Big Data Computing

Master's Degree in Computer Science

2025-2026

Gabriele Tolomei

Department of Computer Science

Sapienza Università di Roma

tolomei@di.uniroma1.it

Sapienza Università di Roma

Recap from Last Lecture

- Focus on hard partitioning clustering
- Formulate hard partitioning clustering as a (**non-convex**) optimization problem
  - Minimizing "some" aggregated internal cluster distance
- Computing exact solution is **NP-hard** due to exponential search space
- Use an iterative (approximate) solution → e.g., **K-means**

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs
- Tries to minimize the SSD between each point and its assigned centroid

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs
- Tries to minimize the SSD between each point and its assigned centroid
- It iteratively repeats **2 steps**:

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs
- Tries to minimize the SSD between each point and its assigned centroid
- It iteratively repeats **2 steps**:
  - **Assignment step** → assign each point to the closest centroid

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs
- Tries to minimize the SSD between each point and its assigned centroid
- It iteratively repeats **2 steps**:
  - **Assignment step** → assign each point to the closest centroid
  - **Update step** → recompute centroids

One-Slide K-means

- A greedy algorithm to find K hard partitions of N inputs
- Tries to minimize the SSD between each point and its assigned centroid
- It iteratively repeats **2 steps**:
    - **Assignment step** → assign each point to the closest centroid
    - **Update step** → recompute centroids
- The final output depends on key choices (**local optimum**)

Alternative Seed Choice: K-means++

- A method to carefully select initial centroids

Alternative Seed Choice: K-means++

- A method to carefully select initial centroids
- Proposed in 2007 by Arthur and Vassilvitskii [paper]

Alternative Seed Choice: K-means++

- A method to carefully select initial centroids
- Proposed in 2007 by Arthur and Vassilvitskii [paper]
- **Intuition:** Spreading out the K initial cluster centers is a good thing

Alternative Seed Choice: K-means++

- A method to carefully select initial centroids
- Proposed in 2007 by Arthur and Vassilvitskii [paper]
- **Intuition:** Spreading out the K initial cluster centers is a good thing
- Select the i-th centroid as the farthest data point to any other already selected centroids

Alternative Seed Choice: K-means++

1. Choose **one** centroid uniformly at random from among initial data points

Alternative Seed Choice: K-means++

1. Choose **one** centroid uniformly at random from among initial data points
2. For each data point $\mathbf{x}$, compute $D(\mathbf{x})$ as the distance between $\mathbf{x}$ and the **nearest centroid** already chosen

Alternative Seed Choice: K-means++

1. Choose **one** centroid uniformly at random from among initial data points
2. For each data point $\mathbf{x}$, compute $D(\mathbf{x})$ as the distance between $\mathbf{x}$ and the **nearest centroid** already chosen
3. Choose one new data point at random as a new centroid with probability proportional to $D(\mathbf{x})^2$

Alternative Seed Choice: K-means++

1. Choose **one** centroid uniformly at random from among initial data points
2. For each data point $\mathbf{x}$, compute $D(\mathbf{x})$ as the distance between $\mathbf{x}$ and the **nearest centroid** already chosen
3. Choose one new data point at random as a new centroid with probability proportional to $D(\mathbf{x})^2$
4. Repeat steps 2. and 3. until K centroids are chosen, then run Lloyd-Forgy

"Vanilla" K-means vs. K-means++

- Random initialization of "vanilla" K-means may give clusters that are **arbitrarily worse** than optimum

"Vanilla" K-means vs. K-means++

- Random initialization of "vanilla" K-means may give clusters that are **arbitrarily worse** than optimum
- K-means++ provides an upper-bound to the approximation obtained w.r.t. the optimal solution

"Vanilla" K-means vs. K-means++

- Random initialization of "vanilla" K-means may give clusters that are **arbitrarily worse** than optimum
- K-means++ provides an upper-bound to the approximation obtained w.r.t. the optimal solution
- At most, clusters obtained with K-means++ initialization are $O(\log K)$ worse than the optimal partitioning

K-means: How Many Clusters?

- Number of clusters K is given
  - Great! Partition N data points into a predetermined number K of clusters
  - Unfortunately, it is very uncommon to know K in advance

K-means: How Many Clusters?

- Number of clusters K is given
  - Great! Partition N data points into a predetermined number K of clusters
  - Unfortunately, it is very uncommon to know K in advance
- Finding the "right" number K of clusters is part of the problem!
  - Trade-off between having too few and too many clusters
  - Total benefit vs. Total cost

K-means: Total Benefit

- Given a clustering, define the benefit b_i for a data point $\mathbf{x}_i$ to be the similarity to its assigned centroid

K-means: Total Benefit

- Given a clustering, define the benefit b_i for a data point $\mathbf{x}_\mathrm{i}$ to be the similarity to its assigned centroid
- Define the total benefit B to be the sum of the individual benefits

K-means: Total Benefit

- Given a clustering, define the benefit b_i for a data point $\mathbf{x}_i$ to be the similarity to its assigned centroid
- Define the total benefit B to be the sum of the individual benefits

NOTE
There is always a clustering whose total benefit B=N
(where N is the number of data points)

Why?

