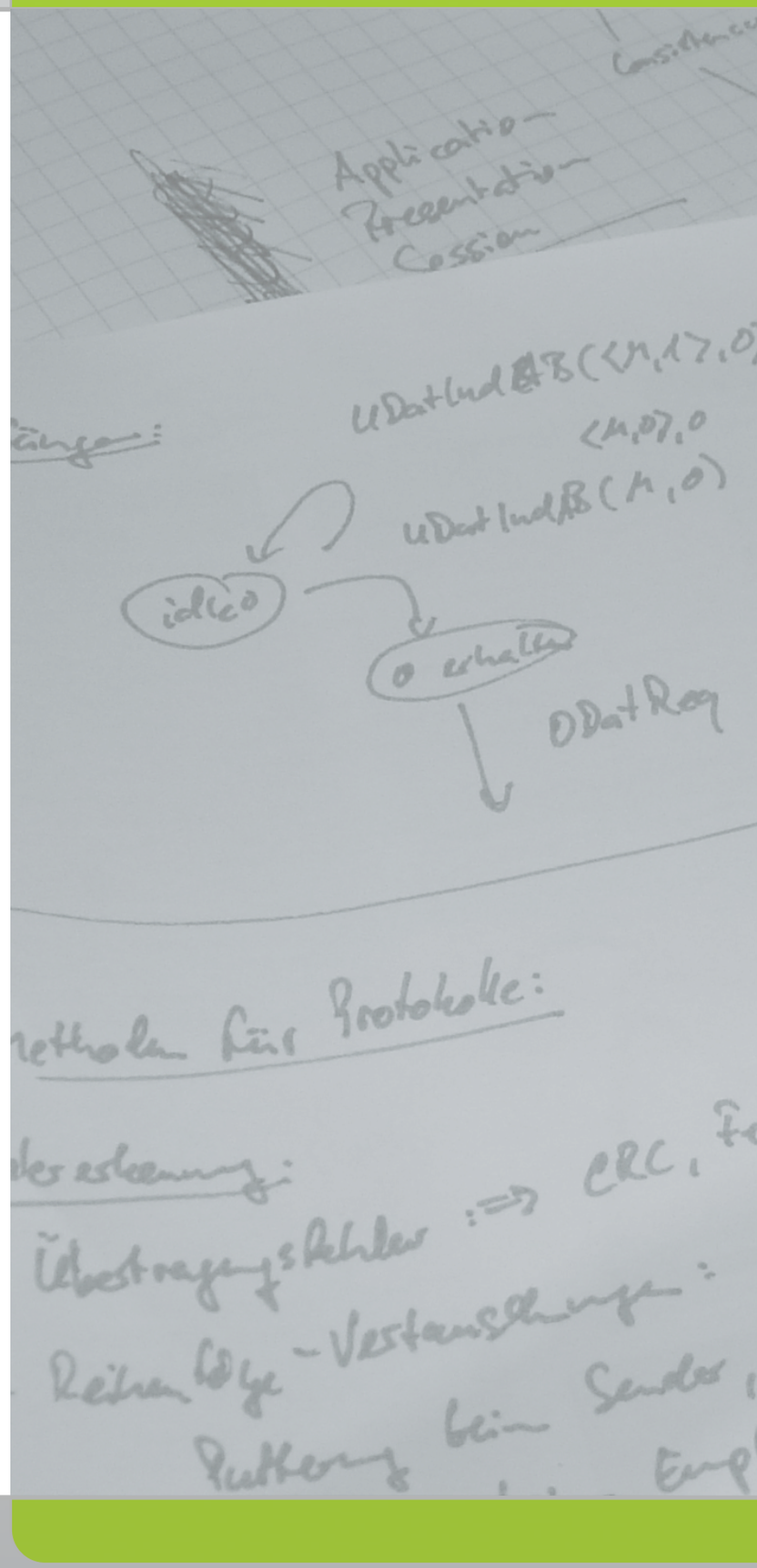
Data Mining mit RapidMiner

technische universität dortmund

Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

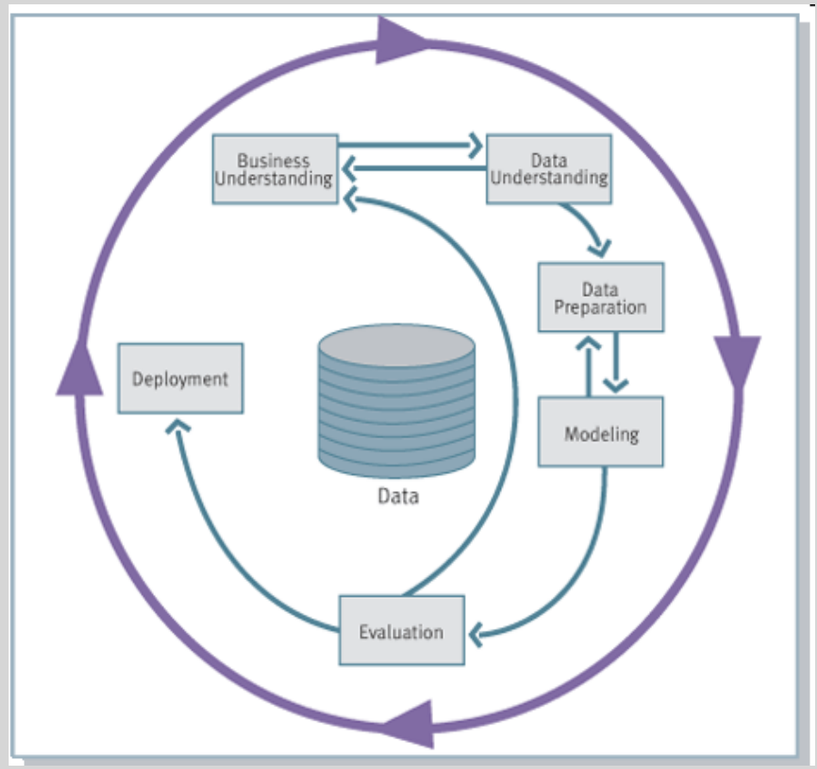
Motivation

- CRISP: DM-Prozess besteht aus unterschiedlichen Teilaufgaben
- Datenvorverarbeitung spielt wichtige Rolle im DM-Prozess
- Systematische Evaluationen erfordern flexible und strukturierte Experimentierumgebung
