

DataScience

...aus Sicht eines BWL-ers

- Ja. Powerpoint. Ich weiß
- Disclaimer: Bin noch kein Data Scientist
- Übersicht im Galopp
- Weitere Folien im Anhang
- Folien selbst zwar nicht verschicken, aber quellenfolie
- Ziele mit Vortrag: Wissen vermitteln, Präsentationsfähigkeit üben, Gelerntes verinnerlichen, überdenken + Feedback

Gliederung

- I. Grundlagen
- II. Definition „Data Science“
- III. Anwendungsgebiete und Industrien
- IV. Vorgehen – und Werkzeugkoffer
- V. Gefahren – von außen
- VI. Probleme – von innen
- VII. Erfolge

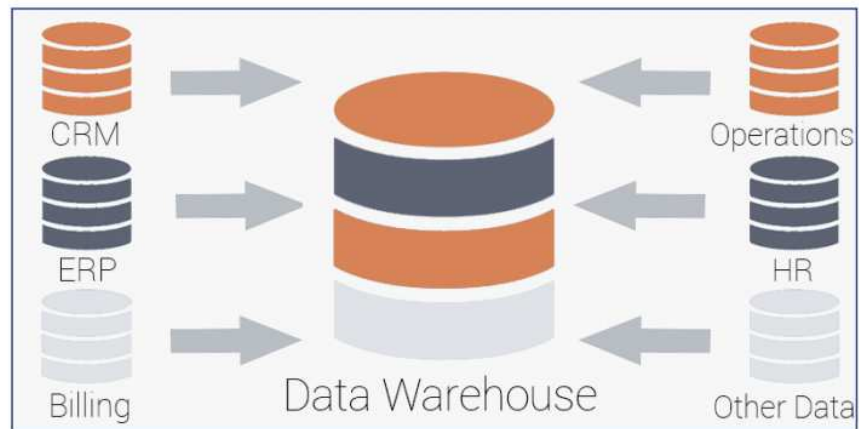
(nicht genauer erläutert):

- VI. Statistik – und weitere Probleme
- VII. Machine Learning, Data Mining, Artificial Intelligence
- VIII. Datenbanken und Parallelisierung

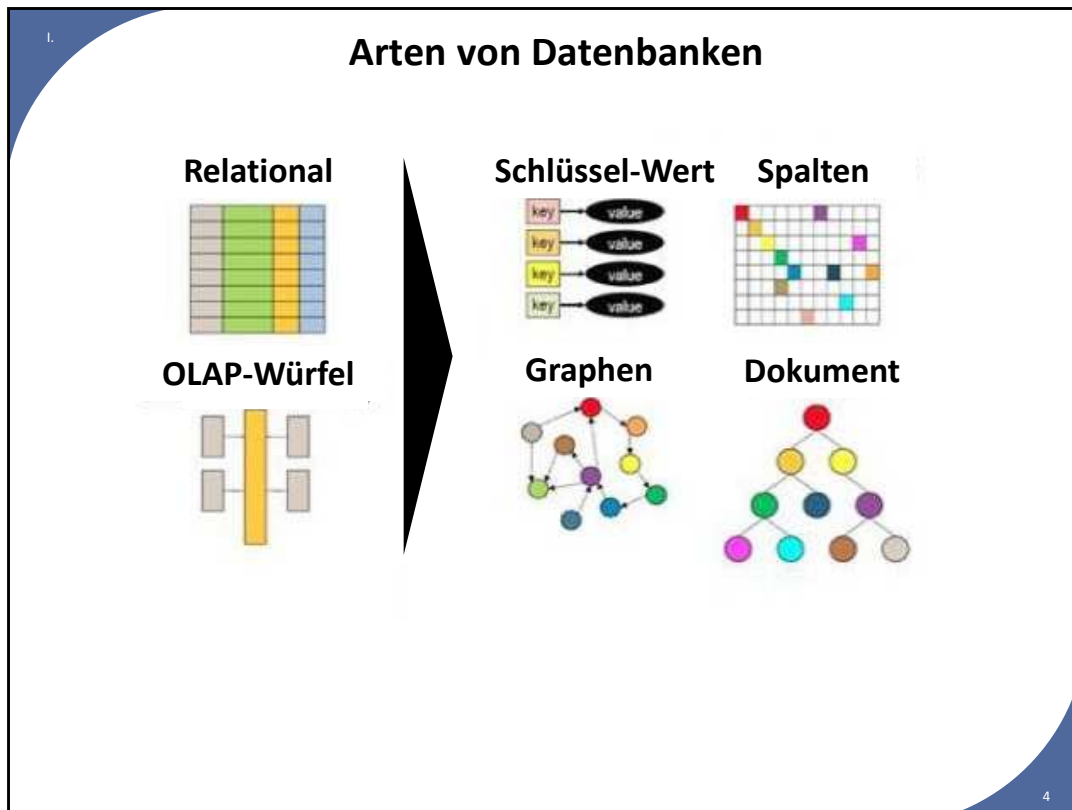
2

- Viele Beispiele
- Nicht genauer erläutert: Diese Punkte würden die Veranstaltungszeit sprengen

Datensammlung

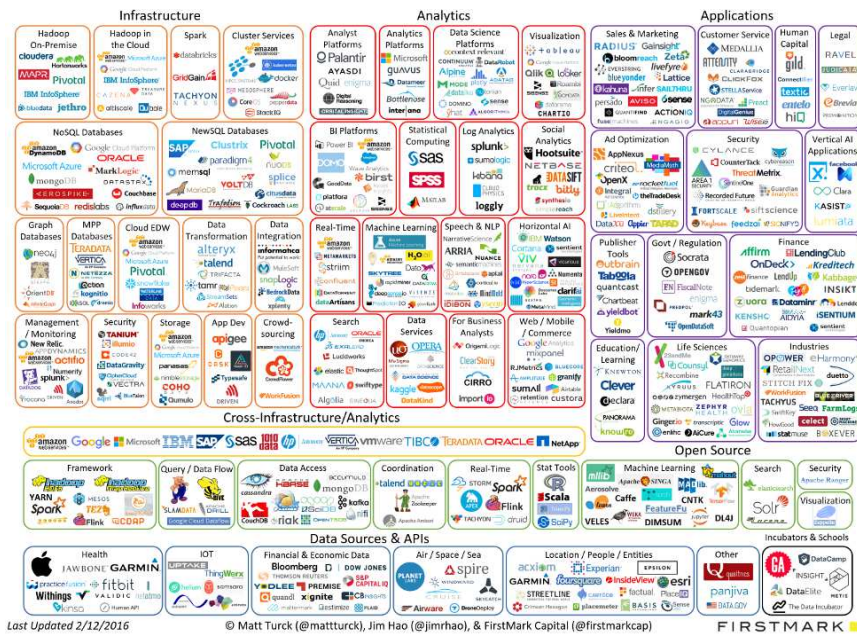


- <http://www.kdnuggets.com/2016/07/database-key-terms-explained.html>
- Buchempfehlung: Alex Wright, Glut: Mastering Information Through the Ages. Weitgefasst, nicht nur Technikentwicklung



- <http://www.kdnuggets.com/2016/07/seven-steps-understanding-nosql-databases.html>
- OLAP: Online Analytical Processing – Daten aus Datawarehouse
- Vs. OLTP (quasi vorgeschaltet)
- structured, semi-structured vs unstructured. Webscraping
- ERD-Diagramm vs. keine Tabellenschemata im vorhinein
- SQL vs. Individual query System
- 3-6 V's. Am Häufigsten: Volume, Velocity, Variety
- ACID within a node and eventually consistent across the cluster
- CAP: Consistency – availability (Verfügbarkeit/Antwortzeit) – partition tolerance (Ausfalltoleranz)

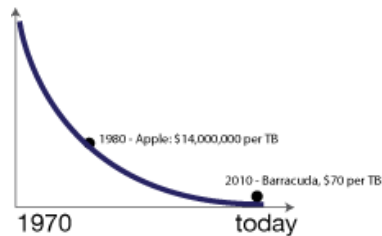
Big Data Landscape



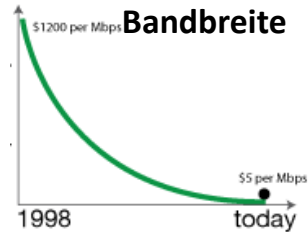
- Trotzdem nur 30% aller Datenbanken
- Riesiges Ökosystem
- Nur Übersichtsfolie, nicht durchlesen
- <http://www.kdnuggets.com/2016/08/big-data-key-terms-explained.html>

Warum jetzt?

