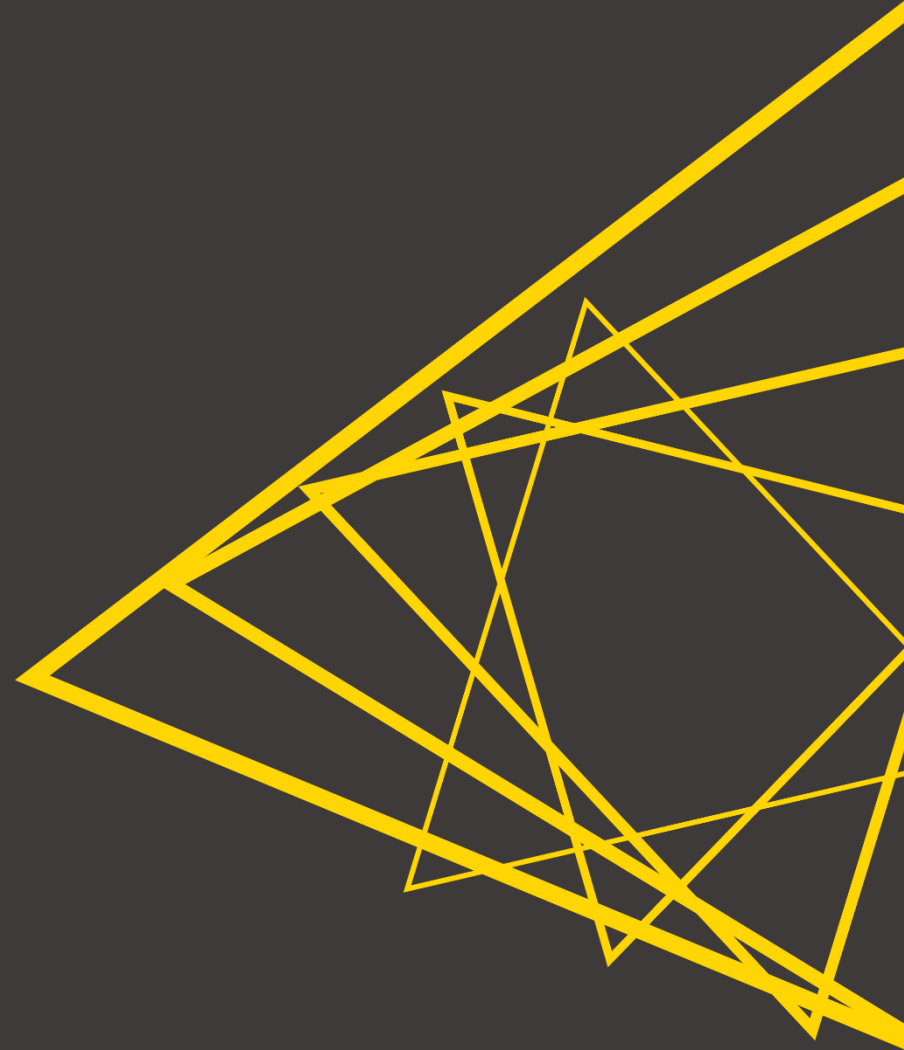


ETH Zürich Workshop

Thursday, January 12 2023

Alice Krebs and Greg Landrum



Before we get started...

... some questions for you:

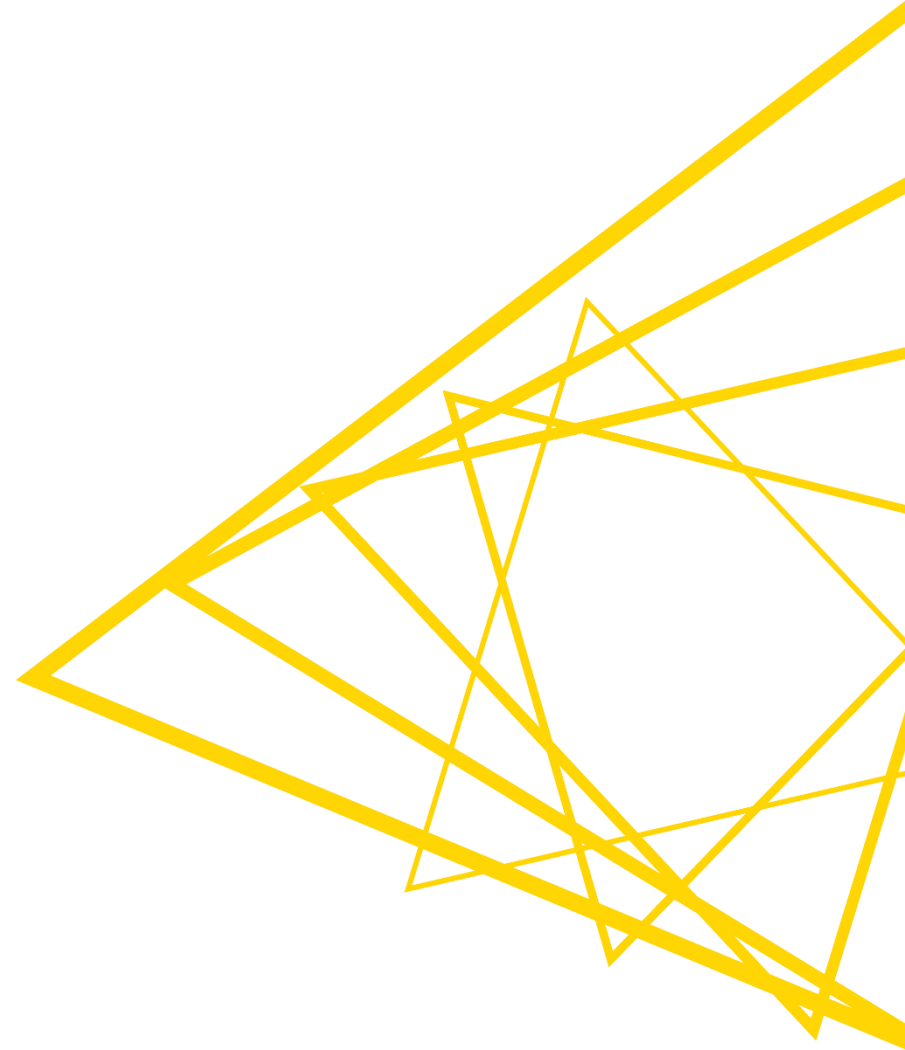
- What's your background? BSc, MSc, PhD?
- Which department are you from?
- What tools are you using for data analysis?
- What do you expect/hope to learn today?
- Did you manage to open the “00_Setup” workflow already?

Change of plan for today

- 9.00 – 10.15: Introduction to KNIME Analytics Platform
- 10.15 – 10.45: Coffee break
- 10.45 – 12.00: Hands-on session

- 9.00 – 9.45: Introduction to KNIME Analytics Platform
- 9.45 – 10.15: Start hands-on session
- 10.15 – 10.45: Coffee break
- 10.45 – 12.00: Hands-on session

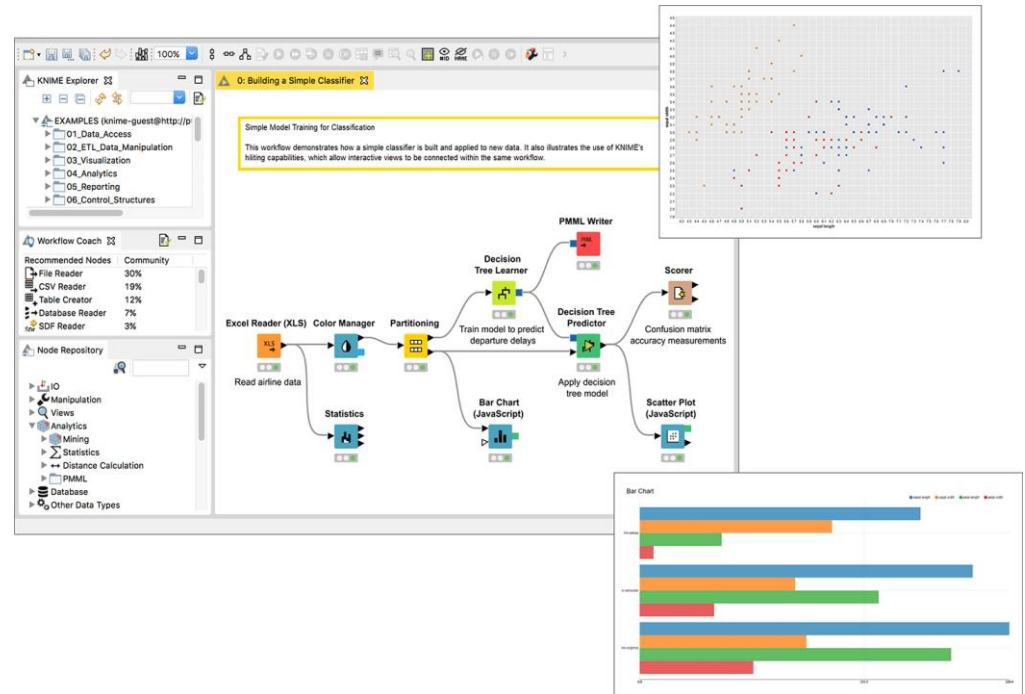
Introduction to KNIME Analytics Platform



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:

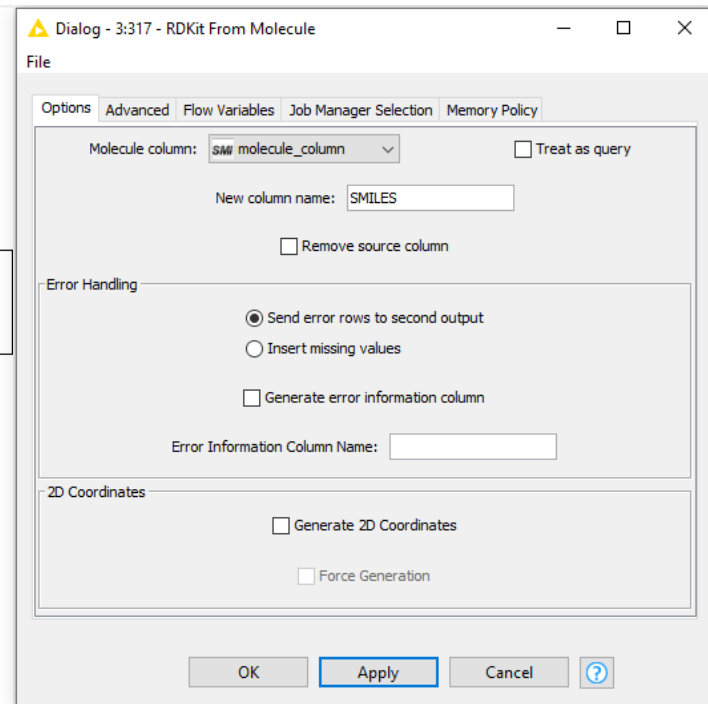
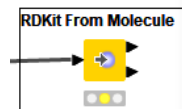
- Text Mining
- Network Mining
- Cheminformatics
- Many integrations, such as Java, R, Python, Weka, Keras, Plotly, H2O, etc.



Graphical programming paradigm?

- Lets users create programs by manipulating program elements graphically rather than by specifying them textually

```
mols = []  
for smi in input_1_pandas[self.molecule_column]:  
    mols.append(Chem.MolFromSmiles(smi))
```



`rdkit.Chem.rdmolfiles.MolFromSmiles((AtomPairsParameters)SMILES, (SmilesParserParams)params) → Mol :`

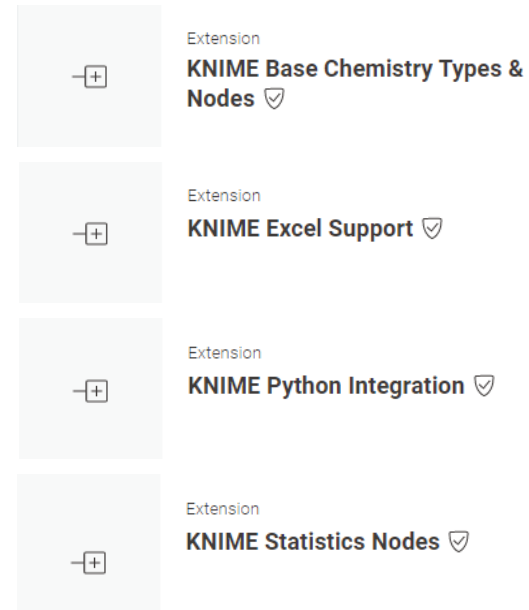
Construct a molecule from a SMILES string.

Extensions?

- KNIME AP comes by default only with basic functionality



- Extensions (and integrations) add specific functionality
- Not default, need to be added



The KNIME Workbench

The screenshot displays the KNIME Analytics Platform interface with the following components:

- KNIME Explorer:** Shows a project tree with folders like '2020_05_NIBR_Chemistry_training_412' and files like '01_Chemistry_basics'.
- Workflow Coach:** Lists recommended nodes: File Reader (28%), CSV Reader (20%), and Table Creator (13%).
- Node Repository:** A categorized list of nodes including IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, Community Nodes, KNIME Labs, Workflow Control, Workflow Abstraction, Reporting, Chemistry, and ChemAxon / Infocom.
- Main Canvas:** Displays a workflow titled '01_Chemistry_basics' with two steps:
 - Step 1: Read data from different sources** includes a 'File Reader', 'Excel Reader (XLS)', and 'SDF Reader' node, all connected to a 'Concatenate' node.
 - Step 2: Remove duplicates** includes 'Molecule Type Cast', 'RDKit Canon SMILES', and 'Duplicate Row Filter' nodes connected in sequence.
- Description Panel:** Provides details for '01_Chemistry_basics', including a title and a description of the workflow's purpose in cheminformatics.
- Console:** Shows log messages from the KNIME Console, including warnings about column specifications and node creation.
- Outline:** A small thumbnail view of the workflow canvas.

The KNIME Workbench – most important windows

The screenshot displays the KNIME Analytics Platform interface. The top bar shows the title 'KNIME Analytics Platform - /Users/daria_knime/knime-workspace_rdkit_20200421'. The main workspace is divided into several panes:

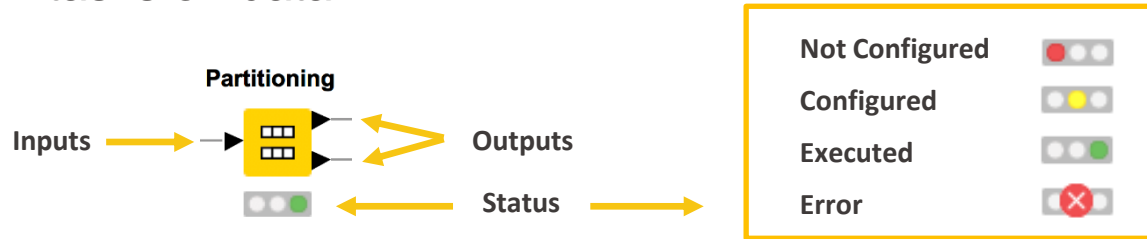
- KNIME Explorer:** Located on the top left, it shows a file tree with folders like '2020_04_21_RDKIT_ChemInformatics' and '2020_05_NIBR_Chemistry_training_412', and files like '01_Chemistry_basics', '02_Structure_Rendering', and '03_Substructure_Search'.
- Workflow Coach:** Below the Explorer, it lists 'Recommended Nodes' with their usage percentages: File Reader (28%), CSV Reader (20%), and Table Creator (13%).
- Node Repository:** On the bottom left, it provides a categorized list of nodes under 'Community Nodes', including IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, Community Nodes, KNIME Labs, Workflow Control, Workflow Abstraction, Reporting, Chemistry, and ChemAxon / Infocom.
- Workflow editor:** The central pane shows a workflow diagram with two steps. Step 1, 'Read data from different sources', contains 'File Reader', 'Excel Reader (XLS)', and 'SDF Reader' nodes, all of which feed into a 'Concatenate' node. Step 2, 'Remove duplicates', contains a 'Duplicate Row Filter' node that receives input from the 'Concatenate' node.
- Description:** On the top right, it shows the '01_Chemistry_basics' workflow description. The title is '01_Chemistry_basics'. The description states: 'This workflow demonstrates basic cheminformatics functionality within KNIME Analytics platform: Reading and writing various chemistry data formats; canonicalization of chemical structure; duplicate filtering; descriptor calculation; interactive filtering on multiple properties. Training sets were collected from ChEMBLdtd. Each set corresponds to a publication in which the molecule was determined experimentally.'
- Console:** At the bottom right, it displays log messages from the 'KNIME Console'. The log file is located at: /Users/daria_knime/knime-workspace_rdkit_20200421/.metadata/knime/knime.log. The messages include warnings about column 'assay_id' and 'Color Manager', and confirmations that nodes created empty data tables.

Yellow callout boxes highlight the following components:

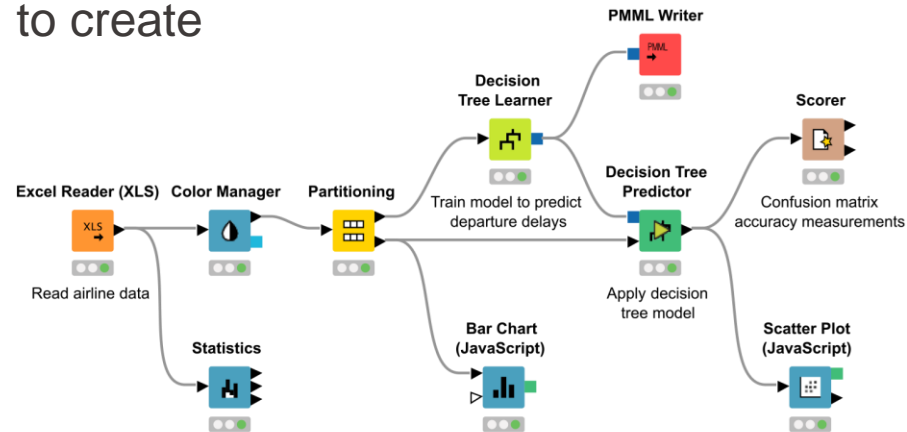
- Workflows:** Points to the KNIME Explorer pane.
- Nodes:** Points to the Node Repository pane.
- Workflow editor:** Points to the central workflow diagram.
- Node description:** Points to the Description pane.

