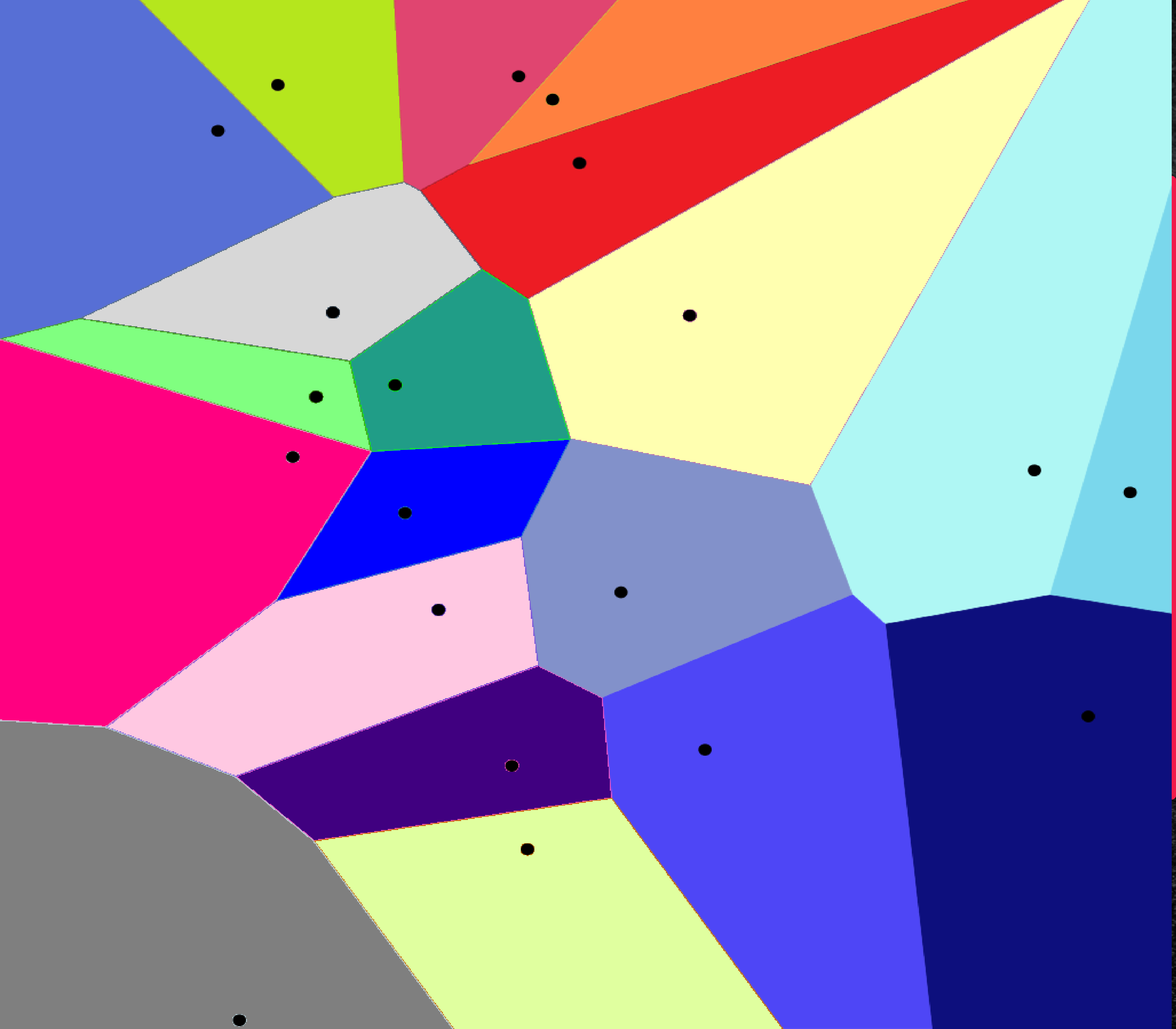


Clustering Analysis

By Karis Kim

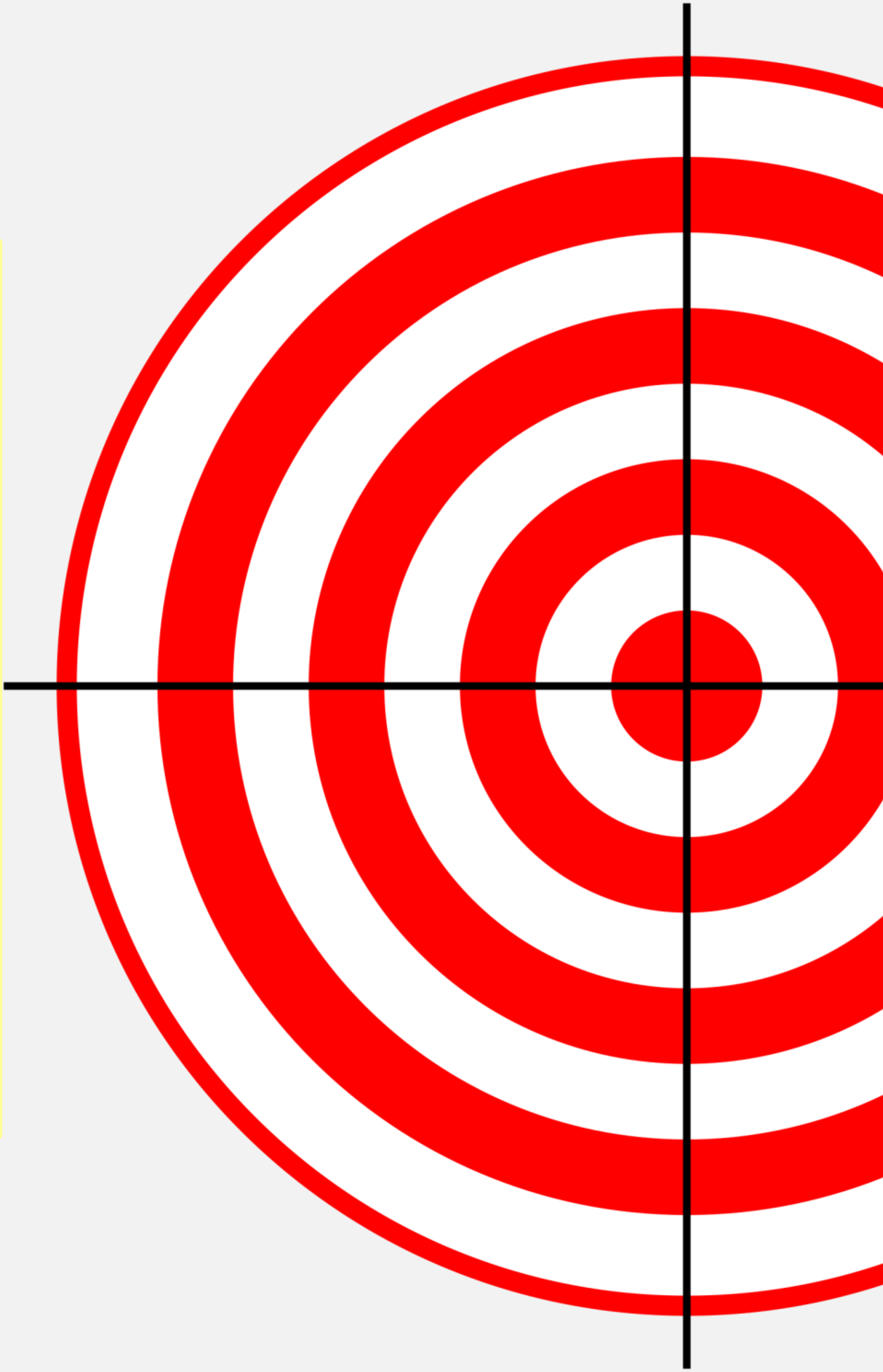
Clustering:

