KNIME TUTORIAL

What is KNIME?

- KNIME = Konstanz Information Miner
- Developed at University of Konstanz in Germany
- Desktop version available free of charge (Open Source)
- Modular platform for building and executing **workflows** using predefined components, called **nodes**
- Functionality available for tasks such as **standard data mining, data analysis** and **data manipulation**
- Extra features and functionalities available in KNIME by extensions
- Written in Java based on the Eclipse SDK platform

KNIME resources

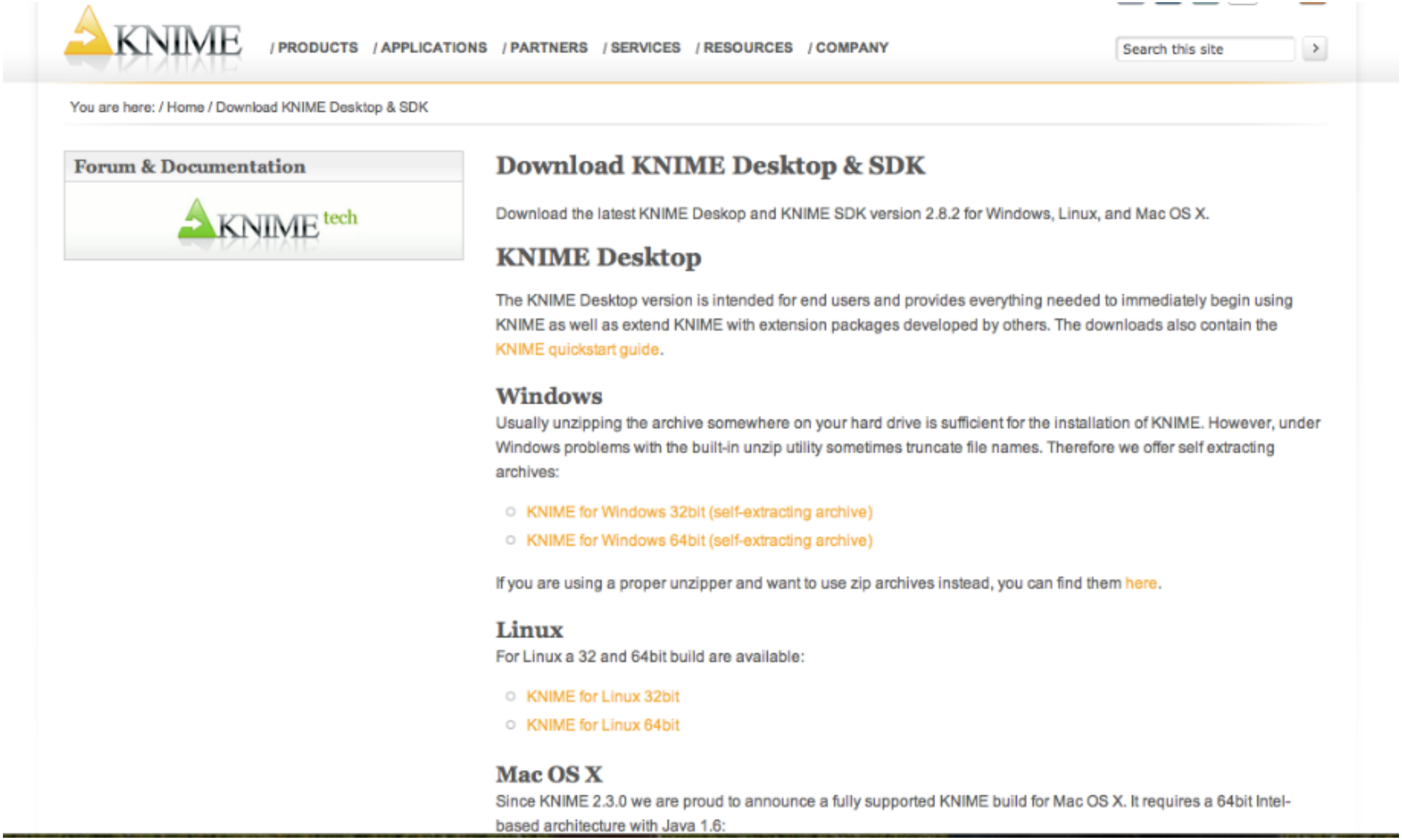
- Web pages containing documentation
- www.knime.org - tech.knime.org – tech.knime.org
- installation-0
- Downloads
  - knime.org/download-desktop
- Community forum
  - tech.knime.org/forum
- Books and white papers
  - knime.org/node/33079

Installation and updates

- Download and unzip KNIME
  - No further setup required
  - Additional nodes after first launch

- Workflows and data are stored in a ***workspace***

- New software (nodes) from update sites
  - http://tech.knime.org/update/community-contributions/realease



You are here: / Home / Download KNIME Desktop & SDK

Download KNIME Desktop & SDK

Download the latest KNIME Deskop and KNIME SDK version 2.8.2 for Windows, Linux, and Mac OS X.

KNIME Desktop

The KNIME Desktop version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others. The downloads also contain the KNIME quickstart guide.

Windows

Usually unzipping the archive somewhere on your hard drive is sufficient for the installation of KNIME. However, under Windows problems with the built-in unzip utility sometimes truncate file names. Therefore we offer self extracting archives:

- KNIME for Windows 32bit (self-extracting archive)
- KNIME for Windows 64bit (self-extracting archive)

If you are using a proper unzipper and want to use zip archives instead, you can find them here.

Linux

For Linux a 32 and 64bit build are available:

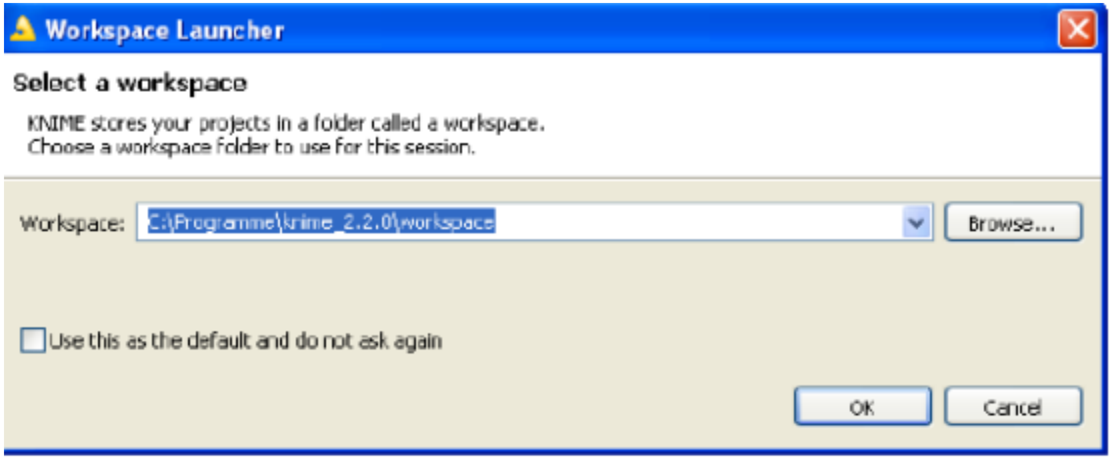
- KNIME for Linux 32bit
- KNIME for Linux 64bit

Mac OS X

Since KNIME 2.3.0 we are proud to announce a fully supported KNIME build for Mac OS X. It requires a 64bit Intel-based architecture with Java 1.6:

Workspace

- The workspace is the directory where all your workflows and preferences are saved in the next KNIME session.
- The workspace directory can be located anywhere on your hard-disk.
- By default, the workspace directory is "**[KNIME]\workspace**". But, you can change it, by changing the path requested at the beginning, before starting the KNIME working session.



