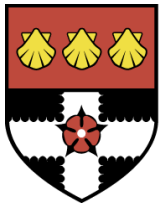


The University of Reading

CSMDM16 - Data Analytics and Mining

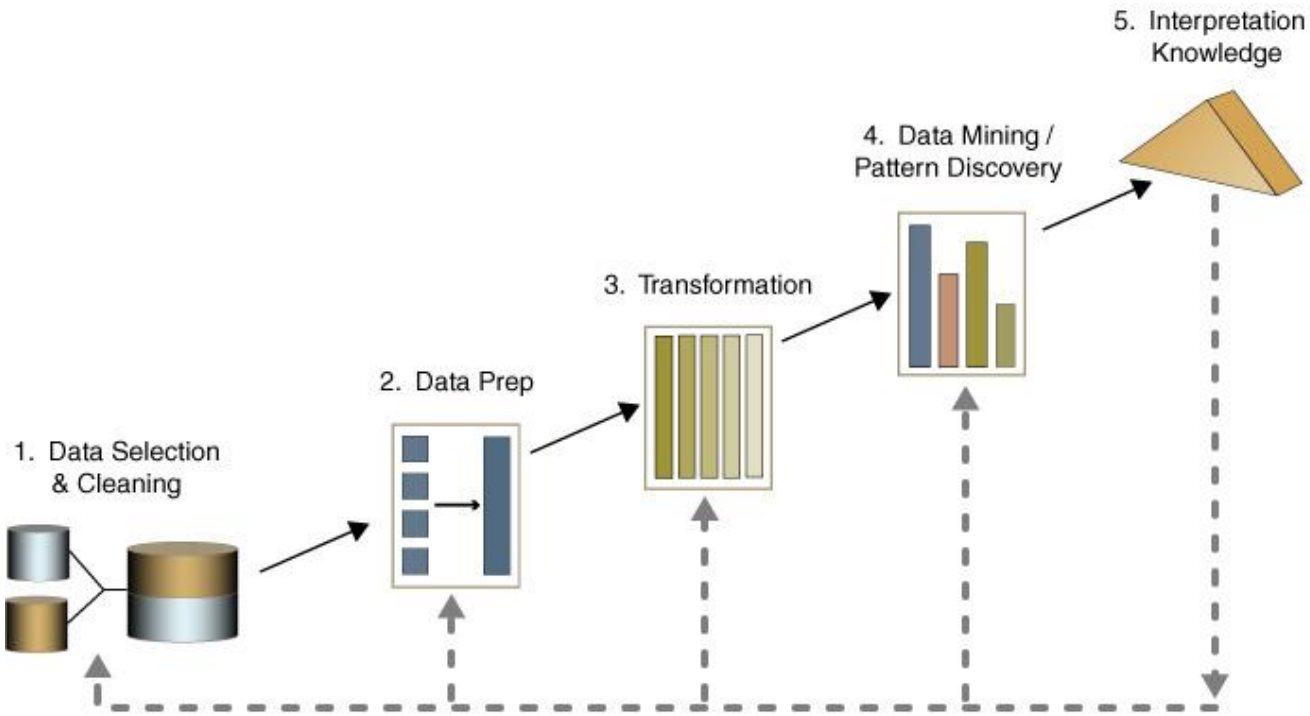
KDD Development Environments: KNIME

Dr. Giuseppe Di Fatta
Associate Professor
Department of Computer Science
G.DiFatta@reading.ac.uk

KDD Development Environments

- Increasing demand for integrated environments to facilitate the KDD process

- **Data mining workflow systems** that integrates analytical data mining methods for prediction, discovery, classification, etc., with data management and information visualization.



An Overview of the Steps That Compose the KDD Process

KDD Development Environments – Open Source

- **Weka** 3, Data Mining Software in Java

- **Orange**, a component-based data mining software (C++)

- **MLC++** is a library of C++ classes for supervised machine learning

- **D2K**, Data to Knowledge (Java)

- **KNIME**, Konstanz Information Miner (Java)

KDD Development Environments - Commercial

- **RapidMiner** (formerly **YALE**, Yet Another Learning Environment) (Java) – free trial version available

- Pentaho – free trial/light version available
  Also free community edition: http://community.pentaho.com/
  (Note: Pentaho Data Mining is based on Weka)

- IBM SPSS (Statistical Package for Social Science)

- SAS

- STATISTICA (Dell acquired StatSoft in March 2014)

2014 KDnuggets Poll

http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html

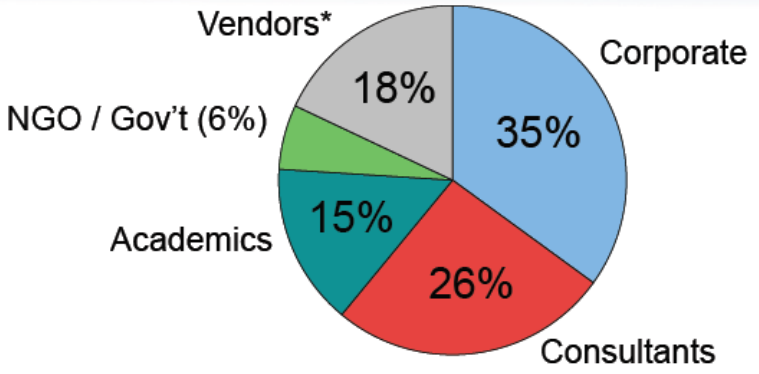
- The 15th annual KDnuggets Software Poll with over 3,000 voters
- Poll: what Analytics, Data Mining, Data Science software/tools you used in the past 12 months for a real project

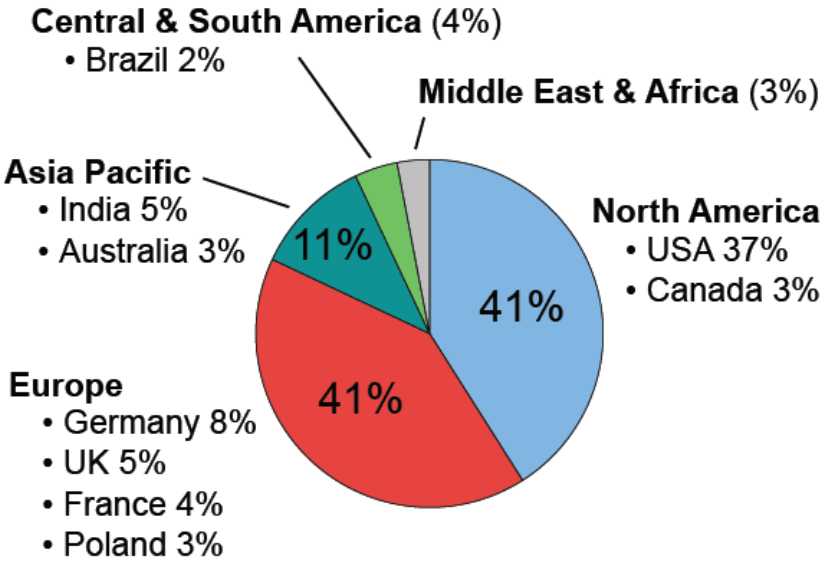
- The top 10 tools by share of users were:
  - RapidMiner, 44.2% share (39.2% in 2013)
  - **R, 38.5% (37.4% in 2013)**
  - Excel, 25.8% (28.0% in 2013)
  - SQL, 25.3% (na in 2013)
  - Python, 19.5% (13.3% in 2013)
  - Weka, 17.0% (14.3% in 2013)
  - **KNIME, 15.0% (5.9% in 2013)**
  - Hadoop, 12.7% (9.3% in 2013)
  - SAS base, 10.9% (10.7% in 2013)
  - Microsoft SQL Server, 10.5% (7.0% in 2013)

2013 Data Miner Survey: Overview

Vendors are included in this analysis.

- 6th survey since 2007
- 68 questions
- 10,000+ invitations emailed, plus promoted by newsgroups, vendors, and bloggers
- Respondents: 1,259 data miners from 75 countries
- Data collected in first half of 2013





*Data from software vendors is excluded from analyses in this presentation unless otherwise noted.

Some Key Findings:

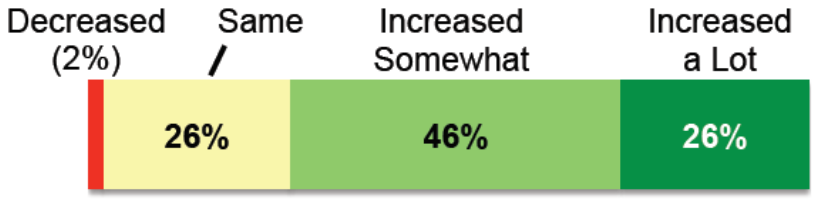
- **BIG DATA**: Many in the field are talking about the phenomena of Big Data. There are clearly some areas in which the volume and sources of data have grown. However it is unclear how much Big Data has impacted the typical data miner. While data miners believe that the size of their datasets have increased over the past year, data from previous surveys indicate that the size of datasets have been fairly consistent over time.
- **THE ASCENDANCE OF R**: The proportion of data miners using R is rapidly growing, and since 2010, R has been the most-used data mining tool. While R is frequently used along with other tools, an increasing number of data miners also select R as their primary tool.
- **ENGAGEMENT & JOB SATISFACTION**: The Data Miners in our survey are highly engaged with the analytic community: consuming and producing content, entering competitions and searching for education and growth within their jobs. All of these activities lead to high job satisfaction, which has been increasing over time.
- **ANALYTIC SOFTWARE**: Data miners are a diverse group who are looking for different things from their data mining tools. Ease-of-use and cost are two distinguishing dimensions. Software packages vary in their strengths and features. STATISTICA, KNIME, SAS JMP and IBM SPSS Modeler all receive high satisfaction ratings.

Big Data: Hype or Reality?

There is a lot of talk in the business and technical press about Big Data. Clearly some businesses and scientific areas are working with very large data sets. However, it is unclear how much Big Data has impacted the typical data miner.

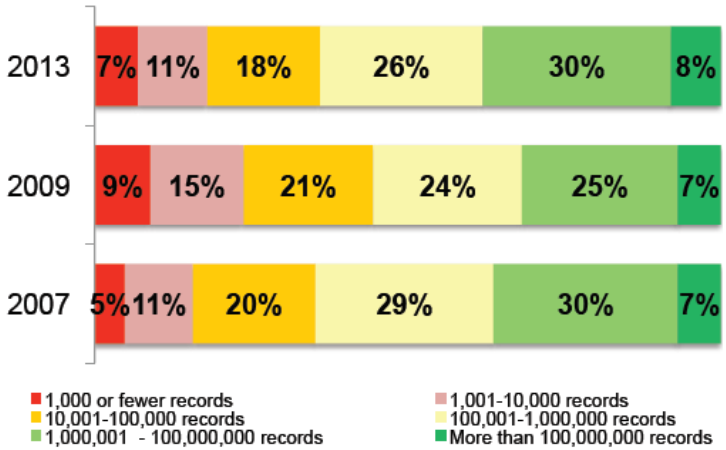
In 2013, the general perception among data miners is that data volumes have increased (72% say it has). However, the datasets they report using are of similar size to what was reported in 2007. Additionally, only 13% report that their company has an active big data program.

2013: Perception of Data Size Increase



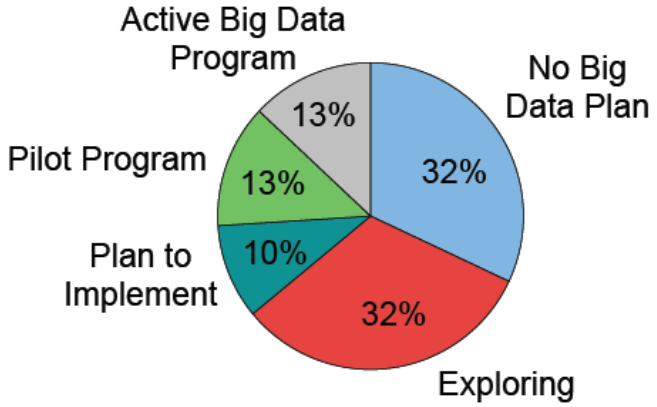
Question: Has the volume/size of data that you use in your analyses increased in the last two years?

Typical Data Set Size



Question: What size data sets did you typically data mine in the past year?

2013: Your Company's Big Data Plan



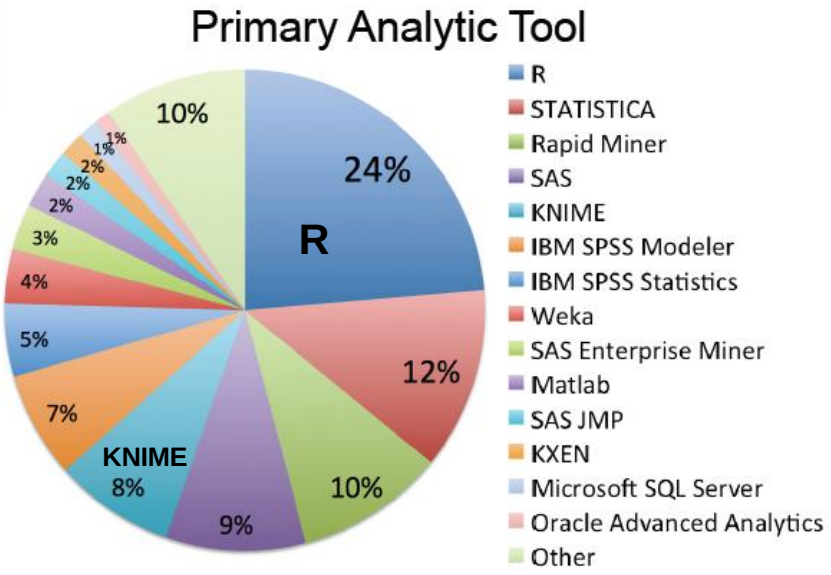
Question: What is your company / organization doing with regards to Big Data?

