Clustering

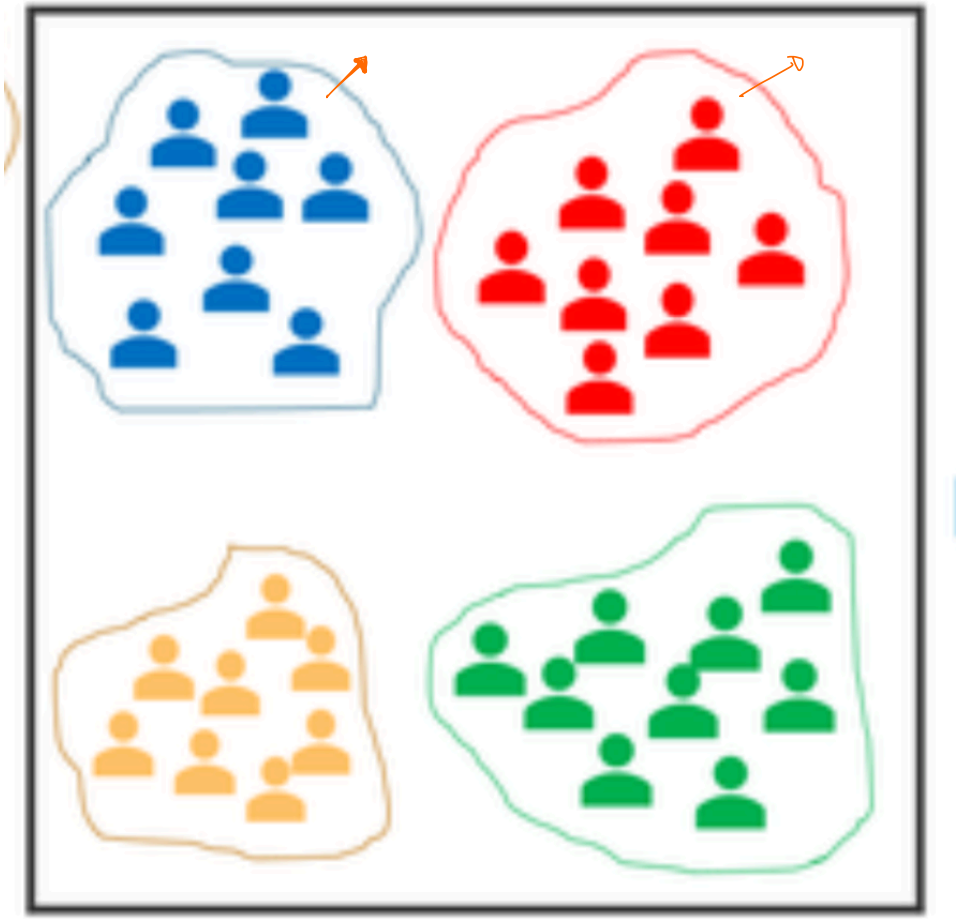
ECE 4200

What is clustering

Given a bunch of examples, group them together.

Ad targeting:

NETFLIX?



