Data Mining

Einführung

Dr. Hanna Köpcke
Wintersemester 2020

Abteilung Datenbanken, Universität Leipzig
http://dbs.uni-leipzig.de

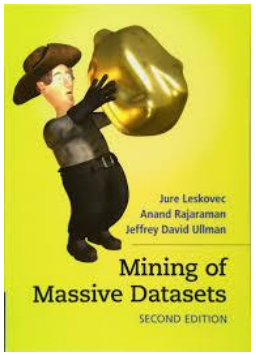
Inhaltsverzeichnis

Organisation

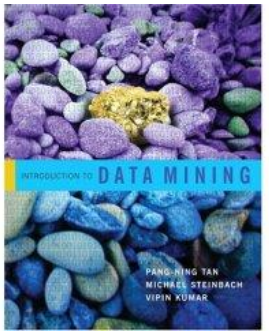
- Anmeldung zur Vorlesung und Prüfung via Almaweb
- Moodle Kurs: https://moodle2.uni-leipzig.de/course/view.php?id=28297
  - Sämtliche Materialien werden dort bereitgestellt
  - Einschreibeschlüssel: dm_ws_2020
- 12 Vorlesungsblöcke bereitgestellt als Videopodcasts (Moodle) jeweils bis zum urspr. Vorlesungstermin (Donnerstags 11:15)
  - Weiterverteilen/Kopieren/Veröffentlichen von Vorlesungsvideos ist nicht gestattet!
- Live-Fragezeit und ggf. praktische Übungen über BigBlueButton
  - Jeder 2. Donnerstag 11:15-12:45
  - Start: 12.11.2020
  - Hier können Sie zu den Inhalten der bereits eingestellten Vorlesungsvideos Fragen stellen
- Prüfung
  - Klausur am Ende des Semesters, 90 Minuten

Literatur

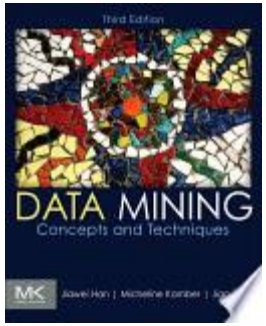
- „Mining of Massive Datasets“
  von Leskovec, Rajaraman und Ullman, Stanford University
  - Buchkapitel, Originalfolien und Videos: http://www.mmds.org/



- „Introduction to data mining”
  von Tan, Steinbach & Kumar (2006)



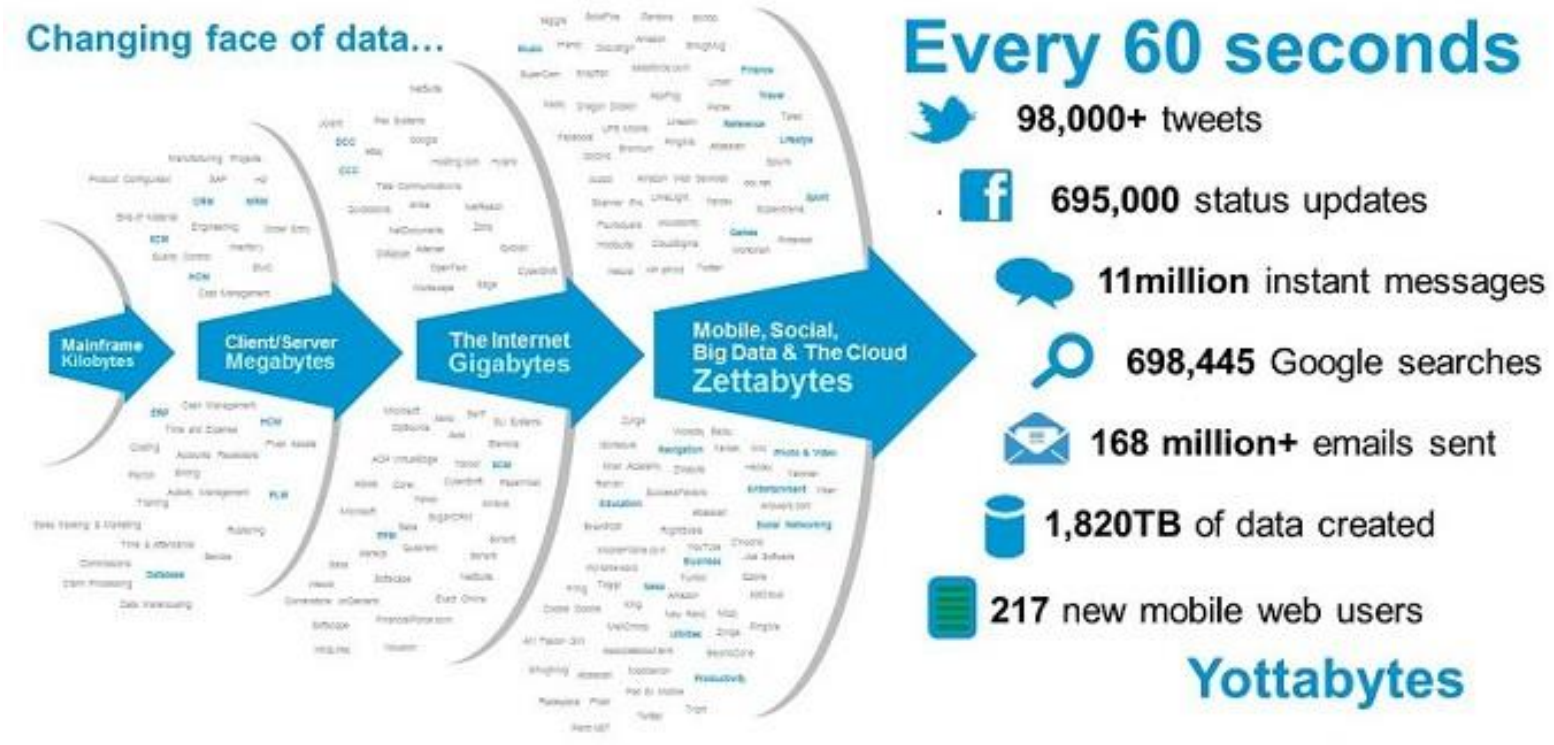
- „Data Mining: Concepts and Techniques”
  von Han, Kamber & Pei (2012)
  - https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques



Lehrveranstaltungen WS 2020

- Lehrveranstaltungen: https://dbs.uni-leipzig.de/stud
  - Datenbanksysteme 1 (DBS1)
  - Vorlesung: Implementierung von Datenbankystemen I
  - Vorlesung: Cloud and Big Data Management
  - Vorlesung: **Data Mining**
  - Praktikum: Data Warehousing
  - Seminar: New Trends in Machine Learning and Data Analytics
  - Oberseminar (Vortrag über laufende Bachelor-/Masterarbeit)

- Verwendung der Data-Mining-Vorlesung:
  - Data-Science-Modul Skalierbare Datenbanktechnologien (SDBT) 1
  - Bachelor-Modul „Realisierung von Informationssystemen“ (5LP, zwei Vorlesungen)
  - Master-Modul „Moderne Datenbanktechnologien”
    - (kleines) Kernmodul (5LP, zwei Vorlesungen)
    - (großes) Vertiefungsmodul (10LP, zwei Vorlesungen + Praktikum/Seminar)

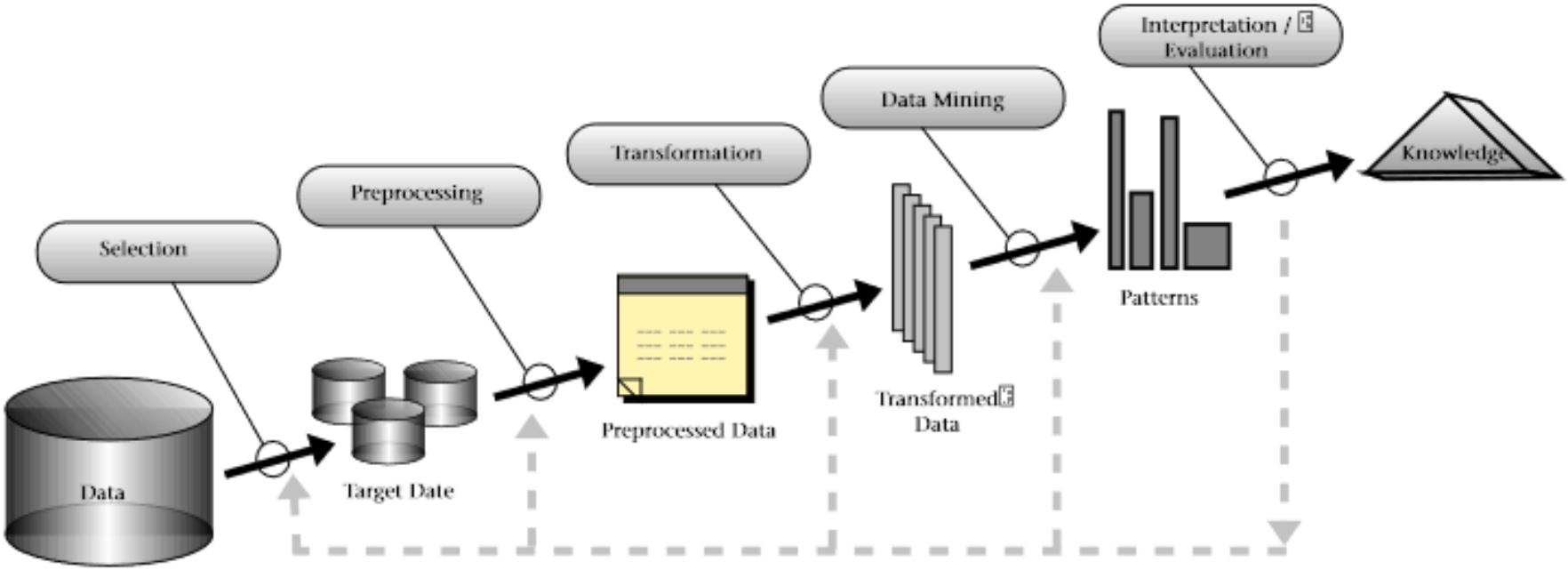


Bildquelle: https://community.hpe.com/t5/image/serverpage/image-id/32999iD48A69B4124853D6?v=v2

Data Mining: Definition

