CHA 2555
Artificial Intelligence

Clustering – Part 1

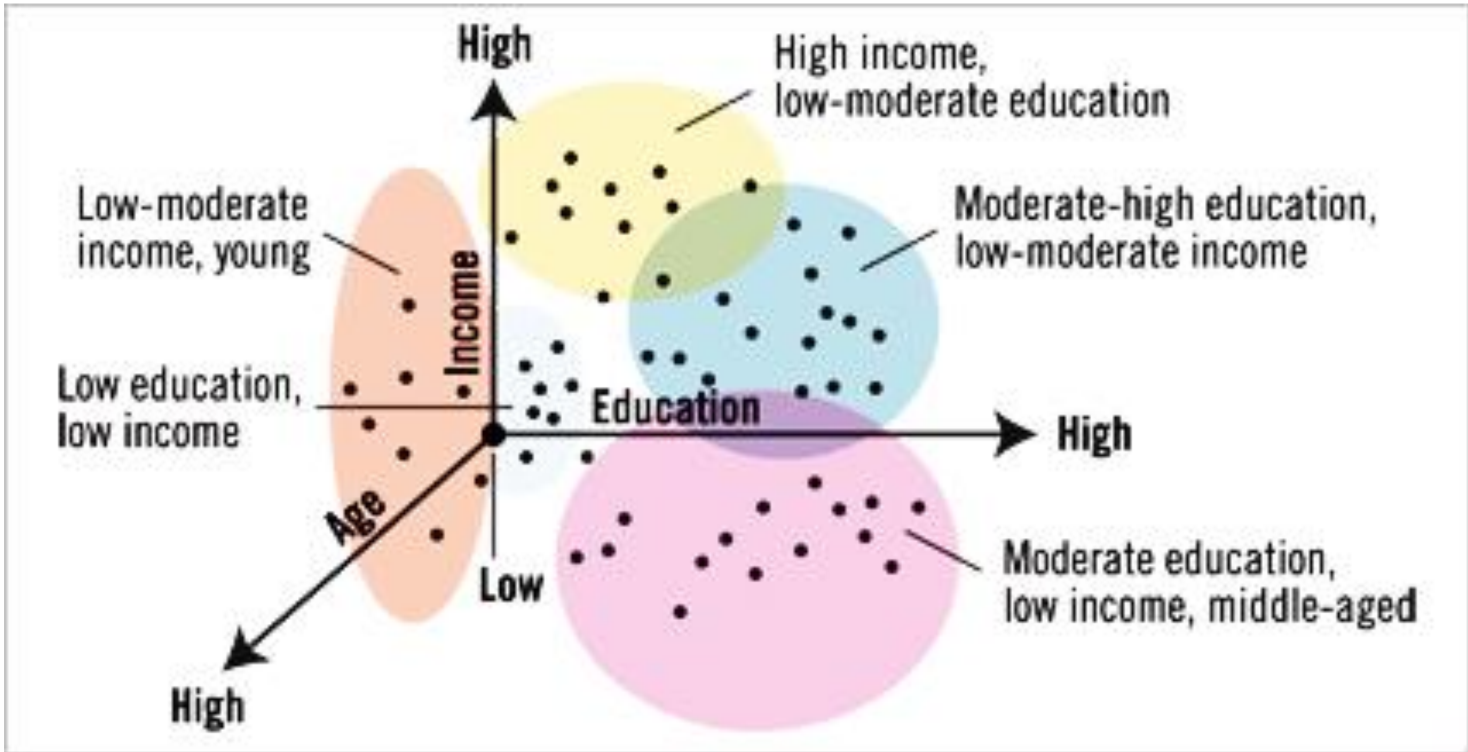
Dr Tianhua Chen

Department of Computer Science
University of Huddersfield

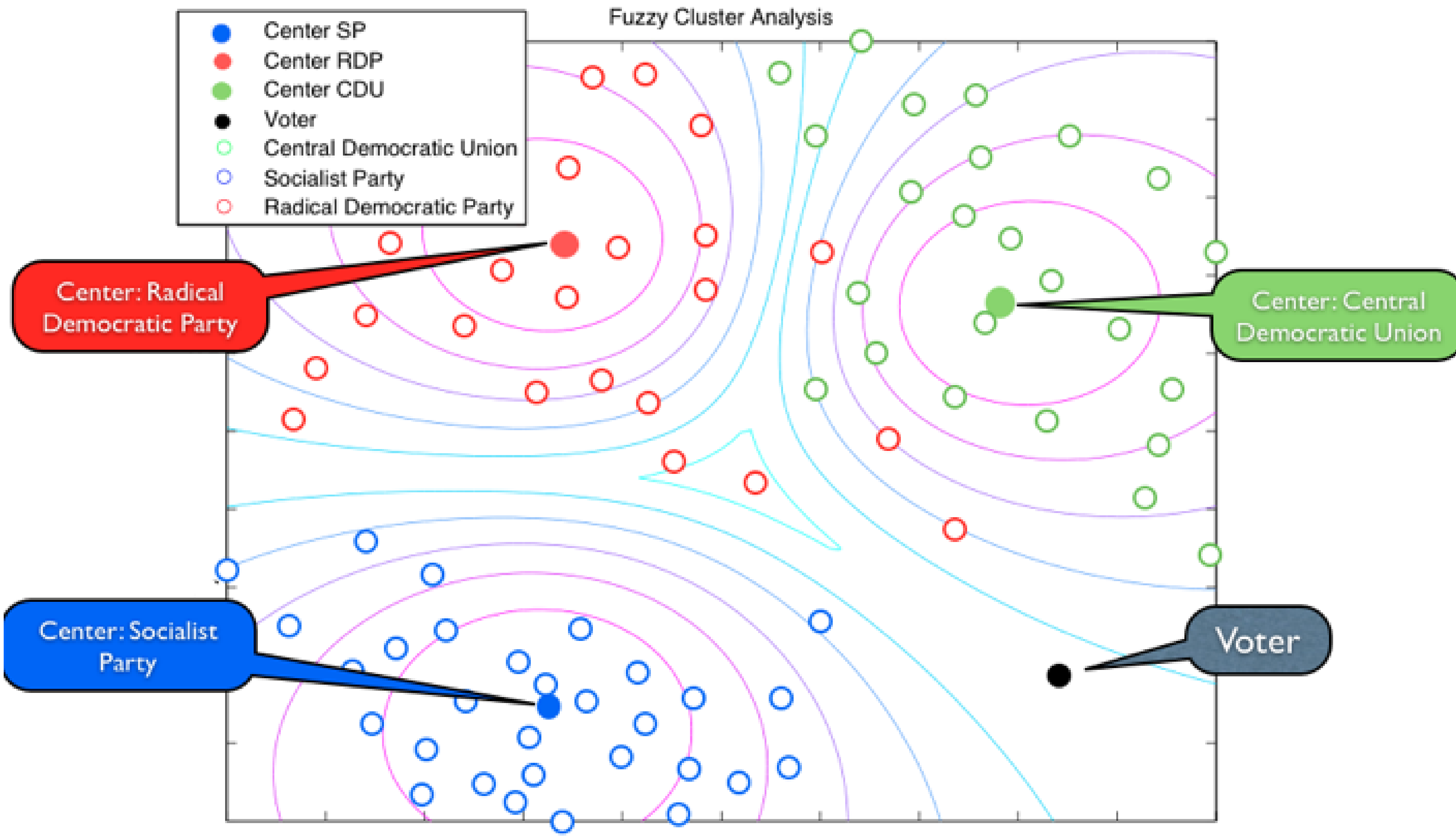
What we'll look at

- Preliminaries: datasets, data points, features, distance
- What is clustering?
- Partitional clustering
  - *k*-means algorithm
  - Extensions (fuzzy)

