Unsupervised ML

Institute of Technology of Cambodia

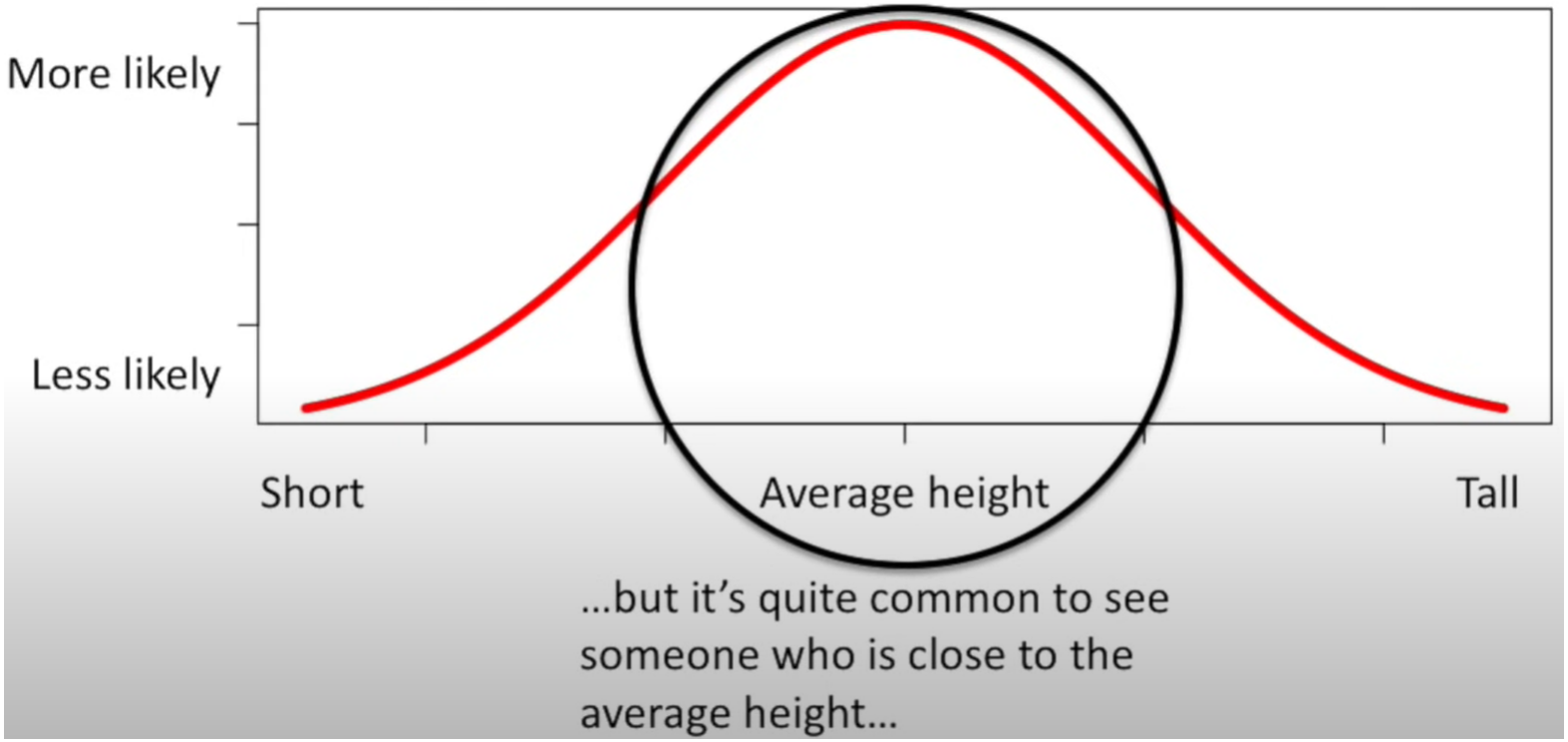
November 7, 2023

Gaussian Mixtures

Gaussian Mixtures

- Generalize K-means concept
  - Clusters are sets of points that are close in euclidean space
  - Different clusters tend to be far apart
- Translate it statistically
  - Each cluster can be described using a normal distribution centered on its centriod, with the probability of observing points decreasing with the distance to the centriod

Normal Distribution



Gaussian Mixtures

We define a **generative model** for k clusters.

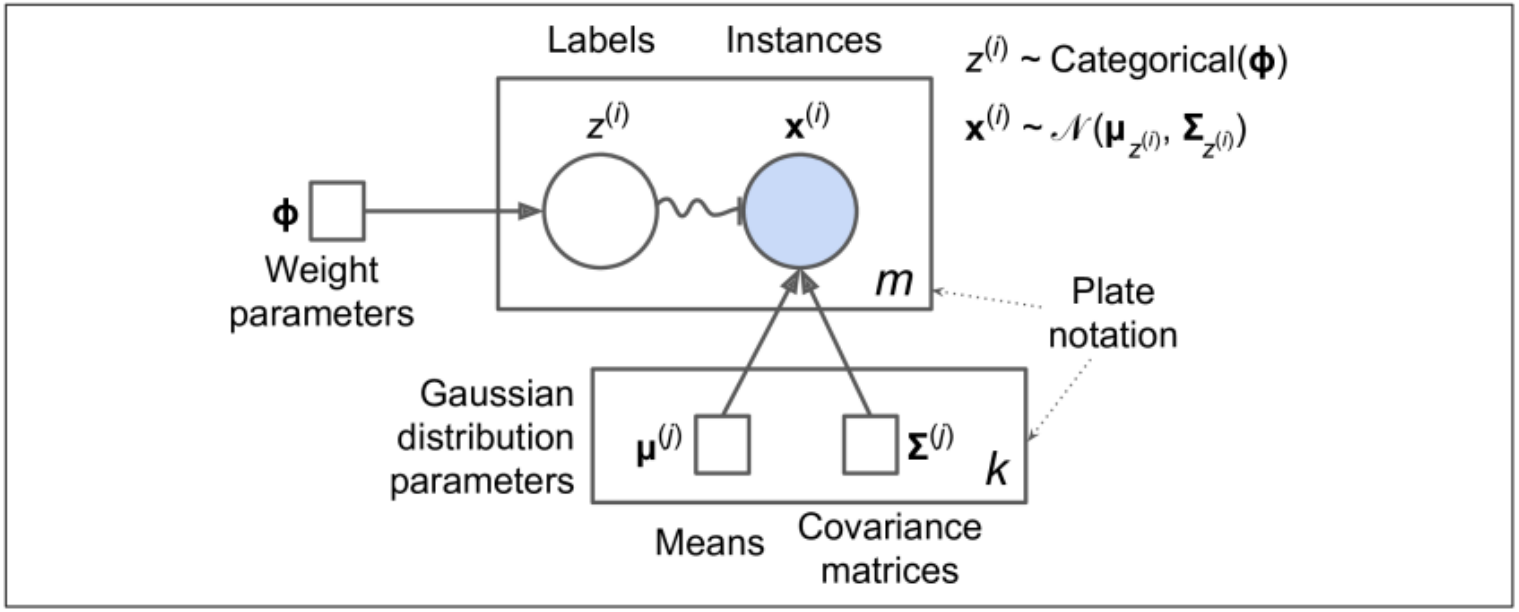
- Each cluster corresponds to a Gaussian distribution, defined by a center and a variance, or covariance matrix
- The problem to solve is to find the parameters (centers, variances) that maximize the likelihood of the corresponding model to generate the observed item X.

