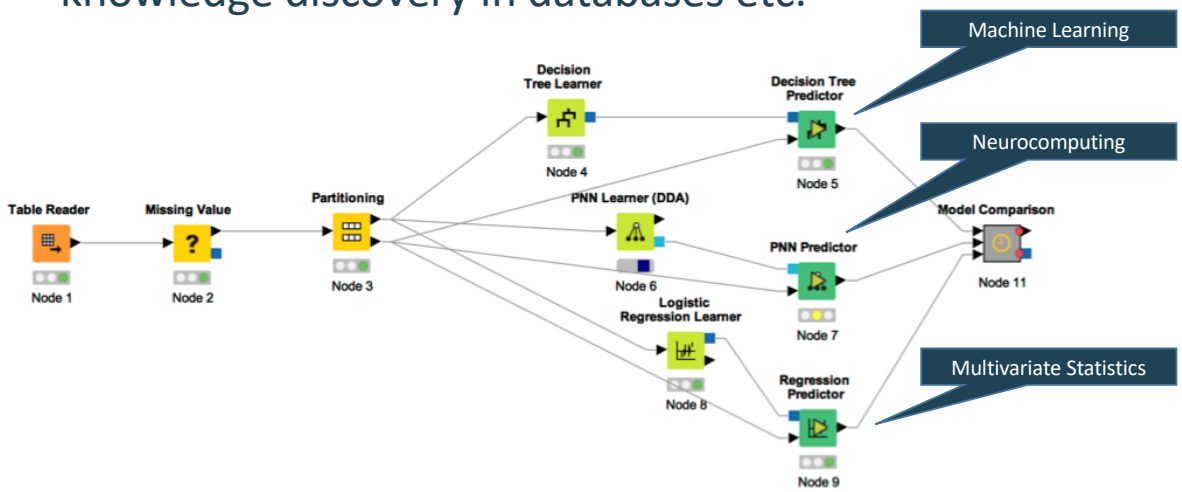
Agenda

1. Introduction to Advanced Analytics
2. Process Models
3. Myths and Pitfalls
4. Categorization of Advanced Analytics Algorithms
5. Fields of Application (exemplary use cases)
6. Challenges
7. Roles in a Data Science Team
8. The Value of Analytics

1.
Introduction to Advanced Analytics

1. What is Advanced Analytics?

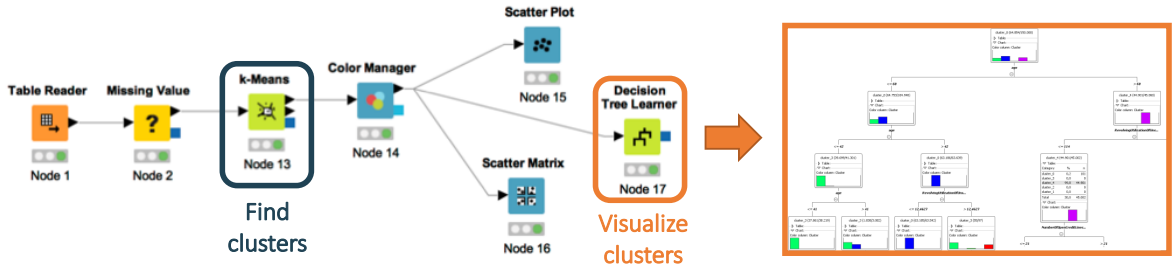
- ...is a **combination** of different academic fields like multivariate statistics, artificial intelligence, machine learning, pattern recognition, neurocomputing, knowledge discovery in databases etc.



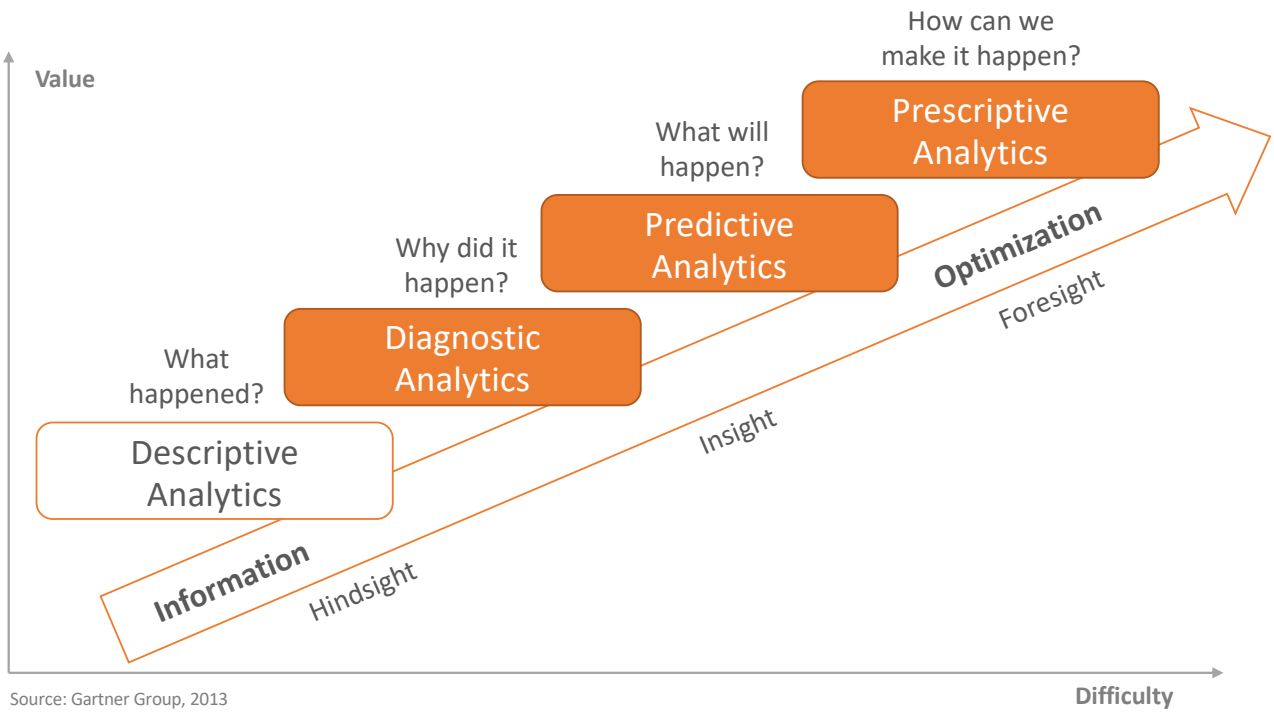
1. What is Advanced Analytics?