K-means: Total Cost

- Assign a cost p to each cluster, thereby a clustering with K clusters has a total cost $P=Kp$

K-means: Total Cost

- Assign a cost p to each cluster, thereby a clustering with K clusters has a total cost $P=Kp$
- Define the value V of a clustering to be total benefit-total cost

$$\boxed{V = B - P}$$

K-means: Total Cost

- Assign a cost p to each cluster, thereby a clustering with K clusters has a total cost $P=Kp$
- Define the value V of a clustering to be total benefit-total cost

$$\boxed{V = B - P}$$

Goal:
Find the clustering which maximizes V, over all choices of K

K-means: Total Cost

- Assign a cost p to each cluster, thereby a clustering with K clusters has a total cost $P=Kp$
- Define the value V of a clustering to be total benefit-total cost $\boxed{V = B-P}$

Goal:
Find the clustering which maximizes V, over all choices of K

B increases with larger values of K, but P allows to stop that

K-means: "Elbow" Method

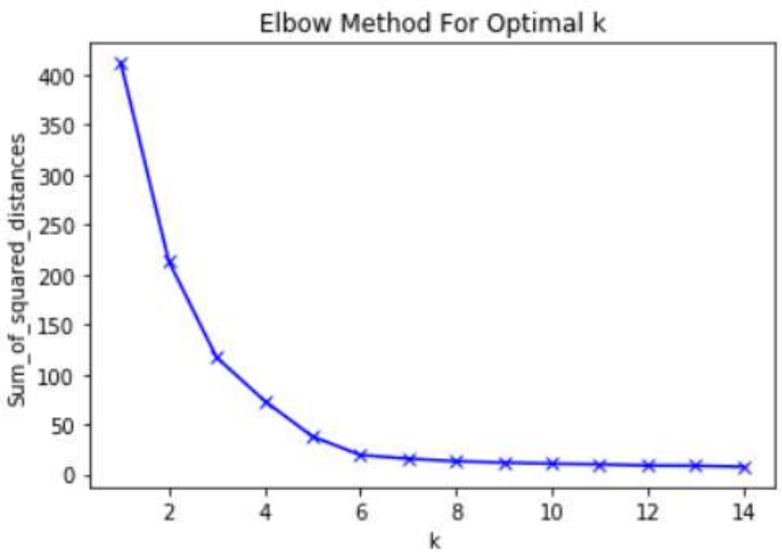
- Empirical method to figure out the right number K of clusters

K-means: "Elbow" Method

- Empirical method to figure out the right number K of clusters
- Trade-off between total benefit and total cost

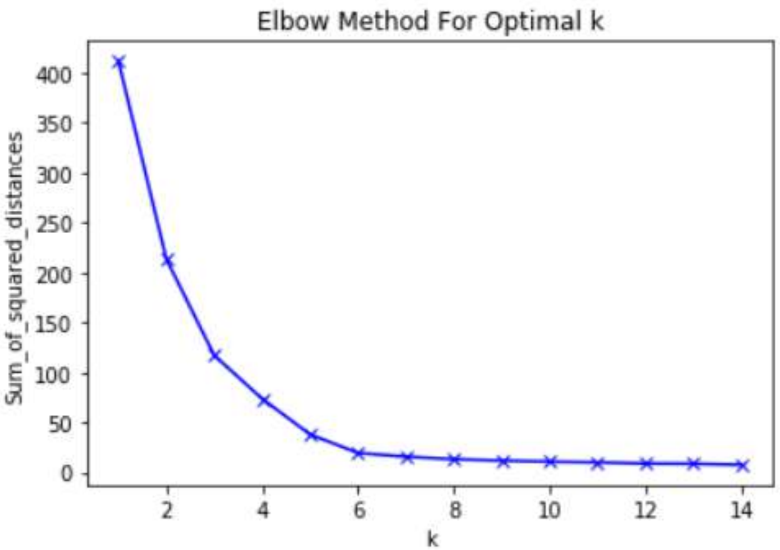
K-means: "Elbow" Method

- Empirical method to figure out the right number K of clusters
- Trade-off between total benefit and total cost
- Try multiple values of K and look at the change of the SSD



K-means: "Elbow" Method

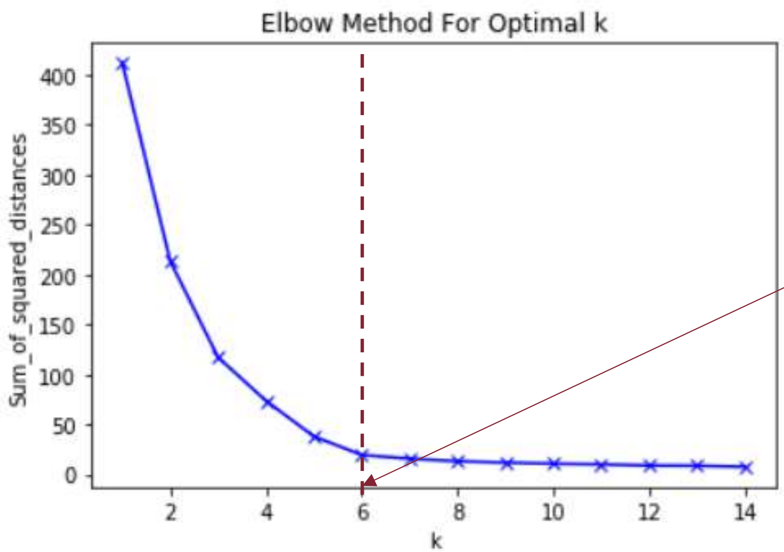
- Empirical method to figure out the right number K of clusters
- Trade-off between total benefit and total cost
- Try multiple values of K and look at the change of the SSD



