Mach🤖ne Learn🤖ng.

#Clustering

Eng. Hassan Elseoudy

Table of Contents.

1. Introduction
2. K-Mean Clustering
3. Spectral Clustering
4. Clustering Validation
5. Python.

Introduction

What is Clustering?

What is Clustering?

Clustering is the task of **dividing the population or data points** into a **number of groups** such that data points in the same groups are more **similar** to other data points in the same group than those in other groups.

What is Clustering?

the aim is to **make groups** with **similar attributes** and assign them into **clusters**.

Clustering Types

Hard Clustering: each data point either belongs to a cluster **completely or not**.
Examples : **K – means clustering**.

Soft Clustering: a probability or likelihood of that data point to be in those clusters is assigned.
Examples : **Fuzzy C-means** .

Supervised vs Unsupervised learning

Supervised learning:

Given (x_i , y_i), $i = 1, . . . , n$.

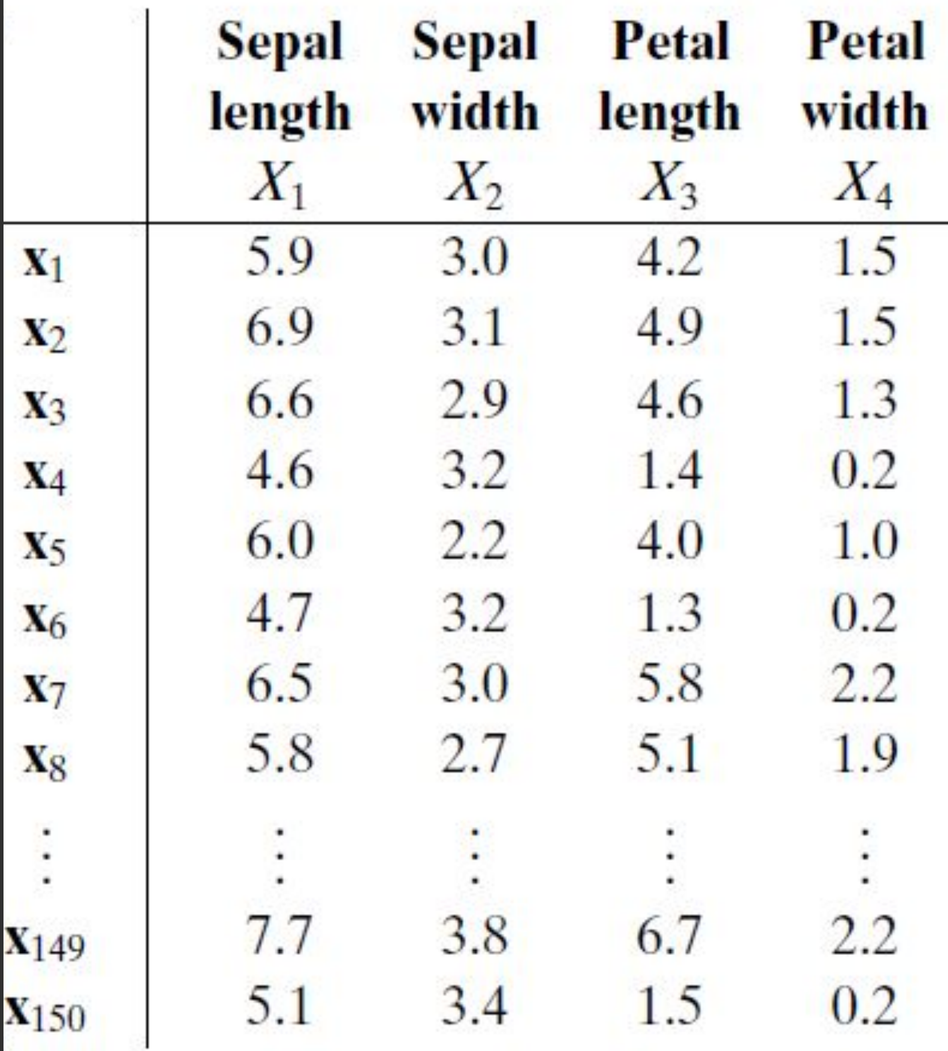
$f(x) : X \to Y$.

- Categorical -> classification.
- Continuous -> regression.

Unsupervised learning: Given only (x_i), $i = 1, . . . , n$, can we infer the underlying structure of X?

Unsupervised learning

Example
Iris Dataset



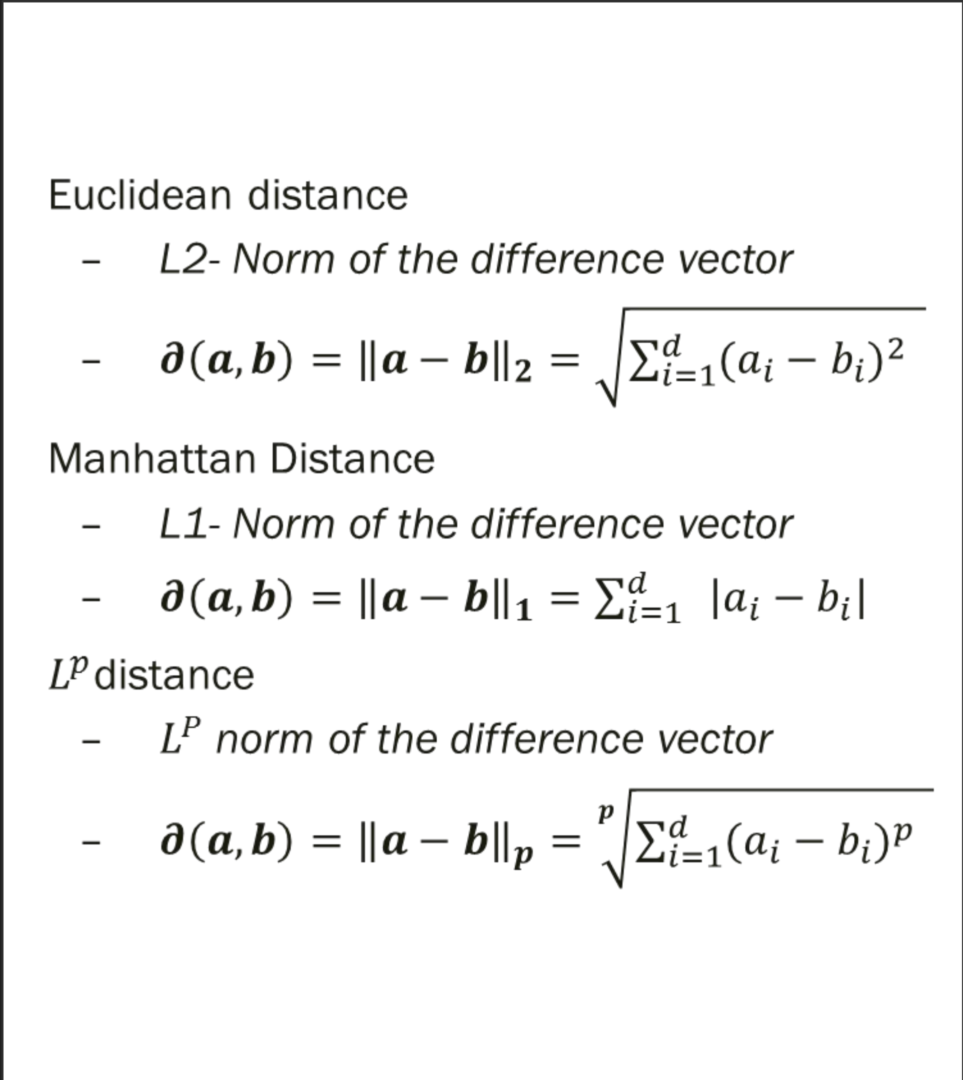
	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4
$\mathbf{x}_1$	5.9	3.0	4.2	1.5
$\mathbf{x}_2$	6.9	3.1	4.9	1.5
$\mathbf{x}_3$	6.6	2.9	4.6	1.3
$\mathbf{x}_4$	4.6	3.2	1.4	0.2
$\mathbf{x}_5$	6.0	2.2	4.0	1.0
$\mathbf{x}_6$	4.7	3.2	1.3	0.2
$\mathbf{x}_7$	6.5	3.0	5.8	2.2
$\mathbf{x}_8$	5.8	2.7	5.1	1.9
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2

Unsupervised learning

Why do unsupervised learning?

- **Raw data** is cheap.
- **Save** memory/computation.
- **Reduce noise** in high-dimensional data.
- Often a **pre-processing step** for supervised learning.

