Big Data Anwendungen

Einführung

Agenda

- Einführung
  - Organisatorisches
  - Big Data
  - Data Mining
  - Ausblick

- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

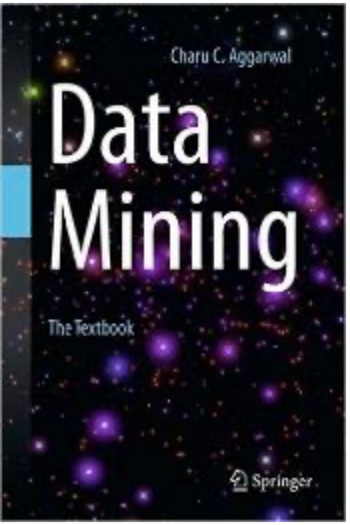
Über mich

- Organisatorisches
  - Vorlesung und Übung
  - Kontaktadresse: stephan.schosser@ovgu.de

- Mini CV
  - Studium: Wirtschaftsinformatik (Schwerpunkt Daten)
  - Promotion: Informatik (Schwerpunkt Daten / Verhalten)
  - Habilitation: Wirtschaftswissenschaft (Schwerpunkt Verhaltensökonomie)
  - Beruf
    Datenanalyst / Data Scientist bei deutschen Versendern und Beratern (Otto-Gruppe, Schwarz Gruppe, Klingel Gruppe, parsionate GmbH)

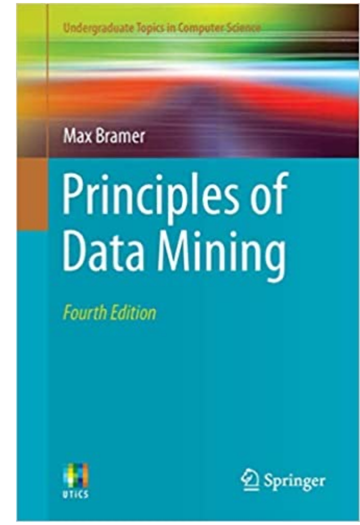
- Warum hier?
  - Habilitation erfordert 2 SWS pro Semester
  - Habe Spaß an der Lehre

Literatur



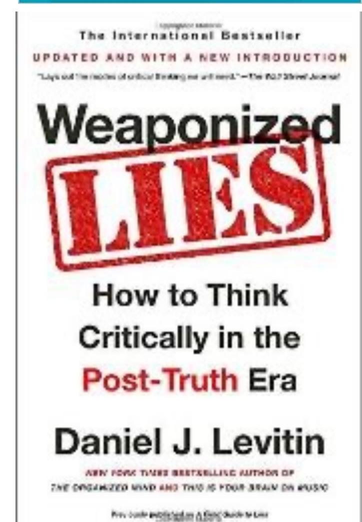
C. Aggarwal (2015):
Data Mining: The Textbook, Springer,
ISBN: 978-3319381169

M. Bramer (2020)
Principles of Data Mining
ISBN: 978-1447173069



R. Bachmann, G. Kemper, T. Gerzer (2014):
Big Data – Fluch oder Segen?,
ISBN: 978-3826696909

D. Levitin (2017):
Weaponized Lies,
ISBN: 978-1101983829



Veranstaltungstermine (Planung)

Vorlesungen			Übungen		
16.06.	10:00	Einführung			
16.06.	11:45	Deskriptive Methoden der Datenanalyse			
16.06.	13:30	Datenqualität			
16.06.	15:15	Klassifikation I			
17.06.	10:00	Klassifikation II			
17.06.	11:45	Recommender Systems	17.06.	13:30	Desktiptive Methoden
			17.06.	15:15	Klassifikation
18.06.	10:00	Clusteringverfahren I			
18.06.	11:45	Clusteringverfahren I	18.06.	13:30	Recommender Sys.
			18.06.	15:15	Clusteringverfahren
23.06.	10:00	Stream Mining			
23.06.	11:45	Social Network Analysis	23.06.	13:30	Stream Mining
23.06.	15:15	Technische Lösungen	24.06.	10:00	Social Network Anal.
24.06.	11:45	Datenschutz und Gesellschaft	24.06.	15:15	Praxisübung – KNIME (3h)
			25.06.	10:00	Praxisübung – Python (3h)
			25.06.	13:30	Bonusblätter 1 bis 3

Vorlesungsunterlagen

- Foliensatz zur Vorlesung
  - Online verfügbar
  - https://elearning.ovgu.de/course/view.php?id=9517

- Übungsinhalte
  - Aufgabenbeschreibung, Datensätze
  - http://goschosser.de/de/bda/2023/

- Zugangsdaten
  - Benutzername: bda
  - Passwort: BigData

Sonstiges

- Klausur
  - Durchführung
    - Lösen von Verständnisfragen
    - Anwenden von Algorithmen
  - Vorbereitung
    - Teilnahme an den Übungen, Vorlesungen
  - Termin
    - 29.07.2023

- Kommunikation
  - Fragen, Anmerkungen, Kritik an stephan.schosser@ovgu.de
  - Mindestanforderungen für Beantwortung von E-Mails
    - Kurz, prägnant formuliert
    - Keine Anfragen à la
      - Ich habe die Übung verpasst. Wie ist die Lösung für Aufgabe X?
      - Ich war das Semester über surfen. Was kommt in der Klausur?

