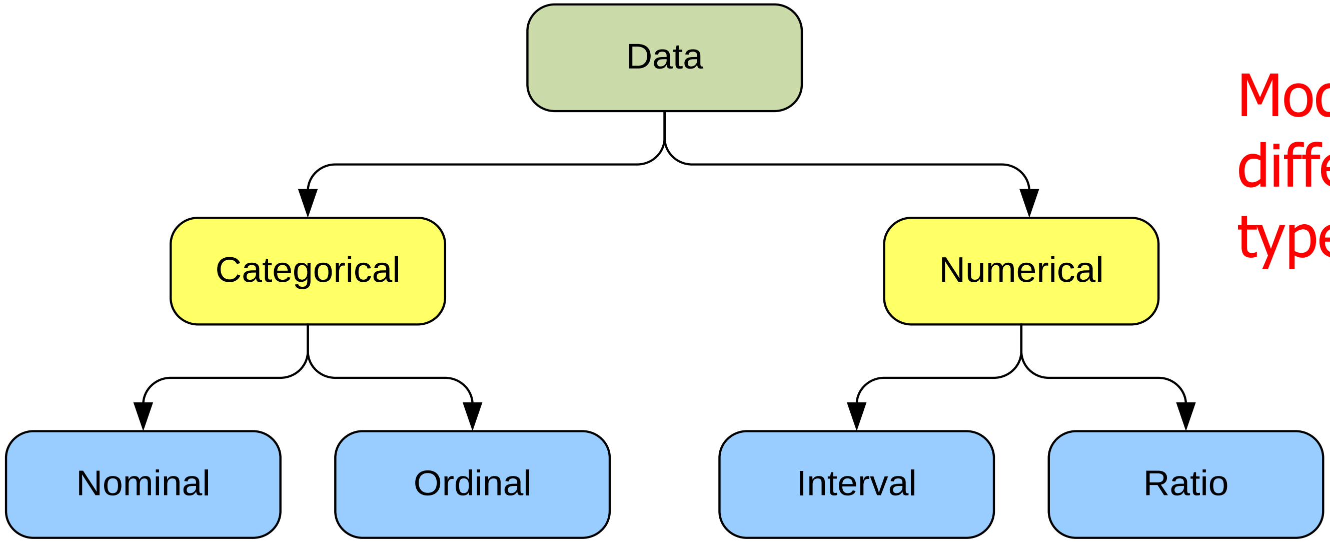
K-Means Clustering

Data

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, sounds, videos & images
- Data: lowest level of abstraction (from which information and knowledge are derived)



Models with different data types?

Type of patterns?

- Classification
- Association Rule
- Prediction
- Clustering (segmentation)
- Sequential (or time series) relationships
- Outer or outlier detection
- Seasonal patterns