Common Distances

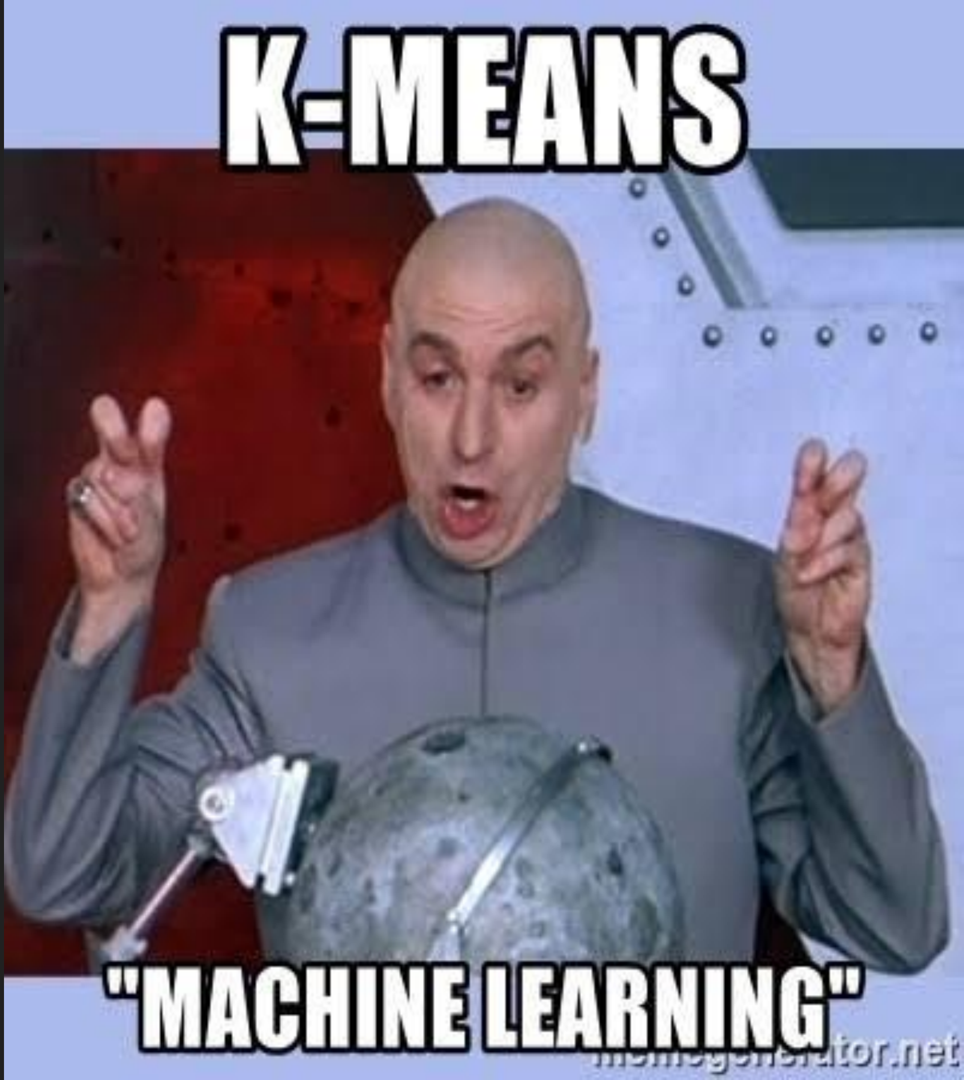


- Euclidean distance
  - *L2- Norm of the difference vector*
  - $\boldsymbol{\partial}(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}-\boldsymbol{b}\|_{\mathbf{2}} = \sqrt{\sum_{i=1}^{d}(a_i-b_i)^2}$
- Manhattan Distance
  - *L1- Norm of the difference vector*
  - $\boldsymbol{\partial}(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}-\boldsymbol{b}\|_{\mathbf{1}} = \sum_{i=1}^{d} |a_i-b_i|$
- L^p distance
  - L^P *norm of the difference vector*
  - $\boldsymbol{\partial}(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}-\boldsymbol{b}\|_{\boldsymbol{p}} = \sqrt[\boldsymbol{p}]{\sum_{i=1}^{d}(a_i-b_i)^p}$

Question?

K-Means Clustering

What is K-Means?



K-Means

Introduction

- Uses Distance Function
- Uses Mean as representative. Called **centroid**
- It has a parameter **K** that you need to guess before clustering.
- Iterative **two-step** approach
    - Cluster Assignment.
    - Centroid Update

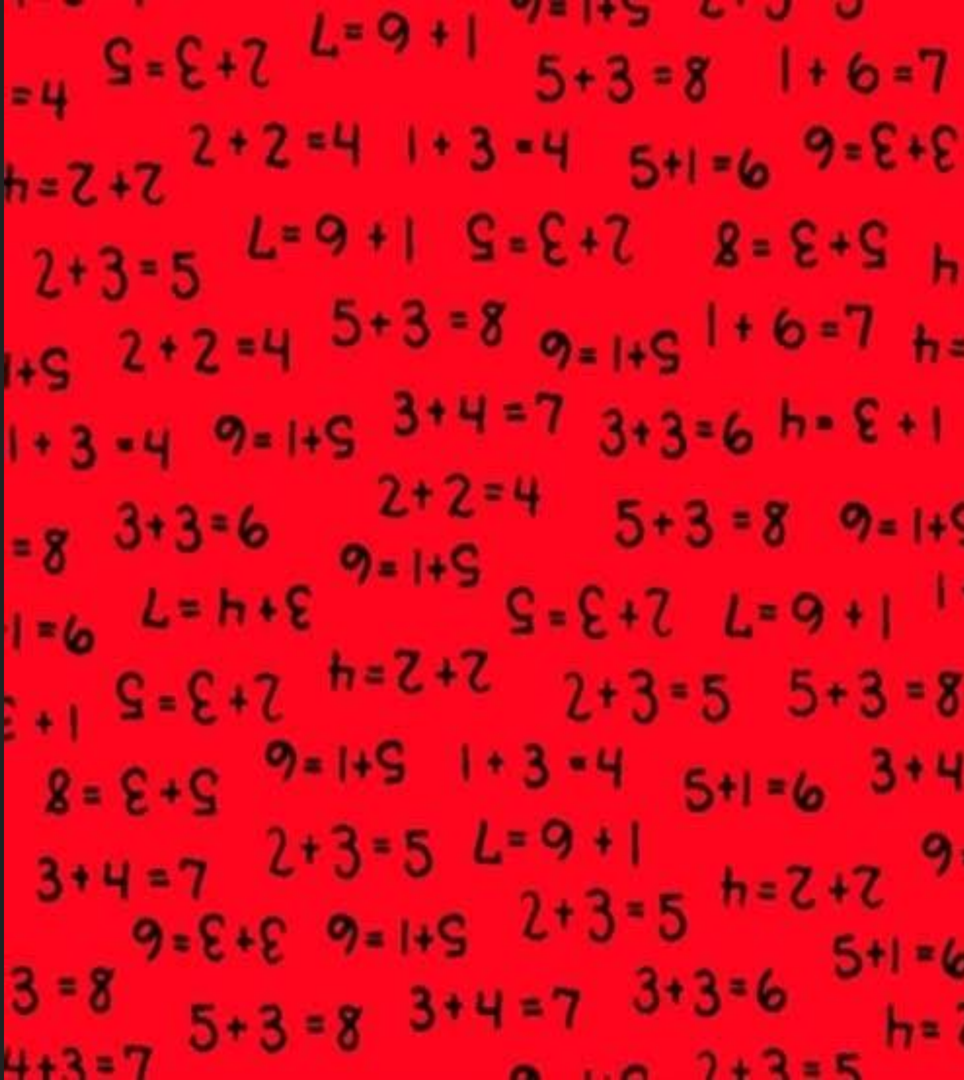
K-Means

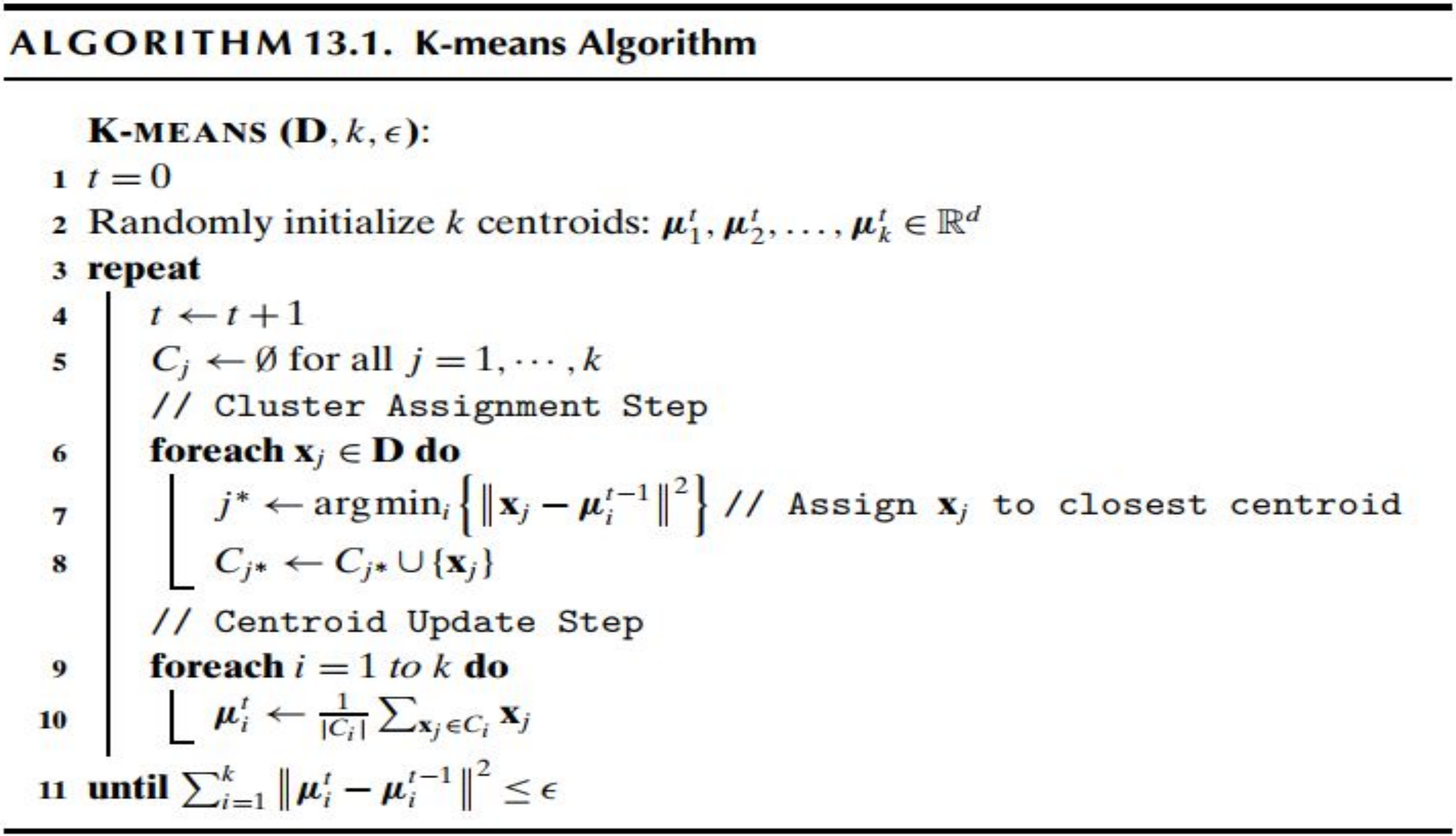
Objective Function

Minimize the sum of the errors between samples in a cluster and their representative (**centroid**).

