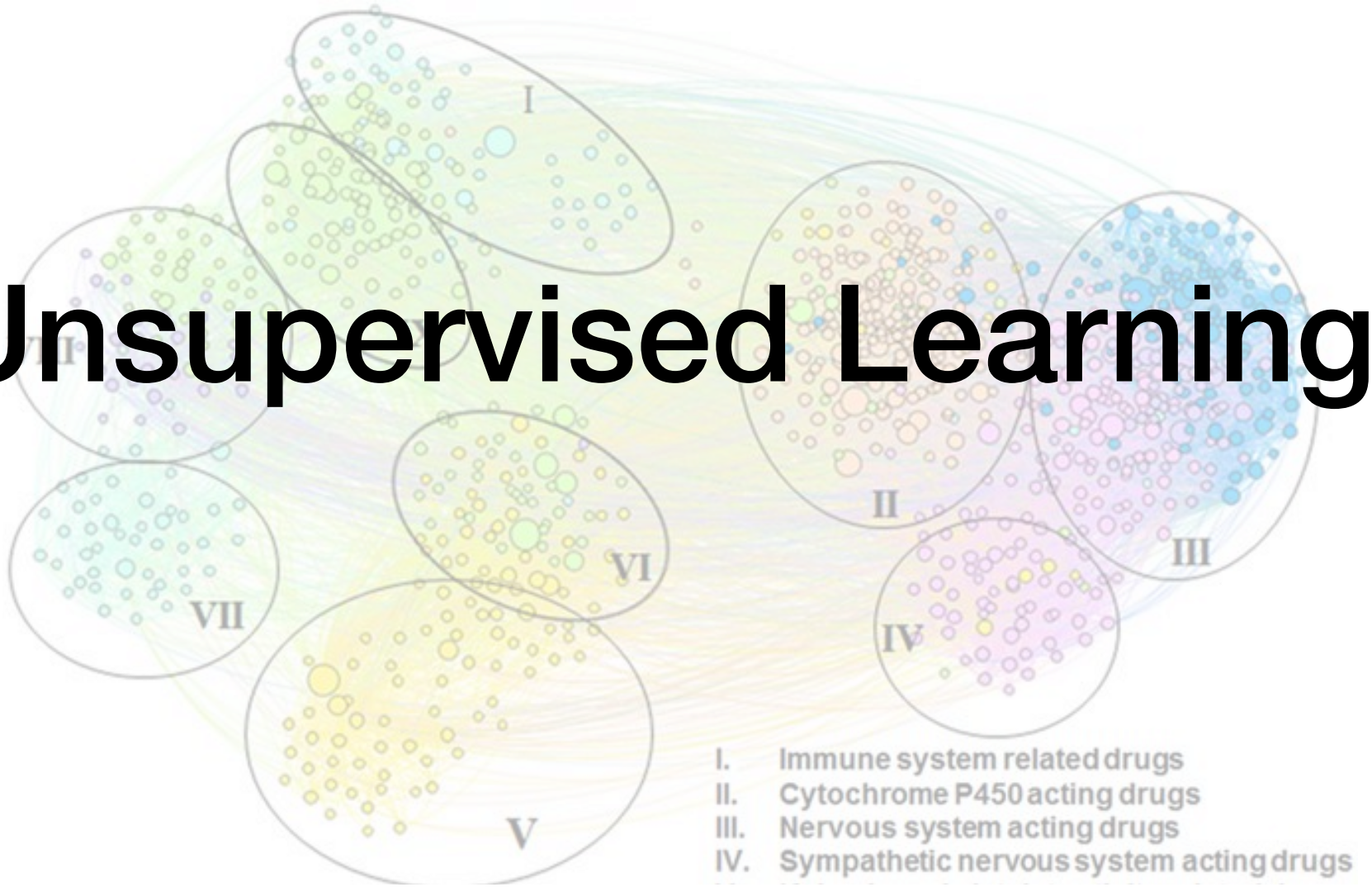
Unsupervised Learning



Dr. Fabien Plisson

Chemoinformatics in Drug Discovery

LANGEBIO, UGA CINVESTAV

October 15-18, 2019 - Irapuato, Mexico

Program

Introduction to Unsupervised Learning

Clustering methods

Dimension(ality) Reduction methods

Program

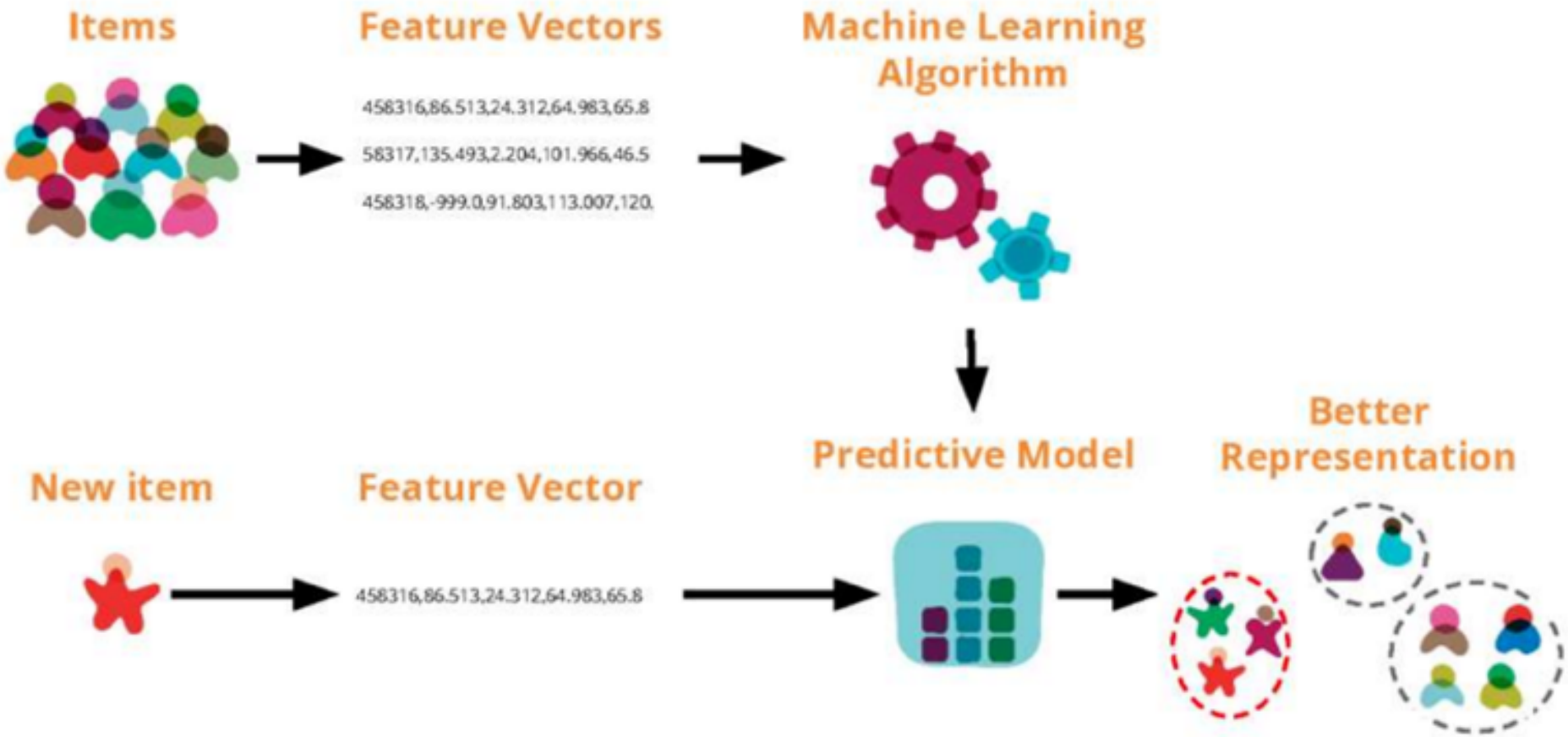
Introduction to Unsupervised Learning

Clustering methods

Dimension(ality) Reduction methods

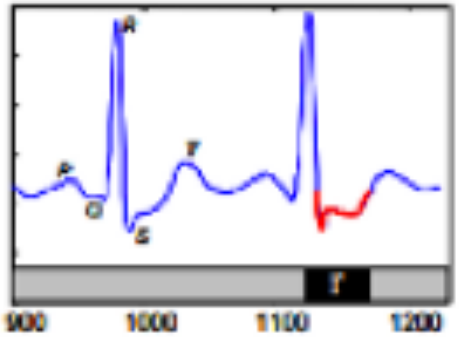
What is Unsupervised Learning

Discover the underlying patterns (or structure) of the data



Anomaly Detection

This makes the source of the anomaly apparent. Note that in the four normal ST waves, after the brief descending section, the signal rises monotonically. However, the anomalous ST wave has an additional local peak caused by a premature beat, thus justifying the cardiologist's diagnosis of premature ventricular contraction.



