UNIVERSITAS INDONESIA

Veritas, Probitas, Iustitia

pusilkom ui

quality results

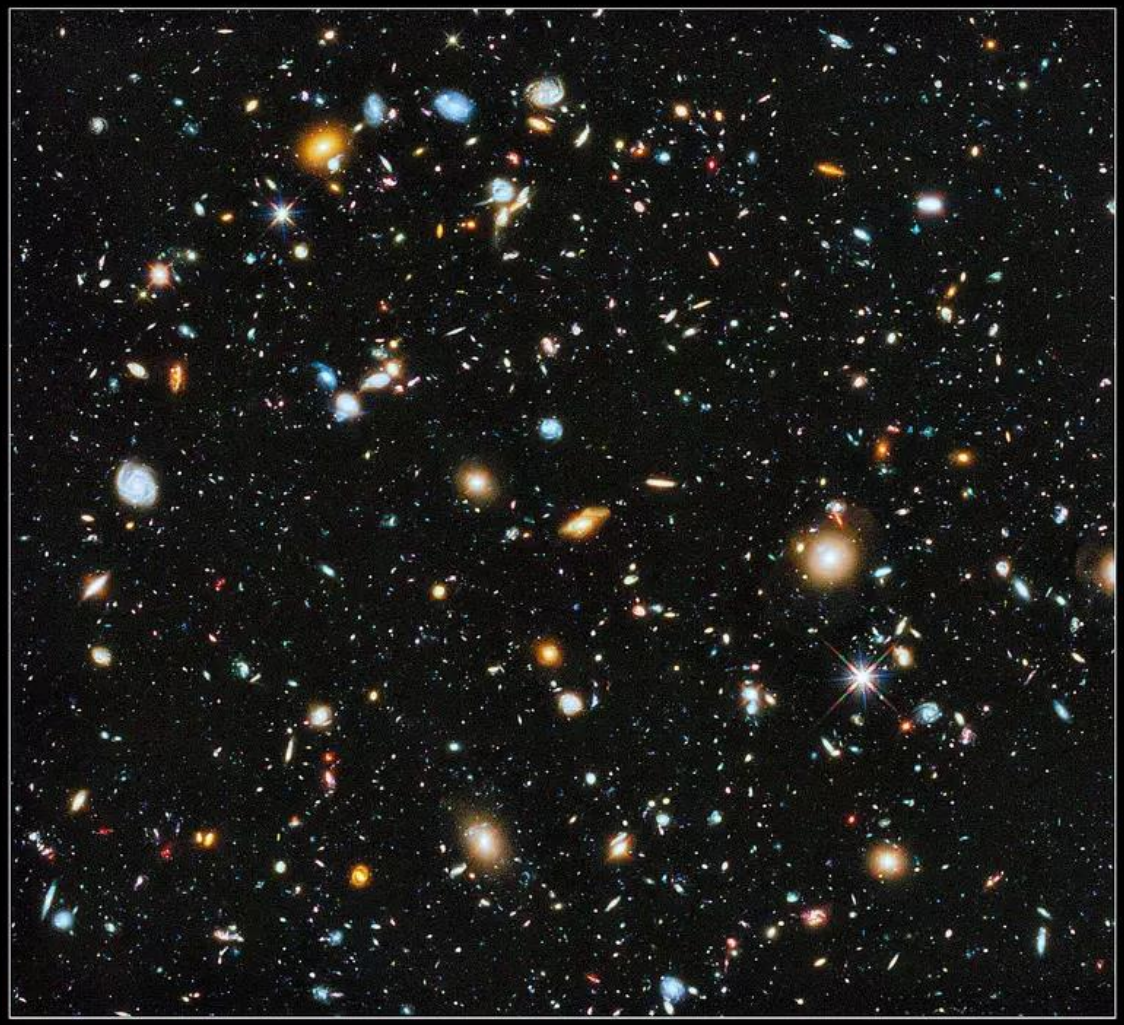
Cluster Analysis

Denny, Ph.D.

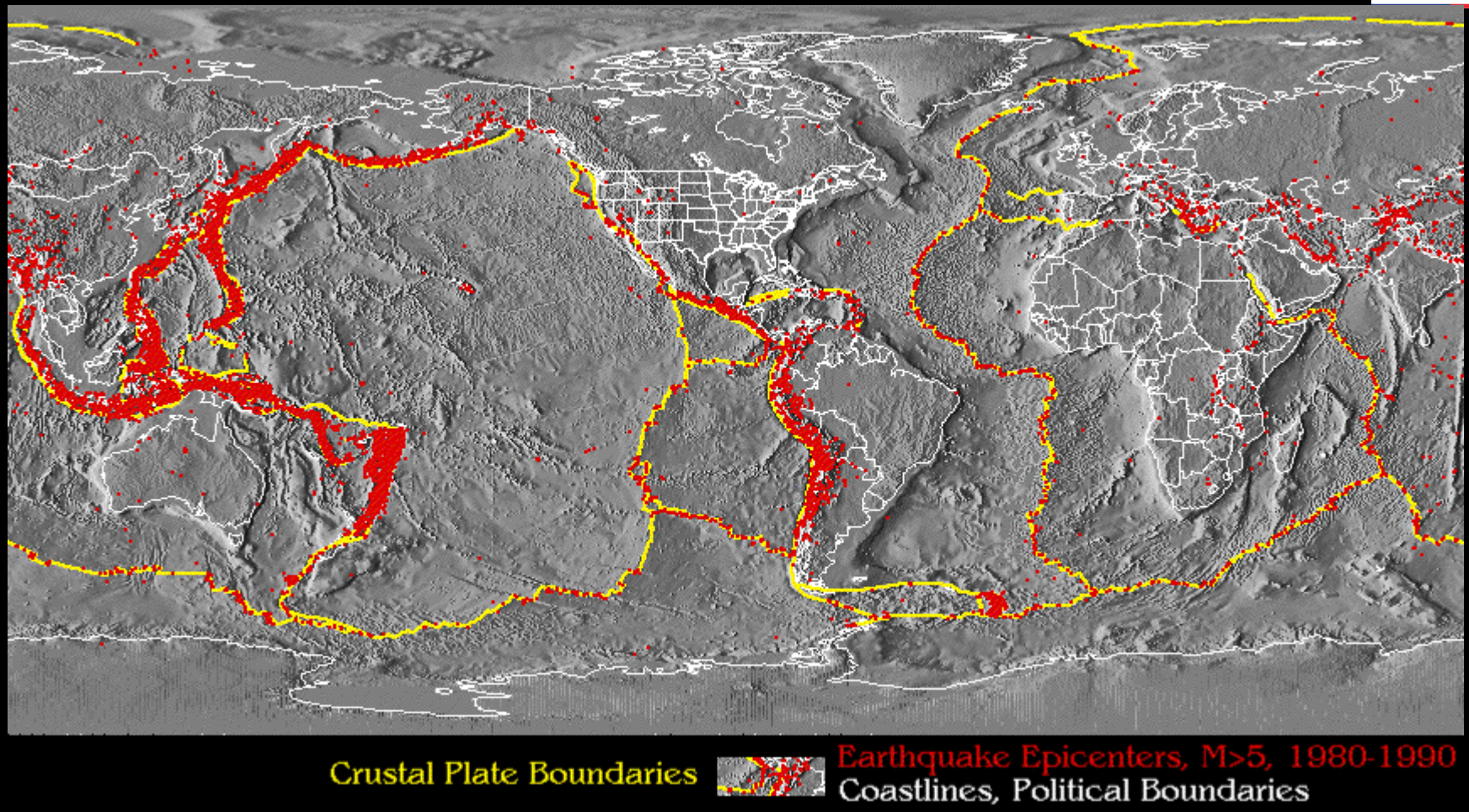
Outlines

- **What is Clustering**
- **Clustering Applications**
- **Major Clustering Approaches**
  - Partitioning Methods (i.e., K-Means Clustering)
  - Hierarchical Methods (i.e., Agglomerative Nesting)
- **Data Types and Normalization**

