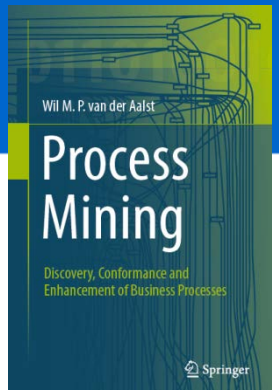
Process Mining: Data Science in Action

Cluster Analysis

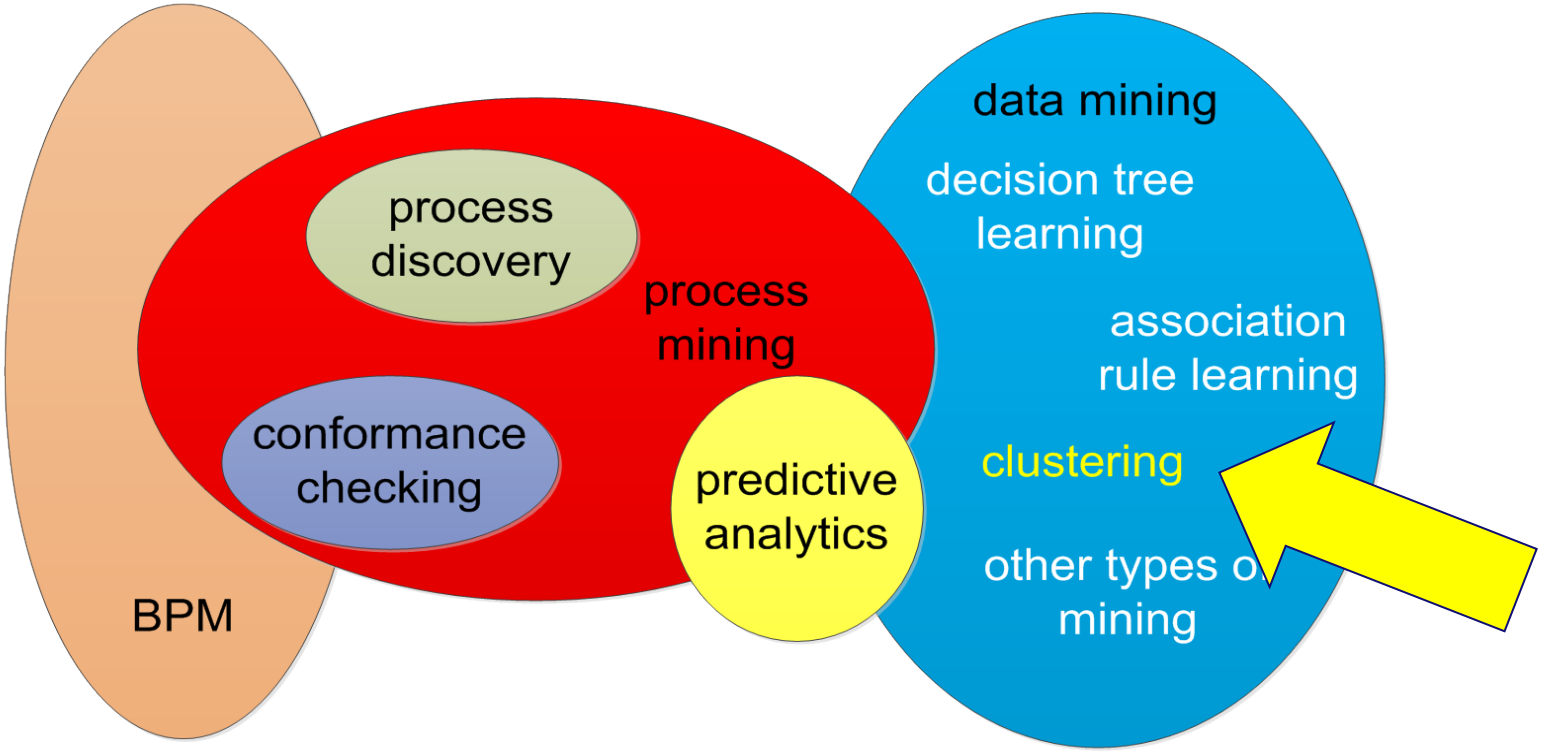


prof.dr.ir. Wil van der Aalst
www.processmining.org

Technische Universiteit Eindhoven University of Technology

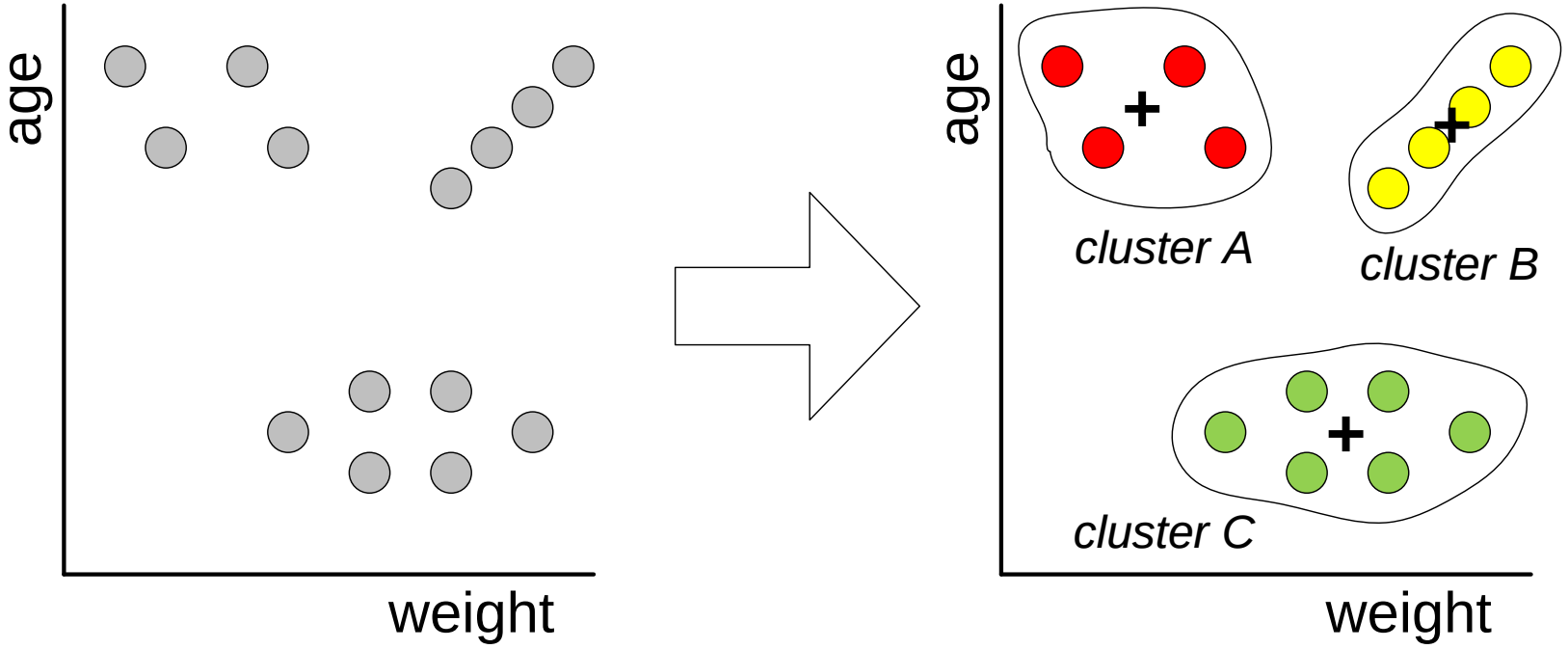
Where innovation starts

Clustering
(unsupervised learning: no response variable)