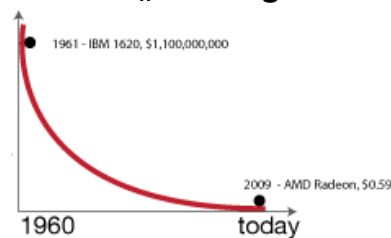
Kosten Datenspeicher



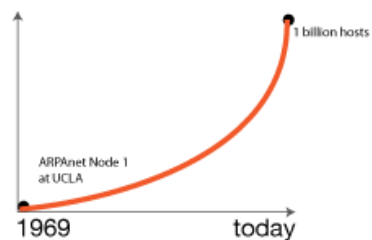
Kosten Übertragungs-Bandbreite



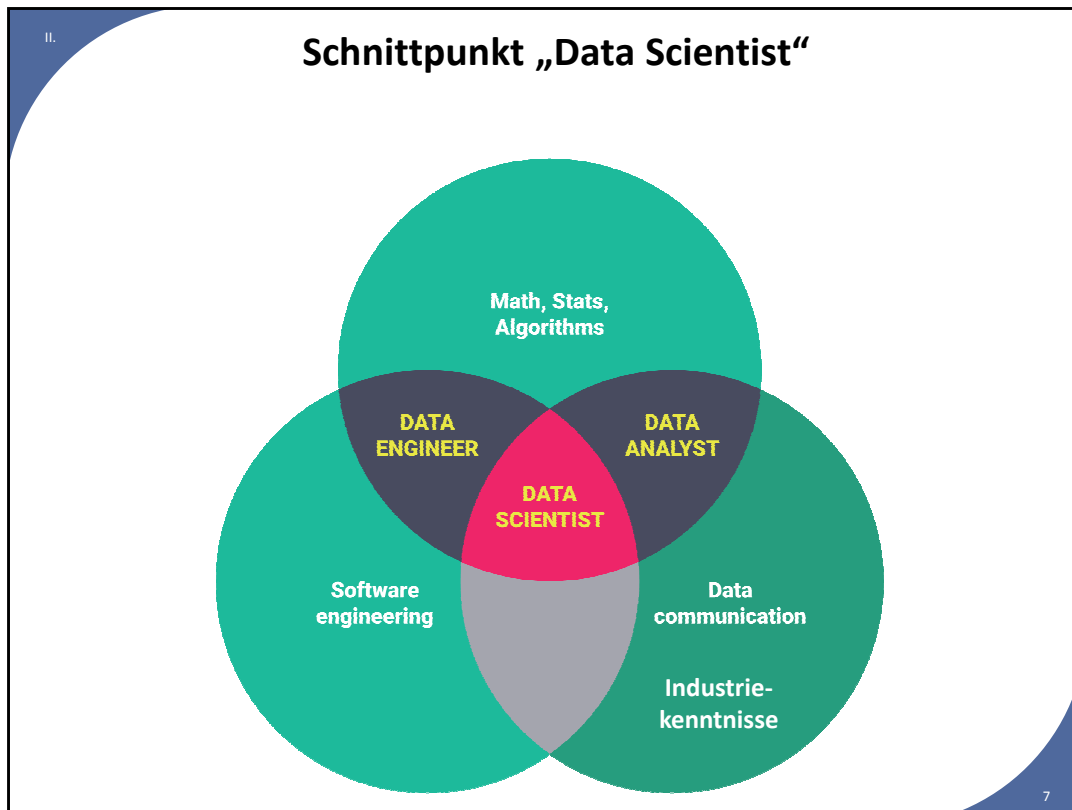
Kosten „Leistung“



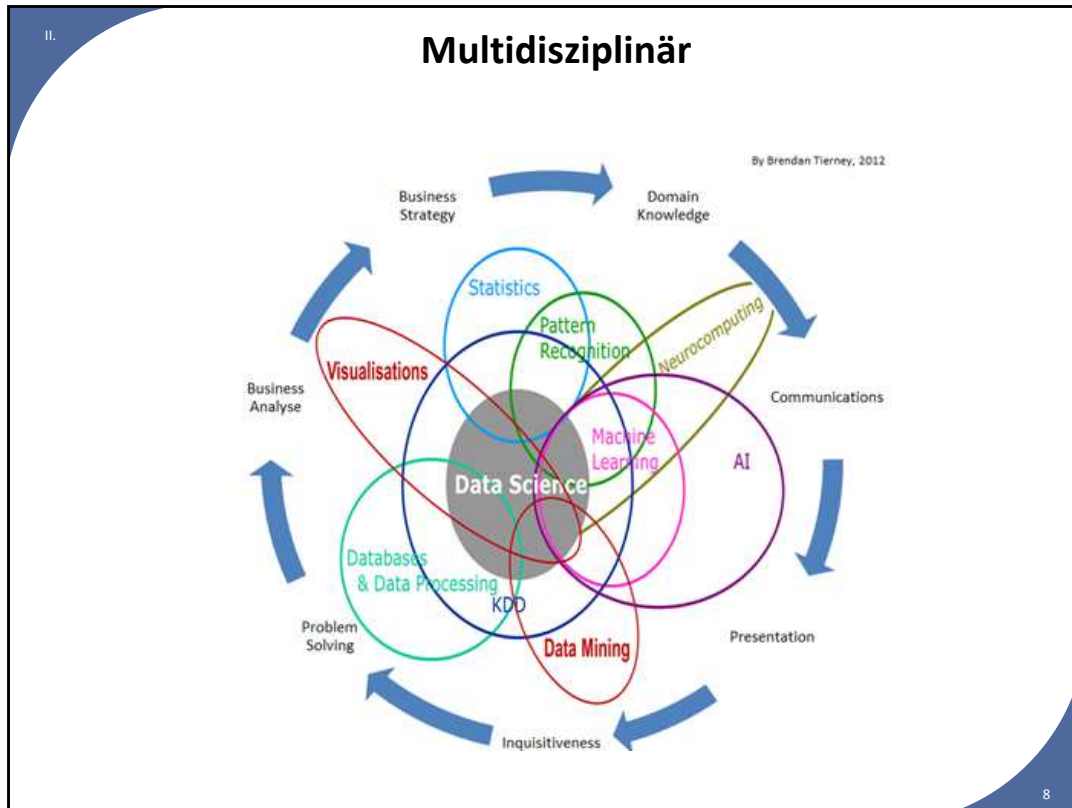
Netzwerk-Größe



- Fehlt: CPU installierte Leistung + RAM ging hoch
- <http://radar.oreilly.com/2011/08/building-data-startups.html>
- Big Data gestern – heute – morgen
- GB: 0,7 CD – 4,7 DVD (single) bzw. 8,5 (double), 25 Bluray (single) bzw. 50 (double)
- Gigabyte, 10^9 , Terabyte 10^{12} , Petabyte 10^{15} , Exa, Zetta, Yota
- 1936 bzw. 1938: World Brain, H.G. Wells: "any student, in any part of the world, would be able to sit with his projector in his own study at his or her convenience to examine *any* book, *any* document, in an exact replica" (p. 54).
- Moores Law. Quanteneffekte. Andere Computerarchitektur, z.B. HP's "The Machine"
- Siehe auch "Wissensgenerierung"



- „sexiest job des 21. Jhd“: teuer - wird aber auch automatisiert
- Jobrollen überschneiden sich. Titel z.B. Analytics XY, Business Solution Architect, Data Consultant
- Josh Wills 2012: Jemand der besser ist in Statistik als ein Software Entwickler und ein Besserer Softwareentwickler ist als ein Statistiker
- Head of Data Engineering at Slack. Vorher bei Cloudera, Google
- Bekannteres Venn-Diagramm: 2010, Drew Conway (R-Blogger)
- Keine einheitliche Definition
- Interdisziplinär