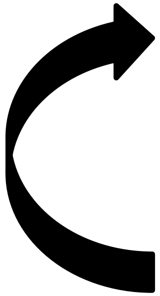
Predictive Modeling

Step 1: Research Problem

Step 2: Data Understanding

Step 3: Data Preparation & cleaning

Accounts for ~85% of total project time



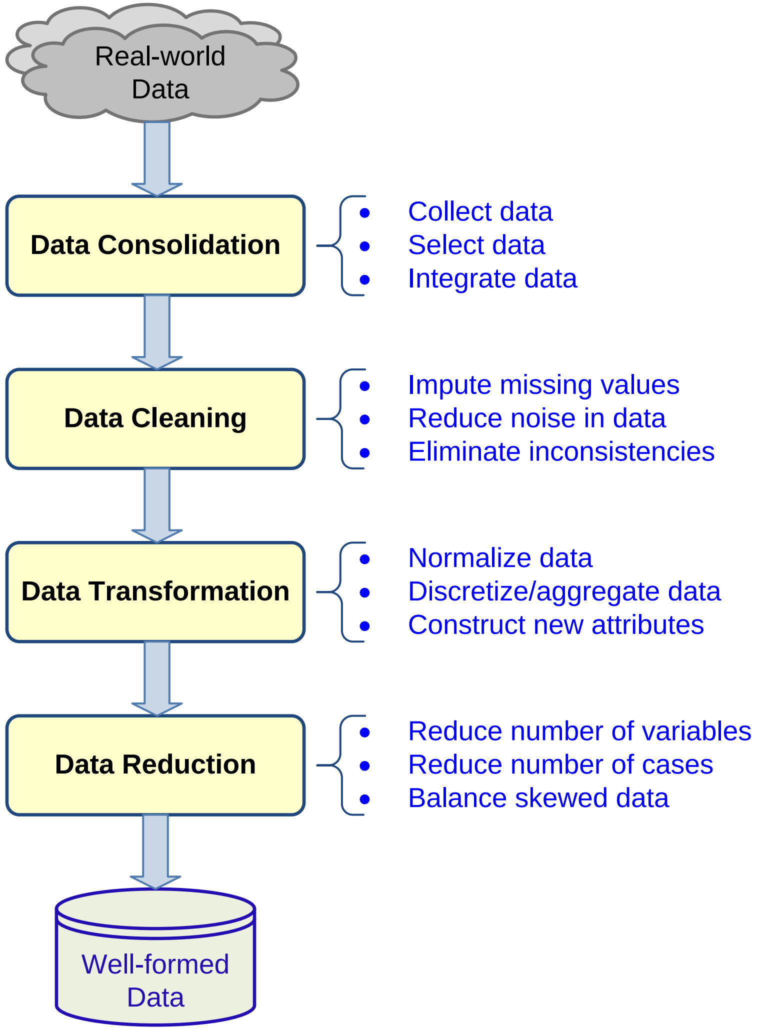
Step 4: Propose model

Step 5: Model Learning

Step 5: Testing and Evaluation

Step 6: Deployment

Data Preparation – A Critical DM Task



- Collect data
- Select data
- Integrate data

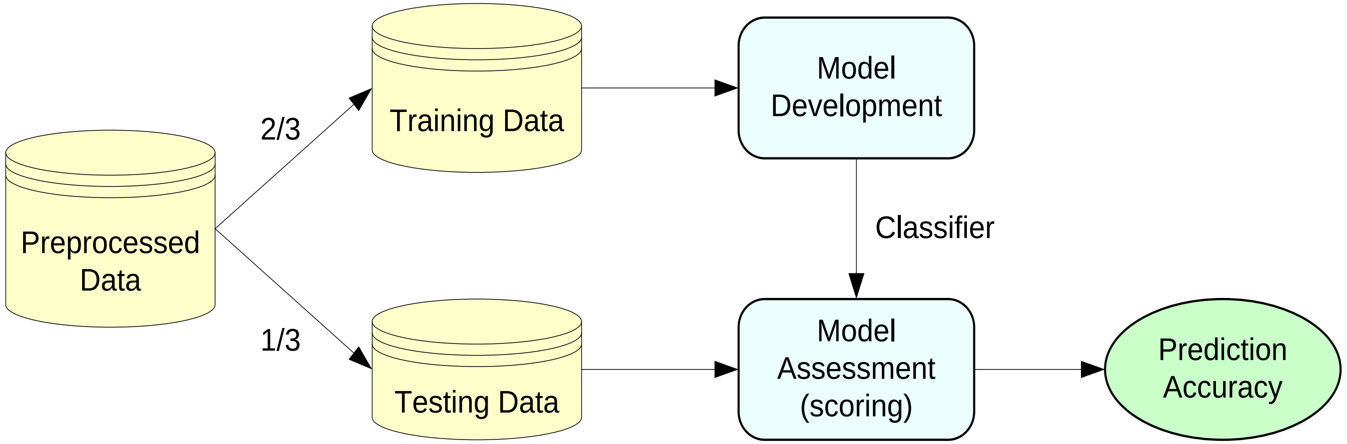
- Impute missing values
- Reduce noise in data
- Eliminate inconsistencies

- Normalize data
- Discretize/aggregate data
- Construct new attributes

- Reduce number of variables
- Reduce number of cases
- Balance skewed data

Data sets

- Training data set
- Testing data set
- New data set



Learning Model

- Supervised learning
- Unsupervised learning

Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Cluster Analysis

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Based on distance measures
- There is no output variable
- Also known as segmentation

Cluster Analysis

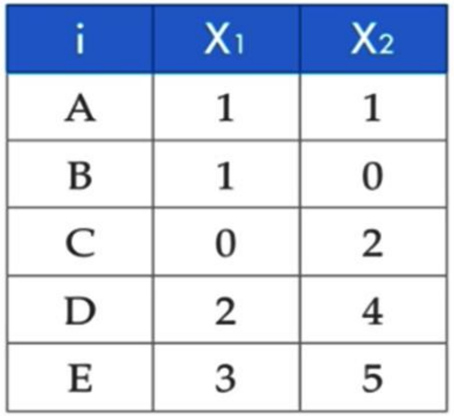
- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on.
  - Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms

Cluster Analysis

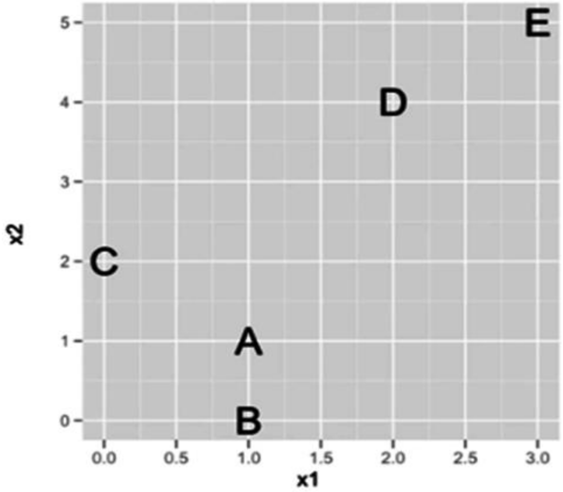
- How many clusters?
  - There is no “truly optimal” way to calculate it
  - Heuristics are often used
    - Look at the sparseness of clusters
    - Number of clusters = $(n/2)^{1/2}$ (n: no of data points)
    - Use Akaike information criterion (AIC)
    - Use Bayesian information criterion (BIC)
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items.
  - Euclidian versus Manhattan (rectilinear) distance

