



AdvancedAnalytics
.Academy

www.advancedanalytics.academy

Introduction to Advanced Analytics

Stefan Weingaertner
Stuttgart, 28/02/2019



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

3

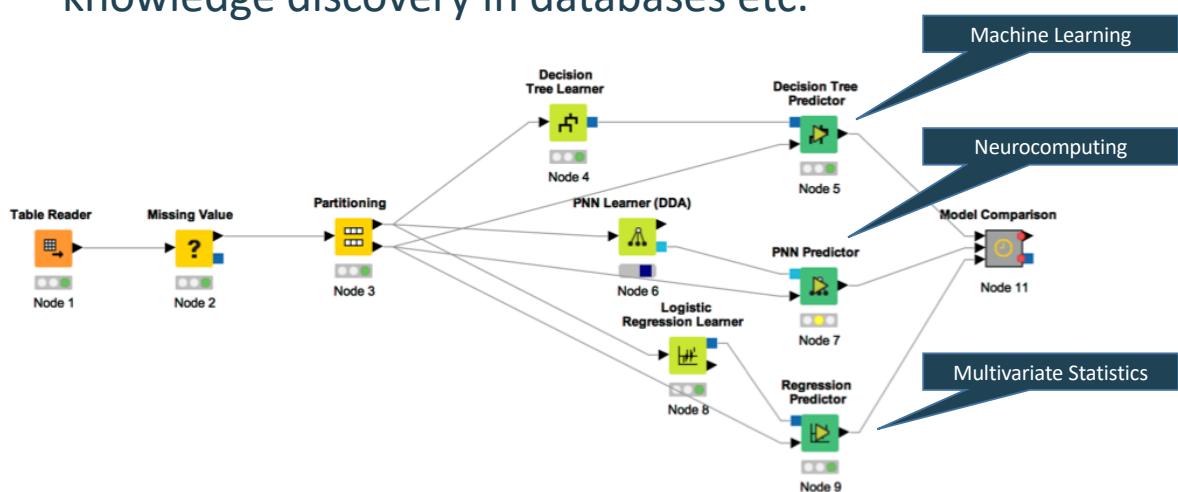
This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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.