Agenda

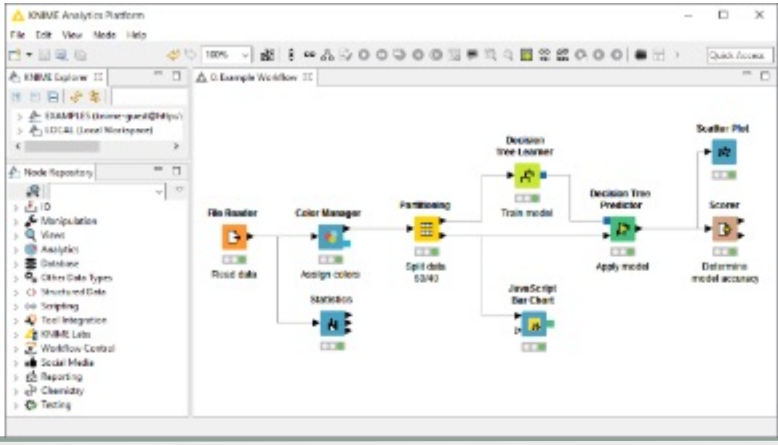
- Einführung
  - Organisatorisches
  - Big Data
  - Data Mining
  - Ausblick

- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Überblick über die Veranstaltung

- Inhalt der Veranstaltung
  - Methoden des Data Mining
    (Überblick über Verfahren zu Clustering, Klassifikation und Empfehlungen)
  - Anwendung mit Data Mining-Tools
    (Einführung in ein exemplarisches Tool - KNIME)
  - Überblick über Technologien zur Datenverarbeitung
    (Überblick über zu Grunde liegende Technologien)
  - Überblick über Big Data
    (Im Kern: Was sind die Innovationen der letzten 15 Jahre)

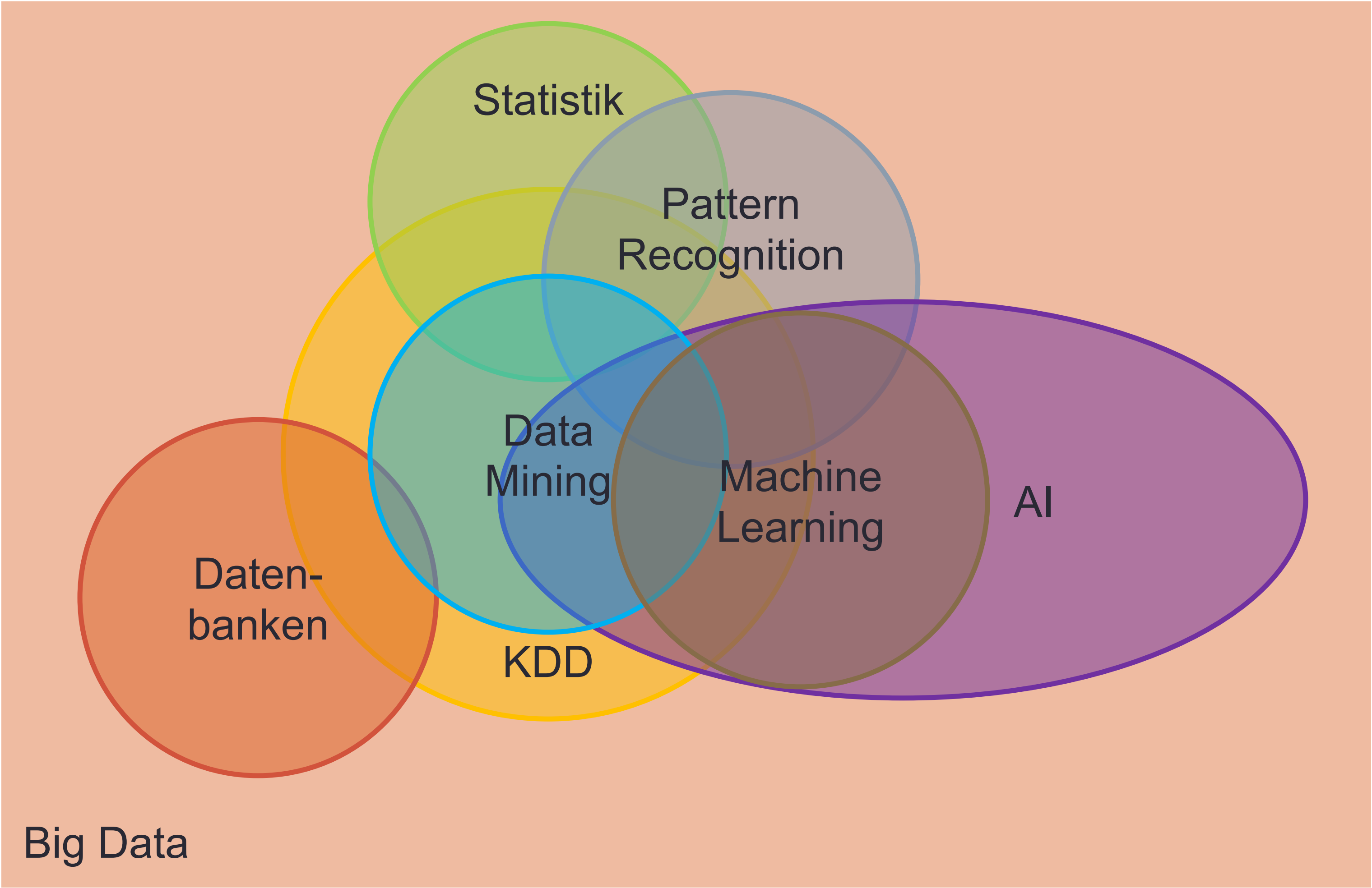
- Eingesetzte Software
  - KNIME Analytics Platform
    - Fokus auf Data Mining
    - Frei downloadbare Software
    - http://www.knime.org
  - Python
    - Programmiersprache
    - Erlaubt einfache Analysen



