

Open for Innovation

KNIME

Welcome to Sharing and Deploying Data Science with KNIME Server

Going live at:

Berlin 6:00 PM (CET)

New York City 12:00 PM (EDT)

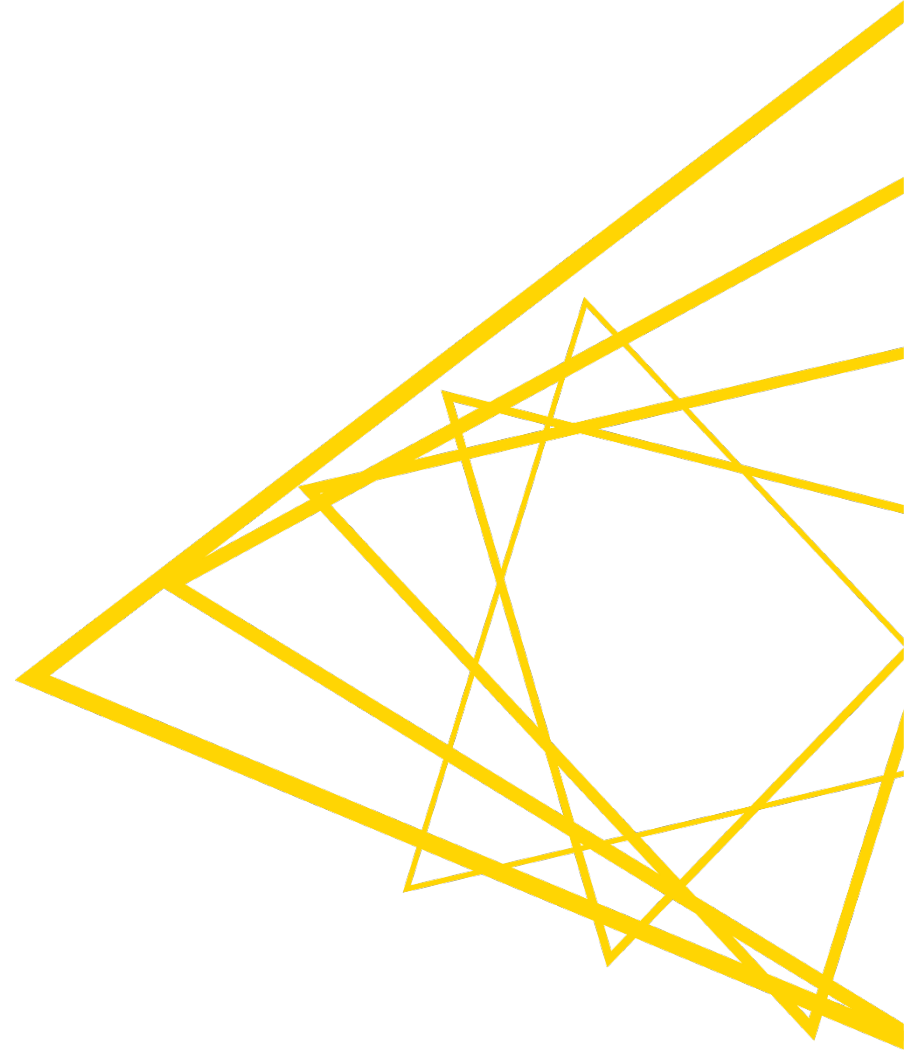
Austin 11:00 AM (CDT)

London 5:00 PM (GMT)



Before we start...

- Please use the Q&A section to post your questions.
- Upvote for your favorite questions.



Challenges Operationalizing Data Science

Creating Great Data Science

Large,
complex
variety of
technologies

Providing
skills and
resources

Continuous,
high quality
delivery

Mainstreaming
across the
organization

Data Science Driven Applications

R&D

MFG

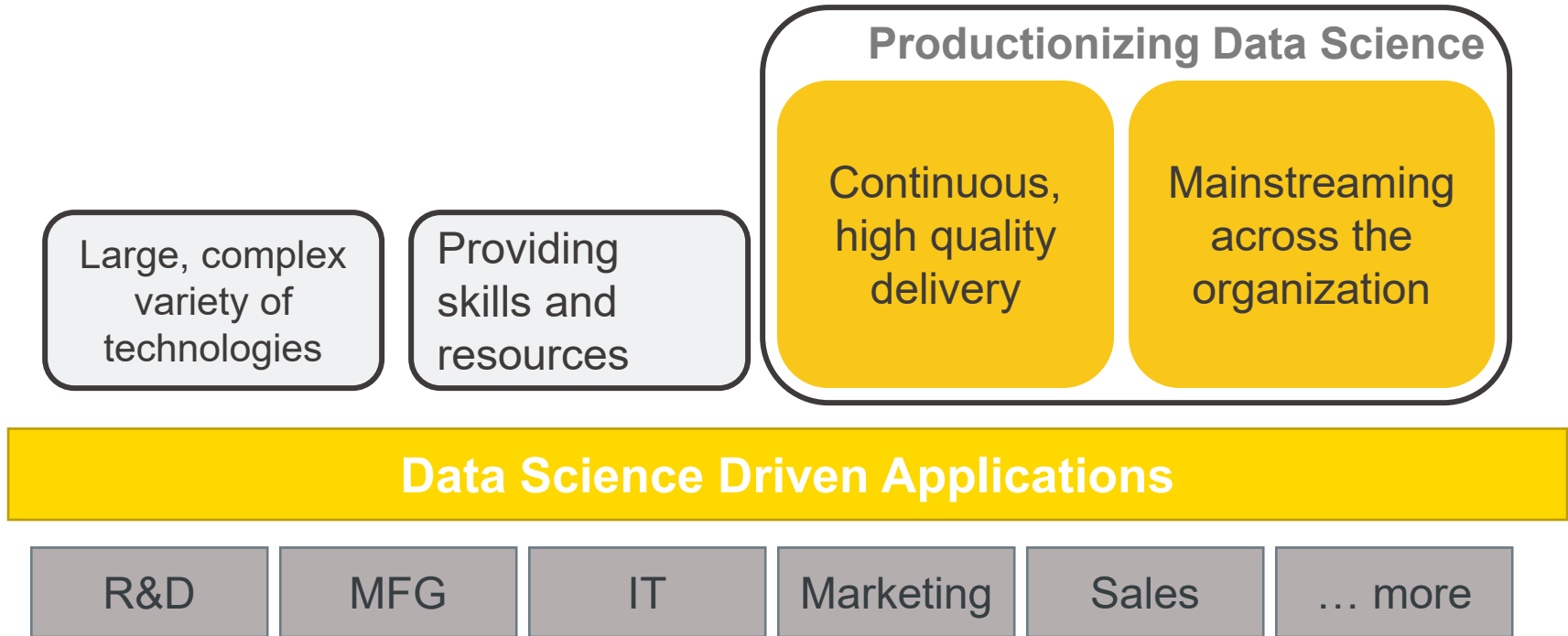
IT

Marketing

Sales

... more

Challenges Operationalizing Data Science



KNIME Software: for Individuals & the Business

Create

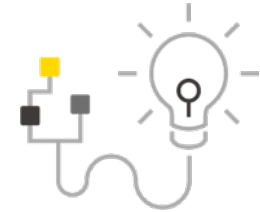
Productionize

Gather &
Wrangle

Model &
Visualize

Deploy &
Manage

Consume &
Optimize



**Open Source
KNIME Analytics Platform**

KNIME
Extensions

KNIME
Integrations

Community
Extensions

Partner
Extensions

**Commercial
KNIME Server**

KNIME
WebPortal

Data
Science as
a Service

KNIME Server

- Collaboration
 - Share Expertise and Templates/Blueprints
- Automation
 - Schedule, Monitor, Update
- Deployment
 - To Humans: Analytical Applications (“Guided Analytics”)
 - To Machines: via REST and custom APIs
- Management
 - Manage distributed setups
 - Integrate with existing enterprise setup (authentication)

Collaboration

The diagram illustrates the components of a workflow system, organized into five main sections:

- Workflow Hub:** A screenshot of a web-based interface for managing workflows, showing a list of workflows and their details.
- Workflows:** A diagram showing a sequence of nodes connected by arrows, representing a workflow process.
- Metanodes, Nodes:** A diagram showing a node labeled "Partitioning" and a node labeled "My Predictor", representing the building blocks of a workflow.
- Data:** A diagram showing various data formats (XLS, GDF, TEXT, CSV, XML, MATLAB) and a "Data" label, representing the input and output of the workflow.

