
Non-Standard-Datenbanken und Data Mining

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

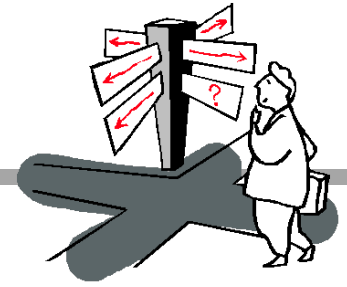
Was sind Merkmale von
Standard-Datenbanken?



Merkmale von Standard-Datenbanken

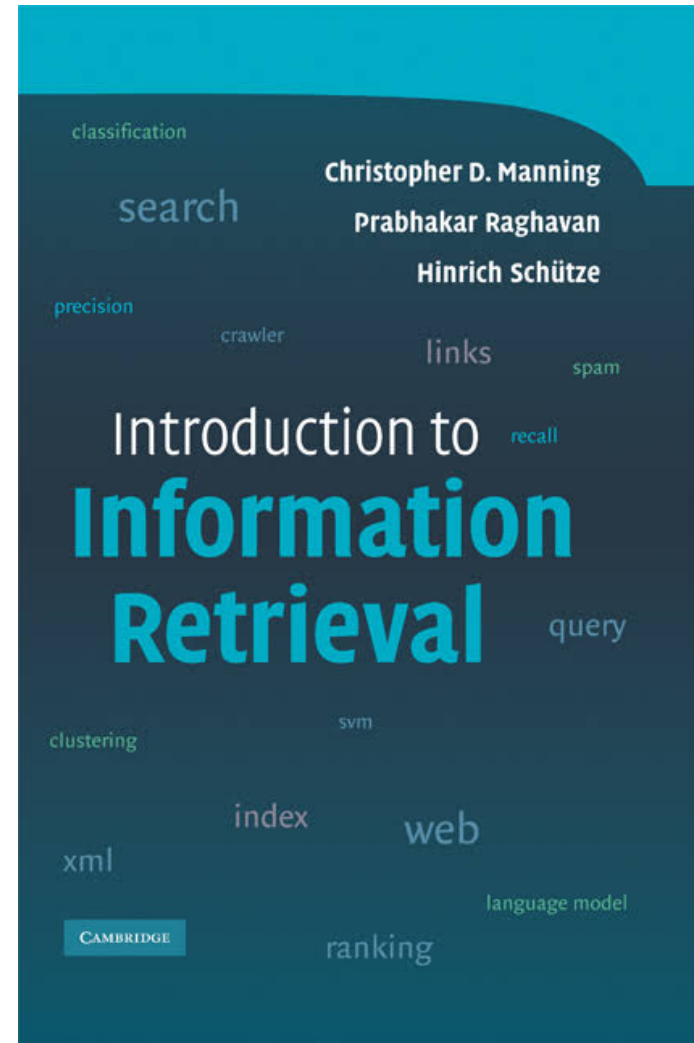
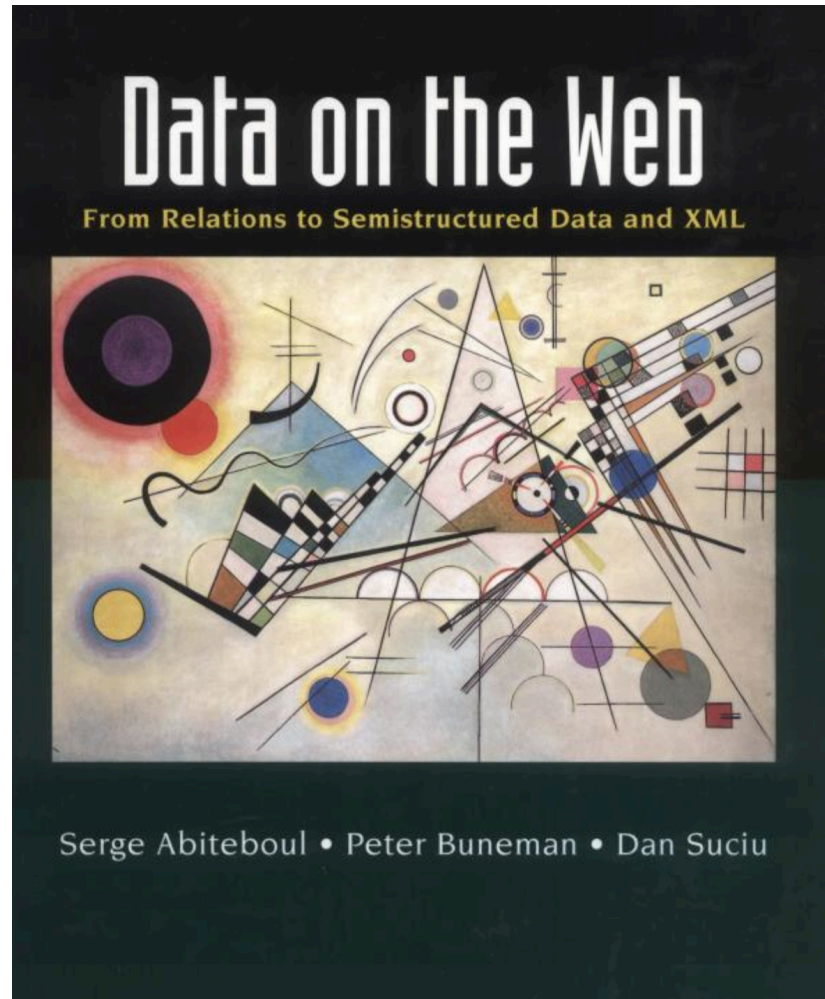
- Datenmodell: relational („Tabellen“ und „Tupel“)
- Annahmen:
 - Strukturen fix (Schemaänderung möglich, aber aufwendig)
 - Verwerfen d. A. führt zu semistrukturierten Datenbanken und zu Graphdatenbanken
 - Abstrakte Entitäten mit abstr. Assoziationen (Tupel in Relationen)
 - Verwerfen führt zu temporalen, sequenzorientierten, räumlichen, und multimodalen¹ 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

