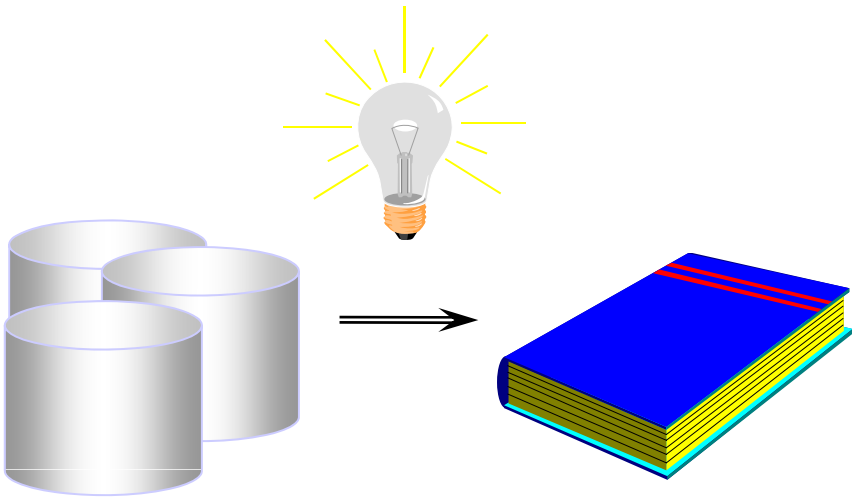
Einführung Data Mining

Praktische Anwendungen für automatisierte und lernende Informationsverarbeitung

Wolfgang Konen, FH Köln

November 2009

Data Mining (DM): Entdecken von Wissen in Datenbanken



- Unternehmen und Institutionen sammeln ungeheure Datenmengen
- Data Mining: Identifikation von wettbewerbsrelevantem Wissen aus grossen Datenbanken
- Automatische Erkennung von **Mustern**
  - nicht-trivial
  - bisher unbekannt
  - potentiell nützlich

  [Fayyad, 1996]

Analysebeispiel: Mobiltelefonie

©Plambeck/ PIXELIO

Ein Problem ...

- Sie sind Marketingmanager im Mobilfunk
  - Problem: zu viele Abgänge bei Vertragsende: 40%
  - Neukunden erhalten kostenlos ein Telefon
  - Ihr Unternehmen zahlt 250 EUR Provision pro Abschluss
  - Jedem Kunden bei Vertragsende eine neues Telefon zu geben ist zu teuer

Eine Lösung ...

- Drei Monate vor Vertragsende vorhersagen, welche Kunden nicht verlängern
  - Den Kunden, die man behalten will, bietet man ein neues Telefon an

Wie kann man künfiges Verhalten vorhersagen?

- Kartenlegen?
- Würfeln?
- **Data Mining**?

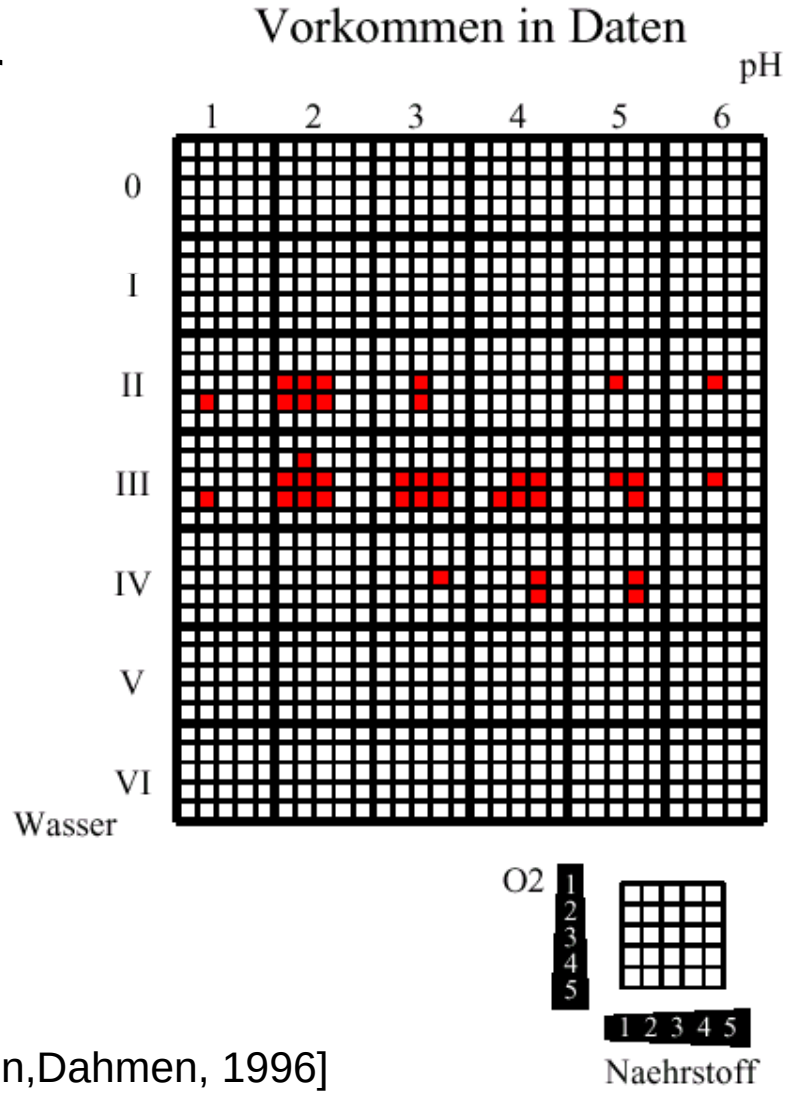
Beispiel aus der Automobilindustrie

- Daten über 7 - 10 Jahre Historie für 7 Mio. Fahrzeuge
  - Fahrzeugdaten (Produktionsdaten; Daten über Motor, Getriebe, ...)
  - Beanstandungen (Schadensteil, Schadensart, ...)
  - Werkstattaufenthalte
- Frage: Wie kann man das Auto zuverlässiger machen?
- Mustererkennung: Suche in Datenbank nach möglichen Gründen für Ausfälle
- Umsetzung des Wissens:
  - Änderung in Konstruktion
  - Wechsel des Zulieferers
  - Kundendienst: vorbeugende Wartung
  - usw.

