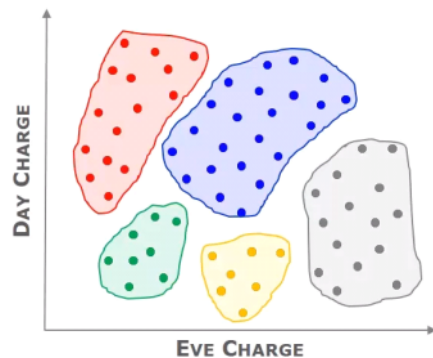


# Clustering: Evaluation

mercoledì 2 gennaio 2019 19:10

Supponiamo di voler fare una clusterizzazione su gruppi di utenti di una compagnia telefonica, il cliente dice che possono esistere secondo loro 3 o 4 tipologie di clienti, effettuando un'analisi di clusterizzazione si scopre che:



You assume that the two attributes are independent as it emerges from the plot on the left.

A clustering algorithm is designed to find clusters and thus it will find clusters in a data set, even if that data set has no natural cluster structure.

Come misurare e valutare un algoritmo di clustering?

## Most Important Issues for Cluster Validation

- **Determinare la tendenza del cluster:** di un set di dati, ad esempio distinguere strutture non random attualmente esistenti nel dataset.
- **Determinare corretto numero di clusters** (quasi si direbbe)

Per rispondere al problema della validazione dei clusters, vengono introdotti misure di validazione o indici.

Queste metriche appartengono solitamente a tre tipi:

- **External or supervised** = misure che estendono la struttura scoperta dal cluster, matchano alcune strutture esterne (Business requirements)
- **Internal or Unsupervised** = misura la bontà di una struttura di clustering senza rispettare le informazioni esterne, possono essere:
  - o **Cohesion Measure** (compactness, tightness): determinano quanto sono uniti tra loro (vicini) gli oggetti all'interno dello stesso cluster
  - o **Separation Measure** (isolation), determina come sono distinti o ben separati i cluster tra di loro
- **Relative indexes** = permette di comparare diversi clusters. Può essere supervisionato o non supervisionato. Questo tipo di misure non sono differenti rispetto alle misure di valutazione dei singoli clusters, ma sono delle misurazione specifiche.

## External or supervised

Let

$$P = \{P_1, \dots, P_R\}$$

be a partitioning, of a data set consisting of " $m$ " objects (records), into " $R$ " categories.

$$C = \{C_1, \dots, C_K\}$$

be the partition obtained with a clustering algorithm into " $K$ " clusters.