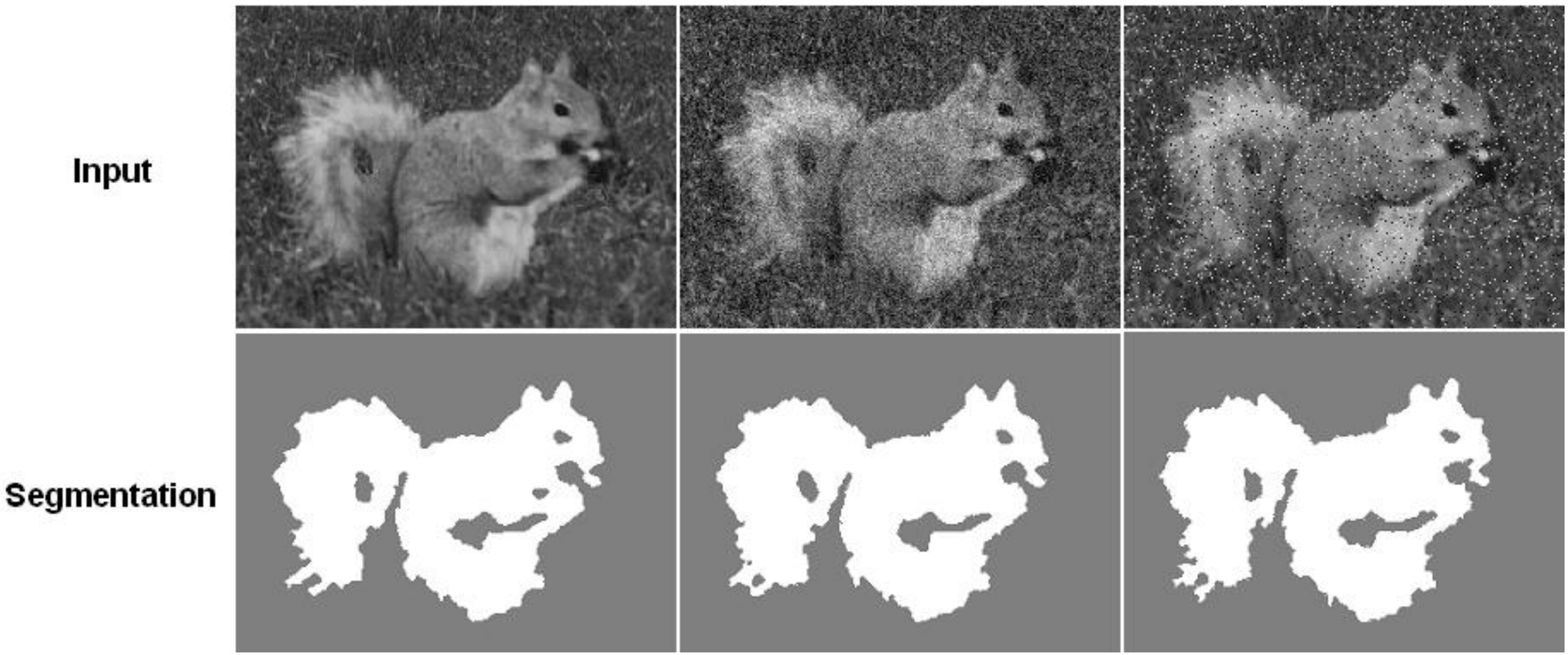
Why clustering is useful



Why clustering is useful

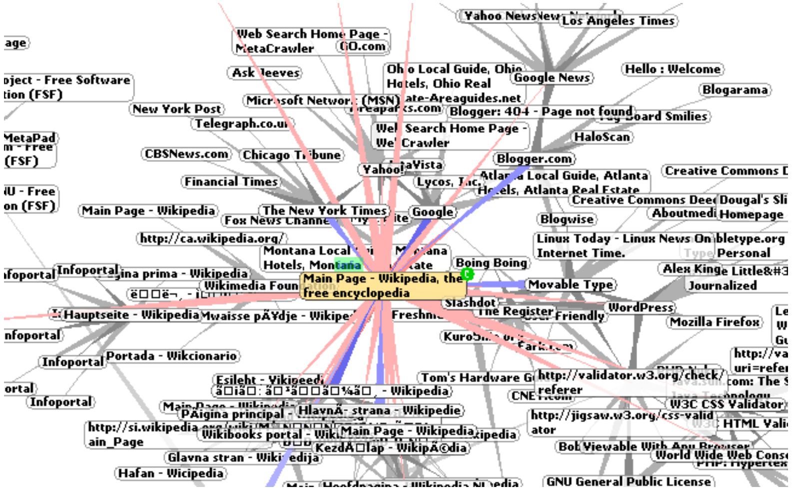


Why clustering is useful



Clustering applications

- Image segmentation (e.g. object recognition)
- Social network analysis (recognise communities)
- Bioinformatics ([gene expression] analysis)
- Data mining/retrieval (e.g. search result grouping)
- Recommender systems (clusters of users)
- …



Original

Kmeans : 8

Datasets - labelled

	Headache	*Muscle pain*	*Temp.*	*Flu*
1	Yes	Yes	37.2	No
2	Yes	Yes	38.1	Yes
3	Yes	Yes	39.0	Yes
4	No	Yes	36.9	No
5	No	No	37.9	No
6	No	Yes	39.2	Yes

