KNIME TUTORIAL

Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

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

Forum & Documentation

KNIME tech

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:

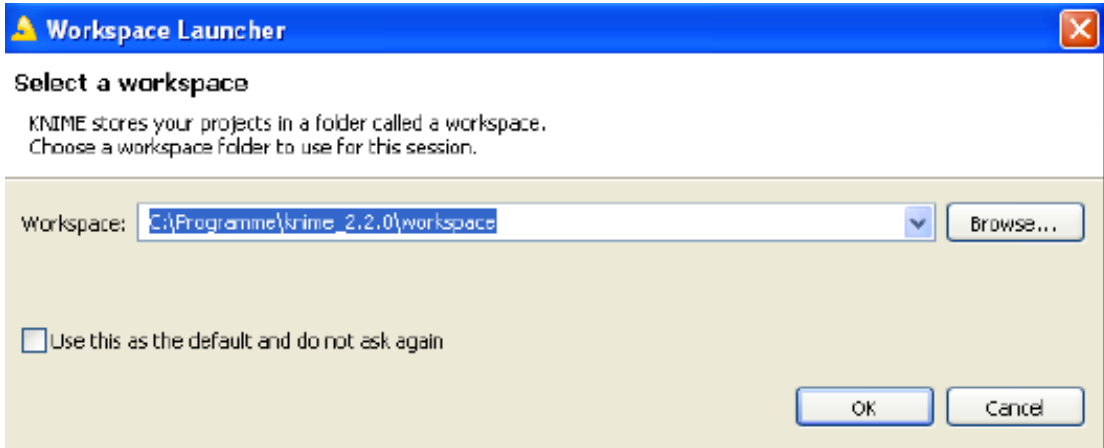
- KNIME for Linux 32bit
- KNIME for Linux 64bit

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:

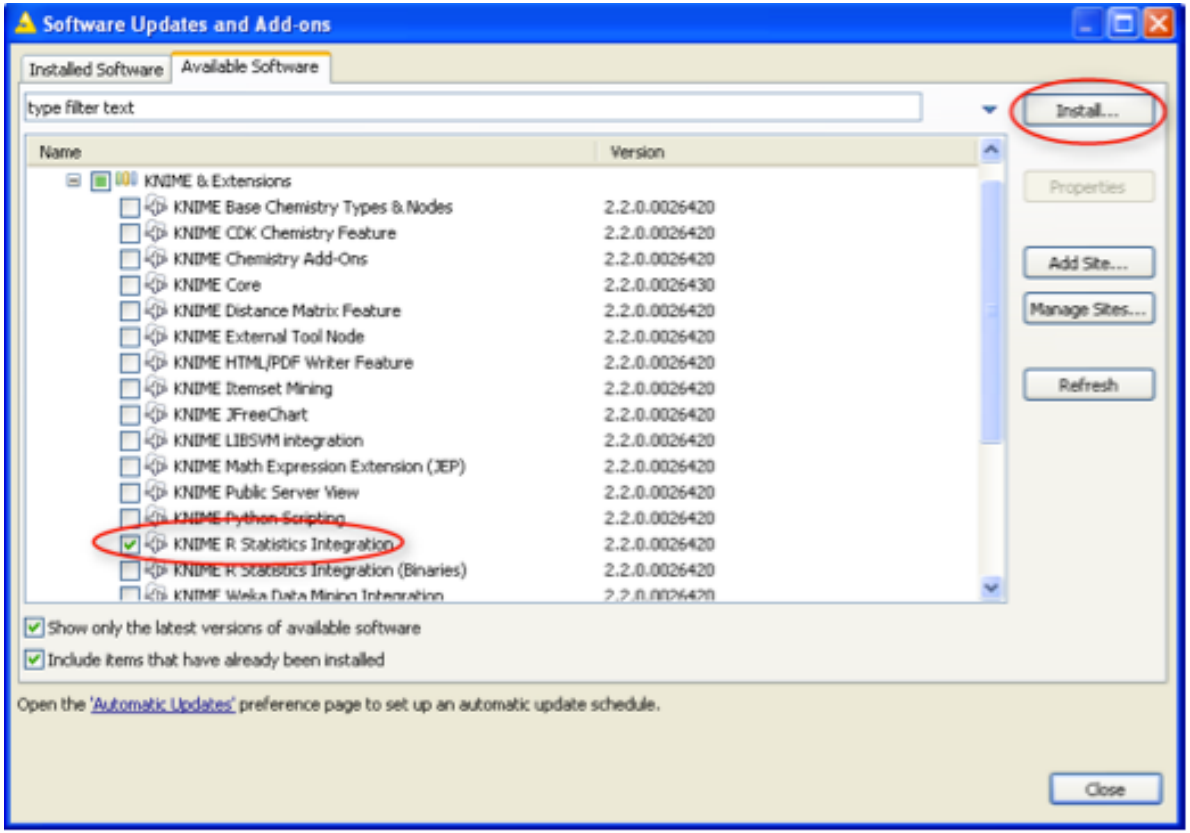
Workspace

- 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.



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

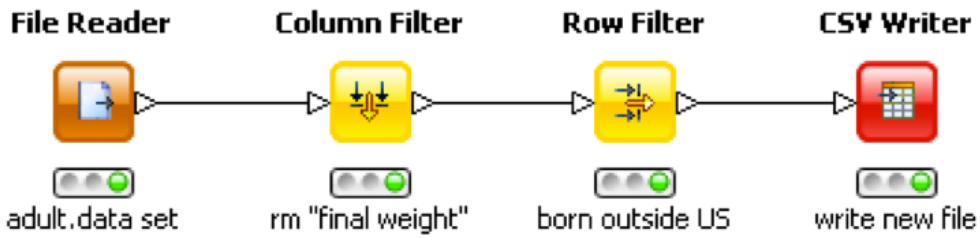


What can you do with KNIME?

- **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

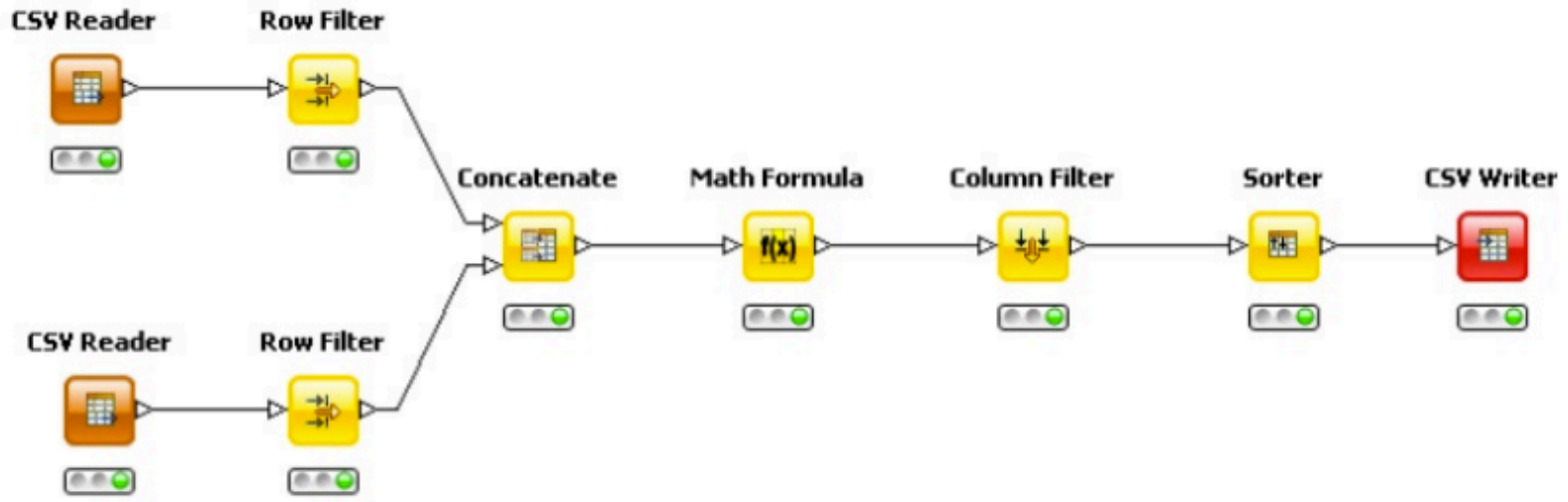
- KNIME does not work with scripts, **it works with workflows.**
- A workflow is an analysis flow, which is the sequence of the analysis steps necessary to reach a given result:
  1. Read data
  2. Clean data
  3. Filter data
  4. Train a model



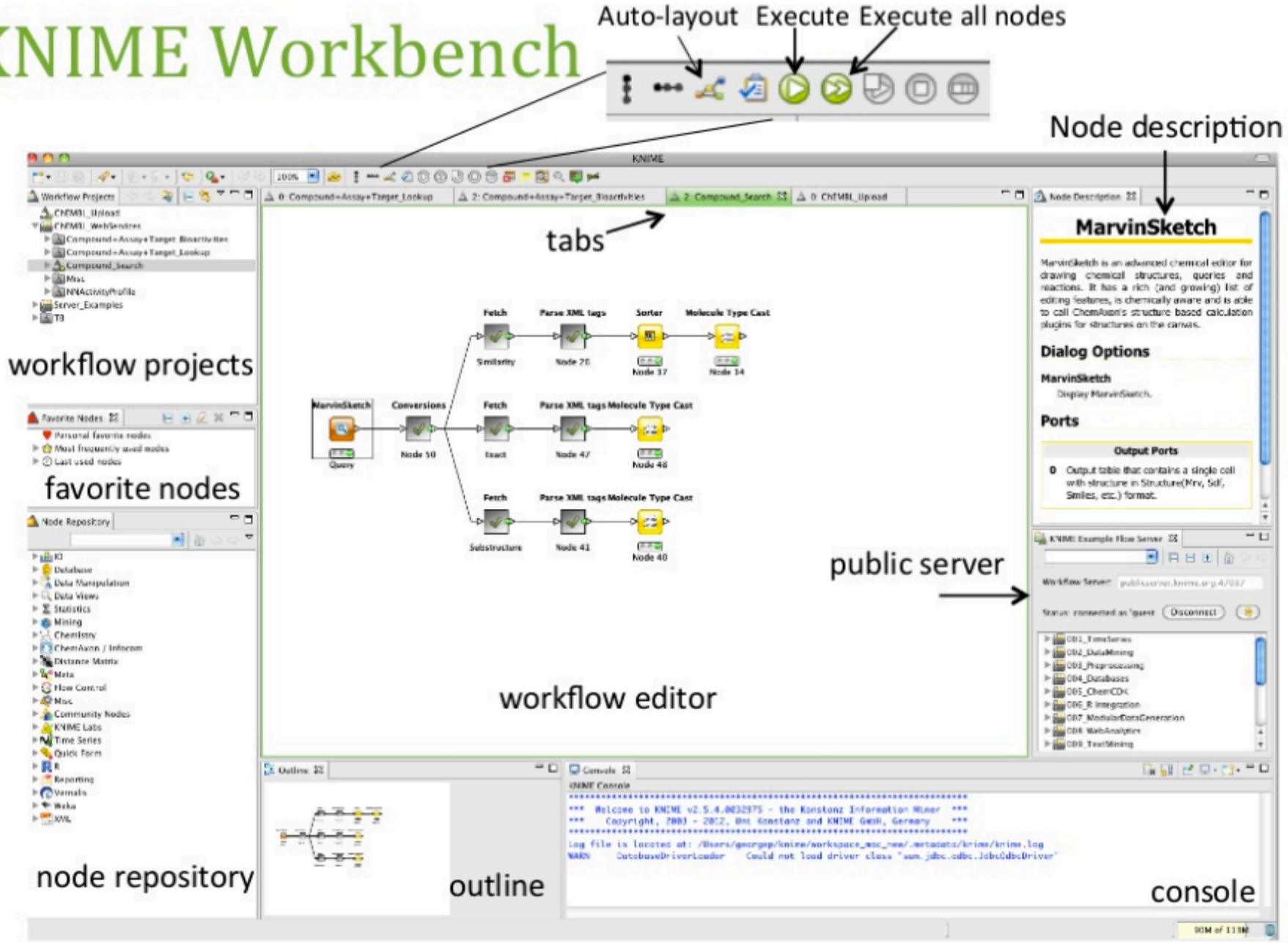
- KNIME implements its workflows **graphically**.
- Each step of the data analysis is executed by a little box, called a **node**.
- **A sequence of nodes makes a workflow.**

