Reminders

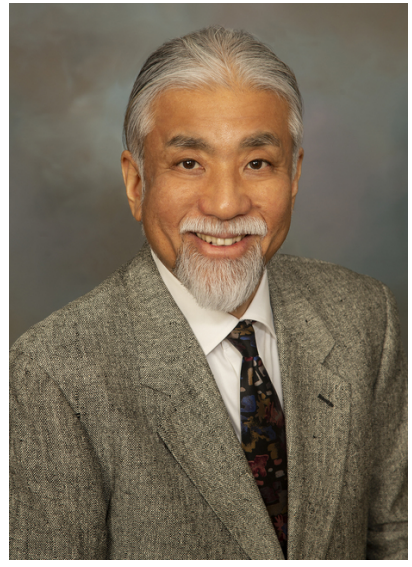
- Final project due June 8th at 11:59 pm, latest June 9th, 11:59 pm
- Submit as a single PDF to Gradescope; assignment for final project will be posted this weekend.

Clustering and the K-means method

Based on Prof. Naoki Saito's notes for MAT167

Melissa Zhang, MAT 167, UC Davis

Lecture 25



Reference

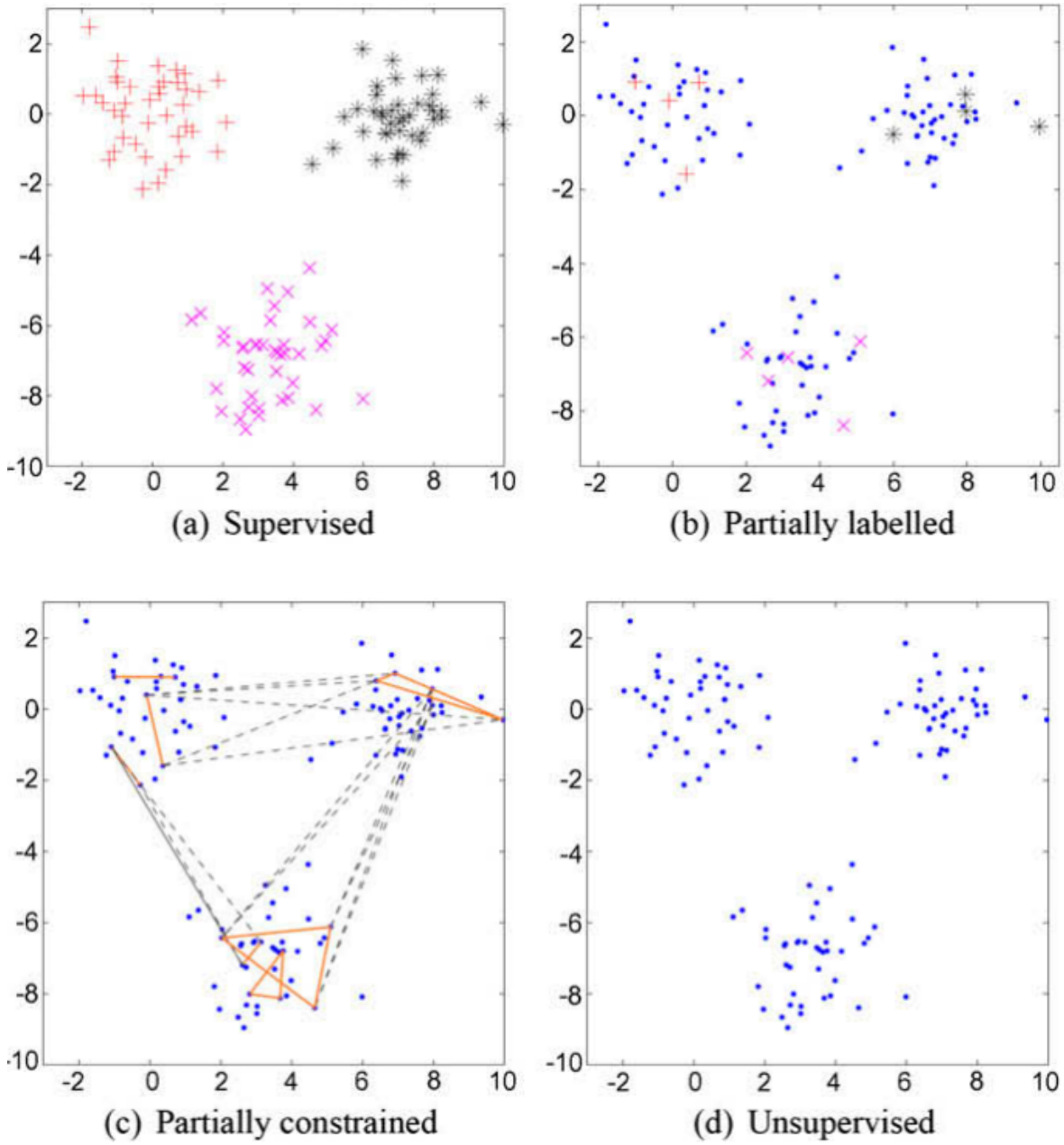
The images from this lecture are from:

Anil K. Jain, **Data clustering: 50 years beyond K-means**, *Pattern Recognition Letters, Volume 31, Issue 8, 2010, Pages 651-666, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2009.09.011. (https://www.sciencedirect.com/science/article/pii/S0167865509002323)*

Why Data Clustering?

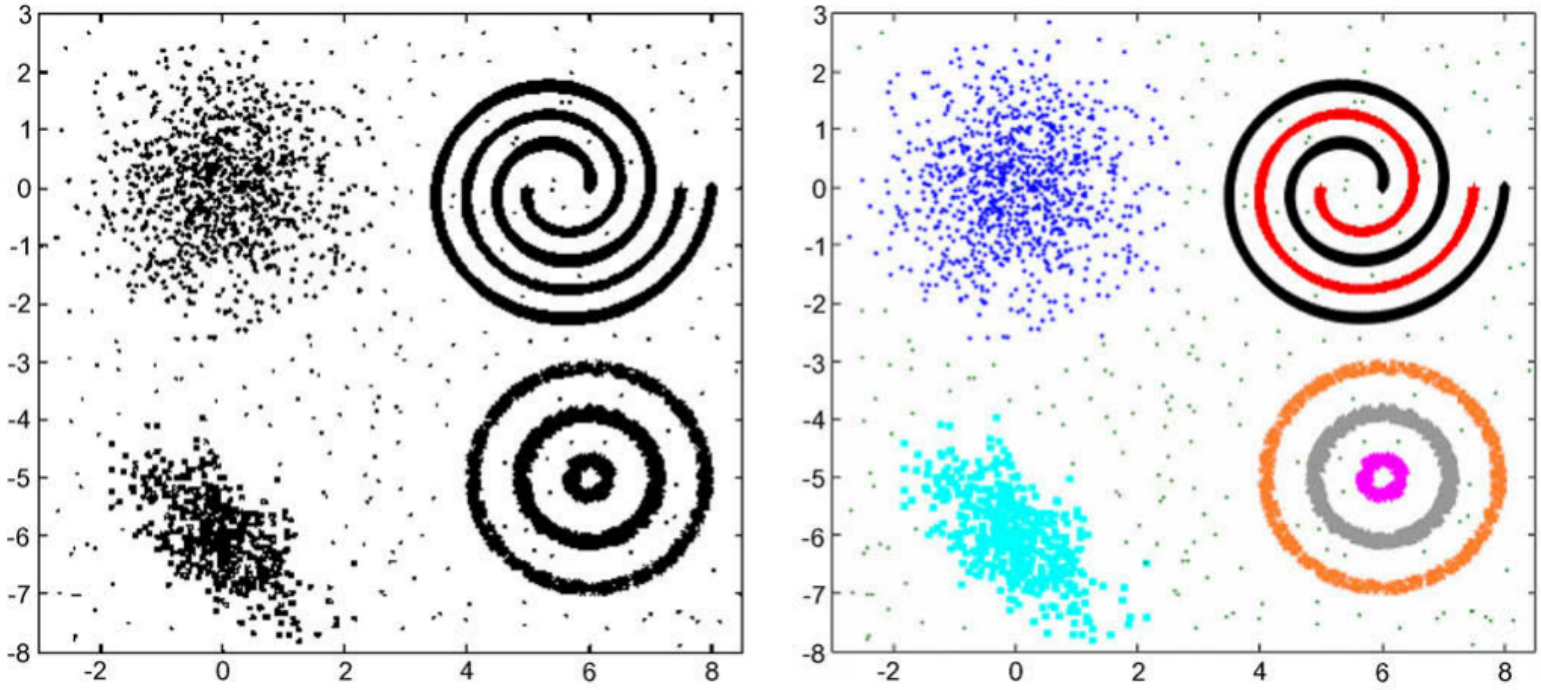
- Gain insight into the **underlying structure** of a large volume of data
  - exploration of novel data
  - identify new phenomena
  - identify new relationships
- **Classify** by similarity
  - handwriting recognition (this Friday)
  - taxonomy, phylogenetic relationships
  - populations and age brackets to advertise to
- **compression** of data
  - organize and summarize data
  - create cluster prototypes