Each row is an object/data point

Each column is a feature/symptom/measurement = dimensions

Have a *class* feature = decision/diagnosis

Datasets - unlabelled

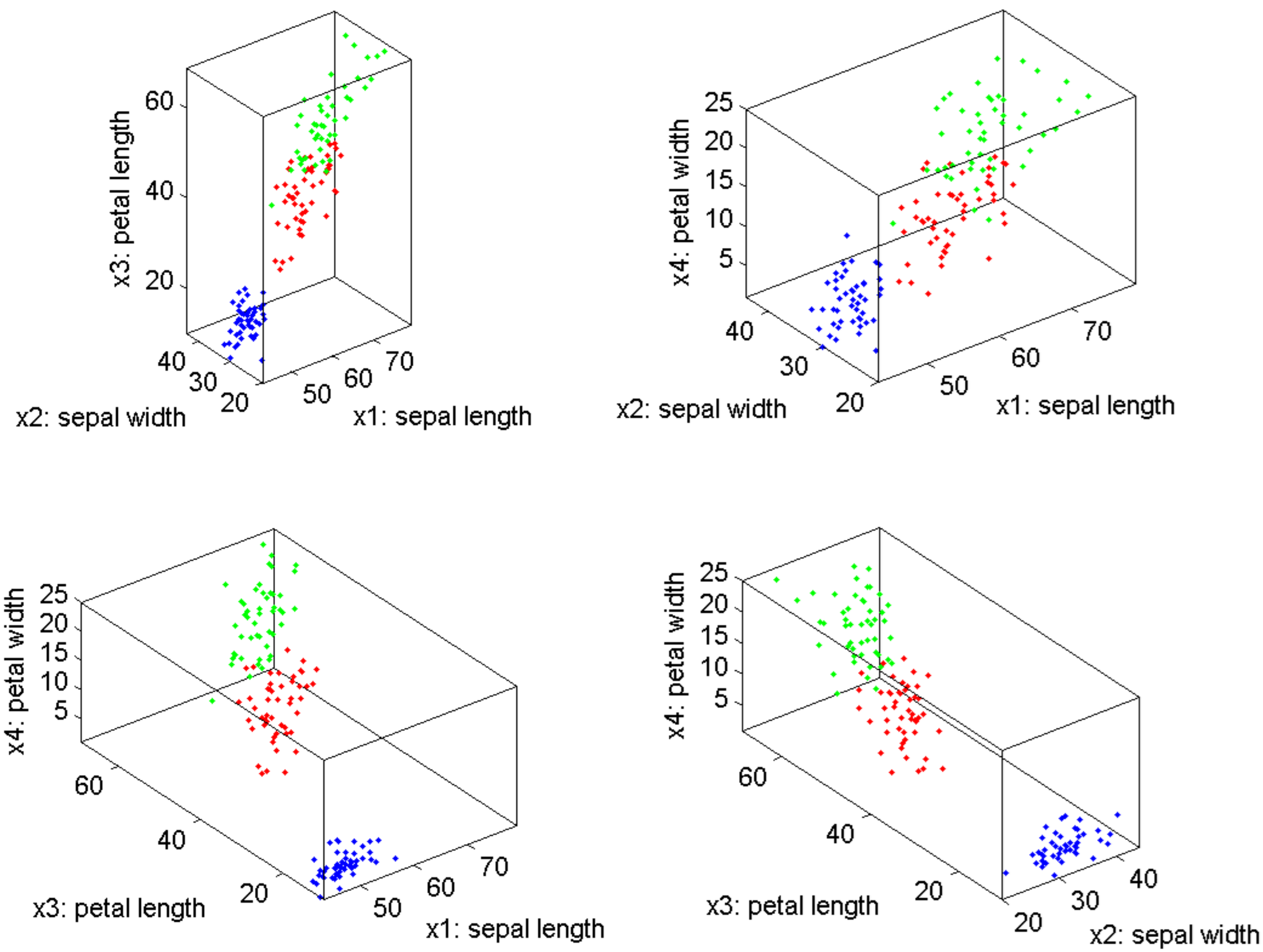
	Headache	*Muscle pain*	*Temp.*	?
1	Yes	Yes	37.2	
2	Yes	Yes	38.1	
3	Yes	Yes	39.0	
4	No	Yes	36.9	
5	No	No	37.9	
6	No	Yes	39.2	

Example: Iris dataset

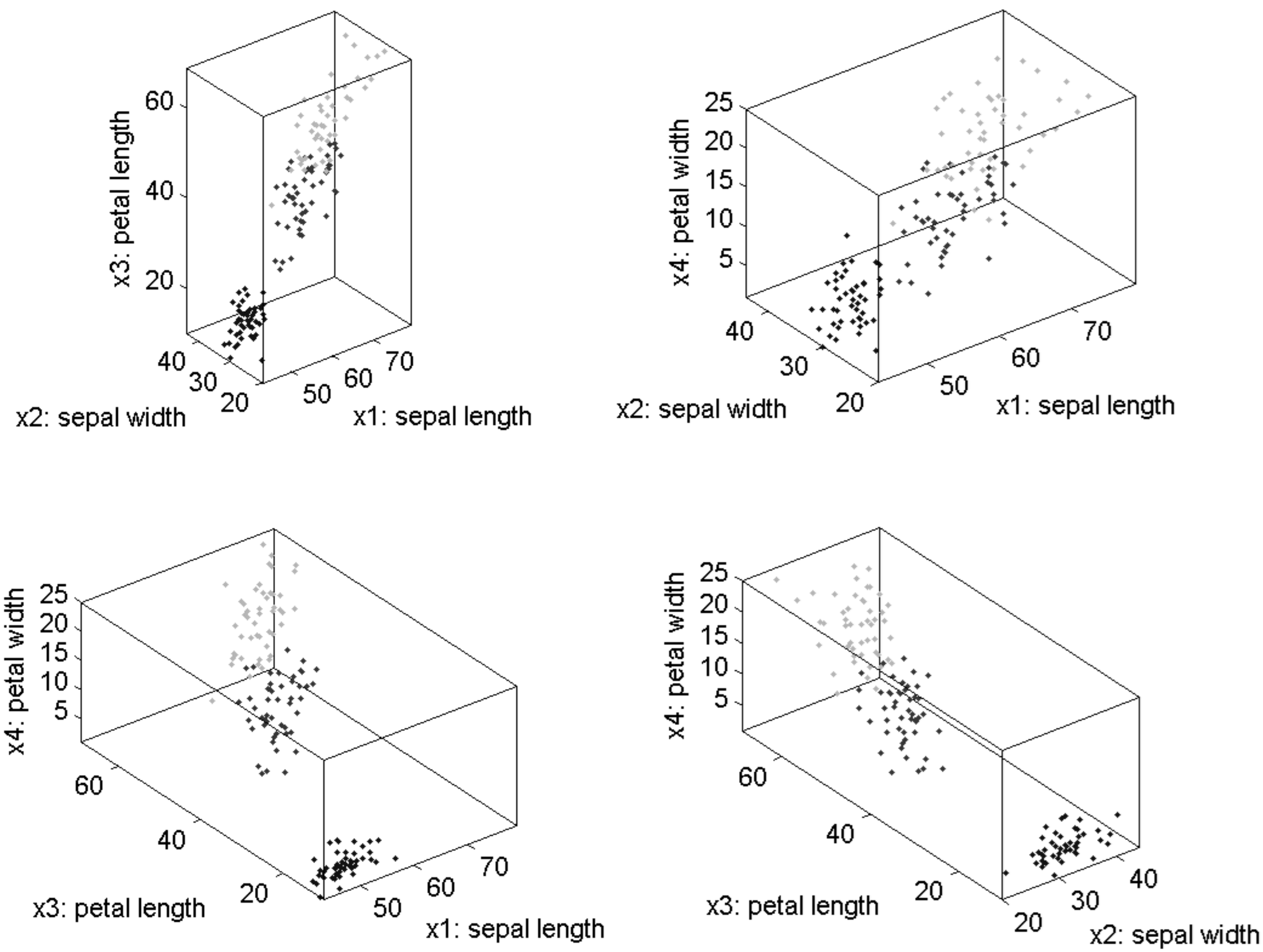
Example: Iris dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa

Example: Iris dataset



Example: Iris dataset



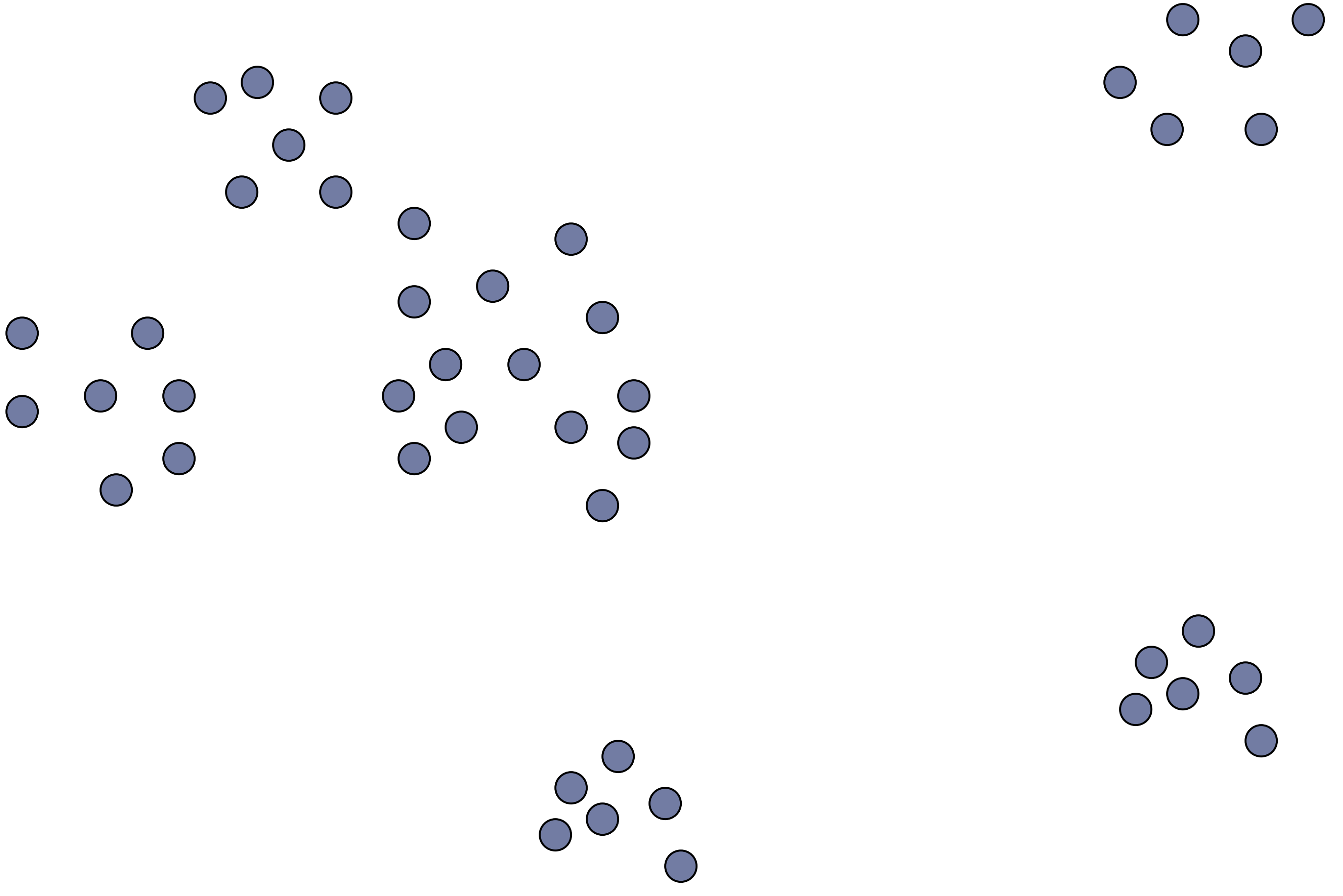
What is clustering?

- **Clustering**: the process of grouping a set of objects into classes of similar objects

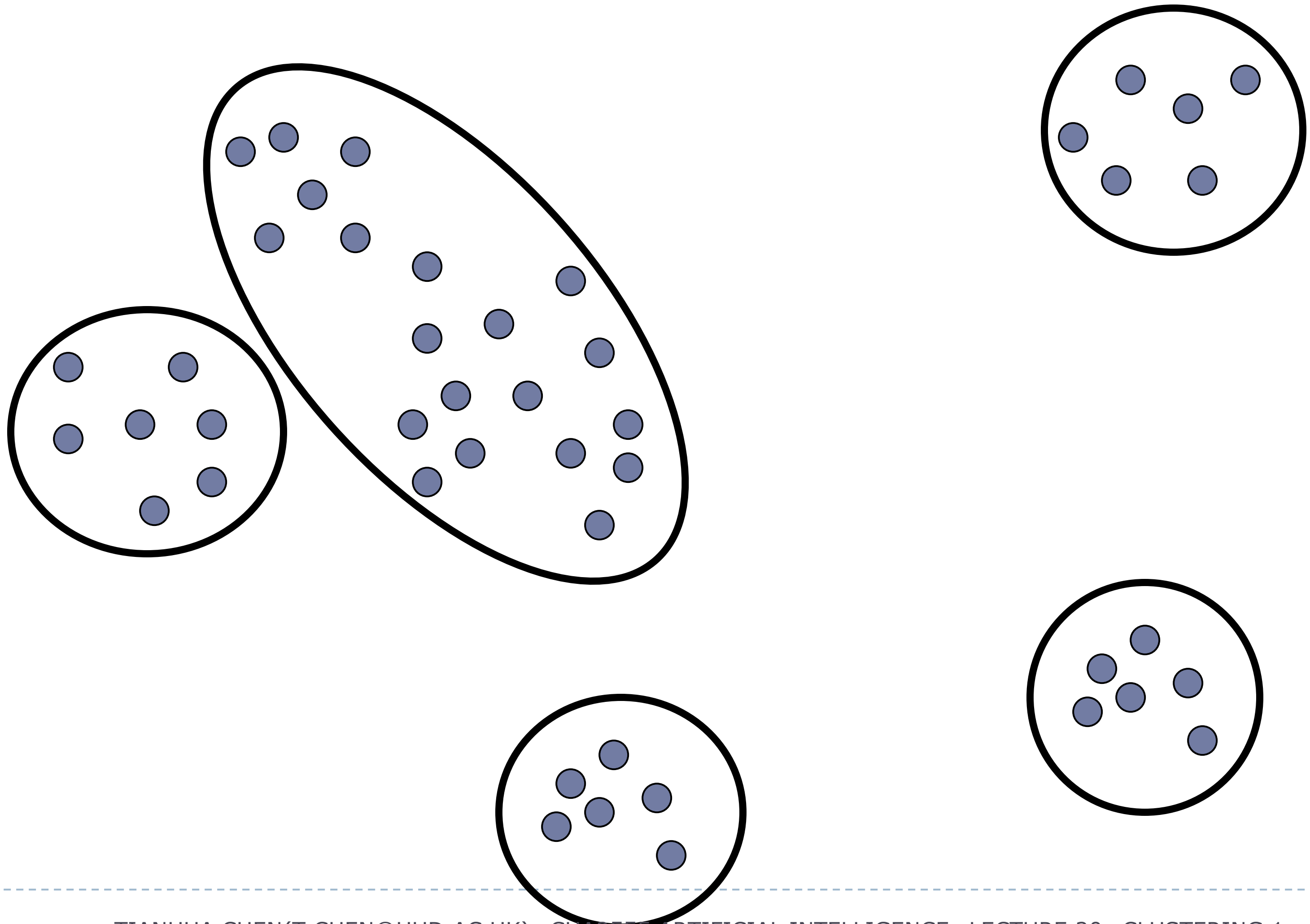
- Most common form of ***unsupervised learning***
  - Unsupervised learning = learning from raw data
  - …as opposed to supervised data where a classification of examples is given

