Clustering

Data Mining & Analytics

Prof. Zach Pardos

INFO254/154: Spring ’19

Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

o, p C_i

Clustering: Terminology

Instance, row, data point, object,
o, p

cluster, group, partition
C_i

Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

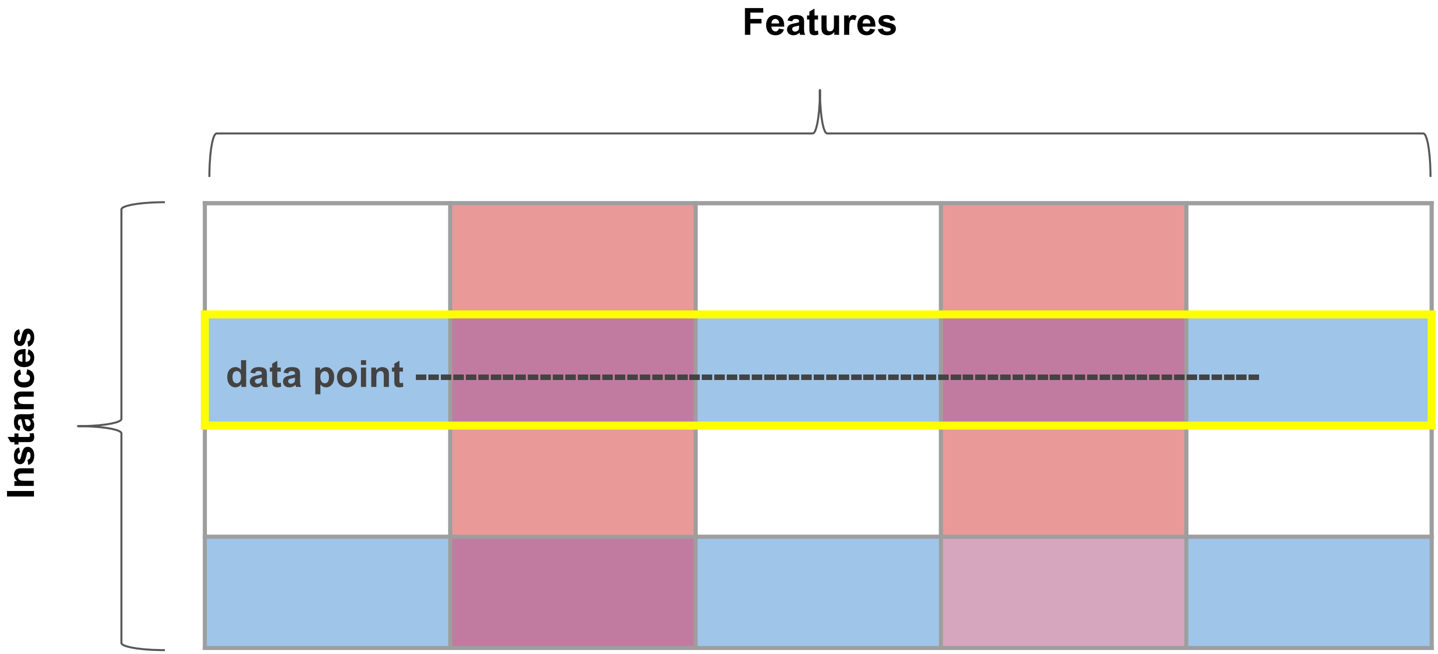
o, p C_i

Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

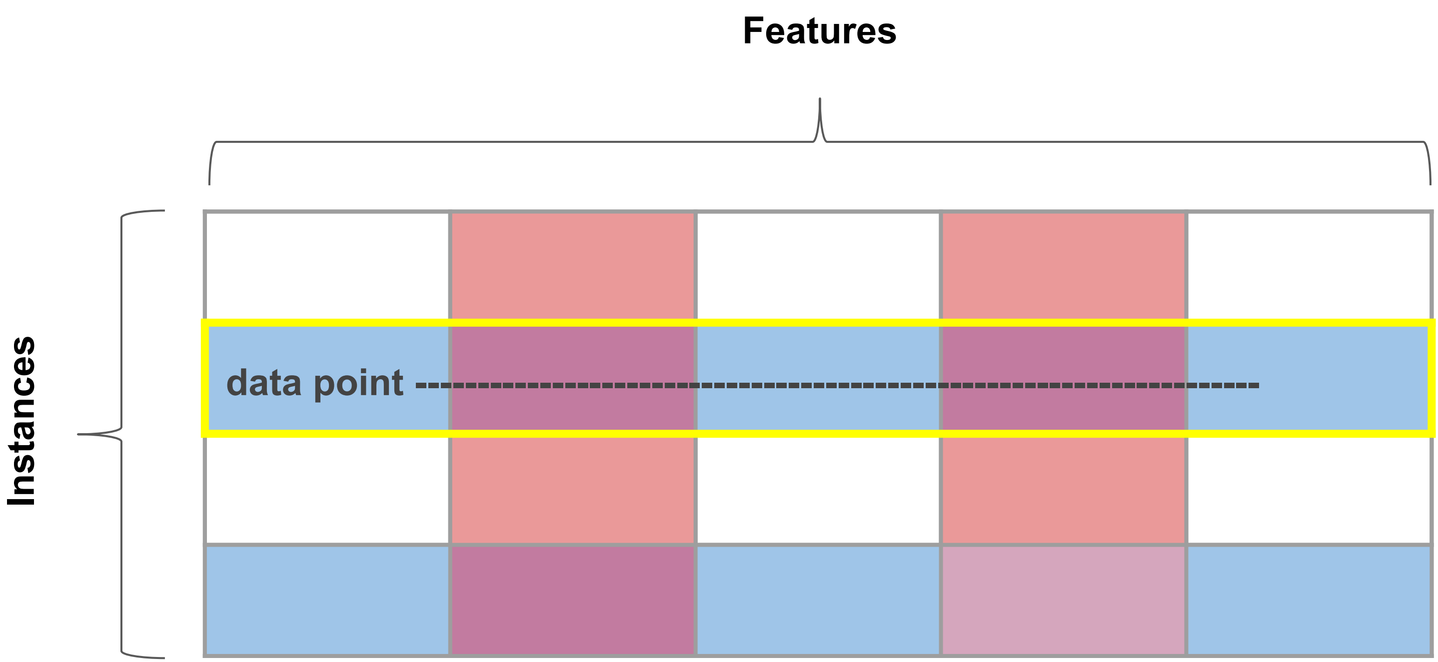
Clustering: Terminology

Instance, row, data point, object, cluster, group, partition



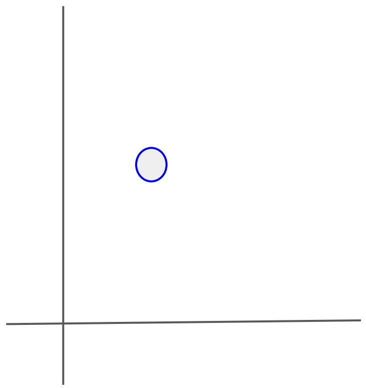
Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