MATH ALERT!

Algorithm



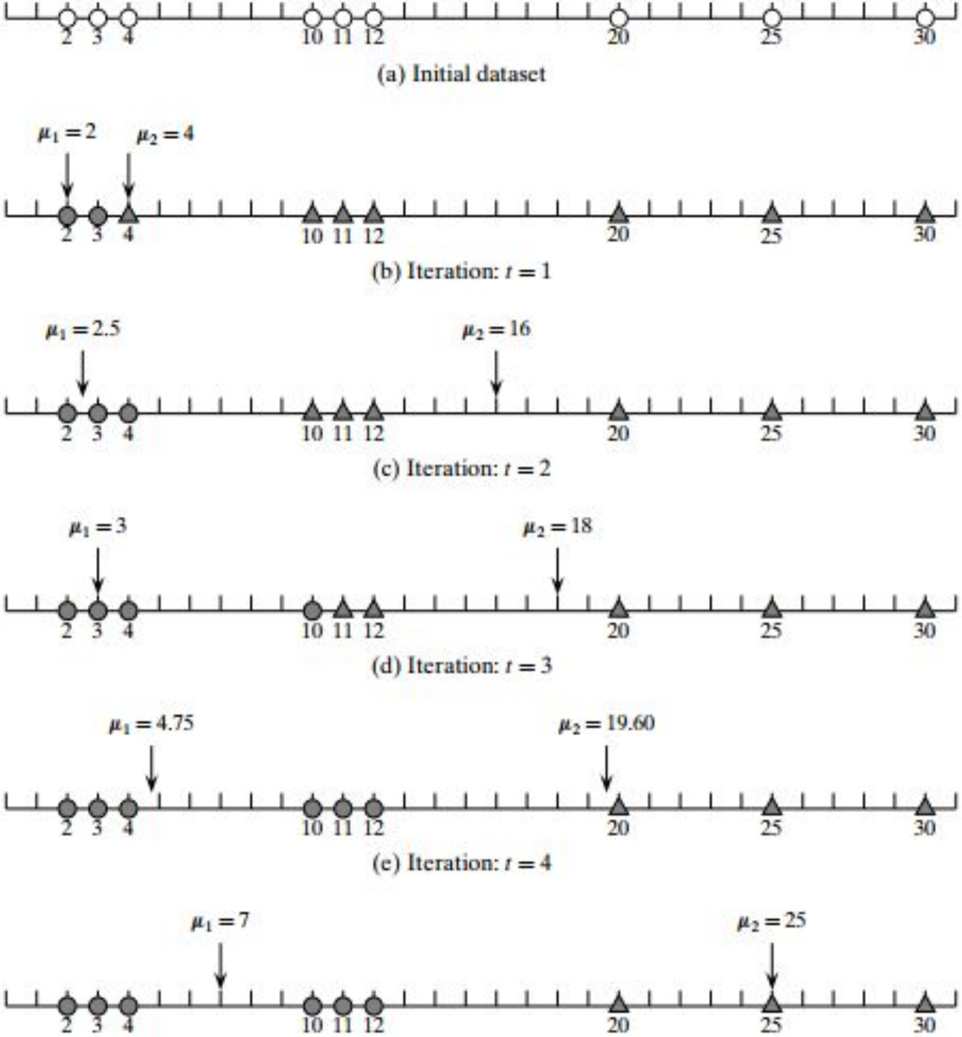


ALGORITHM 13.1. K-means Algorithm

```
K-MEANS ($\mathbf{D}, k, \epsilon$):
$t = 0$
Randomly initialize $k$ centroids: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$
repeat
    $t \leftarrow t + 1$
    $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$
    // Cluster Assignment Step
    foreach $\mathbf{x}_j \in \mathbf{D}$ do
        $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^{t-1} \right\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest centroid
        $C_{j*} \leftarrow C_{j*} \cup \{\mathbf{x}_j\}$
    // Centroid Update Step
    foreach $i = 1$ to $k$ do
        $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
until $\sum_{i=1}^k \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \le \epsilon$
```

