Knowledge Discovery in Databases

Prof. Dr. Gerd Stumme
Dipl.-Wi.-Inf. Andreas Hotho
FG Wissensverarbeitung
FB Mathematik/Informatik

Organisatorisches

Vorlesung

- Beginn: 20. April 2004
- Dienstag, 10 – 12 Uhr in Raum 1332

Übungen

- Mittwoch, 16 – 18 Uhr in Raum -1606
- Beginn: 28. April 2004
- fällt aus am 5. Mai
- wird als Präsenzübung abgehalten (s. nächste Folie)

Organisatorisches

Präsenzübung bedeutet

- **selbständiges Bearbeiten** des Übungsblattes in Kleingruppen à 3-4 Personen unter Betreuung des Assistenten
- **kein prinzipielles Wiederholen** des Vorlesungsstoffs
- **kein Vorrechnen** der Musterlösung etc. (Diese wird später zur Verfügung gestellt.)

- **Nötig dafür:**
  - selbständige Vorlesungsnachbereitung **vor** der Übung
  - Mitbringen des Skriptes
  - eigene Aktivität entfalten

Organisatorisches

Warum ein neues Übungskonzept?

- aktives Erarbeiten des Vorlesungsstoffes bringt mehr
- Zusammenhänge im Stoff erkennen
- strukturiertes Denken und selbständiges Arbeiten lernen
- Teamarbeit lernen
- Erklären lernen (als Tutor und als Teilnehmer)
- Klausurtraining ;-)
- *Ihr Studium der ... haben Sie abgeschlossen. Zu Ihren persönlichen Stärken zählen Sie Eigeninitiative, Kommunikations- und Kooperationsbereitschaft, Teamarbeit.* (Typischer Anzeigentext)

Organisatorisches

Sprechstunden nach Absprache:

Prof. Dr. Gerd Stumme (Vorlesung): stumme@cs.uni-kassel.de, 0561/804-6251
Dipl.-Wi.-Inf. Andreas Hotho (Übungen): hotho@cs.uni-kassel.de, 0561/804-6252

FG Wissensverarbeitung, FB Mathematik/Informatik
Raum 0439, Wilhelmshöher Allee 73

Informationen im Internet: http://www.kde.cs.uni-kassel.de
Hier ist u.a. folgendes zu finden:

- aktuelle Ankündigungen
- Folienkopien
- Übungsblätter
- Literaturempfehlungen
- Termine

Ausgewählte Literatur

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurasamy. **Advances in Knowledge Discovery and Data Mining**. Cambridge, London. MIT press, 1996.
- T.M. Mitchell. **Machine Learning.** McGraw-Hill. 1997.
- P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth: **CRoss Industry Standard Process for Data Mining**, 1999, http://www.crisp-dm.org/
- Weitere Literatur findet sich auf der Homepage der Vorlesung.

Die Folien wurden im wesentlichen vom Institut AIFB der Universität Karlsruhe übernommen. Bei der Erstellung der Folien haben u.a. mitgewirkt: R. Engels, M. Erdmann, A. Hotho, A. Mädche, S. Staab, R. Studer, G. Stumme

Übersicht über die Vorlesung

I. **Einführung**
- Allgemeines & Organisatorisches
- Fallstudien von Knowledge Discovery Anwendungen
- CRISP-DM Prozessmodell

II. **Datenbereitstellung**
- Data Warehousing / Data Mart

III. **Vertrautmachen mit Daten**
- Online Analytical Processing (OLAP)
- Visualisierung großer Datenmengen
- Datencharakteristiken (DCT)

IV. **Preprocessing**
- Datenreduktion
- Datenableitung
- Datentransformation
- Diskretisierung

Übersicht über die Vorlesung

V. **Einführung in das Text Mining**

VI. **Überwachte Data und Text Mining Verfahren**
- Entscheidungsbaumverfahren C4.5
- Induktives Logisches Programmieren (ILP)
- Künstliche Neuronale Netzwerke

VII. **Unüberwachte Data und Text Mining Verfahren**
- Clustering: Self Organizing Maps
- Formale Begriffsanalyse
- Assoziationsregeln
- Generalisierte Assoziationsregeln mit Taxonomien

VIII. **Modellierung (Zusammenfassung)**

IX. **Evaluierung**

X. **Anwendung**

I.1 Problemstellung (Fayyad et al. 1996)

- **Möglichkeiten zur Sammlung und Generierung von Daten wächst explosionsartig:**
  - **Database Marketing**
    - Verkaufsdaten (Grundlage: bar codes)
    - Kreditkartentransaktionen
    - Telefongespräche
  - **Umweltüberwachung**
    (Grundlage: Sensoren + Vernetzung)
  - **Produktdatenbanken**
  - **Internet- und Intranetdokumente**
    - Semi-strukturierte Dokumente (HTML, XML)
    - unstrukturierte Dokumente

- **Gigabytes an neuen Daten pro Tag/Woche:**
  - welche Daten sind tatsächlich nützlich?
  - Datenfriedhof
- **Standardanalysemethoden**:
  - Spreadsheets
  - ad-hoc DB-Anfragen (SQL)

  sind nicht mehr hinreichend
- **Methoden und Werkzeuge zur Unterstützung des Menschen bei der Generierung nützlichen Wissens aus großen Datenbeständen und Dokumenten werden benötigt**
- **Ziel ist der Aufbau von (interpretierbaren) Modellen**

