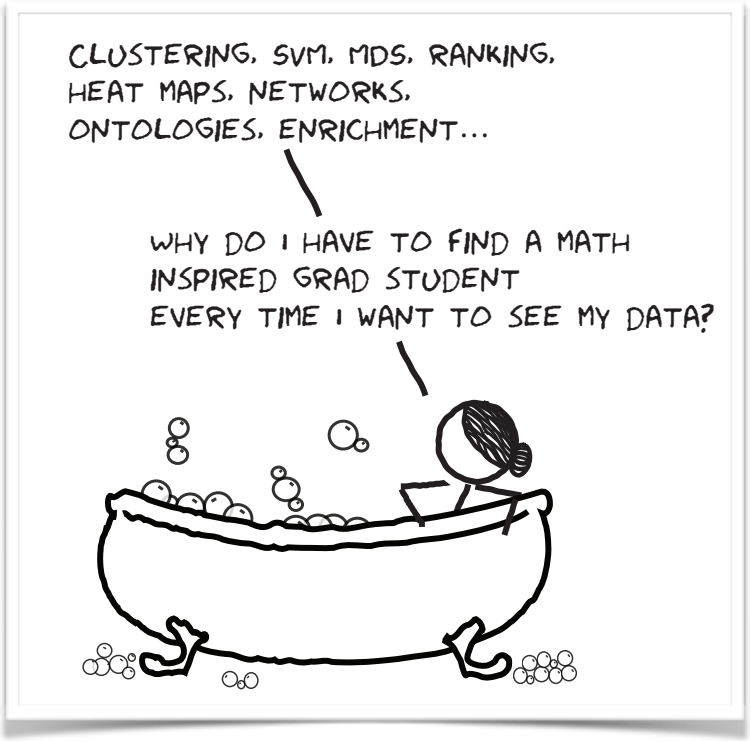


Data Mining

Working notes for Introduction to Data Mining Course

These notes include Orange workflows and visualizations we will construct during the course.

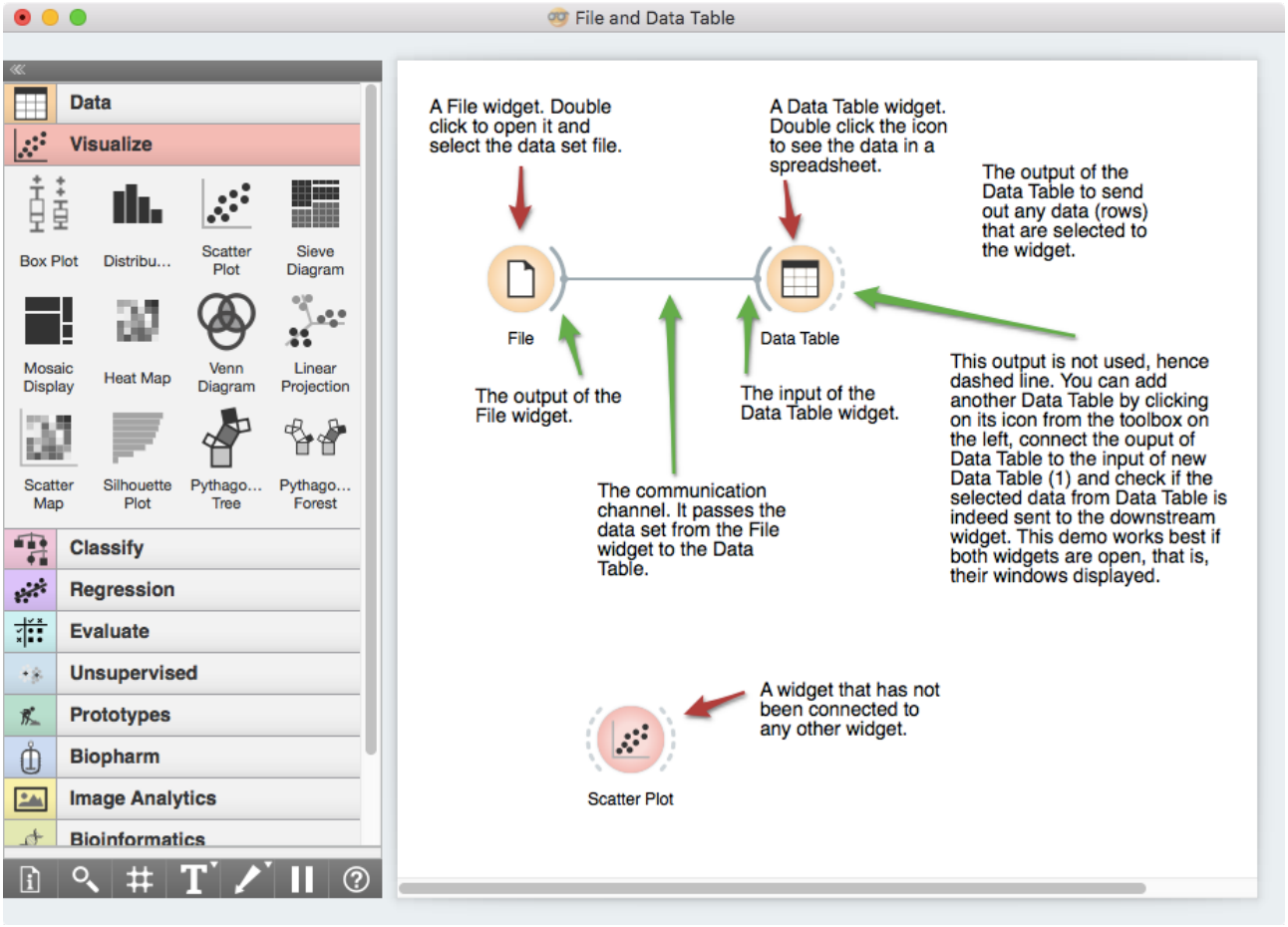
The notes were written by Blaž Zupan and Janez Demšar with huge help from the members of the Bioinformatics Lab in Ljubljana that develop and maintain Orange.