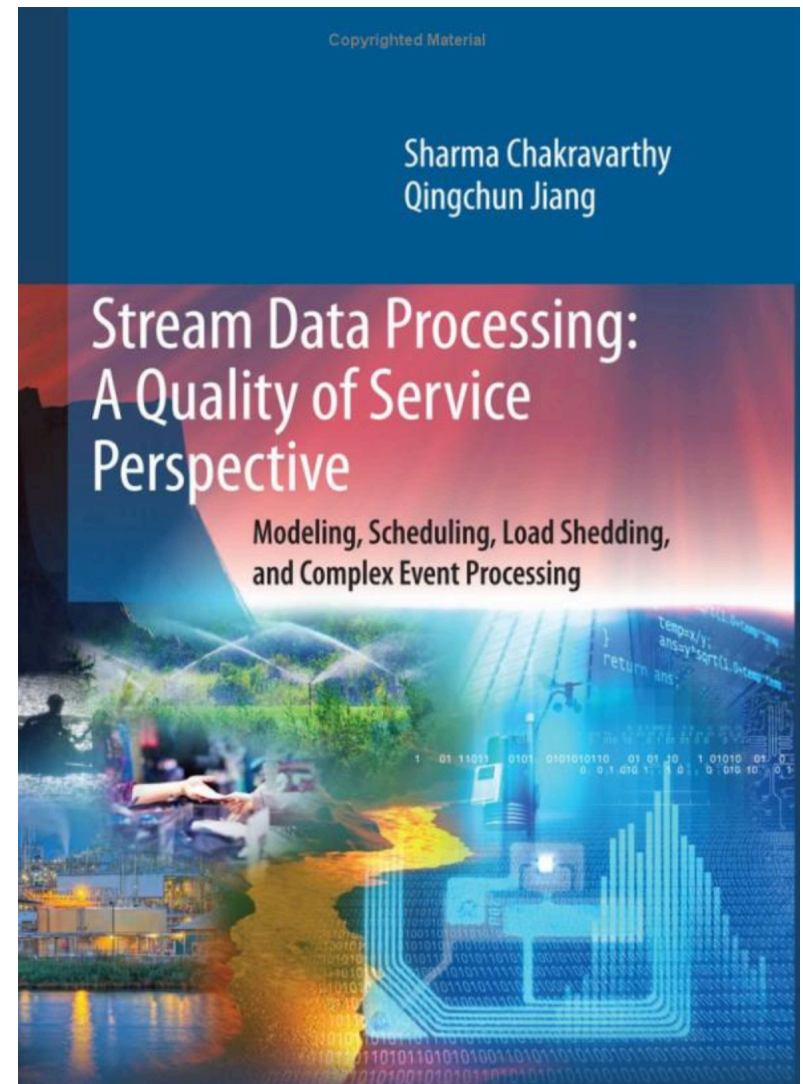
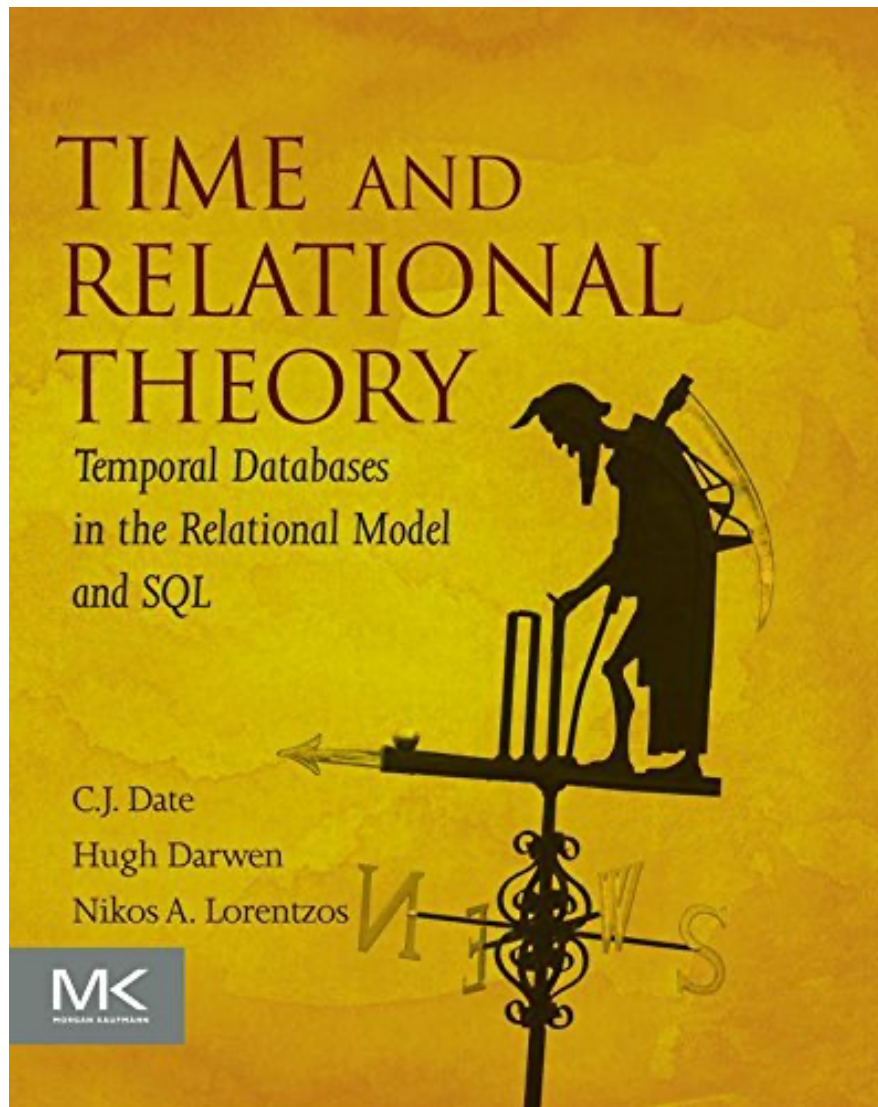
Non-Standard-Datenbanken

