DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Knowledge Discovery in Databases

WS 2017/18

Kapitel 1: Einleitung

Vorlesung: Prof. Dr. Peer Kröger

Übungen: Anna Beer, Florian Richter

Organisatorisches

- Wöchentliche Veranstaltungen
  - Vorlesung:
    - Mittwoch, 09:30 – 12:00 h, Raum S 004 (Schellingstr. 3)
  - Übungen (Beginn: siehe Vorlesungs-Homepage):
    - Do, 14:00 - 16:00 Uhr , Raum D Z005 (HGB)
    - Do, 16:00 - 18:00 Uhr, Raum D Z005 (HGB)
    - Fr, 12:00 - 14:00 Uhr, Raum A 015 (HGB)
    - Fr, 14:00 - 16:00 Uhr, Raum A 015 (HGB)
- Klausur: tbd

Organisatorisches

- Material (Folien, Übungsblätter, Links, etc.):
  - Vorlesungs-Homepage:
    http://www.dbs.ifi.lmu.de/cms/studium_lehre/lehre_master/kdd1718/index.html
  - Übungen:
    - Bitte vor der Übung downloaden und zu hause vorbereiten
    - Präsentation und Diskussion der Lösung in den Übungen
  - Klausur:
    - Stoff: alles, was in der Vorlesung und Übung bis dahin dran kam.
    - Registrierung via UniWorX (in den nächsten Tagen)

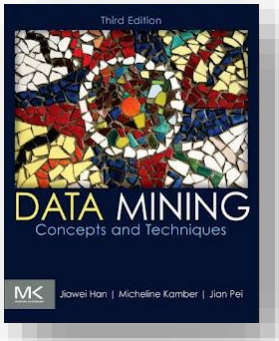
Inhalt von KDD I (Plan)

1. Einleitung
2. Daten Repräsentation, Data Warehousing, Business Intelligence (BI)
3. Frequent Pattern Mining
4. Clustering
5. Outlier Detection
6. Classification
7. Regression
8. Further Topics

Literatur Acknowledgements

Diese Folien verwenden u.a. verändertes Material aus den geschützten Folien derAuthoren der folgenden Bücher:

© Jiawei Han, Micheline Kamber, Jian Pei:
Data Mining – Concepts and Techniques,
3rd ed., Morgan Kaufmann Publishers, 2011.
http://www.cs.uiuc.edu/~hanj/bk3



© Martin Ester and Jörg Sander:
Knowledge Discovery in Databases – Techniken und Anwendungen
Springer Verlag, 2000 (in German).

An vielen Stellen wird zudem auf die Original-Literatur verwiesen.

Motivation

- Data Mining = Extraktion von Mustern aus Daten
- Muster
  - „Normale“ Muster („Regularities“) – z.B. häufige Itemsets, Clusters
  - „Abnormalitäten“ („Irregularities“) – z.B. Ausreißer
- Hervorstechende Muster
  - Viele Muster sind trivial oder repräsentieren bereits bekannte Informationen
    - „all mothers in our database are female“
  - Viele Muster sind redundant
    - „Ist die Kombi {bread, butter, salt} häufig, so ist die Kombi „{bread, butter} auch häufig“
- Aggregation der Daten kann hier Abhilfe schaffen

Was ist Data Mining?

- Knowledge Discovery in Databases (Data Mining):
  - Extraction of interesting (*non-trivial*, *implicit*, *previously unknown* and *potentially useful*) information or patterns from data in *large databases*

- Alternative Name:
  - Data mining: eine Fehlbezeichnung?
  - knowledge extraction, data/pattern analysis, data archeology, data dredging ("Ausbaggern"), information harvesting, business intelligence, etc.
  - Neuerdings: Data Science, Machine Learning (nicht so neu), Artificial Intellegence (noch weniger neu)
- Roots of data mining
  - Statistics
  - Machine learning
  - Database systems
  - Information visualization

Warum Data Mining?

"Necessity is the mother of invention"

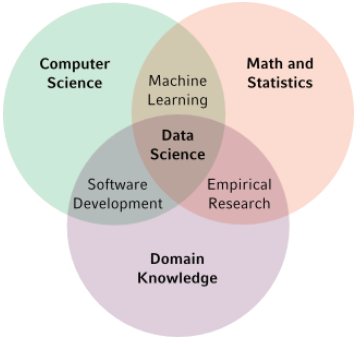
- Datenexplosion, Digitalisierung
  - "Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories"
- "We are drowning in data, but starving for knowledge!"
- Lösungen für dieses Problem: data warehousing and data mining
  - Data Warehousing and on-line analytical processing (OLAP), BI
  - Data Mining um interessante Muster in Daten zu identifizieren(rules, regularities, patterns, constraints) from data in large databases

