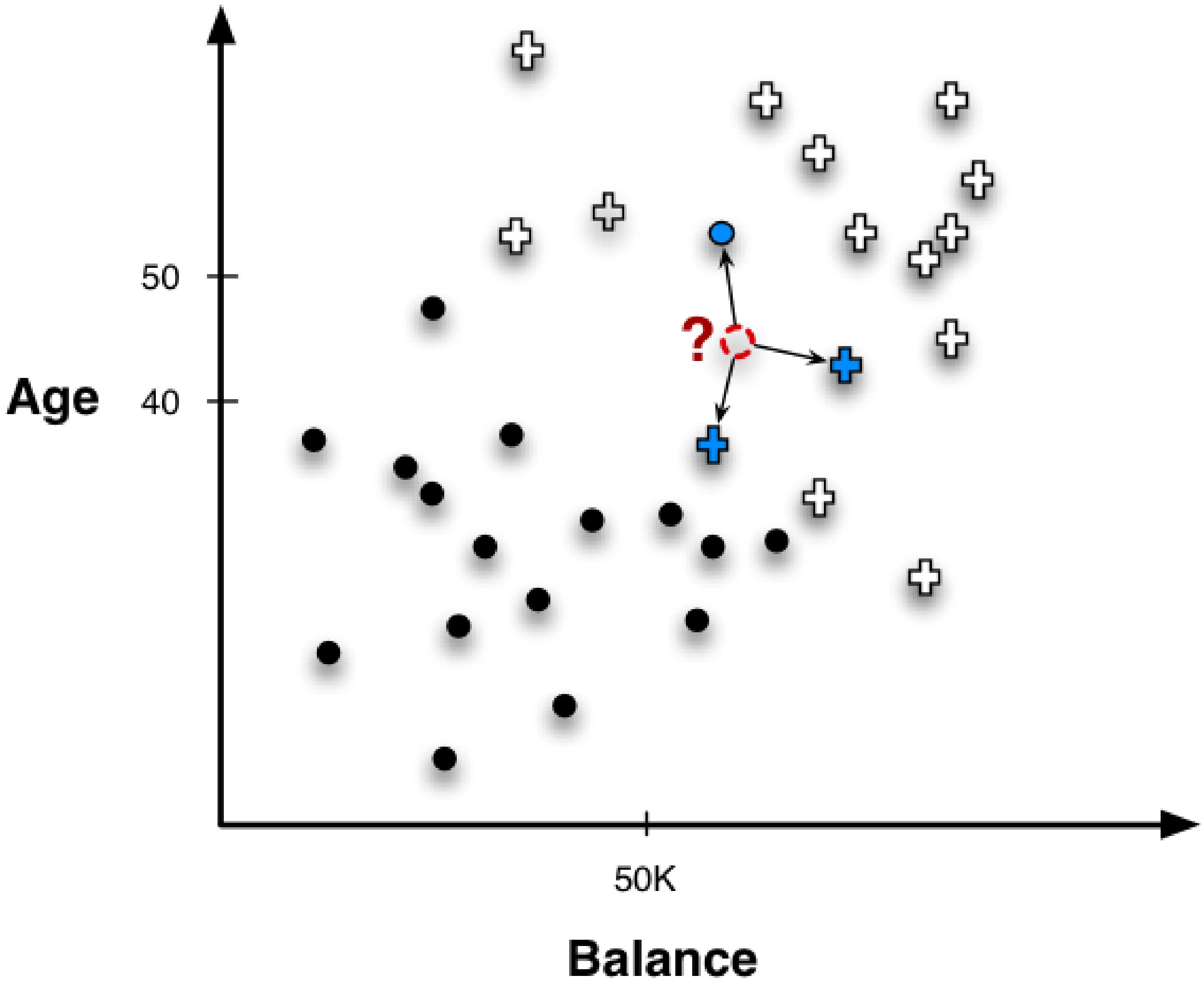
- Finding look-alike customers is one of the practical way to identify potential customers
- The steps include

1. Find the target group
2. Calculate distance from all customer to ..
   - All customers in the target group, or
   - The centroid of the group
3. Rank the customers based on the distance

Nearest neighbor for predictive modeling

- Given a new example whose target variable we want to predict, we scan through all the training examples and choose several that are most similar to the new examples
- Then we predict the new example's target value, based on the nearest neighbors' (known) target values.
- The prediction is based on a combining function

Classification



What should be our combining function?

