HTWK Leipzig

Oberseminar Data Mining

Systeme und Tools zum Data Mining: RapidMiner

Ricardo Hofmann, Matthias Neubert | Leipzig | 11. Mai 2010

Motivation

Inhalt

Entwicklung

- entwickelt an der Technischen Universität Dortmund
- erschienen im Jahre 2001
- anfangs unter dem Namen YALE - (”Yet Another Learning Environment”) veröffentlicht
- 2007 umbenannt in RapidMiner
- zu diesem Zeitpunkt Version 4.0
- seit Februar 2010 Version 5.0

Produktübersicht I

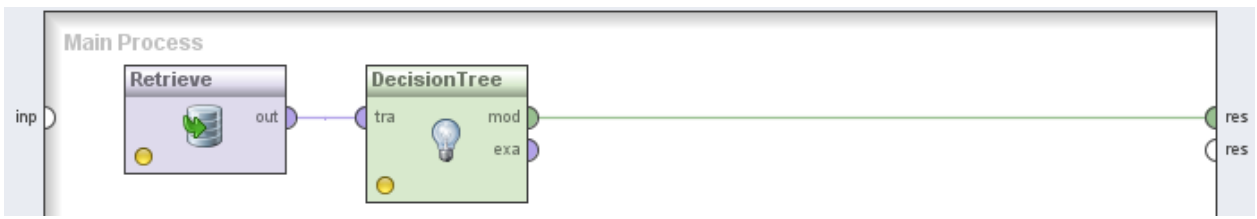
- lizensiert unter der AGPL bzw. proprietär
- Open-Souce-Software
- erhältlich in der Community- oder Enterprise Edition
- komplett in Java geschrieben und damit auf allen großen Plattformen lauffähig
- bietet die Möglichkeit über Java API von externen Programmen genutzt zu werden

Produktübersicht II

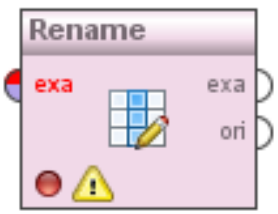
- Umgebung zum maschinellen Lernen und zur Umsetzung des KDD-Prozesses (insbesondere des Data Minings)
- Realisierung mittels einer Reihe von Operatoren (z.Z. ca. 500 verschiedene), z.B.:
  - Algorithmen zum maschinellen Lernen
  - Datenvorverarbeitungsoperatoren
  - Meta-Operatoren
  - Operatoren zur Visualisierung
  - Operatoren zum Im- und Export
  - ...
- RapidMiner nutzt XML um Operatorbäume darzustellen, die den KDD-Prozess modellieren

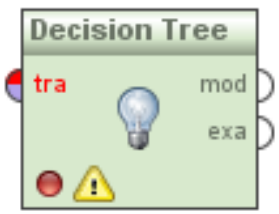
XML-Operatorbaum

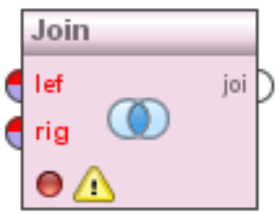
```
1  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
2  <process version="5.0">
3    <!-- [...] -->
4    <operator activated="true" class="process" expanded="true" name="Root">
5      <process expanded="true" height="541" width="675">
6        <operator activated="true" class="retrieve" expanded="true" height="60
               " name="Retrieve" width="90" x="45" y="30">
7          <parameter key="repository_entry" value="../../data/Golf" />
8        </operator>
9        <operator activated="true" class="decision_tree" expanded="true"
               height="76" name="DecisionTree" width="90" x="180" y="30" />
10       <connect from_op="Retrieve" from_port="output" to_op="DecisionTree"
               to_port="training set" />
11       <connect from_op="DecisionTree" from_port="model" to_port="result 1" />
12       <!-- [...] -->
13     </process>
14   </operator>
15 </process>
```



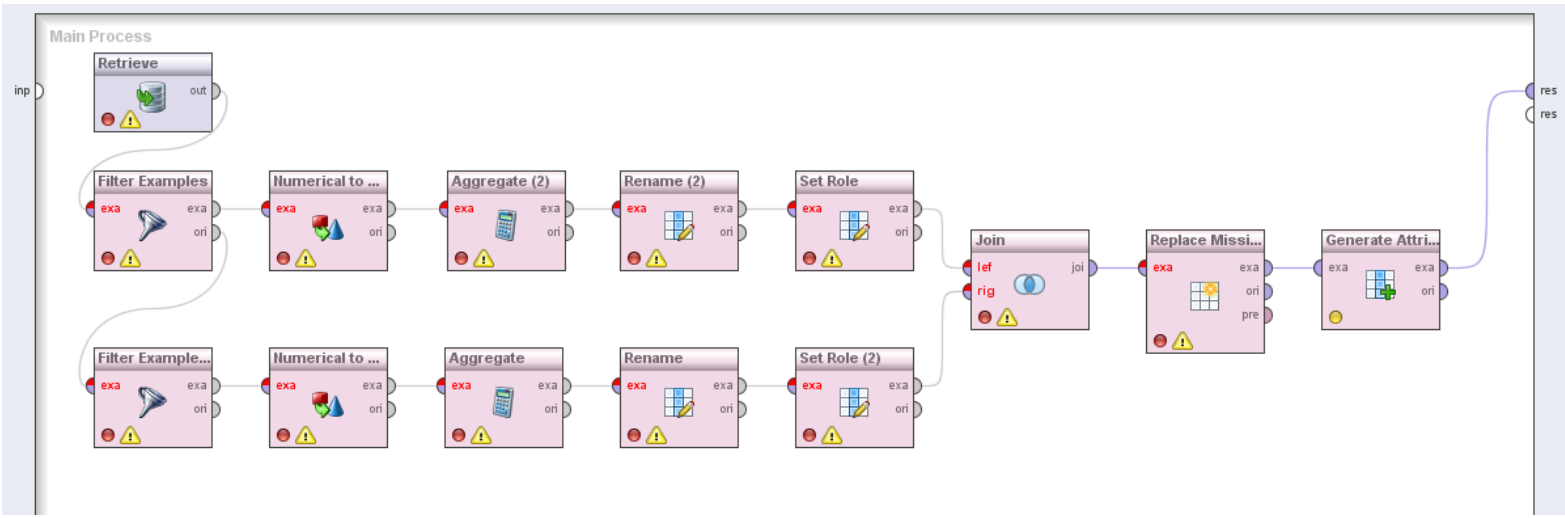
Beispiel-Operatoren







Beispiel-Operatorkette



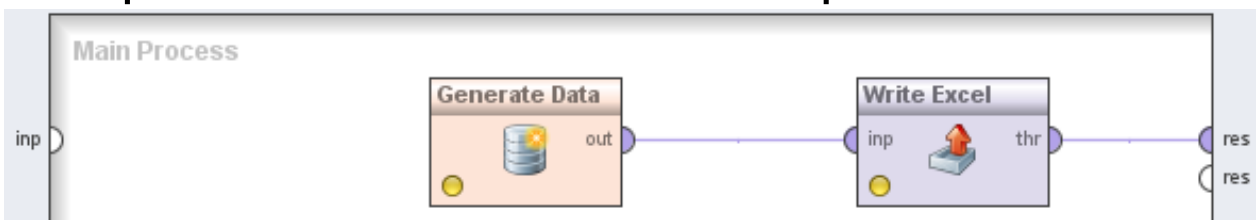
Überblick

Schnittstellen

- 3 Möglichkeiten zur Bedienung:
  - Server Modus (Kommandozeile)
  - über Java API aus externen Programmen
  - GUI Modus

Server Modus

- Voraussetzungen:
  - Umgebungsvariable „RAPIDMINER_HOME“ auf Verzeichnis der Installation setzen
  - optional: PATH-Variable zu „rapidminer.bat“ setzen
- Beispiel: Datei TestProcess.rmp



- Aufruf allg.: `rapidminer [-f] Prozessname`
- am Beispiel: `rapidminer -f TestProcess.rmp`

Einbindung in externes Programm

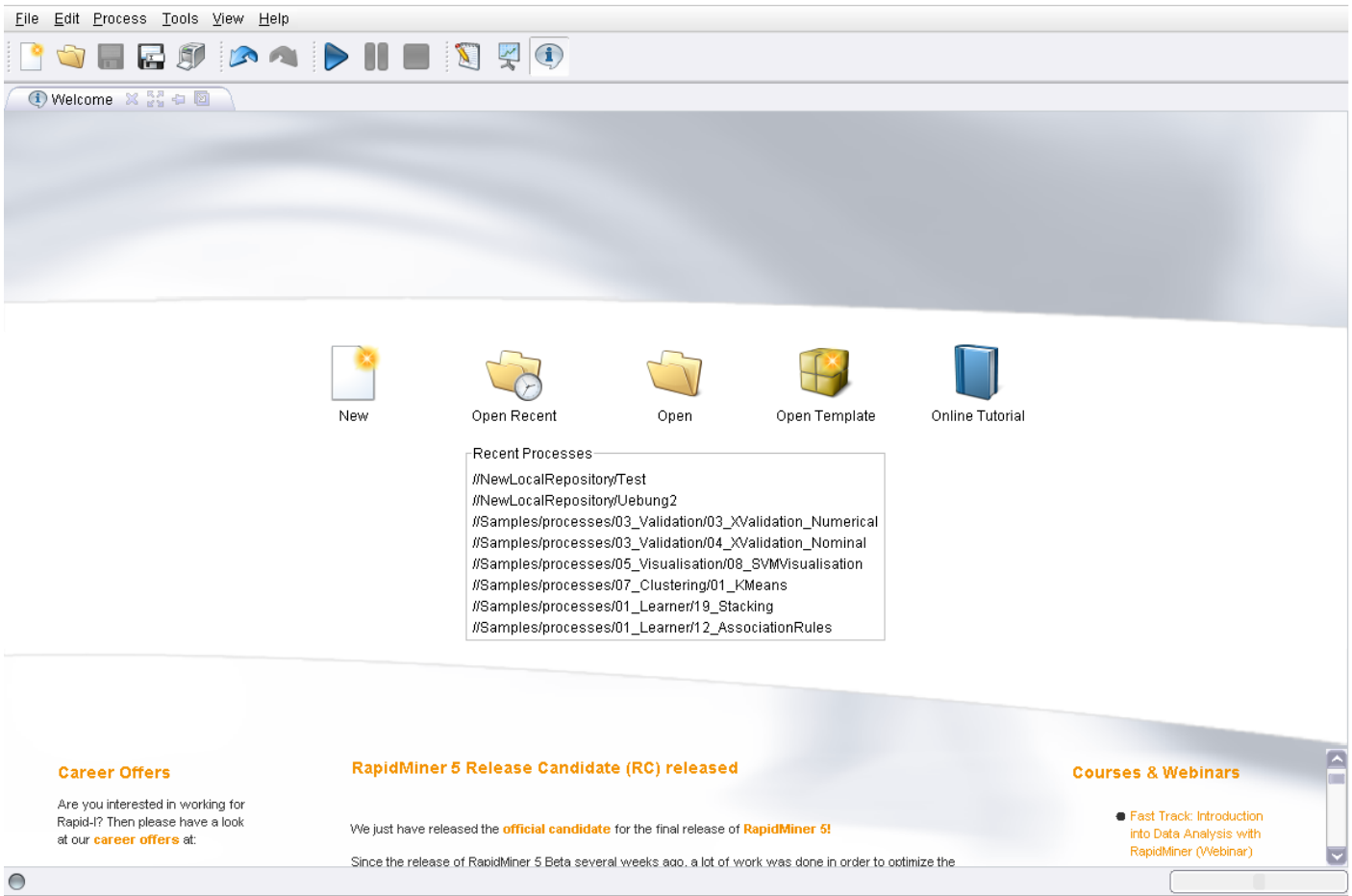
```
import com.rapidminer.Process;
import com.rapidminer.RapidMiner;
import com.rapidminer.operator.Operator;
import com.rapidminer.operator.OperatorException;
import com.rapidminer.operator.generator.ExampleSetGenerator;
import com.rapidminer.tools.OperatorService;

public class ProcessCreator {

  public static void main(String[] argv) {
    Process process = createProcess();
    System.out.println(process.getRootOperator().createProcessTree(0));

    try {
      process.run();
    } catch (OperatorException e) {
    e.printStackTrace();
    }
  }

// [...]
```