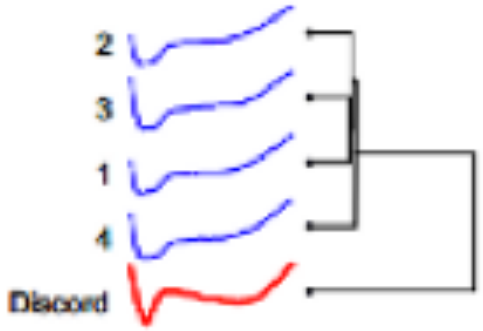
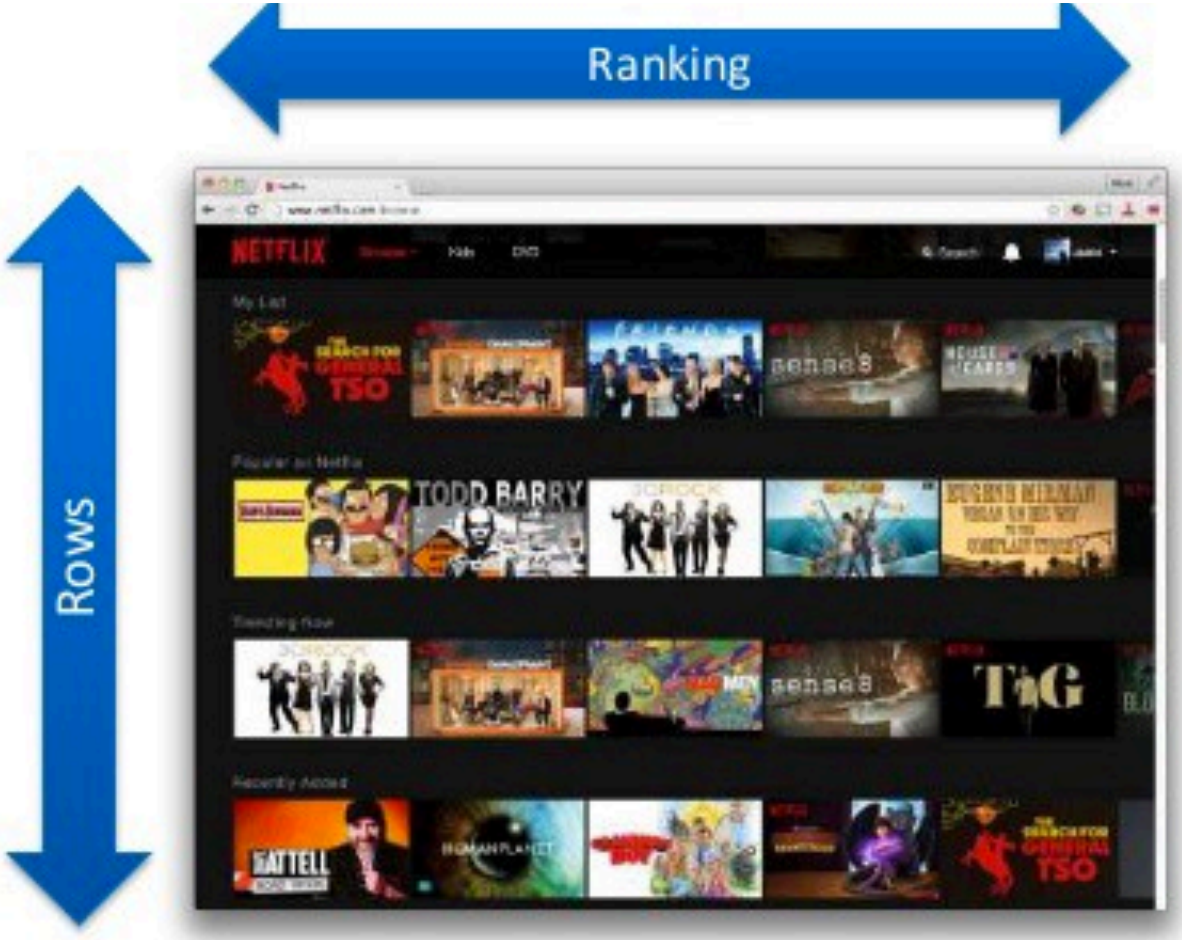


Figure 14: (*left*) A zoom-in of a section of Figure 13. The first heartbeat has been annotated with the classic notation. (*right*) Five ST waves from Figure 13 (including the discord) hierarchically clustered

Recommendation Systems



Features/Variables

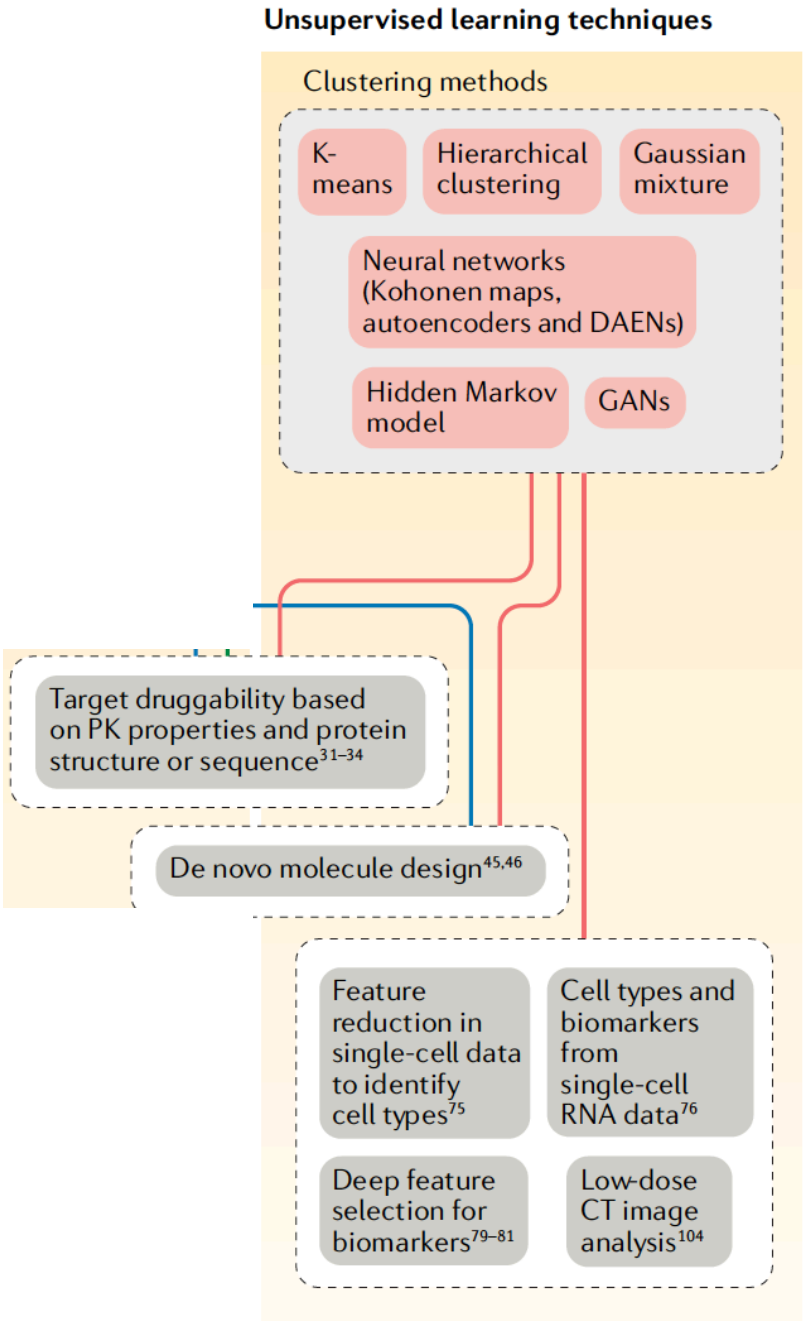
Tracking movies choices (before/now/after)
Tags e.g. sci-fi, 1h39min, Ryan Reynolds…
Movies Images
Family and friends