	Headache	*Muscle pain*	*Temp.*	?
1	Yes	Yes	37.2	
2	Yes	Yes	38.1	
3	Yes	Yes	39.0	
4	No	Yes	36.9	
5	No	No	37.9	
6	No	Yes	39.2	

Clustering



Clustering



Clustering algorithms

- Partitional algorithms
  - Usually start with a random (partial) partitioning
  - Refine it iteratively
    - *k*-means clustering
    - Model-based clustering
- Hierarchical algorithms
  - Bottom-up, agglomerative
  - Top-down, divisive

Clustering considerations

- **Clustering**: the process of grouping a set of objects into classes of similar objects

- What does it mean for objects to be similar? How do we measure this?

- What algorithm and approach do we take?
  - Partitional
  - Hierarchical

Clustering considerations

- **Clustering**: the process of grouping a set of objects into classes of similar objects

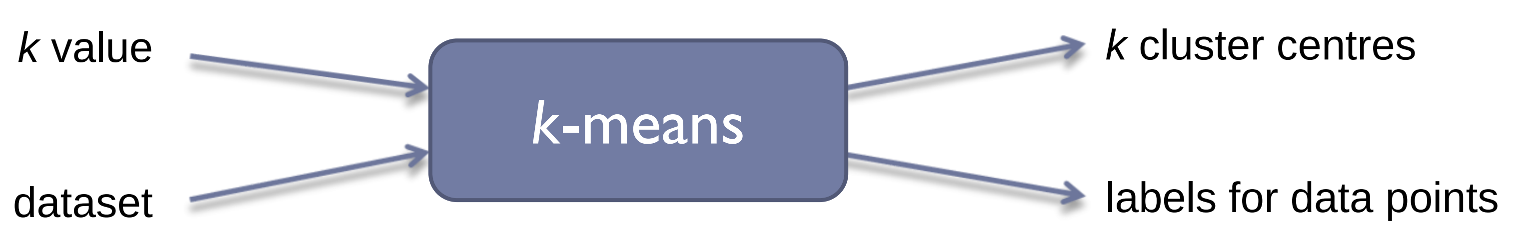
- How many clusters?

- Can we label or name the clusters?

- How do we make it efficient and scalable?

k–means algorithm(s)

- Terminology: **centroid** = a point that is considered to be the center of a cluster
- Start by picking *k*, the number of clusters (centroids)
- Initialise clusters by picking one point per cluster (seeds)
  - E.g., pick data points at random
  - Could also generate these randomly



Populating clusters

Iterate until converged

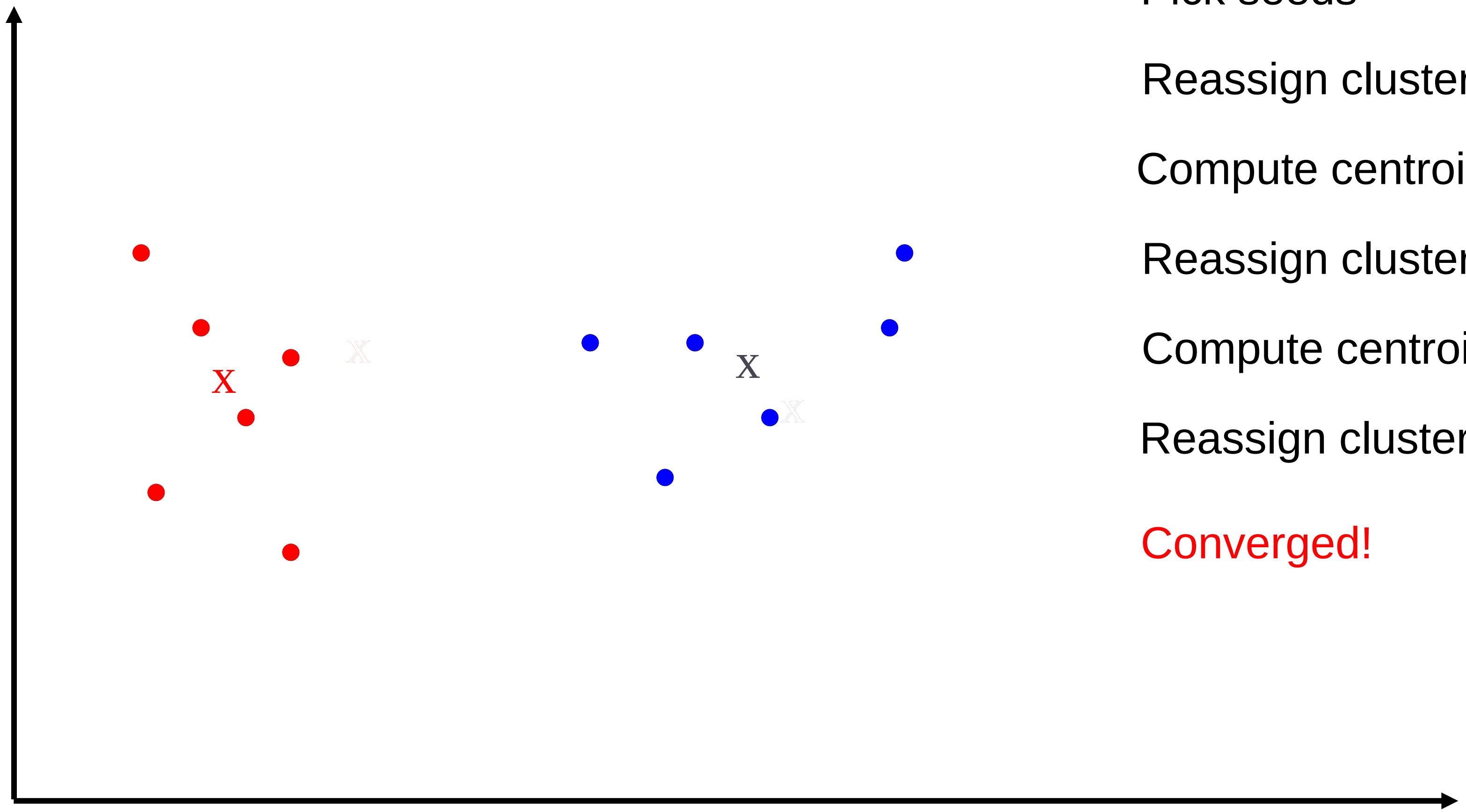
1. Compute **distance** from all data points to all *k* centroids

2. For each **data point**, assign it to the cluster whose current centroid it is nearest

3. For each **centroid**, compute the average (mean) of all points assigned to it

4. Replace the *k* centroids with the new averages

k-means example (k = 2)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

Converged!