4

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

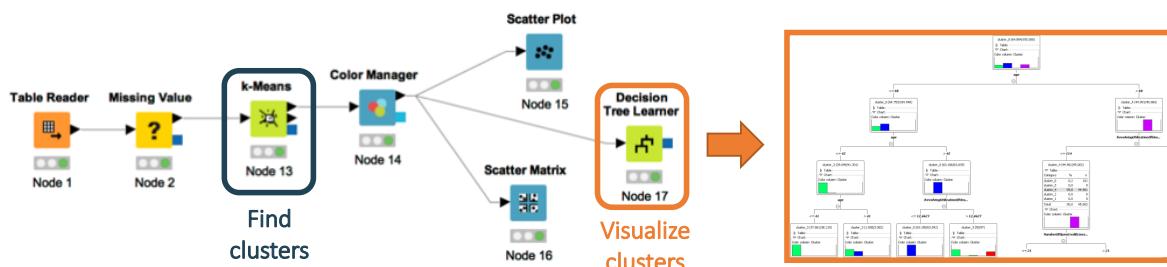
© AdvancedAnalytics.Academy GmbH



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



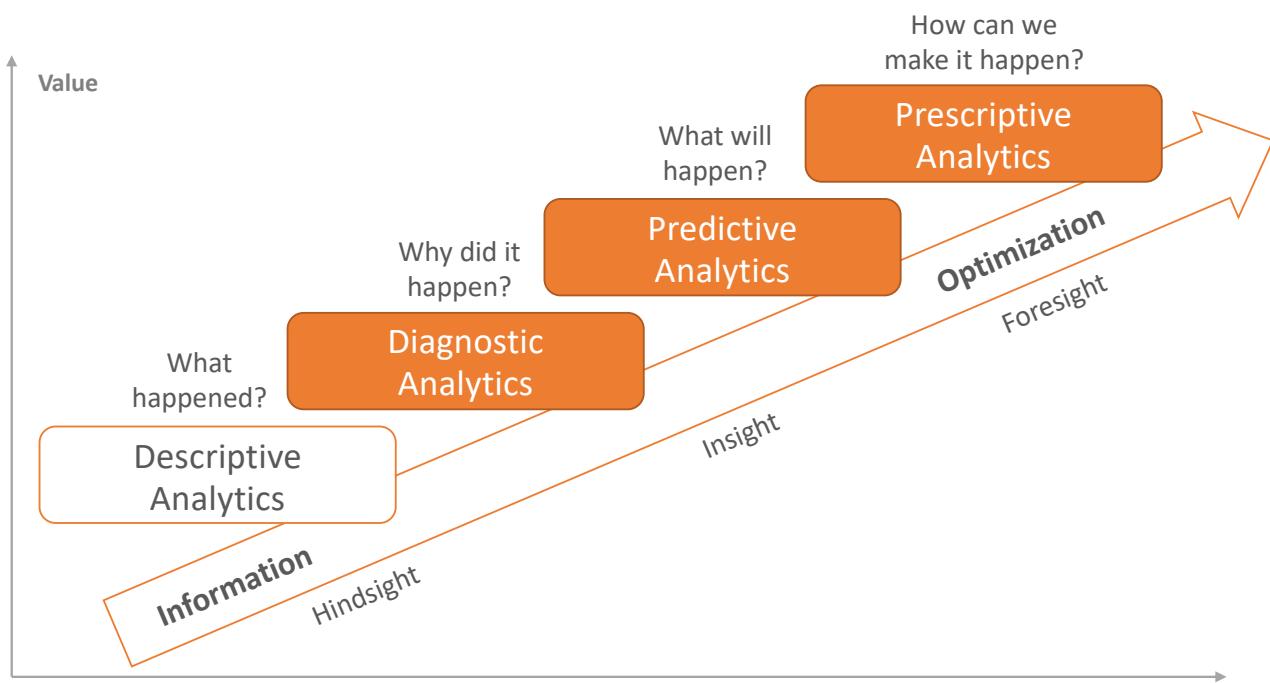
5

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



Advanced Analytics – Beyond Descriptive Analytics



Source: Gartner Group, 2013

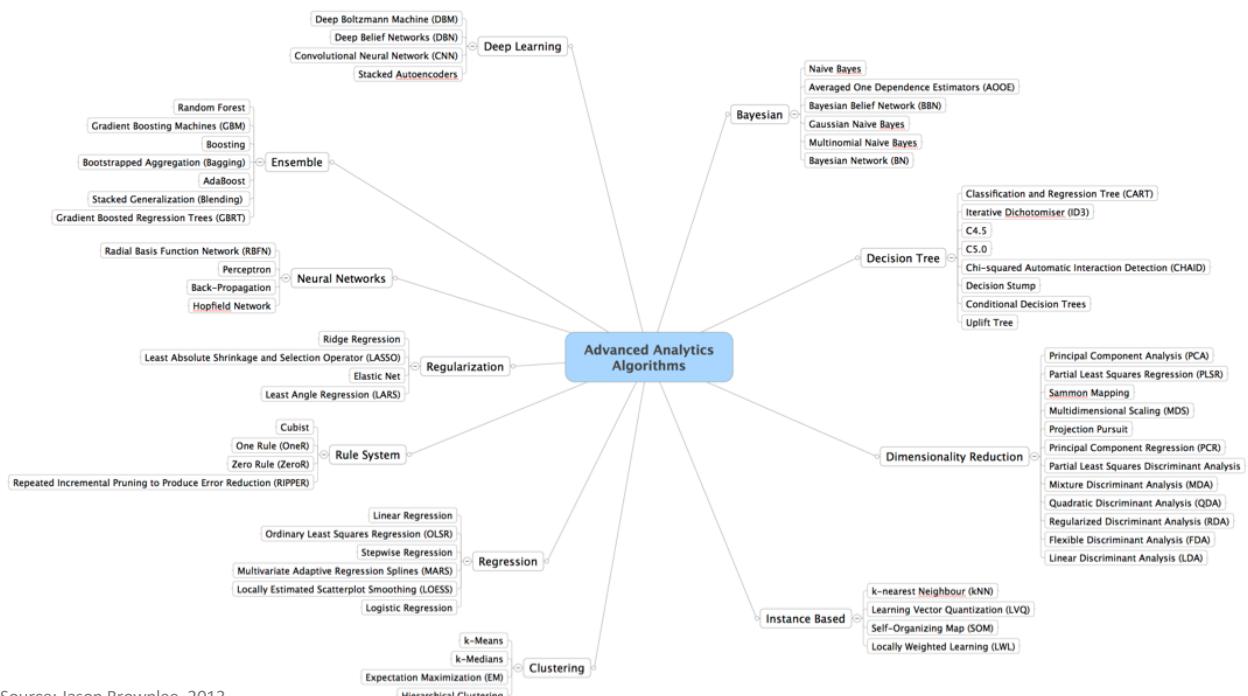
6

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



An overview of Advanced Analytics Algorithms



Source: Jason Brownlee, 2013

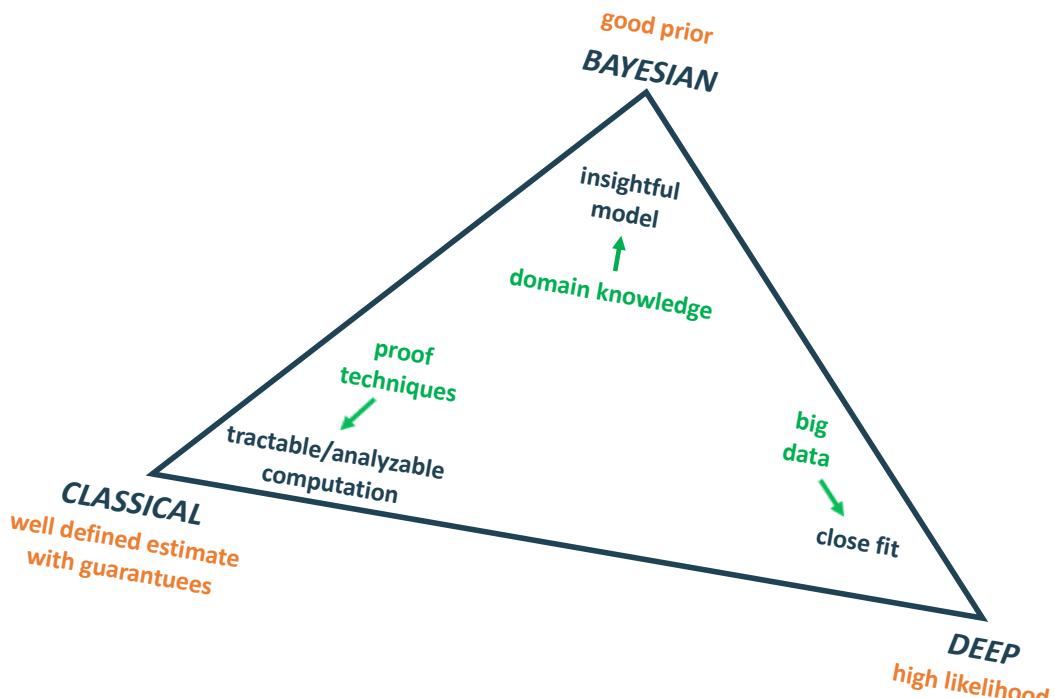
7

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



The Three Cultures of Machine Learning



Source: Jason Eisner, 2015

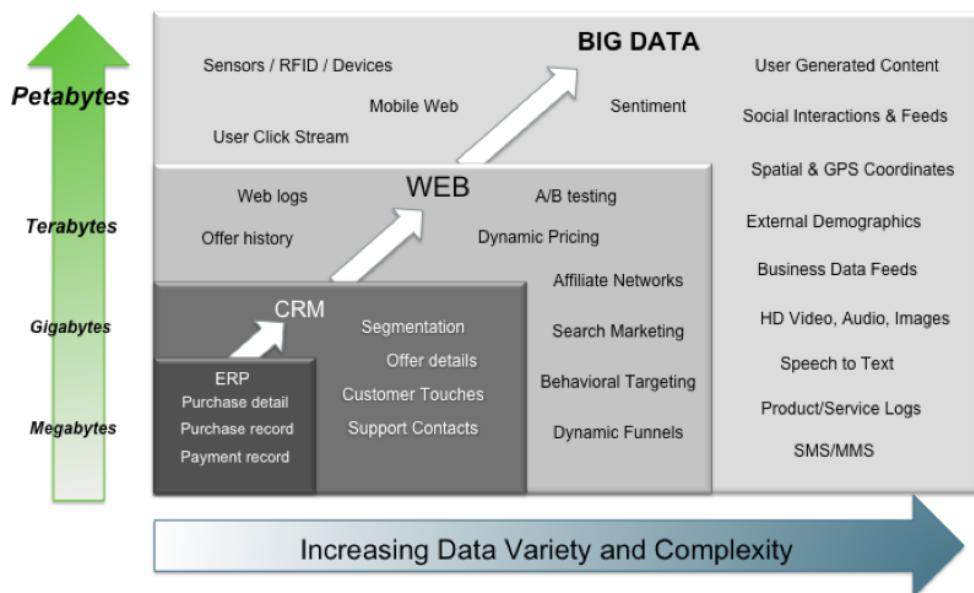
8

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



Determining factors for the rise of Big Data Analytics



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