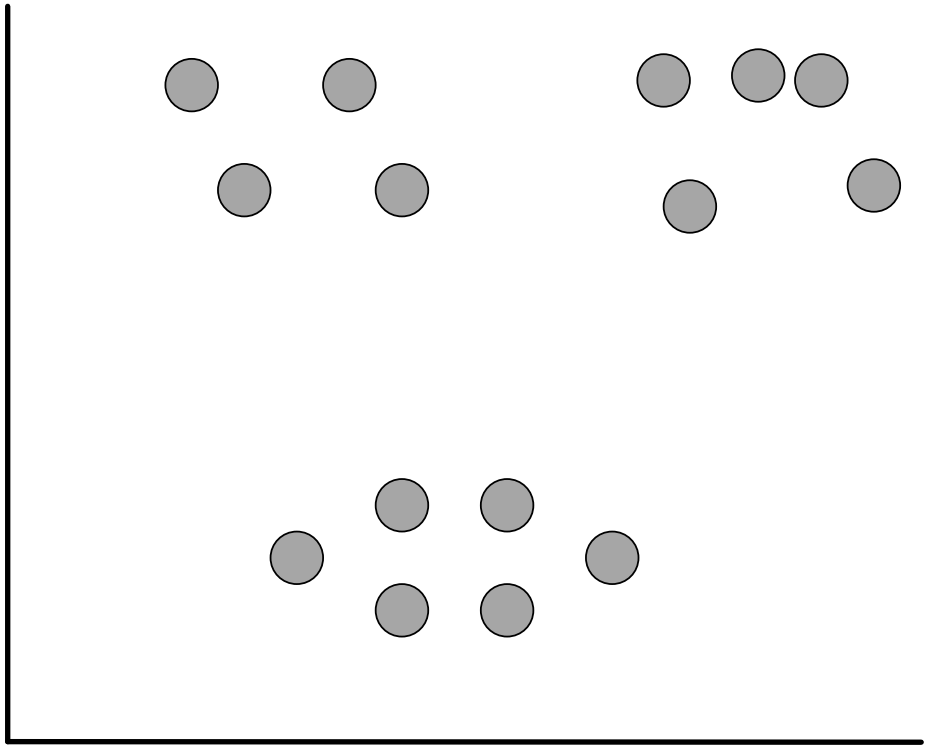


TU/e

Clustering



k-means clustering

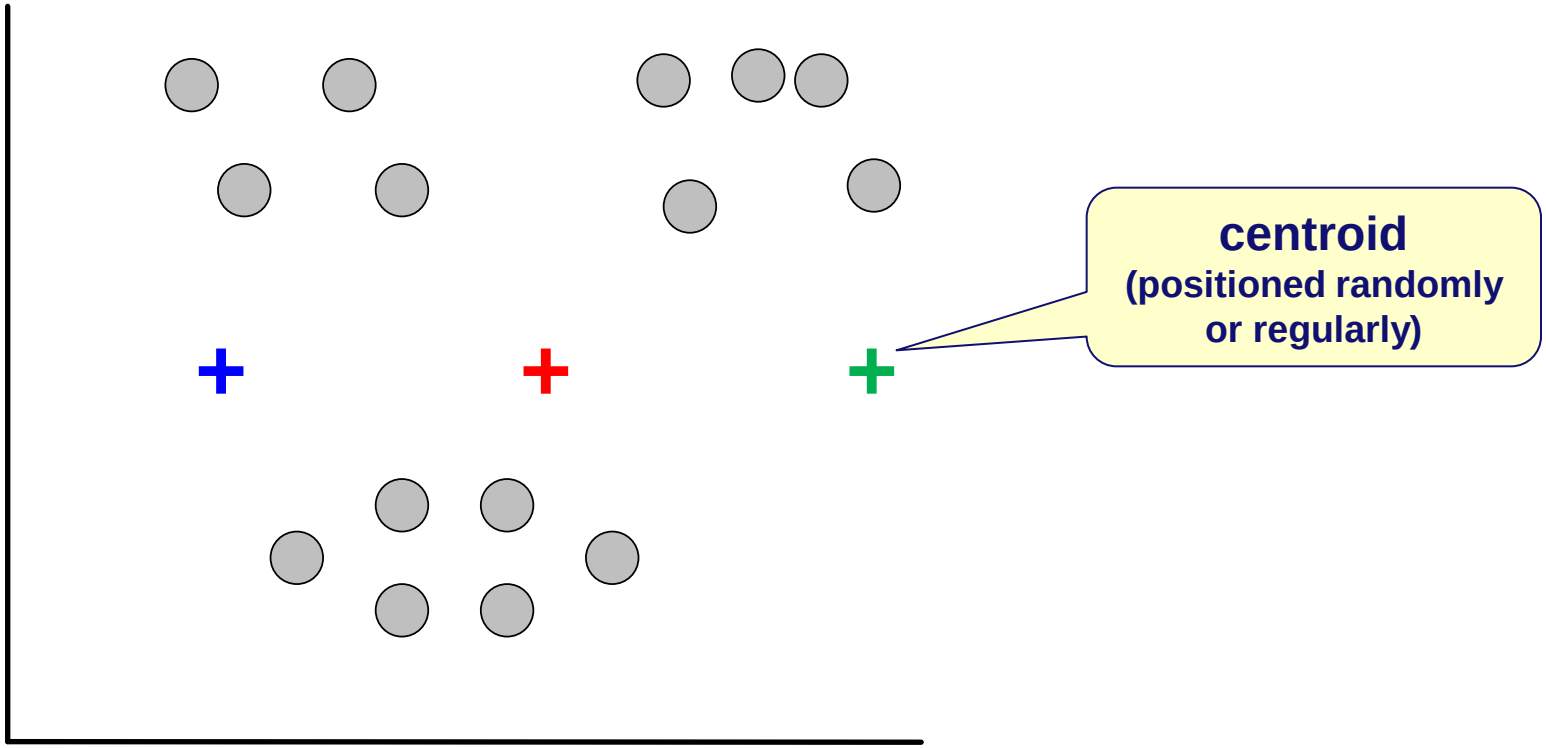


assume *k*=3

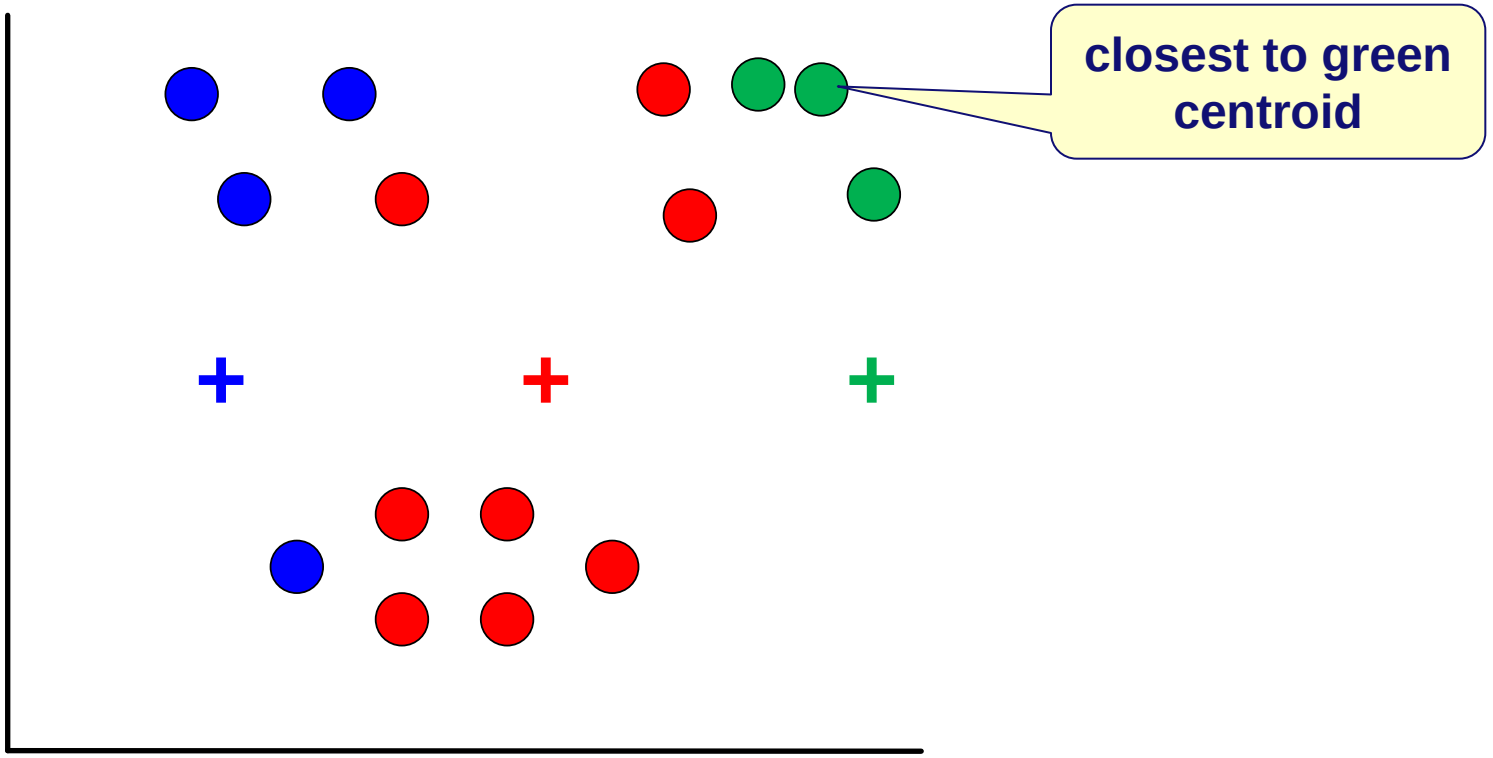
Just an illustration. May be misleading: often many dimensions!