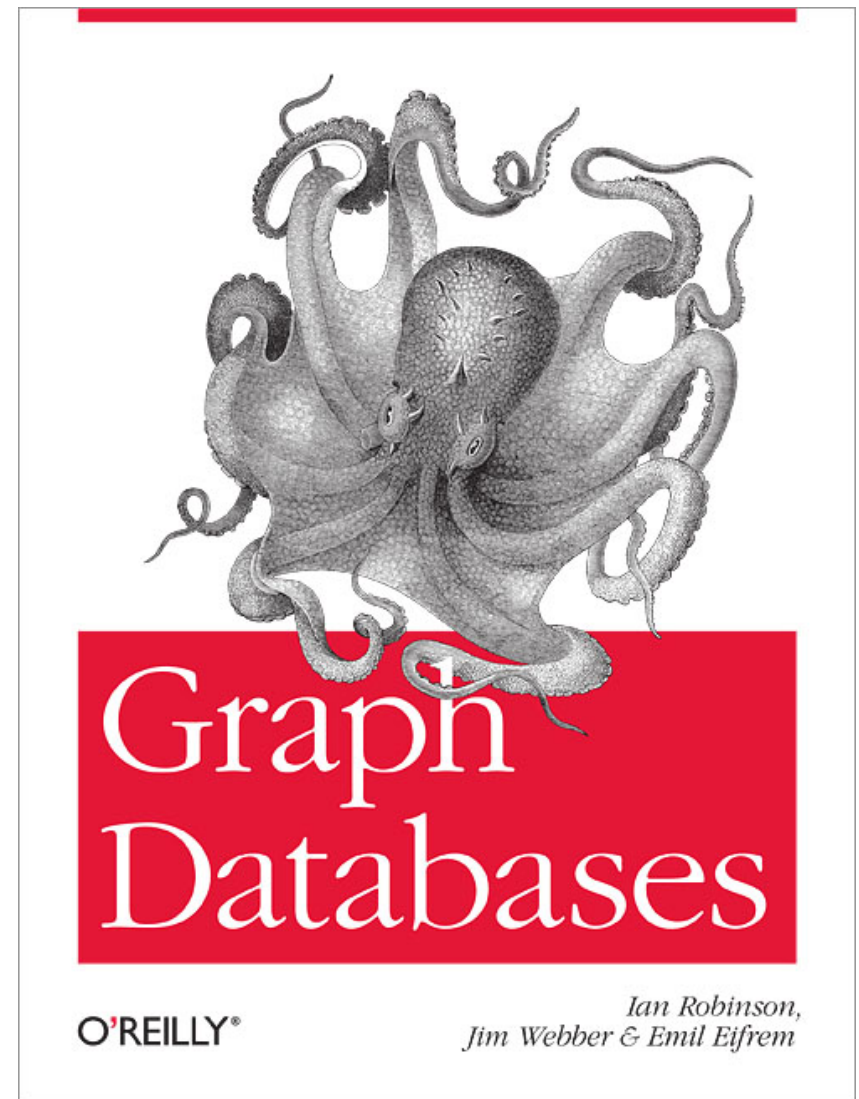
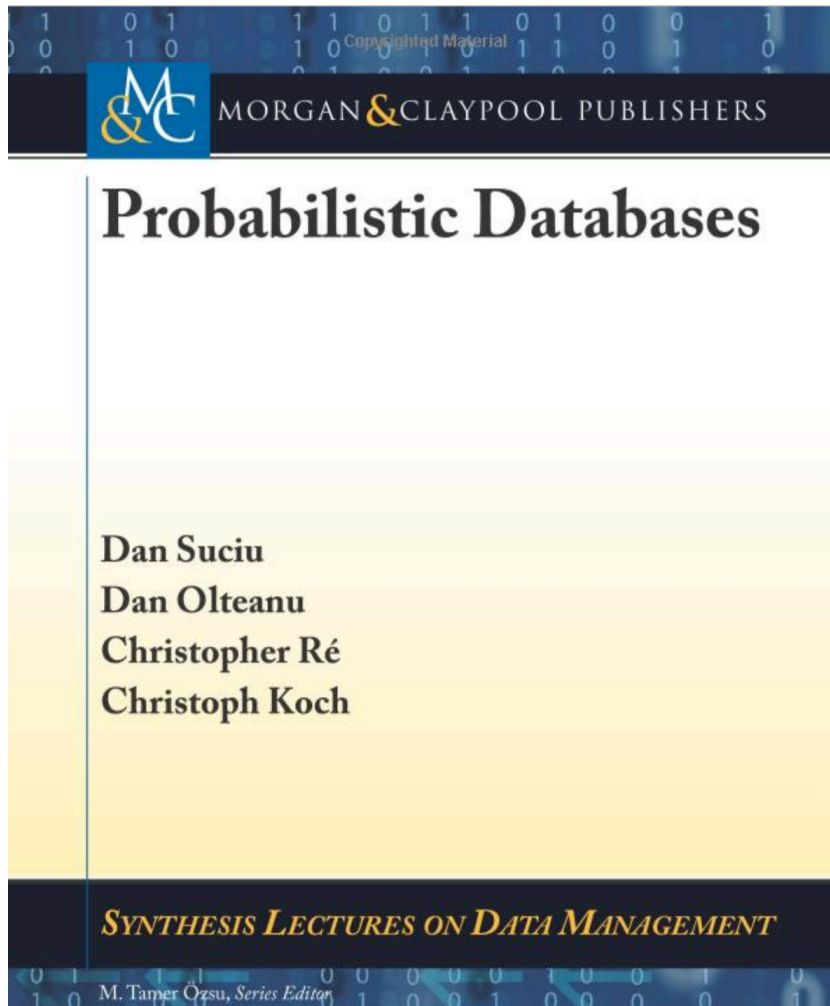