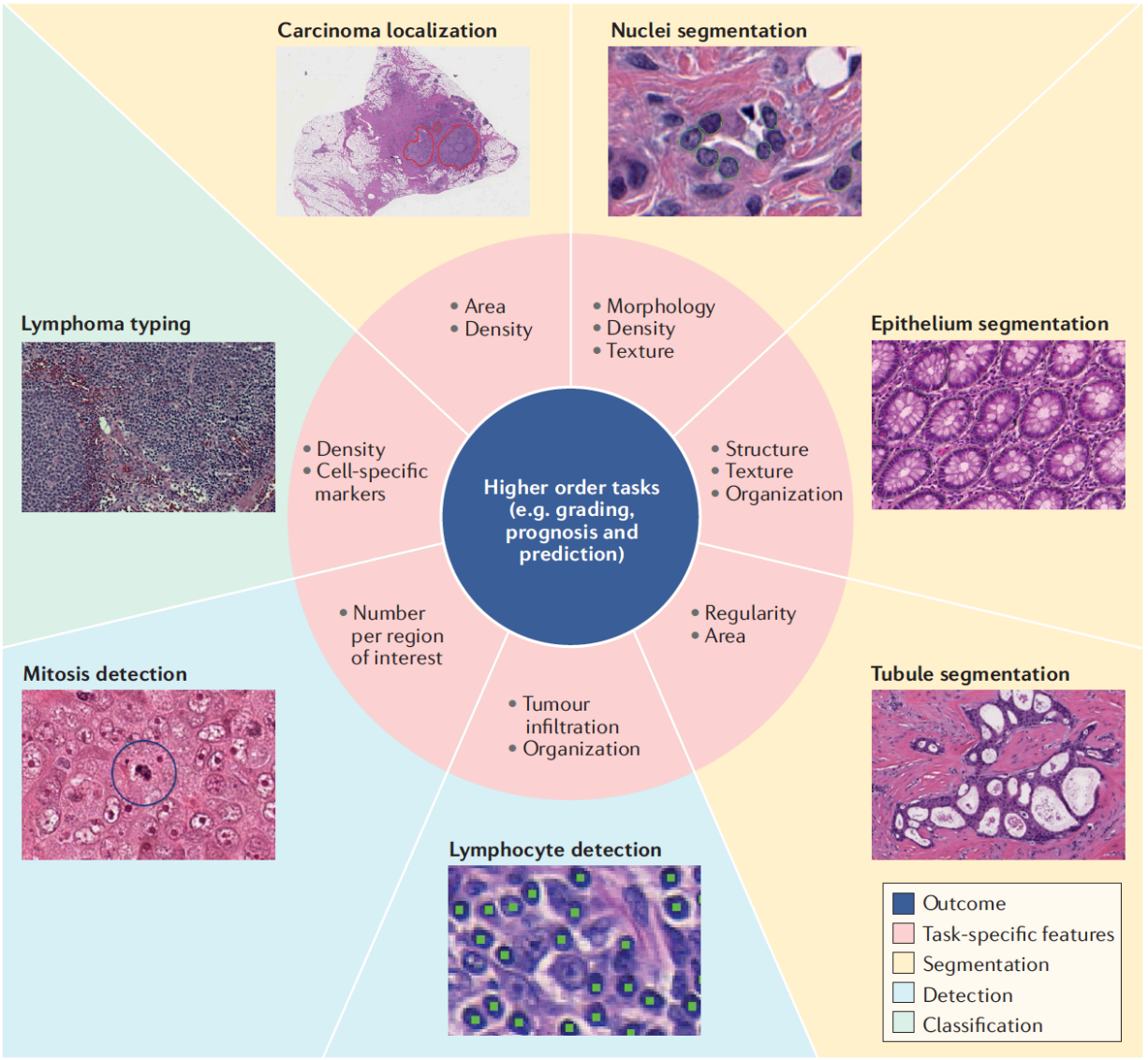
> NETFLIX Taste communities

Observations

100+ millions subscribers
A/B test products with 300,000 users

Cell segmentation





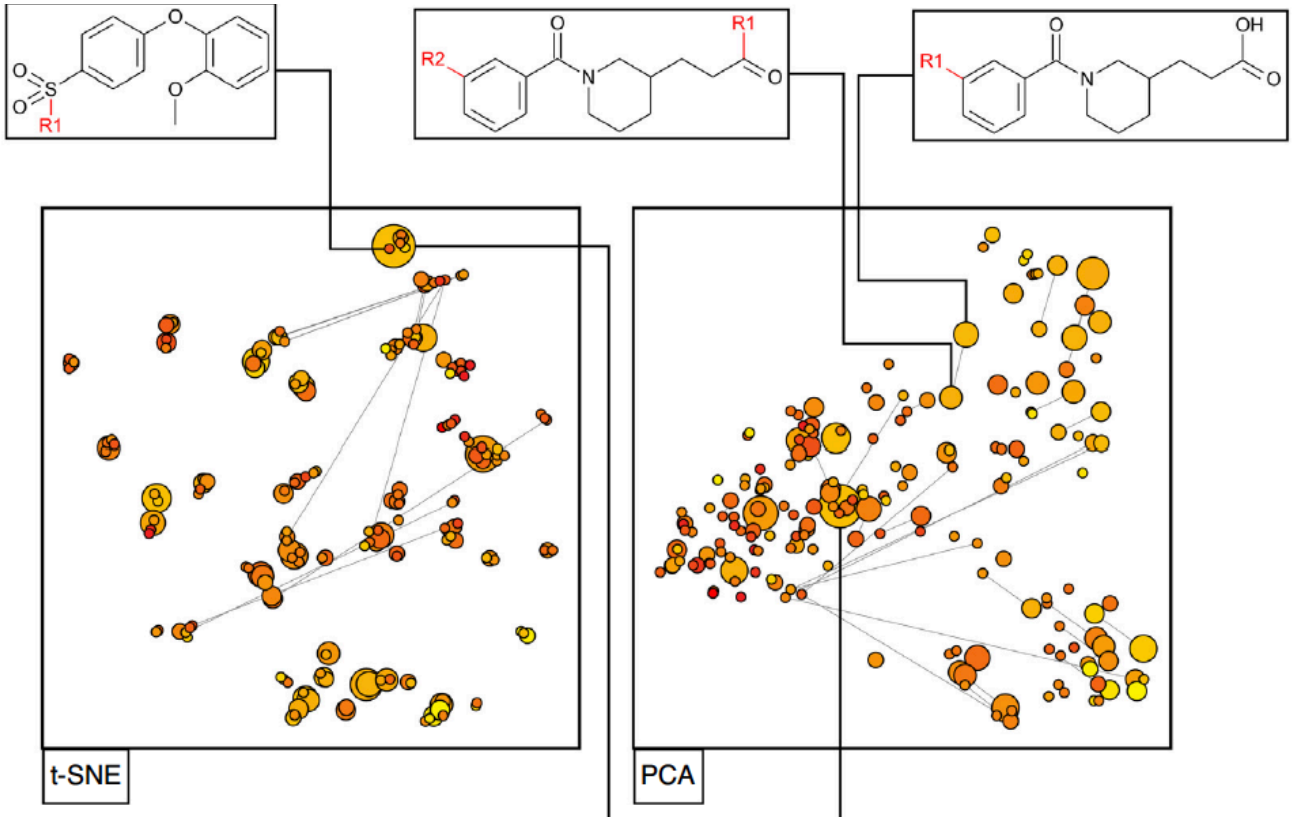
Vamathevan et al. Nature Reviews Drug Discovery (2019)

Organising Structure-Activity Data (Chemical Space)

Reaching for the bright StARs in chemical space

José L. Medina-Franco[1], J. Jesús Naveja[1,2] and Edgar López-López[1]

[1] Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico
[2] PECEM, School of Medicine, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico



Medina-Franco et al. Drug Discovery (2019) - Hot off the press

Two main families of algorithms

Clustering

Centroid-based clustering – **K-means Clustering**

Connectivity-based clustering – **Hierarchical Clustering**

Distribution-based clustering – **Expectation Maximization (EM)**

Density-based clustering – **Spatial Clustering Approach with Noise (DBSCAN)**

Dimension Reduction

Principal Component Analysis (PCA) Variants

Linear & Generalized Discriminant Analysis (LDA, GDA)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

https://lvdmaaten.github.io/tsne/

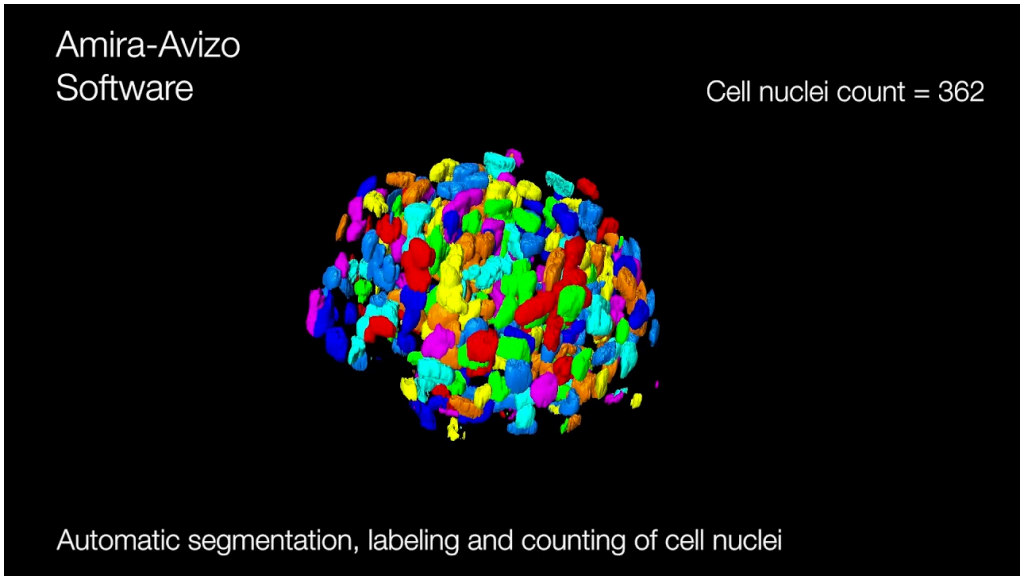
Factor Analysis

Program

Introduction to Unsupervised Learning

Clustering methods

Dimension(ality) Reduction methods



Monothetic, Polythetic, Hard or Soft Clustering

Monothetic Cluster = Data sub-populations have common property

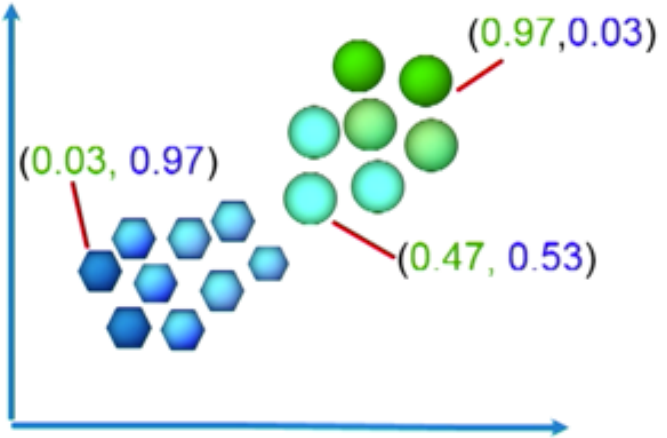
Polythetic Cluster = Data sub-populations have similar property (distance = membership)

