CSMDM21 - Data Analytics and Mining

Introduction to the Data Science Platform KNIME

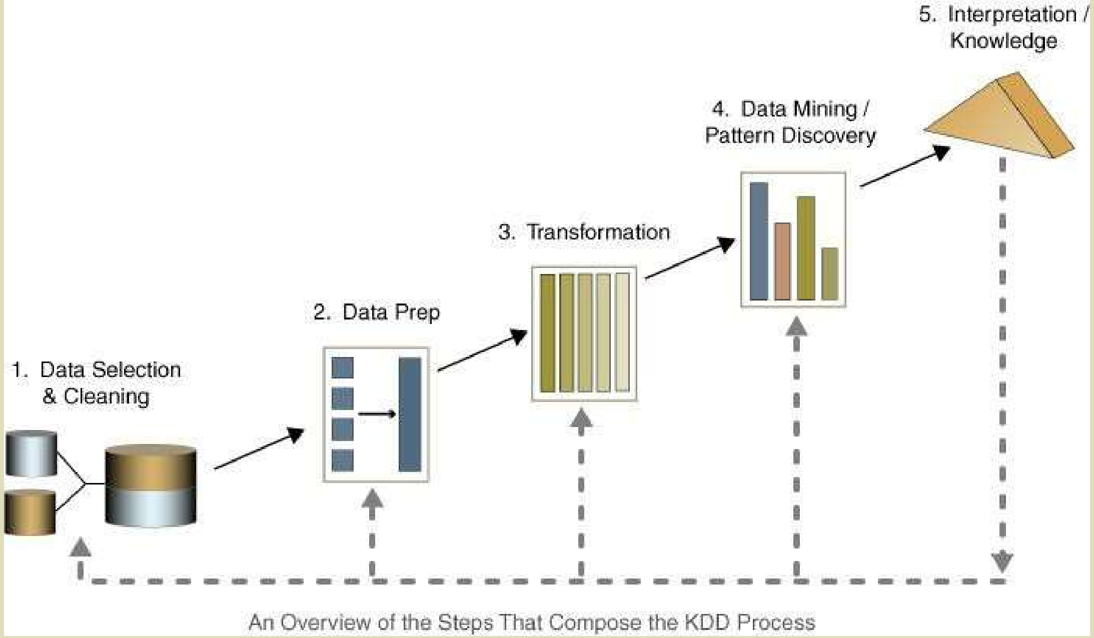
Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos created by
Prof. Giuseppe Di Fatta

Development Environments for Data Science

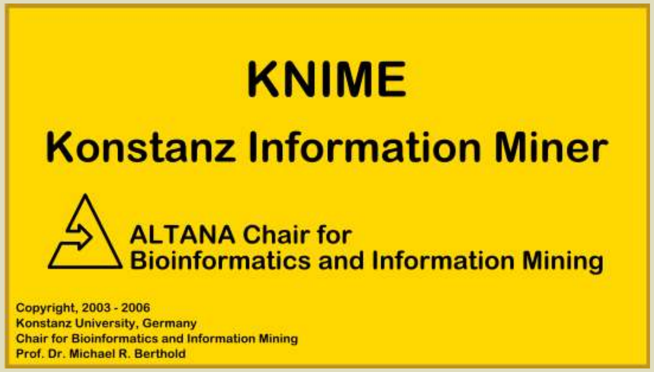
- ❑ Data Science
- ❑ Knowledge Discovery Process (KDD)
- ❑ Data Mining and Machine Learning algorithms

- An open and integrated environment to facilitate the KDD process.
- A **workflow management system for data analytics and mining** that integrates methods for data management and information visualization with algorithms for prediction, pattern discovery, classification.


5. Interpretation / Knowledge
4. Data Mining / Pattern Discovery
3. Transformation
2. Data Prep
1. Data Selection & Cleaning

An Overview of the Steps That Compose the KDD Process

Introducing KNIME



KNIME

Konstanz Information Miner

ALTANA Chair for Bioinformatics and Information Mining

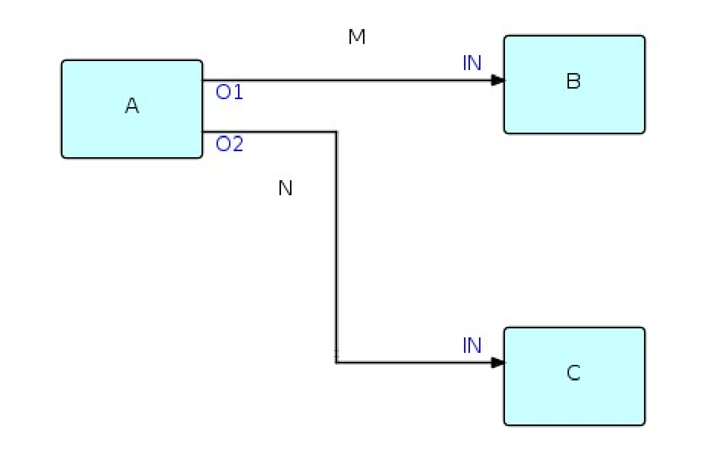
Copyright, 2003 - 2006
Konstanz University, Germany
Chair for Bioinformatics and Information Mining
Prof. Dr. Michael R. Berthold

- Developed at the **ALTANA-Chair for Bioinformatics and Information Mining**, Department of Computer and Information Science, University of Konstanz (Germany)
- 1st release: ~140 MM effort in two years
- Under continuous evolution and extension
  - April 2006: 1st release, version 1.0.0
  - Current release: version 4.2.1 (July 2020)
    - "KNIME Analytics Platform" – An open source software for creating data science applications
      - extensions and integrations
      - KNIME SDK
    - KNIME server (proprietary, enterprise software for teams)
  - Available for Windows, Linux and Mac OS X
  - An open source license (GPL) allows KNIME to be downloaded, distributed, and used freely.

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

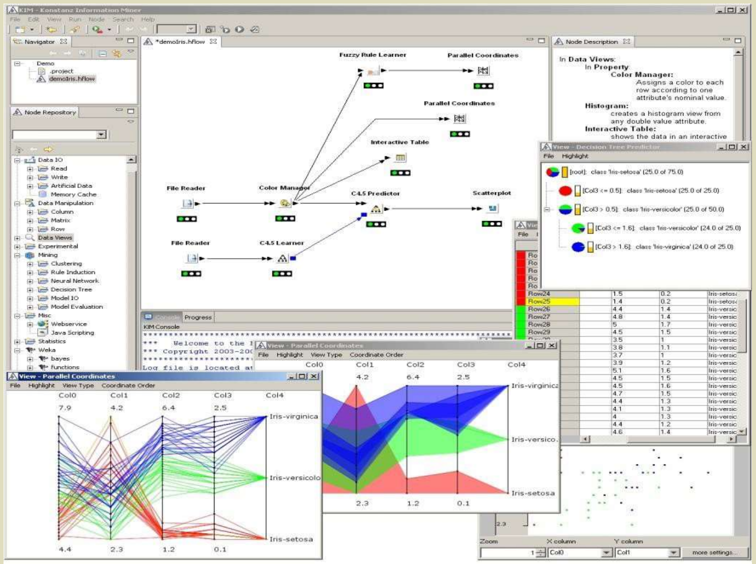
- Flow-based Programming (FBP) is a programming paradigm that defines applications as networks of "black box" processes, which exchange data across predefined connections by message passing, where the connections are specified externally to the processes.
- These black box processes can be reconnected 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

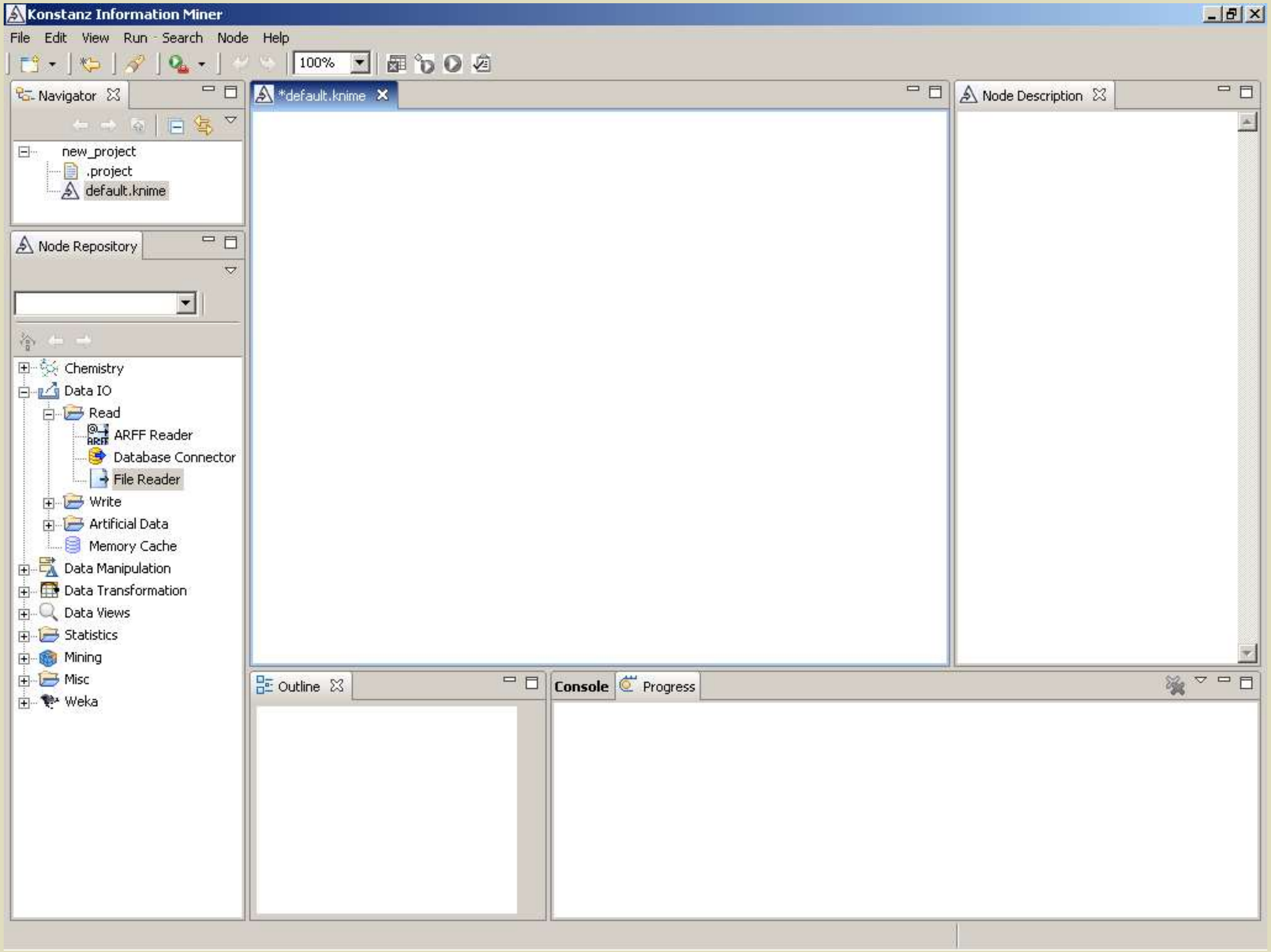
Next:

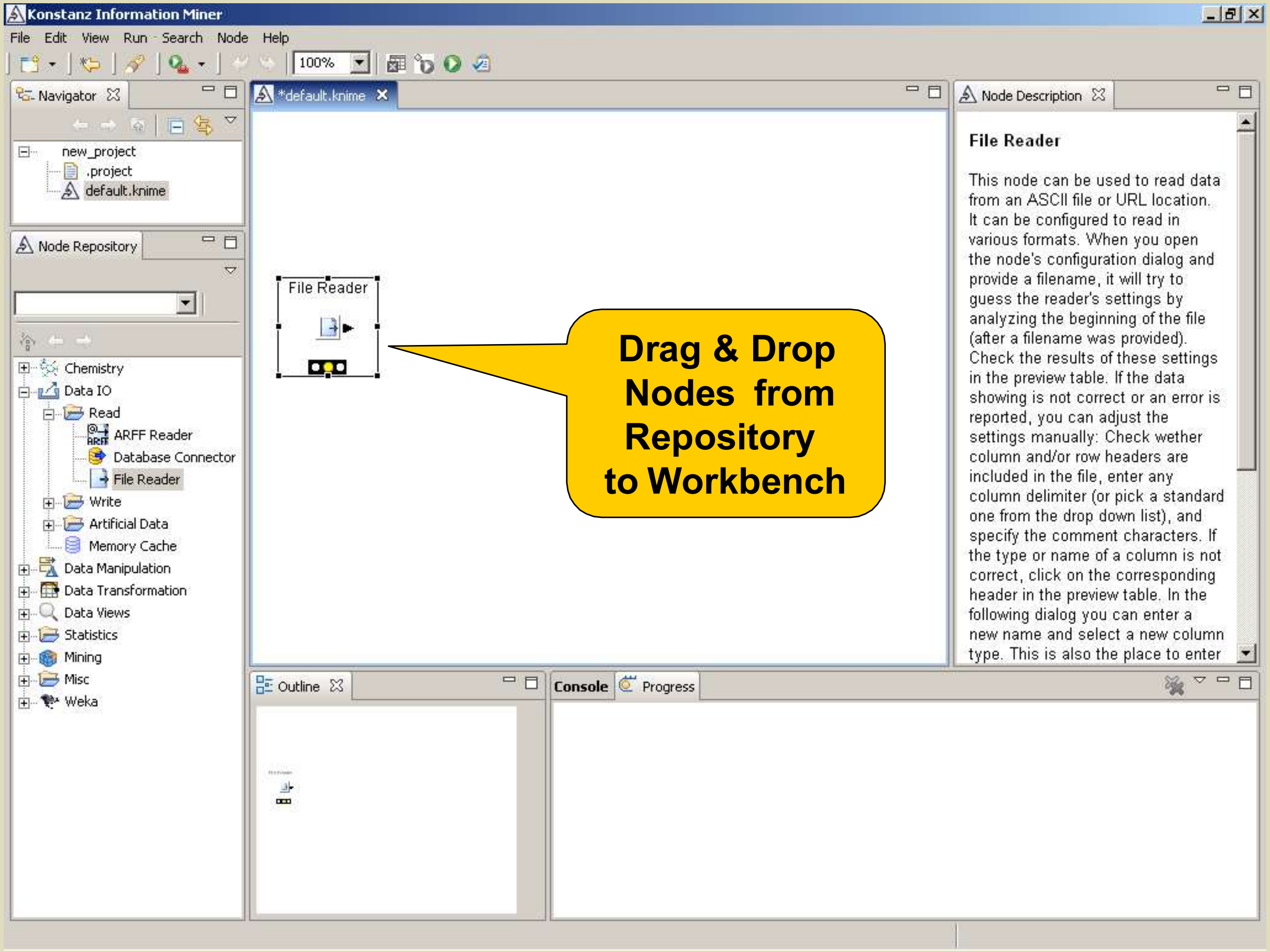
- ➢Introduction to KNIME - Demo

Demo - start

The User Interface of KNIME 1.0

The original user interface was so simple and intuitive that is still very similar in the latest version.

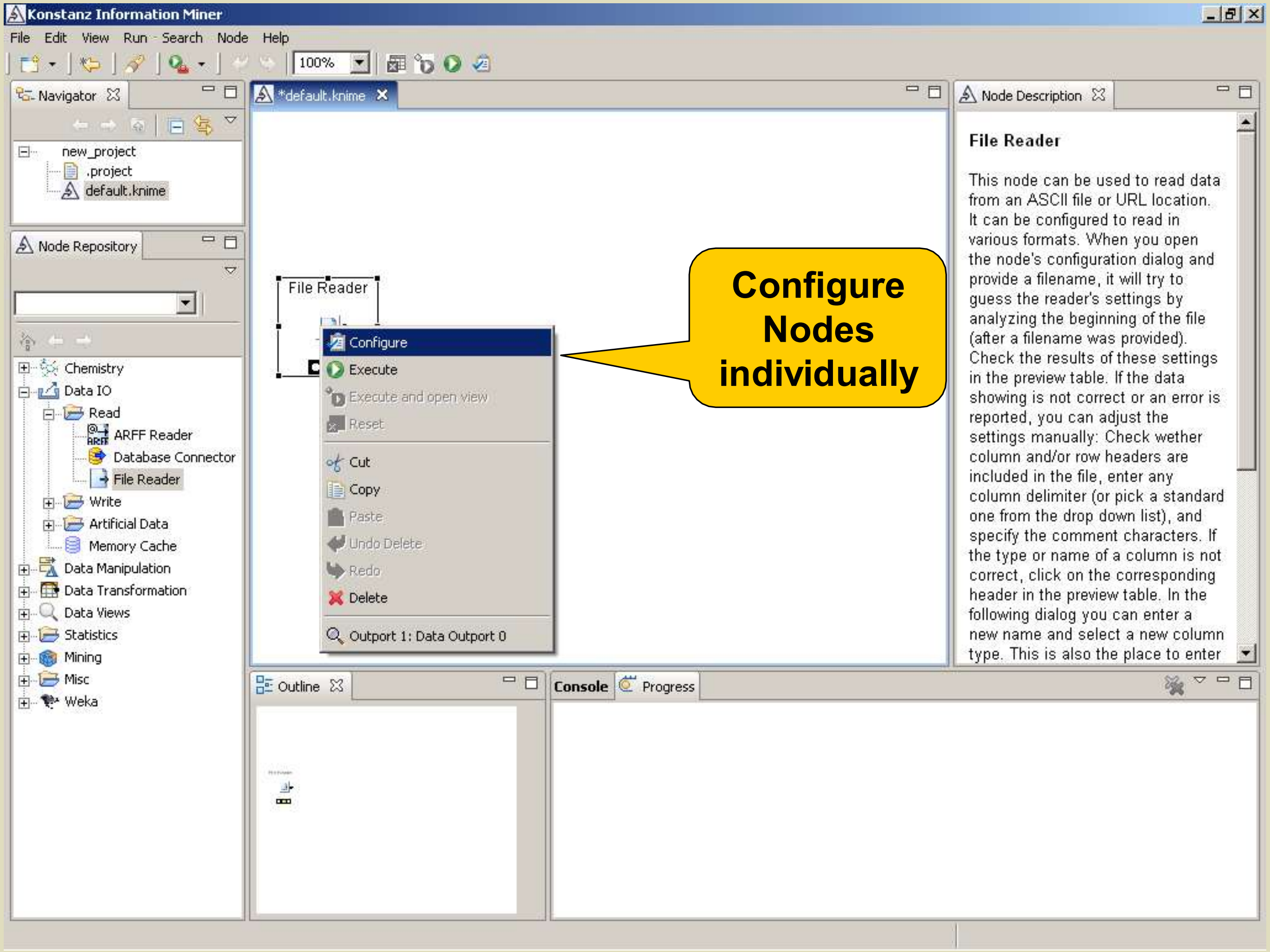




File Reader

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

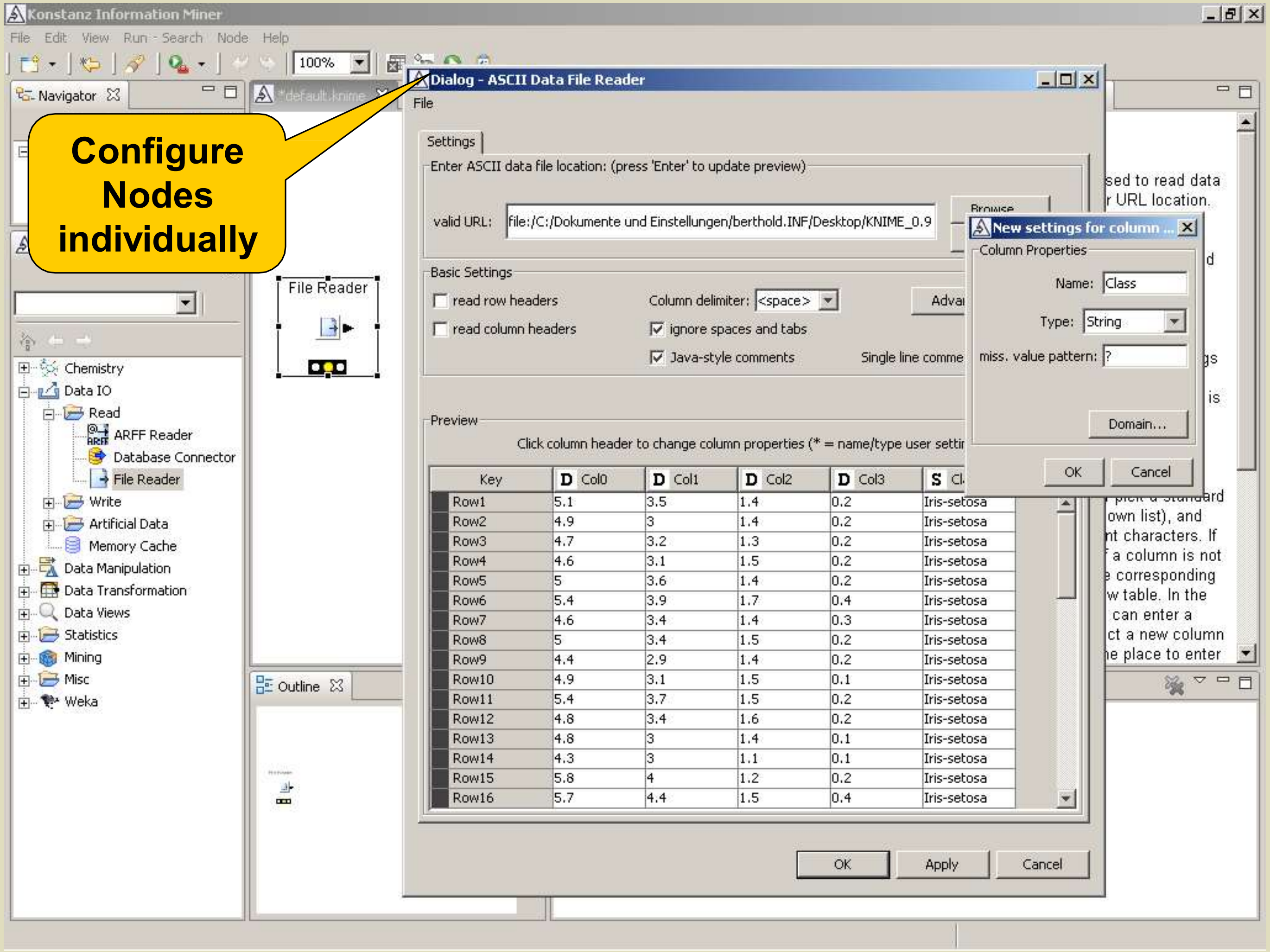
Drag & Drop Nodes from Repository to Workbench



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

Configure Nodes individually



Configure Nodes individually

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

Basic Settings

□ read row headers

Column delimiter: <space>

□ read column headers

☑ ignore spaces and tabs

☑ Java-style comments

Single line comme

Preview

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

Key	D Col0	D Col1	D Col2	D Col3	S 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

New settings for column ...