© Glathe / PIXELIO

Beispiel Umweltanwendung: Data Mining für ökologische Standortbewertung Pflanzen

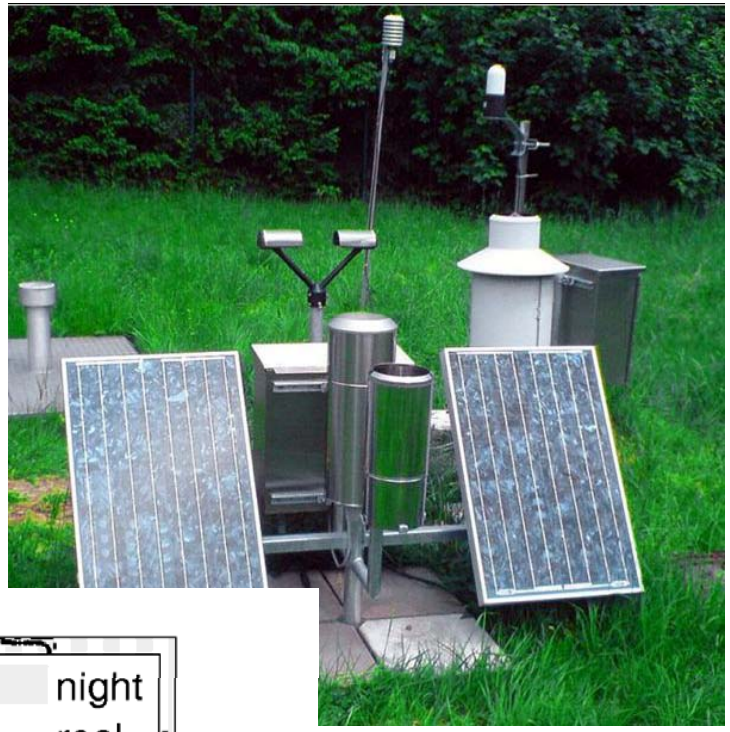
- „Wo wächst was?“ – relevant für Umweltverträglichkeitsprüfung
- bisher: aufwendig manuell erstellte Ökodiagramme
- multifaktorielle Daten:
  - Wasser, Säuregehalt, Nährstoffe, Lichtintensität,...
- mit Data Mining können vergleichbare Ergebnisse umfassender, schneller (aktueller) und kostensparender erzielt werden

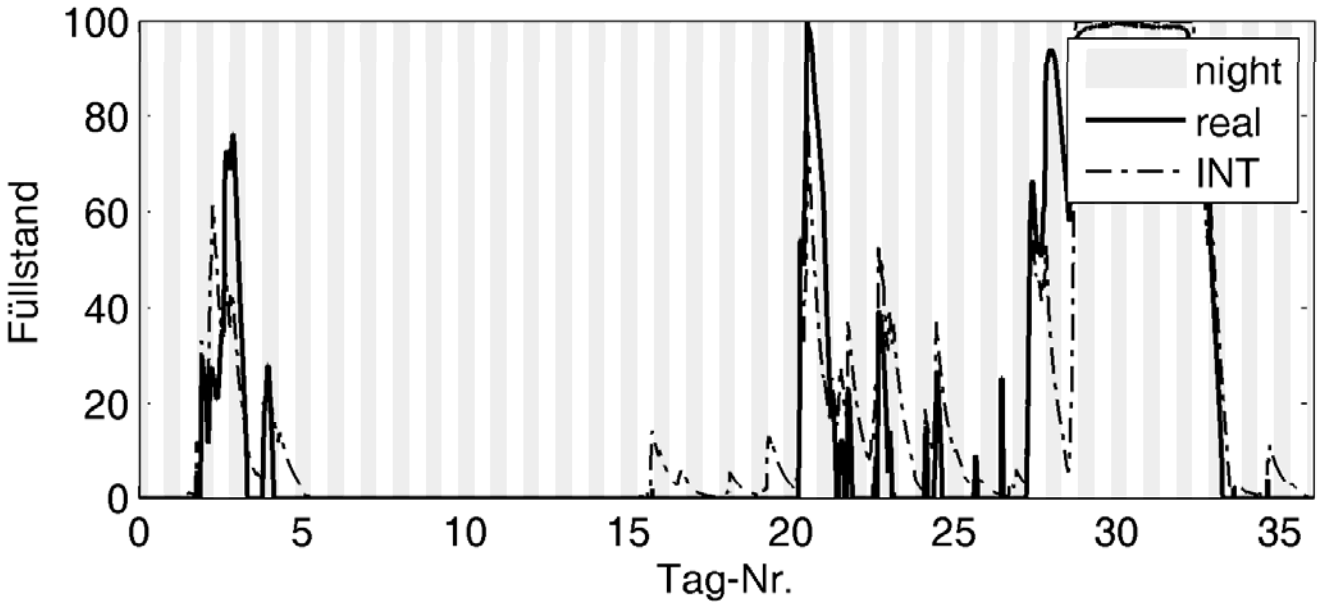


[Kirsten,Wrobel,Dahmen,Dahmen, 1996]

Beispiel Wassermanagement

- Vorhersage Füllstand RÜB (Regenüberlaufbecken)
- Modellbildung für Forecast, Output abhängig von
  - Regenmenge
  - Bodenzustand & Grundwasser (hidden states)





[Konen, Zimmer, Bartz-Beielstein, 2009]