Knowledge Discovery in Databases (KDD) :
(Wissensgewinnung aus Datenbanken)

Definition (Fayyad et al. 1996)

"Knowledge Discovery in Databases (KDD)
is the non-trivial process of
identifying valid, novel, potentially useful,
and ultimately understandable
patterns in data"

- **Daten:**
  - Menge F von Fakten (Fällen, Beispielen) („cases, examples")

    z.B. Tupel einer relationalen DB
    Sätze in einer Datei
    Text-Dokumente aus dem Web

- **Muster** („pattern", generiertes Wissen):
  - Ausdruck E einer Sprache L zur Beschreibung von Beziehungen in F

    z.B.:
    - Wertebeschränkung für DB-Felder
    - Beziehung zwischen DB-Feldern
    - Regeln zwischen Werten / Worten
    - „**interessante**" Worte
  - E ist einfacher als die Aufzählung der Faktenmenge F und lässt sich auf neue Daten übertragen

- **Verständlichkeit** (ultimately understandable):
  - gefundene Muster müssen für den Menschen verständlich sein
  - wie kann man Muster beschreiben ?
- **Gültigkeit** (validity):
  - gefundenes Muster sollte mit gewisser Sicherheit für neue Daten zutreffend sein
- **Prozess** (process):
  - Prozess ist mehrstufig, u.a.
    - Business Understanding
    - Data Preparation
    - Modeling
  - nicht-trivial
    z.B. *nicht* Berechnung Mittelwert

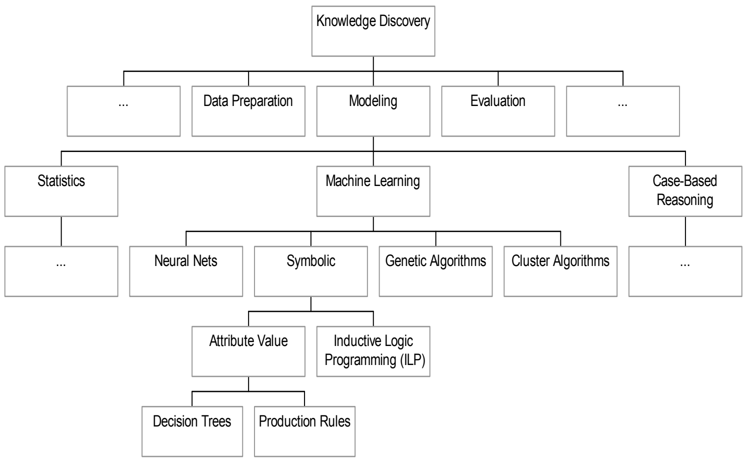
Data Mining

- **zwei alternative Bedeutungen**
- **Bedeutung (1):**
  - Synonym für KDD: beinhaltet alle Aspekte des Prozesses der Wissensgewinnung
  - diese Bedeutung ist insbesondere in der Praxis verbreitet
- **Bedeutung (2):**
  - Teil des KDD-Prozesses:
    Mustergewinnung / Modellierung, Interpretation
  - Anwendung von Algorithmen, die unter gewissen Ressourcenbeschränkungen Muster / Modelle E bei gegebener Faktenmenge F erzeugen

"Data Archeology"

(Brachman)

Typen von Verfahren



(CRISP: **http://www.crisp-dm.org/**)

- **Task Types [Aufgabentypen] may be defined**
  - from a **method-oriented** view
    - what is the methodological approach?
  - from an **application-oriented** view
    - which business problem has to be solved?
- **Up to now there does not exist a standard of task types**

– Segmentation [Segmentierung]

- separate data into interesting and meaningful subgroups
- all members of a subgroup share common characteristics
- segmentation may be a data preparation step or the main modeling step
- segmentation may result in
  an enumeration of the members of the subgroups
  or in a conceptual description of the subgroups

- **appropriate techniques (among others):**
  - conceptual clustering [begriffliches Clustern]
  - statistical clustering [statistisches Clustern]
  - Self-Organizing Maps (SOM)

– Classification [Klassifikation]

- assignment of objects to predefined classes
- each class label is a discrete (symbolic) value
- objective is to learn classification models (classifiers)
  which assign the correct class label to previously
  unseen and unlabeled examples
- the class label is known for the training examples

- **appropriate techniques (among others):**
  - decision tree learning [Entscheidungsbäume]
  - inductive logic programming (ILP)
  - k-nearest neighbour

– Prediction (Forecasting) [Vorhersage]

- similar to classification
- target attribute (class label) is continous attribute
- determine the numerical value of the target attribute
  for unseen examples

- **appropriate techniques (among others):**
  - regression analysis [Regressionsanalyse]
  - neuronal networks [Neuronale Netze]

– Dependency Analysis [Abhängigkeitsanalyse]

- find a model that describes significant dependencies between data items or events
- dependencies are strict or probabilistic
- associations are a special case of dependencies
  - describe data items or events which frequently occur together
- **sequential patterns are also a special kind of dependencies where sequences of events are analysed**