TU/e

Generate *k*=3 centroids (e.g., random)

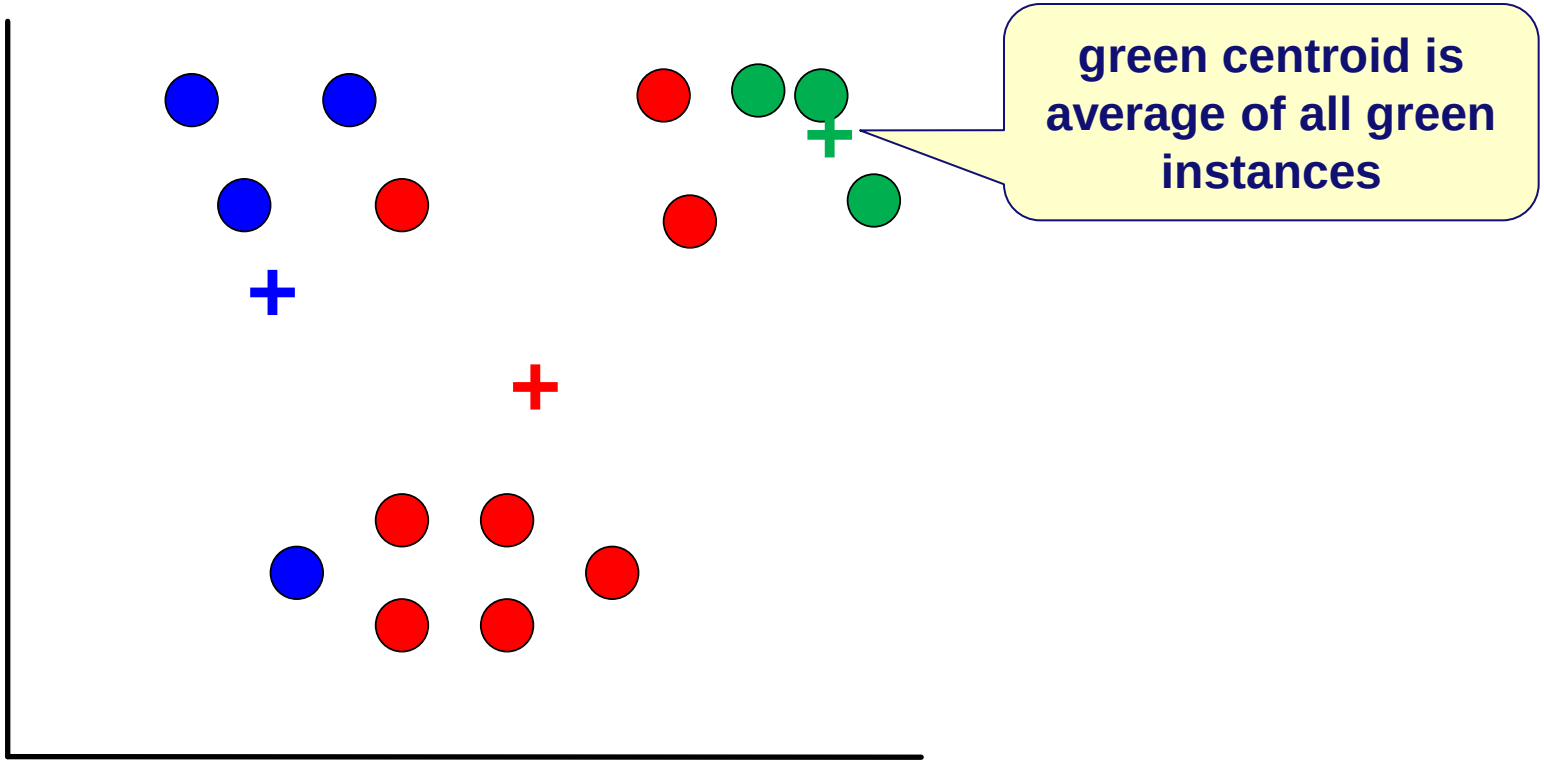


Assign instances to closest centroid



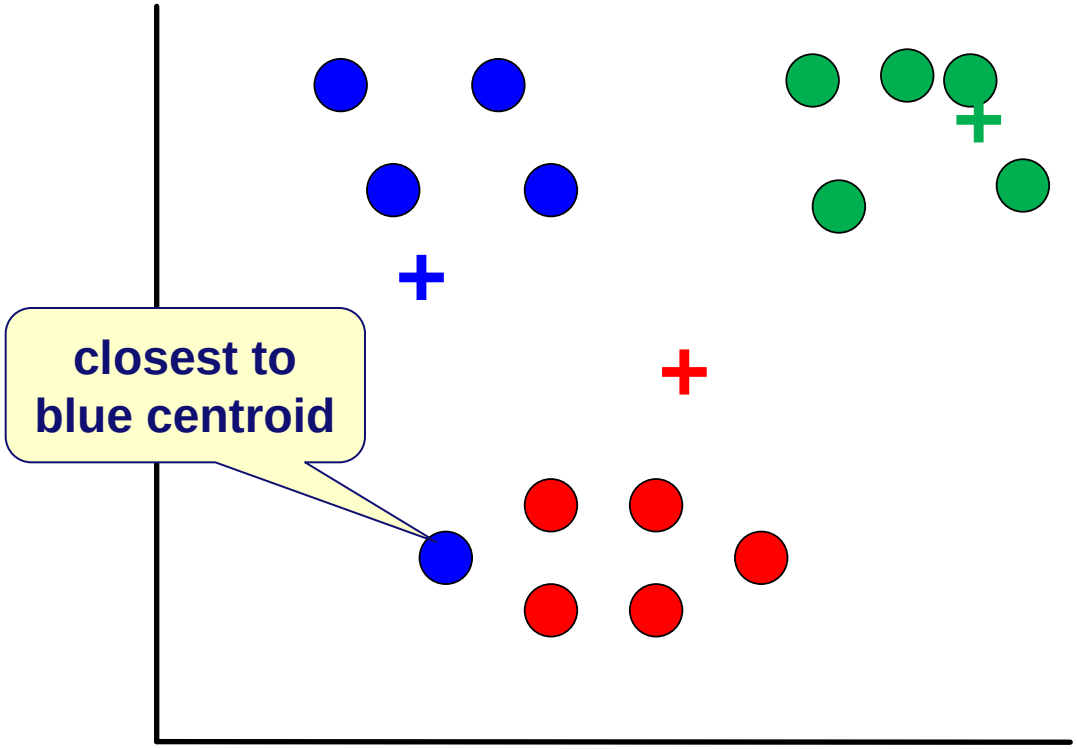
TU/e

Recompute centroids based on assigned instances



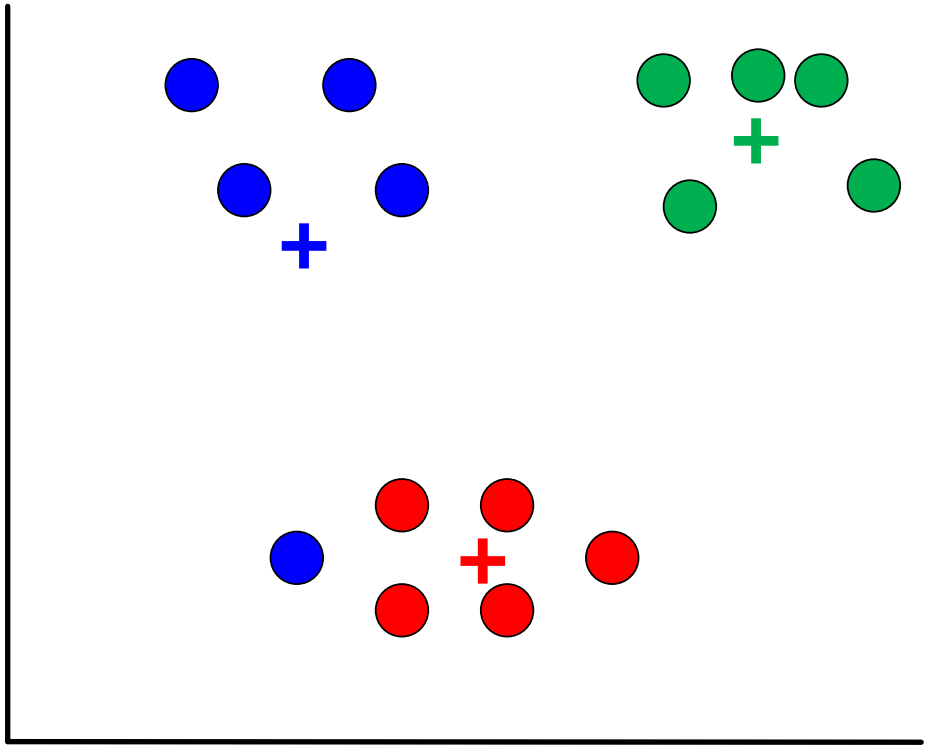
TU/e

Assign instances to closest centroid

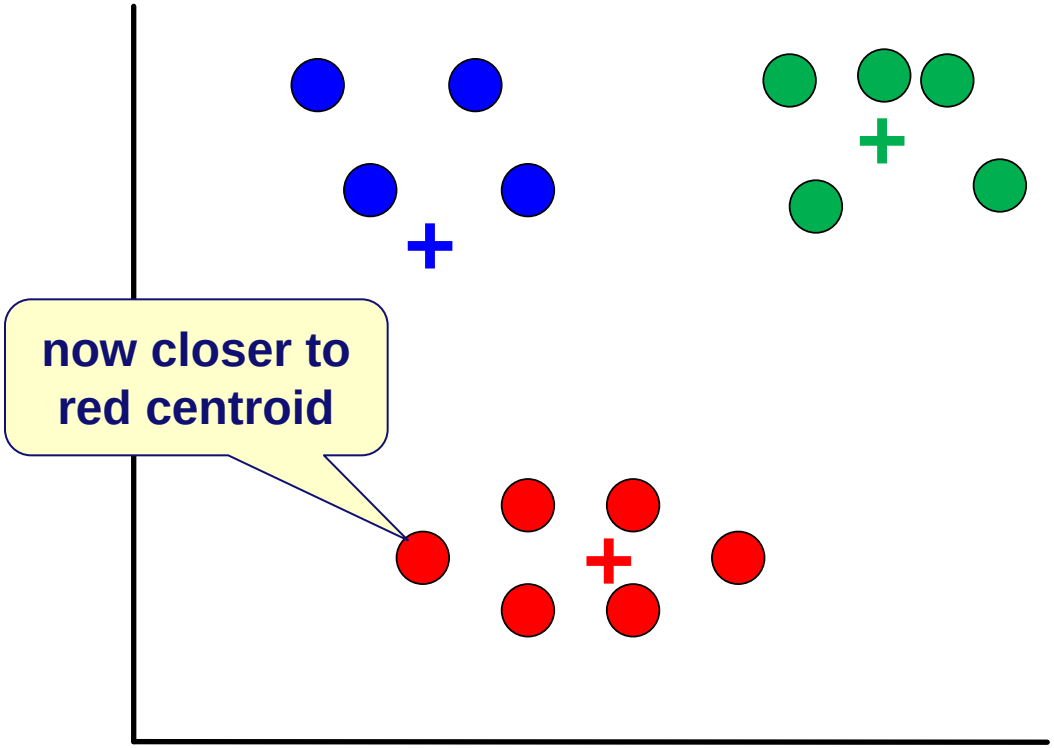


TU/e

Recompute centroids based on assigned instances

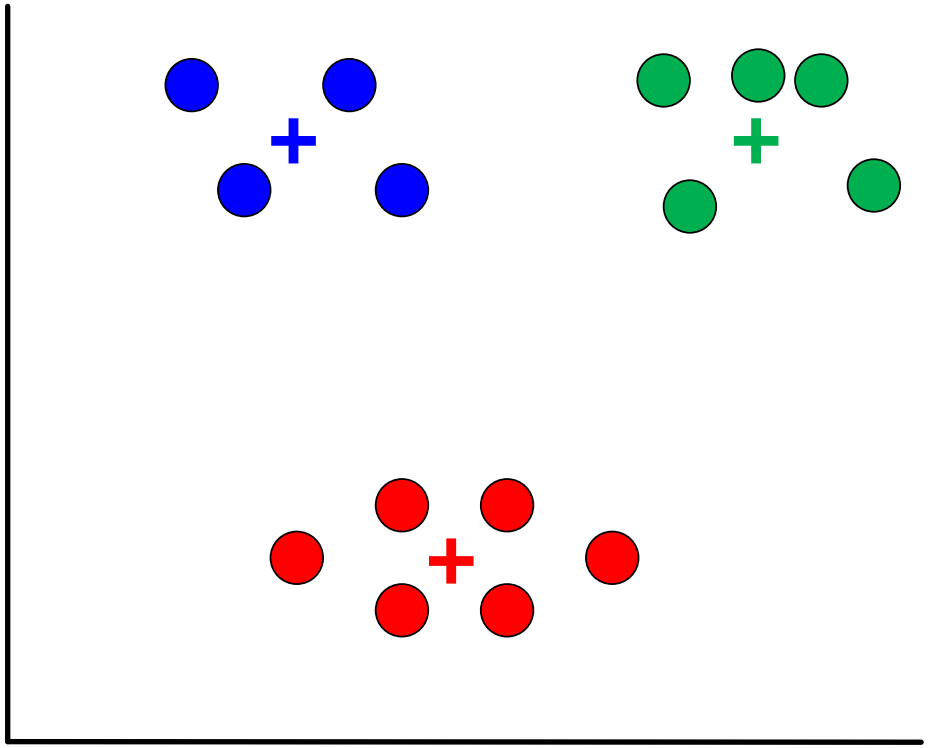


Assign instances to closest centroid



