CM50264: Machine Learning 1

Unsupervised learning

Kwang In Kim
k.kim@bath.ac.uk

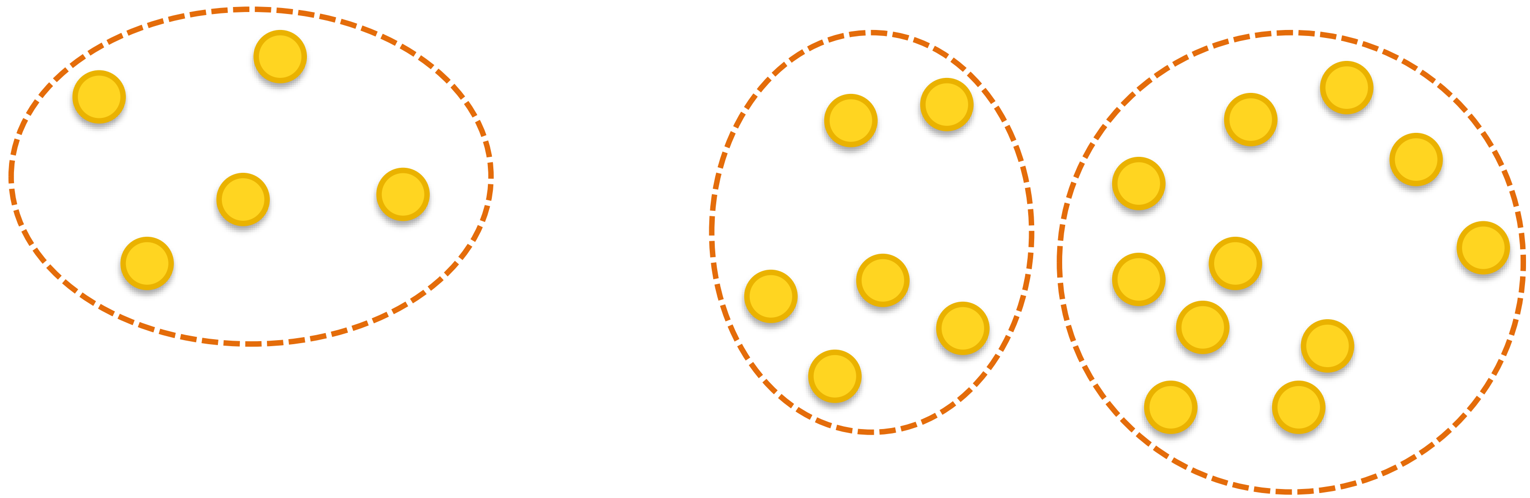
Supervised / unsupervised learning

- Supervised learning
  - learn from labelled examples:
    $$D = \{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^N, y^N)\}$$
    Pairs of input and the corresponding desired output.
- Unsupervised learning
  - learn from unlabelled examples:
    $$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$$
    Input data only; no desired outputs.
  - Detect underlying structure in data.
  - E.g., clustering and dimensionality reduction.

Clustering

$$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \boldsymbol{X}$$

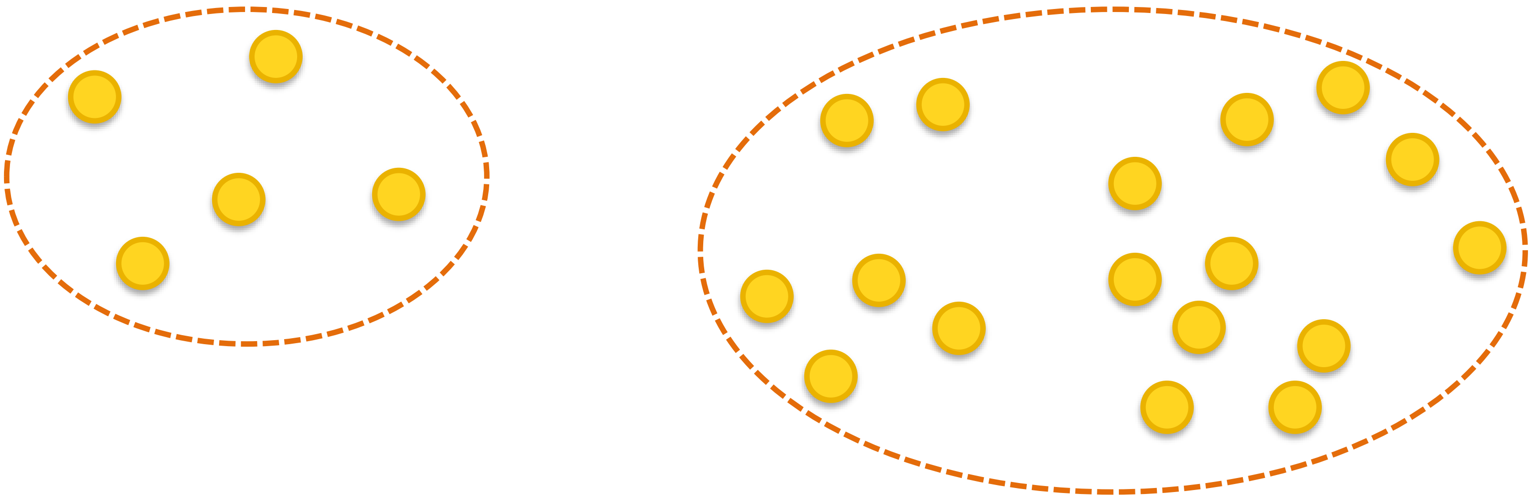
- Group together similar data instances.
- E.g., $\boldsymbol{X} \subset \mathbf{R}^2$



Clustering

$$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \boldsymbol{X}$$

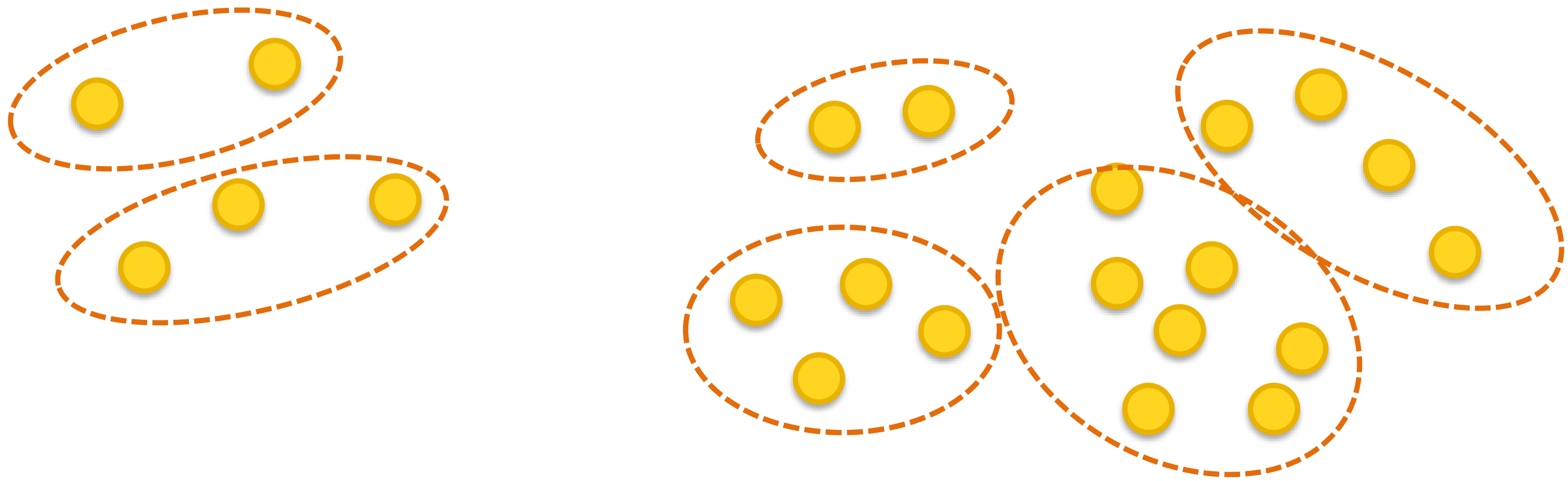
- Group together similar data instances.
- E.g., $\boldsymbol{X} \subset \mathbf{R}^2$



Clustering

$$D = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\} \subset \boldsymbol{X}$$

- Group together similar data instances.
- E.g., $\boldsymbol{X} \subset \mathbf{R}^2$



Clustering applications

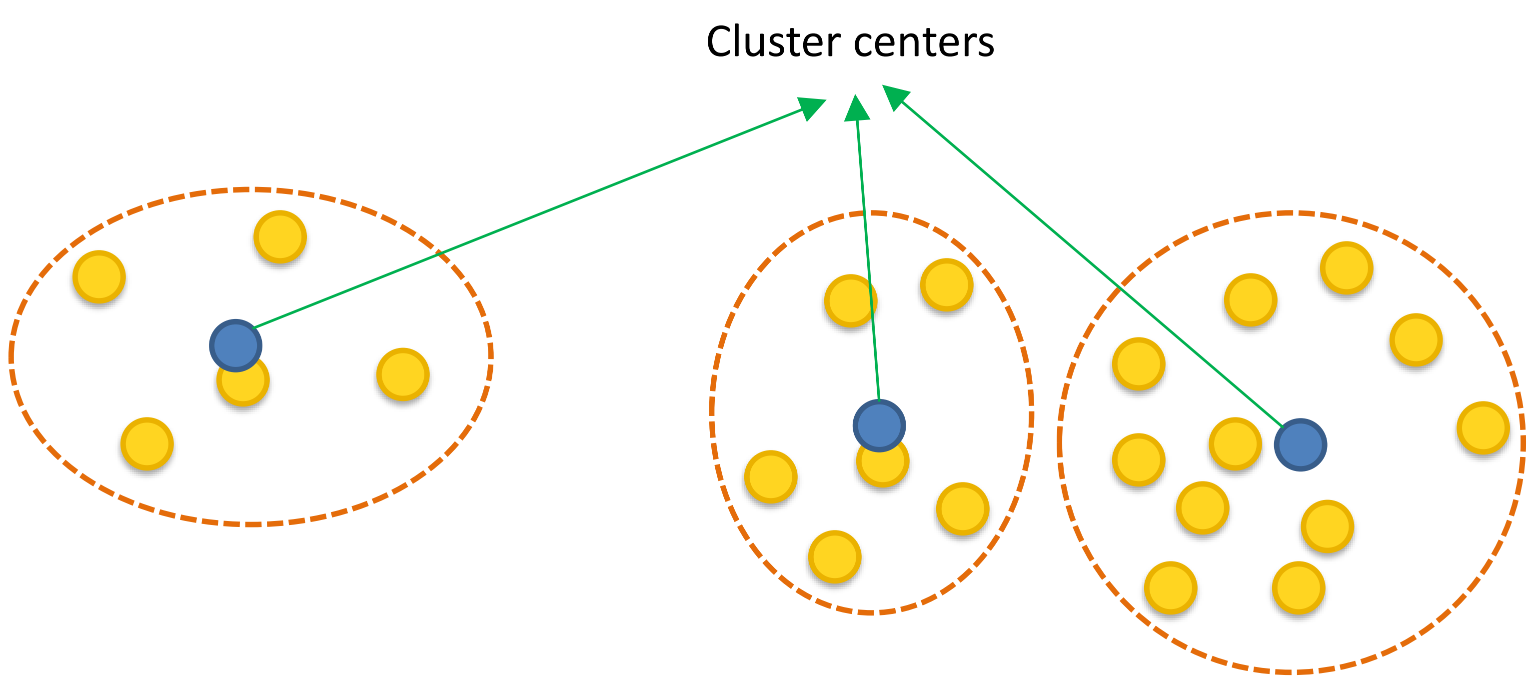
- Marketing: help to discover distinct groups in customer bases, and develop targeted marketing programs.
- Social network analysis: recognize communities within large groups of people.
- Animal ecology: discover and compare communities of organisms.
- Gene sequence analysis: group homologous sequences into gene families.

Image segmentation

- N (# data points): # pixels.
- Feature $\mathbf{x}^i$: color values (R,G,B) + row and column values of i-th pixel.

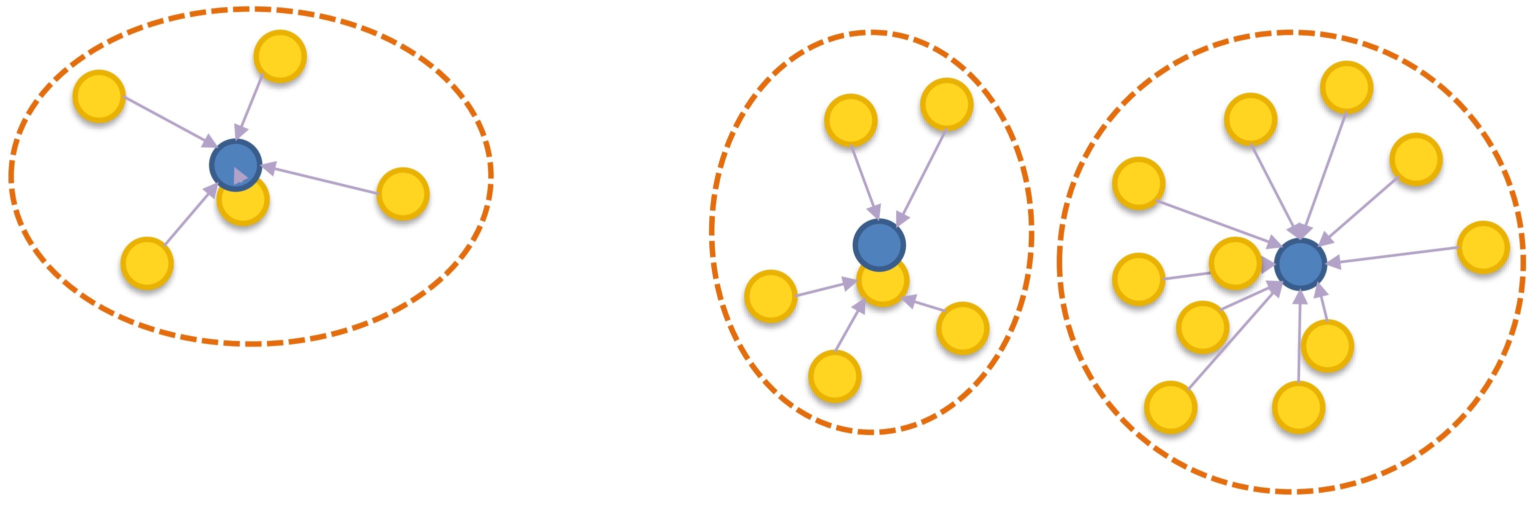

Vector quantization

- Calculate the mean vector (center) for each cluster.