9

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



Developments

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

10

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



2. Process Models

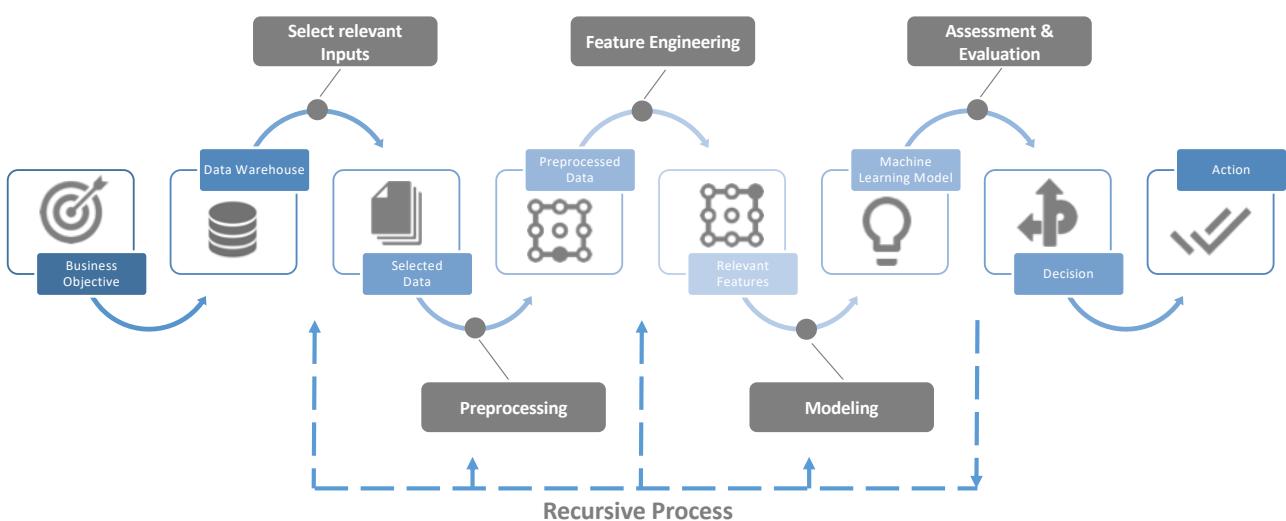
11

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



2. A typical Advanced Analytics Process



12

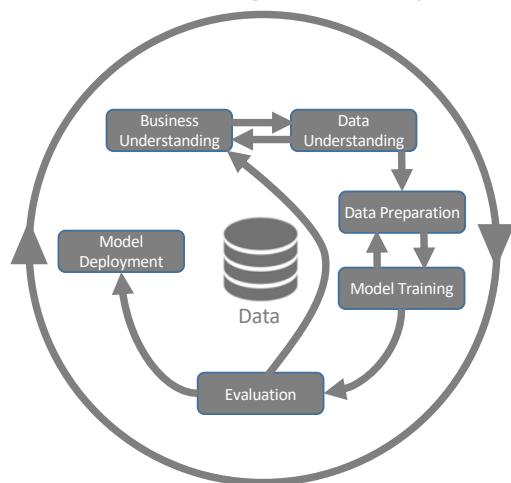
This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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



13

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



3. Advanced Analytics Myths and Pitfalls

14

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH





“

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

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

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

17

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



4.