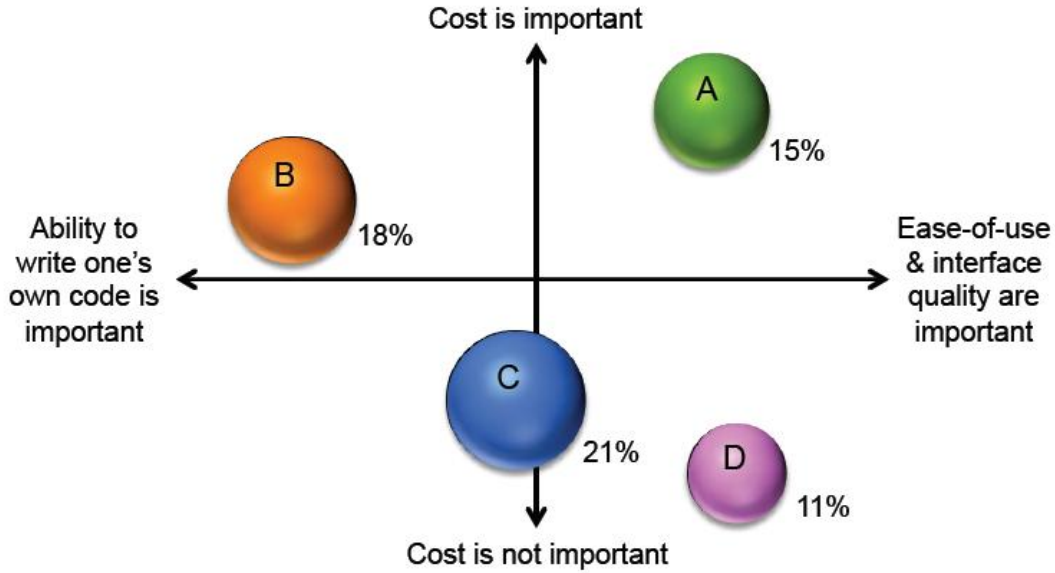
Rexer Analytics - 2013 Data Miner Survey

Tool Selection

Data miners are a diverse group who are looking for different things from their data mining tools. They report using multiple tools to meet their analytic needs, and even the most popular tool is identified as their primary tool by just 24% of data miners. Over the years, R and Rapid Miner have shown substantial increases.

Cluster analysis* reveals that, in their tool-selection preferences, data miners fall into 5 groups. The primary dimensions that distinguish them are price sensitivity and code-writing / interface / ease-of-use preferences.



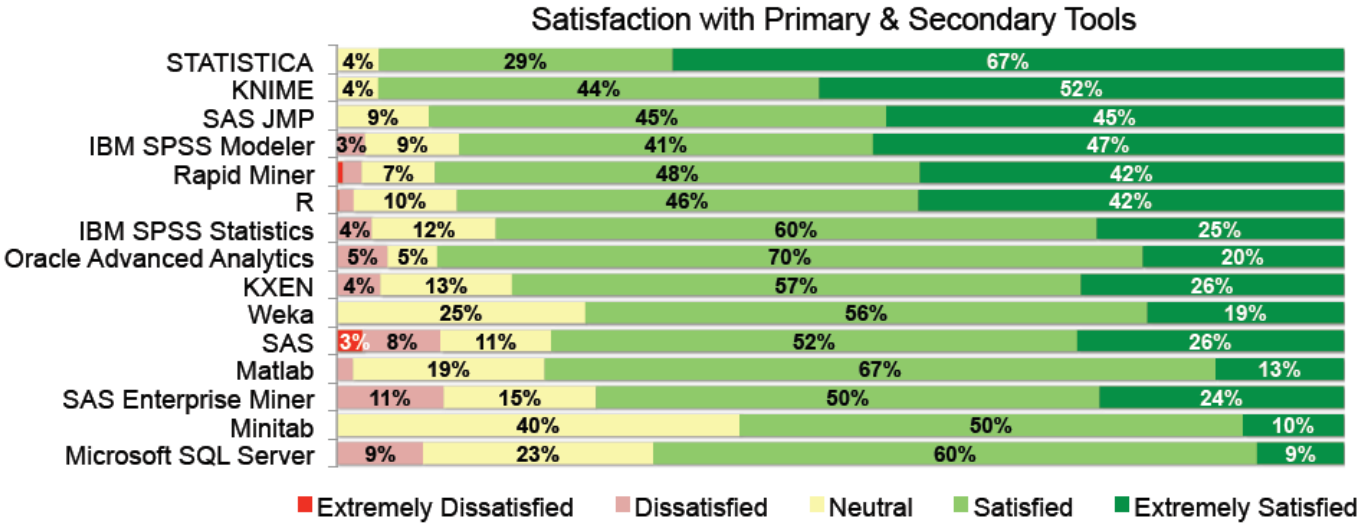


*Cluster analysis was conducted on data miners' ratings of the importance of 22 tool selection factors.

Tool Satisfaction

Most data miners are happy with their analytic software. STATISTICA and KNIME have particularly high satisfaction ratings (they also had the highest ratings in the 2011 survey). SAS JMP, IBM SPSS Modeler, Rapid Miner and R also have high ratings. While people are more satisfied with their primary tools, the patterns of primary and secondary tool satisfaction are generally similar. However, people choosing IBM SPSS Statistics as their secondary tool give it high ratings, while people using SAS Enterprise Miner and IBM SPSS Modeler as their secondary tools give these tools lower ratings.

Most people also report that they will continue using their primary tools – the highest continuation rate is among people choosing KNIME as their primary tool: 85% report that they are "extremely likely" to continue using it as their primary tool for the next 3 years. R and STATISTICA users also report especially high continuation plans. Across all tools, when people say they are likely to switch primary tools, many are choosing R (see page 16).



Satisfaction question: Please rate your overall satisfaction with [insert name of previously identified software package].

Tool Satisfaction: Details

Overall, data miners express the most satisfaction with the quality and accuracy of their tools' model performance and with the variety of algorithms their tools make available to them. Data miners are least satisfied with their tools' help functions, their graphical visualization of models, and their ability to handle large data sets. STATISTICA received strong ratings across many dimensions.

	Overall	IBM SPSS Statistics	IBM SPSS Modeler	KNIME	R	Rapid Miner	SAS	SAS Enterprise Miner	STATISTICA	Weka
Quality and accuracy of model performance	4.28	3.96	4.15	4.30	4.39	4.25	4.20	4.48	4.62	4.16
Variety of available algorithms	4.27	3.66	4.05	4.36	4.74	4.55	3.91	4.23	4.59	4.46
Data manipulation capabilities	4.19	3.91	4.36	4.54	4.24	4.07	4.50	3.74	4.52	3.48
Dependability/Stability of software	4.19	4.02	3.96	4.27	4.24	4.07	4.28	4.16	4.51	4.03
Ability to automate repetitive tasks	4.18	3.79	3.76	4.42	4.35	4.18	4.26	4.10	4.44	3.76
Quality of output / Ease of interpretation	4.11	3.87	3.89	4.17	4.10	4.18	3.84	4.10	4.59	3.82
Ease of use	4.11	4.10	4.67	4.58	3.59	4.39	3.77	4.27	4.58	4.03
Good metrics of model quality	4.08	3.72	3.89	3.91	4.19	4.17	4.01	4.17	4.50	4.06
Data quality assessment & data preparation capabilities	4.05	3.72	4.27	4.37	4.02	4.00	4.26	3.77	4.41	3.47
Ability to easily incorporate data at different levels of granularity (e.g. transaction data and customer data)	4.03	3.87	4.25	4.21	3.94	4.04	4.14	4.10	4.30	3.59
Cost of software	4.03	3.02	2.89	4.85	4.93	4.86	2.33	2.70	3.91	4.89
Ability to modify algorithm options to fine-tune analyses	4.01	3.26	3.63	3.80	4.35	4.10	3.91	3.94	4.28	4.18
Good variable discovery, profiling and selection	4.00	3.64	4.16	4.07	4.03	4.06	3.78	4.23	4.42	3.77
Quality of user interface	3.97	4.02	4.47	4.54	3.49	4.37	3.66	4.10	4.53	3.54
Ease of model deployment (scoring to other data sets)	3.97	3.42	4.01	4.21	3.87	4.19	3.92	4.00	4.43	3.75
Speed	3.95	3.62	4.01	4.00	3.69	3.95	3.93	3.97	4.54	3.70
Enables mining within one's database	3.92	3.59	4.18	4.08	3.92	3.83	3.78	3.93	4.26	3.69
Ability to handle very large data sets	3.84	3.65	4.19	3.90	3.27	3.59	4.35	4.30	4.56	3.18
Strong graphical visualization of models	3.83	2.94	3.60	3.90	4.14	4.01	3.09	3.77	4.58	3.38
Useful help menu, demos and tutorials	3.82	3.87	3.82	4.05	3.86	3.54	3.67	3.90	4.23	3.50

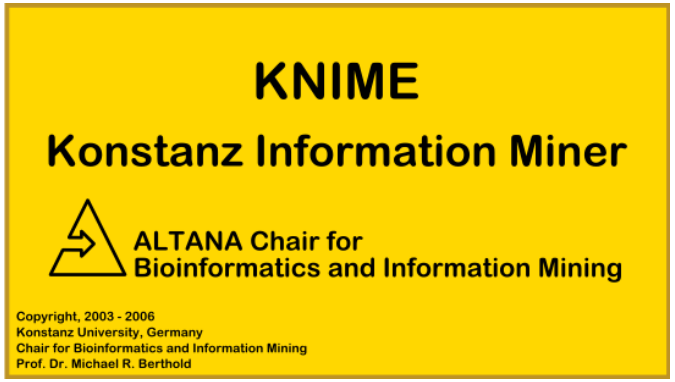
Mean satisfaction rating on 1-5 scale

Higher Satisfaction

Lower Satisfaction

Question: Rate how satisfied you are with the performance of your primary data mining package (identified earlier) on each of these factors.

KNIME



- Developed at the Department of Computer and Information Science, University of Konstanz, Germany

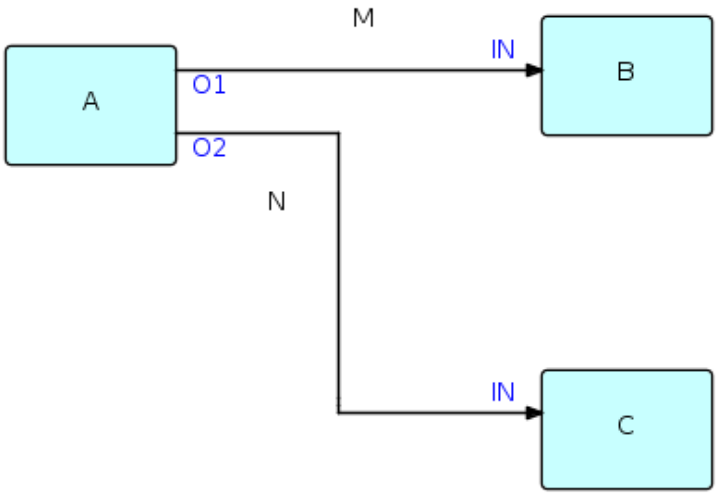
- Under continuous evolution and extension
  - 1st release in April 2006
  - ver.3 released in Dec. 2015
  - current version: 3.2.1

First release and first publication in 2006:

M. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, B. Wiswedel, **"KNIME: the Konstanz Information Miner"**, Proc. of Workshop on Multi-Agent Systems and Simulation (MAS&S), 4th Annual Industrial Simulation Conference (ISC), Palermo, Italy, June 5-7, 2006, pp.58-61.

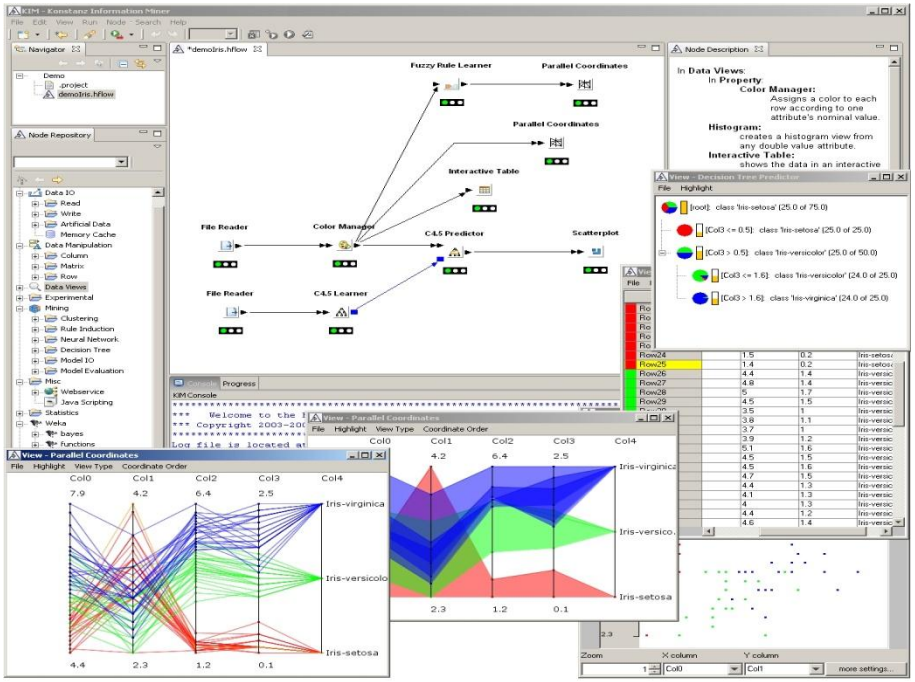
Flow-based Programming

- Flow-based Programming (FBP) is a programming paradigm that defines applications as networks of "black box" processes, which exchange data across predefined connections by message passing, where the connections are specified externally to the processes. These black box processes can be reconnected endlessly to form different applications without having to be changed internally. FBP is thus naturally component-oriented.