Ein bißchen (historisches) Buzzword-Bingo

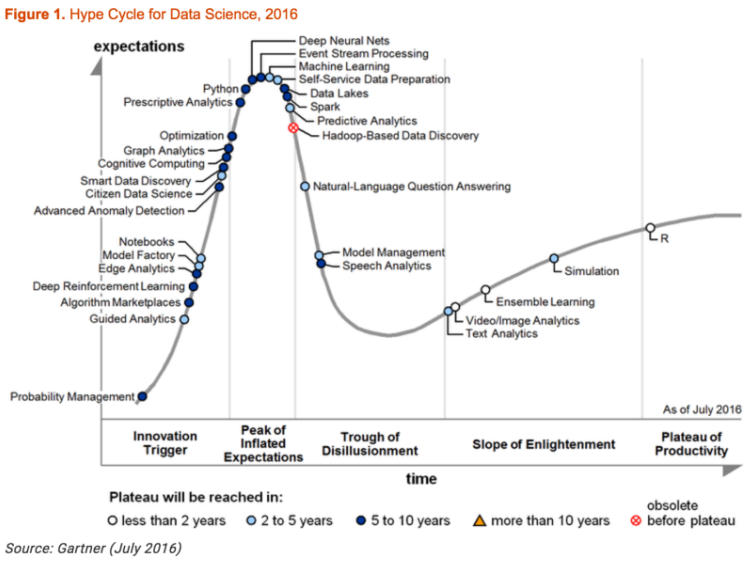
- Big Data
  - McKinsey-Report 2011
- Data Science
  - Die Eierlegende-Woll-Milch-Sau
- Machine Learning und KI (AI)
  - Altes in neuem Gewand?
  - Klassisches ML: induktives Lernen (aus Fakten = Beobachtungen Abhängigkeiten lernen)
  - Klassische AI: deduktives Lernen (aus Fakten, Regeln, Axiome, neue Theoreme ableiten)

    Paul ist Schotte, alle Schotten sind geizig

    => Paul ist geizig
- Data Lake
- Digitalisierung, Industrie 4.0, …



Data Science



Und was jetzt "Data Mining"?

- Nochmal ML
  - Basierend auf Fakten/Beobachtungen werden Abhängigkeiten/Muster abgeleitet („generalisiert")

    Paul ist Schotte, Tom ist Schotte, Alex ist Schotte,

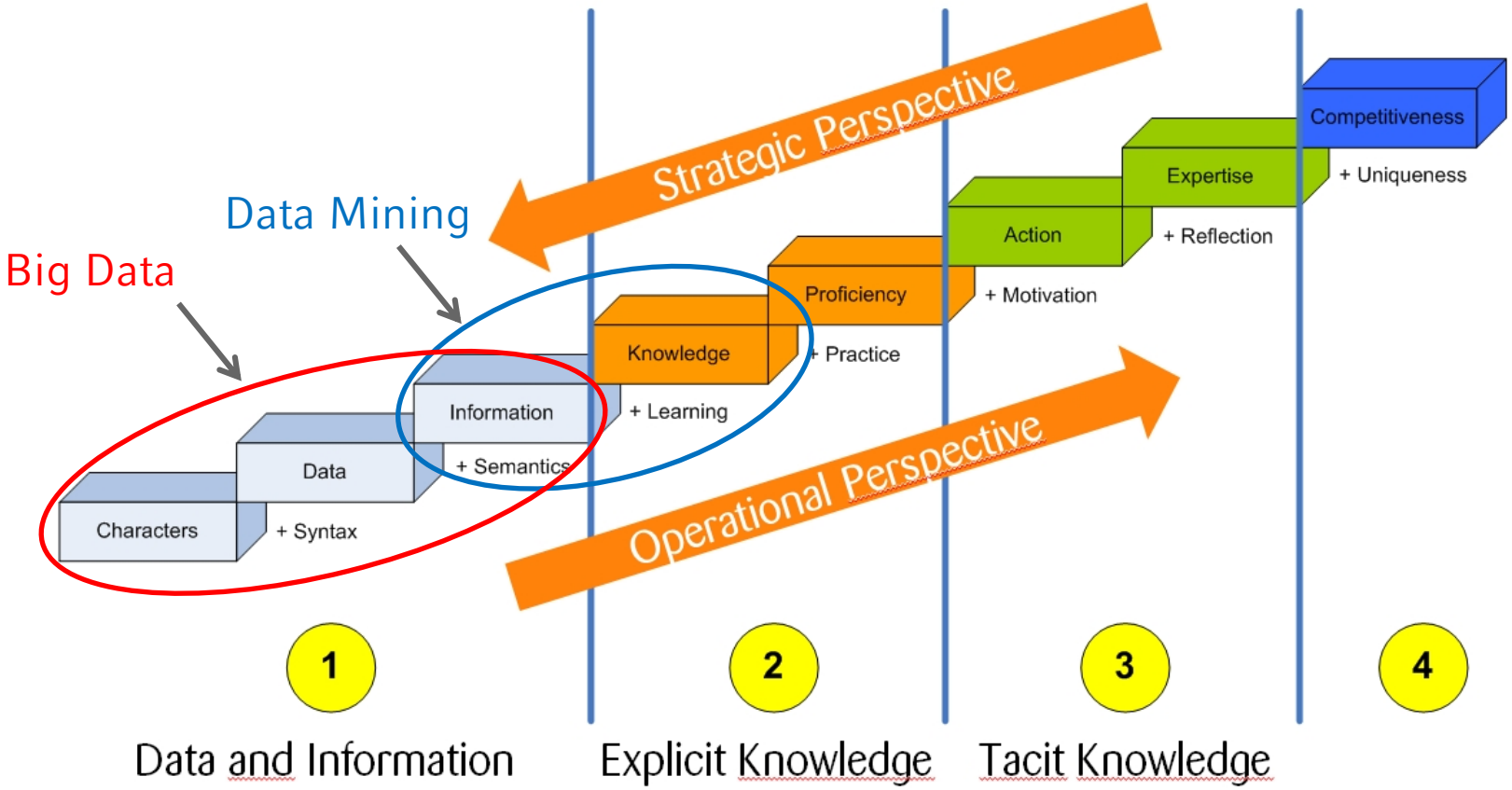
    Paul ist geizig, Tom ist geizig, Alex ist geizig

