What’s new ...

Bernd Wiswedel
KNIME.com AG, Zurich, Switzerland
• Two feature releases last year: 2.10 & 2.11

• Documented in “What’s new summary”, YouTube, Changelogs
What’s new summary pages

New node to implement the DBCSCAN (density based clustering) algorithm.

The node has two input ports: one for the data and one for the distance formula. With the appropriate distance function, DBCSCAN is able to cluster oddly shaped groups of data.

kNN now supports more distance functions

k Nearest Neighbors (Distance Function) implements kNN and supports more distance functions, besides the Euclidean function. The node provides an additional input port for the distance formula.

Target Shuffling

This node shuffles the values randomly inside a selected column to assess the statistical accuracy of data mining results.

IO

Writer Nodes Improvements

video [http://youtube.be/9cL1UtZjg](http://youtube.be/9cL1UtZjg)
Learning Hub & YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.

Beliebte Kanäle auf YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.

Beliebte Kanäle auf YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.

Beliebte Kanäle auf YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.

Beliebte Kanäle auf YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.

Beliebte Kanäle auf YouTube

Welcome to KNIME TV
2,800 Ausaude vor 8 Monaten
This is a short introductory video to what you can expect to find in this KNIME TV channel.
Release Stats

• KNIME 2.10
  – 23 new nodes (node sets)
  – 89 enhancements

• KNIME 2.11
  – 35 new nodes (node sets)
  – 68 enhancements
Outline

• Interactive feature demos
• ... by the developers*
• Questions? Right after the session

*) with exceptions
Agenda

- Social Media Connectors: Google API & Twitter
- Database & Big Data
- Analytics: PMML
- Analytics: Custom Distance Measures, etc
- Enhanced Time Series Analysis
- Interactive Python Scripting
- Tips & Tricks: UI & Utility Nodes
- JSON Processing
Social Media Connectors: Google API & Twitter

- Aaron Hart -
Google API/Analytics Connectors

Dialog - 4:1 - Google API Connector

- Service account email:
  809096066640-54mcsq5n459sn1tda228borloem20

- P12 key file location:

- Scopes:

Dialog - 4:2 - Google Analytics Connector

- Account:
  www.knime.org

- Webproperty:
  http://www.knime.org

Dialog - 4:2 - Google Analytics Query

- User Type:
  A boolean indicating if a user is new or returning. Possible values: New Visitor, Returning Visitor.
  dimension = userType

- Metrics:
  - pageviews
  - users
  - newUsers
  - percentNewSessions
  - sessions
  - sessionDuration
  - bounceRate
  - pagePath:
    - blog

- Sort:
  - pageviews

- Start date: 2013-01-01
- End date: 2016-01-01

- Default:

- "start-date" and "end-date" are controlled by variables.
Twitter Nodes

Twitter API Connector

Twitter Timeline

Twitter Search

Twitter Users
Twitter Data Collection

Twitter API Connector -> Twitter Search -> Database Update -> Row Filter -> Column Filter -> Database Writer

MySQL Connector -> Connect to DB
#independence #bettertogether #BBC #voteyes #news #yes #UK #scottishindependence
Databases & Big Data

- Tobias Koetter -
Databases
Database Connectors

• Preconfigured connectors
  – DB specific dialog
  – DB specific behavior/capability
    • SQL dialects
    • Aggregation methods
  – Bundle necessary JDBC drivers

• General database connector
  – Can connect to any JDBC source
New Utility Nodes

- Drop table
  - missing table handling
  - cascade option
- Execute any SQL statement also DDL
- Manipulate existing queries
New Manipulation Nodes

- Filter rows and columns
- Join tables/queries
- Sort your data
- Aggregate your data
- User friendly dialogs
Big Data
Pre-processing on Hadoop

- KNIME Big Data Extension (commercial)
  - Package required drivers/libraries for specific HDFS, Hive, Impala access
  - Runs on Hadoop
- Preconfigured connectors
  - Hive (Big Data Extension)
  - Cloudera Impala (Big Data Extension)
  - HP Vertica (OS)
  - ParStream (Partner)
Hive/Impala Loader

- Upload a KNIME data table to Hive/Impala
- Part of the commercial Big Data Extension
Pre-processing Example Smart Meters: Classic

This workflow reads Ireland's electricity data, converts the dates from the proprietary format into datetime values, and groups kW values by:
- day
- hour
- intra-day times
- month
- year
- week

It also aggregates average and % values for the k-Means procedure

Irish Smart Energy Meter Trials
- July 2009 – Dec 2010
- 6000 meters
- roughly 176m rows of data
Pre-processing Example Smart Meters: on Hadoop

This workflow reads Ireland’s electricity data, converts the dates from the proprietary format into datetime values, and groups kW values by, day, hour, intra-day times, month, year, week. It also aggregates average and % values for the k Means procedure.