At the bottom, a central label reads "Shared Repository, Versioning", indicating the underlying infrastructure for the workflow system.

Automation

The diagram illustrates the workflow of a distributed system, divided into three main sections:

- Provisioning:** Shows a network diagram with various nodes and a clock icon, indicating 'Scheduled Execution'.
- Remote, Client Control:** Shows a 'Remote Execution' window with a list of tasks and a 'View Info' button.
- Distributed Executors:** Shows three server icons with arrows indicating communication between them.

Deployment

The diagram illustrates three distinct access methods for a data science application:

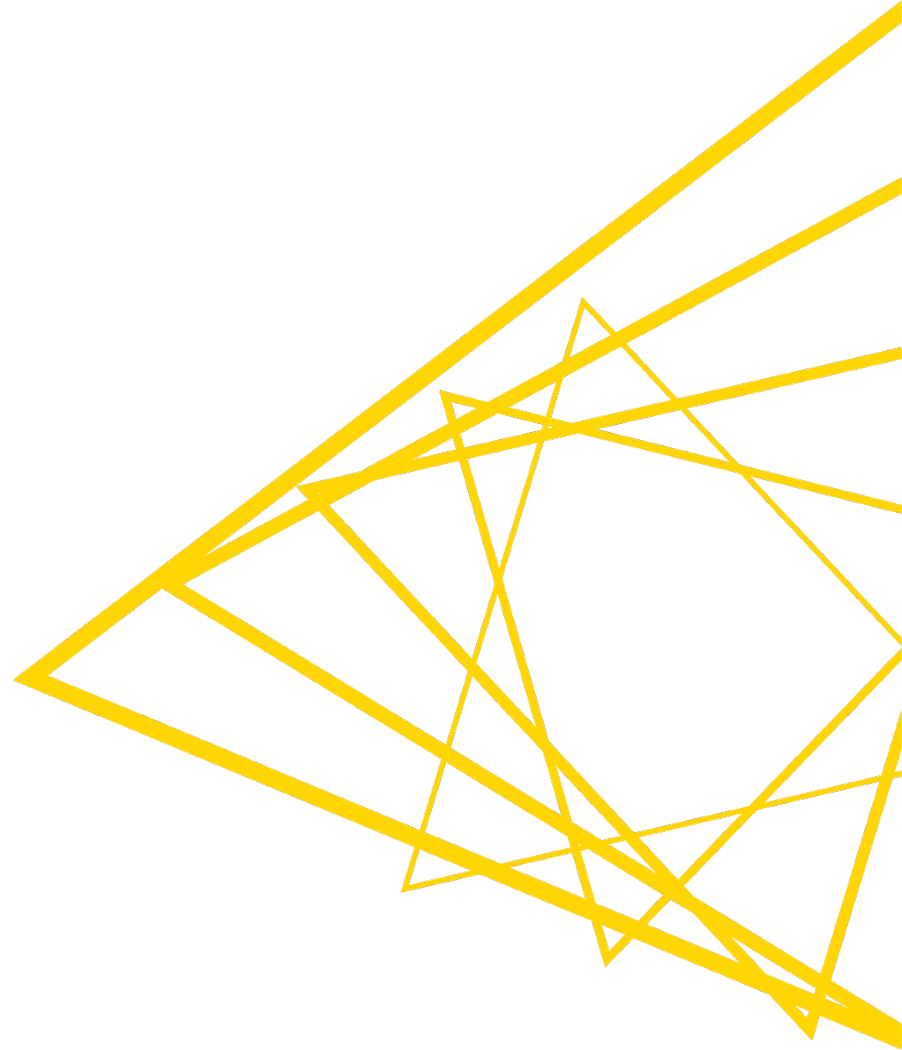
- Web Front-ends:** A screenshot of a web browser displaying a data visualization interface with a tree chart and a bar chart. Below the screenshot is the text "Web Browser Access".
- Guided Analytic Apps:** A screenshot of a guided analytics application showing a pie chart, a bar chart, and a scatter plot. Below the screenshot is the text "Custom Application Access".
- RESTful Services:** A flow diagram of a data pipeline. It starts with "JSON Input" (Node 1), followed by "JSON to Table" (Node 5), "Transpose" (Node 6), "Columns Rename" (Node 10), "RScript Description Calculation" (Node 17), "Table to JSON" (Node 18), and finally "JSON Output" (Node 2). Below the flow diagram is the text "Custom Application Access".

Management

The diagram illustrates the three main components of the KNIME workflow, each represented by a yellow box with a blue header and a white body containing icons and text.

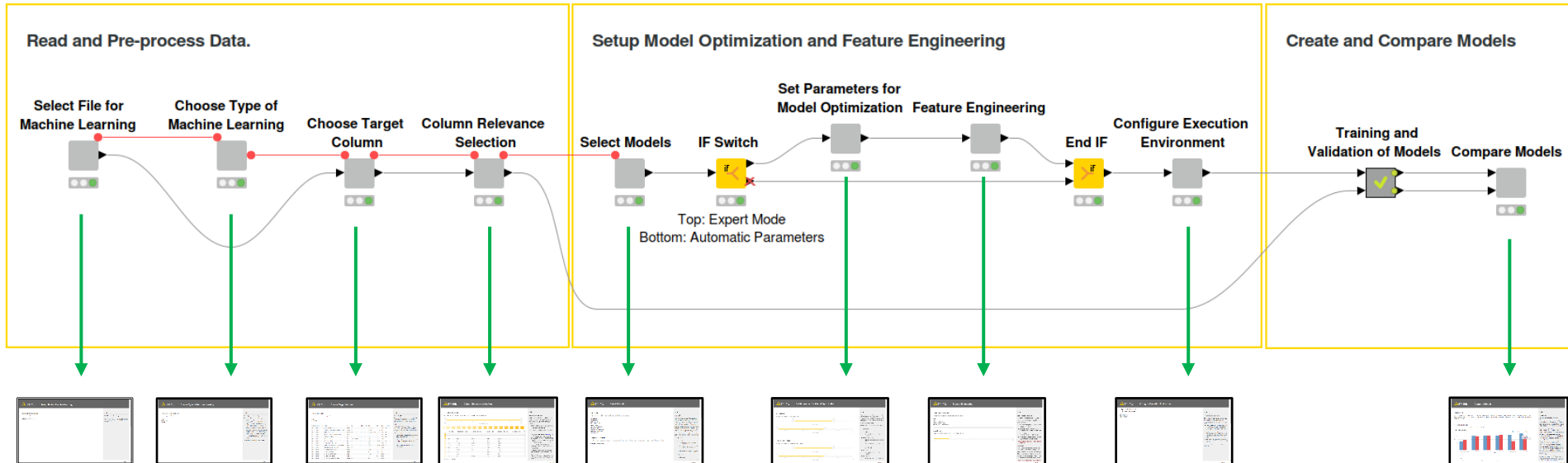
- Security:** The first box on the left, titled "Security", contains icons for a padlock, LDAP, and Java, with the text "Nodes, Files, Applications" below them.
- Configuration & Versioning:** The middle box, titled "Configuration & Versioning", contains a screenshot of the KNIME Desktop Viewer interface showing a workflow tree and a context menu.
- Client & User Preferences:** The third box on the right, titled "Client & User Preferences", contains an icon of a computer monitor displaying the KNIME logo.

