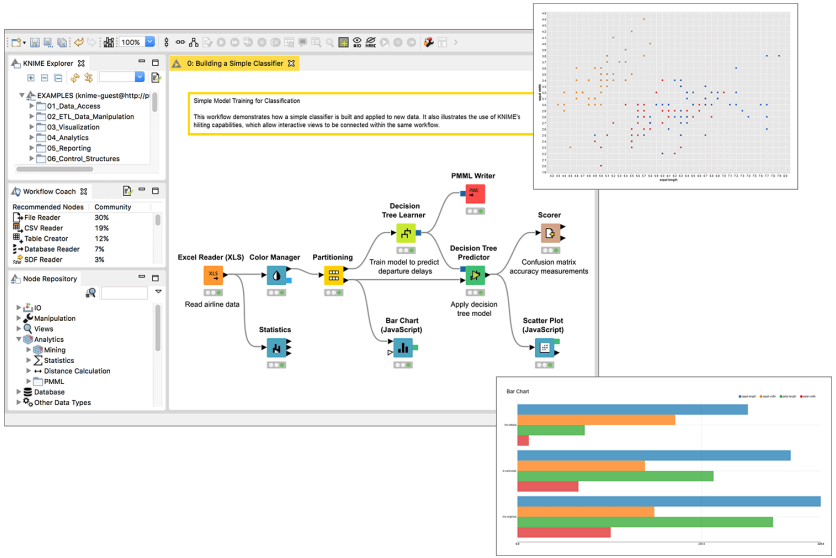

# [L1-DS] KNIME Analytics Platform for Data Scientists: Basics

KNIME AG

# Overview
# KNIME Analytics Platform

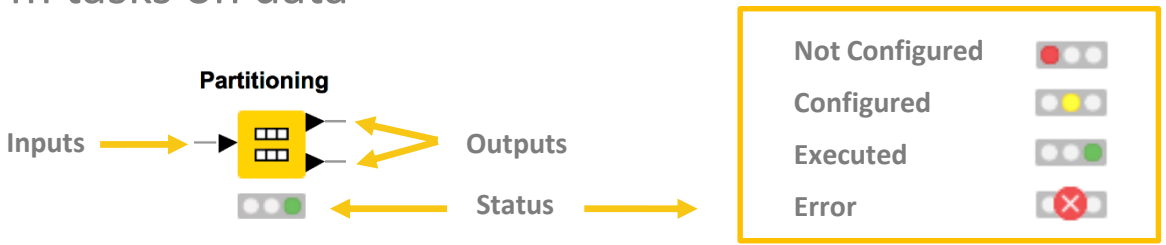
# What is KNIME Analytics Platform?

- 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, Keras, Plotly, H2O, etc.

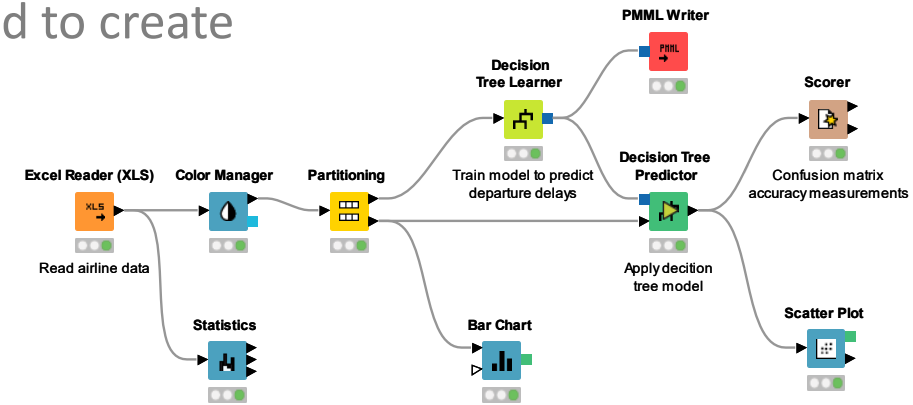


# Visual KNIME Workflows

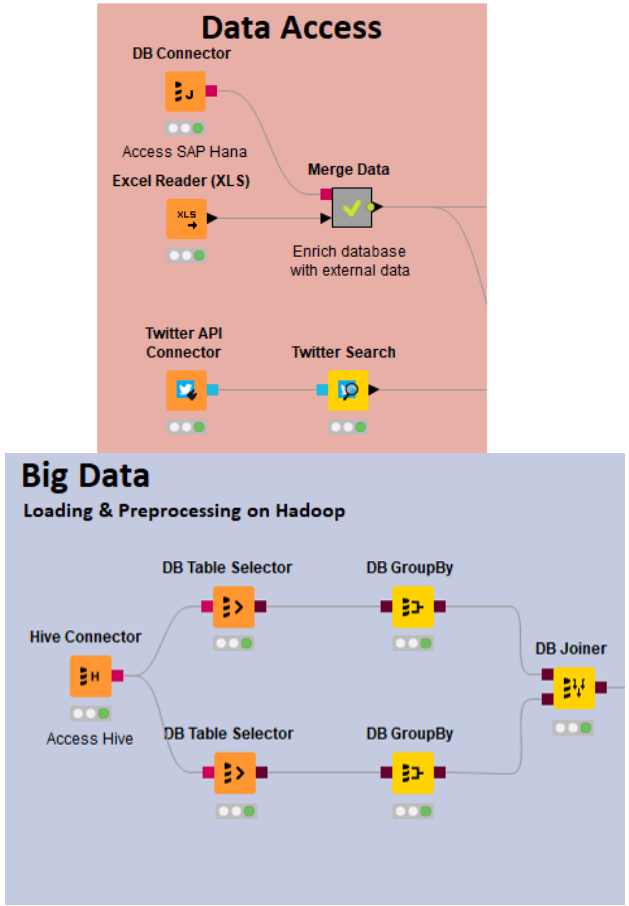
## **NODES** perform tasks on data




## Nodes are combined to create **WORKFLOWS**

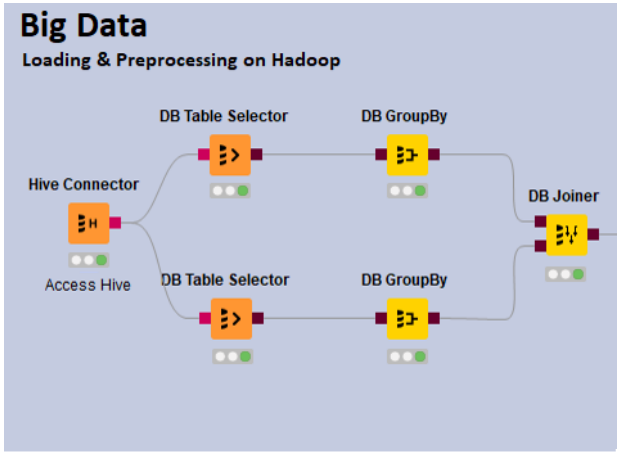


# Data Access



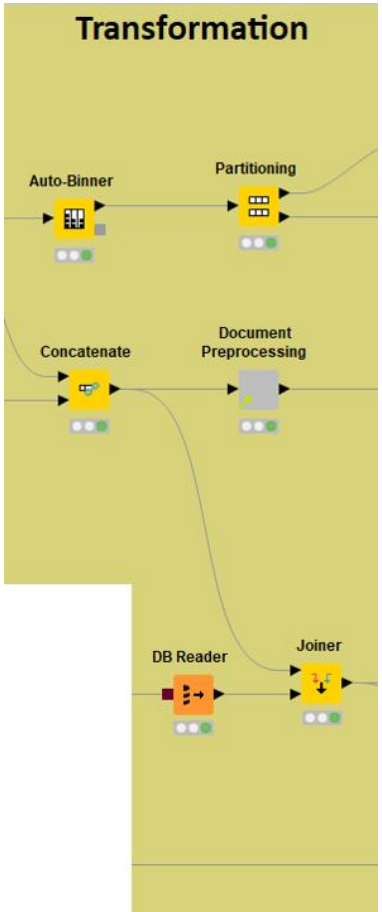

- Databases
  - MySQL, PostgreSQL
  - any JDBC (Oracle, DB2, MS SQL Server)
- Files
  - CSV, txt
  - Excel, Word, PDF
  - SAS, SPSS
  - XML
  - PMML
  - Images, texts, networks, chem
- Web, Cloud
  - REST, Web services
  - Twitter, Google

# Big Data



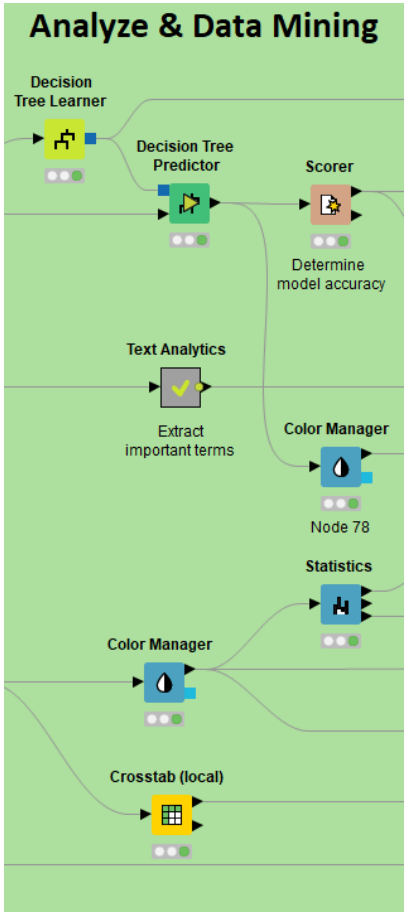

- Spark & Databricks
- HDFS support
- Hive
- Impala
- In-database processing

# Transformation



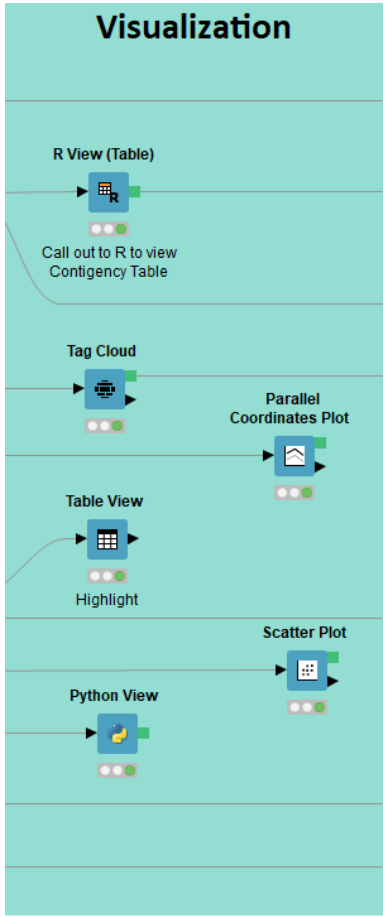

- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature Creation and Selection

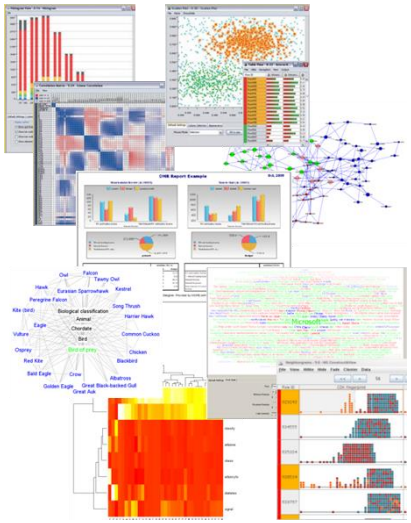
# Analysis & Data Mining




- 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, Keras

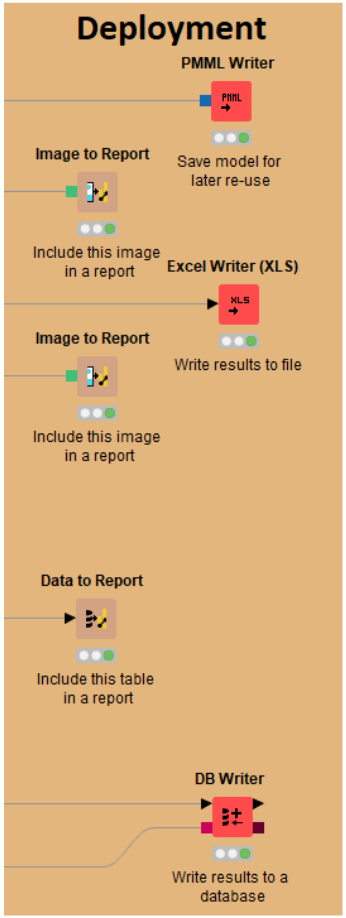
# Visualization



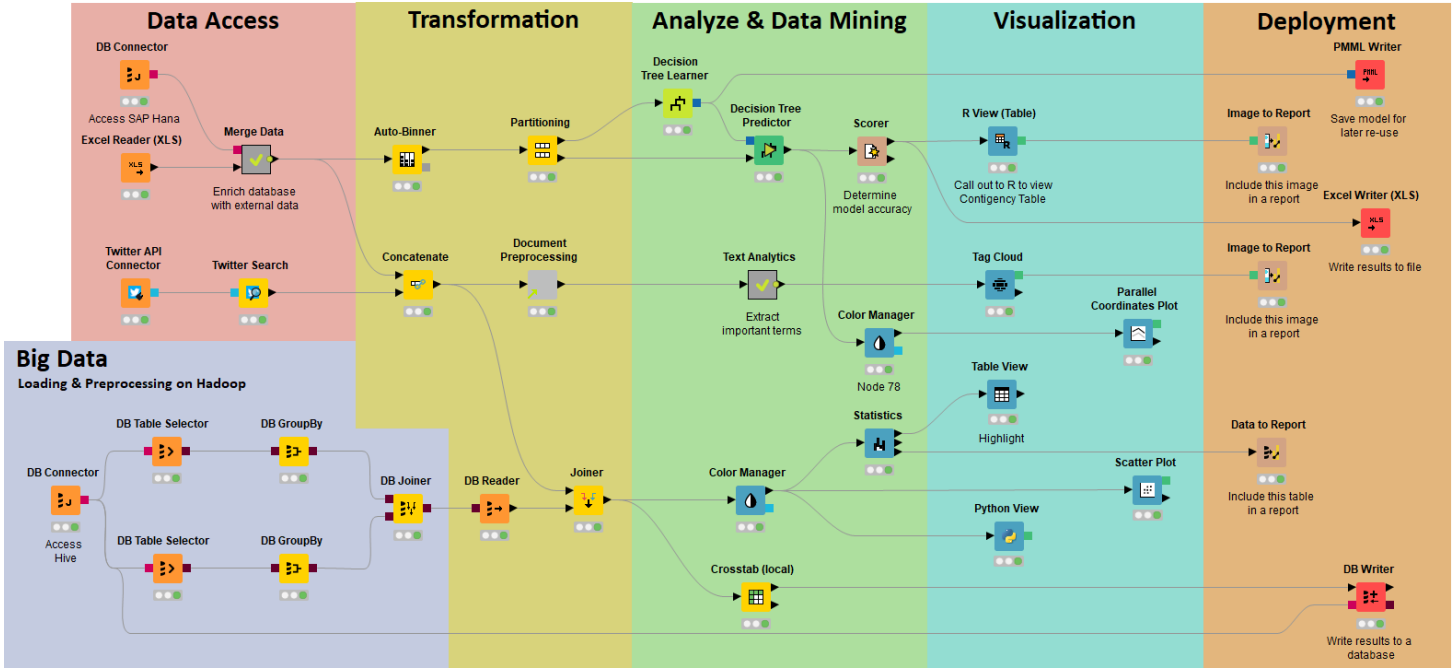
- Interactive Visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

# Deployment




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

# Over 2000 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
Plotly
Community / 3rd

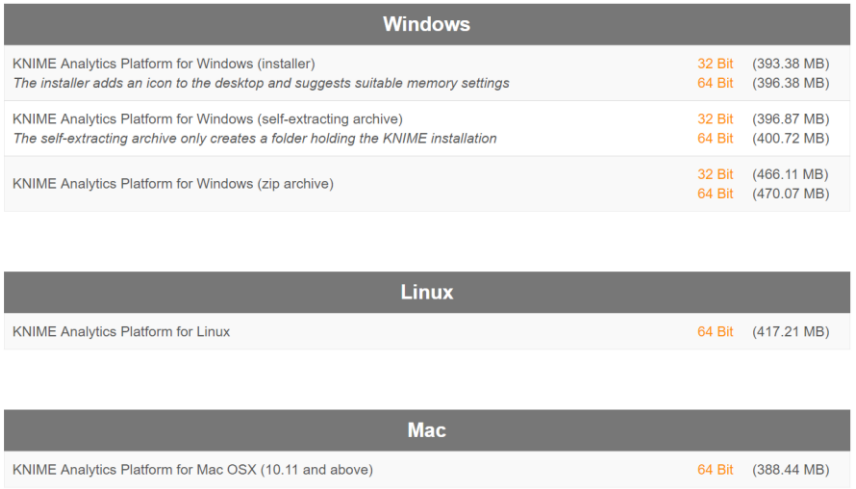
**Deployment**
via BIRT
PMML
XML, JSON
Databases
Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd

# Overview

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

# Install KNIME Analytics Platform

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



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

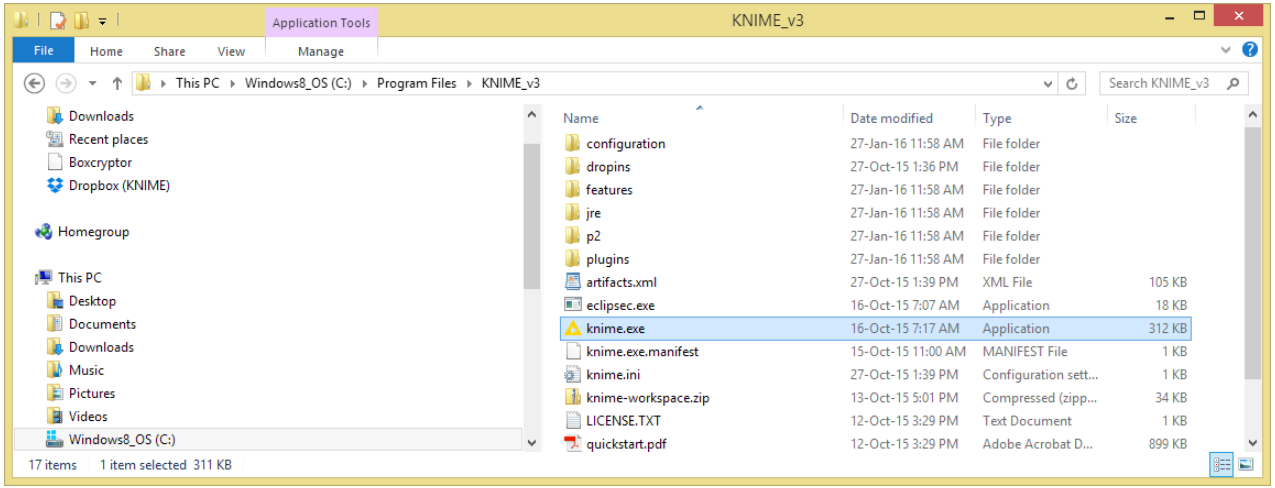
| Linux | | |
|---|---|---|
| KNIME Analytics Platform for Linux | 64 Bit | (417.21 MB) |

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

# Start KNIME Analytics Platform

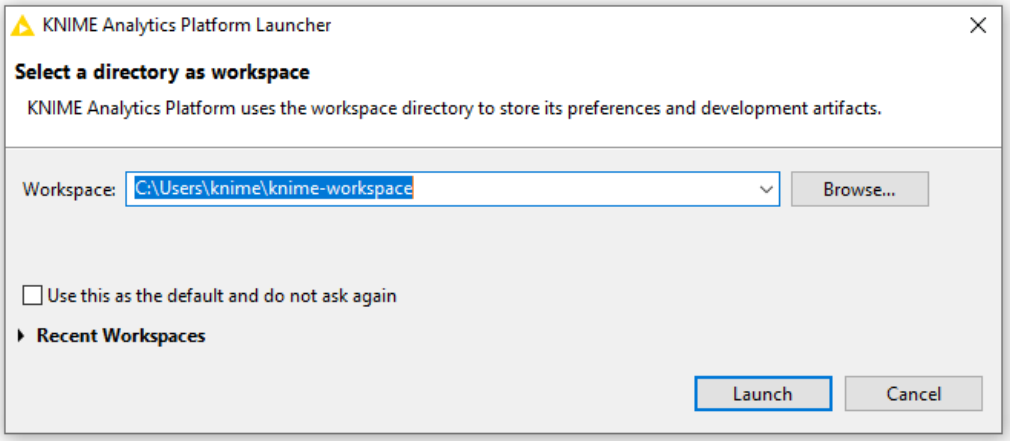
- Use the shortcut created by the installer


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

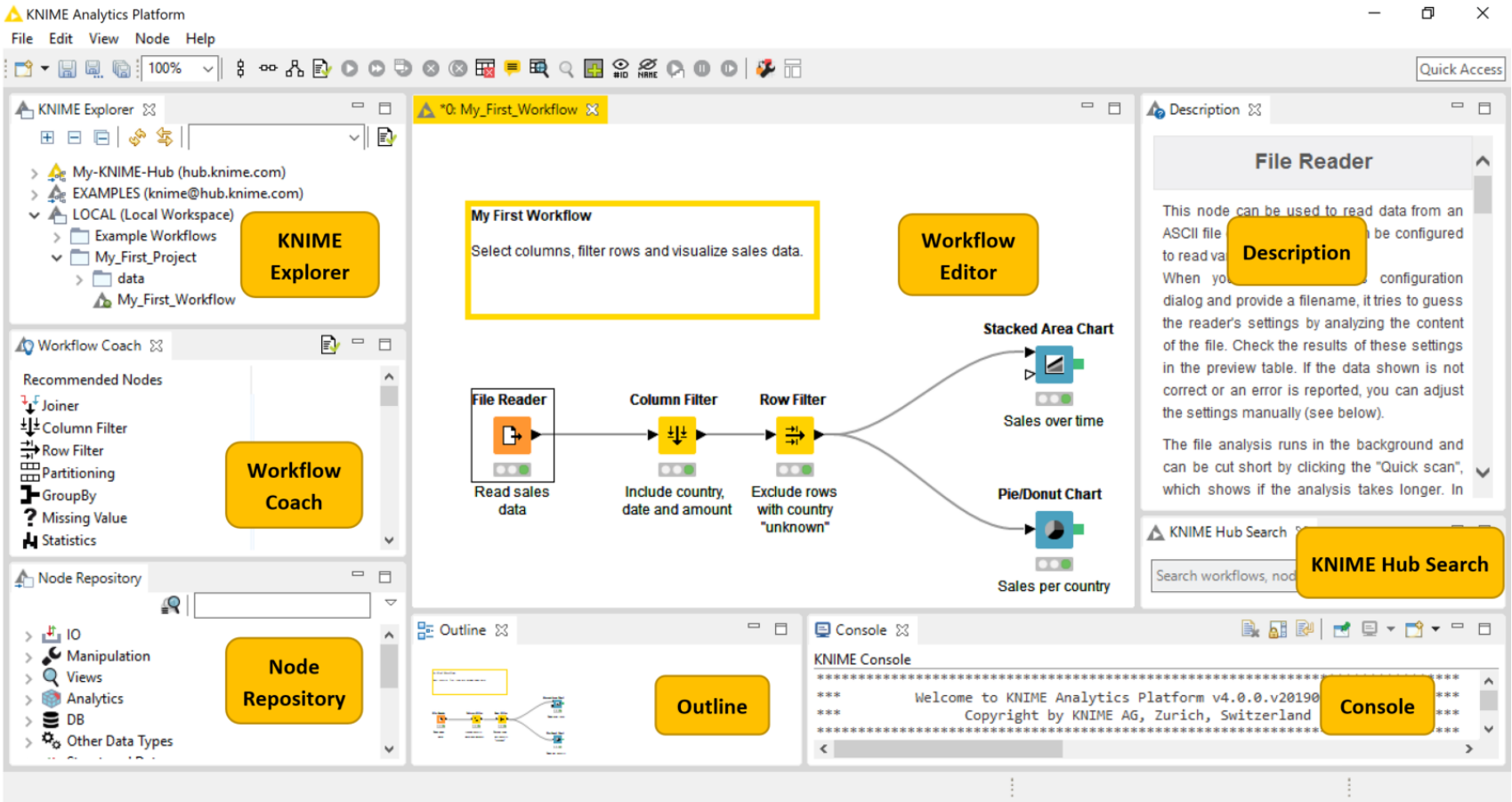


# The KNIME Workspace

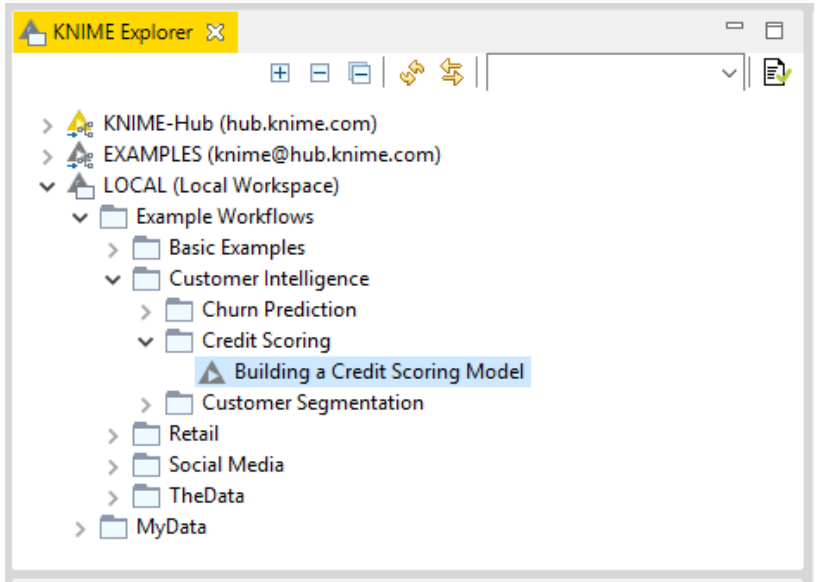
- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session.
- Workspaces are portable (just like KNIME)



# The KNIME Workbench



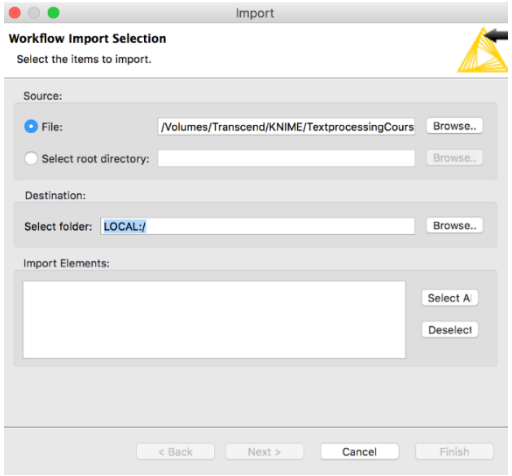
# KNIME Explorer

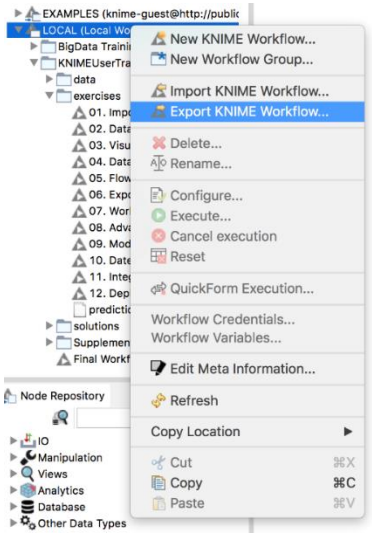


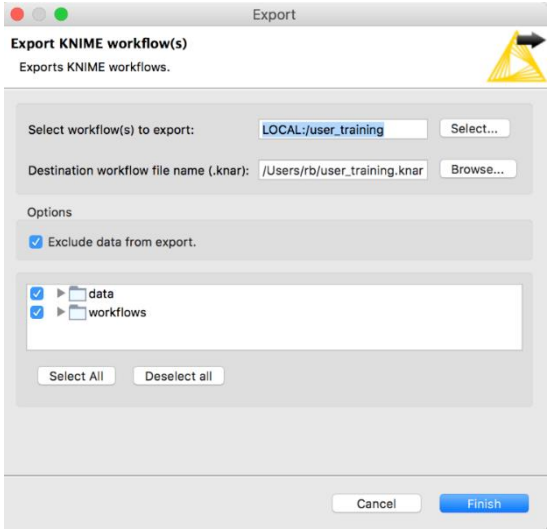

- In LOCAL you can access your own workflow projects.
- The Explorer toolbar on the top has a search box and buttons to
  - select the workflow displayed in the active editor
  - refresh the view
- The KNIME Explorer can contain 4 types of content:
  - Workflows
  - Workflow groups
  - Data files
  -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

# Creating New Workflows, Importing and Exporting

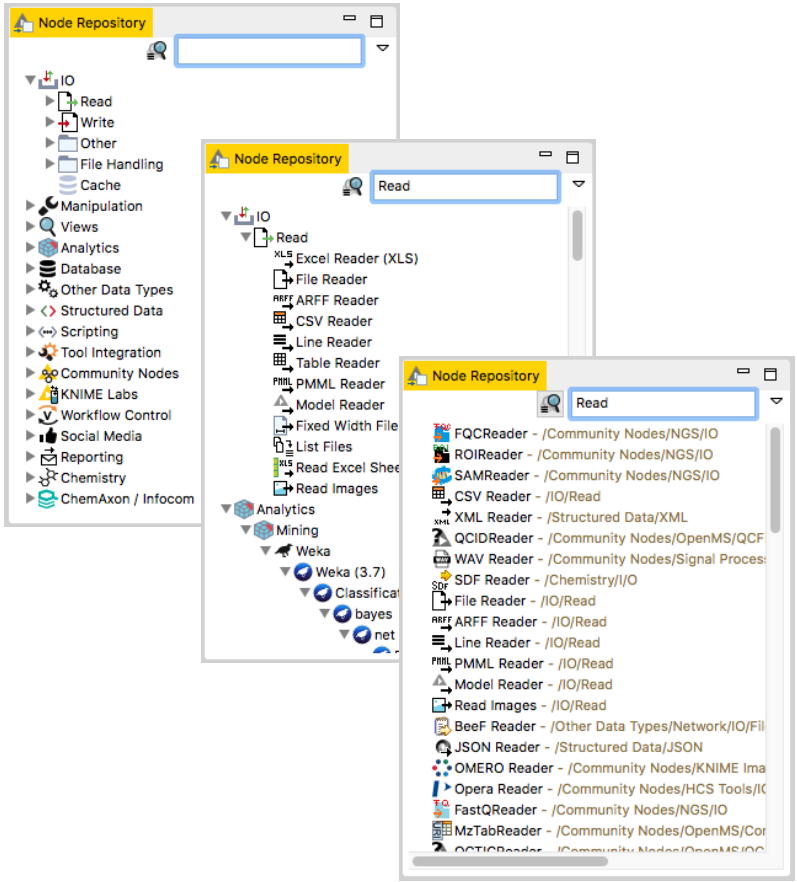
- Right-click in KNIME Explorer to create new workflow or workflow group or to import workflow
- Right-click on workflow or workflow group to export



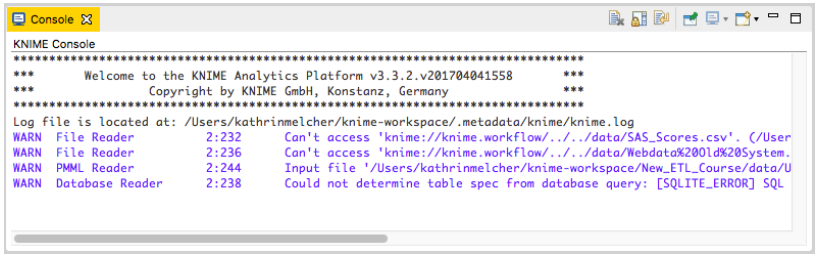



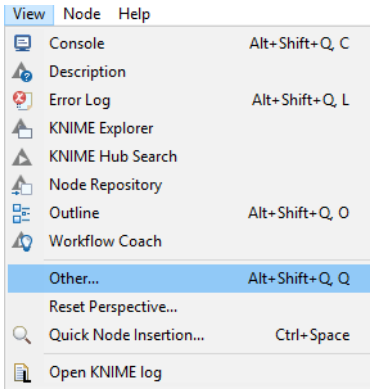
# Node Repository

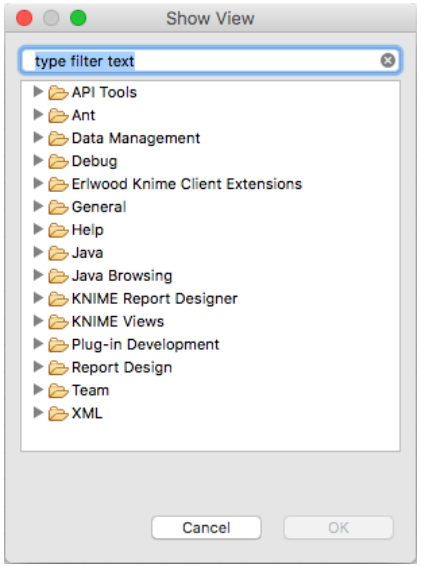



- The Node Repository lists all KNIME nodes
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name
- Nodes can be added by drag and drop from the Node Repository to the Workflow Editor.

# Console and Other Views

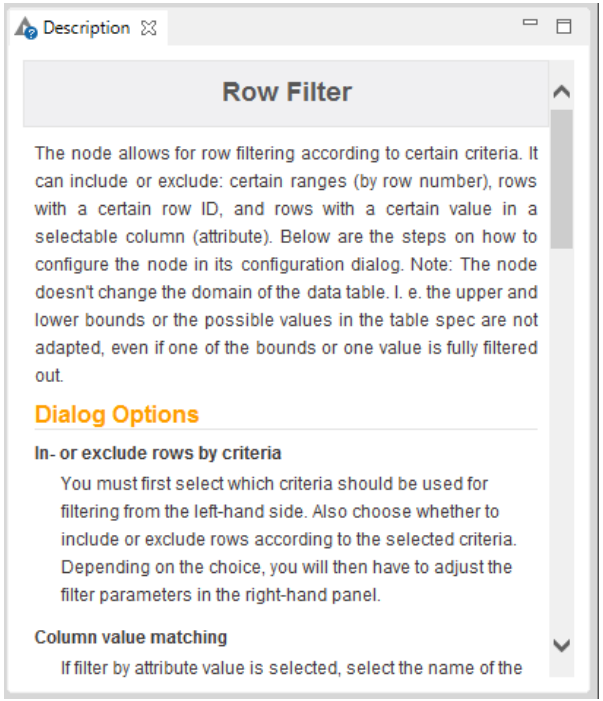


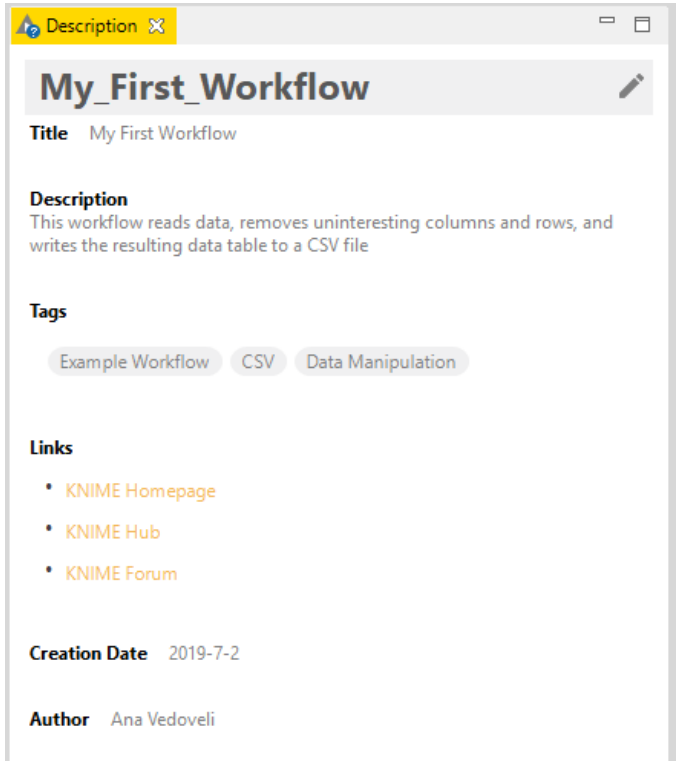

- Console view prints out error and warning messages about what is going on under the hood
- Click on View and select Other… to add different views
  - Node Monitor, Licenses, etc.

# Description




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

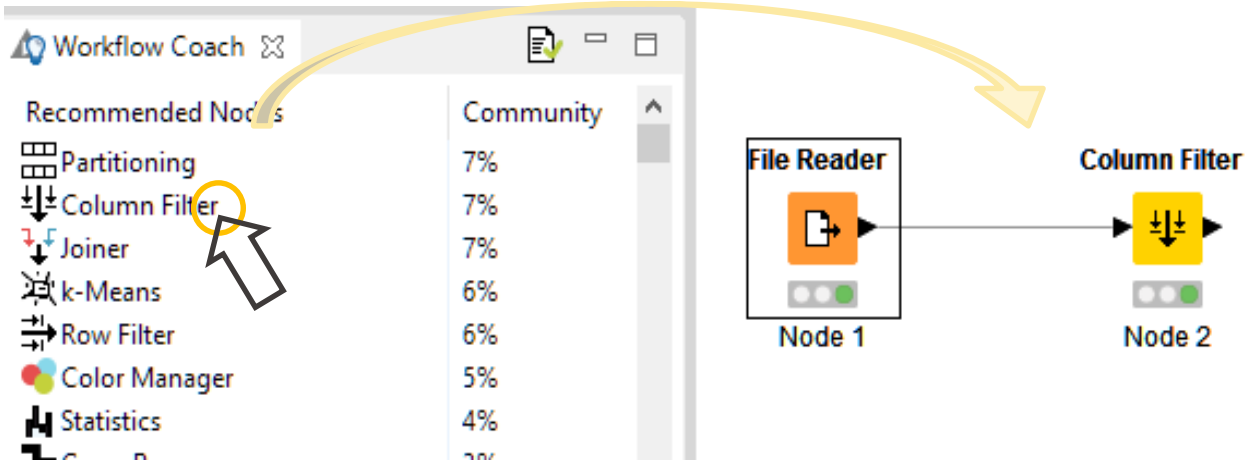
# Workflow Description




- When selecting the workflow, the Description window gives information about the workflow's:
  - Title
  - Description
  - Associated Tags and Links
  - Creation Date
  - Author

# Workflow Coach

- Node 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



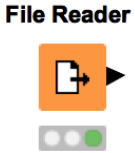

- ***.knar** for KNIME Archive Files

# More on Nodes…

A node can have 3 states:


**Not Configured**:
The node is waiting for configuration or incoming data.

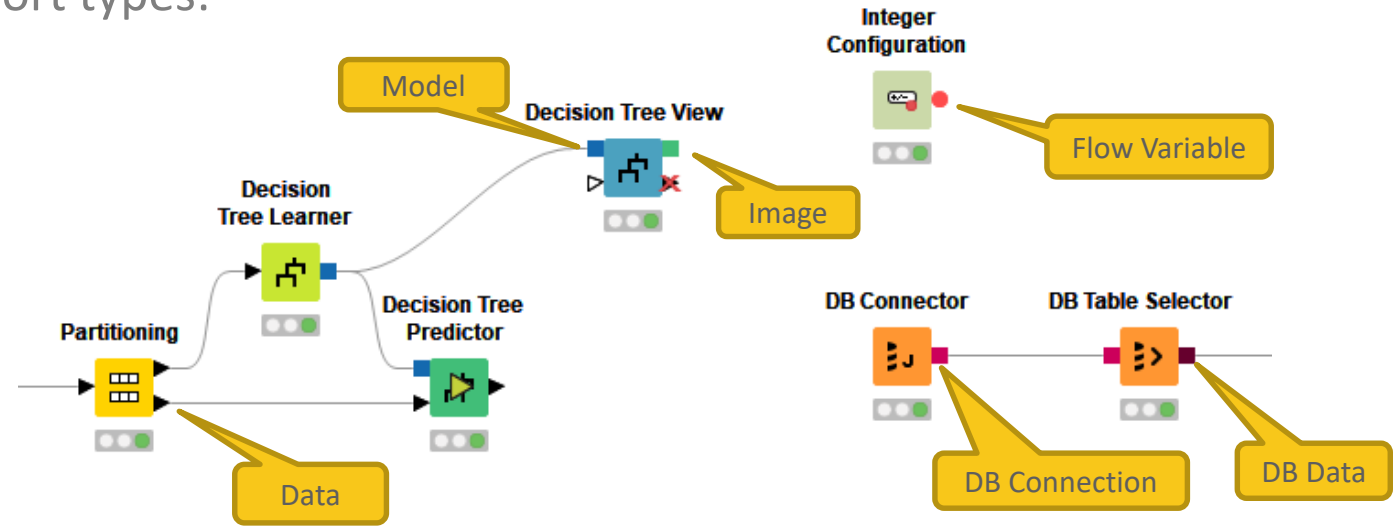

**Configured**:
The node has been configured correctly, and can be executed.




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

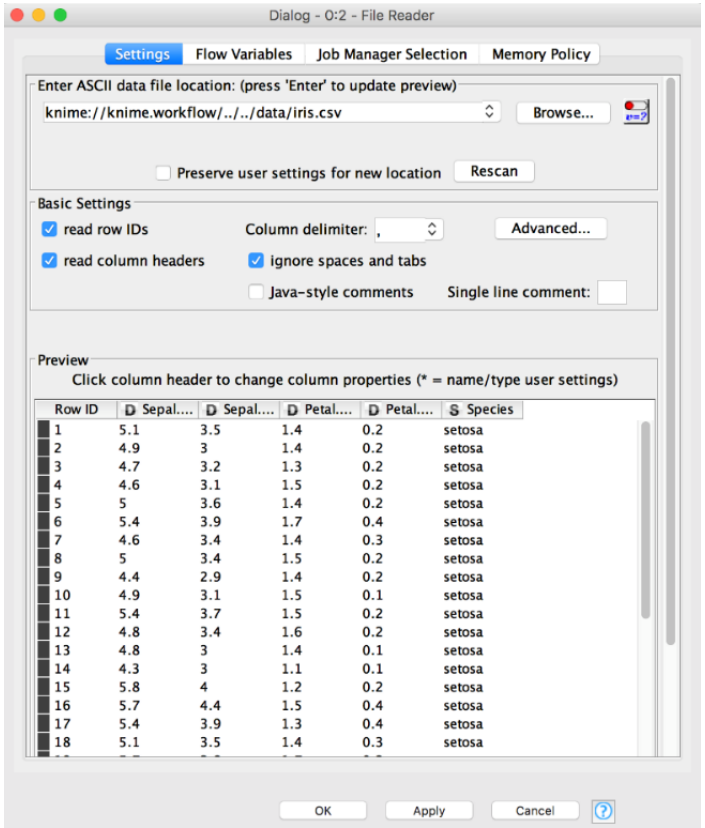
# Inserting and Connecting Nodes

- 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:



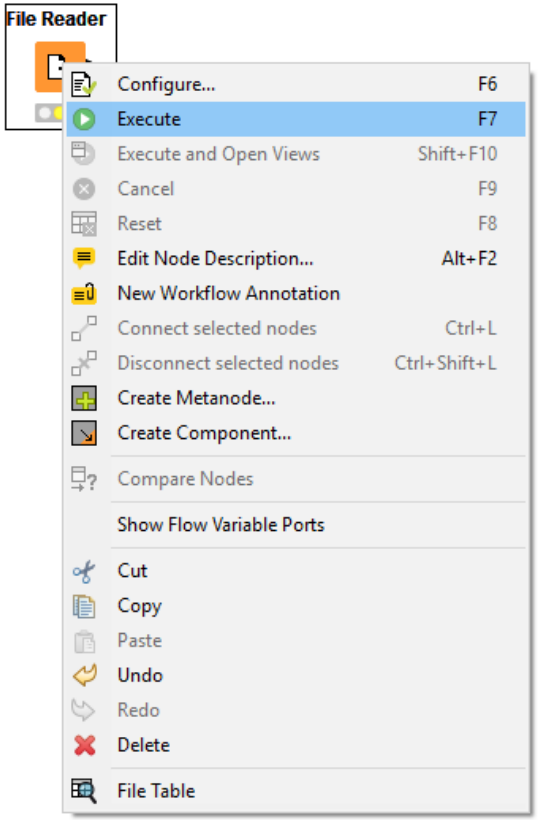
# Node Configuration

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



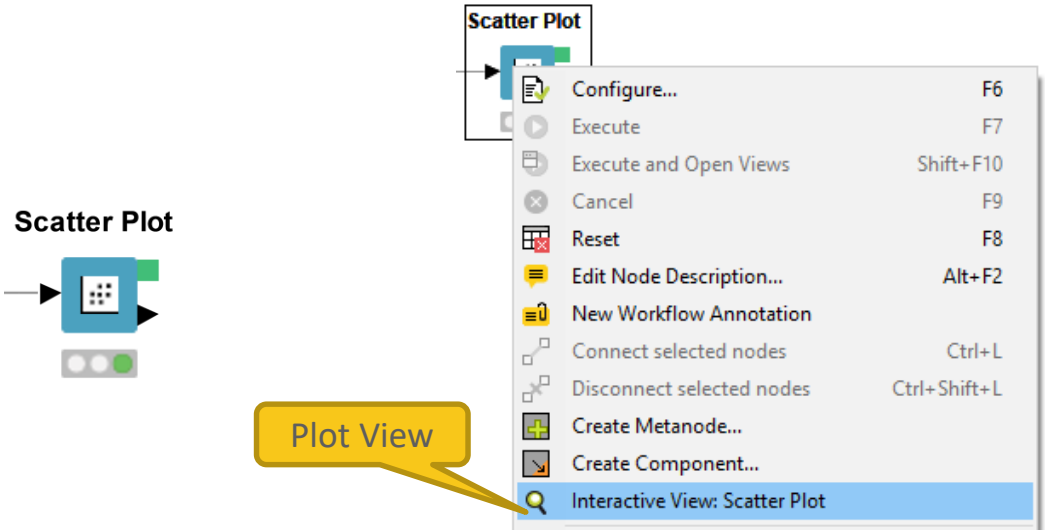
# Node Execution

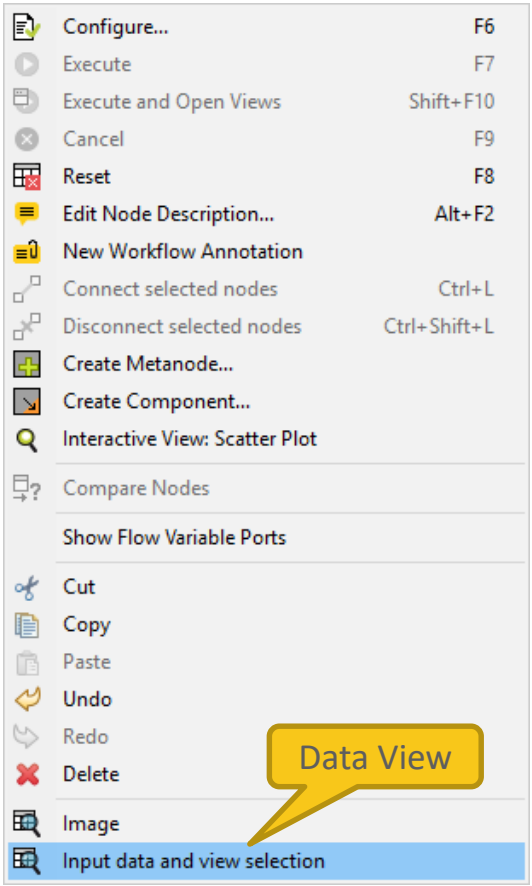
- Right-click node
- Select Execute in the context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