As K increases, SSD sharply decreases

K-means: "Elbow" Method

- Empirical method to figure out the right number K of clusters
- Trade-off between total benefit and total cost
- Try multiple values of K and look at the change of the SSD



As K increases, SSD sharply decreases

up to a certain value

Non-Euclidean Distances

- So far, we have focused on **Euclidean distance** (i.e., δ = L^2-Norm)

Non-Euclidean Distances

- So far, we have focused on **Euclidean distance** (i.e., δ = L^2-Norm)
- The same hard clustering framework can be used with other δ

Non-Euclidean Distances

- So far, we have focused on **Euclidean distance** (i.e., δ = L^2-Norm)
- The same hard clustering framework can be used with other δ
- Some of them just resemble Euclidean distance, and centroids (i.e., means) still minimize those

Non-Euclidean Distances

- So far, we have focused on **Euclidean distance** (i.e., δ = L^2-Norm)
- The same hard clustering framework can be used with other δ
- Some of them just resemble Euclidean distance, and centroids (i.e., means) still minimize those
  - δ = **Cosine distance** = Euclidean distance on normalized input points
  - δ = **Correlation** = Euclidean distance on standardized input points

Non-Euclidean Distances

- So far, we have focused on **Euclidean distance** (i.e., δ = L^2-Norm)
- The same hard clustering framework can be used with other δ
- Some of them just resemble Euclidean distance, and centroids (i.e., means) still minimize those
  - δ = **Cosine distance** = Euclidean distance on normalized input points
  - δ = **Correlation** = Euclidean distance on standardized input points
- Others, require specific minimizers
  - δ = **Manhattan distance** (L^1-Norm) → median is the minimizer (**K-medians**)

Alternative Formulations: K-medoids

- Similar to K-means yet chooses input data points as centers (**medoids**)

Alternative Formulations: K-medoids

- Similar to K-means yet chooses input data points as centers (**medoids**)
- A medoid is the closest object to any other point in the cluster

Alternative Formulations: K-medoids

- Similar to K-means yet chooses input data points as centers (**medoids**)
- A medoid is the closest object to any other point in the cluster
- Works with **any arbitrary distance** δ

Alternative Formulations: K-medoids

- Similar to K-means yet chooses input data points as centers (**medoids**)
- A medoid is the closest object to any other point in the cluster
- Works with **any arbitrary distance** δ
- **PAM** (**P**artitioning **A**round **M**edoids) greedy Algorithm, introduced by Kaufman and Rousseeuw in 1987 [paper] vs. Lloyd-Forgy

Alternative Formulations: K-medoids

- Similar to K-means yet chooses input data points as centers (**medoids**)
- A medoid is the closest object to any other point in the cluster
- Works with **any arbitrary distance** δ
- PAM (Partitioning Around Medoids) greedy Algorithm, introduced by Kaufman and Rousseeuw in 1987 [paper] vs. Lloyd-Forgy
- Robust to outliers yet computationally expensive $O(K(N-K)^2)$

Bradley-Fayyad-Reina (BFR) K-means

- A variant of K-means explicitly thought for large datasets

Bradley-Fayyad-Reina (BFR) K-means

- A variant of K-means explicitly thought for large datasets
- Works better in high-dimensional Euclidean space

Bradley-Fayyad-Reina (BFR) K-means

- A variant of K-means explicitly thought for large datasets
- Works better in high-dimensional Euclidean space
- (Strong) Assumption on the shape of clusters:
  - Normally distributed around the centroid
  - Independence between data dimensions

