Non-Standard-Datenbanken und Data Mining

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun und Felix Kuhr (Übungen)

Organisatorisches: Übungen

- **Start**: Freitag, 18. Oktober 2017
- **Zwei Übungen**: ab 18.10. Fr. 8-10 Uhr , IFIS, Geb. 64, Raum Banach und Cook/Karp (Anmeldung über Moodle notwendig)
  - Erste Ausgabe eines Übungsblattes am 17.10. ab 18 Uhr über Moodle
- **Übungsaufgaben** stehen jeweils nach der Vorlesung am Donnerstag ca. ab 18 Uhr über Moodle bereit
- **Abgabe der Lösungen** erfolgt bis Mittwoch 12 Uhr in der IFIS-Teeküche (jeweils in der Woche nach der Ausgabe, 1 Kasten pro Übungsgruppe)
  - Erste Abgabe am 23.10.
- Aufgaben können in einer **2-er Gruppe** bearbeitet werden (also bitte Name(n), Matrikelnummer(n) und Übungsgruppennummer vermerken)
- In den Übungen am Freitag wird der Übungszettel besprochen, dessen Lösungen bis zum jeweils vorigen Mittwoch abgegeben wurde, und es werden auch Fragen zum jeweils neuen Übungszettel geklärt (ggf. mit Präsenzaufgaben als Hilfestellung)

Organisatorisches: Prüfung

- Die **Eintragung in den Kurs** und in eine Übungsgruppe ist **Voraussetzung**, um an dem Modul Non-Standard-Datenbanken teilnehmen zu können
- Am Ende des Semesters findet eine **Klausur** statt
- **Voraussetzung** zur Teilnahme an der Klausur sind mindestens **50% der gesamtmöglichen Punkte aller Übungszettel**

Teilnehmerkreis und Voraussetzungen

Studiengänge

- Bachelor **Informatik**
- Bachelor **IT-Sicherheit (?)**
- Bachelor **Mathematik in Medizin und Lebenswissenschaften**
- Bachelor **Medieninformatik**
- Bachelor **Medizinische Informatik**
- Master **Informatik**
- Master **MML**

Voraussetzungen (Bestehen der unten genannten Veranstaltungen ist keine Teilnahmebedingung für NDB-Klausur)

- Algorithmen und Datenstrukturen
- Lineare Algebra und Diskrete Strukturen
- Datenbanken
- Kontextfreie Grammatiken

Vorteilhaft

- Einführung in die Logik

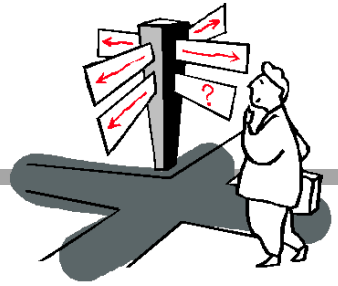

Was sind Merkmale von
Standard-Datenbanken?

Merkmale von Standard-Datenbanken

- Datenmodell: relational („Tabellen“ und „Tupel“)
- Annahmen:
  - Strukturen fix (Schemaänderung möglich, aber aufwendig)
    - Verwerfen d. A. führt zu semistrukturierten Datenbanken und zu Graphdatenbanken
  - Abstrakte Entitäten mit abstr. Assoziationen (Tupel in Relationen)
    - Verwerfen führt zu temporalen, sequenzorientierten, räumlichen, und multimodalen[1] Datenbanken
  - Daten persistent, Anfragen bzgl. Schnappschuss einmal beantwortet
    - Verwerfen führt zu stromorientierten Datenbanken
  - Daten enthalten feste Werte bzw. Referenzen
    - Verwerfen führt zu Datenbanken für unsichere und unvollständige Information

[1] Datenbanken mit Text-, Bild-, Audio-, Videodaten mit relationalen Metadaten sowie relationalen Annotationen