    => Alle Schotten sind geizig
  - Wesentlich: Modellierung des (Lern-)Problems (meist als Optimierungsproblem) steht im Vordergrund
  - Lösung wird dann oft Standard-Algorithmen überlassen (d.h. die „Maschine lernt") => funktionaler Ansatz
- Data Mining
  - Im Fokus steht die Berechnung der Muster (durch ein Verfahren/einen Algorithmus); dazu werden oft keine Fakten vorab benötigt
  - Dieser Algorithmus wird von einer Person geliefert (d.h. hier lernt nicht die Maschine sondern der Mensch steuert das Lernen)

    => prozeduraler Ansatz

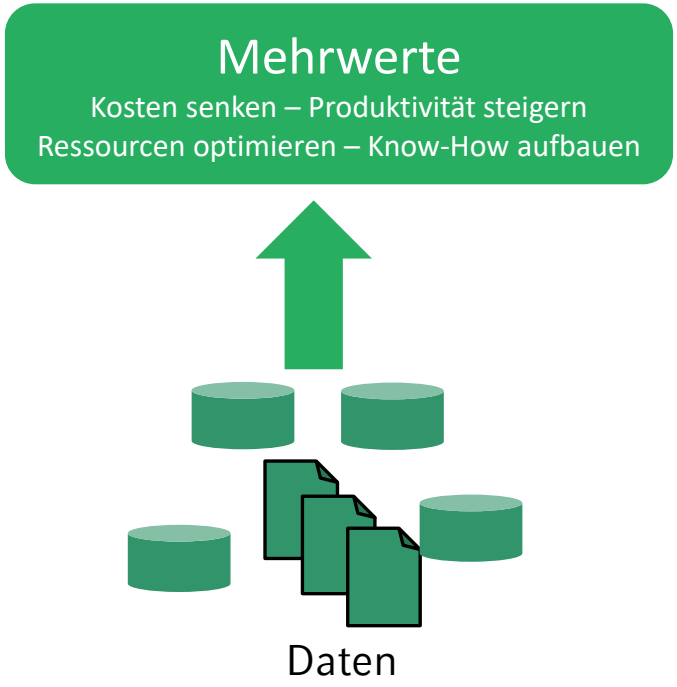
Wertschöpfung durch Wissen

- Stairs of Knowledge (K. North):