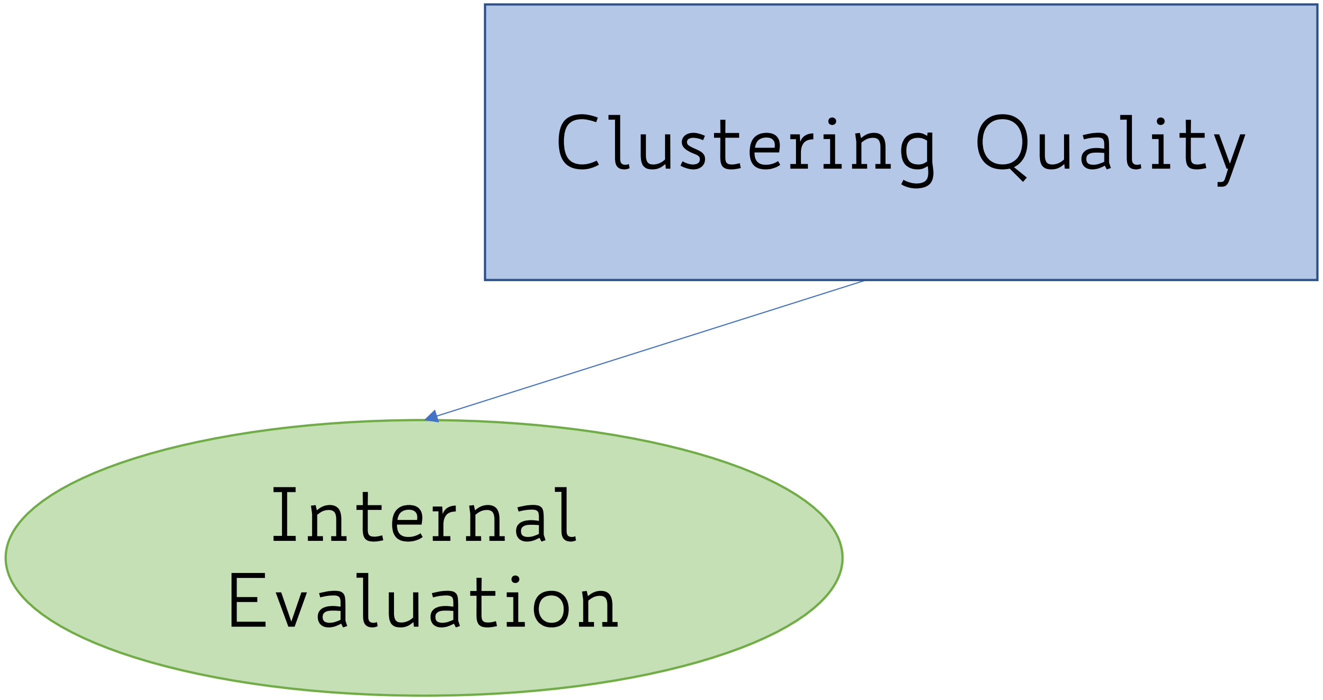
Bradley-Fayyad-Reina (BFR) K-means

- A variant of K-means explicitly thought for large datasets
- Works better in high-dimensional Euclidean space
- (Strong) Assumption on the shape of clusters:
  - Normally distributed around the centroid
  - Independence between data dimensions
- Reference to the original paper

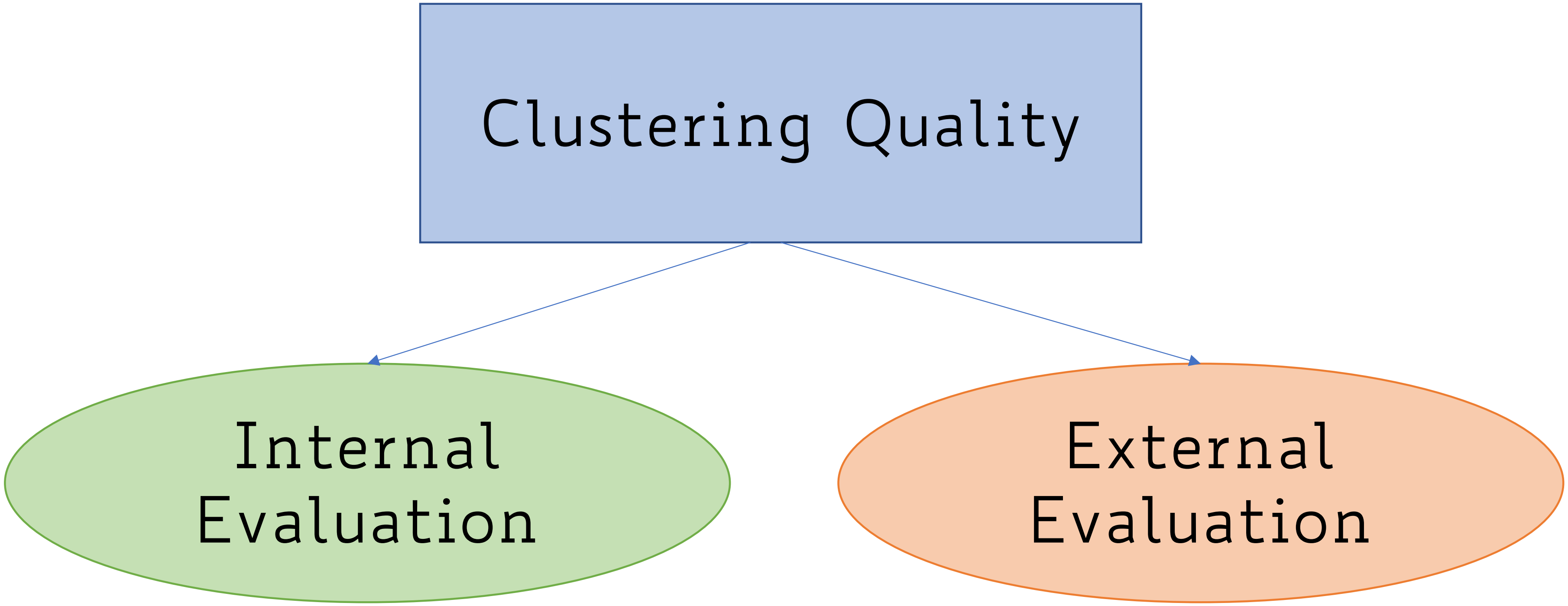
Measures of Clustering Quality

Clustering Quality

Measures of Clustering Quality



Measures of Clustering Quality



Internal Evaluation

- Clustering is evaluated based on the data that was clustered itself

Internal Evaluation

- Clustering is evaluated based on the data that was clustered itself
- A good clustering will produce high quality clusters with:
  - high intra-cluster similarity
  - low inter-cluster similarity

Internal Evaluation

- Clustering is evaluated based on the data that was clustered itself
- A good clustering will produce high quality clusters with:
  - high intra-cluster similarity
  - low inter-cluster similarity
- The measured quality of a clustering depends on
  - data representation
  - similarity measure