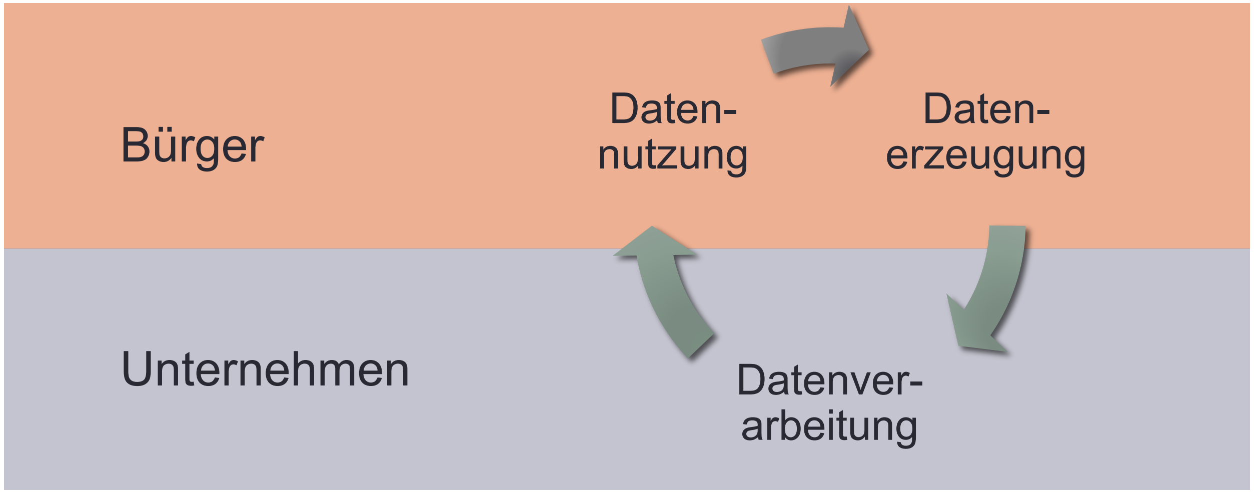
Zitate

- D. Patil (Chief Data Scientist, Greylock Partners, San Francisco)
„[Bei Big Data] geht es darum, im Datenwust Muster zu erkennen und richtig zu deuten. Wir selbst sind jetzt ein Datenprodukt.“
- Gartner IT Glossar (http://www.gartner.com/it-glossary/big-data)
Big Data sind sehr umfangreiche, stark volatile und stark unterschiedliche Daten, die kosteneffiziente, innovative Formen der Informationsverarbeitung benötigen und verbessertes Verständnis, verbesserte Entscheidungen und verbesserte Prozessautomation ermöglichen.
- D. Ariely (Duke University, Blog: http://danariely.com)
„Mit Big Data ist es wie mit Sex im Teenager-Alter. Jeder spricht darüber. Keiner weiß wirklich, wie es geht. Alle denken, dass die anderen es tun, also behauptet jeder, dass er es auch tut.“
- A. Merkel (Bundeskanzlerin, Aussage im Kontext von Big Data)
„Wir müssen auch eine gesellschaftliche Debatte darüber führen, dass Daten der Rohstoff der Zukunft sind und dass das [...] Prinzip der Datensparsamkeit nicht mehr zur heutigen Wertschöpfung passt.“

Einordnung


Statistik
Pattern Recognition
Data Mining
Machine Learning
AI
Daten-banken
KDD
Big Data

Primärkreislauf



- Datenerzeugung
  - Benutzung von Produkten, Services, Kommunikationsmedien
  - „Online sein“
- Datenverarbeitung
  - Schaffung neuer Produkte
  - Schaffung technischer Innovationen
- Datennutzung
  - Einsatz neuer Produkte und Innovationen liefert neue Daten
  - Datenkreislauf wird beschleunigt ⇒ Big Data

Datenerzeugung

- Internet
  E-Mails, Anfragen an Suchmaschinen, Verhalten auf Streamingplatformen, Kaufhistorien, Facebook, TikTok, Twitter, Chatbots
- Mobile Endgeräte
  Ortungsdaten von Smartphones, Fahrzeugdaten- und Sensoren, Bilder in der Cloud (inkl. Geodaten!), Smart Watches, VR Brillen, Spielwaren
- Stationäre Endgeräte
  Smart Meter, Amazon Echo, Thermostate, Waschmaschinen, Fernseher
- Bezahlmittel
  EC-Karten, Kreditkarten, Rabattkarten wie Payback, Lidl Pay
- Behördliche Daten
  Daten von Meldebehörden, Finanzämtern, Gesundheitskarte