- Ggf. periodische Wiederholungen von Analysen notwendig

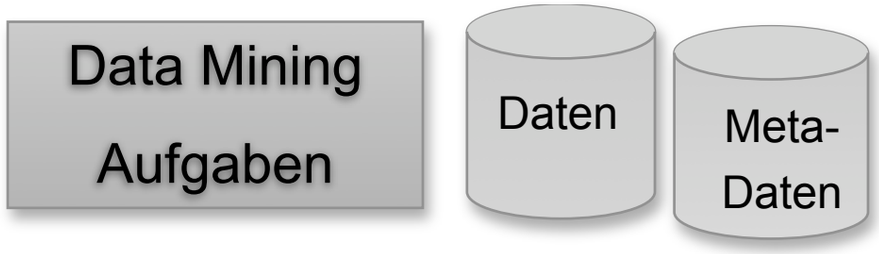


Anforderungen

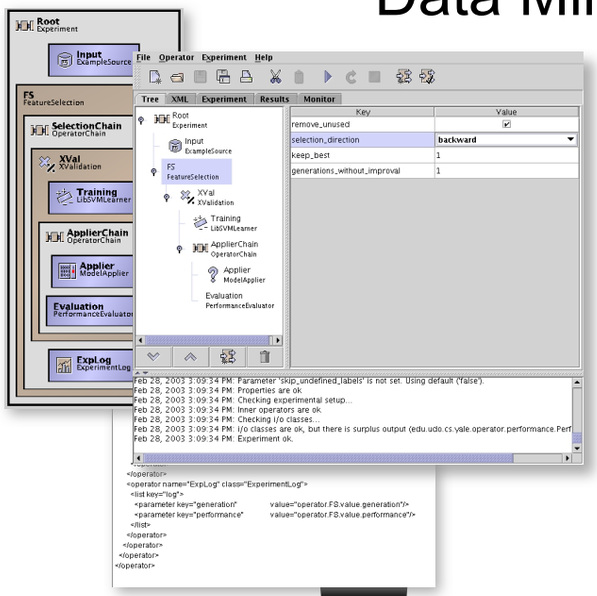
- Einfache wiederverwendbare Spezifikation von DM-Prozessen
- Austauschbarkeit von Lern-Verfahren, insbesondere:
  - Durchführen von Verfahrensvergleichen
- Kombination/Verschachtelung von Verfahren
- Verfahren zur Merkmalsauswahl und -generierung



