Introduction to

Open for Innovation
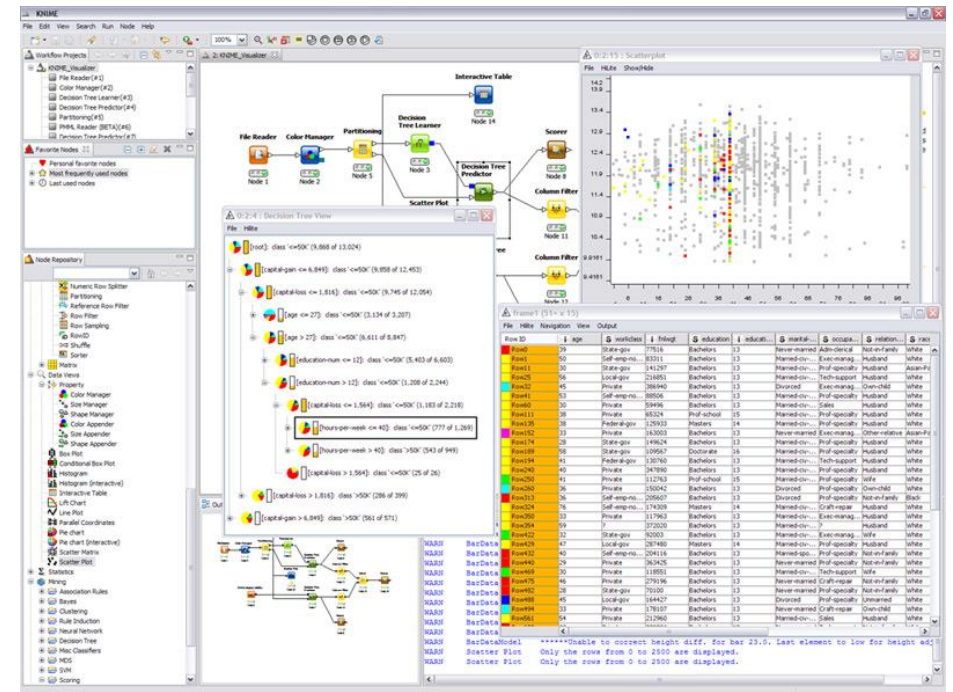
KNIME

January 23, 2017
Bernardo Najlis

Agenda

- What is Knime
- Where to get it and online resources
- What can I do with Knime
- How does it compare with similar tools
- Knime Lingo
- Knime Workbench
- How to build a Workflow
- Samples and Questions

Open for Innovation
KNIME

What is **Knime**?

- KNIME stands for Konstanz Information Miner
- It is an Open Source Data Analytics, Reporting and Integration platform
- Use a GUI to assembly 'nodes' for data preprocessing (ETL), modelling and data analysis and visualization
- Modules for:
  - Data Mining
  - Data Analysis
  - Data Manipulation
  - More modules and extensions can be added!
- Written in Java and based on Eclipse

Open for Innovation
KNIME

Where to get it and other online resources

- http://knime.org/downloads/overview
  - Skip the registration form, go straight to step (2) and download the version **with all free extensions (~2Gb)**
- **Community Forum and Online Self Training**
- **Books**

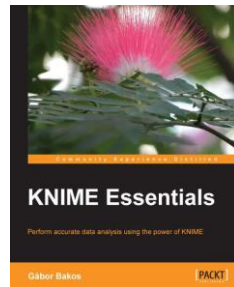 **KNIME Essentials**
  By: Gábor Bakos
  Publisher: Packt Publishing
  Pub. Date: October 16, 2013
  Print ISBN-13: 978-1-84969-921-1
  Web ISBN-13: 978-1-84969-922-8
  Pages in Print Edition: 148