Categorization of Advanced Analytics Algorithms

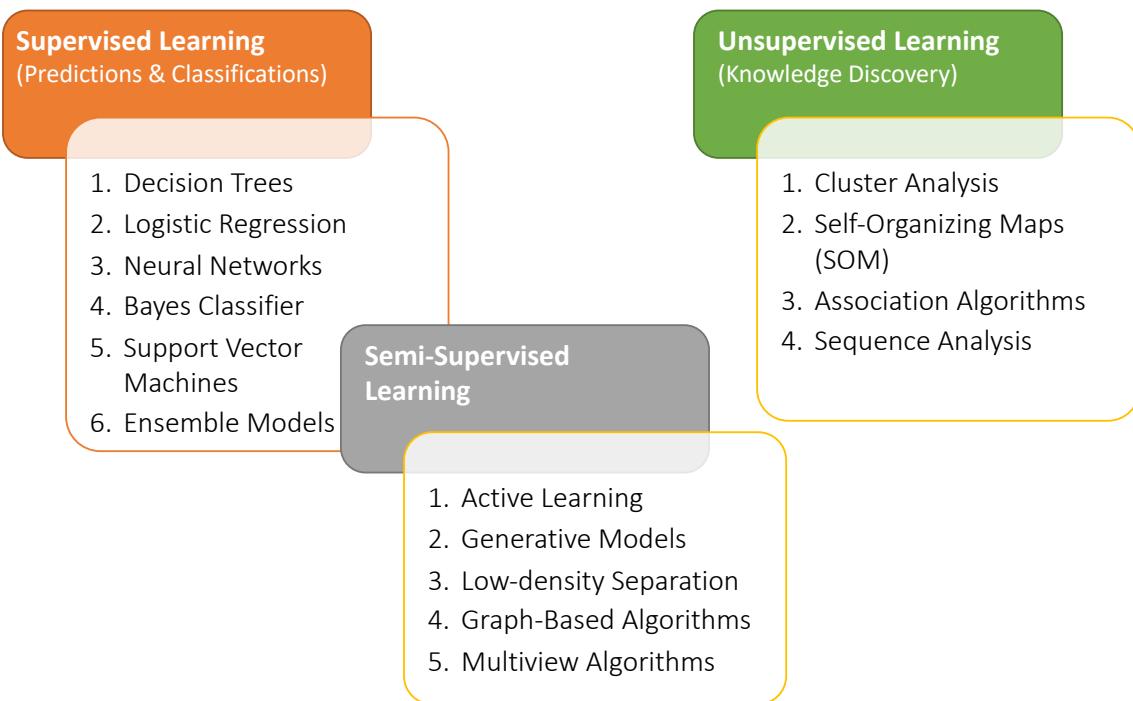
18

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



4. Categorization of Advanced Analytics Algorithms



19

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



5. Fields of Application (exemplary Use Cases)

20

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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.

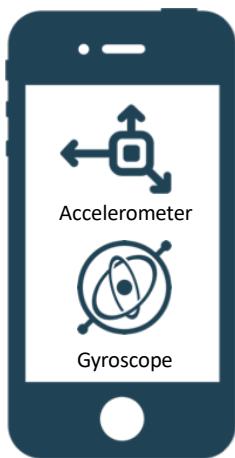
Fraud Detection & Money Laundering

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

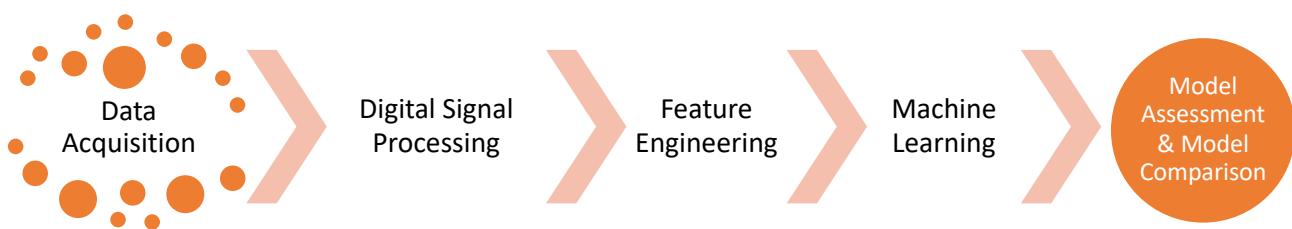


Digital Signal Processing & Machine Learning



Activity Classification

Analytical Process



Confusion Matrix - 0.90 - Score (Mean)						
	File	Hide	Activity	Prediction (Actual)	WALKING	WALKING_UPSTAIRS
				WALKING	19	14
				WALKING_UPSTAIRS	0	279
				SITTING	0	0
				SITTING_STANDING	0	0
				STANDING	256	235
				STANDING_WALKING	123	123
				WALKING_DOWNSTAIRS	0	0
				WALKING_UPSTAIRS_DOWNSTAIRS	0	0
				DOWNSTAIRS	0	0
				LAYING	0	0
						537

Correct classified: 540	Wrong classified: 526
Accuracy: 82.0 %	Error: 17.44 %
Cohen's kappa is: 0.29	

6.

Roles in a Data Science Team

25

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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



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

26

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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.

27

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



7.

