

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

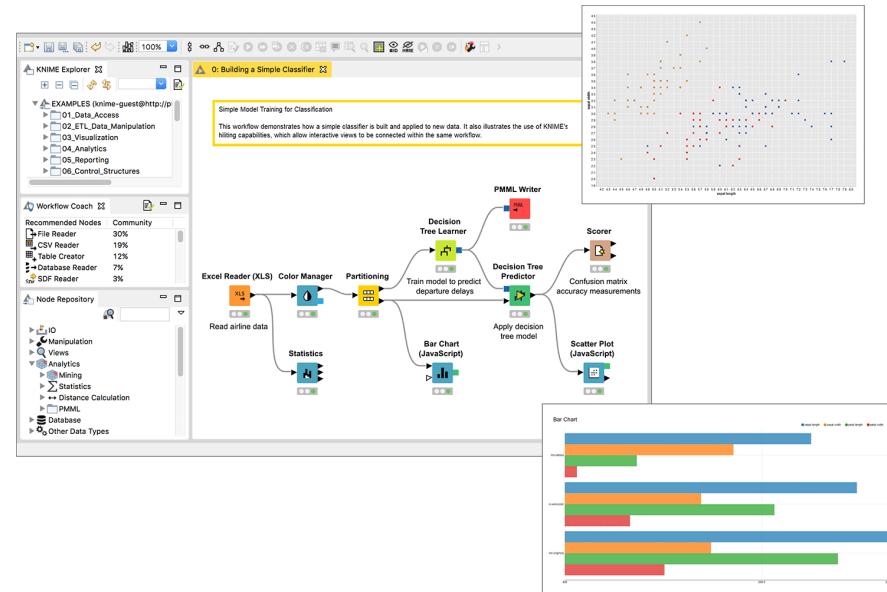
KNIME AG

# **Overview**

## **KNIME Analytics Platform**

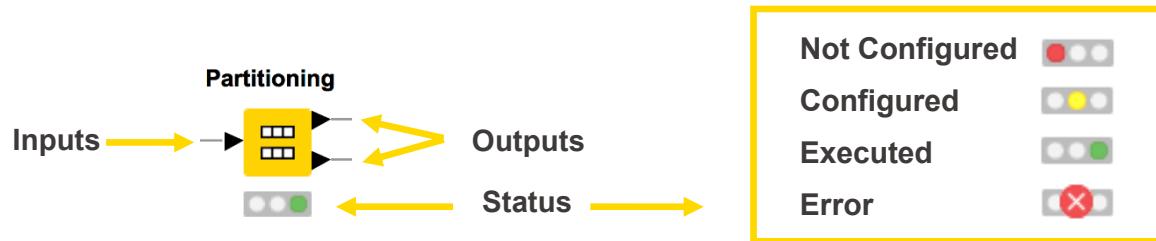
# What is KNIME Analytics Platform?

- 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.

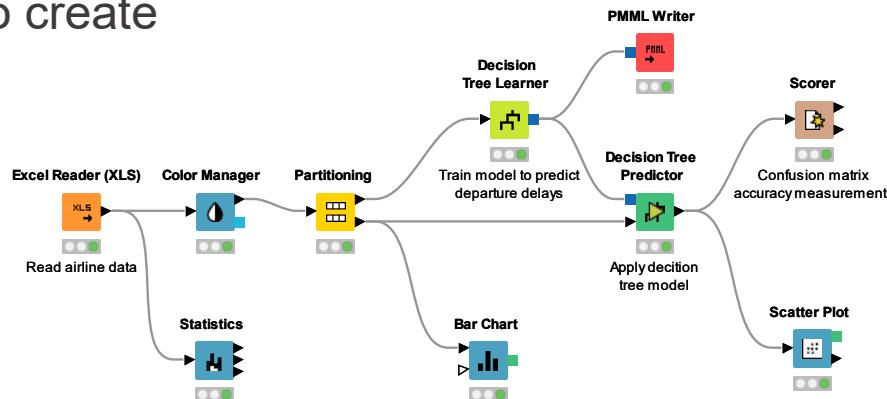


# Visual KNIME Workflows

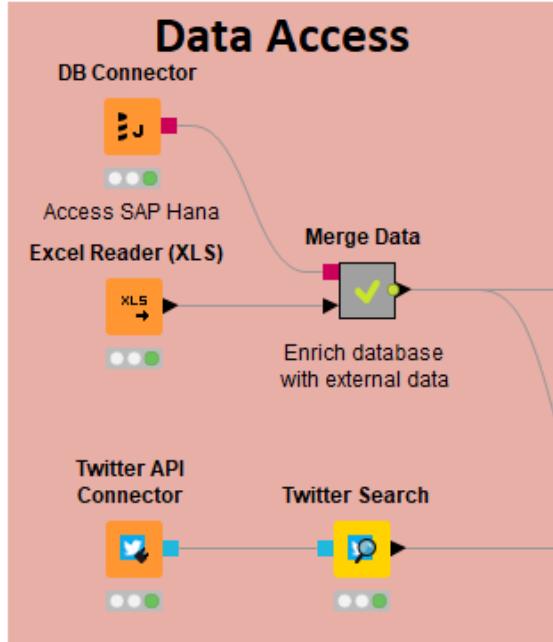
**NODES** perform tasks on data



Nodes are combined to create  
**WORKFLOWS**

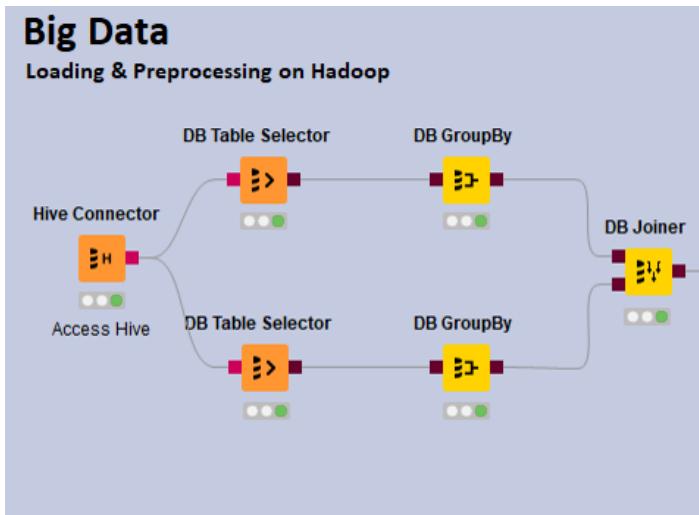


# Data Access



- Databases
  - MySQL, PostgreSQL, Oracle
  - Theobald
  - any JDBC (DB2, MS SQL Server)
  - Amazon DynamoDB
- Files
  - CSV, txt, Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON, PMML
  - Images, texts, networks
- Other
  - Twitter, Google
  - Amazon S3, Azure Blob Store
  - Sharepoint, Salesforce
  - Kafka
  - REST, Web services

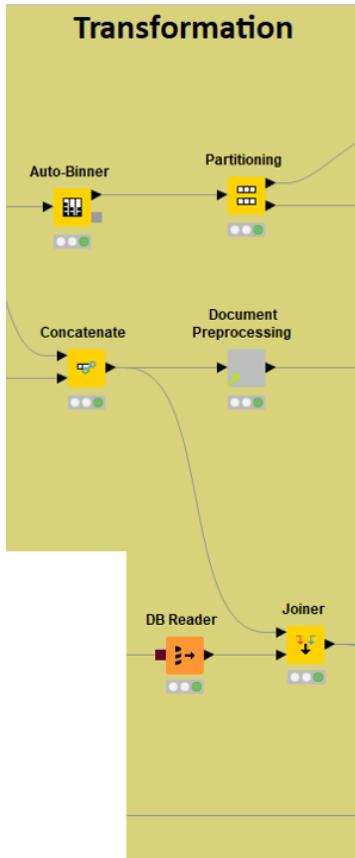
# Big Data



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

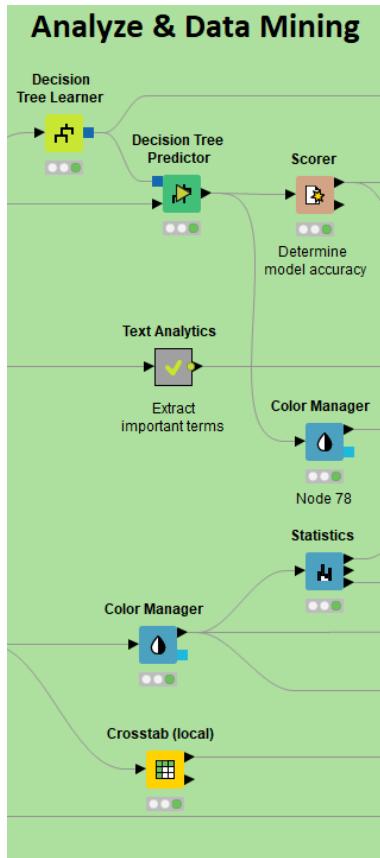


# Transformation



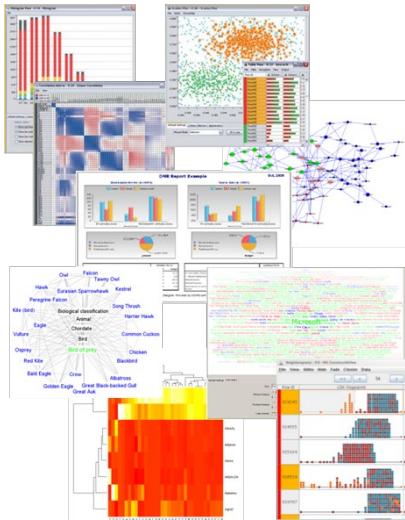
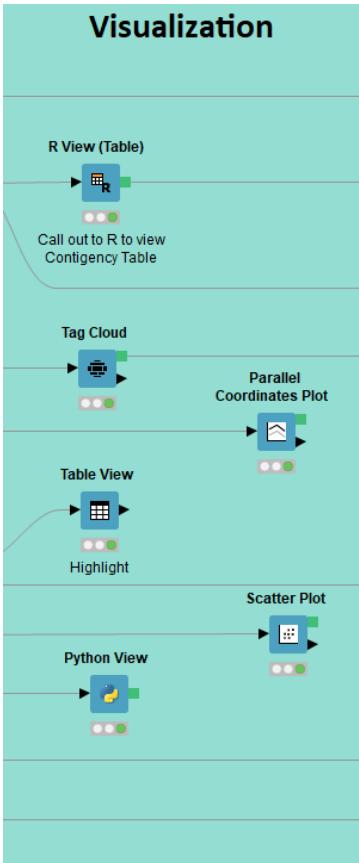
- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature Creation and Selection

# Analysis & Data Mining



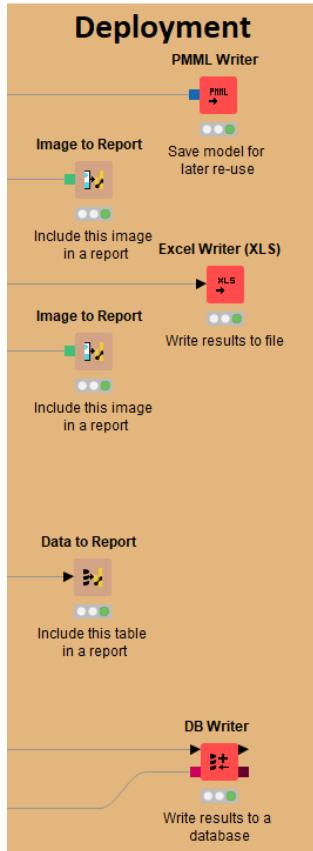
- Regression
  - Linear, logistic
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Deep Learning
  - Keras, DL4J
- External
  - R, Python, Weka, H2O, Keras

# Visualization



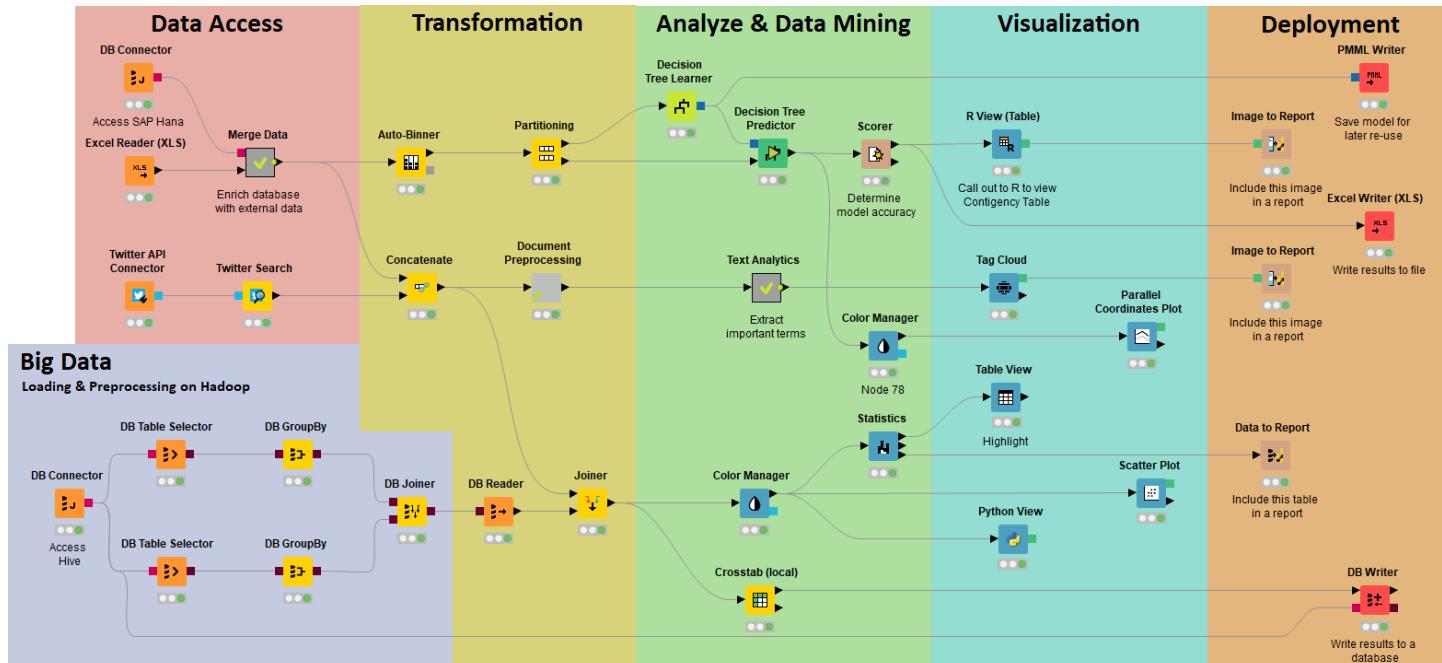
- Interactive Visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

# Deployment



- Database
- Files
  - Excel, CSV, txt
  - XML
  - PMML
  - to: local, KNIME Server, Amazon S3, Azure Blob Store
- 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

# 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

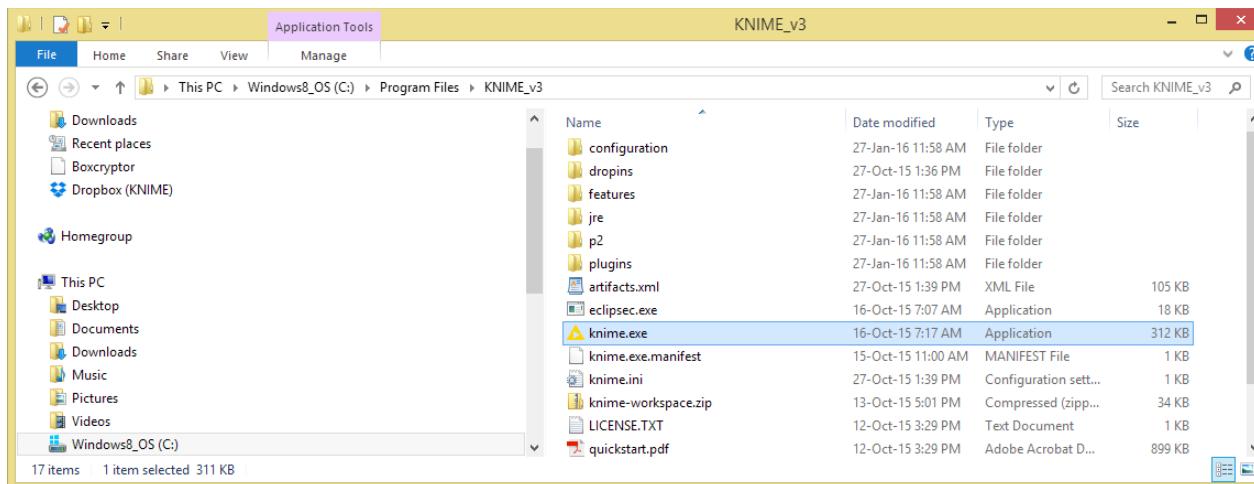
Windows
KNIME Analytics Platform for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>
32 Bit (393.38 MB) 64 Bit (396.38 MB)
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding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

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

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

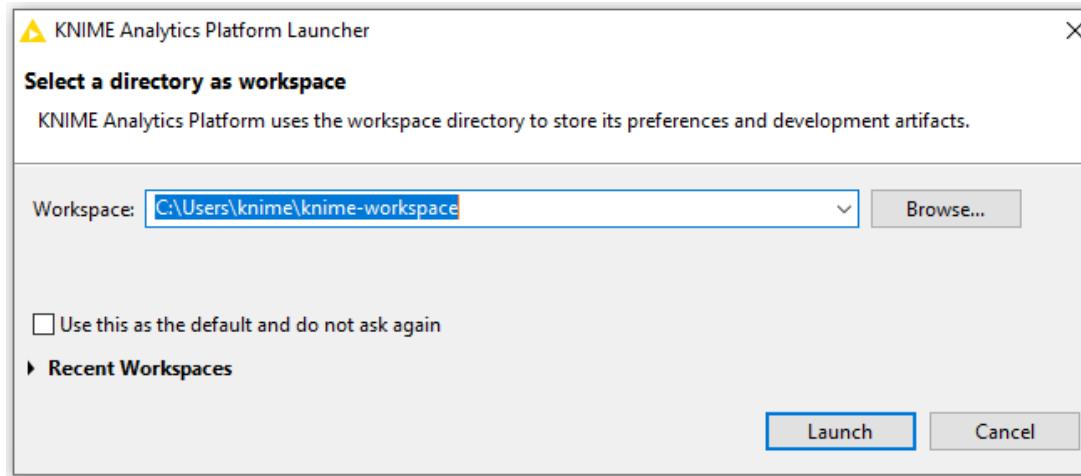
# Start KNIME Analytics Platform

- Use the shortcut created by the installer
- Or go to the installation directory and launch KNIME via the knime.exe



# The KNIME Workspace

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

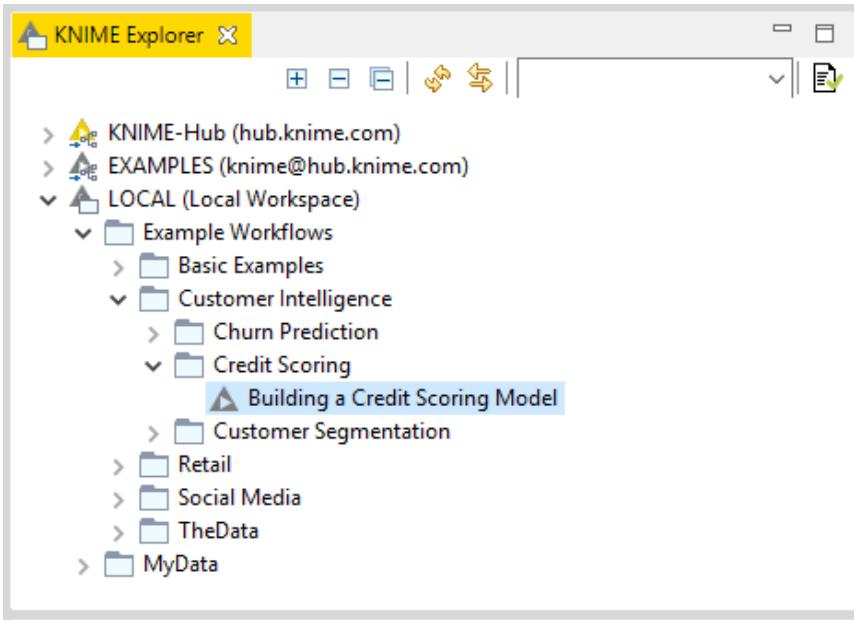


# The KNIME Analytics Platform Workbench

The screenshot shows the KNIME Analytics Platform Workbench interface with several panels highlighted by yellow boxes:

- KNIME Explorer**: Shows the project structure with sections like My-KNIME-Hub, EXAMPLES, LOCAL (Local Workspace), and Example Workflows. A specific workflow named "KNIME Explorer" is selected.
- Workflow Coach**: Displays recommended nodes categorized by community usage.
- Node Repository**: Lists categories such as IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, and Community Nodes.
- Workflow Editor**: Shows a workflow titled "My first Workflow" with four nodes: File Reader (read adult.csv), Row Filter (keep only records born in the US), Column Filter (remove gender), and Table Writer (Write table).
- Outline**: Shows a hierarchical outline of the workflow steps.
- Console & Node Monitor**: Displays the state of the "Row Filter" node, which is EXECUTED. It shows Port Output Port 0 and a preview of the data table.
- Description**: Provides a detailed description of the Row Filter node, including its purpose and configuration steps.
- KNIME Hub Search**: A search bar for finding workflows, nodes, and more.

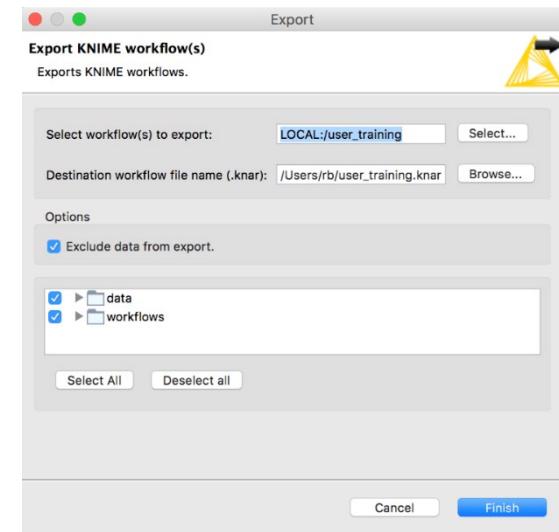
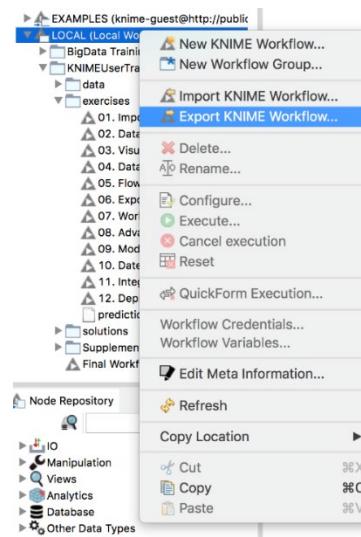
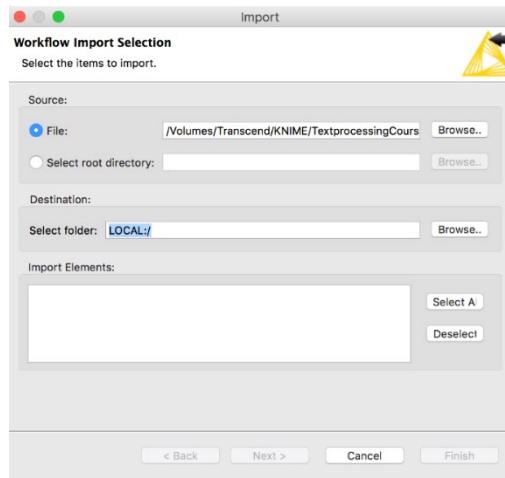
# KNIME Explorer



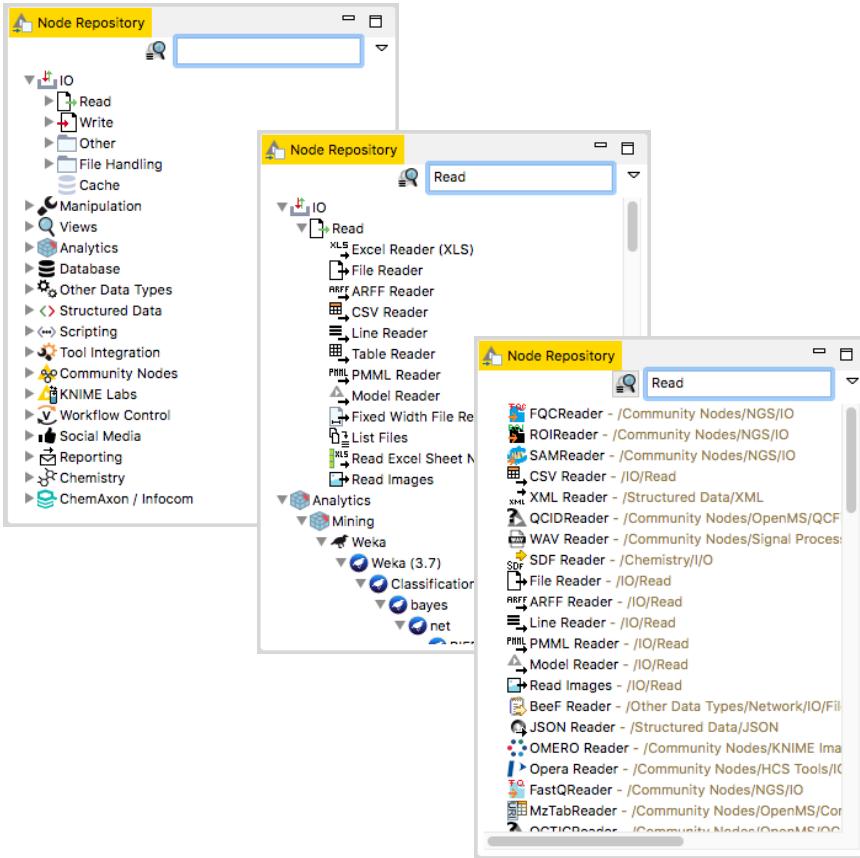
- In LOCAL you can access your own workflow projects.
- The Explorer toolbar on the top has a search box and buttons to
  - ➡ select the workflow displayed in the active editor
  - ⟳ refresh the view
- The KNIME Explorer can contain 4 types of content:
  - Workflows
  - Workflow groups
  - Data files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

# Creating New Workflows, Importing and Exporting

- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the 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

# Description

The screenshot shows the 'Row Filter' configuration dialog. At the top, there's a title bar with the node name 'Row Filter'. Below the title, a descriptive text block explains the node's functionality: it allows for row filtering according to certain criteria, such as ranges by row number, specific row IDs, or values in a column. It notes that the node doesn't change the domain of the data table. A section titled 'Dialog Options' follows, with a heading 'In- or exclude rows by criteria'. The text here instructs the user to select filtering criteria and adjust parameters. Another section, 'Column value matching', is partially visible at the bottom.

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

# Workflow Description

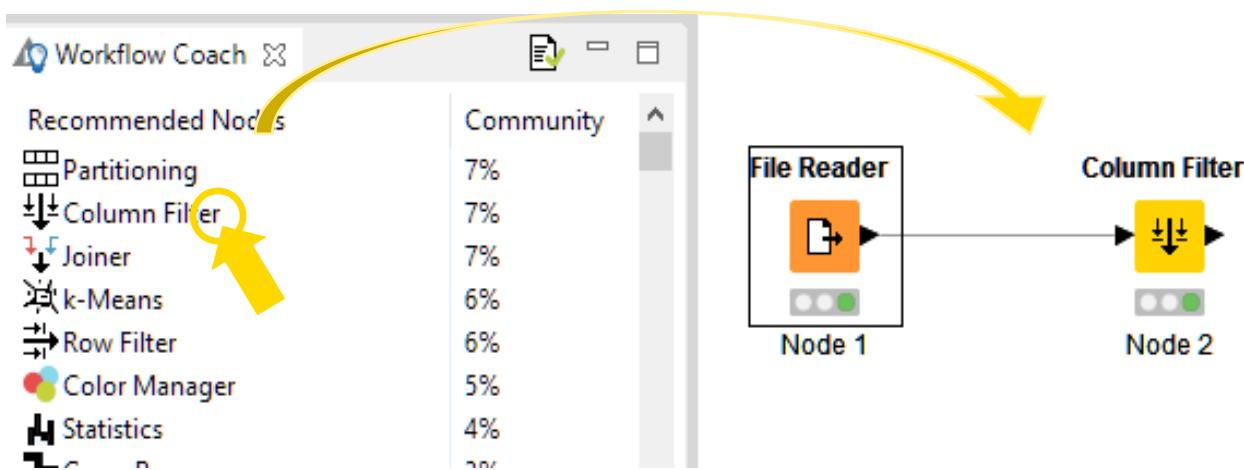
The screenshot shows the 'Description' window for a workflow named 'My\_First\_Workflow'. The window has a yellow header bar with a question mark icon, the title 'Description', and a close button. The main content area contains the following information:

- Title:** My First Workflow
- Description:** This workflow reads data, removes uninteresting columns and rows, and writes the resulting data table to a CSV file.
- Tags:** Example Workflow, CSV, Data Manipulation
- Links:**
  - KNIME Homepage
  - KNIME Hub
  - KNIME Forum
- Creation Date:** 2019-7-2
- Author:** Ana Vedoveli

- 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



# Node Monitor

- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.

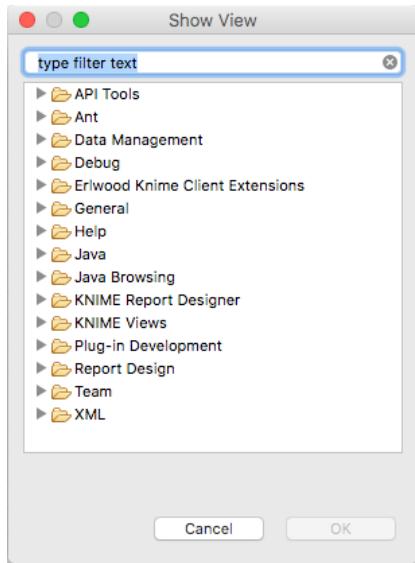
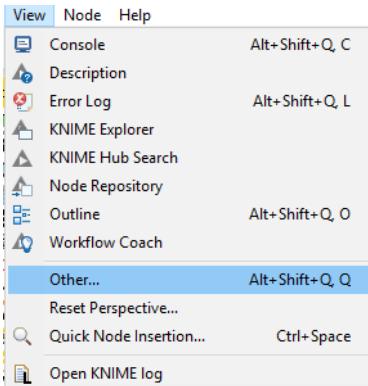
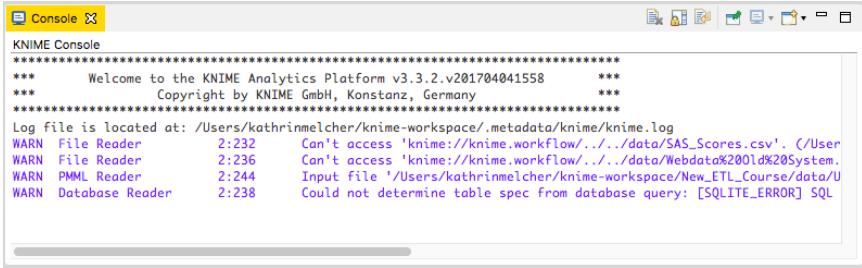
The screenshot shows the KNIME Node Monitor interface. At the top, there are tabs for 'Console' and 'Node Monitor'. Below the tabs, it displays the node information: 'Node: Get Customers from Database (0:1207)' and 'State: EXECUTED'. A dropdown labeled 'Port Output' is set to 'Port 0', and a button labeled 'Load data' is visible. To the right of the table, a context menu is open with the following options:

- Show Output Table
- Show Variables
- Show Configuration
- Show Entire Configuration
- Show Node Timing Information
- Show Graph Annotations

The main area contains a table with the following data:

ID	CustomerID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV	Products
	CustomerID: 722204	S	F	80000	4	42	1	1	4
	CustomerID: 489847	M	M	60000	2	46	1	1	4
	CustomerID: 8444723	M	M	40000	1	32	1	2	3
	CustomerID: 1487427	M	M	30000	2	63	1	1	2
	CustomerID: 4693433	M	M	20000	2	63	1	1	3
	CustomerID: 7724940	M	M	30000	2	33	1	2	3
	CustomerID: 9784443	M	M	60000	2	34	1	2	3
	CustomerID: 3177757	M	M	70000	2	57	1	1	5

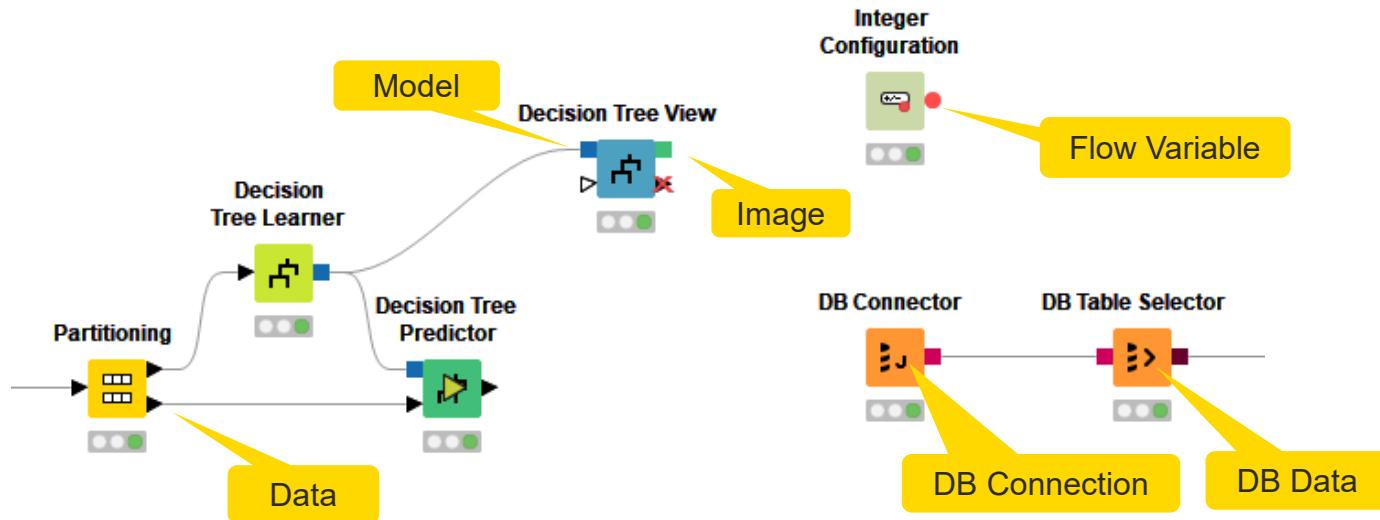
# Console and Other Views



- 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.

# Inserting and Connecting Nodes

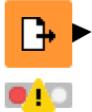
- Insert nodes into workspace by dragging them from Node Repository or by double-clicking in Node Repository
- Connect nodes by left-clicking output port of Node A and dragging the cursor to (matching) input port of Node B
- Common port types:



# More on Nodes...

- A node can have 4 states:

File Reader



## Not Configured:

The node is waiting for configuration or incoming data.

File Reader



## Configured:

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

File Reader



## Executed:

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

File Reader

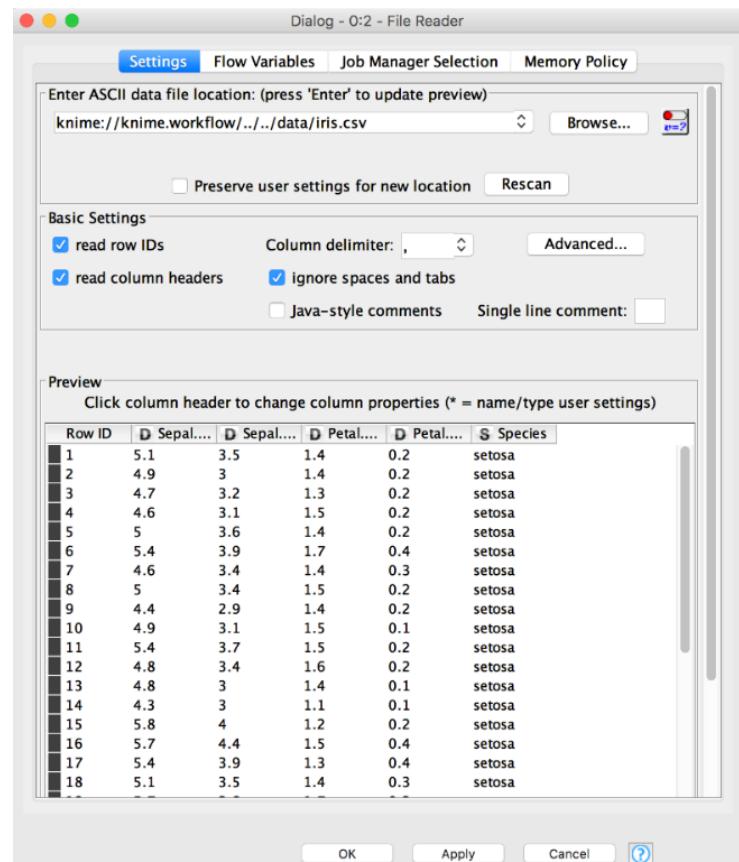


## Error:

The node has encountered an error during execution.

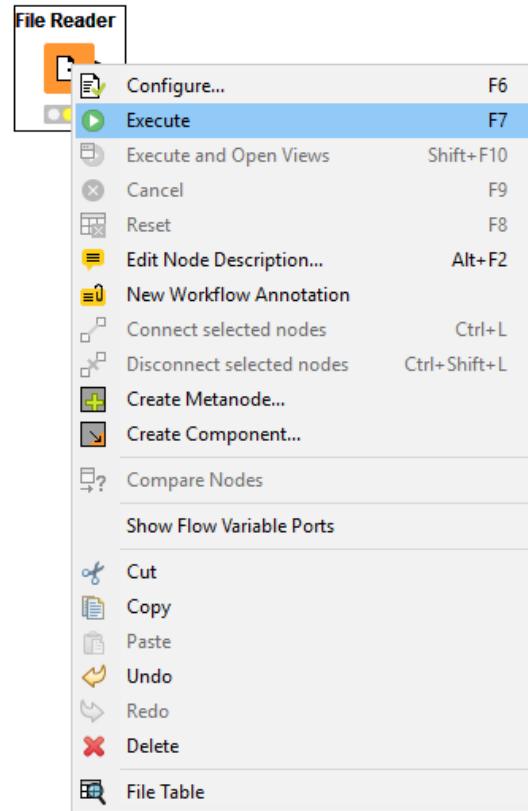
# Node Configuration

- 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



# Tool Bar

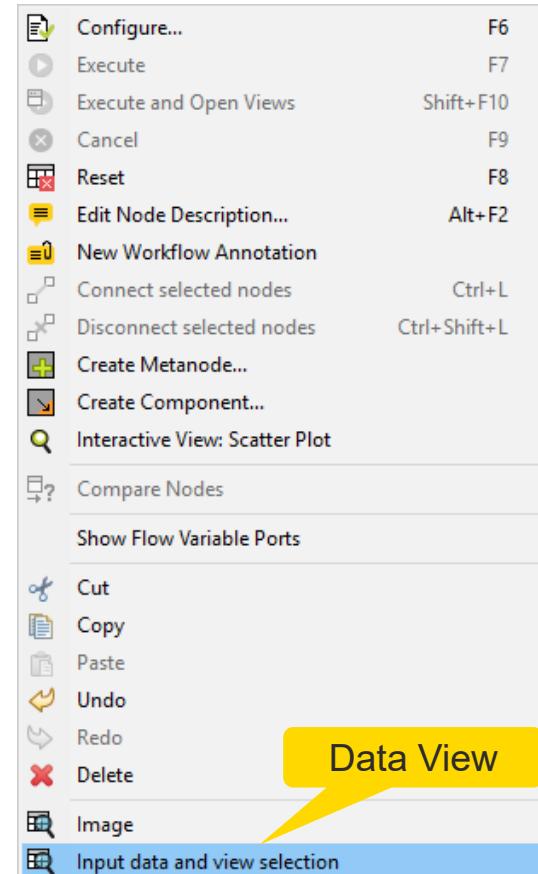
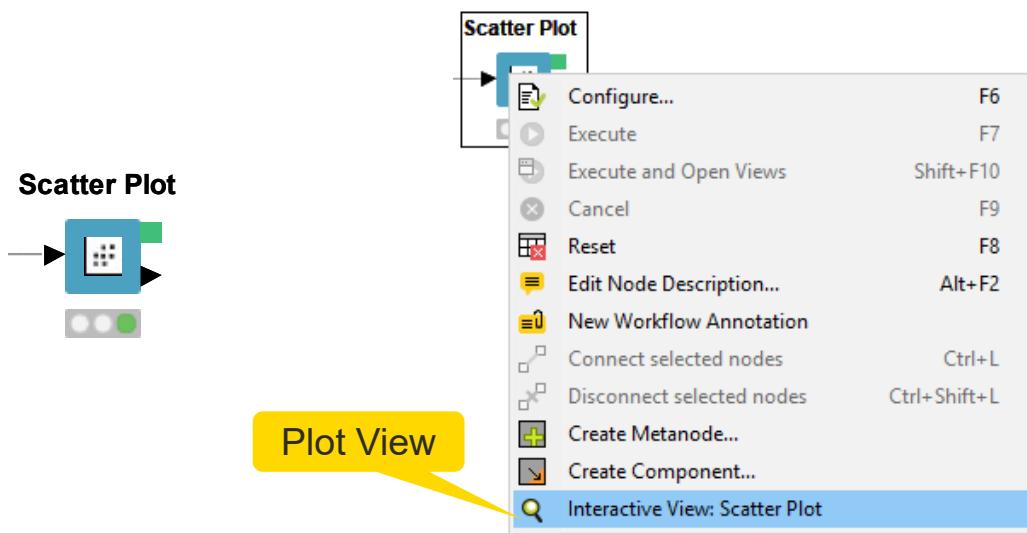


The buttons in the toolbar can be used for the active workflow. The most important buttons are:

- 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

# Node Views

- Right-click node to inspect the execution results by
  - selecting output ports (last option in the context menu) to inspect tables, images, etc.
  - selecting Interactive View to open visualization results in a browser



# 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



# Getting Started: KNIME Hub

- Place to search and share
  - Workflows
  - Nodes
  - Components
  - Extensions

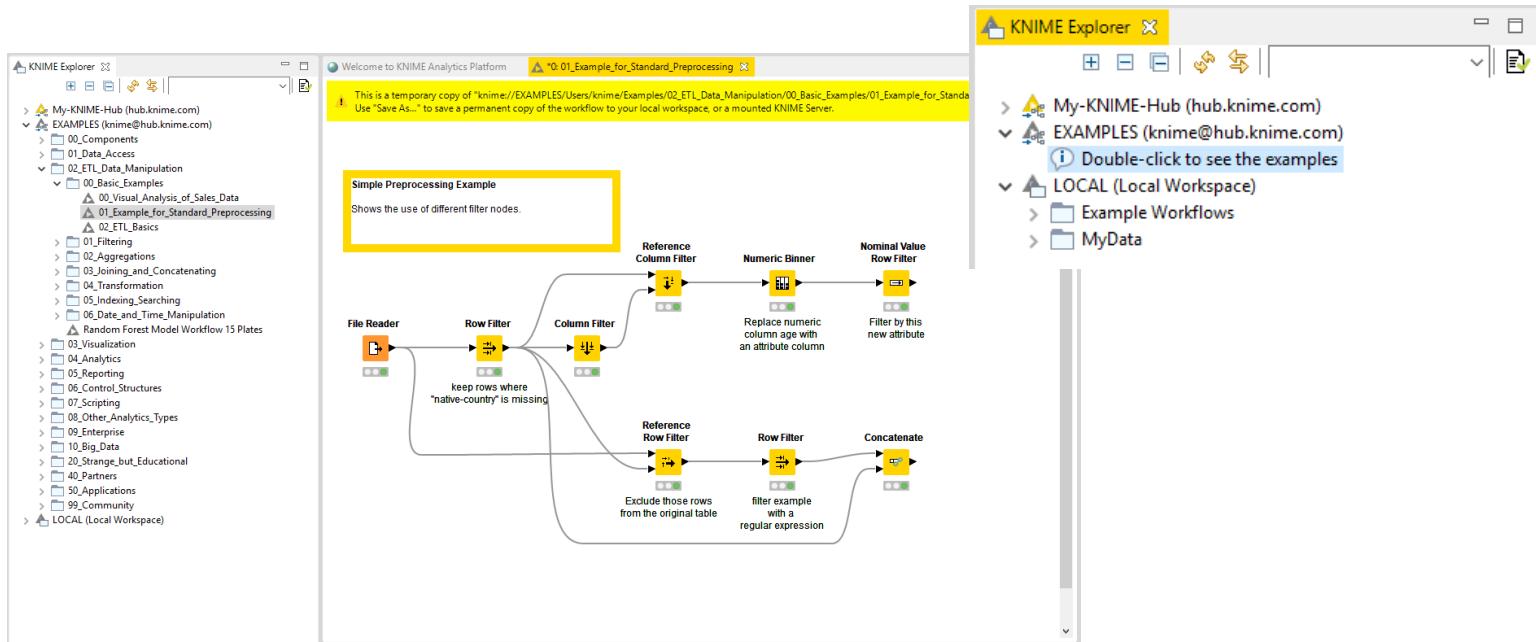
The screenshot shows the KNIME Hub search results for the query "Sentiment Analysis". The results page displays 350 items. At the top, there are filters for All, Nodes, Components, Workflows, and Extensions. Below the filters, three workflow cards are shown:

- Sentiment Analysis**: This workflow shows how to train a simple neural network for text classification, in this case sentiment analysis. The used network learns a 128 dimensional word embedding followed by an LSTM. This example uses deep learning, keras, and text classification nodes. It has 11 nodes.
- Sentiment Analysis**: This workflow shows how to train a simple neural network for text classification, in this case sentiment analysis. The used network learns a 128 dimensional word embedding followed by an LSTM. This example uses deep learning, keras, and text classification nodes. It has 12 nodes.
- Sentiment Analysis (Classification) of Documents**: This workflow shows how to import text from a csv file, convert it to documents, preprocess the documents and transform them

The screenshot shows the main homepage of the KNIME Hub. It features a search bar at the top right with the placeholder "Search workflows, nodes and more...". Below the search bar, there are four large statistics: 3 993 Nodes, 265 Components, 2 541 Workflows, and 211 Extensions. On the left, there is a "How to Getting started" section with a diagram showing a workflow node flow. On the right, there is a "Forum" section with the text "Get help from our community and help others". At the bottom, the URL <https://hub.knime.com> is displayed.

# Getting Started: KNIME Example Server

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



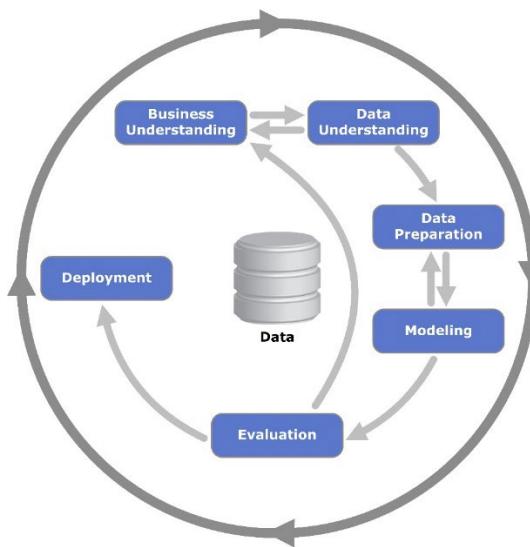
# Hot Keys (for Future Reference)

Task	Hot key	Description
Node Configuration	F6	opens the configuration window of the selected node
	F7	executes selected configured nodes
	Shift + F7	executes all configured nodes
Node Execution	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
	F8	resets selected nodes
Workflow Operations	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

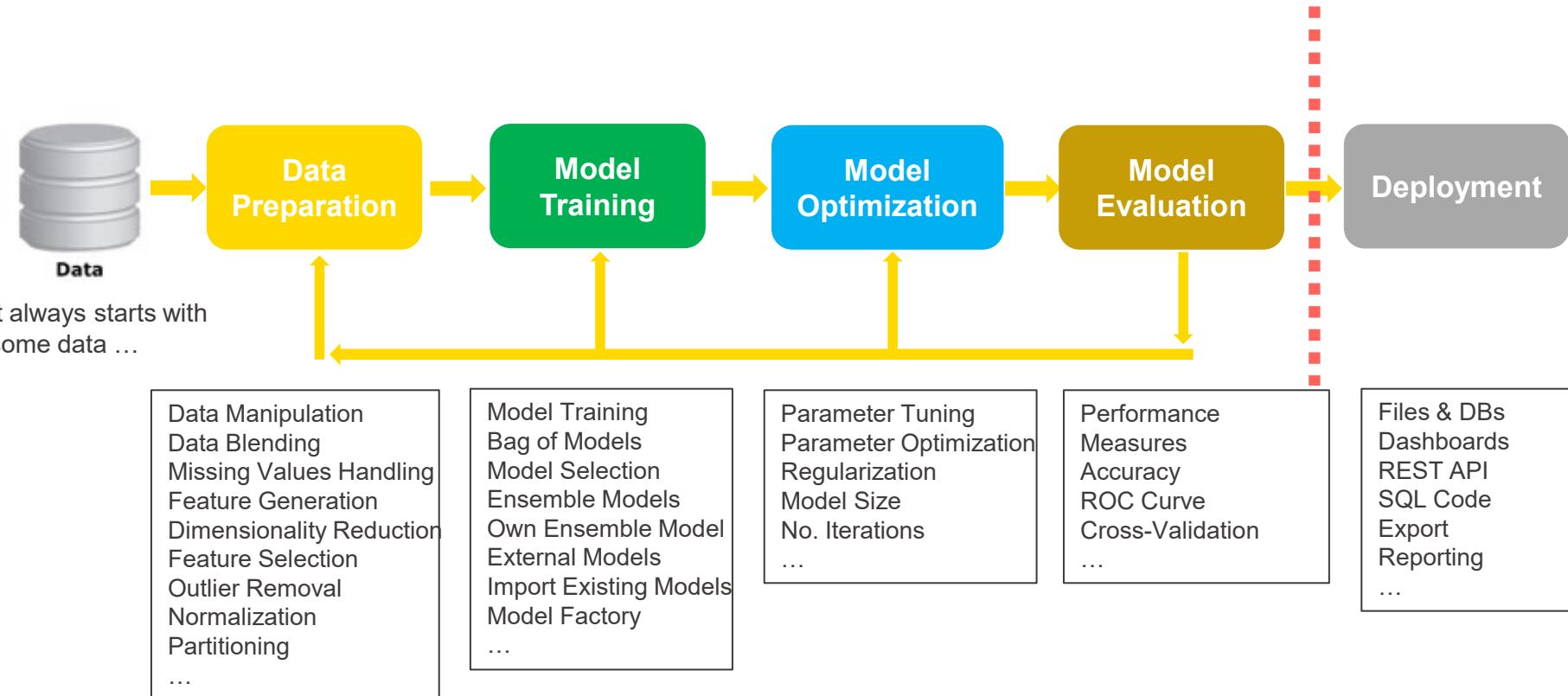
# Today's Example

# 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



# 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

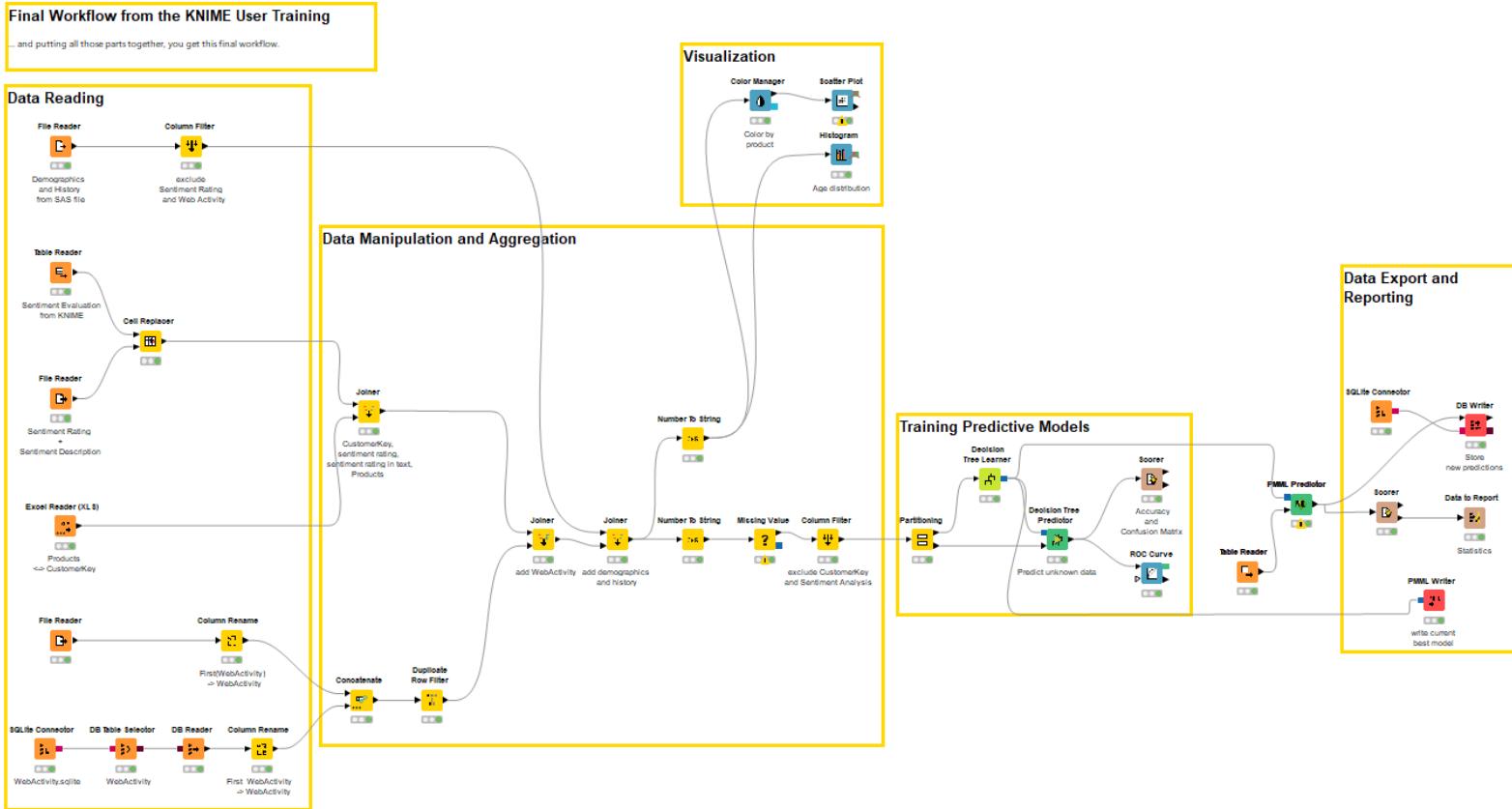
The diagram illustrates the Data Explorer node and its connection to the Data Explorer View window. On the left, a blue rectangular node labeled "Data Explorer" has two output ports. The top port is connected by a grey arrow to a yellow arrow pointing towards a central window. The bottom port is connected to three small colored circles (grey, green, and blue). The central window is titled "Data Explorer View" and displays a table of data statistics. The table has columns for Column, Exclude Column, Minimum, Maximum, Mean, and Standard Deviation. The data rows are as follows:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation
CustomerKey		11000	27336	19281.750	5319.909
WebActivity		0	5	1.000	1.524
SentimentRating		0	5	1.846	1.619
EstimatedYearlyIncome		10000	170000	57066.921	32242.624
NumberOfContracts		0	4	1.493	1.145
Age		29	100	48.288	11.382
Target		0	1	0.489	0.500

To the right of the Data Explorer View window is a file browser interface showing the contents of the "data" folder. The folder structure is as follows:

- KNIMEUserTraining
  - data
    - temp
      - 50IPs.table
      - books.json
      - books.xml
      - ContractValues.csv
      - CurrentDetailData.table
      - database.mv.db
      - database.trace.db
      - DecTree.pmml
      - location\_data.table
      - meter\_data.csv
      - onelineDeploymentData.csv
      - Product Data2.xls
      - ProductPropensitymodel.pmml
      - sales.csv
      - sampled\_meter\_data.table
      - SAS\_Scores.csv
      - Sentiment Analysis.table
      - Sentiment Rating.csv
      - UpsellTriplePlay.pmml
      - weather.table
      - WebActivity.sqlite
      - Webdata Old System.csv
    - exercises
    - solutions
    - Supplementary Workflows
    - Final Workflow

# Today's Example: Churn Prediction



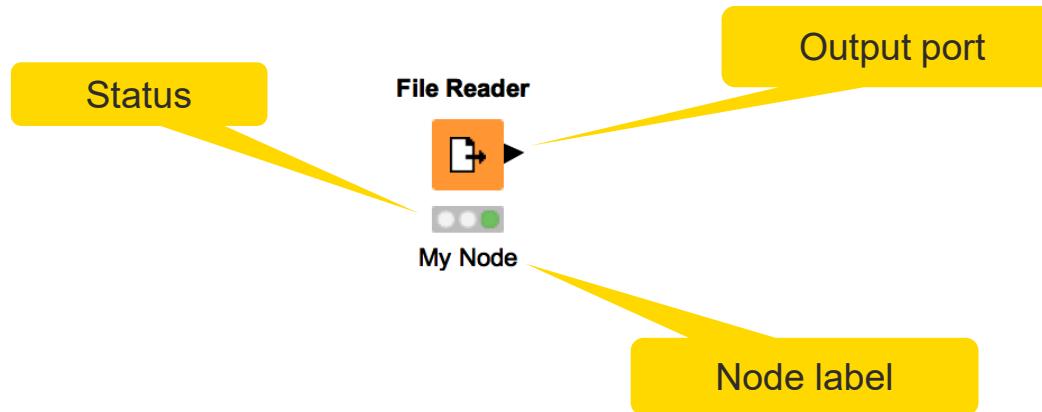
# **Importing Data**

## **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

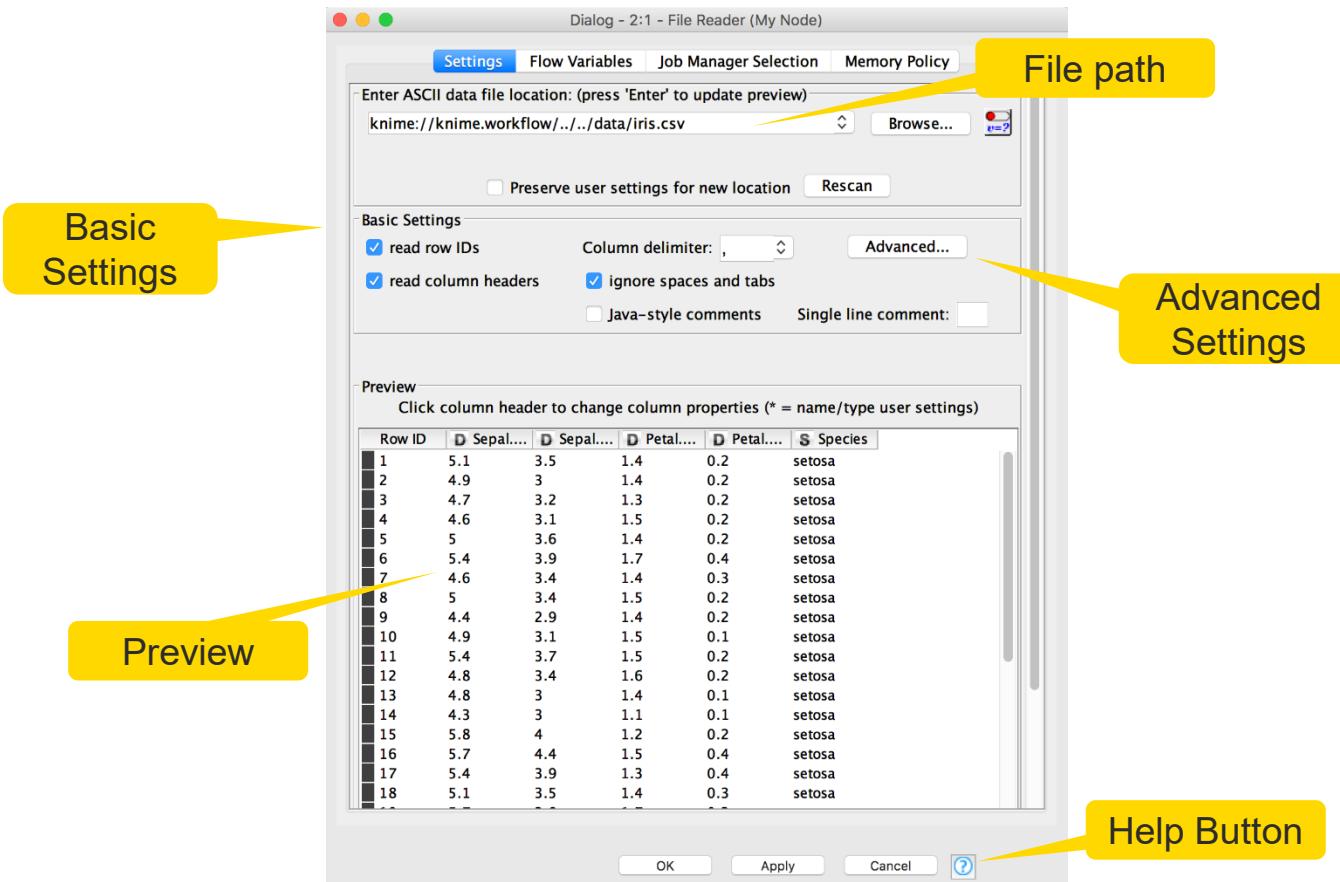
- Reads all text based files (e.g. csv, txt, etc.)
- Many advanced features allow it to read most ‘weird’ files
  - Short lines, inline comments, headers and special encoding



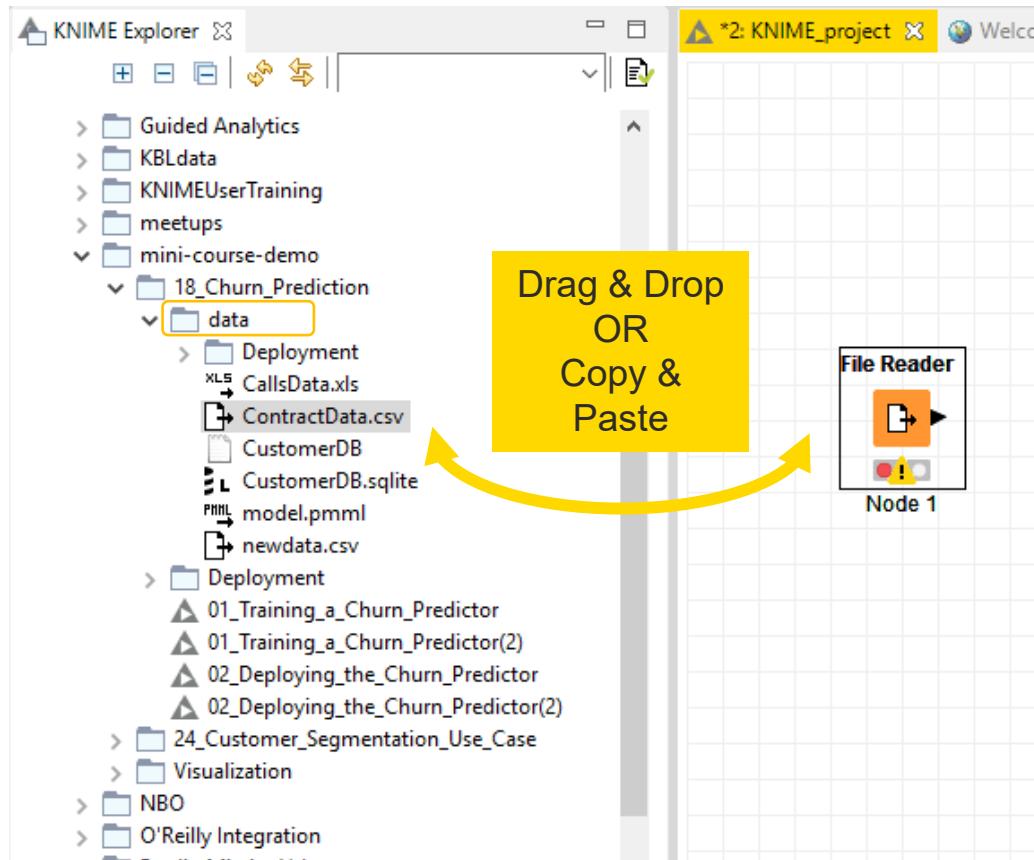
YouTube KNIME TV Channel video:

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

# File Reader Configuration

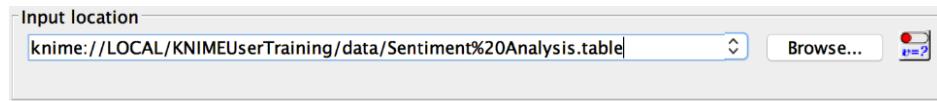


# Alternative Faster Way ...

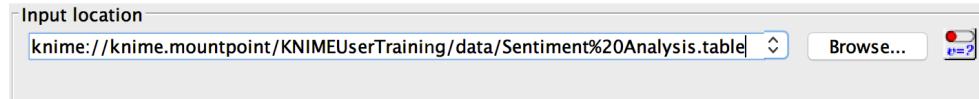


# Filenames and the knime:// Protocol

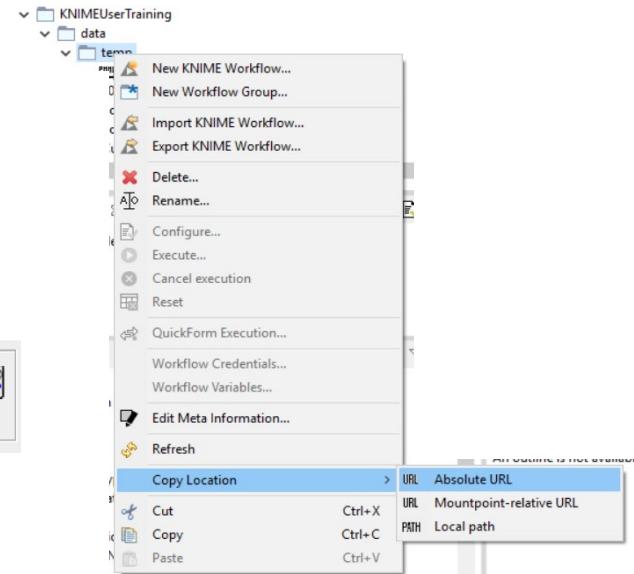
- Absolute URL



- Mountpoint-relative URL

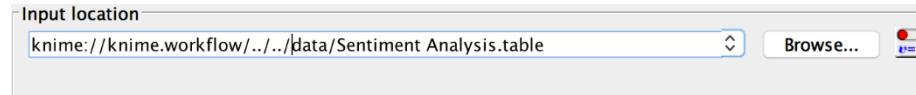
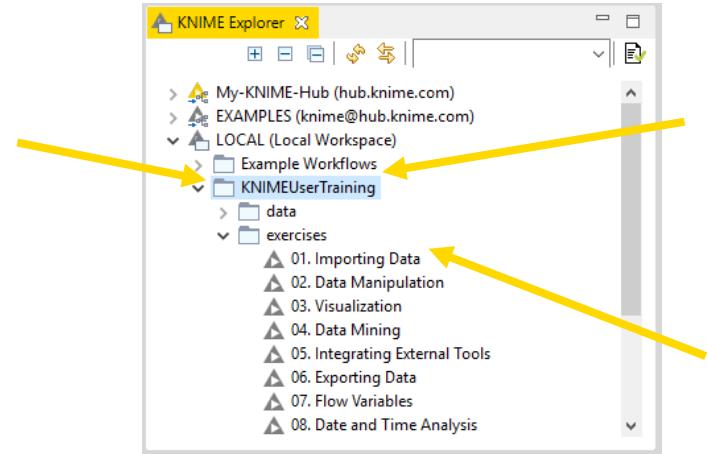


- Local path



# 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: Simple File Reader

- Faster compared to the File Reader node
- Only basic settings

**Simple File Reader**



Preview

File path

Basic settings

Preview

Dialog - 0:1210 - Simple File Reader

Settings Advanced Settings Limit Rows Encoding Flow Variables Memory Policy

Input location: cher/knime-workspace-dw-course/L1-DW KNIME Analytics Platform for Data Wranglers - Ba Browse...

Connection timeout [s]: 1

Reader options

Format

Autodetect format

, Column delimiter \n Row delimiter

" Quote char " Quote escape char

# Comment char

Has column header  Has row ID

Support short data rows

Preview

The suggested column types are based on the first 50 rows only. See 'Advanced Settings' tab.

Row ID	Accou...	Churn	Int'l Plan	VMail ...	State	Area ...	Phone
Row0	128	0	0	1	KS	415	382-4657
Row1	107	0	0	1	OH	415	371-7191
Row2	137	0	0	0	NJ	415	358-1921
Row3	84	0	1	0	OH	408	375-9999
Row4	75	0	1	0	OK	415	330-6626
Row5	118	0	1	0	AL	510	391-8027
Row6	121	0	0	1	MA	510	355-9993
Row7	147	0	1	0	MO	415	329-9001
Row8	117	0	0	0	LA	408	335-4719
Row9	141	0	1	1	WV	415	330-8173
Row10	65	1	0	0	IN	415	329-6603
Row11	74	0	0	0	RI	415	344-9403
Row12	168	0	0	0	IA	408	363-1107
Row13	95	0	0	0	MT	510	394-8006
Row14	62	0	0	0	IA	415	366-0239

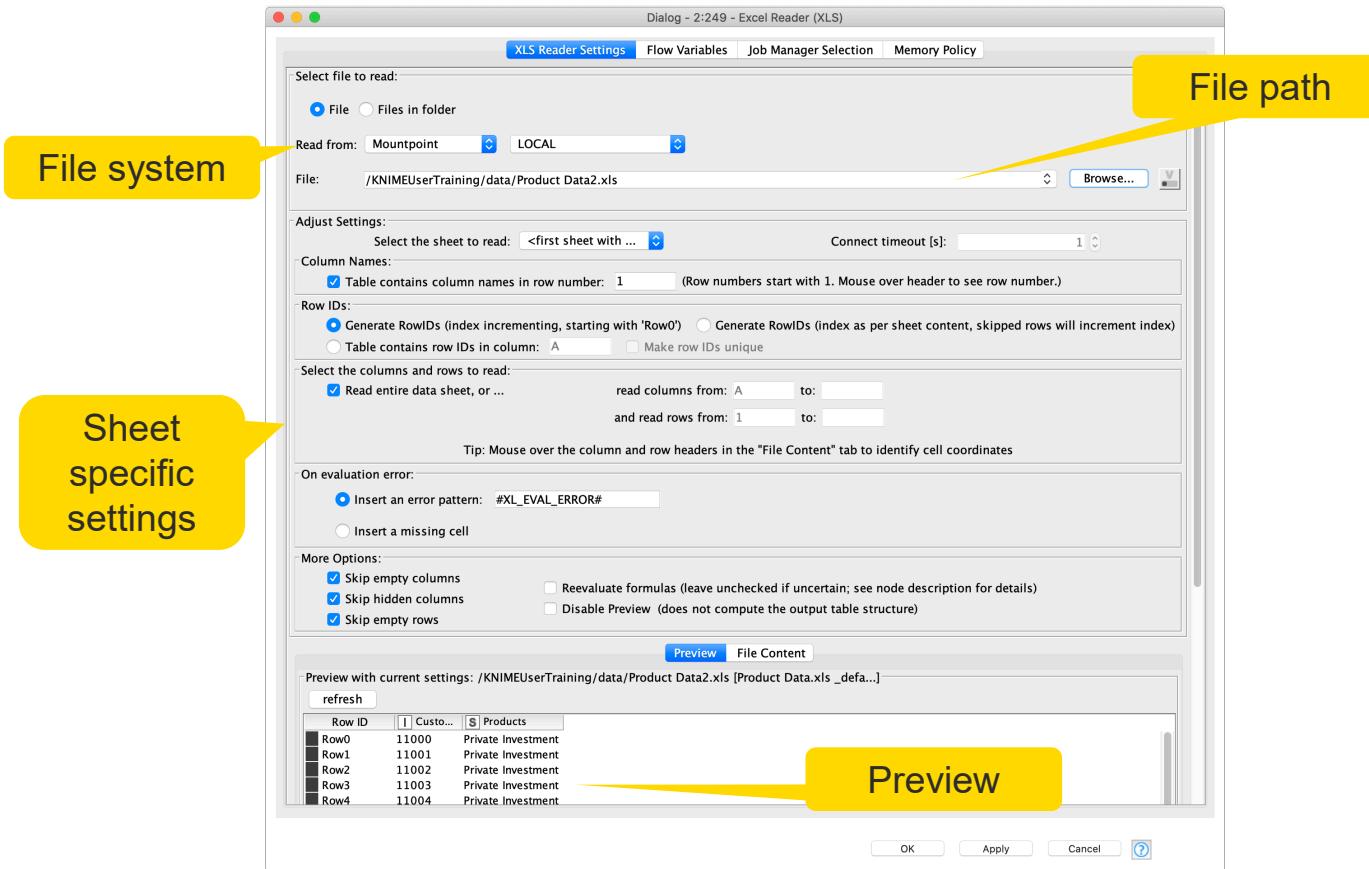
OK Apply Cancel ?

# 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

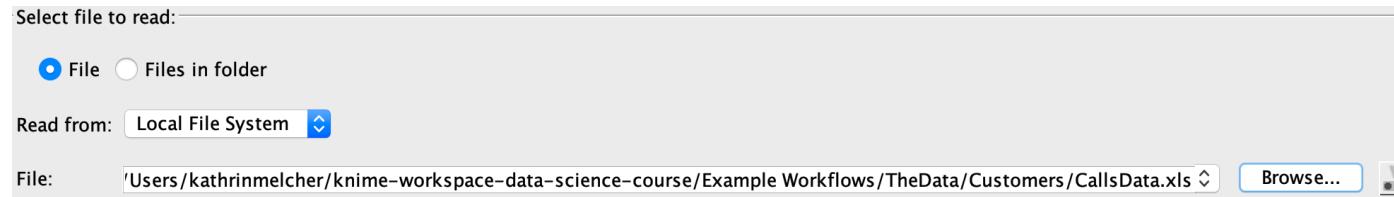
## ▪ Local File System

Select file to read:

File  Files in folder

Read from: Local File System

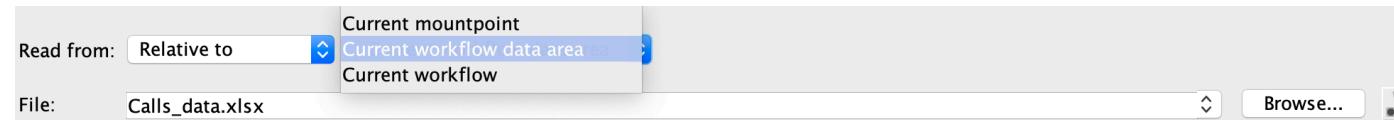
File: 'Users/kathrinmelcher/knime-workspace-data-science-course/Example Workflows/TheData/Customers/CallsData.xls'



## ▪ Relative to ...

Read from: Relative to

File: Calls\_data.xlsx



## ▪ Mountpoint

Read from: Mountpoint

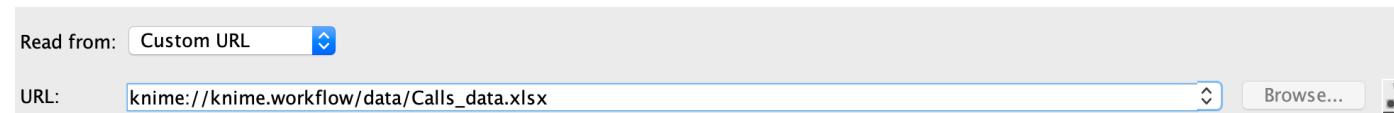
File: /Example Workflows/TheData/Customers/CallsData.xls



## ▪ Custom URL

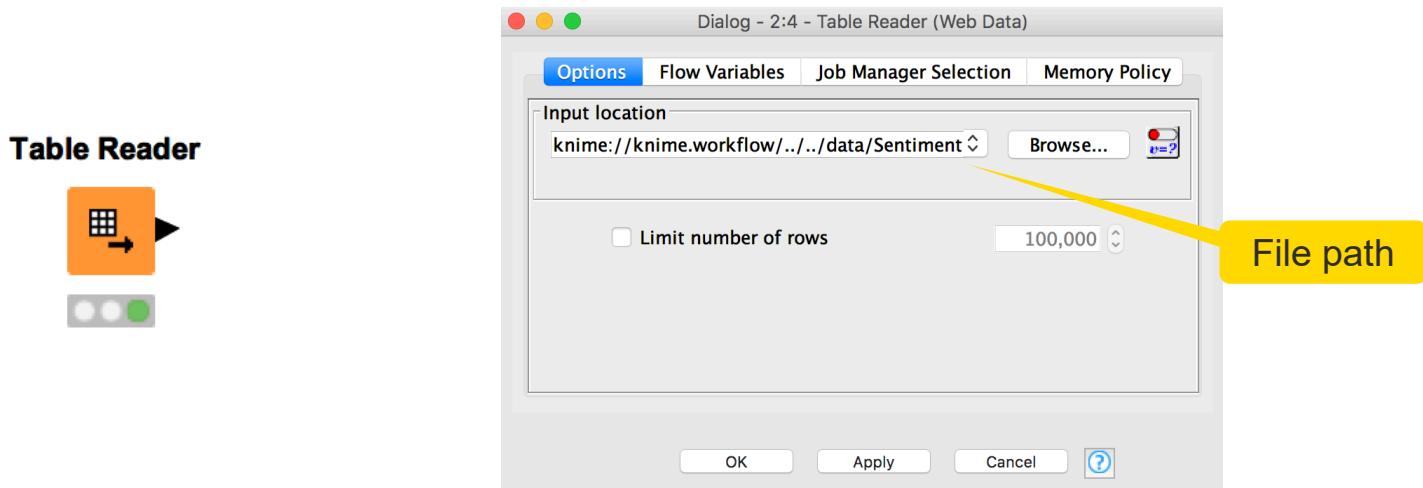
Read from: Custom URL

URL: knime://knime.workflow/data/Calls\_data.xlsx



# New Node: Table Reader

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

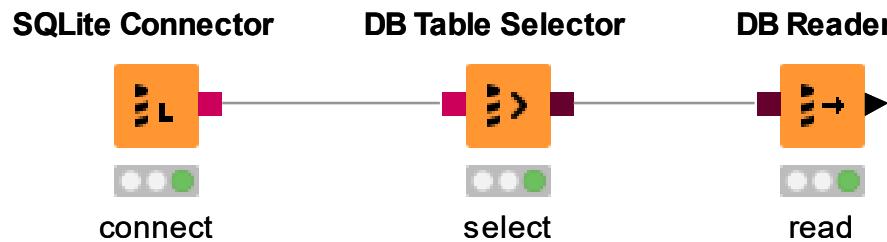


YouTube KNIME TV channel video:

<https://youtu.be/tid1qi2HAOo>

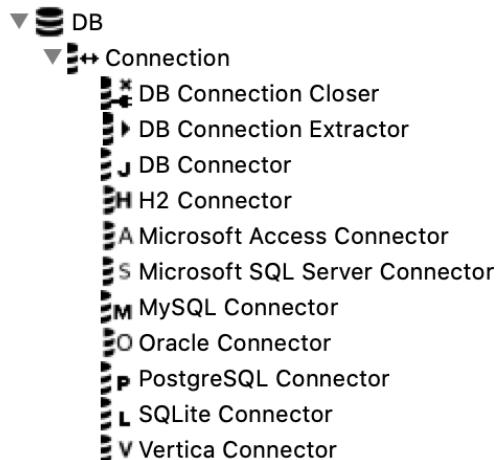
# Database Connectivity

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

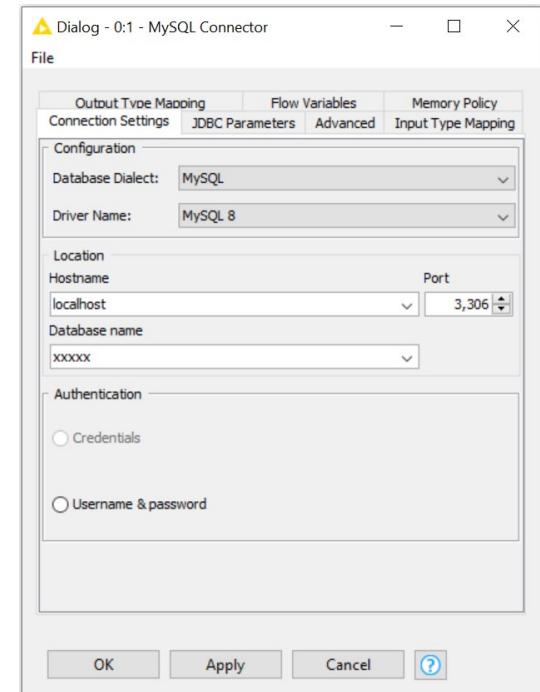


# New Nodes: Database Connectors

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



MySQL Connector



# 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/spaces/Education/latest/Courses/>

Table Reader



Sentiment Evaluation  
from KNIME

File Reader



Sentiment Rating  
+  
Sentiment Description

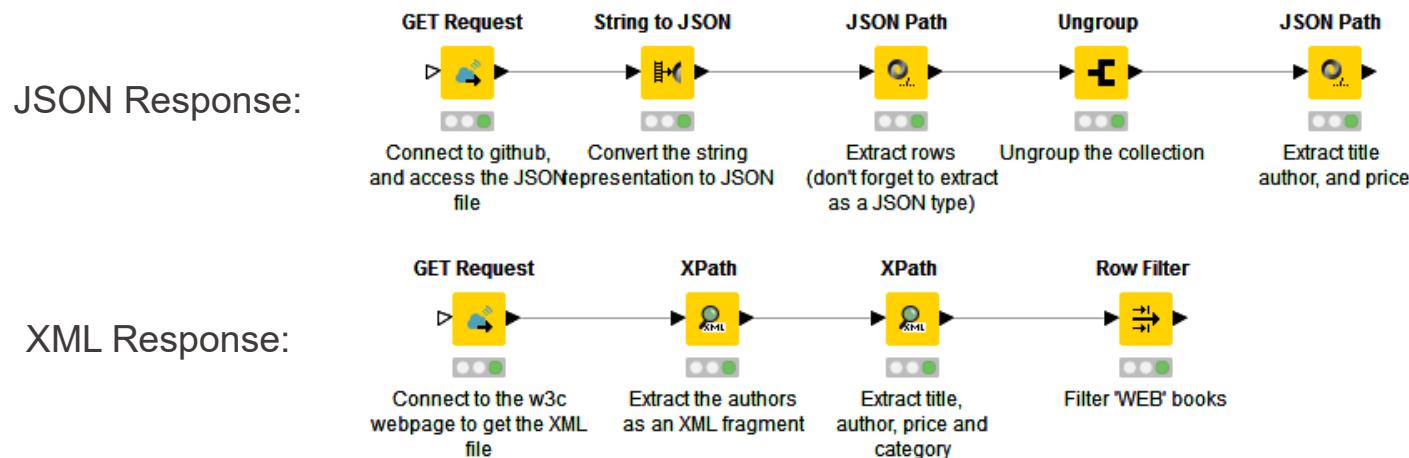
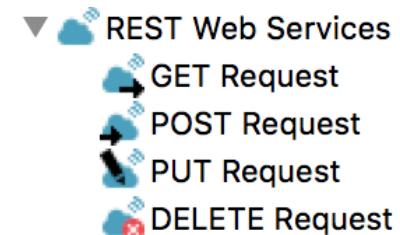
Excel Reader (XLS)



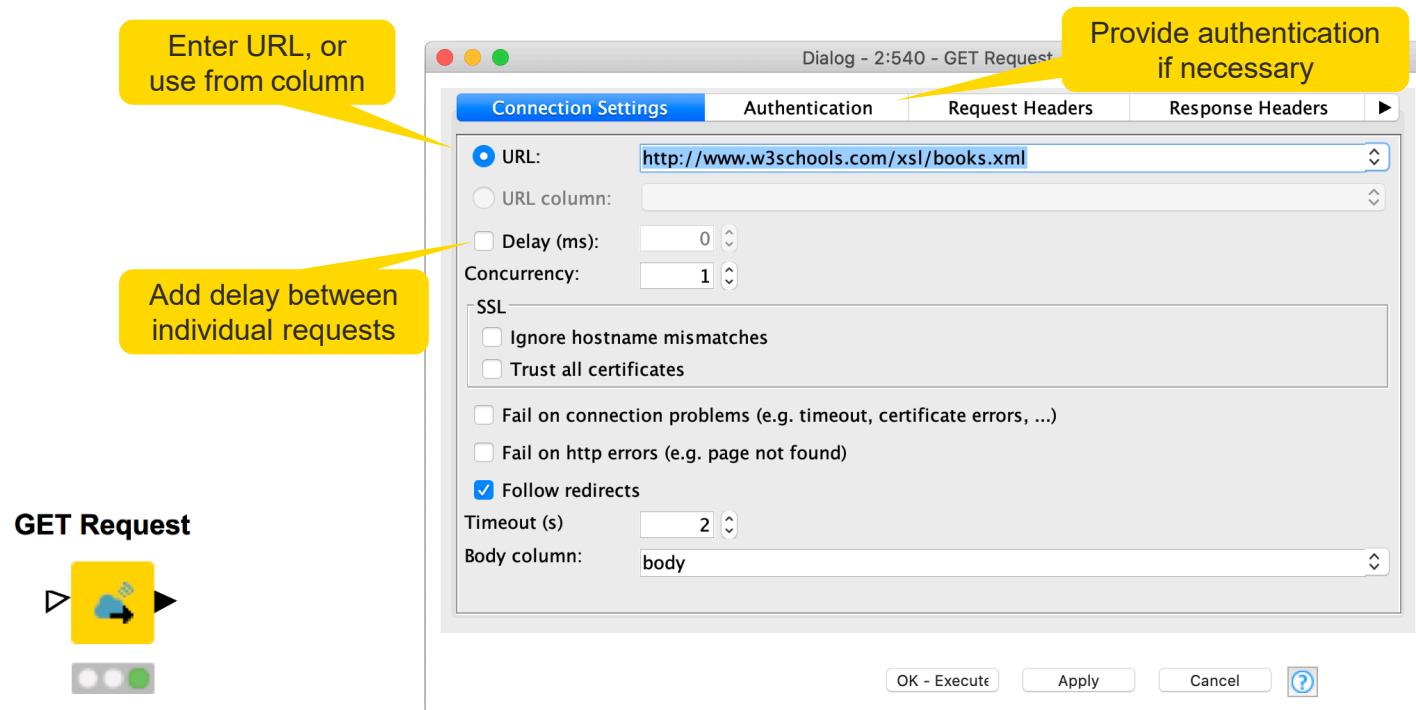
Products  
<->  
Customer

# RESTful Web Services

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



# RESTful Web Services

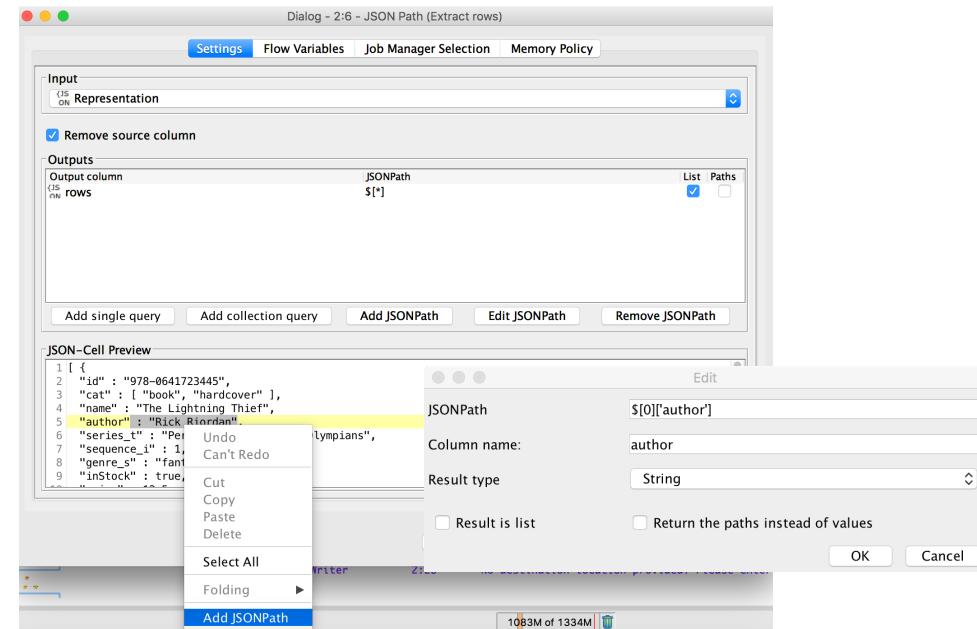


<https://www.knime.com/blog/a-restful-way-to-find-and-retrieve-data>

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

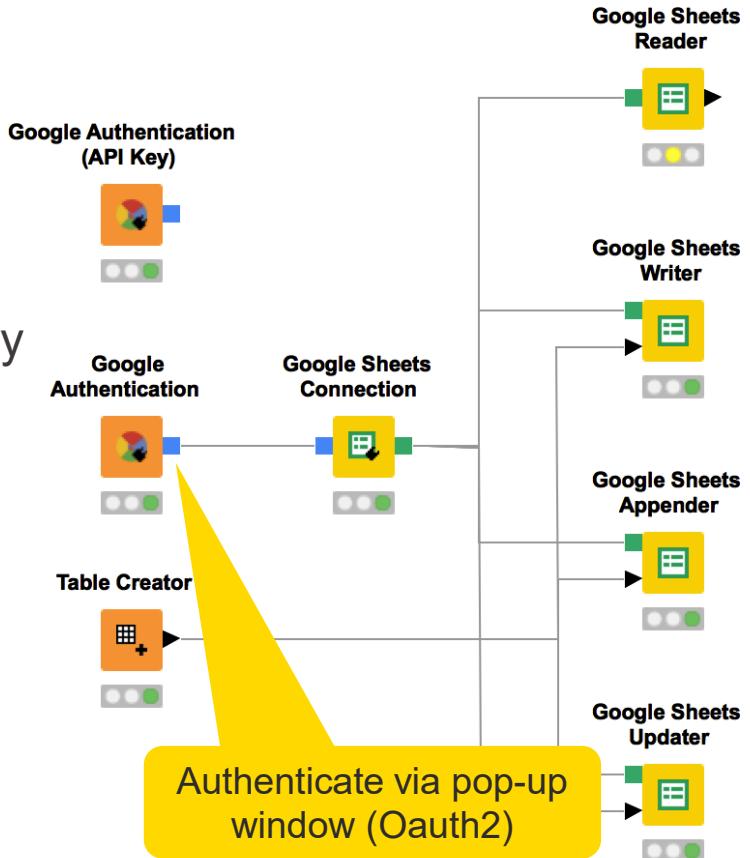
# JSON Reader and JSON Path nodes

- 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)



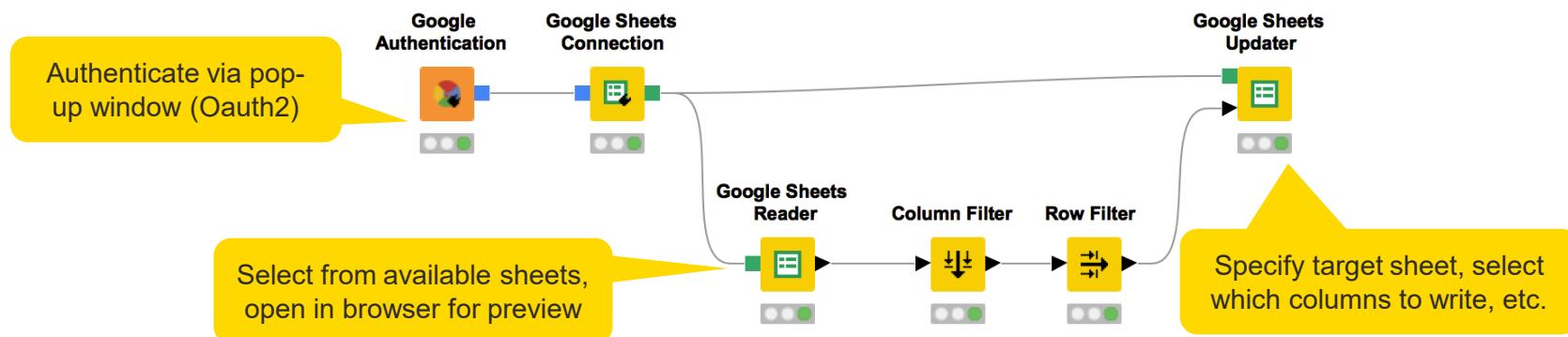
# Google Sheets

- 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)

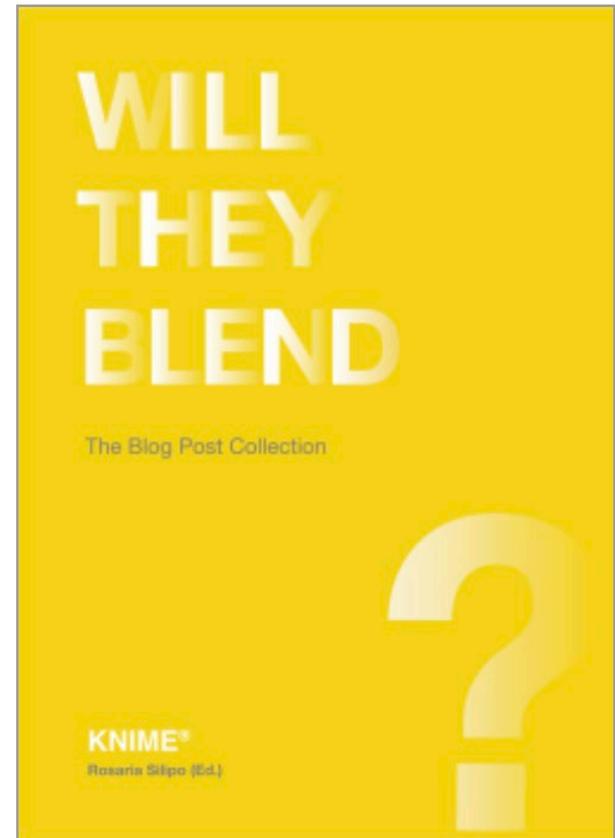


# Other Useful Data Sources

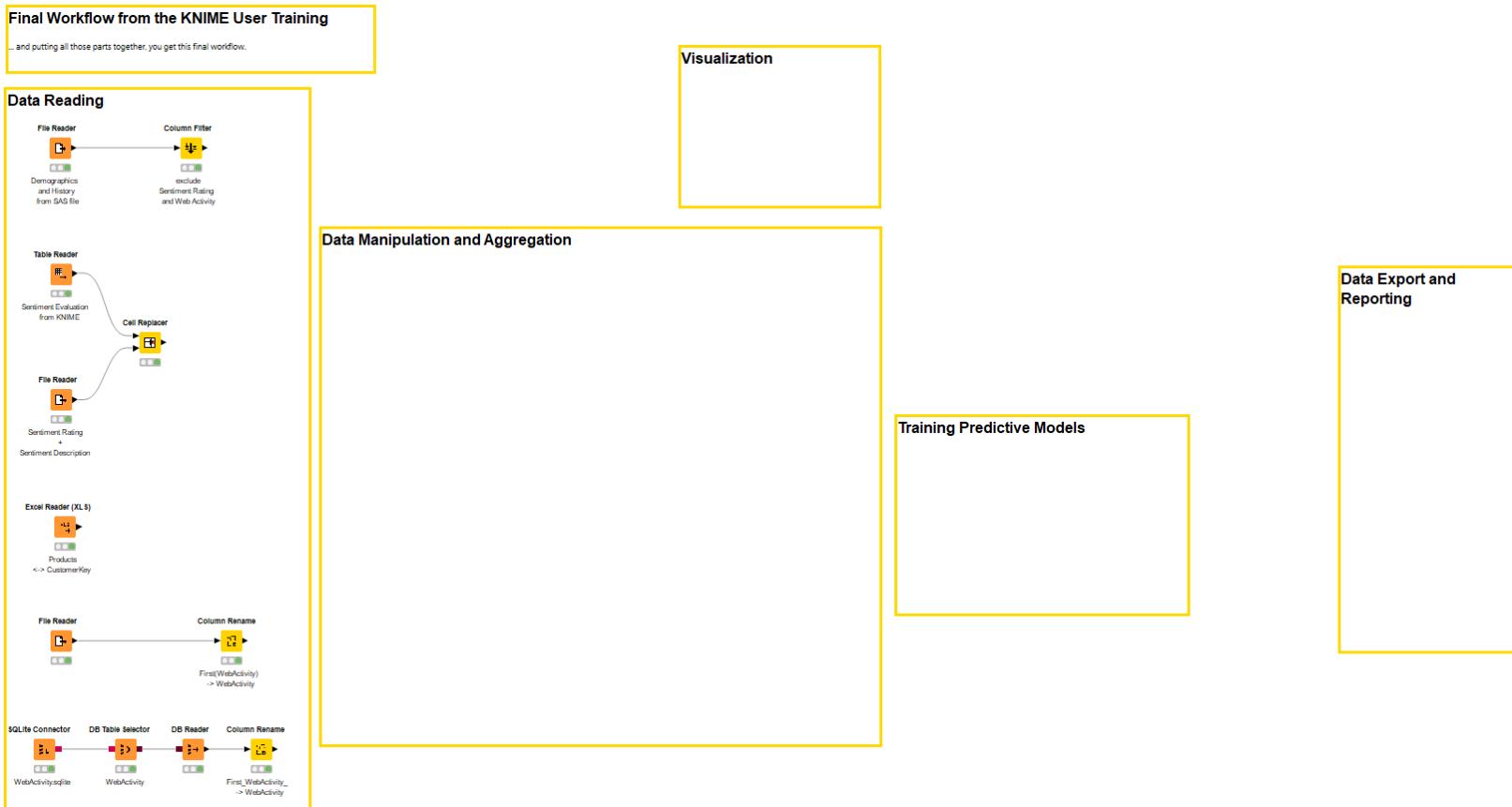
- KNIME Analytics Platform provides many more options to access data:
  - PMML Reader – reads standard predictive models
  - XML Reader with XPATH support
  - Python/R Source nodes
  - Tika Parser – extracts textual data from 200+ file types
  - REST Web Services, and many more



- Find out more in by downloading the free book “Will they blend”  
<https://www.knime.com/knimepress/download-will-they-blend>



# Today's Example

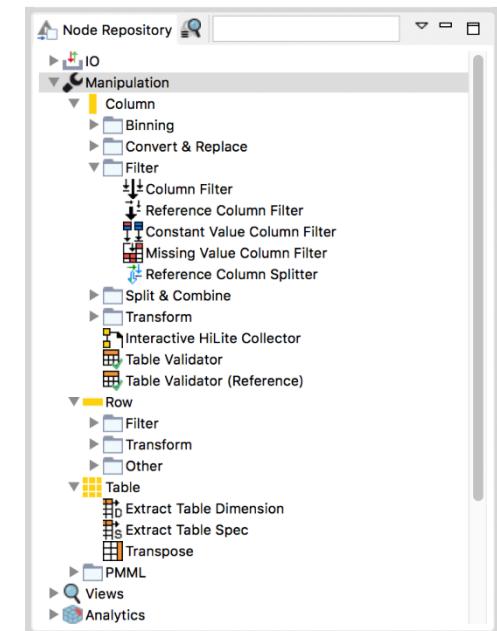
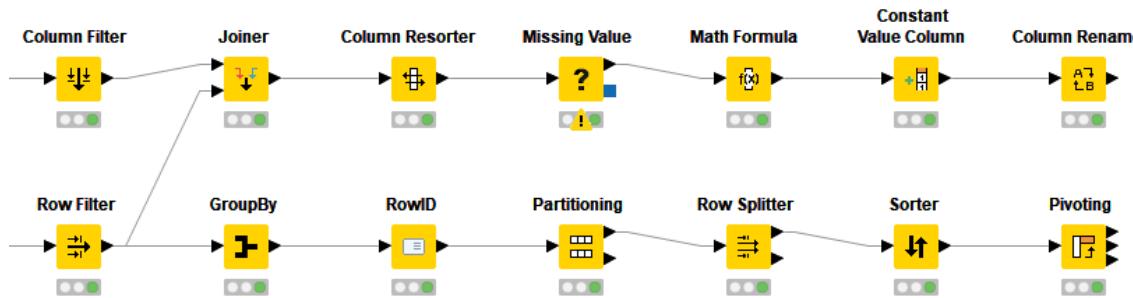


# Data Manipulation

## Clean, Join, Aggregate

# Data Manipulation Nodes

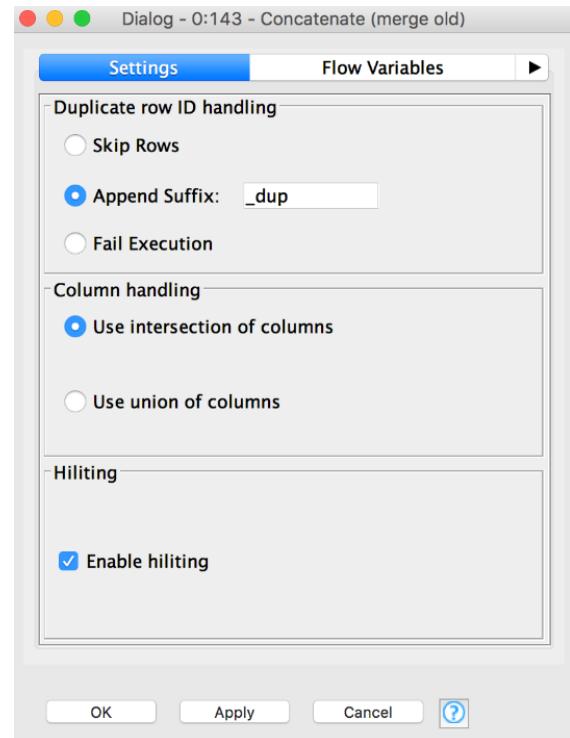
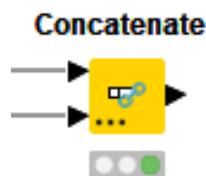
- Yellow color with a variety of input and output ports
- Apply a transformation to input data
- Many, many nodes!



# New Node: Concatenate

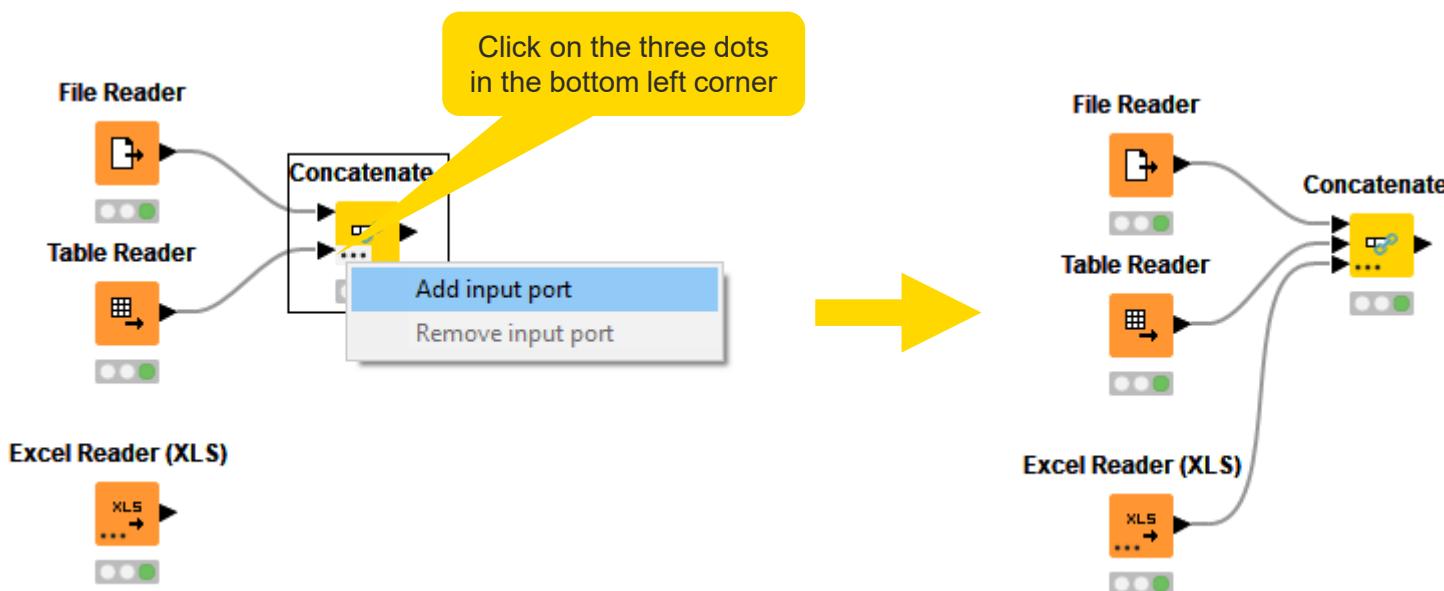
Combine rows from 2 or more tables with shared columns

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



# Dynamic Ports

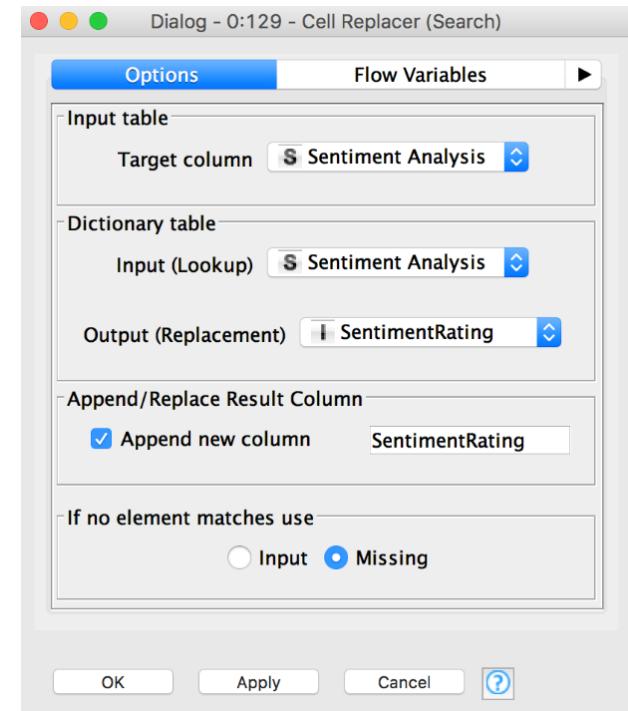
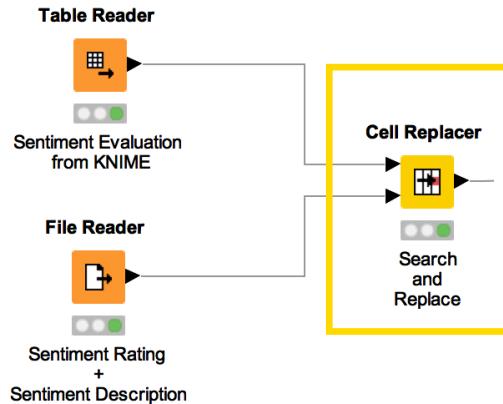
Add and remove node ports based on your needs, e.g. in order to concatenate three or more tables



# New Node: Cell Replacer

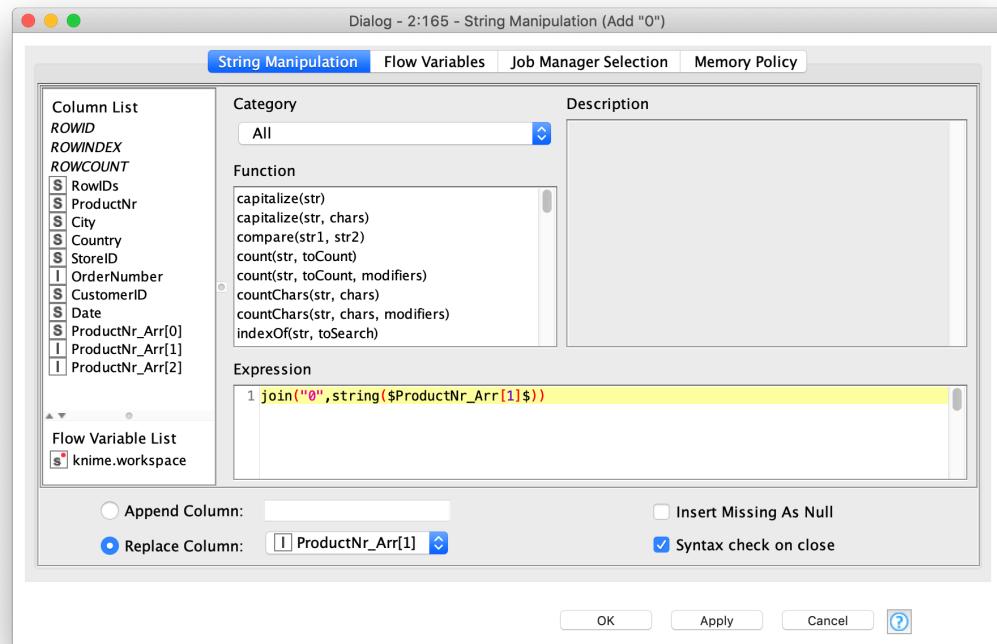
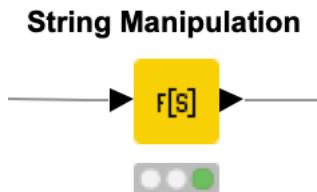
Replaces the content of a column based on a lookup

- 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 a String Column
  - Cleans up capitalization
  - Joins string values
  - Pads strings, e.g. padLeft
  - Replaces string values



# 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

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

Join by CustomerKey

Inner Join

Left Outer 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

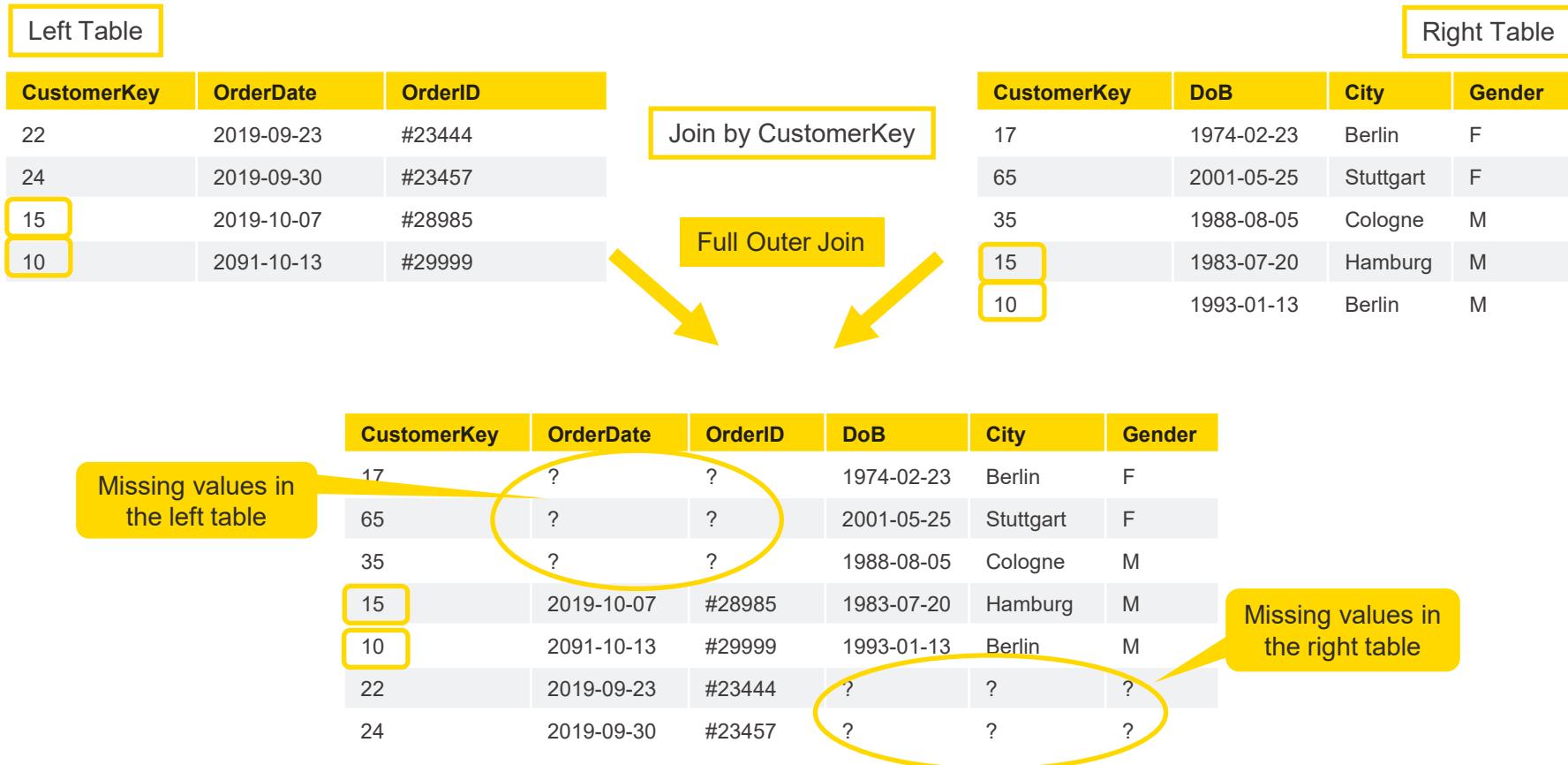
Right Outer Join

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

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

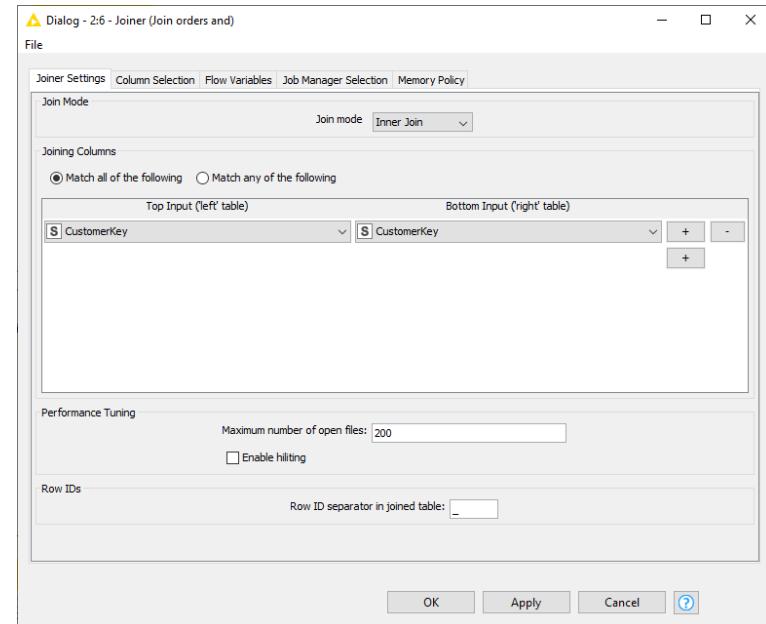
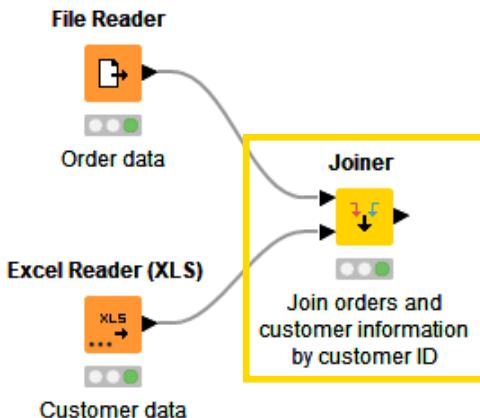
# Joining Columns of Data



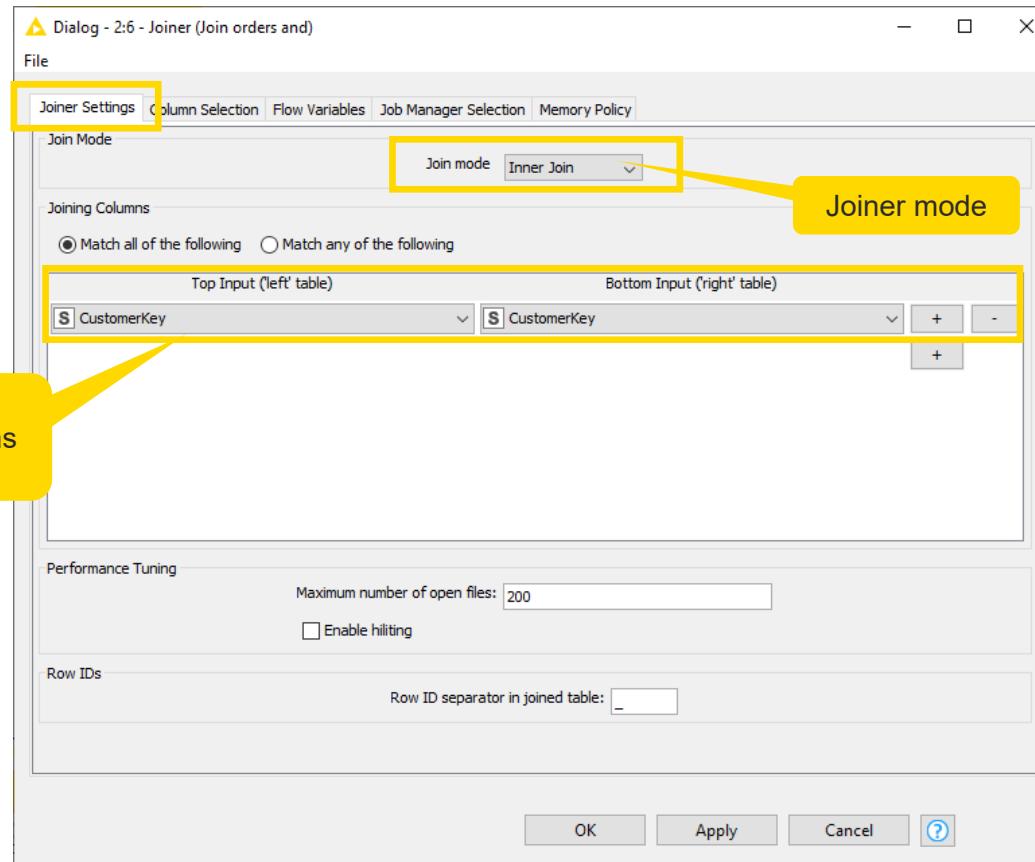
# New Node: Joiner

Combines columns from 2 different tables

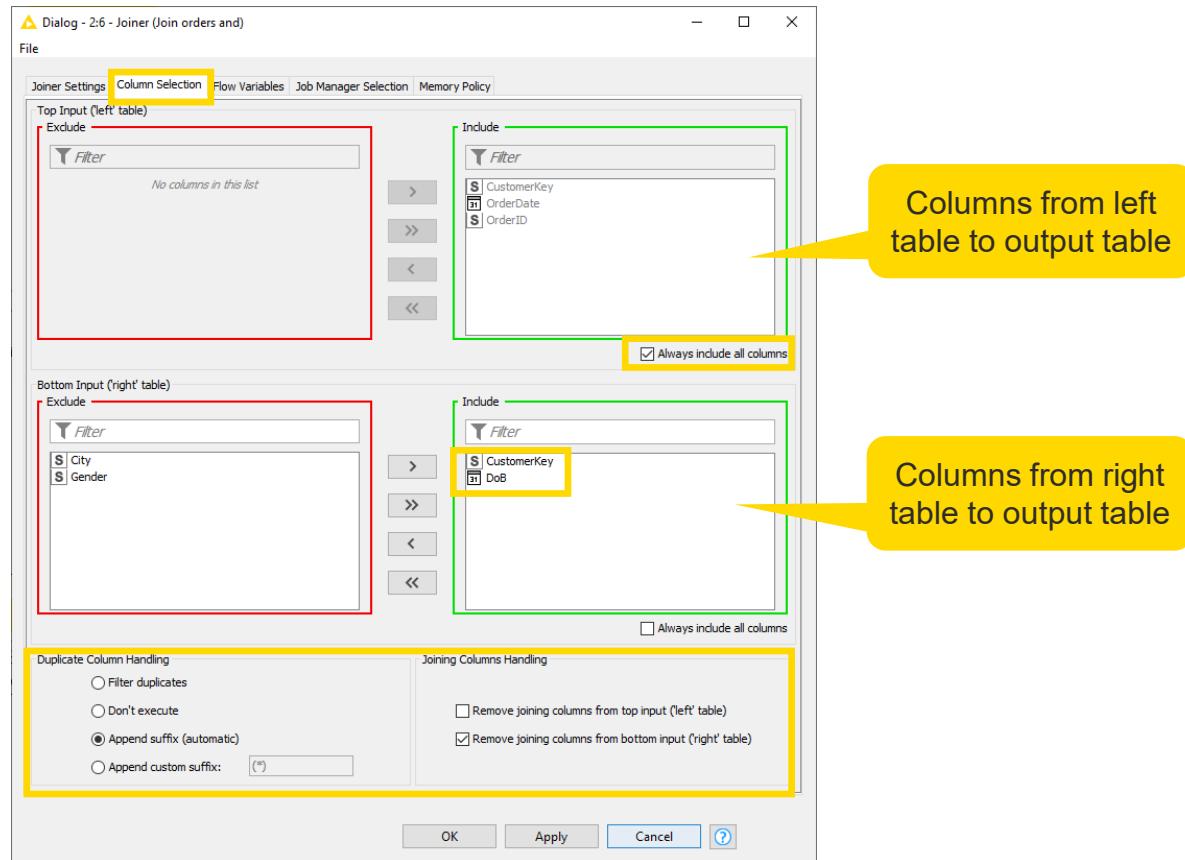
- Top port contains “Left” data table
- Bottom port contains “Right” data table



# Joiner Configuration – Linking Rows



# Joiner Configuration – Column Selection



# 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



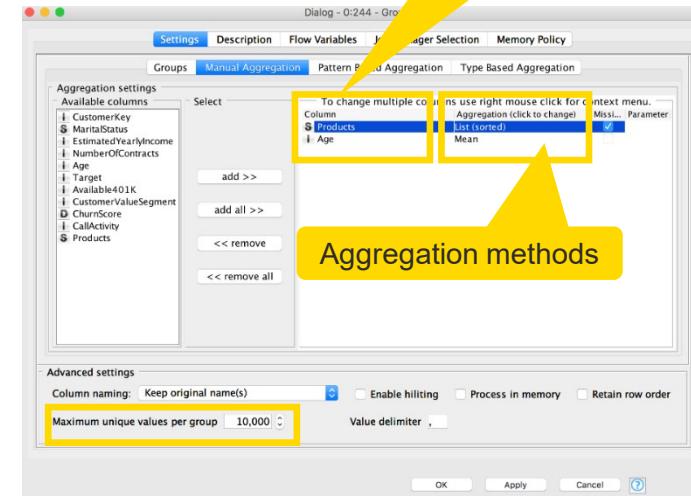
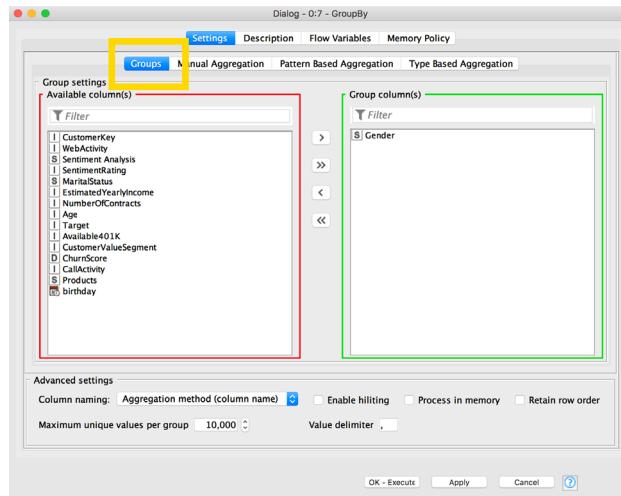
Group	Sum(# Ordered Items)
Clothing	8
Home	3
Electronics	12

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

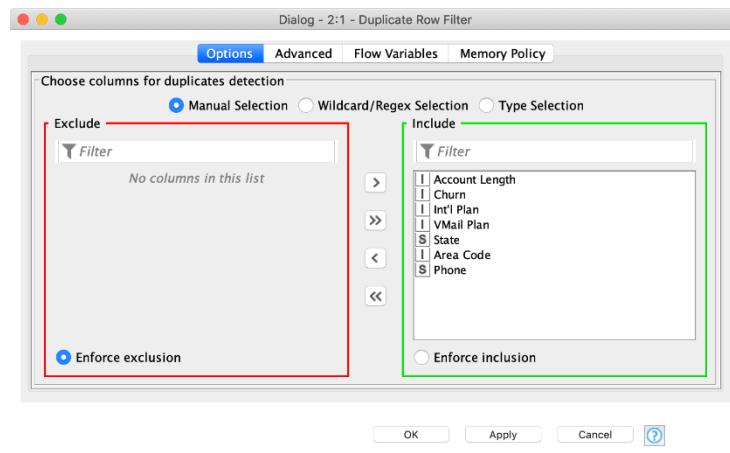


YouTube KNIME TV video: <https://youtu.be/bDwF-TOMtWw>

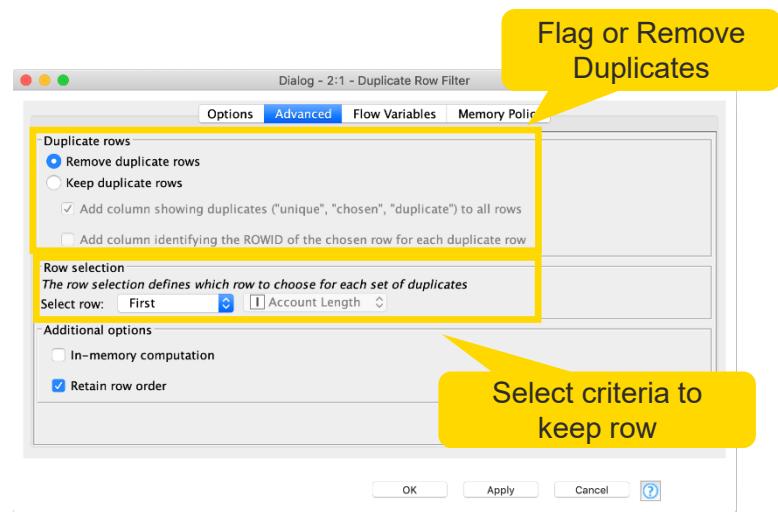
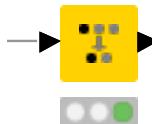
# New Node: Duplicate Row Filter

Detect duplicate rows and apply a selected treatment

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

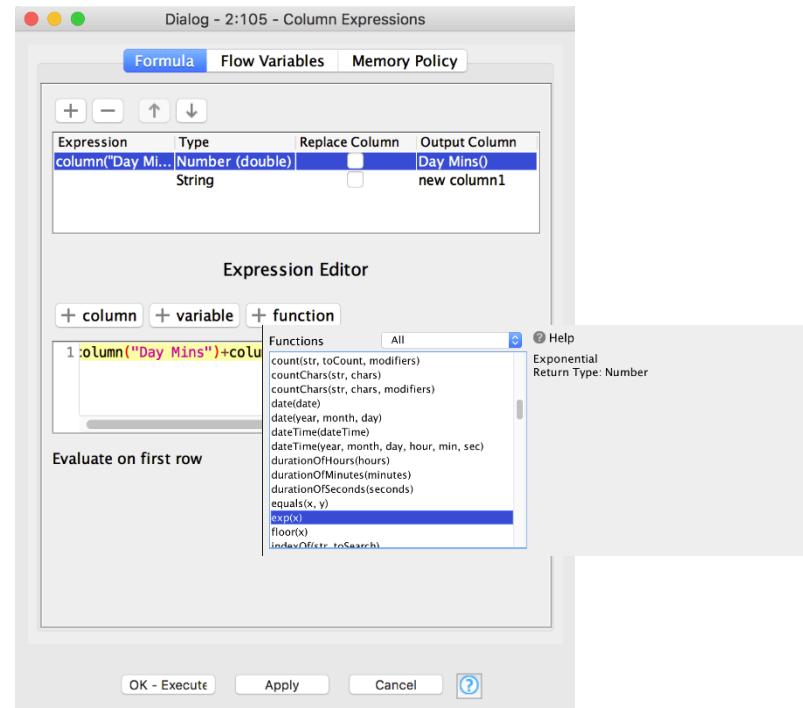
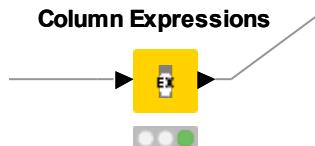


## Duplicate Row Filter



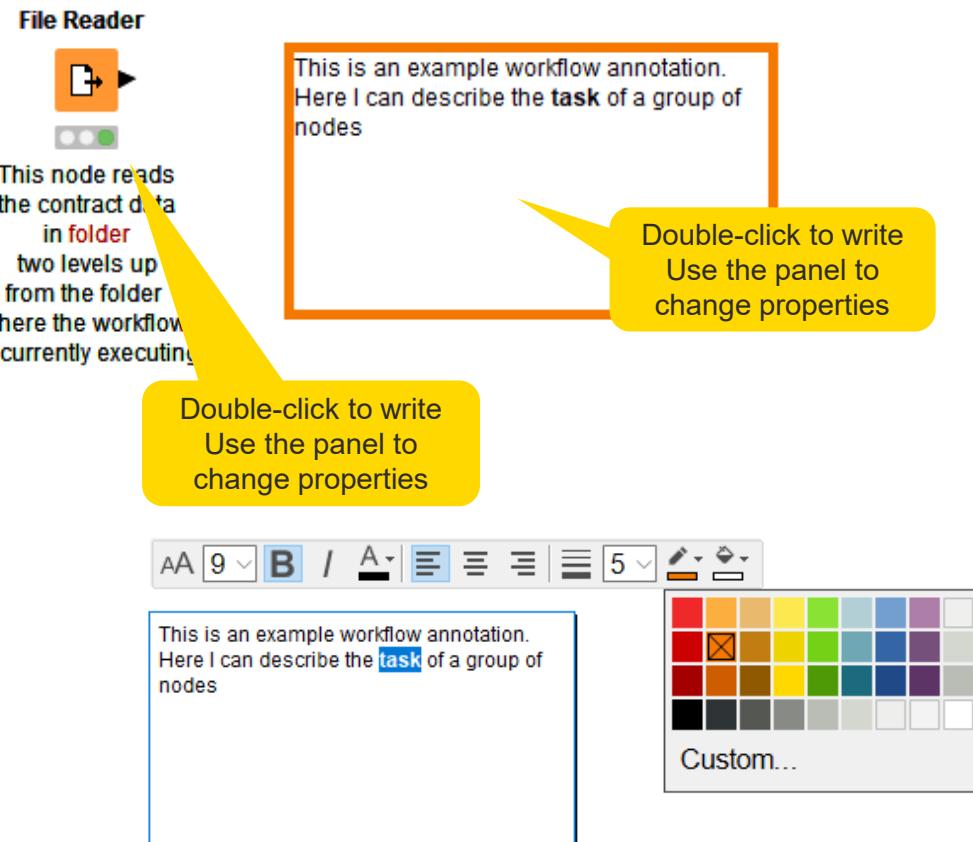
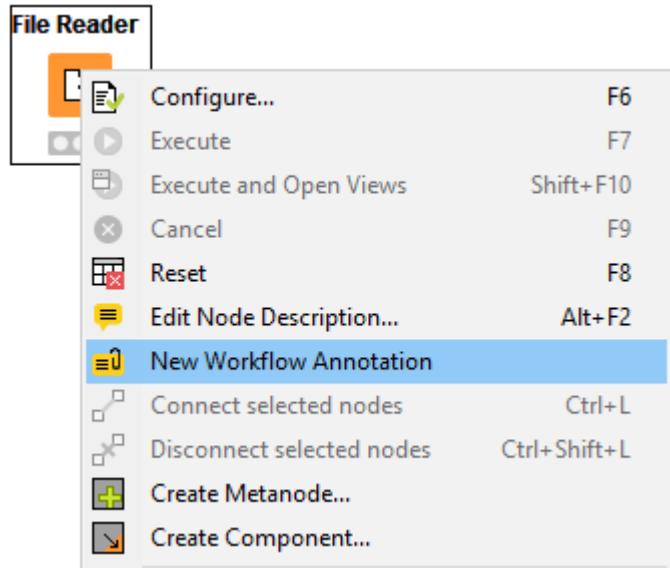
# New Node: Column Expression

- 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**

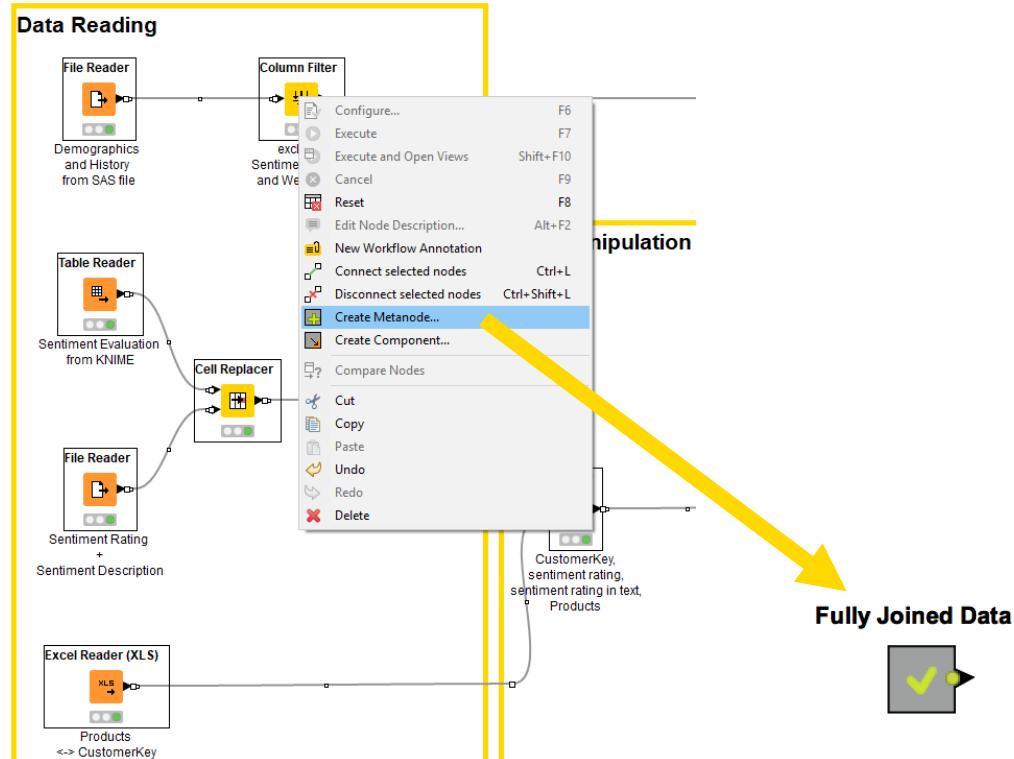
# Comments & Annotations



YouTube KNIME TV Channel:  
[https://youtu.be/AHURYB\\_O8sA](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

The screenshot shows the KNIME Hub interface. At the top, there is a search bar with the text "column filter". Below the search bar, the text "36 results" is displayed. A navigation bar below the search bar includes tabs for "All", "Nodes", "Components" (which is highlighted with a yellow oval), "Workflows", and "Extensions".  
  
The search results list two components:

- Interactive Column Filter**: This component creates an interactive view to filter and select columns for your model. It is developed by paclotamag > Public. A "Component" button is shown to its right.
- Column Filter (by Index)**: This component allows to filter columns by their index (i.e. position). An example can be: select the second and the last columns (e.g. with "2,-1"). Configuration takes a string of column positions... It is developed by knime > Examples > 00\_Components > Data Manipulation > Column Filter (by Index). A "Component" button is shown to its right.

A large yellow callout bubble points from the text "Drag and drop from the KNIME Hub to your workflow" towards the "Column Filter (by Index)" component card. The component card itself also has a yellow border around its bottom section, which contains detailed configuration information and examples.

# KNIME WorkflowDiff

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

The screenshot shows the KNIME Node Comparison interface. At the top, there are two 'Column Filter' nodes. Below them, two tables show the configuration details for each node.

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

The screenshot shows the KNIME Workflow Comparison interface. It displays a tree view of nodes from two workflows: LOCAL/0:2\_Sentiment\_Classification and LOCAL/0:3\_Sentiment\_Classification\_v2. The tree includes various nodes like Decision Tree Learner, Document vector, Extract Table Dimension, Term to String, Category to class, and many others. Below the tree, a 'Node Settings Comparison' table is shown for the 'Snowball Stemmer' node.

Name	Type	Value	Name	Type	Value
Snowball Stemmer (34)			Snowball Stemmer (34)		
Node Settings	sub-config		Node Settings	sub-config	
Document Column_Interval	sub-config		Document Column	sub-config	
Document Column	string	Preprocessed Document	Document Column	string	Preprocessed Document
> Preprocess Unmodifiable_Interval	sub-config		> Preprocess Unmodifiable_Interval	sub-config	
Preprocess Unmodifiable	boolean	false	Preprocess Unmodifiable	boolean	false
> Replace Document_Interval	sub-config		> Replace Document_Interval	sub-config	
Replace Document	boolean	true	Replace Document	boolean	true
> New Document Column_Nam_Interval	sub-config		> New Document Column_Nam_Interval	sub-config	
New Document Column_Nam	string	Preprocessed Document	New Document Column_Nam	string	Preprocessed Document
> Stemmer Name_Interval	sub-config		> Stemmer Name_Interval	sub-config	
Stemmer Name	string	Porter	Stemmer Name	string	German
> System Node Settings	sub-config		> System Node Settings	sub-config	

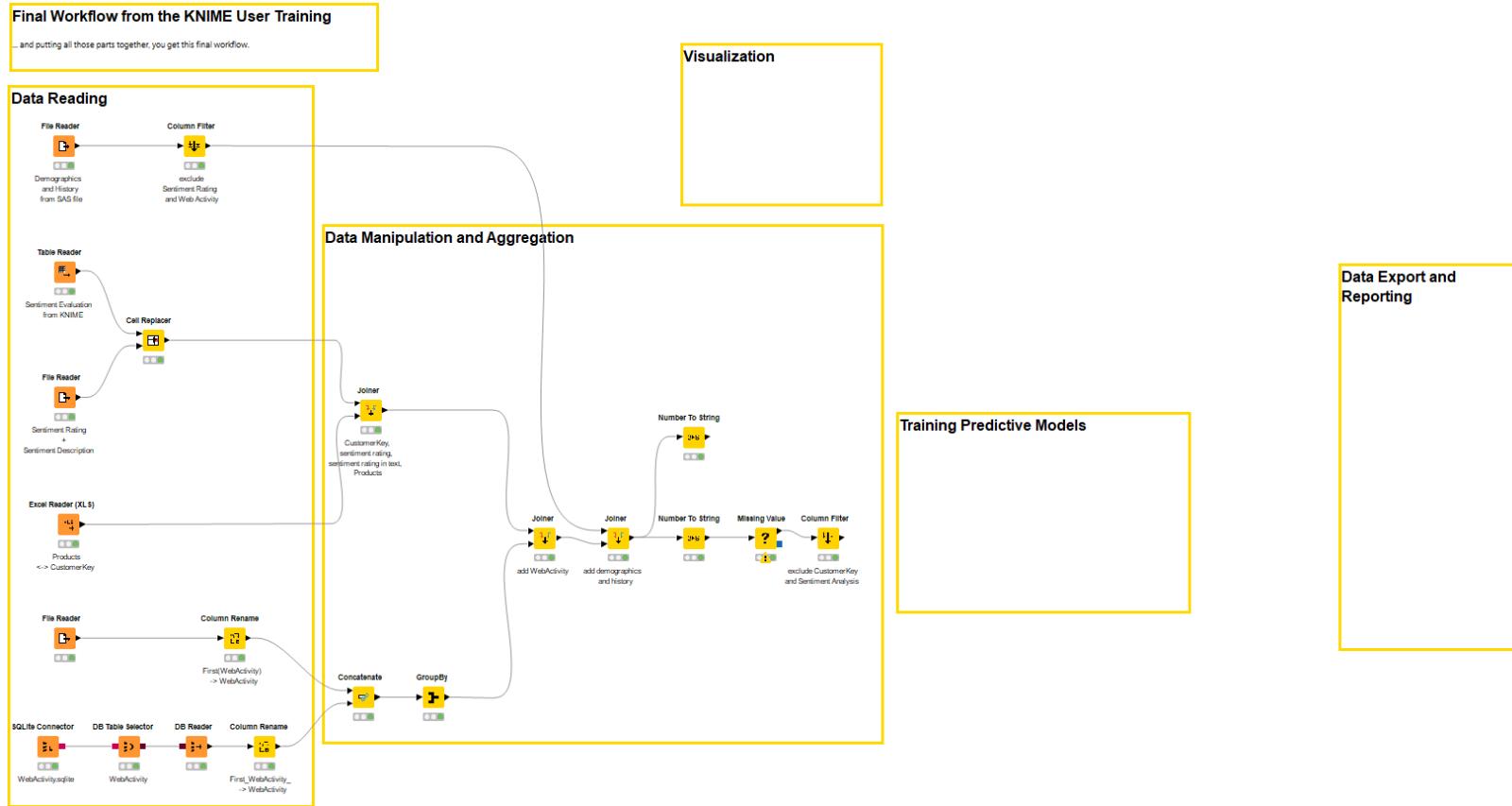
# Data Manipulation Exercise, Activity II

---

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

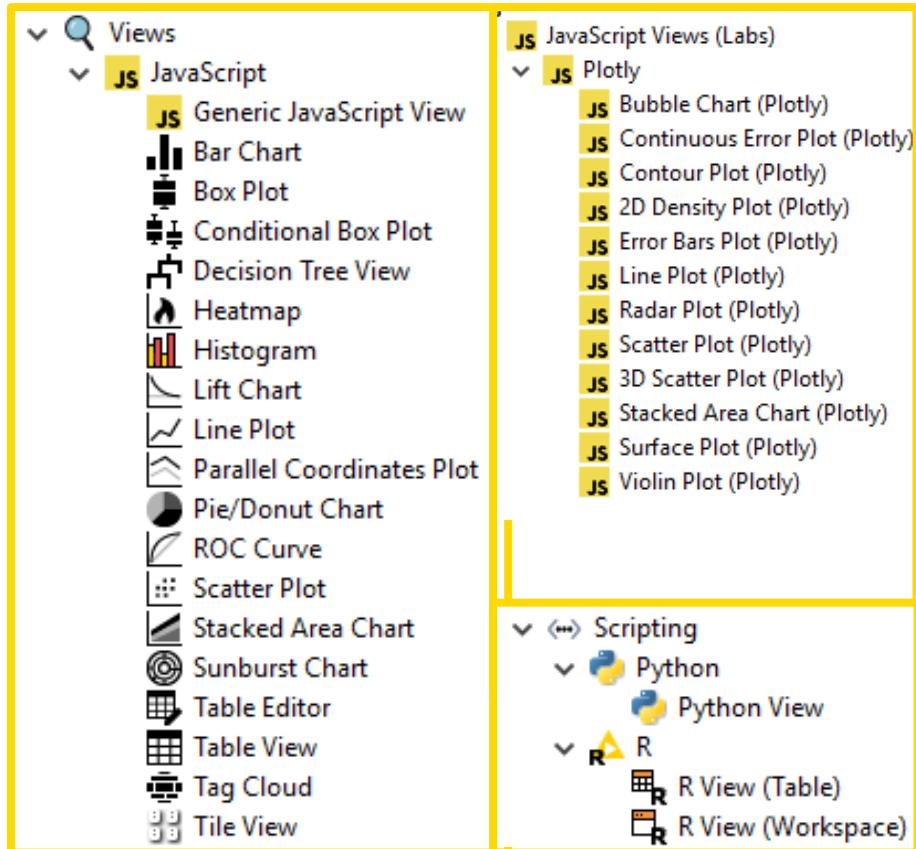


# **Data Visualization**

## **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

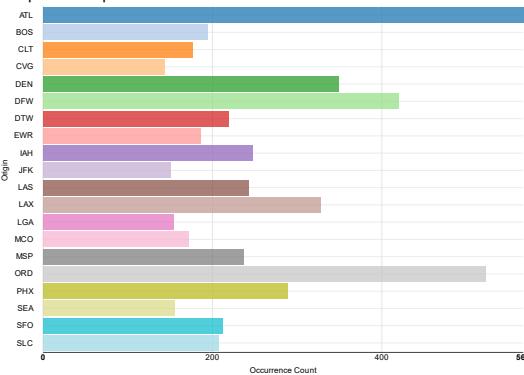


# Visualizations Using One Column

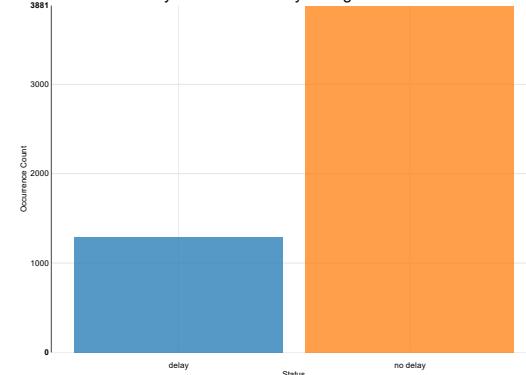
Number of Flights by Date



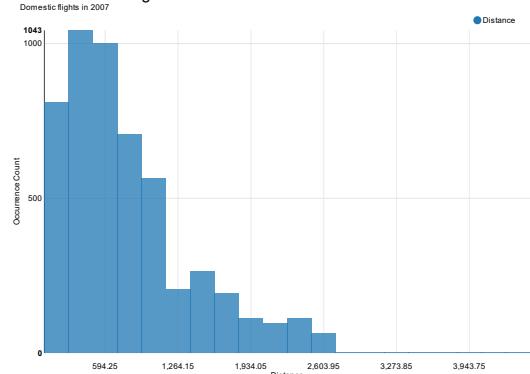
Departure Airports



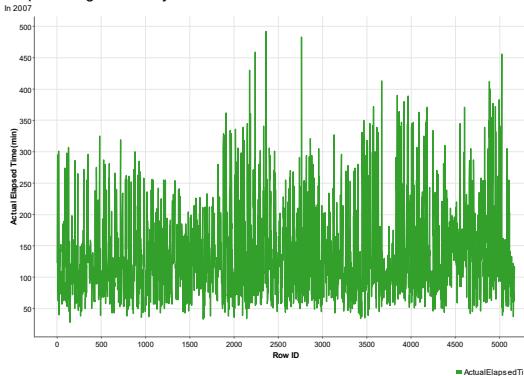
Distribution of Delayed and Non-Delayed Flights



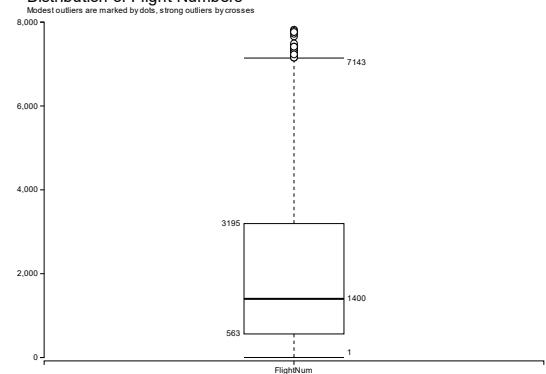
Distribution of Flight Distances



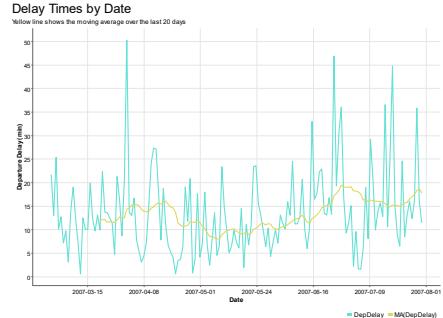
Elapsed Flight Time by Row ID



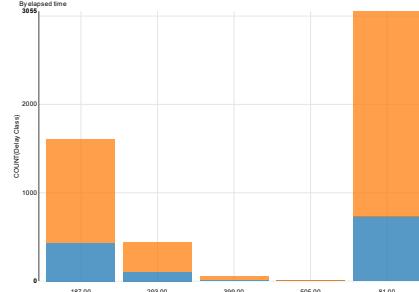
Distribution of Flight Numbers



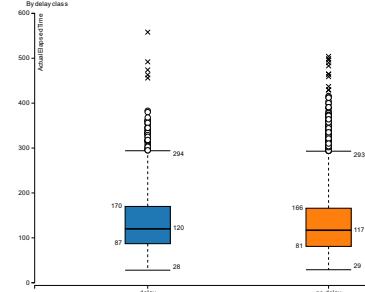
# Visualizations Using Two Columns



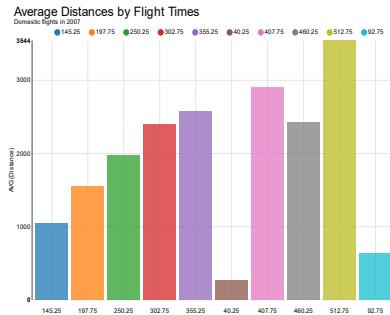
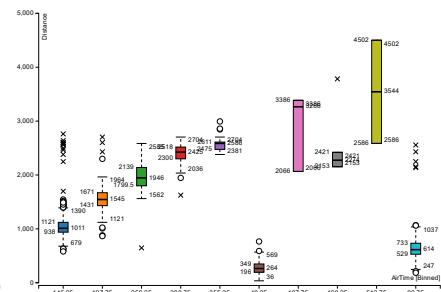
Number of Delayed and Non-Delayed Flights



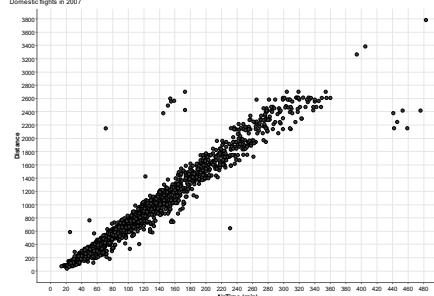
Distribution of Elapsed Time



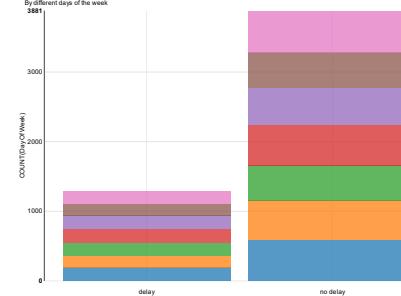
Distribution of Flight Distances by Flight Times



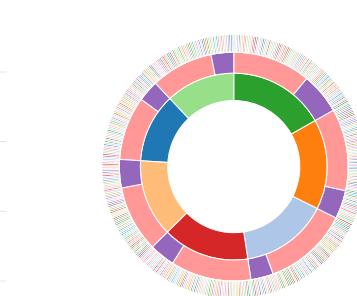
Correlation between Flight Time and Distance



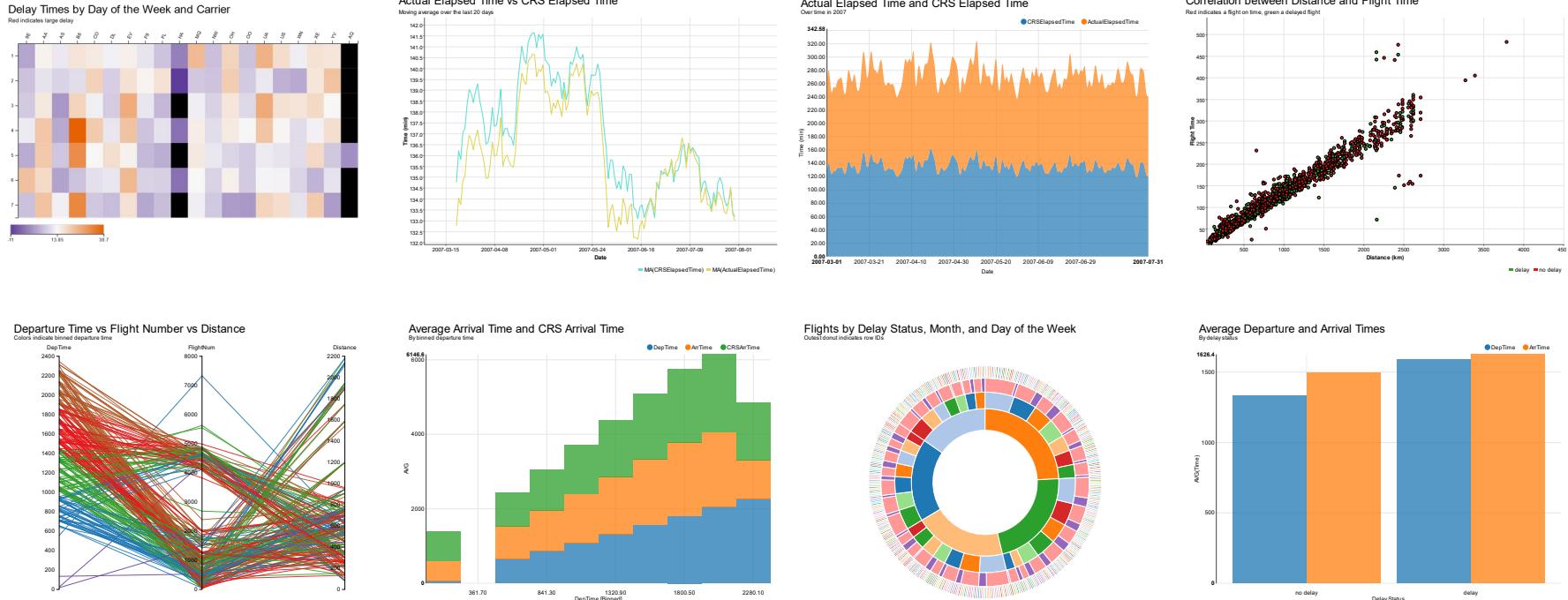
Number of Delayed and Non-Delayed Flights



Flights by Day of the Week and Delay Status

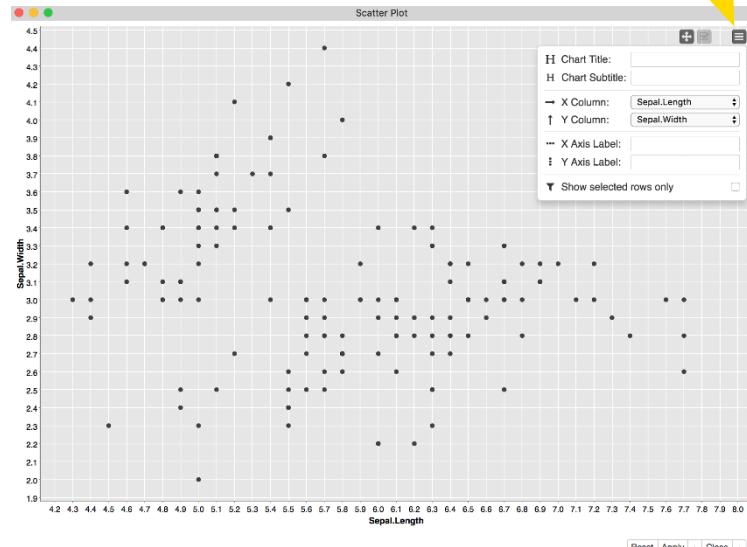
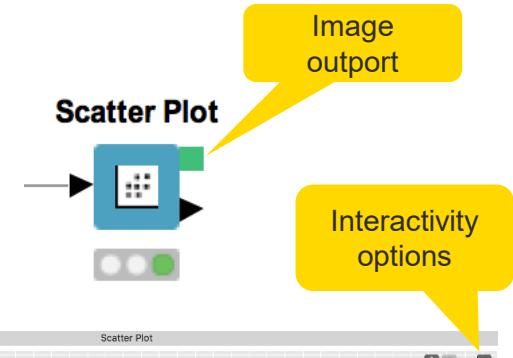


# Visualizations Using Three Columns



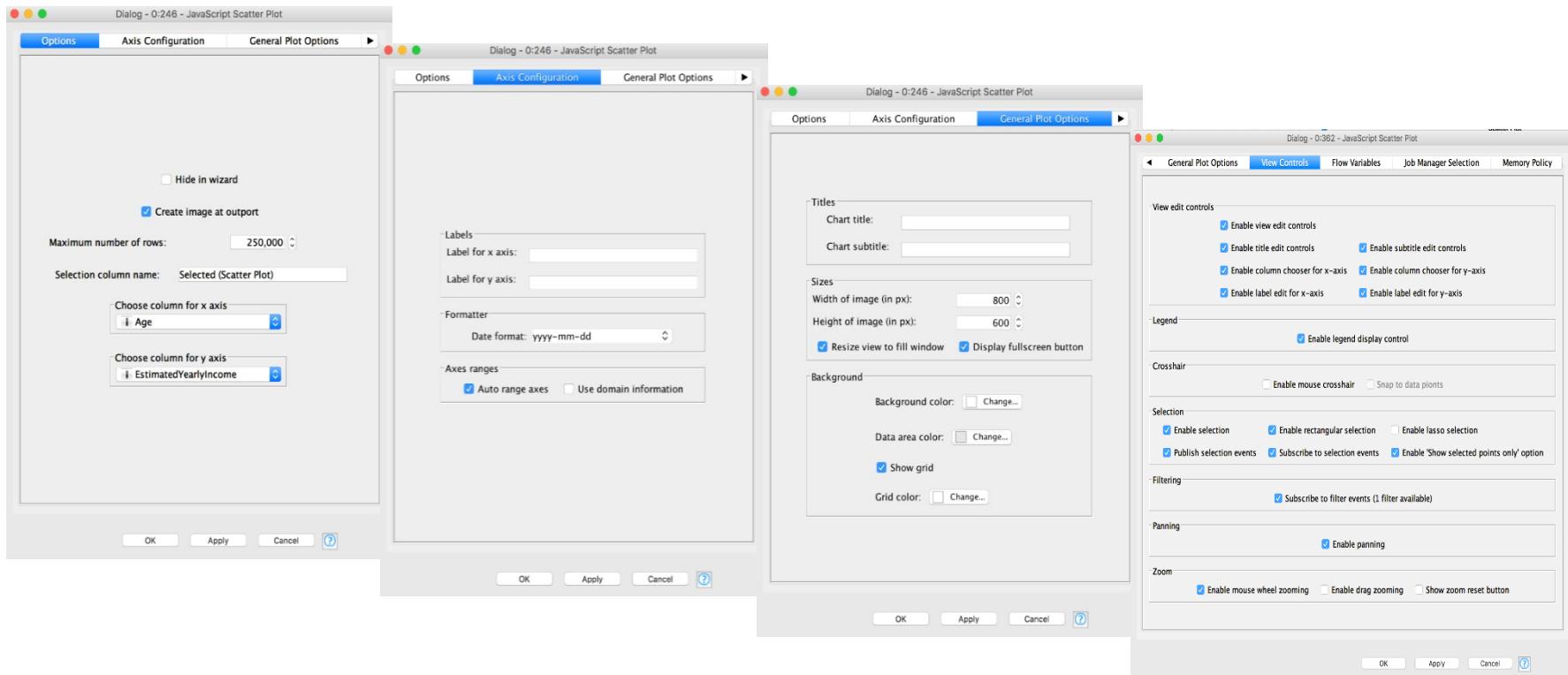
# New Node: Scatter Plot

- 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



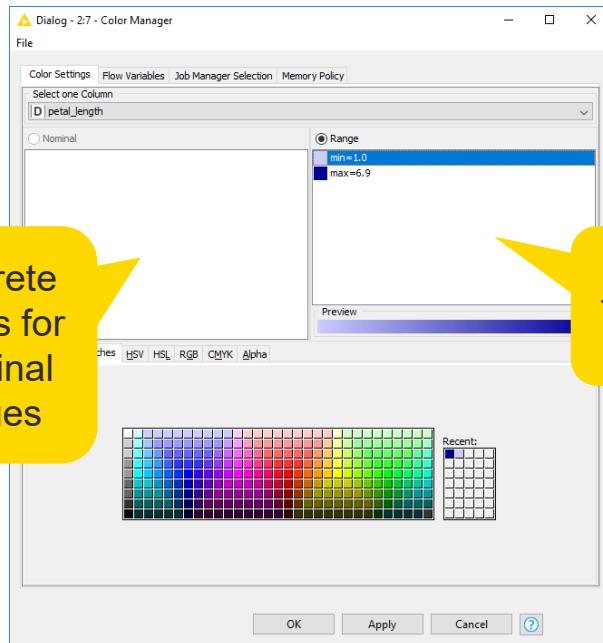
# New Node: Scatter Plot

- Four configuration tabs



# New Node: Color Manager

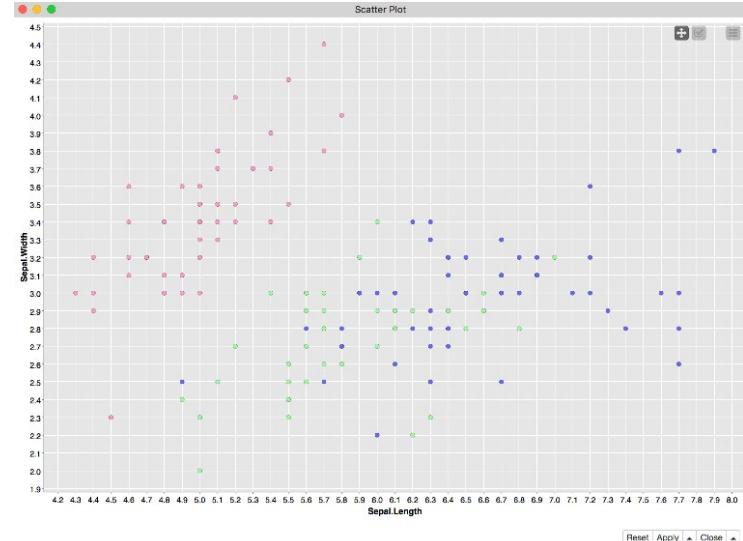
- Color by nominal or continuous values
- Sync colors between views using the color model port and Color Appender node



Discrete colors for nominal values

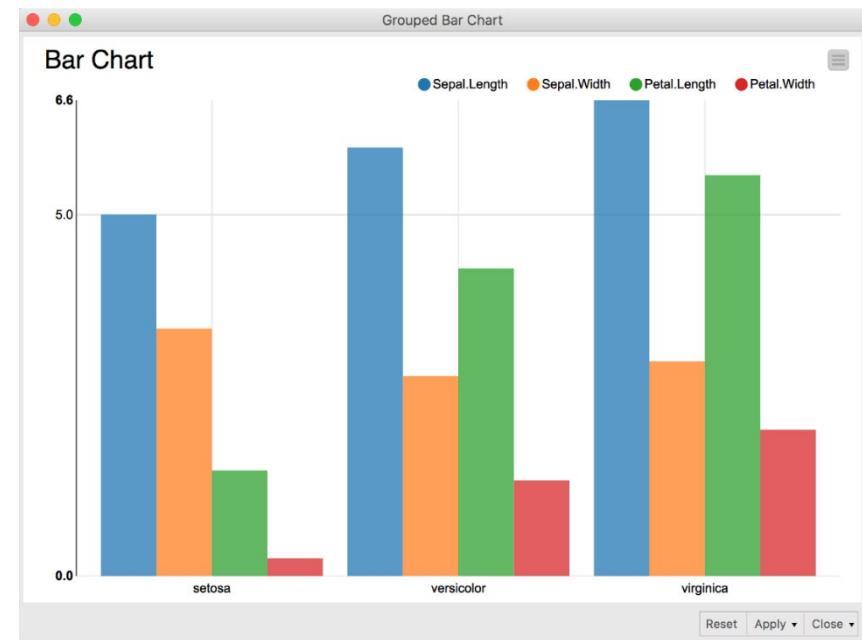
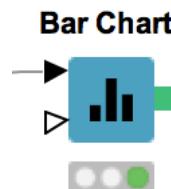


Color range for numerical values



# New Node: Bar Chart

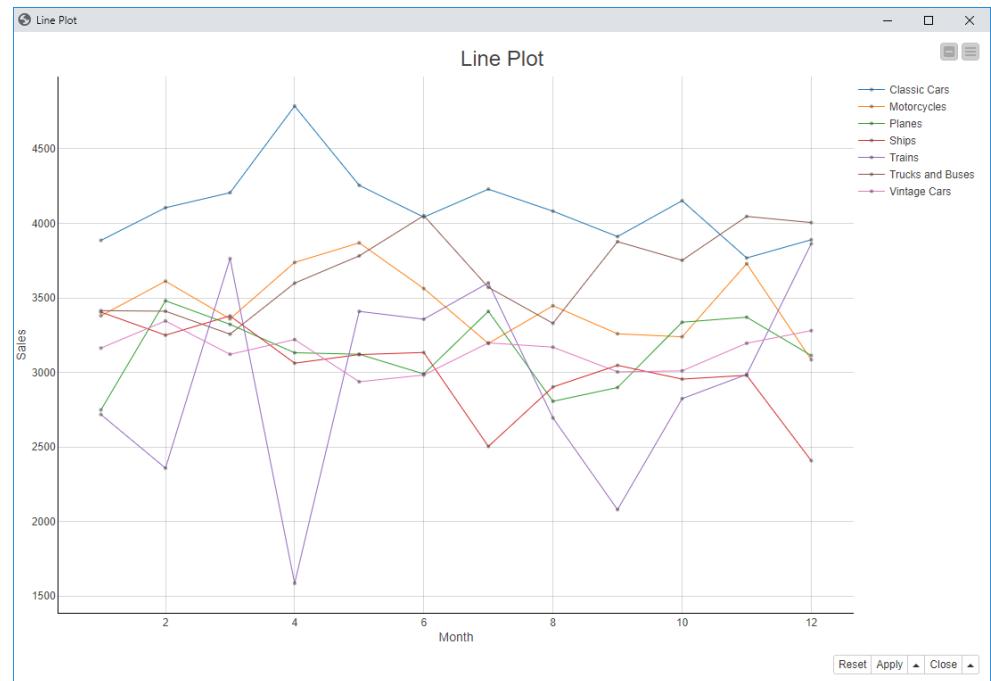
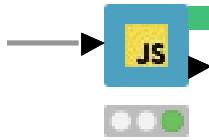
- Show numerical values across categories
- Vertical or horizontal bars
- Bars can be grouped or stacked



# New Node: Line Plot

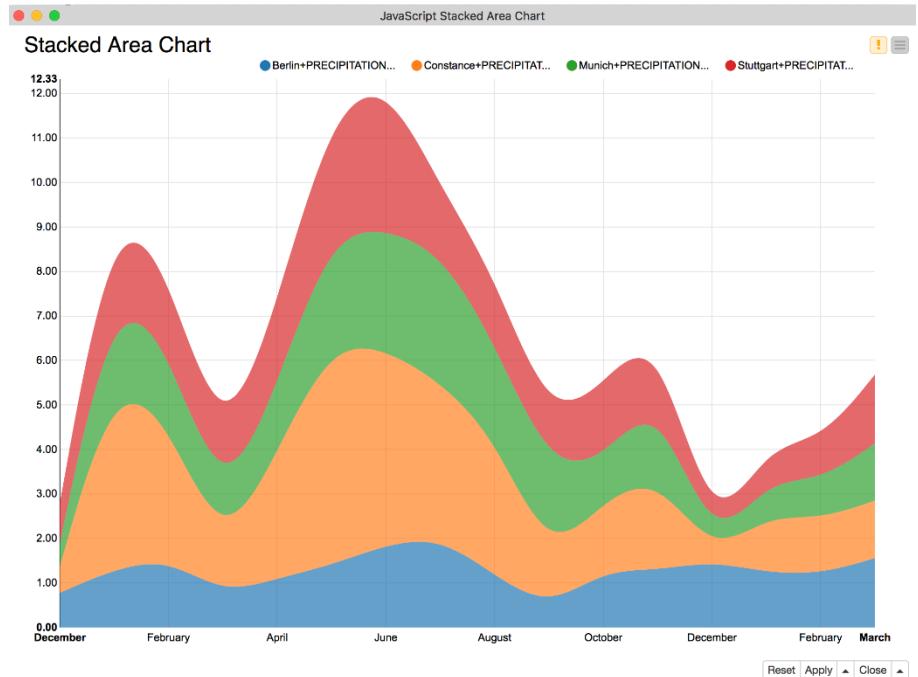
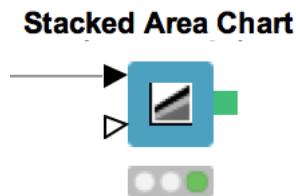
- Plot sequence of values, e.g. over time
- Useful to identify trends, also between groups

Line Plot (Plotly)



# New Node: Stacked Area Chart

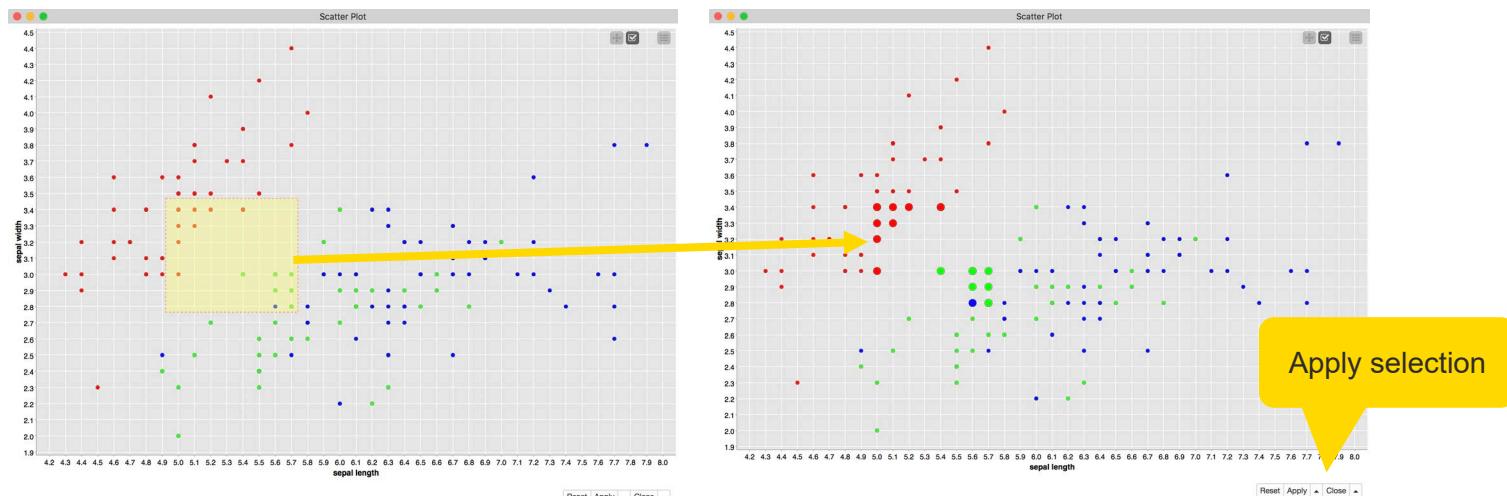
- Visualizes numerical values from multiple columns as stacked areas
- Great for plotting distributions over time



# Selection & Filtering in JavaScript Views

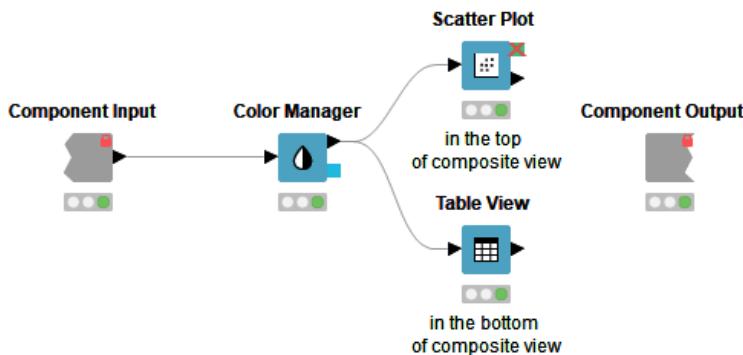
Interactivity allows you to select data points in views

- 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

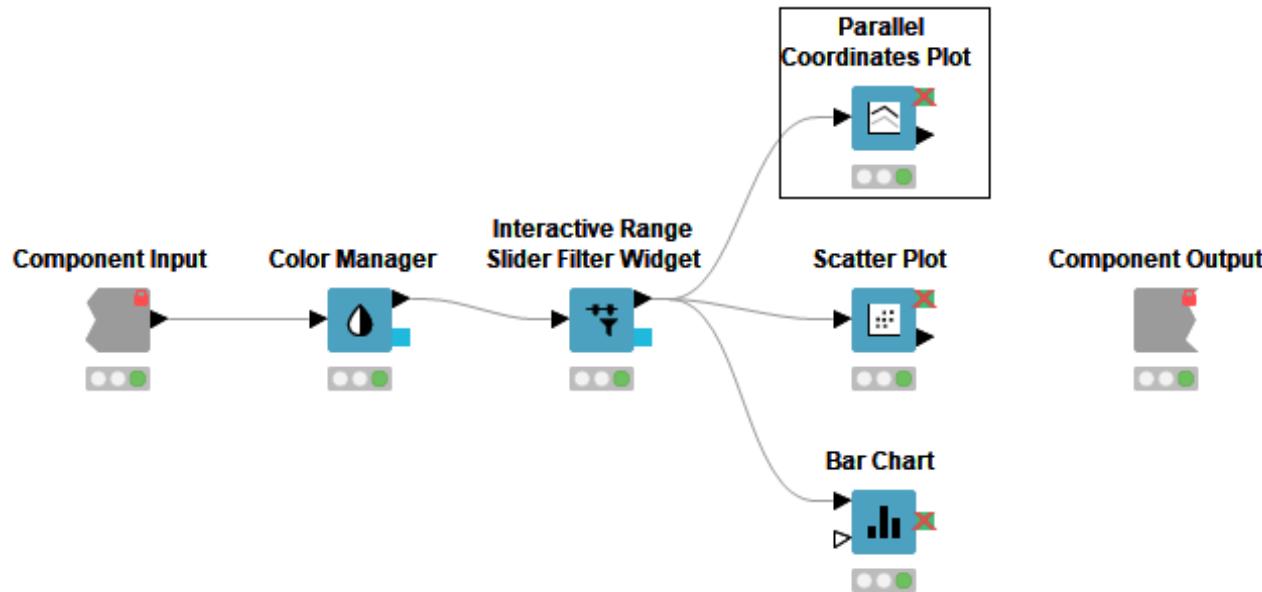


# Components – Combined Views

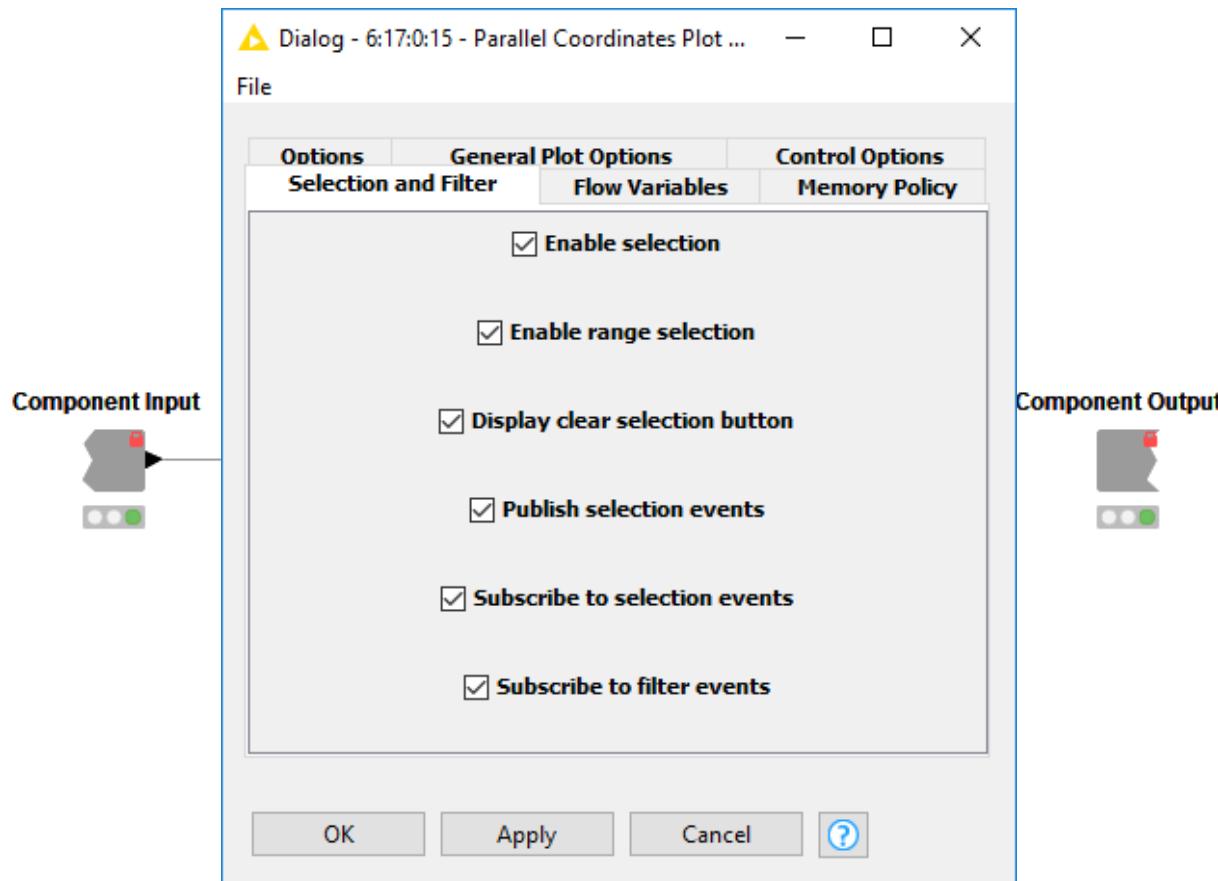
- Multiple JavaScript View nodes can be combined in Components
- Selections are transmitted to all other views
- Also for use on the KNIME



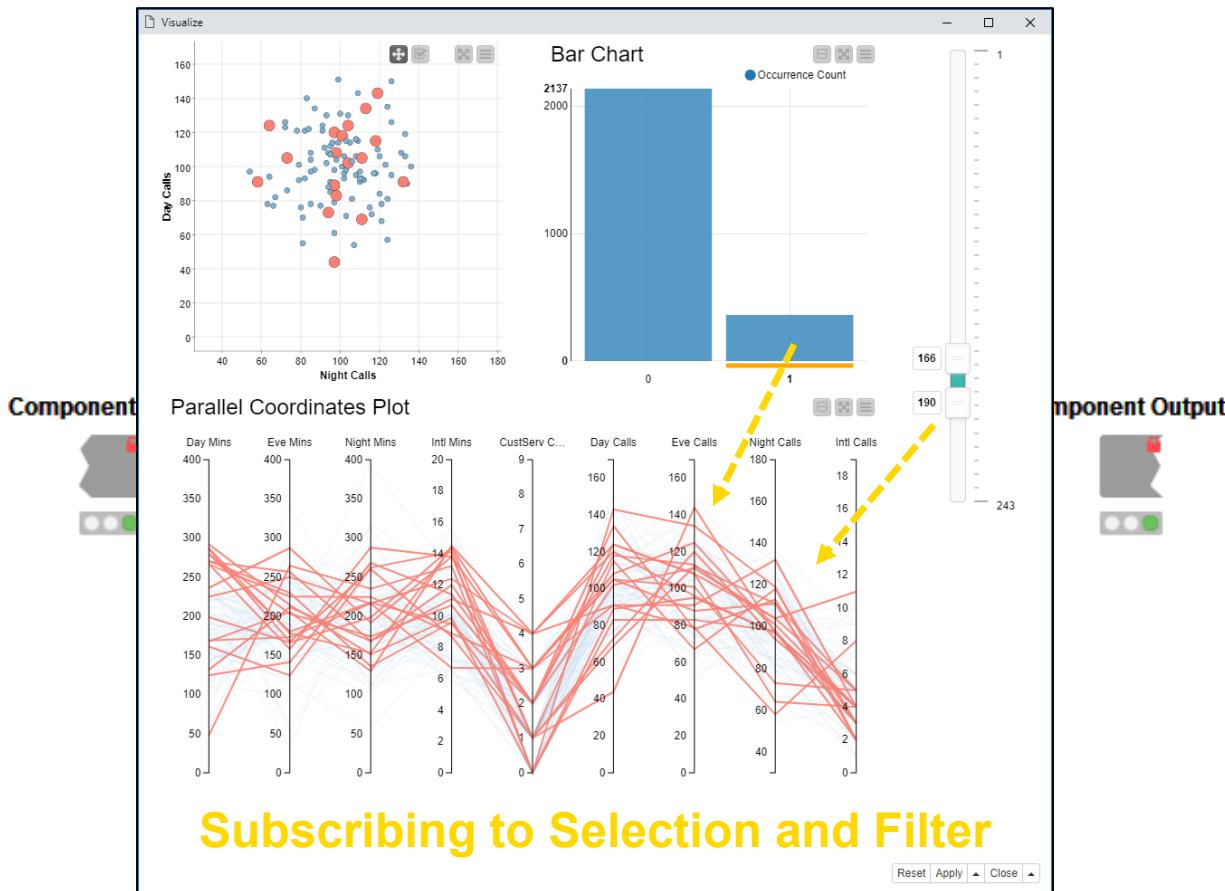
# Interactivity across Charts: Selection and Filter Events



# Interactivity across Charts: Selection and Filter Events



# Interactivity across Charts: Selection and Filter Events

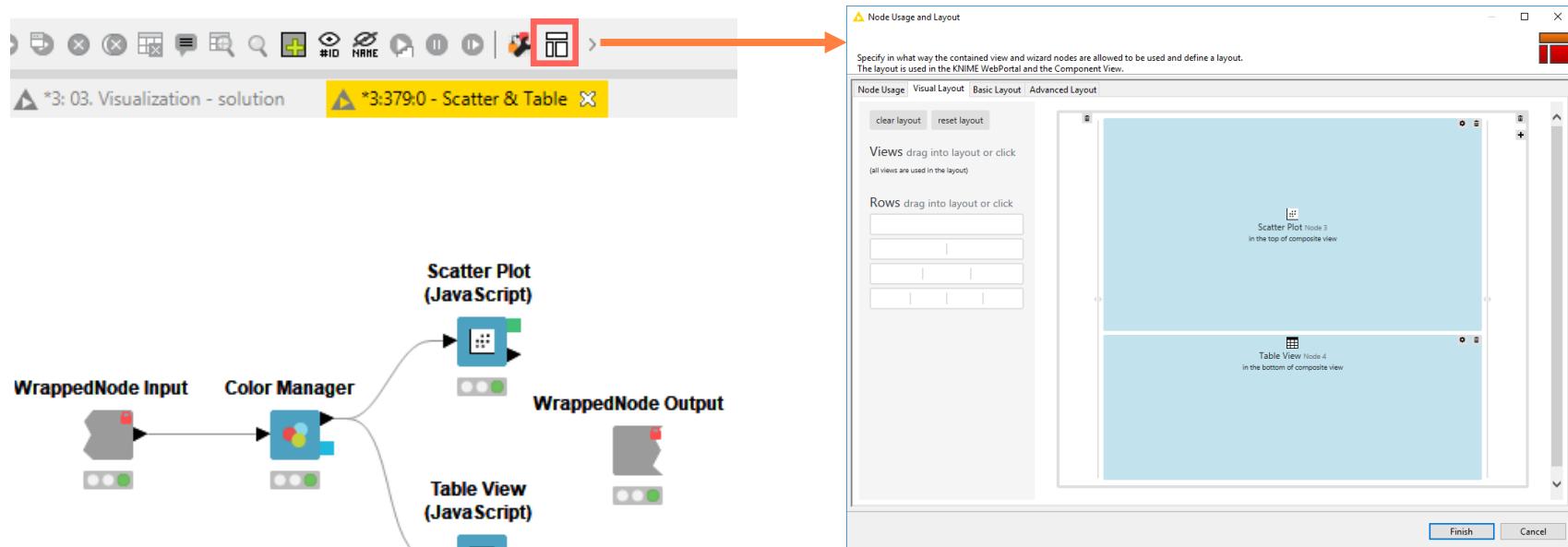


# Interactivity across Charts: Selection and Filter Events



# 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)

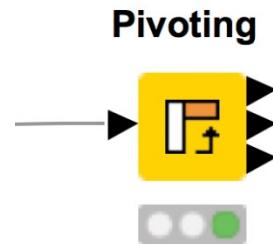
Category	Online	Onsite
Clothing	7	1
Home	0	3
Electronics	12	0

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

The image displays four overlapping windows from the KNIME interface, each showing different aspects of the Pivoting node configuration:

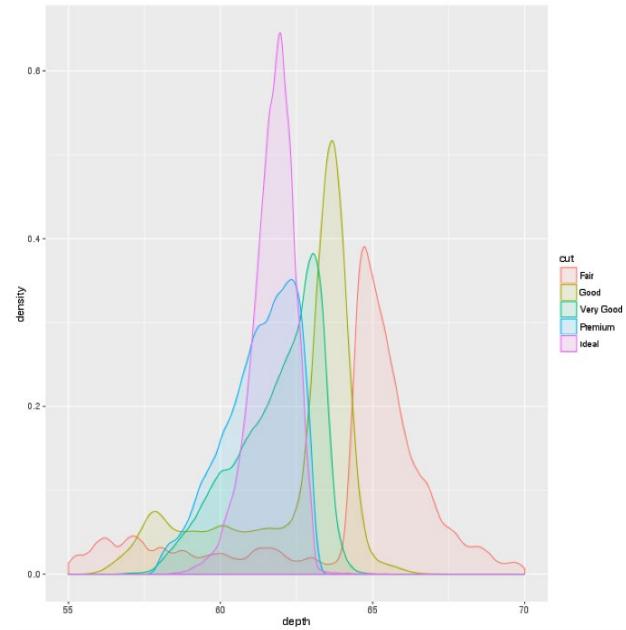
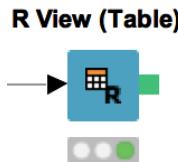
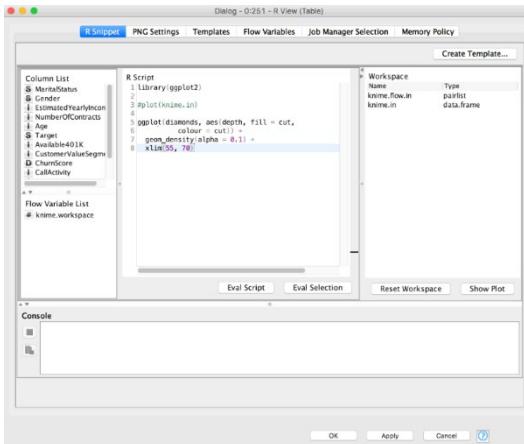
- Groups ~ Rows**: Shows the "Groups" tab with "Available column(s)" ProductID, OrderedItems, and Store.
- Pivots ~ Columns**: Shows the "Pivots" tab with "Available column(s)" ProductID, OrderedItems, and Category.
- Aggregation**: Shows the "Aggregation settings" tab with "Available columns" ProductID and OrderedItems. Under "Select", "OrderedItems" is listed with "Aggregation (click to change)" set to "Sum".
- Pivot table - 0:35 - Pivoting**: A preview window showing a pivot table with three rows (Row0, Row1, Row2) and three columns. The first column is Row ID, the second is Category, and the third is a sum of OrderedItems. The preview shows:

Row ID	Category	Value
Row0	Clothing	1823
Row1	Electronics	10754
Row2	Home	7180

A yellow arrow points from the "Aggregation" window towards the pivot table preview, indicating the final output of the pivoted data.

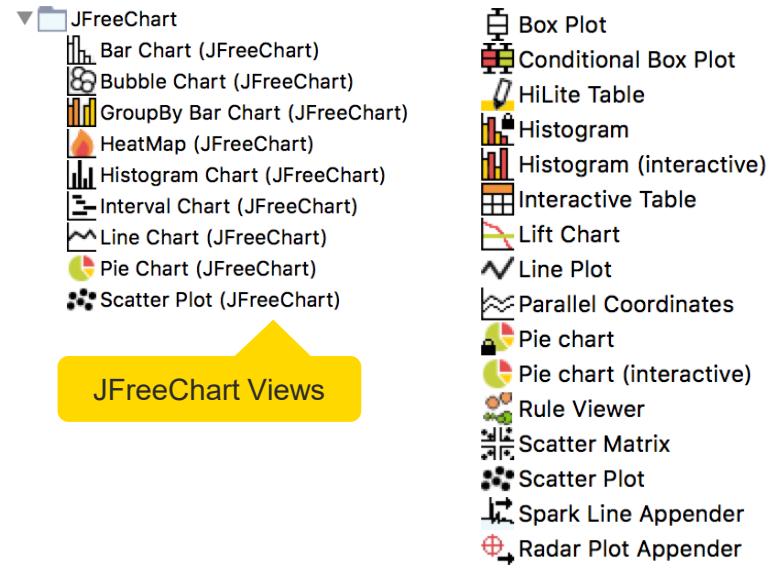
# Script-based View Nodes

- 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



# Legacy View Nodes: JFreeChart & KNIME Views

- KNIME provides three types of visualizations
  - **JavaScript Views**
  - JFreeChart Views
  - Local Views
- Active development only for JavaScript Views -> use those!
- JFreeChart and Local Views still useful when visualizing locally



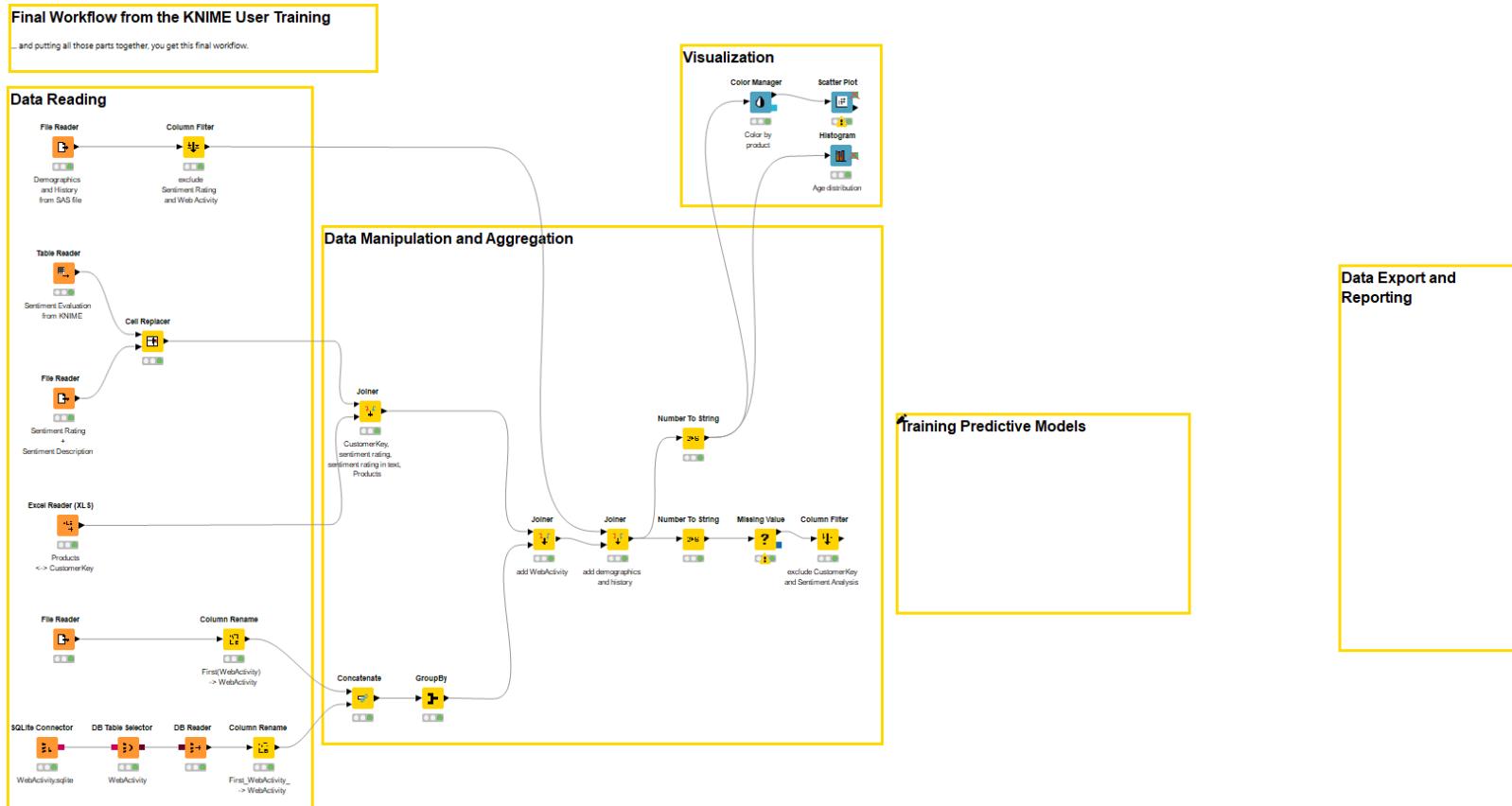
# Visualization Exercise

---

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**

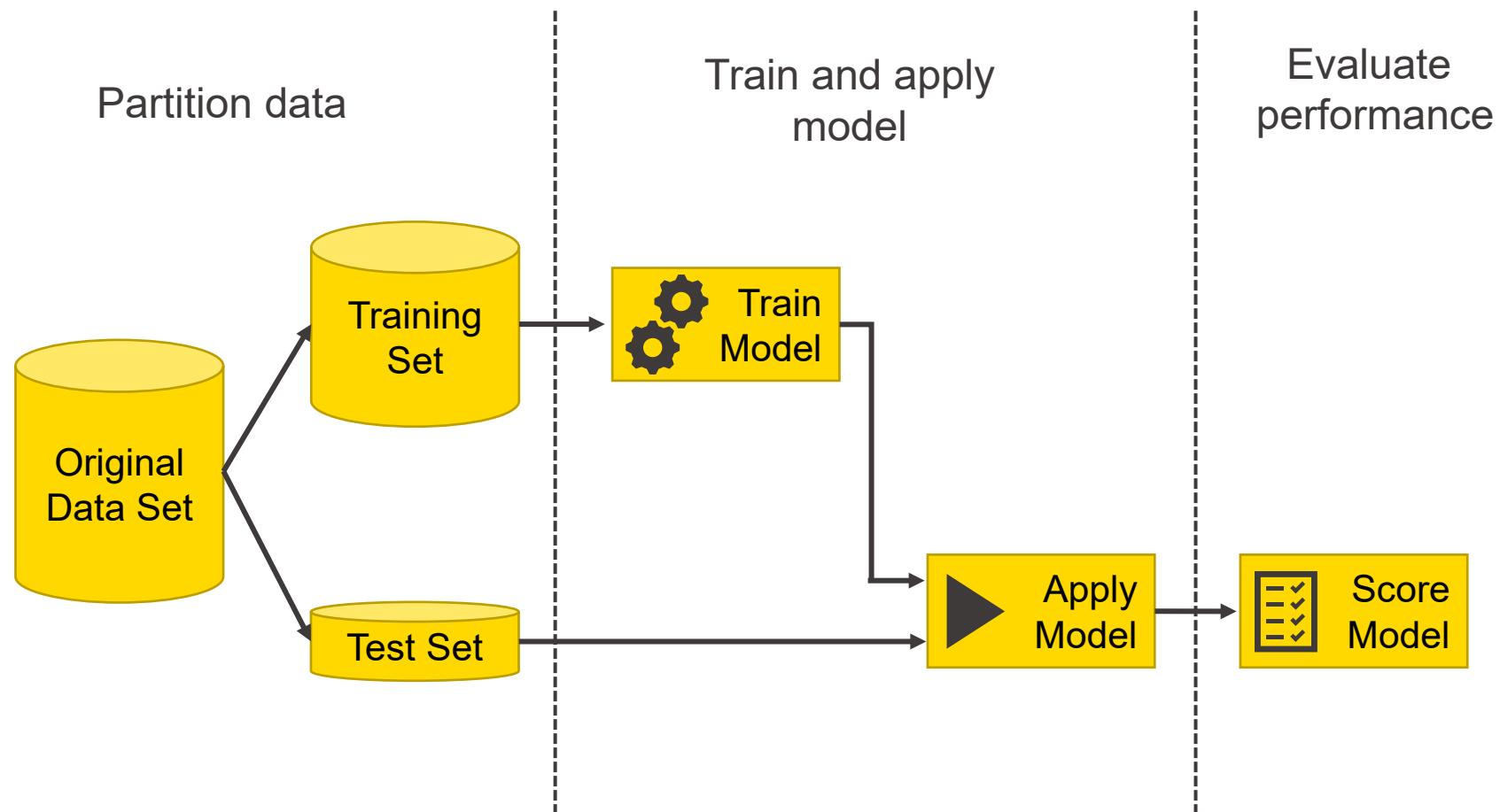
# Data Mining Strategies

---

## Example Applications:

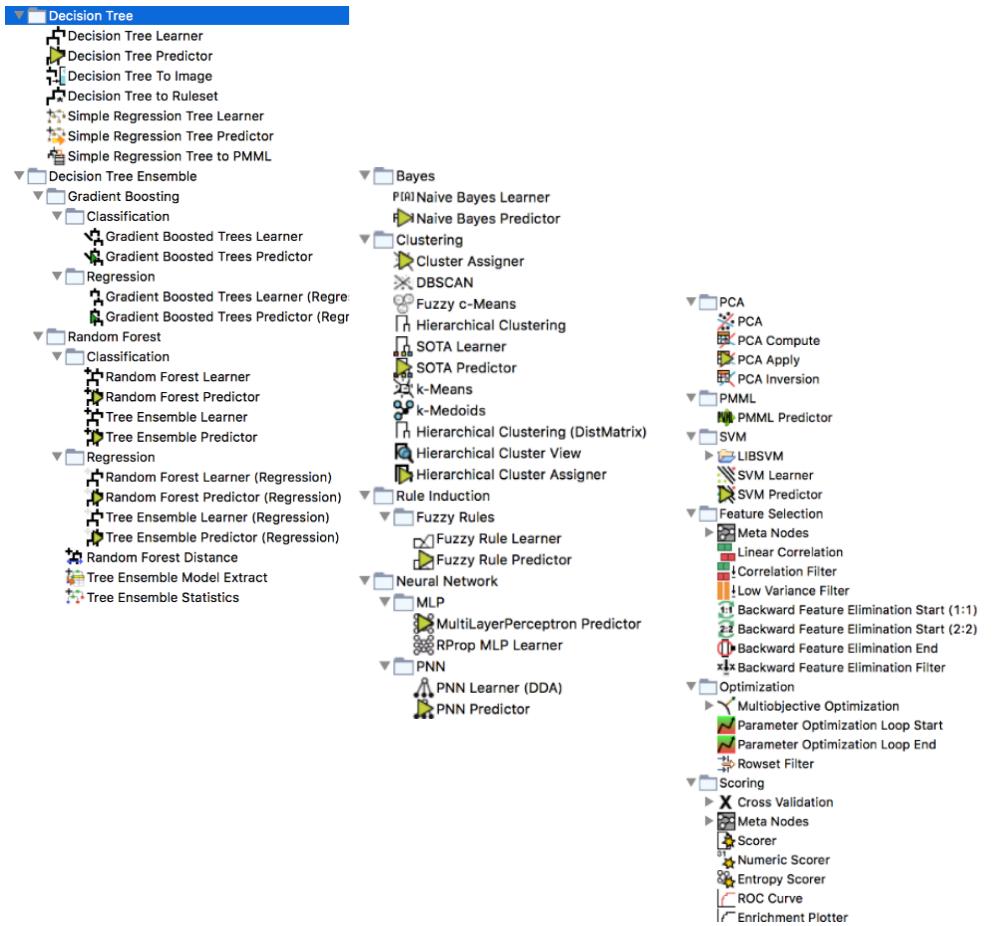
- Anomaly Detection (fraud, predictive maintenance)
- Association Rule Learning (market basket analysis)
- Clustering (market segmentation)
- Classification (next best offer, churn preventions)
- Regression (trend estimation)

# Data Mining: Process Overview



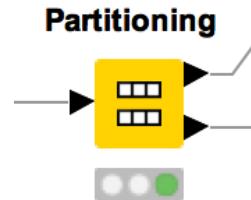
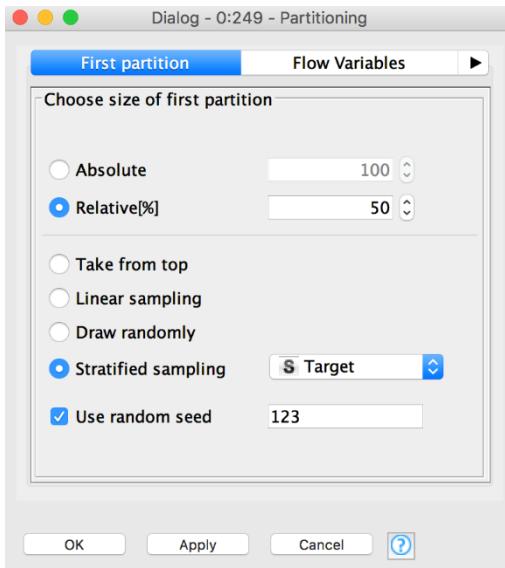
# Data Mining in KNIME

- 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



The screenshot shows the 'First partition (as defined in dialog) - 0:249 - Partitioning' window. It displays a table titled 'Table "default" - Rows: 5775 Spec - Columns: 13'. The columns are Row ID, \$ Marita..., \$ Gender, Estim..., Numb..., and Age. The data includes rows from Row0 to Row20, showing various gender and age distributions.

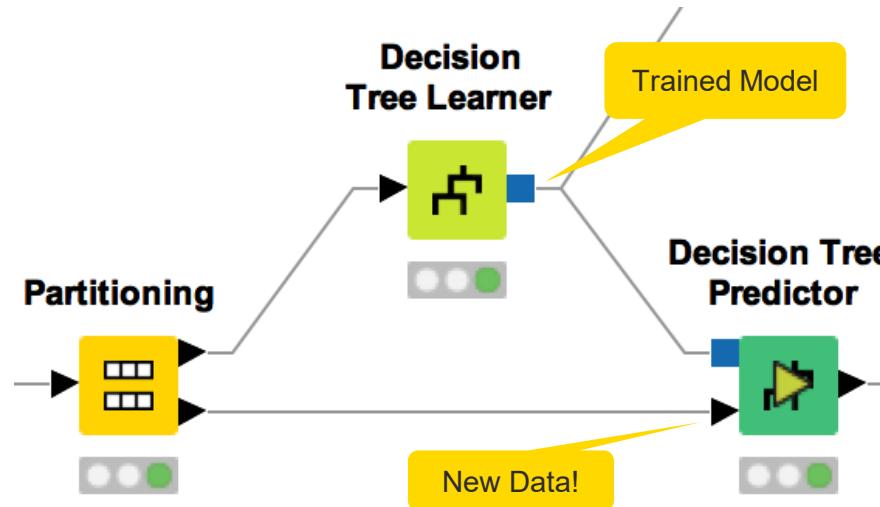
Row ID	\$ Marita...	\$ Gender	Estim...	Numb...	Age
Row0	M	M	90000	0	44
Row7	M	M	60000	2	46
Row9	S	M	70000	1	46
Row10	S	F	70000	1	46
Row13	M	M	100000	3	42
Row14	S	F	100000	3	42
Row15	S	F	30000	1	31
Row17	S	F	20000	2	66
Row18	S	M	30000	2	66
Row20	S	M	40000	2	32
...					

The screenshot shows the 'Second partition (remaining rows) - 0:249 - Partitioning' window. It displays a table titled 'Table "default" - Rows: 5775 Spec - Columns: 13'. The columns are Row ID, \$ Marita..., \$ Gender, Estim..., Numb..., and Age. The data includes rows from Row1 to Row19, showing the remaining data after the first partition was taken.

Row ID	\$ Marita...	\$ Gender	Estim...	Numb...	Age
Row1	S	M	60000	1	45
Row2	M	M	60000	1	45
Row3	S	F	70000	1	42
Row4	S	F	80000	4	42
Row5	S	M	70000	1	45
Row6	S	F	70000	1	44
Row8	S	F	60000	3	46
Row11	M	M	60000	4	46
Row12	M	F	100000	2	42
Row16	M	M	30000	1	31
Row19	S	M	40000	2	32
...					

# 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

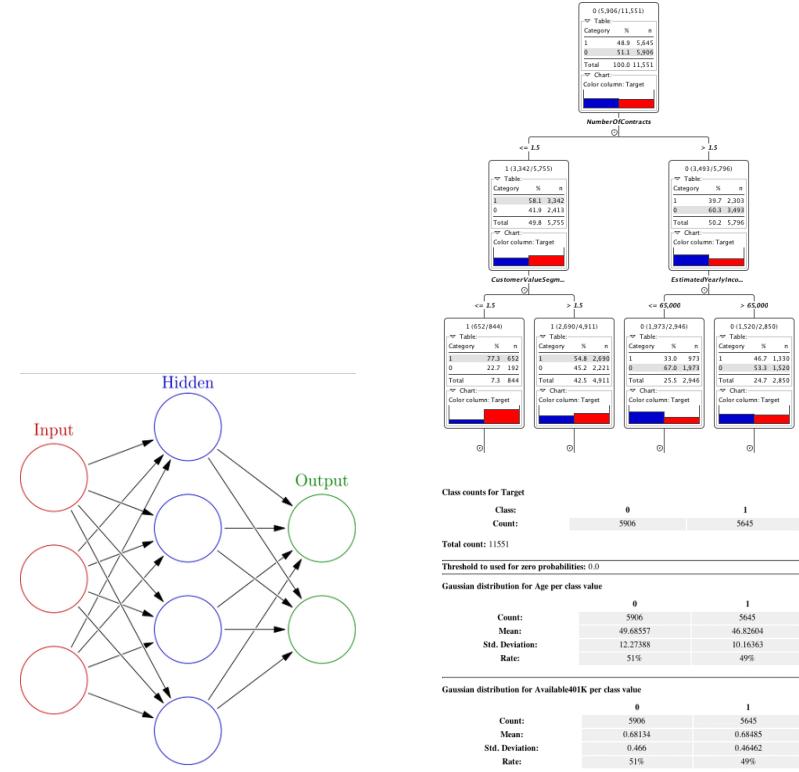
Predict *nominal* outcomes on existing data (supervised)

## ■ 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



# 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

Output data - 0:311 - Column Resorter

File Hilite Navigation View

Table "default" - Rows: 5776 Spec - Columns: 17 Properties Flow Variables

R...	I CustomerKey	S Marital...	S Gender	S Target	S 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

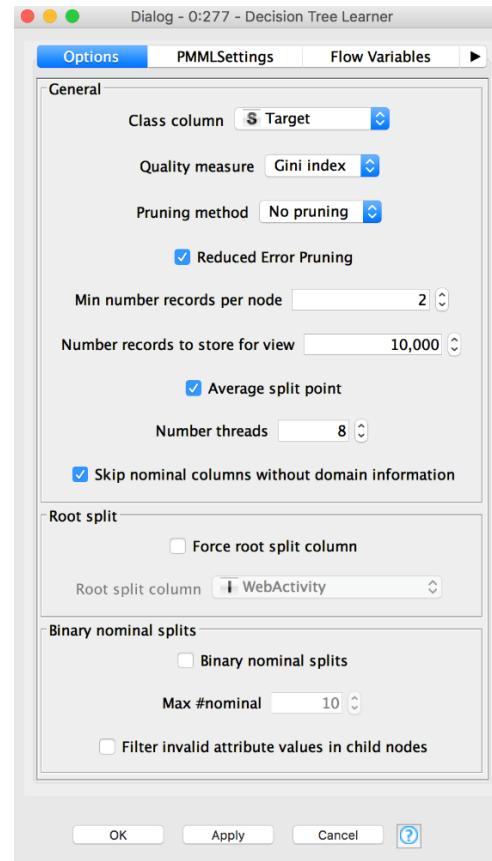
# KNIME's Decision Tree

J.R. Quinlan, “C4.5 Programs for machine learning”

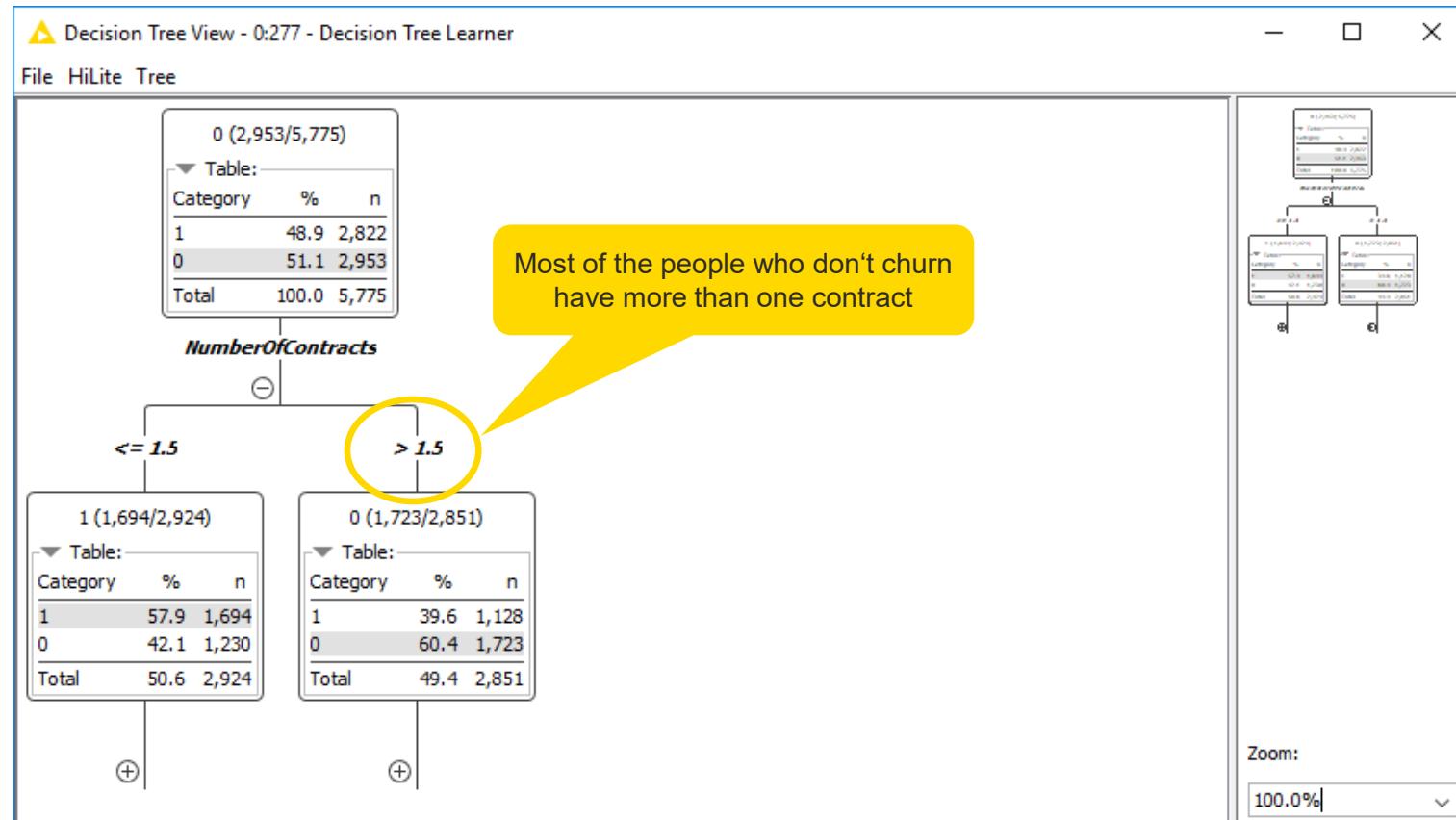
J. Shafer, R. Agrawal, M. Mehta, “SPRINT: A Scalable Parallel Classifier for Data Mining”

- 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

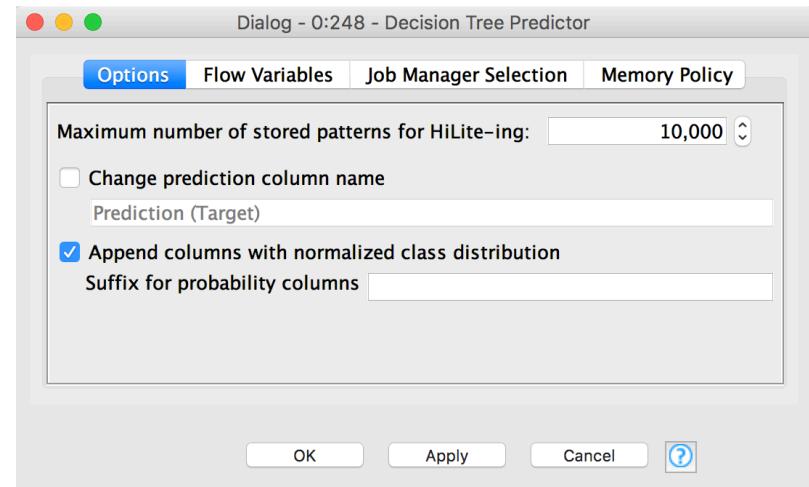
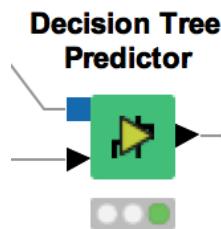


# Decision Tree View



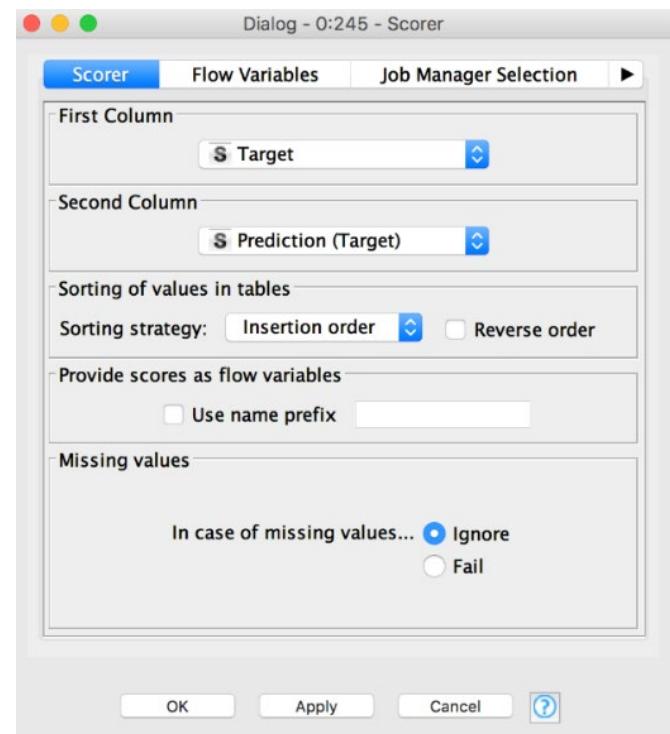
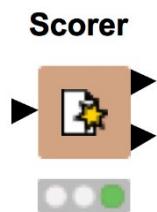
# New Node: Decision Tree Predictor

- 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

- An accuracy statistics table provides a detailed analysis of model quality

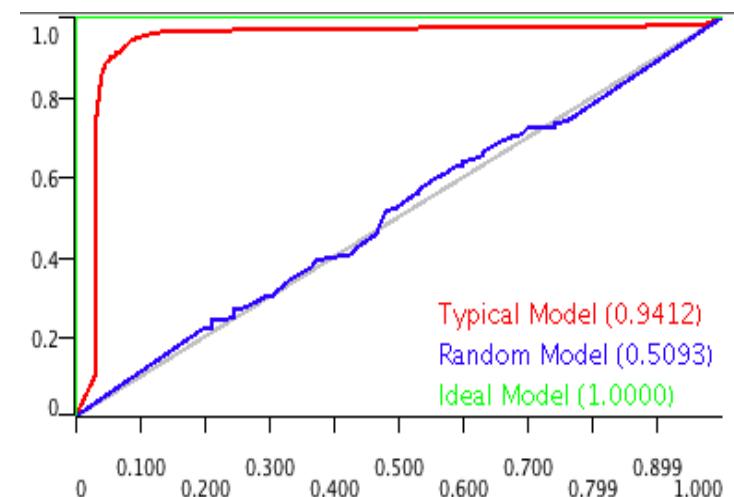
Accuracy statistics - 0:297 - Scorer												
File	Hilite	Navigation	View	Table "default" – Rows: 3			Spec – Columns: 11	Properties		Flow Variables		
Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa	
1	2073	759	2193	750	0.734	0.732	0.734	0.743	0.733	?	?	
0	2193	750	2073	759	0.743	0.745	0.743	0.734	0.744	?	?	
Overall	?	?	?	?	?	?	?	?	?	0.739	0.477	

# Confusion Matrix

	Predicted class <b>POSITIVE</b> (churn)	Predicted class <b>NEGATIVE</b> (no churn)
Actual class <b>POSITIVE</b> (churn)	TRUE POSITIVE (TP) 2073	FALSE NEGATIVE (FN) 750
Actual class <b>NEGATIVE</b> (no churn)	FALSE POSITIVE (FP) 759	TRUE NEGATIVE (TN) 2193

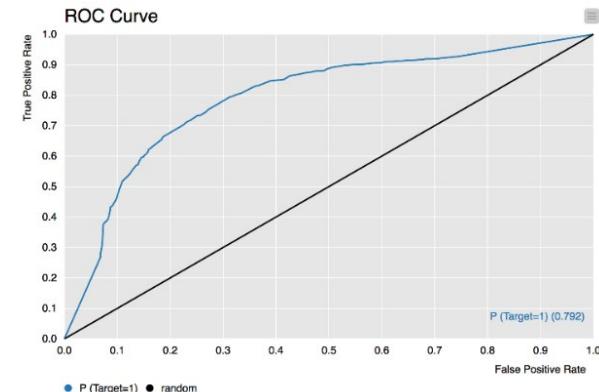
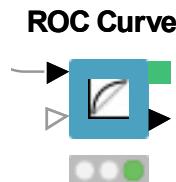
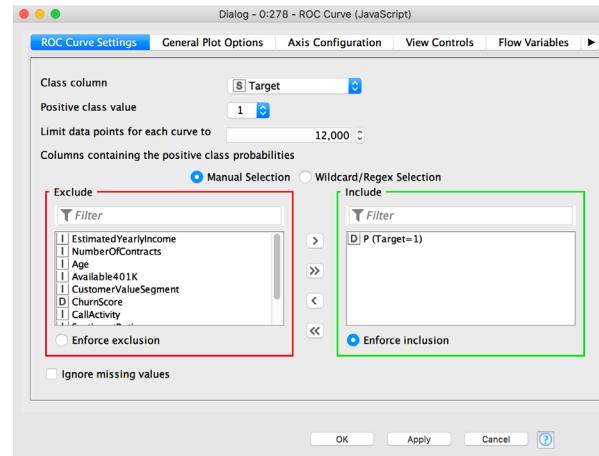
# Receiver Operating Characteristics

- 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:
  - Original class column
  - Positive class value
  - Probability for the selected positive class value for one or multiple models



# Data Mining Exercise, Activity I

---

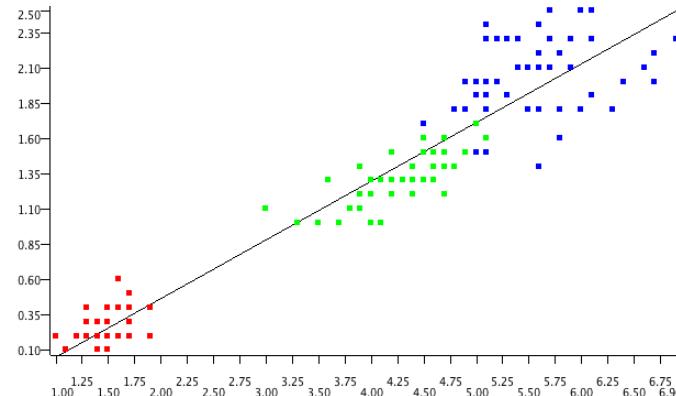
Start with exercise: *Data Mining, Activity I*:

- 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)

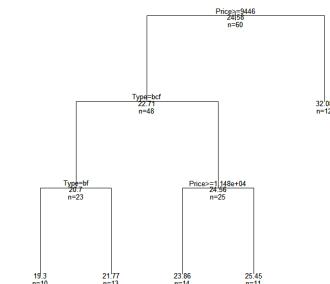
- Applications
  - Forecasting
  - Quantitative Analysis
  
- Methods
  - Linear
  - Polynomial
  - Regression Trees
  - Partial Least Squares



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

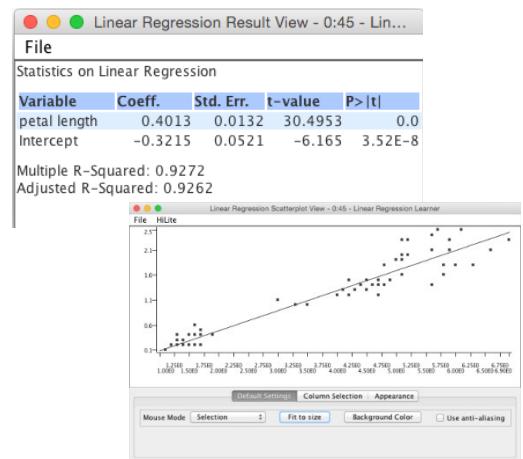
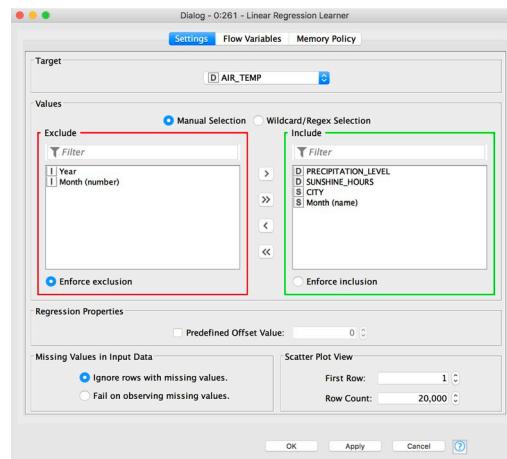
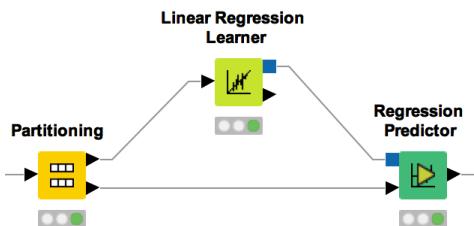
Multiple R-Squared: 0.9271  
Adjusted R-Squared: 0.9266



# New Nodes: Linear Regression Learner & Regression Predictor

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

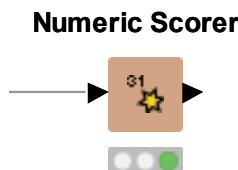
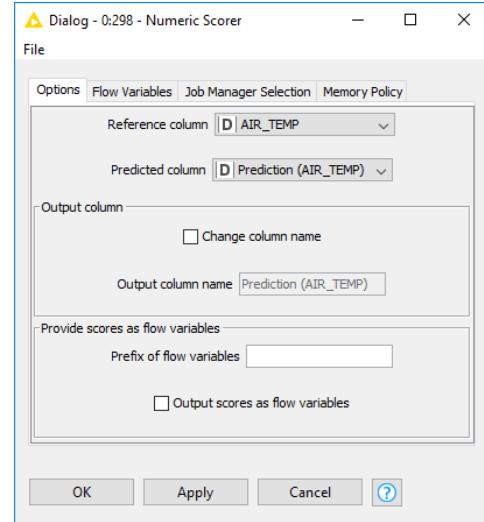
- Model coefficients provided in 2nd output port
- Also available: Polynomial and Tree Ensemble Regression nodes



# 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.



The "Statistics - 0:298 - Numeric Scorer" view displays a table titled "Scores" with 6 rows. The table has columns for "Row ID" and "Prediction (AIR\_TEMP)". The data is as follows:

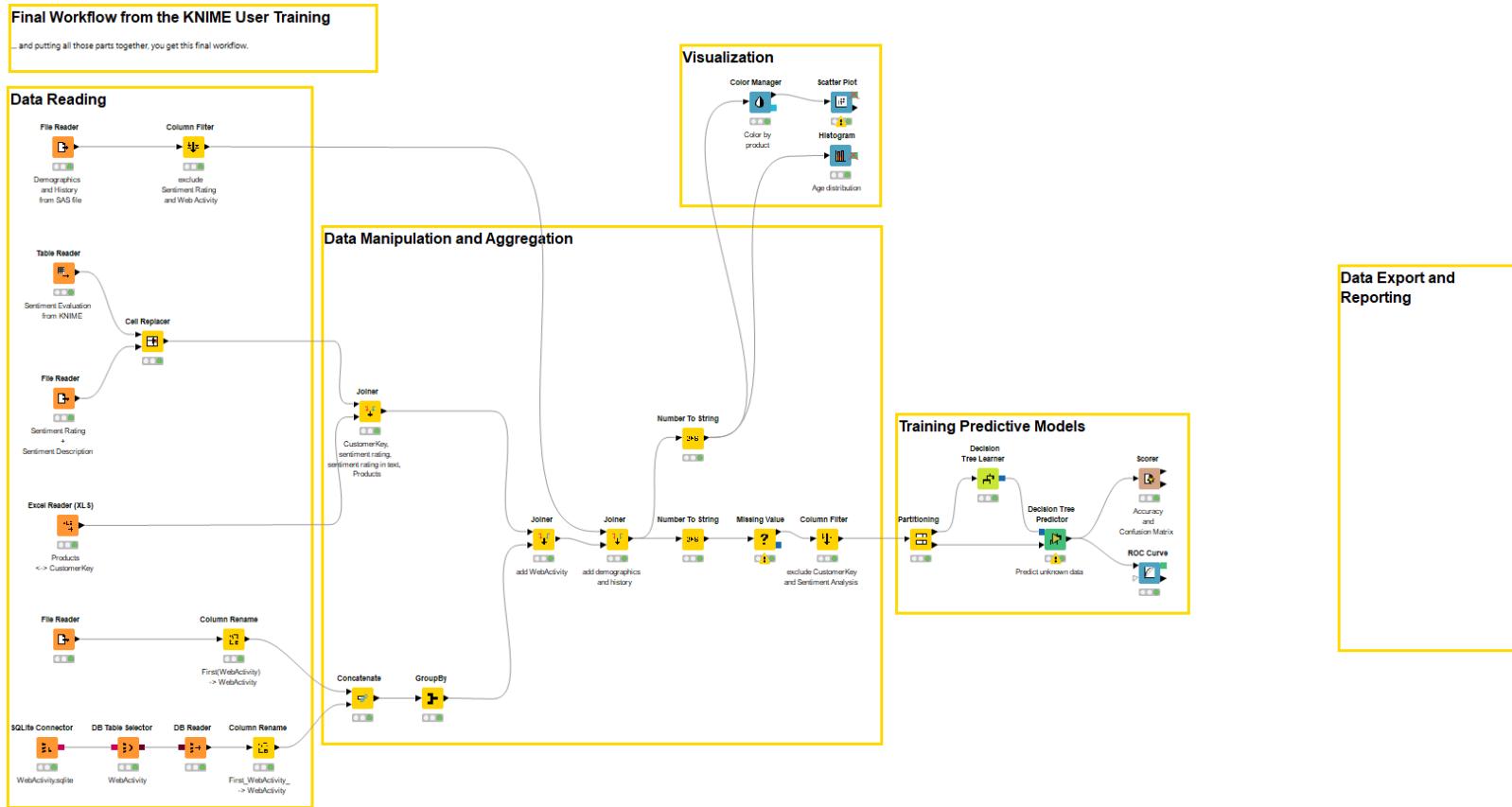
Row ID	Prediction (AIR_TEMP)
$R^2$	0.333
mean absolute error	3.574
mean squared error	21.329
root mean squared error	4.618
mean signed difference	1.048
mean absolute percentage error	NaN

# 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

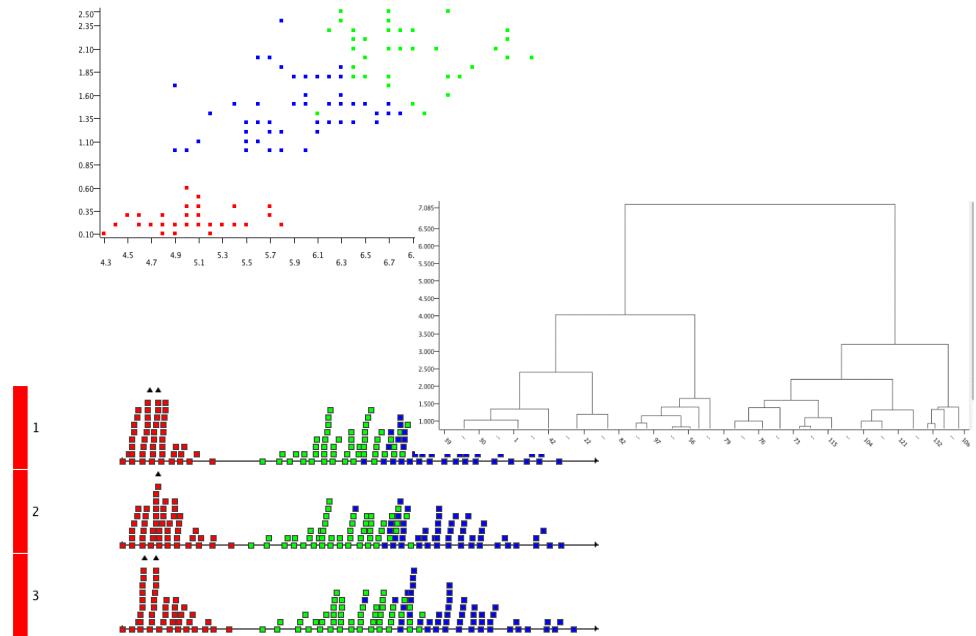
# Today's Example



# Clustering

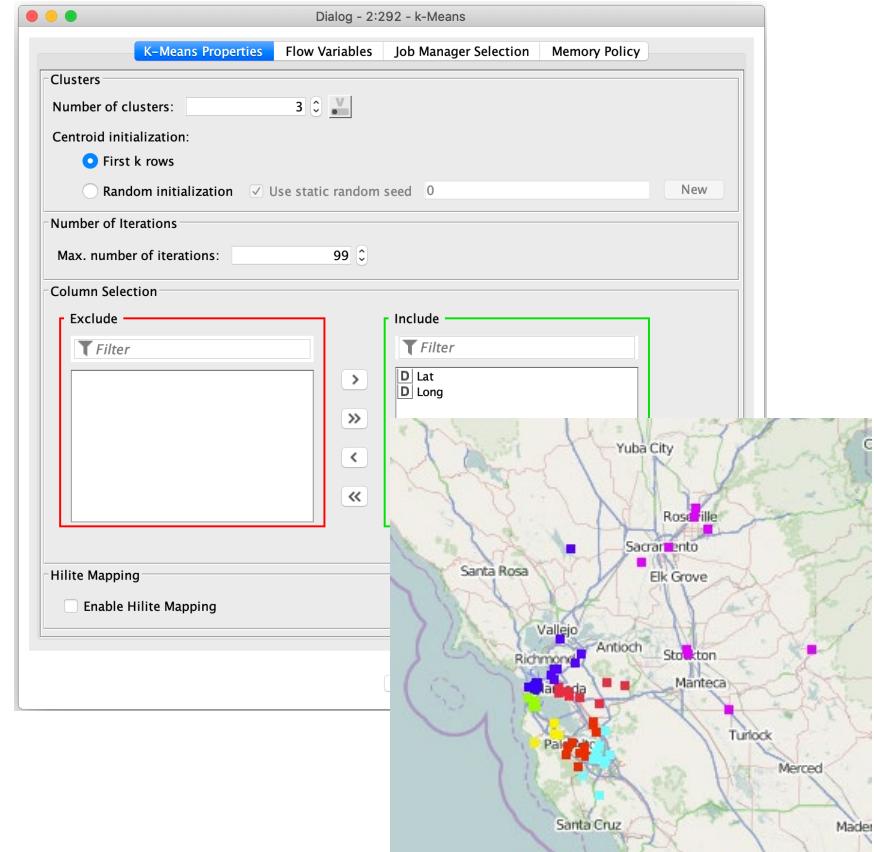
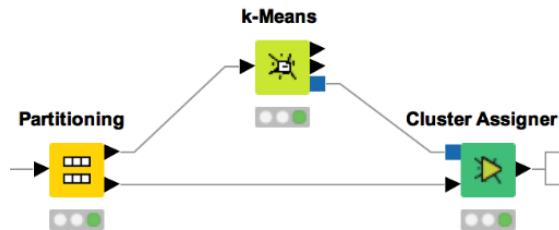
Discover hidden structure in **unlabeled** data (unsupervised)

- 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

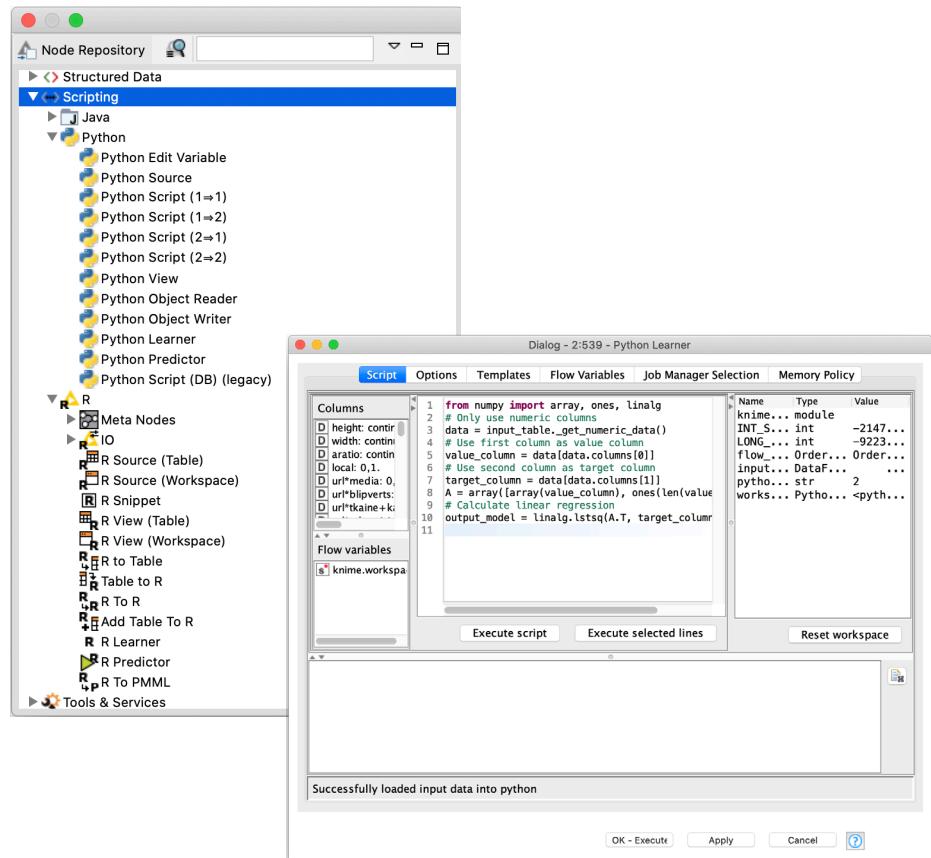
---

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

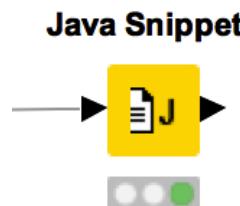
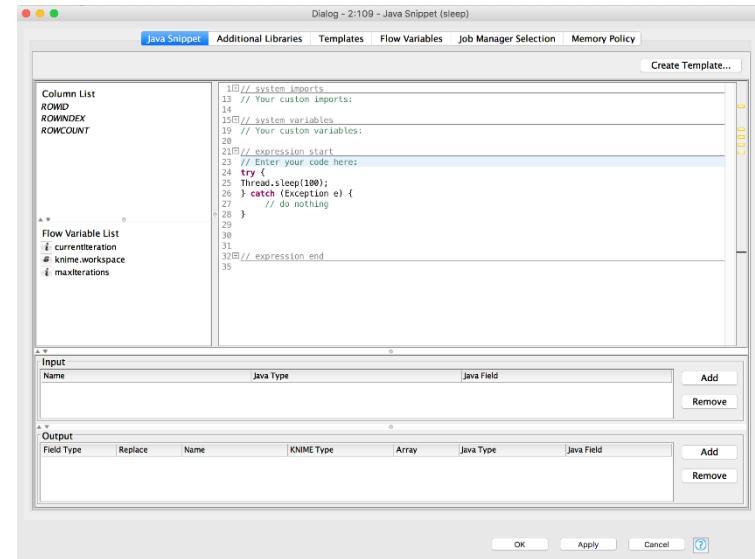
# Scripting Integrations: R and Python

- 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

---

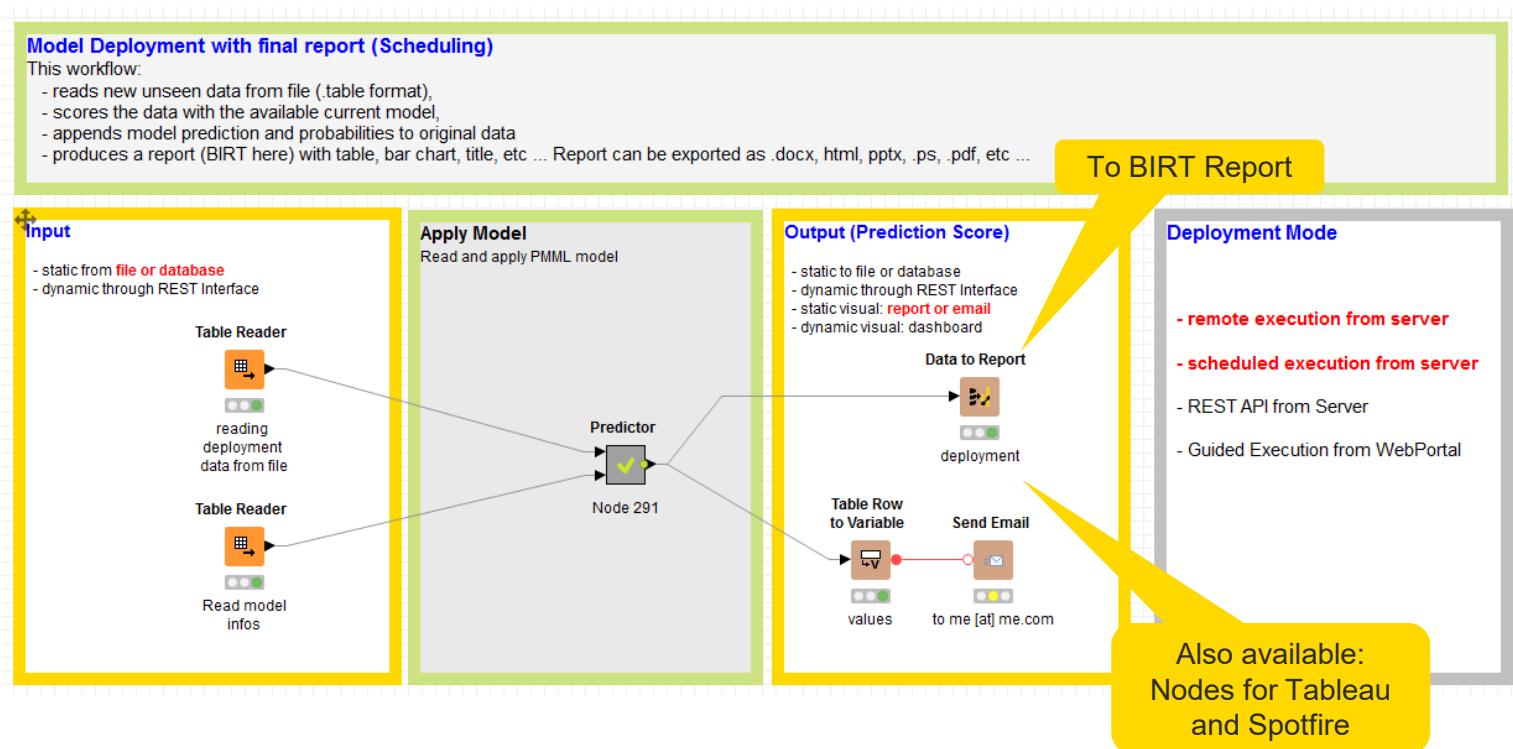
## Input

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

## Output

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

# To Report / Email

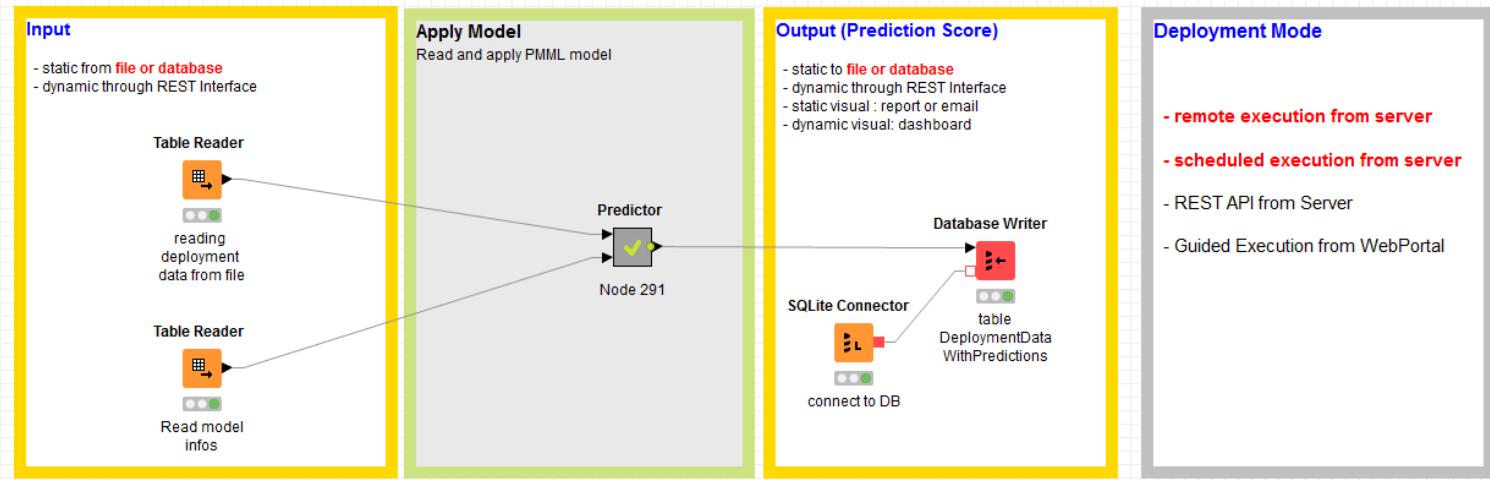


# To File / Database

## Model Deployment File to Database (Scheduling)

This workflow:

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

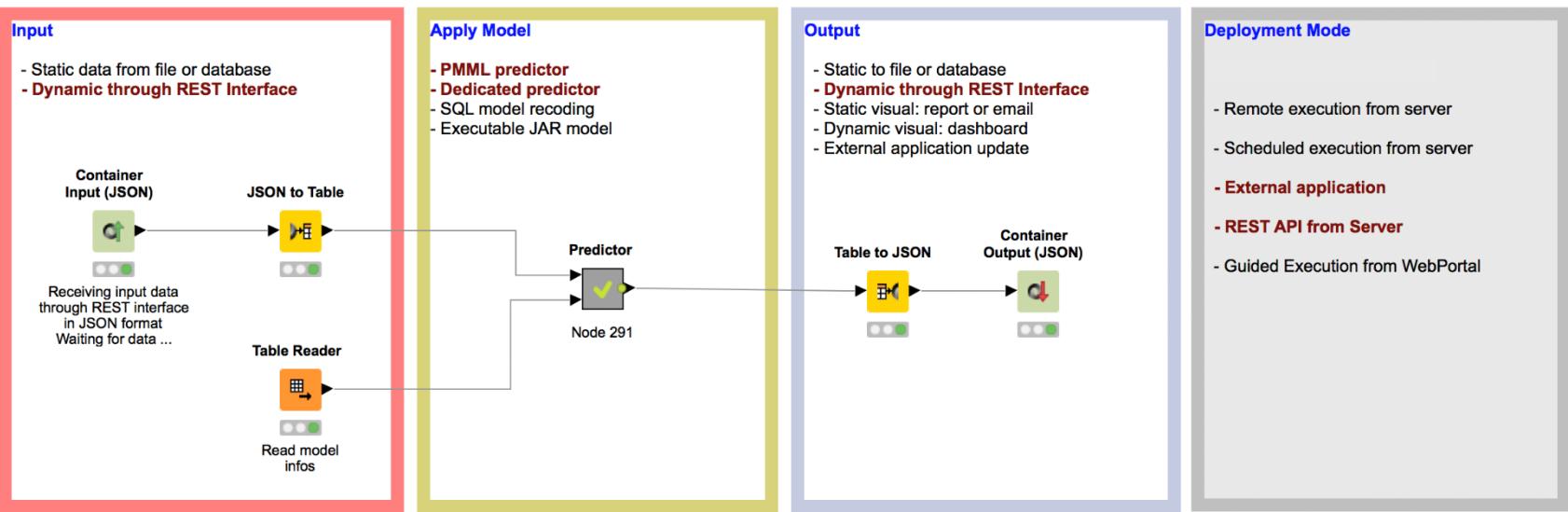


# REST API (Available on KNIME Server)

## Model Deployment as REST API

This workflow:

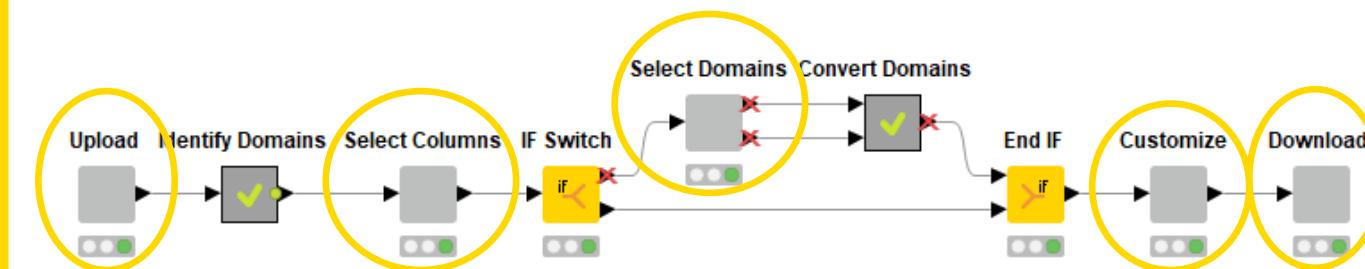
- receives new unseen data via REST interface (JSON format),
- scores the data with the available current model,
- appends model prediction and probabilities to original data,
- makes results available at the output REST interface.



# 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

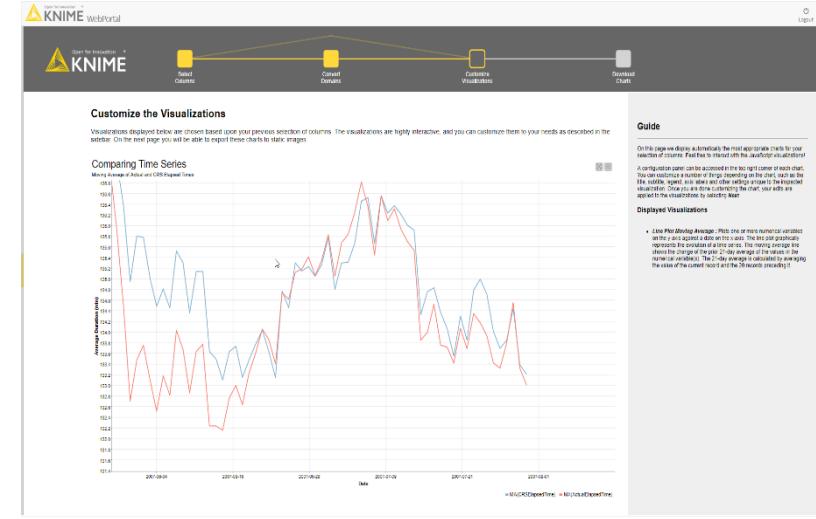
**Step 4**  
Interactive View

**Step 5**  
Download Image

# Workflow on KNIME WebPortal

The screenshot shows the 'Data Upload' step of the KNIME WebPortal. On the left, there's a form titled 'Data Upload' with a dropdown menu set to 'adult.csv'. On the right, a 'Guide' section provides instructions: 'Upload the dataset to visualize. The file must be in CSV format. The file will be appended to the server for further processing.' The KNIME logo is at the top left, and a 'Logout' button is at the top right.

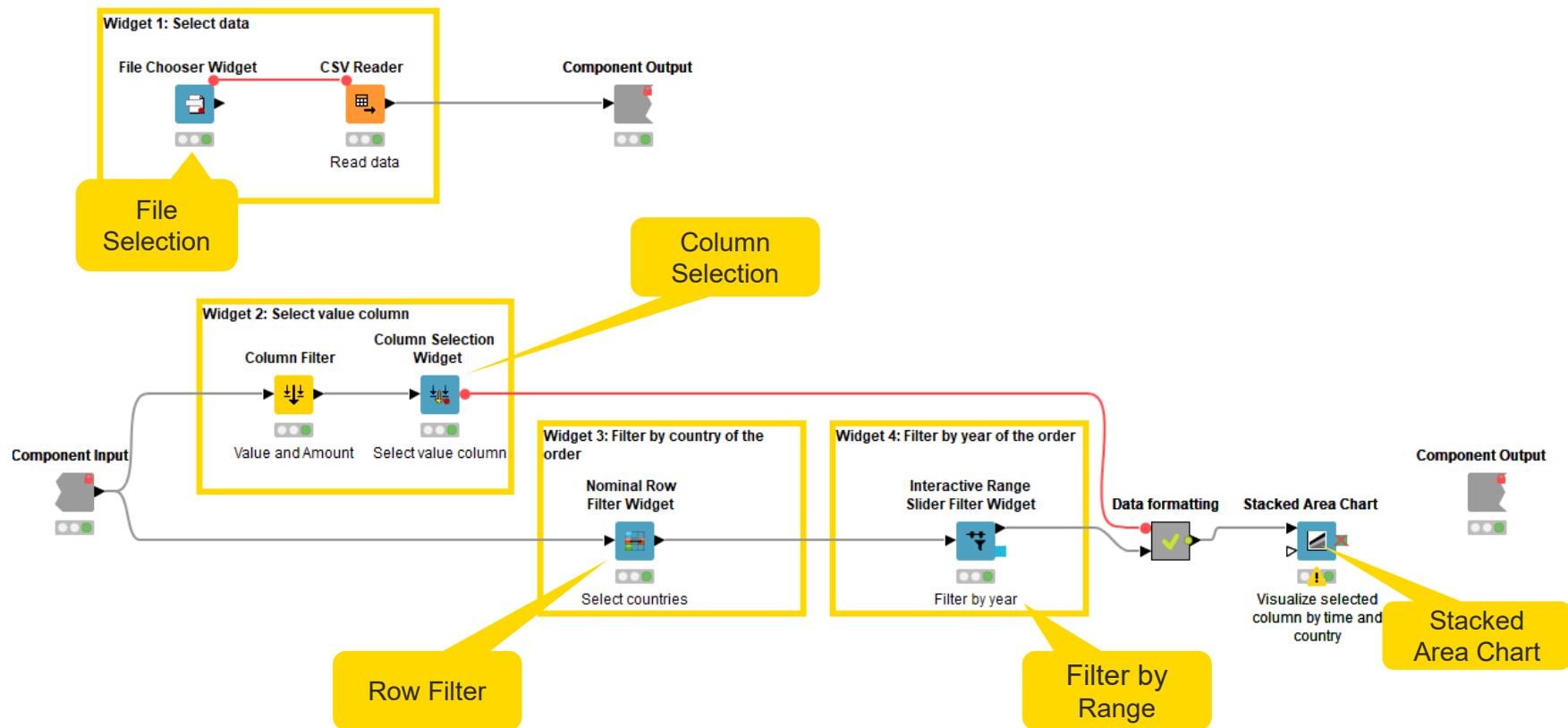
WebPortal Page  
(Step 1)  
Upload File



WebPortal Page  
(Step 4)  
Interactive View

Available in  
KNIME Server

# Components to Produce Dashboard on Web Page

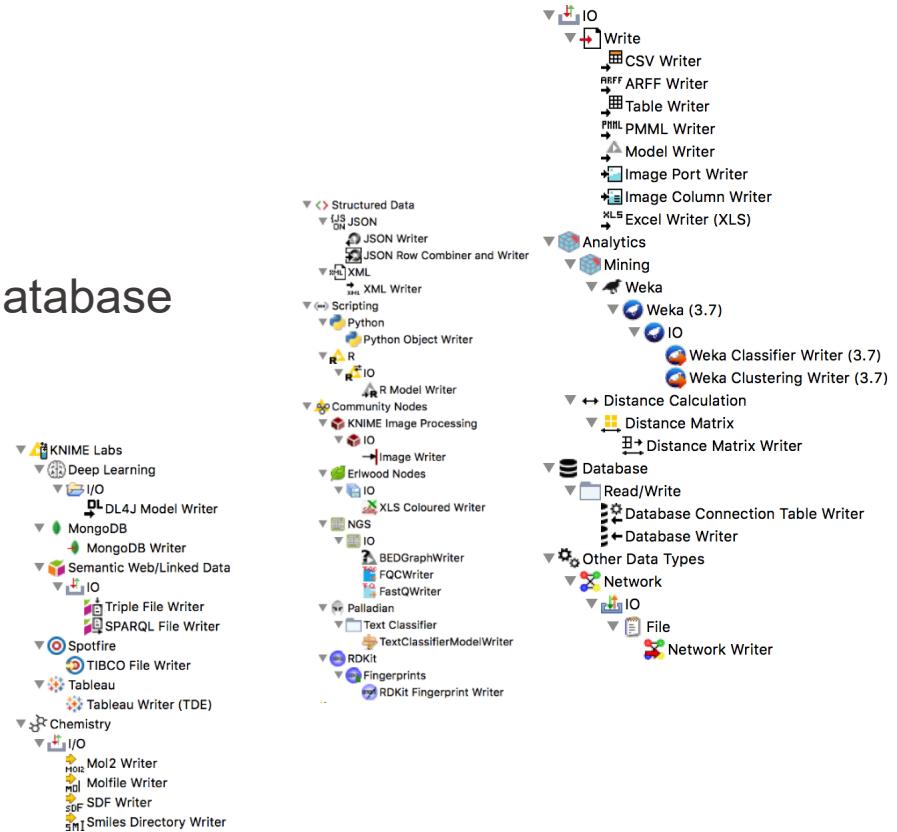
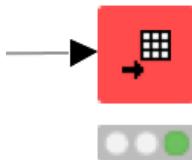


# Data Export Nodes

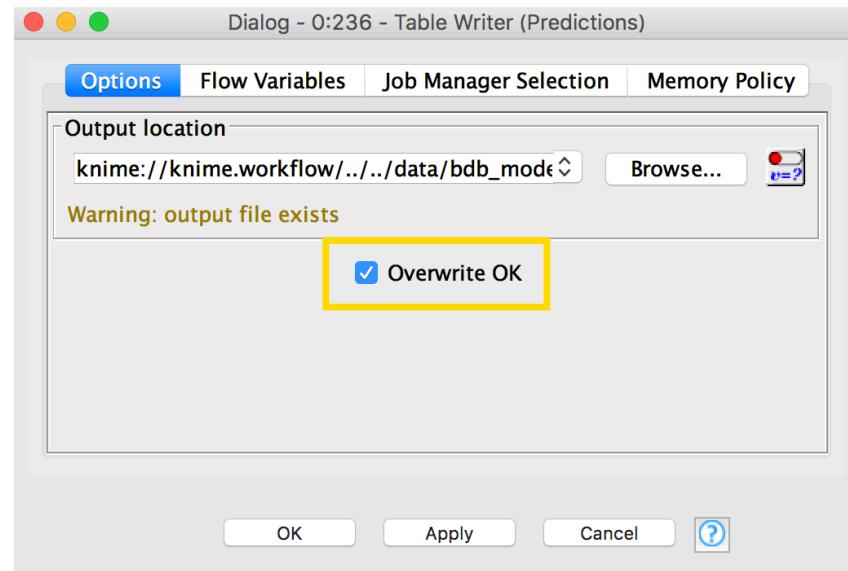
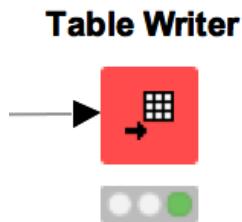
Typically characterized by:

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

## Table Writer



# New Node: Table Writer



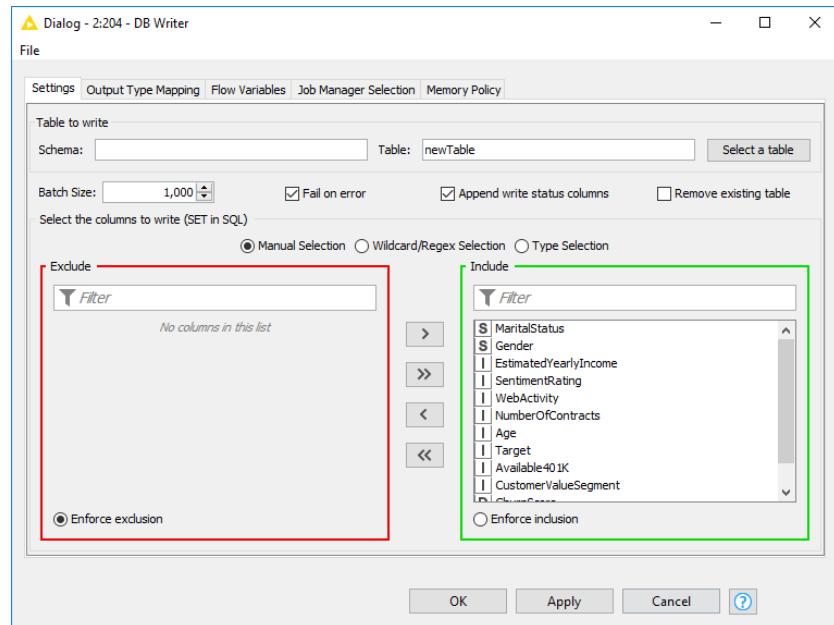
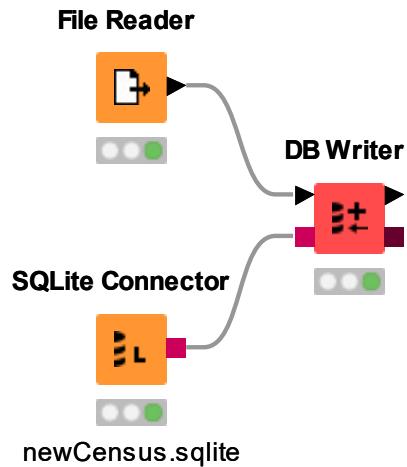
# New Node: XLS Writer

The image shows the KNIME interface. On the left, there is a red 'Excel Writer (XLS)' node with three output ports. The top port is highlighted with a red box. On the right, a configuration dialog for the 'Excel Writer (XLS)' node is open, titled 'Dialog - 3:260 - Excel Writer (XLS)'. The dialog has several tabs: 'Settings' (selected), 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Settings' tab contains fields for 'Output location' (Write to: Relative to knime.workflow, File: ./data/example.xlsx, Browse...), 'Sheet name' (Name of the sheet: default), 'Add names and IDs' (checkboxes for add column headers and add row ids), 'Missing value pattern' (checkbox for For missing values write: [ ]), 'Layout' (Portrait selected, Landscape, US Letter 8 1/2 x 11 in), and selection options for 'Exclude' and 'Include' columns.

**Exclude:** A red box highlights the 'Exclude' section. It contains a 'Filter' input field and a message 'No columns in this list'. Below it is a radio button 'Enforce exclusion'.

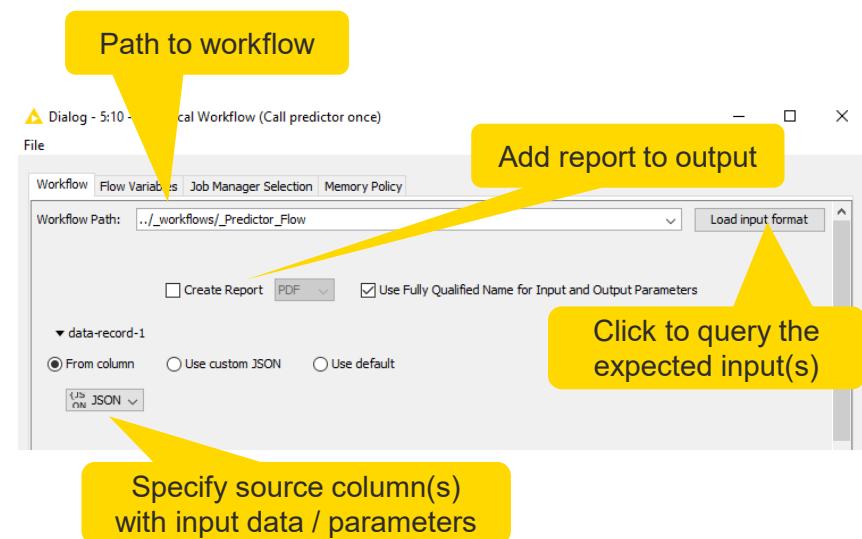
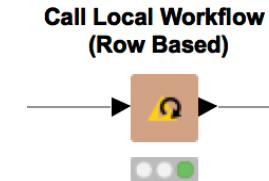
**Include:** A green box highlights the 'Include' section. It contains a 'Filter' input field and a list of columns: CustomerKey, MaritalStatus, Gender, EstimatedYearlyIncome, NumberOfContracts, Age, Target, Available401K, and CustomerValueSegment. Below the list is a radio button 'Enforce inclusion'.

# New Node: Database Writer



# Automation: Call Local Workflow

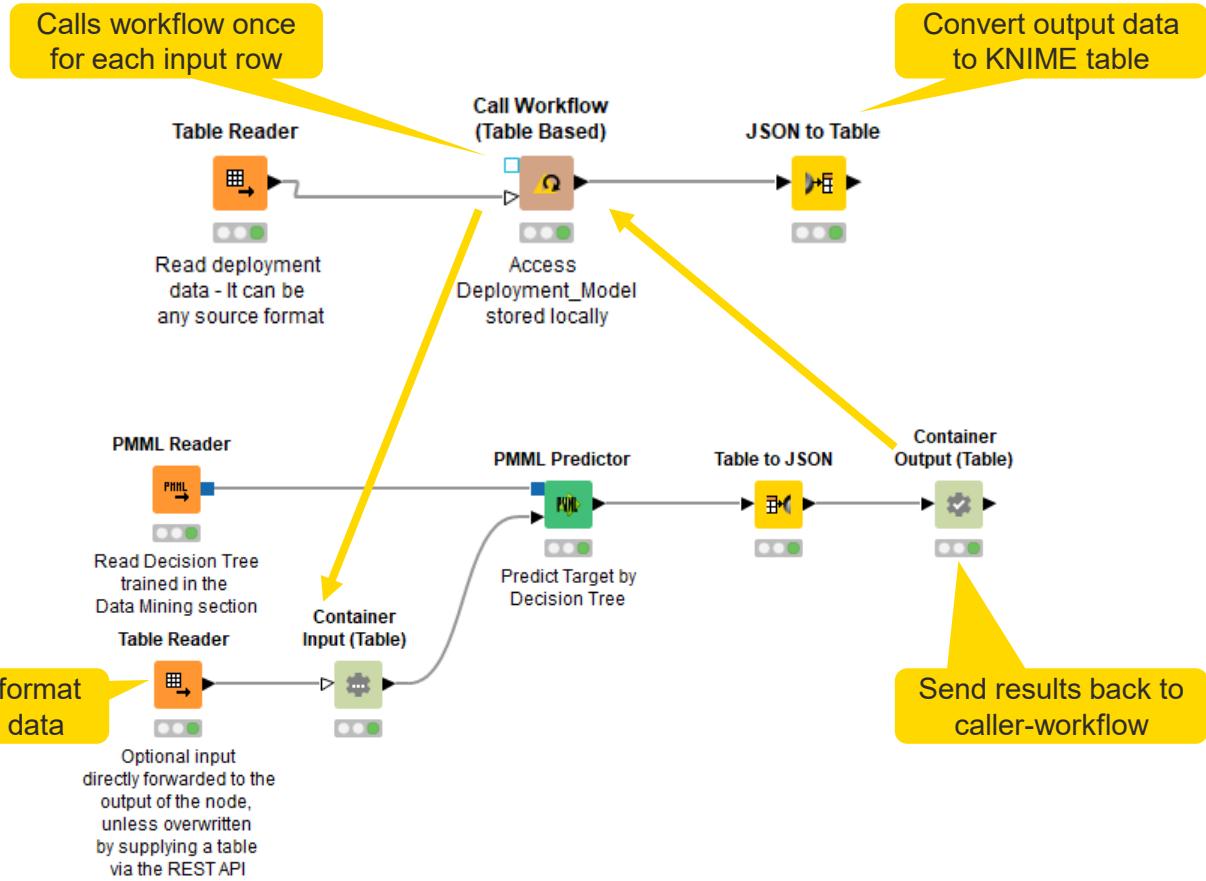
- Use Call Local Workflow node to send data and parameters to other workflows and trigger execution
  - Send results back to caller-workflow
  - Include report from called workflow
- Create modular workflows
  - E.g. separate workflows for ETL and prediction
- Alternative: Call Remote Workflow
  - Trigger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



# Automation: Call Local Workflow

ETL

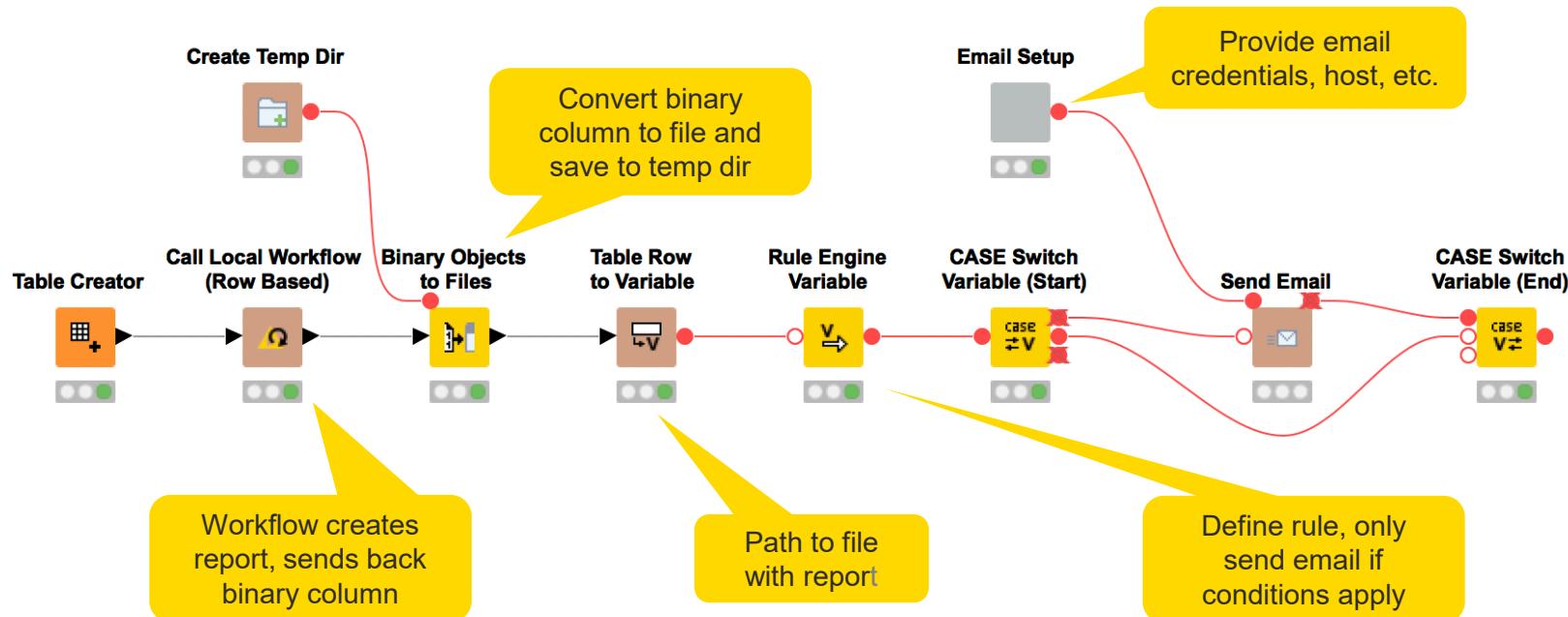
Prediction



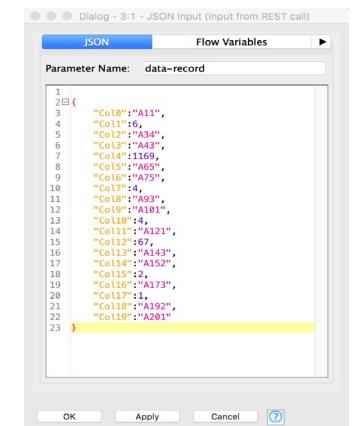
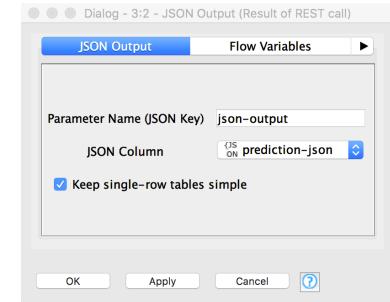
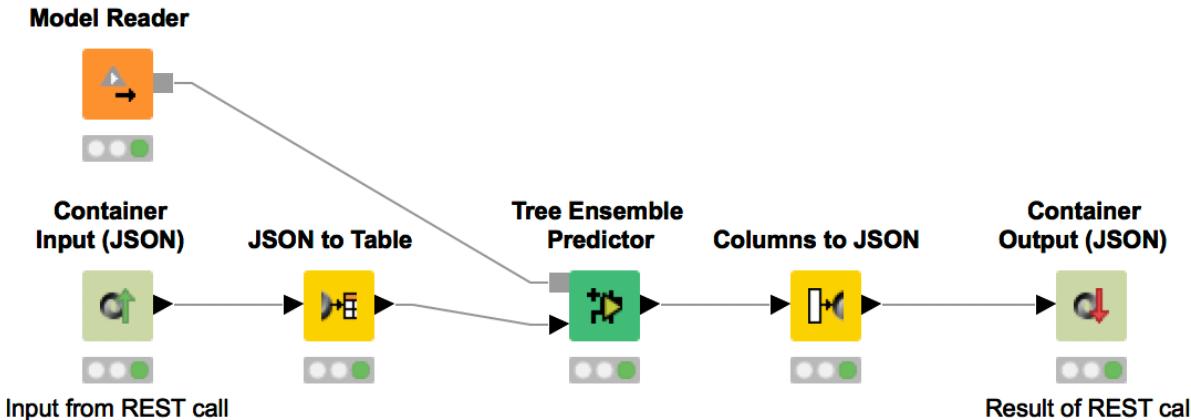
# Use Call Local Workflow to Send Conditional Emails with Report

Sometimes, report should be sent under specific circumstances

- E.g. if some KPI is below threshold



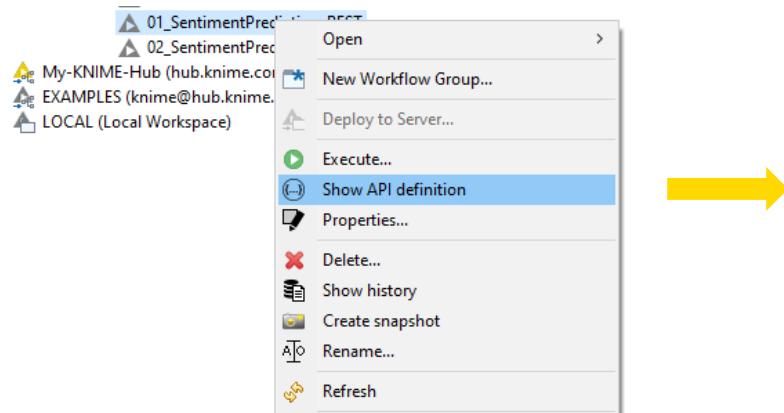
# KNIME Server as a REST Resource



<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



The screenshot shows the Swagger UI interface for the KNIME REST API. The URL in the address bar is `https://datascience1.knime.com/knime/rest/doc/index.html?url=https%3A//datascience1.knime.com/knime/rest/v4/repository/Users/moritz.heine/LaPanca/01_SentimentPrediction_REST:execution`. The interface is organized into sections: **metadata**, **job-control**, and **execution**.

**execution**

**POST /v4/repository/Users/moritz.heine/LaPanca/01\_SentimentPrediction\_REST:execution** Executes a job from this workflow

This call combines loading, executing, and deleting a job in one call. You can pass input parameter for quickform nodes defined in the workflow. All input parameters are suffixed with their unique node ID in order to make the parameters unique themselves. If a parameter name is unique without the node ID suffix you can also omit the suffix when sending it to the server. For example, if the fully qualified parameter name is `int-input-1` and there is no other input parameter that begins with `int-input` you can use `int-input` as the name in your request.

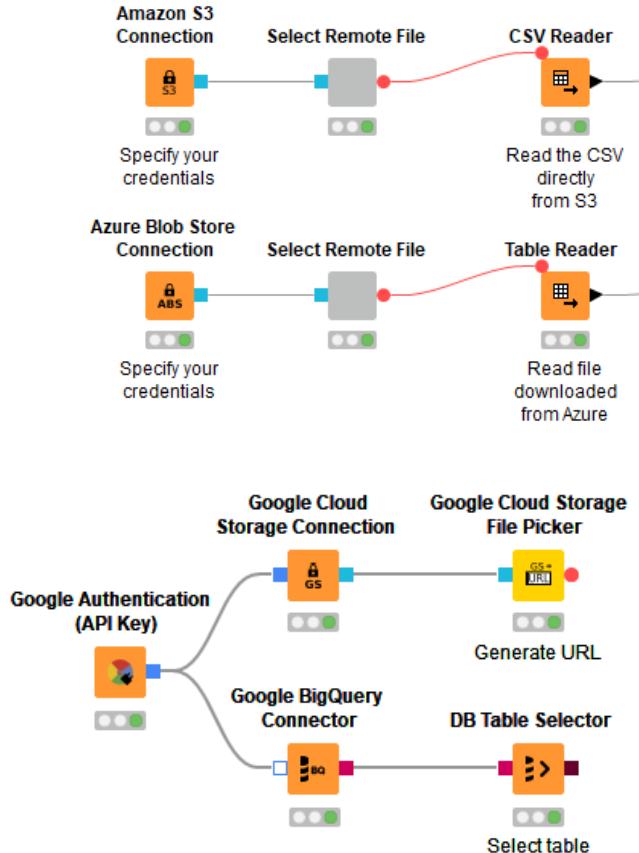
**Parameters**

Name	Description
timeout integer (query)	Sets a timeout in milliseconds that the call should wait for the job being loaded. If the workflow doesn't load within the time a 504 error will be returned.  timeout - Sets a timeout in milliseconds that t
format string (query)	If the workflow creates a report you can specify the desired report format. If no report format is provided no report will be generated.  PDF
reset boolean (query)	True if the job should be reset before execution. If false (the default) job execution continues from its saved state.  --

# Remote File Handling – Cloud Storage

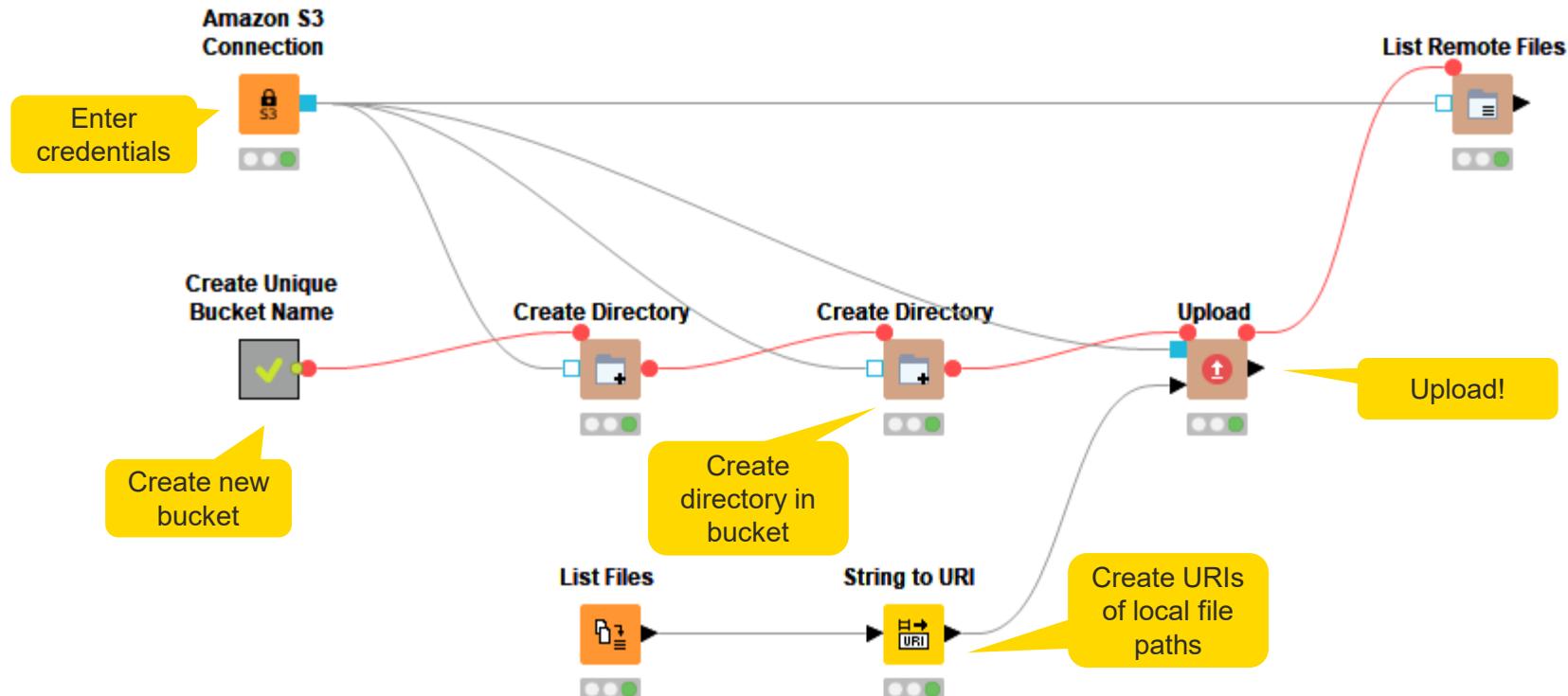
Integrate remote data sources from Amazon AWS, Microsoft Azure, and Google Cloud

- 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

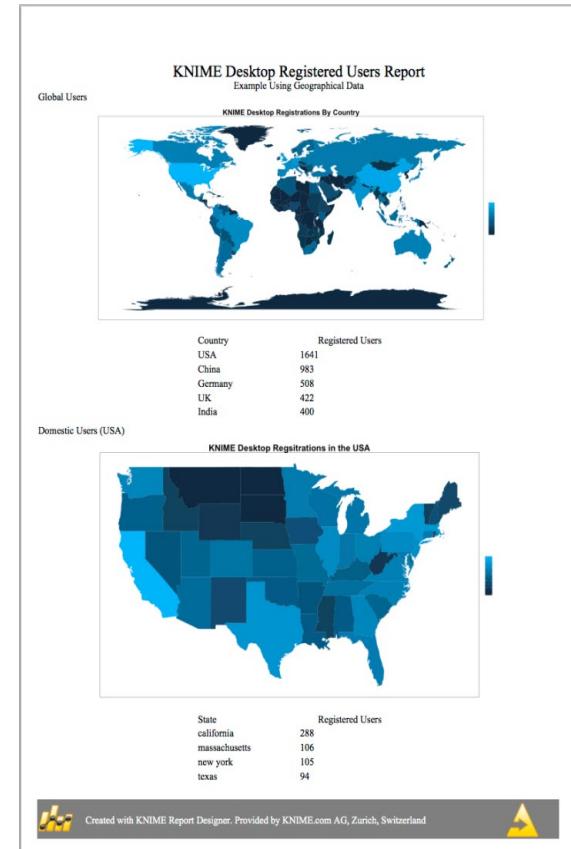
Example: Upload all files from a local directory to Amazon S3



# **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

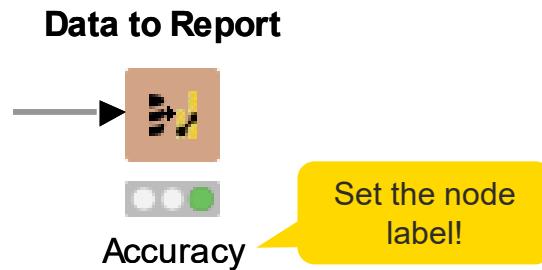
---

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

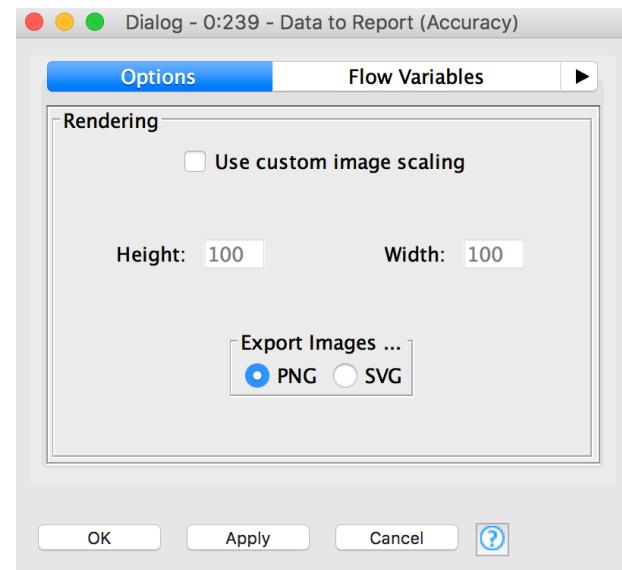


# New Node: Data to Report

Send a data table to BIRT



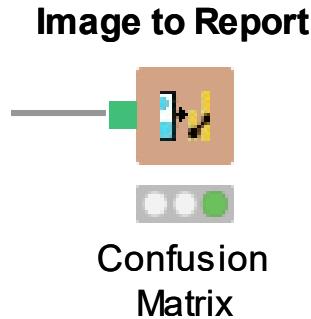
Hint: The node label will be used to identify the data source in the reporting view -> Make sure to use understandable labels if you have more than one data source



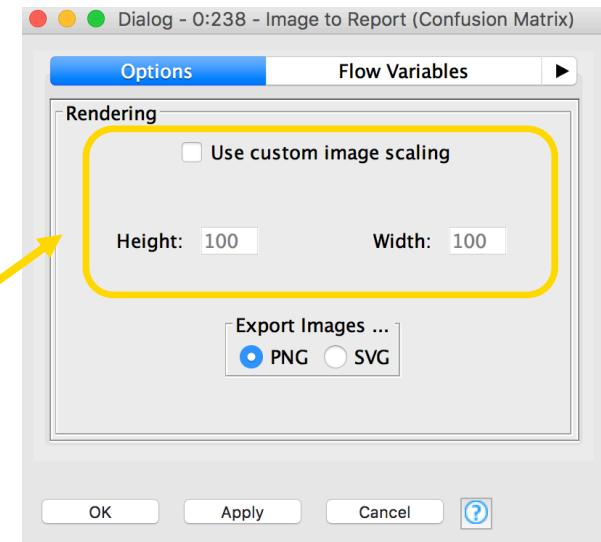
# New Node: Image to Report

Send an image to BIRT

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

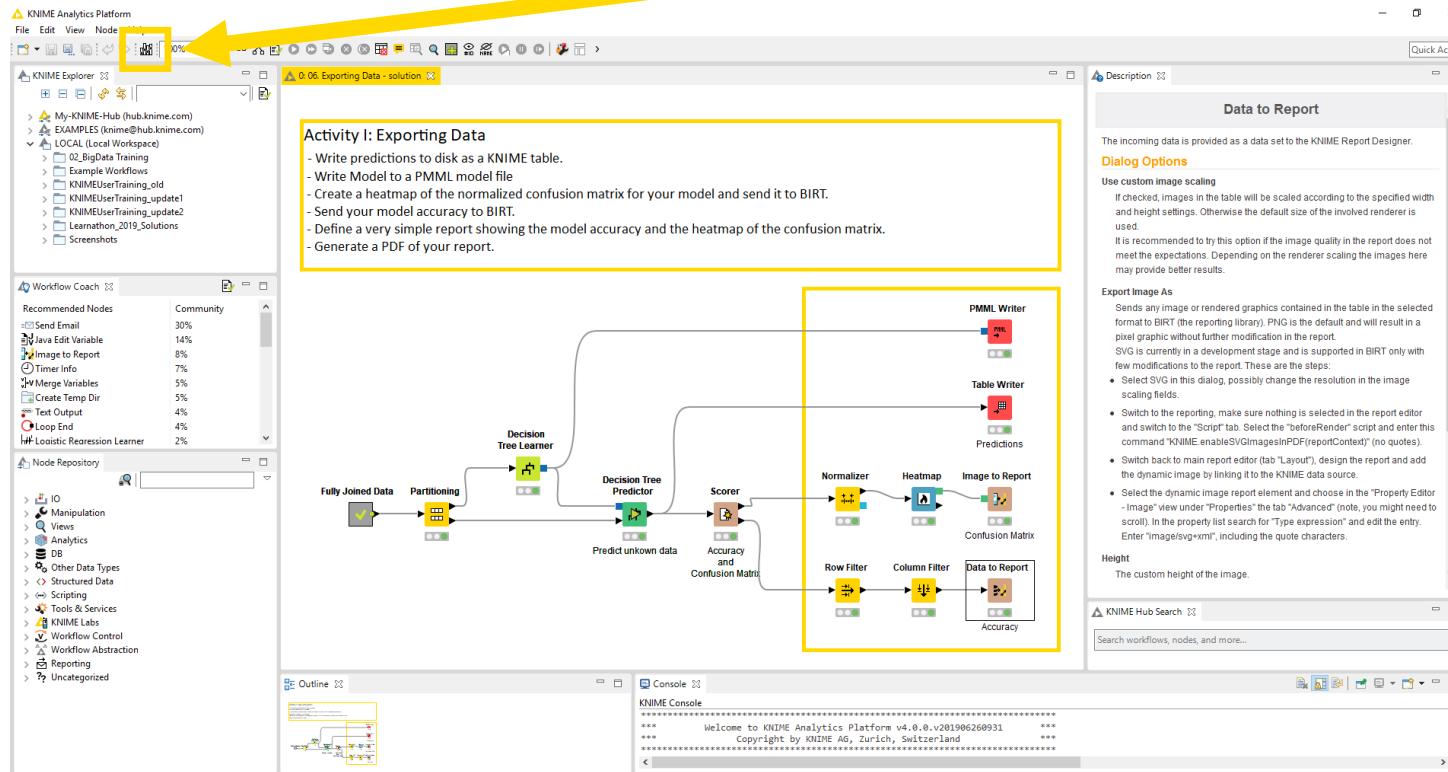


Hint: Customize the image size in the Data to Report node to fit the report



# Edit the Report

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



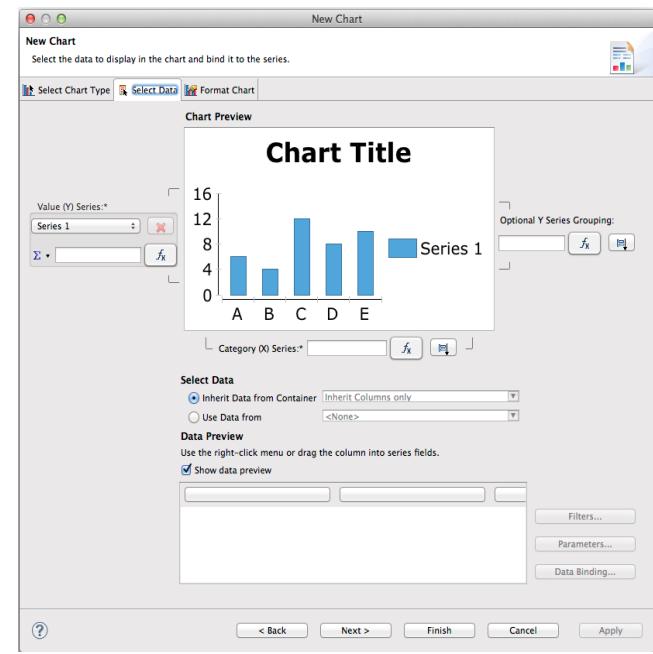
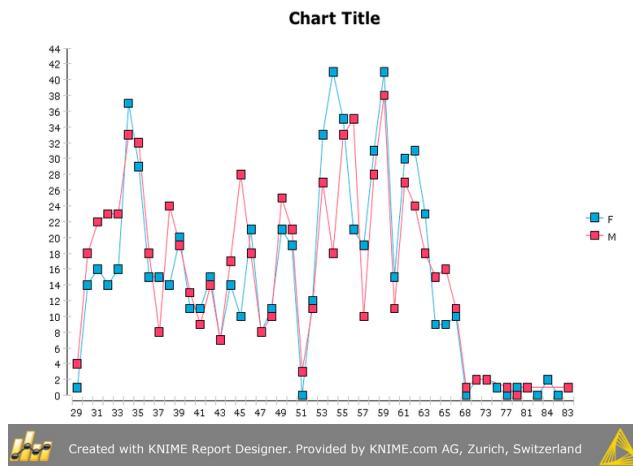
# Reporting Perspective

The screenshot shows the KNIME Reporting Perspective. On the left, there's a tree view of a workspace with projects like '01\_KNIMEUserTraining' and 'exercises'. Below it is a palette with various report items: Label, Aa Text, Dynamic Text, Data, Image, Grid, List, Table, Chart, Cross Tab, Quick Tools, Aggregation, and Relative Time Period. A yellow callout points to this palette with the text: "Add report items via drag and drop". In the center, there's a main area titled "My Data Mining Report" containing sections for "Overall Model Accuracy" and "Confusion Matrix". A yellow callout points to this area with the text: "Data from KNIME - names of data sources are taken from node label". At the top, there's a toolbar with a "Create Report" button, which is highlighted with a yellow callout and the text: "Click button to create report". At the bottom, there are tabs for "Layout", "Master Page", "Script", "XML Source", and "Properties". The "Properties" tab is selected, showing a "General" section with fields for Author, Created by, Path, Title, Themes, Report Orientation, Display Name, and Thumbnail. A yellow callout points to this tab with the text: "View tabs". On the right, there's a "Report layout – only structure, data is filled in when creating the report" callout pointing to the main report area.



# 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.)

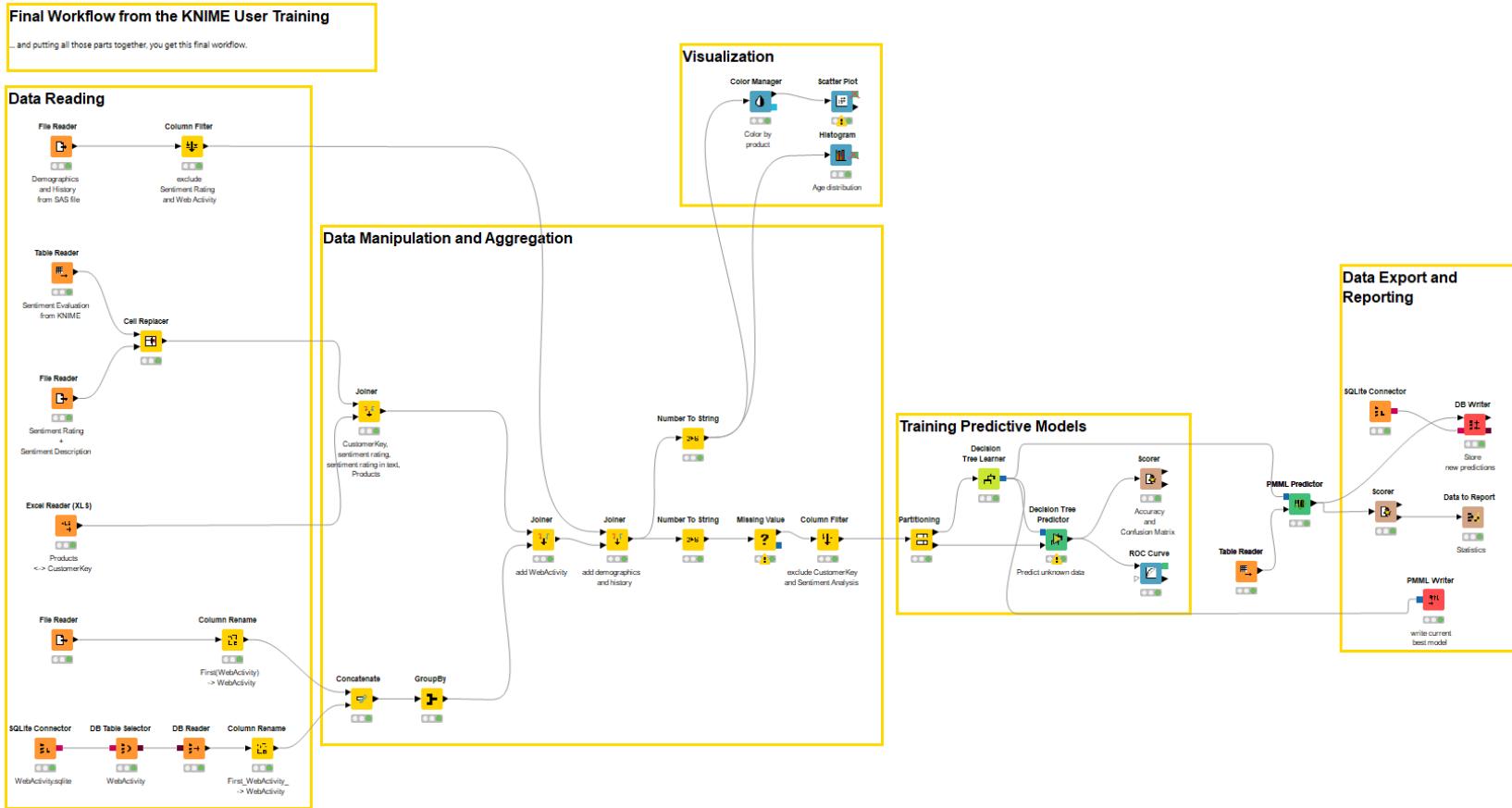
# Exporting Data Exercise

---

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](mailto:education@knime.com)