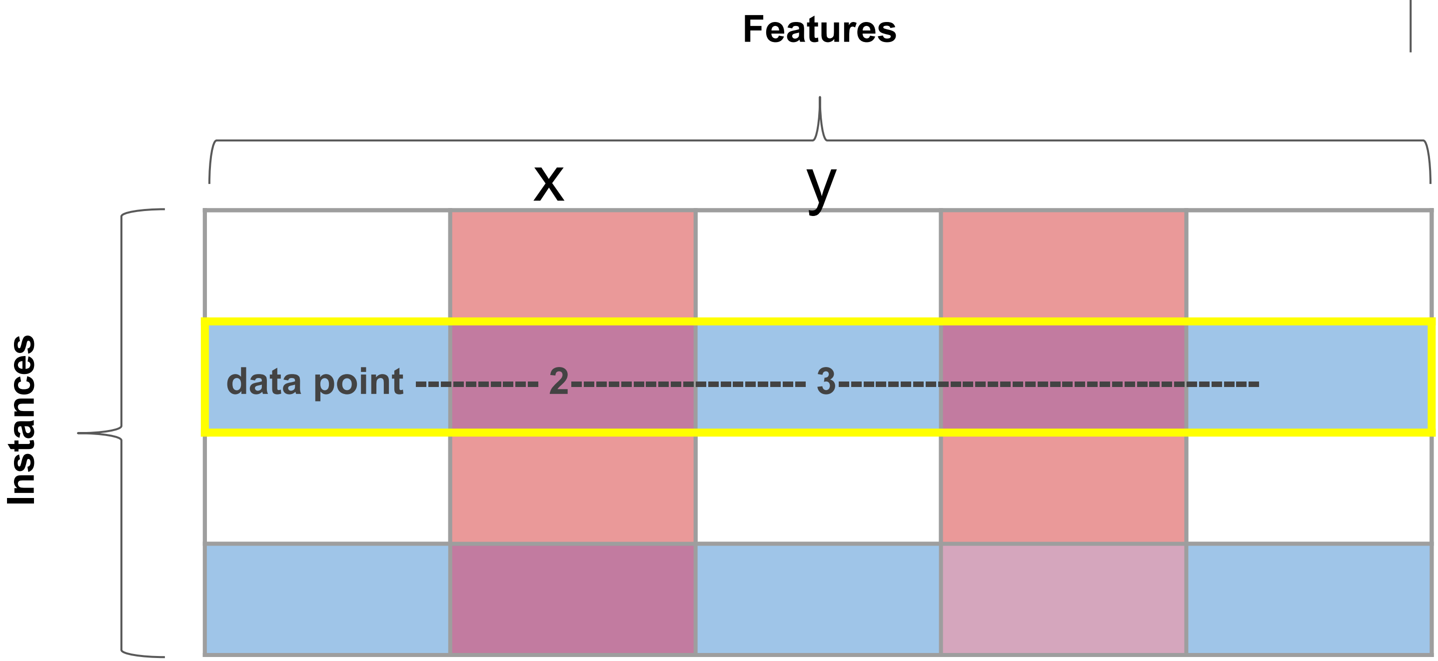


Why “data point”? Think Euclidean space

Clustering: Terminology



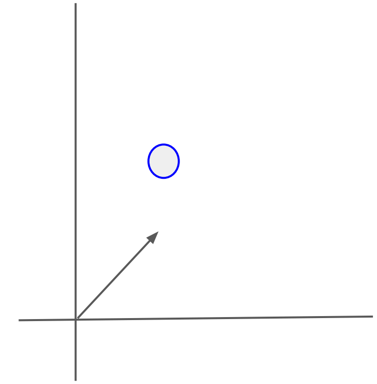
Instance, row, data point, object, cluster, group, partition

