What’s New and Cooking in KNIME Analytics Platform

Bernd Wiswedel
Data Science: More Than Just Methods!

Create

Gather & Wrangle

Model & Visualize

Productionize

Deploy & Manage

Consume & Optimize

KNIME Analytics Platform

KNIME Extensions

KNIME Integrations

Community Extensions

Partner Extensions

KNIME Server

KNIME WebPortal

Data Science as a Service
Components are for sharing

Languages

Workflows

Building Blocks

Services

Applications
Components are for sharing...
Components are for sharing

Components in KNIME:

• **Continued Abstraction of Data Science:**
  – Machine Learning Automation
  – Feature Selection & Engineering
  – Interpretability
  – Model Management and Deployment
  – …

• **Sharing of Reusable Pieces**
  – Integrating & Blending Corporate Data
  – Standardize Data Access
  – Define Corporate Reporting Schema
  – …
Reusable Components on KNIME Hub

- Components can be used similar to real nodes
- Including configuration, error messages and documentation

EXAMPLES (knime@hub.knome.com)
- Components
  - Automation
    - Dimensionality Reduction (LDA)
    - Parameter Optimization
  - Data Manipulation
    - Column Filter (by Index)
    - Column Name Editor
    - Quantile Normalization
  - Financial Analysis
    - Asset prices to one-period (log) returns
    - One-period log returns to multi-period log returns
    - One-period returns to multi-period returns
  - Guided Analytics
    - Classification Threshold Analysis
  - Life Sciences
  - Model Interpretability
    - Compute LIME
    - Partial Dependence Pre-processing
  - Text Processing
    - Document Pre-processing
    - PubMed Document Extractor
  - Time Series
    - Aggregation Granularity
    - ARIMA Learner
    - ARIMA Predictor
    - Auto ARIMA Learner
(New) Database Integration
Recap: Database Extension

- Visually assemble complex SQL statements
- Connect to almost all JDBC-compliant databases
- Harness the power of your database within KNIME
- Operations are performed within the database
Usability Improvements

• Improved schema handling
• Flexible type handling
• Rich SQL editor
Framework Improvements

• Driver management
• Parallel execution
• Streaming execution
Big Data Extensions
KNIME Extension for Apache Spark

• Read, Process and analyse data in Apache Spark
• Data stays in cluster, KNIME nodes manage Spark jobs
• Revised Spark model learner nodes based on current spark.ml APIs
• Revised nodes can handle categorical columns
• Spark Predictor (Classification) node provides conditional class probabilities
Apache Livy Support

• Gives out-of-the-box Spark compatibility with:
  – Hortonworks HDP
  – Amazon EMR
  – Microsoft Azure HDInsight
• Cloudera packaging provided by KNIME
• Revised node dialog
PySpark Script Nodes

- **Apache Spark**
  - IO
  - Column
  - Mining
- **Misc**
  - Java Snippet
  - Management
- **PySpark**
  - PySpark Script (1 to 1)
  - PySpark Script (1 to 2)
  - PySpark Script (2 to 1)
  - PySpark Script (2 to 2)
  - PySpark Script Source

**PySpark Script Source**

```
# Custom pySpark code
# SparkSession can be used with variable spark
# The input dataFrame(s) [dataframe]
# The output dataFrame(s) must be: [resultDataFrame]
resultDataFrame = dataFrame
```

Execution finished.

```
resultDataFrame(10 of 50 rows):
Row 0: [0.490014491038, 0.50779002433, 0.12092496075, 0.316210587440, 0.03229267943, 0.12092496075, 0.316210587440] Cluster_3
Row 1: [0.430014491038, 0.54779002433, 0.12092496075, 0.316210587440, 0.03229267943, 0.12092496075, 0.316210587440] Cluster_3
```
Simplified Kerberos Support

- User authentication against Kerberos or Active Directory
- Mostly used Big Data & Database Nodes
Integrations
Integrations / Extensions: Google Drive

Google Authentication & Google Drive Connectivity
Integrations / Extensions: Tableau

- Supporting Hyper
- Write Extracts and push them to Tableau Server
- New capabilities
  - Appending files
  - Date & Time Support
Integrations / Extensions: Jupyter