Einbindung in externes Programm

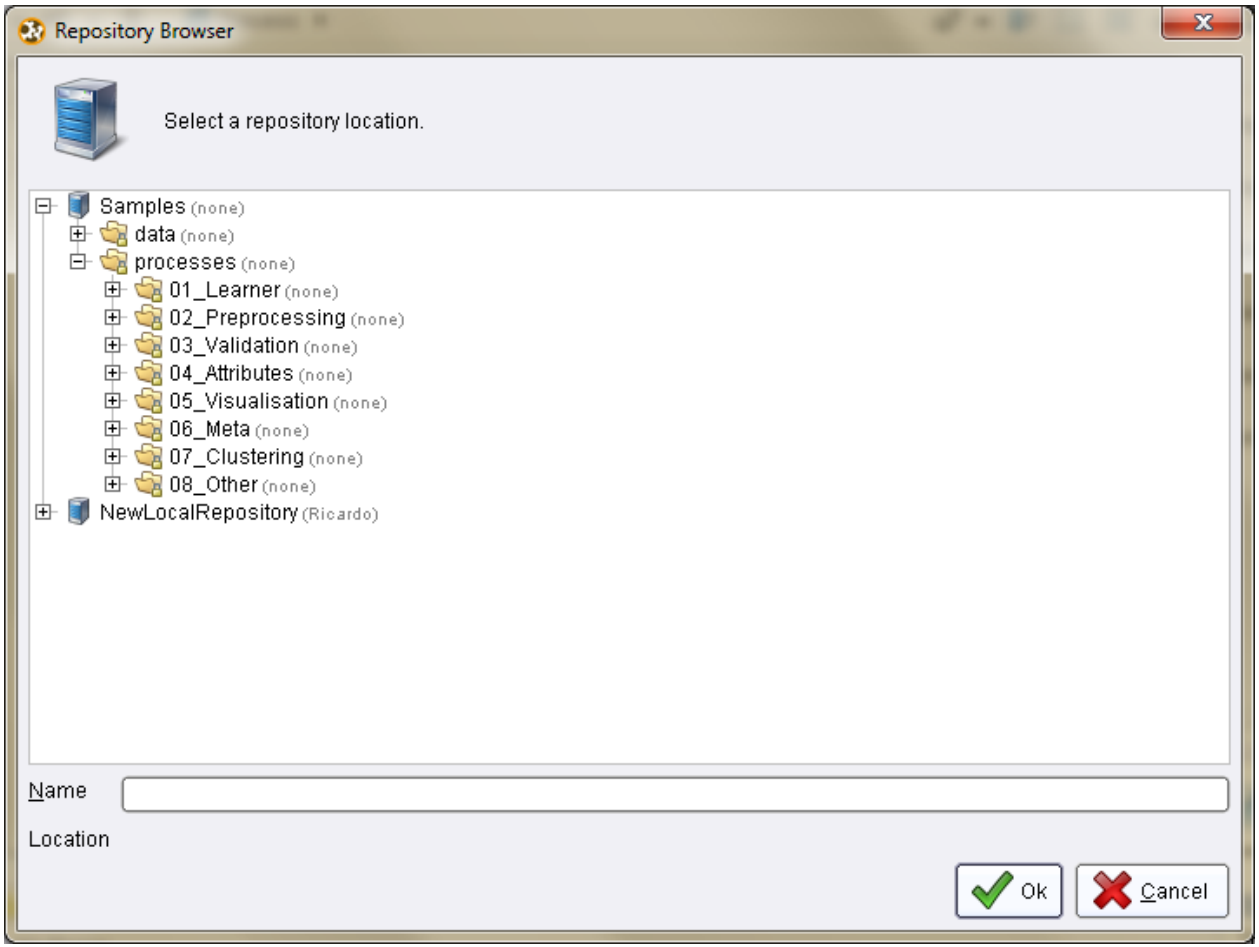
```
// [...]

  public static Process createProcess() {
    RapidMiner.init();
    Process process = new Process();
    try {
      Operator inputOperator = OperatorService.createOperator(
            ExampleSetGenerator.class);
      inputOperator.setParameter("target_function", "sum_classification");
      process.getRootOperator().getSubprocess(0).addOperator(inputOperator);
    } catch (Exception e) { e.printStackTrace(); }
      return process;
  }
  /* Ausgabe:
   *
   * Process[0] (Process)
   *   subprocess 'Main Process'
   *     +- Generate Data[0] (Generate Data)
   */
}
```

(Quelle: nach [RI09])

