Der KDD-Prozess

Knowledge Discovery in Databases

Reihenfolge

1. Der allgemeine KDD-Prozess
   - Grundlegendes und Vertiefungen an einigen Stellen
2. Der KDD-Prozess durch das CRISP-Modell dargestellt
   - Darstellung als Kochbuch
3. Der KDD-Prozess durch Case-Based-Reasoning dargestellt
   - Anwenderfreundlichkeit für Endbenutzer
   - Wiederverwendung durch Case-Betrachtung

KDD-Prozess:

Wissensentdeckung in Datenbanken ist der nichttriviale Prozess der Identifikation gültiger, neuer, potentiell nützlicher und schlussendlich verständlicher Muster in (großen) Datenbeständen.

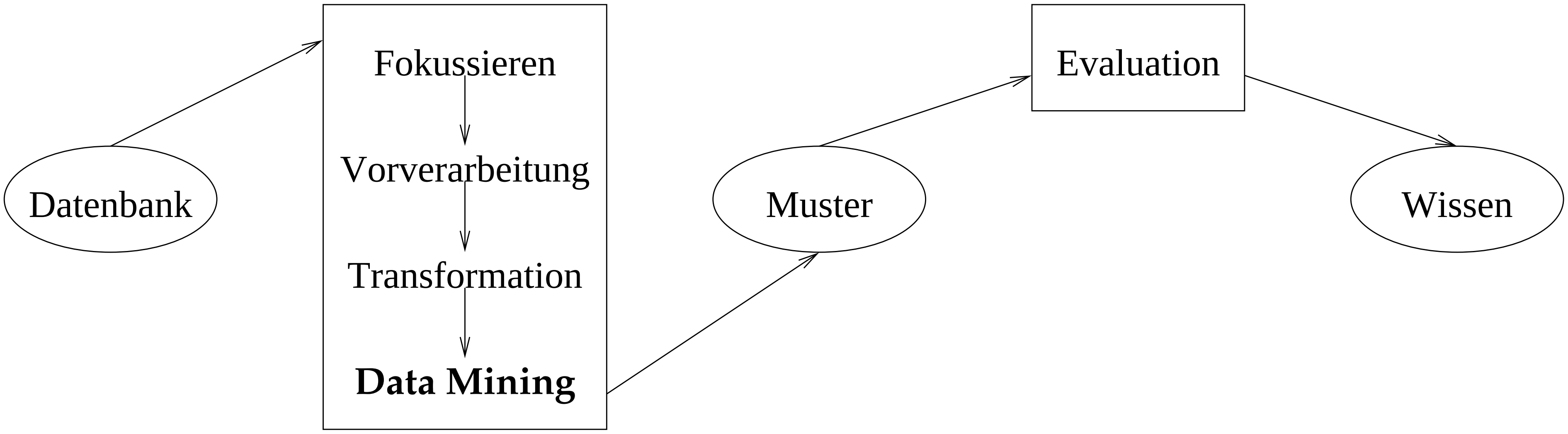
- Muster: kann als Wissen aufgefasst werden, Regeln, Assoziationen, Objektgruppierungen, Entscheidungsbäume, ...

- Gültigkeit: richtige Beschreibung durch die Daten

- Nichttrivialität: keine simplen Datenbankabfragen

- Datenbestand: meistens Tabellen, Spalten sind Attribute, Variablen oder Merkmale, Zeilen sind Datensätze, Objekte oder Instanzen

Allgemeine KDD-Prozess



Fokussieren

- Verständnis der Anwendung erlangen
  - Was wird gemacht?
- Verständnis des momentanen Anwendungswissens
- Ziele festlegen
  - Was soll gelernt werden?
- Datenquelle muss ausfindig gemacht werden

Vorverarbeitung

- Integration
  - Daten aus unterschiedlichen Quellen
  - verschiedene Konventionen für Daten
  - seien Körpergröße und Menschenlänge zwei verschiedene Attribute
  - ein Arzt misst den Blutdruck 10mal am Tag, ein anderer nur 3mal

Vorverarbeitung

- Konsistenz
  - keine verschiedenen Werte für dasselbe Attribut
  - Schreibfehler für Namen dürfen nicht auftreten
  - keine verschiedenen Date-Formate für ein Attribut