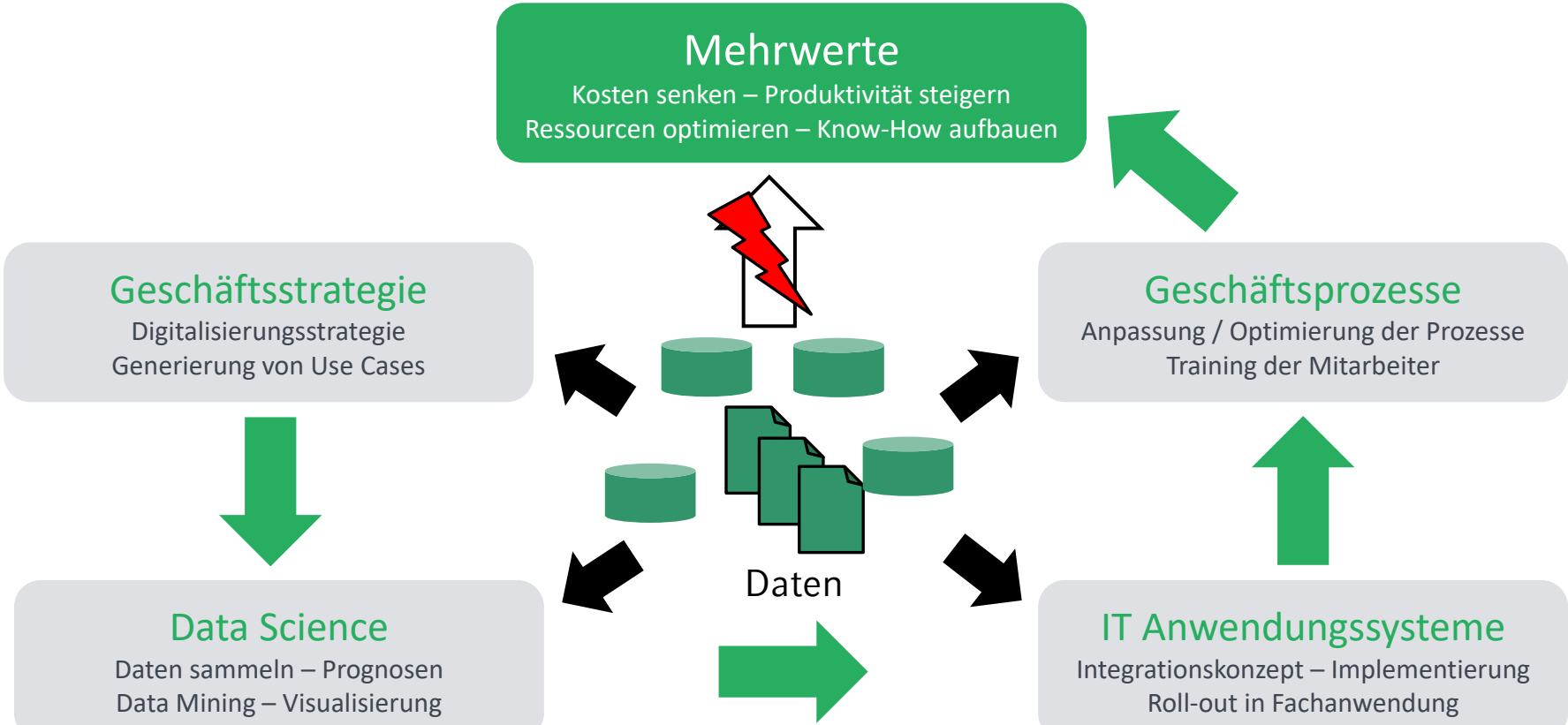


Stairs of Knowledge: North, K.: Wissensorientierte Unternehmensführung - Wertschöpfung durch Wissen. Gabler, Wiesbaden 1998.
Picture from: http://wissensarbeiter.wordpress.com/2012/10/29/information-wissen-und-expertise-dazwischen-liegen-welten/