**SUPERVISED OR EXTERNAL INDICES** compare **P** to **C**:

**3(a)** **Case 1:** **x** and **y** belong to the same cluster of **C**

**SUPERVISED OR EXTERNAL INDICES** compare **P** to **C**:

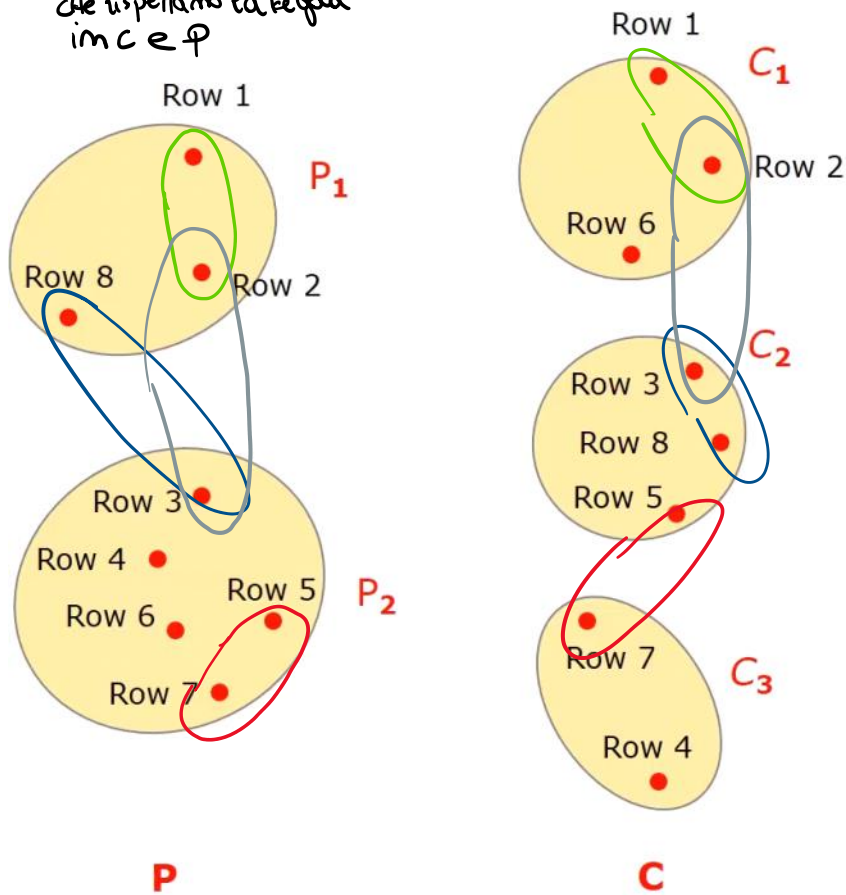
3(a) **Case 1:** **x** and **y** belong to the **same cluster** of **C**  
and to the **same category** of **P**

4(b) **Case 2:** **x** and **y** belong to the **same cluster** of **C**  
but to **different categories** of **P**

10(c) **Case 3:** **x** and **y** belong to **different clusters** of **C**  
but to the **same category** of **P**

11(d) **Case 4:** **x** and **y** belong to **different clusters** of **C**  
and to **different categories** of **P**

numero di pairs  
che rispettano la regola  
in  $C \neq P$



Possiamo quindi definire le seguenti misure:

The overall number of pairs amounts to

$$M = \frac{m \times (m-1)}{2} = a + b + c + d$$

**RAND**  $R = \frac{a + d}{M}$   **$R \in [0,1]$**

**JACCARD**  $J = \frac{a}{a + b + c}$   **$J \in [0,1]$**

**FOWLKES AND MALLOWS**  $FM = \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}}$   **$FM \in [0,1]$**

**Γ STATISTICS**  $\Gamma = \frac{M \times a - (a+b) \times (a+c)}{\sqrt{((a+b) \times (a+c)(M-a-b) \times (M-a-c))}}$   **$\Gamma \in [-1,1]$**

The larger the values, the more similar are **C** and **P**.

↖ Più il valore di questi indici è alto  
più è alta la quantità dei clusters generati

## Internal or Unsupervised

Molte misure della validità interna o di validità per gli schemi di partizione dei cluster sono basati sulla nozione di **Coesione e Separazione**

In generale, la formula utilizzata per questo tipo di metriche è la seguente:

Set di  $k$  clusters :  $C_1, \dots, C_k$

overall validity :  $\sum_{i=1}^k w_i \cdot \text{validity}(C_i)$

La funzione di validità può essere: **Coesione, Separazione** oppure una di queste combinazioni

I Pesi ( $w$ ) dipendono dalla misura di validità del cluster, in alcuni casi sono settati a 1, oppure possono essere la cardinalità del cluster corrispondente, mentre in altri casi rispecchiano proprietà più complicate, come la radice quadrata della coesione.