- **Videos**

  Introduction to Data Analytics with **KNIME**
  By: Rosaria Silipo
  Publisher: Infinite Skills
  Publication Date: 20-SEP-2016
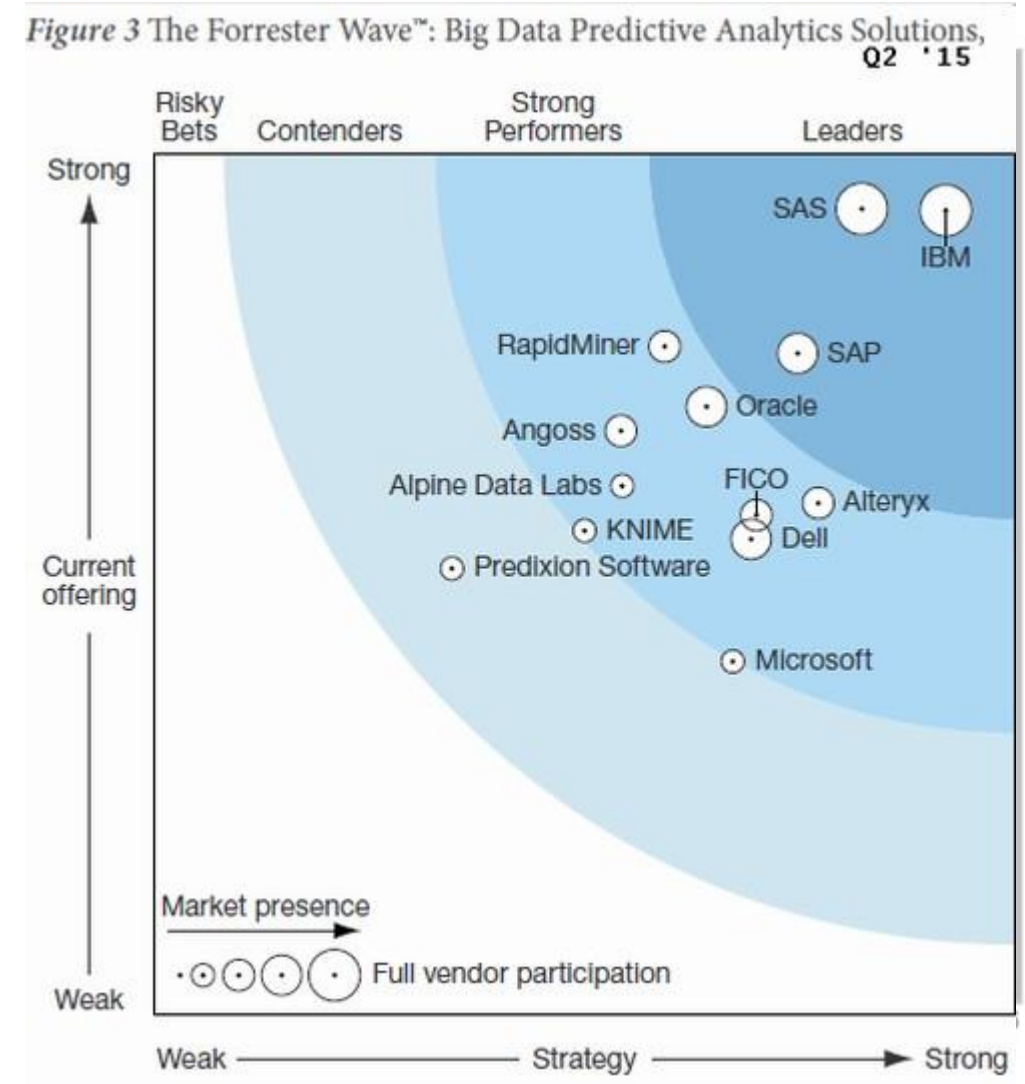 **Insert Date:** 26-SEP-2016

Open for Innovation
KNIME

What can I do with Knime?