Vorverarbeitung

- Vervollständigung
  - Fehlende Werte: Der Wert für die mögliche Variable wurde nicht eingegeben. Dieser Wert kann jedoch eingegeben werden. Vom Benutzer vergessen.
  - Leere Werte: Es kann kein Wert für diese Variable eingegeben werden. Schwarz und weiß aber nicht grau.
  - Hier nur Betrachtung von **fehlenden Werten.**
    - Ersetzungen der fehlenden Werte soll keine Verzerrung darstellen.
    - 1,2,3,x,5 ist Zahlenfolge. Wenn x=100, dann passt x nicht in den Kontext.
    - Zur Kontexterhaltung: Datenmenge als Ganzes sehen

Möglichkeit zur Ersetzung: **unvoreingenommener Schätzer**

- 1,2,3,x,5 ist gegeben
- Sei der Mittelwert der unvoreingenommene Schätzer $\Rightarrow$ x=2.75
- Sei die Standardabweichung der unvoreingenommene Schätzer $\Rightarrow$ x=4.659
- Der Benutzer muss wissen, welcher Schätzer den wenigsten Schaden anrichtet.
- Es existieren natürlich noch andere Berechnungen für Schätzer.

Transformation

- Transformation der vorverarbeiteten Daten im Hinblick auf das Ziel des KDD-Prozesses

- und auf das anstehende Lernverfahren (Data-Mining).

Hier Unterteilung in Diskretisierung und Feature Construction, Extraction and Selection.

- Diskretisierung
  - viel tausend Werte können auf eine geringe Anzahl von Intervallen reduziert werden -> Datenreduktion
  - erst durch Diskretisierung ist überhaupt Zugang zu speziellen Lernverfahren möglich

Diskretisierung

- Beispiel:
  - Attribut Farbe=[1,10]
  - 1=weiß, 10=rot, $2.\overline{33}$=helles rot
  - nach der Diskretisierung nur noch kategorische Werte
  - Farbe={weiß, hellrot, mittelrot, dunkelrot}
  - weiß= [0,1], hellrot=[2,4],...,rot=[9,10]
  - es existiert nun eine sinnvolle Vergröberung der numerischen Attribute

Equal-Width

- Gegeben: Anzahl N der Werte und x_{min} und x_{max} bzgl. der gesamten Werte
- Der Benutzer bestimmt die Anzahl n der Intervalle
- Gesucht: Die einzelnen Trennpunkte $T_i = x_{min} + \frac{i(x_{max} - x_{min})}{n}$ mit i=0...n
- Es ergeben sich folgende Intervalle: $[T_i, T_{i+1}]$
- Vorteil: Einfaches Vorgehen und gute Laufzeit, weil nur x_{min} und x_{max} nötig sind.
- Nachteil: Keine Berücksichtigung der Verteilung der Werte, es wird eine Gleichverteilung der Werte über den ganzen Bereich angenommen.

NUM_INT

- Voraussetzung: sortierten Wertefolge
  - aufsteigend Suche nach Trennstellen
- Auffassung des Begriffs Trennstelle $\Rightarrow$ 2 Modi
- Der Unterschied zu Equal-Width besteht darin, dass nun auch die Anordnung der Werte zur Intervallbildung hinzugezogen wird.

1. Gap:
   - Lücken suchen
   - größte Differenz zweier aufeinander folgender Werte feststellen
   - das initiale Intervall teilen
   - neuen Intervalle auch so betrachten

2. Discontinuity:
   - kleiner Wertebereich im vergleich zur Anzahl der Werte
   - Betrachtung der Häufigkeitsverteilung der Werte
   - Diskontinuität in Bezug auf die Anzahl der Tupel pro Wert kann als Indiz für eine geeignete Trennstelle gewertet werden

Feature Constructin, Extraction and Selection

- Merkmalskonstruktion (feature construction)
  - es existieren zu wenig aussagekräftige Merkmale
  - Merkmalskonstruktion
  - hier werden fehlende Relationen ausfindig gemacht
  - neue Merkmale werden kreiert und zu den Originalen hinzugefügt
  - Merkmalsanzahl steigt