Validity = **COHESION** → Higher values are better

Validity = **SEPARATION** → lower values are better

Validity = **SEPARATION**  $\rightarrow$  **lower** values are better

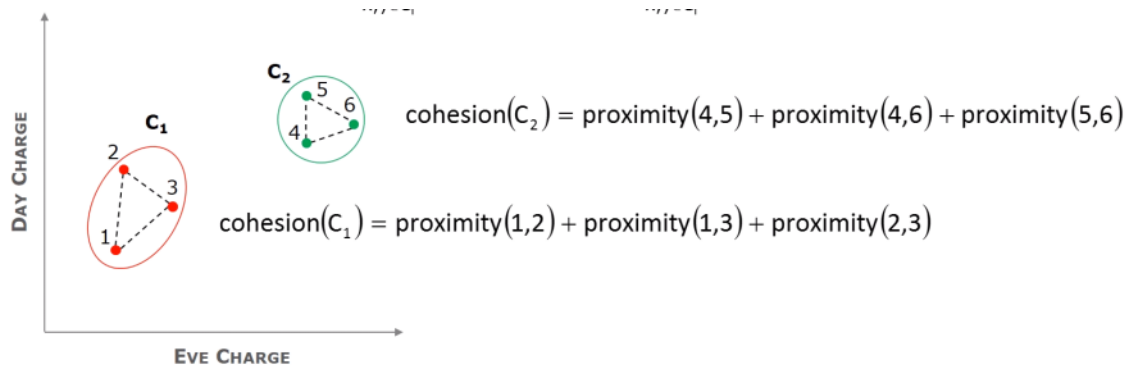
### Graph Based Clusters: Cohesion

Per i **Graph Based Clusters**, la coesione di un cluster, può essere definita come:

- Somma dei pesi dei collegamenti nel grafico di prossimità che collegano i punti all'interno del cluster

$$\text{cohesion}(C_i) = \sum_{x,y \in C_i} \text{proximity}(x,y) = \sum_{x,y \in C_i} \text{similarity}(x,y)$$

Esempio:



Attenzione: Quando si considera lo spazio degli attributi, è importante e utile ricordare che la **similarità è inversamente proporzionale alla dissimilarità/distanza**

Quindi la coesione e la similarità sono massimizzate quando la dissimilarità/distanza è minimizzata.

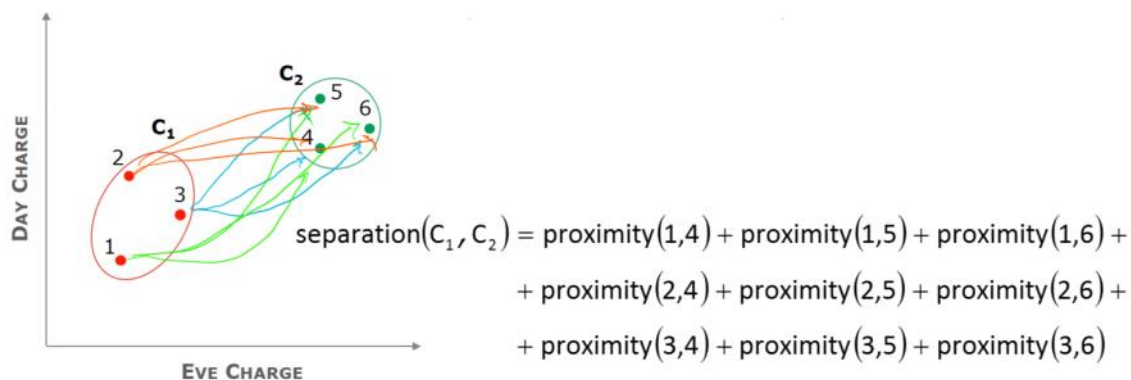
### Graph Based Clusters: Separation

La separazione invece può essere definita come segue:

- Somma dei pesi dei collegamenti da punti in un cluster a punti nell'altro cluster

$$\text{separation}(C_i, C_j) = \sum_{x \in C_i, y \in C_j} \text{proximity}(x,y) = \sum_{x \in C_i, y \in C_j} \text{similarity}(x,y)$$

Esempio:




Attenzione = anche in questo caso vale la relazione tra similarità e dissimilarità/distanza precedente

### Prototype Based Clusters: Cohesion

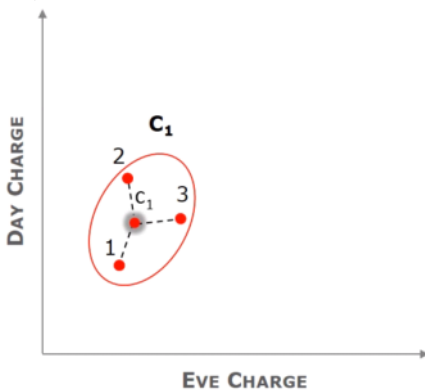
Per i **Prototype Based Clusters**, la coesione di un cluster può essere definita come:

