Merge Data

Merges two datasets, based on values of selected attributes.

Inputs

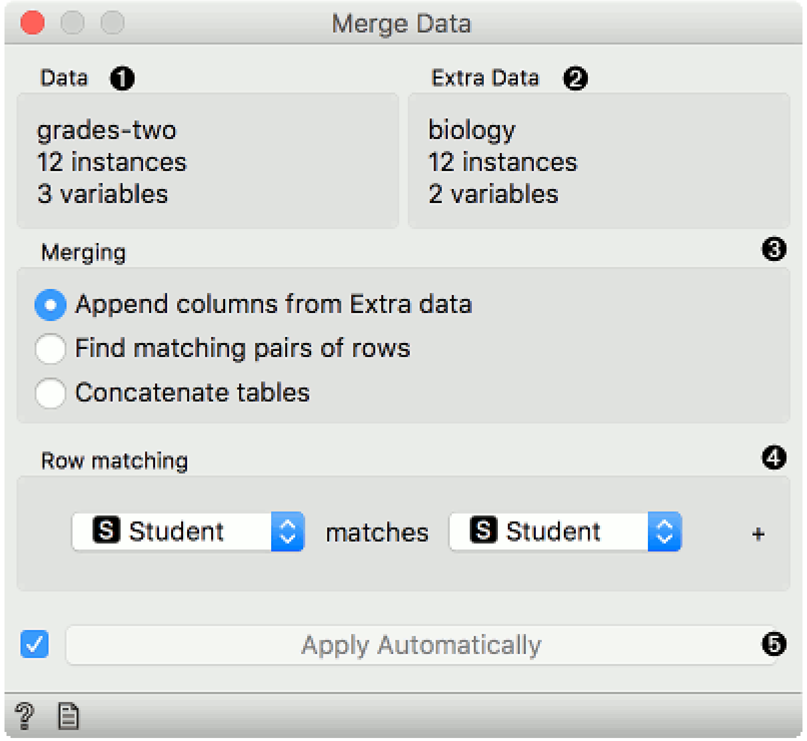
- Data: input dataset
- Extra Data: additional dataset

Outputs

- Data: dataset with features added from extra data

The **Merge Data** widget is used to horizontally merge two datasets, based on the values of selected attributes (columns). In the input, two datasets are required, data and extra data. Rows from the two data sets are matched by the values of pairs of attributes, chosen by the user. The widget produces one output. It corresponds to the instances from the input data to which attributes (columns) from input extra data are appended.

If the selected attribute pair does not contain unique values (in other words, the attributes have duplicate values), the widget will give a warning. Instead, one can match by more than one attribute. Click on the plus icon to add the attribute to merge on. The final result has to be a unique combination for each individual row.



Merge Data

Data ❶

grades-two
12 instances
3 variables

Extra Data ❷

biology
12 instances
2 variables

Merging ❸

● Append columns from Extra data
○ Find matching pairs of rows
○ Concatenate tables

Row matching ❹

S Student matches S Student +

☑ Apply Automatically ❺

1. Information on main data.
2. Information on data to append.
3. Merging type:
    - **Append columns from Extra Data** outputs all rows from the Data, augmented by the columns in the Extra Data. Rows without matches are retained, even where the data in the extra columns are missing.
    - **Find matching pairs of rows** outputs rows from the Data, augmented by the columns in the Extra Data. Rows without matches are removed from the output.
    - **Concatenate tables** treats both data sources symmetrically. The output is similar to the first option, except that non-matched values from Extra Data are appended at the end.
4. List of attributes from Data input.
5. List of attributes from Extra Data input.
6. Produce a report.

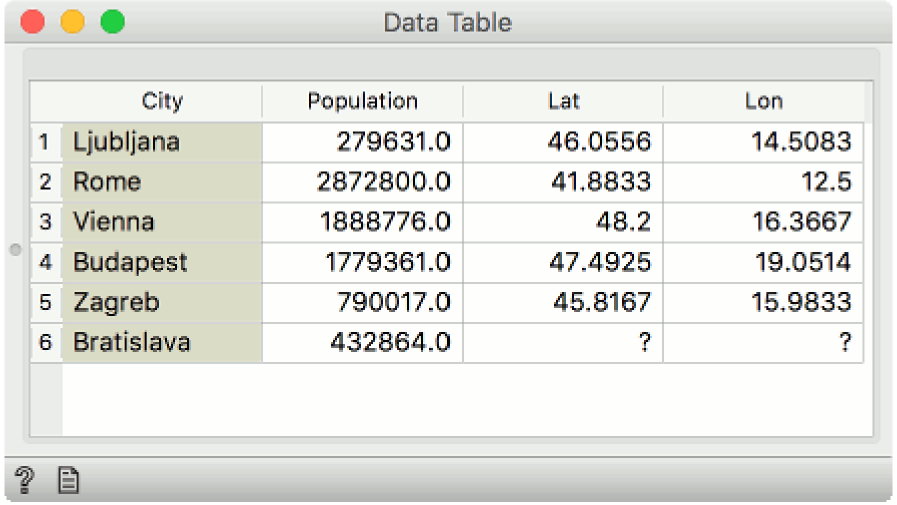
Merging Types

#####Append Columns from Extra Data (left join)

Columns from the Extra Data are added to the Data. Instances with no matching rows will have missing values added.