Gaussian Mixture

There are several GMM variants: in the simplest variant, implemented in the `GaussianMixture` class, you must know in advance the number k of Gaussian distributions. The dataset $\mathbf{X}$ is assumed to have been generated through the following probabilistic process:

- For each instance, a cluster is picked randomly among k clusters. The probability of choosing the jth cluster is defined by the cluster's weight $\phi^{(j)}$.[7] The index of the cluster chosen for the ith instance is noted $z^{(i)}$.
- If $z^{(i)}=j$, meaning the ith instance has been assigned to the jth cluster, the location $\mathbf{x}^{(i)}$ of this instance is sampled randomly from the Gaussian distribution with mean $\boldsymbol{\mu}^{(j)}$ and covariance matrix $\boldsymbol{\Sigma}^{(j)}$. This is noted $\mathbf{x}^{(i)} \sim \mathcal{N}\left(\mu^{(j)}, \Sigma^{(j)}\right)$.

Gaussian Mixture



- The circles represent random variables
- The squares represent fixed values (i.e., parameters of the model)
- The large rectangles are called plates: they indicate that their content is repeated several times

K-Means Equivalence

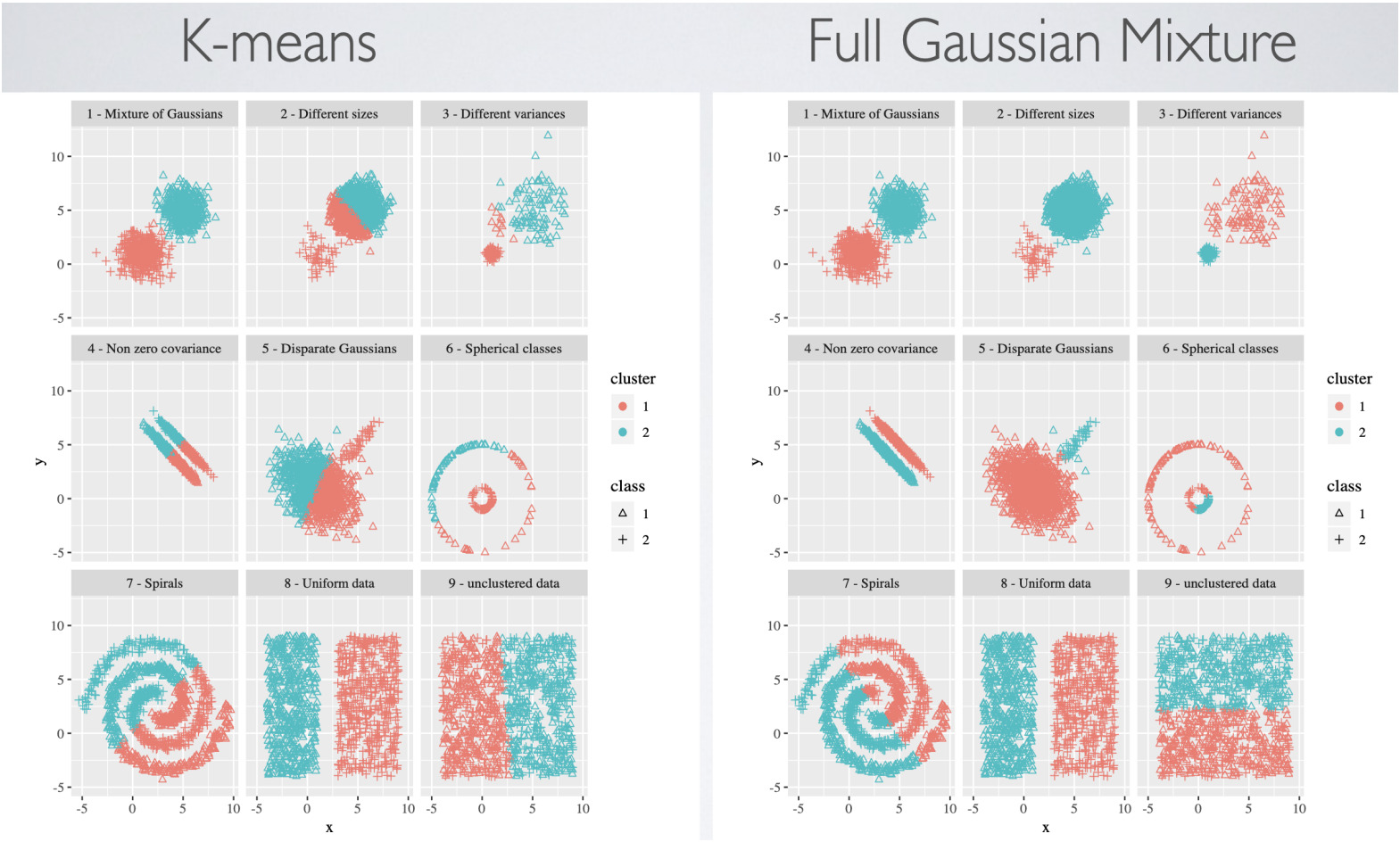
If we assume that:

- The Gaussian distributions are defined only by their variance, not by complete covariance matrices
  - Similar in all directions, "spherical"
- The variance value is the same for all Gaussian distributions
  - Spheres of the same "size"
- The probability for each item to be generated by each of the Gaussian distribution is identical

Then it can be shown that the objective is equivalent to the k-means objective!