- La somma delle prossimità rispetto al prototipo (centroide o medoide) del cluster

$$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i) = \sum_{x \in C_i} \text{similarity}(x, c_i)$$

  
**CENTROID OR MEDOID**  
**of cluster  $C_i$**

Esempio:



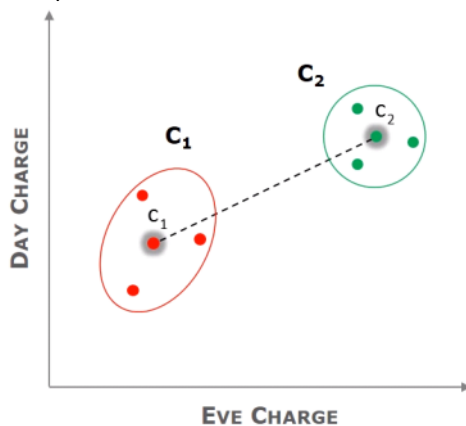
$$\text{cohesion}(C_1) = \sum_{x \in C_1} \text{proximity}(x, c_1) = \text{proximity}(1, c_1) + \text{proximity}(2, c_1) + \text{proximity}(3, c_1)$$

### Prototype Based Clusters: Separation

Per i Prototype Based Clusters, la separazione tra due clusters può essere misurata come

$$\text{separation}(C_i, C_j) = \text{proximity}(c_i, c_j) = \text{similarity}(c_i, c_j)$$

Esempio:



$$\text{separation}(C_1, C_2) = \text{proximity}(c_1, c_2)$$

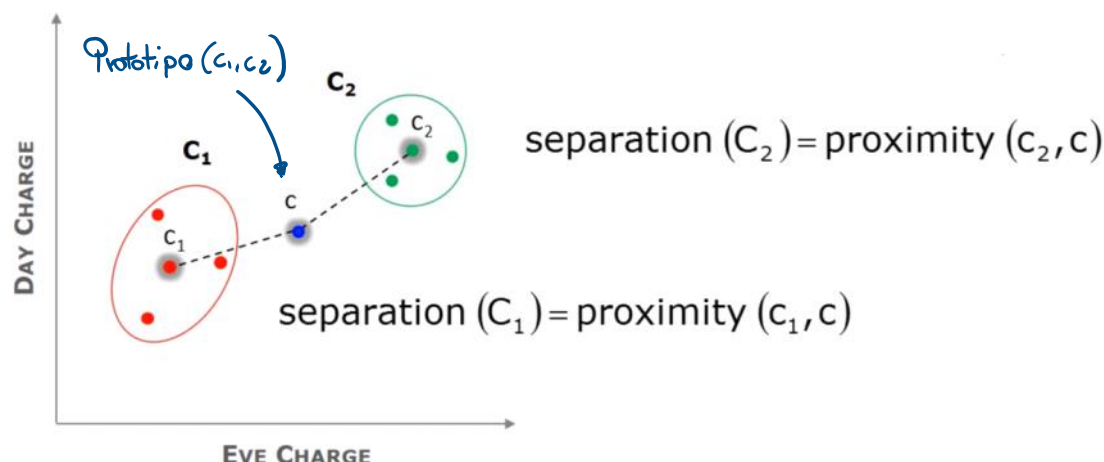
Esiste anche una possibile alternativa per rappresentare la separazione tra due clusters:

- Utilizzare un **Centroide Prototipo**

$$\text{separation}(C_i) = \text{proximity}(c_i, c) = \text{similarity}(c_i, c)$$

↓  
**OVERALL PROTOTYPE**

Ad esempio:



### *Pesi associati all'overall Validity*

Abbiamo definito quindi la cohesione e la separazione per la validità dei cluster, ma per completare la formula di overall validity introdotta precedentemente:

$$\text{overall validity} = \sum_{i=1}^k w_i \cdot \text{validity}(C_i)$$

Abbiamo bisogno di definire ancora il concetto di peso ( $w$ ) da usare.

