



Open for Innovation[®]

KNIME

KNIME Analytics Platform Course for Beginners

KNIME AG

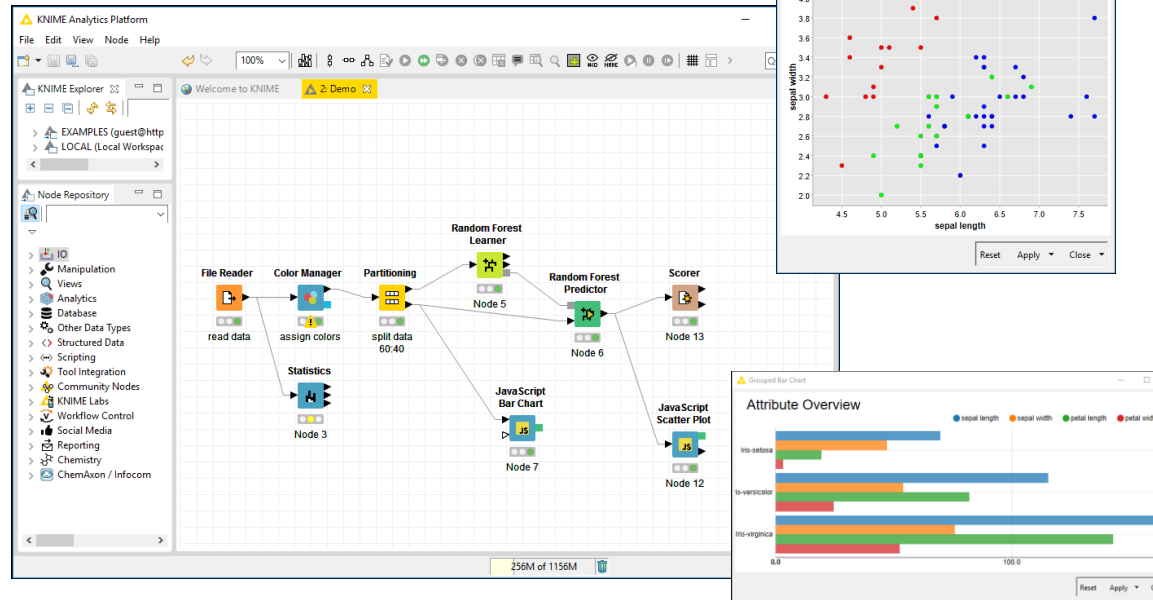
Overview

KNIME Analytics Platform



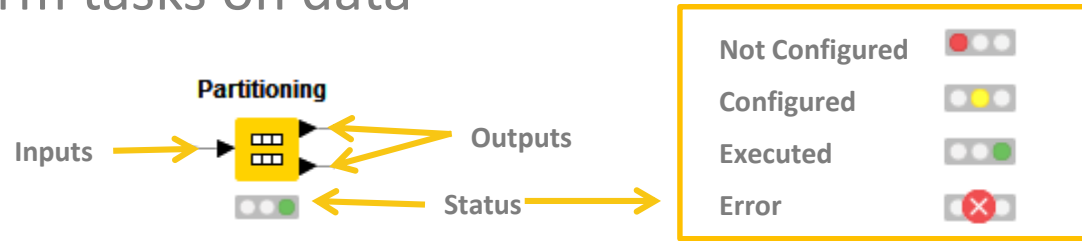
What is KNIME Analytics Platform?

- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
 - Text Mining
 - Network Mining
 - Cheminformatics
 - Many integrations, such as Java, R, Python, Weka, H2O, etc.

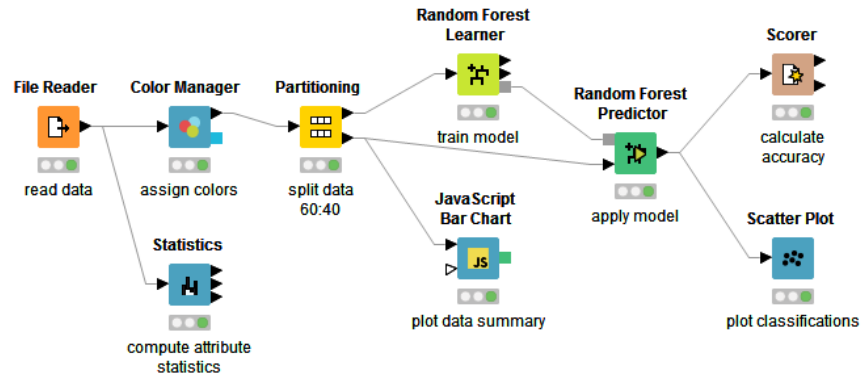


Visual KNIME Workflows

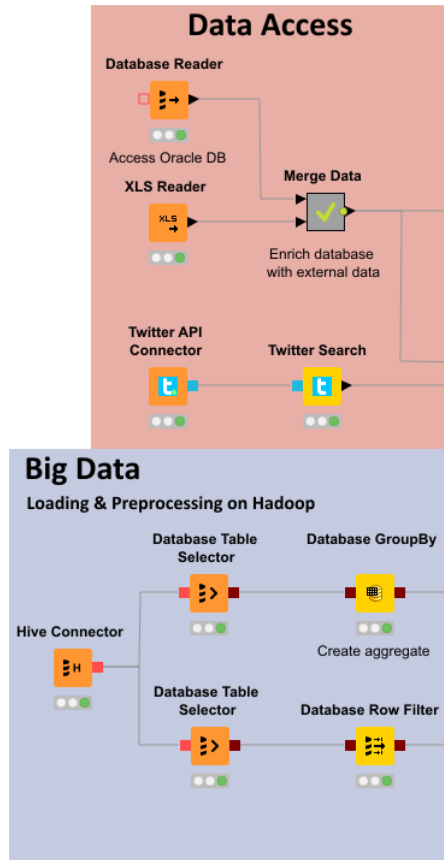
NODES perform tasks on data



Nodes are combined to create **WORKFLOWS**

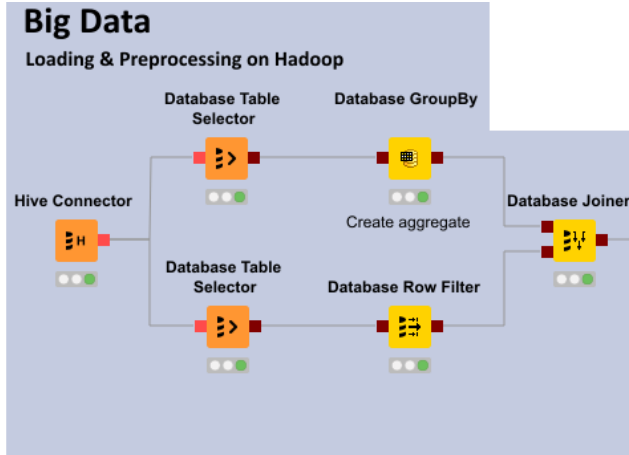


Data Access



- Databases
 - MySQL, MS SQL Server, PostgreSQL
 - any JDBC (Oracle, DB2, ...)
- Files
 - CSV, txt
 - Excel, Word, PDF
 - SAS, SPSS
 - XML, JSON
 - PMML
 - Images, texts, networks, chem
- Web, Cloud
 - REST, Web services
 - Twitter, Google

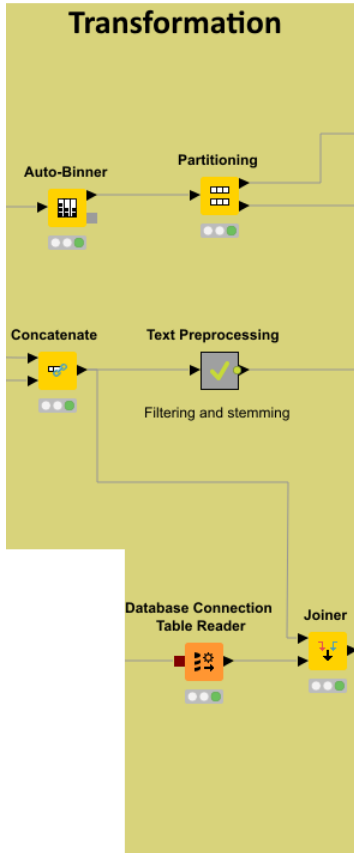
Big Data



- Spark
- HDFS support
- Hive
- Impala
- Vertica
- In-database processing

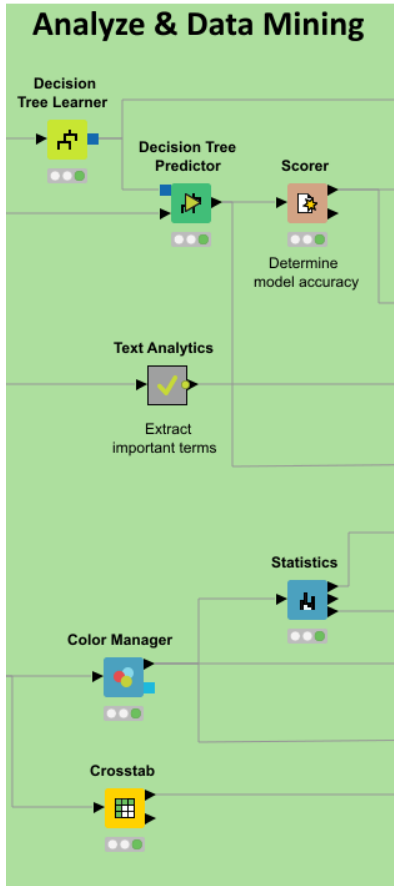


Transformation



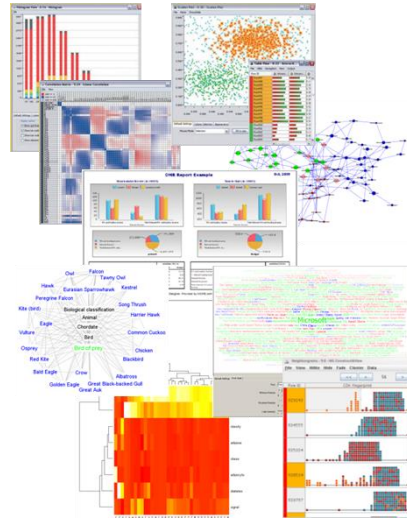
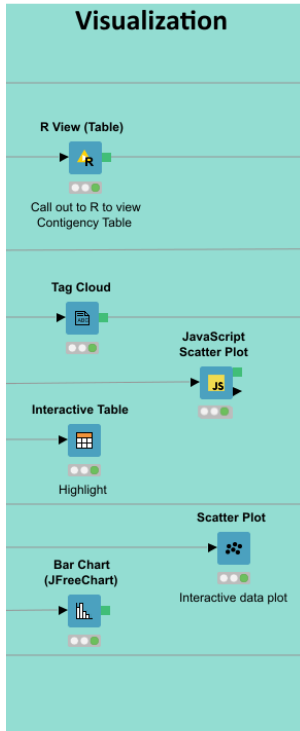
- Preprocessing
 - Row, column, matrix based
- Data blending
 - Join, concatenate, append
- Aggregation
 - Grouping, pivoting, binning
- Feature Creation and Selection

Analysis & Data Mining



- Regression
 - Linear, logistic
- Classification
 - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
 - k-means, DBSCAN, hierarchical
- Validation
 - Cross-validation, scoring, ROC
- Deep Learning
 - Keras, DL4J
- External
 - R, Python, Weka, H2O

Visualization



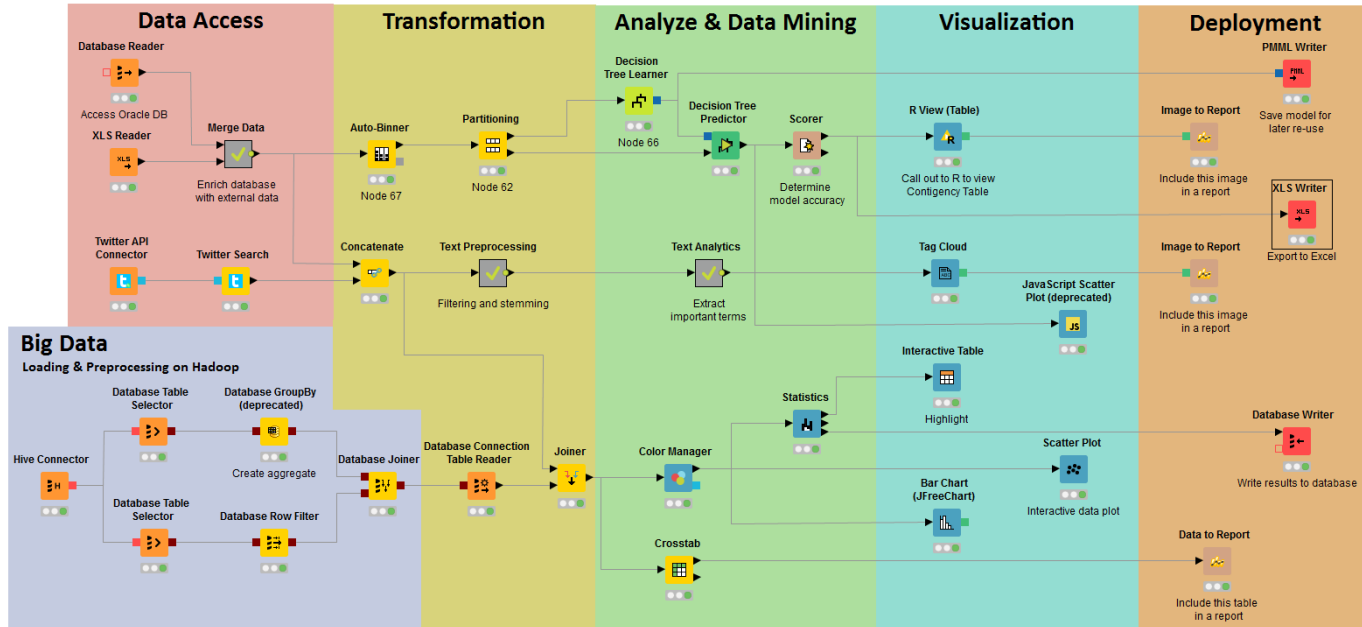
- Interactive Visualizations
- JavaScript-based nodes
 - Scatter Plot, Box Plot, Line Plot
 - Networks, ROC Curve, Decision Tree
 - Adding more with each release!
- Misc
 - Tag cloud, open street map, molecules
- Script-based visualizations
 - R, Python

Deployment



- Database
- Files
 - Excel, CSV, txt
 - XML
 - PMML
 - to: local, KNIME Server, SSH-, FTP-Server
- BIRT Reporting

Over 1500 native and embedded nodes included:



Data Access

MySQL, Oracle, ...
 SAS, SPSS, ...
 Excel, Flat, ...
 Hive, Impala, ...
 XML, JSON, PMML
 Text, Doc, Image, ...
 Web Crawlers
 Industry Specific
 Community / 3rd

Transformation

Row
 Column
 Matrix
 Text, Image
 Time Series
 Java
 Python
 Community / 3rd

Analysis & Mining

Statistics
 Data Mining
 Machine Learning
 Web Analytics
 Text Mining
 Network Analysis
 Social Media Analysis
 R, Weka, Python
 Community / 3rd

Visualization

R
 JFreeChart
 JavaScript
 Community / 3rd

Deployment

via BIRT
 PMML
 XML, JSON
 Databases
 Excel, Flat, etc.
 Text, Doc, Image
 Industry Specific
 Community / 3rd

Overview

- Installing KNIME Analytics Platform
- The KNIME Workspace
- The KNIME File Extensions
- The KNIME Workbench
 - Workflow editor
 - Explorer
 - Node repository
 - Node description
- Installing new features

Install KNIME Analytics Platform

- Select the KNIME version for your computer:
 - Mac, Win, or Linux and 32 / 64bit
- Download archive and extract the file, or download installer package and run it

Windows		
KNIME Analytics Platform for Windows (installer)	32 Bit	(393.38 MB)
<i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	64 Bit	(396.38 MB)
KNIME Analytics Platform for Windows (self-extracting archive)	32 Bit	(396.87 MB)
<i>The self-extracting archive only creates a folder holding the KNIME installation</i>	64 Bit	(400.72 MB)
KNIME Analytics Platform for Windows (zip archive)	32 Bit	(466.11 MB)
	64 Bit	(470.07 MB)

Linux		
KNIME Analytics Platform for Linux	64 Bit	(417.21 MB)

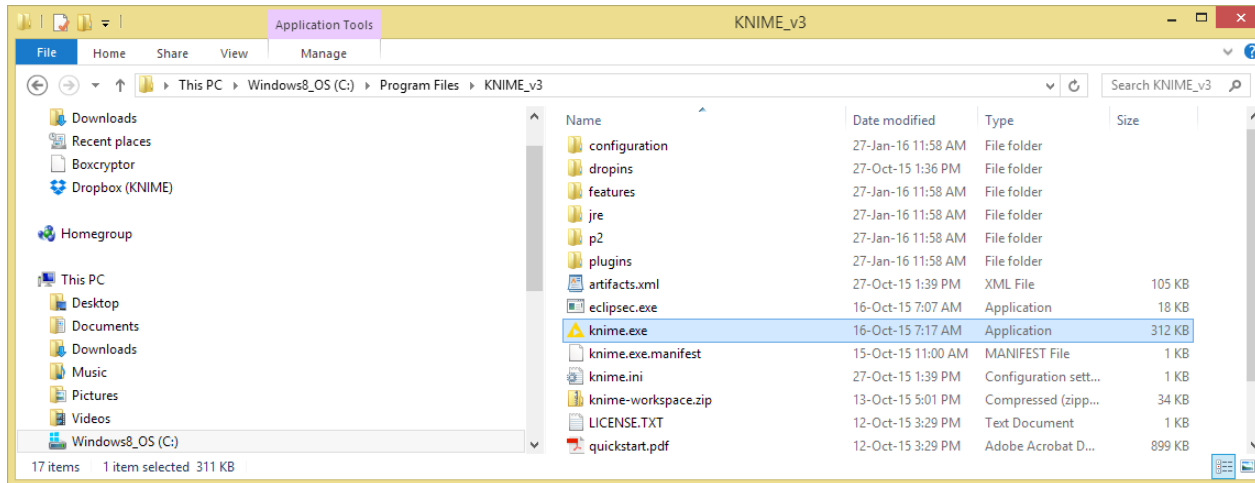
Mac		
KNIME Analytics Platform for Mac OSX (10.11 and above)	64 Bit	(388.44 MB)

Start KNIME Analytics Platform

- Use the shortcut created by the installer

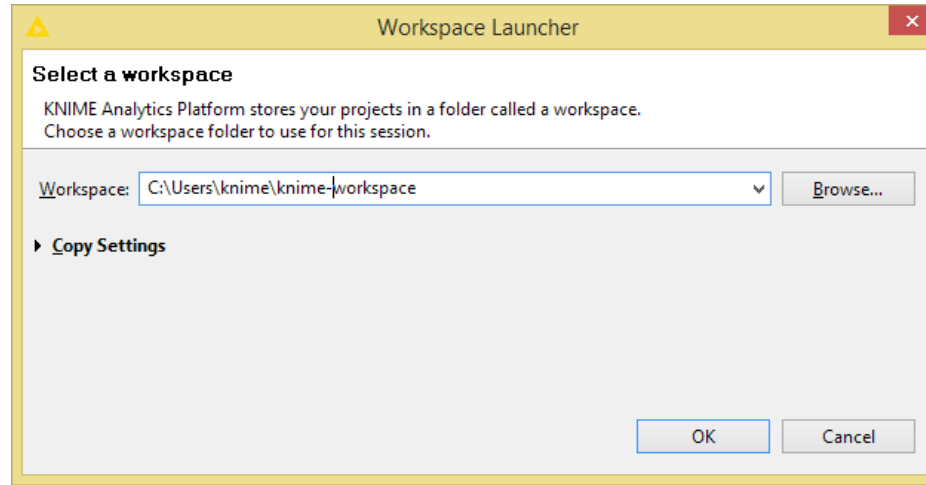


- Or go to the installation directory and launch **KNIME** via the `knime.exe`



The KNIME Workspace

- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session.
- Workspaces are portable (just like KNIME)



Welcome Page



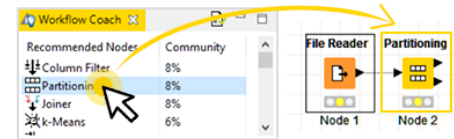
Welcome to KNIME Analytics Platform!

New to KNIME? Looking for resources to get started?

- Register for emails with introductory tips [here](#).
- Explore our [Quickstart Guide](#).
- Check out [7 things to do after installing KNIME Analytics Platform](#)
- Find more hints and how-tos in the [Learning Hub](#).
- And register for our release and event emails right [here](#).

This page will be displayed upon startup but you can customize the content using the checkboxes at the bottom.

NEW since 3.2: Workflow Coach recommends matching nodes.



Updates for the following components are available:

- DYMATRIX Uplift Modeling Extensions
- Palladian for KNIME

Click [here](#) in order to install updates.

Where to go from here

- [Create new workflow](#)
- [Learning Hub](#)
- [Browse example workflows](#)
- [Get additional nodes](#)
- [Go to my workflows](#)
- [Mount KNIME Cloud Server](#)

Most recently used workflows

- ModelSelection_WebPortal_Part1
- ModelSelection_WebPortal_Part1
- ModelSelection_BasicWorkflow
- DataCleaning_WebPortal_v2.0
- KNIME_project2
- Sexy ETL_v2.0

Tips & Tricks

Specialist Nodes

Did you know there are a whole variety of specialist nodes available from KNIME Labs and the Community around Scripting, Image Processing, Text Processing, Internet Mining, Network Mining, Cell Biology and Genetics, and Chemistry. To access them, go to Help Menu and choose Install New<- Software.

- Show intro text at next start
- Show update notifications at next start
- Show links and most recently used workflows at next start

The KNIME Workbench

The screenshot displays the KNIME Analytics Platform interface with several key components highlighted by yellow callout boxes:

- KNIME Explorer:** Located in the top-left corner, it shows a hierarchical tree view of the workspace. A callout points to the '04. Data Mining - solution' folder.
- Workflow Editor:** The central workspace where workflows are built. A callout points to the main workflow area.
- Node Recommendations:** A panel on the left side of the workflow editor that suggests relevant nodes. A callout points to this panel.
- Node Repository:** A panel at the bottom-left showing a categorized list of available nodes. A callout points to this panel.
- Node Description:** A panel on the right side that provides detailed information about the selected node. A callout points to the 'Partitioning' node description.
- Console:** A panel at the bottom-right showing system logs and messages. A callout points to the console output.
- Outline:** A small panel at the bottom-left of the workflow editor showing a simplified view of the workflow structure. A callout points to this panel.

The workflow shown in the editor includes nodes for 'Fully Joined Data', 'Partitioning', 'Decision Tree Learner', 'Decision Tree Predictor', 'Scorer', 'JavaScript ROC Curve', and 'Node 278 ROC Curve'. The 'Node Description' panel for 'Partitioning' contains the following text:

Partitioning

The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports. The following options are available in the dialog:

Dialog Options

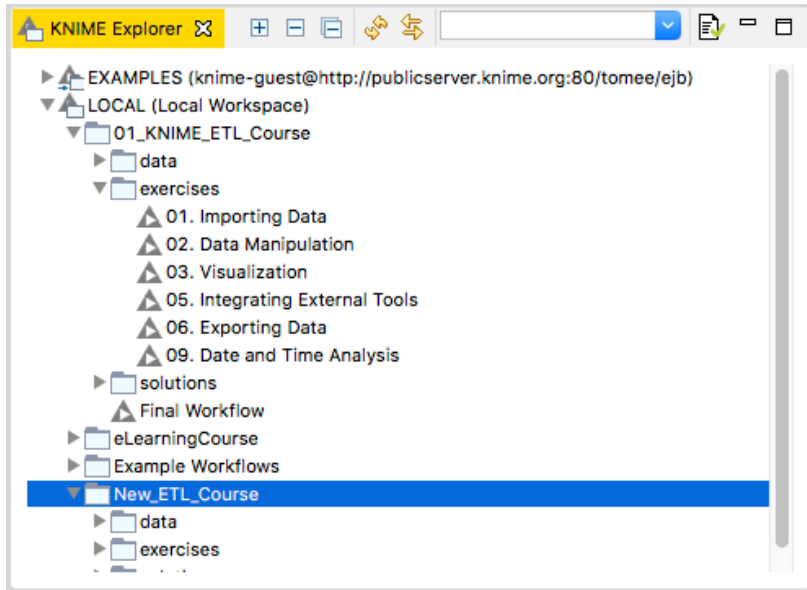
Absolute
Specify the absolute number of rows in the first partition. If there are less rows than specified here, all rows are entered into the first table, while the second table contains no rows.



Relative
The percentage of the number of rows in the input table that are in the first partition. It must be between 0 and 100, inclusively.

From top
This mode puts the top-most rows into the first output table and the remainder in the second table.

Linear sampling

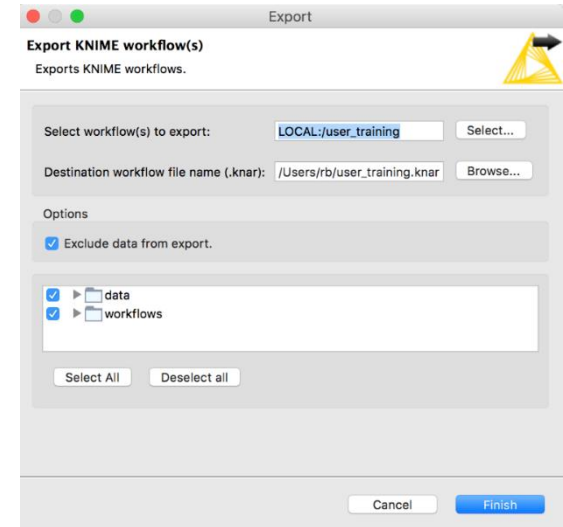
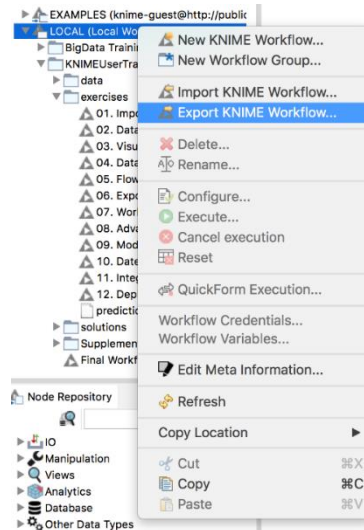
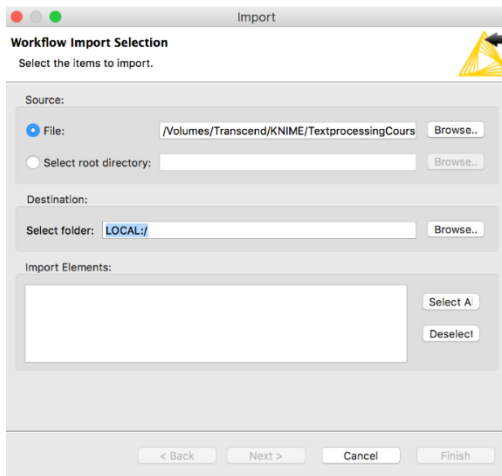
KNIME Explorer



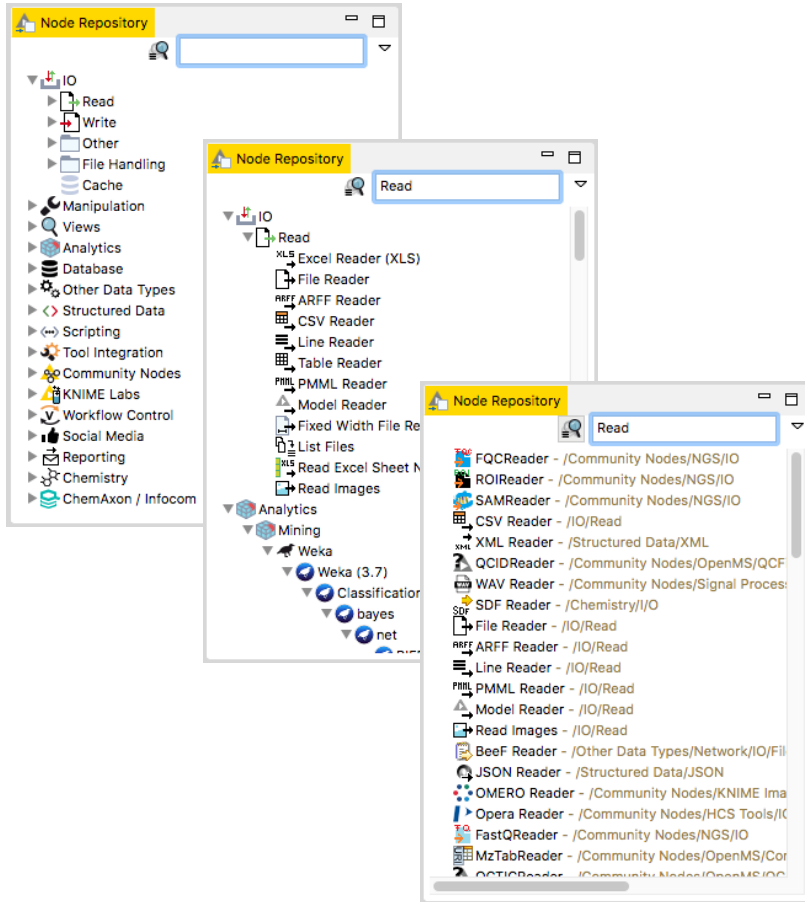
- In LOCAL you can access your own workflow projects.
- The Explorer toolbar on the top has a search box and buttons to
 -  select the workflow displayed in the active editor
 -  refresh the view
- The KNIME Explorer can contain 4 types of content:
 - Workflows
 - Workflow groups
 - Data files
 - Metanode templates

Creating New Workflows, Importing and Exporting

- Right-click in KNIME Explorer to create new workflow or workflow group or to import workflow
- Right-click on workflow or workflow group to export

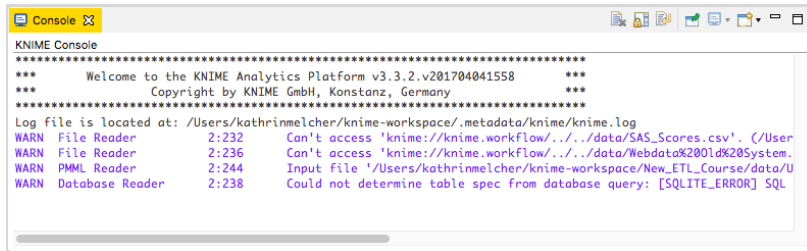


Node Repository



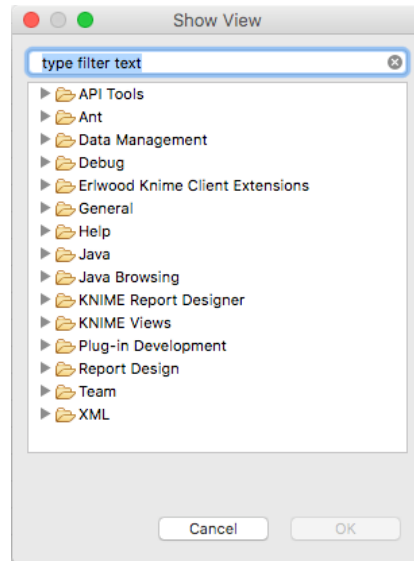
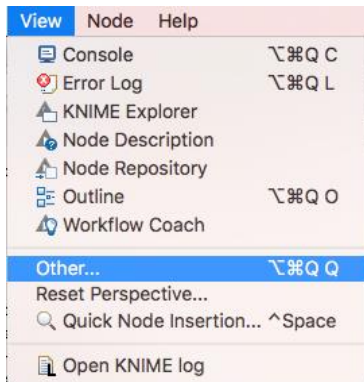
- The Node Repository lists all KNIME nodes
- The search box has 2 modes
 - 🔍 **Standard Search** – exact match of node name
 - 🔍 **Fuzzy Search** – finds the most similar node name
- Nodes can be added by drag and drop from the Node Repository to the Workflow Editor.

Console and Other Views

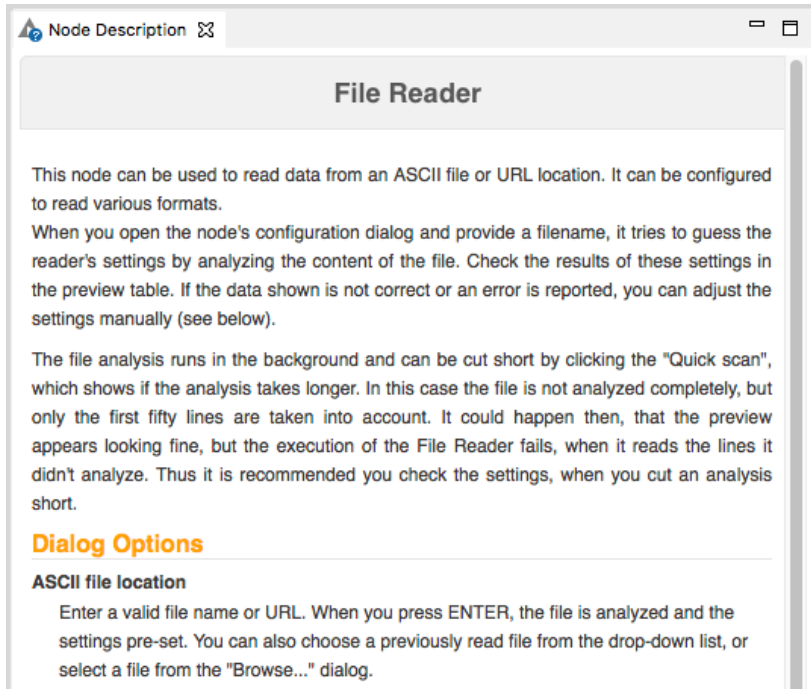


```
KNIME Console
*****
*** Welcome to the KNIME Analytics Platform v3.3.2.v201704041558 ***
*** Copyright by KNIME GmbH, Konstanz, Germany ***
*****
Log file is located at: /Users/kathrinmelcher/knime-workspace/.metadata/knime/knime.log
WARN File Reader 2:232 Can't access 'knime://knime.workflow/../../data/SAS_Scores.csv': (/User
WARN File Reader 2:236 Can't access 'knime://knime.workflow/../../data/Webdata%201d%20System.
WARN PMML Reader 2:244 Input file '/Users/kathrinmelcher/knime-workspace/New_ETL_Course/data/U
WARN Database Reader 2:238 Could not determine table spec from database query: [SQLITE_ERROR] SQL
```

- Console view prints out error and warning messages about what is going on under the hood.
- Click on View and select Other... to add different views
 - Node Monitor, Licenses, etc.



Node Description



The screenshot shows a window titled "Node Description" with a sub-header "File Reader". The main content area contains the following text:

This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Dialog Options

ASCII file location

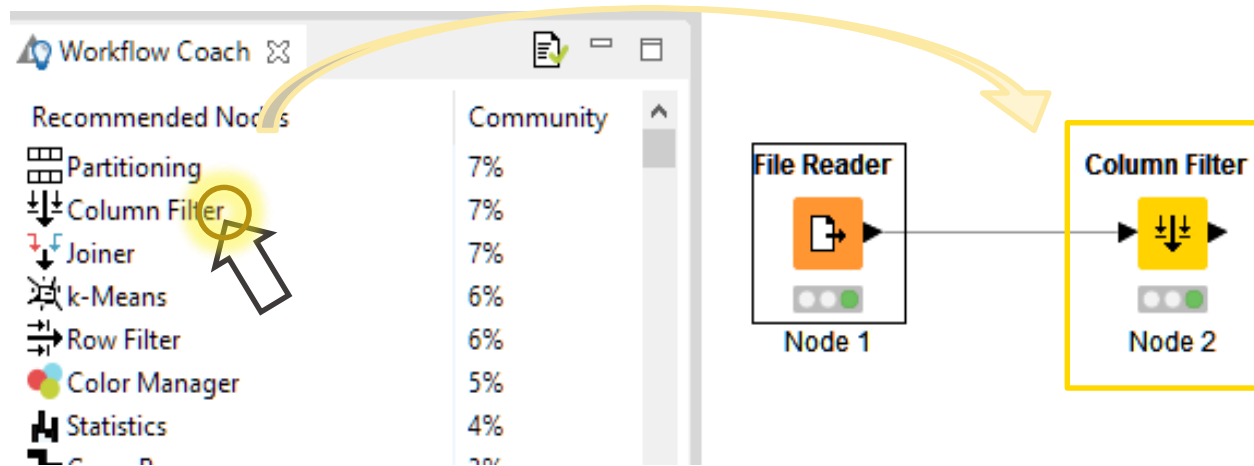
Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set. You can also choose a previously read file from the drop-down list, or select a file from the "Browse..." dialog.

- The Node Description window gives information about:
 - Node Functionality
 - Input & Output
 - Node Settings
 - Ports
 - References to literature

Workflow Coach

Recommendation engine






- Gives hints about which node use next in the workflow
- Based on KNIME communities' usage statistics
- Based on own KNIME workflows



Tool Bar



The buttons in the toolbar can be used for the active workflow. The most important buttons:

-  Execute selected and executable nodes (F7)
-  Execute all executable nodes
-  Execute selected nodes and open first view
-  Cancel all selected, running nodes (F9)
-  Cancel all running nodes

KNIME File Extensions

- Dedicated file extensions for Workflows and Workflow groups associated with KNIME Analytics Platform

- ***.knwf** for KNIME Workflow Files



- ***.knar** for KNIME Archive Files



More on Nodes...

A node can have 3 states:

File Reader



Not Configured:

The node is waiting for configuration or incoming data.

File Reader



Configured:

The node has been configured correctly, and can be executed.

File Reader

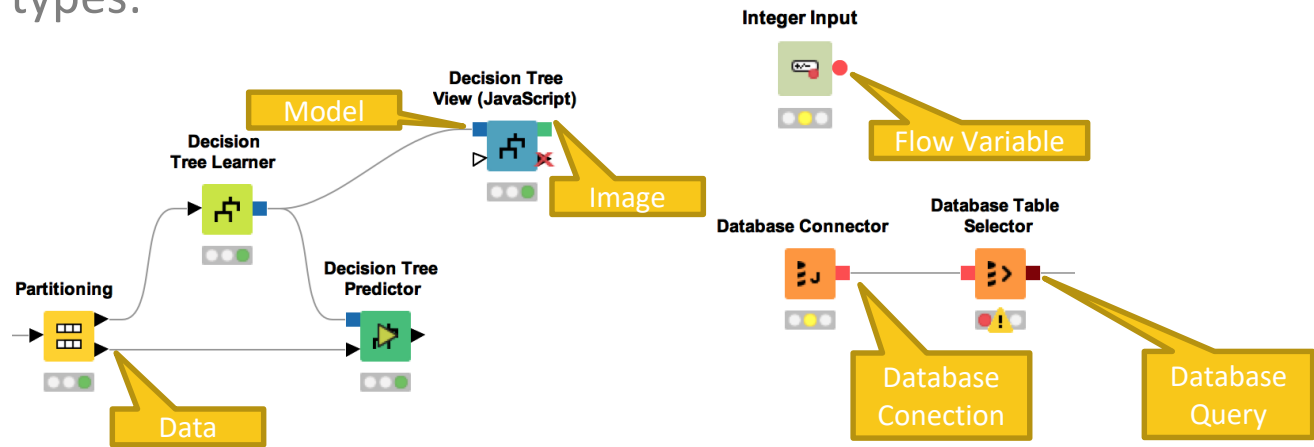


Executed:

The node has been successfully executed. Results may be viewed and used in downstream nodes.

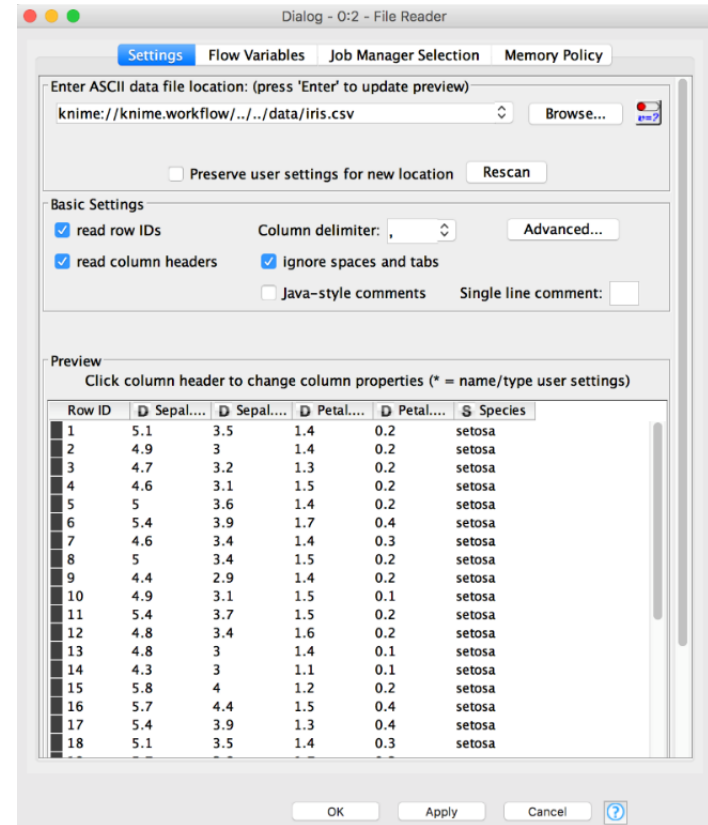
Inserting and Connecting Nodes

- Insert nodes into workspace by dragging them from Node Repository or by double-clicking in Node Repository
- Connect nodes by left-clicking output port of Node A and dragging the cursor to (matching) input port of Node B
- Common port types:



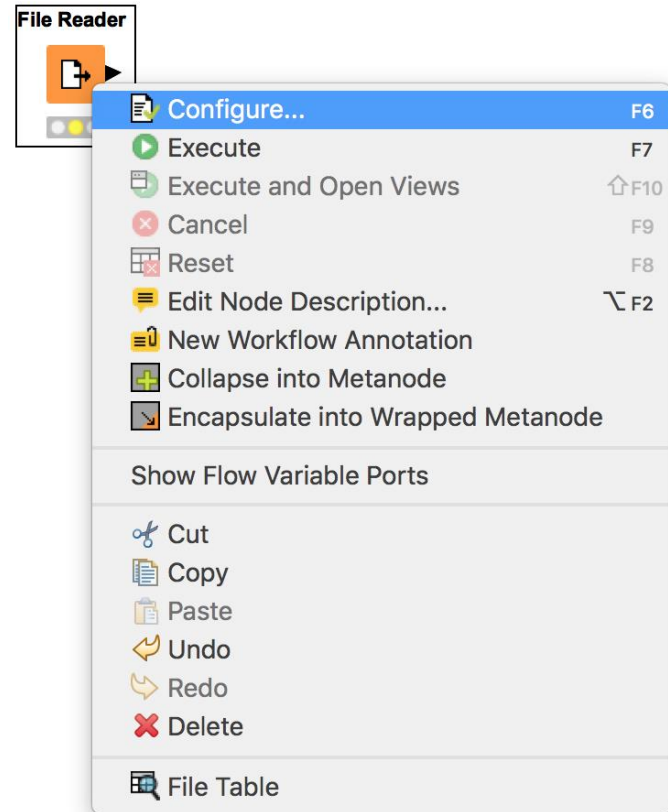
Node Configuration

- Most nodes require configuration
- To access a node configuration window:
 - Double-click the node
 - Right-click > Configure



Node Execution

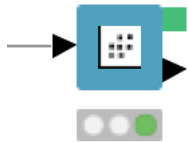
- Right-click node
- Select Execute in context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



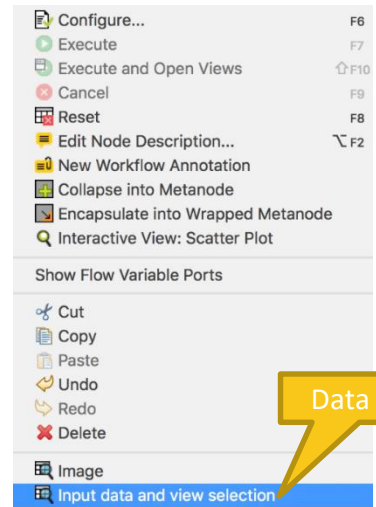
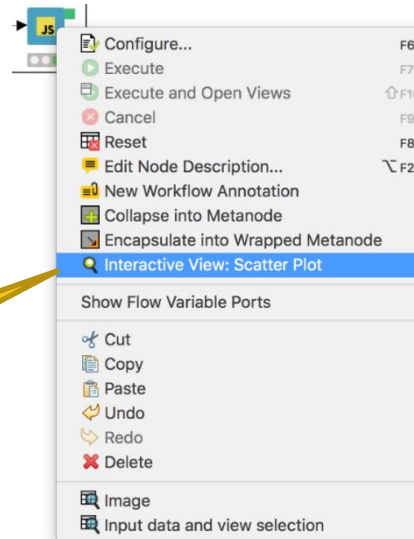
Node Views

- Right-click node
- Select Views in context menu
- Select output port to inspect execution results

Scatter Plot (JavaScript)



Plot View



Data View

Curved Connections!

The image illustrates the process of enabling curved connections in the KNIME Workflow Editor. It is divided into three main sections:

- Top:** A toolbar with a yellow arrow pointing to the 'Curved Connections' icon (a red circle with a white arrow).
- Middle:** A 'Workflow Editor Settings' dialog box. The 'Node Connections' section has the 'Curved connections' checkbox checked. Other settings include 'Enable Grid' (checked), 'Snap to grid' (checked), 'Grid Size' (horizontal spacing: 20, vertical spacing: 20), and 'Connection line width: 1'. A yellow arrow points from the 'Curved connections' checkbox to the workflow diagram below.
- Bottom:** Two workflow diagrams. The left diagram shows a workflow with straight connections between nodes: Table Reader (Customer data) -> Value Selection -> Row Filter -> GroupBy (by Products) -> Sorter (desc by count) -> Table Row to Variable (first row) -> Row Filter (filter by Products). A large yellow arrow points from this diagram to the right diagram. The right diagram shows the same workflow but with curved connections between the nodes.

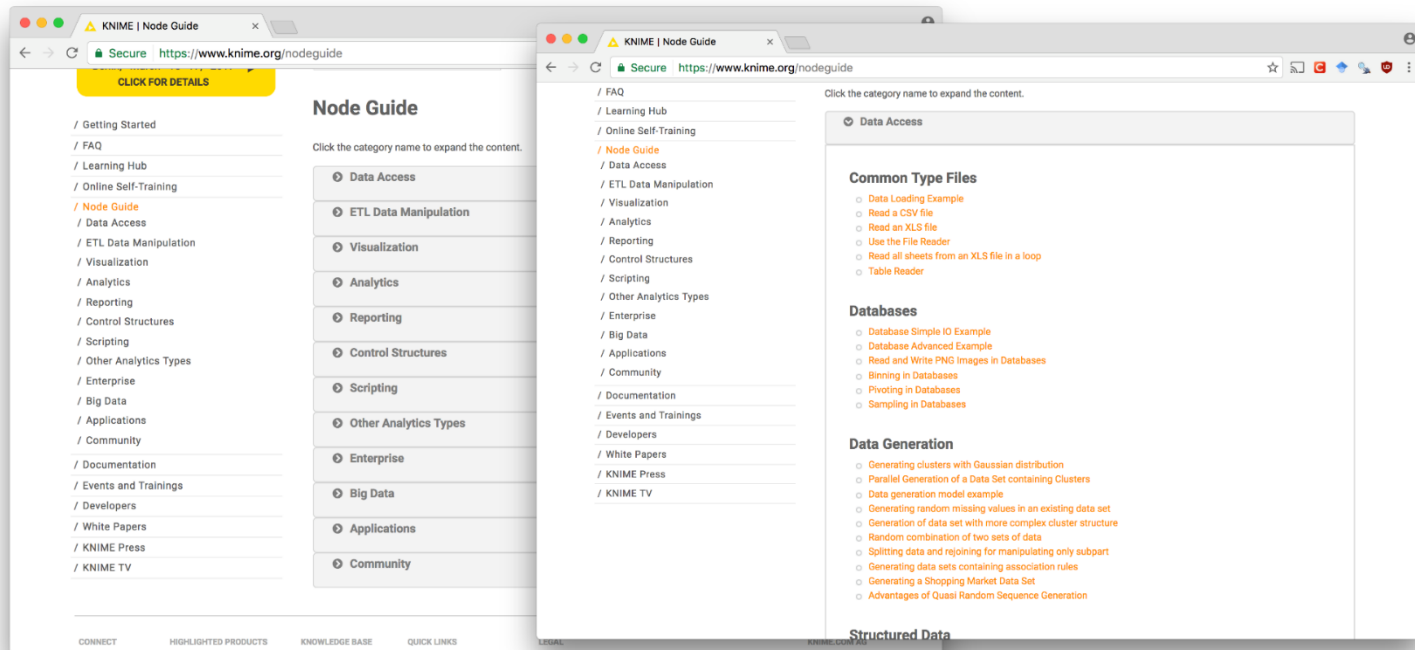
Getting Started: KNIME Example Server

- Public repository with large selection of example workflows for many, many applications
- Connect via KNIME Explorer

The image shows a screenshot of the KNIME Explorer application. On the left, a file tree is visible under the 'EXAMPLES' folder, with '02_Database_Advanced_Example' selected. The main area displays a workflow diagram titled 'Database - Advanced Usage'. The workflow consists of the following nodes: 'Metanode: Setup SQLite database in temporary folder', 'Database Table Connector', 'Database Row Filter', 'Database Column Filter', 'Database Connection Table Reader', 'Database Query', 'Database Connection Table Reader', and 'Database Writer'. A yellow box highlights the text: 'This workflow shows a number of database nodes that directly work inside a database. It's using the file based "sqlite" database (the entire database is written to a file on the hard disk.)'. A 'Login' dialog box is open in the top right corner, and a warning message is displayed at the top of the workflow area.

Online Node Guide

- Workflows from Example Server also available online
 - <https://www.knime.com/nodeguide>



Hot Keys (for future reference)

Task	Hot key	Description
Node Configuration	F6	opens the configuration window of the selected node
Node Execution	F7	executes selected configured nodes
	Shift + F7	executes all configured nodes
	Shift + F10	executes all configured nodes and opens all views
	F9	Cancels selected running nodes
	Shift + F9	Cancels all running nodes
Move Nodes and Annotations	Ctrl + Shift + Arrow	moves the selected node in the arrow direction
	Ctrl + Shift + PgUp/PgDown	moves the selected annotation in the front or in the back of all overlapping annotations
Workflow Operations	F8	resets selected nodes
	Ctrl + S	saves the workflow
	Ctrl + Shift + S	saves all open workflows
	Ctrl + Shift + W	closes all open workflows
Meta-node	Shift + F12	opens meta-node wizard

Additional Resources

KNIME pages (<https://www.knime.com>)

- **SOLUTIONS** for example workflows
- RESOURCES/**LEARNING HUB** <https://www.knime.com/learning-hub>
- RESOURCES/**NODE GUIDE** <https://www.knime.com/nodeguide>
- Book **WILL THEY BLEND** <https://www.knime.com/knimepress/will-they-blend>

KNIME Tech pages

- **FORUM** for questions and answers <https://forum.knime.com>
- **DOCUMENTATION** for docs, FAQ, changelogs, ...
- **COMMUNITY CONTRIBUTIONS** for dev instructions and third party nodes

KNIME TV on YouTube <https://www.youtube.com/user/KNIMETV>

Today's Example: Next Best Offer (NBO)

- Traditional Direct Marketing advertises a single product to a specific audience. The Next Best Offer (NBO) approach focuses on taking existing customers (and their data) and using upsell models to find interesting new products for them.
- Today we construct a workflow that joins diverse data sources into a set of complete customer records. Using this, we will build and deploy a predictive model to find people who might be interested in a newly available product.

The data

numeric Nominal Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞	No. -∞	Histogram
CustomerKey	<input type="checkbox"/>	11000	29483	20241.500	5336.016	28473061.667	0	-1.200	374143886	0	0	0	0	0	
EstimatedYearlyIncome	<input type="checkbox"/>	10000	170000	57305.778	32285.842	1042375574.469	0.822	0.646	1059240000	0	0	0	0	0	
SentimentRating	<input type="checkbox"/>	0	5	1.844	1.612	2.600	0.473	-0.947	34091	5165	0	0	0	0	
WebActivity	<input type="checkbox"/>	0	5	1.004	1.523	2.318	1.394	0.688	18559	11116	0	0	0	0	
NumberOfContracts	<input type="checkbox"/>	0	4	1.503	1.138	1.296	0.402	-0.434	27776	4238	0	0	0	0	
Age	<input type="checkbox"/>	29	100	48.232	11.261	126.812	0.569	-0.112	891521	0	0	0	0	0	
Target	<input type="checkbox"/>	0	1	0.494	0.500	0.250	0.024	-2.000	9132	9352	0	0	0	0	
Available401K	<input type="checkbox"/>	0	1	0.676	0.468	0.219	-0.754	-1.432	12501	5982	1	0	0	0	
CustomerValueSegment	<input type="checkbox"/>	1	3	2.103	0.694	0.481	-0.141	-0.926	38880	0	0	0	0	0	
ChurnScore	<input type="checkbox"/>	0	1	0.274	0.334	0.112	1.216	0.193	5071.000	6310	0	0	0	0	
CallActivity	<input type="checkbox"/>	1	5	3.215	1.262	1.592	-0.298	-0.928	59424	0	1	0	0	0	

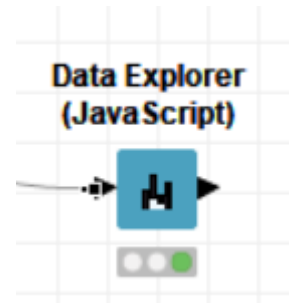
The data

Numeric Nominal **Data Preview**

Search:

Column	Exclude Column	No. missings	Unique values	All nominal values	Histogram
MaritalStatus	<input type="checkbox"/>	0	2	M, S	
Gender	<input type="checkbox"/>	0	2	M, F	

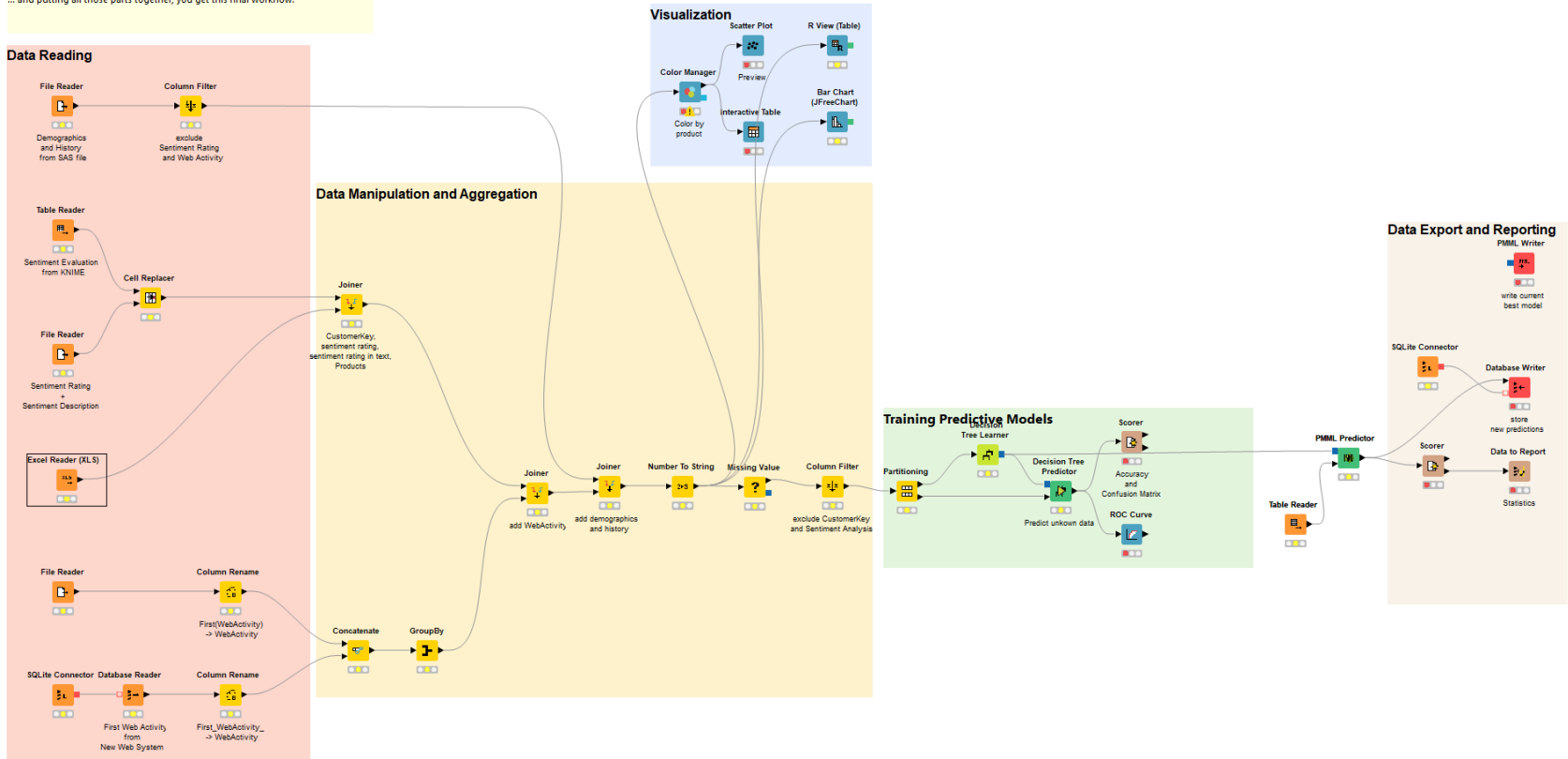
Showing 1 to 2 of 2 entries



Today's Example: Next Best Offer (NBO)

Final Workflow from the KNIME User Training

... and putting all those parts together, you get this final workflow.



Importing Data

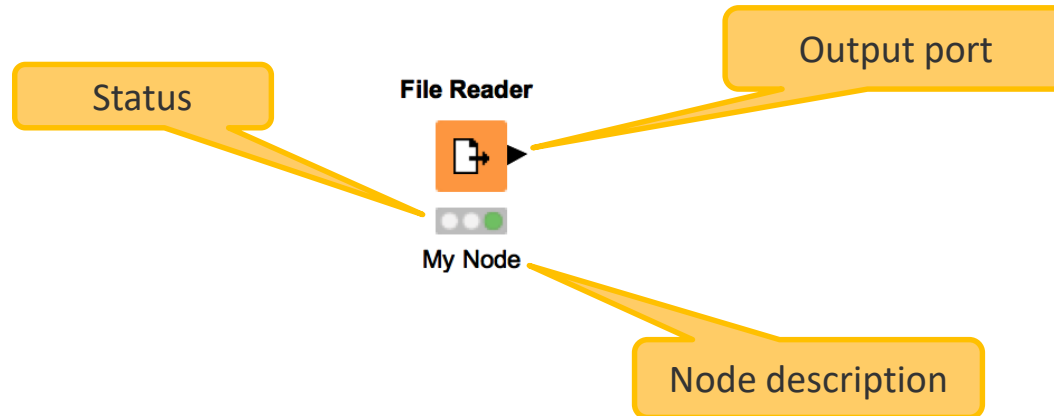
Accessing files and databases



Data Source Nodes

Typically characterized by:

- Orange color
- No input ports, 1-2 output ports



New Node: File Reader

Workhorse of the KNIME Source nodes

- Reads all text based files (e.g. csv, txt, etc.)
- Many advanced features allow it to read most 'weird' files
 - Short lines, inline comments, headers and special encoding

File Reader



My Node

YouTube KNIME TV Channel video:

<https://youtu.be/flaHQw-Qhlg>

File Reader Configuration

The screenshot shows the 'File Reader (My Node)' configuration dialog. It has four tabs: 'Settings', 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Settings' tab is active. At the top, there is a text field for 'Enter ASCII data file location' with the value 'knime://knime.workflow/../../data/iris.csv' and a 'Browse...' button. Below this is a checkbox for 'Preserve user settings for new location' and a 'Rescan' button. The 'Basic Settings' section contains several options: 'read row IDs' (checked), 'read column headers' (checked), 'Column delimiter' (set to comma), 'ignore spaces and tabs' (checked), 'Java-style comments' (unchecked), and 'Single line comment' (empty). An 'Advanced...' button is also present. The 'Preview' section shows a table with 18 rows of data. At the bottom are 'OK', 'Apply', and 'Cancel' buttons, along with a 'Help Button' (question mark icon).

File path

Basic Settings

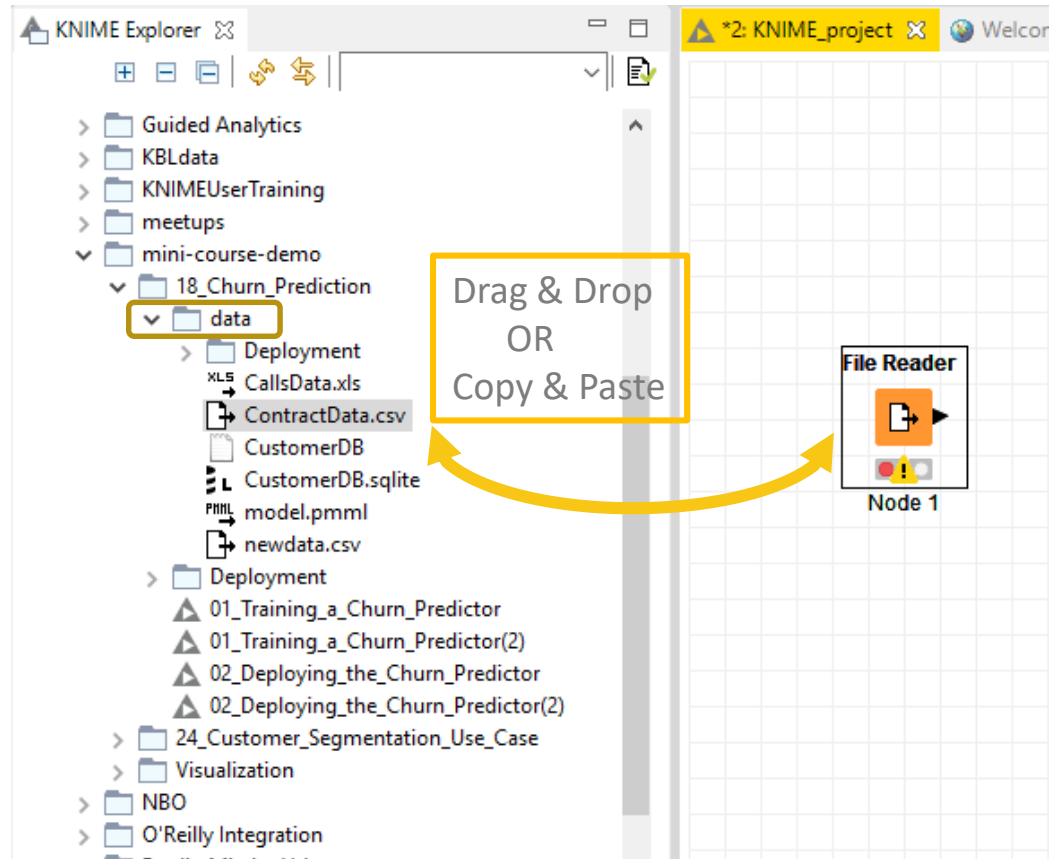
Advanced Settings

Preview

Help Button

Row ID	D Sepal....	D Sepal....	D Petal....	D Petal....	S Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa


Alternative Faster Way ...



Filenames and the knime:// protocol


Absolute URL

Input location

Browse...

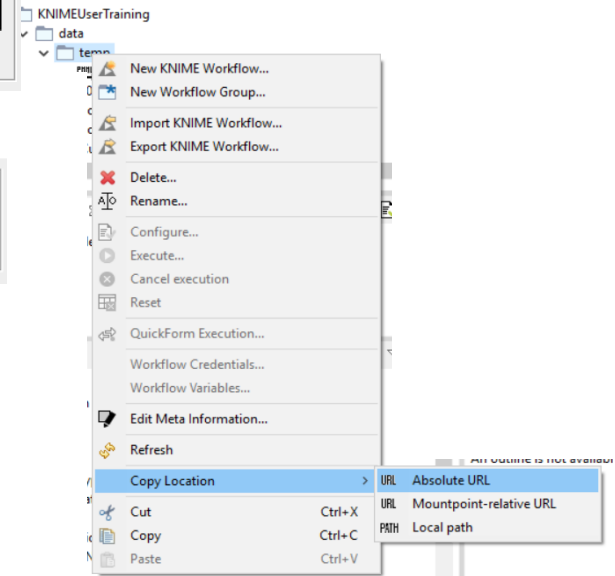

Mountpoint-relative URL

Input location

Browse...

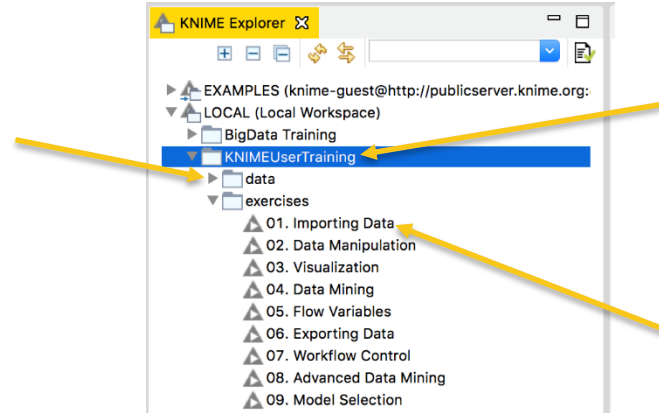
Local path

Input location

Browse...

Workflow Relative File Paths

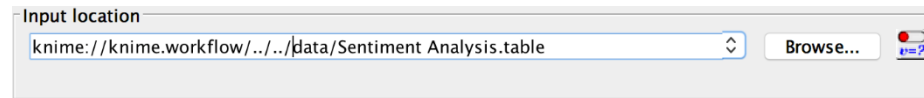
- Best choice if workflows are to be shared
- Requires matching folder structure within workflow group
 - Independent of environment outside of workflow group



Example: Path to „Sentiment Analysis.table“

- Local path:
C:\Users\rb\knime-workspace\KNIMEUserTraining\data\Sentiment Analysis.table

- Workflow relative:



YouTube KNIME TV Channel:

<https://youtu.be/U9sP4g4yGwY>

New Node: Excel Reader (XLS)

- Reads .xls and .xlsx file from Microsoft Excel
 - Supports reading from multiple sheets

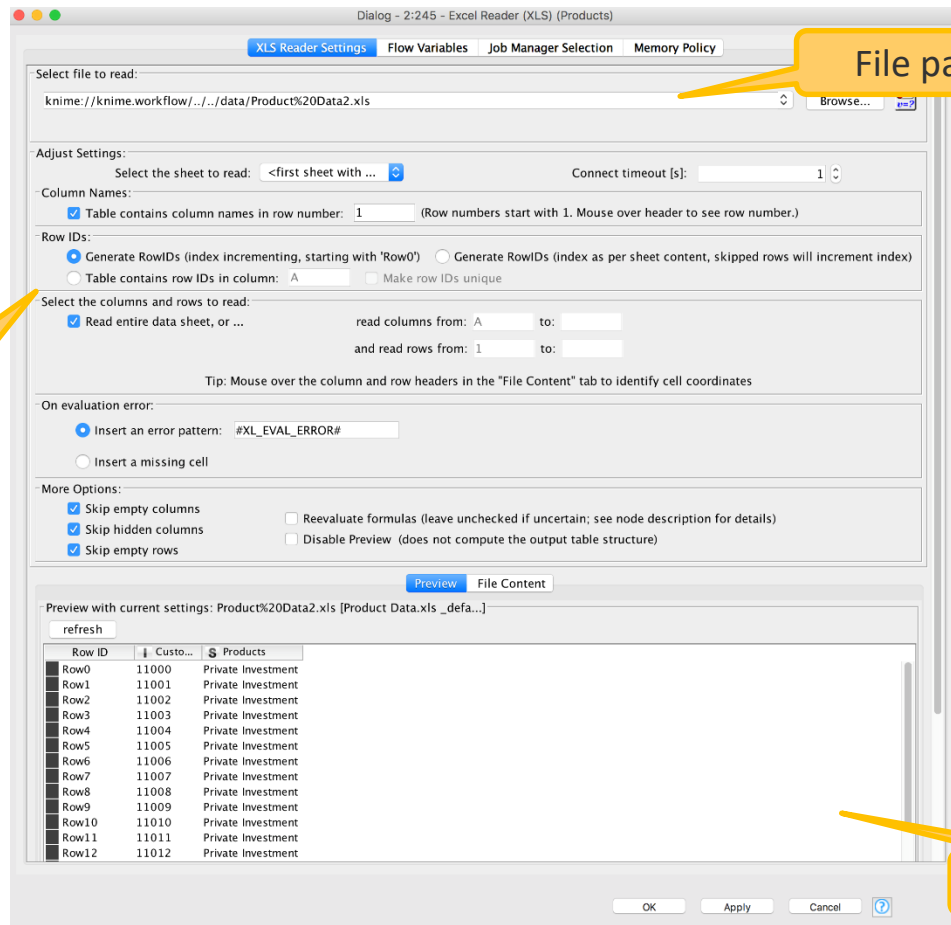
Excel Reader (XLS)



Read Excel Sheet Names (XLS)



Excel Reader Configuration



File path

Sheet specific settings

Preview

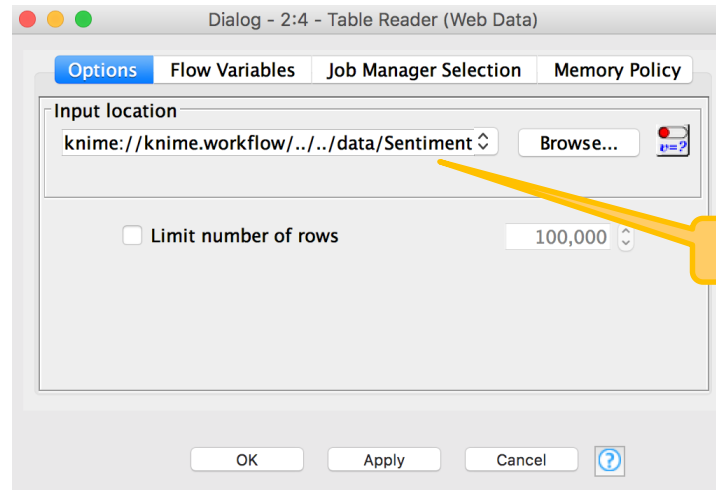
New Node: Table Reader

- Reads tables from the native KNIME Format.
 - Maximum performance, minimum configuration

Table Reader



Web Data

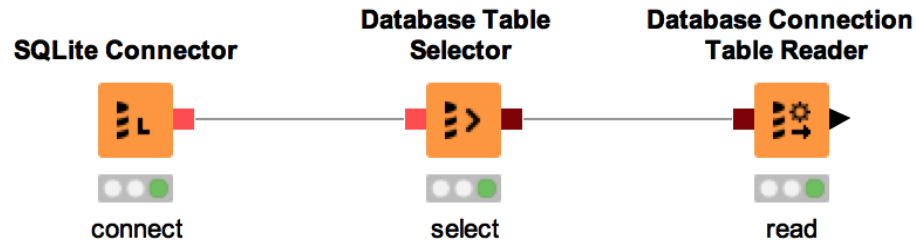


YouTube KNIME TV channel video:

<https://youtu.be/tid1qi2HAOo>

Database Connectivity

- Read data from any JDBC enabled database
- Write your own SQL or model it using dedicated nodes

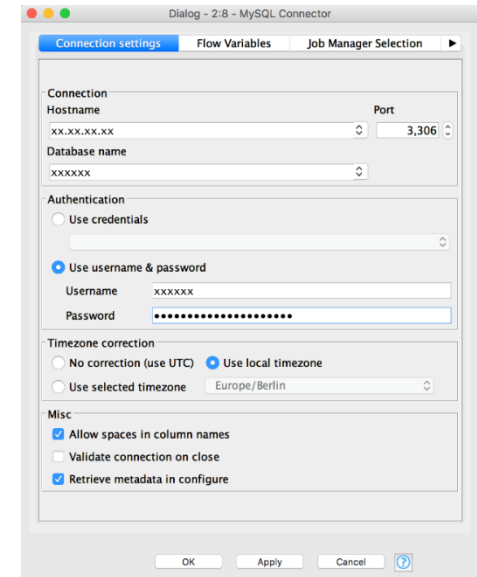


New Nodes: Database Connectors

- Native: Postgres, MySQL, MS SQL Server, SQLite
- Database Connector (e.g. Oracle, DB2, HANA).
- Big Data: HIVE and Impala

- ▼ Connector
 - A Amazon Athena Connector
 - R Amazon Redshift Connector
 - J Database Connector
 - H H2 Connector
 - H Hive Connector
 - I Impala Connector
 - S Microsoft SQL Server Connector
 - M MySQL Connector
 - P PostgreSQL Connector
 - L SQLite Connector
 - V Vertica Connector

MySQL Connector



Other Useful Data Sources

- PMML Reader – reads standard predictive models
- XML Reader with XPATH support
- Python/R Source nodes
- Tika Parser – extracts textual data from 200+ file types
- REST Web Services, and many more

XML Reader



CSV Reader



PMML Reader



R Source (Table)



Importing Data Exercise

Start with exercise: *Importing Data*

Read the following files

- Sentiment Analysis.table
- Sentiment Rating.csv
- Product Data2.xls

Optional: Read table *web_activity* from the database *WebActivity.sqlite*

(hint: drag and drop the files from the KNIME Explorer panel to get started)

Table Reader



Sentiment Evaluation
from KNIME

File Reader



Sentiment Rating
+
Sentiment Description

Excel Reader (XLS)



Products
<->
Customer

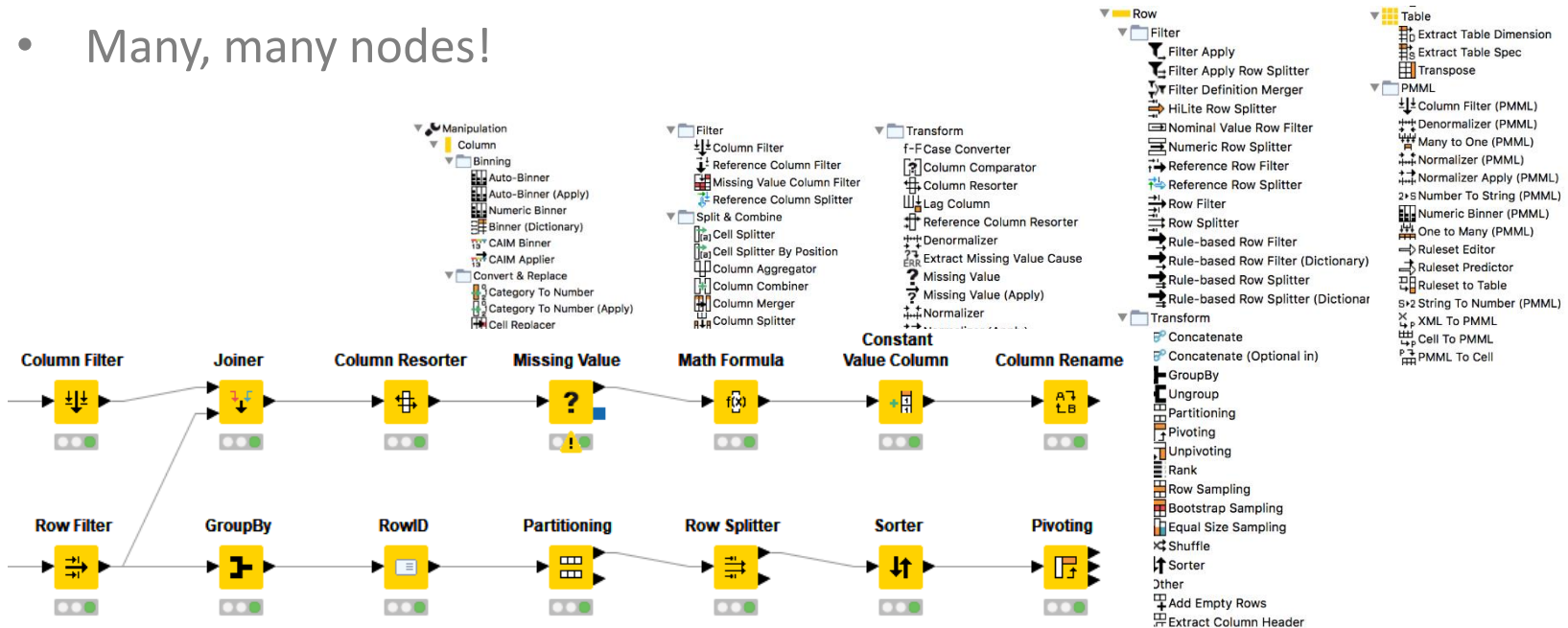
Data Manipulation

Clean, join, aggregate



Data Manipulation Nodes

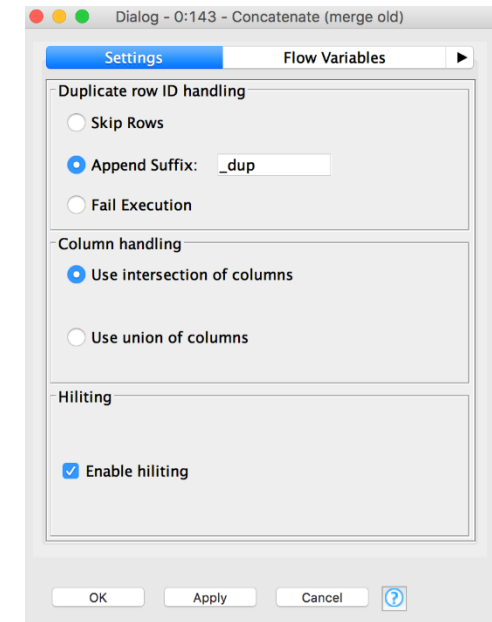
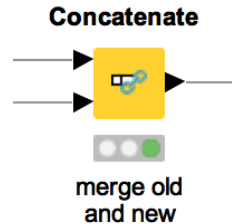
- Yellow color with a variety of input and output ports
- Apply a transformation to input data
- Many, many nodes!



New Node: Concatenate

Combine rows from 2 tables with shared columns

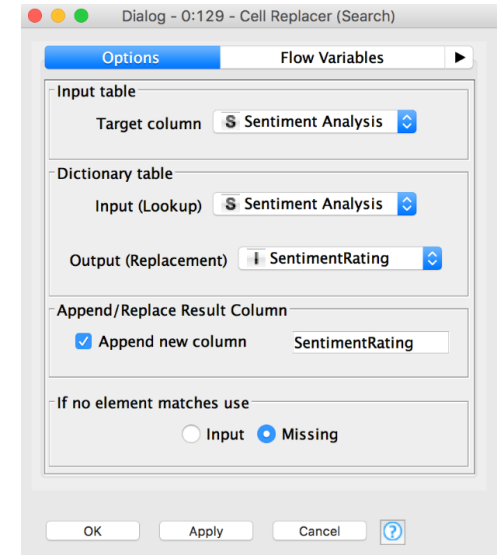
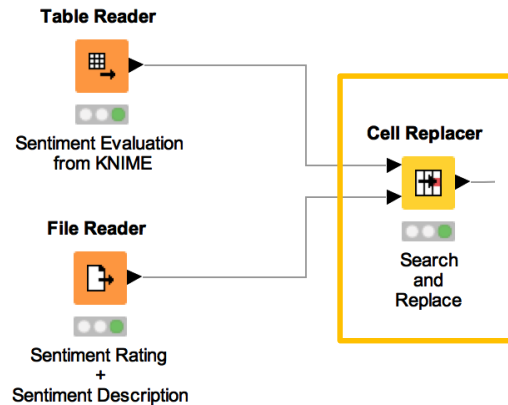
- Handles duplicate row keys gracefully
- Take the union or intersection of columns



New Node: Cell Replacer

Replaces the content of a column based on a lookup

- Top port references the table to be searched
- Bottom port holds the lookup table (search keys and replacement values)

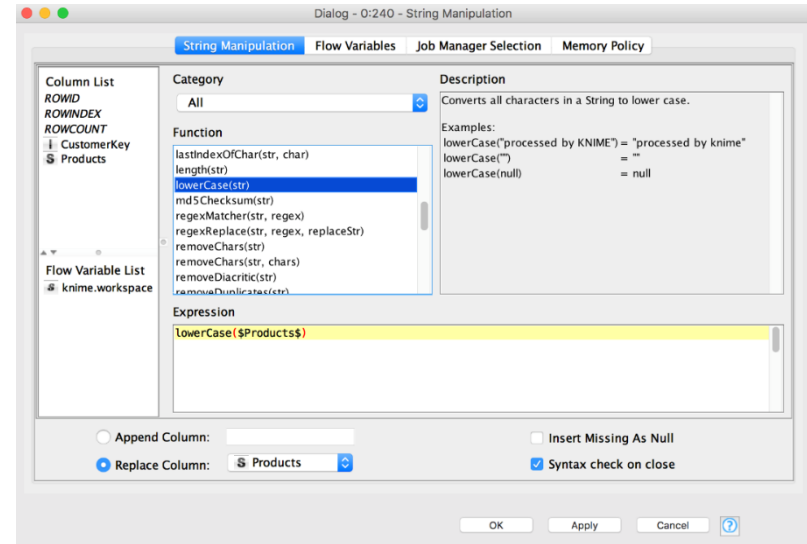
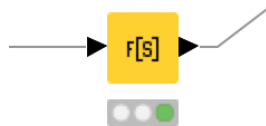


New Node: String Manipulation

Create and edit values in String columns

- Clean up capitalization (eg. Lowercase)
- Replace strings
- Modify existing strings or create new columns

String Manipulation



Data Manipulation Exercise, Activity I

Start with exercise: *Data Manipulation, Activity I*

- Concatenate web activity data from old and new systems
- Replace sentiment evaluation (strings) with corresponding numeric values
- Use String Manipulation to ensure that all entries of the Products column are lower case from the product data spreadsheet.

Joining Columns of Data

Left Table

Manually created table - 3:49 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 4

Row ID	ID	age	income	class
Row0	1	23	<=50K	F1
Row1	3	25	<=50K	F3
Row2	6	22	>50K	A4
Row3	8	21	<=50K	C3

Join by ID

Right Table

Manually created table - 3:50 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 6 Spec - Columns: 4

Row ID	ID	age	income	sex
Row0	1	23	<=50K	M
Row1	2	25	<=50K	F
Row2	4	23	>50K	M
Row3	5	21	<=50K	F
Row4	6	25	>50K	M
Row5	7	24	<=50K	M

Inner Join

Joined table - 3:51 - Joiner

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 7

Row ID	ID	age	income	class	age (#1)	income...	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M

Left Outer Join

Joined table - 3:51 - Joiner

File Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 7

Row ID	ID	age	income	class	age (#1)	income...	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M
Row1_?	3	25	<=50K	F3	?	?	?
Row3_?	8	21	<=50K	C3	?	?	?

Missing values in the right table.

Missing values in the left table.

Right Outer Join

Joined table - 3:51 - Joiner

File Hilite Navigation View

Table "default" - Rows: 6 Spec - Columns: 7

Row ID	ID	age	income	class	age (#1)	income...	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M
?_Row1	?	?	?	?	?	?	?
?_Row2	?	?	?	?	?	?	?
?_Row3	?	?	?	?	?	?	?
?_Row5	?	?	?	?	?	?	?

Joining Columns of Data

Left Table

Manually created table - 3:49 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 4

Row ID	ID	age	income	class
Row0	1	23	<=50K	F1
Row1	3	25	<=50K	F3
Row2	6	22	>50K	A4
Row3	8	21	<=50K	C3

Right Table

Manually created table - 3:50 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 6 Spec - Columns: 4

Row ID	ID	age	income	sex
Row0	1	23	<=50K	M
Row1	2	25	<=50K	F
Row2	4	23	>50K	M
Row3	5	21	<=50K	F
Row4	6	25	>50K	M
Row5	7	24	<=50K	M

Join by ID

Full Outer Join

Joined table - 3:51 - Joiner

File Hilite Navigation View

Table "default" - Rows: 8 Spec - Columns: 7 Properties Flow Variables

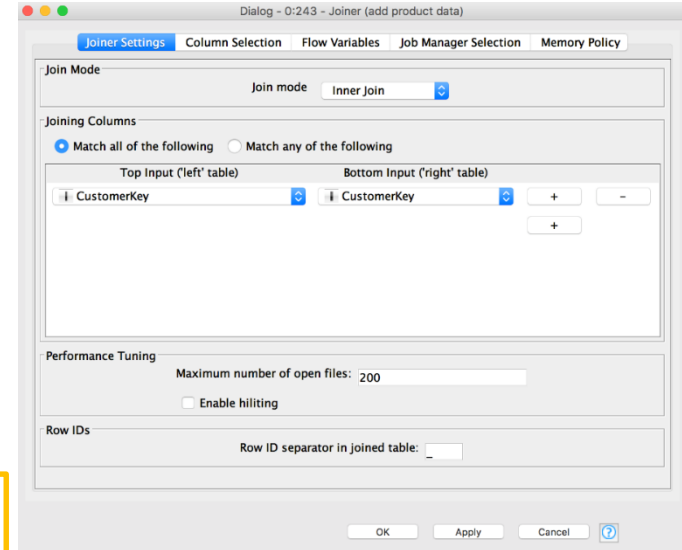
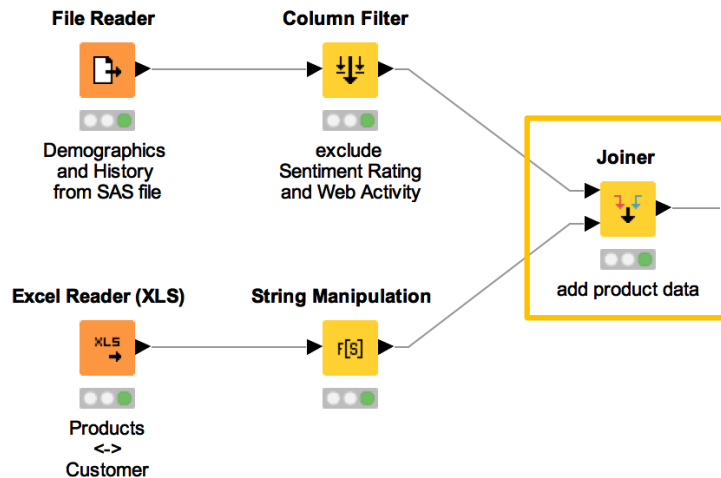
Row ID	ID	age	income	class	age (#1)	income...	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M
Row1_?	3	25	<=50K	F3	?	?	?
Row3_?	8	21	<=50K	C3	?	?	?
?_Row1	?	?	?	?	25	<=50K	F
?_Row2	?	?	?	?	23	>50K	M
?_Row3	?	?	?	?	21	<=50K	F
?_Row5	?	?	?	?	24	<=50K	M

Missing values in the left table.

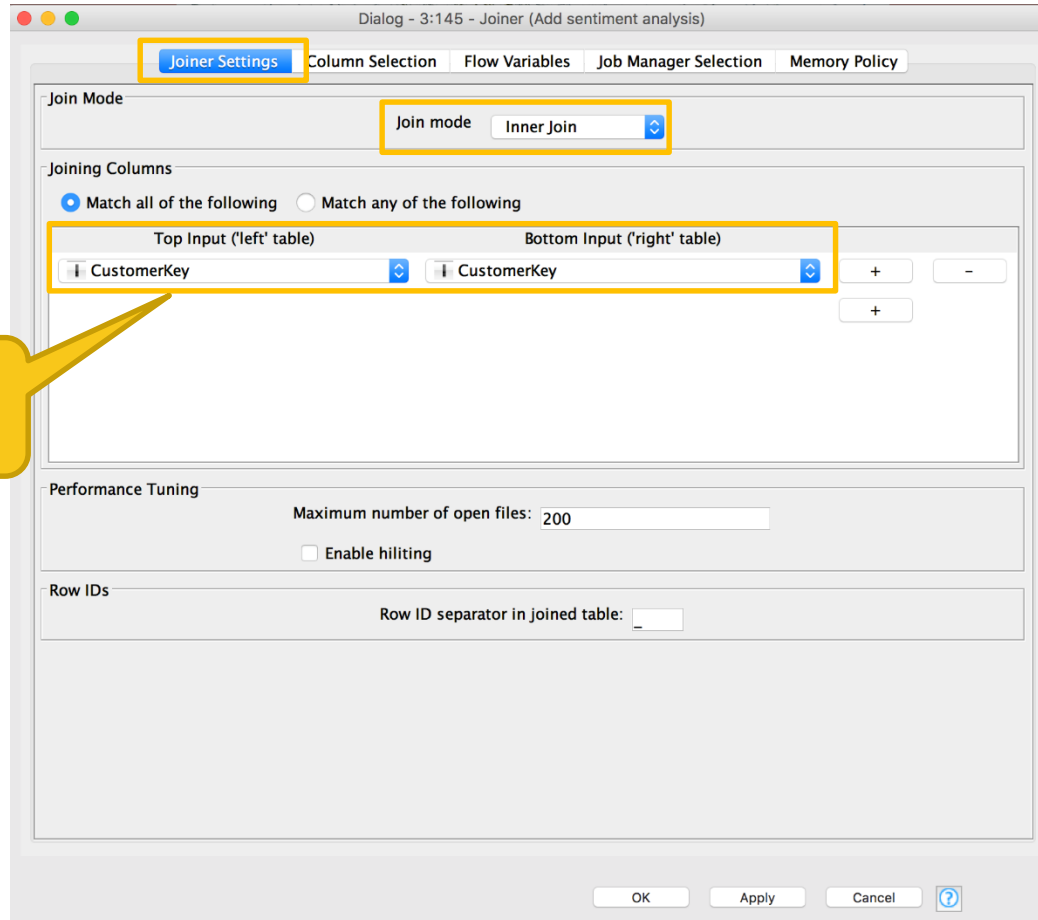
Missing values in the right table.

New Node: Joiner

- Combines columns from 2 different tables
- Top port contains “Left” data table
- Bottom port contains “Right” data table

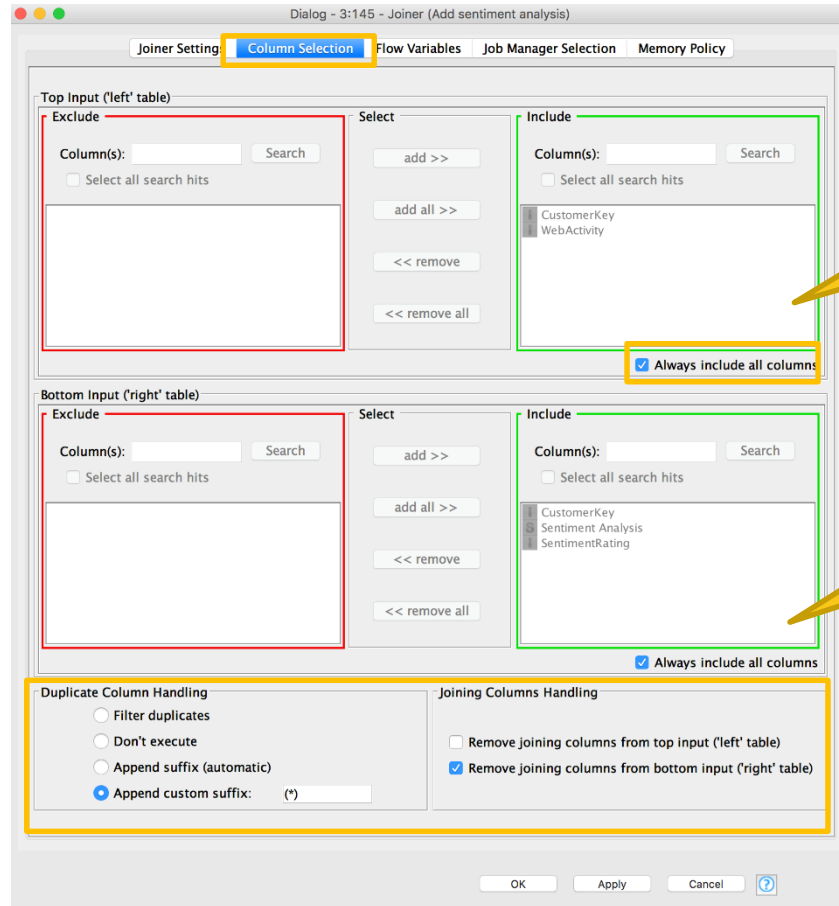


Joiner Configuration – Linking Rows



Values to join on.
Multiple joining
columns are allowed.

Joiner Configuration – Column Selection

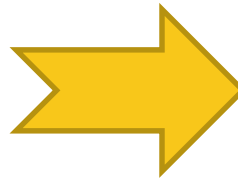


Columns from left table to output table

Columns from right table to output table

Data Aggregation

RowID	Group	Value
R1	M	2
R2	F	3
R3	M	1
R4	F	5
R5	F	7
R6	M	5



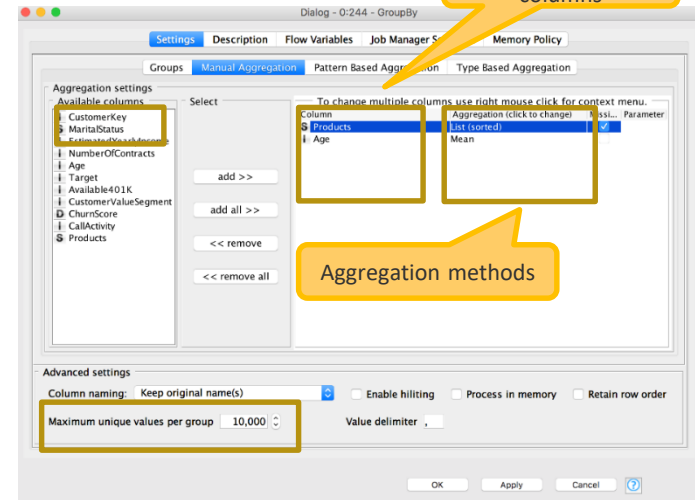
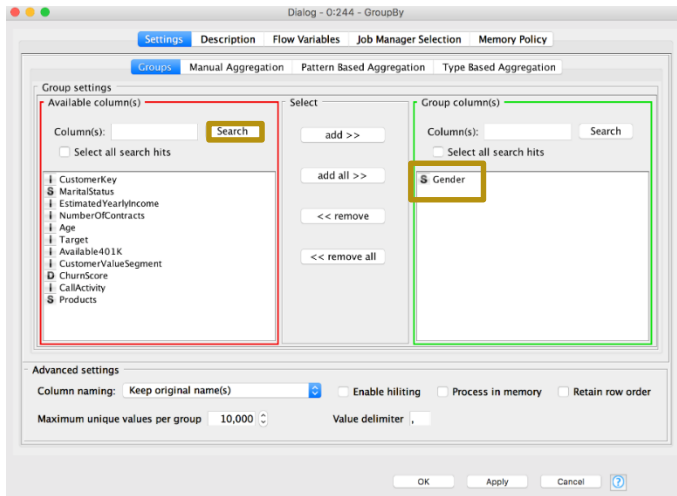
RowId	Group	Sum(value)
R1+R3+R6	M	8
R2+R4+R5	F	15

Aggregated on “group” by method:
sum(“value”)

New Node: GroupBy

Aggregate to remove duplicates or summarize data

- First tab provides grouping options
- Second tab provides control over aggregation details



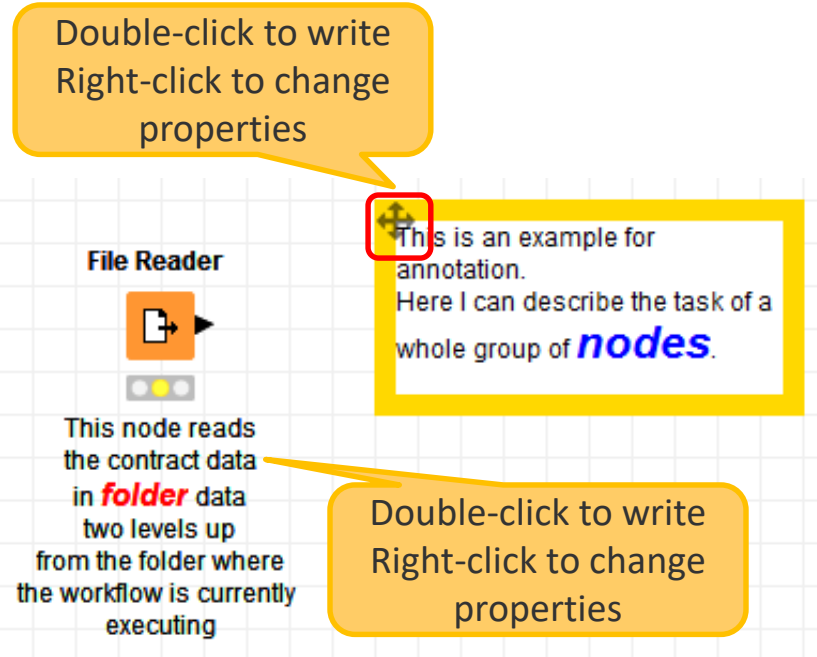
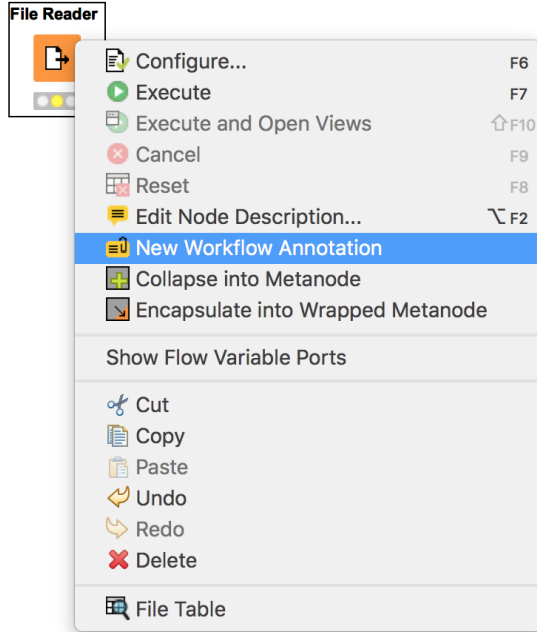
YouTube KNIME TV video:

<https://youtu.be/bDwF-TOMtWw>

Workflow organization and documentation



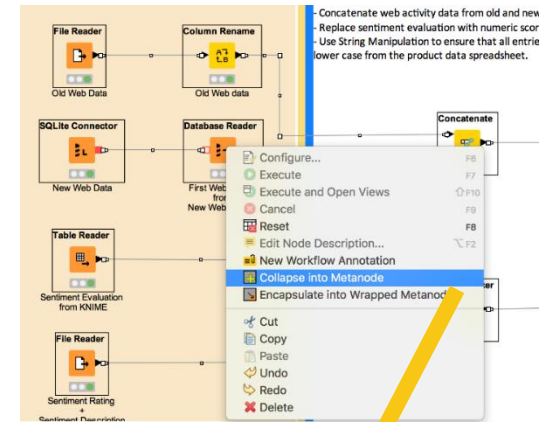
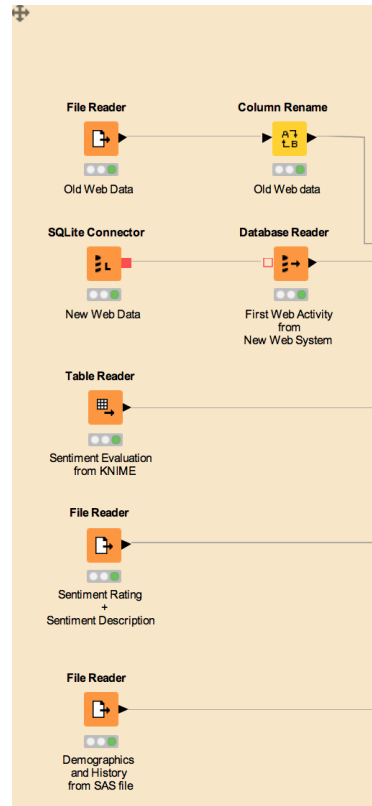
Comments & Annotations



YouTube KNIME TV Channel:
https://youtu.be/AHURYB_O8sA

Workflow Organisation – Good Practices

- Workflow annotations
- Node labels
- Metanodes
 - Right click -> Collapse...
 - Organize workflow by task
 - Hide complexity & improve readability



Fully Joined Data



KNIME WorkflowDiff

- Automates identification and comparison of nodes in a workflow, metanodes, and two different workflows
- Identifies insertions, deletions, substitutions, and parameter changes

WorkflowDiff interface showing a comparison of two Column Filter nodes. The 'Old' node has 'petal length', 'petal width', and 'class' in the included_names list. The 'New' node has 'sepal length', 'sepal width', and 'class' in the included_names list.

Column Filter 0:16			Column Filter 0:15		
Name	Type	Value	Name	Type	Value
Node Settings			Node Settings		
filter-filter	sub-config		filter-filter	sub-config	
filter-type	string	STANDARD	filter-type	string	STANDARD
included_names			included_names		
array-size	int	3	array-size	int	3
0	string	petal length	0	string	sepal length
1	string	petal width	1	string	sepal width
2	string	class	2	string	class
excluded_names			excluded_names		
enforce_option	string	EnforceExclusion	enforce_option	string	EnforceExclusion
name_pattern	sub-config		name_pattern	sub-config	
datatype	sub-config		datatype	sub-config	
System Node Settings			System Node Settings		

WorkflowDiff interface showing a comparison of two Snowball Stemmer nodes. The 'Old' node has 'Porter' as the Stemmer Name. The 'New' node has 'German' as the Stemmer Name.

Snowball Stemmer (34)			Snowball Stemmer (34)		
Name	Type	Value	Name	Type	Value
Node Settings			Node Settings		
Document Column Internals	sub-config		Document Column Internals	sub-config	
Document Column	string	Preprocessed Document	Document Column	string	Preprocessed Document
Preprocess Unmodifiable_Inte	sub-config		Preprocess Unmodifiable_Inte	sub-config	
Preprocess Unmodifiable	boolean	false	Preprocess Unmodifiable	boolean	false
Replace Document Internals	sub-config		Replace Document Internals	sub-config	
Replace Document	boolean	true	Replace Document	boolean	true
New Document Column Nam	sub-config		New Document Column Nam	sub-config	
New Document Column Nam	string	Preprocessed Document	New Document Column Nam	string	Preprocessed Document
Stemmer Name Internals	sub-config		Stemmer Name Internals	sub-config	
Stemmer Name	string	Porter	Stemmer Name	string	German
System Node Settings			System Node Settings		

Data Manipulation Exercise, Activity II

Start with exercise *Data Manipulation, Activity II*

- Join all data together using a series of joiner nodes and the “Customer Key” field
- Resolve duplicates in the joined dataset (hint: GroupBy node)
- Clean up and document your workflow using annotations, node labels, and metanodes

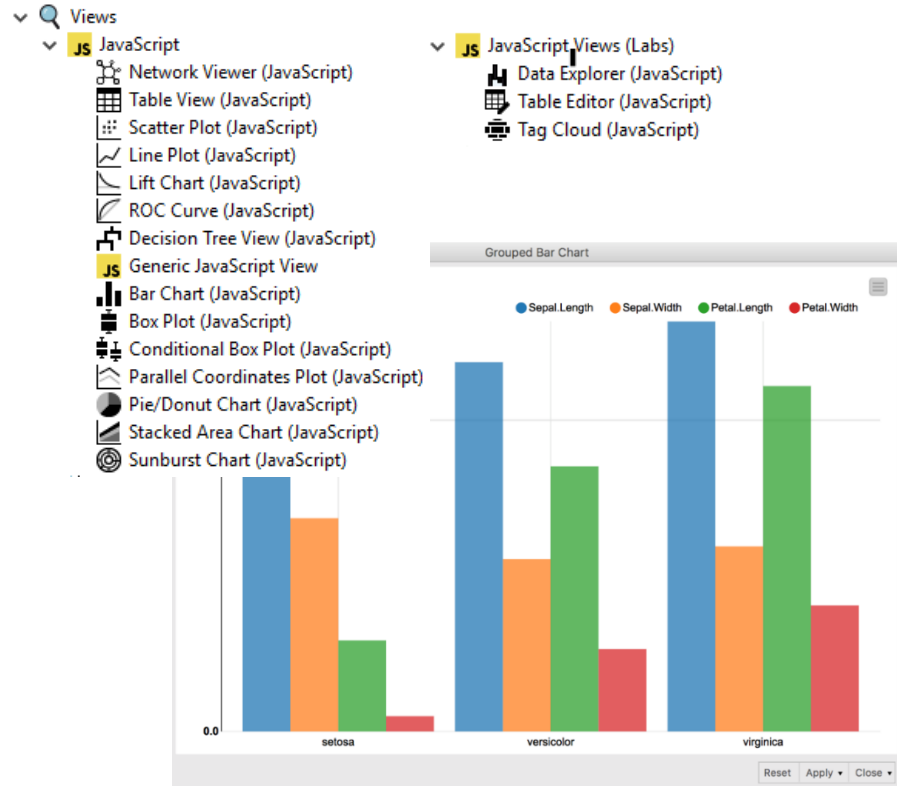
Data Visualization

Charts and tables



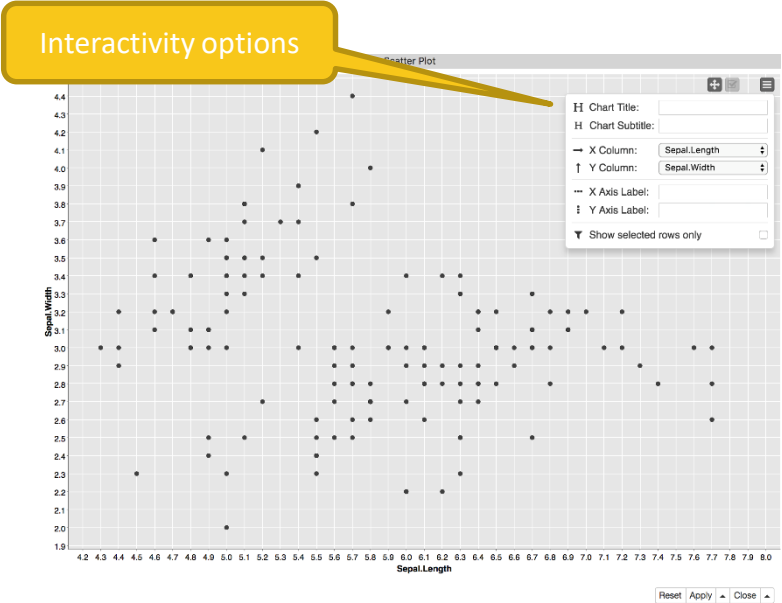
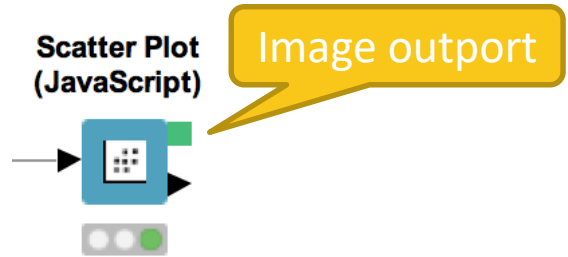
Data Visualization

- Large selection of easy to use visualization nodes
 - JavaScript-based, interactive plots and tables
 - Dedicated nodes, no scripting required
- R and Python View nodes for highly customizable graphics in KNIME
 - Require scripting



New Node: Scatter Plot (JavaScript)

- Plot different columns on X and Y
- Displays data including color information
- Produces an interactive view and an image
- Select data points and publish selection to other views



New Node: Scatter Plot (JavaScript)

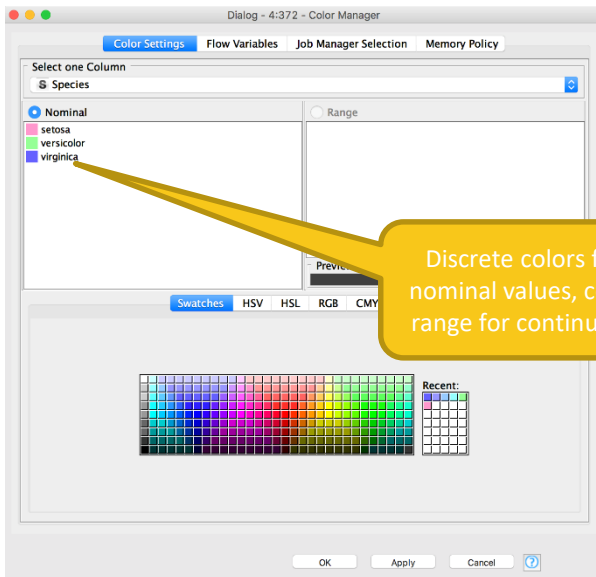
- 4 configuration tabs

The image displays four overlapping screenshots of the Scatter Plot node configuration interface, illustrating the four configuration tabs mentioned in the text:

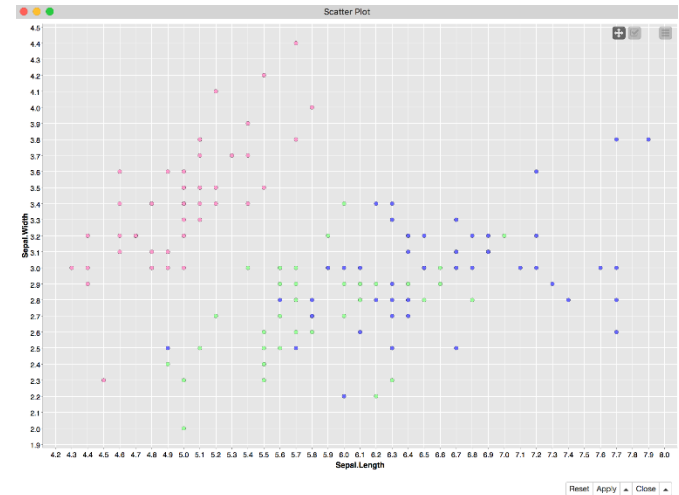
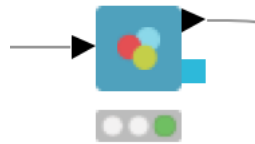
- Options:** Shows settings for output and data handling, including "Create image at output" (checked), "Maximum number of rows" (2,500), "Selection column name" (Selected (Scatter Plot)), "Choose column for x axis" (Sepal.Length), "Choose column for y axis" (Sepal.Width), "Report on missing values" (checked), and "Axes ranges" (Auto range axes checked).
- Axis Configuration:** Shows settings for labels and date/time formatting, including "Label for x axis", "Label for y axis", "Date and Time formatter" (Locale: English (United States)), "Local Date format" (YYYY-MM-DD), "Local Date&Time format" (YYYY-MM-DD), "Local Time format" (HH:mm:ss), "Zoned Date&Time format" (YYYY-MM-DD z), "Time zone (for zoned format)" (Europe/Berlin), "Date&Time (legacy) format" (YYYY-MM-DD), and "Axes ranges" (Auto range axes checked).
- General Plot Options:** Shows settings for titles, features, sizes, and colors, including "Titles" (Chart title, Chart subtitle), "Features" (Show color legend unchecked, Show grid checked), "Sizes" (Width of image (in px): 800, Height of image (in px): 600, Resize view to fill window checked, Display fullscreen button checked), and "Colors" (Background color, Data area color, Grid color).
- View Controls:** Shows settings for view edit controls, legend, and selection, including "View edit controls" (Enable view edit controls, Enable title edit controls, Enable subtitle edit controls, Enable column chooser for x-axis, Enable column chooser for y-axis, Enable label edit for x-axis, Enable label edit for y-axis), "Legend" (Enable legend display control), "able mouse crosshair" (Snap to data points), "nable rectangular selection" (Enable lasso selection), "subscribe to selection events" (Enable 'Show selected points only' option), "subscribe to filter events (no filters available)", "Enable panning" (checked), "ooming" (Enable drag zooming, Show zoom reset button).

New Node: Color Manager

- Color by nominal or continuous values
- Sync colors between views using the color model port and Color Appender node

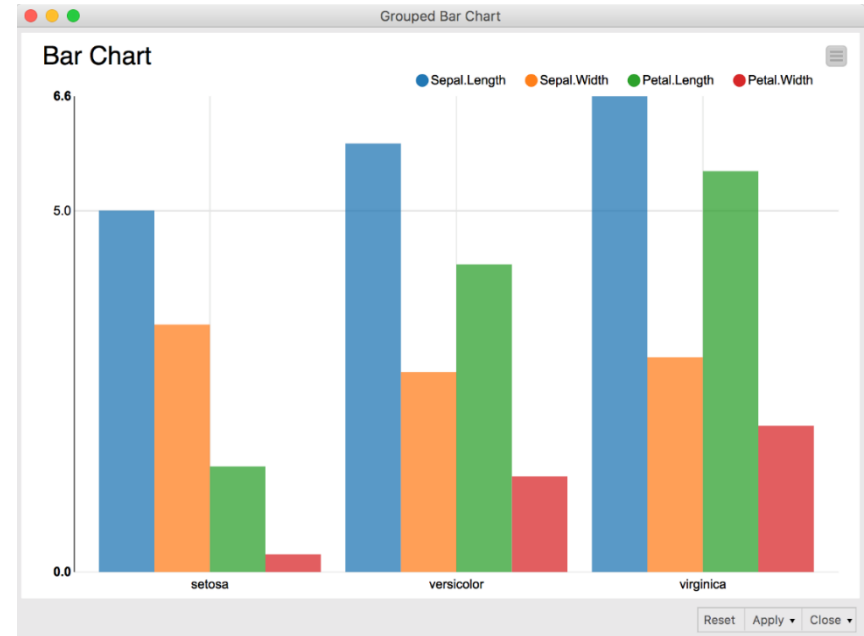


Color Manager



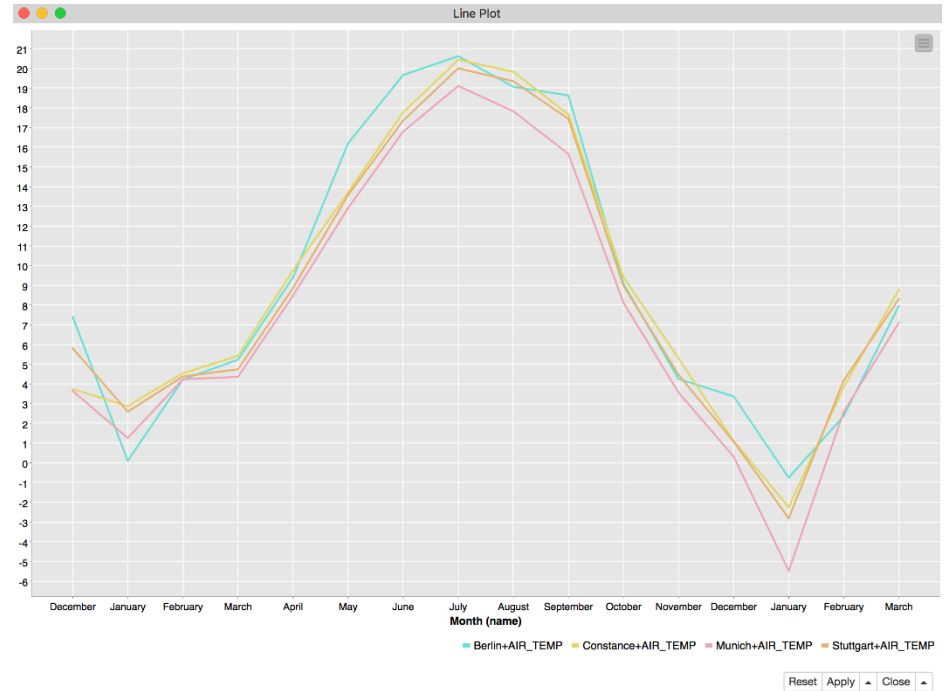
New Node: Bar Chart (JavaScript)

- Show numerical values across categories
- Vertical or horizontal bars
- Bars can be grouped or stacked



New Node: Line Plot (JavaScript)

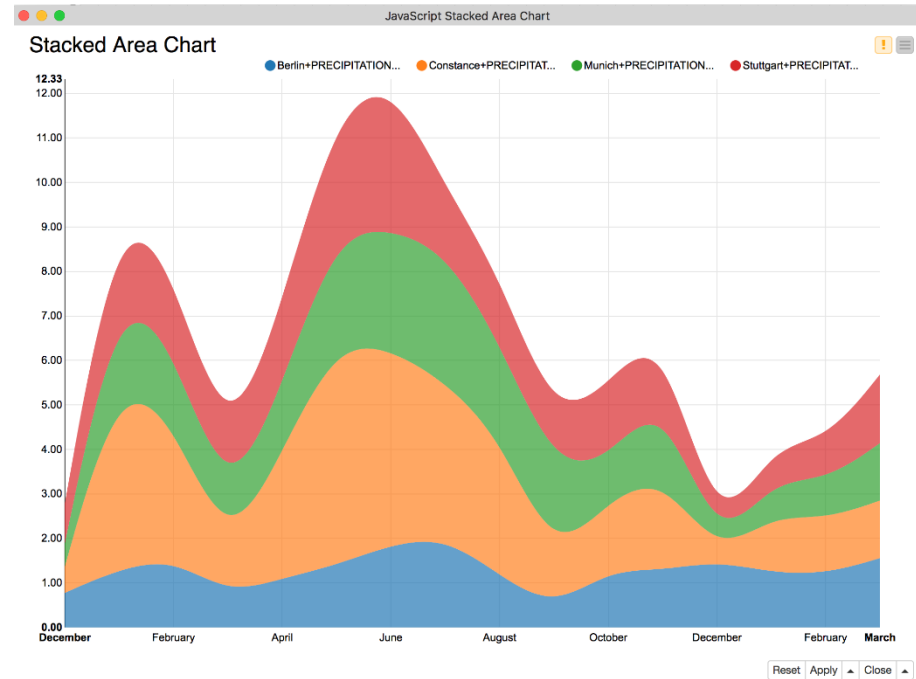
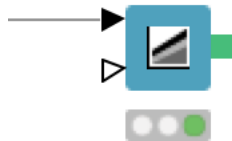
- Plot sequence of values, e.g. over time
- Useful to identify trends, also between groups



New Node: Stacked Area Chart (JavaScript)

- Visualizes numerical values from multiple columns as stacked areas
- Great for plotting distributions over time

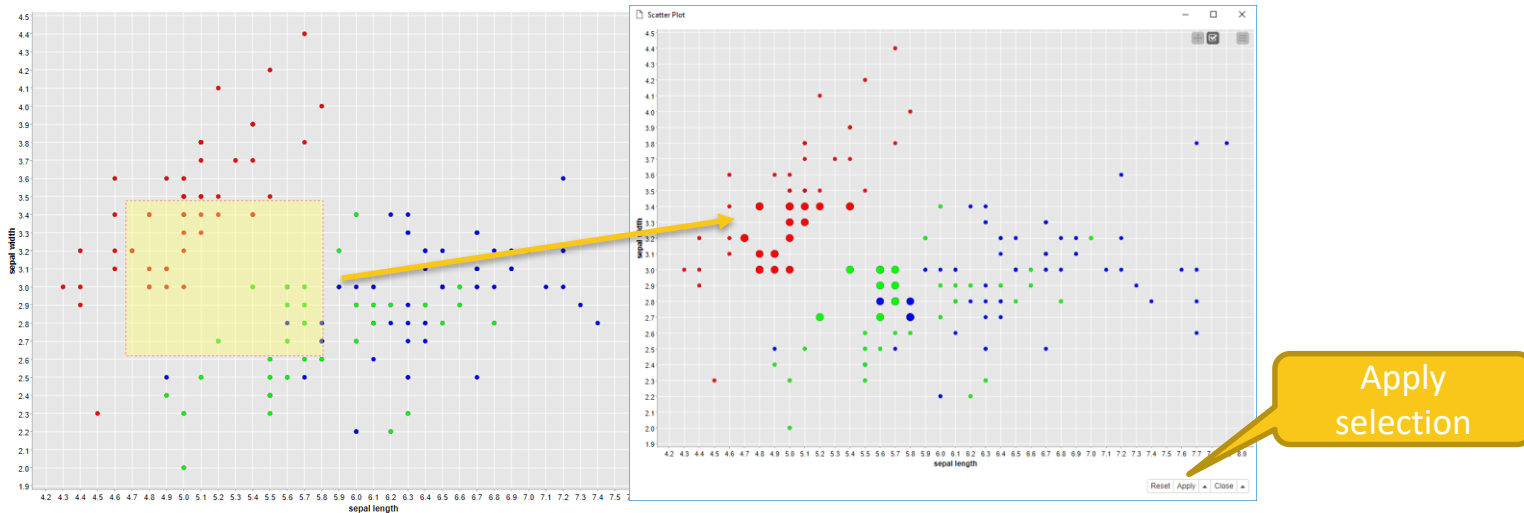
**Stacked Area Chart
(JavaScript)**



Selection & Filtering in JavaScript Views

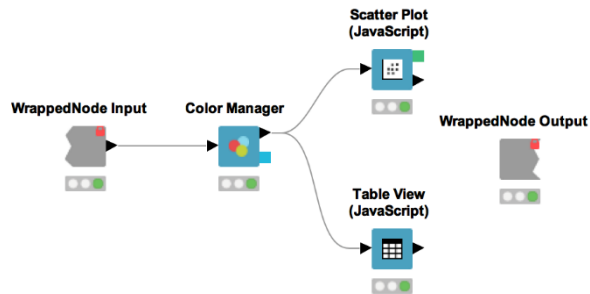
Interactivity allows you to select data points in views

- Selection is propagated to other views.
- Highlight selected rows or filter them
- Click “Apply” to add column to data that indicates selection (true/false) for use in downstream nodes

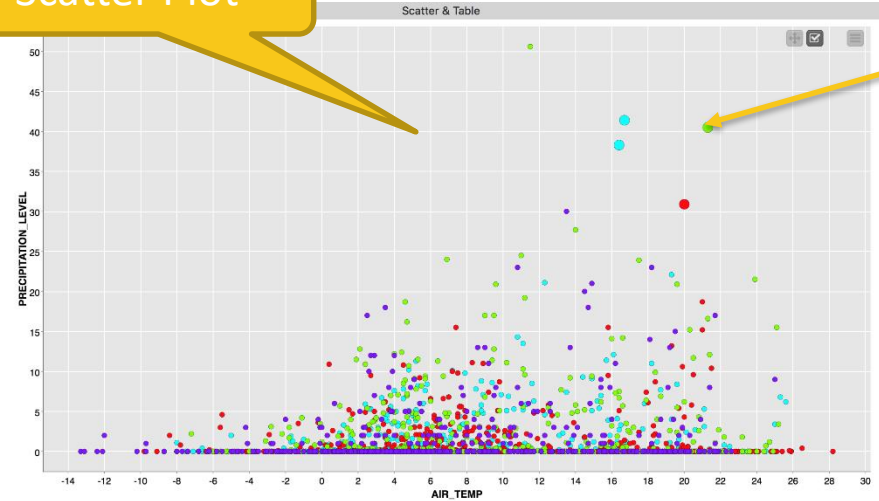


Wrapped Metanodes – Combined Views

- Multiple JavaScript View nodes can be combined in wrapped metanode
- Selections are transmitted to all other views
- Also for use on the KNIME WebPortal



Scatter Plot



Show 10 entries

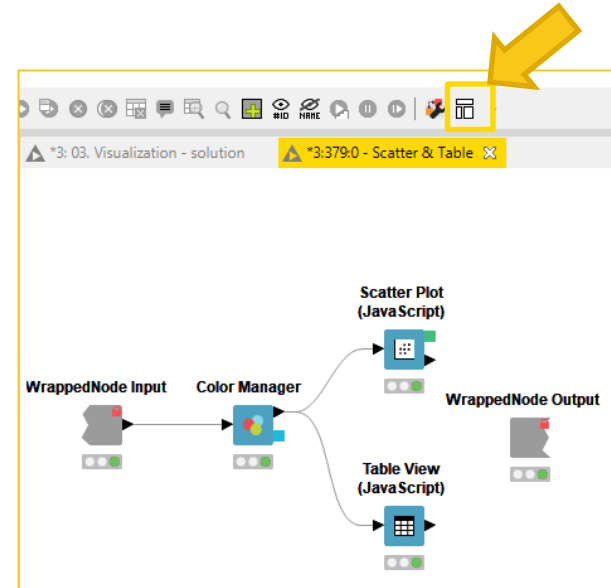
Search:

<input checked="" type="checkbox"/>	DATE	AIR_TEMP	PRECIPITATION_LEVEL	SUNSHINE_HOURS	CITY
<input checked="" type="checkbox"/>	2016-05-29 CEST	16.7	41.4	1.317	Stuttgart
<input checked="" type="checkbox"/>	2016-06-08 CEST	16.4	38.3	0.317	Stuttgart
<input checked="" type="checkbox"/>	2016-07-27 CEST	20	30.9	2.483	Berlin
<input checked="" type="checkbox"/>	2016-08-04 CEST	21.3	40.5	0.417	Constance

Table View

Configure content and views layout

- Click layout button when inside wrapped node to assign views to rows and columns
- Views underneath each other?
 - Same column, different rows
- Side-by-side views?
 - Same row, different columns
- Define width of element
 - Distribute width among elements in a row, sum up to 12
 - E.g. two elements: 6 each, or 9 and 3, etc.
 - Total width can be < 12 if content should not span whole row
 - Single element in row, width = 6 -> takes only half the space, rest is empty



The screenshot shows the 'Advanced Layout' configuration panel. It has three tabs: 'Node Usage', 'Basic Layout', and 'Advanced Layout'. The 'Advanced Layout' tab is active. It contains a table with columns for 'Node', 'Row', 'Column', and 'Width'. Below the table is a 'Reset' button.

Node	Row	Column	Width
Scatter Plot (JavaScript) ID: 377	1	1	12
Table View (JavaScript) ID: 378	2	1	12

Configure content and views layout

Node Usage and Layout

Specify in what way the contained view and Quickform nodes are allowed to be used and define a layout. The layout is used in the KNIME WebPortal and the Wrapped Metanode View.

Node Usage Basic Layout Advanced Layout

Node	Row	Column	Width
Scatter Plot (JavaScript) ID: 377	1	1	12
Table View (JavaScript) ID: 378	2	1	12

Default: Views underneath each other, same column, full width

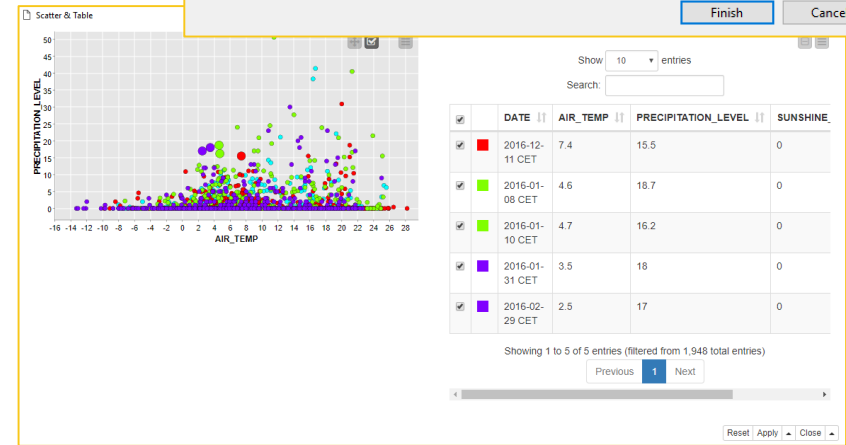
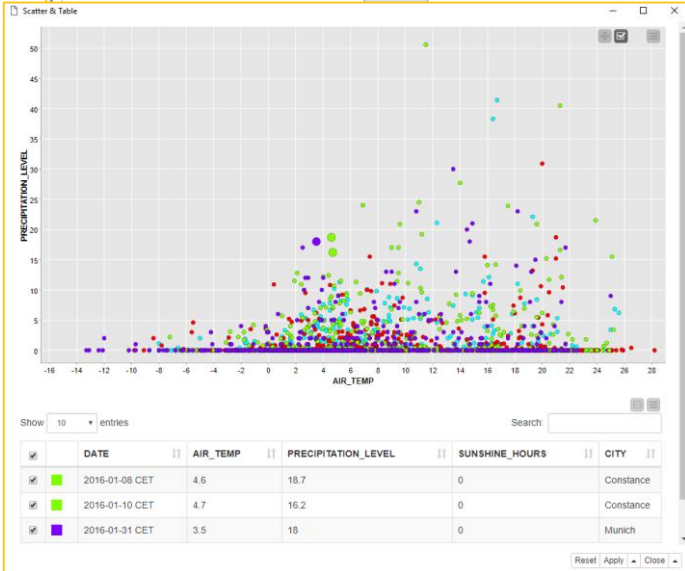
Node Usage and Layout

Specify in what way the contained view and Quickform nodes are allowed to be used and define a layout. The layout is used in the KNIME WebPortal and the Wrapped Metanode View.

Node Usage Basic Layout Advanced Layout

Node	Row	Column	Width
Scatter Plot (JavaScript) ID: 377	1	1	6
Table View (JavaScript) ID: 378	1	2	6

Same metanode, different layout: side-by-side, equal width



Data Aggregation

Sex	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55

Aggregation: Count

Sex	blond	brown	black	red
f	2	1	1	0
m	1	1	0	2

Aggregation: Mean(Age)

Sex	blond	brown	black	red
f	42	53	22	0
m	13	16	0	38,5

Solution: Pivoting Node

Data Aggregation

Sex	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55



Aggregation: Mean(Age)

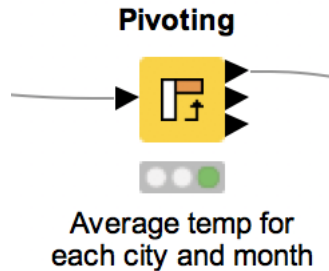
Sex	blond	brown	black	red
f	42	53	22	0
m	13	16	0	38,5

Pivoting Node: **Group** - **Pivot** - **Aggregate**

New Node: Pivoting

Performs pivoting on selected columns for grouping and pivoting

- Values of group columns become unique rows
- Values of the pivot columns become unique columns for each set of column combination together with each aggregation
- Many aggregation methods are provided (similar to GroupBy)



New Node: Pivoting

The image shows the configuration interface for the Pivoting node in KNIME. It is divided into three main sections: **Groups**, **Pivots**, and **Manual Aggregation**.

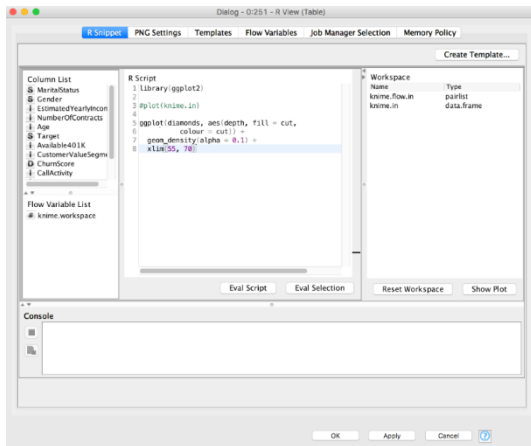
- Groups ~ Rows:** This section is used to define the rows of the pivot table. The 'Available column(s)' list contains 'CITY' and 'AIR_TEMP'. The 'Group column(s)' list contains 'Year' and 'Month (number)'. A yellow callout box points to this section with the text 'Groups ~ Rows'.
- Pivots ~ Columns:** This section is used to define the columns of the pivot table. The 'Pivot column(s)' list contains 'CITY'. A yellow callout box points to this section with the text 'Pivots ~ Columns'.
- Aggregation:** This section is used to define the aggregation function for the data. The 'Available columns' list contains 'AIR_TEMP'. The 'Aggregation (click to change)' dropdown is set to 'Mean'. A yellow callout box points to this section with the text 'Aggregation'.

Below the configuration interface, a preview table is shown. The table has 8 rows (Row0 to Row7) and 6 columns: Year, Month, Berlin+AIR_TEMP, Constance+AIR_TEMP, Munich+AIR_TEMP, and Stuttgart+AIR_TEMP. A yellow arrow points from the 'Aggregation' callout box to the data table.

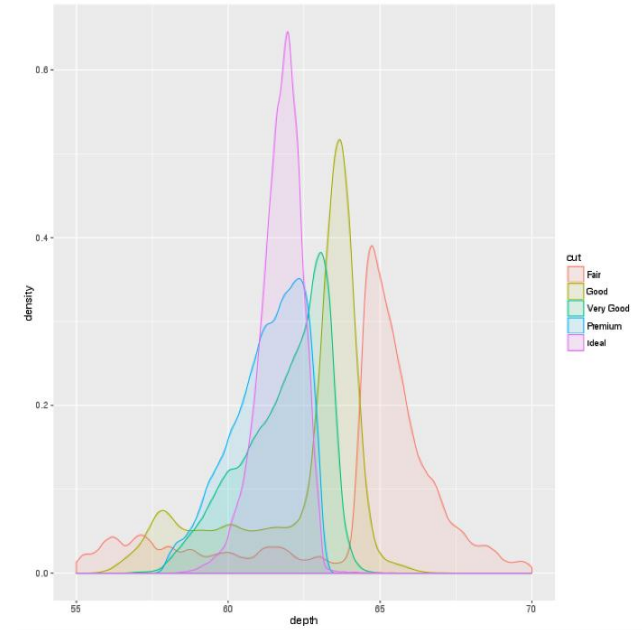
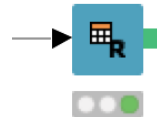
Row ID	Year	Month (..)	Berlin+AIR_TEMP	Constance+AIR_TEMP	Munich+AIR_TEMP	Stuttgart+AIR_TEMP
Row0	2015	12	2.777	2.874	1.255	2.594
Row1	2016	1	2.224	4.545	4.238	4.379
Row2	2016	2	2.229	5.423	4.358	4.742
Row3	2016	3	2.407	9.76	8.477	8.863
Row4	2016	4	6.206	13.71	12.919	13.581
Row5	2016	5	9.663	17.777	16.797	17.357
Row6	2016	6	10.619	20.435	19.11	20
Row7	2016	7				

Script-based View Nodes

- R View nodes for greater customizability
 - Use your favorite libraries, e.g. ggplot2
- If you prefer Python: Python View node
- For JS developers: Generic JavaScript View

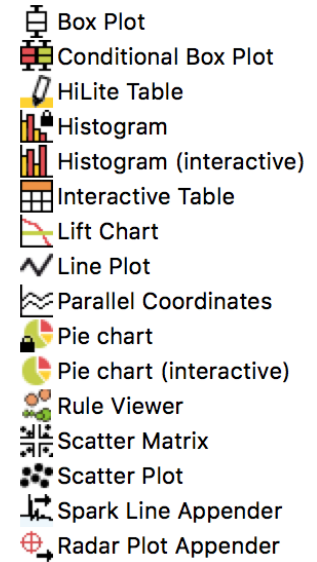
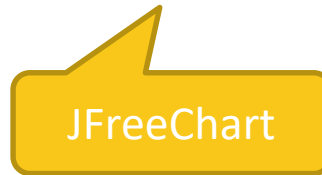
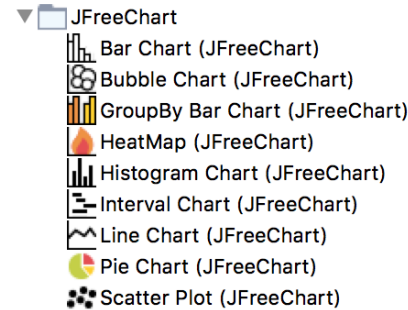


R View (Table)



Legacy View Nodes: JFreeChart & KNIME Views

- KNIME provides three types of visualizations
 - **JavaScript Views**
 - JFreeChart
 - KNIME Views
- Active development only for JavaScript Views -> use those!
- JFreeChart and KNIME Views still useful until all plot types are implemented in JS (we're on it)



Visualization Exercise

Start with exercise: *Visualization*

- Read weather.table
- Use a Color Manager to color by cities, then plot AIR_TEMP against the SUNSHINE_HOURS using Scatter Plot (JavaScript)
- Compare the temperature between cities over time in a Line Plot and a Stacked Area Chart (use Pivoting first!)
- (Use the pivoting node to get the average temperature per month and city and use the month as x-axis)
- Create a composite view by combining a Scatter Plot (JavaScript) and a TableView (JavaScript) in a Wrapped Metanode
 - Select nodes -> right-click -> Encapsulate into Wrapped Metanode

Data Mining

Partition, learn, predict, score

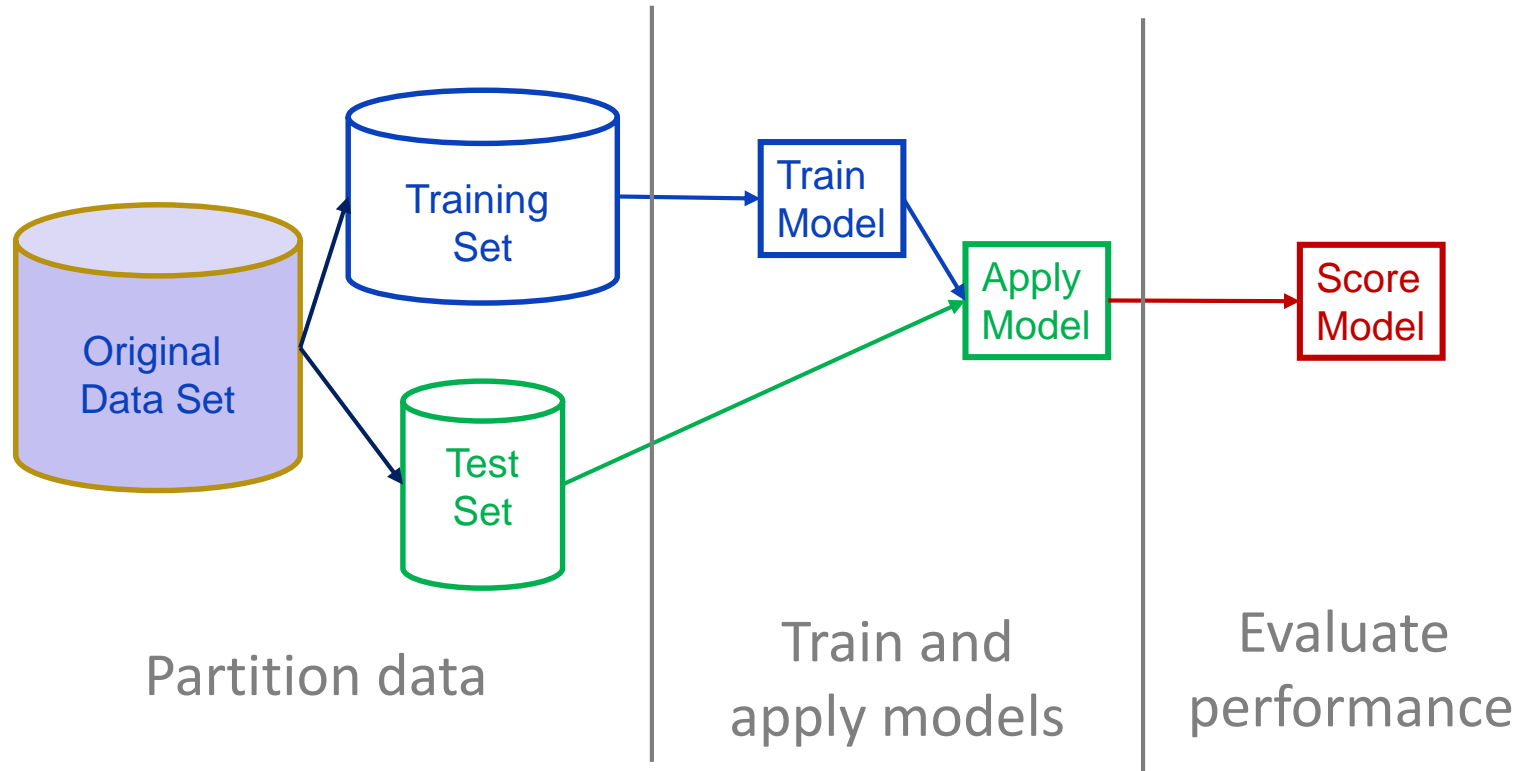


Data Mining Strategies

Example applications:

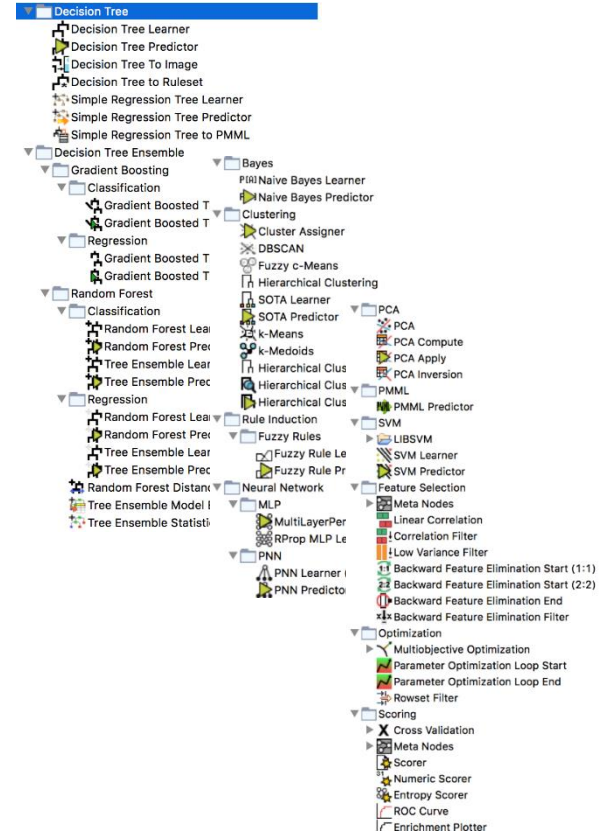
- Anomaly Detection (fraud, predictive maintenance)
- Association Rule Learning (market basket analysis)
- Clustering (market segmentation)
- Classification (next best offer, churn preventions)
- Regression (trend estimation)

Data Mining: Process Overview



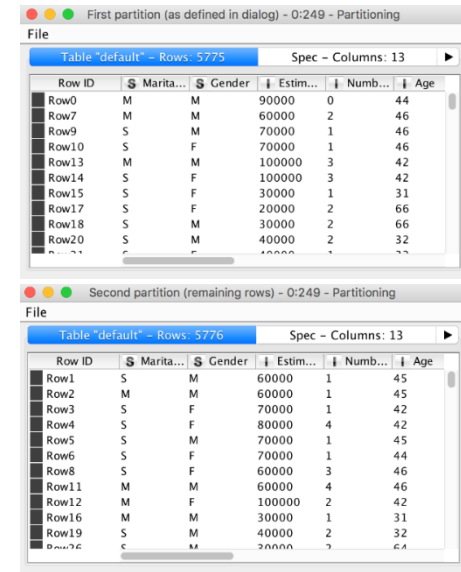
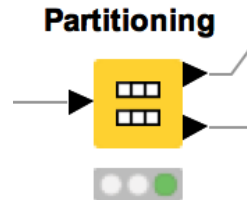
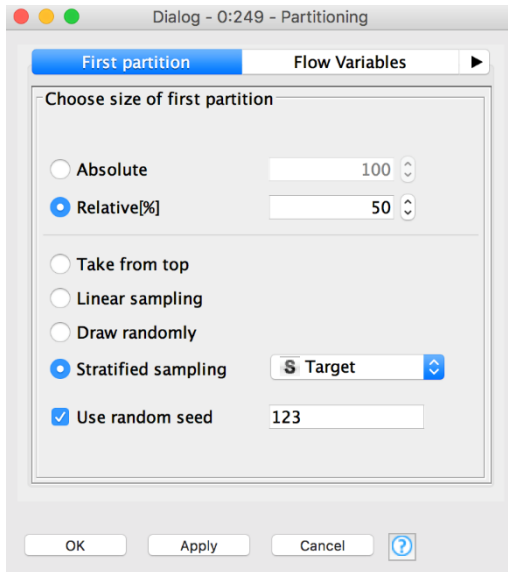
Data Mining in KNIME

- KNIME has many modeling tools!
 - Decision tree, random forest, SVM, regression, neural networks, clustering, ...
 - and integrations with other libraries: R, Python, H2O, WEKA, libSVM, etc.
- And many model evaluation nodes
 - ROC, standard, numeric and entropy scorers
 - Feature elimination
 - Cross validation



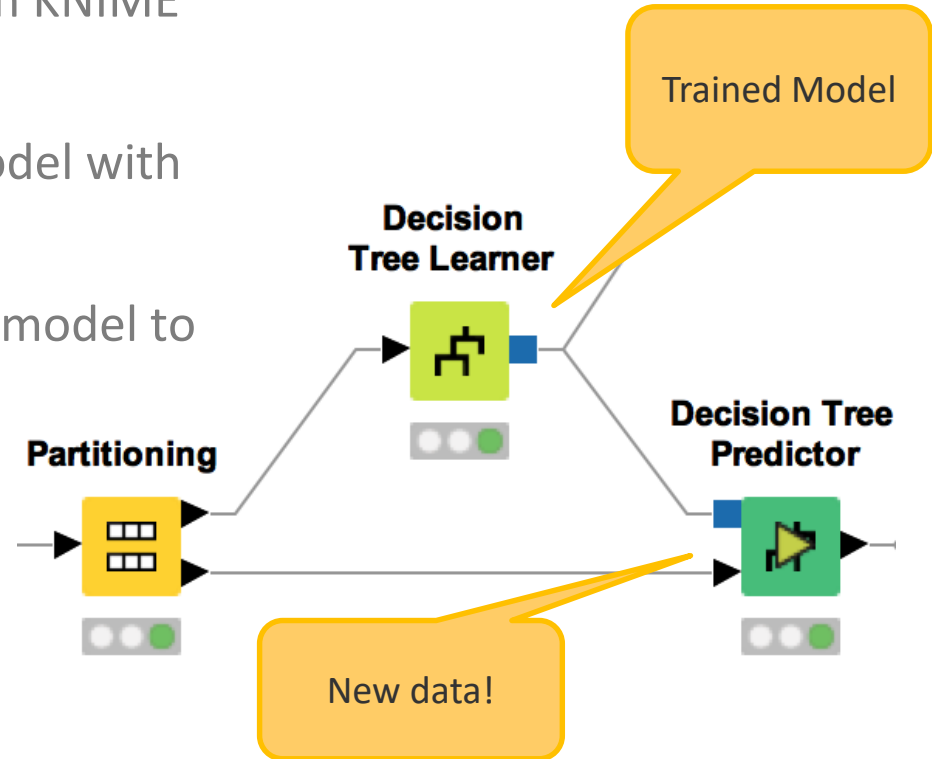
New Node: Partitioning

- Use to split data into training and evaluation sets
 - Partition by count (e.g. 10 rows) or fraction (e.g. 10%)
 - Sample by a variety of methods; random, linear, stratified



Learner-predictor Motif

- Most data mining approaches in KNIME use a Learner-predictor motif.
- The Learner node trains the model with its input data.
- The Predictor node applies the model to a different subset of data.



Classification

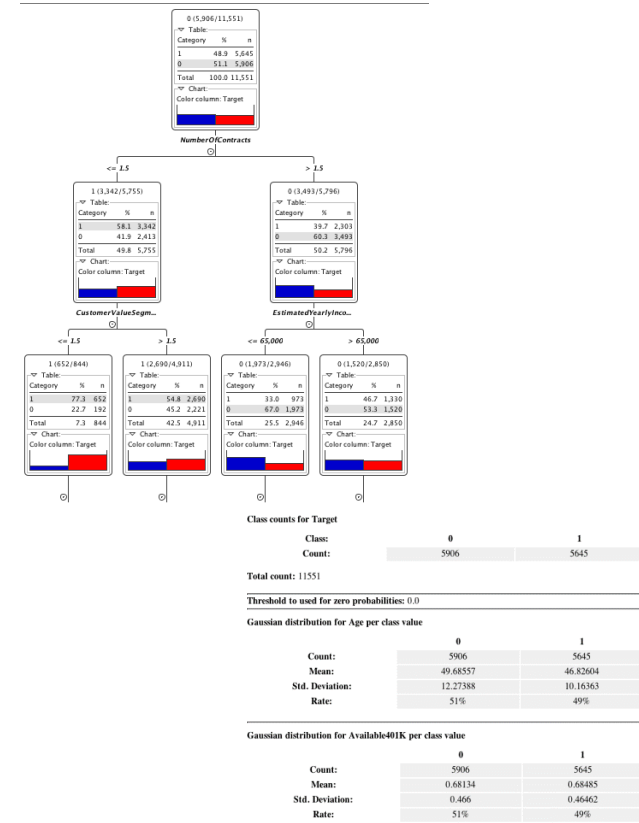
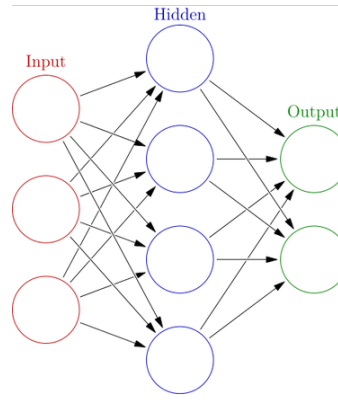
Predict *nominal* outcomes on existing data (supervised)

- Applications

- Churn analysis (yes/no)
- Chemical activity (active/inactive)
- Spam detection (spam/not spam)
- Optical character recognition (A-Z)

- Methods

- Decision Trees
- Neural Networks
- Naïve Bayes
- Logistic Regression



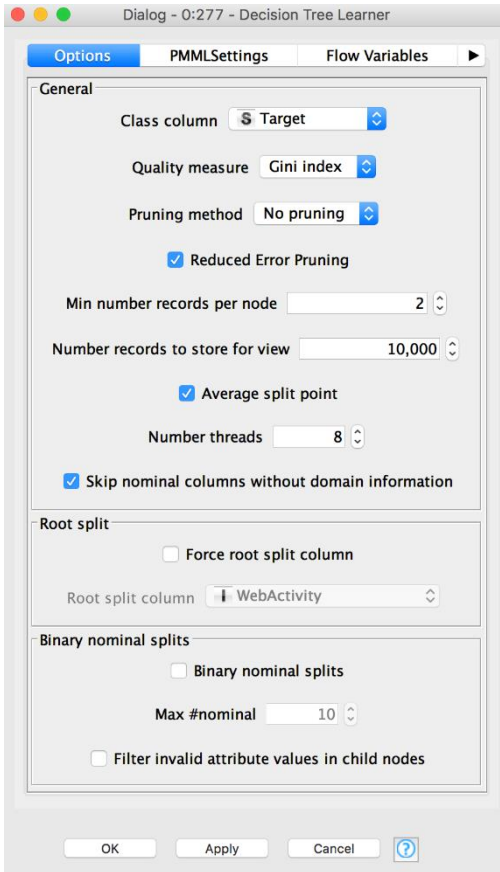
KNIME's Decision Tree

J.R. Quinlan, "C4.5 Programs for machine learning"

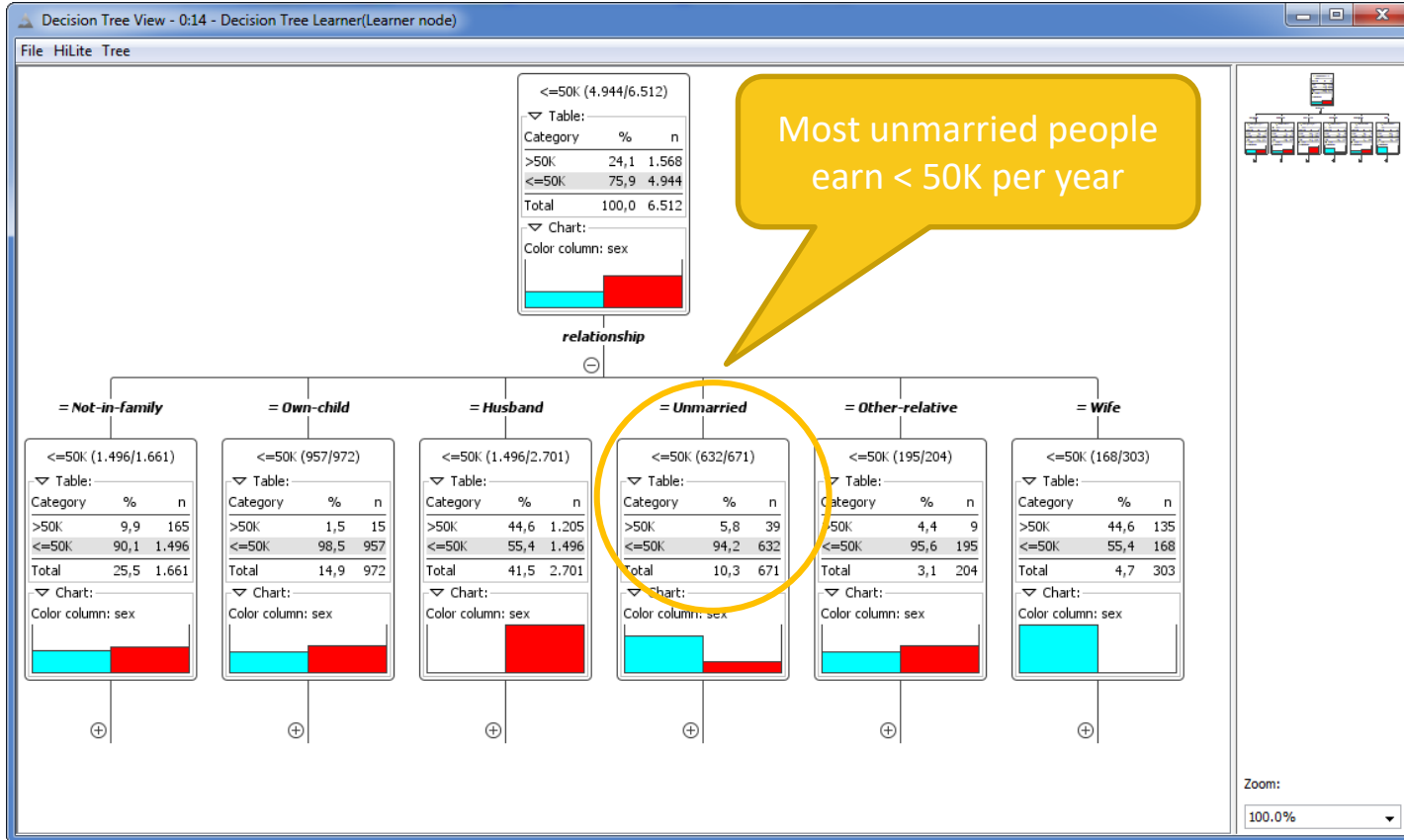
J. Shafer, R. Agrawal, M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining"

- C4.5 builds a tree from a set of training data using the concept of information entropy.
- At each node of the tree, the attribute of the data with the highest **normalized information gain** (difference in entropy) is chosen to split the data.
- The C4.5 algorithm then recurses on the smaller sub lists.

New Node: Decision Tree Learner

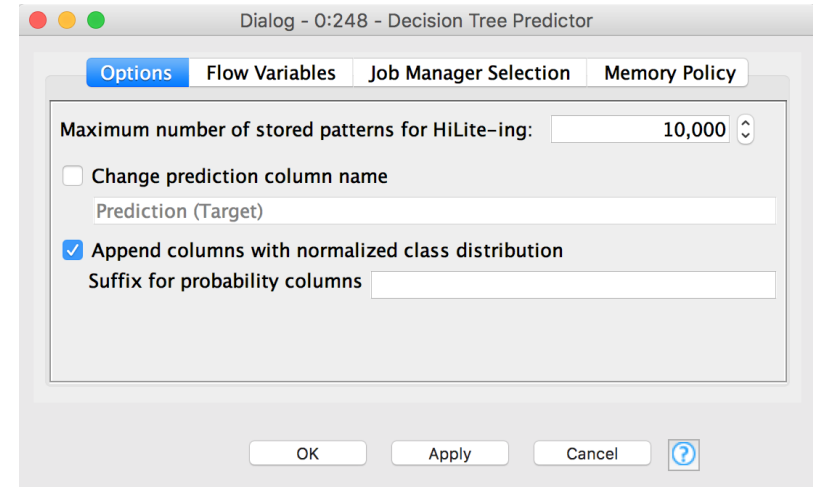
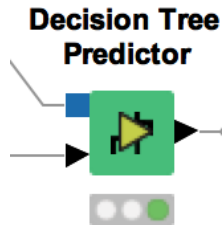


Decision Tree View



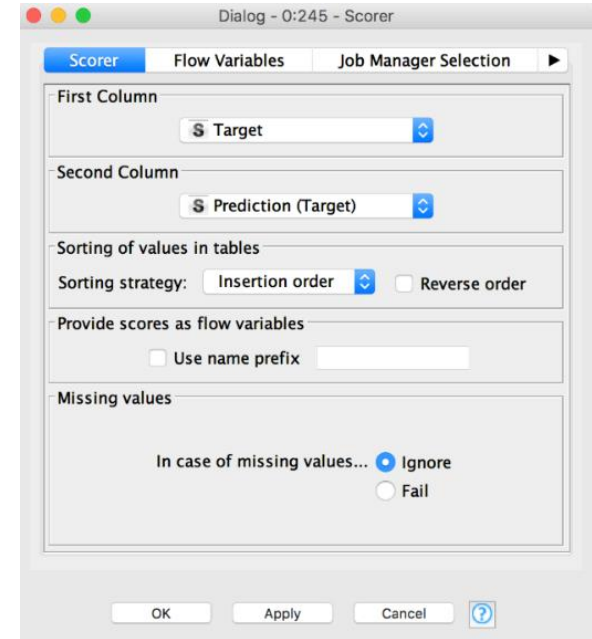
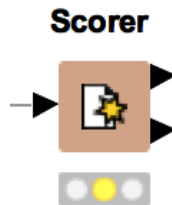
New Node: Decision Tree Predictor

- Takes a decision tree model & applies it to new data
- Check the box to append class probabilities

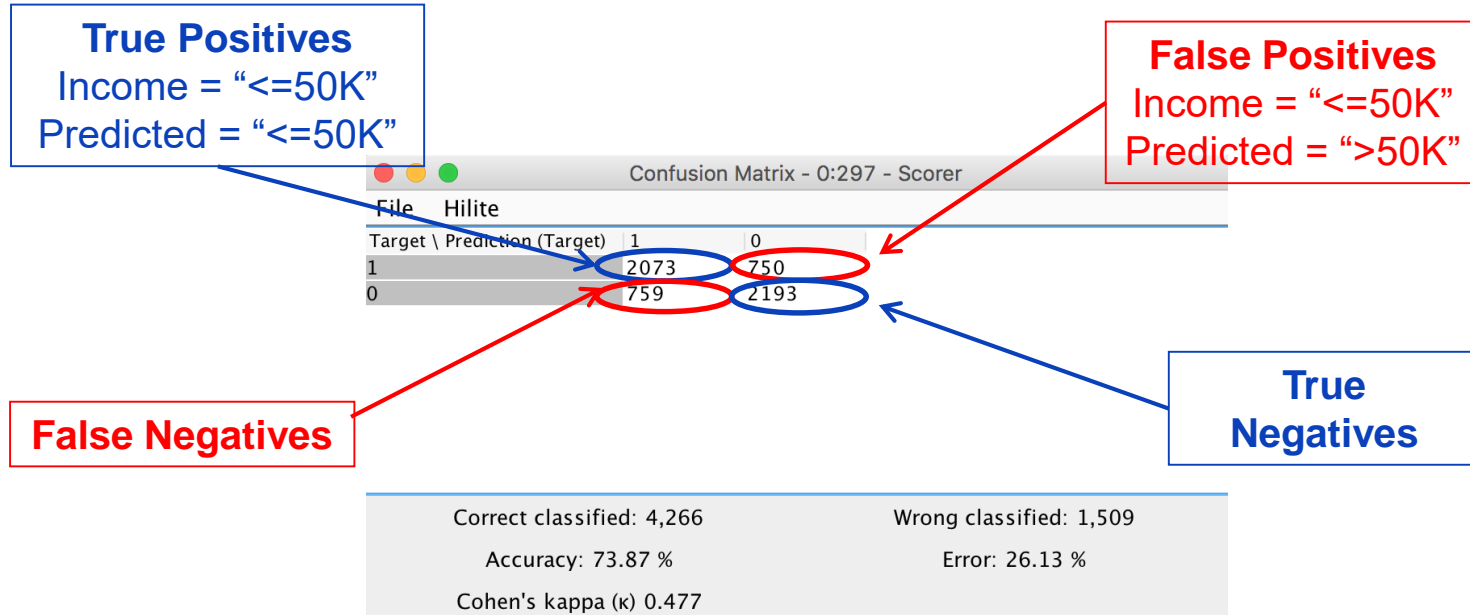


New Node: Scorer

- Compare predicted results to known truth in order to evaluate model quality
 - Confusion matrix shows the distribution of model errors
 - An accuracy statistics table provides a detailed analysis of model quality.



New Node: Scorer



Scorer: Accuracy Measures

Accuracy statistics - 0:297 - Scorer

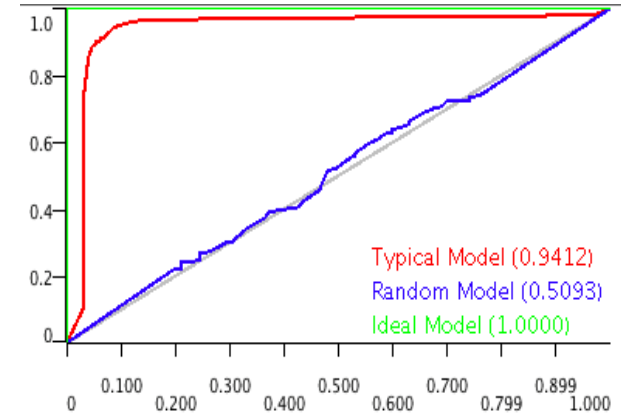
File Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
1	2073	759	2193	750	0.734	0.732	0.734	0.743	0.733	?	?
0	2193	750	2073	759	0.743	0.745	0.743	0.734	0.744	?	?
Overall	?	?	?	?	?	?	?	?	?	0.739	0.477

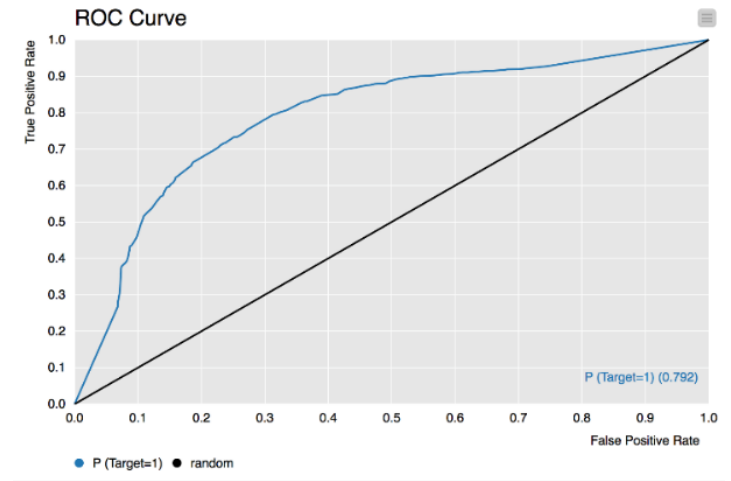
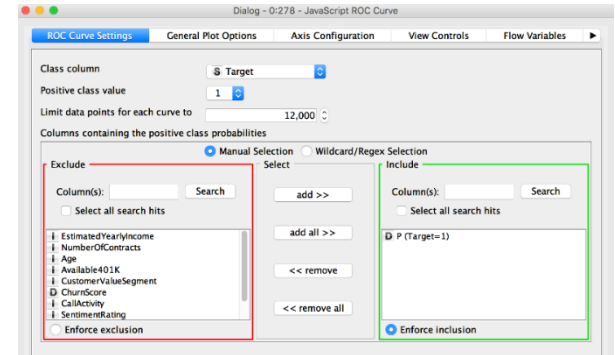
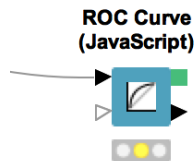
Receiver Operating Characteristics

- Sort by confidence in target class
- Plot true positive rate vs false positive rate
- Ideal models achieve 100% TPR with 0% FPR
- Area under the curve indicates model quality
 - (1=ideal model, 0.5 = random outcome)



New Node: ROC Curve (JavaScript)

- Requires individual class probabilities from a preceding predictor
- User must define:
 1. Original class column
 2. Positive class value
 3. Probability for the selected positive class value for one or multiple models



Data Mining Exercise, Activity I

Start with exercise: *Data Mining, Activity I*:

- Partition the fully joined data
 - 50%, Stratified Sampling
- Train a decision tree on the training data
 - (Learn against “Target” column)
- Use the model to predict the upsell potential for remaining records.
- Evaluate the quality of a model with a Scorer.
- Optional: Find AUC for the model using ROC curve.

Regression

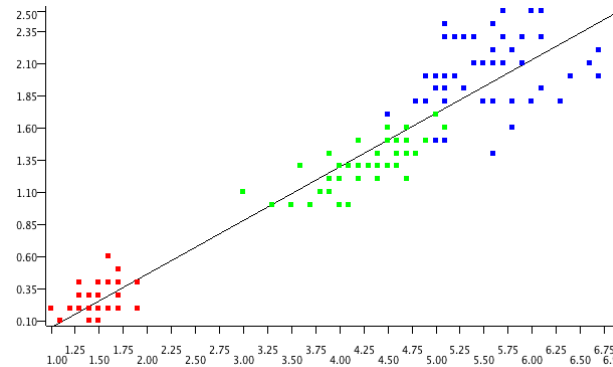
Predict *numeric* outcomes on existing data (supervised)

Applications

- Forecasting
- Quantitative Analysis

Methods

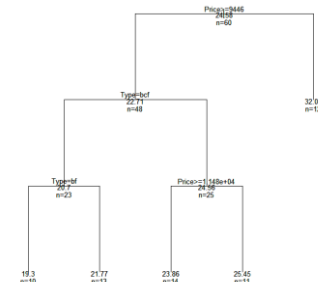
- Linear
- Polynomial
- Regression Trees
- Partial Least Squares



Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
Petal.Length	0.4158	0.0096	43.3872	0.0
Intercept	-0.3631	0.0398	-9.1312	4.44E-16

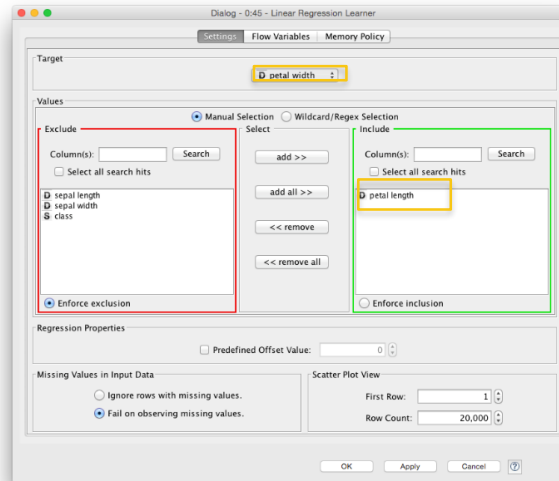
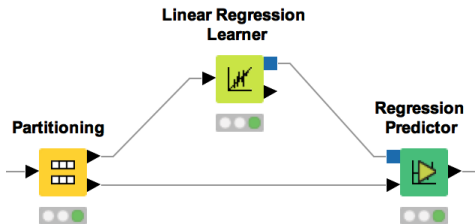
Multiple R-Squared: 0.9271
Adjusted R-Squared: 0.9266



New Nodes: Linear Regression Learner & Regression Predictor

A linear model relating a dependent variable to 1 or more independent variables

- Model coefficients provided in 2nd output port
- Also available: Polynomial and Tree Ensemble Regression nodes



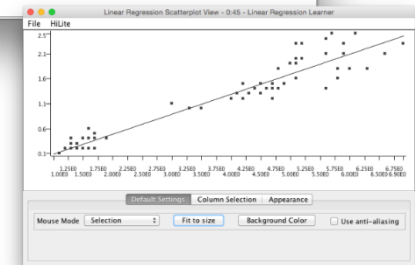
Linear Regression Result View - 0:45 - Lin...

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
petal length	0.4013	0.0132	30.4953	0,0
Intercept	-0.3215	0.0521	-6.165	3.52E-8

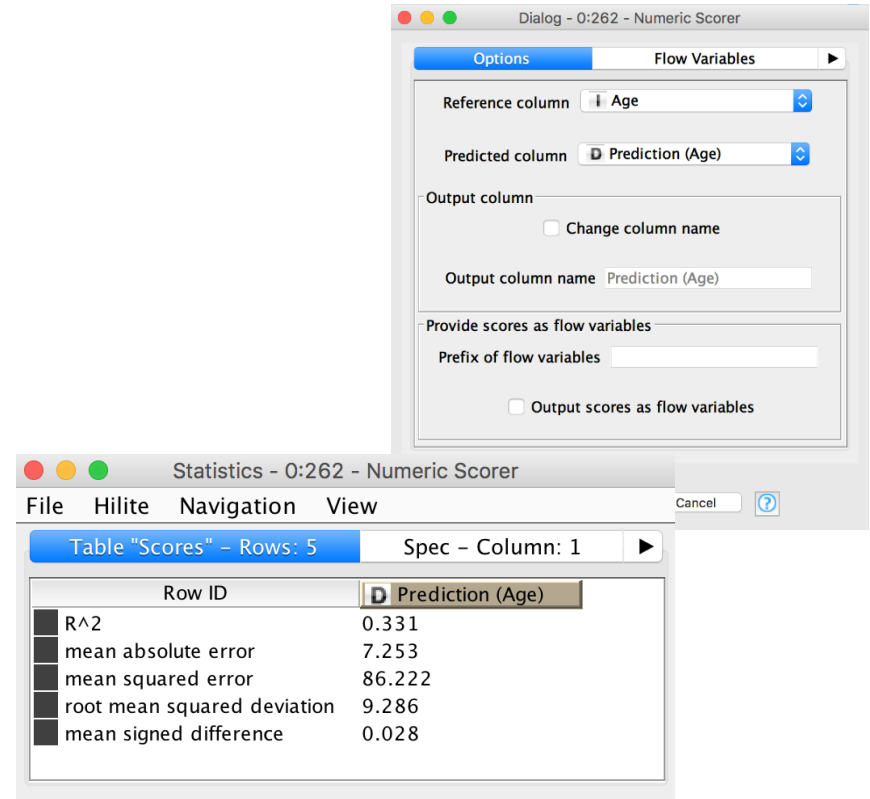
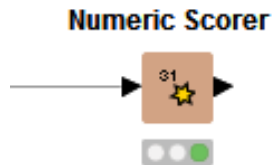
Multiple R-Squared: 0.9272
Adjusted R-Squared: 0.9262



New Node: Numeric Scorer

Similar to scorer node, but for nodes with *numeric* predictions (e.g. linear/polynomial regression)

- Compare dependent variable values to predicted values to evaluate goodness of fit.
- Report R^2 , RMSD, SEM etc.



The image shows two screenshots of the KNIME software interface. The top screenshot is the 'Dialog - 0:262 - Numeric Scorer' window. It has two tabs: 'Options' and 'Flow Variables'. Under 'Options', there are fields for 'Reference column' (set to 'Age'), 'Predicted column' (set to 'Prediction (Age)'), and 'Output column' (with a 'Change column name' checkbox). Under 'Flow Variables', there is a 'Prefix of flow variables' field and an 'Output scores as flow variables' checkbox.

The bottom screenshot is the 'Statistics - 0:262 - Numeric Scorer' window. It shows a table titled 'Table "Scores" - Rows: 5' with one column 'Spec - Column: 1'. The table contains the following data:

Row ID	Prediction (Age)
R^2	0.331
mean absolute error	7.253
mean squared error	86.222
root mean squared deviation	9.286
mean signed difference	0.028

Data Mining Exercise, Activity II

Start with exercise: *Data Mining, Activity II*:

- Read the weather.table
- Split the data into 2016 for training and use 2017 as test data
- Train a linear regression model that predicts the AIR_TEMP as a function of all other parameters in the data set
- Use the model to predict the temperature in 2017 and evaluate it with the Numeric Scorer
- Optional: Calculate mean temperature per month on the training data
 - Join the mean temperature to the test data set (2017)
 - Use the Numeric Scorer to see if the easiest model is better than the Linear Regression

Clustering

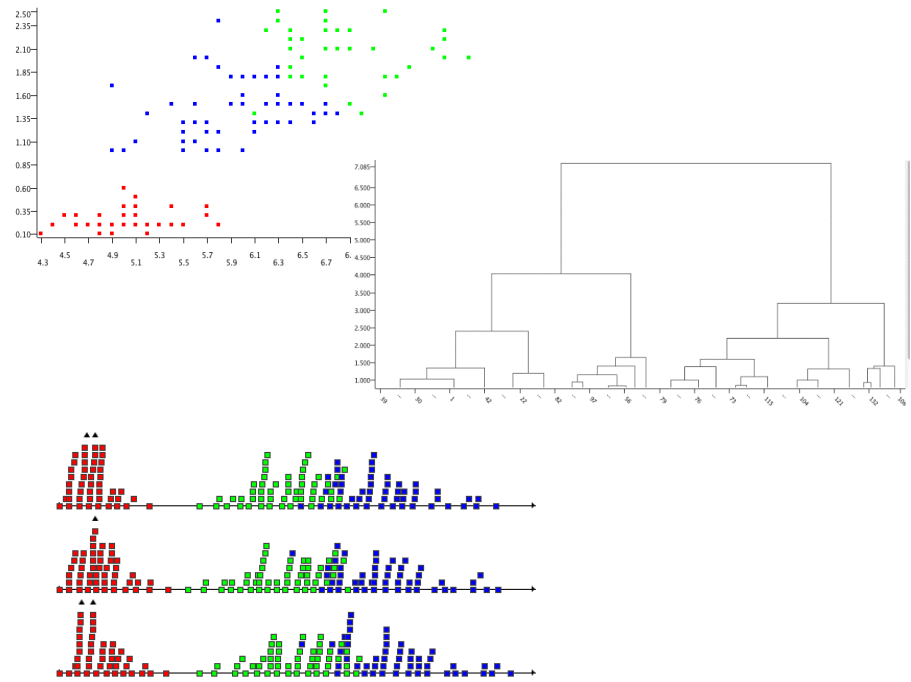
Discover hidden structure in **unlabeled** data (unsupervised)

Applications

- Market Segmentation
- Diversity picking

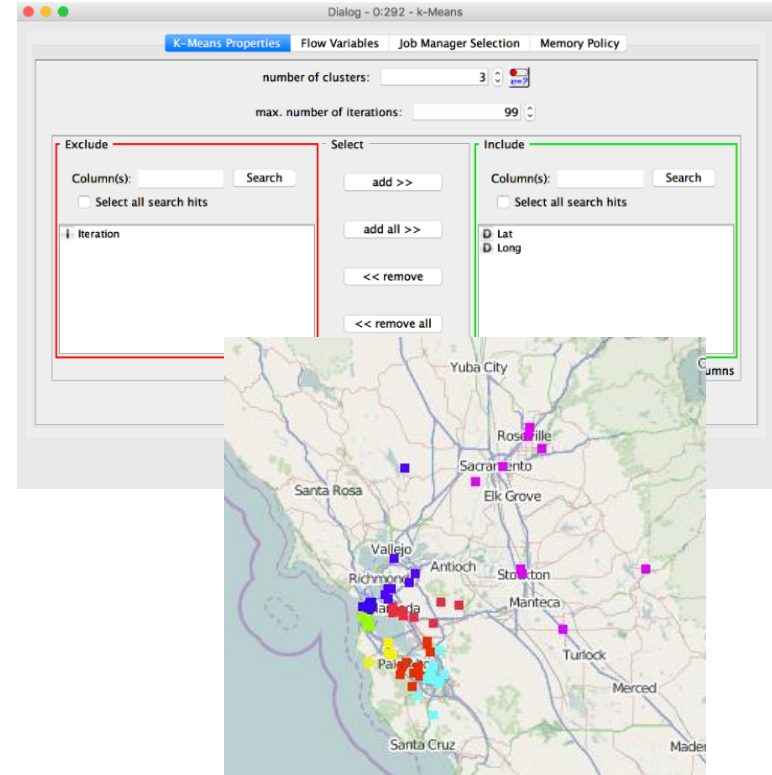
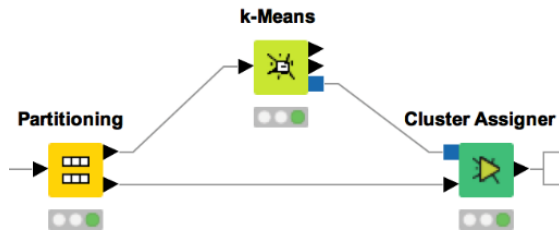
Methods

- K-means/medoids
- Hierarchical
- DBScan
- OPTICS
- Neighbourgrams



New Nodes: k-Means Clustering

- Looks at n observations to define the means for k clusters.
- Each observation is then assigned to its closest cluster center.
- You must provide k .



Data Mining Exercise, Activity III

Start with exercise: *Data Mining, Activity III*

- Read the `location_data.table` file
- Filter to entries from California (`region_code = CA`)
- Train a k-means model with $k=3$. Use only position data for clustering (latitude and longitude)
- Optional: Plot latitude and longitude in a view (OSM Map or Scatter Plot) and use that to help you visually optimize k .

Integrating External Tools

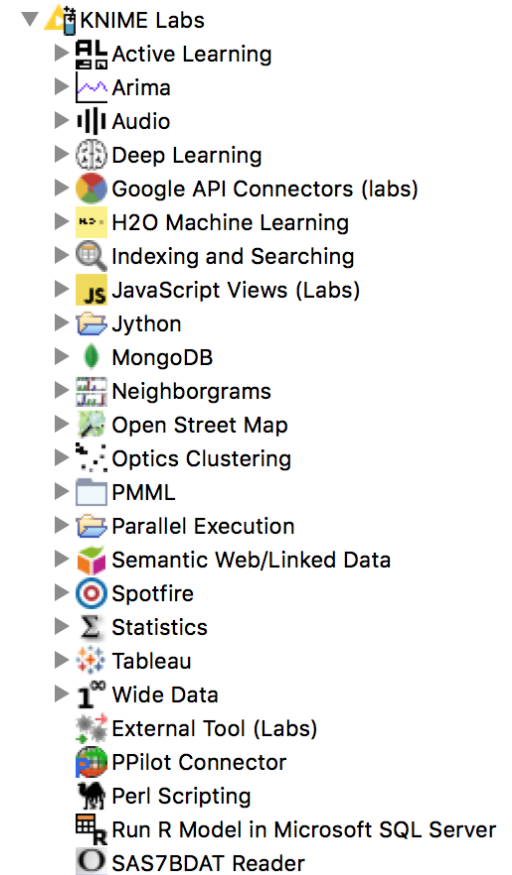


Goal of This Session

- This session gives a quick overview of the external tools that can be called within KNIME, e.g.:
 - Java, R, Python
 - Web services

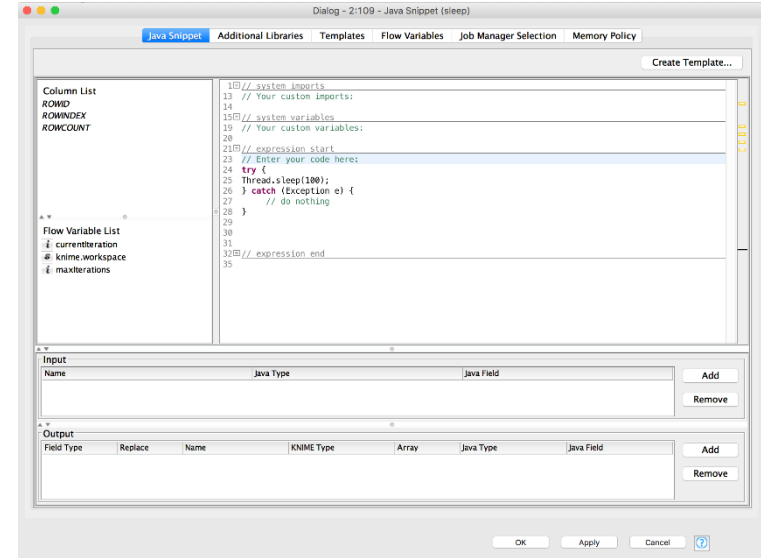
KNIME Labs

- KNIME Labs enable you to preview new KNIME features and plug-ins that are still under development.
- The nodes provided in KNIME Labs are not (yet) part of the official KNIME version because the functionality and/or API may not be finalized.
- You can get these plug-ins by installing the extension from the KNIME Labs extensions category.

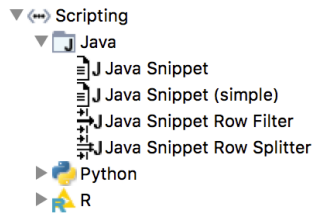
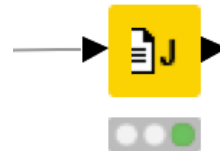


Java Snippet

- Fastest running scripting node in KNIME
- Syntax highlighting, auto completion, error checking
- Templates allow you to save scripts for later re-use
- Import custom libraries

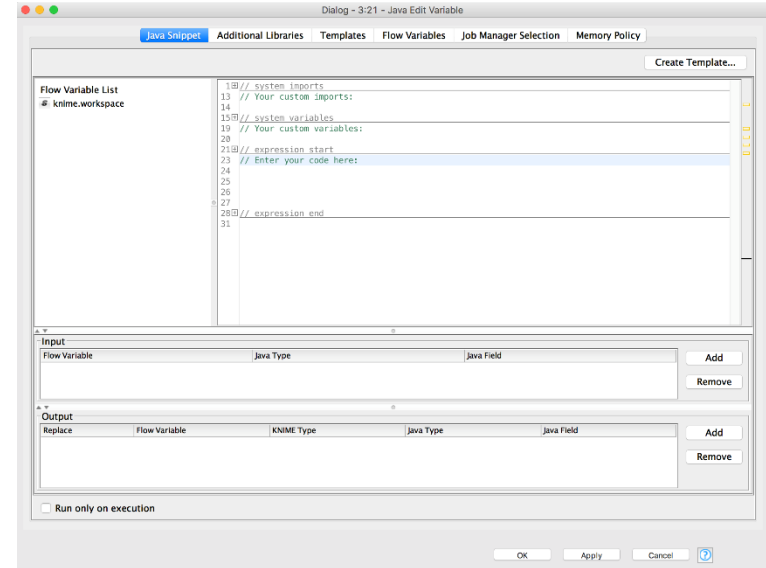


Java Snippet



Java Edit Variable

- Same as Java snippet, but with flow variable ports



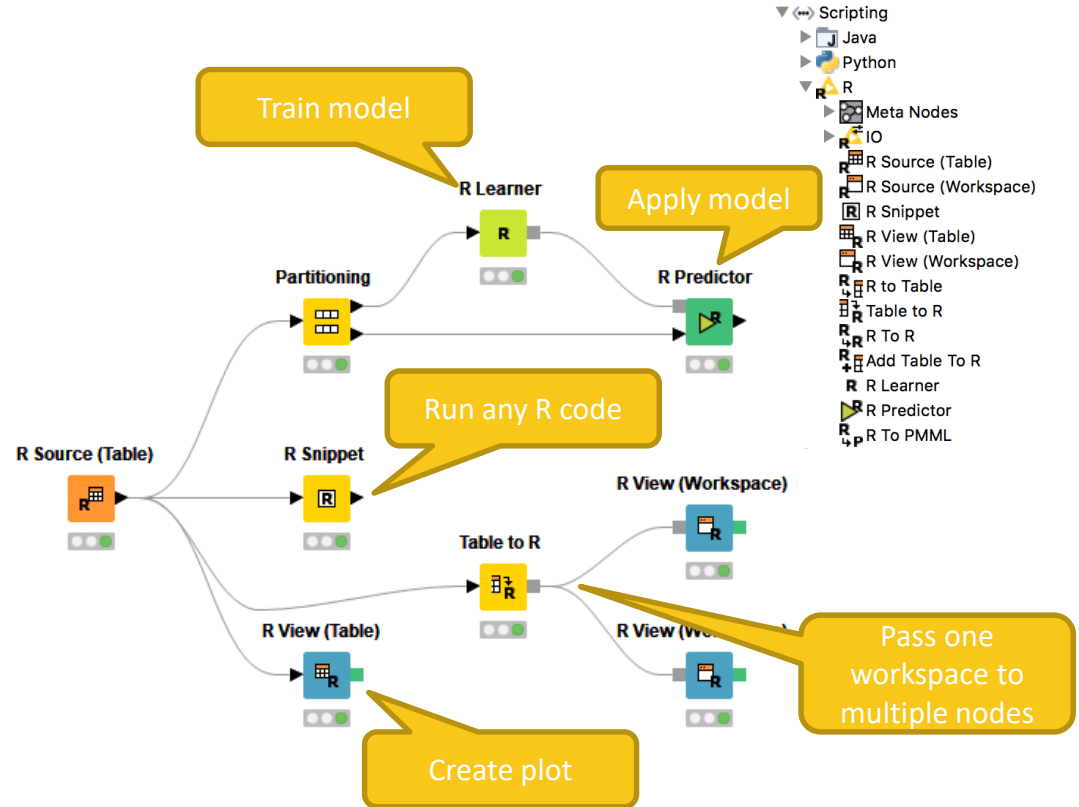
Java Edit Variable



R Integration

- Run any R code from KNIME
- Works with existing R installations
- Nodes for many tasks
- First run:
install.packages('Rserve')
and
install.packages('Cairo')*

*mac only



R Integration

The image shows the KNIME R Integration interface with several callouts highlighting key features:

- Syntax Highlighting:** Points to the R script editor where code is color-coded.
- Create and store templates:** Points to the 'Create Template...' button at the top right.
- R workspace:** Points to the 'Workspace' table on the right side of the editor.
- Show Results:** Points to the 'Show Plot' button at the bottom right of the editor.
- Evaluate script:** Points to the 'Eval Script' button at the bottom center of the editor.
- R console output:** Points to the 'Console' area at the bottom of the interface.

The R script in the editor is as follows:

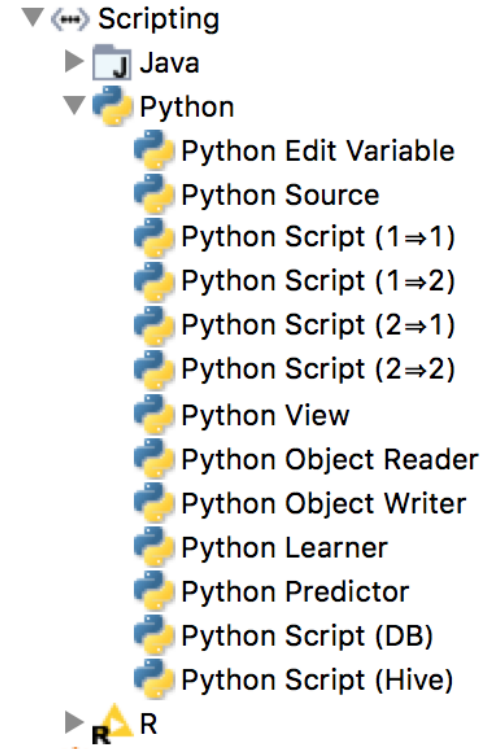
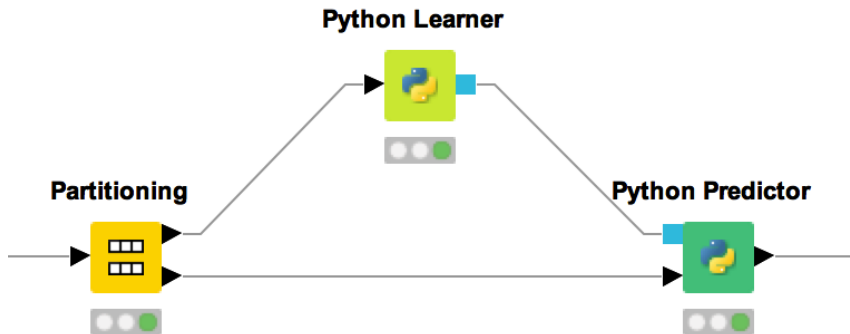
```
1 # This example relies on ggplot2 and grid,  
2 # your R installation, please add them!  
3 library(ggplot2)  
4 library(grid)  
5  
6 # Insert column references here.  
7 # Note: variable names are used as plot la  
8  
9 x = knime.in$"Universe_0_0"  
10 y = knime.in$"Universe_0_1"  
11 #Column to color by:  
12  
13 class =knime.in$"Cluster Membership"  
14  
15 # Use a flow variable for a title  
16  
17 title = 'foo'  
18  
19 # define a plot theme  
20 # http://docs.ggplot2.org/0.9.2.1/theme.ht  
21 clean_theme = theme(panel.background = ele
```

The 'Workspace' table contains the following data:

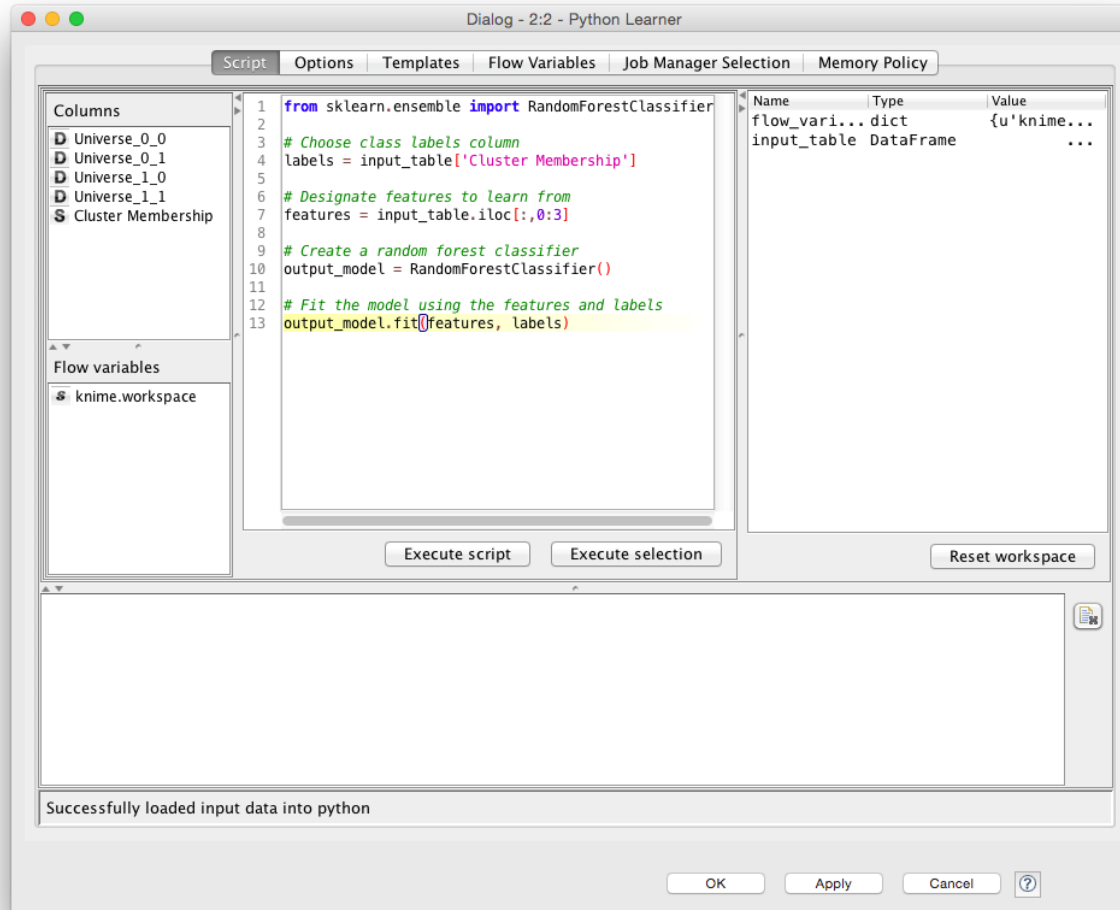
Name	Type
knime.flow.in	pairlist
knime.in	data.frame

Python Integration

- Run Python inside KNIME
- Works with existing installations
- UI modeled after R integration

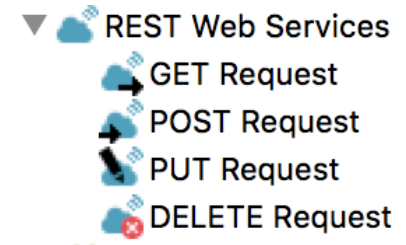


Python Scripting UI

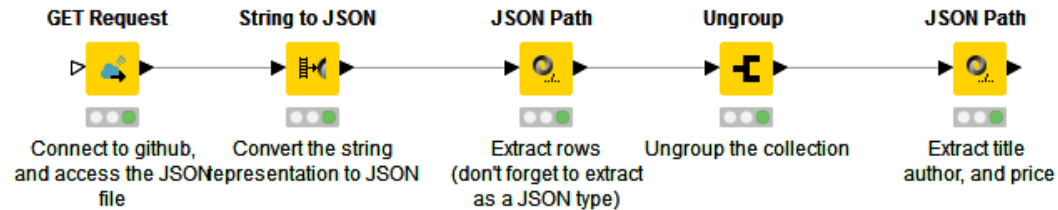


RESTful Web Services

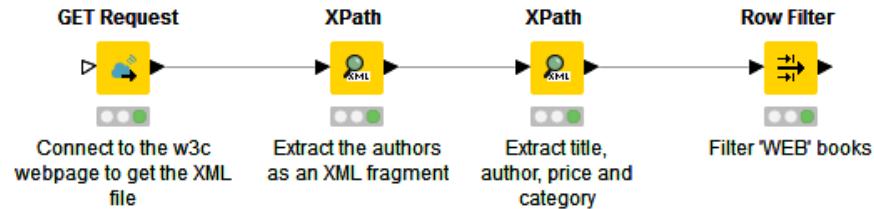
- Use KNIME nodes to interact with RESTful web services
- Send requests using standard HTTP methods



JSON Response:

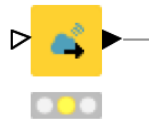


XML Response:



RESTful Web Services

GET Request



Enter URL, or use from column

Add delay between individual requests

Provide authentication if necessary

Dialog - 2:26 - GET Request (Connect to the w3c)

File

Connection Settings Authentication Request Headers Response Headers Flow Variables Job Manager Selection Memory Policy

URL: http://www.w3schools.com/xsl/books.xml

URL column:

Delay (ms): 0

Concurrency: 1

SSL

Ignore hostname mismatches

Trust all certificates

Fail on connection problems (e.g. timeout, certificate errors, ...)

Fail on http errors (e.g. page not found)

Follow redirects

Timeout (s): 2

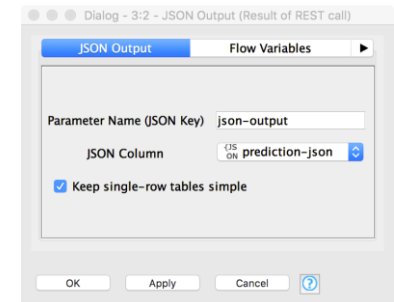
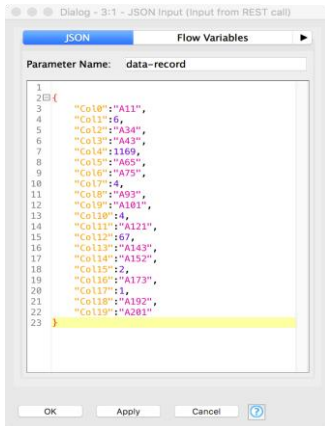
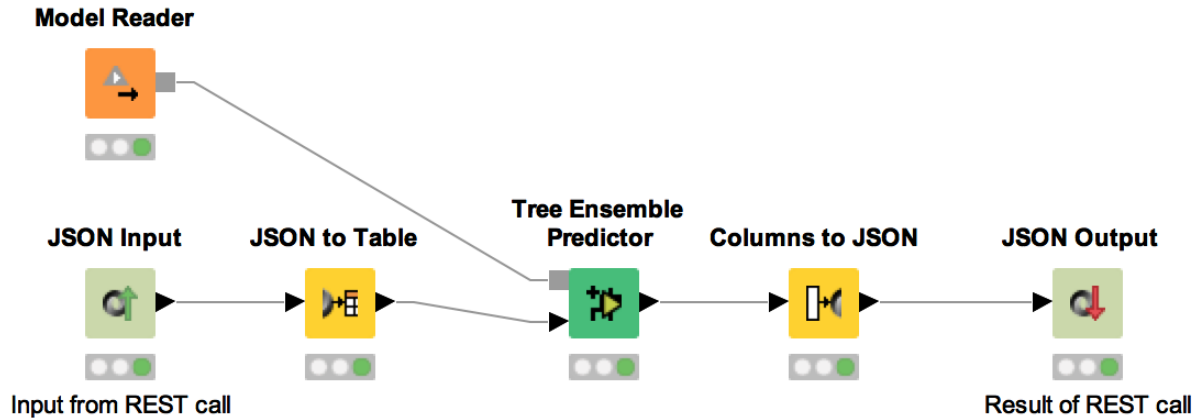
Body column: body

OK Apply Cancel ?

<https://www.knime.com/blog/a-restful-way-to-find-and-retrieve-data>

<https://www.knime.com/blog/OSM-meets-CSV-file-and-Google-API>

KNIME Server as a REST resource



<https://www.knime.org/blog/giving-the-knime-server-a-rest>

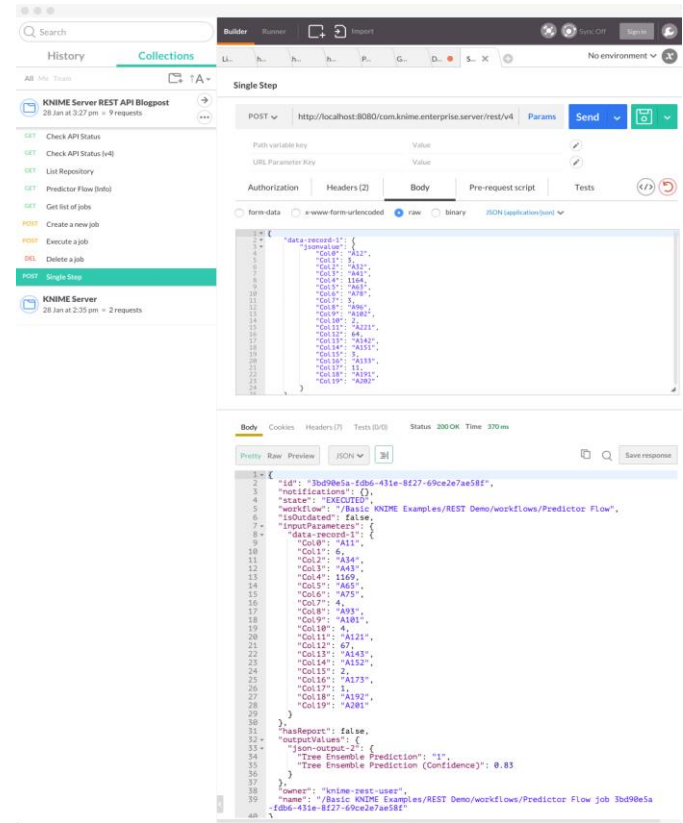
KNIME Server as a REST resource

- Use cURL, SOAPUI or Chrome extension Postman to explore the REST API

```
Jons-MacBook-Pro:~ jon$ curl -u 'knime-rest-user:knockknock' -i http://localhost:8080/com.knime.enterprise.server/rest
HTTP/1.1 200 OK
Server: Apache-Coyote/1.1
Cache-Control: private
Expires: Thu, 01 Jan 1970 00:00:00 UTC
Cache-Control: no-transform,max-age=86400
Date: Thu, 28 Jan 2016 13:27:45 GMT
KNIME-Class: com.knime.enterprise.server.rest.api.ent.ServerVersion
Link: <http://localhost:8080/com.knime.enterprise.server/rest/_profile/knime-server-doc.xml>;rel="profile"
Content-Type: application/vnd.mason+json;charset=UTF-8
Content-Length: 481
```

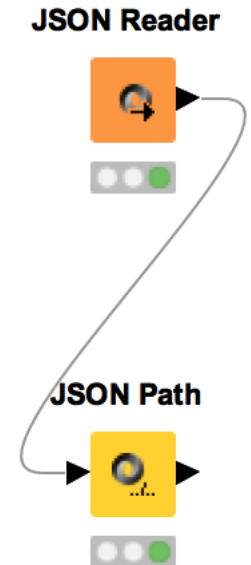
```
{
  "@controls": {
    "self": {
      "href": "http://localhost:8080/com.knime.enterprise.server/rest/",
      "method": "GET"
    },
    "knime:v4": {
      "href": "http://localhost:8080/com.knime.enterprise.server/rest/v4/",
      "title": "KNIME Server API v4",
      "method": "GET"
    }
  },
  "version": {
    "major": 4,
    "minor": 2,
    "revision": 2
  },
  "@namespaces": {
    "knime": {
      "name": "http://www.knime.com/server/rels#"
    }
  }
}
```

```
Jons-MacBook-Pro:~ jon$
```



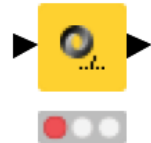
JSON and JSON Path

- Use the JSON Reader (or the GET Resource) nodes to get an JSON cell
- Use JSONPath nodes to query the JSON and extract certain parameters
- Editor window simplifies construction of JSON queries by auto-generating them (click on properties)



JSON Path

JSON Path



Dialog - 2:6 - JSON Path (Extract rows)

Settings | Flow Variables | Job Manager Selection | Memory Policy

Input
JSON Representation

Remove source column

Outputs

Output column	JSONPath	List	Paths
rows	[\$*]	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Add single query | Add collection query | Add JSONPath

JSON-Cell Preview

```
1 [ {
2   "id" : "978-0641723445",
3   "cat" : [ "book", "hardcover" ],
4   "name" : "The Lightning Thief",
5   "author" : "Rick Riordan",
6   "series_t" : "Percy Jackson and the Olympians",
7   "sequence_i" : 1,
8   "genre_s" : "fantasy",
9   "inStock" : true,
```

Undo
Can't Redo
Cut
Copy
Paste
Delete
Select All
Folding
Add JSONPath

Edit

JSONPath:

Column name:

Result type:

Result is list Return the paths instead of values

OK Cancel

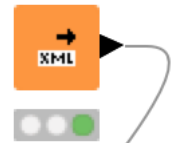
OK Apply Cancel ?

1083M of 1334M

XML and XPath

- Use the XML Reader (or the GET Resource) nodes to get an XML cell
- Use XPath nodes to query the XML and extract certain parameters
- Editor window simplifies construction of XPath queries by auto-generating them (click on XML elements)

XML Reader



XPath



XPath



Dialog - 13:46 - XPath

Settings Namespace Flow Variables

XML column: XML XML

Remove source column.

XPath summary

Column name	XPath query	Type
author	/book/author	String(SingleCell)

Selected XPath: /book/author

Add XPath Edit XPath Remove XPath

XML-Cell Preview

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <book year="2001">
3   <author>Vincent Bugliosi</author>
4   <title>The Betrayal of America: How the Supreme Court Unde
5   <category>About Elections</category>
6   <isbn>1-56025-355-X</isbn>
7 </book>
8
```

OK Apply Cancel ?

XPath Query Settings

Column name:

New column name: author

XPath query for column name: name (relative to value query)

XPath value query

/book/author

Return type

XPath data type: String cell

Options:

Return missing cell on empty string.

Multiple tag options

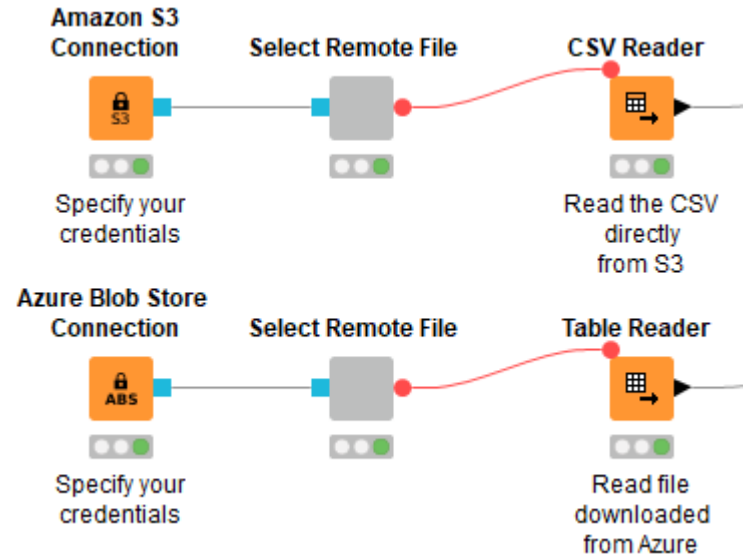
Single Cell Collection Cell

Multiple Columns Multiple Rows

Ok Cancel ?

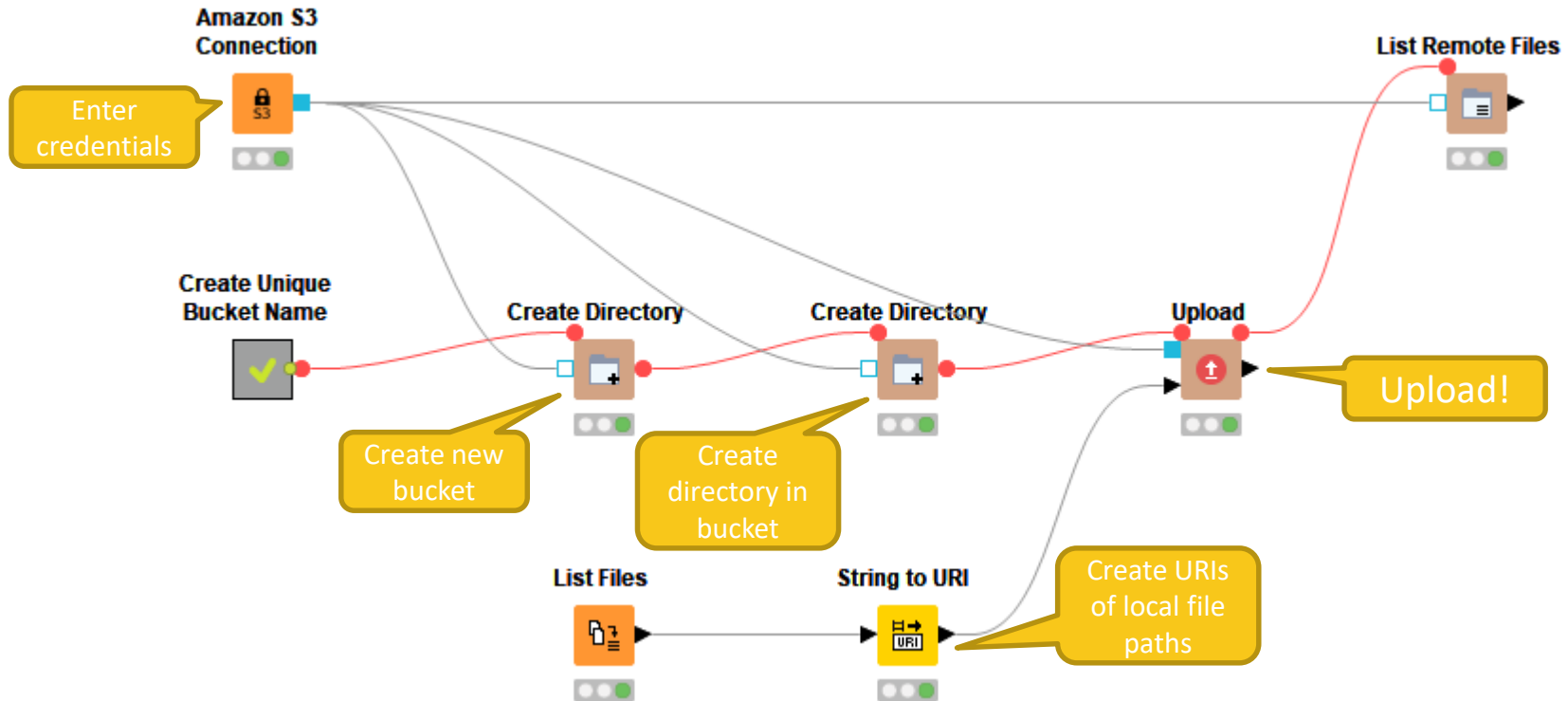
Remote File Handling – Cloud Storage

- Integrate remote data sources from Amazon AWS and Microsoft Azure
 - Upload files
 - Download files, or read their content directly into KNIME
 - List files in remote directories
 - Create directories
 - Delete files / directories



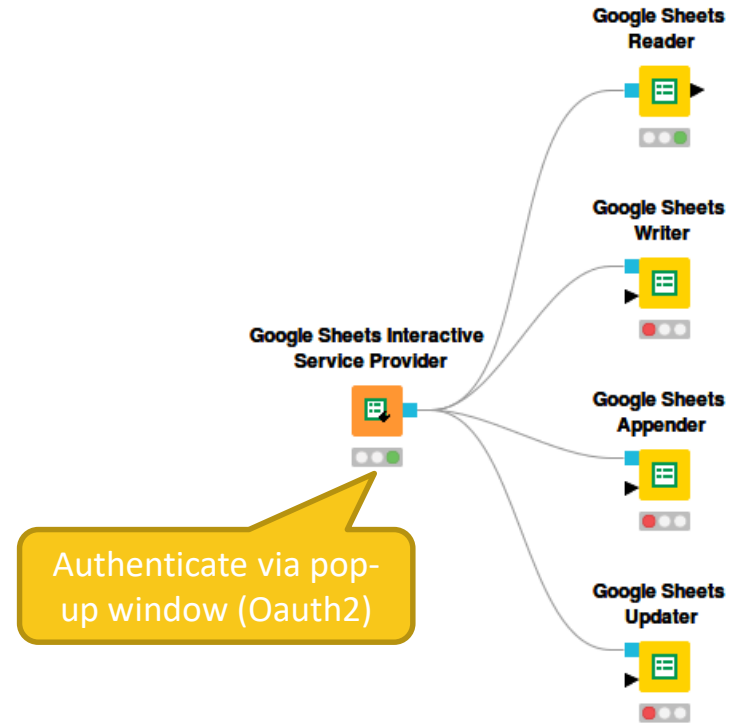
Remote File Handling – Cloud Storage

Example: Upload all files from a local directory to Amazon S3



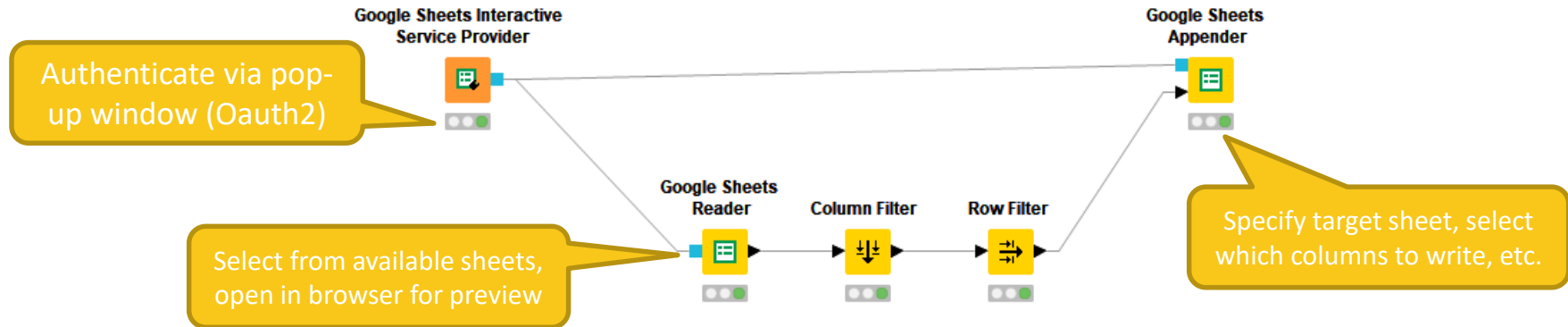
Google Sheets

- Access your data stored in Google Services
 - Read data from Google Sheets
 - Write data to new sheets
 - Modify existing sheets
- Makes collaboration and sharing of data easy
 - (especially vs. sending Excel sheets via email...)



Google Sheets

- Select from available sheets on Google Drive
- Transform data in KNIME, or enrich with new data
- Create new sheet or update existing sheets
 - Allows to read from / write to specific range of sheet (e.g. A1:G10)



Exercises

Start with exercise: *Integrating External Tools*

- Use the GET Request node to call an external web service
 - <https://raw.githubusercontent.com/tamingtext/book/master/apache-solr/example/exampledocs/books.json>
- Read books.json and use the JSON Path node to extract the book name, author, and price

Exporting Data & Deployment



Exporting Data

After an analysis is completed, what next?

- Write results to a file
- Create/update a database
- Save the model for use elsewhere
- Generate a rich report
- Deploy via KNIME WebPortal
- Deploy via workflow as RESTful web service

Input/Output in Deployment

Input

- File (CSV, Table, XLS, ...)
- Database
- JSON for REST API

Output

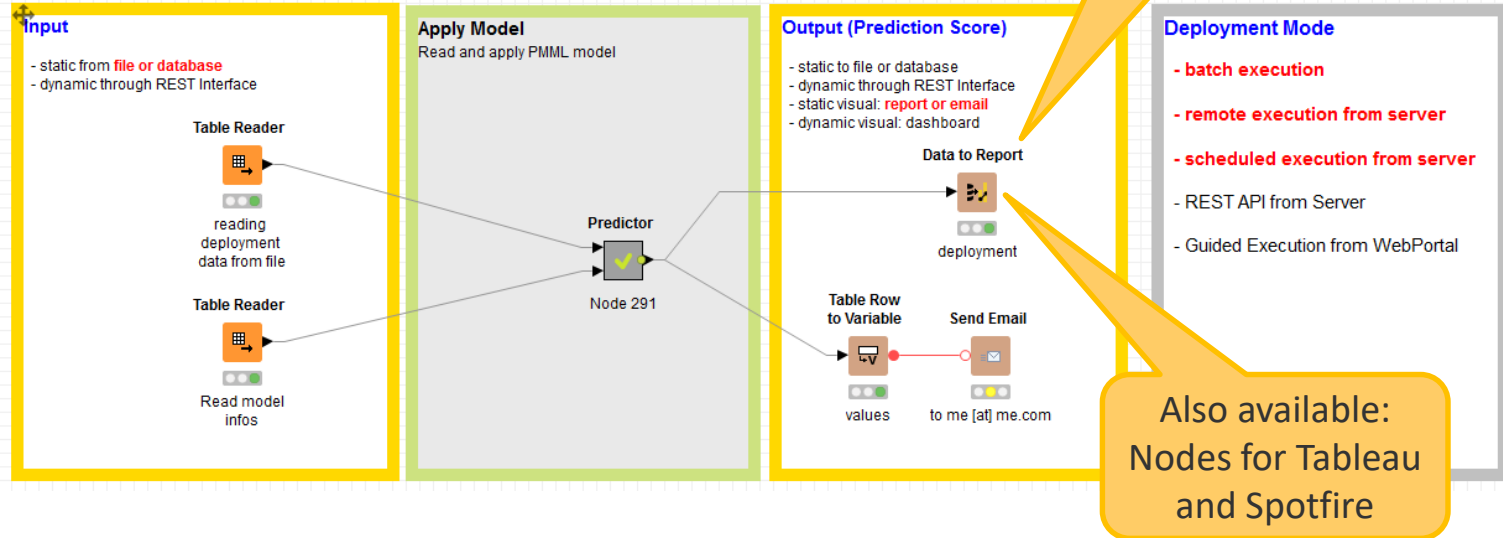
- Report (BIRT, Tableau, Spotfire)
- Email
- File (CSV, Table, XLS, ...)
- WebPortal

To Report / Email

Model Deployment with final report (Scheduling)

This workflow:

- reads new unseen data from file (.table format),
- scores the data with the available current model,
- appends model prediction and probabilities to original data
- produces a report (BIRT here) with table, bar chart, title, etc ... Report can be exported as .docx, html, pptx, .ps, .pdf, etc ...



To File / Database

Model Deployment File to Database (Scheduling)

This workflow:

- reads new unseen data from file (.table format),
- scores the data with the available current model,
- appends model prediction and probabilities to original data
- writes results to database

Input

- static from **file or database**
- dynamic through REST Interface

Table Reader



reading
deployment
data from file

Table Reader



Read model
infos

Apply Model

Read and apply PMML model

Predictor

Node 291

Output (Prediction Score)

- static to **file or database**
- dynamic through REST Interface
- static visual : report or email
- dynamic visual: dashboard

Database Writer



connect to DB

table
DeploymentData
WithPredictions

Deployment Mode

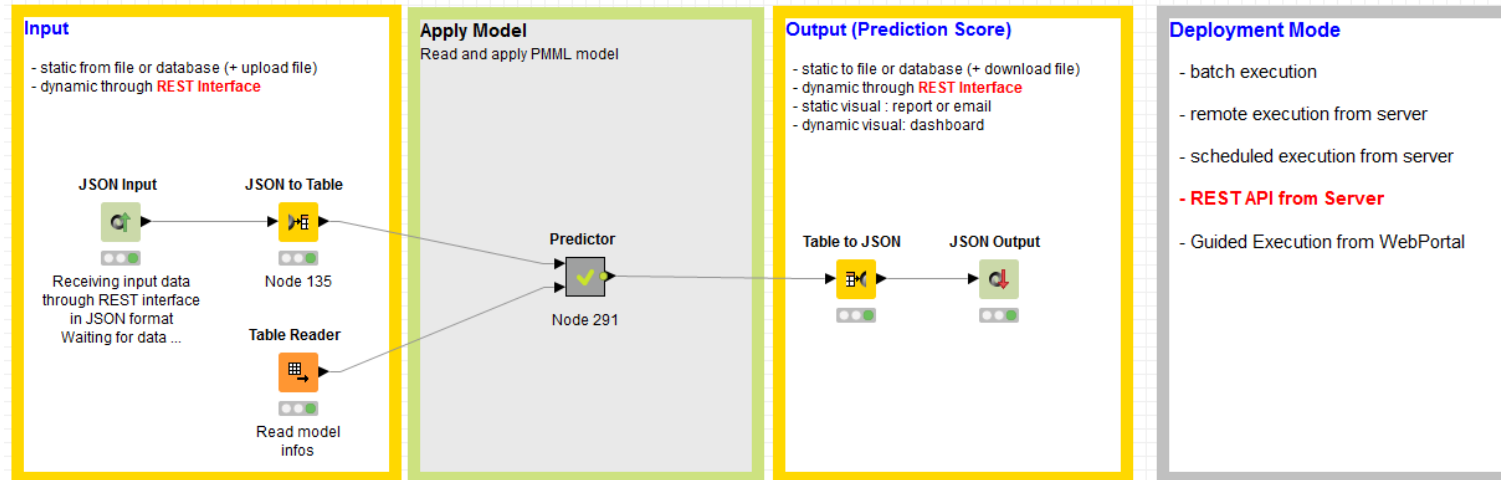
- **batch execution**
- **remote execution from server**
- **scheduled execution from server**
- REST API from Server
- Guided Execution from WebPortal

REST API (available on KNIME Server)

Model Deployment as REST API

This workflow:

- receives new unseen data via REST interface (JSON format),
- scores the data with the available current model,
- appends model prediction and probabilities to original data
- makes results available at the output REST interface



To Dashboard on WebPortal

Model Deployment Guided Execution on WebPortal

This workflow:

- uploads new unseen data from local file (.table or .csv format) from web browser,
- scores the data with the available current model,
- appends model prediction and probabilities to original data
- displays **web page** on web browser with description text, result data table, bar chart of class probabilities, and link to file download

Input

- static from **upload file**
- dynamic through REST Interface

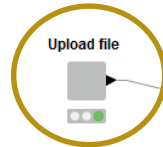


Table Reader



Read model
infos

Apply Model

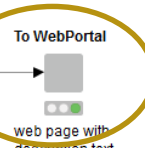
Read and apply PMML model

Predictor

Node 291

Output (Prediction Score)

- static to file or database
- dynamic through REST Interface
- static visual : report or email
- dynamic visual: **dashboard**



web page with
description text
data table
bar chart
link to file download

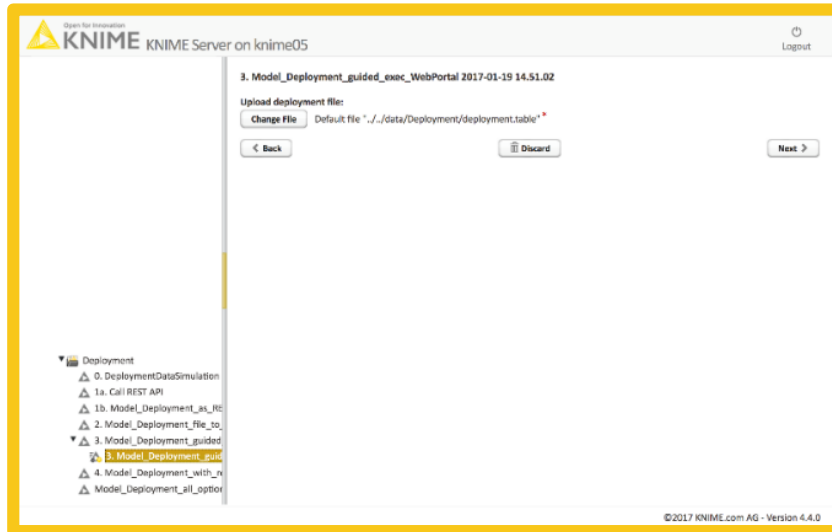
Deployment Mode

- batch execution
- remote execution from server
- scheduled execution from server
- REST API from Server
- **Guided Execution from WebPortal**

Step 1
Upload File

Step 2
Dashboard

Workflow on KNIME WebPortal



Open for Innovation
KNIME KNIME Server on knime05 Logout

3. Model_Deployment_guided_exec_WebPortal 2017-01-19 14.51.02

Upload deployment file:
Change File Default file: ../data/Deployment/deployment.table*

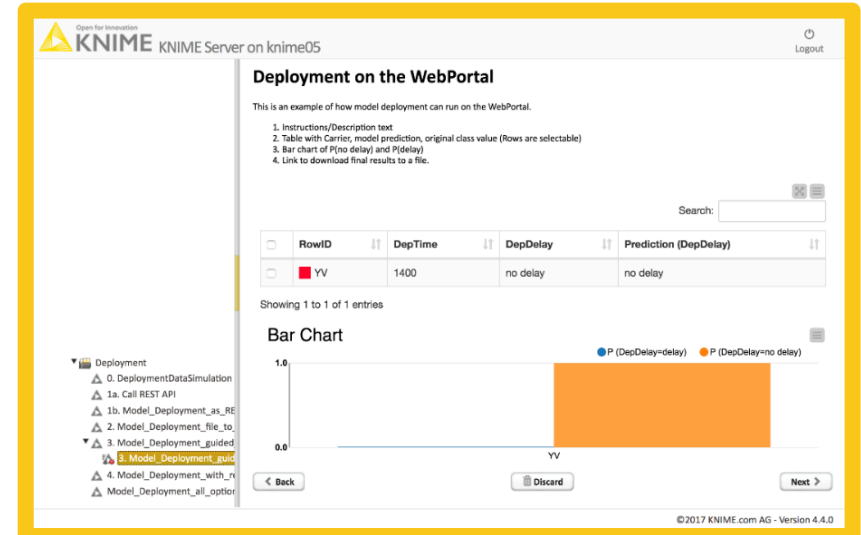
< Back Discard Next >

Deployment

- 0. DeploymentDataSimulation
- 1a. Call REST API
- 1b. Model_Deployment_as_RE
- 2. Model_Deployment_file_to
- 3. Model_Deployment_guided
- 3. Model_Deployment_guided**
- 4. Model_Deployment_with_n
- Model_Deployment_all_option

©2017 KNIME.com AG - Version 4.4.0

Step 1
Upload File



Open for Innovation
KNIME KNIME Server on knime05 Logout

Deployment on the WebPortal

This is an example of how model deployment can run on the WebPortal.

- Instructions/Description text
- Table with Carrier, model prediction, original class value (Rows are selectable)
- Bar chart of P(no delay) and P(delay)
- Link to download final results to a file.


Search:

<input type="checkbox"/>	RowID	DepTime	DepDelay	Prediction (DepDelay)
<input type="checkbox"/>	YV	1400	no delay	no delay

Showing 1 to 1 of 1 entries

Bar Chart

● P (DepDelay=delay) ● P (DepDelay=no delay)



< Back Discard Next >

Deployment

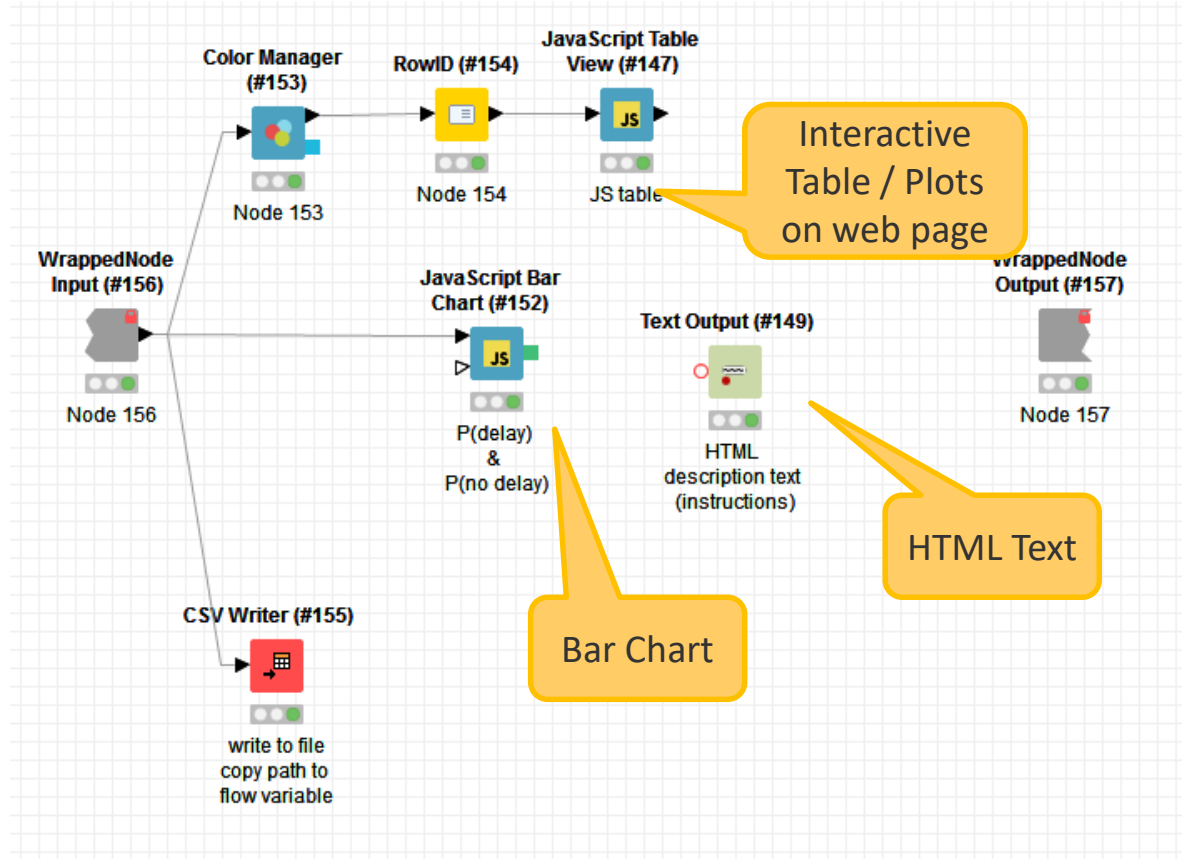
- 0. DeploymentDataSimulation
- 1a. Call REST API
- 1b. Model_Deployment_as_RE
- 2. Model_Deployment_file_to
- 3. Model_Deployment_guided
- 3. Model_Deployment_guided**
- 4. Model_Deployment_with_n
- Model_Deployment_all_option

©2017 KNIME.com AG - Version 4.4.0

Step 2
Dashboard

Available in
KNIME
Server

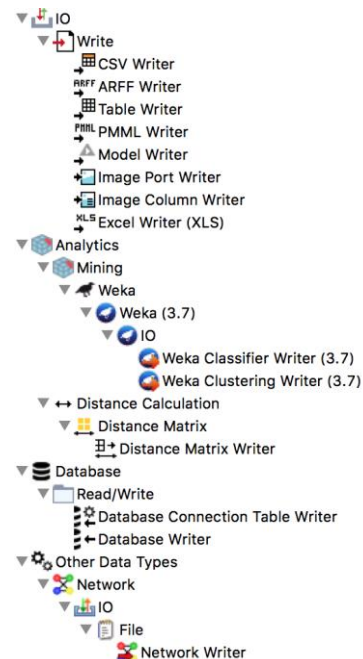
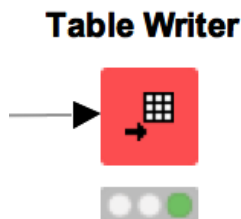
Wrapped Node to produce Dashboard on Web Page



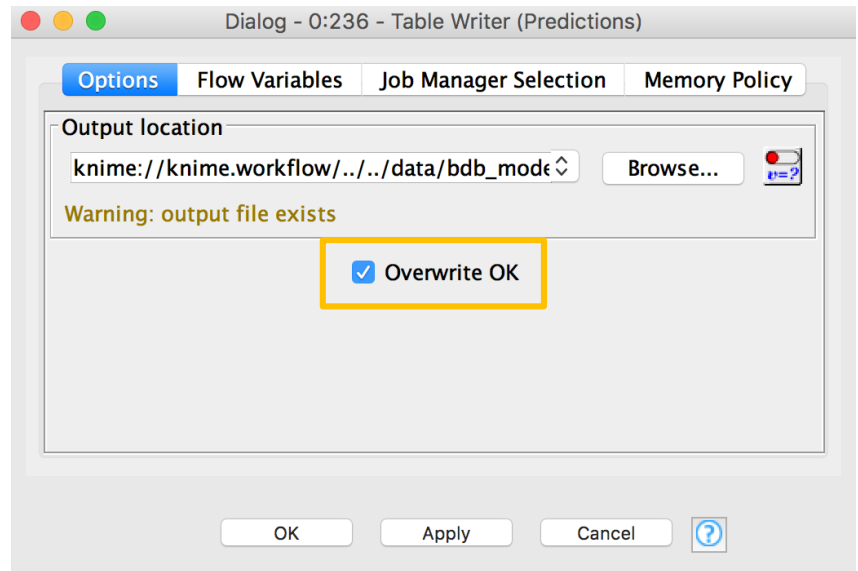
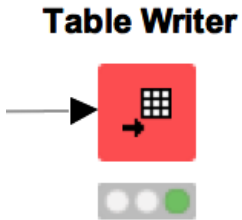
Data Export Nodes

Typically characterized by:

- Magenta color
- 1 input port, no output ports
- Create file on file system or write to database

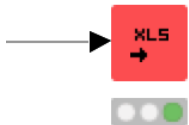


New Node: Table Writer

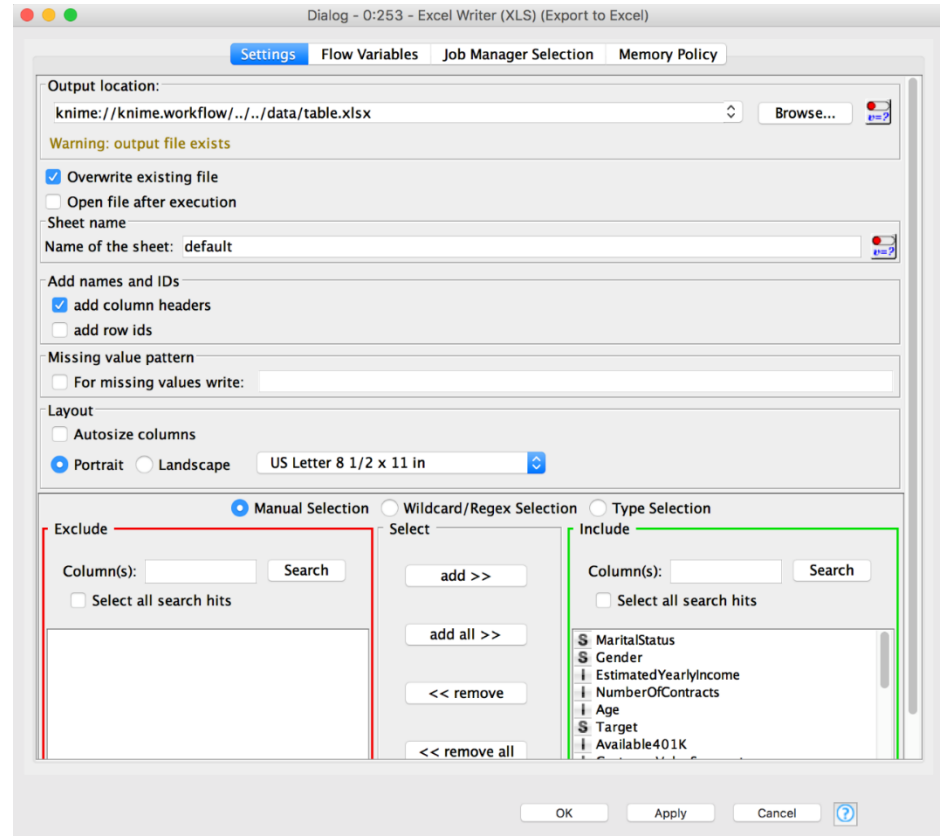


New Node: XLS Writer

Excel Writer (XLS)



Export to Excel



New Node: Database Writer

Only if no Connector node

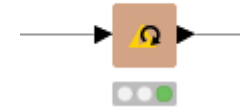
The diagram illustrates a workflow where data from a File Reader is sent directly to a Database Writer. The File Reader node is labeled 'new data'. The Database Writer node is labeled 'create table or append data'. A SQLite Connector node, labeled 'adult data set db', is shown but not connected to the Database Writer. A yellow callout bubble points to the Database Writer node with the text 'Only if no Connector node'.

The screenshot shows the 'Database Writer (create table)' dialog box. The 'Database Driver' is set to 'org.sqlite.JDBC'. The 'Database URL' is 'jdbc:sqlite://<host>:<port>/<database_name>'. The 'User Name' and 'Password' fields are empty. The 'Timezone' is set to 'Use local Timezone'. The 'Misc' section has 'Allow spaces in column names' checked and 'Append Data' checked with '... to existing table (if any!)' selected. The 'Table Name' is 'data_new'. The 'Append Data' section has 'Insert null for missing columns' unchecked. The dialog has 'OK', 'Apply', and 'Cancel' buttons.

Automation: Call Local Workflow

- Use Call Local Workflow node to send data and parameters to other workflows and trigger execution
 - Send results back to caller-workflow
 - Include report from called workflow
- Create modular workflows
 - E.g. separate workflows for ETL and prediction
- Alternative: Call Remote Workflow
 - Trigger execution of workflows on KNIME Server via REST API

Call Local Workflow



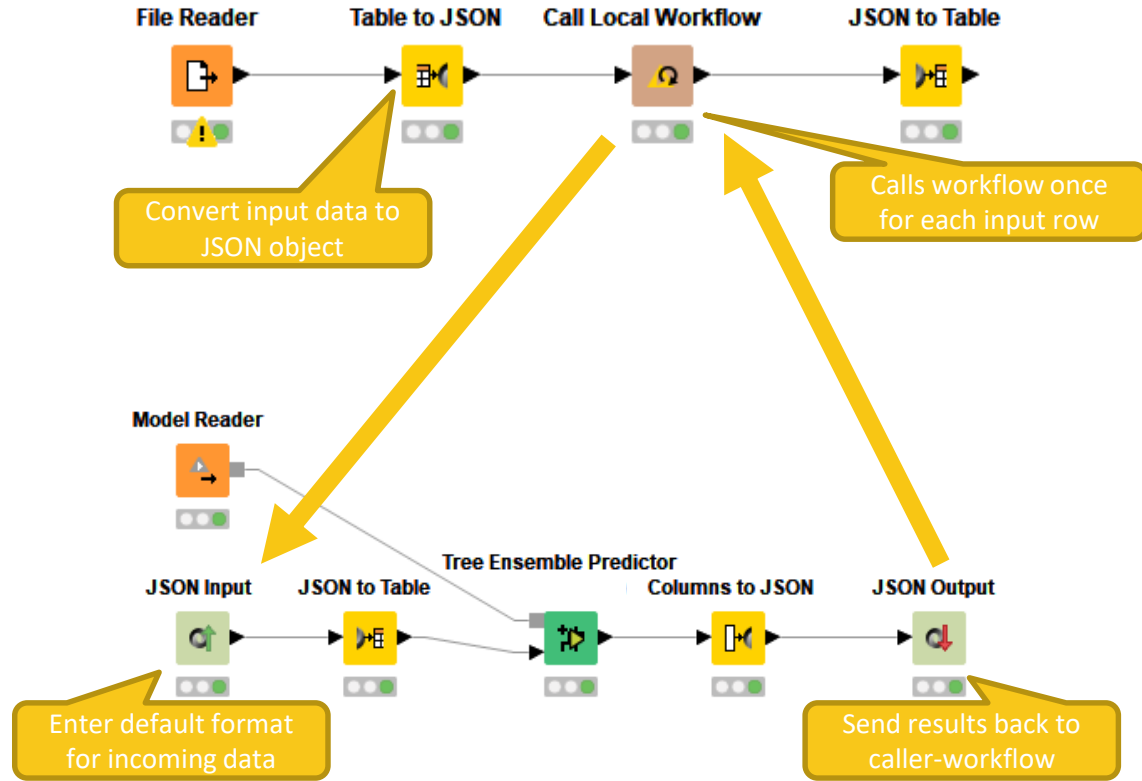
The screenshot shows the 'Call Local Workflow' dialog box with several callouts:

- Path to workflow**: Points to the 'Workflow Path' field, which contains the path `../workflows/_Predictor_Flow`.
- Add report to output**: Points to the 'Create Report' checkbox, which is checked, and the 'PDF' dropdown menu.
- Click to query the expected input(s)**: Points to the 'Load input format' button.
- Specify source column(s) with input data / parameters**: Points to the 'From column' radio button and the 'JSON' dropdown menu.

Automation: Call Local Workflow

ETL

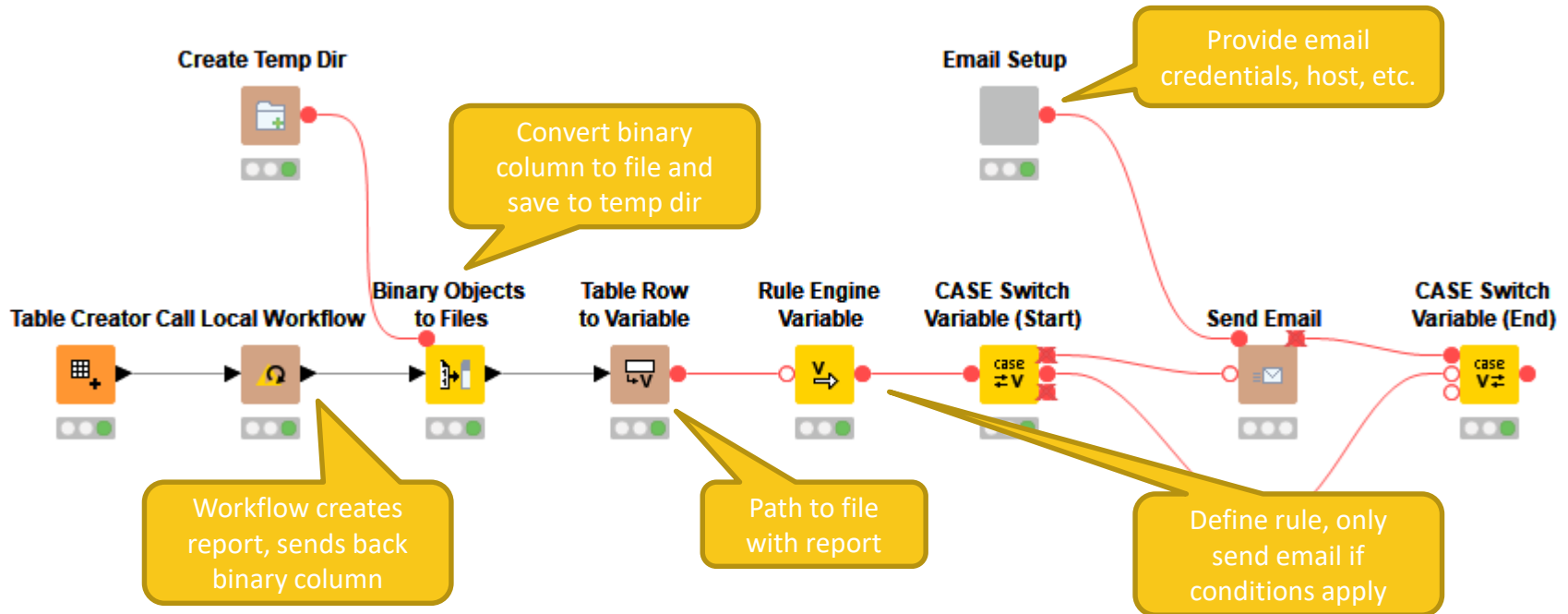
Prediction



Use Call Local to send conditional emails with report

Sometimes, report should be sent under specific circumstances

- E.g. if some KPI is below threshold

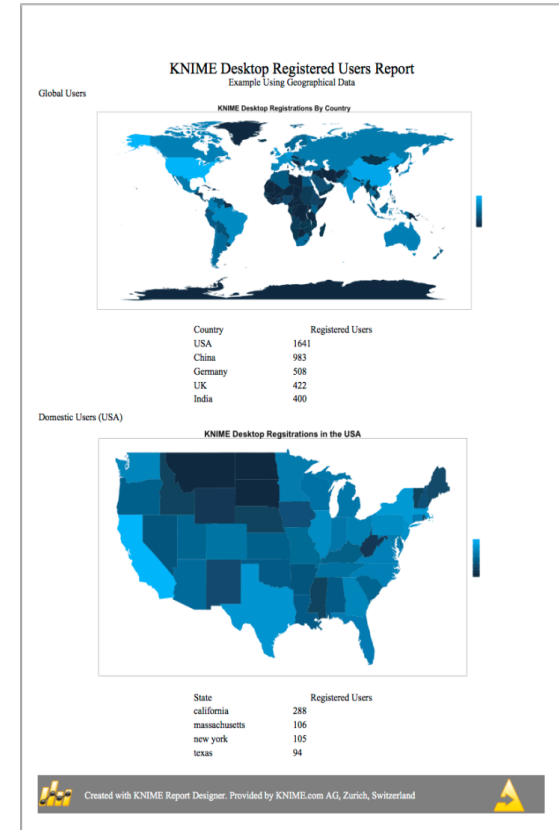


Reporting in KNIME



Reporting in KNIME

- Reporting in KNIME is done via a 3rd party application named BIRT (Business Intelligence Reporting Tool)
- Data is sent to BIRT from KNIME using special nodes.
- Reports in BIRT are constructed from report items, which may include images, tables, charts and labels.
- Reports may be generated in a variety of formats (html, pdf, pptx, xlsx, docx, ...)



Installation

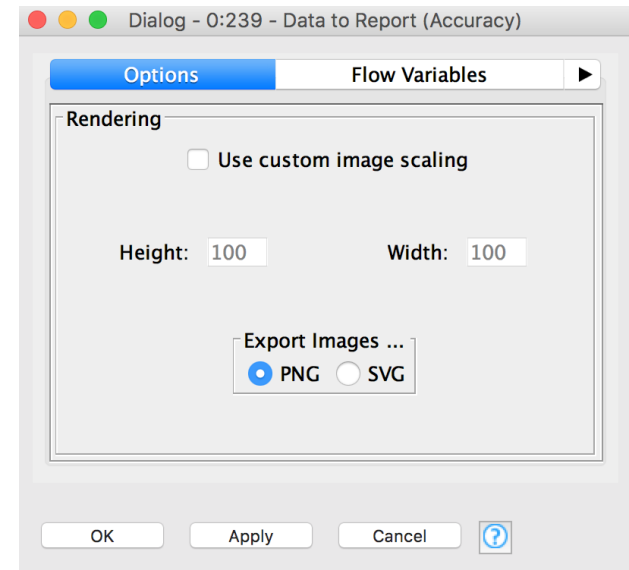
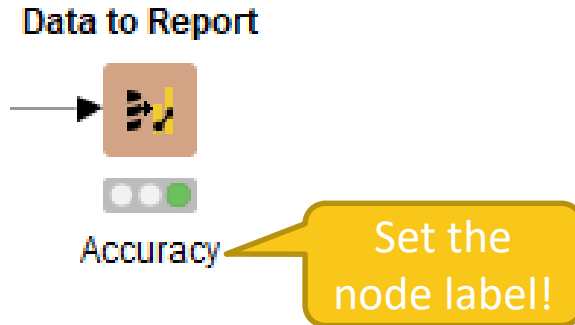
- Can be installed via KNIME -> Install KNIME Extension
- Install the two extensions below

- ▼  KNIME Report Designer
 - >  Eclipse BIRT Report Designer XML Tab Editor
 - >  KNIME Reporting Runtime

New Node: Data to Report

Send a data table to BIRT

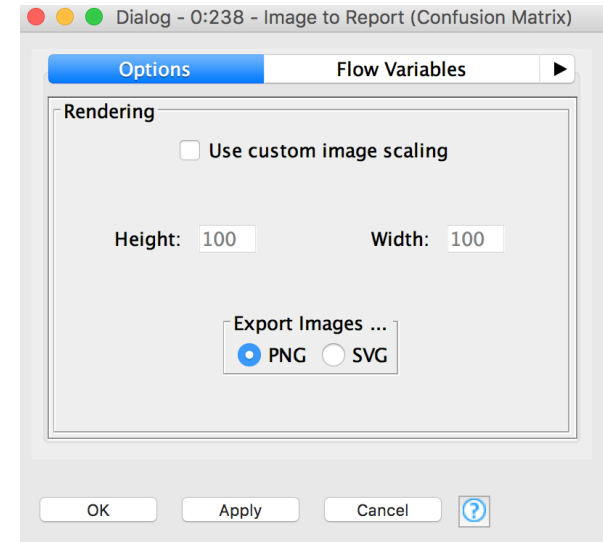
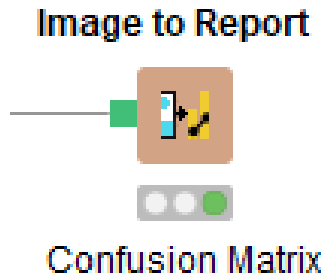
- Hint: The node label will be used to identify the data source in the reporting view -> Make sure to use fitting labels if you have more than one data source



New Node: Image to Report

Send an image to BIRT

- PNG and SVG are supported formats (see node description for details)



Edit the Report

Open the workflow > Click the Report Editor button in the tool bar

The screenshot displays the KNIME Analytics Platform interface. On the left, the 'KNIME Explorer' pane shows a project tree with 'solutions' > '08. Exporting Data' selected. A yellow arrow points to the 'Report Editor' icon in the top toolbar. The main workspace shows a workflow diagram with nodes: Fully Joined Data, Partitioning, Decision Tree Learner (Node 250), Decision Tree Predictor (Predict unknown data), Scorer (Accuracy and Confusion Matrix), Normalizer, HeatMap (JFreeChart), Image to Report, Confusion Matrix, Row Filter, Column Filter, Data to Report, Accuracy, PMML Writer, and Table Writer. A blue box highlights the report editor area, which contains a text-based report titled 'Solution: Exporting Data' with instructions like 'Write predictions to disk as a KNIME table' and 'Generate a PDF of your report.' The bottom console shows a log of warnings and errors, including 'WARN Missing value' and 'ERROR Database Update'.

Reporting Perspective

The screenshot shows the KNIME Reporting Perspective interface. The main window displays a report titled "My Data Mining Report" with a table structure. The interface includes a menu bar (Data, Page, View, Run, Help), a toolbar, a left sidebar with a project tree and a palette, and a bottom panel with tabs (Layout, Master Page, Script, XML Source) and a Properties editor.

Click button to create report: A callout points to a button in the top toolbar.

Data from KNIME - names of data sources are taken from node label: A callout points to the "Data set view" pane showing a tree structure of data sources like "Confusion Matrix" and "Accuracy".

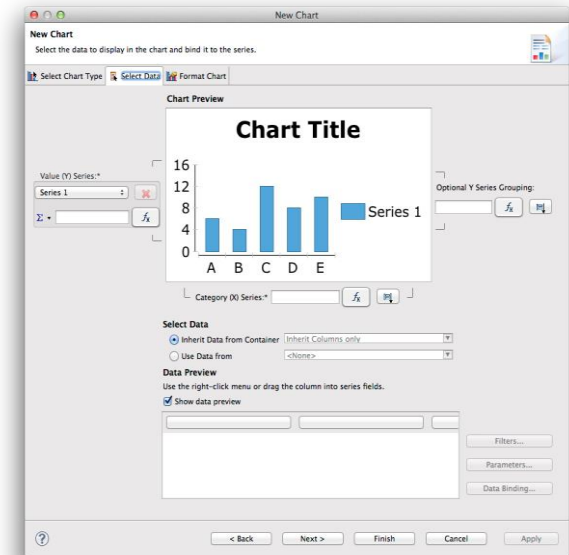
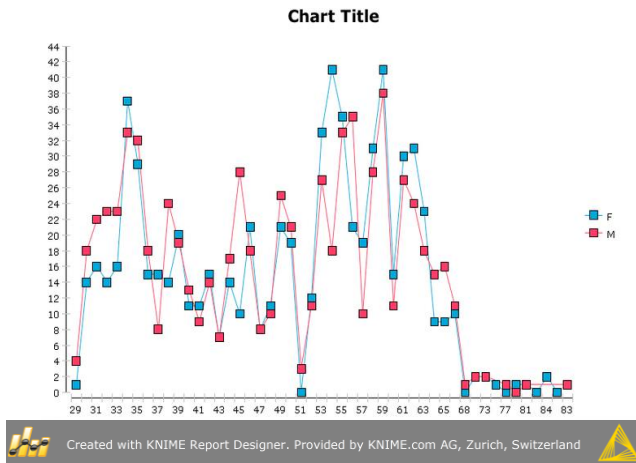
Add report items via drag and drop: A callout points to the "Report Items" palette on the left sidebar.

Report layout – only structure, data is filled in when creating the report: A callout points to the main report content area.

View tabs: A callout points to the tabs at the bottom of the interface.

Charting in BIRT

- Many chart types
- Fine control of plot appearance
- Familiar 'Excel Like' interface
- Supports interactivity



Tips & Tricks

- Use a underlying grid to structure the report
- Names of columns should not change
- Use the grouping function to combine results
- Use the Master Layout Tab (For footers etc.)

Exporting Data Exercise

Start with exercise: *Exporting Data*

- Send heatmap to report via Image to Report node
- Send model accuracy table via Data to Report node
- Create a report that includes the following elements:
 - A report title
 - A table with the model accuracy
 - The heatmap image
- Generate a PDF of your report

The End

education@knime.com