Volumen Datenerzeugung

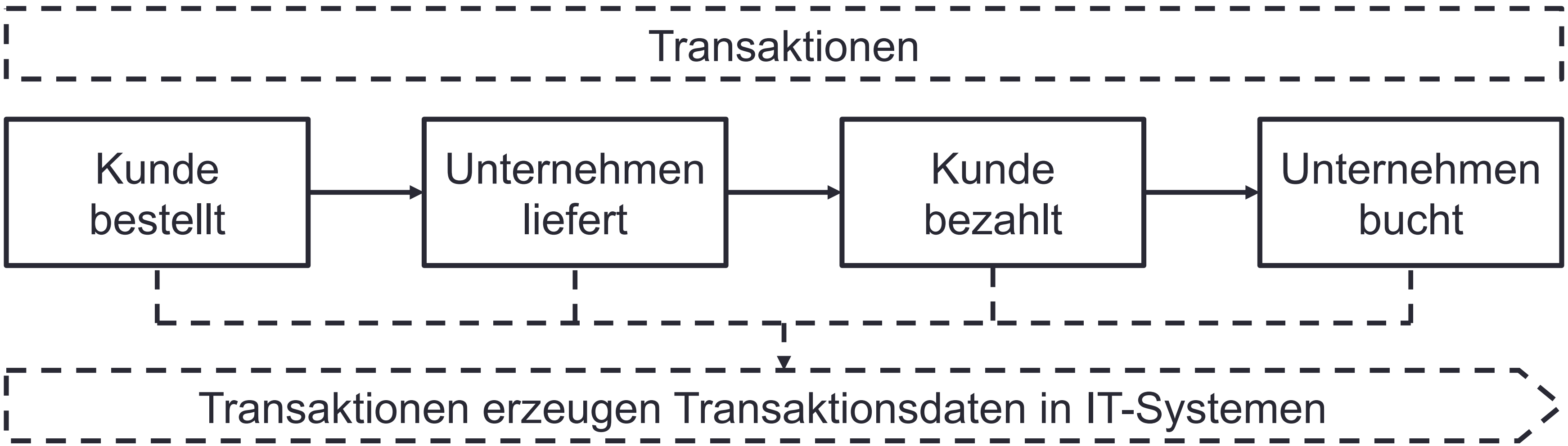
- Menge der Daten
  - Facebook sammelt täglich 4 Petabyte (2019)
  - Walmart sammelt täglich 40 Petabyte (2017)
  - Zahlen für Apple, Google, Amazon schwer zu finden...
    ... bei den aktuellen Datenschutzdiskussionen nicht überraschend
- Zum Vergleich
  - Menschliches Gehirn insgesamt: ca. 2.5 Petabyte
  - Traditioneller Versender vor Big Data: ca. 1 Terabyte
- Insgesamt (Quelle: Statista, 2022)
  - Alle 24 Monate: Verdopplung der weltweit verarbeiteten Datenmenge
  - 2020: 64 Zettabytes, Erwartung für 2025: 181 Zettabytes
- Unternehmen können nur noch überleben...
  ... wenn sie lernen Daten zu nutzen

Wort vor „-byte“	Giga	Tera	Peta	Exa	Zetta
Anzahl Nullen hinter 1	9	12	15	18	21

Geschwindigkeit bei Datenverarbeitung

- Typische Aufgaben vor Big Data
  - Churn Detection
    Erkennen von Kündigungswahrscheinlichkeit bspw. bei Telefontarifen
  - Kundensegmentierung
    Unterteilung der Kunden in Gruppen zur gezielteren Ansprache
  - Dispositionsplanung
    Vorhersage der Nachfrage zur Einkaufssteuerung
- Typische Aufgaben mit Big Data
  - Stauprognose
    Vorhersage der Fahrzeit, wenn Nutzer Büro Richtung Auto verlässt
  - Nachfrageprognose
    Vorhersage des Kundengeschmacks, wenn dieser Seite im Browser öffnet
  - Heizungssteuerung
    Reduktion der Heizleistung bei Öffnen des Fensters
- Unterschiede
  - Anforderung an Vorhersagen in „Echtzeit“ – ohne Überprüfung – steigt
  - Struktur der Daten wesentlich vielfältiger: „Event Store“

Transaktionsdaten



Stammdaten

- Name, Adresse, Geburtsdatum, ...
- Kaum Änderung über Zeit
- Sukzessive ergänzt (z.B. Kleidungsgrößen)

Bewegungsdaten

- Bestelldatum, bestellte Produkte, Menge, Farbe, Größe, Preis, ...
- Fallen bei jeder Transaktion an
- In Bezug zu konkretem Vorgang

Art der anfallenden Transaktionsdaten änderte sich in den letzten Jahren nicht!

Beobachtungs- und Interaktionsdaten

- „Traditionelle“ Vorhersagen
  - Vorschlag von Produkten, die „ähnliche“ Kunden kauften
  - Vorschlag ähnlicher Produkte zum letzten Kauf

- Neue Datenart – neben den „Tranksationsdaten“

- Typische „neue“ Daten
  - Aufenthaltsort zum Zeitpunkt der Interaktion
  - Bewegungsprofile
  - Kommunikationspartner und –medium
  - Gesundheitszustand
  - Medienkonsumprofil

- Mögliche Vorhersagen
  - Änderung des Arbeitgebers, Wohnorts (Umzugsunternehmen)
  - Änderung des Beziehungsstatus (Verkauf von Kleidung)
  - Prognose anstehender Krankheiten (Pharmaindustrie)
  - Geschmackspräferenzen (Verkauf von Modeartikeln)

