What’s new

Bernd Wiswedel
What’s new...

• 2+1 feature releases in the last year: (3.0), 3.1, 3.2

• Changes documented online...
What’s new pages and YouTube
Changelog ...
## Changelog...

### New nodes/features by version:

<table>
<thead>
<tr>
<th>Version</th>
<th># new nodes/sets</th>
<th># features</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2.12</td>
<td>27</td>
<td>53</td>
</tr>
<tr>
<td>v3.0</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>v3.1</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>V3.2</td>
<td>18</td>
<td>60</td>
</tr>
</tbody>
</table>
Outline

Interactive feature demos
... by the team
Outline

• Workbench & User Interface
• Analytics / Mining
• PMML - Standardizing predictive models
• Streaming Executor
• Linked Data & Semantic Web

• KNIME Server & Cloud Products
• KNIME Big Data Extensions
Workbench & User Interface
Workflow Coach!

NEW since 3.2: Workflow Coach recommends matching nodes.

- Three options
  - Based on community statistics
  - Analyze your workspace*
  - From local server**

* Requires KNIME Personal Productivity License
** Requires KNIME Server License

For more information contact us: info@knime.com
Automated installation of features

- If you open a workflow with a missing node
  - The correct plugin is automatically proposed for installation
Easier import/export of workflows

New File Extension for workflows and groups

Use the KNIME protocol for starting workflows
Analytics / Mining
Analytics / Mining

- Trees and Tree Ensembles
- (Forward) Feature Selection
- Deep Learning in KNIME: DeepLearning4J Extension
Analytics / Mining
Trees and Tree Ensembles

Greg Landrum
The general idea is to take advantage of the “wisdom of the crowd”: combining predictions from a large number of weak predictors leads to a more accurate predictor. This is called “bagging”.

Typically: for classification the individual models vote and the majority wins; for regression, the individual predictions are averaged.
How does bagging work?

Pick a different random subset of the training data for each model in the ensemble (bag).
An extra benefit of bagging: out of bag estimation

Allows testing the model using the training data: when validating, each model should only vote on data points that were not used to train it.
Random Forests

- Bags of decision trees, but an extra element of randomization is applied when building the trees: each node in the decision tree only “sees” a subset of the input columns, typically $\sqrt{N}$.

- Random forests tend to be very robust w.r.t. overfitting (though the individual trees are almost certainly overfit)

- Extra benefit: training tends to be much faster
Gradient Boosting

• Another algorithm for creating ensembles of decision trees
• Starts with a tree built on a subset of the data
• Builds additional trees to fit the residual errors
• Typically uses fairly shallow trees
• Can introduce randomness in choice of data subsets ("stochastic gradient boosting") and in variable choice.
Trees and Tree Ensembles: Changes “under the hood”

• Support of binary splits for nominal attributes
• Missing value handling
• Support of byte vector data (high-dimension count fingerprints)
• Code optimization
  – Runtime
  – Memory
Trees and Tree Ensembles: New nodes

• Gradient Boosting
  – Also based on tree ensembles
  – Boosting: Improving an existing model by adding a new model
  – Shallow trees

• Random Forest Distance
  – Distance measure induced by a random forest
  – Based on proximity
Demo: Tree Ensembles
Gradient Boosting dialog
Wine-quality prediction workflow

White wine results

<table>
<thead>
<tr>
<th>File</th>
<th>Confusion Matrix - 0:44 - Scorer (Random Forest OOB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality \ q...</td>
<td>low</td>
</tr>
<tr>
<td>low</td>
<td>854</td>
</tr>
<tr>
<td>high</td>
<td>206</td>
</tr>
<tr>
<td>Correct classified: 3,264</td>
<td>Wrong classified: 654</td>
</tr>
<tr>
<td>Accuracy: 83.308 %</td>
<td>Error: 16.692 %</td>
</tr>
<tr>
<td>Cohen's kappa (k) 0.605</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File</th>
<th>Confusion Matrix - 0:48 - Scorer (Random Forest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality \ Pr...</td>
<td>low</td>
</tr>
<tr>
<td>low</td>
<td>224</td>
</tr>
<tr>
<td>high</td>
<td>49</td>
</tr>
<tr>
<td>Correct classified: 817</td>
<td>Wrong classified: 163</td>
</tr>
<tr>
<td>Accuracy: 83.367 %</td>
<td>Error: 16.633 %</td>
</tr>
<tr>
<td>Cohen's kappa (k) 0.614</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File</th>
<th>Confusion Matrix - 0:50 - Scorer (Stochastic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality \ Pr...</td>
<td>low</td>
</tr>
<tr>
<td>low</td>
<td>245</td>
</tr>
<tr>
<td>high</td>
<td>61</td>
</tr>
<tr>
<td>Correct classified: 826</td>
<td>Wrong classified: 154</td>
</tr>
<tr>
<td>Accuracy: 84.286 %</td>
<td>Error: 15.714 %</td>
</tr>
<tr>
<td>Cohen's kappa (k) 0.644</td>
<td></td>
</tr>
</tbody>
</table>
Wine-quality prediction workflow