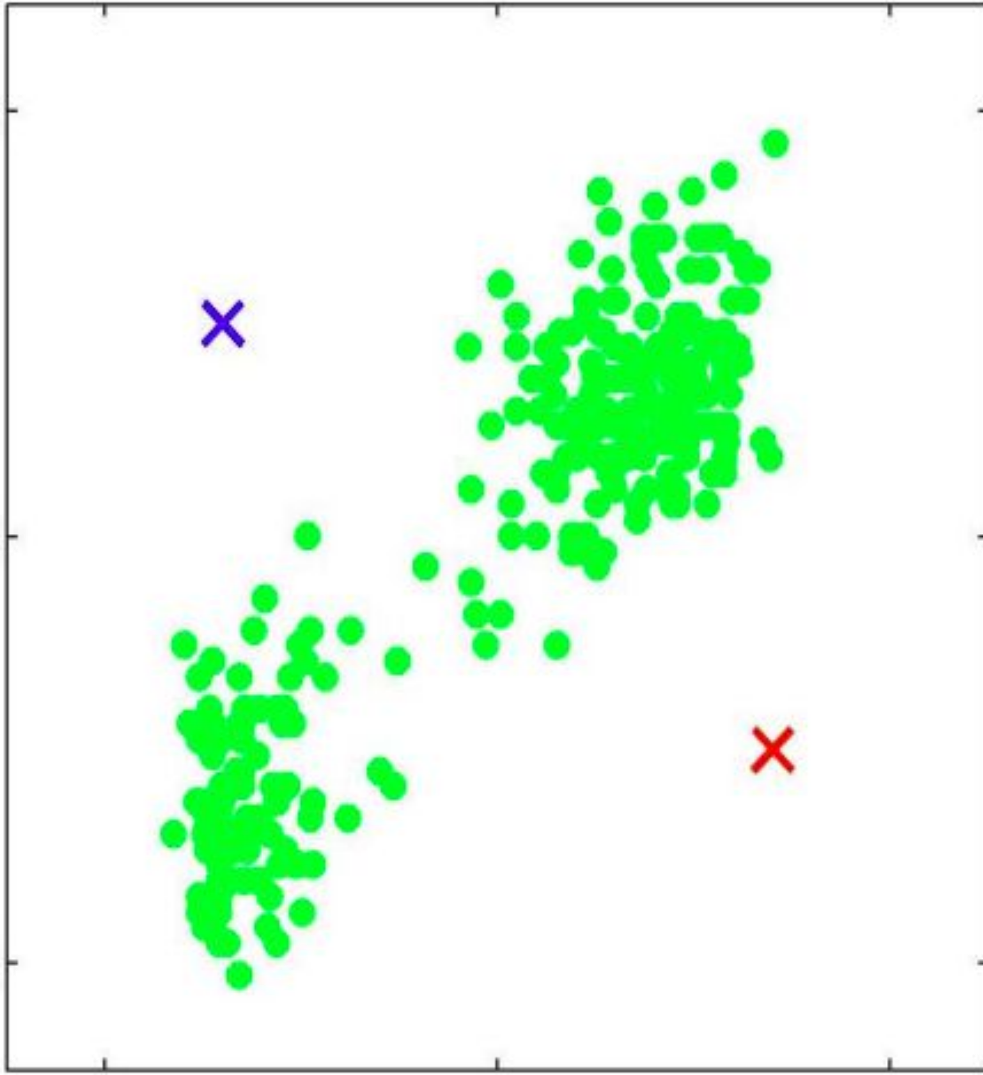
K-Means

Example



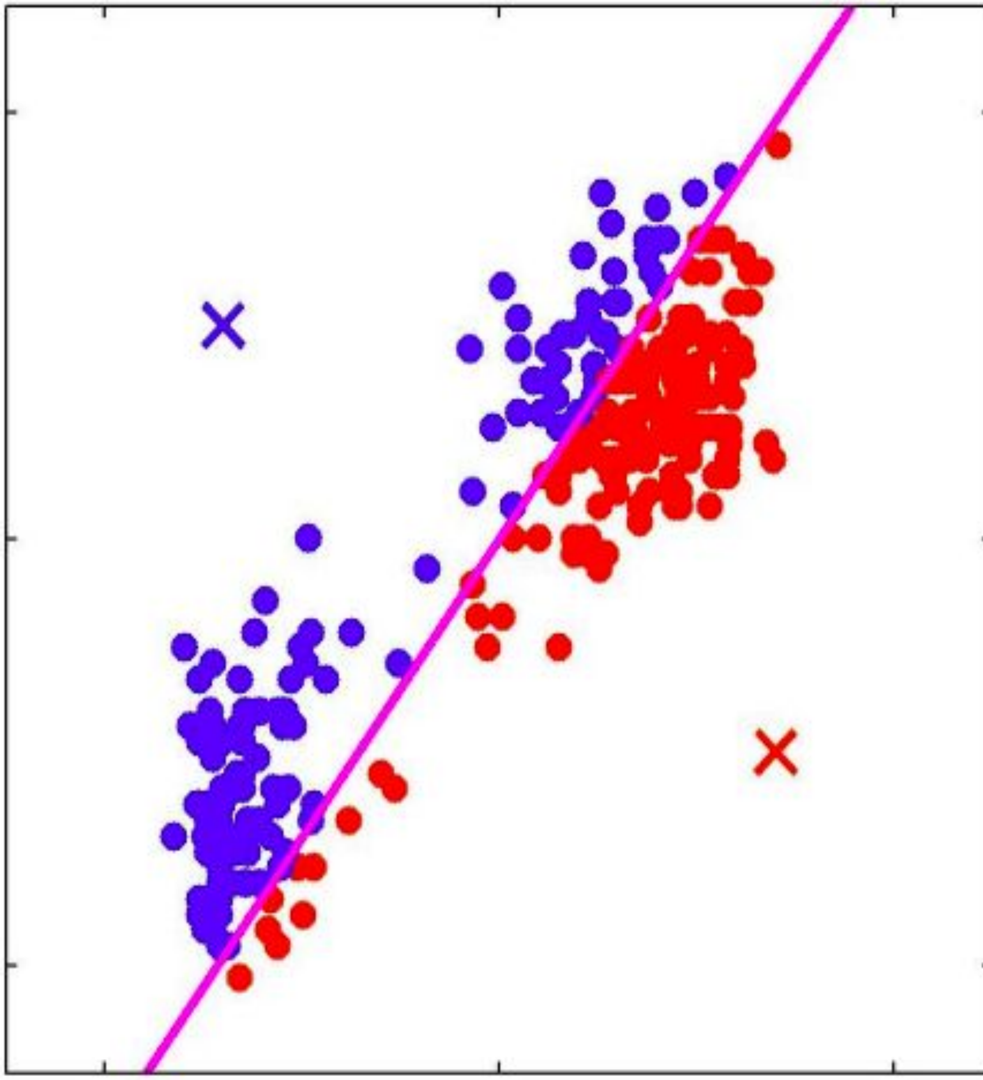
K-Means

Example



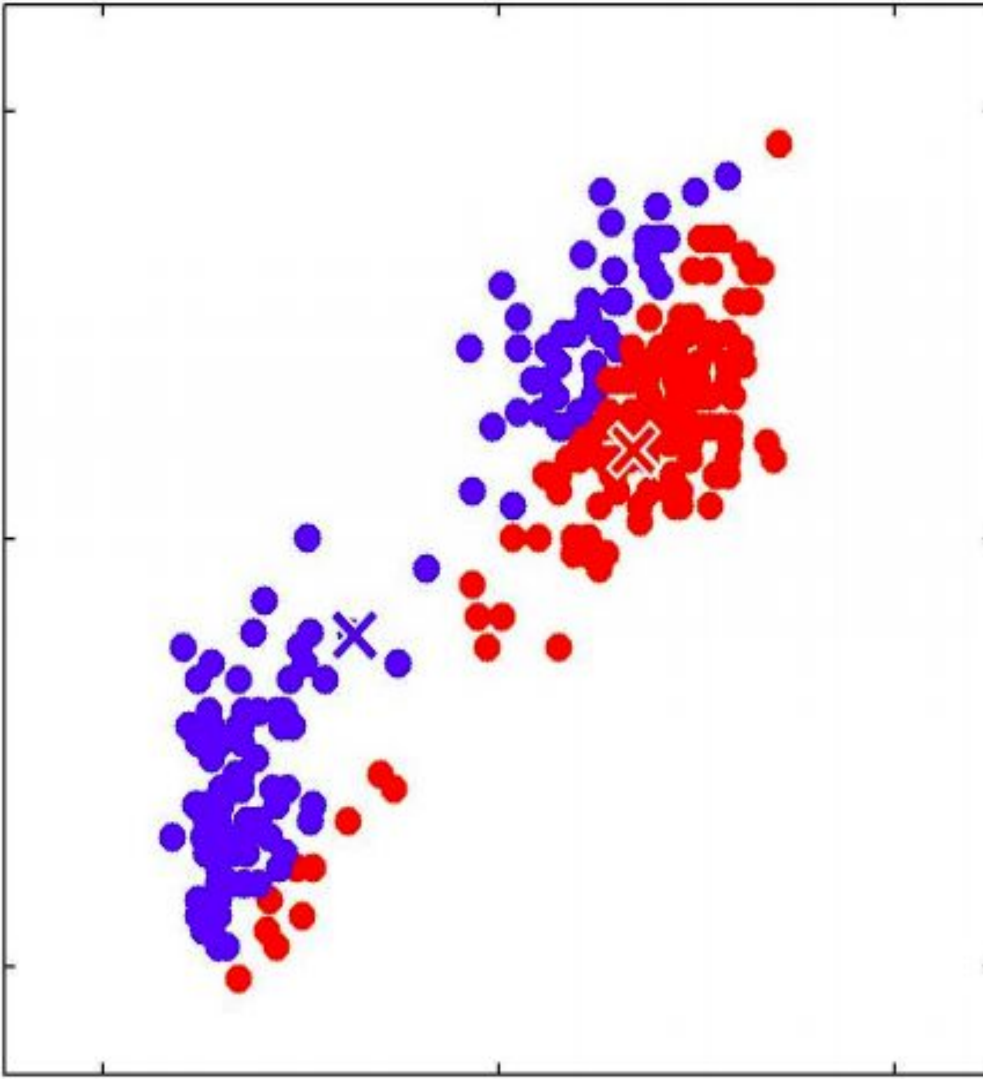
K-Means

Example



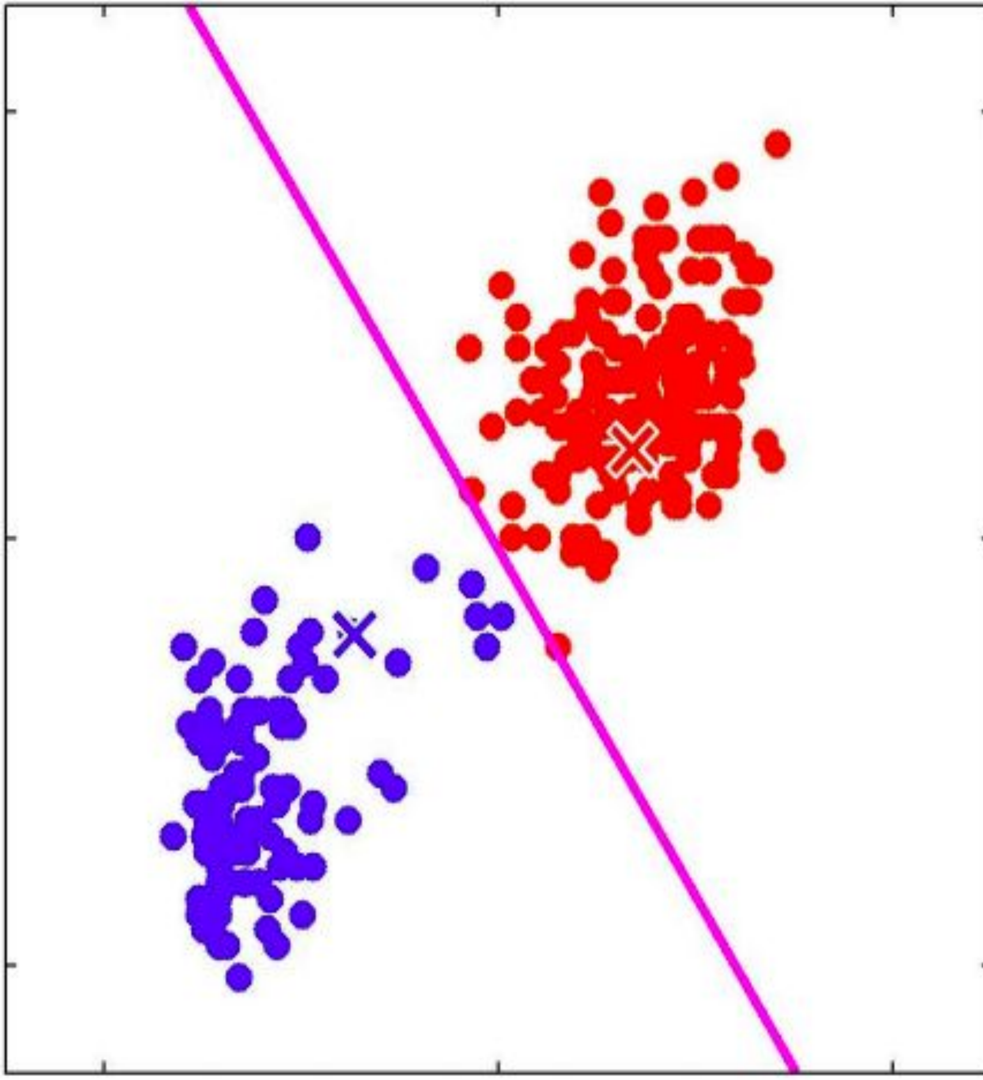
K-Means

Example



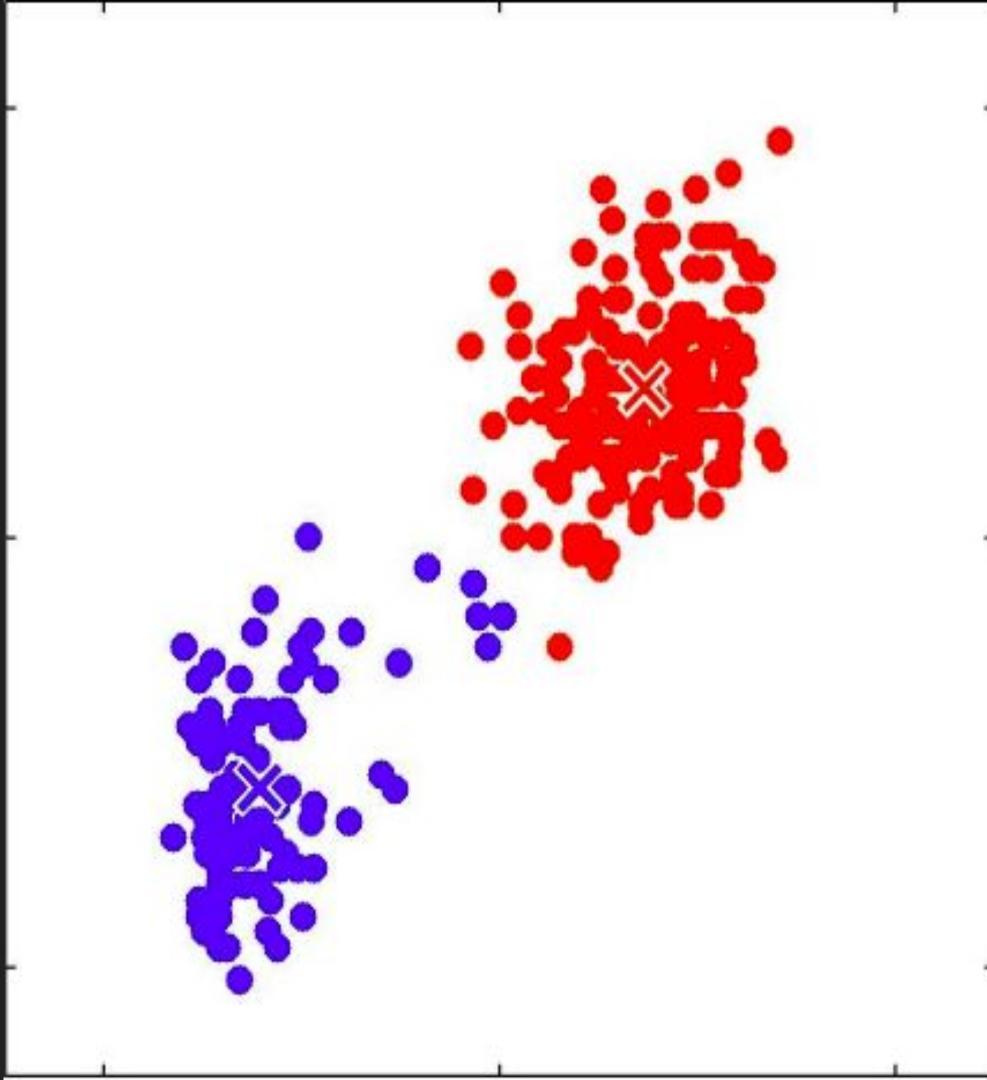
K-Means

Example



K-Means

Example



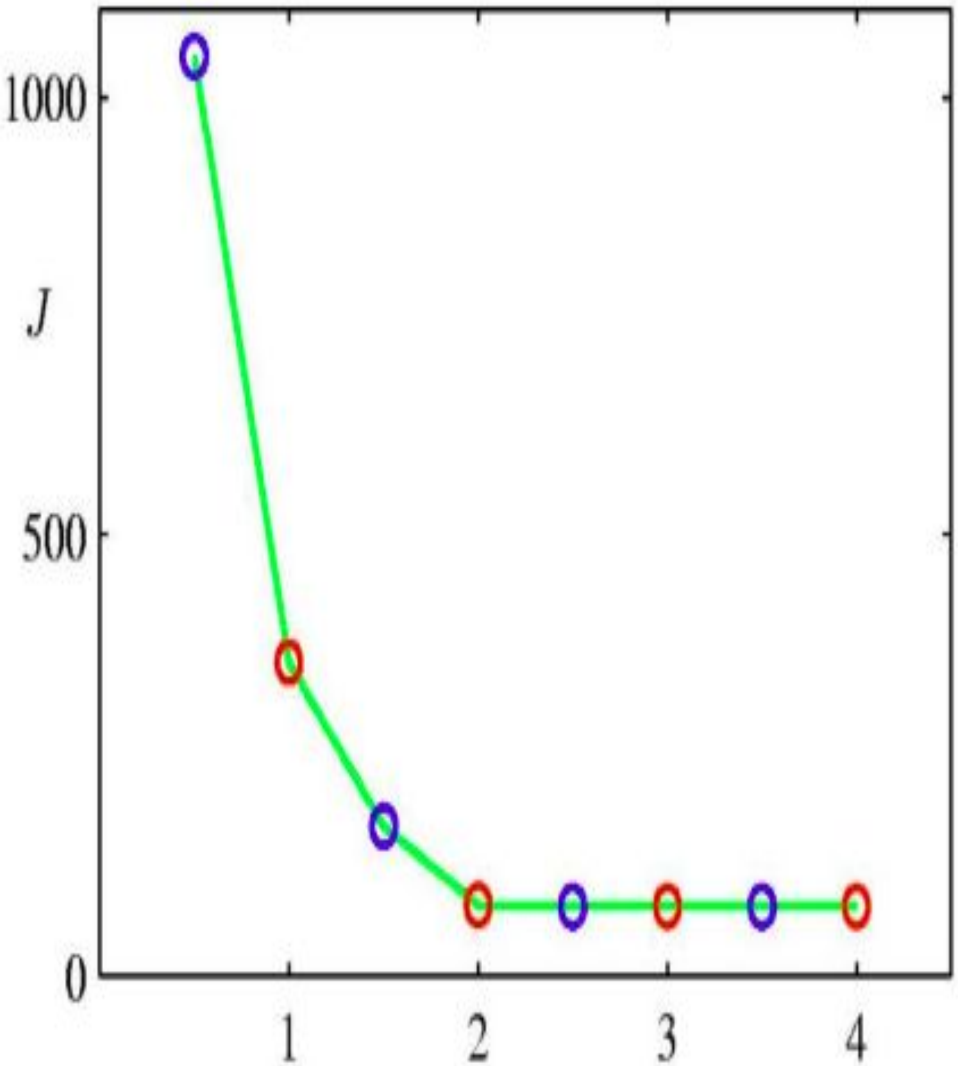
K-Means

Choosing K?