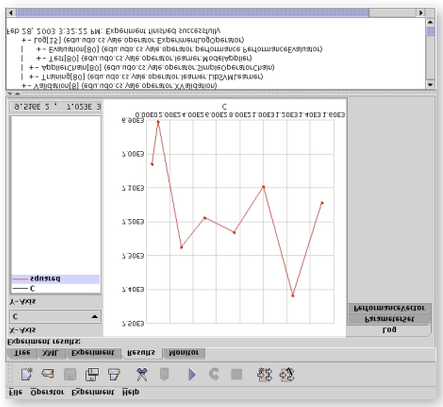
Konzept



Data Mining Experiment

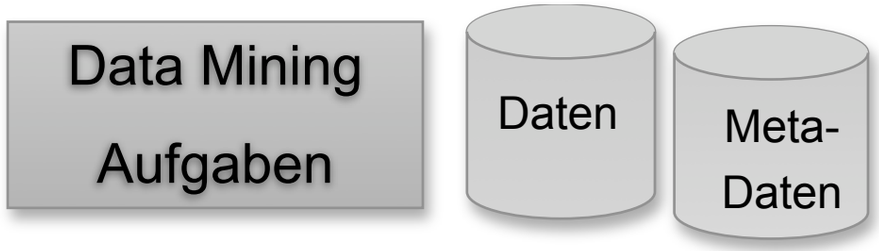


Durchführung



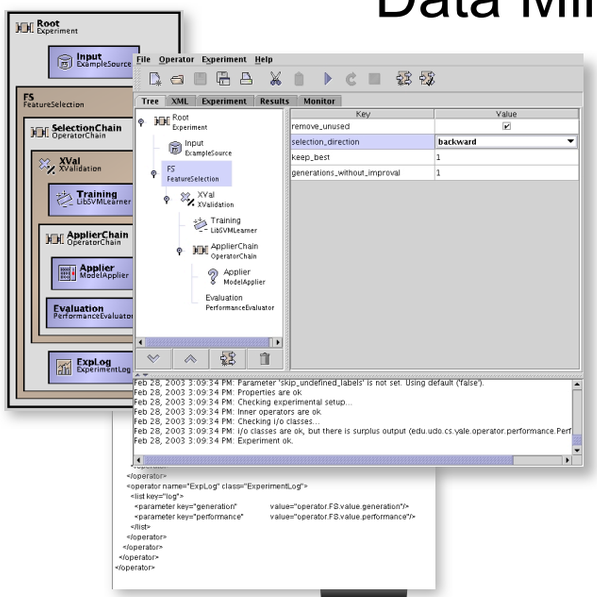
Ergebnis

Konzept



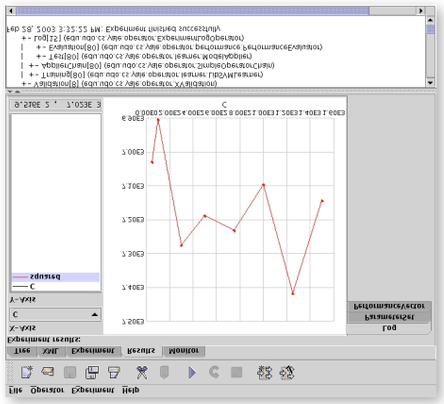
Entwurf

Data Mining Experiment



Durchführung

Anpassung des Entwurfes



Ergebnis

Konzept

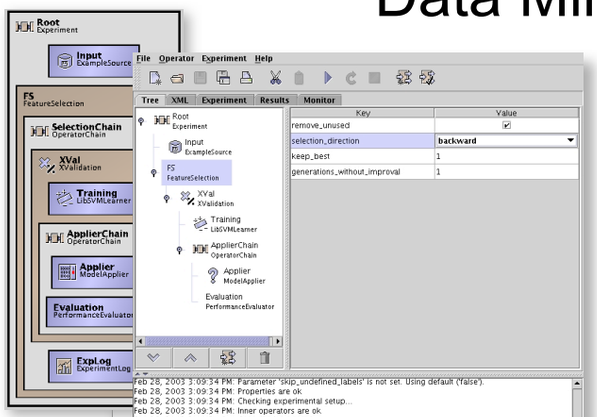
Data Mining Aufgaben

Daten

Meta-Daten

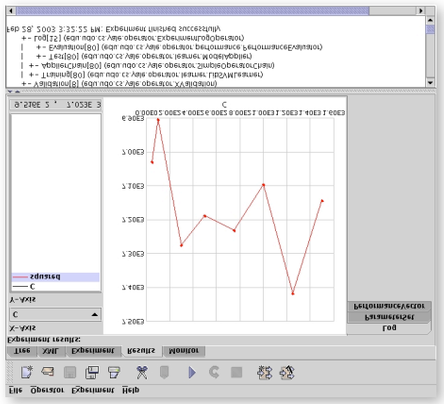
Entwurf

Data Mining Experiment



Durchführung

Anpassung des Entwurfes



Ergebnis

Data-Mining Experiment

Data Mining Aufgaben

- Klassifikation/Regression
- Transduktion