- Data Science Prozess im Unternehmen



- Data Science Prozess im Unternehmen

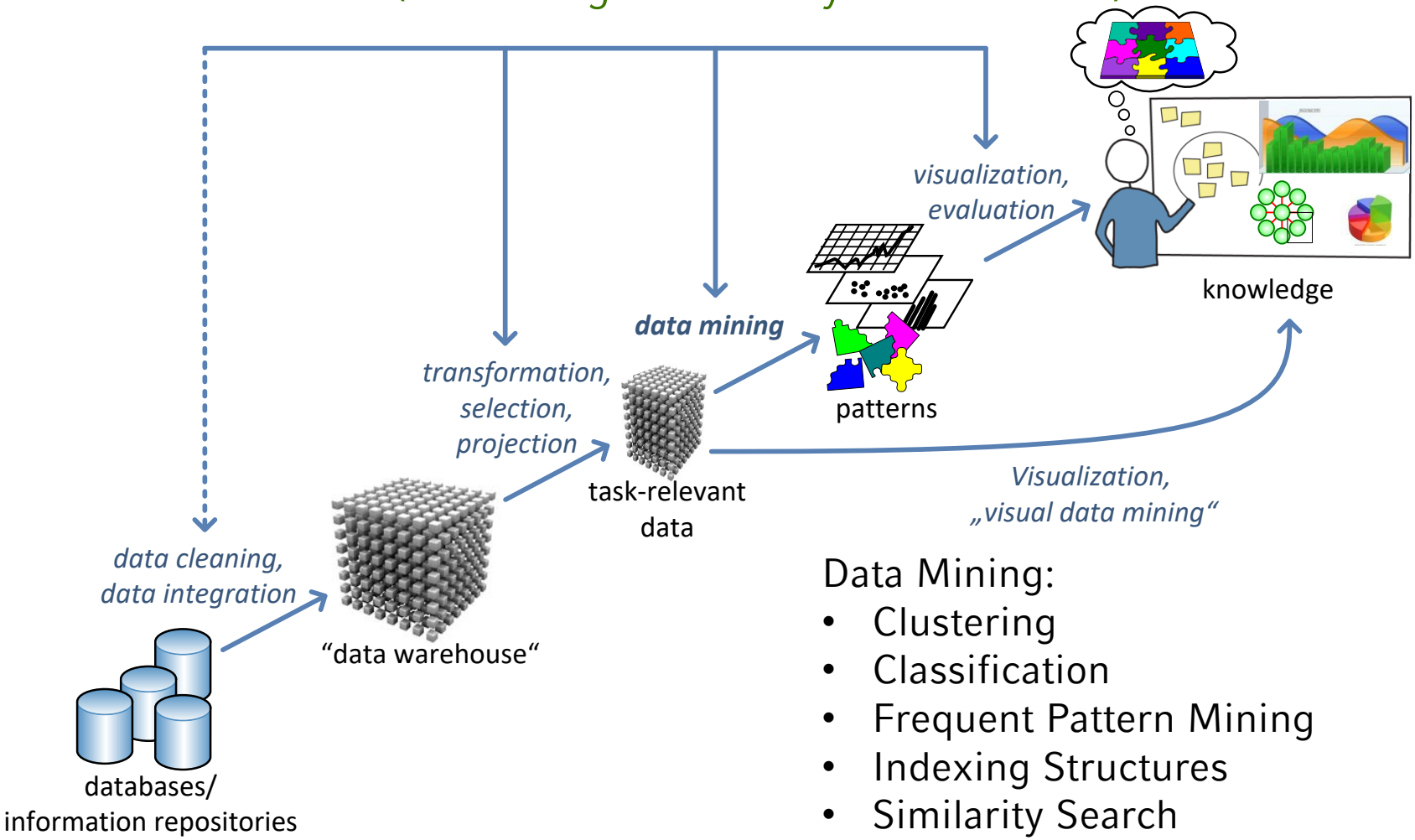


Data Mining: Anwendungen

- Datenanalyse und Entscheidungsfindung (Decision Support)
  - Markt-Analyse und Markt-Management
    - target marketing, customer relation management (CRM), market basket analysis, cross selling, market segmentation
  - Risiko-Analyse und Risiko-Management
    - Forecasting, customer retention (“Kundenbindung”), improved underwriting, quality control, competitive analysis
  - Betrugserkennung und Betrugs-Management
  - …
- Weitere Anwendungen
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering
  - …

Der Knowledge Discovery Prozess

- The KDD-Prozess (**K**nowledge **D**iscovery in **D**atabases)

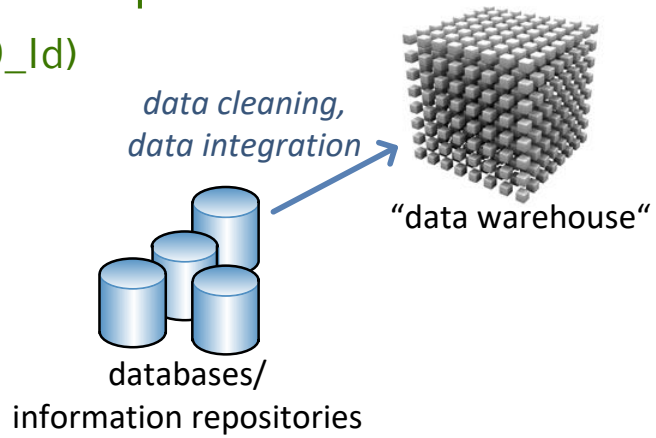


Data Mining:

- Clustering
- Classification
- Frequent Pattern Mining
- Indexing Structures
- Similarity Search

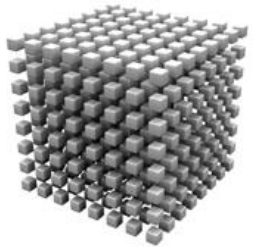
Data Cleaning und Datenintegration

- … macht meist 60% und mehr des Gesamtaufwands aus!!!!
- Integration von Daten aus verschiedenen Datenquellen
  - Mapping von Attributsnamen (z.B. C_Nr → O_Id)
  - Join von verschiedenen Tabellen
    (e.g. Table1 = [C_Nr, Info1]
    and Table2 = [O_Id, Info2] ⇒
    JoinedTable = [O_Id, Info1, Info2])
- Finden & Eliminieren von Inkonsistenzen
- Eliminieren von Rauschen
- Umgang mit fehlenden Werten
  - Wenn möglich, fehlende Werte ersetzen (z.B. mit Default Werten, Mittelwert, Anwendungsspezifische Berechnungen, …)
- …



Data Cleaning und Datenintegration

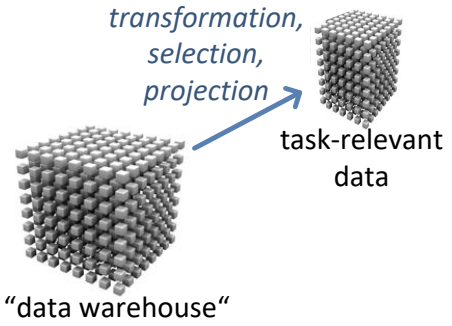
- Im Data Warehouse Bereich spricht man vom ETL-Prozess (Extract-Transform-Load)
- Das klassische Data Warehouse besteht aus einem Data Cube
  - Multidimensionales Datenmodell
  - Intern meist als relationales Schema realisiert
    (3NF; Snowflake- vs. Star-Schema)



- Business Intelligence (BI) setzt hier an mit:
  - Reports: Verschiedene Visualisierungen des Data Cube
    - Entlang verschiedener Dimensionen
    - Auf verschiedenen Aggregationsstufen
    - …

