Data Graduate Program N°2
Cdiscount Academy

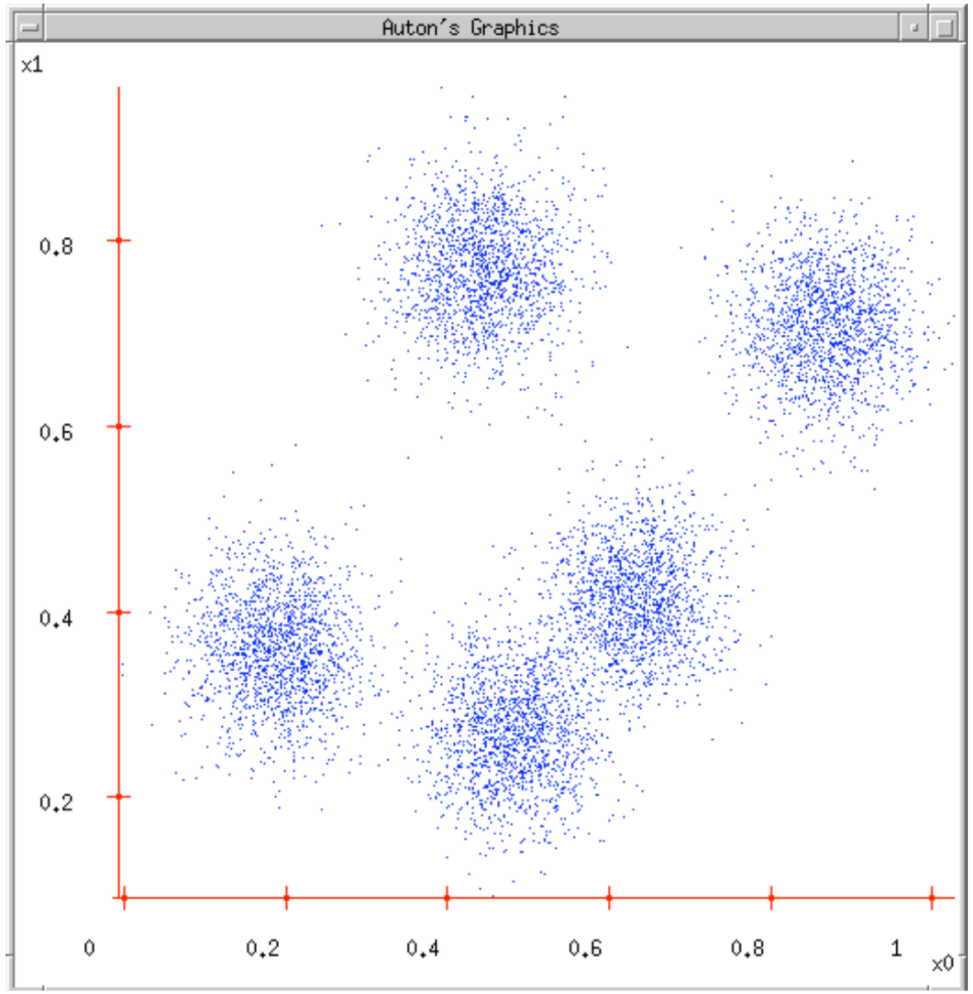
Lundi 20/09 – Mercredi 23/10

SOMMAIRE

- K-means and K-nn

K-means

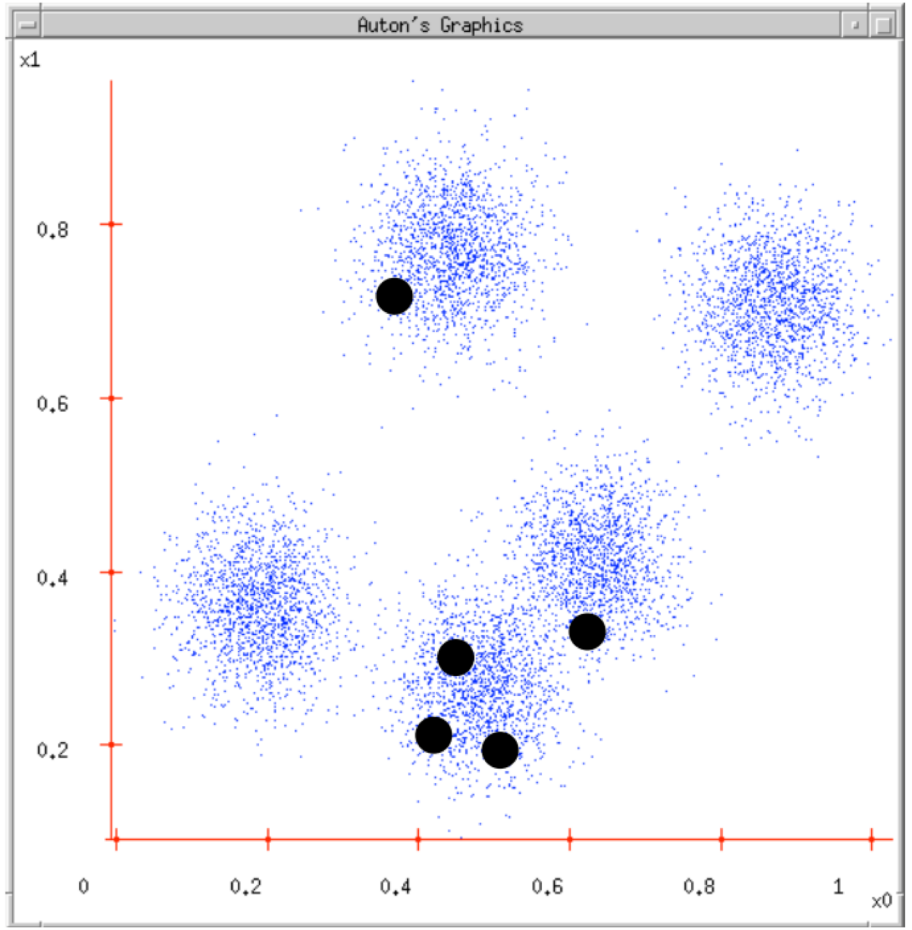
K-MEANS