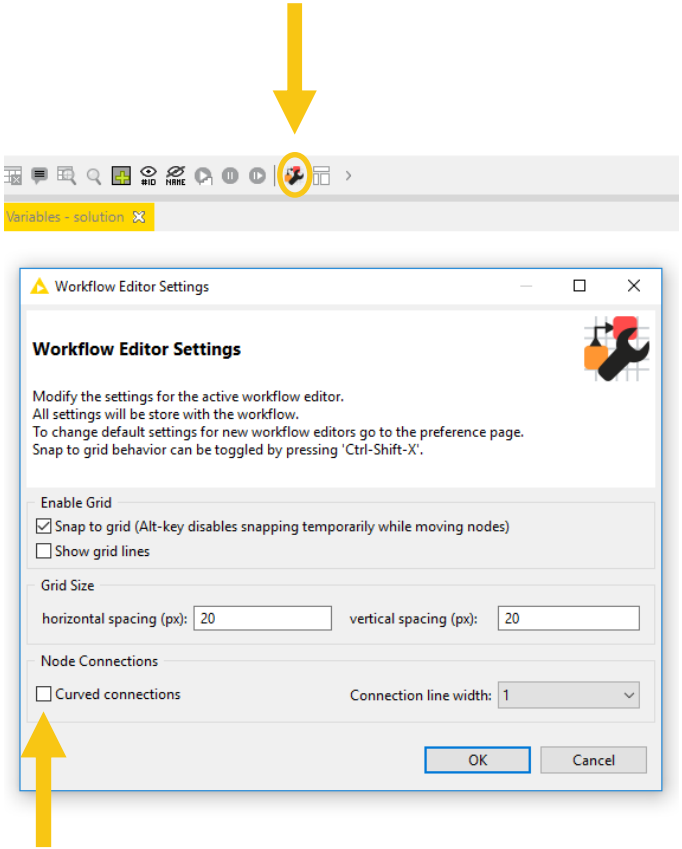
# Node Views

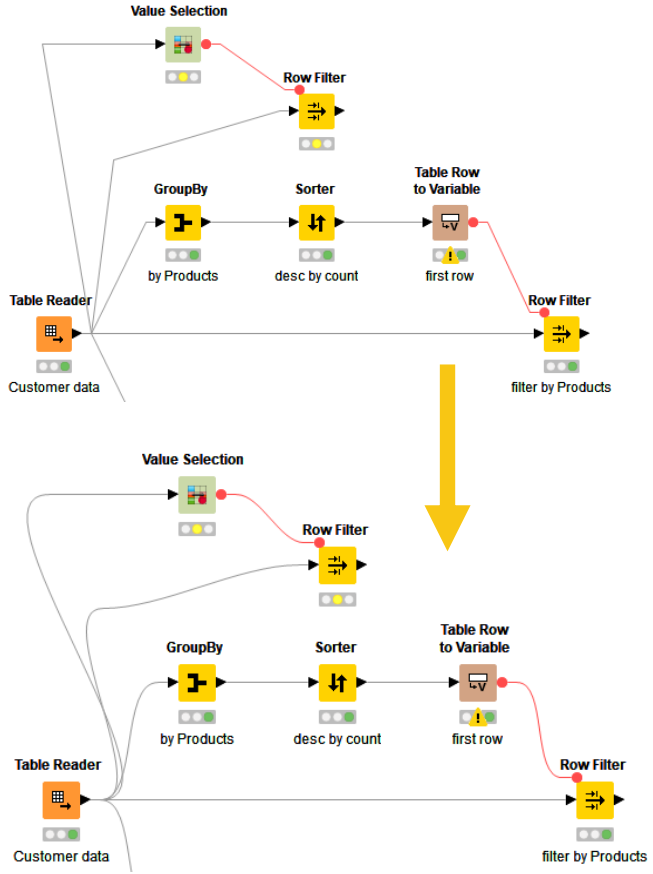
- Right-click node
- Select Views in context menu
- Select output port to inspect execution results



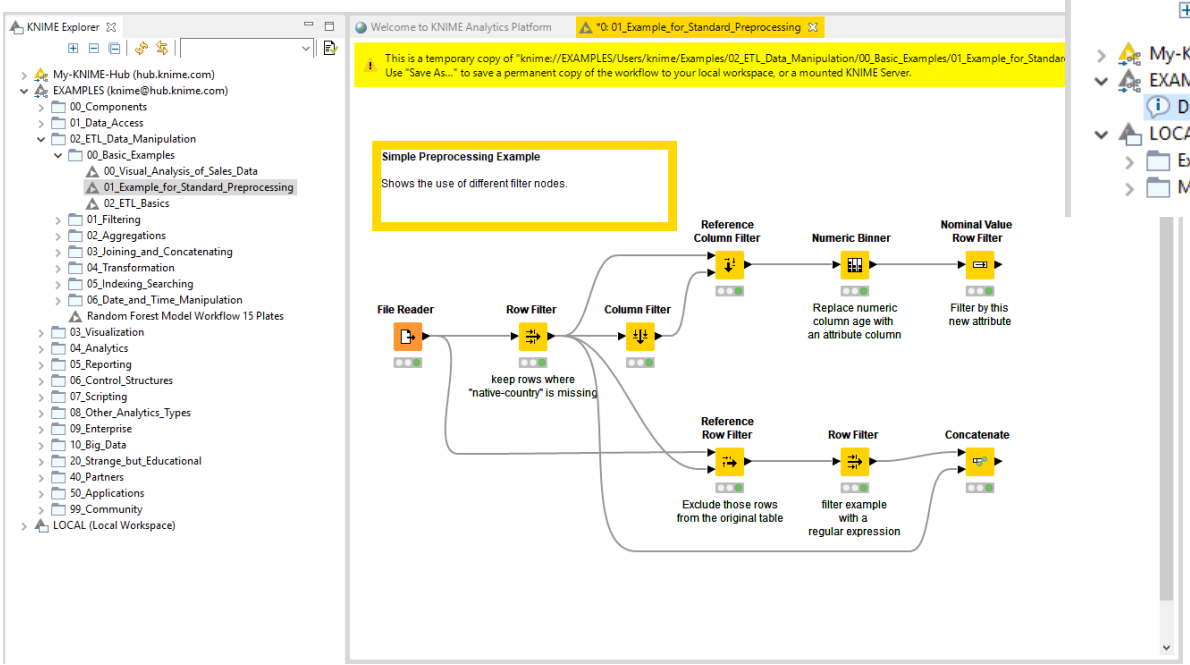
# Curved Connections!



# Getting Started: KNIME Example Server

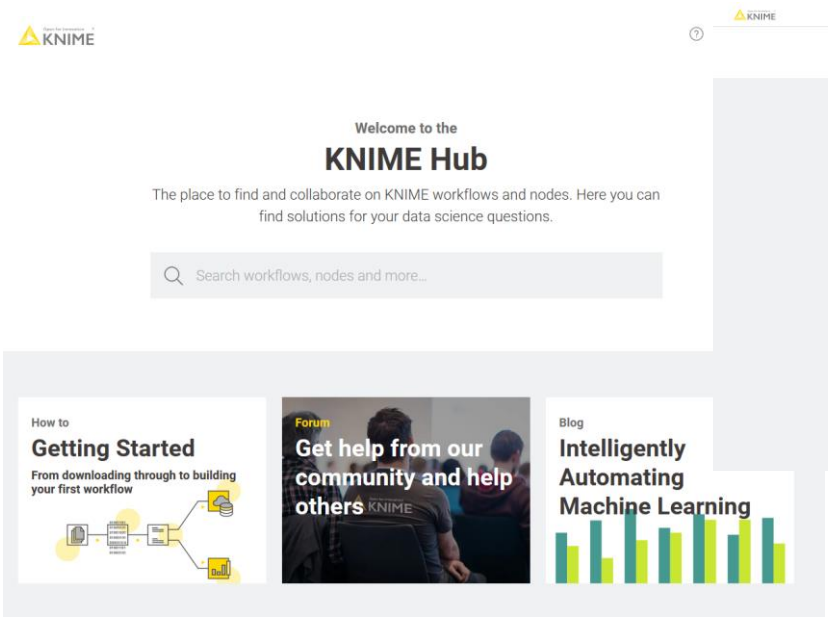
- Connect via KNIME Explorer to a public repository with large selection of example workflows for many, many applications
- Workflows also available on KNIME Hub

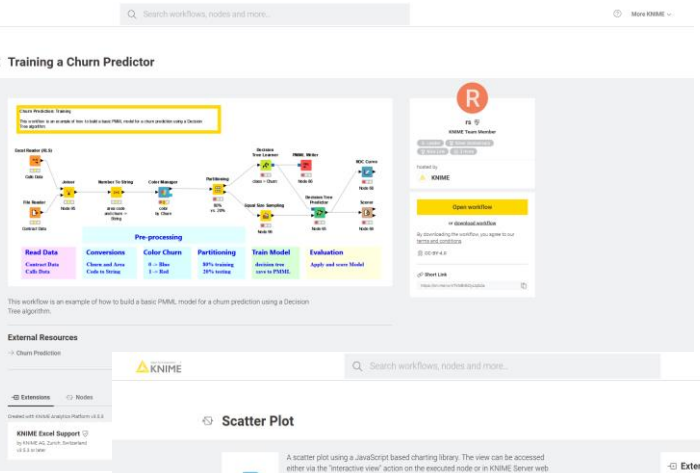


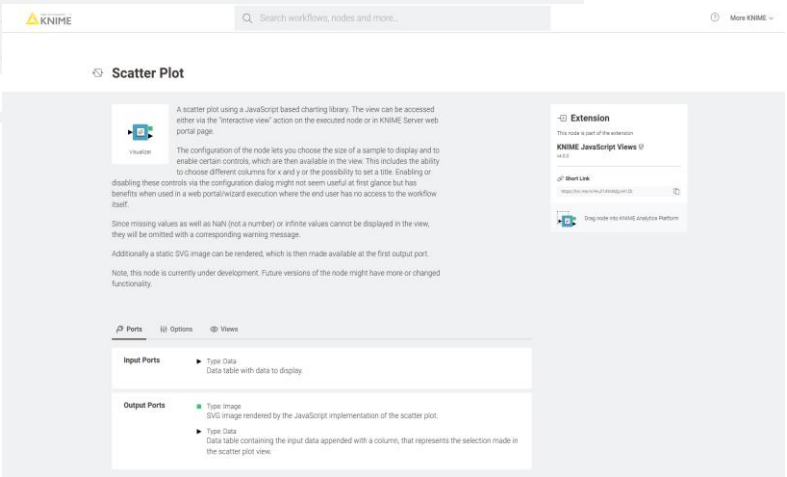
# Sharing Workflows

## How to use the KNIME Hub

# KNIME Hub



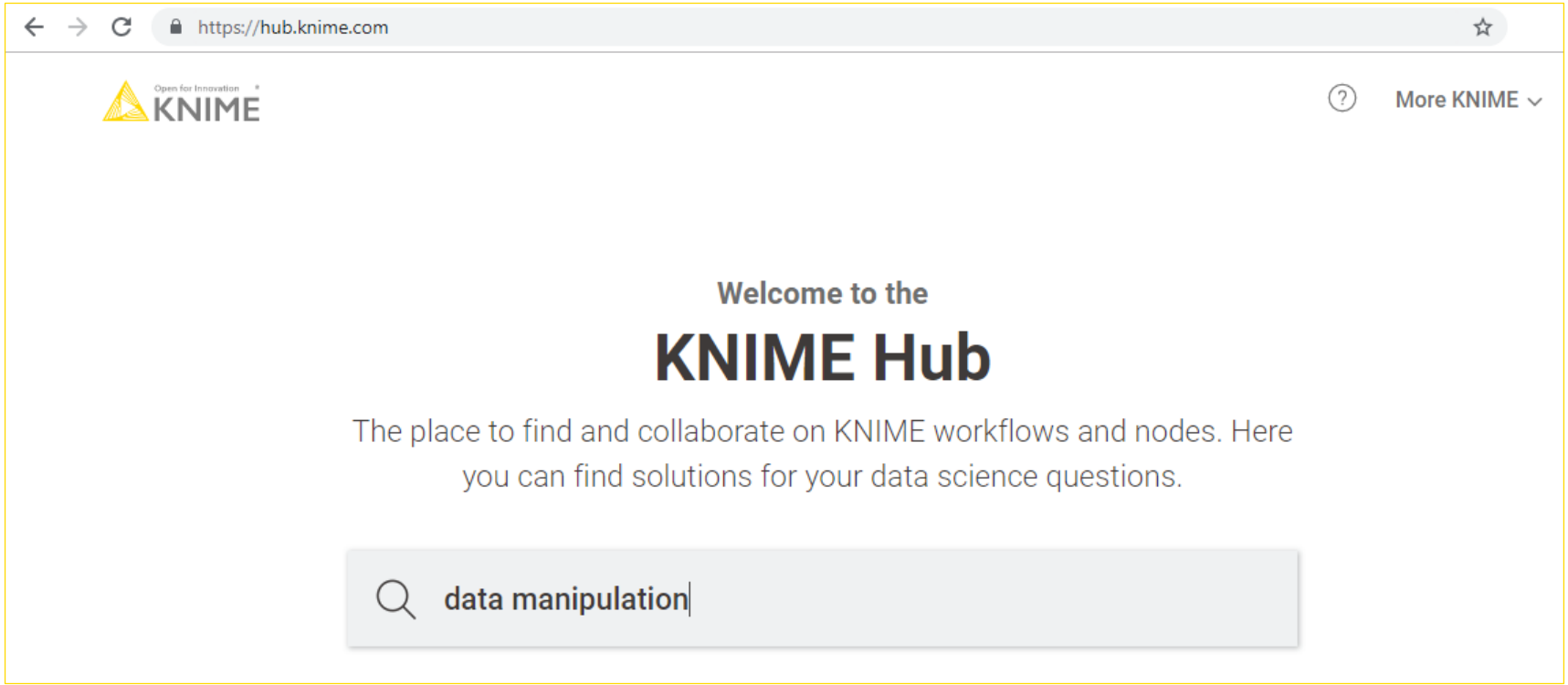



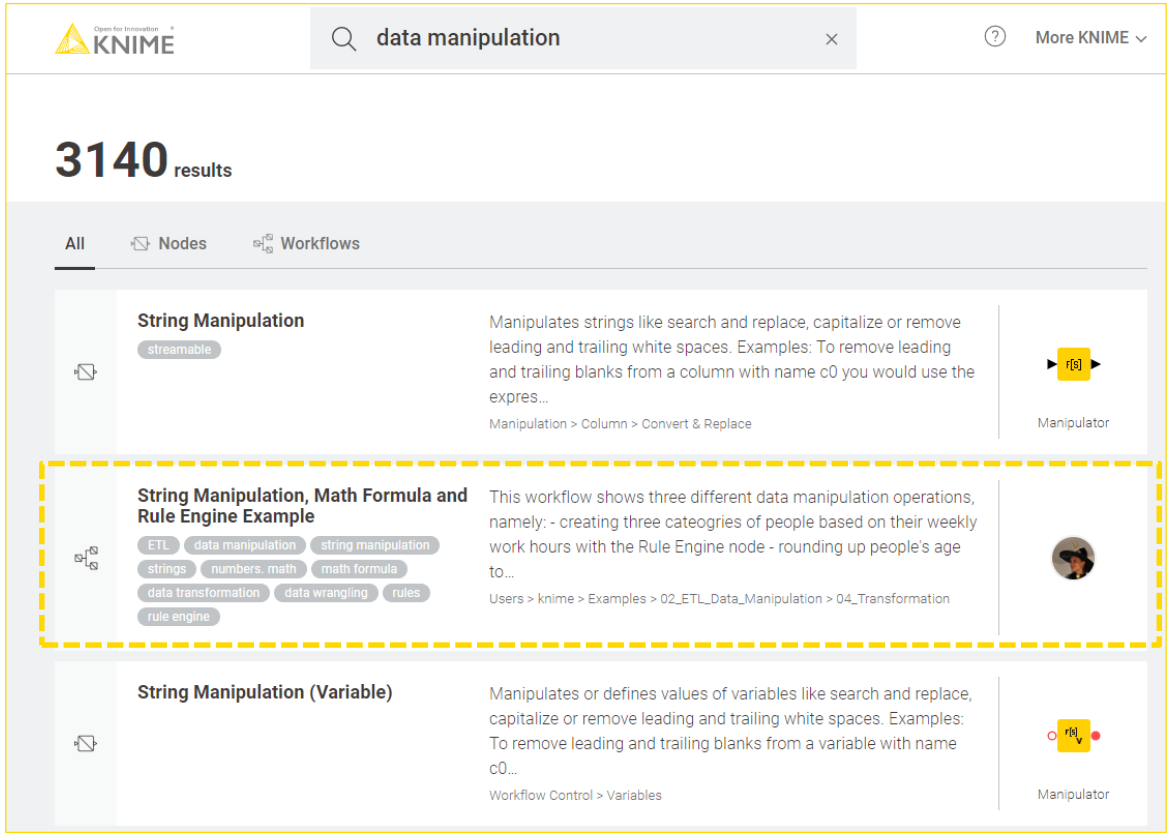

A place to share knowledge about Workflows and Nodes

https://hub.knime.com

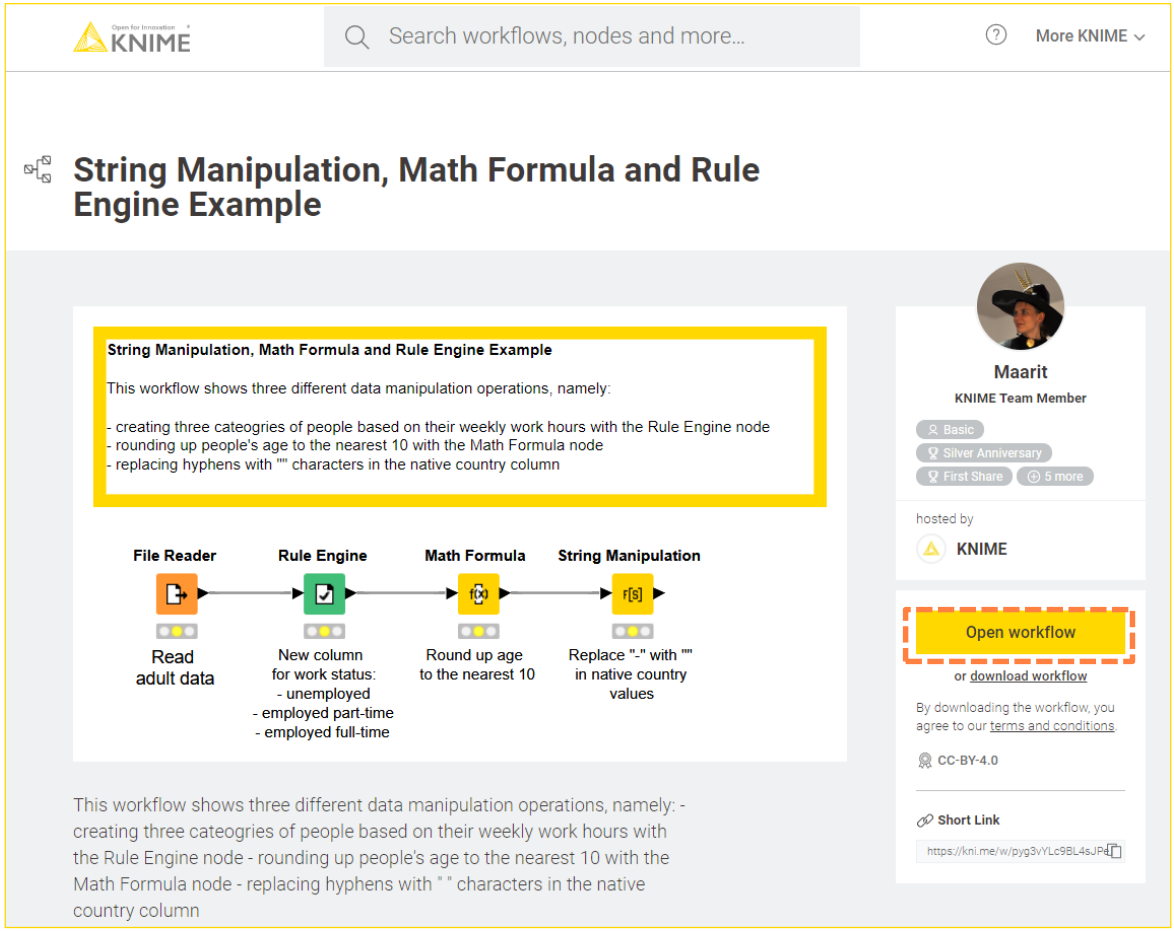
# The KNIME Hub



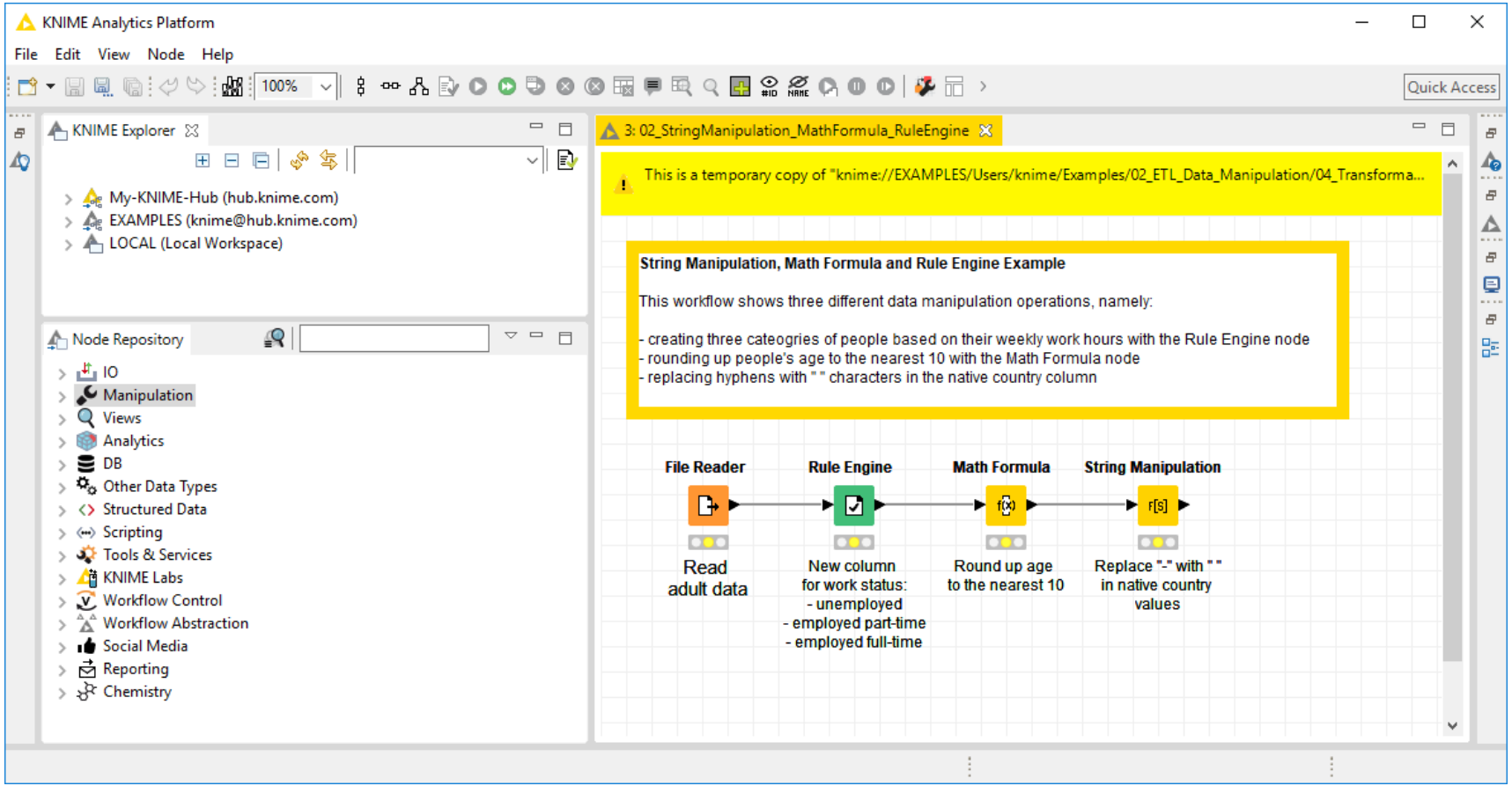
# Searching Nodes and Workflows



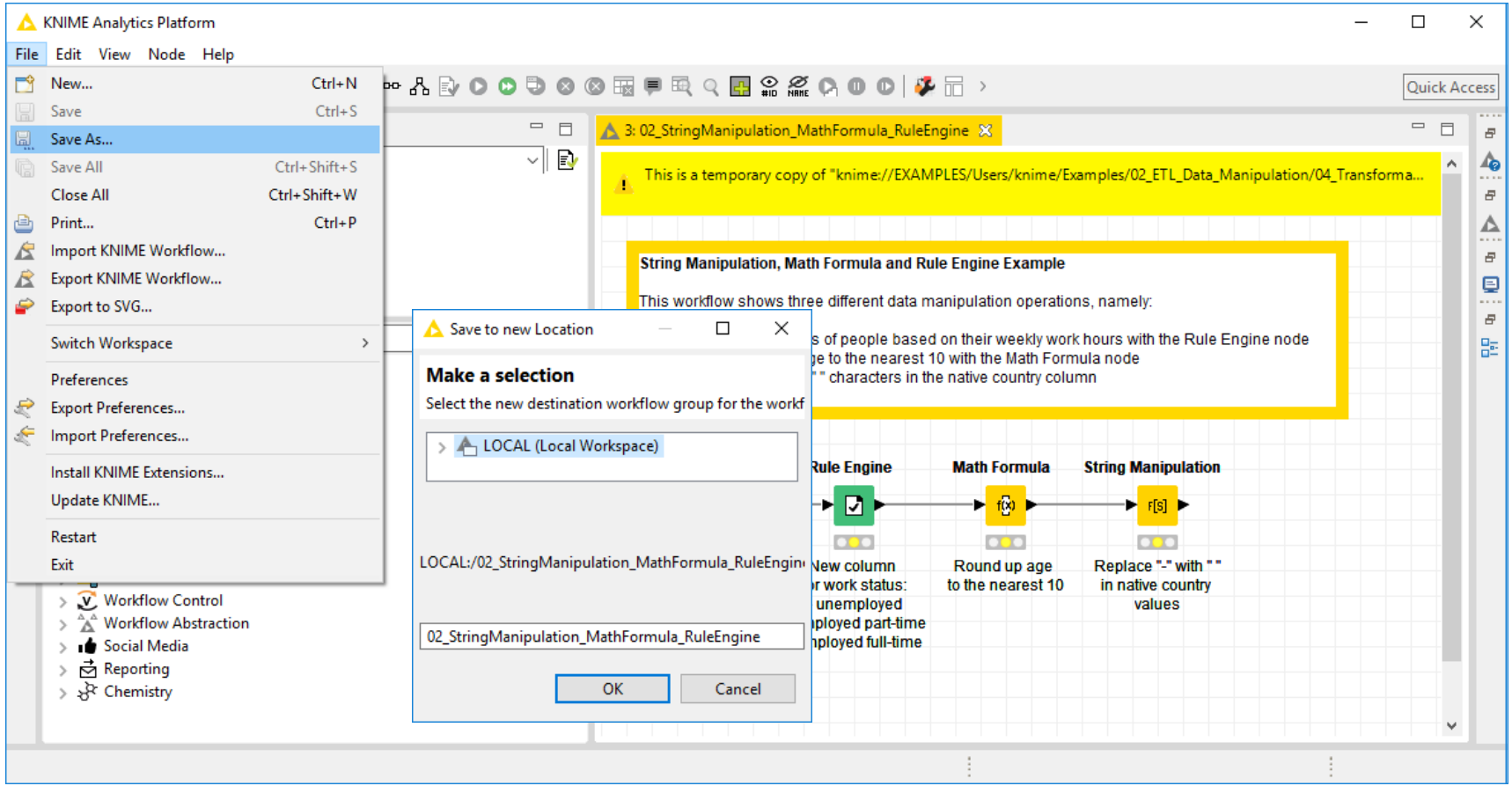
# Opening a Workflow from the Hub



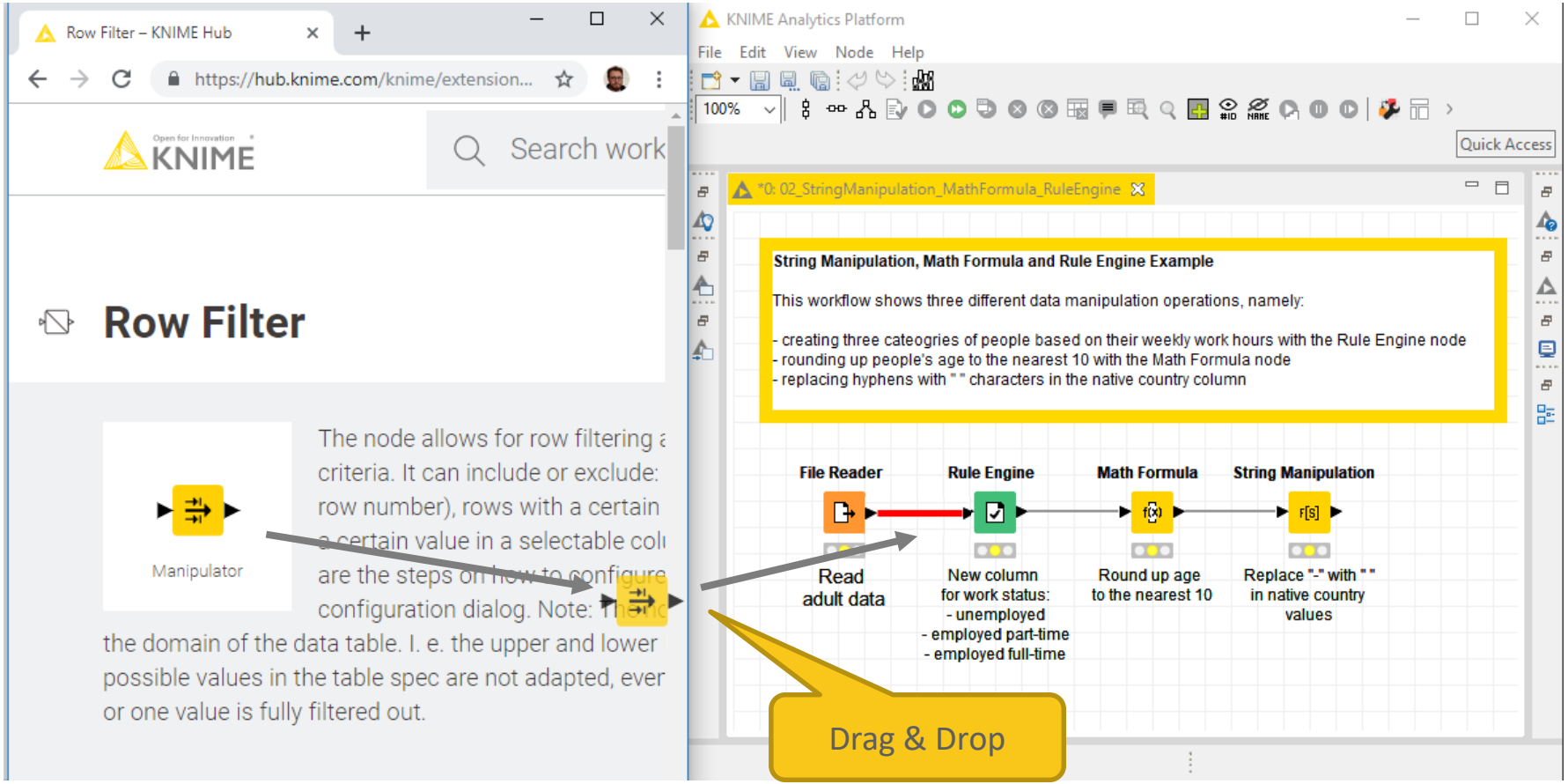
# Open Workflow in KNIME Analytics Platform



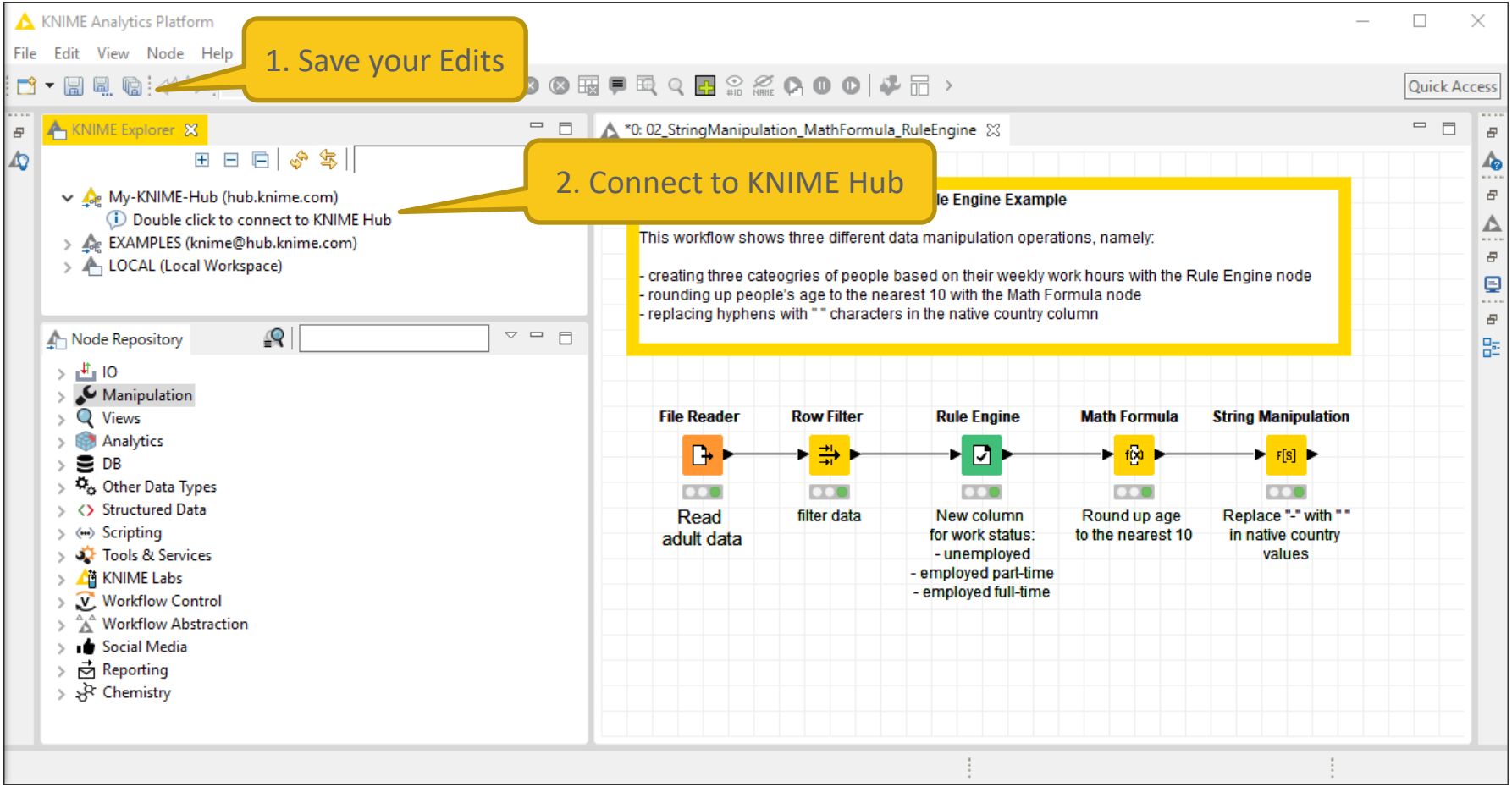
# Saving the Workflow



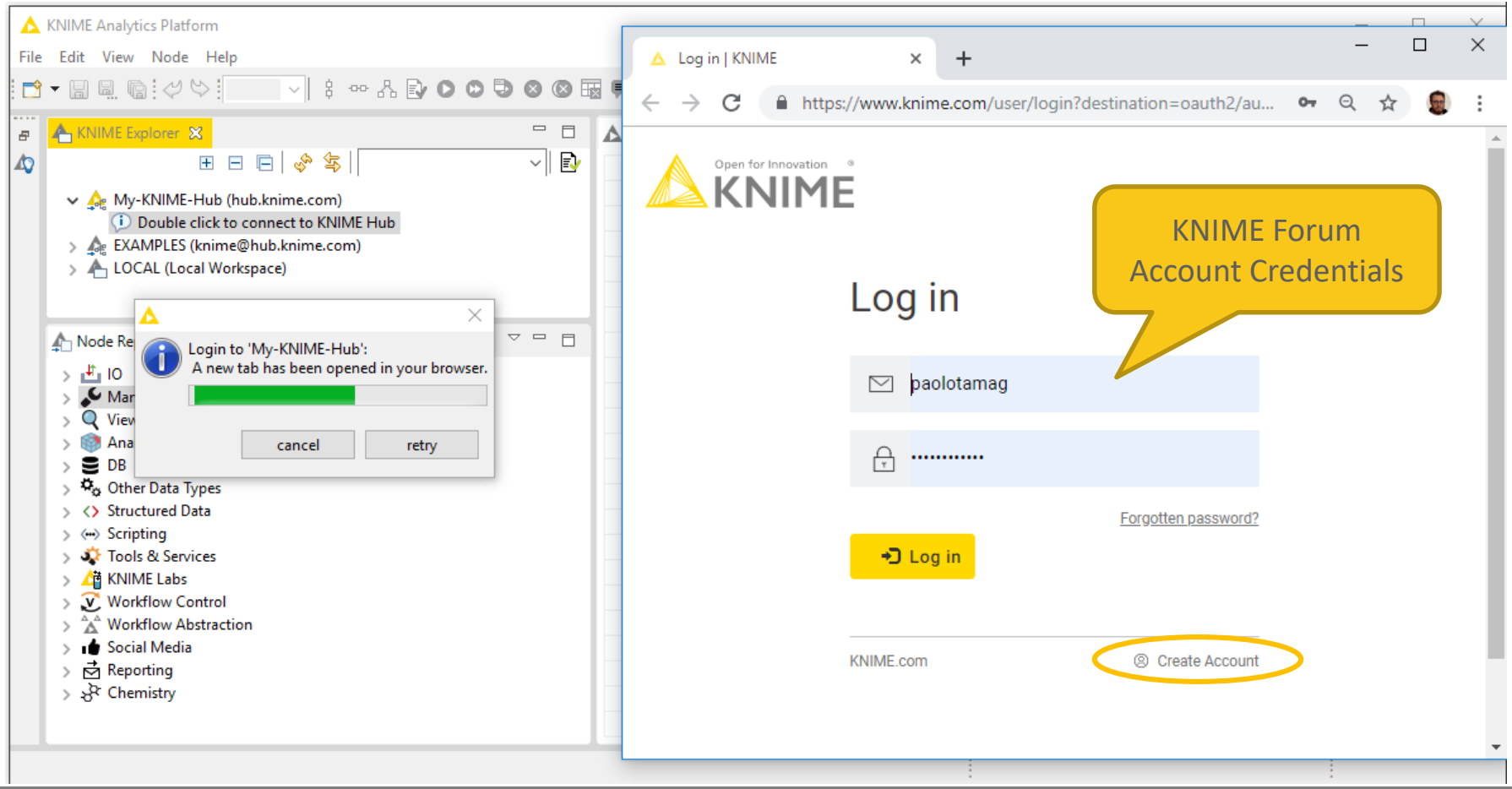
# Edit the Workflow



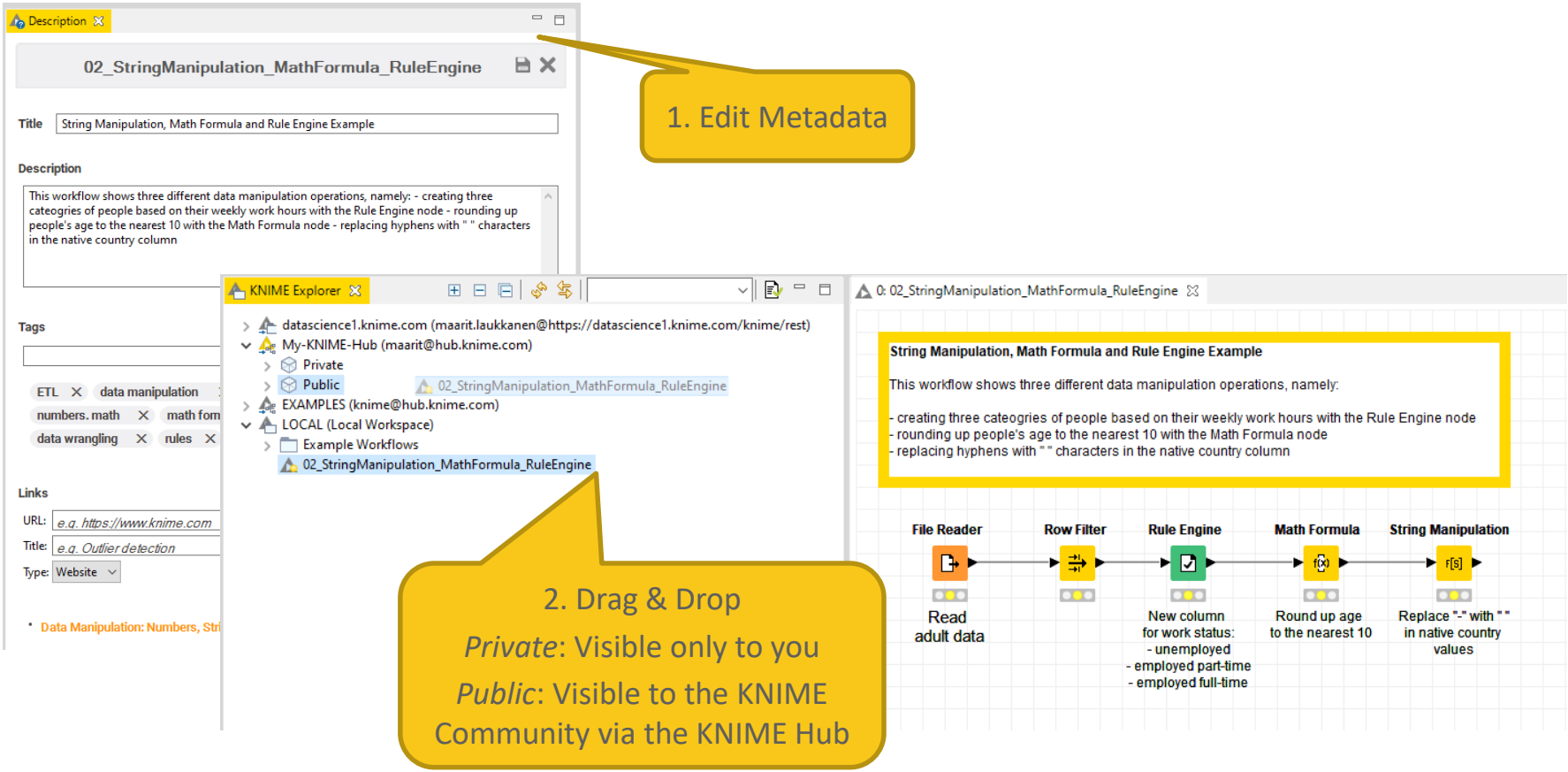
# Sharing the Workflow



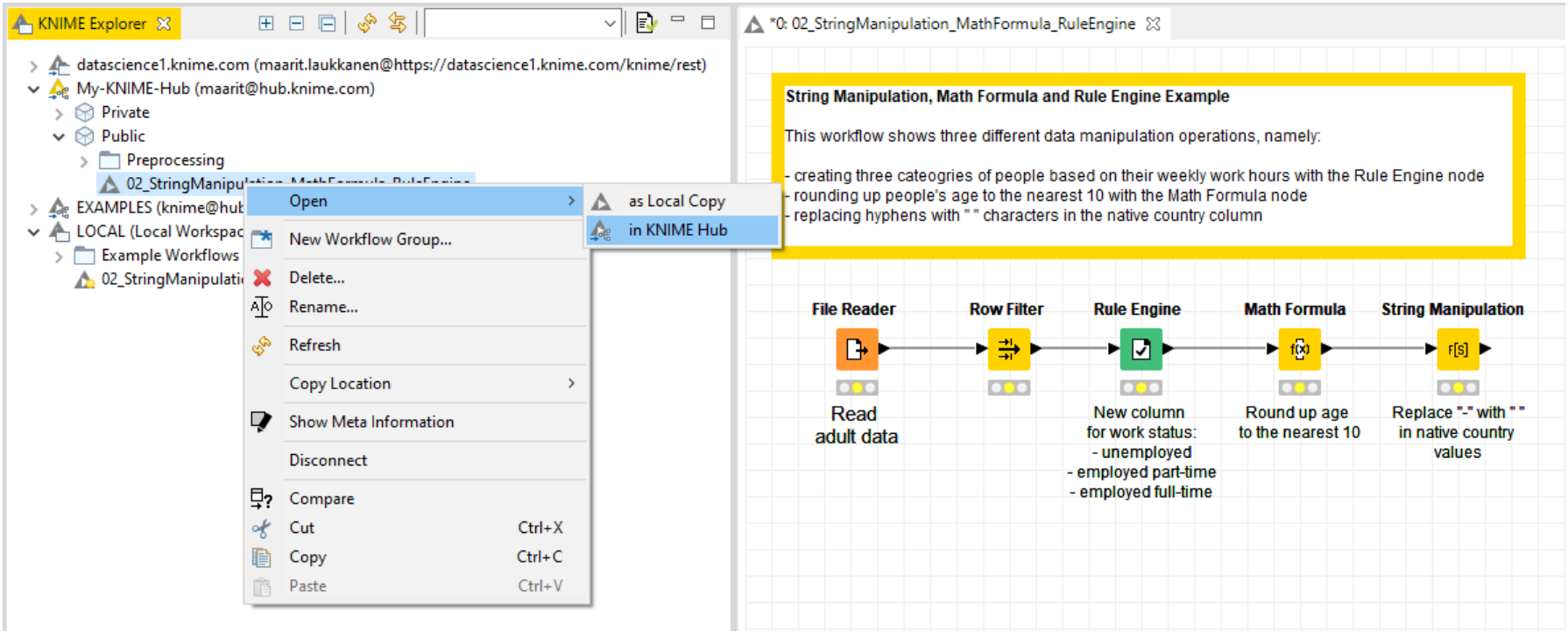
# Log in the Hub



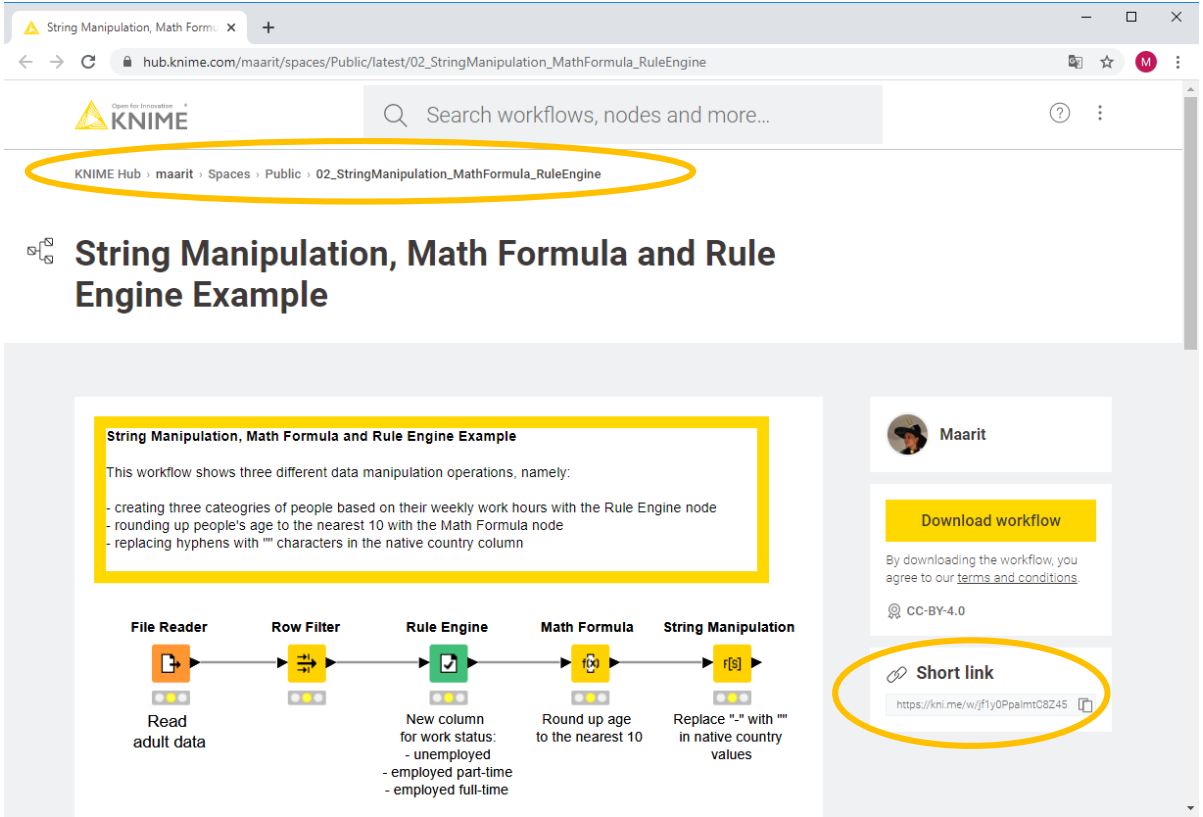
# Publish your Workflow



# Open your Workflow in the Hub



# Open your Workflow in the Hub



# 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 |
| Node Connections | Ctrl + L | connects selected nodes |
| | Ctrl + Shift + L | disconnects selected 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 |
| Metanode | Shift + F12 | opens metanode wizard |