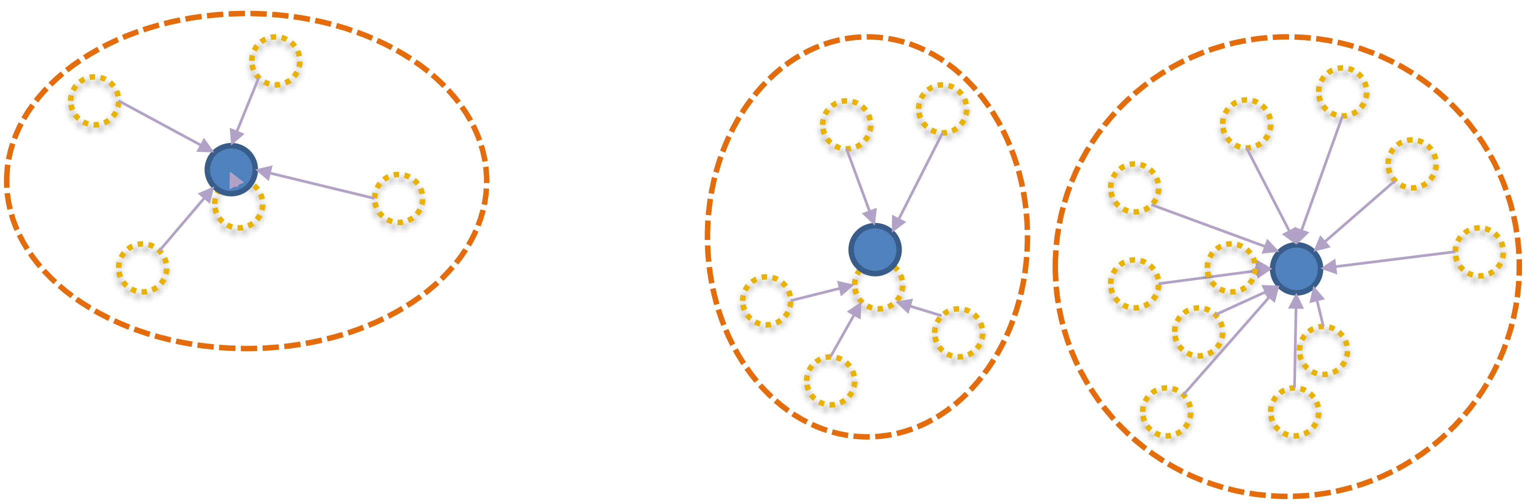
Vector quantization

- Calculate the mean vector (center) for each cluster.
- Replace all elements in each cluster by their respective center.



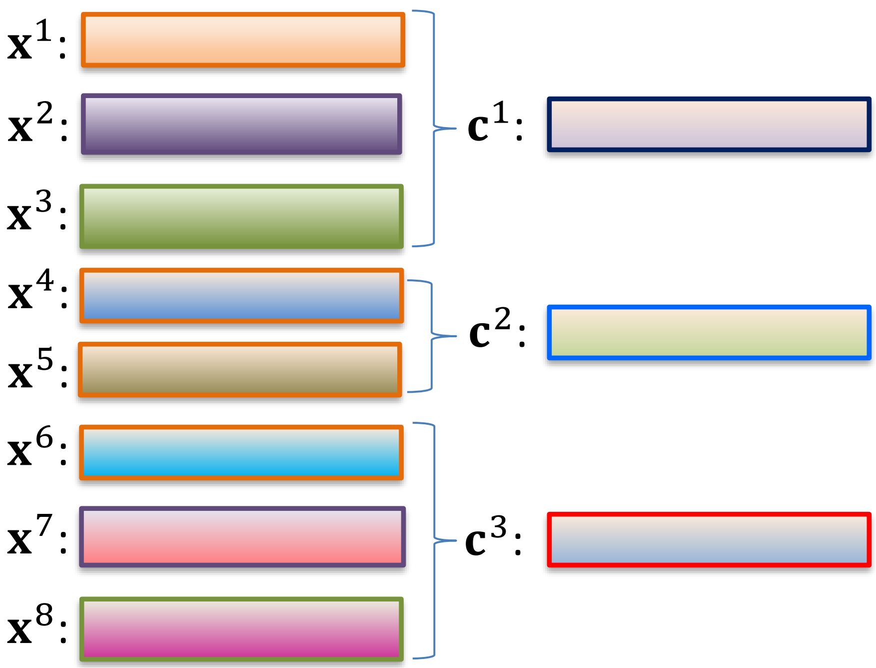
Vector quantization

- Calculate the mean vector (center) for each cluster.
- Replace all elements in each cluster by their respective center.



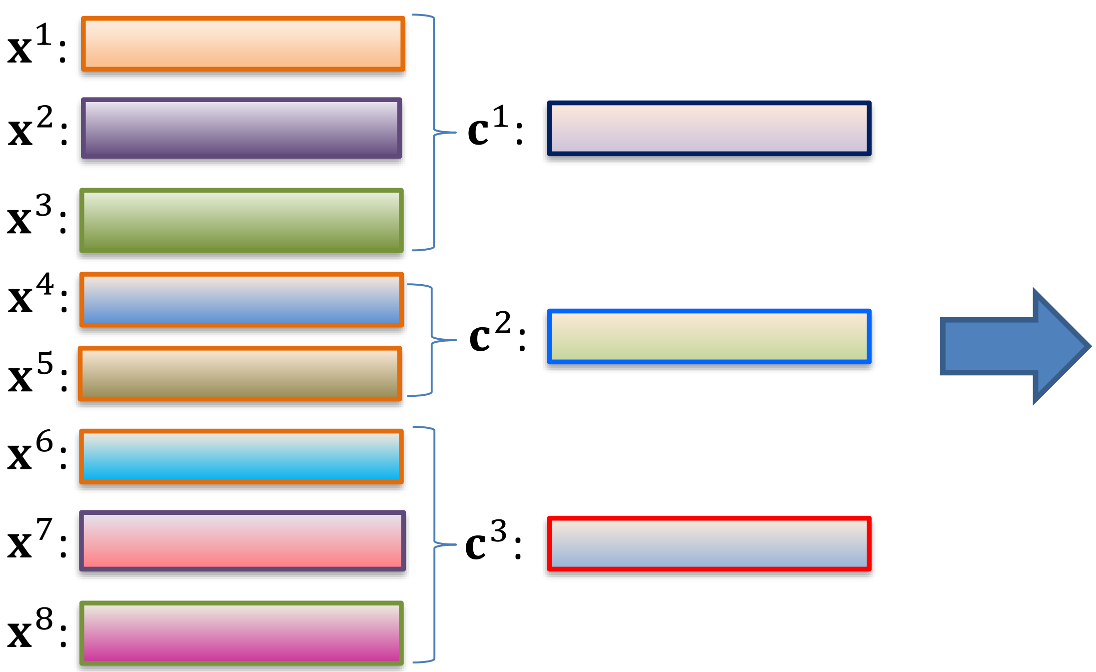
Vector quantization

Data compression: instead of strong all data points, we store the cluster centers plus cluster index for each data entry.



Vector quantization

Data compression: instead of strong all data points, we store the cluster centers plus cluster index for each data entry.



$\mathbf{x}^1$: $I(\mathbf{c}^1)$

$\mathbf{x}^2$: $I(\mathbf{c}^1)$

$\mathbf{x}^3$: $I(\mathbf{c}^1)$

$\mathbf{x}^4$: $I(\mathbf{c}^2)$

$\mathbf{x}^5$: $I(\mathbf{c}^2)$

$\mathbf{x}^6$: $I(\mathbf{c}^3)$

$\mathbf{x}^7$: $I(\mathbf{c}^3)$

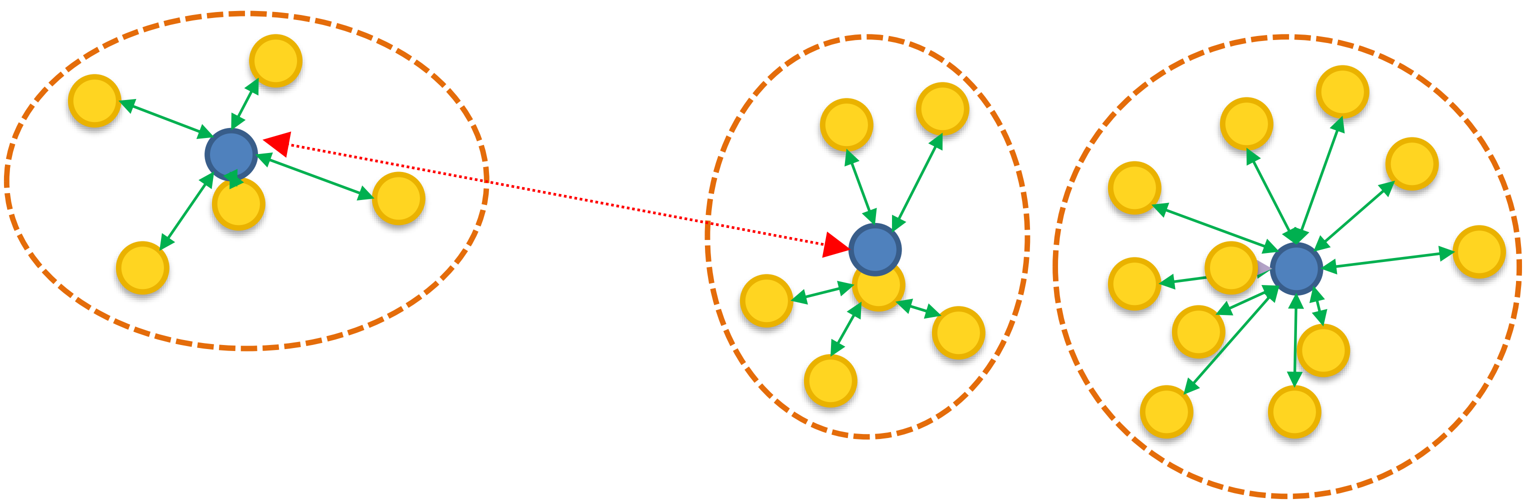
$\mathbf{x}^8$: $I(\mathbf{c}^3)$

What is good clustering?

- A good clustering algorithm will produce
  - Low intra-cluster distances.
  - High inter-cluster distances.
- The quality of a clustering depends on the distance measure, e.g., Euclidean distance.
- Objective evaluation is challenging: done by human / expert inspection.

What is good clustering?

- A good clustering algorithm will produce
  - Low intra-cluster distances.
  - High inter-cluster distances.
- Finding such a clustering is NP-hard.