- J. Paul Morrison, Flow-Based Programming, 2nd Edition: A New Approach to Application Development, CreateSpace, 2010

Knime: Interactive Data Exploration



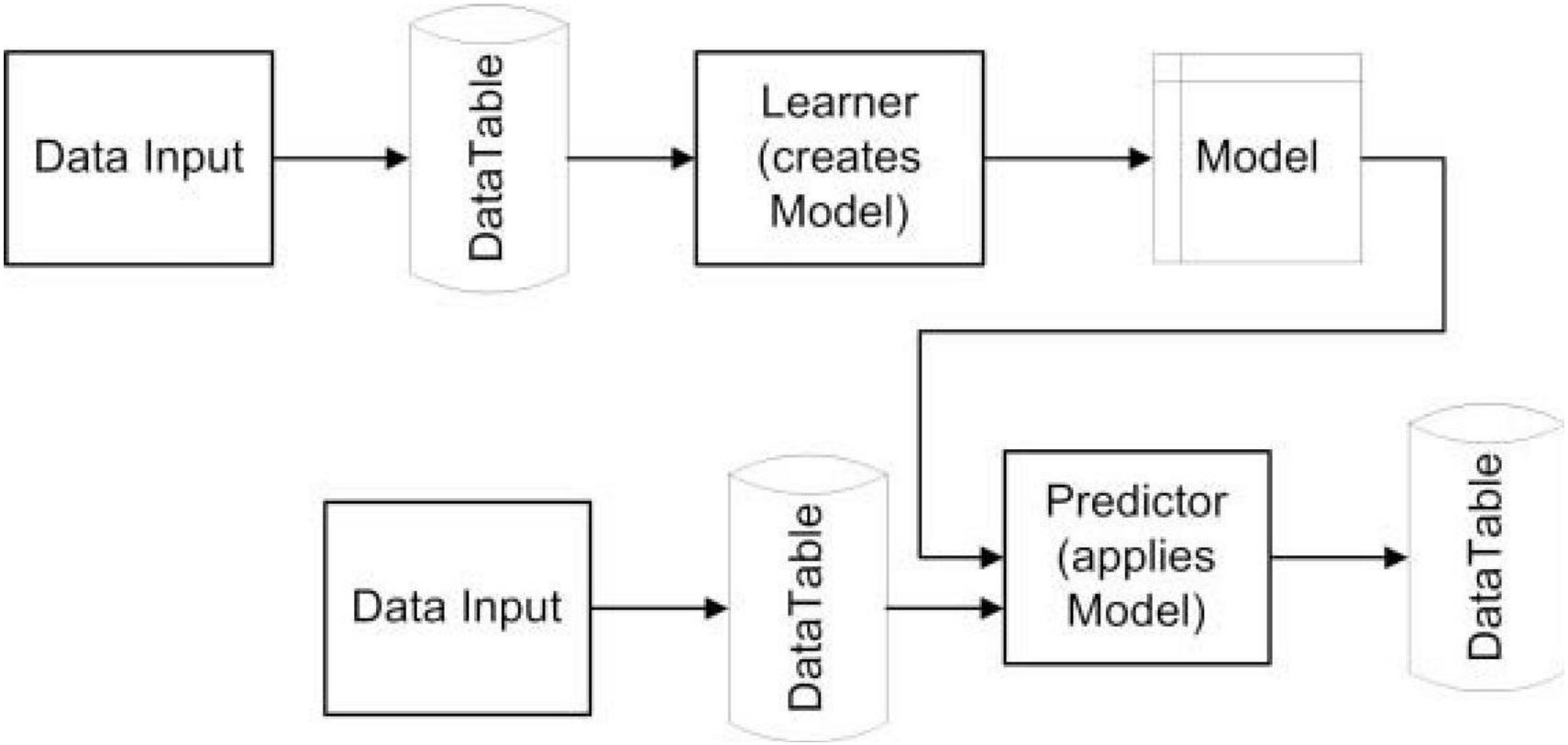
Features:

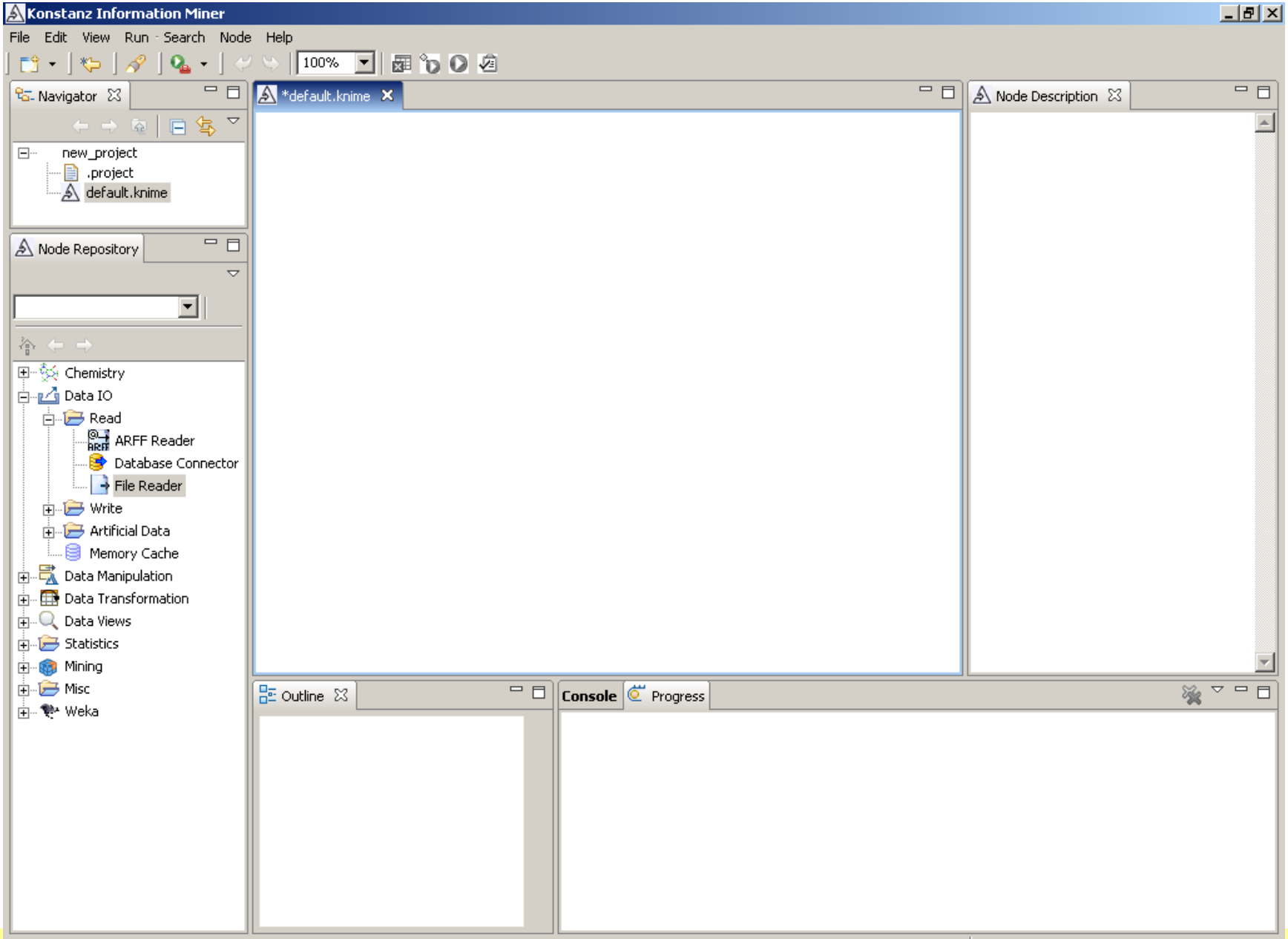
- Modular Data Pipeline Environment
- Large collection of Data Mining techniques
- Data and Model Visualizations
- Interactive Views on Data and Models
- Java Code Base as Open Source Project
- Seamless Integration: R Library, Weka, etc.
- Based on the Eclipse Plug-in technology

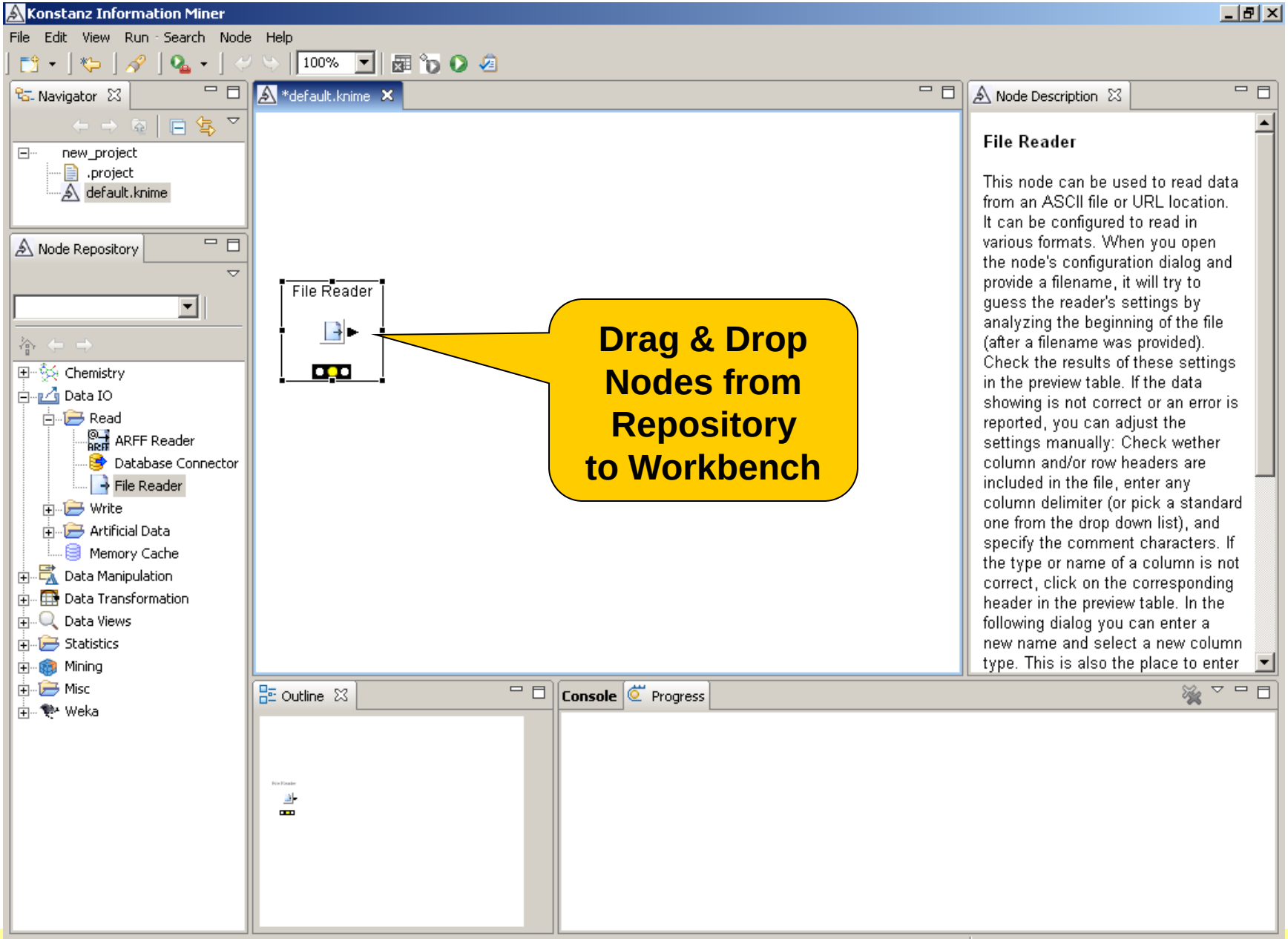
Easy extendibility

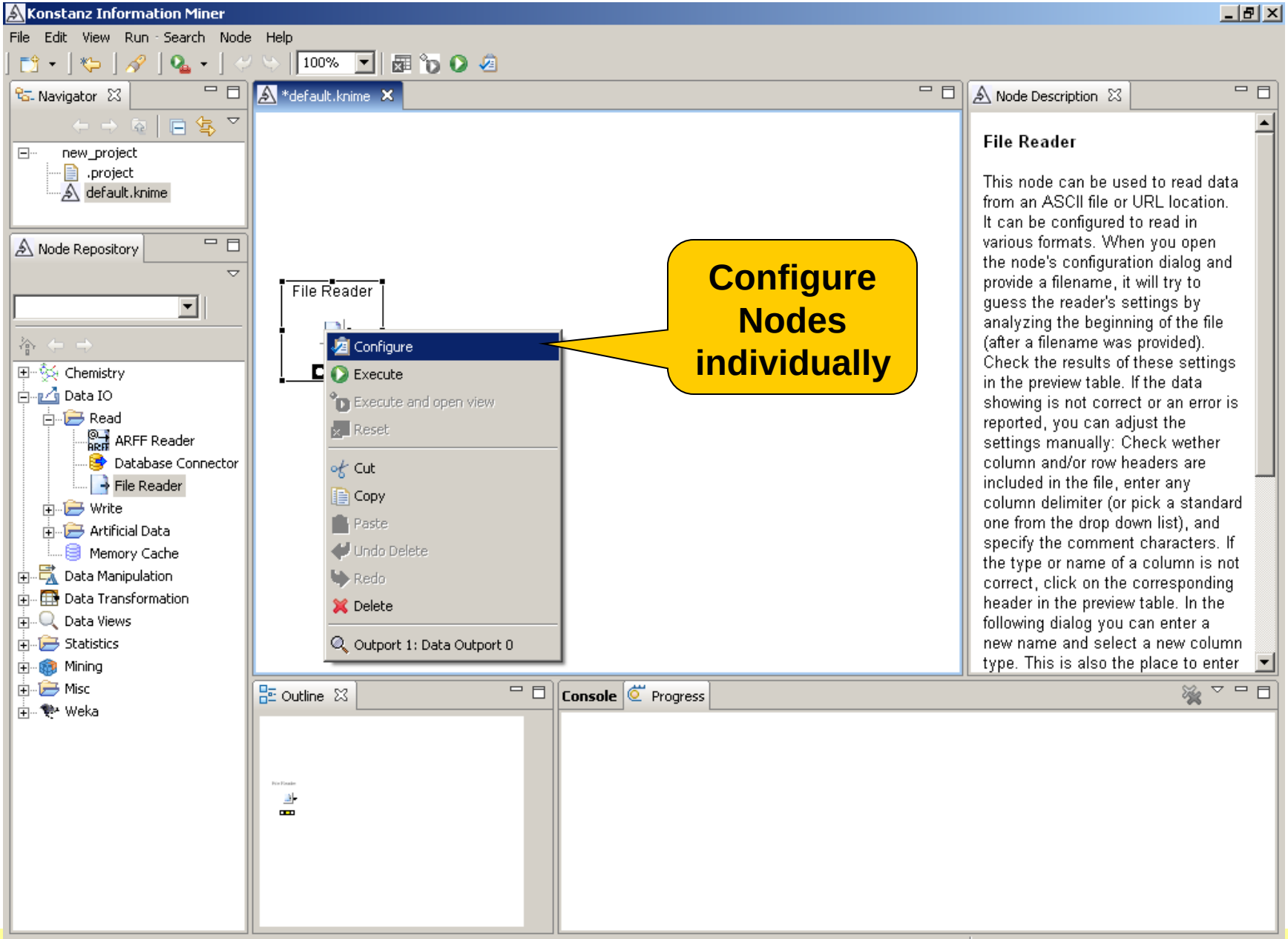
New nodes via open API and integrated wizard

