Oberseminar Data Mining
07. April 2010
Methodik des Data Mining

Knowledge Discovery In Databases

oder auch

Data Mining - Der moderne Goldrausch?

Data Mining...?

Geo-informations-systeme

Digitale Signalverarbeitung

Data Warehouse

Mustererkennung

Hochleistungs-rechnen

Information Retrieval

Maschinelles Lernen

Neuronale Netze

Statistik

Data Mining vs. KDD

Data Mining bezeichnet die Auswertung vorhandener Daten mit dem Ziel, bisher nicht explizit hergestellte Zusammenhänge offenzulegen (*Knowledge Discovery*).
[Herden 2007]

Data Mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data.
[Decker, Focardi 1995]

Knowledge Discovery in Databases describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
[Fayyad et al. 1996]

Arbeitsdefinition

Knowledge Discovery in Databases ist der Prozess der Gewinnung von Wissen, von den Rohdaten bis hin zu den Zusammenhängen.

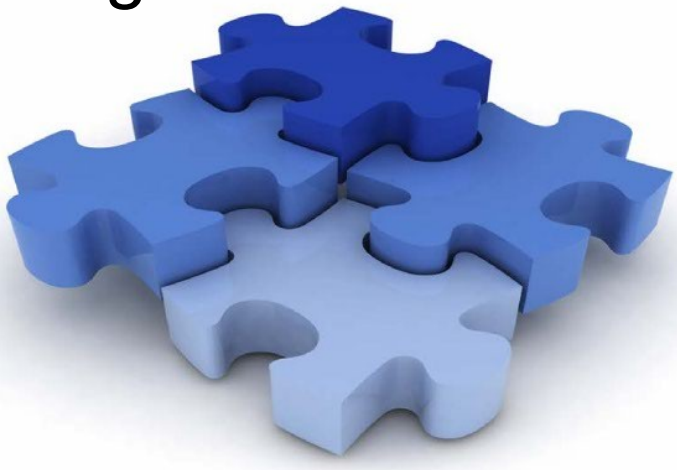
Data Mining ist ein Teilschritt, der nützliche und neuartige Muster in Daten identifiziert.

(*Muster* = Beziehungen zwischen Entitäten, räumliche Zusammenhänge, zeitliche Verläufe, mathematische Gesetzmäßigkeiten, …)