K-MEANS

K-Means (k , X)

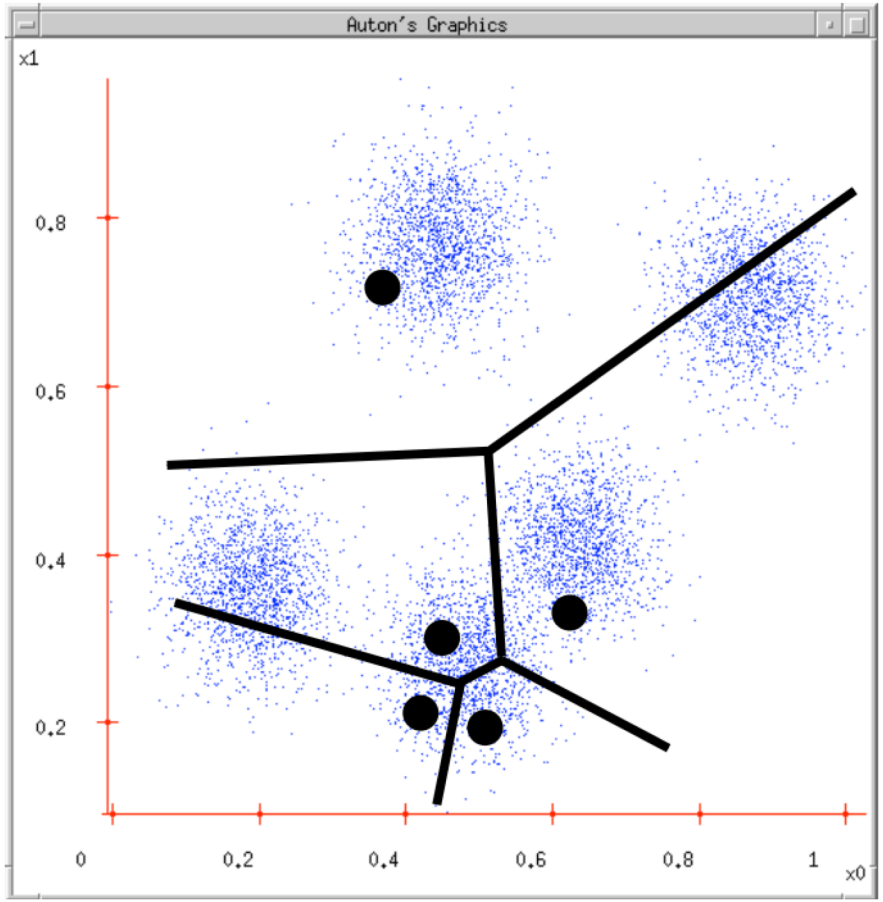
- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re-estimate the cluster centroids based on the data assigned to each cluster