- Merkmalsextrahierung (feature extraction)
  - durch eine Funktion entsteht eine neue Menge von Merkmalen
  - es werden nur die Merkmale benutz, welche durch die Funktion extrahiert wurden
  - andere Merkmale fallen weg
  - weniger Merkmale durch die Funktion

- Merkmalsauswahl (feature selection)
  - nicht alle Merkmale sind für ein bestimmtes Lernverhalten nötig
  - Merkmalsauswahl
  - weniger Merkmale
  - erzeugt keinen neuen Merkmale

Voraussetzung für Feature construction, extraction and selection

- das Lernverfahren muss mit einbezogen werden

- es ist ein Zusammenspiel zwischen Transformation und Data-Mining

Data Mining

Data Mining ist die Anwendung effizienter Algorithmen (Lernalgorithmen), die die in einer Datenbank enthaltenen gültigen Muster finden.

- Clusteringverfahren: Partionierung in Gruppen (Cluster) von Objekte
- Klassifikationsverfahren
  - Entscheidungsbaumverfahren: Bäumen, mit denen sich kategorische Variablen vorhersagen lassen
  - Regellernverfahren: Wenn-Dann-Regeln werden aus den Daten erzeugt
  - Regressionsverfahren: numerisches Vorhersageverfahren

- Assoziationsregelverfahren: 80% der Käufer von A und C, kaufen auch B, D und E

- Probabilistische Netze: Graph mit Kantenbewertung zur Abhängigkeit von Variablen

Evaluation

- Die erkannten Muster sollen, in Bezug auf ihre Zielsetzung, bewertet, gesäubert und visualisiert werden.

KDD-Anwendungen

- Kundensegmentierung
  - Profile von Kunden gewinnen.
- Kündigungsanalyse
  - Mögliche Kündigungsabsichten sollen vorhergesagt werden.
- Kreditwürdigkeitsentscheidungen
  - Soll der Kunde einen Kredit bekommen?

- Warenkorbanalyse
  - Klassische Anwendung der Assoziationsregeln.
- Medizin
  - Nebenwirkungen erkennen und Diagnosen erstellen.
- Antwortvorhersagen
  - Auswahl von Kunden für die direkte Ansprache, hier soll die positive Antwort des angeschriebenen Kunden am wahrscheinlichsten sein

CRISP-Modell

- KDD-Prozess zur Zeit mehr eine Kunst, als ein gut verstandener, zuverlässiger Prozess