- …is the **process** of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or identify business critical hidden patterns.

Combination of Cluster Analysis & Decision Trees

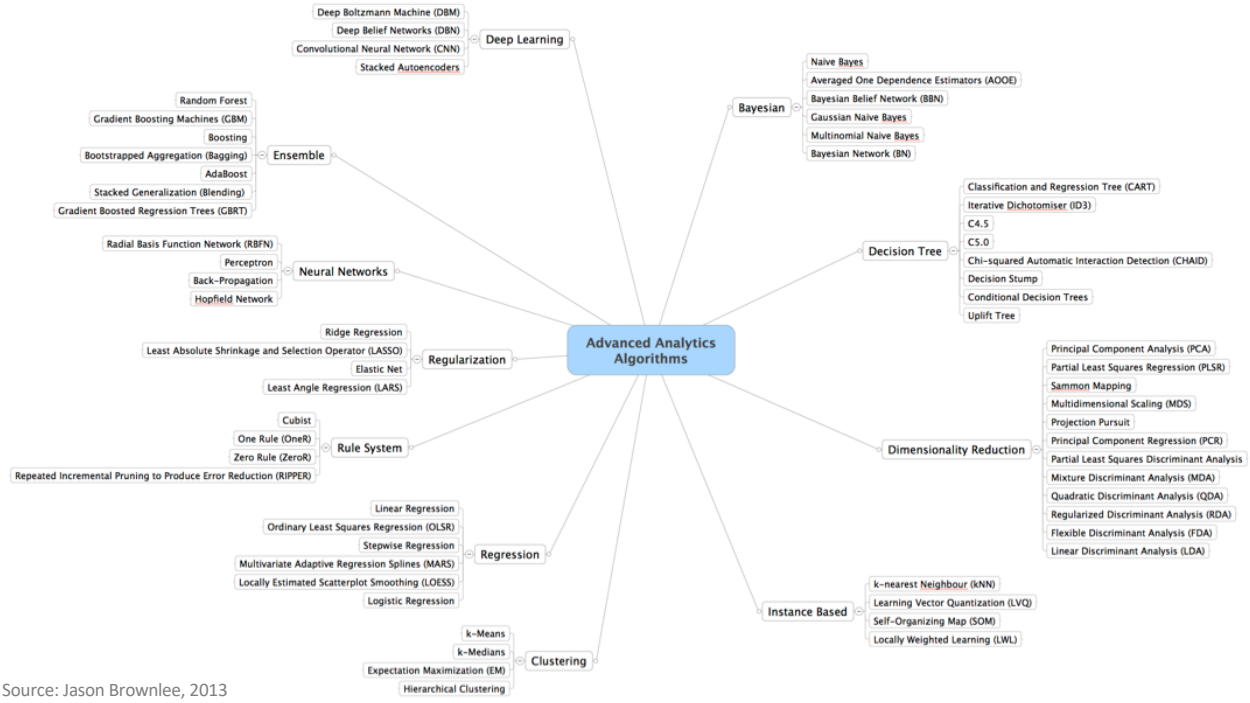


Advanced Analytics – Beyond Descriptive Analytics



Source: Gartner Group, 2013

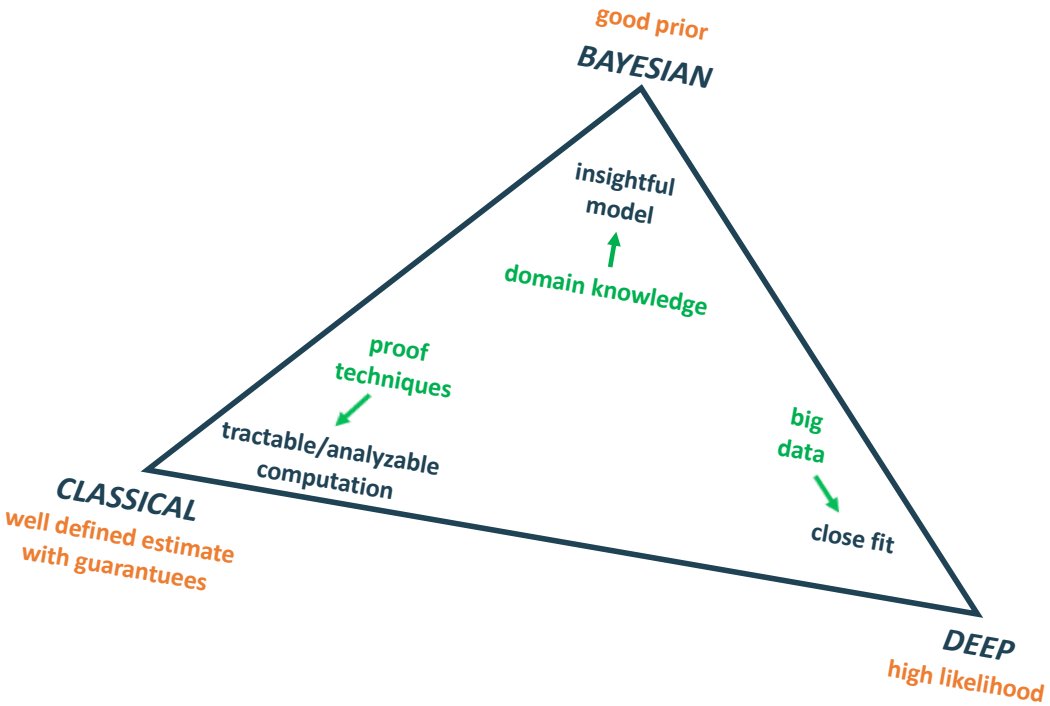
An overview of Advanced Analytics Algorithms