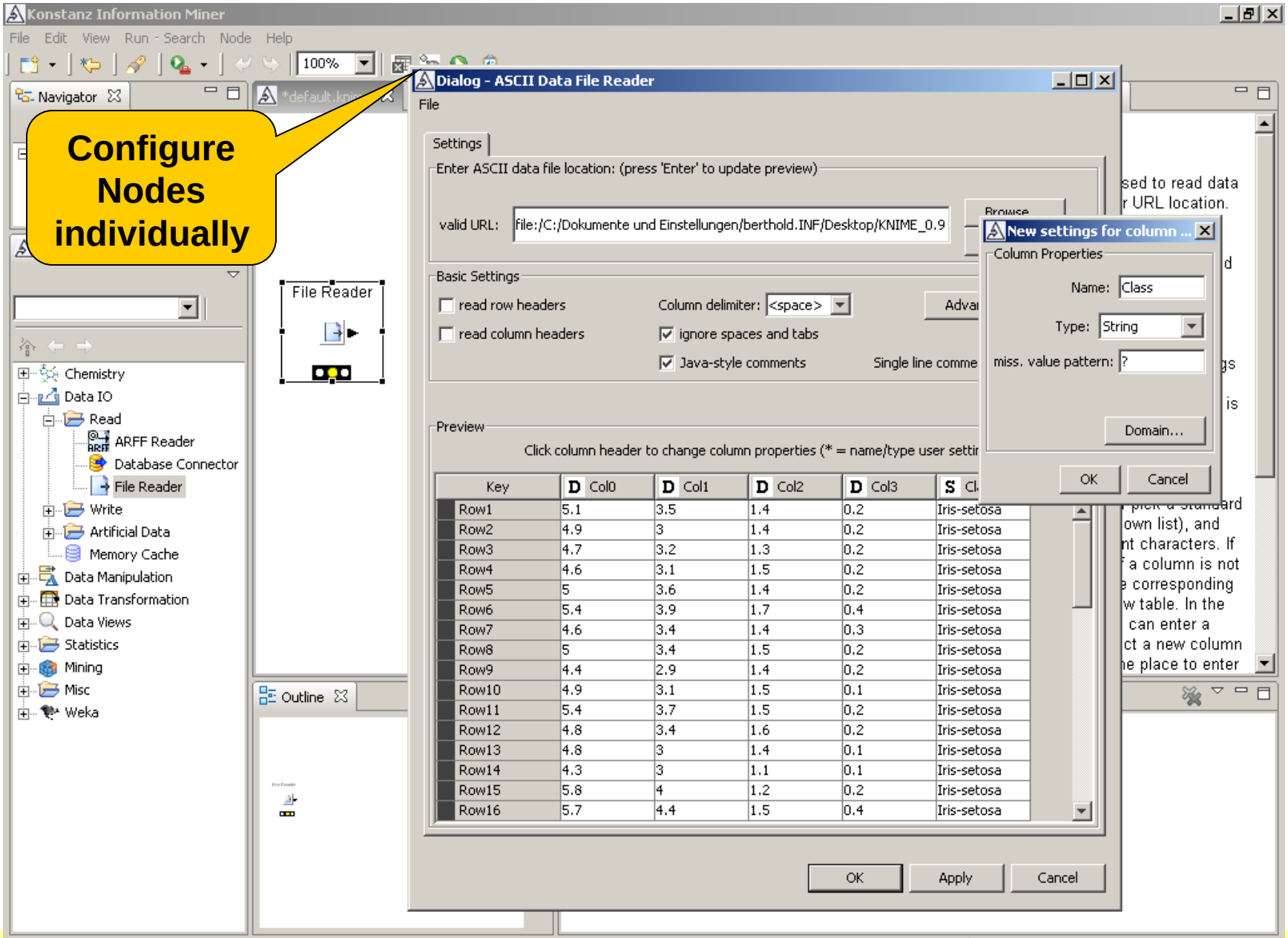
Data Pipeline



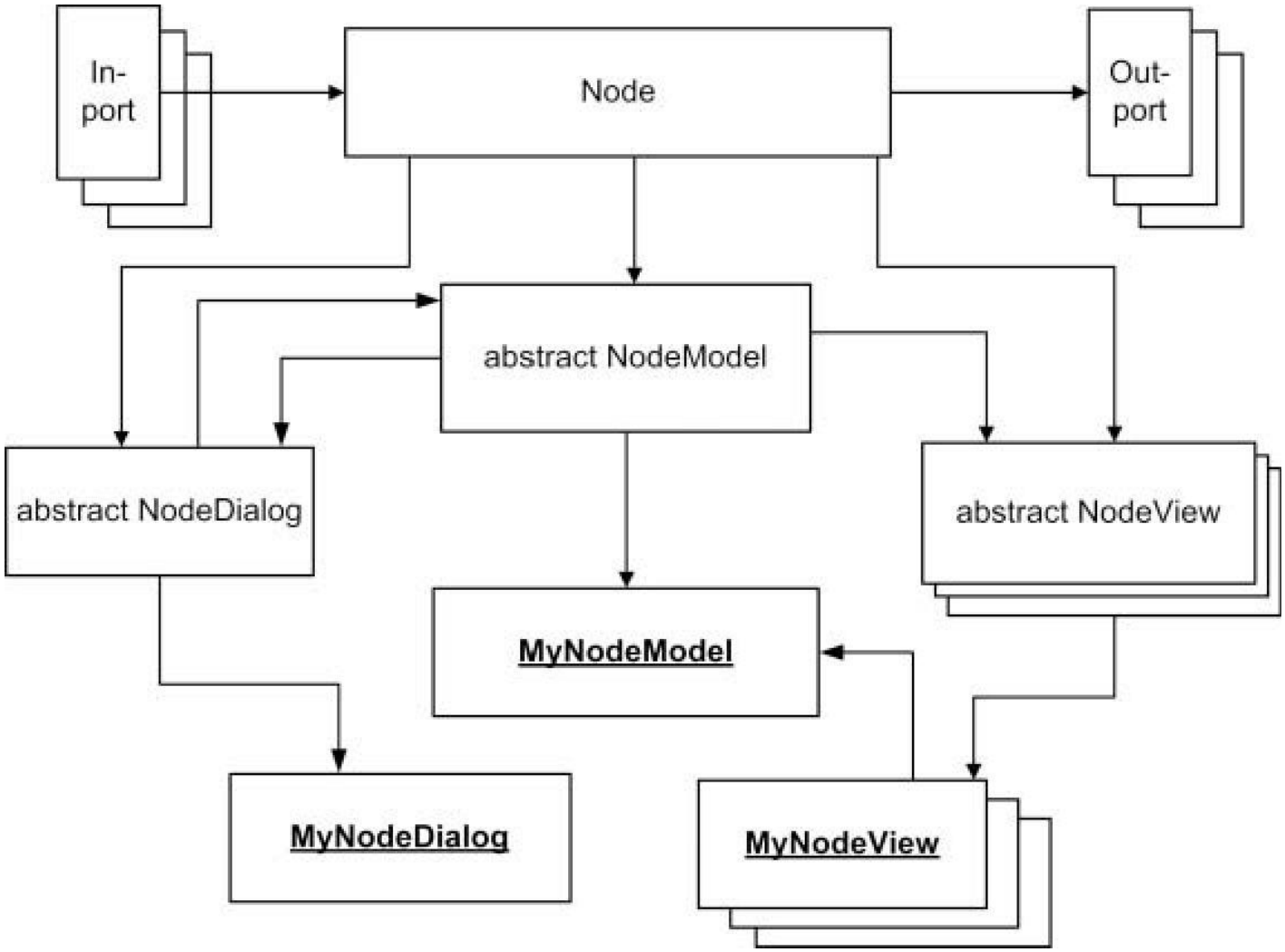

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Statistics
Mining
Misc
Weka
*default.knime
Node Description
Outline
Console
Progress

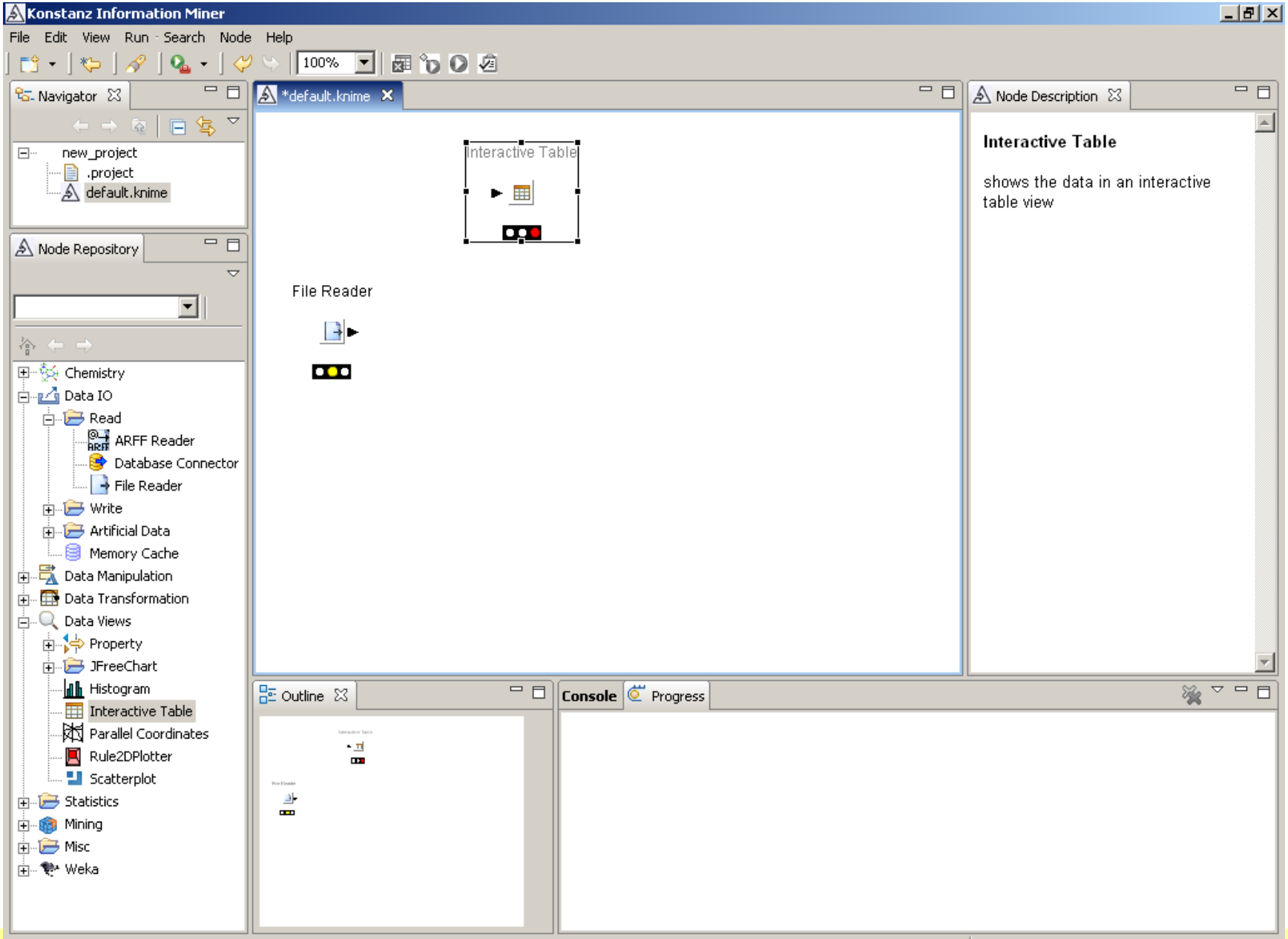

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Statistics
Mining
Misc
Weka
*default.knime
File Reader
Drag & Drop Nodes from Repository to Workbench
Node Description
File Reader
This node can be used to read data from an ASCII file or URL location. It can be configured to read in various formats. When you open the node's configuration dialog and provide a filename, it will try to guess the reader's settings by analyzing the beginning of the file (after a filename was provided). Check the results of these settings in the preview table. If the data showing is not correct or an error is reported, you can adjust the settings manually: Check wether column and/or row headers are included in the file, enter any column delimiter (or pick a standard one from the drop down list), and specify the comment characters. If the type or name of a column is not correct, click on the corresponding header in the preview table. In the following dialog you can enter a new name and select a new column type. This is also the place to enter
Outline
Console
Progress