Red wine results

- **Random Forest**
  - Confusion Matrix - 0:32 - Scorer (Random Forest OOB)
  - Correct classified: 1,049
  - Wrong classified: 230
  - Accuracy: 82.017 %
  - Error: 17.983 %
  - Cohen's kappa (κ): 0.637

- **Gradient boosting**
  - Confusion Matrix - 0:37 - Scorer (Random Forest)
  - Correct classified: 260
  - Wrong classified: 60
  - Accuracy: 81.25 %
  - Error: 18.75 %
  - Cohen's kappa (κ): 0.621

- **Stochastic Gradient Boosting**
  - Confusion Matrix - 0:41 - Scorer (Stochastic)
  - Correct classified: 256
  - Wrong classified: 64
  - Accuracy: 80 %
  - Error: 20 %
  - Cohen's kappa (κ): 0.596
Random Forest distance workflow
Similarity search dialog

![Similarity search dialog](image)

- **Connected Distance Measure**
  - Random Forest Distance
  - for columns: "minx", "miny", "maxx", "maxy", ... 64 more>

- **Search Options**
  - Coefficient Type: distance
  - Neighbors selection: nearest (most similar)
  - Neighbor Count: 4
  - Use range filter (min, max distance)
  - Minimum: 0.70
  - Maximum: 1.00

- **Output Options**
  - Output column prefix: nearest neighbor
  - Representative Column (2nd input?): <RowID>
  - RowID Suffix Separator: _
## Input table

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Item</th>
<th>Color</th>
<th>Size</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>A</td>
<td>Red</td>
<td>S</td>
<td>$10</td>
</tr>
<tr>
<td>Item 2</td>
<td>B</td>
<td>Blue</td>
<td>M</td>
<td>$15</td>
</tr>
<tr>
<td>Item 3</td>
<td>C</td>
<td>Green</td>
<td>L</td>
<td>$20</td>
</tr>
</tbody>
</table>

Note: The table displays the details of various items with their respective colors, sizes, and prices.
Nearest Neighbors
# Nearest Neighbors

<table>
<thead>
<tr>
<th>Row ID</th>
<th>Image</th>
<th>g</th>
<th>r</th>
<th>distance</th>
<th>g (iso)</th>
<th>r (iso)</th>
<th>distance (iso)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row01</td>
<td>Zebra</td>
<td>3</td>
<td>0.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row02</td>
<td>Zebra</td>
<td>0</td>
<td>0.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row03</td>
<td>Zebra</td>
<td>1</td>
<td>0.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row04</td>
<td>Zebra</td>
<td>2</td>
<td>0.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Row05</td>
<td>Zebra</td>
<td>3</td>
<td>0.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analytics / Mining
Feature Selection
Feature Selection

• “Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction” [Wikipedia]

• Why?
  – Better generalization
  – Simplification of the model
  – Shorter training times

• Dozens of methods to do that...
Feature Selection

- Backward Feature Elimination:
  Start with full feature set, iteratively remove ‘worst’ feature

- Forward Feature Selection:
  Start with empty feature set, iteratively add ‘best’ feature
Feature Selection nodes

• Same loop structure as former Backward Feature Elimination nodes
• Different strategies
  – Forward selection
  – Backward elimination
• Uses Flow Variable as score
  – Flexibility
• Preconfigured meta nodes for both strategies
Analytics / Mining
Deep Learning in KNIME: DeepLearning4J Extension

Christian Dietz
What is Deep Learning?

• State-of-the-art algorithms for learning tasks on images, videos, text or sound
• Multi-Layer Neural Networks
• Regression and Classification, Unsupervised Learning, Reinforcement Learning, ...
What is DeepLearning4J?