Welcome to the course on Data Mining! This course is designed for students and researchers of life sciences. You will see how common data mining tasks can be accomplished without programming. We will use Orange to construct visual data mining workflows. Many similar data mining environments exist, but the lecturers prefer Orange for one simple reason—they are its authors.

If you haven't already installed Orange, please download the installation package from http://orange.biolab.si.

Lesson 1: Workflows in Orange

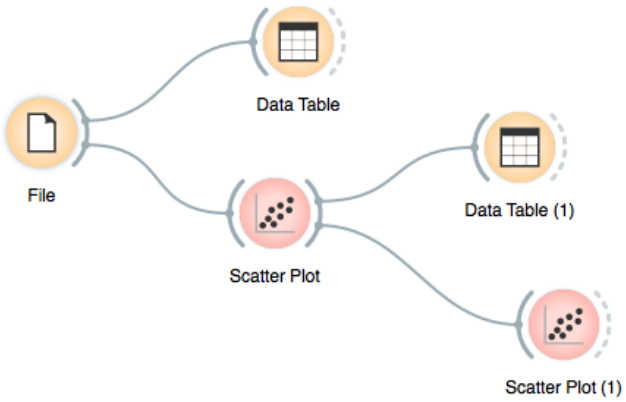
Orange workflows consist of components that read, process and visualize data. We call them "widgets". Widgets are placed on a drawing board (the "canvas"). Widgets communicate by sending information along a communication channel. Output from one widget is used as input to another.



A simple workflow with two connected widgets and one widget without connections. The outputs of a widget appear on the right, while the inputs appear on the left.

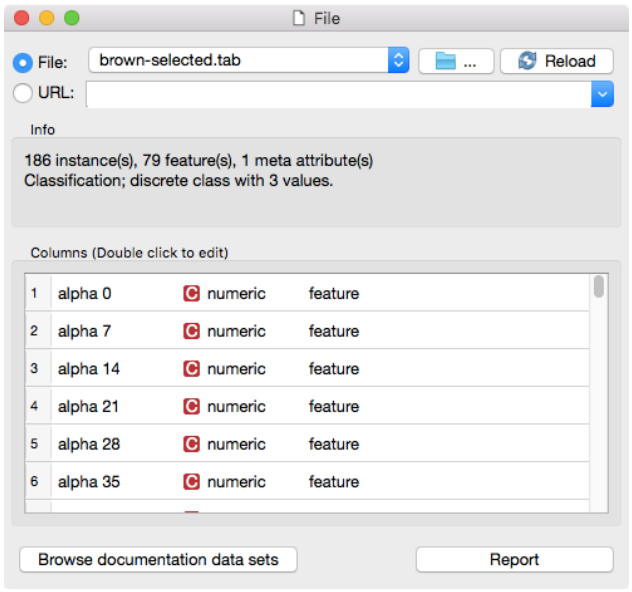
We construct workflows by dragging widgets onto the canvas and connecting them by drawing a line from the transmitting widget to the receiving widget. The widget's outputs are on the right, and the inputs on the left. In the workflow above, the File widget sends data to the Data Table widget.