- KDD = Knowledge Discovery from Data
- Gefällt mir, gerade weil komplex und unübersichtlich
- Alles überlappt sich – Es gibt nicht „die eine Definition“ von Data Science

Meine Definition

- Erkenntnisse/ datenbasierte Entscheidungen/
datenbasierte Produkte
- Datenaufbereitung und -auswertung – oft viel & parallel
- statistische Methoden, programmieren
- Mustererkennung, Automatisierung, Skalierung
- Englisch, Datenbanken, Statistik, Machine Learning,
Programmieren, Industrie, Kommunikation

<http://www.kdnuggets.com/2016/11/big-data-data-science-explained.html>

<https://www.oreilly.com/ideas/what-is-data-science>

- Voraussetzungen: Neugierig, Lernbegeistert – Leseratte hilft, Zahlenaffin, kommunizieren können
- IQ, Background: nicht soo wichtig
- Jose Quesada, der Initiator des Berliner Data Science Retreat, hat dies wie folgt formuliert: "A good GitHub profile is ten times better than any certification".
- <http://www.datascienceglossary.org>
- Covert overt data product:
- Again, a covert data product; and the fact that it's covert is precisely what makes it valuable. A human can't deal with the raw data, and digesting the data into hourly summaries so that humans can use it makes it less useful, not more. What doctors and nurses need isn't data, they need to know that the sick baby is about to get sicker. Do we want products that deliver data? Or do we want products that deliver results based on data? We're evolving toward the latter, though we're not there yet.

The data is still in the foreground, but we're starting to look beyond the data to the bigger picture: better quality of life.

Anwendungsgebiete

(z.B.):

- Textkorrektur und -Voraussage
- Analysen – Marketing, Controlling, Social Media...
- Voraussagen und Empfehlungen
- Fraud Detection
- (Künstliche Intelligenz / AI)
- ...



**Alle
(möglich)**

- Anwendungsgebiete, Industrien, Vorgehen
- ... Szenarienanalyse
- Sprachverständnis
- Gesichtserkennung auf Fotos
- Umwelt Erkennung für selbstfahrende Autos

III.

Industrien

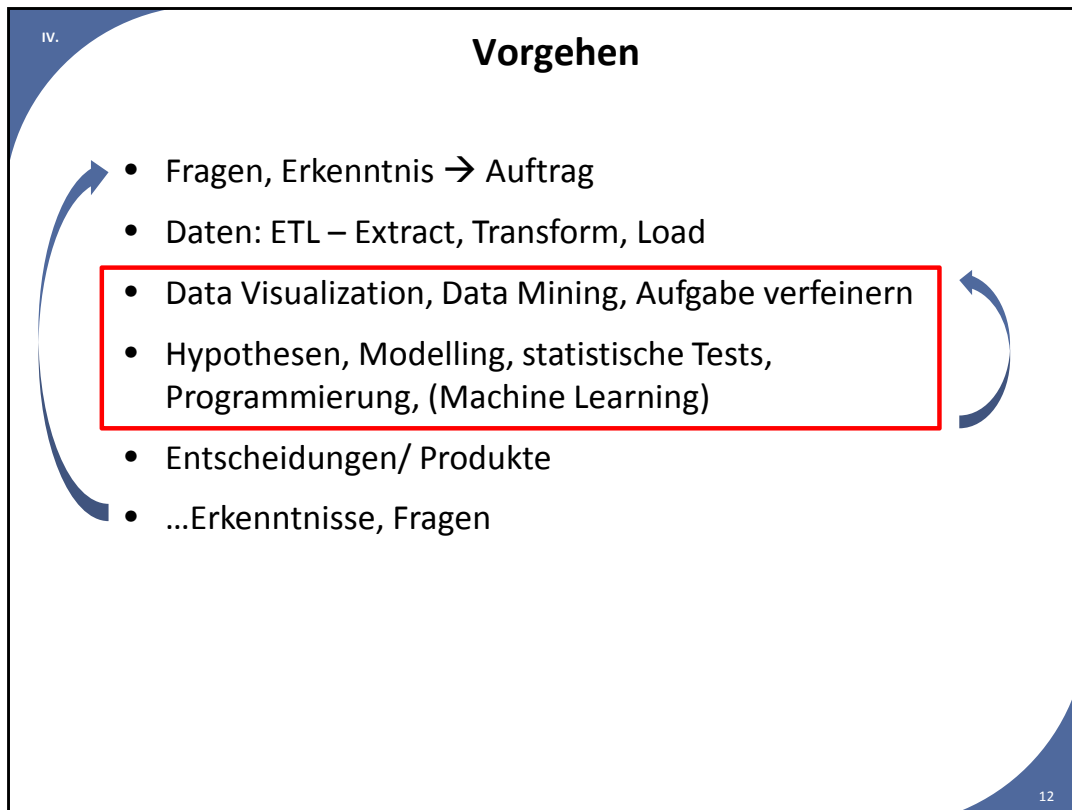
(z.B.):

- Handel
- Medizin
- Mobilität
- Wissenschaft
- Landwirtschaft
- Militär
- ...

➔ **Alle
(möglich)**

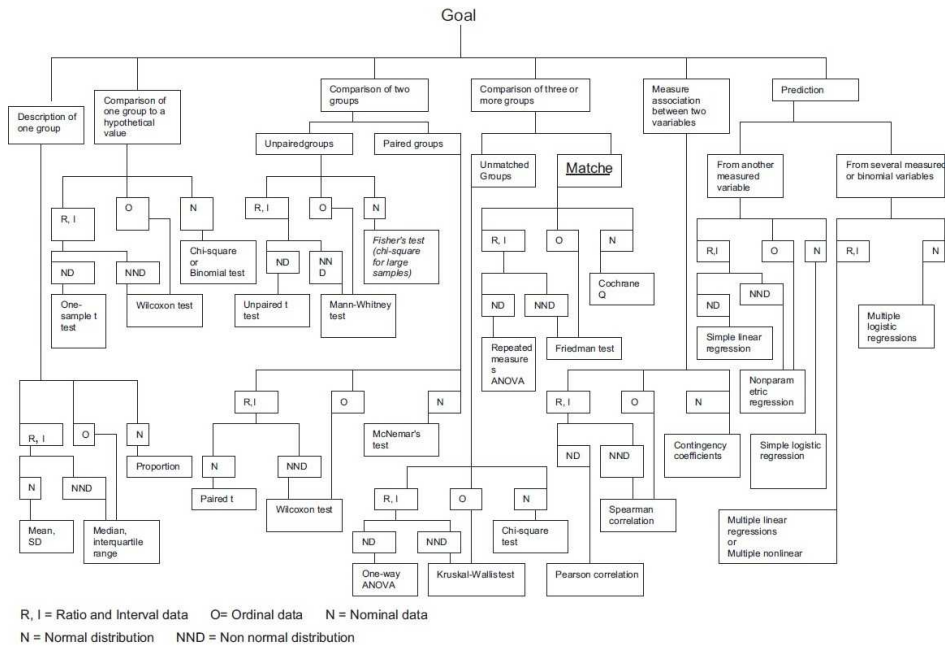
11

- Produktion - Einsparen von Lagerfläche - keine Überproduktion – oder Predictive Maintenance
- Fertigung
- Logistik - Flottenmanagement
- Kryptographie
- X 47-B
- Higgs im CERN + Gravitationswellen durch 5 Sigma-Abweichung



- ETL: Auslesen, Bereinigen/ Vereinheitlichen, Umformen Aufteilen, Unterbereiche („Data Marts“) einlesen in spezielle Programme. Mit
- Garbage in, Garbage out
- Meetings + Brainstorming im Team – Research - Modelling - Coding – Bug Hunting – Dokumentieren – Planen – Reports
- Ausrollen - Bug Hunting
- Evaluieren

Statistik: Welche Formel?



