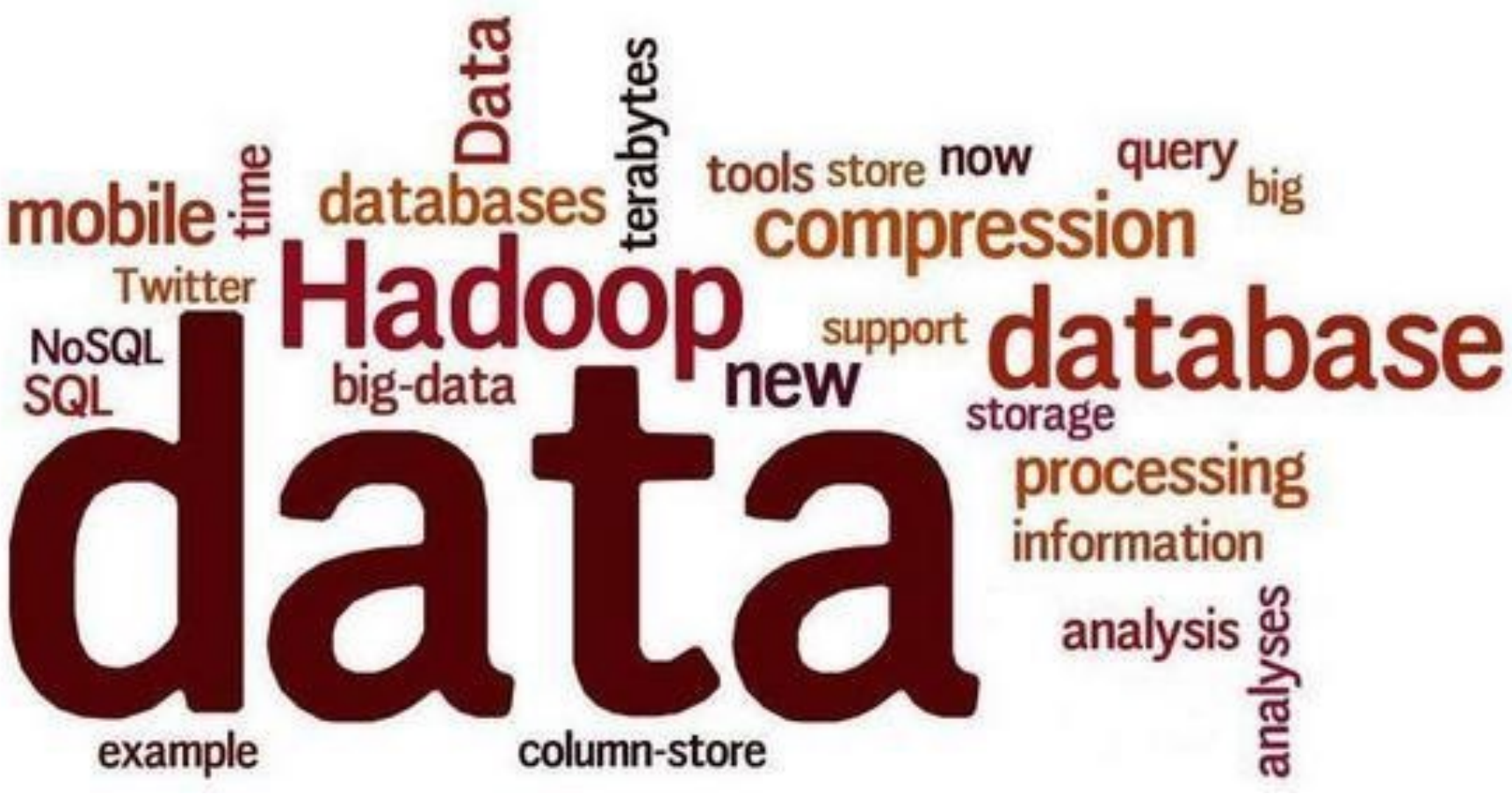
DHBW Mannheim

Big Data

Lecture 1

Frank Schulz


Data
terabytes
mobile
time
databases
tools
store
now
query
big
compression
Twitter
Hadoop
support
database
NoSQL
SQL
big-data
new
storage
data
processing
information
analysis
analyses
example
column-store

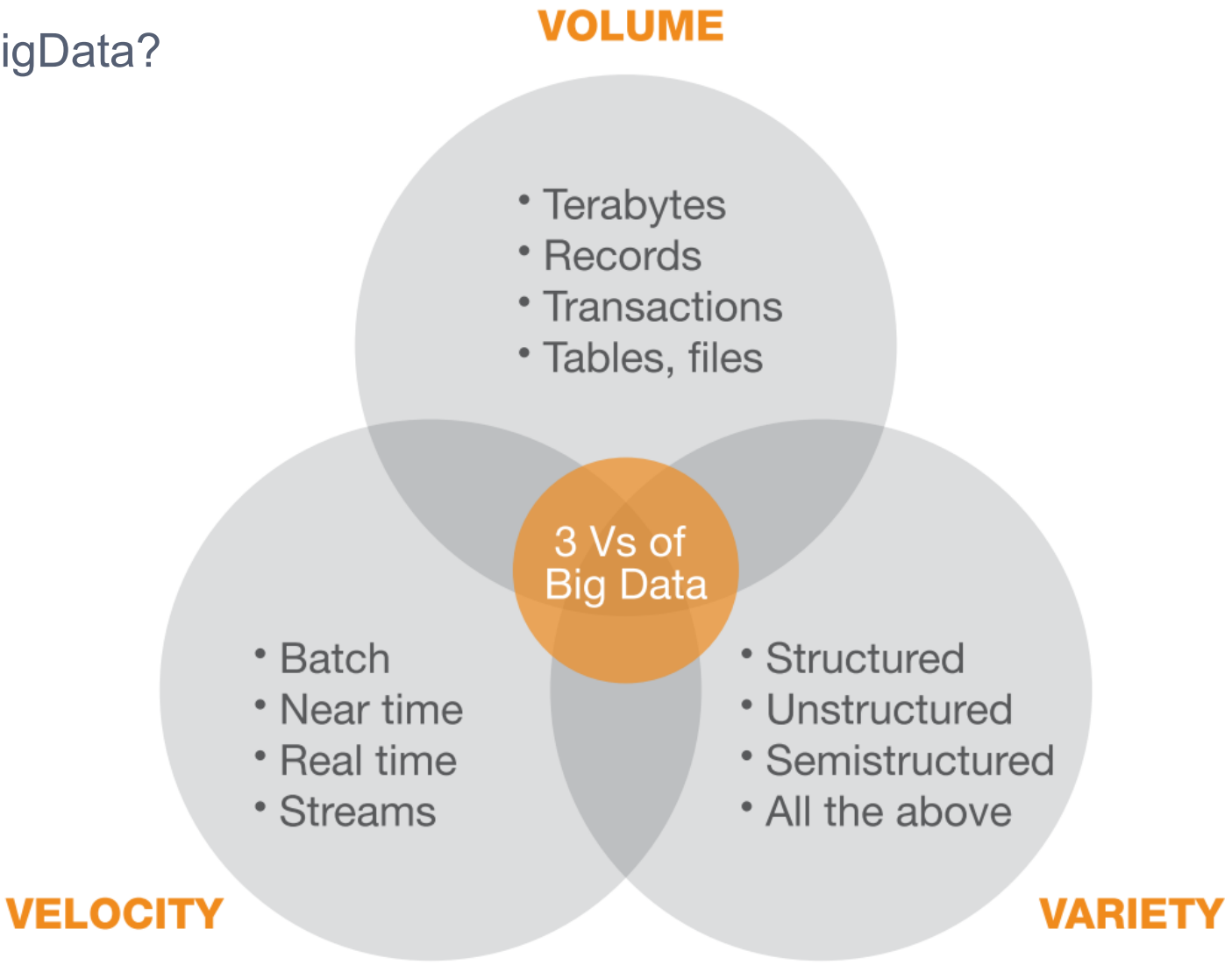
Was ist BigData?