- Data Access
  - File
  - Database I/O
- Transformation
  - Filtering, Grouping, Joining
- Analyze and Data mining
  - Weka
  - R
  - Python
  - Mathlab
- Visualization
 - Different types of charts
- Deployment
  - Text mining
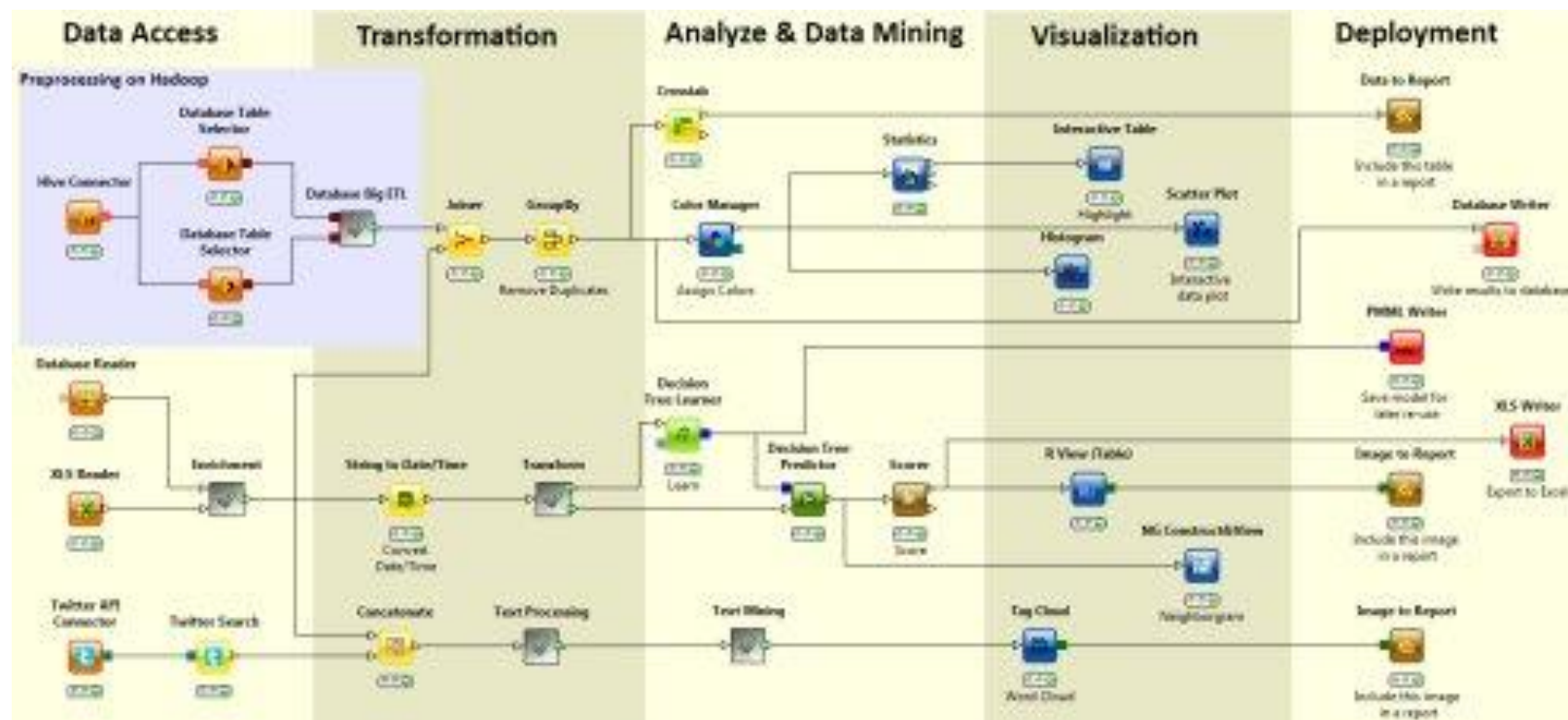
How does Knime compare with others?

