Arne Scheffer, AG3.2+3, Zentrum für Informationsverarbeitung

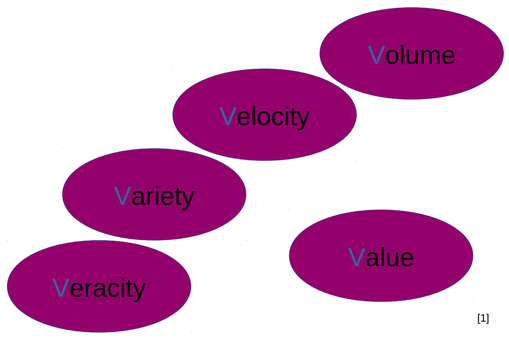
Data Science @ eScience WWU

Einleitung

Die Art, in welchem Umfang und in welcher Form Daten zur Verfügung stehen, hat sich mit flächedeckender Social-Media-Nutzung, extensiven Logdaten und sogenannten Mehrwertdiensten stark verändert.
Allen ist gemein, dass die Daten der Kunden den Hauptwertbestand darstellen.
Exemplarisch sind das nebst bekannten Social-Media-Vertretern Payback-Kartensysteme, Health-Tracker, GSM und GPS-Tracker in Autos und Handys, Internet-of-Things-Daten und Tracking-Systeme für das Klickverhalten der Nutzer im Web.
Aufgabe des Data-Science ist zum einen die infrastrukturelle Herausforderung, mit den zumeist umfangreichen und z.T. als Datenstrom eintreffenden Daten umzugehen.
Zum anderen gilt es jedoch, Wert aus diesen Daten in Form von Zusammenhängen zu generieren. Dies ist die eigentliche Kernaufgabe.
Die anfallenden Daten sind dabei häufig unstrukturiert,und die zu findenden Zusammenhänge befinden sich z.T im Bereich bisher unbeachteter Fragestellungen, für die es keine vorgedachten Lösungen gibt.
Als Data-Mining-Methoden kommen daher u.a. solche des unüberwachten, aber auch des überwachten machinellen Lernens zum Einsatz.
Dies geschieht z.B. in Form schwacher KI auf Basis einer Deep-Learning-Architektur. Typische statistische Themenfelder hierbei sind die Regression, die Kategorisierung, Cluster-Bildung/Ähnlichkeiten und Überlebensmodelle.

Big Data

- Die 5V: Größe, Geschwindigkeit (Datenstrom), Heterogenität, Gültigkeit, Wert



[1]

- Umgang mit Volume und Velocity

MapReduce
Object storage API
Apache Hadoop
Apache Spark
Massive parallel processing

- Umgang mit unstrukturierten Daten – Beispiel Text Mining:

Sentiment-Analyse
Topic modeling
Bag of Words
Stemming
Lemmatisierung
Stop words

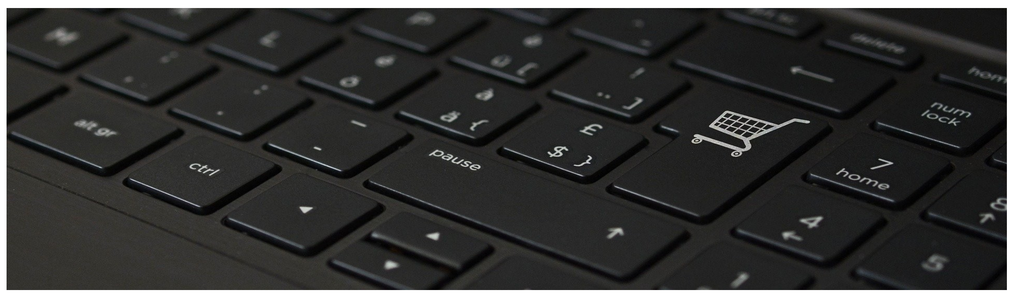
Ansprechpartner

Für Data Science:
Arne Scheffer
- Einsteinstraße 60
- Raum 104

Daten heute

- Social media und sogenannte Mehrwertdienste
- Persönliche (Profil-)Daten als Haupt-Wertgegenstand

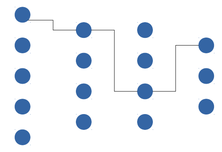
Payback-Karten
Zahlungsverhalten
Health-Tracker
Klickverhalten
Social media
Kaufverhalten
GPS-/GSM-Daten
IoT-Logdaten



Methoden

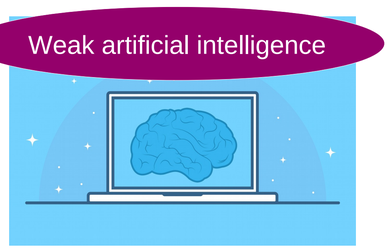
- Aus dem Data Mining

Kategorisierung
Regression
Clustering, Ähnlichkeiten
Ereigniszeitanalyse



- Unüberwachtes und überwachtes maschinelles Lernen

k-means/-median
Entscheidungsbaum
Random forest
Support-vector machine
Neuronales Netz



Quellenangaben:

[1] Textverweise zu den 5V:
- ursprünglich 3V: vgl. „3D Data Management: Controlling Data Volume, Velocity, and Variety“
  - Doug Laney, 06.02.2001
- Ergänzt um Veracity: vgl. „Harness the Power of Big Data The IBM Big Data Platform“, S. 14ff
  - Paul Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, David Corrigan, 08.11.2012
- Ergänzt um Value: vgl. „Transforming Big Data into Smart Data: Deriving value via harnessing Volume, Variety, and Velocity using semantic techniques and technologies“
  - Amit Sheth, 2014 IEEE 30th International Conference on Data Engineering

Bilder Pixabay:
- matrix-1735640:Jonny Lindner, matrix-4646234:Gerd Altmann, analytics-3088958:xresch, cloud-3805852:Gerd Altmann, watch-4638673:George Dyulgerow, board-3700116:athree23, ecommerce-3546296:Pete Linforth, work-731198:Free-Photos, artificial-4082314:Mohammed Hassan