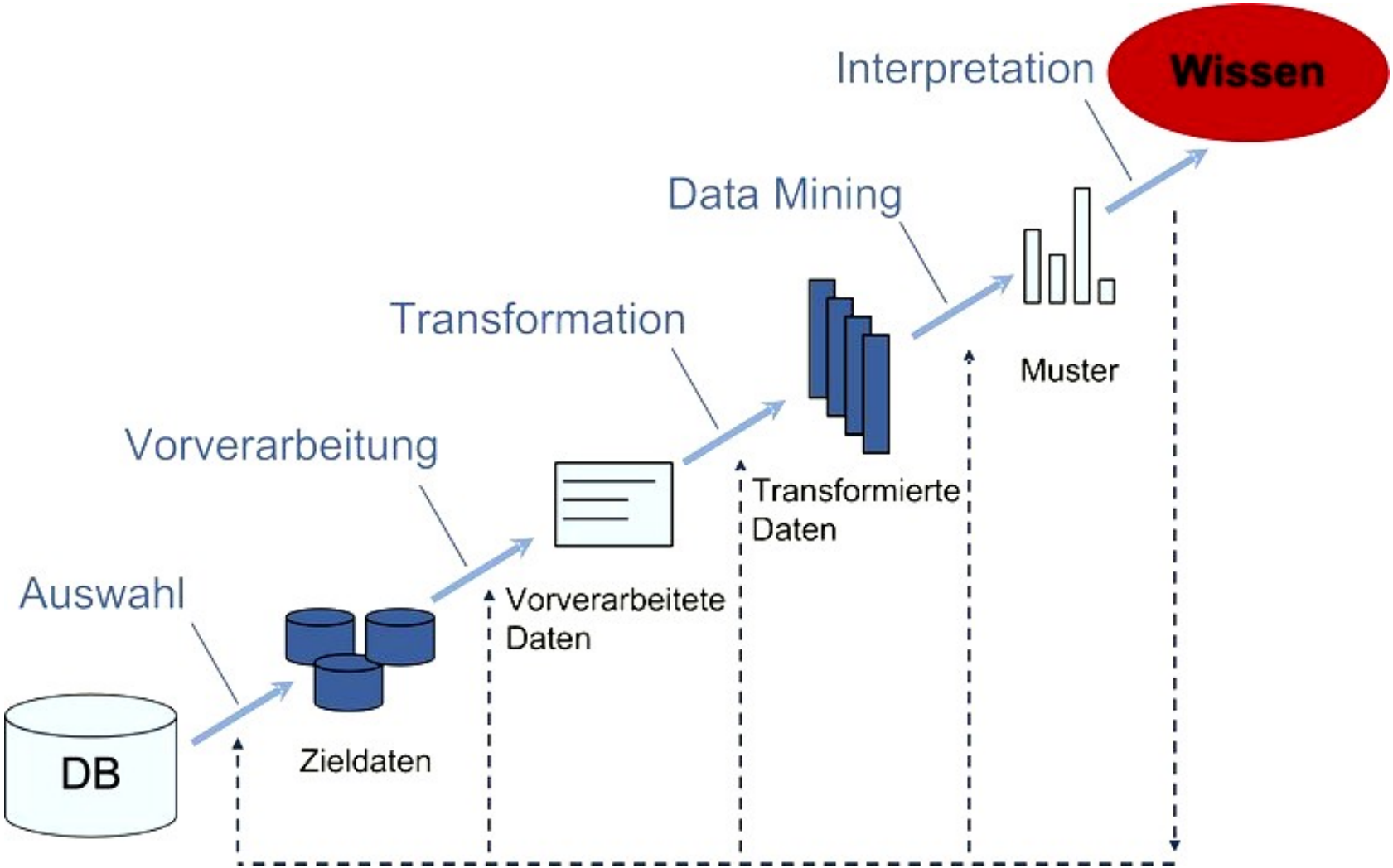
Beide Begriffe werden meist synonym verwendet.

KDD vs. Statistik vs. OLAP

KDD	Datenbestände nach Regelmäßigkeiten, Mustern und Abweichungen und Beziehungen untersuchen – Generieren von Hypothesen
Statistik	Überprüfung neuer Hypothesen, Analyse empirischer Daten
OLAP	Auswertung von Daten – Beantworten von vorher festgelegten Fragestellungen



Von Daten zum Wissen



[nach Fayyad et al. 1996]

Ziele des Data Mining

- Daten → Wissen, Entscheidungen schneller treffen
  - Kundenzufriedenheit
  - Marktkenntnis
  - Vorsprung vor der Konkurrenz
  - Erschließung neuer Vertriebskanäle

Alle 18 Monate verdoppelt sich die Speicherkapazität in Unternehmen!

[IBM 2009]

(1) Selektion

- Auswahl des Datenbestandes

(2) Vorverarbeitung

- Beseitigung von Qualitätsmängeln

(3) Transformation + Reduktion

- In die benötigte Form gebracht

(4) Data Mining

- Mustererkennung durch math. Verfahren

(5) Interpretation

- Evaluation durch Experten

Selektion

- Auswahl der zu analysierenden Daten aus einer Rohdatenmenge, vertikal und horizontal
- Zusammenfügen von Daten aus mehreren Quellen, z.B. in ein Data Warehouse (Datenbanken, Data Cubes, einzelne Zeilen, nichtformatierte Daten, Formulare, *overlay data*...)
  - Problem: Heterogene Daten
    - Wie kann man sicher sein, dass einzelne Entities den selben Inhalt haben? (cust_id und cust_number)
    - Redundanzen
    - ...

(1) Selektion

- Auswahl des Datenbestandes

(2) Vorverarbeitung

- Beseitigung von Qualitätsmängeln

(3) Transformation + Reduktion

- In die benötigte Form gebracht

(4) Data Mining

- Mustererkennung durch math. Verfahren

(5) Interpretation

- Evaluation durch Experten

Warum die Vorverarbeitung?