Using clustering algorithms on the customer base

What is clustering

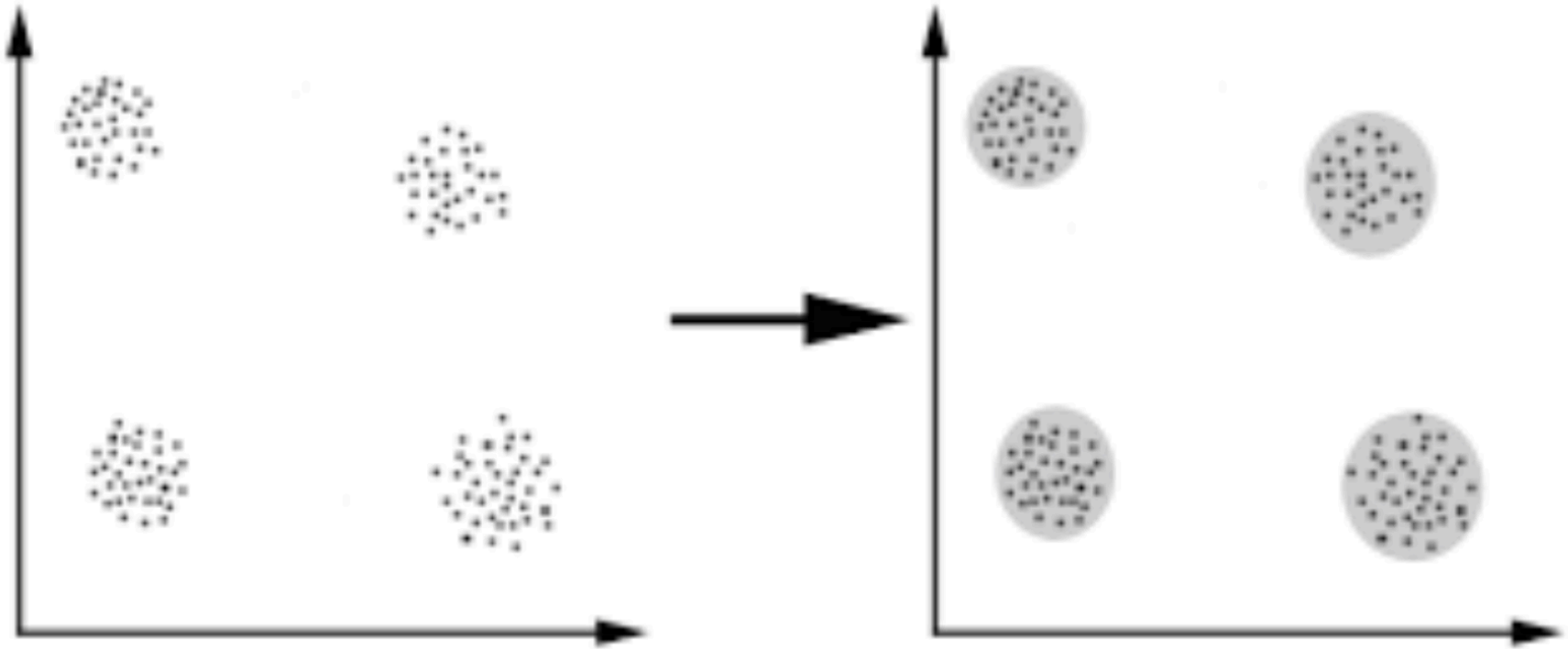
Given a bunch of examples, group them together.

Discover underlying patterns in data

Summarization

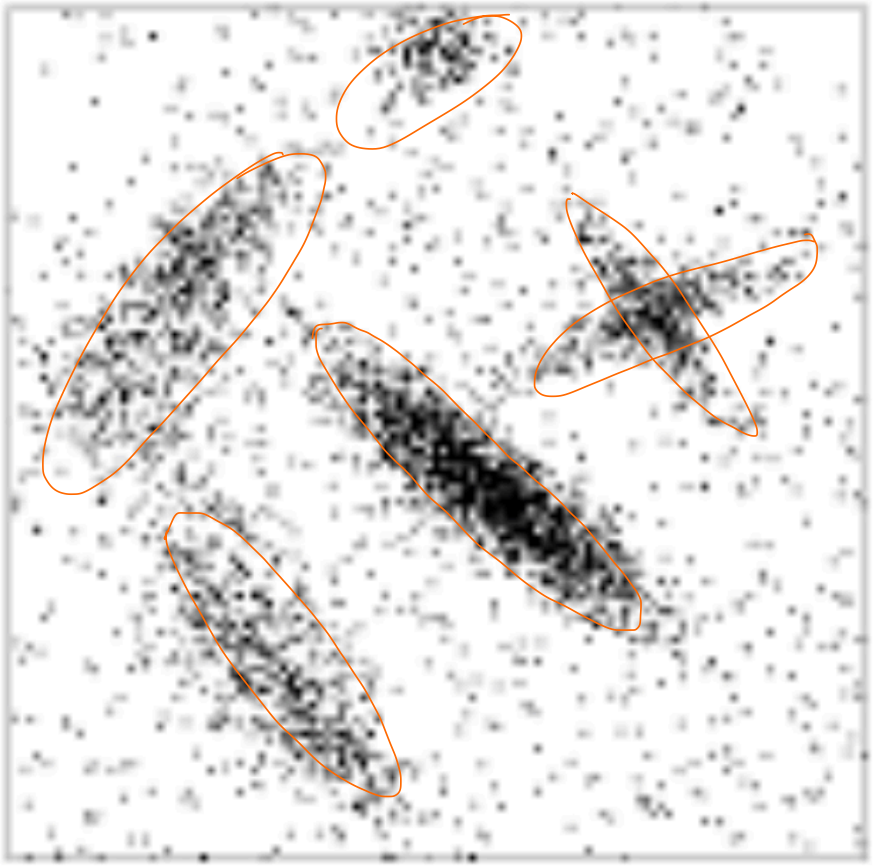
What is clustering

Given a bunch of examples, group them together.