- Gartner's Magic Quadrant for Advance Analytics Platforms

- Leaders quadrant in 2016 with SAS, IBM and Dell

- Strong Performer / Contender in Forrester's Wave

Open for Innovation
KNIME

Figure 3 The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 '15

Knime Lingo

- Store your work in a **workspace**
- **Workspace** can contain **workflow groups** built using the **workflow editor**
- **Workflows** can contain **nodes, meta nodes, connections, workflow variables, workflow credentials** and **annotations**
- Each **node** has a type, which identifies the algorithm associated with it
- **Nodes** have **parameters, inports** and **outports**, and can have any of these states:
  - **Misconfigured**
  - **Configured**
  - **Queued for Execution**
  - **Running**
  - **Executed**
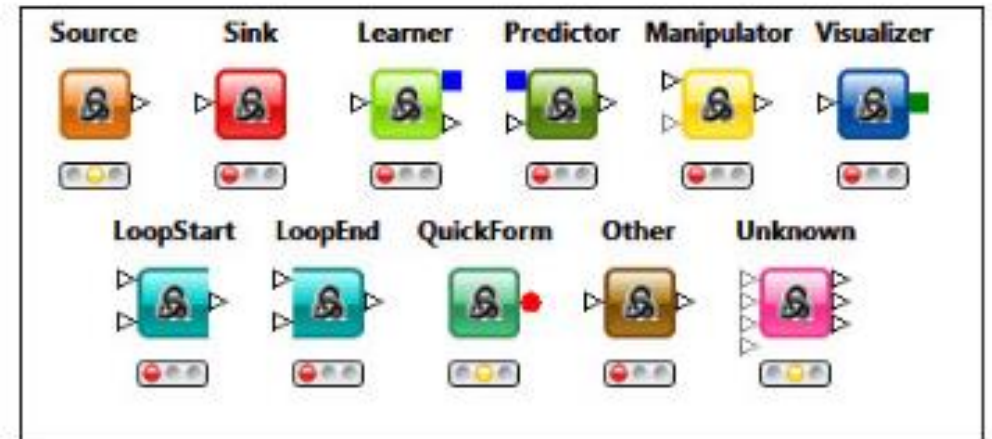
Knime Workbench

- Workflow Projects
- Favorite Nodes
- Node Repository
- Workflow Editor
- Outline
- Node Description
- Console

Open for Innovation
KNIME
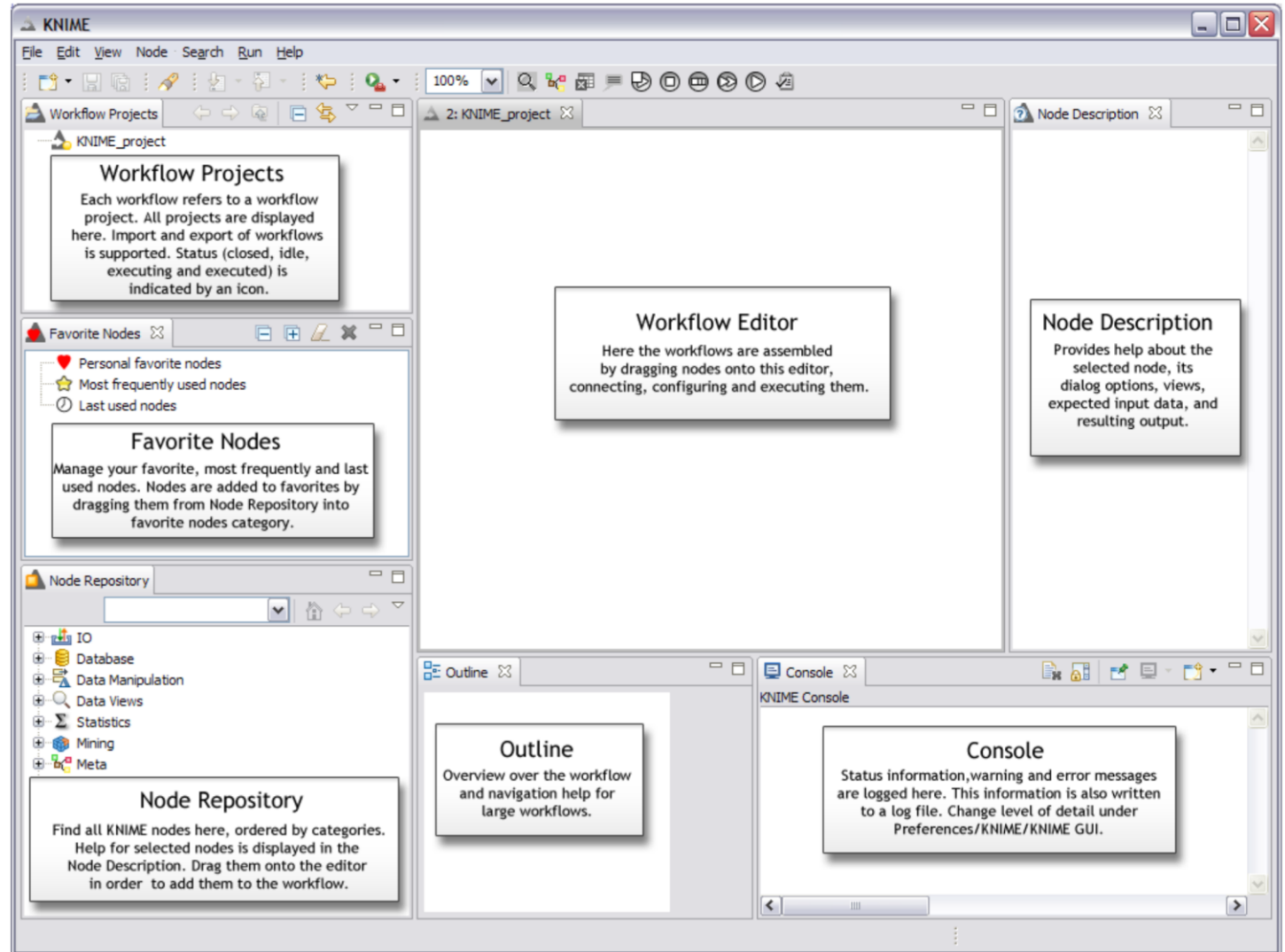
How to Build a Knime Workflow