- Verschiedene Grafiken – hier eine der detaillierteren, aber auch verwirrenderen
- Entscheidungsbäume: Automatisierung

Werkzeugkoffer

- SQL, R, Python...
- Hadoop / HDFS + MapReduce, Hive, Pig, Spark,...
- HBase, Cassandra, Neo4j, Hana...
- Talend, Clover...
- Tableau, PowerBI, Qlik, Cognos...
 - oder ggplot2, Shiny, matplotlib, Bokeh...
- Rapid Miner, Knime, SPSS, SAS...

- <http://www.kdnuggets.com/2015/08/big-data-question-hadoop-spark.html>
- <https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/>
- <https://data-science-blog.com/blog/2015/05/24/top-10-der-python-bibliotheken-fur-data-science/>
- Cognos TM1, Insight, Express, Automation
- Pentaho, Alteryx: EIGENTLICH kein ETL
- Emcien, hyper anna
- Sisense, Periscope Data
- Quid, Recorded Future, Palantir
- KNIME: 15.3. 20 Mio. € Investment
- Andere Sprachen: julia, Scala, Haskell, Wolfram Language
- ETL: man kann auch z.B. Pentaho Kettle, Alteryx...
- D3.js, Gephi,
- MongoDB, Couchbase,

Zwischenübersicht

- I. Grundlagen
- II. Definition „Data Science“
- III. Anwendungsgebiete und Industrien
- IV. Vorgehen – und Werkzeugkoffer

- V. Gefahren – von außen
- VI. Probleme – von innen
- VII. Erfolge

v.

KEINE PANIK!!!

Datenschutz

Missbrauch

Identitätsdiebstahl

news bubble

fake news

Alternative Fakten

data exhaust

Überwachung



Datenkrake

ausspionieren

Regularien

16

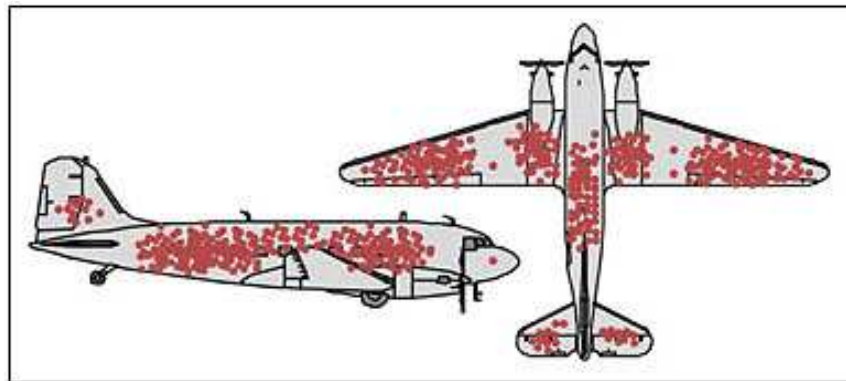
- „Allgemeine“ Gefahren von außen
- Keep Calm, do Whatever you normally do and say 42
- <https://en.wikipedia.org/wiki/USA-247>
- <http://www.zeit.de/digital/internet/2014-09/nrol-39-logo-krake>
- Kabelstück, das für den Test von Weltraumfahrzeugen verwendet und Octopus Harness genannt wird.
- Offiziell soll der Oktopus symbolisch für die Fähigkeit stehen, Probleme überwinden und aus jeder Situation einen Ausweg finden zu können. Dafür sei das Tier bekannt, befanden die NRO-Mitarbeiter, zudem gelte das Tier unter Seefahrern als besonders intelligent.
- Auch die nach dem Globus greifenden Tentakel würden gut passen. Sie sollen zeigen, dass sich die Feinde der USA in Zukunft nirgendwo mehr verstecken können. Oder wie der Manager der Mission in einem für den internen Gebrauch vorgesehenen Artikel zitiert wird: "Der Oktopus steht für die Idee. [...] Wir haben unsere Finger überall, zu jeder Zeit.,"
- 2012 – also vor Snowden

Widerstände im Management



- „Die zweite Option fühlt sich Richtig an. Die nehmen wir“
- Dilbert wagt sarkastischen Widerspruch: „Sollen wir immer ignorieren, was die Daten uns sagen – oder ist das eine einmalige Ausnahme“
- Das nennt sich „Intuition“
- Dilbert: Das ist nur einen Schritt von Hexerei entfernt“
- Schattenplanungssystem
- Dilbert oder auch xkcd sind Fundgruben

Selbsttäuschung

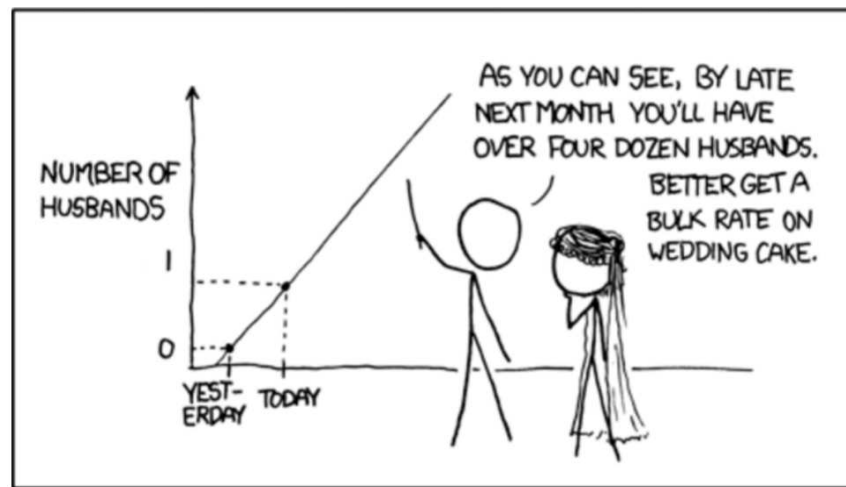


Credit: Cameron Moll

https://en.wikipedia.org/wiki/List_of_cognitive_biases

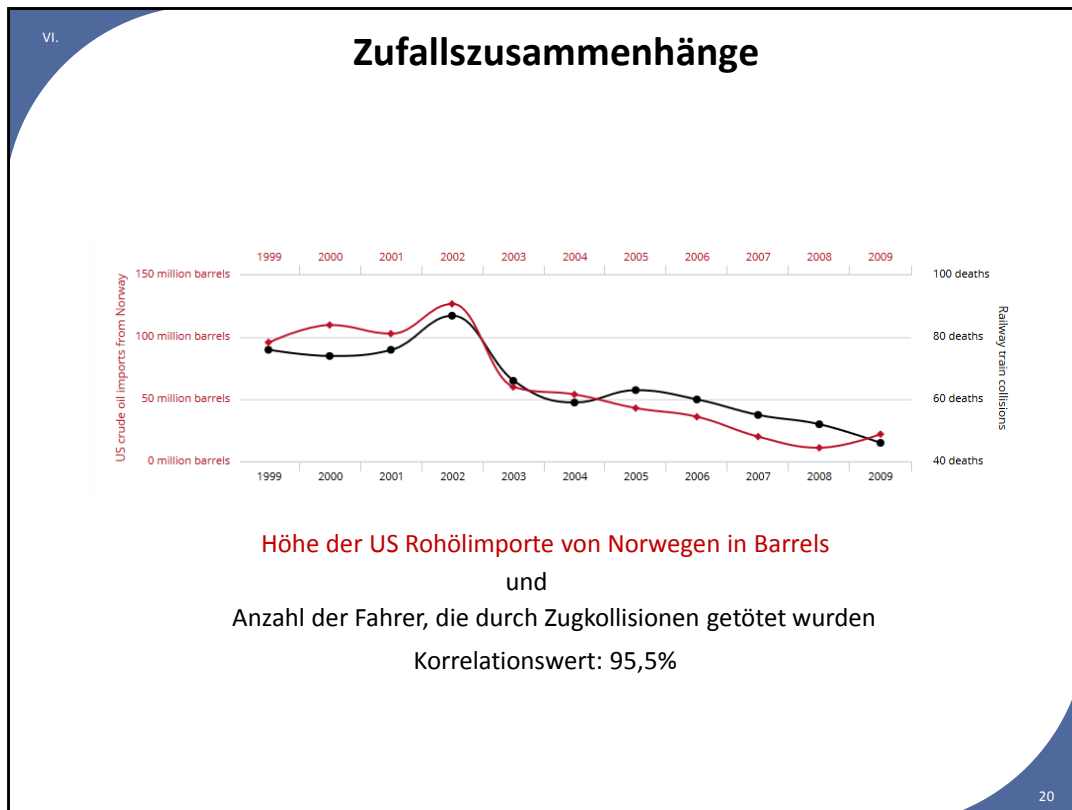
- Fehlschlüsse, Denkfallen
- Airforce: Überlebensquote von Bombern im 2. WW
- Ok, Bild ist Douglas C-47 Skytrain statt B24 Liberator – aber who cares
- Panzerung im Bereich Einschusslöcher
- Überlebensbias – eine von vielen Selbsttäuschungen
- Daten oft statistisch → immer mit Wahrscheinlichkeiten verbunden
- Augenzeugen sind sehr ungenau: Wir sehen was wir zu sehen erwarten, z.B. durch Vorprägung oder Beeinflussung