Example: K-mean Clustering

Step 0



i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

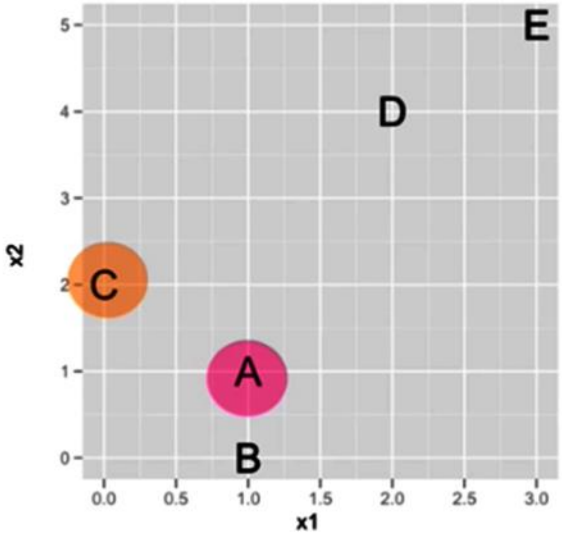


Use K=2. Also A and C are selected as initial means

K-mean Clustering

Step 0

	i	X_1	X_2
$\bar{X}_1^0$ ●	A	1	1
	B	1	0
$\bar{X}_2^0$ ●	C	0	2
	D	2	4
	E	3	5



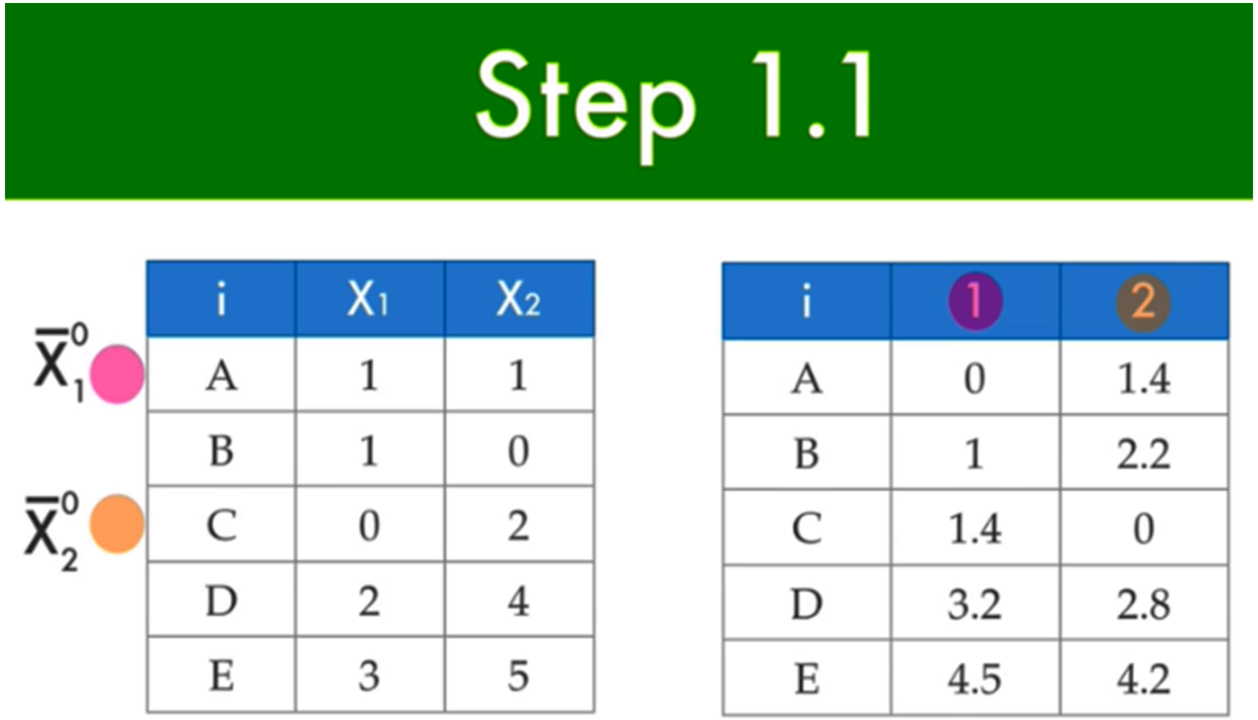
Use K=2. Also A and C are selected as initial cluster means

K-mean Clustering

Step 1.1

Compute the distances of dataset from first cluster mean and from the second cluster mean