Visual KNIME Workflows

NODES perform tasks on data



Nodes are combined to create **WORKFLOWS**

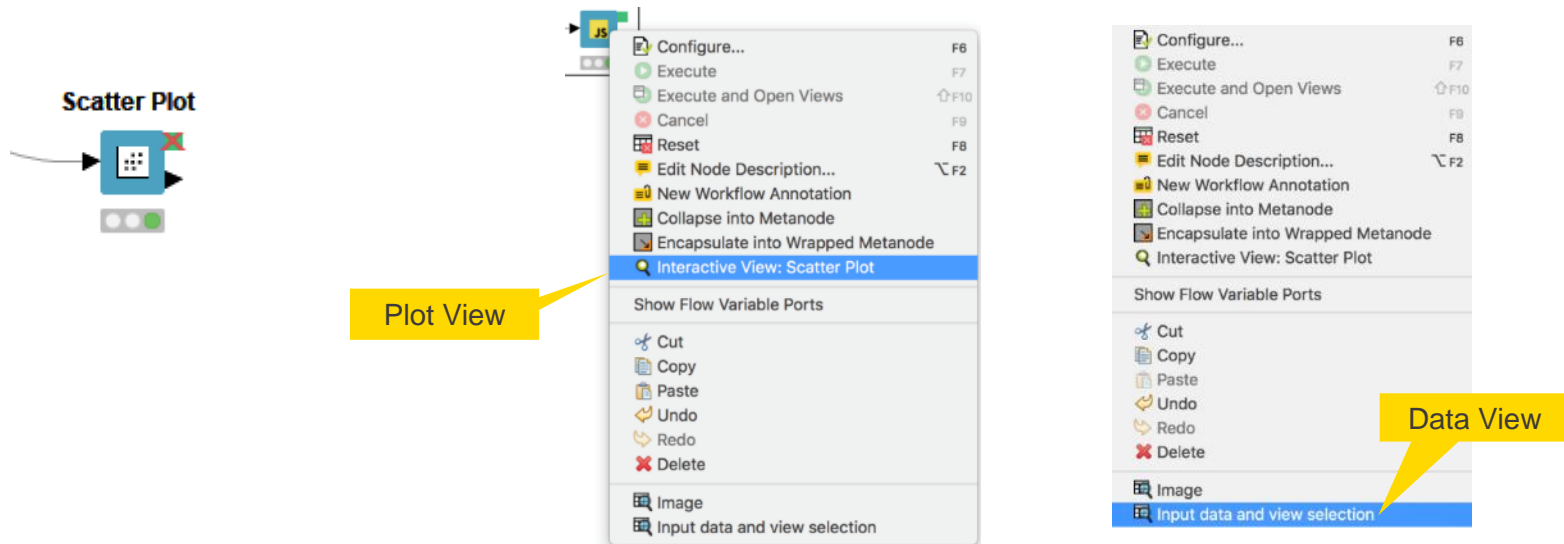


Node Outputs and Views

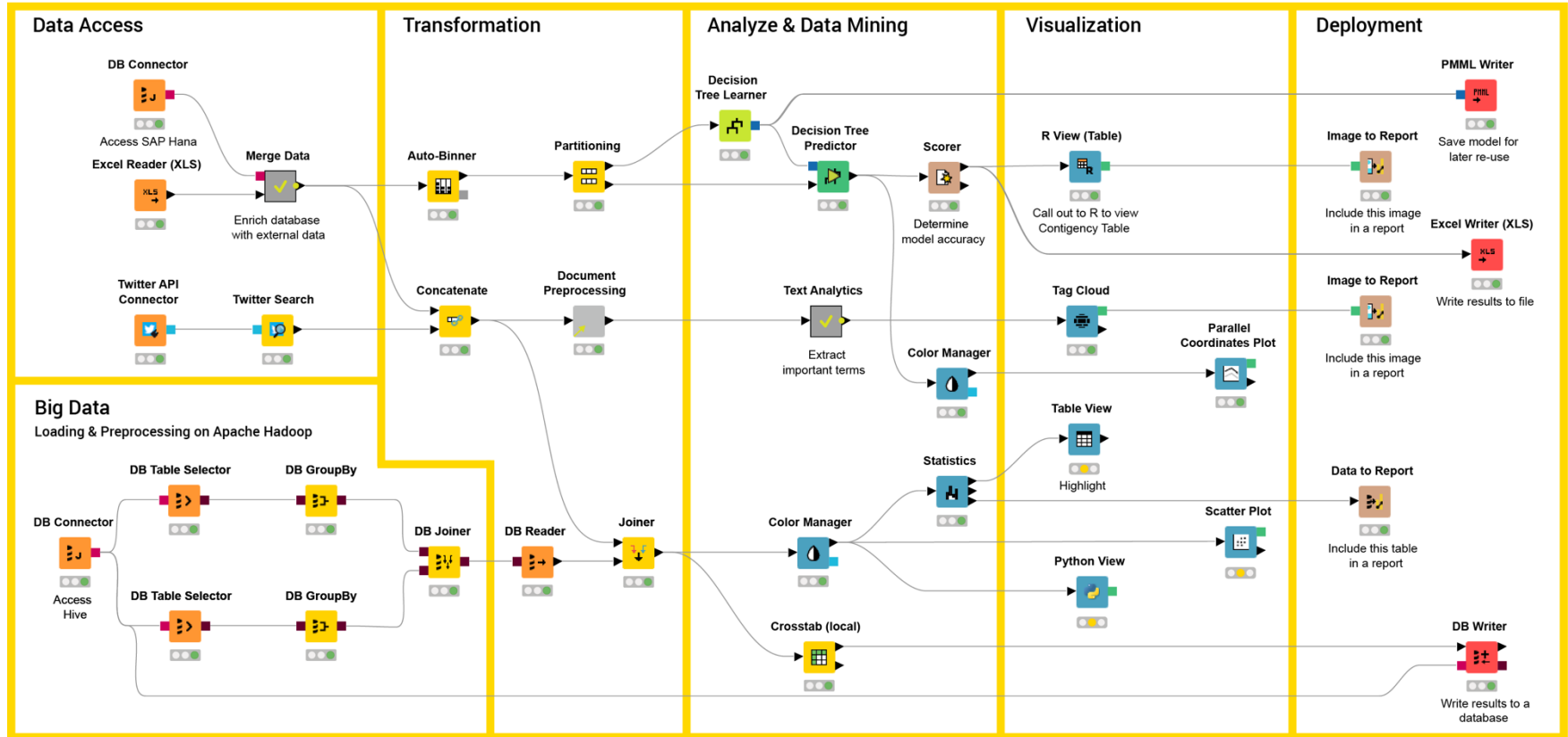
- Right-click executed node
- Select View option in context menu

OR

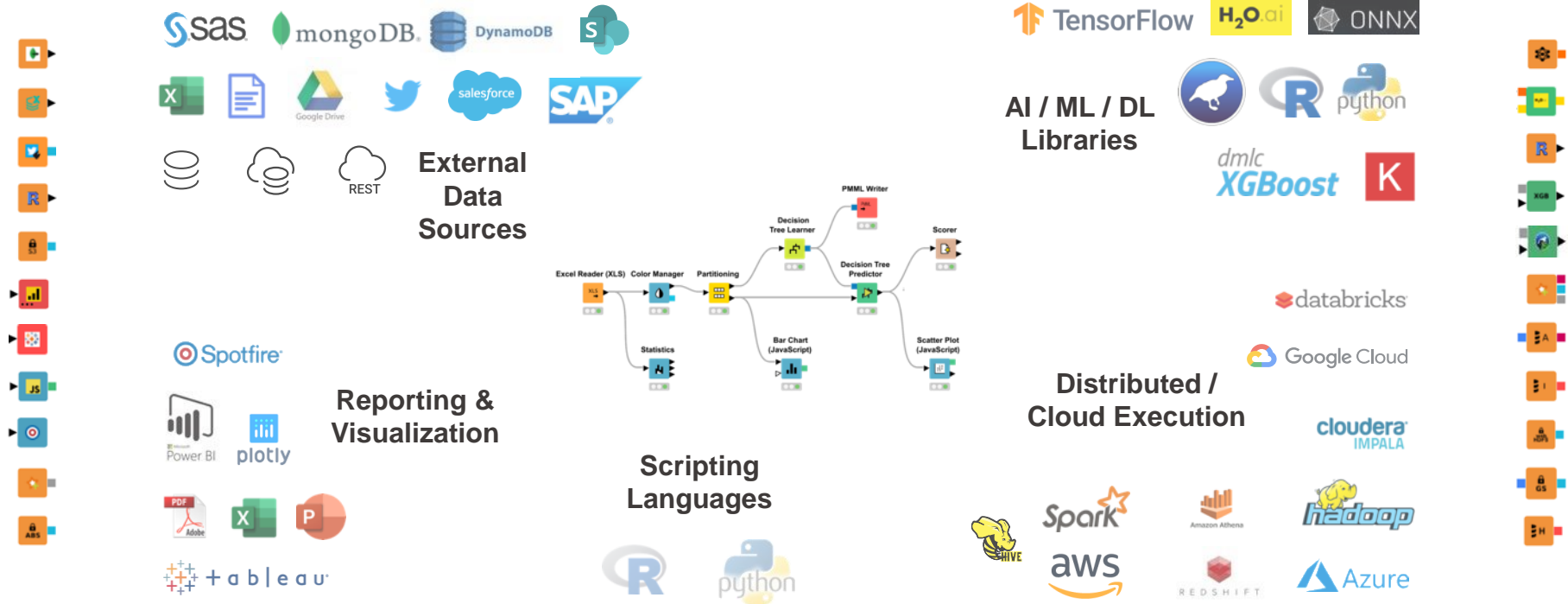
- Select output port (last item) to inspect execution results



4000+ Nodes for all Steps of End-To-End Data Science

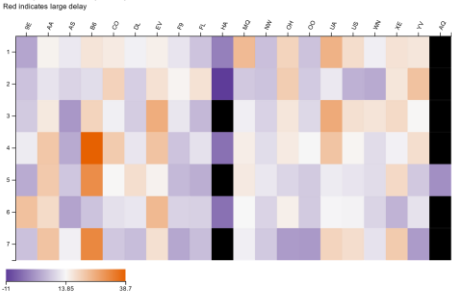


Mix & Match Data Sources, Technologies, and Execution

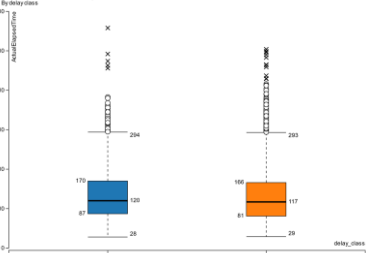


Data Visualization

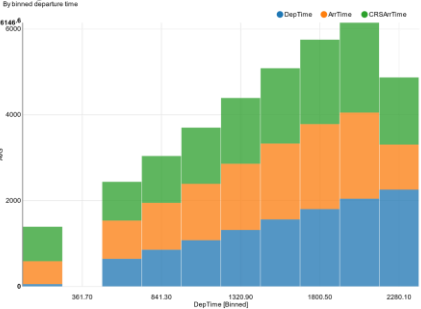
Delay Times by Day of the Week and Carrier



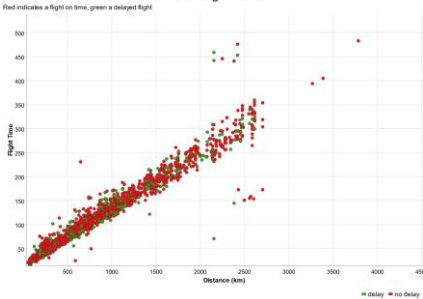
Distribution of Elapsed Time



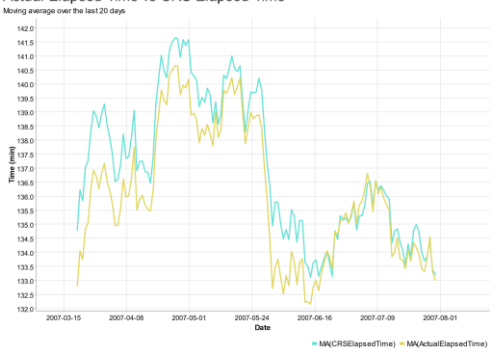
Average Arrival Time and CRS Arrival Time



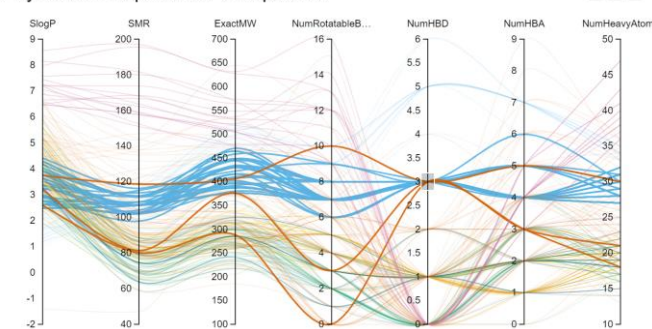
Correlation between Distance and Flight Time



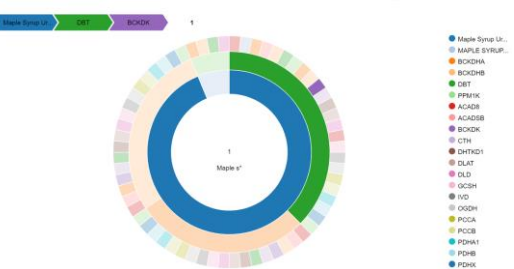
Actual Elapsed Time vs CRS Elapsed Time



PhysChem Properties of Compounds



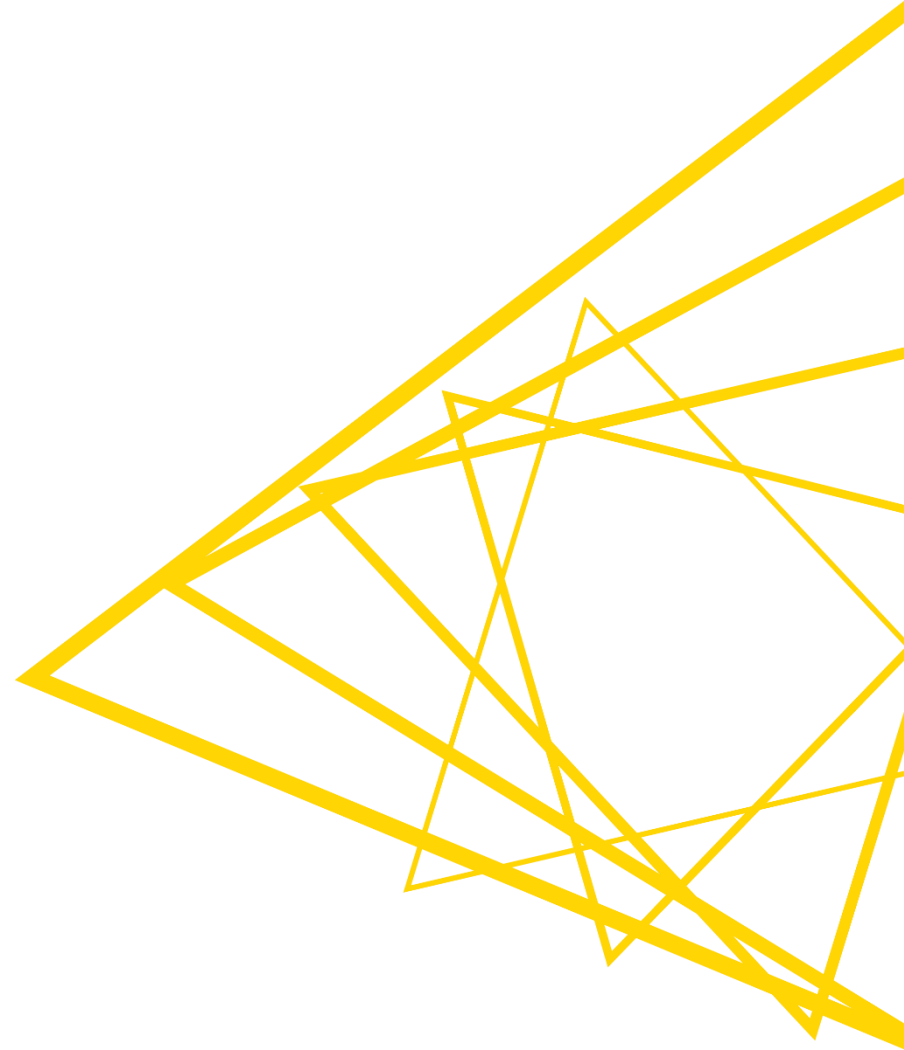
Interactions



Why KNIME?

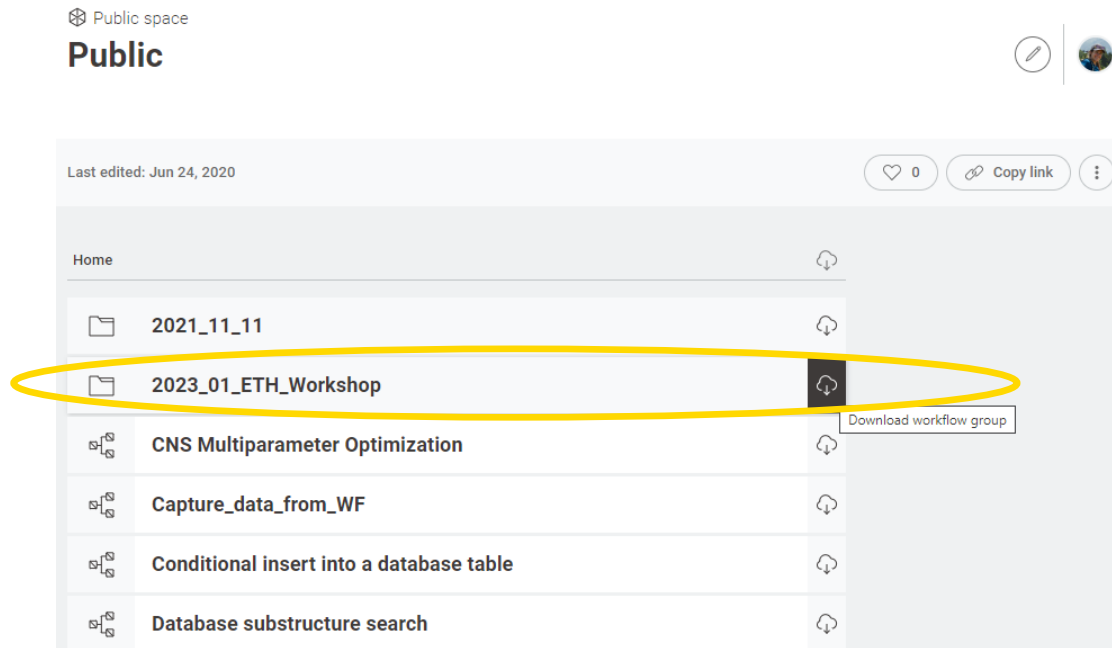
- Self-documenting
 - Every workflow is a track of what you did to your data
- Reproducible
 - Once configured, it will run the same way every time and have the same results
- Data is available at every manipulation step
 - Trace and control how the data table is changing throughout the workflow
- Interactive data visualization
- Chemistry capabilities

Setup KNIME AP for today



Set up KNIME Analytics Platform

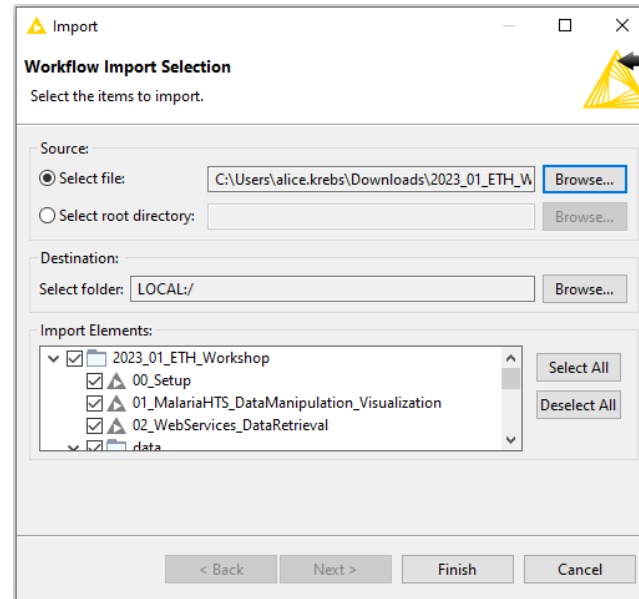
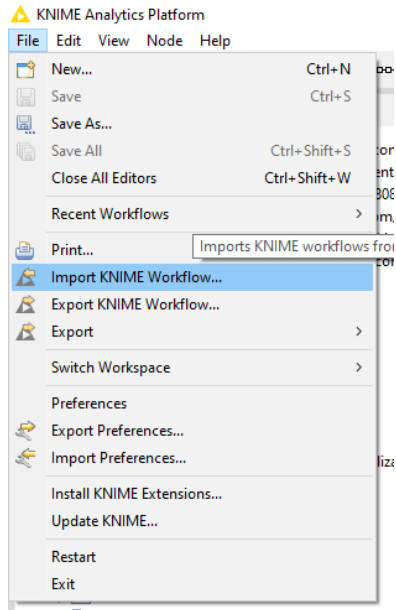
- Download workflows from the KNIME Community Hub
 - <https://kni.me/s/5Goi-yKrwMLzAU4o>