Extrapolation



Quelle: xkcd

- Wenn man naiv einen Trend in die Zukunft fortschreibt, kommt raus
- Gestern: 0 Ehemänner. Heute: 1. Nächster Monat: über 4 Dutzend
- ...besser Mengenrabatt bei Hochzeitskuchen aushandeln
- Trends mit Konfidenzintervall – statt einer Linie – und Methode (Lineare Regression, Multivariate Adaptive Regression Splines, Polynomgraph X Dimensionen)
- MARS = siehe Salford Systems



- 10 Jahre hoher Gleichlauf
- Data dredging
- Tylervigen
- S&P 500 mit Butterproduktion, Anzahl von Schafen, Käseproduktion in Bangladesh und US
- Storchproblem
- Google correlate
- Signifikanz ist durch Volumen irrelevant
- Bigger equals weirder
- Data dredging = data fishing = p-hacking
- Data dredging: Ohne Hypothese Daten untersucht. Ungleich data mining
- Korrelation/ Gleichlauf bedeutet nicht ursächlicher Zusammenhang...nur stärkerer oder schwächerer ANZEICHEN für Zusammenhang...Logik hinterfragen
- Was ist Ursache, was Effekt – was kommt zuerst
- Unterschiede: Was, Warum, Ob
- Bayes

Ideales Ergebnis – Selbstfahrende Autos

[Tesla Model X, autopilot avoids a crash in The Netherlands](https://www.youtube.com/watch?v=FadR7ETT_1k)

12 sek

- Hyperlink: https://www.youtube.com/watch?v=FadR7ETT_1k
 - Idealerweise
 - 12 sek
 - Wann ertönt das Warnsignal, und wann bremsst das Auto ab
 - Gab ja auch schon einen tödlichen Unfall auf Grund falscher Umgebungserkennung
 - False Positivs
 - Mobileye, an Israeli automotive company that makes vehicle safety systems used by dozens of carmakers, including Tesla Motors
- “If a self-driving car follows the law precisely, then during rush hour I might wait in a merge situation for an hour,” Shalev-Shwartz says
- Intel acquired Mobileye

Ideales Ergebnis - Robotik

[Introducing Handle](#)

90 sek

- Hyperlink: <https://www.youtube.com/watch?v=-7xvqQeoA8c>
- Boston Dynamics (z.B. BigDog, Cheetah, Atlas): von Google gekauft – letztes Jahr an Toyota verkauft
- ...oder SWORDS Special Weapons Observation Reconnaissance Detection System,
- Nr. 5 lebt
- Baxter, Sawyer von Rethink Robotics
- Computer sind nicht dumm – sie sind “differentially enabled”
- <https://www.youtube.com/watch?v=cR-YIZ9NdIA>

Weitere Einführungen/Workshops

<https://andcode.de/cook>

<https://www.facebook.com/cookandcode/>

23

- Auf Nicht-Itler ausgerichtet
- Wordpress, -Blockchain, Chatbots, Arbeitsalltag automatisieren
- SQL, Java, PHP

**Für Fragen stehe ich
gerne zur Verfügung.**

Michael Schmidt, ms@grauschattierung.de

Quellen und weitergehende Informationen

- 2-min-papers (Youtube)
- Analyticsvidhya The Art of Data Science,
Roger Peng
- Crash Courses (Youtube)
- Datacentral
- data-science-blog (**d**) The Master Algorithm,
Pedro Domingos
- Kdnuggets
- MOOCS: Coursera/ Udemy...
- O'Reilly Six honest serving-men,
Rudyard Kipling
- Quora
- Slideshare
- Stackexchange
- TED-Videos

25

- I KEEP six honest serving-men
(They taught me all I knew);
Their names are What and Why and When
And How and Where and Who.
I send them over land and sea,
I send them east and west;
But after they have worked for me,
I give them all a rest.

I let them rest from nine till five,
For I am busy then,
As well as breakfast, lunch, and tea,
For they are hungry men.
But different folk have different views;
I know a person small—
She keeps ten million serving-men,
Who get no rest at all!

She sends'em abroad on her own affairs,
From the second she opens her eyes—
One million Hows, two million Wheres,
And seven million Whys!

The Elephant's Child

- Nick Bostrom: Superintelligence. Paths, Dangers, Strategies
- Eric Siegel: Predictive Analytics: The Power to Predict
- Carlos Bueno: Lauren Ipsum_ A Story About Computer Science (Kinderbuch)

Die Zukunft

- VR (Oculus, Vive), AR (Pokemon, Hololens)
- Exoskeleton, Enhancements, Cyborgs
- Mensch-Maschine-Interfaces
- Sawyer, Asimov, Talon
- Unmöglich: Foundation Trilogie, Sherlock, Minority Report
- Data Exhaust, Always on, DNA Computing
- Automatisierung, Jobverlust, Klassen. Grundeinkommen?
- De-Augmented Reality
- Arbeitserleichterung. WALL·E? Ready Player One?
- Bewusstsein? Sklaverei? Ethik? Maschinenaufstand?

„Die Zukunft ist noch nicht geschrieben“

- Manches mit ., manches mit ?
- Arthur C. Clarke: Jede hinreichend fortschrittliche Technologie ist von Magie nicht zu unterscheiden
- Memristoren
- Neuristoren
- HP The Machine
- De-Augmented Reality: <https://www.youtube.com/watch?v=gpum4nK2wOM>
- Mglw. auch Hivemind?

Zahlenverständnis

A) Verarbeitung des Gehirns

B) Komplexität des Themas

Durchschnitt (Avg) von x: 9

Standardabweichung von x: 11

Durchschnitt (Avg) von y: 7,50

Standardabweichung von y: 4,1

Korrelationswert: 0,816

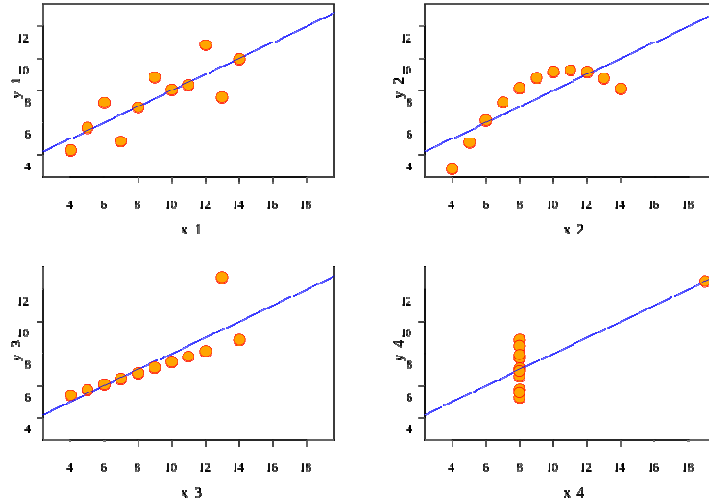
Regressionsgleichung: $y = 3,00 + 0,500x$