- „Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.” (Han, Kamber & Pei)
- „Data mining is the study of collecting, cleaning, analyzing, and gaining useful insights from data” (Aggarwal)
- „Data mining is the process of automatically discovering useful information in large data repositories” (Tan, Steinbach & Kumar)
- „Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data.” (Zaki, Meira)

Data Mining und Knowledge Discovery

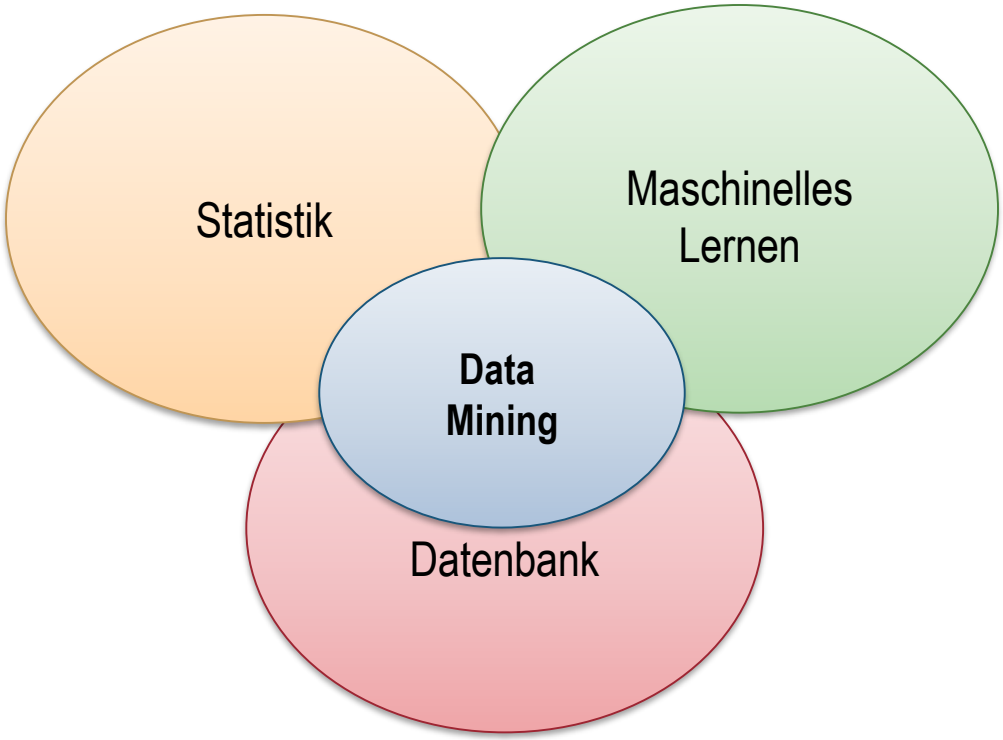


Source: Fayyad, et.al., 1996

Data Mining: Modell

- Ergebnis eines DM-Verfahrens: **Modell** (auch: Muster)
  - **Deskriptiv**: für Menschen verständliche Zusammenfassung
  - **Prädiktiv**: Vorhersage unbekannter Werte aus bekannten Werten
  - Statistisch: Annahme einer Wahrscheinlichkeitsverteilung und Schätzen der Parameter

Die Ursprünge



CRISP-DM: Standard-Vorgehensmodell für Data Mining

Bildquelle: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#/media/File:CRISP-DM_Process_Diagram.png