Import/export of workflow

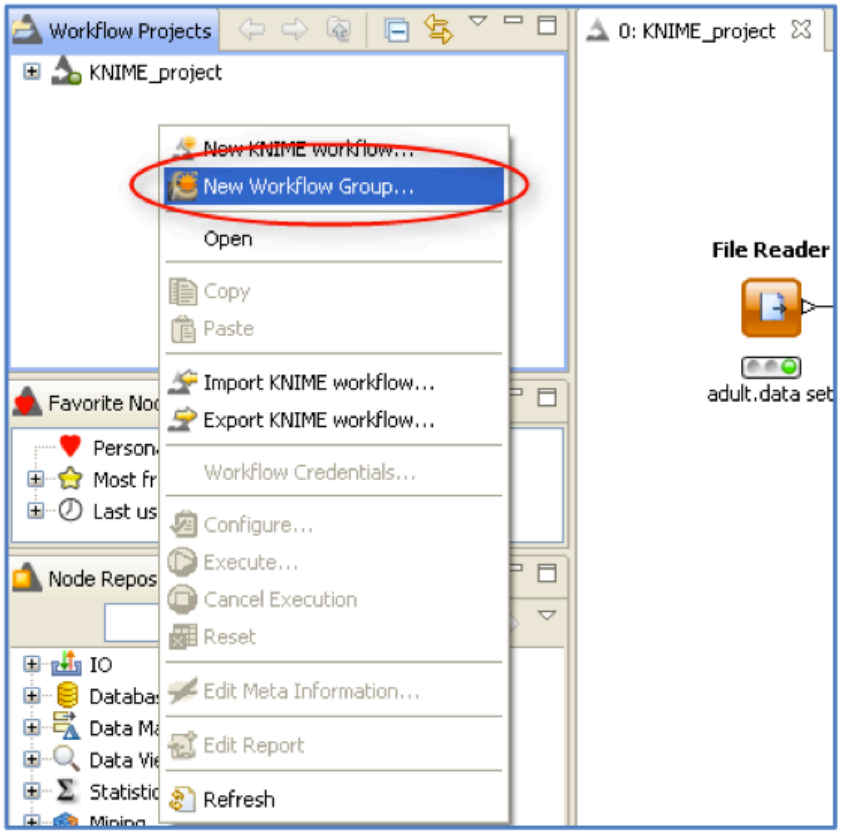
- Workflows can be imported and exported as .zip files
  - With or without the underlying data
  - File → Import KNIME workflow…
  - File → Export KNIME workflow…