[Cabena et al. 1997]:

- 10% des Zeitaufwandes im KDD entfallen auf die Ausführung von Data-Mining-Methoden
- 90% werden für Datenaufbereitung und Nachbearbeitung aufgewendet
- Untersuchungen belegen Fehlerwahrscheinlichkeit in Rohdaten von bis zu 30%

Ziel: einheitliche Struktur und Format, Steigerung der Datenqualität besonders bei heterogenen Datenquellen

Vorverarbeitung - Inkonsistenzen

Semantische Probleme

- Synonyme, Homonyme
- ➔ Individuelle Lösung mit Hilfe von Metadaten und bereichsspezifischem Wissen

Syntaktische Probleme

- Verschiedene Schreibweisen
- ➔ Nutzung eines einheitlichen Schemas und von Katalogen (bsp. Straßenverzeichnis)

Zeitreihenprobleme

- Fehlende Aktualität der Daten

Vorverarbeitung - Inkonsistenzen

Redundanzen

Fehlende Werte

- Unbestimmbar vs. nicht bestimmt
- ➔ Säuberung

Falschwerte

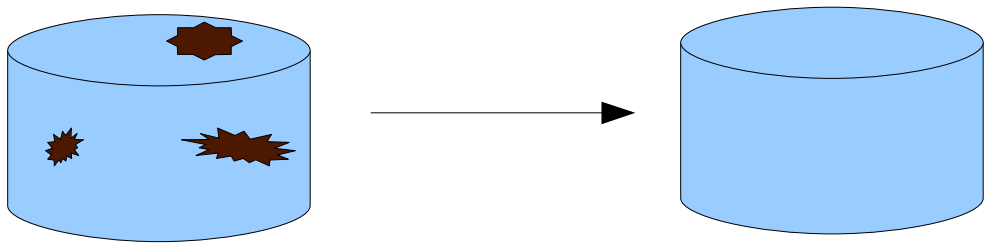
- ➔ Transformation

Zu „genaue“ Werte

- ➔ Aggregation

Säuberung bei fehlenden Werten

- Tupel ignorieren
- Wert manuell einsetzen
- Globale Konstante (bsp. *unknown)* einsetzen
- Mittelwert einsetzen
- Wahrscheinlichsten Wert einsetzen – Ermittlung mit Hilfe von Entscheidungsbäumen o.ä.



