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
- New software (nodes) from update sites
  - http://tech.knime.org/update/community-contributions/realease
- Workflows and data are stored in a *workspace*

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

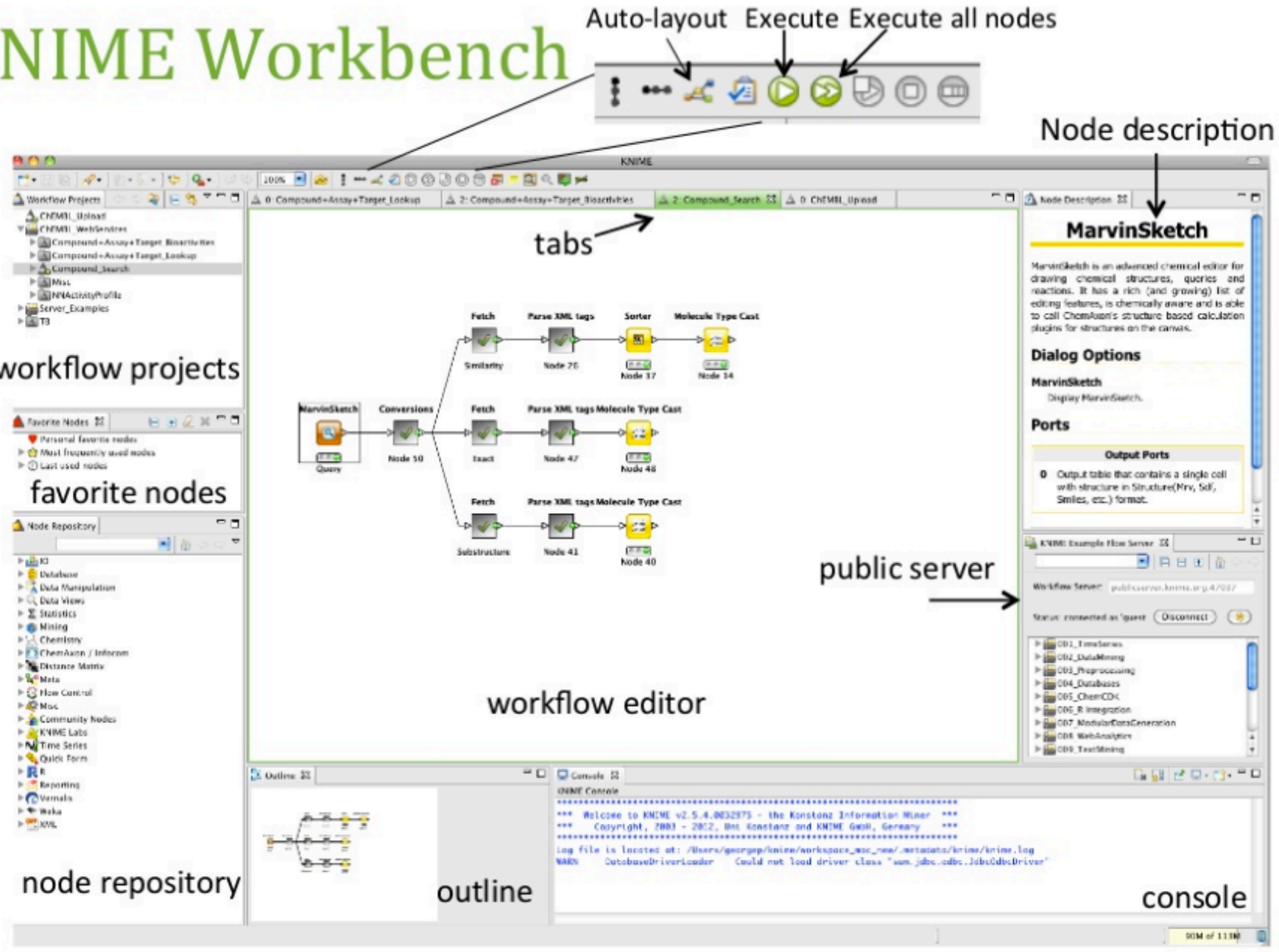
Mac OS X

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:

What can you do with KNIME?

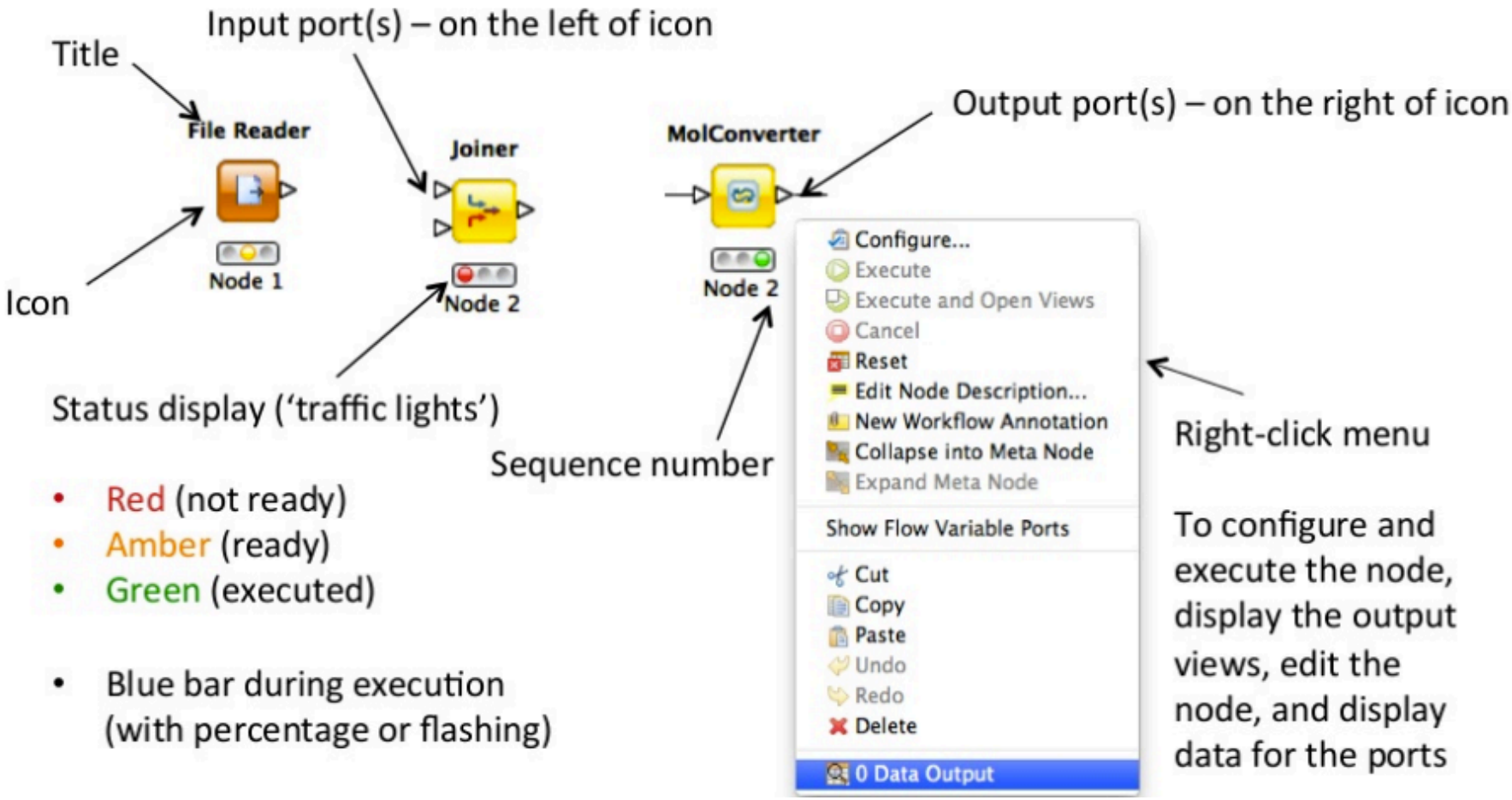
- **Data manipulation and analysis**
  - File & database I/O, filtering, grouping, joining, ….
- **Data mining / machine learning**
  - WEKA, R, Interactive plotting
- **Scripting Integration**
  - R, Perl, Python, Matlab …
- **Much more**
  - Bioinformatics, text mining and network analysis