Clusters in Nature

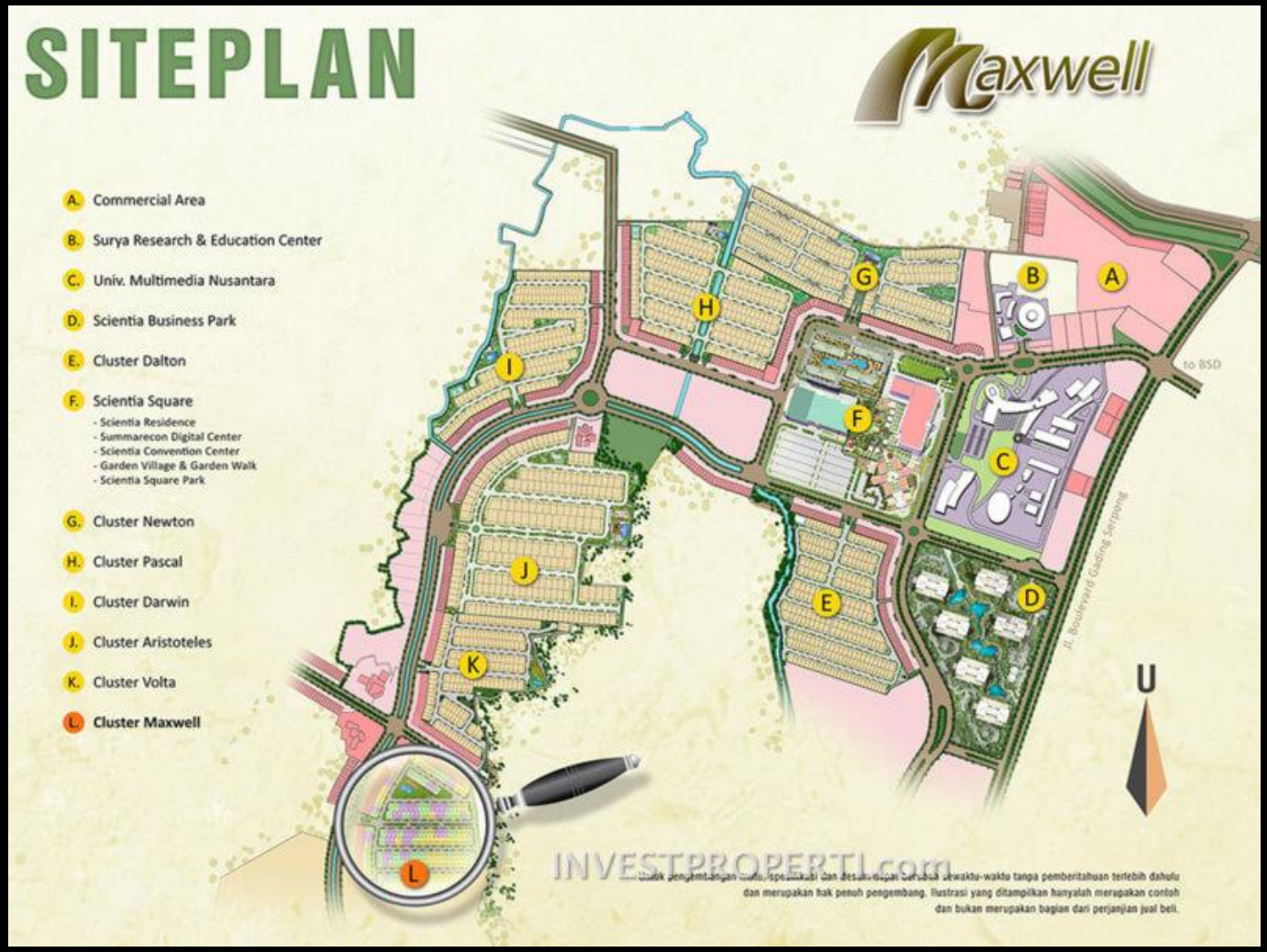


Clusters in Nature

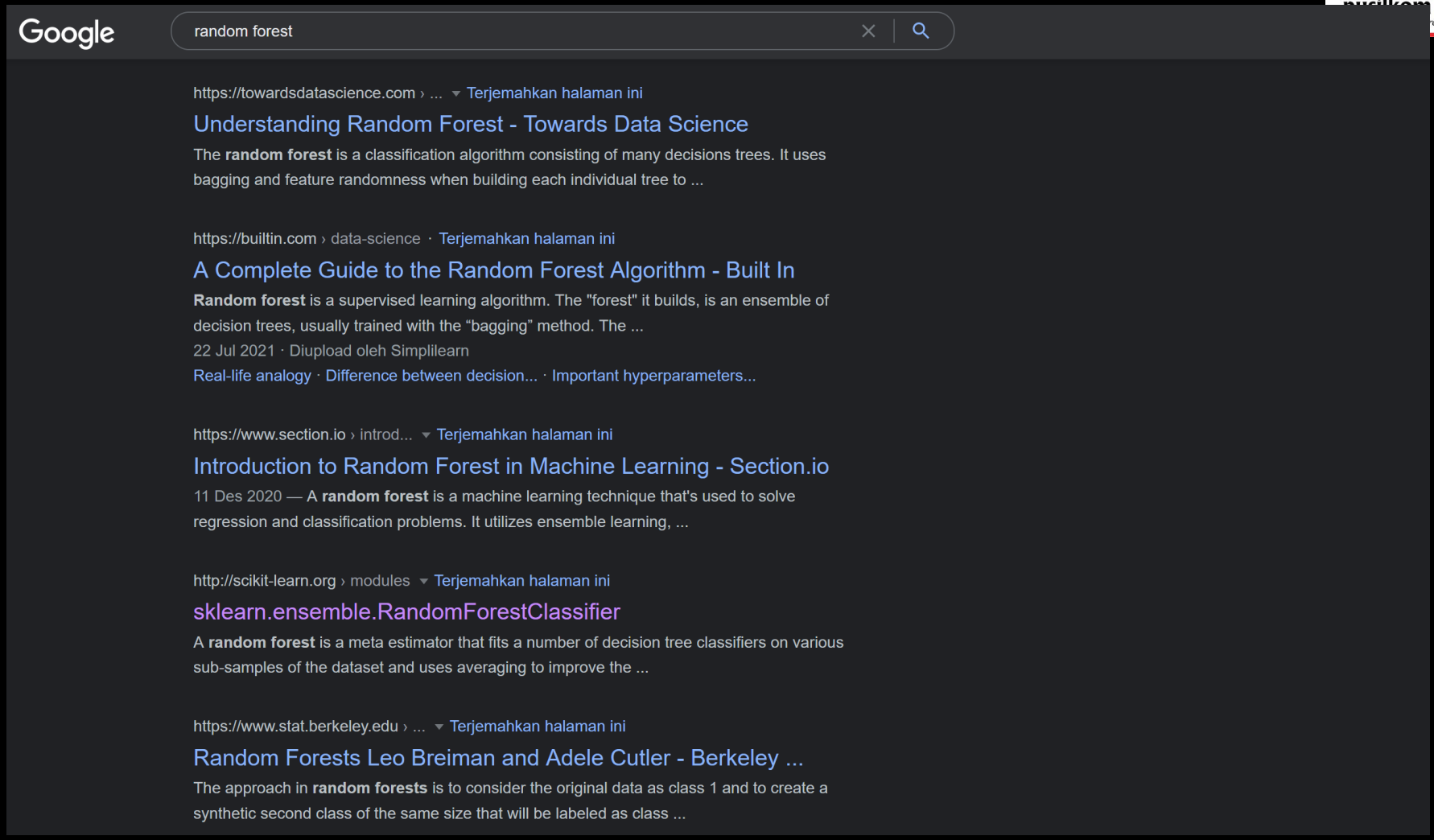


Clusters of Humans

Clusters

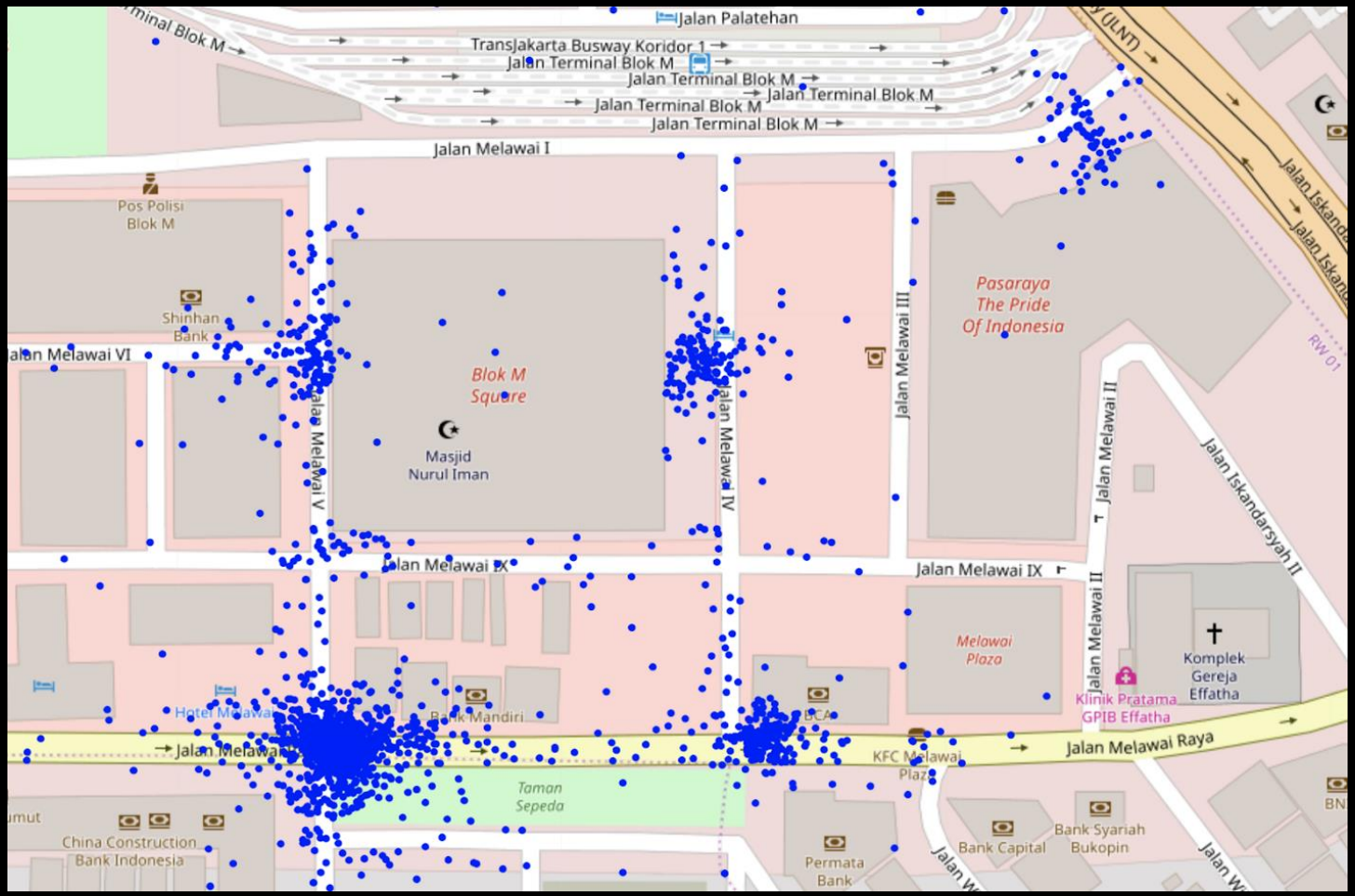

SITEPLAN
Maxwell
A. Commercial Area
B. Surya Research & Education Center
C. Univ. Multimedia Nusantara
D. Scientia Business Park
E. Cluster Dalton
F. Scientia Square
- Scientia Residence
- Summarecon Digital Center
- Scientia Convention Center
- Garden Village & Garden Walk
- Scientia Square Park
G. Cluster Newton
H. Cluster Pascal
I. Cluster Darwin
J. Cluster Aristoteles
K. Cluster Volta
L. Cluster Maxwell
to BSD
Jl. Boulevard Gading Serpong
U
INVESTPROPERTI.com

Text Clusters: Google Search

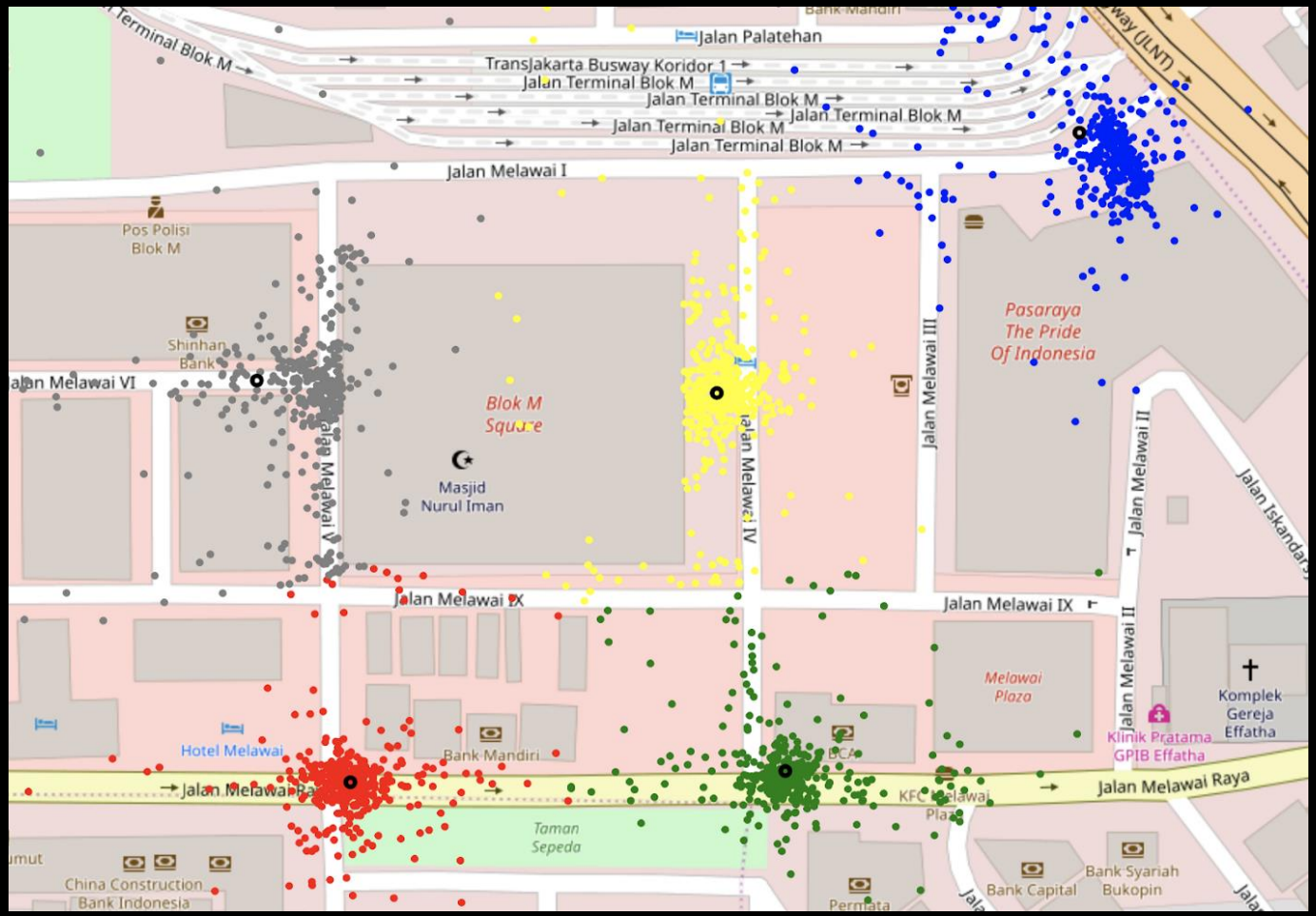


INTRODUCTION

Gojek: Pickup Points



Gojek: Clustering Pickup Points



What s Cluster Analysis?

- **Cluster: a collection of data objects**
  - such that objects in a cluster are similar to one another
  - Yet dissimilar to objects in other clusters
- **Cluster analysis:**
  - Grouping a set of data objects into clusters
- **Clustering is a family of unsupervised learning that attempt to group data points into subsets data point**
  - no predefined classes

Why Clustering?

- **There is no label is our dataset**
  - cannot use supervised methods like classifications, regressions

- **Too much data, difficult to understand the data**
  - instead of looking at millions of records,
  - we group similar data into a small number of clusters
  - Then we analyze the behaviour of the clusters

Clustering Applications

- **Typical applications**
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms
- **Example**
  - In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics, so that it can be enhanced customer relationship management (develop targeted marketing programs).
  - To improve project management, clustering can be applied to partition projects into categories based on similarity so that project auditing and diagnosis can be conducted effectively.

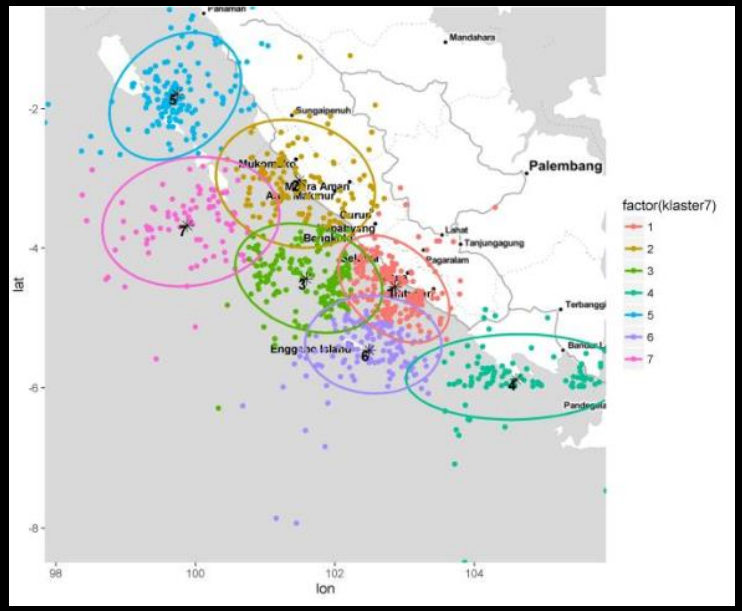
Clustering Applications

- **General Applications:**
  - Pattern Recognition
  - Spatial Data Analysis
    - Create thematic maps in GIS by clustering feature spaces
    - Detect spatial clusters and explain them in spatial data mining
  - Image Processing
  - Economic Science (especially market research)
  - WWW
    - Document classification
    - Cluster weblog data to discover groups of similar access patterns

15

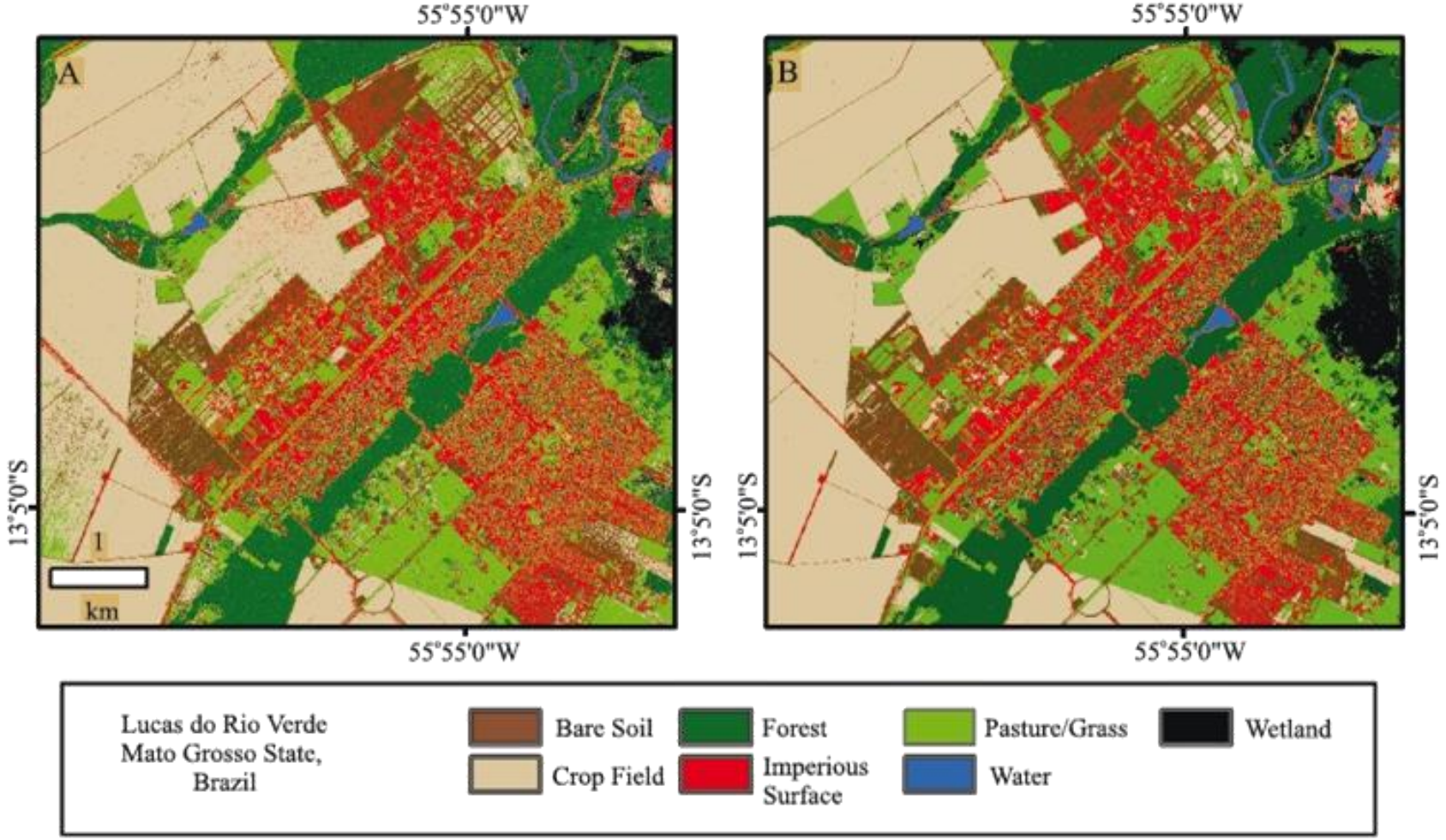
Clustering Applications

- **Insurance: Identifying groups of motor insurance policy holders with a high average claim cost**
- **City-planning: Identifying groups of houses according to their house type, value, and geographical location**
- **Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults**