Start by constructing a workflow that consists of a File widget, two Scatter Plot widgets and two Data Table widgets:



Workflow with a File widget that reads data from disk and sends it to the Scatter Plot and Data Table widget. The Data Table renders the data in a spreadsheet, while the Scatter Plot visualizes it. Selected data points from the Scatterplot are sent to two other widgets: Data Table (1) and Scatter Plot (1).

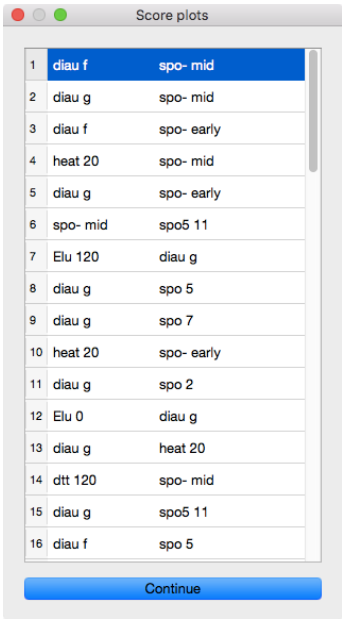
The File widget reads data from your local disk. Open the File Widget by double clicking its icon. Orange comes with several preloaded data sets. From these ("Browse documentation data sets..."), choose brown-selected.tab, a yeast gene expression data set.

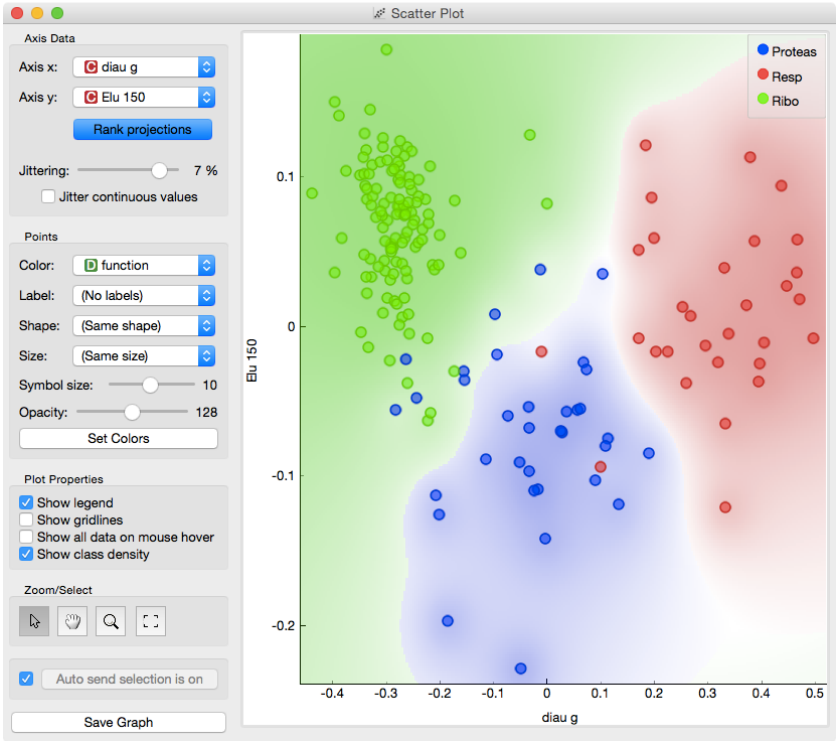


Orange workflows often start with a File widget. The brown-selected data set comprises 186 rows (genes) and 81 columns. Out of the 81 columns, 79 contain gene expressions of baker's yeast under various conditions, one column (marked as a "meta attribute") provides gene names, and one column contains the "class" value or gene function.

After you load the data, open the other widgets. In the Scatter Plot widget, select a few data points and watch as they appear in widget Data Table (1). Use a combination of two Scatter Plot widgets, where the second scatterplot shows a detail from a smaller region selected in the first scatterplot.

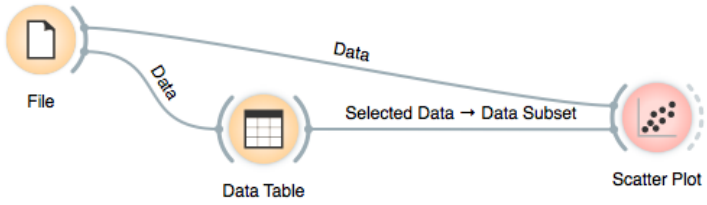
Following is more of a side note, but it won't hurt. Namely, the scatter plot for a pair of random features does not provide much information on gene function. Does this change with a different choice of feature pairs in the visualization? Rank projections (the button on the top left of the Scatter Plot widget) can help you find a good feature pair. How do you think this works? Could the suggested pairs of features be useful to a biologist?