Weitere Anwendungsbeispiele

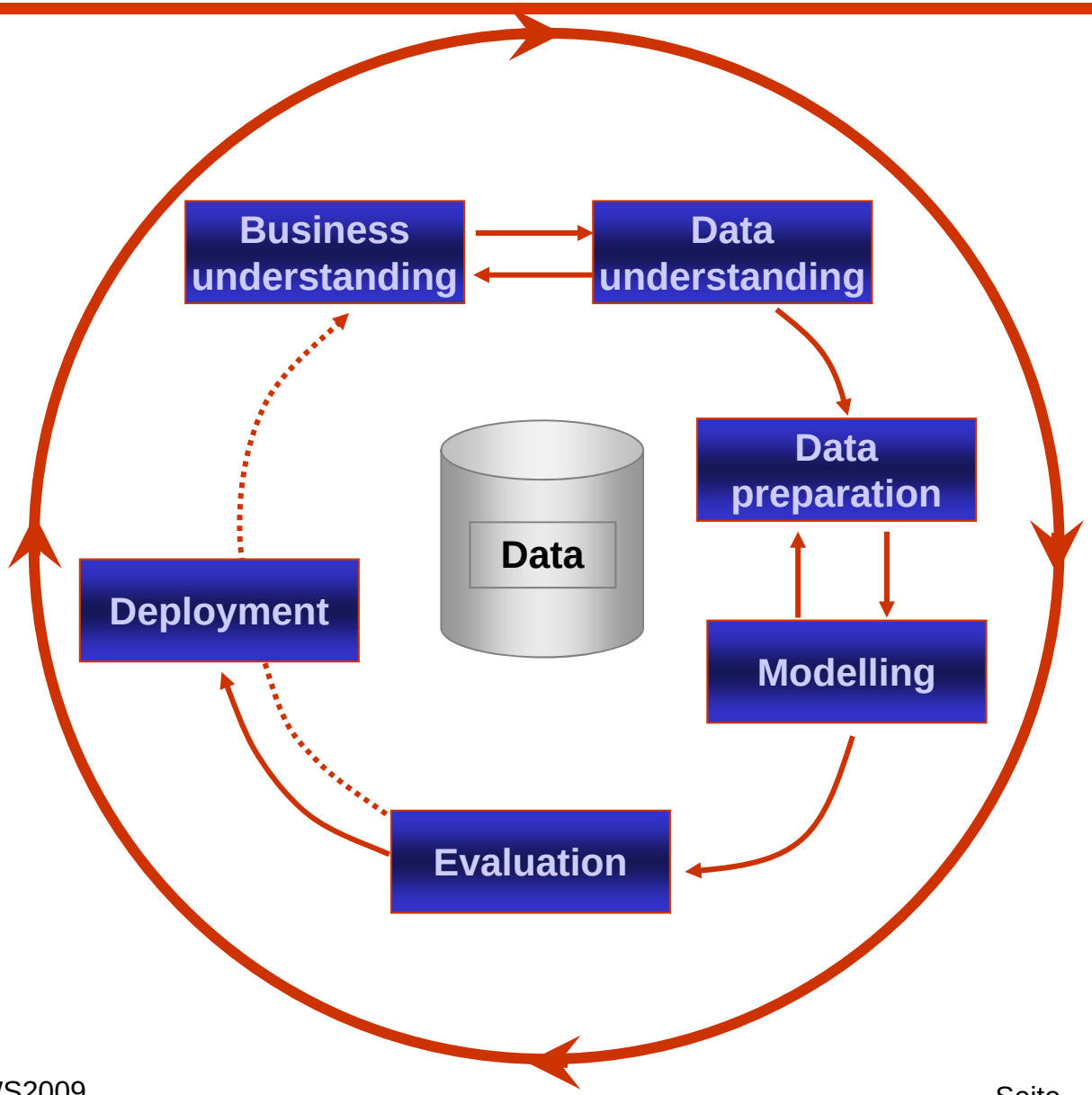
- Betrugserkennung
  - Beispiel: Erkennung typische Muster zur Identifikation von Kreditkartenbetrug.
- Kreditbeurteilung
  - Identifikation von Kriterien für Kreditwürdigkeit von Kunden
- Nachfrageprognose
  - Wieviele Einheiten von Produkt X setzen wir in der KW42 ab?

⇒Computerwoche 03/2007: „Zweiter Frühling Data Mining“
http://www.computerwoche.de/produkte_technik/business_intelligence/590688

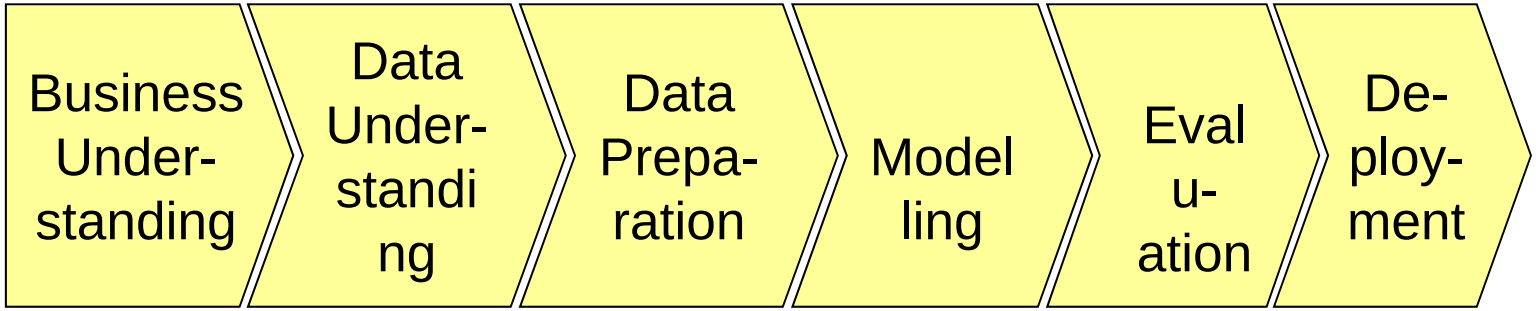
Der CRISP-DM Standard

- CRISP-DM: Cross-Industry Standard Process for Data Mining
- Entwickelt im Rahmen eines EU-Projekts von 1996-99
  - Partner: DaimlerChrysler (Deutschland)
    NCR Systems Copenhagen (USA, Dänema
    OHRA Bank Groep B.V. (Niederlande)
    SPSS Inc. (USA)
  - Gründung einer Special Interest Group
- Der CRISP-DM 1.0 Report beschreibt
  - die CRISP-DM Methodology
  - das CRISP-DM Referenzmodell
  - den CRISP-DM User Guide
  - den jeweiligen Resultate/Reports der einzelnen Phasen
- Für Informationen zu CRISP-DM siehe http://www.crisp-dm.org