Hard clustering

Each observation belongs to exactly one cluster

Soft clustering

An observation can belong to more than one cluster to a certain degree (e.g. likelihood of belonging to the cluster)

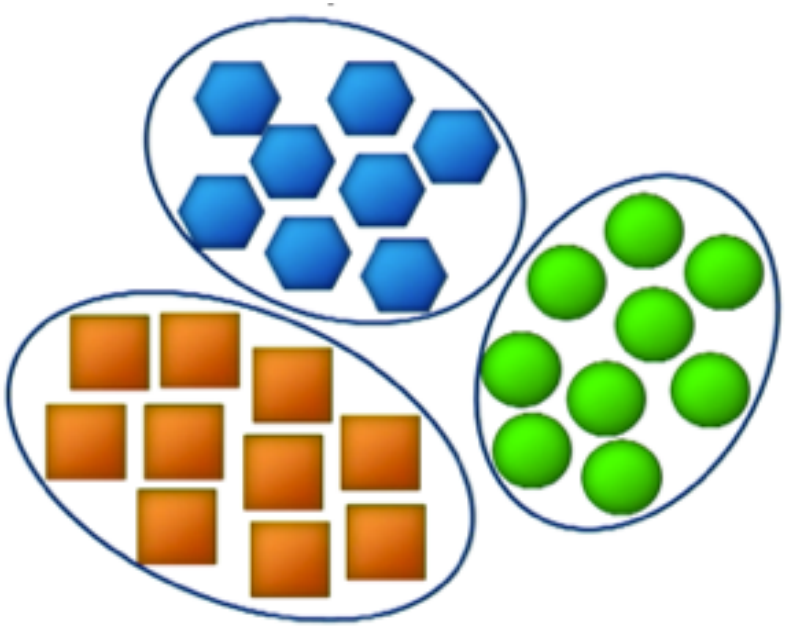


Partition or Hierarchical Clustering

Partitioning (Flat) Clustering

Construct partitions and evaluate them by some criterion.

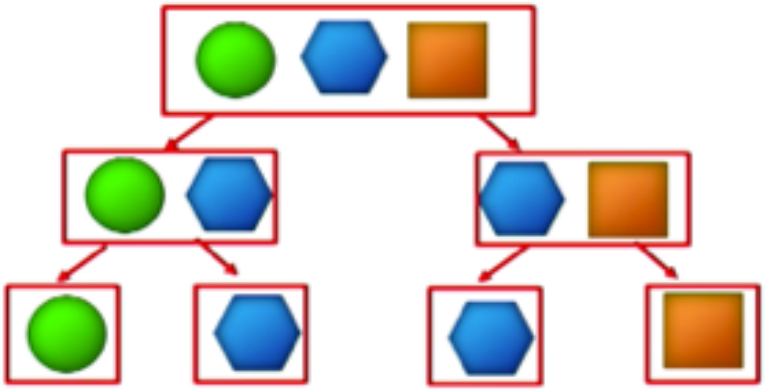
e.g. K-means / distance



Hierarchical Clustering

Decompose the set of data (into objects) using predefined rules.

e.g. Population / gender



How similar two observations are from one another?

Euclidean distance (most popular, easiest to compute)

Manhattan distance

Cosine distance

Jaccard distance

Minkowski distance

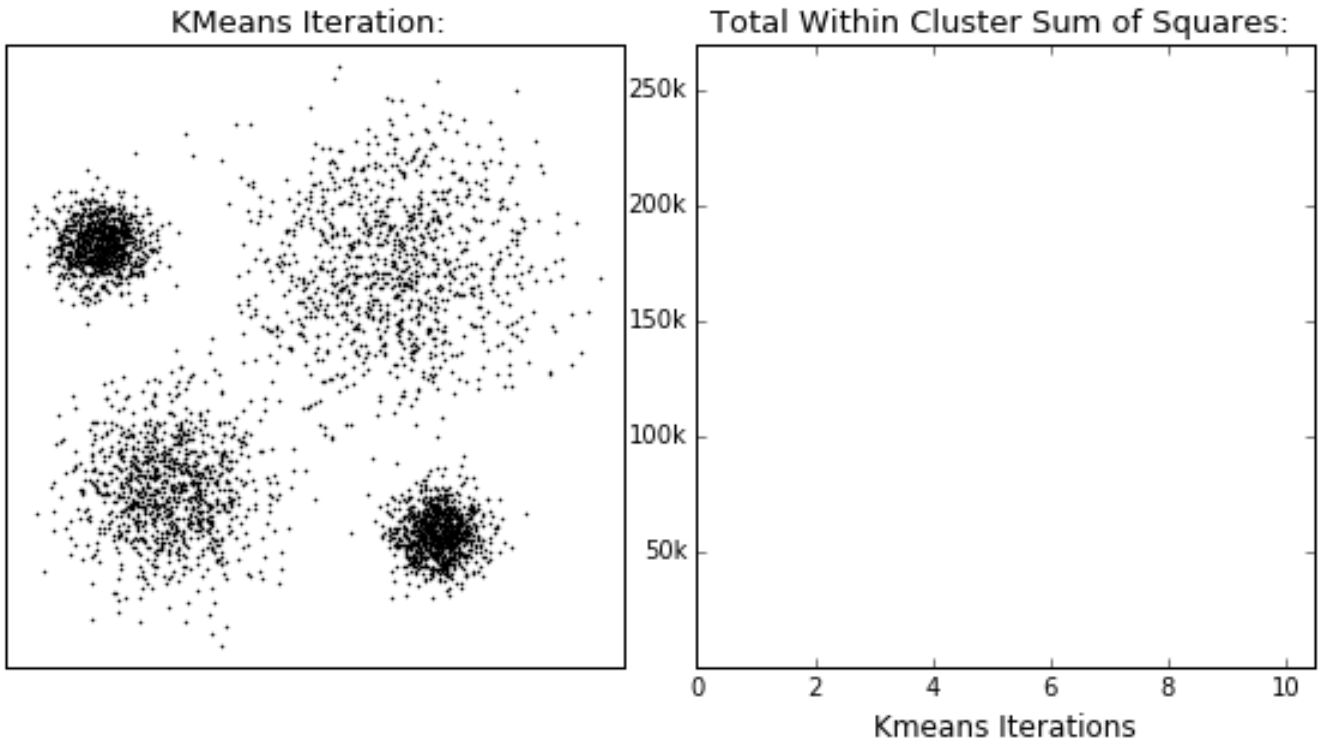
Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

K-Means Clustering



K-Means / K-Medias Clustering

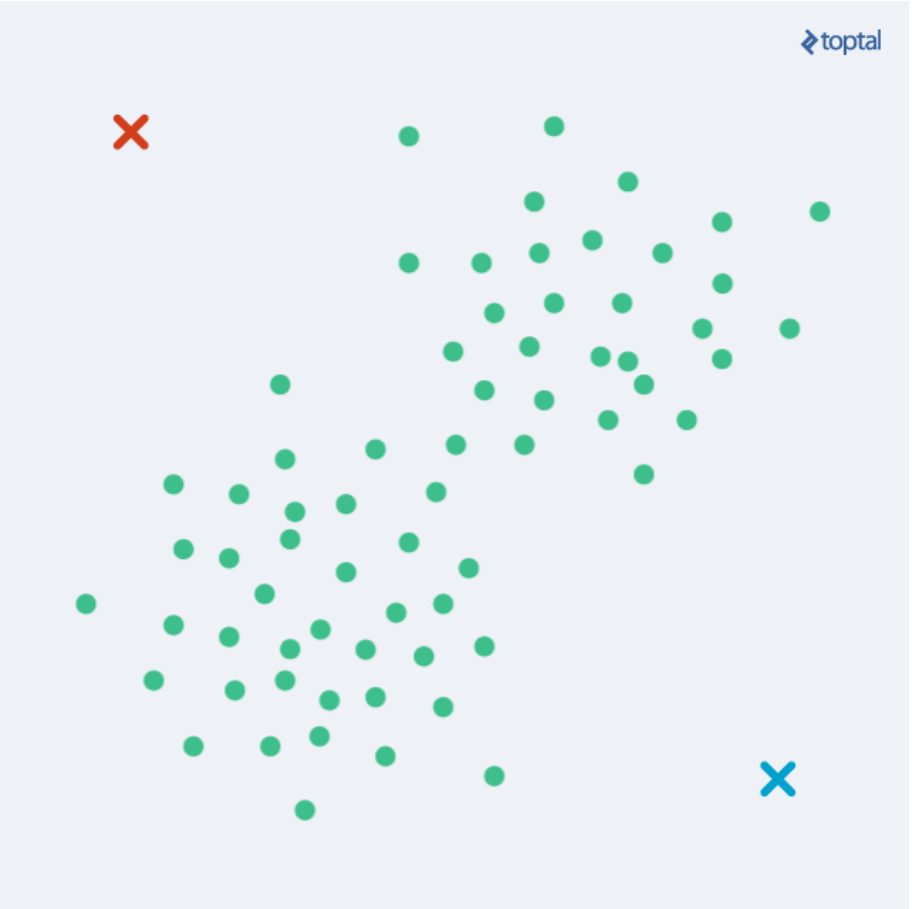
Polythetic, Hard, Partitioning method
> Splits the data into a defined number of clusters