I pesi possono essere molti, ma tipicamente esprimono: **la misura della grandezza del cluster**

Alcuni di questi esempi sono:

CLUSTER MEASURE	CLUSTER WEIGHT	TYPE
$\text{cohesion}(C_i) = \sum_{x, y \in C_i} \text{proximity}(x, y)$	$\frac{1}{m_i}$	Graph-Based cohesion
$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i)$	1	Prototype-Based cohesion
$\text{separation}(C_i) = \text{proximity}(c_i, c)$	$m_i$	Prototype-Based separation

Potenzialmente, tutte le misure non supervisionata della validità di un cluster possono essere usate come una funzione obiettivo per un altro algoritmo di clustering e viceversa.

Ci siamo quindi concentrati sulla coesione e sulla separazione per la valutazione complessiva di un gruppo di cluster. Molte di queste misure di validità dei clusters, possono essere anche usate per valutare singoli clusters e oggetti (records)

Potremmo inoltre classificare i singoli cluster in base al loro valore di validità specifico del cluster (come ad esempio la loro coesione o separazione).

Un cluster che ha un alto valore di coesione, può essere considerato meglio di un cluster che ha un basso valore.

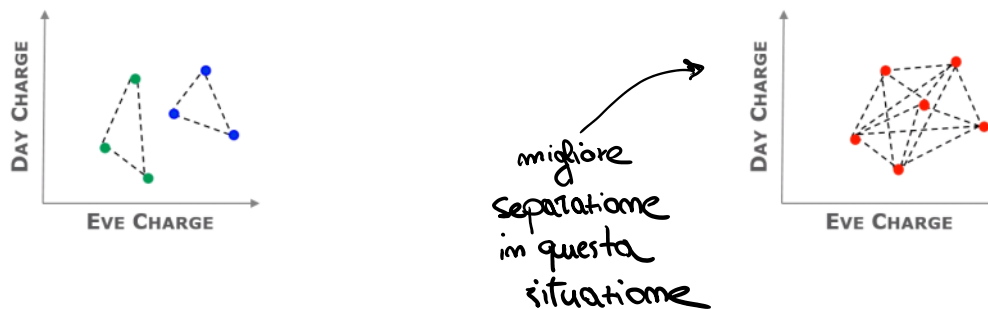
Questa informazione è utile per **Migliorare la qualità del processo di cluster analysis**

**Cluster not very cohesive** → **split into several clusters**



Un altro esempio può essere il seguente:

**Two clusters are relatively cohesive but not well separated** → **merge them into a single cluster**



È possibile inoltre cercare di misurare gli oggetti all'interno di un cluster per scoprire il **loro contributo** in termini di coesione e separazione all'interno del cluster

Objects that contribute more to the overall cohesion or separation of a cluster



near the interior of the cluster

Objects that contribute less to the overall cohesion or separation of a cluster



near the edge of the cluster

### Silhouette Coefficient:

Il coefficiente di Silhouette è una misura di valutazione che sfrutta il concetto di: **interno e bordo** di un cluster per valutare i punti, clusters e l'intero set di clusters.

Si applica principalmente al: **Partitional Clustering**

La misura di silhouette, combinata con coesione e separazione per un i-esimo oggetto (record) può essere rappresentata in questo modo:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, +1]$$

Dove:

- $A_i$  = distanza media dall'i-esimo oggetto rispetto a tutti gli altri oggetti all'interno del cluster
- $B_i$  = minimo della distanza media tra l'i-esimo oggetto, rispetto a tutti gli altri oggetti in ogni cluster differente rispetto al cluster a cui ci riferiamo (i-esimo cluster considerato)

**Valore negativo silhouette** = significa che la media della distanza dal punto nel suo cluster ( $a_i$ ) è maggiore rispetto alla minima distanza media tra tutti i punti in un altro cluster ( $b_i$ )

Quindi noi cerchiamo che il valore di silhouette sia **positivo** ( $a_i < b_i$ ) e per  $a_i$  che sia prossimo allo 0 il più possibile.