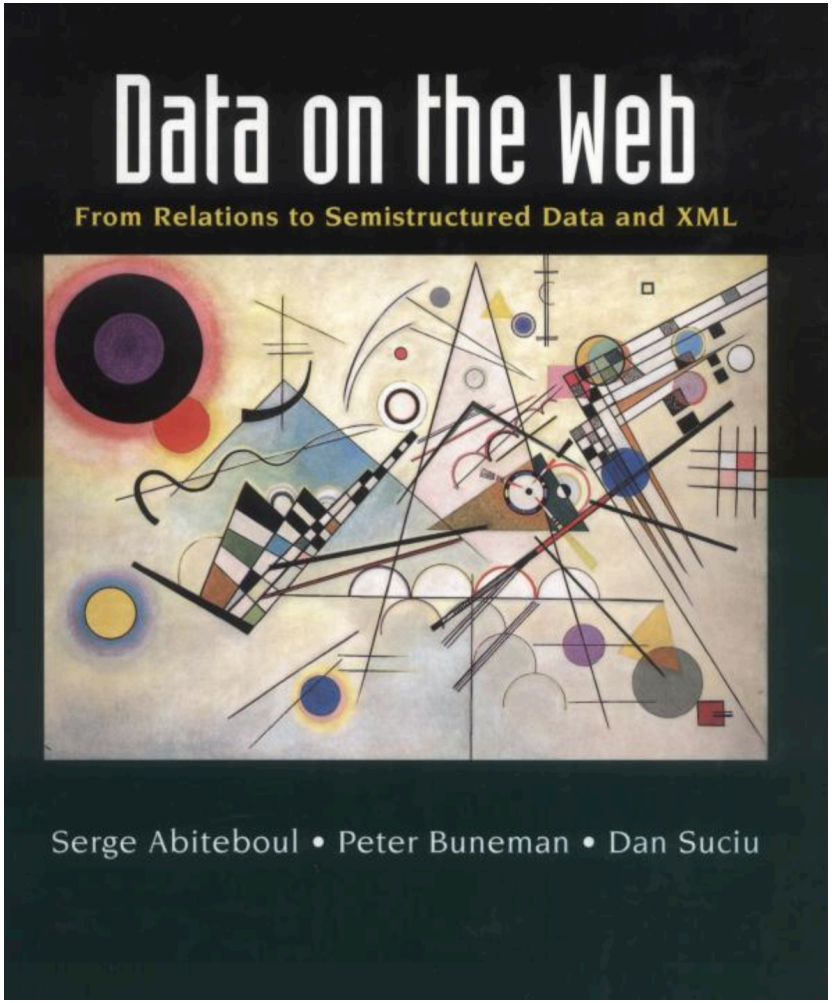
Non-Standard-Datenbanken



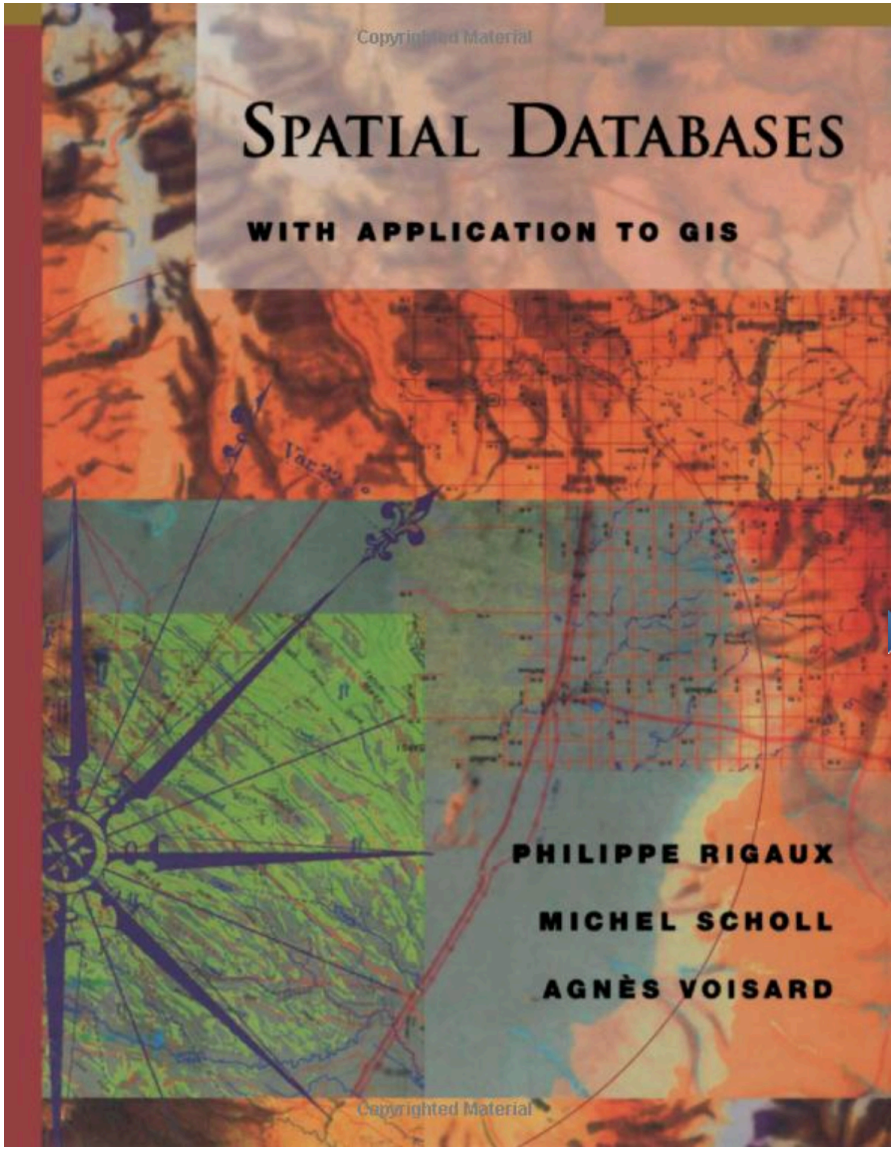
- **Semistrukturierte** Datenbanken (XML)
- **Räumliche und multimodale** Datenbanken
  - lineare und mehrdimensionale Strukturen
- **Temporale Datenbanken**
  - zeitlich beschränkte Gültigkeiten
- Datenbanken für **Datenströme** (Fensterkonzept)
- **Bewertung** von Antworten (Top-k-Anfragen)
- **Probabilistische** Datenbanken zur Repräsentation **unsicherer Information**
- Next Generation Databases, **Graphdatenbanken**