K-mean Clustering



Step 1.1

	i	X_1	X_2
$\bar{X}_1^0$ ●	A	1	1
	B	1	0
$\bar{X}_2^0$ ●	C	0	2
	D	2	4
	E	3	5

i	1	2
A	0	1.4
B	1	2.2
C	1.4	0
D	3.2	2.8
E	4.5	4.2

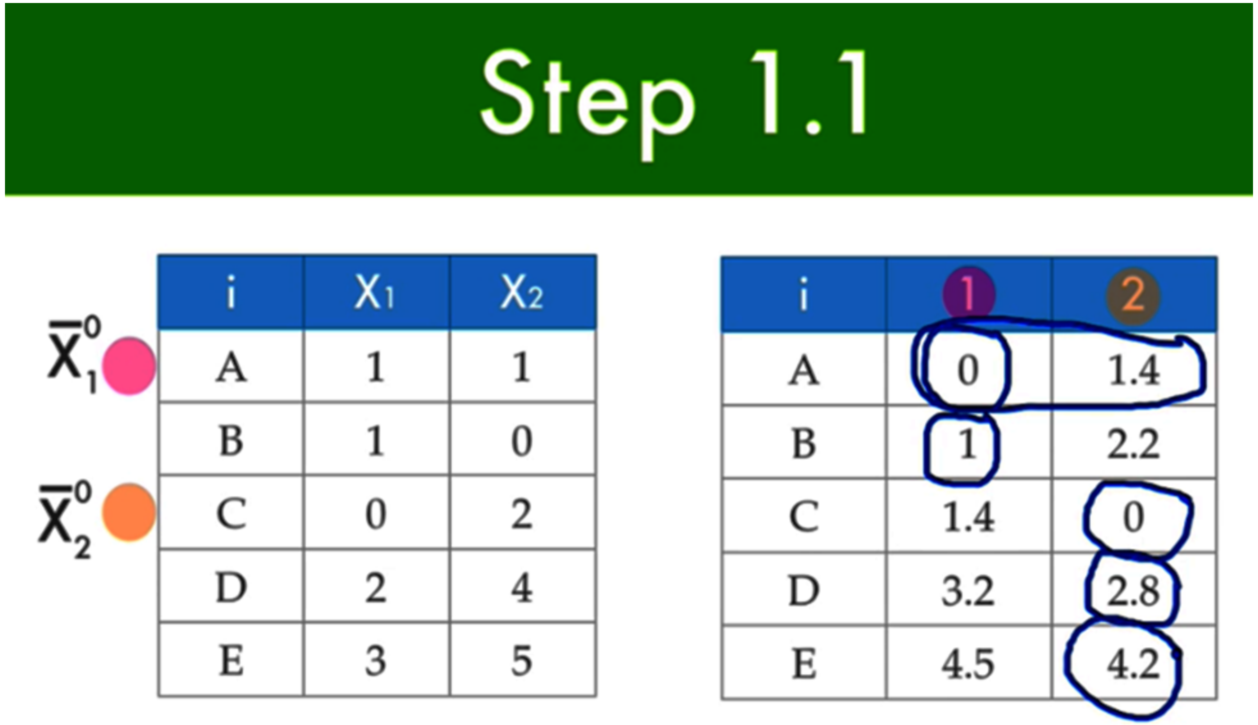
First column contains distances of dataset to first mean and second column are distances of dataset from second mean

K-mean Clustering

Step 1.1

Compare the distances in two columns and assign the element having smaller distance to respective cluster