For example, the first table may contain city names and the second would be a list of cities and their coordinates. Columns with coordinates would then be appended to the data with city names. Where city names cannot be matched, missing values will appear.

In our example, the first Data input contained 6 cities, but the Extra Data did not provide Lat and Lon values for Bratislava, so the fields will be empty.

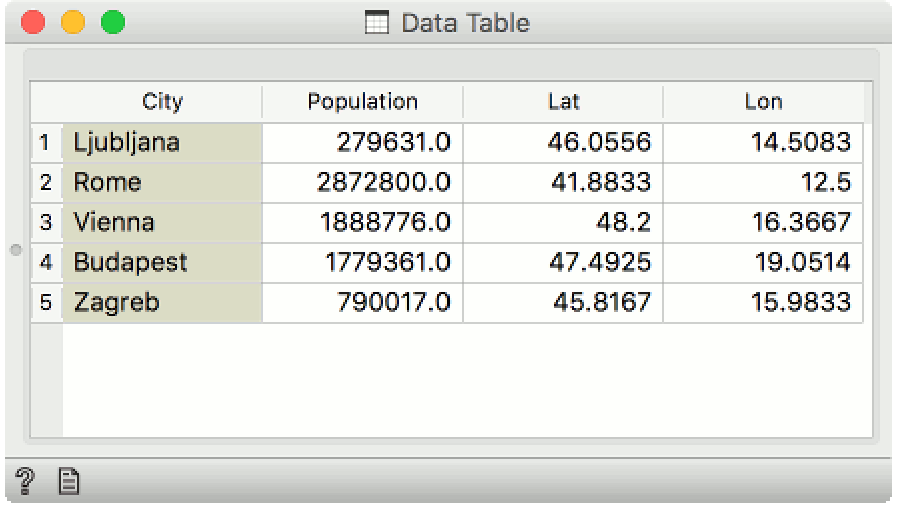


	City	Population	Lat	Lon
1	Ljubljana	279631.0	46.0556	14.5083
2	Rome	2872800.0	41.8833	12.5
3	Vienna	1888776.0	48.2	16.3667
4	Budapest	1779361.0	47.4925	19.0514
5	Zagreb	790017.0	45.8167	15.9833
6	Bratislava	432864.0	?	?

#####Find matching pairs of rows (inner join)

Only those rows that are matched will be present on the output, with the Extra Data columns appended. Rows without matches are removed.

In our example, Bratislava from the Data input did not have Lat and Lon values, while Belgrade from the Extra Data could not be found in the City column we were merging on. Hence both instances are remove - only the intersection of instances is sent to the output.

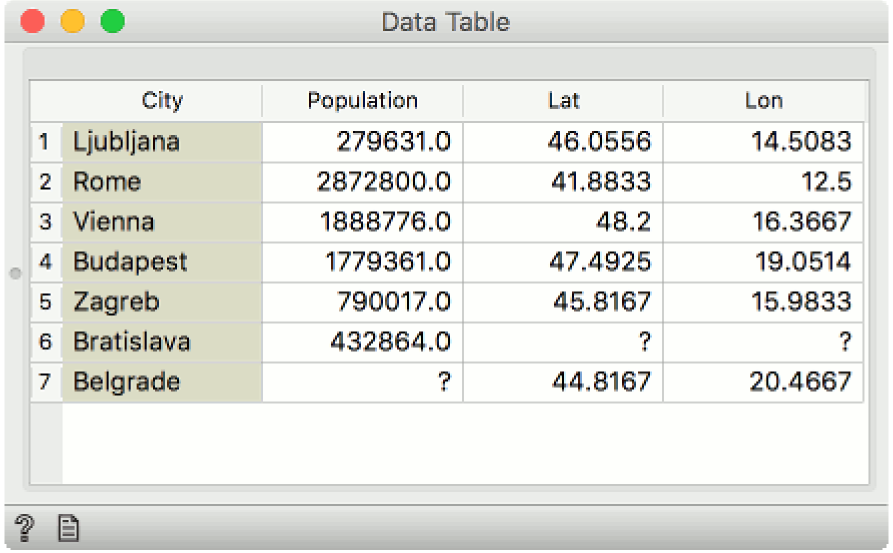


	City	Population	Lat	Lon
1	Ljubljana	279631.0	46.0556	14.5083
2	Rome	2872800.0	41.8833	12.5
3	Vienna	1888776.0	48.2	16.3667
4	Budapest	1779361.0	47.4925	19.0514
5	Zagreb	790017.0	45.8167	15.9833

#####Concatenate tables (outer join)

The rows from both the Data and the Extra Data will be present on the output. Where rows cannot be matched, missing values will appear.