Source: Jason Brownlee, 2013

AdvancedAnalytics .Academy

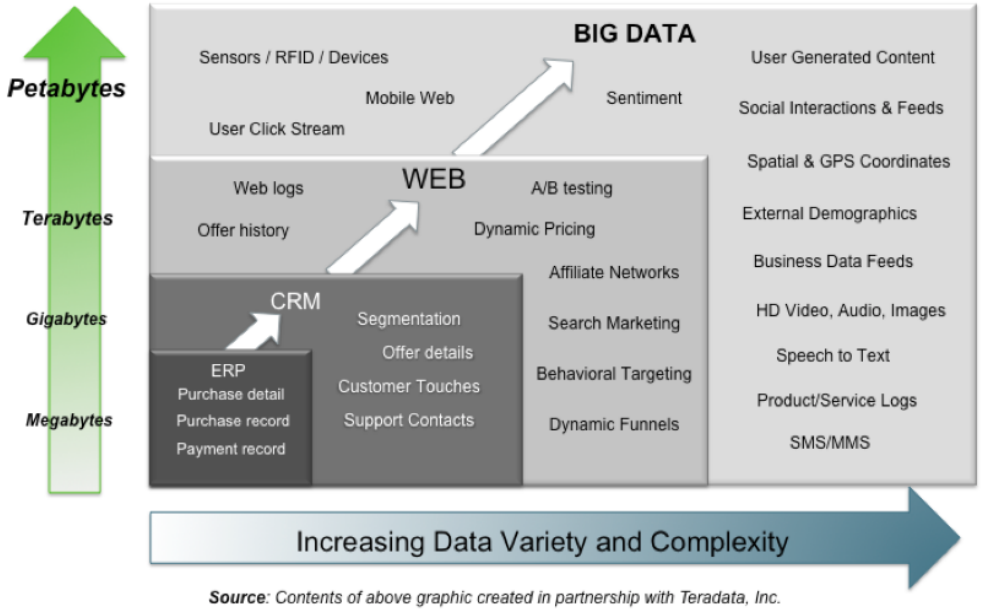
The Three Cultures of Machine Learning



Source: Jason Eisner, 2015

AdvancedAnalytics .Academy

Determining factors for the rise of Big Data Analytics



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

AdvancedAnalytics .Academy

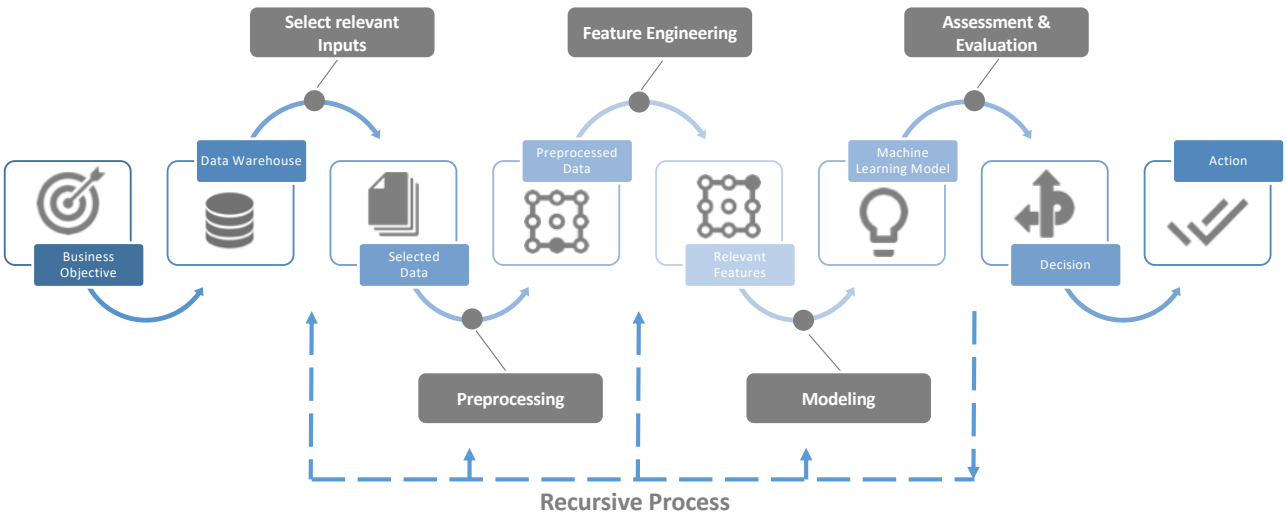
Developments

- The rise of open source…
- Algorithm development, e.g.
  - Ensemble Models
  - Deep Learner
- Hyper Parameterization
- Parallel Processing
- Online Learning
- Analytics Automation

AdvancedAnalytics .Academy

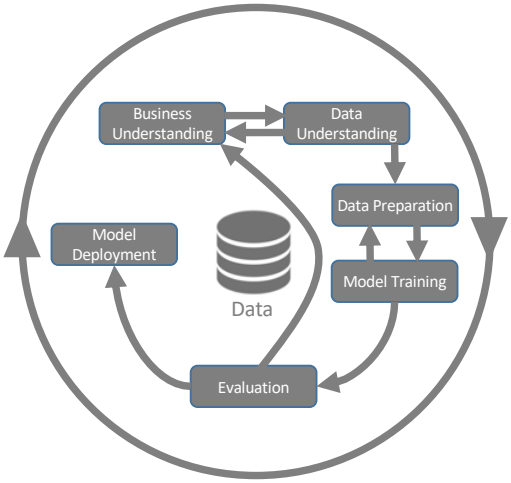
2.
Process Models

2. A typical Advanced Analytics Process



2. CRISP-DM

- CRISP-DM = Cross Industry Standard Process for Data Mining.
- A data mining process model that describes commonly used approaches that expert data miners use to tackle problems.
- CRISP-DM breaks the process of data mining into six phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Model Training
  - Evaluation
  - Model Application



AdvancedAnalytics .Academy

3.
Advanced Analytics Myths and Pitfalls

AdvancedAnalytics .Academy

“

If you’ve got terabytes of data, and you’re relying on data mining to find interesting things in there for you, you’ve lost before you’ve even begun.