/ˈkləstərNG/

Process of finding meaningful groups in data

Goal:

Exploring,
not predicting



Applications of Clustering

Describing

1. Marketing
2. Document Clustering
3. Session Grouping

Preprocessing

1. Reduce Dimensionality
2. Object Reduction

Clustering Types

1 Exclusive/Strict Partitioning

2 Overlapping/ Multiview

3 Hierarchical

4 Fuzzy/ Probabilistic

Clustering Algorithm Types

1

Prototype/ Centroid/ Center Based

2

Density

3

Hierarchical

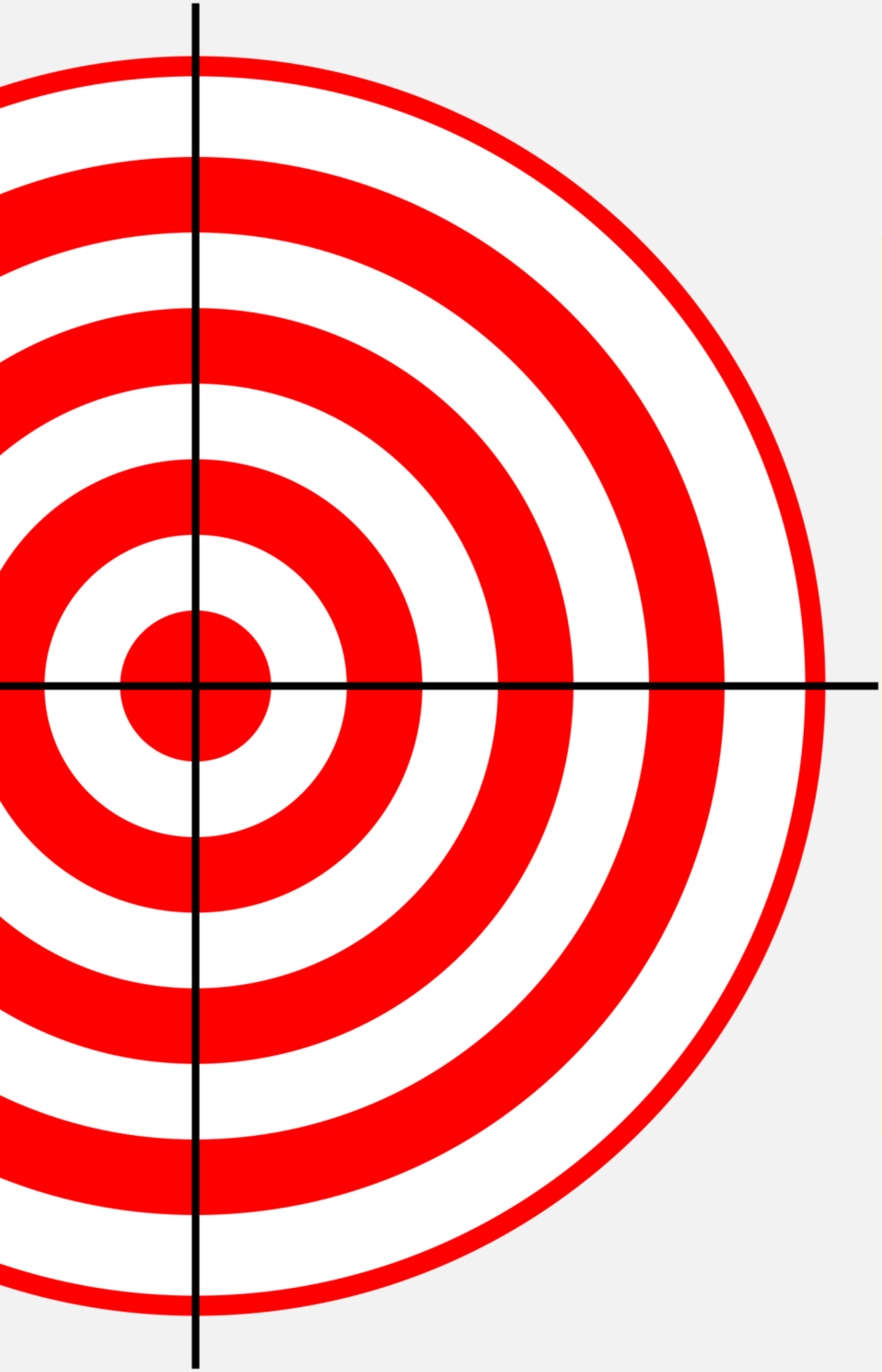
4

Model/ Distribution Based

k-Means:

/k-meens/

Data set is divided into user-specified *k* clusters around the nearest centroid

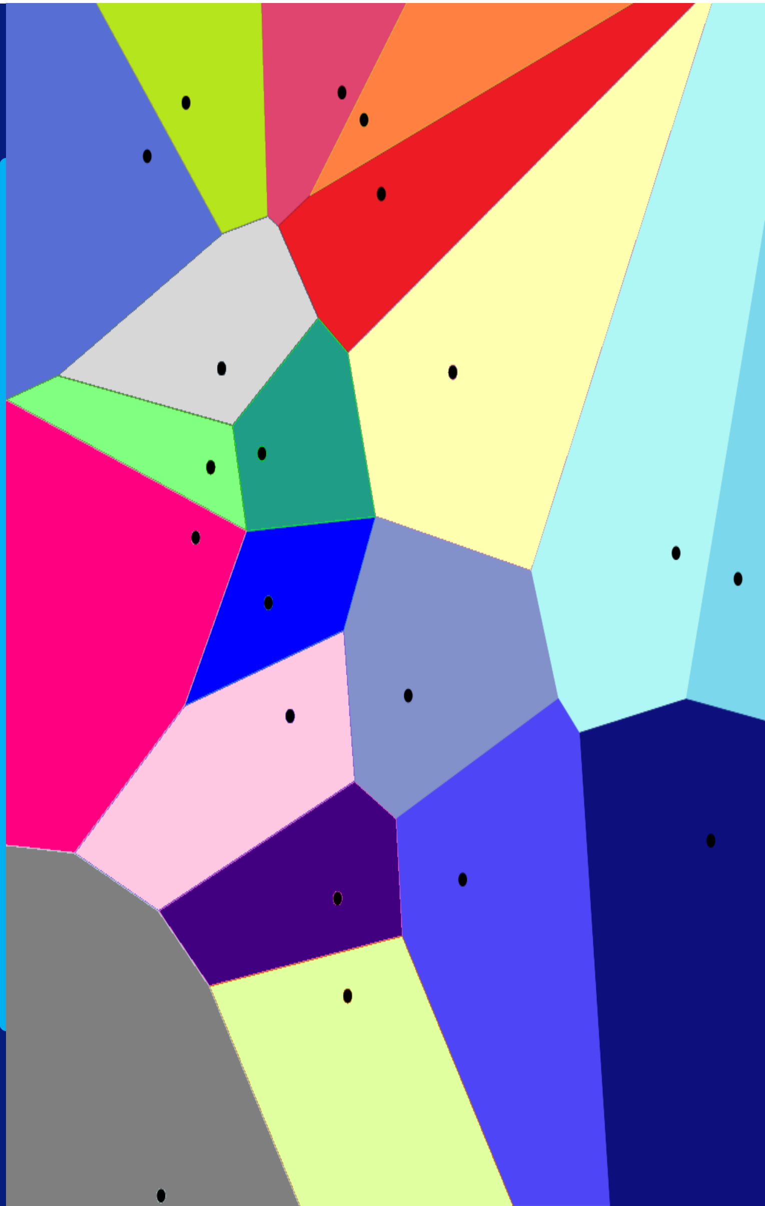


Goal:

To find the prototype data point for each cluster

k-Means Process

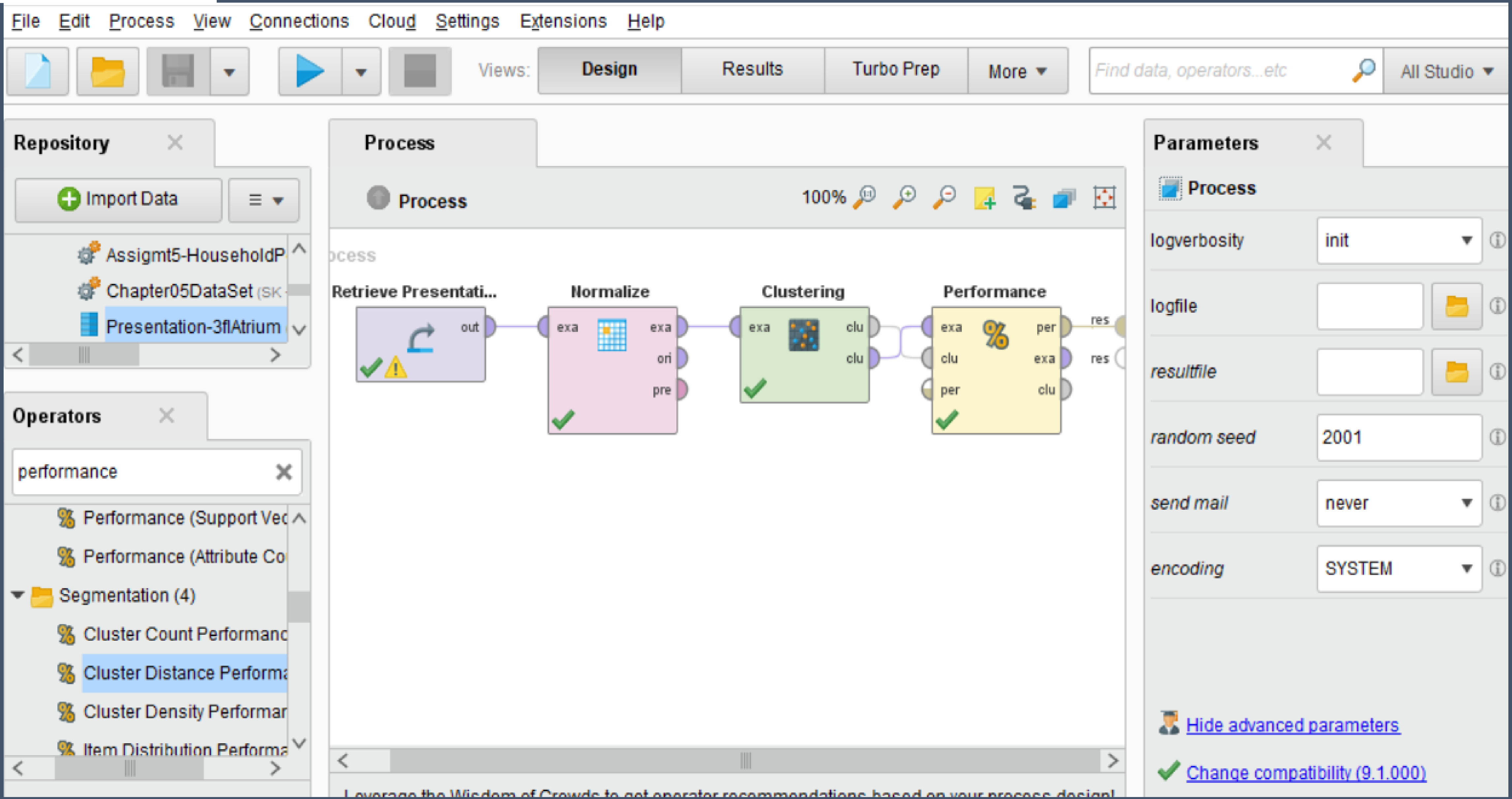
1. Set *k* # of clusters,
   initiate *k* random centroids
2. Assign data points to nearest centroid
3. Find new centroids (most representative)
   as centroid w/minimal SSE
   = new mean of cluster
4. Repeat step 2 and 3
5. When no significant change,
   declare final centroid as prototype



HOW TO PARTITION?