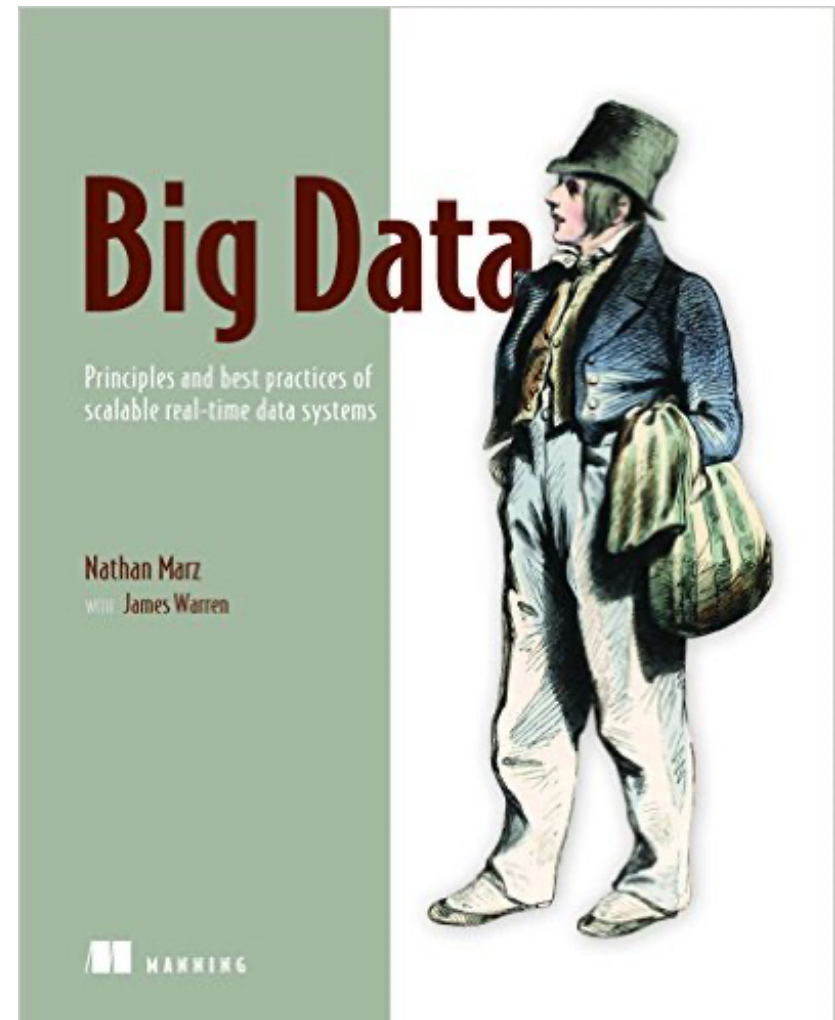
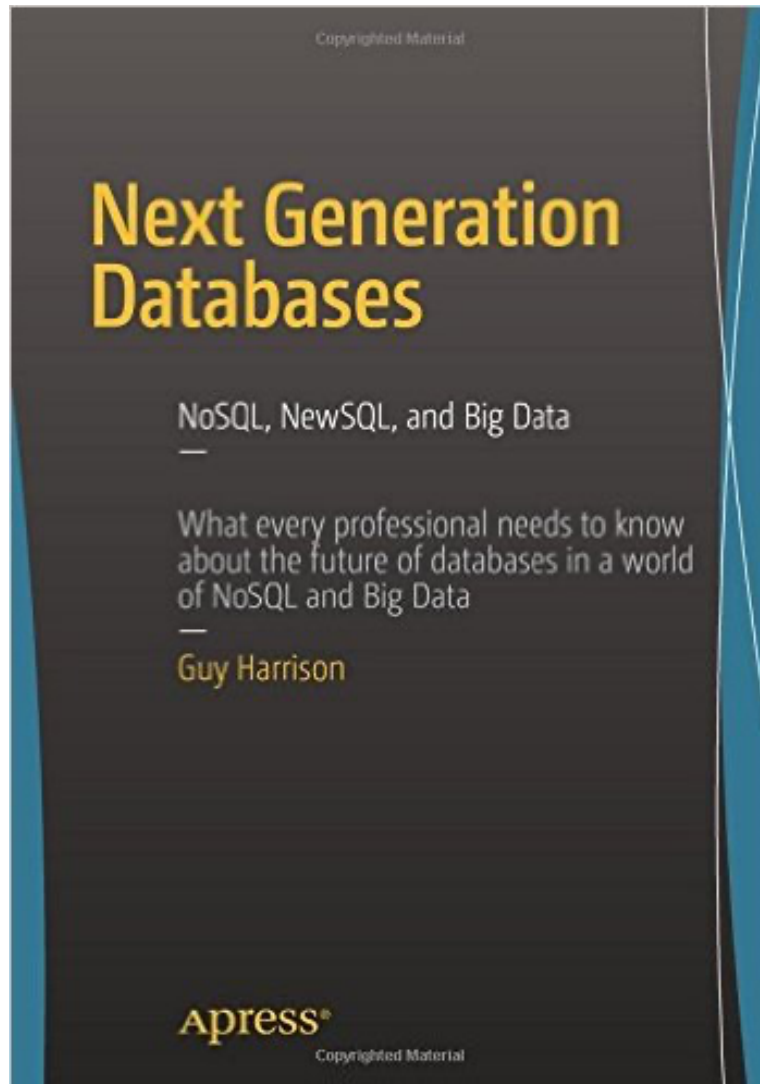
- **Semistrukturierte** Datenbanken (XML)
- **Räumliche und multimodale** Datenbanken
 - lineare und mehrdimensionale Strukturen
- **Temporale Datenbanken**
 - zeitlich beschränkte Gültigkeiten
- Datenbanken für **Datenströme** (Fensterkonzept)
- **Bewertung** von Antworten (Top-k-Anfragen)
- **Probabilistische**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