Das CRISP-DM Referenzmodell

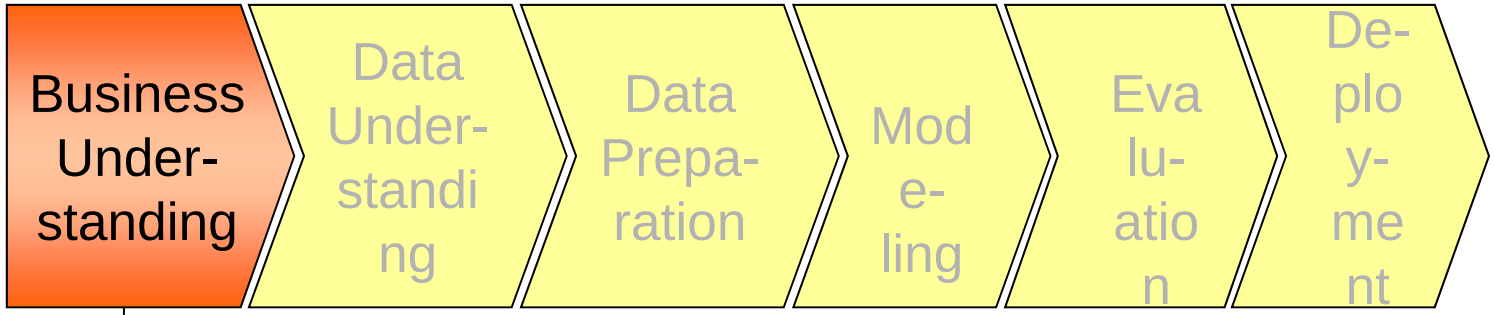


Die 6 Schritte des KDD-Prozesses



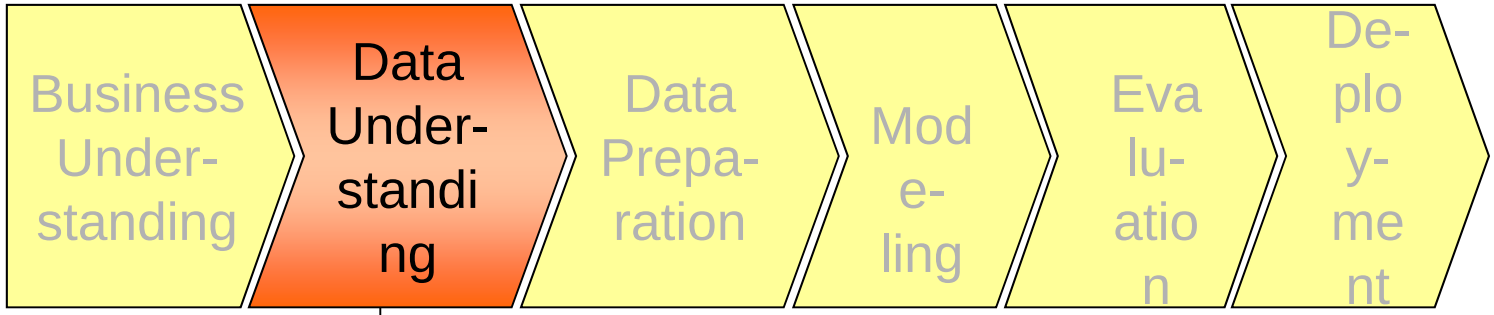
- Wichtig: Jeder Schritt ist zu dokumentieren!

1. Schritt: Geschäftsmodell verstehen



- Geschäftsmodell / -ziele verstehen (z.B. Abwanderung von Kunden verhindern)
- Ressourcen, Zeit, Risiken, Chancen
- Genaue Spezifikation der Ziele (z.B. 70% der Abwanderer erkennen bei 20% Fehlalarmen)
- Projektplan mit Meilensteinen

2. Schritt: Daten verstehen



- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality

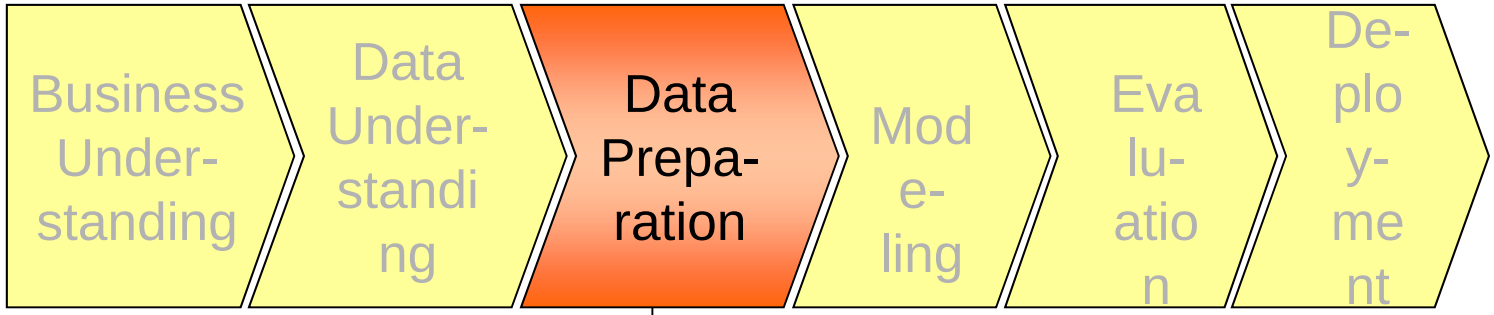
Wo kommen die Daten her? Wie? Joins über mehrere DBs notwendig?