Measuring distance

- Euclidean distance most often used to determine how far points are from each other (although alternatives exist)

- Defined as (features indexed from 1 to *n*):

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

- E.g. only two features: $\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2}.$

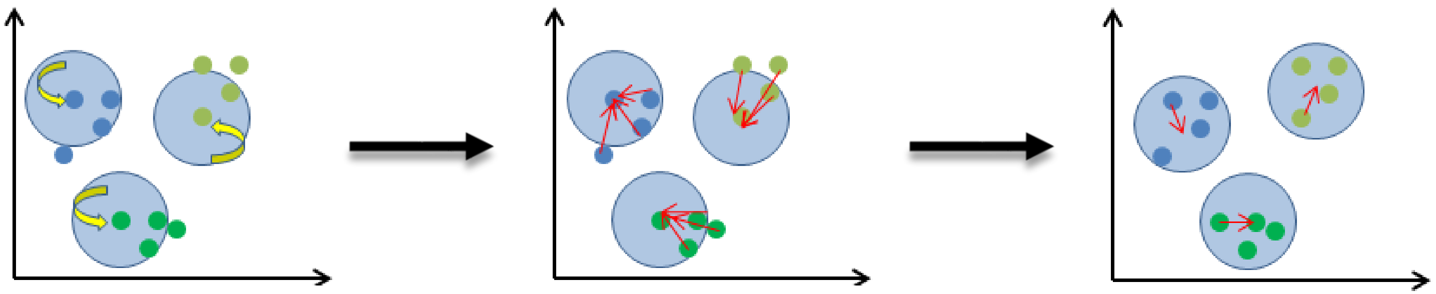
- E.g. only one feature: $\sqrt{(x-y)^2} = |x-y|.$

Distance between data points

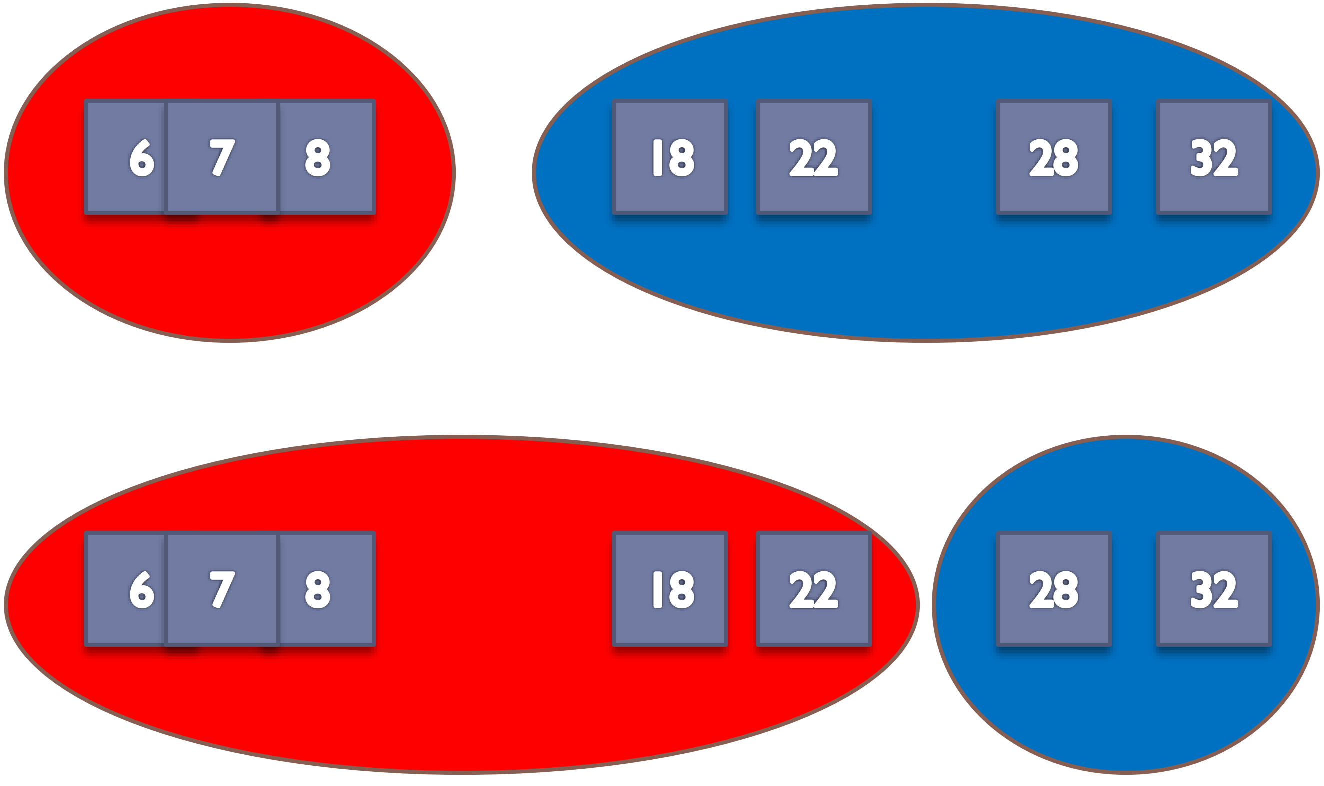
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2