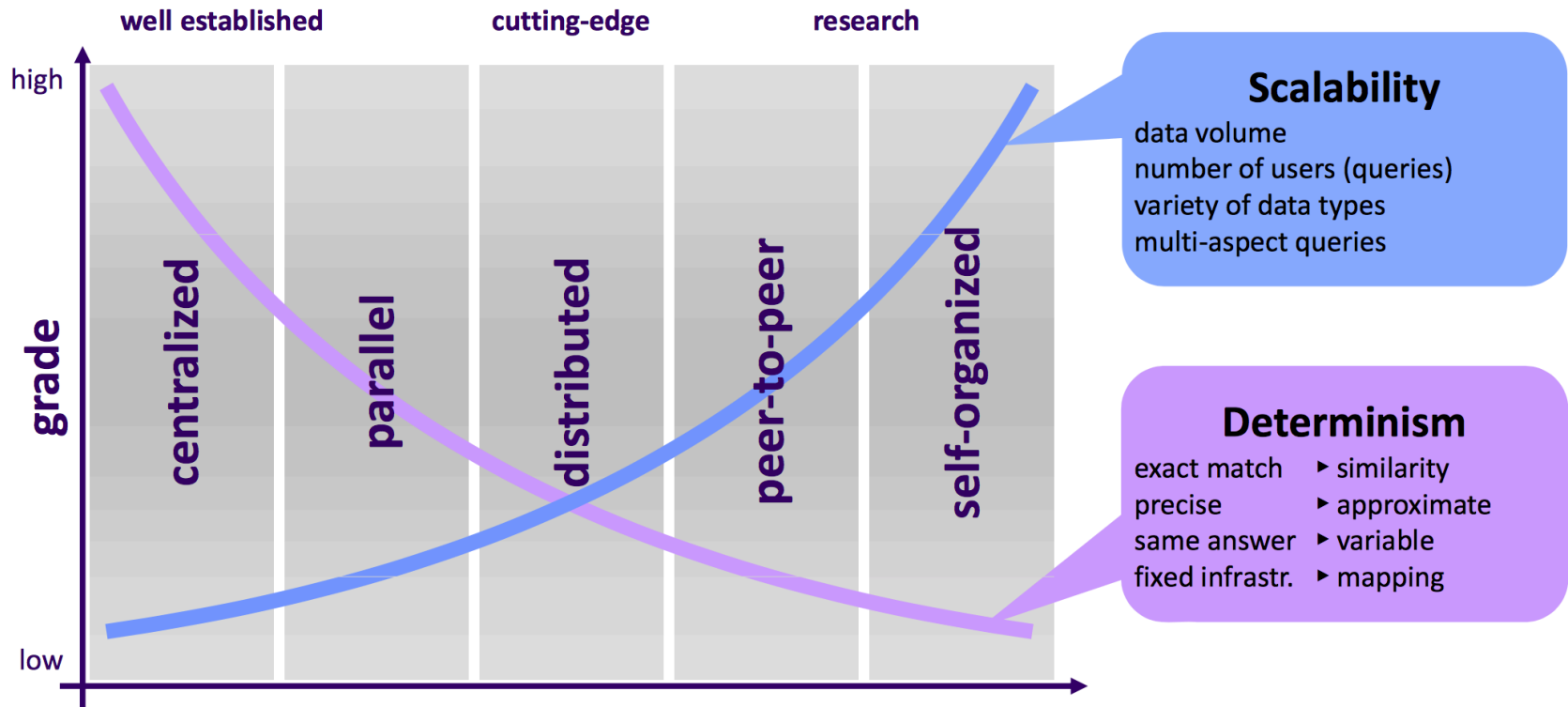






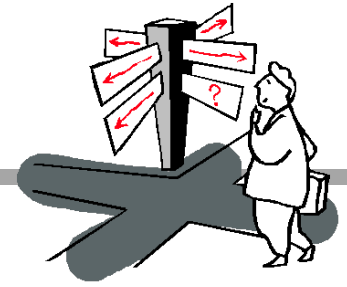


Datenhaltungstypen und Anfragebeantwortung



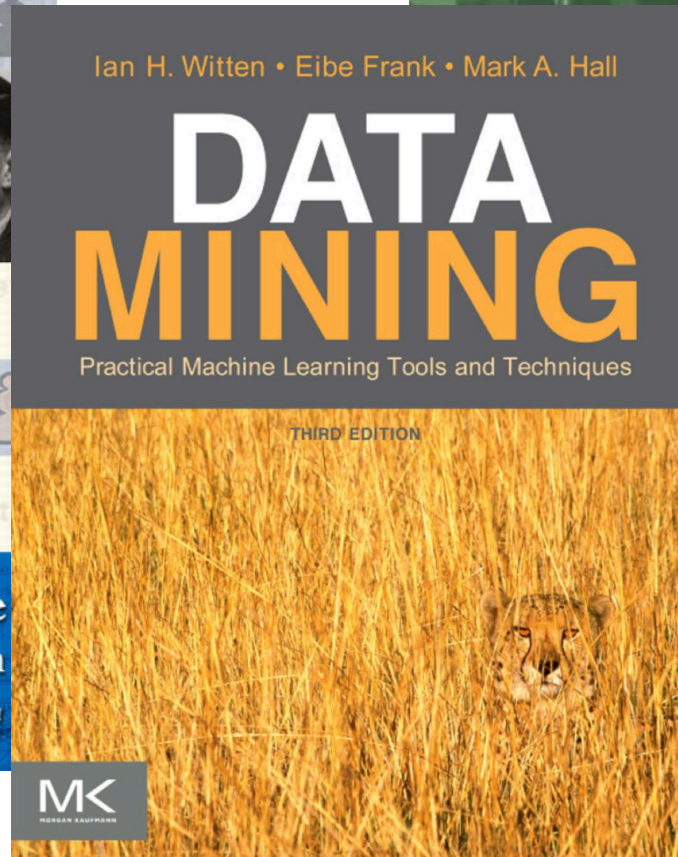
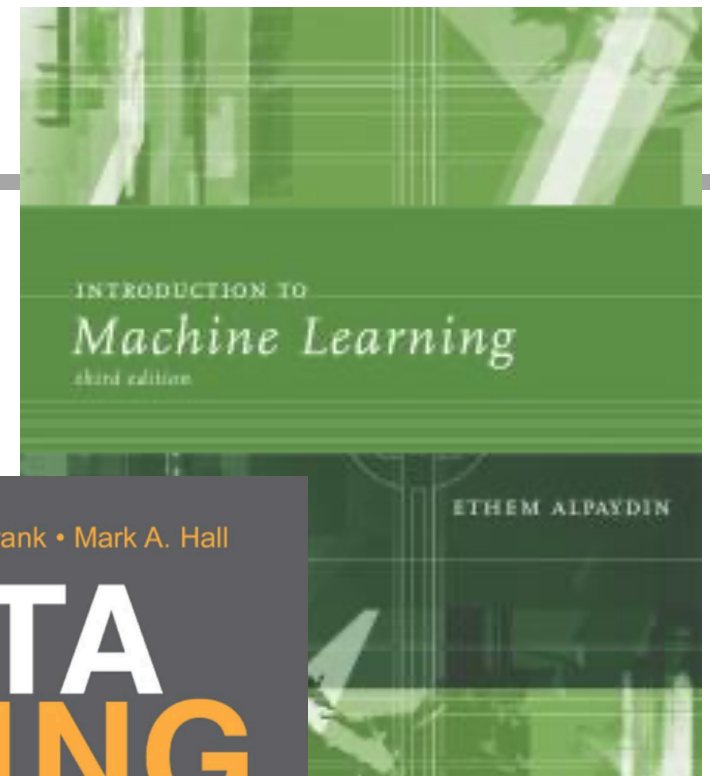