What is clustering

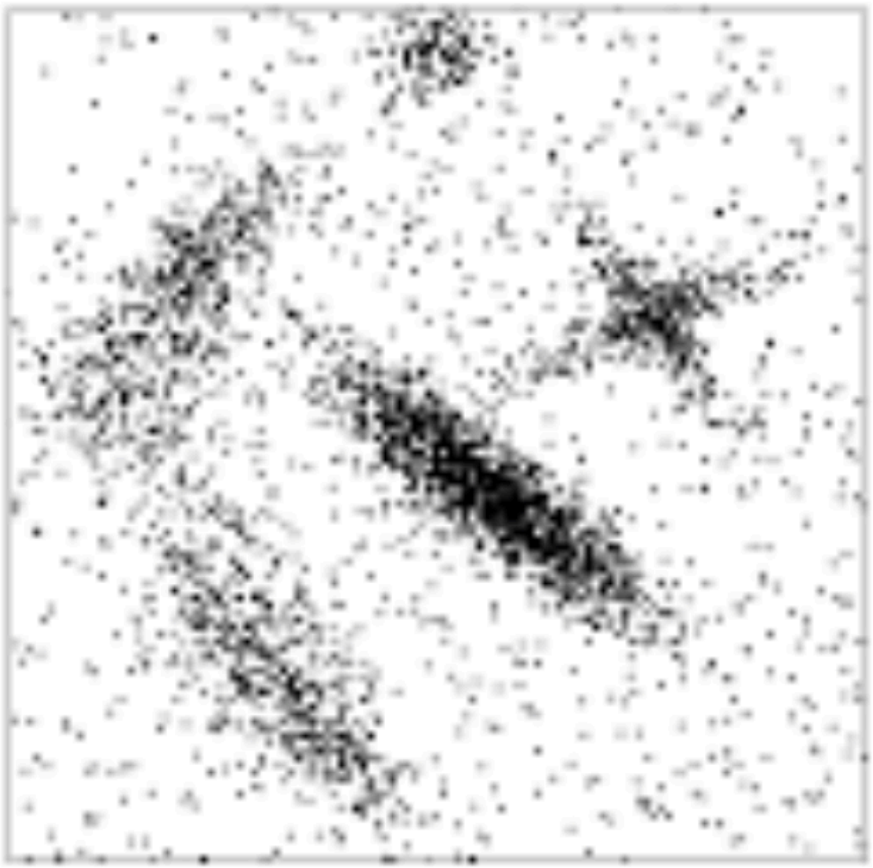
Given a bunch of examples, group them together.