Agenda

- Einführung
  - Organisatorisches
  - Big Data
  - Data Mining
  - Ausblick

- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Von Daten zur Entscheidung (Gianotti und Pedreschi)

Daten
- Kundendaten
- Daten aus den Filialen
- Demographische Daten
- Geographische Daten

Inspektion

Information
- X lebt in Z
- S ist Y Jahre alt
- X und S sind umgezogen
- W hat Geld in Z

Aggregation (zielgerichtet)

Wissen
- Anzahl Y des Produkts A wird in Gebiet Z verwendet
- Kunden der Klasse Y verwenden x% von C in Zeitraum D

Interpretation

Entscheidung
- Sonderangebot für Produkt A in Gebiet Z
- Mailings an Kunden mit Profil P
- Cross-Selling von Produkt an Kunden C

Data Mining Verfahren I

- Klassifikation
  - Jede Beobachtung ist über n Variablen $(x_1, x_2, \ldots, x_n)$ beschrieben
  - Klassifikation schließt aus diesen Variablen eine Klasse $c \in \{1, \ldots, C\}$
  - Typische Beispiele
    - Kreditwürdigkeitsprüfung
    - Churn detection

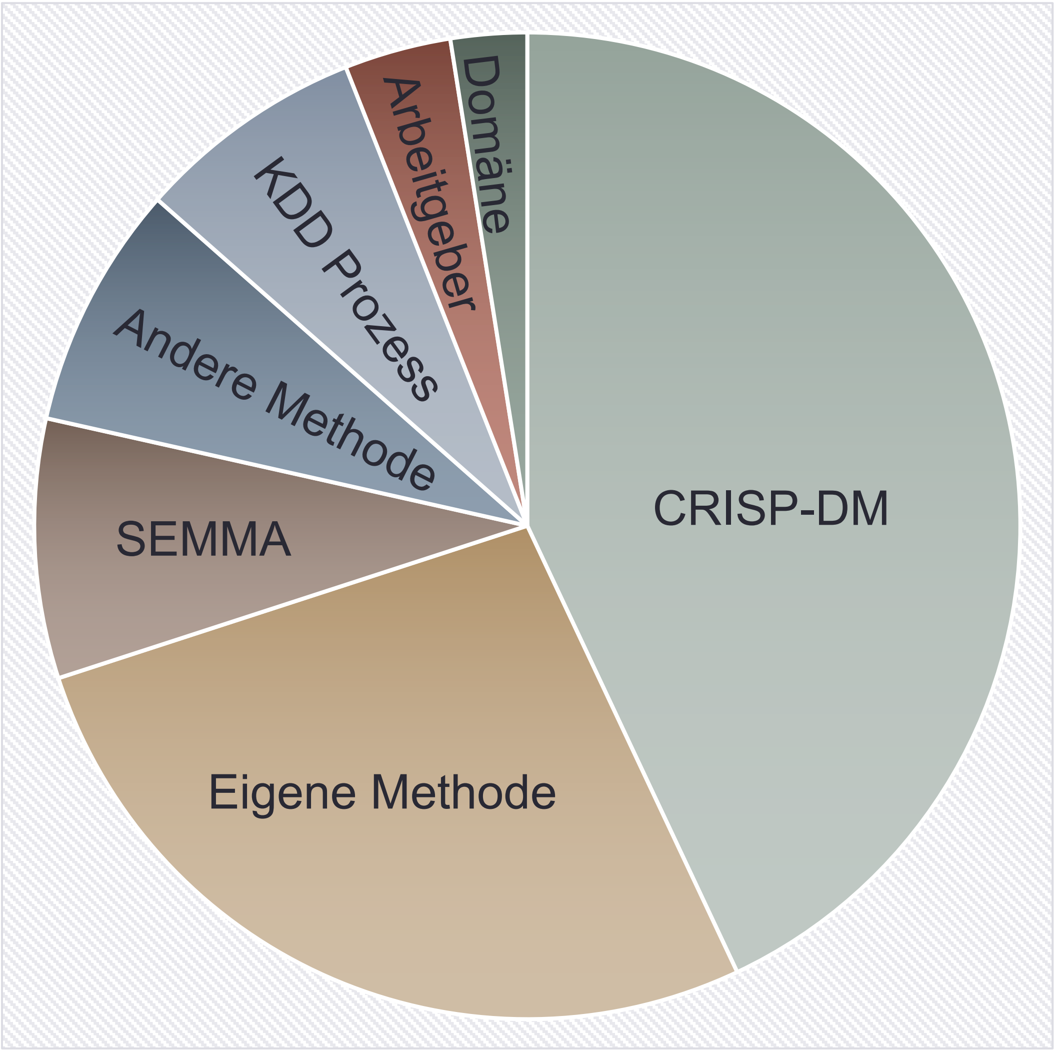
- Clustering
  - Jede Beobachtung ist über n Variablen $(x_1, x_2, \ldots, x_n)$ beschrieben
  - Clustering identifiziert vorher unbekannte “Cluster“, d.h. Gruppen
    - ... mit ähnlichen Beobachtungen
    - ... die sich untereinander möglichst stark unterscheiden
  - Typische Beispiele
    - Segmentierung von Kunden auf Basis Verhaltensmustern
    - Identifikation von Warensortimenten