A simple combining function would be majority vote

The predicted class would be positive

Example: similarity in prediction

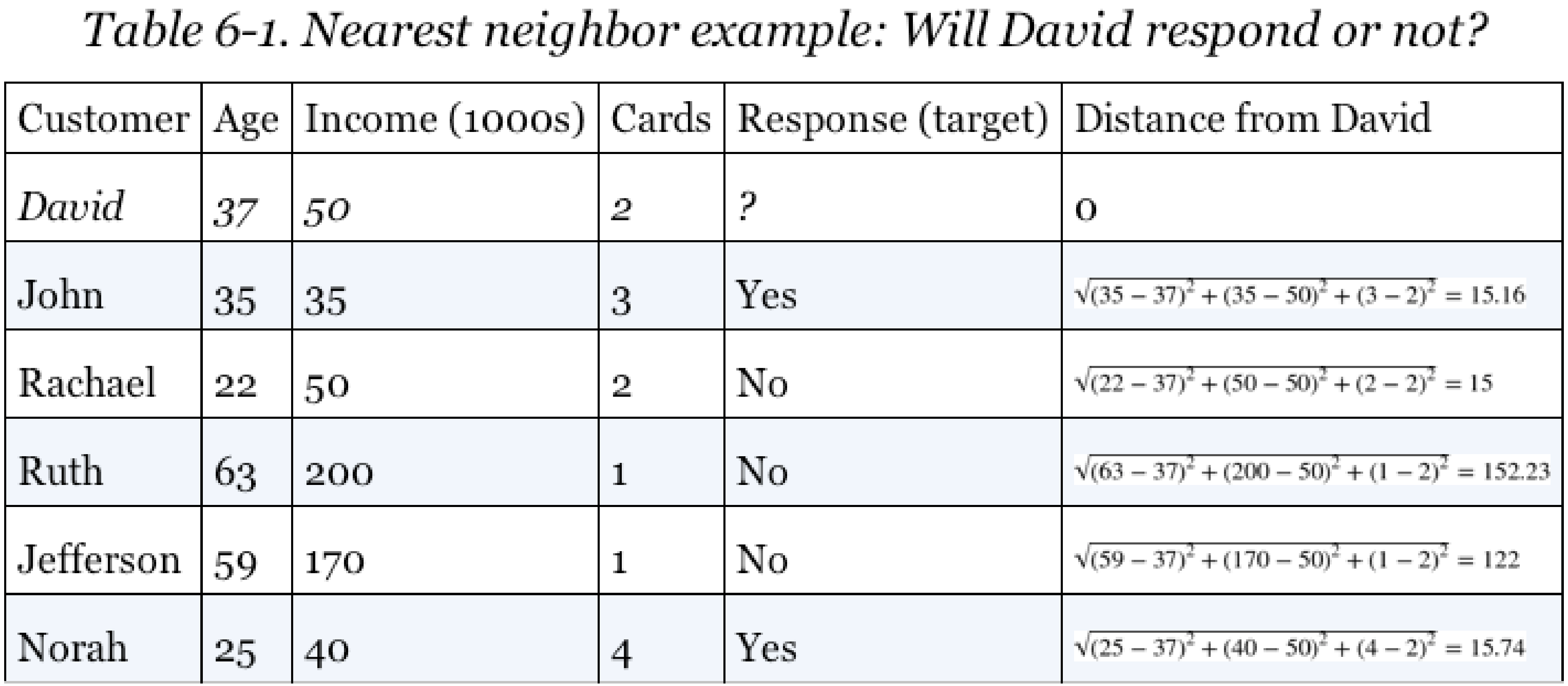


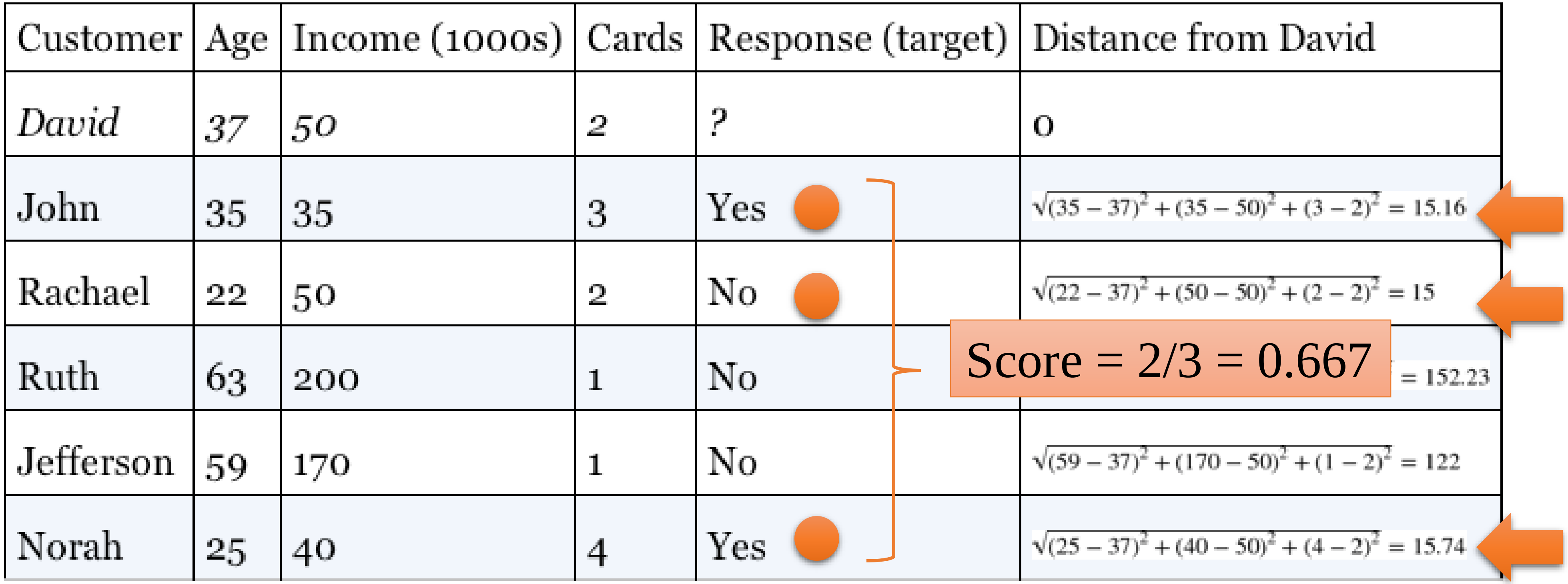
Table 6-1. Nearest neighbor example: Will David respond or not?

Customer	Age	Income (1000s)	Cards	Response (target)	Distance from David
David	*37*	*50*	*2*	*?*	0
John	35	35	3	Yes	$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2} = 15.16$
Rachael	22	50	2	No	$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2} = 15$
Ruth	63	200	1	No	$\sqrt{(63-37)^2+(200-50)^2+(1-2)^2} = 152.23$
Jefferson	59	170	1	No	$\sqrt{(59-37)^2+(170-50)^2+(1-2)^2} = 122$
Norah	25	40	4	Yes	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2} = 15.74$

How many neighbors should we use?
Should they have equal weights in the combining function?

Probability estimation

- Probability estimation is as important as Yes/No decision
- Consider again the previous example of deciding whether David will be a responder or not

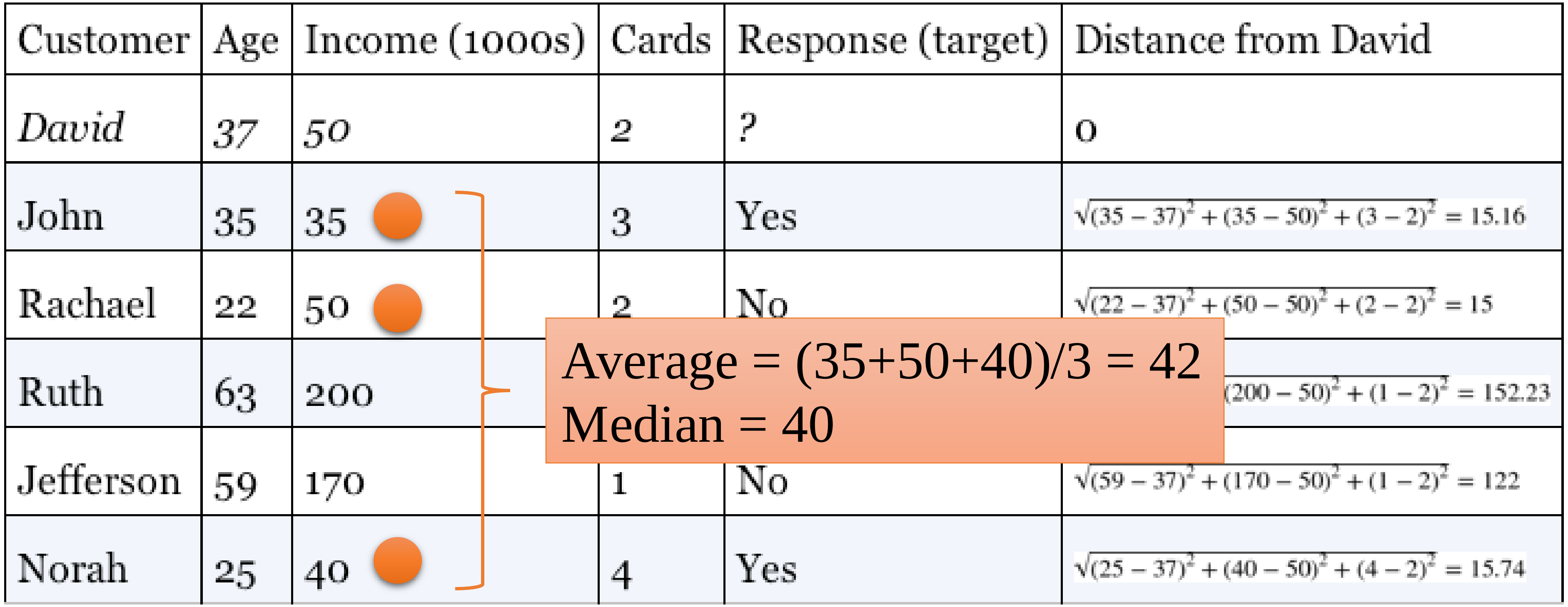


Customer	Age	Income (1000s)	Cards	Response (target)	Distance from David
David	*37*	*50*	*2*	*?*	0
John	35	35	3	Yes ●	$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2} = 15.16$
Rachael	22	50	2	No ●	$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2} = 15$
Ruth	63	200	1	No	$= 152.23$
Jefferson	59	170	1	No	$\sqrt{(59-37)^2+(170-50)^2+(1-2)^2} = 122$
Norah	25	40	4	Yes ●	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2} = 15.74$

Score = 2/3 = 0.667

Regression

- Once we can retrieve nearest neighbors, we can use them for any predictive mining task by combining them in different ways
- Note that, in retrieving neighbors, we are not using the target variable in prediction.