Import the workflow group

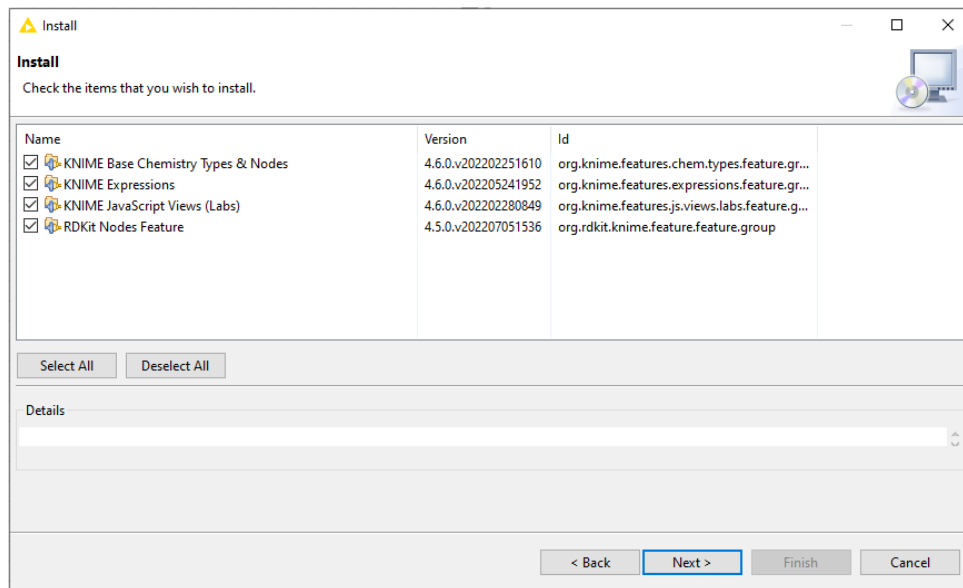
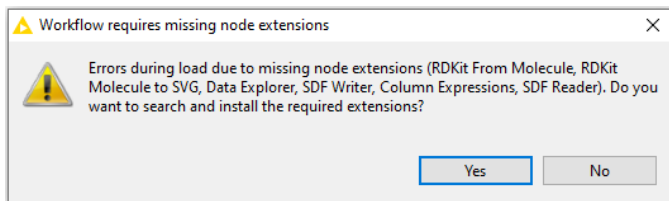
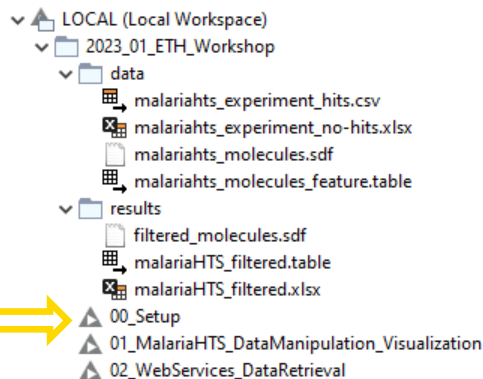
- Navigate to your download folder and choose the “2023_01_ETH_Workshop.knar” file

.knwf → single workflow
.knar → group of workflows



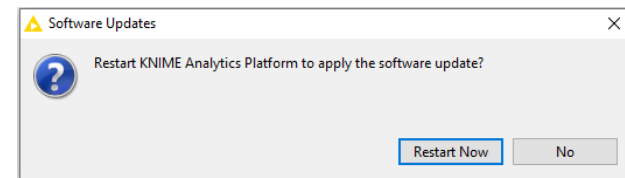
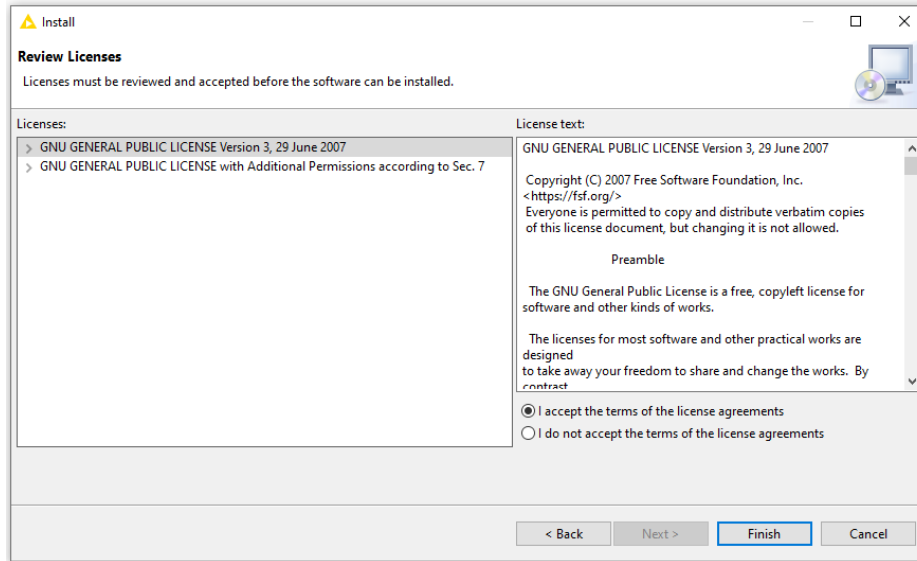
Getting started

- Open the “00_Setup” workflow to install all the extensions you need today

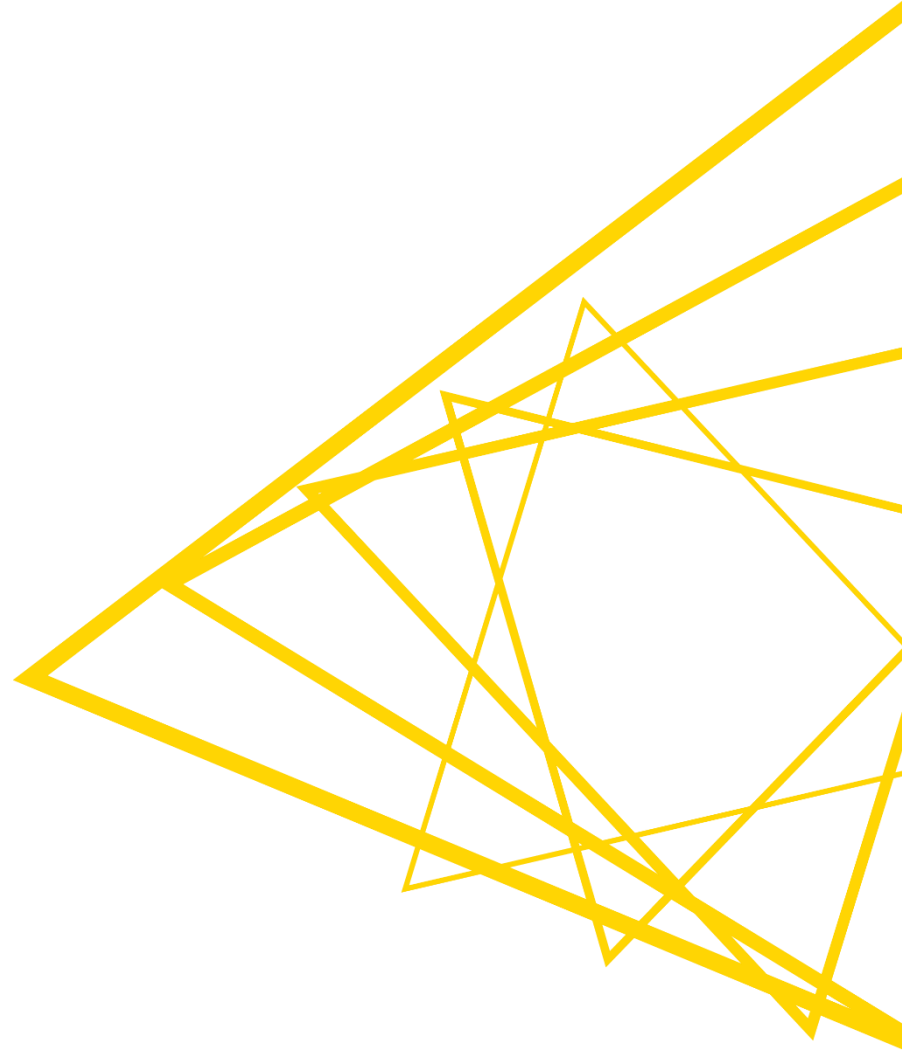


Getting started

- Open the “00_Setup” workflow to install all the extensions you need today

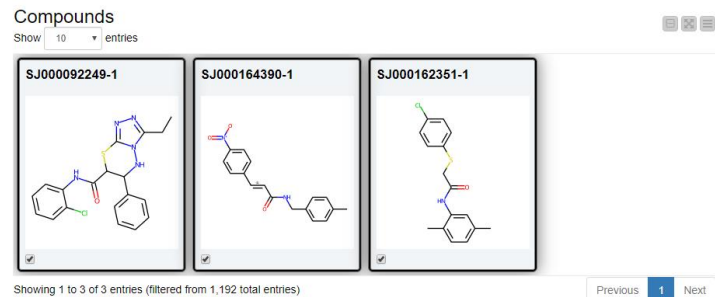
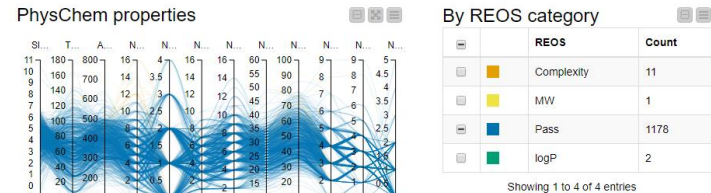
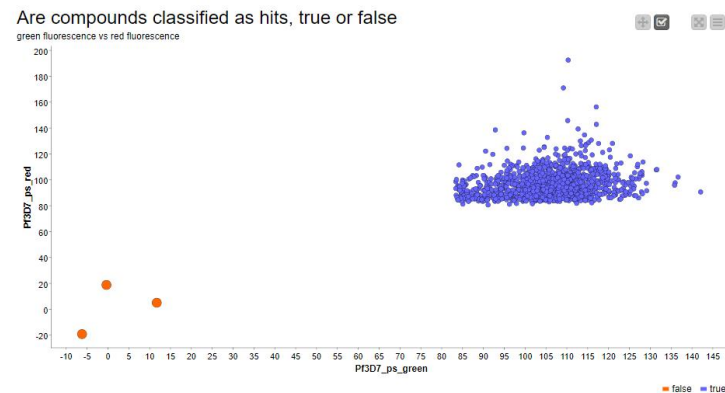
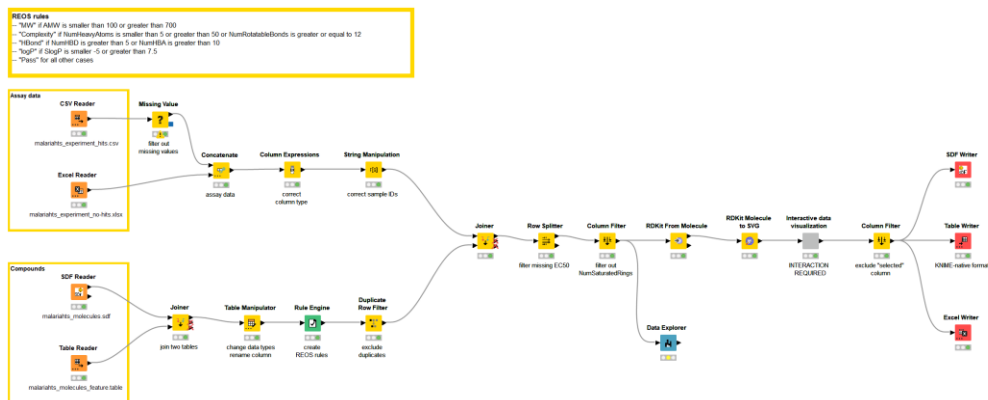


Ready to go!



- 01_MalariaHTS_DataManipulation_Visualization

- 01_MalariaHTS_DataManipulation_Visualization

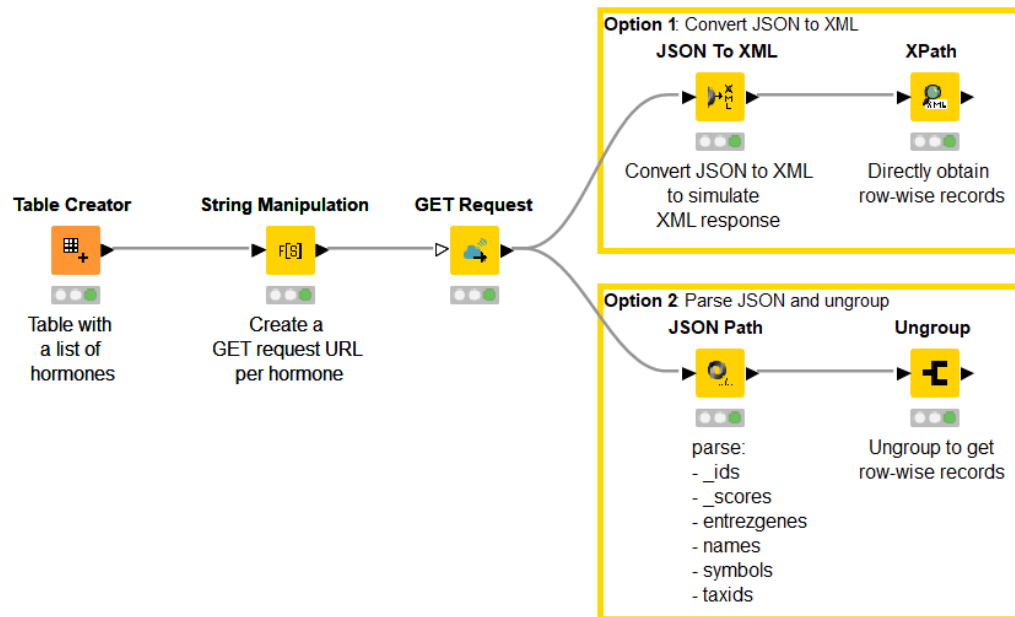


Today's workflows II

02_WebServices_DataRetrieval

In this workflow snippet, we will use the REST service provided by MyGene.info to obtain a list of human genes related to specific hormones. Then, we parse the JSON response into a table that is easy to read.

Sometimes, responses come in XML format. We have also included a way to parse XML responses by first converting the JSON response directly to XML.



01_MalariaHTS_DataManipulation_Visualization

■ Step 1: import data from 4 different sources

Assay data

CSV Reader

malariahts_experiment_hits.csv

Excel Reader

malariahts_experiment_no-hits.xlsx

Compounds

SDF Reader

malariahts_molecules.sdf

Table Reader

malariahts_molecules_feature.table

File Table - 4:276 - CSV Reader (malariahts_experiment_hits.csv)

File Edit Hilite Navigation View

Table "default" - Rows: 1528 Spec - Columns: 5 Properties Flow Variables

Row ID	[S] Sample	[D] Pf3D7_ps_green	[D] Pf3D7_ps_red	[S] Pf3D7_ps_hit	[D] Pf3D7_pEC50
Row0	SJ000259230-1	102.523	80.893	true	?
Row1	sj000282539-1	113.112	88.908	true	6.227
Row2	SJ000033142-1	122.304	104.185	true	6.069
Row3	sj000079671-1	105.88	97.726	true	4.824
Row4	sj000179372-1	98.684	84.861	true	6.142
Row5	sj000276817-1	98.44	86.477	true	5.14
Row6	SJ000273047-1	100.665	80.332	true	?
Row7	sj000260256-1	102.336	96.501	true	4.824
Row8	sj000123502-1	113.261	101.839	true	5.433
Row9	SJ000170548-1	107.117	111.154	true	5.416
Row10	SJ000092590-1	105.653	98.151	true	6.121
Row11	SJ000033131-1	116.077	96.145	true	7.284
Row12	sj000257328-1	111.211	122.335	true	?
Row13	sj000117911-1	114.424	84.23	true	6.239
Row14	sj000018305-1	82.184	89.21	true	?
Row15	SJ000217742-1	100.72	94.014	true	5.64
Row16	sj000128935-1	98.061	90.995	true	5.565
Row17	sj000114920-1	114.096	87.336	true	5.272
Row18	sj000225626-1	111.027	93.789	true	5.309

File Table - 4:275 - Excel Reader (malariahts_experiment_no-hits...)

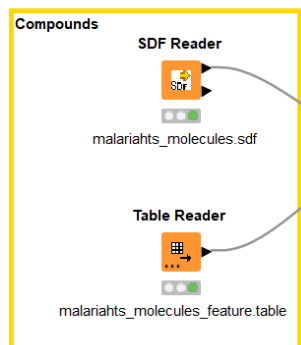
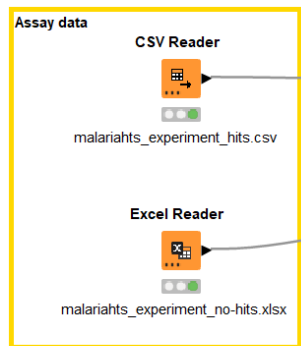
File Edit Hilite Navigation View

Table "default" - Rows: 10000 Spec - Columns: 5 Properties Flow Variables

Row ID	[S] Sample	[D] Pf3D7_ps_green	[D] Pf3D7_ps_red	[S] Pf3D7_ps_hit	[S] Pf3D7_pEC50
Row0	SJ000167487-1	-4.5	1.987	false	?
Row1	SJ000072922-1	-16.842	-0.682	false	?
Row2	SJ000223430-1	1.715	2.431	false	?
Row3	SJ000147674-1	2.178	11.964	false	?
Row4	sj000127323-1	-11.221	0.67	false	?
Row5	SJ000018465-1	5.524	5.319	false	?
Row6	sj000178506-1	-4.619	-6.703	false	?
Row7	SJ000007934-1	8.507	0.532	false	?
Row8	SJ000072074-1	11.421	1.934	false	?
Row9	sj000105546-1	-7.541	-3.692	false	?
Row10	sj000066091-1	10.913	0.62	false	?
Row11	sj000301657-1	15.532	-1.949	false	?
Row12	SJ000248175-1	-7.671	-0.126	false	?
Row13	SJ000183613-1	-4.996	3.268	false	?
Row14	SJ000172736-1	-8.381	-27.308	false	?
Row15	sj000291738-1	-4.754	9.031	false	?
Row16	sj000242990-1	-7.57	-10.372	false	?