Workspace Launcher

Select a workspace

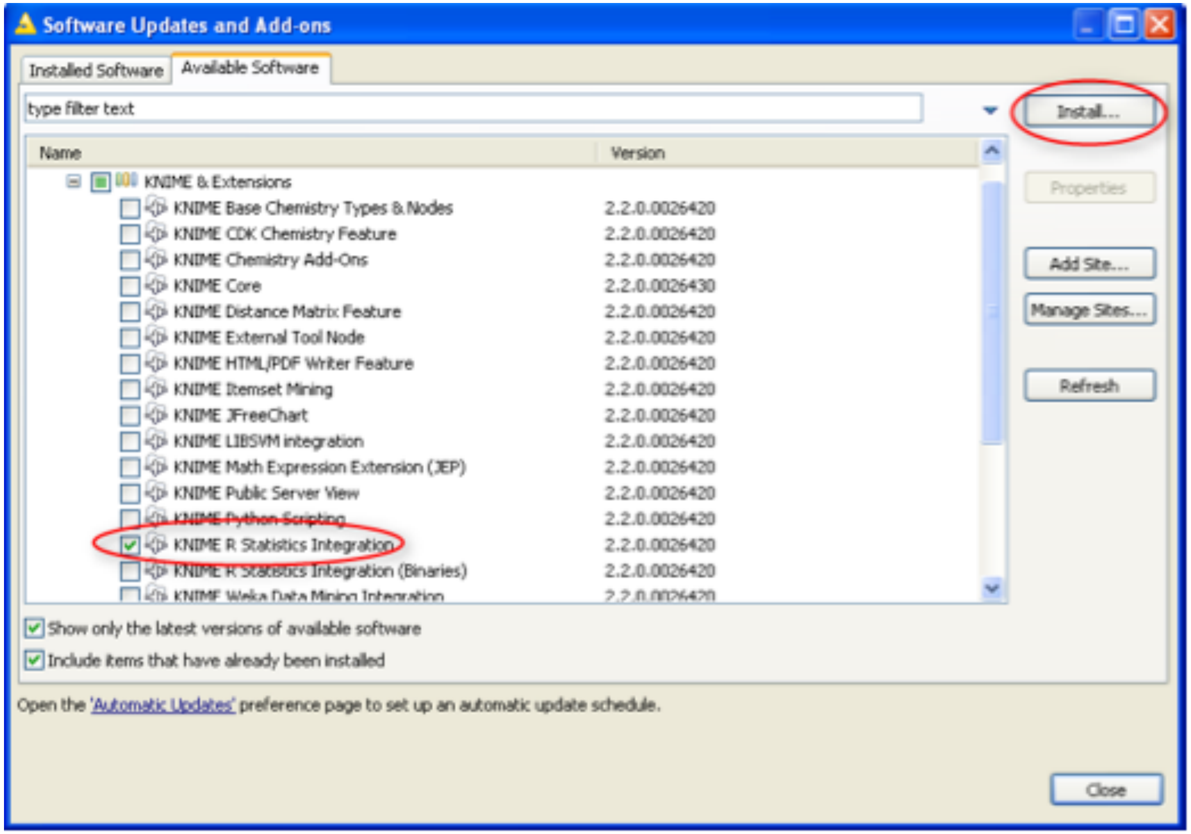
KNIME stores your projects in a folder called a workspace.
Choose a workspace folder to use for this session.

Workspace: C:\Programme\knime_2.2.0\workspace

□ Use this as the default and do not ask again

Download Extensions

- From the Top Menu, select **Help -> Software Updates**
- In the “Software Updates” window, select Tab **Available Software**
- Open the sites and **select the extensions**
- Click the **Install** button on the top right
- Restart KNIME
- In the **Node Repository** you can see the new nodes



Name	Version
⊟ ■ KNIME & Extensions	
□ KNIME Base Chemistry Types & Nodes	2.2.0.0026420
□ KNIME CDK Chemistry Feature	2.2.0.0026420
□ KNIME Chemistry Add-Ons	2.2.0.0026420
□ KNIME Core	2.2.0.0026430
□ KNIME Distance Matrix Feature	2.2.0.0026420
□ KNIME External Tool Node	2.2.0.0026420
□ KNIME HTML/PDF Writer Feature	2.2.0.0026420
□ KNIME Itemset Mining	2.2.0.0026420
□ KNIME JFreeChart	2.2.0.0026420
□ KNIME LIBSVM integration	2.2.0.0026420
□ KNIME Math Expression Extension (JEP)	2.2.0.0026420
□ KNIME Public Server View	2.2.0.0026420
□ KNIME Python Scripting	2.2.0.0026420
☑ KNIME R Statistics Integration	2.2.0.0026420
□ KNIME R Statistics Integration (Binaries)	2.2.0.0026420
□ KNIME Weka Data Mining Integration	2.2.0.0026420

☑ Show only the latest versions of available software

☑ Include items that have already been installed

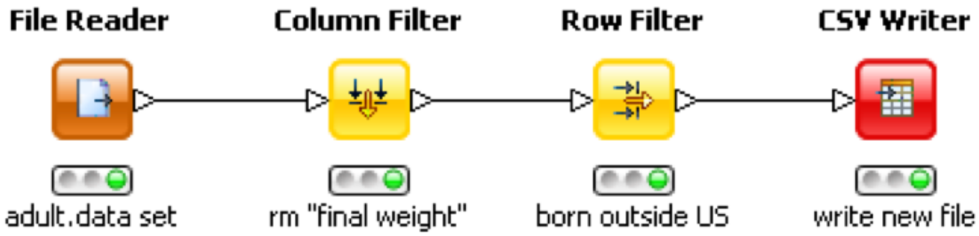
Open the 'Automatic Updates' preference page to set up an automatic update schedule.

What can you do with KNIME?

- **Data manipulation and analysis**
  - File & database I/O, filtering, grouping, joining, ….
- **Data mining / machine learning**
  - WEKA, R, Interactive plotting
- **Scripting Integration**
  - R, Perl, Python, Matlab …
- **Much more**
  - Bioinformatics, text mining and network analysis

KNIME Workflow

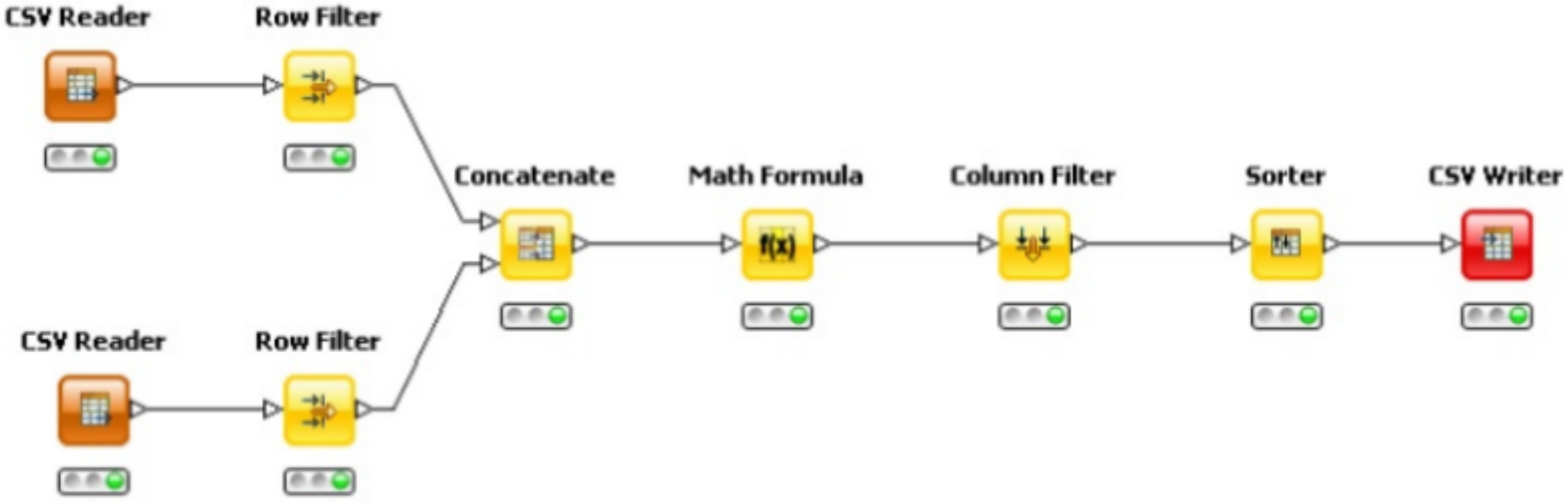
- KNIME does not work with scripts, **it works with workflows.**
- A workflow is an analysis flow, which is the sequence of the analysis steps necessary to reach a given result:
  1. Read data
  2. Clean data
  3. Filter data
  4. Train a model