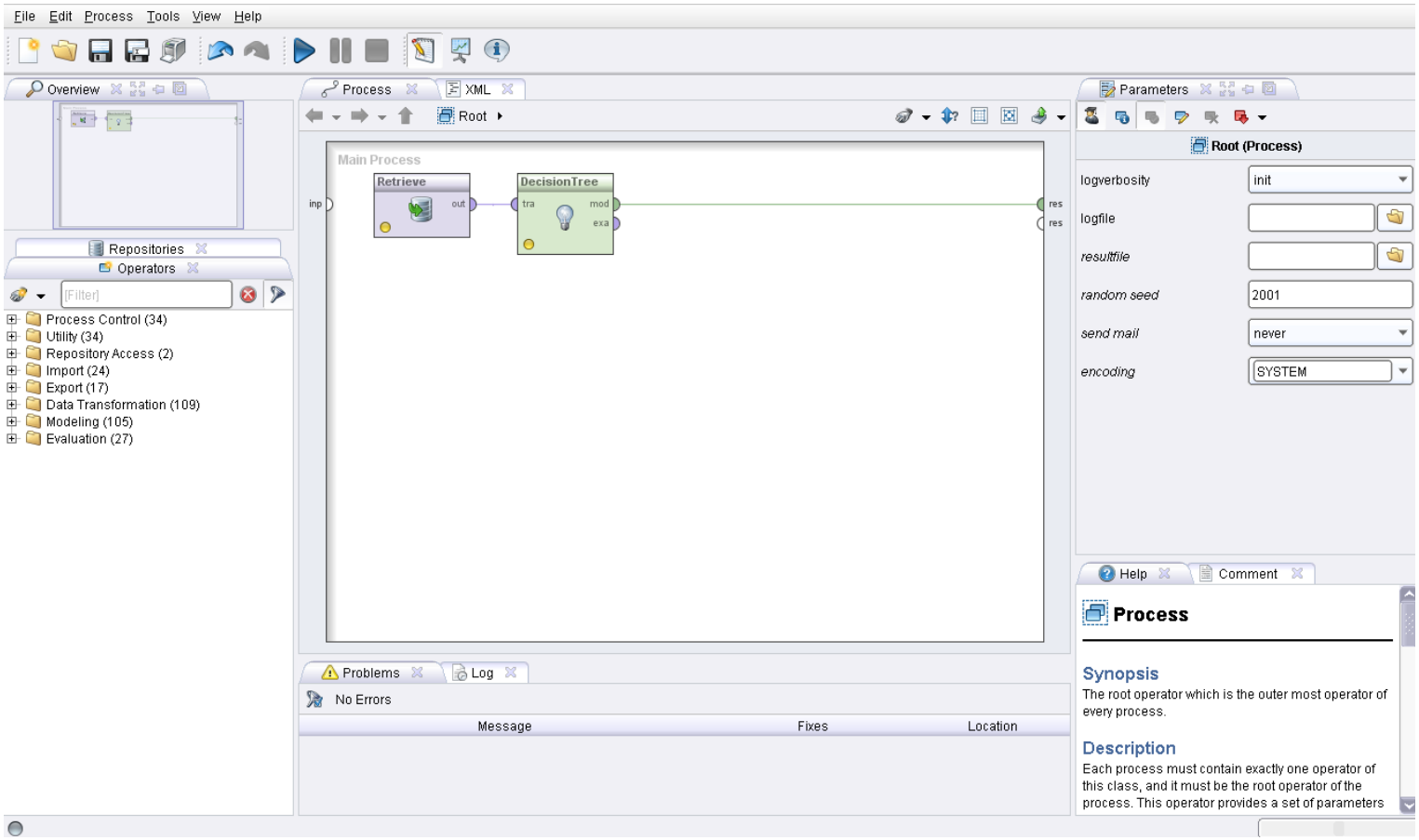
Oberfläche - Start



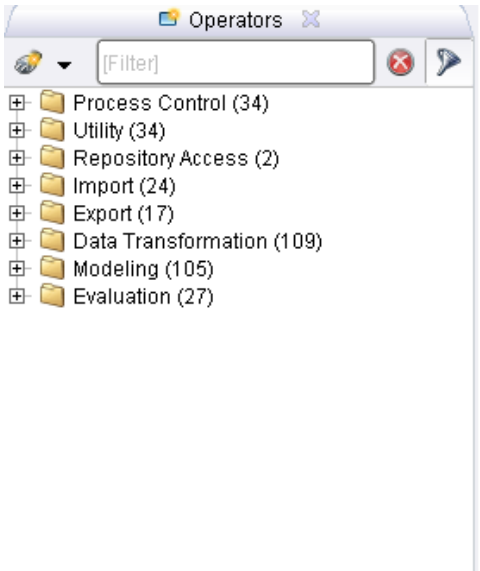
Oberfläche - Neuer Prozess

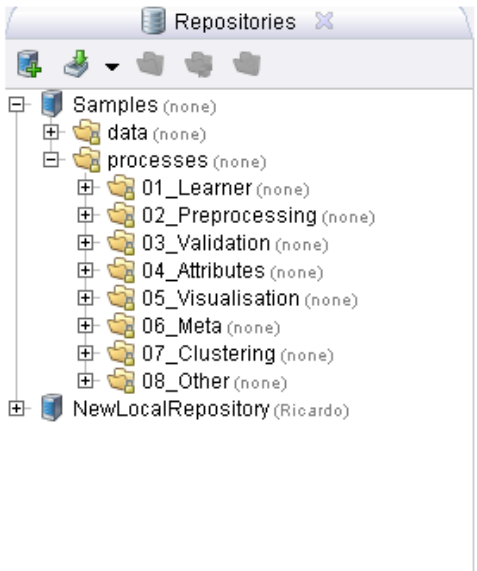


Oberfläche - Design Workspace

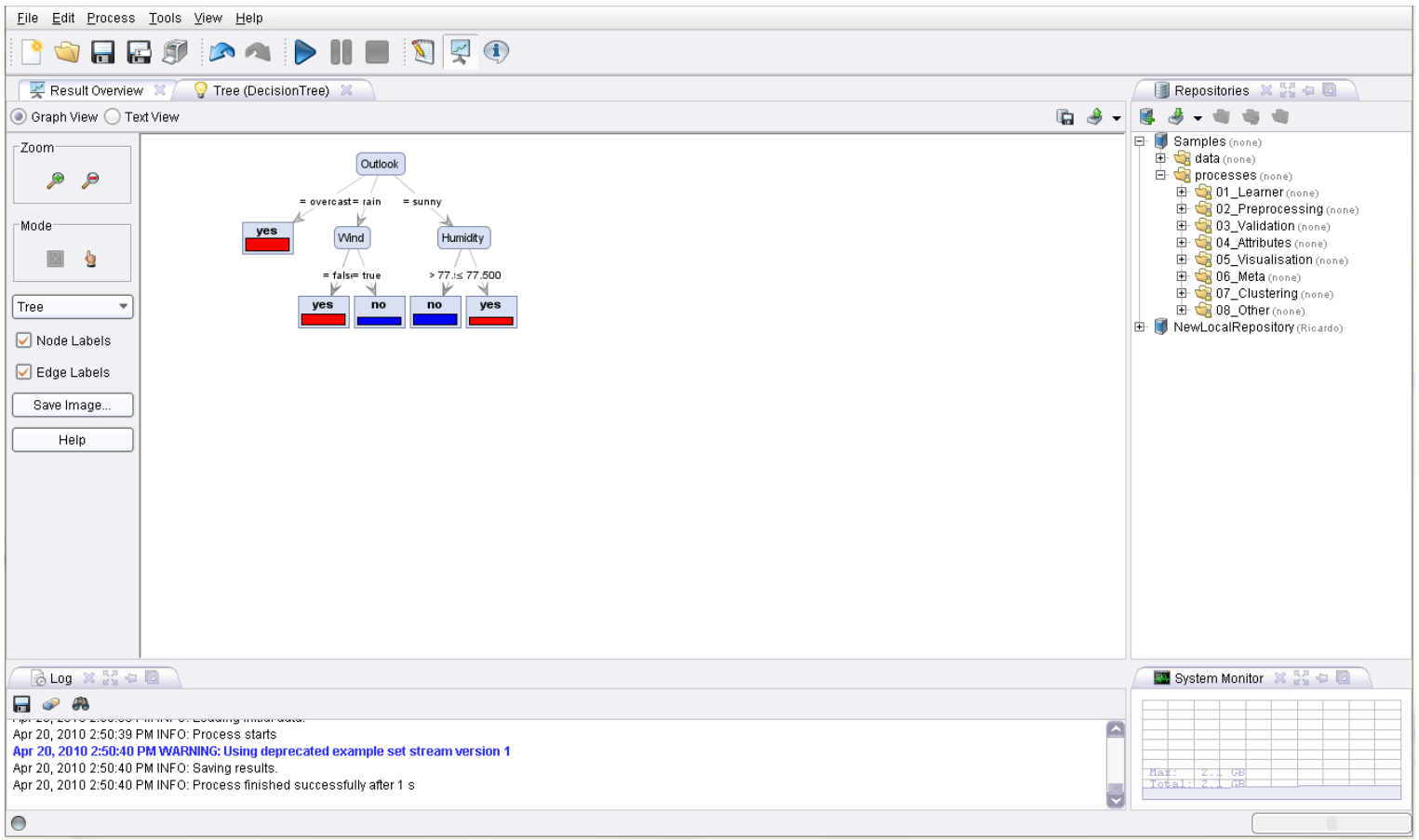


Oberfläche - Operatoren und Repositories





Oberfläche - Result Workspace



Visualisierung

- es bestehen 3 Möglichkeiten der Visualisierung von Ergebnissen
  - Meta-Daten-Sicht (*Meta Data View*)
  - Daten-Sicht (*Data View*)
  - grafische Darstellungs-Sicht (*Plot View*)
- bei der grafischen Darstellung besteht die Möglichkeit diverse Visualisierung mittels 2D- und 3D-Grafiken zu erzeugen

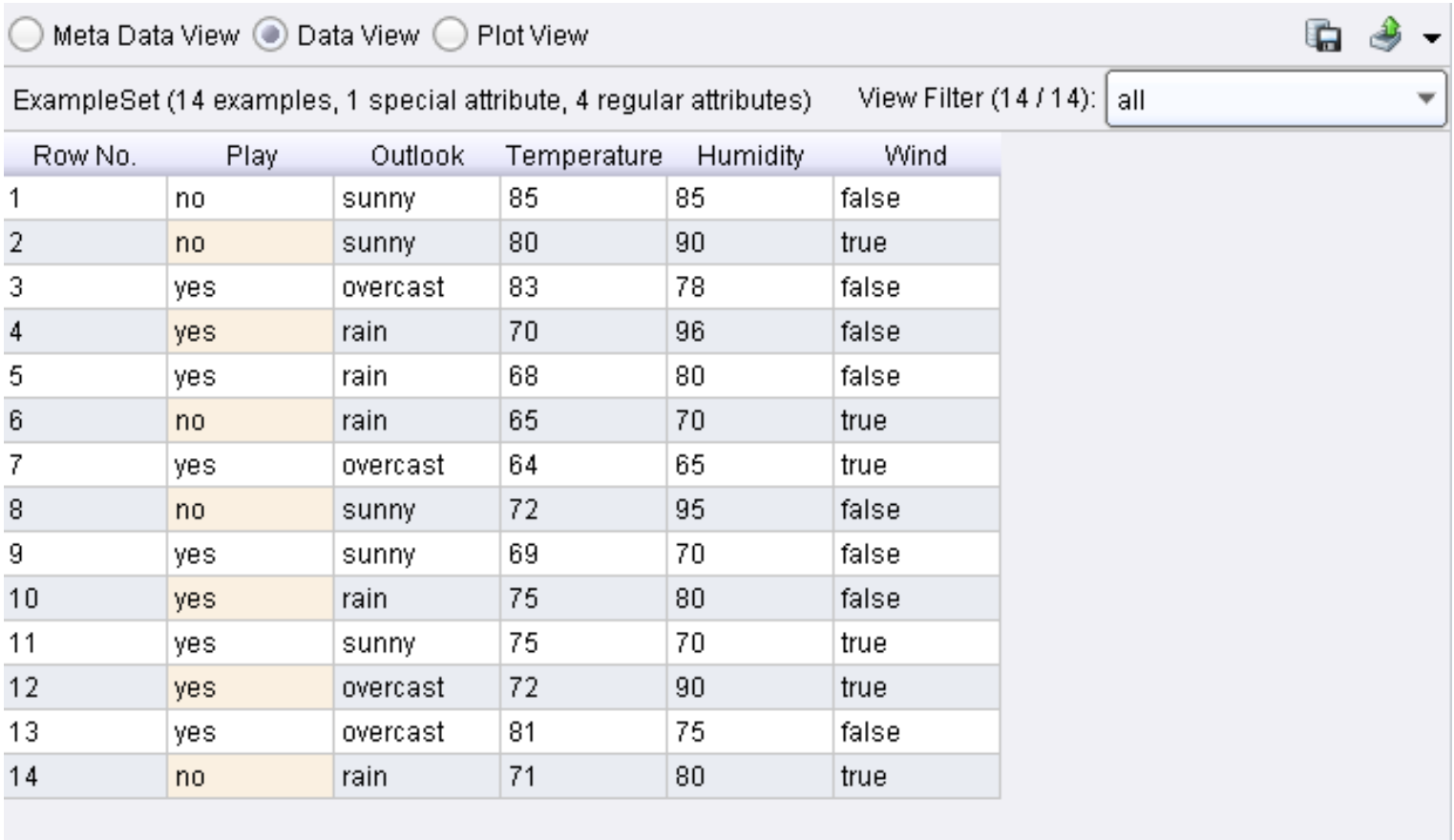
Beispiel - Meta Data View

Meta Data View | Data View | Plot View

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	Play	nominal	mode = yes (9), least = no (5)	no (5), yes (9)	0
regular	Outlook	nominal	mode = sunny (5), least = overcast (4)	rain (5), overcast (4), sunny (5)	0
regular	Temperature	integer	avg = 73.571 +/- 6.572	[64.000 ; 85.000]	0
regular	Humidity	integer	avg = 80.286 +/- 9.840	[65.000 ; 96.000]	0
regular	Wind	nominal	mode = false (8), least = true (6)	true (6), false (8)	0