In our example, both Bratislava and Belgrade are now present. Bratislava will have missing Lat and Lon values, while Belgrade will have a missing Population value.



Data Table

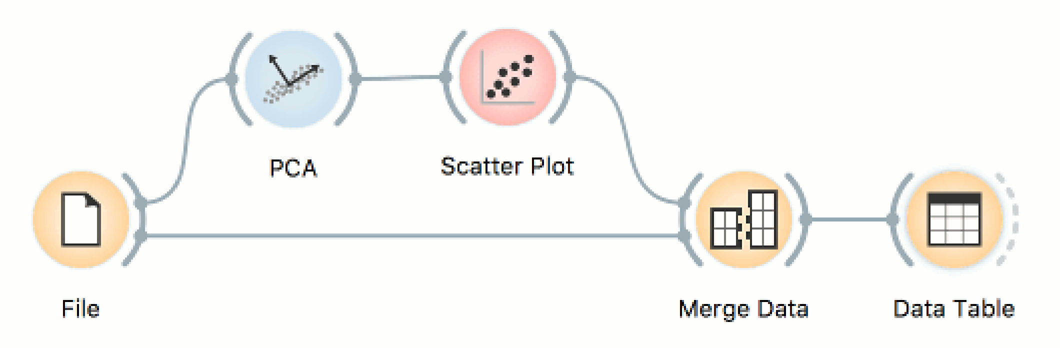
	City	Population	Lat	Lon
1	Ljubljana	279631.0	46.0556	14.5083
2	Rome	2872800.0	41.8833	12.5
3	Vienna	1888776.0	48.2	16.3667
4	Budapest	1779361.0	47.4925	19.0514
5	Zagreb	790017.0	45.8167	15.9833
6	Bratislava	432864.0	?	?
7	Belgrade	?	44.8167	20.4667

#####Row index

Data will be merged in the same order as they appear in the table. Row number 1 from the Data input will be joined with row number 1 from the Extra Data input. Row numbers are assigned by Orange based on the original order of the data instances.

#####Instance ID

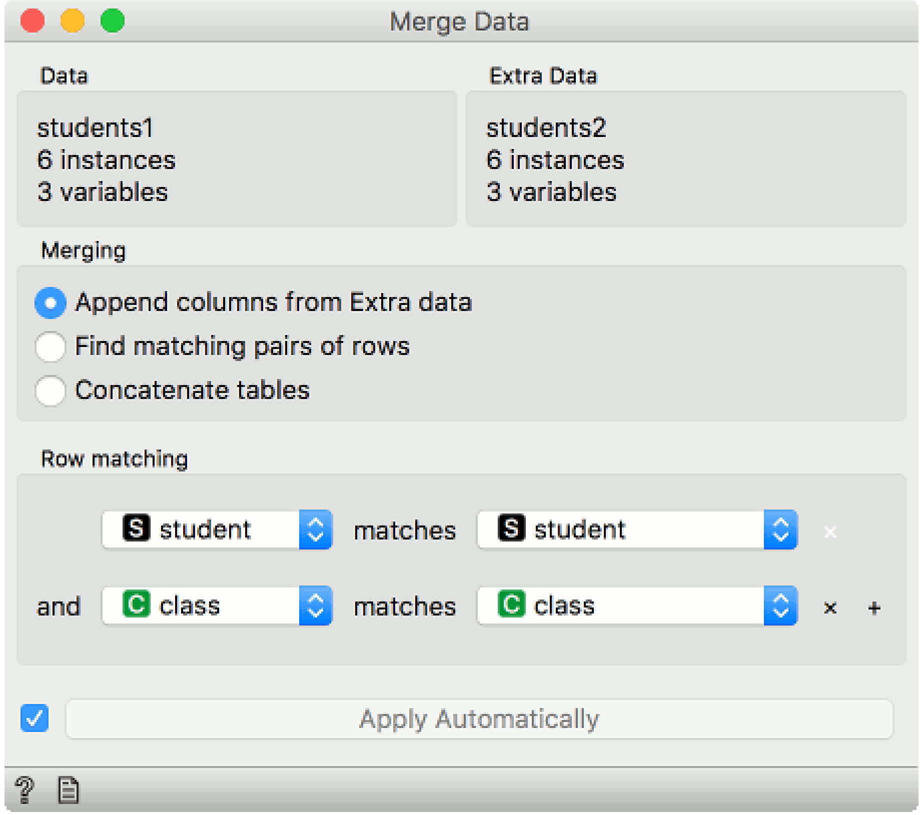
This is a more complex option. Sometimes, data in transformed in the analysis and the domain is no longer the same. Nevertheless, the original row indices are still present in the background (Orange remembers them). In this case one can merge on instance ID. For example if you transformed the data with PCA, visualized it in the Scatter Plot, selected some data instances and now you wish to see the original information of the selected subset. Connect the output of Scatter Plot to Merge Data, add the original data set as Extra Data and merge by Instance ID.