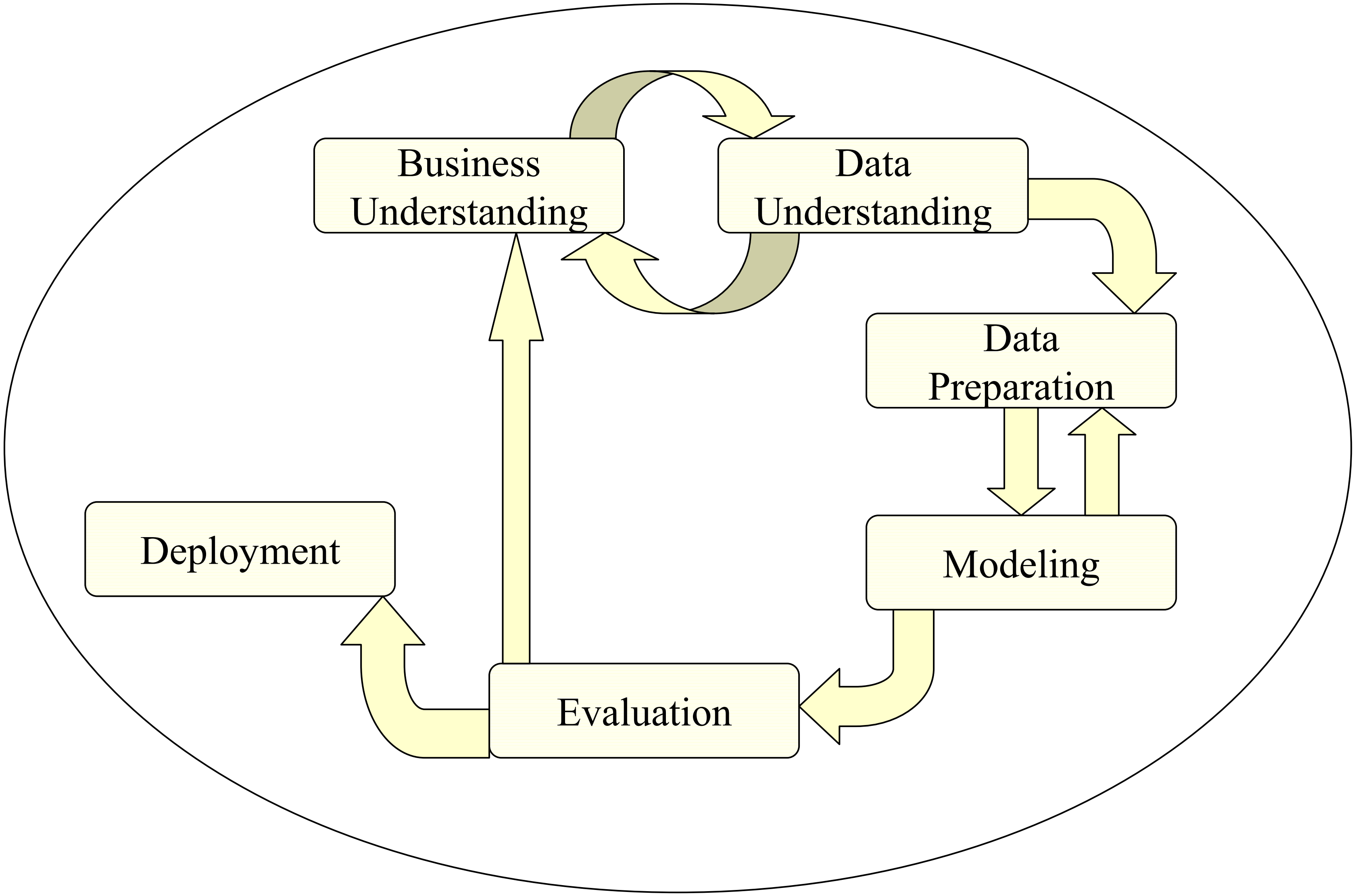
- CRISP-DM - "CRoss-Industry Standard Process for Data Mining" - soll abhilfe schaffen