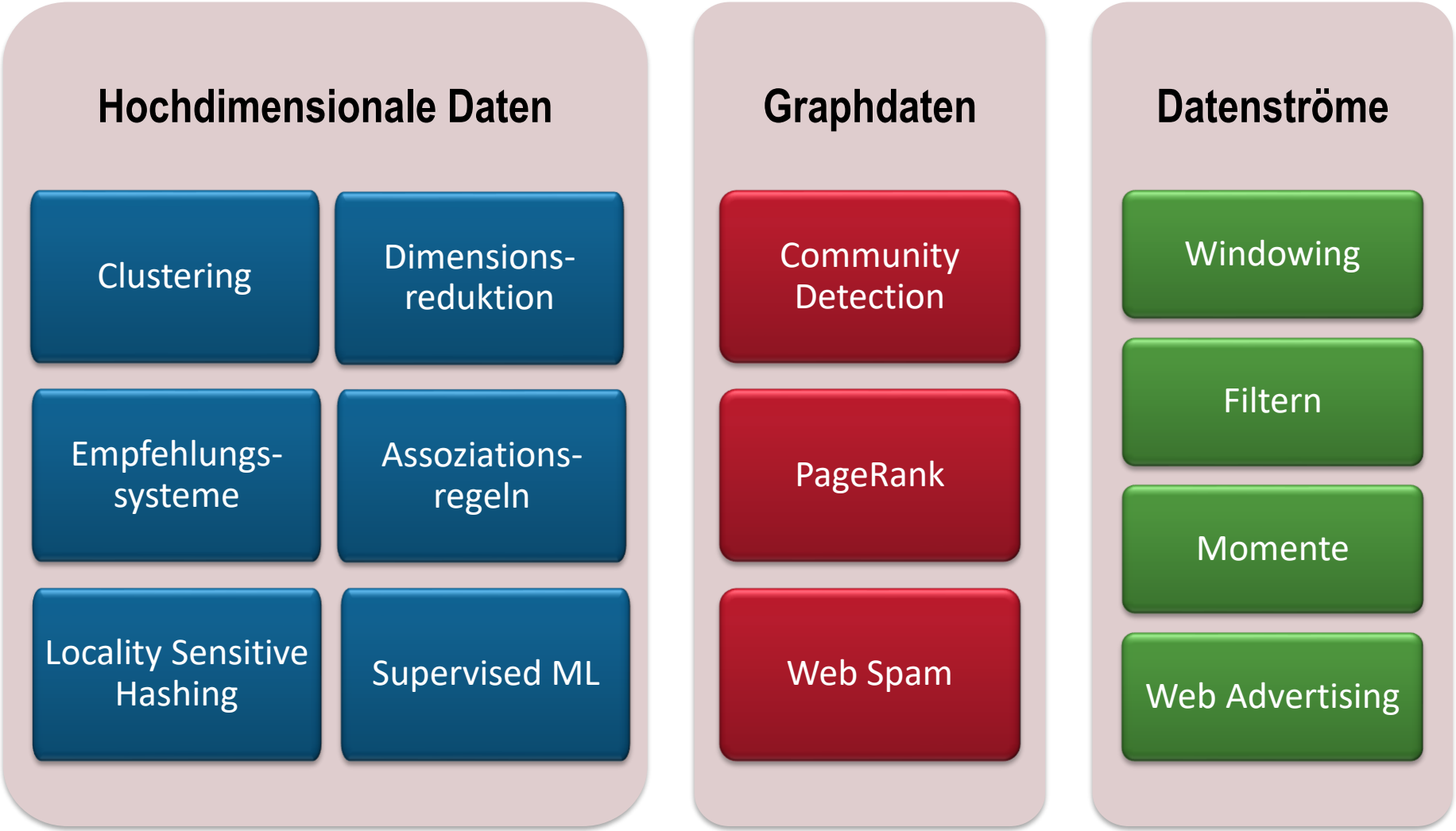
Abgrenzung zu anderen Vorlesungen

- **Schwerpunkt dieser Vorlesung: skalierbare Algorithmen**
- Data Mining ist letztes Kapitel der Vorlesung „*Data Warehousing*“
- *Cloud und Big Data Management*: Ausführliche Behandlung von Technologien für verteilte Datenhaltung und -verarbeitung
- **Keine Datenbanken** (DBS 1+2, *Mehrrechner-Datenbanken, NoSQL*)
- **Keine Inferenz** (*Statistisches Lernen*)
- Fokus auf das **Gebiet der Datenanalyse**

Lernziele

- Kenntnis skalierbarer **Algorithmen** zur **Analyse von Daten**
  - Nachvollziehen der **Funktionsweise** der Algorithmen
  - **Anwendung** der Algorithmen an Beispieldaten
  - Beurteilung der **Anwendbarkeit** von Algorithmen (Komplexität, Engpässe)
  - **Vergleich** verschiedener Algorithmen

Übersicht


Hochdimensionale Daten
Clustering
Dimensions-reduktion
Empfehlungs-systeme
Assoziations-regeln
Locality Sensitive Hashing
Supervised ML
Graphdaten
Community Detection
PageRank
Web Spam
Datenströme
Windowing
Filtern
Momente
Web Advertising