27

- A picture is worth a thousand words. An interface is worth a thousand pictures. Ben Shneiderman
Warum Datenvisualisierung? Ausspruch „Ein Bild...“
- Verarbeitungsgeschwindigkeit – und Aufwand – des Gehirns
- Einem angehenden Mathematiker 4 verschiedene Datenbündel geben
- Durchschnitt (Average), Standardabweichung, Korrelation/Gleichlauf, Regressionsgleichung/Geradengleichung

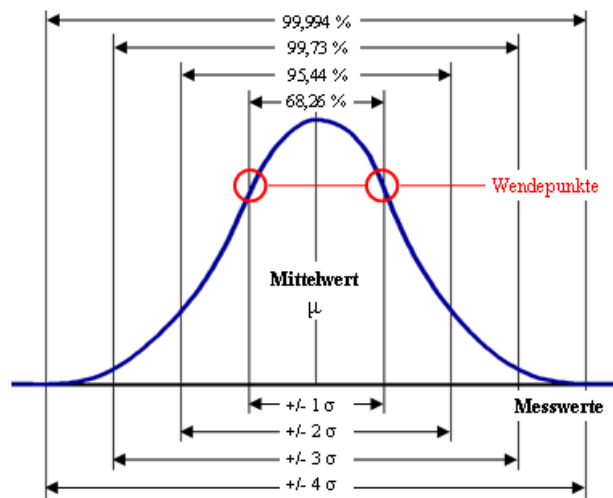
Visualisierung

Francis Anscombes Quartett, 1973



- Seitdem weitere Beispiele veröffentlicht
- Summenstatistiken
- Nächste Folie: komplexe Zusammenhänge nur anhand summierter Aussagewerte von oftmals (statistischen) Daten zu verstehen

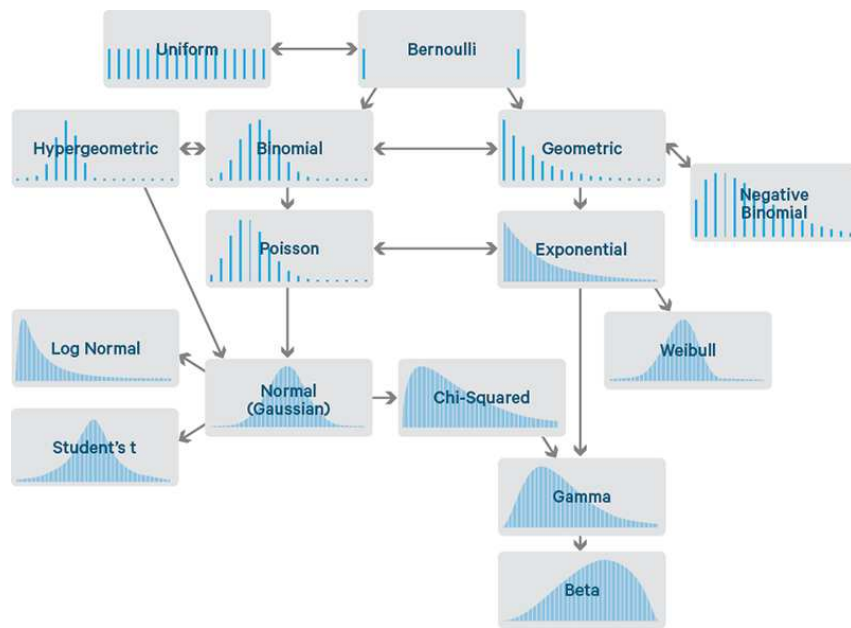
Normalverteilung



29

- Jugendliche: Passt in Mathe auf + Mathe KANN Spaß machen
- http://www.ted.com/talks/conrad_wolfram_teaching_kids_real_math_with_computers
- Wolfram Language (Mathematica, Wolfram Alpha, Siri)
https://www.youtube.com/watch?v=_P9HqHVPeik
- R ist weiter verbreitet. Häßlich, aber es gibt auch RStudio, RCmdr,
- Physics:
https://www.youtube.com/playlist?list=PL8dPuuaLjXtN0ge7yDk_UA0ldZJdhwkoV
- Higgs, Gravitationswellen: über Statistik auf 5 Sigma bewiesen

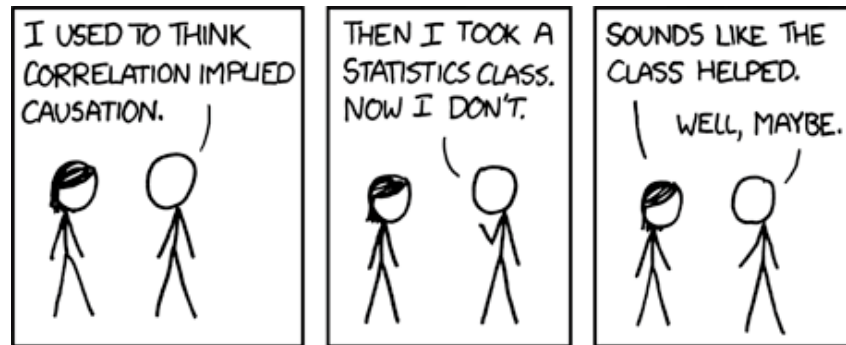
Verteilungen



30

- Sample vs. Population
- Diskret vs. Kontinuierlich
- Diskret:
 - Hypergeometrische Verteilung
 - Binomialverteilung
 - Poisson-Verteilung
- Kontinuierlich:
 - Normalverteilung
 - Logarithmische Normalverteilung
 - Betragsverteilung 1. Art
 - Rayleigh-Verteilung (Betragsverteilung 2. Art)
 - Weibull Verteilung

Unsicherheit

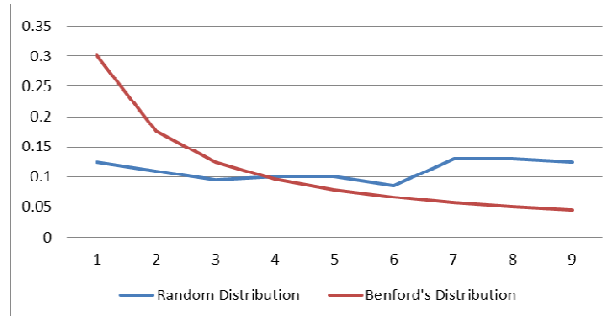


31

- xkcd

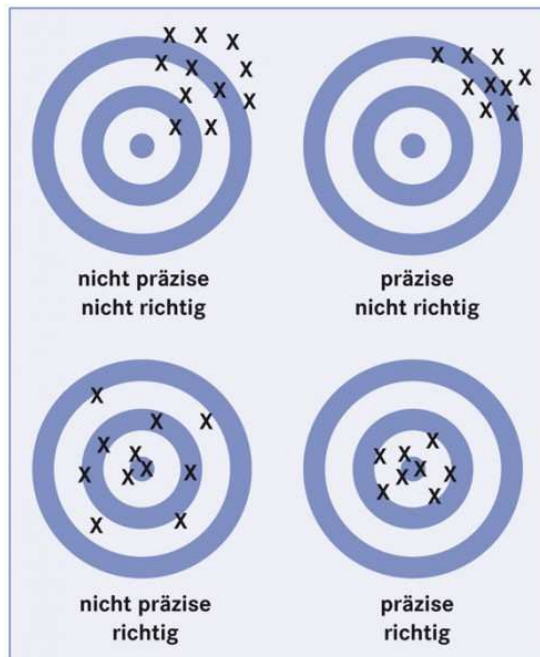
Benford

Leading Digit	Occurring Frequency
1	30.10%
2	17.60%
3	12.50%
4	9.70%
5	7.90%
6	6.70%
7	5.80%
8	5.10%
9	4.60%