Customer	Age	Income (1000s)	Cards	Response (target)	Distance from David
David	*37*	*50*	*2*	*?*	0
John	35	35 ●	3	Yes	$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2} = 15.16$
Rachael	22	50 ●	2	No	$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2} = 15$
Ruth	63	200			$(200-50)^2+(1-2)^2 = 152.23$
Jefferson	59	170	1	No	$\sqrt{(59-37)^2+(170-50)^2+(1-2)^2} = 122$
Norah	25	40 ●	4	Yes	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2} = 15.74$

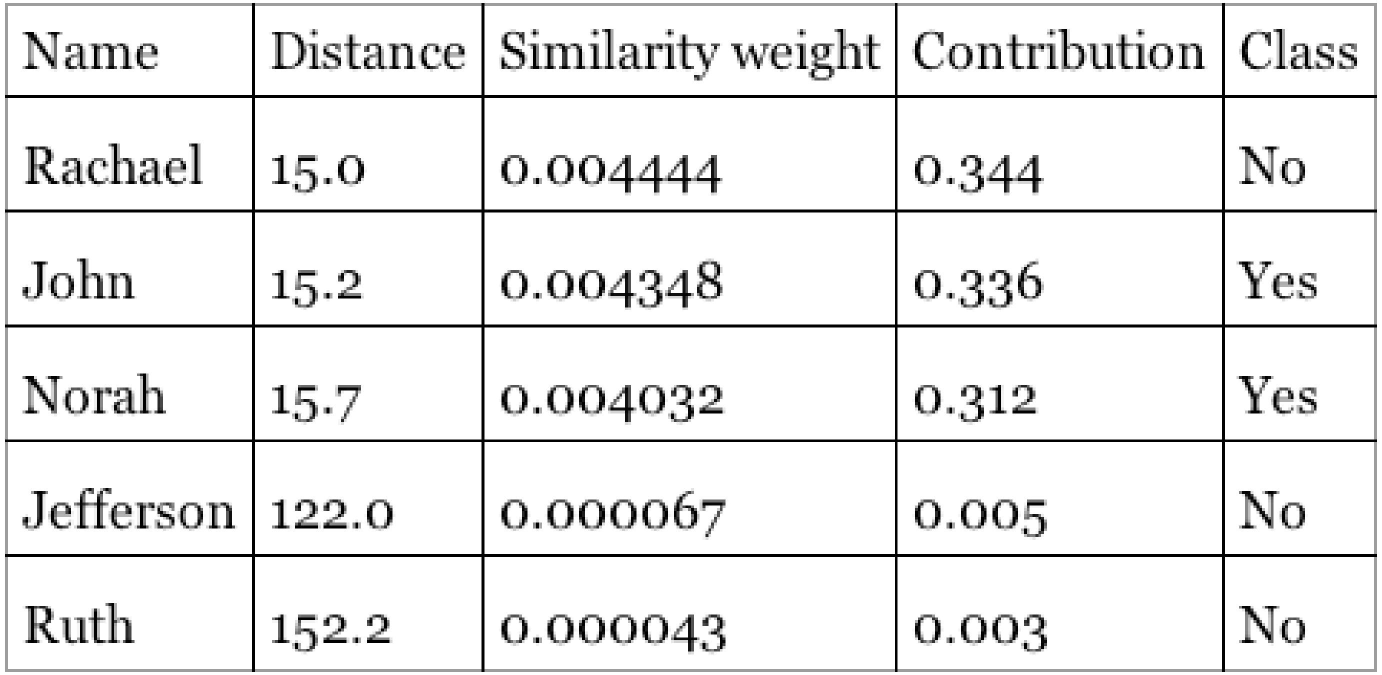
Average = (35+50+40)/3 = 42
Median = 40

How many neighbors and how much influence?

- No simple answer to how many neighbors should be used
- Nearest neighbor algorithms are often referred by k-NN, where k is the number of neighbors
- The greater k is, the more estimates are smoothed out among neighbors
- If $k = n$ (number of training samples), the majority is the class probability
- But the answer is sensitive to the number k

Weighted voting

- Nearest neighbor method often uses weighted voting or similarity moderated voting by scaling neighbor contribution by distance



Name	Distance	Similarity weight	Contribution	Class
Rachael	15.0	0.004444	0.344	No
John	15.2	0.004348	0.336	Yes
Norah	15.7	0.004032	0.312	Yes
Jefferson	122.0	0.000067	0.005	No
Ruth	152.2	0.000043	0.003	No

Unsupervised segmentation

- In some applications, we may want to find groups of objects, that are not driven by predefined targets
- Can we develop better products, better marketing campaigns, better sales methods, or better customer service, if we understand customer's natural subgroup better?
- This task is called unsupervised segmentation, or more simply clustering

What is clustering?

Clustering: a task of dividing up data into groups (clusters), so that points in any one group are more “similar” to each other than to points outside the group.

Why clustering

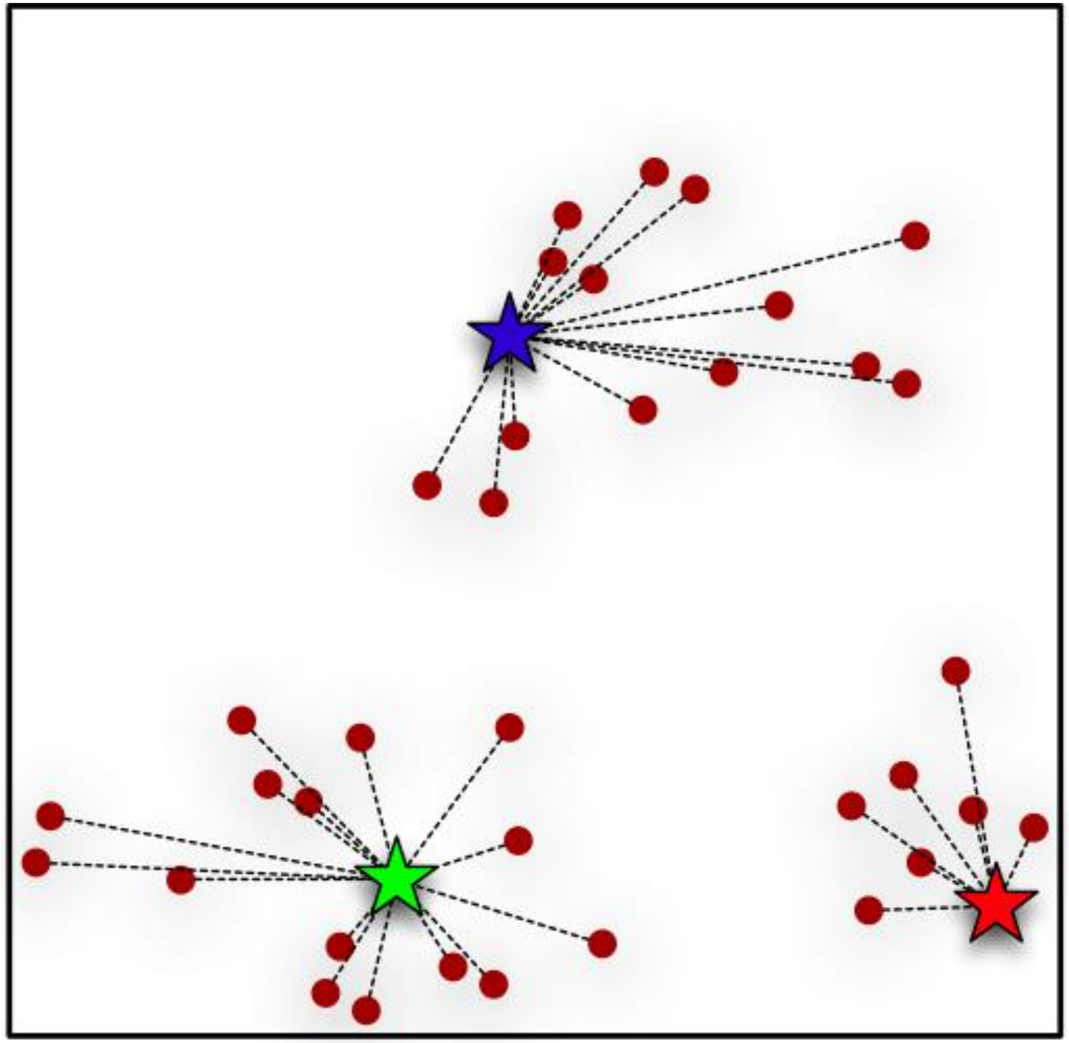
- Summary: deriving a reduced representation of the full data set. E.g., centroid representation
- Discovery: looking for new insights into the structure of the data. E.g., finding groups of students that commit similar mistakes, or groups of 80s songs that sound alike