"the size of the data itself becomes part of the problem" (Mike Loukides)

"Man merkt es, sobald man ein BigData Problem hat."

"data that becomes large enough that it cannot be processed using conventional methods" (Edd Dumbill)

Was ist BigData?



Quelle: Philip Russom, TDWI Report "Big Data Analytics"

3V

Volume

- Große Datenmenge
- Viele Datenquellen

Variety

- Strukturierte Daten: relationale Tabellen, CSV (comma separated values)
  einfache Struktur
- Semi-strukturierte Daten: JSON
  mittelkomplexe Struktur
- Unstrukturierte Daten: Text, Bilder, Audio, Video
  komplexe Struktur

Velocity

- Schnell eintreffende Daten
- Batch, Near Time, Real Time, Data Streams

Von 3V zu 4V zu 5V

Volume

- Verteilung der Datenmenge auf mehrere Rechner notwendig

Variety

- Verschiedene Datenformate erschweren die Auswertung und Integration

Velocity

- Schnell eintreffende oder sich rasch ändernde Daten

Veracity („Aufrichtigkeit, Wahrhaftigkeit“)

- Interpretation von Rohdaten nicht klar bzw. nicht festgelegt, Fehler in Daten

Value

- Wert der Daten (monetär, wissenschaftlich, …)

Skalierbarkeit

Skalierbarkeit

- Fähigkeit eines Systems, eine wachsende Last durch hinzugefügte Ressourcen aufzufangen
- Ideal: lineare Skalierbarkeit, d.h. mit doppelten Ressourcen kann man doppelt so große Probleme lösen

Beispiele

- 2011: Map Reduce Cluster sortiert 1 Petabyte (10^{15} = 10^{13} Zufallsstrings von je 100 Bytes) in 33 Minuten auf einem Cluster aus 8000 Rechnern (Google)
- 2013: Hadoop Cluster mit 2100 Rechnern sortiert 100 TB in 72 Minuten (Yahoo)
- 2014: Spark Cluster mit 207 Rechnern (AWS EC2) sortiert 100 TB in 23 Minuten (Databricks)

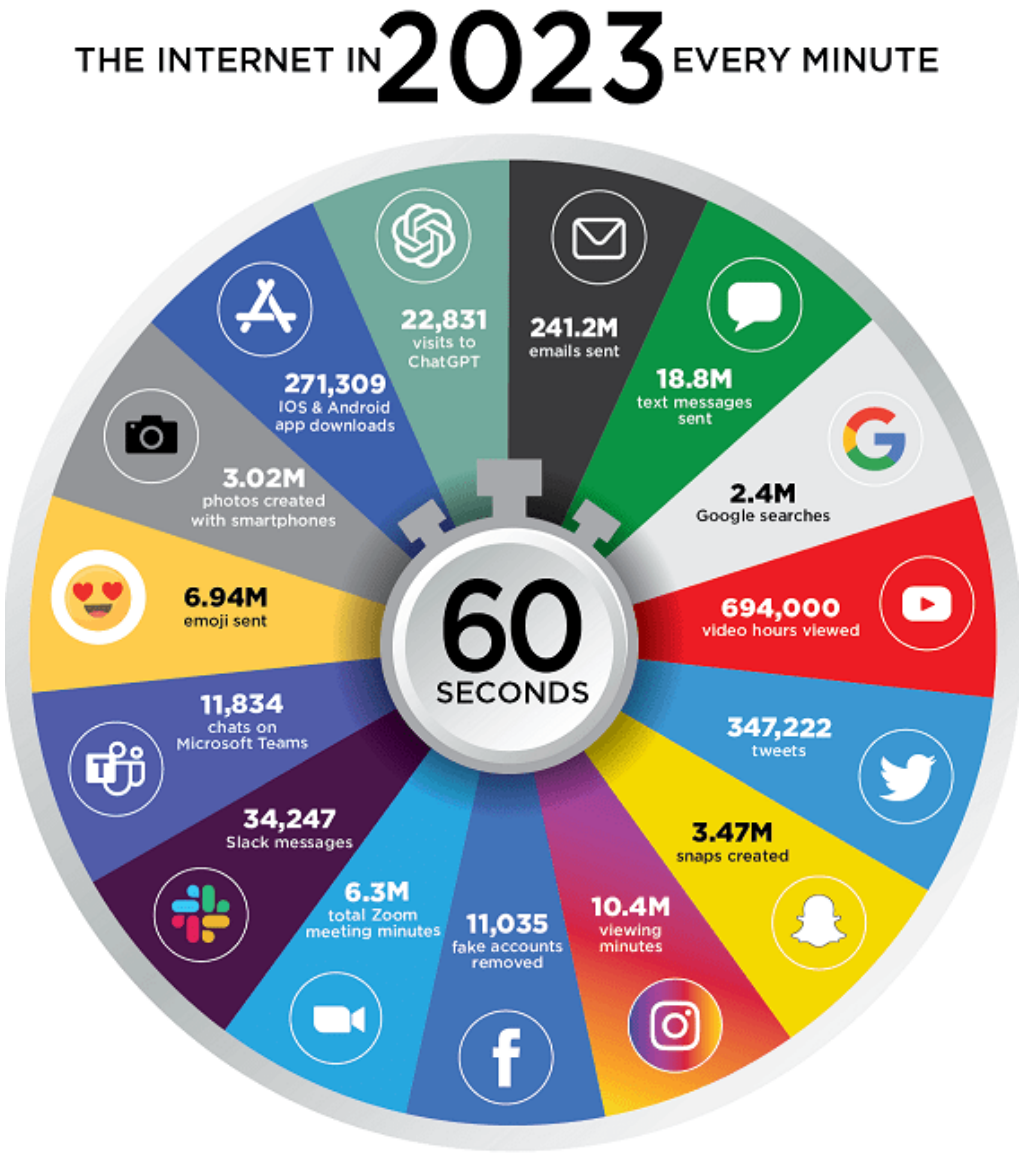
Datenvolumen

	Einheit	Faktor	Name (deutsch)	Name (englisch)
B	Byte	**1**	Eins	One
KB	Kilobyte	**10^3**	Tausend	Thousand
MB	Megabyte	**10^6**	Million	Million
GB	Gigabyte	**10^9**	Milliarde	Billion
TB	Terabyte	**10^{12}**	Billion	Trillion
PB	Petabyte	**10^{15}**	Billarde	Quadrillion
EB	Exabyte	**10^{18}**	Trillion	Quintillion
ZB	Zettabyte	**10^{21}**	Trilliarde	Sextilion

Anwendungen

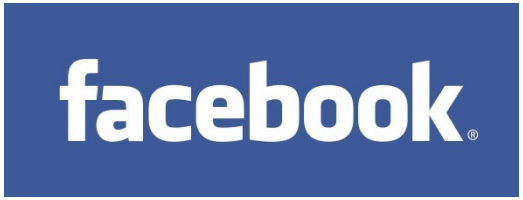
Datenquellen

- Webseiten und Social Media
  - Web Seiten, Tweets, Blog Posts
- Business Data
  - ERP Daten, Shopping Basket Analysis, Customer Segmentation
- Sensordaten (Internet of Things)
  - Smart Factory, Smart Home, Smart Mobility
- Öffentliche Daten
  - Statistiken von Verwaltungen und Regierungen
- Biomedizinische Daten
  - Personalisierte Medizin, "Precision Medicine"