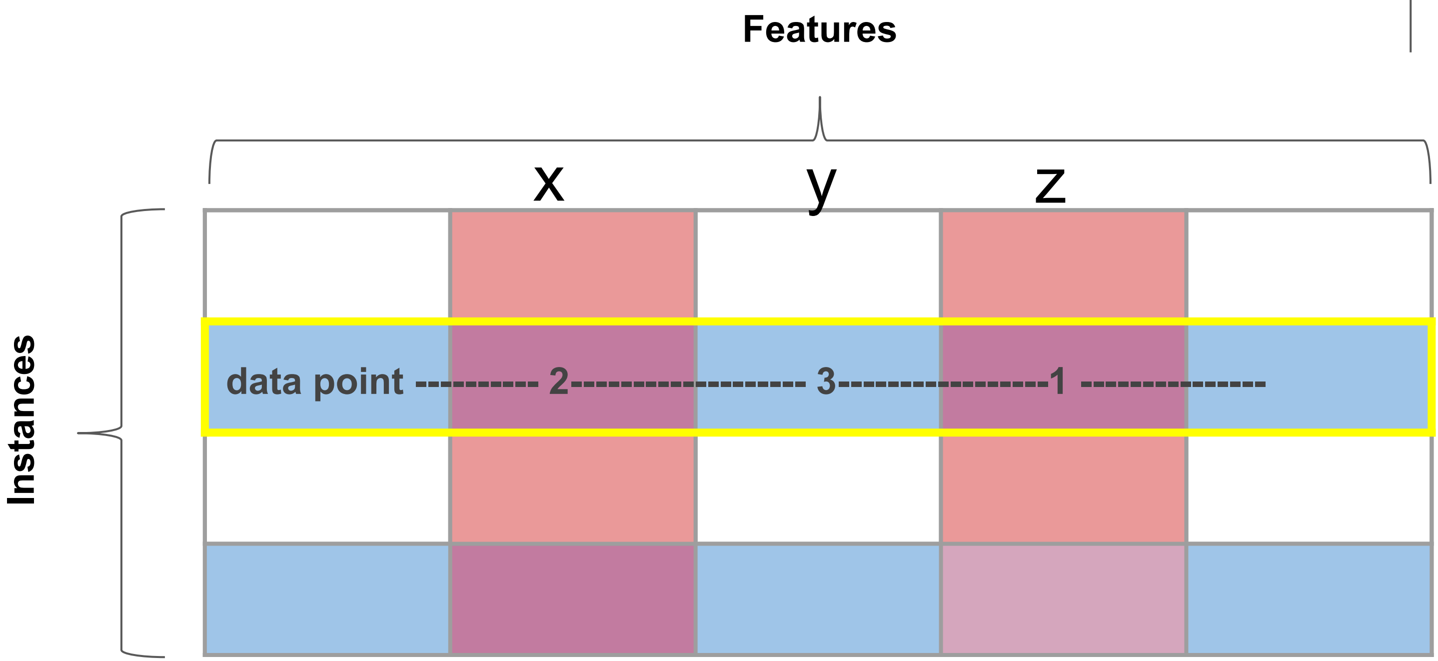


Why “data point”? Think Euclidean space

Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

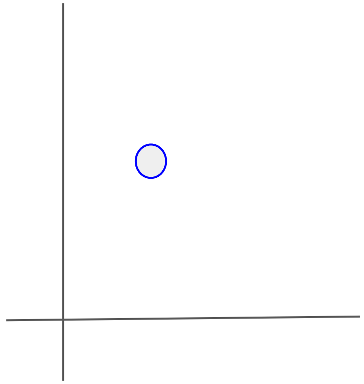


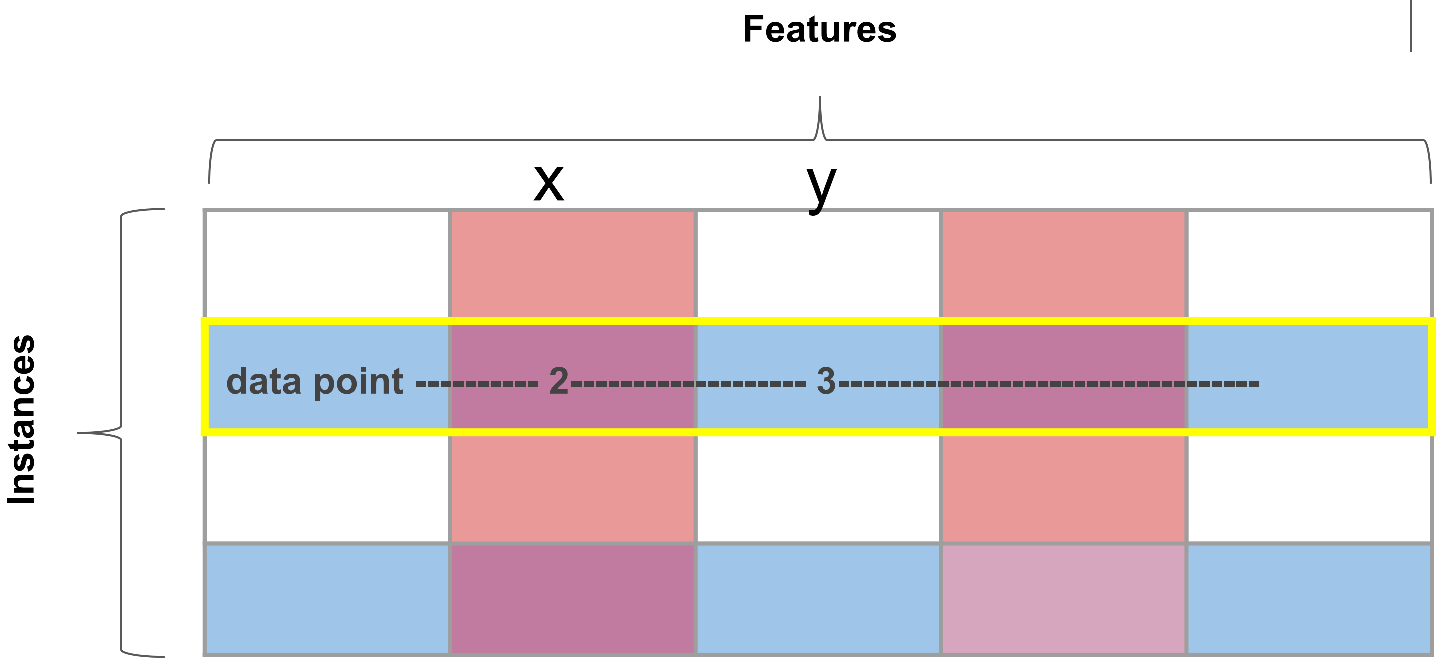


Why “data point”? Think Euclidean space

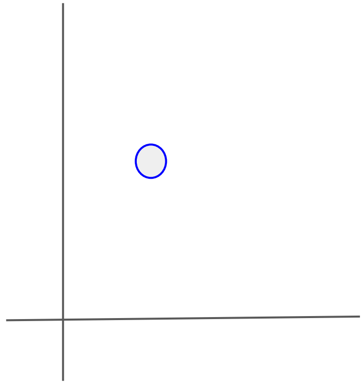
Clustering: Terminology

Instance, row, data point, object, cluster, group, partition

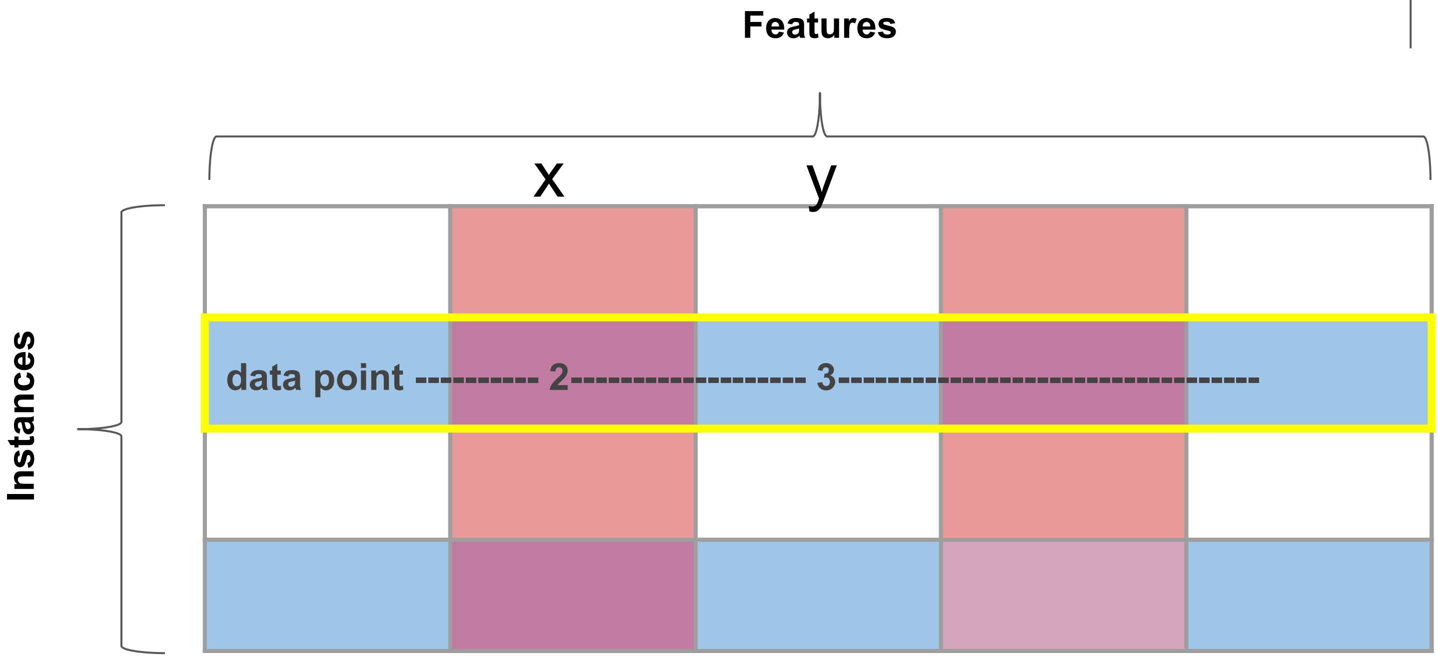




Clustering: Theory

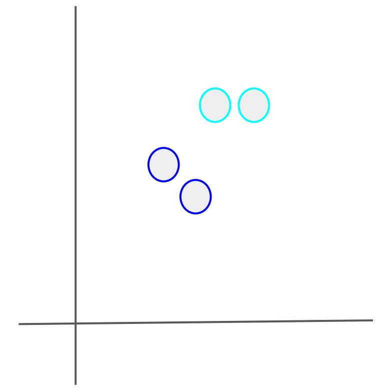