Literatur

1999

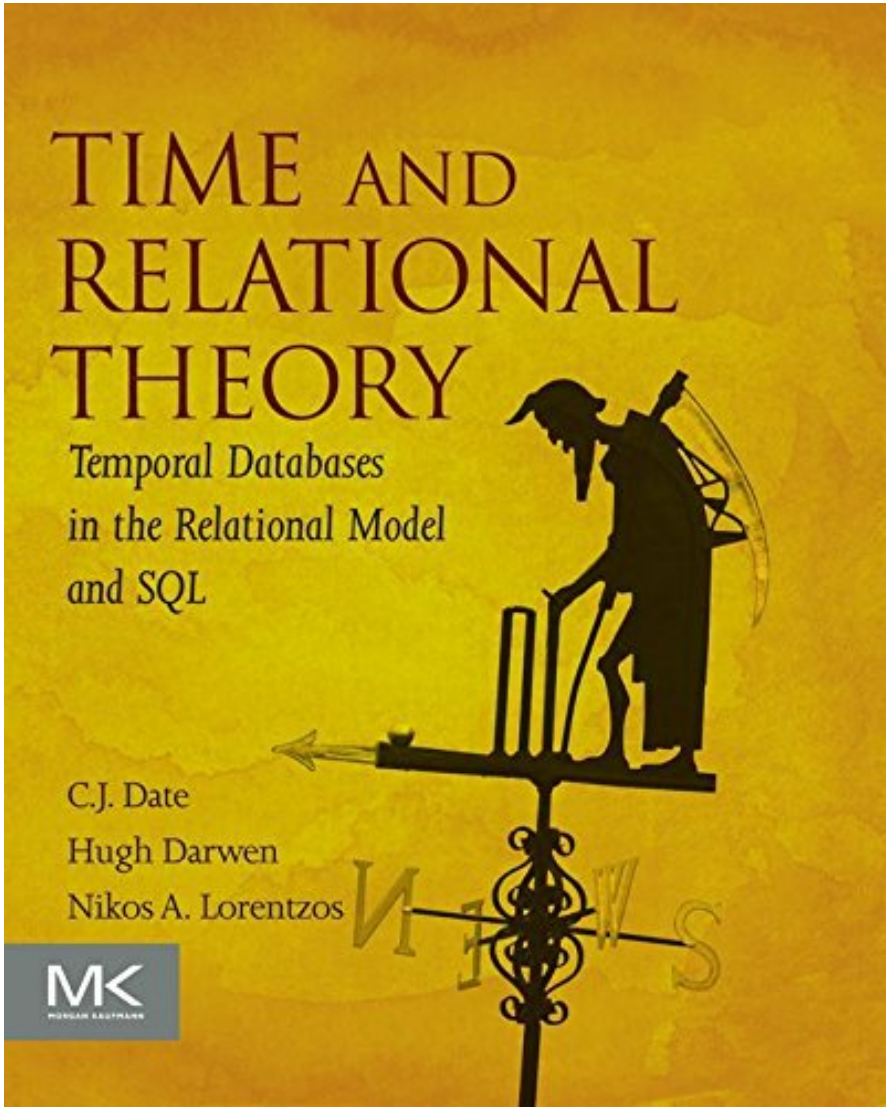


2001

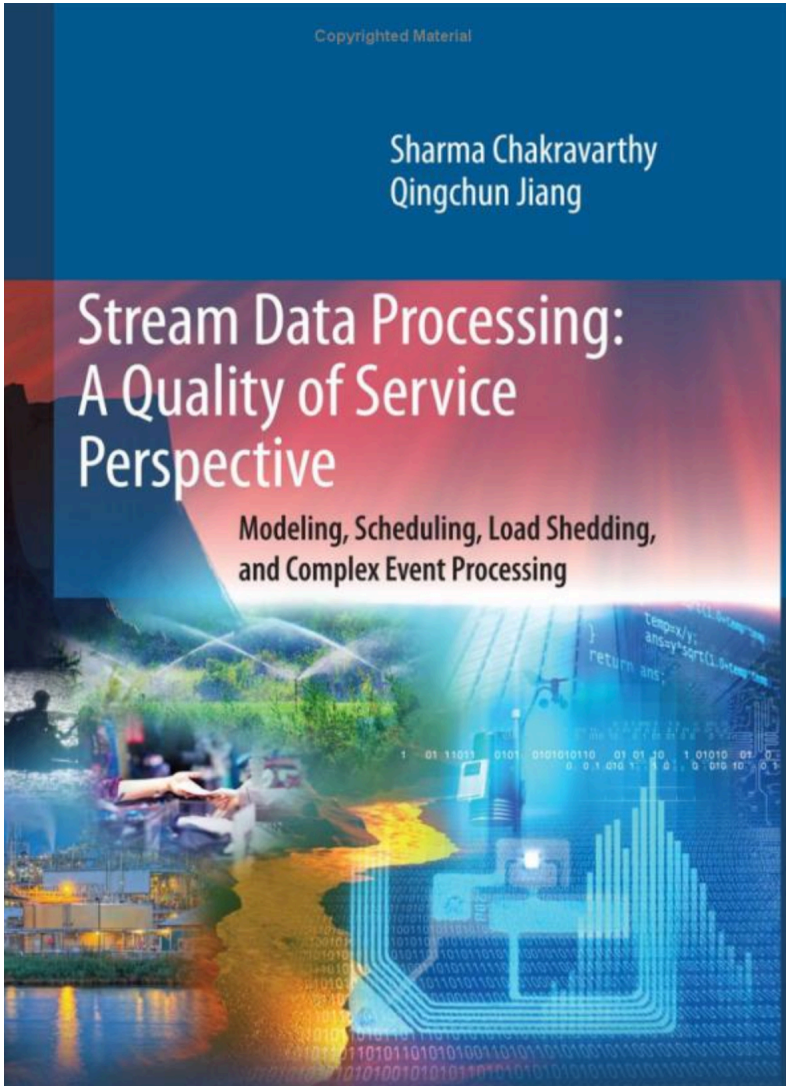


Literatur

2014

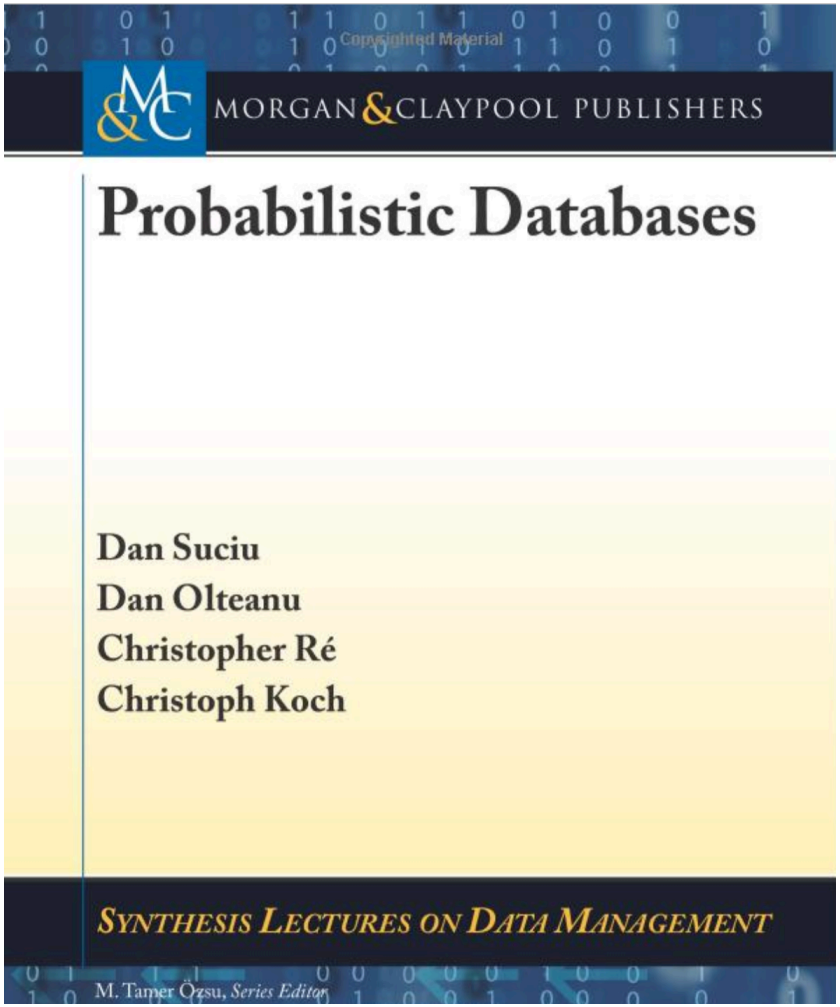


2009

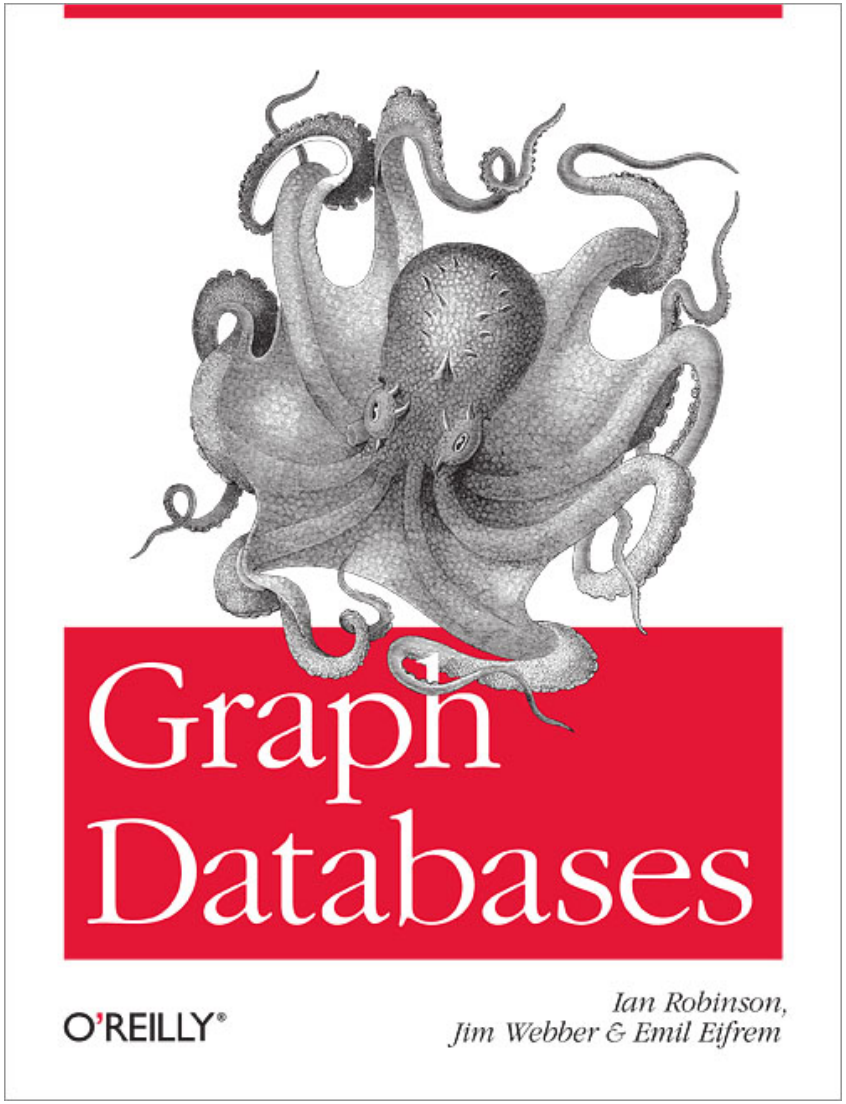