Quelle: https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

Facebook



Datenvolumen

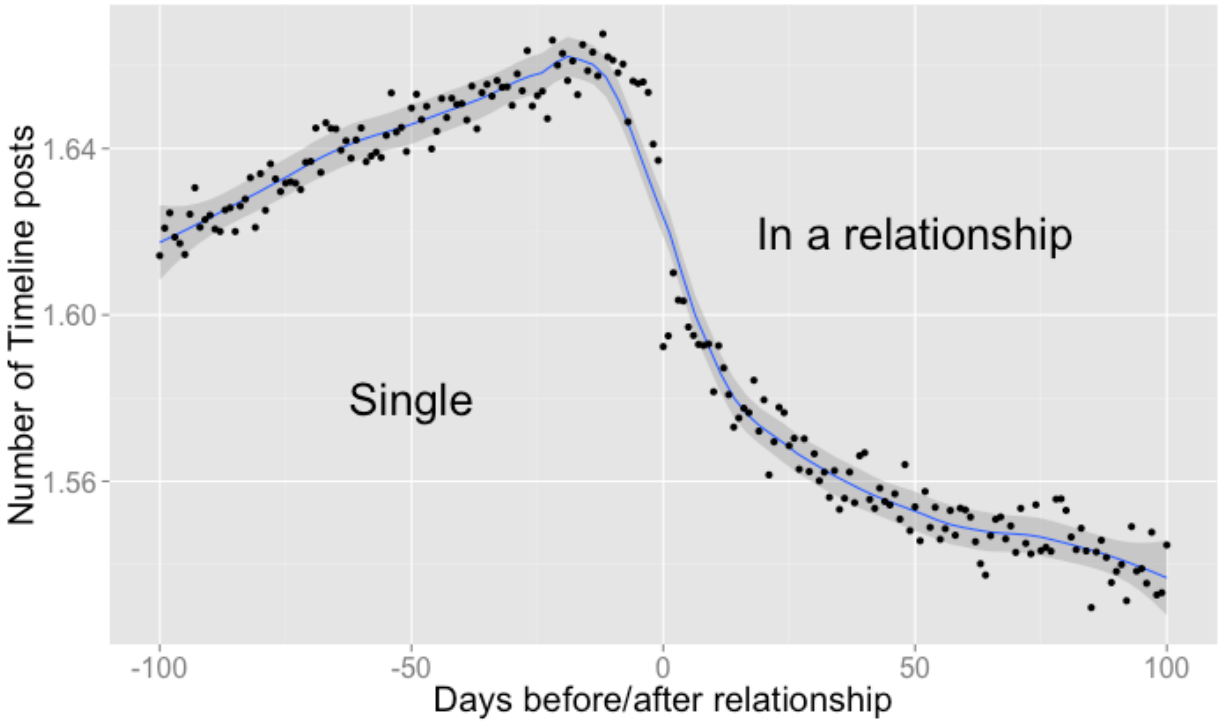
- 2014 hatte Facebook im Data Warehouse 300 PB (300.000 TB) gespeichert, mit einer Wachstumsrate von 600 TB/Tag

Datenstrukturen

- Facebook Social Graph
  - Definiert die Vernetzung von Usern,
    muss sehr schnell gelesen werden können (random access read)
- Facebook Messages
  - Muss schnell gelesen und geschrieben werden können,
    Wachstum: 6 TB / Tag
- Facebook Photo Store
  - Muss effizient gespeichert werden, schnelles Wachstum

Quelle: Dhruba Borthakur, „Petabyte scale data at Facebook“ (2012)
“Scaling the Facebook data warehouse to 300 PB“ (2014)

Facebook

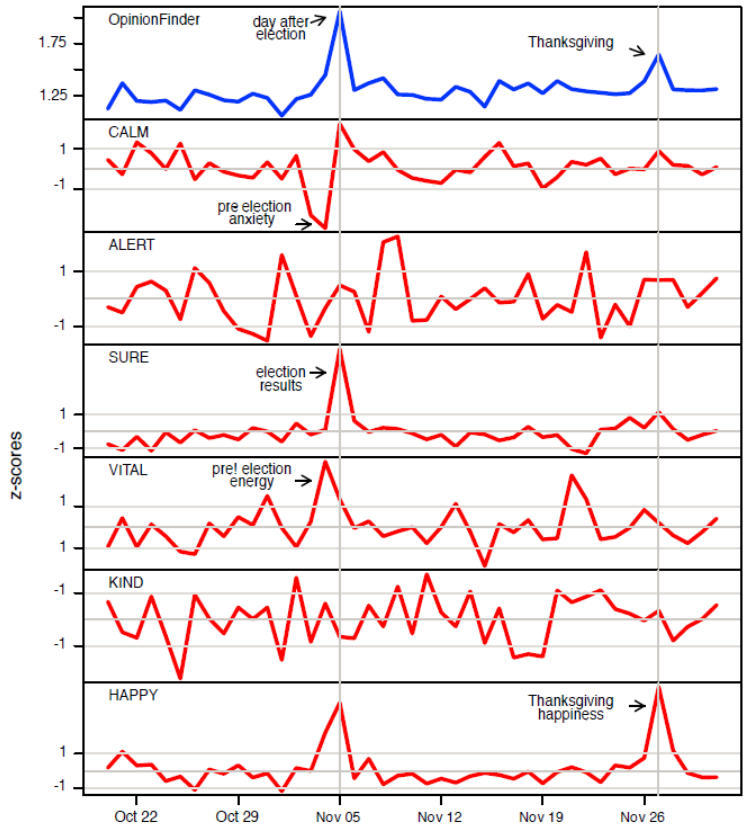


During the 100 days before the relationship starts, we observe a slow but steady increase in the number of timeline posts shared between the future couple. When the relationship starts ("day 0"), posts begin to decrease.

Quelle: https://www.facebook.com/notes/facebook-data-science/the-formation-of-love/10152064609253859, 2014

Twitter mood predicts the stock market

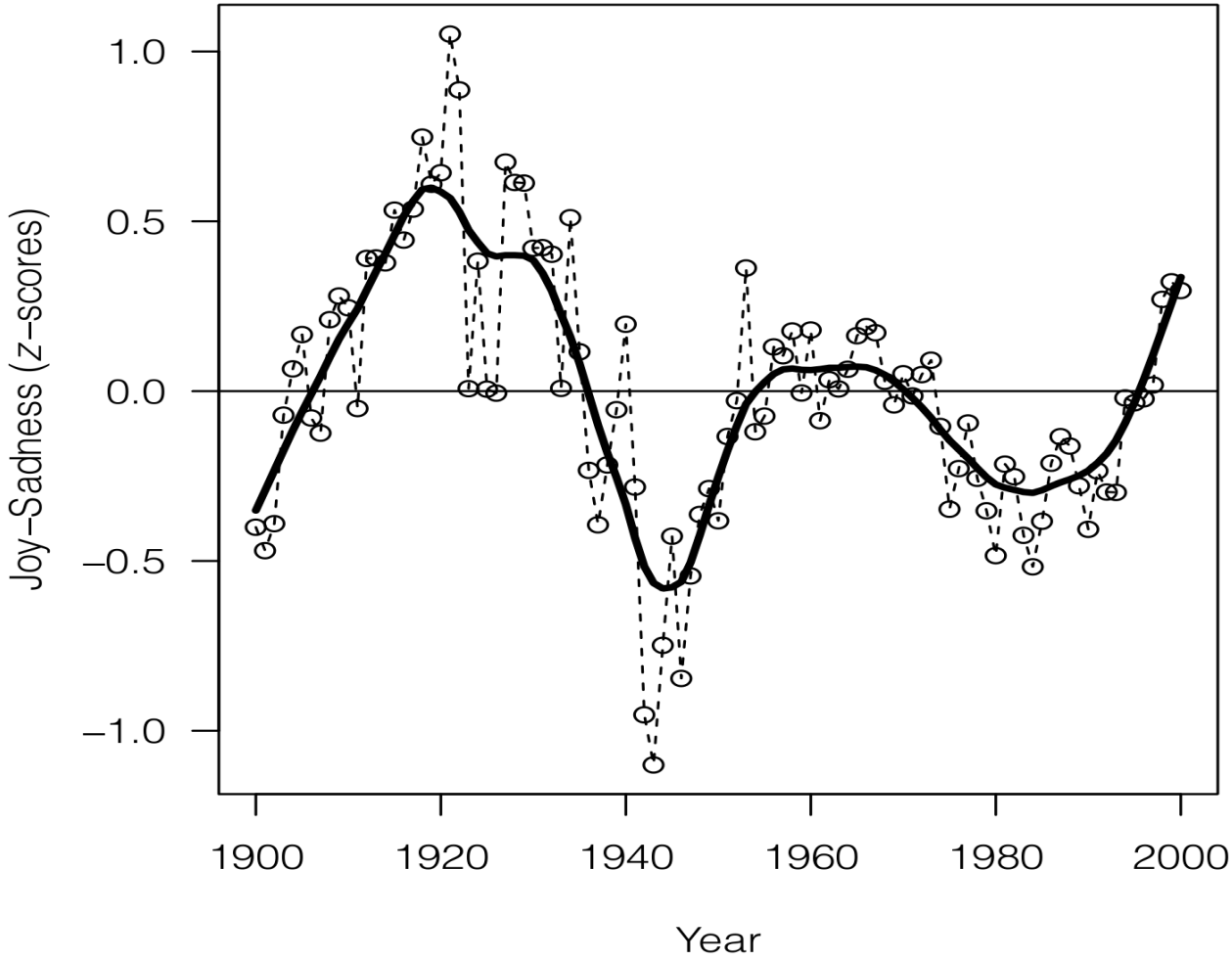
- Ziel: Vorhersage des „Dow Jones Industrial Average“ Börsenindex anhand von Emotionen in Twitter Nachrichten
- **Sentiment Analysis**: Wörter und Texte nach emotionaler Stimmung beurteilen
  - In diversen Python Packages enthalten