- Search in Node Repository

- Dragging nodes into Workflow Editor

- Connecting Nodes

- Configuring Nodes

- Executing (per node or one-shot)
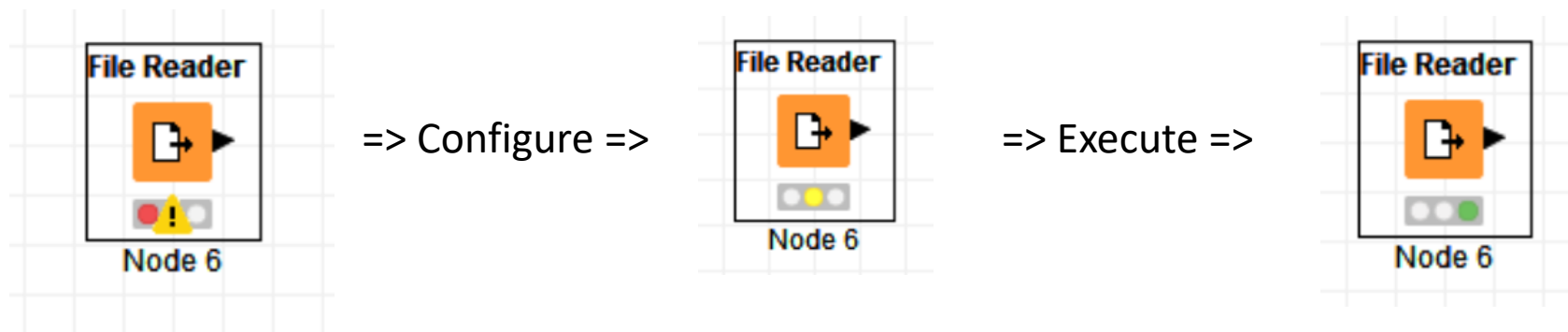