- We can relax some of those constraints to get better results

K-Means Comparison

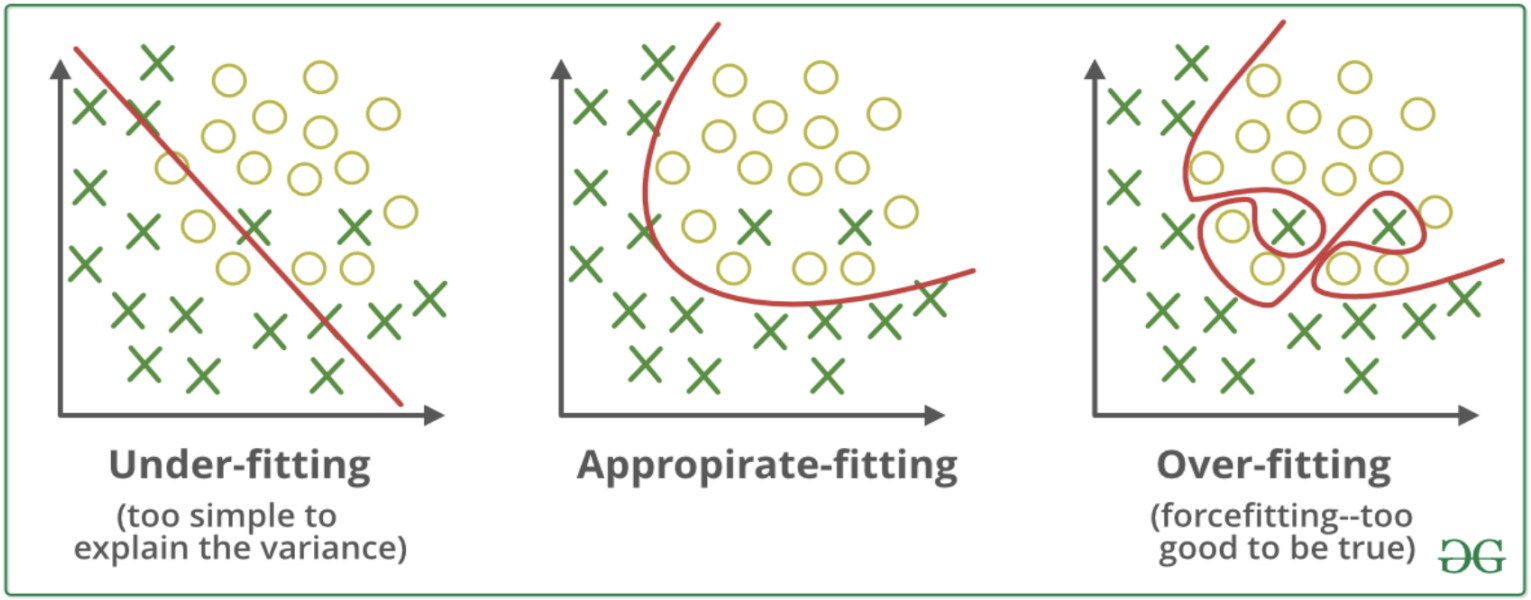


https://smorbieu.gitlab.io/gaussian-mixture-models-k-means-on-steroids/

Pros and Cons

Gaussian mixture seems an improvement over k-means. Why not always using it?

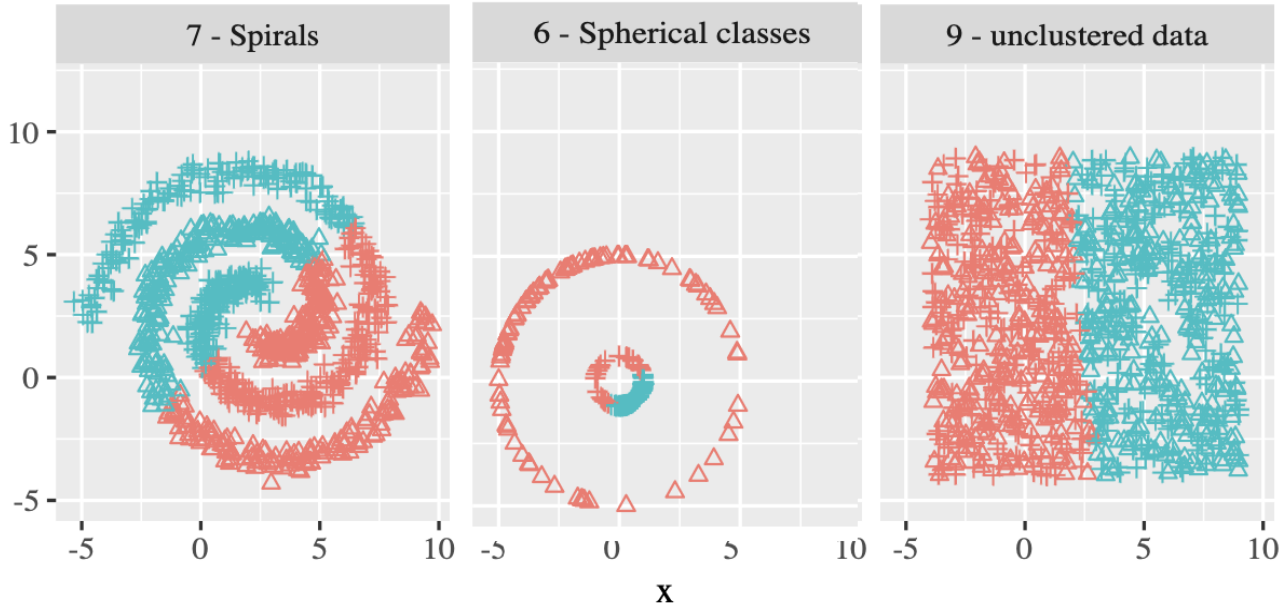
- Force of habits
- Higher computational cost (more parameters, more complex problem)
- Higher possibility of overfitting (more parameters, more overfit risk)



Remaining Problems

We can mention 3 problems remaining (at least)

- The number of clusters still need to be provided
  - If allowed to change, it will always converge to the trivial solution with each item in its own cluster
- If the data is completely random, the method still finds clusters
- Impossible to discover non-convex structures, such as circles or spirals.