- **appropriate techniques (among others):**
  - regression analysis [Regressionsanalyse]
  - association rules [Assoziationsregeln]
  - Bayesian networks [Bayes'sche Netze]

– **Deviation Detection [Abweichungsanalyse]**
- identify deviation of values compared to previous values or normative values
- when is a deviation significant?
  - cause an action

- **appropriate techniques (among others):**
  - neuronal networks [Neuronale Netze]

(Dueck 1999)

– **There exists a large amount of potential application areas for Knowledge Discovery, sometimes called Business Intelligence Applications**
- **banks / insurance**
  - customer centric view (behaviour, risk, cross selling)
  - product view (portfolio analysis, cross selling)
- **commerce / retail**
  - market basket analysis, customer behaviour, analysis of regions
- **telecommunication**
  - customer relationship management
  - fraud detection
- **transportation**
  - one-to-one selling
  - aircraft maintenance

Application types

– **Customer Relationship Management (CRM)**
- **collection of various activities (based on a well-developed Data Warehouse), among others:**
- **customer retention:**
  - to acquire a new customer is much more expensive than to keep a customer
  - what are good characteristics of customers that might switch to a competitor?
    - cellular phone market
    - discount hopping with respect to credit cards

– **Customer Relationship Management (CRM) (continued)**
- **customer segmentation:**
  - What kind of customers do we have?
  - How many classes of customers?,
    (e.g. normal customers, techno freaks, ...)
  - use different campaigns for different classes
- **basket analysis:**
  - What products are bought together?
  - What types of customers do we have?
    - adjust product offerings
  - Is the customer behaviour different during the week?
- **marketing campaign management:**
  - What are promising target groups for specific product types?

I.2.2 Application-oriented view

– **Customer Relationship Management (CRM) (continued)**

- **fraud detection:**
  - How to avoid unpaid bills?
  - How to identify illegal use of cellular phones?

- **one-to-one business**
  - collect information about individual customers
  - have exactly those products available that are bought by your customers
  - clear relationship to cross selling

- **customer life cycle**
  - distinguish bad customers from customers that are potentially interesting in the future, e.g. students, grandchildren, ...

I.3 Examples from Real Life

Case Studies:

A. Data Warehousing / Data Mining in telecommunications

B. Adaptive Fraud Detection using Neural Networks

C. Determining process sequences for the manufacturing of work pieces using ILP

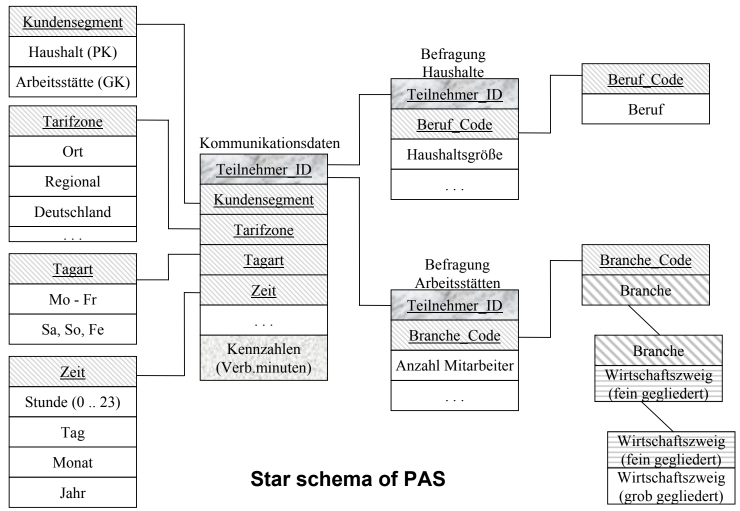
D. Text Mining on Reuters financial news

I.3.A Data Warehousing / Data Mining in Telecommunications

A. Example from Real Life: Data Warehousing / Data Mining in Telecommunications

- **panel** = customer inquiry using cross section and longitudinal section data (Quer- und Längsschnittdaten)
- approx. 5000 households
- **Data Mart "Panel Analysis System"** (PAS) containing
  - call detail records
  - social-demographic data

I.3.A Data Warehousing / Data Mining in Telecommunications



Star schema of PAS

Example: call detail record

customerID	distance	type of day	date/time	comm. minutes
1	Ort	Mo-Fr	19.11.98/9:55	20 min
1	Ort	Mo-Fr	20.11.98/10:10	18 min
2	Regional	Mo-Fr	19.11.98/21:00	120 min
2	Regional	Mo-Fr	20.11.98/17:00	2 min

Idea:

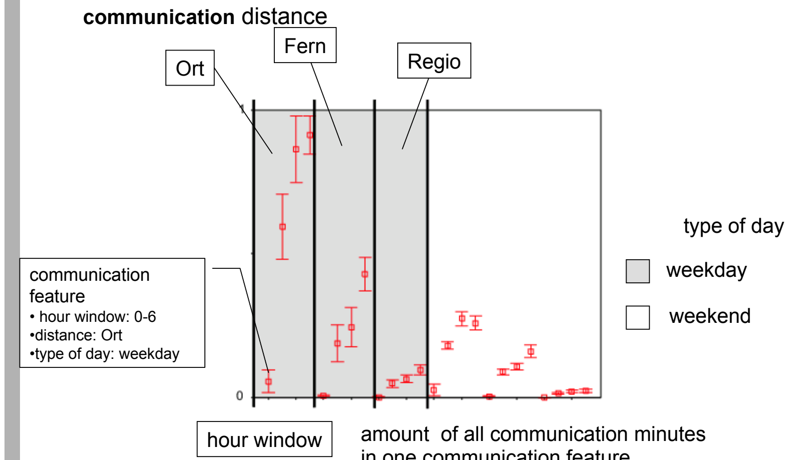
- Use call detail records to derive communications profiles of customers
- Identify customers, which have similar communications profiles => construct customer segments
- investigate the customer segments using social-demographic features

Data Preparation:

Application of OLAP-functionality to preprocess the customer detail records

- Exploratory analysis to derive suitable aggregation level
- Operation „pivot“ for „turning“ the data set, i.e., customerID becomes database key, communication minutes are summarized
- Operation „slice & dice“ for eliminating uninteresting attribute values (e.g. communication distances)

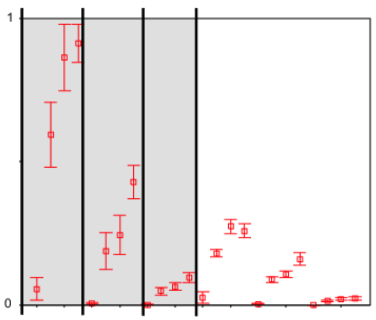
Average communication profile of panel customers



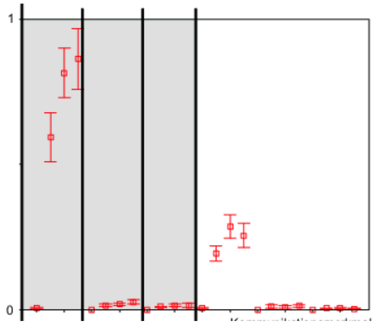
Identification of customer segments

- **Summarization of communication minutes for three months for all customers in reference to the 24 communication features**
- **Use partitioning cluster technique k-means**
- **5000 panel customers are separated into 10 clusters,**
- **The largest cluster contains 777 panel customers, the smallest 103 panel customers**

Profiles of customer segments



average over all
panel customers



1 cluster containing 777
cluster members

Interpretation of customer segments

- **Add social-demographic features, like**
  - size of household
  - profession
  - number of children
  - age of persons
  - nationality
  - ...
- **E.g. decision tree technique C5.0 delivers the rule**

WENN HH > 4 und Beruf = „Beamter“
DANN Cluster_Nr = 1

B. Example from real life: Adaptive Fraud Detection

(Fawcett and Provost 1997)

- Detecting fraudulent usage of cellular telephones

- fraud caused by cloning (Mobile Identification No., Electronic Serial No.)

- typical example of deviation detection

- detect unusual patterns of behavior
  ⇒ indicator for potentially fraudulent usage
- basis: typical profile of behavior
  e.g.
  - no. of calls
  - duration of calls (airtime)
  - origin of calls

- general approach

(1) Start from call data of the customers

(2) for each account learn rules which indicate fraudulent behavior

e.g. (TIME_OF_DAY = NIGHT) AND
(LOCATION = BRONX) → FRAUD
[Certainty factor = 0.89]

(3) select a subset of all generated rules

(tens of thousands of rules may be generated in step (2))

- select rules which cover a minimum number of accounts

(choose appropriate threshold)

I.3.B Adaptive Fraud Detection

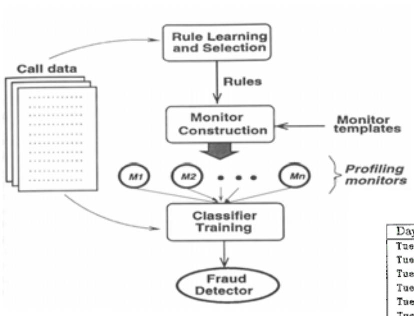


Figure 1: The framework for automatically constructing fraud detectors.

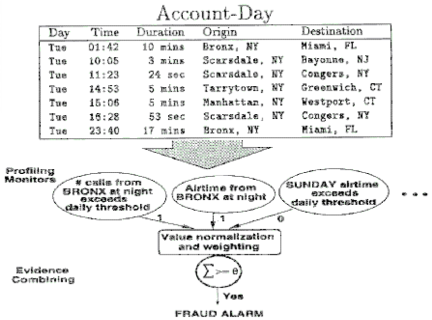


Figure 2: A DC-1 fraud detector processing a single account-day of data.