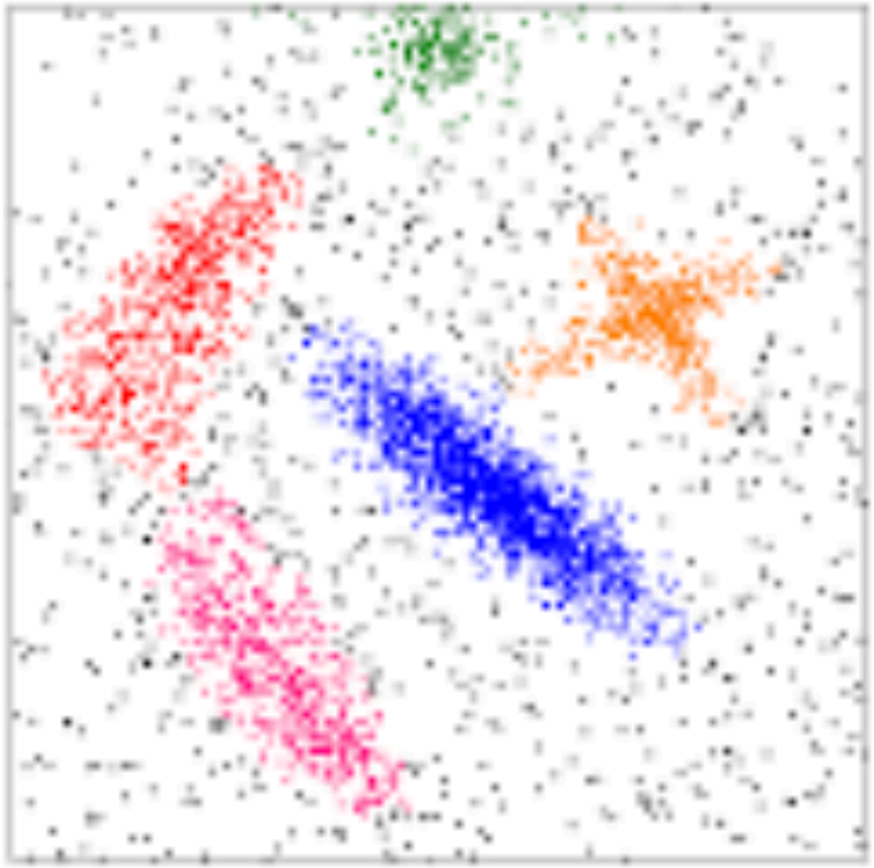


Gaussian Mixture Models (GMM).

What is clustering

Given a bunch of examples, group them together.





Clustering

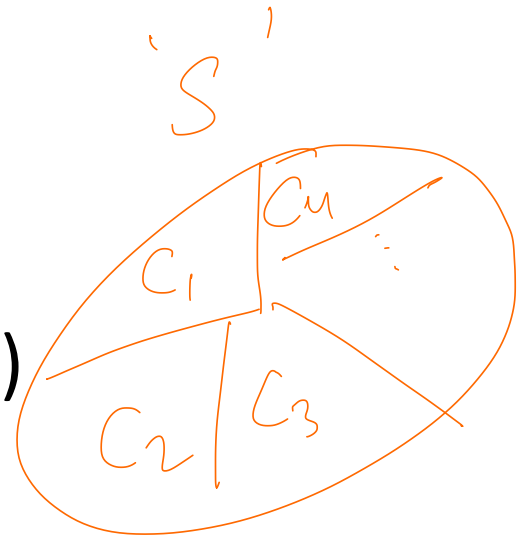
k: number of clusters (say is known)

$S = \{\bar{X}_1, \ldots, \bar{X}_n\}$ a set of n points (examples) $\bar{X}_i \in \mathbb{R}^d$

Goal: Partition S into k groups

How to do that?

Clustering



k: number of clusters (say is known)

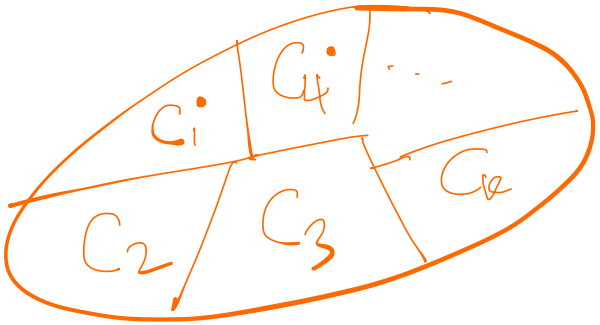
$S = \{\bar{X}_1, \ldots, \bar{X}_n\}$ a set of n points (examples)

Goal: Output a partition $C_1, \ldots, C_k$ of S

HOW?

* points inside the cluster are close to each other
* clusters are far from each other.

k-means clustering



Input: $k, S = \{\bar{X}_1, \ldots, \bar{X}_n\}$

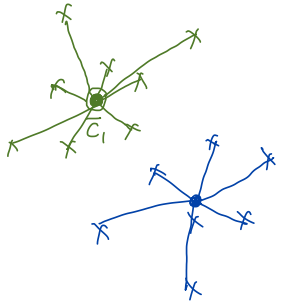
$\in \mathbb{R}^d$ (mean of the cluster)

Output: $C_1, \ldots, C_k$, and $\bar{c}_1, \ldots, \bar{c}_k$ (means) to **minimize**:

$$\sum_{j=1}^{k} \sum_{\bar{X}_i \in C_j} \|\bar{X}_i - \bar{c}_j\|_2^2$$

goes over clusters $C_1 \ldots C_k$.