1. Data is splitted into K subset
K is specified in advance

2. Create K points ‘centroids’ =
prototypical attribute in a subset

3. Measure the similarity (minimum
distance) between individual points
and centroids

4. Assign individual points with K
value of nearest centroid



Algorithm

1. Partition data into K subsets

2. Compute seed points as centroids c_1 … c_K
at random locations, means of the K clusters

3. Repeat until convergence:
   - for each individual point x_i :
     find nearest centroid c_J
     assign the point x_i to cluster J
   - for each cluster J = 1, … K:
       new centroid c_J = mean of all points x_i
       assigned to cluster J in previous step

4. Stop when none of the cluster assignments change

Assumption, Pros and Cons

Balanced clusters (~ size, density) within dataset
Each cluster is spherical *(independent features, features of equal variance)*

Fast computational cost O(n)
Easy to implement, easy to interpret
Scalability - work on large datasets

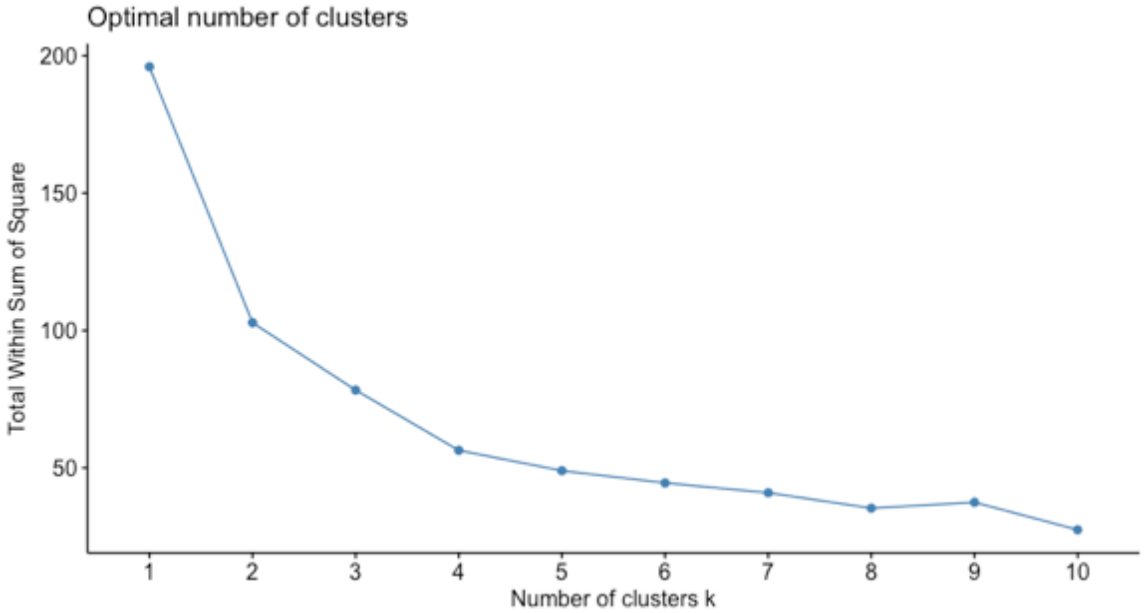
Requires prior knowledge of K
Repeatability – random choice of K, lack consistency
Not recommended for small datasets (<100 observations)
Sensitive to outliers

Tuning parameters

- Intra-cluster **distance** metric e.g Euclidean, Cosine, Manhattan
- Number of **k** clusters
- Number of **iterations**

- Searching k using

the Elbow method
the Silhouette Width
the Gap statistic



Intrinsic vs. Extrinsic K-means

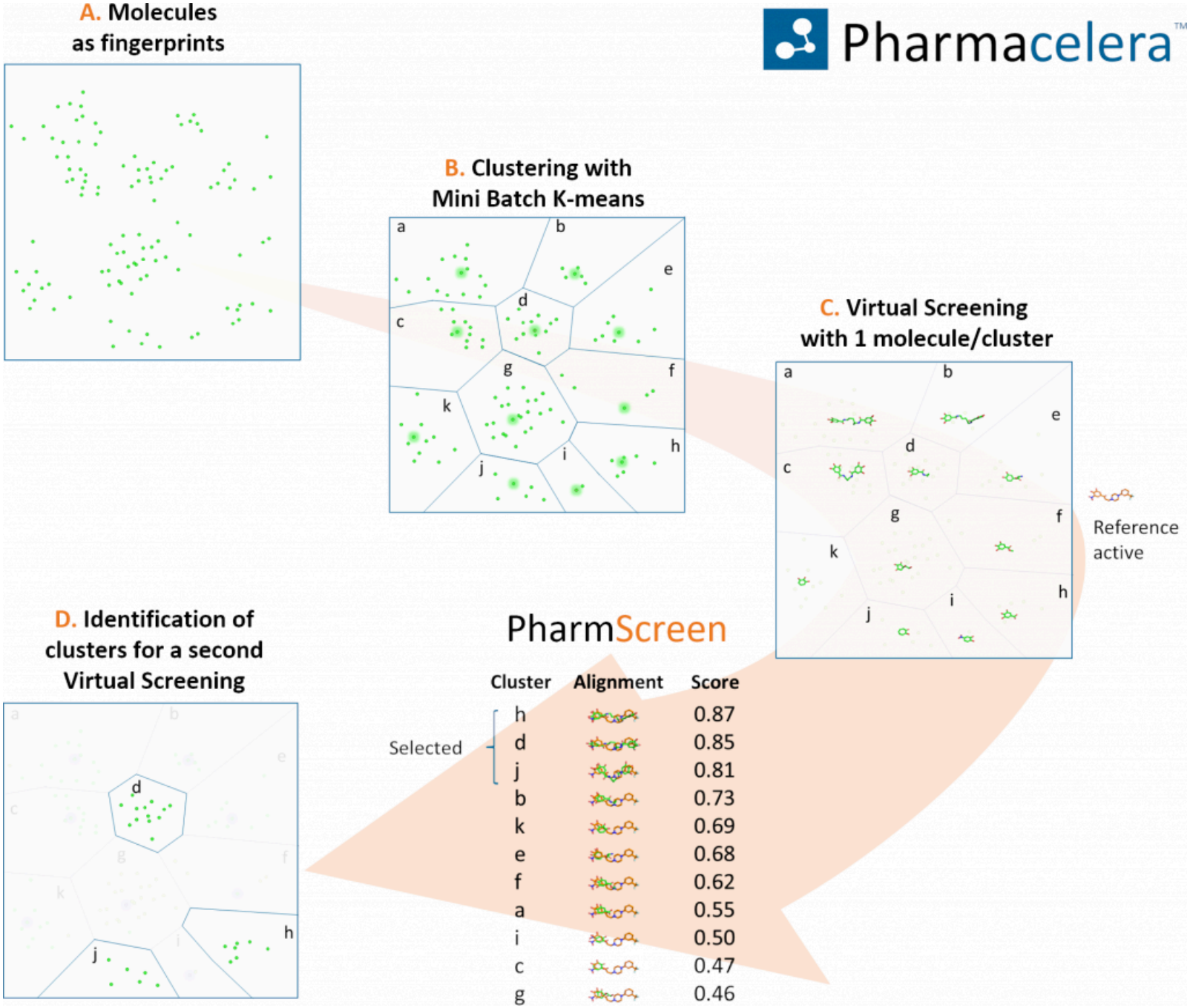
Intrinsic Cluster

- Helps understand data patterns (qualitative)
- Clusters = classes (similar to a classifier)

Extrinsic Cluster – *to solve another problem*

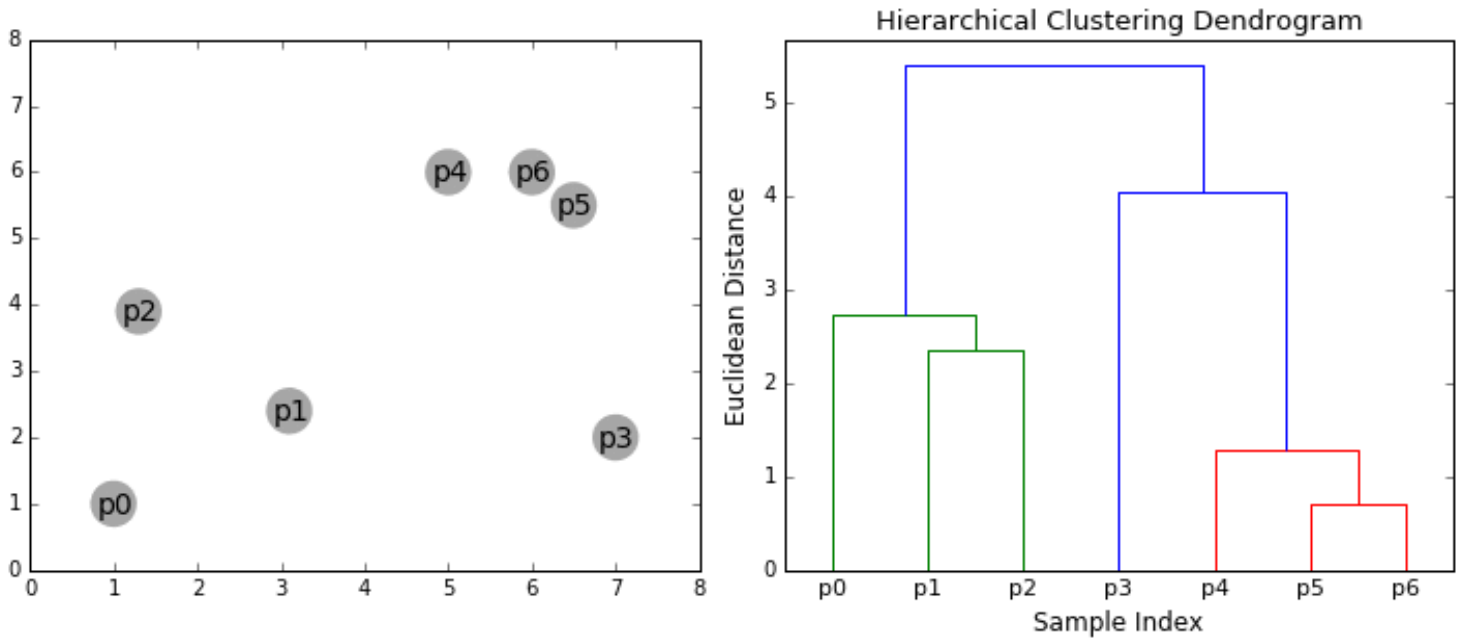
- Represents images with cluster features e.g. Image classification
- Train different classifiers for each data sub-population
- Identify and eliminate outliers