# Stay connected with KNIME

**Blog:** knime.com/blog

**Forum**: forum.knime.com

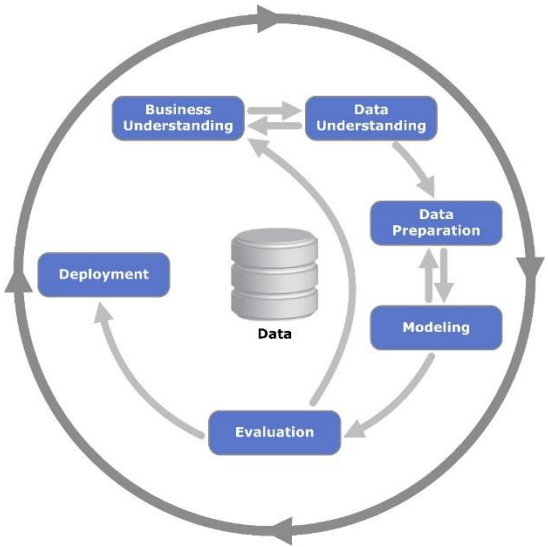
**KNIME Hub**:
hub.knime.com

**KNIME E-Learning Course**:
www.knime.com/e-learning-course

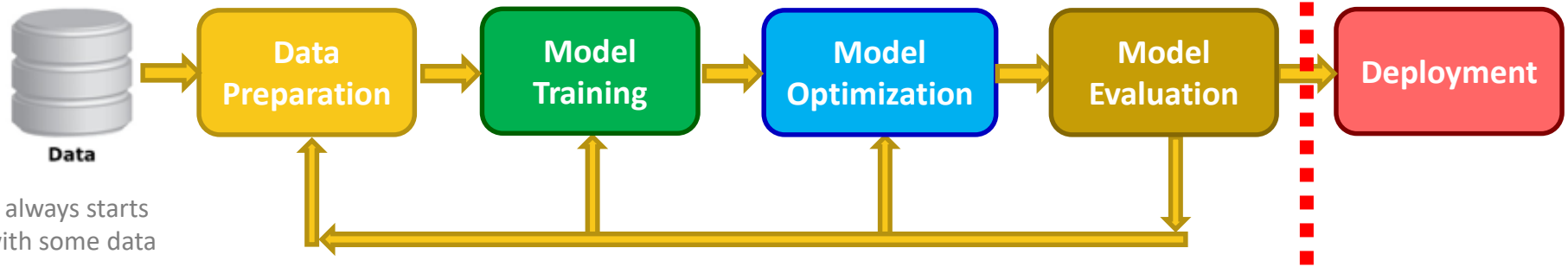
Follow us on social media:

# Today's Example: Churn Prediction

- Build a data science application step by step
- Each section of the course has an associated workflow with exercises
- The exercises complete the steps in the CRISP-DM cycle



# Today's Example: Churn Prediction




It always starts with some data …

- Data Manipulation
- Data Blending
- Missing Values Handling
- Feature Generation
- Dimensionality Reduction
- Feature Selection
- Outlier Removal
- Normalization
- Partitioning
- …

- Model Training
- Bag of Models
- Model Selection
- Ensemble Models
- Own Ensemble Model
- External Models
- Import Existing Models
- Model Factory
- …

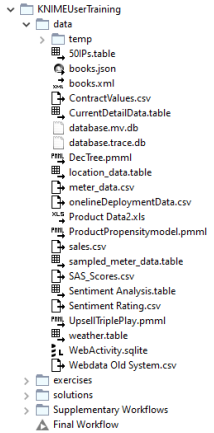
- Parameter Tuning
- Parameter Optimization
- Regularization
- Model Size
- No. Iterations
- …

- Performance Measures
- Accuracy
- ROC Curve
- Cross-Validation
- …

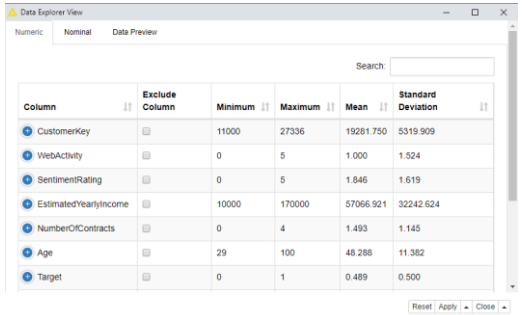
- Files & DBs
- Dashboards
- REST API
- SQL Code Export
- Reporting
- …

# The Data

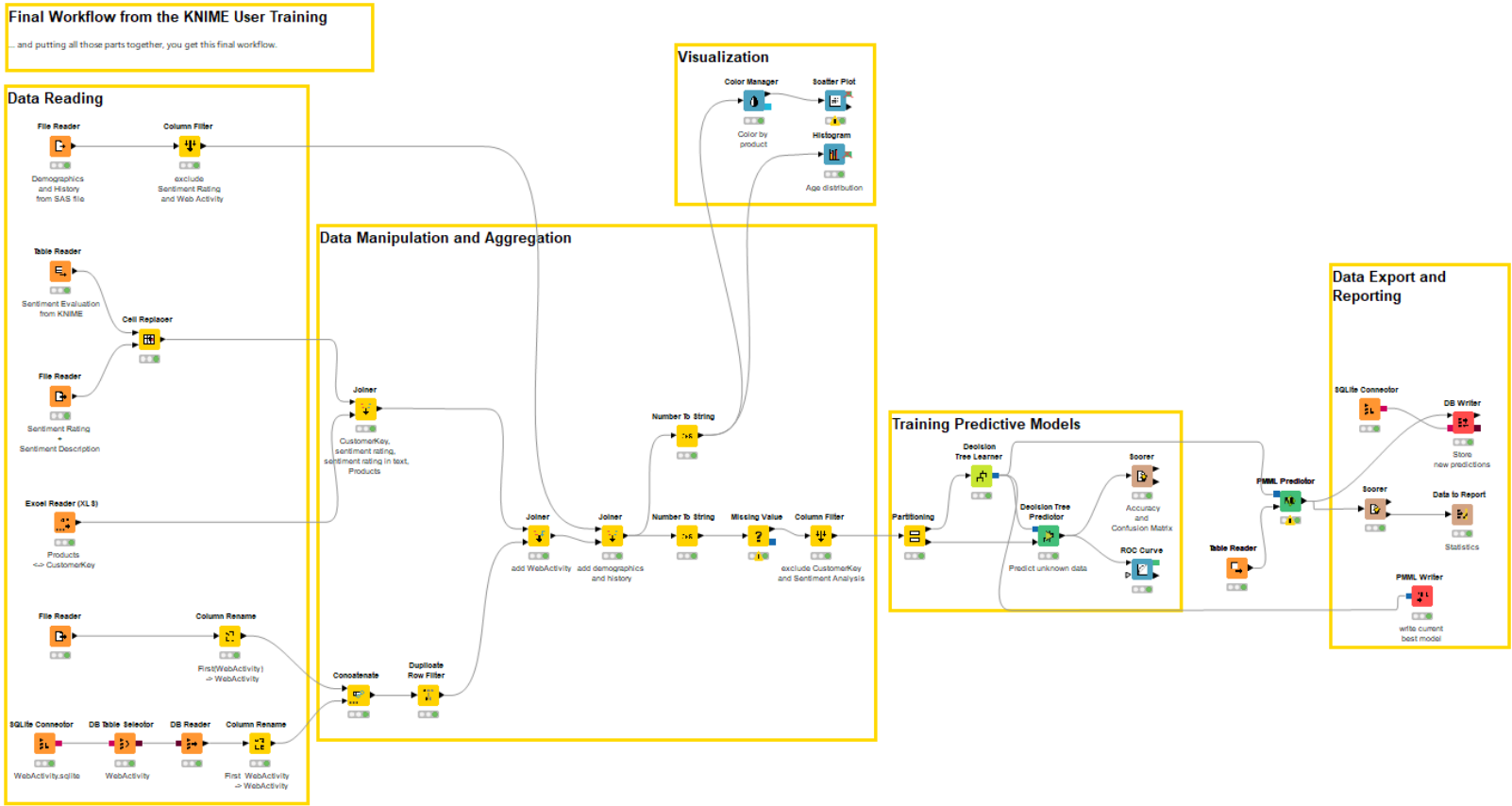
- The data files used in the exercises are available in the “data” folder: data files in different file formats, web-based data, data on a database, etc.




- For churn prediction, customer data are blended from different sources
- The Data Explorer node is helpful in inspecting data

# Today's Example: Churn Prediction



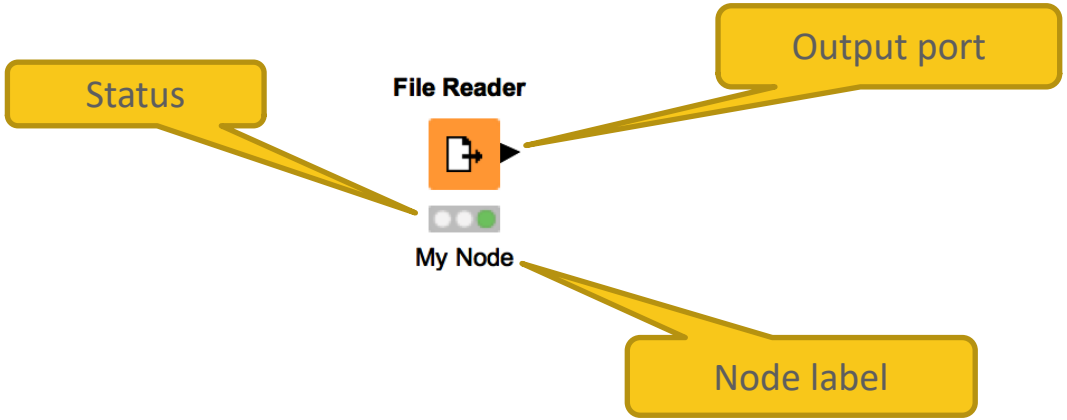
# Importing Data

## Accessing Files and Databases

# Data Source Nodes

Typically characterized by:

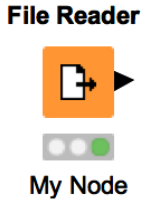
- Orange color
- No input ports, 1-2 output ports



# New Node: File Reader

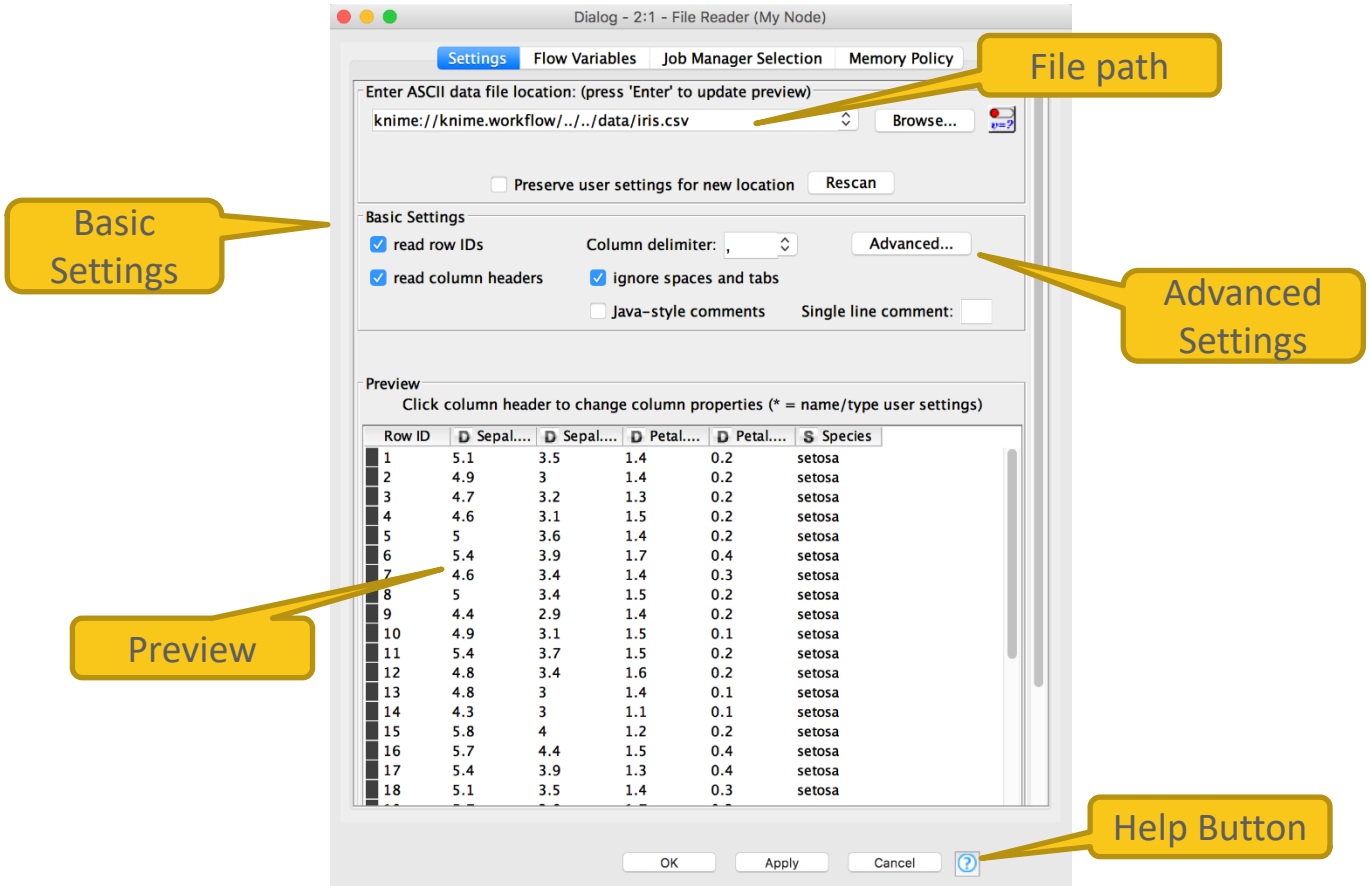
Workhorse of the KNIME Source nodes

- 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

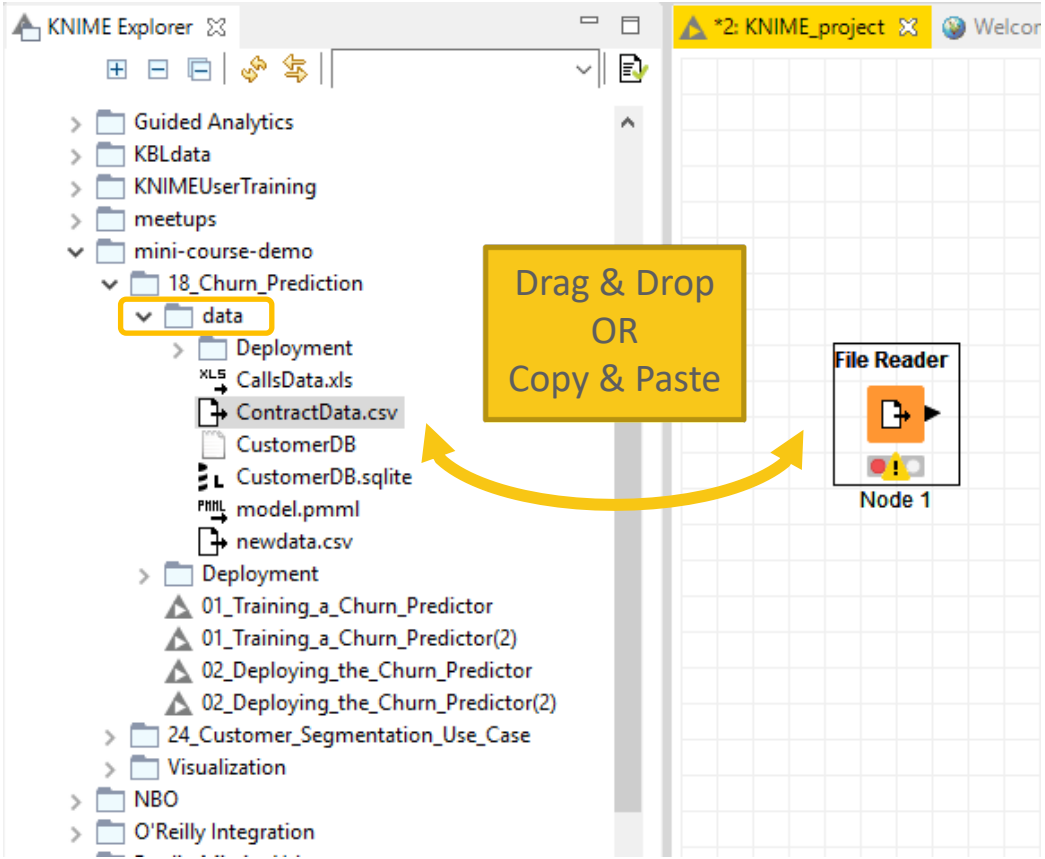



YouTube KNIME TV Channel video:
https://youtu.be/flaHQw-Qhlg

# File Reader Configuration

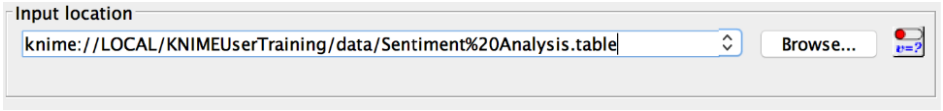


# Alternative Faster Way …



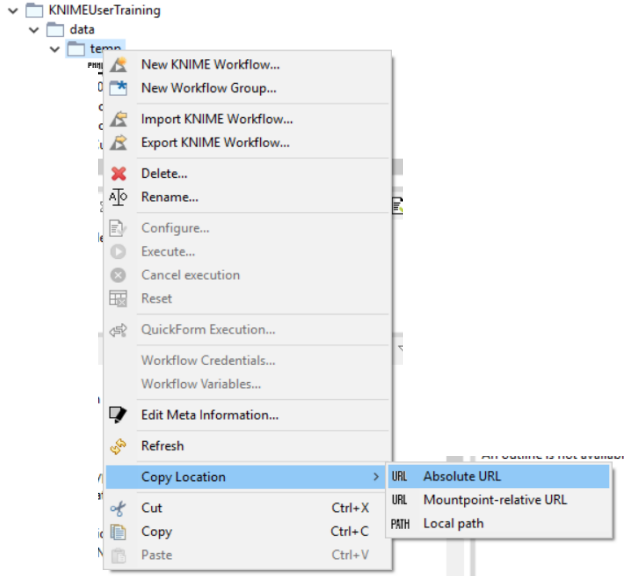
# Filenames and the knime:// Protocol

## Absolute URL



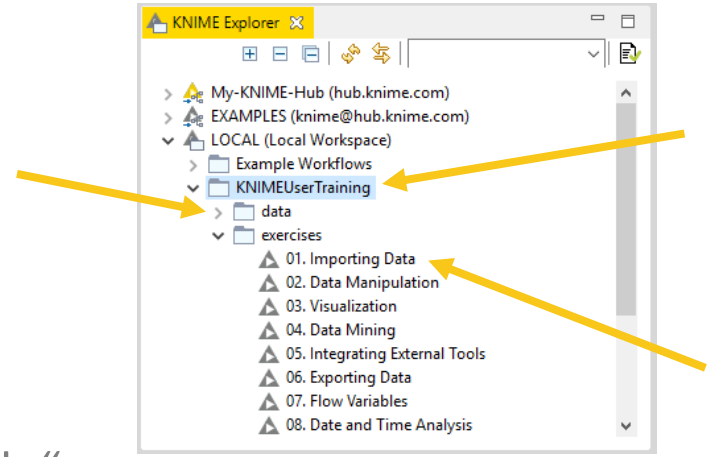

## Mountpoint-relative URL

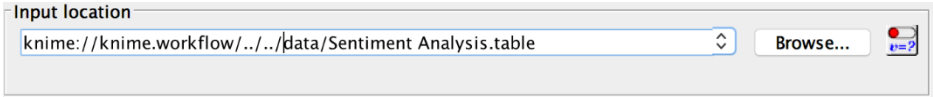

## Local path

# 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:



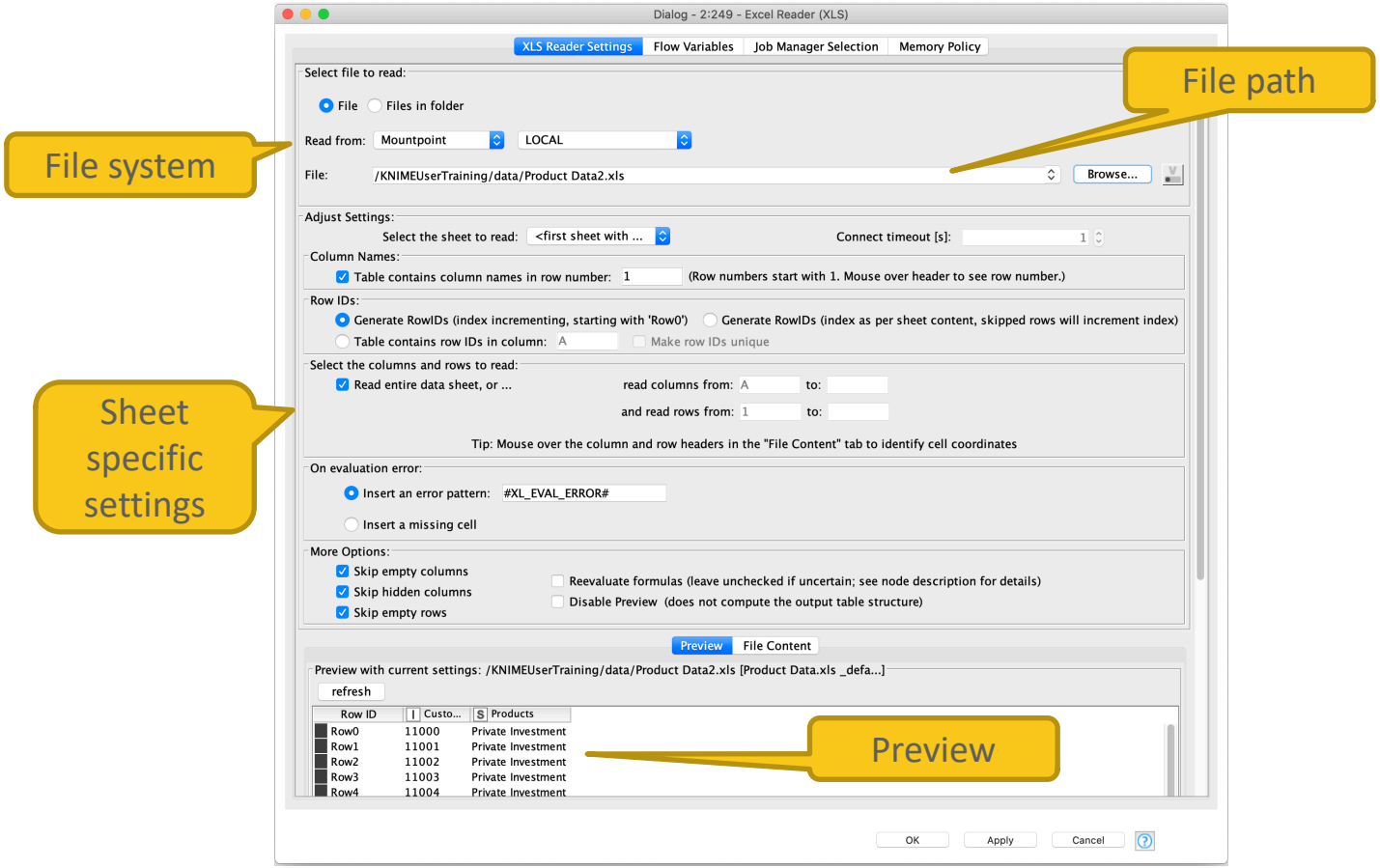




YouTube KNIME TV Channel: https://youtu.be/U9sP4g4yGwY

# New Node: Excel Reader (XLS)

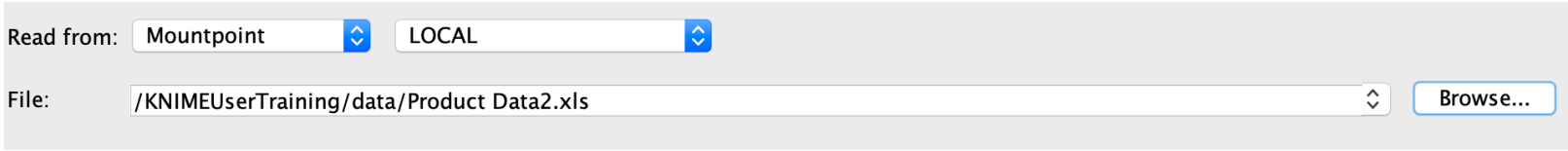
- Reads .xls and .xlsx file from Microsoft Excel
- Supports reading from multiple sheets

# Excel Reader Configuration

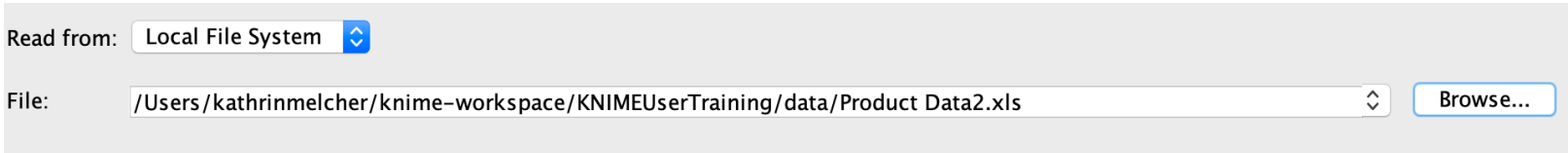


# Filenames and the knime:// Protocol

## Absolute URL



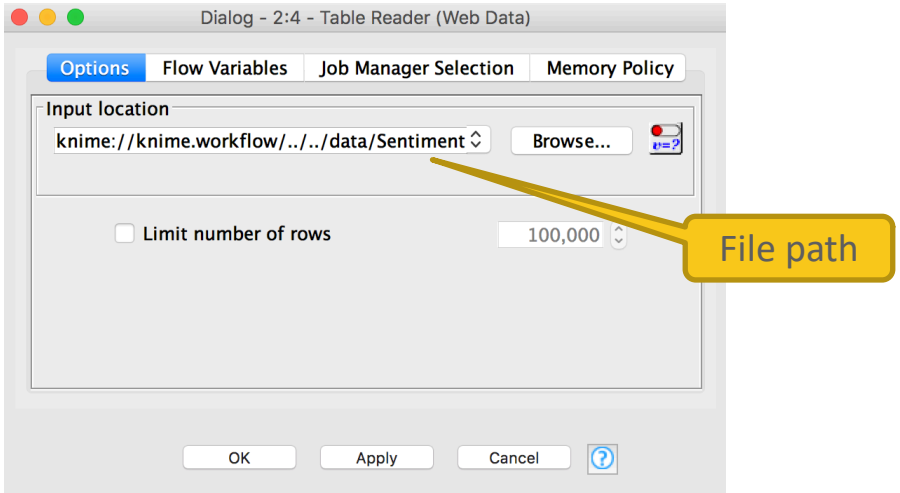

## Local Path




## Custom URL

# New Node: Table Reader

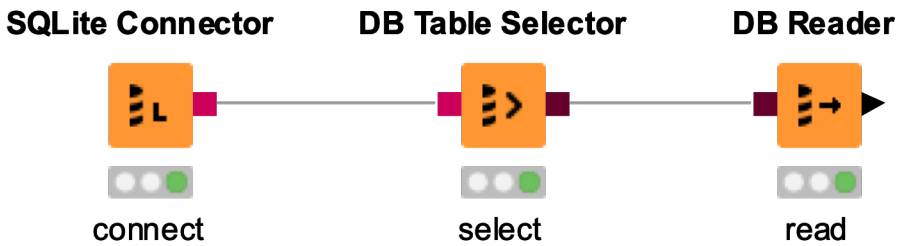
- Reads tables from the native KNIME Format.
- Maximum performance, minimum configuration


YouTube KNIME TV channel video:
https://youtu.be/tid1qi2HAOo

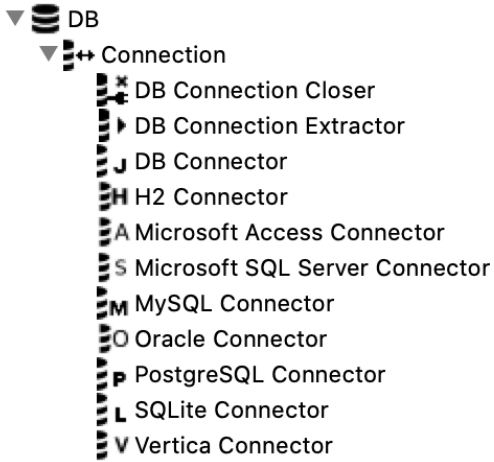
# Database Connectivity

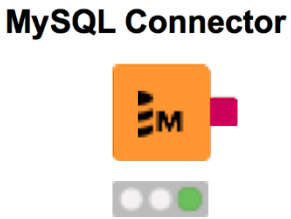
- Read data from any JDBC enabled database
- Write your own SQL or model it using dedicated nodes

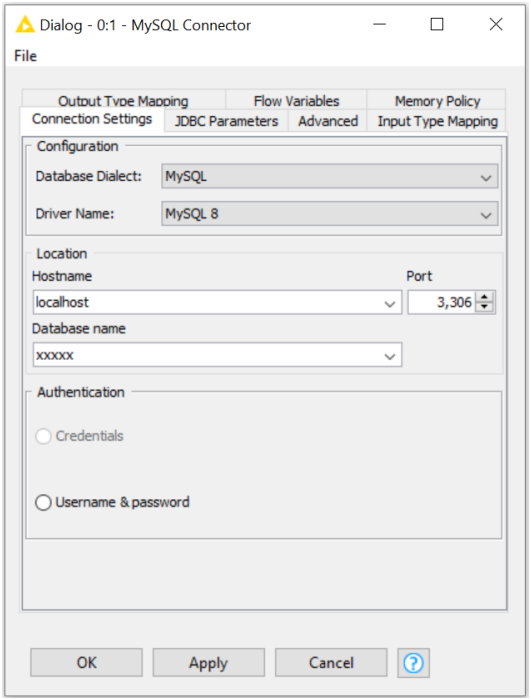


# New Nodes: Database Connectors

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

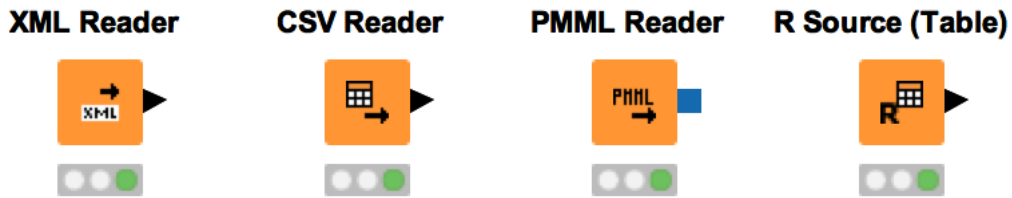


# Other Useful Data Sources

- 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



# Importing Data Exercise

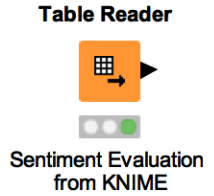
Start with exercise: *Importing Data*

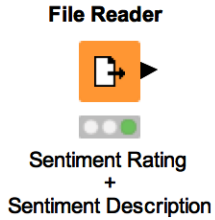
Read the following files

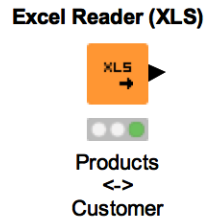
- *Sentiment Analysis.table*
- *Sentiment Rating.csv*
- *Product Data2.xls*

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

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



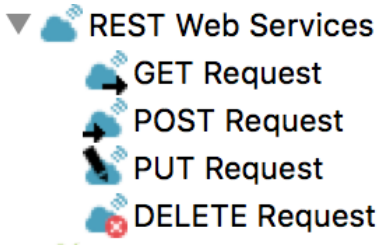



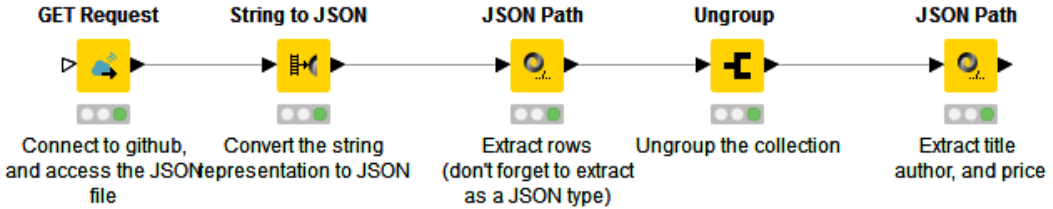

You can download the training workflows from the KNIME Hub:
https://hub.knime.com/knime/space/Education/01%20KNIME%20User%20Training/

# RESTful Web Services

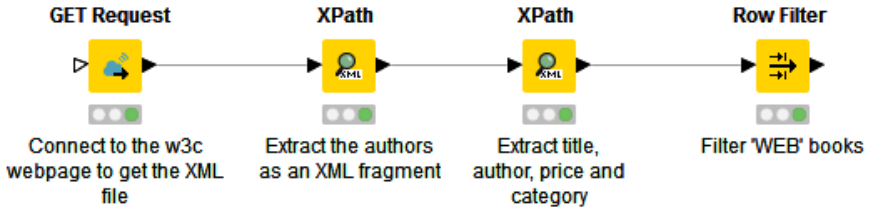
- Use KNIME nodes to interact with RESTful web services
- Send requests using standard HTTP methods



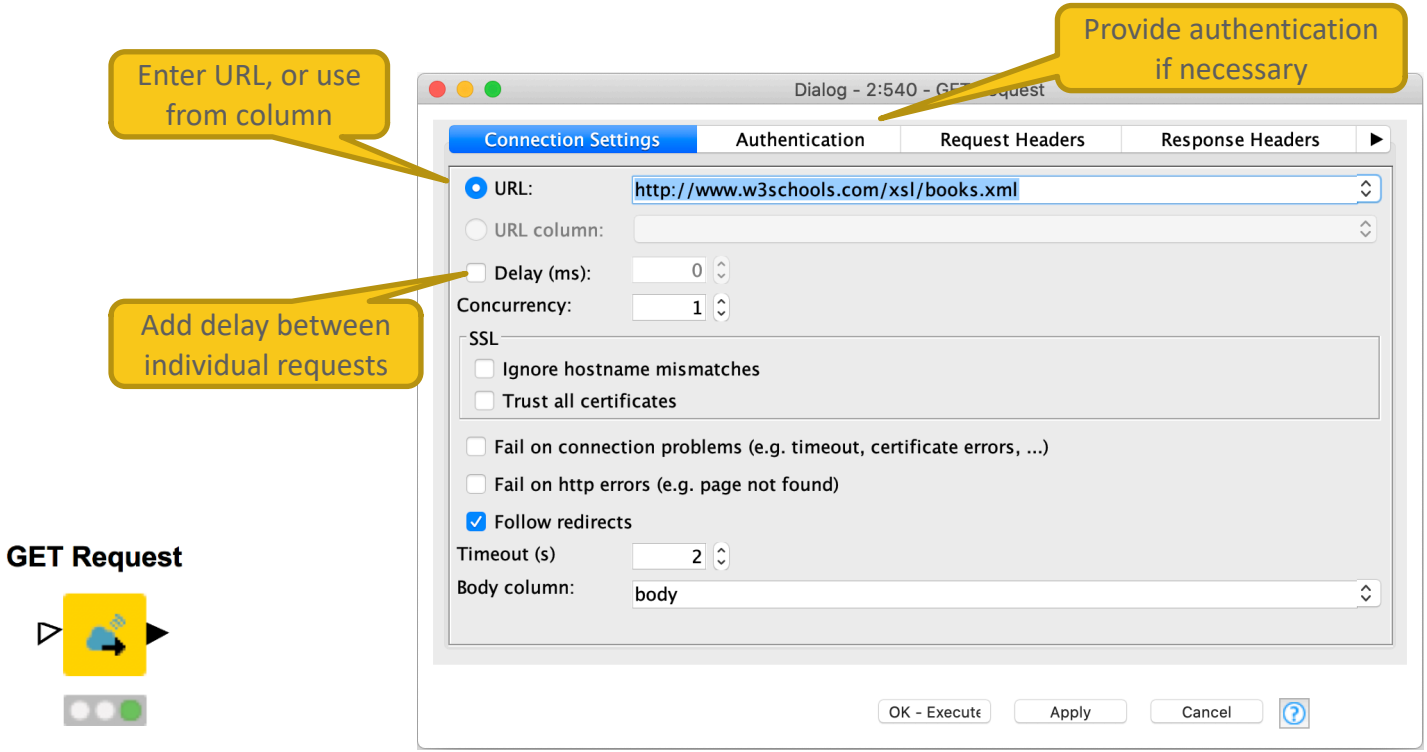

JSON Response:




XML Response:



# RESTful Web Services

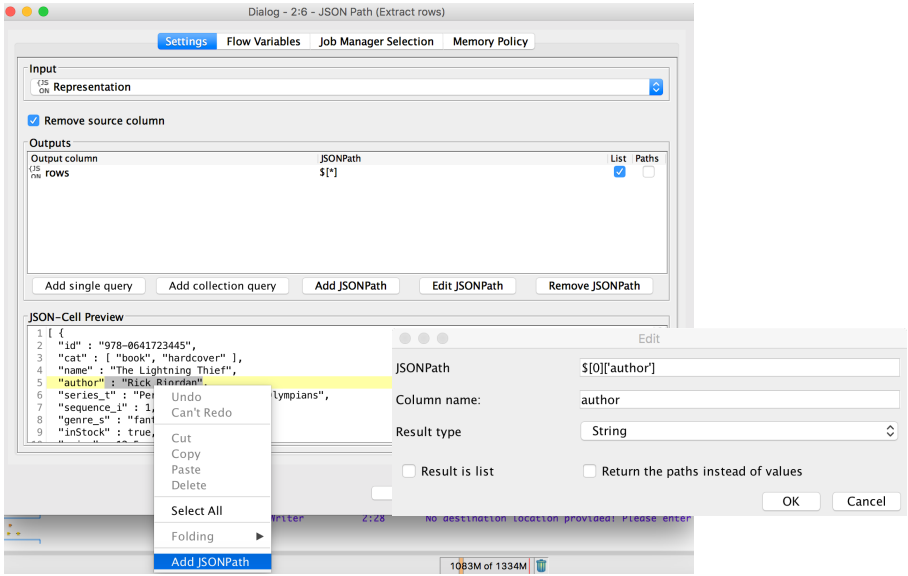



https://www.knime.com/blog/a-restful-way-to-find-and-retrieve-data

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

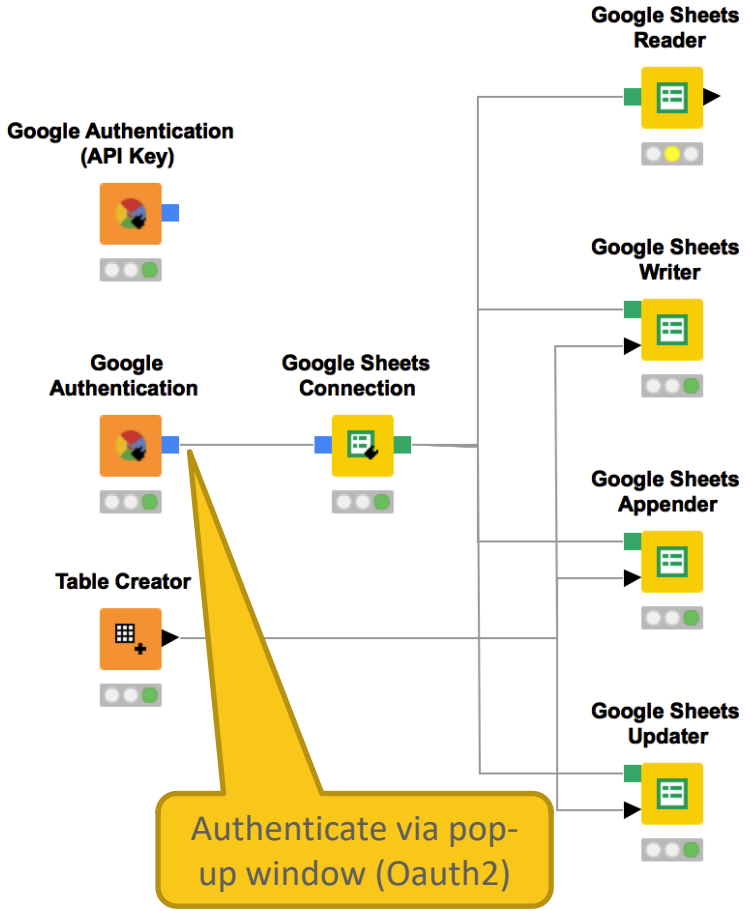
# JSON Reader and JSON Path nodes

- Use the JSON Reader (or GET Request) node to get a JSON cell
- Use the JSON Path node to query the JSON file and extract parameters
- Editor window simplifies construction of JSON queries by auto-generating them (click on properties)