Instance, row, data point, object, cluster, group, partition

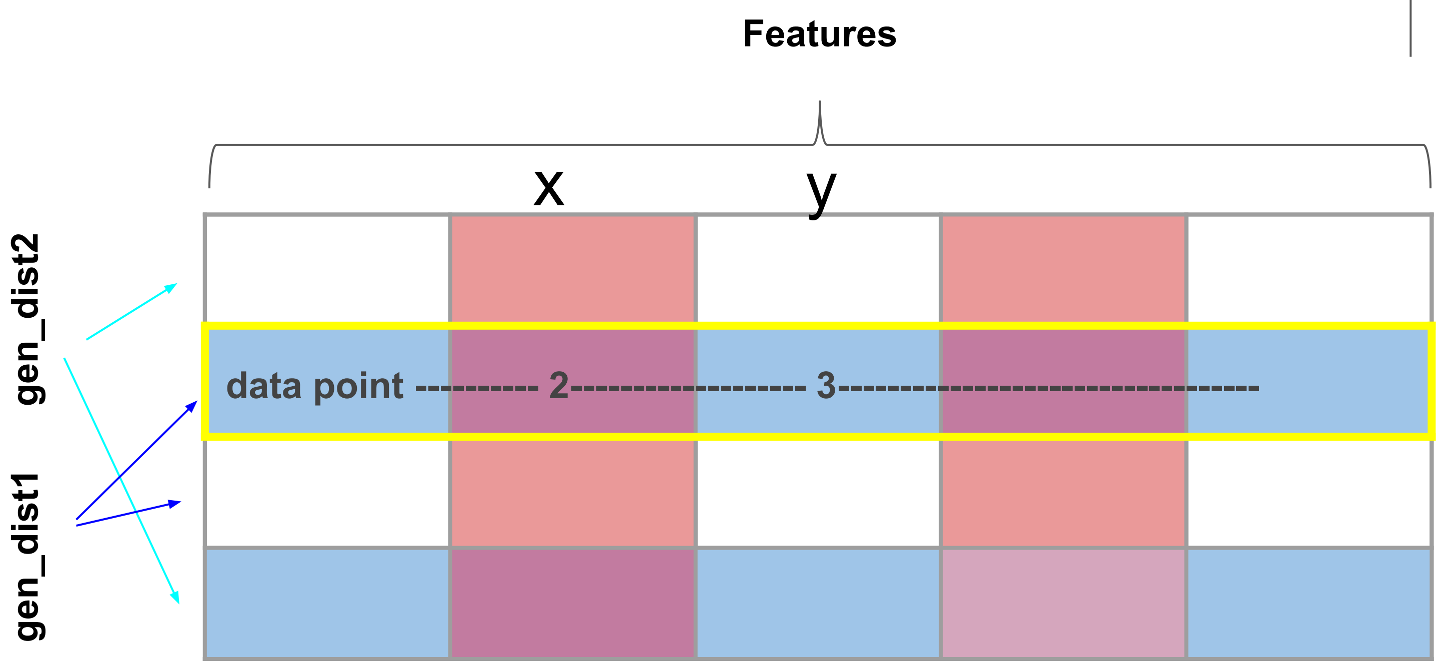


What is the hypothesis behind clustering?

Clustering: Theory



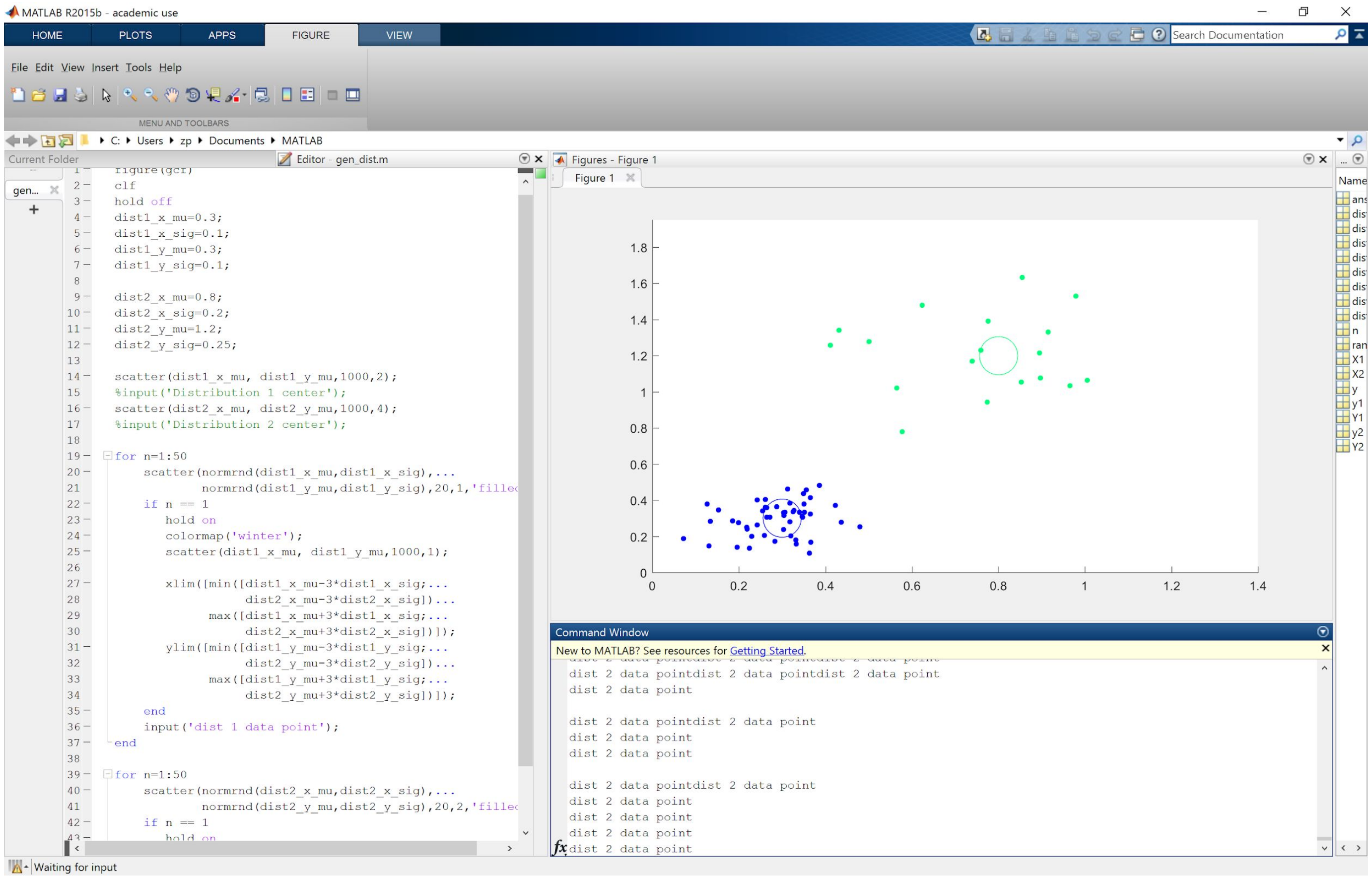
Instance, row, data point, object, cluster, group, partition



What is the hypothesis behind clustering?

That there is a set (K) of generating distributions from which the data were created

Clustering: MATLAB Demo



[link to example code (MATLAB)]

Clustering (in-class exercise)

Height

Wearing glasses?