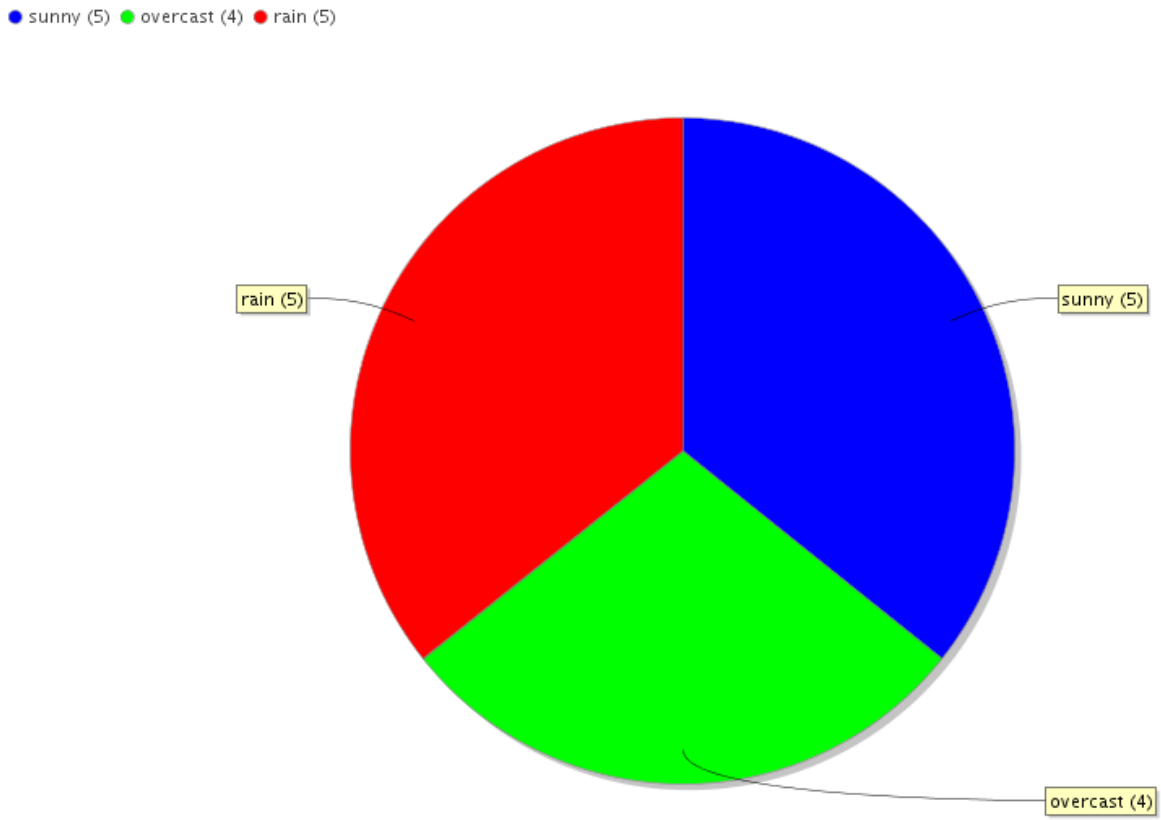
Beispiel - Data View


Meta Data View | Data View | Plot View

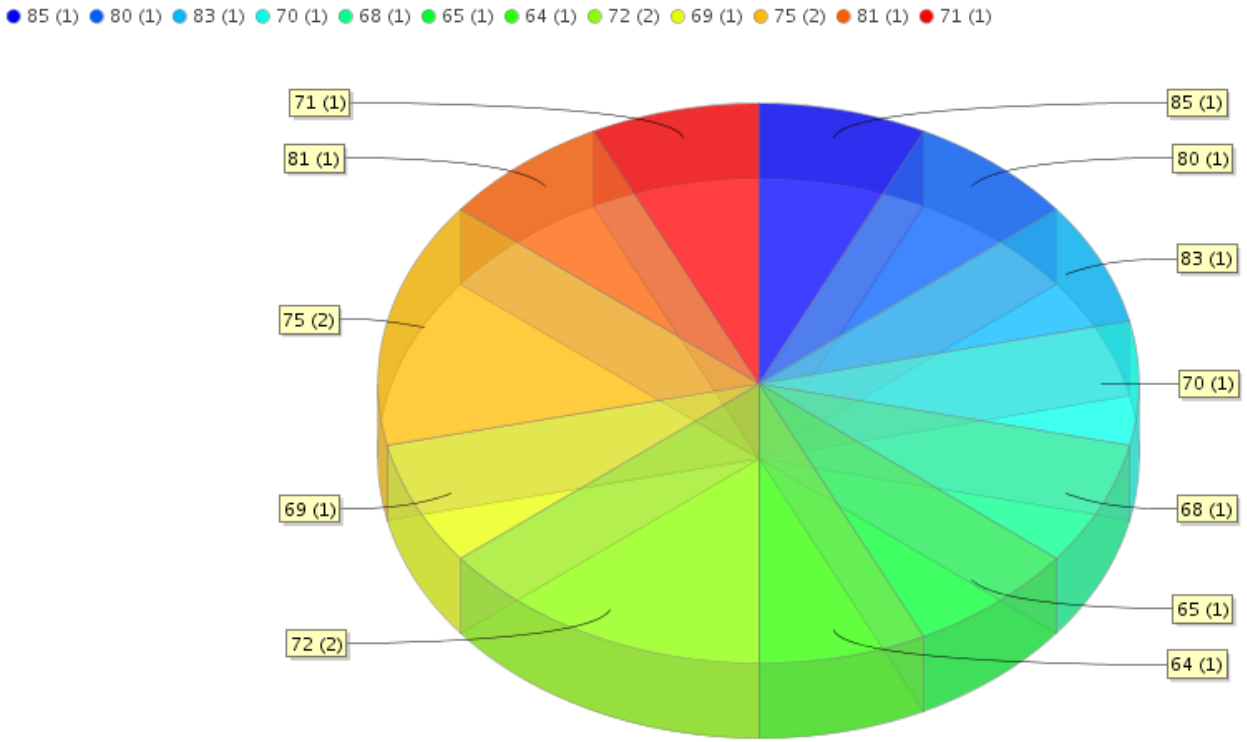
ExampleSet (14 examples, 1 special attribute, 4 regular attributes) View Filter (14 / 14): all

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

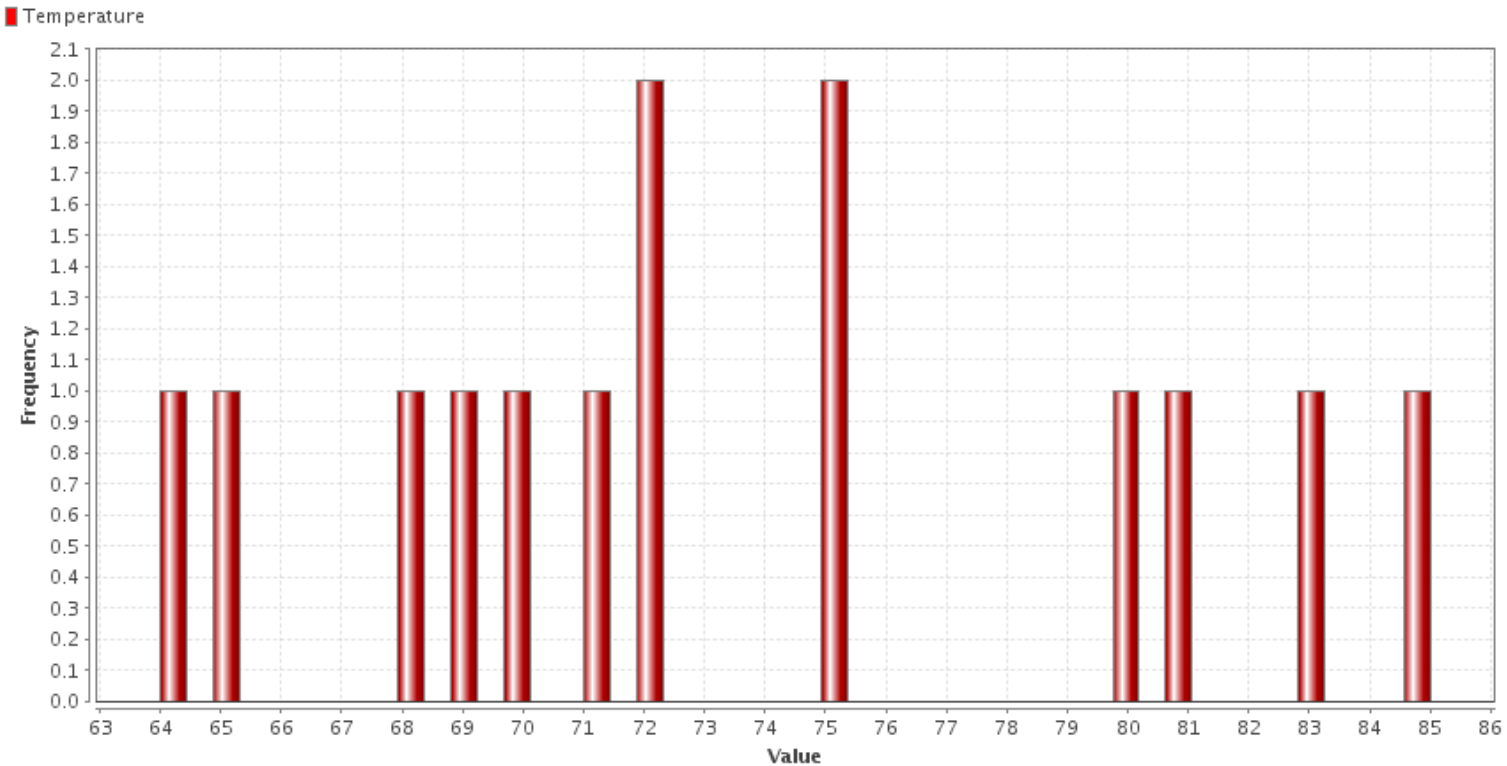
Beispiel - Plot View (Pie)



Beispiel - Plot View (Pie 3D)



Beispiel - Plot View (Histogramm)



Erweiterungsmöglichkeiten

- RapidMiner bietet die Möglichkeit über Erweiterungen den Funktionsumfang zu vergrößern
- Beispiele für Erweiterungen sind:
  - Weka Extension
  - Parallel Processing
  - Text Processing
  - Web Mining
  - Reporting Extension
  - Series Processing
  - PMML

Datenformate

Funktion	Formate
Import	CSV, Excel, Access, BibTeX, Database, DBase, URL, SPSS, AML, ARFF, XRFF, Stata, Sparse, C4.5, DasyLab
Export	CSV, Excel, Access, AML, ARFF, XRFF, Database

Überblick

Wiederholung KDD-Prozess

- KDD = Knowledge Discovery in Databases
- Schritte:
  1. Datenselektion und - extraktion
  2. Datenbereiningung und -transformation
  3. Data Mining
  4. Interpretation
- Umsetzung in RapidMiner als Operatoren-Kette

Funktionsweise im RapidMiner

- Austausch von IOObjects zwischen Operatoren
- Datenmenge als ExampleSet bezeichnet
  - entspricht Tabelle
  - Examples sind die Zeilen
  - Attribute sind die Spalten