Other uses, e.g.,

- Checking up on someone else’s work/decisions, investigating the validity of pre-existing group assignments
- Helping with prediction, i.e., in classification or regression

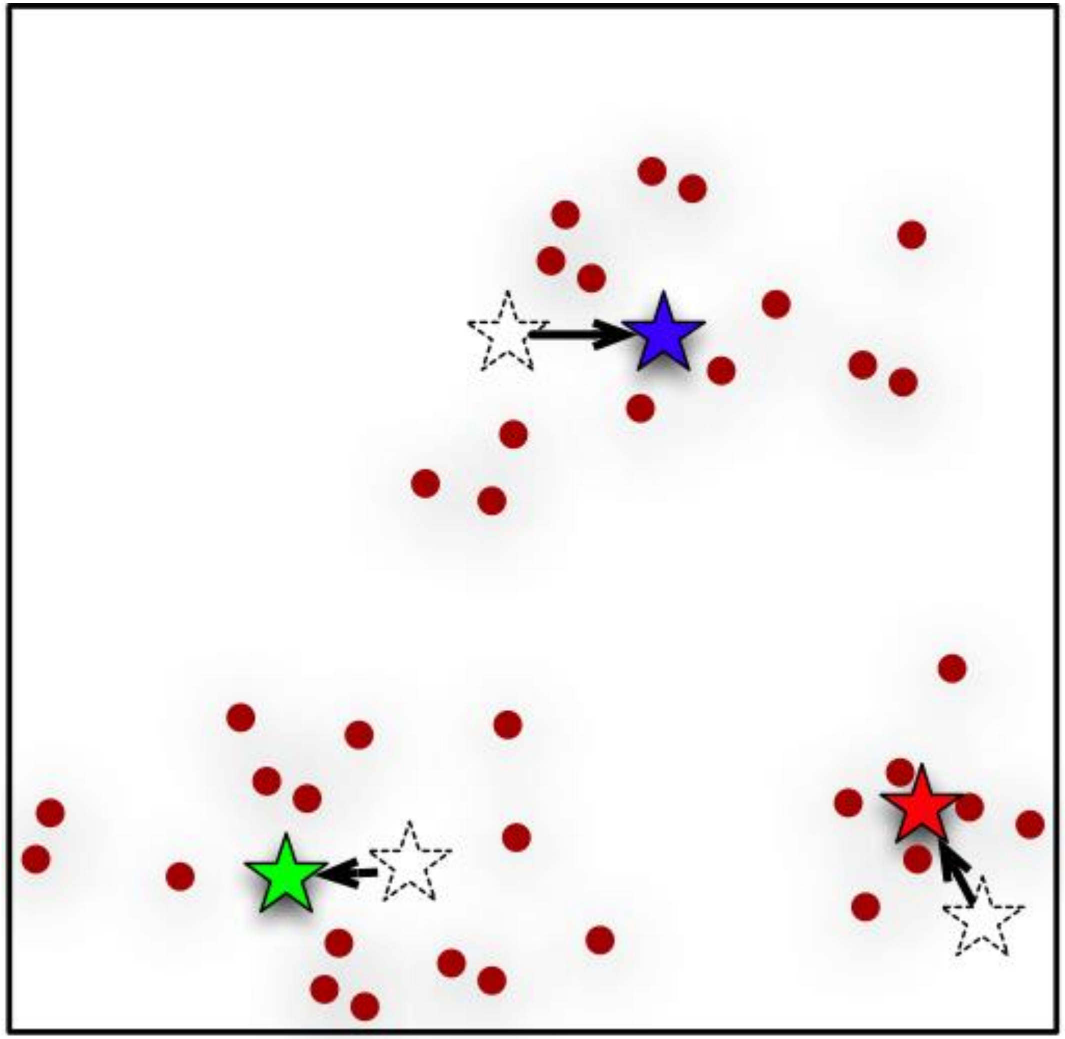
Centroid

- Hierarchical clustering focuses on similarities between the individual instances
- A difference way is to focus on the center of the cluster themselves, or centroid
- The star is not exactly the point but rather the center



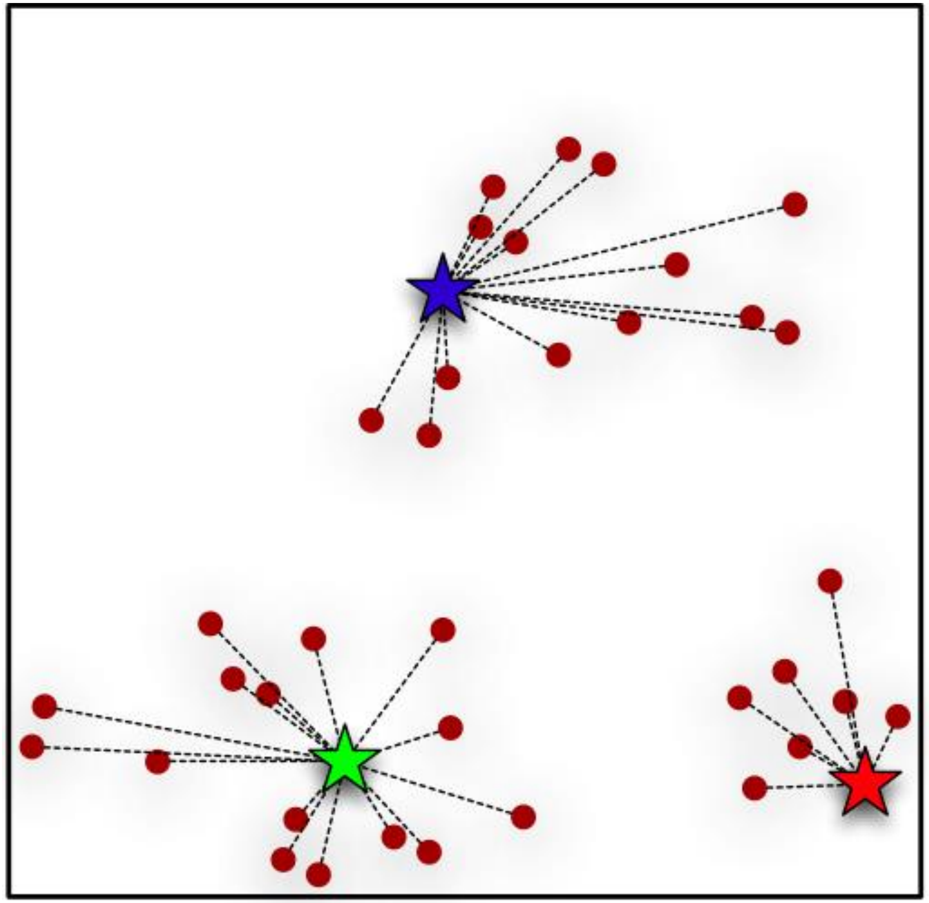
k-means clustering

- The most popular centroid-based algorithm is called *k-means* clustering
- In k-means, the “means” are the centroids, represented by the arithmetic means (average) of the values along each dimension