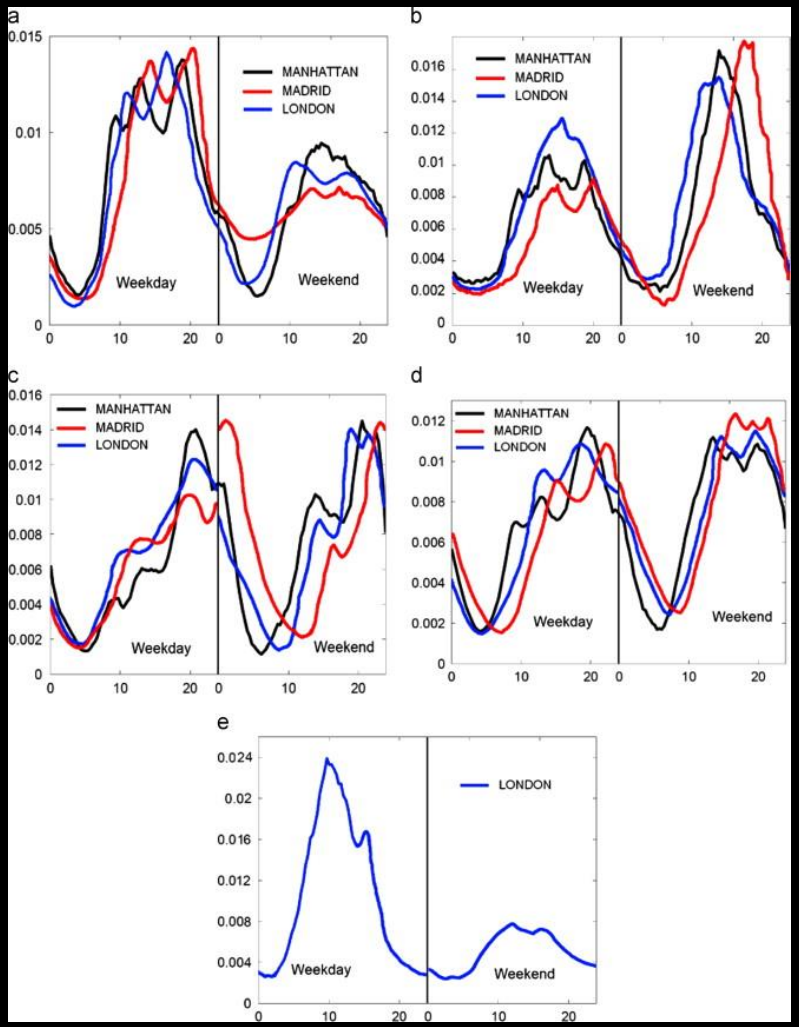


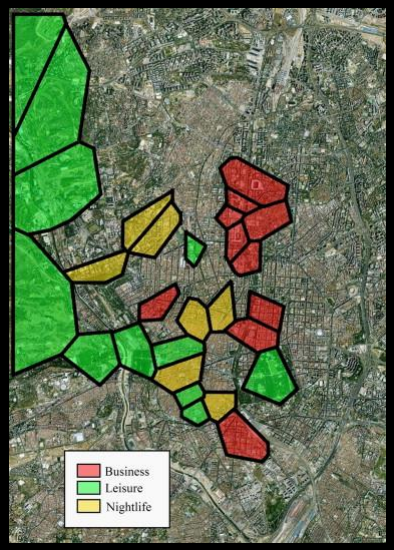
Land use/cover classification in the Brazilian Amazon

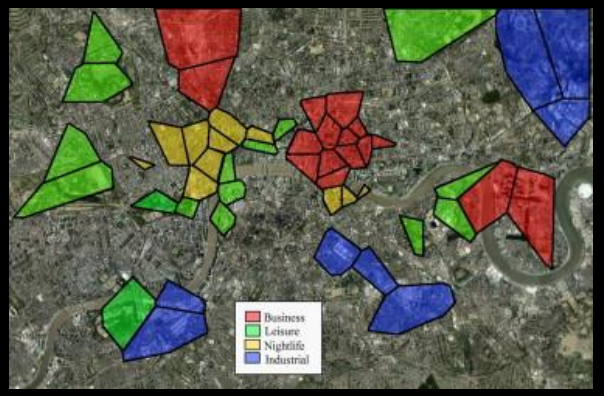
- **Land use: Identification of areas of similar land use in an earth observation database**



Spectral clustering for sensing urban land use using Twitter activity





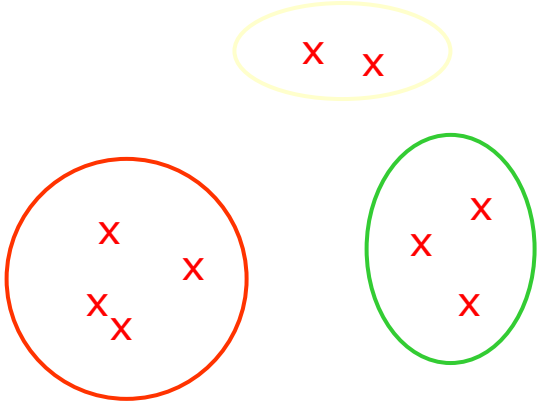


(Frias-Matiniez and Frias-Matiniez, 2014)

Major Clustering Approaches

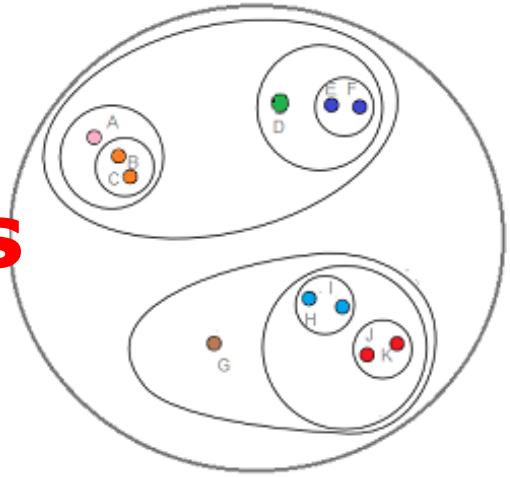
- Partitioning methods
  - Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$.
  - Most partitioning methods are distance-based
- Hierarchical methods
  - Creates a hierarchical decomposition of the given set of data objects.
  - Can be distance-based or density- and continuity-based.

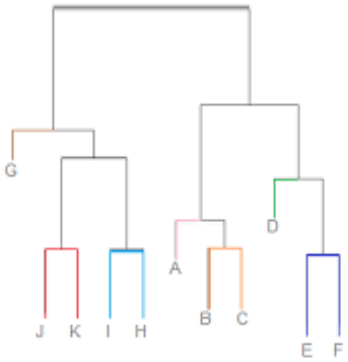
Clustering's Clusters ☺



FLAT CLUSTERING

VS

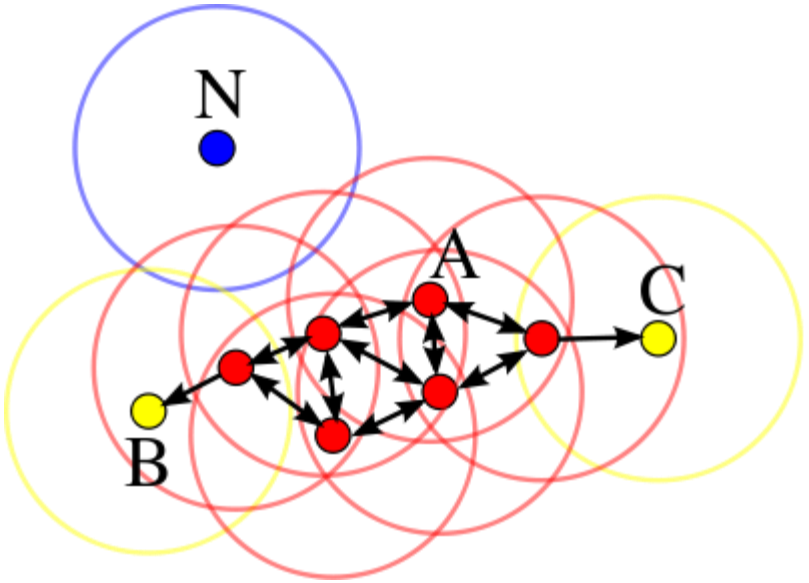




HIERARCHICAL CLUSTERING

Major Clustering Approaches

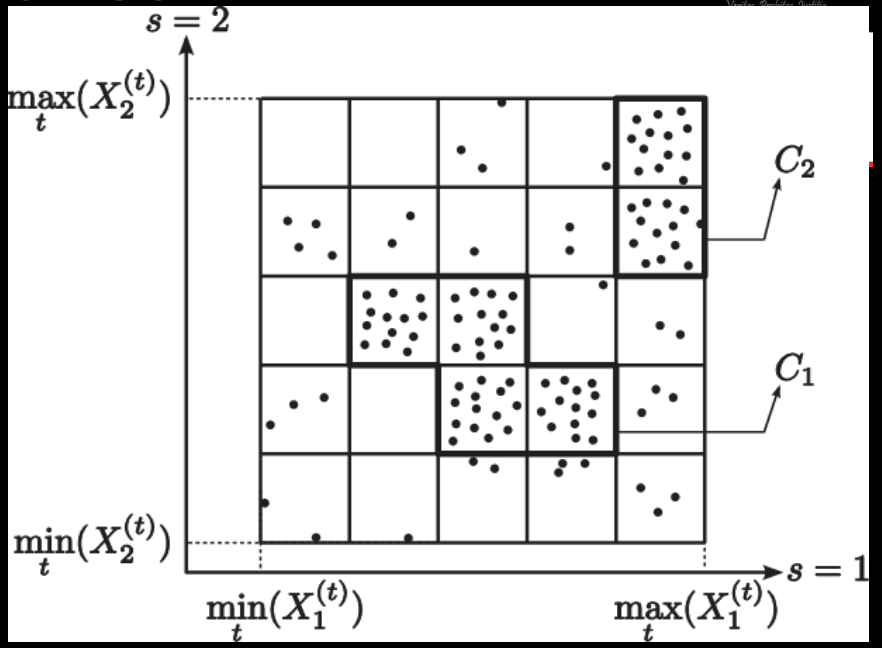
- Density-based methods: connectivity and density functions
  - The general idea is to continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold.

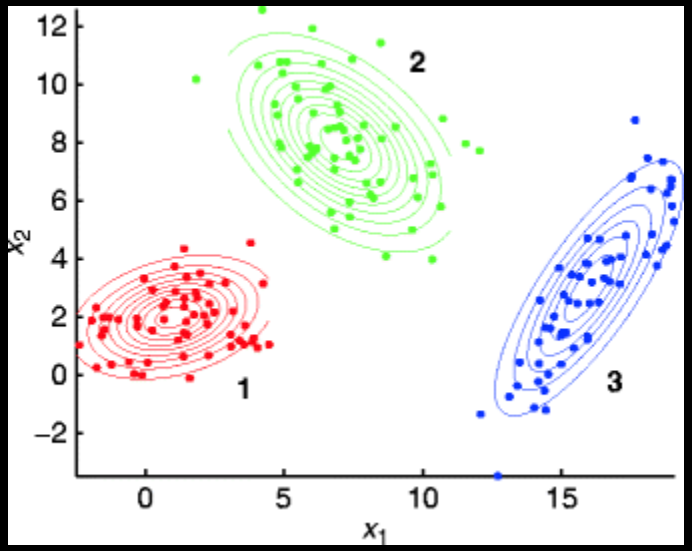


Major Clustering Approaches

- **Grid-based methods: Quantize the object space into a finite number of cells that form a grid structure.**

- **Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other**





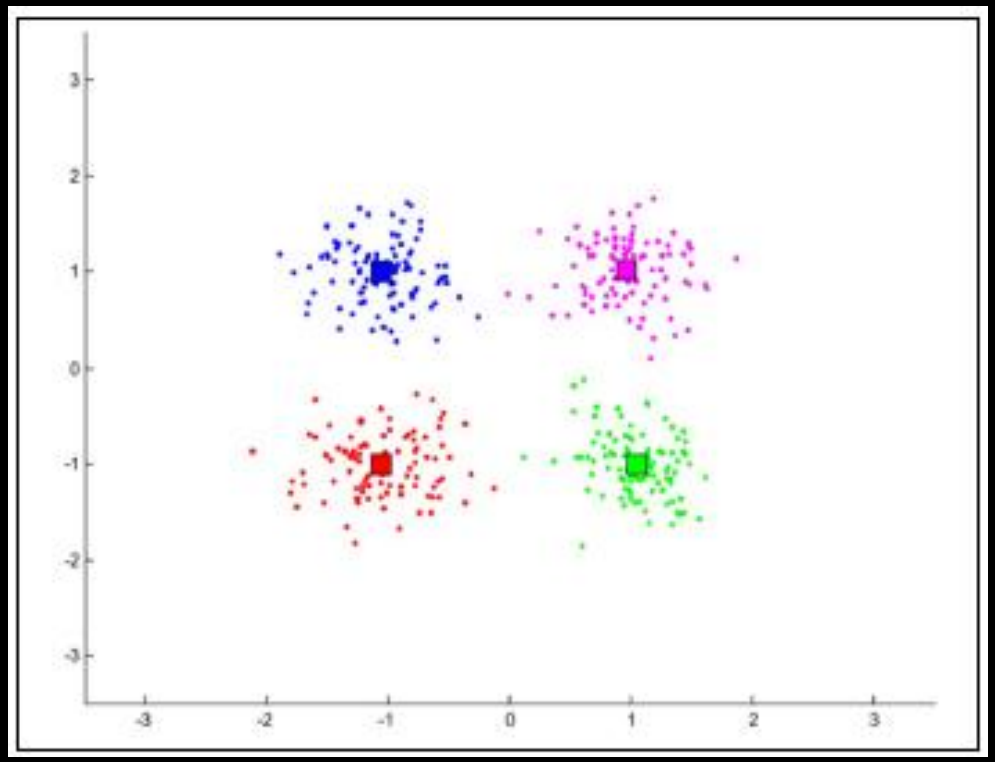
PARTITIONING METHODS (K-MEANS CLUSTERING)