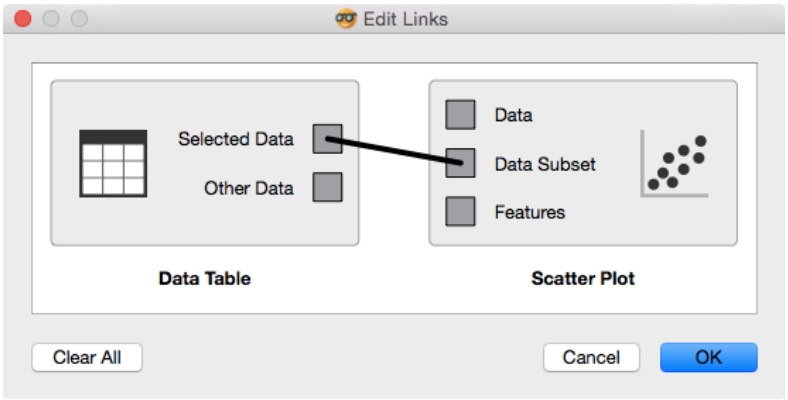


We can connect the output of the Data Table widget to the Scatter Plot widget to highlight the chosen data instances (rows) in the scatterplot.

In this workflow, we have switched on the option "Show channel names between widgets" in File→Preferences.

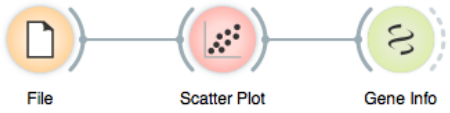


How does Orange distinguish between the primary data source and the data selection? It uses the first connected signal as the entire data set and the second one as its subset. To make changes or to check what is happening under the hood, double click on the line connecting the two widgets.



Orange comes with a basic set of widgets for data input, preprocessing, visualization and modeling. For other tasks, like text mining, network analysis, and bioinformatics, there are add-ons. Check them out by selecting Add-ons… from the options menu.

The rows in the data set we are exploring in this lesson are gene profiles. We can use the Gene Info widget from the Bioinformatics add-on to get more information on the genes we selected in any of the Orange widgets.

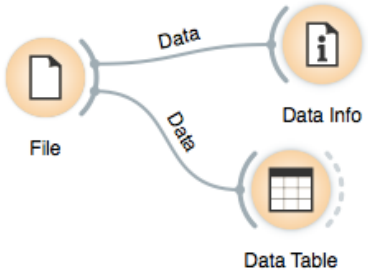


Lesson 2: Basic Data Exploration

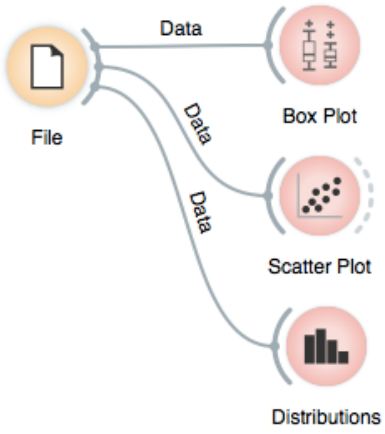
Let us consider another problem, this time from clinical medicine. We will dig for something interesting in the data and explore it a bit with visualization widgets. You will get to know Orange better, and also learn about several interesting visualizations.

We will start with an empty canvas; to clean it from our previous lesson, use either File→New or select all the widgets and remove them (use the backspace/delete key, or Cmd-backspace if you are on Mac).

Now again, add the File widget and open another documentation data set: heart_disease. How does the data look like?