Metadaten (Anzahl Attribute, Werte, Format,Typen, Mengen)

Beispiele anschauen, Visualisierung (z.B. Verteilungen, Korrelationen)

Datenqualität bestimmen, Fehler erkennen, Vollständigkeit

3. Schritt: Daten aufbereiten



- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format & TransformData

Auswahl der (wichtigen) Variablen, ggf. Sampling der Records

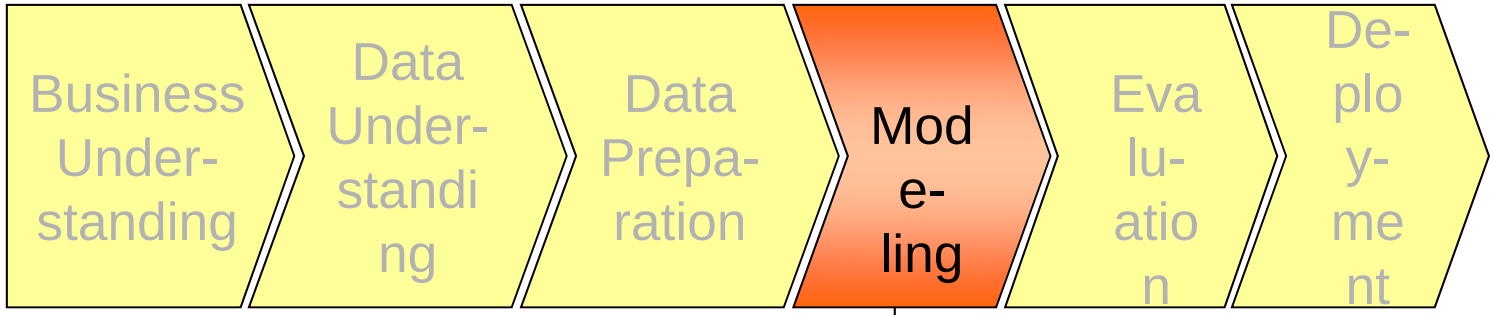
Ausreisser erkennen, fehlende Werte behandeln

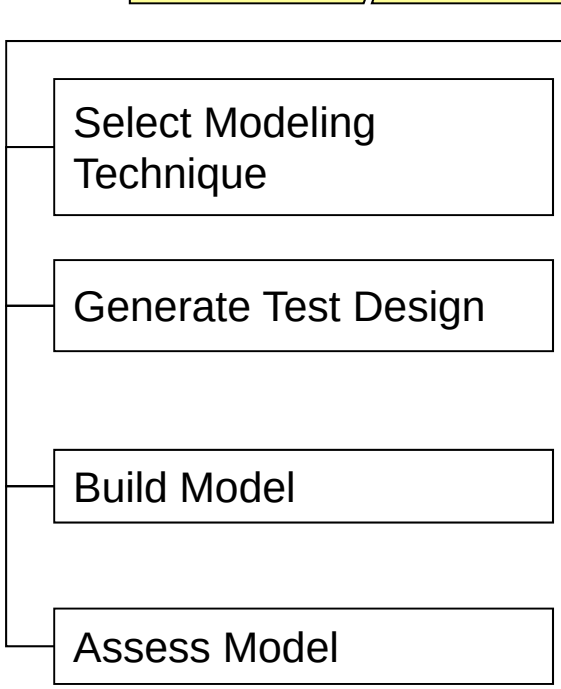
Abgeleitete Variablen (z.B. Summe oder logische Verknüpfungen)

Daten aus verschiedenen Tabellen zusammenführen

Formatierung (z.B. Datum), Transformation (z.B. normierte Verteilung oder PCA)

4. Schritt: Modellierung





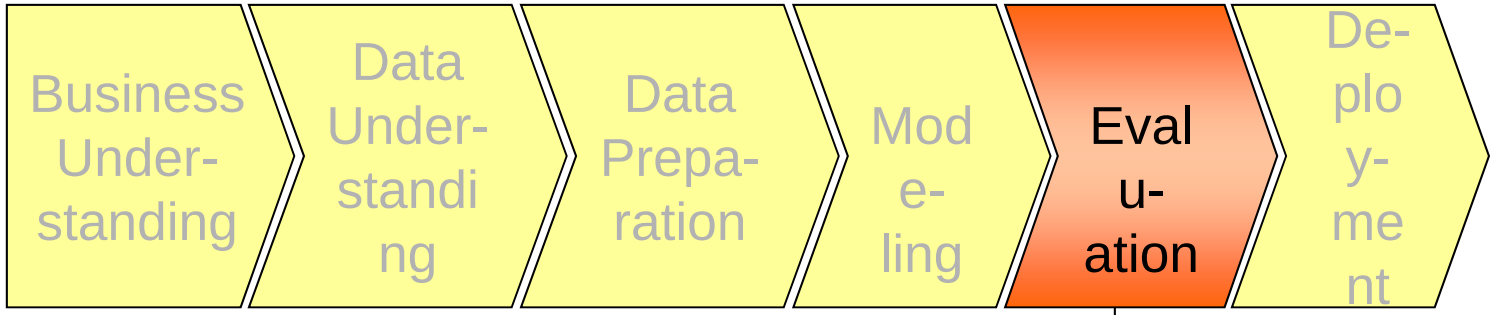
Auswahl der Methode, ggf. mehrfach

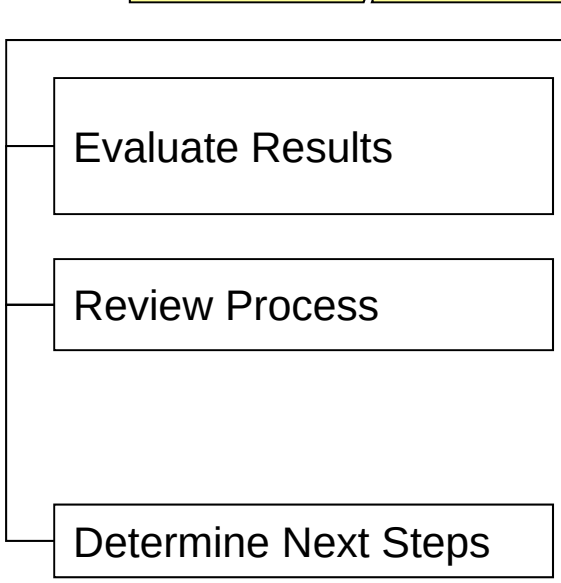
Wie messen wir Modellgüte?
Aufteilung der Daten in Trainings-, Test- und Validierungsmenge

Modell bauen, Parameter einstellen und begründen

Technische Bewertung der Modellgüte, ggf. mit anderen Parametern wdh.