Define a proximity measure using Euclidean distance

rapidminer



logverbosity: init

logfile:

resultfile:

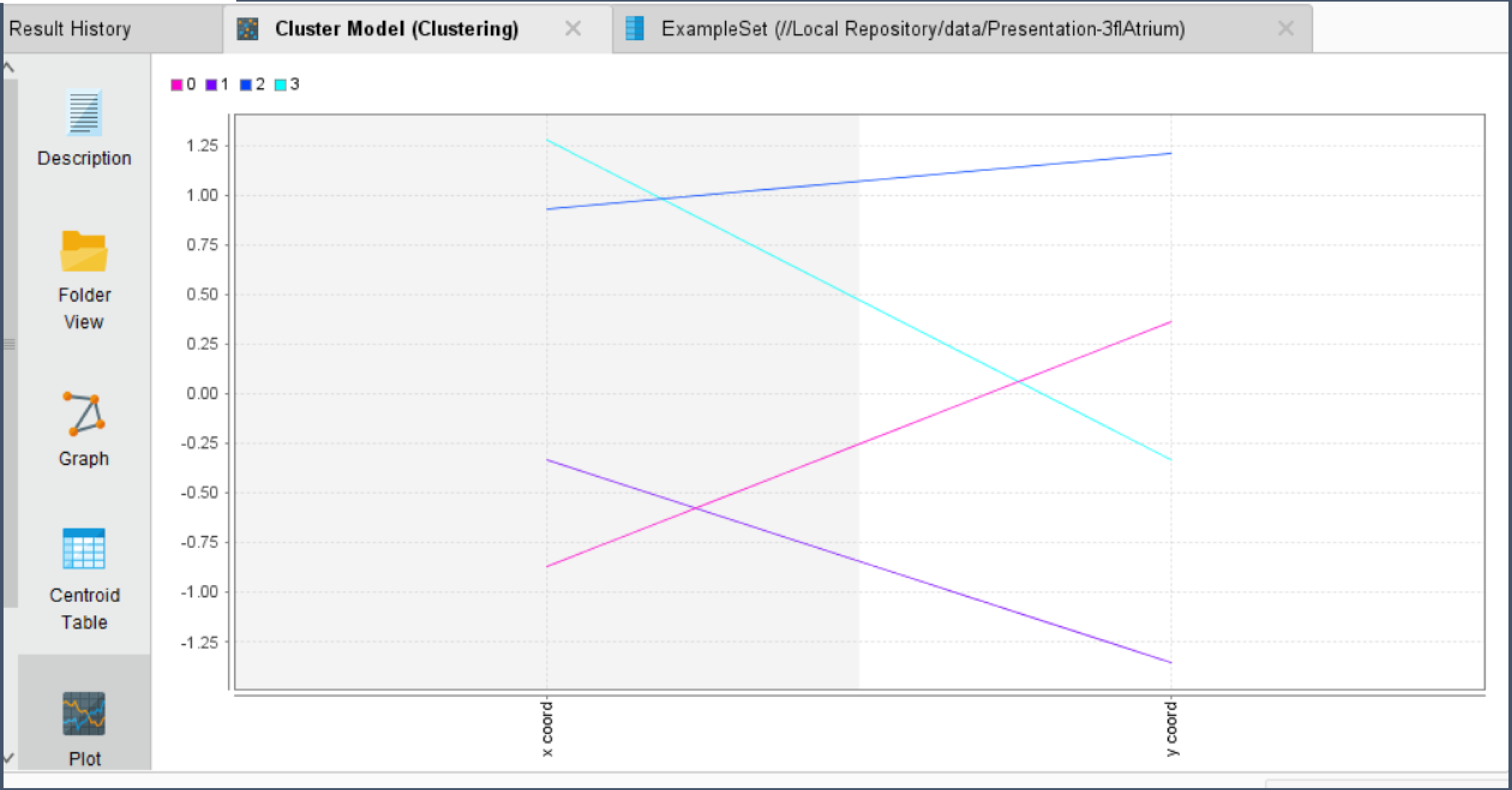
random seed: 2001

send mail: never

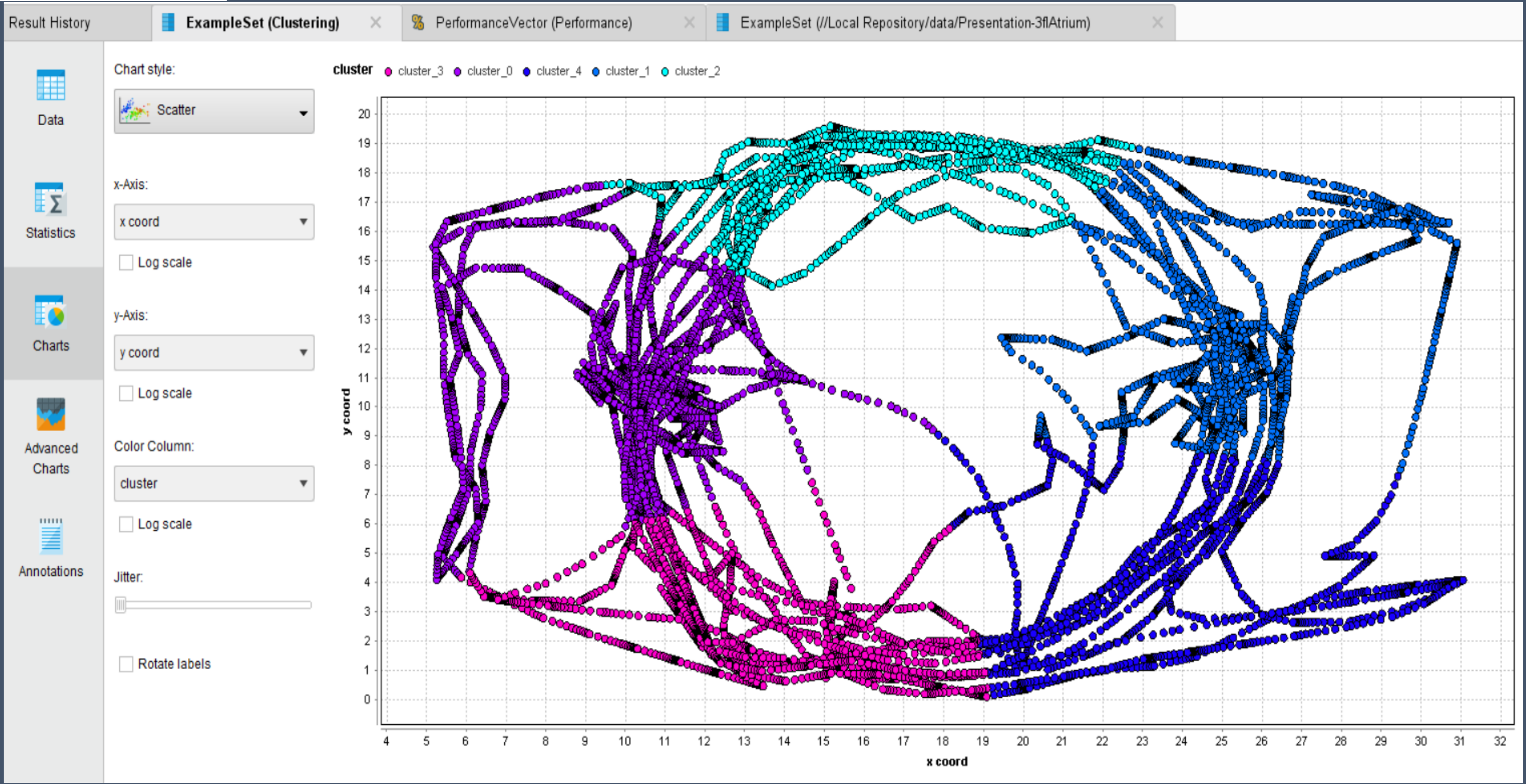
encoding: SYSTEM

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

rapidminer


Result History
Cluster Model (Clustering)
ExampleSet (//Local Repository/data/Presentation-3flAtrium)
Description
Folder View
Graph
Centroid Table
Plot
0
1
2
3
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
-1.25
x coord
y coord

rapidminer



Result History | ExampleSet (Clustering) | PerformanceVector (Performance) | ExampleSet (//Local Repository/data/Presentation-3flAtrium)