- Merkmalsextraktion
- Merkmalsgenerierung
- Merkmalsselektion
- Concept Drift
- Zeitreihenanalyse
- Text-Mining

RapidMiner

- Modellierung von DM-Prozessen als Abfolge von Operatoren (Ketten)
- Verschachtelung von Operatoren
- Transparente/effiziente Datenhaltung
- Leichte Erweiterbarkeit
- GUI-Modus/Batch-Modus
- Einbindung externer Programme (z.B. Weka, SVM-Imlementierungen)

Integrierte Operatoren

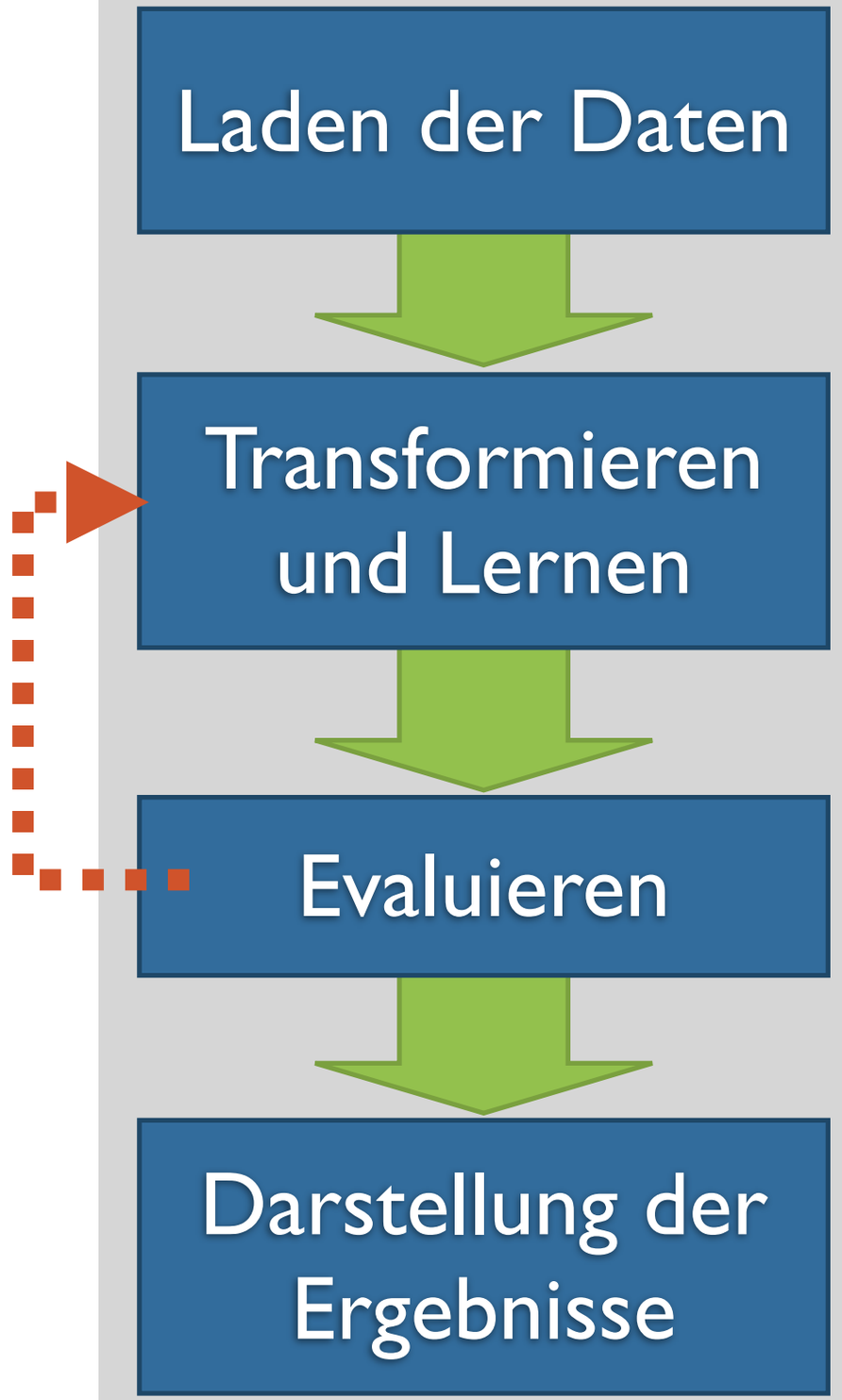
- Operatoren zur Ein-/Ausgabe
- Datenvorverarbeitung
- Zahlreiche Lernverfahren
  (Weka-Lerner, Clustering, ...)
- Performanzbewertung von
  Lernverfahren
- Verwaltung/Ausgabe von
  Lernergebnissen