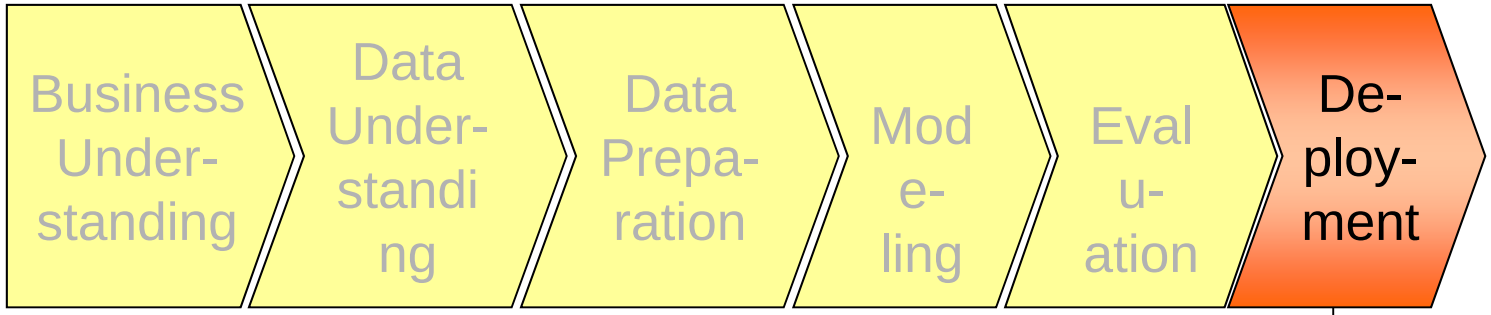
5. Schritt: Evaluation





- Bewertung aller Resultate in Bezug auf betriebswirtschaftliche Ziele
- Begutachtung aller Schritte. Wurden nur Daten verwendet, die auch in Zukunft verfügbar sind? Was wurde übersehen?
- Wie geht es weiter?

6. Schritt: Einsatz



- Plan Deployment
- Plan Monitoring and Maintenance
- Produce Final Report
- Review Project

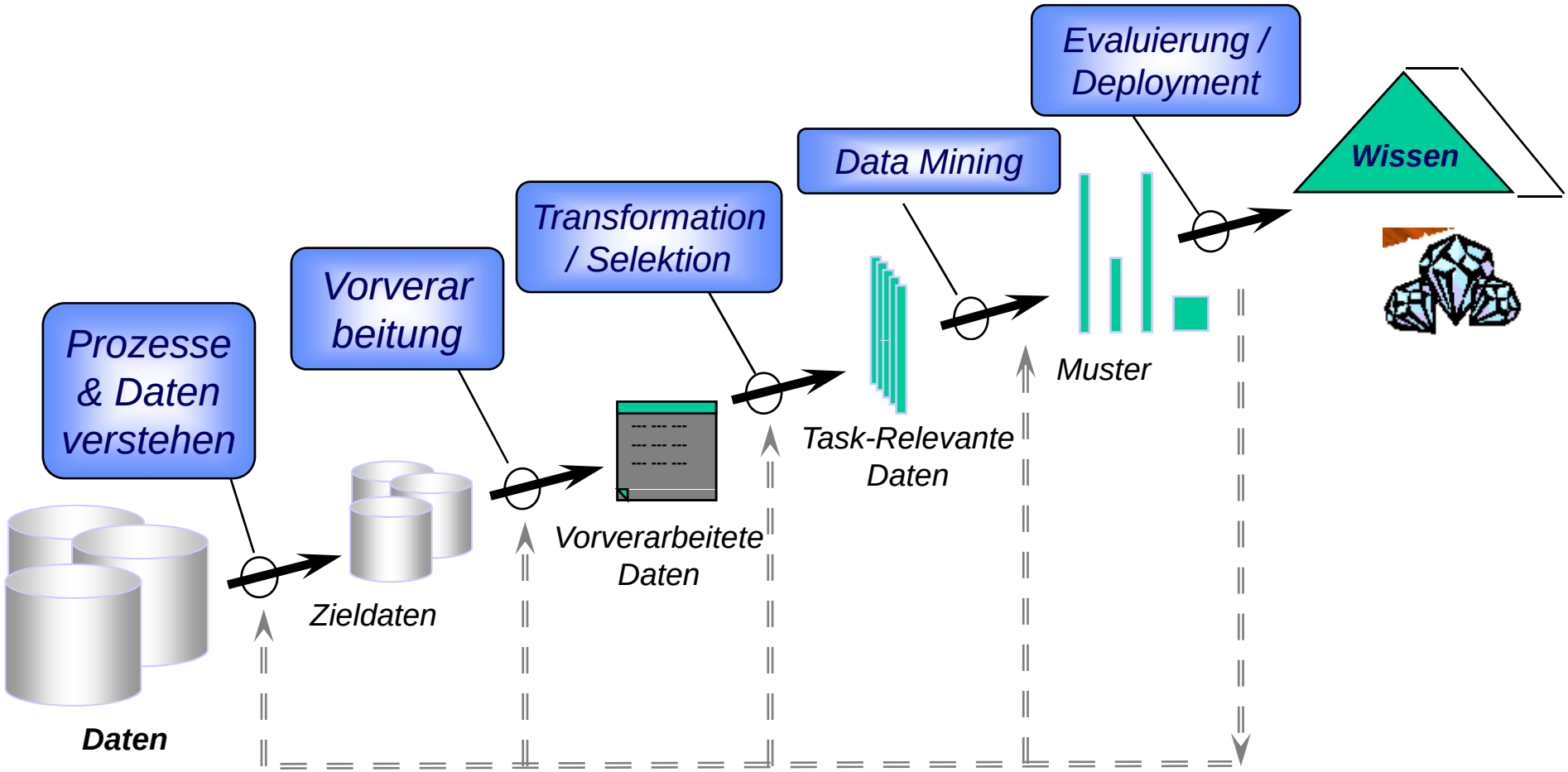
Implementierungsstrategie, Einsatz DM in Praxis, ggf. Mitarbeiterschulung

Überwachung der Gültigkeit der Modelle, Strategie für Überarbeitungen

Zusammenfassender Bericht / Präsentation

Begutachtung des Gesamtprojektes. Lessons Learned (f. weitere Projekte): Was lief schief? Was lief gut?