Gaussian Mixture

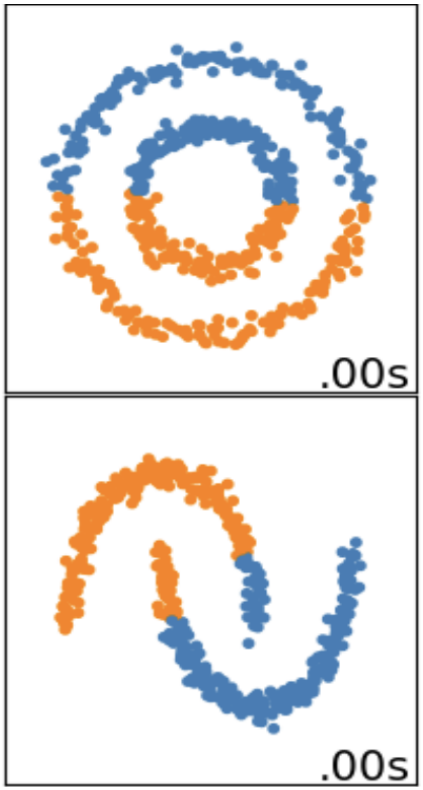
What are Gaussian mixture models used for?

1. Clustering
2. Density estimation
3. Anomaly detection
4. Feature extraction

DBSCAN

K-Means/GM limits

The problem of spiral/Circular/weird shaped clusters comes from the assumption that items of a cluster should be "normally distributed" around their mean.



Local Definitions

To overcome this problem, several methods propose local definitions of clusters

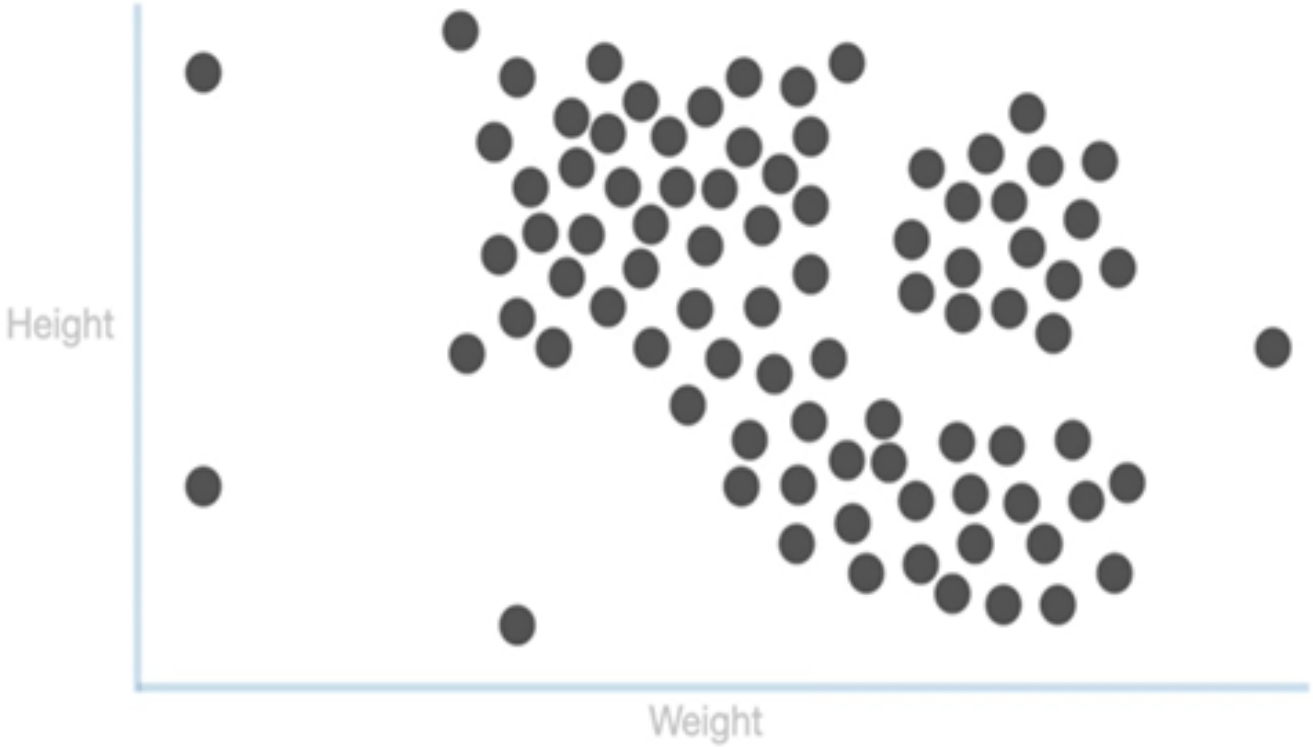
- Does not explicitly optimize a global function
- Items belong to clusters because they are close enough, locally, to other items in that cluster
- Clusters exist because there is continuum between all items in it, locally