Column Properties

Name: Class

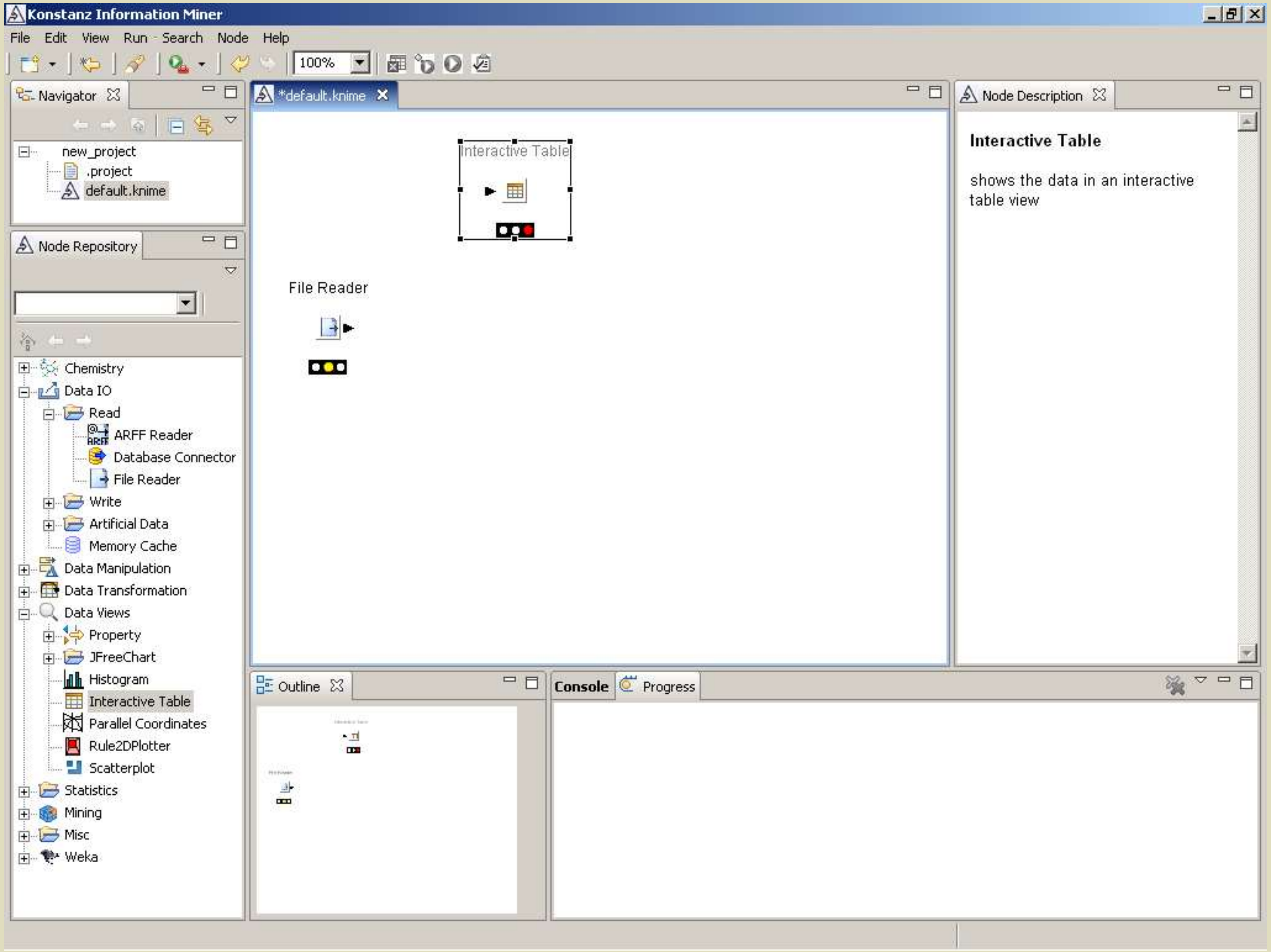
Type: String

miss. value pattern: ?

Domain...

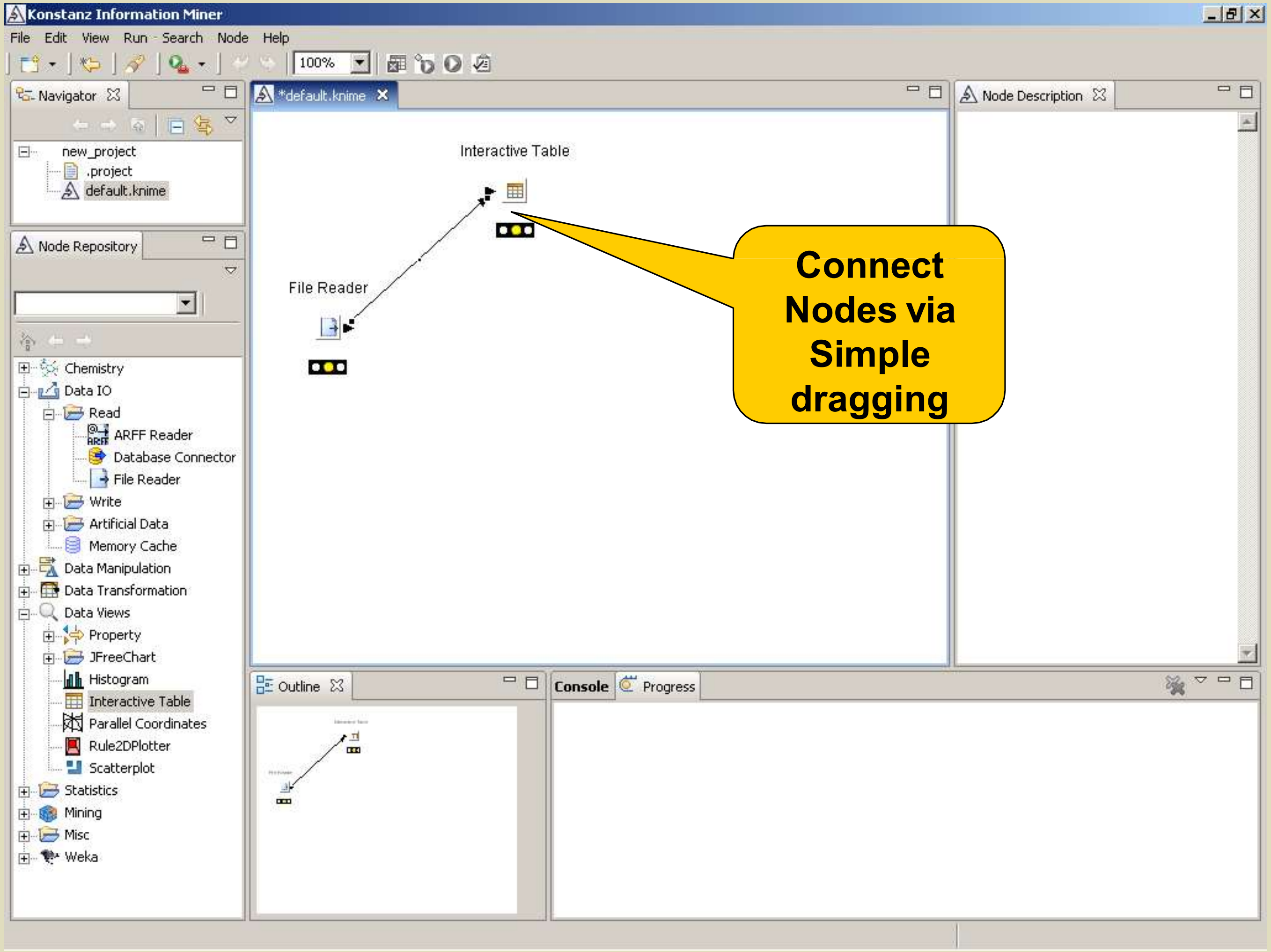
OK Cancel

sed to read data
r URL location.
d
gs
is
pick a standard
own list), and
nt characters. If
a column is not
e corresponding
w table. In the
can enter a
ct a new column
e place to enter