Live Demo



Guided Analytics for ML/AI Automation

- Interaction & Automation
- Data Scientist's Choice: right mix for target audience



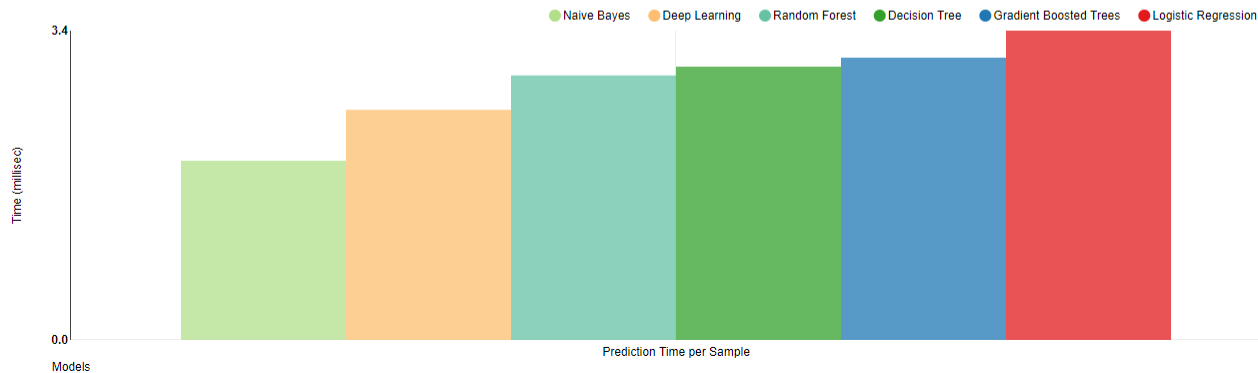
Compare Models Time

Compare in those bar charts the time each model took to train and to compute a sample prediction.

Training Time



Prediction Time per Sample



Details for Experts

You can compare the models now by different metrics. Above you can see a bar plot with 3 main measures of performance: *Accuracy* is the percentage of correct predictions among all predictions. *Area Under Curve (AUC)* measures of the area under the [Receiver Operating Characteristic](#) curve.

The ROC plot describes the Receiver operating characteristic curves, one for each model. On the y-axis you have the true positive rate, on the x-axis you have the false positive rate based on the class you previously provided. The greater the area under the model curve (AUC) the better is the performance of the model. More info about this plot can be found at this [link](#).

Advanced Assessment of Models

The advanced assessment of models sections shows four additional charts per model.

- 1. Performance Metrics Bar Charts

For this visualization we measured the following metrics:

- [Recall](#) (or True Positive Rate) (% of "casualty" rows correctly classified)
- [Precision](#) (or Positive Predicted Value) (% of predicted "casualty" rows correctly classified)
- [Specificity](#) (or True Negative Rate) (% of not "casualty" rows correctly classified)
- [F-measure](#) (harmonic average between Recall and Precision)

- 2. Cumulative Gain Chart and Lift Chart

This chart is drawing a curve that reflects how well the model is doing compared to a random classifier. You are selecting rows from the test ranked by the probability of class "casualty". On the x-axis you have the percentage of top ranked rows by the model that define the partition of rows you are considering. On the y-axis you measure the response as the percentage of "casualty" rows over their total number in your partition of top ranked rows. If the model is bad, the curve will be close to the black line, where the percentage of original "casualty" rows is exactly equal the percentage of selected rows, (e.g. where the model ranking is random). The cumulative gain curve should be above the bisector line and the greater the area between the cumulative gain curve and the bisector line is, the better the model is.

Reset Apply Close

Advanced Assessment of Models

Each row represents a series of additional information about each created model.

Decision Tree

Figure 1: Performance Metrics

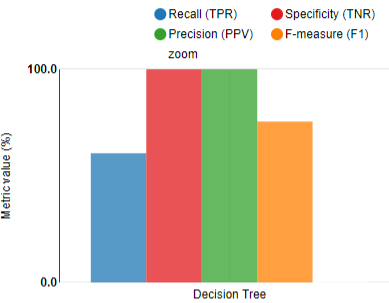


Figure 2: Cumulative Gain Chart

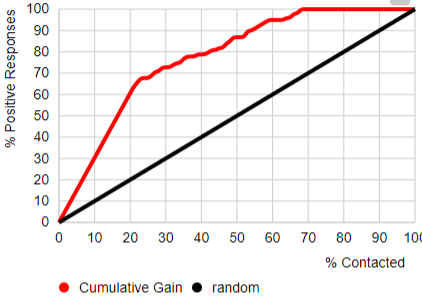


Figure 3: Global Feature Importance

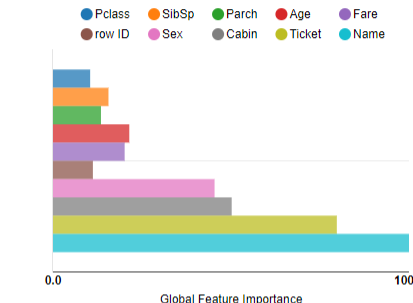
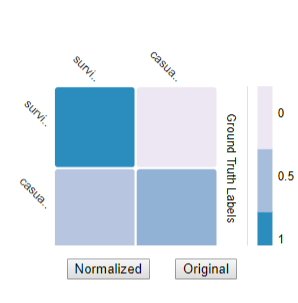


Figure 4: Confusion Matrix Heatmap



Gradient Boosted Trees

Figure 1: Performance Metrics

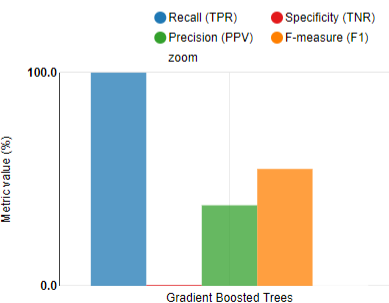


Figure 2: Cumulative Gain Chart

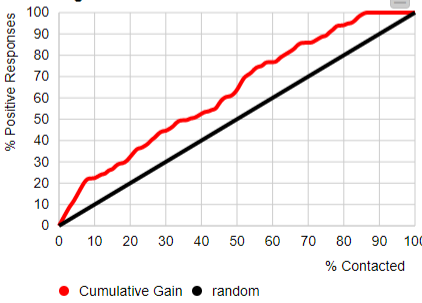


Figure 3: Global Feature Importance

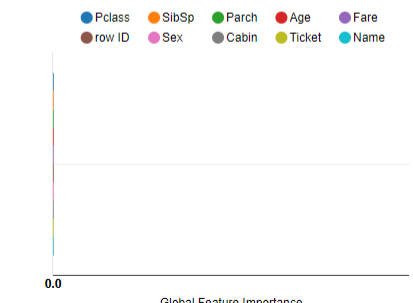
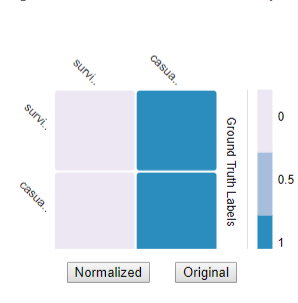


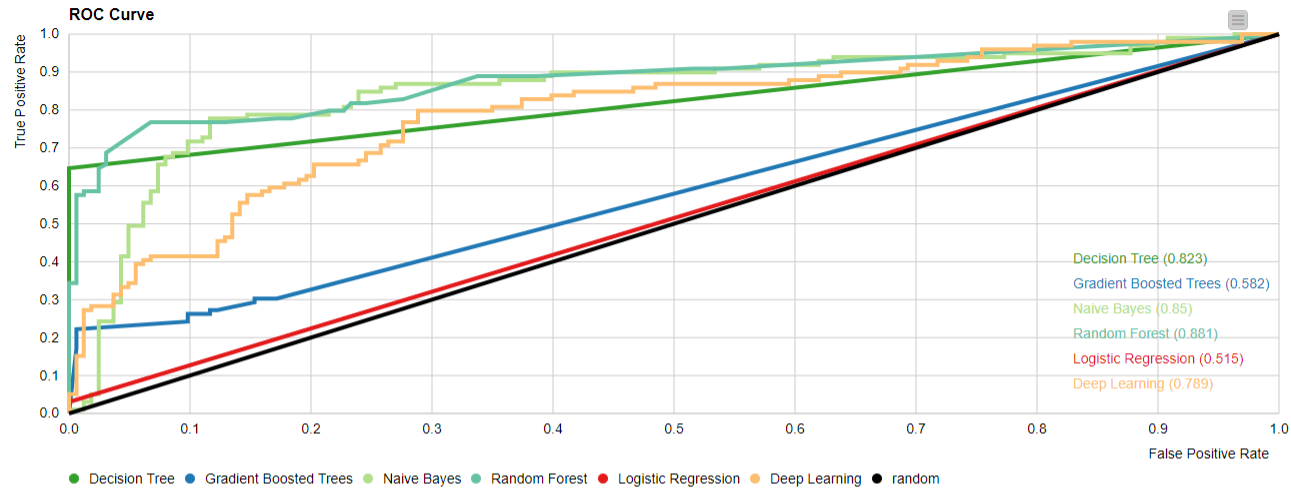
Figure 4: Confusion Matrix Heatmap



Naïve Bayes

ROC Curve Plots

Compare the curves, one for each model. The greater the area under a curve the better the model is.