Internal Evaluation: Davies-Bouldin Index

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{\delta(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)}\right)$$

$$\begin{aligned} K &= \text{number of clusters} \\ \boldsymbol{\mu}_k &= \text{centroid of cluster } C_k \\ \sigma_k &= \text{avg. distance of all elements of cluster } C_k \text{ from its centroid } \boldsymbol{\mu}_k \\ \delta(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) &= \text{distance between centroids of } C_i \text{ and } C_j \end{aligned}$$

Internal Evaluation: Davies-Bouldin Index

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{\delta(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)} \right)$$

$$K = \text{number of clusters}$$
$$\boldsymbol{\mu}_k = \text{centroid of cluster } C_k$$
$$\sigma_k = \text{avg. distance of all elements of cluster } C_k \text{ from its centroid } \boldsymbol{\mu}_k$$
$$\delta(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \text{distance between centroids of } C_i \text{ and } C_j$$

The smaller the better

Internal Evaluation: Dunn Index

$$D = \frac{\min_{1 \leq i < j \leq K} \delta(C_i, C_j)}{\max_{1 \leq k \leq K} \delta'(C_k)}$$

$$K = \text{number of clusters}$$
$$\delta(C_i, C_j) = \text{distance between cluster } C_i \text{ and } C_j$$
$$\delta'(C_k) = \text{intra-cluster distance of cluster } C_k$$

Distance between centroids

Max distance between any pair of objects

Internal Evaluation: Dunn Index

$$D = \frac{\min_{1 \le i < j \le K} \delta(C_i, C_j)}{\max_{1 \le k \le K} \delta'(C_k)}$$

K = number of clusters

$\delta(C_i, C_j)$ = distance between cluster C_i and C_j

$\delta'(C_k)$ = intra-cluster distance of cluster C_k

Distance between centroids

Max distance between any pair of objects

The higher the better

Internal Evaluation: Silhouette Coefficient

mean distance between i and all other data points in the same cluster C_i

Internal Evaluation: Silhouette Coefficient

mean distance between i and all other data points in the same cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} \delta(i, j)$$

Internal Evaluation: Silhouette Coefficient

smallest mean distance of i to all points in any other cluster C_k != C_i

Internal Evaluation: Silhouette Coefficient

smallest mean distance of i to all points in any other cluster C_k != C_i

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} \delta(i, j)$$

Internal Evaluation: Silhouette Coefficient

mean distance between i and all other data points in the same cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} \delta(i, j)$$

smallest mean distance of i to all points in any other cluster C_k != C_i

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} \delta(i, j)$$

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i) \end{cases}$$

Internal Evaluation: Silhouette Coefficient

mean distance between i and all other data points in the same cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} \delta(i, j)$$

smallest mean distance of i to all points in any other cluster C_k != C_i

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} \delta(i, j)$$

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i) \end{cases}$$

The higher the better

External Evaluation

- Clustering is evaluated based on data that was not used for clustering, yet pre-classified (**gold standard** data)

External Evaluation

- Clustering is evaluated based on data that was not used for clustering, yet pre-classified (**gold standard** data)
- Quality measured by the ability to discover some or all of the hidden patterns in gold standard data

External Evaluation

- Clustering is evaluated based on data that was not used for clustering, yet pre-classified (**gold standard** data)
- Quality measured by the ability to discover some or all of the hidden patterns in gold standard data
- Hard as it requires labeled data typically provided by human experts