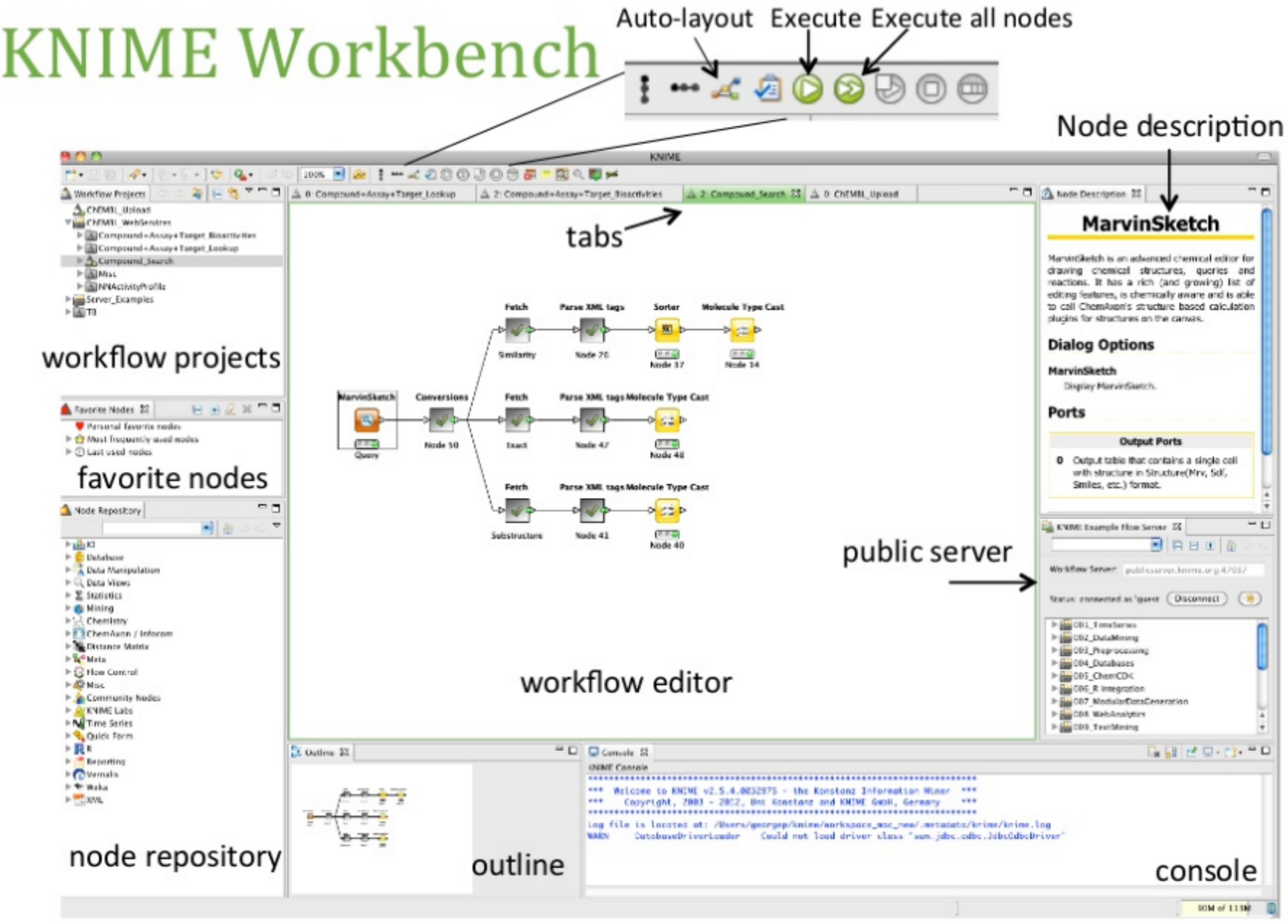
- KNIME implements its workflows **graphically**.
- Each step of the data analysis is executed by a little box, called a **node**.
- **A sequence of nodes makes a workflow.**

Import/export of workflow

- Workflows can be imported and exported as .zip files
  - With or without the underlying data
  - File → Import KNIME workflow...
  - File → Export KNIME workflow...



KNIME Workbench



MarvinSketch

MarvinSketch is an advanced chemical editor for drawing chemical structures, queries and reactions. It has a rich (and growing) list of editing features, is chemically aware and is able to call ChemAxon's structure based calculation plugins for structures on the canvas.

Dialog Options

MarvinSketch

Display MarvinSketch.

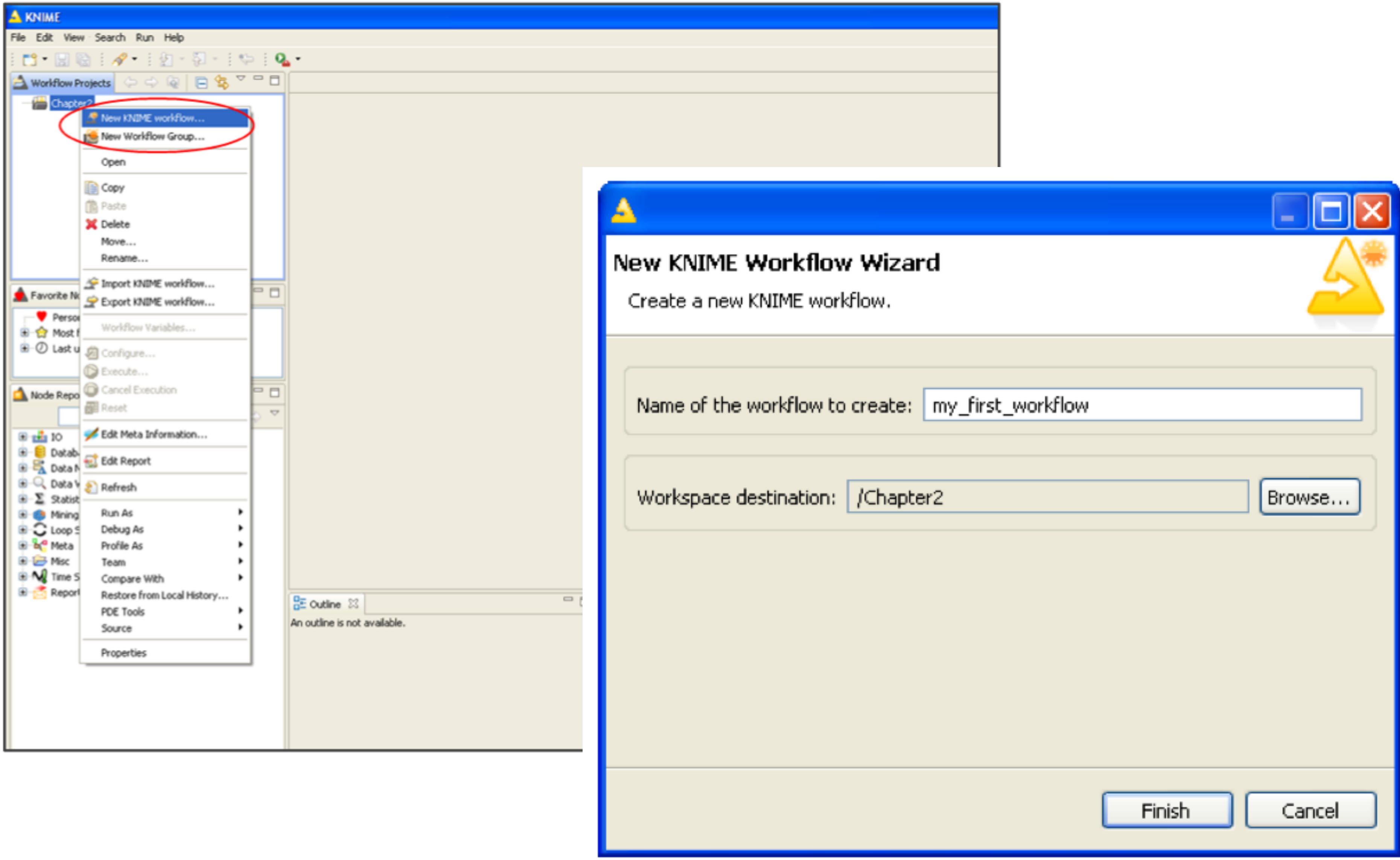
Ports

Output Ports	
0	Output table that contains a single cell with structure in Structure(Mrv, Sdf, Smiles, etc.) format.

KNIME Console

```
**************************************************************************
***  Welcome to KNIME v2.5.4.0032975 - the Konstanz Information Miner  ***
***    Copyright, 2003 - 2012, Uni Konstanz and KNIME GmbH, Germany    ***
**************************************************************************
Log file is located at: /Users/georgep/knime/workspace_mac_new/.metadata/knime/knime.log
WARN     DatabaseDriverLoader    Could not load driver class "sun.jdbc.odbc.JdbcOdbcDriver"
```

Create a new workflow



New KNIME Workflow Wizard

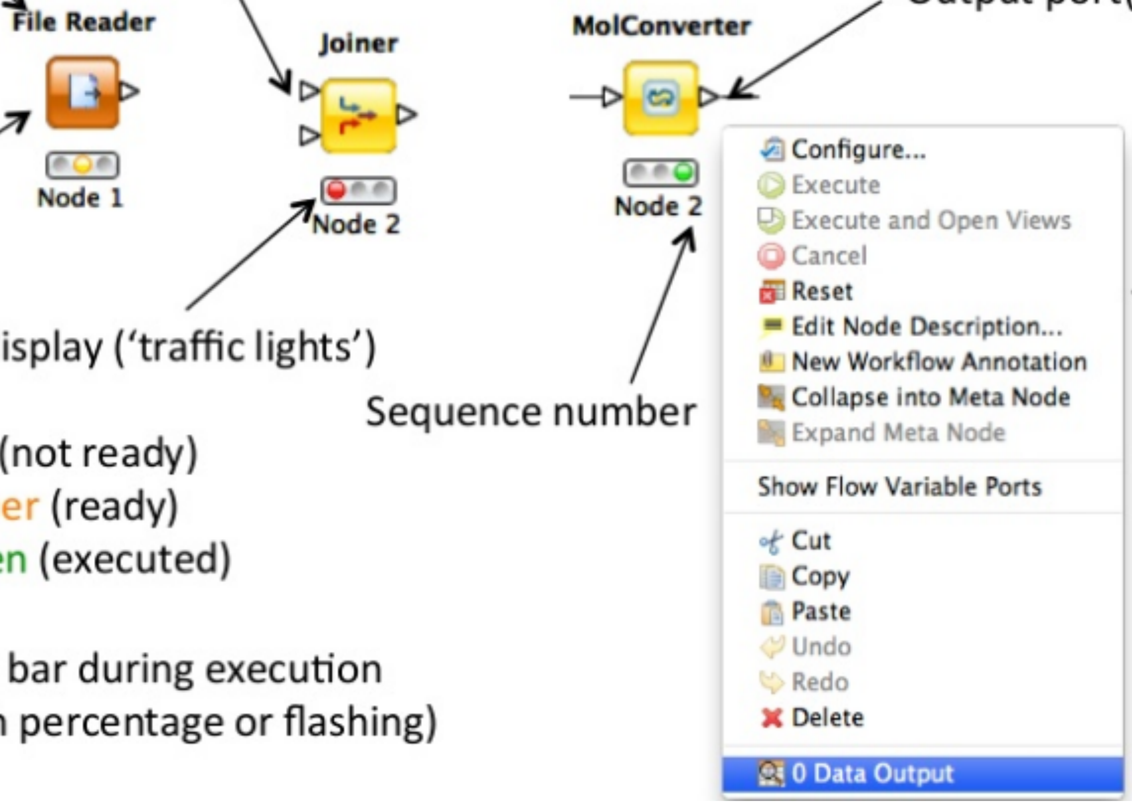
Create a new KNIME workflow.