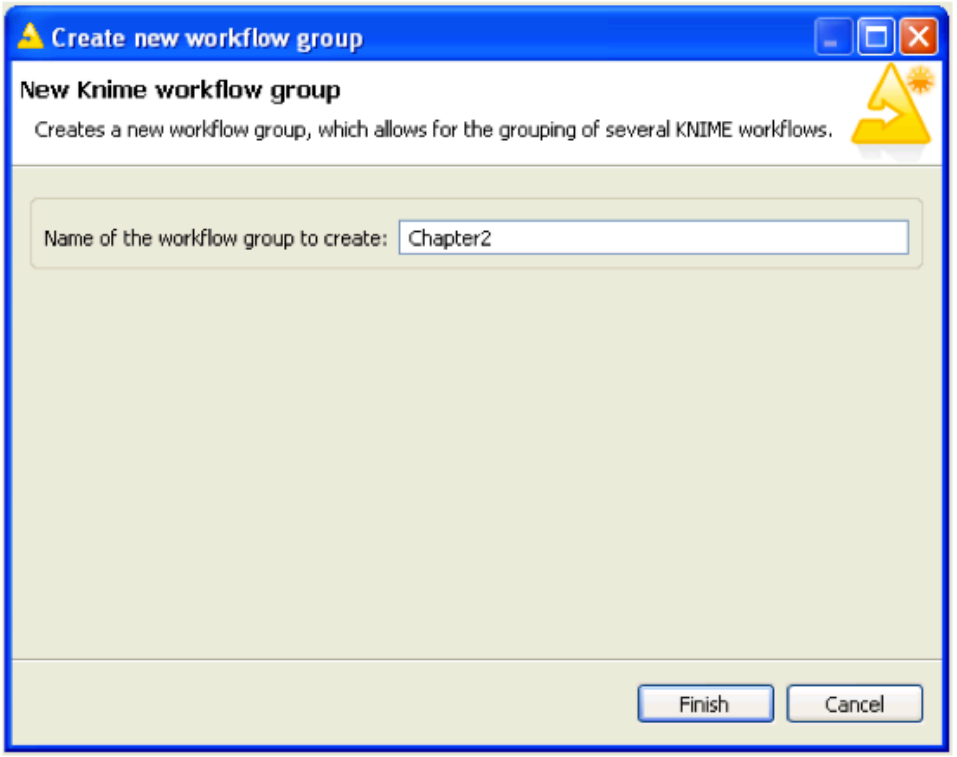


KNIME Workbench

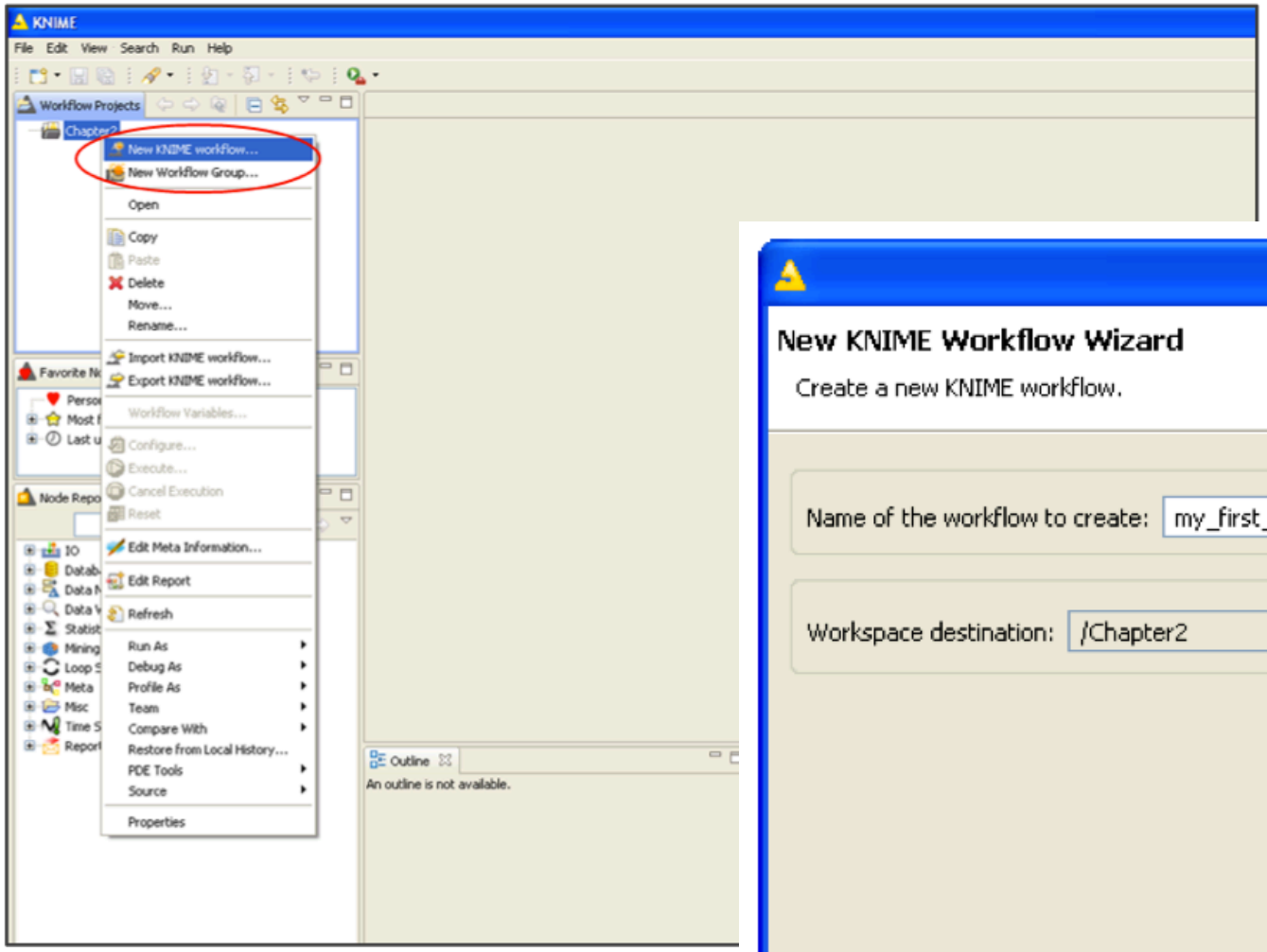


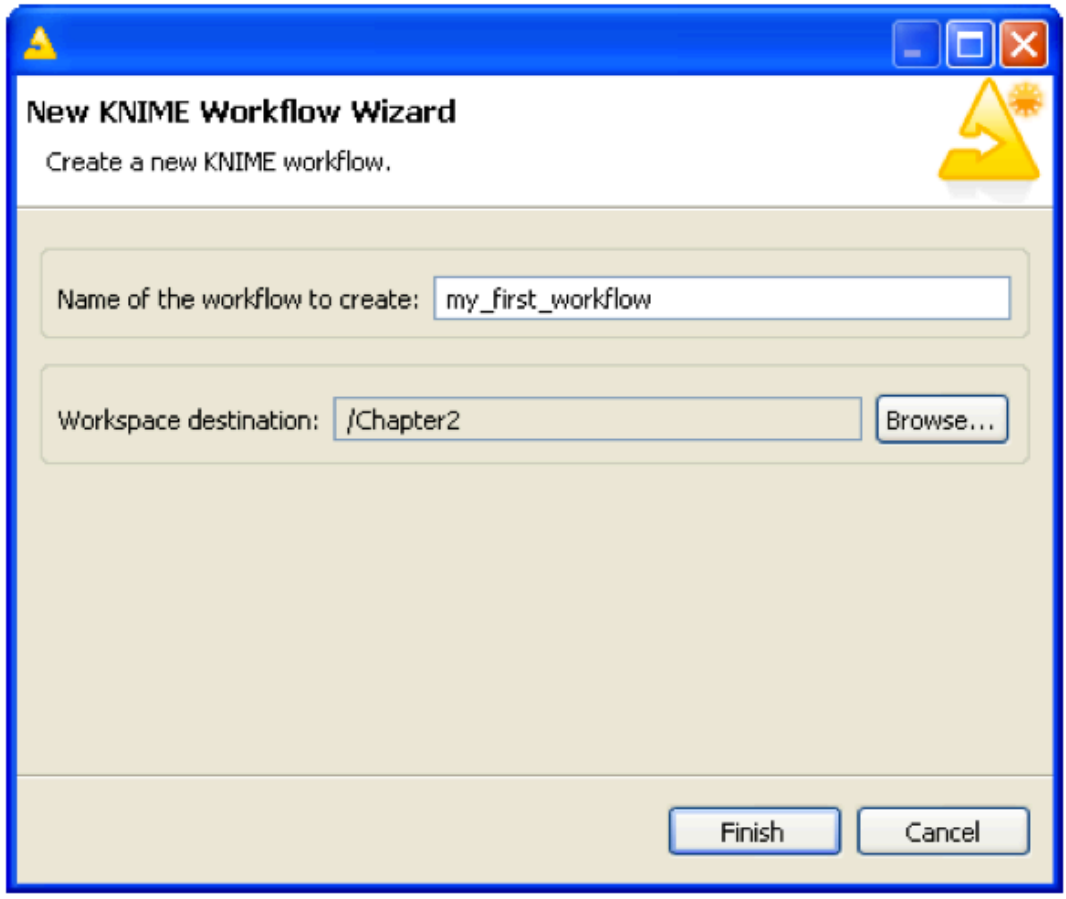
Create a new workflow group





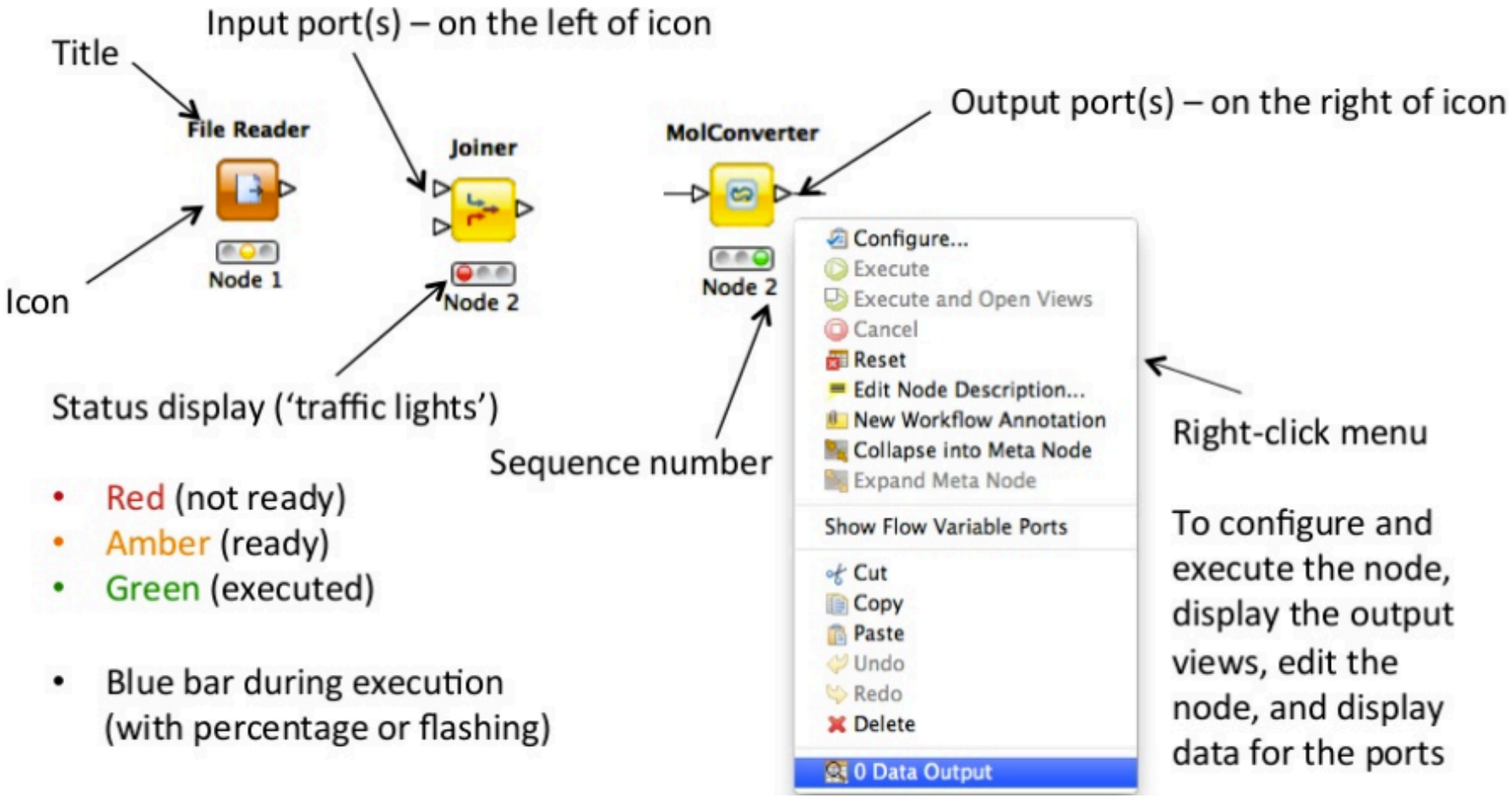
Create a new workflow





KNIME nodes: Overview

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



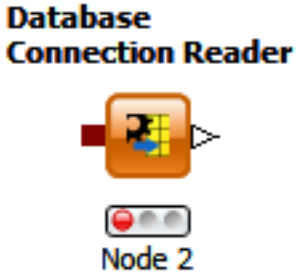
- Red (not ready)
- Amber (ready)
- Green (executed)

- Blue bar during execution (with percentage or flashing)

Ports

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

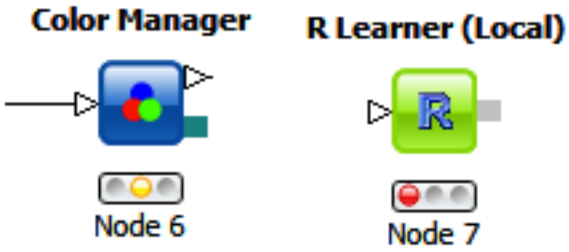
- **Database Port:** Nodes executing commands inside a database are recognized by their database ports (brown square)



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

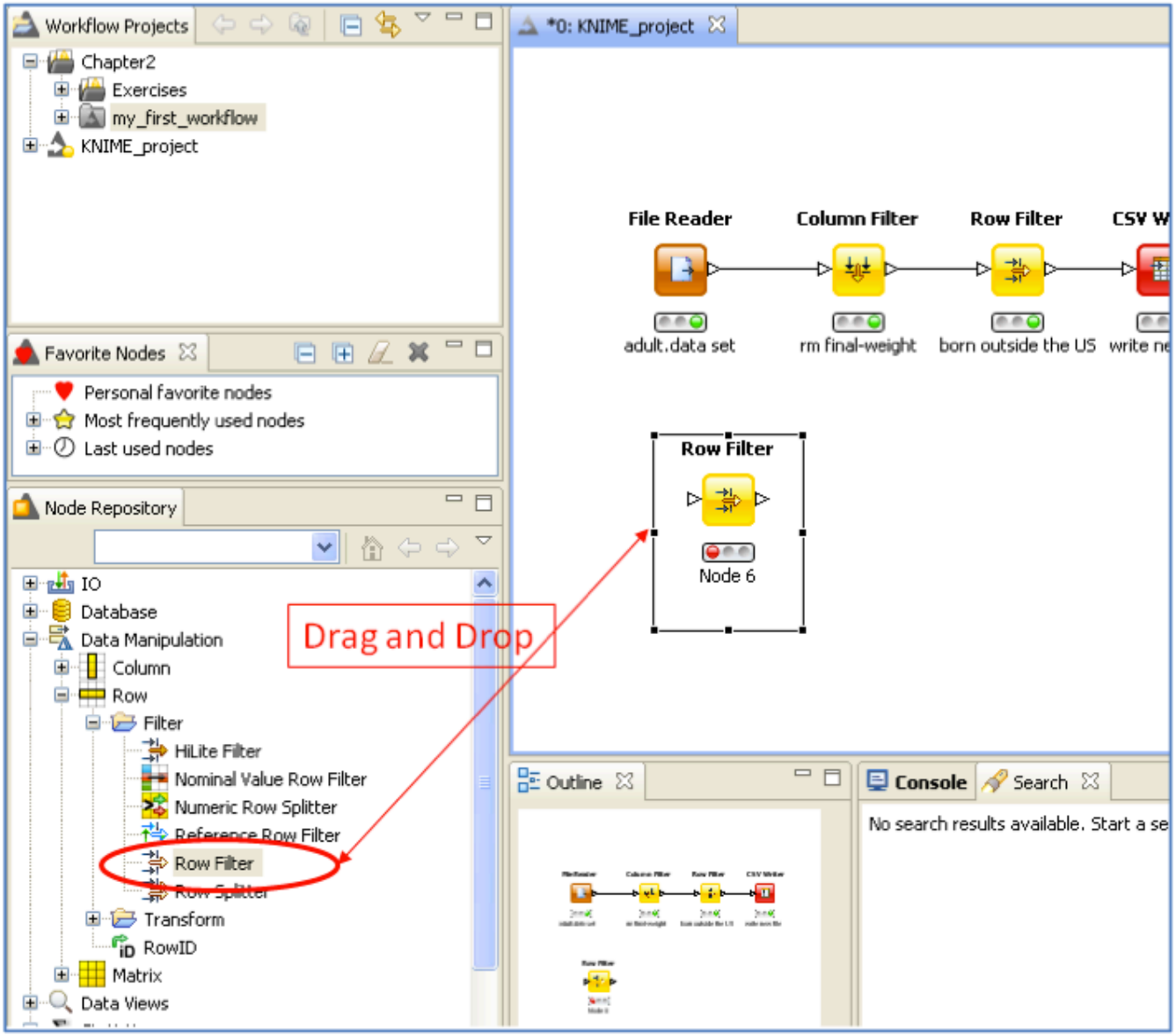
Other Ports

- 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).