Il valore di Silhouette assume il massimo valore di 1 quando il valore di  $a_i$  è = 0

È possibile inoltre calcolare **la media del coefficiente di silhouette per un cluster** semplicemente prendendo la media del coefficiente di silhouette per tutti i punti (records) all'interno di un singolo cluster.

Una misura globale di efficacia e riuscita di un cluster = può essere quindi ottenuta calcolando: **la media del coefficiente di silhouette per tutti i punti**

### Cophenetic Correlation Coefficient



Viene utilizzato per calcolare la bontà e la valutazione dei clusters ottenuti. È simile al coefficiente di Silhouette, ma viene utilizzato per il clustering gerarchico

Si applica principalmente al: **Hierarchical Clustering**

Questo coefficiente misura il **grado di similarità** tra:

- Matrice di Prossimità P
- Matrice Cophenetic Q

Gli elementi della matrice Cophenetic Q sono: **i records che misura il livello di prossimità in cui le coppie di punti sono raggruppate nello stesso cluster la prima volta**

Anche il valore del Cophenetic coefficient assume valori in un range da  $[-1, 1]$ , quando l'indice di tale valore è più vicino all'1 indica una similarità significativa tra P e Q e una buona misura per le gerarchie rispetto ai dati.

Tuttavia, per il legame medio, anche per valori molto alti di Cophenetic Correlation Coefficient, non è assicurata una similarità sufficiente tra le due matrici.

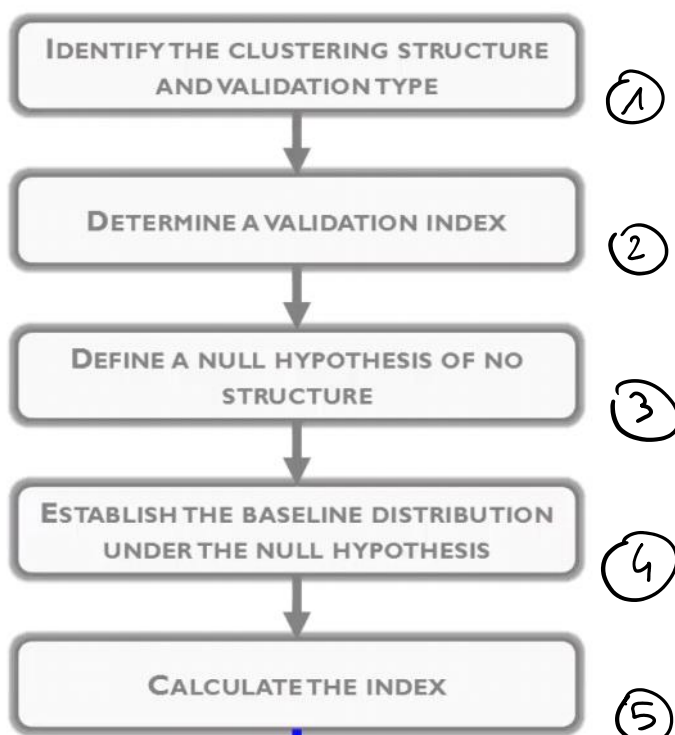
## Cluster Validity

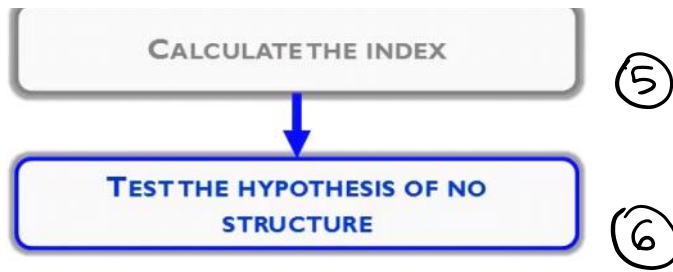
### Paradigma di validità:

I criteri esterni ed interni per il clustering sono molto strettamente legati ai metodi statistici e ai test di ipotesi.