KNIME Workbench



KNIME nodes: Overview

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



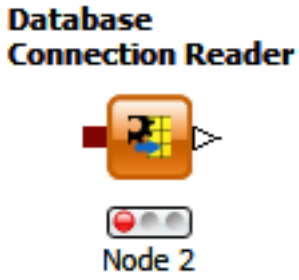
- Red (not ready)
- Amber (ready)
- Green (executed)

- Blue bar during execution (with percentage or flashing)

Ports

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

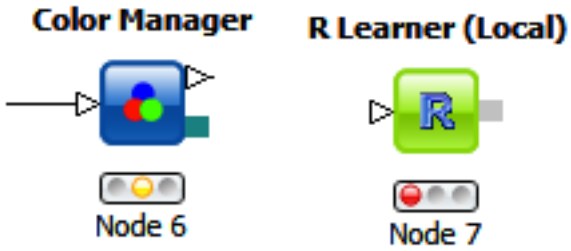
- **Database Port:** Nodes executing commands inside a database are recognized by their database ports (brown square)



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

Other Ports

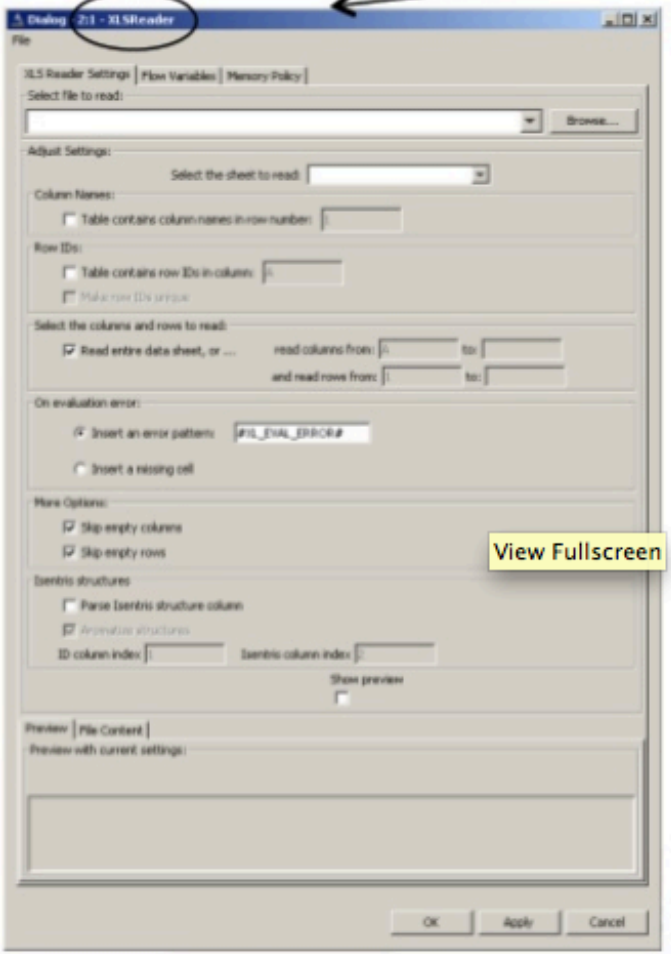
- Whenever a node provides data that does not fit a flat data table structure, a general purpose port for structured data is used (dark cyan square).
- All ports not listed above are known as "unknown" types (gray square).