Name of the workflow to create: my_first_workflow

Workspace destination: /Chapter2

KNIME nodes: Overview

Node = basic processing unit of KNIME workflow which performs a particular task



Title

Input port(s) – on the left of icon

Output port(s) – on the right of icon

Icon

Status display ('traffic lights')

- Red (not ready)
- Amber (ready)
- Green (executed)

- Blue bar during execution (with percentage or flashing)

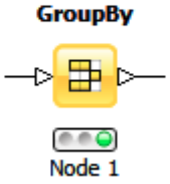
Sequence number

Right-click menu

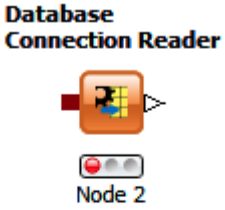
To configure and execute the node, display the output views, edit the node, and display data for the ports

Ports

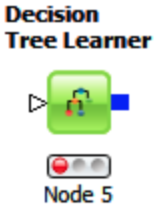
- **Data Port:** a white triangle which transfers flat data tables from node to node



- **Database Port:** Nodes executing commands inside a database are recognized by their database ports (brown square)



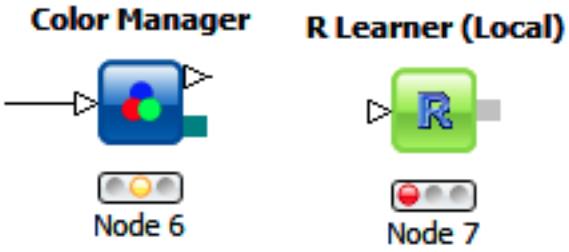
- **PMML Ports:** Data Mining nodes learn a model which is passed to the referring predictor node via a blue squared PMML port



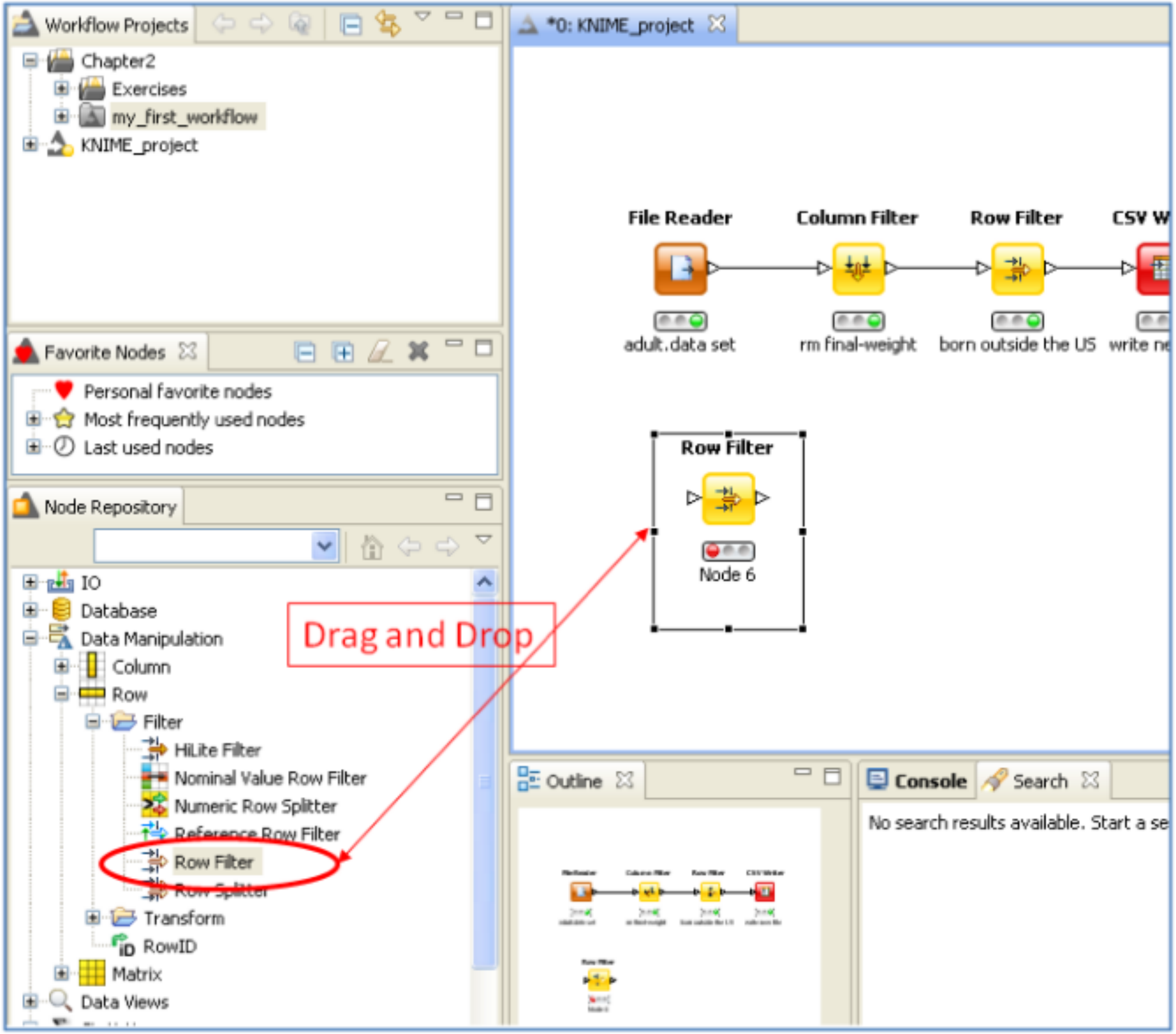
Other Ports

- Whenever a node provides data that does not fit a flat data table structure, **a general purpose port for structured** data is used (dark cyan square).

- All ports not listed above are known as **"unknown" types** (gray square).