KNIME calls Jupyter  

Jupyter calls KNIME

**Python Script (1⇒1)**

```
In [1]:
import pandas as pd
import knime

knime.executable_path = "C:\Users\glandrum\Desktop\KNIME_3.7\knime.exe"

In [2]:

workspace = "C:\Users\glandrum\KNIME workspaces\KNIME nightly"

Now pick a workflow and make sure it's the right one by displaying an SVG of the workflow:

In [3]:

workflow = "KNIME Store/Py/KNIME Jupyter/ChEMBL Target Data"

knime.Workflow(workflow_path=workflow, workspace_path=workspace)

Out[3]:

```

In [4]:

interesting_targets = {'target_chebi_id': ['CHEMBL214', 'CHEMBL219', 'CHEMBL1982']}

df = pd.DataFrame(interesting_targets)
```
Integrations / Extensions: KNIME Plotly Integration

- (KNIME Labs) Plotly: An integration with the well known open-source visualization library
Visualizations: Heatmap, Tile View & Dendrogram

4:00 PM  **Session 3: Components are for Sharing**
What's New and Cooking in KNIME Analytics Platform
Bernd Wiswedel (KNIME)

Auditing Human Decisions in the Background Check Industry
SJ Porter (HireRight)

**Visualizing with Guided Analytics: For End Users AND Experts**
Phil Winters and Team (KNIME)

Friday, November 8 / 204 Amphitheater

8:30 AM  Registration, coffee, breakfast

9:00 AM  **Session 4: Life and Other Data Science**
What's New and Cooking in the KNIME WebPortal
Jeany Prinz and Paolo Tamagnini (KNIME)
Utility Nodes: T-SNE (Dimensionality Reduction)
Utility Nodes: Duplicate Row Filter

- Find duplicates and process them
- Based on KNIME Community wishlist
Model Learning & Model Interpretability
Integrations / Extensions: XGBoost

- Popular open-source library for optimized distributed gradient boosting
- Often used in machine learning competitions
Model Interpretability: Lime, Shap & Co

• Nodes supporting model interpretability on a per record basis: LIME, SHAP, and Shapley Values
Model Interpretability: Partial Dependence Plot

• Partial Dependence/ICE Plot show marginal effect a feature has on the predicted outcome.
Model Interpretability

Dean Abbott
Co-Founder and Chief Data Scientist at SmarterHQ
3d · Edited

I love all of the new machine learning features and think I will be using all of them soon, especially model interpretation — record level and partial dependency plots

KNIME
6,035 followers
4d

Check out the features around new #machinelearning functionality for model interpretation, #automation, and a number of new #algorithms released in #KNIME Analytics Platform 4.0. For more info, visit https://bit.ly/2Xd4c5T
Amazon Comprehend & Translate

Integration with AWS for text mining tasks

Amazon Comprehend Syntax Tagger: Part-of-speech tagger based on AWS Comprehend (https://aws.amazon.com/comprehend/)

Amazon Authentication

KNIME Labs
- Amazon Web Services
  - Amazon Comprehend
    - Amazon Comprehend (Dominant Language)
    - Amazon Comprehend (Entity Tagger)
    - Amazon Comprehend (Key Phrases)
    - Amazon Comprehend (Sentiment)
    - Amazon Comprehend (Syntax Tagger)
  - Amazon Translate
  - Amazon Translate
  - Amazon Authentication

Table Creator → RSS Feed Reader → Column Filter → Amazon Comprehend (Syntax Tagger) → Tagged Document Viewer (JavaScript)
KNIME Machine Learning

H2O Isolation Forest for Anomaly Detection

ONNX Integration via TensorFlow
One last thing...
What’s cooking...
What’s cooking (I): Connectivity

• Google BigQuery Connector

• File Handling improved (e.g. bulk read from S3)

• PowerBI Connector Node
What’s cooking (II): ML related...

• Active Learning revised (loops and nodes) & Model building via Weak Supervision

• Integration with Amazon Personalize
Some highlights in KNIME Analytics Platform 4.1...

• Components, Component, Components (easier editing, sharing, parameterization)
• ... and tons more
What’s Cooking

Check out nightly builds:

KNIME Analytics Platform – Nightly Builds

• Nightly Builds publicly available:

https://www.knime.com/form/nightly-build