Predominant color of clothing

How did you balance the 3 values?
How did you choose K?

Clustering

Instance, row, data point, object, cluster, group, partition

- Classification:
  - grouping data points with respect to a target
- Clustering:
  - grouping data points with respect to a similarity metric

Clustering

Instance, row, data point, object, cluster, group, partition

Clustering

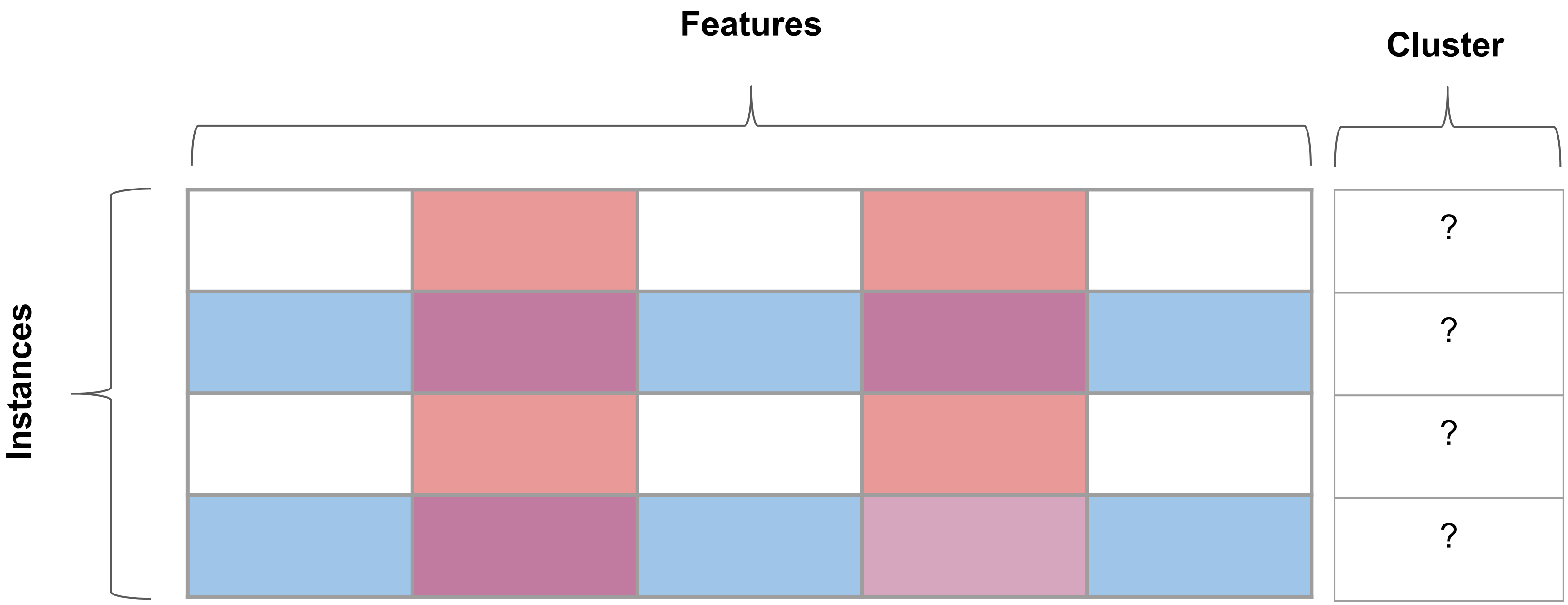
Instance, rows, feature, attribute, column, target, label

The cluster labels may be:

1) An existing feature included in the clustering

Clustering

Instance, rows, feature, attribute, column, target, label

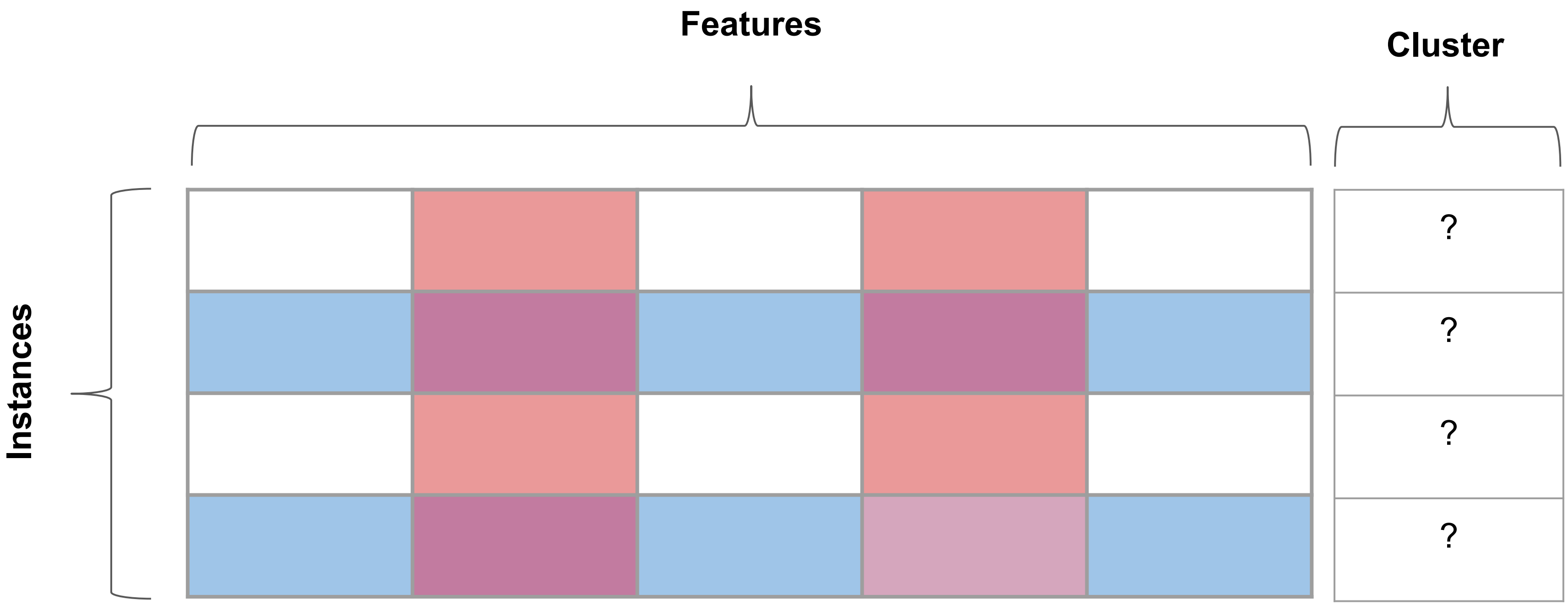


The cluster labels may be:

1) An existing feature included in the clustering
2) An existing feature not included in the clustering (i.e.target)

Clustering

Instance, rows, feature, attribute, column, target, label



The cluster labels may be:

1) An existing feature included in the clustering
2) An existing feature not included in the clustering (i.e.target)
3) A latent attribute you don't have direct access to