TU/e

Recompute centroids based on assigned instances

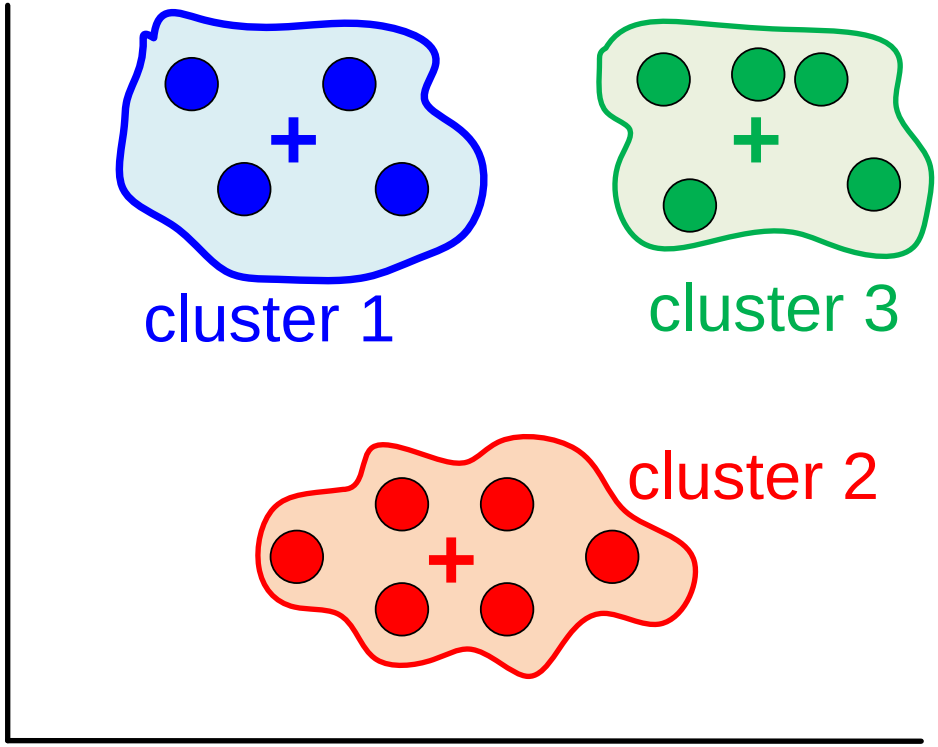


Fixpoint has been reached: Nothing changes anymore.

Due to non-deterministic nature (random initial centroids), experiment may be repeated multiple times. Select "best clustering" at end.

TU/e

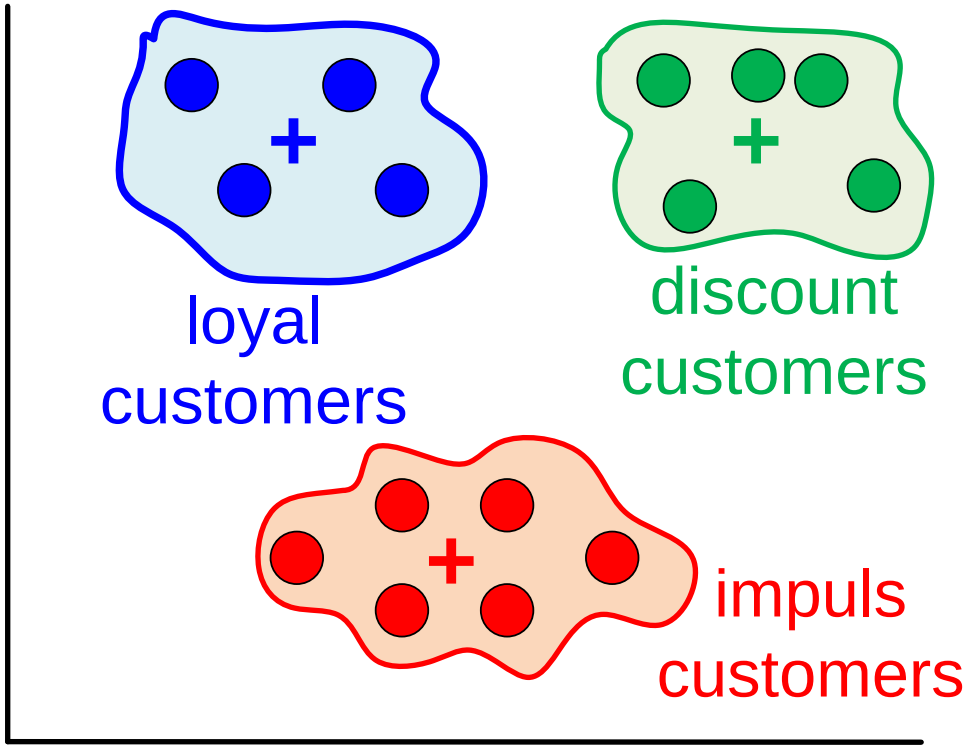
Clusters returned



Main idea:
The instances in a cluster are more similar to each other than to those in other clusters.