K-mean Clustering

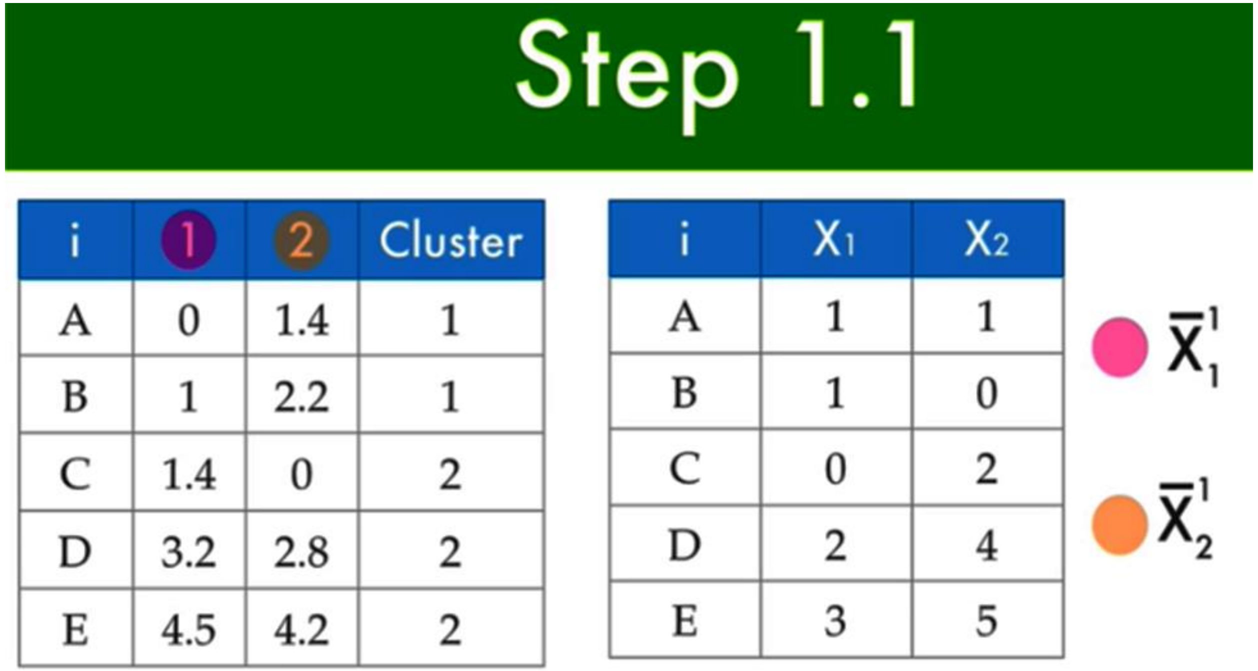


Step 1.1

	i	X1	X2
$\bar{X}_1^0$ ●	A	1	1
	B	1	0
$\bar{X}_2^0$ ●	C	0	2
	D	2	4
	E	3	5

i	1	2
A	0	1.4
B	1	2.2
C	1.4	0
D	3.2	2.8
E	4.5	4.2

K-mean Clustering



Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

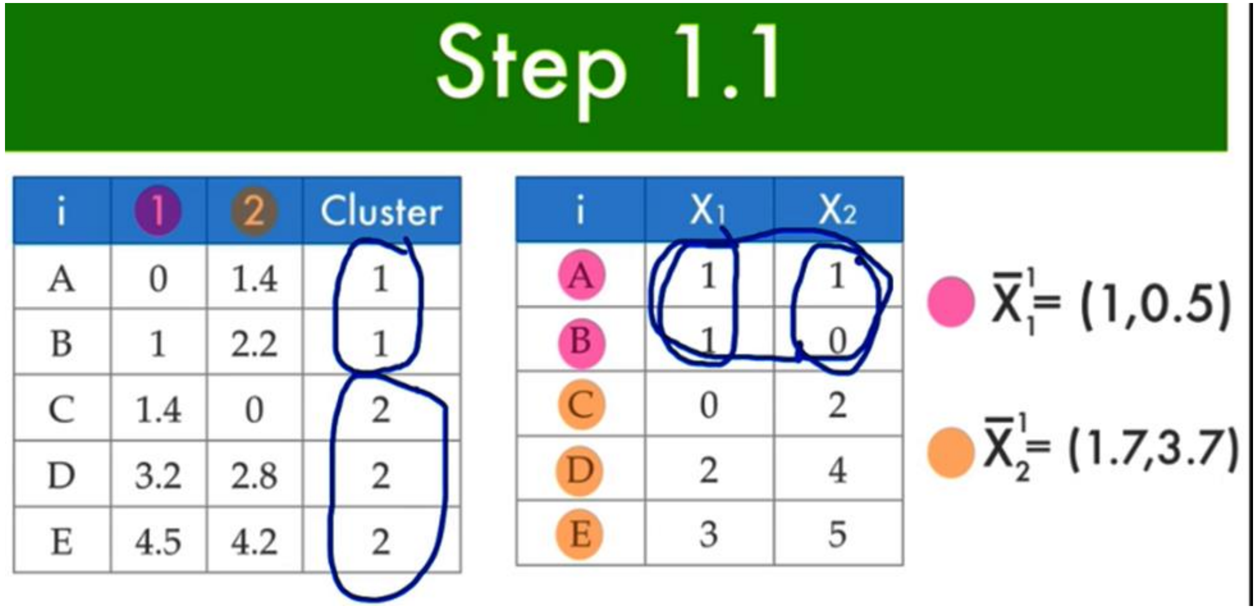
i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

● $\bar{X}_1^1$

● $\bar{X}_2^1$

A and B are assigned to cluster 1 and C, D and E to cluster 2. Recalculate the cluster means

K-mean Clustering



Step 1.1

i	1	2	Cluster
A	0	1.4	1
B	1	2.2	1
C	1.4	0	2
D	3.2	2.8	2
E	4.5	4.2	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