DBSCAN

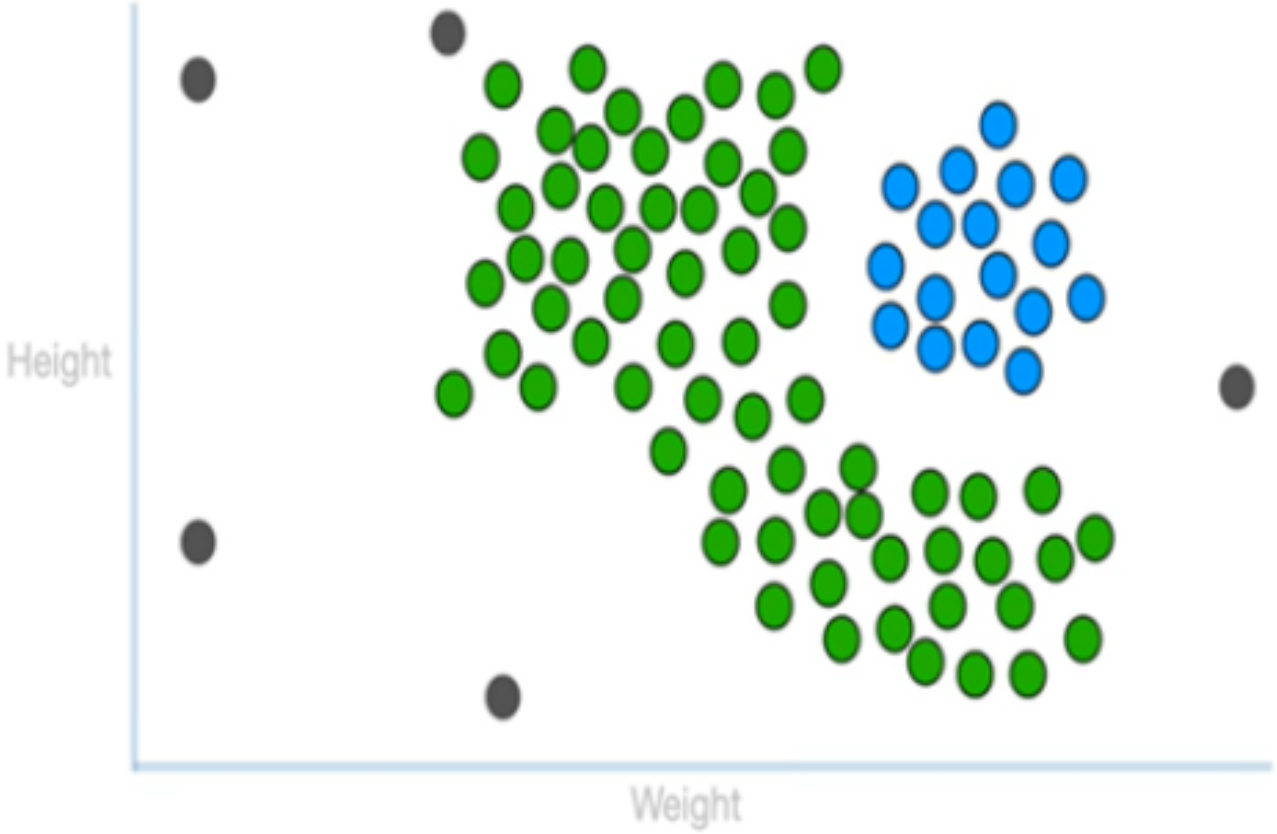
Now, imagine we collected **Weight** and **Height** measurements from a bunch of people…