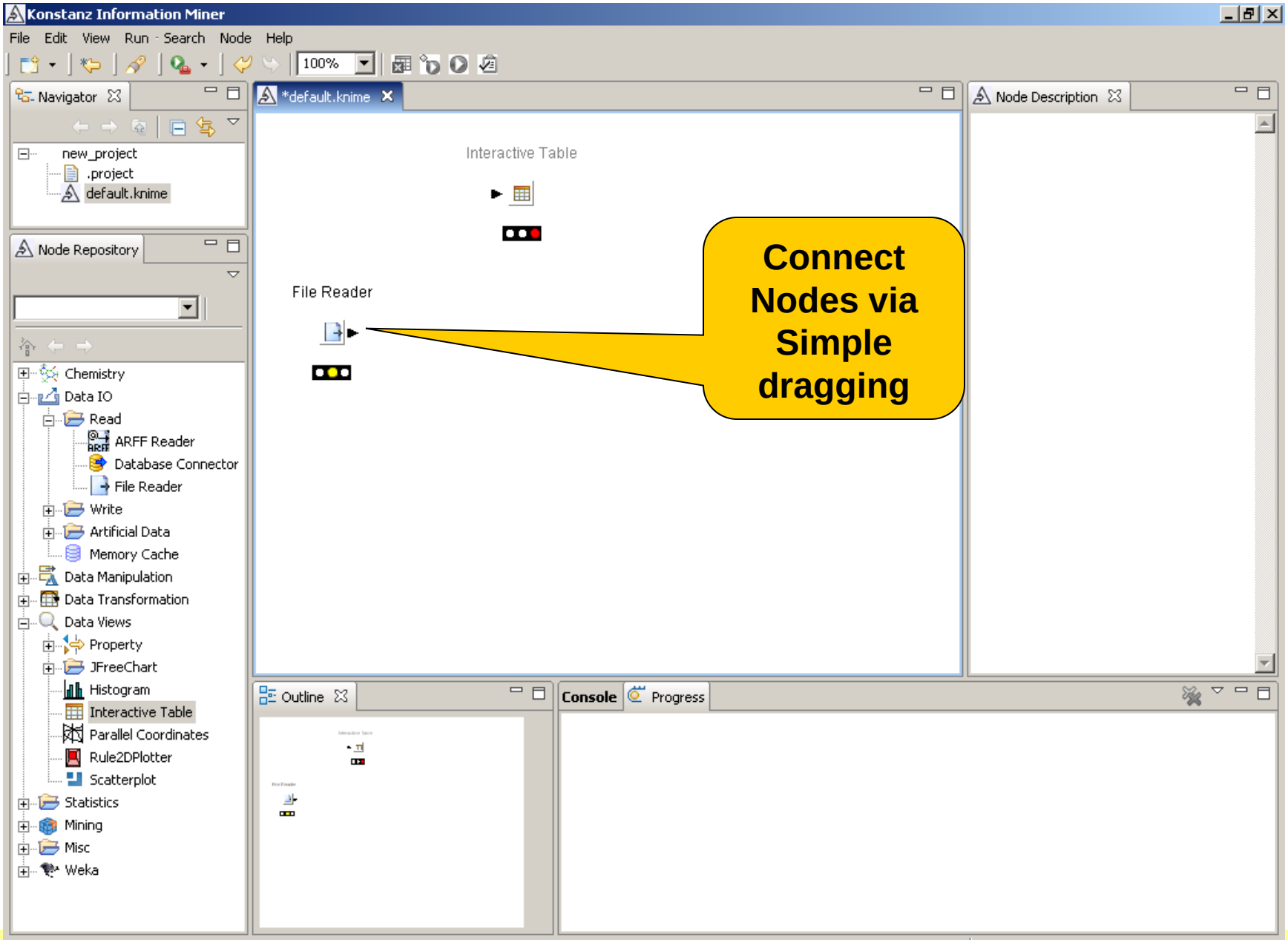

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Statistics
Mining
Misc
Weka
*default.knime
File Reader
Configure
Execute
Execute and open view
Reset
Cut
Copy
Paste
Undo Delete
Redo
Delete
Outport 1: Data Outport 0
Configure Nodes individually
Node Description
File Reader
This node can be used to read data from an ASCII file or URL location. It can be configured to read in various formats. When you open the node's configuration dialog and provide a filename, it will try to guess the reader's settings by analyzing the beginning of the file (after a filename was provided). Check the results of these settings in the preview table. If the data showing is not correct or an error is reported, you can adjust the settings manually: Check wether column and/or row headers are included in the file, enter any column delimiter (or pick a standard one from the drop down list), and specify the comment characters. If the type or name of a column is not correct, click on the corresponding header in the preview table. In the following dialog you can enter a new name and select a new column type. This is also the place to enter
Outline
Console
Progress

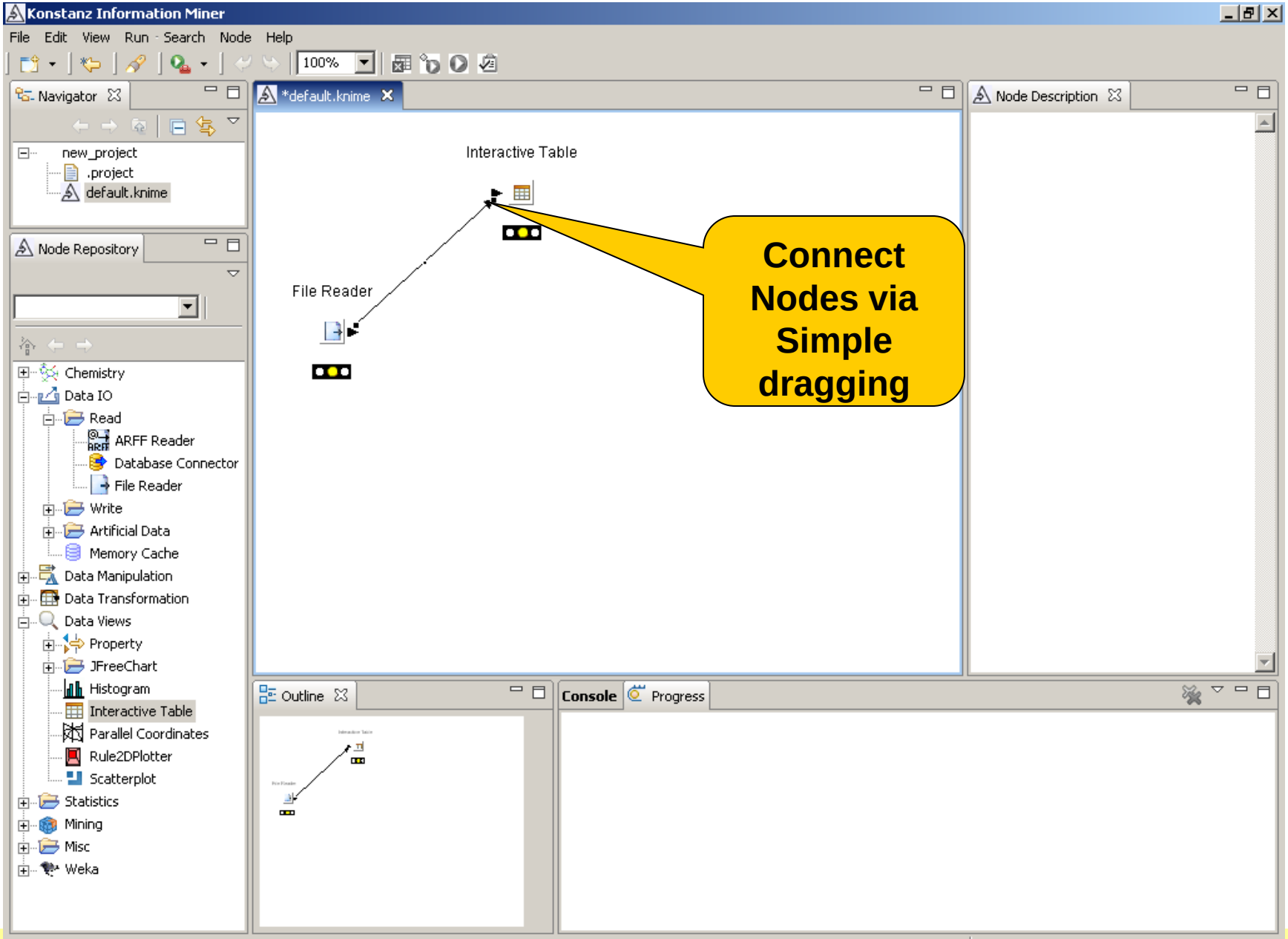

Configure Nodes individually
Konstanz Information Miner
File Edit View Run Search Node Help
Navigator
Dialog - ASCII Data File Reader
File
Settings
Enter ASCII data file location: (press 'Enter' to update preview)
valid URL: file:/C:/Dokumente und Einstellungen/berthold.INF/Desktop/KNIME_0.9
Browse
Basic Settings
read row headers
read column headers
Column delimiter: <space>
ignore spaces and tabs
Java-style comments
Single line comme
Preview
Click column header to change column properties (* = name/type user settir
Key | Col0 | Col1 | Col2 | Col3 | Cl
Row1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa
Row2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa
Row3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa
Row4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa
Row5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa
Row6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa
Row7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa
Row8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa
Row9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa
Row10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa
Row11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa
Row12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa
Row13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa
Row14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa
Row15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa
Row16 | 5.7 | 4.4 | 1.5 | 0.4 | Iris-setosa
OK
Apply
Cancel
New settings for column ...
Column Properties
Name: Class
Type: String
miss. value pattern: ?
Domain...
OK
Cancel
File Reader
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Statistics
Mining
Misc
Weka
Outline

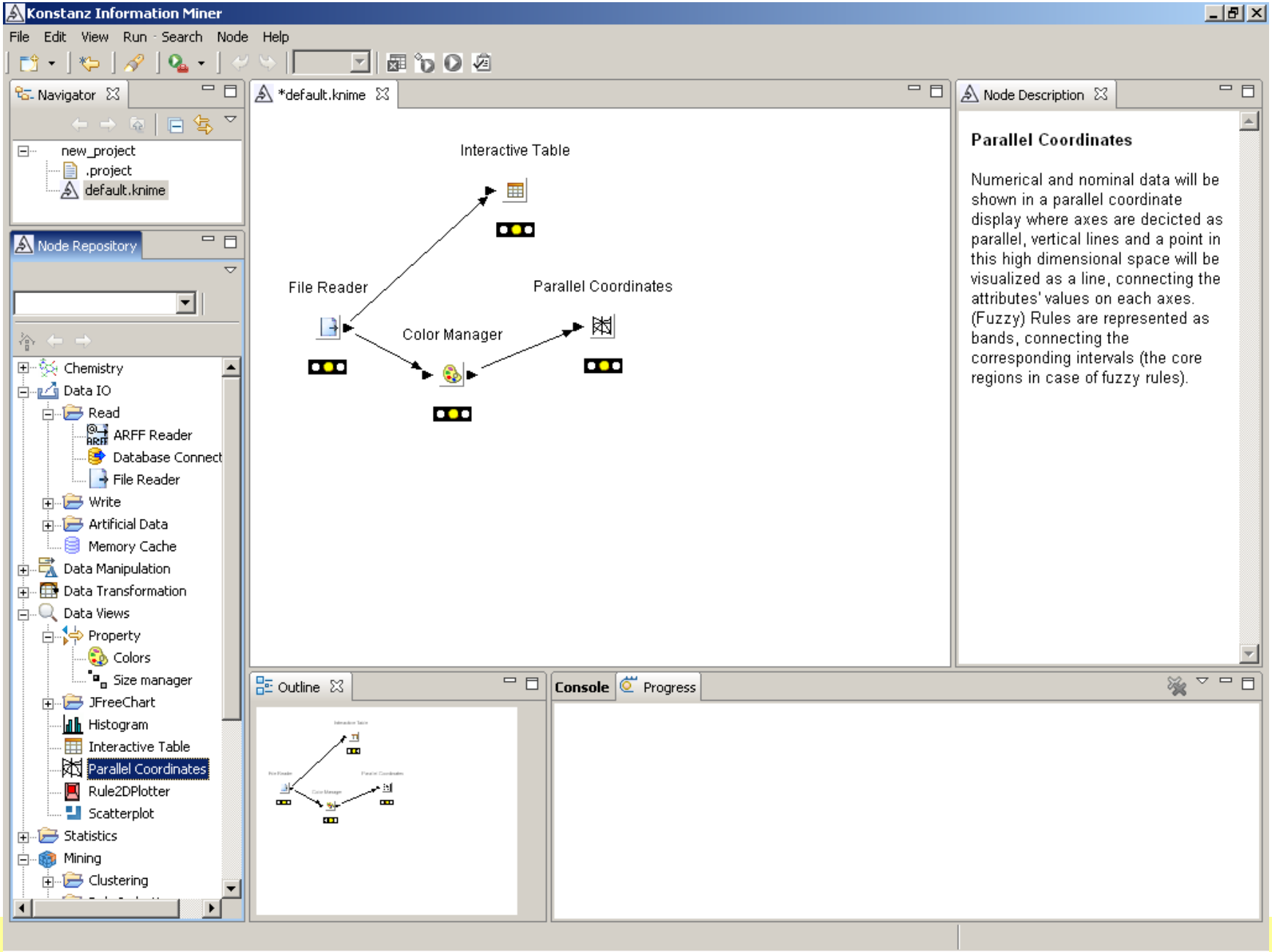
Node Model

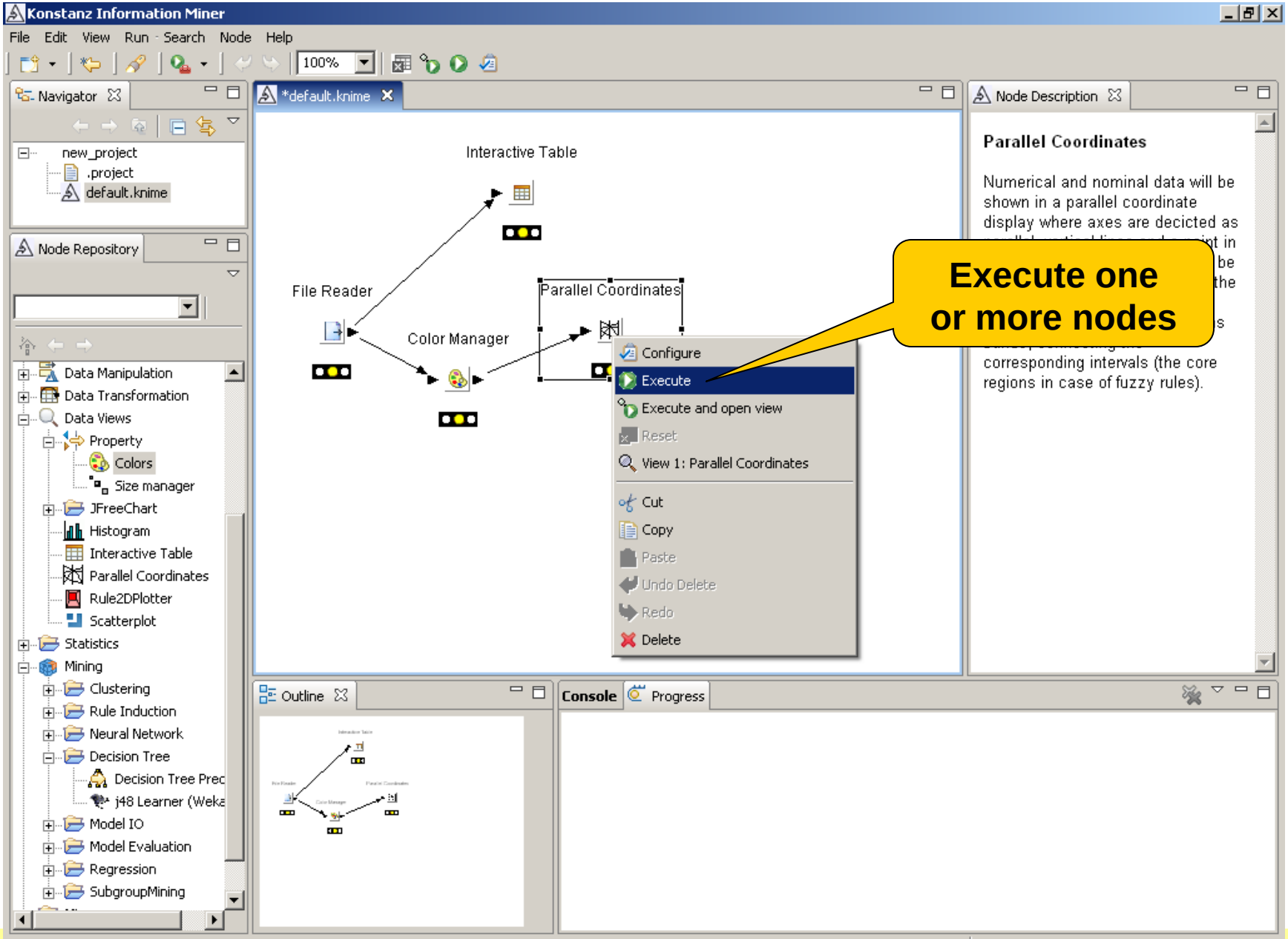

In-port
Node
Out-port
abstract NodeModel
abstract NodeDialog
abstract NodeView
MyNodeModel
MyNodeDialog
MyNodeView

