

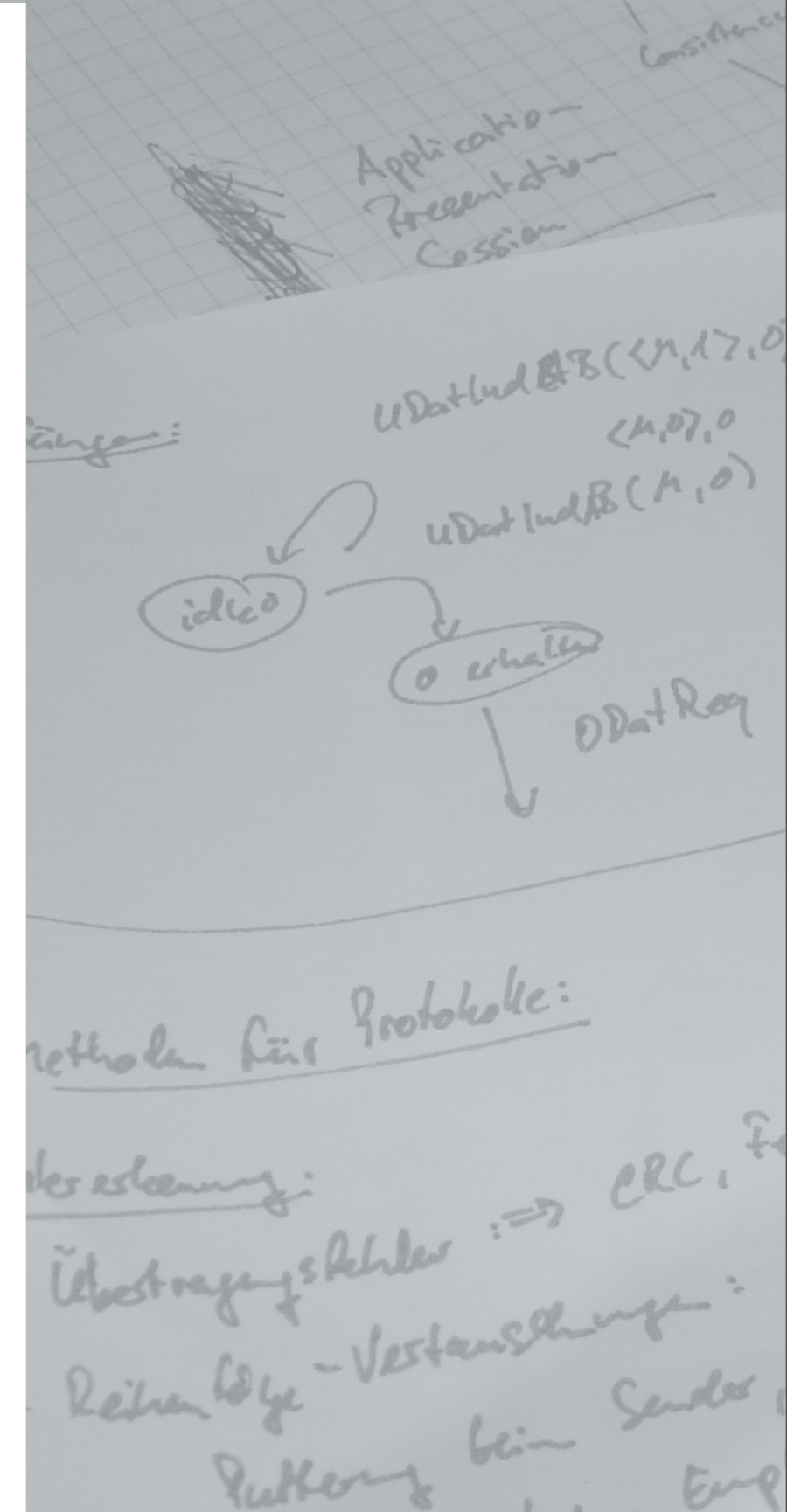
# Data Mining mit RapidMiner





# Motivation

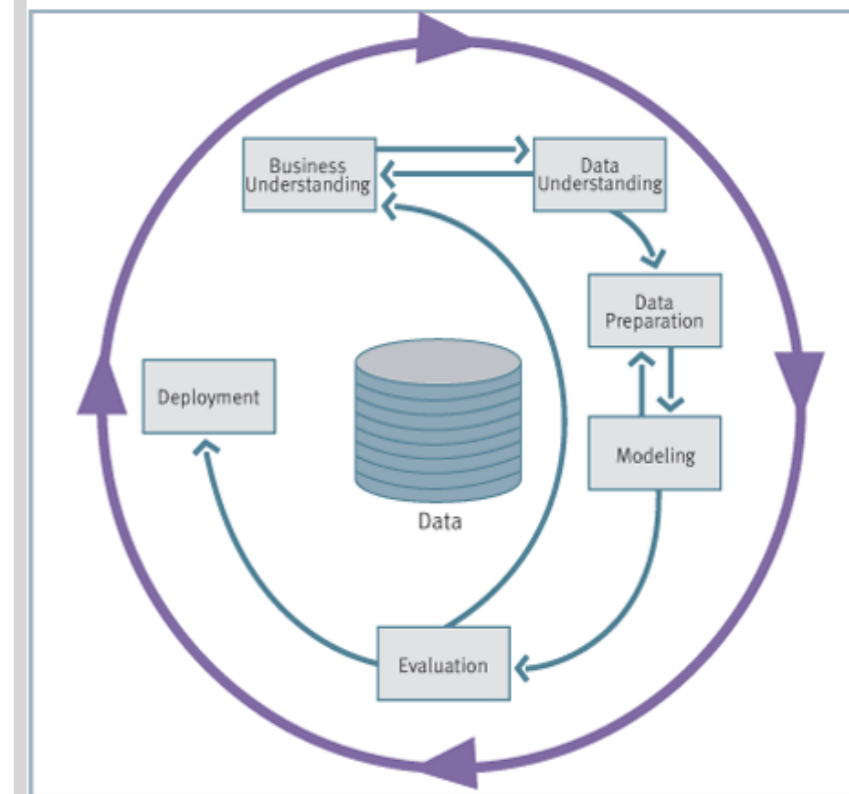
- CRISP: DM-Prozess besteht aus unterschiedlichen Teilaufgaben
- Datenvorverarbeitung spielt wichtige Rolle im DM-Prozess
- Systematische Evaluationen erfordern flexible und strukturierte Experimentierumgebung
- Ggf. periodische Wiederholungen von Analysen notwendig





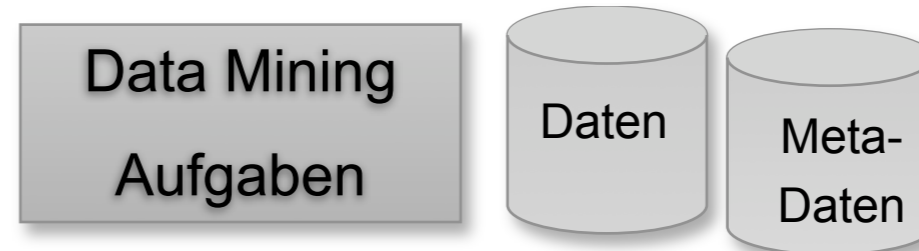
# Anforderungen

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





# Konzept



Entwurf

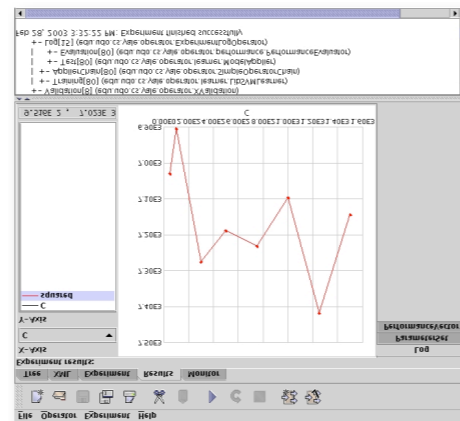
Anpassung des Entwurfes

Data Mining Experiment

Durchführung

## Data-Mining Experiment

Ergebnis



Tree

- Root Experiment
  - Input ExampleSource
  - FS FeatureSelection
    - SelectionChain OperatorChain
      - XVal XValidation
      - Training LibSVMTrainer
      - ApplierChain OperatorChain
        - Applier ModelApplier
        - Evaluation PerformanceEvaluator
        - ExpLog ExperimentLog