K-MEANS

K-Means (k , X)

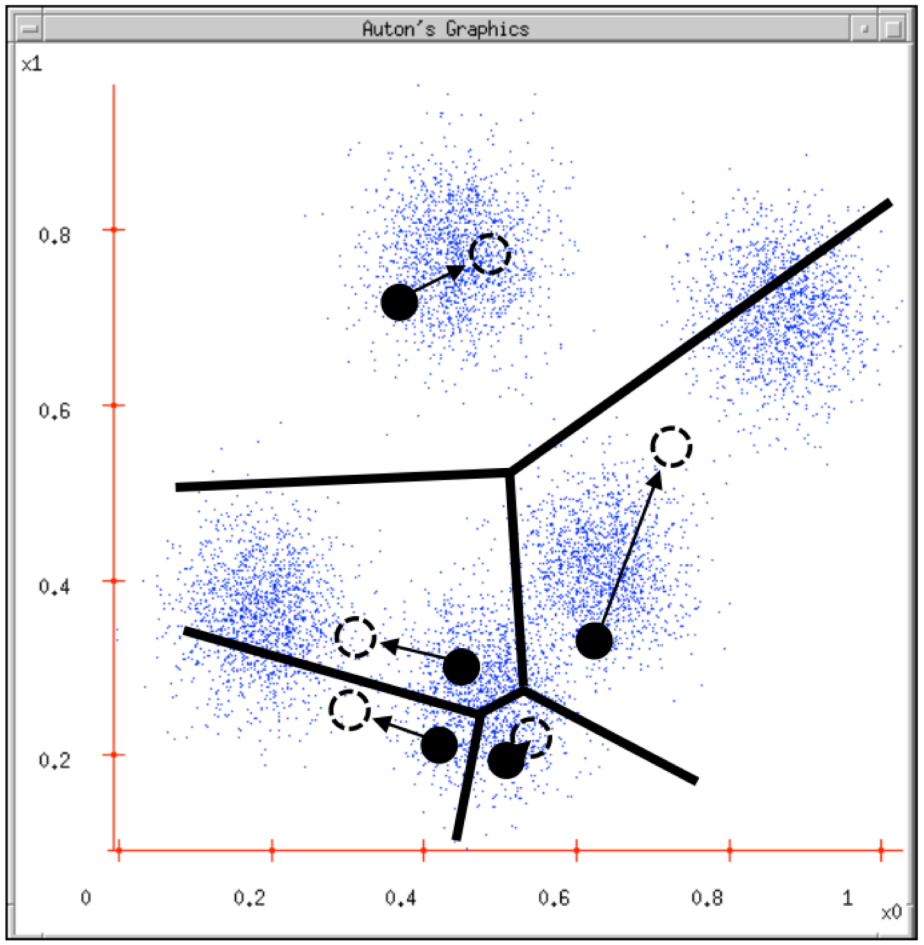
- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re-estimate the cluster centroids based on the data assigned to each cluster