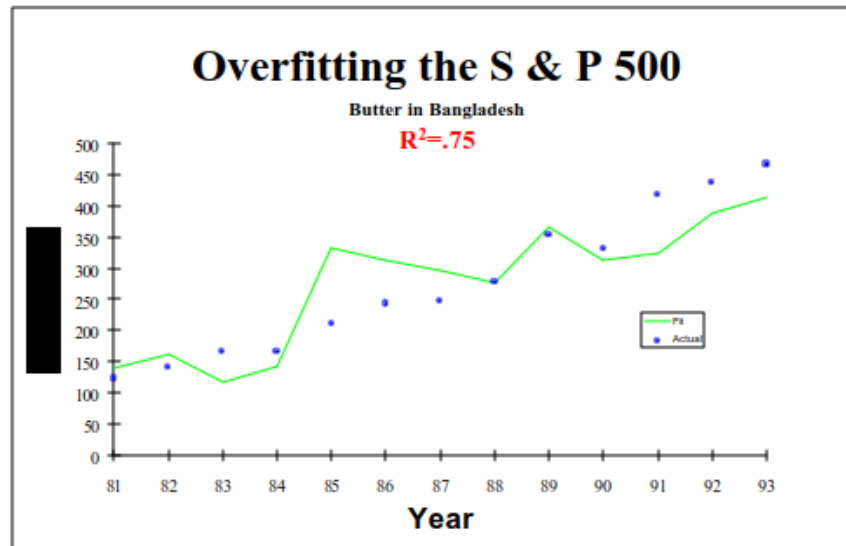


- Math4uandme.com

Präzision vs. Streuung



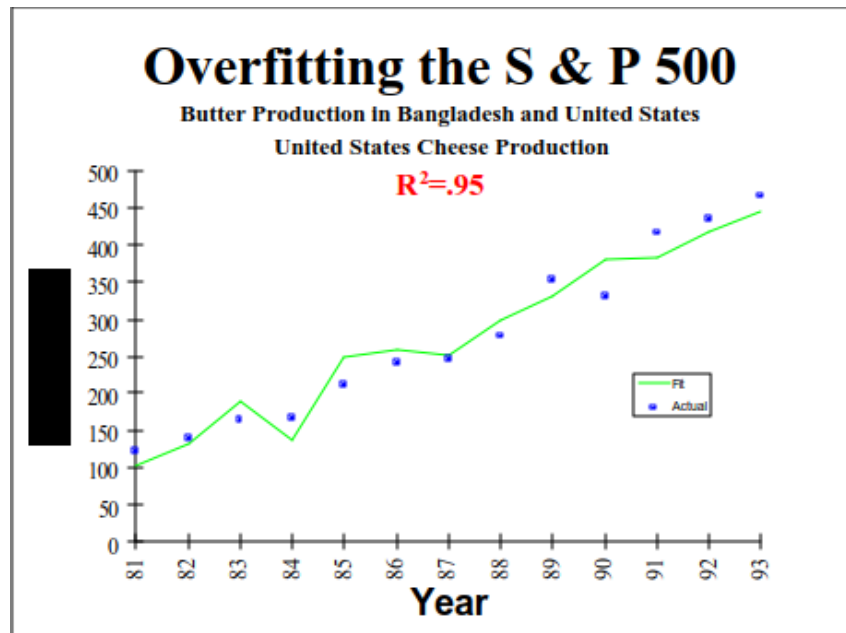
Overfitting 1/3



34

- STUPID DATA MINER TRICKS: OVERFITTING THE S&P 500
- David J. Leinweber, Ph.D. Caltech

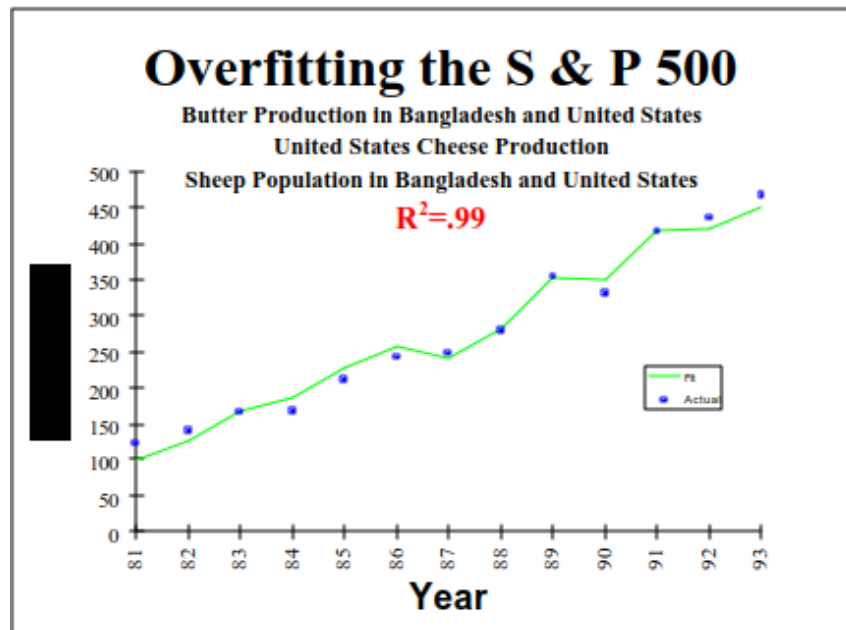
Overfitting 2/3



35

- STUPID DATA MINER TRICKS: OVERFITTING THE S&P 500
- David J. Leinweber, Ph.D. Caltech

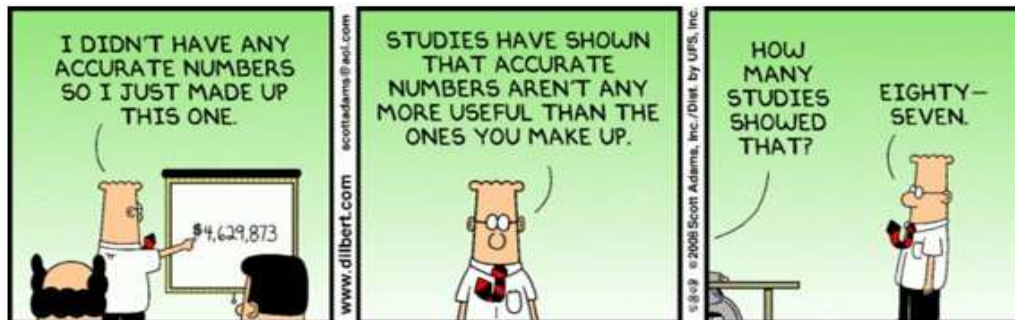
Overfitting 3/3



36

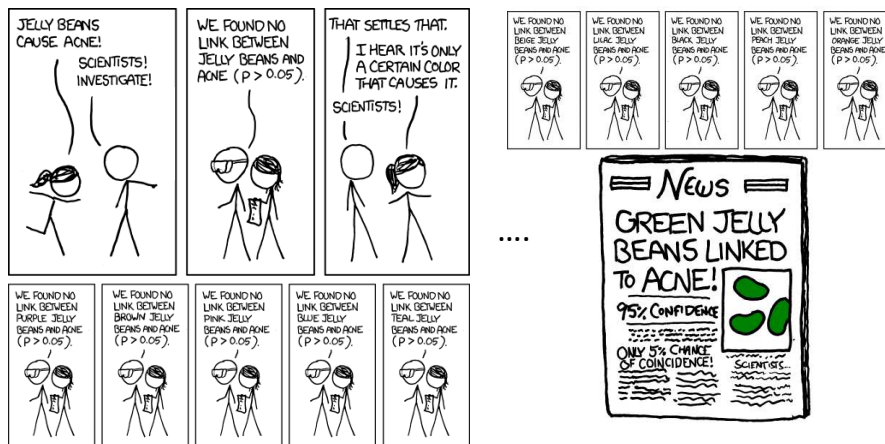
- STUPID DATA MINER TRICKS: OVERFITTING THE S&P 500
- David J. Leinweber, Ph.D. Caltech