(1) Selektion

- Auswahl des Datenbestandes

(2) Vorverarbeitung

- Beseitigung von Qualitätsmängeln

(3) Transformation + Reduktion

- In die benötigte Form gebracht

(4) Data Mining

- Mustererkennung durch math. Verfahren

(5) Interpretation

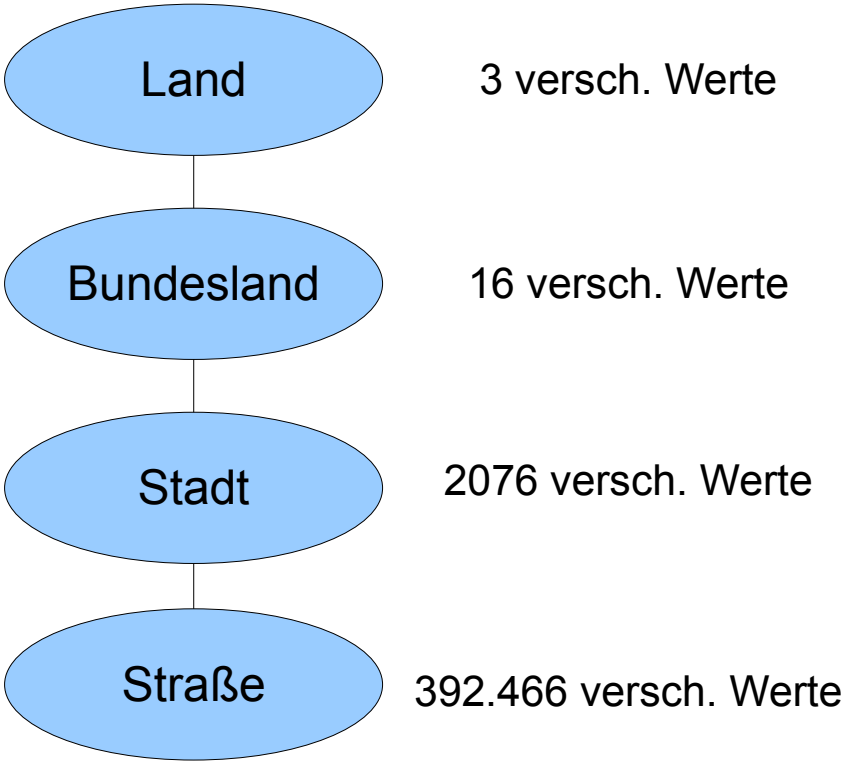
- Evaluation durch Experten

Transformation

- Glättung, d.h. „Ausreißer“ entfernen
  - Nützlich für Entscheidungsbäume, Hierarchien, ...
- Aggregation / Summenbildung
  - Daten in unterschiedlichen Granularitäten
- Generalisierung
  - „low level“ nach „higher level“ (street → city → country), numerische zu abstrakten Werten (young – middle-aged – senior) und umgekehrt
- Normierung
  - Vergleichbarkeit herstellen, günstig für für KNN
- Konstruktion von Attributen

Datenreduktion

- Data cube aggregation
- Attribute/Feature subset selection
- Dimensionality reduction
- Numerosity reduction
- Discretization