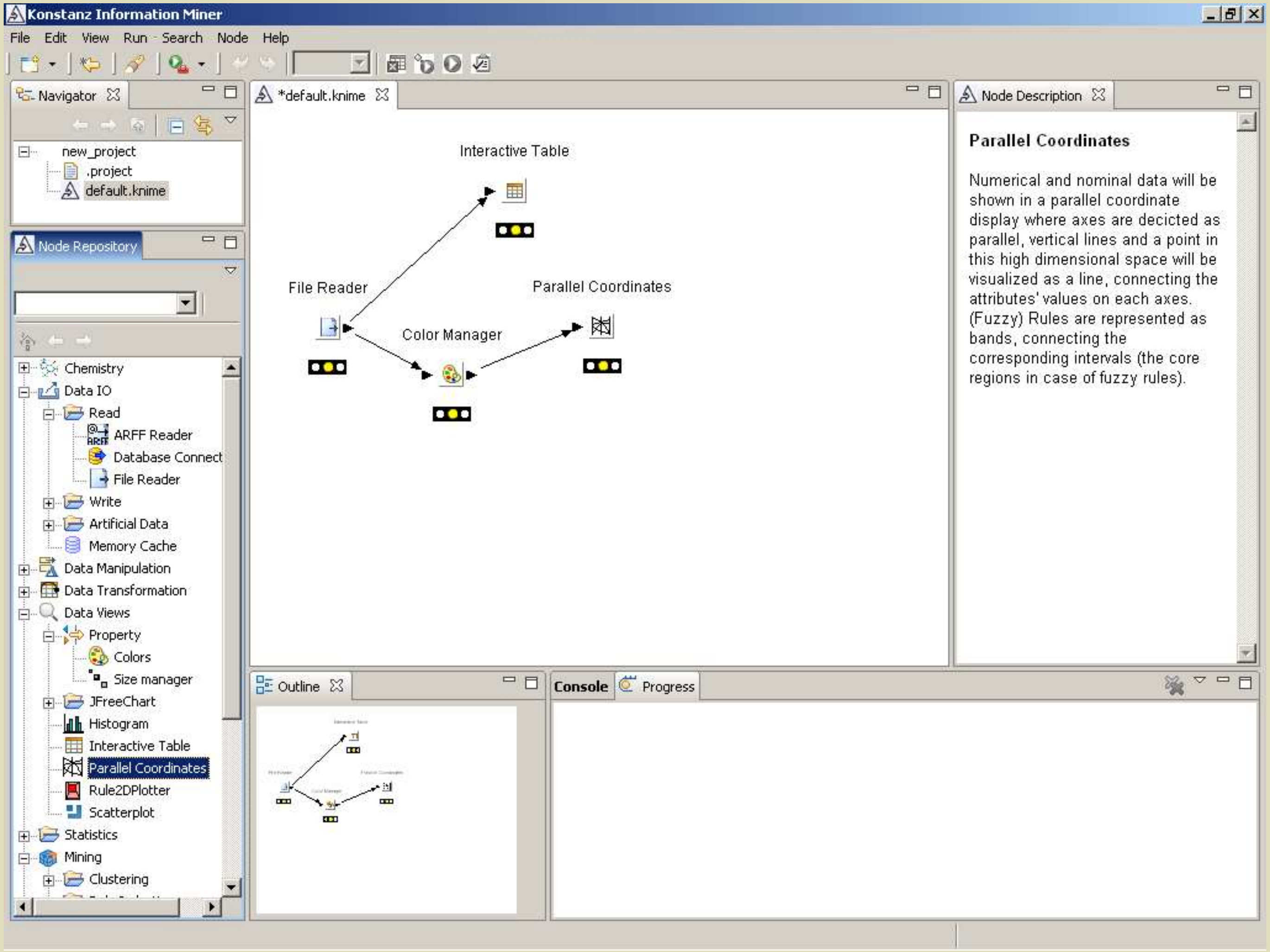


Interactive Table

shows the data in an interactive table view

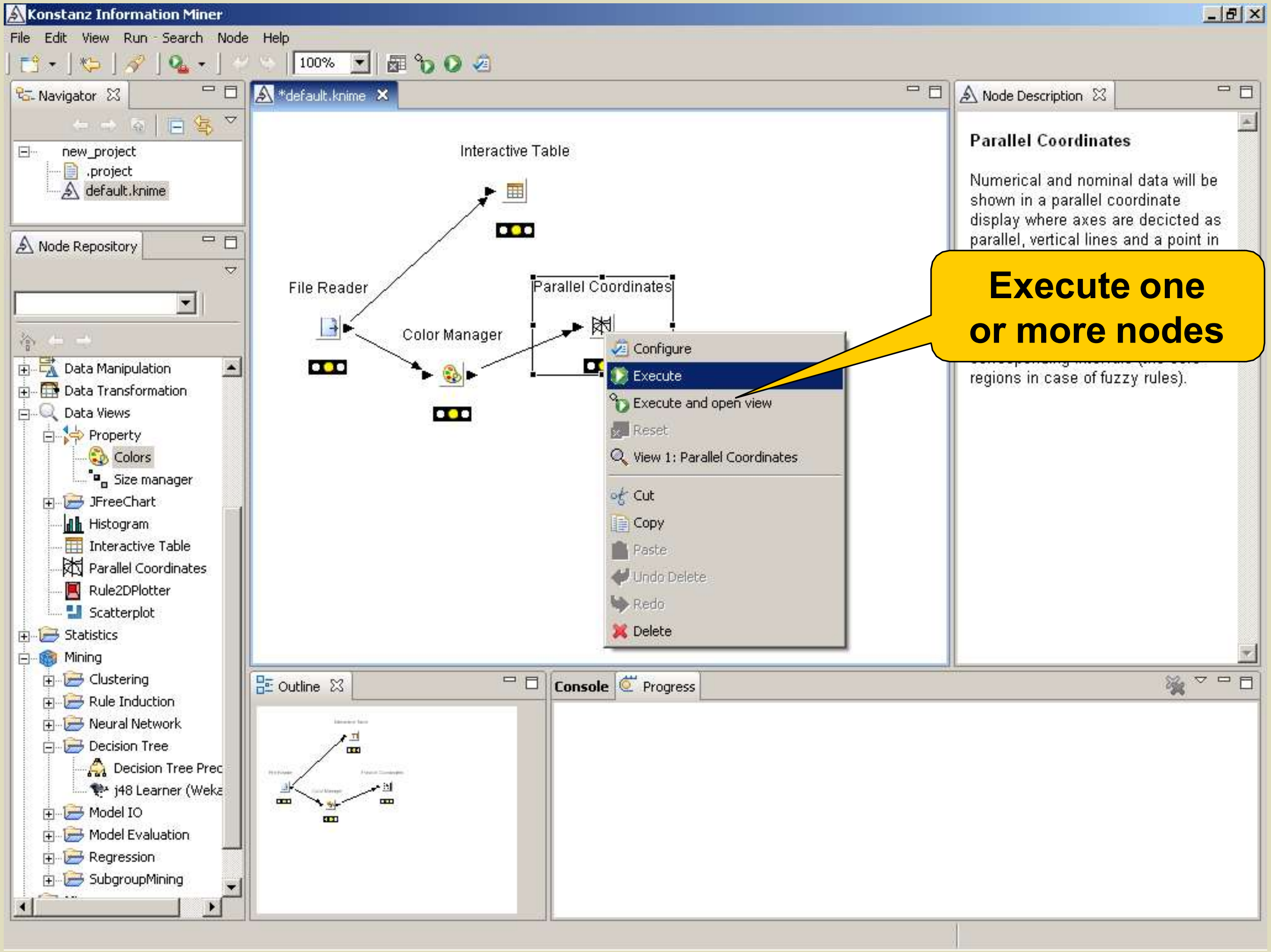


Connect Nodes via Simple dragging



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).