TU/e

Many use cases



For example: finding homogenous groups of customers, patients, sessions, students, etc.

After clustering: apply additional mining techniques on the partitioned input data.

TU/e

Simplified data set: only 6 items

- 5000 parties ate at the Italian restaurant.
- Menu is restricted to pizza margherita, pizza siciliana, lasagna, spaghetti carbonara, vino rosso, and birra.

<new process> – RapidMiner 5.3.015 @ nbwin1027

File Edit Process Tools View Help

Result Overview | ExampleSet (//Local Repository/data/food-poisoning-simple)

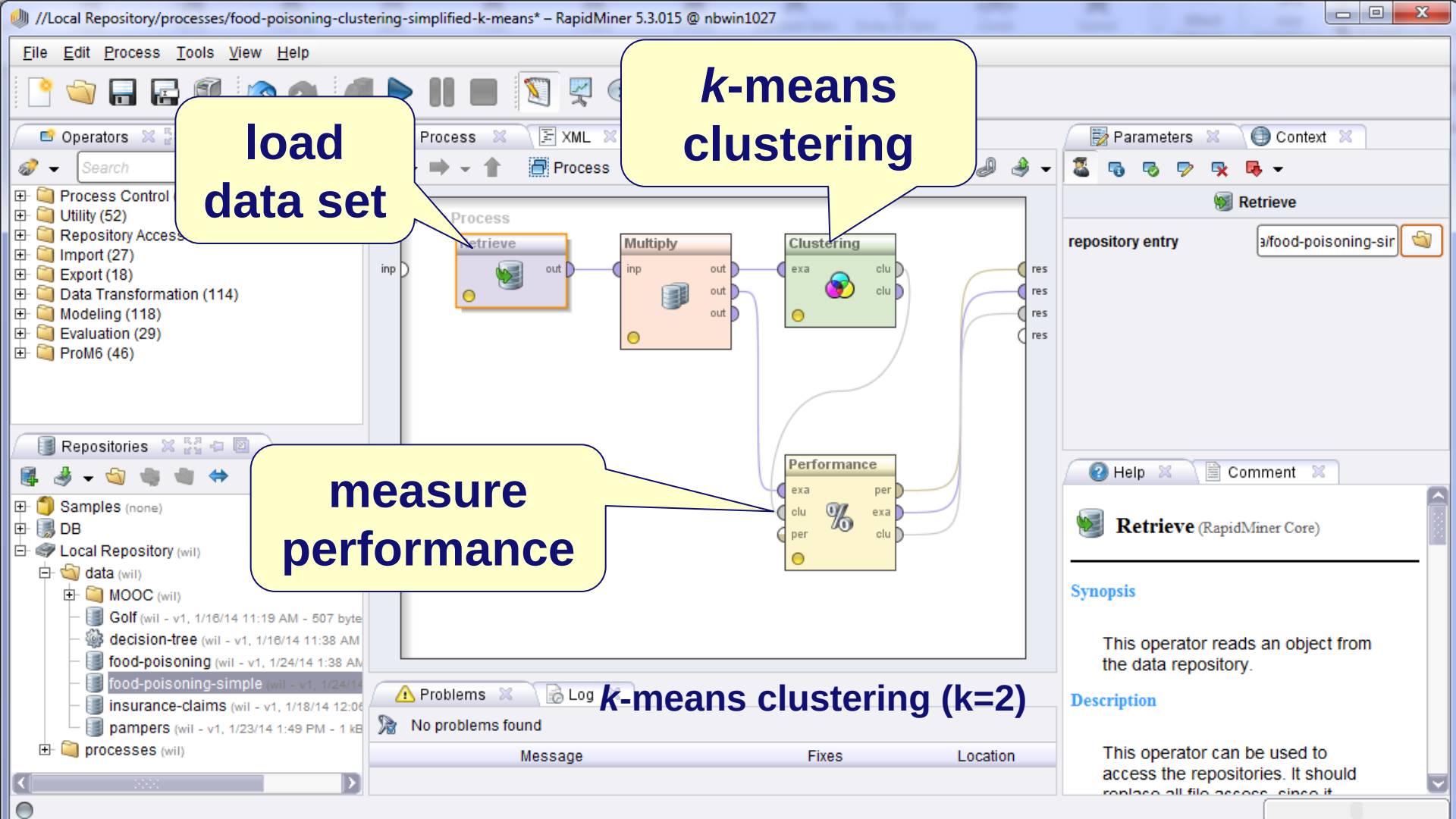
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (5000 examples, 1 special attribute, 6 regular attributes)

View Filter (5000 / 5000): all

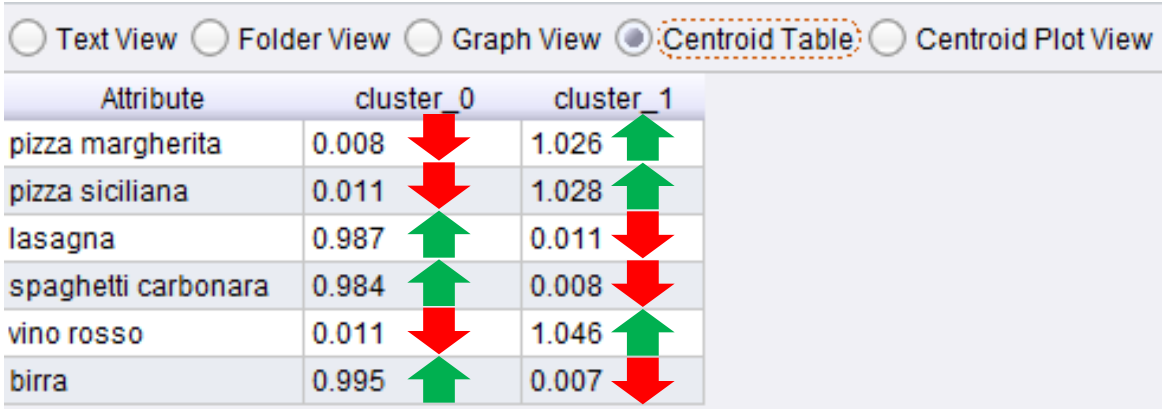
Row No.	class	pizza margh...	pizza sicilia...	lasagna	spaghetti c...	vino rosso	birra
1	not sick	0	0	1	2	0	2
2	not sick	0	0	1	2	0	1
3	not sick	1	1	0	0	1	0
4	not sick	2	1	0	0	2	0
5	not sick	1	0	0	0	1	0
6	not sick	0	0	2	0	0	1
7	not sick	0	0	1	1	0	1
8	not sick	2	1	0	0	2	0
9	not sick	0	0	2	0	0	2
10	not sick	1	1	0	0	1	0
11	not sick	0	0	2	2	0	1
12	not sick	0	1	0	0	2	0
13	not sick	1	1	0	0	0	0
14	not sick	1	1	0	0	2	0
15	not sick	0	0	1	1	0	2

only 6 items


//Local Repository/processes/food-poisoning-clustering-simplified-k-means* – RapidMiner 5.3.015 @ nbwin1027
File Edit Process Tools View Help
load data set
k-means clustering
measure performance
k-means clustering (k=2)
Operators
Process
XML
Parameters
Context
Search
Process Control
Utility (52)
Repository Access
Import (27)
Export (18)
Data Transformation (114)
Modeling (118)
Evaluation (29)
ProM6 (46)
Retrieve
Multiply
Clustering
Performance
repository entry
Repositories
Samples (none)
DB
Local Repository (wil)
data (wil)
MOOC (wil)
Golf (wil - v1, 1/16/14 11:19 AM - 507 byte
decision-tree (wil - v1, 1/16/14 11:38 AM
food-poisoning (wil - v1, 1/24/14 1:38 AM
food-poisoning-simple
insurance-claims (wil - v1, 1/18/14 12:06
pampers (wil - v1, 1/23/14 1:49 PM - 1 kB
processes (wil)
Problems
Log
No problems found
Message
Fixes
Location
Help
Comment
Retrieve (RapidMiner Core)
Synopsis
This operator reads an object from the data repository.
Description
This operator can be used to access the repositories. It should

Results *k*-means clustering (k=2)

- **Two clusters:**
  - **Cluster 0: 2551 instances**
  - **Cluster 1: 2449 instances**
- **Centroids**