#####Merge by two or more attributes

Sometimes our data instances are unique with respect to a combination of columns, not a single column. To merge by more than a single column, add the *Row matching* condition by pressing plus next to the matching condition. To remove it, press the x.

In the below example, we are merging by *student* column and *class* column.


Merge Data

Data
students1
6 instances
3 variables

Extra Data
students2
6 instances
3 variables

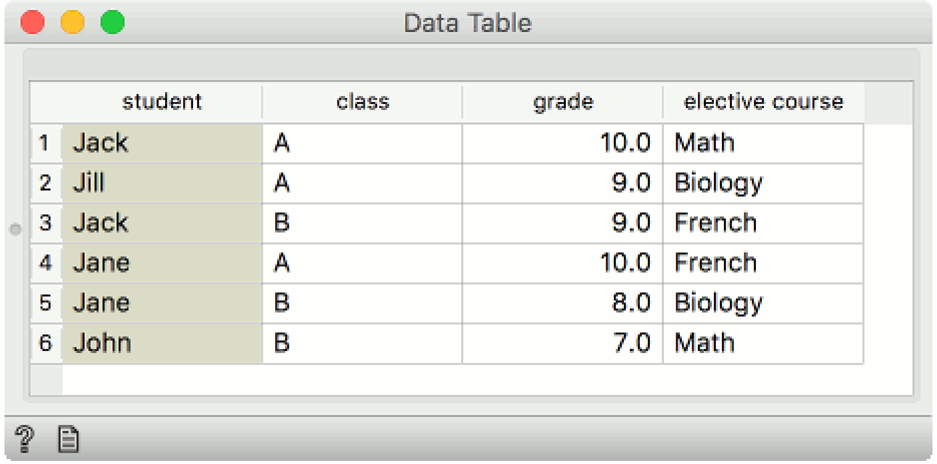
Merging
● Append columns from Extra data
○ Find matching pairs of rows
○ Concatenate tables

Row matching
S student matches S student ×
and C class matches C class × +

☑ Apply Automatically

Say we have two data sets with student names and the class they're in. The first data set has students' grades and the second on the elective course they have chosen. Unfortunately, there are two Jacks in our data, one from class A and the other from class B. Same for Jane.

To distinguish between the two, we can match rows on both, the student's name and her class.

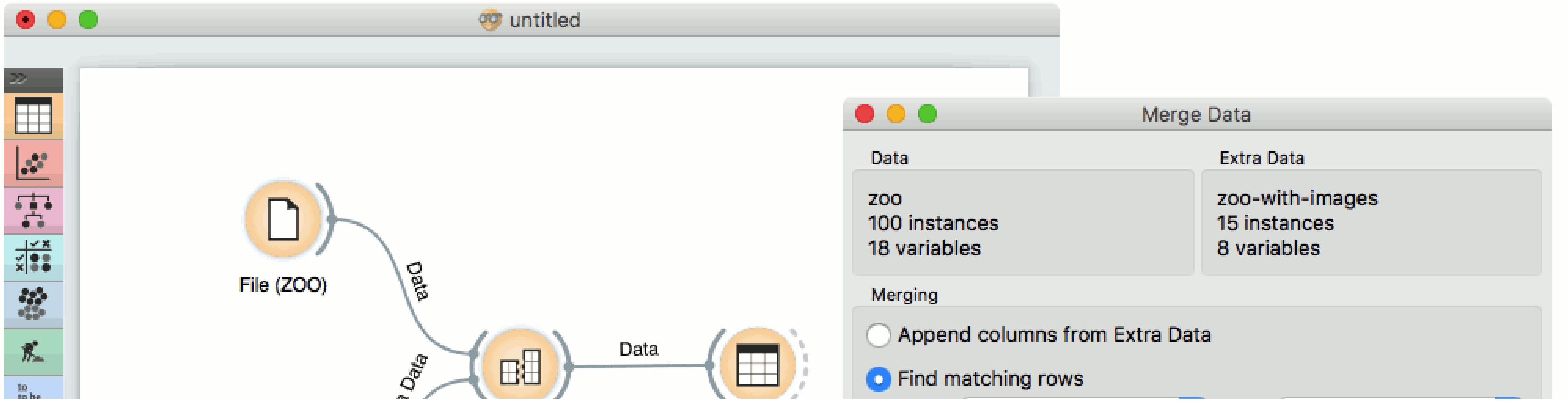


Data Table

	student	class	grade	elective course
1	Jack	A	10.0	Math
2	Jill	A	9.0	Biology
3	Jack	B	9.0	French
4	Jane	A	10.0	French
5	Jane	B	8.0	Biology
6	John	B	7.0	Math

Examples

Merging two datasets results in appending new attributes to the original file, based on a selected common attribute. In the example below, we wanted to merge the **zoo.tab** file containing only factual data with zoo-with-images.tab containing images. Both files share a common string attribute *names*. Now, we create a workflow connecting the two files. The *zoo.tab* data is connected to **Data** input of the **Merge Data** widget, and the *zoo-with-images.tab* data to the **Extra Data** input. Outputs of the **Merge Data** widget is then connected to the Data Table widget. In the latter, the **Merged Data**channels are shown, where image attributes are added to the original data.