Data Clustering $\subset$ Machine Learning



Data Clustering $\subset$ Machine Learning

Just using the simplest algorithm without thinking about the data might not get you the best results, though.



(a) Input data

(b) Desired clustering

We'll discuss how to work with this issue later today.

The K-means algorithm

- most popular
- quite simple!
- tried and true, still used in practice despite decades of research into other clustering algorithms

The K-means algorithm

- **Setup:** Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_j \in \mathbb{R}^d$ for all $1 \leq j \leq n$.
- We want to group them into a set of K clusters ($K \ll n$), called $C = \{c_1, \ldots, c_k, \ldots, c_K\}$.
- The K-means algorithm finds a partition of the points such that *the squared error between the empirical mean of a cluster and the points in the cluster is minimized.*
- This is like a "least squares" fitting to K points in $\mathbb{R}^d$, as opposed to a line in $\mathbb{R}^2$.

The K-means algorithm

- $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_j \in \mathbb{R}^d$ for all $1 \leq j \leq n$
- $C = \{c_1, \ldots, c_k, \ldots, c_K\}$
- The K-means algorithm tries to minimize J below:
  Let μ_k = mean of cluster c_k.
  Define

$$J(c_k) := \sum_{\mathbf{x}_j \in c_k} \|\mathbf{x}_j - \mu_k\|^2$$

  and

$$J(C) := \sum_{k=1}^{K} J(c_k).$$

The K-means algorithm

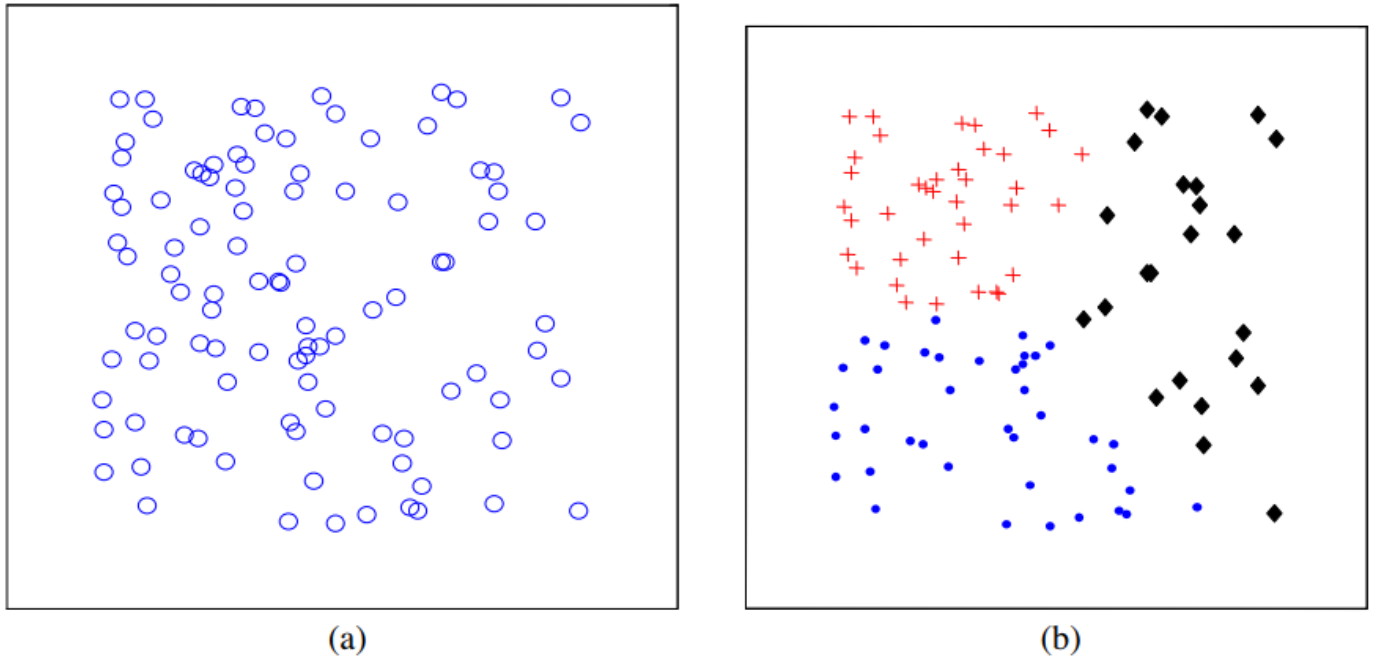


Fig. 8. Cluster validity. (a) A dataset with no "natural" clustering; (b) K-means partition with $K = 3$.

Note that if $K = n$, then $J = 0$; J always decreases as K grows.
We fix a small $K \ll n$.

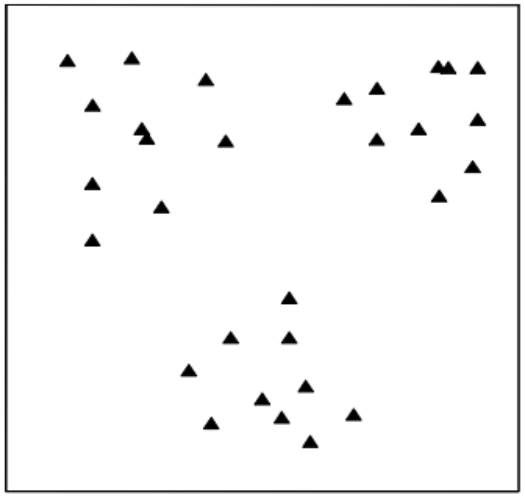
- for taxonomy, maybe set $K = 2$
- for handwriting digits, maybe set $K =$ around 10

The K-means algorithm

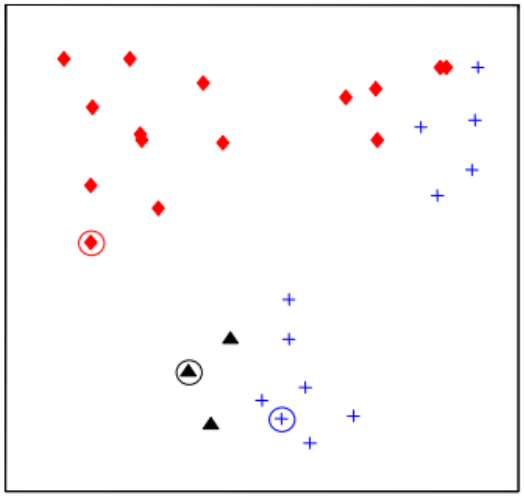
Here are the main steps:

1. Select an initial partition with K clusters
2. Generate a new partition by assigning each point (i.e., vector) to its closest cluster center
3. Compute new cluster centers
4. Repeat Steps 2 and 3 until cluster membership stabilizes.