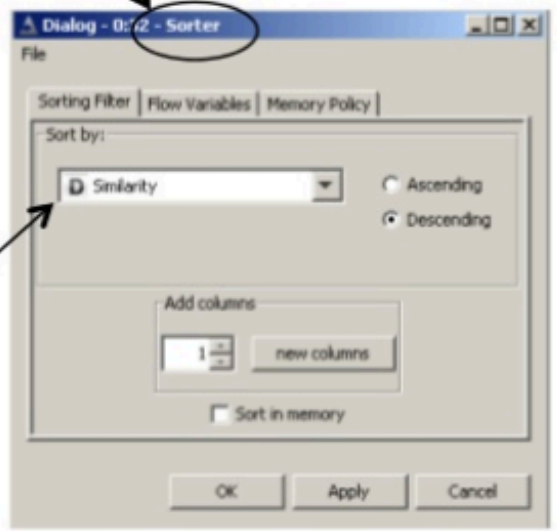


KNIME nodes: Dialogs

Double click to configure…

Configuration menus for selected nodes

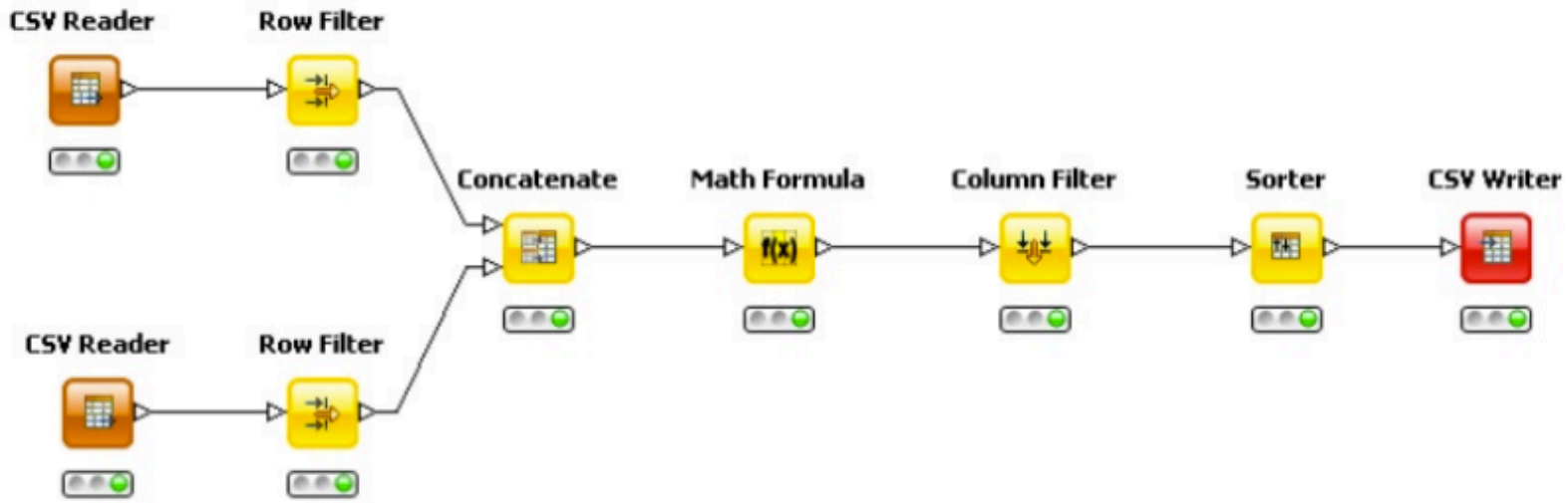




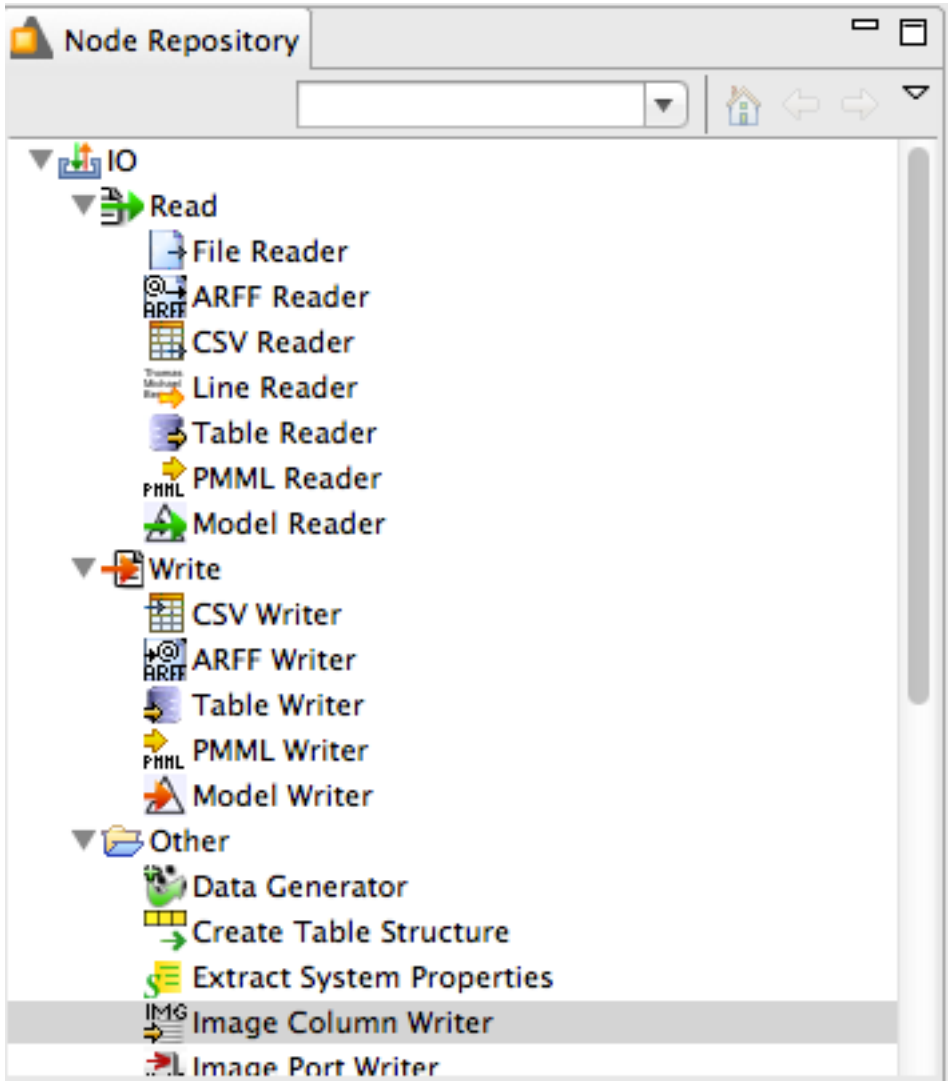
Explicit column type

An example 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…



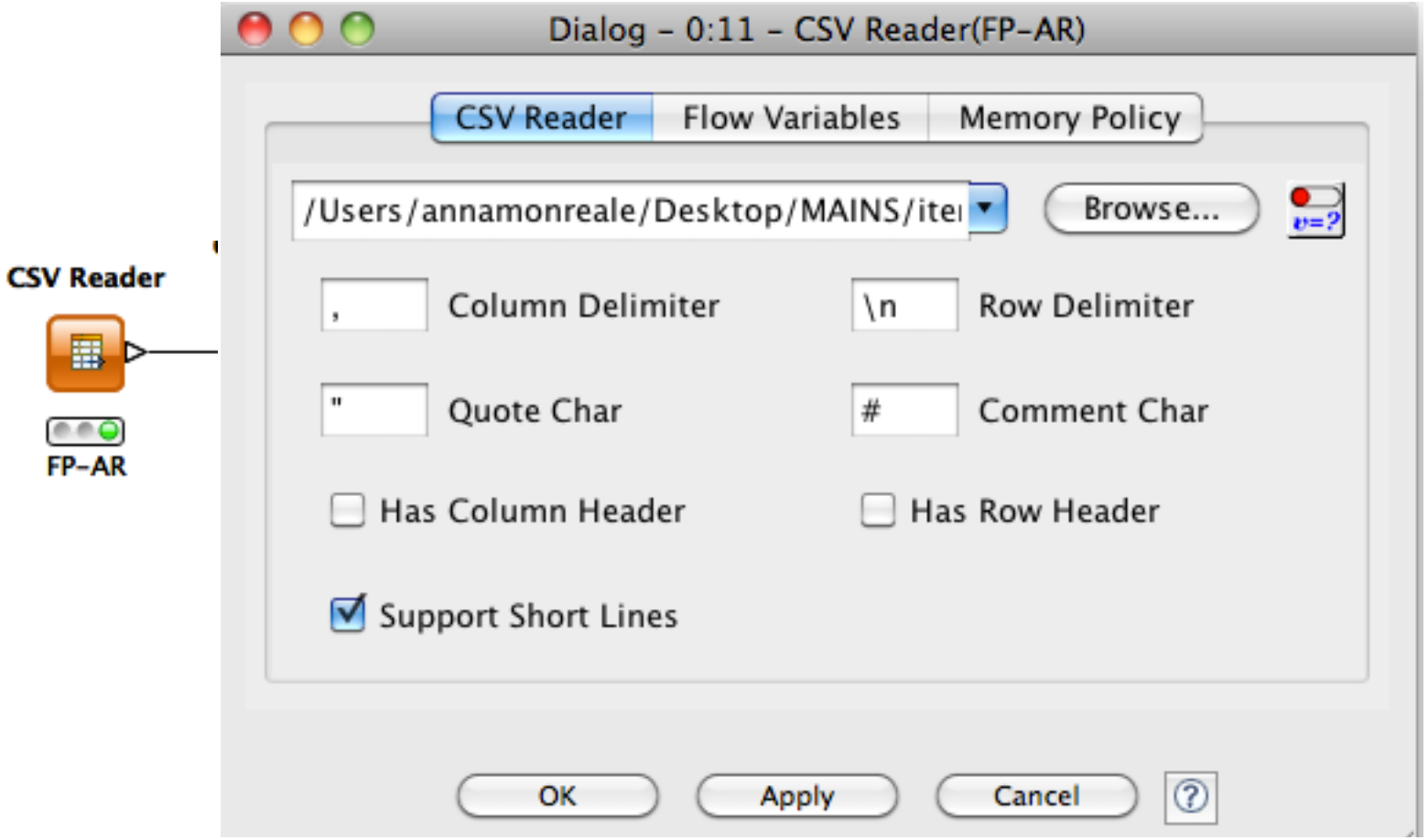
I/O Operations



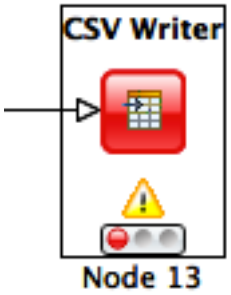
ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