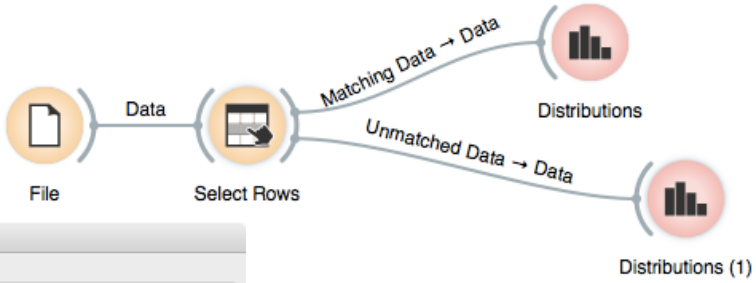


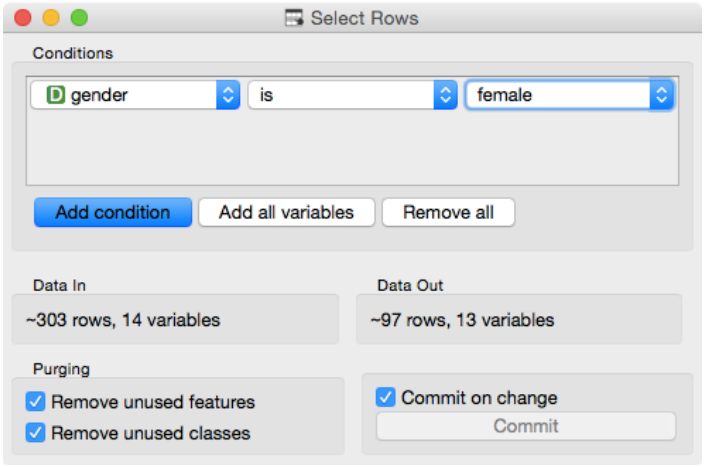
Let us check whether common visualizations tell us anything interesting. (Hint: look for gender differences. These are always interesting and occasionally even real.)



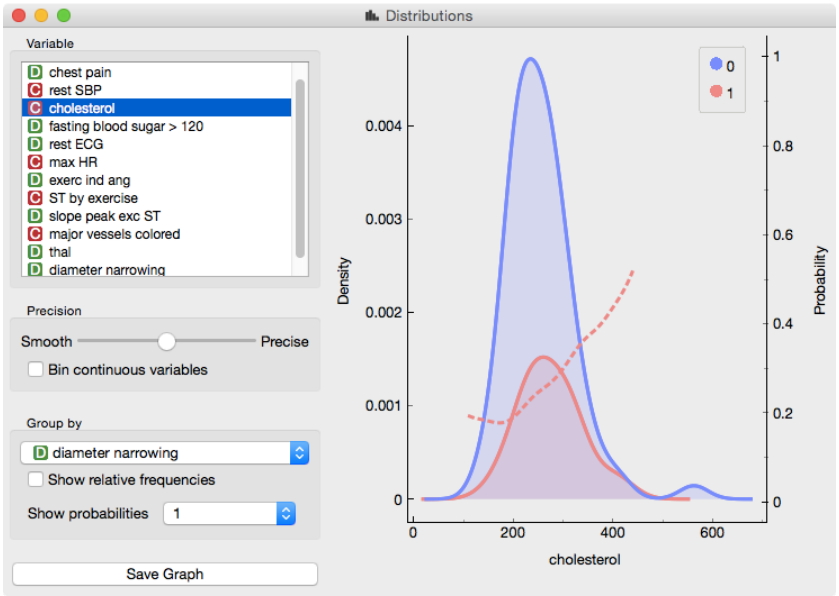
The two Distributions widgets get different data: the upper gets the selected rows and the lower gets the others. Double-click the connection between the widgets to access setup dialog, as you've learned in the previous lesson.

Data can also be split by the value of features — in this case gender — and analyze it separately.



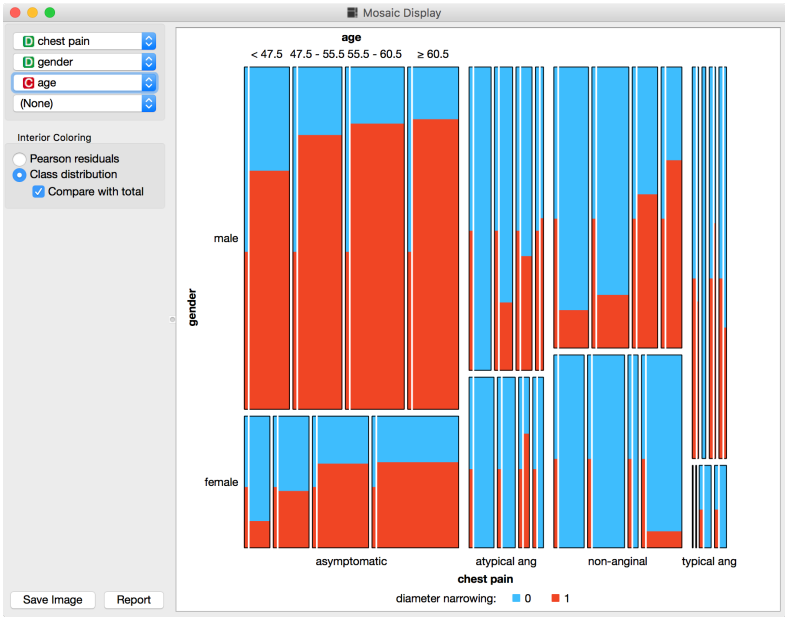


In the Select Rows widget, we select the female patients. You can also add other conditions. Selection of data instances provides a powerful combination with visualization of data distribution. Try having at least two widgets open at the same time and explore the data.