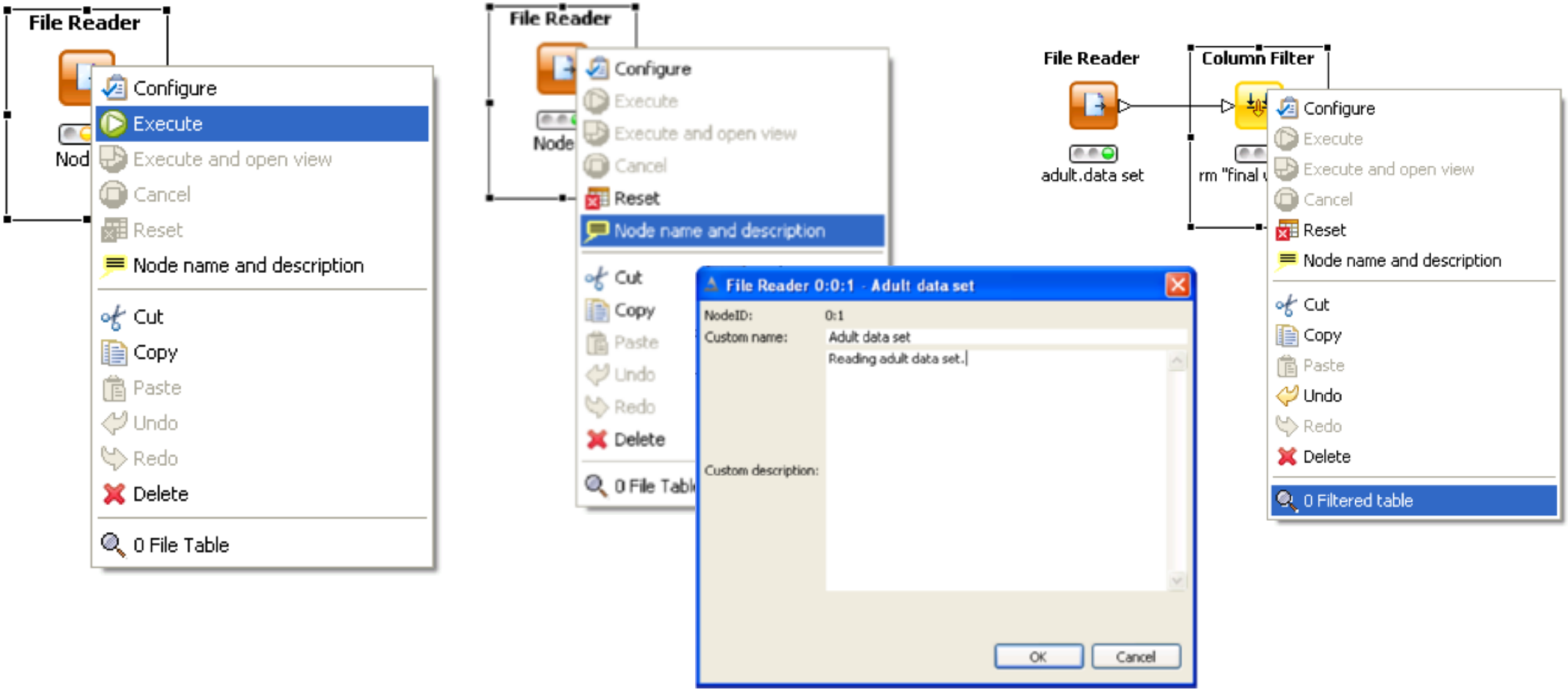


Node Creation



Drag and Drop

Node Operations

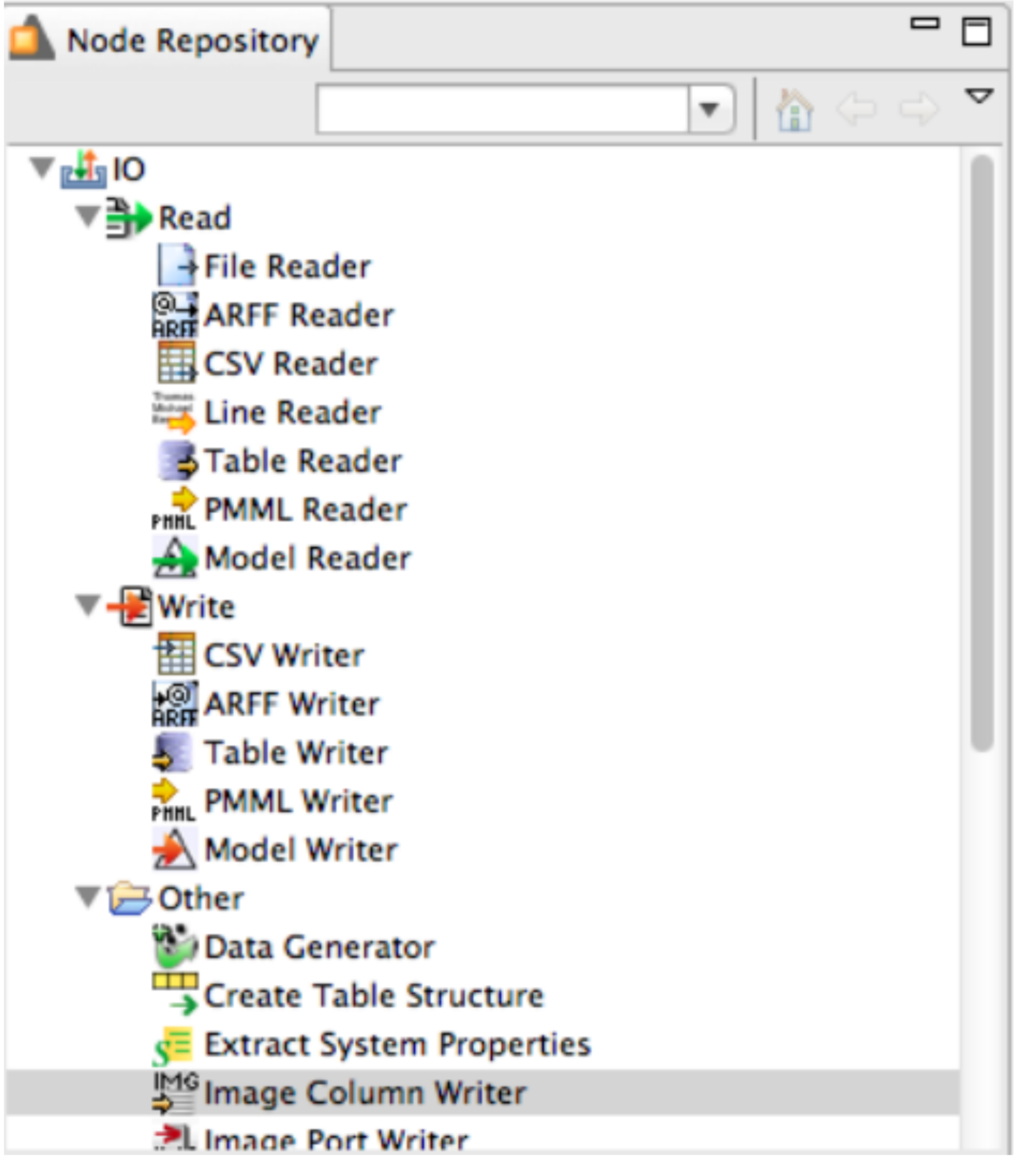


NodeID: 0:1

Custom name: Adult data set

Custom description: Reading adult data set.

I/O Operations



Node Repository

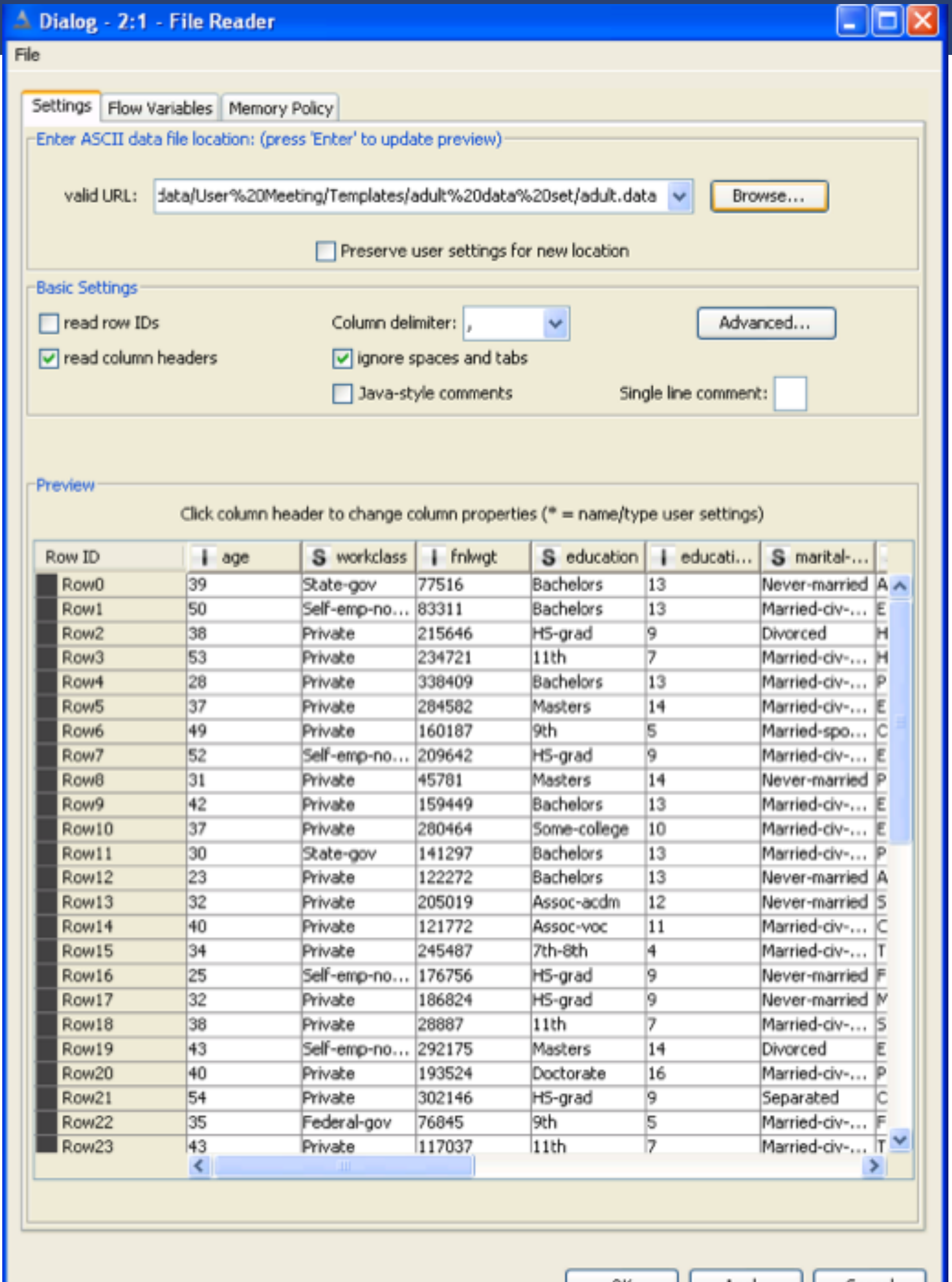
- IO
  - Read
    - File Reader
    - ARFF Reader
    - CSV Reader
    - Line Reader
    - Table Reader
    - PMML Reader
    - Model Reader
  - Write
    - CSV Writer
    - ARFF Writer
    - Table Writer
    - PMML Writer
    - Model Writer
  - Other
    - Data Generator
    - Create Table Structure
    - Extract System Properties
    - Image Column Writer
    - Image Port Writer

ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

CSV (Comma-Separated Values) file stores tabular data (numbers and text) in plain-text form.