One way to select **K** for the **K-means** algorithm is to try different values of **K**, plot the K-means objective versus K, and look at the “**elbow-point**” in the plot.

K-Means

Choosing K?



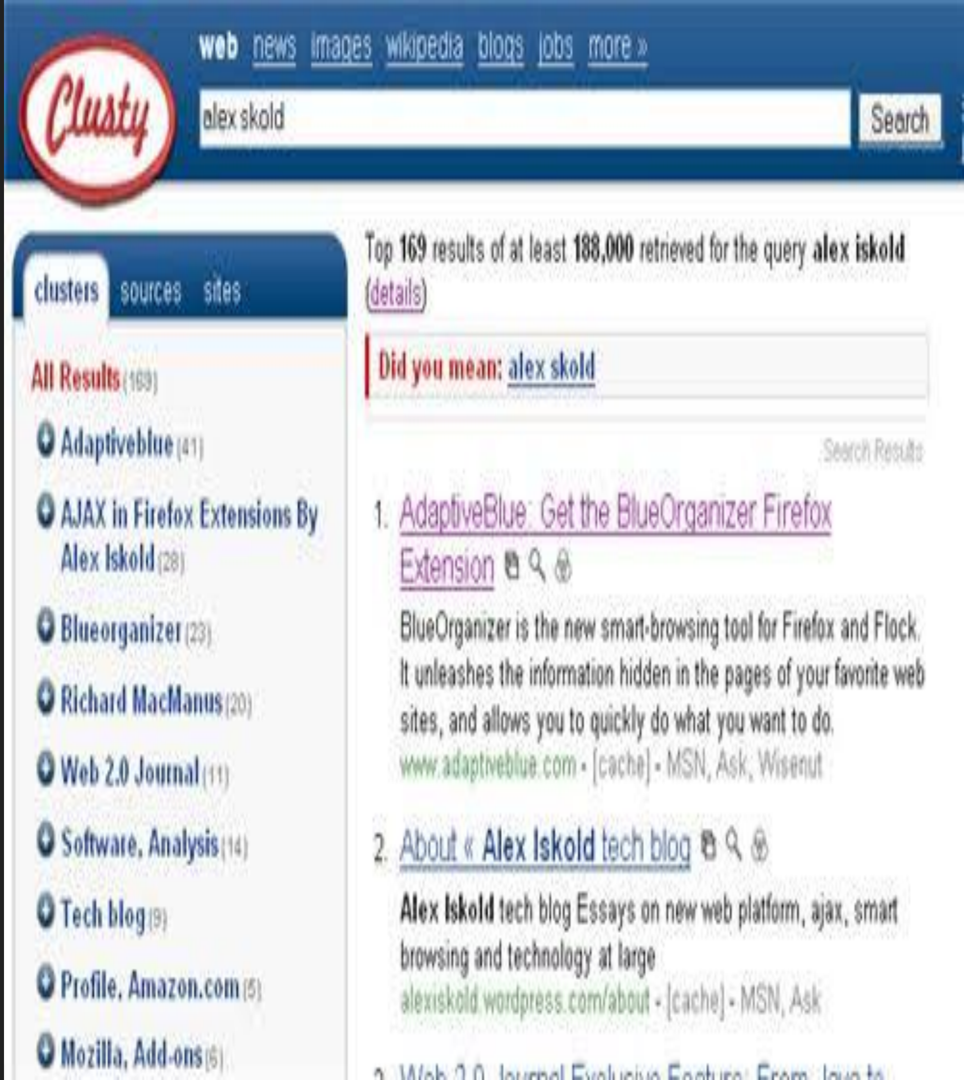
K-Means

Limitations

- **Hard assignments** of data points to clusters.
- **Sensitive** to outliers.
- Works poorly on **non-convex** clusters.