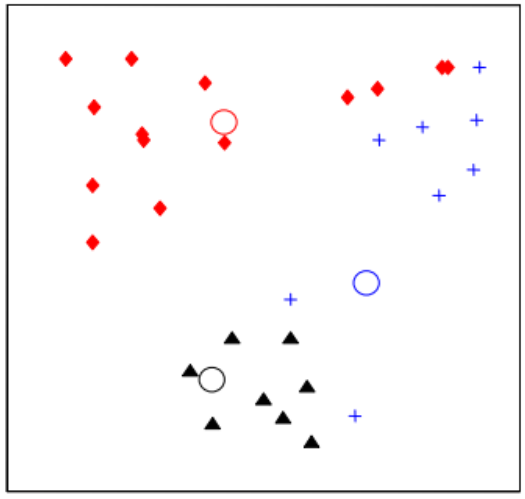
K-means algorithm example



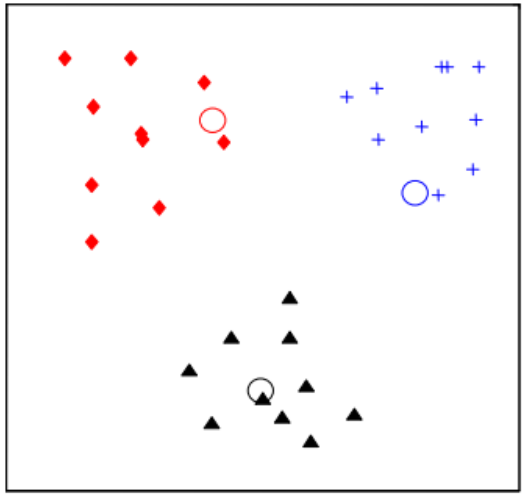
(a) Input data



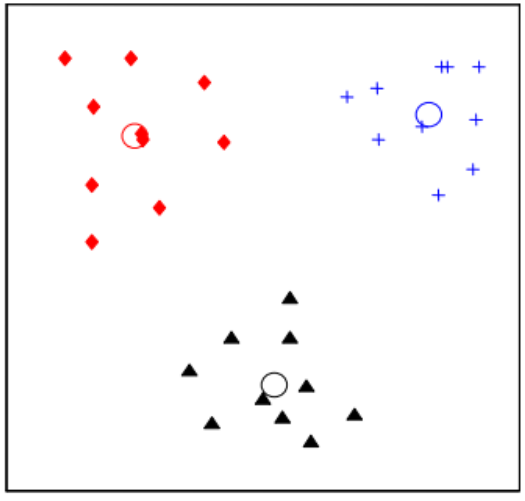
(b) Seed point selection



(c) Iteration 2



(d) Iteration 3



(e) Final clustering

The K-means algorithm: issues

Warning: This algorithm doesn't necessarily give you the *global minimum*, only a *local minimum*!

This type of minimization problem is known to be **NP-hard**, i.e.

- If you show tell me "correct" clustering that minimizes J, then I can certify that you are correct.
- However, in order to *find* the "correct" clustering, I would need to do an exhaustive search.

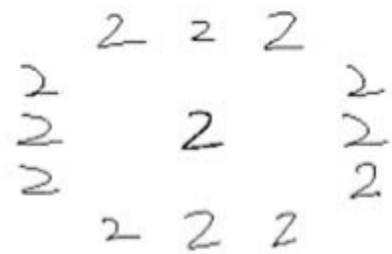
The K-means algorithm: issues

Other issues:

- How to preset K?
  - This could introduce bias at the start.
  - Or, you might just have too much data to even pick meaningful means.
  - You *could* randomly pick initial partitions, and run the algorithm many times...
- How many times should you run the algorithm to try to find the global minimum?

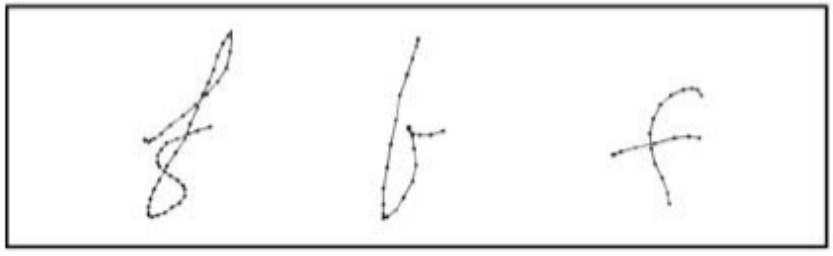
Handwriting variation

Just because there are 26 letters in English doesn't necessarily mean you should choose $K = 26$:

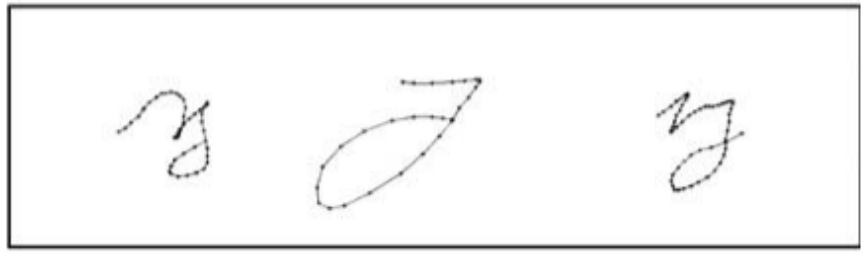


(a)

(b)



(c)



(d)

Clustering ensembles

What about shapes that are not blobs?

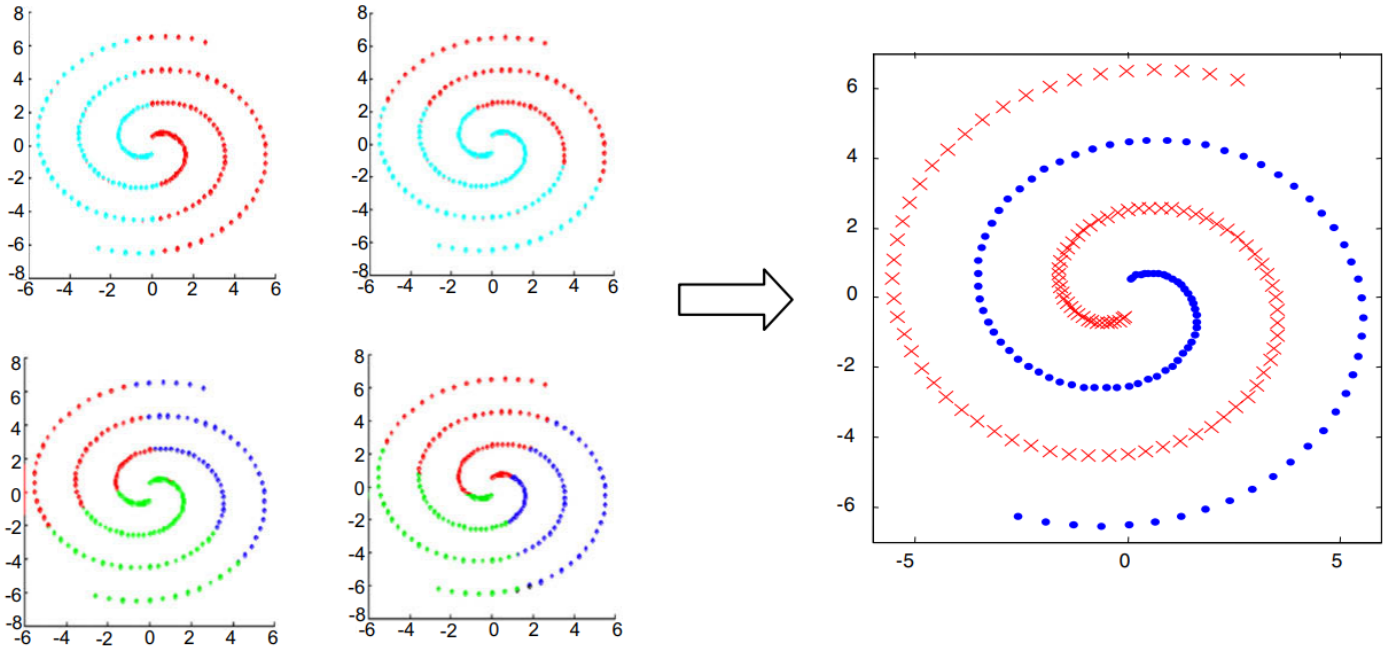


Fig. 11. Clustering ensembles. Multiple runs of K-means are used to learn the pair-wise similarity using the "co-occurrence" of points in clusters. This similarity can be used to detect arbitrary shaped clusters.

Even more applications

- Measure d characteristics of skin cells and to determine characteristics that indicate whether a cell is cancerous / benign.
- Choose K $= 2$ to "binarize" a photo, i.e. make each pixel either black or white.
  - May choose K=3 for a kind of grayscale, repaint these clusters with different colors