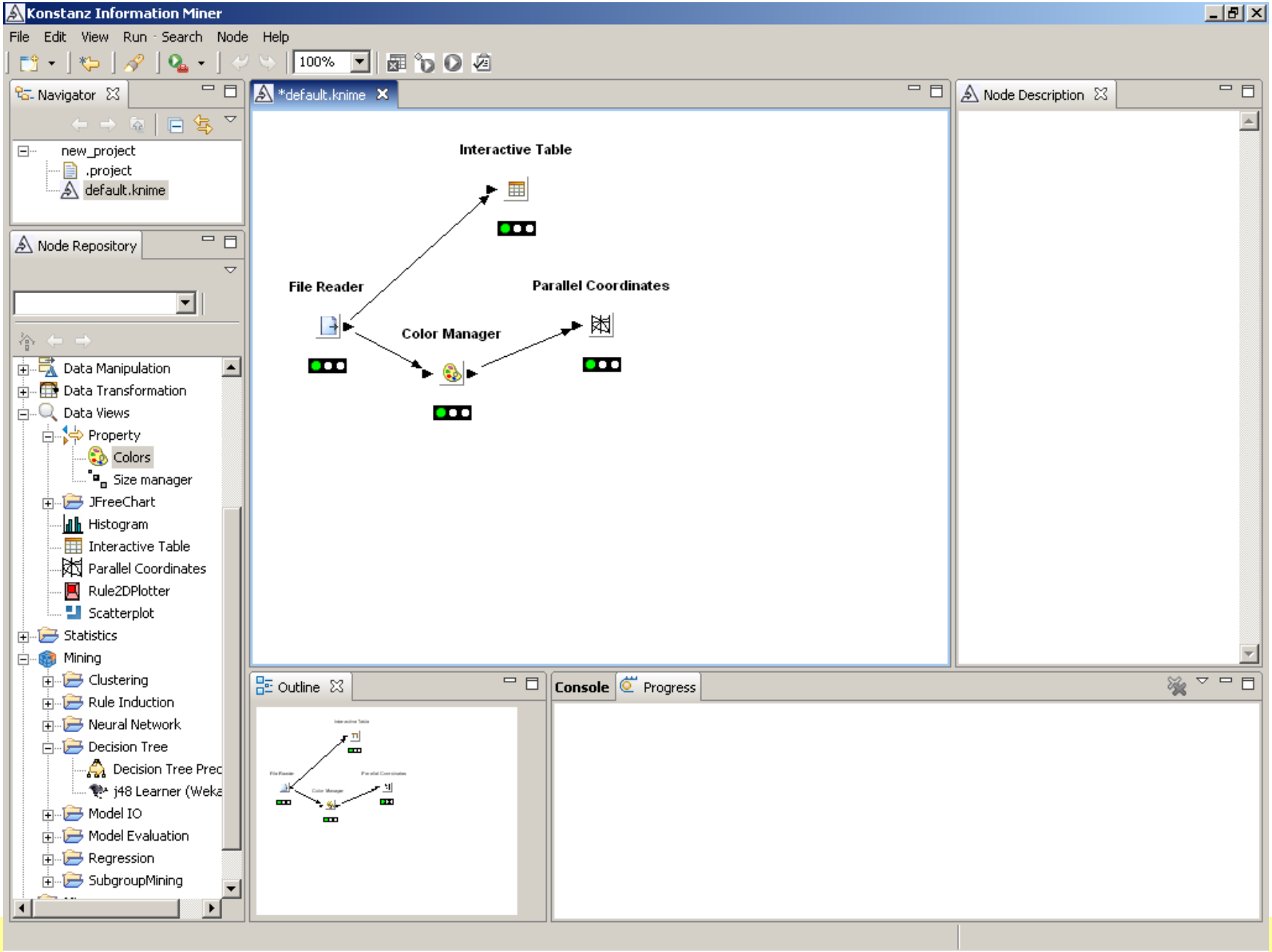

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Property
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Misc
Weka
*default.knime
Interactive Table
File Reader
Node Description
Interactive Table
shows the data in an interactive table view
Outline
Console
Progress

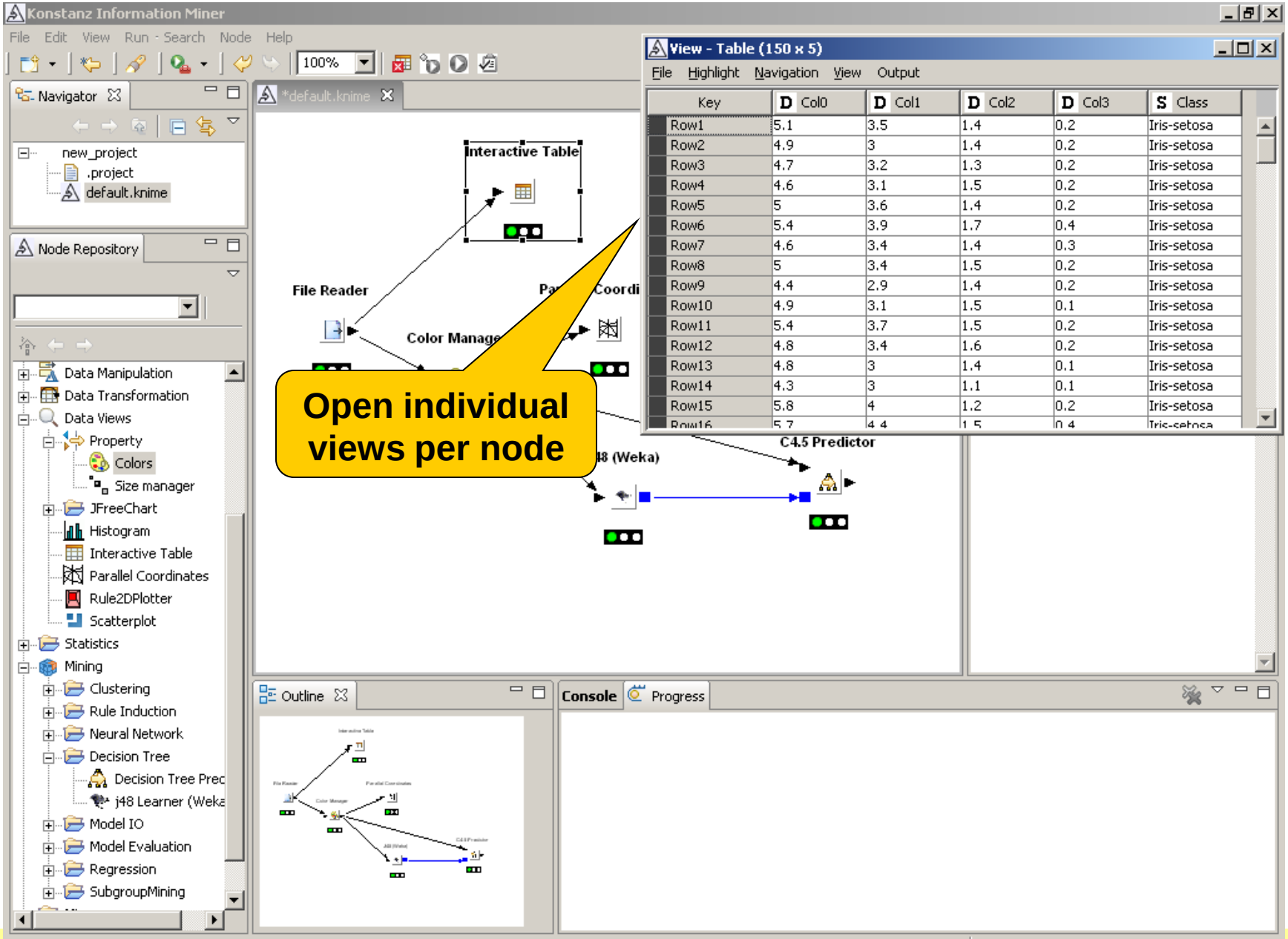

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Property
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Misc
Weka
*default.knime
Interactive Table
File Reader
Connect Nodes via Simple dragging
Node Description
Outline
Console
Progress

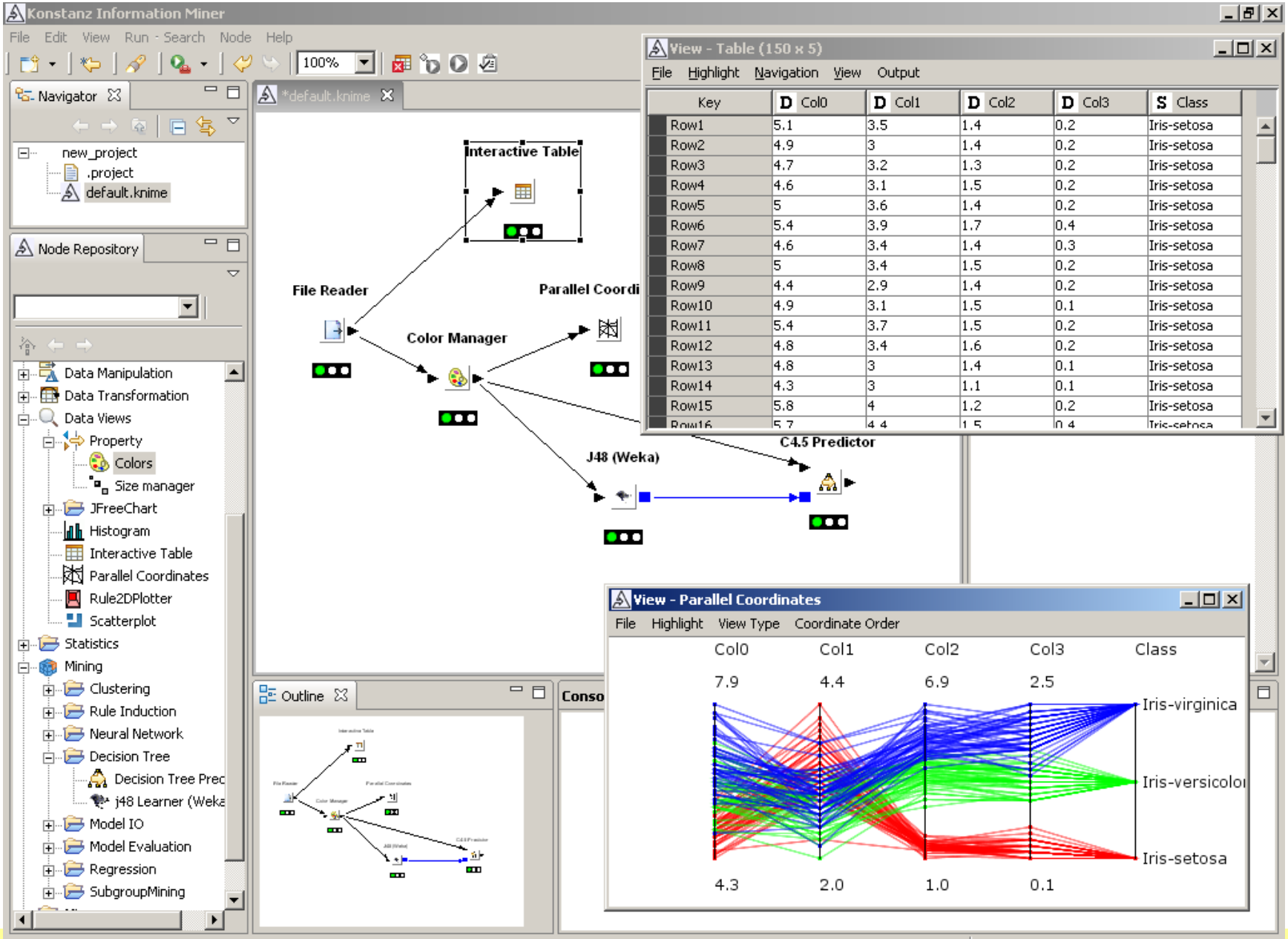

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connector
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Property
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Misc
Weka
*default.knime
Interactive Table
File Reader
Connect Nodes via Simple dragging
Node Description
Outline
Console
Progress