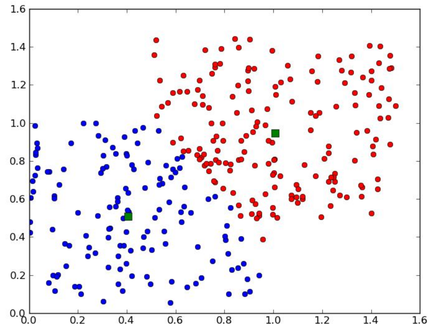
Remaining lecture outline:

- Types of clustering methods
- Intrinsic measures of the goodness of a clustering
- The k-means algorithm
- Heuristics for choosing K

Clustering: Types of clustering methods

Types

Partitioning



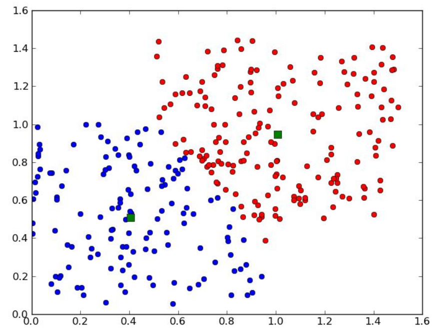
- Find mutually exclusive clusters of (hyper) spherical shape
- Distance-based
- May use mean or medoid to represent cluster center

Han, Camber, Pei (2011), Sec 10.1

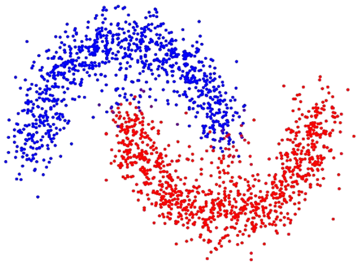
Clustering: Types of clustering methods

Types

Partitioning



Density-based

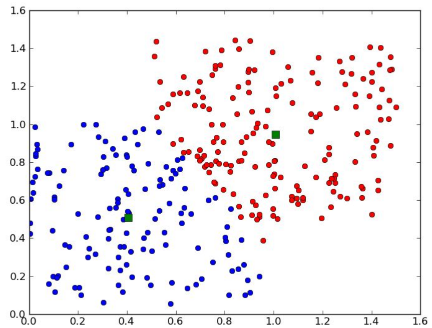


Han, Camber, Pei (2011), Sec 10.1

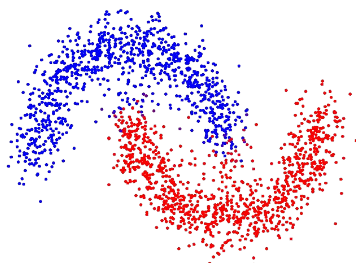
Clustering: Types of clustering methods

Types

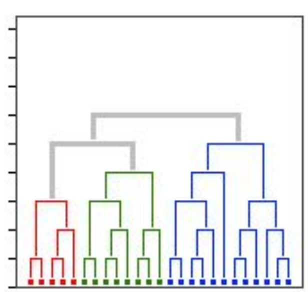
Partitioning



Density-based



Hierarchical

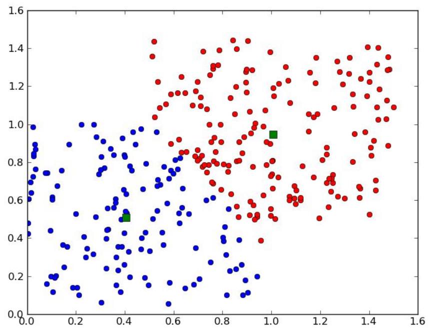


Han, Camber, Pei (2011), Sec 10.1

Clustering: Types of clustering methods

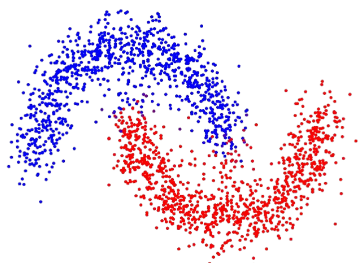
Types

Partitioning



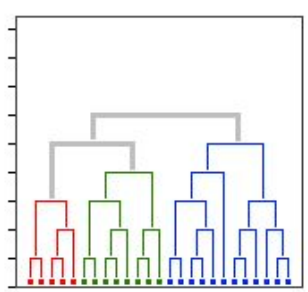
k-means
(covered today)

Density-based



spectral
(covered later in semester)

Hierarchical



Han, Camber, Pei (2011), Sec 10.1

Clustering: Terminology

Instance, row, <u>data point</u>, object, *o*, *p*

<u>cluster</u>, group, partition C_i

Clustering (partitioning) formal definition:

D is a dataset containing *n* data points which can be represented in euclidean space

Partitioning distributes data points in *D* into *k* clusters, $C_1, \ldots, C_k$ such that $C_i \subset D$ and $C_i \cap Cj = \varnothing$ *for 1 <= i,j <= k*

Clustering: Measuring the goodness of a Clustering

Within-cluster Variance

- The sum of squared error between data points and their respective cluster center
- The lower the sum the higher quality the clustering
- Brute-force in this scenario is prohibitively expensive
  - How expensive?
  - What happens as k reaches |*D*|?