Download Model

The following table summarizes the information in the charts. Please select from the table the model you would like to download and use.

Model	Accuracy (%)	Training Time (mins)	Prediction Time (millisec)	Area Under Curve (%)	Workflow	Parameters
Decision Tree	94.276	0.4	3	82.323	Download	Download
Logistic Regression	86.819	0.4	3.4	51.515	Download	Download
Random Forest	84.812	0.8	2.9	88.062	Download	Download
Naive Bayes	82.713	0.4	2	85.047	Download	Download
Deep Learning	81.28	4	2.6	78.881	Download	Download
Gradient Boosted Trees	50.036	0.9	3.1	58.226	Download	Download

cumulative gain chart curve and the bisector line.

3. Global Feature Importance Bar Chart

We consider for each case only the original selected features for the relative model. Then we train a surrogate random forest model overfitting the test set predicted classes. From such model it is possible to measure how often each feature is useful to outcome a prediction. More info at this [link](#).

4. Normalized Confusion Matrix Heatmap

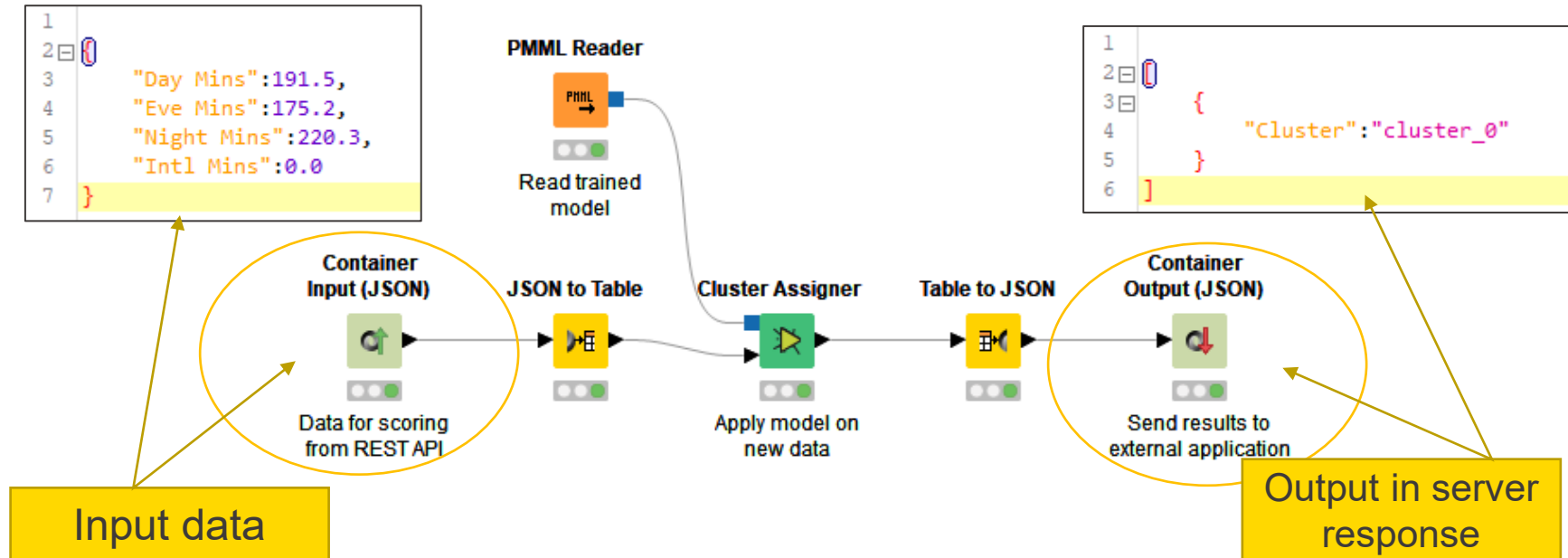
A confusion matrix is summarizing all the outcomes on the test set by considering how many instance fall in each case given by prediction and original class. The heatmap is encoding with shades of blue which cell has the most instances in each row. A model with a good performance should have most of the dark blue cells on the diagonal from top-left corner to bottom right corner of the squared image. More info at this [link](#).

Reset Apply Close

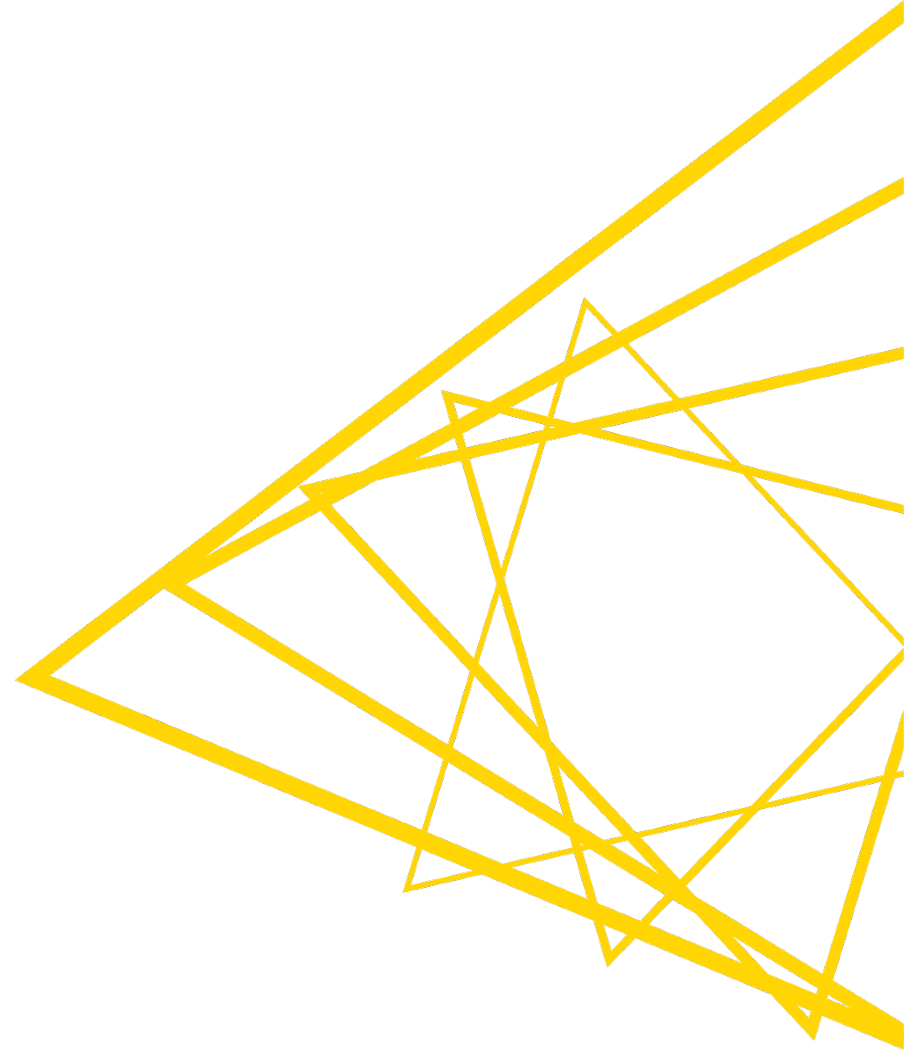
KNIME Server REST API

- Enables external integration
- Build applications around KNIME Server
- e.g. for Microservices and real time scoring