Information

- Open-Source (GPL-Lizenz)
- Erfolgreiche Anwendung auf unterschiedliche Lernaufgaben
- Weltweite Verbreitung (Anwender / Wissenschaftler in über 30 Ländern)
- Dokumentation/Download/uvm unter
  http://rapid-i.com

DM-Experiment

- Laden der Daten
  - Datenbank, Datei
- Transformieren und Lernen
  - Fehlende Werte? Normierung? Klassifikation? Clustering?
- Optimierung:
  - Verfahrensauswahl, Parameter
- Ausgabe der Ergebnisse
  - Performanz, Regeln, Cluster



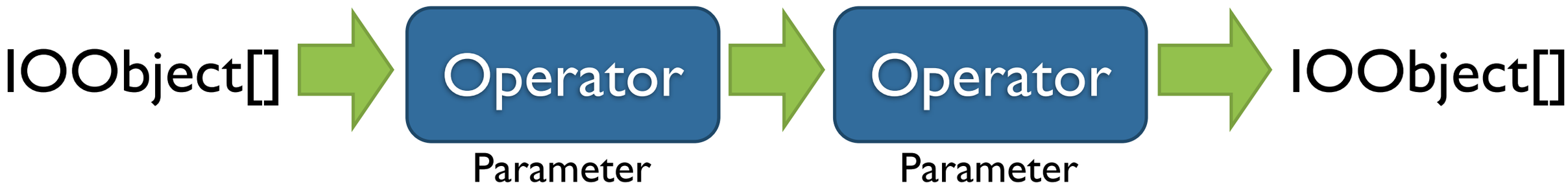
Operator/OperatorChain

Operator/OperatorChain

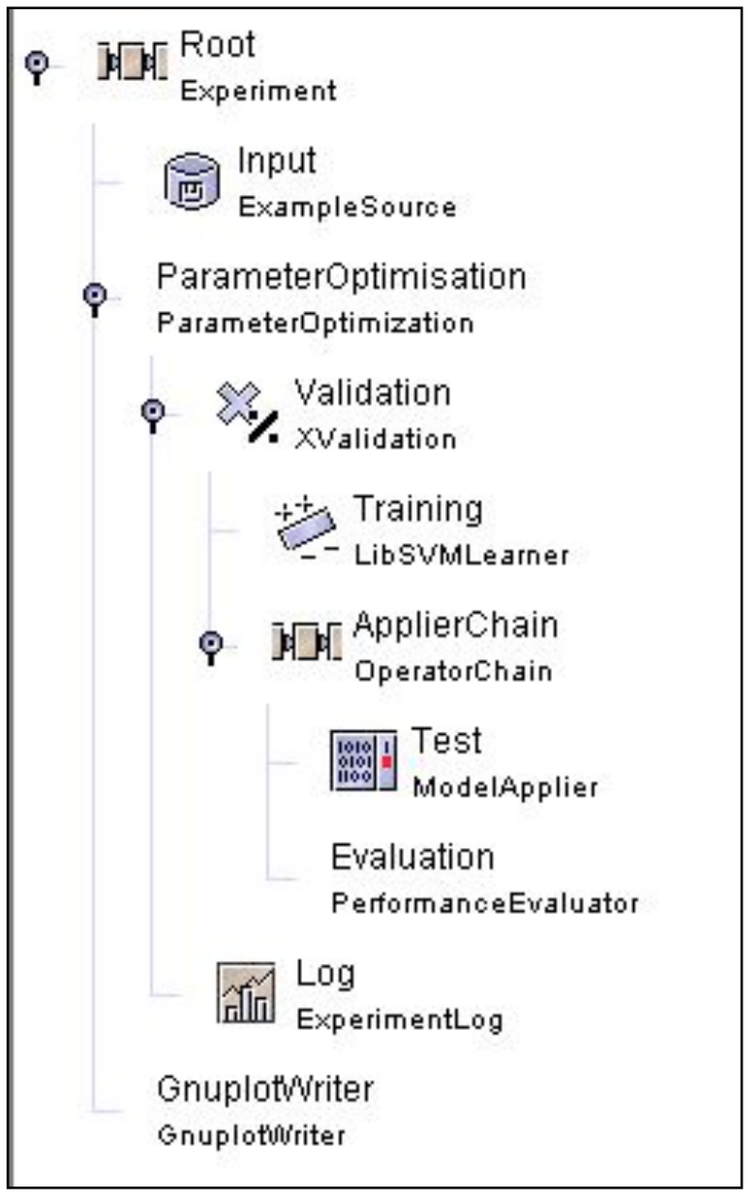
- Operator: Typ, Name, Eingabe, Ausgabe, Parameter