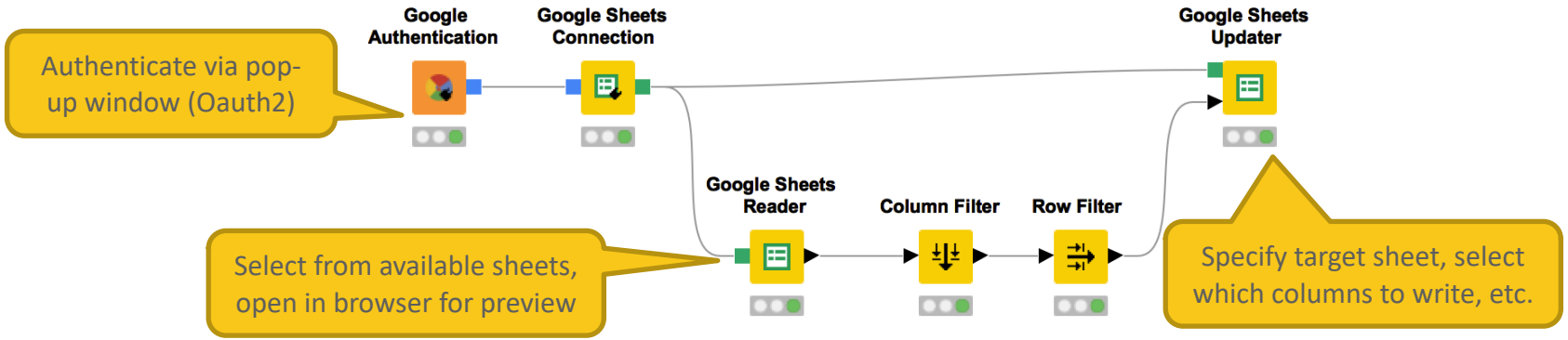
# Google Sheets

- 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)

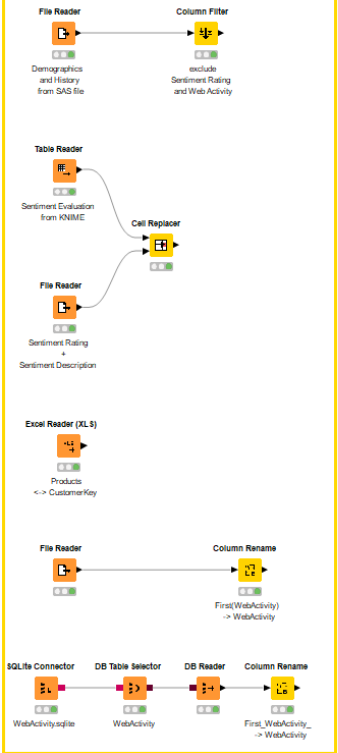


# Today's Example

**Final Workflow from the KNIME User Training**

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

**Data Reading**




**Visualization**

**Data Manipulation and Aggregation**

**Training Predictive Models**

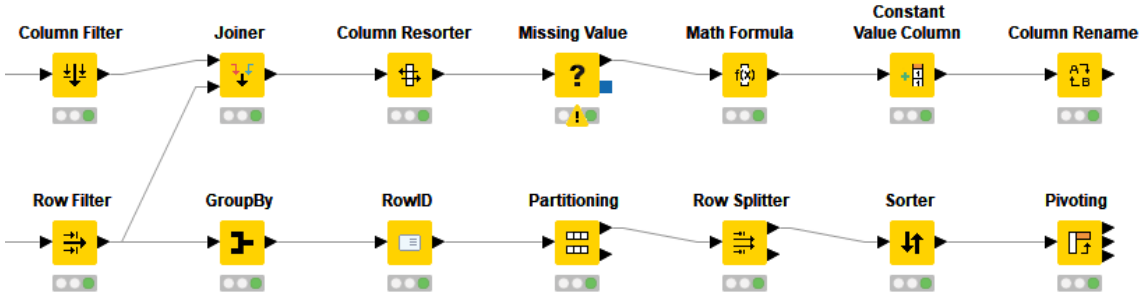
**Data Export and Reporting**

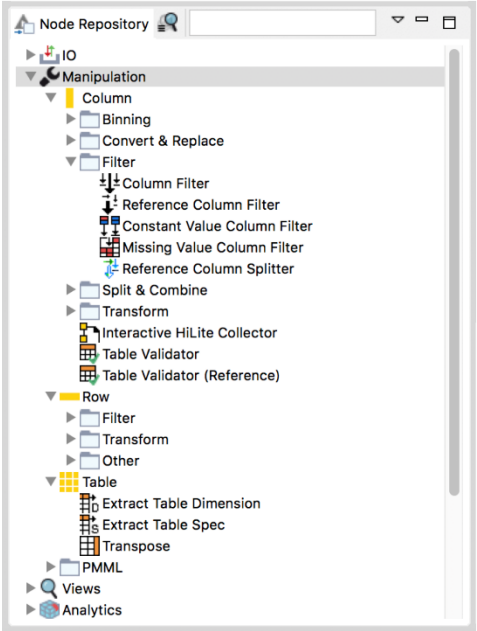
# Data Manipulation

## 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!

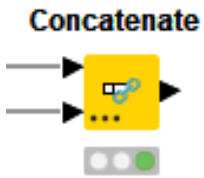


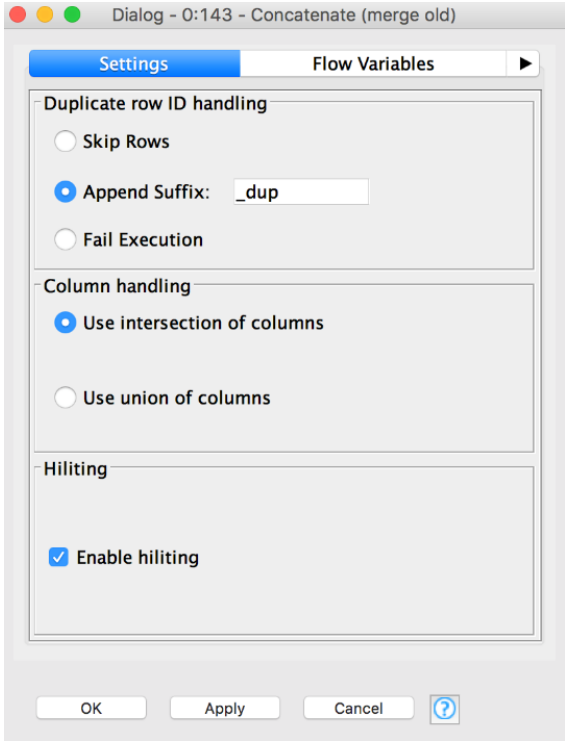



# New Node: Concatenate

Combine rows from 2 or more tables with shared columns

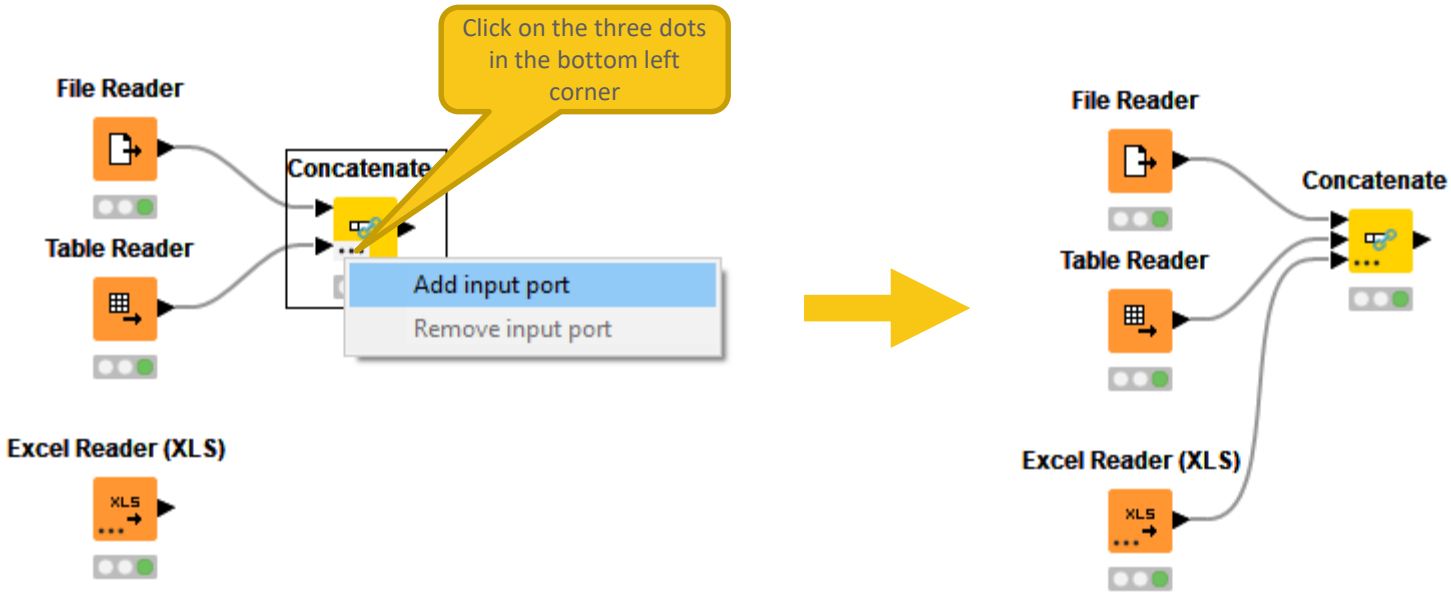
- Handles duplicate row keys gracefully
- Take the union or intersection of columns



# Dynamic Ports

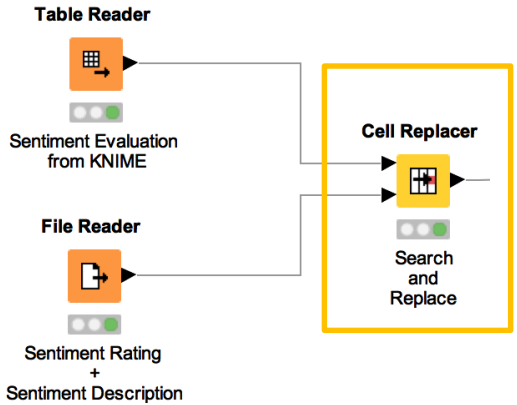
Add and remove node ports based on your needs, e.g. in order to concatenate three or more tables

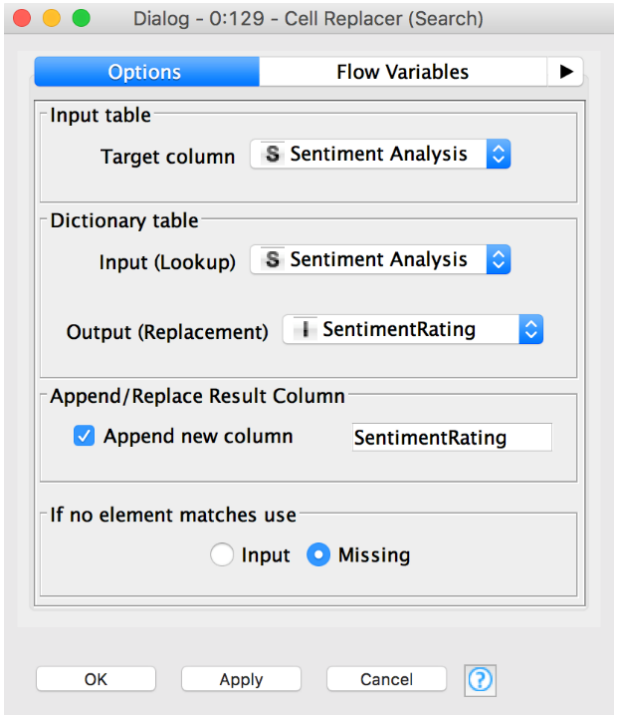


# New Node: Cell Replacer

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)

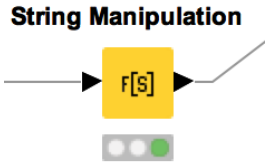


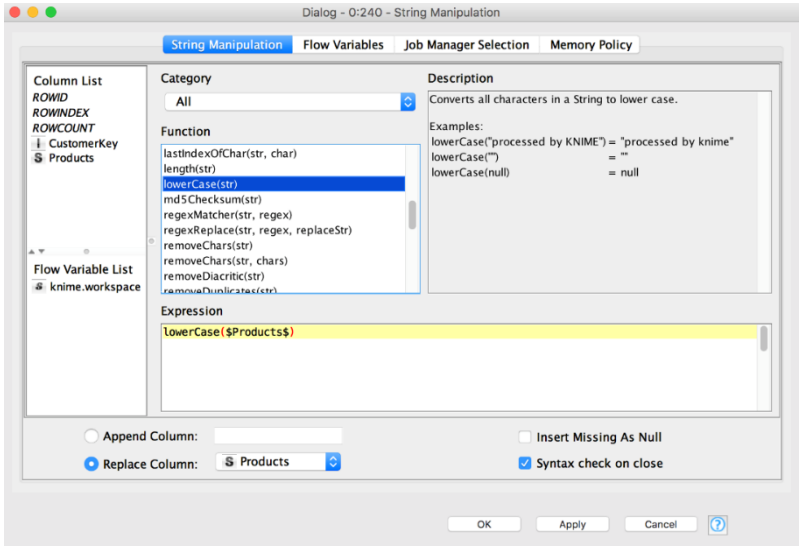



# New Node: String Manipulation

Create and edit values in String columns

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



# Data Manipulation Exercise, Activity I

Start with exercise: *Data Manipulation, Activity I*

- Concatenate web activity data from the old and new systems
- Replace the written sentiment values with the numeric sentiment scores
- Make sure that all product names in the product data spreadsheet are written in lower case letters

# Joining Columns of Data

Left Table

| CustomerKey | OrderDate | OrderID |
|---|---|---|
| 22 | 2019-09-23 | #23444 |
| 24 | 2019-09-30 | #23457 |
| 15 | 2019-10-07 | #28985 |
| 10 | 2091-10-13 | #29999 |

Join by CustomerKey

Right Table

| CustomerKey | DoB | City | Gender |
|---|---|---|---|
| 17 | 1974-02-23 | Berlin | F |
| 65 | 2001-05-25 | Stuttgart | F |
| 35 | 1988-08-05 | Cologne | M |
| 15 | 1983-07-20 | Hamburg | M |
| 10 | 1993-01-13 | Berlin | M |

Inner Join

| CustomerKey | OrderDate | OrderID | DoB | City | Gender |
|---|---|---|---|---|---|
| 15 | 2019-10-07 | #28985 | 1983-07-20 | Hamburg | M |
| 10 | 2091-10-13 | #29999 | 1993-01-13 | Berlin | M |

Left Outer Join

| CustomerKey | OrderDate | OrderID | DoB | City | Gender |
|---|---|---|---|---|---|
| 22 | 2019-09-23 | #23444 | ? | ? | ? |
| 24 | 2019-09-30 | #23457 | ? | ? | ? |
| 15 | 2019-10-07 | #28985 | 1983-07-20 | Hamburg | M |
| 10 | 2091-10-13 | #29999 | 1993-01-13 | Berlin | M |

Right Outer Join

| CustomerKey | OrderDate | OrderID | DoB | City | Gender |
|---|---|---|---|---|---|
| 17 | ? | ? | 1974-02-23 | Berlin | F |
| 65 | ? | ? | 2001-05-25 | Stuttgart | F |
| 35 | ? | ? | 1988-08-05 | Cologne | M |
| 15 | 2019-10-07 | #28985 | 1983-07-20 | Hamburg | M |
| 10 | 2091-10-13 | #29999 | 1993-01-13 | Berlin | M |

# Joining Columns of Data

Left Table

| CustomerKey | OrderDate | OrderID |
|---|---|---|
| 22 | 2019-09-23 | #23444 |
| 24 | 2019-09-30 | #23457 |
| 15 | 2019-10-07 | #28985 |
| 10 | 2091-10-13 | #29999 |

Join by CustomerKey

Right Table

| CustomerKey | DoB | City | Gender |
|---|---|---|---|
| 17 | 1974-02-23 | Berlin | F |
| 65 | 2001-05-25 | Stuttgart | F |
| 35 | 1988-08-05 | Cologne | M |
| 15 | 1983-07-20 | Hamburg | M |
| 10 | 1993-01-13 | Berlin | M |

Full Outer Join

Missing values in the left table

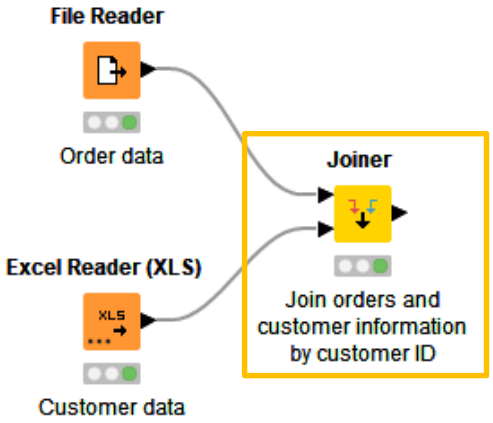
| CustomerKey | OrderDate | OrderID | DoB | City | Gender |
|---|---|---|---|---|---|
| 17 | ? | ? | 1974-02-23 | Berlin | F |
| 65 | ? | ? | 2001-05-25 | Stuttgart | F |
| 35 | ? | ? | 1988-08-05 | Cologne | M |
| 15 | 2019-10-07 | #28985 | 1983-07-20 | Hamburg | M |
| 10 | 2091-10-13 | #29999 | 1993-01-13 | Berlin | M |
| 22 | 2019-09-23 | #23444 | ? | ? | ? |
| 24 | 2019-09-30 | #23457 | ? | ? | ? |

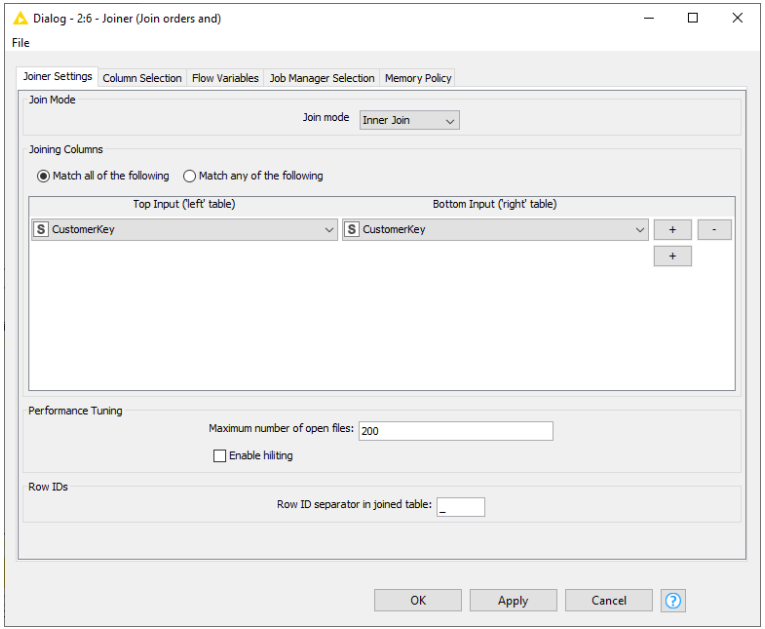
Missing values in the right table

# New Node: Joiner

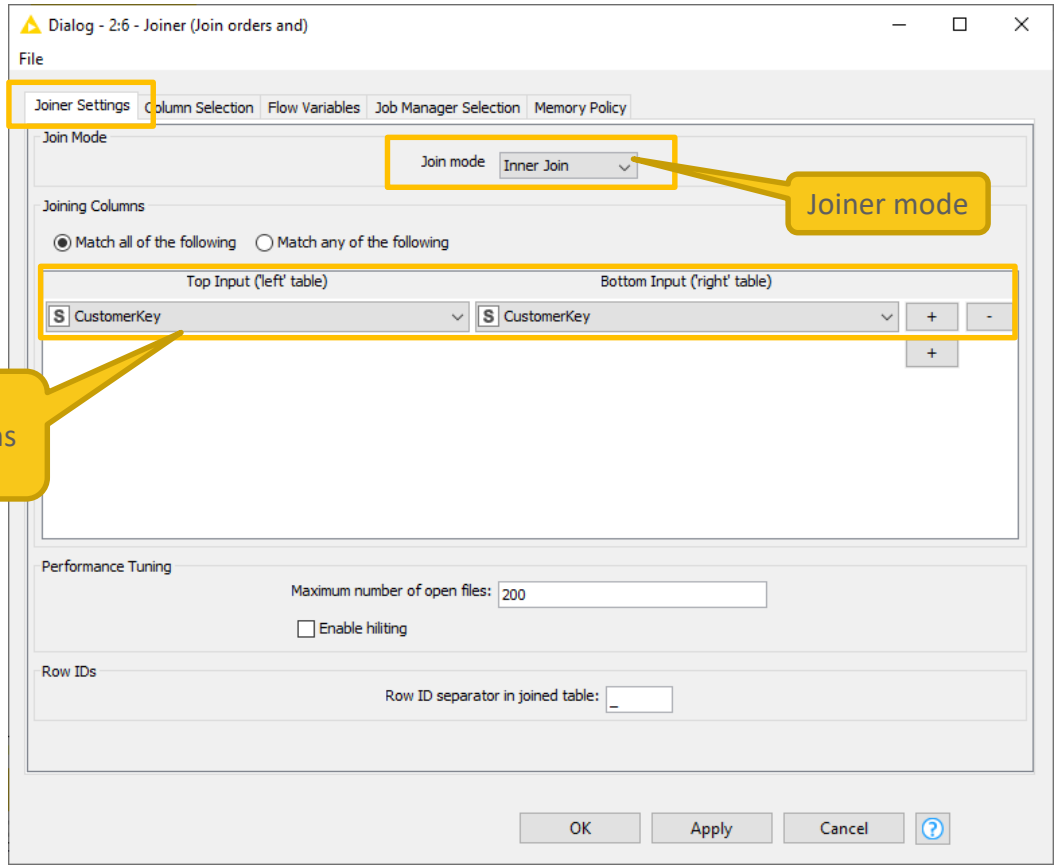
Combines columns from 2 different tables

- Top port contains “Left” data table
- Bottom port contains “Right” data table

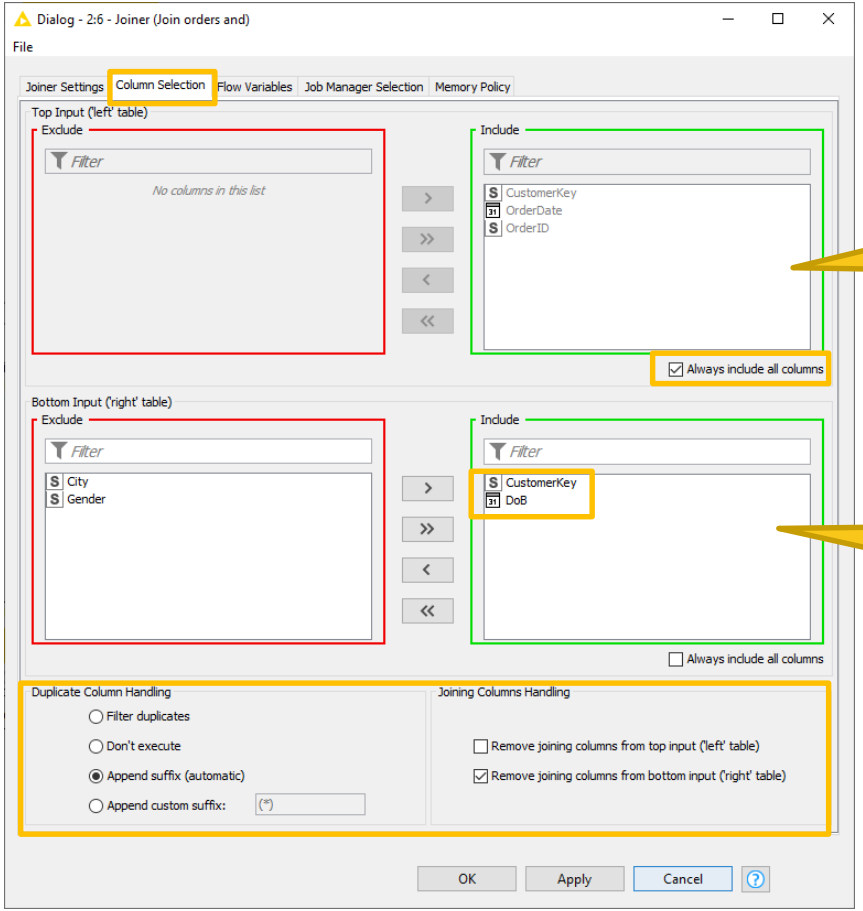


# Joiner Configuration – Linking Rows



# Joiner Configuration – Column Selection




Columns from left table to output table

Columns from right table to output table

# Data Aggregation

| Product ID | Category | # Ordered Items |
|---|---|---|
| P 1 | Clothing | 2 |
| P 2 | Home | 3 |
| P 3 | Clothing | 1 |
| P 4 | Clothing | 5 |
| P 5 | Electronics | 7 |
| P 6 | Electronics | 5 |

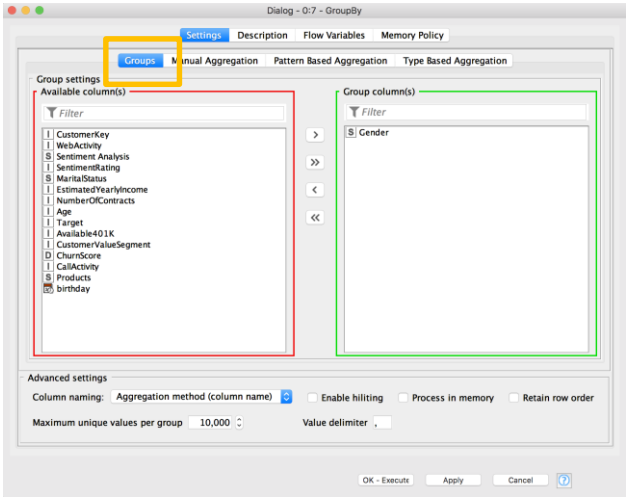
| Group | Sum(# Ordered Items) |
|---|---|
| Clothing | 8 |
| Home | 3 |
| Electronics | 12 |

Aggregated on Category (group) by Sum (aggregation method)

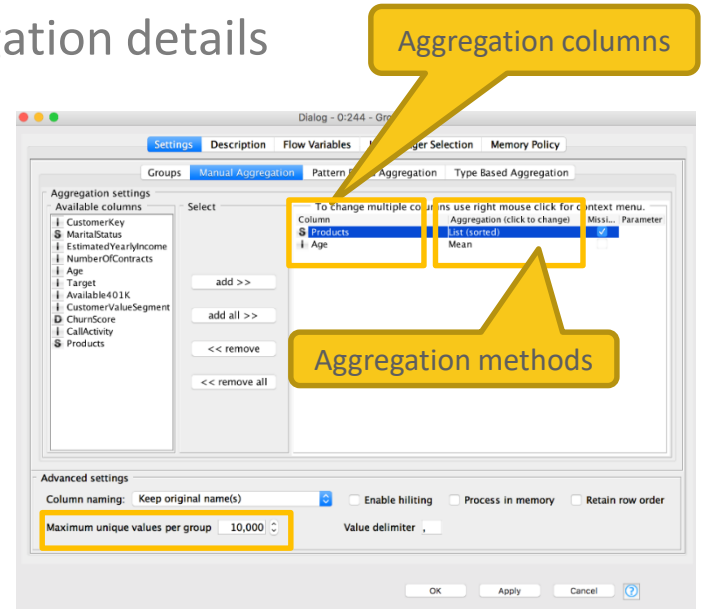
# New Node: GroupBy

Aggregate rows to summarize data

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





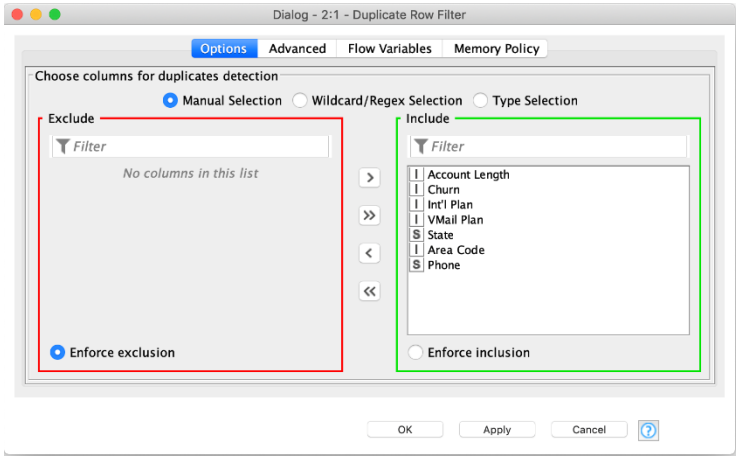


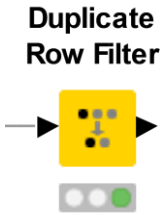

YouTube KNIME TV video:
https://youtu.be/bDwF-TOMtWw

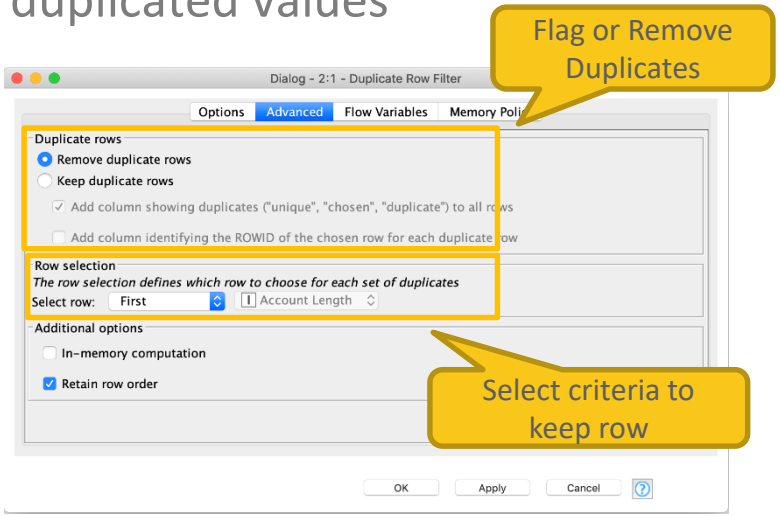
# New Node: Duplicate Row Filter

Detect duplicate row and apply a selected treatment

- First tab provides the option to select columns
- Second tab provides options for treating duplicated values

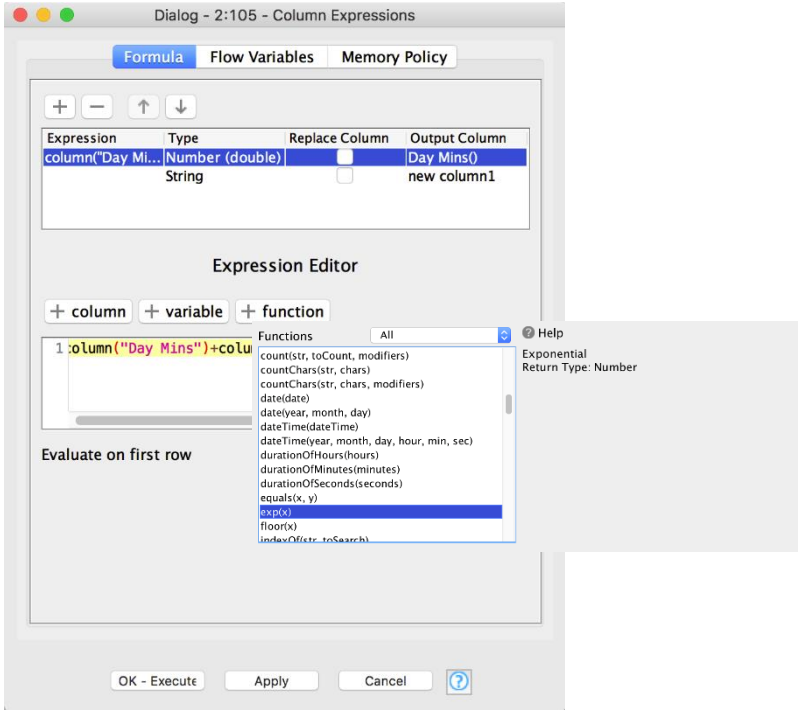


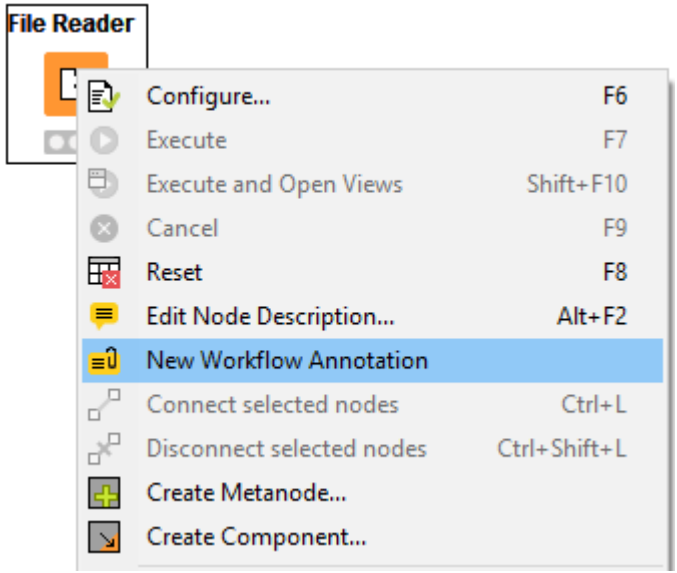
# New Node: Column Expression

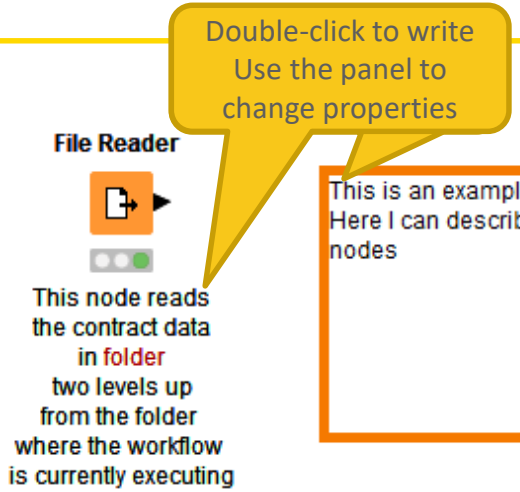
- Append or modify an arbitrary number of columns using expressions
- Many different functions are available
- No restriction on number of lines per expression allow to write complex expressions
- Part of the KNIME Labs extension

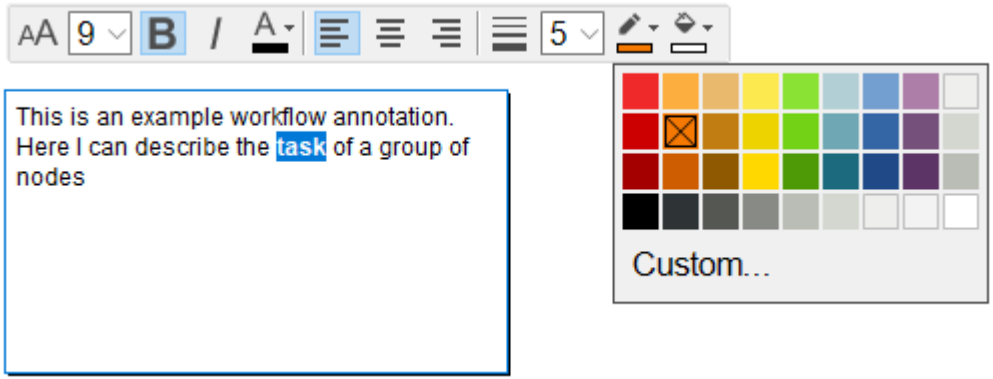



# Workflow Organization and Documentation

# Comments & Annotations



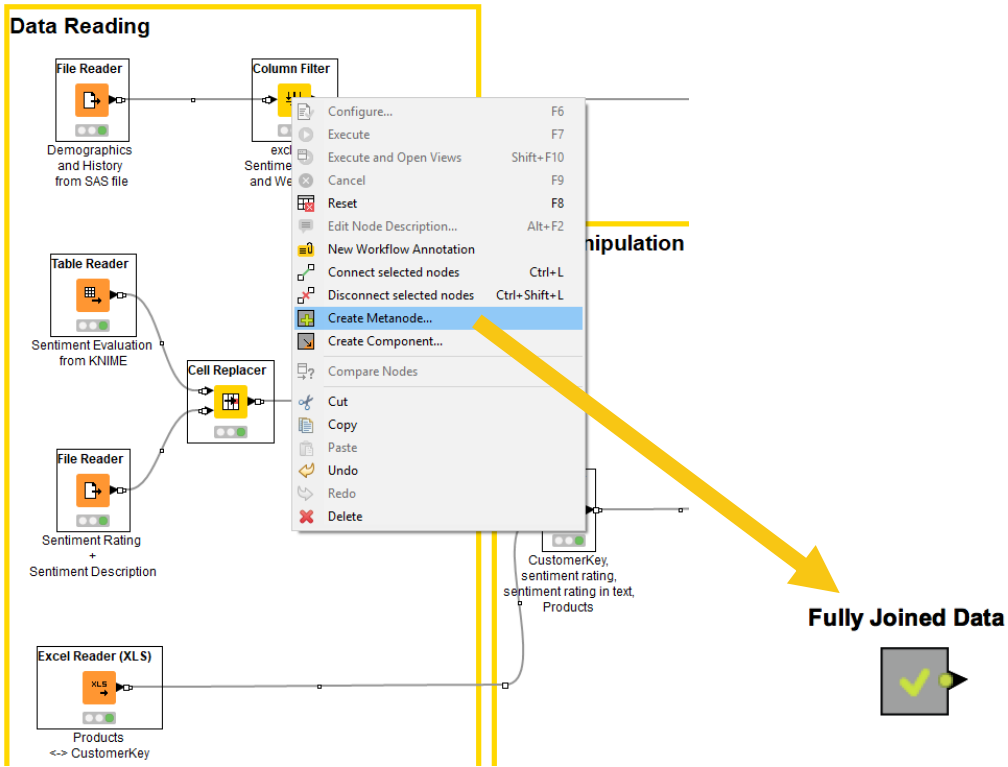




YouTube KNIME TV Channel:
https://youtu.be/AHURYB_O8sA

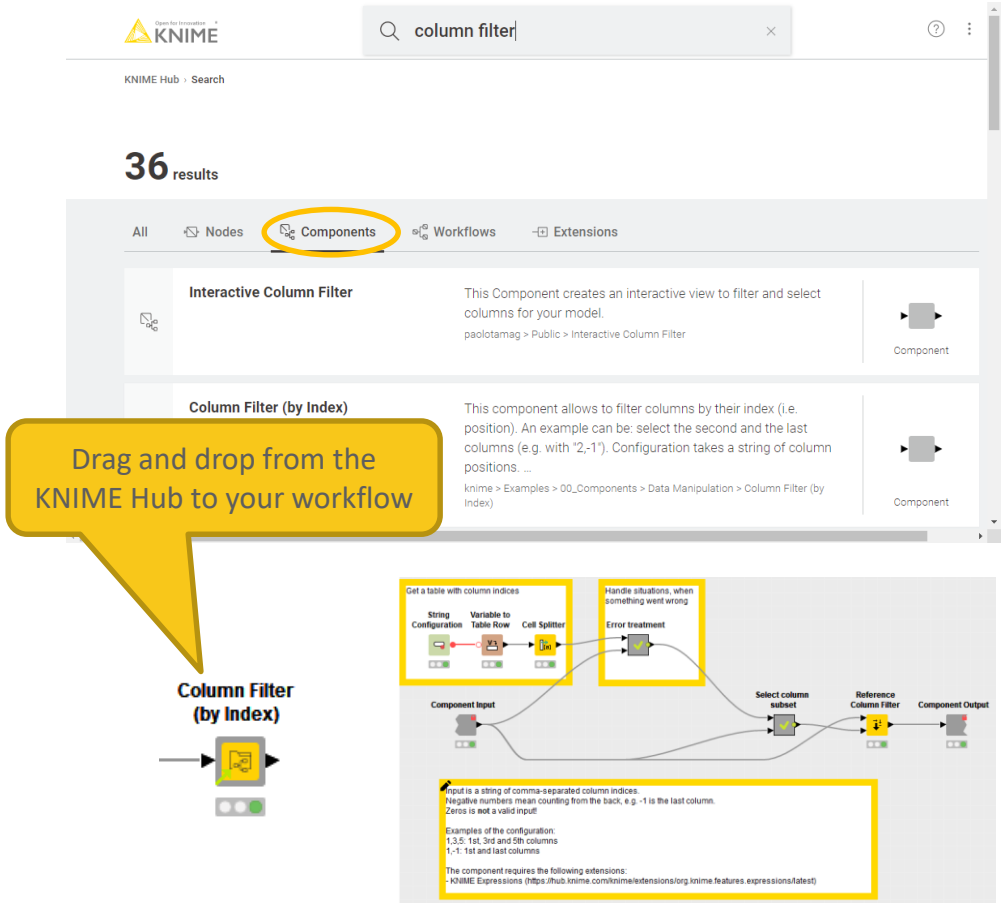
# Workflow Organisation – Good Practices

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



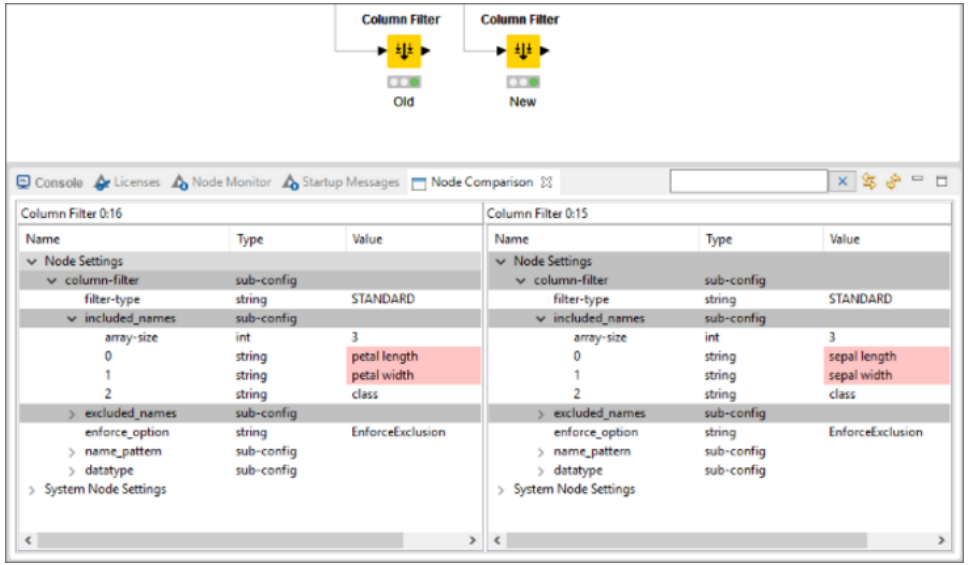
# Workflow Organisation – Components

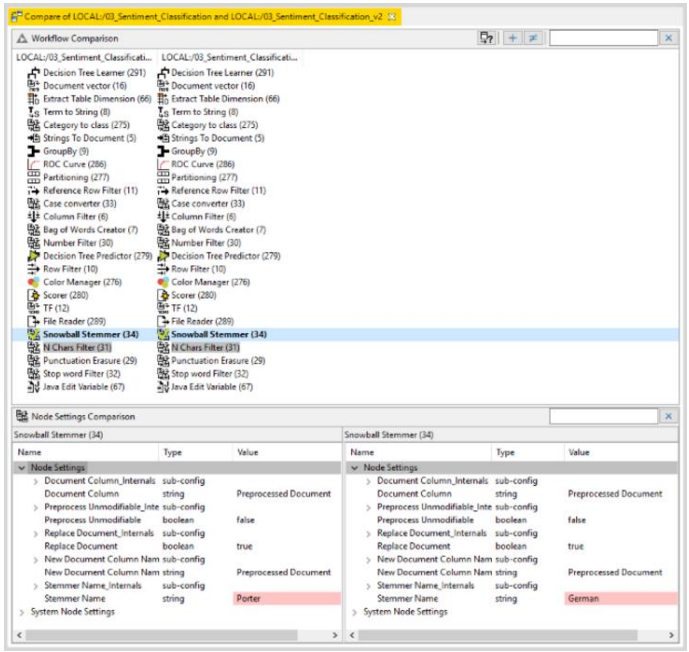
- Component encapsulates a reusable functionality as a KNIME workflow
- Components can be configured as any KNIME nodes
- Access and share components on the KNIME Hub



# KNIME WorkflowDiff

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



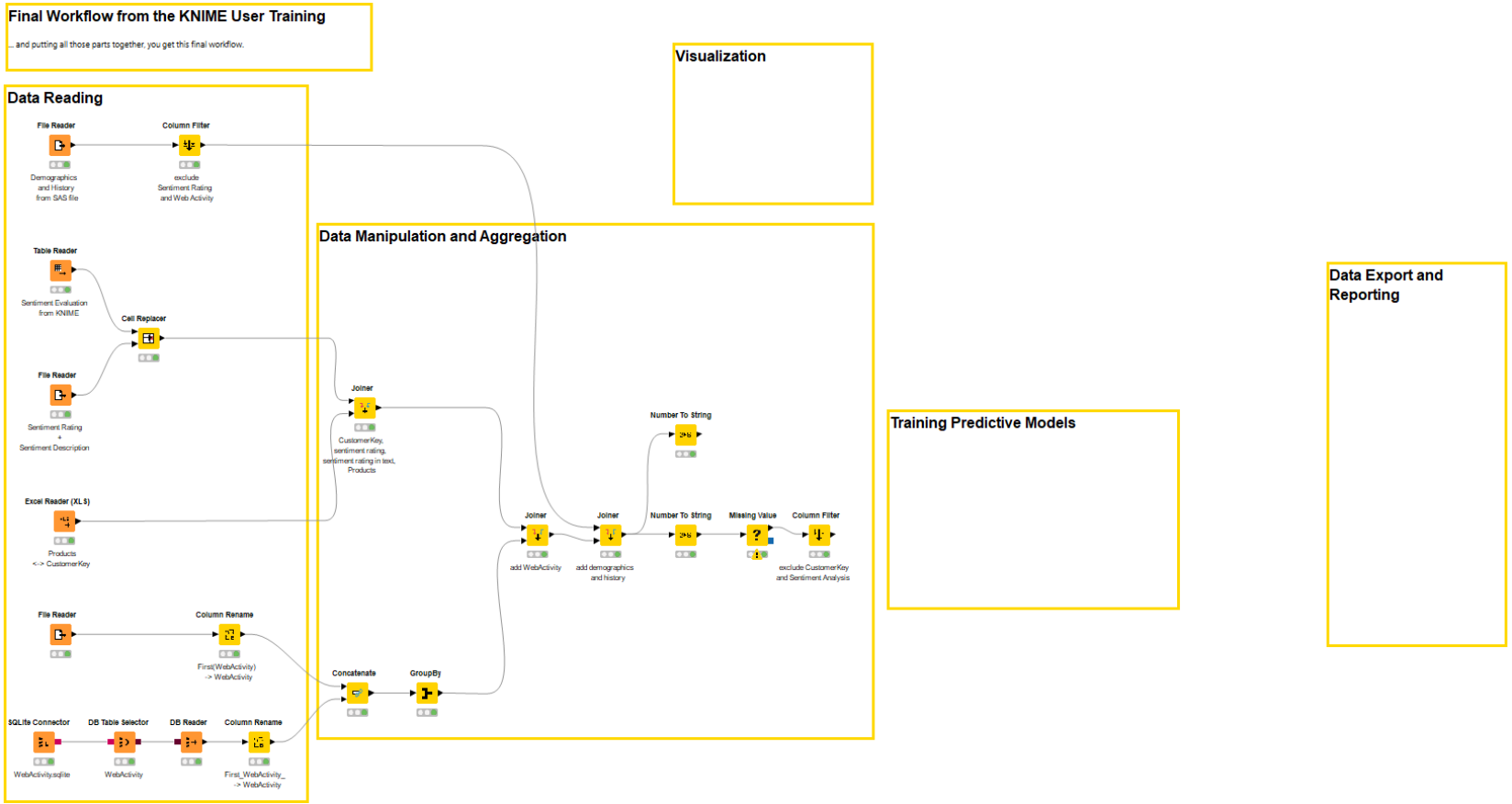



# Data Manipulation Exercise, Activity II

Start with exercise *Data Manipulation, Activity II*

- Join all data into one table using a series of joiner nodes (use "Customer Key" as the joining column)
- Filter out duplicate rows
- Clean up and document your workflow using annotations, node labels, and metanodes

# Today's Example

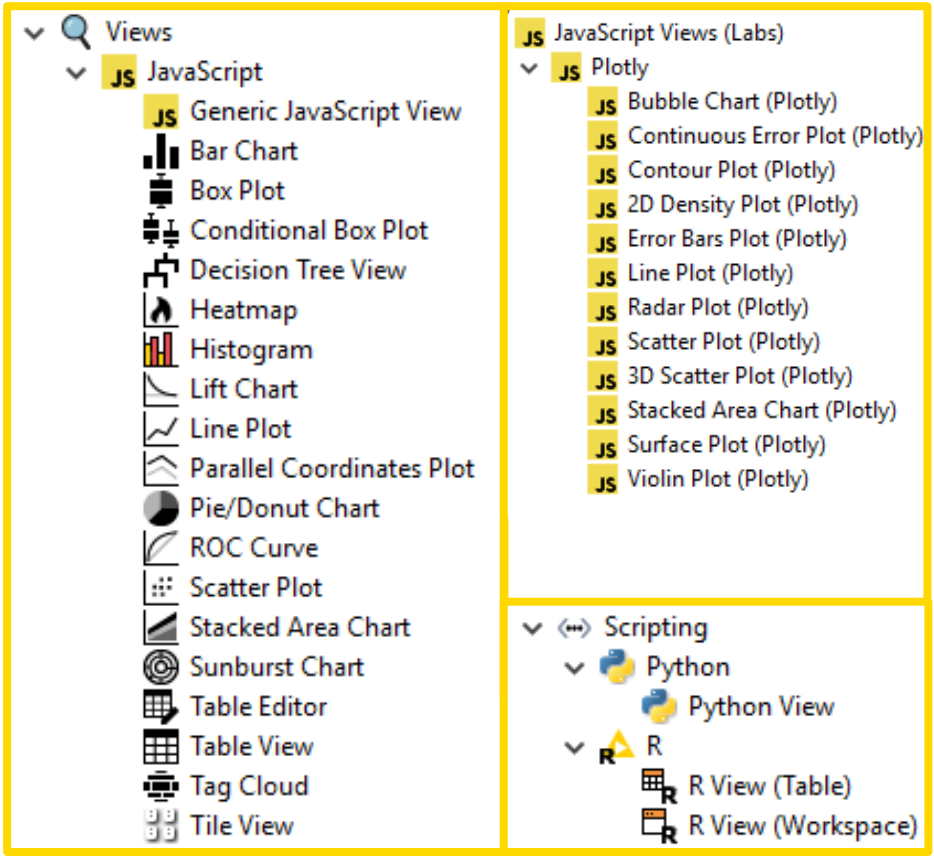


# Data Visualization

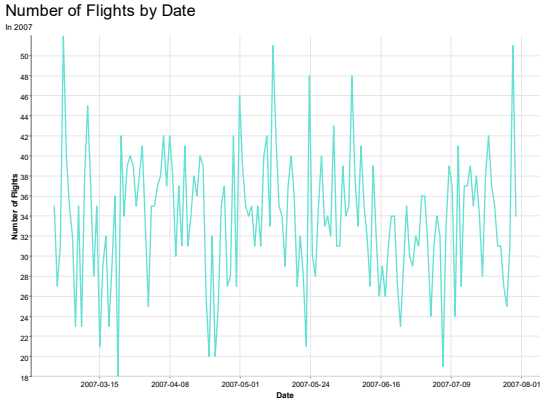
## Charts and Tables

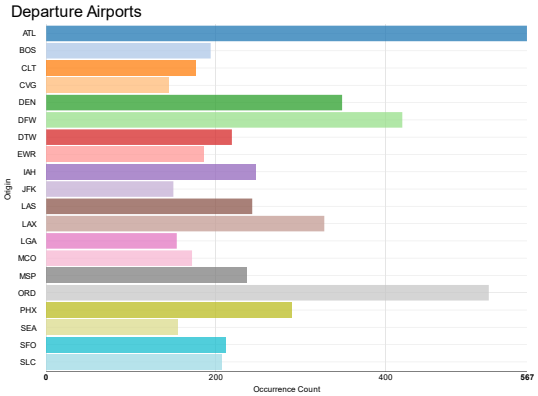
# Data Visualization

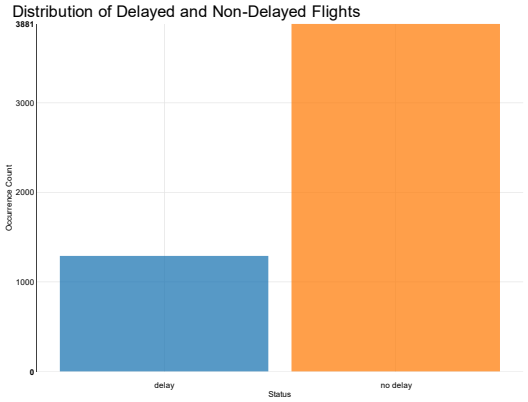
- Large selection of easy to use visualization nodes
  - Web-based and interactive
  - Dedicated nodes, no scripting required
- Plotly nodes
  - Similar but integrated from an external library
- R and Python View nodes for highly customizable graphics
  - Require scripting

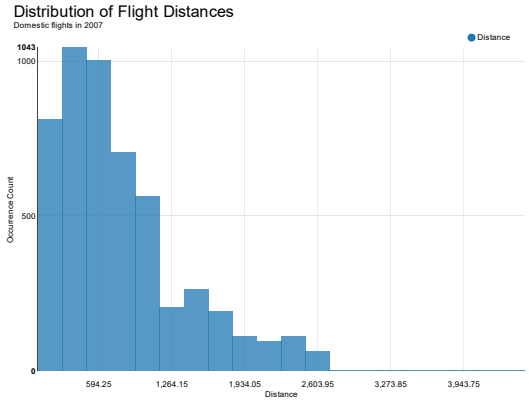


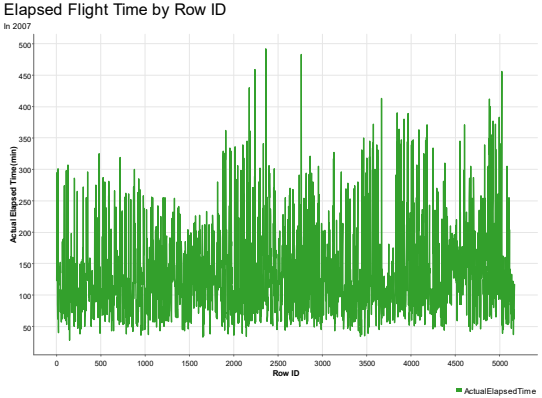
# Visualizations using 1 Column

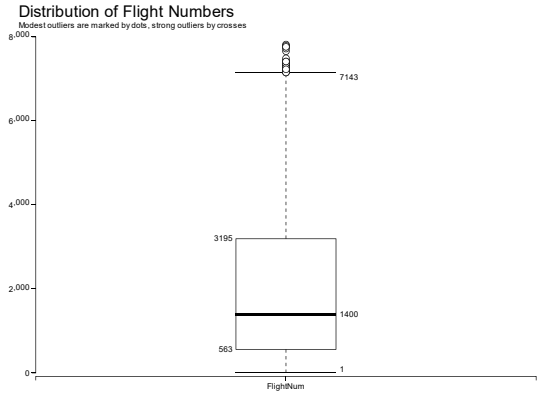


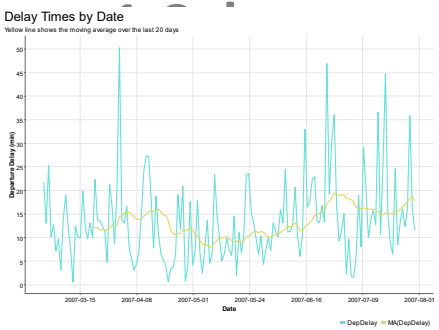



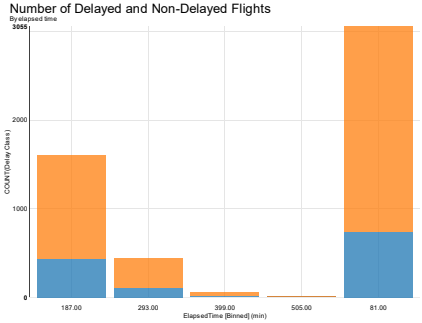



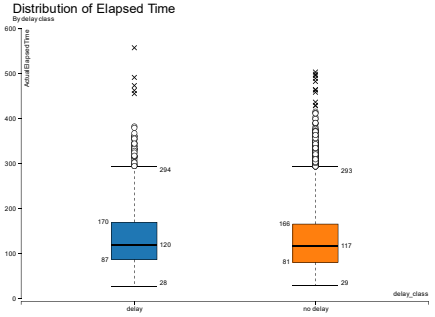



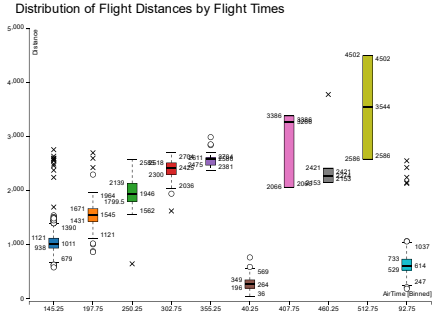



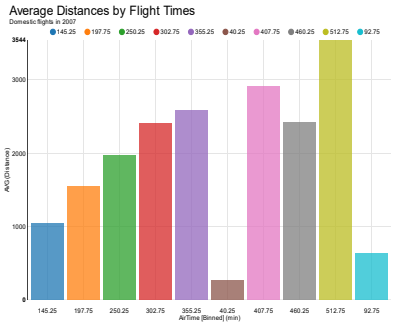
# Visualizations using 2 Columns

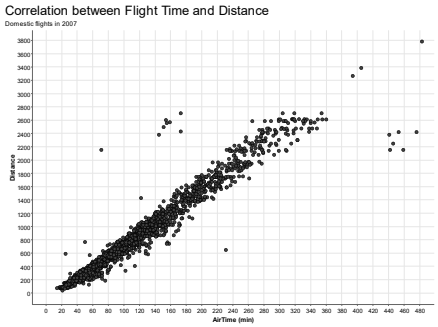


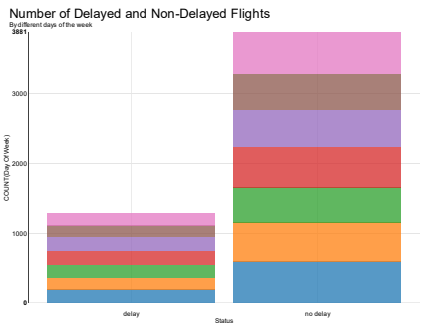



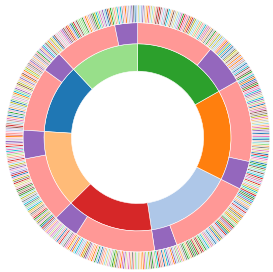



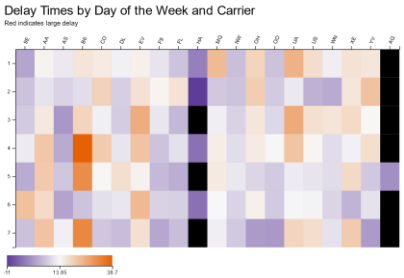



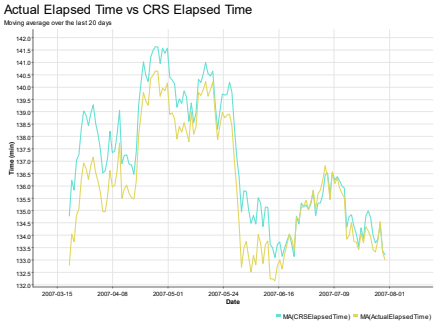



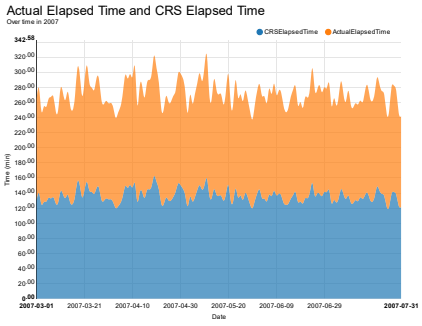



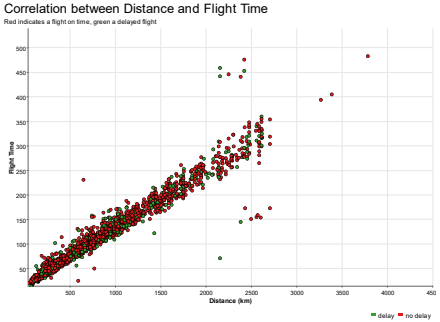



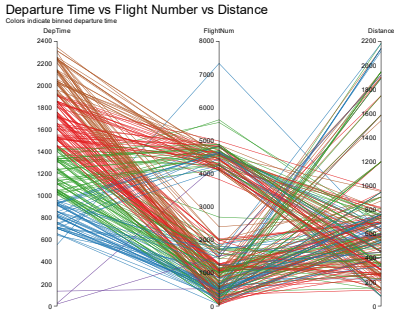
# Visualizations using 3 Columns

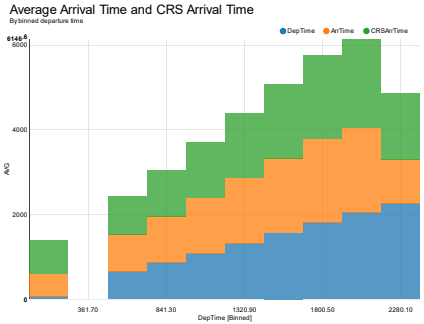


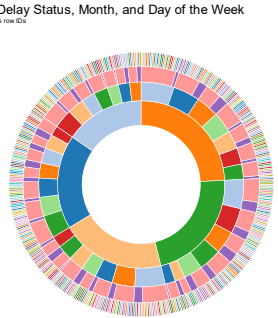



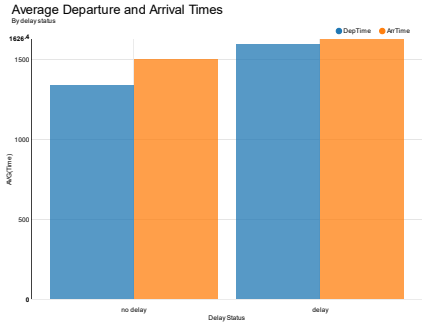



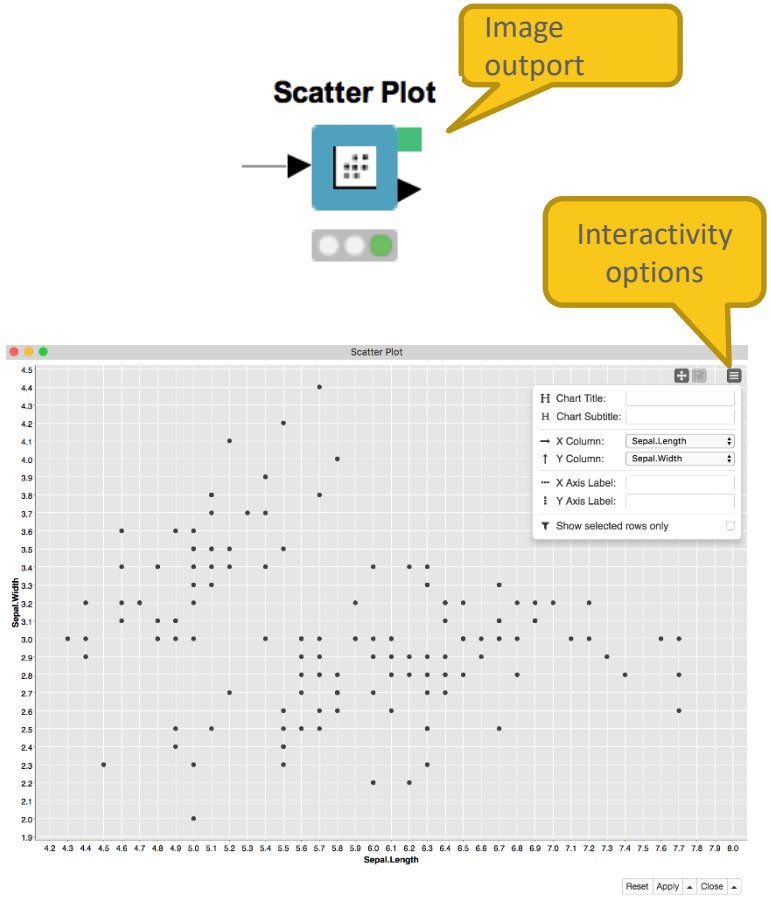



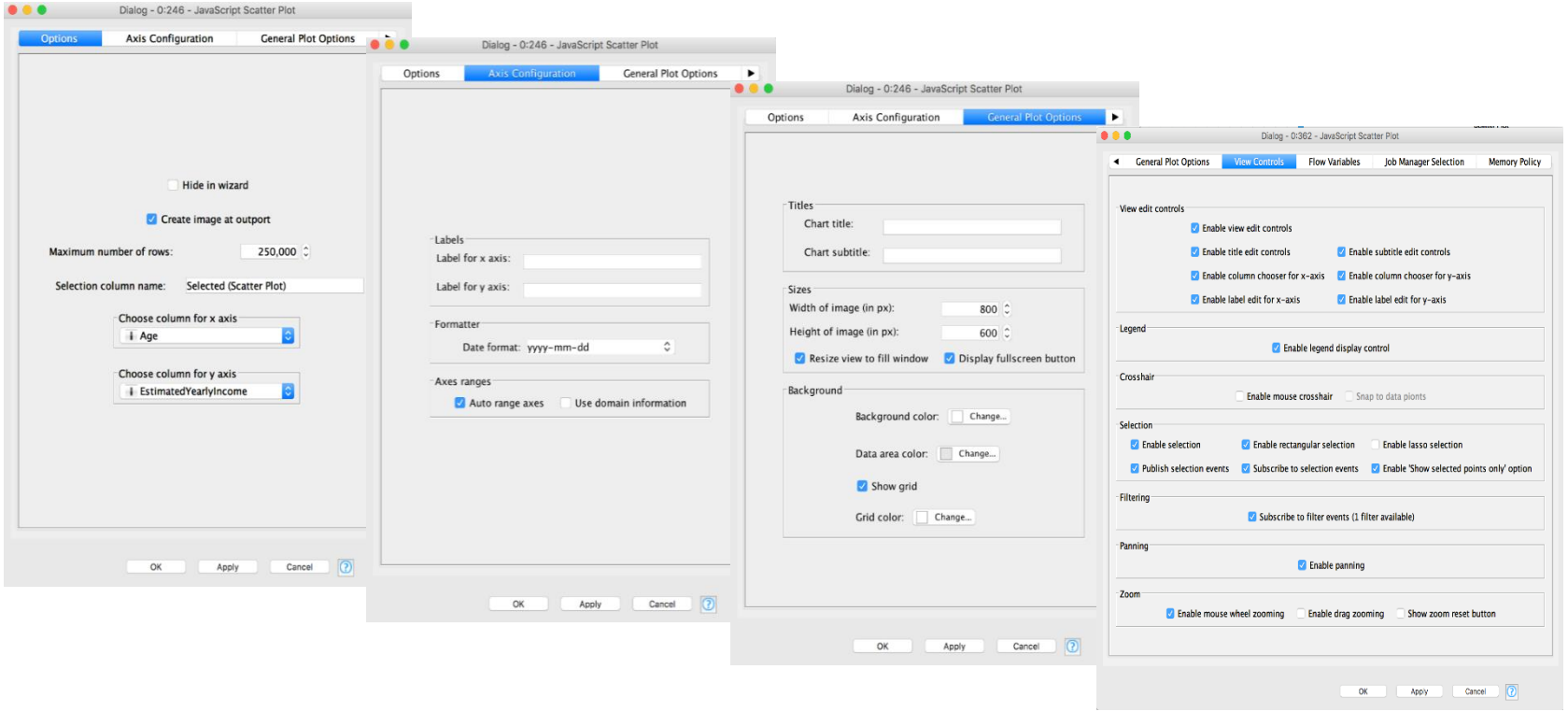
# New Node: Scatter Plot

- Plots 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



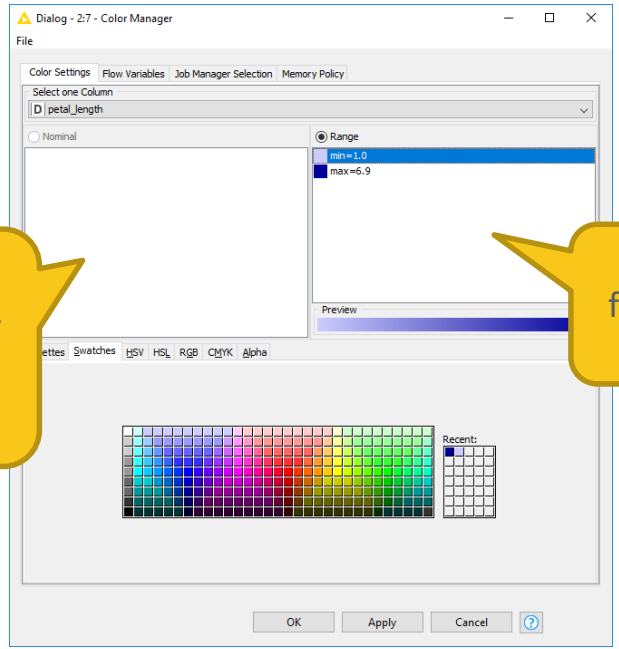
# New Node: Scatter Plot

## Four configuration tabs



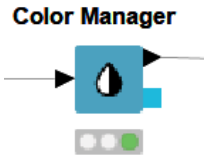
# New Node: Color Manager

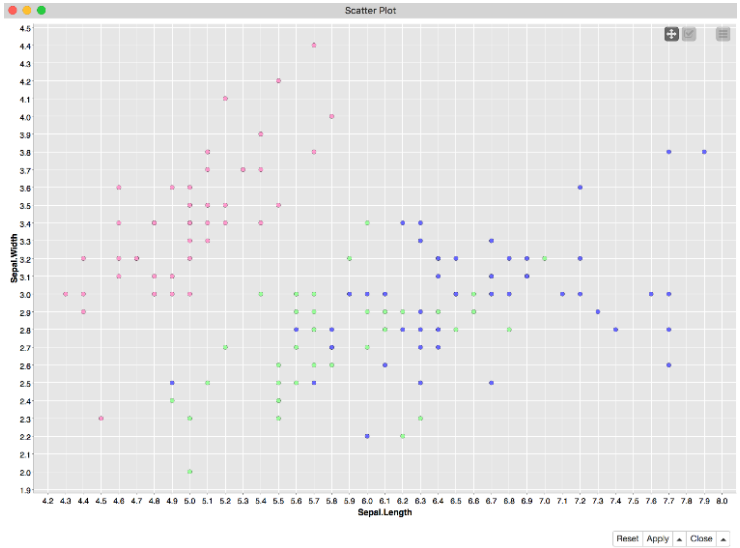
- Color by nominal or continuous values
- Sync colors between views using the color model port and Color Appender node




Discrete colors for nominal values

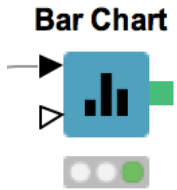
Color range for numerical values

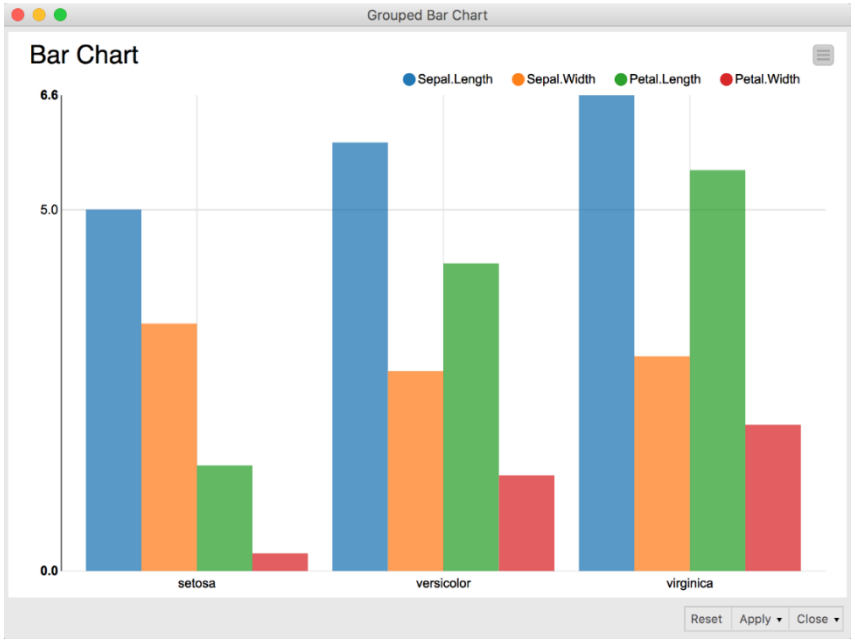


# New Node: Bar Chart

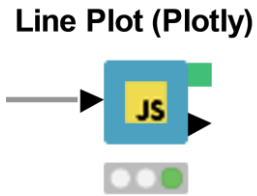
- Show numerical values across categories
- Vertical or horizontal bars
- Bars can be grouped or stacked

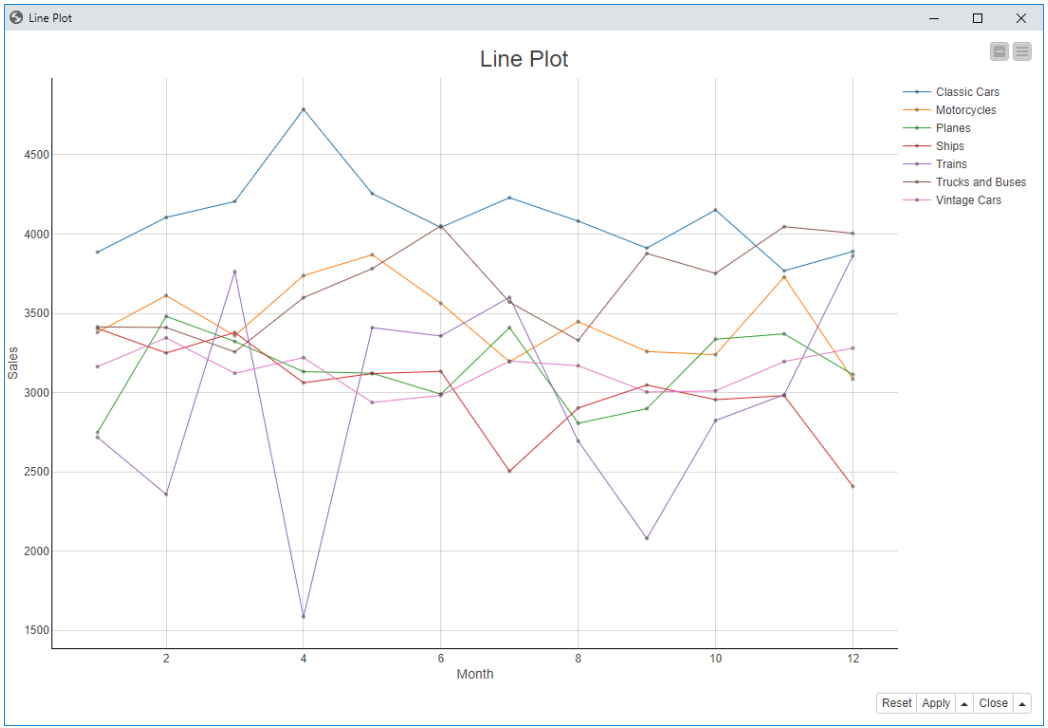


# New Node: Line Plot

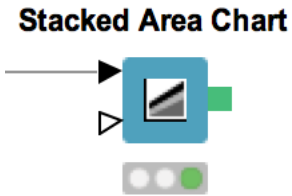
- Plot sequence of values, e.g. over time
- Useful to identify trends, also between groups

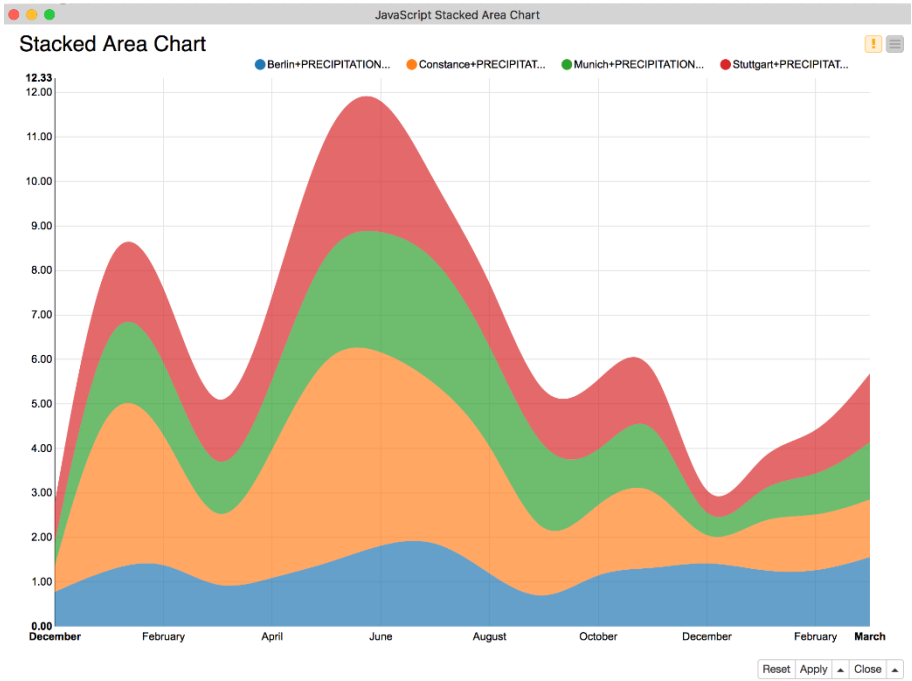


# New Node: Stacked Area Chart

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

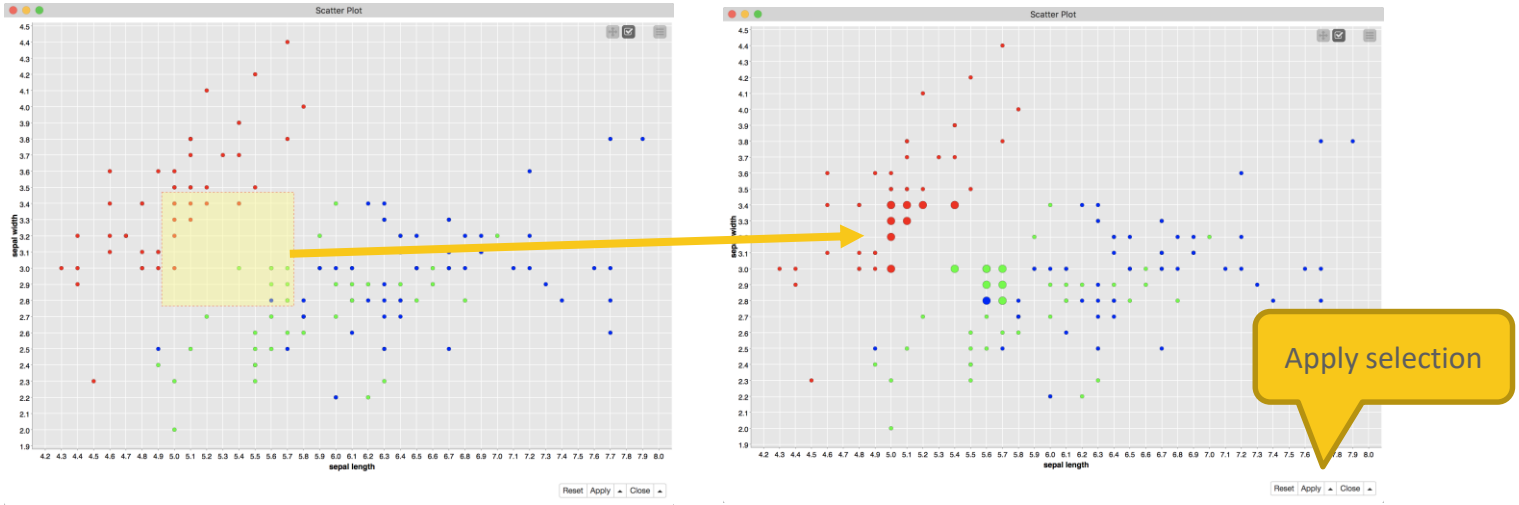


# Selection & Filtering in JavaScript Views

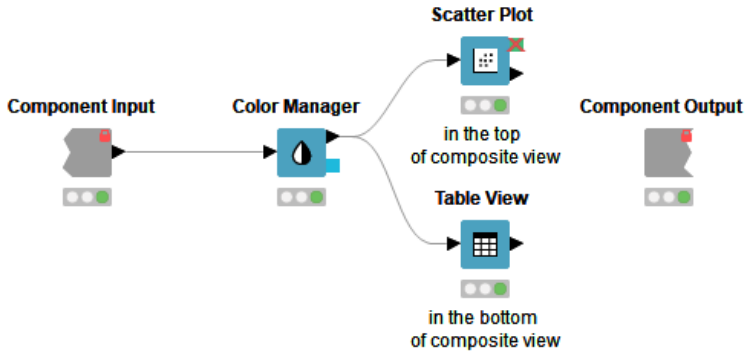
Interactivity allows you to select data points in views

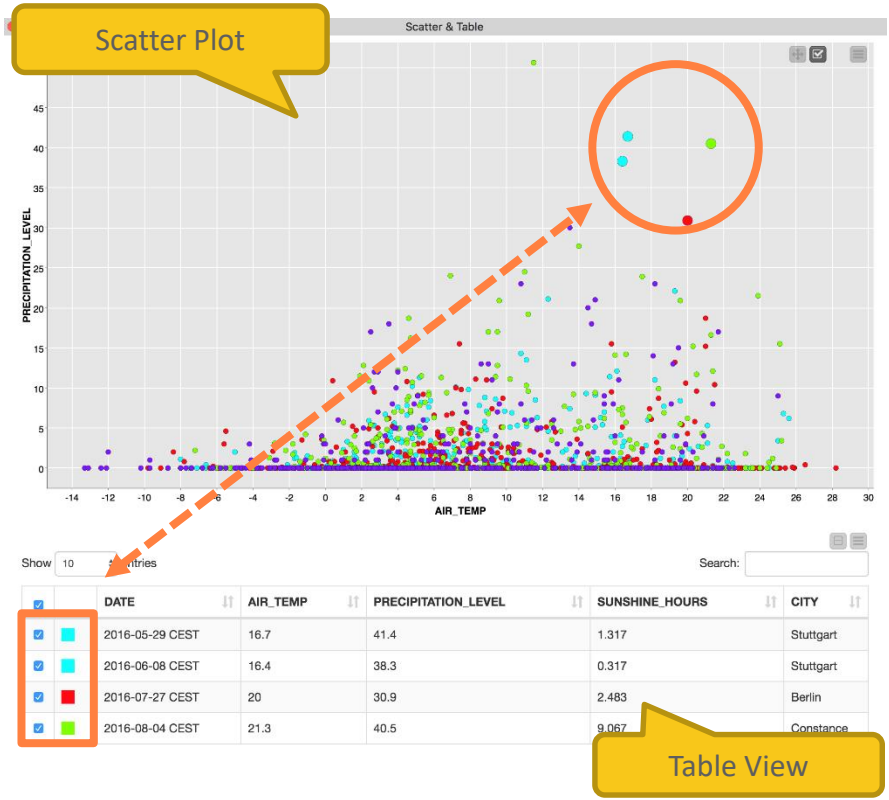
- Selection is propagated to other views
- Highlight selected rows or filter them
- Click “Apply” to add column to data that indicates selection (true/false) for use in downstream nodes



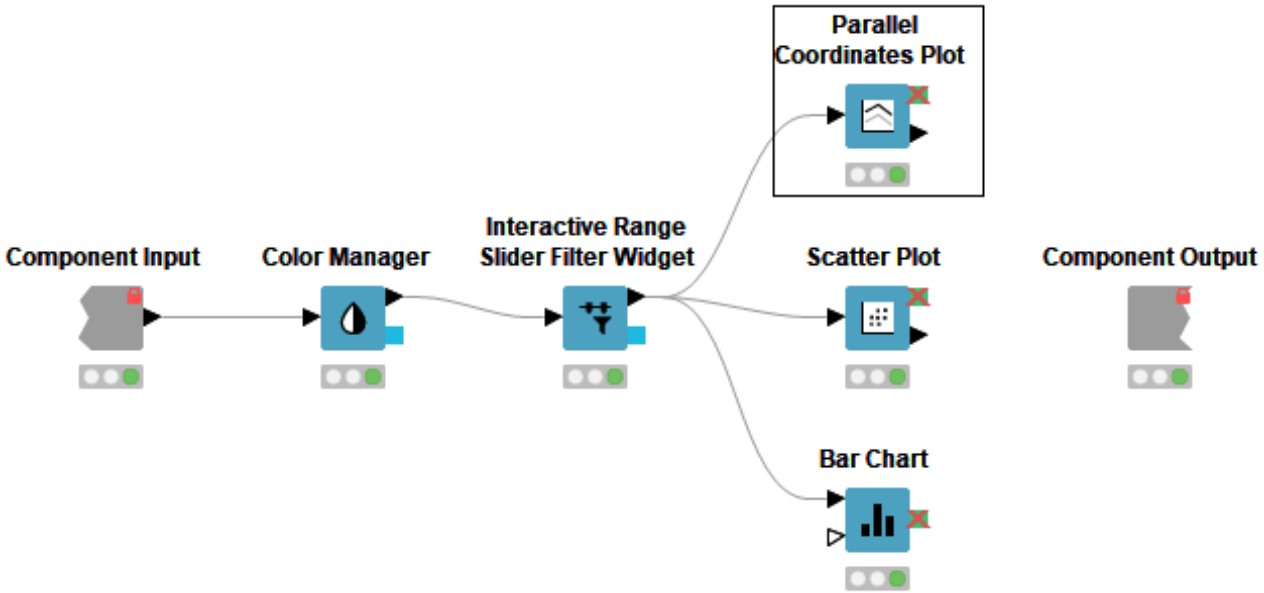
# Components – Combined Views

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

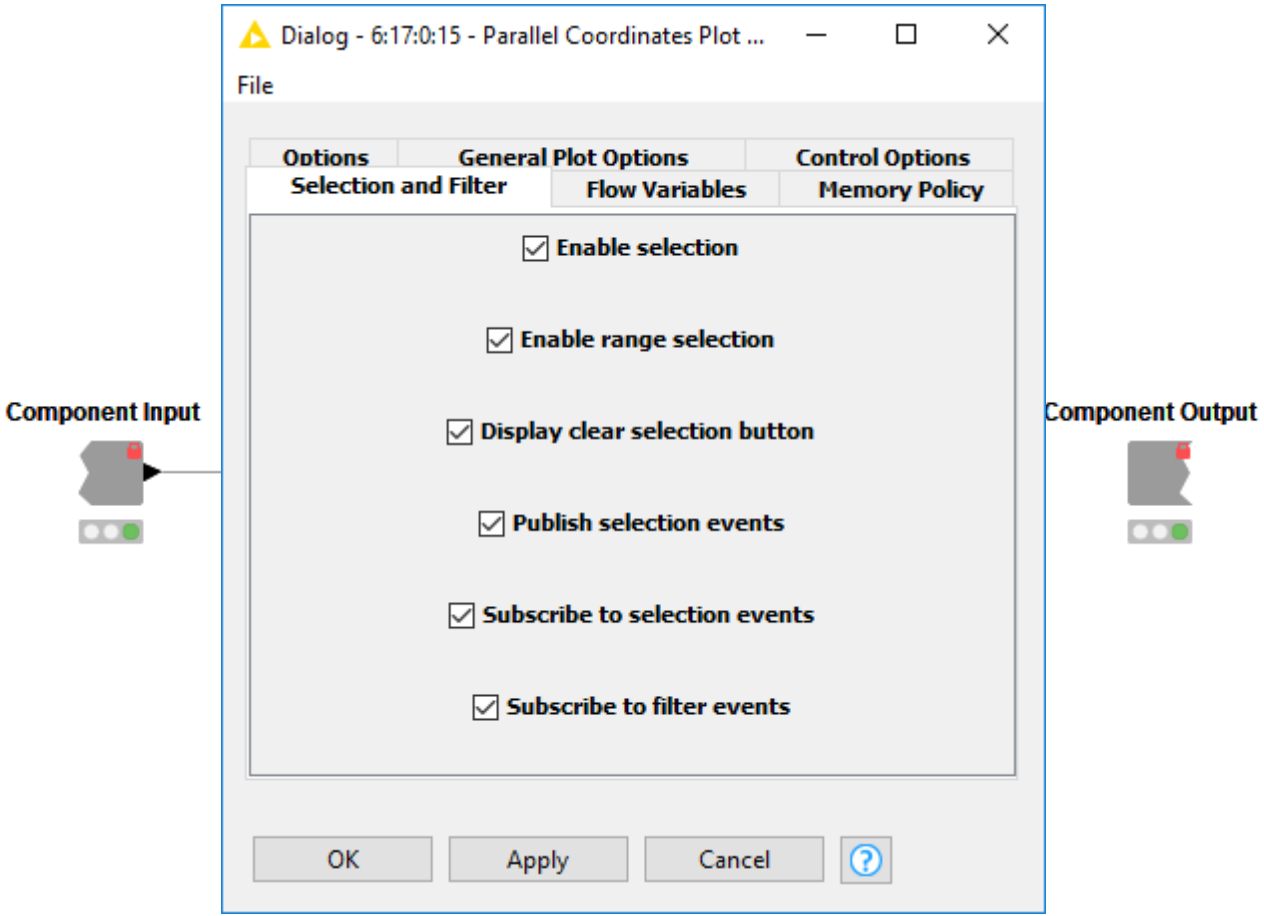


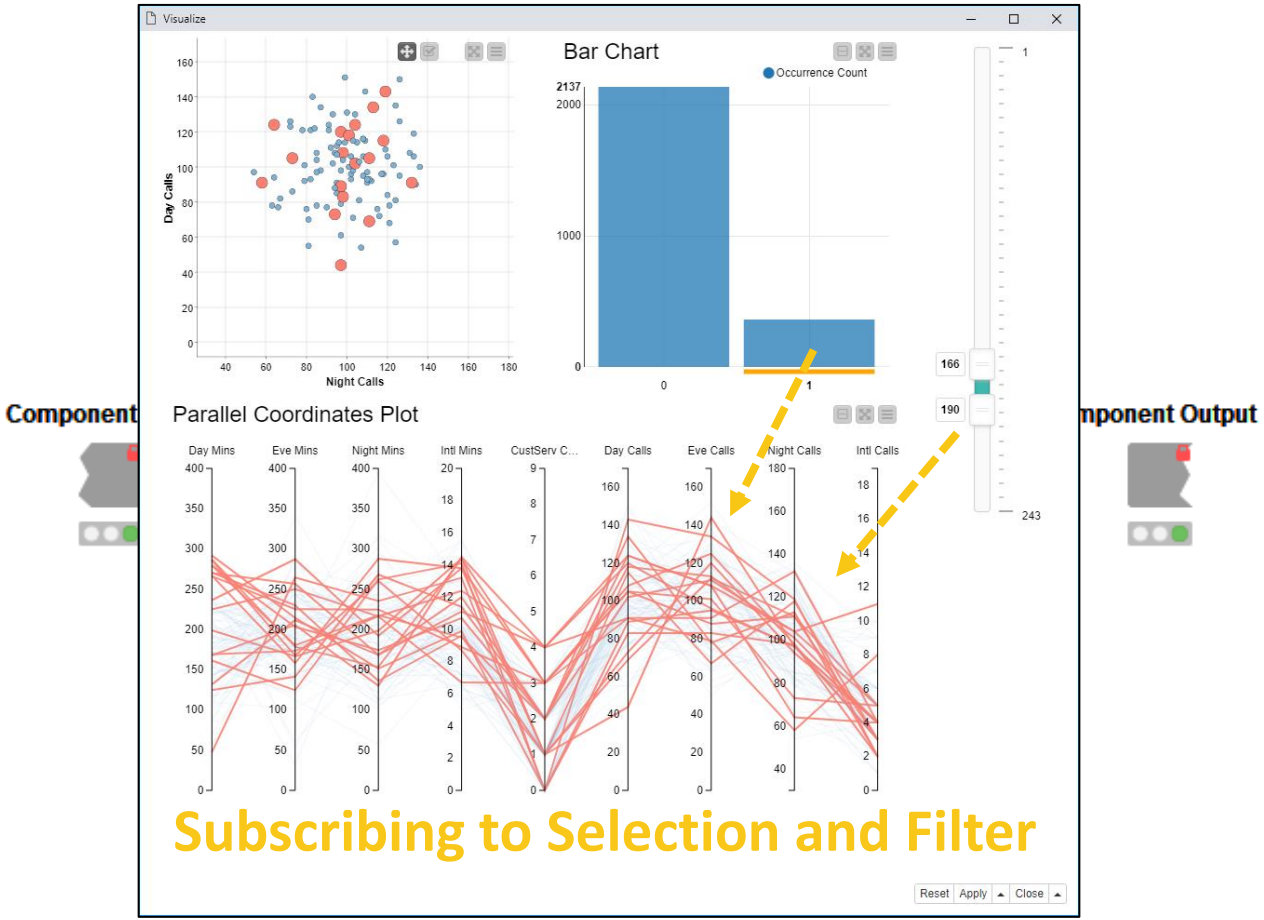
# Interactivity across Charts: Selection and Filter Events



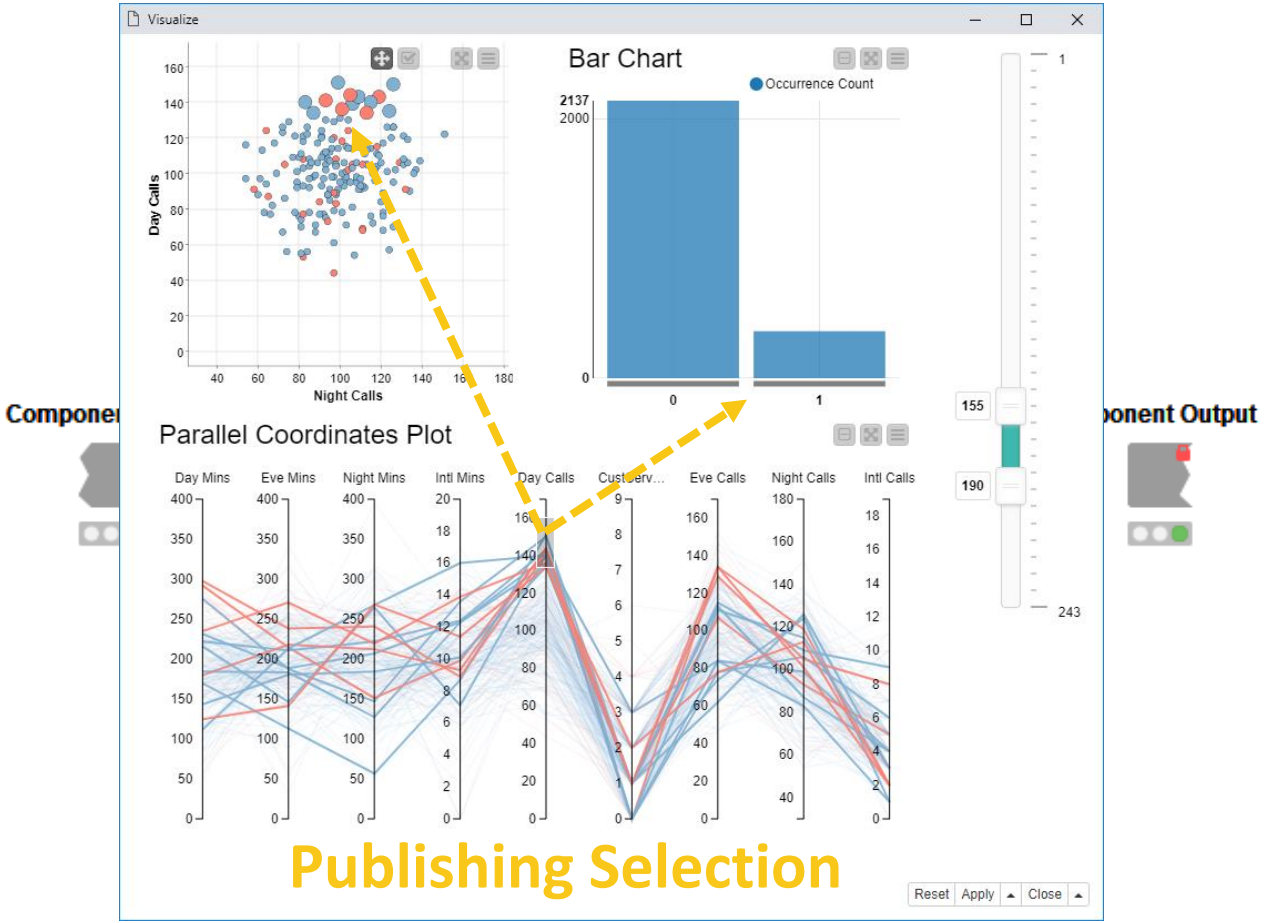
# Interactivity across Charts: Selection and Filter Events



# Interactivity across Charts: Selection and Filter Events

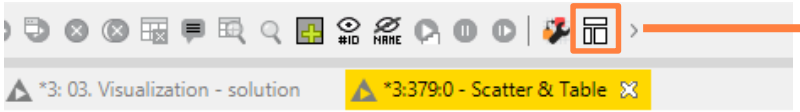


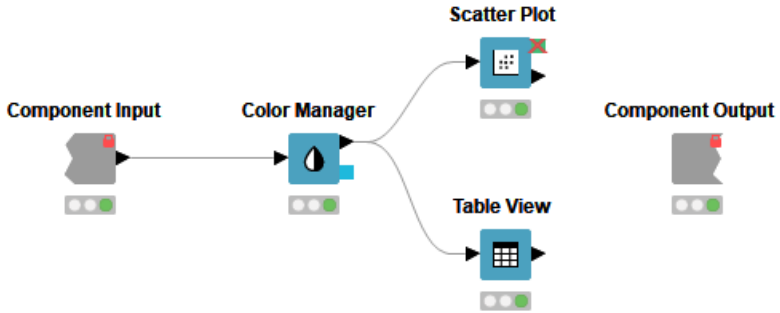
# Interactivity across Charts: Selection and Filter Events



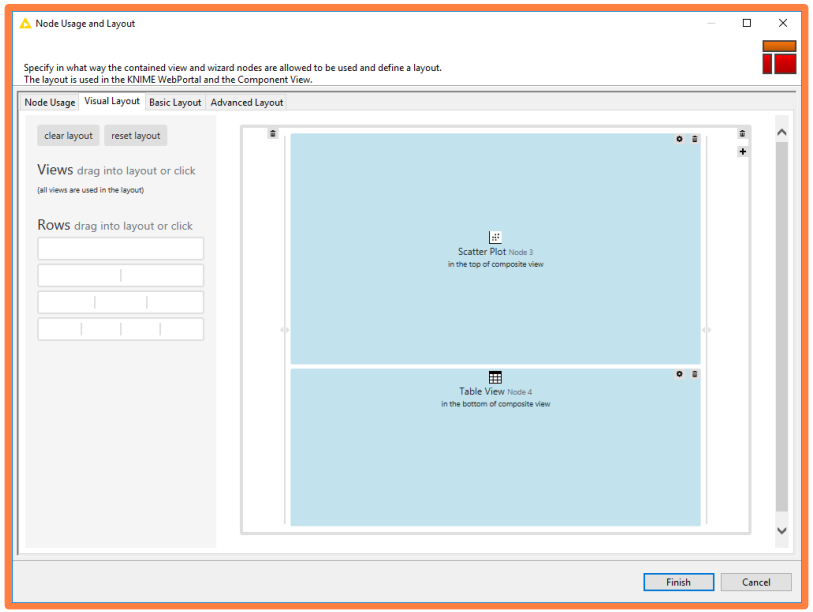
# Configure Content and Views Layout

- Click layout button when inside Component to assign views to rows and columns




- Add views and rows via *drag&drop*
- Add columns using *+* buttons



# Data Aggregation

| Product ID | Store | Category | # Ordered Items |
|---|---|---|---|
| P 1 | Online | Clothing | 2 |
| P 2 | Onsite | Home | 3 |
| P 3 | Onsite | Clothing | 1 |
| P 4 | Online | Clothing | 5 |
| P 5 | Online | Electronics | 7 |
| P 6 | Online | Electronics | 5 |

Aggregation: Count

| Category | Online | Onsite |
|---|---|---|
| Clothing | 2 | 1 |
| Home | 0 | 1 |
| Electronics | 2 | 0 |

Aggregation: Sum (# Ordered Items)

| Category | Online | Onsite |
|---|---|---|
| Clothing | 7 | 1 |
| Home | 0 | 3 |
| Electronics | 12 | 0 |

**Solution: Pivoting Node**

# Data Aggregation

| Product ID | Store | Category | # Ordered Items |
|---|---|---|---|
| P 1 | Online | Clothing | 2 |
| P 2 | Onsite | Home | 3 |
| P 3 | Onsite | Clothing | 1 |
| P 4 | Online | Clothing | 5 |
| P 5 | Online | Electronics | 7 |
| P 6 | Online | Electronics | 5 |

Aggregation: Sum (# Ordered Items)

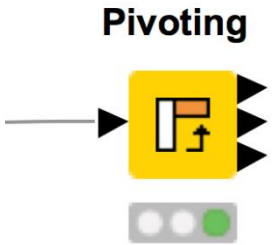
| Category | Online | Onsite |
|---|---|---|
| Clothing | 7 | 1 |
| Home | 0 | 3 |
| Electronics | 12 | 0 |

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

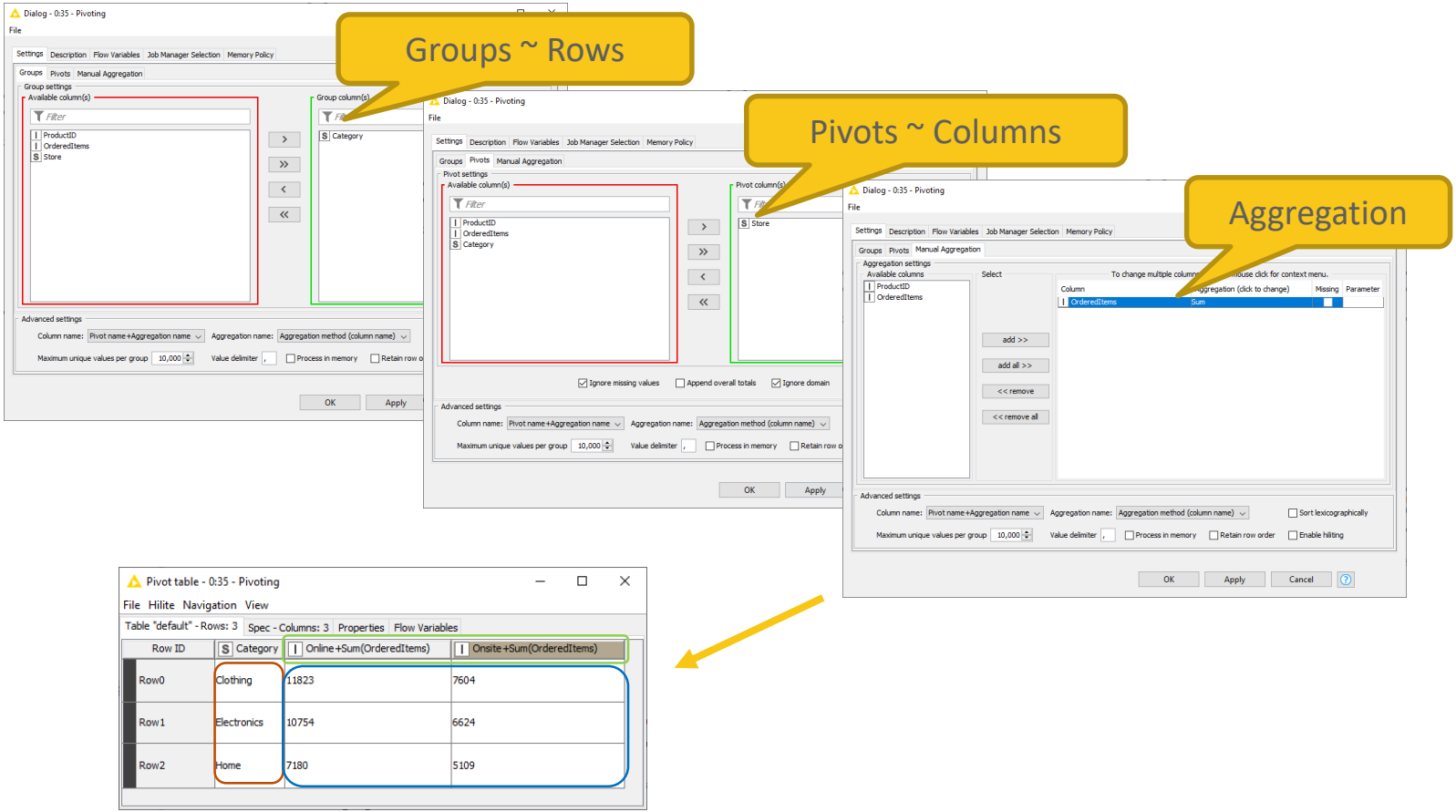
# New Node: Pivoting

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)

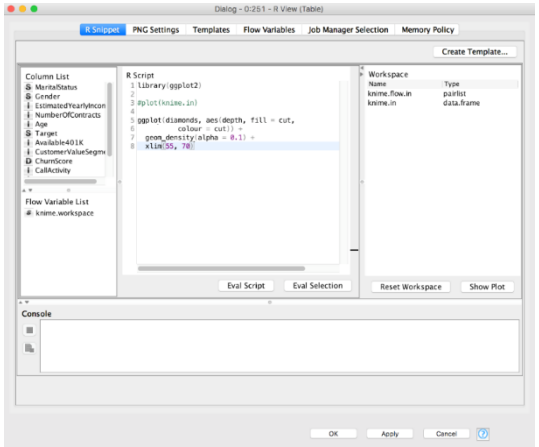


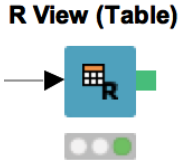
# New Node: Pivoting

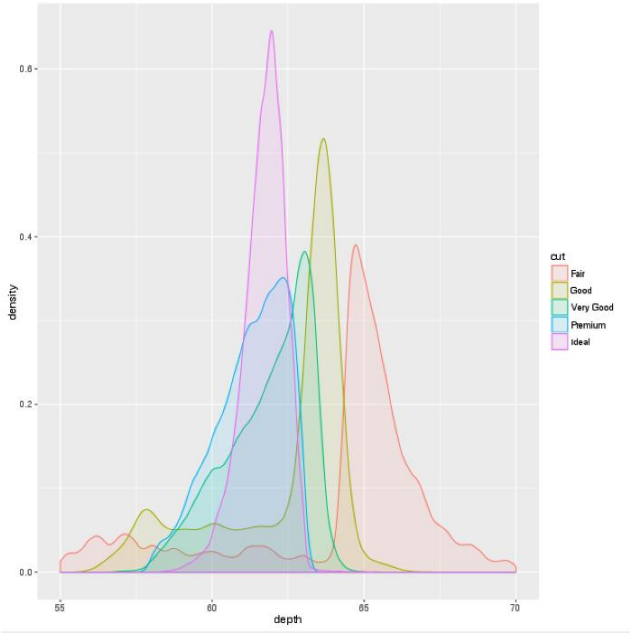


# Script-based View Nodes

- 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



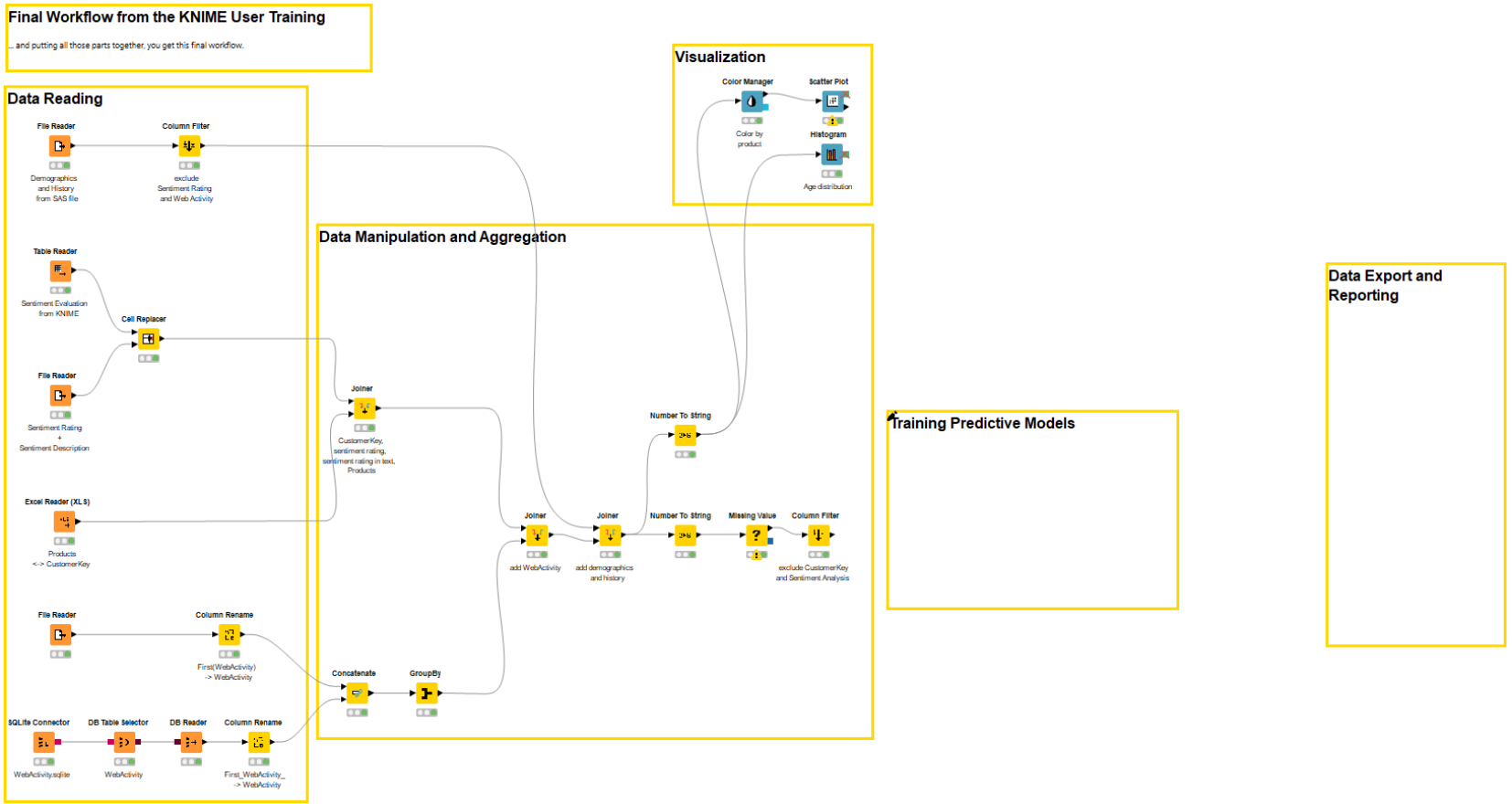
# Visualization Exercise

Start with exercise: *Visualization*

- Read *sales.csv* data
- Assign a different color to each product
- Plot BasketValue against BasketSize using the Scatter Plot node
- Show the total BasketValue by time and product in a Line Plot and a Stacked Area Chart (Use the Pivoting node to get the sum of sales by Quarter and Product!)

- Execute the *Fully Joined Data* metanode
- Show the number of customers in the different web activity categories in a Bar Chart
- Show the age distribution of the customers in a Histogram
- Create a composite view by combining the Bar Chart and Histogram
- Select one web activity class in the Bar Chart. Which age classes are represented in the selected web activity class?

# Today's Example



# Data Mining

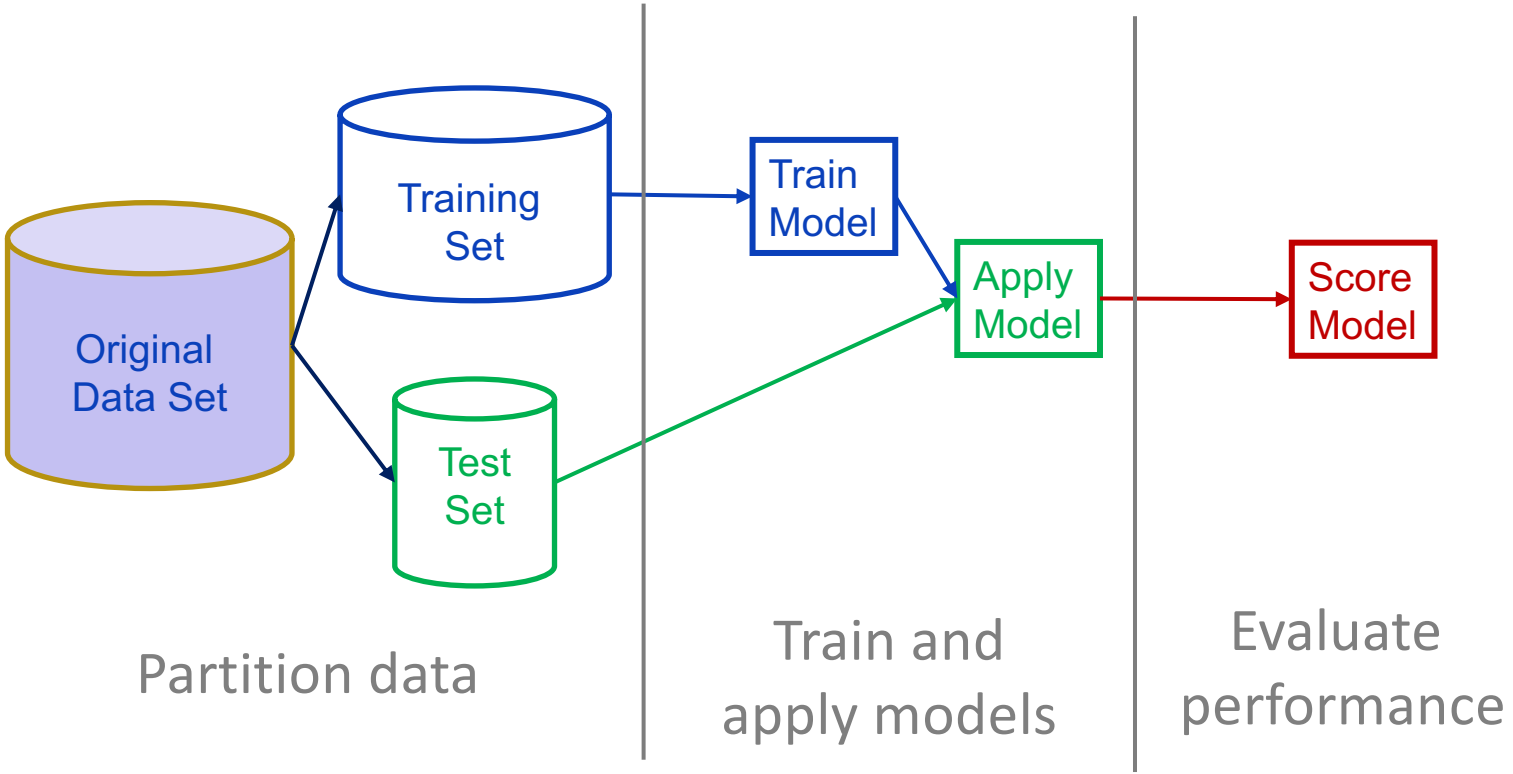
## Partition, Learn, Predict, Score

# Data Mining Strategies

Example Applications:

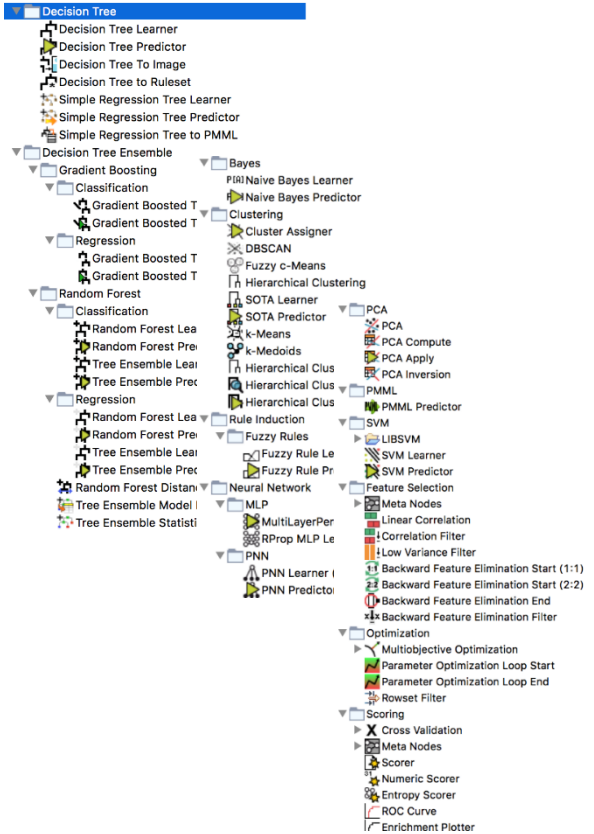
- Anomaly Detection (fraud, predictive maintenance)
- Association Rule Learning (market basket analysis)
- Clustering (market segmentation)
- Classification (next best offer, churn preventions)
- Regression (trend estimation)

# Data Mining: Process Overview



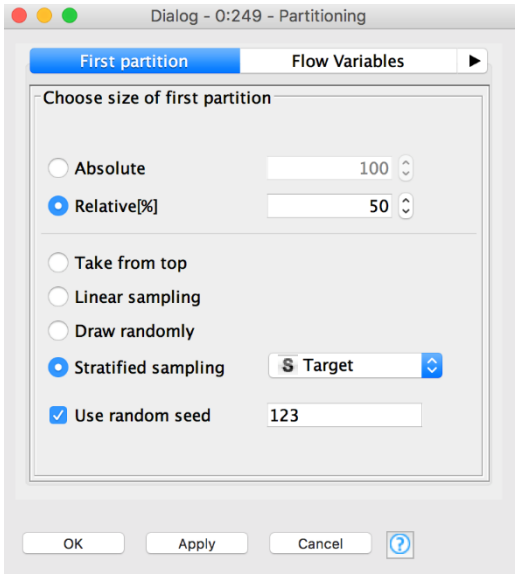
# Data Mining in KNIME

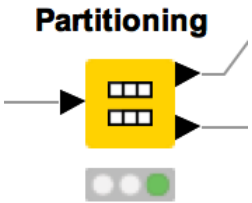
- KNIME has many modeling tools!
  - Decision tree, random forest, SVM, regression, neural networks, clustering, …
  - and integrations with other libraries: R, Python, H2O, WEKA, libSVM, etc.
- And many model evaluation nodes
  - ROC, standard, numeric and entropy scorers
  - Feature elimination
  -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

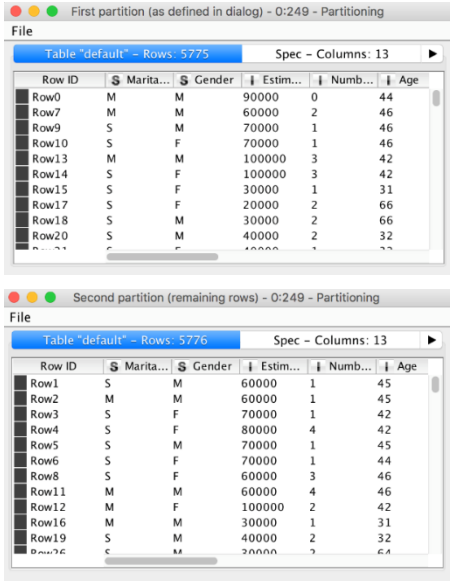


# New Node: Partitioning

- 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

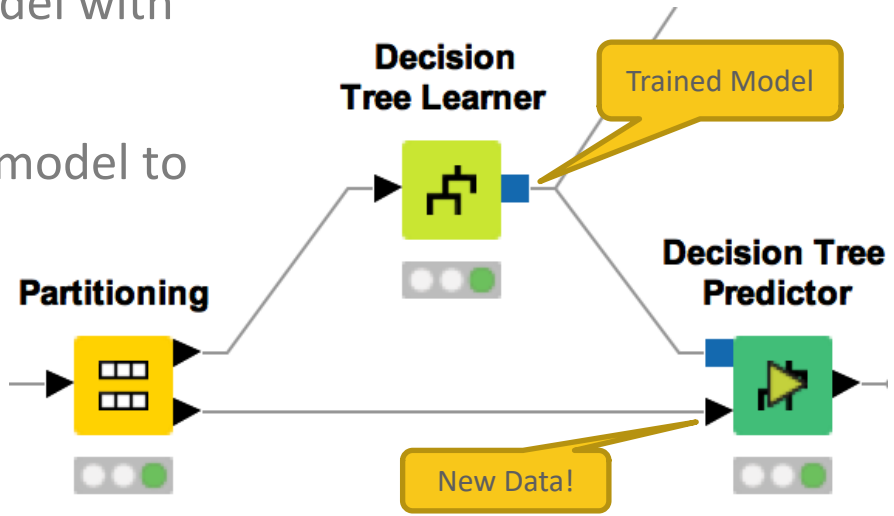


# Learner-Predictor Motif

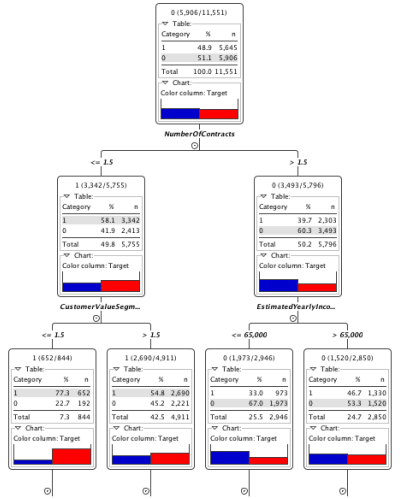
- Most data mining approaches in KNIME use a Learner-predictor motif.
- The Learner node trains the model with its input data.
- The Predictor node applies the model to a different subset of data.



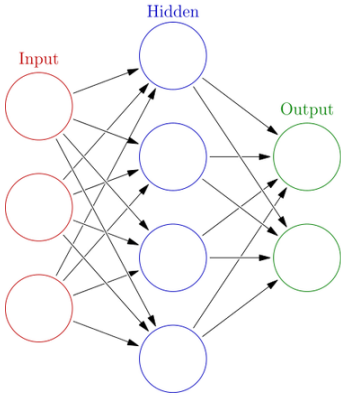
# Classification

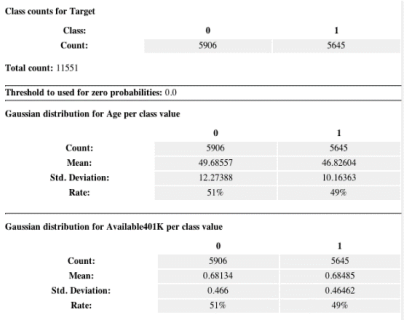
Predict *nominal* outcomes on existing data (supervised)

- 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



**Class counts for Target**

| Class: | 0 | 1 |
|---|---|---|
| Count: | 5906 | 5645 |

**Total count:** 11551

**Threshold to used for zero probabilities:** 0.0

**Gaussian distribution for Age per class value**

| | 0 | 1 |
|---|---|---|
| Count: | 5906 | 5645 |
| Mean: | 49.68557 | 46.82604 |
| Std. Deviation: | 12.27388 | 10.16363 |
| Rate: | 51% | 49% |

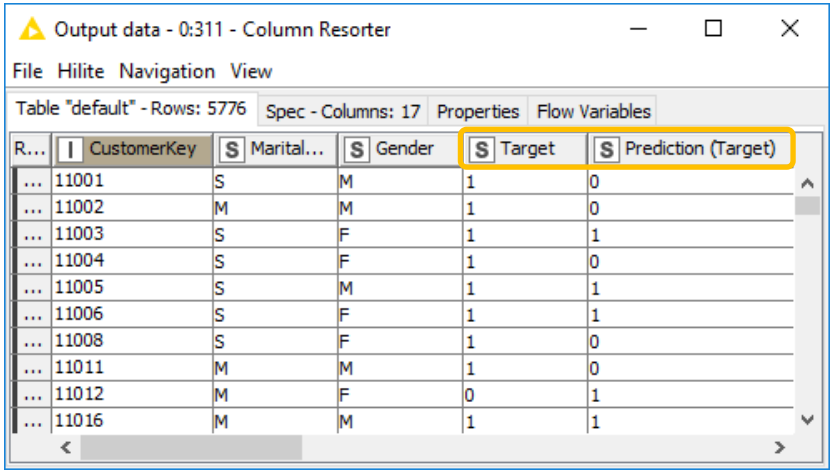
**Gaussian distribution for Available401K per class value**

| | 0 | 1 |
|---|---|---|
| Count: | 5906 | 5645 |
| Mean: | 0.68134 | 0.68485 |
| Std. Deviation: | 0.466 | 0.46462 |
| Rate: | 51% | 49% |

# Target Column

- Target column contains values that are predicted by the classification model
- Binomial target values are often encoded to 1 and 0

| Application | Target Column | Target Values |
|---|---|---|
| Churn analysis | Churn | Yes/No or 1/0 |
| Chemical activity | Active | Yes/No or 1/0 |
| Spam Detection | Spam | Yes/No or 1/0 |
| Optical Character Recognition | Character | A-Z |




| R... | CustomerKey | Marital... | Gender | Target | Prediction (Target) |
|---|---|---|---|---|---|
| ... | 11001 | S | M | 1 | 0 |
| ... | 11002 | M | M | 1 | 0 |
| ... | 11003 | S | F | 1 | 1 |
| ... | 11004 | S | F | 1 | 0 |
| ... | 11005 | S | M | 1 | 1 |
| ... | 11006 | S | F | 1 | 1 |
| ... | 11008 | S | F | 1 | 0 |
| ... | 11011 | M | M | 1 | 0 |
| ... | 11012 | M | F | 0 | 1 |
| ... | 11016 | M | M | 1 | 1 |

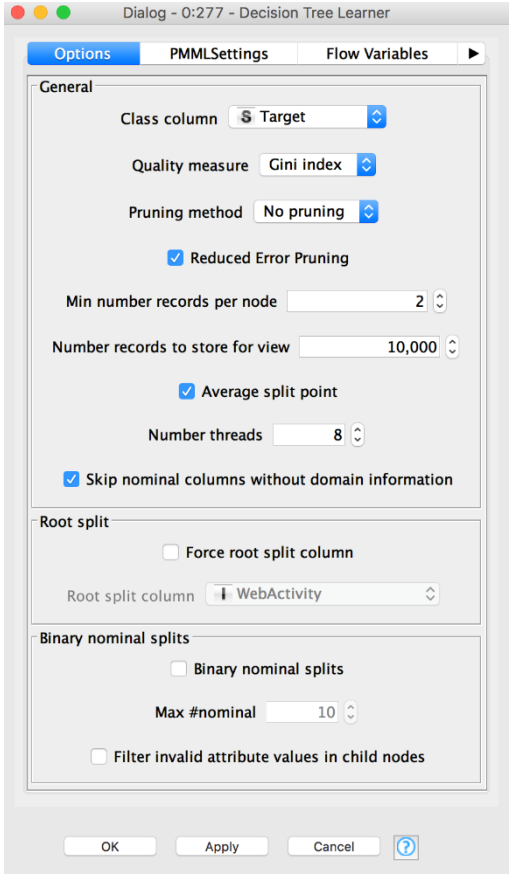
# KNIME's Decision Tree

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

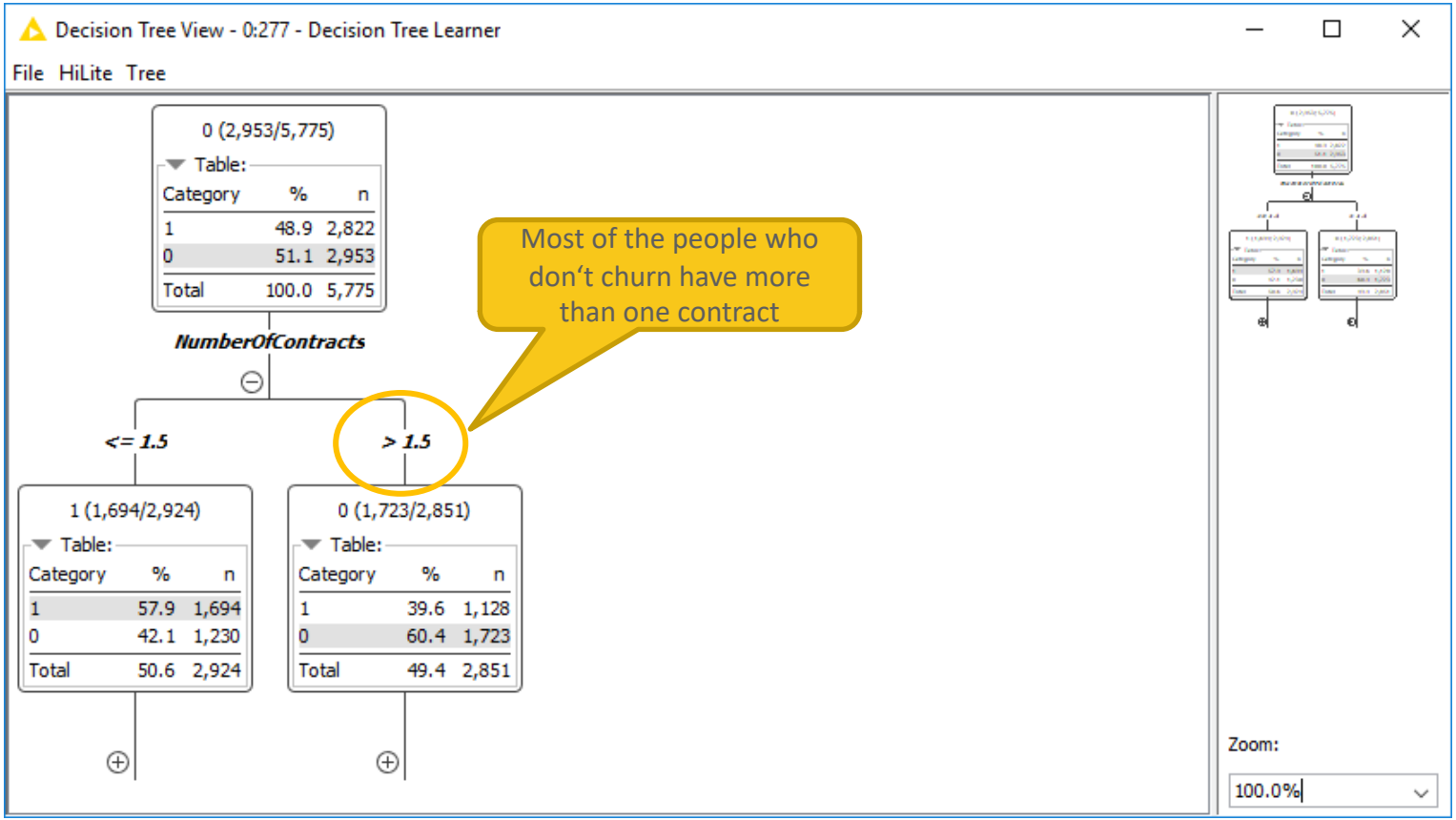
J. Shafer, R. Agrawal, M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining"

- 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

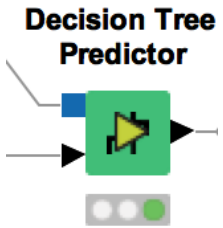



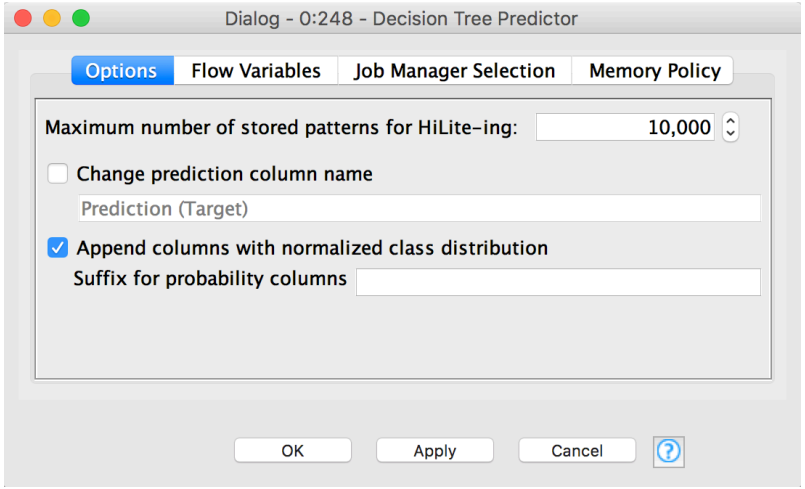
# Decision Tree View



# New Node: Decision Tree Predictor

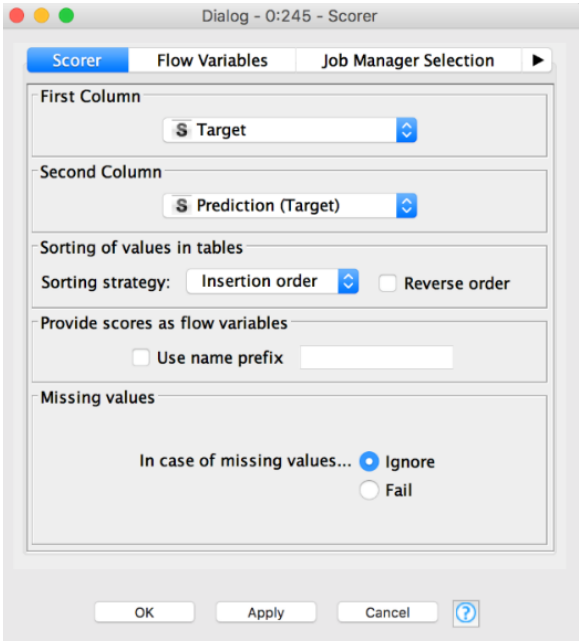
- Takes a decision tree model & applies it to new data
- Check the box to append class probabilities



# New Node: Scorer

Compare predicted results to known truth in order to evaluate model quality

# New Node: Scorer

Confusion matrix shows the distribution of model errors

Confusion Matrix - 0:297 - Scorer

File Hilite

| Target \ Prediction (Target) | 1 | 0 |
|---|---|---|
| 1 | 2073 | 750 |
| 0 | 759 | 2193 |

| | |
|---|---|
| Correct classified: 4,266 | Wrong classified: 1,509 |
| Accuracy: 73.87 % | Error: 26.13 % |
| Cohen's kappa (κ) 0.477 | |

An accuracy statistics table provides a detailed analysis of model quality

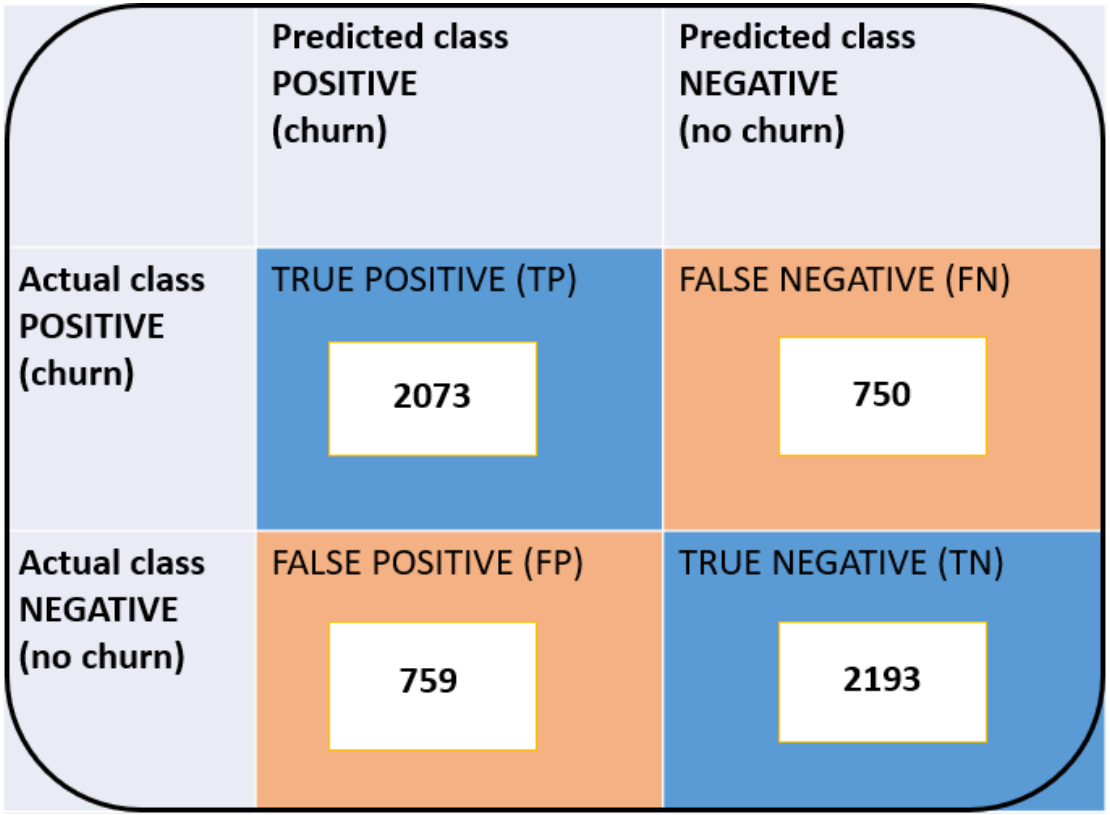
Accuracy statistics - 0:297 - Scorer

File Hilite Navigation View

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

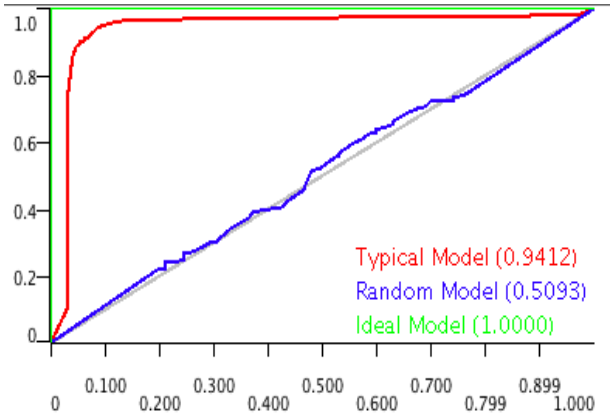
| Row ID | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specifity | F-measure | Accuracy | Cohen's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2073 | 759 | 2193 | 750 | 0.734 | 0.732 | 0.734 | 0.743 | 0.733 | ? | ? |
| 0 | 2193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

# Confusion Matrix



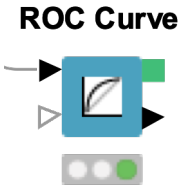
# Receiver Operating Characteristics

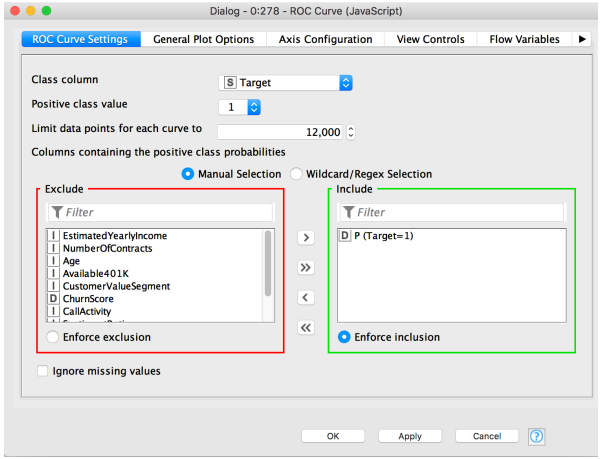
- Sort by confidence in target class
- Plot true positive rate vs false positive rate
- Ideal models achieve 100% TPR with 0% FPR
- Area under the curve indicates model quality
  - (1=ideal model, 0.5 = random outcome)

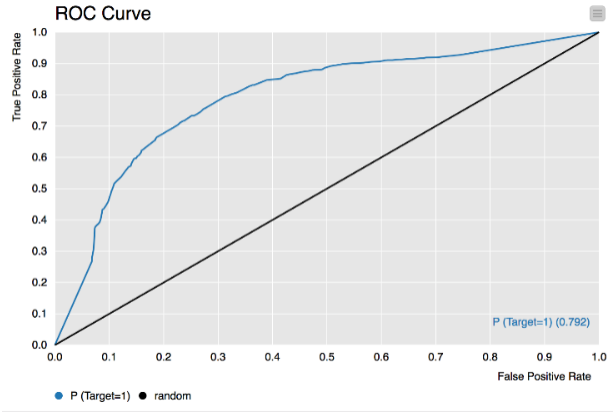


# New Node: ROC Curve

- 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

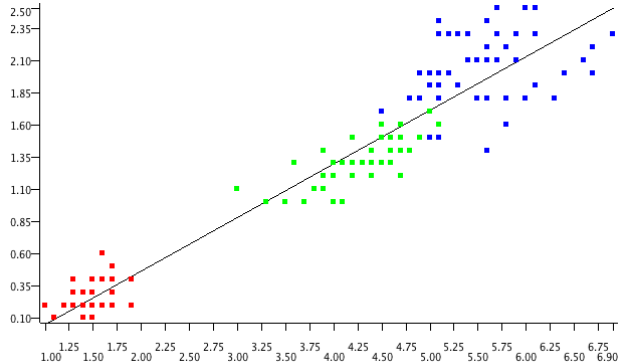
Start with exercise: *Data Mining, Activity I*:

- Partition the fully joined data into a training and test set (50%, Stratified Sampling on Target)
- Train a decision tree on the training set to predict Target
- Use the trained model to predict Target in the test set
- What is the overall accuracy of your model?
- Optional: Evaluate the accuracy and robustness of the model with the ROC Curve node

# Regression

Predict *numeric* outcomes on existing data (supervised)

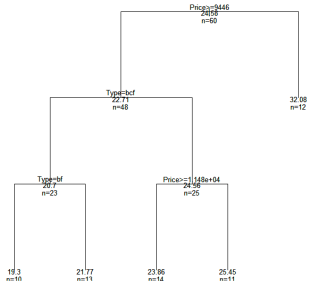
- Applications
  - Forecasting
  - Quantitative Analysis
- Methods
  - Linear
  - Polynomial
  - Regression Trees
  - Partial Least Squares




**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 |

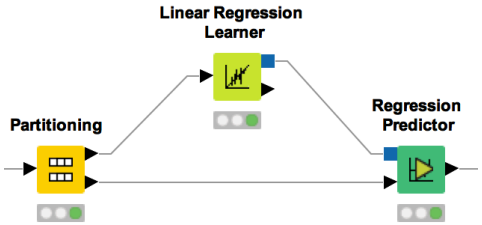
Multiple R-Squared: 0.9271
Adjusted R-Squared: 0.9266

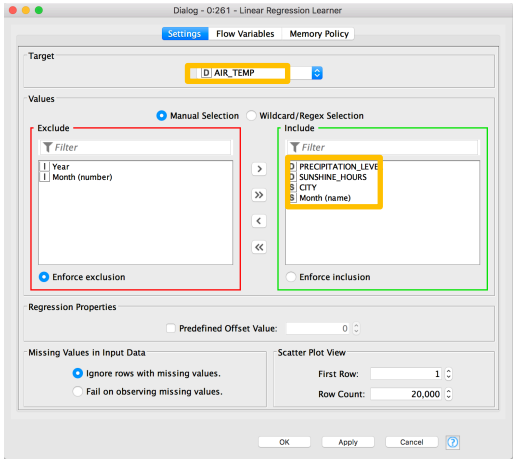


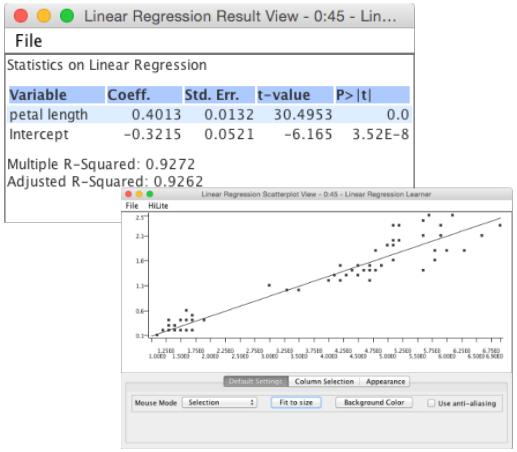
# New Nodes: Linear Regression Learner & Regression Predictor

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



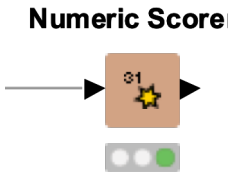



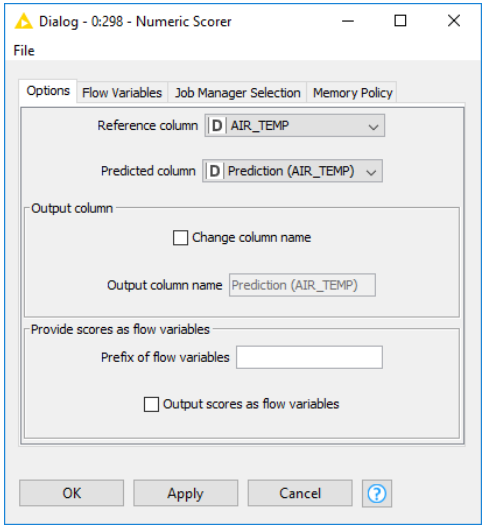



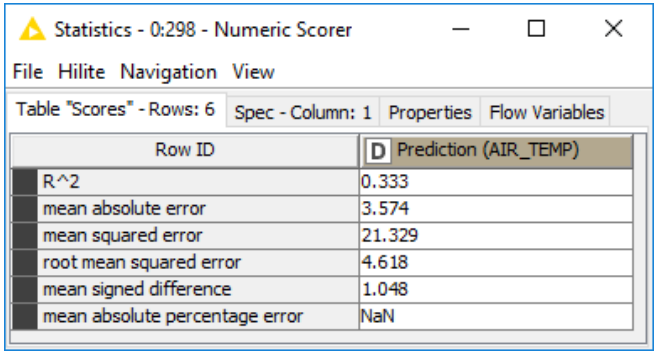
# New Node: Numeric Scorer

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 $R^2$, MAE, MSE, RMSE etc.



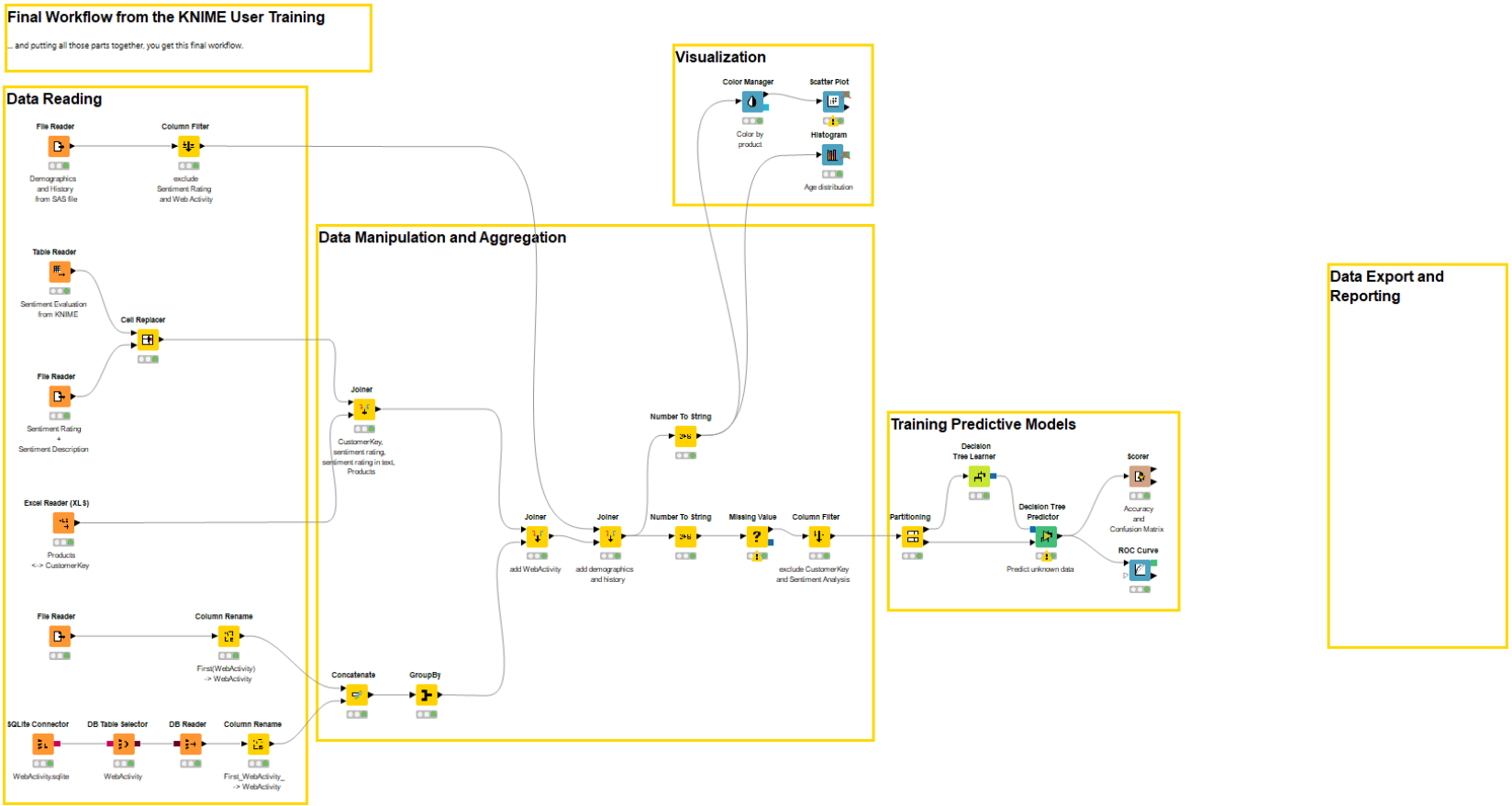

| Row ID | D Prediction (AIR_TEMP) |
|---|---|
| R^2 | 0.333 |
| mean absolute error | 3.574 |
| mean squared error | 21.329 |
| root mean squared error | 4.618 |
| mean signed difference | 1.048 |
| mean absolute percentage error | NaN |

# Data Mining Exercise, Activity II

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

- Read *weather.table* data
- Split the data into rows up to 2016 (training set) and rows from 2017 on (test set)
- Train a linear regression model that predicts the AIR_TEMP as a function of all other features in the dataset
- Use the model to predict the temperature in 2017 and evaluate the model with the Numeric Scorer node
- Optional:
  - Calculate the mean temperature per month in the training data
  - Join the mean temperature per month to the test set
  - Use the Numeric Scorer to see if the average monthly temperature provides a better prediction than the Linear Regression model

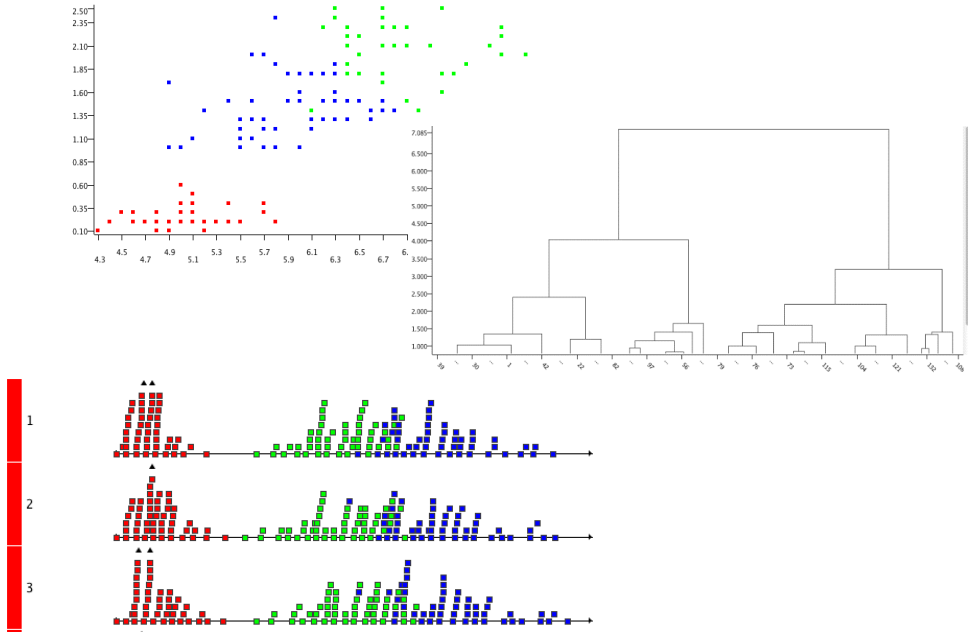
# Today's Example



# Clustering

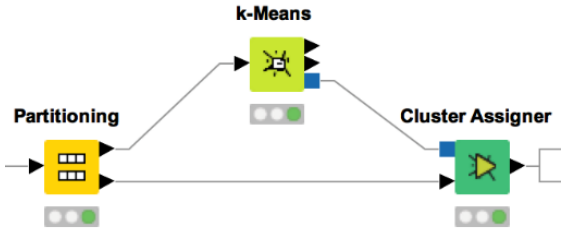
Discover hidden structure in unlabeled data (unsupervised)

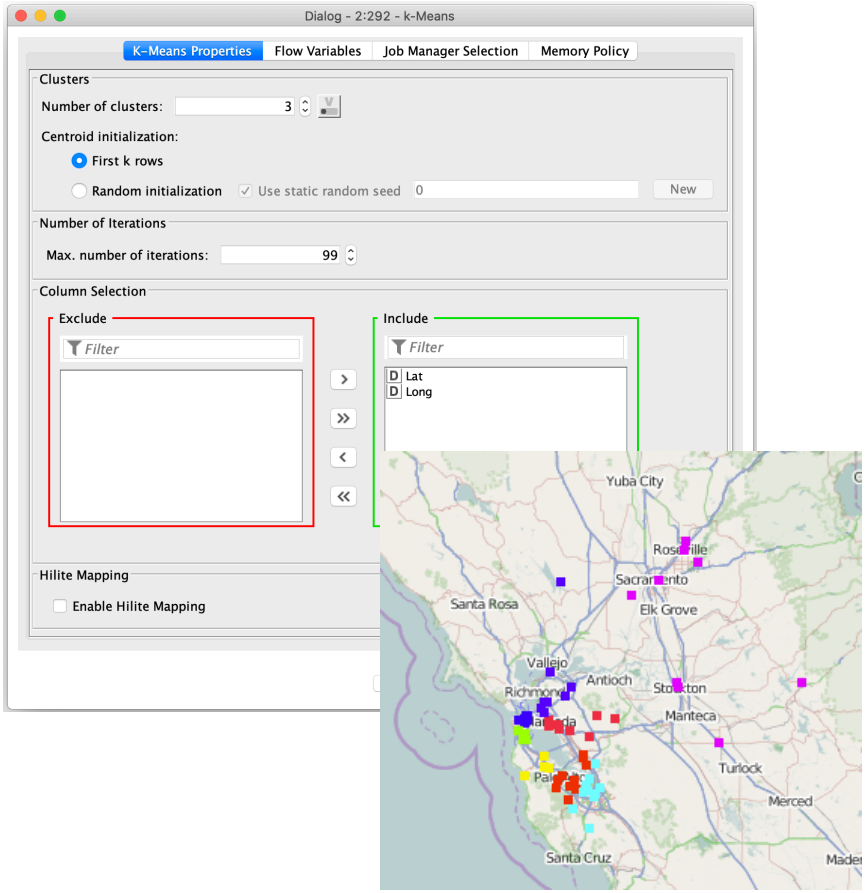
- Applications
  - Market Segmentation
  - Diversity picking
- Methods
  - K-means/medoids
  - Hierarchical
  - DBScan
  - OPTICS
  - Neighbourgrams



# 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.



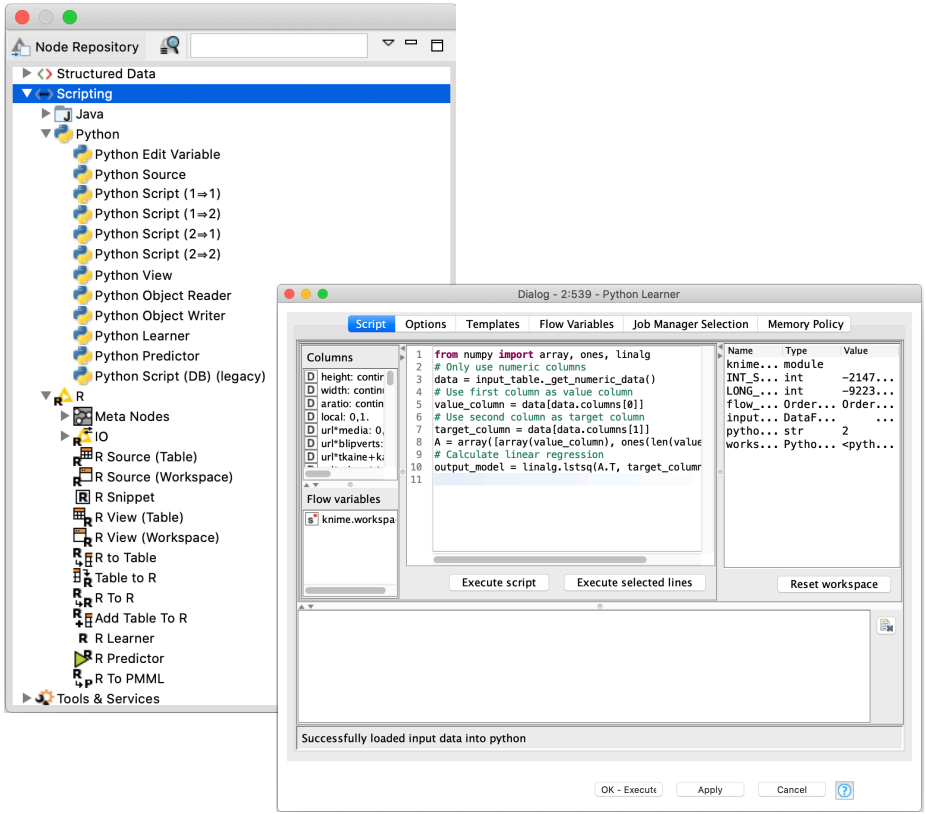
# Data Mining Exercise, Activity III

Start with exercise: *Data Mining, Activity III*

- Read *location_data.table* data
- Filter the data to entries from California (region_code = CA)
- Perform k-means clustering with k=3. Use only latitude and longitude for clustering.
- Optional: plot latitude and longitude in a view (OSM Map or Scatter Plot) and use the view to visually optimize k

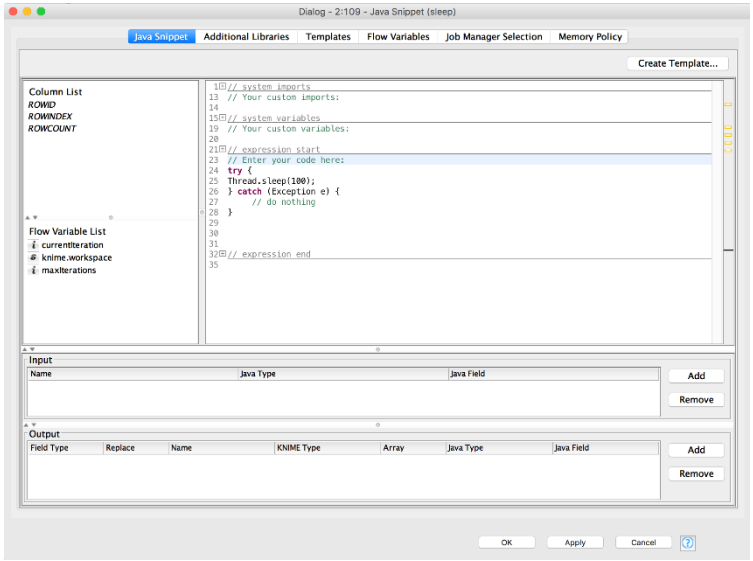
# Scripting Integrations: R and Python

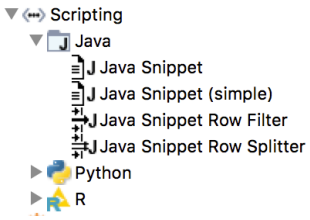
- Run R or Python code in KNIME Analytics Platform
- Works with existing Python and R installations
- Syntax highlighting support
- Different nodes for many tasks, e.g training a model using an algorithm available in Python

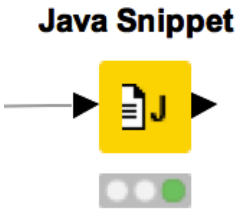


# Java Snippet

- 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



# 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

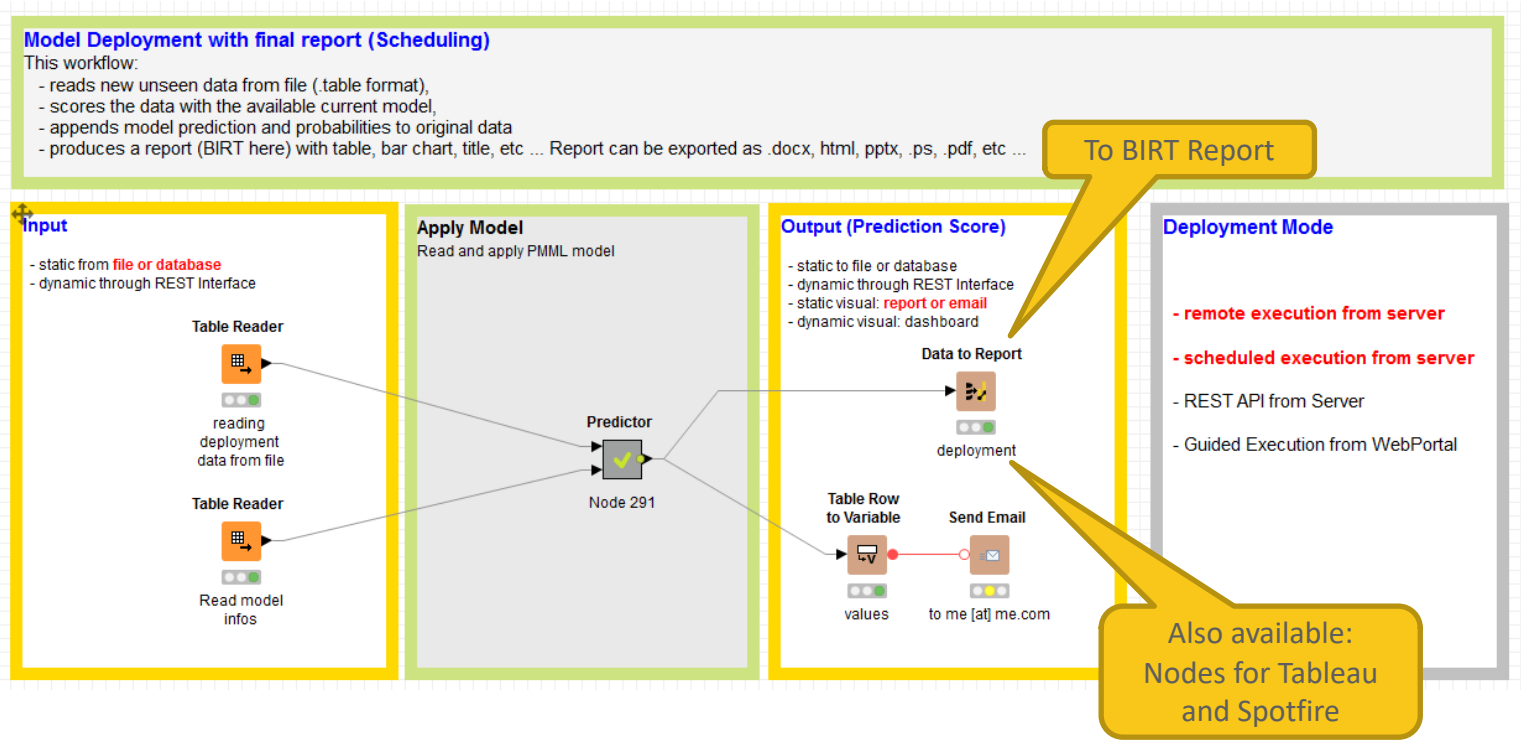
## Input

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

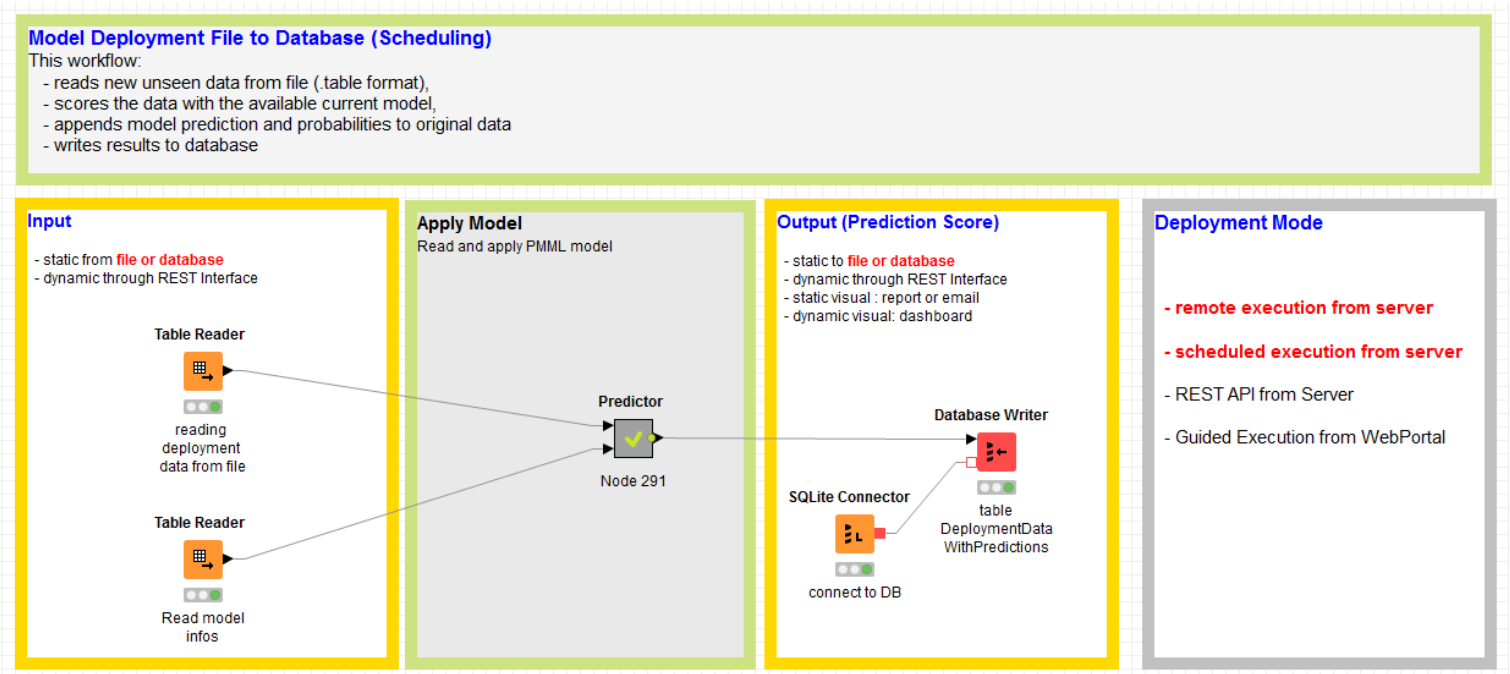
## Output

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

# To Report / Email

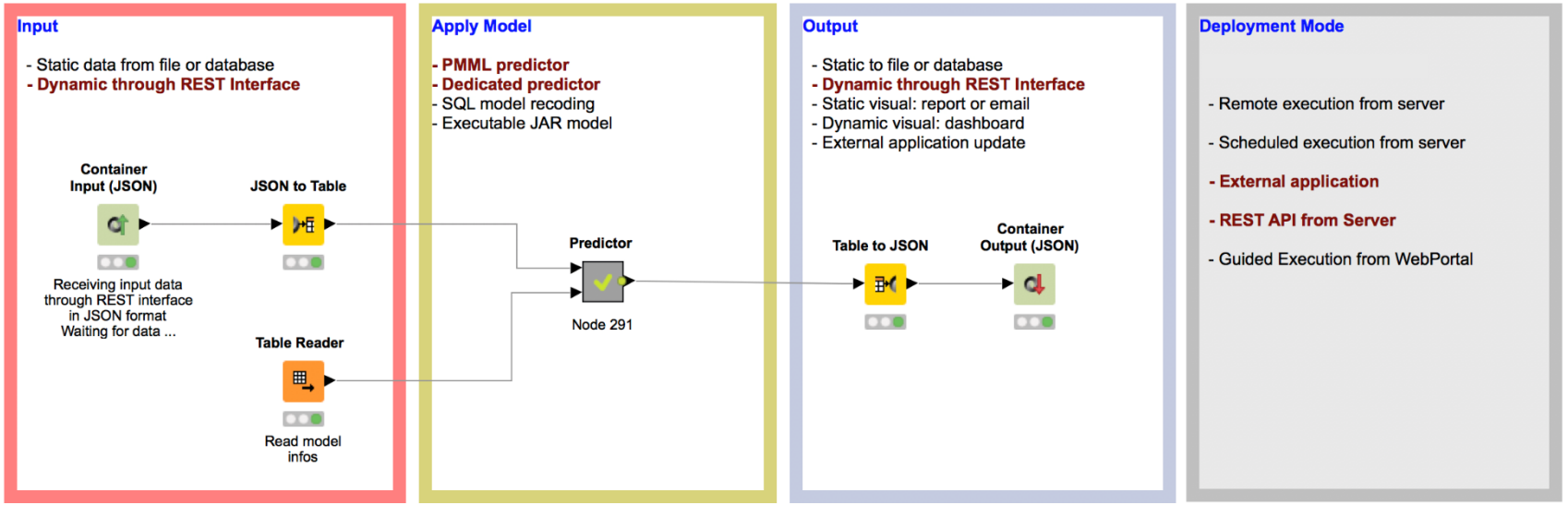


# To File / Database



# REST API (Available on KNIME Server)

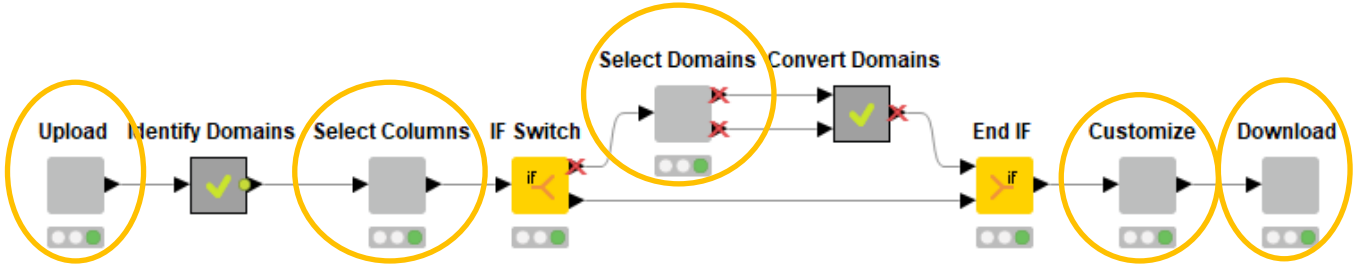
**Model Deployment as REST API**
This workflow:
- receives new unseen data via REST interface (JSON format),
- scores the data with the available current model,
- appends model prediction and probabilities to original data,
- makes results available at the output REST interface.



# To Dashboard on WebPortal

**The Process Step by Step**

1. Upload your data / Select one of the available datasets
2. Select the columns to visualize (maximum 3)
3. Convert the domain of the columns (OPTIONAL)
4. Customize the visualizations interactively
5. Download the images of the customized charts




**Step 1**
Upload File

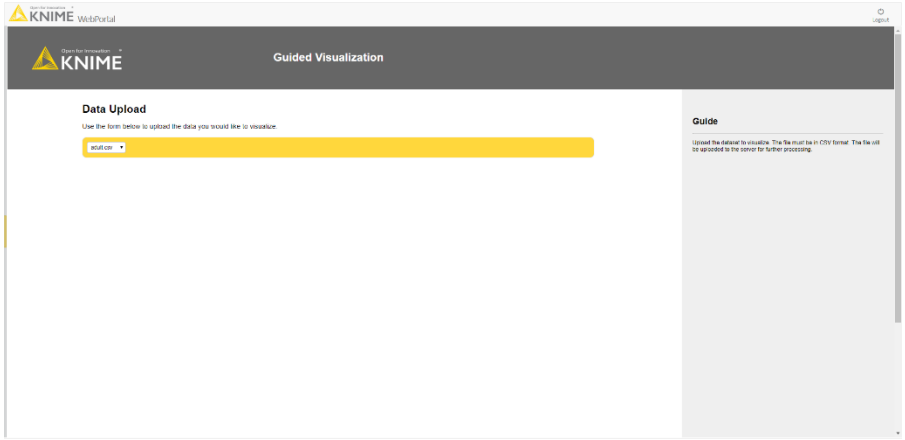
**Step 2**
Select Columns

**Step 3**
Customize
Column Domains

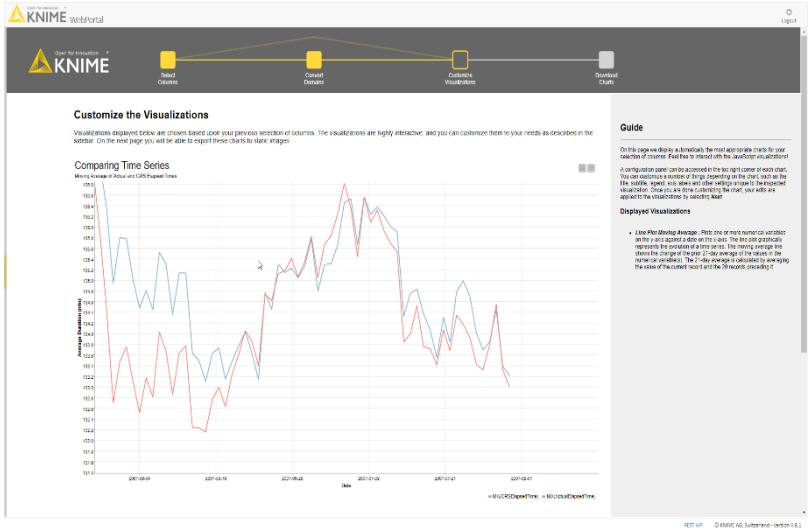
**Step 4**
Interactive View

**Step 5**
Download Image

# Workflow on KNIME WebPortal



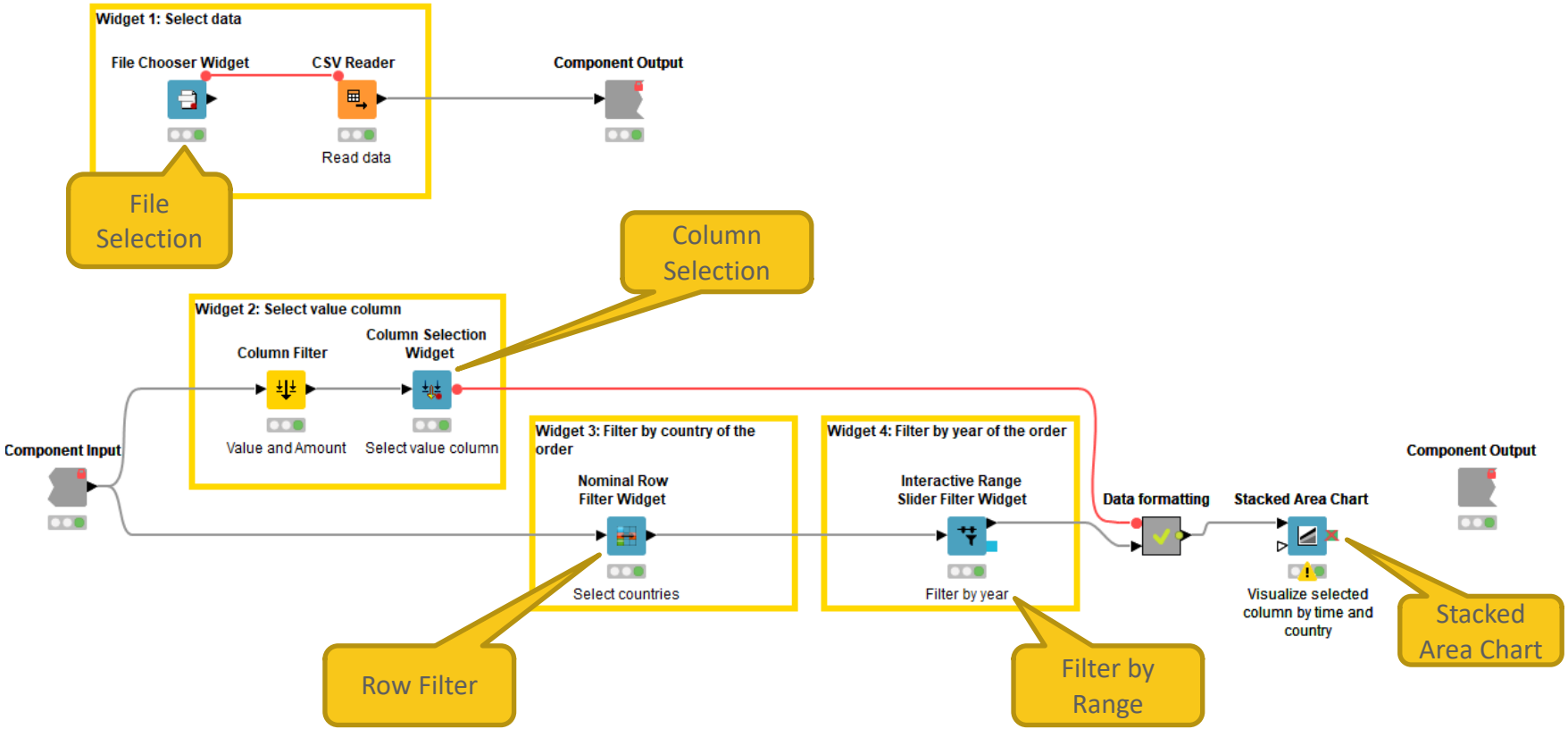

**WebPortal Page (Step 1)**
Upload File




**WebPortal Page (Step 4)**
Interactive View

**Available in KNIME Server**

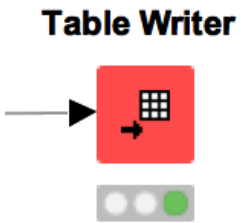
# Components to Produce Dashboard on Web Page

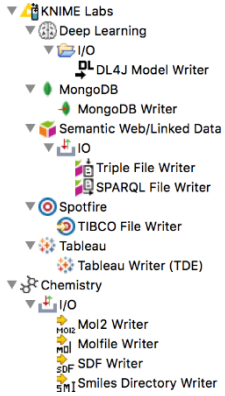


# Data Export Nodes

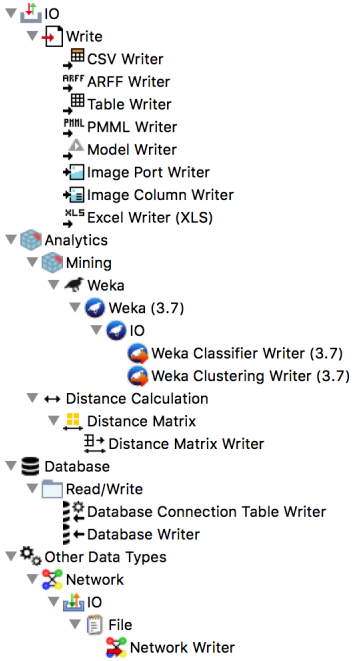
Typically characterized by:

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

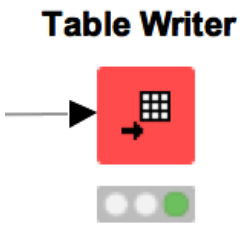


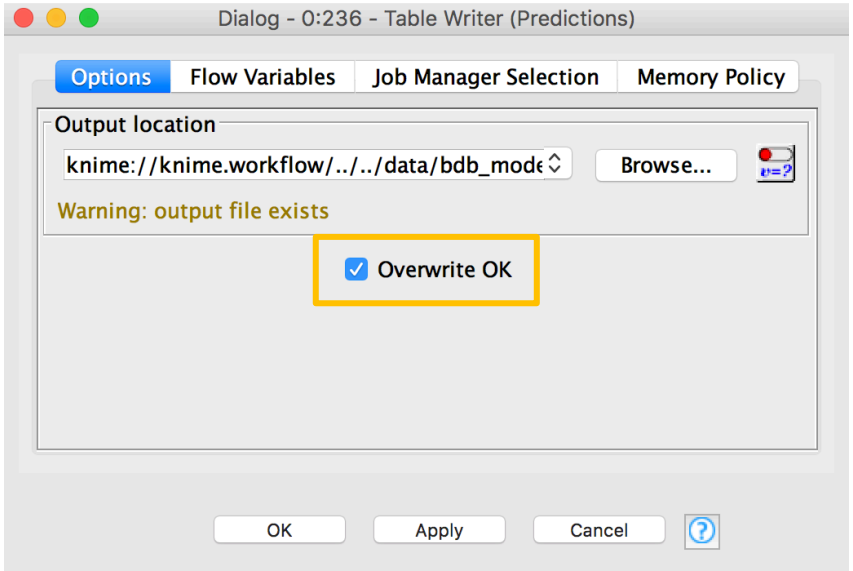





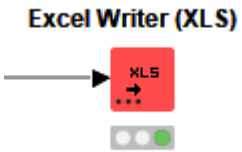


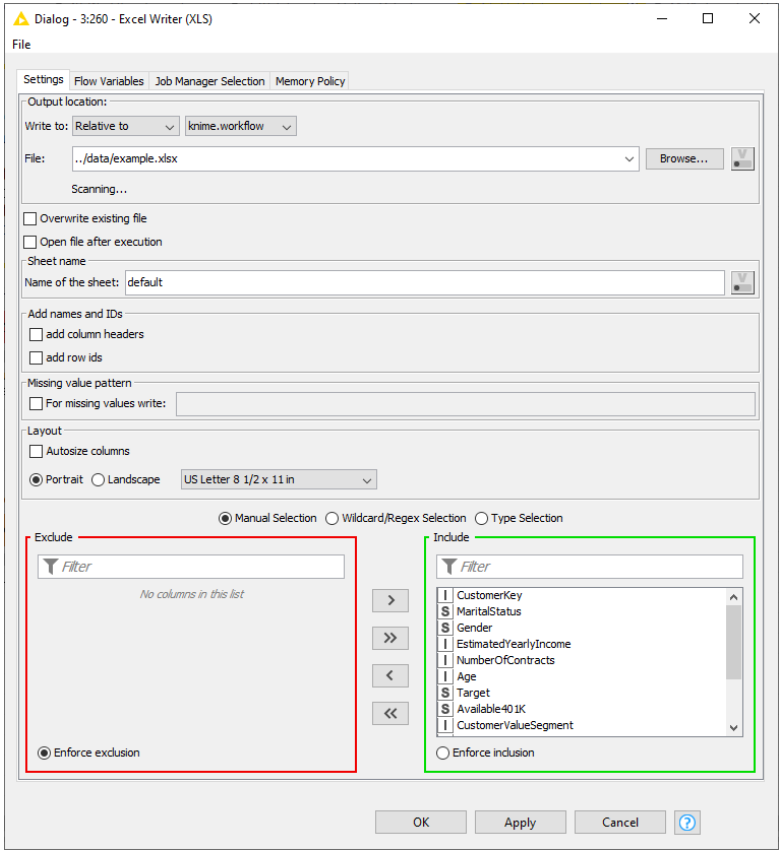
# New Node: Table Writer



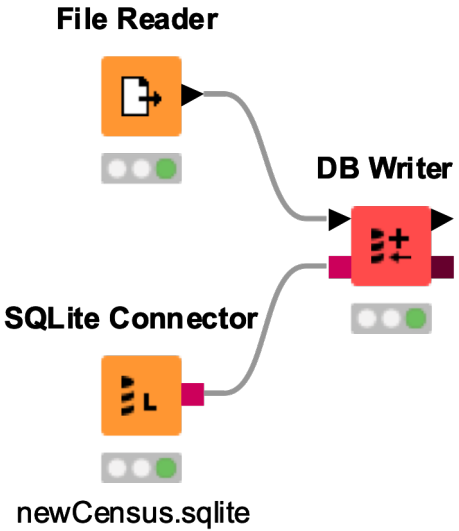



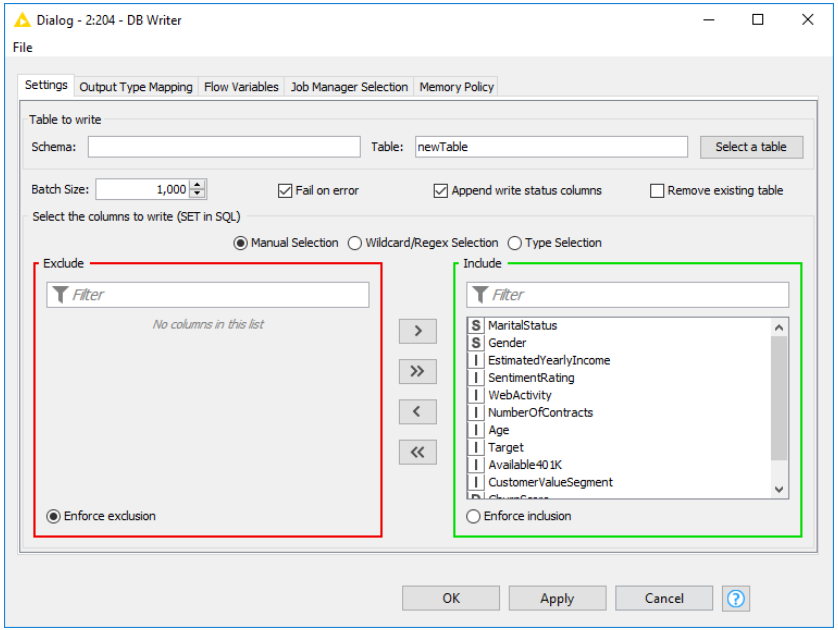
# New Node: XLS Writer



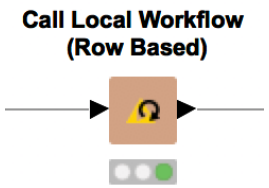
# New Node: Database Writer

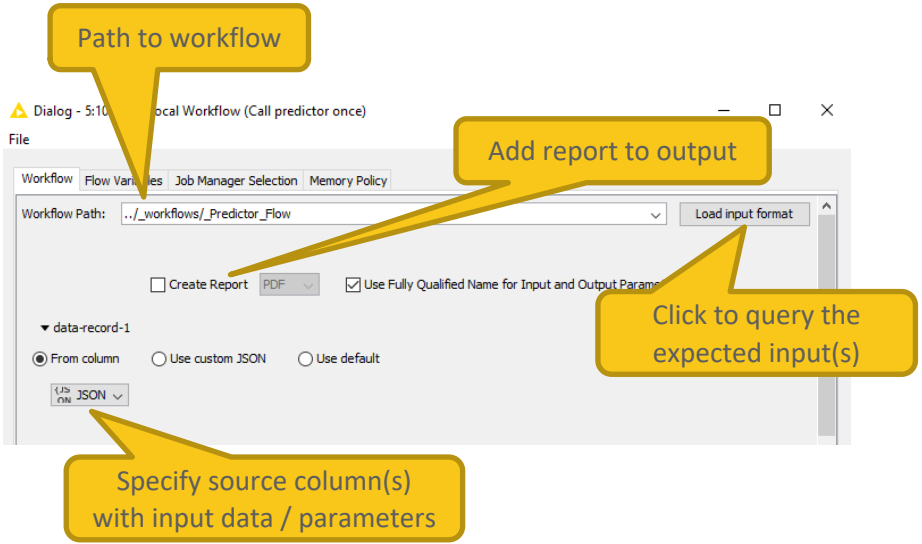


# Automation: Call Local Workflow

- 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

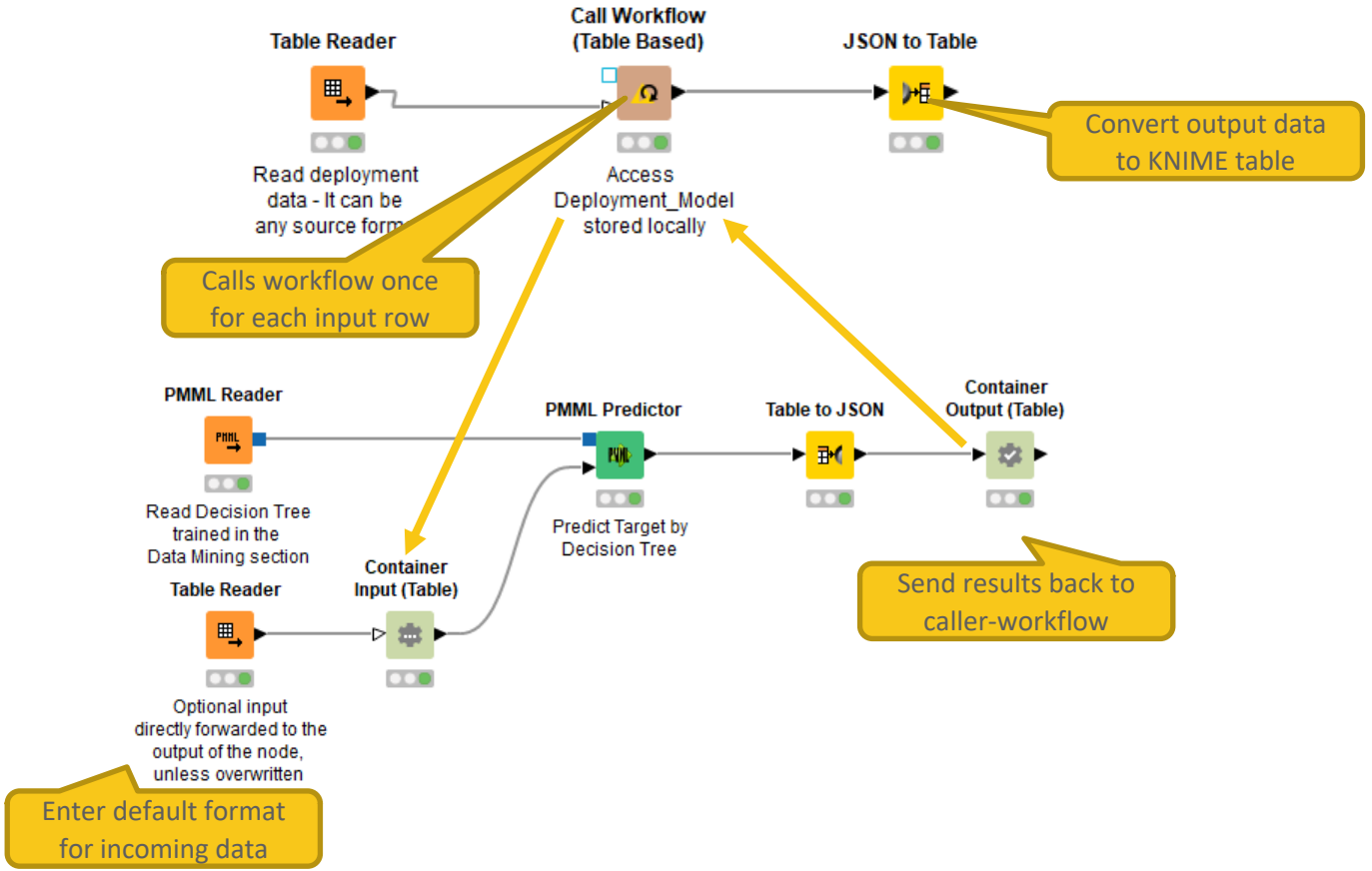


# Automation: Call Local Workflow

ETL

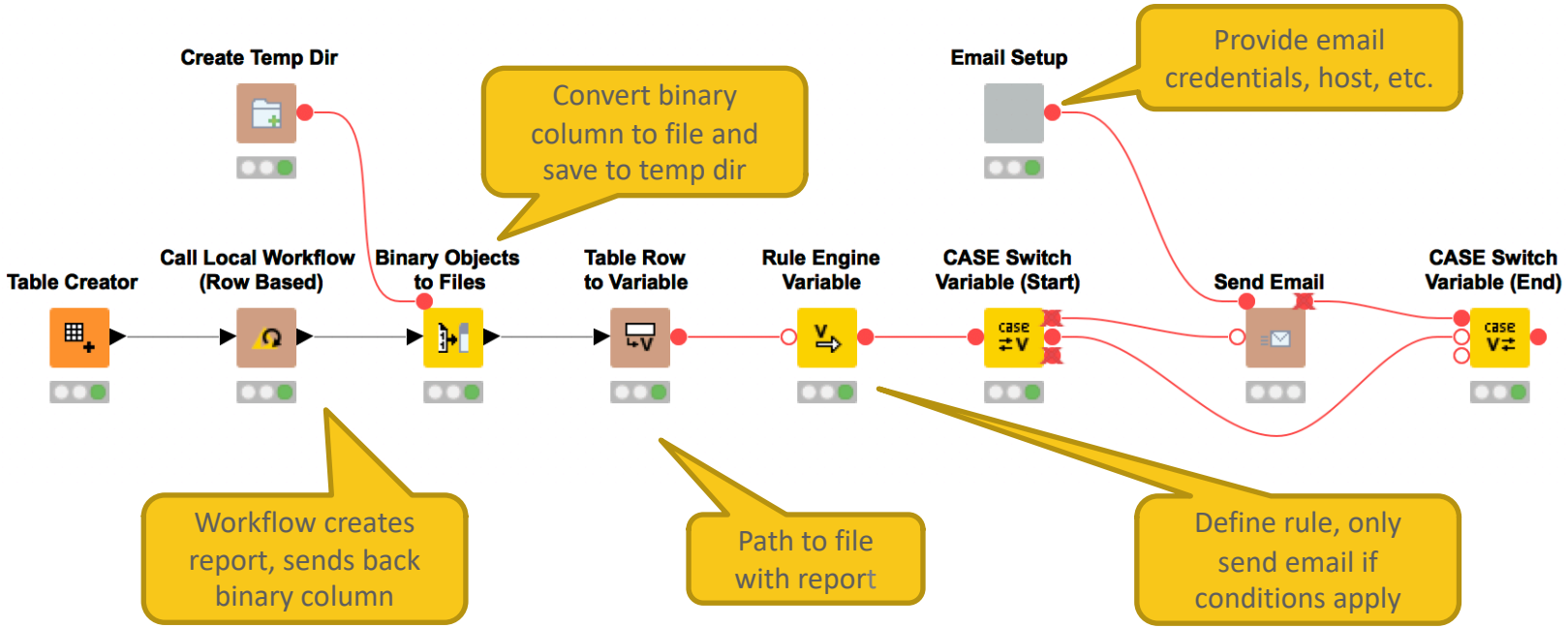
Prediction



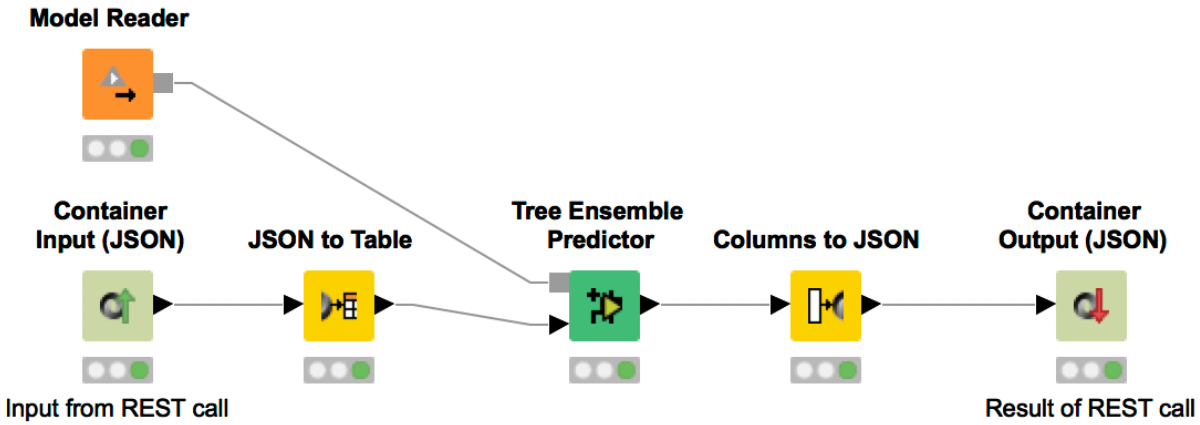
# Use Call Local Workflow to Send Conditional Emails with Report

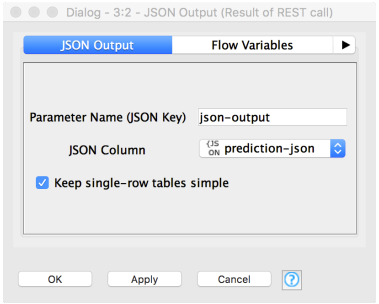
Sometimes, report should be sent under specific circumstances

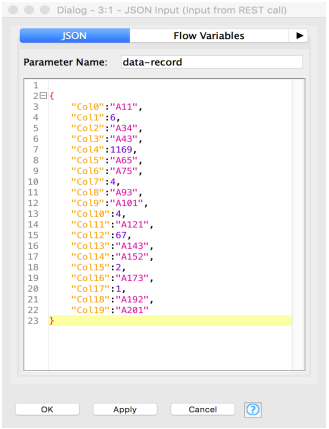
- E.g. if some KPI is below threshold



# KNIME Server as a REST Resource



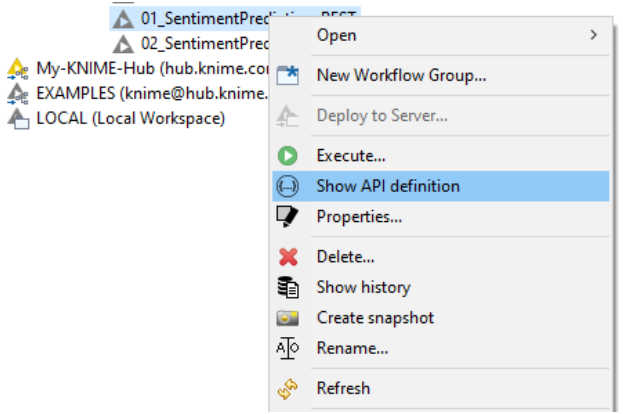



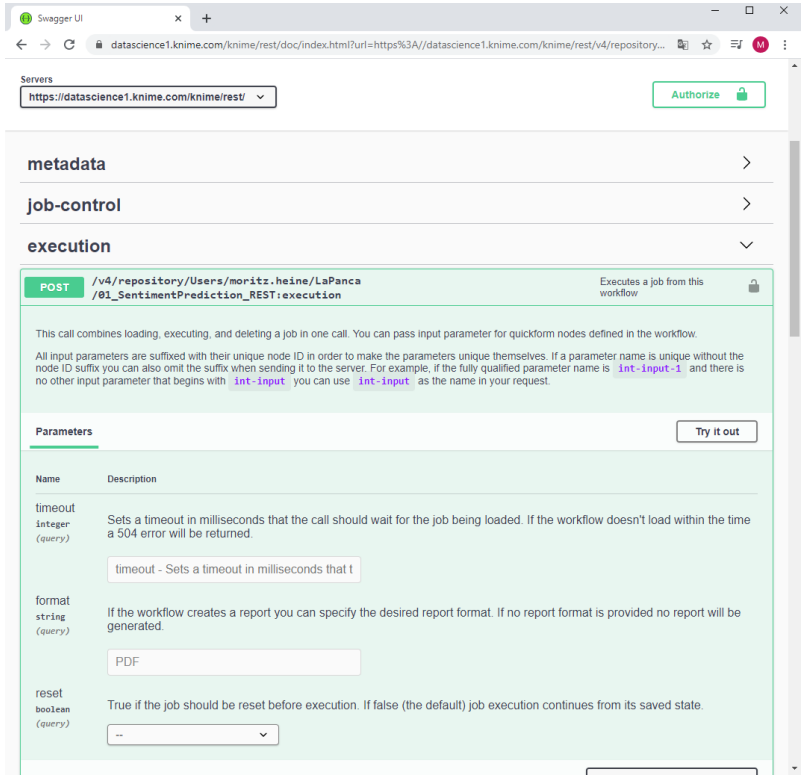




https://www.knime.org/blog/giving-the-knime-server-a-rest

# KNIME Server as a REST resource

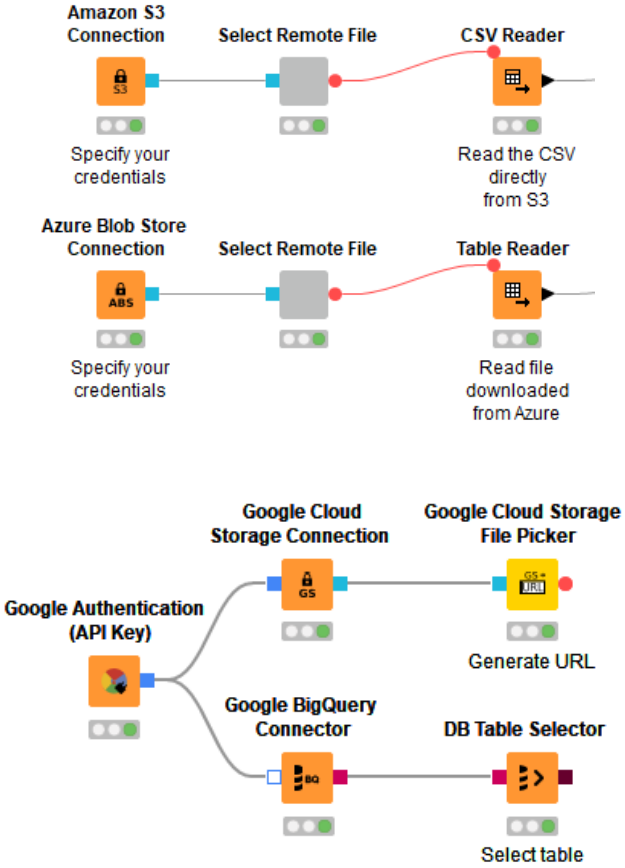
- Use Swagger, SOAPUI or Chrome extension Postman to explore the HTTP requests and test them



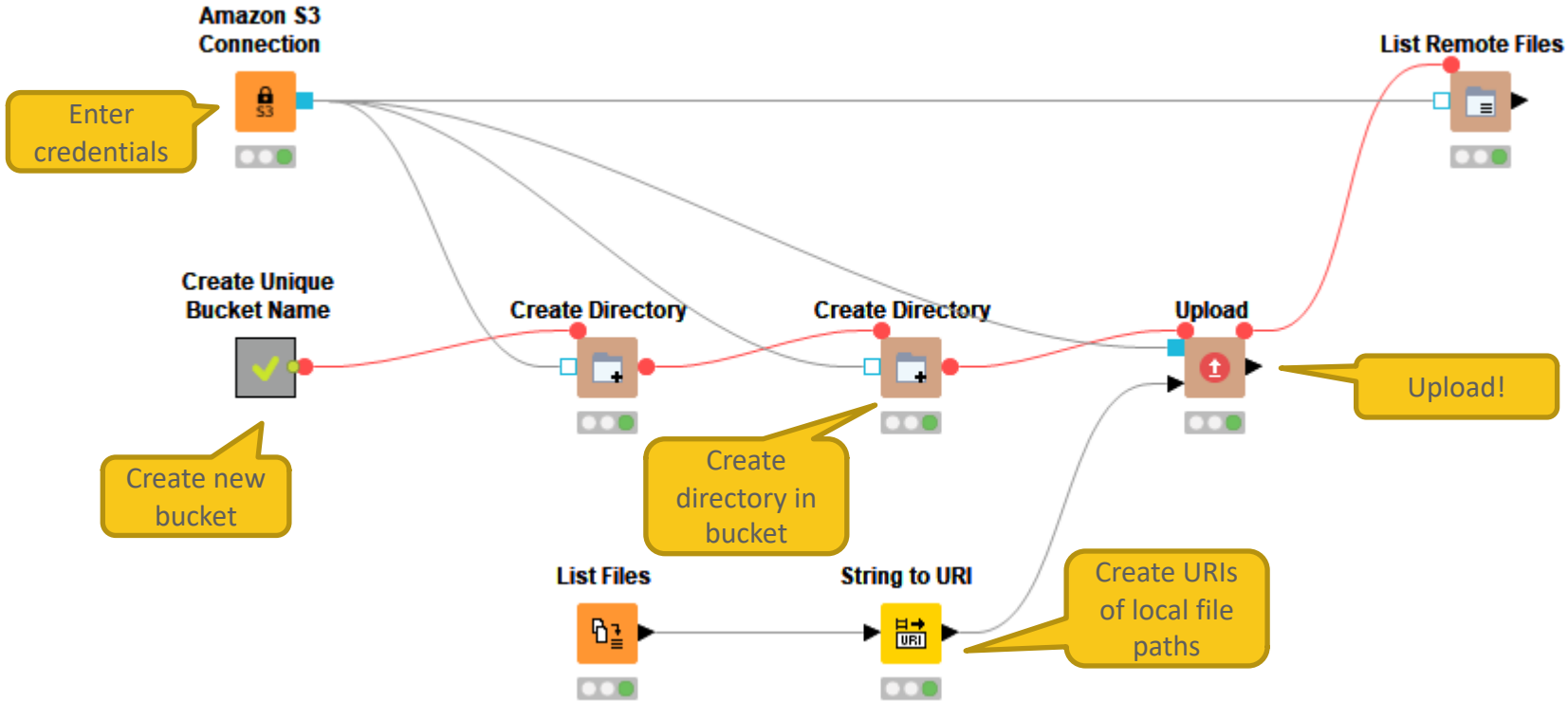
# Remote File Handling – Cloud Storage

- Integrate remote data sources from Amazon AWS, Microsoft Azure, and Google Cloud
  - 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

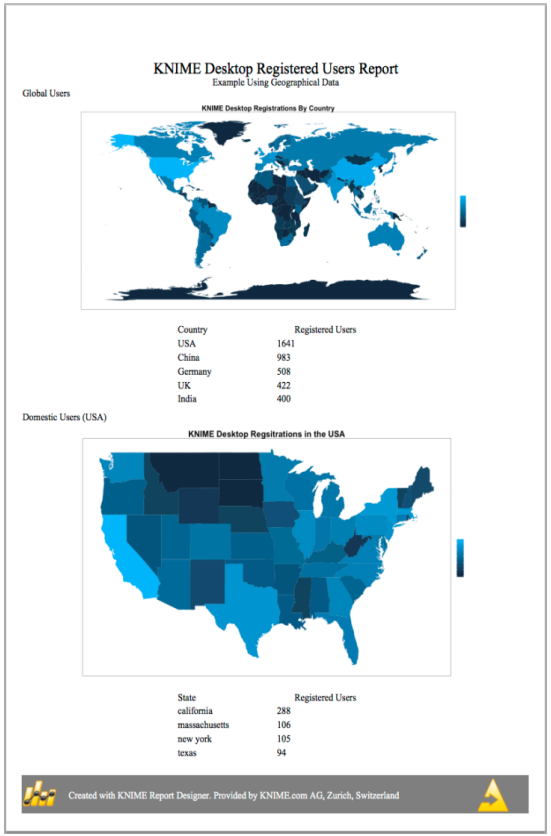
Example: Upload all files from a local directory to Amazon S3



# Reporting in KNIME

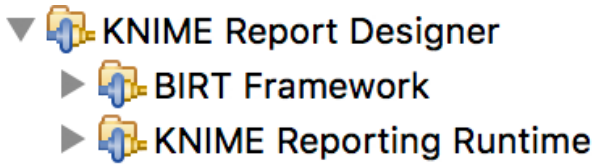
# Reporting in KNIME

- 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, …)



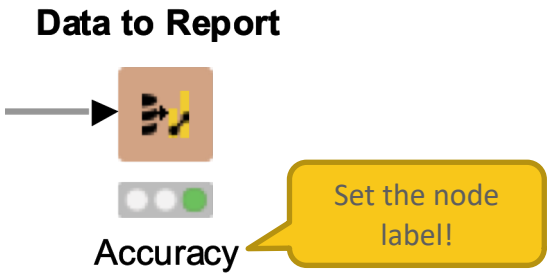
# Installation

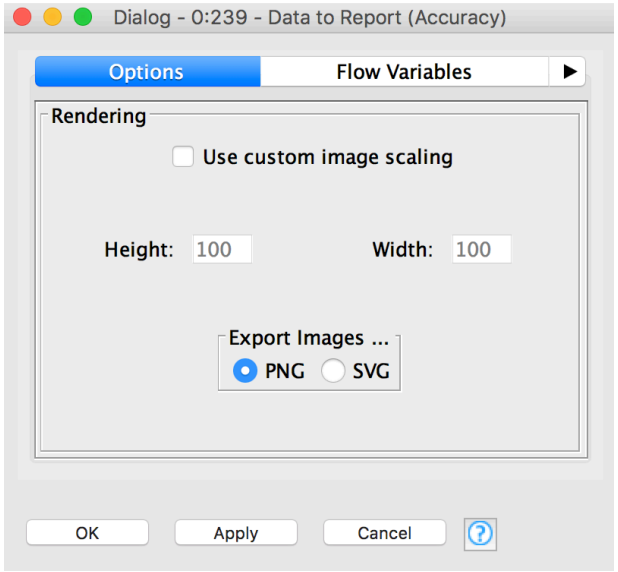
- Can be installed via KNIME -> Install KNIME Extension
- Install the KNIME Report Designer



# New Node: Data to Report

Send a data table to BIRT



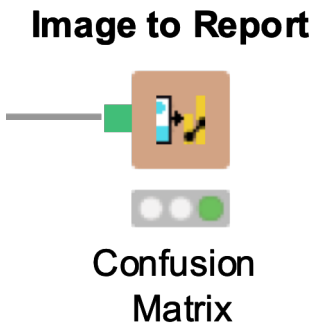



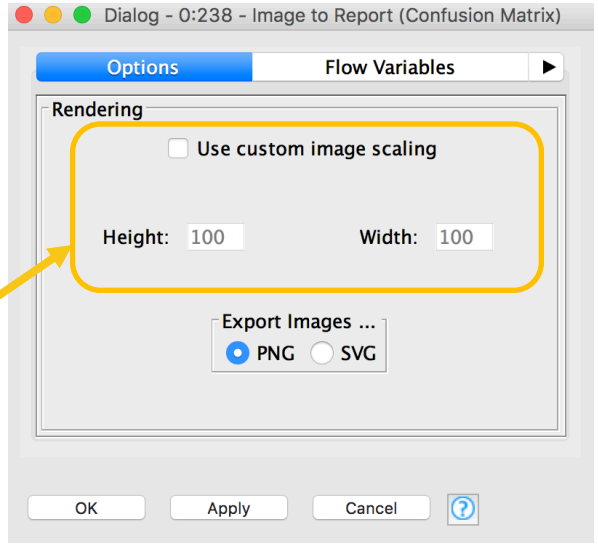

Hint: The node label will be used to identify the data source in the reporting view -> Make sure to use understandable labels if you have more than one data source

# New Node: Image to Report

Send an image to BIRT

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

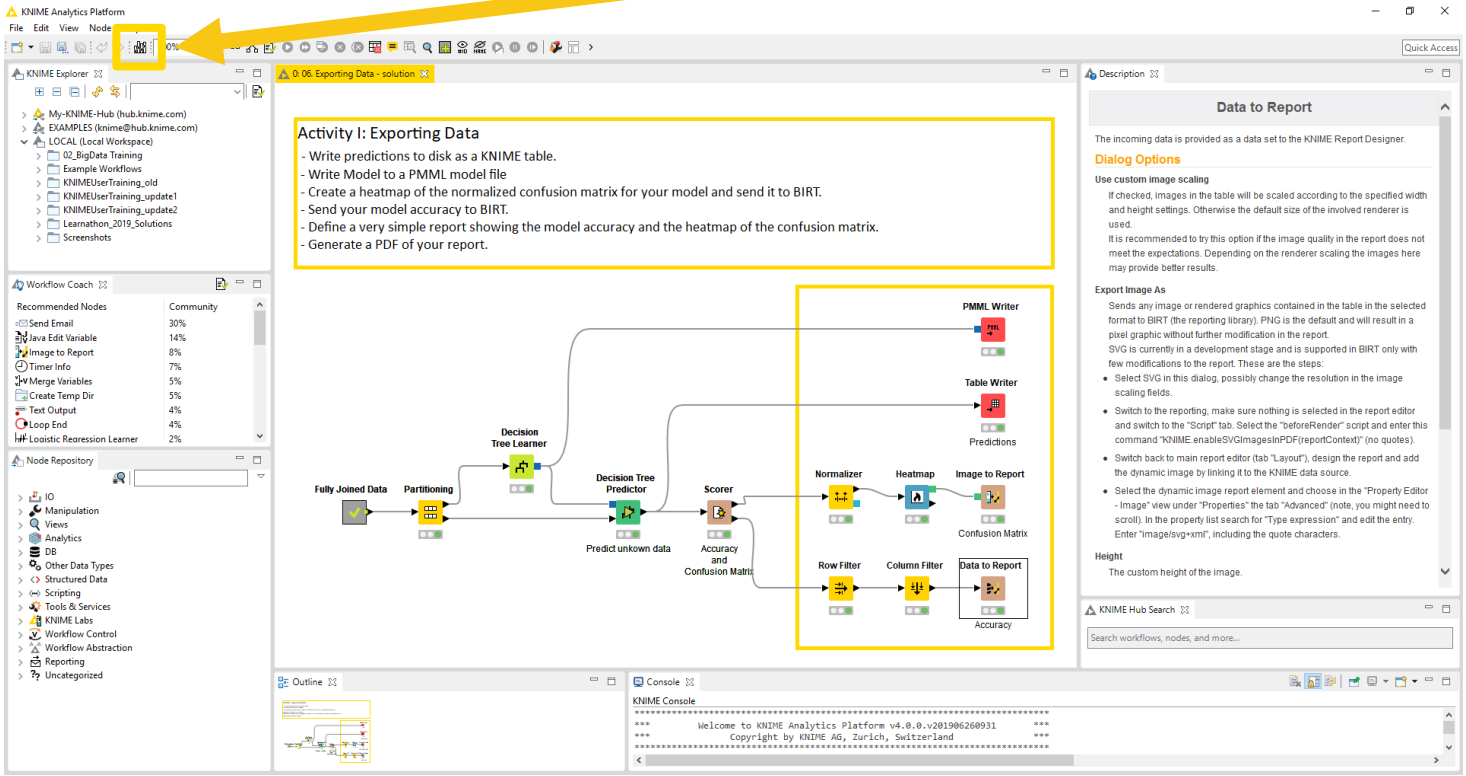


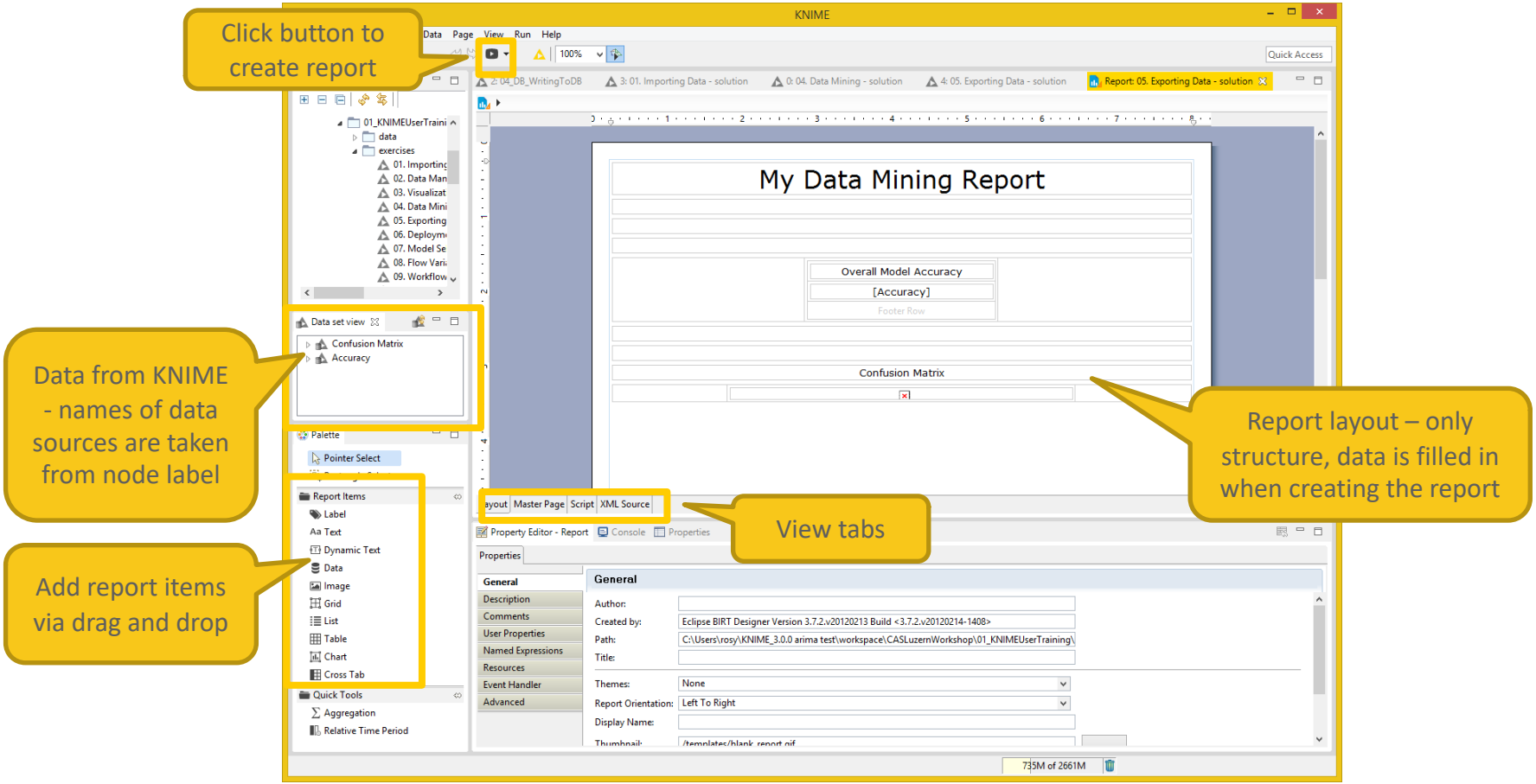

Hint: Customize the image size in the Data to Report node to fit the report

# Edit the Report

Open the workflow and click the Report Editor button in the tool bar

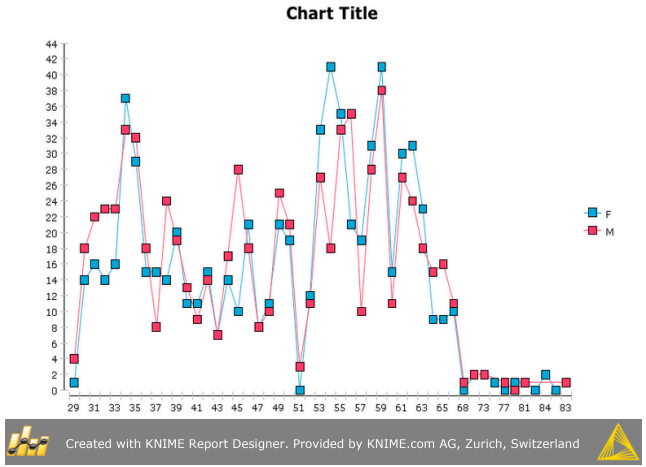


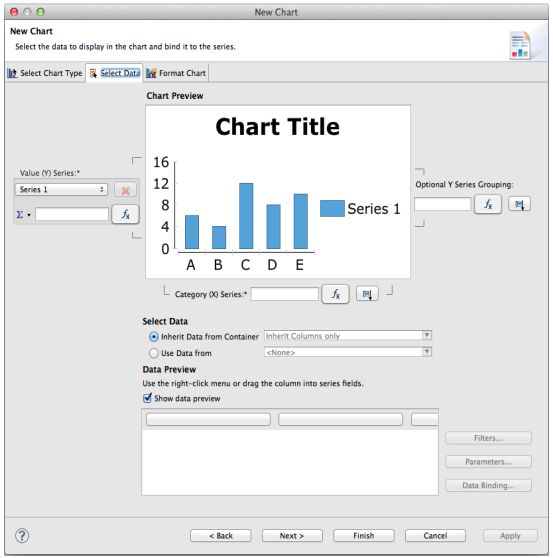
# Reporting Perspective



# Charting in BIRT

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



# Tips & Tricks

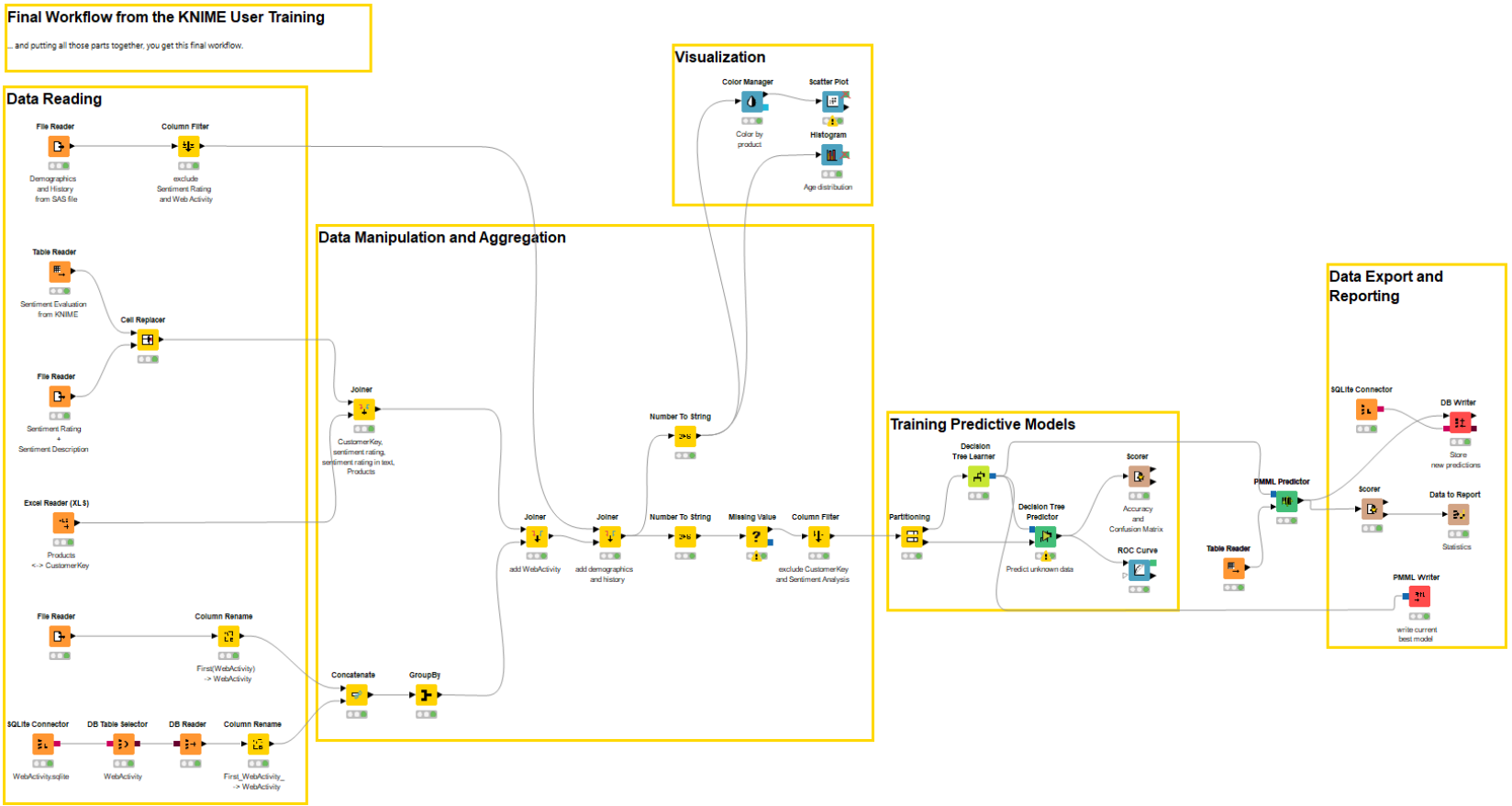
- Use an underlying grid to structure the report
- Names of columns should not change
- Use the grouping function to combine results
- Use the Master Layout Tab (For footers etc.)

# Exporting Data Exercise

Start with exercise: *Exporting Data*

- Write the predictions to a KNIME table
- Write the decision tree model to a PMML model file
- Create a heatmap of the normalized confusion matrix of your model and send it to a BIRT report
- Send your model accuracy to a BIRT report
- Create a simple report showing the overall accuracy and the heatmap of the confusion matrix
- Generate a PDF of your report

# Today's Example




# Thank You!

education@knime.com