Mini Batch K-means



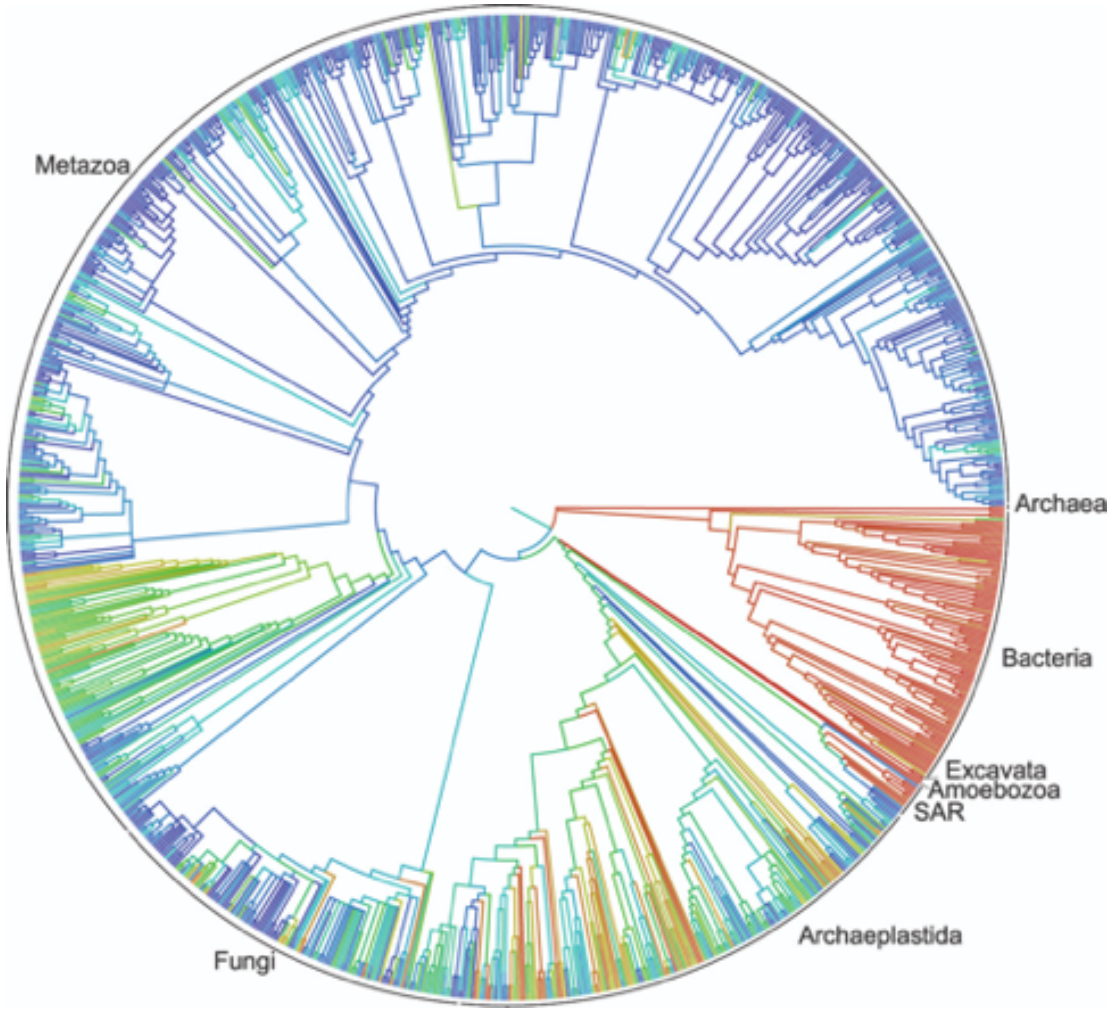
https://new.pharmacelera.com/science/clustering-methods-big-library-screening/

Hierarchical Clustering



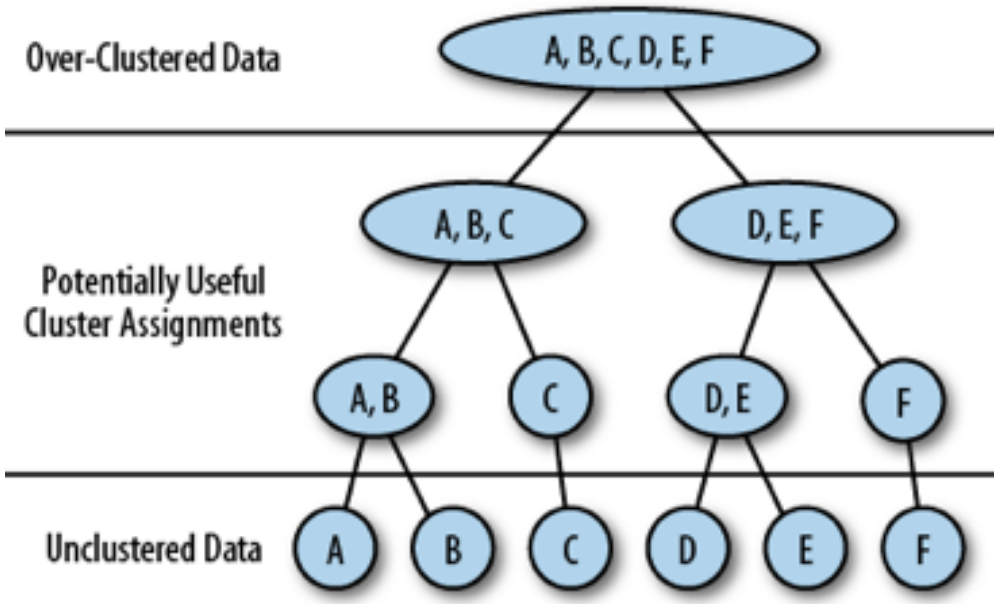
Hierarchical Clustering

Monothetic, Hard, Hierarchical method
> Splits or combines the data into a given number of clusters



How does it work?

Monothetic, Hard, Hierarchical method
> Splits or combines the data into a given number of clusters



Build tree-like clusters **dendrograms** that either combine or split clusters (of observations) using **distance similarity measure**

The classification consists of a series of partitions, which may run from a single cluster containing all individuals to n clusters each containing a single individual

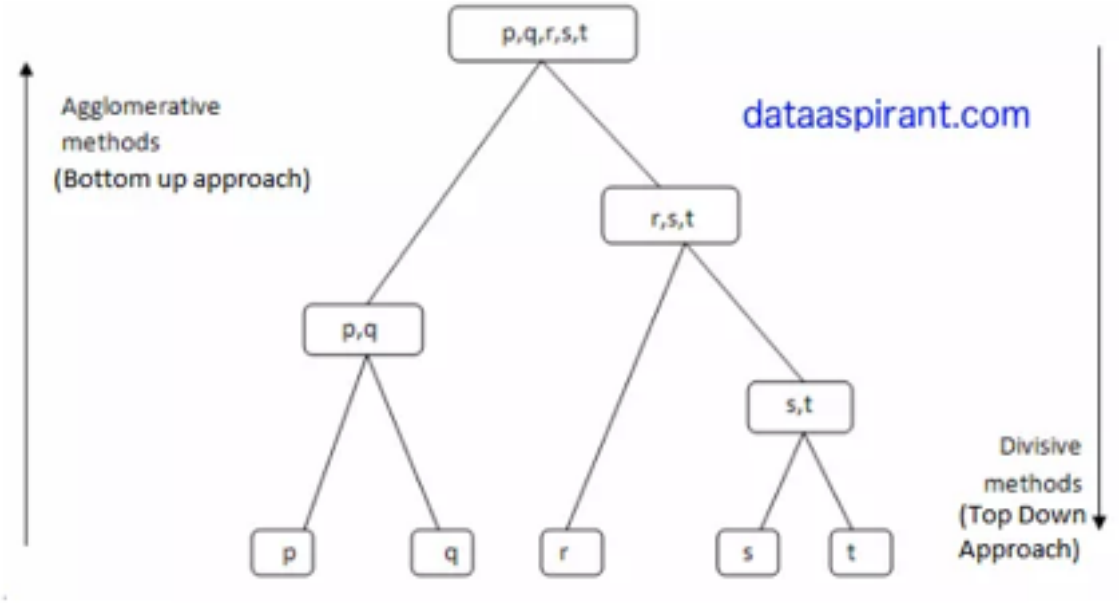
Two Types

Agglomerative Nesting (AGNES)