Data Mining Verfahren II

- Recommender Systeme
  - Für jede Beobachtung ist eine Menge von konsumierten Artikeln bekannt
  - Recommender Systeme schließen, welche Artikel auch konsumiert wird
  - Typische Beispiele
    - Verbesserung der Layout-Planung von Supermärkten
    - Cross Selling von Produkten im Call Center
    - Generierung von Playlists bei Spotify oder Netflix

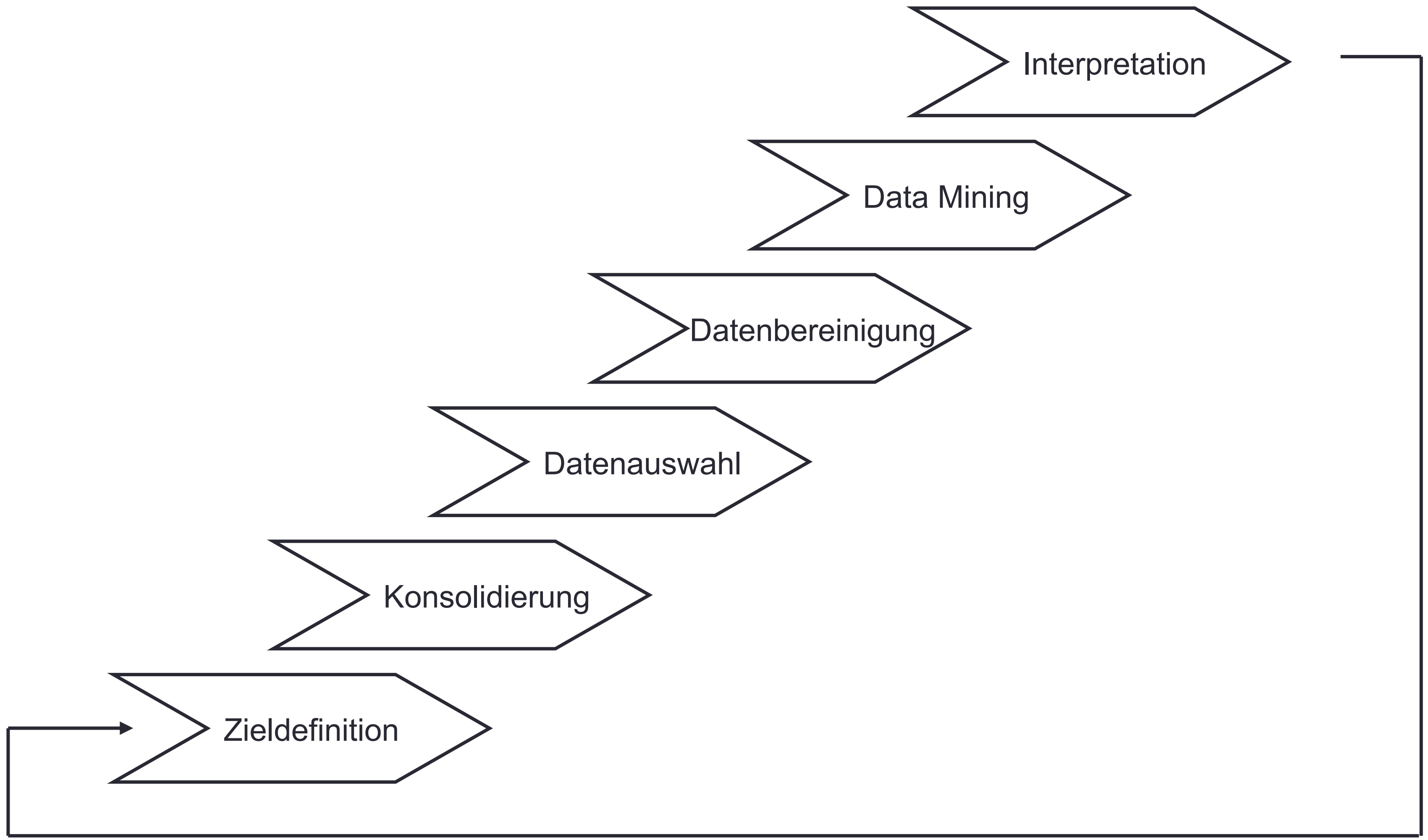
- Regressionsanalysen (hier nicht vorgestellt)
  - Jede Beobachtung ist über n Variablen $(x_1, x_2, \dots, x_n)$ beschrieben
  - Vorhersage eines numerischen oder binären Attributs
  - Typische Beispiele
    - Vorhersagen von Dispositionsmengen
    - Vorhersage von Tagestemperatur

Data Science Methoden



- Beschreiben „Vorgehen“ von den Daten zum Wissen
- In der Regel basieren „Data Mining“ Tools auf einer Methode
- Verfahren unterscheiden sich geringfügig
- Hier Überblick über SEMMA, CRISP-DM und KDD Prozess

