Cluster analysis

Petr Ocelík

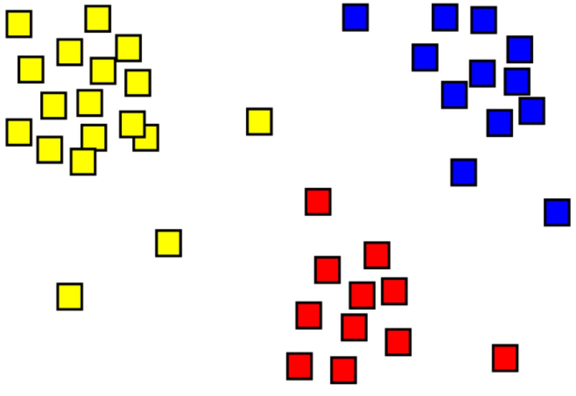
MVPd002 Quantitative Research in International and European Politics

Plan for today

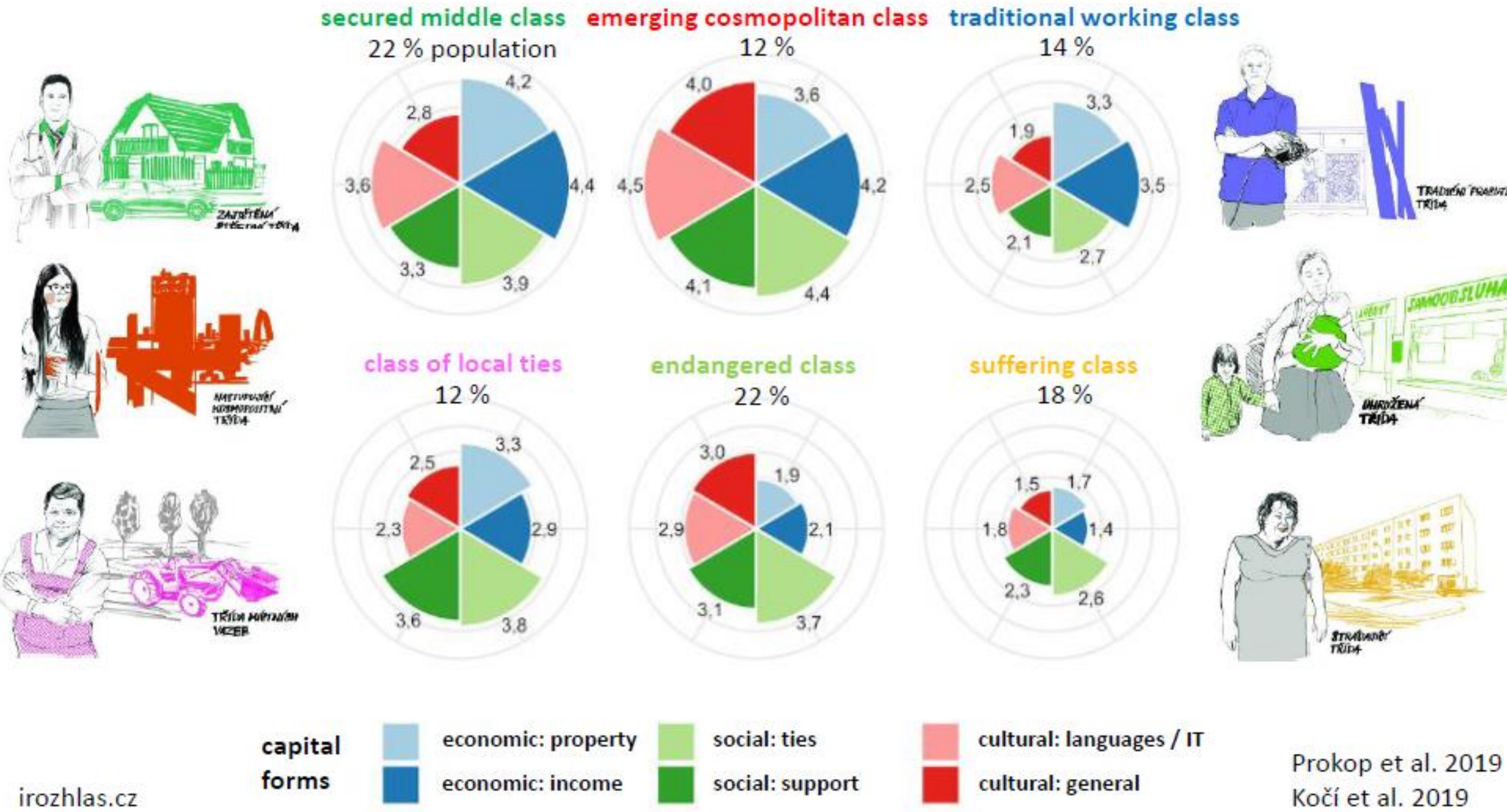
- Intuition
- Cluster analysis step-by-step
- Exercise

Cluster analysis

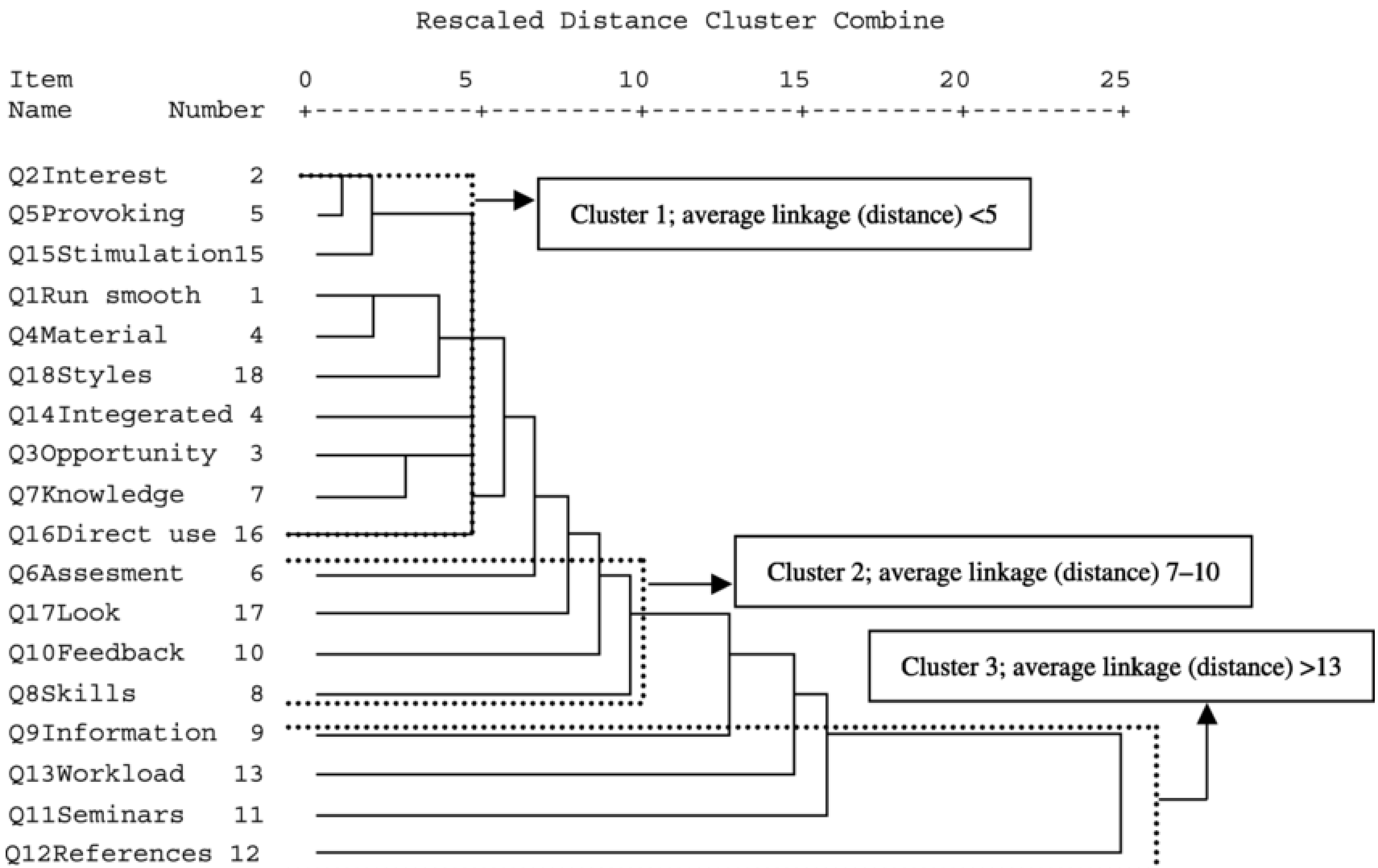
- **Data reduction** technique
- **Cluster:** a grouping of similar objects
- Basic idea: identifying groups of mutually **similar objects** based on particular variable(s)
- **Unsupervised** technique