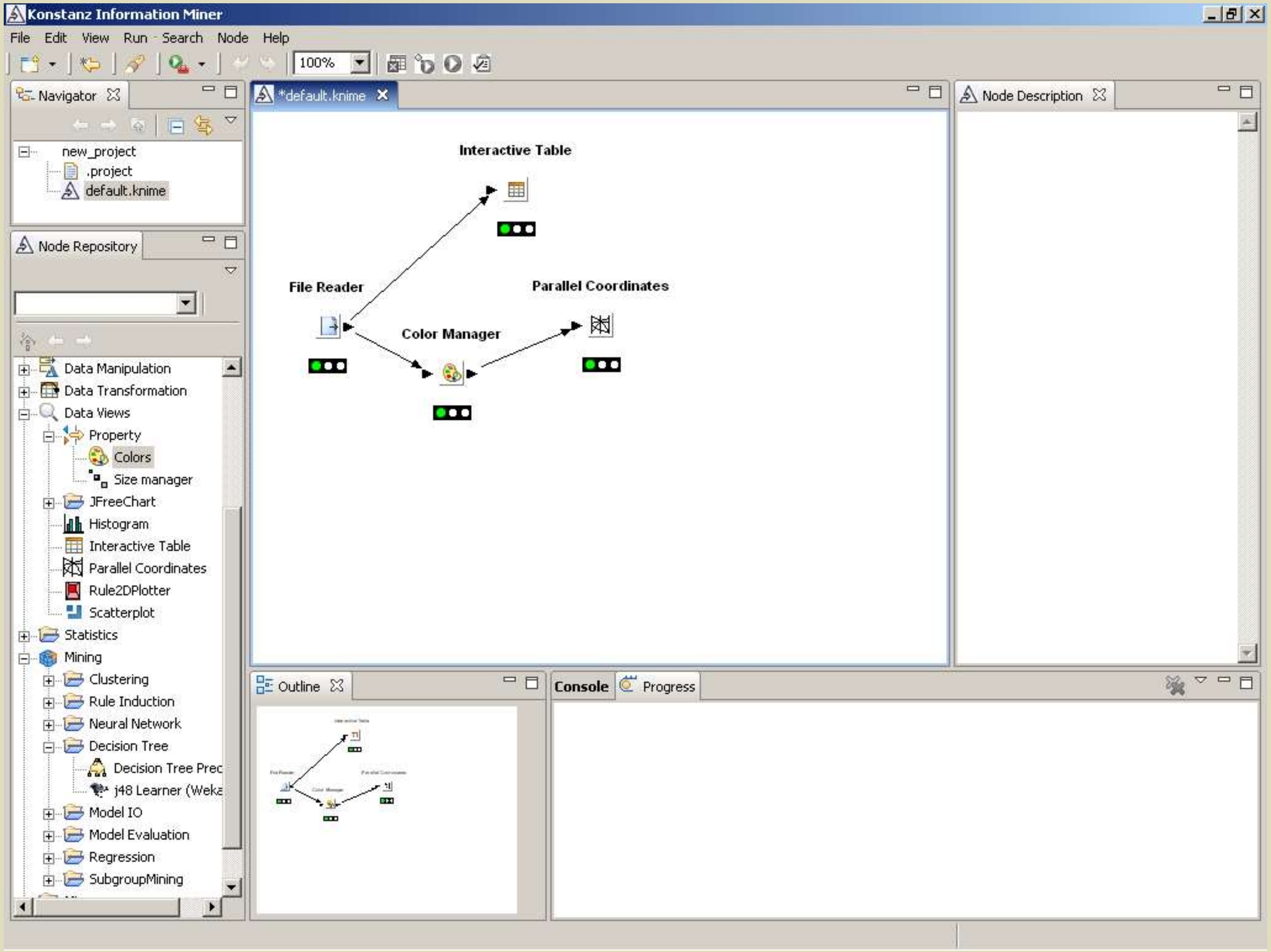


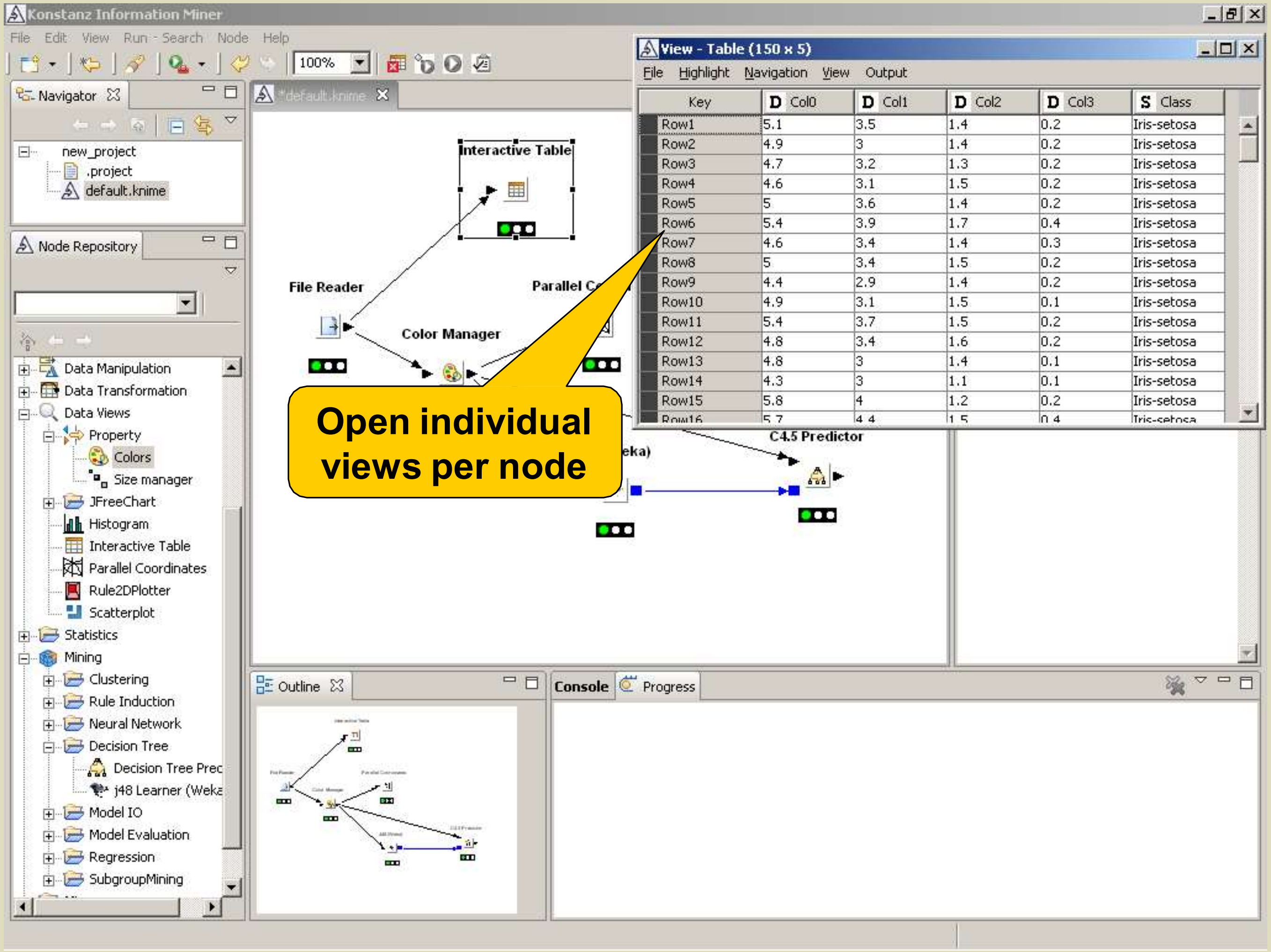
Parallel Coordinates

Numerical and nominal data will be shown in a parallel coordinate display where axes are decicted as parallel, vertical lines and a point in

regions in case of fuzzy rules).

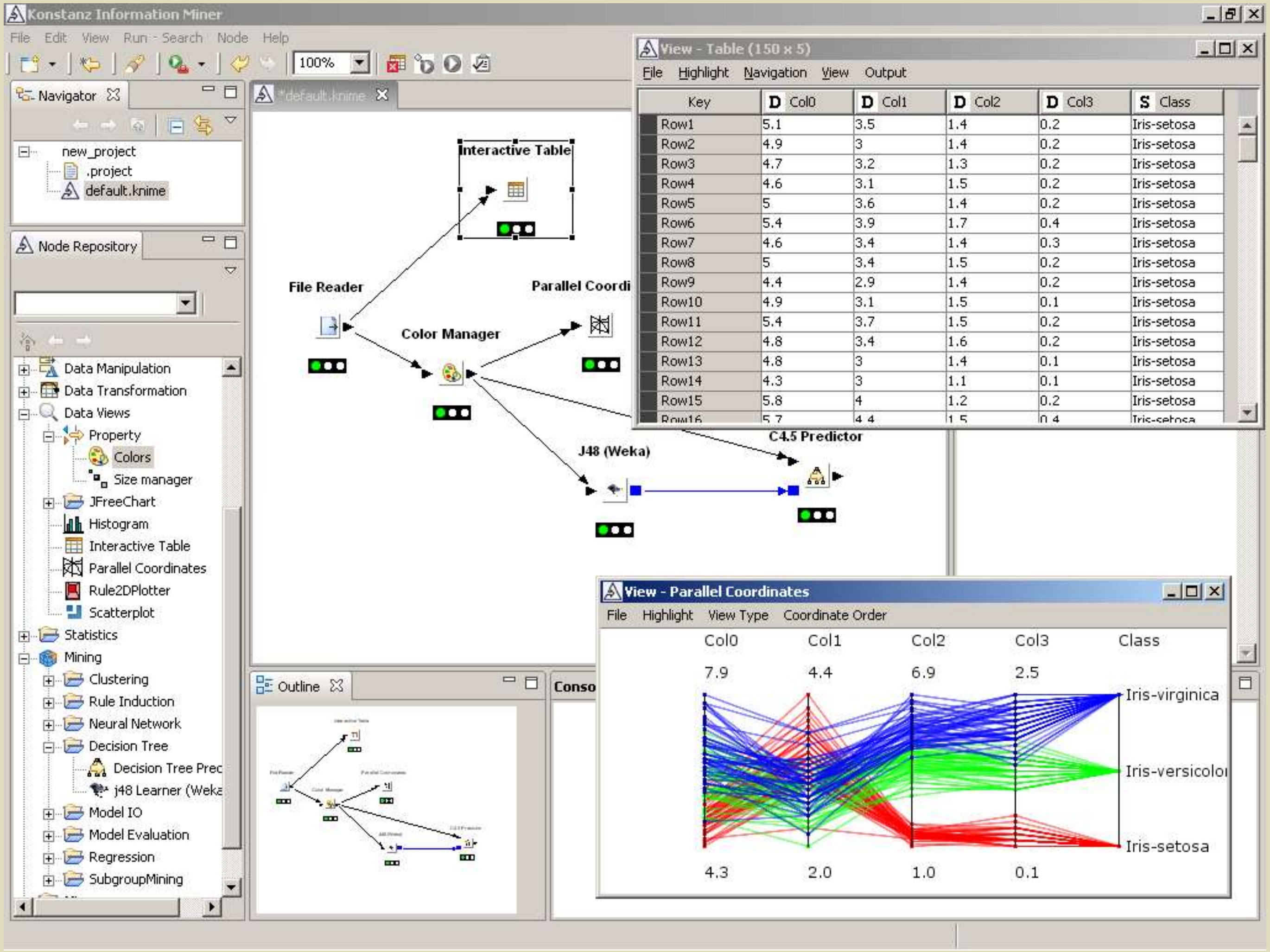
Execute one or more nodes


Konstanz Information Miner
File Edit View Run Search Node Help
100%
Navigator
new_project
.project
default.knime
*default.knime
Node Description
Interactive Table
File Reader
Parallel Coordinates
Color Manager
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
Outline
Console
Progress

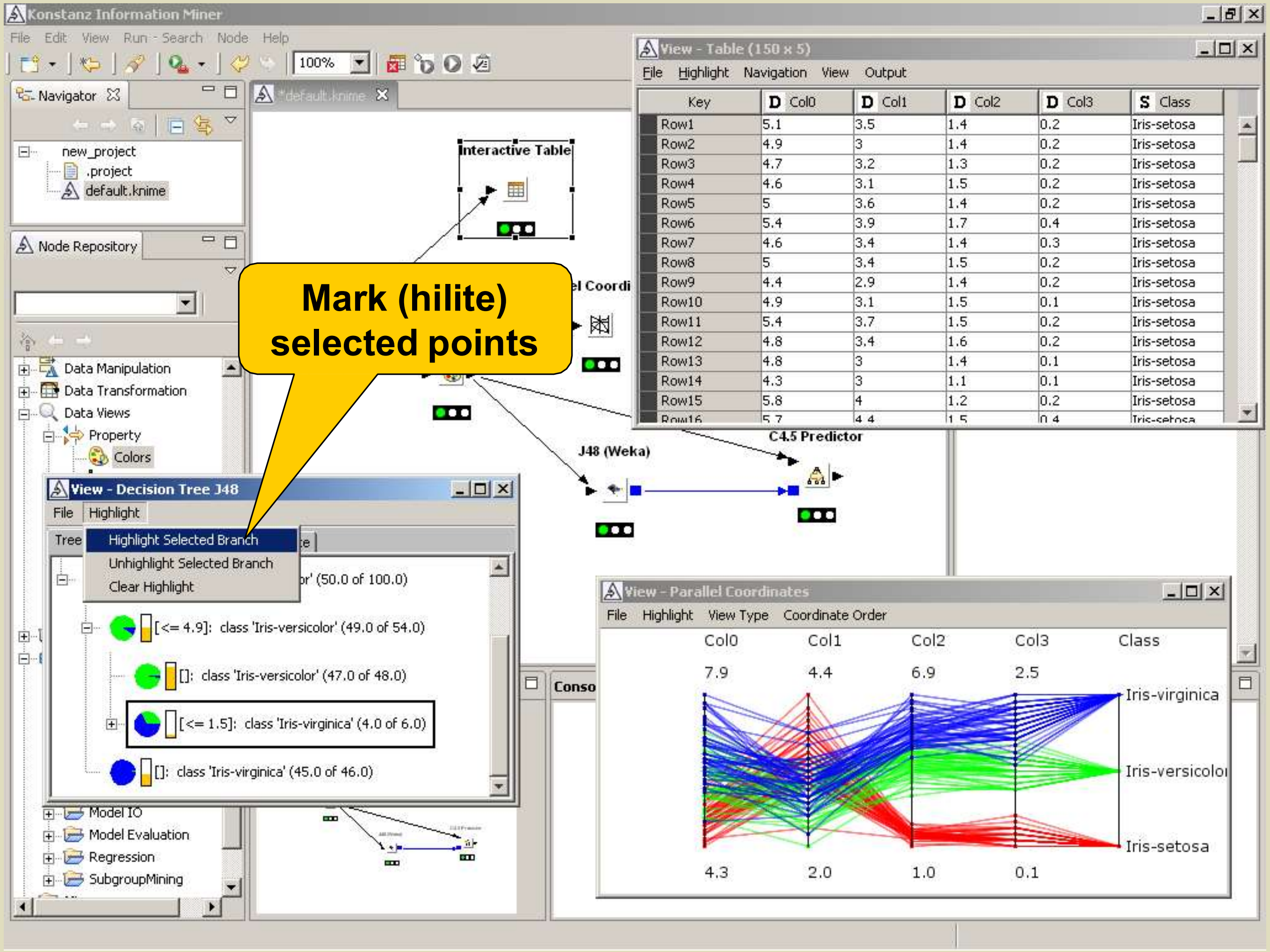


Key	D Col0	D Col1	D Col2	D Col3	S Class
Row1	5.1	3.5	1.4	0.2	Iris-setosa
Row2	4.9	3	1.4	0.2	Iris-setosa
Row3	4.7	3.2	1.3	0.2	Iris-setosa
Row4	4.6	3.1	1.5	0.2	Iris-setosa
Row5	5	3.6	1.4	0.2	Iris-setosa
Row6	5.4	3.9	1.7	0.4	Iris-setosa
Row7	4.6	3.4	1.4	0.3	Iris-setosa
Row8	5	3.4	1.5	0.2	Iris-setosa
Row9	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
Row16	5.7	4.4	1.5	0.4	Iris-setosa