Attribute

Rollen

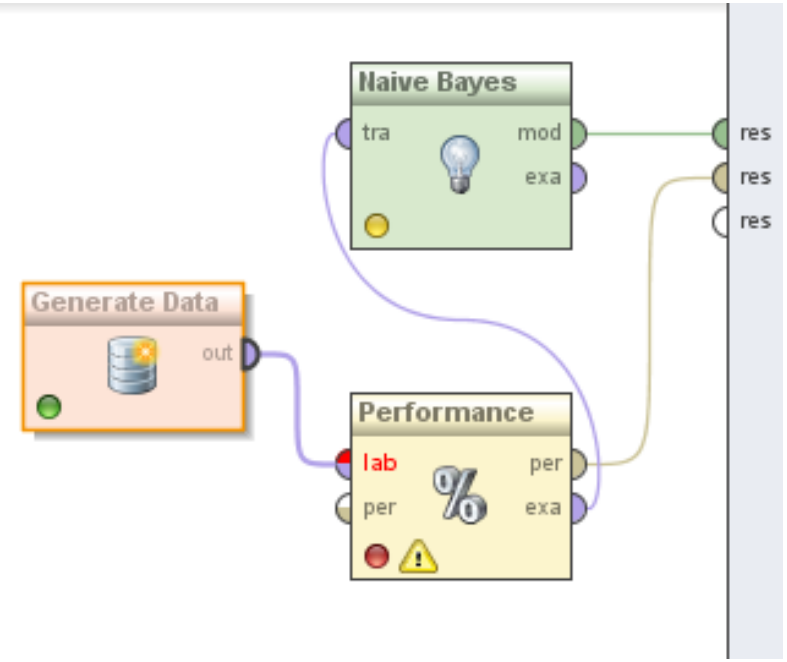
- regular attributes
- special attributes
  - ID
  - Label
  - Prediction
  - Cluster
  - Weight
  - Batch

Typen

- (bi-/poly-)nominal
- numeric
- date
- text

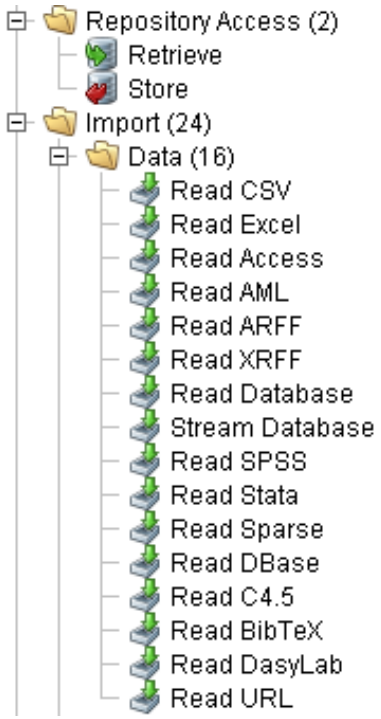
Farbliche Markierung in GUI

- Violett = ExampleSet
- Grün = Model
- Braun = PerformanceVector
- Pink = Merkmalsgewicht

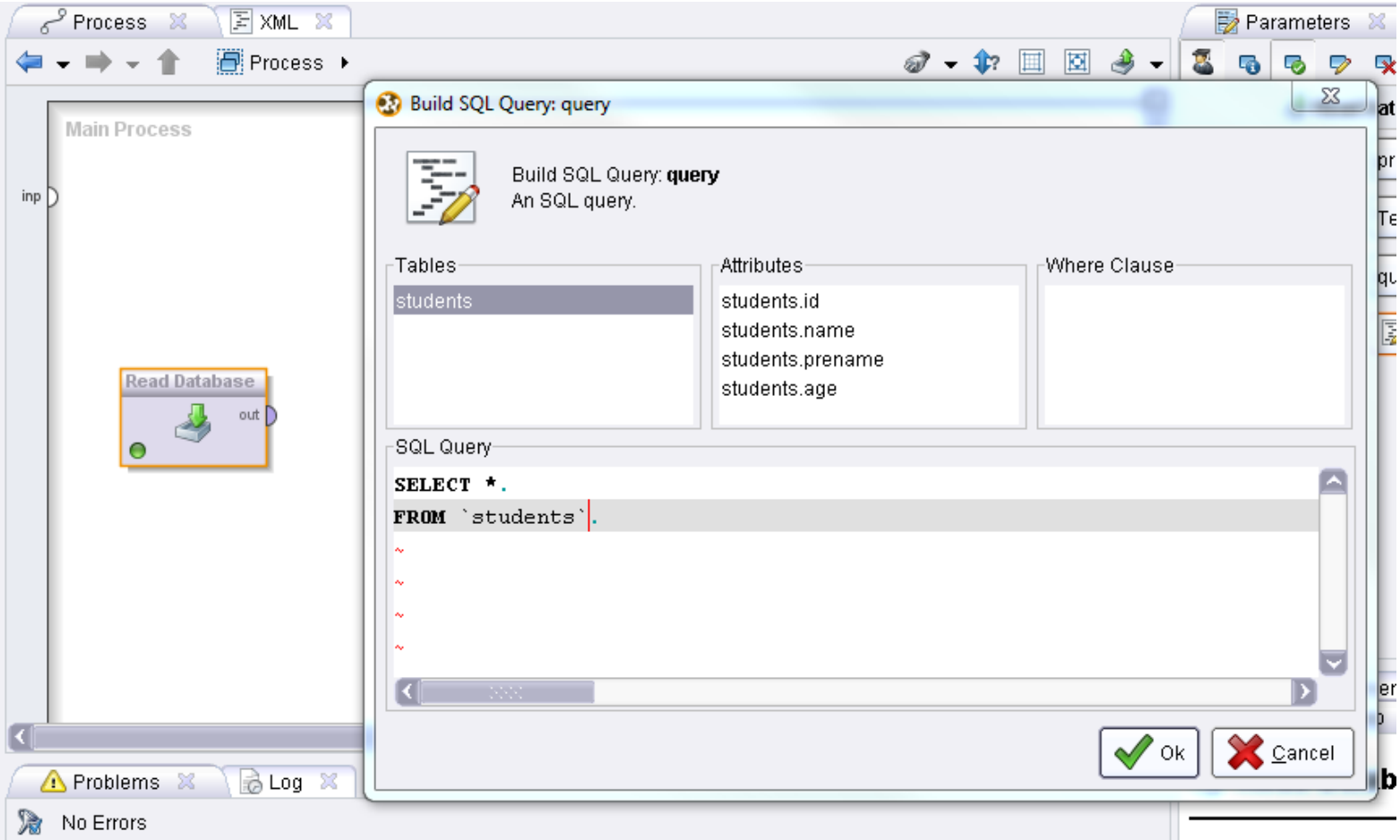


Datenselektion und -extraktion

- Import aus
  - Repository
  - verschiedenen Dateiformaten
- Generierung von Daten
- Ausgabe als ExampleSet
- Speicherung im Repository möglich

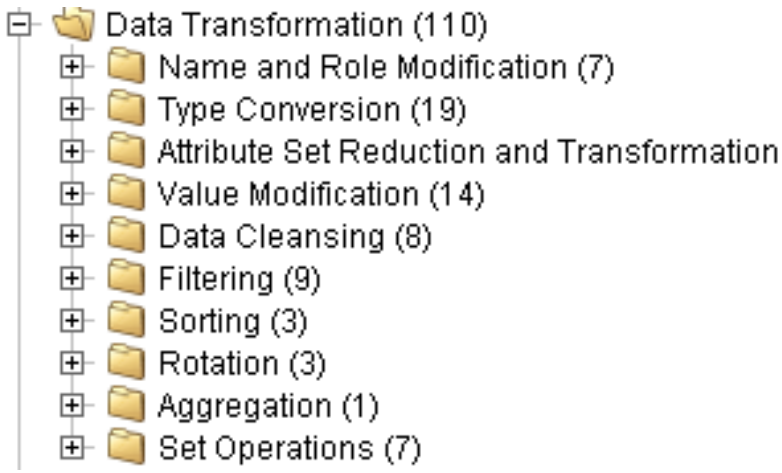


Lesen aus Datenbank



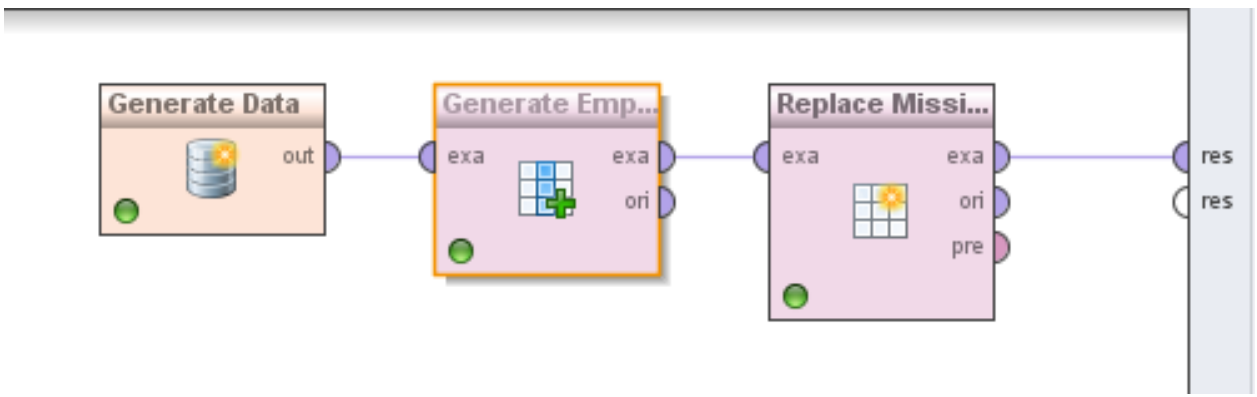
Datenbereiningung und -transformation

- Umbennung, Rollenzuweisung
- Typumwandlung
- Transformation von Attributen
- Wertmodifikation
- Datenbereinigung
- Filterung
- Sortierung
- Rotation
- Aggregation
- Operatoren (z.B. Joins)



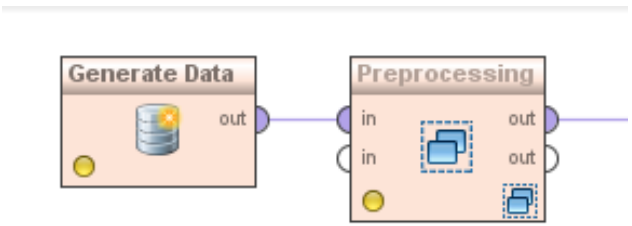
Beispiel Data Cleansing

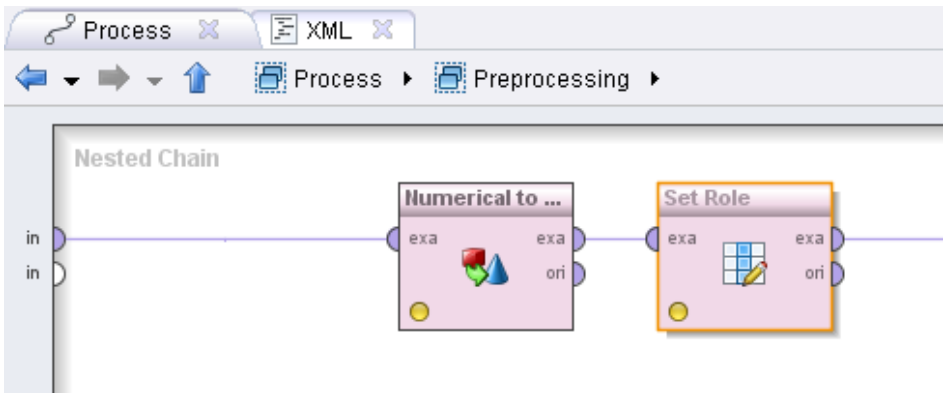
- Operator: Replace Missing Values
- Ersetzung fehlender Werte durch
  - Minimum
  - Maximum
  - Durchschnitt
  - Null
  - Wert