Merge Data

Data

zoo
100 instances
18 variables

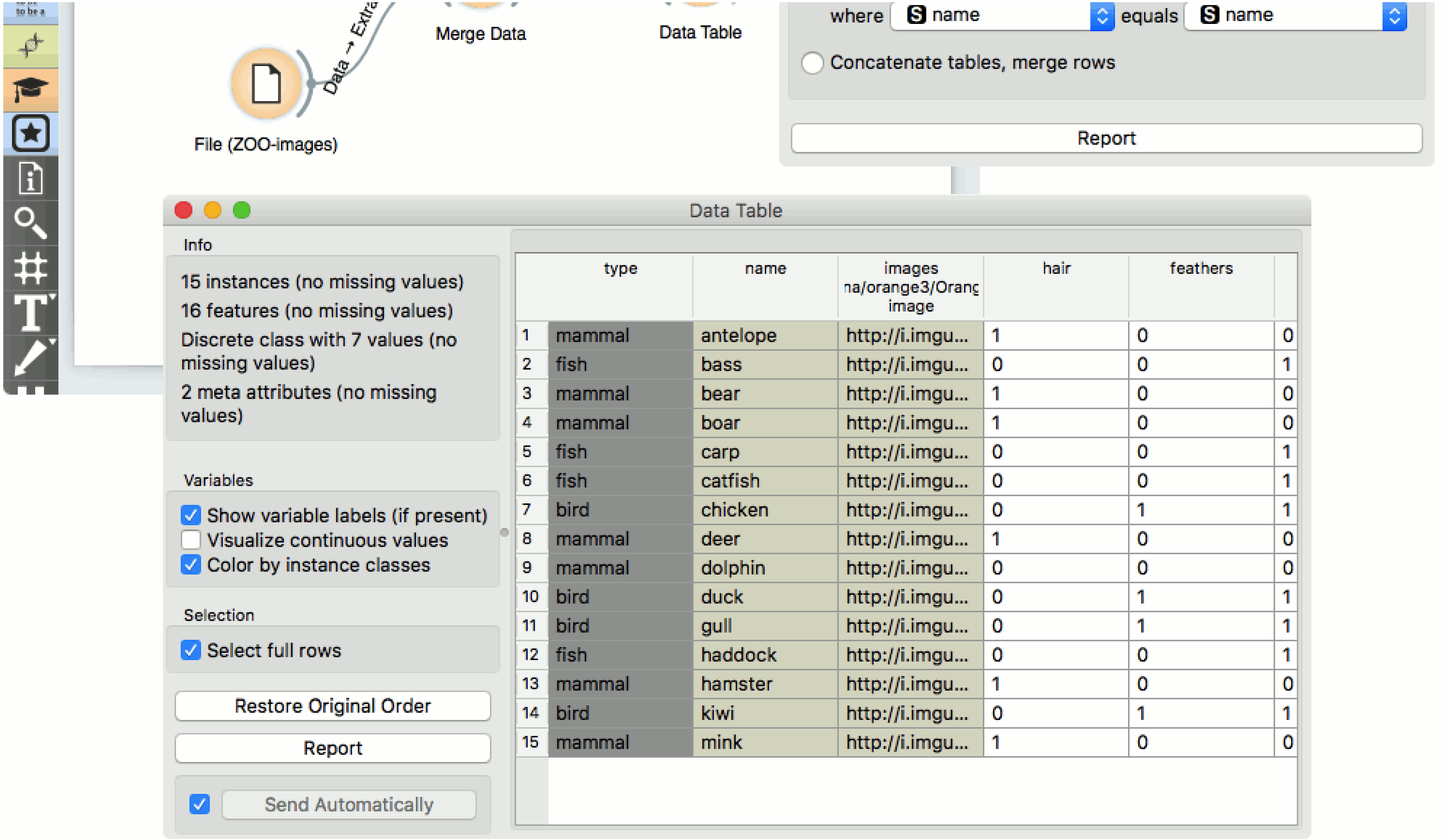
Extra Data

zoo-with-images
15 instances
8 variables

Merging

○ Append columns from Extra Data

● Find matching rows



where S name equals S name

○ Concatenate tables, merge rows

Report

Data Table

Info

15 instances (no missing values)
16 features (no missing values)
Discrete class with 7 values (no missing values)
2 meta attributes (no missing values)