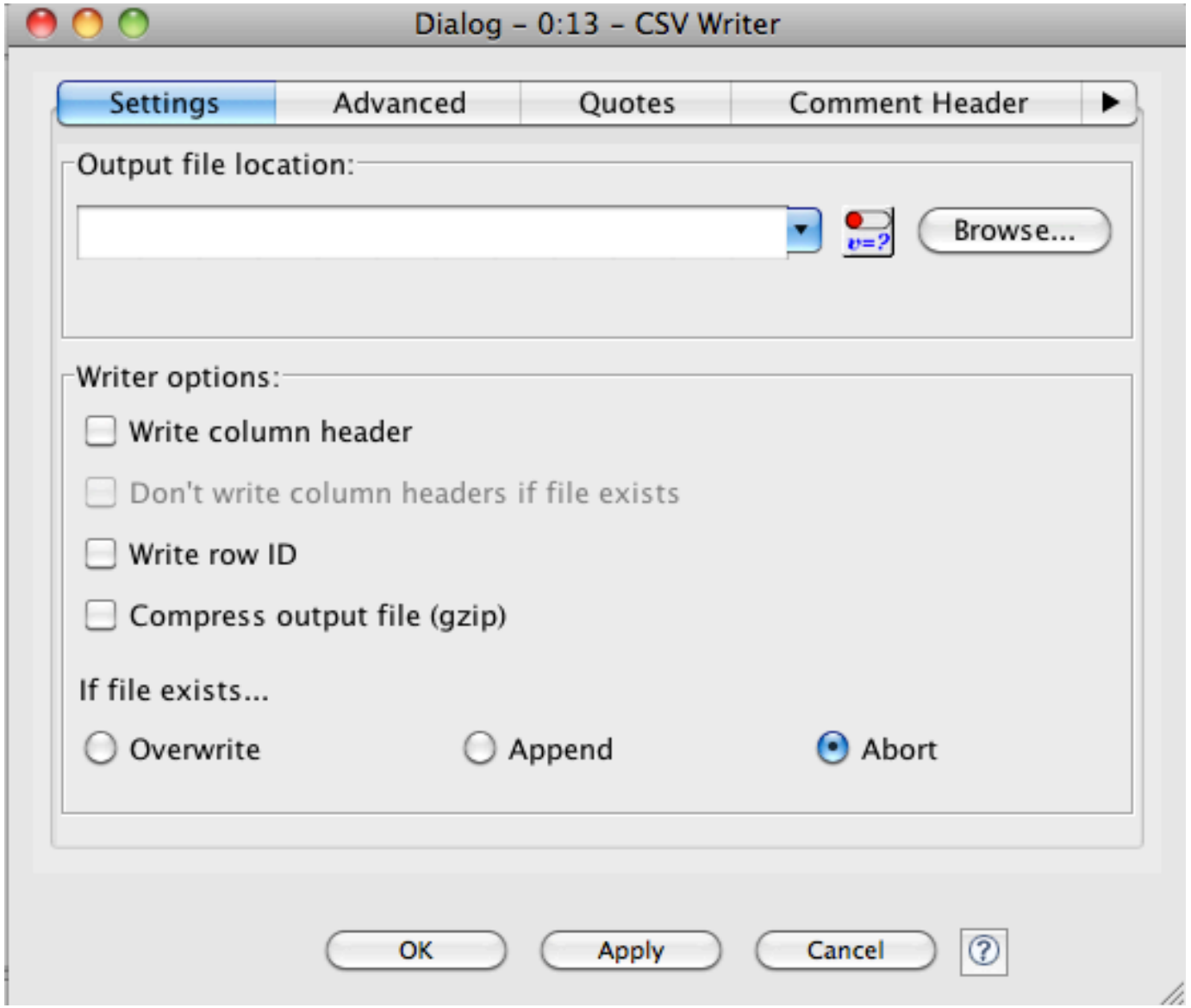
CSV (Comma-Separated Values) file stores tabular data (numbers and text) in plain-text form.

CSV Reader


CSV Reader
FP-AR
Dialog - 0:11 - CSV Reader(FP-AR)
CSV Reader
Flow Variables
Memory Policy
/Users/annamonreale/Desktop/MAINS/ite
Browse...
,
Column Delimiter
\n
Row Delimiter
"
Quote Char
#
Comment Char
Has Column Header
Has Row Header
Support Short Lines
OK
Apply
Cancel

CSV Writer


CSV Writer
Node 13

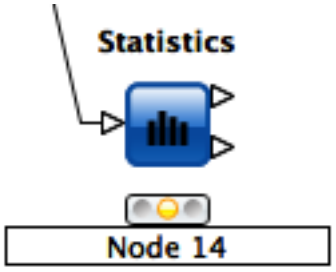

Dialog – 0:13 – CSV Writer
Settings
Advanced
Quotes
Comment Header
Output file location:
Browse...
Writer options:
Write column header
Don't write column headers if file exists
Write row ID
Compress output file (gzip)
If file exists...
Overwrite
Append
Abort
OK
Apply
Cancel

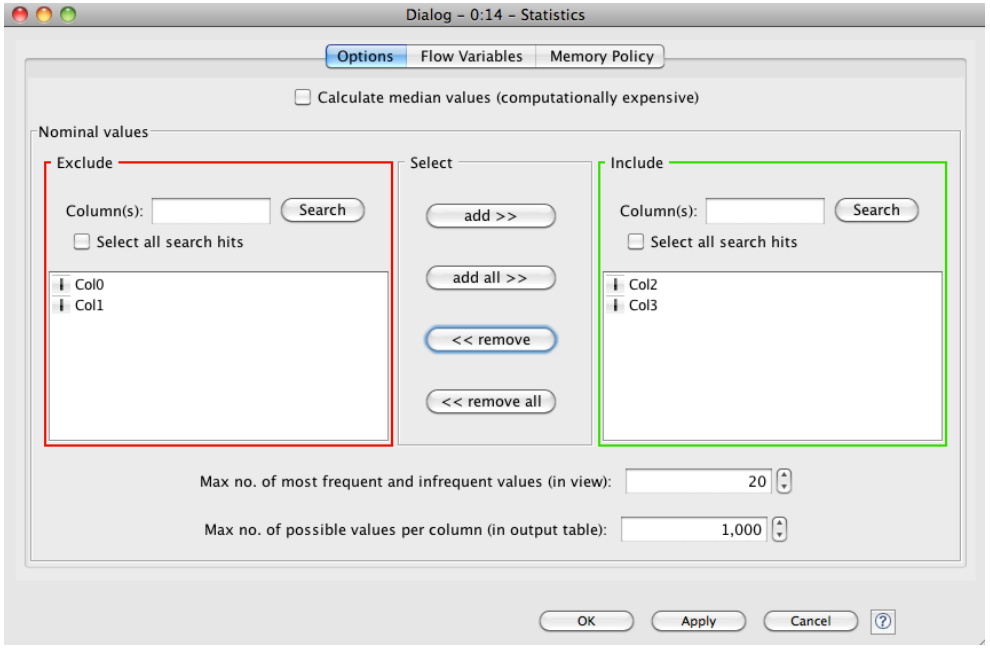
Data Manipulation

- 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.