Literatur

2011

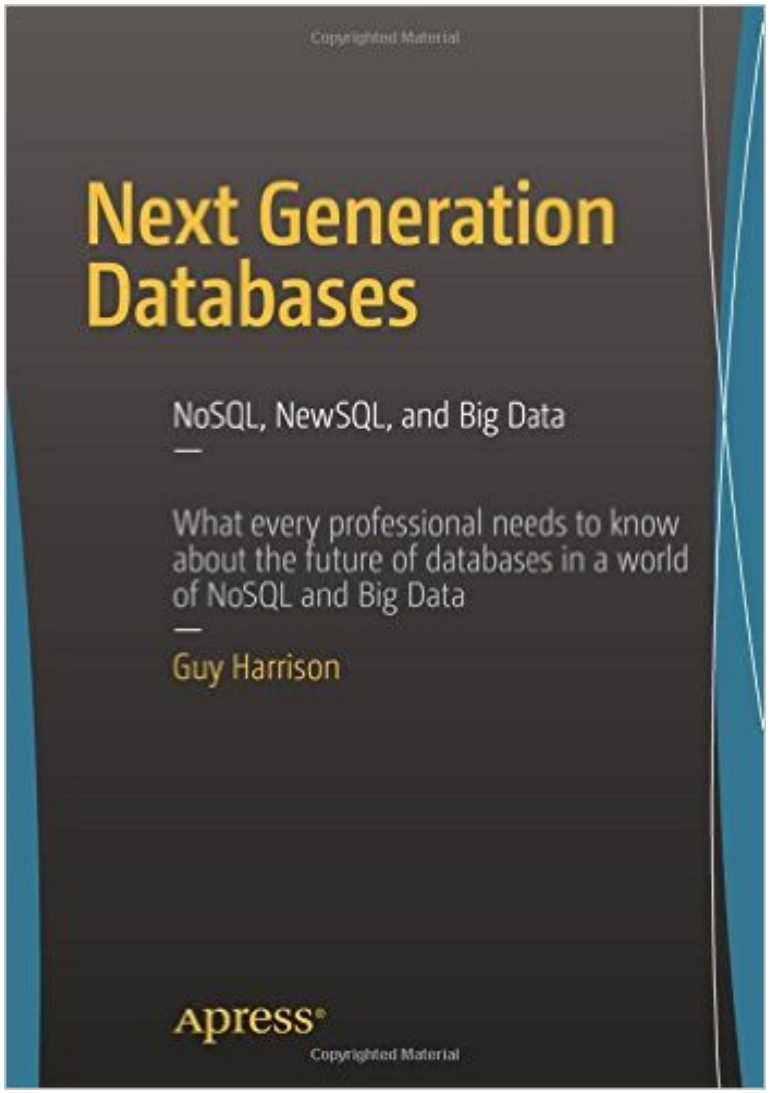


2016

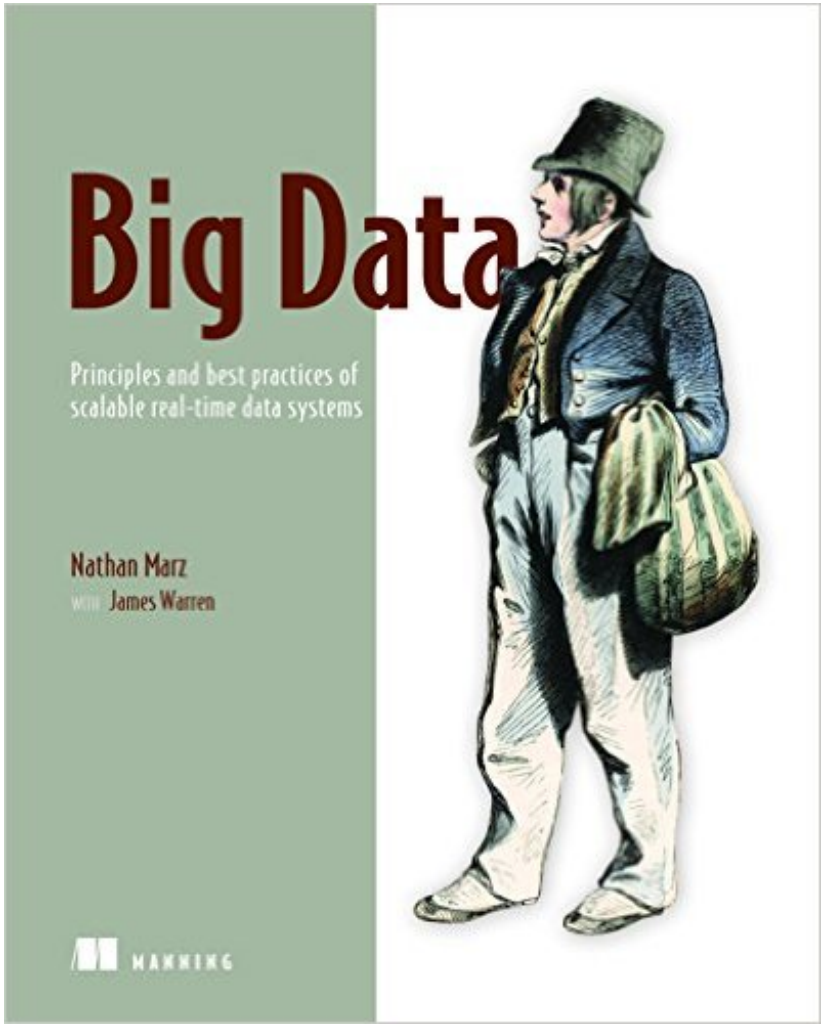


Literatur

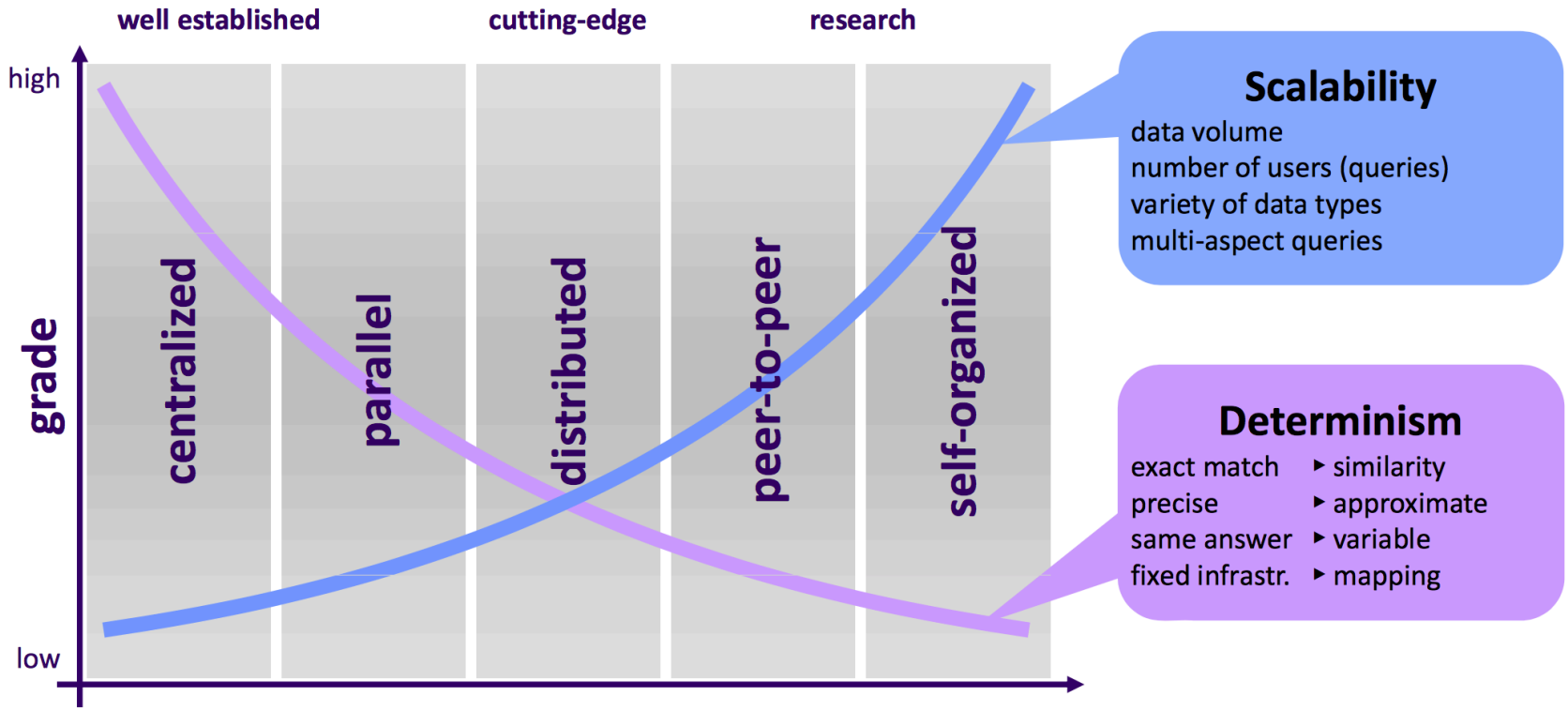
2015



2015



Datenhaltungstypen und Anfragebeantwortung

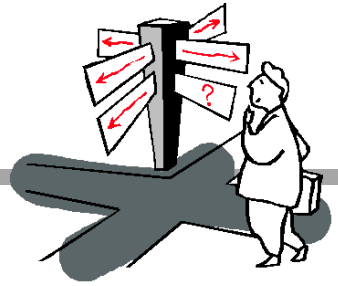


G.Navarro and V.Pestov(Eds.): SISAP2012, LNCS7404, pp.8–24, 2012. © Springer-Verlag Berlin Heidelberg 2012