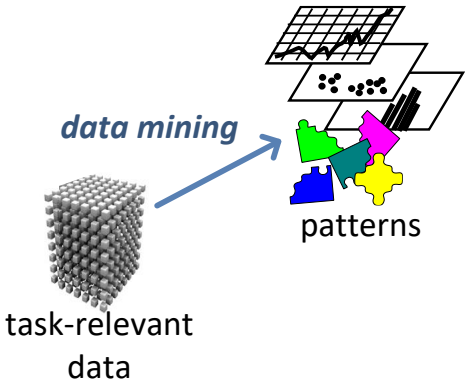
Beim Data Mining gehts weiter: Fokus auf Task-relevante Daten

- Bestimmen von wichtigen Merkmalen (Features), Dimensionsreduktion, invariante Repräsentationen
- Selektionen
  - Auswahl relevanter Tupel/Reihen
    (z.B., Sales-Daten von 2001)
- Projektionen
  - Auswahl relevanter Attribute/Spalten
    (z.B., "id", "date" "amount" from (Id, name, date, location, amount))
- Transformationen, z.B.:
  - Normalisieren (z.B., age:[18, 87] → n_age:[0, 100]
  - Diskretisieren von numerischen Attributen (z.B., amount:[0, 100] → d_amount:{low, medium, high}
  - Ableiten von Tupeln und Attributen
    - Aggregation von Mengen von Tupeln (z.B., total amount per months)
    - Neue Attribute (z.B., diff = sales current month – sales previous month)



Basic Data Mining Aufgaben (Tasks)

- Suche nach interessanten Mustern oft abhängig von der Art der Daten
- Data Mining „Funktionen“ (Verfahrensklassen):
  - Frequent Patterns (häufige Muster)
  - Clustering
  - Classification
  - Characterization and Discrimination
  - Weitere (spezialisierte) Verfahrensklassen
    - Outlier detection (Ausreißererkennung)
    - Sequential patterns (typw. in Sequenz-Daten)
    - Trends and analysis of changes (typw. in zeitlichen Daten)
    - Spatial data mining, web mining
    - …
- Eine wichtige Aufgabe des Data Scientists
  - Die richtige Verfahrensklasse bestimmen
  - Den richtigen Data Mining Algorithmus auswählen



Frequent Pattern Mining

- Die einfachste Art von Mustern: Häufigkeiten von Vorkommen (einzelner Werte, Wertkombinationen, etc.) zählen
- Bestimme häufige Muster in Transaktionsdatenbanken
  - Daten: Menge von Transaktionen/"Itemsets" (z.B. Warenkörbe)
  - Items, die häufig gemeinsam in Transaktionen vorkommen (*frequent itemsets*): Hinweis auf Korrelationen/Kausalitäten

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

- Anwendugen:
  - Market-basket analysis
  - Cross-marketing
  - Catalog design
  - Manchmal auch als Basis Operation für Clustering und Klassifikation
  - Association rule mining: Bestimme Korrelationen zwischen Itemsets

Beispiele:

buys(x, "diapers") → buys(x, "beers") [support: 0.5%, confidence: 60%]

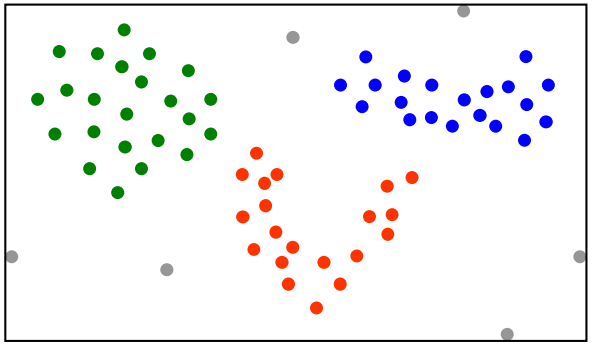
major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [support: 1%, confidence: 75%]

Identität vs. Ähnlichkeit

- Oft kann man nicht einfach nur Häufigkeiten von Werten/Wertekombinationen/etc. zählen (z.B. bei numerischen Werten)
- Dann kann man z.B. Ähnlichkeits-Maße verwenden, um zu entscheiden, ob zwei Werte/Wertkombinationen/etc. „gleich“ sind
- Anstatt häufige Werte/Wertekombinationen zu finden, möchte man dann Werte/Wertekombinationen (Datenobjekte) finden, die ähnlich zueinander sind

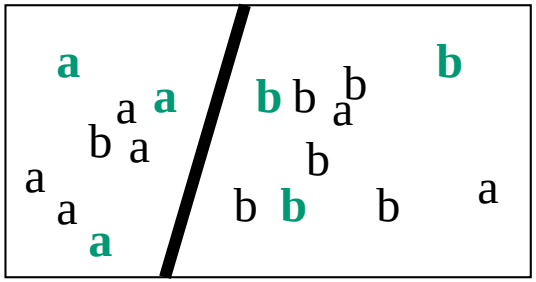
Clustering

- Cluster = Gruppe von Datenobjekten (Tupel, …) die ähnlich zu einander sind
- Die Cluster sind *a priori* unbekannt („unsupervised“), insbesondere
  - die Semantik der Cluster
  - die Charakteristik der Cluster
- Wesentliche „Zutat“: Ähnlichkeits-Maß (oder Distanz-Maß)
  - Quantifiziert die Ähnlichkeit von Objekten
  - Beispiel: Objekte sind Punkte im 2D; Ähnlichkeit ist die Nähe der Punkte zueinander (z.B. Euklidische Distanz)
- Anwendung
  - Customer profiling/segmentation
  - Document or image collections
  - Web access patterns
  - . . .