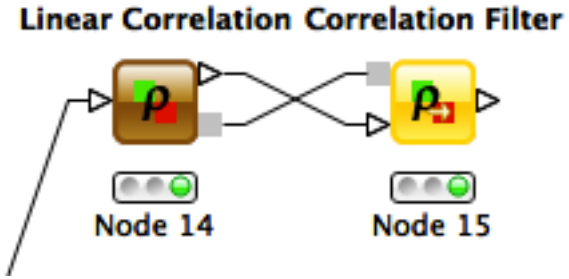




Correlation Analysis

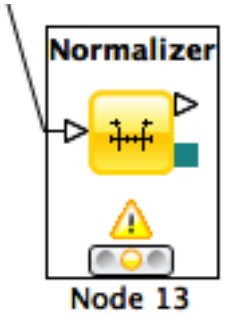
- **Linear Correlation node** computes for each pair of selected columns a correlation coefficient, i.e. a measure of the correlation of the two variables
  - Pearson Correlation Coefficient

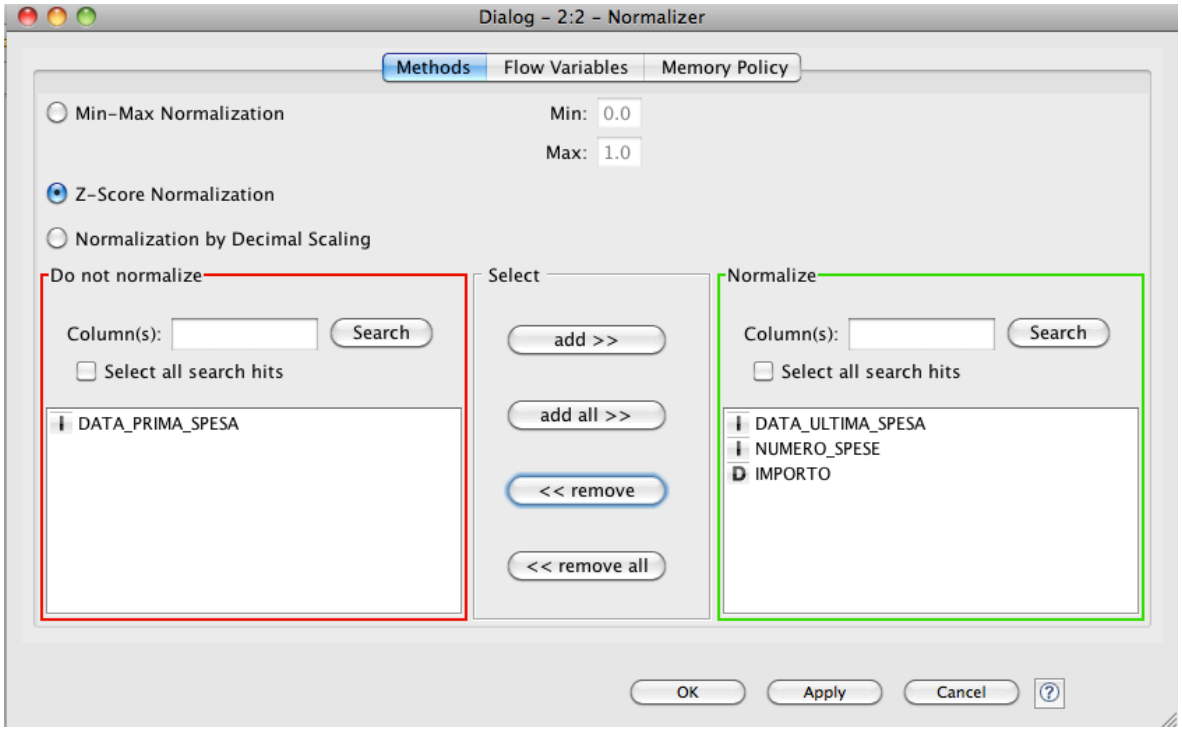
- **Correlation Filtering node** uses the model as generated by a Correlation node to determine which columns are redundant (i.e. correlated) and filters them out.
  - **The output table will contain the reduced set of columns.**



Data Normalization

- Data normalization
  - Min-max normalization
  - Z-score normalization





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
- Weka

MARKET BASKET ANALYSIS

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

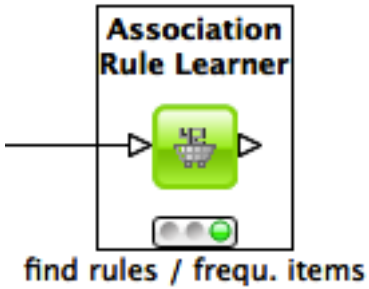
Market Basket Analysis

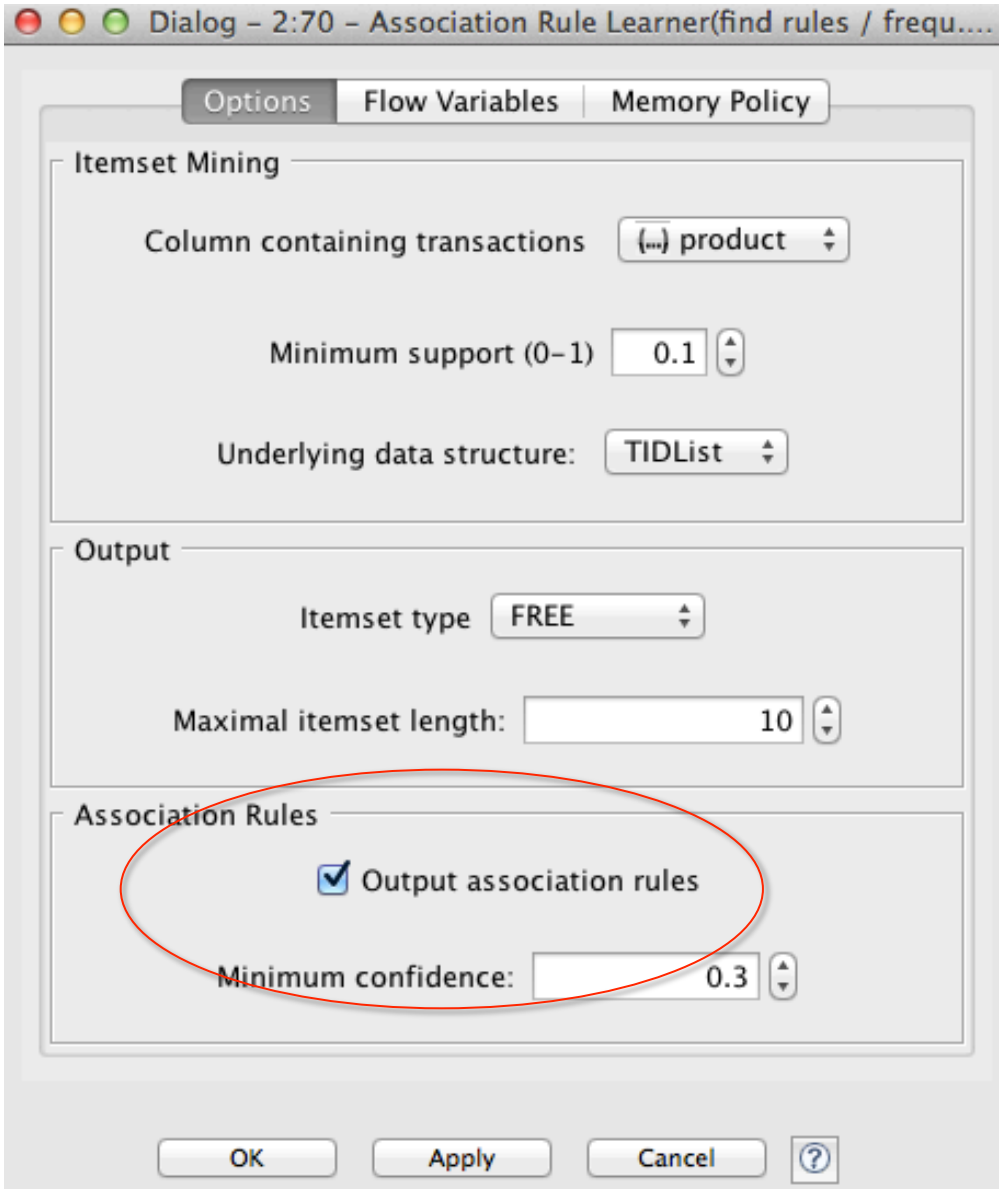
- **Problem**: given a database of transactions of customers of a supermarket, find **the set of frequent items co-purchased** and analyze the **association rules** that is possible to derive from the frequent patterns.