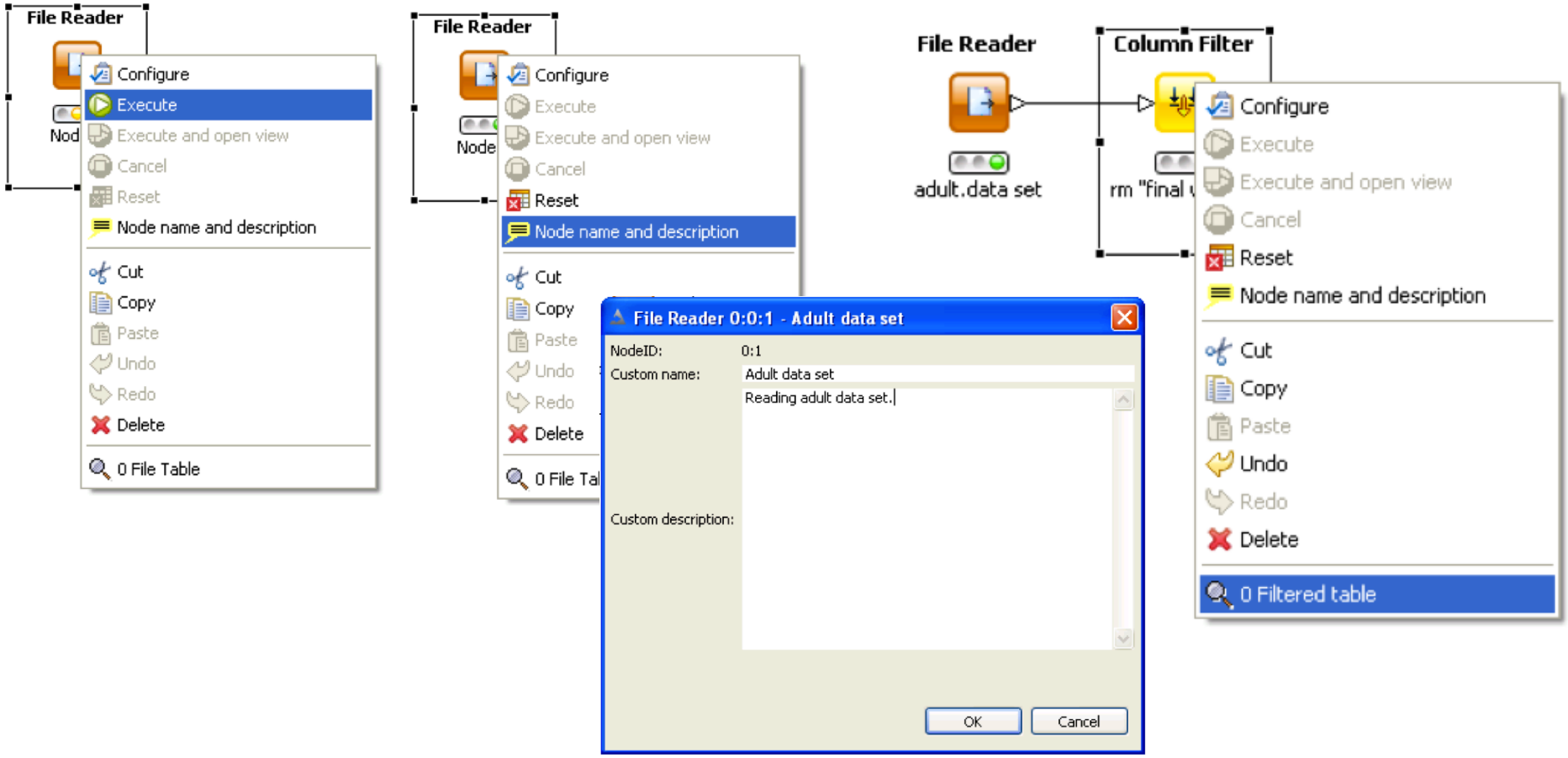


- All ports not listed above are known as **"unknown" types** (gray square).

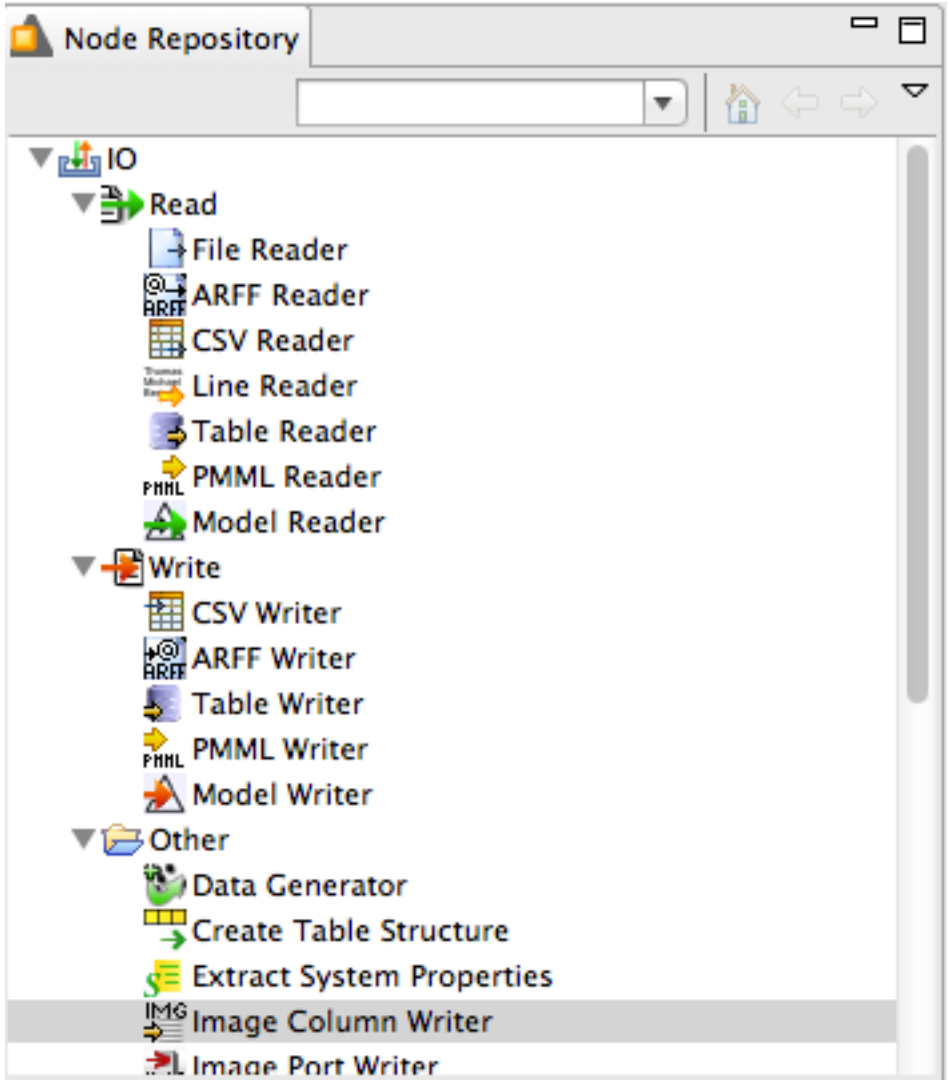
Node Creation



Node Operations



I/O Operations



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

Dialog - 2:1 - File Reader

File

Settings | Flow Variables | Memory Policy

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

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

Preserve user settings for new location

Basic Settings

read row IDs — Column delimiter: , — Advanced...

read column headers — ignore spaces and tabs

Java-style comments — Single line comment:

Preview

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

Row ID	age	workclass	fnlwgt	education	educati...	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