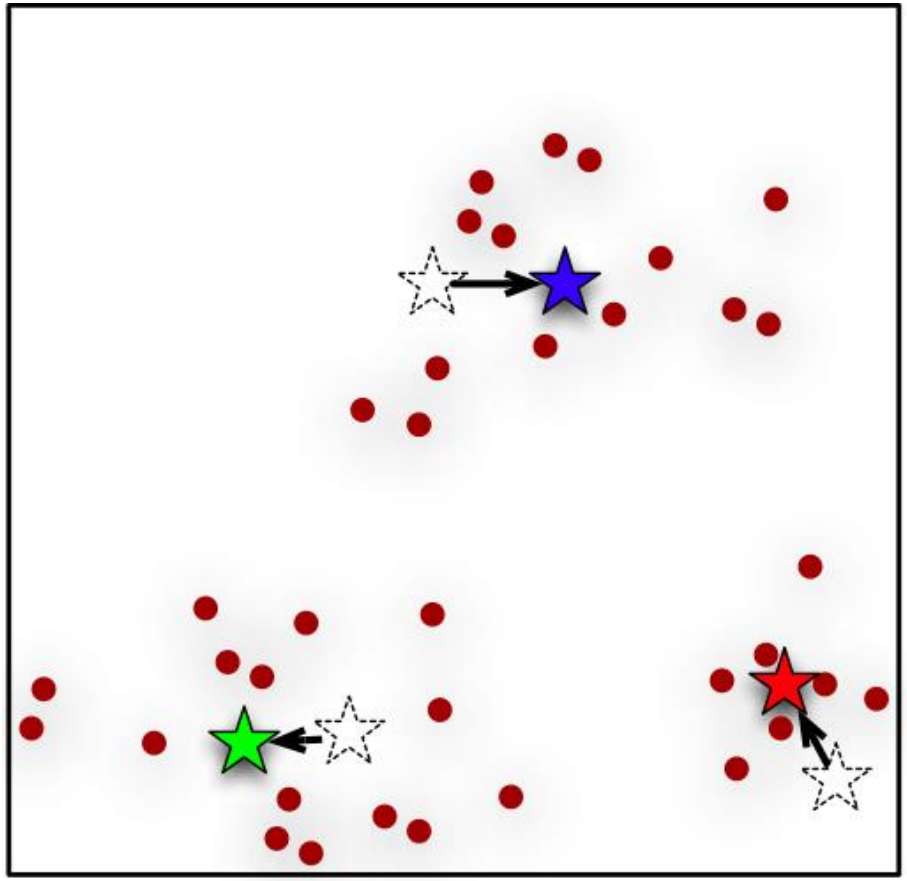
k-means clustering: step 1

1. Create *k* initial cluster center, usually randomly, but sometimes by randomly choosing *k* actually data points



k-means clustering: step 2

2. For each cluster, its center is relocated by finding the actual centroid point of the cluster



Algorithm

Given an initial set of k means $m_1{}^{(1)},\ldots,m_k{}^{(1)}$

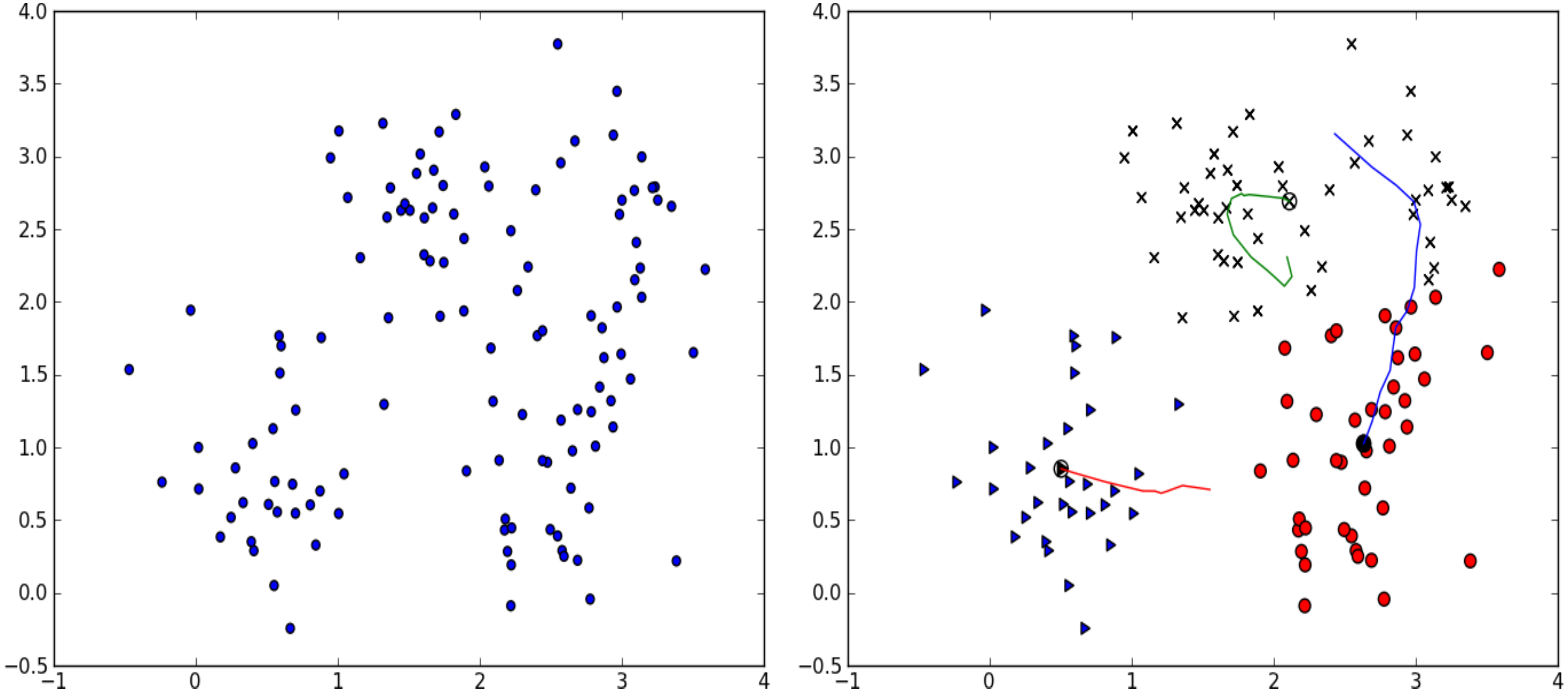
- **Assignment**: Assign each observation to the nearest cluster by calculating distance to the centroid