I.3.B Adaptive Fraud Detection

(4) construct <u>profiling monitors</u>
- rules are not universal since each account has its own typical behavior
- profiling monitors are trained for each account
  - identify normal behavior of a customer, e.g. ‘‘customer calls from Bronx an average of 5 minutes per night with a standard deviation of 2 minutes)

(5) <u>usage</u> of profiling monitor
- compare current customer behavior with normal behavior from step (4)
- indicate fraud if current behavior is above threshold
  e.g. ‘‘15 minute call at night from Bronx‘‘

(6) <u>combine evidence</u> from different monitors
- monitor output is weighted
- threshold is learned on the sum of the weighted outputs
- use a <u>neural net</u> for this step

I.3.B Adaptive Fraud Detection

- <u>results</u>:

- **initially 3630 rules were generated**
- **subset of 99 rules was selected**
- **finally 9 profiling monitors were used**
- **<u>quality</u> of fraud detection comparable to <u>hand-crafted</u> profiling methods**

I.3.C Determining process sequences for the manufacturing of work pieces

<u>C. Example from real life:</u> Determining process sequences for the manufacturing of work pieces (Wiese 1998)

- work pieces are described by relations between form elements of the work pieces

- form elements are described by attributes like
- ‘diameter‘
- ‘kind_of_form_element‘

- relations are e.g.
- ‘neighbor‘
- ‘precede‘

- relations represent <u>background knowledge</u> of the domain

- **approach**:
  - **use Inductive Logic Programming (ILP) approach**
    - **algorithm JoJo-Fol**
    - **exploit background knowledge**
    - **use restricted form of predicate logic to describe learned model**
      - e.g.
        ‘precede(X,Y) :- outside(X), inside(Y)‘
        ‘‘X precedes Y in the manufacturing process if X is on the ‘outside‘ and Y is on the ‘inside‘ “
    - **background knowledge defines**
      - terminology, i.e. predicates, which may be used within the rules
      - facts that are known to be true

- **training set**
  - around 2500 positive facts
    e.g. ‘precede (form_element1, form_element3)‘
  - around 2500 negative facts
- **background knowledge**
  - form elements are described by ground facts (attribute value pairs)
    e.g. ‘diameter(form_element3, 66)‘
  - relations between form elements are specified by ground facts
    e.g. ‘neighbor(form_element1, form_element2)‘
  - around 4900 facts are provided as background knowledge

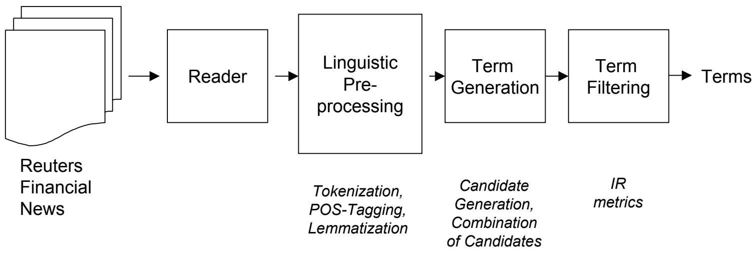
- **results**
  - **JoJo-Fol generates 51 rules with 164 premises**
  - **it takes several hours to generate these rules**
  - **achieved accuracy around 95%**

D. Example from real life: Text Mining at Term Level

(Feldman et al. 1998)

- **Reuters Financial News of years 1995-96**
- **in total 51.725 documents containing over 170.000 unique words**
- **size of collection is approx. 120 MB; each document contained on average 864 words**
- **mining goal: extract rules concerning interesting joint ventures**

Architecture



Term Generation

- **at this stage sequences of tagged lemmas are selected as potential term candidates on the basis of relevant morpho-syntactic patterns**

Term Filtering

- **reduce number of term candidates on the basis of some statistical relevance-scoring schema**
- **approx. 45 terms per document remain**

General association rule algorithm

- **generates rules between pairs of terms rather than individual terms**
- **constructed taxonomy enables the user to specify the mining task in a concise way**
- **user interest: business alliances between companies**

=> 12.000 frequent sets were generated
(support threshold 5 documents, confidence threshold 0.1)

=> frequent sets generated 575 associations

Sample of generated rules

america online inc., bertelsman ag. => joint venture (13/0.72)

apple computer inc., sun microsystems inc => merger talk (22/0.72)

apple computer inc., taligent inc. => joint venture (6/0.75)

sprint corp., tele-communications inc. => alliance (8/0.25)

burlington northern inc., santa fe pacific corp. => merger (14/0.4)

I.4 The KDD Process

(CRISP: **http://www.crisp-dm.org/pub-paper.pdf**)
(Fayyad et al. 1996, chapter 2)
(Engels 1999)

“Knowledge discovery is a knowledge-intensive task consisting of complex interactions, protacted over time, between a human and a (large) database, possibly supported by a heterogeneous suite of tools”

(Brachman/Anand 1996)

- **KDD process has to be oriented towards application task and user (process developer)**
- **development requires some knowledge about data bases, data analysis methods and application area**
- **KDD process is composed of a sequence of different steps**
- **KDD process is interactive and iterative**
  - user has to take decisions
  - some steps have to be carried out several times

- **in the literature you find different proposals for structuring the KDD process**
- **examples:**
  - process model of (Brachman/Anand 1996)
  - process model of (Engels 1999)
  - CRISP-DM methodology

    (Cross-Industry Standard Process Model for Data Mining)

    http://www.crisp-dm.org/
- **subsequently, we discuss the CRISP-DM methodology**

I.4.1 The CRISP-DM Methodology

- **hierarchical process model at four levels of abstraction:**
  - **phase: top level process decomposition**
    - business understanding
    - data understanding
    - data preparation
    - modelling
    - evaluation
    - deployment
  - **generic task: each phase is decomposed into several generic tasks**
    - cover the whole process (complete)
    - cover all possible applications (stable)