- **Volfefe Index**
- Messwert für die Volatilität von amerikanischen Anleihen in Abhängigkeit von Tweets von Donald Trump
- https://en.wikipedia.org/wiki/Volfefe_index

Sentiment Analysis

Stimmung im 20. Jahrhundert gemessen am Vorkommen von positiven und negativen emotionalen Ausdrücken in Büchern, sortiert nach Erscheinungsjahr



Sentiment Analysis

Zwei Datengrundlagen

- Google n-Grams:
  - Statistik über Wörter (1-Gram) oder Wortfolgen der Länge *n* (n-Gram) in den von Google digitalisierten Büchern
  - http://books.google.com/ngrams
- WordNet / WordNet Affect:
  - Datenbank und Klassifikation von Wörtern
  - http://wndomains.fbk.eu/wnaffect.html

Quelle: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

Smart Metering

Feingranulare Messung des Stromverbrauchs

Vorteile

- Vorteile für Energieerzeuger (Optimierung durch bessere Vorhersage)
- Vorteile für Energieverbraucher (flexible Kostenmodelle)

Sicherheitsbedenken

- Electrical devices exhibit a specific pattern of electricity consumption, most typically when switching on. This pattern represents a "fingerprint" of the device and allows to determine a family's lifestyle ("non-intrusive load monitoring")
- Even specific movies can be determined if the power consumption is measured with a resolution of ½ second, because the consumption varies with the amount of dark and light areas on the screen. A five-minute chunk of data is sufficient to determine the movie.

Quelle: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

Intelligenter Stromzähler
https://de.wikipedia.org/wiki/Intelligenter_Z%C3%A4hler

Customer Churn / Customer Attrition

Predicting and avoiding that existing customers cancel their contracts

Companies want to avoid that customers move away or renew their contract

- Especially for contracts with fixed duration (mobile phone, pay TV, utilities, ...)
- Depending on expected customer's value (CLV - customer lifetime value)

Various approaches, for example

- Classification of customers into categories
- Prediction of customer behaviour for each category
- Tailored offers for customers to renew their contracts

Analysis of interaction patterns

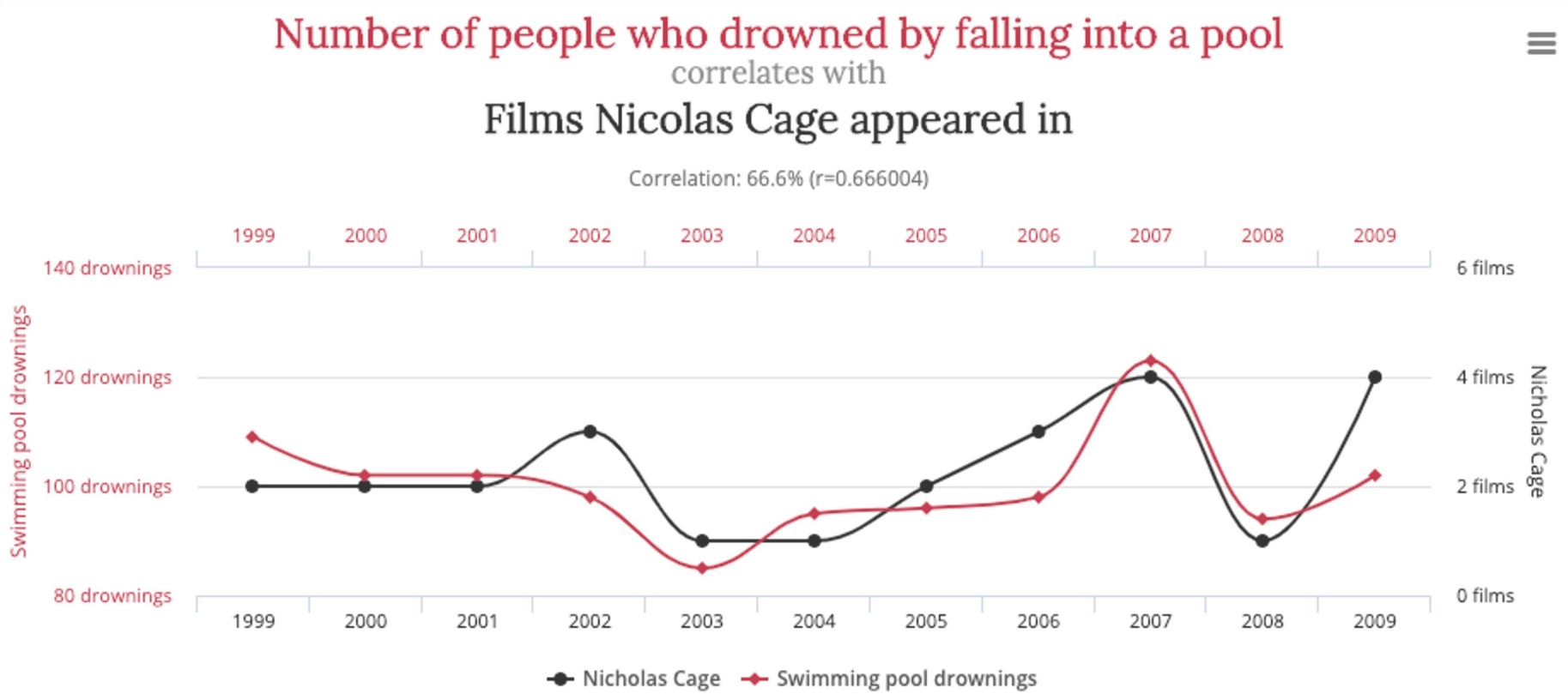
- Example: it has been observed that customer activity depends on the weather

Google Trends & Co.