“

Herb Edelstein

AdvancedAnalytics .Academy

3. Advanced Analytics Myths

- Advanced Analytics provides instant crystal ball-predictions
- Advanced Analytics is all about algorithms
- Advanced Analytics is all about predictive accuracy
- A big data warehouse is a requirement for Advanced Analytics
- Advanced Analytics is all about vast quantities of data
- Advanced Analytics is for large companies with lots of customer data
- Only those who know every aspect of the technology must perform Advanced Analytics

AdvancedAnalytics .Academy

3. Advanced Analytics Pitfalls

- Buried under mountains of data
- The Mysterious Disappearing Terabyte
- Disorganized Advanced Analytics
- Insufficient business knowledge
- Insufficient data knowledge
- Erroneous assumptions, courtesy of the experts
- Incompatibility of Advanced Analytics tools

4.
Categorization of Advanced Analytics Algorithms

4. Categorization of Advanced Analytics Algorithms

Supervised Learning
(Predictions & Classifications)

1. Decision Trees
2. Logistic Regression
3. Neural Networks
4. Bayes Classifier
5. Support Vector Machines
6. Ensemble Models

Unsupervised Learning
(Knowledge Discovery)

1. Cluster Analysis
2. Self-Organizing Maps (SOM)
3. Association Algorithms
4. Sequence Analysis

Semi-Supervised Learning

1. Active Learning
2. Generative Models
3. Low-density Separation
4. Graph-Based Algorithms
5. Multiview Algorithms

5.
Fields of Application (exemplary Use Cases)

5. Fields of Application (examples)

Customer Segmentation

Customer segmentation, also referred to as market segmentation, is the process of finding **homogenous** sub-groups within a **heterogeneous** aggregate customer base.

Propensity Modeling (Churn, Next Best Offer etc)

1. Predict customer churn by assessing their propensity of risk to churn.
2. Predict customer need and behavior by assessing their propensity to buy a product.

Association Analysis (eg. Market Basket Analysis)

Association rules are employed today in many application areas including market basket analysis, web usage mining, intrusion detection and bioinformatics.

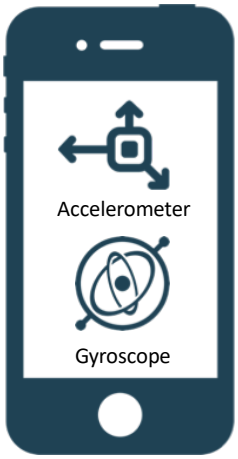
Fraud Detection & Money Laundering

Anticipate illegal or suspicious activities and transactions – such as identity theft, insurance fraud and money laundering by applying predictive analytics methods.

IoT Analytics Use Case Activity Detection with DSP & Machine Learning

IoT Analytics

Overview



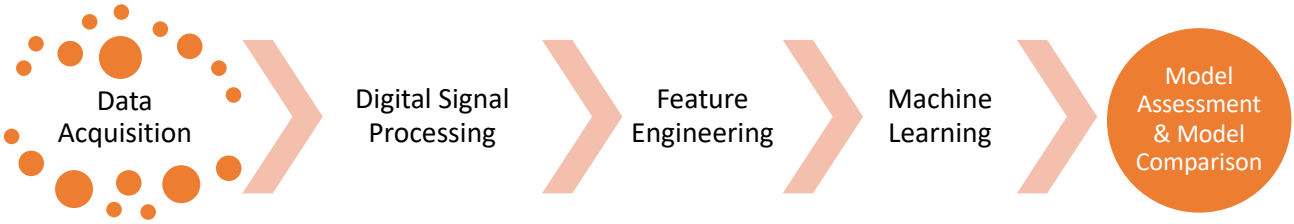
Smart Phone

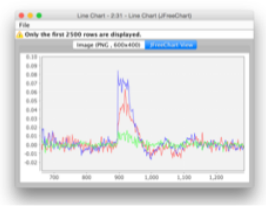
Digital Signal Processing & Machine Learning

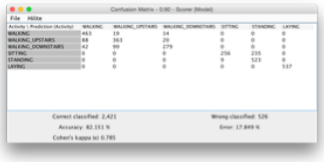
Activity Classification

AdvancedAnalytics .Academy

Analytical Process







AdvancedAnalytics .Academy

6.
Roles in a Data Science Team

The role of the Citizen Data Scientist (Gartner, 2017)

- Many users who traditionally work with BI tools
- Focus on descriptive analytics

- Few experts with deep analytical skills
- Focus on predictive and prescriptive analytics

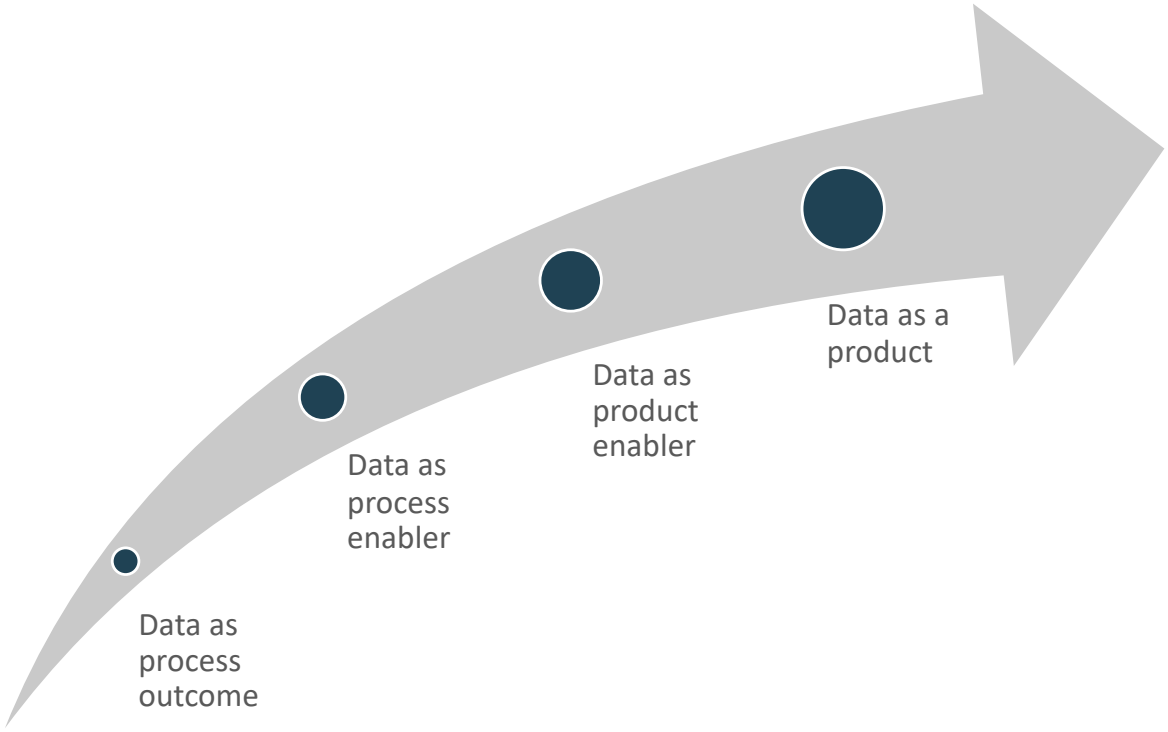
Citizen Data Scientists bridge the gap between mainstream self-service analytics by business users and the advanced analytics techniques of data scientists.