The Value of Analytics

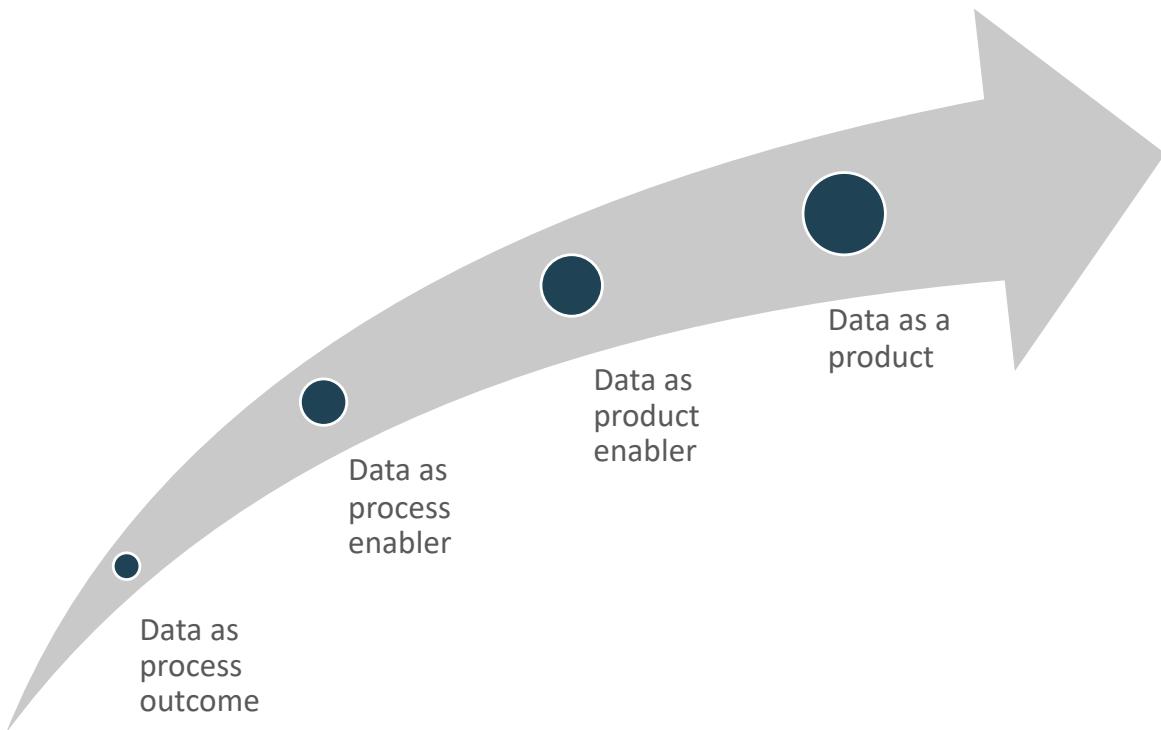
28

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



Different roles of data



29

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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.

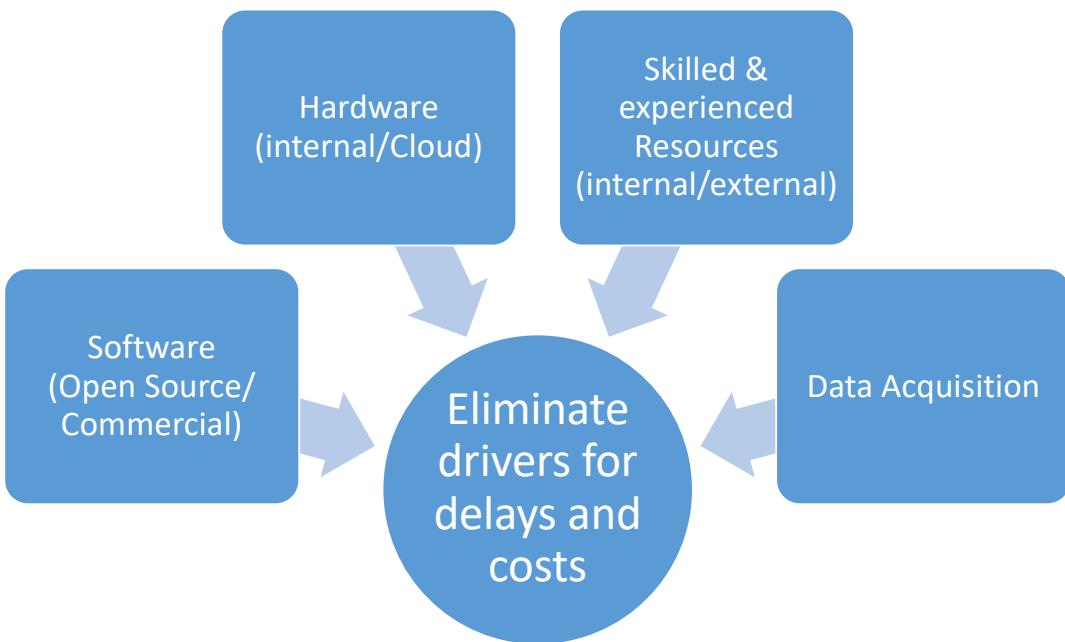
30

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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



31

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy GmbH



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, H2O, Keras/TensorFlow etc.