Monitor

Key	Value
remove_unused	<input checked="" type="checkbox"/>
selection_direction	backward
keep_best	1
generations_without_improval	1

```

Feb 28, 2003 3:09:34 PM: Parameter 'skip_undefined_labels' is not set. Using default (false).
Feb 28, 2003 3:09:34 PM: Properties are ok.
Feb 28, 2003 3:09:34 PM: Checking experimental setup...
Feb 28, 2003 3:09:34 PM: Inner operators are ok.
Feb 28, 2003 3:09:34 PM: Checking I/O classes.
Feb 28, 2003 3:09:34 PM: I/O classes are ok, but there is surplus output (edu.uco.cs.yale.operator.performance.Perf
Feb 28, 2003 3:09:34 PM: Experiment ok.
    
```



# 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-Implementierungen)







# Integrierte Operatoren

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



# Information

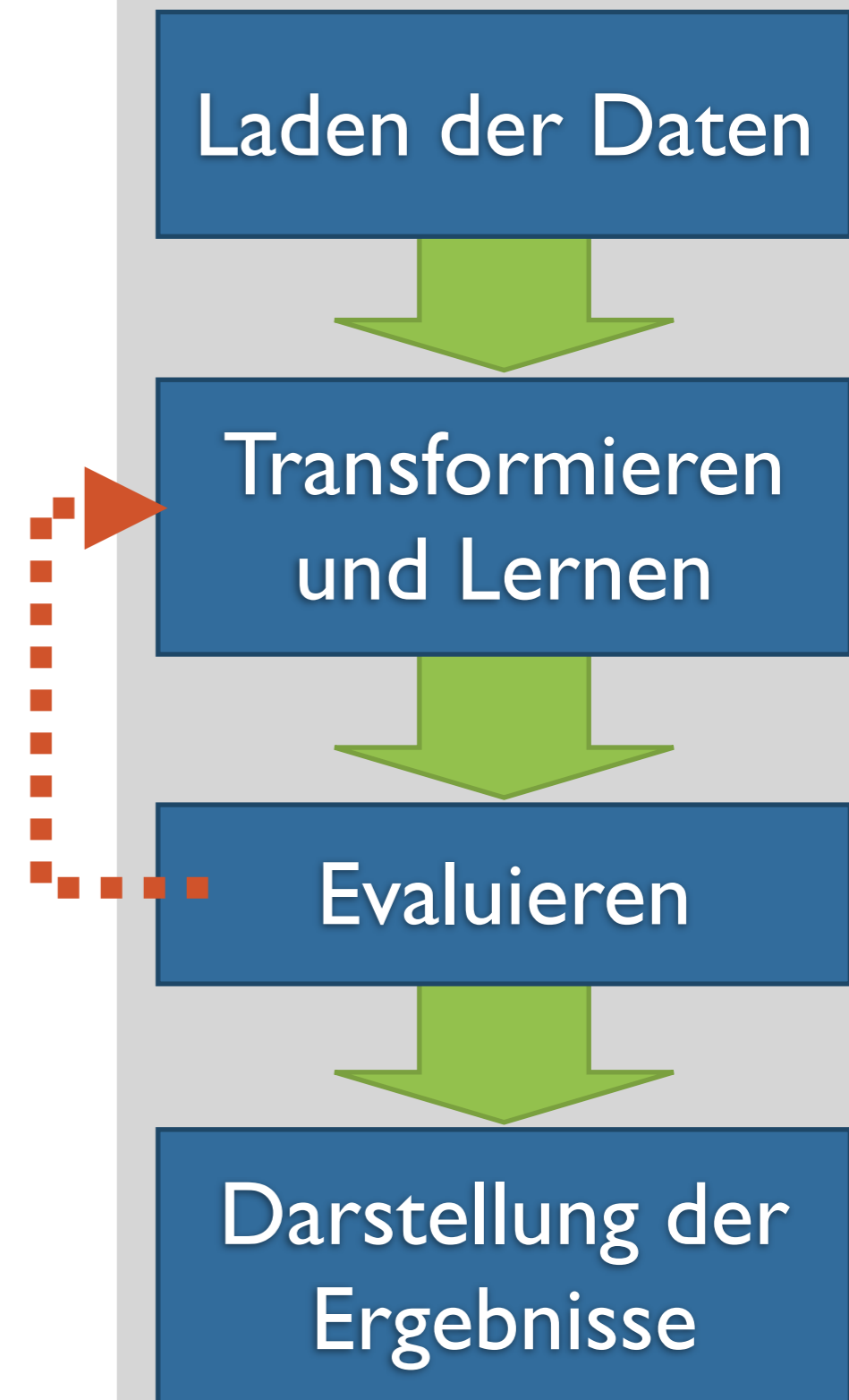
- Open-Source (GPL-Lizenz)
- Erfolgreiche Anwendung auf unterschiedliche Lernaufgaben
- Weltweite Verbreitung (Anwender und Forscher in über 30 Ländern)
- Dokumentation/Download/uvm unter **<http://rapid-i.com>**