K-Means

Document Classification



Top 169 results of at least 188,000 retrieved for the query **alex iskold** (details)

Did you mean: alex skold

All Results (169)

- Adaptiveblue (41)
- AJAX in Firefox Extensions By Alex Iskold (28)
- Blueorganizer (23)
- Richard MacManus (20)
- Web 2.0 Journal (11)
- Software, Analysis (14)
- Tech blog (9)
- Profile, Amazon.com (5)
- Mozilla, Add-ons (6)

1. AdaptiveBlue: Get the BlueOrganizer Firefox Extension

   BlueOrganizer is the new smart-browsing tool for Firefox and Flock. It unleashes the information hidden in the pages of your favorite web sites, and allows you to quickly do what you want to do.

   www.adaptiveblue.com - [cache] - MSN, Ask, Wisenut

2. About « Alex Iskold tech blog

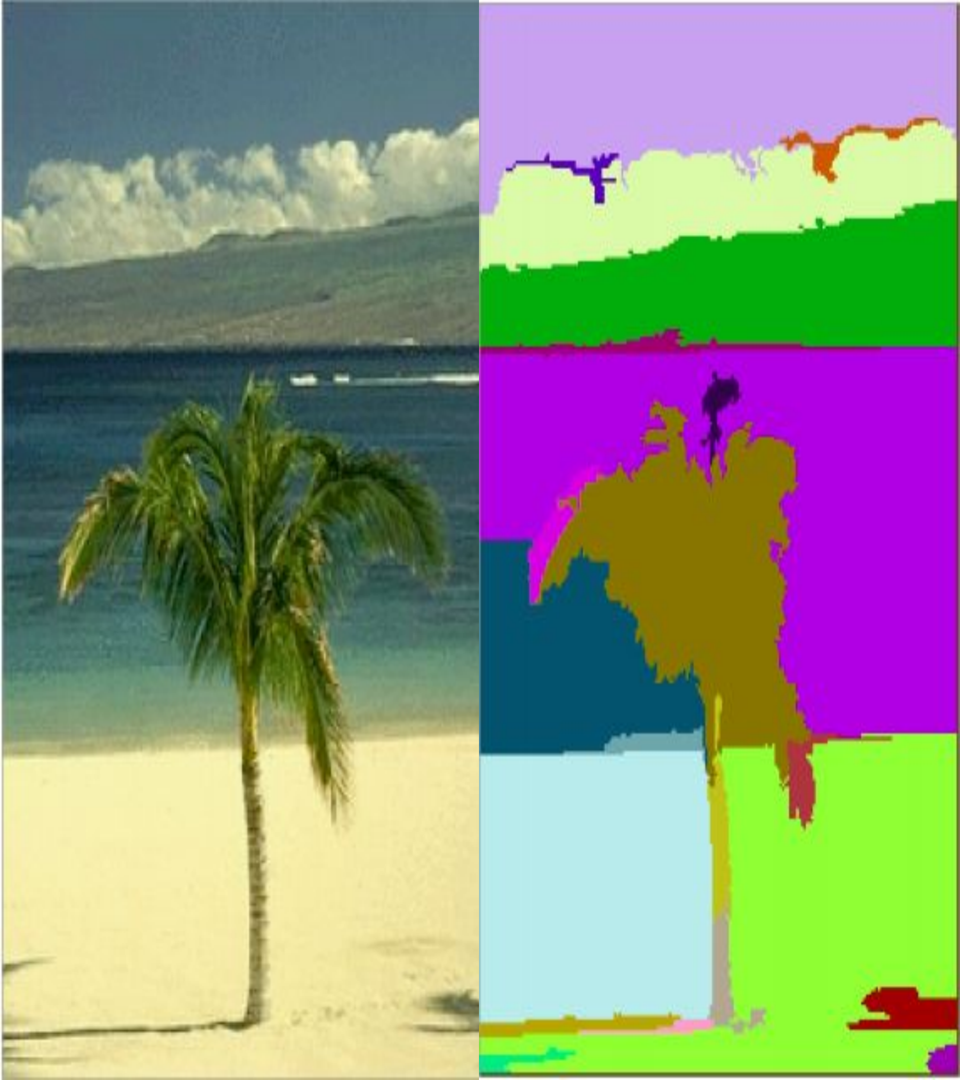
   Alex Iskold tech blog Essays on new web platform, ajax, smart browsing and technology at large

   alexiskold.wordpress.com/about - [cache] - MSN, Ask

3. Web 2.0 Journal Exclusive Feature: From Java to

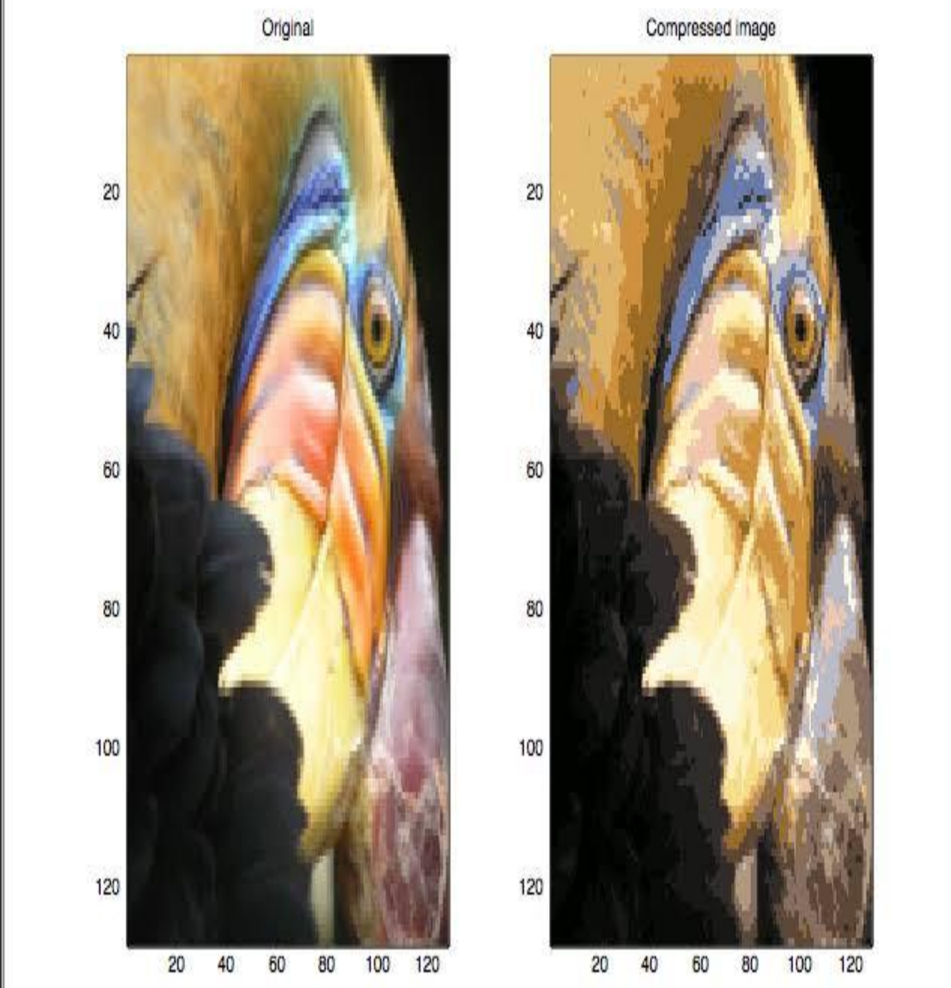
K-Means

Image Segmentation



K-Means

Image Compression



Question?

PUZZLE-TIME

You have 15 L.E with you. You go to a shop and shopkeeper tells you price as 1 L.E per chocolate. He also tells you that you can get a chocolate in return of 3 wrappers. How many maximum chocolates you can eat?

Spectral Clustering

What is Spectral Clustering?

Spectral Clustering

Spectral clustering is a technique with roots in **graph theory**, where the approach is used to identify **communities of nodes** in a graph based on the edges connecting them. The method is **flexible** and allows us to **cluster non graph data** as well.

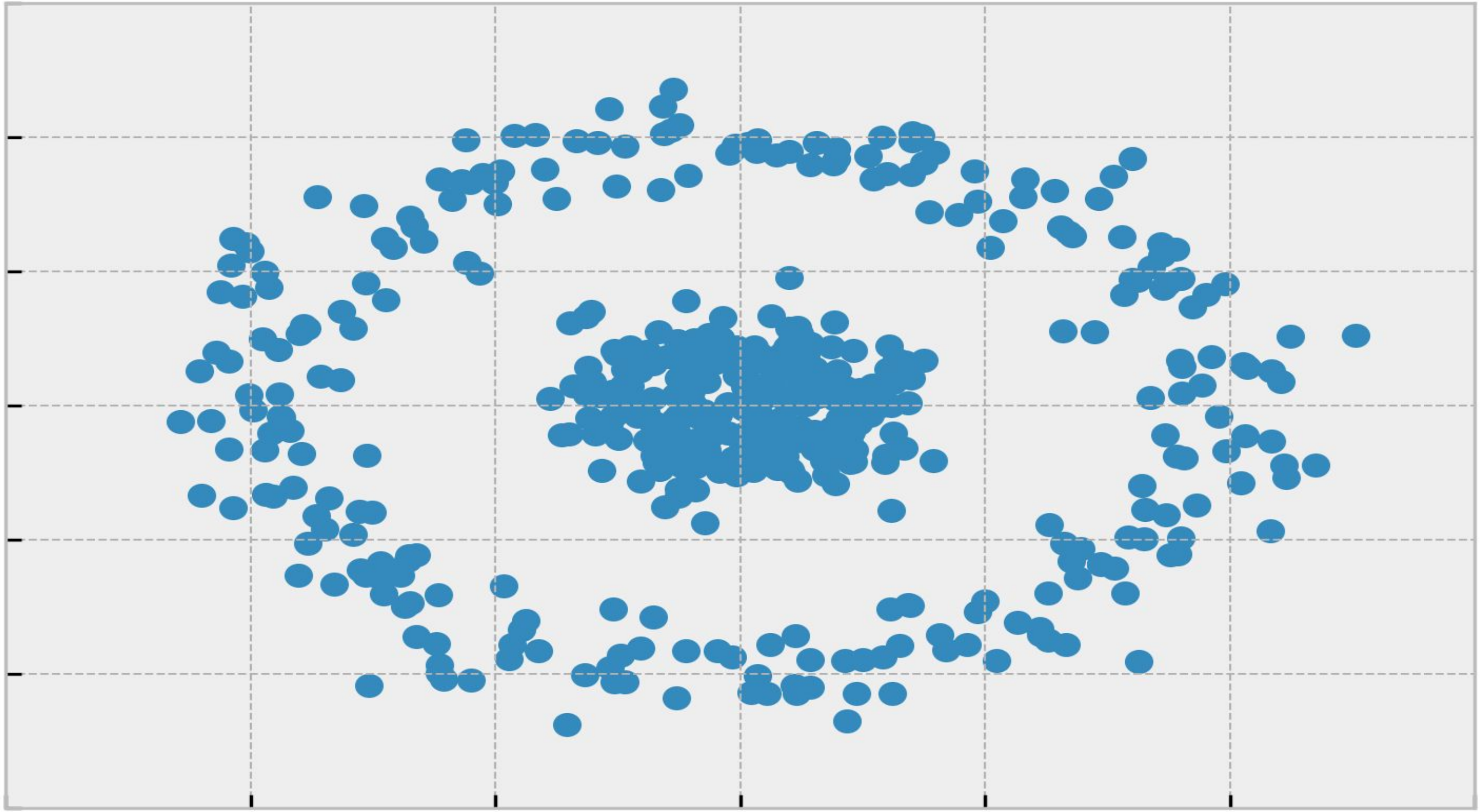
Spectral Clustering

Spectral clustering uses information from the **eigenvalues** (spectrum) of special matrices built from the graph or the data set.

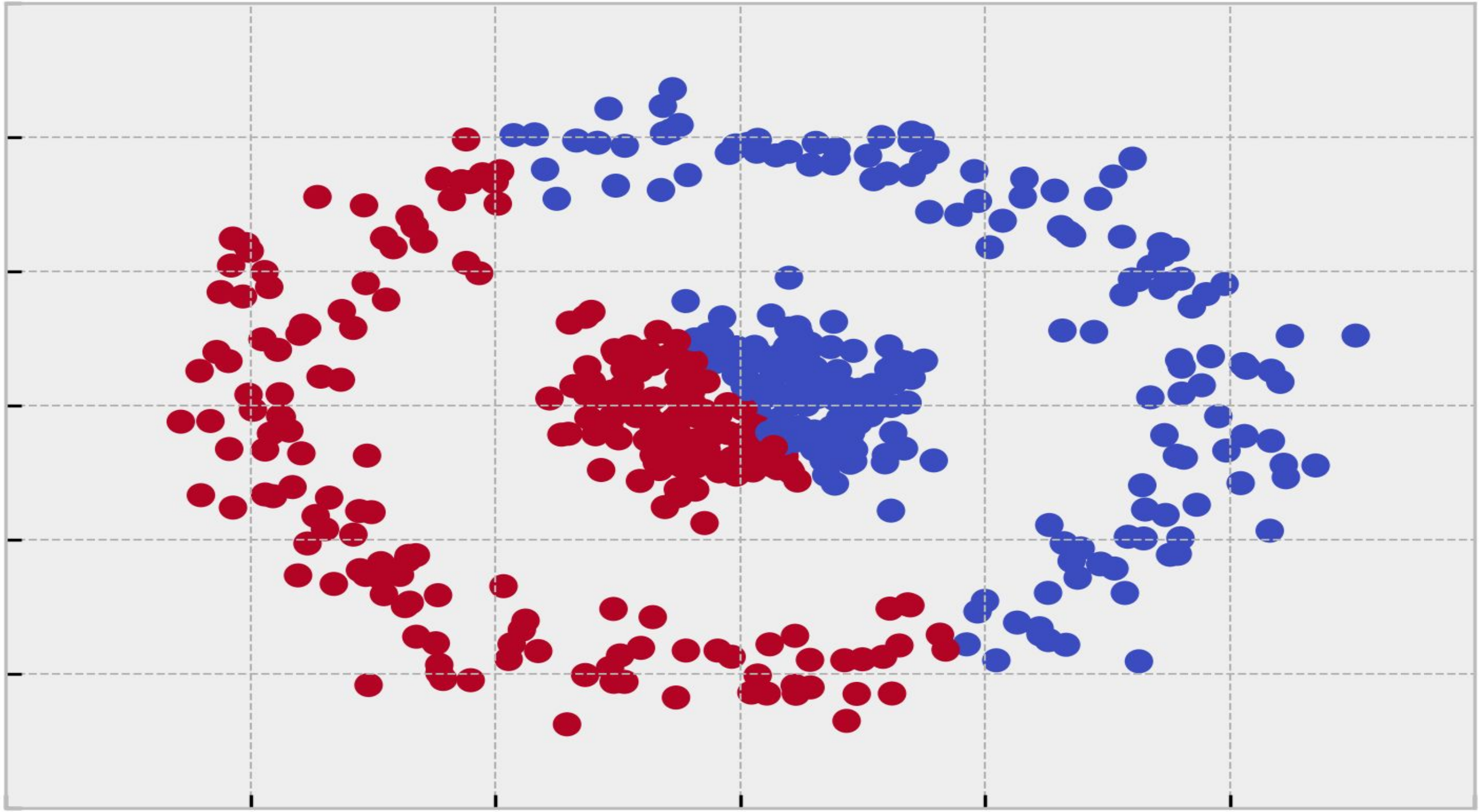
Spectral Clustering

There are a many ways to treat our data as a **graph**. The easiest way is to construct a **k-nearest neighbors graph**. A k-nearest neighbors graph treats every data point as a node in a graph. An edge is then drawn from each node to its k nearest neighbors in the original space. Generally, the algorithm isn’t too sensitive of the choice of k. Smaller numbers like 5 or 10 usually work pretty well.

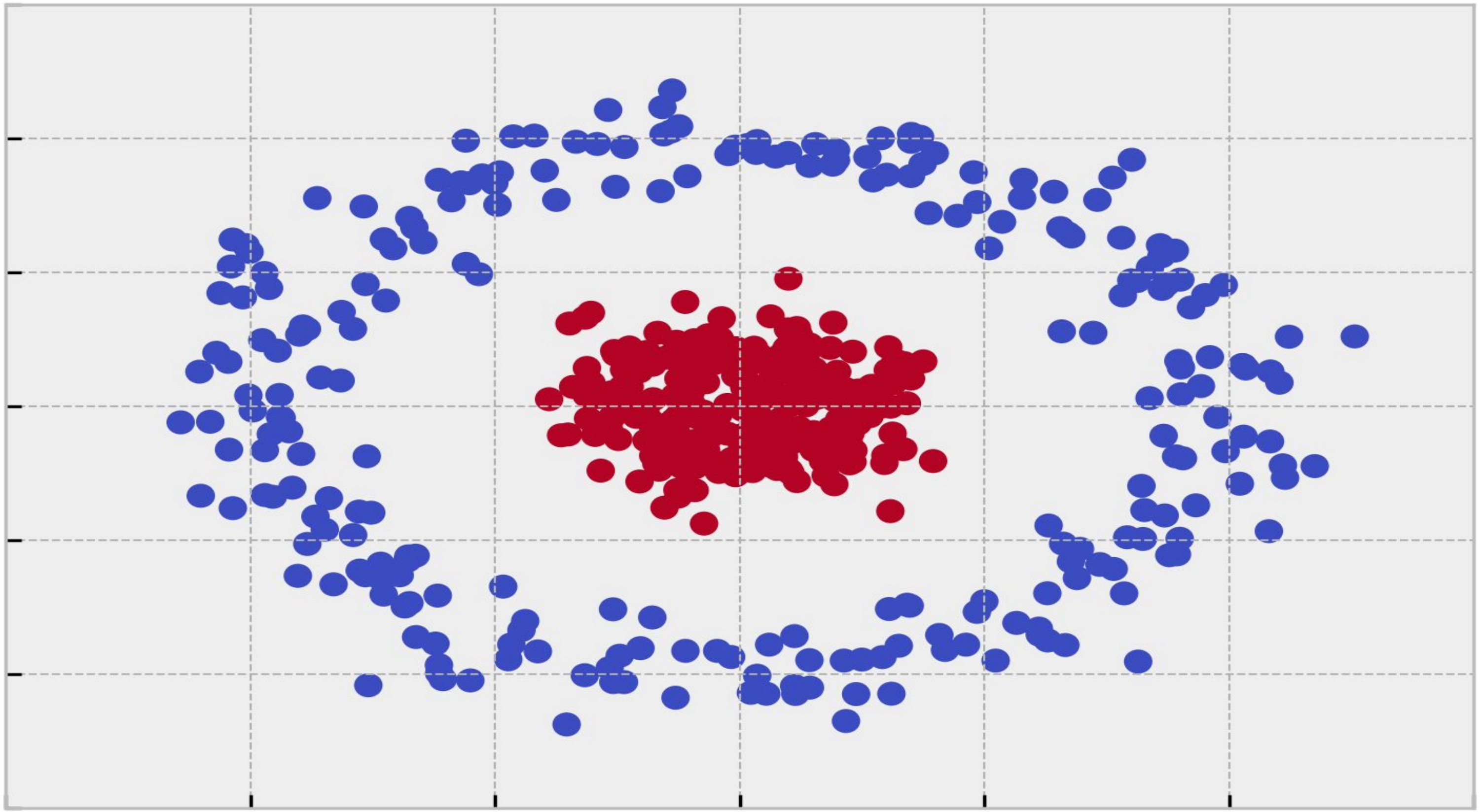
Circles



K-Means Circles



Spectral Circles



Question?