Anfragesprachen für Datenbanken

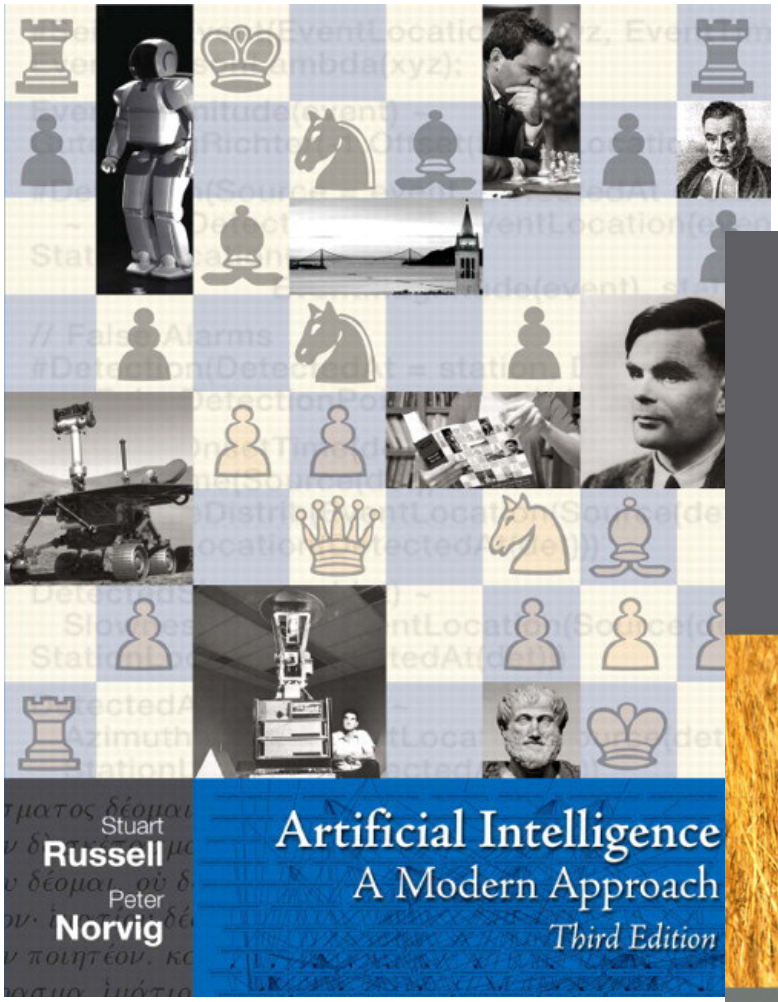
- Semantik und Ausdrucksstärke
  - Was braucht man in der Anwendung?
- Ausdrucksstärke vs. Skalierbarkeit
  - Betrachtung der Komplexität des Anfragebeantwortungsproblems für eine gegebene Anfragesprache
  - Datenkomplexität
    - Wie wirkt sich eine Verdopplung des Datenbestandes bei fixer Anfrage auf die Worst-Case-Laufzeit der besten Anfragebeantwortungsalgorithmen aus?
  - Kombinierte Komplexität
    - Laufzeit bezogen auf Anfragelänge und Datenmenge (relevant aber selten betrachtet, da Anfragelänge klein)

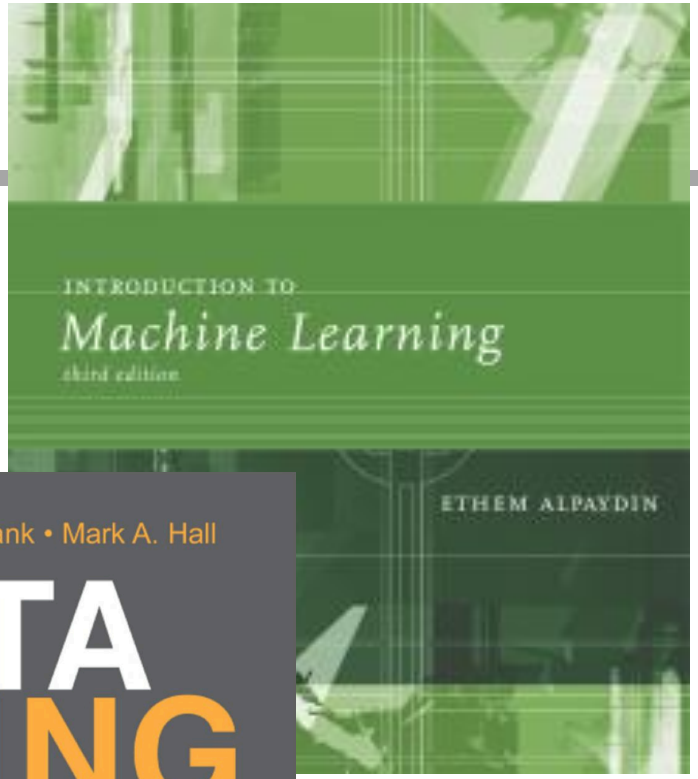
Data Mining

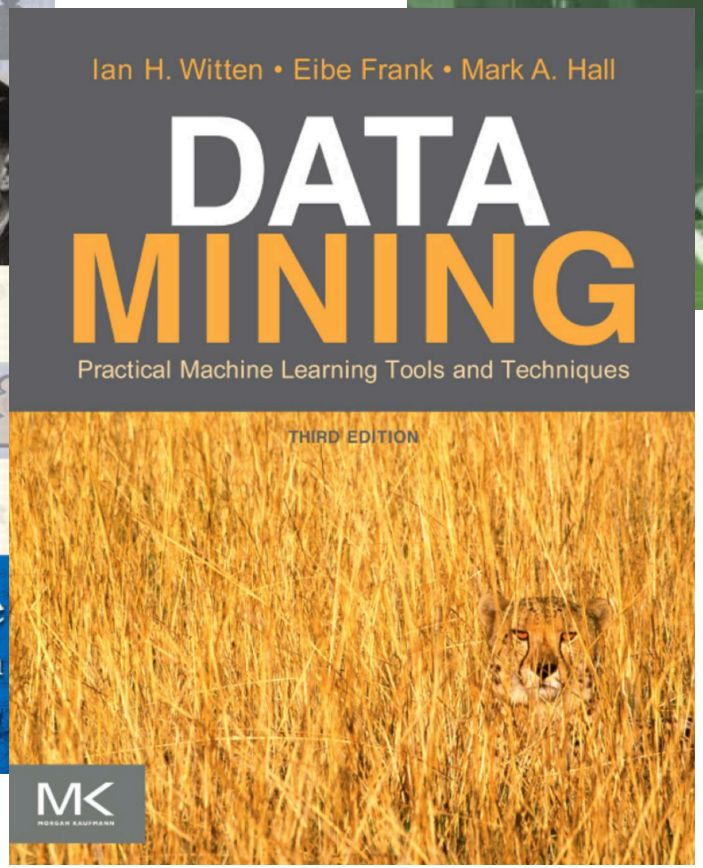


- Gewinnung von
  Entscheidungsfunktionen (Klassifikation) oder
  Berechnungsfunktionen (Regression) aus Daten
  - Statische Daten
  - Über der Zeit eintreffende Daten (Stream Data Mining)
  - Daten mit Zeitstempel oder auch Gültigkeitsinformation (Historical Data Mining)
- Extraktion von **relationalen Beschreibungen** aus Texten bzw. Vektordaten (z.B. Bildern, Audiodaten, Videodaten)
- Umgang mit unsicheren und unvollständigen Daten