Partitioning Algorithms

- **Partitioning method: construct a partition of database *D* of *n* objects into a set of *k* clusters.**
  - Each group must contain at least one object and each object must belong to exactly one group.
- **Given a *k*, find a partition of *k clusters* that optimized the chosen partitioning criterion**
  - Global optimal: exhaustively enumerate all partitions: C(n,k)
  - Heuristic methods: *k-means* and *k-medoids* algorithms
    - *k-means* (MacQueen'67): Each cluster is represented by the mean value of the objects in the cluster
    - *k-medoids* (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects located near the center of the clusters.
- **These heurustic methods work well for finding spherical-shaped clusters in small to medium-sized databases.**

K-means Clustering

- Cluster: A group of data points that behave similarly
- Centroid: the center of a cluster (for K-means clustering, centroid represented by mean value of the objects in the cluster)



K-means Clustering Algorithm

- Input
  - k : the number of clusters
  - D : a data set containing n objects
- Output
  - A (non empty) set of k clusters
- Algorithm
  1. Arbitrarily choose k objects from D as initial centroid
  2. For each data point assign the point to the nearest centroid
  3. Recalculate the centroid positions
     - Mean value of the data in clusters
  4. Repeat steps 2-3 until **stopping criteria is met**

K-means Clustering

- Nearest centroid can be calculate using distance formula, for example using Euclidean distance,

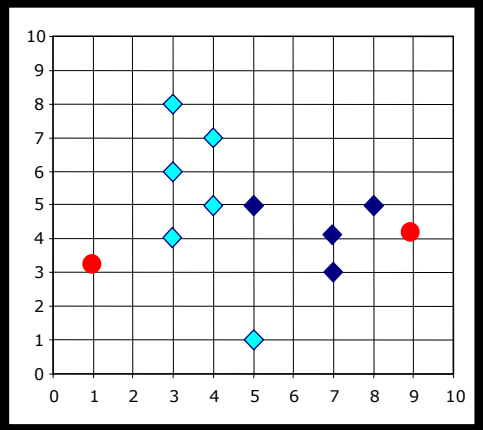
$$dist(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- The quality of cluster C_i can be measured within-cluster variation
  - Which is the *sum of the squared error* between all objects (p) in C_i and the centroid c_i

$$ESS = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

Partitioning Algorithms: *k*-Means

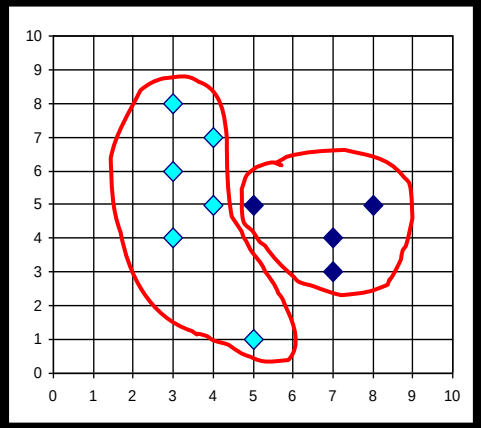
- **Example:**



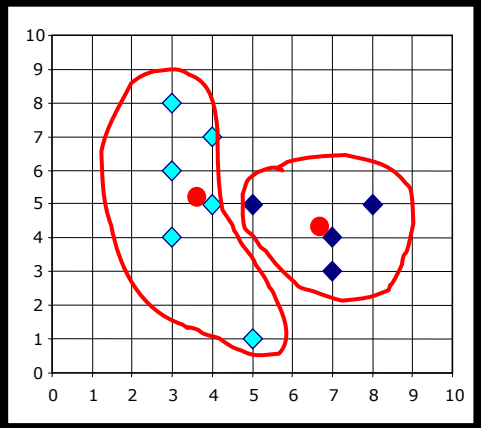
K=2

Arbitrarily choose K object as initial cluster center

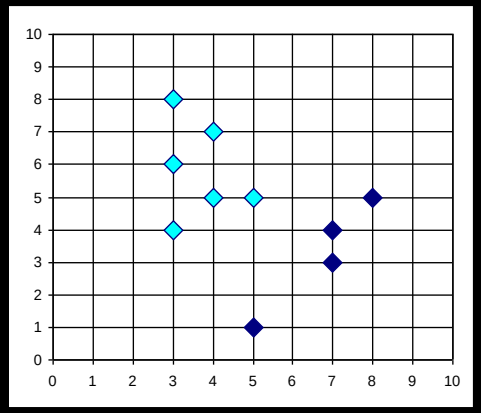
Assign each objects to most similar center



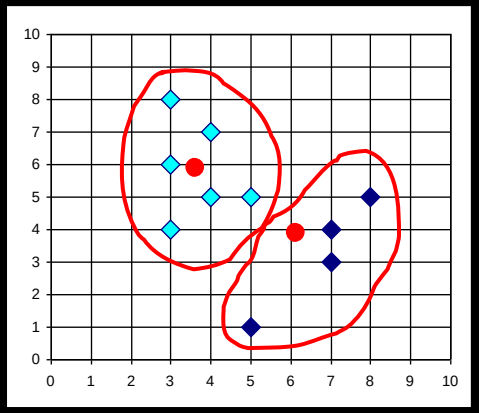
Update the cluster means



reassign

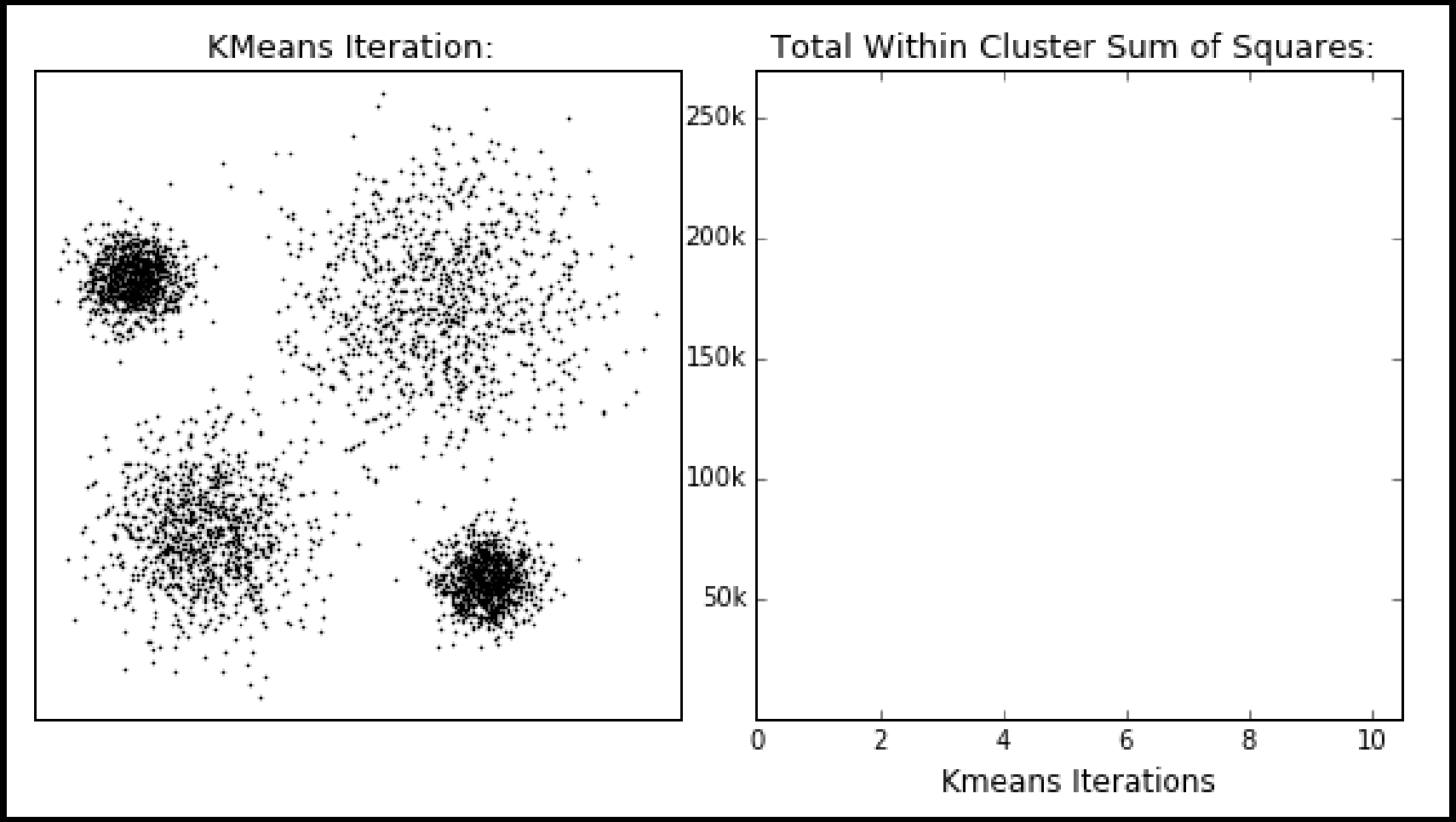


Update the cluster means



reassign

K-means Clustering Illustration

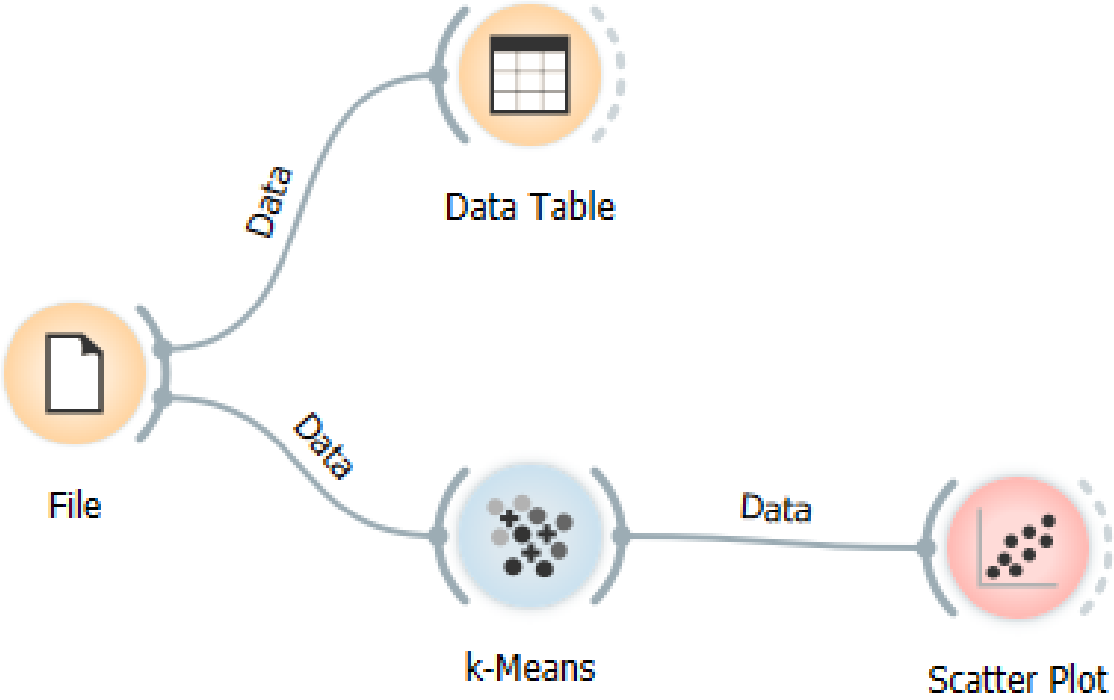


https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

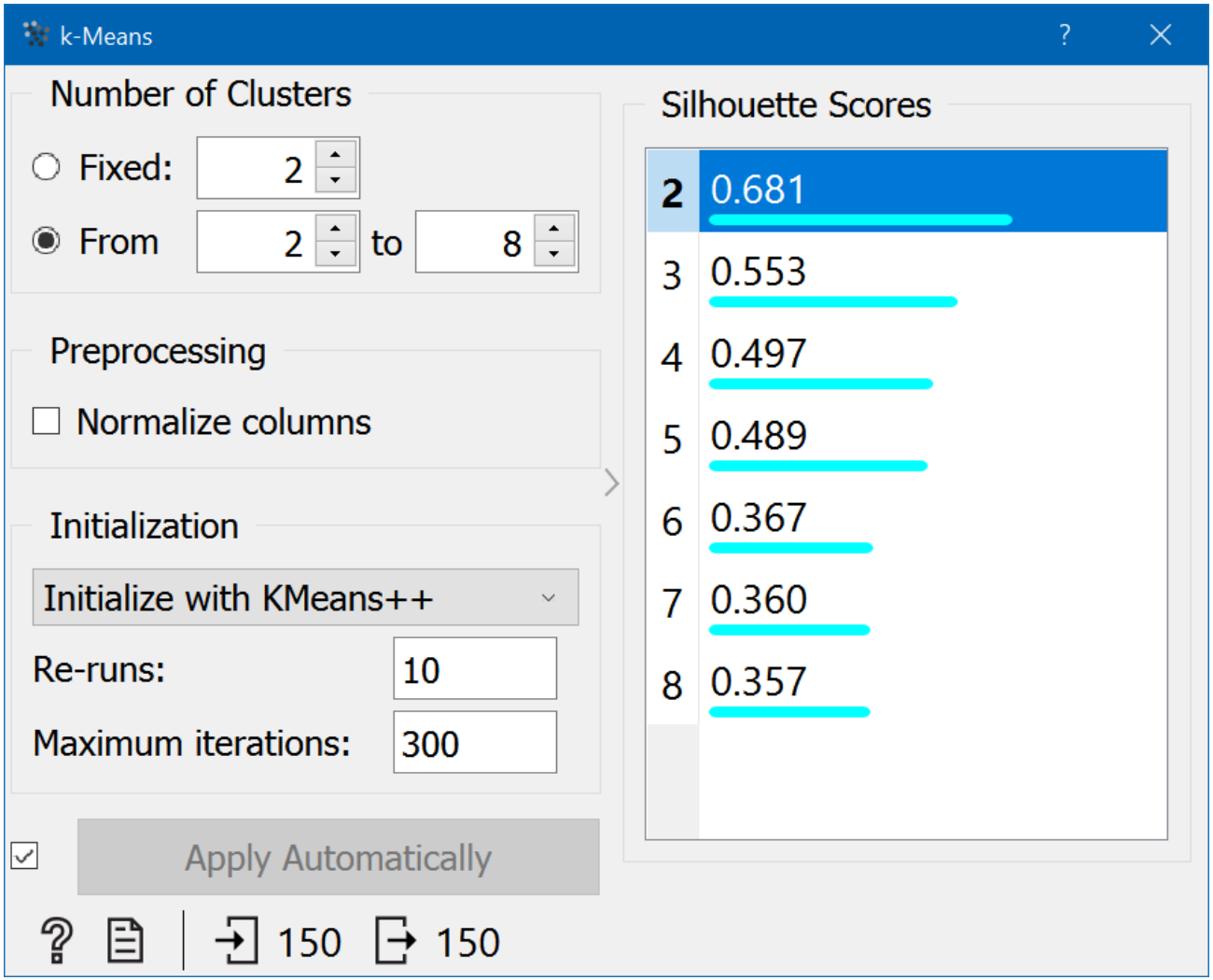
K-means Clustering

- The *k*-means method is not guaranteed to converge to the global optimum and often terminates at a local optimum.
- The result may depend on the initial random selection of centroid
- To obtain good results
  - It is common to run the k-means algorithm multiple times with different initial cluster centers
- Weakness:
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Orange Workflow: k-Means