Read data from file

File Reader
Node 20



Enter ASCII data file location: (press 'Enter' to update preview)

valid URL: data/User%20Meeting/Templates/adult%20data%20set/adult.data

□ Preserve user settings for new location

Basic Settings

□ read row IDs

☑ read column headers

Column delimiter: ,

☑ ignore spaces and tabs

□ Java-style comments

Single line comment:

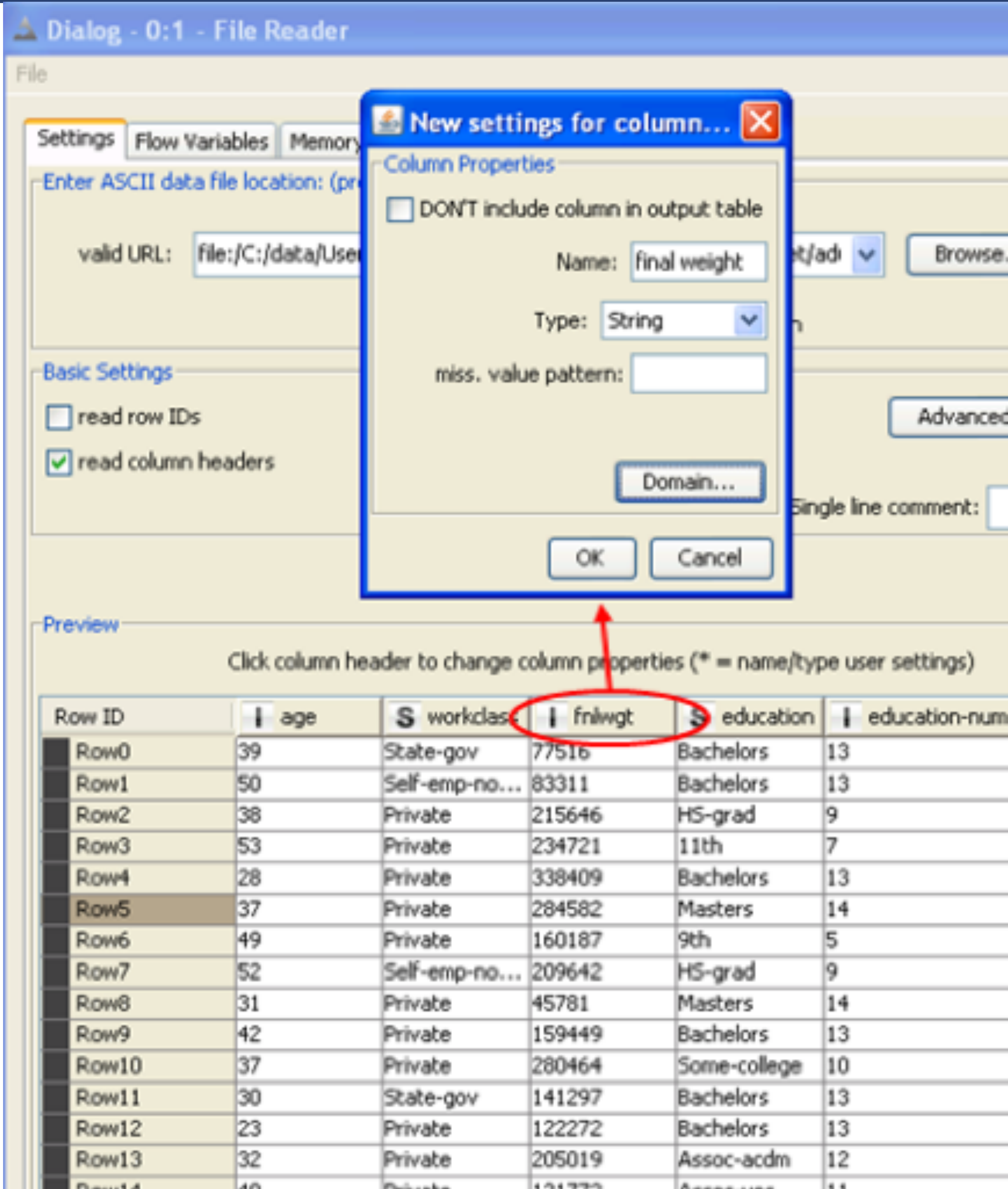
Preview

Click column header to change column properties (* = name/type user settings)

Row ID	I age	S workclass	I fnlwgt	S education	I educati...	S marital-...	
Row0	39	State-gov	77516	Bachelors	13	Never-married	A
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	E
Row2	38	Private	215646	HS-grad	9	Divorced	H
Row3	53	Private	234721	11th	7	Married-civ-...	H
Row4	28	Private	338409	Bachelors	13	Married-civ-...	P
Row5	37	Private	284582	Masters	14	Married-civ-...	E
Row6	49	Private	160187	9th	5	Married-spo...	C
Row7	52	Self-emp-no...	209642	HS-grad	9	Married-civ-...	E
Row8	31	Private	45781	Masters	14	Never-married	P
Row9	42	Private	159449	Bachelors	13	Married-civ-...	E
Row10	37	Private	280464	Some-college	10	Married-civ-...	E
Row11	30	State-gov	141297	Bachelors	13	Married-civ-...	P
Row12	23	Private	122272	Bachelors	13	Never-married	A
Row13	32	Private	205019	Assoc-acdm	12	Never-married	S
Row14	40	Private	121772	Assoc-voc	11	Married-civ-...	C
Row15	34	Private	245487	7th-8th	4	Married-civ-...	T
Row16	25	Self-emp-no...	176756	HS-grad	9	Never-married	F
Row17	32	Private	186824	HS-grad	9	Never-married	M
Row18	38	Private	28887	11th	7	Married-civ-...	S
Row19	43	Self-emp-no...	292175	Masters	14	Divorced	E
Row20	40	Private	193524	Doctorate	16	Married-civ-...	P
Row21	54	Private	302146	HS-grad	9	Separated	C
Row22	35	Federal-gov	76845	9th	5	Married-civ-...	F
Row23	43	Private	117037	11th	7	Married-civ-...	T

Read data from file

- Click in the column name
  - Change column name
  - Change type



Enter ASCII data file location: (pr

valid URL: file:/C:/data/Use …t/adi

Basic Settings

□ read row IDs

☑ read column headers

Single line comment:

New settings for column...