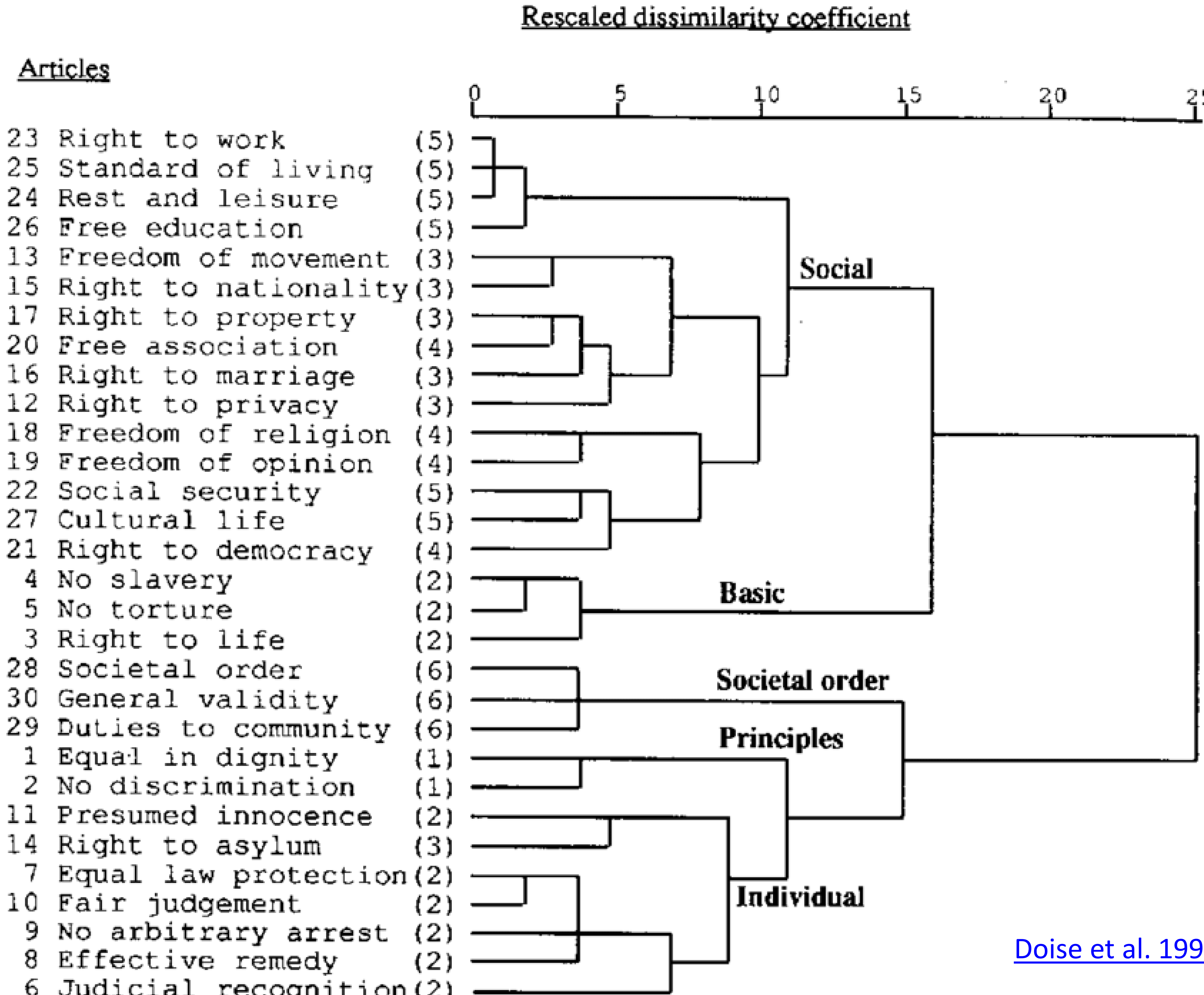
wikimedia commons



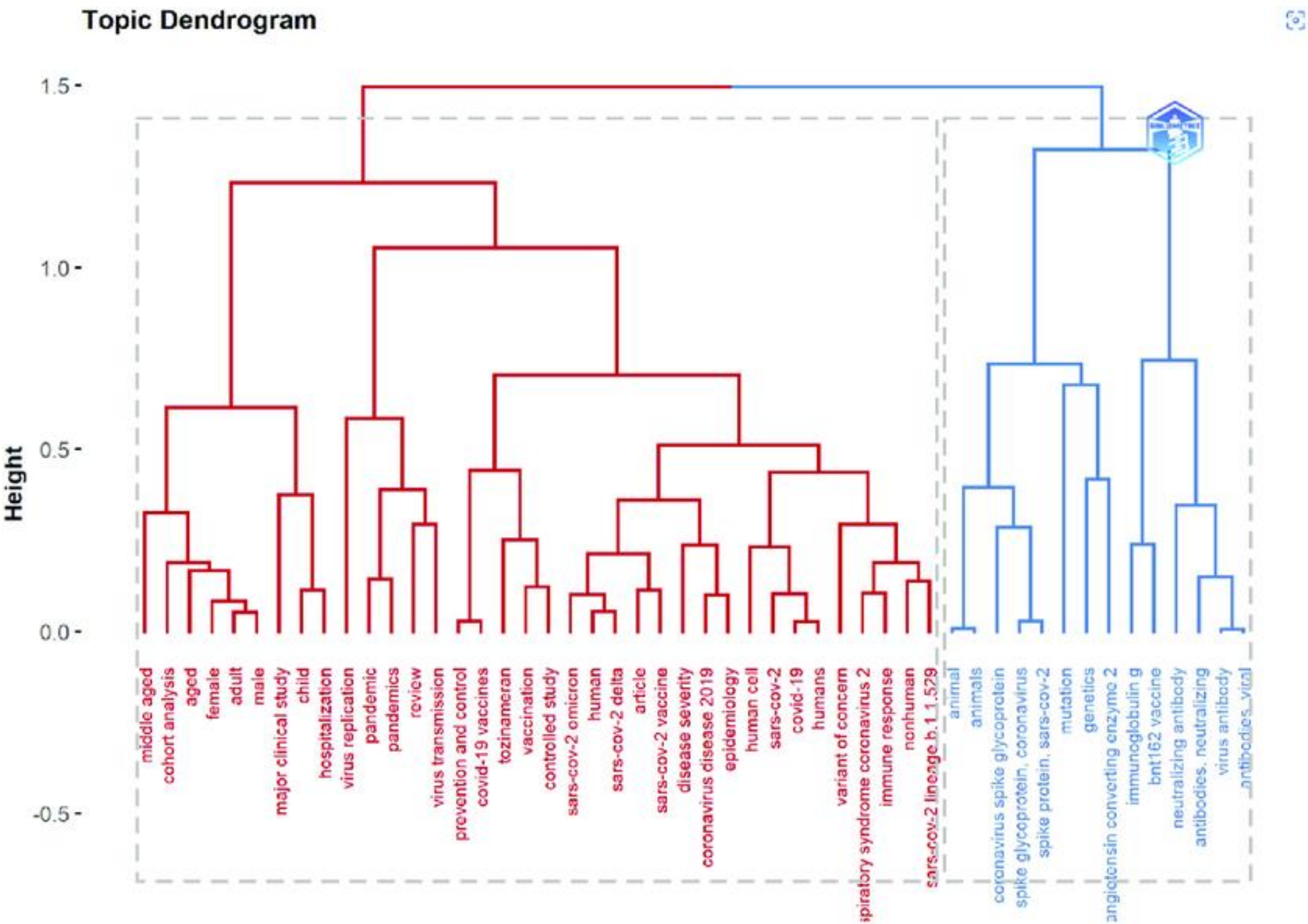
secured middle class
22 % population
4,2
4,4
3,9
3,3
3,6
2,8
emerging cosmopolitan class
12 %
3,6
4,2
4,4
4,1
4,5
4,0
traditional working class
14 %
3,3
3,5
2,7
2,1
2,5
1,9
class of local ties
12 %
3,3
2,9
3,8
3,6
2,3
2,5
endangered class
22 %
1,9
2,1
3,7
3,1
2,9
3,0
suffering class
18 %
1,7
1,4
2,6
2,3
1,8
1,5
capital forms
economic: property
economic: income
social: ties
social: support
cultural: languages / IT
cultural: general
irozhlas.cz
Prokop et al. 2019
Kočí et al. 2019



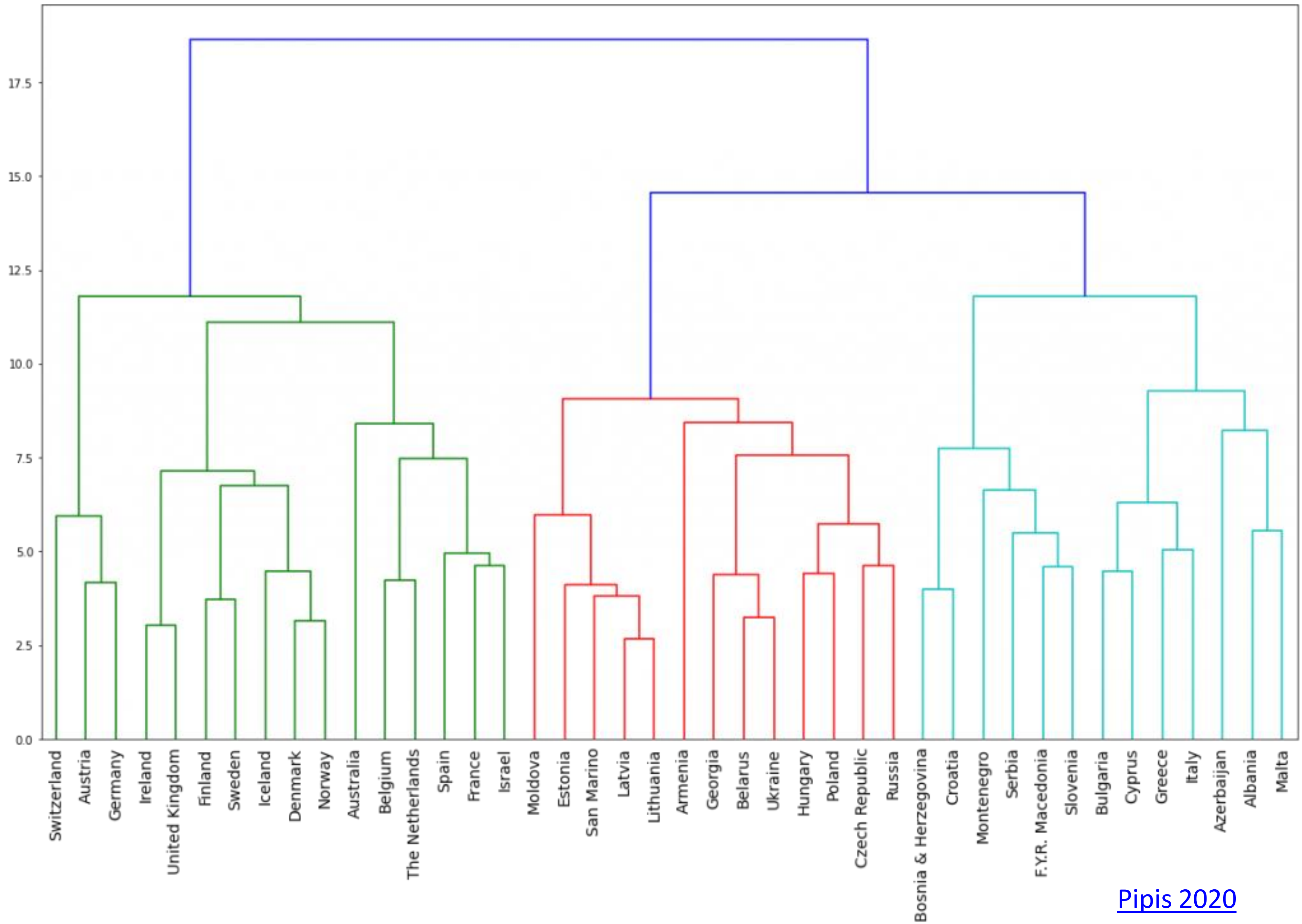
[Ansari & Oskrochi 2006]



Doise et al. 1999



Radanliev et al. 2020


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Switzerland
Austria
Germany
Ireland
United Kingdom
Finland
Sweden
Iceland
Denmark
Norway
Australia
Belgium
The Netherlands
Spain
France
Israel
Moldova
Estonia
San Marino
Latvia
Lithuania
Armenia
Georgia
Belarus
Ukraine
Hungary
Poland
Czech Republic
Russia
Bosnia & Herzegovina
Croatia
Montenegro
Serbia
F.Y.R. Macedonia
Slovenia
Bulgaria
Cyprus
Greece
Italy
Azerbaijan
Albania
Malta

Pipis 2020

Cluster analysis: process

1. Sampling and data collection

2. Similarity measures

3. Clustering methods

4. Cluster solution interpretation

5. Cluster solution diagnostics

1. Sampling and data collection

- What is a **target population**?
- What is the **level of analysis**?
- What is the **unit of observation**?
- What **set of variables** are we interested in?
- Practical considerations

2. Similarity measures

- (Dis)similarity of objects is quantified by using **similarity measures**

- Choice of similarity measure needs to consider:

1. **level of measurement**: categorical vs continuous

2. **data dimensionality**: number of variables (vars)

3. **scale sensitivity**: small vs large data, vars scales

- We distinguish between **association-based** and **distance-based** similarity measures (not exhaustive)

2. Similarity measures: representations

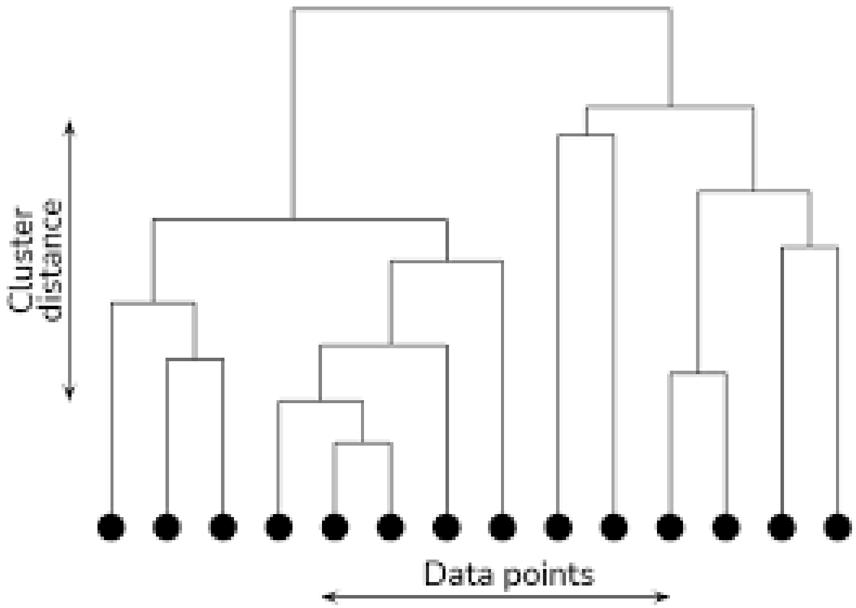
- The degree of (dis)similarity can be captured numerically or **graphically**.