KNIME – A Leader in Data Science

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



Source: Gartner (January 2019)

35

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

© AdvancedAnalytics.Academy



The KNIME® Analytics Platform

The screenshot shows the KNIME Analytics Platform interface. On the left, there's a "Node Repository" panel with various nodes categorized under "Analytics", "Database", "Other Data Types", "Scripting", "Tool Integration", "Community Nodes", "KNIME Labs", "Workflow Control", "Social Media", "Reporting", "Chemistry", and "ChemAxon / Infocom". The main workspace displays a workflow for a Random Forest Predictor. The workflow starts with a "File Reader" node (labeled "read data") connected to a "Color Manager" node (labeled "assign colors"). Both lead to a "Partitioning" node (labeled "split data 60:40"). The output of the Partitioning node goes to a "Random Forest Learner" node (labeled "Node 5"), which then connects to a "Random Forest Predictor" node (labeled "Node 6"). The predictor node connects to a "Scorer" node (labeled "Node 13"). The "Scorer" node connects to a "JavaScript Bar Chart" node (labeled "Node 3") and a "JavaScript Scatter Plot" node (labeled "Node 7"). The "JavaScript Bar Chart" node and the "JavaScript Scatter Plot" node both connect to a "Grouped Bar Chart" visualization window titled "Attribute Overview". This visualization shows bar charts for "sepal length", "sepal width", "petal length", and "petal width" across three categories: "Iris-setosa", "Iris-versicolor", and "Iris-virginica". The "Grouped Bar Chart" window has "Reset", "Apply", and "Close" buttons at the bottom. The top right of the workspace shows a "Scatter Plot" visualization titled "sepel width" vs "sepal length", with data points colored by category. The bottom right of the workspace shows a "Grouped Bar Chart" visualization titled "Attribute Overview". The bottom of the interface shows a status bar with "256M of 1156M" and a trash bin icon.

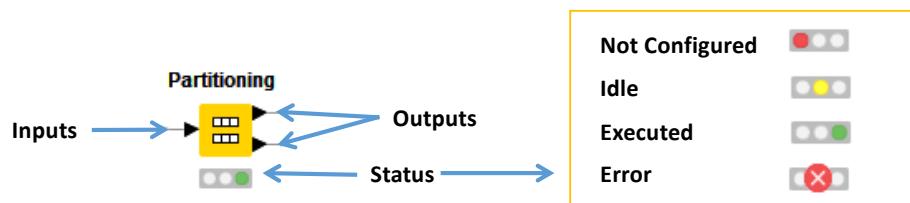
36

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

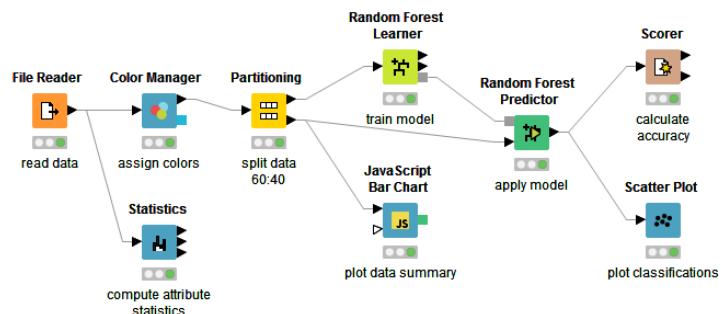


Visual KNIME Workflows

NODES perform tasks on data



Nodes are combined to create
WORKFLOWS

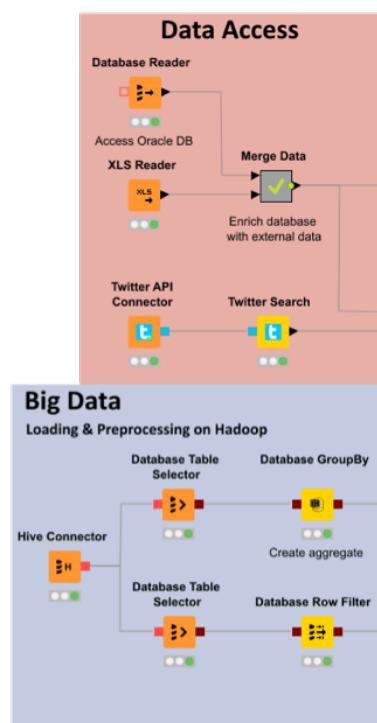


37

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Data Access



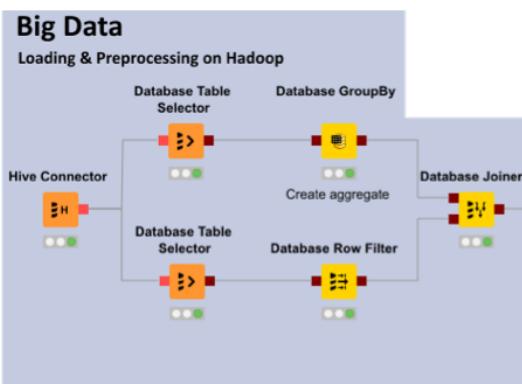
- **Databases**
 - MySQL, PostgreSQL
 - any JDBC (Oracle, DB2, MS SQL Server)
- **Files**
 - CSV, TXT
 - Excel, Word, PDF
 - SAS, SPSS
 - XML
 - PMML
 - Images, texts, networks, chem
- **Web, Cloud**
 - REST, Web services
 - Twitter, Google