The role of the Citizen Data Scientist (Gartner, 2017)

- A Citizen Data Scientist is a person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities, but whose primary job function is outside the field of statistics and analytics.
- Citizen Data Scientists bridge the gap between mainstream self-service analytics by business users and the advanced analytics techniques of data scientists.
- They are able to perform sophisticated analysis that would previously have required more expertise, enabling them to deliver advanced analytics without having the skills that characterize data scientists.

7.
The Value of Analytics

Different roles of data

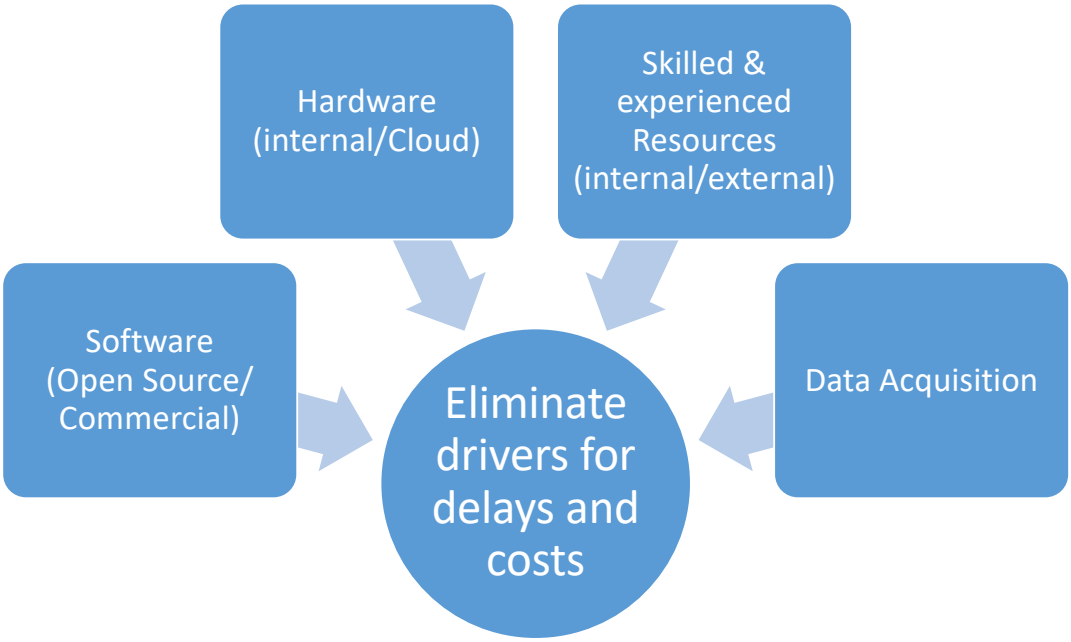


Data as…

- **process outcome**
  - Measurements, KPI
  - Quality assurance
  - Support main processes
- **process enabler**
  - virtual data
  - used to control MES, ERP, PP
  - E-commerce, electronic cash
- **product enabler**
  - IoT, e.g. smart home, smart wearables, low-cost health/medical equipment
  - Sharing economy, e.g. ride-sharing, esp. mobile/hyper local
- **the product itself**
  - market research data, data broker
  - GPS navigation systems
  - ”iTunes/Netflix”

➡ **Depending on the overall data strategy the impact and setup of advanced analytics initiatives are completely different.**

Best practice: Establish an analytics playground for a permanent test&learn environment



Overview
KNIME Analytics Platform

Additional Resources

KNIME pages (www.knime.org)

- **SOLUTIONS** for example workflows
- RESOURCES/**LEARNING HUB** www.knime.org/learning-hub
- RESOURCES/**NODE GUIDE** https://www.knime.org/nodeguide

KNIME Tech pages (tech.knime.org)

- **FORUM** for questions and answers
- **DOCUMENTATION** for docs, FAQ, changelogs, ...
- **COMMUNITY CONTRIBUTIONS** for dev instructions and third party nodes

KNIME TV on **YouTube** https://www.youtube.com/user/KNIMETV

What is KNIME Analytics Platform?