Attribute	cluster_0	cluster_1
pizza margherita	0.008	1.026
pizza siciliana	0.011	1.028
lasagna	0.987	0.011
spaghetti carbonara	0.984	0.008
vino rosso	0.011	1.046
birra	0.995	0.007

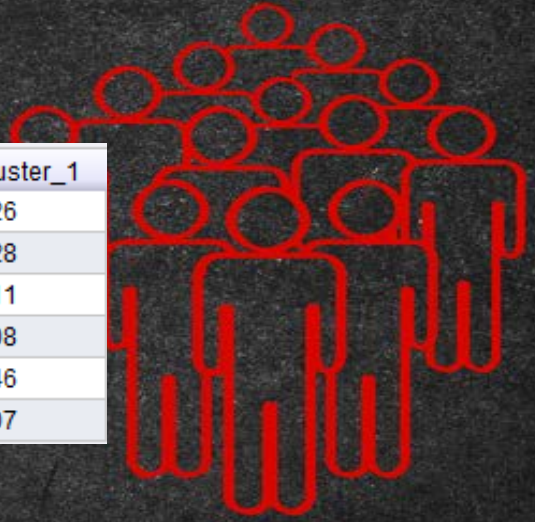
TU/e

Cluster 0: 2551 instances

Cluster 1: 2449 instances

birra

vino rosso



Attribute	cluster_0	cluster_1
pizza margherita	0.008	1.026
pizza siciliana	0.011	1.028
lasagna	0.987	0.011
spaghetti carbonara	0.984	0.008
vino rosso	0.011	1.046
birra	0.995	0.007