	Weight	Height
Person 1	56	150
Person 2	62	170
Person 3	71	168
…	…	…

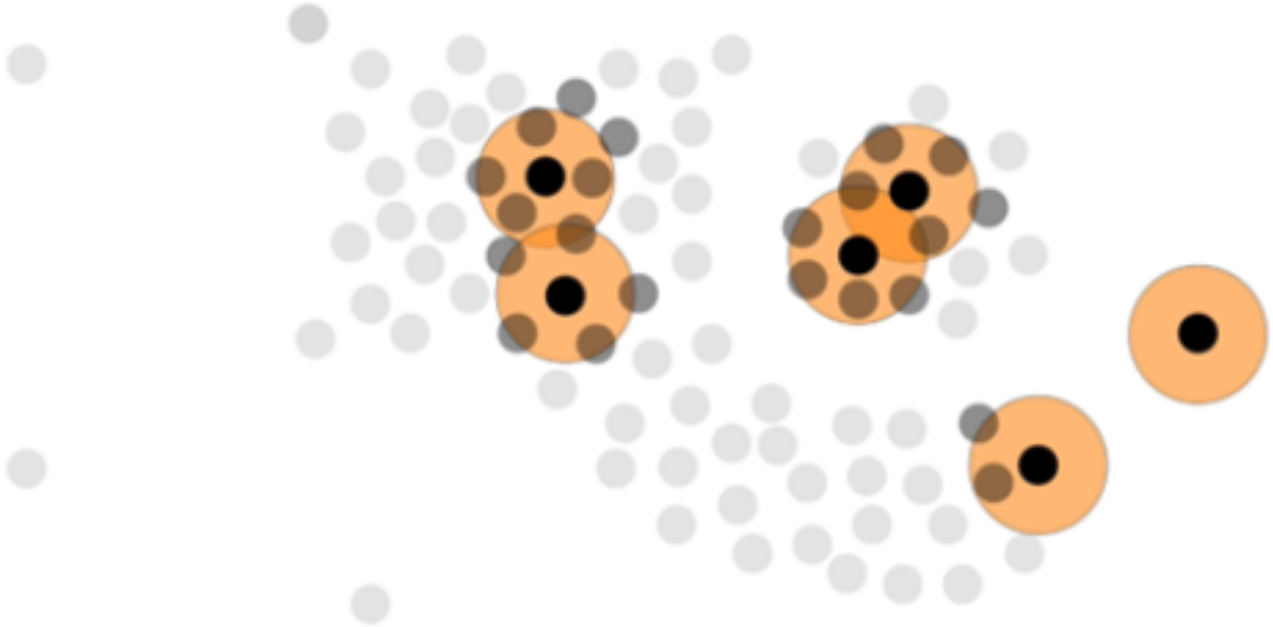
DBSCAN



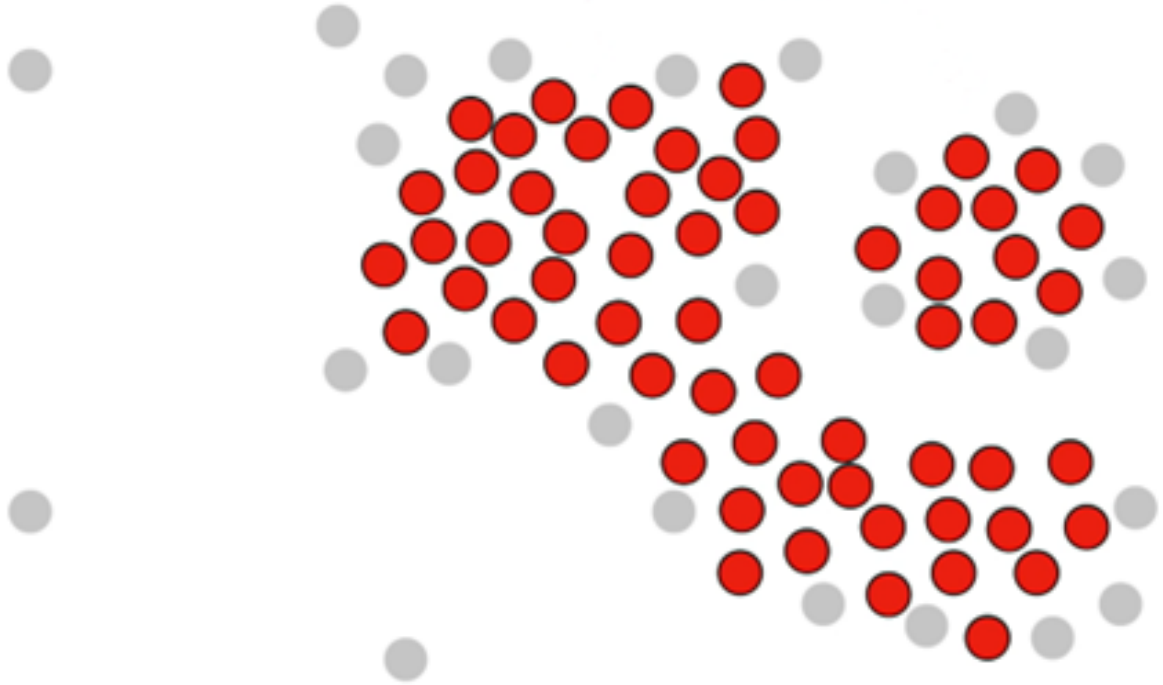
DBSCAN



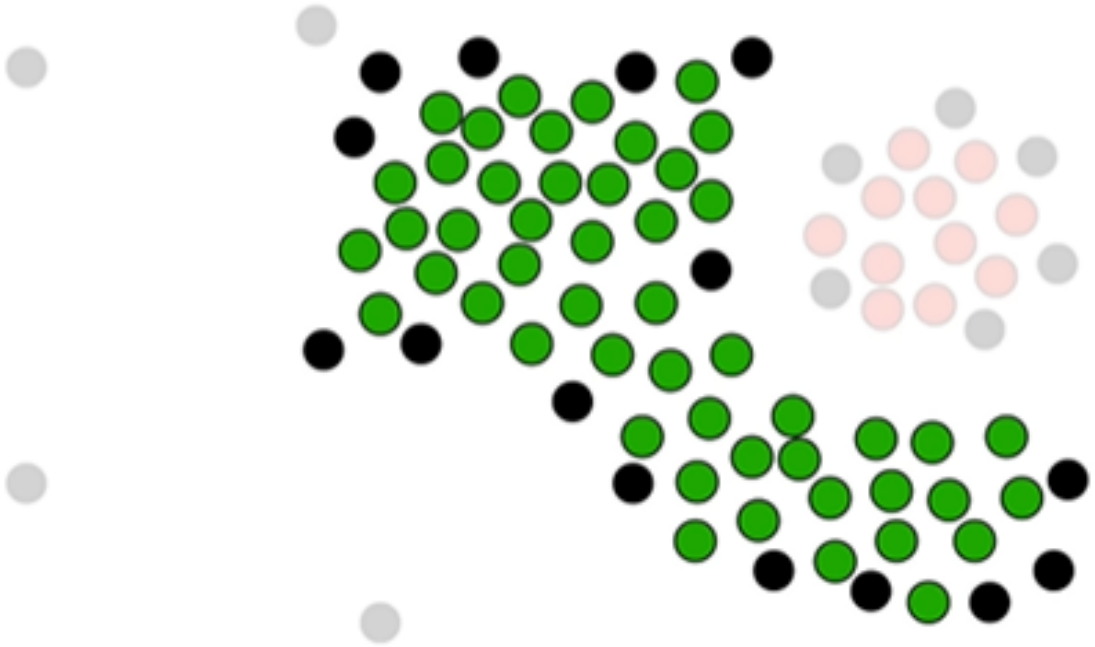
DBSCAN



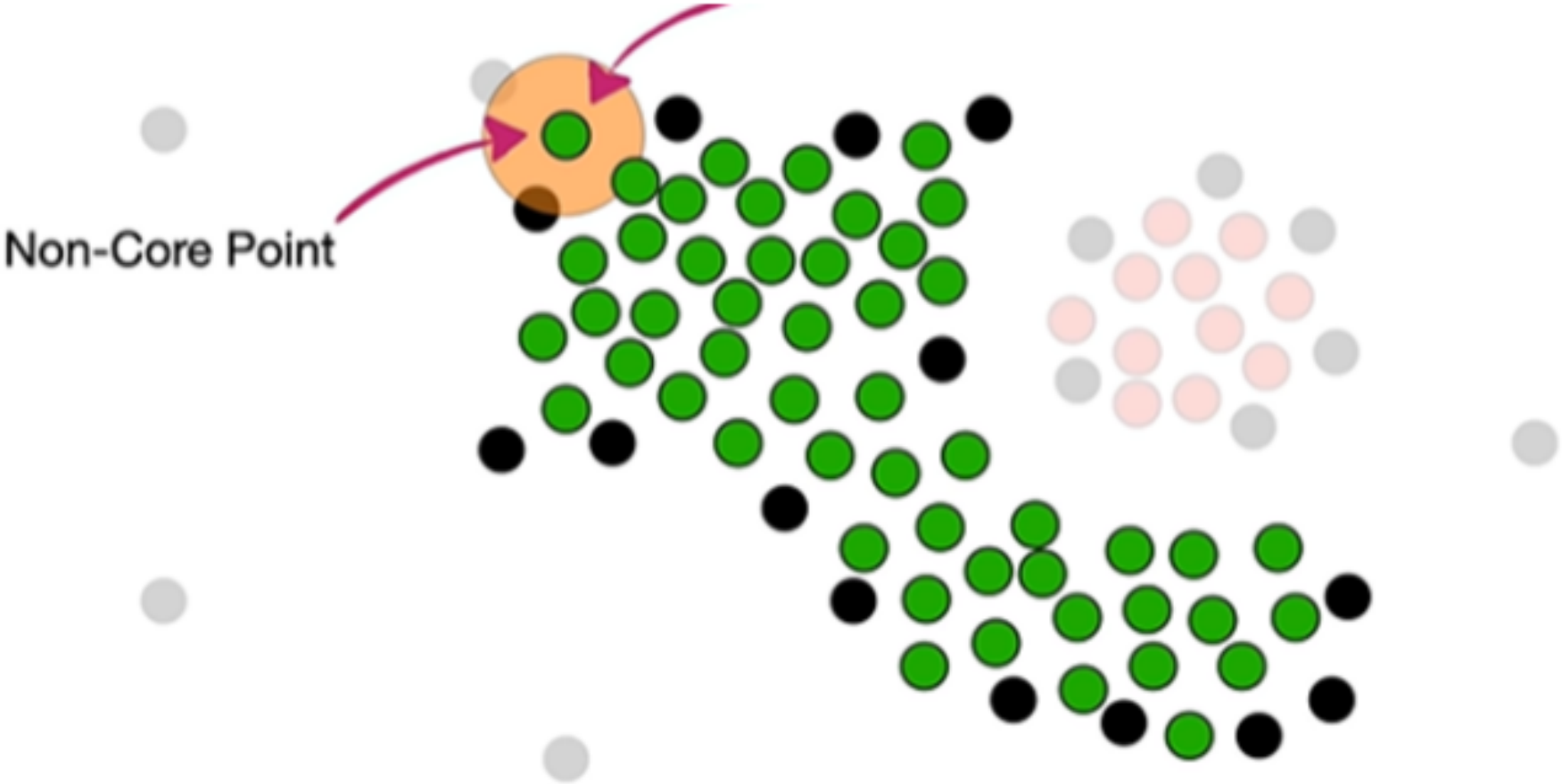
DBSCAN



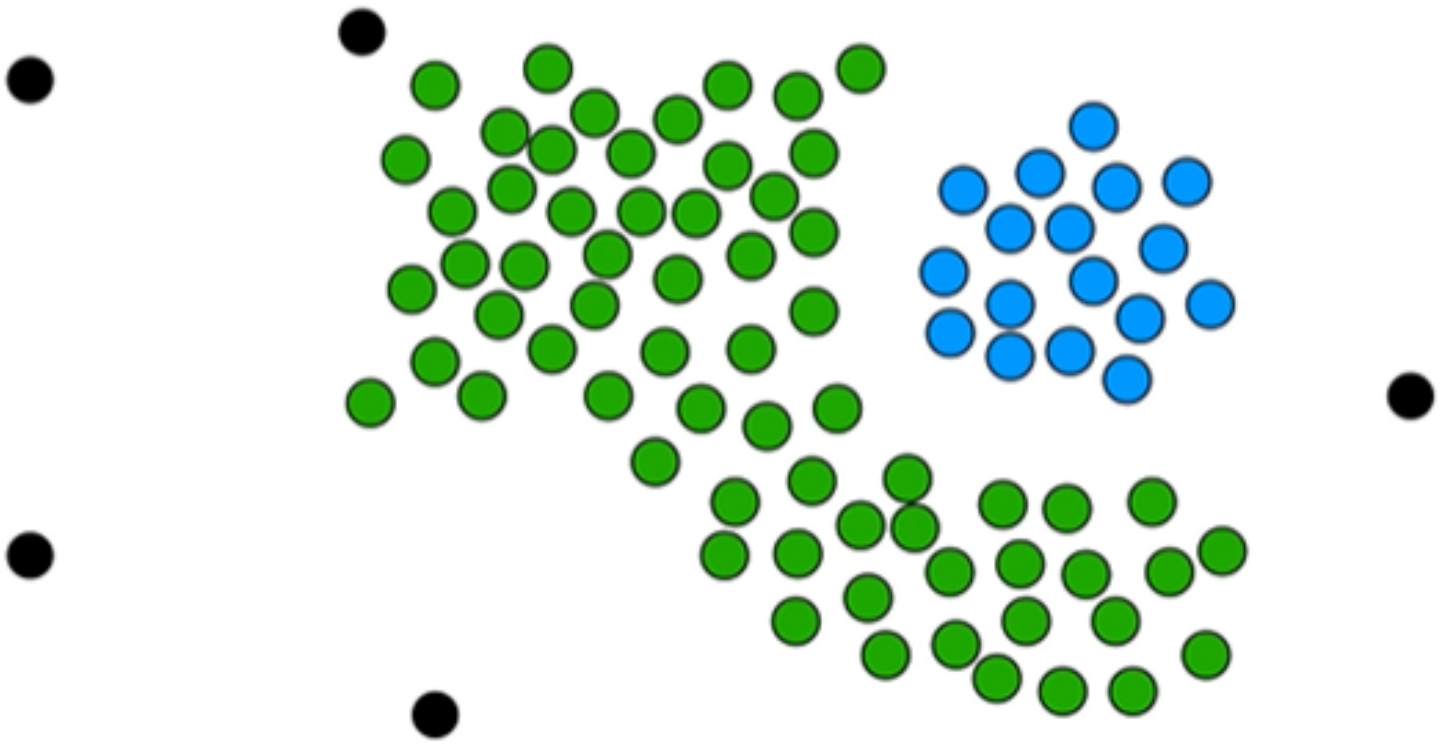
DBSCAN



DBSCAN



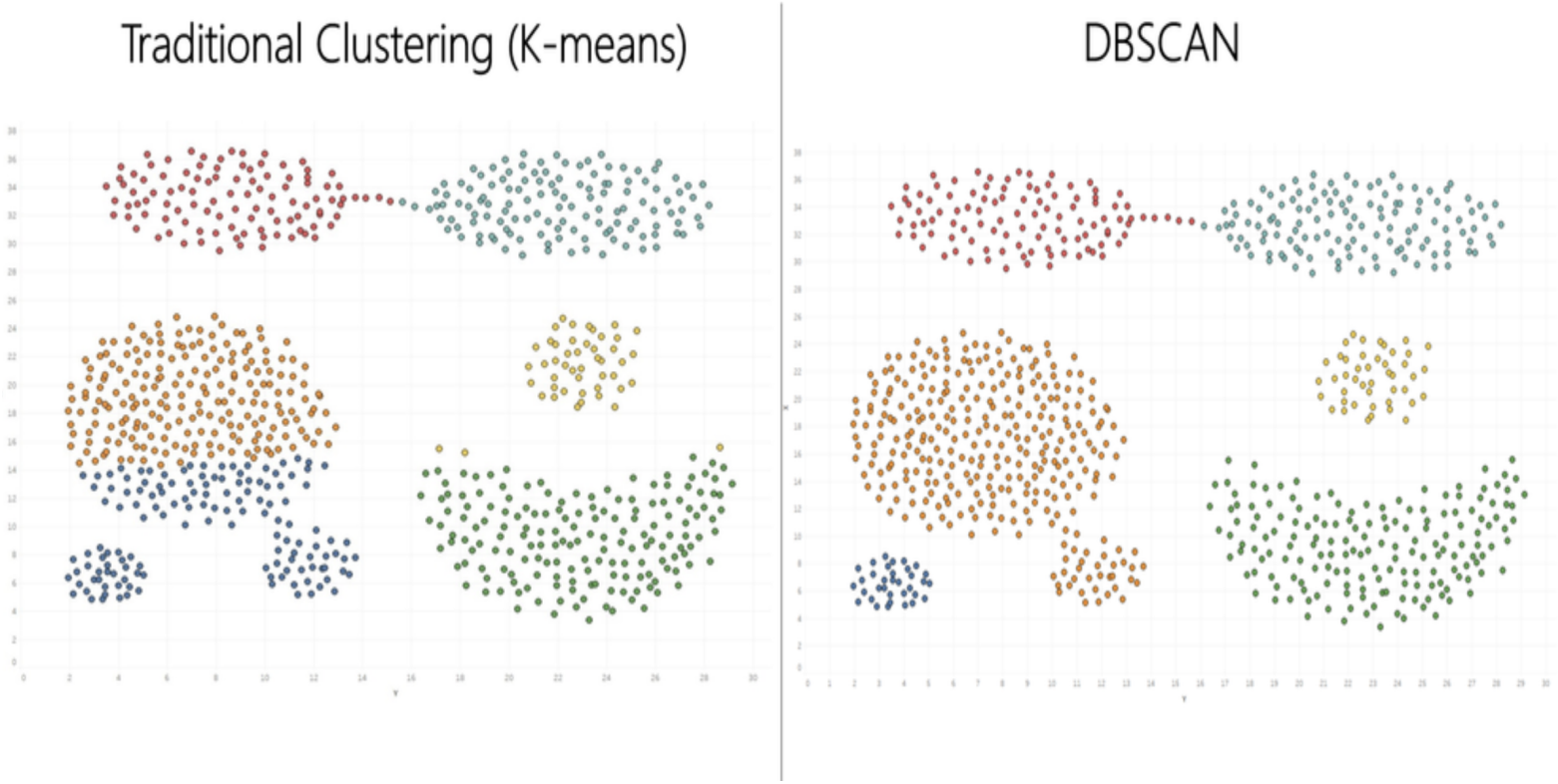
DBSCAN



DBSCAN

1. Build a graph such as
   - Each core node is a node
   - A link exist between core nodes if they are at smaller than distance threshold
2. Detect the connected components of the graph
   - 2 nodes belong to the same connected components if there is a path between them
3. For all non-core nodes
   - If they have no core points directly reachable, discard them as noise
   - Else, attribute them to (one of) the clusters for which one core point is directly reachable
     - Variant DBSCAN*, ignore those points as noise

DBSCAN



https://community.alteryx.com/t5/Data-Science/Partitioning-Spatial-Data-with-DBSCAN/ba-p/446273

DBSCAN

Strength