(1) Selektion

- Auswahl des Datenbestandes

(2) Vorverarbeitung

- Beseitigung von Qualitätsmängeln

(3) Transformation + Reduktion

- In die benötigte Form gebracht

(4) Data Mining

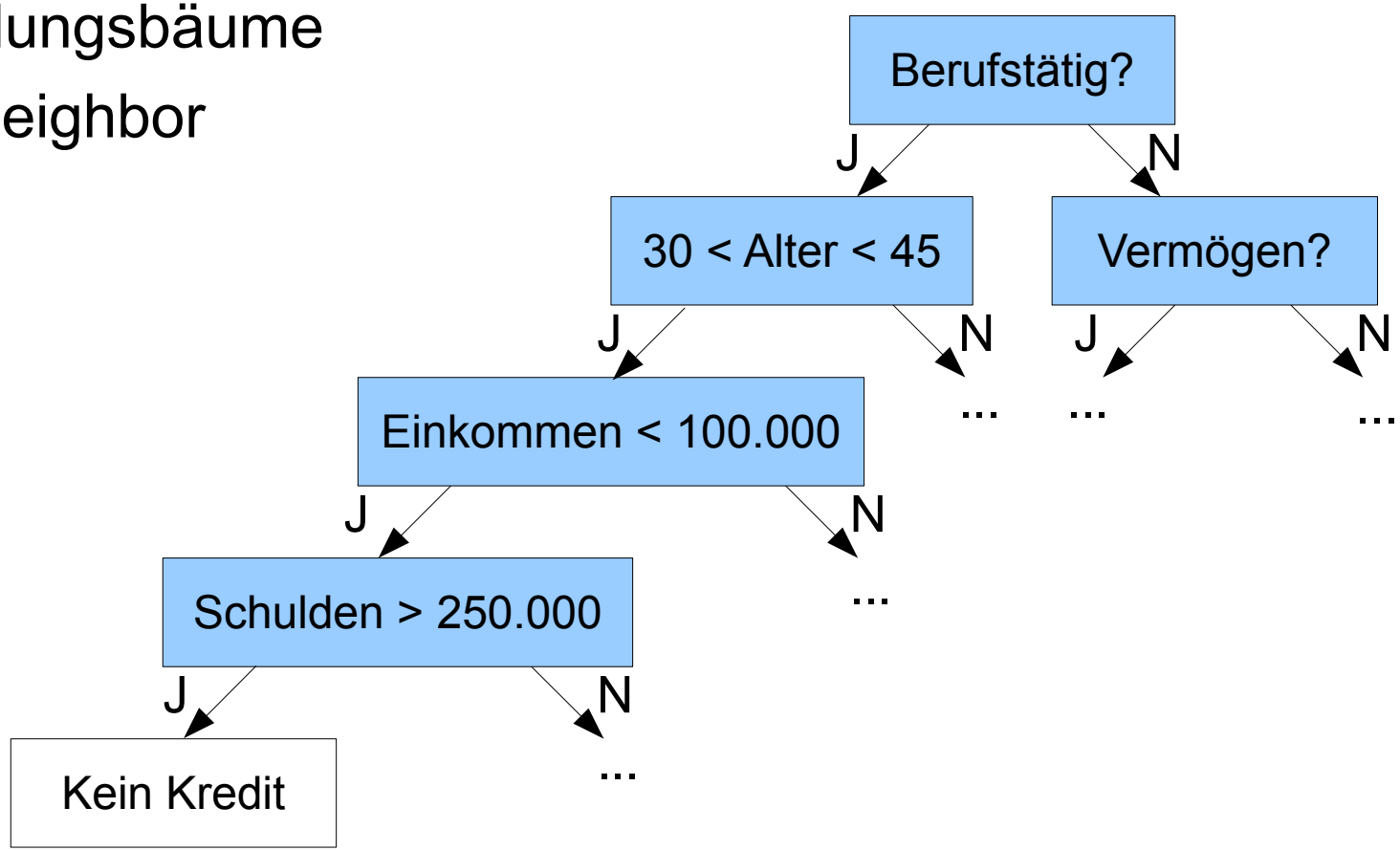
- Mustererkennung durch math. Verfahren

(5) Interpretation

- Evaluation durch Experten

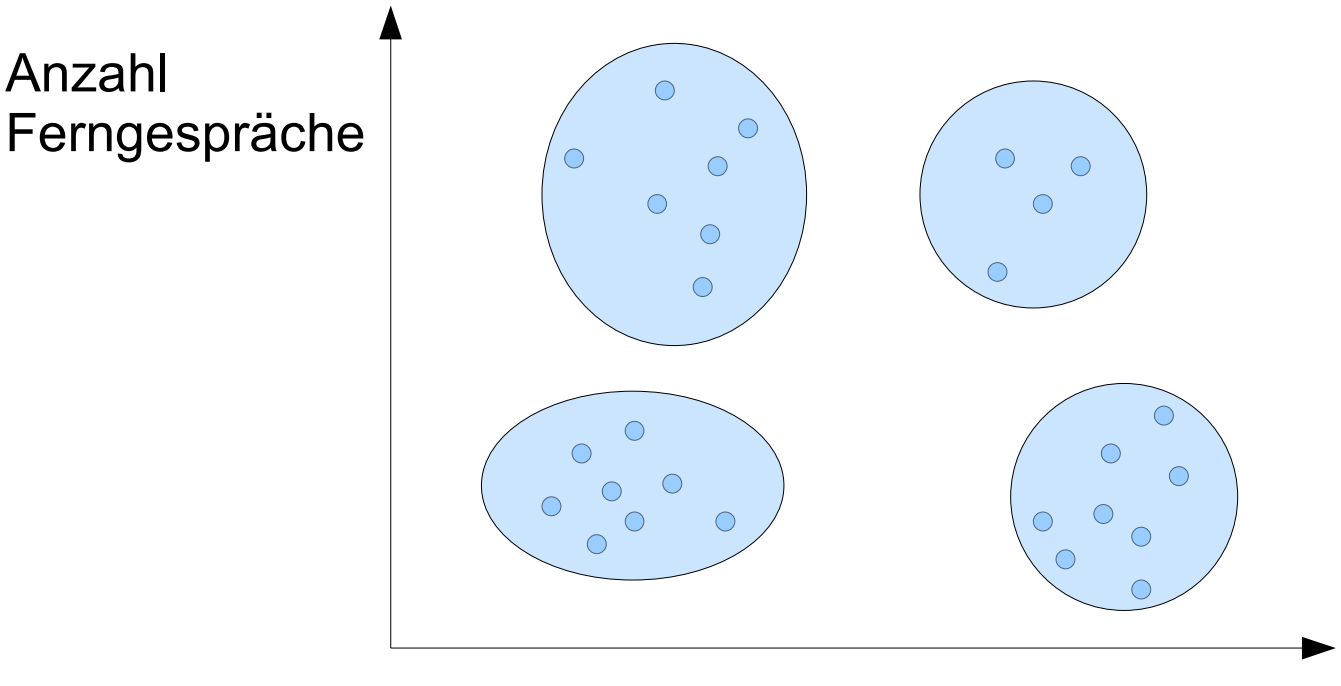
Data-Mining-Analysen

- Klassifikationsanalysen
  - Bsp. Kreditwürdigkeit von Bankkunden feststellen
    - Entscheidungsbäume
    - Nearest neighbor
    - ...