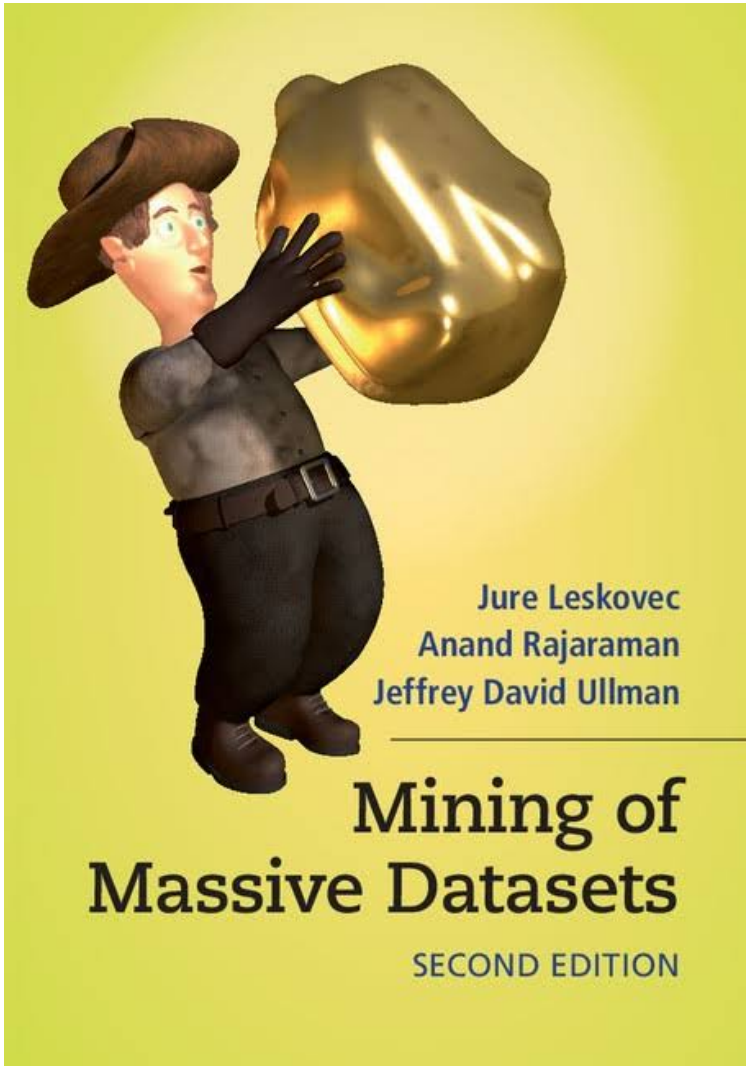
Literatur

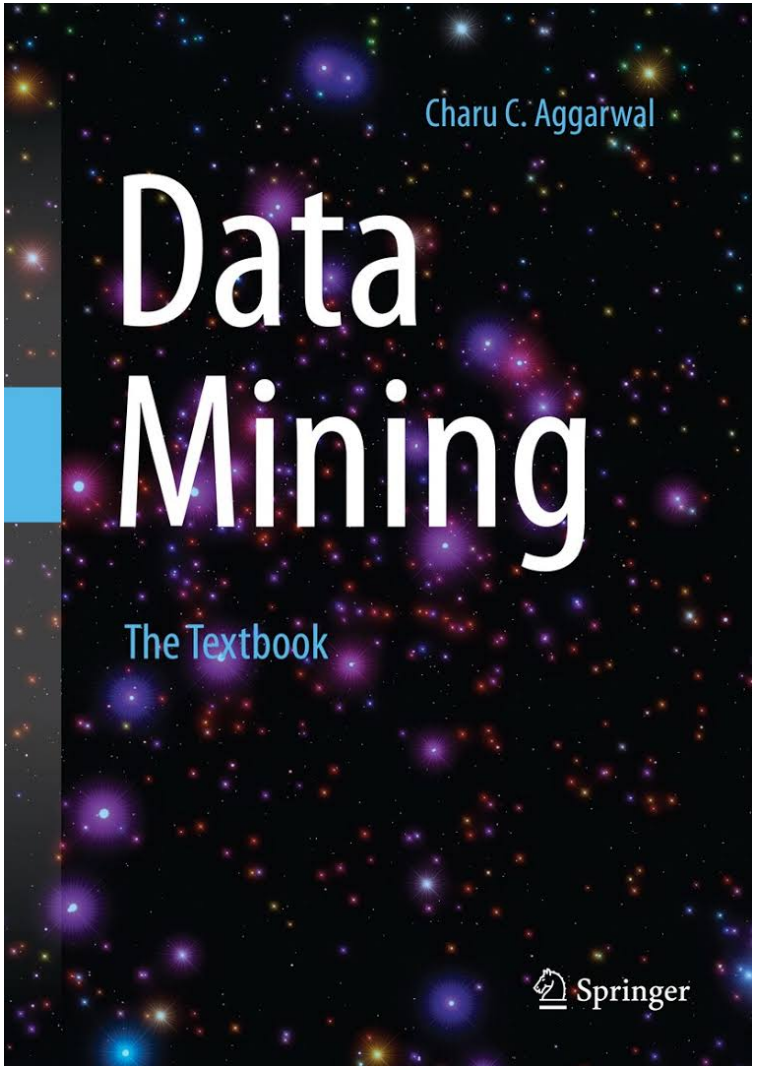

Stuart
Russell
Peter
Norvig
Artificial Intelligence
A Modern Approach
Third Edition