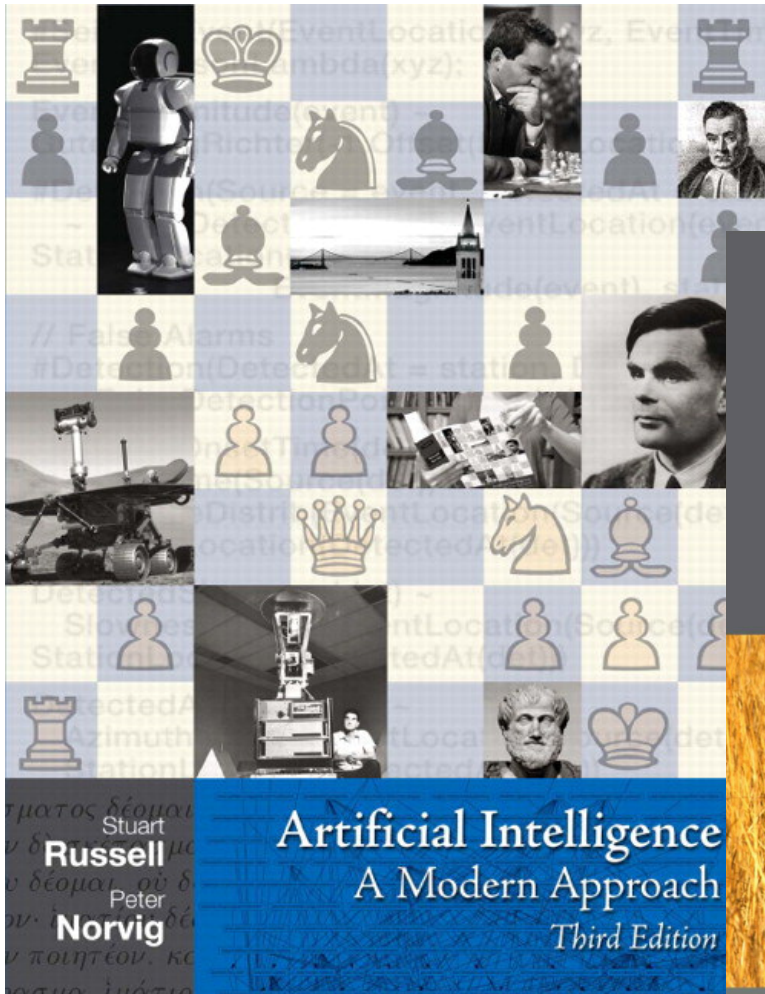
Anfragesprachen für Datenbanken

- Semantik und Ausdruckstärke
 - Was braucht man in der Anwendung?
- Ausdruckstärke 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)