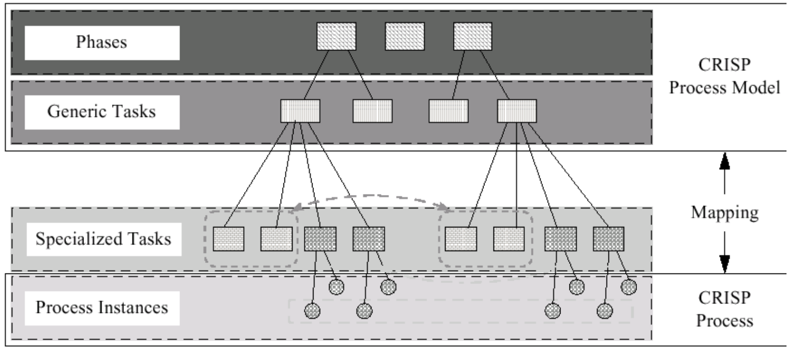


Figure 1: Four Level Breakdown of the CRISP-DM Methodology

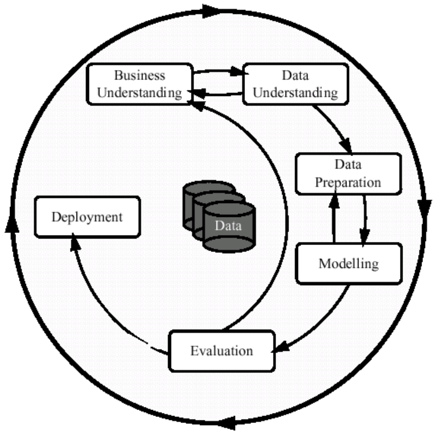


Figure 2: Phases of the CRISP-DM Reference Model

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 3: Generic Tasks (bold) and Outputs (italic) of the CRISP-DM Reference Model

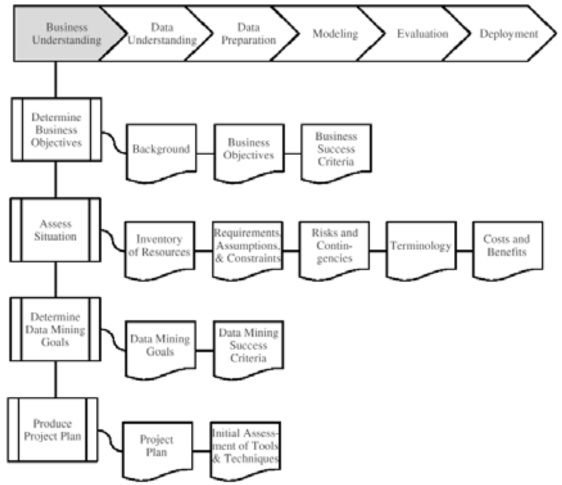
- **specialized task:**
  - **mapping of the generic tasks to specialized tasks that are adapted to the specific situation at hand**
  - **mapping is driven by Data Mining Context that is defined by 4 dimensions:**
    - **application domain**
    - **problem type**
    - **technical aspect**
    - **applied tool and techniques**
- **process instance:**
  - **record of the actions, decisions and results of an actually performed KDD process**

I.4.1 The CRISP-DM Methodology

Table 1: Dimensions of Data Mining Contexts and Examples

	Data Mining Context			
Dimension	*Application Domain*	*Data Mining Problem Type*	*Technical Aspect*	*Tool and Technique*
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		

I.4.1 (i) Business Understanding



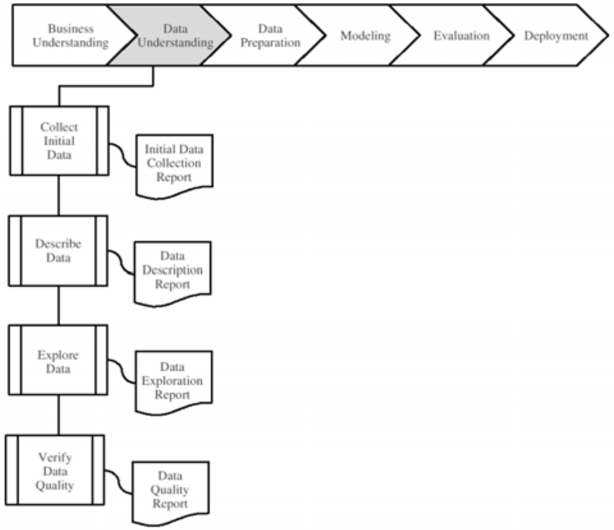
I.4.1 (i) Business Understanding

- **Determine Business Objectives**
  - understand from a business perspective what the client really wants to accomplish
    ⇔ do not produce the right answers to the wrong question
  - identify key persons (management, finance, domain expert, user)
  - define success criteria - related to business objectives
- **Assess Situation**
  - identify available resources as well as constraints and assumptions (e.g. legal issues)
  - identify risks (business, organisational, technical)