Orange Workflow: k-means

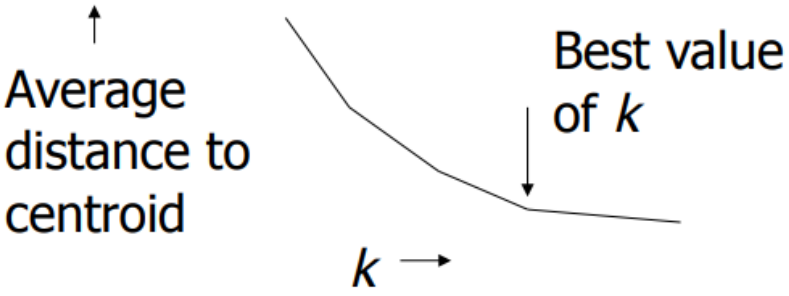


Praktek dengan Orange

- **Ujicoba kmeans dengan data:**
  - Iris
  - cluster-kmeans-1.tab dari lms
  - cluster-kmeans-2.tab dari lms

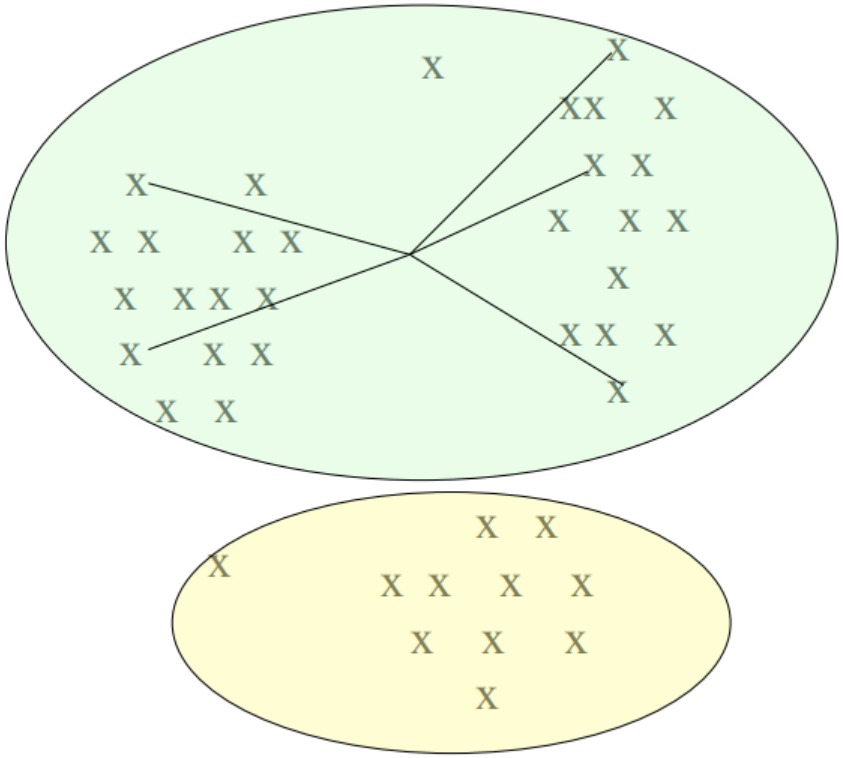
How many clusters? Use the Elbow method.

- **Try different k number of clusters, look at the changes in the average distance to centroid, as k increases**
- **Average falls rapidly until right k, then changes little**

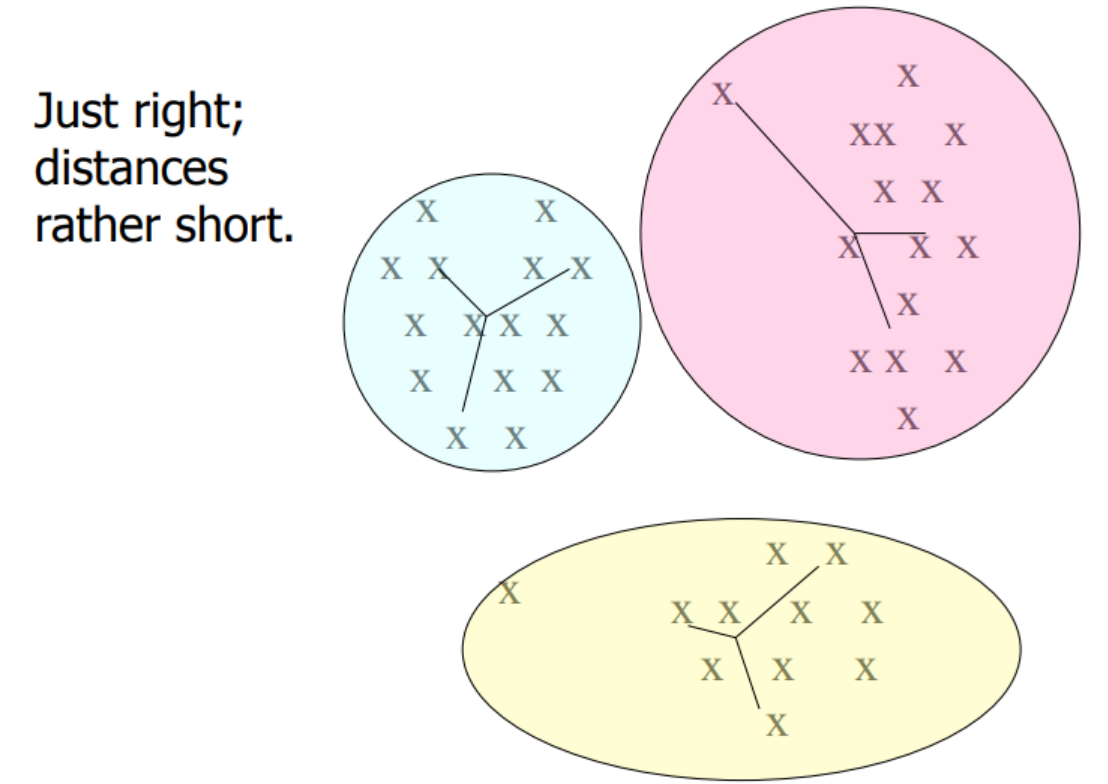


How many clusters? Use the Elbow method.

Too few;
many long
distances
to centroid.

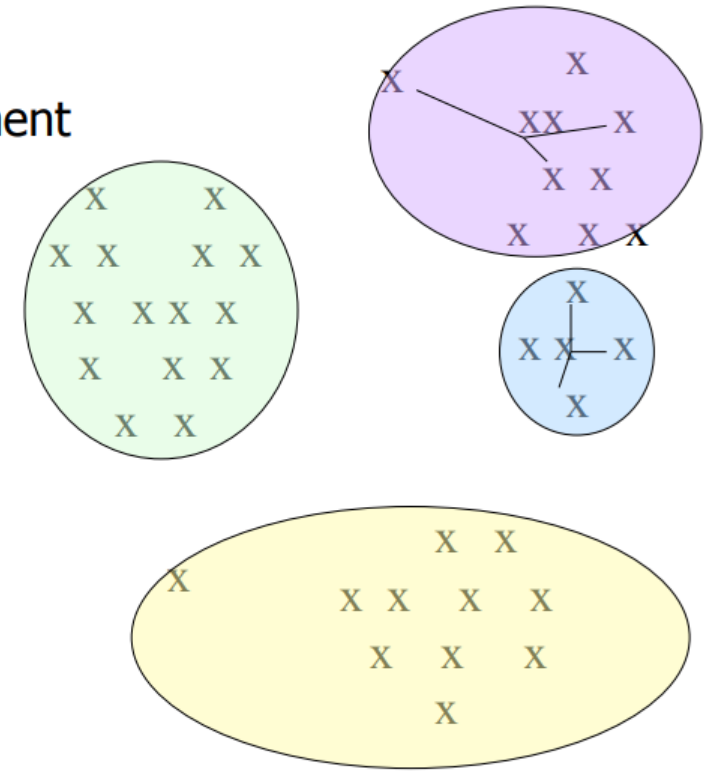


How many clusters? Use the Elbow method.



How many clusters? Use the Elbow method.

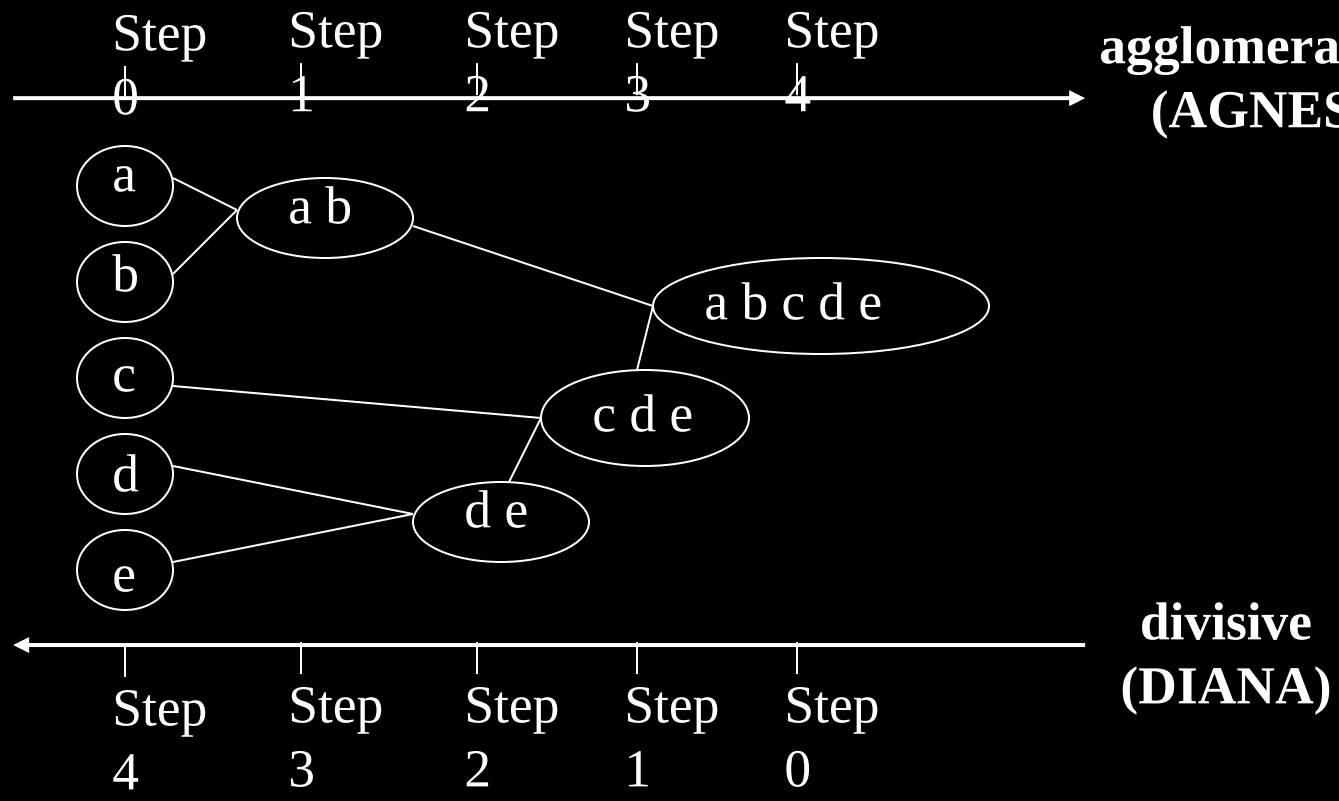
Too many;
little improvement
in average
distance.