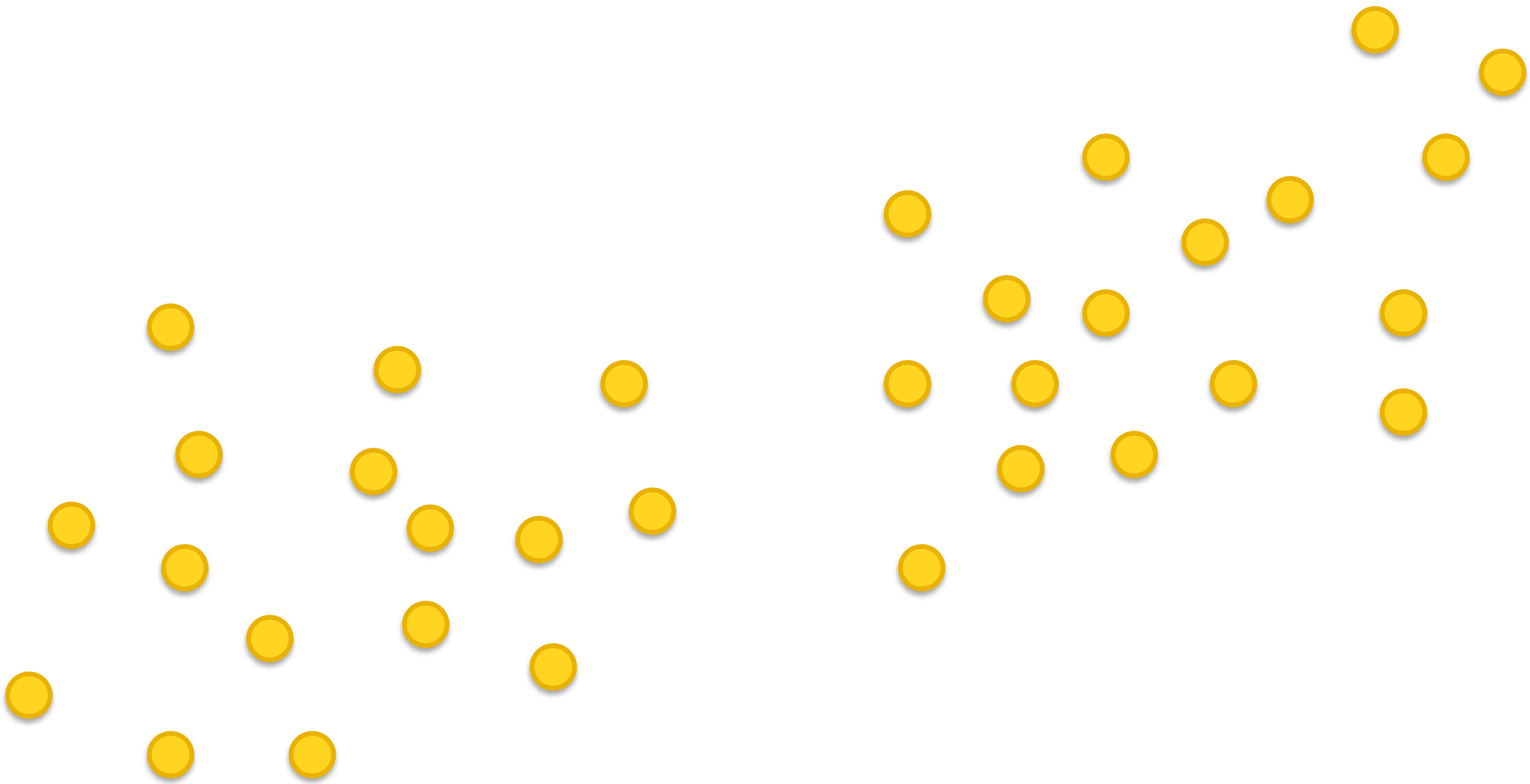


K-means clustering

An iterative clustering algorithm

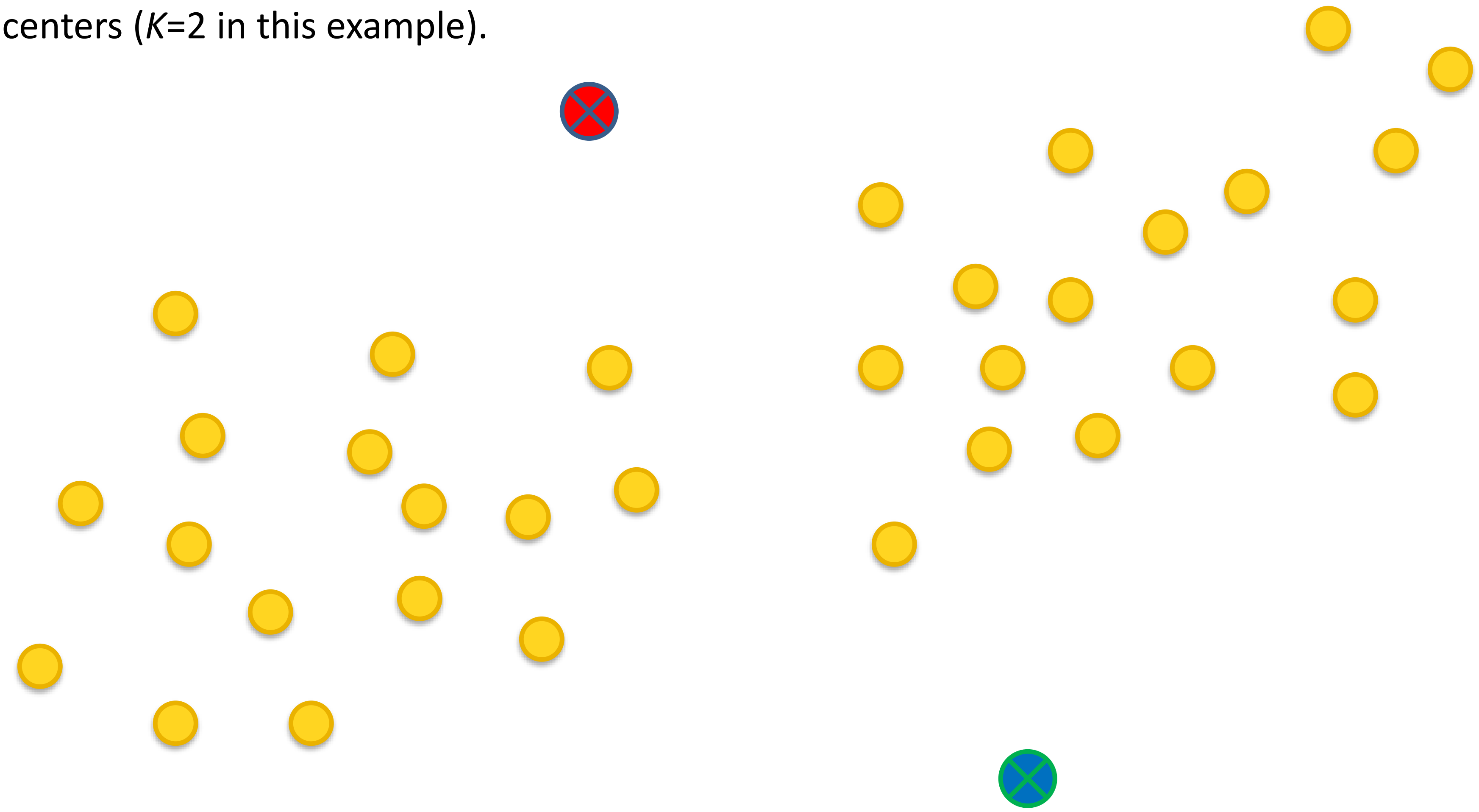
1. Initialize: pick *K* random points as cluster centers.
2. Iterate
   - Assign data points to closest cluster center.
   - Change the cluster center to the average of its assigned points.
3. Stop when no point assignments change.

K-means clustering example



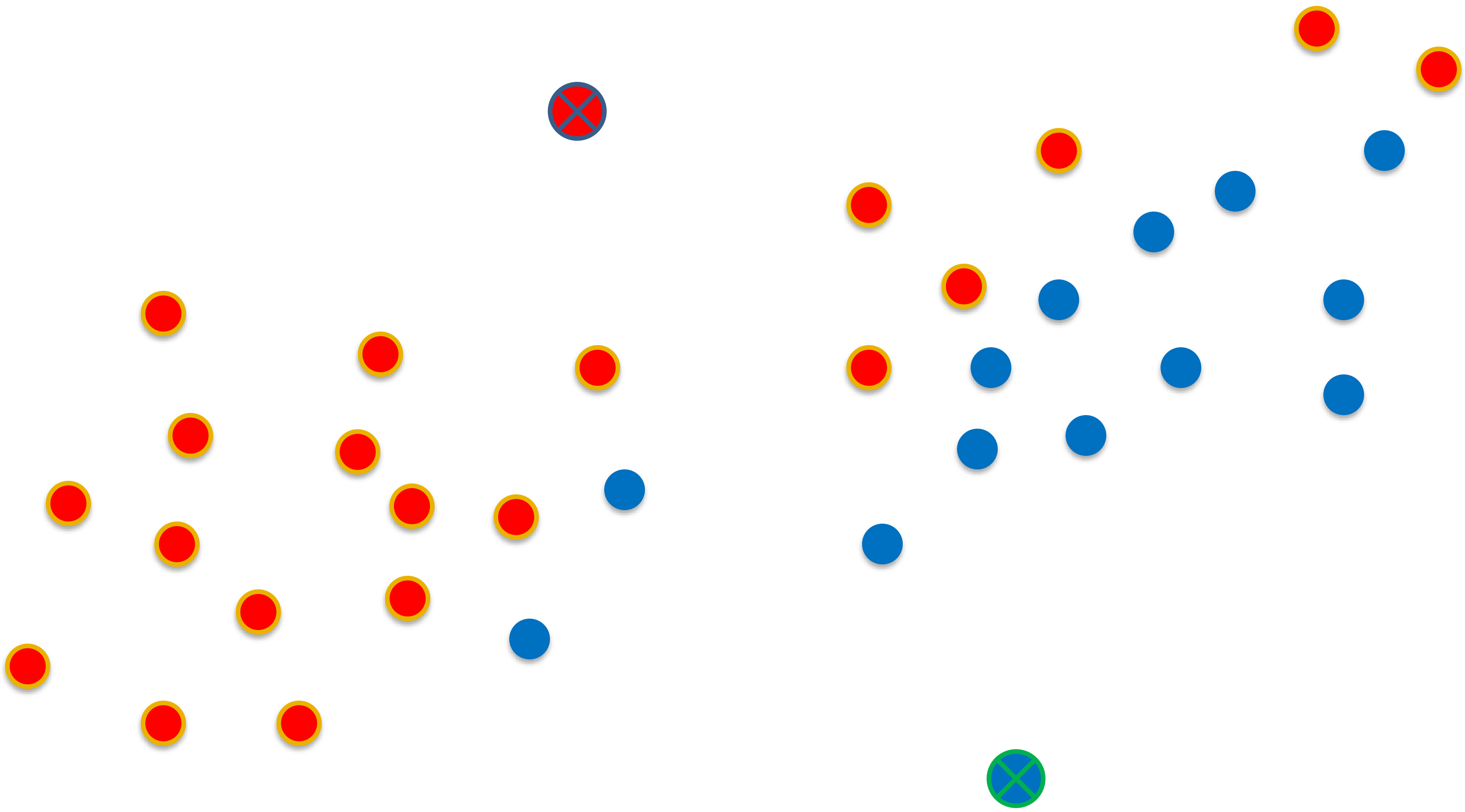
K-means clustering example

Initialize: pick *K* random points as cluster centers (*K*=2 in this example).



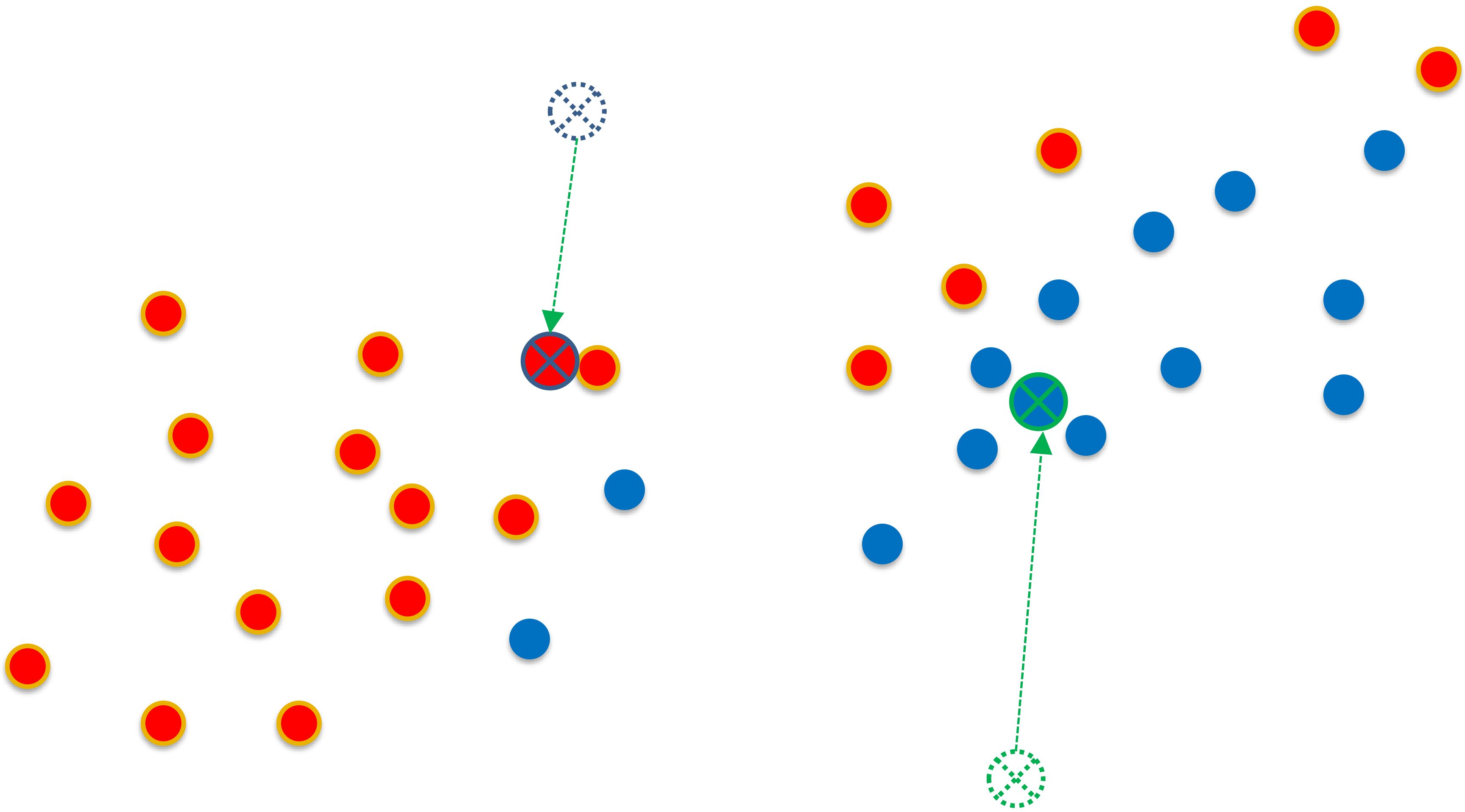
K-means clustering example

Assign data points to closest cluster center.