Vorverarbeitung als Subprozess

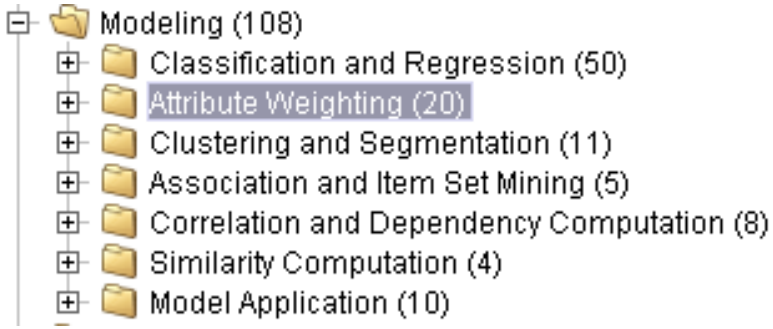
- Vorverarbeitungsschritte als Subprozess gekapselt
  ➔ bessere Übersicht
- Utility/Subprocess



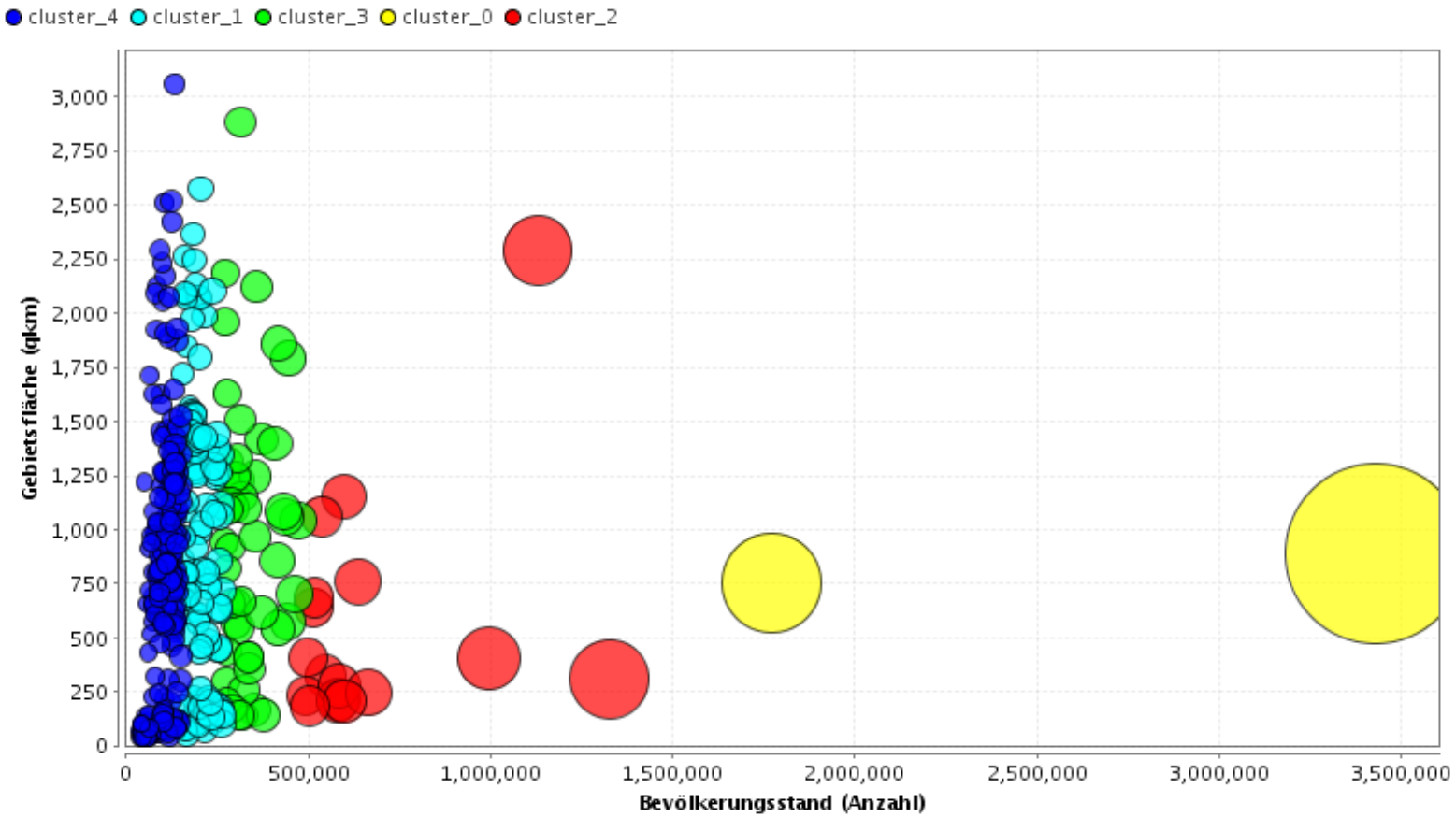


Data Mining

- Klassifikation
- Attributgewichtung
- Clustering
- Assoziationsanalyse
- Korrelation
- Ähnlichkeitsberechnung



Beispiel Clustering



Interpretation

- visuelle Darstellung in verschiedenen Graphen- und Diagrammtypen
- Bewertung durch Benutzer
  - gefundene Muster beurteilen
  - Aussagekraft des Ergebnisses
- evtl. erneutes Data Mining

Produktpräsentation

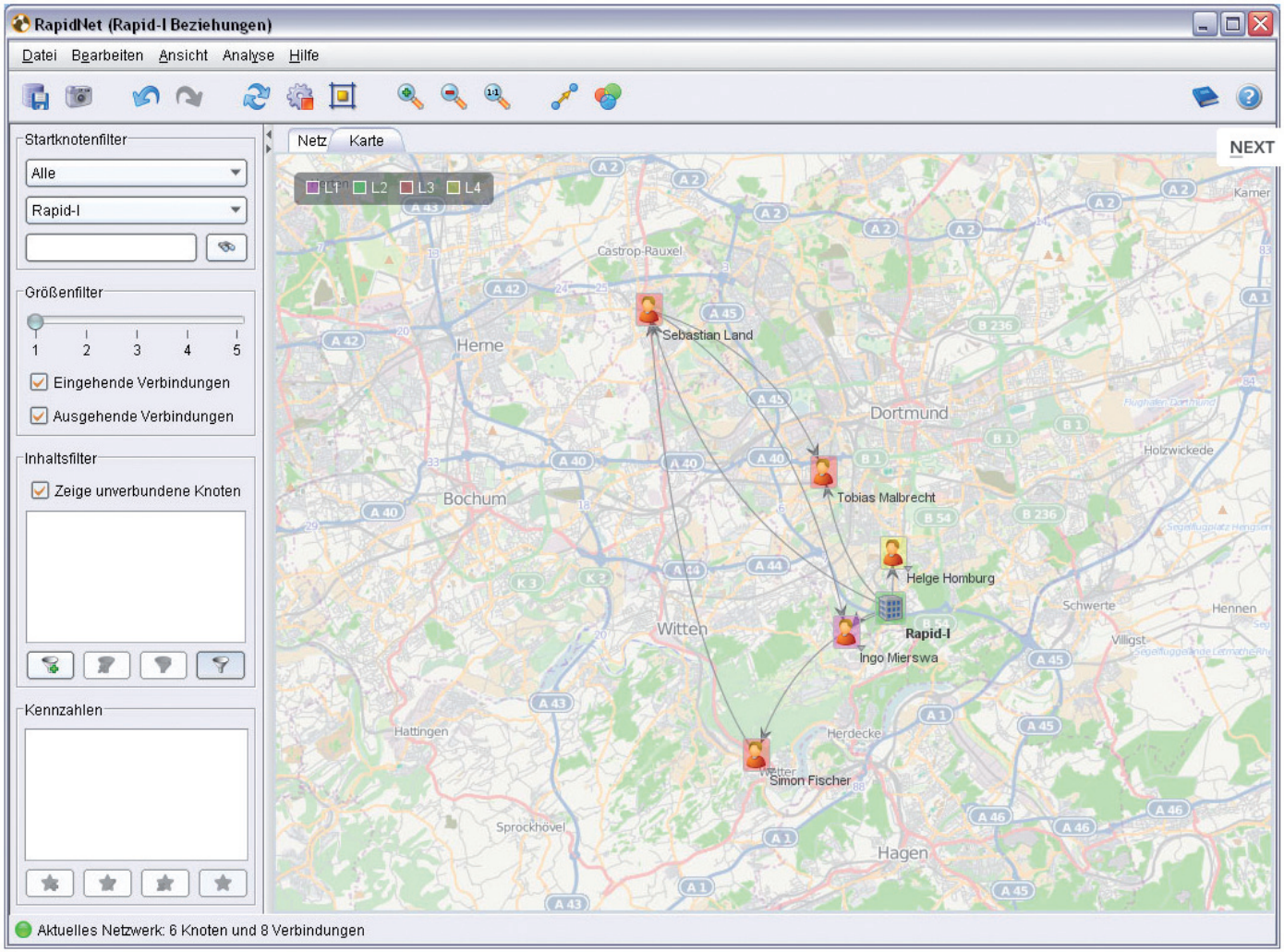
1. Allgemeine Übersicht
2. Warenkorbanalyse (FP-Growth)
3. Clustering nach Einwohnerzahlen
4. Text Mining