**Validity Paradigm** = per una struttura di clustering, è basata sulla seguente ipotesi nulla:

- **There is no structure in the dataset**





1. Decidere se usare indici interni o esterni: in questo modo è possibile identificare la struttura degli algoritmi di clustering e la loro tipologia di validazione
2. Quale specifico indice di validazione dobbiamo utilizzare?
3. È importante definire delle null hypothesis:
  - a. **Random Position Hypothesis** = tutte le posizioni dei punti **m** in una determinata regione di uno spazio n-dimensionale sono ugualmente probabili *[Associata a Ratio Data]*
  - b. **Random Graph Hypothesis** = tutte le matrici **[m x m]** di prossimità hanno un rank ugualmente probabile *[Associata a prossimità di dati ordinali tra coppie di oggetti e dati]*
  - c. **Random Label Hypothesis** = tutte le permutazioni delle labels su **m** data points sono ugualmente probabili *[Possibile per tutti i tipi di dato]*
4. Stabilire una distribuzione (baseline) rispetto alle ipotesi nulle, utilizzando **Monte carlo analysis** e **bootstrapping**
5. Calcolare lo specifico indice rispetto al cluster utilizzato nella soluzione sviluppata
6. Convertiamo il valore dell'indice calcolato al valore della distribuzione della baseline calcolata nel passo 4, più specificatamente ad un preciso quartile della distribuzione empirica.  
Il quantile dipende dall'ipotesi di significatività del test che vogliamo eseguire

Solitamente quindi si spera di dover **rigettare l'ipotesi nulla**: *che non esiste una struttura definita nel cluster che abbiamo sviluppato.*

Se rigettiamo quindi questa ipotesi, possiamo dire che il cluster che abbiamo creato: è significativo

<http://www.turingfinance.com/clustering-countries-real-gdp-growth-part2/>

## Problema Fondamentale

Il problema fondamentale dell'analisi di clusterizzazione è: **determinare il numero reale di cluster k**

Per risolvere questo problema, ci sono diverse soluzioni.

La più semplice è utilizzare la proiezione dei dati in uno spazio euclideo in 2 o 3 dimensioni utilizzando tecniche di visualizzazione per scoprire insight e determinare il numero di clusters esistenti.

Questo tipo di analisi molto semplice non può essere impiegato su tutti i tipi di dati, ma solamente ad un numero ristretto di applicazioni

È possibile utilizzare alcuni indici visti in precedenza per cercare di risolvere questo problema.

Indici esterni e interni richiedono l'utilizzo e l'applicazione di test statistici che sono computazionalmente intensivi.

## Relative Criteria

Elimina questi requisiti di computazione, e si concentra sulla comparazione dei risultati di clustering generati da differenti algoritmi di clustering sullo stesso algoritmo con differenti input parameters

### Indici di misurazione del numero ottimale di clusters

Possiamo quindi avere degli indici che ci permettono di identificare il corretto numero di cluster per risolvere il problema fondamentale, alcuni di questi indici principali sono:

- **Calinski e Harabasz:**
  - o Il valore di K corrispondente al massimo è ottenuto per essere il numero ottimale di clusters
- **Dunn:**
  - o Il valore di k corrispondente al massimo è ottenuto per essere il numero ottimale di clusters (un valore molto elevato suggerisce la presenza di clusters compatti e ben separati)
- **David-Bouldin:**
  - o Il valore di K corrispondente al minimo è preso per essere l'ottimo numero di clusters

Per i **Probabilistic mixture model-based clustering** gli indici principali sono:

- **Akaike information criterion (AIC):**
  - o Il valore di K corrispondente al minimo è preso per essere il valore ottimale dei clusters
- **Minimum Description Length (MDL):**
  - o Il valore di k corrispondente al minimo è preso per essere il valore ottimale dei clusters
- **Bayesian Information Criterion (BIC):**
  - o Il valore di K corrispondente al massimo è preso per essere il valore ottimale di clusters