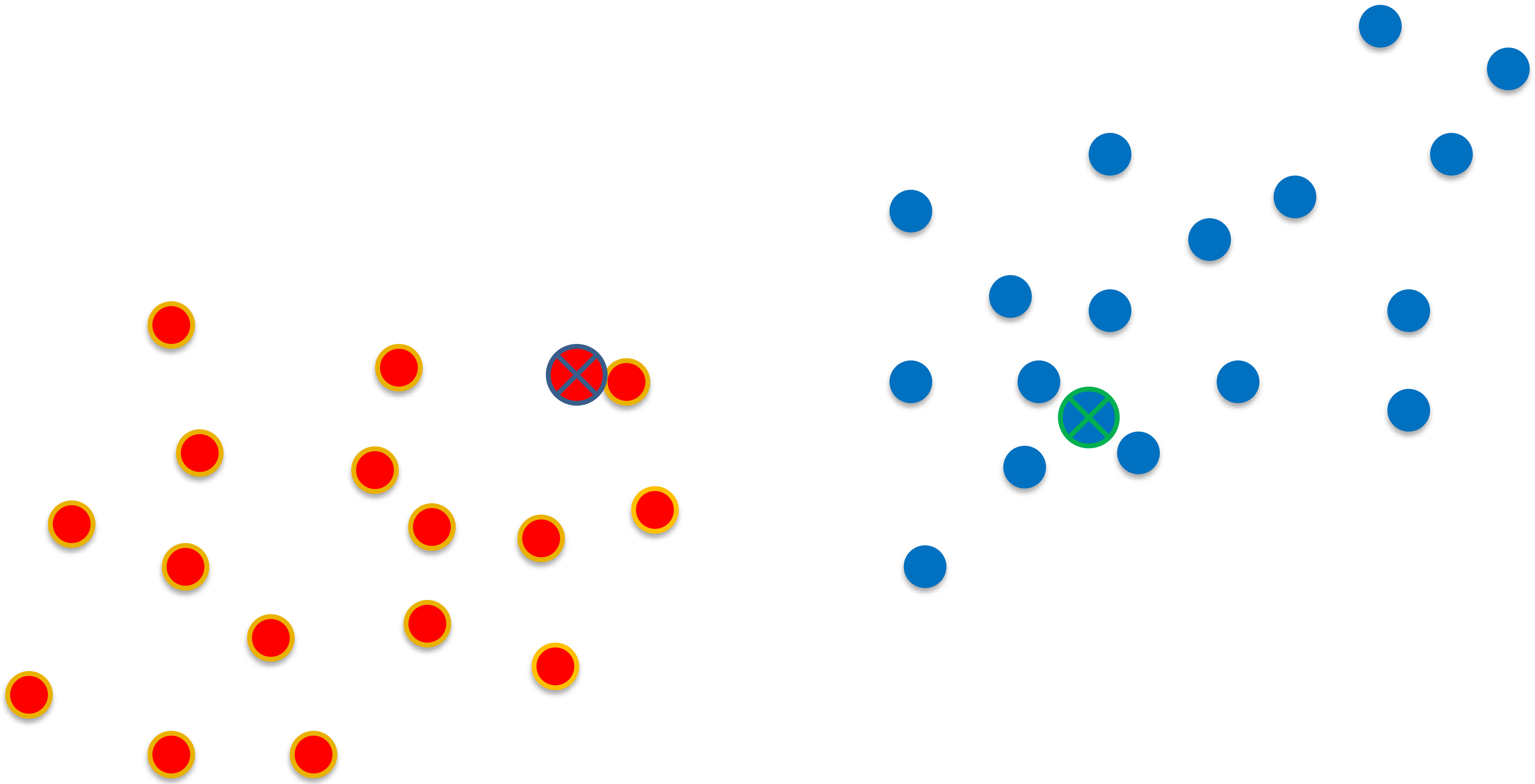
K-means clustering example

Change each cluster center to the average of its assigned points.



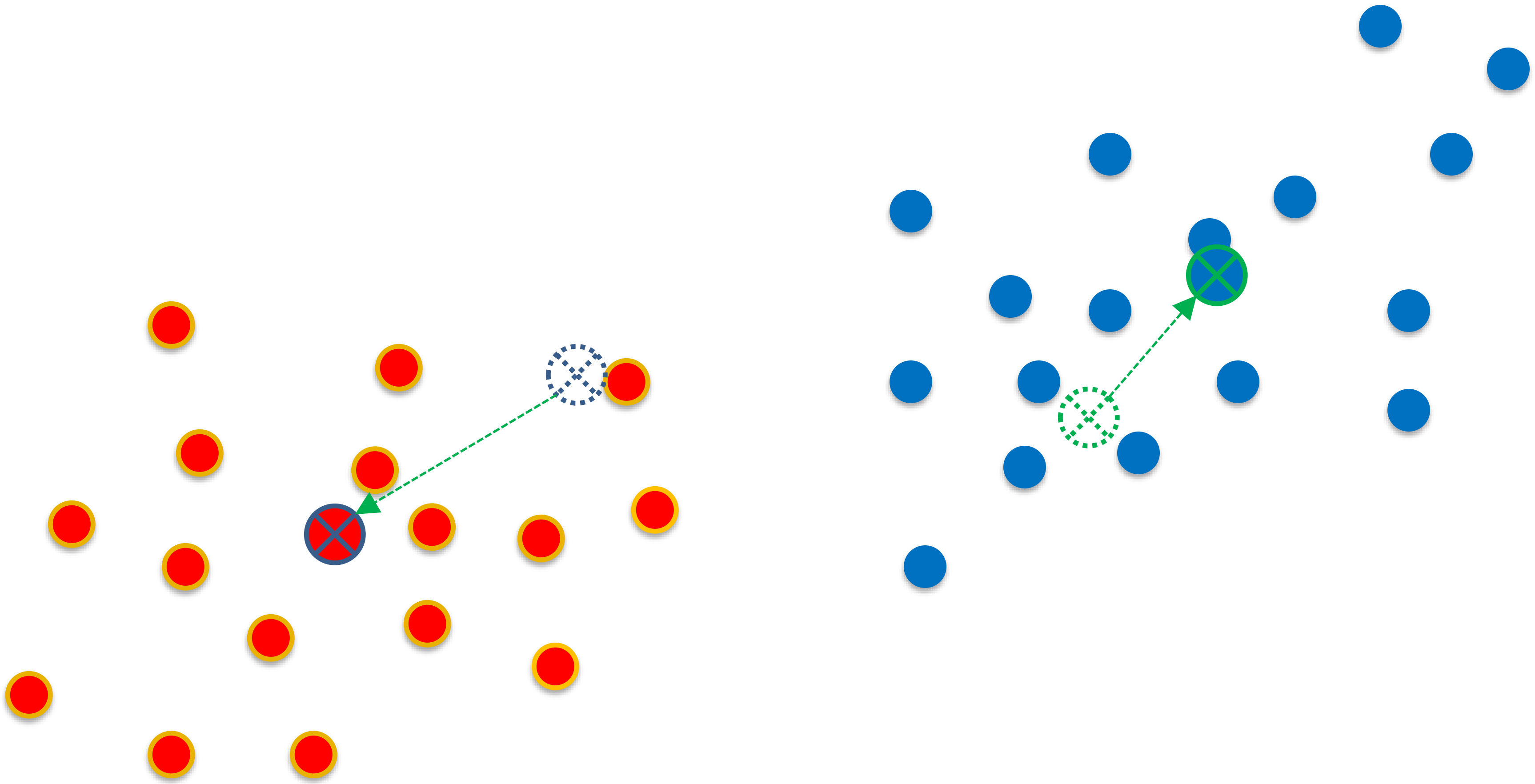
K-means clustering example

Assign data points to closest cluster center.



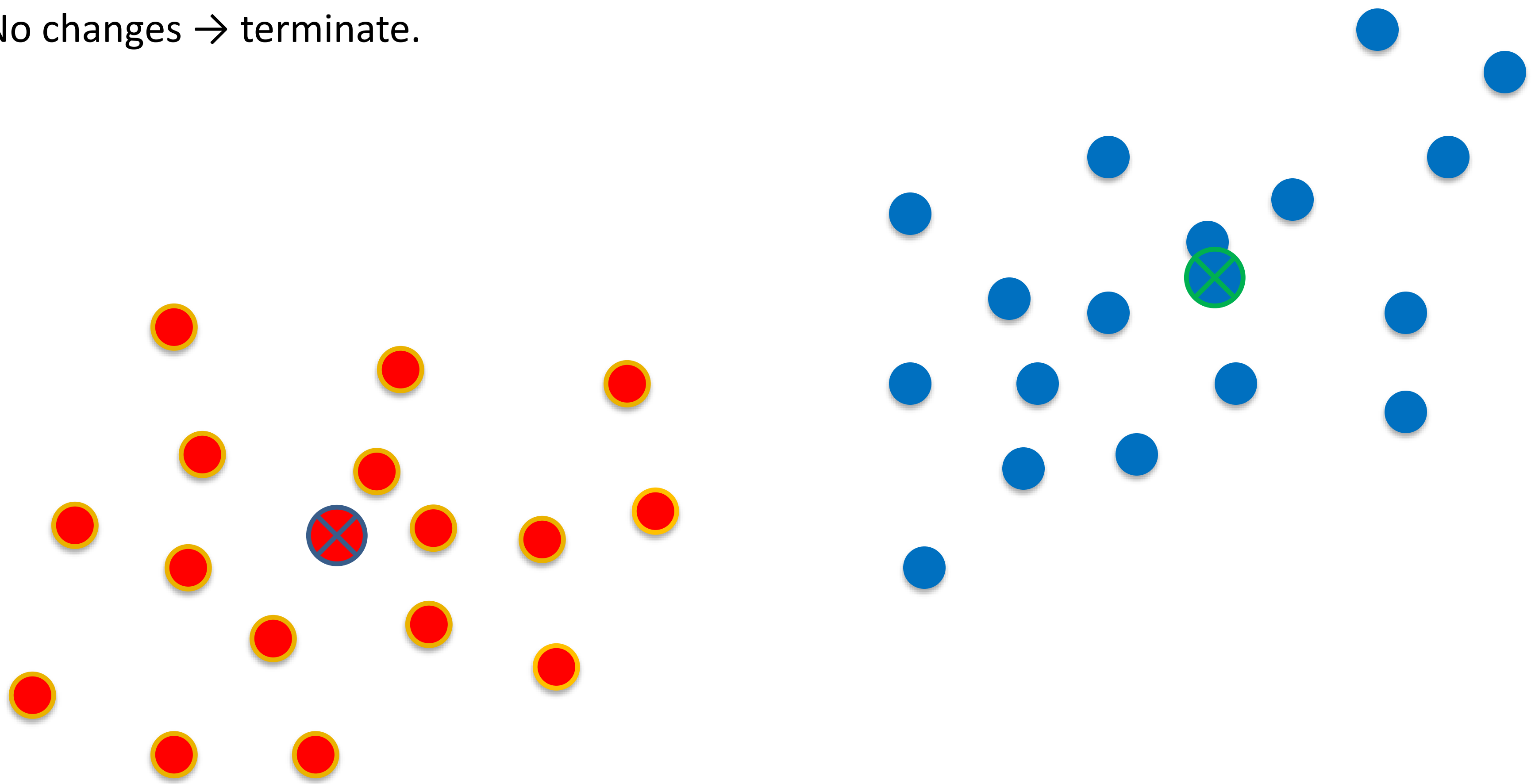
K-means clustering example

Change the cluster center to the average of its assigned points.



K-means clustering example

Assign data points to closest cluster center:
No changes → terminate.

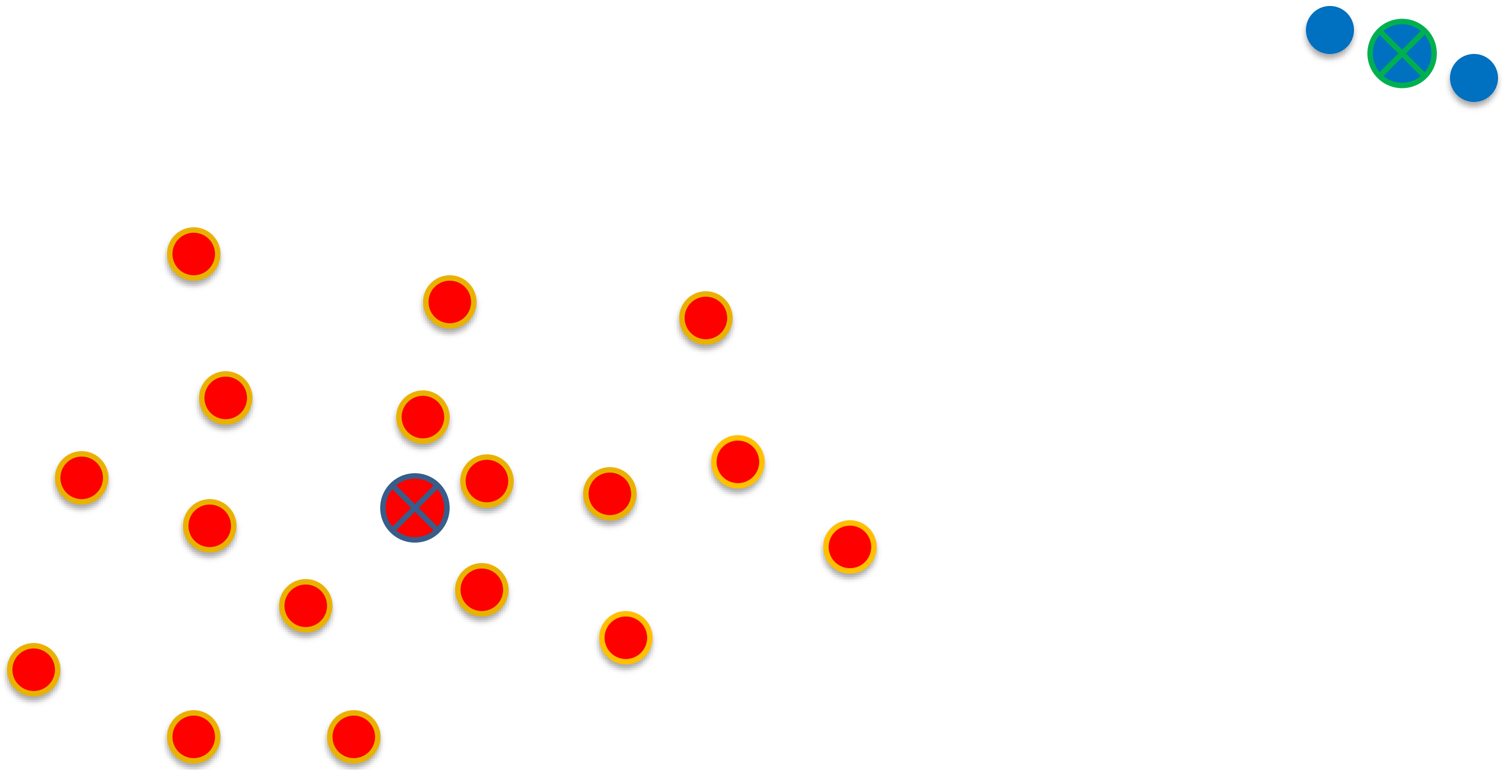


Properties of K-means algorithm

$$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\} \subset \boldsymbol{X} \subset \mathbf{R}^m$$

- Guaranteed to converge in a finite number of iterations.
- Run-time per iteration:
  K: # clusters; *N*: # data points; *m*: data dim.
  - Assign data points to closest cluster center: O(*KNm*).
  - Change the cluster center to the average of its assigned points: O(*Nm*).
- Non-deterministic: depends on center initialization.
- Requires *K* as a hyper-parameter.