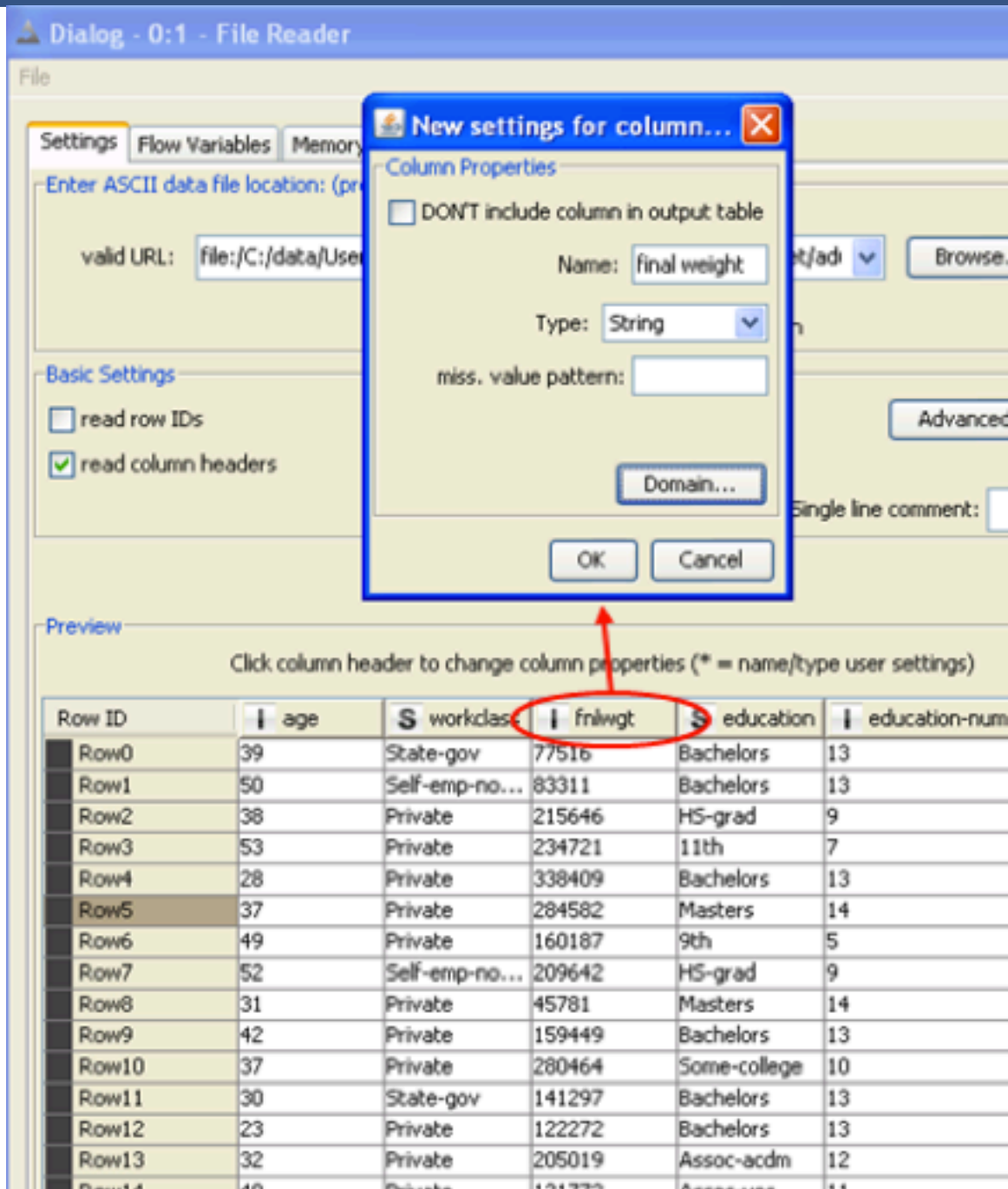
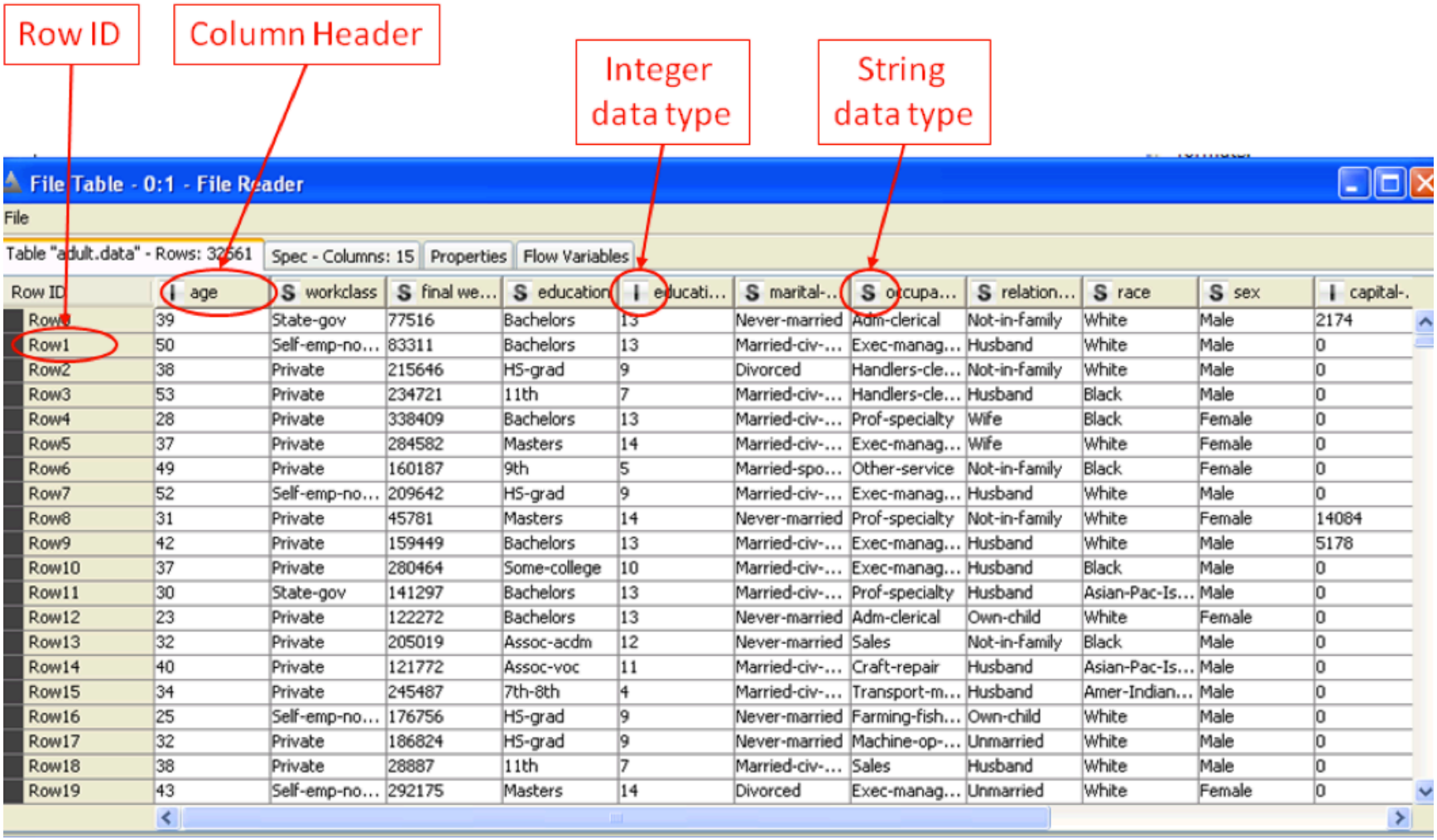
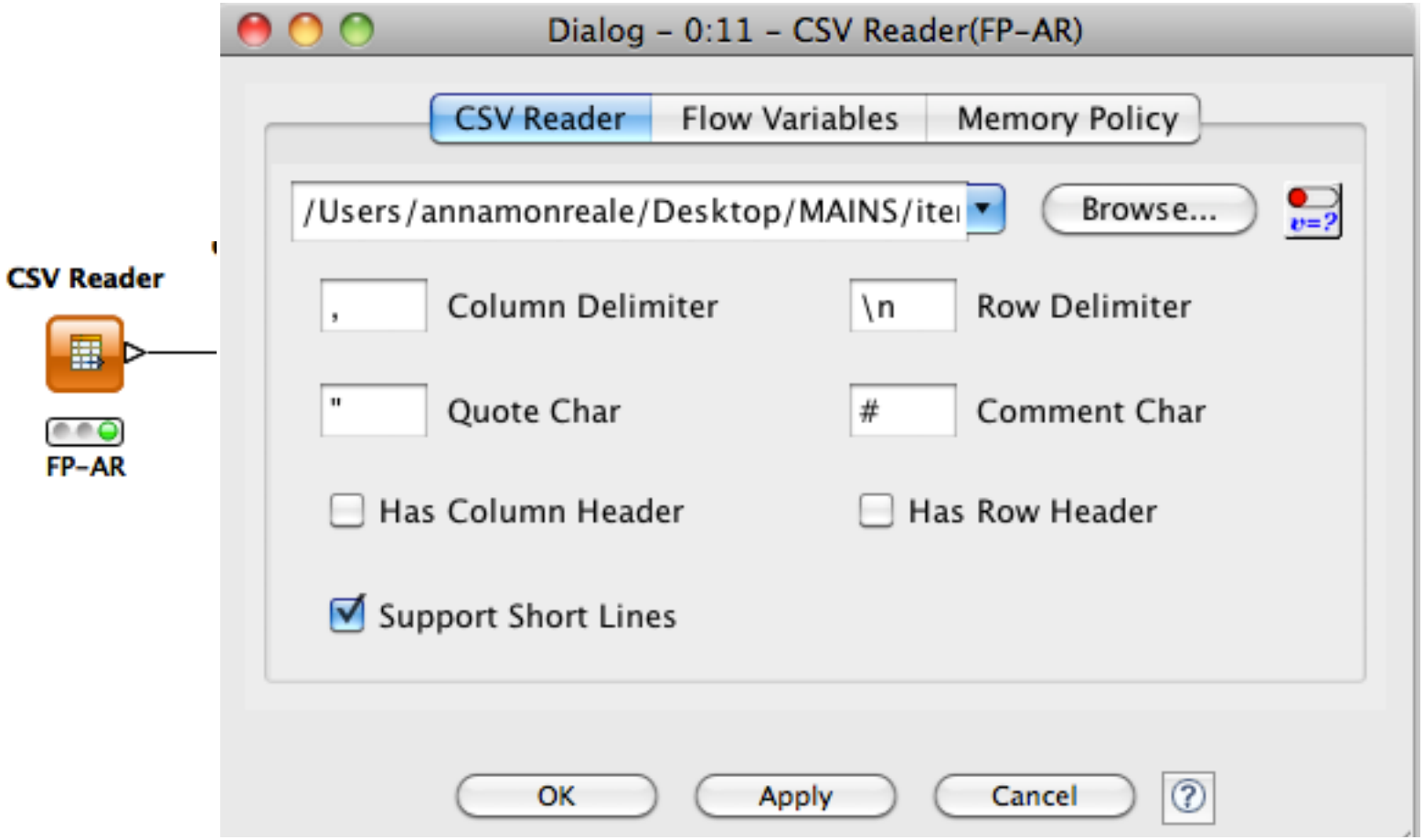


Table Data

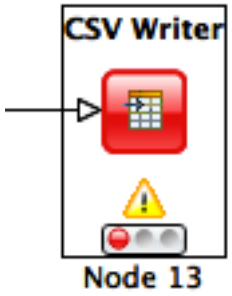


Row ID	age	workclass	final we...	education	educati...	marital-...	occupa...	relation...	race	sex	capital-.
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	Exec-manag...	Husband	White	Male	0
Row2	38	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



CSV Writer



Dialog – 0:13 – CSV Writer

Settings | Advanced | Quotes | Comment Header

Output file location:

Browse...