Termination conditions

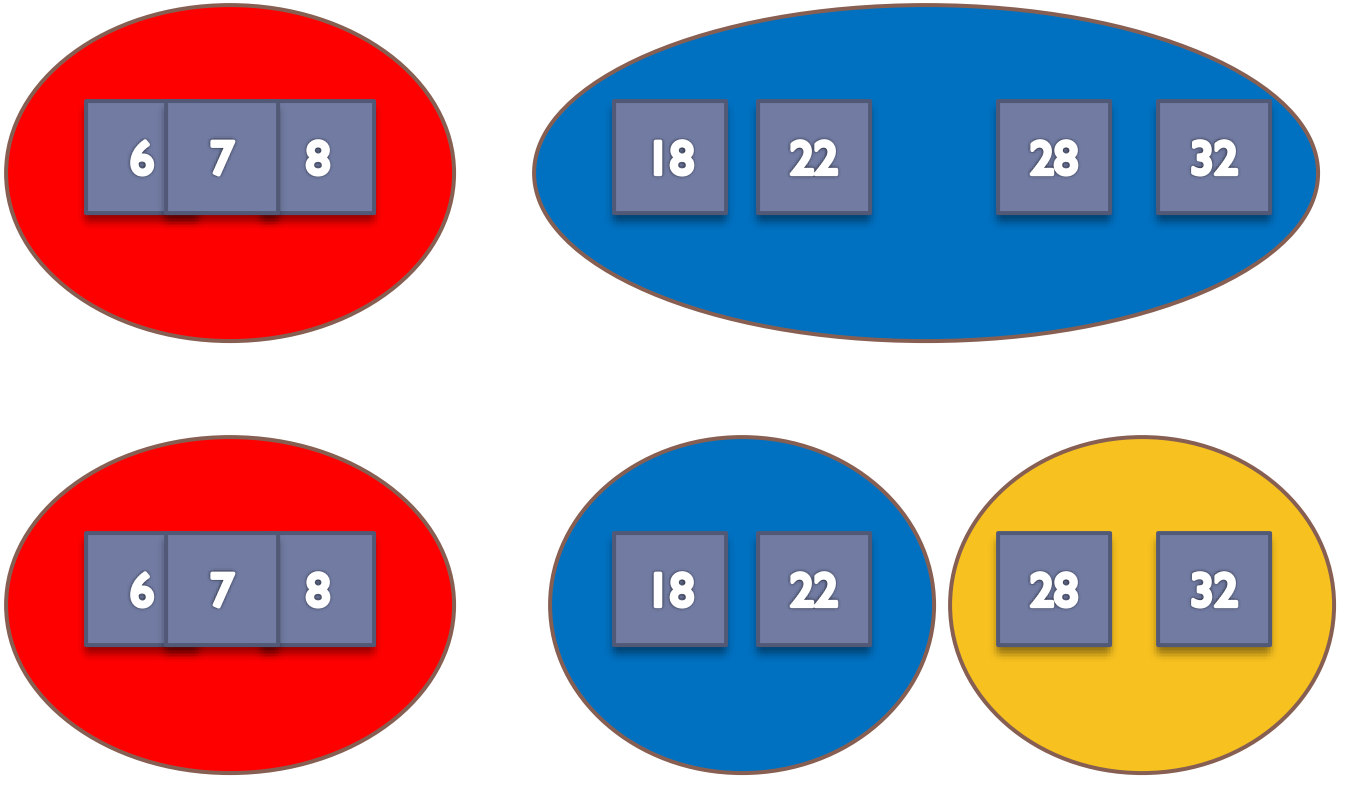
- Several possibilities, e.g.,
  - A fixed number of iterations
  - Centroid positions don't change (can be proven to converge)
  - Clusters look reasonable



Cluster validity



Cluster validity



Cluster validity: what we want!

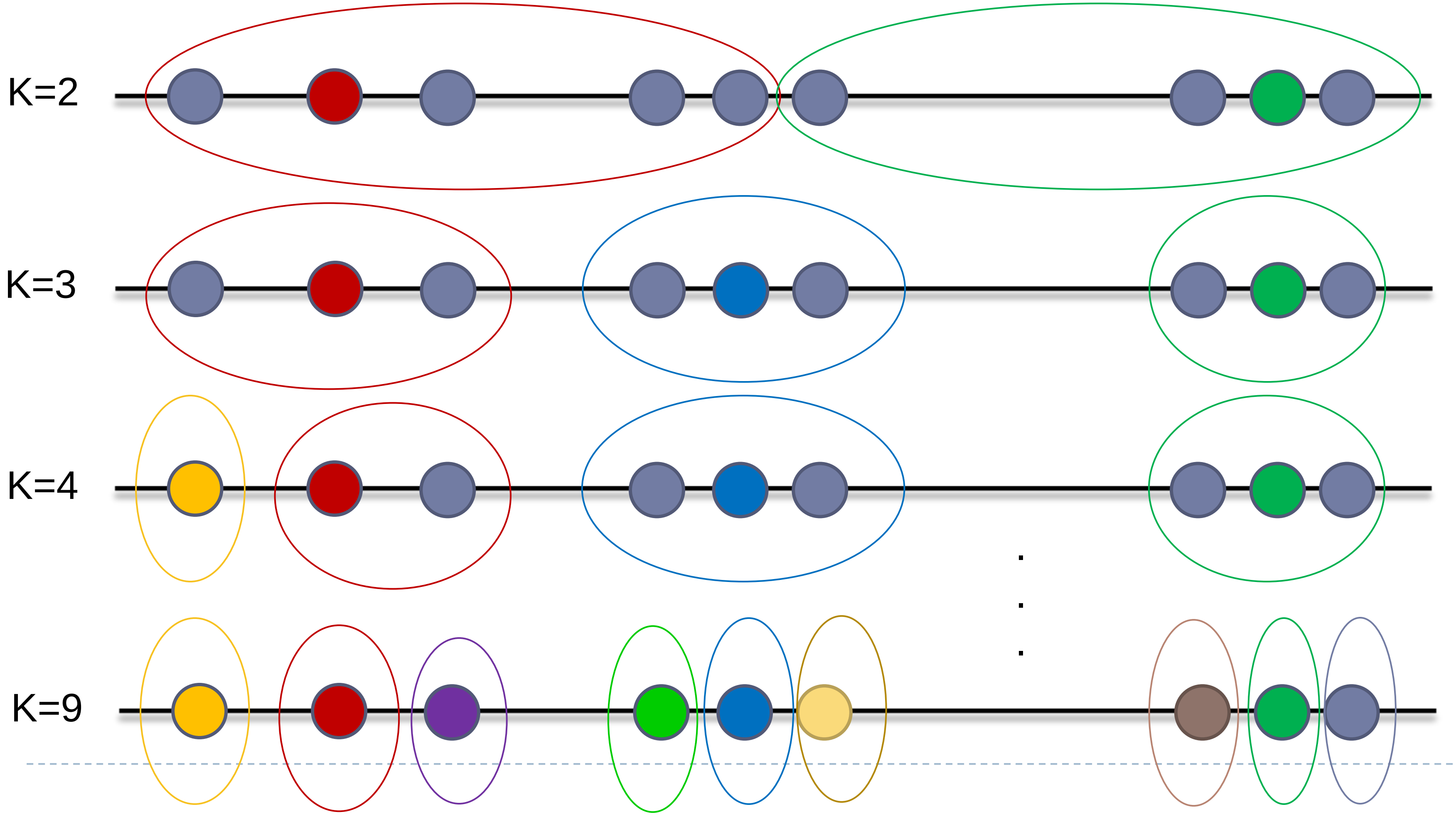
- High **inter-cluster** distances
  - Large distance between clusters
  - Otherwise known as good *separability*

- Low **intra-cluster** distances
  - Distances between data points within a cluster should be relatively low
  - Otherwise known as good *compactness*
  - *Adequate distortion (e.g. mean distance between centroid and points)*

- Many cluster validity measures have been developed, often based on these distances
  - But beyond the scope of this module