Data-Mining-Analysen

- Assoziationsanalysen
  - Bsp. Untersuchung des Kaufverhaltens von Kunden
- Clustering
  - Gruppeneinteilung von Kunden



(1) Selektion

- Auswahl des Datenbestandes

(2) Vorverarbeitung

- Beseitigung von Qualitätsmängeln

(3) Transformation + Reduktion

- In die benötigte Form gebracht

(4) Data Mining

- Mustererkennung durch math. Verfahren

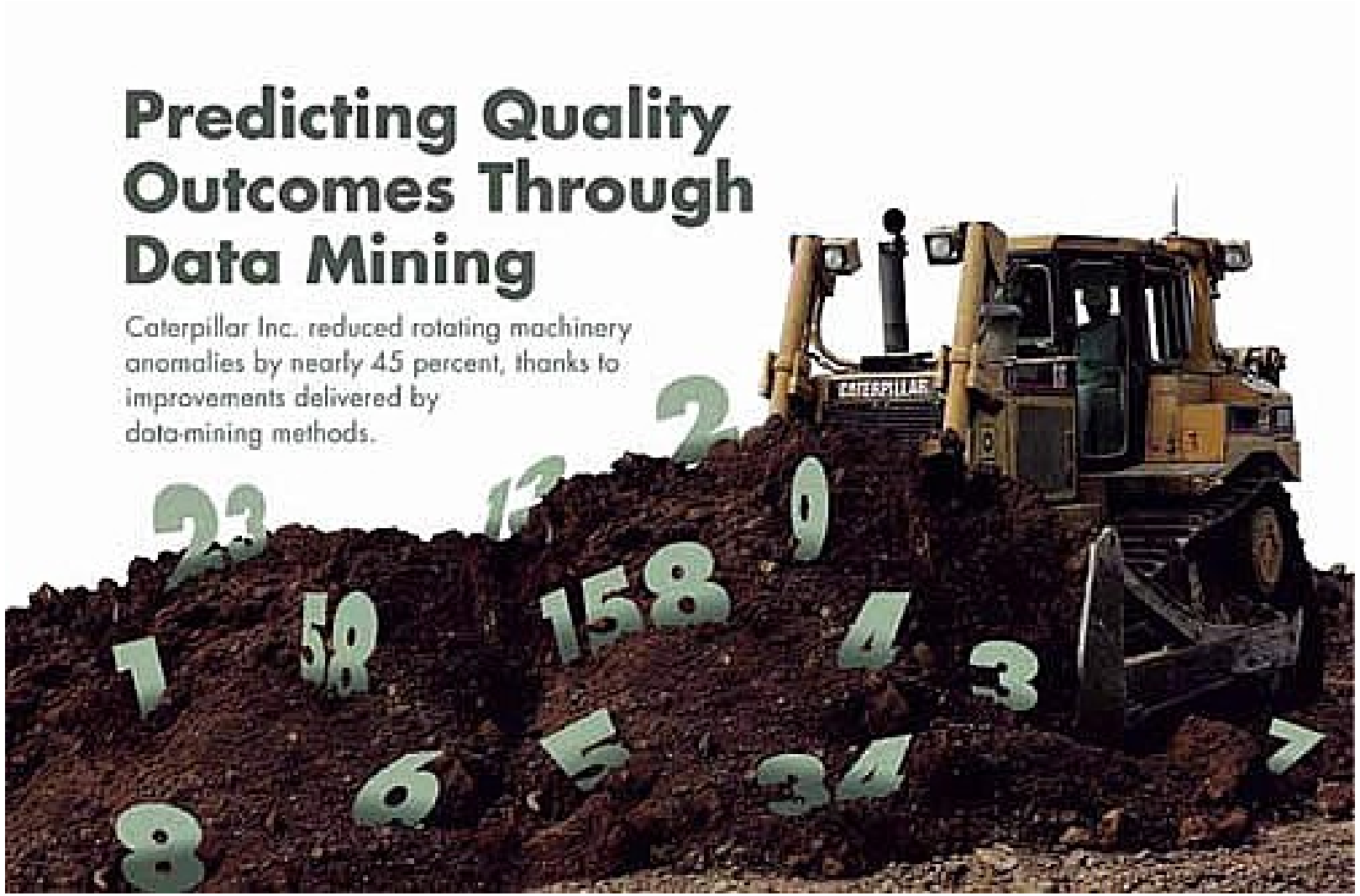
(5) Interpretation

- Evaluation durch Experten

Anforderungen

- Effizienz
- Skalierbarkeit
- Verarbeitung von unterschiedlichen Datentypen
- Verarbeitung von großen Datenmengen
- *Interestingness* (Neuigkeitswert, Relevanz)
- Darstellung der Unsicherheit aufgrund fehlerhafter oder unvollständiger Daten

Anwendungsgebiete



Anwendungsgebiete

- Marketing
  - Kundensegmentierung nach Kaufverhalten und Interessen → gezielte Werbemaßnahmen, verbesserte Kundenansprache, Steigerung der Verkaufszahlen
  - Warenkorbanalyse zur Optimierung von Preisen und der Platzierung von Produkten im Supermarkt
  - Kundenbindung bzw. Neukundengewinnung
  - Cross Selling
- Betrugsaufdeckung bei Banken
- Web Usage Mining
  - Personalisierung von Internetpräsenzen durch die Erstellung von Zugriffsprofilen

Anwendungsgebiete

- Text Mining
  - Anwendung von Data-Mining-Verfahren auf Textdokumenten zum Wissensmanagement, Kundenmanagement, ...
- Player-Tracking
  - „optimiertes“ Glücksspiel durch Spielerverfolgung bis hin zur Anpassung von Spielabläufen
- Pharmakovigilanz
  - Arzneimittelüberwachung nach Marktzulassung im Hinblick auf unbekannte Nebenwirkungen
- Aufdeckung von Zusammenhängen in ungeklärten Straftaten

[vgl. Wikipedia 2010]

Probleme beim Data Mining

- Datenqualität