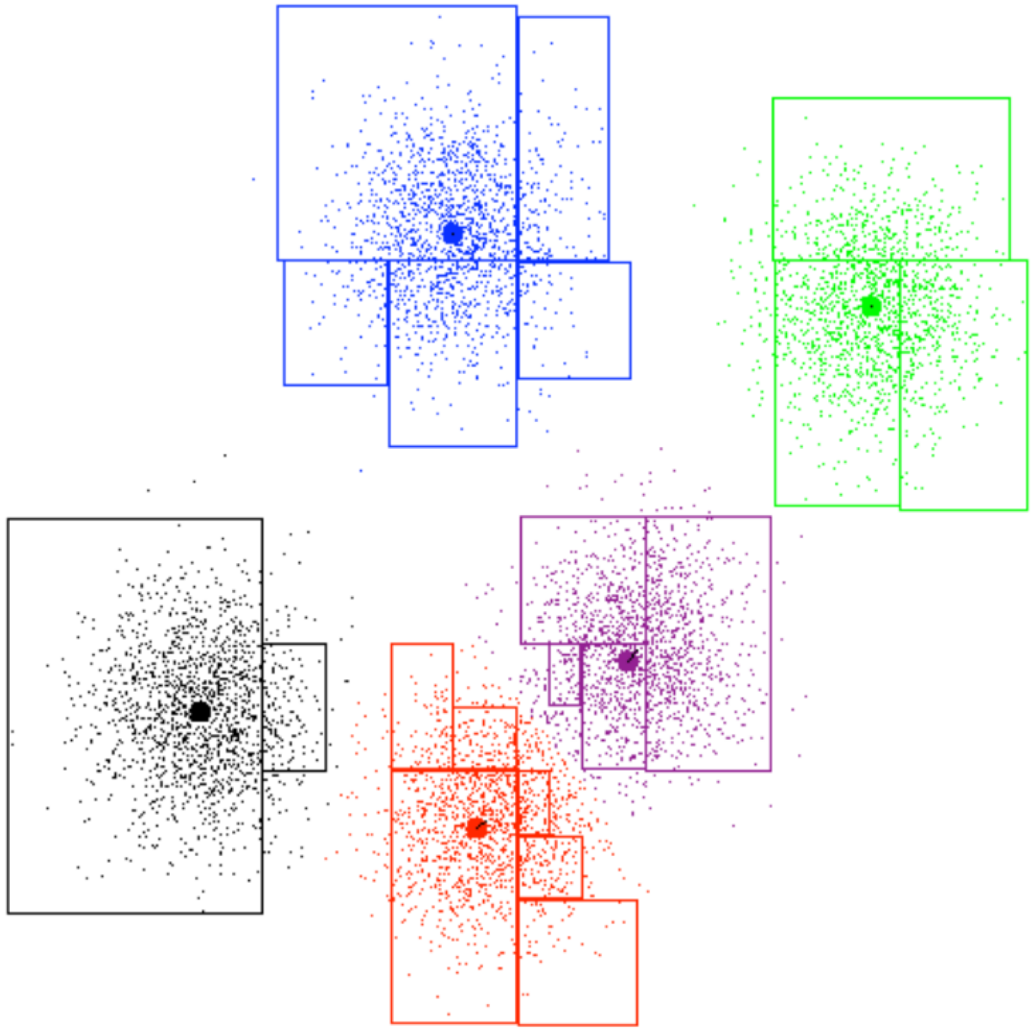
K-MEANS

K-Means (k, X)

- Randomly choose k cluster center locations (centroids)
- Loop until convergence
  - Assign each point to the cluster of the closest centroid
  - Re-estimate the cluster centroids based on the data assigned to each cluster