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

Clustering: K-means algorithm

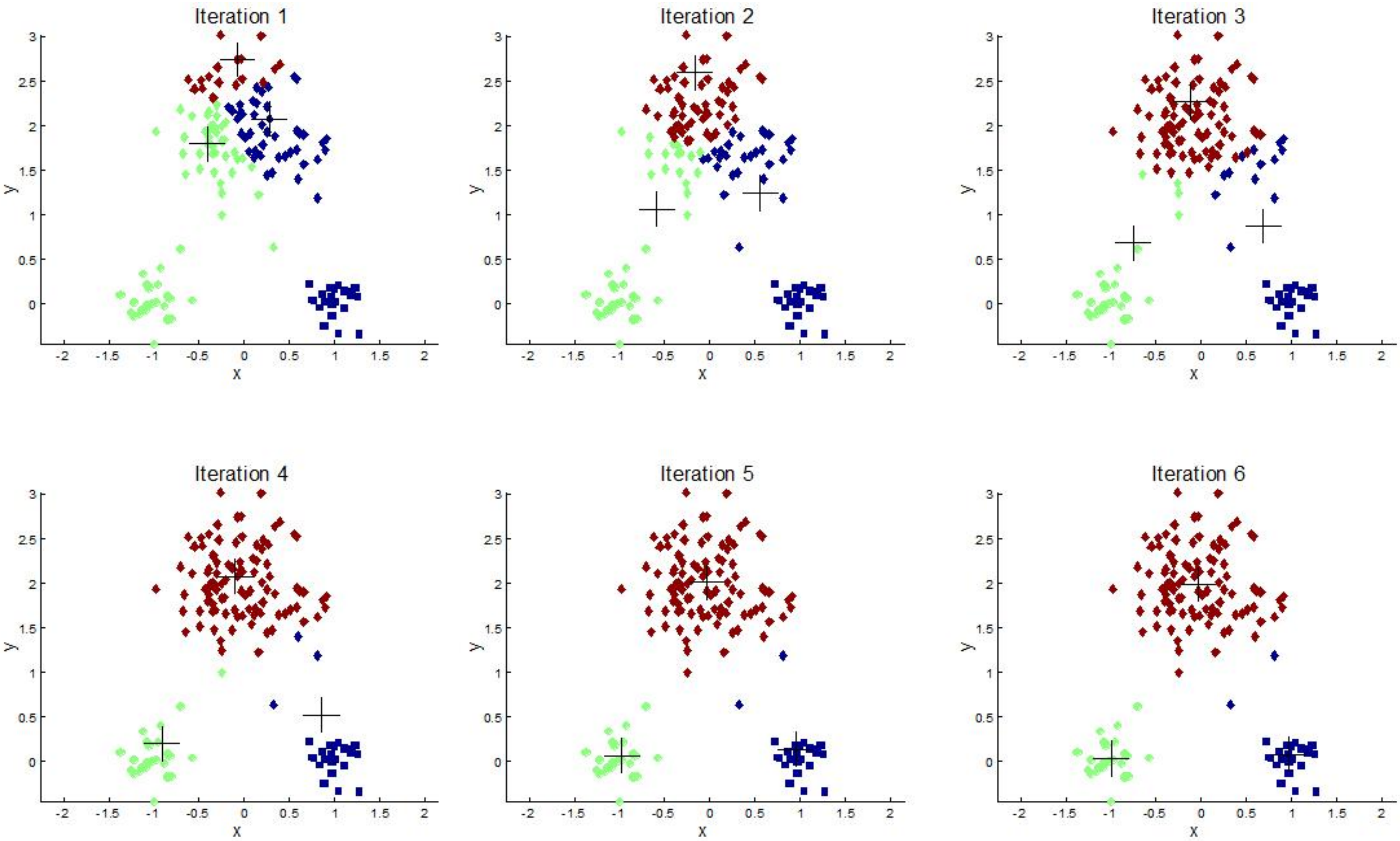
algorithm

Take as input: ***k*** the number of clusters and ***D*** a set of data points

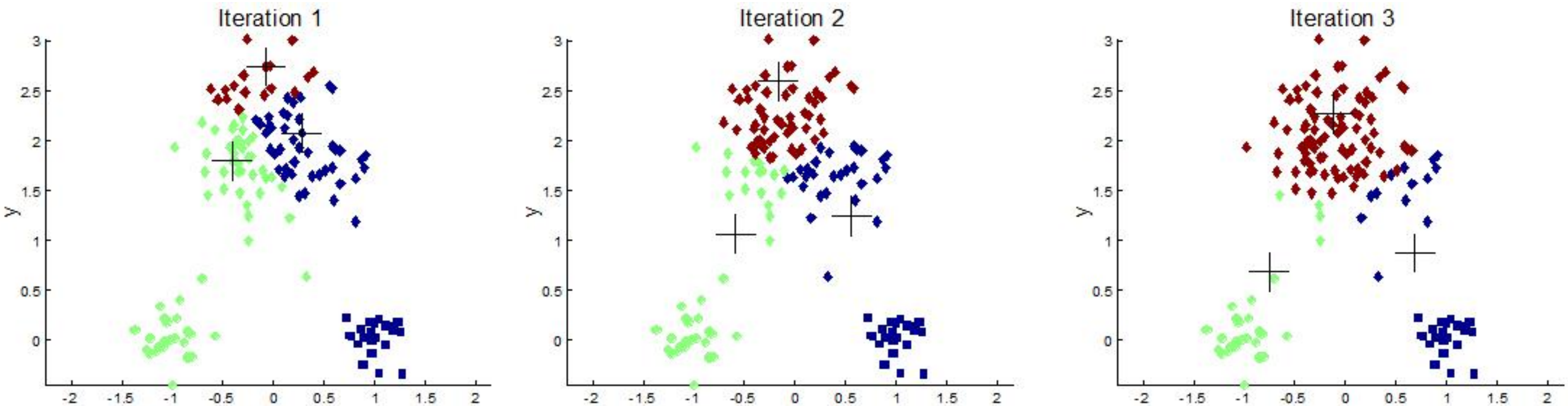
1. Start by randomly choosing k data points to serve as initial centroids
2. Until there is no change in cluster assignments (or set a max iteration):
    i. (re)assign each data point to the cluster centroid to which the data point is closest to in euclidean space
    ii. update the cluster centroid values to represent the means of the data points of each cluster
3. Return the cluster membership for all data points