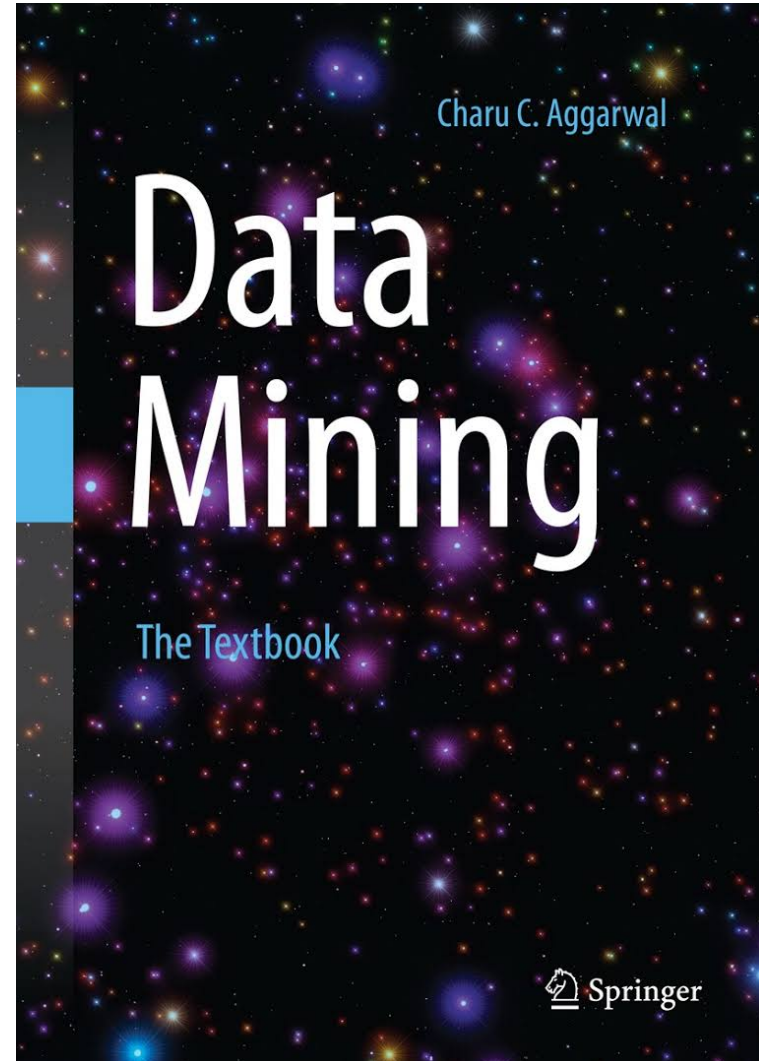
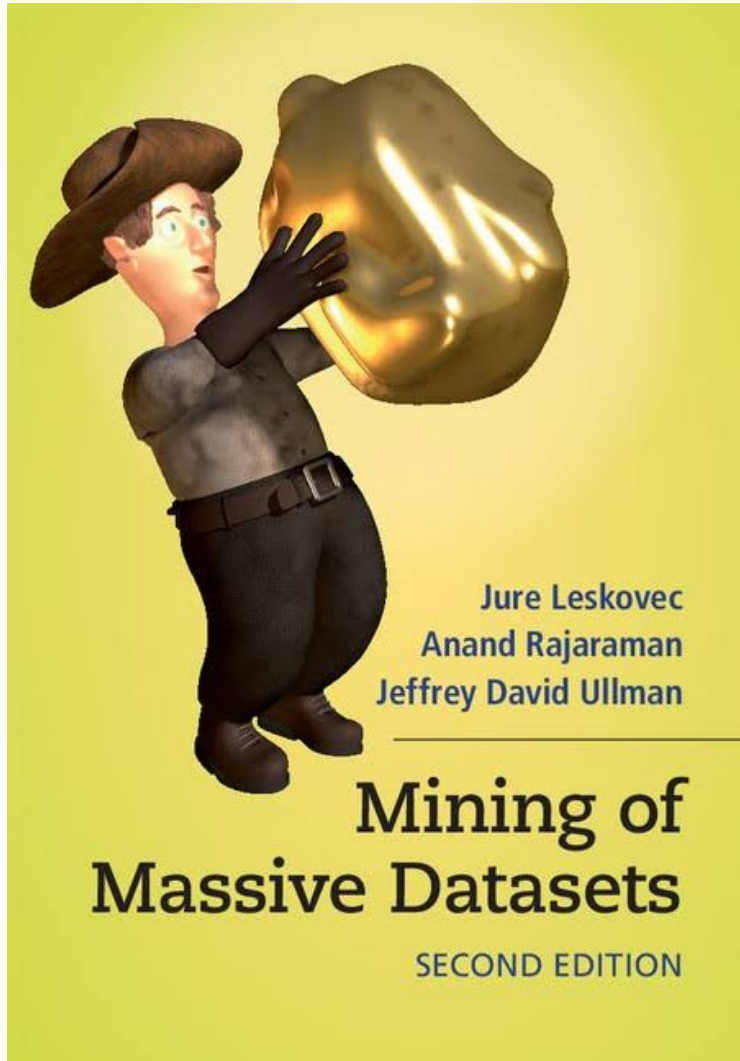


- Gewinnung von **Entscheidungsfunktionen** (Klassifikation) oder **Berechnungsfunktionen** (Regression) aus Daten
 - Statische Daten
 - Über der Zeit eintreffende Daten (Stream Data Mining)
 - Daten mit Zeitstempel oder auch Gültigkeitsinformation (Historical Data Mining)
- Extraktion von **relationalen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