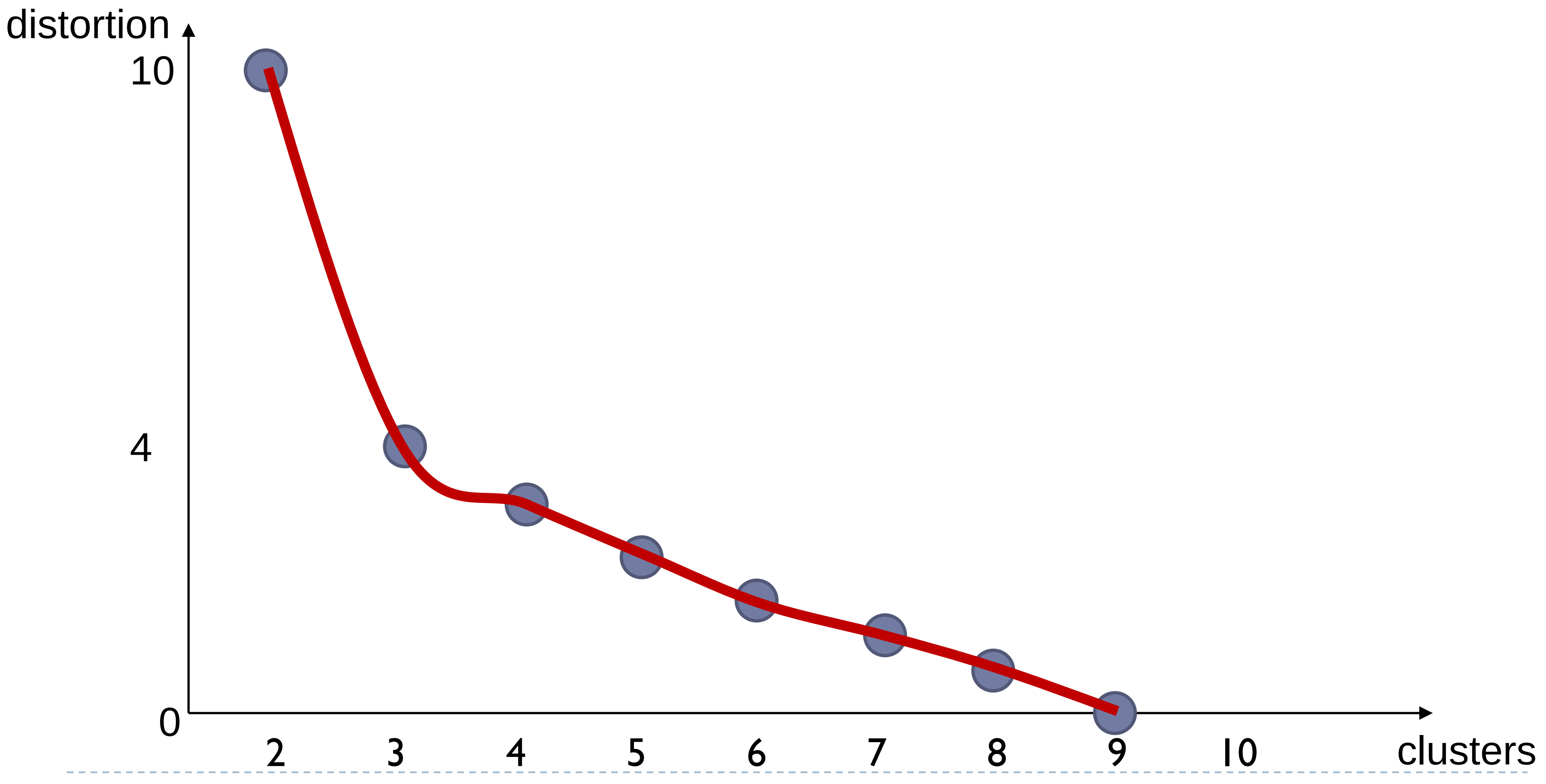
Find optimal k – elbow technique

- Run K-means for multiple number of clusters and plot the cluster distortion



Find optimal k – elbow technique

- Run K-means for multiple number of clusters and plot the cluster distortion

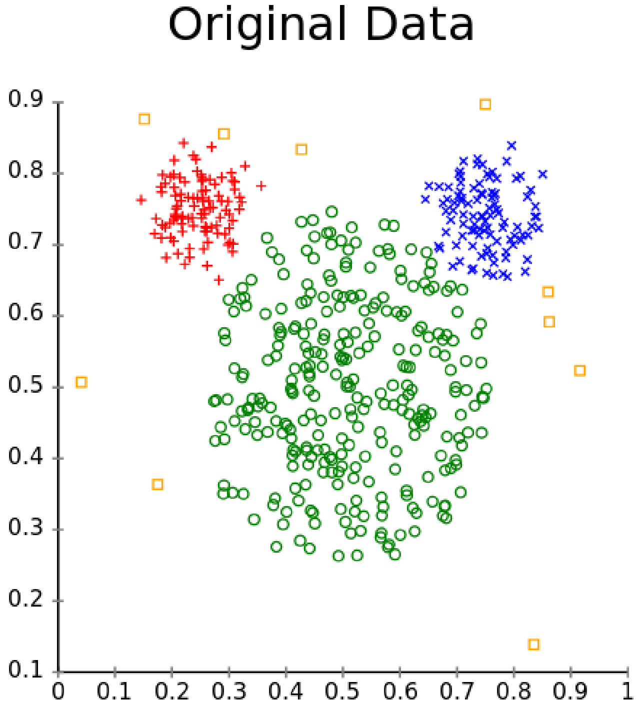


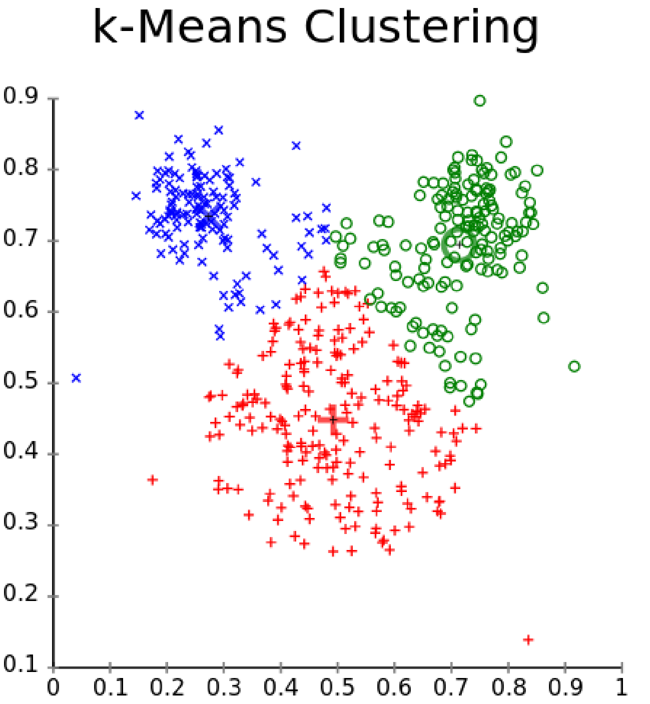
Limitations

- Must choose parameter *k* in advance, or try many values
  - This is a particular problem for *k*-means as often the optimal number of clusters is not known
- Data must be numerical and must be compared via a suitable distance measure
  - E.g. ‘Fruit’ feature: how do you compare apples and oranges? How far is a banana from a pineapple?
- The algorithm is sensitive to outliers/points which do not belong in any cluster
  - These can distort the centroid positions and ruin the clustering

Limitations

- The algorithm works best on data which contains spherical clusters; clusters with other geometry may not be found
  - Even then, it tends to generate clusters of similar size...





Next…

- Hierarchical clustering