Column Properties

□ DON'T include column in output table

Name: final weight

Type: String

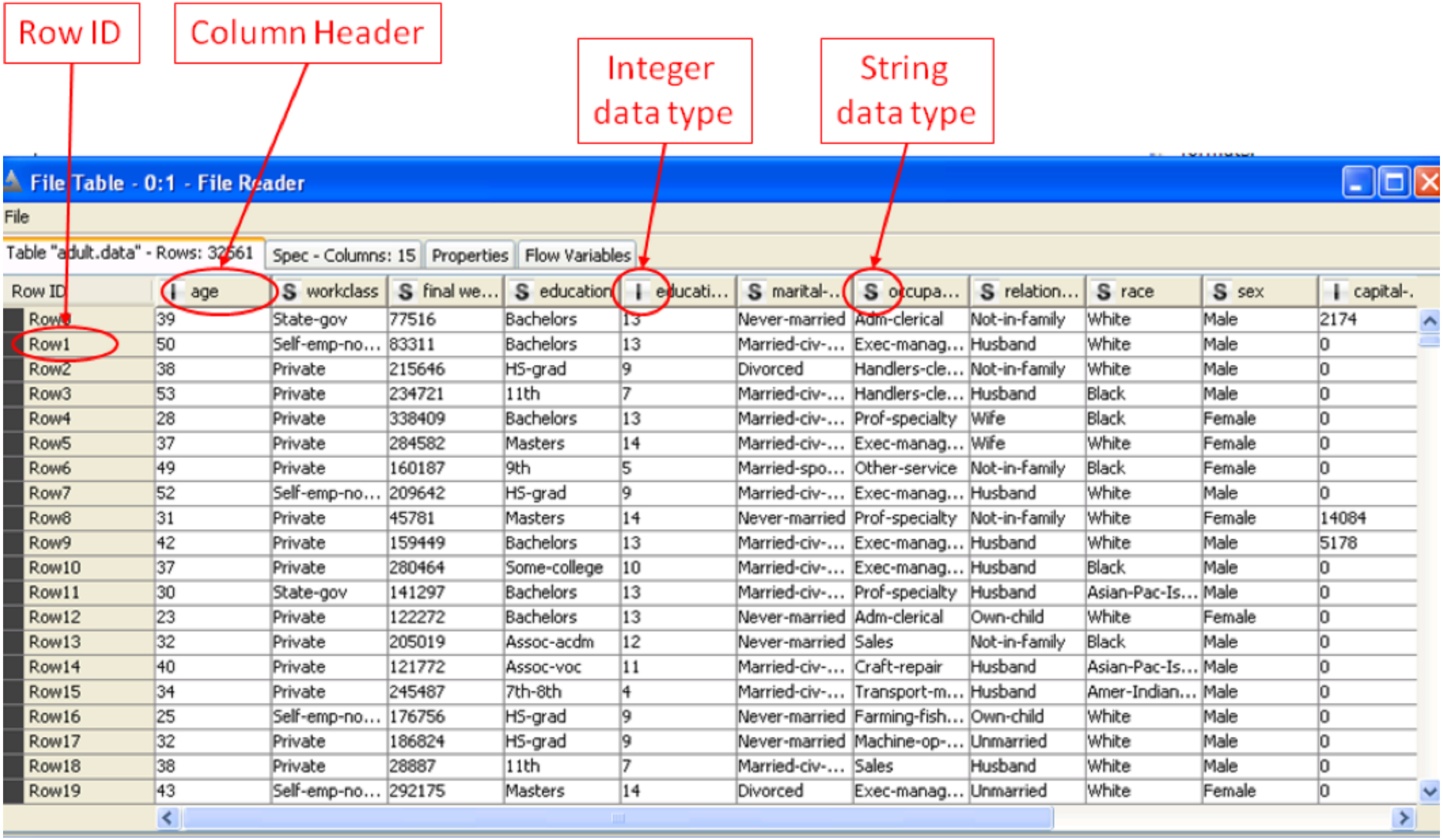
miss. value pattern:

Preview

Click column header to change column properties (* = name/type user settings)

Row ID	I age	S workclass	I fnlwgt	S education	I education-num
Row0	39	State-gov	77516	Bachelors	13
Row1	50	Self-emp-no...	83311	Bachelors	13
Row2	38	Private	215646	HS-grad	9
Row3	53	Private	234721	11th	7
Row4	28	Private	338409	Bachelors	13
Row5	37	Private	284582	Masters	14
Row6	49	Private	160187	9th	5
Row7	52	Self-emp-no...	209642	HS-grad	9
Row8	31	Private	45781	Masters	14
Row9	42	Private	159449	Bachelors	13
Row10	37	Private	280464	Some-college	10
Row11	30	State-gov	141297	Bachelors	13
Row12	23	Private	122272	Bachelors	13
Row13	32	Private	205019	Assoc-acdm	12

Table Data



Row ID

Column Header

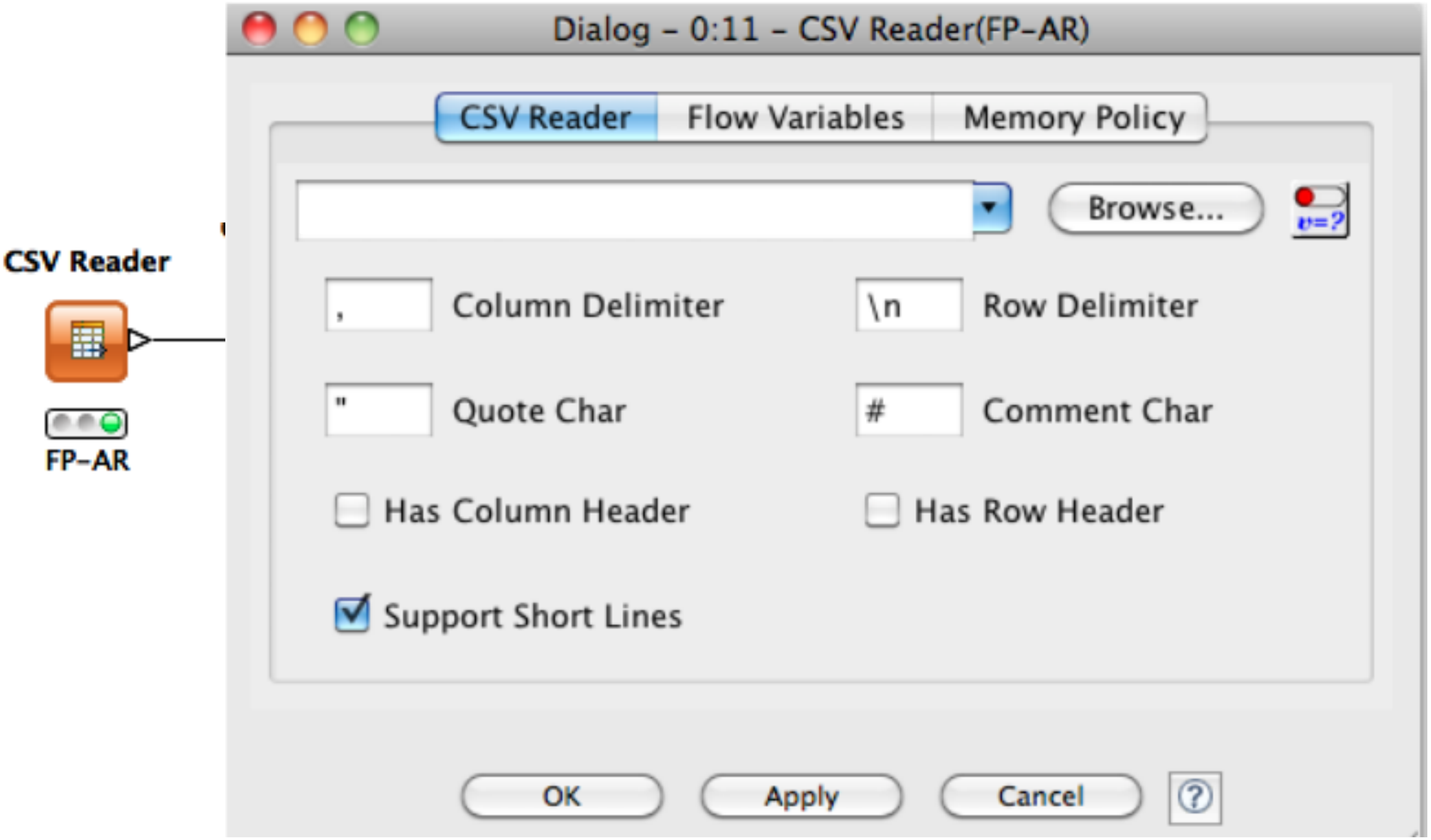
Integer data type

String data type

File Table - 0:1 - File Reader
File
Table "adult.data" - Rows: 32561
Spec - Columns: 15
Properties
Flow Variables

Row ID	I age	S workclass	S final we...	S education	I educati...	S marital-...	S occupa...	S relation...	S race	S sex	I capital-.
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	Exec-manag...	Husband	White	Male	0
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cle...	Not-in-family	White	Male	0
Row3	53	Private	234721	11th	7	Married-civ-...	Handlers-cle...	Husband	Black	Male	0
Row4	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specialty	Wife	Black	Female	0
Row5	37	Private	284582	Masters	14	Married-civ-...	Exec-manag...	Wife	White	Female	0
Row6	49	Private	160187	9th	5	Married-spo...	Other-service	Not-in-family	Black	Female	0
Row7	52	Self-emp-no...	209642	HS-grad	9	Married-civ-...	Exec-manag...	Husband	White	Male	0
Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084
Row9	42	Private	159449	Bachelors	13	Married-civ-...	Exec-manag...	Husband	White	Male	5178
Row10	37	Private	280464	Some-college	10	Married-civ-...	Exec-manag...	Husband	Black	Male	0
Row11	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specialty	Husband	Asian-Pac-Is...	Male	0
Row12	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0
Row13	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0
Row14	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband	Asian-Pac-Is...	Male	0
Row15	34	Private	245487	7th-8th	4	Married-civ-...	Transport-m...	Husband	Amer-Indian...	Male	0
Row16	25	Self-emp-no...	176756	HS-grad	9	Never-married	Farming-fish...	Own-child	White	Male	0
Row17	32	Private	186824	HS-grad	9	Never-married	Machine-op-...	Unmarried	White	Male	0
Row18	38	Private	28887	11th	7	Married-civ-...	Sales	Husband	White	Male	0
Row19	43	Self-emp-no...	292175	Masters	14	Divorced	Exec-manag...	Unmarried	White	Female	0