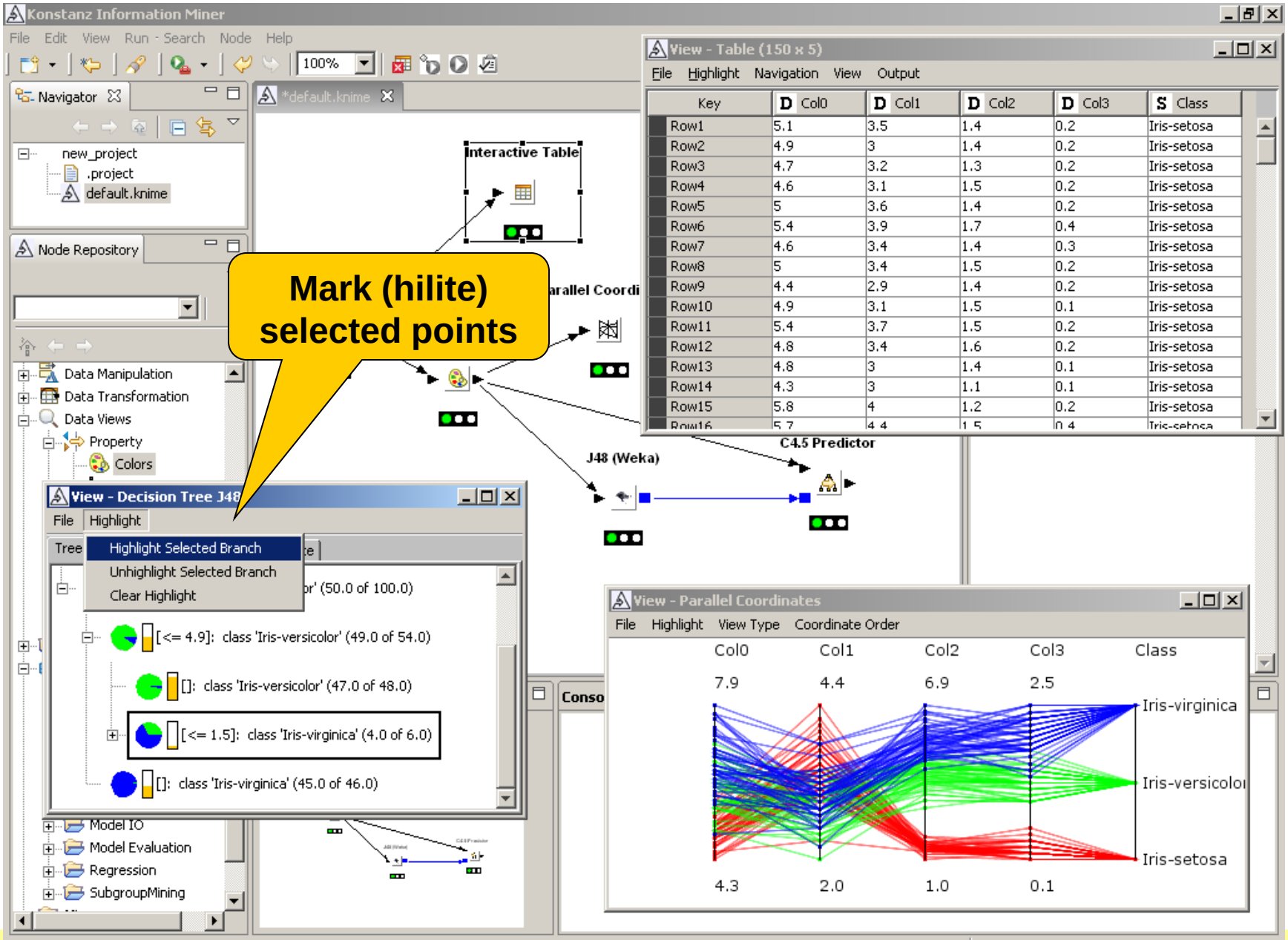

Konstanz Information Miner
File Edit View Run Search Node Help
Navigator
new_project
.project
default.knime
Node Repository
Chemistry
Data IO
Read
ARFF Reader
Database Connect
File Reader
Write
Artificial Data
Memory Cache
Data Manipulation
Data Transformation
Data Views
Property
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Clustering
*default.knime
Interactive Table
File Reader
Parallel Coordinates
Color Manager
Node Description
Parallel Coordinates
Numerical and nominal data will be shown in a parallel coordinate display where axes are decicted as parallel, vertical lines and a point in this high dimensional space will be visualized as a line, connecting the attributes' values on each axes. (Fuzzy) Rules are represented as bands, connecting the corresponding intervals (the core regions in case of fuzzy rules).
Outline
Console
Progress

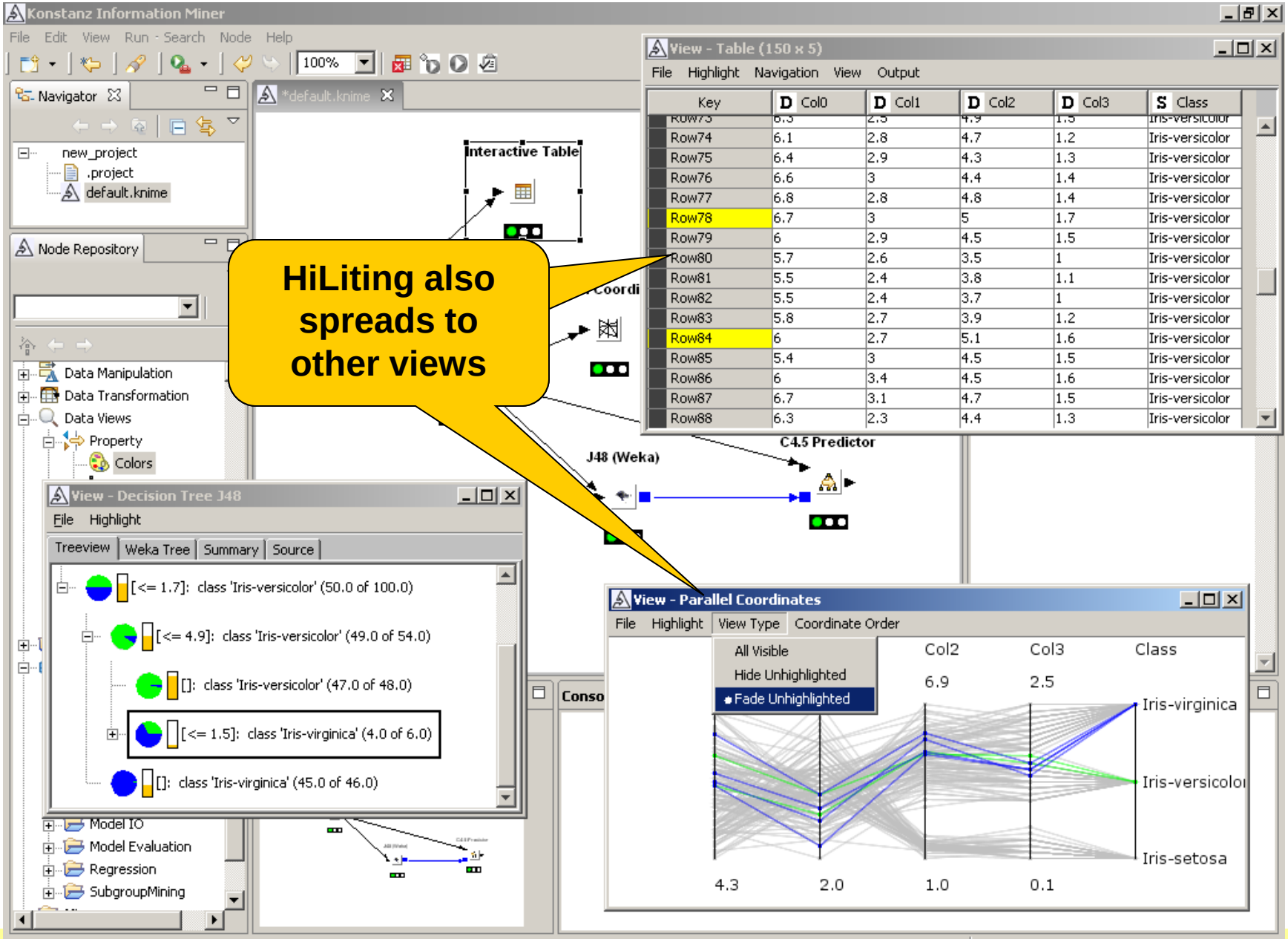

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Clustering
Rule Induction
Neural Network
Decision Tree
Decision Tree Prec
j48 Learner (Weka
Model IO
Model Evaluation
Regression
SubgroupMining
*default.knime
Interactive Table
File Reader
Color Manager
Parallel Coordinates
Configure
Execute
Execute and open view
Reset
View 1: Parallel Coordinates
Cut
Copy
Paste
Undo Delete
Redo
Delete
Execute one or more nodes
Node Description
Parallel Coordinates
Numerical and nominal data will be shown in a parallel coordinate display where axes are decicted as
corresponding intervals (the core regions in case of fuzzy rules).
Outline
Console
Progress

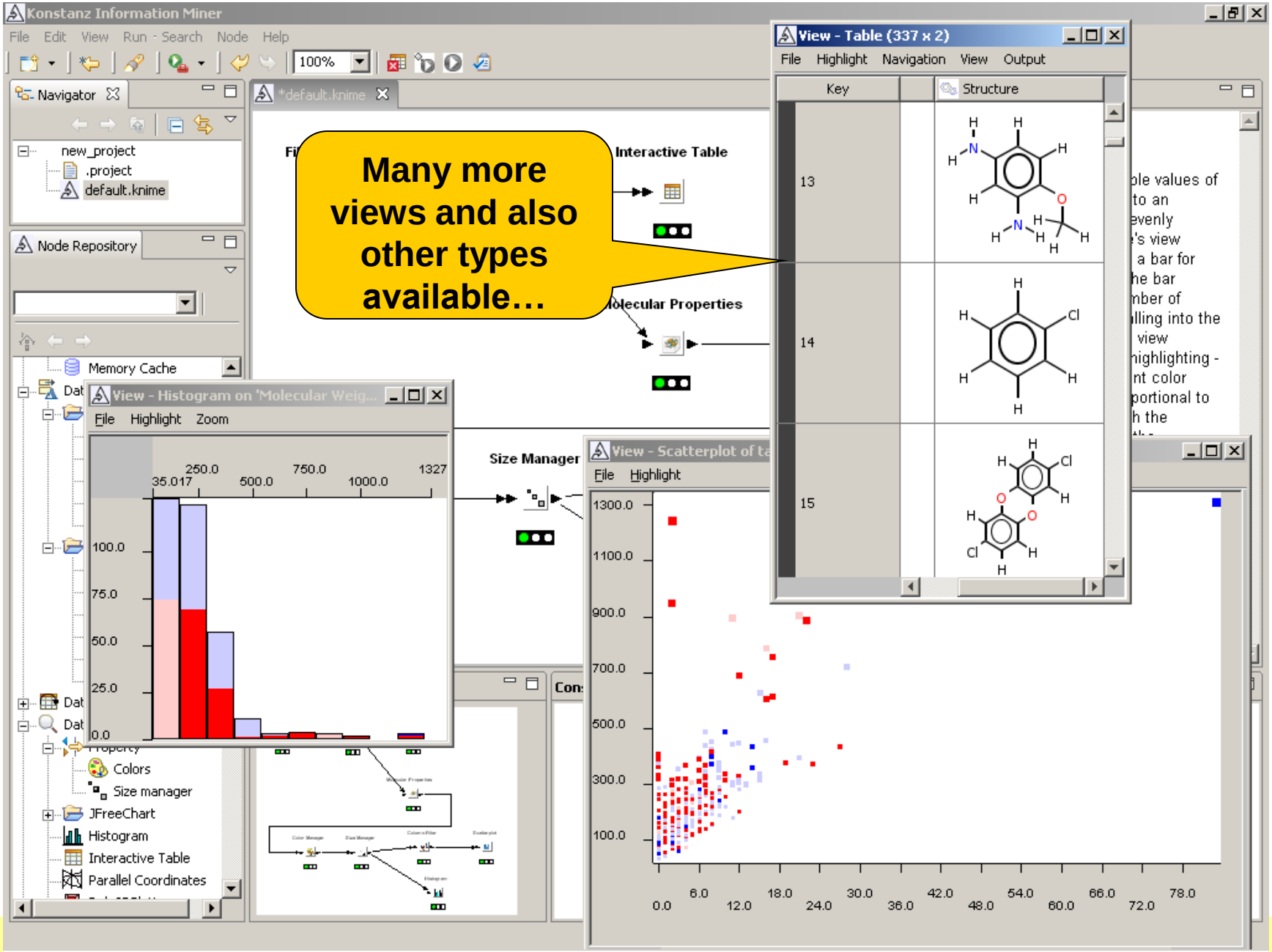

Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Clustering
Rule Induction
Neural Network
Decision Tree
Decision Tree Prec
j48 Learner (Weka
Model IO
Model Evaluation
Regression
SubgroupMining
*default.knime
Interactive Table
File Reader
Parallel Coordinates
Color Manager
Node Description
Outline
Console
Progress


Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Clustering
Rule Induction
Neural Network
Decision Tree
Decision Tree Prec
j48 Learner (Weka
Model IO
Model Evaluation
Regression
SubgroupMining
*default.knime
Interactive Table
File Reader
Color Manage
C4.5 Predictor
Open individual views per node
View - Table (150 x 5)
File Highlight Navigation View Output
Key | Col0 | Col1 | Col2 | Col3 | Class
Row1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa
Row2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa
Row3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa
Row4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa
Row5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa
Row6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa
Row7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa
Row8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa
Row9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa
Row10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa
Row11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa
Row12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa
Row13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa
Row14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa
Row15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa
Outline
Console
Progress


Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
Rule2DPlotter
Scatterplot
Statistics
Mining
Clustering
Rule Induction
Neural Network
Decision Tree
Decision Tree Prec
j48 Learner (Weka
Model IO
Model Evaluation
Regression
SubgroupMining
*default.knime
Interactive Table
File Reader
Parallel Coordi
Color Manager
J48 (Weka)
C4.5 Predictor
Outline
View - Table (150 x 5)
File Highlight Navigation View Output
Key | Col0 | Col1 | Col2 | Col3 | Class
Row1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa
Row2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa
Row3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa
Row4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa
Row5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa
Row6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa
Row7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa
Row8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa
Row9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa
Row10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa
Row11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa
Row12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa
Row13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa
Row14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa
Row15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa
View - Parallel Coordinates
File Highlight View Type Coordinate Order
Col0 Col1 Col2 Col3 Class
7.9 4.4 6.9 2.5
Iris-virginica
Iris-versicolor
Iris-setosa
4.3 2.0 1.0 0.1


Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Model IO
Model Evaluation
Regression
SubgroupMining
*default.knime
Interactive Table
Parallel Coordi
J48 (Weka)
C4.5 Predictor
Mark (hilite) selected points
View - Table (150 x 5)
File Highlight Navigation View Output
Key | Col0 | Col1 | Col2 | Col3 | Class
Row1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa
Row2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa
Row3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa
Row4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa
Row5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa
Row6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa
Row7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa
Row8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa
Row9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa
Row10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa
Row11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa
Row12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa
Row13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa
Row14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa
Row15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa
View - Decision Tree J48
File Highlight
Highlight Selected Branch
Unhighlight Selected Branch
Clear Highlight
[<= 4.9]: class 'Iris-versicolor' (49.0 of 54.0)
[]: class 'Iris-versicolor' (47.0 of 48.0)
[<= 1.5]: class 'Iris-virginica' (4.0 of 6.0)
[]: class 'Iris-virginica' (45.0 of 46.0)
View - Parallel Coordinates
File Highlight View Type Coordinate Order
Col0 Col1 Col2 Col3 Class
7.9 4.4 6.9 2.5
Iris-virginica
Iris-versicolor
Iris-setosa
4.3 2.0 1.0 0.1


Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
*default.knime
Interactive Table
Node Repository
Data Manipulation
Data Transformation
Data Views
Property
Colors
Model IO
Model Evaluation
Regression
SubgroupMining
J48 (Weka)
C4.5 Predictor
HiLiting also spreads to other views
View - Table (150 x 5)
File Highlight Navigation View Output
Key | Col0 | Col1 | Col2 | Col3 | Class
Row74 | 6.1 | 2.8 | 4.7 | 1.2 | Iris-versicolor
Row75 | 6.4 | 2.9 | 4.3 | 1.3 | Iris-versicolor
Row76 | 6.6 | 3 | 4.4 | 1.4 | Iris-versicolor
Row77 | 6.8 | 2.8 | 4.8 | 1.4 | Iris-versicolor
Row78 | 6.7 | 3 | 5 | 1.7 | Iris-versicolor
Row79 | 6 | 2.9 | 4.5 | 1.5 | Iris-versicolor
Row80 | 5.7 | 2.6 | 3.5 | 1 | Iris-versicolor
Row81 | 5.5 | 2.4 | 3.8 | 1.1 | Iris-versicolor
Row82 | 5.5 | 2.4 | 3.7 | 1 | Iris-versicolor
Row83 | 5.8 | 2.7 | 3.9 | 1.2 | Iris-versicolor
Row84 | 6 | 2.7 | 5.1 | 1.6 | Iris-versicolor
Row85 | 5.4 | 3 | 4.5 | 1.5 | Iris-versicolor
Row86 | 6 | 3.4 | 4.5 | 1.6 | Iris-versicolor
Row87 | 6.7 | 3.1 | 4.7 | 1.5 | Iris-versicolor
Row88 | 6.3 | 2.3 | 4.4 | 1.3 | Iris-versicolor
View - Decision Tree J48
File Highlight
Treeview Weka Tree Summary Source
[<= 1.7]: class 'Iris-versicolor' (50.0 of 100.0)
[<= 4.9]: class 'Iris-versicolor' (49.0 of 54.0)
[]: class 'Iris-versicolor' (47.0 of 48.0)
[<= 1.5]: class 'Iris-virginica' (4.0 of 6.0)
[]: class 'Iris-virginica' (45.0 of 46.0)
View - Parallel Coordinates
File Highlight View Type Coordinate Order
All Visible
Hide Unhighlighted
Fade Unhighlighted
Col2
Col3
Class
6.9
2.5
Iris-virginica
Iris-versicolor
Iris-setosa
4.3
2.0
1.0
0.1


Konstanz Information Miner
File Edit View Run Search Node Help
Navigator
new_project
.project
default.knime
Node Repository
Memory Cache
Colors
Size manager
JFreeChart
Histogram
Interactive Table
Parallel Coordinates
*default.knime
Interactive Table
Molecular Properties
Size Manager
Many more views and also other types available…
View - Table (337 x 2)
File Highlight Navigation View Output
Key
Structure
13
14
15
View - Histogram on 'Molecular Weig...
File Highlight Zoom
35.017
250.0
500.0
750.0
1000.0
1327
100.0
75.0
50.0
25.0
0.0
View - Scatterplot of ta
File Highlight
1300.0
1100.0
900.0
700.0
500.0
300.0
100.0
0.0
6.0
12.0
18.0
24.0
30.0
36.0
42.0
48.0
54.0
60.0
66.0
72.0
78.0

Data Table

- Contains meta information (spec)
  - data types
  - domains
  - # of rows/cols
- Large tables are buffered on disc
- Blob cell support for large data cells e.g. images

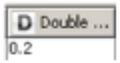
BR vector data - 4:8 - Bitvector Generator

File

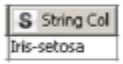
Table "default" - Rows: 150 | Spec - Columns: 5 | Properties |

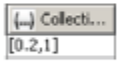
Row ID	Double ...	String Col	Integer ...	Collecti...	BitVectors
Row0	0.2	Iris-setosa	1	[0,2,1]	10
Row1	0.2	Iris-setosa	1	[0,2,1]	10
Row2	0.2	Iris-setosa	1	[0,2,1]	10
Row3	0.2	Iris-setosa	1	[0,2,1]	10
Row4	0.2	Iris-setosa	1	[0,2,1]	10
Row5	0.4	Iris-setosa	1	[0,4,1]	10
Row6	0.3	Iris-setosa	1	[0,3,1]	10
Row7	0.2	Iris-setosa	1	[0,2,1]	10
Row8	0.2	Iris-setosa	1	[0,2,1]	10
Row9	0.1	Iris-setosa	1	[0,1,1]	10
Row10	0.2	Iris-setosa	1	[0,2,1]	10
Row11	0.2	Iris-setosa	1	[0,2,1]	10
Row12	0.1	Iris-setosa	1	[0,1,1]	10
Row13	0.1	Iris-setosa	1	[0,1,1]	10
Row14	0.2	Iris-setosa	1	[0,2,1]	10
Row15	0.4	Iris-setosa	1	[0,4,1]	10
Row16	0.4	Iris-setosa	1	[0,4,1]	10
Row17	0.3	Iris-setosa	1	[0,3,1]	10
Row18	0.3	Iris-setosa	1	[0,3,1]	10
Row19	0.3	Iris-setosa	1	[0,3,1]	10
Row20	0.2	Iris-setosa	1	[0,2,1]	10

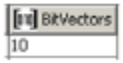
Data Types

- Common data types
  - Double Value 

  - Int Value 

  - String Value 

  - Collections 

    - Sets
    - Lists
  - Bit vectors 

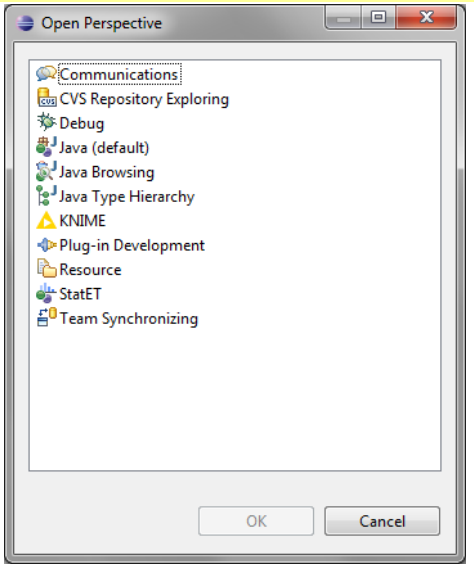
- Additional data types
  - Terms and Documents
  - Image
  - Network
  - Chemical types
    - Molecules i.e. CDK, Smiles, SDF, …
  - Distance Matrix
  - Custom data types

KNIME Features

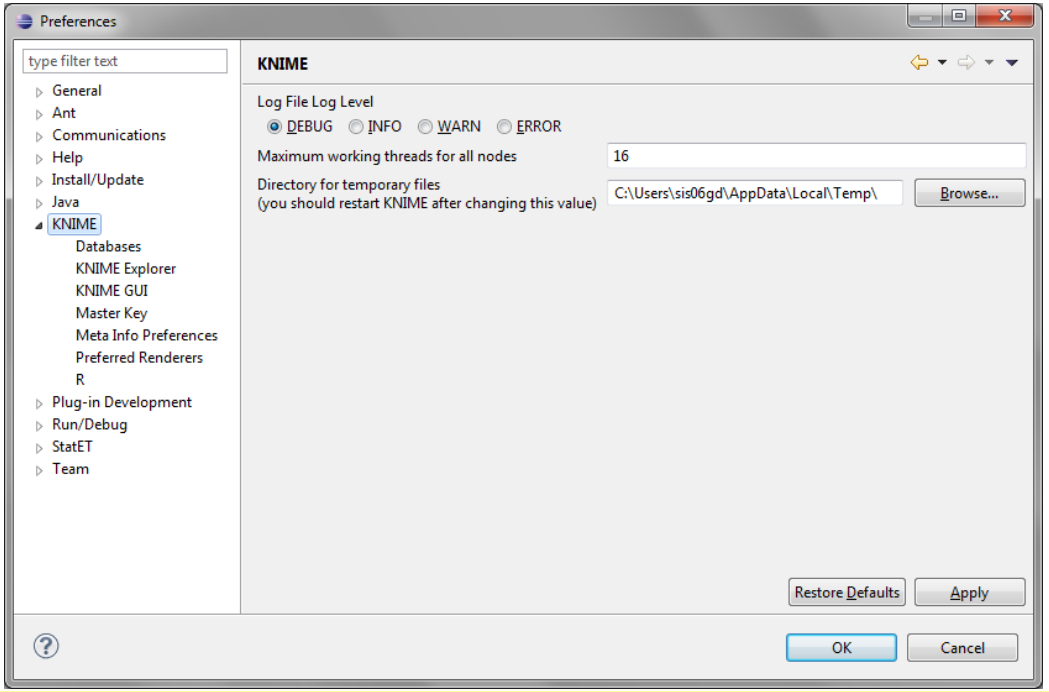
- Node types
  - I/O
  - Data manipulation
  - Learners
  - Predictors
  - Views
- Highlighting
- Metanodes
- Quickforms
- Loops and flow variables
- Error handling: “try-catch" nodes
- Extensions (KNIME plugins)

Perspective and Preferences

- In the SDK version (Eclipse), you need to select the KNIME perspective.



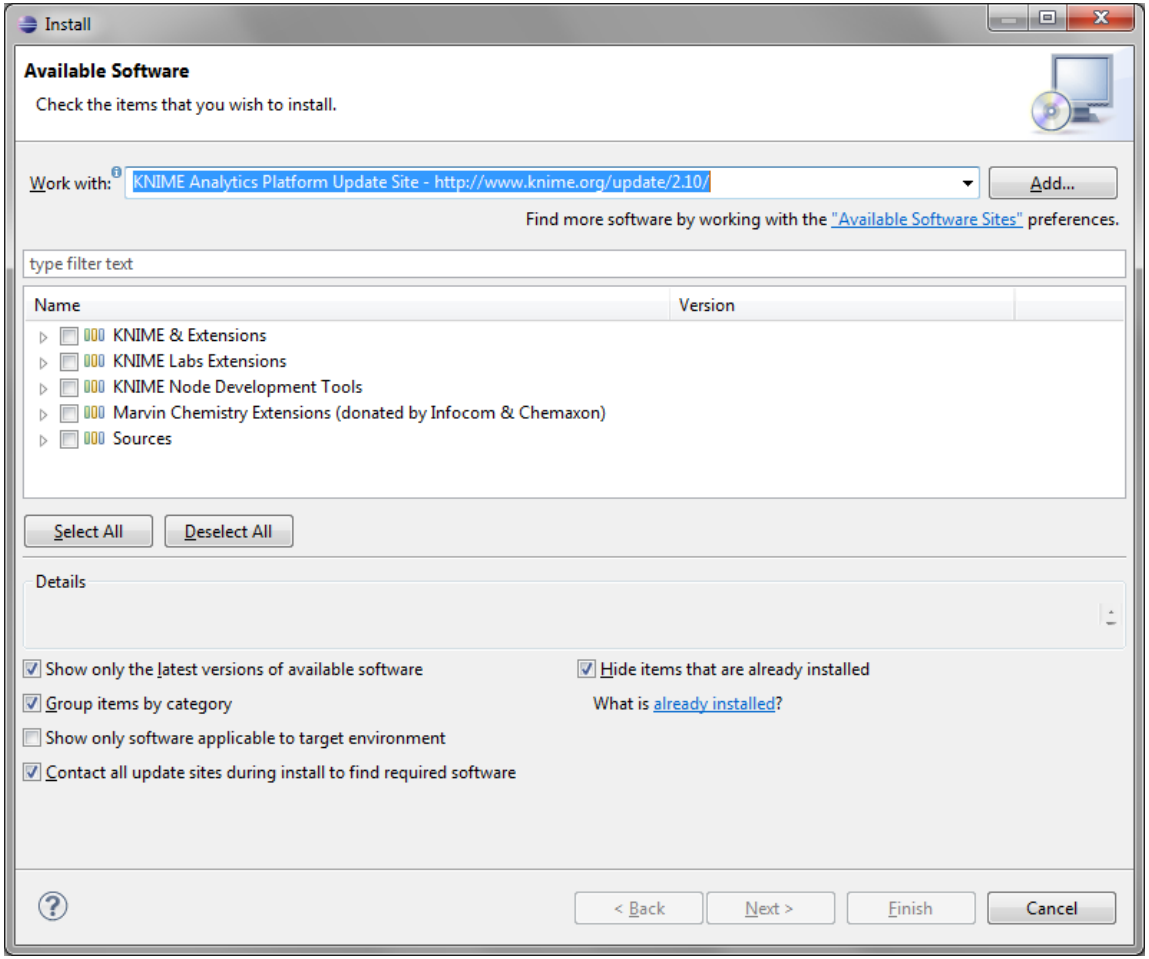
- In Menu Windows, select Preferences



KNIME Updates and Extensions

KNIME can be updated and extended by means of the Eclipse "Install New Software" mechanism.

- select the KNIME update site



KNIME Extensions

- Experimental Extensions: http://tech.knime.org/knime-labs
  - Modular Data Generators
  - Network Mining
  - Perl Scripting
  - Text Processing
  - etc.

- Community Contributions: http://tech.knime.org/community
  - Chemoinformatics
  - High Content Screening
  - Image Processing
  - Next Generation Sequencing
  - R/Groovy/Matlab/Python Scripting
  - STARK
  - etc.

- KNIME is designed to be extended!
  - You just need to use the SDK version.

KNIME Extensions (Plugins)

Some available plugins include:

- Chemistry types and features
- Distance Matrix
- Ensemble Learning
- HTML/PDF Writer
- Item Set Mining
- R Statistics Integration
- Report Designer
- Webservice Client
- Weka Data Mining Integration
- XLS Support
- XML Processing

Conclusions on KNIME

- Modularity and extendibility
  - General and extendible data structure (DataTable and DataCell)
  - Nodes encapsulate computational processing tasks (algorithms)

- A workflow management system
  - directed edges connects nodes to create data pipelines
  - a workflow is, in general, a directed acyclic graph
  - multi-threading
  - Meta-nodes (nested workflows)

- New releases
  - Enhanced GUI and performance
  - Include more and more modules and features

KNIME Useful Resources

KNIME User → desktop version: “KNIME Analytics Platform”

- http://www.knime.org/knime
- http://www.knime.org/downloads/datasets
- http://www.knime.org/introduction/examples

Workflow examples:

- parallel coordinates on Iris data
- PCA on Iris data
- decision tree on Iris data

KNIME Developer → SDK version (Eclipse):

- http://tech.knime.org/developer-guide
- http://tech.knime.org/developer/example
- API: for example see the DataTable interface in
  http://tech.knime.org/docs/api/org/knime/core/data/package-summary.html
- Simple exercise: create a new KNIME node to compute column stats

Rosaria’s blog with lots of examples:

- http://www.dataminingreporting.com/