Chart style: Scatter

x-Axis: x coord

☐ Log scale

y-Axis: y coord

☐ Log scale

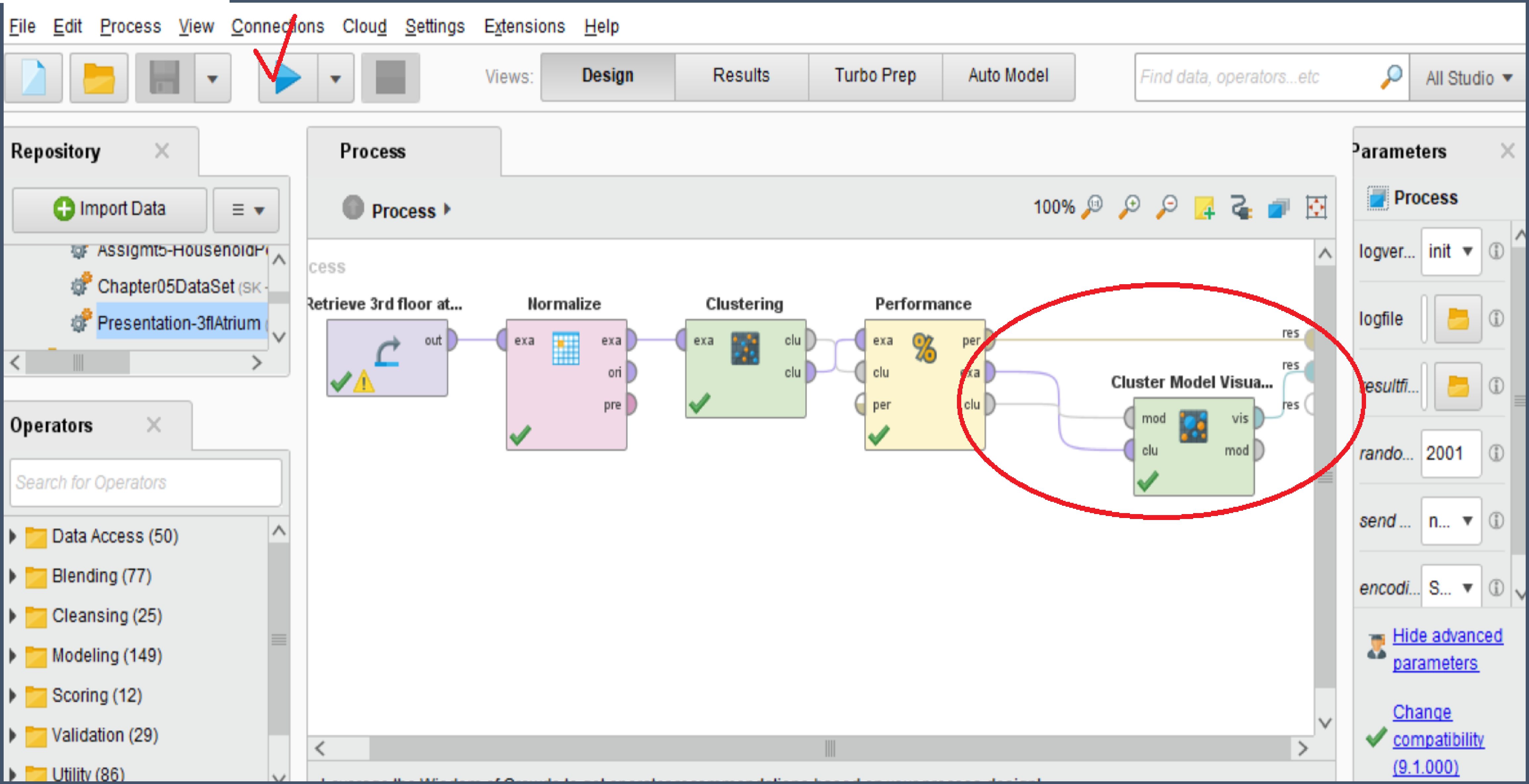
Color Column: cluster

☐ Log scale

Jitter:

☐ Rotate labels

rapidminer



logver... init

logfile

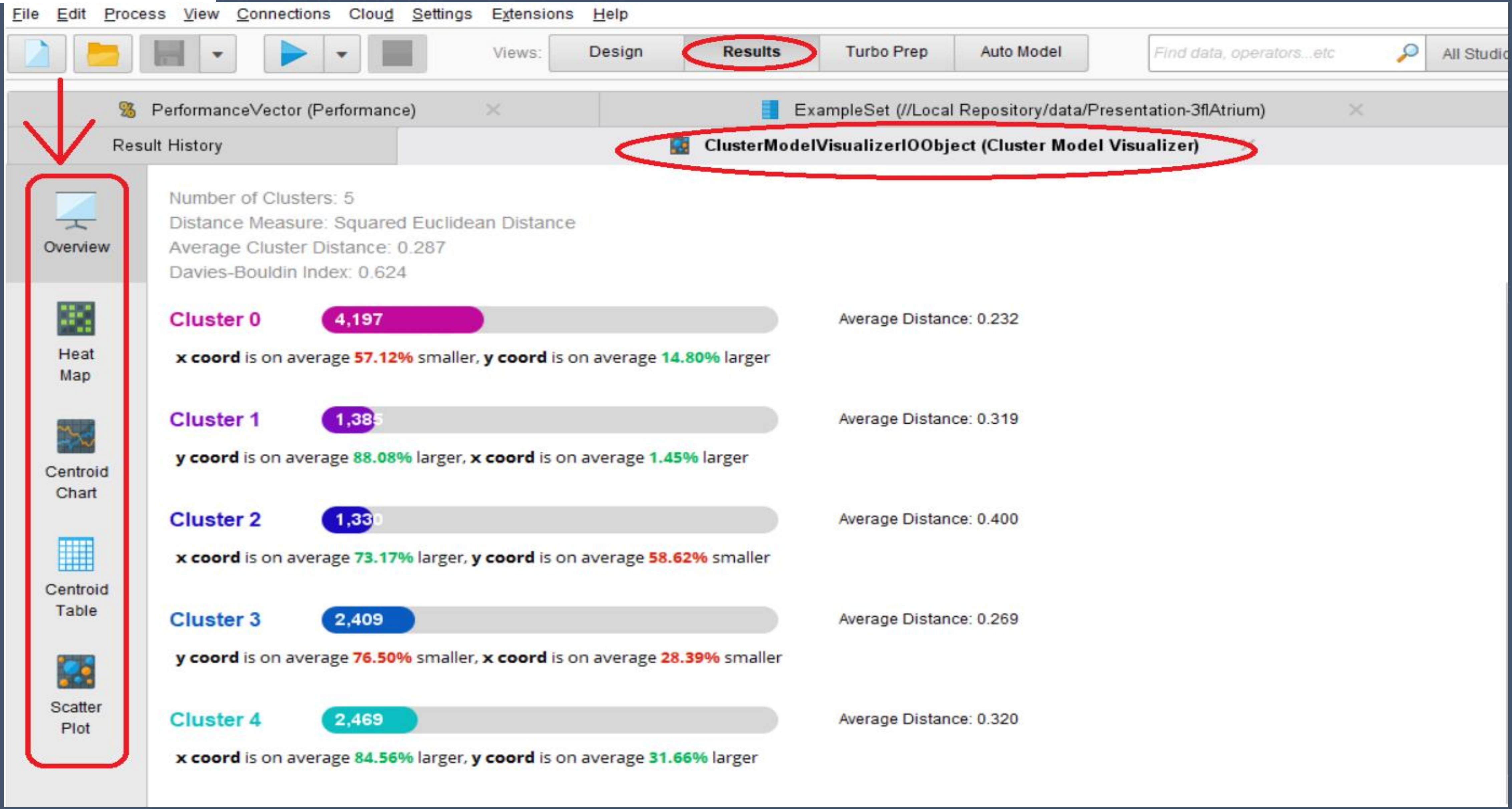
resultfi...