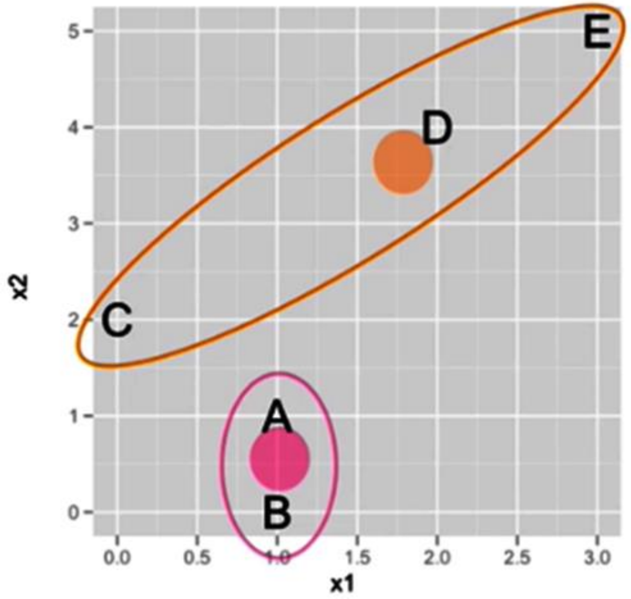
● $\bar{X}_1^1 = (1,0.5)$

● $\bar{X}_2^1 = (1.7,3.7)$

New cluster means

K-mean Clustering

Step 1.1 - Plots



● $\bar{X}_1^1 = (1,0.5)$

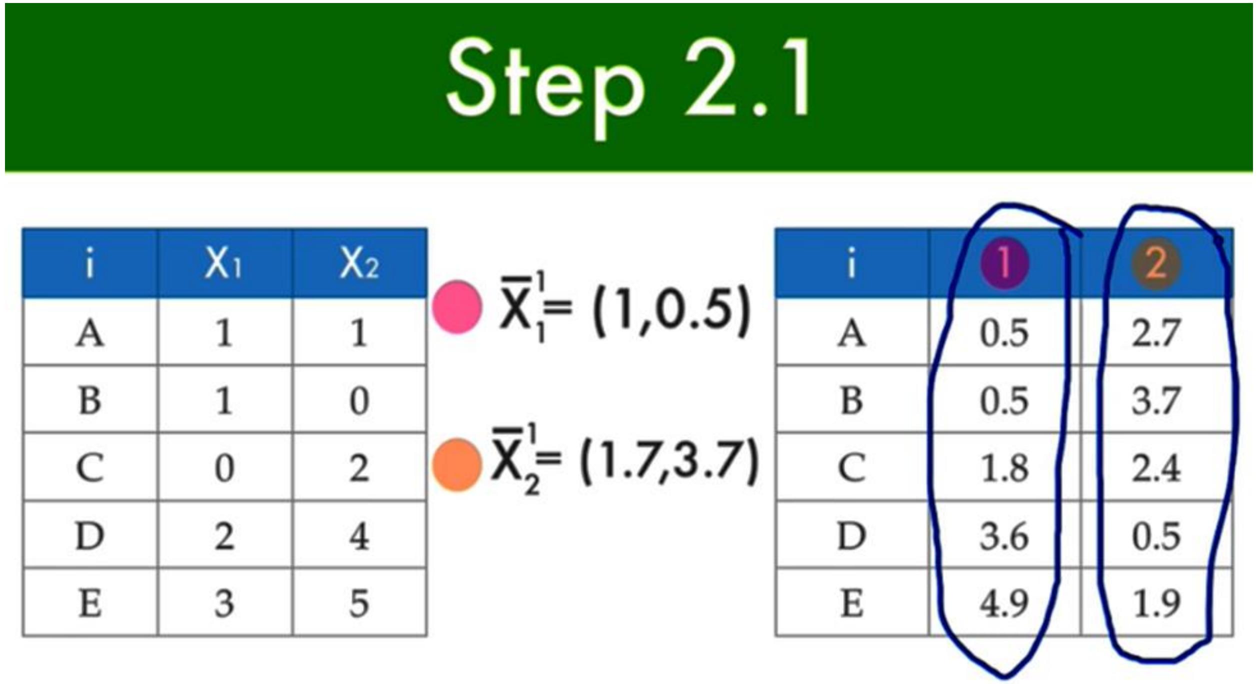
● $\bar{X}_2^1 = (1.7,3.7)$

New sketch of the clusters

K-mean Clustering

Once again calculate the distances

K-mean Clustering



Step 2.1

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

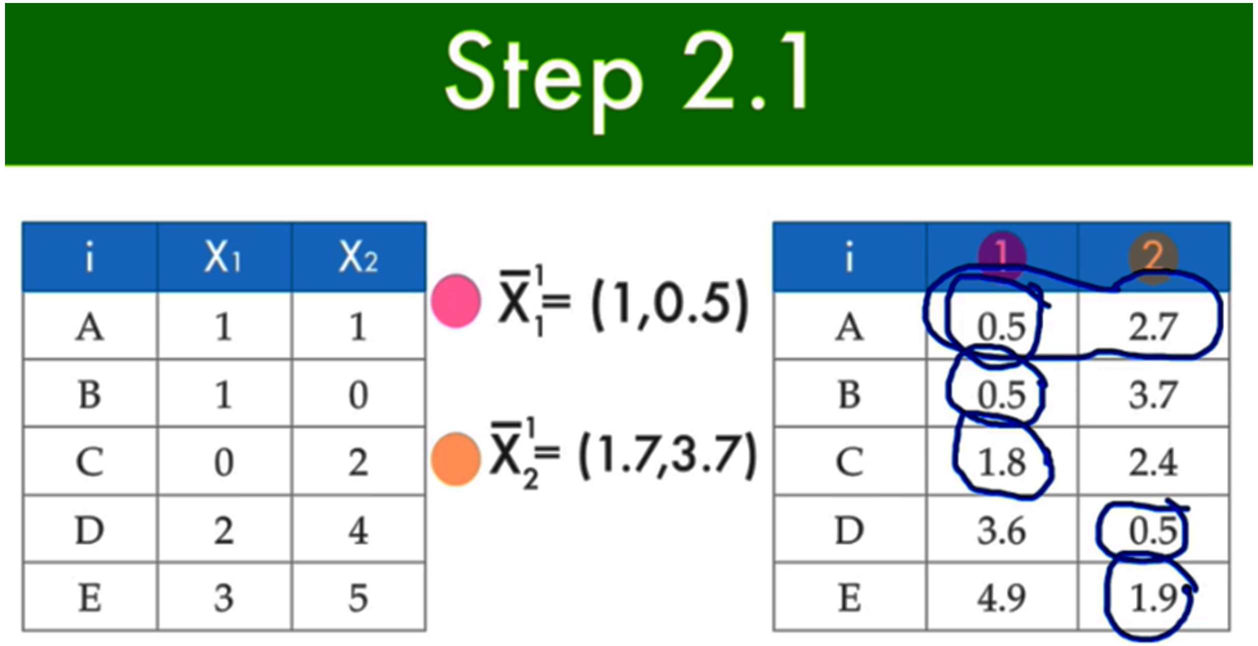
● $\bar{X}_1^1 = (1,0.5)$

● $\bar{X}_2^1 = (1.7,3.7)$

i	1	2
A	0.5	2.7