K-means clustering failure case



Measuring the clustering performance

Reconstruction error:

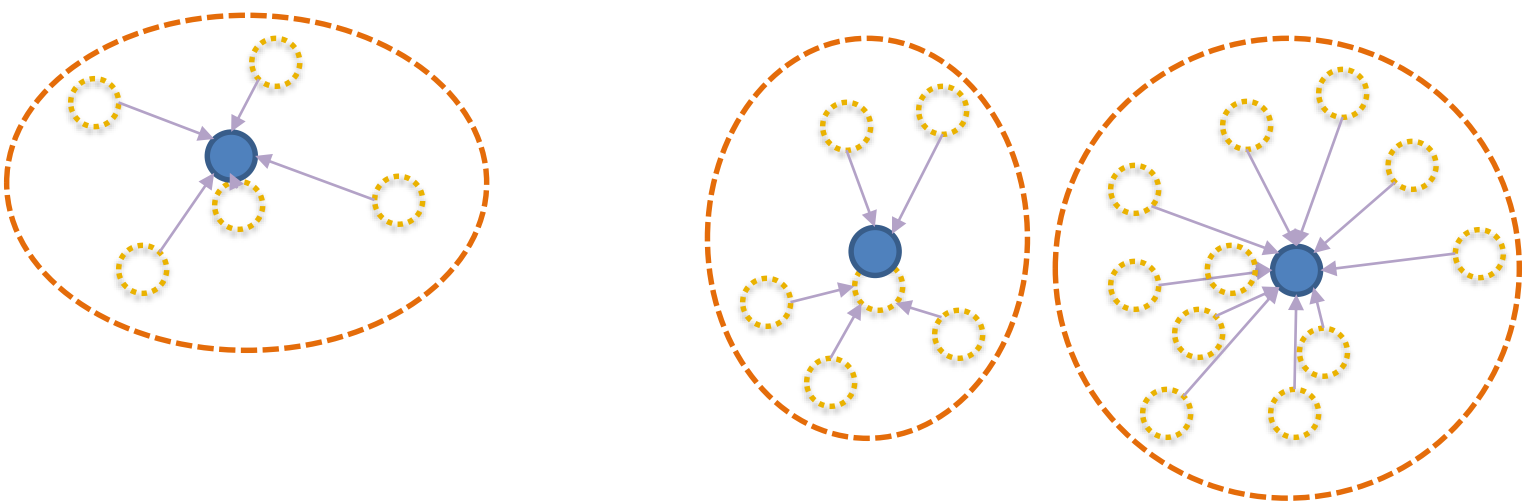
$$\begin{aligned}
\mathbf{x}^1 &\to \mathbf{c}^1 \\
\mathbf{x}^2 &\to \mathbf{c}^1 \\
\mathbf{x}^3 &\to \mathbf{c}^1 \\
\mathbf{x}^4 &\to \mathbf{c}^2 \\
\mathbf{x}^5 &\to \mathbf{c}^2 \\
\mathbf{x}^6 &\to \mathbf{c}^3 \\
\mathbf{x}^7 &\to \mathbf{c}^3 \\
\mathbf{x}^8 &\to \mathbf{c}^3
\end{aligned}$$

$$\begin{aligned}
&\|\mathbf{x}^1 - \mathbf{c}^1\| \\
&\|\mathbf{x}^2 - \mathbf{c}^1\| \\
&\|\mathbf{x}^3 - \mathbf{c}^1\| \\
&\|\mathbf{x}^4 - \mathbf{c}^2\| \\
&\|\mathbf{x}^5 - \mathbf{c}^2\| \\
&\|\mathbf{x}^6 - \mathbf{c}^3\| \\
&\|\mathbf{x}^7 - \mathbf{c}^3\| \\
&\|\mathbf{x}^8 - \mathbf{c}^3\|
\end{aligned}$$

$$\sum_{i=1}^{N} \|\mathbf{x}^i - \mathbf{c}(\mathbf{x}^i)\|$$

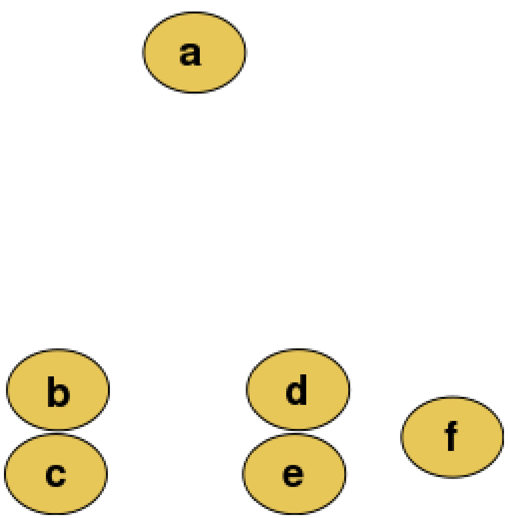
or

$$\sum_{i=1}^{N} \|\mathbf{x}^i - \mathbf{c}(\mathbf{x}^i)\|^2$$

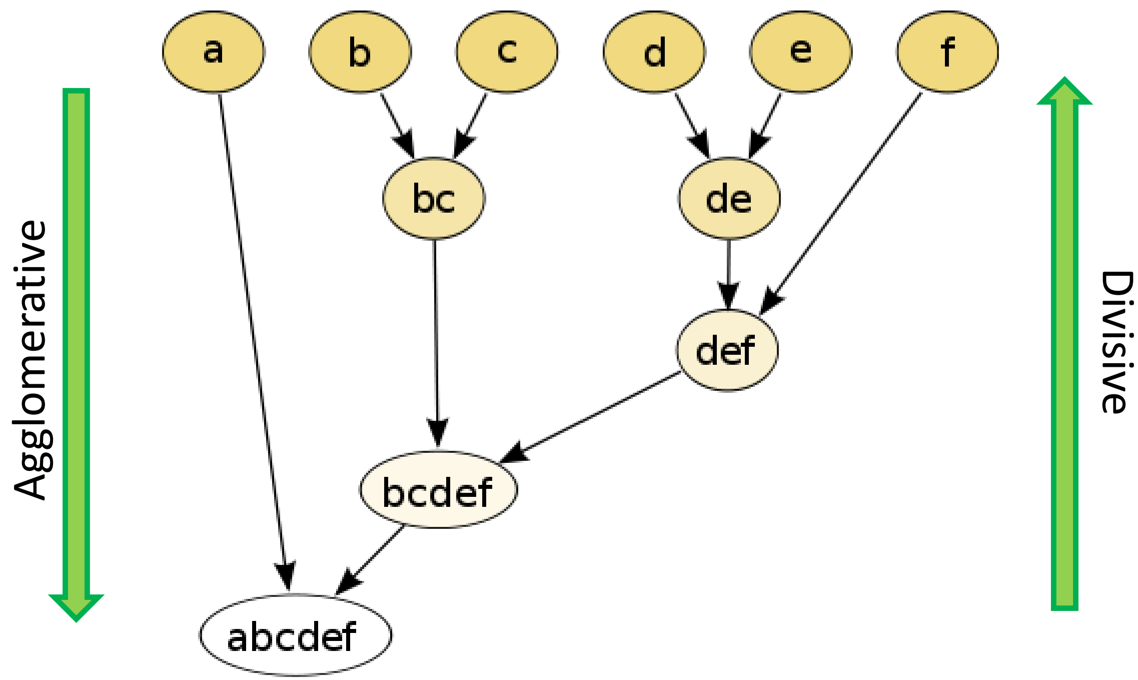


Other clustering approaches: Hierarchical clustering

builds a hierarchy of clusters



Raw data



Clustering tree

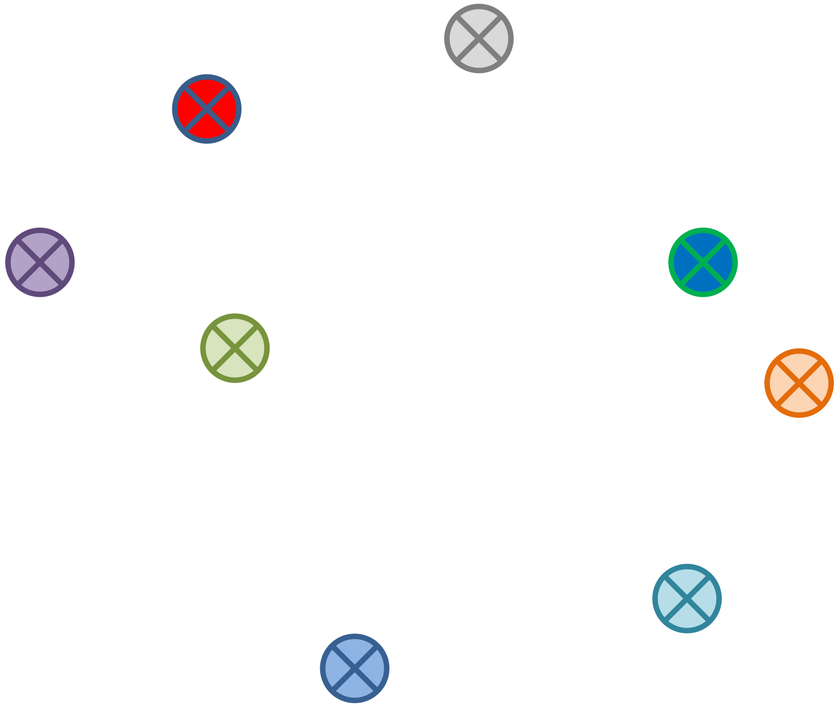
[Wikipedia]

Single-linkage clustering

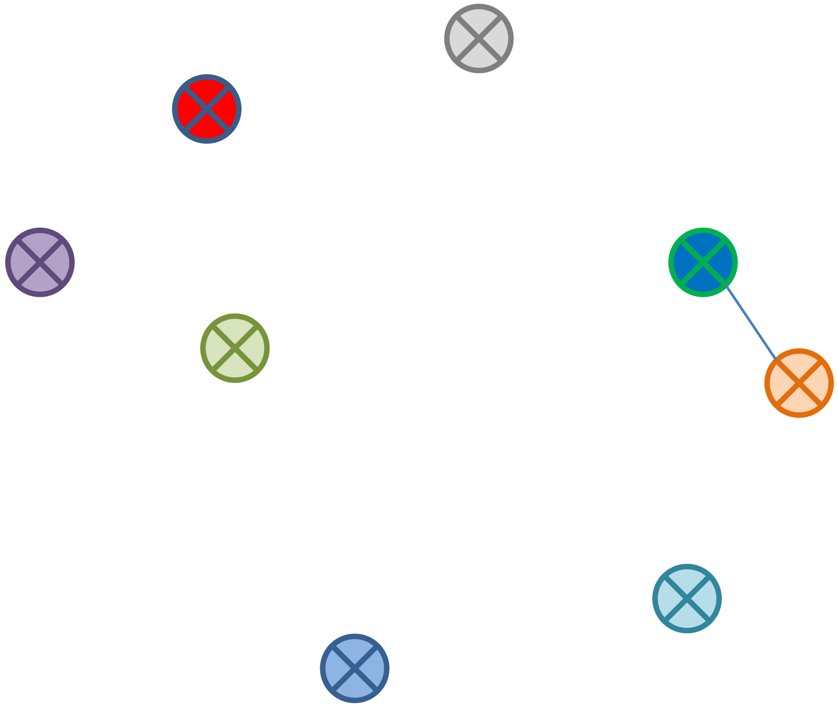
An iterative clustering algorithm

1. Initialize: assign a cluster center to each data point.
2. Iterate
   - Find the two closest pair of cluster centers.
   - Merge the two clusters and define a new center.
3. Stop when # number of clusters < Threshold.