38

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Big Data



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

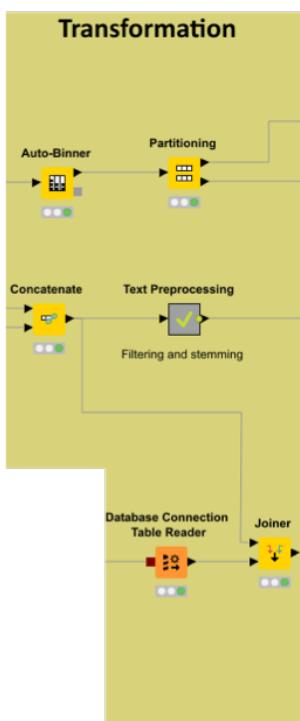


39

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

 AdvancedAnalytics
.Academy

Transformation



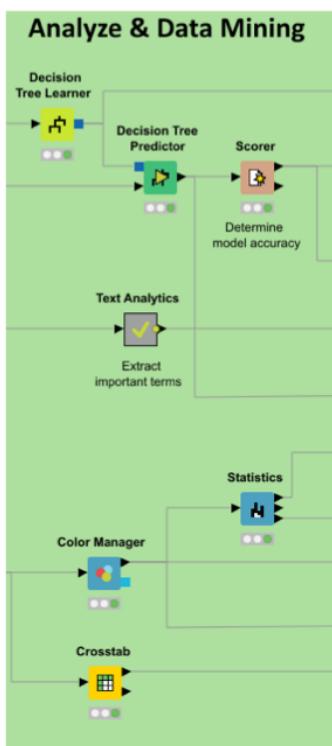
- Preprocessing
 - Row, column, matrix based
- Data blending
 - Join, concatenate, append
- Aggregation
 - Grouping, pivoting, binning
- Feature Creation and Selection

40

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

 AdvancedAnalytics
.Academy

Analyze & Data Mining



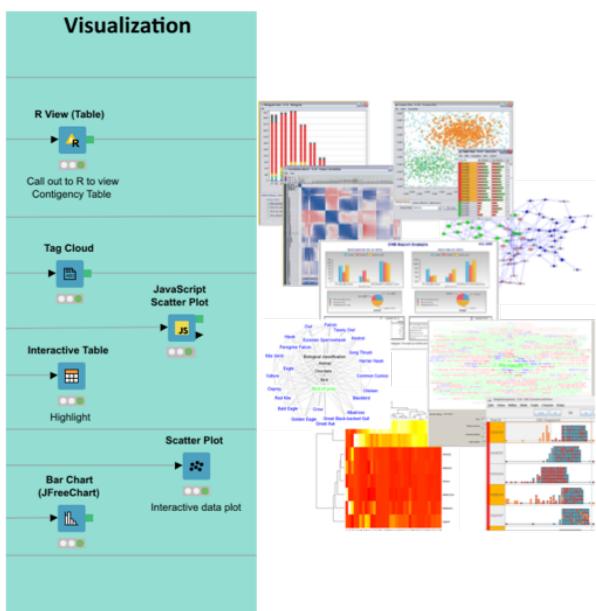
- Regression
 - Linear, logistic
- Classification
 - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
 - k-means, DBSCAN, hierarchical
- Validation
 - Cross-validation, scoring, ROC
- Misc
 - PCA, MDS, item set mining
- External
 - Python, R, Weka

41

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Visualization



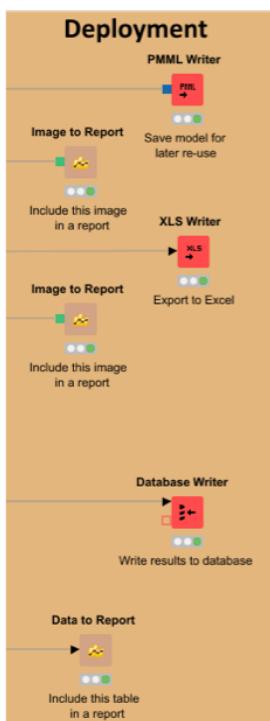
- Interactive
 - Scatter plot, histogram, pie charts, box plot
 - Highlighting (brushing)
- JFreeChart
- JavaScript
- Misc
 - Tag cloud, open street map, networks, molecules
- External
 - R

42

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Deployment



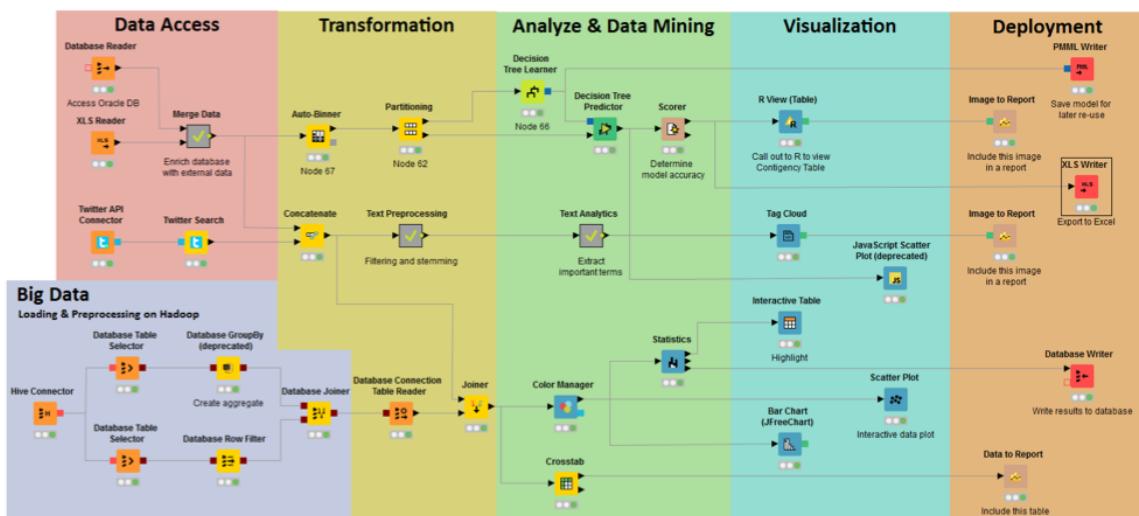
- Database
- Files
 - Excel, csv, txt
 - XML
 - PMML
 - to: local, KNIME Server, SSH-, FTP-Server
- BIRT Reporting