K-MEANS



K-MEANS

- K-means finds a local optimum of the following objective function:

$$\arg\min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{S}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$$

where $\boldsymbol{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ is a partitioning over $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ s.t. $X = \bigcup_{i=1}^{k} \mathcal{S}_i$ and $\boldsymbol{\mu}_i = \text{mean}(\mathcal{S}_i)$

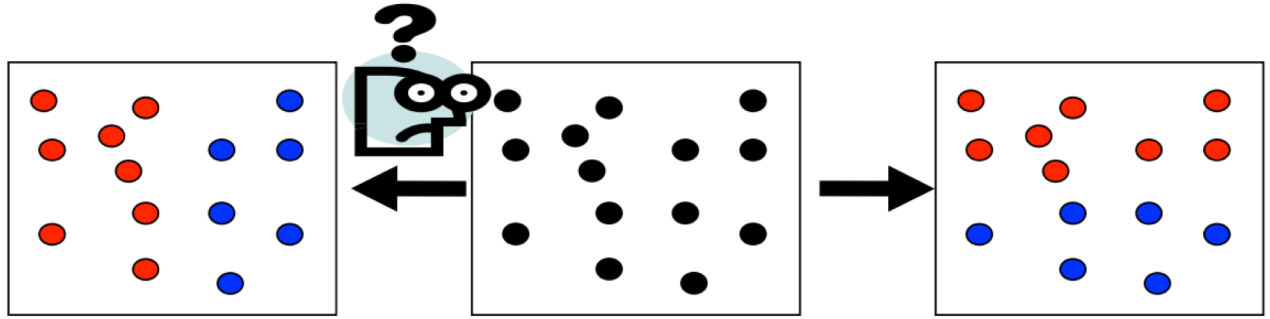
PROBLEM WITH K-MEANS

- ***Very*** sensitive to the initial points
  - Do many runs of K-Means, each with different initial centroids
  - Seed the centroids using a better method than randomly choosing the centroids
    - e.g., Farthest-first sampling

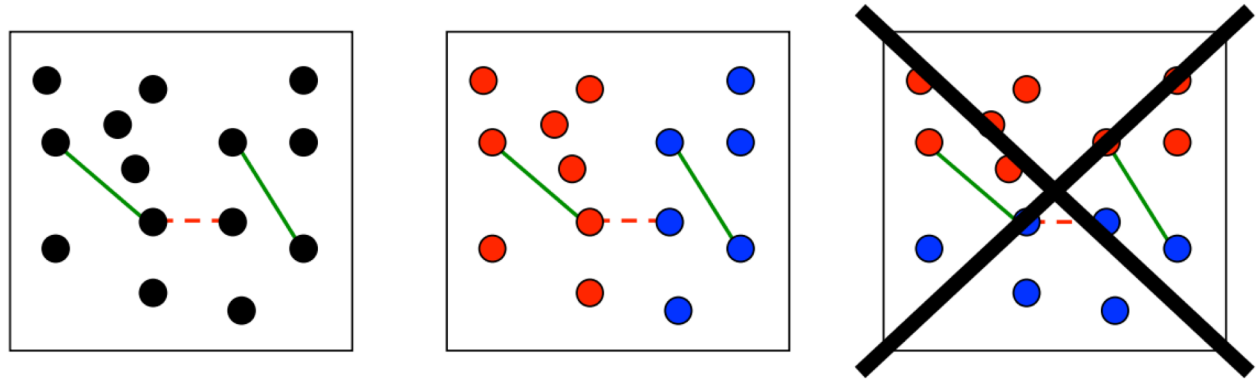
- Must manually choose k
  - Learn the optimal k for the clustering
    - Note that this requires a performance measure

PROBLEM WITH K-MEANS

- How do you tell it which clustering you want?