I.4.1 (i) Business Understanding

- **Determine Data Mining Goals**
  - derive data mining goals from business objectives
  - define data mining success criteria (e.g. model accuracy, model performance, ...)
- **Produce Project Plan**
  - take iterations into account
  - typical effort distribution:
    - 50% - 70% in Data Preparation Phase
    - 20% - 30% in Data Understanding Phase
    - 10% - 20% in Modeling, Evaluation and Business Understanding Phase
    - 5% - 10% in Deployment Phase

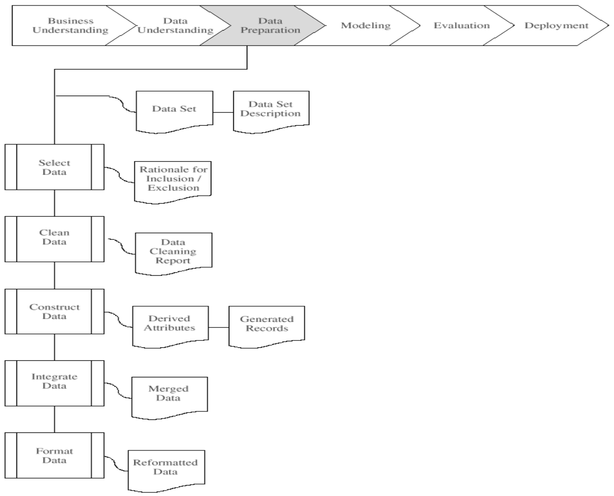
I.4.1 (ii) Data Understanding



I.4.1 (ii) Data Understanding

- **Collect Initial Data**
  - identify relevant attributes
  - identify inconsistencies between sources
- **Describe Data**
  - characterize attributes (relevance, statistical characteristics, ...)
- **Explore Data**
  - querying, visualization
- **Verify Data Quality**
  - identify errors in data
  - number of missing values
    - identify false encodings of missing values (e.g. 1.11.[19]11 as birthday)

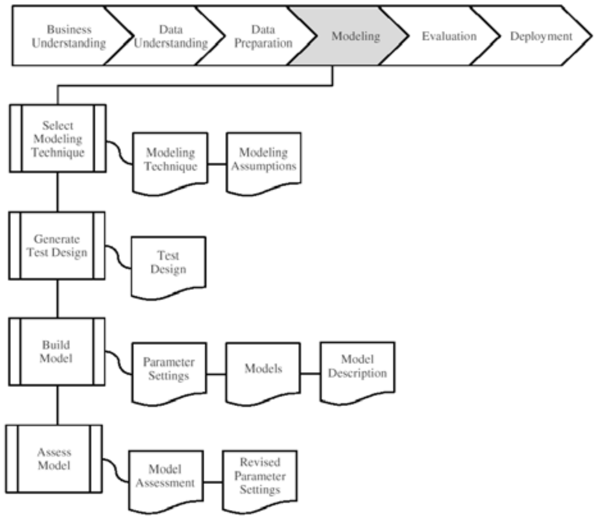
I.4.1 (iii) Data Preparation



I.4.1 (iii) Data Preparation

- **Select Data**
  - includes focusing
- **Clean Data**
  - correct false values
  - (insert suitable defaults)
  - (estimate missing values)
- **Construct Data**
  - define derived attributes (if needed)
  - normalize / transform single attributes (if needed)
- **Integrate Data**
  - combine data from different sources
  - be aware of syntactic / semantic inconsistencies
- **Format Data**

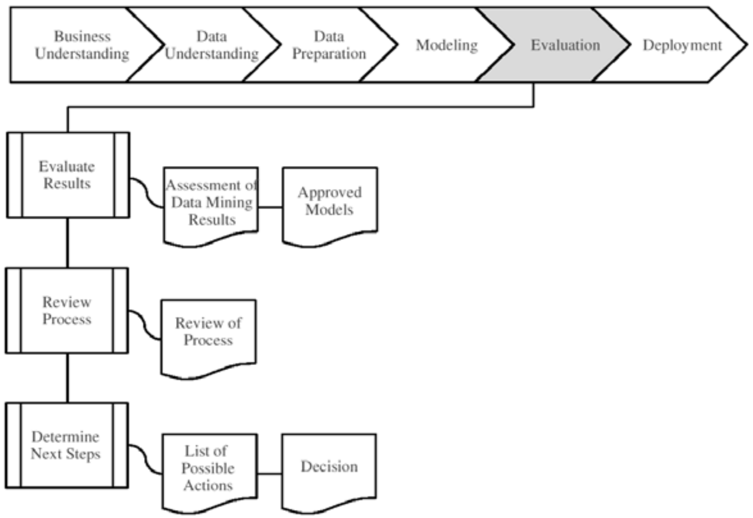
I.4.1 (iv) Modeling



I.4.1 (iv) Modeling

- **Select Modeling technique**
  - take into account:
    - experience with specific techniques
    - experience with specific tools
    - „political requirements“
- **Generate Test Design**
  - divide data sets into training data, test data and evaluation data
- **Build Model**
  - select appropriate parameter settings (typically, several iterations are needed)
- **Assess Model**
  - evaluate results with respect to data mining success criteria
  - check model against already known knowledge
  - revise parameter settings (if needed) and go back to „Build Model“
  - rank the generated models with respect to success criteria