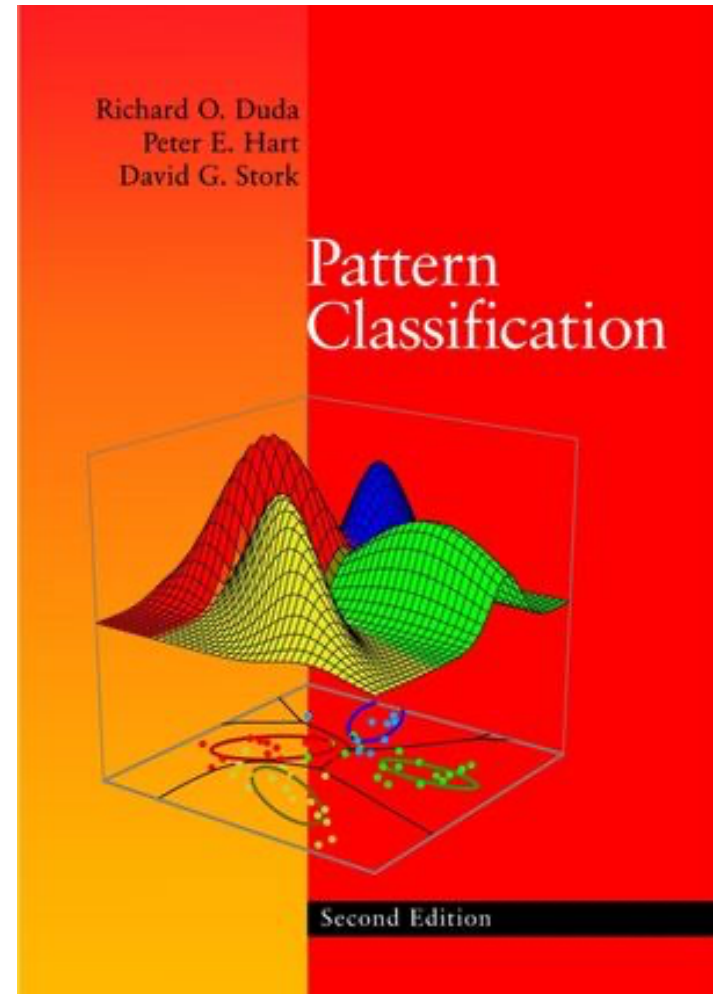
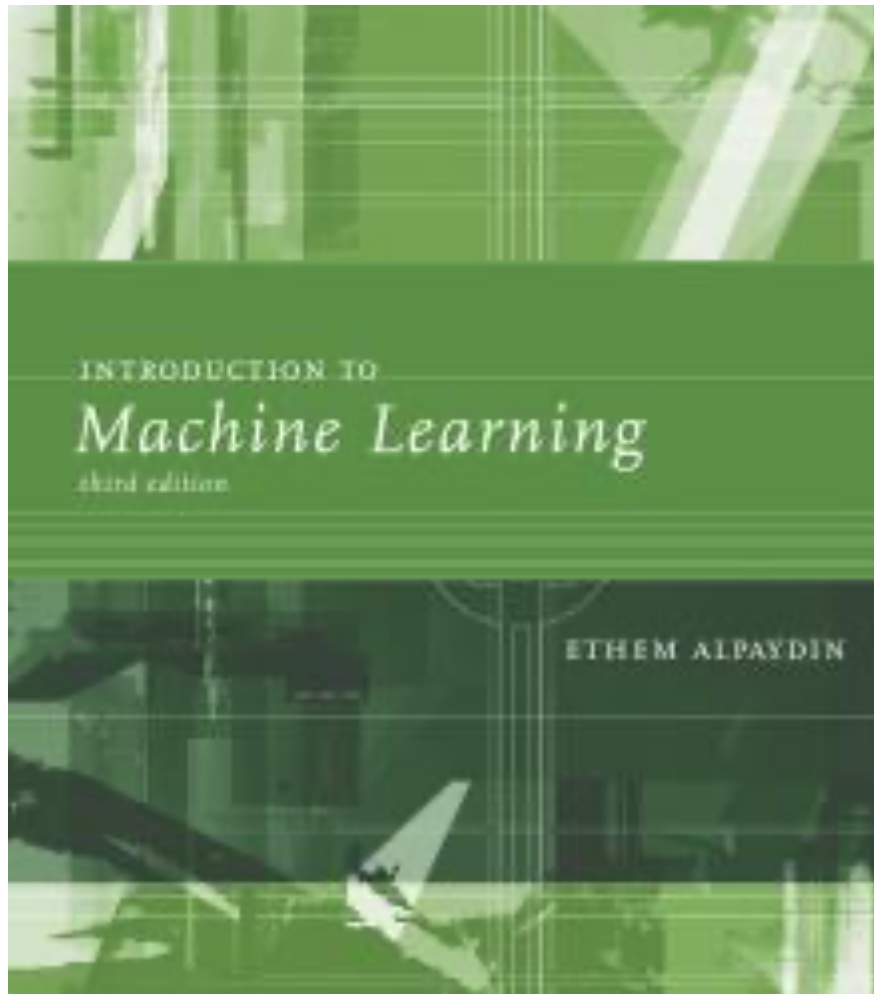
Literatur



Literatur



Literatur



Übersicht

- Semistrukturierte Datenbanken (am Beispiel von XML) und Volltextsuche
- Information Retrieval
- Mehrdimensionale Indexstrukturen
- Cluster-Bildung
- Einbettungstechniken
- First-n-, Top-k-, und Skyline-Anfragen
- Probabilistische Datenbanken, Anfragebeantwortung, Top-k-Anfragen und Open-World-Annahme
- Probabilistische Modellierung, Bayes-Netze, Anfragebeantwortungsalgorithmen, Lernverfahren,
- Temporale Datenbanken und das relationale Modell,
- Probabilistische Temporale Datenbanken
- SQL: neue Entwicklungen (z.B. JSON-Strukturen und Arrays), Zeitreihen (z.B. TimeScaleDB)
- Stromdatenbanken, Prinzipien der Fenster-orientierten inkrementellen Verarbeitung
- Approximationstechniken für Stromdatenverarbeitung, Stream-Mining
- Probabilistische raum-zeitliche Datenbanken und Stromdatenverarbeitungssysteme: Anfragen und Indexstrukturen, Raum-zeitliches Data Mining, Probabilistische Skylines
- Von NoSQL- zu NewSQL-Datenbanken, CAP- und CALM-Theorem, CRON-Hypothese
Blockchain-Datenbanken