HIERARCHICAL METHODS

Hierarchical Clustering

- **Use distance matrix as clustering criteria. This method does not require the number of clusters *k* as an input, but needs a termination condition**



42

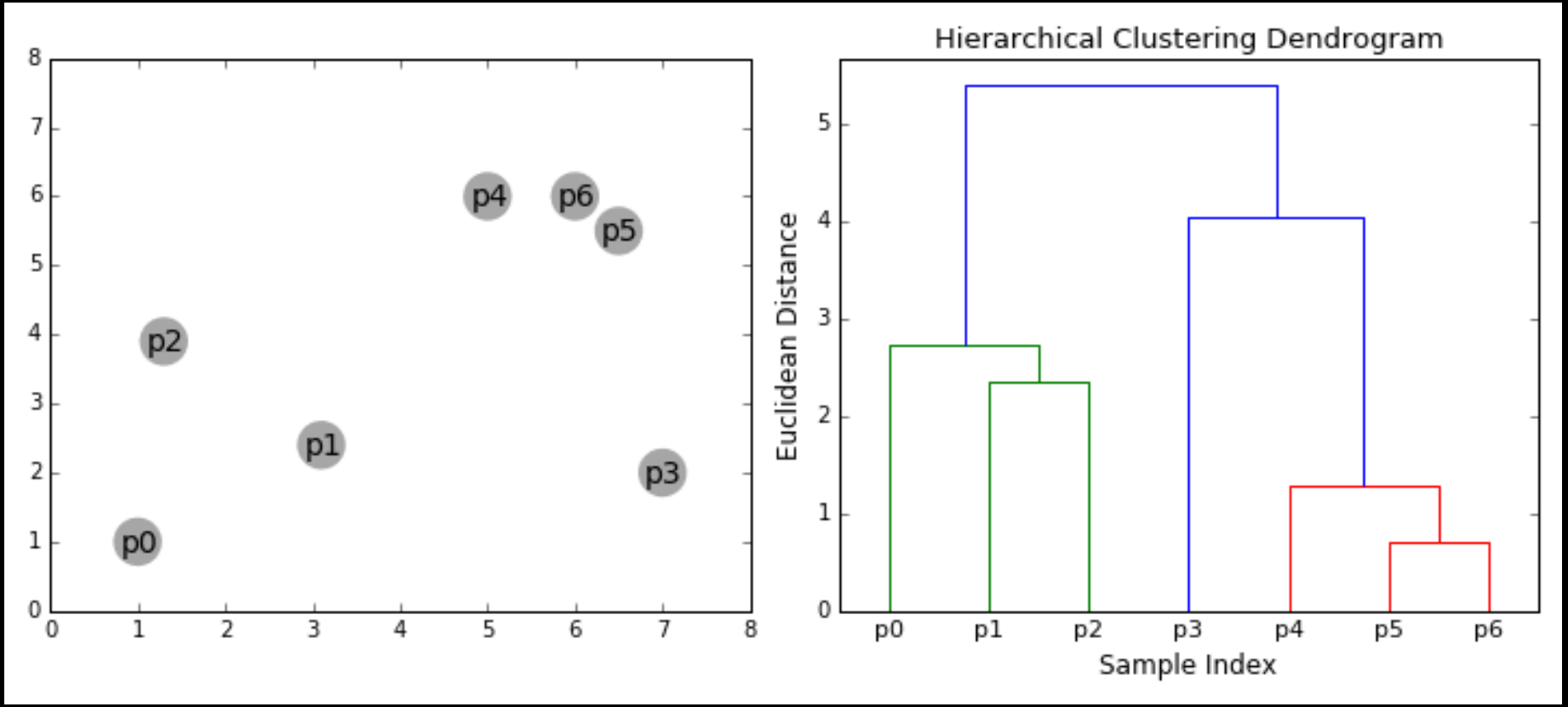
Hierarchical Clustering

- works by grouping data objects into a hierarchy or “tree” of clusters
- Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone.
- This clustering technique is divided into two types:
  - Agglomerative
    - Also called the *bottom-up* approach
    - Initially each data point is considered as individual cluster
    - At each iteration, the similar cluster merge with other cluster until one cluster or *k* cluster are formed
  - Divisive
    - Also called the *top-down* approach
    - Consider all data points as a single cluster
    - At each iteration, separate the data points which are not similar and treat as an individual cluster

Agglomerative Nesting

- Places each object into a cluster of its own
- The cluster are then merged step-by-step according to some criterion
  - For example, using minimum Euclidean distance between any two objects
- A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering
  - It shows how objects are grouped together (in an agglomerative method) or partitioned (in a divisive method) step-by-step
- Two key factor of Hierarchical Clustering are,
  - distance formula
  - Linkage criterion: determined how each cluster will be merge

Hierarchical Clustering



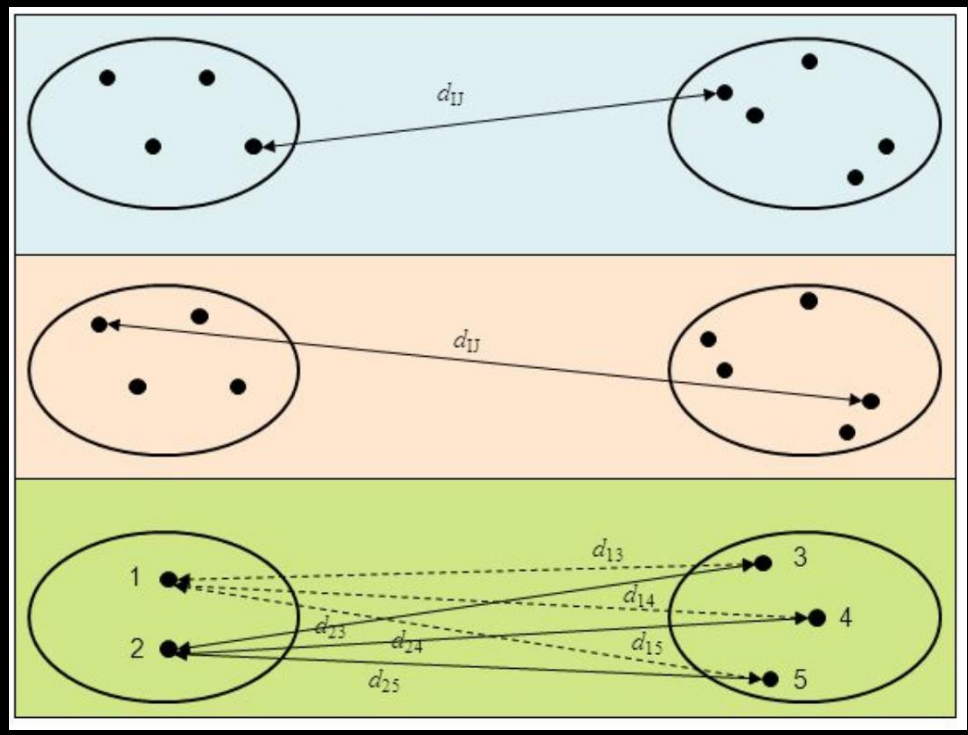
https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Linkage Criterion

- **Single Linkage**
  - Minimum distance between points in each cluster
- **Complete Linkage**
  - Maximum distance between points in each cluster
- **Average Linkage**
  - Average distance between points in each cluster
- **Centroid Linkage**
  - Distance between cluster centroids
- **Ward Linkage**
  - Minimize increase in variance for each cluster

Distance between Clusters

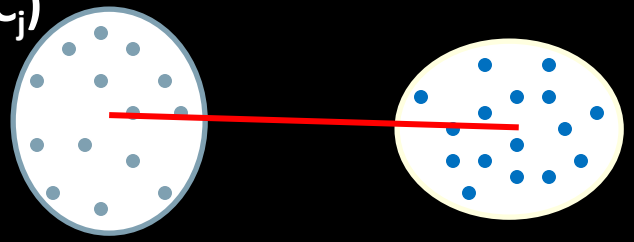
- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = min(t_{ip}, t_{jq})
- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = max(t_{ip}, t_{jq})
- **Average:** avg distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = avg(t_{ip}, t_{jq})



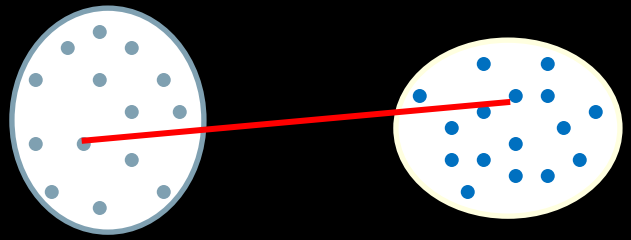
49

Distance between Clusters

- **Centroid: distance between the centroids of two clusters, i.e., dist(K_i, K_j) = dist(C_i, C_j)**

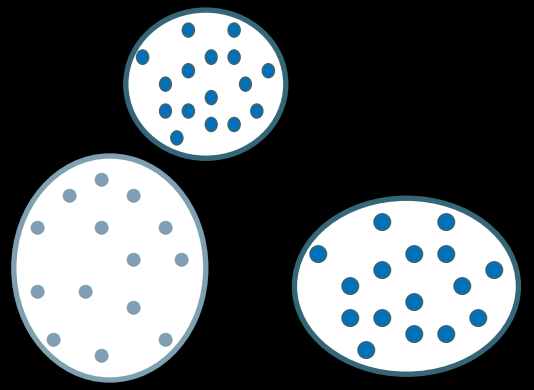


- **Medoid: distance between the medoids of two clusters, i.e., dist(K_i, K_j) = dist(M_i, M_j)**
  - Medoid: a chosen, centrally located object in the cluster



50

Distance between Clusters



- **Ward's Linkage**
  - Ward's minimum variance method
  - the objective function is the Error Sum of Squares (ESS)

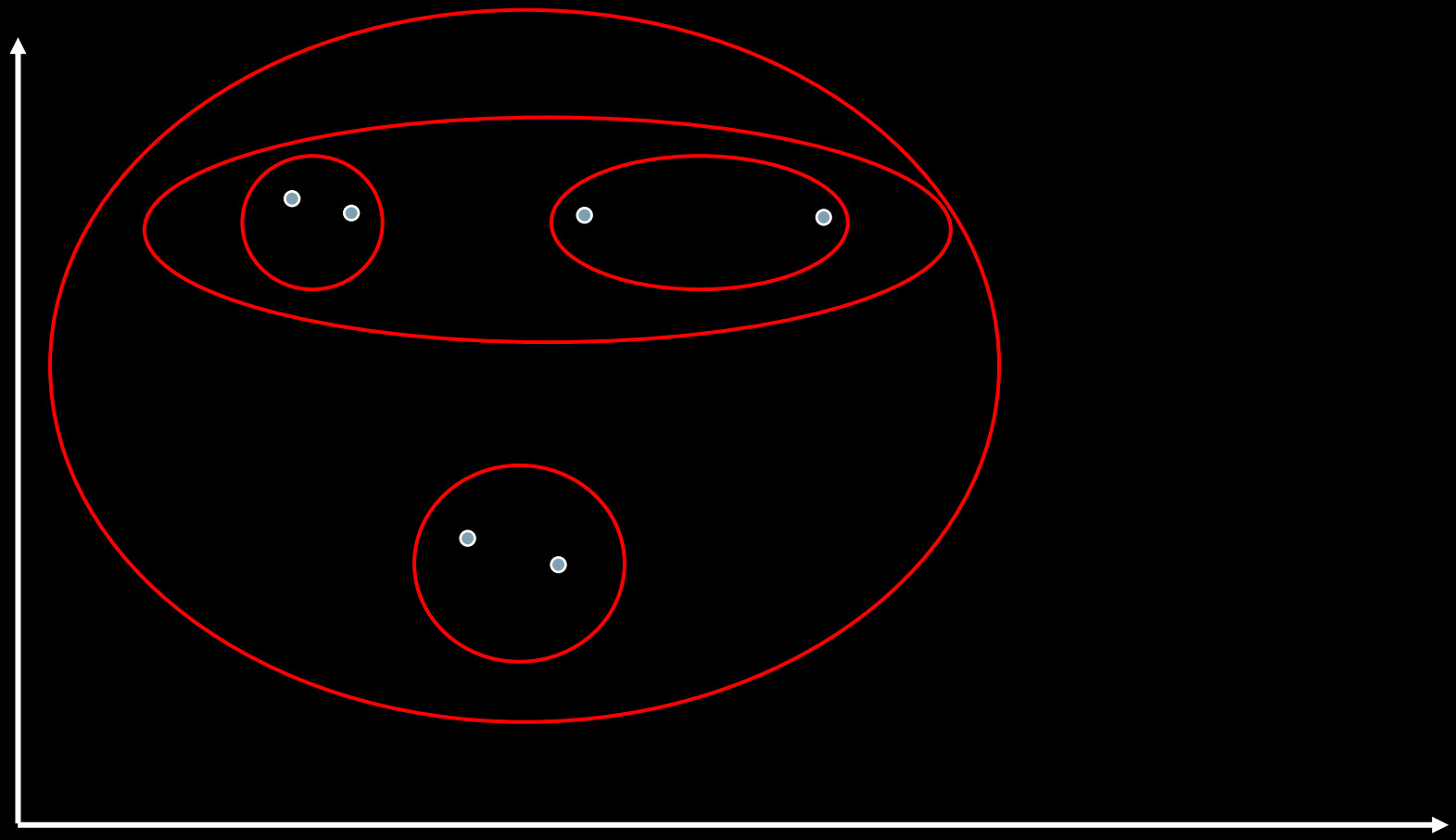
$$ESS(C_j) = \sum_{x_i \in C_j} \|x_i - \mathrm{m_j}\|^2$$

  - Ward's minimum variance criterion minimizes the total within-cluster variance.
    - At each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.

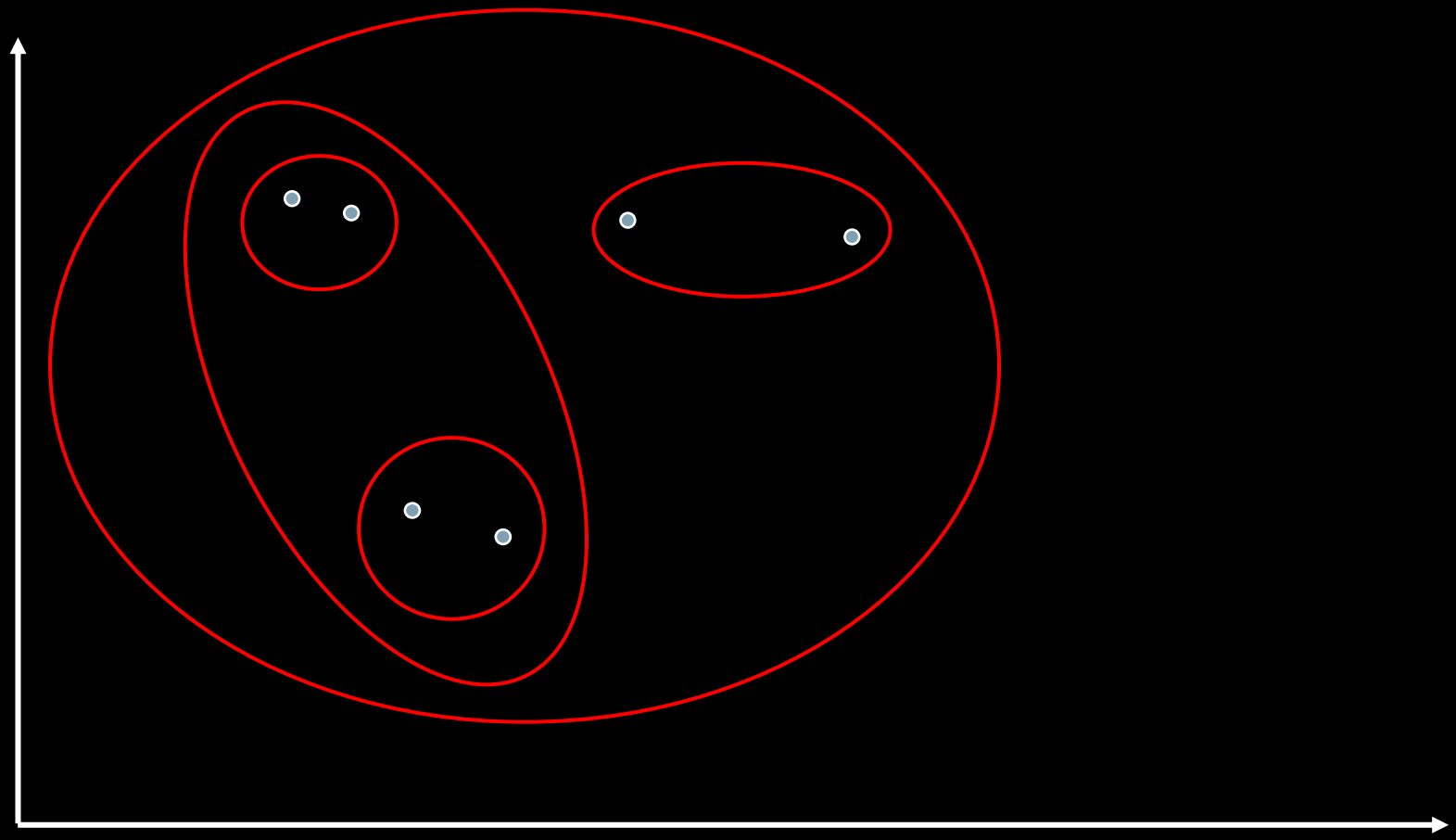
$$d_{ward}(C_i, C_j) = ESS(C_i \cup C_j) - [ESS(C_i) + ESS(C_j)]$$