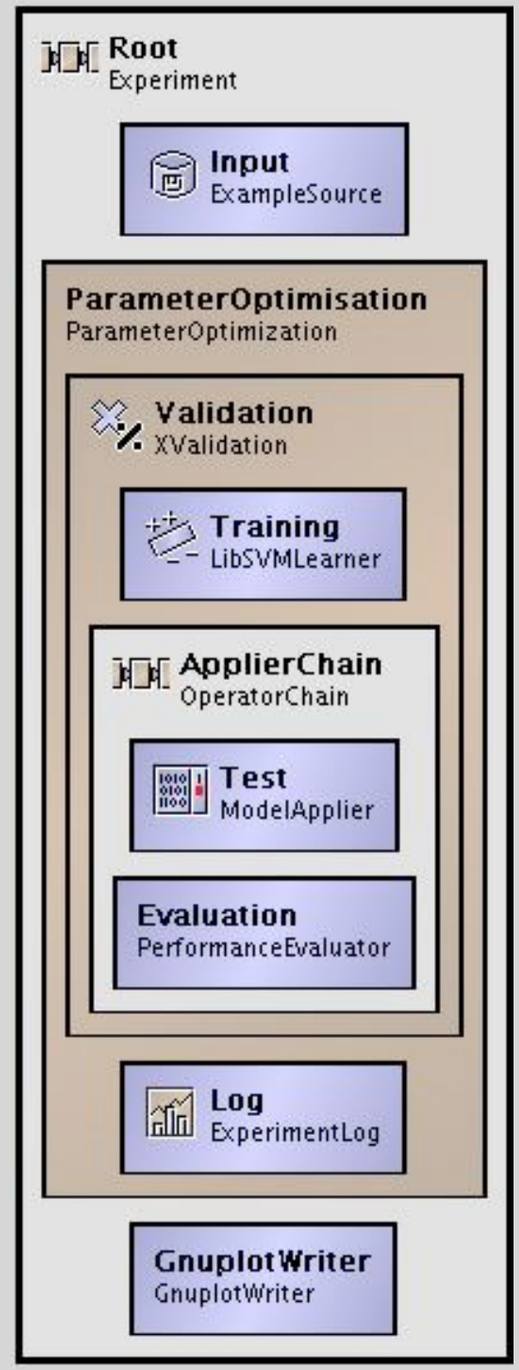
Operator/OperatorChain

- Operator: Typ, Name, Eingabe, Ausgabe, Parameter



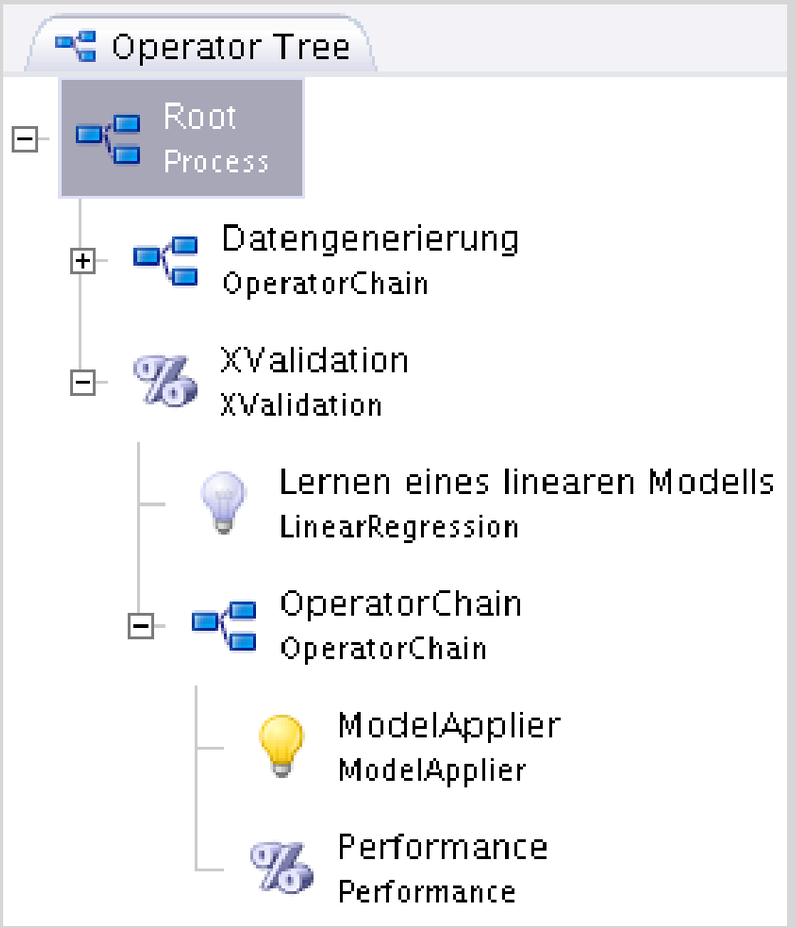
Beispiel: Operatorbaum





Beispiel: X-Validierung

- Zur Erinnerung:
  - Aufteilen der Daten in *n* Teilmengen
  - Lernen eines Modells auf *n-1* Mengen, Testen des Modells auf der übriggebliebenen Menge



Aufbau/ Ablauf

Parameter

Logfenster

IOObject

- Objekte, die zwischen Operatoren ausgetauscht werden
- Beispiele:
  - ExampleSet (eine Menge von Daten)
  - Model (gelerntes Model)
  - PerformanceVector (Menge von Leistungsmaßen)
  - Ähnlichkeit
  - Merkmalsgewichte
  - ...

ExampleSet (IOObject)

- Beschreibung der Attribute (Metadaten):
  - Name
  - Skala: nominal, integer, real, ...
  - Einheit
  - Typ: Einzelwert, Zeitreihe, ...
  - Position (Spalte) in der Datendatei
- Sicht auf Daten

ExampleSet (IOObject)

- Spezielle Attribute:
  - Label
  - Predicted label
  - Id
  - Cluster
- Beliebig erweiterbar...