Google Trends

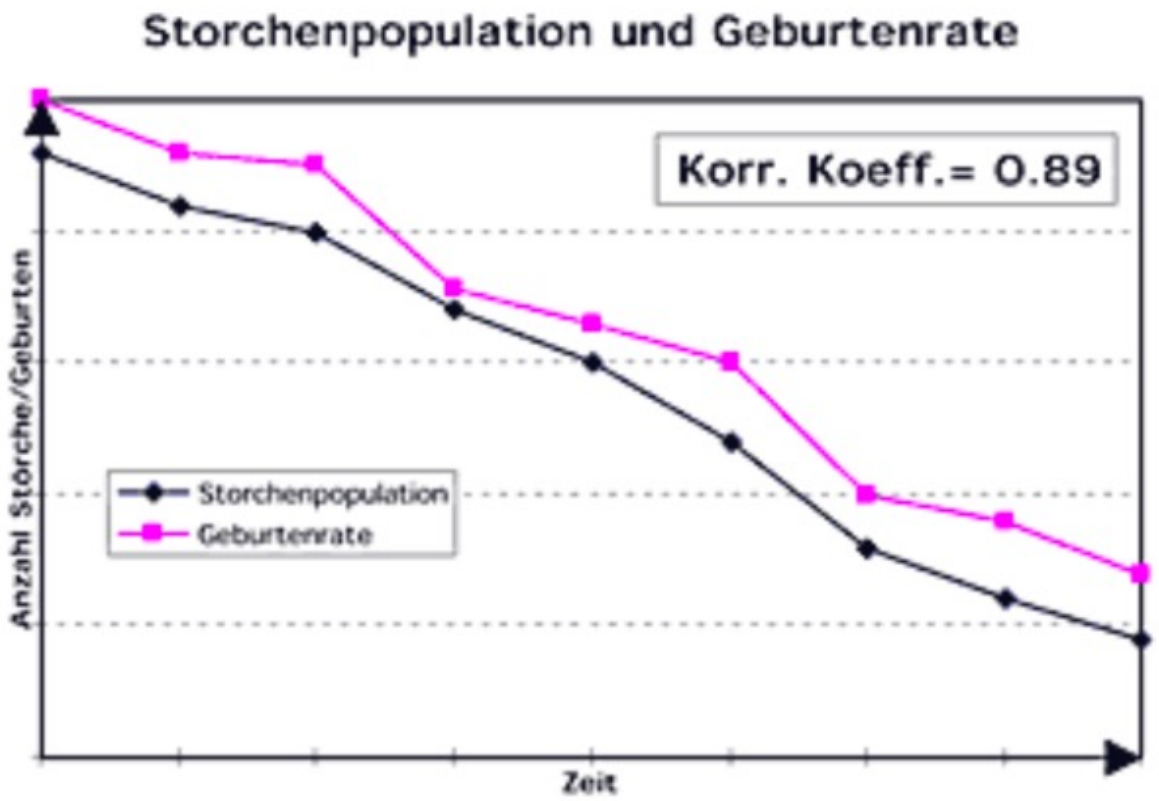
- Ermittelt den zeitlichen Verlauf der Häufigkeit von beliebigen Suchbegriffen
- Zeitliche Auflösung hängt vom gewählten Zeitraum ab, bis zu 1 Minute möglich
- Es wird nur die relative Häufigkeit angegeben (Wert zwischen 0 und 100), nicht die absolute Zahl der Suchanfragen
- http://www.google.com/trends

Korrelation vs Kausalität



Quelle: http://www.tylervigen.com/spurious-correlations

Korrelation vs Kausalität



Aus einem ähnlichen Muster darf man nicht auf einen kausale Zusammenhang schließen.

Siehe Scheinkorrelation, https://de.wikipedia.org/wiki/Cum_hoc_ergo_propter_hoc

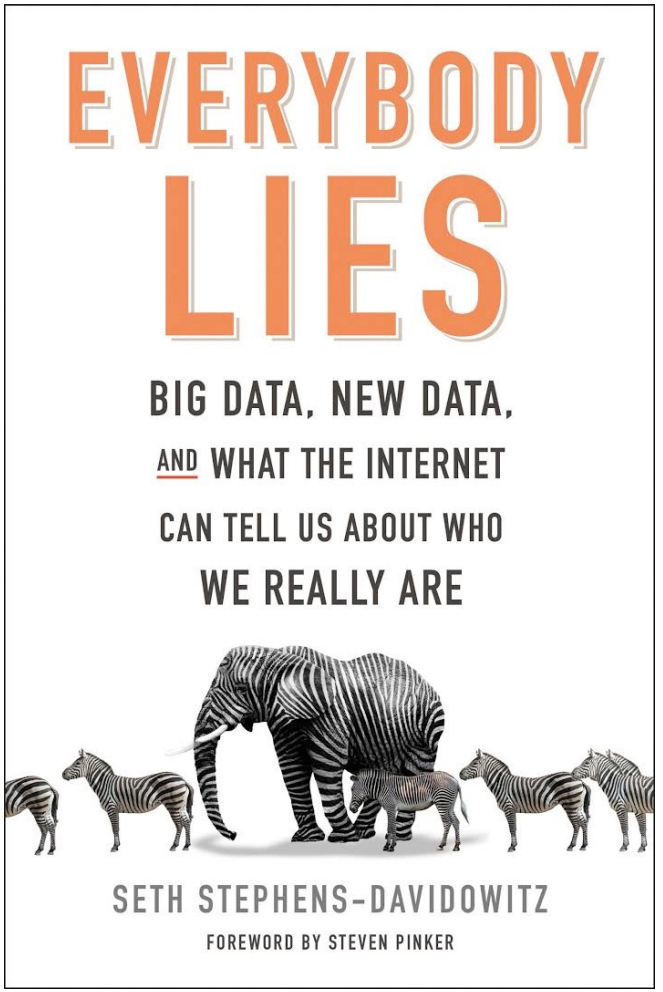
Daten enthalten die Wahrheit

Eine Suche bei Google enthüllt viel über die Person, die die Suche ausführt.

Anders als bei einem Fragenbogen oder einem persönlichen Interview repräsentieren die Suchanfragen die tatsächlichen Interessen und Fragen.

Die Analyse von Suchanfragen ist ein neuer quantitativer Zugang für Psychologie, Soziologie usw.

"In God we trust, everybody else must provide data." (Quelle unbekannt)

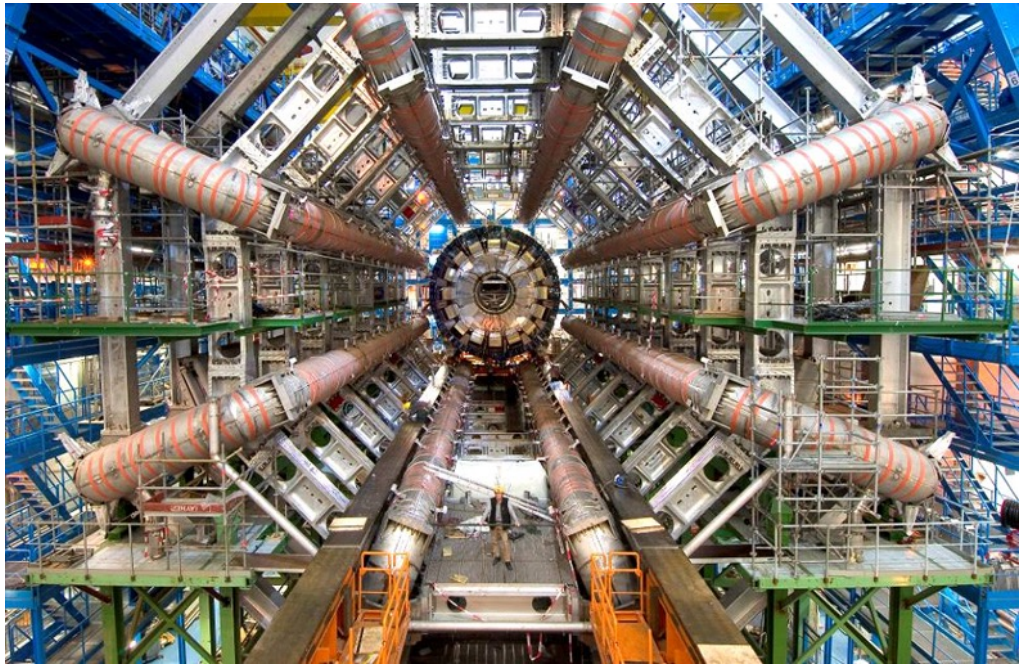


Wissenschaftliche Anwendungen

Physik

Teilchenbeschleuniger Large Hadron Collider (LHC) am CERN in Genf

- gilt als komplexeste jemals von Menschen gebaute Anlage (gebaut ab 1998)
- Entdeckung des Higgs-Bosons (genauer: Messungen konsistent mit Theorie) im Juli 2012, Nobelpreis für Physik 2013
- produziert 25 PB (25.000 TB) pro Jahr