Constrained clustering techniques (semi-supervised)



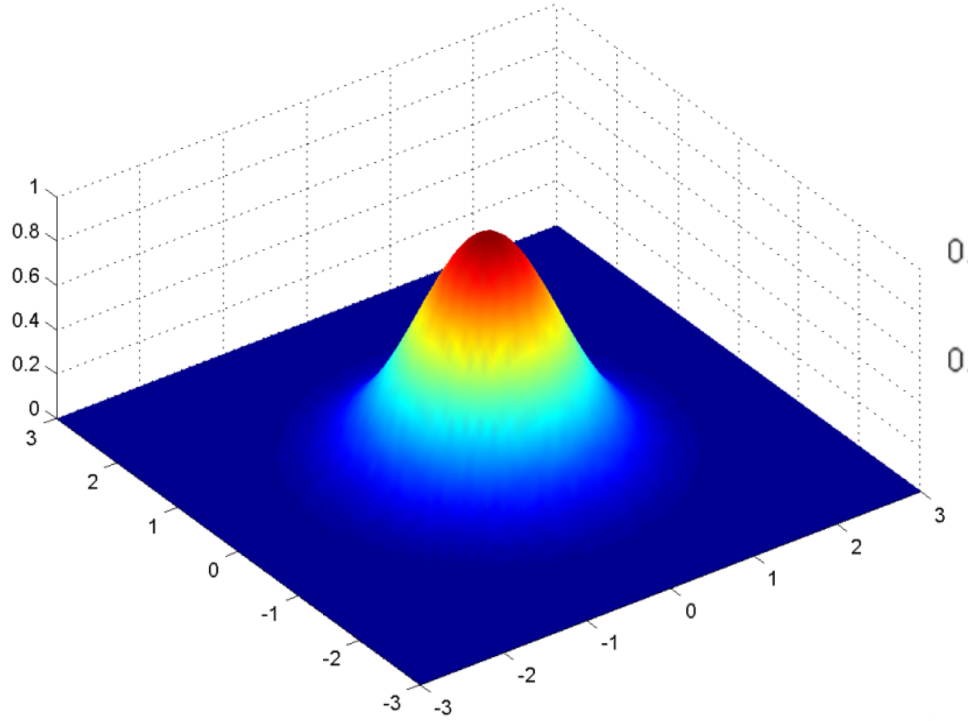
Same-cluster constraint (must-link)

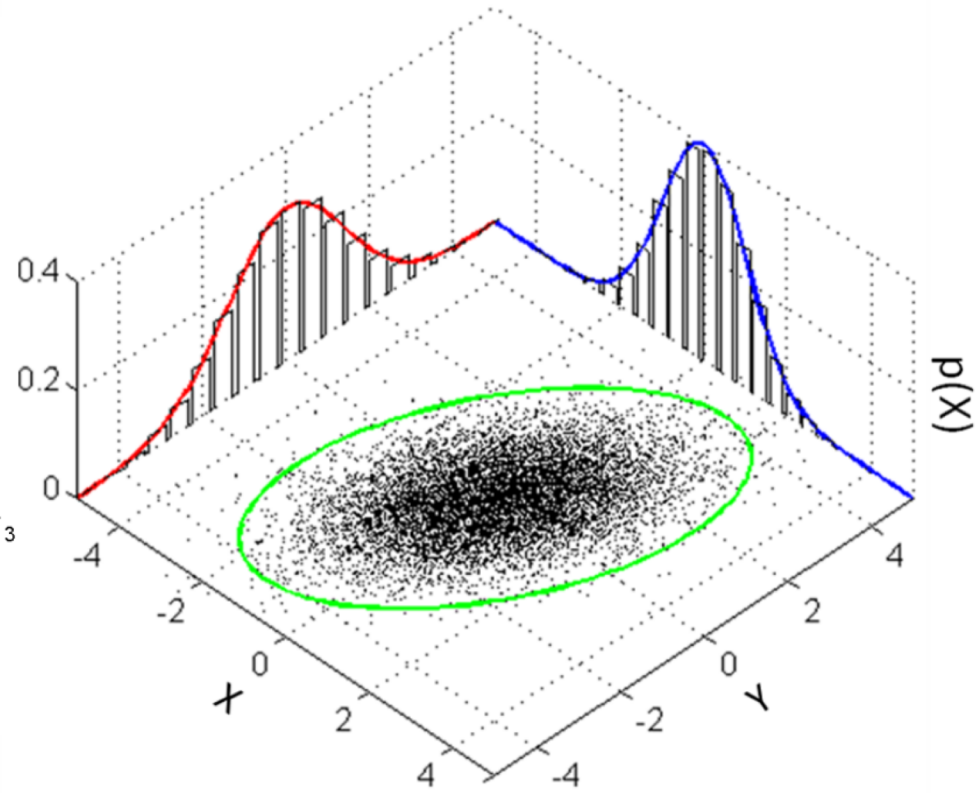
Different-cluster constraint (cannot-link)