- Open-source Deep Learning framework
- Supports state-of-the-art network architectures
- GPU/CPU support
- Distributed computations on Apache Spark and Hadoop
- Word2Vec for Text Mining
Deep Learning in KNIME

• Integration of DeepLearning4J
• Visually assemble networks using KNIME nodes
• Integrates with other KNIME extensions, e.g. KNIME Image Processing / KNIME Text Mining
• Networks can be trained and executed on GPU and CPU
Basic Deep Learning workflow

Start
Network

Create Architecture

Learn on Data

Create Predictions
Celebrity Face Recognition

• Problem Description
  – Recognize faces of celebrities in web images.

• Solution
  – Image classification using a state-of-the-art deep convolutional network architecture (AlexNet).
MSRA-CFW: Data set of celebrity faces

- 27210 faces of 20 celebrities
Workflow

- Read Images
- Joiner
  - join with label
- Preprocess Images
- Partitioning
  - split in train and test set
- Shuffle
- AlexNet
- DL4J Feedforward Learner (Classification)
- DL4J Feedforward Predictor
- Scorer
Workflow
Workflow
Workflow
Workflow
Workflow
Workflow
Active Learning

- Labs Extension
- Involve user to construct training data set
- Workflow loop to query and label ‘interesting’ data points
- Used user-labeled data set on remaining data
Active Learning (example from Node.Pedia)
Statistics nodes

• Several new useful statistic nodes in KNIME Labs.
• Thanks to Bob Muenchen (University of Tennessee).
• Work in progress! We are still adding nodes.

• Missing anything? See R integration
R Integration

• Rewrite of infrastructure
  – Significantly faster
  – Concurrent execution
• No change of usage model
PMML - Standardizing Predictive Models
What is PMML?

- Predictive Model Markup Language
- XML based standard for predictive models
- KNIME can export most of its models as PMML
- To consume 3rd-party models, a scoring engine such as Zementis Adapa/UPPI is more suitable
PMML Creation in KNIME

- Special port for PMML models
- Supported by most KNIME learners
  - Decision Trees, Neural Nets, ...
  - Ensembles
- Also used for Preprocessing
  - Normalizing, Binning, Missing Values, ...
- Modular PMML
  - Built step by step parallel to the data flow
Demo: Modular PMML
Decision Tree to Ruleset

• Transforms a decision tree to a PMML ruleset model
  – Easier to interpret

• Also outputs rules as a KNIME table
  – Easier to export & deploy
  – Can be manipulated using standard KNIME nodes
Applying Rulesets

- New node: Rule Engine (Dictionary)
  - Input: data and ruleset table
  - Output: Results and optional PMML model
- Import rules from other sources
- Mix rules from multiple sources
Streaming Executor
Streaming

- Default Execution
Streaming

• Streaming Execution
Streaming

- Row-wise
- Process, pass & forget → Faster with less I/O overhead
- Concurrent execution
Demo: Streaming Executor
Streaming – Pros & Cons

Advantages
• Less I/O overhead (process, pass & forget)
• Parallelization

Disadvantages
• No intermediate results, no interactive execution
• Not all nodes can be streamed
Streaming – Streamed nodes

- More than 100 nodes
- Text Processing nodes
- Image Processing nodes
- ...
Semantic Web / Linked Data

Tobias Kötter
Semantic Web

• Access the wealth of Semantic Web from within KNIME
• Create your own Semantic Web with the Memory Endpoint
• Read/write support for Semantic Web file formats
• Manipulate triple stores via SPARQL
• Usage model similar to database integration

Sponsored development by Boehringer-Ingelheim, Germany
Wrap-up – KNIME Analytics Platform
Utility Nodes & Integrations
KNIME RESTful Web Service Client Nodes

• REST Client nodes: Get / Post / Put / Delete Resources
• Follow-up of famous KREST community extensions
• Integrating with KNIME’s XML/JSON Processing nodes
• Powerful Configuration
• Extensible (e.g. custom auth types)

Follow-up on KREST community extension
KNIME Tableau Integration

- Tableau: Popular (commercial) visualization and dashboard application

- New KNIME nodes to:
  - Write native Tableau files (TDE)
  - Send data to Tableau server

Thanks for testing and feedback goes to Forest Grove Technology
KNIME’s partner in Australia
KNIME Server