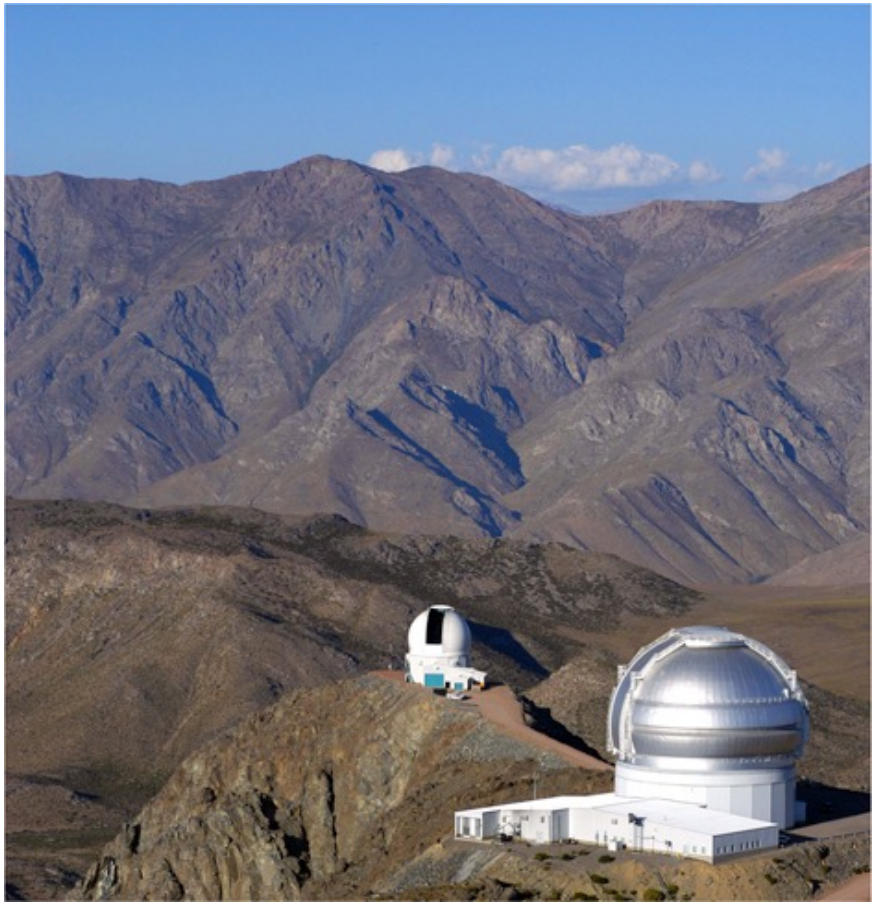


CERN in Google Street View: https://www.google.com/maps/views/streetview/cern?gl=de

Astronomie

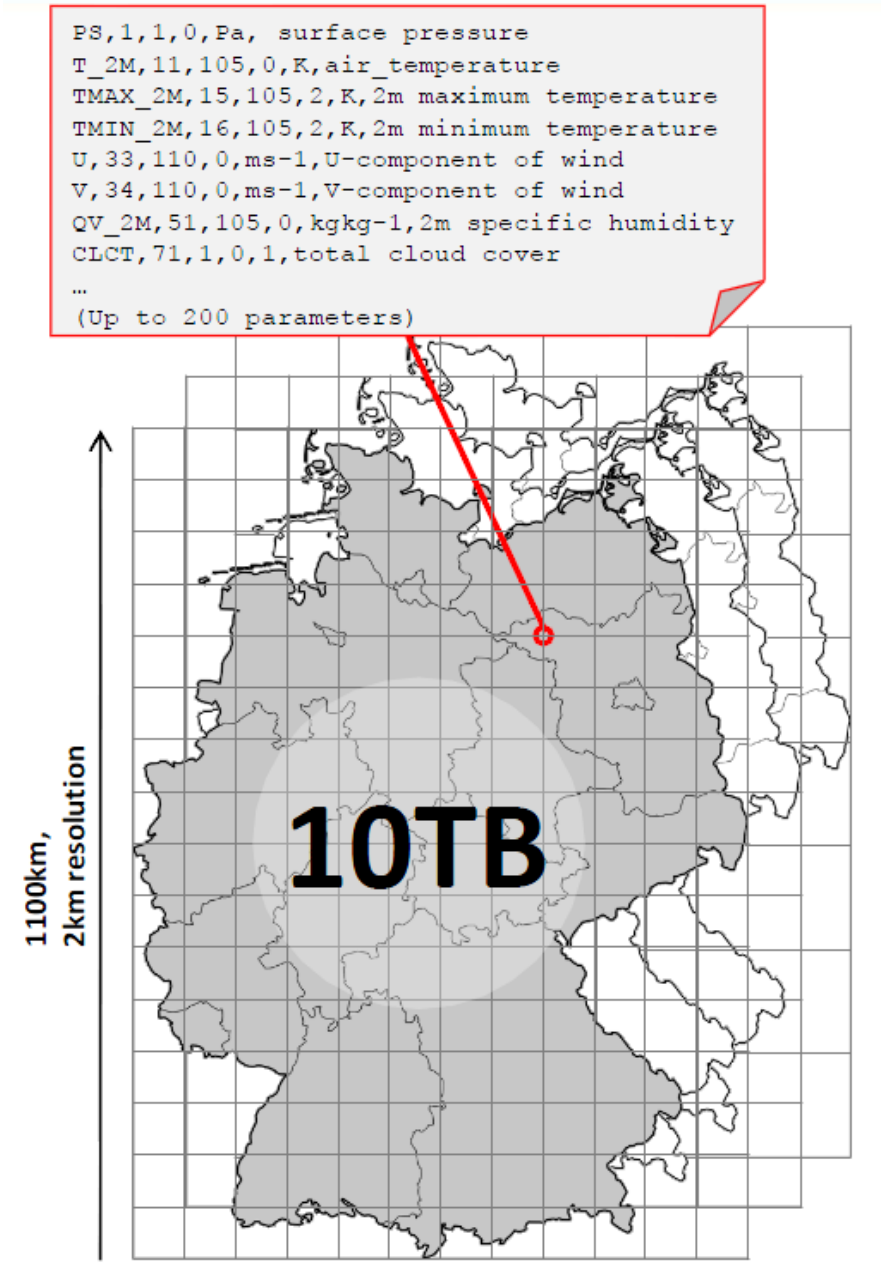
Large Synoptic Survey Telescope (LSST)
https://www.lsst.org/

- "the widest, fastest, deepest eye"
- auf dem Cerro Pachón in Chile
- Inbetriebnahme 2023 geplant
- Spiegeldurchmesser 8 m
- Kamera mit 3,2 Gigapixel
- produziert
  - 15 TB pro Nacht
  - 1.28 PB pro Jahr