- Knime gives two options:
  - **Item Set Finder node & AR Learner** node implementing Borgelt's algorithms (additional nodes to be installed)
  - **Association Rule node:** computes bothe frequent itemsets and AR (default node in standard Knime installation)

Frequent Patterns and AR in KNIME

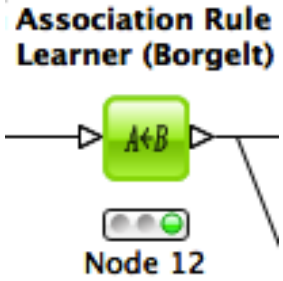
- One node for both task:
  - Association rule learner
  - Frequent pattern extraction





Alternative nodes for the same tasks…

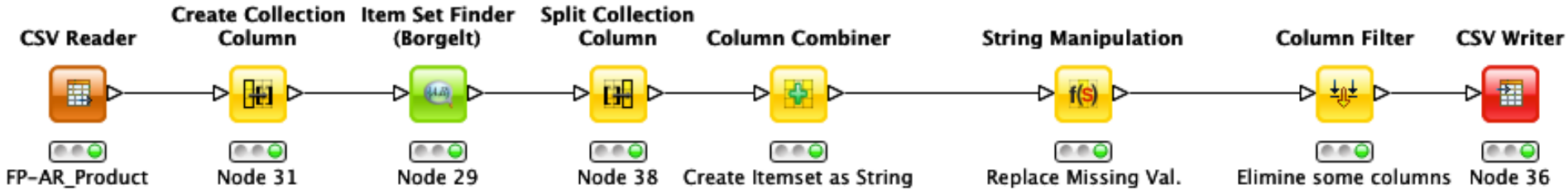
- The two nodes implement the Borgelt's Algorithms:



- **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**

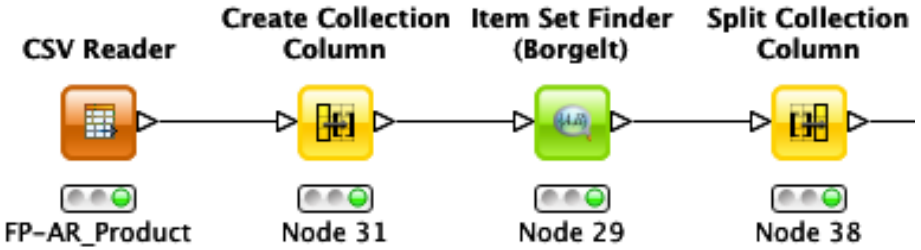
Write the itemsets in a file

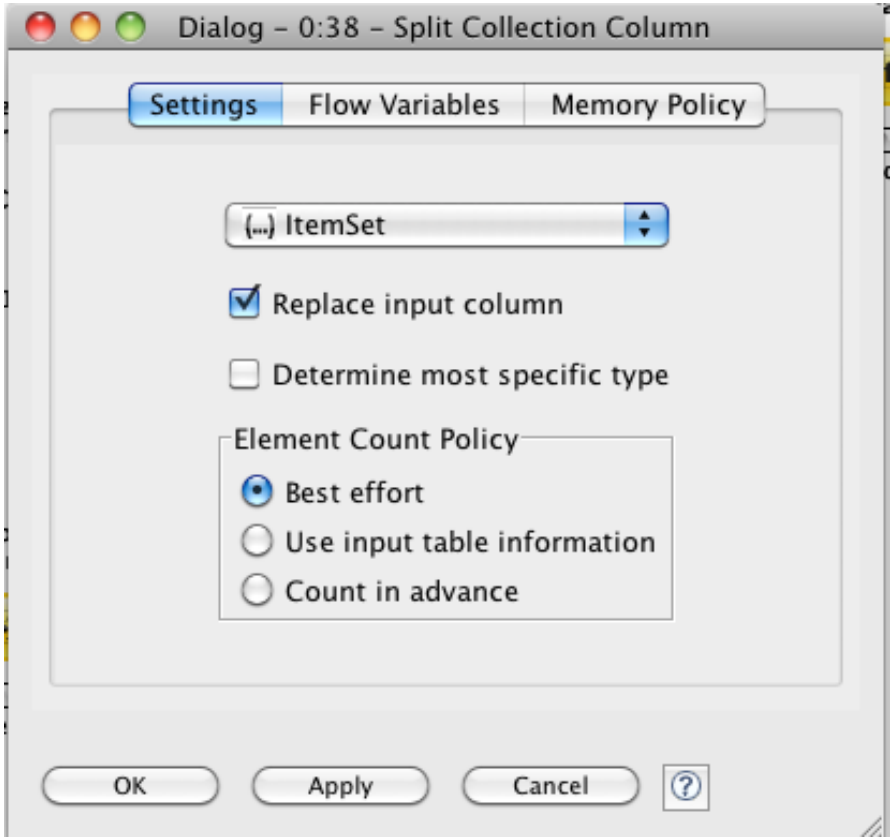
- Given the output of the Item set Finder node sometimes you cannot see all the components of the itemset
  - we need to transform it in a string and
  - then, we can also write the result in a file