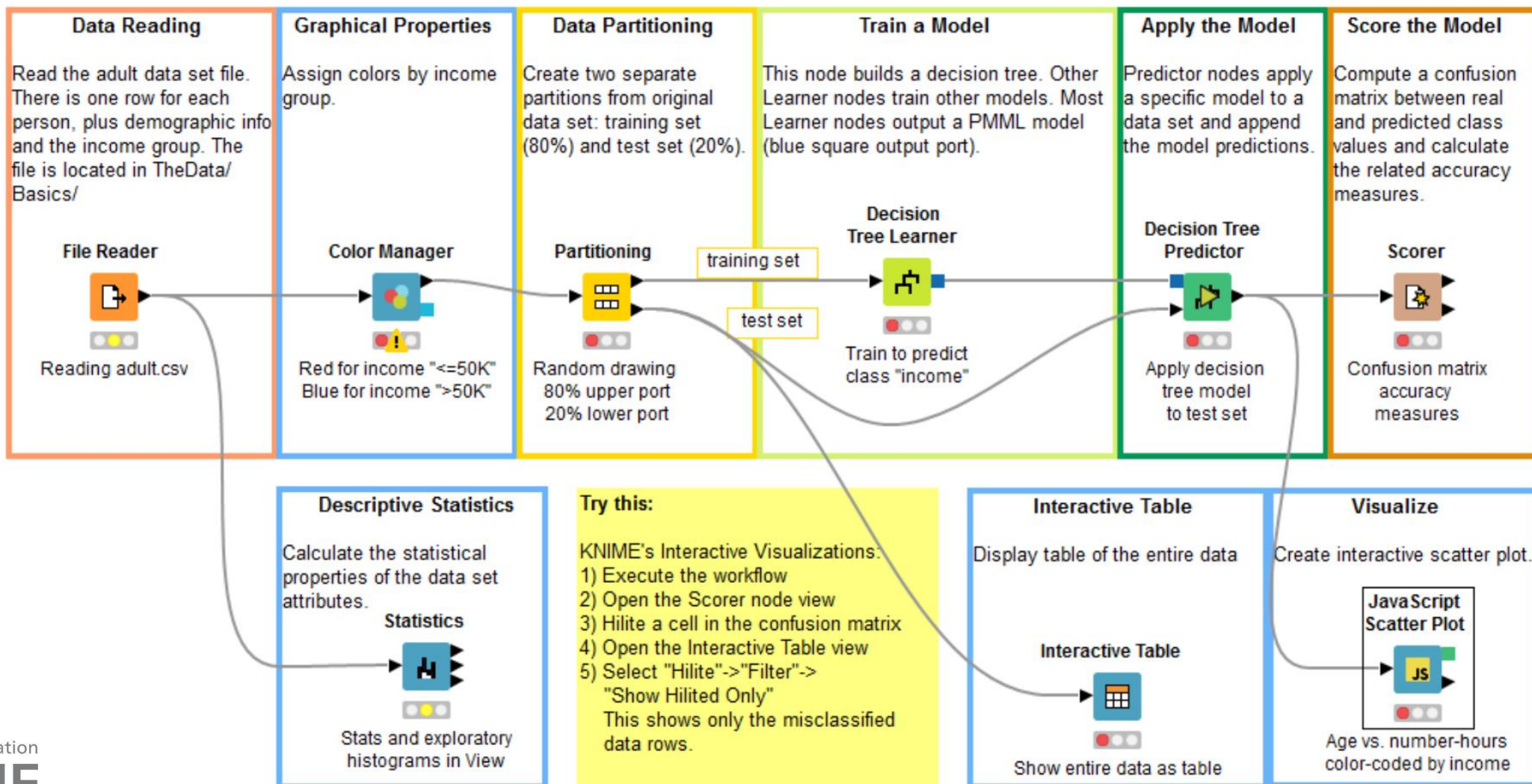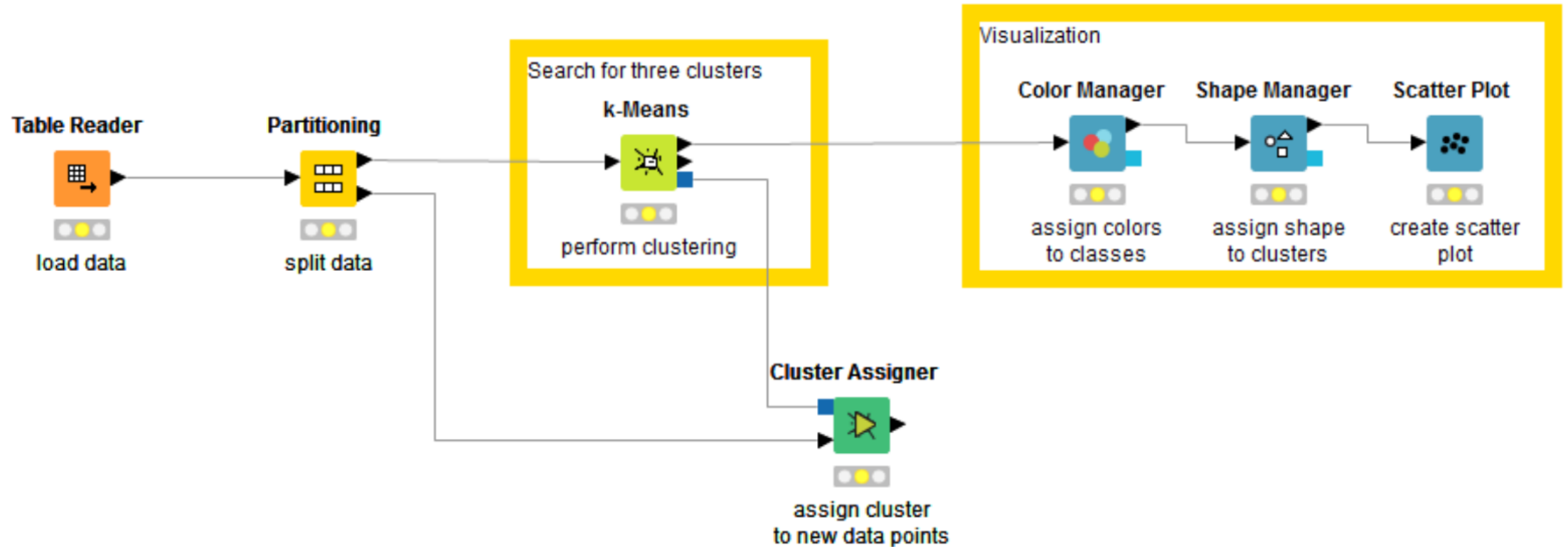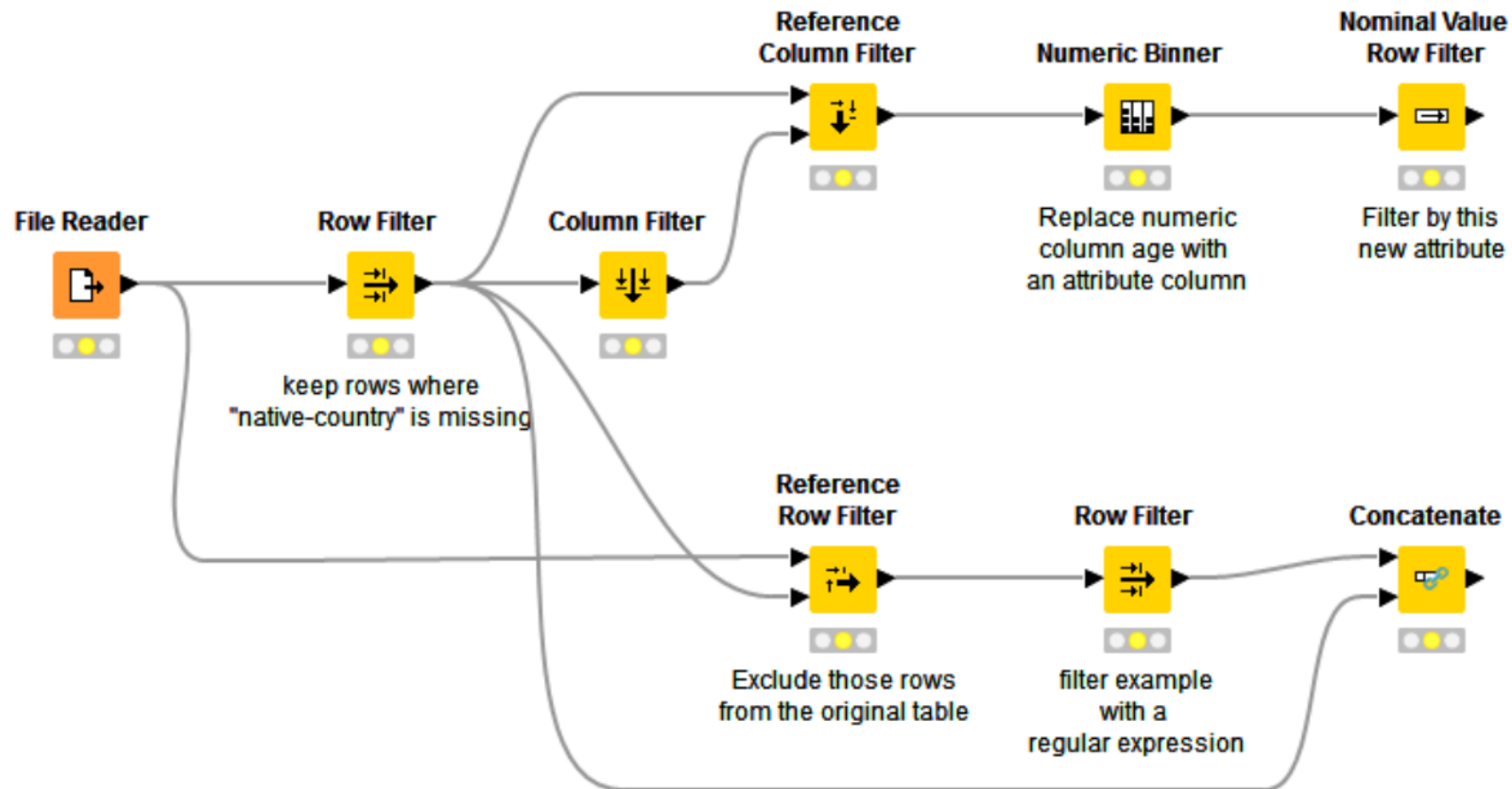