Überblick

RapidNet

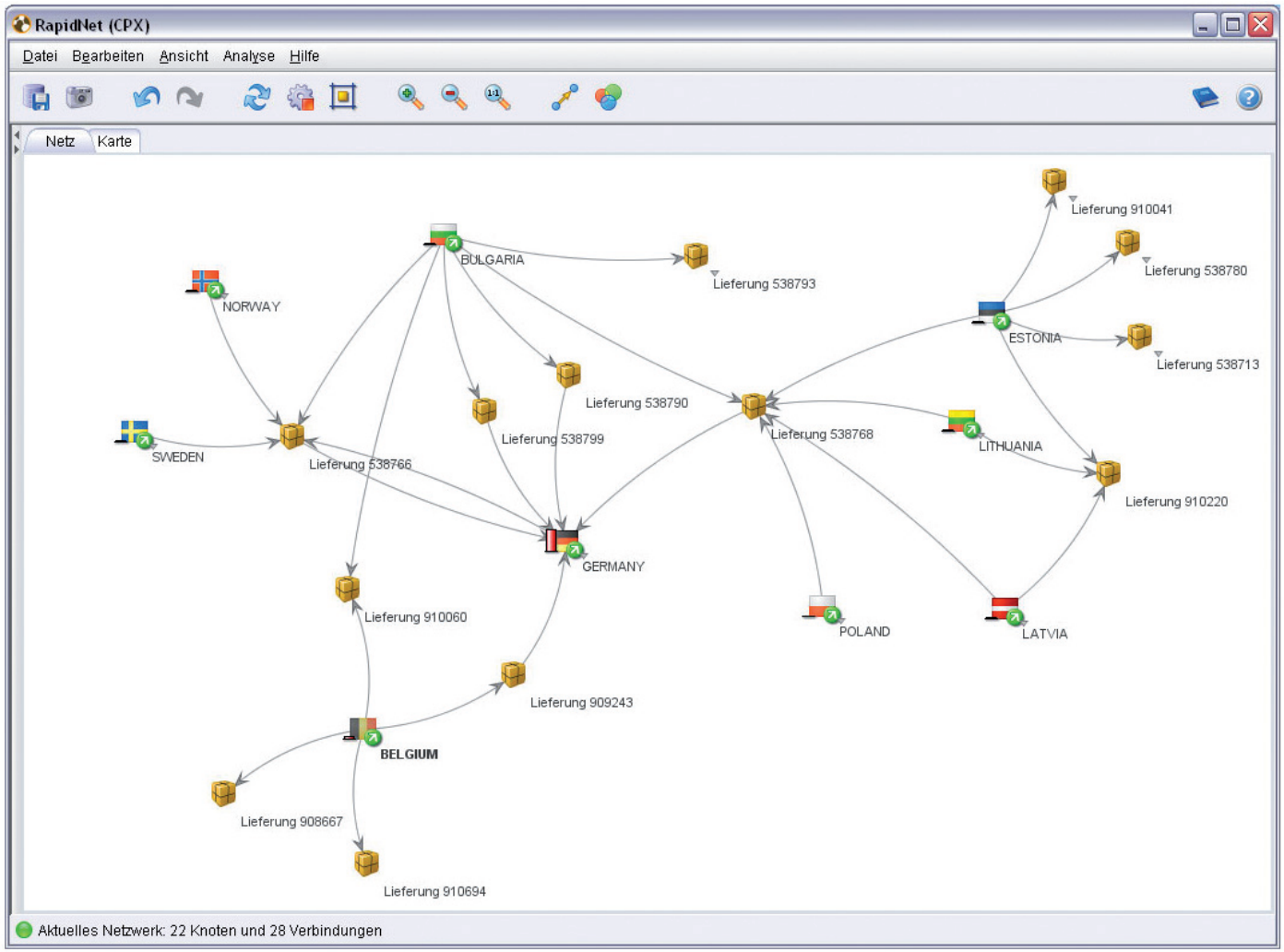
- Allgemein: Struktur- und Relations-Explorer
- zahlreiche Möglichkeiten zur Visualisierung
- basiert auf Funktionen des RapidMiner
- einsatzfähig auf allen gängigen Plattformen
- Möglichkeiten
  - Strukturelle Clusteranalysen
  - Darstellung von hierarchischen Relationen
  - Visualisierung von geographischen Informationen auf Karten
  - ...

Weitere Werkzeuge von Rapid-I ➤ RapidNet



(Quelle: [RI10b])

Weitere Werkzeuge von Rapid-I ➤ RapidNet

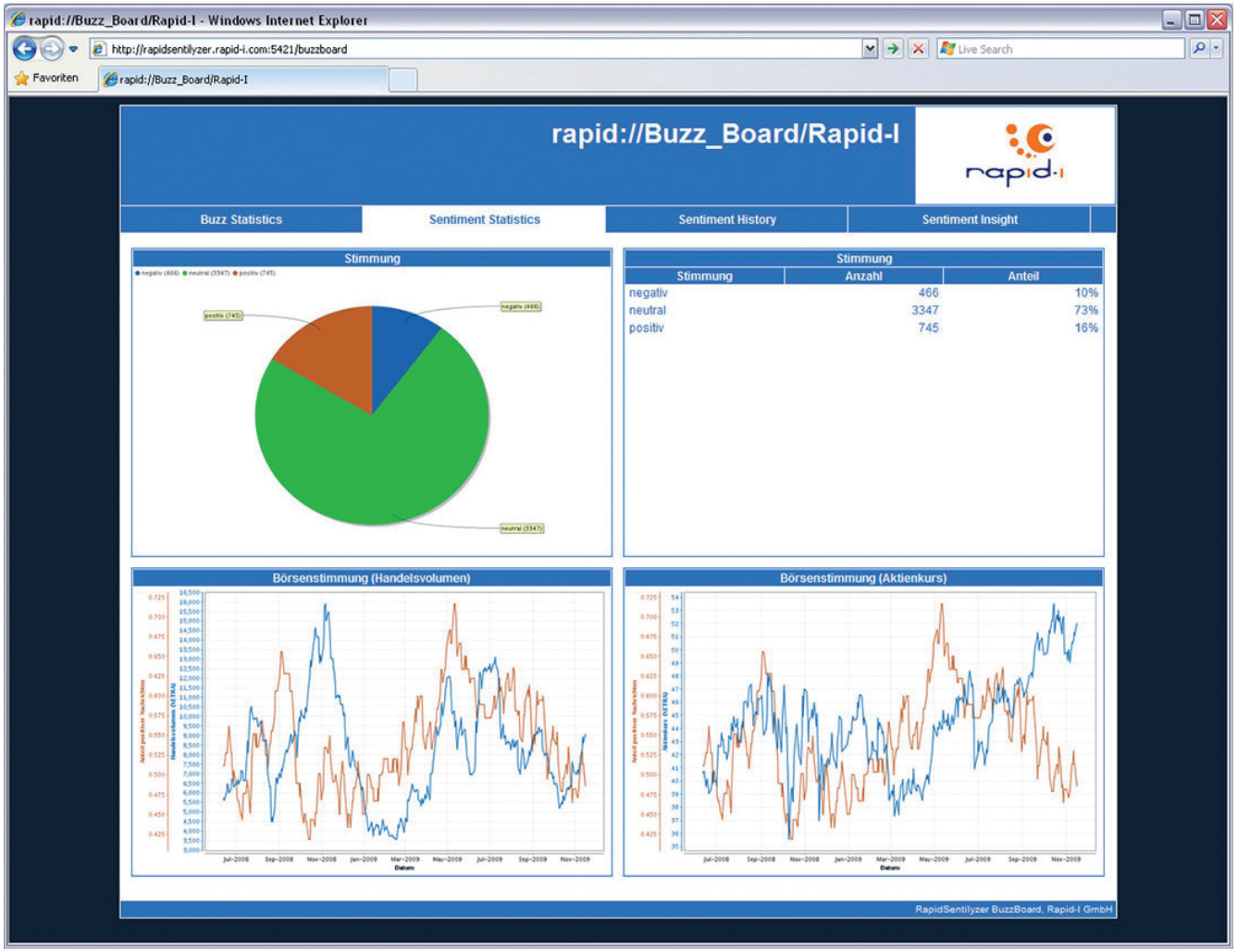


(Quelle: [RI10b])

RapidSentilyzer

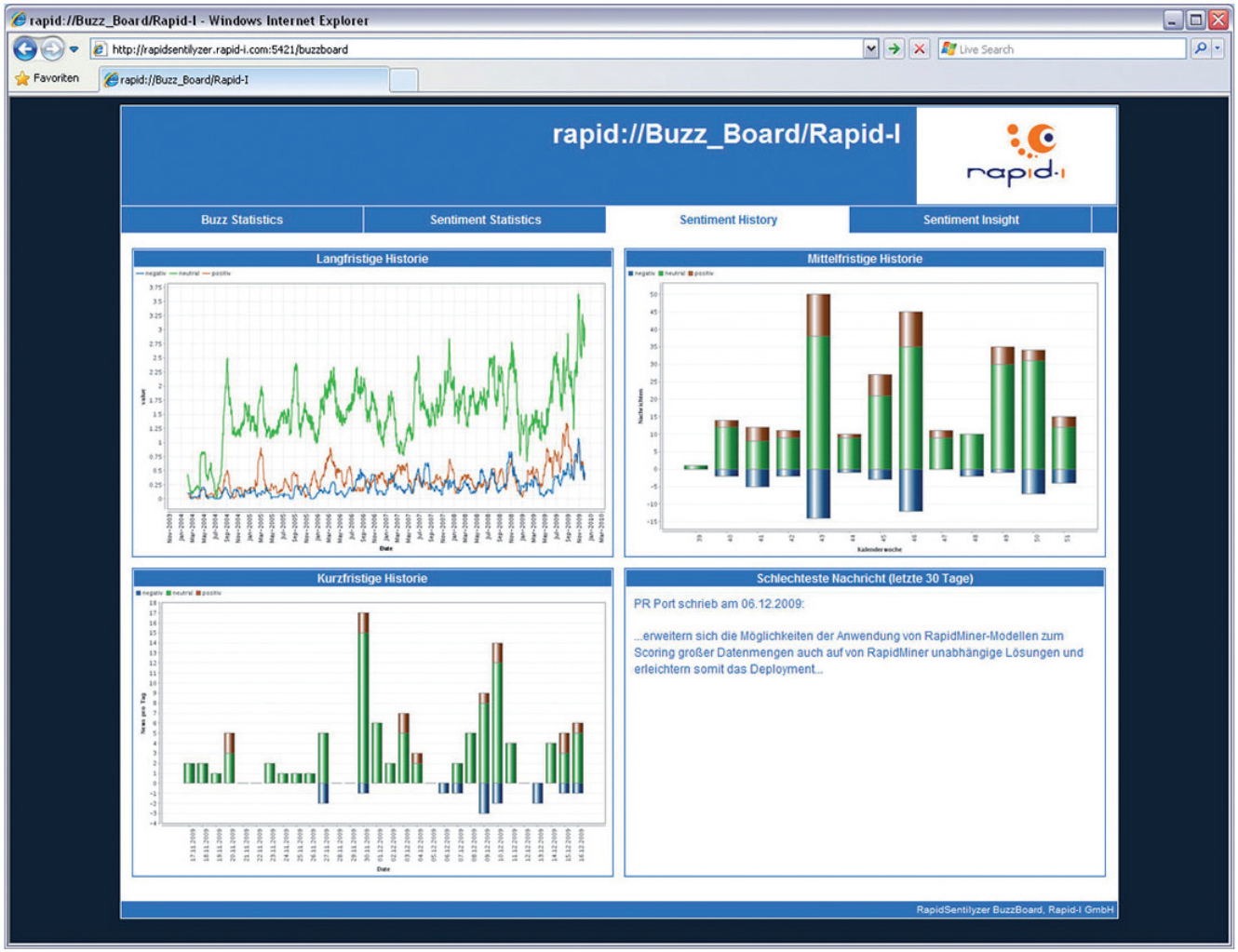
- dient zur automatischen Sammlung von Informationen
- Verwendung von Crawling-Techniken in Kombination mit Data- und Text Mining
- basiert auf Funktionen des RapidMiner
- zentrale Zusammenfassung der Informationen im sogenannten „RapidSentilyzer BuzzBoard“, bestehend aus:
  - Buzz Statistics
  - Sentiment Statistics
  - Sentiment History
  - Sentiment Insight