Klassifikation

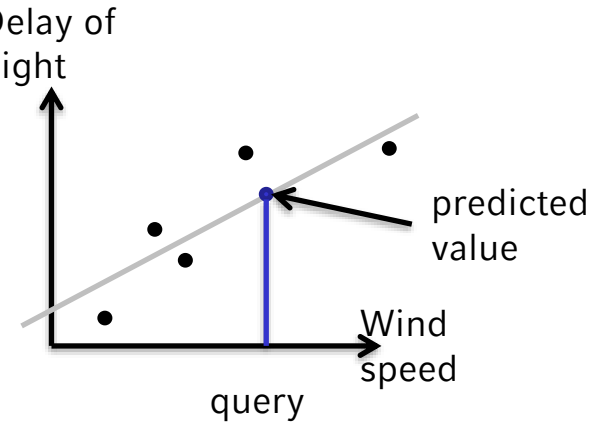
- Grundidee ähnlich wie Clustering, nur
  - die Gruppen (=Klassen) sind *a priori* bekannt („supervised“): sog. „class labels“
  - Es existieren bereits Beispiel-Objekte für die Klassen (“training data”), d.h. die Charakteristik/Semantik der Klassen ist (z.T.!!!) bekannt
- Aufgabe:
  - Finde ein Modell (Funktion/Regel), das
    (basierend auf den Trainingsdaten)
    - Die einzelnen Klassen beschreibt und trennt
    - Die Klassenzugehörigkeit “neuer” Objekte vorhersagt



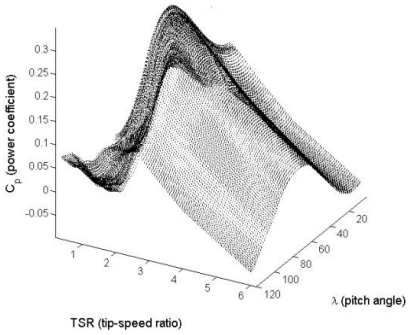
- Anwendungen
  - Classify gene expression values for tissue samples to predict disease type and suggest best possible treatment
  - Automatic assignment of categories to large sets of newly observed celestial objects
  - Predict unknown or missing values (→ KDD pre-processing step)
  - . . .

Prediction/Regression

- Ähnlich wie Klassifikation, der Output ist aber nicht kategorisch sondern nummerisch
- Aufgabe: finde ein Modell/Funktion (basierend auf Trainingsdaten), das
  - Den Zusammenhang zwischen Input und Output darstellt
  - Den nummerischen Output-Wert für ein "neues" Objekt vorhersagt



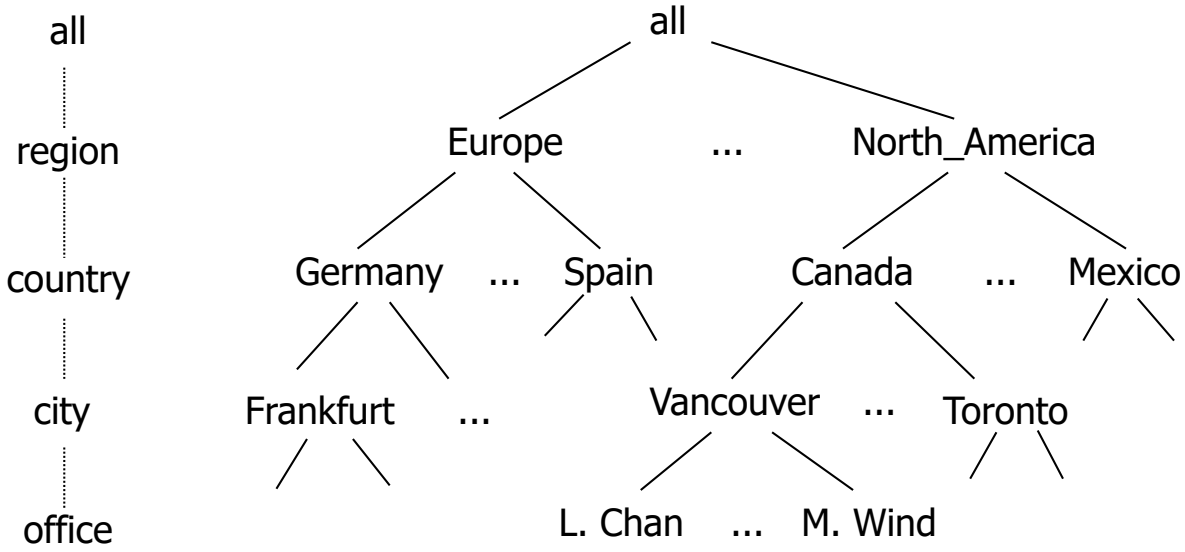
- Applications
  - Build a model of the housing values, which can be used to predict the price for a house in a certain area
  - Build a model of an engineering process as a basis to control a technical system
  - . . .