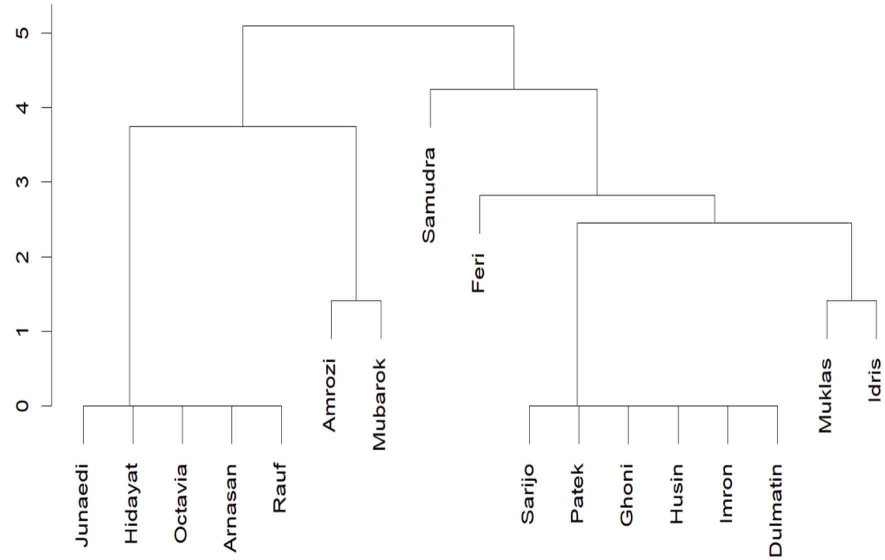
- Dendrogram

- Heatmap

- Cluster profile

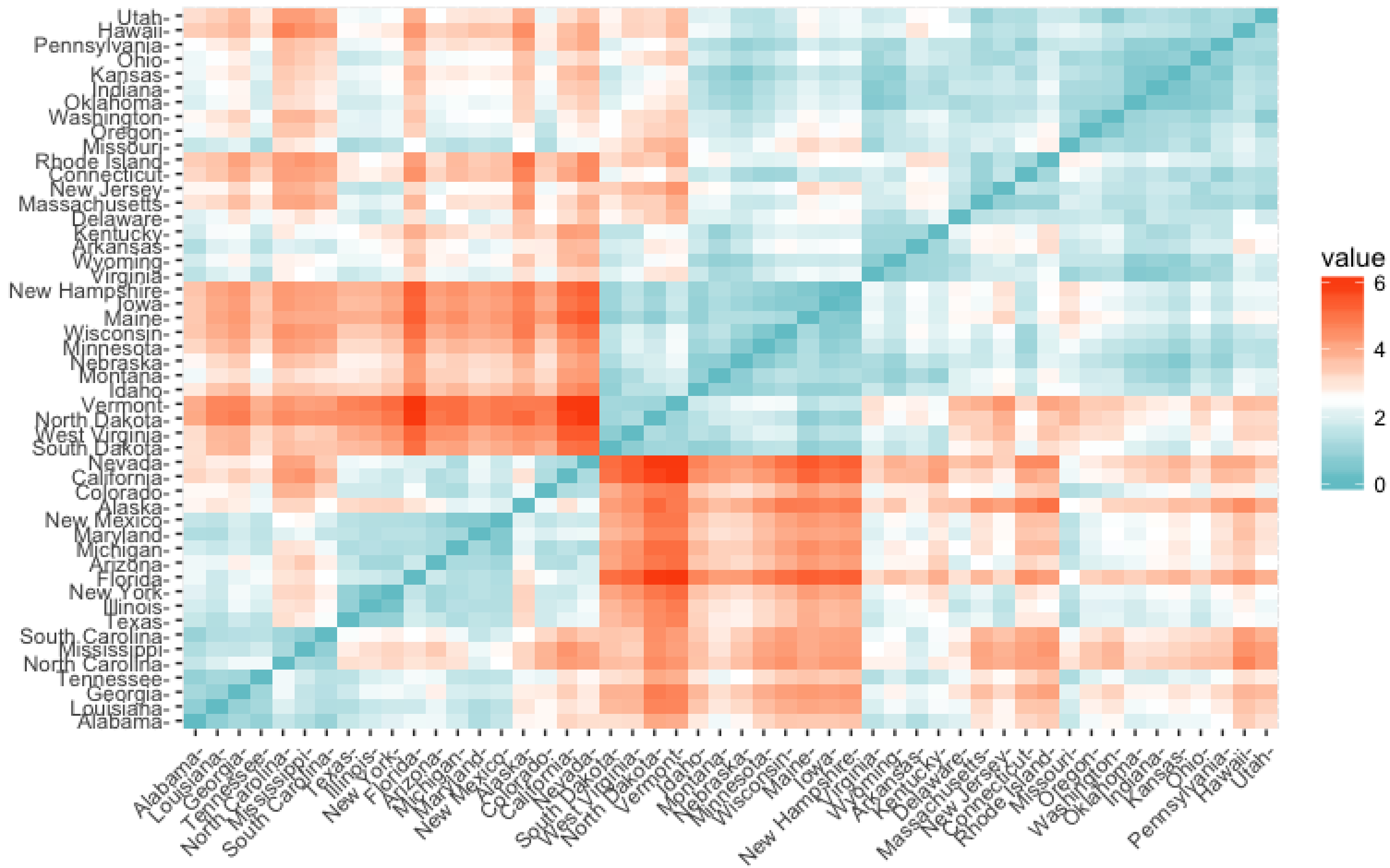
Dendrogram





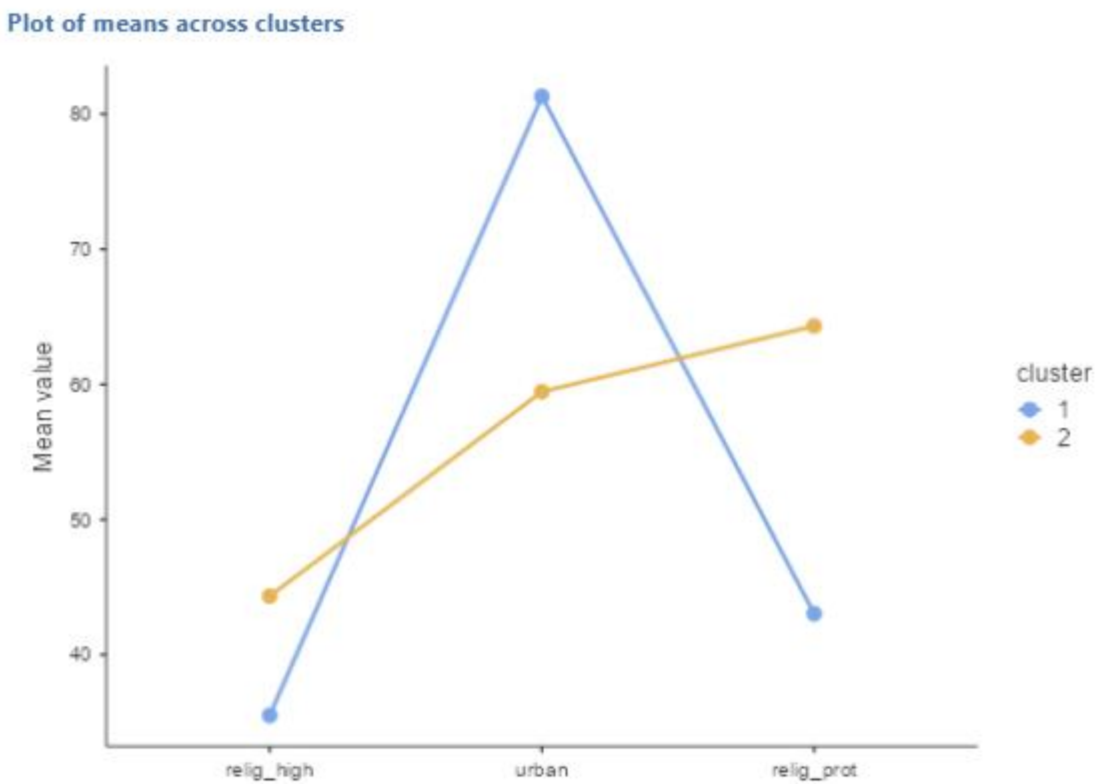
[Pai 2021](#)

Heatmap



[UC Business 2024]

Cluster profile



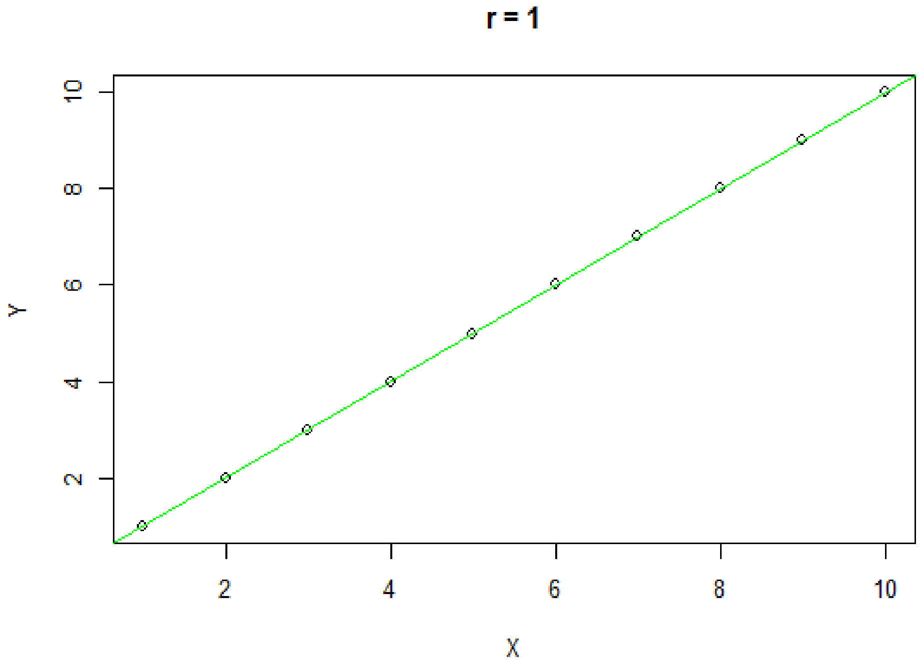
2.1 Pearson's coefficient

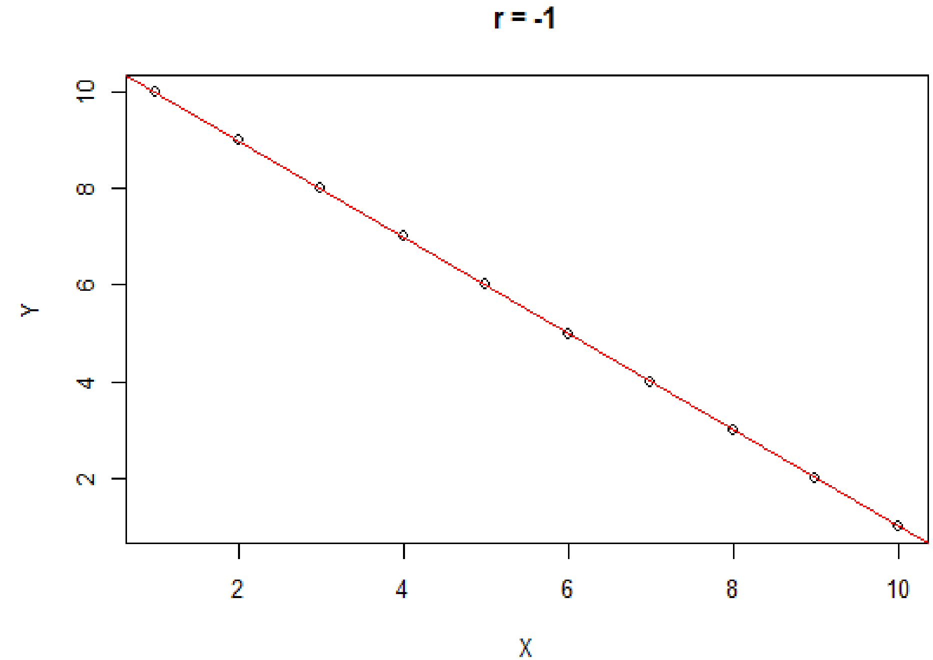
- Pearson's r measures the existence, strength and direction of the **linear relationship** between two variables

measurement level	number of values	range	coefficient
continuous-continuous	many-many	<-1, 1>	Pearson's r