52

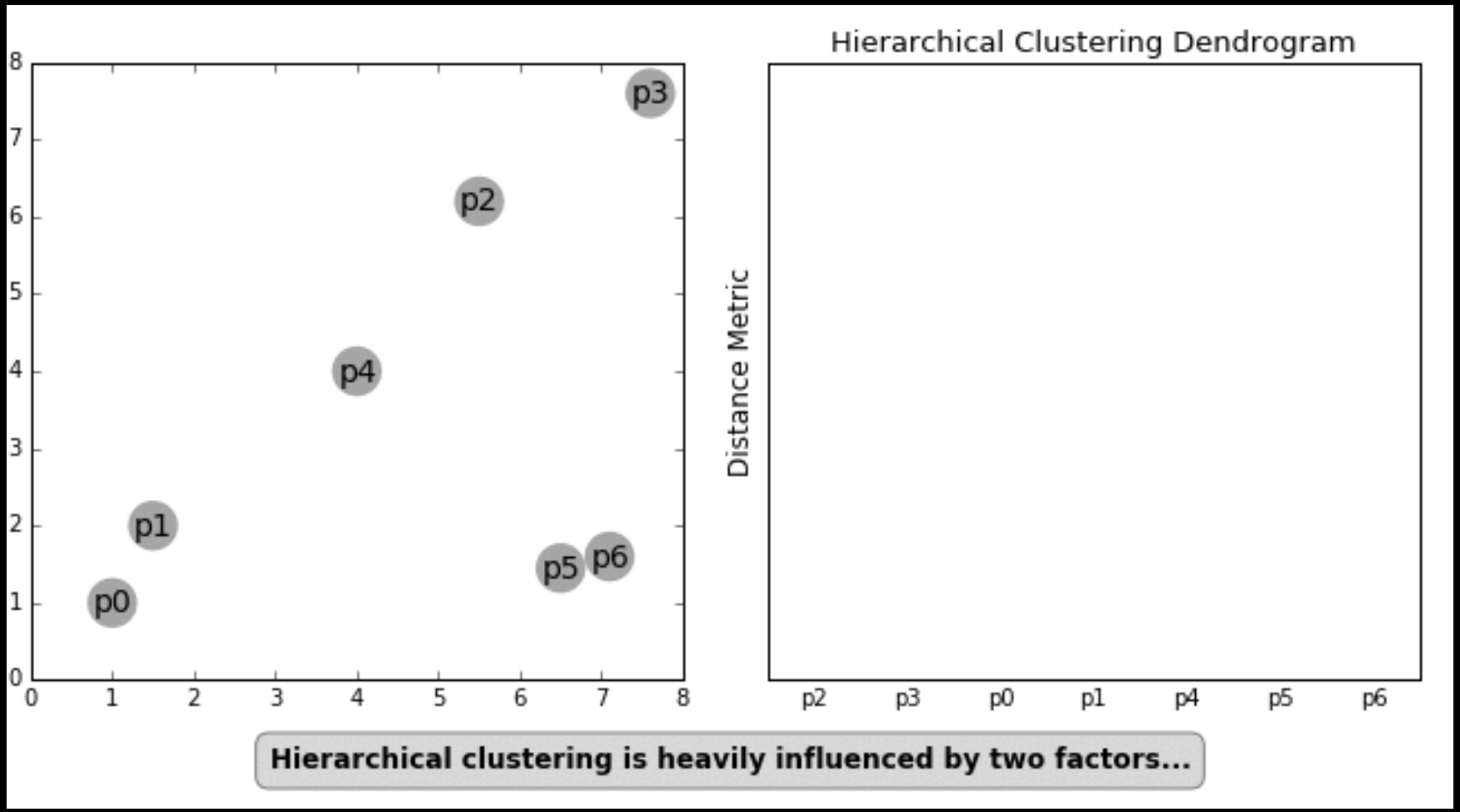
Single Link Example



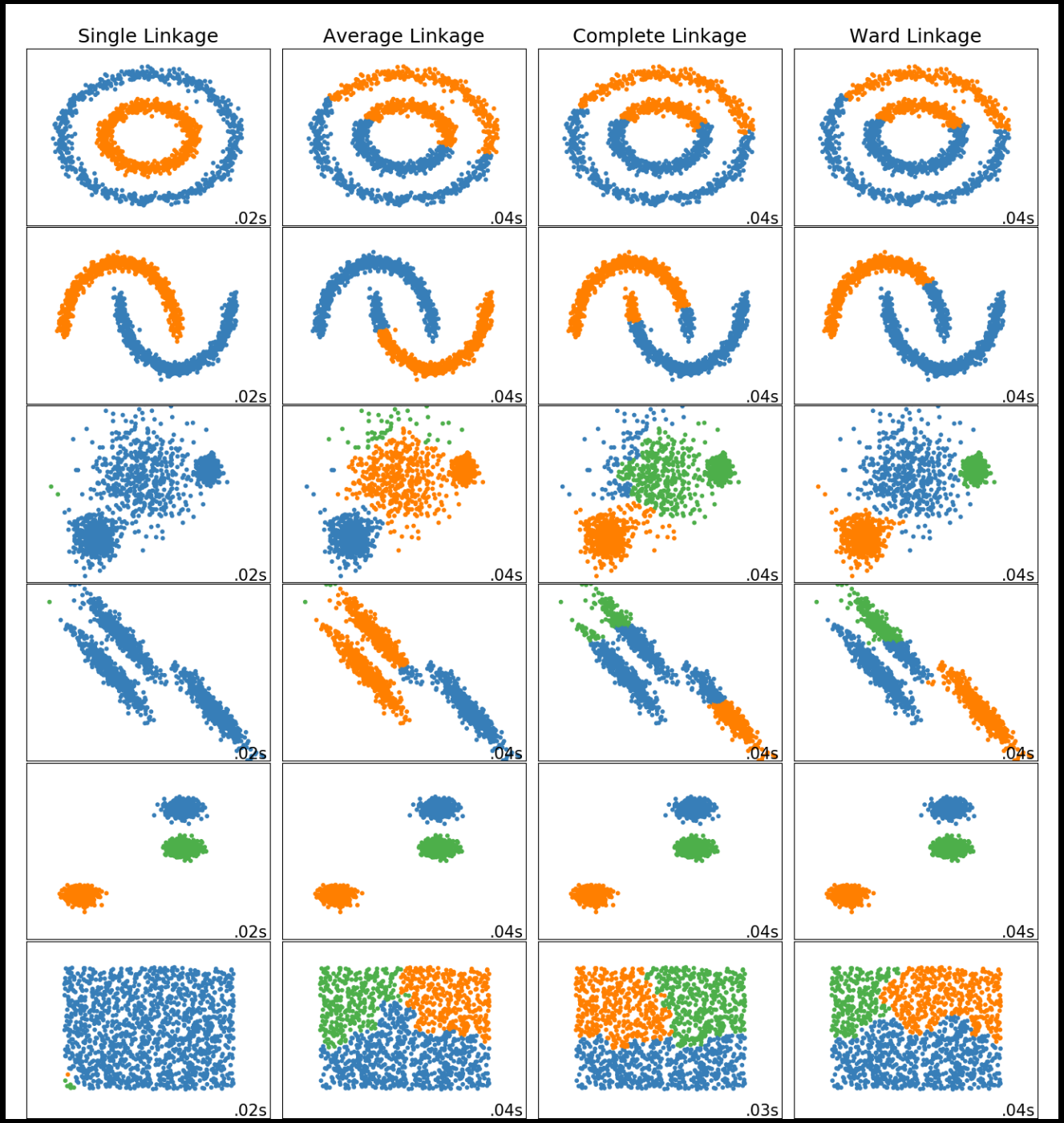
Complete Link Example



Agglomerative Nesting

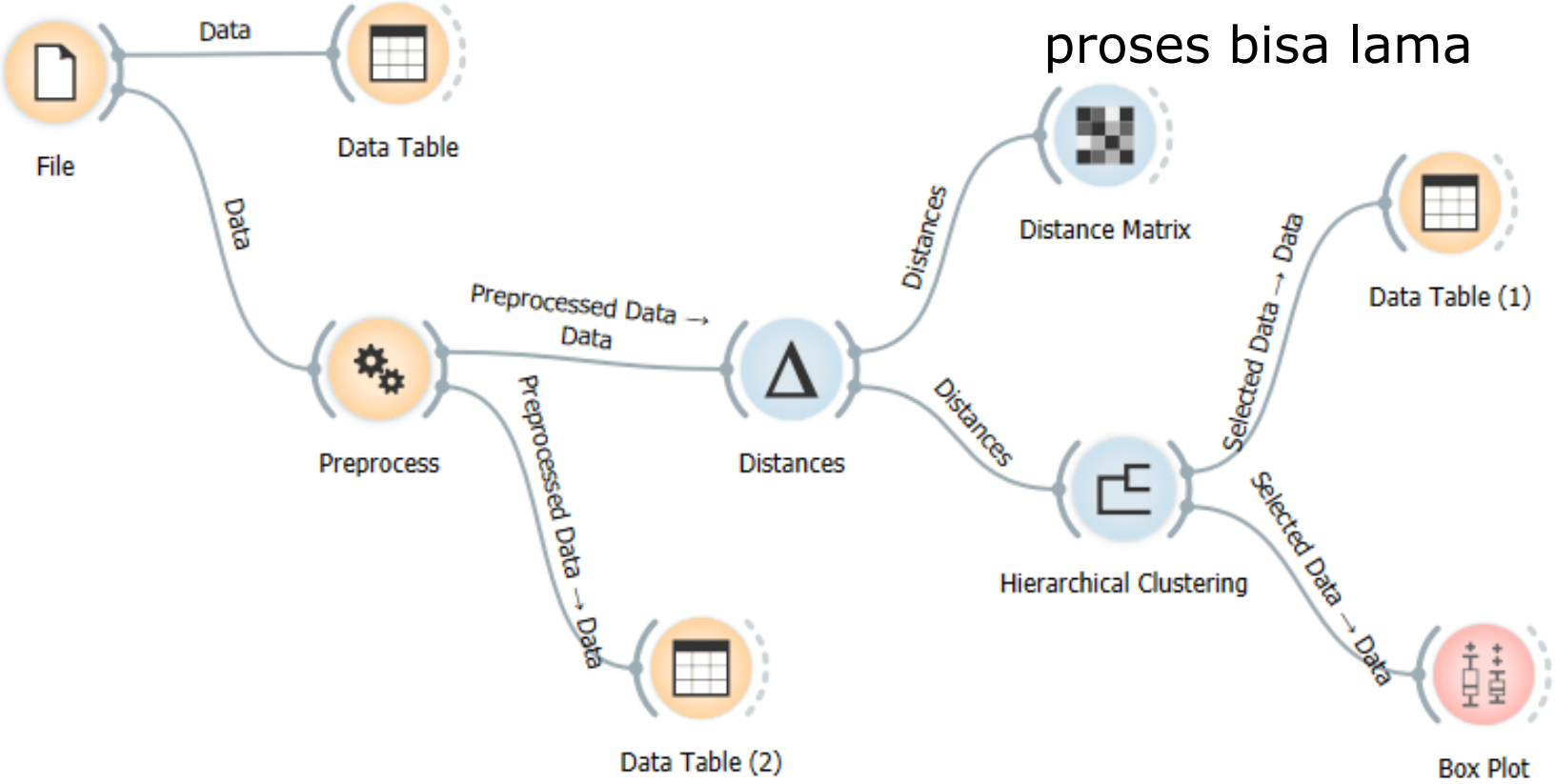


https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/


Single Linkage
Average Linkage
Complete Linkage
Ward Linkage
.02s
.04s
.04s
.04s
.02s
.04s
.04s
.04s
.02s
.04s
.04s
.04s
.02s
.04s
.04s
.04s
.02s
.04s
.04s
.04s
.02s
.04s
.03s
.04s

56

Orange Workflow: Hierarchical Clustering



CLUSTER EVALUATION

Cluster Validity

“Good Clustering”

- **A good clustering method will produce high quality clusters with**
  - high intra-class similarity
  - low inter-class similarity
- **The quality of a clustering result depends on both the similarity measure used by the method and its implementation.**

Cluster Validity

- **How to evaluate the "goodness" of the resulting clusters?**
- **Clustering is ill defined**
  - Unlike supervised learning where labels lead to crisp performance criteria such as accuracy and squared error, clustering quality depends on how the clusters will be used.
  - Devising clustering criteria that capture what users need is difficult. Most clustering algorithms search for optimal clustering based on a pre-specified clustering criterion.
- **Then why do we want to evaluate them?**
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters
  - To decide whether there is noise in the data

Clustering Evaluation

- Davies-Bouldin Index
- Silhouette Index
- Dunn Index

- Basically, they measures the compactness / cohesiveness of the clusters, and separation between clusters.