Die einzelnen Schritte

- Business Understanding
  - Problemdefinition
  - Verständnis des Projektziels

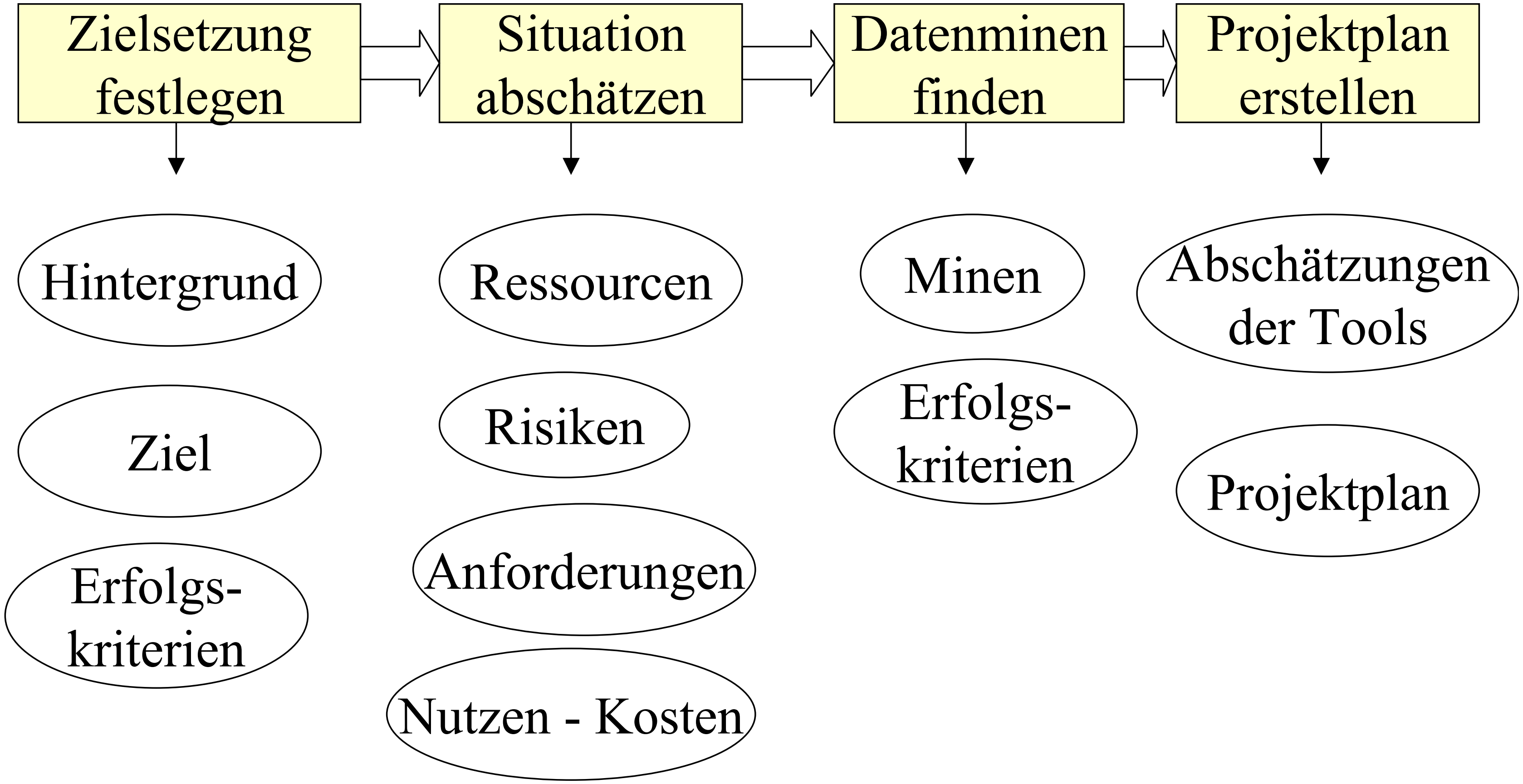
- Data Understanding
  - mit den Daten vertraut zu werden
  - Datenqualitätsprobleme identifizieren
  - Teilmengen von Daten erkennen
- Data Preparation
  - Datenversorgung
  - Attributsauswahl
  - Transformationen fürs Modeling

CRISP-Modell



- Modeling:
  - Lernalgorithmen aussuchen und anwenden
- Evaluation:
  - Abgleich mit dem Ziel im Business Understanding
- Deployment:
  - Organisation des neuen Wissens für den Anwender

Zur Verdeutlichung der Präzision des CRISP-Modells. So genau wie ein Kochbuch. Macht es einfacher für den Anwender.



KDD-Prozess anhand des Case-Based-Reasoning Grundgerüsts

Problem

- heterogene Daten
- kein universelles Mining-Tool
- Anpassung an das spezielle Tool ist nötig

Abhilfe

- Case-Based-Reasoning, als Grundgerüst für den KDD-Prozess
- Case: Ansammlung von Businessproblemen und Daten
- jeder Case besitzt Vorverarbeitung und passende Analysemethoden
- Wiedereinsatz für einen ähnlichen Fall wird einfacher

Endbenutzer und Experte

- Endbenutzer verwendet lediglich vorhandene Cases
  - er muss bekannte Cases wiedererkenne
  - er kann einfache Anpassungen machen
- Experte erstellt neue Cases für den Endbenutzer
  - greift natürlich auf Vorhandene zurück und übernimmt ganze Analyseketten
  - es ergibt sich auch hier eine Beschleunigung für den Experten

Datenhaltung

Data Warehouse (DWH)

- normalisierte, multi-relationale Datenbank als Datenquelle, wohingegen die meisten existierenden Lernalgorithmen auf einer Tabelle basieren

- viele Merkmale (features) und Tabellen, welche nur teilweise für den gegenwärtigen Case relevant sind

- DWH enthält auch die bereits bearbeiteten Cases und deren Arbeitsschritte