$k = 2$



Within a cluster the points should be close to each other!

k-means clustering

Input: $k, S = \{\bar{X}_1, \ldots, \bar{X}_n\}$

Output: $C_1, \ldots, C_k$, and $\bar{c}_1, \ldots, \bar{c}_k$ (means) to **minimize**:

error ⟵ **minimize** $\sum_{j=1}^{k} \sum_{\bar{X}_i \in C_j} \|\bar{X}_i - \bar{c}_j\|_2^2$

$C_1 \cdots C_k$, Take a cluster 'C' ⟶ $\{\bar{X}_1, \ldots, \bar{X}_{|C|}\}$

$\bar{c}_1 \quad \bar{c}_k$

$\bar{c}$ ⟶ minimize $\sum_{\bar{X}_i \in C} \|\bar{X}_i - \bar{c}\|_2^2$

Observation 1

4, 4, 5, 4, 6, 6, 6, 5, 4, 6, 6 → prove

'x' $(4-x)^2 \times 4 + (5-x)^2 \cdot 2 + (6-x)^2 \cdot 5$

$\rightarrow \frac{4\times4+5\times2+6\times5}{11}$

Given a cluster $C = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{|C|}\}$, how to choose $\bar{c}$?

$$\textbf{minimize} \sum_{\bar{X}_i \in C} \|\bar{X}_i - \bar{c}\|_2^2$$

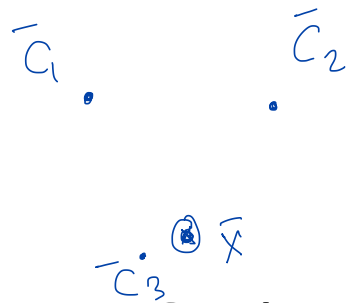
Answer: The **best** $\bar{c}$ is

$$\frac{1}{|C|} \sum_{1 \le i \le |C|} \bar{X}_i$$

Once we have $\{C_1, C_2, \ldots, C_k\}$ we obtain best $\{\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k\}$

$$\bar{c}_j = \frac{1}{|C_j|} \cdot \sum_{\bar{X}_i \in C_j} \bar{X}_i$$

Observation 2



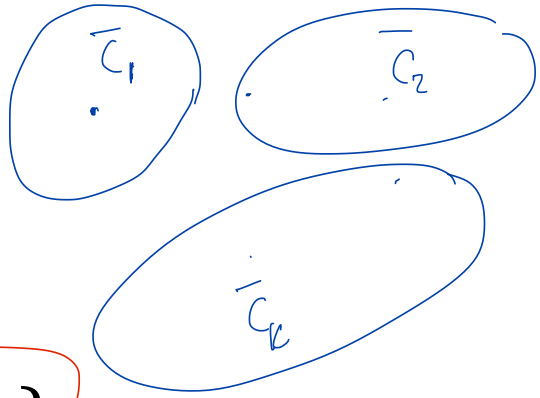
Given $\{\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k\}$, which cluster C_j should $\bar{X}_i$ be assigned?

Answer: The **best possible** cluster is

$$j^* = \arg\min_j \|\bar{X}_i - \bar{c}_j\|$$

Once we have $\{\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_k\}$ we obtain best $\{C_1, C_2, \ldots, C_k\}$

k-means algorithm



Initialize:

Choose k data points as initial $\{\bar{c}_1, \ldots, \bar{c}_k\}$

Repeat:

- Assign each data point $\bar{X}_i$ to the closest $\bar{c}_j$ to get clusters
- Change $\{\bar{c}_1, \ldots, \bar{c}_k\}$ to the **new** cluster means

Until no cluster assignment changes

possible clusterings $< k^n$ ($k!$)

$\boldsymbol{How\ good\ is\ this\ algorithm?}$

- Does the k means algorithm solve k means clustering?

NO. In fact k means clustering is NP HARD!

- It gives some approximate solution (not very bad)
- Widely used in practice

Issues/Questions

Runtime:

Convergence: → Yes!!!

What is the k to use? →

take 7 points in 2-d.

run k = 2

k means++: smart initialization of centers to get better approximation.

Look at the paper ***k* means ++** Ziteng

Vasilvitskii, ...