Data Mining ist eine Phase im Prozess der Wissensentdeckung aus Datenbanken



Prozeßschritte der Wissensentdeckung

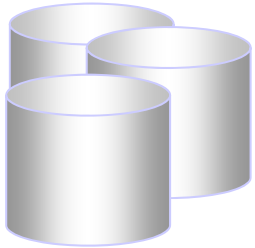
- Prozesse und Daten verstehen
  - Geschäftsprozesse mit den Anwendern besprechen
  - Fokussierung, Auswahl relevanter Daten
- Vorverarbeitung / Transformation
  - Bereinigung der Daten, z.B. was tun mit leeren Feldern,
  - Verdichtung der Daten durch Datenreduktion und -projektion
  - Datenformat vereinheitlichen
- Data Mining
  - Auswahl von Techniken und Methoden
  - evtl. viele Testläufe mit verschiedenen Parametern
- Evaluierung
  - Beurteilung der Ergebnisse bzgl. festgelegter Kriterien
  - Dokumentation, Visualisierung der Ergebnisse
- Deployment
  - Überführung in die Anwendung

Warum Data Mining?

- Vereinfachung und **Automatisierung** statistischer Prozesse
  - Datenanalyse
  - Anwendung der Modelle
- Bessere, verständlichere Modelle
- Data Mining bringt viele Verfahren und Werkzeuge
- Data Mining bringt erprobte statistische Verfahren ins Spiel (wo vorher evtl. noch wenig Statistik vorkam)
  - Data Mining Verfahren basieren auf statistischen Verfahren

Welche Formen von Daten kann Data Mining analysieren?

- strukturierte Daten (Data Warehouses, DWH)
  - z.B. Zeit-/Messreihen, Kundendaten

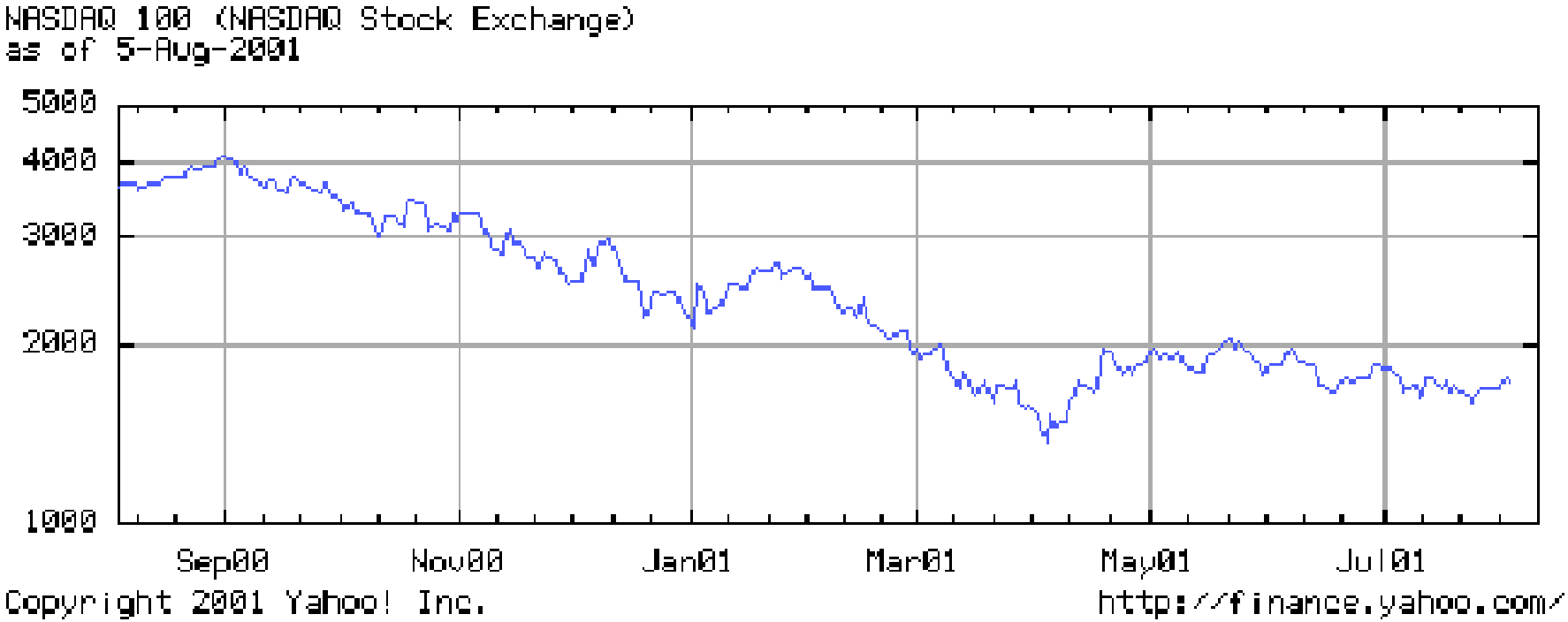


- Multimedia-Daten (Bild, Ton)

- Geografische Daten (GIS, Spatial Databases)

Welche Formen von Daten kann Data Mining analysieren?

- Zeitreihen

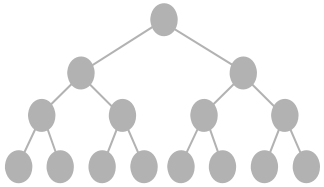


Welche Formen von Daten kann Data Mining analysieren?