01_MalariaHTS_DataManipulation_Visualization

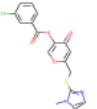


■ Step 1: import data from 4 different sources



Read molecules - 4:267 - SDF Reader (malariahts_molecules.sdf)

File Edit Hilite Navigation View

Table "default" - Rows: 11528 Spec - Columns: 6 Properties Flow Variables

Row ID	sdf Molecule	[S] Sample	[D] SlogP	[D] TPSA	[D] ExactMW	[D] AverageMolecularWeight
SJ000263455-1		SJ000263455-1	3.538	74.33	376.028	376.821
SJ000136517-1		SJ000136517-1	2.78	71.95	307.045	307.4
SJ000244874-1		SJ000244874-1	5.721	71.53	477.118	477.611

Read table - 4:7 - Table Reader (malariahts_molecules_feature.t...)

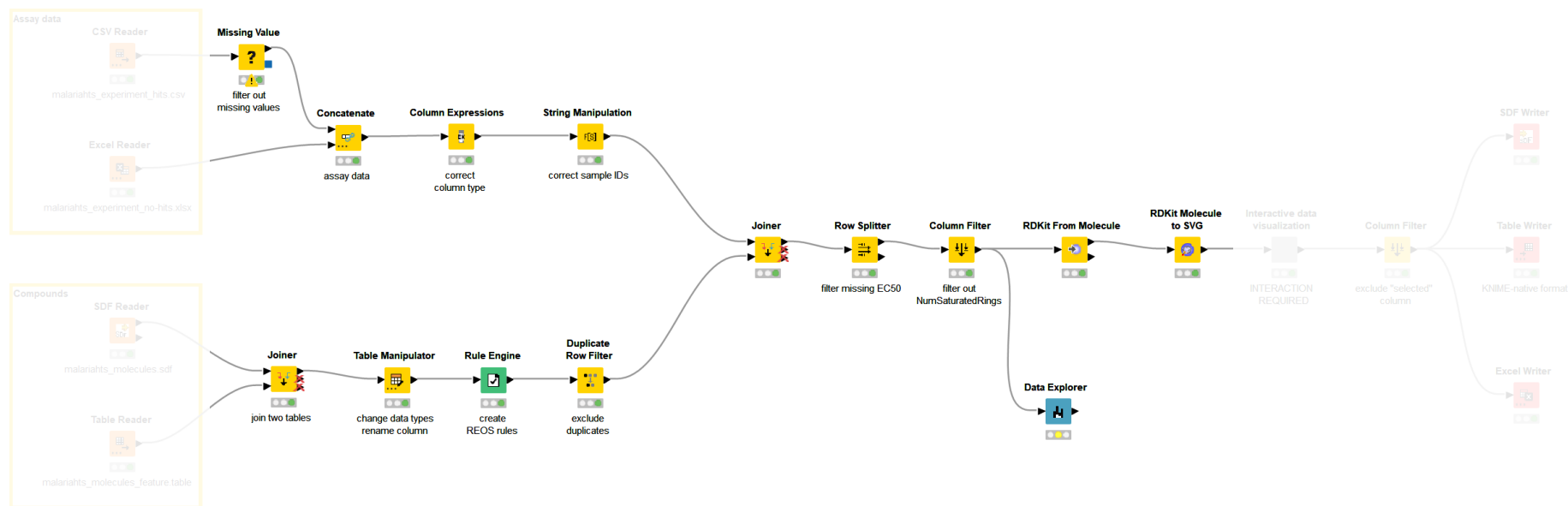
File Edit Hilite Navigation View

Table "default" - Rows: 11628 Spec - Columns: 11 Properties Flow Variables

Row ID	[S] Sample	[S] NumRotatableBonds	[I] NumHBD	[I] NumHBA	[I] NumHeteroAtoms	[S] NumHeavyAtoms	[I] NumAtoms	[I] NumRings	[I] NumAromaticRings	[I] NumSaturatedRings	[I] NumAliphaticRings
Row0	SJ000263455-1	5	0	7	8	25	38	3	3	0	0
Row1	SJ000136517-1	5	1	6	7	20	33	2	2	0	0
Row2	SJ000244874-1	5	1	6	8	33	56	5	4	0	1
Row3	SJ000185650-1	7	0	9	10	27	43	2	1	0	1
Row4	SJ000295142-1	2	0	3	5	21	39	3	1	1	2
Row5	SJ000007375-1	5	1	4	6	28	45	4	4	0	0
Row6	SJ000243289-1	6	2	6	7	24	45	3	3	0	0
Row7	SJ000116582-1	4	1	4	8	30	52	4	3	1	1
Row8	SJ000293545-1	3	2	4	7	27	39	4	4	0	0
Row9	SJ000057137-1	6	1	4	7	25	42	3	3	0	0
Row10	SJ000079475-1	4	1	7	10	28	40	4	4	0	0
Row11	SJ000038023-1	5	2	4	10	27	46	2	1	0	1
Row12	SJ000055177-1	7	0	5	7	34	36	3	2	0	1

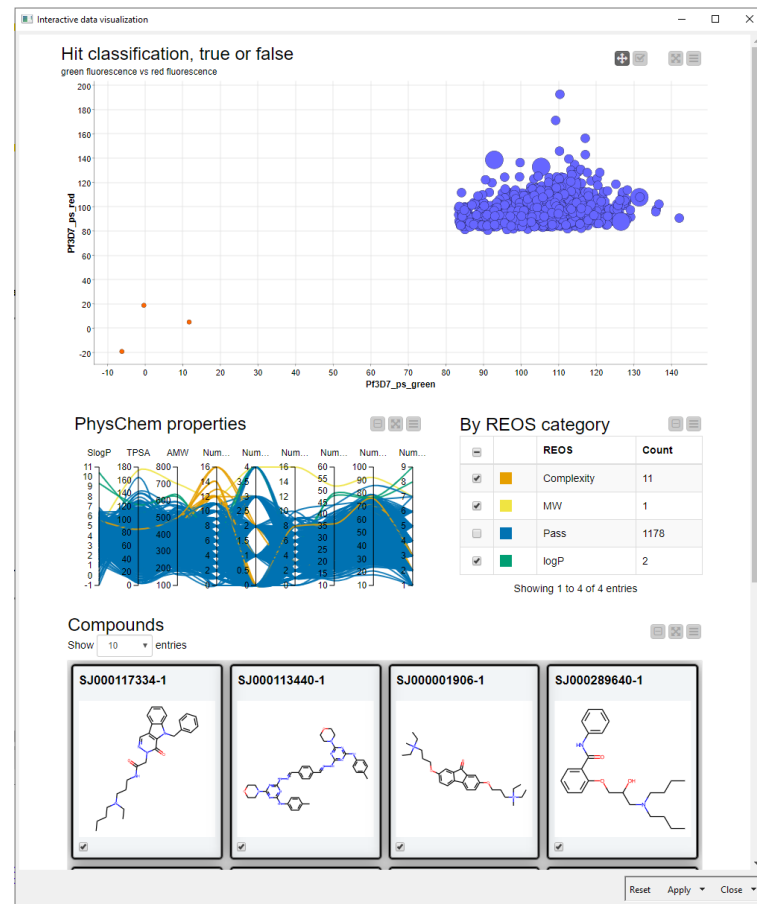
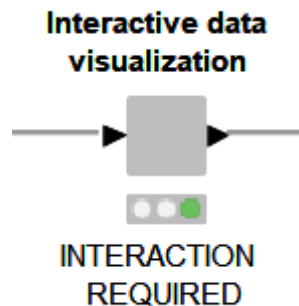
01_MalariaHTS_DataManipulation_Visualization

- Step 2:** data cleaning, merging datasets, manipulate data, create classification



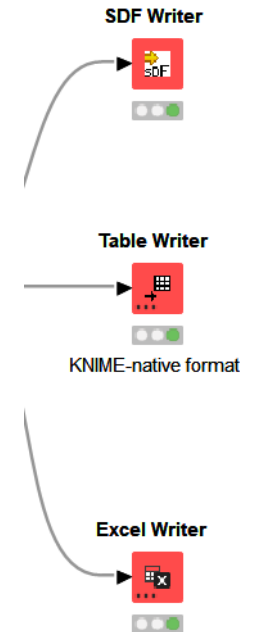
01_MalariaHTS_DataManipulation_Visualization

- **Step 3:** visualize and interactively select data
 - build a component to combine different plots in one view



01_MalariaHTS_DataManipulation_Visualization

- **Step 4:** Export the data
 - SD file
 - KNIME-native table (only re-usable in KNIME)
 - Excel



Content, theory, background on...

- Import data
 - Local file system
 - Workflow relative path
- Merging data sets
 - Concatenate
 - Join
- Exclude data
 - Filter
 - Splitter
- Handle missing values and duplicates
- Change data
 - Change assigned data type
 - Capitalize the letters in the sample ID
 - Rename columns
 - Create classification of REOS rules
- Visualize data in components

Import data

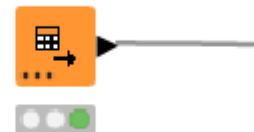
- Different Reader nodes for different file formats
- “Table” is a KNIME native format
- Drag and drop the files from the data folder in the KNIME Explorer

Excel Reader



malariahts_experiment_no-hits.xlsx

CSV Reader



malariahts_experiment_hits.csv

SDF Reader



malariahts_molecules.sdf

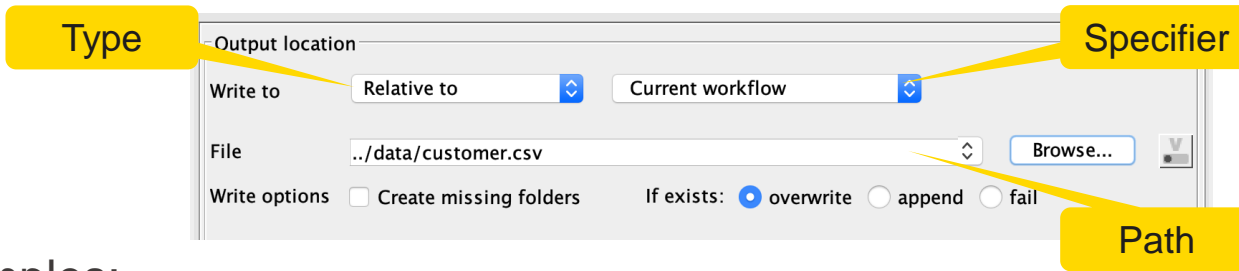
Table Reader



malariahts_molecules_feature.table

Common Settings: File Path

- A path consists of three parts:
 - **Type**: Specifies the file system type e.g. local, relative, mountpoint, custome_url or connected.
 - **Specifier**: Optional string with additional file system specific information e.g. relative to which location (knime.workflow)
 - **Path**: Specifies the location within the file system



- Examples:
 - (LOCAL, , C:\Users\username\Desktop)
 - (RELATIVE, knime.workflow, file1.csv)
 - (MOUNTPOINT, MOUNTPOINT_NAME, /path/to/file1.csv)
 - (CONNECTED, amazon-s3:eu-west-1, /mybucket/file1.csv)

Common Settings: Four Default File Systems

■ Local File System

Input location

Read from: Local File System

Mode: ☒ File ☐ Files in folder

File: /Users/kathrinmelcher/Desktop/course_data.csv Browse...

■ Relative to ...

Read from: Relative to

File: Calls_data.xlsx Browse...

- Current mountpoint
- Current workflow data area
- Current workflow

■ Mountpoint

Read from: Mountpoint LOCAL

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

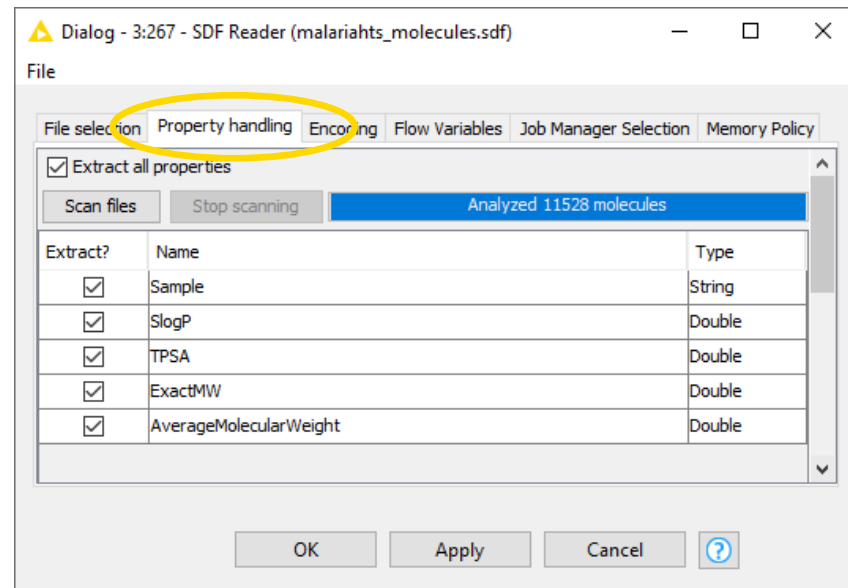
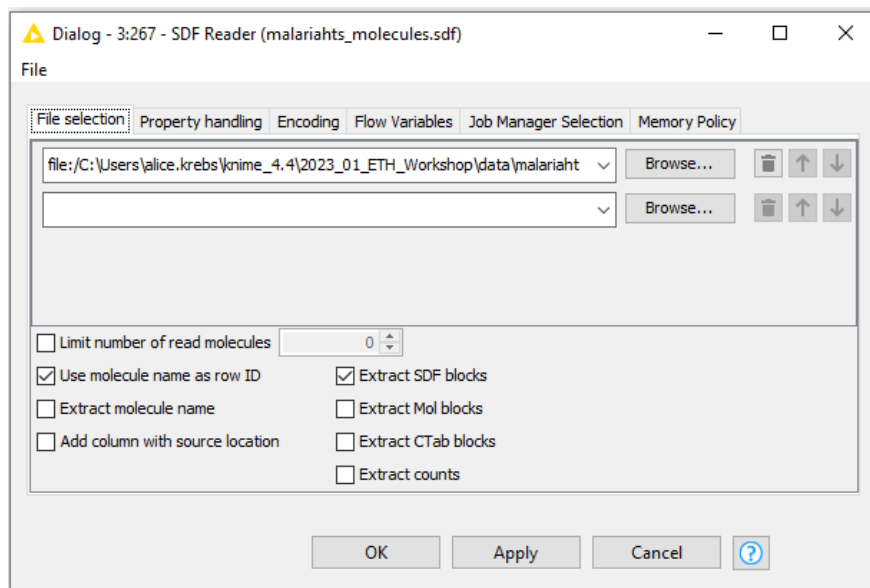
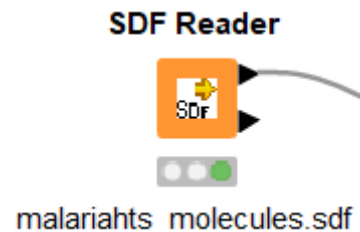
■ Custom URL

Read from: Custom URL

URL: knime://knime.workflow/data/Calls_data.xlsx Browse...

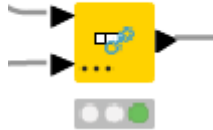
Special case – SDF reader node

- No new file handling
- Extract properties to get all data

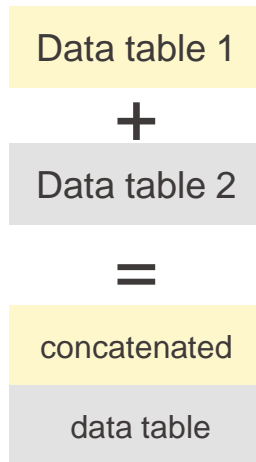


Concatenate vs. Join

Concatenate



- Simply links to tables, appends them underneath each other, like in a chain



Joiner

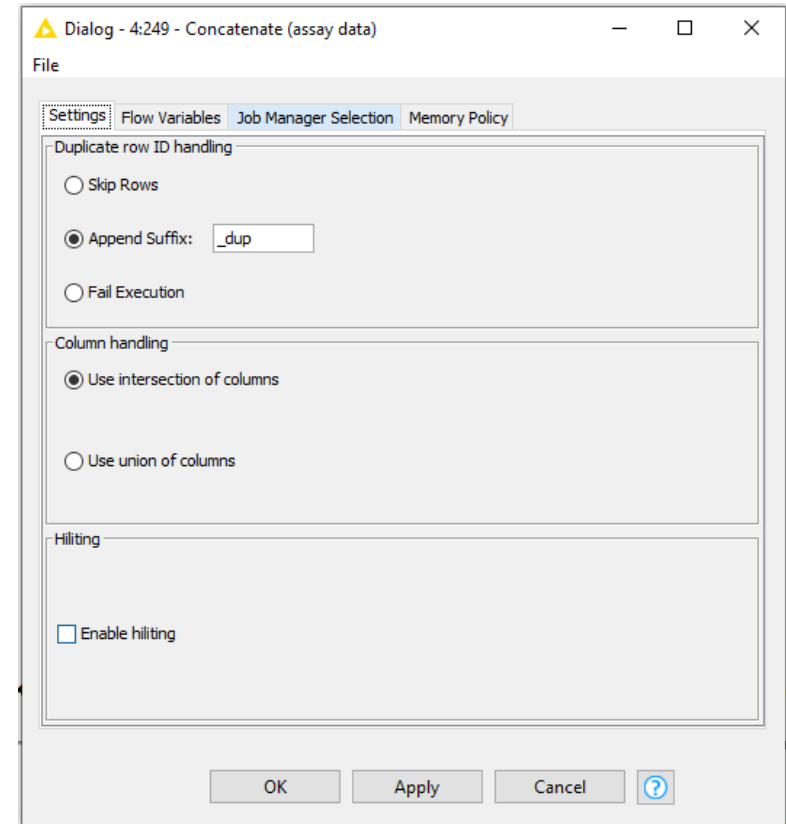
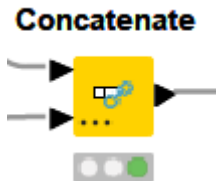


- Combines two tables row wise

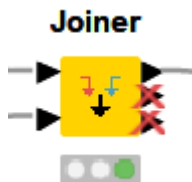


Concatenate node

- Intersection
 - use only the columns that appear in both input tables
- Union
 - use all columns available in the input tables



Joiner node



Dialog - 4:299 - Joiner

File

Joiner Settings | Column Selection | Performance | Flow Variables | Job Manager Selection | Memory Policy

Join columns

Match ☒ all of the following ☐ any of the following

Top Input (left table) Bottom Input (right table)

[S] Sample [S] Sample + -

Compare values in join columns by ☒ value and type ☐ string representation ☐ making integer types compatible

Include in output

☒ Matching rows

☐ Left unmatched rows

☐ Right unmatched rows

Inner join

Output options

☐ Split join result into multiple tables (top = matching rows, middle = left unmatched rows, bottom = right unmatched rows)

☐ Merge join columns

☐ Hiding enabled

Row Keys

☒ Concatenate original row keys with separator

☐ Assign new row keys sequentially

☐ Keep row keys

OK Apply Cancel ?

Dialog - 4:299 - Joiner

File

Joiner Settings | Column Selection | Performance | Flow Variables | Job Manager Selection | Memory Policy

Top Input (left table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

[S] Sample
[D] PF3D7_ps_green
[D] PF3D7_ps_red
[S] PF3D7_ps_hit
[D] PF3D7_pEC50

☐ Enforce inclusion

Bottom Input (right table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

[S] Sample

☒ Enforce exclusion

Include

Filter

[S] Molecule
[D] SlogP
[D] TPSA
[D] AMW
[I] NumRotatableBonds
[I] NumHBD
[I] NumHBA
[I] NumHeteroAtoms

☐ Enforce inclusion

Duplicate column names

☐ Do not execute

☒ Append custom suffix

OK Apply Cancel ?

Joining Columns of Data – Inner Join

Left Table

molregno	chembl_id	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Right Table

molregno	Ki_value	Ki_relation	Ki_unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM



Inner Join

molregno	chembl_id	SMILES	Ki_value	Ki_relation	Ki_unit
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Joining Columns of Data – Left Outer Join

Left Table

molregno	chembl_id	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Right Table

molregno	Ki_value	Ki_relation	Ki_unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM



Left Outer Join

molregno	chembl_id	SMILES	Ki_value	Ki_relation	Ki_unit
22	CHEMBL1794855	CCCN(CCC)	?	?	?
24	CHEMBL278751	CCN(C)	?	?	?
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Joining Columns of Data – Right Outer Join

Left Table

molregno	chembl_id	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Right Table

molregno	Ki_value	Ki_relation	Ki_unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM

Right Outer Join

molregno	chembl_id	SMILES	Ki_value	Ki_relation	Ki_unit
17	?	?	76.0	=	nM
65	?	?	6.56	=	nM
35	?	?	100	>	nM
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Joining Columns of Data – Full Outer Join

Left Table

molregno	chembl_id	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Right Table

molregno	Ki_value	Ki_relation	Ki_unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM

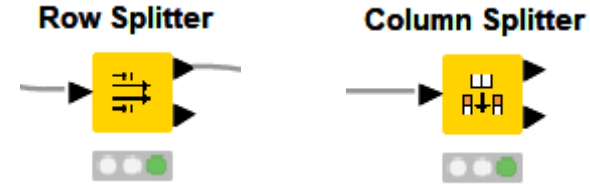
Values missing in the left table

Full Outer Join

molregno	chembl_id	SMILES	Ki_value	Ki_relation	Ki_unit
17	?	?	76.0	=	nM
65	?	?	6.56	=	nM
35	?	?	100	>	nM
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM
22	CHEMBL1794855	CCCN(CCC)	?	?	?
24	CHEMBL278751	CCN(C)	?	?	?

Values missing in the right table

Filter vs. Splitter



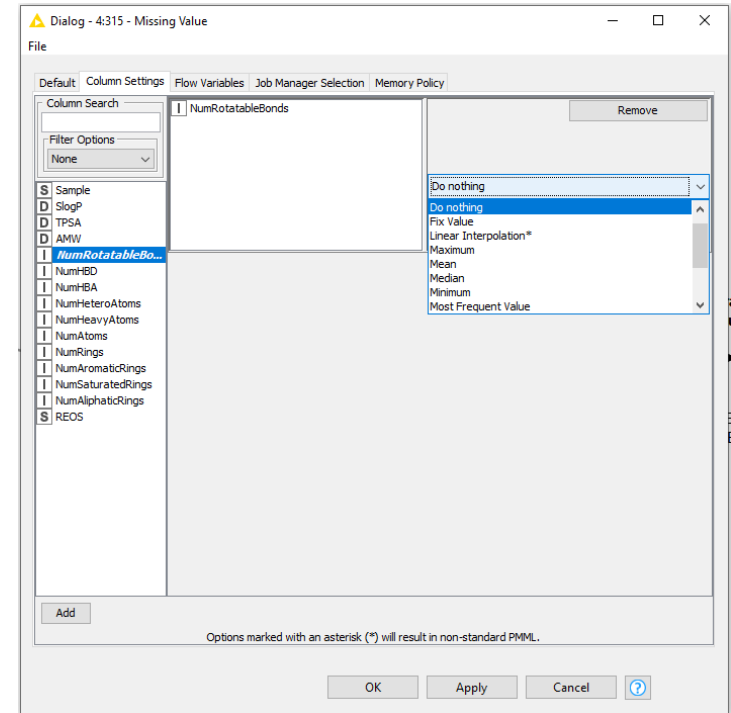
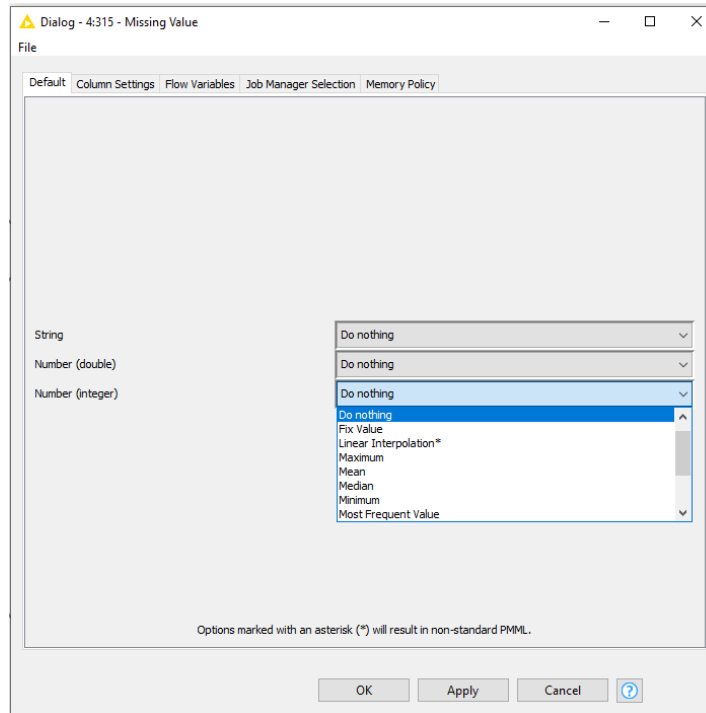
- User-defined criteria
- Only one output port
- Excluded rows or columns are not available for downstream processing

- User-defined criteria
- Two output ports
- Splits a dataset into two
- The 'excluded' data is available at the lower output port for downstream processing

Missing value handling

- First tab: define table-wide action depending on data type
- Second tab: define action for individual columns

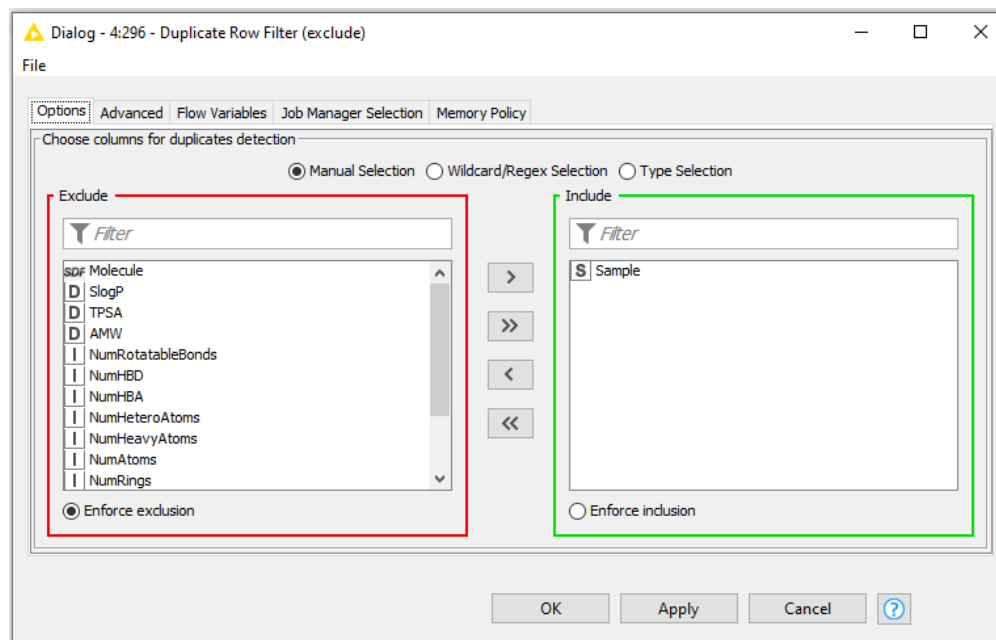
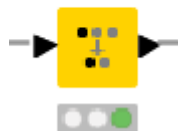
Missing Value



Exclude duplicates

- identifies duplicate rows
- duplicate rows have identical values in certain columns
- user defines the column(s) for duplicate detection

Duplicate Row Filter



Change assigned data type

- modify existing columns using expressions

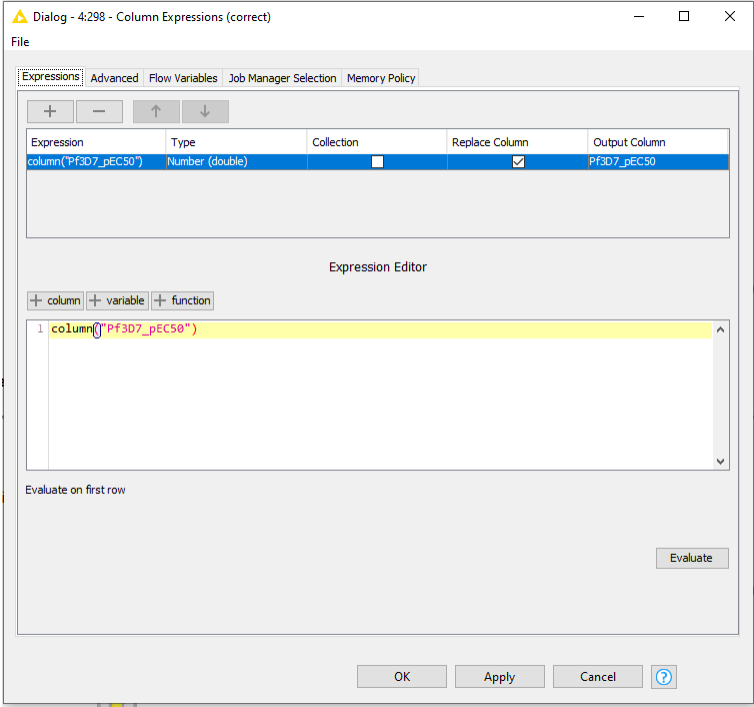


Table "default" - Rows: 11189						Spec - Columns: 5	Properties	Flow Variables
Row ID	[S] Sample	[D] Pf3D7_...	[D] Pf3D7_...	[S] Pf3D7_...	[?] Pf3D7_pEC50			
Row1	sj000282539-1	113.112	88.908	true	6.22650610772...			
Row2	SJ000033142-1	122.304	104.185	true	6.06905096883...			
Row3	sj000079671-1	105.88	97.726	true	4.82390874094...			
Row4	si000179372-1	98.684	84.861	true	6.14158312227...			



Table "default" - Rows: 11189						Spec - Columns: 5	Properties	Flow Variables
Row ID	[S] Sample	[D] Pf3D7_...	[D] Pf3D7_...	[S] Pf3D7_...	[D] Pf3D7_pEC50			
Row1	sj000282539-1	113.112	88.908	true	6.227			
Row2	SJ000033142-1	122.304	104.185	true	6.069			
Row3	sj000079671-1	105.88	97.726	true	4.824			
Row4	si000179372-1	98.684	84.861	true	6.142			

Capitalize the letters in the sample ID

- Manipulates strings
- Many different functions available
- Can also be used for type converting

String Manipulation

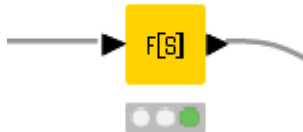
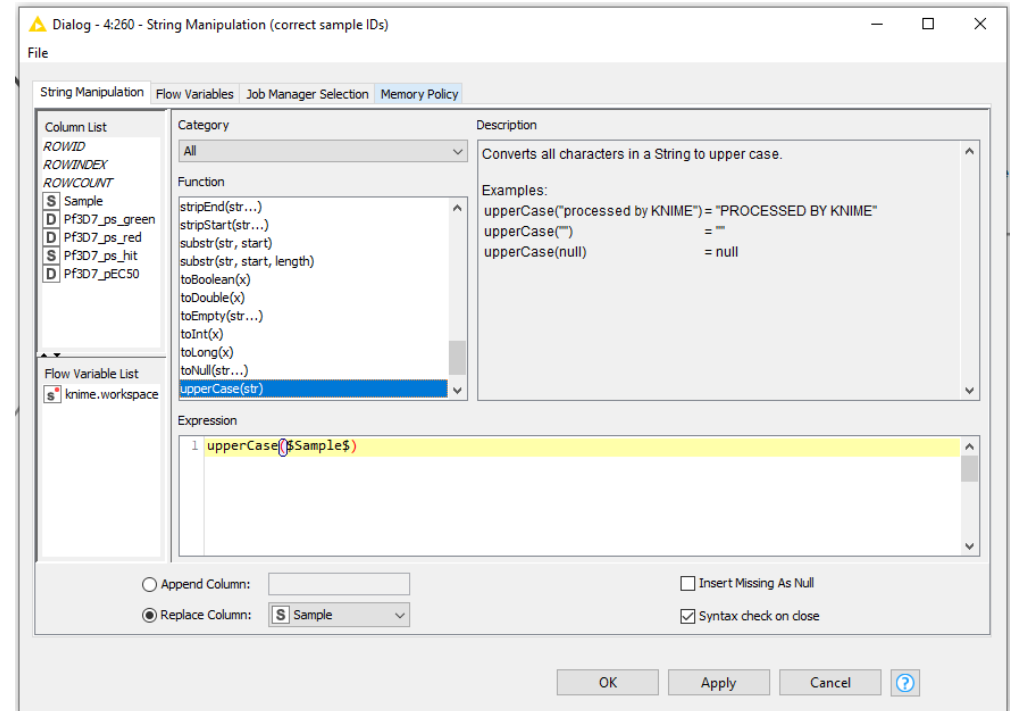


Table "default" - Rows: 1528 Spec -

Row ID	[S] Sample
Row0	SJ000259230-1
Row1	sj000282539-1
Row2	SJ000033142-1
Row3	sj000079671-1
Row4	sj000170377-1



Rename column

- Transformations of input columns
 - Renaming
 - Filtering
 - Re-ordering
 - Type changing
- Check data manipulation in the preview

Table Manipulator



Dialog - 4:278 - Table Manipulator (change data types)

File

Settings | Flow Variables | Job Manager Selection | Memory Policy

Row ID handling
☐ Use existing row ID ☐ Prepend table index to row ID

Transformations
Reset actions ↑ Move up ↓ Move down ☒ Enforce types Take columns from: ☒ Union ☐ Intersection

	Column	New name	Type
<input checked="" type="checkbox"/>	Molecule		sdf SDF
<input checked="" type="checkbox"/>	Sample		S String
<input checked="" type="checkbox"/>	SlogP		D Number (double)
<input checked="" type="checkbox"/>	TPSA		D Number (double)
<input type="checkbox"/>	ExactMW		D Number (double)
<input checked="" type="checkbox"/>	AverageMolecularWeight	AMW	D Number (double)
<input checked="" type="checkbox"/>	NumRotatableBonds		I String → Number (integer)
<input checked="" type="checkbox"/>	NumHBD		I Number (integer)
<input checked="" type="checkbox"/>	NumHBA		I Number (integer)
<input checked="" type="checkbox"/>	NumHeteroAtoms		I Number (integer)
<input checked="" type="checkbox"/>	NumHeavyAtoms		I String → Number (integer)
<input checked="" type="checkbox"/>	NumAtoms		I Number (integer)
<input checked="" type="checkbox"/>	NumRings		I Number (integer)
<input checked="" type="checkbox"/>	NumAromaticRings		I Number (integer)
<input checked="" type="checkbox"/>	NumSaturatedRings		I Number (integer)
<input checked="" type="checkbox"/>	NumAliphaticRings		I Number (integer)

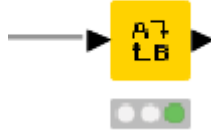
Preview
✔ Data analysis successfully completed.

Row ID	sdf Molecule	S Sample	D SlogP	D TPSA	D AMW	I NumRo...	I NumHBD	I NumHBA	I Num
Row0		S3000263455-1	3.538	74.33	376.821	5	0	7	8
Row1		S3000136517-1	2.78	71.95	307.4	5	1	6	7
Row2		S3000244874-1	5.721	71.53	477.611	5	1	6	8

OK Apply Cancel ?

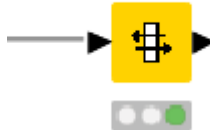
There isn't just one way of doing it...

Column Rename



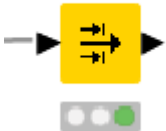
- Rename columns

Column Resorter

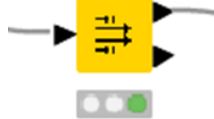


- Change the column order

Row Filter

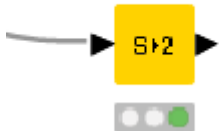


Row Splitter

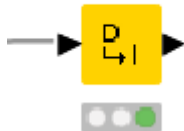


- Exclude missing values

String To Number

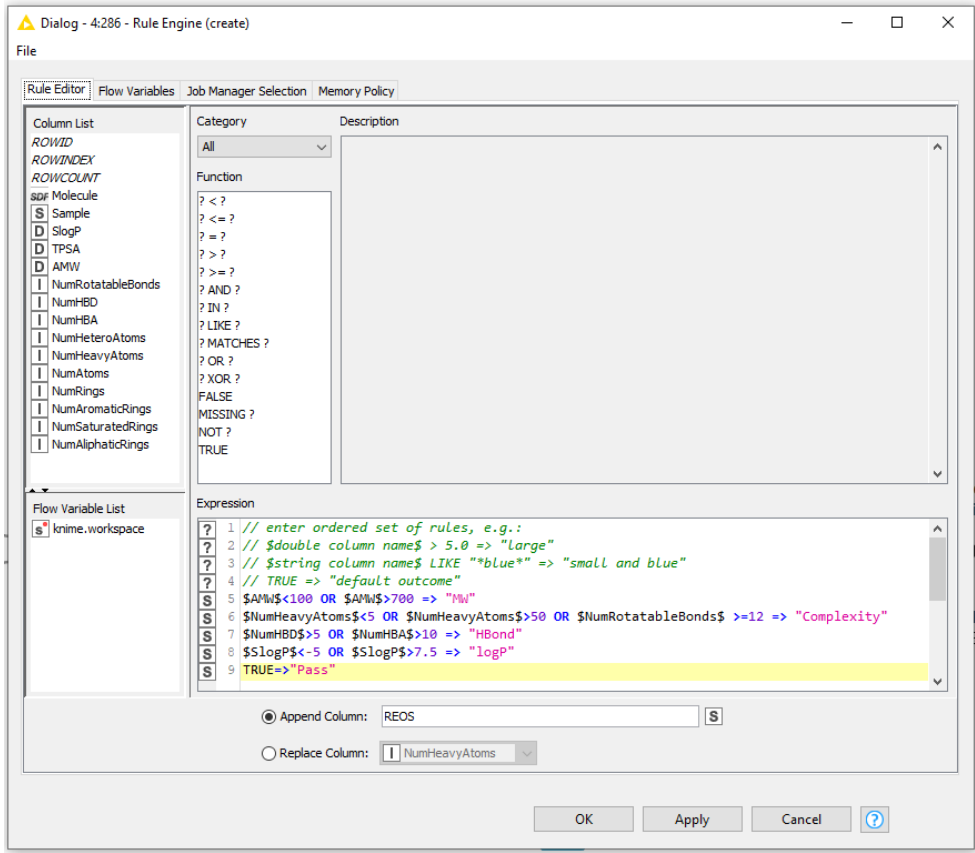
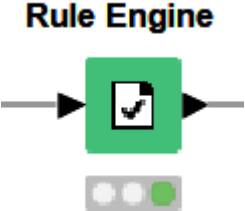


Double To Int



- Convert data types

Create classification REOS rules



*Define
Condition => Outcome
if condition
is met*


If not do this

IF... THEN... ELSE

GroupBy node

- Aggregate rows to summarize data
- Different aggregation methods available

Product ID	REOS	AMW
P001	Pass	462.55
P002	Pass	397.92
P003	Complexity	431.28
P004	Pass	431.28
P005	MW	371.46
P006	logP	389.20




Group	Count*(AMW)
Pass	3
Complexity	1
MW	1
logP	1

GroupBy node

- Aggregate rows to summarize data
- Different aggregation methods available

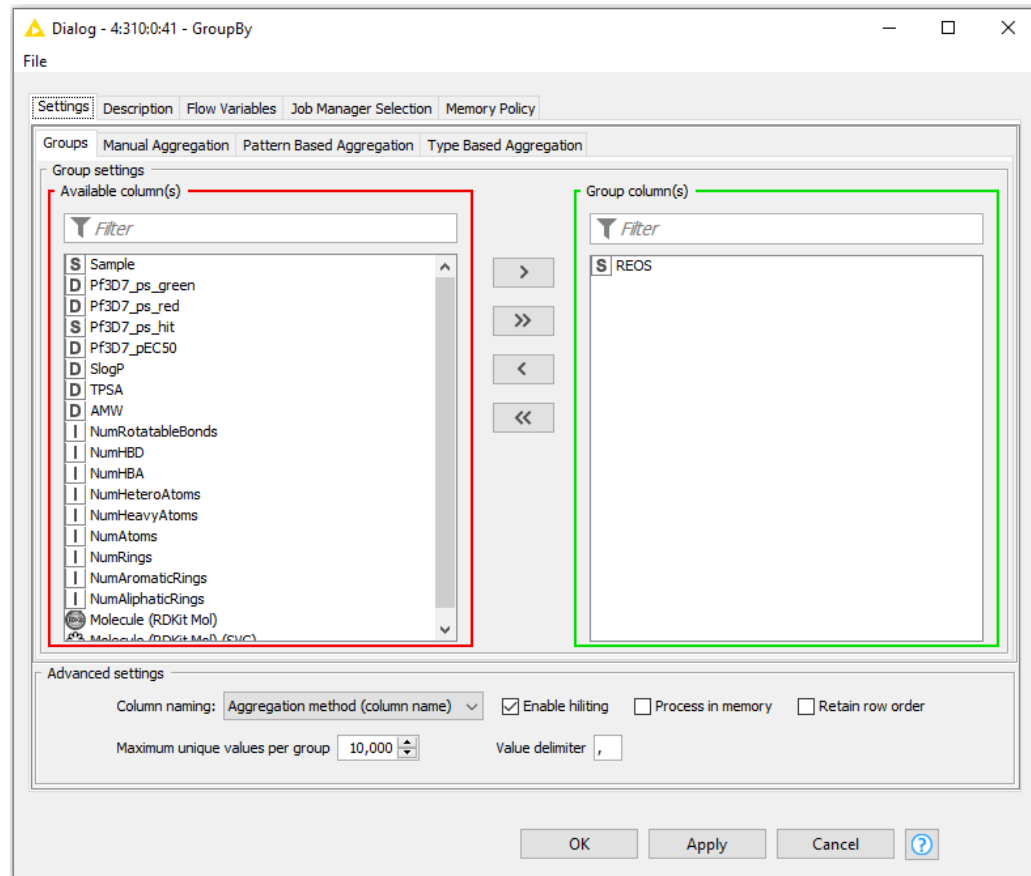
Product ID	REOS	AMW
P001	Pass	462.55
P002	Pass	397.92
P003	Complexity	431.28
P004	Pass	431.28
P005	MW	371.46
P006	logP	389.20



Group	Mean(AMW)
Pass	430.58
Complexity	431.28
MW	371.46
logP	389.20

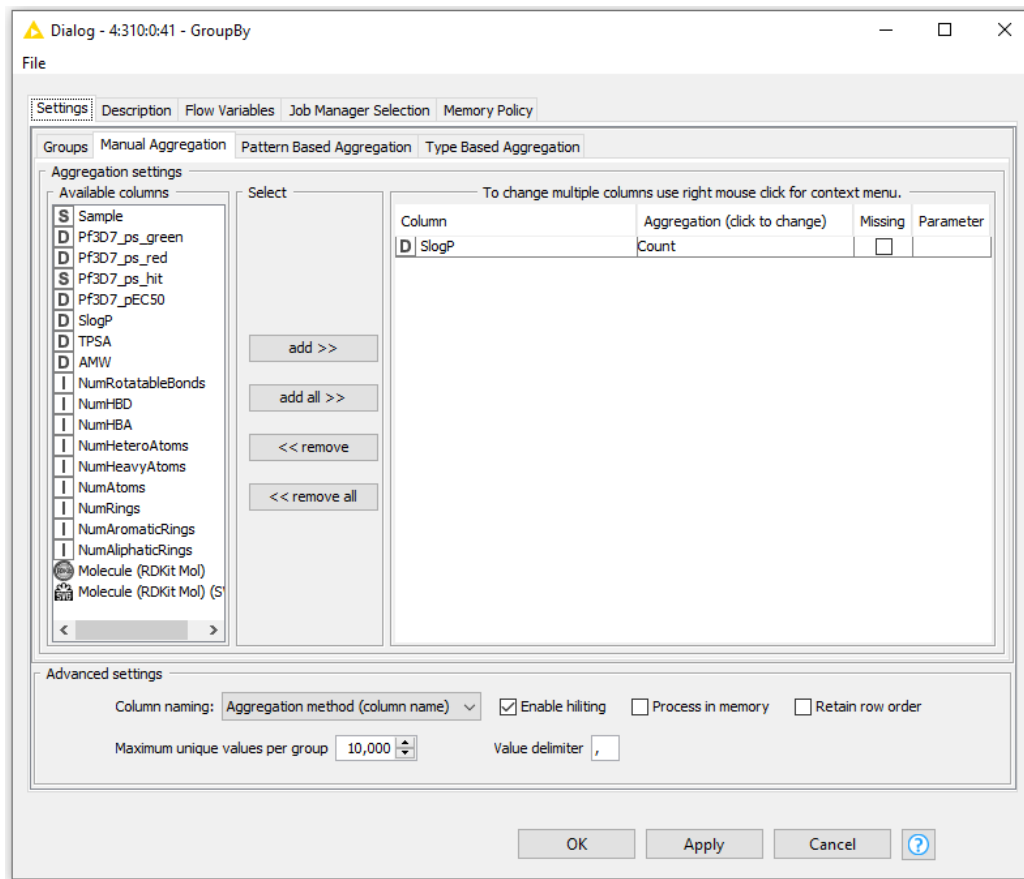
Groupby node configuration

- Define the column to group on in the first tab



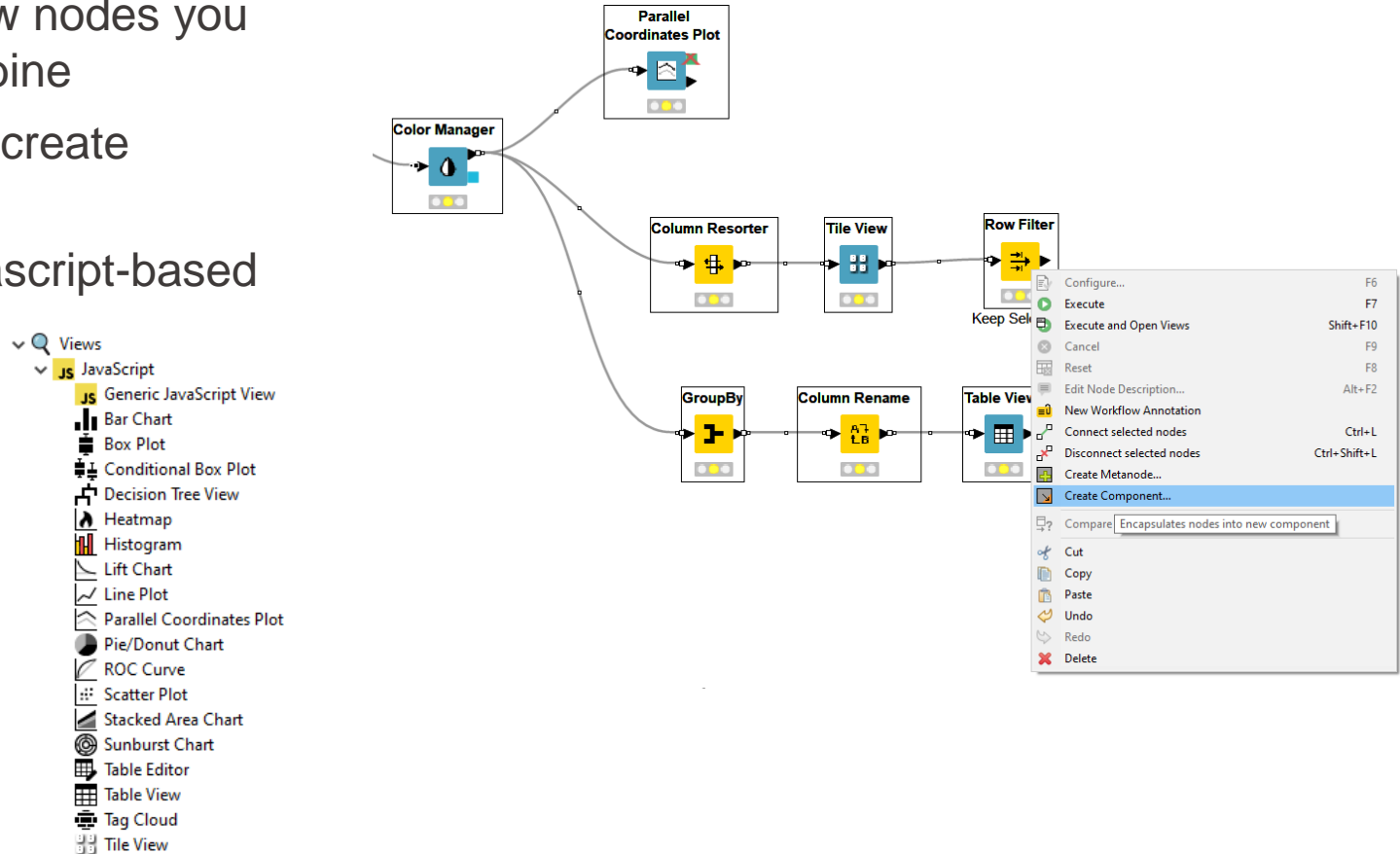
Groupby configuration

- Define the aggregation method in the second tab
- Information about the methods is provided in the Description tab



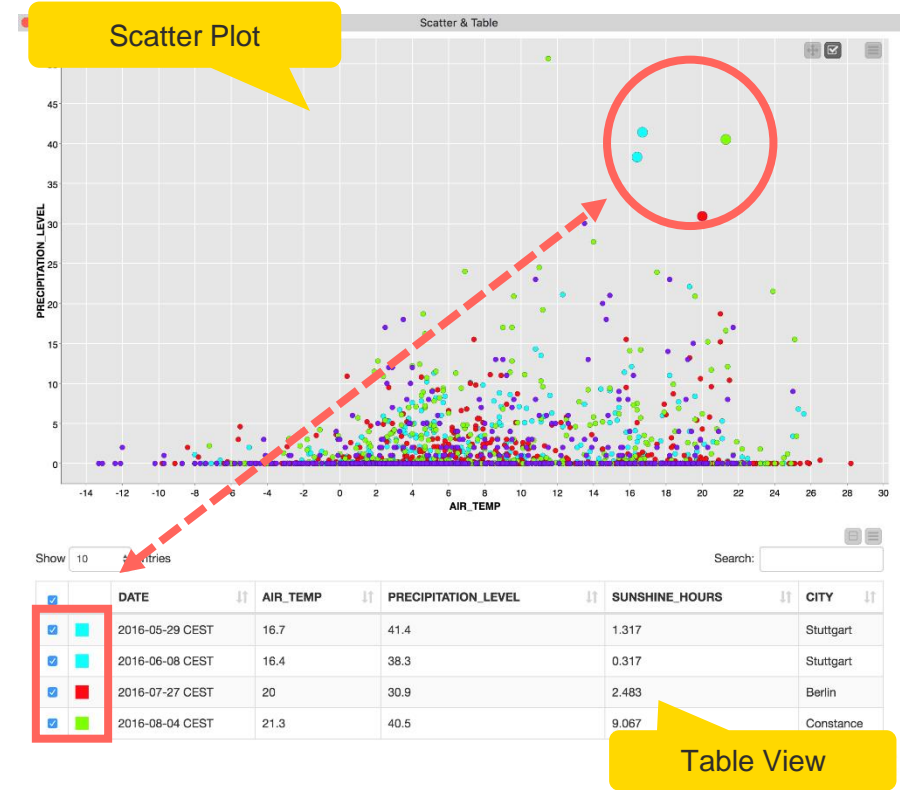
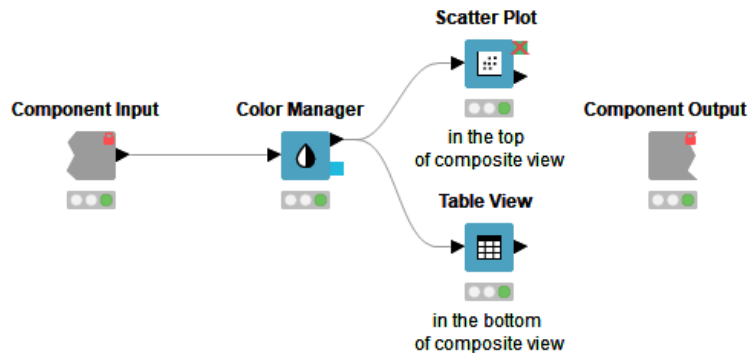
Visualization in components

- Mark the view nodes you want to combine
- Right-click > create component
- Use the Javascript-based nodes

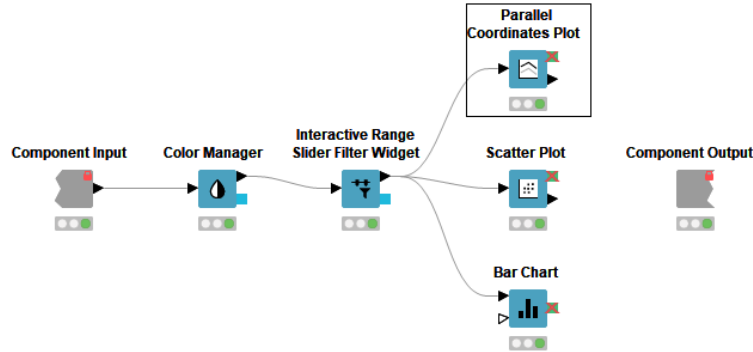


Components – Combined Views

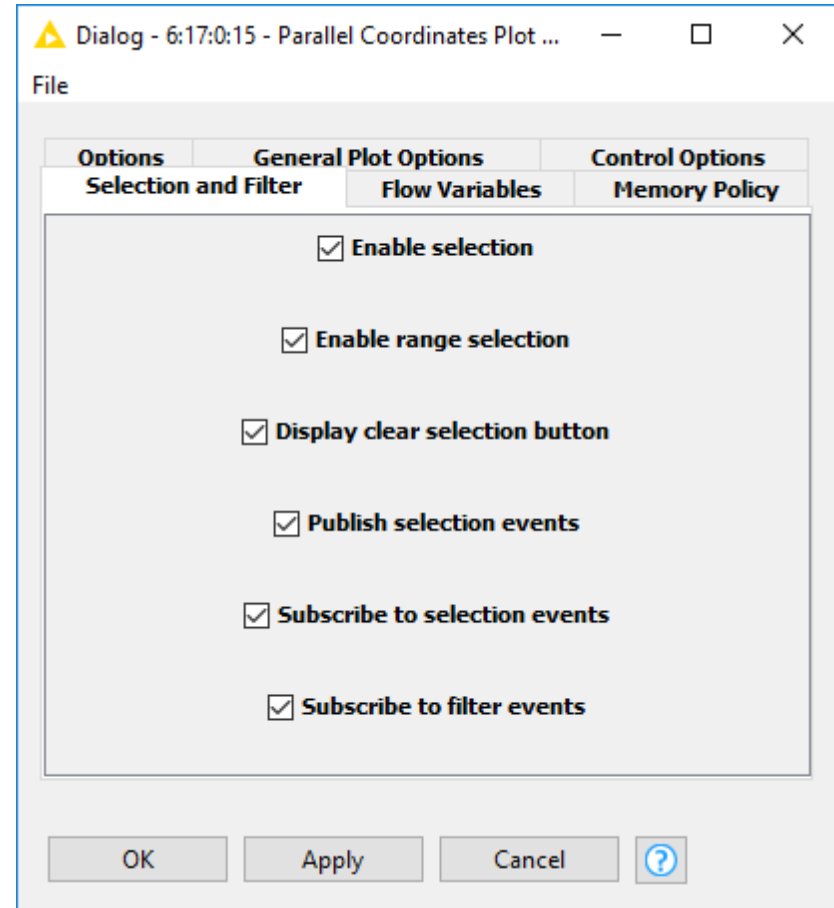
- Multiple JavaScript View nodes can be combined in components
- Selections are transmitted to all other views
- Also for use on the KNIME WebPortal (commercial product)



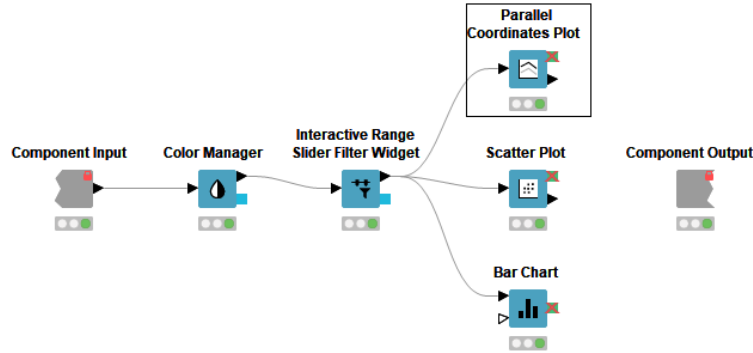
JS-based nodes: Selection and Filter Events



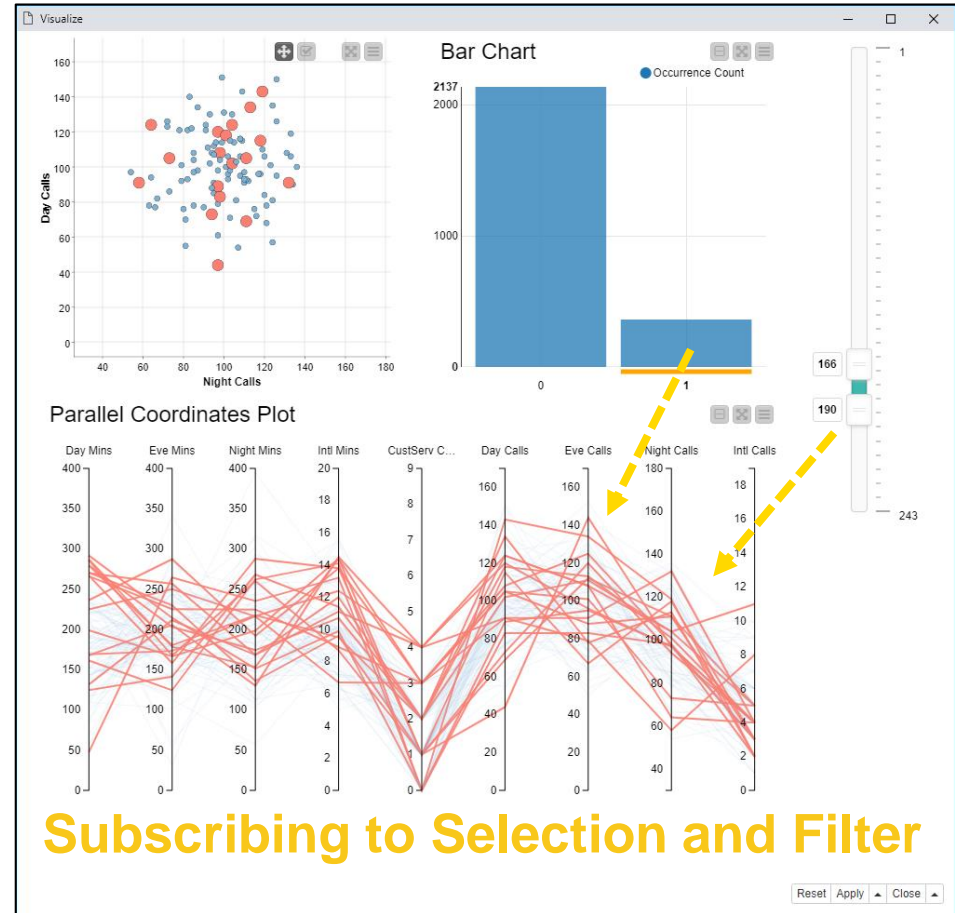
- Selection and Filter tab in many Javascript-based view nodes



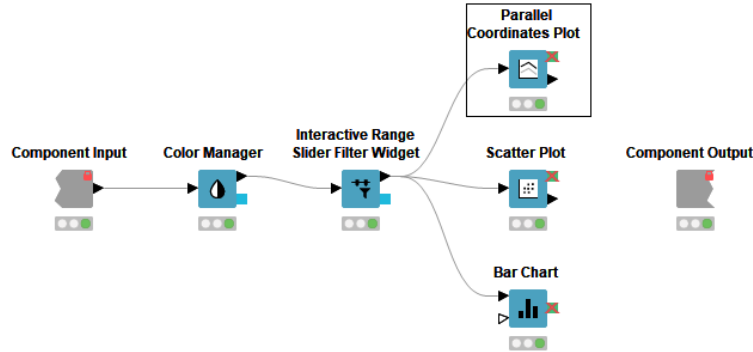
JS-based nodes: Selection and Filter Events



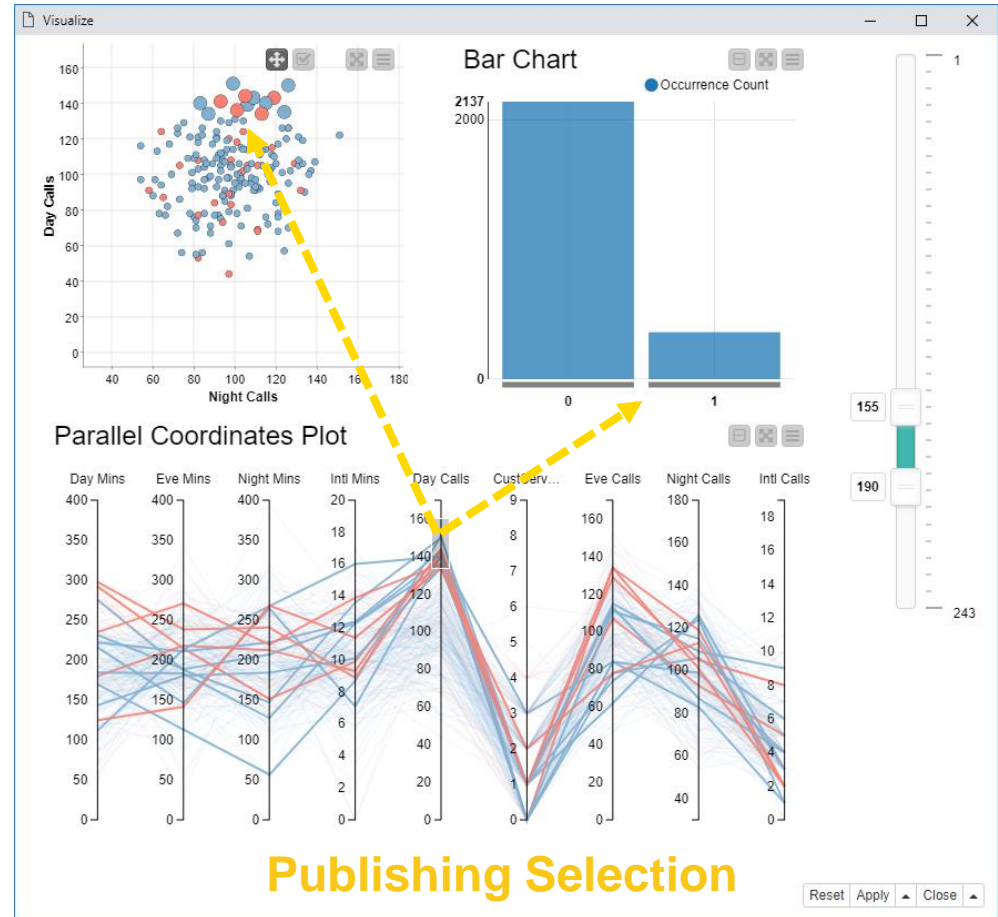
- Listens to the selection made in other view



JS-based nodes: Selection and Filter Events

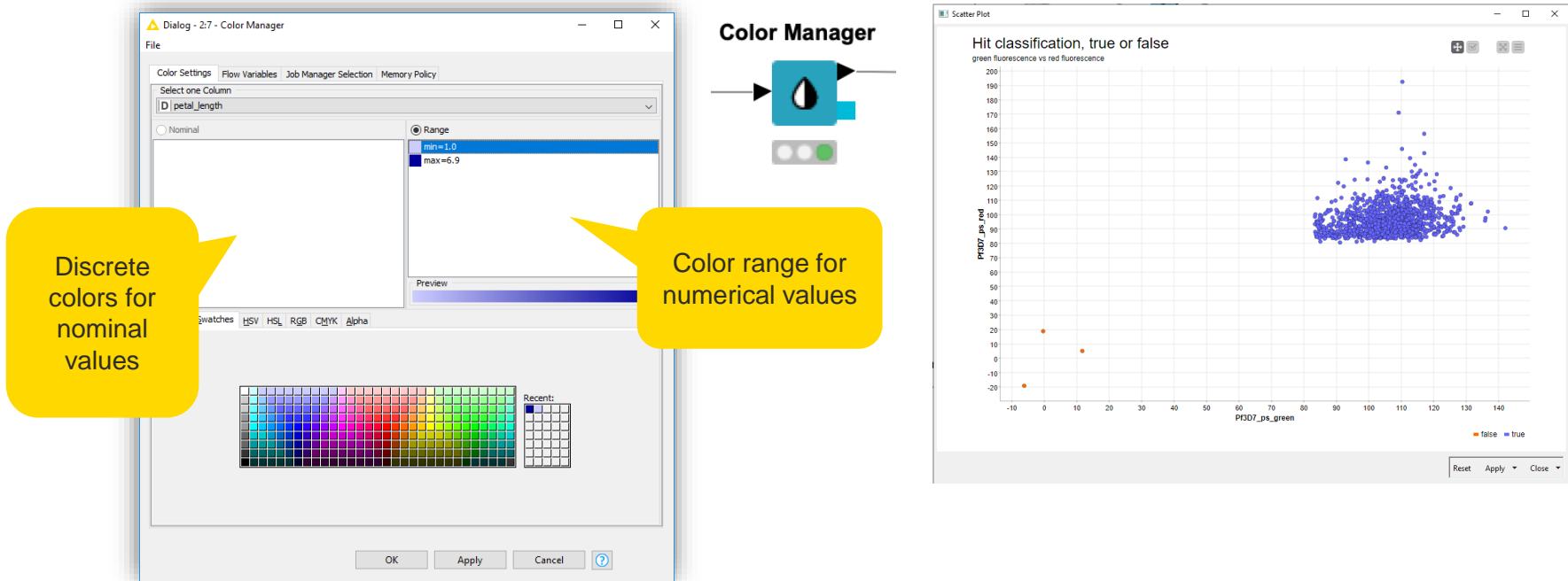


- Makes the other views listen to the selection made in this view



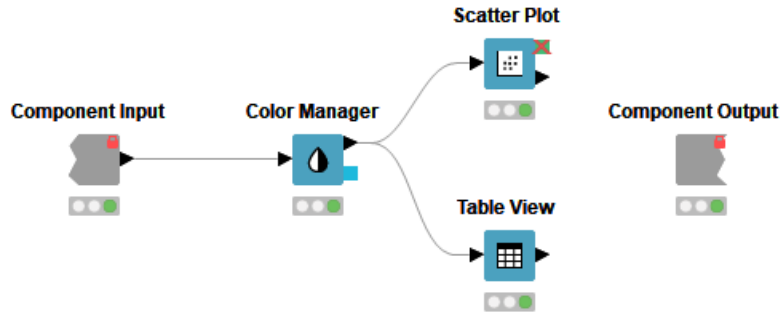
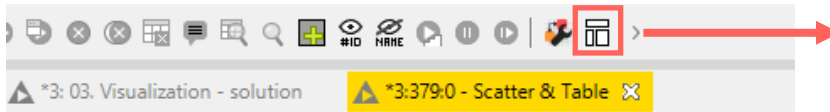
Color Manager

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

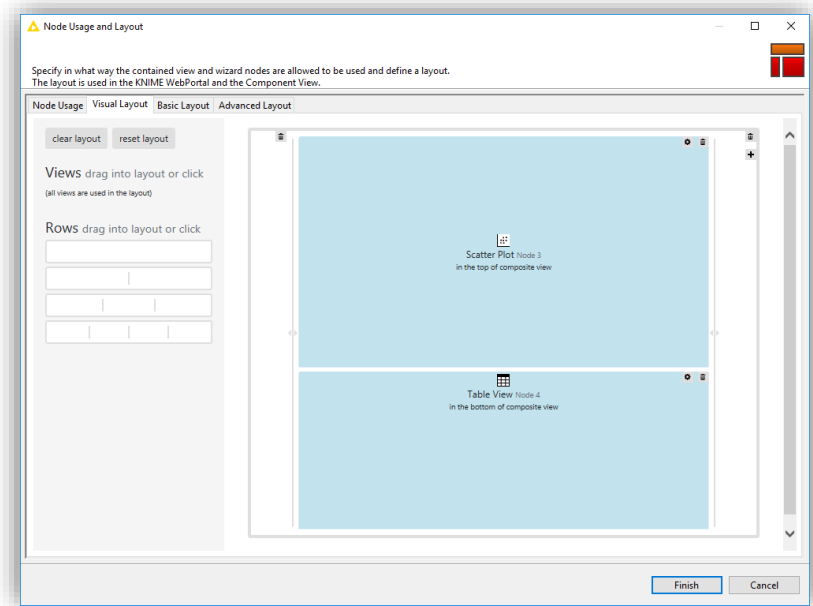


Configure Content and Views Layout

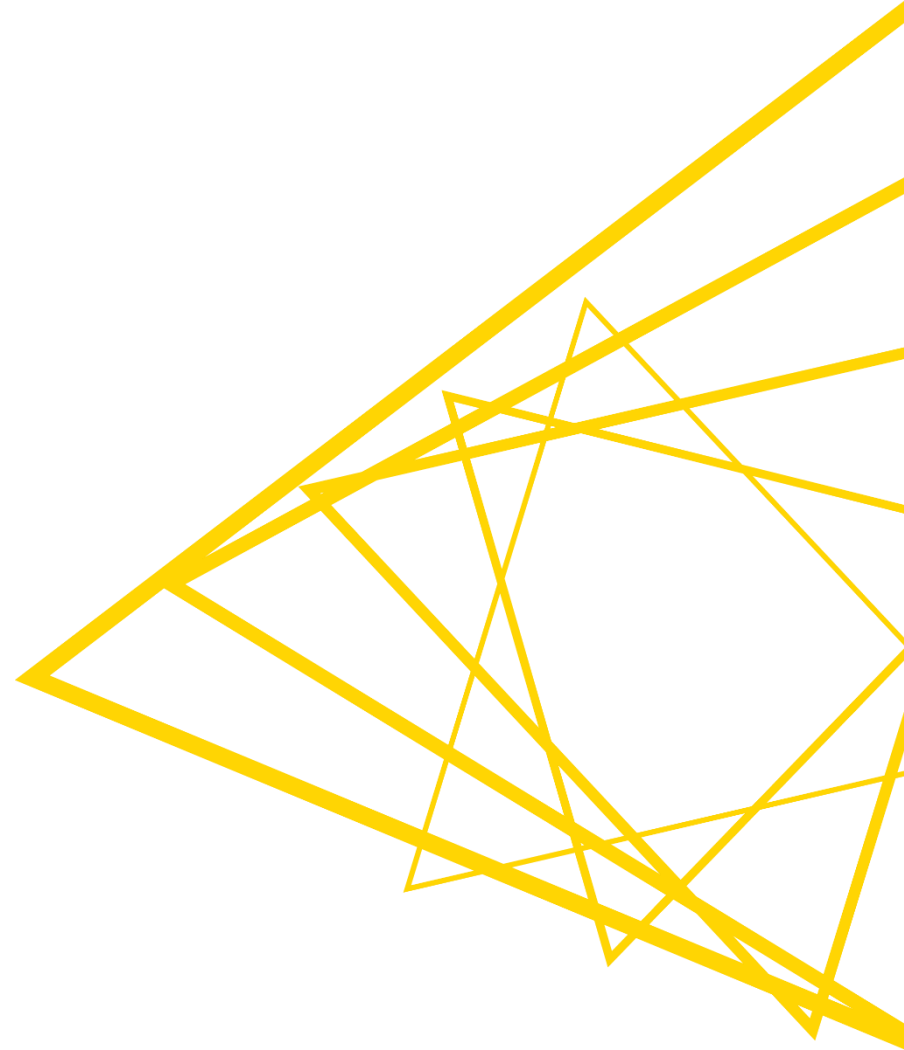
- Click layout button when inside Component to assign views to rows and columns



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



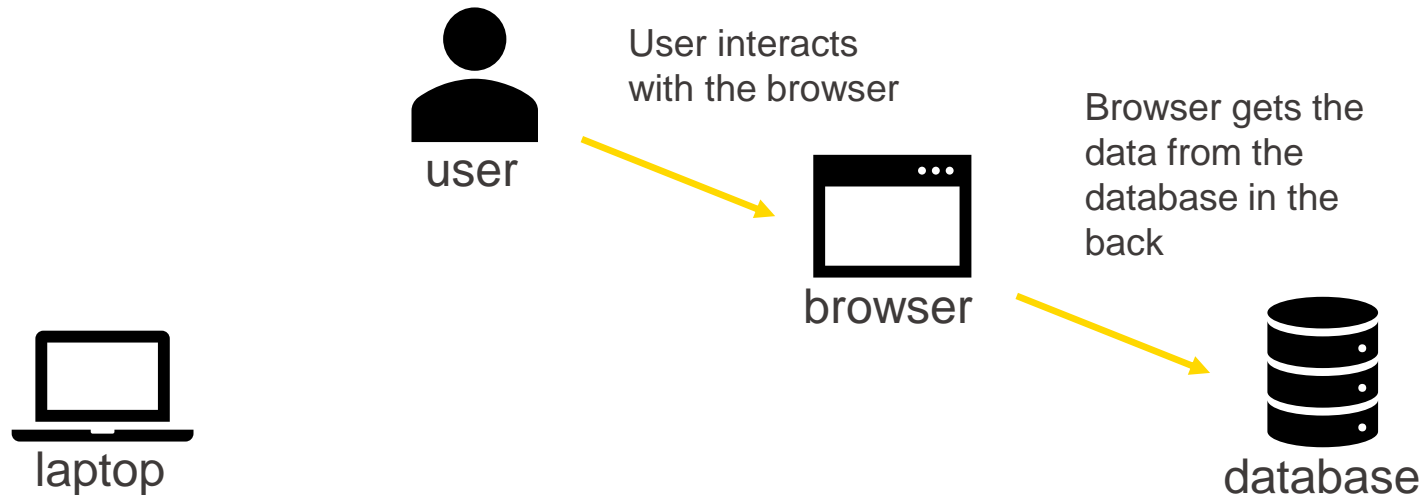
That's it for workflow I...



Workflow II – web service data retrieval

- Instead of retrieving data manually, I can do it in an automated way
- Especially useful for large amounts of data
- Many data sources offer the 'endpoint' to do so (→ REST API, web service)

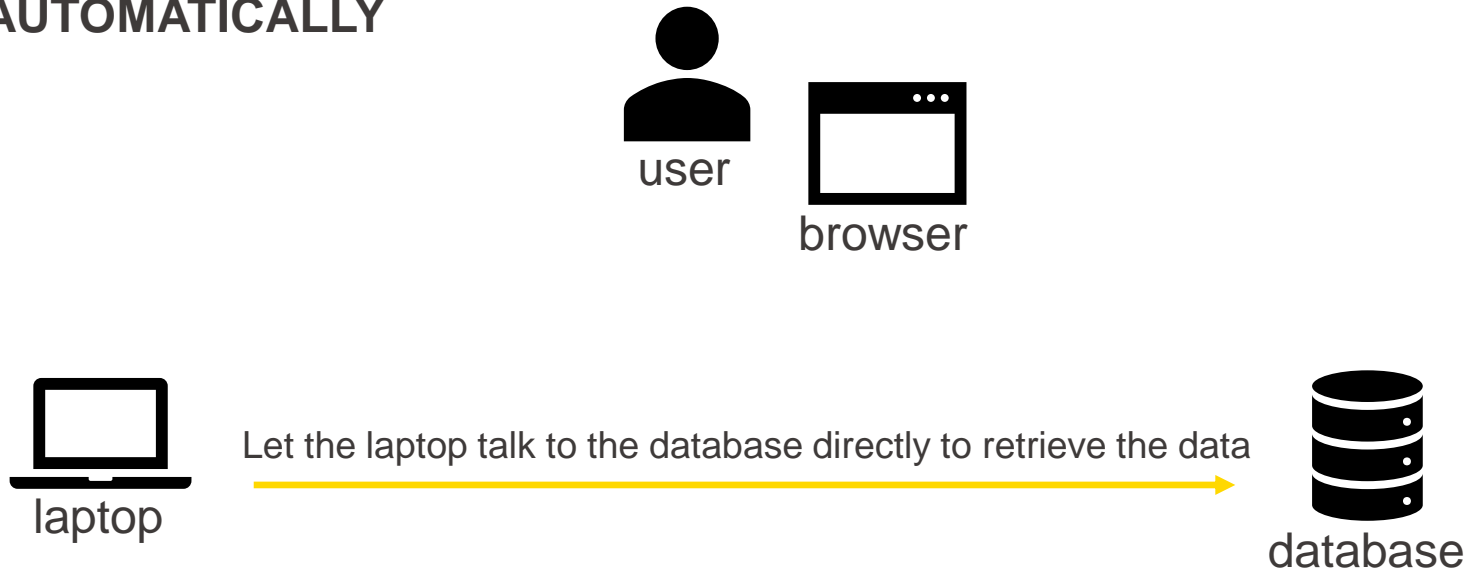
MANUALLY



Workflow II – web service data retrieval

- Instead of retrieving data manually, I can do it in an automated way
- Especially useful for large amounts of data
- Many data sources offer the 'endpoint' to do so (→ REST API, web service)

AUTOMATICALLY



Workflow II – web service data retrieval

- Examples of data sources with an API:
 - <https://docs.mygene.info/en/latest/index.html>
 - <https://chembl.gitbook.io/chembl-interface-documentation/web-services/chembl-data-web-services>
- Construct the according URL to retrieve data for according query

Quick start

MyGene.info provides two simple web services: one for gene queries and the other for gene annotation retrieval. Both return results in JSON format.