- Softwarequalität
- Aussagekraft der Ergebnisse
- Datenschutz
- ...

Noch Fragen?


HEY, MEINE MIGRÄNE IST WEG.
www.nichtlustig.de


EINE SAMMELKLAGE WEGEN SEXUELLER BELÄSTIGUNG AM ARBEITSPLATZ?
SIE NENNEN ES MELKEN, ICH NENNE ES KRANK!
www.nichtlustig.de

Quellen

[Cabena et al. 1997] Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A. : Discovering Datamining: From Concept to Implementation. Prentice Hall, 1997.

[Decker, Focardi 1995] Decker, K.; Focardi, S.: Technology overview: a report on data mining. Swiss Federal Institute of Technology (ETH Zurich) Technical Report CSCS TR-95-02, Zürich, 1995.

[Fayyad et al. 1996] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Hrsg.): Advances in knowledge discovery and data mining. AAAI Press, 1996.

[Herden 2007] Herden, O.: Data Mining. In: Kudraß, T. (Hrsg.): Taschenbuch Datenbanken. Hanser-Verlag, 2007.

[IBM 2009] http://www.redbooks.ibm.com/redpapers/pdfs/redp4529.pdf

[Kneip 2008] Kneip, M.: Data Mining. GRIN Verlag, 2008.

[Wikipedia 2010] http://de.wikipedia.org/wiki/Data_mining

Bildquellen

Goldminer S.3: http://www.flickr.com/photos/jansochor/

Kuh Eröffnungsbildschirm: http://www.welt.de/multimedia/archive/1219770910000/00652/kuh_klima_DW_Wissen_652247g.jpg

Mining S. 9 ff. (Gliederung): http://thomaslarock.com/wp-content/uploads/2009/06/datamining.jpg

Mining S. 5: http://www.engr.sjsu.edu/lwesley/images/DataMining.jpg

Puzzle S. 6: http://www.darteam.co.in/Market%20Research%20Reports/product_small.jpg

Commercial S. 24: http://www.qualitydigest.com/sept06/Images/Data%20Mining/0906Datamining.jpg

Mining S. 27: http://www.monash.edu.au/pubs/monmag/issue7-2001/img/datamining7.jpg

Kühe S. 28: http://nichtlustig.de