External Evaluation: Purity

$$C_1 \ldots, C_K = \text{set of } K \text{ clusters}$$

$$L_1 \ldots, L_J = \text{set of } J \text{ labels}$$

$$n_{i,j} = \text{number of items with label } L_j \text{ clustered in } C_i$$

$$n_i = \sum_{j=1}^{J} n_{i,j} \text{ number of items clustered in } C_i$$

$$\text{purity}(C_i) = \frac{1}{n_i} \max_{j \in \{1,\ldots,J\}} n_{i,j}$$

$$\text{purity} = \frac{1}{K} \sum_{i=1}^{K} \text{purity}(C_i)$$

External Evaluation: Purity

$$C_1 \ldots, C_K = \text{set of } K \text{ clusters}$$

$$L_1 \ldots, L_J = \text{set of } J \text{ labels}$$

$$n_{i,j} = \text{number of items with label } L_j \text{ clustered in } C_i$$

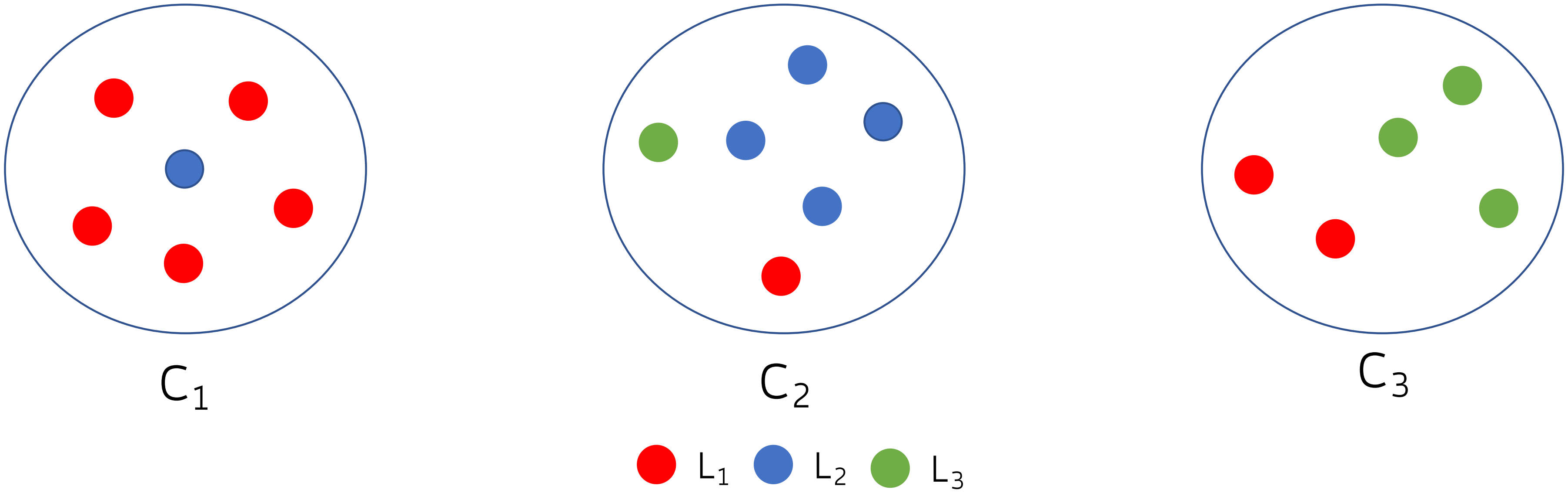
$$n_i = \sum_{j=1}^{J} n_{i,j} \text{ number of items clustered in } C_i$$

$$\text{purity}(C_i) = \frac{1}{n_i} \max_{j \in \{1,\ldots,J\}} n_{i,j}$$

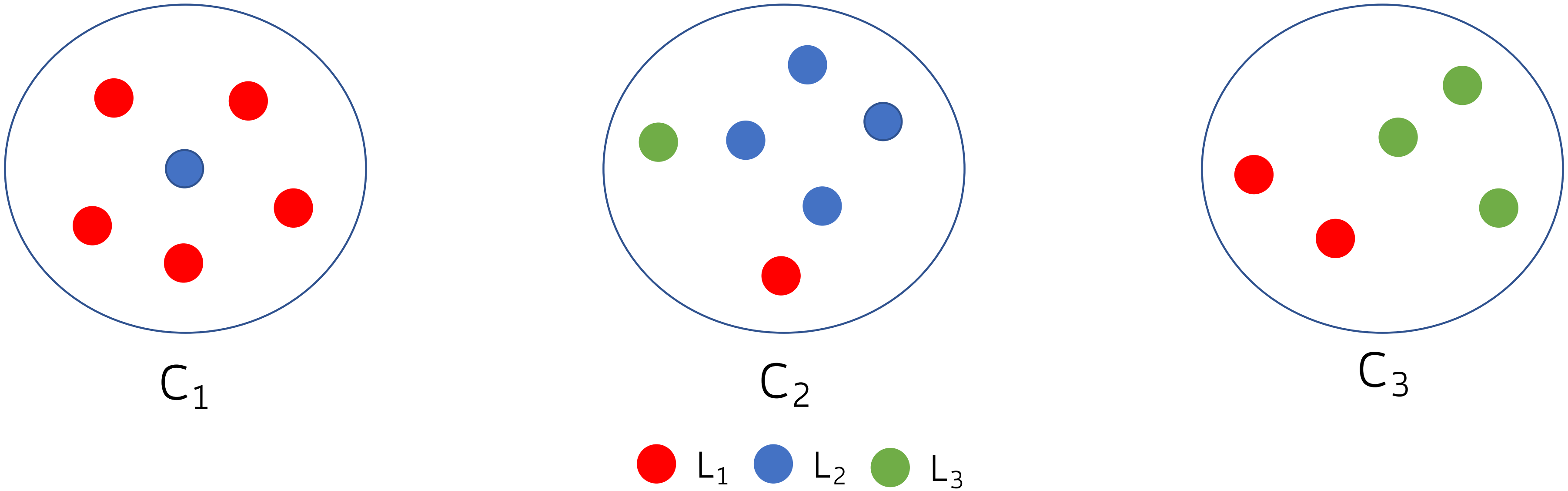
$$\text{purity} = \frac{1}{K} \sum_{i=1}^{K} \text{purity}(C_i)$$