# DM-Experiment

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



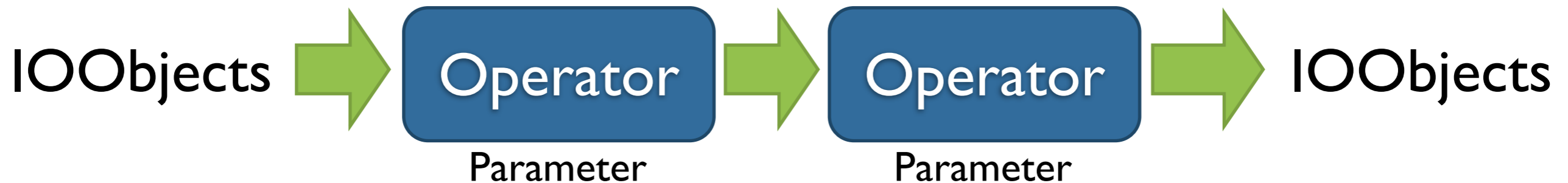


# Operator/OperatorChain

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

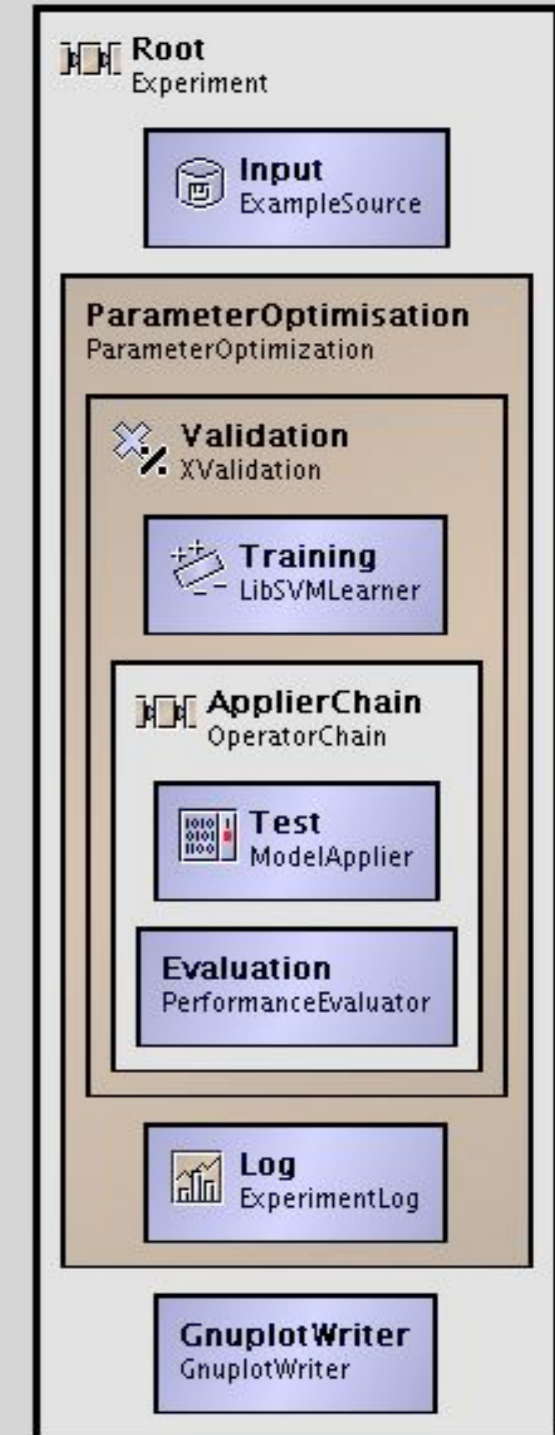
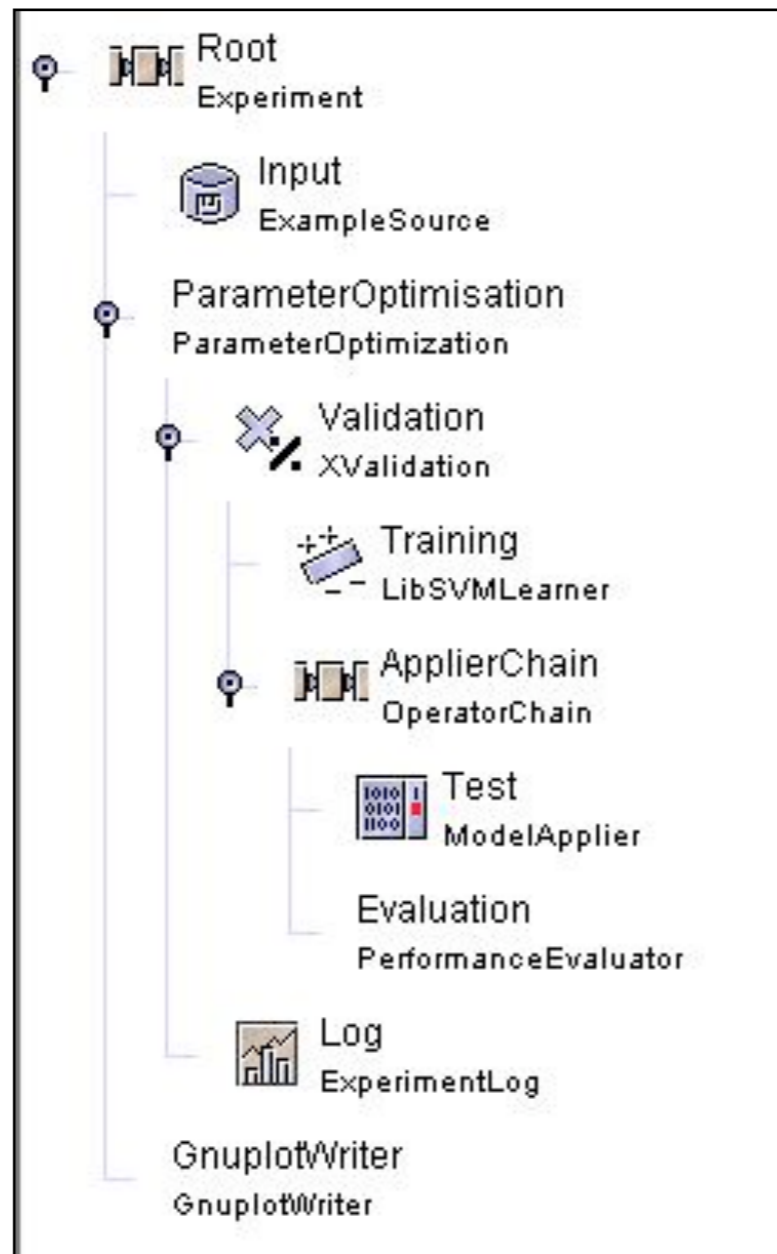