There are two less known — but great — visualizations for observing interactions between features.

You can play with the widget by trying different combinations of 1-4 features.

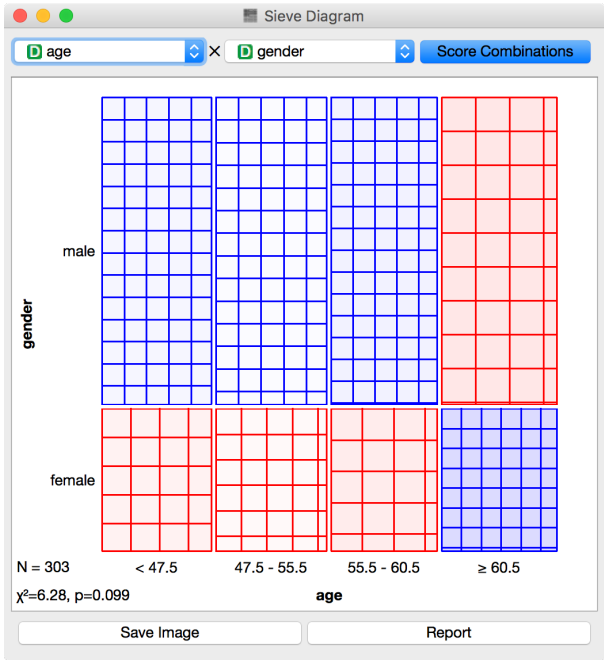


Mosaic display shows a rectangle split into columns with widths reflecting the prevalence of different types of chest pain. Each column is then further split vertically according to gender distributions within the column. The resulting rectangles are split again horizontally according to age group sizes. Within the resulting bars, the red and blue areas represent the outcome distribution for each group and the tiny strip to the left of each shows the overall distribution.

What can you read from this diagram?

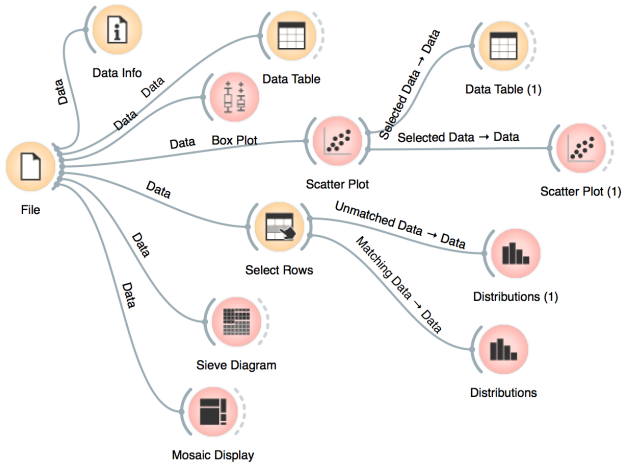
Another visualization, Sieve diagram also splits a rectangle horizontally and vertically, but with independent cuts, so the areas correspond to the expected number of data instances if the observed variables were independent. For instance, 1/4 of patients are older than 60, and 1/3 of patients are female, so the area of the bottom right rectangle is 1/12 of the total area. With roughly 300 patients, we would expect 1/12 × 300 = 25 older women in our data. As a matter of fact, there are 34. Sieve diagram shows the difference between the expected and the observed frequencies by the grid density and the color of the field.

See the Score Combinations button? Guess what it does? And how it scores the combinations? (Hint: there are some Greek letters at the bottom of the widget.)