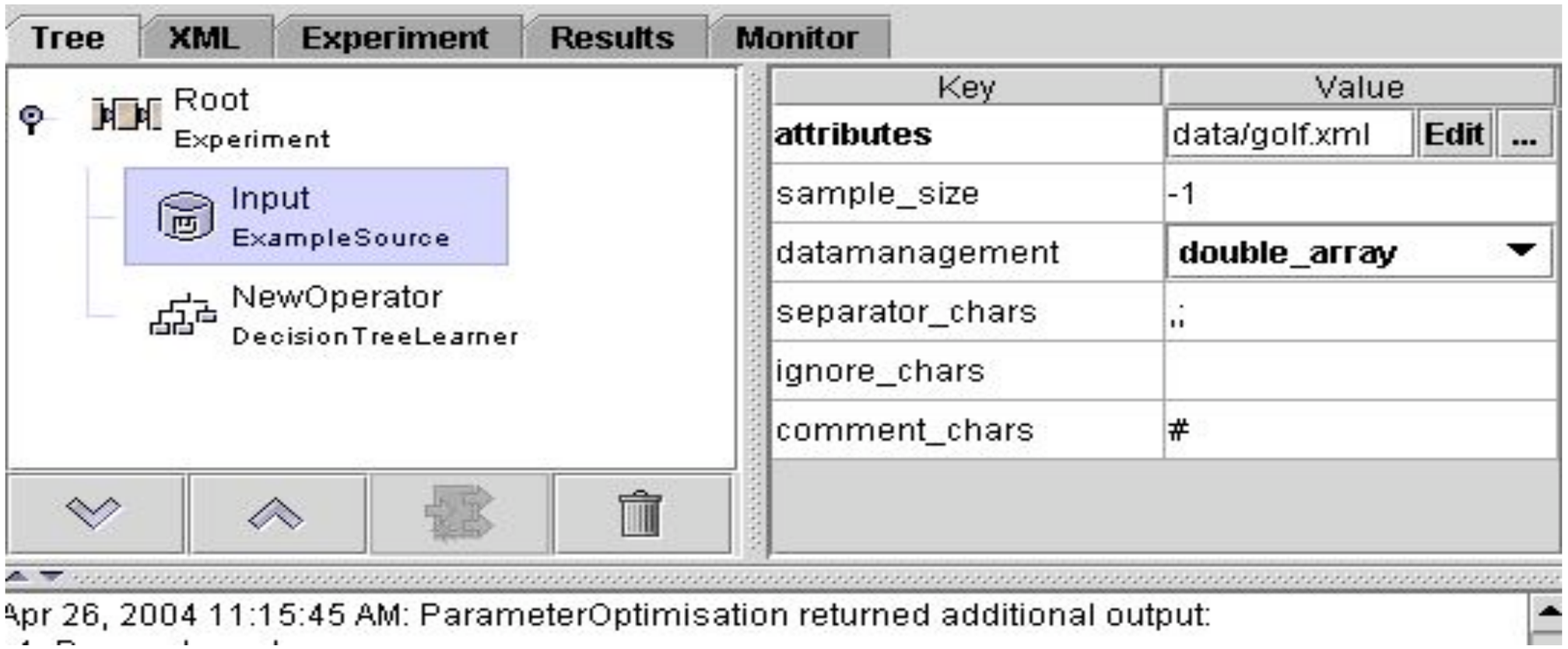
ExampleSet (IOObject)

File Table

golf.data (1)	golf.data (2)	golf.data (3)	golf.data (4)	golf.data (5)
Outlook	Temperature	Humidity	Wind	Play
attribute	attribute	attribute	attribute	label
[unit]	[unit]	[unit]	[unit]	[unit]
nominal	integer	integer	nominal	nominal
single...	single...	single...	single...	single...
sunny	85.0	85.0	false	no
sunny	80.0	90.0	true	no
overcast	83.0	78.0	false	yes
rain	70.0	96.0	false	yes
rain	68.0	80.0	false	yes
rain	65.0	70.0	true	no
overcast	64.0	65.0	true	yes
sunny	72.0	95.0	false	no
sunny	69.0	70.0	false	yes

ExampleSet (IOObject)

- Input: -
- Output: ExampleSet
- Parameter: Attributdatei, Datendatei, Sampling, ...



IOObjects/Resultate