Open individual views per node



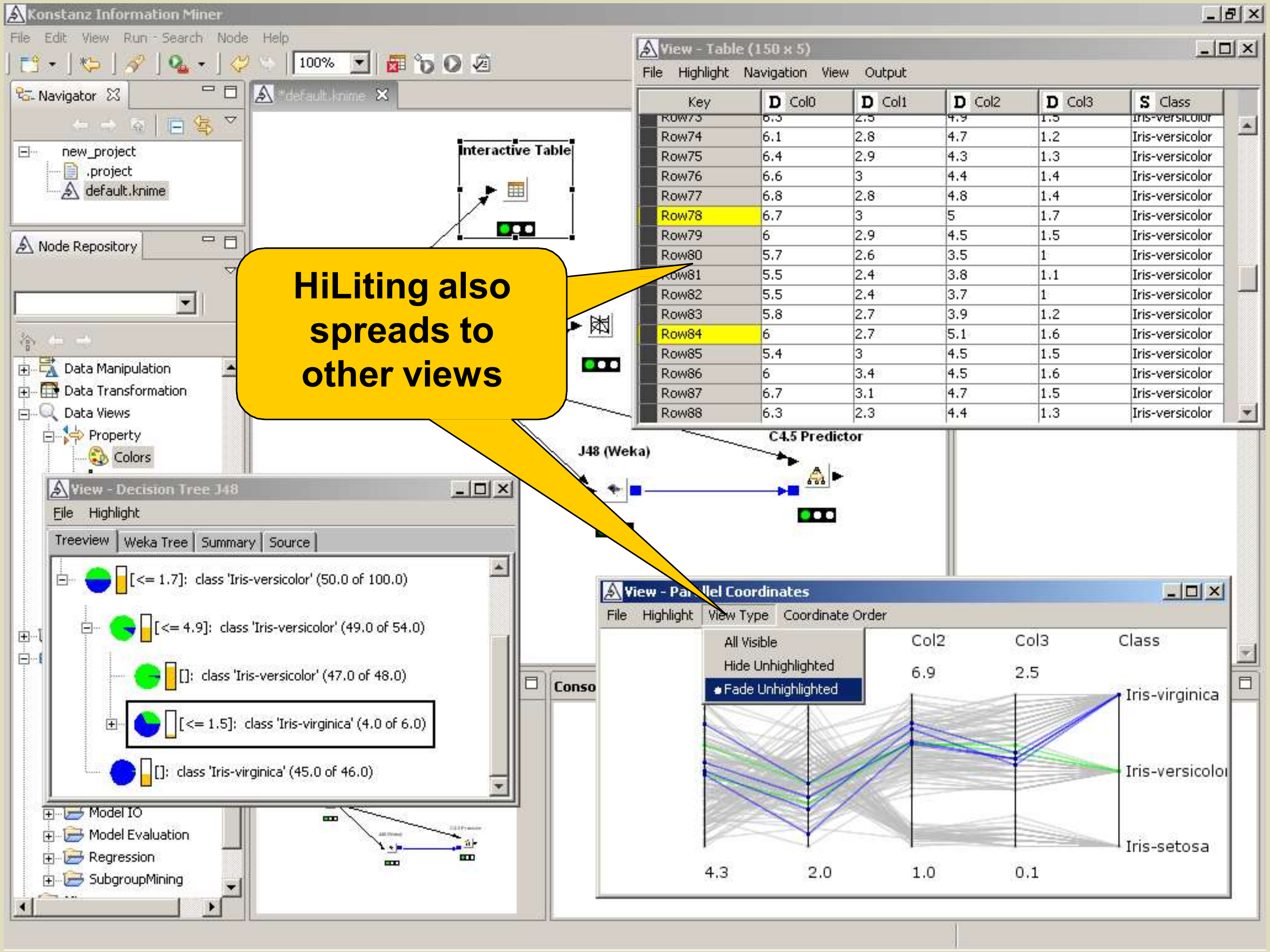
Key	D Col0	D Col1	D Col2	D Col3	S Class
Row1	5.1	3.5	1.4	0.2	Iris-setosa
Row2	4.9	3	1.4	0.2	Iris-setosa
Row3	4.7	3.2	1.3	0.2	Iris-setosa
Row4	4.6	3.1	1.5	0.2	Iris-setosa
Row5	5	3.6	1.4	0.2	Iris-setosa
Row6	5.4	3.9	1.7	0.4	Iris-setosa
Row7	4.6	3.4	1.4	0.3	Iris-setosa
Row8	5	3.4	1.5	0.2	Iris-setosa
Row9	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
Row16	5.7	4.4	1.5	0.4	Iris-setosa



Mark (hilite) selected points

Key	D Col0	D Col1	D Col2	D Col3	S 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
Row16	5.7	4.4	1.5	0.4	Iris-setosa

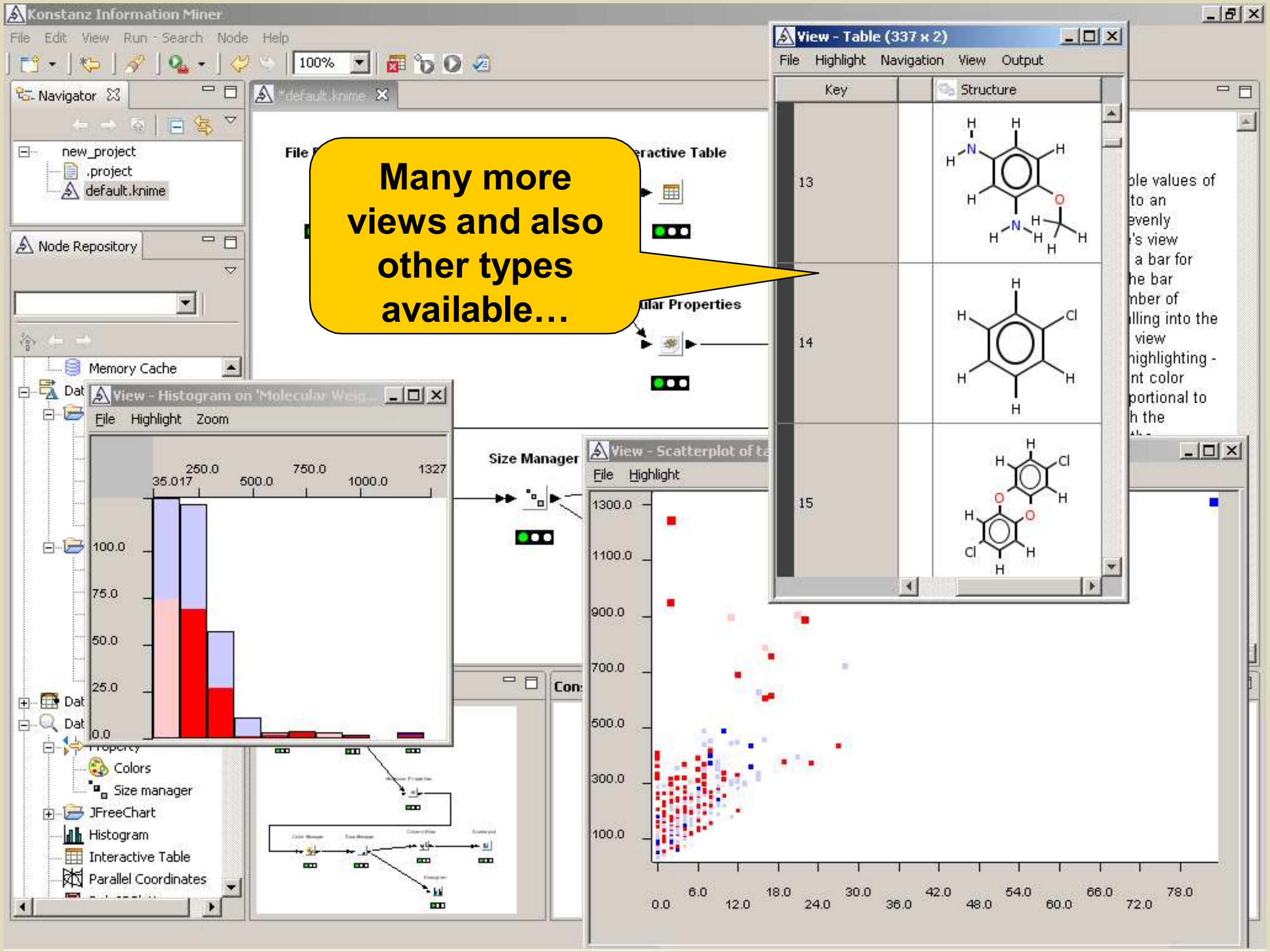
```
r' (50.0 of 100.0)
  [<= 4.9]: class 'Iris-versicolor' (49.0 of 54.0)
    []: class 'Iris-versicolor' (47.0 of 48.0)
    [<= 1.5]: class 'Iris-virginica' (4.0 of 6.0)
  []: class 'Iris-virginica' (45.0 of 46.0)
```



Key	D Col0	D Col1	D Col2	D Col3	S Class
Row73	6.3	2.5	4.9	1.5	Iris-versicolor
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

```
[<= 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)
```

HiLiting also spreads to other views



Many more views and also other types available…

Demo - end

University of Reading

CSMDM16 - Data Analytics and Mining

Introduction to KNIME - continued

Dr. Carmen Lam
Department of Computer Science
carmen.lam@reading.ac.uk

The Four States of a KNIME Node

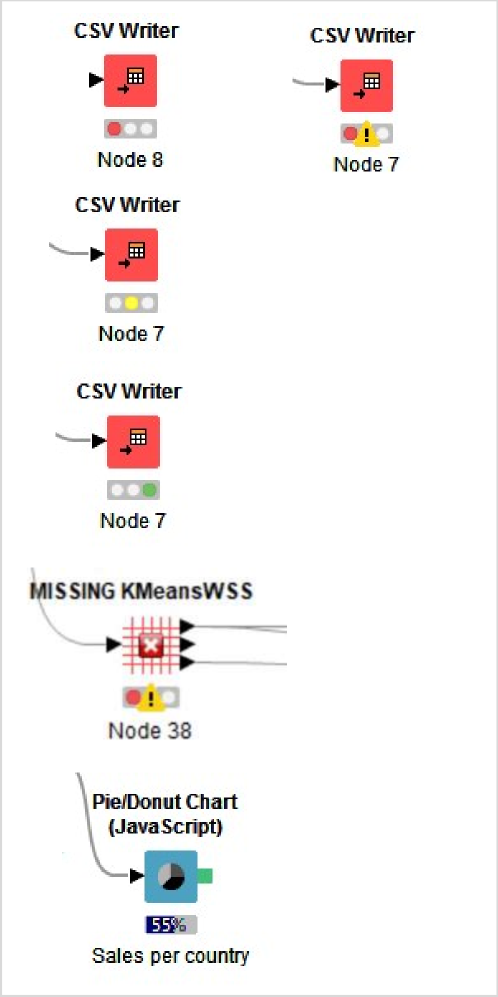
- **Not Configured** (red light): node has not been configured yet or the configuration is not valid, or the node has been executed unsuccessfully.

- **Configured** (yellow light): node configuration is valid

- **Executed** (green light): node has been executed successfully and, if applicable, output data and views are ready to be used.

- **Error**: the node is not available: the platform is missing the extension that provides the node.

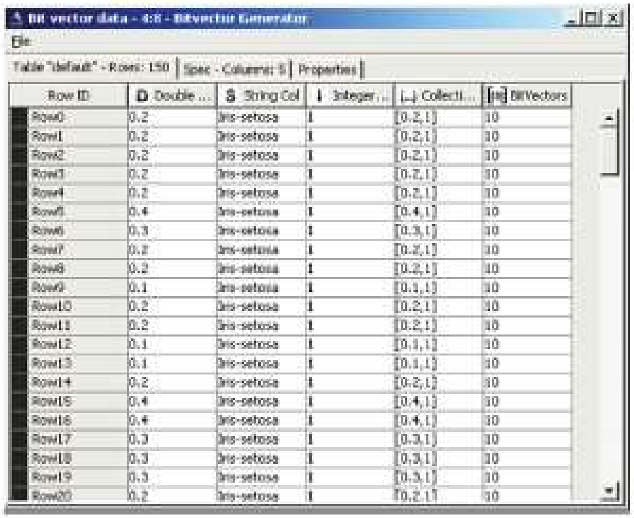
❑ Progress bar: most nodes typically provide information about the progress of their execution (%).