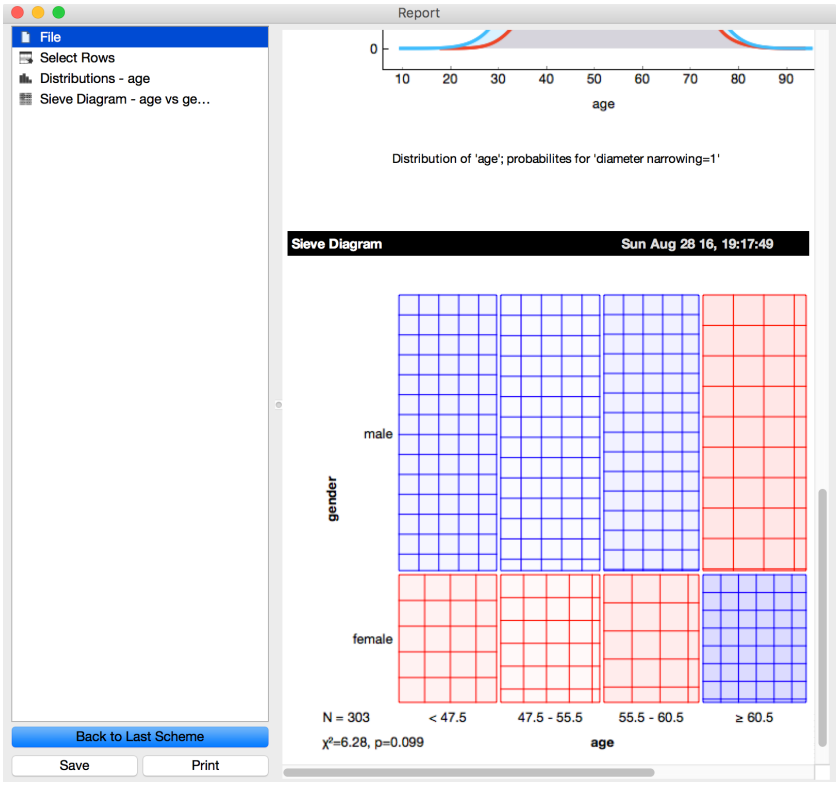
Lesson 3: Saving Your Work

If you followed the instructions so far — except for those about removing widgets — your workflow may look like this.



You can save it (File→Save) and share it with your colleagues. Just don't forget to put the data files in the same directory as the file with the workflow.

Widgets also have a Report button, which you can use to keep a log of your analysis. When you find something interesting, like an



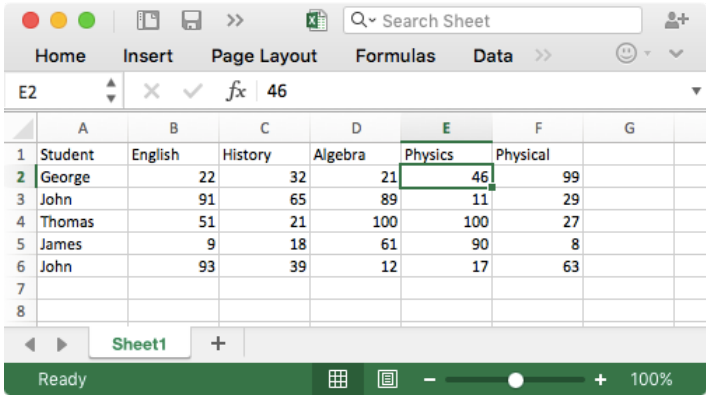
unexpected Sieve Diagram, just click Report and the graph will be added to your log. You can also add reports from the widgets on the path to this one, to make sure you don't forget anything relevant.

Clicking on a part of the report also allows you to add a comment.

You can save the report as HTML or PDF, or a report file that includes all workflows related report items that you can later open in Orange. In this way, you and your colleagues can reproduce your analysis results.

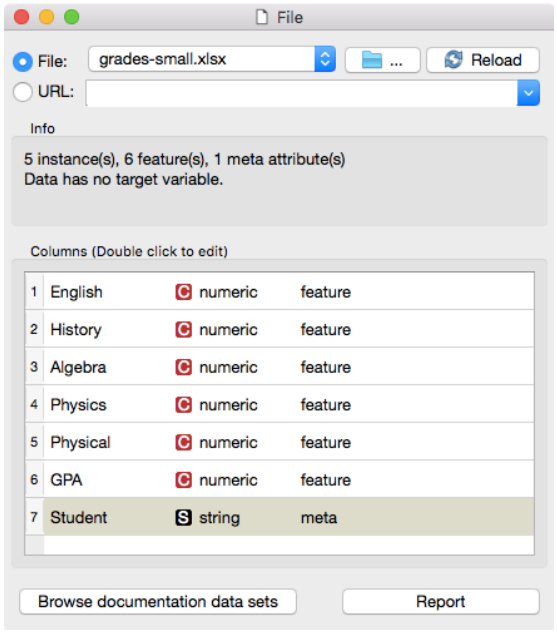
Lesson 4: Loading Your Own Data Set

The data sets we have worked with in previous lessons come with Orange installation. Orange can read data from spreadsheet file formats which include tab and comma separated and Excel files. Let us prepare a data set (with school subjects and grades) in Excel and save it on a local disk.