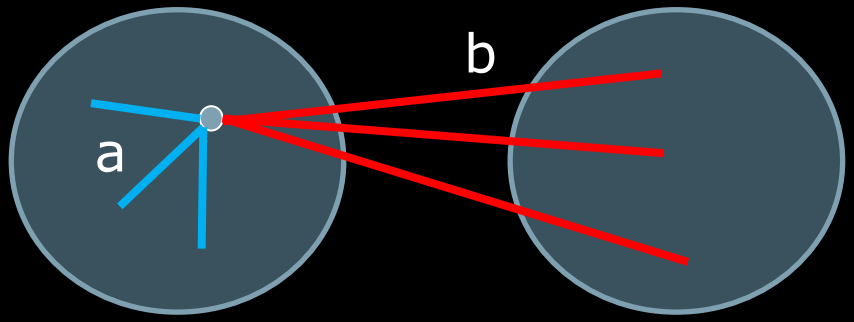
- A good clustering results should have compact and separated clusters.

Silhouette Coefficient

- **Silhouette Coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clustering**
- **For an individual point, *i***
  - ***a*** = average distance of ***i*** to the points in the same cluster
  - ***b*** = min (average distance of ***i*** to points in another cluster)
  - silhouette coefficient of ***i***:

    s = 1 – a/b if ***a < b***

  - Typically between -1 (worst) and 1 (best).



- **Can calculate the average Silhouette width for a cluster or a clustering**

How many clusters? Silhouette method

It's a normalized difference in distance, capturing how close a point is to other points in its own cluster compared to points in the next nearest cluster.

b(i): distance between i and its **next nearest cluster centroid**

(cluster labels assigned from ground-truth classes)

a(i): average distance between i and all of the other points in **its own cluster**

For a single point, i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Sprawling, overlapped clusters

Tight, well-separated clusters

Silhoutte Coefficient

- **measure of cohesiveness**

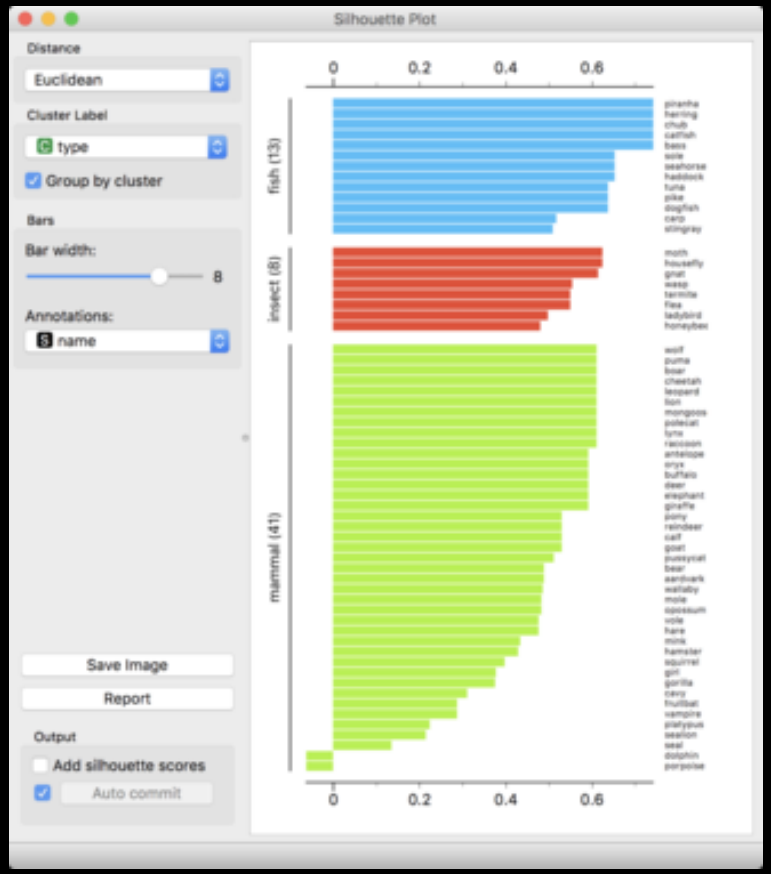
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

- **measure of separation**

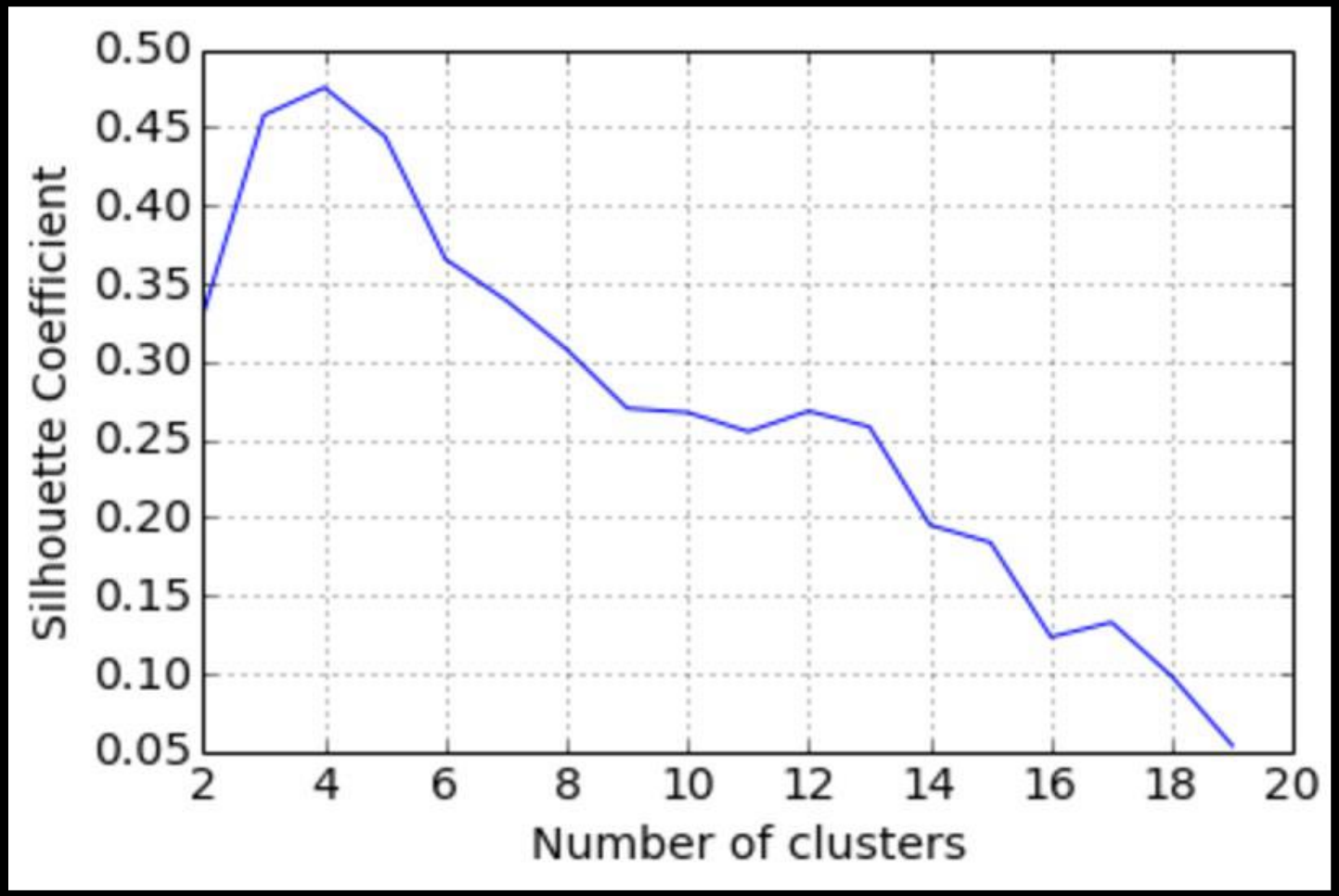
$$\mathrm{b(i)} = \min_{\mathrm{k} \neq \mathrm{i}} \frac{1}{|\mathrm{C_k}|} \sum_{\mathrm{j} \in \mathrm{C_k}} \mathrm{d(i,j)}$$

- **combine both measures**

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



Choosing an optimal number of K and cluster validation



Optimal number of cluster is 4

DATA TYPES

Type of data in clustering analysis

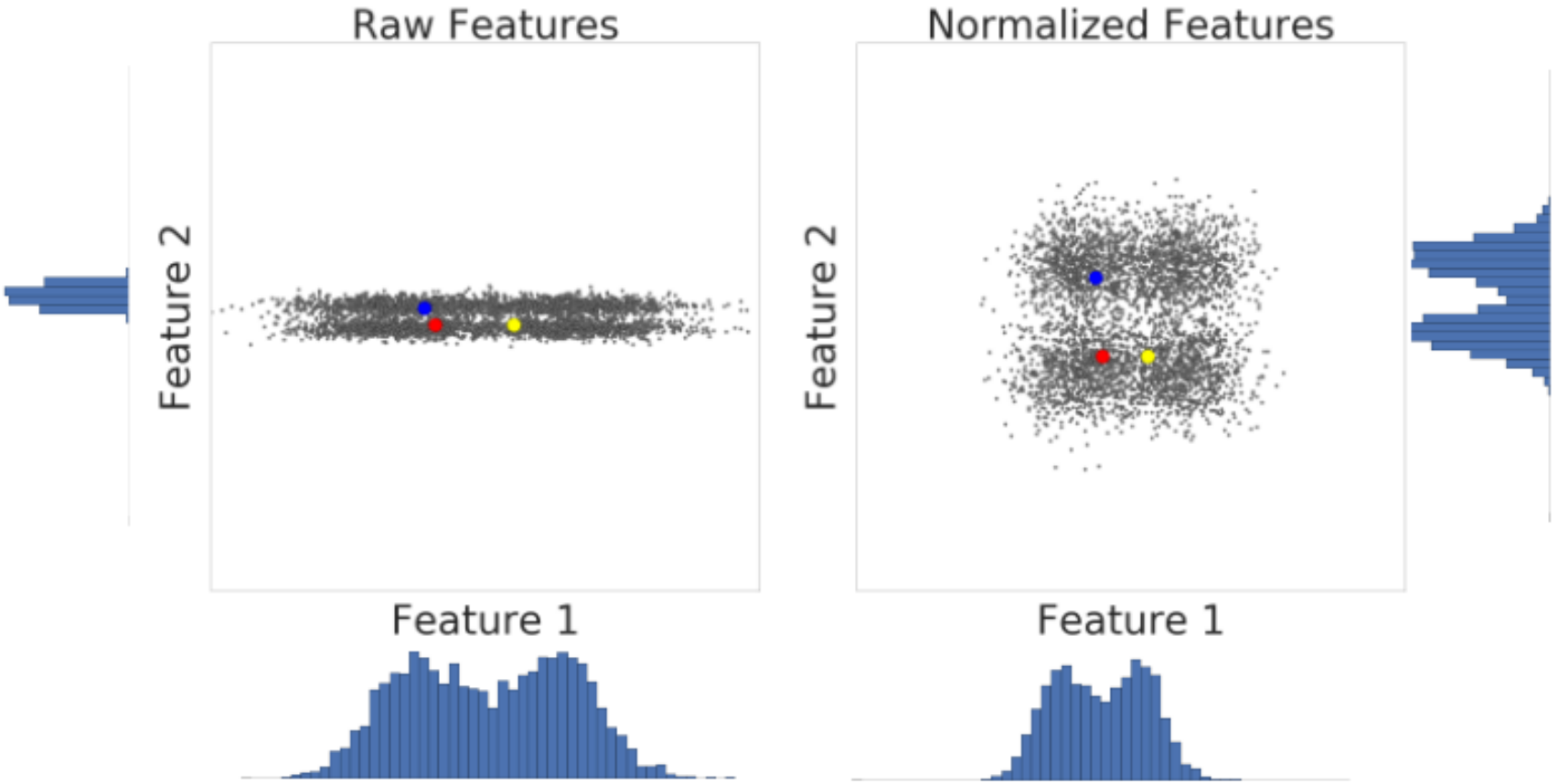
- **Nominal:**
  - Binary variables
- **Ordinal**
- **Interval-scaled variables**
- **Ratio variables**
- **Variables of mixed types**

- **Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: d(i,j)**
- **The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.**

Scale	Data Type	Operations	Distinct Features	Central Tendency
Nominal	Discrete	=, ≠	Categories Only	Mode Only
Ordinal	Discrete	=, ≠ ≤, ≥	Ordered Categories	Mode & Median
Interval	Continuous	=, ≠, ≤, ≥ +, -	Meaningful Intervals	Mode, Median & Mean
Ratio	Continuous	=, ≠, ≤, ≥ +, -, ×, ÷	Absolute Zero Value	Mode, Median & Mean

Interval-scaled variables

- **Interval-scaled variables are continuous measurements of roughly linear scale.**
  - weight, height, latitude and longitude coordinates, and weather temperature.
- **Measurement unit used can affect the clustering analysis**
  - height (from meters to inches), Weight (from kilograms to pounds). The clustering structure may be very different.
- **Expressing a variable in smaller units will lead to a larger range (meters → cms) for that variable, and thus a larger effect on the resulting clustering structure.**
- **To help avoid dependence on the choice of measurement units, the data should be standardized.**
- **Standardizing measurements attempts to give all variables an equal weight → useful when given no prior knowledge of the data.**


Raw Features
Feature 2
Feature 1
Normalized Features
Feature 2
Feature 1

Interval-scaled variables (2)

- **Clusting basketball player → may prefer to give more weig to the variable *height*.**
- ***How can the data for a variable be standardized?***
  - Convert the original measurements to unitless variables.
  - Calculate the **mean absolute deviation**:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

  - where x_{1f}, ..., x_{nf} are n measurements of f, and m_f is the **mean** value of f:

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - Calculate the **standardized measurement**:

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Interval-scaled variables (3)

- ***Once I have standardized the data, how can I compute the dissimilarity between objects?***
  - Dissimilarity (or simmilarity) between two data objects is computed based on the *distance* between each pair of objects:
    - Euclidean distance, Manhattan distance, Minkowski distance
- **Minkowski Distance**

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

where *q* is a positive integer and

***i* = (x_{i1}, x_{i2}, …, x_{ip}) and *j* = (x_{j1}, x_{j2}, …, x_{jp})**

are two *p*-dimensional data objects,

- **If *q* = 1 → Manhattan, If *q* = 2 → Euclidean**
- **Weighting can also be applied.**

Properties

- $d(i,j) \geq 0$
- $d(i,i) = 0$
- $d(i,j) = d(j,i)$
- $d(i,j) \leq d(i,k) + d(k,j)$

Summary

- **Cluster analysis groups objects based on their similarity and has wide applications**
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- **Measure of similarity can be computed for various types of data**
- **Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods**