National Technical University of Athens

Using KNIME for modelling toxicity in nanoparticles

Georgios Drakakis, PhD

National Technical University of Athens

eNanoMapper

The Konstanz Information Miner

- Available at https://www.knime.org/
- KNIME is an open source data analytics platform.
- Uses pipeline philosophy.
- Nodes for machine learning and data mining.
- Modeling, data analysis, visualization and reporting.

Berthold MR, Cebron N, Dill F & Gabriel TR. The Konstanz Information Miner. in *Studies in Classification, Data Analysis, and Knowledge Organization* (GfKL 2007); 11: 319–326 (Springer, 2007)

Data for this workshop

Please download files:

- Iris: https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/datasets/
- Walkey: https://data.enanomapper.net/substanceowner/FCSV-319611C6-E7DA-3977-A5AC-EB74D49A4319/dataset

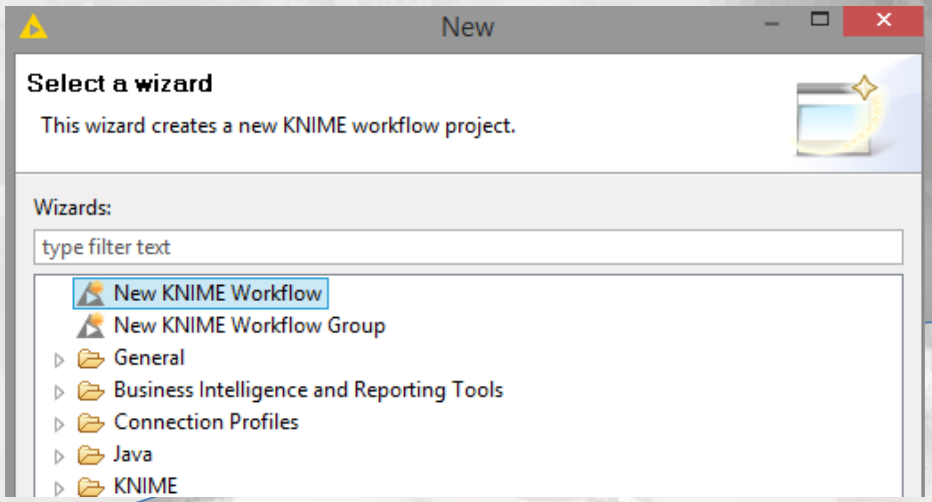
  - Export as CSV
- Gene to Protein IDs Dictionary: distributed locally

eNM eNanoMapper

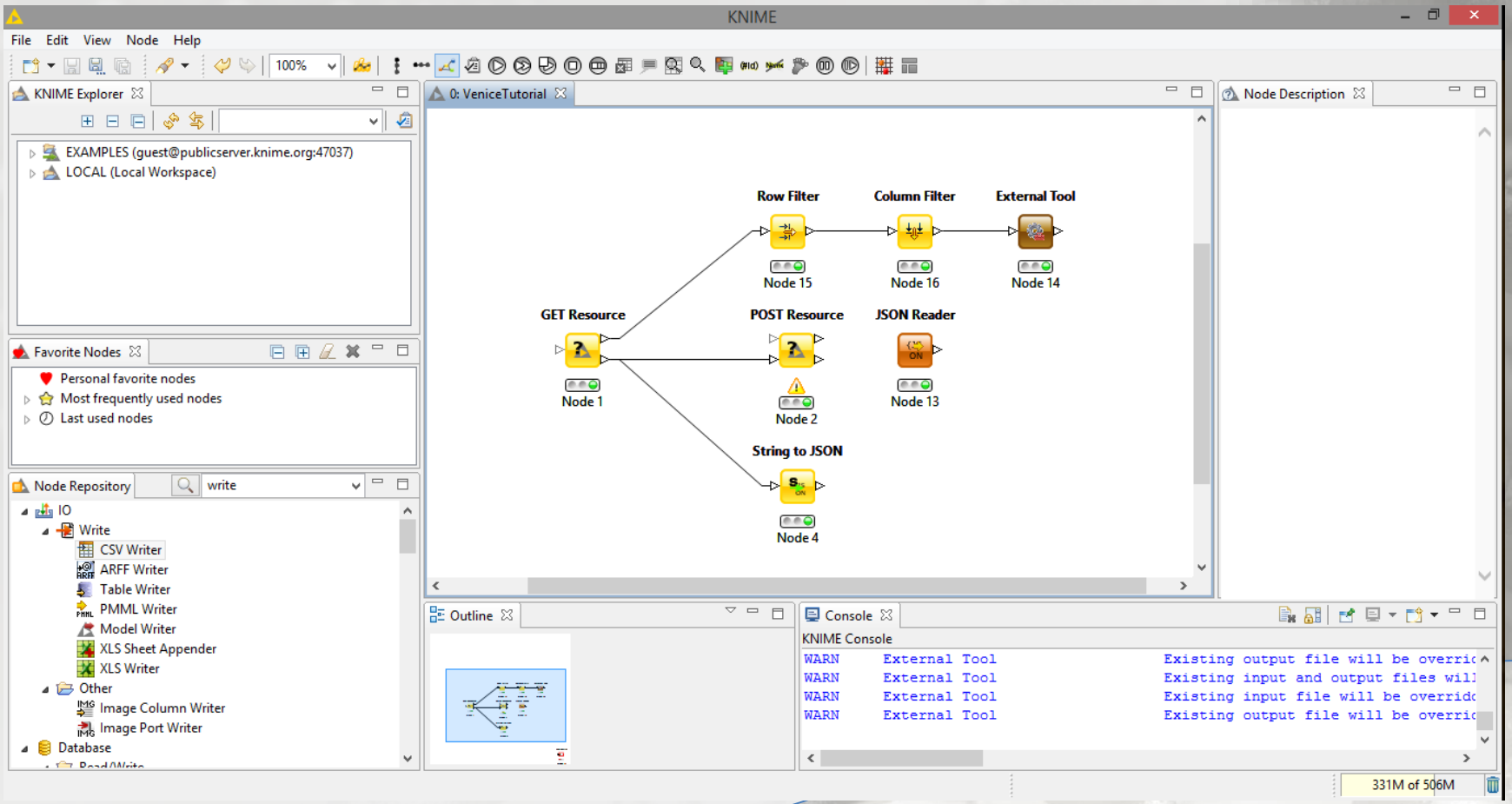
Getting Started

- 1st time users: Open KNIME, Enter Name of default workspace
- Returning users: Open KNIME, Select default workspace

- File ->New …
- New KNIME Workflow -> Next
- Type Name -> Finish



What the interface looks like

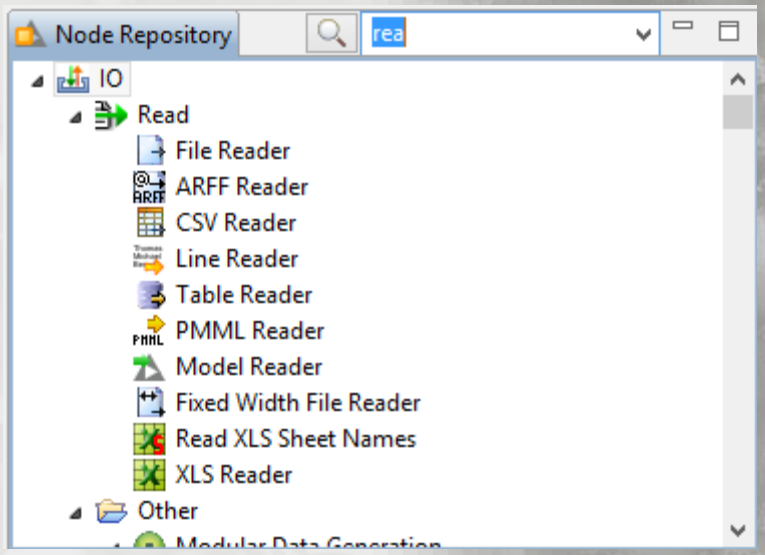


KNIME Windows (from View Menu)

- KNIME Explorer: Workflows saved previously
- Favorite Nodes: Most frequently used
- Node repository: All nodes available
- Outline: Map of current workflow
- Console: Messages from KNIME (warnings/errors)
- Node Description: Info about the node functionality and ports
- Workflow window

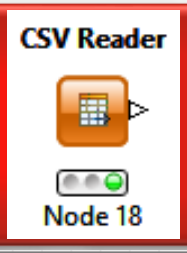
Node Repository

- Nodes Under Categories
- I/O, RDKit, KNIME Labs, Weka, etc…
- I/O contains readers and writers

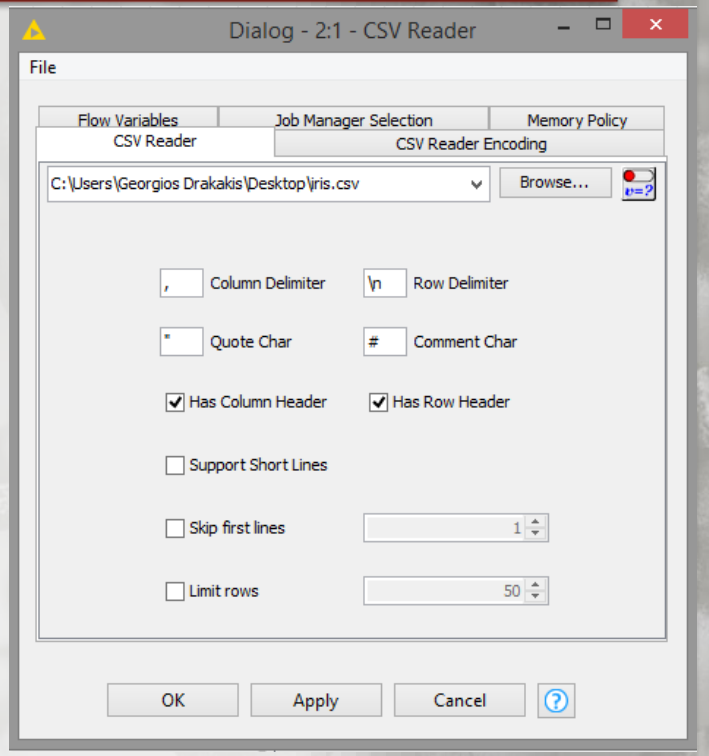


- Drop down menu or *type* in search
- To insert into workflow
- Double Click on node or drag and drop
- Try CSV Reader

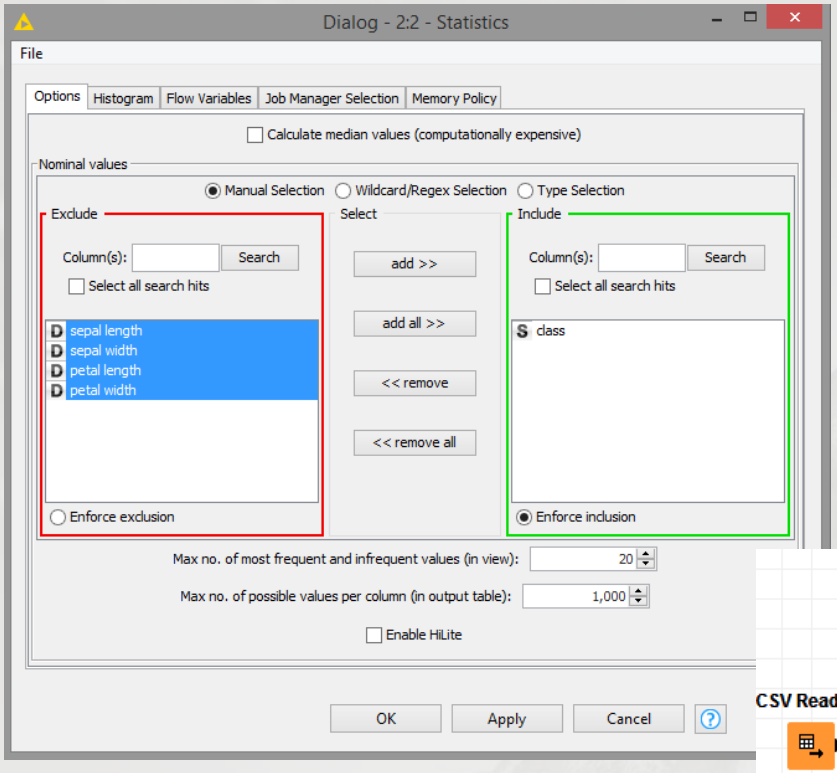
CSV Reader

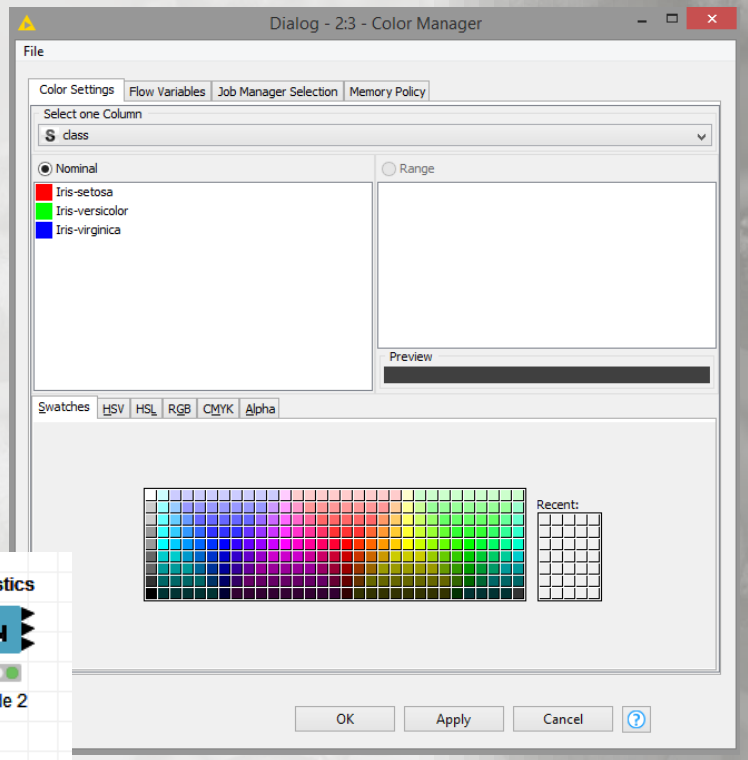


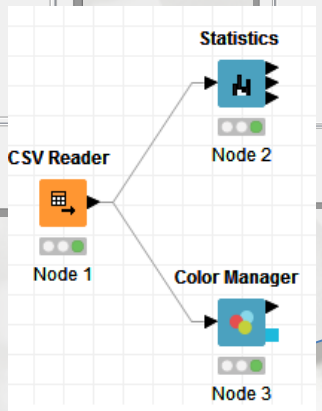
- See:
- What it does (Right: Node description)
- How to configure
- Use iris.csv
- How to execute
- View results
- Connect to another node



Statistics and Color Nodes



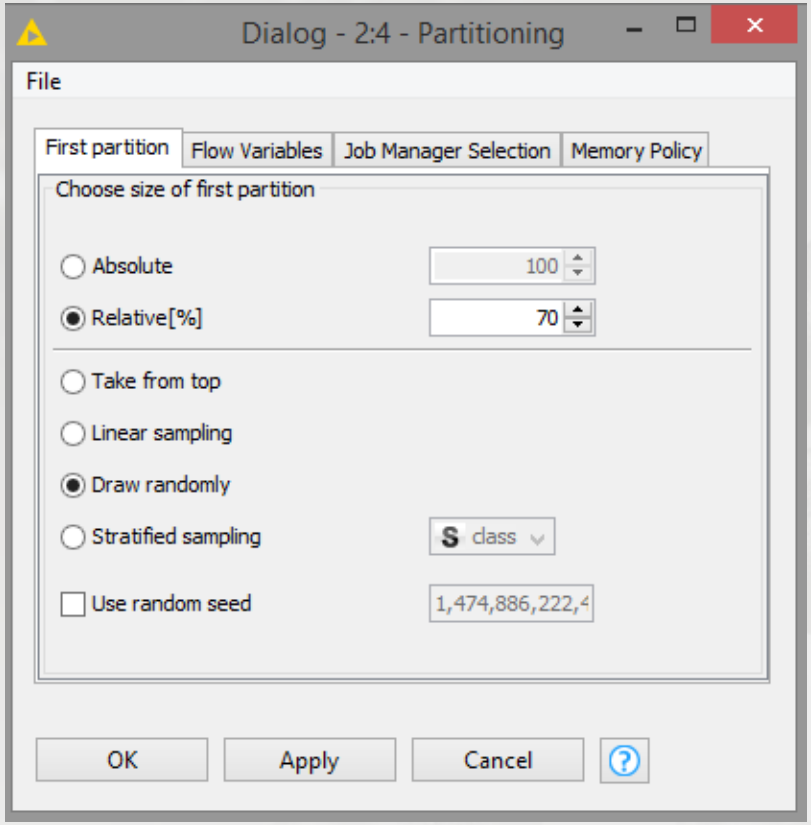


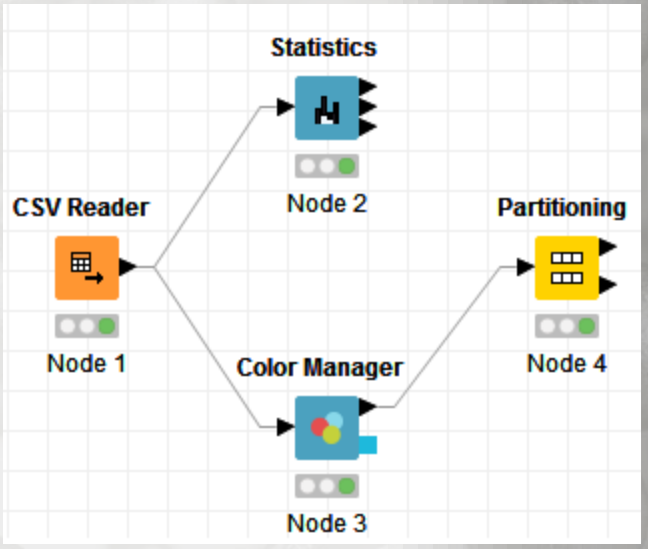


eNanoMapper

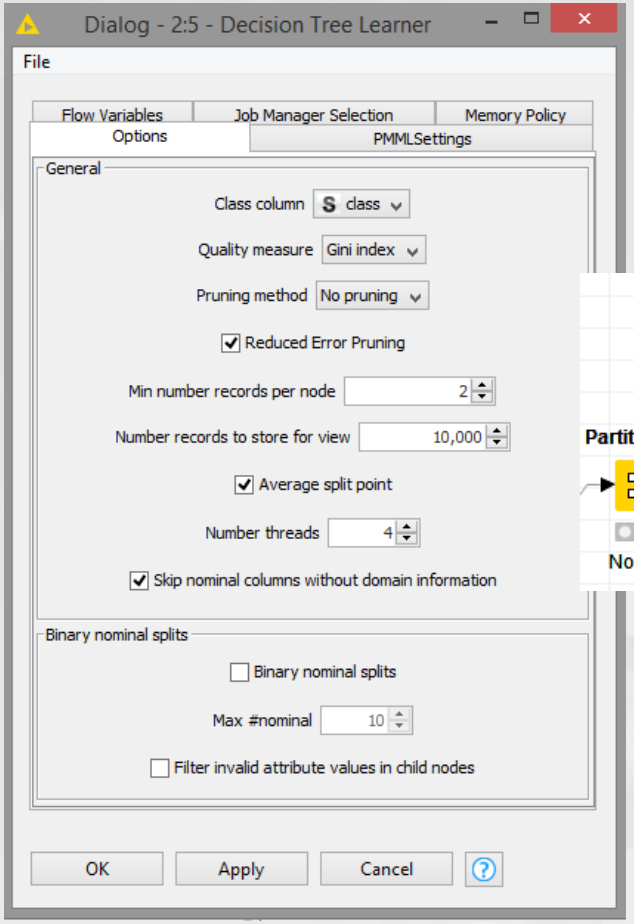
Splitting for training/predicting

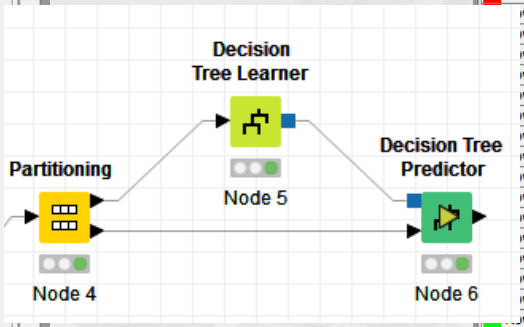
- Why do we split data?

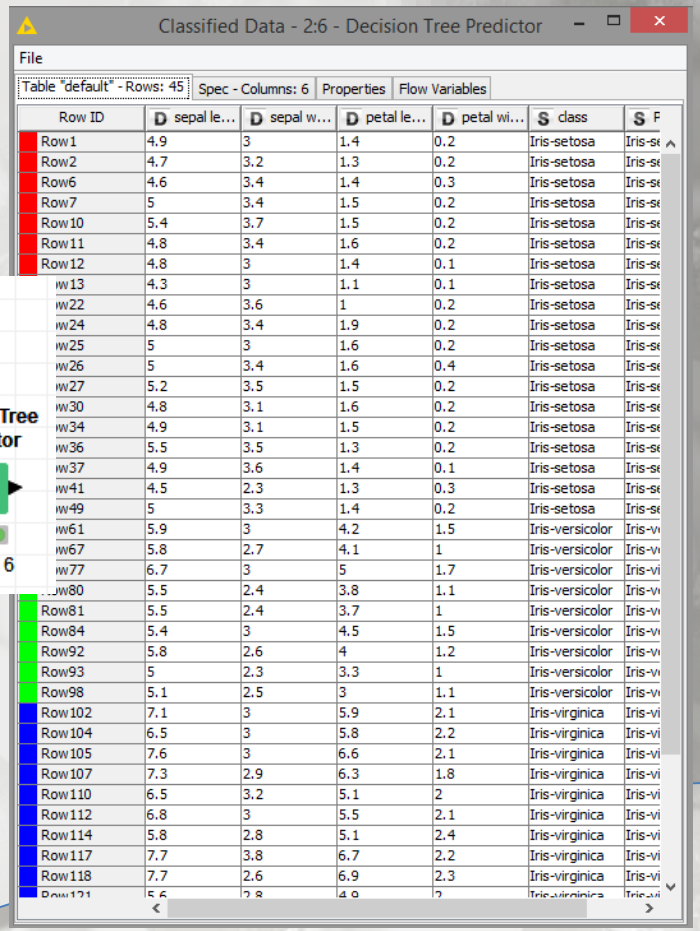




Task 1: Modelling a Dummy Dataset

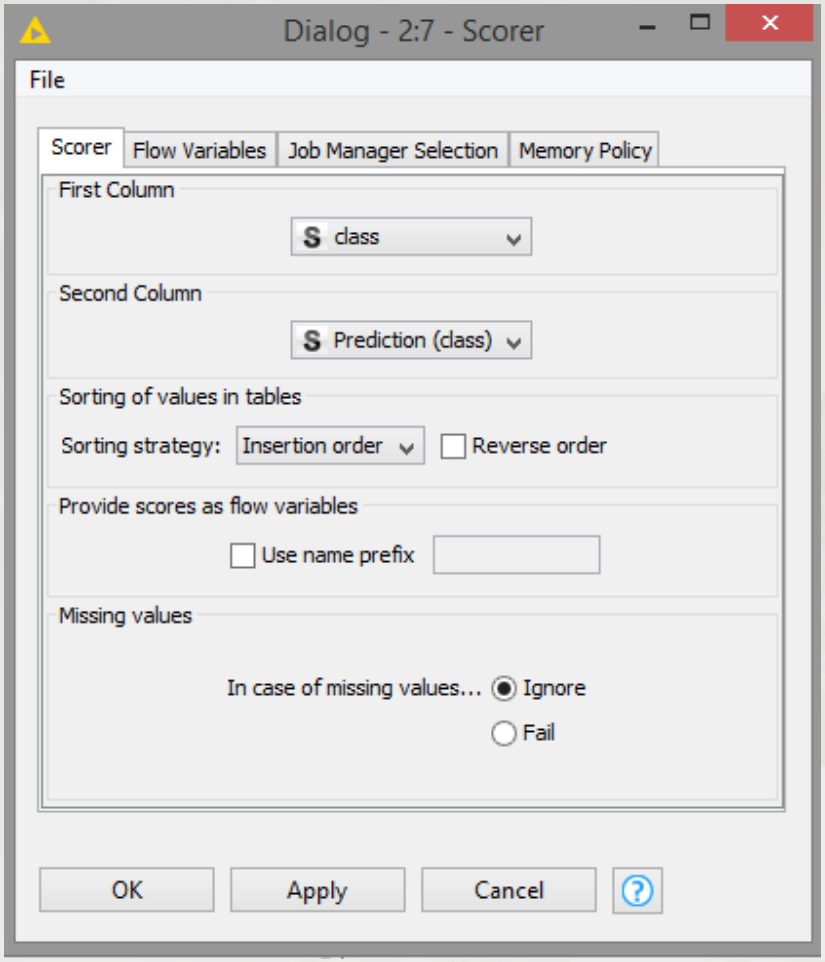


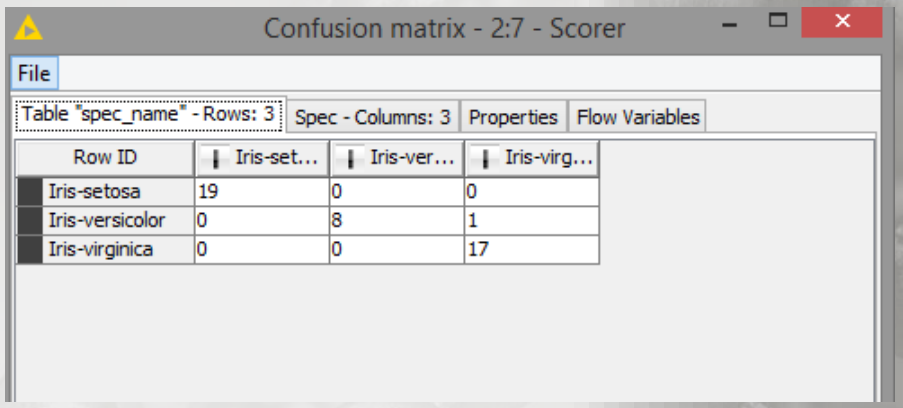




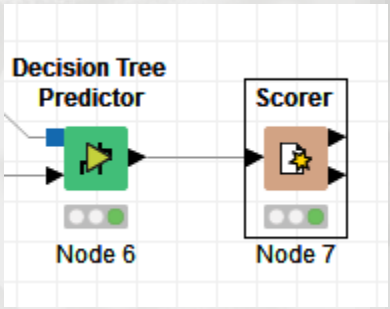
eNanoMapper

Scoring the model

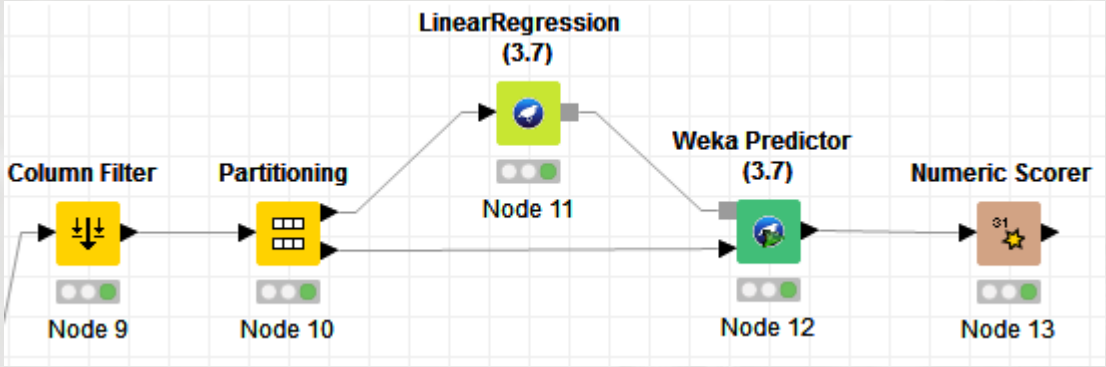




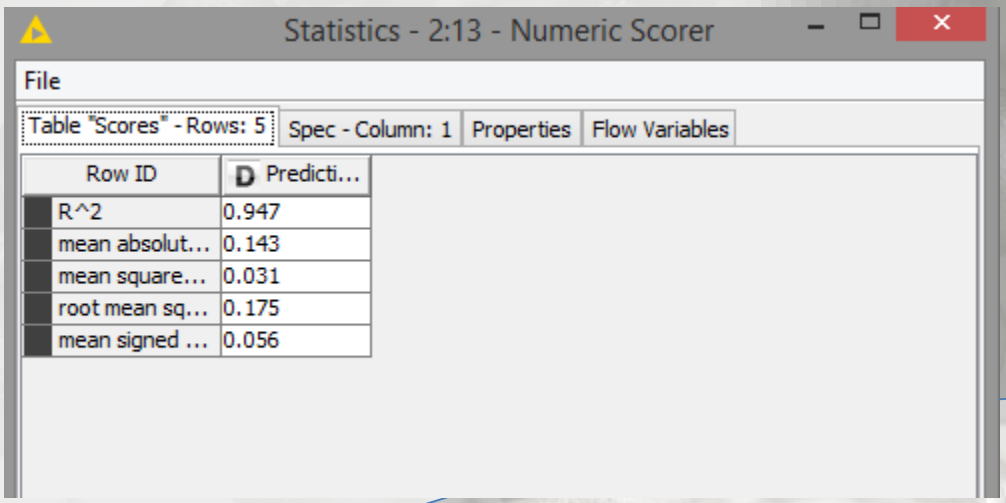
Row ID	Iris-set...	Iris-ver...	Iris-virg...
Iris-setosa	19	0	0
Iris-versicolor	0	8	1
Iris-virginica	0	0	17



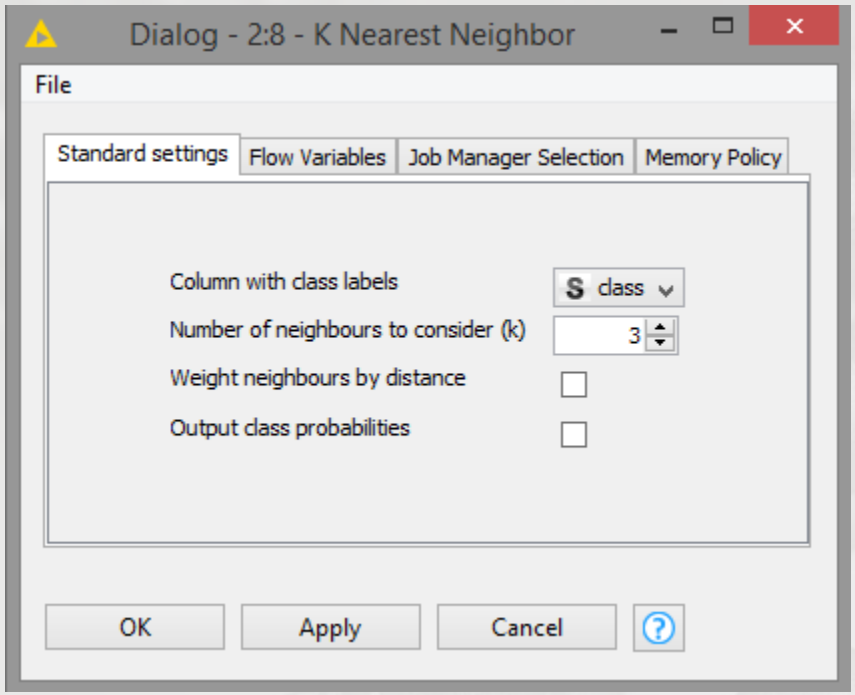
DIY: Linear Regression Example

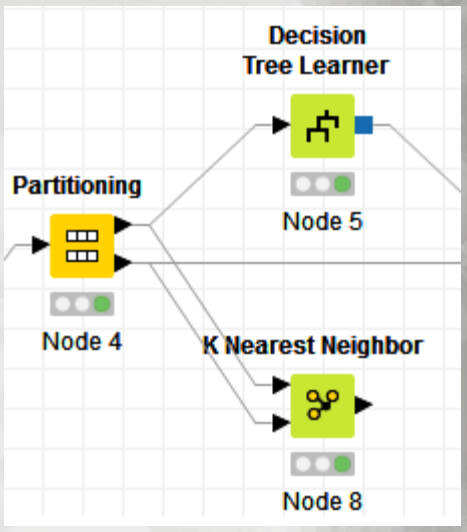



```
LinearRegression

Linear Regression Model

petalwidth =

     -0.2142 * sepallength +
      0.206  * sepalwidth +
      0.5235 * petallength +
     -0.1348
```



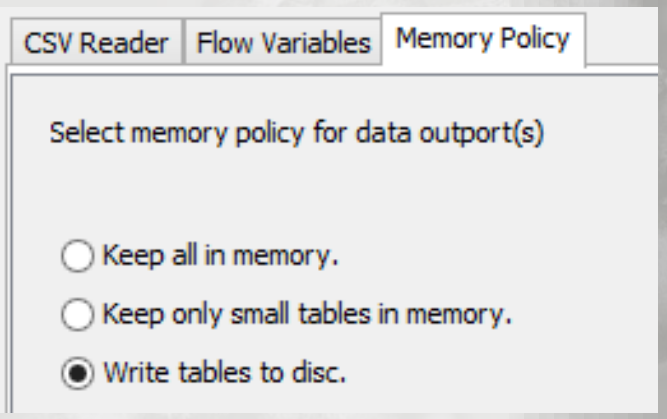
Clustering the Iris Dataset





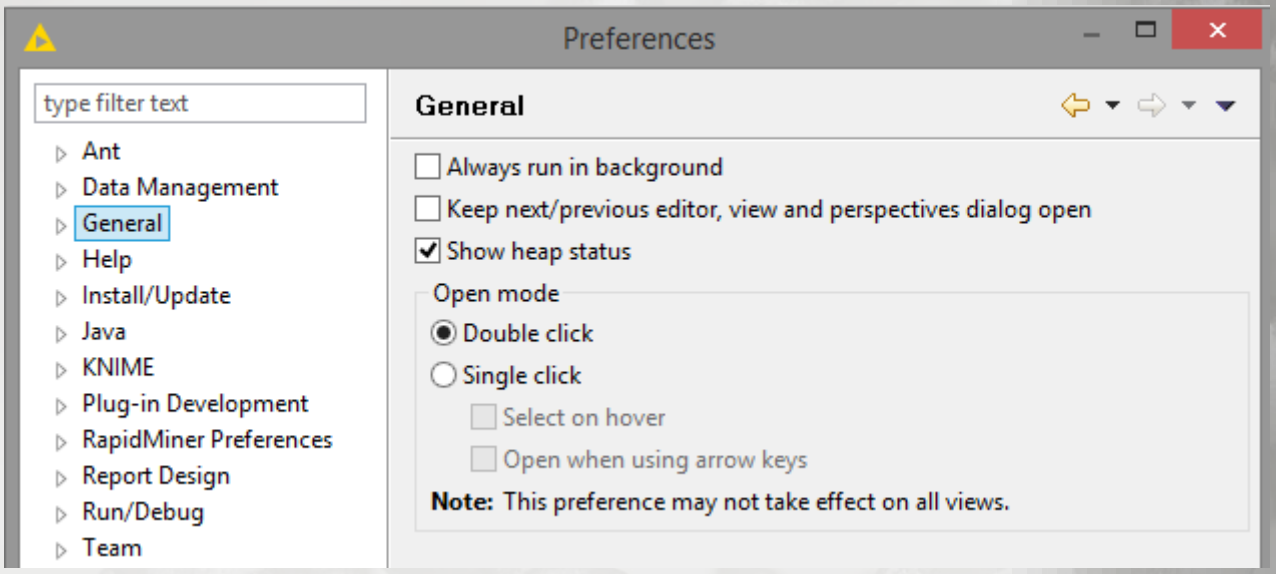
Writing Tables to Disc

- Helps save memory
- Takes longer for calculations



Clear Memory after calculations

- File -> Preferences -> General -> Show Heap Status



Dataset for Task 2

- http://pubs.acs.org/doi/abs/10.1021/nn406018q

Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles

Carl D. Walkey†§, Jonathan B. Olsen‡§, Fayi Song†§, Rong Liu∇⊗, Hongbo Guo‡§, D. Wesley H. Olsen†‡§, Yoram Cohen∇⊗, Andrew Emili‡§, and Warren C. W. Chan†§⊥∥#*
†Institute of Biomaterials and Biomedical Engineering, ‡Banting and Best Department of Medical Research, §Donnelly Centre for Cellular and Biomolecular Research, ⊥Department of Chemical Engineering, ∥Department of Chemistry, #Department of Materials Science and Engineering, University of Toronto, Toronto, Ontario, Canada M5S 3G9
∇Center for Environmental Implications of Nanotechnology, ⊗Chemical and Biomolecular Engineering Department, University of California, Los Angeles, California 90095, United States

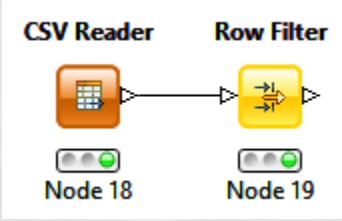
ACS Nano, **2014**, *8* (3), pp 2439–2455
DOI: 10.1021/nn406018q
Publication Date (Web): February 11, 2014

*Address correspondence to warren.chan@utoronto.ca.

- Made available at https://data.enanomapper.net/

Row Filter to remove unwanted

- Option: Include Rows by number



File Table - 2:1 - CSV Reader

File

Table "walkey.csv" - Rows: 100 | Spec - Columns: 823 | Properties | Flow Variables

Row ID	NP ID	Element	Abbrevi...	Classifi...	Net cell.
Row81	G60.Ser-SH_1	[Au]	Ser-SH	Anionic	0.061
Row82	G60.SPP_1	[Au]	SPP	Anionic	0.046
Row83	G60.Trp-SH_1	[Au]	Trp-SH	Anionic	0.065
Row84	?	?	?	?	?
Row85	?	?	?	?	?
Row86	?	?	?	?	?
Row87	?	?	?	?	?
Row88	?	?	?	?	?
Row89	?	?	?	?	?
Row90	?	?	?	?	?
Row91	?	?	?	?	?
Row92	?	?	?	?	?
Row93	?	?	?	?	?
Row94	?	?	?	?	?
Row95	?	?	?	?	?
Row96	?	?	?	?	?

Filtered - 2:2 - Row Filter

File

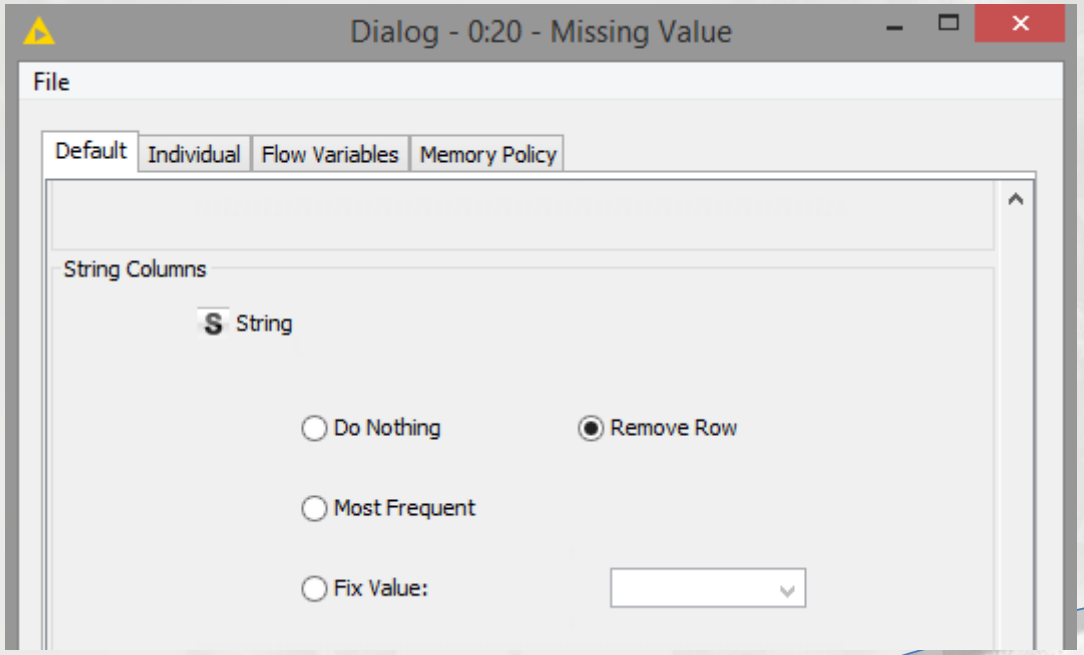
Table "walkey.csv" - Rows: 84 | Spec - Columns: 823 | Properties | Flow Variables

Row ID	NP ID	Element	Abbrevi...	Classifi...	Net cell.
Row67	G60.CIT_1	[Au]	CIT	Anionic	0.037
Row68	G60.CTAB_1	[Au]	CTAB	Cationic	0.06
Row69	G60.CVVIT_1	[Au]	CVVIT	Anionic	0.04
Row70	G60.DDT@B...	[Au]	DDT@BDHDA	Cationic	0.053
Row71	G60.DDT@D...	[Au]	DDT@DOTAP	Cationic	0.81
Row72	G60.DTNB_1	[Au]	DTNB	Anionic	0.017
Row73	G60.HDA_1	[Au]	HDA	Cationic	0.497
Row74	G60.MBA_1	[Au]	MBA	Anionic	0.155
Row75	G60.MPA_1	[Au]	MPA	Anionic	0.118
Row76	G60.MUTA_1	[Au]	MUTA	Cationic	2.509
Row77	G60.NT@PS...	[Au]	NT@PSMA-AP	Anionic	0.049
Row78	G60.ODA_1	[Au]	ODA	Cationic	0.097
Row79	G60.Phe-SH_1	[Au]	Phe-SH	Anionic	0.056
Row80	G60.PVA_1	[Au]	PVA	Anionic	0.024
Row81	G60.Ser-SH_1	[Au]	Ser-SH	Anionic	0.061
Row82	G60.SPP_1	[Au]	SPP	Anionic	0.046
Row83	G60.Trp-SH_1	[Au]	Trp-SH	Anionic	0.065

eNanoMapper

Alternatively Use the Missing Value Node

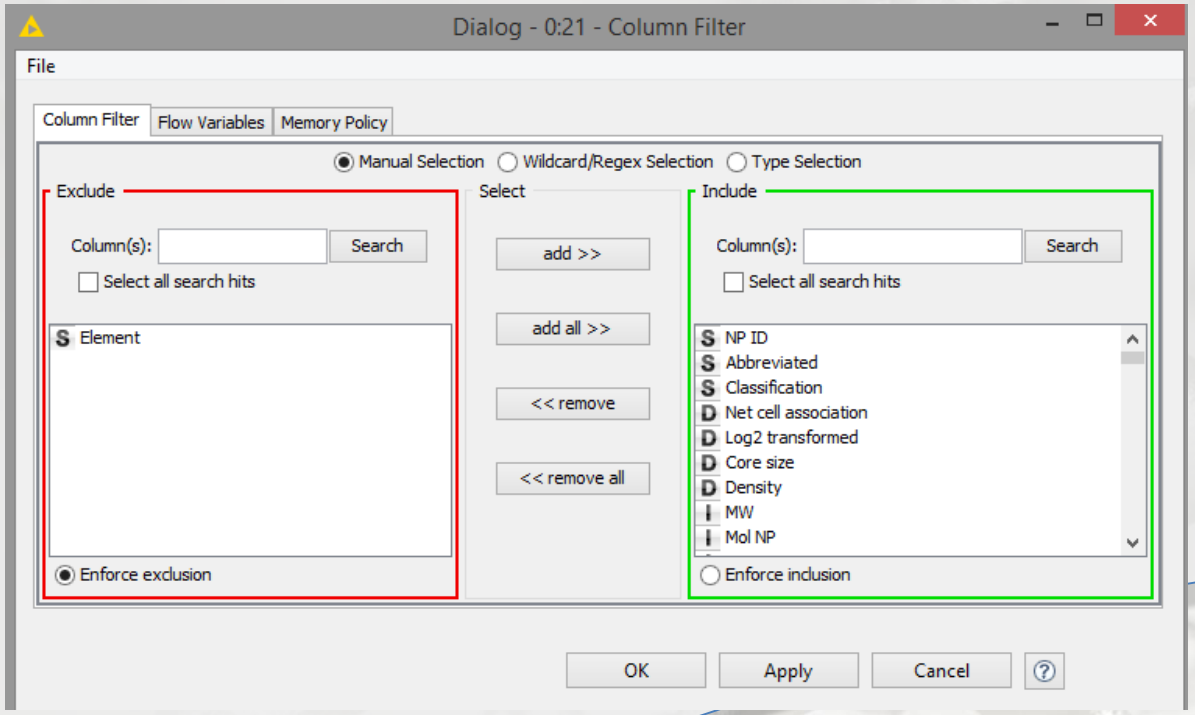
- Type Missing into the Node repository
- Configure to remove rows when a missing value occurs



Column Filter

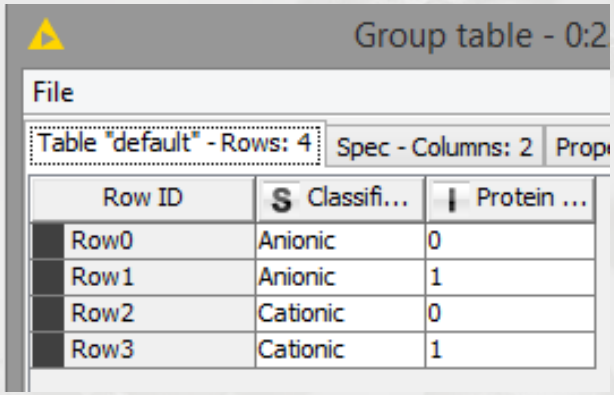
- Remove Unwanted Properties

  i.e. Element (Au in all entries)



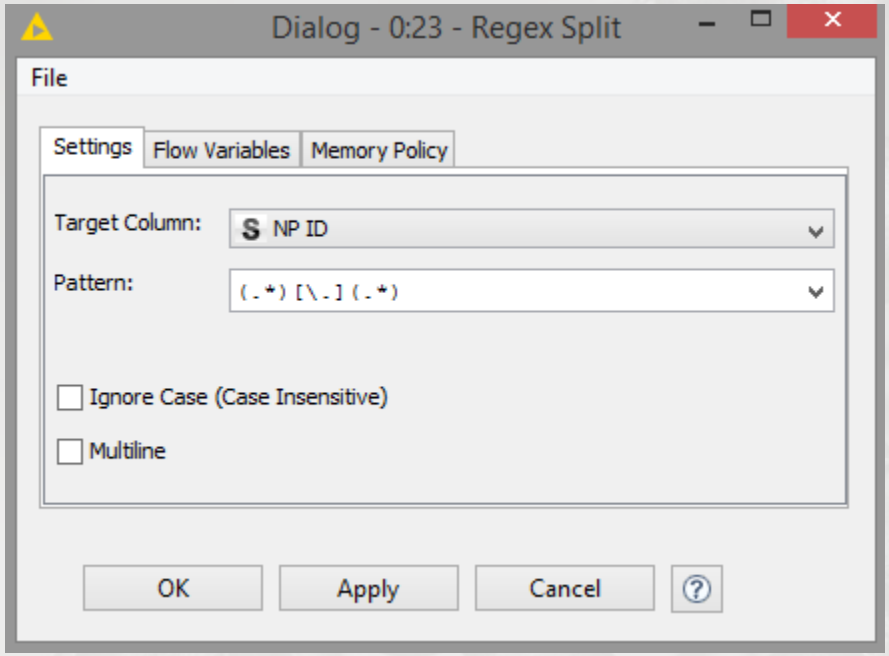
GroupBy node

- Groups per entries of a particular column or sets of columns
- If all columns are selected it is a essentially a check for duplicate rows
- In the second tab, one can select manual aggregation methods

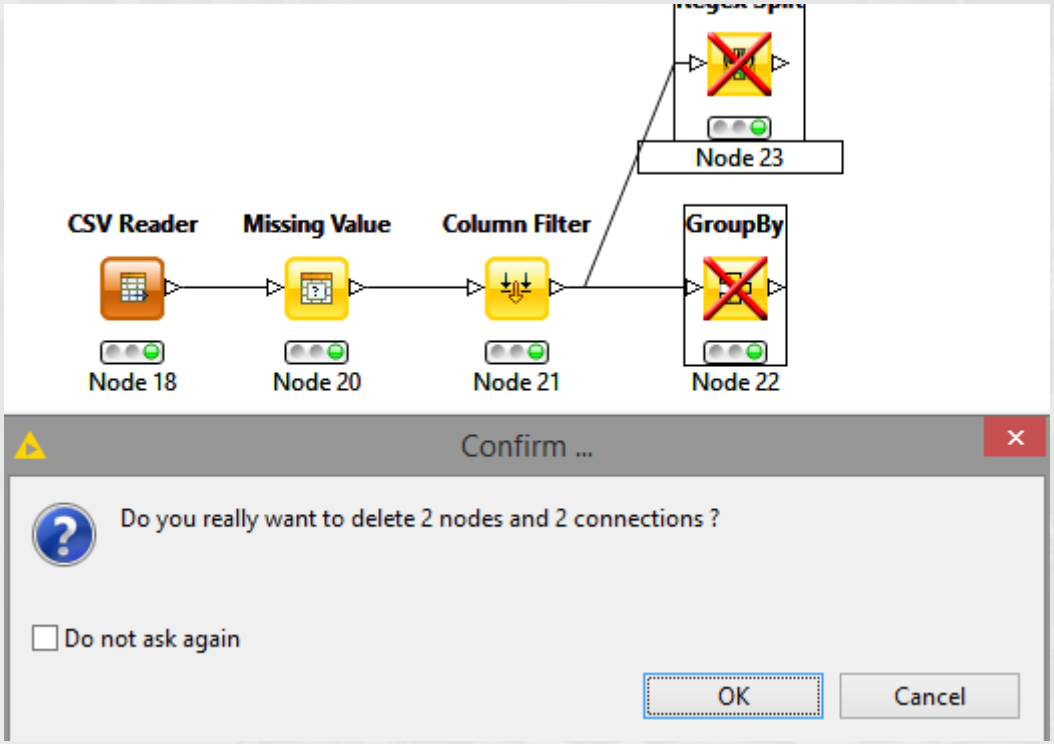


RegEx Split

- Separate NP ID based on dot
- (.*)[\.](.*)

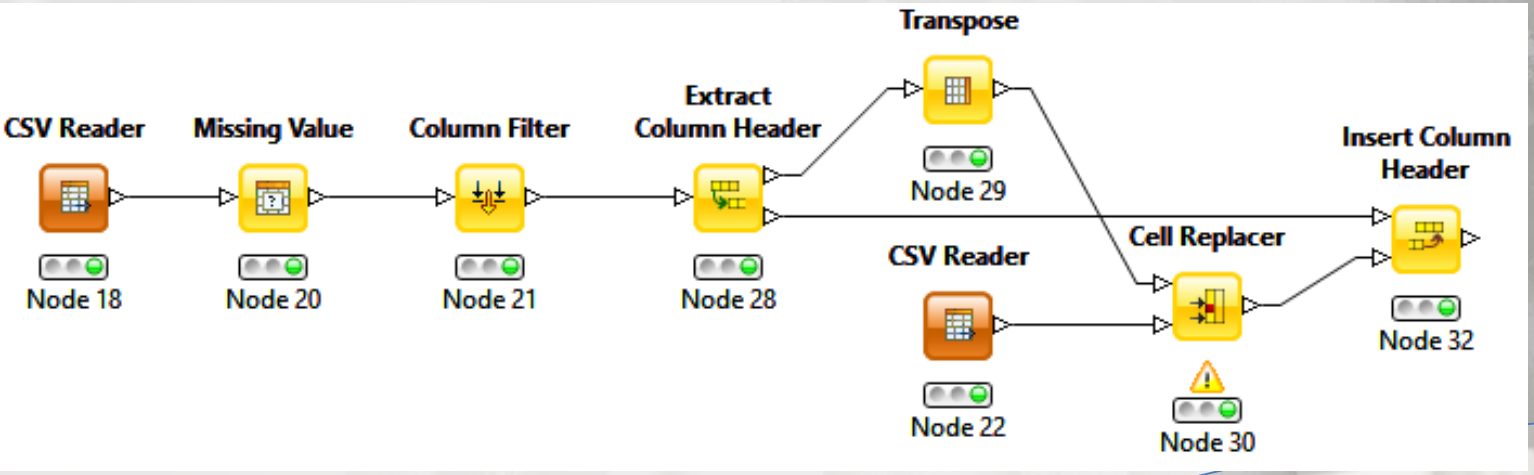


Deleting nodes (Highlight & Delete)



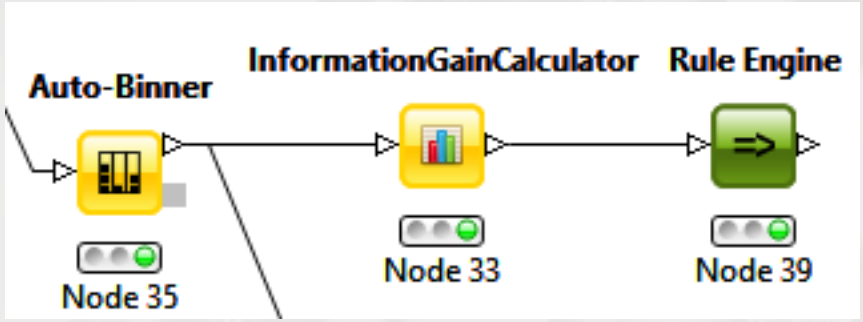
Replace Gene IDs with Protein IDs

- You will need Extract Column Header, Transpose, Cell Replacer and Insert Column Header
- Try to derive the correct configurations, call me if you are having issues



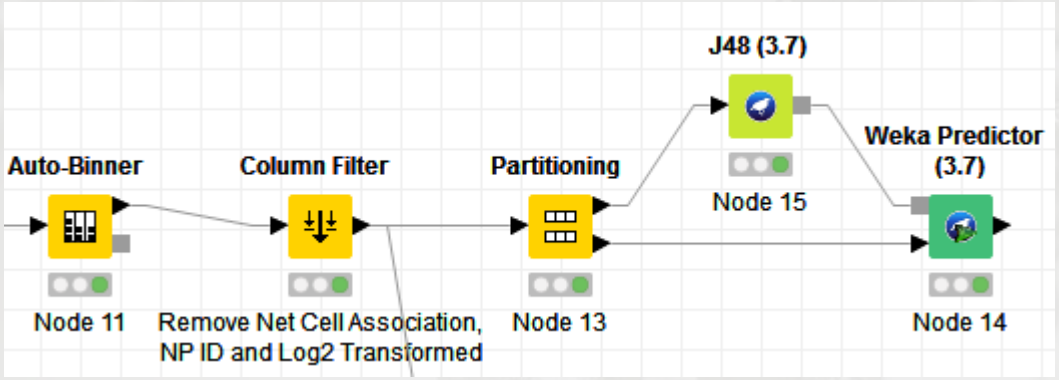
3 more useful nodes

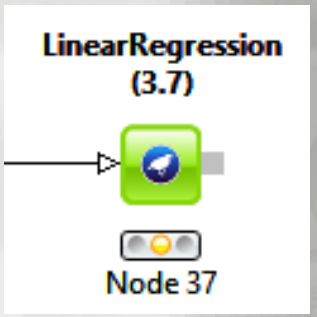
- Auto-binner: helps with simplification/categorization, visualization and classification algorithms
- Information gain: can pinpoint significant properties linked to the class
- Rule engine: can create a new attribute based on a custom rule on one or more attributes



Building simple models

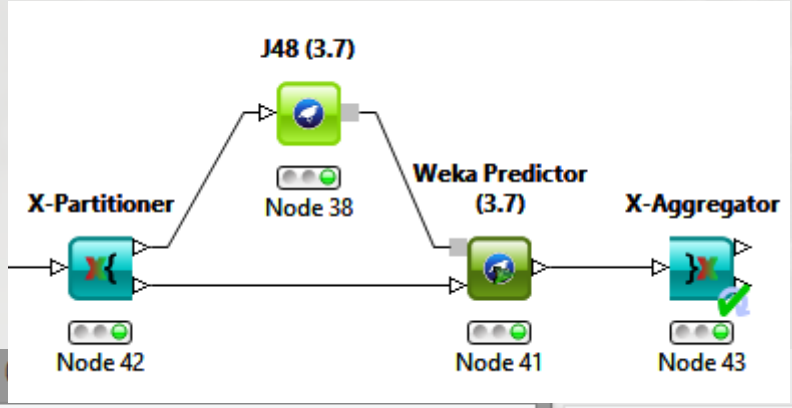
- Try to build a model using continuous and categorical Y (preferably from WEKA: J48 and Linear Regression*)
- Use the scorer nodes to report accuracy/R^2





*make sure variable is switched to continuous

Cross Validation



Prediction table -

File

Table "default" - Rows: 84 | Spec - Columns: 821 | Properties | Flow Variables

Row ID	Net cell...	Log2 tr...	Core size	Density	MW
Row8	0.006	-7.294	14.9	19.1	197
Row16	0.032	-4.975	14.9	19.1	197
Row22	0.016	-5.928	14.9	19.1	197
Row23	0.019	-5.736	14.9	19.1	197
Row33	0.012	-6.361	14.9	19.1	197
Row44	0.01	-6.611	14.9	19.1	197
Row47	0.04	-4.652	14.9	19.1	197
Row49	0.092	-3.436	31.6	19.1	197
Row54	0.005	-7.599	31.6	19.1	197
Row9	0.005	-7.59	14.9	19.1	197
Row10	0.458	-1.128	14.9	19.1	197

Error rates - 0:43 - X-Aggregato

File

Table "default" - Rows: 10 | Spec - Columns: 3 | Properties | Flow Vari

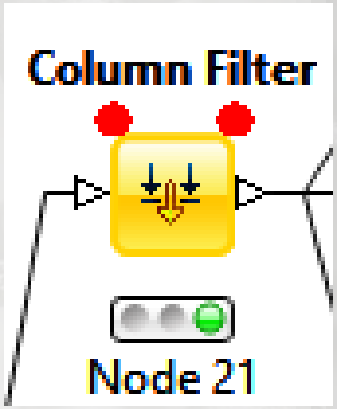
Row ID	Error in %	Size of ...	Error C...
fold 0	11.111	9	1
fold 1	25	8	2
fold 2	11.111	9	1
fold 3	0	8	0
fold 4	0	8	0
fold 5	22.222	9	2
fold 6	12.5	8	1
fold 7	0	9	0
fold 8	12.5	8	1
fold 9	25	8	2

Task 3

- Build at least 3 different models using stratified 5 and 10-fold cross validation
- Report your best results

If we have time – KNIME Variables

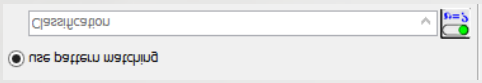
- Can use entries as variables and vice versa
- For filtering based on attribute
- For Looping
- For writing multiple files
- etc.



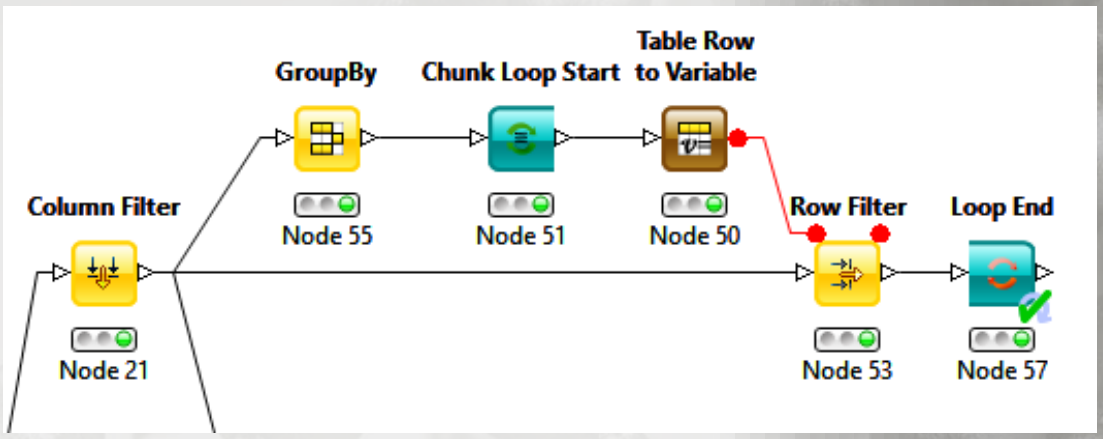
- Right Click -> Show variable ports

Variables in KNIME

- GroupBy Anionic/Cationic
- Loop (Start/End)
- Make value a variable
- Use it to filter rows

()

- Check collected results
- (per iteration/last Col)



Acknowledgements

Prof Harry Sarimveis

The Sarimveis Group at NTUA:

- Hampos Homenidis
- Dr Philip Doganis
- Dr Georgia Tsiliki
- Evangelia Anagnostopoulou
- Angelos Valsamis

The eNanoMapper consortium

Contact email: gdrakakis356@gmail.com