Agglomerative clustering

Hierarchical clustering

- For each value of k, we obtain a k clustering

Algorithm:

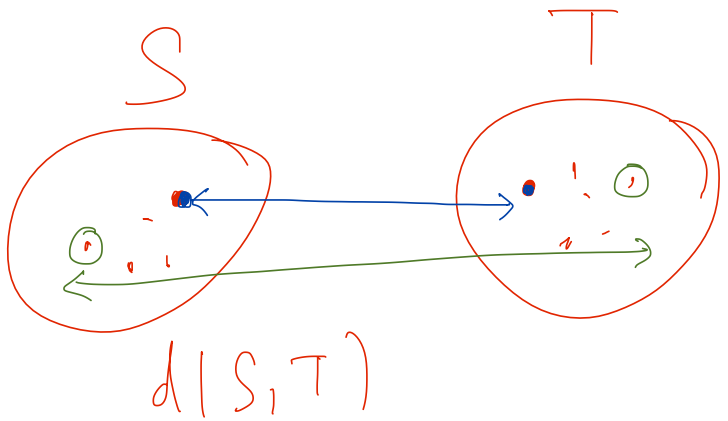
Initialize:

Each point is a cluster of its own

Repeat:

Merge **two closest** clusters into one cluster

Until we have one cluster



Closest?

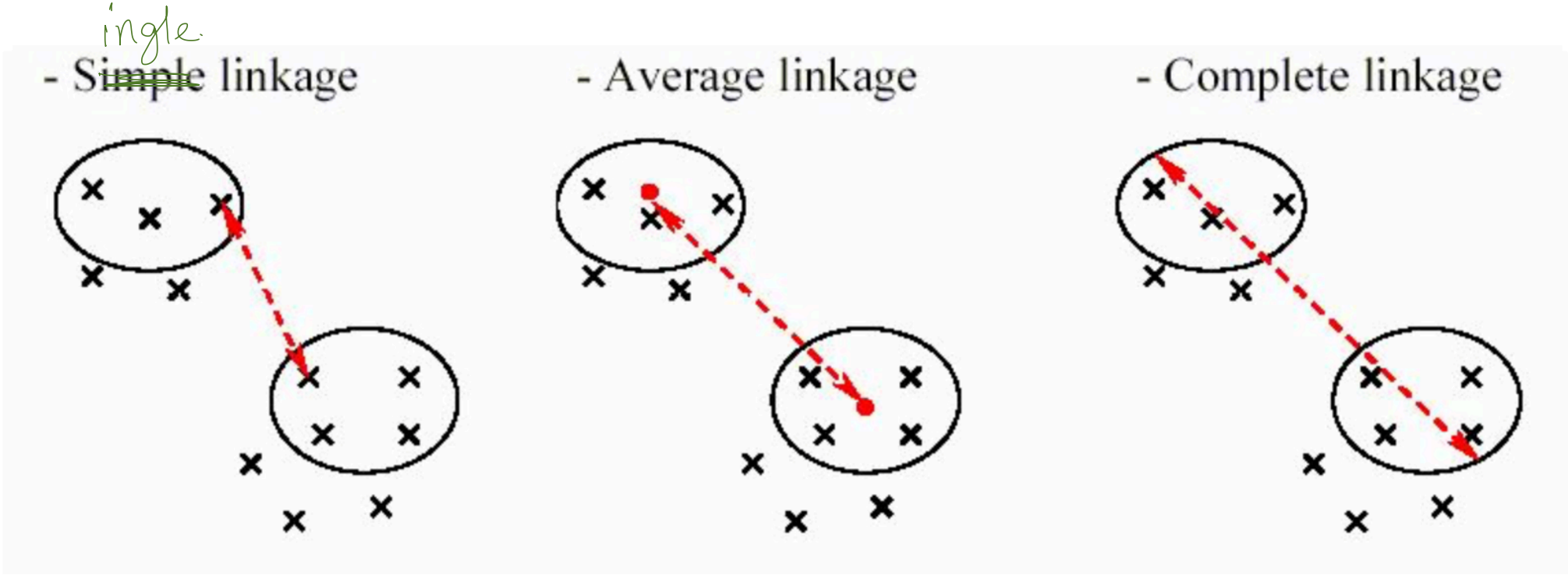
Merge **two closest** clusters into one cluster

What does closest mean for two clusters?

- Single linkage (closest points)
- Average linkage
- Complete linkage (farthest points)

Closest?

- Single linkage (closest points)
- Average linkage
- Complete linkage (farthest points)



Soft clustering

Assign probabilities to the instances being in clusters

collaborative filtering.

Recommendation Systems