Deploy KNIME workflows as web services

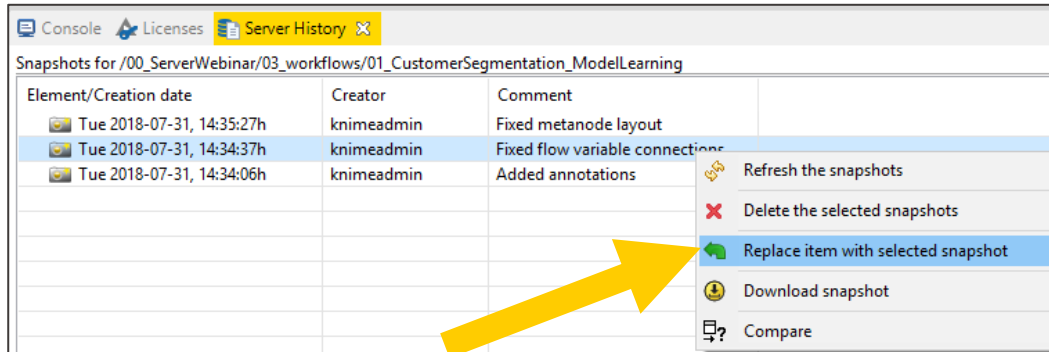


Live Demo



Version Control

- Create snapshots of workflow revisions
- Allow roll-back to previous versions
- Full workflow history available in client and on Workflow Hub

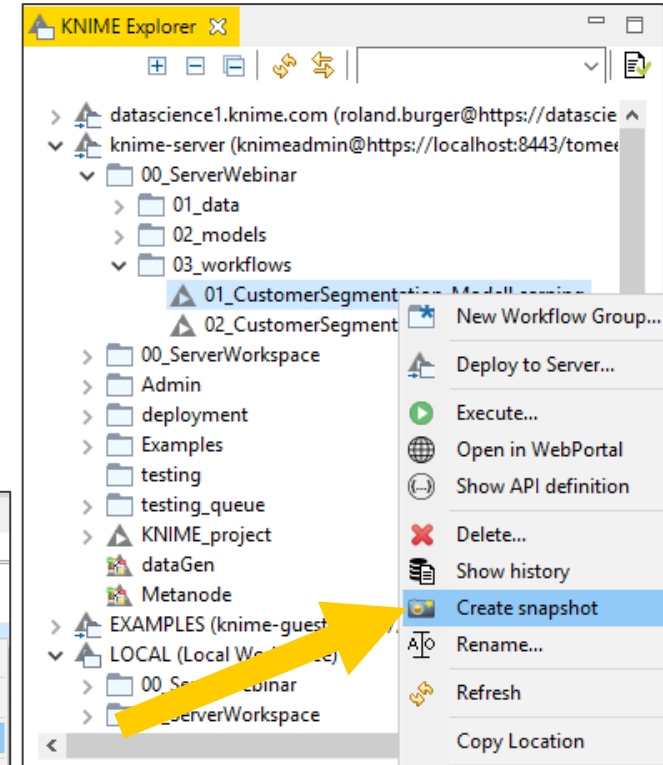


The screenshot shows the 'Server History' window in KNIME. It displays a table of snapshots for the workflow '00_ServerWebinar/03_workflows/01_CustomerSegmentation_ModelLearning'. A context menu is open over the table, with a yellow arrow pointing to the 'Replace item with selected snapshot' option.

Element/Creation date	Creator	Comment
Tue 2018-07-31, 14:35:27h	knimeadmin	Fixed metanode layout
Tue 2018-07-31, 14:34:37h	knimeadmin	Fixed flow variable connections
Tue 2018-07-31, 14:34:06h	knimeadmin	Added annotations

Context menu options:

- Refresh the snapshots
- Delete the selected snapshots
- Replace item with selected snapshot
- Download snapshot
- Compare



Workflow Difference

- Inspect differences between versions of the same workflow

Workflow Comparison

knime-server/00_ServerWebinar/...	LOCAL:/00_ServerWebinar/03_wo...
Color Manager (101)	Color Manager (101)
Select Data File (158)	Select Data File (158)
Display Cluster Result (130)	Display Cluster Result (130)
Define Cluster Parameters (12)	Define Cluster Parameters (126)
k-Means (138)	k-Means (138)
PMML Writer (156)	PMML Writer (156)
File Reader (91)	File Reader (91)
	Bar Chart (JavaScript) (159)

Highlight differences:

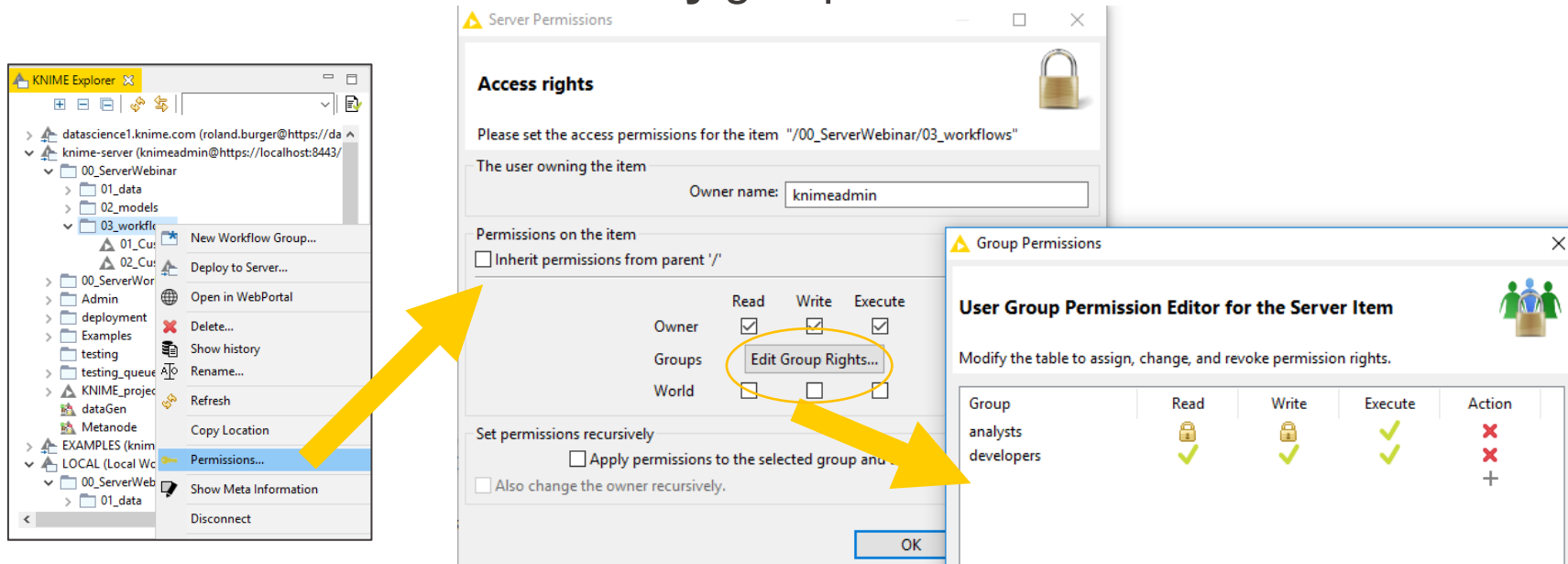
- Nodes included/excluded
- Node configurations

Node Settings Comparison

k-Means (138)			k-Means (138)		
Name	Type	Value	Name	Type	Value
Node Settings					
nrClusters_Internals	sub-config		nrClusters_Internals	sub-config	
nrClusters	int	3	nrClusters	int	3
maxNrIterations_Internals	sub-config		maxNrIterations_Internals	sub-config	
maxNrIterations	int	99	maxNrIterations	int	999
cfgColsms_Internals	sub-config		cfgColsms_Internals	sub-config	
cfgColsms	sub-config		cfgColsms	sub-config	
enableHilite_Internals	sub-config		enableHilite_Internals	sub-config	
enableHilite	boolean	false	enableHilite	boolean	false
System Node Settings					

Permission Management

- Manage permissions for workflows and data on KNIME Server
- Control which groups can access, modify, or execute a workflow
- Use **LDAP / Active Directory** groups for centralized administration



Managing Preferences – local

How do I
configure
Python?