Jon Fuller
KNIME Server

Shared Repositories
- Data
  - XLS, GDF, TXT, CSV, XML, many more
- Nodes & Metanodes
- Workflows

Access Management
- Nodes, Files, Applications
  - LDAP, Java
- Flexible Execution
  - Remote, Scheduled

Web Enablement
- Content, Services
  - REST API
KNIME Server Improvements

• WebPortal enhancements
• Advanced scheduled execution
• Extended REST API
• For the sysadmins...
  – KNIME Server installer (Windows/Linux)
  – Administration pages – make admin easy
• Distribute license files for Analytics Platform extensions
  E.g. Big Data Extensions, Personal Productivity (incl. Workflow Difference)
Advanced scheduled execution
KNIME WebPortal enhancements (JavaScript Views)

• “Paged Table” view to page-wise scroll through data supporting sorting, search and selection

Sponsored development by Genentech, USA
WebPortal templates
KNIME Server – Extended REST API

- Integrate KNIME Server functionality with IT infrastructure
- Execute workflows, check server status, and more

See Blog Posts for detailed tutorials:
  - https://www.knime.org/blog/giving-the-knime-server-a-rest
  - https://www.knime.org/blog/the-knime-server-rest-api
Execute workflow via REST API

- Add Quickforms to define workflow API
Workflow Automation
Demo: Workflow Automation
Workflow orchestration via REST API

- Calling a remote workflow

Part of the Personal Productivity Extensions
For the Sysadmins...
KNIME Server Installer

Step-by-step guided Server installation (Windows and Linux)
KNIME Server – Admin made easy

- KNIME Administrator is often not a KNIME Analytics Platform user

- Make tasks like user administration easier

- Get an overview of the KNIME Server health
Go to the administration portal
KNIME Server – Admin made easy

![KNIME Server Administration](image)

**Status Info**

- **Host:** unknown
- **Version:** 4.3.0
- **Uptime:** 0 days, 0 hours, 6 minutes, 46 seconds,
- **Executors Info:** KNIME Executors associated with user: (none), 0: RMI Executor 0 (user: (none), port: 50100, uptime: 00:00:20, status: Accepting Jobs, #jobs: 1)

**License and Users info**

- **License Type:** KNIME Server
- **Expiration Date:** 2015-08-15
- **Company:** KNIME.com AG
- **Customer:** Jon Fuller
- **Host Identifiers:** 
  - *(AWS Instance IDs)*
- **Comment:**
- **Users**

©2016 KNIME.com AG, support@knime.com
### License Info

**License Type:** KNIME Server  
**Expiration Date:** 2016-07-08  
**Company:** KNIME.com AG  
**Customer:** Jon Fuller  
**Host Identifiers:** { "Host IP": "10.0.0.26" }  
**Comment:** Demo license

#### Users

**Desktop**
- limit: 5  
- current: 0

**Webportal**
- limit: 50  
- current: 1  
  **active Users**
  - name: admin  
  - last login: 2016-06-09T10:23:32.268

**Webservice**
- limit: 50  
- current: 1  
  **active Users**
  - name: admin  
KNIME Server – Admin made easy
KNIME Server – Admin made easy
KNIME Server – Admin made easy
KNIME Server – License Distribution

- License files no longer required in client installation
- Checked out from KNIME Server
- Centrally managed, less configuration required
KNIME Workflow Difference
Demo: KNIME Workflow Difference
KNIME Workflow Difference - Summary

• Identifies changes in workflow-structure
• Aligns workflows to identify differences
• Part of KNIME Personal Productivity Extensions
• KNIME Server license includes Personal Productivity license
KNIME in the Cloud
KNIME Cloud Analytics Platform

- Get started quickly
- Bring your analytics to your cloud hosted data
- Scale your analytics workflow up to 32 cores and 448 Gb RAM
KNIME Cloud Analytics Platform

At KNIME®, we build software for fast, easy and intuitive access to advanced data science, helping organizations drive innovation. KNIME Cloud Analytics Platform brings the power and flexibility of the open source KNIME Analytics Platform to the cloud for the first time on Microsoft Azure.

With more than 1000 modules, hundreds of ready-to-run examples, community contributions, and tool integrations, KNIME Cloud Analytics Platform helps any data scientist solve their most complex data puzzles in a reliable and scalable framework with plenty of room to grow.