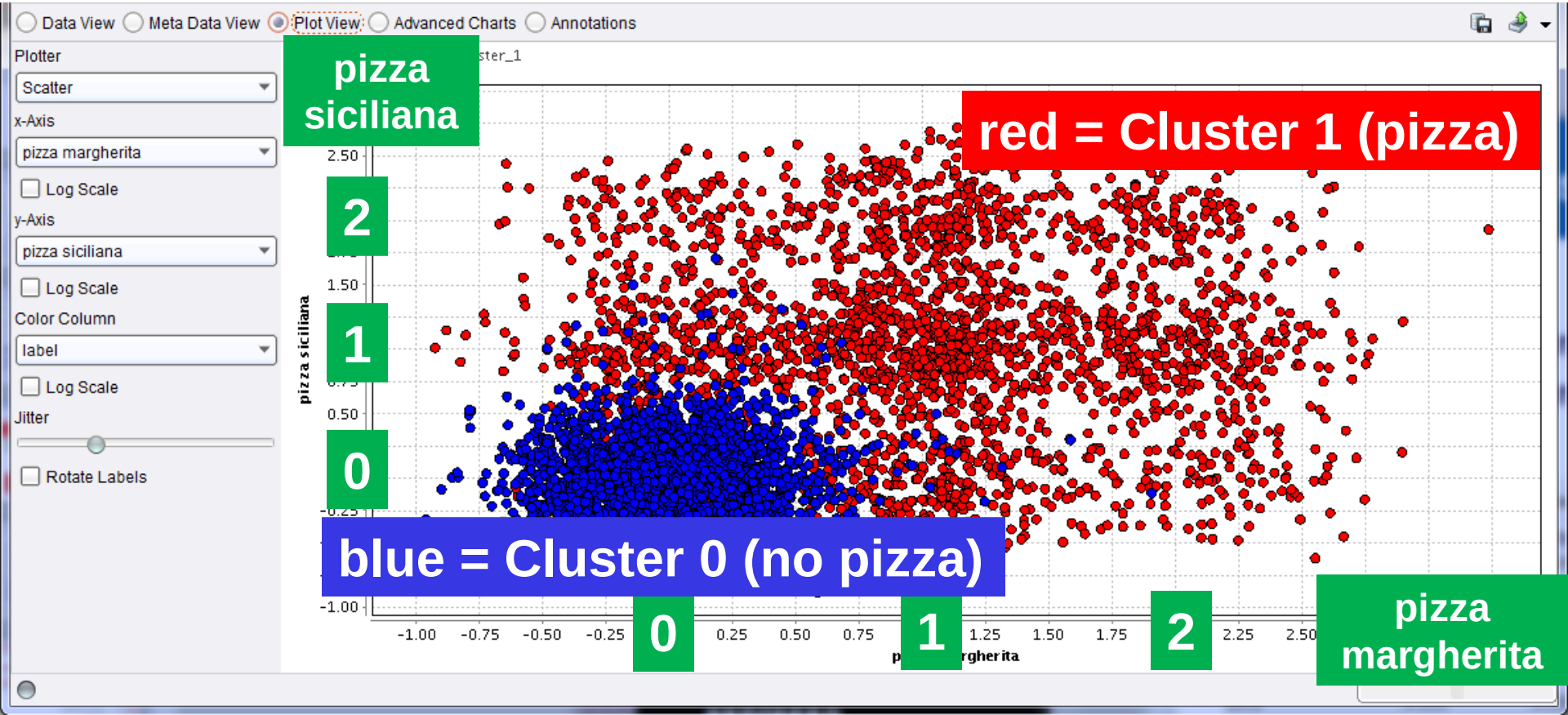
lasagna

pizza siciliana

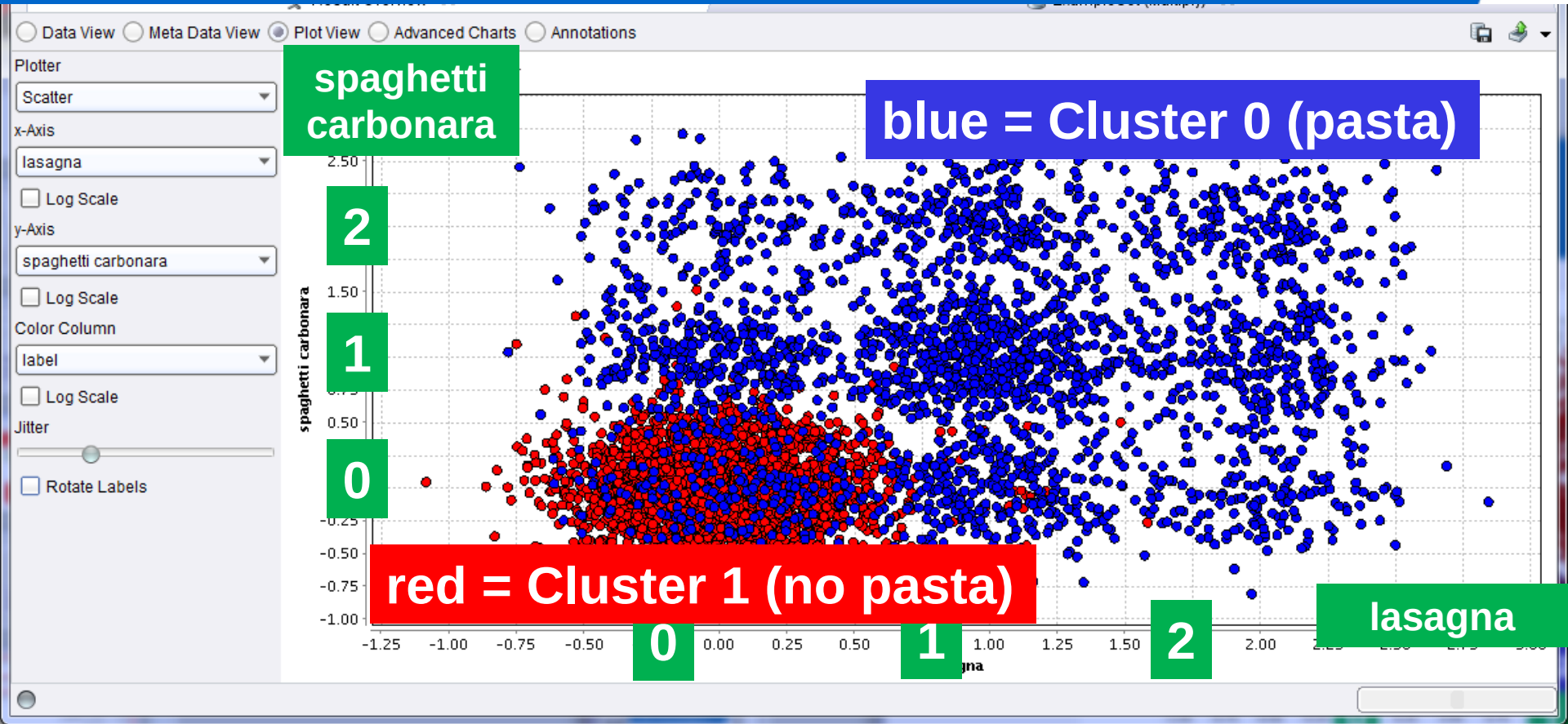
spaghetti carbonara

pizza margherita

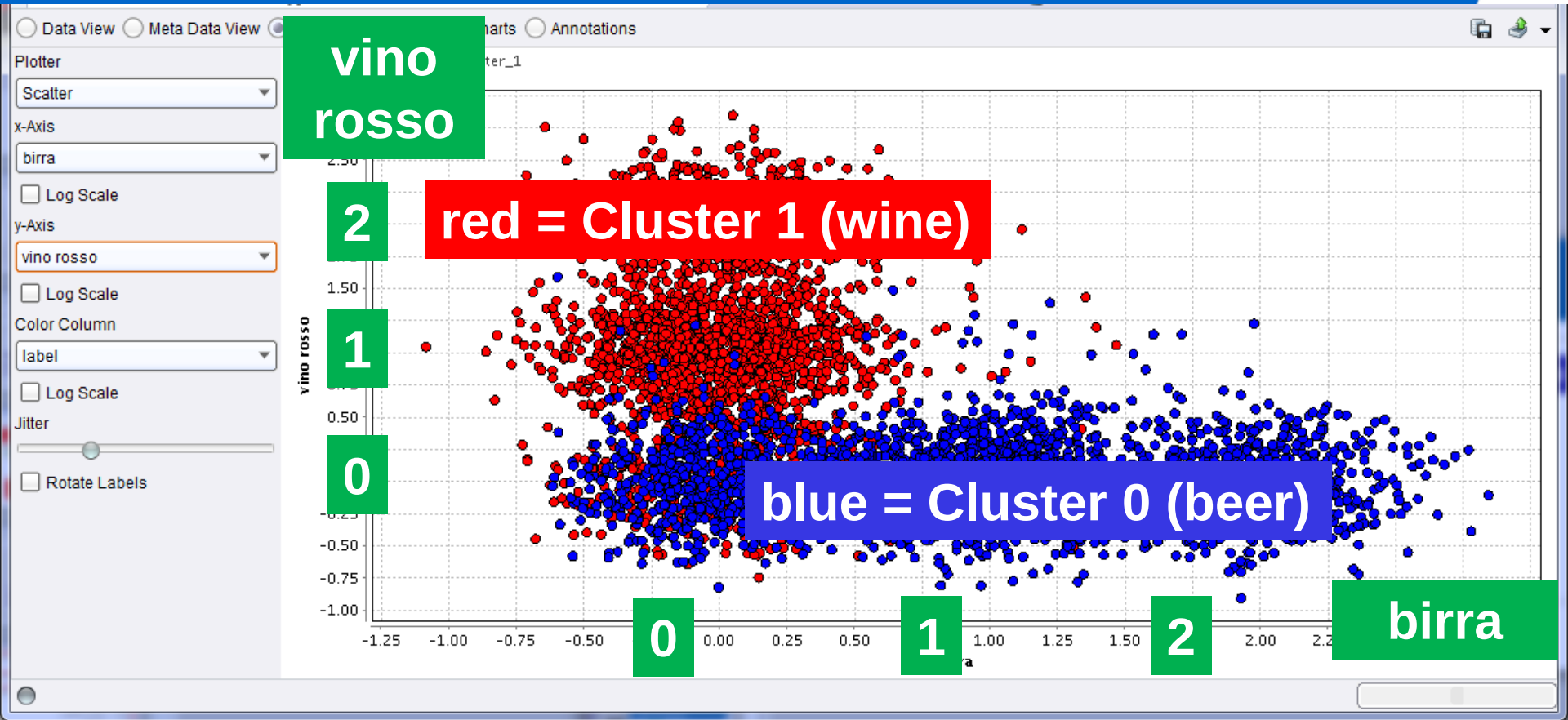
Scatter plot (with jitter)



Scatter plot (with jitter)

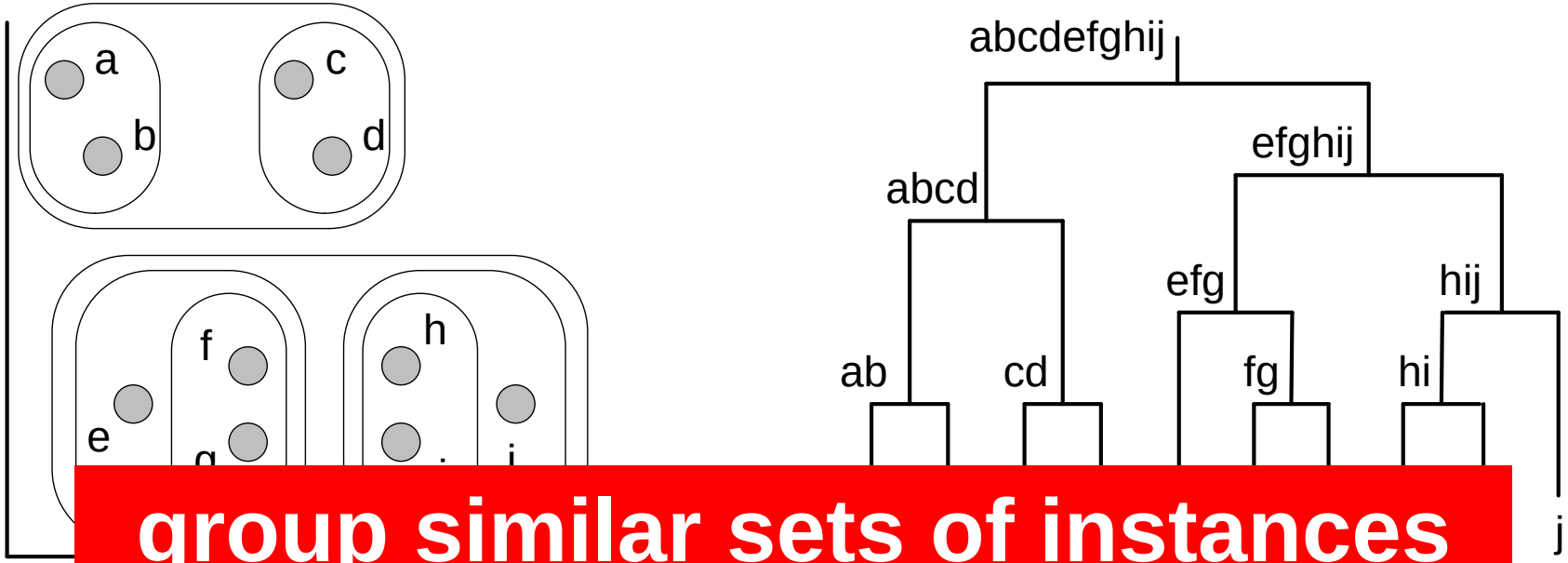


Scatter plot (with jitter)




