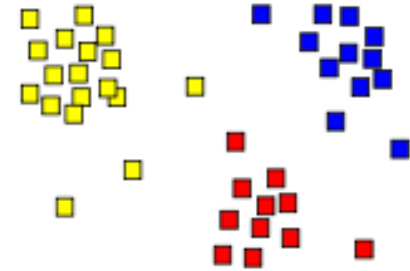# Clustering
CS1D6: Introduction to data and statistics

Dr. Fayyaz Minhas

Department of Computer Science

University of Warwick

The problem

- How to group objects?

- Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)

Simplest Clustering

- K-Means Clustering

- Represent each sample as a vector
- Each cluster will be represented by an "exemplar" vector
  - Can have "K" such examplars
- A given sample will be associated with the cluster whose exemplar is the closest to it in terms of Euclidean distance
- Question
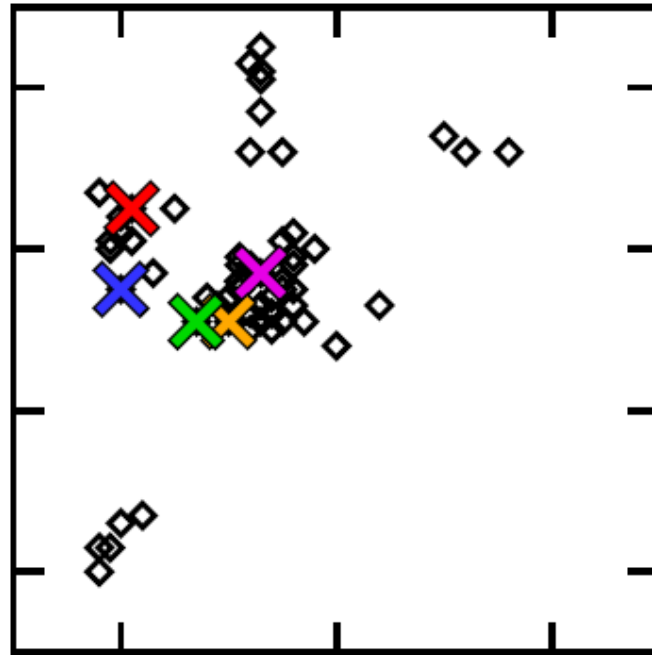 - How do we determine the exemplars?
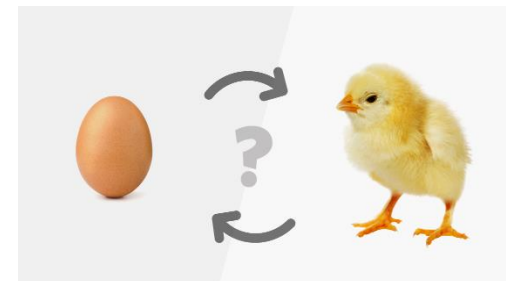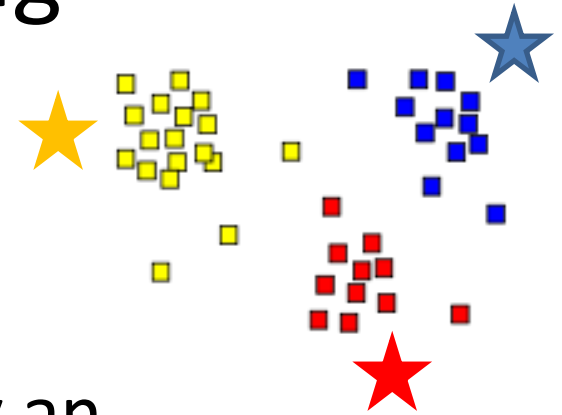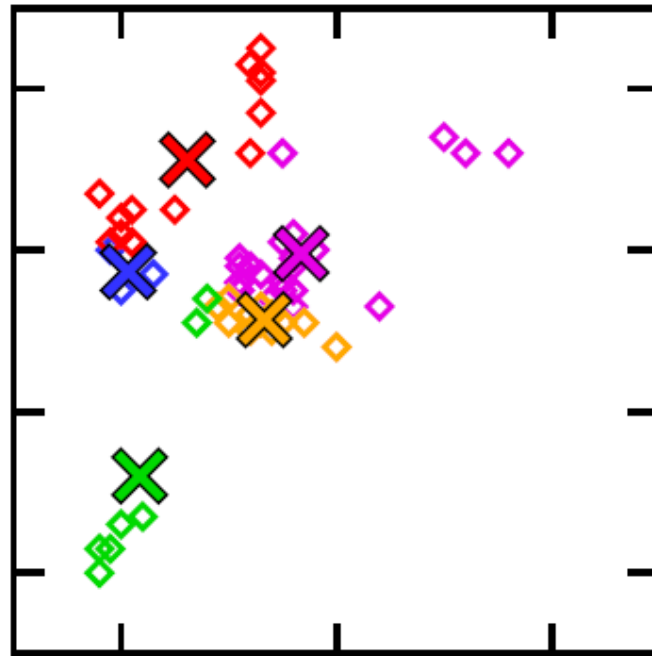
Simplest Clustering