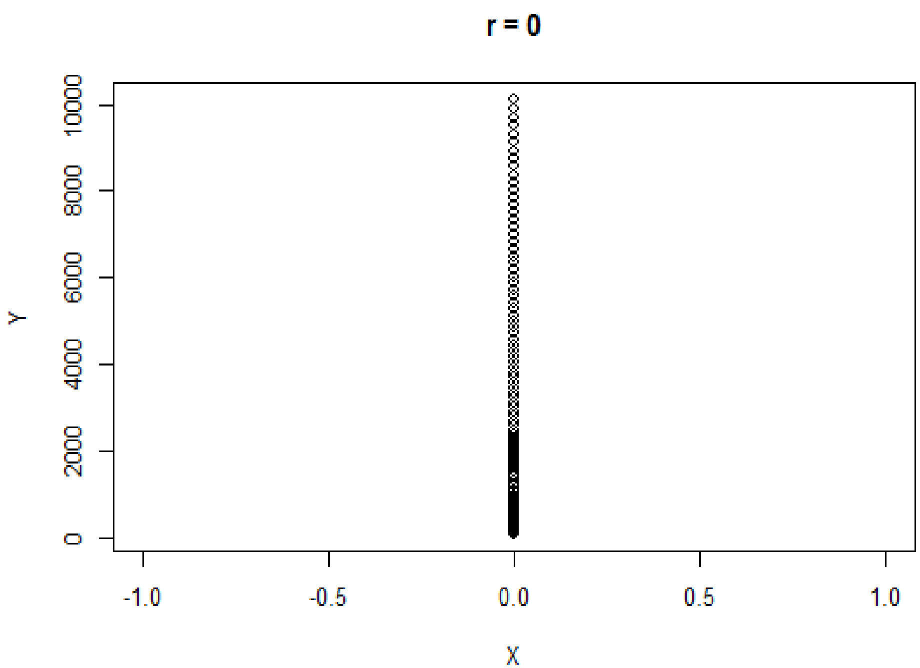
Soukup et al. 2022

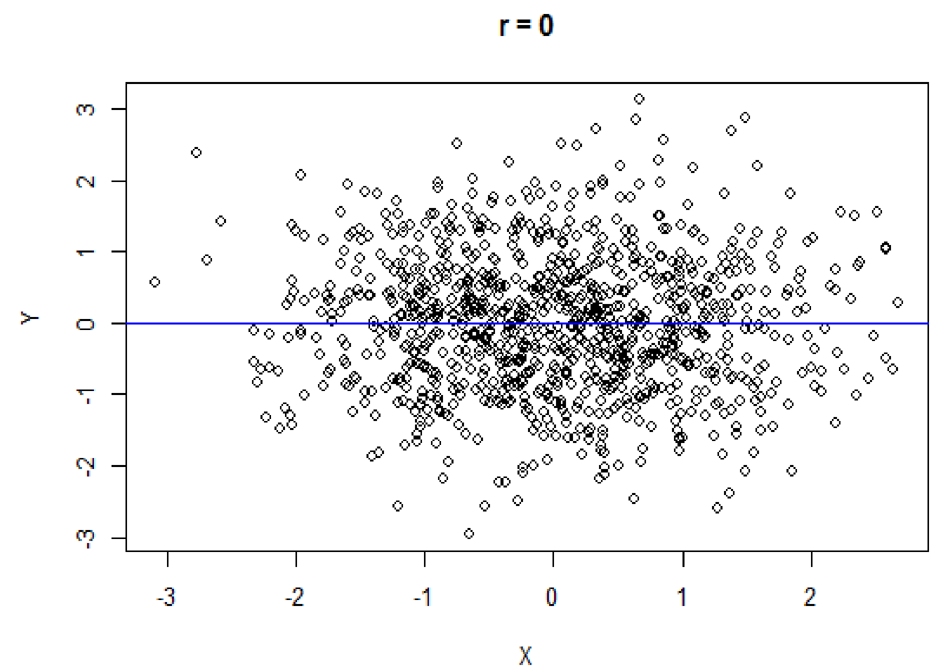
- Not suitable for heterogeneous data and/or nonlinear relationships between variables