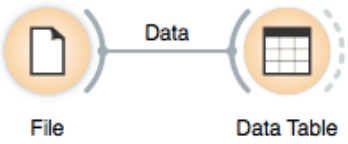
	A	B	C	D	E	F	G
1	Student	English	History	Algebra	Physics	Physical	
2	George	22	32	21	46	99	
3	John	91	65	89	11	29	
4	Thomas	51	21	100	100	27	
5	James	9	18	61	90	8	
6	John	93	39	12	17	63	
7							
8							

In Orange, we can use the File widget to load this data set.

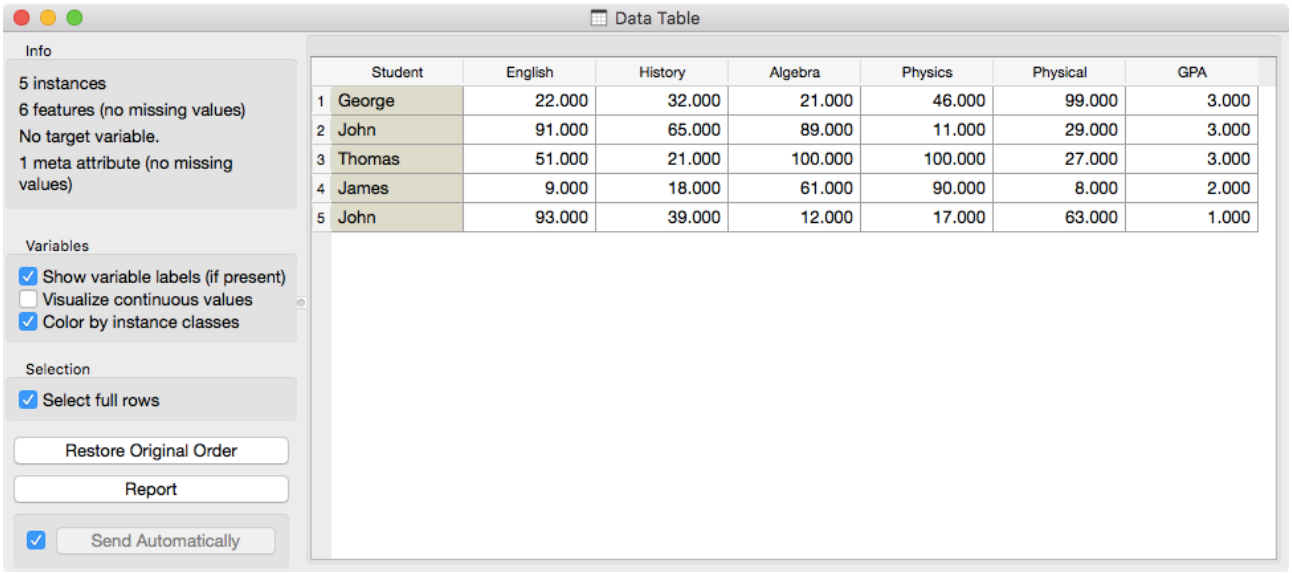


Looks ok. Orange has correctly guessed that student names are character strings and that this column in the data set is special, meant to provide additional information and not to be used for any kind of modeling (more about this in the coming lectures). All other columns are numeric features.

It is always good to check if all the data was read correctly. We can connect our File widget with the Data Table widget,



and double click on the Data Table to see the data in the spreadsheet format.



	Student	English	History	Algebra	Physics	Physical	GPA
1	George	22.000	32.000	21.000	46.000	99.000	3.000
2	John	91.000	65.000	89.000	11.000	29.000	3.000
3	Thomas	51.000	21.000	100.000	100.000	27.000	3.000
4	James	9.000	18.000	61.000	90.000	8.000	2.000
5	John	93.000	39.000	12.000	17.000	63.000	1.000

Nice, everything is here.

Instead of using Excel, we could also use Google Sheets, a free on-line spreadsheet alternative. Then, instead of finding the file on the local disk, we would enter its URL address to the File widget URL entry box.

There is more to input data formatting and loading. Say, we can define the type and kind of the data column, specify that the column is actually a web address of an image, and more. But enough for the first day. If you would really like to dive in for more, check out the documentation page on Loading your Data, or a video on this subject.