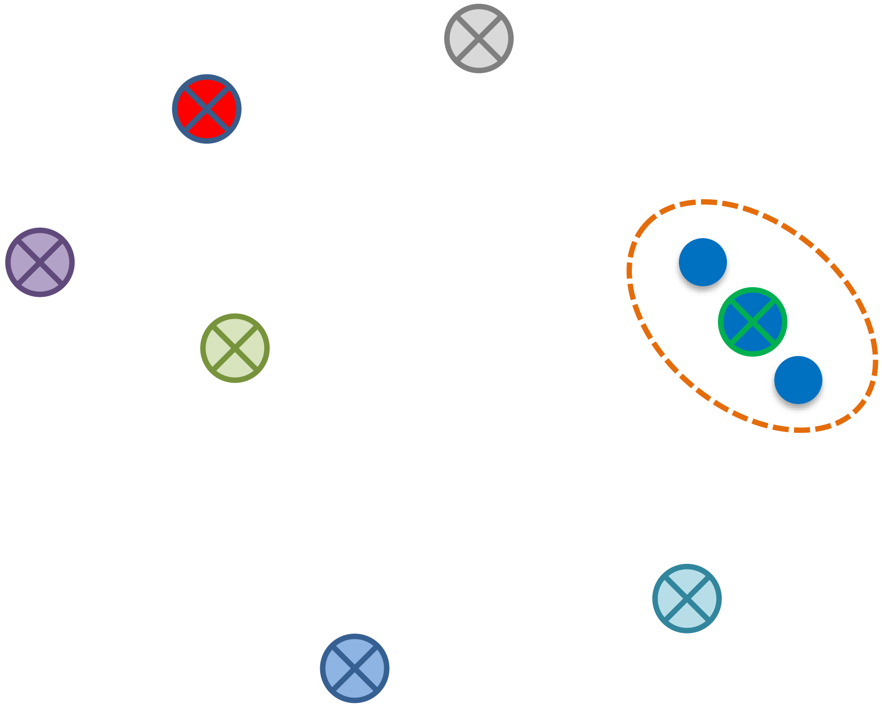
Single linkage clustering example



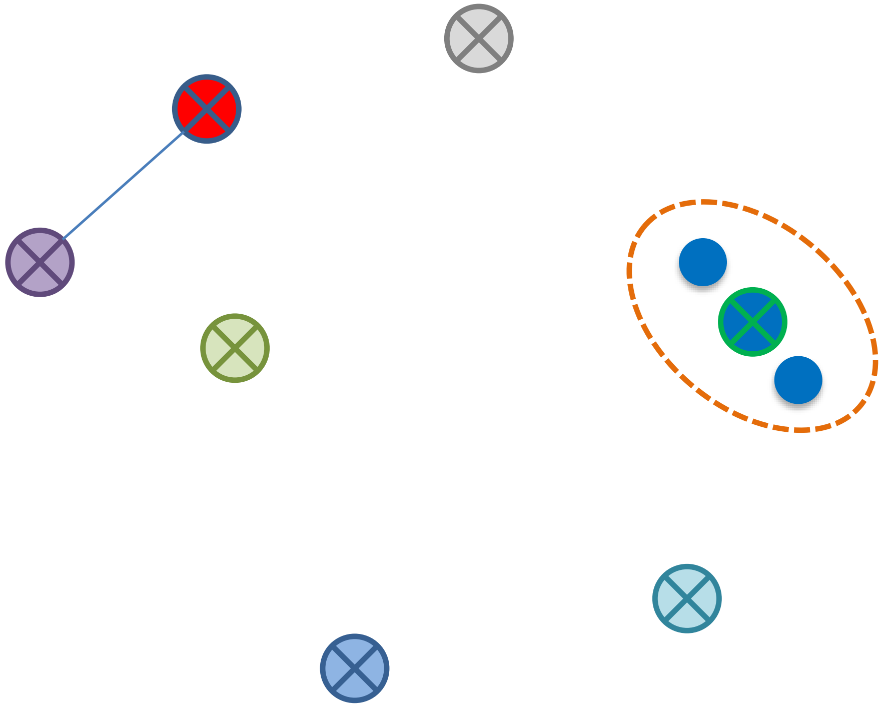
Single linkage clustering example



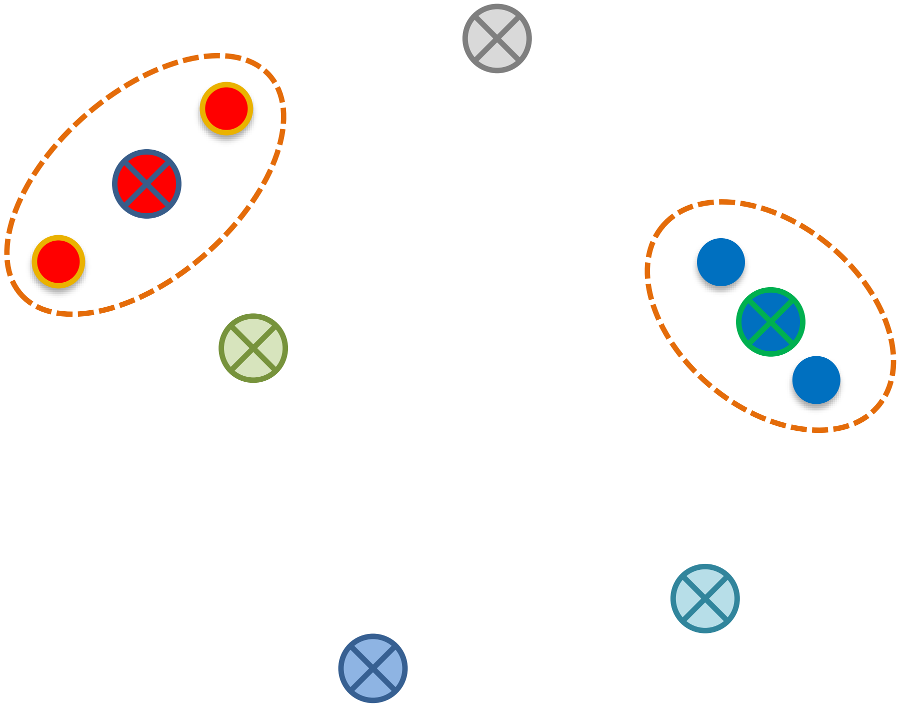
Single linkage clustering example



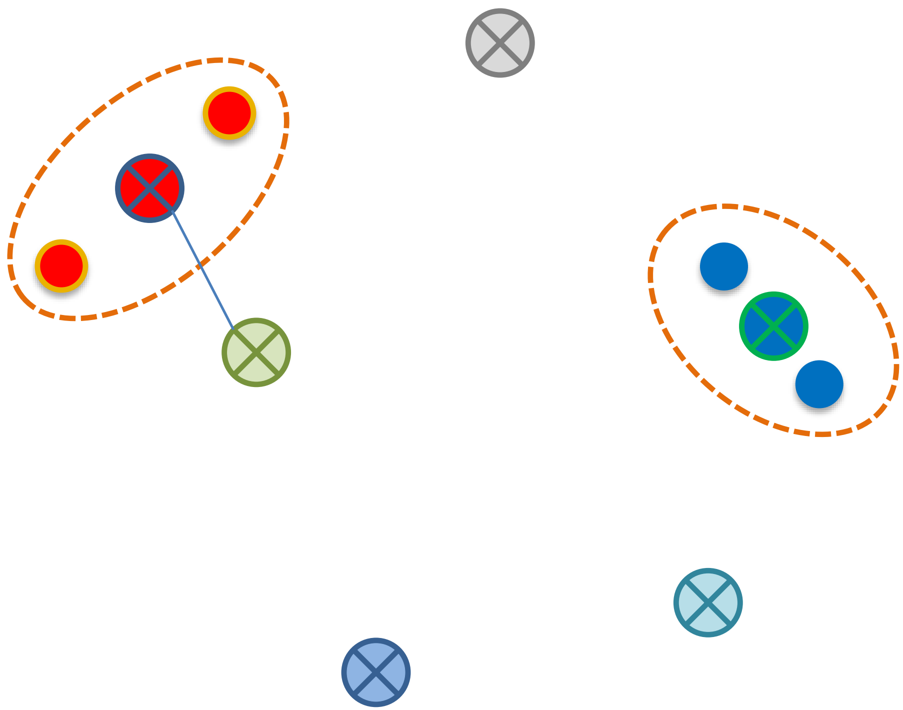
Single linkage clustering example



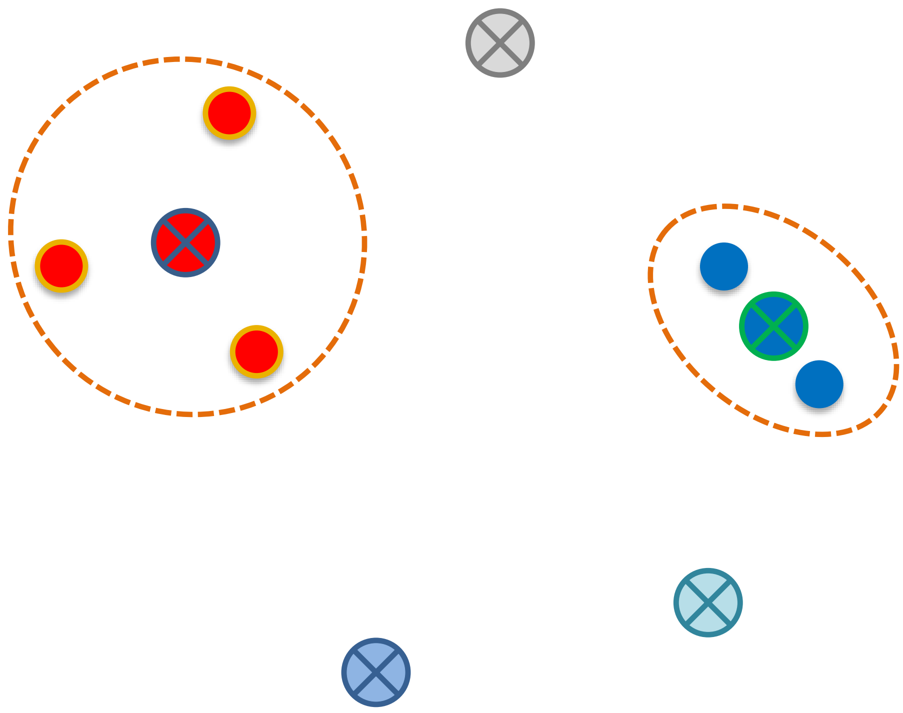
Single linkage clustering example



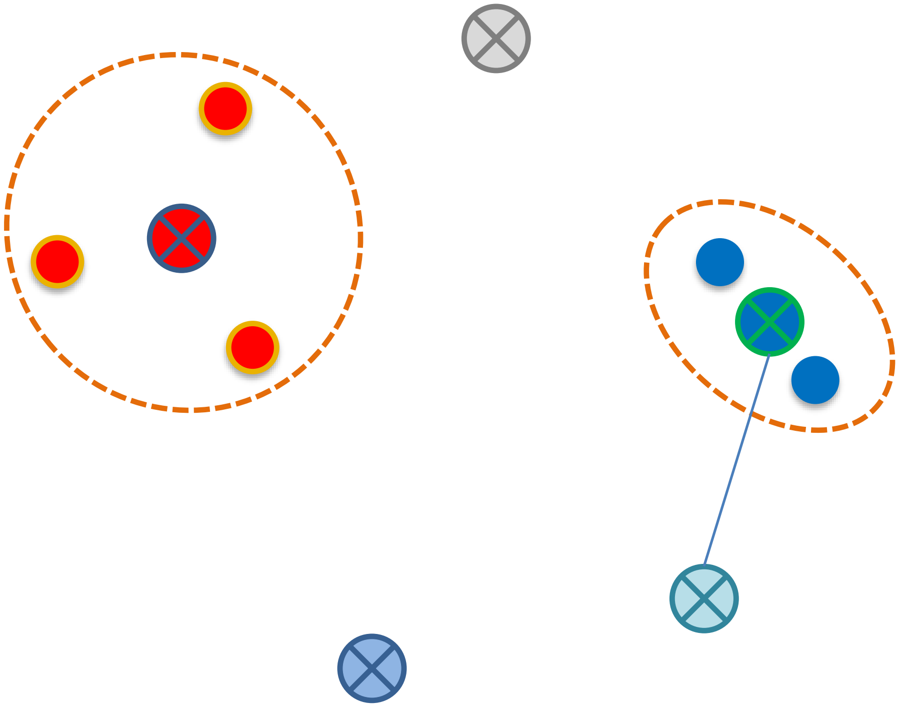
Single linkage clustering example



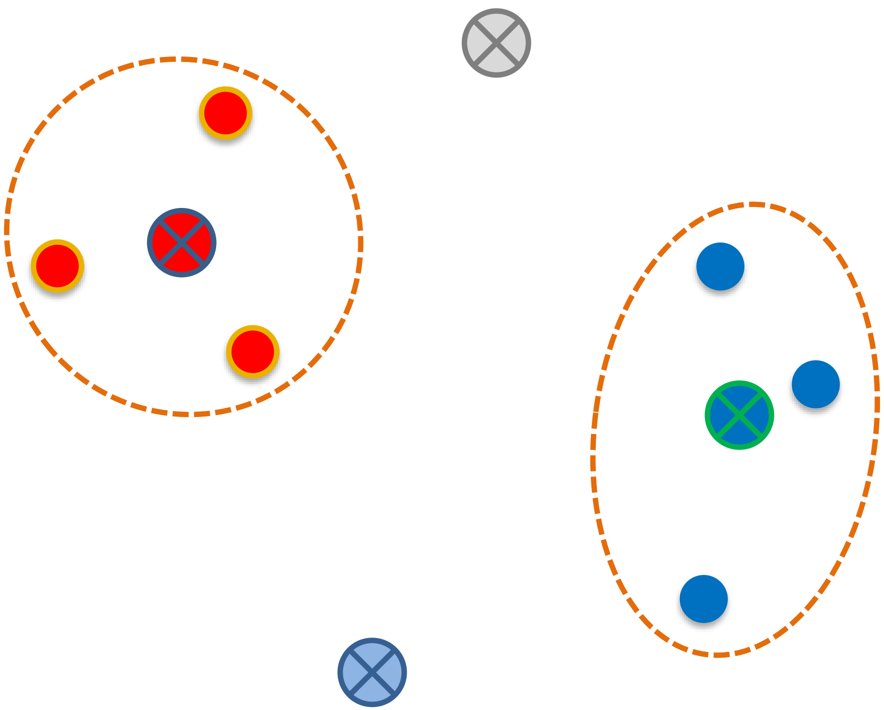
Single linkage clustering example



Single linkage clustering example

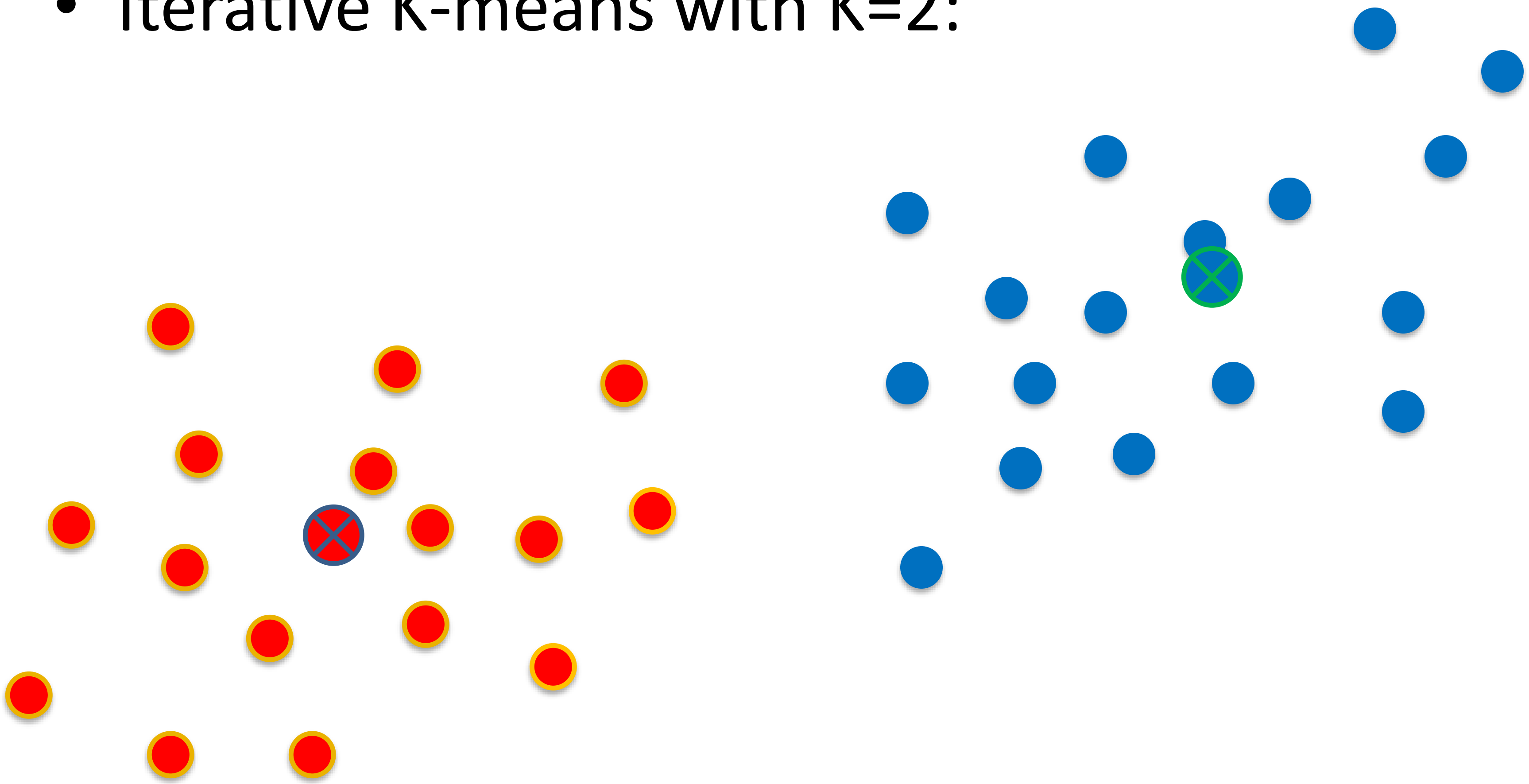


Single linkage clustering example



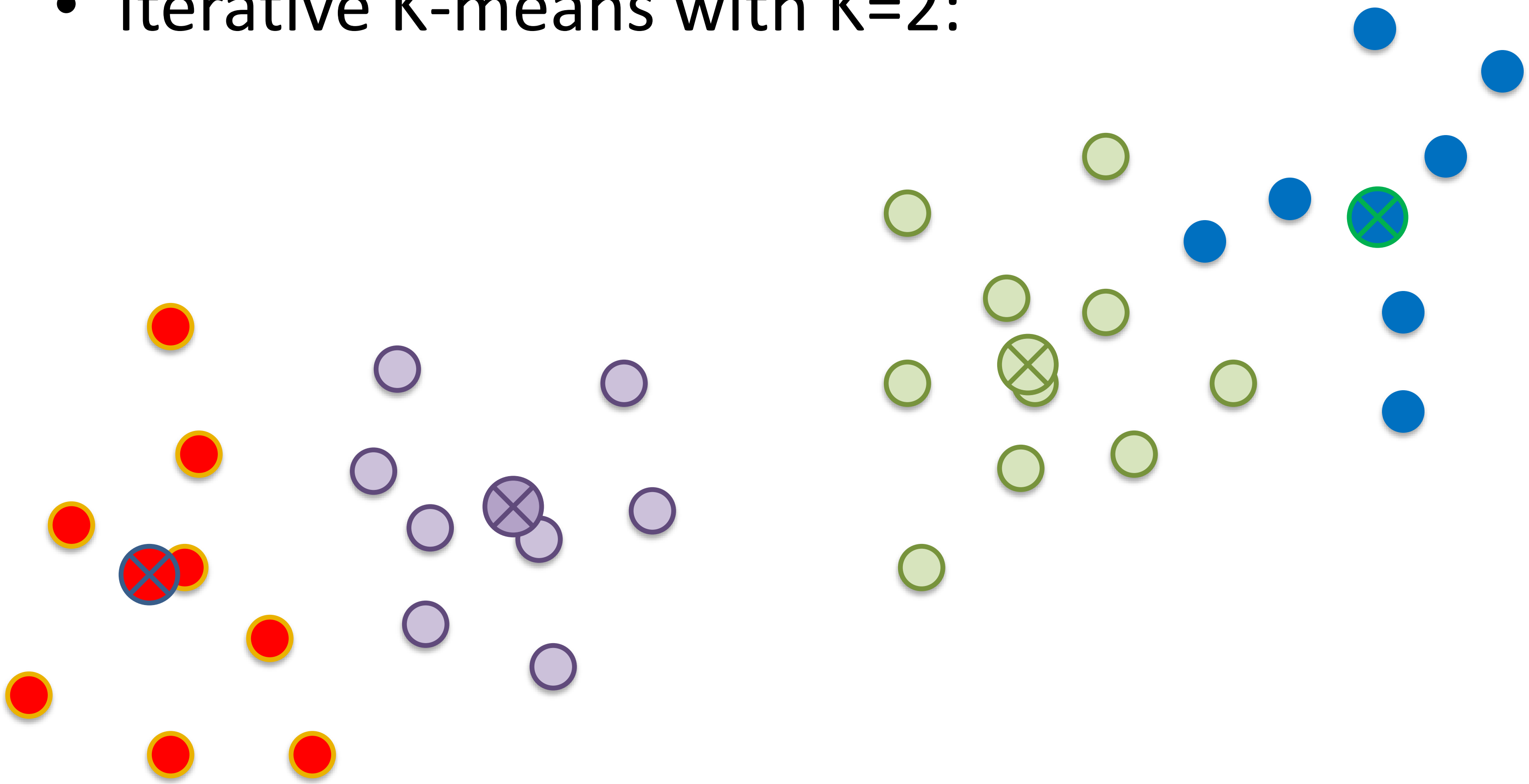
Divisive clustering example

- Iterative K-means with K=2:



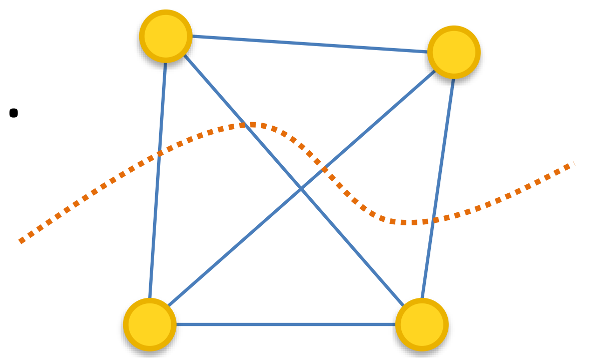
Divisive clustering example

- Iterative K-means with K=2:



Other clustering approaches: Spectral clustering

- A good clustering algorithm will produce
  - Low intra-cluster distances.
  - High inter-cluster distances.
- Finding such a clustering is NP-hard.



- Spectral clustering approximates the above objective and form a clustering by solving an approximate optimization problem.
- Uses a graph representation of data.

Supervised, unsupervised, semi-supervised learning

- Supervised learning
  - learn from labelled examples:

$$D = \{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^N, y^N)\}$$

Pairs of input and the corresponding desired output.

- Unsupervised learning
  - learn from unlabelled examples:

$$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$$

Input data only; no desired outputs.

Supervised, unsupervised, semi-supervised learning

- Supervised learning
  - learn from labelled examples:

$$D = \{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^N, y^N)\}$$

Pairs of input and the corresponding desired output.

- Unsupervised learning
  - learn from unlabelled examples:

$$D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$$

Input data only; no desired outputs.

- Semi-supervised learning
  - learn from labelled examples:

$$D = \{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^N, y^N)\}$$

and unlabelled examples:

$$U = \{\mathbf{x}^{N+1}, \dots, \mathbf{x}^{N+U}\}.$$

Semi-supervised Learning

- In semi-supervised learning, we are given labeled data points

$$D = \{(\mathbf{x}^1, \boxed{y^1}), \ldots, (\mathbf{x}^N, \boxed{y^N})\}$$

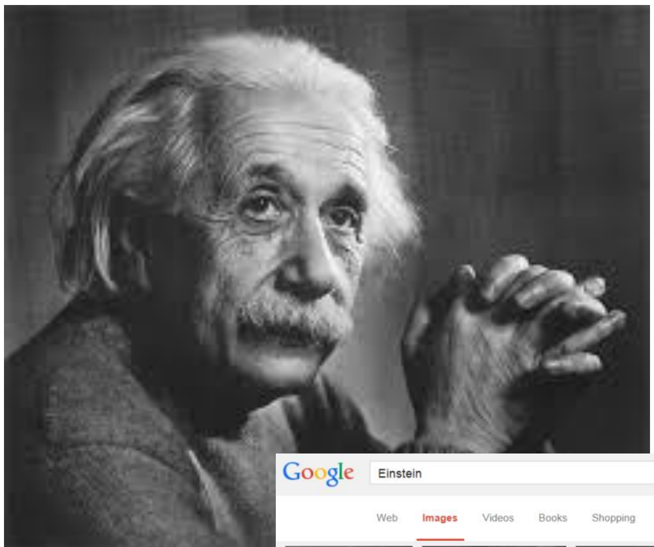
as well as unlabeled data points

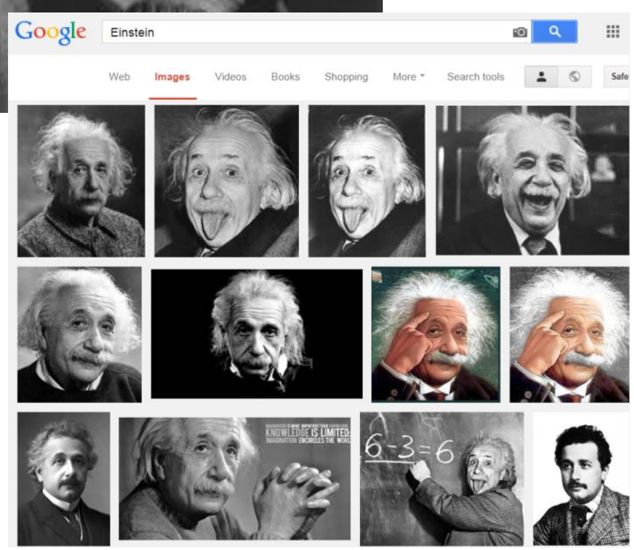
$$U = \{\mathbf{x}^{N+1}, \ldots, \mathbf{x}^{N+1}\}.$$

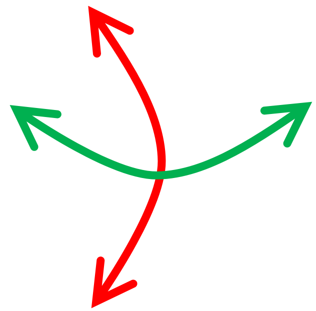
- Our task is to assign labels to U.

Sometimes, called *transductive learning*.

Why semi-supervised learning?