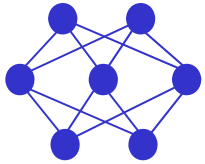
- Textdokumente (Text Mining)

- Web Mining (z.B. Klickpfad-Analysen)

Wichtige Methoden des Data Mining

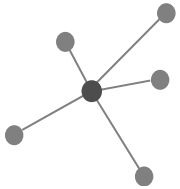


Entscheidungsbäume

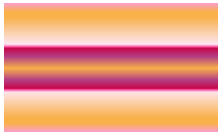


Neuronale Netze

Regelinduktion



Nearest Neighbor



Genetische Algorithmen

Taxonomie von Aufgaben im Data Mining (Auszug)

- **Klassifikation**
  - Lernen, die Datensätze aufgrund ihrer Attribute vorgegebenen Klassen zuzuordnen
  - Beispiel: Klassifikation von Bankkunden ob kreditwürdig oder nicht
- **Regression und Vorhersage**
  - Modellierung einer kontinuierlichen Outputgröße, z.B. zeitliche Veränderungen
  - Beispiele: Zeitreihenanalyse, Aktienkurse, Vorhersage von Systemausfällen auf Basis von Messdaten
- Clustering
  - Aufteilung einer Datenmenge in disjunkte Gruppen ähnlicher Objekte
  - Beispiel: Gruppen von Autos mit ähnlichen Ausfällen, Kundengruppierung für CRM (customer relationship mngmt)

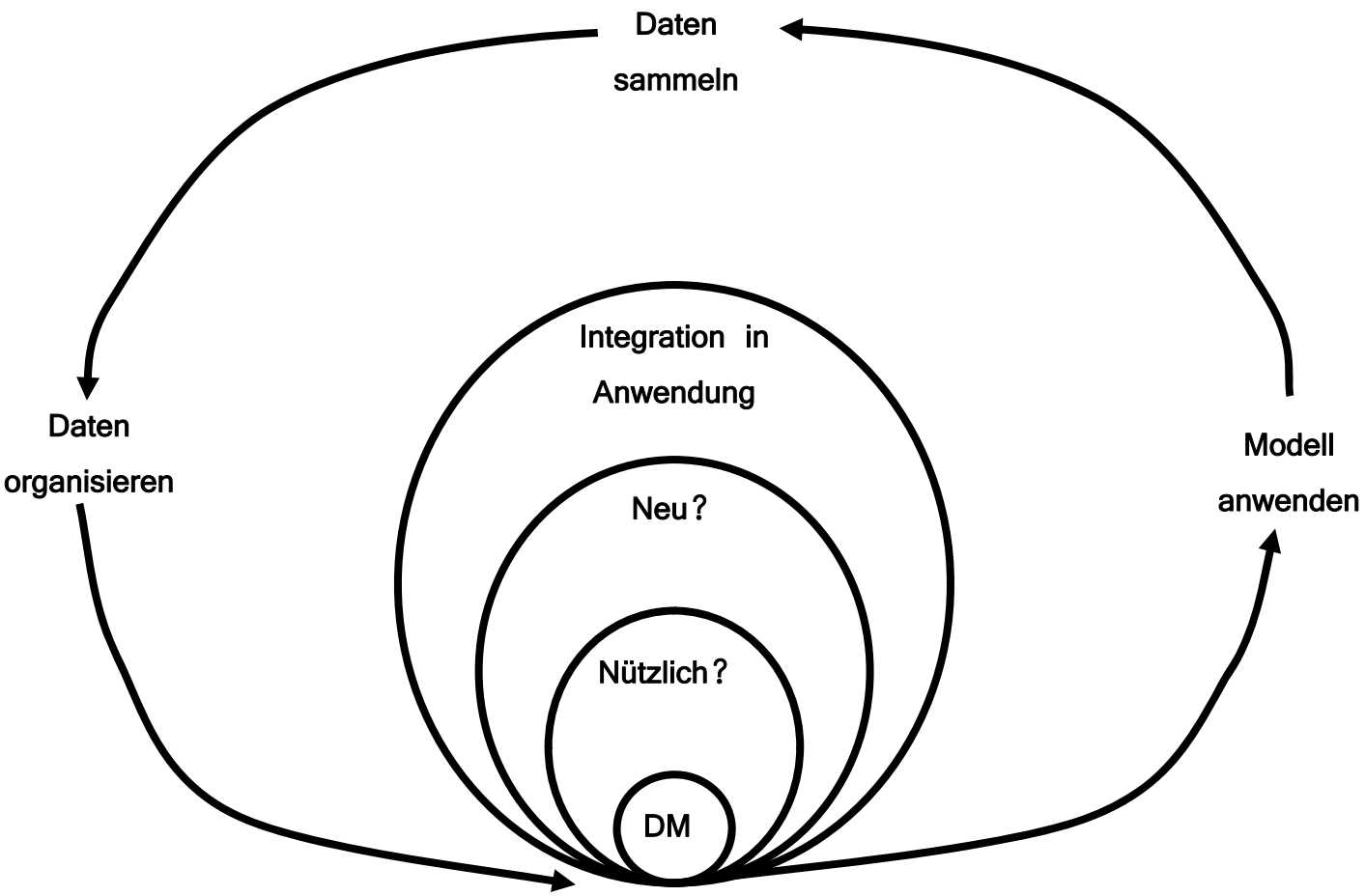
Übungen

- Welches sind die 6 wichtigen Phasen in einem Data Mining Projekt?

- Welche Typen (Formen) von Daten?

- Welche Gründe können den Einsatz von Data Mining motivieren?

Technologie ist nicht alles



Technologie ist nicht alles

- Data Mining ist ein kreativer Prozess
  - es reicht nicht: Datenbank + Softwaretool = fertig

- Wichtig für erfolgreiche Data-Mining-Projekte
  - gesunder Menschenverstand bei der Prüfung der Daten auf Plausibilität / Validität
  - gutes Verständnis der Prozesse, der Teilnehmer und deren **Ziele**
  - dies alles **kreativ** einbauen in die Modellbildung, die für aktuellen Prozess betrieben wird

⇒ *Data Mining macht Spass !*