Biased because having as many clusters as items maximizes purity

External Evaluation: Purity Example

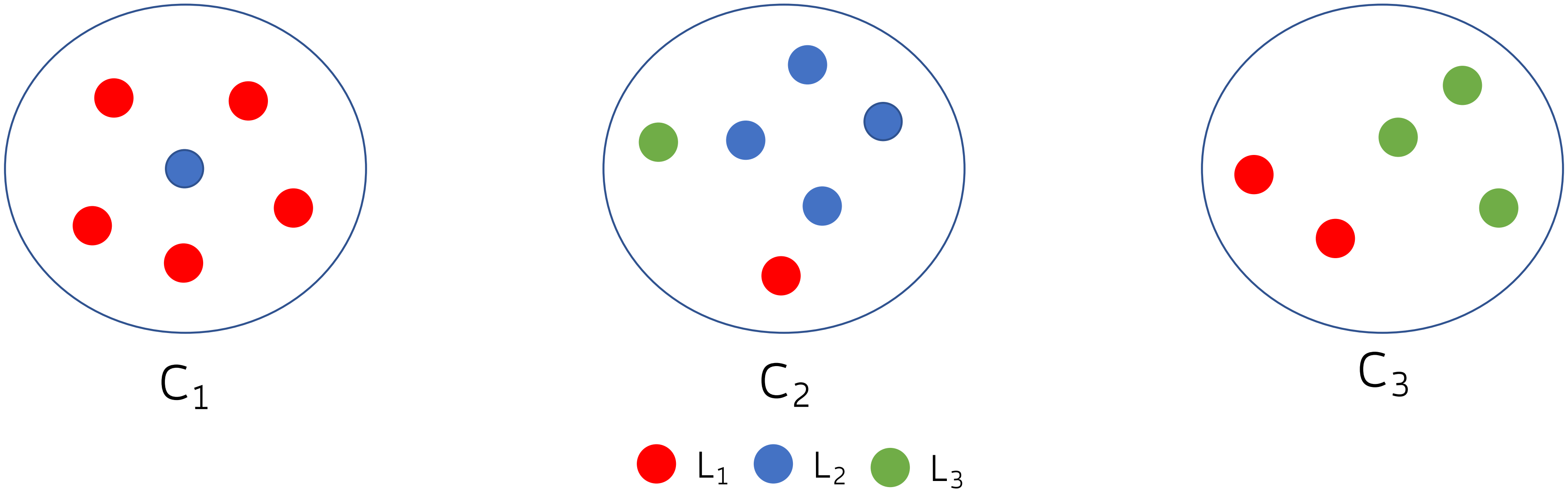


External Evaluation: Purity Example



purity(C_1) = 1/6 * max{5, 1, 0} = 5/6

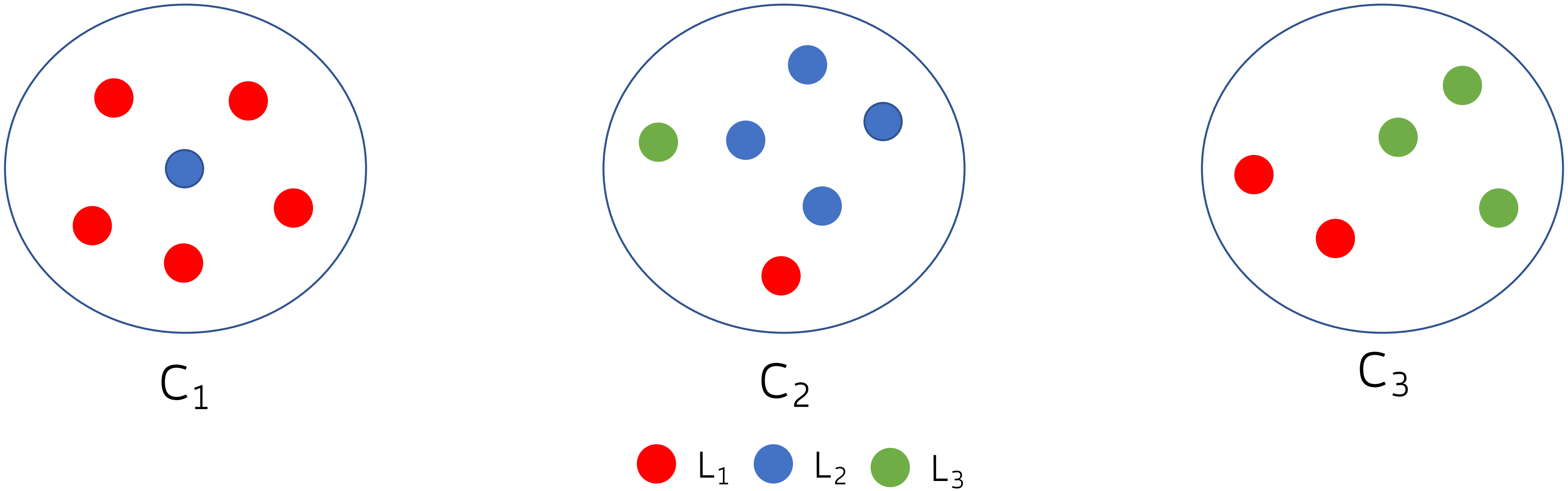
External Evaluation: Purity Example



purity(C_1) = 1/6 * max{5, 1, 0} = 5/6

purity(C_2) = 1/6 * max{1, 4, 1} = 4/6 = 2/3

External Evaluation: Purity Example

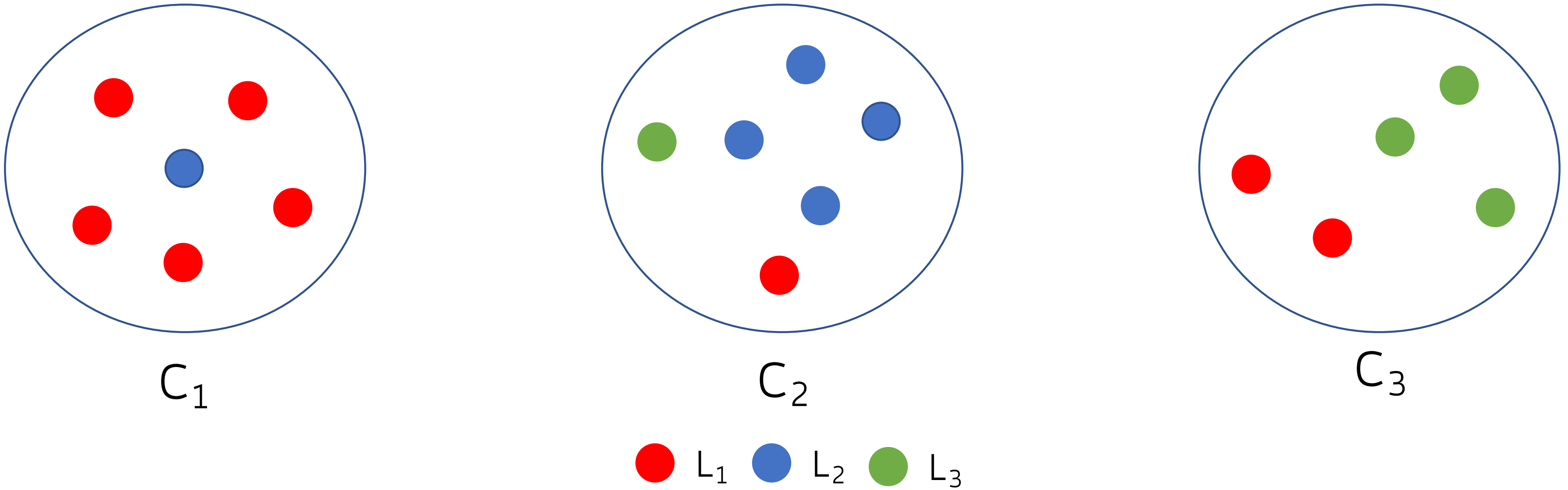


$\text{purity}(C_1) = 1/6 * \max\{5, 1, 0\} = 5/6$

$\text{purity}(C_2) = 1/6 * \max\{1, 4, 1\} = 4/6 = 2/3$

$\text{purity}(C_3) = 1/5 * \max\{2, 0, 3\} = 3/5$

External Evaluation: Purity Example



purity(C_1) = 1/6 * max{5, 1, 0} = 5/6

purity(C_2) = 1/6 * max{1, 4, 1} = 4/6 = 2/3

purity(C_3) = 1/5 * max{2, 0, 3} = 3/5

purity = 1/3 * purity(C_1)+purity(C_2)+purity(C_3) = 7/10

External Evaluation: Rand Index

$$\text{Rand} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP = number of *true positives*
TN = number of *true negatives*
FP = number of *false positives*
FN = number of *false negatives*

External Evaluation: Rand Index

$$\text{Rand} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\left.\begin{aligned}
TP &= \text{number of } \textit{true positives} \\
TN &= \text{number of } \textit{true negatives} \\
FP &= \text{number of } \textit{false positives} \\
FN &= \text{number of } \textit{false negatives}
\end{aligned}\right\}$$

All computed from pairs of elements

External Evaluation: Rand Index

$$\text{Rand} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$TP = \text{number of } \textit{true positives}$$
$$TN = \text{number of } \textit{true negatives}$$
$$FP = \text{number of } \textit{false positives}$$
$$FN = \text{number of } \textit{false negatives}$$

All computed from pairs of elements

Measures the level of agreement between clustering and ground truth

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth		
Different Clusters in Ground-Truth		

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth	TRUE POSITIVES (TP)	