Klimaforschung

- Datenraster für Deutschland (Beispiel)
  - Räumlich: 1100 km x 950 km
    Auflösung 2 km
  - Zeitlich: 3 Monate
    Auflösung 1 Stunde
  - ⇒ Datenvolumen 10 TB
- Klimamodelle mit verschiedenen Parametern
- Ziel: Erkennen von speziellen Phänomenen
  - Monsun, Dürren, Überschwemmungen

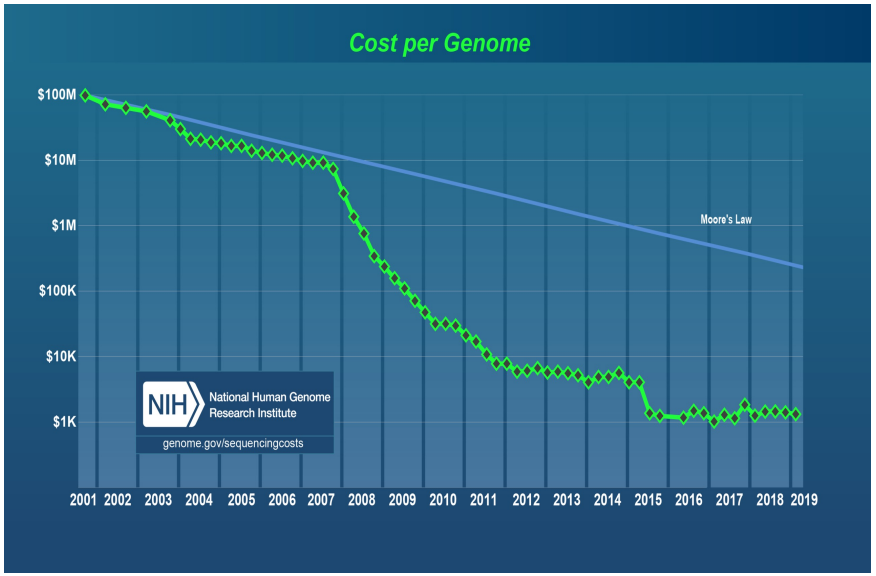


Quellen:
Volker Markl (TU Berlin)
Potsdam-Institut für Klimafolgenforschung (PIK)

Pesonalisierte Medizin

Precision Medicine

- Medizin basierend auf dem individuellen Genom des Patienten
- Prävention: Ermitteln von individuellen Risikofaktoren, um persönliche Gesundheitsrisiken zu bewerten und zu vermeiden
- Therapie: Medikamente, die speziell für eine genetisch definierte Gruppen von Patienten oder sogar für einen einzelnen Patienten entwickelt werden
- Kosten für whole genome sequencing (WGS) bei ungefähr 200 $ (September 2022)



Illumina NovaSeq X

Quellen
https://en.wikipedia.org/wiki/Personalized_medicine
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494865/

Grenzen von Big Data

Google Flu Trends

Google Flu Trends

Zeitlicher Verlauf und geographische Häufigkeit von Suchbegriffen, die mit Grippe in Verbindung gebracht werden, soll einen Grippe-Ausbruch vorhersagen

Februar 2009: Jeremy Ginsberg et al.
"Detecting influenza epidemics using search engine query data" (Nature Vol. 457)

Projekt war 2008 - 2014 aktiv

https://www.google.org/flutrends/about/

Google Flu Trends

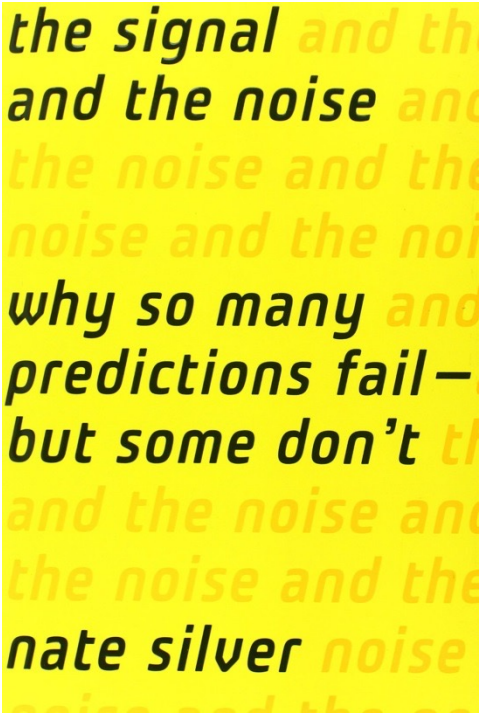
- März 2014: David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis" (Science Vol. 343)

- Die nichtsaisonale H1N1-Pandemie des Jahres 2009 wurde von Google Flu Trends schlicht übersehen. Danach hat ein verbessertes GFT das Ausmaß der saisonalen Epidemien 2011/2012 und 2012/2013 um mehr als 50 Prozent überschätzt, in einem Teilintervall war die vorhergesagte Zahl der Arztbesuche doppelt so hoch wie tatsächlich.
- Da Google weder die verwendeten Suchbegriffe noch den Algorithmus offenlegt, fällt die Suche nach den Gründen für diese Fehlprognosen schwer. Die Autoren vermuten eine "Big-Data-Hybris", die dazu führt, dass die Google-Forscher sich angesichts der Menge der Daten nicht ausreichend um deren Validität und Reliabilität kümmern.

Quelle: Süddeutsche Zeitung 2014, http://sz.de/1.1912226

The Signal and the Noise

- Nate Silver: „The Signal and the Noise – Why so many predictions fail, but some don‘t“ (2012)



- Große Datenmengen sind kein Wundermittel.
- Noise (Rauschen – irrelevante Daten) wächst mindestens genauso schnell wie Signale (relevante Daten). Dadurch helfen sehr großen Datenmengen nicht immer, um nützliche Informationen zu extrahieren.
- Vorhersagen von hochkomplexen Sachverhalten bleiben schwierig
  - z.B. Chaotische Systeme („Schmetterlings-Effekt“)
- Das Wissen von menschlichen Experten ist wichtiger denn je, um die richtigen Fragen an die Daten zu stellen und sinnvolle Modelle zu entwickeln.

Alle Modelle sind falsch, aber manche sind nützlich.

George Box, britischer Statistiker (1919-2013)

Nutzen und Wert von Daten

Daten sind wertvoll

- "When hardware became commoditized, software was valuable.
  Now that software is being commoditized, data is valuable.“
  (Tim O'Reilly, 2011)

- commodity = Handelsware, Massenware

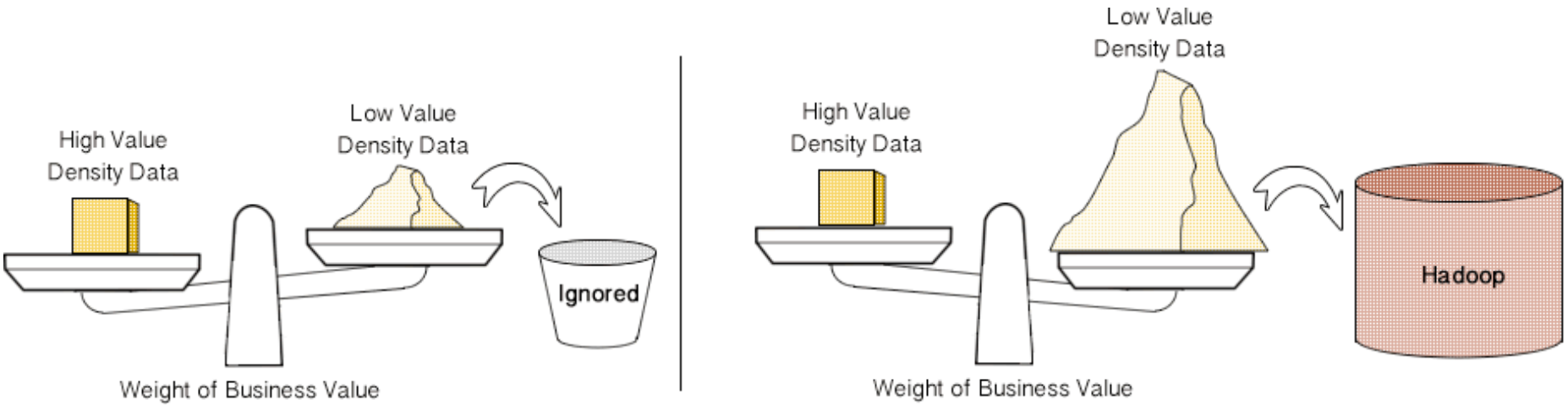
Marktplätze für Daten entstehen:

z.B. https://foursquare.com/ (Location Data for Mobile Advertising)