Variables

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

Selection

☑ Select full rows

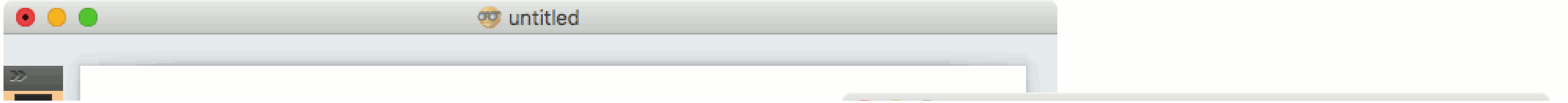
Restore Original Order

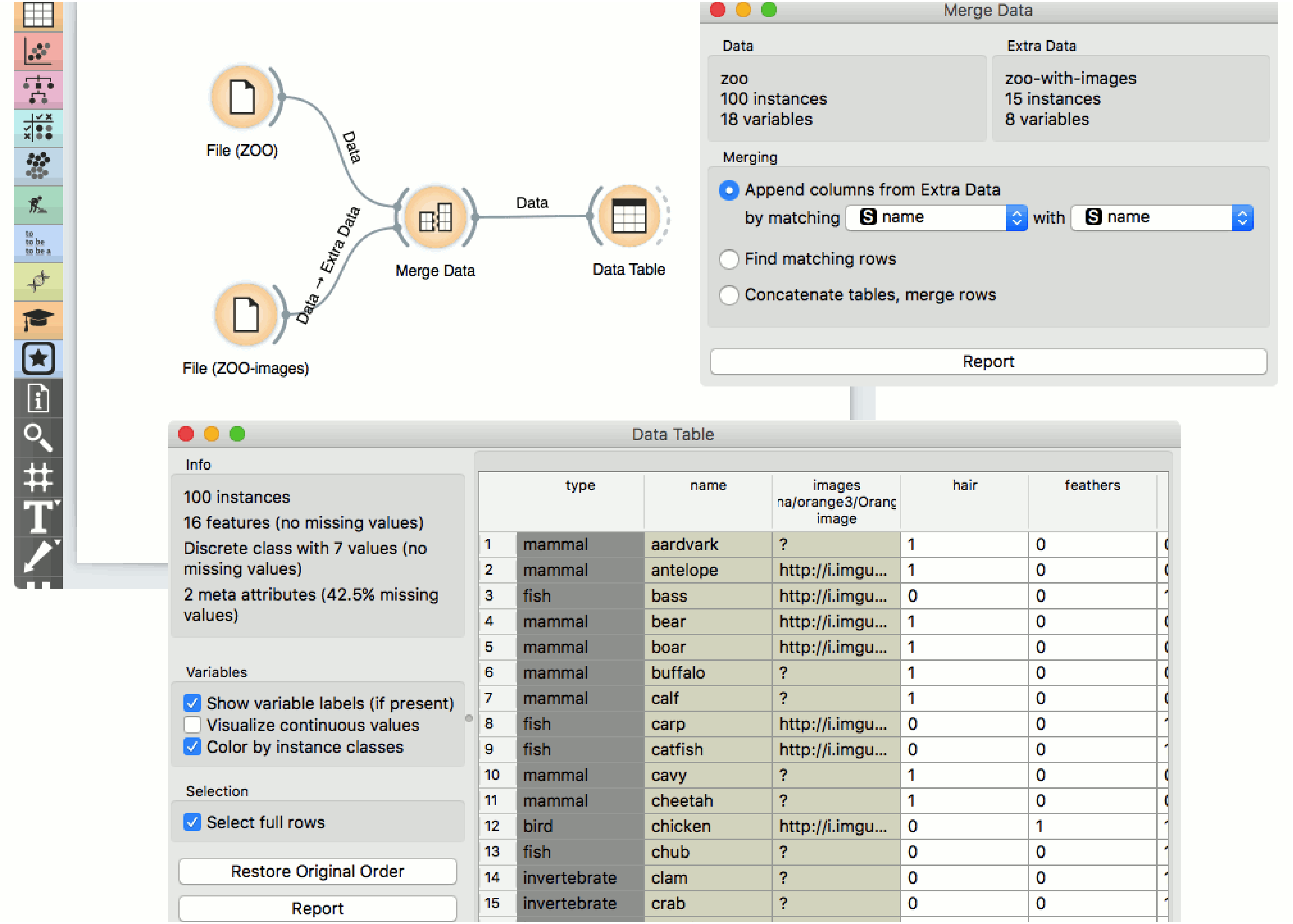
Report

☑ Send Automatically

	type	name	images na/orange3/Orang image	hair	feathers	
1	mammal	antelope	http://i.imgu...	1	0	0
2	fish	bass	http://i.imgu...	0	0	1
3	mammal	bear	http://i.imgu...	1	0	0
4	mammal	boar	http://i.imgu...	1	0	0
5	fish	carp	http://i.imgu...	0	0	1
6	fish	catfish	http://i.imgu...	0	0	1
7	bird	chicken	http://i.imgu...	0	1	1
8	mammal	deer	http://i.imgu...	1	0	0
9	mammal	dolphin	http://i.imgu...	0	0	0
10	bird	duck	http://i.imgu...	0	1	1
11	bird	gull	http://i.imgu...	0	1	1
12	fish	haddock	http://i.imgu...	0	0	1
13	mammal	hamster	http://i.imgu...	1	0	0
14	bird	kiwi	http://i.imgu...	0	1	1
15	mammal	mink	http://i.imgu...	1	0	0