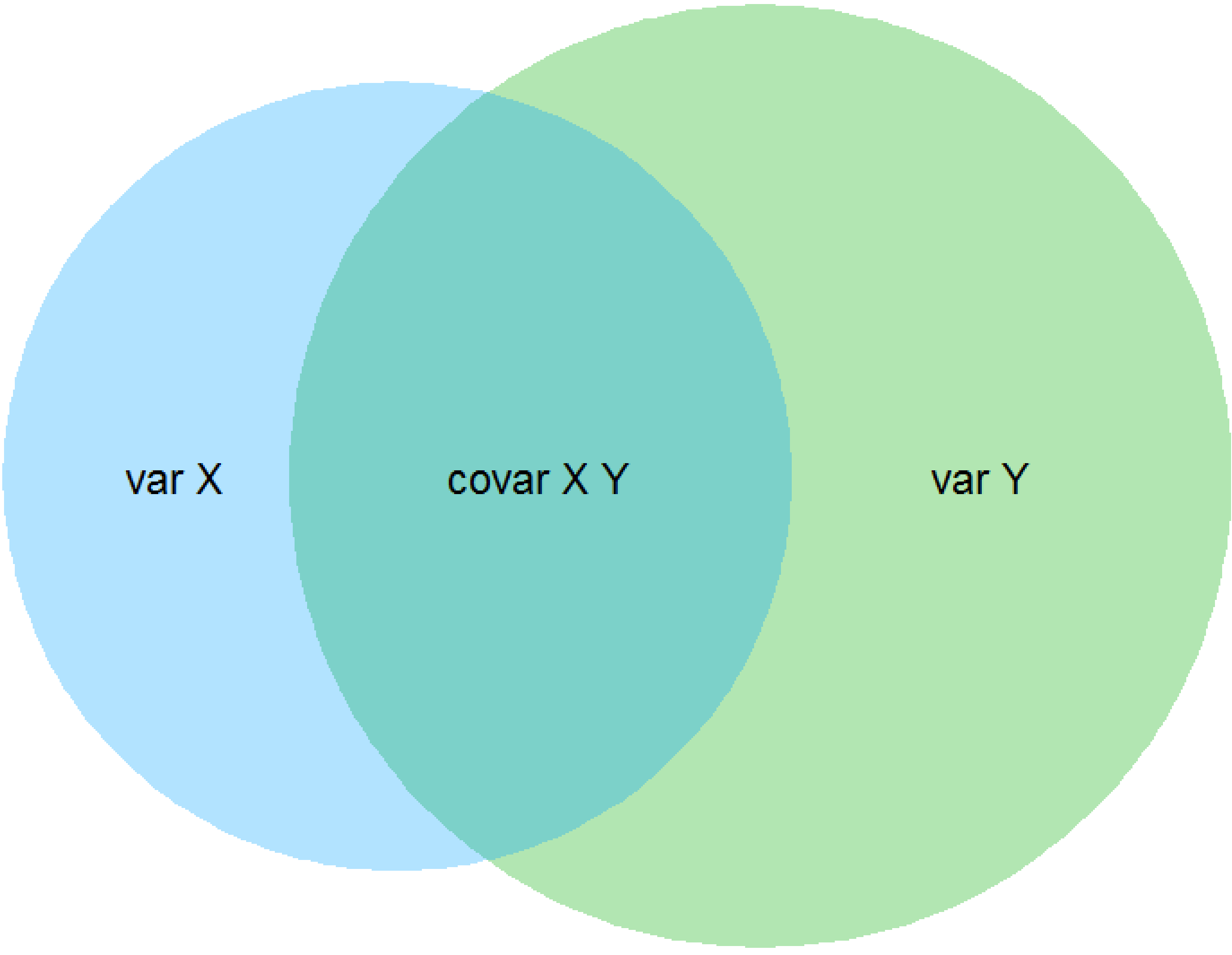


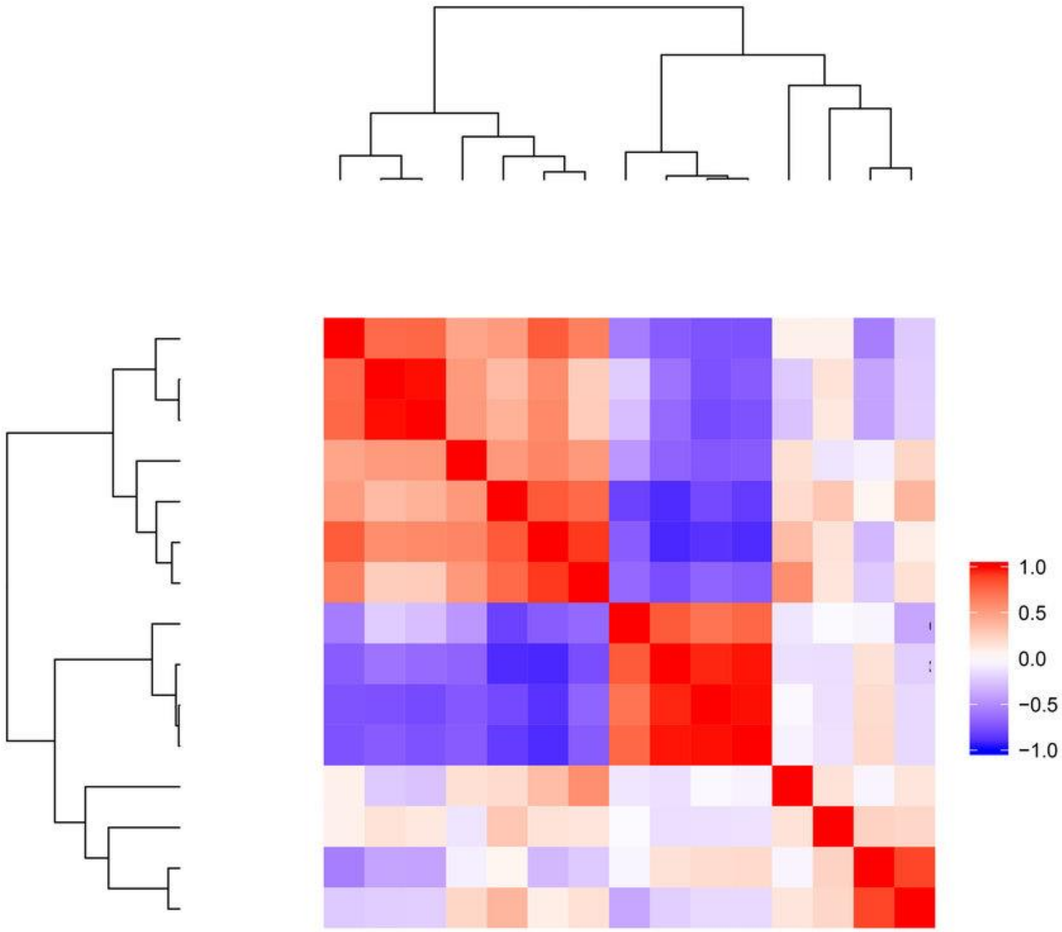
http://guessthecorrelation.com/





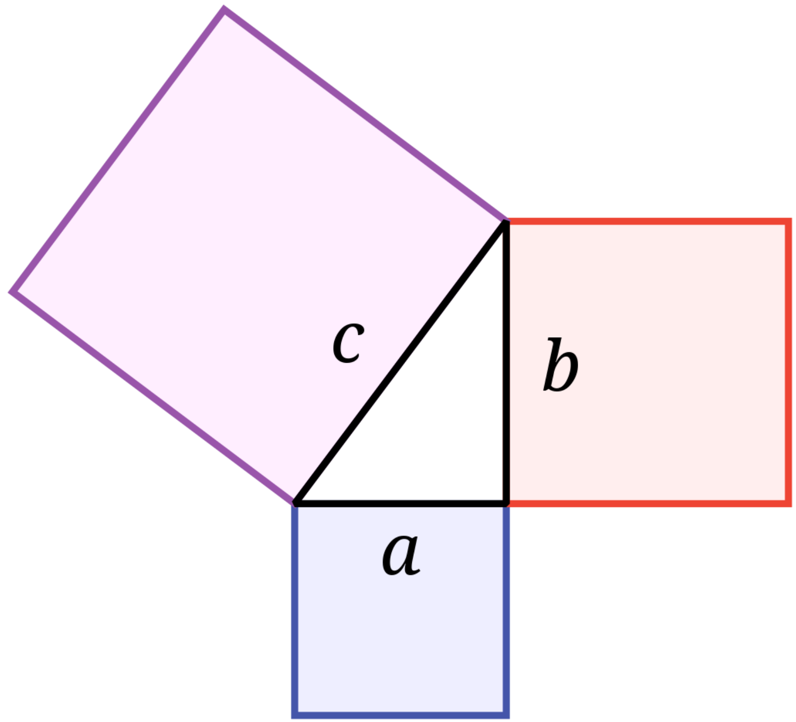
- r = covariance / combined total variance




1.0
0.5
0.0
−0.5
−1.0

2.2 Euclidean distance

- Do you recall this?

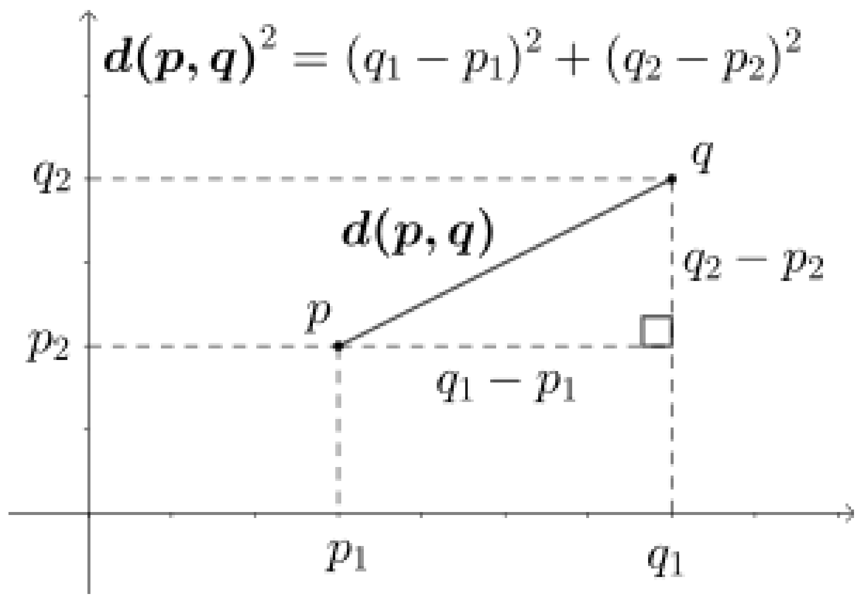


$$a^2 + b^2 = c^2$$

wikimedia commons

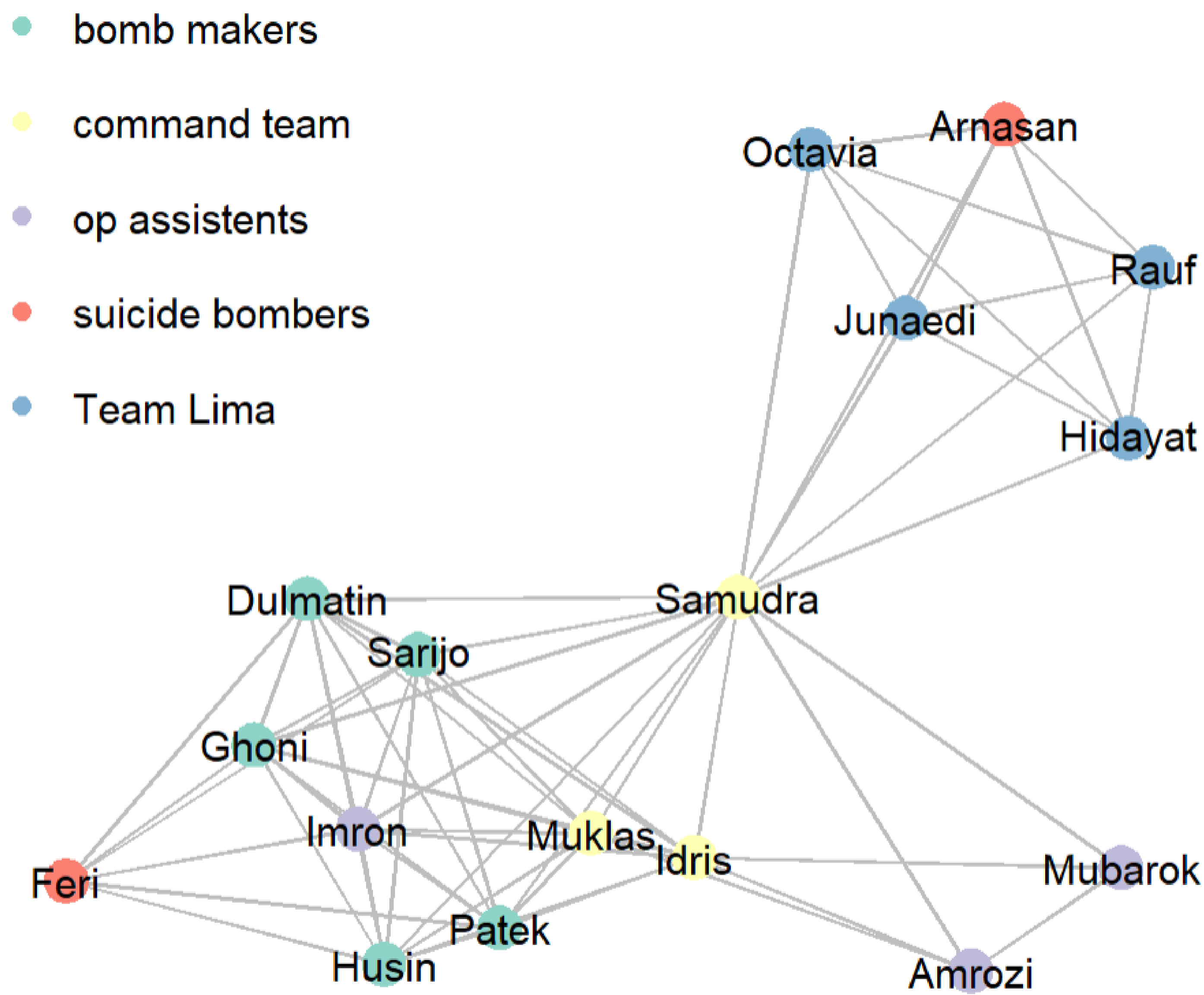
2.2 Euclidean distance

- **Euclidean distance (d)** is a generalization of the Pythagorean theorem
- d(p, q), of two objects (p, q) equals **the length of the straight line between** them



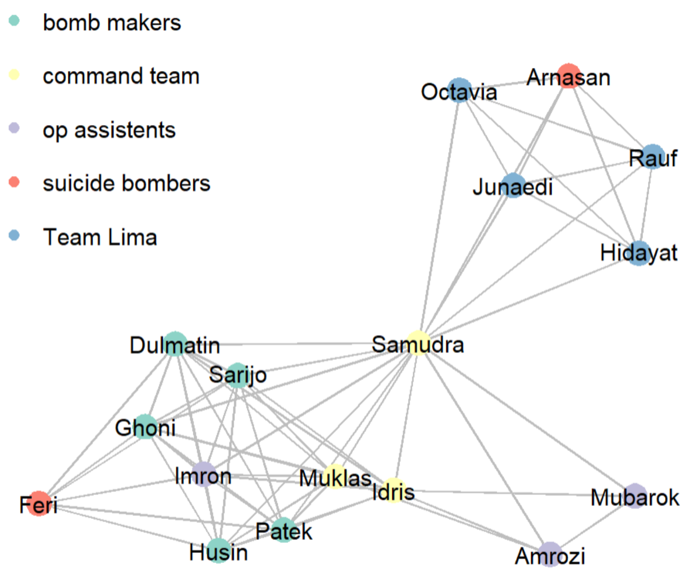
$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

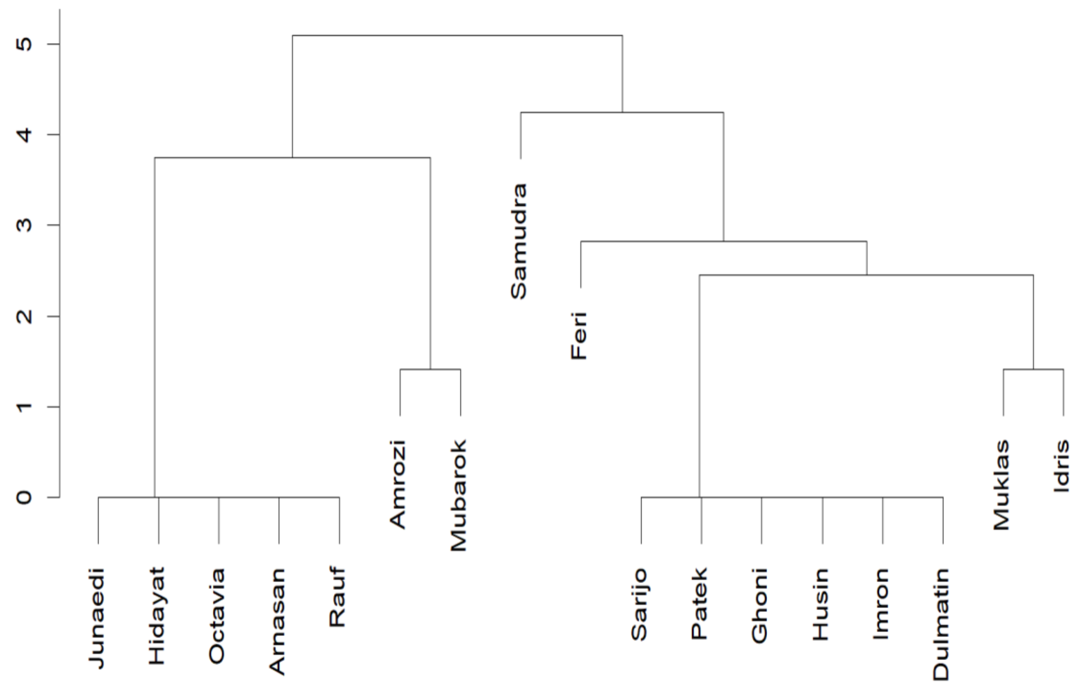
wikimedia commons



Koschade 2007

2.2 Euclidean distance

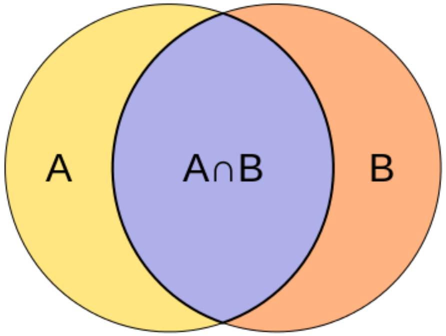


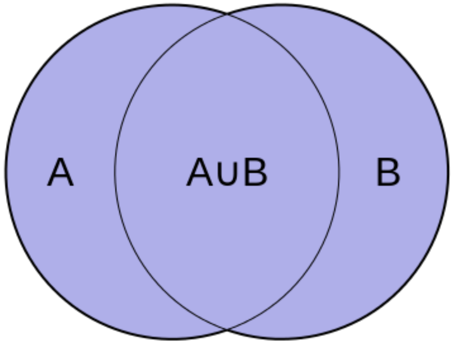


2.3 Jaccard's coefficient

measurement level	number of values	range	coefficient
categorical-categorical	binary	<0, 1>	Jaccard's