Gene query service

URL

```
http://mygene.info/v3/query
```

Examples

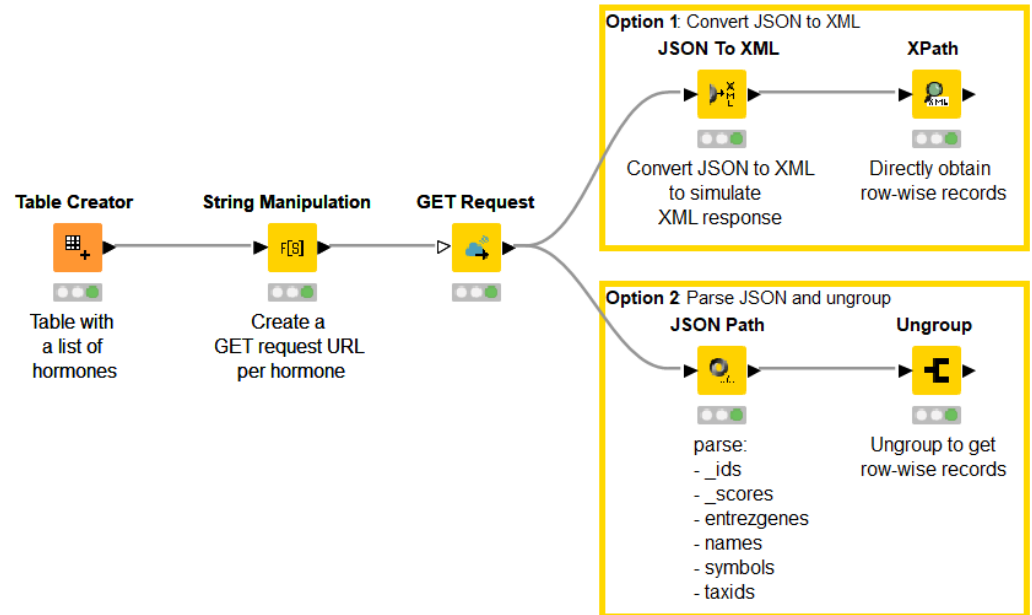
```
http://mygene.info/v3/query?q=cdk2
http://mygene.info/v3/query?q=cdk2&species=human
http://mygene.info/v3/query?q=cdk?
http://mygene.info/v3/query?q=.*
http://mygene.info/v3/query?q=entrezgene:1017
http://mygene.info/v3/query?q=ensemblgene:ENSG00000123374
http://mygene.info/v3/query?q=cdk2&fields=symbol,refseq
```

Steady part

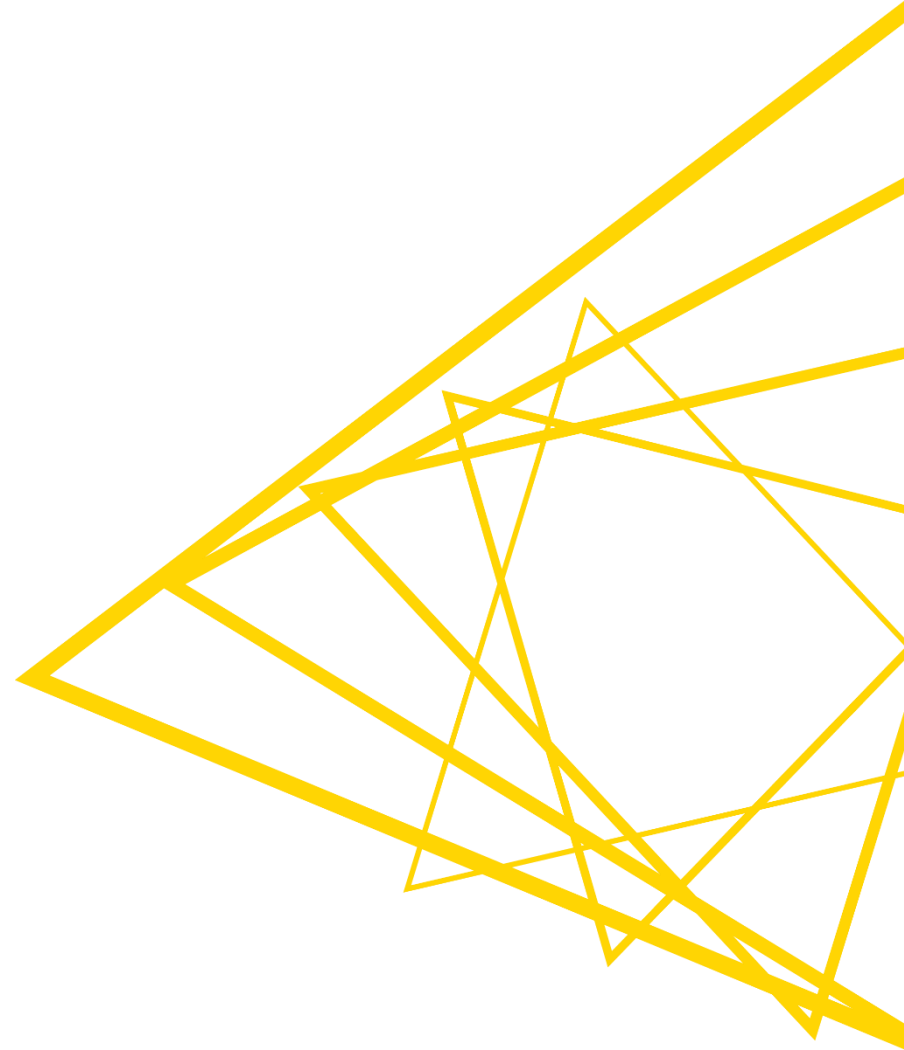


Workflow II – web service data retrieval

- Join the steady part of the URL with your query using the String Manipulation
- The GET Request node ‘talks’ to the DB and retrieves the data
- The result is often not very human-friendly (JSON or XML data format)
- Parse the results to get a neat table



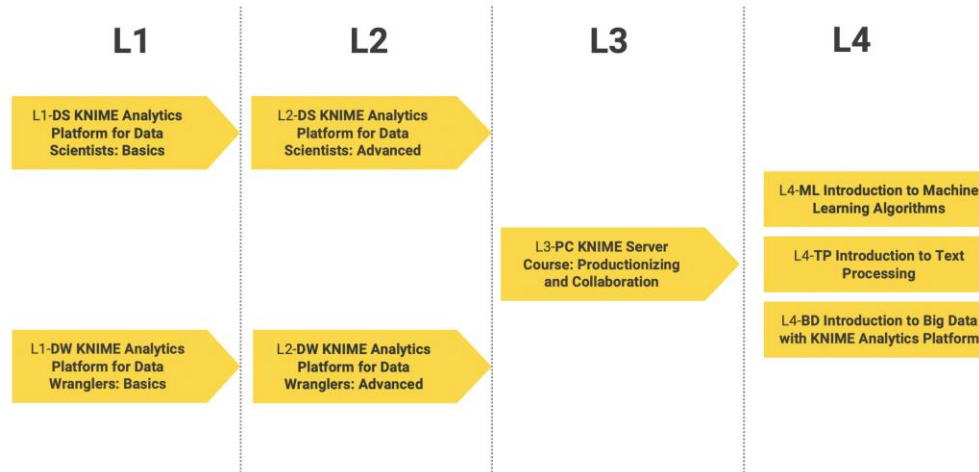
That's it for workflow II...



KNIME courses

■ Self-paced courses

- Courses are organized by level L1 (basic) - L4 (specialized)
- lessons with ~5 minutes videos, hands-on exercises, and knowledge-check questions



■ Instructor-lead online courses

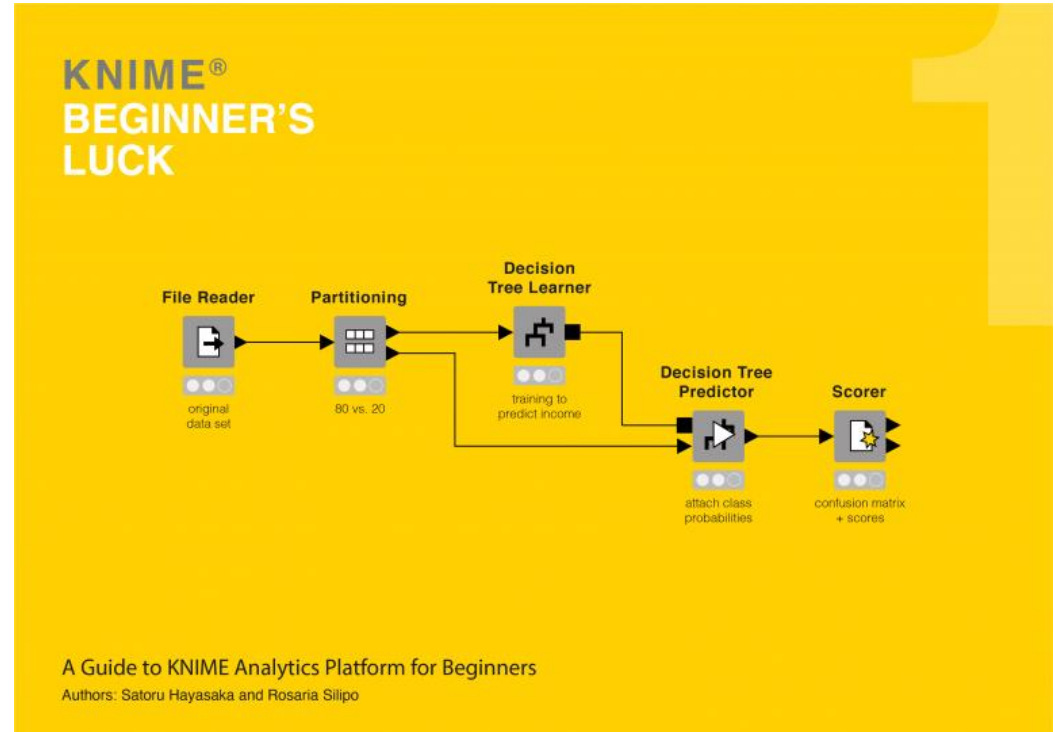
- One week, 1h15 per day
- Possibility to ask questions
- Discount available for academics

<https://www.knime.com/knime-courses>

KNIME book

- KNIME Beginners Luck

CODE: ETH-0123



<https://www.knime.com/knimepress/beginners-luck>

More resources

■ KNIME Community Hub

- Find example workflows you can adapt
- Find out e.g. how others use a certain nodes

Welcome to the
KNIME Community Hub

Solutions for data science: find workflows, nodes and components, and collaborate in spaces.

Search workflows, nodes and more...

Workflows	Nodes	Components	Extensions
12 952	4 293	1 270	222

... and Youtube

<https://www.youtube.com/knimetv>

■ KNIME Forum

- Get and find answers to your questions

Open for Innovation
KNIME

Community Hub Blog Forum Educators Events Solutions Careers Contact Download

SOFTWARE / PRICING / COMMUNITY / LEARNING / PARTNERS / ABOUT

Contributor Hall of Fame Just KNIME It Challenge FAQ About the Forum

all categories all tags Categories Latest New (28) Unread (33) Top + New Topic

Category	Topics	Latest
KNIME Analytics Platform For discussions related to KNIME Analytics Platform	215 / month 22 unread 13 new	Update to Eclipse 2022-06 for KNIME AP 4.7.0 KNIME Development 15 12m
KNIME Extensions For discussions related to KNIME Extensions and Integrations Text Processing 1 unread 1 new Scripting Reporting Image Processing REST Big Data Deep Learning 1 unread	38 / month 4 unread 1 new	NGS Tools missing for Kyme 4.7 Community Extensions ngs 3 15m
KNIME Announcements This category is used to announce new releases and important messages to the KNIME Community	1 / month	External Tools issues - Bash Nodes unusable - KNIME Community Hub KNIME Hub 2 18m
		External Tool "issues" and high difficulty to 0

Cheat Sheets

- Building a KNIME Workflow for Beginners
- Building Components for your Team or the KNIME Community
- Control and Orchestration with KNIME AP
- Data Wrangling with KNIME AP
- Connectors with KNIME AP
- Machine Learning with KNIME AP

Cheat Sheet: Data Wrangling with KNIME Analytics Platform

ACCESS DATA

File Reads a CSV file from either your local file system or another connected file system. Click the three dots in the lower left corner to add a dynamic connection input port to connect to an external file system. See KNIME AP Access Data Storage, etc.

Excel Reads data from one or more Excel files. One sheet from each Excel file. The first sheet is used to read multiple sheets from one Excel file.

Table Reads data from a table in the database. The table files are represented using a KNIME proprietary format, including the full file structure, and are represented by space and line feeds.

SQL Executes a SQL query to retrieve data from a database. The query is represented by space and line feeds.

Database Connects to the JDBC-compliant database. The JDBC driver must be added to the KNIME Preferences and then selected in the Node configuration window.

Database Connects to the JDBC-compliant database. The JDBC driver must be added to the KNIME Preferences and then selected in the Node configuration window.

Common settings of Reader and Writer nodes:

File path: All Reader and Writer nodes require a file path. The file path can be expressed as an absolute path or as a relative path to a key location in the current KNIME installation, or a path defined in an external file system. A relative path is used.

Multiple files: Reader nodes can read and concatenate multiple files, according to a selected file selection or a name pattern.

Transformation: All Reader nodes include a Transformation tab for receiving, filtering, re-ordering, and type changing of the columns.

COMBINE DATA

Join Concatenates the rows of all input tables by adding a new column to the first table. The new column is named after the input table. The new column is named after the input table.

Join Concatenates the rows of all input tables by adding a new column to the first table. The new column is named after the input table. The new column is named after the input table.

Join Concatenates the rows of all input tables by adding a new column to the first table. The new column is named after the input table. The new column is named after the input table.

FILTER DATA

Filter Filters rows in or out of the input table according to a set of rules. The rules are defined in the configuration window. The rules are evaluated from top to bottom. Using TRUE as the condition always returns all rows.

Filter Filters rows in or out of the input table according to a set of rules. The rules are defined in the configuration window. The rules are evaluated from top to bottom. Using TRUE as the condition always returns all rows.

Filter Filters rows in or out of the input table according to a set of rules. The rules are defined in the configuration window. The rules are evaluated from top to bottom. Using TRUE as the condition always returns all rows.

WRITE DATA

Excel Writes the input data table to a CSV file. Click the three dots in the lower left corner to add a dynamic connection input port to write to an external file system. See KNIME AP Access Data Storage, etc.

Table Writes the input data table to a table in the database. The table files are represented using a KNIME proprietary format, including the full file structure, and are represented by space and line feeds.

SQL Executes a SQL query to insert data into a database. The query is represented by space and line feeds.

Database Inserts the data into the database. The data is represented by space and line feeds.

Database Inserts the data into the database. The data is represented by space and line feeds.

RESHAPE AND AGGREGATE DATA

Group Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the grouped rows. Groups are represented by space and line feeds.

Group Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the grouped rows. Groups are represented by space and line feeds.

Group Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the grouped rows. Groups are represented by space and line feeds.

DATA TYPES & CONVERSIONS

String Sequence of characters, e.g. "This is a string".

Integer Whole real-valued numbers, e.g. 100 or -345.

Double Real-valued numbers, e.g. 0.5 or 0.1234.

Date/Time A date format for time, date, and time. The format is defined in the configuration window.

Boolean Two possible values only, e.g. TRUE and FALSE.

Collection A collection of multiple values of either the same or different types, e.g. a list of values or a set of values. In KNIME, a collection is only used in the configuration window.

Document A document type, e.g. a text file, a PDF file, or a JSON file.

CREATE COLUMNS

Formula Implements a number of math operations across multiple input columns. The math operations can be applied to the corresponding columns in the input table. The first rule that matches is used.

Formula Implements a number of math operations across multiple input columns. The math operations can be applied to the corresponding columns in the input table. The first rule that matches is used.

Formula Implements a number of math operations across multiple input columns. The math operations can be applied to the corresponding columns in the input table. The first rule that matches is used.

DYNAMIC PORT

Dynamic ports Additional input ports can be added by clicking the three dots in the bottom left corner of a node.

Dynamic ports Additional input ports can be added by clicking the three dots in the bottom left corner of a node.

Dynamic ports Additional input ports can be added by clicking the three dots in the bottom left corner of a node.

FORMAT EXCEL SHEETS

Format Applies the KNIME extension to automatically format an existing Excel sheet. The format is defined in the configuration window. The format is defined in the configuration window.

Format Applies the KNIME extension to automatically format an existing Excel sheet. The format is defined in the configuration window. The format is defined in the configuration window.

Format Applies the KNIME extension to automatically format an existing Excel sheet. The format is defined in the configuration window. The format is defined in the configuration window.

DATE/TIME

Time Extracts rows when the time value is in the selected column. The time value is extracted from the selected column. The time value is extracted from the selected column.

Time Extracts rows when the time value is in the selected column. The time value is extracted from the selected column. The time value is extracted from the selected column.

Time Extracts rows when the time value is in the selected column. The time value is extracted from the selected column. The time value is extracted from the selected column.

CLEAN DATA

Remove Removes rows from the input table. The rows are removed from the input table. The rows are removed from the input table.

Remove Removes rows from the input table. The rows are removed from the input table. The rows are removed from the input table.

Remove Removes rows from the input table. The rows are removed from the input table. The rows are removed from the input table.

KNIME HUB

KNIME Hub Browse and share workflows, nodes, and components. Add settings or comments to other workflows at <https://www.knime.com/knime-hub>.

KNIME Hub Browse and share workflows, nodes, and components. Add settings or comments to other workflows at <https://www.knime.com/knime-hub>.

KNIME Hub Browse and share workflows, nodes, and components. Add settings or comments to other workflows at <https://www.knime.com/knime-hub>.

<https://www.knime.com/cheat-sheets>

© 2023 KNIME AG. All rights reserved.

69

Open for innovation
KNIME

Thank You!

Questions? Please reach out

`alice.krebs@knime.com`

`glandrum@ethz.ch`