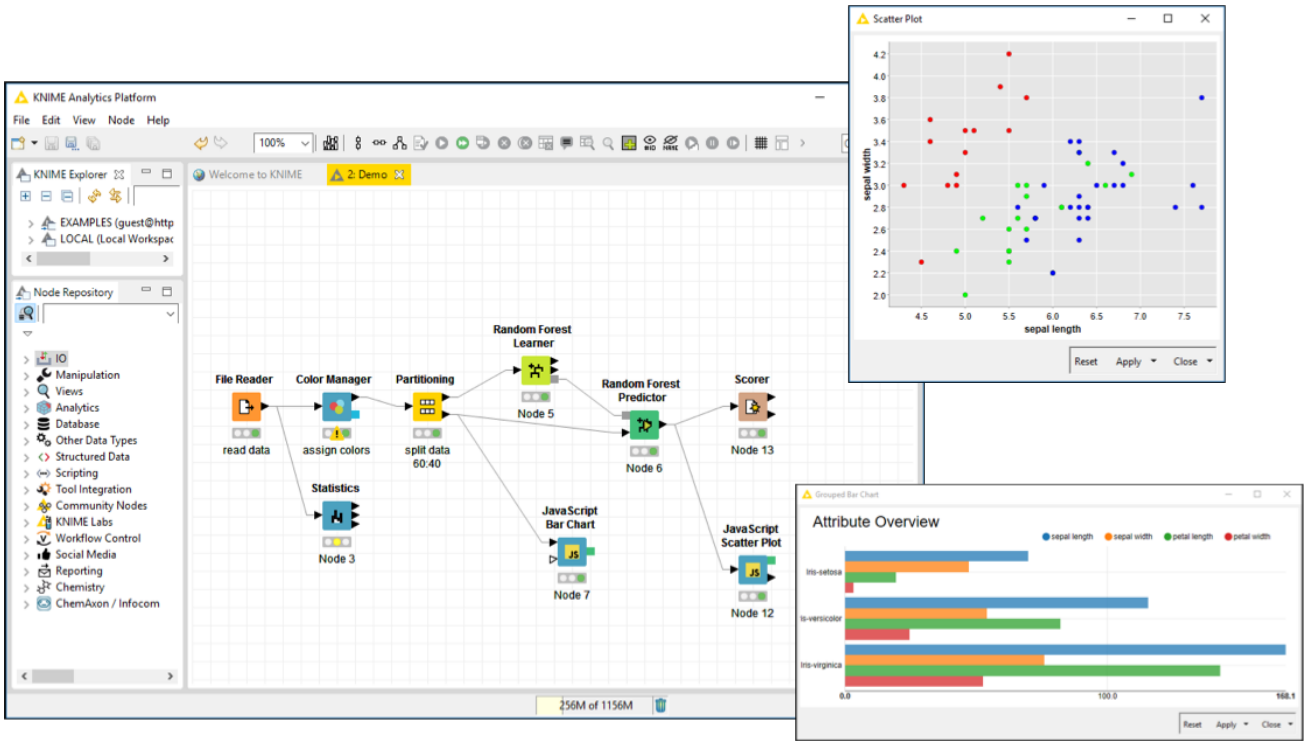
- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Weka machine learning
  - Many integrations, such as Java, R, Python, H20, Keras/TensorFlow etc.

KNIME – A Leader in Data Science

Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms

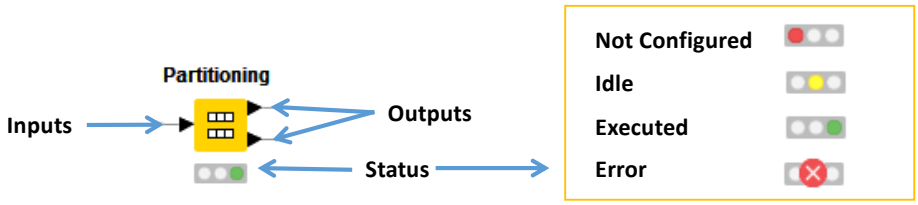
Source: Gartner (January 2019)

The KNIME® Analytics Platform

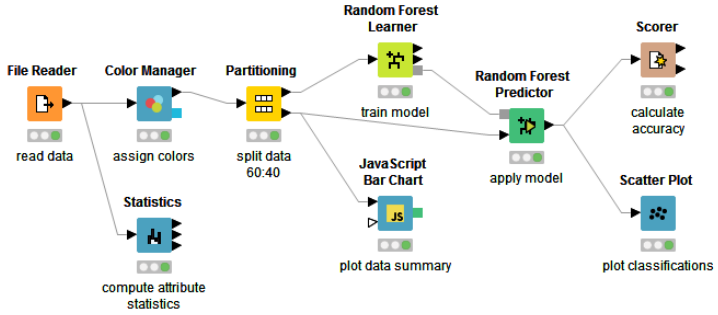


Visual KNIME Workflows

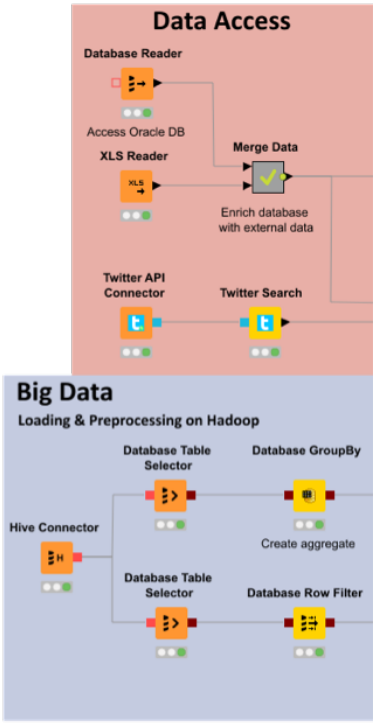
NODES perform tasks on data



Nodes are combined to create **WORKFLOWS**

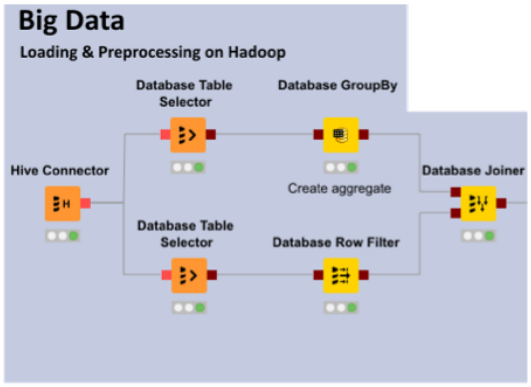


Data Access



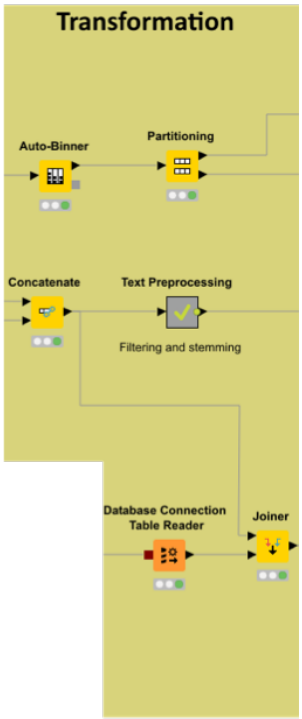
- Databases
  - MySQL, PostgreSQL
  - any JDBC (Oracle, DB2, MS SQL Server)
- Files
  - CSV, TXT
  - Excel, Word, PDF
  - SAS, SPSS
  - XML
  - PMML
  - Images, texts, networks, chem
- Web, Cloud
  - REST, Web services
  - Twitter, Google