INTRODUCTION TO
Machine Learning
third edition
ETHEM ALPAYDIN

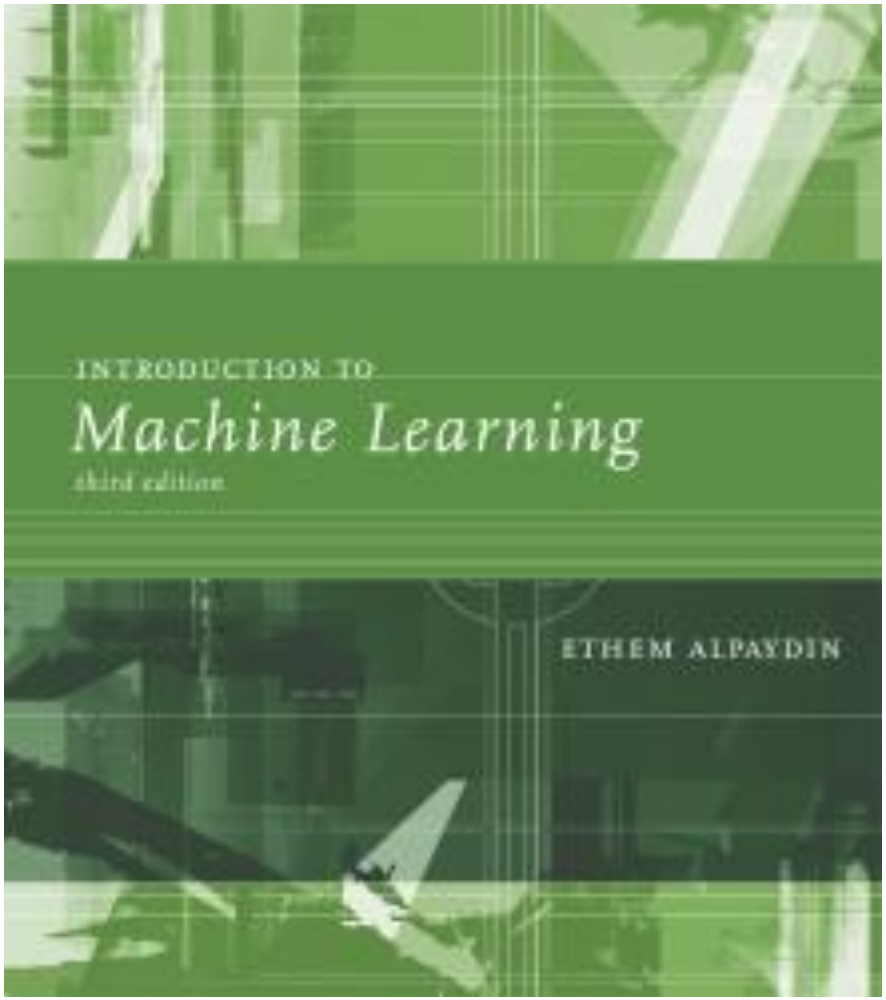

Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

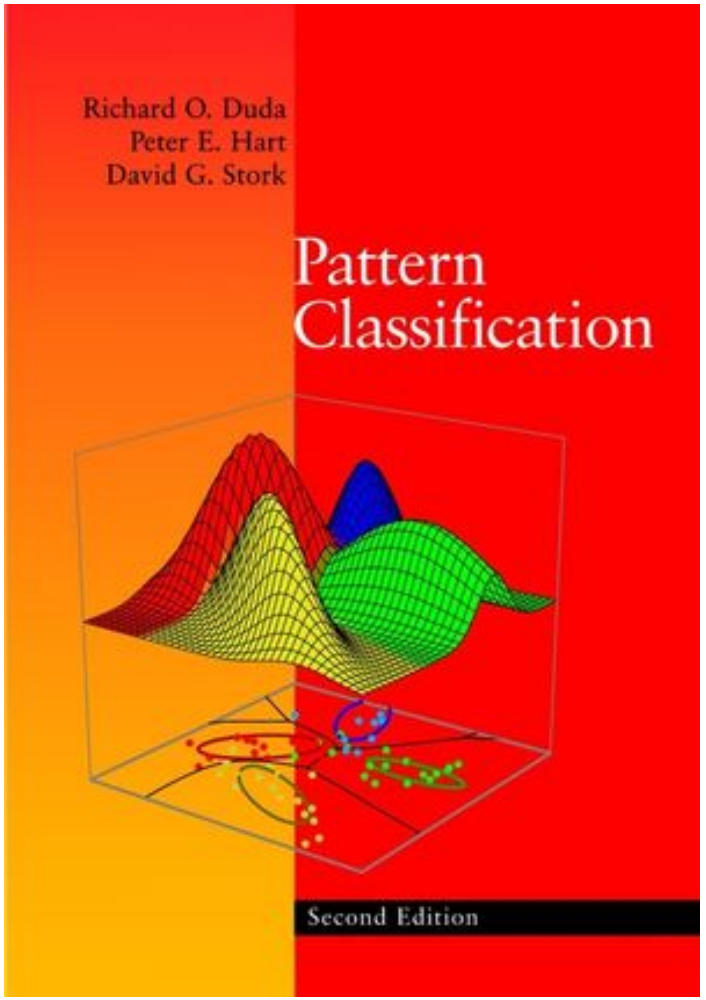
Literatur


Jure Leskovec
Anand Rajaraman
Jeffrey David Ullman
Mining of
Massive Datasets
SECOND EDITION


Charu C. Aggarwal
Data
Mining
The Textbook
Springer

Literatur


INTRODUCTION TO
Machine Learning
third edition
ETHEM ALPAYDIN


Richard O. Duda
Peter E. Hart
David G. Stork
Pattern
Classification
Second Edition

Übersicht

- Semistrukturierte Datenbanken (JSON, XML) und Volltextsuche
- Information Retrieval
- Mehrdimensionale Indexstrukturen
- Cluster-Bildung
- Einbettungstechniken
- Array-Datenbanken
- First-n-, Top-k-, und Skyline-Anfragen
- Probabilistische Datenbanken, Anfragebeantwortung, Top-k-Anfragen und Open-World-Annahme
- Probabilistische Modellierung, Bayes-Netze, Anfragebeantwortungsalgorithmen, Lernverfahren, Verallgemeinerung: Belief Functions, Dempster-Shafer Theorie der Evidenz
- Temporale Datenbanken und das relationale Modell, Zeitreihen in Array-Datenbanken, TimeScaleDB
- Data Mining auf Zeitreihen (SAX, Matrix Product), Probabilistische Temporale Datenbanken
- Dynamische Bayessche Netze, Inferenzalgorithmen und Lernverfahren
- Stromdatenbanken, Prinzipien der Fenster-orientierten inkrementellen Verarbeitung
- Approximationstechniken für Stromdatenverarbeitung, Stream-Mining
- Probabilistische raum-zeitliche Datenbanken und Stromdatenverarbeitungsssysteme: Anfragen und Indexstrukturen, Raum-zeitliches Data Mining
- Von NoSQL- zu NewSQL-Datenbanken, CAP-Theorem, Blockchain-Datenbanken
- Analyse von Graphdaten