KNIME
Analytics
Platform



KNIME
Analytics
Platform

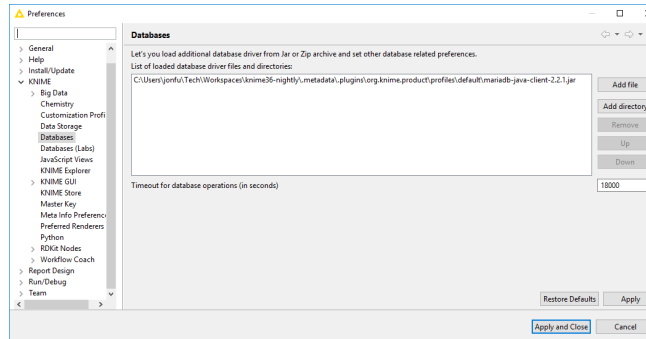
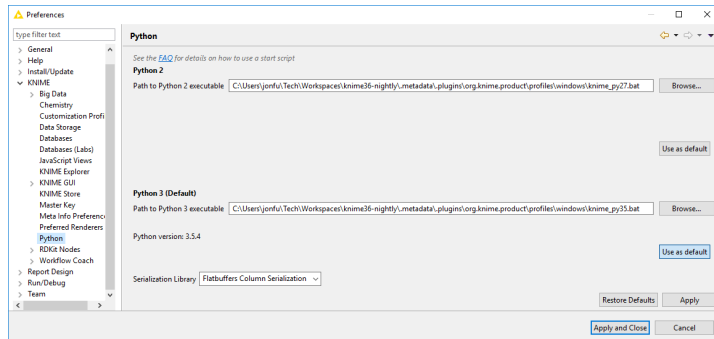
How do I
install
database
drivers?



KNIME
Analytics
Platform

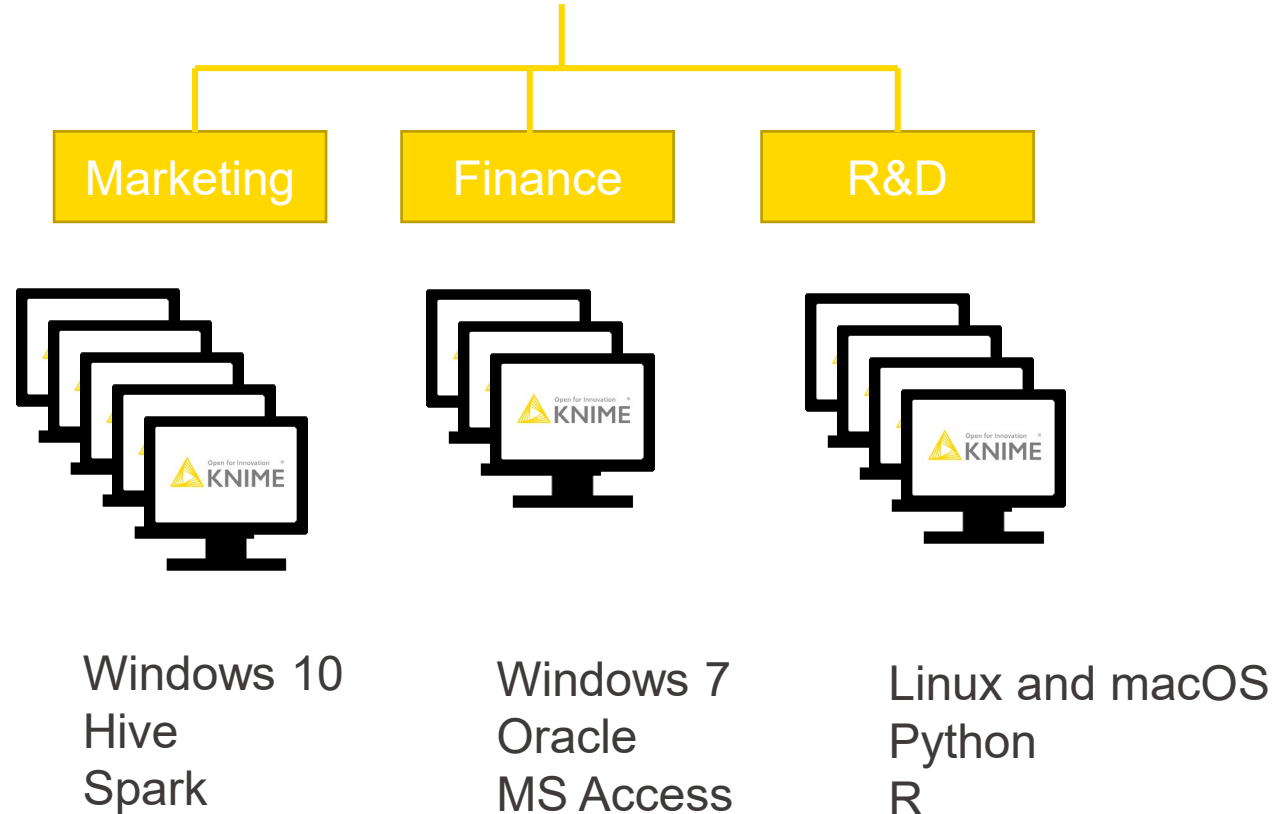


KNIME Website
Internal knowledgebase

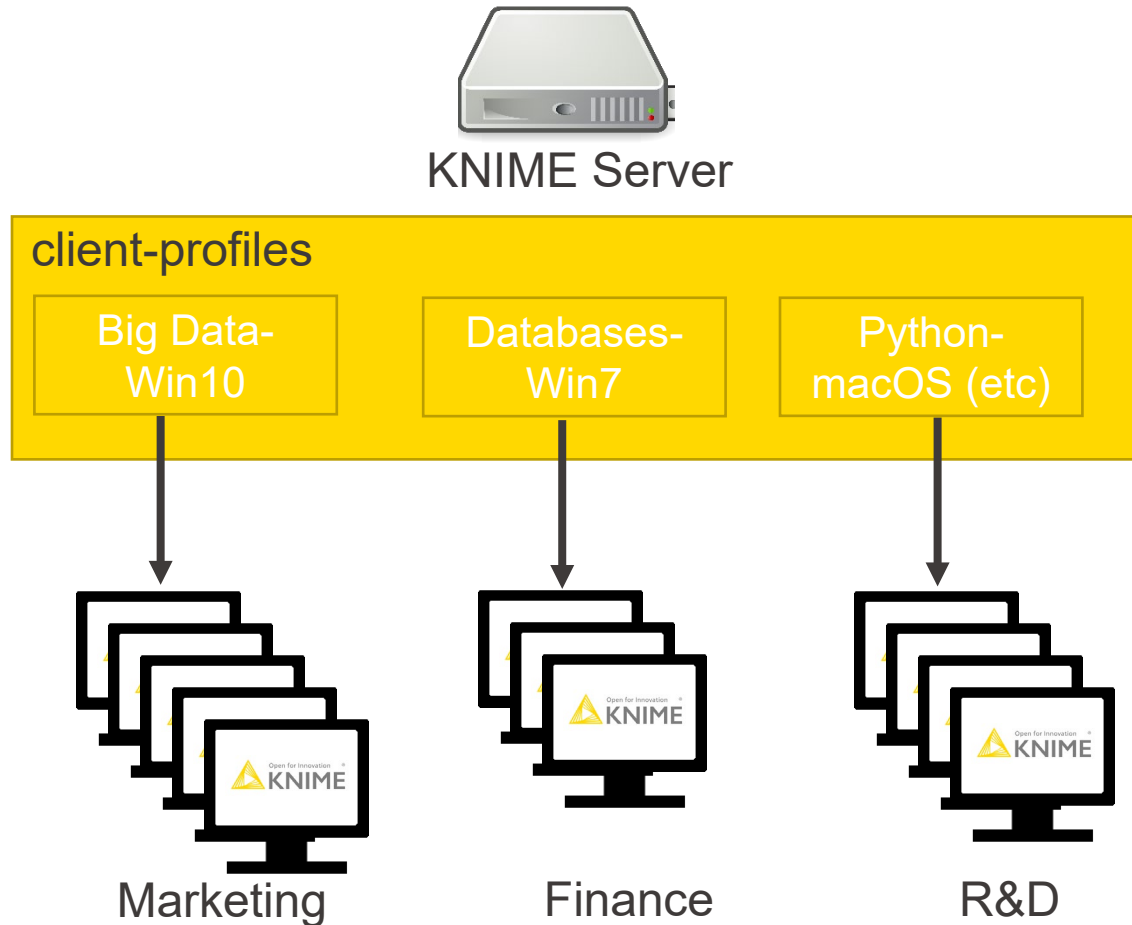


Managing Preferences – KNIME Server

- Different departments/teams have different requirements
- Multiple OS deployments
 - Windows 7
 - Windows 10
 - Linux
 - macOS