Write the itemsets in a file

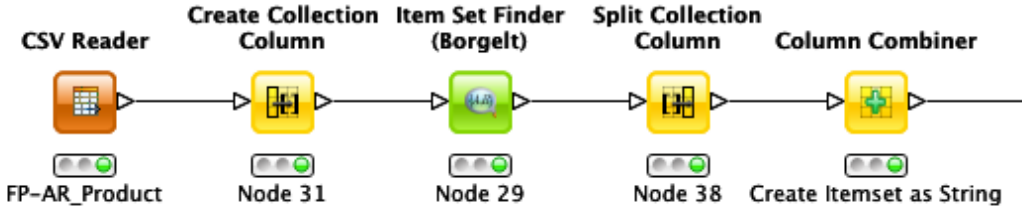
- First we need to split the collection

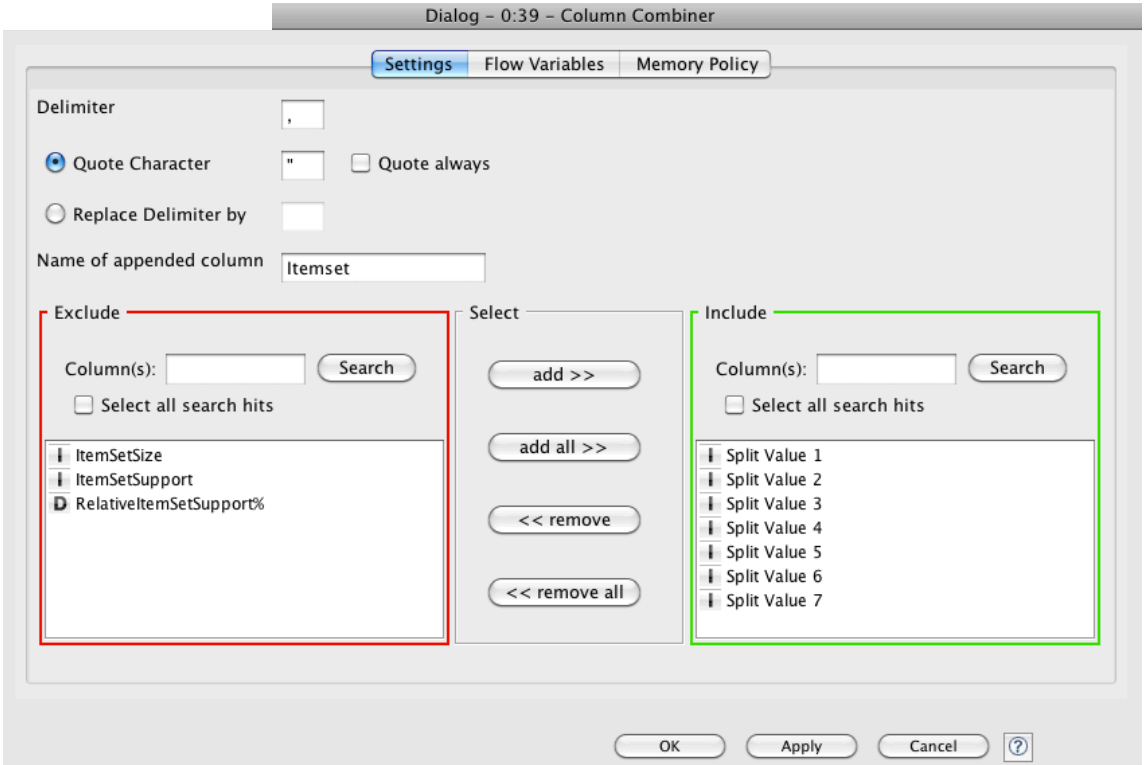




Write the itemsets in a file

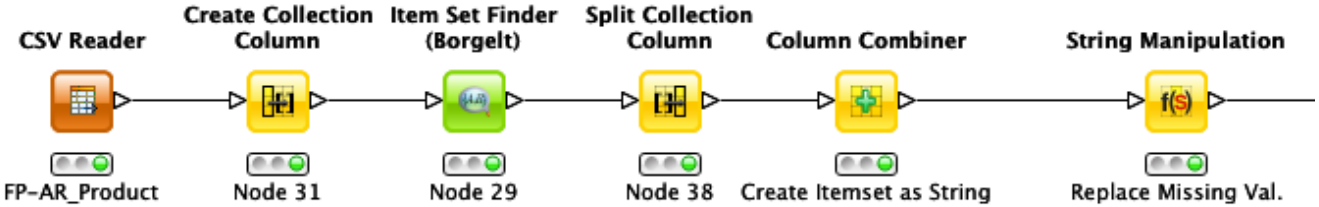
- Second we combine the columns that have to compose the itemset (string)