Knowledge Discovery in Databases I

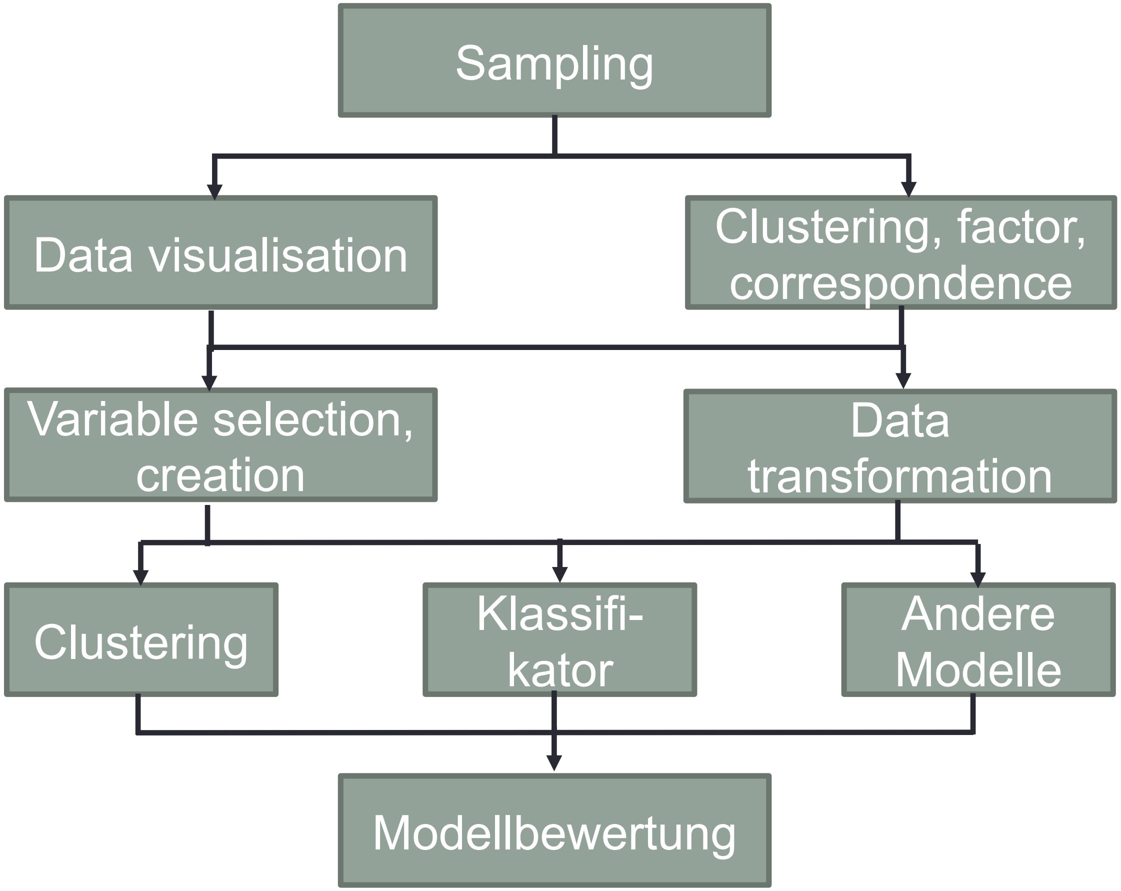


Knowledge Discovery in Databases II

- Zieldefinition
  In welchem Bereich des Unternehmens sollen...
  ... aufgrund erkannter Probleme...
  ... neue Erkenntnisse gewonnen werden
- Konsolidierung und Auswahl der Daten
  - Vorauswahl geeigneter Daten und Variablen und Wahl einer Stichprobe
  - Beurteilung der Datenqualität
    (Fehlende Werte, Ausreißer, Erfassungsfehler)
  - Zusammenführen der Daten
- Datenbereinigung
  - Bilden von Kennzahlen (Verhältniszahlen, Änderungsraten, ...)
  - Behandlung von Ausreißern und fehlenden Werten
- Data Mining
  Einsatz von Data Mining-Verfahren
- Interpretation
  - Überprüfung der gewonnenen Aussagen auf Plausibilität
  - Gegebenenfalls Hypothesentest zur Evaluierung der Aussagen
  - Normalerweise Definition neuer Ziele

SEMMA I

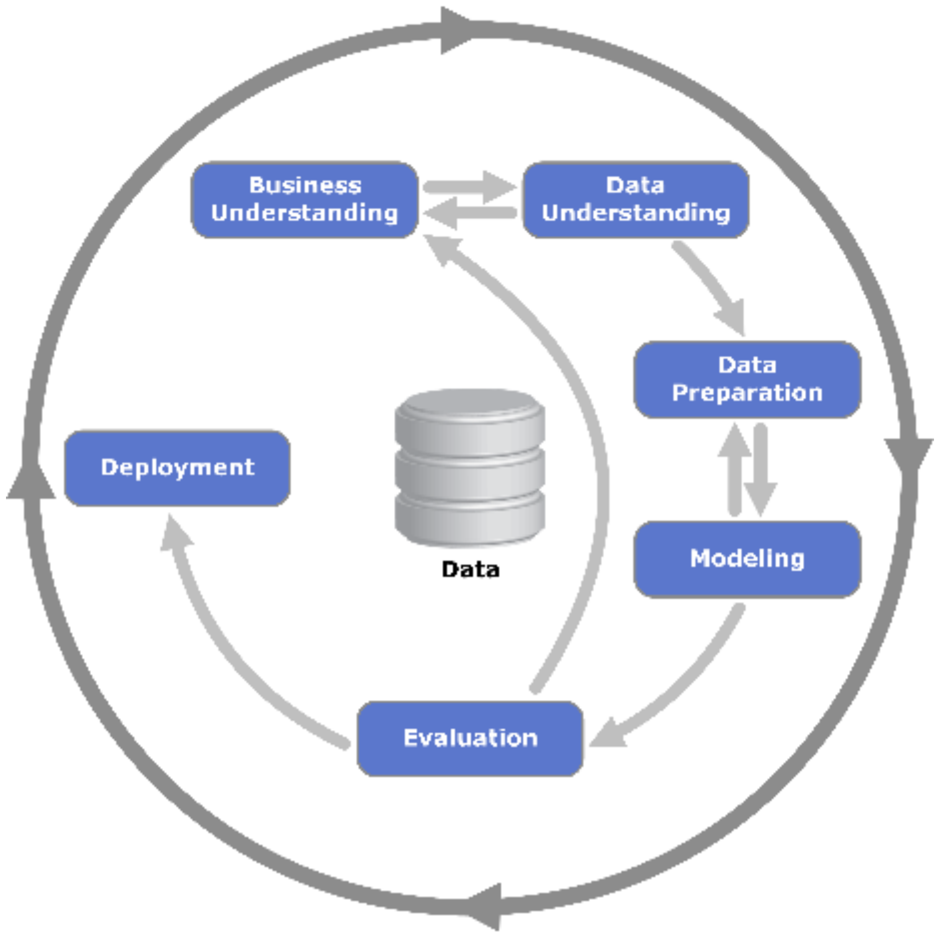
- Sample
- Explore
- Modify
- Model
- Assess