rando... 2001

send ... n...

encodi... S...

rapidminer



Number of Clusters: 5
Distance Measure: Squared Euclidean Distance
Average Cluster Distance: 0.287
Davies-Bouldin Index: 0.624

Cluster 0 4,197 Average Distance: 0.232

x coord is on average 57.12% smaller, **y coord** is on average 14.80% larger

Cluster 1 1,385 Average Distance: 0.319

y coord is on average 88.08% larger, **x coord** is on average 1.45% larger

Cluster 2 1,330 Average Distance: 0.400

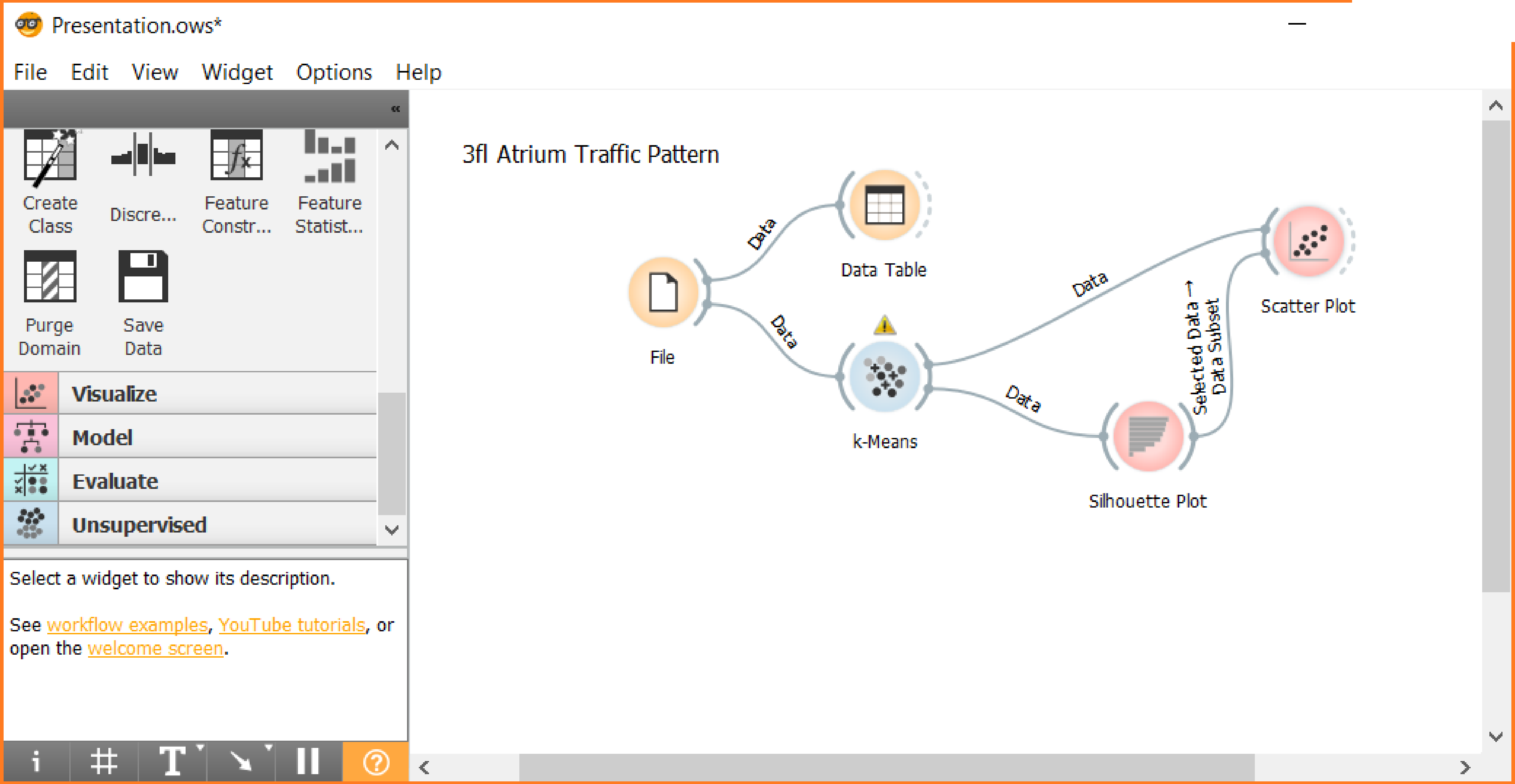
x coord is on average 73.17% larger, **y coord** is on average 58.62% smaller