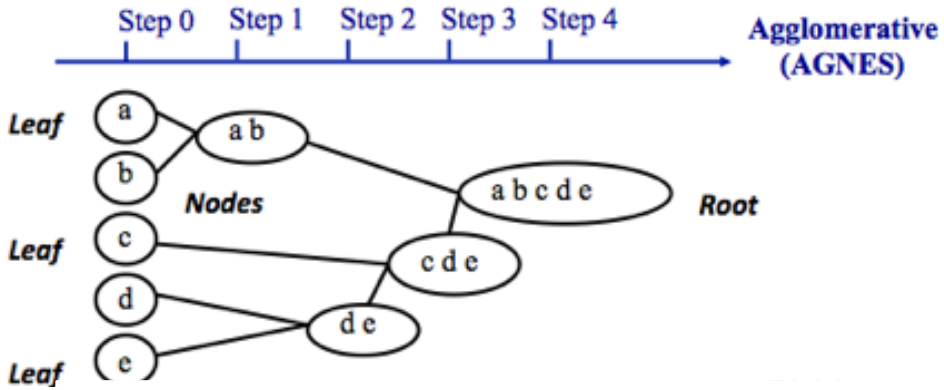
Bottom-up approach: Each observation (leaf) is considered as one cluster. Two 'similar' leaves (shortest distance) are joint into one cluster (node)… until a single cluster remains.

Divisive Analysis (DIANA)

Top-down approach: Starts from root (all observations are in one cluster), the most heterogenous cluster (highest variance/distance) is divided into 2…



Algorithm AGNES (HAC)



1.Assign each point to its own cluster
N elements = *N* clusters

2. Compare elements ~ distance matrix

3. Find the closest (most similar) pair of clusters and merge them into one cluster using **linkage method**

4. Compute distances (similarities) between the new cluster and each of the old clusters.

5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size *N*.

Linkage methods, how dissimilar two clusters are?

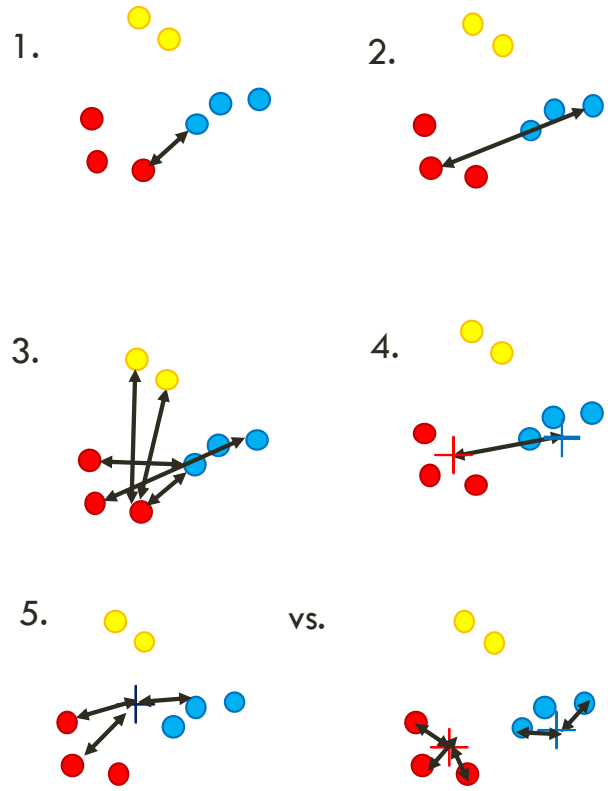
Minimum or single linkage (closest elements)

Maximum or complete linkage (farthest elements)

Mean or average linkage (all pairwise distances)

Centroid linkage (distance between centroids)

Ward's minimum variance (consider 2 clusters, how does the total distance from centroids change?)



Single Linkage	$D_{12} = \min_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$	This is the distance between the closest members of the two clusters.
Complete Linkage	$D_{12} = \max_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$	This is the distance between the members that are farthest apart (most dissimilar)
Average Linkage	$D_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(\boldsymbol{Xi}, \boldsymbol{Yj})$	This method involves looking at the distances between all pairs and averages all of these distances. This is also called UPGMA - Unweighted Pair Group Mean Averaging.
Centroid Method	$D_{12} = d(\underline{x}, \underline{y})$	This involves finding the mean vector location for each of the clusters and taking the distance between these two centroids.
Ward's Method	$D_{12} = \sqrt{\frac{2.\lvert k\rvert.\lvert l\rvert}{\lvert k\rvert + \lvert l\rvert}} \cdot \lVert \bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} \rVert$	This method minimizes the total within-cluster variance. Those clusters are combined whose merger results in minimum information loss (ESS criterion).

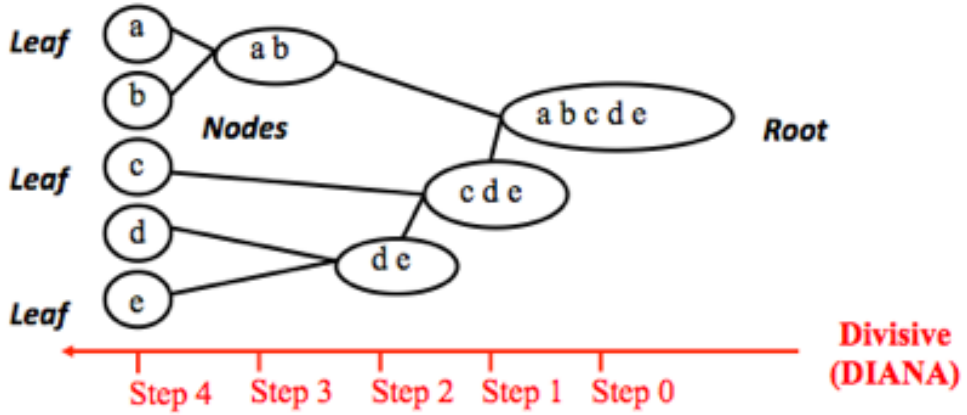
Algorithm DIANA

Opposite to AGNES algorithm

1. Start with a single cluster of all observations (root)

2. Split data points from two clusters

3. Stop when each point is in its own cluster



Pros and Cons

Does not require to know K, number of clusters
Attractive representation – Dendrogram
AGNES – small clusters vs. DIANA – large clusters

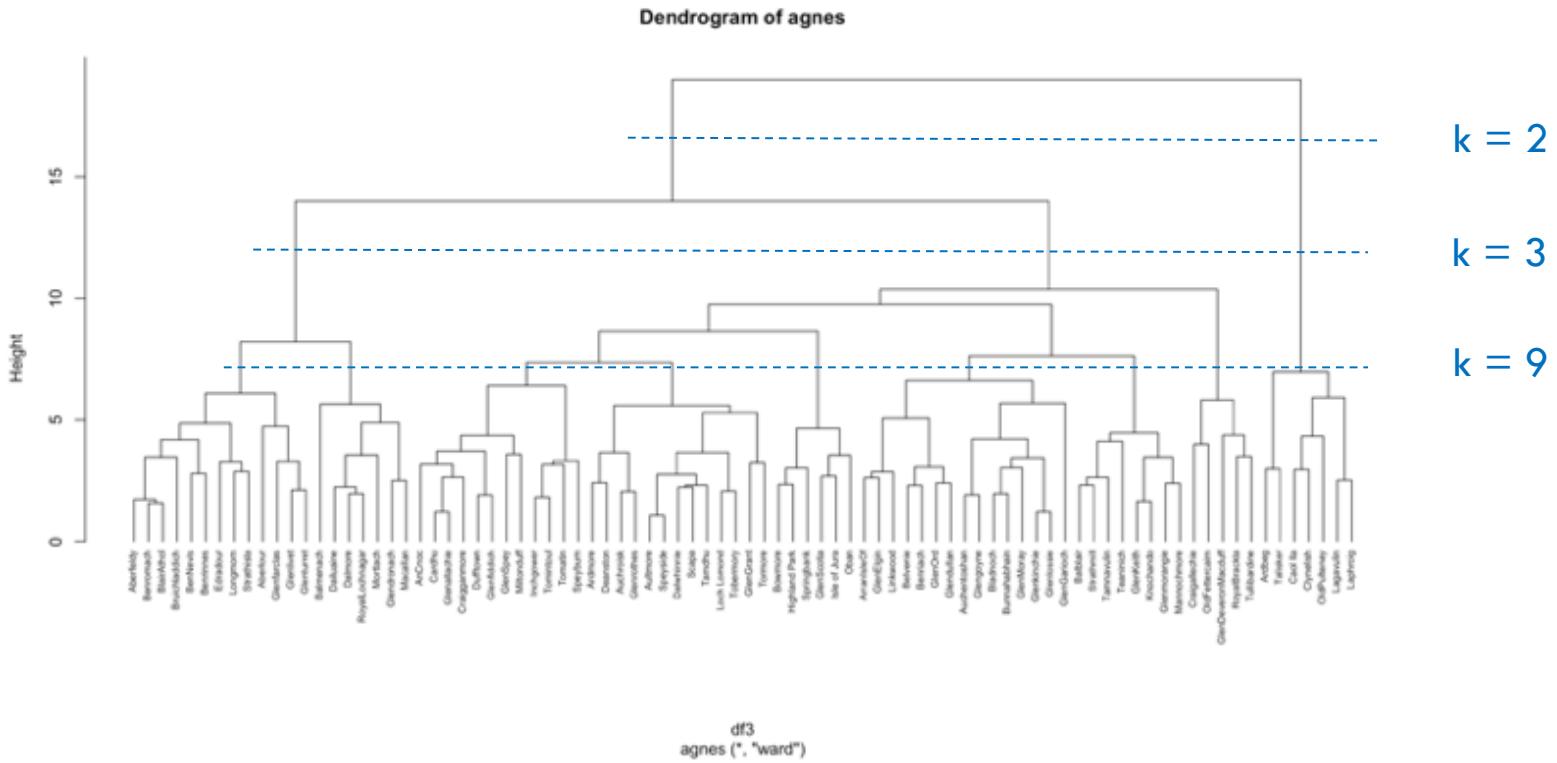
'Greedy' computational cost $O(n^2)$: *finding the best distance at each step*
Preference to smaller datasets