- OperatorChain: Verkettung von Operatoren





# Beispiel: Operatorbaum





Aufbau/  
Ablauf

Key	Value
model_file	
svm_type	epsilon-SVR
kernel_type	poly
degree	6.0
gamma	0.0
coef0	0.0
nu	0.5
cache_size	40
C	200.0
epsilon	0.01
p	0.1
shunking	<input checked="" type="checkbox"/>

```

#2: Humidity:=Humidity[]: integer/single_value/no block[0.0-50.0, avg=00.20371420371429,
#3: Wind:=Wind[]: nominal/single_value/no block [true,false]
label = #4: Play:=Play[]: nominal/single_value/no block [yes,no]
Apr 26, 2004 10:42:10 AM: ID3 learner 'NewOperator': starts learning.
Apr 26, 2004 10:42:10 AM: ID3 learner 'NewOperator': ID3 has successfully learned a decision tree.
Apr 26, 2004 10:42:10 AM: NewOperator returned additional output:
1. Model (type Tree) for label #4: Play:=Play[]: nominal/single_value/no block [yes,no]
Apr 26, 2004 10:42:10 AM: NewOperator: execution time was 153 ms
Apr 26, 2004 10:42:10 AM: Root returned additional output:
1. Model (type Tree) for label #4: Play:=Play[]: nominal/single_value/no block [yes,no]
Apr 26, 2004 10:42:10 AM: Root: execution time was 224 ms
Apr 26, 2004 10:42:10 AM: Experiment finished after 0 seconds
Apr 26, 2004 10:42:10 AM: Experiment:
Rootf11 (Experiment)
    
```

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 (100bject)

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





# ExampleSet (100bject)

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



# ExampleSet (100bject)

File Table				
golf.data (1)	golf.data (2)	golf.data (3)	golf.data (4)	golf.data (5)
Outlook	Temperature	Humidity	Wind	Play
<b>attribute</b> ▼	<b>attribute</b> ▼	<b>attribute</b> ▼	<b>attribute</b> ▼	<b>label</b> ▼
[unit]	[unit]	[unit]	[unit]	[unit]
<b>nominal</b> ▼	<b>integer</b> ▼	<b>integer</b> ▼	<b>nominal</b> ▼	<b>nominal</b> ▼
<b>single...</b> ▼	<b>single...</b> ▼	<b>single...</b> ▼	<b>single...</b> ▼	<b>single...</b> ▼
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



# ExampleSource (Operator)

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

Key	Value
<b>attributes</b>	data/golf.xml <b>Edit</b> ...
sample_size	-1
datamanagement	<b>double_array</b> ▼
separator_chars	::
ignore_chars	
comment_chars	#

Apr 26, 2004 11:15:45 AM: ParameterOptimisation returned additional output:



# 100objects/Resultate

The screenshot shows the YALE (simple.xml\*) application window. The main area displays a decision tree with the following structure:

```

graph TD
    Wind[Wind] -- true --> Outlook1[Outlook]
    Wind -- false --> Temperature1[Temperature]
    Outlook1 -- rain --> no1[no]
    Outlook1 -- overcast --> yes1[yes]
    Outlook1 -- sunny --> Temperature2[Temperature]
    Temperature2 -- ≤75.0 --> yes2[yes]
    Temperature2 -- >75.0 --> no2[no]
    Temperature1 -- ≤83.0 --> yes3[yes]
    Temperature1 -- >83.0 --> no3[no]
    
```

The bottom panel shows the following log output:

```

Apr 26, 2004 10:42:10 AM: Root returned additional output:
1. Model (type Tree) for label #4: Play:=Play[]: nominal/single_value/no block [yes,no]
Apr 26, 2004 10:42:10 AM: Root: execution time was 224 ms
Apr 26, 2004 10:42:10 AM: Experiment finished after 0 seconds
Apr 26, 2004 10:42:10 AM: Experiment:
Root[1] (Experiment)
+- Input[1] (ExampleSource)
+- NewOperator[1] (DecisionTreeLearner)
    
```

The interface includes a menu bar (File, Operator, Experiment, Help), a toolbar with icons for file operations and execution, and a 'Tree' tab selected in the top navigation. A 'Save...' button is visible in the bottom right of the main area.