Managing Preferences – KNIME Server



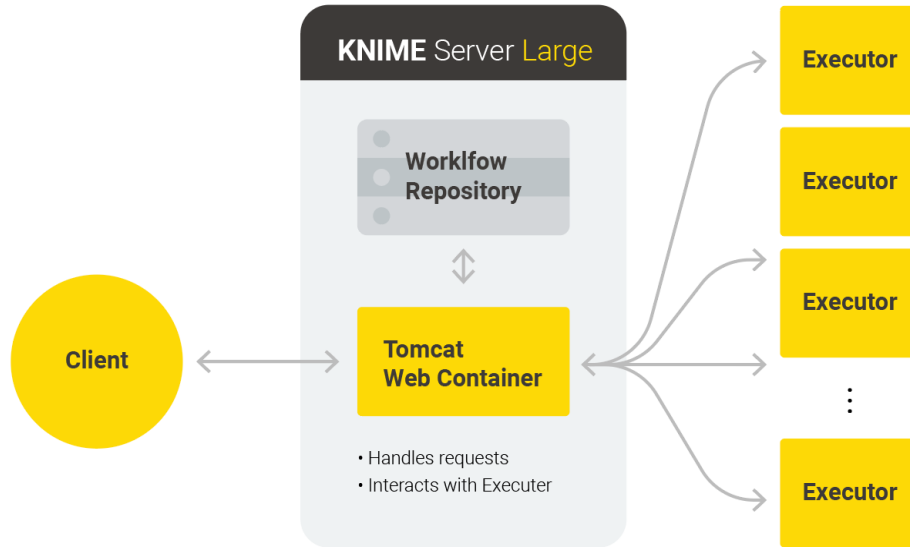
- Client-profiles
 - Python-Linux
 - Python-macOS
 - R-Linux
 - R-macOS
 - Databases-Win7
 - Big Data-Win10



Profiles can include:
Preferences, drivers, and more

On-Prem or Cloud, Scalable, Flexible

KNIME Server Large with multiple Executors

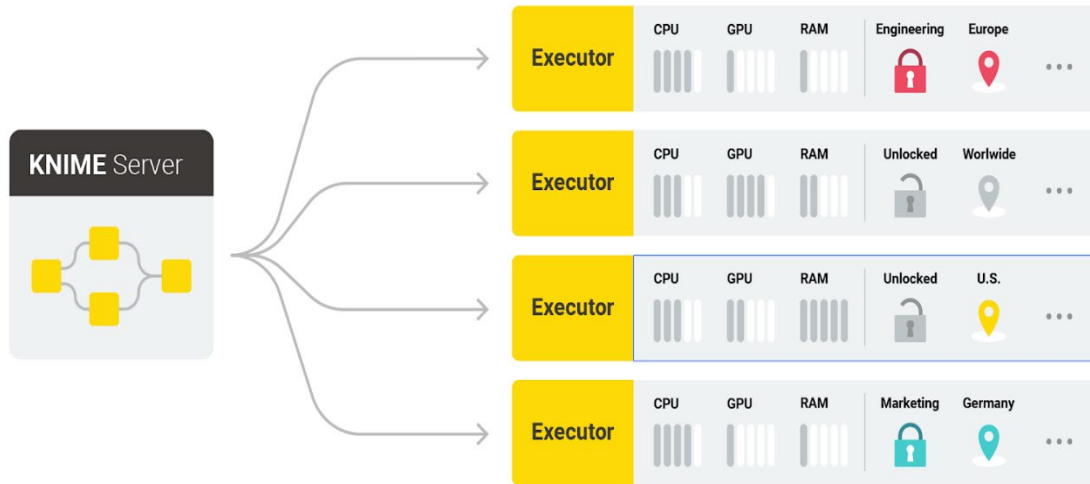


Features:

- Offload compute from KNIME Server
- Scale Executors to compute needs
- Support Executors with varying configurations
- Flexible deployment options

Capacity planning, Routing, Monitoring

Pinning Workflows to KNIME Executors

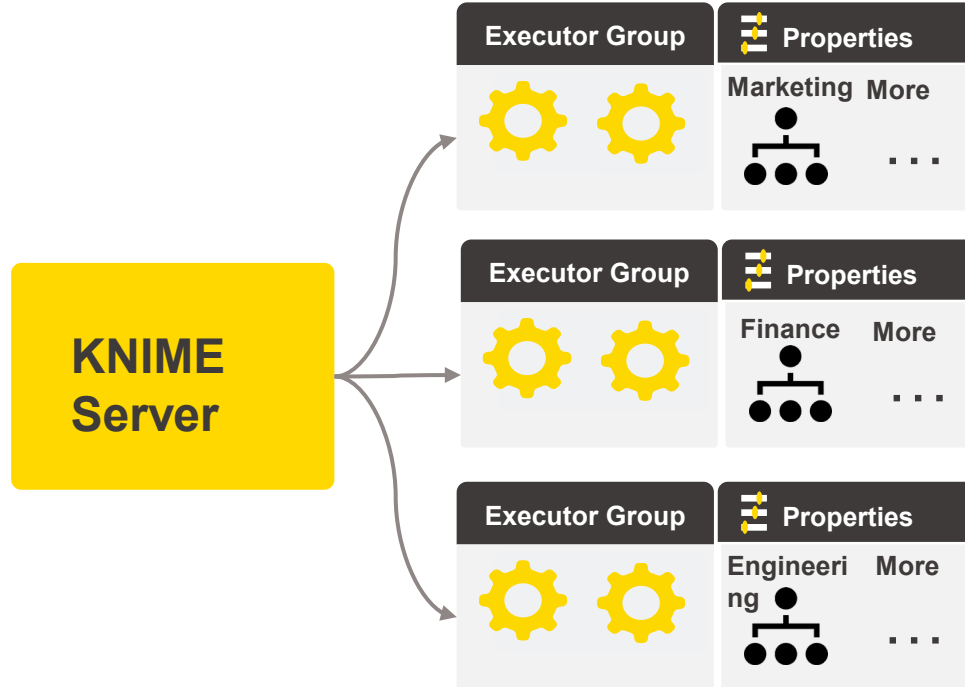


Features:

- Match workflow needs to Executor capabilities
- Partition compute resources by capability, department, usage, ...
- Workflow needs determined by workflow publisher

Expanded routing, Partitioning resources, Monitoring

Supporting Executors in the Enterprise*



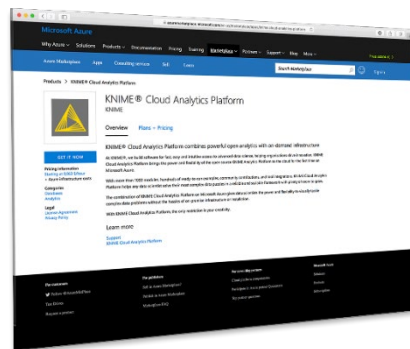
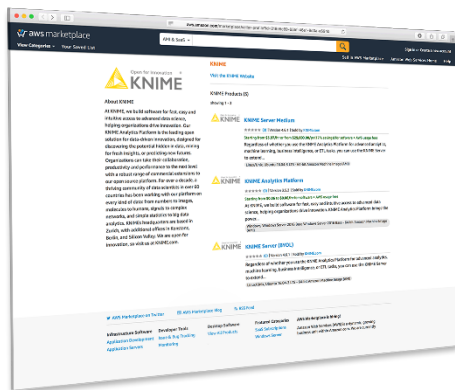
Features:

- Logical groupings of Executors
- Match users/groups to Executor Groups
- Partition compute resources by groups, department, ...
- Partitioning managed by Server administrators