- Initialize the exemplars at random

- Assign points to each cluster based on distance

- Update clusters by assigning the exemplars equal to the mean of the samples in that cluster

- Repeat until bored

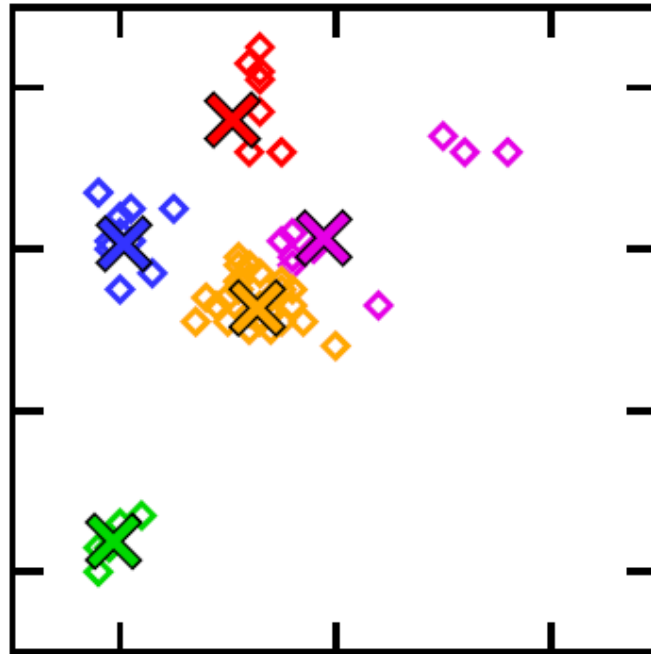https://en.wikipedia.org/wiki/K-means_clustering
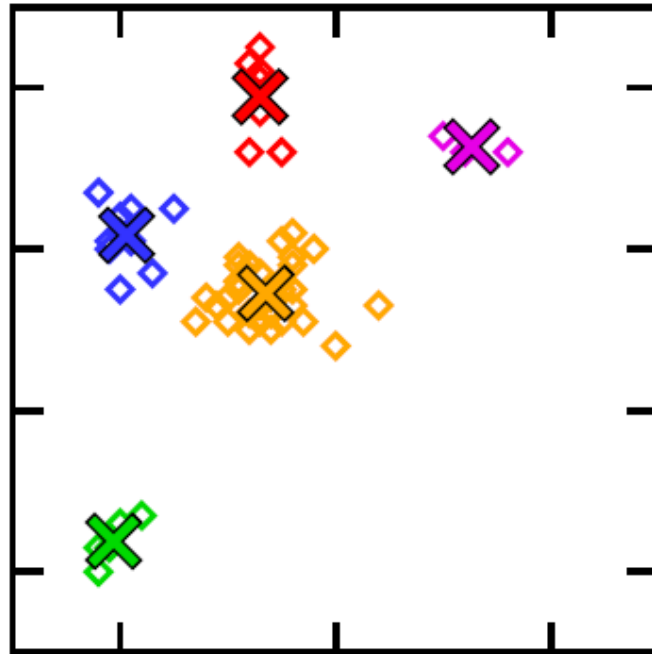
k-means Clustering
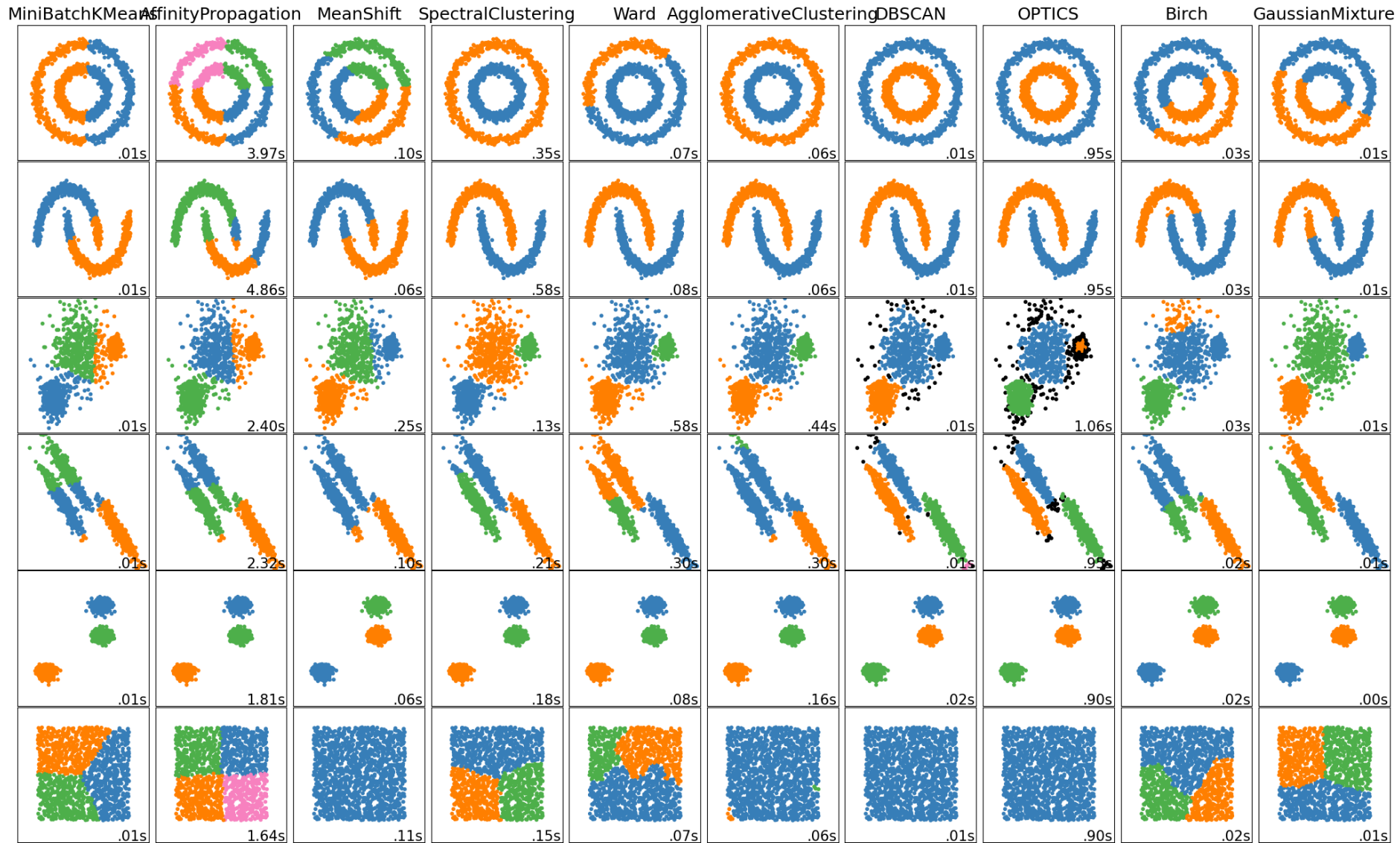
k-means Clustering

k-means Clustering

k-means Clustering

Issues

- Determining How Many Clusters?

- Data Scaling

- Dimensionality Reduction

Other Approaches

Example

- Can we cluster Iris flowers?

  - Can we identify what patterns are common across the three classes without knowing their original labels?

End of Lecture

We want to make a machine that will be proud of us.

- Danny Hillis