SEMMA II

- Sample (Stichprobe)
  - Identifizierung des Input-Datensatzes
  - Ziehung von Stichproben (Sampling)
  - Aufteilung des Datensatzes in Trainings-, Validierungs- und Testdaten
- Explore (Erforschen / Untersuchen)
  - Statistische und grafische Untersuchung des Datensatzes (Datenvisualisierung, deskriptive Statistiken, Zusammenhangsanalyse)
- Modify (Aufbereiten)
  - Datenaufbereitung für Analysen (Erzeugung neuer Variablen, Ausreißer-identifikation, Ersetzen von fehlenden Werten)
- Model (Modellierung)
  - Bestimmen / Schätzen eines Vorhersagemodells (d.h. Anwendung eines Data Mining Verfahrens)
- Assess (Beurteilung)
  - Vergleich konkurrierender Vorhersagemodelle
  - Anwendung des Modells auf „neue“ Daten

CRISP-DM I


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM II

- Business Understanding (Verständnis des Anwendungsgebiets)
  Verständnis der Projektziele und Anforderungen der Anwender
- Data Understanding (Datenverständnis)
  - Datensammlung
  - Erkennen von Datenqualitätsproblemen
  - Entdecken von Hypothesen bzgl. Versteckten Wissen
- Data Preparation (Datenvorbereitung)
  Transformation der Rohdaten zum finalen Datensatz
  (Übertragen in Tabellen, Bereinigung der Daten)
- Modeling (Modellierung)
  Anwendung von Datenmodellen (vgl. Data Mining)
- Evaluation (Evaluierung)
  - Prüfen ob Modell Anwendungsziele erfüllt
  - Entscheidung über Deployment
- Deployment (Umsetzung; oft nicht in der Hand des Data Scientists)
  - Erstellung von Berichten...
  - ... Implementierung des Ergebnisses in Produktivumgebung

Abschließende Bemerkungen

- Vergleich der Ansätze
  - Anwendung in Data Mining Tools
    - SEMMA: SAS Enterprise Miner (und primär in diesem Umfeld genutzt)
    - CRISP-DM: SPSS Modeler (hat aber Ziel allgemeingültig zu sein)
    - KDD Prozess: KNIME, RapidMiner, Weka
  - Bedeutung der „Anwendungsseite“
    - Anwendungsseite bei SEMMA nicht berücksichtigt
    - Bei KDD Prozess und CRISP-DM ist Anwendungsseite zentral
  - Berücksichtigung von Deployment
    - Nur in CRISP-DM enthalten
    - Insbesondere im Bereich Big Data zentral

- Zukunft der Ansätze
  - KDD Prozess ist primär „akademischer“ Ansatz
  - SAS sieht SEMMA lediglich als „Struktur“ für den SAS Enterprise Miner
  - CRISP-DM soll seit 2006 überarbeitet werden – bisher ohne Ergebnis
  - ASUM-DM wurde 2015 von IBM als Nachfolger von CRISP-DM vorgestellt (hat nie Relevanz am Markt erreicht)

Agenda

- Einführung
  - Organisatorisches
  - Big Data
  - Data Mining
  - Ausblick

- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Ausblick

- Zusammenfassung
  - Datenanalyse (egal ob Big Data oder Data Mining) folgt...
    ... ähnliche Prozess (Datenauswahl, deskriptive Statistik, Modellbildung)
  - Big Data stützt sich auf eine Menge verschiedener Verfahren...
    ... diese Verfahren werden (großteils) für Data Mining genutzt
  - Data Mining Verfahren in hochgradig standardisierter Software genutzt
  - Big Data bezüglich moderner Technologien (z.B. bzgl. Performance)...
    ... deutliche Erweiterung gegenüber Data Mining Verfahren

- Lernziel
  - Vorteile von Big Data verstehen und bewerten (!) können
  - Dabei zentral: „Gefühl für Auswertungsverfahren“ (Übung!)
  - Dabei weniger zentral: Kenntnis der technischen Raffinessen im Detail

- Inhalt hier
  - Fokus im praktischen Teil auf Data Mining / Statistik
  - Ergänzung des praktischen Teils um Überblick über technische Lösungen