GAUSSIAN MIXTURE MODELS

- Recall the Gaussian distribution:

$$P(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

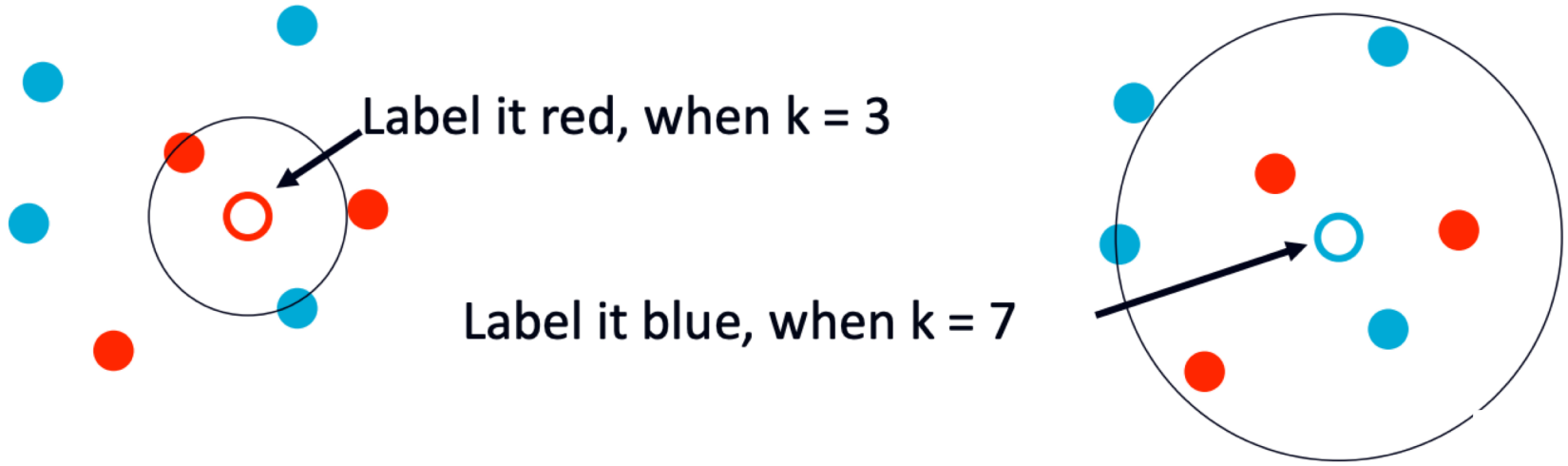




K-NN

K-NN

- Generalizes 1-NN to smooth away noise in the labels
- A new point is now assigned the most frequent label of its *k* nearest neighbors



Shriman TIWARI

Tech Lead/Manager Data Science

Mobile: +33 (0)6 49 71 80 68
shriman.tiwari@keyrus.com

www.keyrus.com

KEYRUS