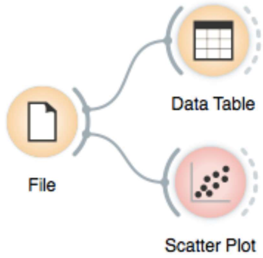
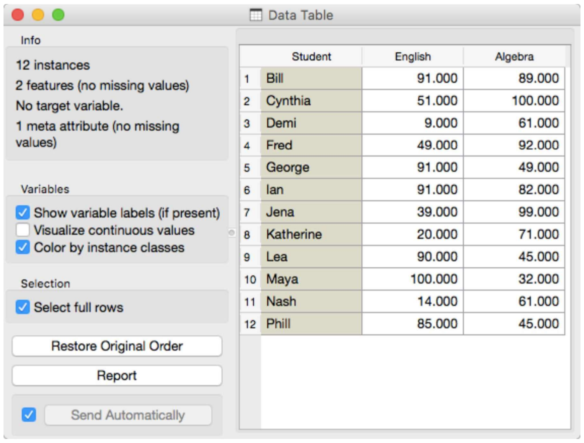
In the class, we will introduce clustering using a simple data set on students and their grades in English and Algebra. Load the data set from http://file.biolab.si/files/grades2.tab.

Lesson 23: Hierarchical Clustering

Say that we are interested in finding clusters in the data. That is, we would like to identify groups of data instances that are close together, similar to each other. Consider a simple, two-featured data set (see the side note) and plot it in the Scatter Plot. How many clusters do we have? What defines a cluster? Which data instances belong to the same cluster? What would a procedure for discovering clusters look like?




Data Table

Info

12 instances
2 features (no missing values)
No target variable.
1 meta attribute (no missing values)

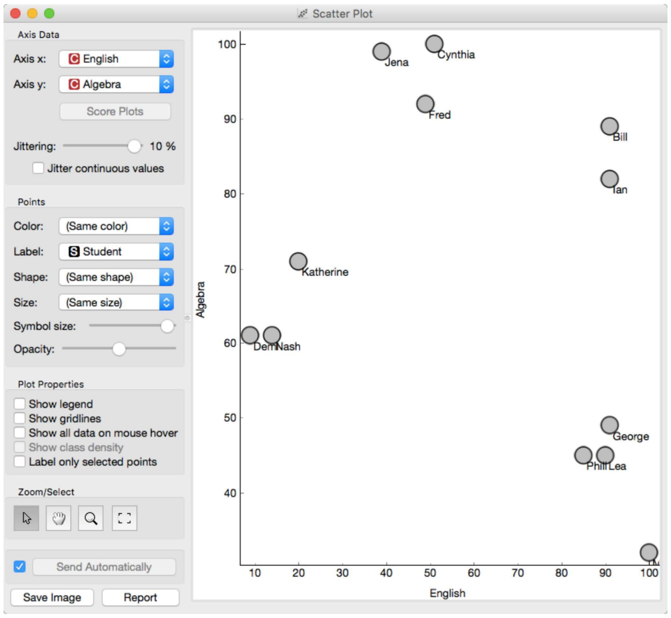
Variables

☑ Show variable labels (if present)
☐ Visualize continuous values
☑ Color by instance classes

Selection

☑ Select full rows

	Student	English	Algebra
1	Bill	91.000	89.000
2	Cynthia	51.000	100.000
3	Demi	9.000	61.000
4	Fred	49.000	92.000
5	George	91.000	49.000
6	Ian	91.000	82.000
7	Jena	39.000	99.000
8	Katherine	20.000	71.000
9	Lea	90.000	45.000
10	Maya	100.000	32.000
11	Nash	14.000	61.000
12	Phill	85.000	45.000


Scatter Plot

Axis Data

Axis x: English
Axis y: Algebra

Jittering: 10 %
☐ Jitter continuous values

Points

Color: (Same color)
Label: Student
Shape: (Same shape)
Size: (Same size)

Plot Properties

☐ Show legend
☐ Show gridlines
☐ Show all data on mouse hover
☐ Show class density
☐ Label only selected points

☑ Send Automatically

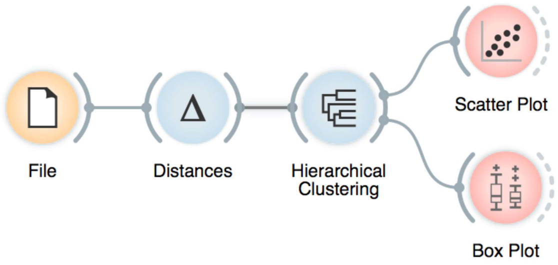
How do we measure the similarity between clusters if we only know the similarities between points? By default, Orange computes the average distance between all their pairs of data points; this is called average linkage. We could instead take the distance between the two closest points in each cluster (single linkage), or the two points that are furthest away (complete linkage).

We need to start with a definition of “similar”. One simple measure of similarity for such data is the Euclidean distance: square the differences across every dimension, some them and take the square root, just like in Pythagorean theorem. So, we would like to group data instances with small Euclidean distances.

Now we need to define a clustering algorithm. We will start with each data instance being in its own cluster. Next, we merge the clusters that are closest together - like the closest two points - into one cluster. Repeat. And repeat. And repeat. And repeat until you end up with a single cluster containing all points.