Big Data



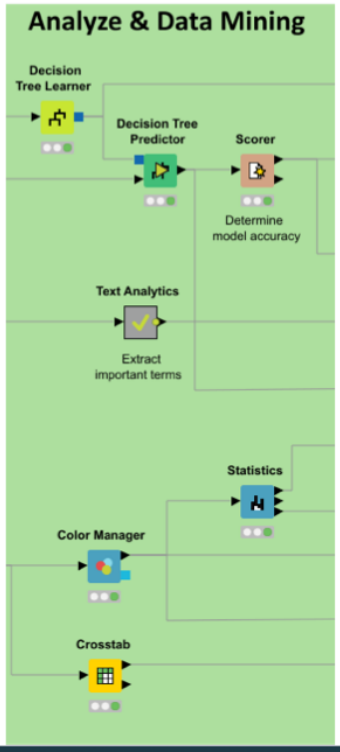
- Spark
- HDFS support
- Hive
- Impala
- HP Vertica
- In-database processing

Transformation



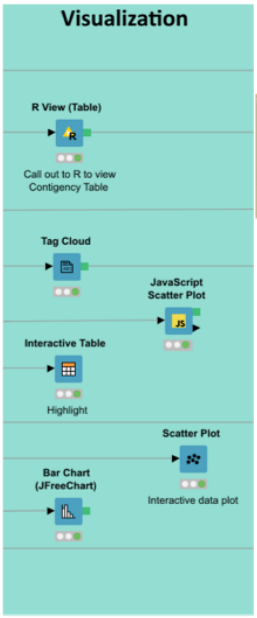
- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature Creation and Selection

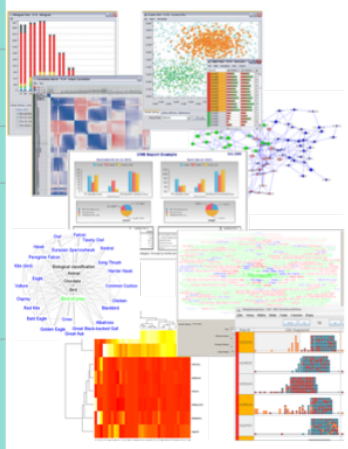
Analyze & Data Mining



- Regression
  - Linear, logistic
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Misc
  - PCA, MDS, item set mining
- External
  - Python, R, Weka

Visualization



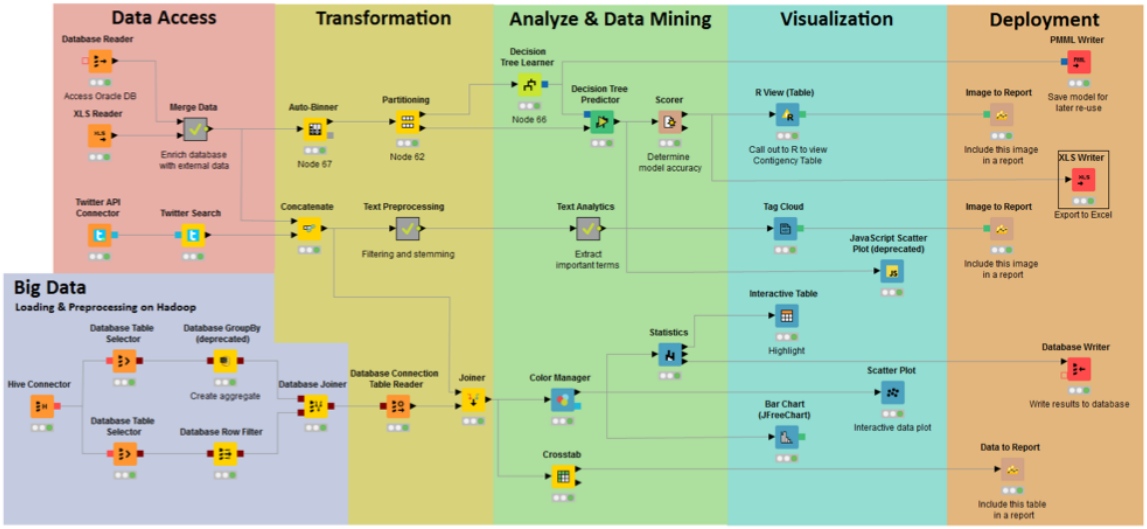


- Interactive
  - Scatter plot, histogram, pie charts, box plot
  - Highlighting (brushing)
- JFreeChart
- JavaScript
- Misc
  - Tag cloud, open street map, networks, molecules
- External
  - R

Deployment

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

Over 1500 native and embedded nodes included:



Data Access	Transformation	Analysis & Mining	Visualization	Deployment
MySQL, Oracle, ...	Row,	Statistics	R	via BIRT
SAS, SPSS, ...	Column	Data Mining	JFreeChart	PMML
Excel, Flat, ...	Matrix	Machine Learning	JavaScript	XML, JSON
Hive, Impala, ...	Text, Image	Web Analytics	Community / 3rd	Databases
XML, JSON, PMML	Time Series	Text Mining		Excel, Flat, etc.
Text, Doc, Image, ...	Java	Network Analysis		Text, Doc, Image
Web Crawlers	Python	Social Media Analysis		Industry Specific
Industry Specific	Community / 3rd	R, Weka, Python		Community / 3rd
Community / 3rd		Community / 3rd		

Overview

- Installing KNIME Analytics Platform
- The KNIME Workspace
- The KNIME File Extensions
- The KNIME Workbench
  - Workflow editor
  - Explorer
  - Node repository
  - Node description
  - Preferences
- Installing new features