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

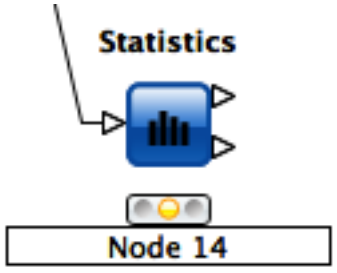
OK Apply Cancel

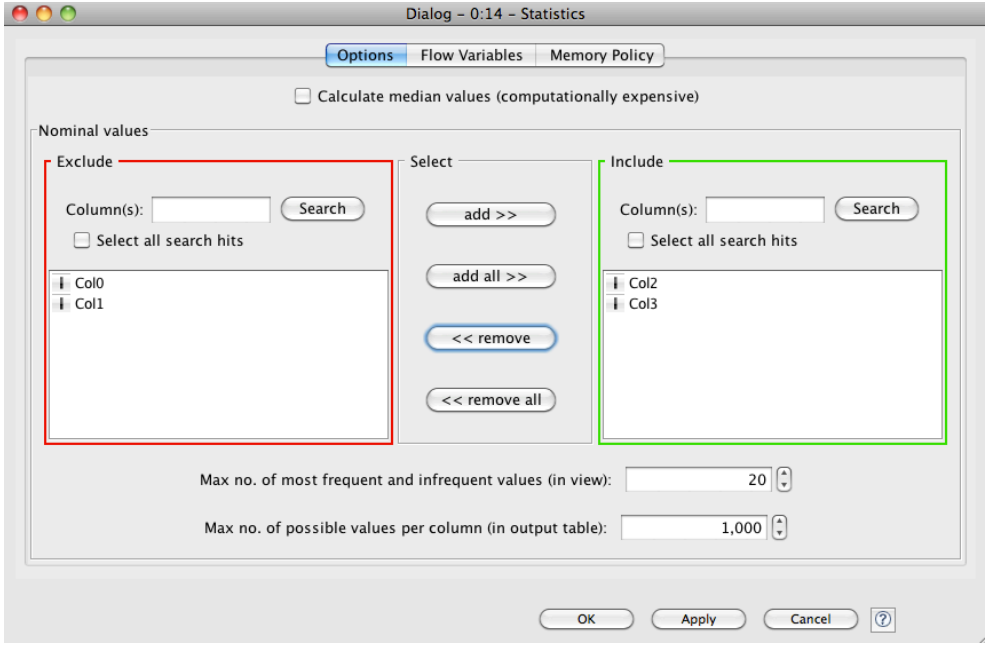
Data Manipulation

- 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.

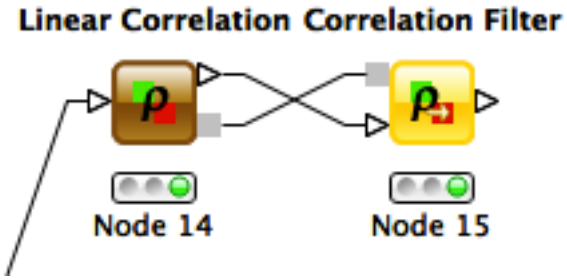




Correlation Analysis

- **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)

EXERCISES

Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

DATA MANIPULATION

Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

Data Manipulation

- See Workflow on the course website

http://didawiki.cli.di.unipi.it/doku.php/dm/mains.santanna.dm4crm.2012

MARKET BASKET ANALYSIS

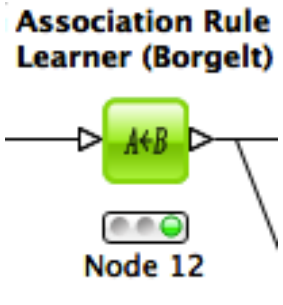
Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

Market Basket Analysis

- **Problem**: given a database of transactions of customers of a supermarket, find **the set of frequent items co-purchased** and analyze the **association rules** that is possible to derive from the frequent patterns.

Frequent Patterns and AR in KNIME

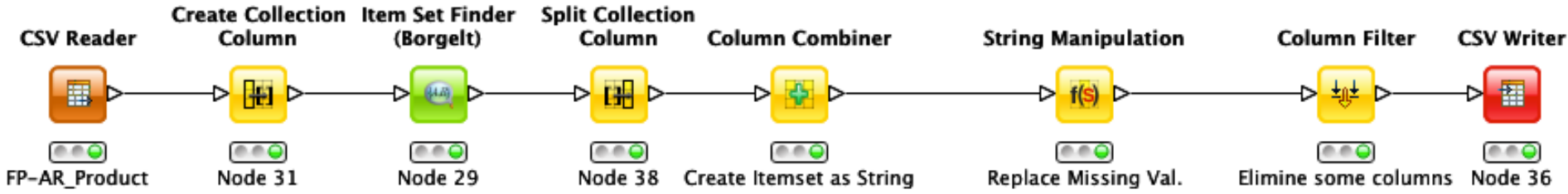
- Use the nodes implementing the Borgelt's Algorithms:



- **Item Set Finder node** provides different algorithms:
  - Apriori (Agrawal et al. 1993)
  - FPgrowth (frequent pattern growth, Han et al 2000)
  - RElim (recursive elimination)
  - SaM (Split and Merge)
  - JIM (Jaccard Item Set Mining)
- **AR Learner uses Apriori Algorithm**

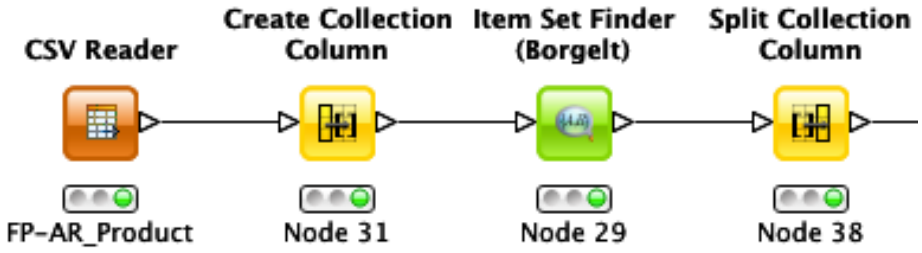
Write the itemsets in a file

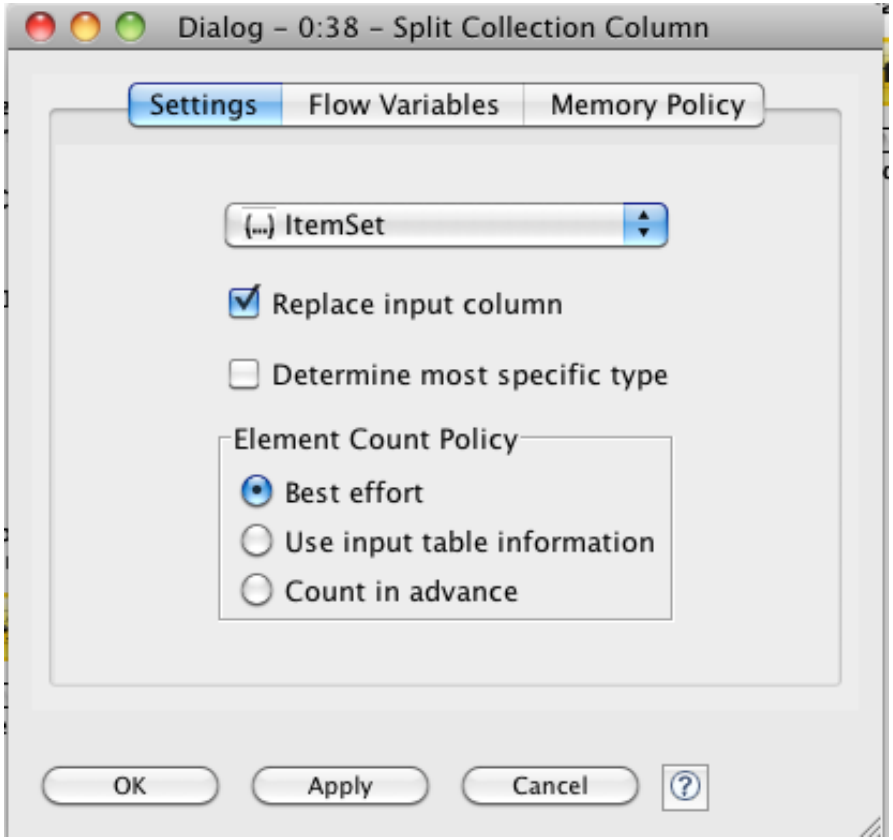
- Given the output of the Item set Finder node sometimes you cannot see all the components of the itemset
  - we need to transform it in a string and
  - then, we can also write the result in a file