		sample B	
		present	absent
sample A	present	A ∩ B	A – B
	absent	B – A	∉ A ∪ B

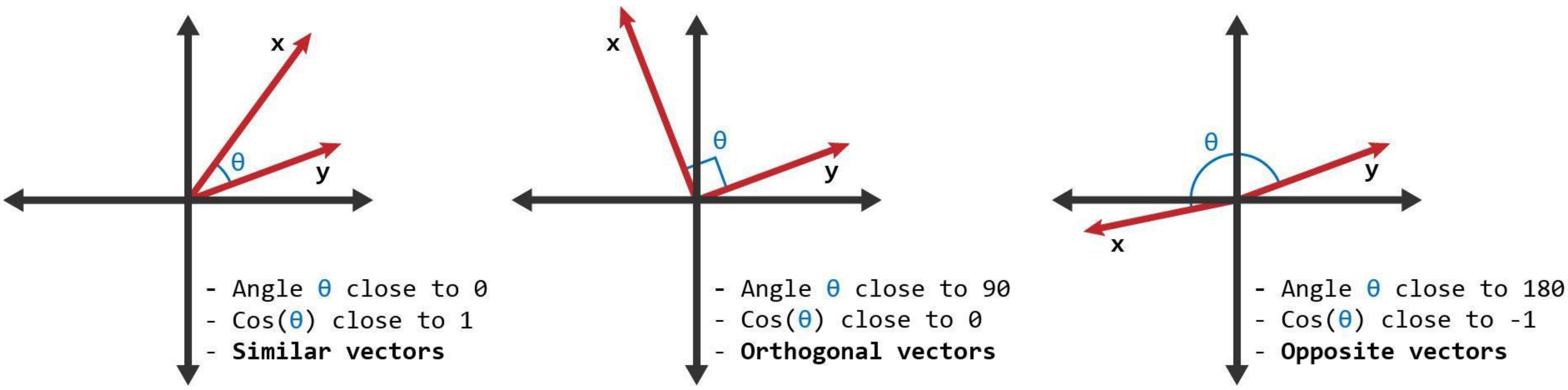




wikimedia commons

- J = the **size of the intersection** (A ∩ B) by the **size of the union** (A + B = A ∪ B) of the samples
- J = A ∩ B / (A ∪ B)
- Does not account for observations missing in both samples (∉ A ∪ B)

2.4 Cosine distance



[Karabiber 2024]

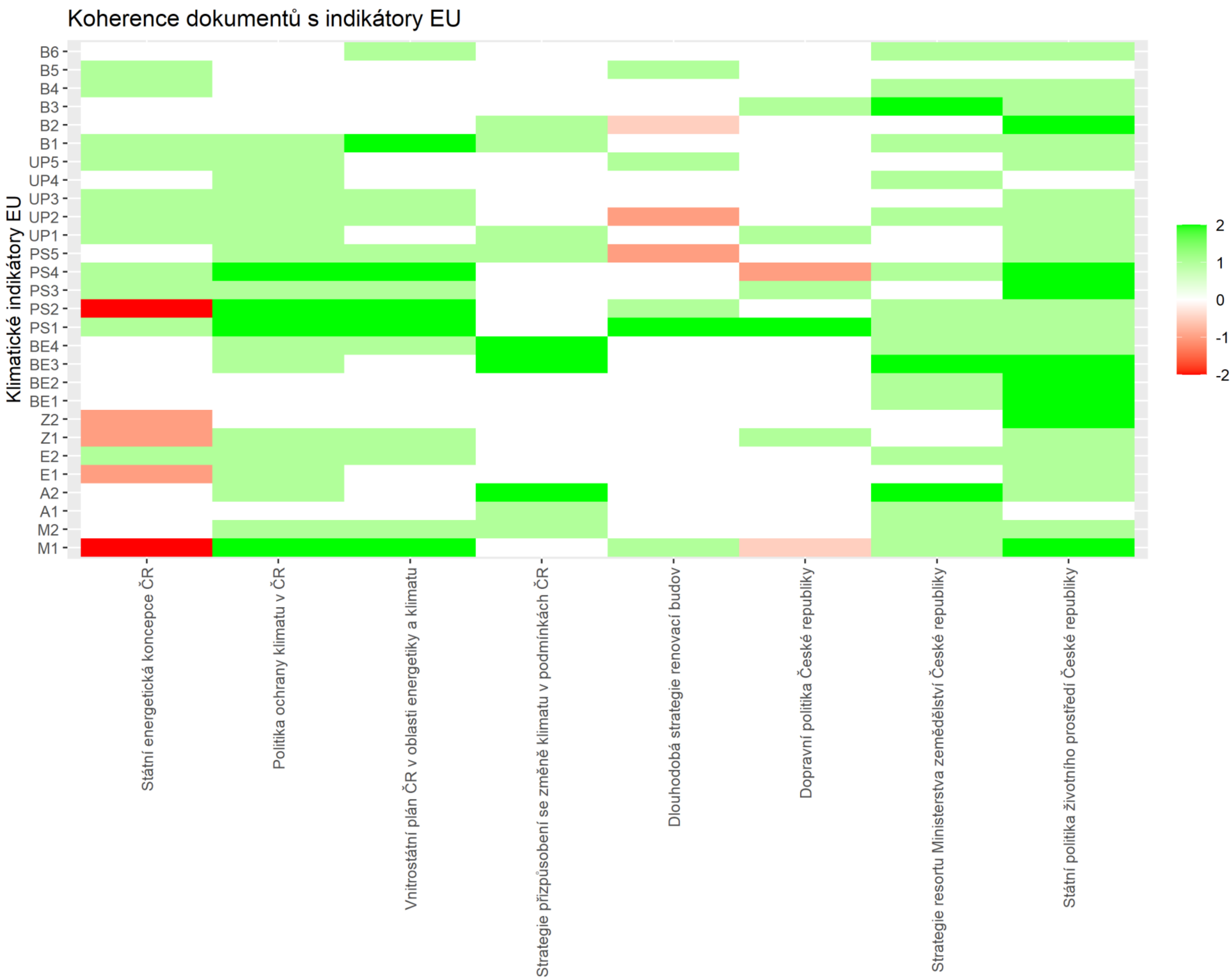
2.4 Cosine distance (CD)

	doc_A	doc_B
geopolitics	4	0
climate change	0	1
Brexit	0	12
Euro	3	9
sovereignty	5	2

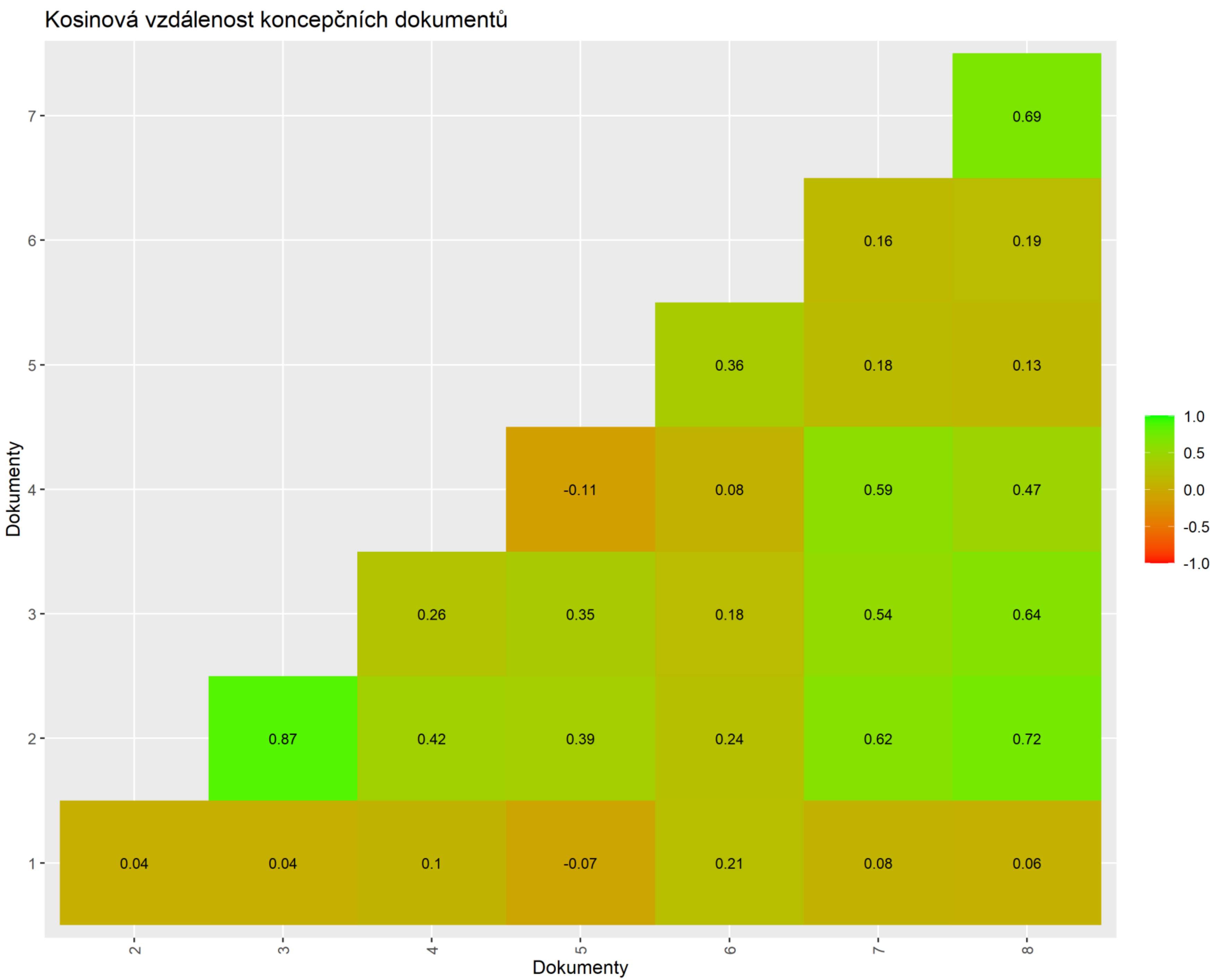
$$cosine\ similarity = |A||B|cos\theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_i^n A_i B_i}{\sqrt{\sum_i^n A_i^2}\sqrt{\sum_i^n B_i^2}}$$

- $\mathrm{CS} = \frac{(4*0 + 0*1 + 0*12 + 3*9 + 5*2)}{\sqrt{(16+0+0+9+2)} * \sqrt{(0+1+144+81+4)}} = 0.345$

- CD = 1 – 0.345 = 0.655



Koherence dokumentů s indikátory EU
Klimatické indikátory EU
B6
B5
B4
B3
B2
B1
UP5
UP4
UP3
UP2
UP1
PS5
PS4
PS3
PS2
PS1
BE4
BE3
BE2
BE1
Z2
Z1
E2
E1
A2
A1
M2
M1
Státní energetická koncepce ČR
Politika ochrany klimatu v ČR
Vnitrostátní plán ČR v oblasti energetiky a klimatu
Strategie přizpůsobení se změně klimatu v podmínkách ČR
Dlouhodobá strategie renovací budov
Dopravní politika České republiky
Strategie resortu Ministerstva zemědělství České republiky
Státní politika životního prostředí České republiky
2
1
0
-1
-2