Clustering Validation

Clustering Validation

Goals

- **Evaluation**
  - *Quality*
- **Stability**
  - *Sensitivity to parameters used*
- **Tendency**
  - *Ability to find groups in data if exists*

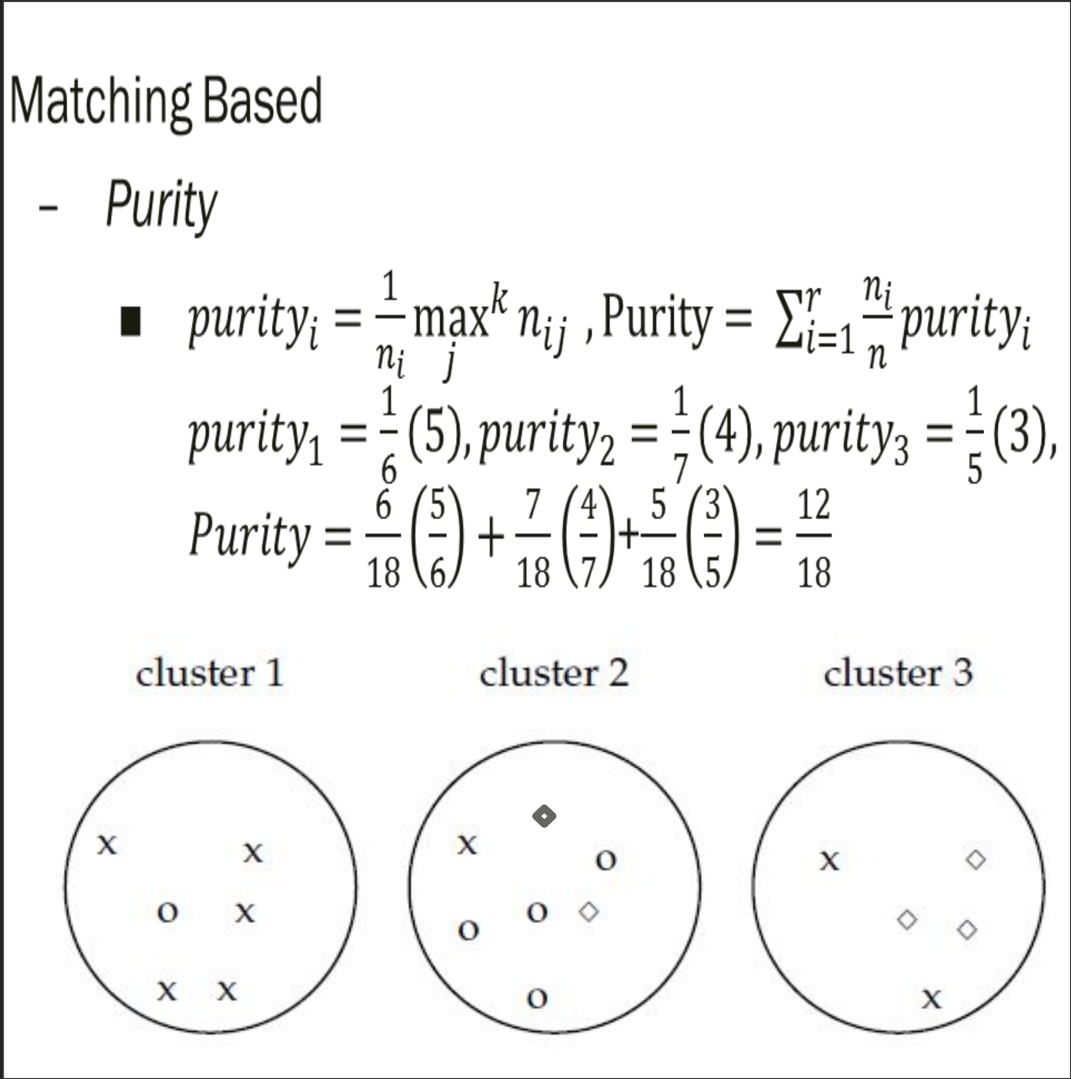
Clustering Validation

Types

- **External**
  - *Expert specified knowledge.*
- **Internal**
  - *Measures derived from the data.*
- **Relative**
  - *Compare different clustering output , to set the best parameters.*

Clustering Validation

Purity



Matching Based

- *Purity*
  - $purity_i = \frac{1}{n_i} \max_{j}^{k} n_{ij}$, Purity = $\sum_{i=1}^{r} \frac{n_i}{n} purity_i$

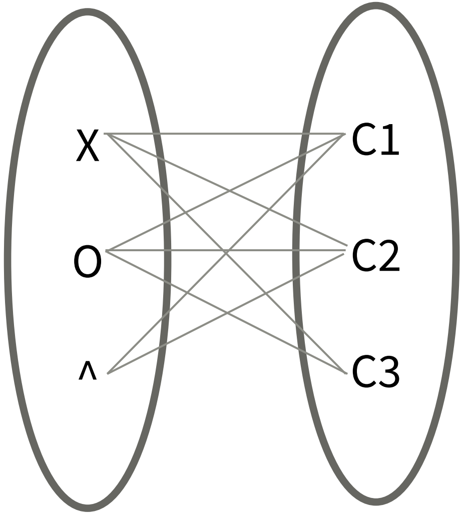
    $purity_1 = \frac{1}{6}(5), purity_2 = \frac{1}{7}(4), purity_3 = \frac{1}{5}(3),$

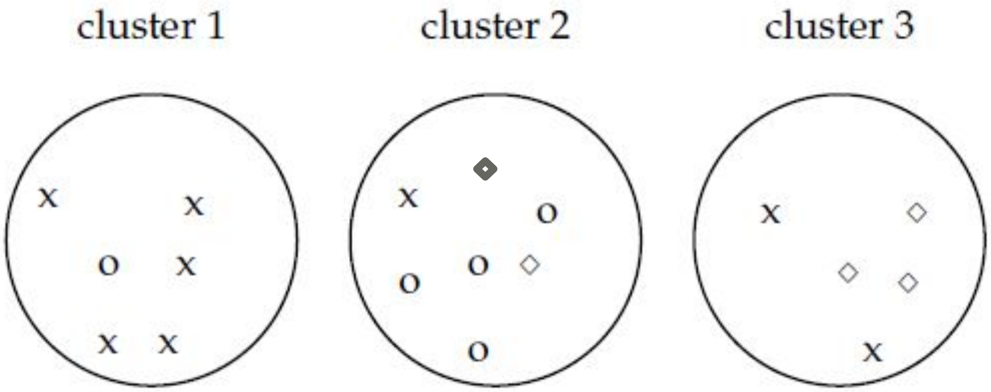
    $Purity = \frac{6}{18}\left(\frac{5}{6}\right) + \frac{7}{18}\left(\frac{4}{7}\right) + \frac{5}{18}\left(\frac{3}{5}\right) = \frac{12}{18}$

Clustering Validation

Maximum Matching

	x	O	^
C1	5	1	0
C2	1	4	2
C3	2	0	3



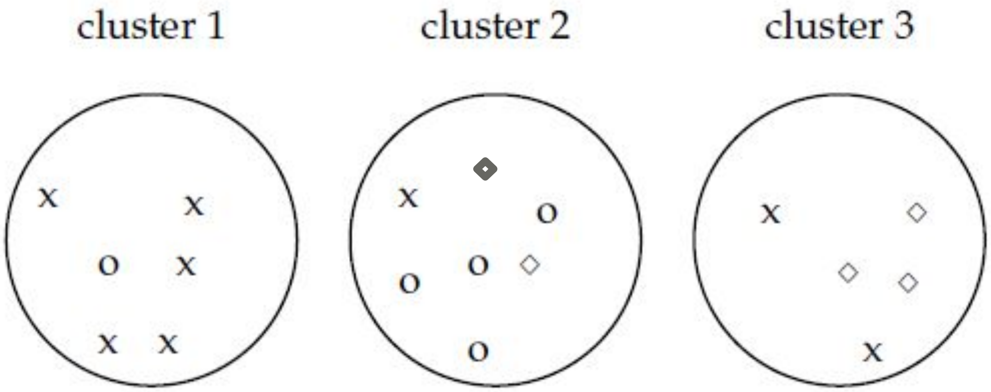


Clustering Validation

F-Measure

- Matching Based
  - *F-measure*
    - For every cluster compute
      - $prec_i = purity_i$
      - $rec_i = \frac{n_{ij_i}}{|Tj_i|}$
      - $F_i = \frac{2*prec_i*rec_i}{prec_i+rec_i}$
      - $F = \frac{1}{r}\sum_{i=1}^{r} F_i$

C1	5/6	5/8	0.714
C2	4/7	4/5	0.666
C3	3/5	3/5	0.666
	F		0.684



PUZZLE-TIME

Know Average Salary without Disclosing Individual Salaries?

Question?

Python (K-Means)

Thank You <3