B	0.5	3.7
C	1.8	2.4
D	3.6	0.5
E	4.9	1.9

Distances in column 1 and column 2 from their respective means

K-mean Clustering



Step 2.1

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

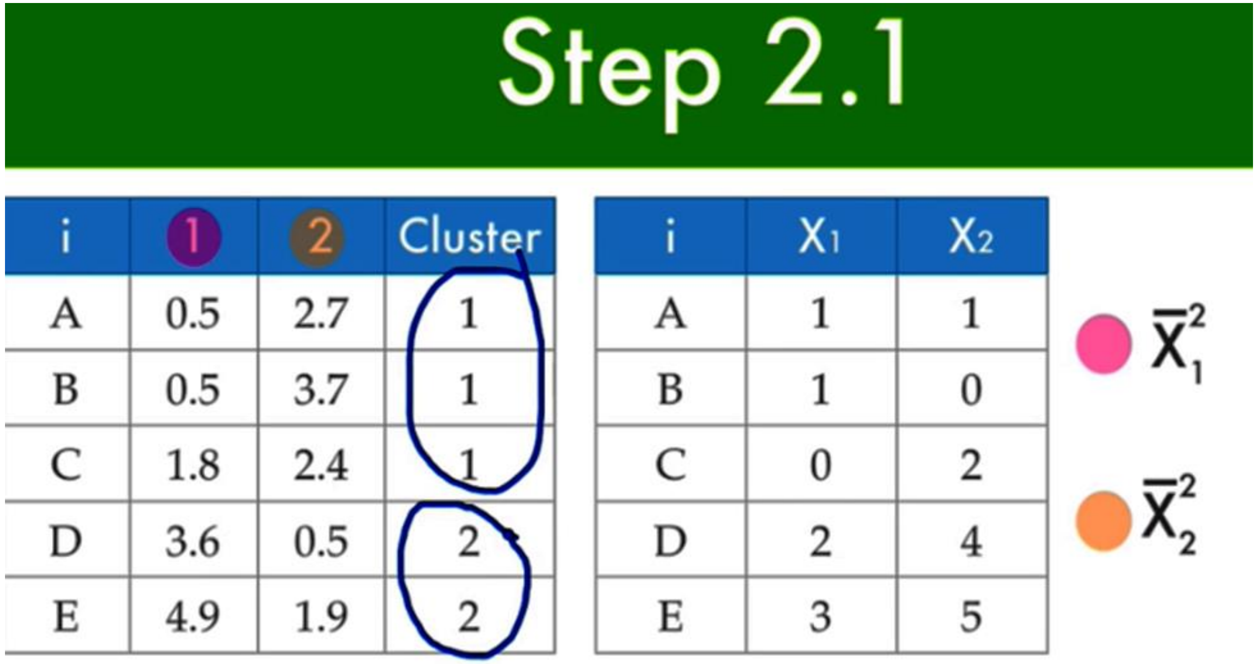
● $\bar{X}_1^1 = (1,0.5)$

● $\bar{X}_2^1 = (1.7,3.7)$

i	1	2
A	0.5	2.7
B	0.5	3.7
C	1.8	2.4
D	3.6	0.5
E	4.9	1.9

Now A, B, and C are assigned to cluster 1 and D, E to Cluster 2

K-mean Clustering



Step 2.1

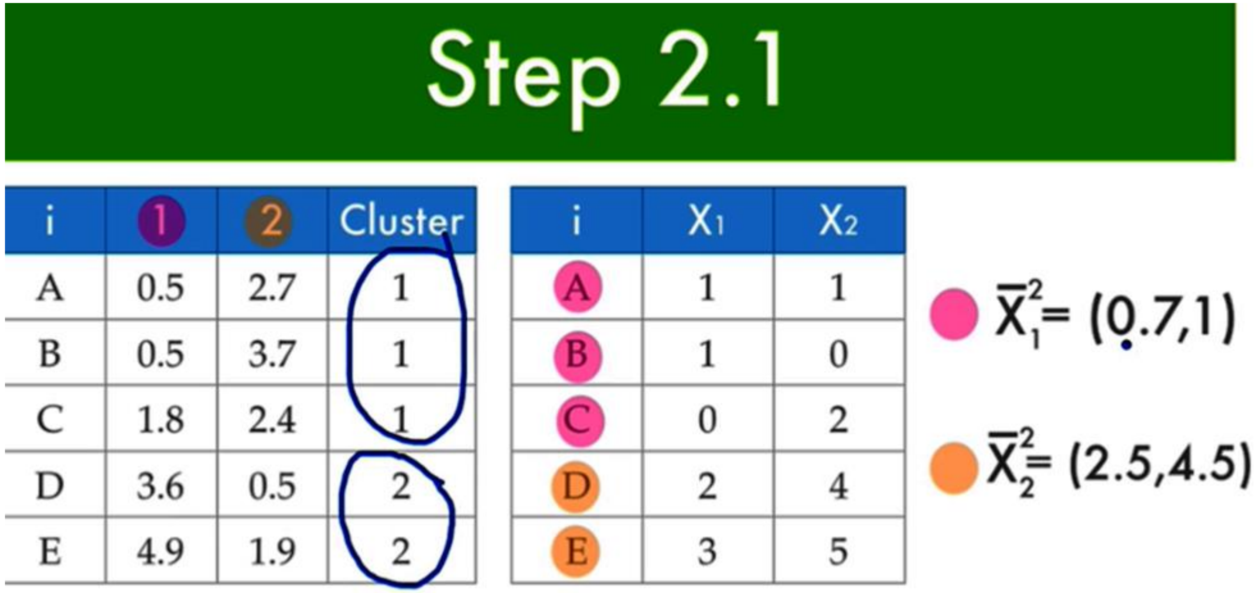
i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