The combination of KNIME Cloud Analytics Platform on Microsoft Azure gives data scientists the power and flexibility to visually tackle complex data problems without the hassles of on-premise infrastructure or installation.

With KNIME Cloud Analytics Platform, the only restriction is your creativity.

KNIME Cloud Analytics Platform - Launch
KNIME Cloud Analytics Platform - Launch

![Microsoft Azure interface for KNIME Cloud Analytics Platform](image)

- **Basics**
  - Name: KNIMECloudAP
  - User name: jon
  - Password: ********
  - Subscription: Pay-As-You-Go
  - Resource group: knimecloud
  - Location: West Europe

- **Create virtual machine**
  - Size: Choose virtual machine size
  - Settings: Configure optional features
  - Summary: KNIME Cloud Analytics Platform
  - Buy:
KNIME Cloud Analytics Platform - Launch
KNIME Cloud Analytics Platform - Launch
KNIME Cloud Analytics Platform - Connect
KNIME Cloud Analytics Platform - Connect
KNIME Cloud Analytics Platform - Connect
KNIME Cloud Analytics Platform - Connect

Welcome to KNIME!

We are happy to help with your first steps if you let us know your email address and register here. We promise to only send a few getting started emails.

If you do want regular updates on new features, releases and KNIME events (and nothing else!) - do register for our official mailing list here - if you haven’t done so already when downloading KNIME.

If you are new to KNIME, why not open one of the example workflows by double clicking one of the workflows in the explorer in the upper left corner? You can also drag&drop nodes from the node repository (bottom left) to the workflow editor after opening an existing or creating a new workflow. But first you may want to install one of the many extensions for processing of additional data types or other tool integrations, such as R and JFreeChart. And don’t forget to take a look at our Quickstart Guide and the Learning Hub, which provides links to great resources for beginners and seasoned KNIMEers.

This page will be displayed upon startup of KNIME but you can customize the content using the checkboxes at the bottom.

Some of the OpenMS KNIME Nodes requirements are missing. Double click for details.
Database Integration and KNIME Big Data Connectors

Tobias Kötter
Database Integration - Recap

- Visually assemble complex SQL statements
- Connect to almost all JDBC-compliant databases
- Preconfigured nodes to connect to various databases
- Harness the power of your database within KNIME
New Database Nodes

• Database Pivot
• Database Numeric-/Auto-Binner and Apply-Binner
• Database Sampling with support for stratified sampling
• Parameterized Database Query
• Python Script (DB)/(Hive)
KNIME Big Data Connectors - Recap

• Package required drivers/libraries for HDFS, Hive, Impala access
• Performs operations on Hadoop
• Extends the open source database integration
• Preconfigured connectors
KNIME Big Data Connectors

- Support for Kerberos secured cluster
- Improved driver handling
- New nodes:
  - httpFS Connector
  - webHDFS Connector
KNIME Spark Executor
KNIME Spark Executor - Recap

• Based on Spark MLlib
• Scalable machine learning on Hadoop
• Algorithms for
  – Classification (decision tree, naïve bayes, ...)
  – Regression (logistic regression, linear regression, ...)
  – Clustering (k-means)
  – Collaborative filtering (ALS)
  – Dimensionality reduction (SVD, PCA)
Familiar Usage Model

• Usage model and dialogs similar to existing nodes
• Spark nodes start and manage Spark jobs
• No coding required
In-Hadoop Processing

- Spark RDDs as input/output format
- Data stays within your cluster
- No unnecessary data movements
- Several input/output nodes e.g. Hive, hdfs files, ...
Combine with Existing KNIME Nodes
Let KNIME Control Your Spark Jobs
47 Spark Nodes and Counting

[Diagram of Spark Nodes and Counting]

© 2016 KNIME.com AG. All Rights Reserved.
KNIME Spark Executor

- Kerberos secured cluster support
- Easier installation
- Supports Spark version 1.2, 1.3, 1.5 and 1.6
Use Case: Smart Meter Analysis

• More than 170 Mio rows with energy usage data from smart meters
• Uses KNIME Analytics Platform, Big Data Connectors and Spark Executor to forecast energy consumption
Demo: KNIME Spark Executor
Summary

• Constantly improving, also thanks to feedback of customers/partners/community

• Questions / Interested in demo / comments? → Talk to us in the breaks / at the booth
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME.com AG under license from KNIME GmbH, and are registered in the United States.