Cluster 3 2,409 Average Distance: 0.269

y coord is on average 76.50% smaller, **x coord** is on average 28.39% smaller

Cluster 4 2,469 Average Distance: 0.320

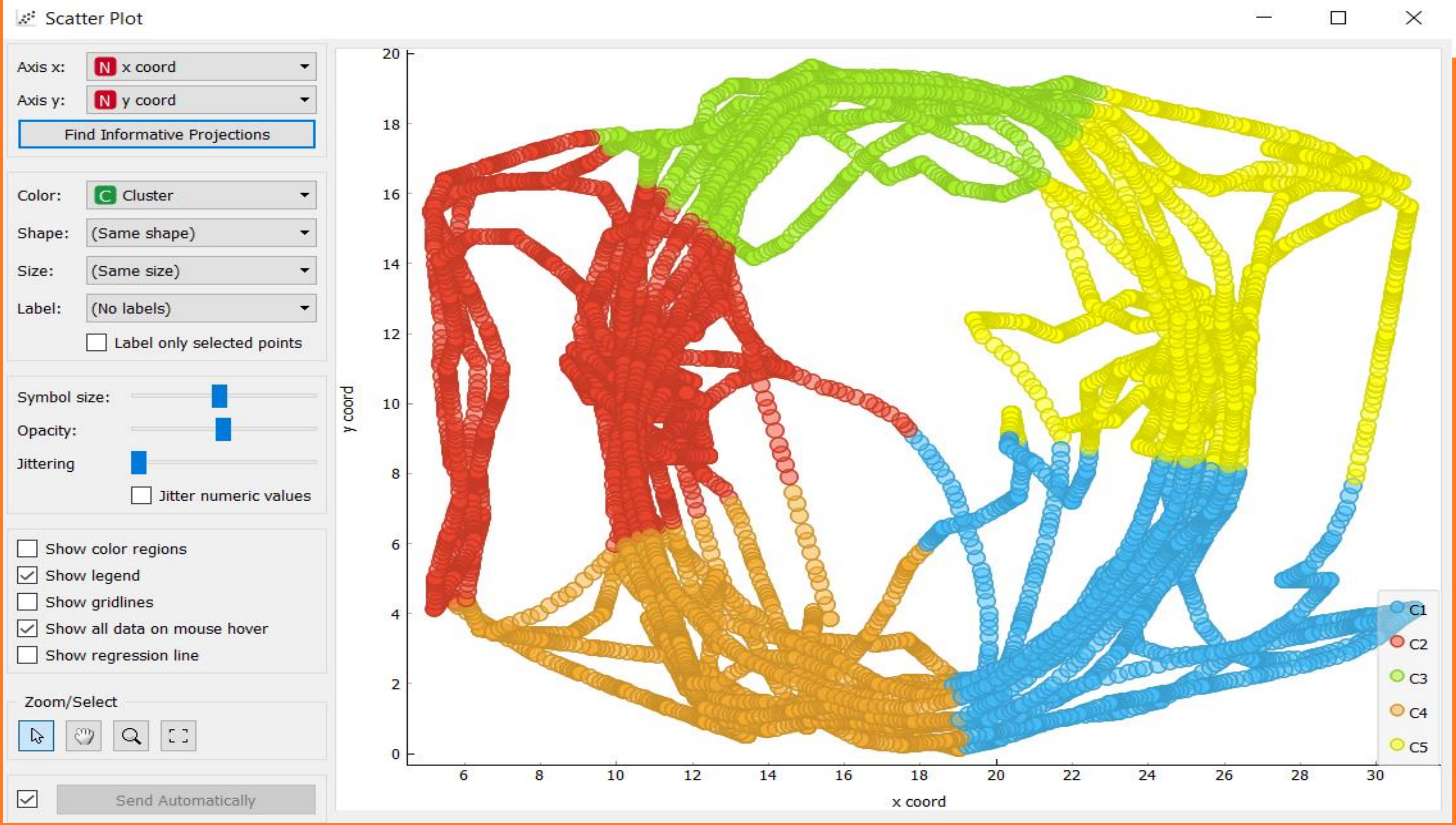
x coord is on average 84.56% larger, **y coord** is on average 31.66% larger



3fl Atrium Traffic Pattern

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.



Scatter Plot

Axis x: N x coord

Axis y: N y coord

Find Informative Projections

Color: C Cluster

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

□ Label only selected points

Symbol size:

Opacity:

Jittering

□ Jitter numeric values

□ Show color regions

☑ Show legend

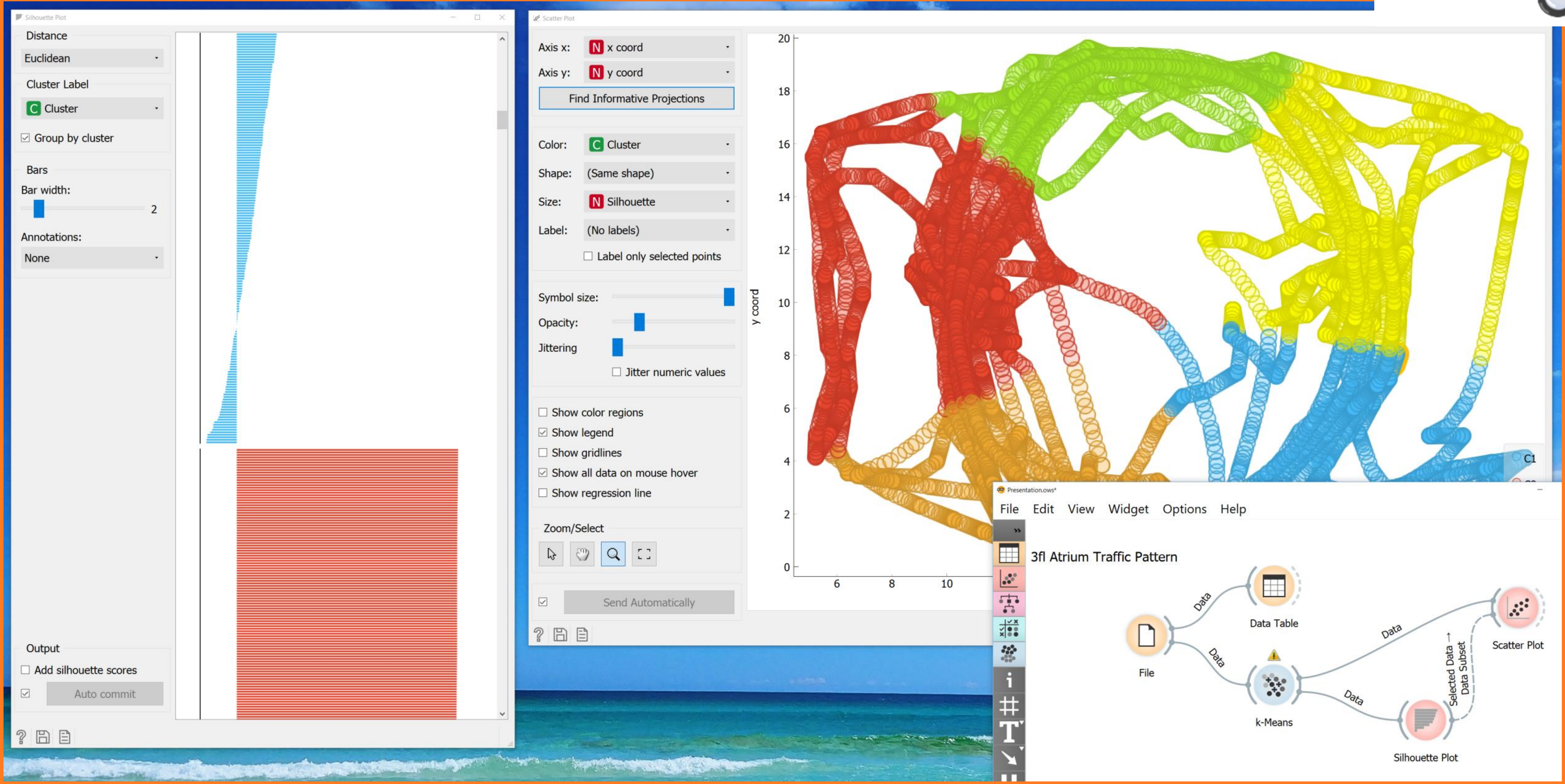
□ Show gridlines

☑ Show all data on mouse hover

□ Show regression line

Zoom/Select

☑ Send Automatically



Distance: Euclidean

Cluster Label: C Cluster

☑ Group by cluster

Bars

Bar width: 2

Annotations: None

Output

□ Add silhouette scores

☑ Auto commit

Axis x: N x coord

Axis y: N y coord

Find Informative Projections

Color: C Cluster

Shape: (Same shape)

Size: N Silhouette

Label: (No labels)

□ Label only selected points

Symbol size:

Opacity:

Jittering

□ Jitter numeric values

□ Show color regions

☑ Show legend

□ Show gridlines

☑ Show all data on mouse hover

□ Show regression line

Zoom/Select

☑ Send Automatically

3fl Atrium Traffic Pattern

RapidMiner vs. Orange

rapidminer

- **Performance operator** measures avr within centroid distance & Davies Bouldin index
- Lower the index the better*
- No limit on sample size
- Cluster Model Visualizer operator to see clusters

orange

- **Silhouette scores** measure how well on avr each data pt fits into designated cluster
- Higher the score the better
- Not computed >5000 samples*
- Cluster visualized automatically; syncs interactively to Silhouette

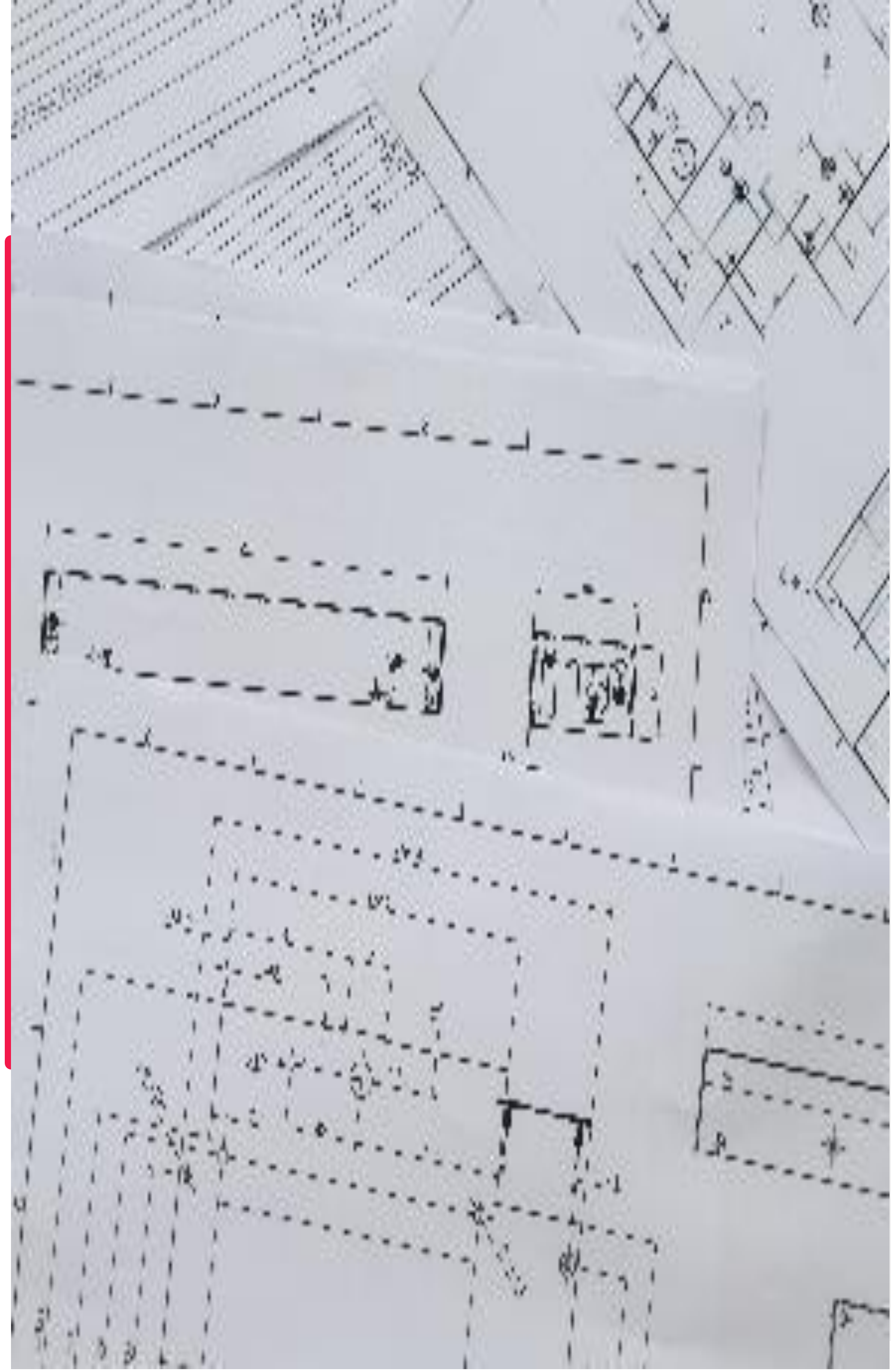
Comparison Summary

	RapidMiner	Orange
Measures cluster quality	Y	Y
Manually select centroids*	N	Y
Sample size limit for cluster quality*	N	Y
Normalize data*	Y	N
Cluster results visualization	Y	Y

Sample Data Information

KSU Marietta Campus Atrium Bldg 3 FL Traffic Pattern Analysis

- 11790 examples with 2 attributes (x & y coordinates)
- Data captured by beacons every 10 sec for 1 hour
- Goal: Monitor traffic pattern clusters to inform virtual/mixed reality application for indoor navigation assistance for the blind
- Idea was to use beacons to detect heavy/light traffic areas indoors and warn blind users, as change from heavy to light/no traffic could indicate construction or path blockage



Thank You

https://www.youtube.com/watch?v=IOeOQyev8Ac ORANGE TUTORIAL