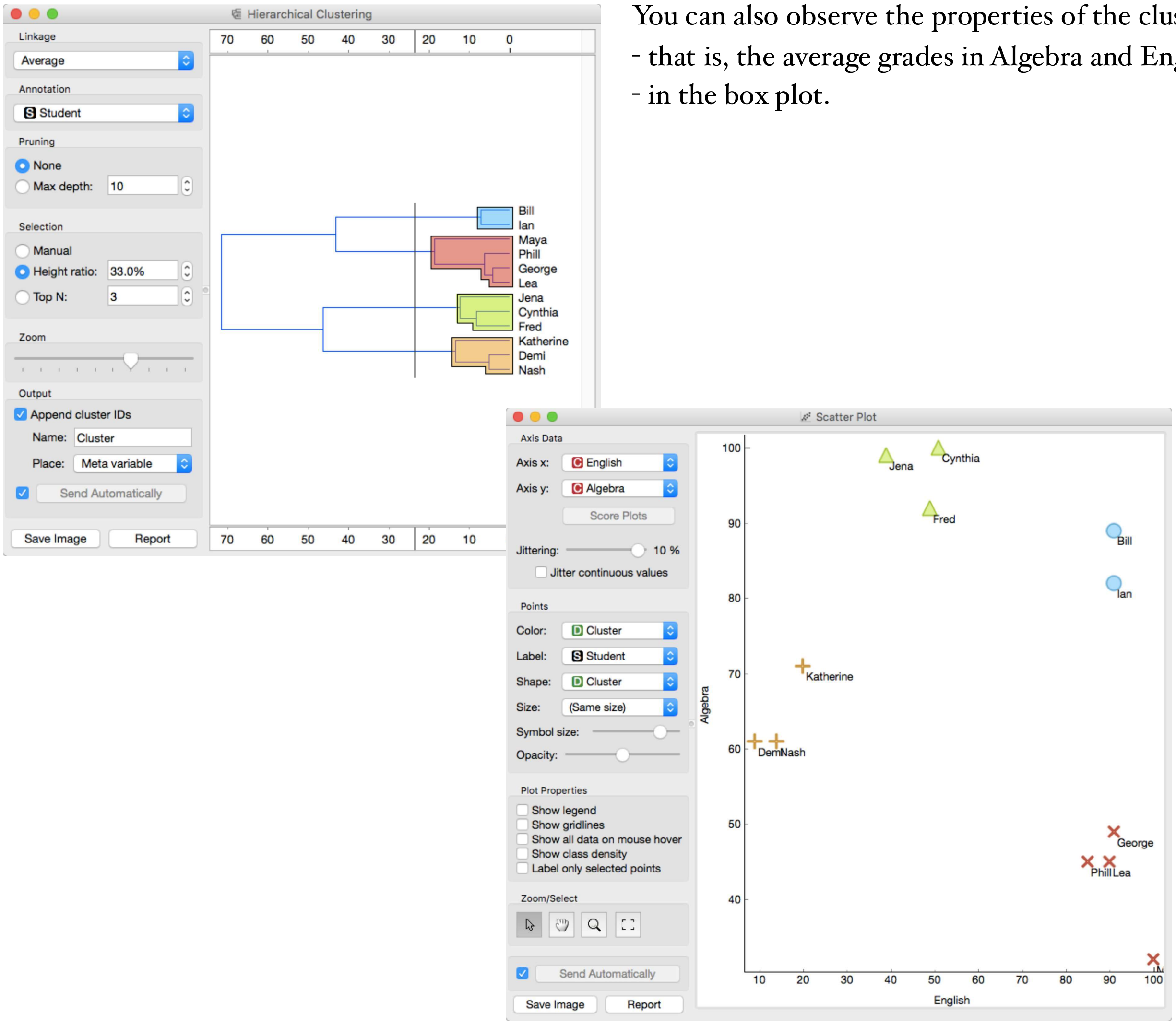
This procedure constructs a hierarchy of clusters, which explains why we call it hierarchical clustering. After it is done, we can observe the entire hierarchy and decide which would be a good point to stop. With this we decide the actual number of clusters.

One possible way to observe the results of clustering on our small data set with grades is through the following workflow:



Let us see how this works. Load the data, compute the distances and cluster the data. In the Hierarchical clustering widget, cut hierarchy at a certain distance score and observe the corresponding clusters in the Scatter plot.

You can also observe the properties of the clusters - that is, the average grades in Algebra and English - in the box plot.


Hierarchical Clustering

Linkage

Average

Annotation

Student

Pruning

● None

○ Max depth: 10

Selection

○ Manual

● Height ratio: 33.0%

○ Top N: 3

Zoom

Output

☑ Append cluster IDs

Name: Cluster

Place: Meta variable

☑ Send Automatically

Save Image | Report

Scatter Plot

Axis Data

Axis x: English

Axis y: Algebra

Score Plots

Jittering: 10 %

□ Jitter continuous values

Points

Color: Cluster

Label: Student

Shape: Cluster

Size: (Same size)

Symbol size:

Opacity:

Plot Properties

□ Show legend

□ Show gridlines

□ Show all data on mouse hover

□ Show class density

□ Label only selected points

Zoom/Select

☑ Send Automatically

Save Image | Report