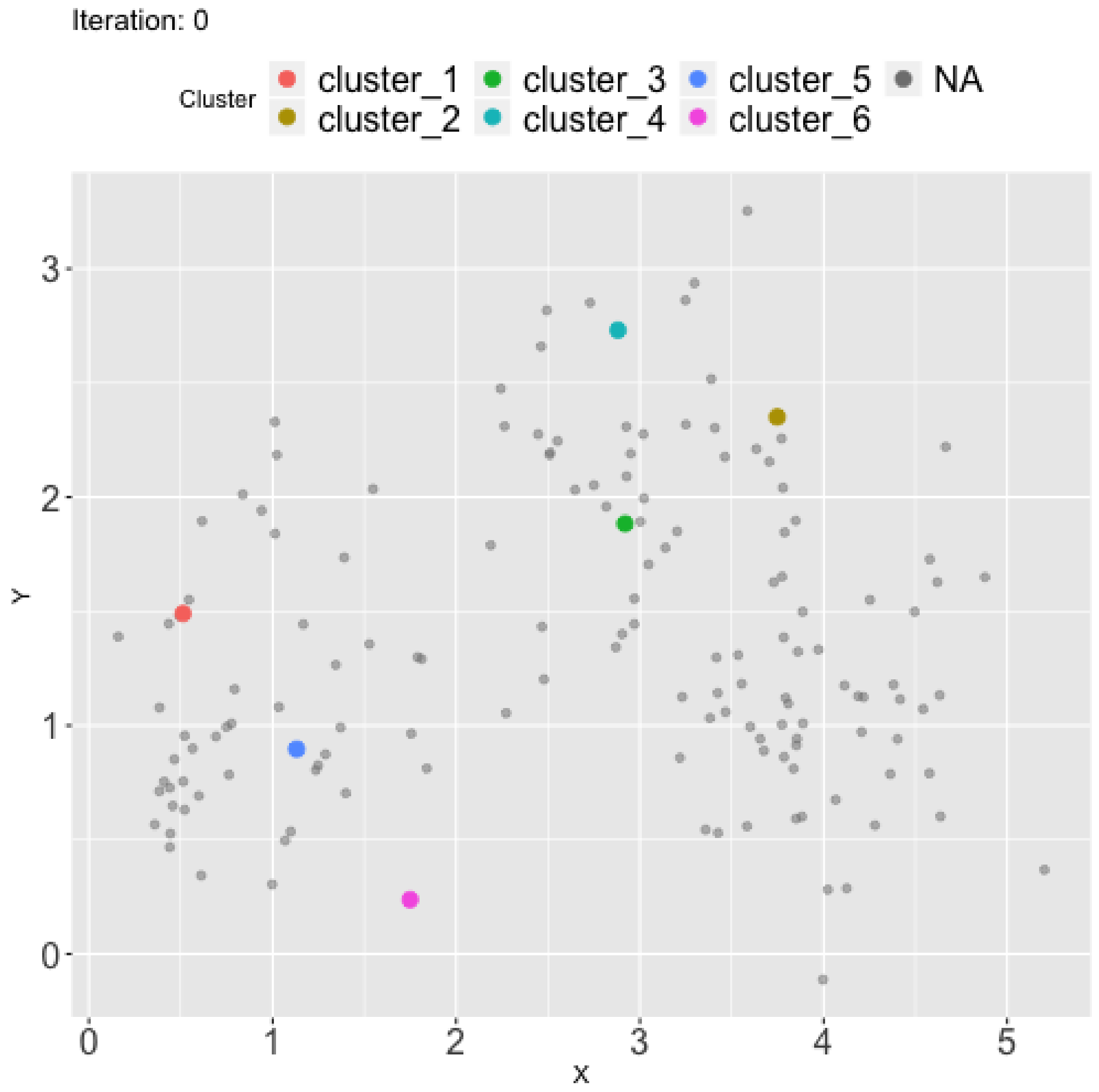
$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2, \forall j, 1 \leq j \leq k \right\}$$

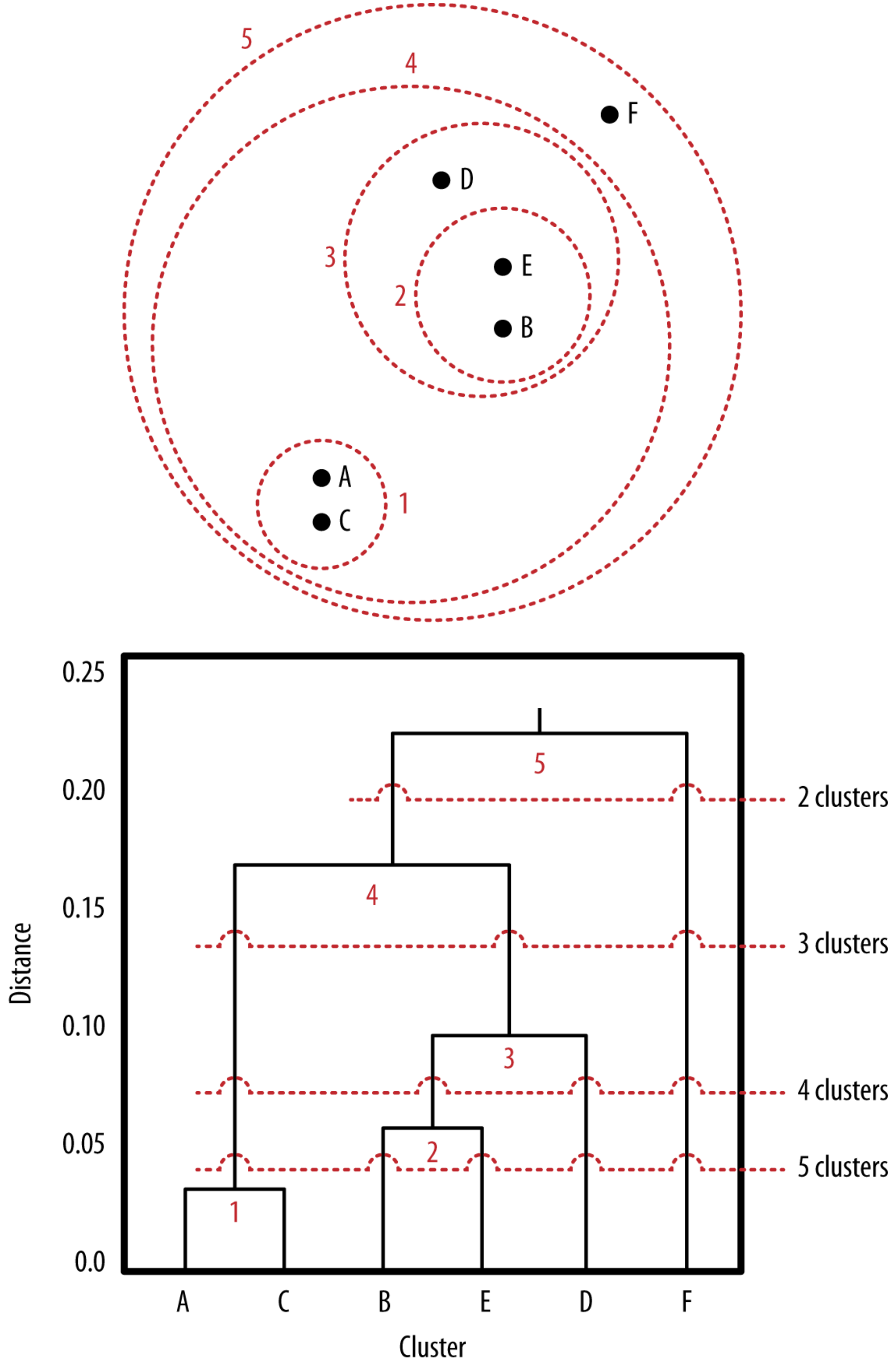
- **Update**: Calculate the new means to be the centroids of the observations in the new clusters