*production July 2020

Elastic Scalability, Cloud

KNIME Cloud Offerings – AWS and Azure

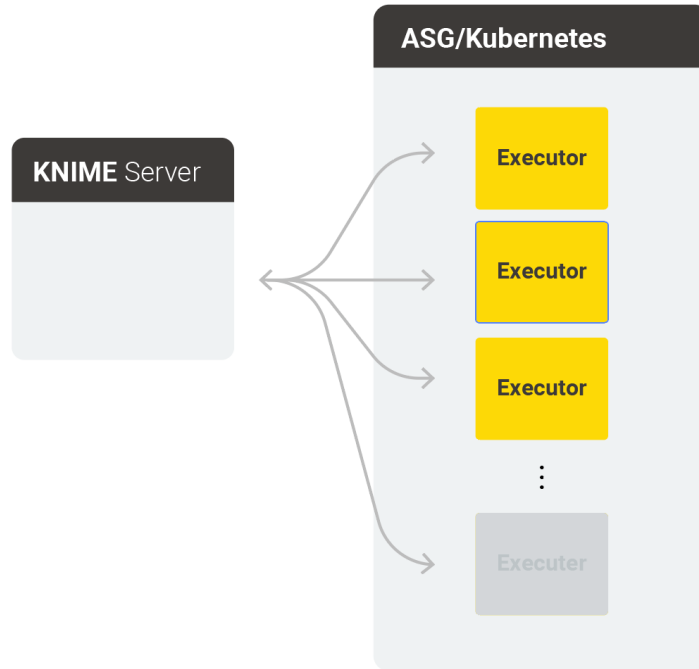


Features:

- KNIME Analytics Platform
- KNIME Server Large BYOL
 - Supports Server Large with multiple Executors
 - Has an embedded Executor so can be stand-alone
- KNIME Executors
 - Multiple Executors that can be used by KNIME Server Large
 - Pay as you go (PAYG) offering supports elastic scaling
 - Bring your own license (BYOL) offering uses cores from your Server license

Elastic Scalability, Cloud

KNIME Executor Auto-scaling

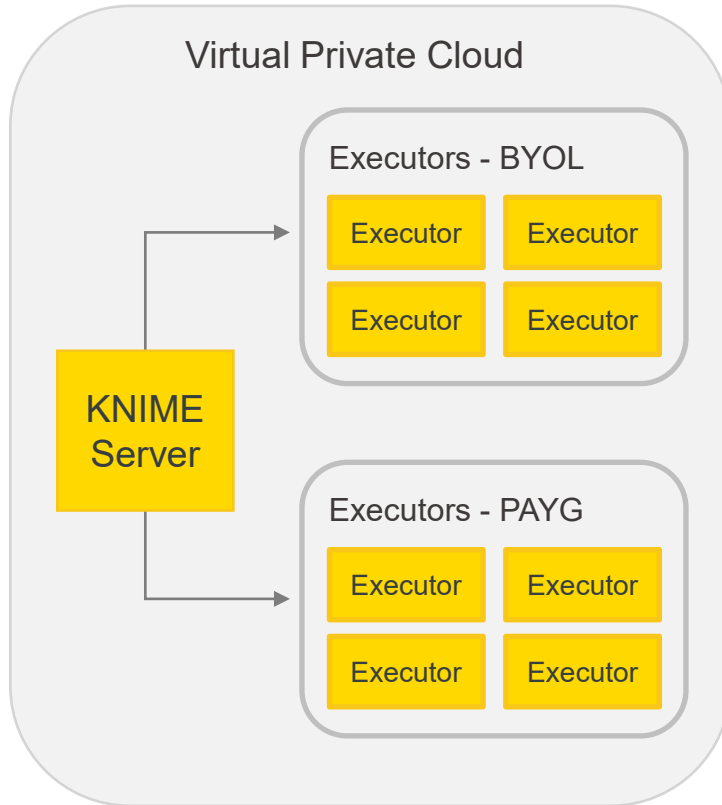


Features:

- Using cloud platforms or Kubernetes* auto-scaling capabilities
- Elastic scaling to meet demand
- Scales out to meet needed capacity
- Scales in to save costs
- Supported on AWS and Azure in PAYG licensing model

Flexible Cloud deployments to meet computing needs

Mixed Cloud Usage

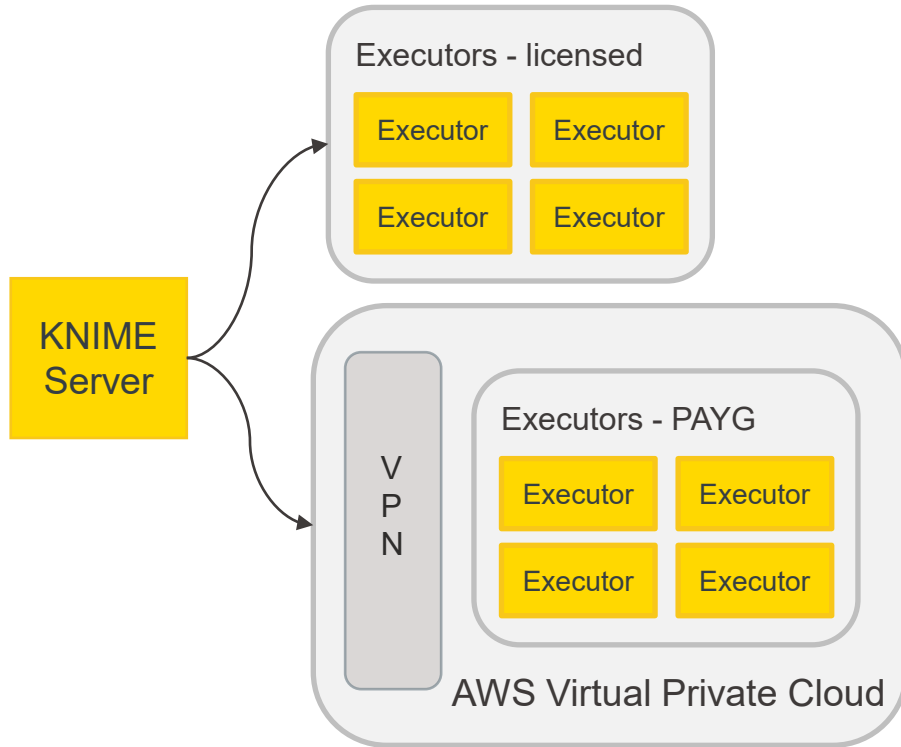


Features:

- Supplement traditionally licensed Executors with Pay-as-you-Go (PAYG) model
- Meet periodic demand peaks
- Fulfill need for speciality hardware (e.g. GPU's)
- Meet budgeting needs

Flexible Hybrid deployments to meet computing needs

KNIME Hybrid Usage Model



Features:

- Mix of Enterprise data center and Cloud deployments
- Meet periodic demand peaks
- Fulfill need for speciality hardware (e.g. GPU's)
- Meet budgeting needs

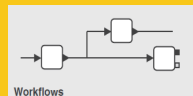
KNIME Server

Collaboration

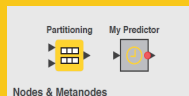
Workflow Hub



Workflows



Metanodes, Nodes



Data



Shared Repository, Versioning

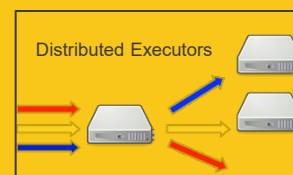
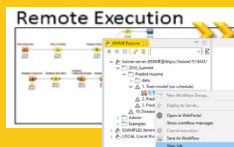
Automation

Provisioning



Scheduled Execution

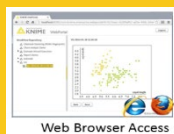
Remote, Client Control



Distributed Executors

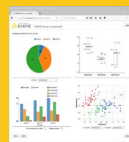
Deployment

Web Front-ends

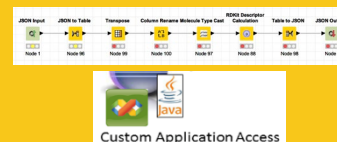


Web Browser Access

Guided Analytic Apps

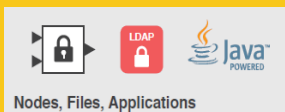


RESTful Services



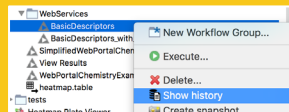
Custom Application Access

Management



Nodes, Files, Applications

Security



Configuration & Versioning



Client & User Preferences

Ready to get started with KNIME Server?
Contact us for a free trial:

contact@knime.com

Thank you!