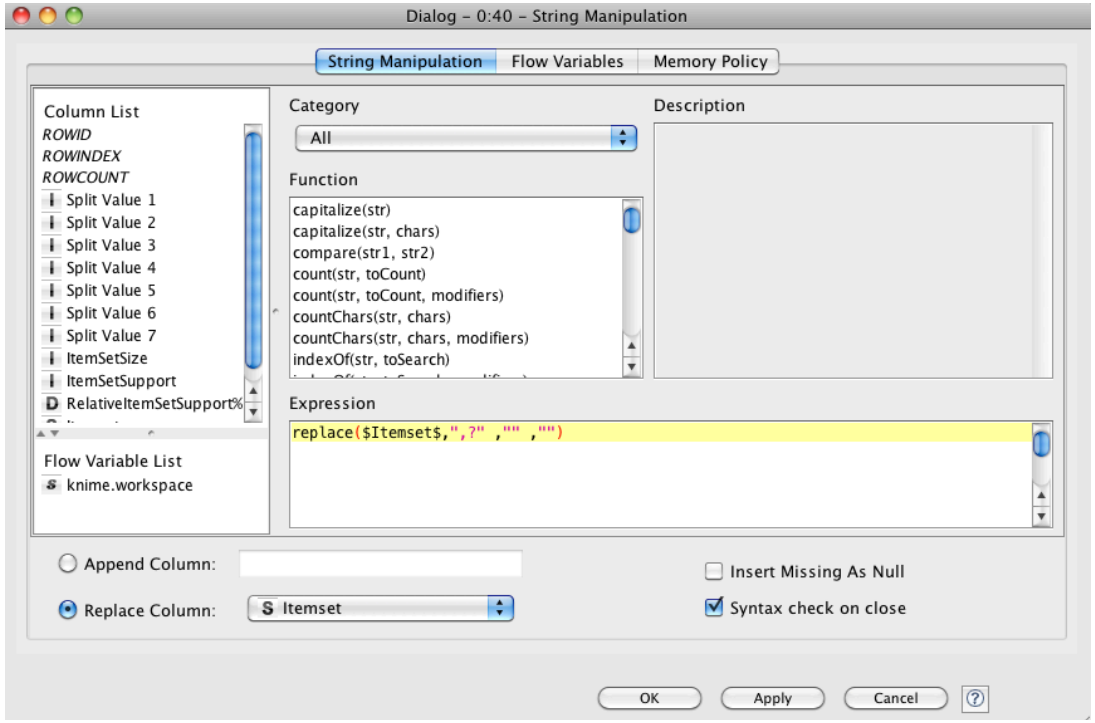


Write the itemsets in a file

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

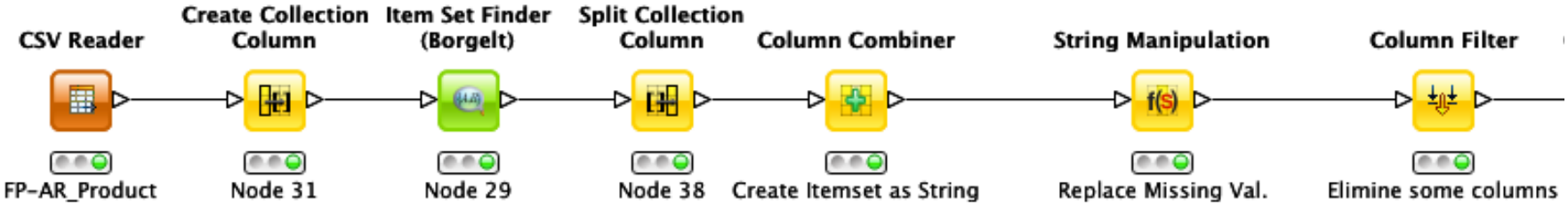


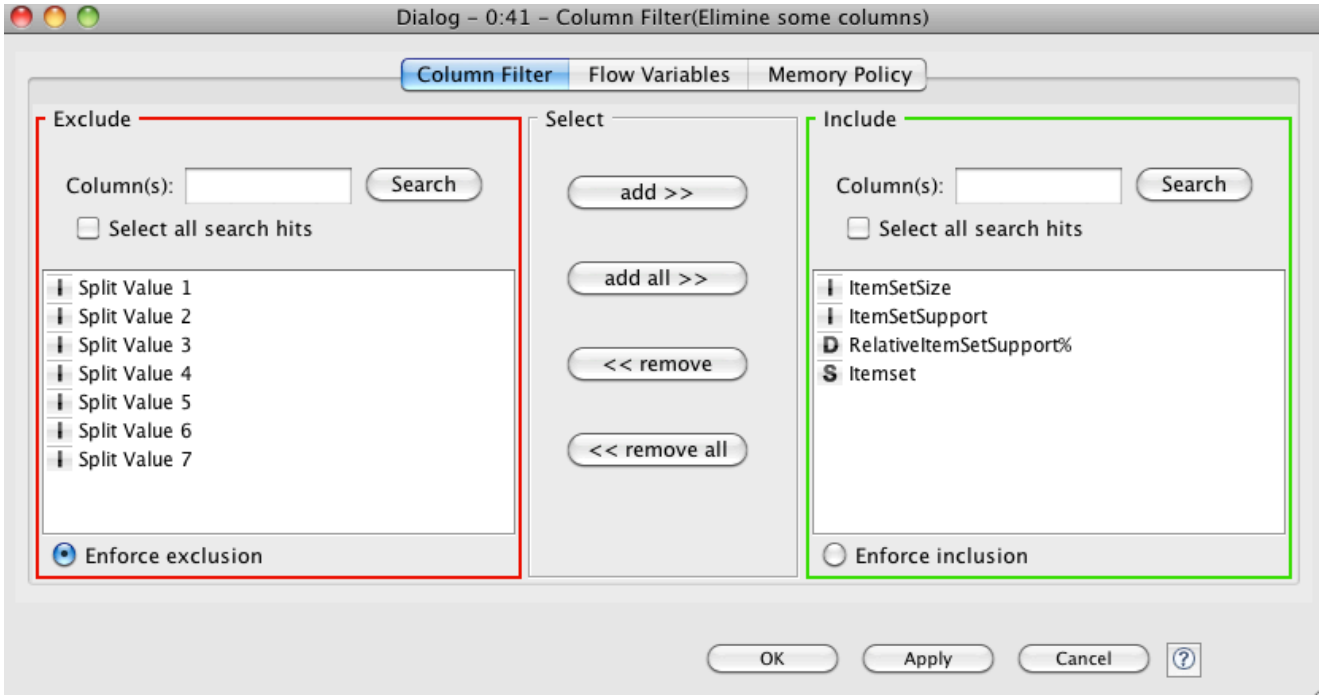
- We use the **replace** operation to eliminate them



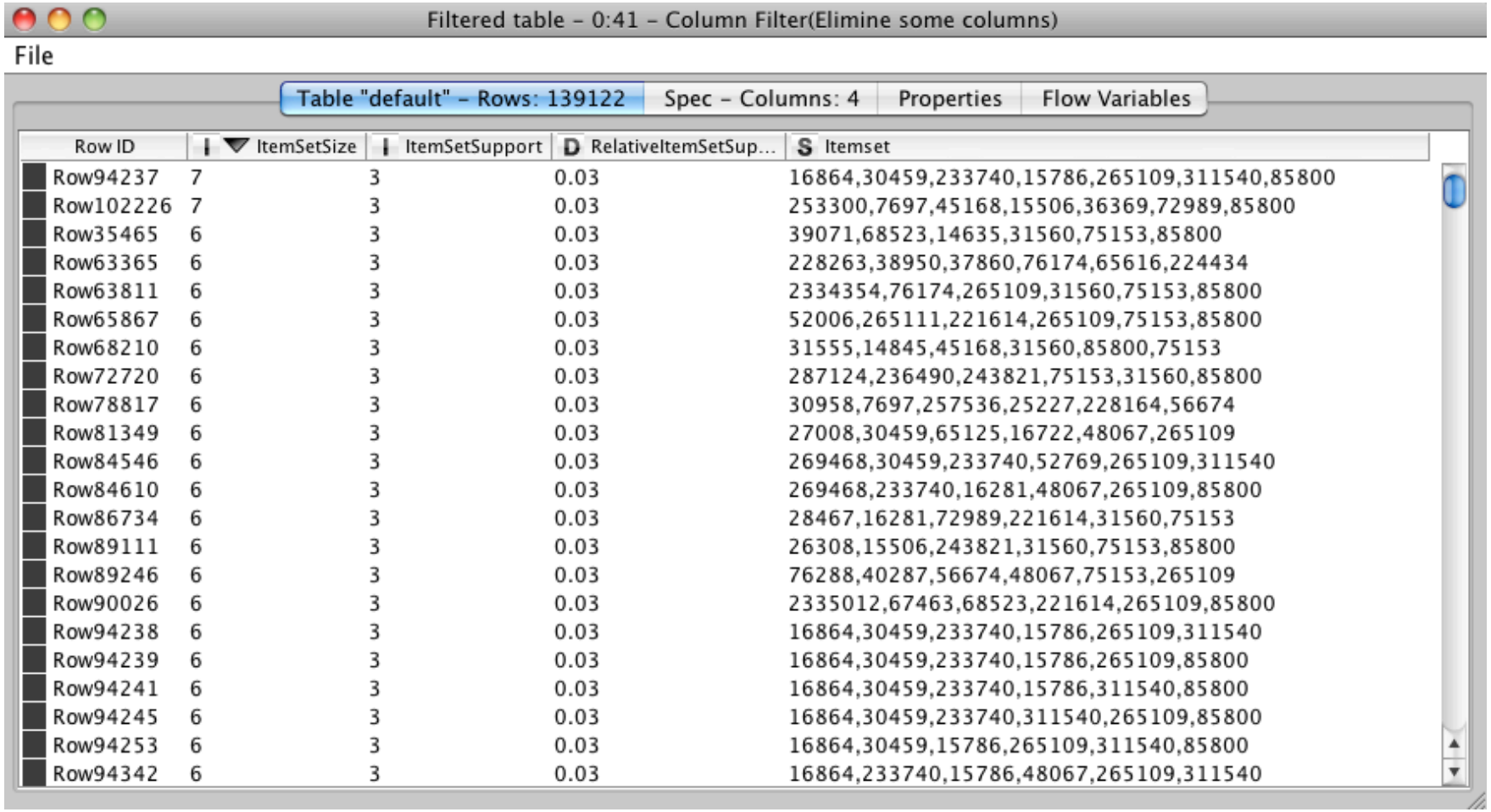
Write the itemsets in a file

- Before writing in a file eliminate the split columns





….. The output table



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**