Wind turbine

Generalization Levels

- Generalisierung/Spezialisierung, Aggregation
  - Basierend auf einer Aggregation von Attributen entlang einer Konzept-Hierarchie (ggfls. Ontologie)
    - Data Cube Anstatz aus dem Data Warehouse-Bereich (OLAP)
    - Attribute-oriented induction Ansatz



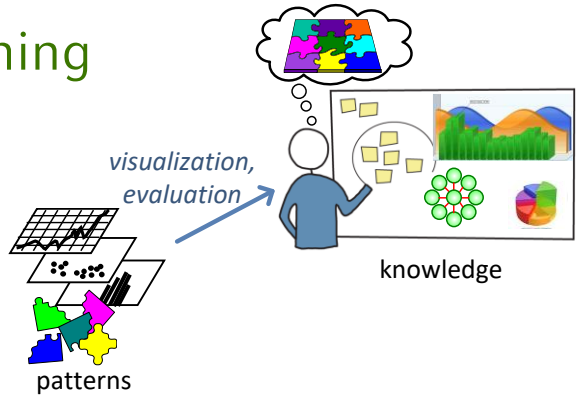
Weitere Data Mining Methoden

- Outlier detection
  - Orthogonal zum Clustering: Finde Objekte, die nicht zur generellen Charakteristik der (Mehrheit der) Daten passen

    (fraud detection, rare events analysis)
- Trends and Evolution Analysis
  - Sequential patterns (finde wiederkehrende Event-Sequenzen)
- Spezielle Datentypen und Anwendungen benötigen eigene Methoden:
  - Spatial data mining
  - Web mining
  - Bio-KDD
  - Graphen
  - . . .

Evaluation und Visualisierung

- Evaluierung von Mustern und Wissens-(re)präsentation:
  - Visualisierung, Transformation, Elimination von redundanten Mustern, etc.
- Integration von Visualisierung und Data Mining
  - Daten Visualisierung
  - Visualisierung von Data Mining Ergebnissen
  - Visualisierung des Data Mining Prozesses
  - Interaktives „visual data mining"



- Verschiedene Arten von 2D/3D Visualisierungsformen (plots, charts, diagrams), z.B. :
  - box-plots, trees, X-Y-Plots, parallel coordinates
- Anwenden des neu gewonnenen Wissens

Summary

- Data Mining: Bestimmung von interessanten Mustern aus großen Datenmengen
- Entstanden aus Datenbank-Technologie, Machine Learning, Statistik, Visualisierung
  - „in great demand"
  - „with wide applications"
- Der KDD Prozess beinhaltet typw. Data Cleaning, Datenintegration, Data Selection, Transformation, ***Data Mining***, Evaluation, und Wissens-(re)präsentation
- Data Mining Methoden: Charakterisierung, Diskriminierung, Assoziation, Klassifikation, Clustering, Outlier- und Trend-Erkennung, etc.

Wie gehts weiter?

1. Einleitung
2. Daten Repräsentation, Data Warehousing, Business Intelligence (BI)
3. Frequent Pattern Mining
4. Clustering
5. Outlier Detection
6. Classification
7. Regression
8. Further Topics

Wie gehts weiter?

- Wir warden im Rahmen der Vorlesung zu Illustrationszwecken das open source data Mining Tool WEKA verwenden
- Sie finden WEKA unter: https://www.cs.waikato.ac.nz/ml/index.html
- Tipp: laden Sie sich die aktuelle Version 3.8 herunter und spielen Sie sich mit dem Tool
- Falls Sie mit echten Daten spielen wollen, sei Ihnen das UCI Machine Learning Repository empfohlen: http://archive.ics.uci.edu/ml/index.php

Bibliographische Ressourcen

- Data mining and KDD:
  - Conference proceedings: KDD, PKDD, PAKDD, SDM, ICDM etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field:
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, CIKM
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, VLDBJ, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI (Comp. Human Interaction), etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Weiterführende Referenzen

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. 2nd ed., Morgan Kaufmann, 2006.
- T. Imielinski and H. Mannila. *A database perspective on knowledge discovery*. Communications of the ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. *From data mining to knowledge discovery: An overview*. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
- G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- M. Ester and J. Sander. *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer Verlag, 2000 (in German).
- M. H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2003.
- D. Hand, H. Mannila, P. Smyth. *Principles of Data Mining*. MIT Press, 2001.