Different Clusters in Ground-Truth		

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth		
Different Clusters in Ground-Truth		TRUE NEGATIVES (TN)

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth		
Different Clusters in Ground-Truth	FALSE POSITIVES (FP)	

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth		FALSE NEGATIVES (FN)
Different Clusters in Ground-Truth		

External Evaluation: Rand Index

n. of pairs	Same Cluster in Clustering	Different Clusters in Clustering
Same Cluster in Ground-Truth	TRUE POSITIVES (TP)	FALSE NEGATIVES (FN)
Different Clusters in Ground-Truth	FALSE POSITIVES (FP)	TRUE NEGATIVES (TN)

Confusion Matrix

External Evaluation: Precision, Recall, F-measure

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN}$$

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

Balances the contribution of false negatives by weighting recall through a parameter $\boldsymbol{\beta}$

External Evaluation: Many Other Measures

- Jaccard index
- Dice index
- Fowlkes-Mallows index
- Mutual information
- etc.

Take-Home Message of Today

- **K-means** is an iterative (**approximated**) clustering method that converges to a local minimum

Take-Home Message of Today

- **K-means** is an iterative (**approximated**) clustering method that converges to a local minimum
- Tries to minimize the internal sum of squared Euclidean distances (Lloyd-Forgy Algorithm)

Take-Home Message of Today

- **K-means** is an iterative (**approximated**) clustering method that converges to a local minimum
- Tries to minimize the internal sum of squared Euclidean distances (Lloyd-Forgy Algorithm)
- Many variants:
  - **K-means++**, **K-medoids** (PAM Algorithm), **BFR K-means**, etc.

Take-Home Message of Today

- **K-means** is an iterative (**approximated**) clustering method that converges to a local minimum
- Tries to minimize the internal sum of squared Euclidean distances (Lloyd-Forgy Algorithm)
- Many variants:
  - **K-means++**, **K-medoids** (PAM Algorithm), **BFR K-means**, etc.
- Internal vs. External measures of **clustering quality**