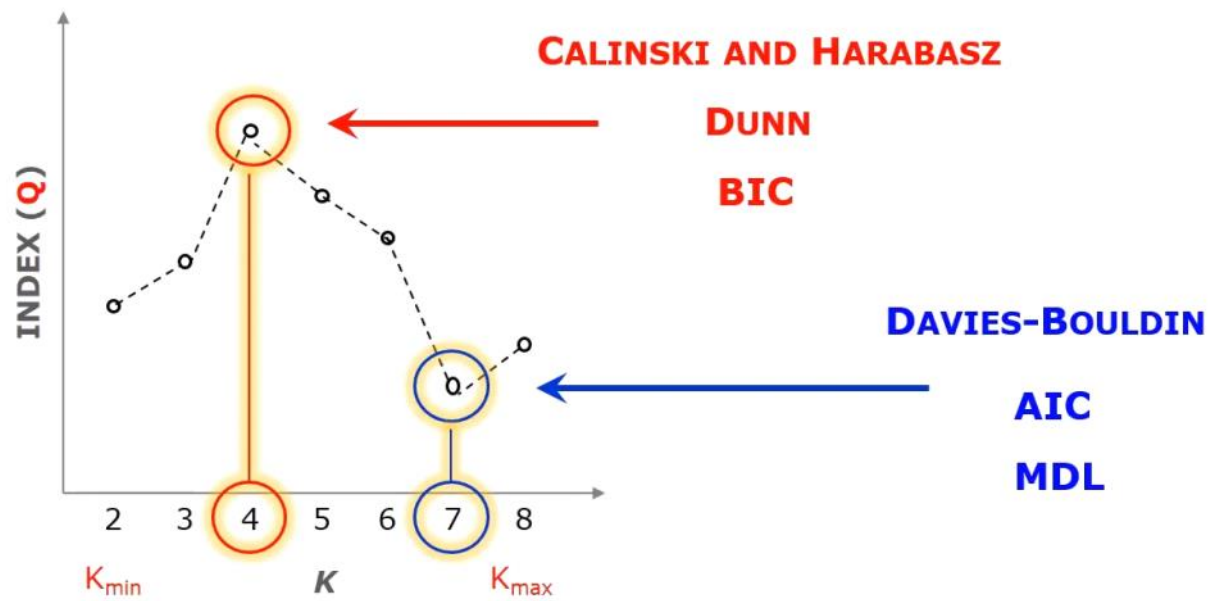
Il Problema Fondamentale della clusterizzazione si può risolvere con una procedura

Per un algoritmo di clustering che **richiede in input un numero di clusters K** da un utente, una sequenza di strutture di clustering si ottiene facendo un algoritmo di clustering diverse volte dove K ha un range che va da  $K_{\min}$  fino a  $K_{\max}$

#### GENERATE SEQUENCE OF CLUSTERING STRUCTURES

1. Choose a **clustering algorithm** and a **validity index**
2. **FOR**  $K=K_{\min}$  to  $K_{\max}$  **DO**
3.     **FOR**  $i=1$  to  $r$  **DO**
4.         run the **clustering algorithm with K** and use parameter values different from the previous running
5.         compute the value **q** of the **validity index** and set  $q(i) = q$
6.     **END FOR**
7.     Choose the **best value q\*** of the **validity index**  $\{q_1, \dots, q_r\}$
8.     set  $Q(K) = q^*$
9. **END FOR**

Le strutture di clusterizzazione sono quindi valutate basandosi sull'indice computato e "l'ottima soluzione di clusterizzazione" si determina scegliendo uno dei migliori valori dell'indice.



In caso si utilizzino delle strutture di clusterizzazione gerarchiche, gli indici saranno anche considerati come **stopping rules**, che ci permetteranno di scoprire qual è il **migliore livello migliore al fine di tagliare in dendrogramma**