Kosinová vzdálenost koncepčních dokumentů
Dokumenty
Dokumenty
1
2
3
4
5
6
7
2
3
4
5
6
7
8
0.69
0.16
0.19
0.36
0.18
0.13
-0.11
0.08
0.59
0.47
0.26
0.35
0.18
0.54
0.64
0.87
0.42
0.39
0.24
0.62
0.72
0.04
0.04
0.1
-0.07
0.21
0.08
0.06
1.0
0.5
0.0
-0.5
-1.0

2. Similarity measures: summary

- Cosine and Euclidean distances appropriate for higher-dimensional data
- Cosine distance used for text-based data
- Euclidean distances and Jaccard coefficient used for network data
- Pearson coefficient appropriate for continuous data and linear relationships
- There are many more measures of similarity

Cluster analysis: process

1. Sampling
2. Similarity measure
3. **Clustering method**
4. Cluster solution interpretation
5. Cluster solution diagnosis

3. Clustering methods

- After the (dis)similarities between objects are calculated, we need to select a particular **clustering method** that **partitions** (clusters) **the data** according specific rules

- There are several clustering approaches, **k-means clustering** and **hierarchical clustering** belong to the most common

3.1 k-means clustering

- **k-means clustering** partitions the data into *k* clusters based on the mean distances in each cluster

1. The number of clusters *k* needs to be **pre-selected**

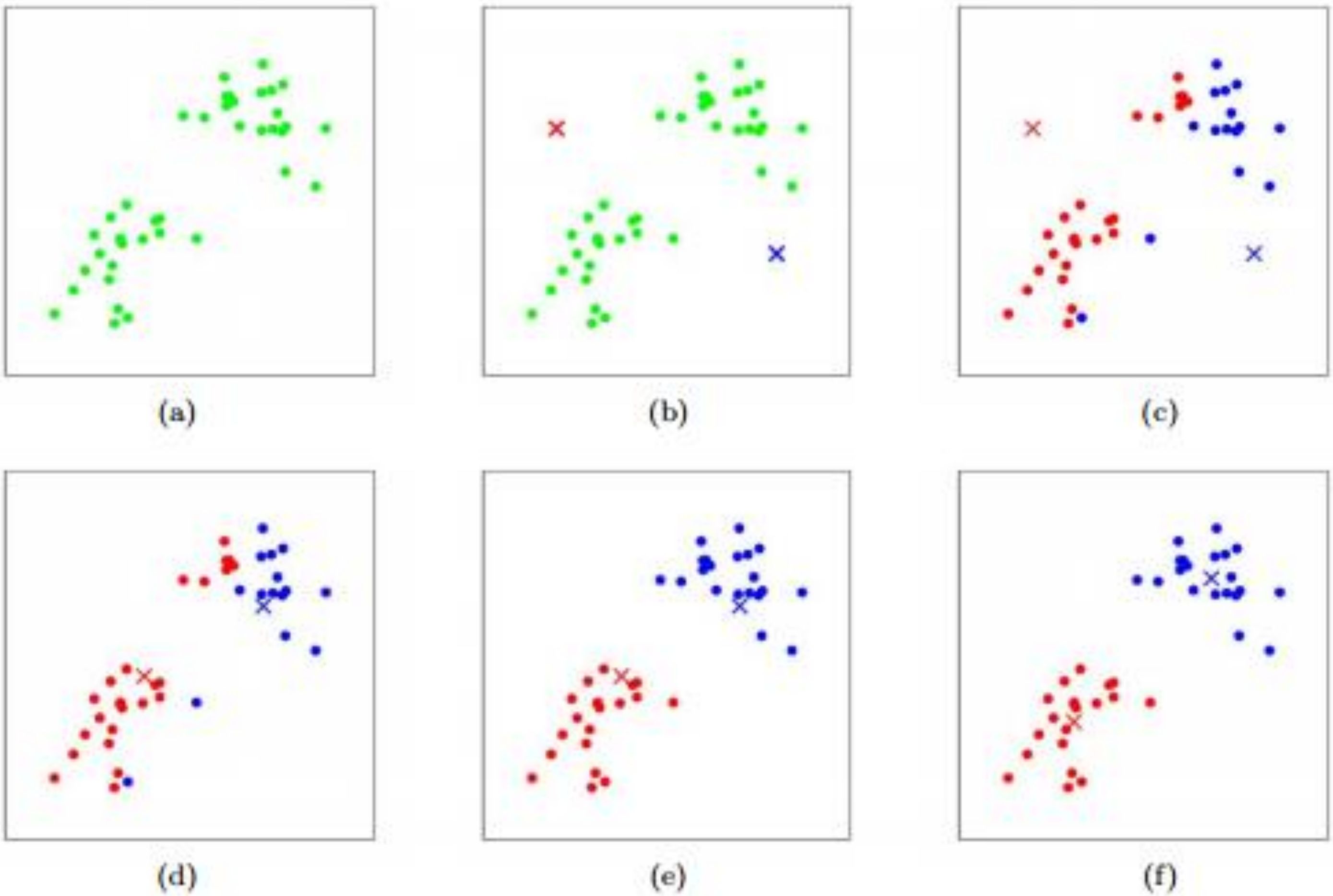
2. The algorithm starts randomly assign cluster centers (**centroids**)

3. Each object is assigned to the nearest centroid based on a particular similarity measure

4. Within the clusters, **the centroids are updated** based on the mean similarity of the objects classified to the respective cluster

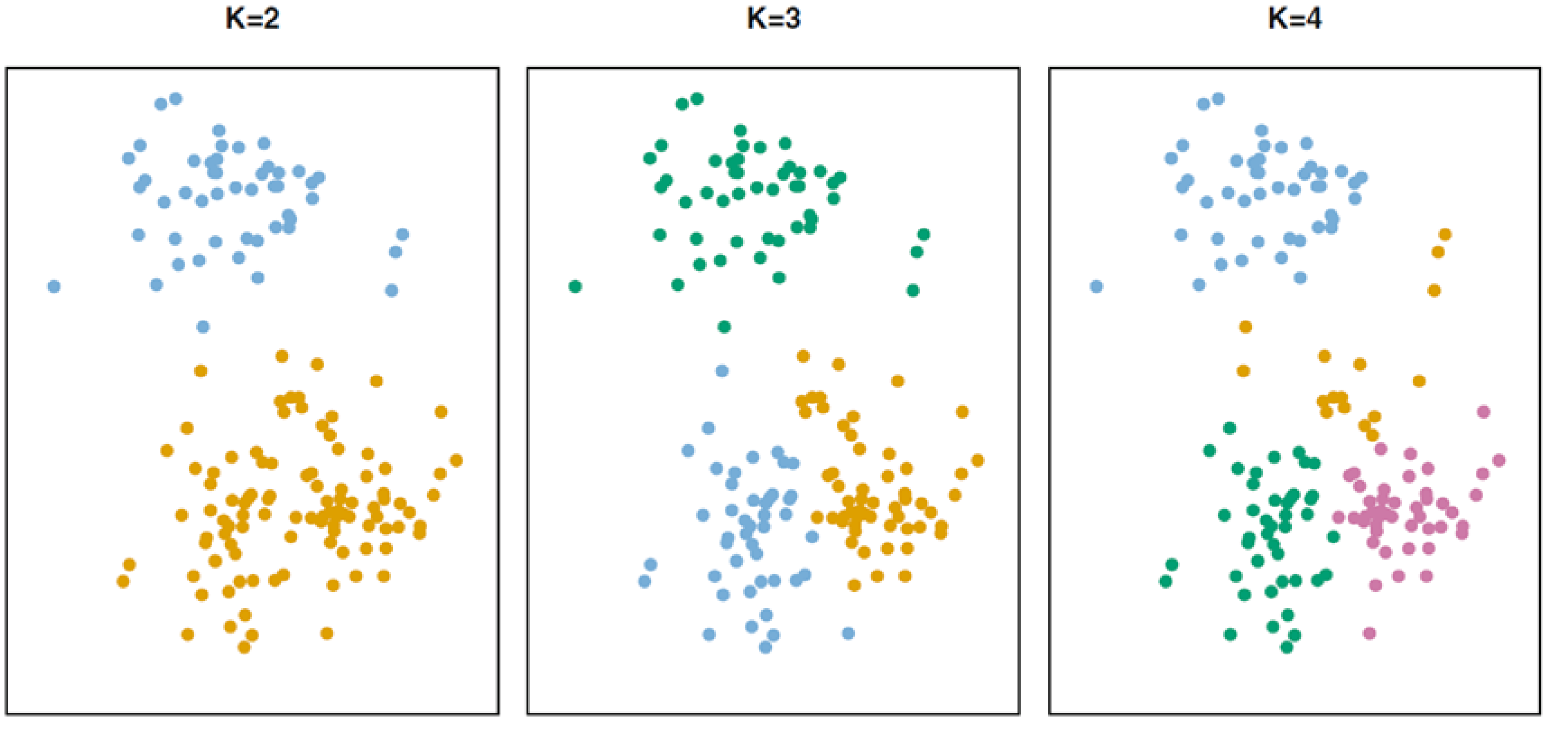
5. The steps 3-4 repeat until the centroids no longer change → solution

2-cluster solution



Jordan 2012

Determining *k*



A simulated data set with 150 observations in two-dimensional space. Each figure show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

James et al. 2013

3.1 Hierarchical clustering

- **Hierarchical clustering** can be performed in **agglomerative** and **divisive** sequence
- The number of clusters is **not pre-selected**
- The **linkage method** (how are clusters merged/split) needs to be defined

1. Treat each objects as a separate cluster (**agglomerative** sequence)

2. The average distances between the clusters are calculated (**average linkage**)

3. The clusters with the lowest average distance are merged

4. The steps 2-3 are repeated until there is only a single cluster

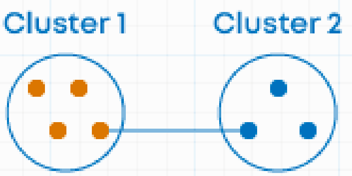
5. The process is represented by **dendrogram**

6. Considering substantive insights, the *k*-cluster solution is identified

- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

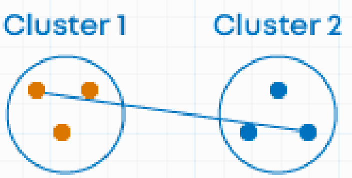
  Minimum distance or distance between closest elements in clusters



- **Complete Linkage**

  $D(c_1,c_2) = \max D(x_1,x_2)$

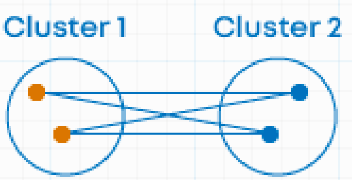
  Maximum distance between elements in clusters



- **Average Linkage**

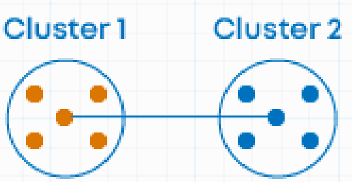
  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum\sum D(x_1,x_2)$