Write the itemsets in a file

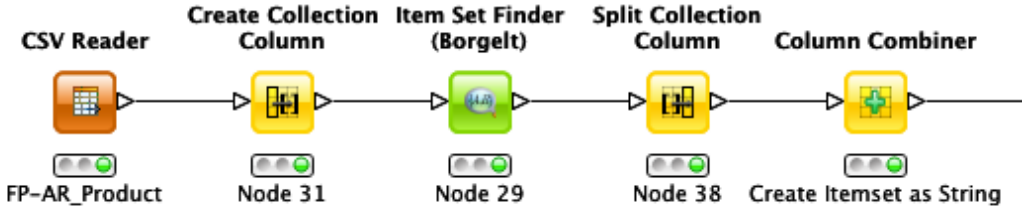
- First we need to split the collection

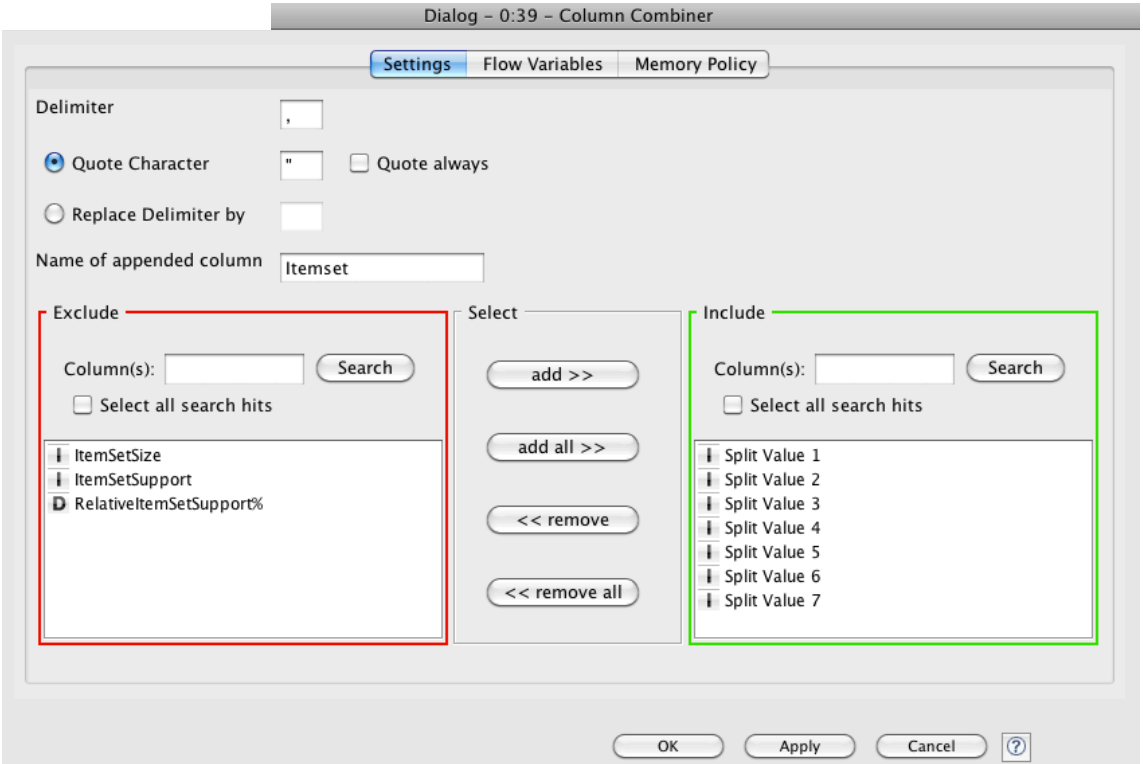




Write the itemsets in a file

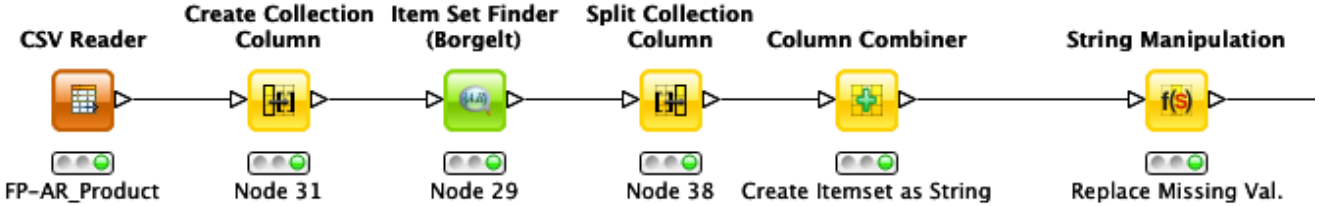
- Second we combine the columns that have to compose the itemset (string)



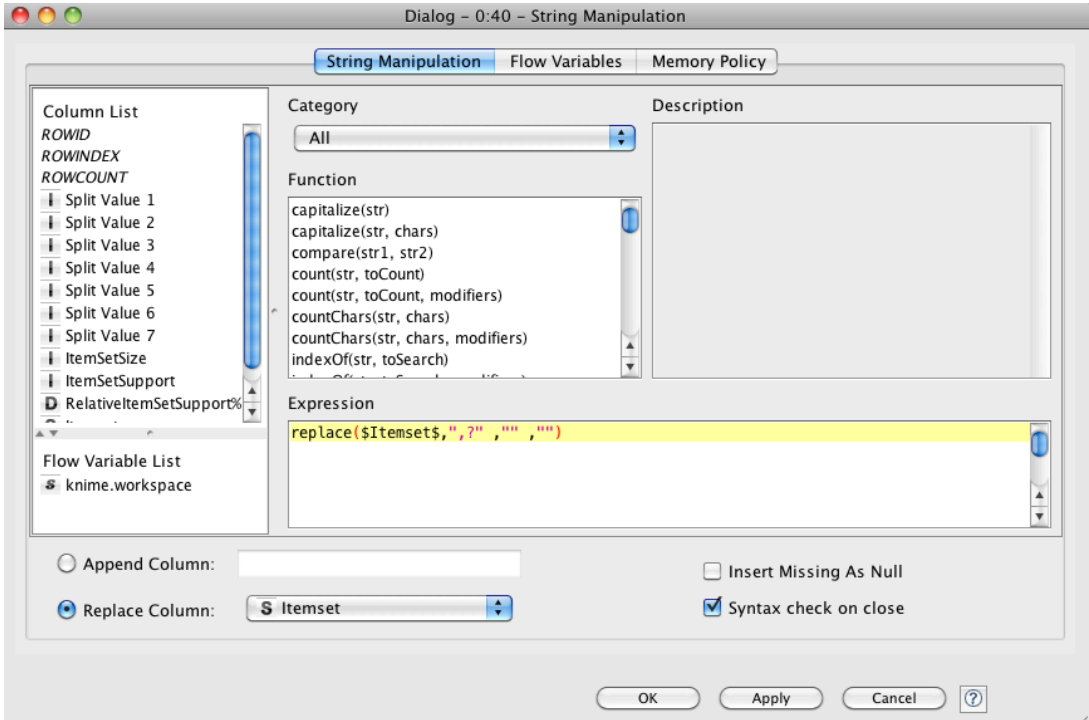


Write the itemsets in a file

- The combiner does not eliminate the missing values “?”
- The combined itemsets contain a lot of “?”

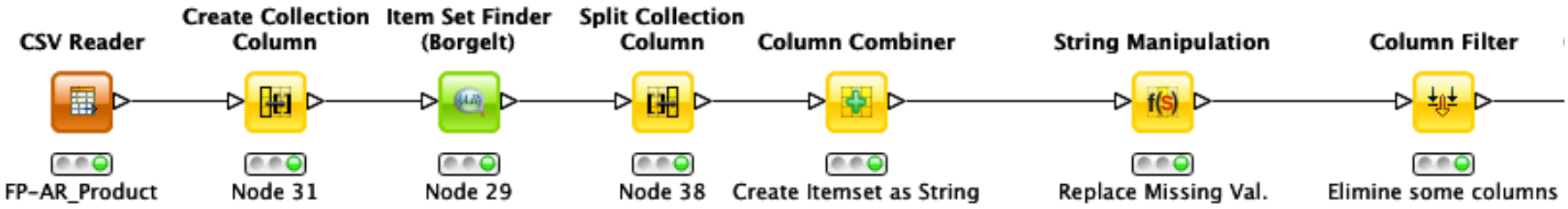


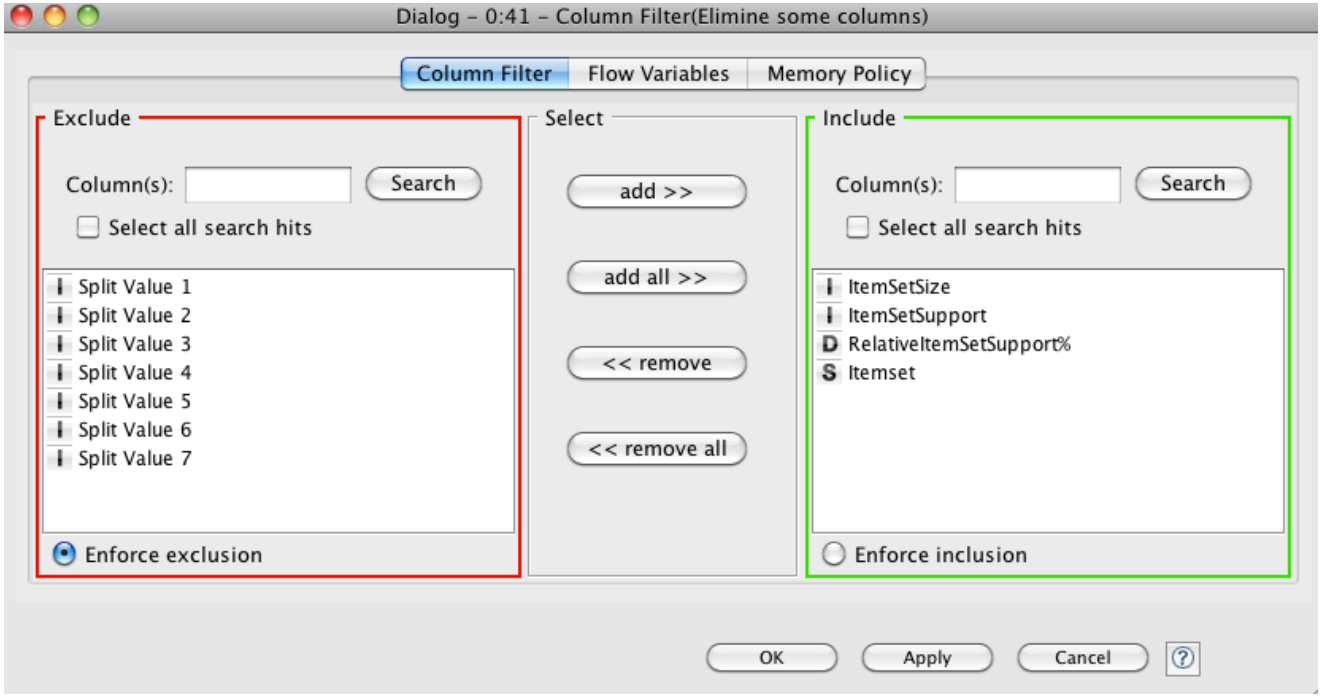
- We use the **replace** operation to eliminate them



Write the itemsets in a file

- Before writing in a file eliminate the split columns





….. The output table

Filtered table – 0:41 – Column Filter(Elimine some columns)

File

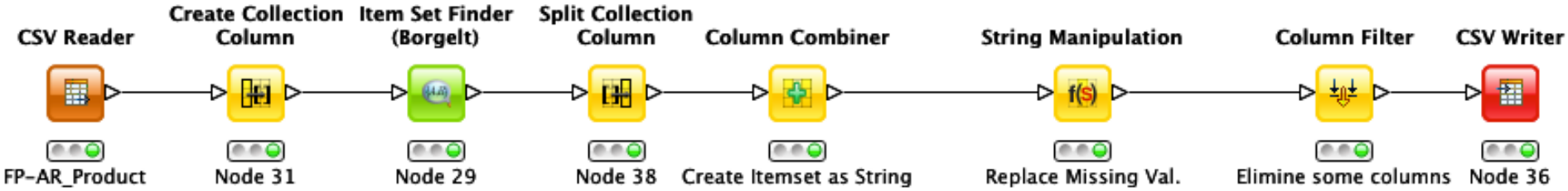
Table "default" – Rows: 139122 | Spec – Columns: 4 | Properties | Flow Variables