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Example: 90 data points




Iteration: 0
Cluster
● cluster_1
● cluster_2
● cluster_3
● cluster_4
● cluster_5
● cluster_6
● NA
y
3
2
1
0
x
0
1
2
3
4
5



Hierarchical Clustering

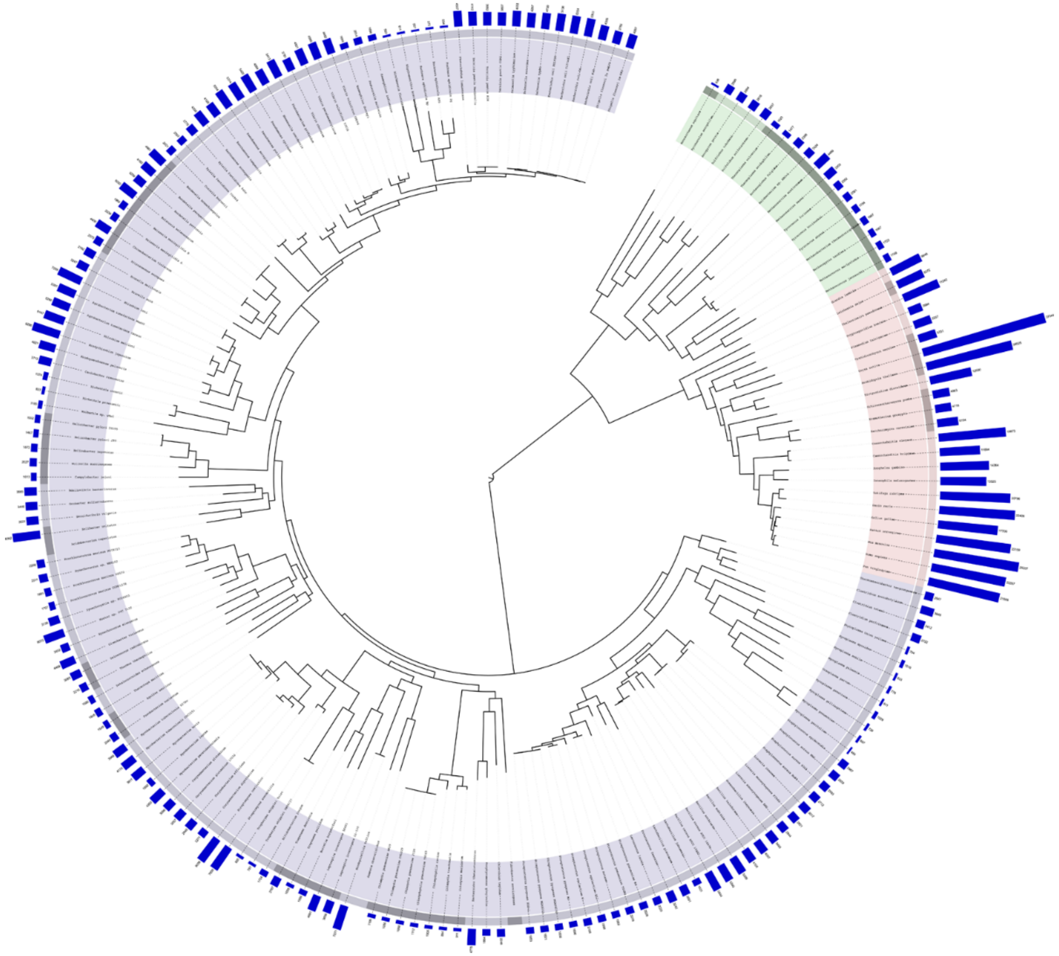
- Nearest neighbor pairs are grouped to clusters
- For example, A and C are closest so they are grouped first. B and E follow that.
- The technique is called “hierarchical” clustering
- The diagram is known as dendrogram

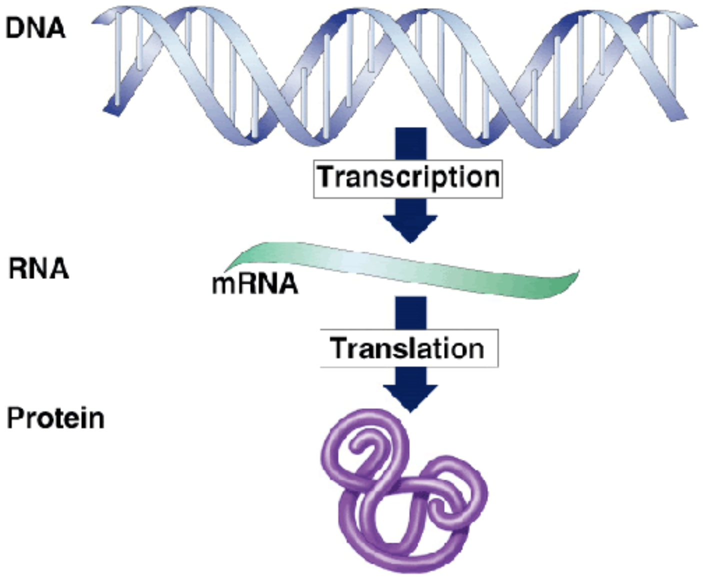
Hierarchical Clustering

- The advantage of hierarchical clustering is that it allows the data analyst to see the groupings before deciding the number of clusters
- Hierarchical clustering are formed by
  - Starting with each node as its own cluster
  - Merging iteratively until only a single cluster remain
  - The clusters are merged based on similarity/distance function that is chosen