Data Table

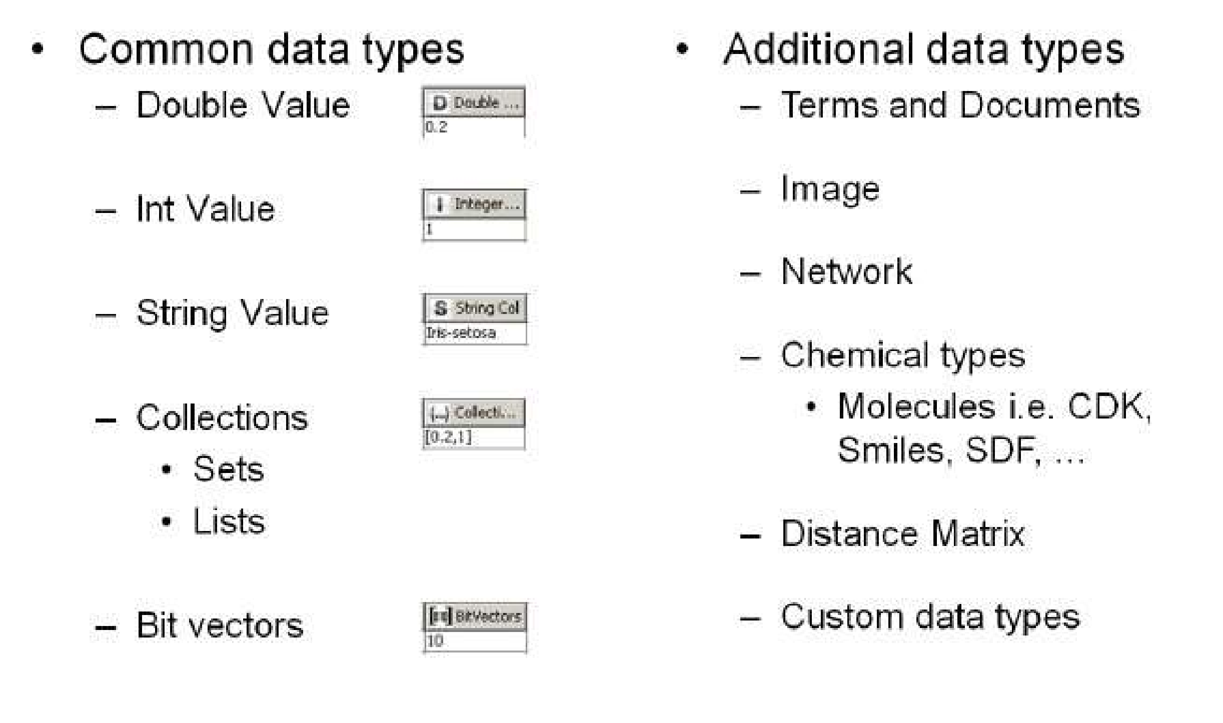
- Contains meta information (spec)
  - data types
  - domains
  - # of rows/cols
- Large tables are buffered on disc
- Blob cell support for large data cells e.g. images

Bitvector Generator


Bit vector data - 4:8 - Bitvector Generator
File
Table "default" - Rows: 150
Spec - Columns: 5
Properties

Row ID	D Double ...	S String Col	I 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

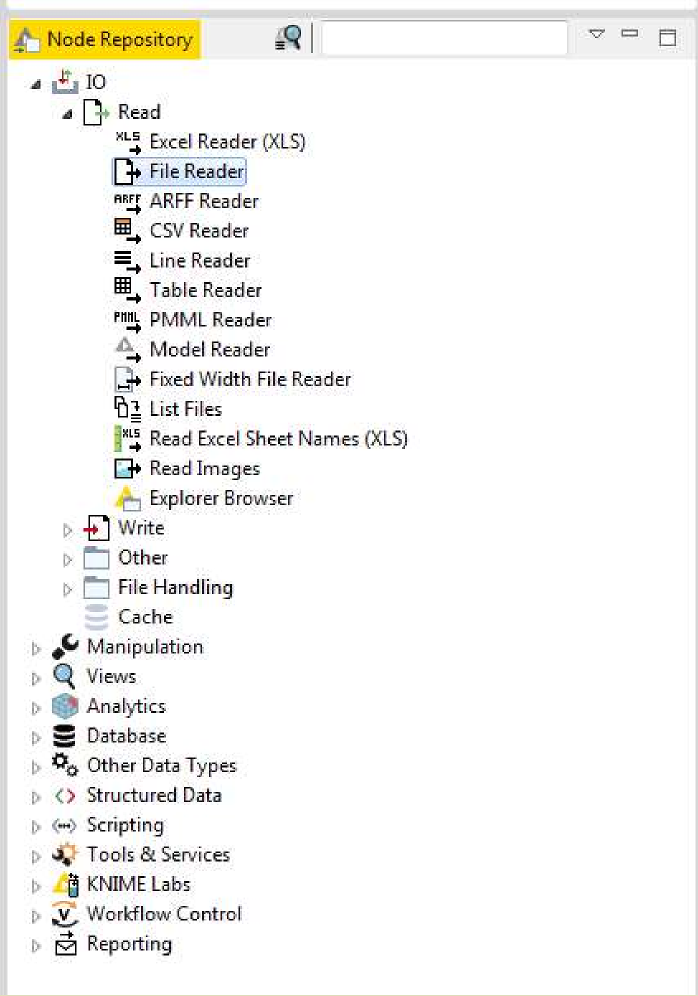
Data Types



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



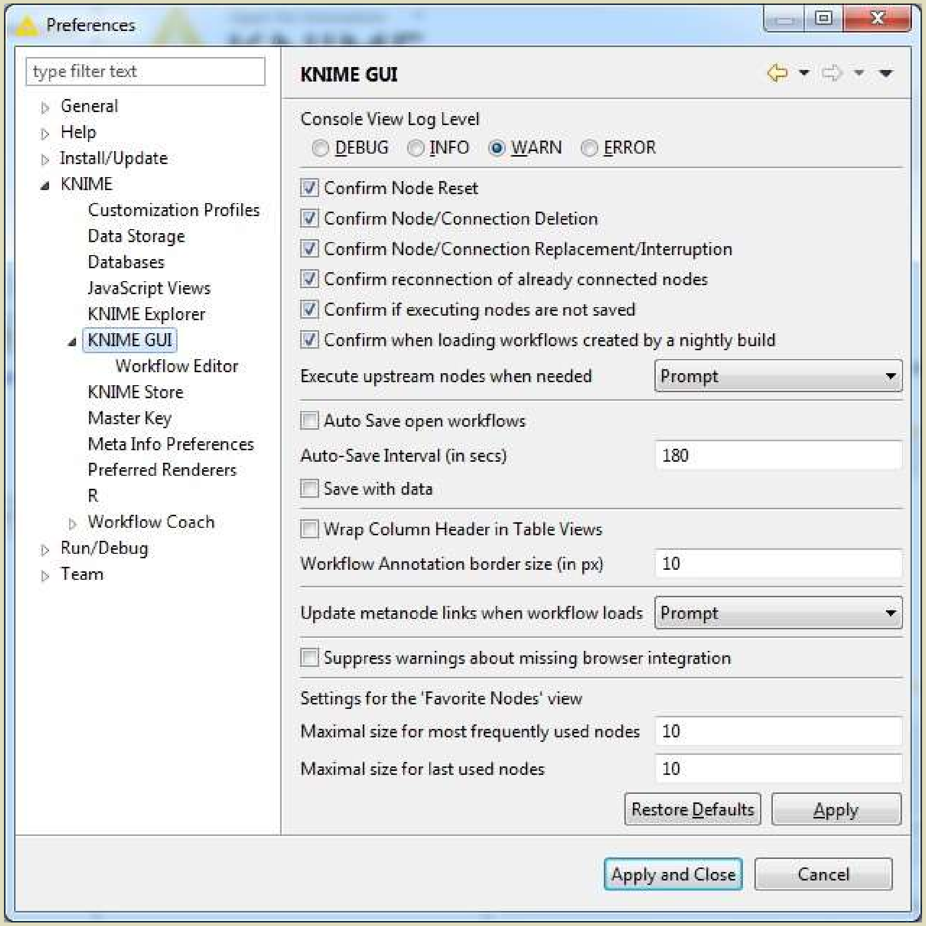
Node Repository

- IO
  - Read
    - Excel Reader (XLS)
    - File Reader
    - ARFF Reader
    - CSV Reader
    - Line Reader
    - Table Reader
    - PMML Reader
    - Model Reader
    - Fixed Width File Reader
    - List Files
    - Read Excel Sheet Names (XLS)
    - Read Images
    - Explorer Browser
  - Write
  - Other
  - File Handling
  - Cache
- Manipulation
- Views
- Analytics
- Database
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- KNIME Labs
- Workflow Control
- Reporting

Preferences

- In the KNIME Analytics Platform
  - In menu File, select Preferences

- E.g., set the level of log messages shown in the console ➔
  - DEBUG
  - INFO
  - WARN
  - ERROR



KNIME GUI

Console View Log Level
○ DEBUG ○ INFO ● WARN ○ ERROR

☑ Confirm Node Reset
☑ Confirm Node/Connection Deletion
☑ Confirm Node/Connection Replacement/Interruption
☑ Confirm reconnection of already connected nodes
☑ Confirm if executing nodes are not saved
☑ Confirm when loading workflows created by a nightly build

Execute upstream nodes when needed: Prompt

☐ Auto Save open workflows
Auto-Save Interval (in secs): 180
☐ Save with data

☐ Wrap Column Header in Table Views
Workflow Annotation border size (in px): 10

Update metanode links when workflow loads: Prompt

☐ Suppress warnings about missing browser integration

Settings for the 'Favorite Nodes' view
Maximal size for most frequently used nodes: 10
Maximal size for last used nodes: 10

File knime.ini and JVM memory

- The file **knime.ini** (in the KNIME installation folder) sets options used by the Java Virtual Machine when KNIME Analytics Platform is launched.
  - You can allocate more memory for the JVM to run KNIME by changing the **Xmx** JVM argument (default: 512 MB)
- Track memory used
  - File -> Preferences -> General -> Tick "Show heap status" checkbox

```
-startup
plugins/org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
--launcher.library
plugins/org.eclipse.equinox.launcher.win32.win32.x86_1.1.551.v20171108-1834
--launcher.defaultAction
openFile
-vm
plugins/org.knime.binary.jre.win32.x86_1.8.0.152-01/jre/bin
-vmargs
-server
-Dsun.java2d.d3d=false
-Dosgi.classloader.lock=classname
-XX:+UnlockDiagnosticVMOptions
-XX:+UnsyncloadClass
-Dsun.net.client.defaultReadTimeout=0
-XX:CompileCommand=exclude,javax/swing/text/GlyphView,getBreakSpot
-Xmx512m
-Dorg.eclipse.swt.browser.IEVersion=10001
-Dsun.awt.noerasebackground=true
-Dequinox.statechange.timeout=30000
```

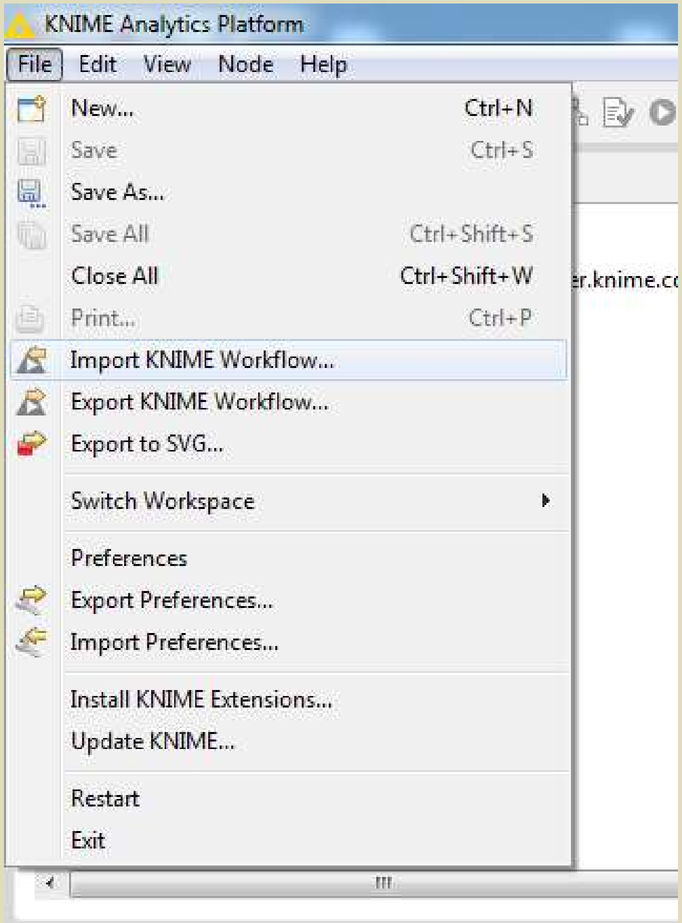
Importing/Exporting KNIME Workflows

❑ Many examples are available at KNIME Hub:

- https://hub.knime.com/
- You can search and import nodes and workflows

❑ Users can share (export and import) workflows via archive files (*archive.knwf*)

- File -> Export KNIME workflow
- File -> Import KNIME workflow



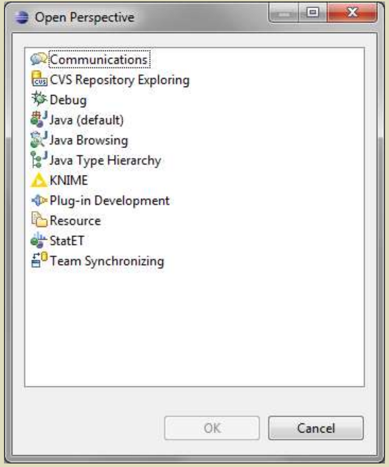
KNIME Analytics Platform

File Edit View Node Help

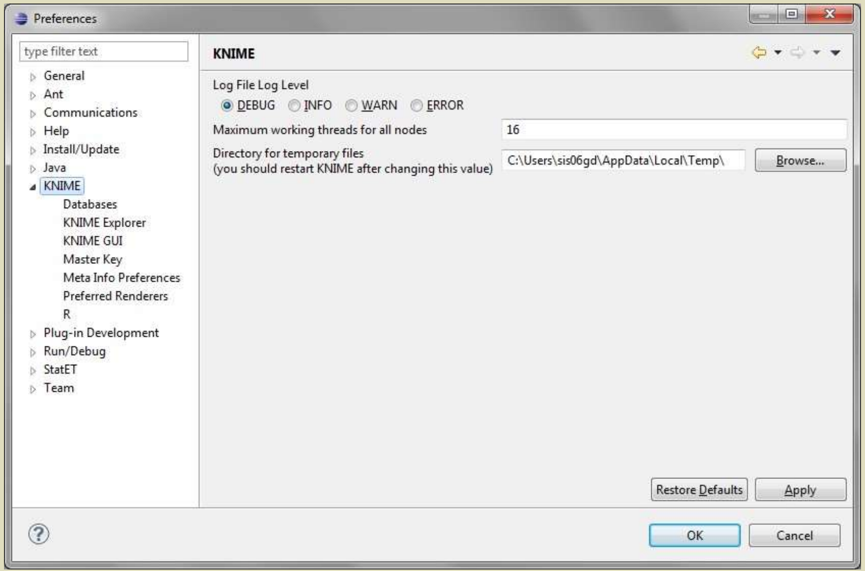
New...	Ctrl+N
Save	Ctrl+S
Save As...	
Save All	Ctrl+Shift+S
Close All	Ctrl+Shift+W
Print...	Ctrl+P
Import KNIME Workflow...	
Export KNIME Workflow...	
Export to SVG...	
Switch Workspace	
Preferences	
Export Preferences...	
Import Preferences...	
Install KNIME Extensions...	
Update KNIME...	
Restart	
Exit	

Perspective and Preferences (in KNIME SDK)

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



- In Menu Windows, select Preferences



KNIME

Log File Log Level

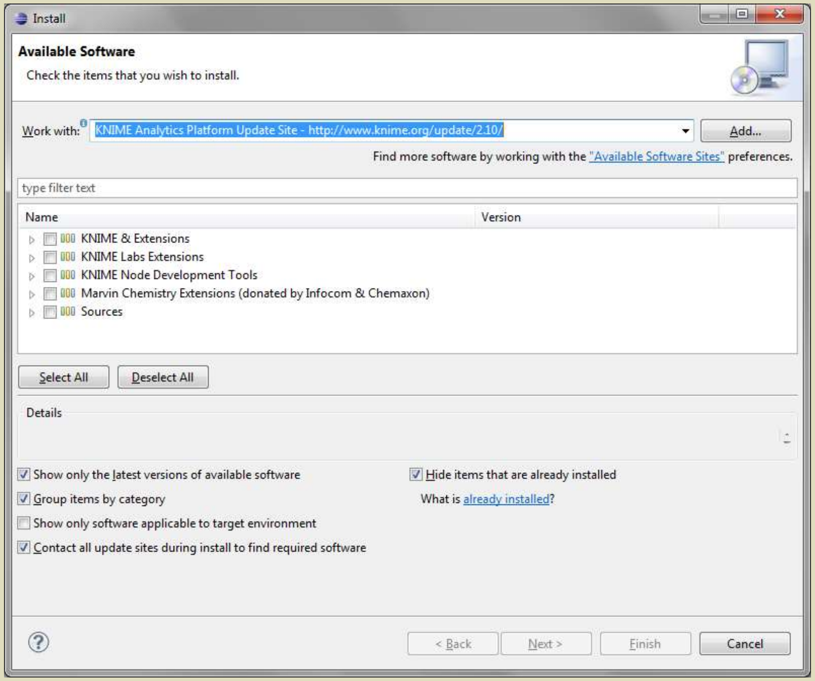
● DEBUG ○ INFO ○ WARN ○ ERROR

Maximum working threads for all nodes: 16

Directory for temporary files
(you should restart KNIME after changing this value): C:\Users\sis06gd\AppData\Local\Temp\ Browse...

KNIME Updates and Extensions

- Extensions are additional features (w.r.t. the basic installation) and are installed via the **KNIME update site**.
  - In KNIME click “Help -> Install KNIME extension...".
  - select the KNIME update site (*http://update.knime.com/analytics-platform/3.7)*
  - select the features you want to install in the dialog.
  - You need to restart KNIME after installing new extensions in order to get them activated.



Install

Available Software

Check the items that you wish to install.

Work with: KNIME Analytics Platform Update Site - http://www.knime.org/update/2.10/ Add...

Find more software by working with the "Available Software Sites" preferences.

type filter text

Name	Version
▷ □ KNIME & Extensions	
▷ □ KNIME Labs Extensions	
▷ □ KNIME Node Development Tools	
▷ □ Marvin Chemistry Extensions (donated by Infocom & Chemaxon)	
▷ □ Sources	

Select All | Deselect All

Details

☑ Show only the latest versions of available software

☑ Group items by category

□ Show only software applicable to target environment

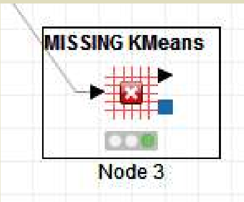
☑ Contact all update sites during install to find required software

☑ Hide items that are already installed

What is already installed?

< Back | Next > | Finish | Cancel

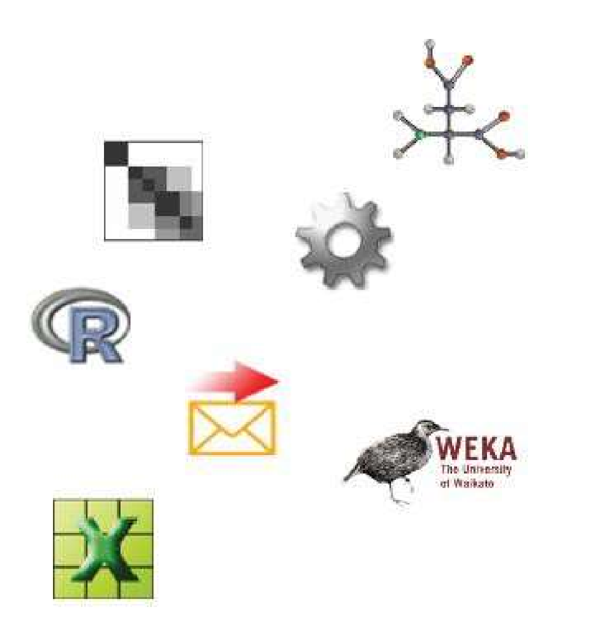
➢ If you import an external workflow and a node is “missing” (warning), you then need to install the specific extension that provides that node.



KNIME Extensions (Plugins/Dropins)

Some available extensions include:

- Chemistry types and features
- Distance Matrix
- Ensemble Learning
- Item Set Mining
- R Statistics Integration
- Python integration
- Weka Data Mining Integration
- HTML/PDF Writer
- Report Designer
- Webservice Client
- XLS Support
- XML Processing
- Cloud connectors (Amazon, Azure, etc.)
- etc.



KNIME Extensions

- Experimental “Lab” Extensions: http://tech.knime.org/knime-labs
  - JavaScript views
  - 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 can create your own KNIME nodes (extensions) by using the KNIME SDK version.

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”

- – https://www.knime.com/downloads

KNIME resources ➔ learning material, examples, etc.:

- – https://www.knime.com/learning
- – https://www.knime.com/learning-hub
- – https://www.knime.com/knimepress
- – https://www.youtube.com/user/KNIMETV

For advanced programmers (based on Java Reflection metaprogramming):

KNIME developer ➔ SDK version (Eclipse):

- – https://www.knime.com/developers
- – https://github.com/knime/knime-sdk-setup
- – 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

Next:

- P01: practical on KNIME Basics

Next week:

- Proximity Measures