=> Configure =>   => Execute =>

Simple Model Training for Classification

Data Reading

Read the adult data set file. There is one row for each person, plus demographic info and the income group. The file is located in TheData/ Basics/

File Reader

Reading adult.csv

Graphical Properties

Assign colors by income group.

Color Manager

Red for income "<=50K" Blue for income ">50K"

Data Partitioning

Create two separate partitions from original data set: training set (80%) and test set (20%).

Partitioning

Random drawing 80% upper port 20% lower port

Train a Model

This node builds a decision tree. Other Learner nodes train other models. Most Learner nodes output a PMML model (blue square output port).

Decision Tree Learner

training set

test set

Train to predict class "income"

Apply the Model

Predictor nodes apply a specific model to a data set and append the model predictions.

Decision Tree Predictor

Apply decision tree model to test set

Score the Model

Compute a confusion matrix between real and predicted class values and calculate the related accuracy measures.

Scorer

Confusion matrix accuracy measures

Descriptive Statistics

Calculate the statistical properties of the data set attributes.

Statistics

Stats and exploratory histograms in View

Try this:

KNIME's Interactive Visualizations:
1) Execute the workflow
2) Open the Scorer node view
3) Hilite a cell in the confusion matrix
4) Open the Interactive Table view
5) Select "Hilite"->"Filter"->
   "Show Hilited Only"
   This shows only the misclassified data rows.

Interactive Table

Display table of the entire data

Interactive Table

Show entire data as table

Visualize

Create interactive scatter plot.

JavaScript Scatter Plot

JS

Age vs. number-hours color-coded by income

Open for Innovation
KNIME

Performing k-means Clustering

Example for Data Preprocessing

Example of R Snippet