- No need to define the number of clusters
- Can discover arbitrarily-shape clusters
- A notion of noise

Weaknesses

- Defining distance threshold is extremely difficult
  - Similar to the number of clusters
  - In fact it determines the number of clusters
- Despite safeguards, risk of the stretched clusters effect

Clustering Evaluation

Internal/External

- Two types of evaluation: internal or external
- External: we have a Ground Truth (GT). We compare what we have found (predictions) with the "truth"
- Internal: No ground truth, we reply only some intrinsic property of our clusters

Internal/External

External Evaluation (extrinsic):

- The ground truth can be exactly the right clustering desired
  - So we are just validating the method, since we already know the answer
- The ground truth can be a proxy to what we want
  - e.g., we want to cluster stars based on their characteristics (size, temperature, color...) We already have a manual historic categorization (red dwarf, Brown dwarfs, Red giants...) We assume that the new categories found should be somewhat similar

External Evaluation

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

- TP: two nodes in same cluster in both GT and solution
- TN: two nodes in different clusters in both GT and solution
- TP+FP+FN+TN = all possible node pairs

Internal Evaluation

AD-HOC SCORES

Several clustering method define their own objective to minimize. This objective can be used as a score for clusters obtained by this method or others

- k-means minimizes inter-cluster variance
- Gaussian mixture maximize likelihood

But can lead to unfair comparison

- Using inter-cluster variance to compare k-means and another method such as DBSCAN is unfair
  - One explicitly minimize this objective, the other no...
- As always, the choice of a score is equivalent to choosing a definition of cluster

Variant: Eblow Method

Another well known method to find automatically the number of cluster consists in plotting a measure of quality such as the inter-cluster variance, and cut at an 'elbow'

- Diminishing returns, less 'worthy' to continue