Tuning parameters

Intra-cluster distance metric
e.g. Euclidean, Manhattan, Cosine, Minkowski…

Inter-cluster distance metric or linkage method
e.g. Single, Complete, Average or Ward's method

Comparison using agglomerative (or divisive) coefficient



Molecular Property eXplorer (MPX)

agglomerative hierarchical clustering

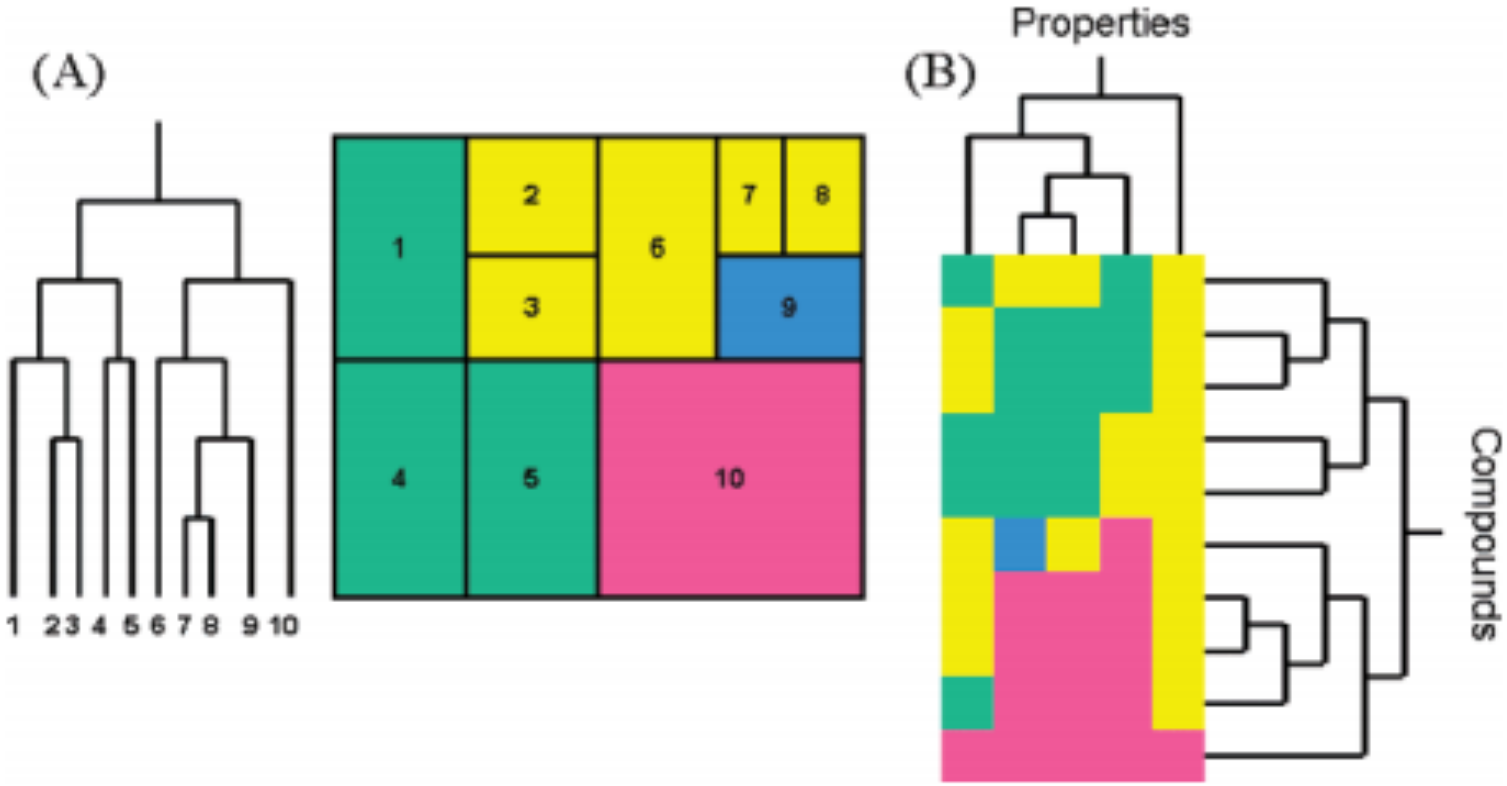
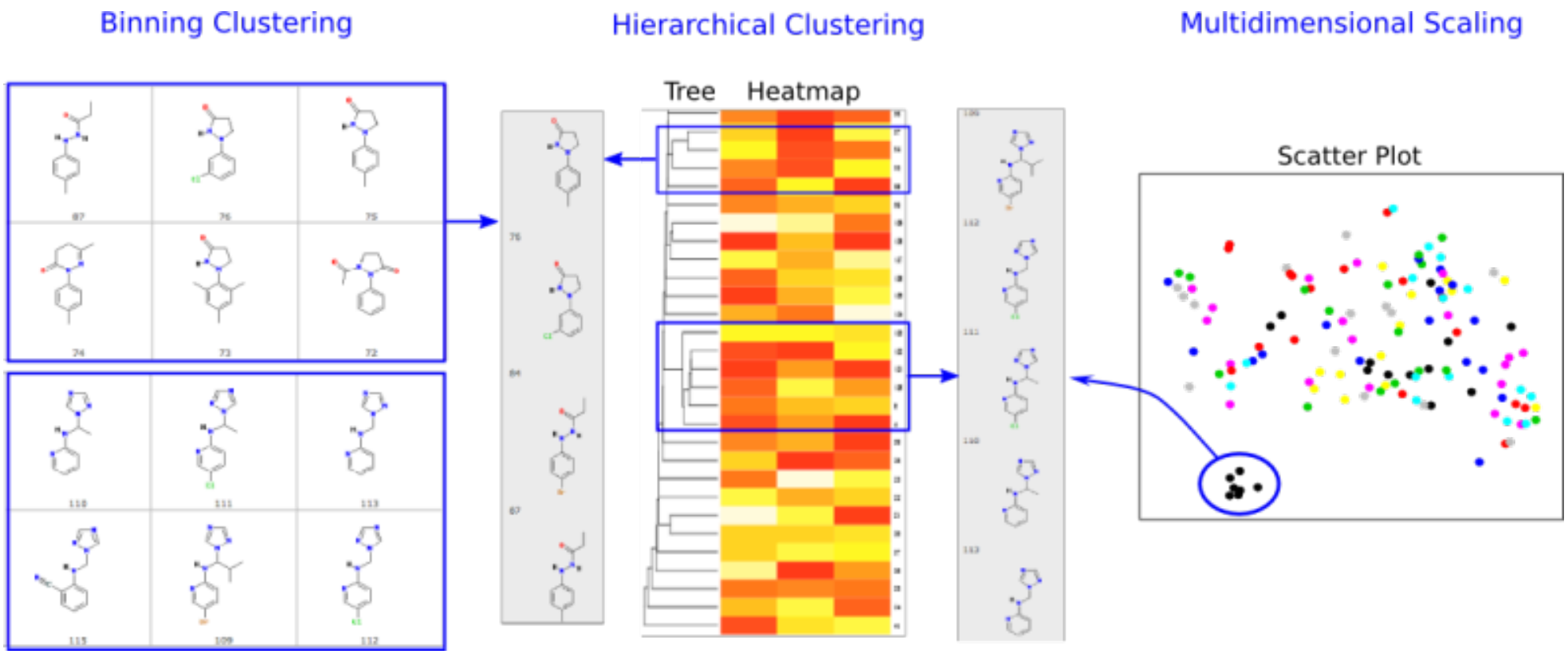


Figure 1. Representation of a hierarchical cluster of 10 compounds as (A) a tree-map and (B) a heatmap.

Kibbey et al. J. Chem. Inf. Model. (2005), 45, 523-532

ChemMine Tools



http://chemminetools.ucr.edu

References

1. Clustering algorithms Data Science for Business
Foster Provost & Tom Fawcett (2013) O'REILLY

2. Chapter 17 – Distances and Similarities in Data Analysis
Dictionary of distances (ISBN: 978-0-444-52087) by E. Deza & M.-M. Deza

3. Applied Predictive Modeling (2013) Kuhn, Max, Johnson, Kjell

4. David Sheehan https://dashee87.github.io

https://github.com/BarbaraDiazE/CABANA_CHEMOINFORMATICS