43

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Over 1500 native and embedded nodes included:



Data Access

MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers
Industry Specific
Community / 3rd

Transformation

Row, Column, Matrix, Text, Image, Time Series, Java, Python, Community / 3rd

Analyze & Mining

Statistics, Data Mining, Machine Learning, Web Analytics, Text Mining, Network Analysis, Social Media Analysis, R, Weka, Python, Community / 3rd

Visualization

R, JFreeChart, JavaScript, Community / 3rd

Deployment

via BIRT, PMML, XML, JSON, Databases, Excel, Flat, etc., Text, Doc, Image, Industry Specific, Community / 3rd

44

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



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

45

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



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) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	32 bit (301,82 MB) 64 bit (305 MB)
KNIME Analytics Platform + all free extensions for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	32 bit (1,74 GB) 64 bit (1,92 GB)
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding the KNIME installation</i>	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)

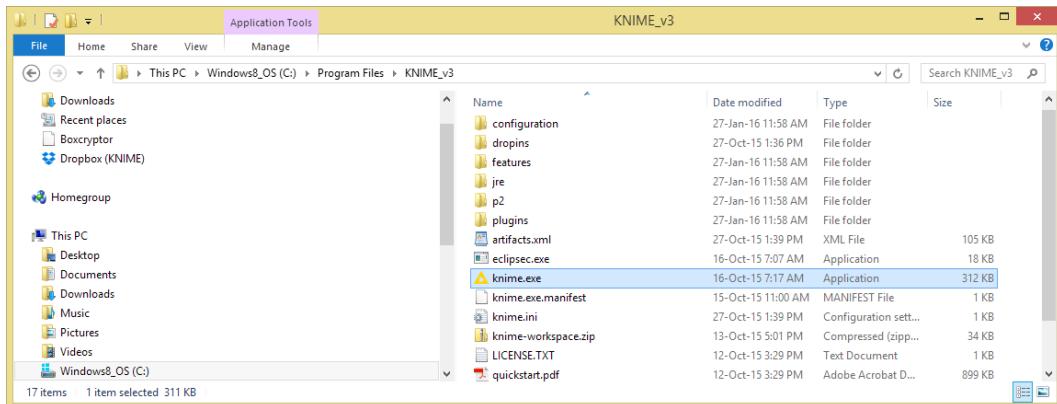
46

This educational material was produced for the Machine Learning Workcamp at
IHK Region Stuttgart February 2019. The copyright is with
AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Start KNIME Analytics Platform

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



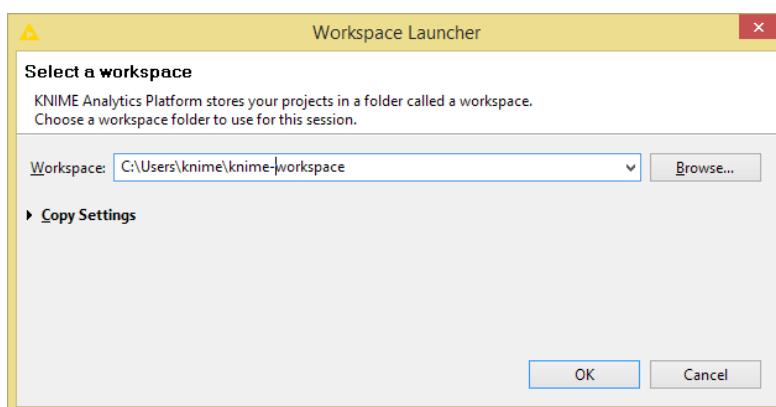
47

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



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)



48

This educational material was produced for the Machine Learning Workcamp at IHK Region Stuttgart February 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.



Welcome Page

The screenshot shows the KNIME Analytics Platform's welcome page. At the top, the KNIME logo is displayed with the tagline "Open for Innovation". Below the logo, the title "Welcome to KNIME Analytics Platform!" is centered. The page contains several sections:

- New to KNIME?** A section for new users with tips like "Register for emails with introductory tips [here](#)." and "Explore our Quickstart Guide."
- Updates for the following components are available:** Lists "DYMATRIZ Uplift Modeling Extensions" and "Palladian for KNIME".
- Workflow Coach:** A sidebar titled "Workflow Coach" showing recommended nodes: Community (8%), Column Filter (8%), Partitioning (8%), Joiner (8%), and k-Means (6%). A yellow arrow points from the "Partitioning" node in the coach to its corresponding node in a workflow diagram below.
- Where to go from here:** A list of links including "Create new workflow", "Learning Hub", "Browse example workflows", "Get additional nodes", "Go to my workflows", and "Mount KNIME Cloud Server".
- Most recently used workflows:** A list of recent workflows: ModelSelection_WebPortal_Part1, ModelSelection_WebPortal_Part1, ModelSelection_BasicWorkflow, DataCleaning_WebPortal_v2.0, KNIME_project2, and Sexy ETL_v2.0.
- Tips & Tricks:** A section with a checkbox for "Show intro text at next start" (checked).
- Specialist Nodes:** A section explaining that there are many specialist nodes available from KNIME Labs and the Community.
- Footer:** Includes a copyright notice for the Machine Learning Workcamp at IHK Region Stuttgart February 2019, and the AdvancedAnalytics.Academy logo.

49

KNIME File Extensions

- Dedicated file extensions for Workflows and Workflow groups associated with KNIME Analytics Platform

- *.knwf for KNIME Workflow Files



Workflow_1.knwf

- *.knar for KNIME Archive Files



Group_wf_1.knar

50



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