BIG DATA EXPERIENCE

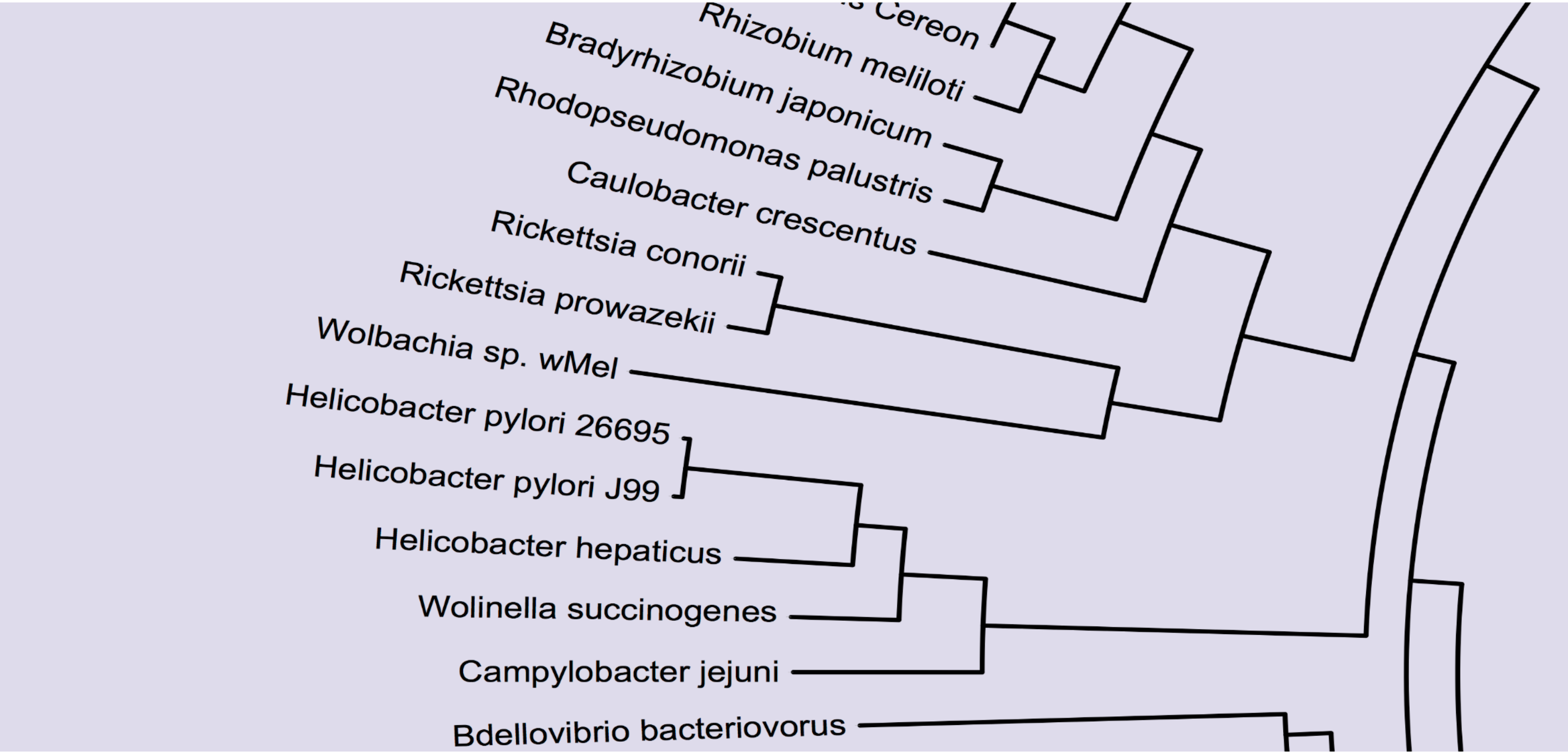
Tree of Life
based on RNA sequence



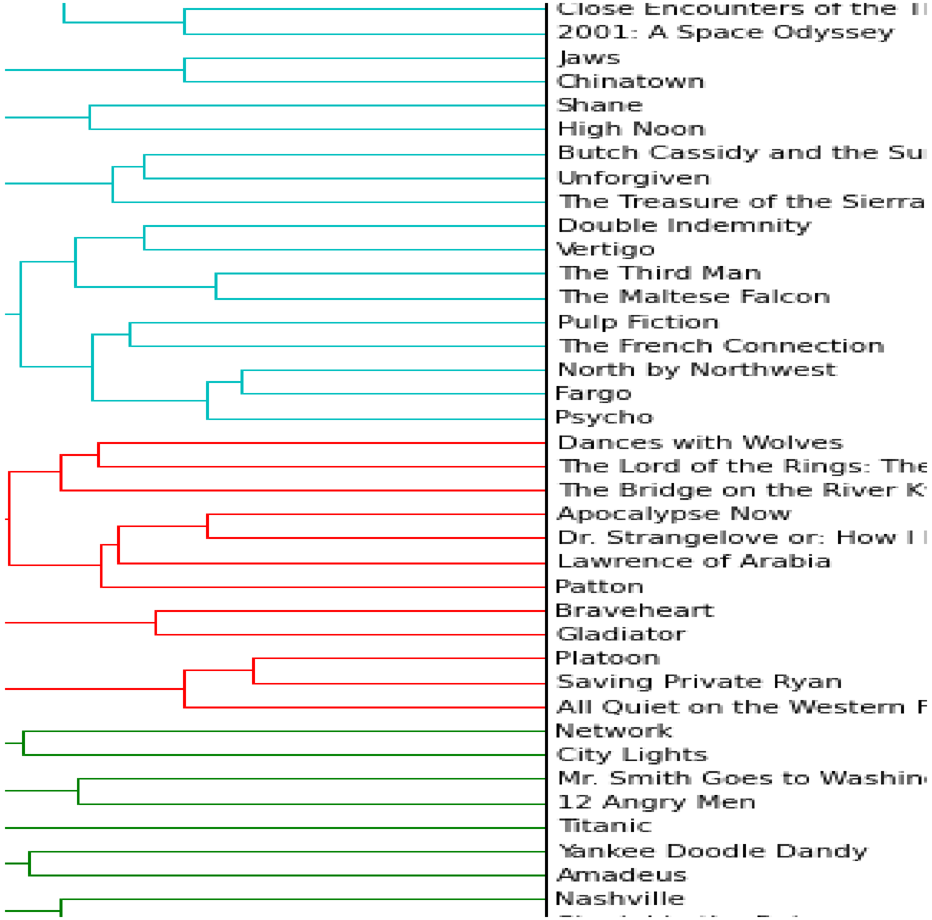


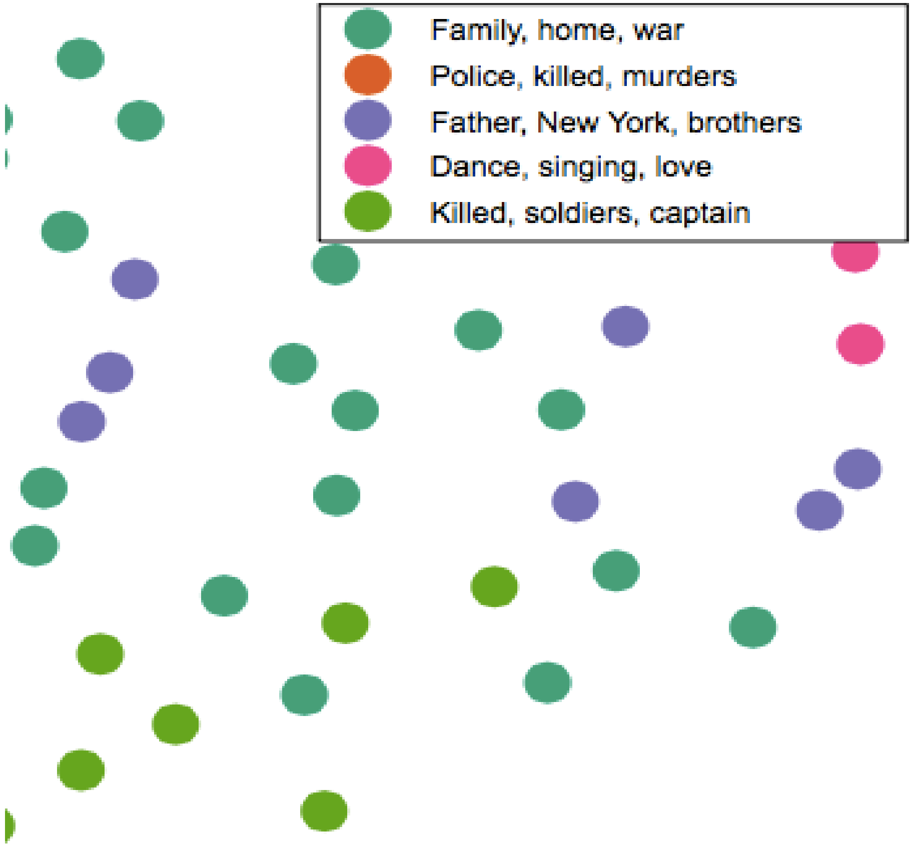
KMUTT

A portion of the Tree of Life



Clustering movies





Lab

- Use mtcars data
- Conduct a clustering study
- Interpret and discuss your results

BIG DATA
EXPERIENCE

Thank you

Question?

BIG DATA
EXPERIENCE
CENTER

KMUTT