Han, Camber, Pei (2011), Sec 10.2

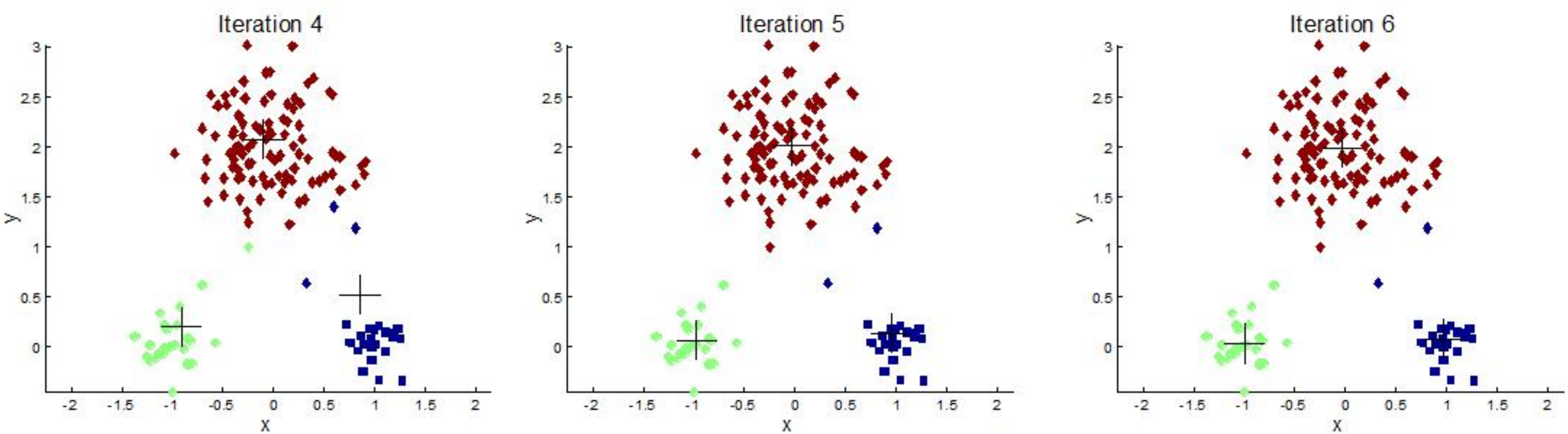
Clustering: K-means algorithm



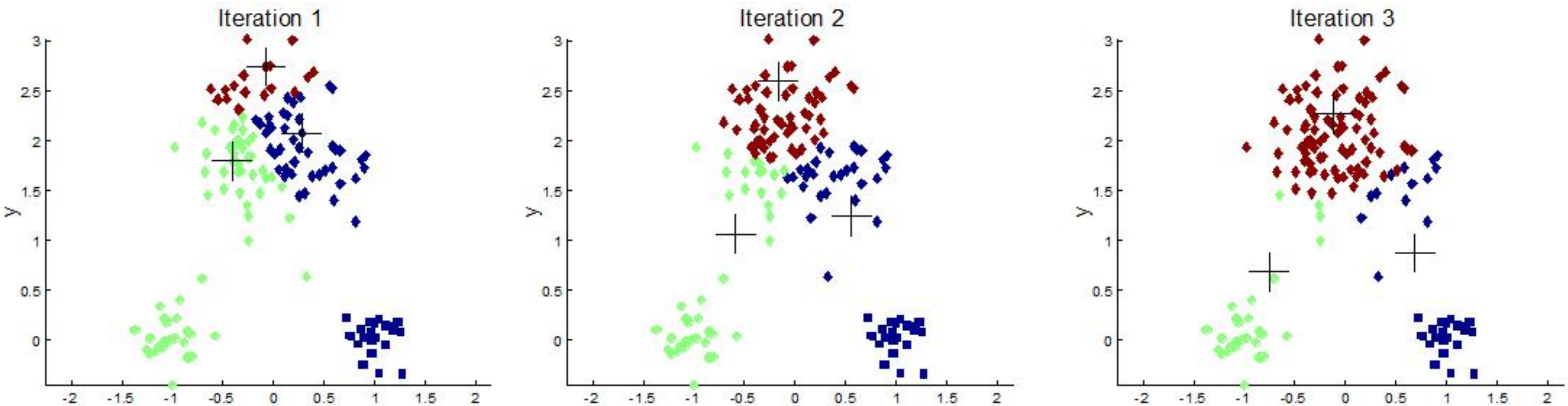
Clustering: K-means algorithm



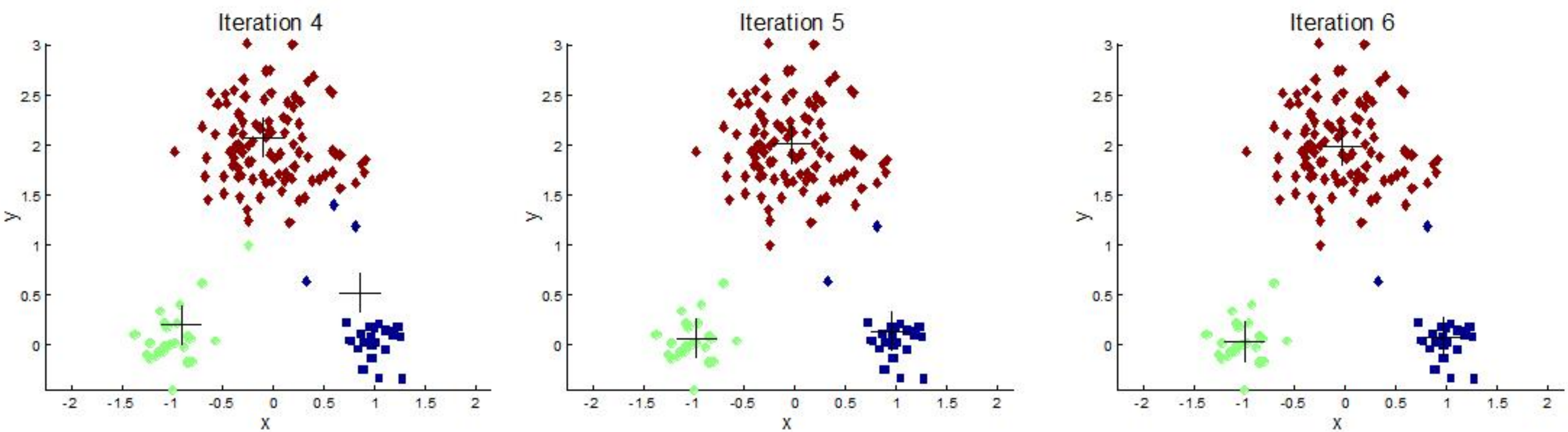
Can SSE be used (for a fixed K) in this scenario?



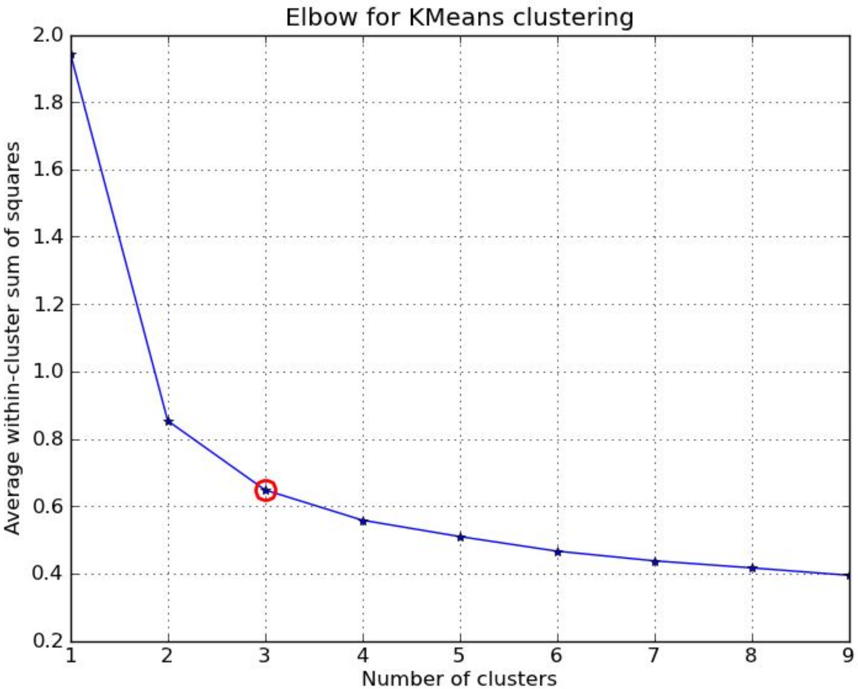
Clustering: K-means algorithm



Can SSE be used to choose a K?



Clustering: Choosing K (heuristics)



$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

<u>Sum of Squared Errors</u>

The Elbow method

- Choose a range of K
- Run K-means for every K in your range
- After each run, calculate sum of squared errors
- Also calculate the change in slope between each consecutive sum
- The “elbow” aka “turning point” is the run where the largest difference in slope is calculated
- Why not use K = #data points?

Clustering: Choosing K (heuristics)

The silhouette score

- For each data point
  - Calculate the average distance between the data point and all of the members of its cluster a(o)

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o, o')}{|C_i| - 1}$$

  - Calculate the minimum average distance between the data point and members of the other clusters

$$b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\}$$

  - Calculate silhouette coefficient

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}.$$

  - Average the coefficients of every data point to calculate the silhouette score

Clustering: Choosing K (heuristics)

The silhouette score

- For each data point
  - Calculate the average distance between the data point and all of the members of its cluster a(o)

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o, o')}{|C_i| - 1}$$

  - Calculate the minimum average distance between the data point and members of the other clusters

$$b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\}$$

  - Calculate silhouette coefficient

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}.$$

  - Average the coefficients of every data point

What is the range of the coefficient/score?

Clustering: Tips

- Silhouette coefficient
  - Does not try to achieve equal cluster sizes
  - Vulnerable to placing outlier data points in their own cluster to maximize coefficient

- Elbow method
  - Error goes to zero as K approaches the number of data points

- Clustering (in general)
  - Sensitive to the scale of the feature. If one feature has a range of 0-100 (eg. age) and the others are between 0 and 1, the euclidean distance metric will be dominated by the distance between age.
  - Solution to this issue is to normalize all features to the same scale

Clustering:

- Ideally, captures generating distributions
- Practically, is an exploration of the structure of your dataset

Office Hours

Starts now in this classroom