The case where we want to include all instances in the output, even those where no match by attribute *names* was found, is shown in the following workflow.


untitled



Merge Data

Data

zoo
100 instances
18 variables

Extra Data

zoo-with-images
15 instances
8 variables

Merging

● Append columns from Extra Data
by matching name with name

○ Find matching rows

○ Concatenate tables, merge rows

Report

Data Table

Info

100 instances

16 features (no missing values)

Discrete class with 7 values (no missing values)

2 meta attributes (42.5% missing values)

Variables

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

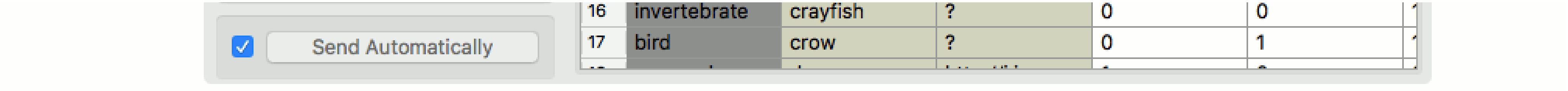
Selection

☑ Select full rows

Restore Original Order

Report

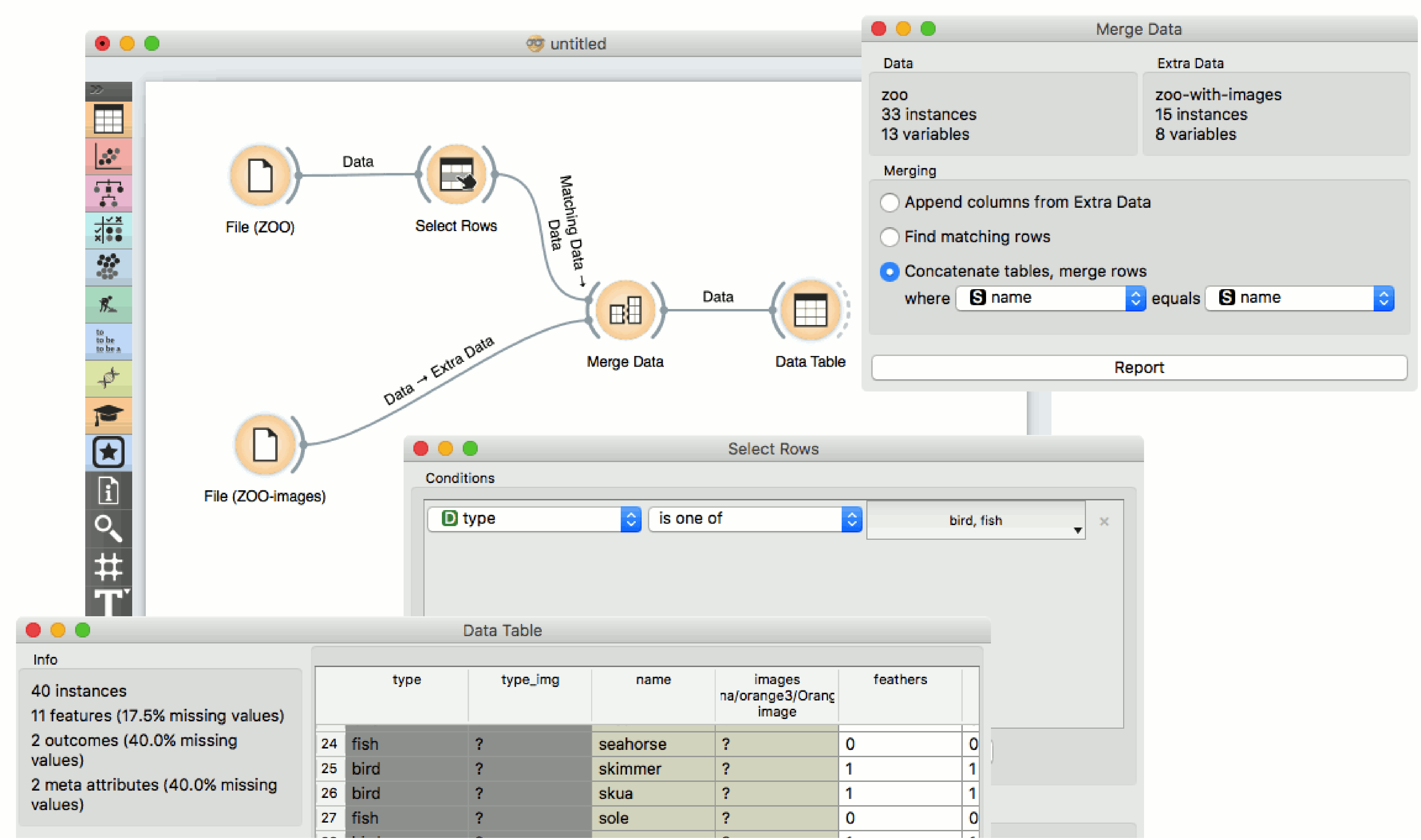
	type	name	images na/orange3/Orang image	hair	feathers
1	mammal	aardvark	?	1	0
2	mammal	antelope	http://i.imgu...	1	0
3	fish	bass	http://i.imgu...	0	0
4	mammal	bear	http://i.imgu...	1	0
5	mammal	boar	http://i.imgu...	1	0
6	mammal	buffalo	?	1	0
7	mammal	calf	?	1	0
8	fish	carp	http://i.imgu...	0	0
9	fish	catfish	http://i.imgu...	0	0
10	mammal	cavy	?	1	0
11	mammal	cheetah	?	1	0
12	bird	chicken	http://i.imgu...	0	1
13	fish	chub	?	0	0
14	invertebrate	clam	?	0	0
15	invertebrate	crab	?	0	0