Group 1
Group 2
Group 3
Group 4
Group 5
Group 8
not only
k-means

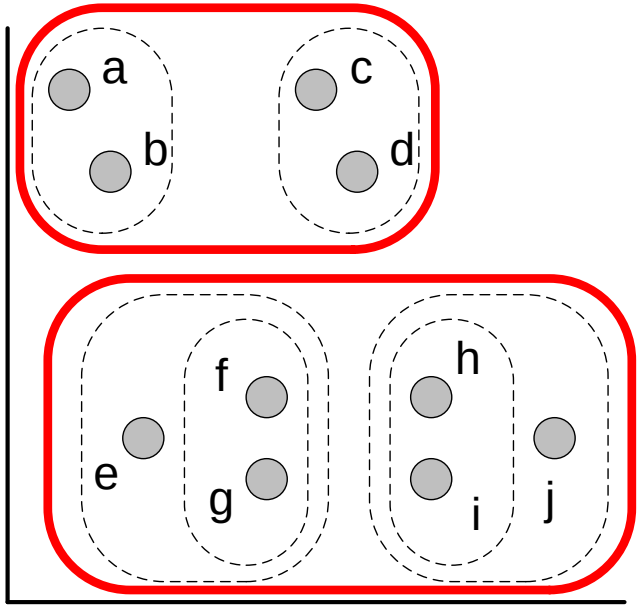
Agglomerative hierarchical clustering

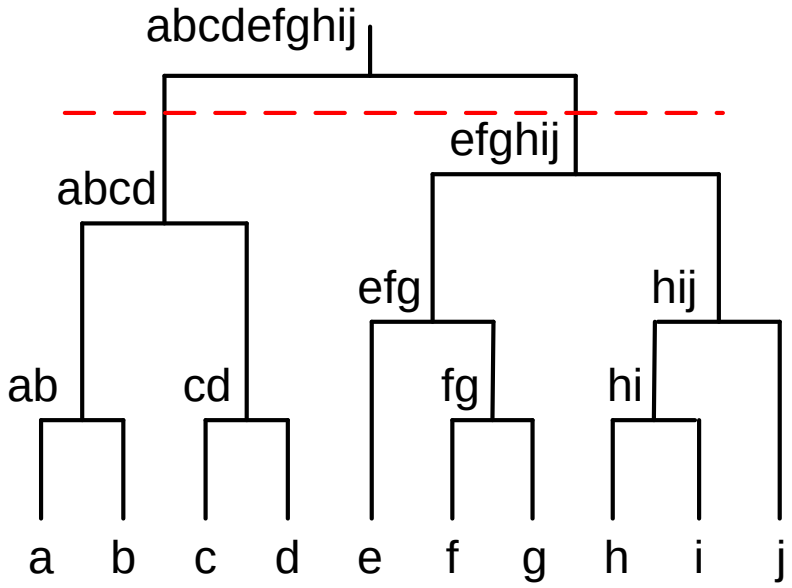


group similar sets of instances in a hierarchical manner

TU/e

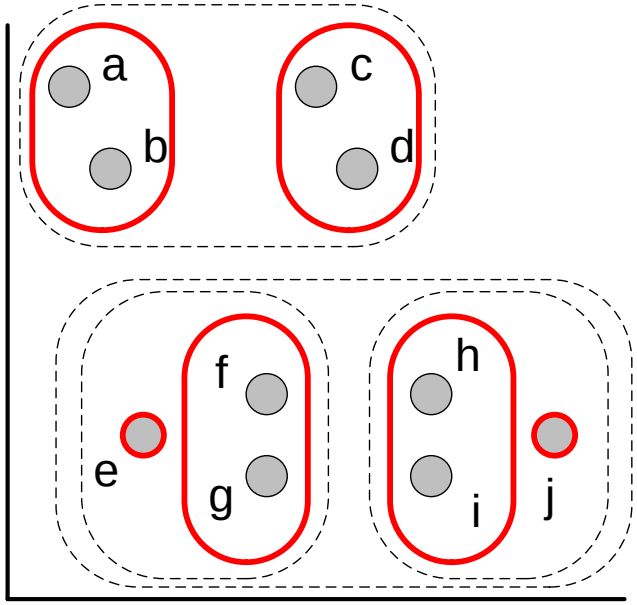
Two larger clusters

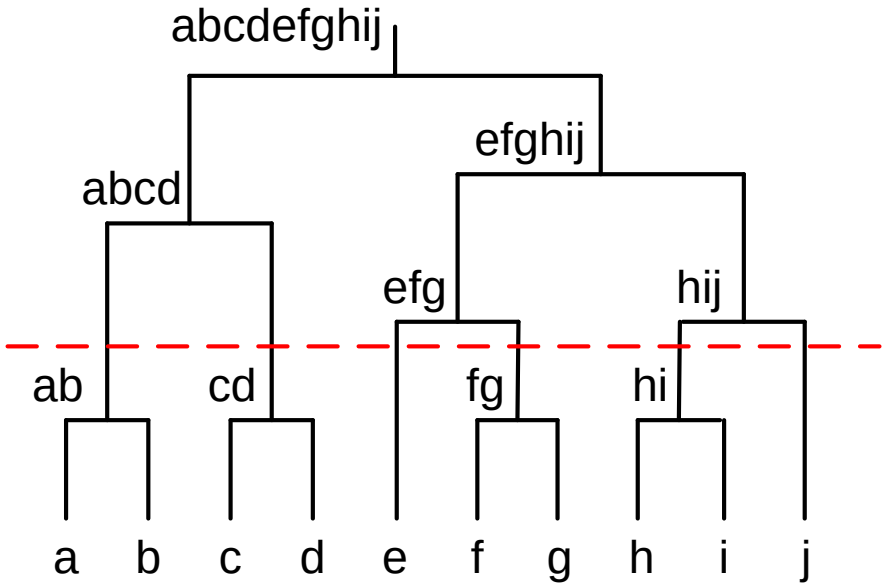




dendrogram

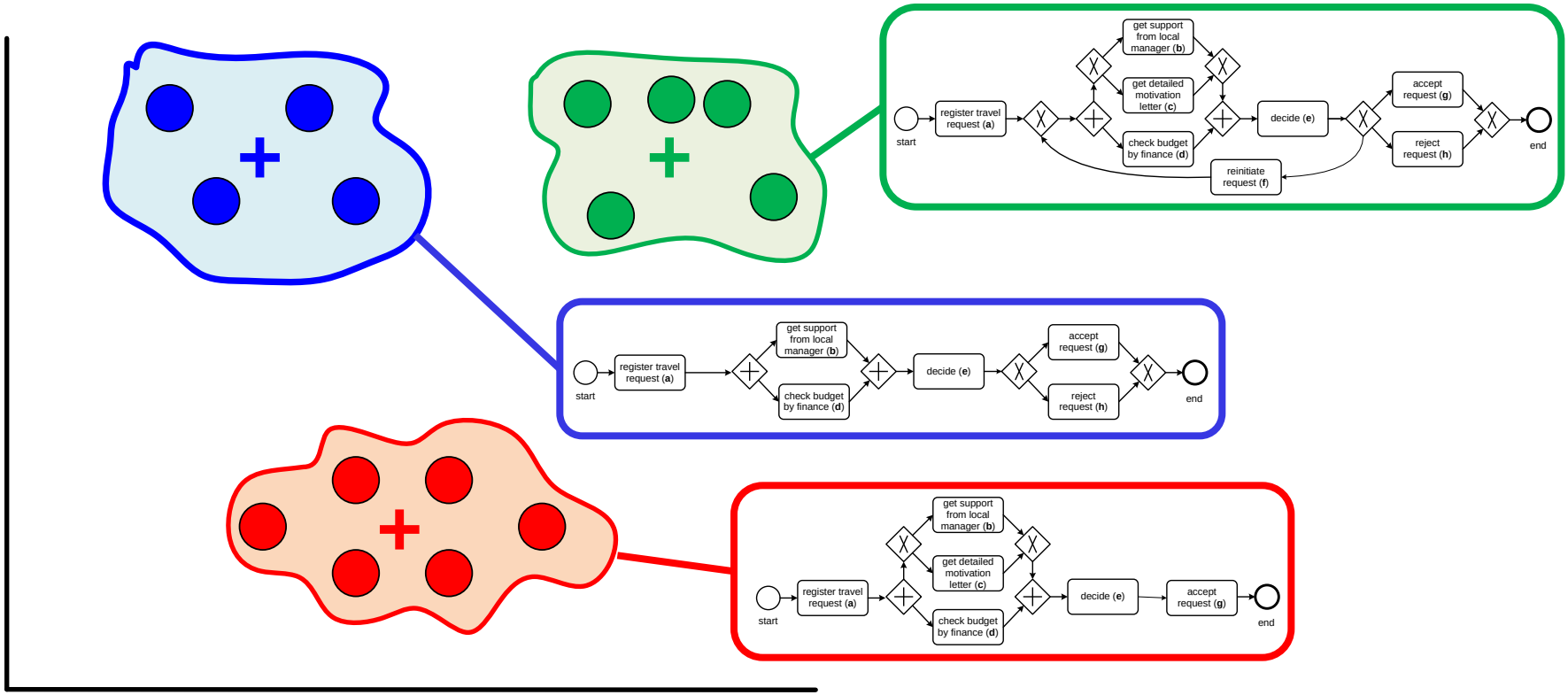
Six smaller clusters





dendrogram

Clustering can be used to split event logs



TU/e

Part I: Preliminaries

- Chapter 1: Introduction
- Chapter 2: Process Modeling and Analysis
- Chapter 3: Data Mining

Part II: From Event Logs to Process Models

- Chapter 4: Getting the Data
- Chapter 5: Process Discovery: An Introduction
- Chapter 6: Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

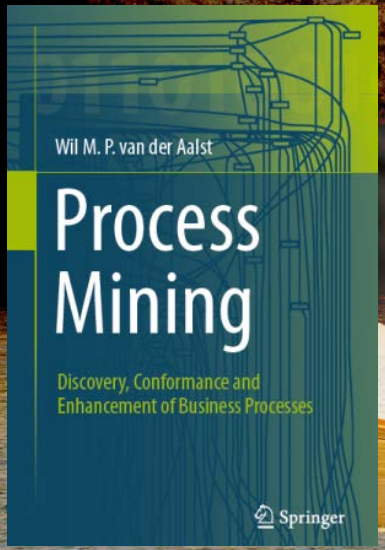
- Chapter 7: Conformance Checking
- Chapter 8: Mining Additional Perspectives
- Chapter 9: Operational Support

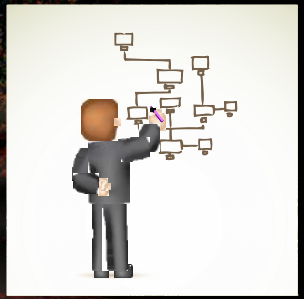
Part IV: Putting Process Mining to Work

- Chapter 10: Tool Support
- Chapter 11: Analyzing "Lasagna Processes"
- Chapter 12: Analyzing "Spaghetti Processes"

Part V: Reflection

- Chapter 13: Cartography and Navigation
- Chapter 14: Epilogue





TU/e