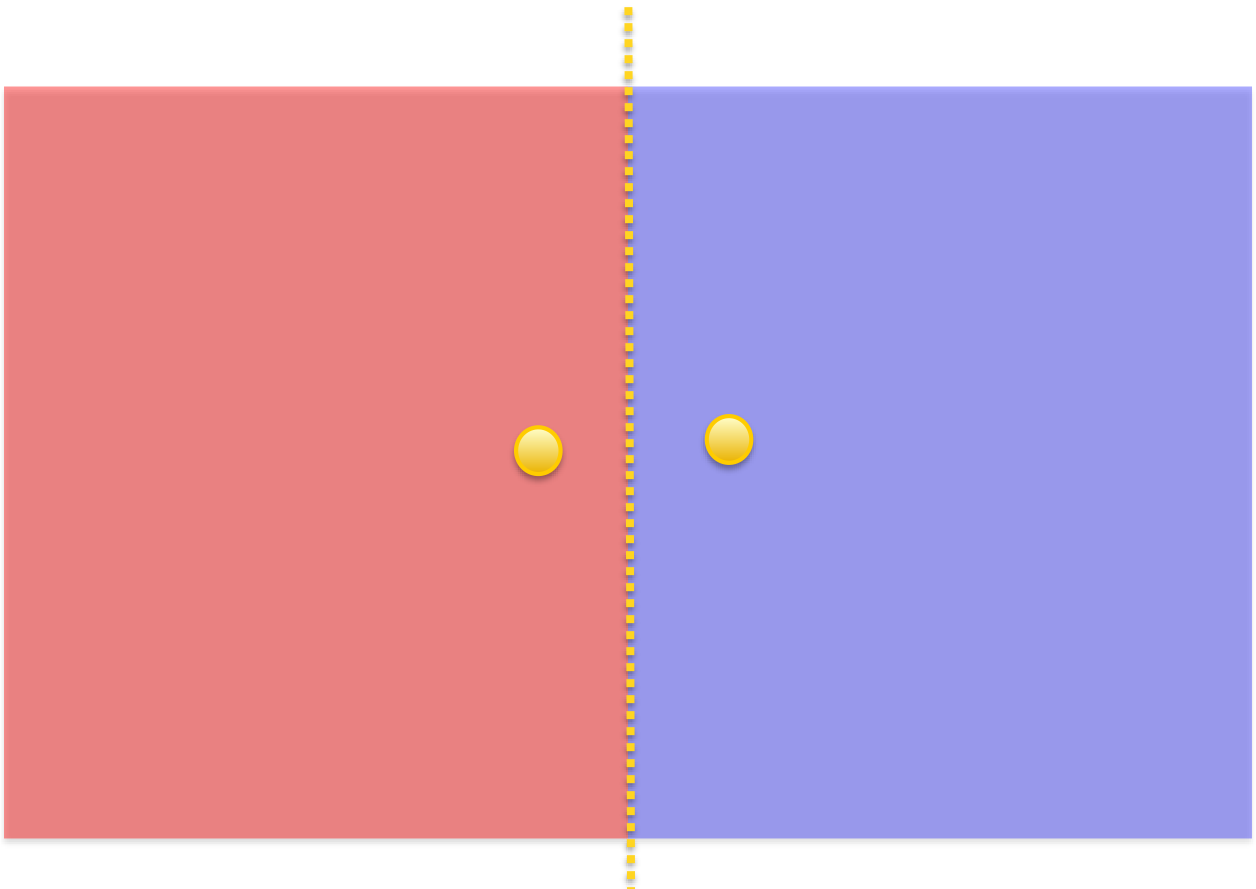




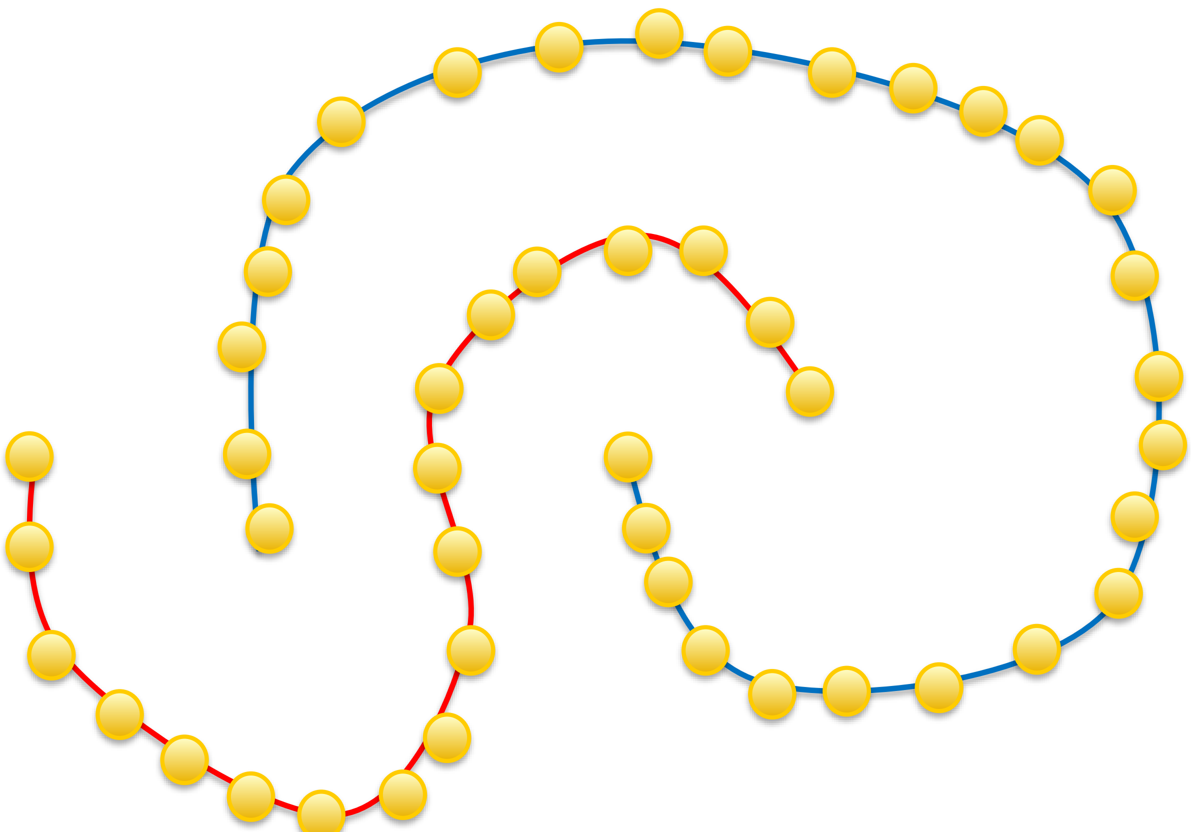
Sometimes, labeling data is difficult; costly.

Obtaining unlabeled data points is easier!

Why semi-supervised learning?

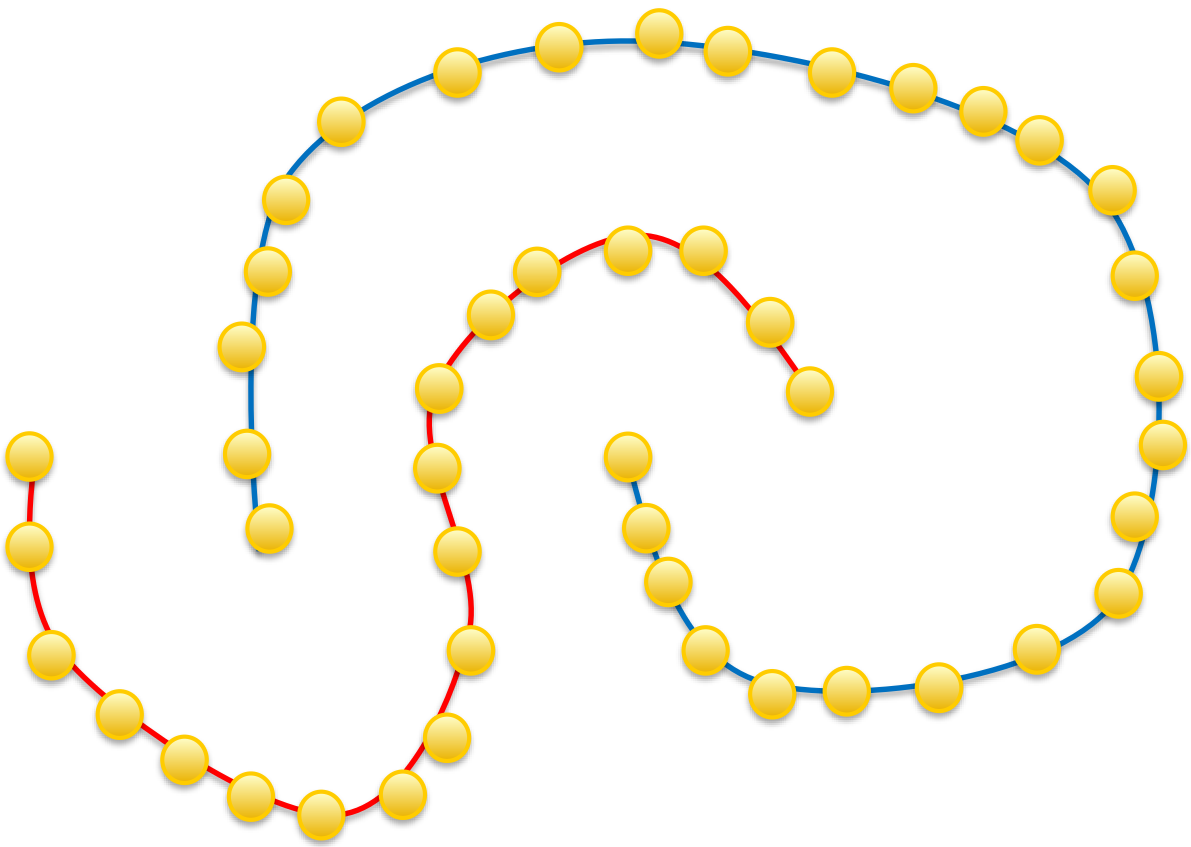


Why semi-supervised learning?



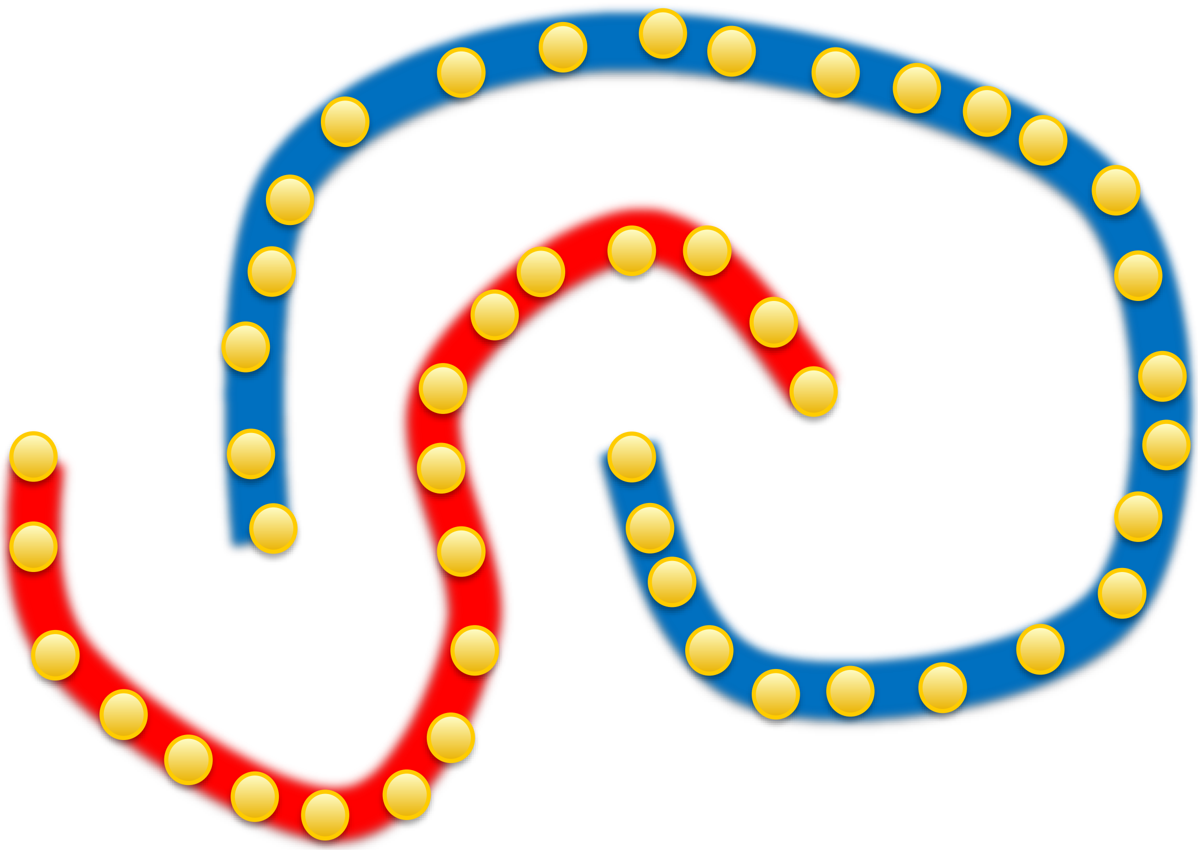
Unlabeled data can provide information on the underlying data generation process.

How to solve semi-supervised learning?



Manifold assumption:
Data are lying on
manifolds.

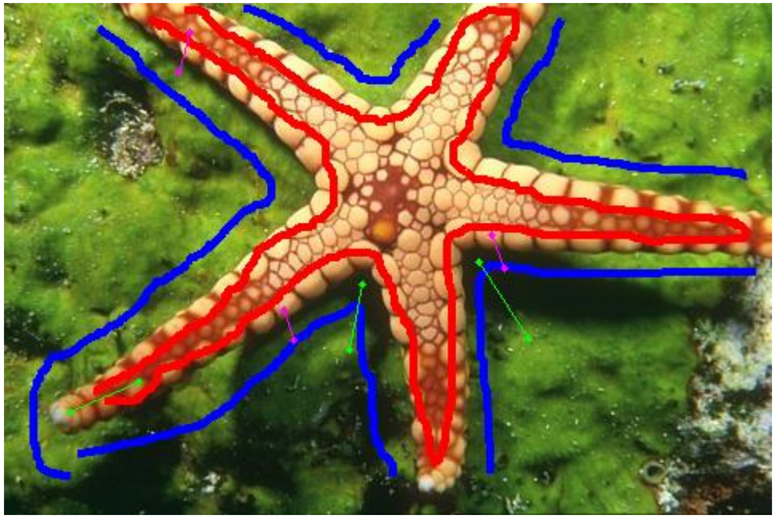
How to solve semi-supervised learning?

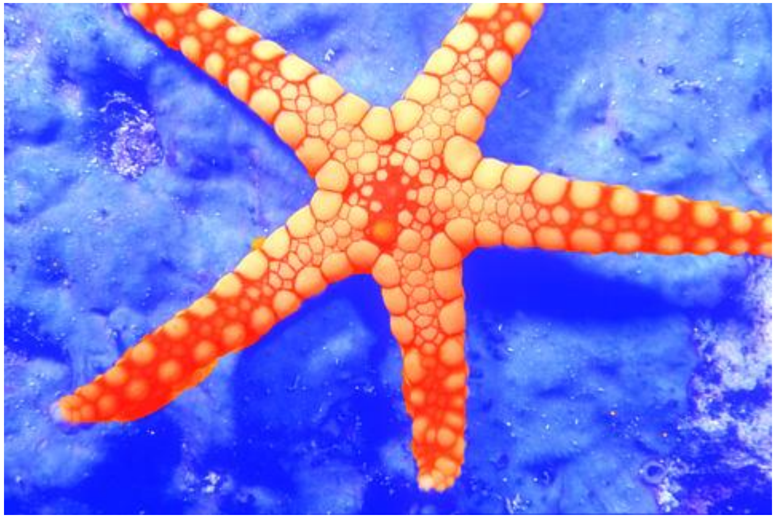


Cluster assumption:
A decision boundary should lie on a low-density region.

Applications: Image colorization

Applications: Interactive image segmentation





Slides references

- David Sontag, *Clustering*, lecture slides.
- Wikipedia article on clustering.