  Average of the distances of all pairs

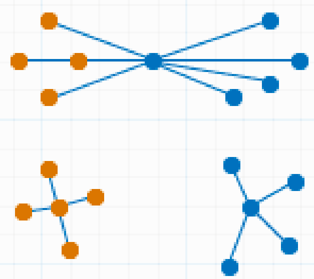


- **Centroid Method**

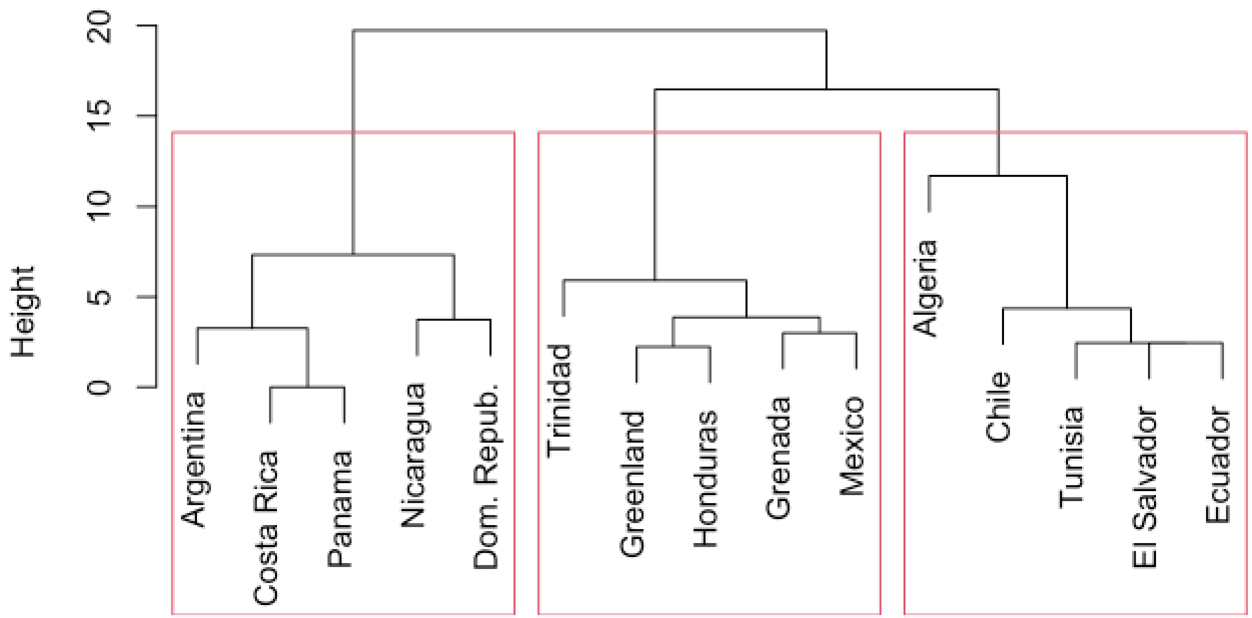
  Combining clusters with minimum distance between the centroids of the two clusters



- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.
  - Objective is to minimize the total within cluster vairance



[NYU 2011]


Height
0
5
10
15
20
Argentina
Costa Rica
Panama
Nicaragua
Dom. Repub.
Trinidad
Greenland
Honduras
Grenada
Mexico
Algeria
Chile
Tunisia
El Salvador
Ecuador

NYU 2011

Cluster analysis: process

1. Sampling
2. Similarity measure
3. Clustering method
4. **Cluster solution interpretation**
5. **Cluster solution diagnosis**

4. Cluster solution interpretation

- Do the clusters have **prima facie** validity (**eyeballing** test)?
- **Substantive** and **theoretical insights** are vital – to what extent the solution aligns with our expectations?
- **Size of the clusters** – do clusters markedly differ in their size?
- **Outliers** – are there any?
- Do clusters **reduce our data well**? → cluster **solution diagnostics**

5. Cluster solution diagnostics

- Important to assess the **quality of our solution**
- Are the clusters internally **cohesive**?
- Are the clusters well **separated**?
- What is the optimal **number of clusters**?

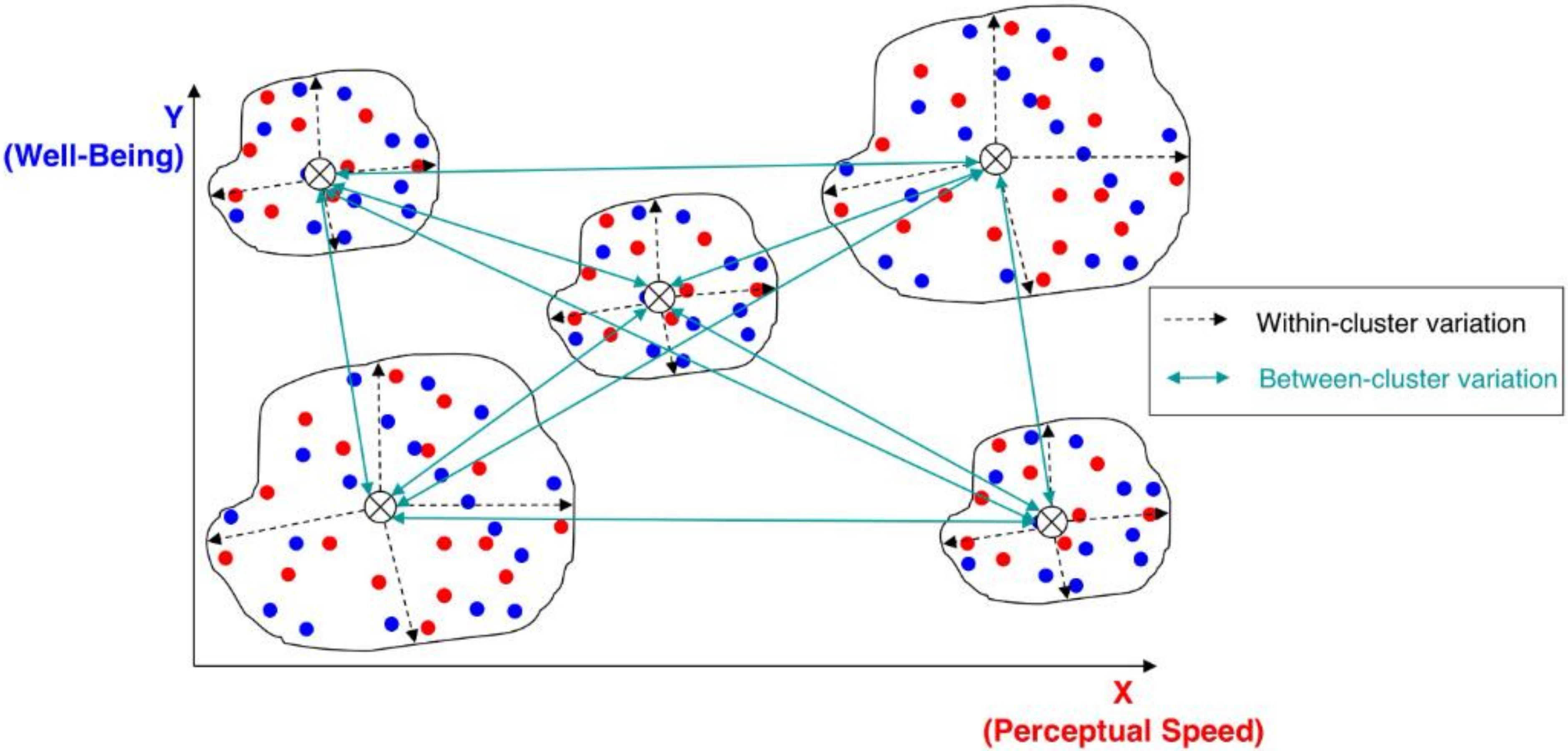
5.1 Within/between sum of squares

- The **within-cluster sum of squares (WCSS**) calculate the sum of squared distances between each object and its cluster centroid (k-means clustering approach)

- The **between-cluster sum of squares** (**BCSS**) measure the sum of squared distance between centroid of each cluster and the overall centroid of all objects

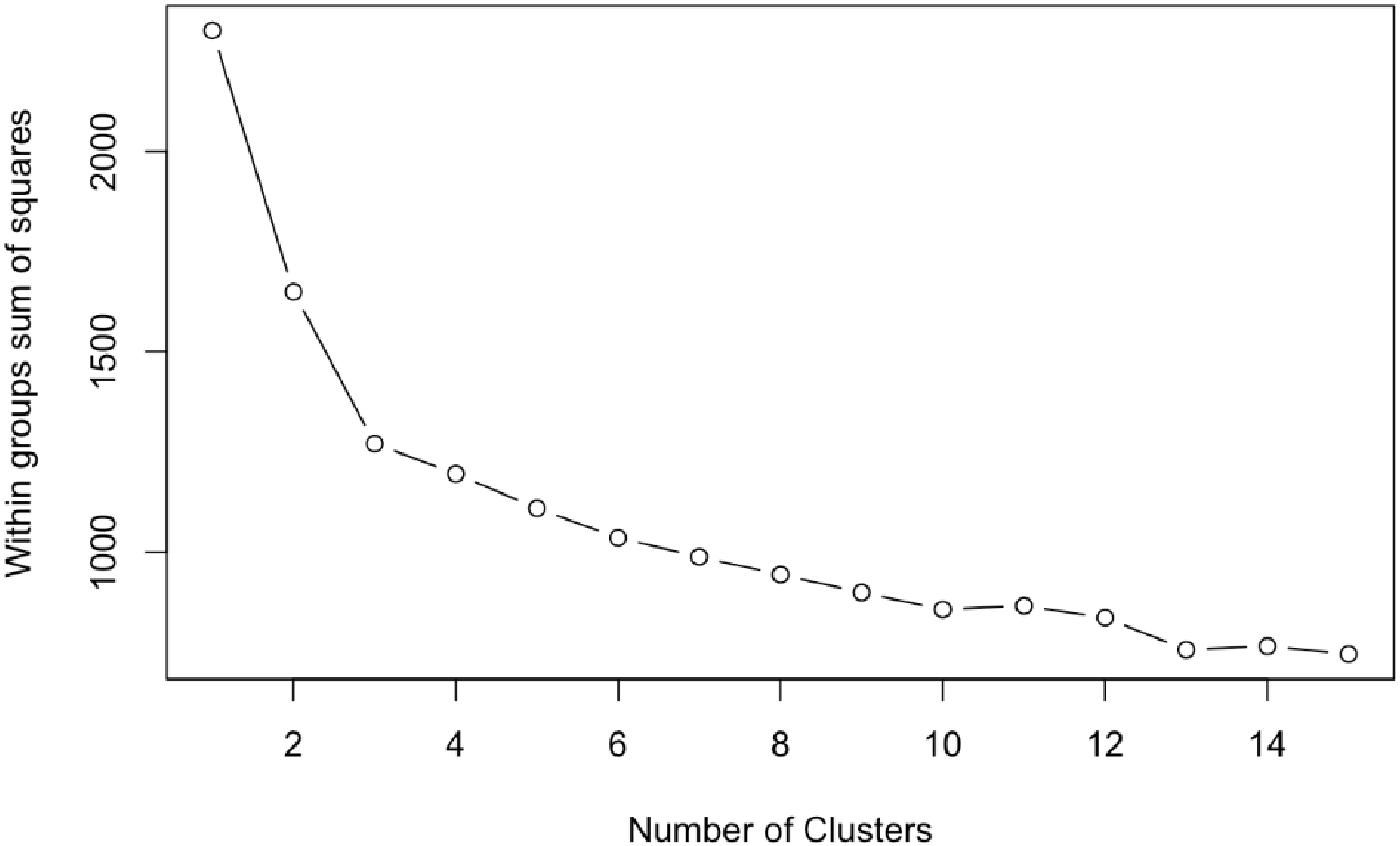
- The **WCSS** capture the **cohesiveness** of clusters, while **BCSS** measure the **separation between clusters**

5.1 WCSS and BCSS



SSRI 2020

Elbow graph



[Kulma 2017]

5.2 Silhouette score

- The **Silhouette score** ranges **<-1, 1>**; where high positive values indicate a good fit of the object within its own cluster, zero indicates a borderline position, and negative values indicate that the object is misclassified

- $\mathrm{s} = \frac{b-a}{\max(b-a)}$; *a* = average within-cluster d, *b* = average between-cluster d

- The Silhouette is calculated for each object and then average is taken to evaluate the cluster solution

- The Silhouette score can be applied both to the **k-means** and **hierarchical clustering**

- Rule of thumb: **s > 0.5** good solution; s < 0.25 bad solution

Exercise

Exercise

- Open Jamovi and install **snowCluster** extension
- Load into Jamovi file “state.csv”
- Check variables relig_prot, urban, and relig_high in the codebook
- Describe the variables
- Apply **(1) k-means** clustering and **(2) hierarchical** clustering
- Interpret the results