Weitere Werkzeuge von Rapid-I ➤ RapidSentilyzer



(Quelle: [RI10c])

Weitere Werkzeuge von Rapid-I ➤ RapidSentilyzer

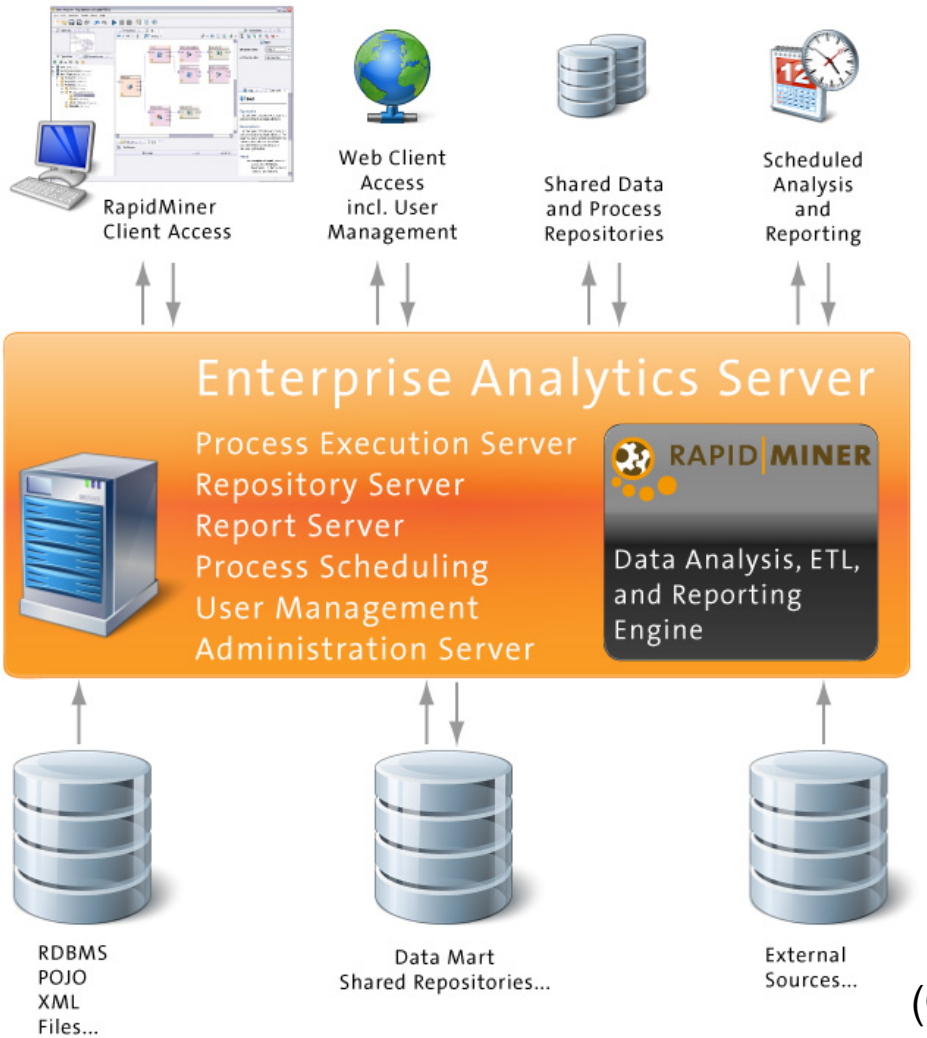


(Quelle: [RI10c])

RapidAnalytics

- Open Source Enterprise Analytics Server
- basierend auf RapidMiner
- Shared Repositories
- Remote und Scheduled Execution
- Zugriff über
  - RapidMiner Client Software
  - Web-Interface
  - Webservices

Weitere Werkzeuge von Rapid-I ➤ RapidAnalytics

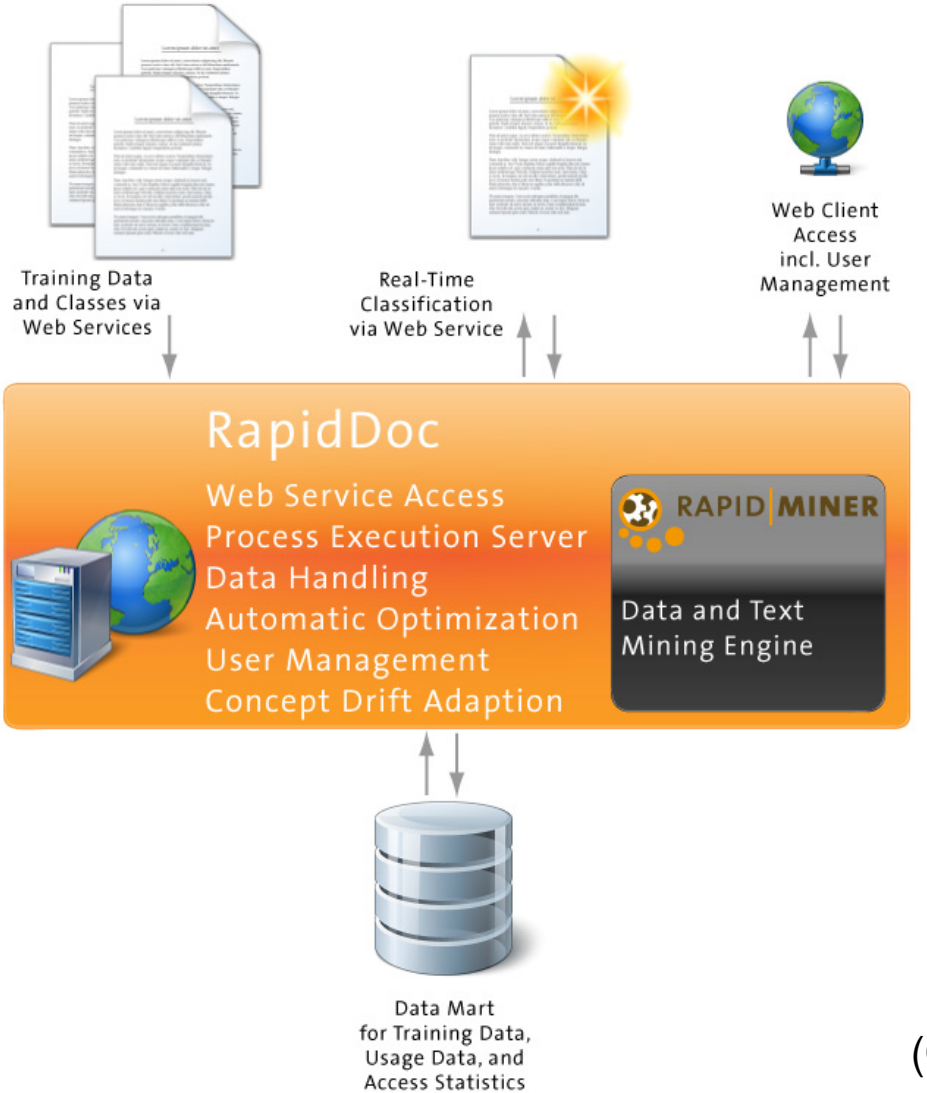


(Quelle: [Rla])

RapidDoc

- automatische Klassifikation von Texten
- Funktionsweise
  - Basis: Webservices
  - Trainingstexte und vordef. Klassen vorgeben
  - Einordnung neuer Text in wahrscheinlichste Klasse
  - Angabe der Sicherheit der Vorhersage
  - Nutzung der RapidMiner Engine
  - Optimierung durch Rapid-I Mitarbeiter

Weitere Werkzeuge von Rapid-I ➤ RapidDoc



(Quelle: [RIb])

Zusammenfassung

- mächtiges Open Source-Data Mining-Tool
- ermöglicht gesamten KDD-Prozess
- viele Operatoren bereits vorhanden
- flexibel einsetz- und erweiterbar
- zahlreiche Visualisierungsvariaten
- weitere Möglichkeiten durch zusätzliche Tools
- Einsatz z.B. bei
  - Allianz
  - Siemens
  - EADS
  - T-Mobile
  - PC-Ware

Quellen I

[RIa] RAPID-I: *Rapid Analytics*. http://rapid-i.com/component/option,com_docman/task,doc_download/gid,48/. – Zugriff: 22.04.2010

[RIb] RAPID-I: *Rapid Doc*. http://rapid-i.com/component/option,com_docman/task,doc_download/gid,49/. – Zugriff: 22.04.2010

[RIc] RAPID-I: *RapidMiner Benutzerhandbuch*. http://sourceforge.net/projects/yale/files/1.%20RapidMiner/5.0/rapidminer-5.0-manual-german.pdf/download. – Zugriff: 09.05.2010

[RI09] RAPID-I: *RapidMiner 4.4*. http://ignum.dl.sourceforge.net/project/yale/1.%20RapidMiner/4.4/rapidminer-4.4-tutorial.pdf. Version: März 2009. – Zugriff: 18.04.2010

Quellen II

[RI10a] RAPID-I: *Rapid - I - RapidMiner*. http://rapid-i.com/content/view/181/190/. Version: April 2010. – Zugriff: 20.04.2010

[RI10b] RAPID-I: *RapidNet*. http://rapid-i.com/component/option,com_docman/task,doc_download/gid,50/. Version: Februar 2010. – Zugriff: 23.04.2010

[RI10c] RAPID-I: *RapidSentilyzer*. http://rapid-i.com/component/option,com_docman/task,doc_download/gid,51/. Version: Februar 2010. – Zugriff: 23.04.2010

[TU] TU DORTMUND: *Data Mining mit RapidMiner*. http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS09/RapidMiner.pdf. – Zugriff: 22.04.2010

[Wik10] WIKIPEDIA: *RapidMiner*. http://de.wikipedia.org/wiki/RapidMiner. Version: April 2010. – Zugriff: 18.04.2010

Vielen Dank für die Aufmerksamkeit!