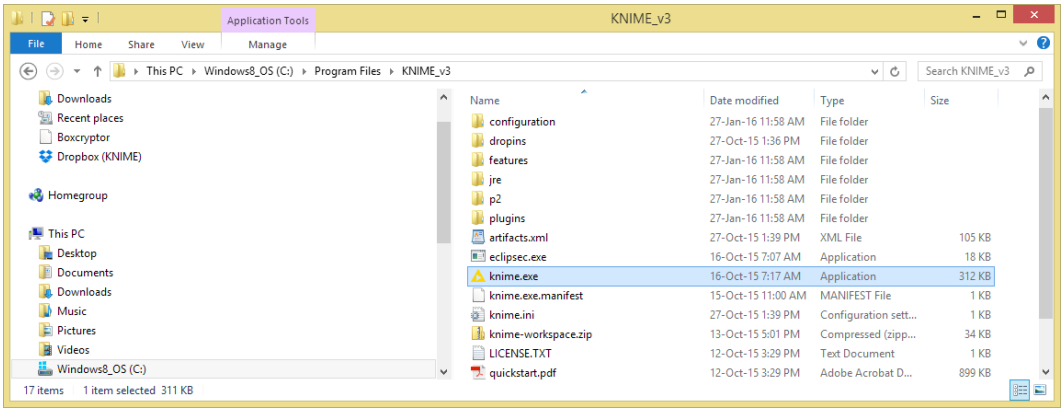
Install KNIME Analytics Platform

- Select the KNIME version for your computer:
  - Mac, Win, or Linux and 32 / 64bit
- Note different downloads (minimal or **full**)
- Download archive and extract the file, or download installer package and run it

Windows	
KNIME Analytics Platform for Windows (installer) *The installer adds an icon to the desktop and suggests suitable memory settings*	32 bit *(301,82 MB)* 64 bit *(305 MB)*
KNIME Analytics Platform + all free extensions for Windows (installer) *The installer adds an icon to the desktop and suggests suitable memory settings*	32 bit *(1,74 GB)* 64 bit *(1,92 GB)*
KNIME Analytics Platform for Windows (self-extracting archive) *The self-extracting archive only creates a folder holding the KNIME installation*	32 bit *(304,18 MB)* 64 bit *(306,45 MB)*
KNIME Analytics Platform for Windows (zip archive)	32 bit *(346,61 MB)* 64 bit *(349,93 MB)*
Linux	
KNIME Analytics Platform for Linux	32 bit *(363,02 MB)* 64 bit *(360,1 MB)*
KNIME Analytics Platform + all free extensions for Linux	32 bit *(1,66 GB)* 64 bit *(2,07 GB)*
Mac OSX	
KNIME Analytics Platform for Mac OSX (10.7 and above)	64 bit *(329,41 MB)*
KNIME Analytics Platform + all free extensions for Mac OSX (10.7 and above)	64 bit *(1,99 GB)*

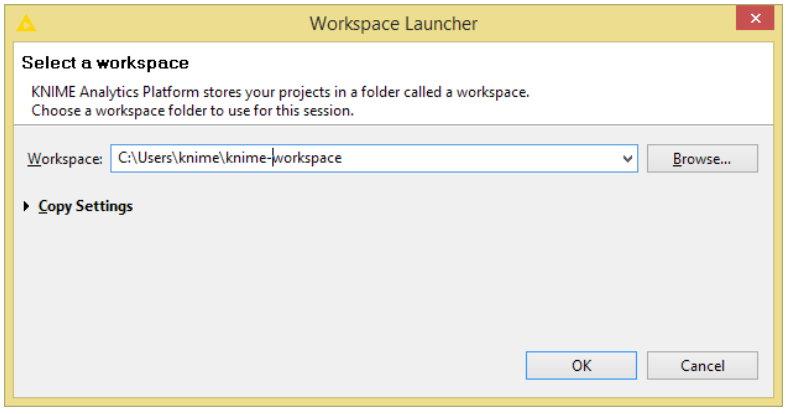
Start KNIME Analytics Platform

- Go to the installation directory and launch **KNIME**, or use the shortcut created on your Desktop.

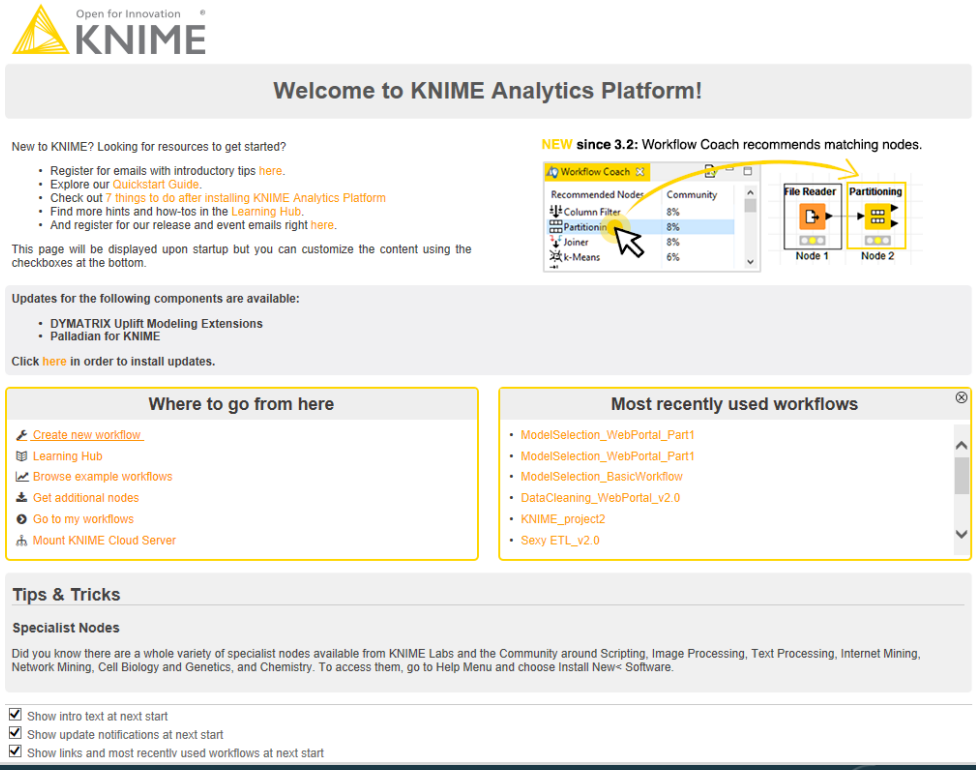


The KNIME Workspace

- 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)



Welcome Page



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


AdvancedAnalytics
.Academy
www.advancedanalytics.academy
Thank you for your attention!
Contact
Stefan Weingaertner
CEO
AdvancedAnalytics.Academy GmbH
E. sw@advancedanalytics.academy
T. +49 711 658 238 80
F. +49 711 658 238 88
M. +49 160 55 63 811