Transformation/Aggregation
- Before: >1 day
- Hadoop/ParStream: <.5h

http://www.knime.org/white-papers#bigdata
HDFS File Handling

• New nodes
  – HDFS Connection
  – HDFS File Permission

• Utilize the existing remote file handling nodes
  – Upload/download files
  – Create/list directories
  – Delete files
HDFS File Handling
Analytics:
Modular PMML & PMML Translation

- Alexander Fillbrunn -
Modular PMML

- **Predictive Model Markup Language**
- Documents your preprocessing and modeling results
- Built in a modular fashion parallel to the data flow
- Start with an empty PMML document, add transformations and finally a model
- Consumable by various predictors
PMML Translation

• PMML to SQL
  – Execution of models directly in the database
  – No data transfer to KNIME required
• PMML to Java
  – Model is compiled into Java bytecode and executed
  – Up to 4 times faster execution
  – Preprocessing and model evaluation in one node
Analytics:
Custom Distance Functions

- Bernd Wiswedel -
Distance Functions

• Used in many data mining algorithms
  – Clustering
  – Classification (kNN)
• Similarity search
• (Address) Deduplication

• New: Extensible framework to define such measures
Custom Distance Functions

- Separate port object representing distance function
- Parameterized distance functions, e.g. Tanimoto vs. Dice → Tversky
Target Shuffling

• Assess validity of your data mining result
• Easier identify “overlearning” with many variables

• Idea: Resample target in training data to break relationship between in- and output
• Learn & score a model
• Do this many times and compare to model on unmodified data
Enhanced time series analysis

- Iris Adä -
New Time Series Features

- New Nodes

- New Meta Nodes
Interactive Python Scripting

- Patrick Winter -
Python: Integrated into KNIME

• General-purpose language
• Wide spread use in chemistry and life science
• Wide range of libraries, including:
  – Scientific computing
  – Text processing
  – Image processing
Python: Pandas as data representation

- Provides data structures and data analysis tools
- Builds on the scientific computing library NumPy
- KNIME table ↔ Pandas DataFrame
Python: Nodes

- Source
- Script
- Script (2:1)
- View
- Learner
- Predictor
- Object Reader
- Object Writer
Demo
Tips & Tricks:
UI & Utility Nodes

- Bernd Wiswedel -
UI Improvements

• Welcome Page: Overview of updates & Shortcuts to most recently used workflows

• Quick Node Insertion: Try “Ctrl-Space” Search and insert nodes quickly

• Auto-Save: Keep shadow copy of workflow
Utility Nodes: Data Validation

- Assert table structure at runtime - fail or fix
  - Useful in workflow automatization
  - Useful in conjunction with “to Report” nodes
Utility Nodes: Auto Type Cast

- Guess (primitive) column types for all-string tables
- Useful for parsing configuration files

This workflow shows one application for the Column Auto Type Cast node. The background is that you want to control a (complex) workflow via several flow variables. The actual values are stored in a configuration file, one variable per line (see also output of the File Reader):

```
variable1: value1
variable2: value2
```

As the variables can be a mixture of string, integer, doubles, dates, etc., the common type of the value column will in most cases be string. However, to use the values in the node configuration, they need to have the correct type. This is where the Column Auto Type Cast node gets involved after transposing the table. It will automatically detect the most specific type in all configured columns. The final Table Row to Variable node will then create flow variables of the correct type.

Choose a date format: `yyyy-MM-dd` `HH:mm`

Missing value pattern: `<none>`

Do a quick scan. Number of rows to consider: 1000
JavaScript Nodes

- Same representation in Client and Webportal
- Interactivity
Bit & Byte Vector Nodes

- Nodes to parse/write bit & byte vectors
- Distance calculation & association rule learning
JSON Type & Utility Nodes

- JSON – JavaScript Object Notation
- Hierarchical data format

```
{   "UGM sessions":{
      "session title":"Money, Love and Gambling",
      "talks":[
        {
          "presenter":"Jörg Neumann",
          "title":"Successful Cross Selling – Automated Data Mining",
          "duration":30
        },
        {
          "presenter":"Cathy Pearl",
          "title":"What's Love Got To Do With It? Analyzing Tweets",
          "duration":30
        },
        {
          "presenter":"Walter Bonifazi",
          "title":"Bet Responsibly. On Your CRM Too.",
          "duration":30
        }
      ]
    }
}```
JSON Type & Utility Nodes - Applications

- Web Service consumption ((K)REST)
- NoSQL databases / MongoDB
- KNIME Web Service
JSON – “JSON Path”

```
{ "UGM sessions": {
   "session title": "Money, Love and Gambling",
   "talks": [ {
      "presenter": "Jörg Neumann",
      "title": "Successful Cross Sell",
      "duration": 30
   }
   ]
}
```

Path: 

```
$.['UGM sessions'].['talks'].[*].['presenter']
```
Wrap-Up

• Many more features but not covered here
• Development driven by community, customers, sponsors

• More is in the works...