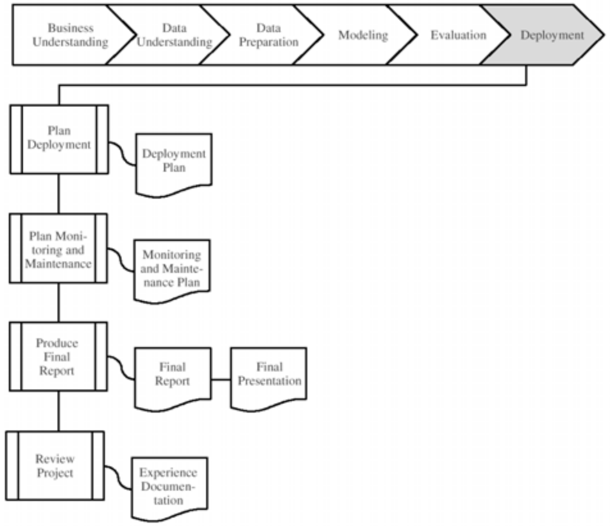
I.4.1 (v) Evaluation



I.4.1 (v) Evaluation

- **Evaluate Results**
  - evaluate results with respect to business objectives
  - what are other findings of the project (e.g. quality of available data should be improved)
- **Review Process**
  - identify failures
- **Determine Next Steps**
  - analyse potential for „Deployment“

I.4.1 (vi) Deployment



I.4.1 (vi) Deployment

- **Plan Deployment**
  - set up deployment plan
- **Plan Monitoring and Maintenance**
  - when should the model not be used any more?
  - will the business objectives change over time?
- **Produce Final Report**
  - what are target groups for final presentations?
- **Review Project**
  - summarize important insights and experiences
  - integrate review results into knowledge management strategy

I.5 Additional aspects

a) data privacy and security

- **The Application of KDD must not break laws like data privacy**
  $\Rightarrow$ refer to OECD Personal Privacy Guidelines

- **data privacy is very important while focussing**
  $\Rightarrow$ the reduction of examples must not allow to draw conclusions on single persons or small groups of persons

  - data must be made anonymous
  - use sufficient number of examples

I.5 Additional aspects

b) criteria to choose a KDD application

(i) application aspects:

- **KDD has to have strong (positive) effects on applications:**
  - **Business Applications**:
    higher turn-over, lower costs,
    higher quality, higher customer satisfaction
  - **Scientific Applications**:
    Access to huge amounts of data
    (readings data, satellite pictures) enables
    new insights.

(ii) Technical aspects:

- sufficient number of examples
- examples contain all relevant attributes
- quality of data is sufficient
  - little number of errors in values
  - little number of missing values
- appropriate algorithms are available
- is language bias of Data Mining-algorithm suiting to posed learning-question?
- possibility to score quality of learned knowledge

(iii) Rechtliche Aspekte:

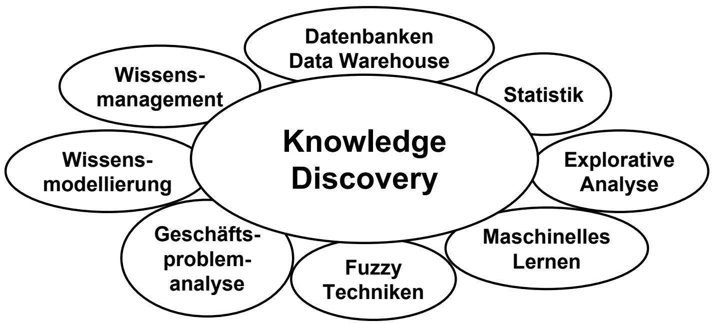
- ist Datenschutz gewährleistet?
- Erlaubt Wettbewerbsrecht Realisierung der Aktionen, die durch KDD-Resultate nahe liegen?

(iv) Personal- / Management Aspekte:

- Liegt explizite Managementunterstützung für den Einsatz neuer Methoden und Techniken vor?
  - Keine Erfahrung vorhanden
  - hoher Zeit- / Kostenaufwand
  - hohes Risiko
- Sind Anwendungsexperten verfügbar?
  - Was sind relevante Attribute?
  - Welche Beziehungen sind schon bekannt?

c) Querbezüge

- **KDD nutzt und integriert eine Vielzahl von Methoden und Techniken aus verschiedenen Gebieten:**



- **Data Warehousing:**
  - **Integration und Abstraktion von Unternehmensdaten aus verschiedenen Datenbanken**
  - **beinhaltet aktuelle und historische Daten**
  - **OLAP-Techniken (On-Line Analytical Processing) bieten flexible Möglichkeiten zur Datenverdichtung und –verfeinerung**
  - **Entscheidungsunterstützung**
  - **siehe Kapitel III dieser Vorlesung**

- **Wissensmanagement:**
  - **Knowledge Discovery sollte Teil einer Gesamtstrategie für das Wissensmanagement sein**
  - **Aufgaben- und Domänenwissen kann zur Verbesserung des KDD-Prozesses verwendet werden:**
    - Was sind potentiell relevante Konzepte und Zusammenhänge?
  - **Resultate des KDD-Prozesses müssen in strukturierten Ansatz im Unternehmen eingebettet werden**
    - Wer nimmt KDD-Resultate wie zur Kenntnis?