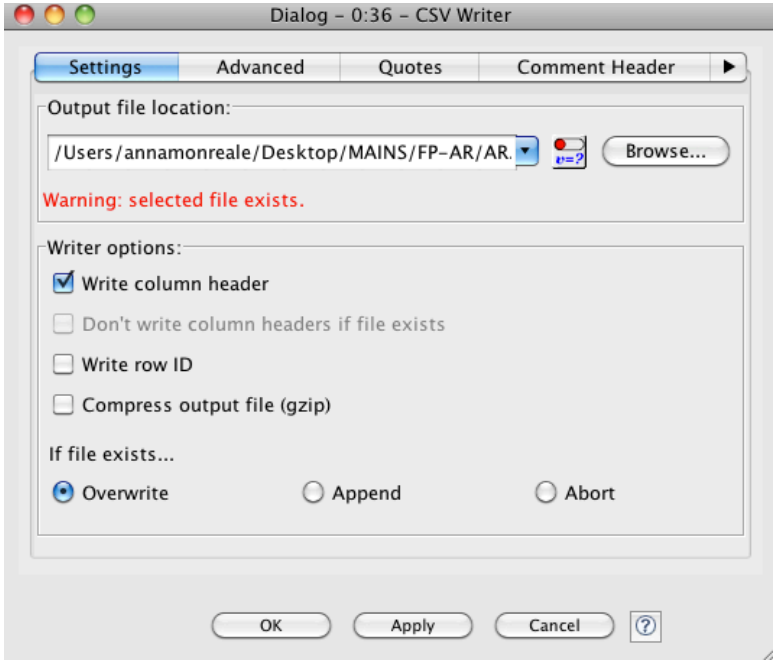
Row ID	ItemSetSize	ItemSetSupport	RelativeItemSetSup…	Itemset
Row94237	7	3	0.03	16864,30459,233740,15786,265109,311540,85800
Row102226	7	3	0.03	253300,7697,45168,15506,36369,72989,85800
Row35465	6	3	0.03	39071,68523,14635,31560,75153,85800
Row63365	6	3	0.03	228263,38950,37860,76174,65616,224434
Row63811	6	3	0.03	2334354,76174,265109,31560,75153,85800
Row65867	6	3	0.03	52006,265111,221614,265109,75153,85800
Row68210	6	3	0.03	31555,14845,45168,31560,85800,75153
Row72720	6	3	0.03	287124,236490,243821,75153,31560,85800
Row78817	6	3	0.03	30958,7697,257536,25227,228164,56674
Row81349	6	3	0.03	27008,30459,65125,16722,48067,265109
Row84546	6	3	0.03	269468,30459,233740,52769,265109,311540
Row84610	6	3	0.03	269468,233740,16281,48067,265109,85800
Row86734	6	3	0.03	28467,16281,72989,221614,31560,75153
Row89111	6	3	0.03	26308,15506,243821,31560,75153,85800
Row89246	6	3	0.03	76288,40287,56674,48067,75153,265109
Row90026	6	3	0.03	2335012,67463,68523,221614,265109,85800
Row94238	6	3	0.03	16864,30459,233740,15786,265109,311540
Row94239	6	3	0.03	16864,30459,233740,15786,265109,85800
Row94241	6	3	0.03	16864,30459,233740,15786,311540,85800
Row94245	6	3	0.03	16864,30459,233740,311540,265109,85800
Row94253	6	3	0.03	16864,30459,15786,265109,311540,85800
Row94342	6	3	0.03	16864,233740,15786,48067,265109,311540

- **Now you can see all the items in a set!!!**

Write the itemsets in a file

- Now we can complete the workflow with the **CSV Writer**





CUSTOMER SEGMENTATION

Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

Customer Segmentation

- **Problem:** given the dataset of RFM (Recency, Frequency and Monetary value) measurements of a set of customers of a supermarket, find a high-quality clustering using K-means and discuss the profile of each found cluster (in terms of the purchasing behavior of the customers of each cluster).

- Applying also the Hierarchical clustering and compare the results

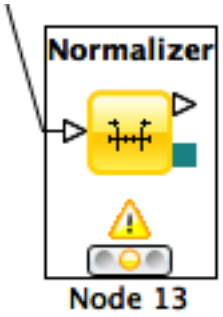
- Provide a short document (max three pages in pdf, excluding figures/plots) which illustrates the input dataset, the adopted clustering methodology and the cluster interpretation.

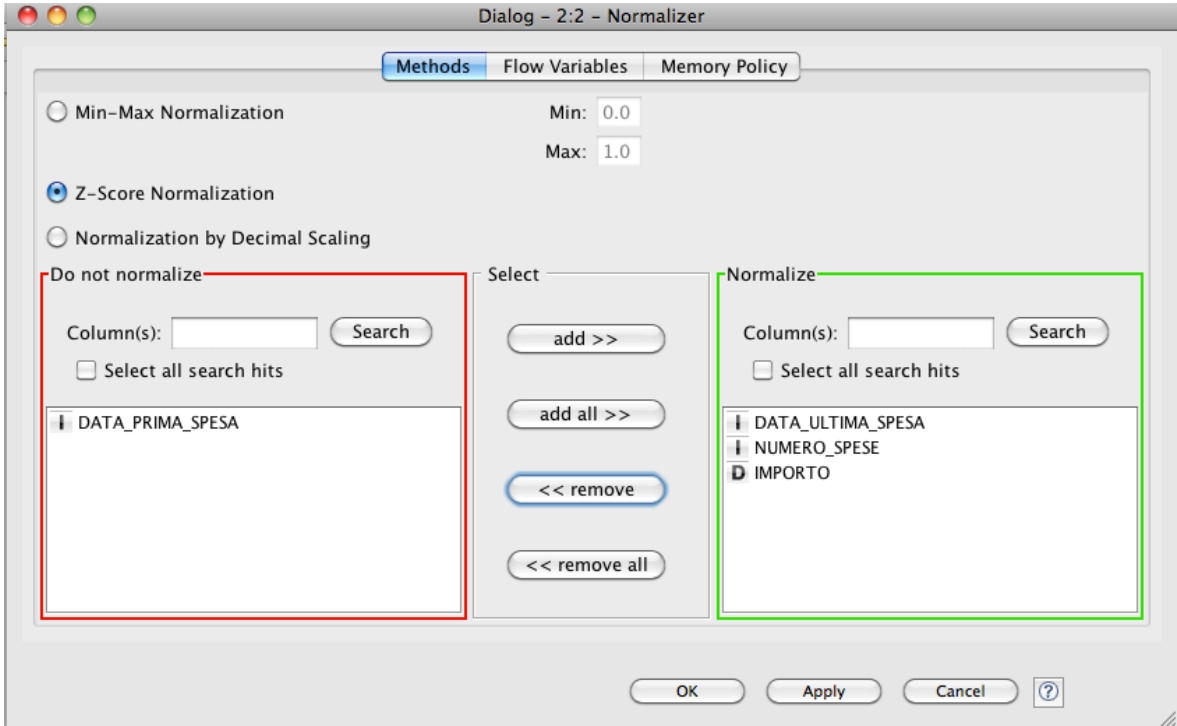
DATA

- **Dataset filename**: rfm_data.csv.
- **Dataset legend**: for each customer, the dataset contains
  - *date_first_purchase*: integer that indicates the date of the first purchase of the customer
  - *date_last_purchase*: integer that indicates the date of the last purchase of the customer
  - *Number of purchases*: number of different purchases in terms of receipts
  - *Amount*: total money spent by the customer
- Need to vompute the columns for
  - *Recency*: no. of days since last purchase
  - *Frequency*: no. of visits (shopping in the supermarket) in the observation period
  - *Monetary value*: total amount spent in purchases during the observation period.

Clustering in KNIME

- Data normalization
  - Min-max normalization
  - Z-score normalization
- Compare the clustering results before and after this operation and discuss the comparison

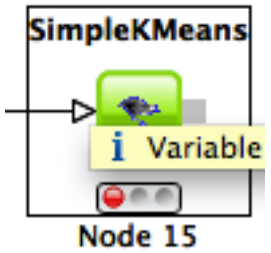




K-Means

- Two options
  - K-means in Mining section of Knime

  - K-means in Weka section of Knime



- The second one allows the SSE computation useful for finding the best k value

CHURN ANALYSIS

Anna Monreale
KDD-Lab, University of Pisa
Email: annam@di.unipi.it

Churn Analysis

- **Problem**: Problem: given a dataset of measurements over a set of customers of an e-commenrce site, find a high-quality classifier, using decision trees, which predicts whether each customers will place only one or more orders to the shop.

- Provide a short document (max three pages in pdf, excluding figures/plots) which illustrates the input dataset, the adopted clustering methodology and the cluster interpretation.

Data

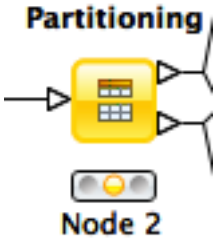
- Filename: *OneShotCustomersEX.csv*
  - Contains transactions from 15,000 online customers

- In the web page of the course you can download the attribute description

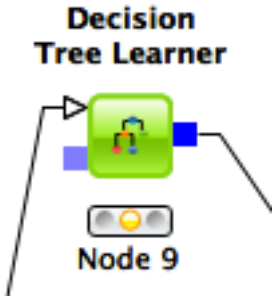
- The class of the data is **Customer Typology** that can be
  - **one shot = only 1 purchase**
  - **loyal = more than one purchase**

Decision Trees in Knime

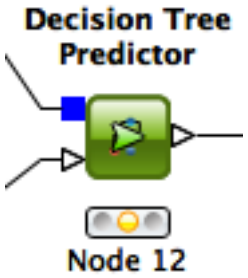
- For Classification by decision trees
  - Partitioning of the data in training and test set



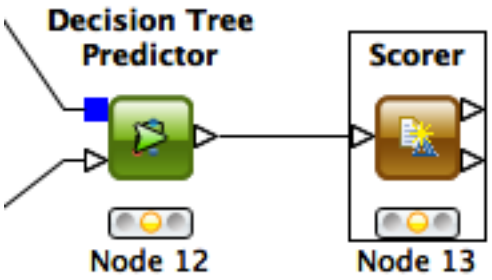
  - On the training set applying the learner

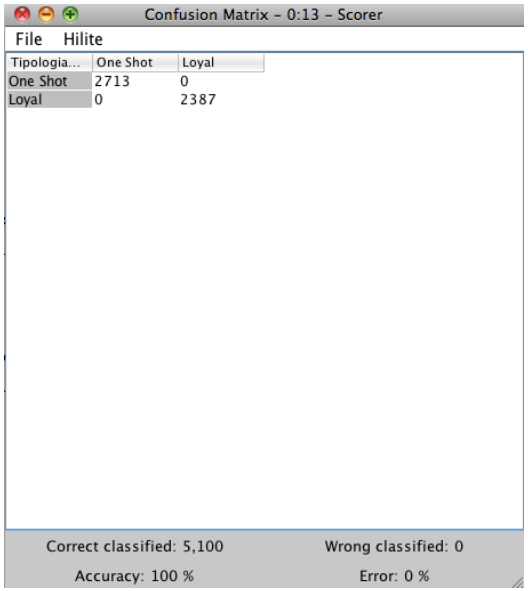


  - On the test set applying the predictor



Evaluation of our classification model





Accuracy statistics - 0:13 - Scorer

File

Table "default" - Rows: 3 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure
One Shot	2713	0	2387	0	1	1	1	1	1
Loyal	2387	0	2713	0	1	1	1	1	1
Overall	?	?	?	?	?	?	?	?	?