Open Data Initiative

- Ziel der freien Verfügbarkeit und Nutzbarkeit von Daten
- Statistiken und administrative Daten (Open Government)
- „Creative Commons Zero“ Lizenz:
  Verzicht auf Copyright bzw. bedingungslose Lizenz („no rights reserverd“)

High Value Density vs. Low Value Density



- Daten mit hoher „Wertdichte“: transaktionale Daten, Kundendaten, ...
- Daten mit geringer “Wertdichte“: Monitoring-Daten, Suchanfragen und Klicks von Benutzern, räumliche und zeitliche Verteilungen...

- Daten mit geringer Wertdichte wurden früher ignoriert, können aber wertvolle Informationen enthalten.

Quelle: McKnight Consulting Group, 2014

Ziele der Datenanalyse

Descriptive Analytics

- Auswertung von Vergangenheitsdaten: *What happened?*
- Beschreibende Statistik, Business Intelligence (BI): Berichte, Reports

Predictive Analytics

- Vorhersage von zukünftigen Entwicklungen: *What will likely happen?*
- Data Science: neue Erkenntnisse aus Daten gewinnen

Prescriptive Analytics

- Empfehlung für zukünftige Aktionen: *What should I do?*
- Data Science: "actionable insights" gewinnen

Beispiel: Vorausschauende Wartung und Instandhaltung

Predictive Maintenance

- Überwachung von technischen Anlagen und Komponenten und rechtzeitige Wartung vor einem Ausfall
- Sensoren liefern regelmäßig Daten über die Anlage
- Training eines statistischen Modells zur Vorhersage der "Remaining Useful Lifetime" (RUL)

Data Analyst

Data Scientist

Data Engineer

Data Scientists

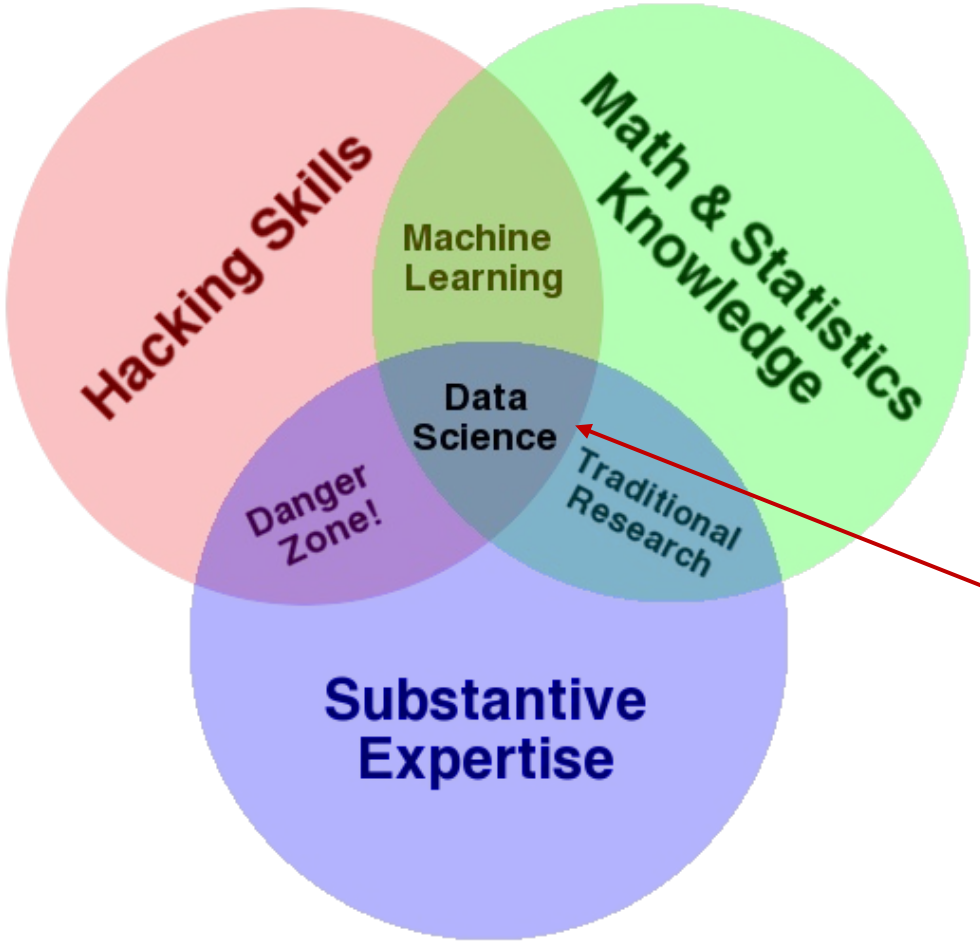
"The sexy job in the next ten years will be statisticians.
People think I'm joking, but who would've guessed that
computer engineers would have been the sexy job of the 1990s?"

"If you are looking for a career where your services will be in high demand,
you should find something where you provide a scarce,
complementary service to something that is getting ubiquitous and cheap.
So what's getting ubiquitous and cheap? Data.
And what is complementary to data? Analysis."

Hal Varian
(University of Berkeley und Chef-Ökonom bei Google)

Quellen:
[1] The McKinsey Quarterly, January 2009.
[2] http://freakonomics.com/2008/02/25/hal-varian-answers-your-questions/

Data Science Venn Diagram



Data Science Unicorn

Quelle: Drew Conway. The Data Science Venn Diagramm, 2010.

Data Science Venn Diagram

Mathematik und Statistik

- Statistische Analysen erstellen und interpretieren

Hacking Skills

- Tools und Programme beherrschen

Substantive Expertise

- Wissen über die Anwendungsdomäne

Danger Zone

- "Know enough to be dangerous"
- Fähigkeit, Analysen zu Fragestellungen der Anwendungsdomäne zu erstellen, aber nicht genug Statistikwissen, um die Analyse zu interpretieren und die begrenzte Gültigkeit der Analyse einzuschätzen.

Aufgabenbeschreibungen

Data Analyst

- Ziel: Business Value durch Datenanalyse
- Aufgaben: Daten aufbereiten, analysieren, visualisieren

Data Scientist

- "Senior Data Analyst"
- Ziel: Business Value durch komplexe Analysen und Vorhersagen
- Aufgaben: wie Data Analyst, plus Statistik und Machine Learning
- Vermittler zwischen Anwendungsdomäne/Fachabteilung und Data Analytics

Aufgabenbeschreibungen

Data Engineer

- Ziel: Management von Big Data Systemen, Bereitstellen von Daten
- Aufgaben: interne und externe Daten integrieren, speichern, bereitstellen

Data Architect

- "Senior Data Engineer"
- Ziel: Design und Architektur von Big Data Systemen
- Aufgaben: Design von Data Management Architekturen

Machine Learning Engineer

- Data Engineer / Data Scientist mit Vertiefung in Machine Learning

Zusammenfassung

- Was ist Big Data?
- Anwendungsfälle und Beispiele
- Grenzen von Big Data
- Data Scientist