☑ Send Automatically

16	invertebrate	crayfish	?	0	0
17	bird	crow	?	0	1

The third type of merging is shown in the next workflow. The output consists of both inputs, with unknown values assigned where no match was found.



Merge Data

Data

zoo
33 instances
13 variables

Extra Data

zoo-with-images
15 instances
8 variables

Merging

○ Append columns from Extra Data

○ Find matching rows

● Concatenate tables, merge rows

where name equals name

Report

Select Rows

Conditions

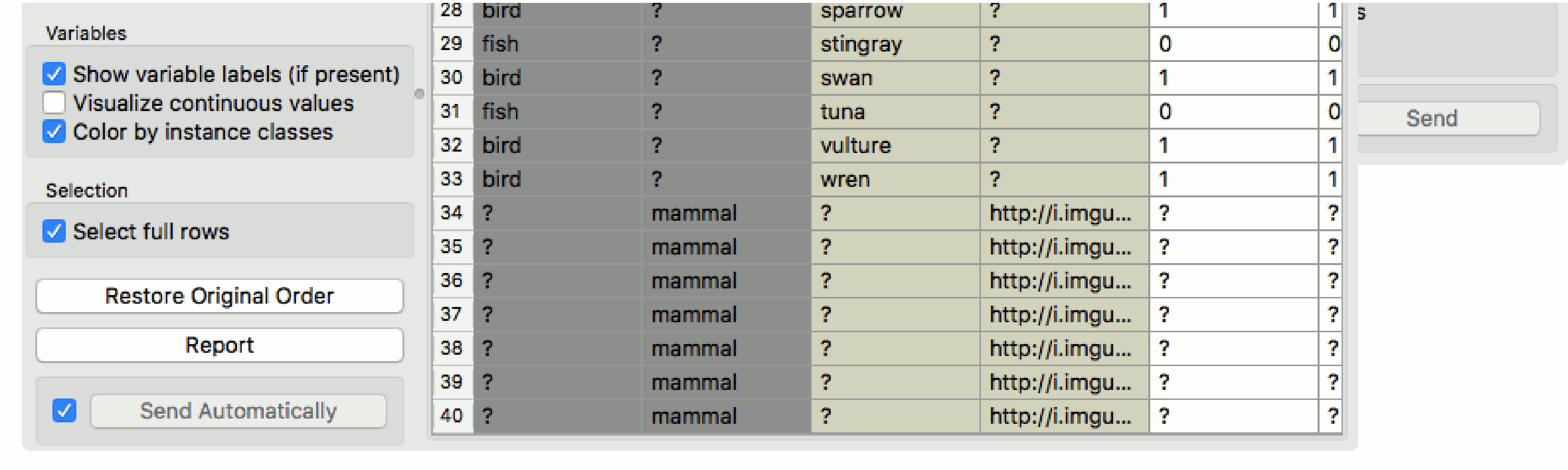
type | is one of | bird, fish

Data Table

Info

40 instances
11 features (17.5% missing values)
2 outcomes (40.0% missing values)
2 meta attributes (40.0% missing values)

	type	type_img	name	images na/orange3/Orang image	feathers	
24	fish	?	seahorse	?	0	0
25	bird	?	skimmer	?	1	1
26	bird	?	skua	?	1	1
27	fish	?	sole	?	0	0



Variables

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

Selection

☑ Select full rows

☑ Send Automatically

28	bird	?	sparrow	?	1	1
29	fish	?	stingray	?	0	0
30	bird	?	swan	?	1	1
31	fish	?	tuna	?	0	0
32	bird	?	vulture	?	1	1
33	bird	?	wren	?	1	1
34	?	mammal	?	http://i.imgu...	?	?
35	?	mammal	?	http://i.imgu...	?	?
36	?	mammal	?	http://i.imgu...	?	?
37	?	mammal	?	http://i.imgu...	?	?
38	?	mammal	?	http://i.imgu...	?	?
39	?	mammal	?	http://i.imgu...	?	?
40	?	mammal	?	http://i.imgu...	?	?