Other input nodes: CSV Reader



, Column Delimiter

\n Row Delimiter

" Quote Char

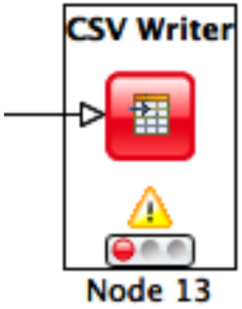
Comment Char

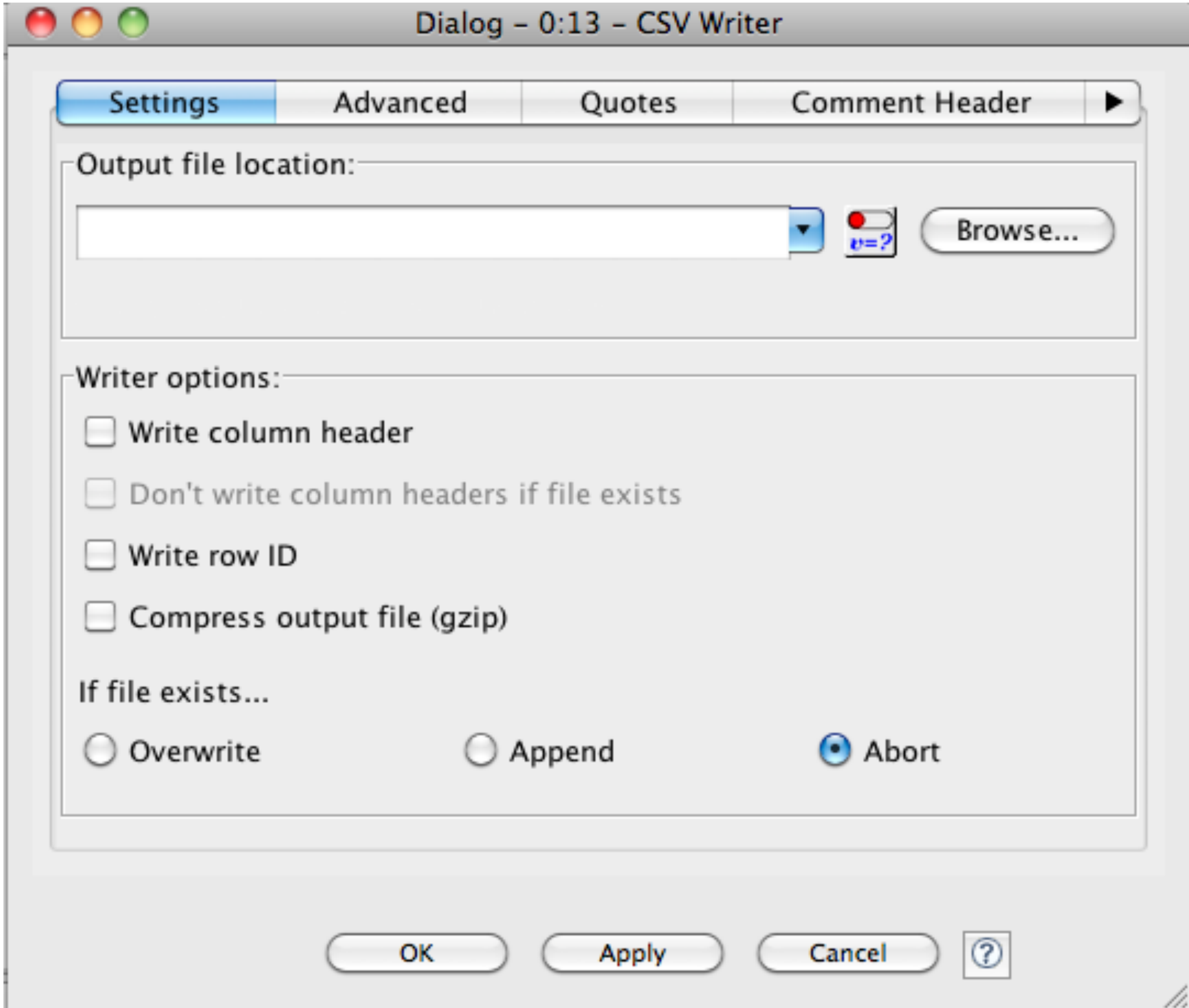
□ Has Column Header

□ Has Row Header

☑ Support Short Lines

CSV Writer





Output file location:

Writer options:

□ Write column header

□ Don't write column headers if file exists

□ Write row ID

□ Compress output file (gzip)

If file exists...

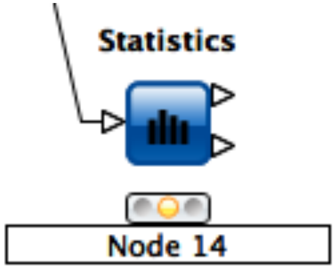
○ Overwrite ○ Append ● Abort

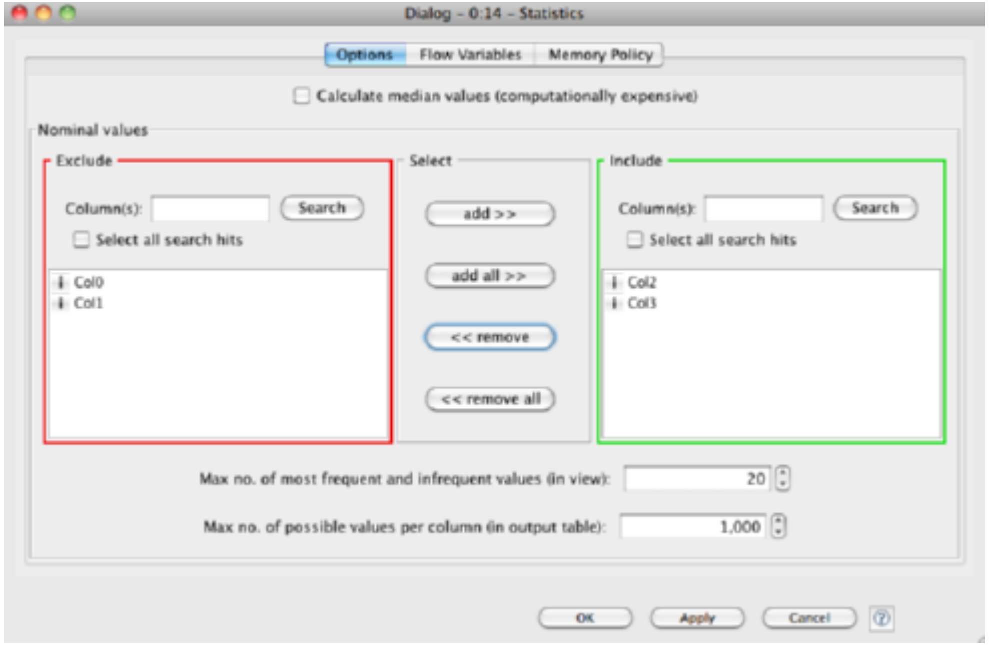
Data Manipulation

- Three main sections
  - **Columns**: binning, replace, filters, normalizer, missing values, …
  - **Rows**: filtering, sampling, partitioning, …
  - **Matrix**: Transpose

Statistics node

- For all numeric columns computes statistics such as
- **minimum**, **maximum**, **mean**, **standard deviation**, **variance**, **median**, **overall sum**, **number of missing values** and **row counts**
- For all nominal values counts them together with their occurrences.





□ Calculate median values (computationally expensive)

Nominal values

Exclude

Column(s): Search

□ Select all search hits

Col0
Col1

Select

Include

Column(s): Search

□ Select all search hits

Col2
Col3

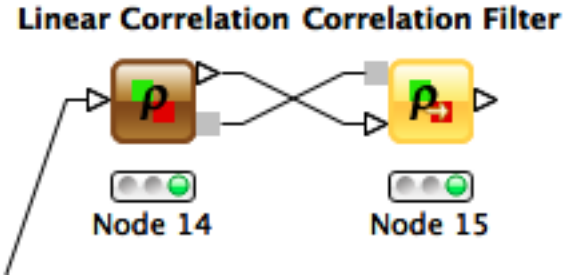
Max no. of most frequent and infrequent values (in view): 20

Max no. of possible values per column (in output table): 1,000

Correlation Analysis

- **Linear Correlation node** computes for each pair of selected columns a correlation coefficient, i.e. a measure of the correlation of the two variables
  - Pearson Correlation Coefficient

- **Correlation Filtering node** uses the model as generated by a Correlation node to determine which columns are redundant (i.e. correlated) and filters them out.
  - **The output table will contain the reduced set of columns.**



Data Views

- Box Plots
- Histograms, Pie Charts, Scatter plots, …
- Scatter Matrix

Mining Algorithms

- Clustering
  - Hierarchical
  - K-means
  - Fuzzy –c-Means
- Decision Tree
- Item sets / Association Rules
  - Borgelt's Algorithms (Extension)
- Weka (Extension)

Data Manipulation

- See Workflow on the course website