● $\bar{X}_1^2$

● $\bar{X}_2^2$

K-mean Clustering



Step 2.1

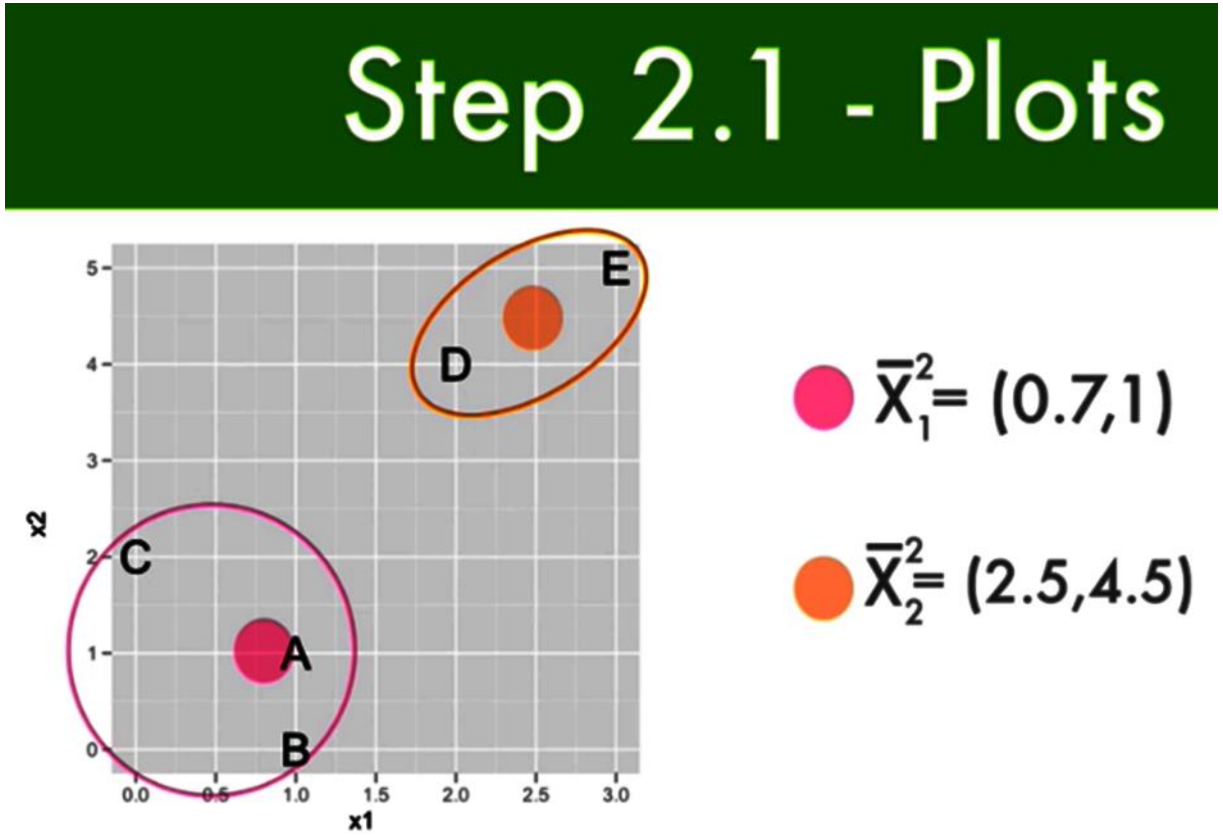
i	1	2	Cluster
A	0.5	2.7	1
B	0.5	3.7	1
C	1.8	2.4	1
D	3.6	0.5	2
E	4.9	1.9	2

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

● $\bar{X}_1^2 = (0.7,1)$

● $\bar{X}_2^2 = (2.5,4.5)$

K-mean Clustering



Step 2.1 - Plots

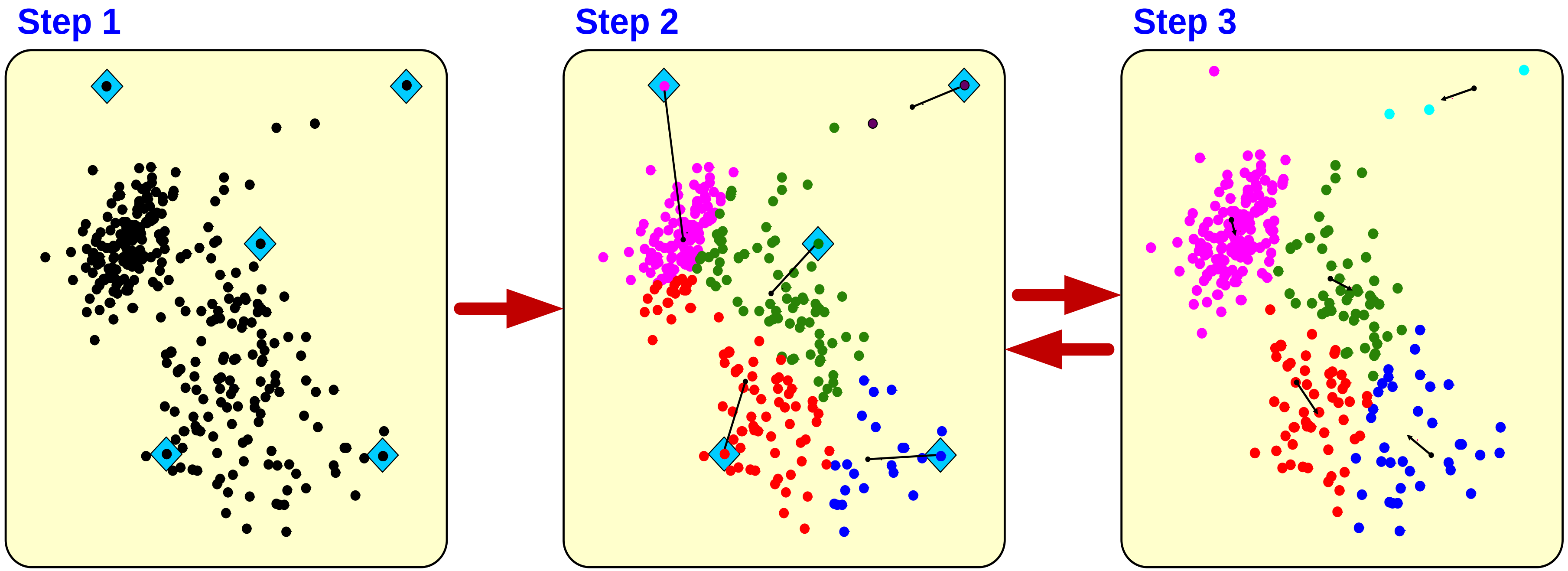
● $\overline{X}_1^2 = (0.7,1)$

● $\overline{X}_2^2 = (2.5,4.5)$

K-mean Clustering

- If we proceed in the same way there will be no change in the cluster means and the algorithm converged.

k-Means Clustering Algorithm



- Questions, comments