Lügen mit Statistik



- Dilbert: Ich hatte keine genauen Zahlen, darum habe ich einfach diese erfunden
- Studien haben gezeigt, dass korrekte Zahlen in keiner Weise nützlicher sind als welche, die man erfunden hat
- Wie viele Studien zeigen das? 87
- Spielt mit „Lügen mit Statistik“
- Lügen mit Statistik geht natürlich – aber es geht ja auch ohne
- Lies, damn lies and statistics Marc Twain

p-hacking



- Forscher unterliegen auch „Veröffentlichungszwang“ → „Publication Bias“
- Why most published research findings are false
- Z.B. Bonferroni - Korrektur
- xkcd

Bayes I

Komplex + Kontraintuitiv

Krebstest, der mit 80% Sicherheit vorhandenen Krebs erkennt

→ Erkennt in 20% aller Fälle nicht, dass Krebs vorhanden ist

+

→ Gibt in 10% aller Fälle ohne Krebs ein falsches „positives“ Signal

In der untersuchten Altersgruppe haben 1% Krebs

Es werden 1000 Personen untersucht

39

- Kontraintuitive Ergebnisse – Bedingte Wahrscheinlichkeiten/ Bayes

Bayes II

Komplex + Kontraintuitiv

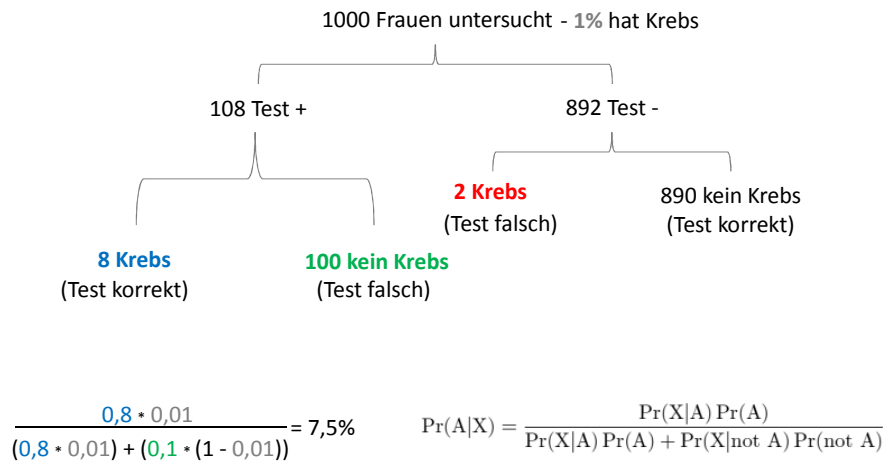


40

- Kontraintuitive Ergebnisse – Bedingte Wahrscheinlichkeiten/ Bayes

Bayes III

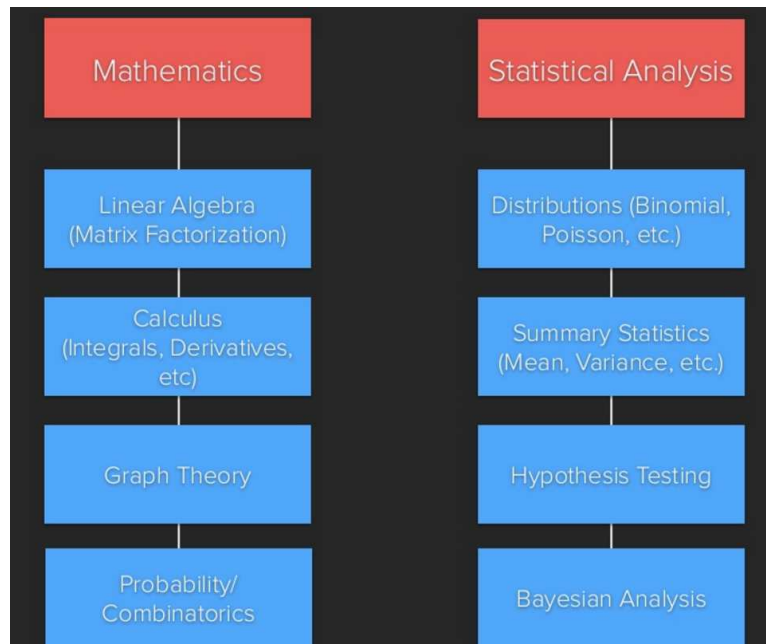
Komplex + Kontraintuitiv



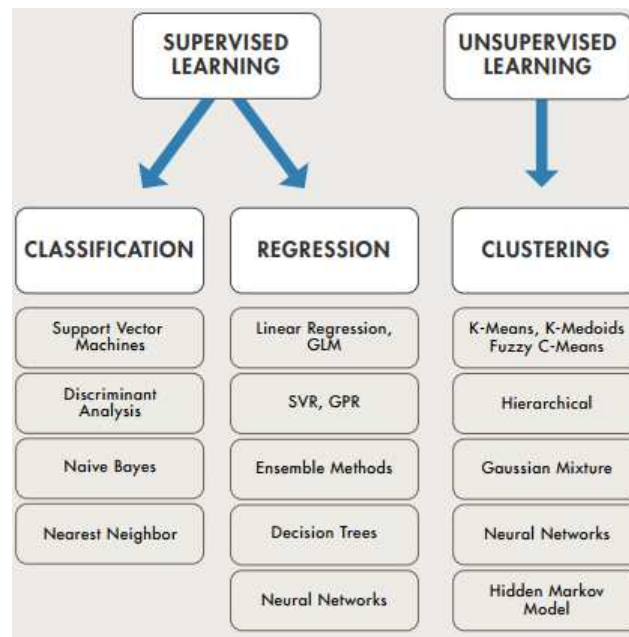
41

- 80% positiv, falls der Patient Krebs hat
- 0 % positiv, falls der Patient keinen Krebs hat
- In der Altersgruppe haben 1% aller Personen Krebs
- True positive: 1% * 80% = 0,8% bzw. 0,008
- False positive: 99% * 10% = 9,9% bzw. 0,099
- Wahrscheinlichkeit, ein positives Resultat zu erhalten: 10,7% bzw. 0,107 (false + true pos)
- Wahrscheinlichkeit, ein negatives Resultat zu erhalten: (false + true neg)
- Wahrscheinlichkeit, dass der Patient wirklich Krebs hat, wenn es der Test anzeigt: 7,5% bzw. 0,075
- Erhöhung von 1% (vor Test) auf 7,5 % nach Test – nicht intuitiv bei Testsicherheit von 80%
- Befinden sich in einem Raum mindestens 23 Personen, dann ist die Chance, dass zwei oder mehr dieser Personen am gleichen Tag (ohne Beachtung des Jahrganges) Geburtstag haben, größer als 50 %
- Oder: Simpsons Paradox
- <https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>
- <https://blog.codinghorror.com/an-initiate-of-the-bayesian-conspiracy/>

Vorwissen



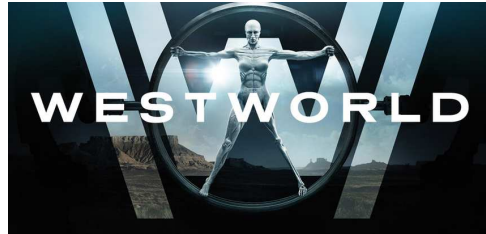
Machine Learning



43

- Definition ML, Definition Data Mining + KDD
- Coursera-Kurs Machine Learning von Andrew Ng
- Neuristoren
- Deep Learning
- Cartoon of the biological brain - a vastly oversimplified version of something we don't even understand
- Recurrent + LSTM: <https://deeplearning4j.org/lstm>
- GPU haben Vorteile ggü. CPU. Nvidia
- Titanic, Iris üben
- Kaggle Hackathons

AI



- Strong vs. Weak AI
- Trotz Fortschritten: Media Hype
- Westworld:
- Intrigen, Gewalt, Sex – mehr als GoT oder House of Cards
- Richtig, Falsch, Gut, Böse, Freier Wille, Sklaverei, Leiden, Emergenz von Bewusstsein durch Komplexität, Erfahrungen
- Bill Gates, Elon Musk, Stephen Hawking u.a. warnen vor AI
- Buch von Nick Bostrom: Superintelligence: Paths, Dangers, Strategies
- <https://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine>
- Voraussage Siri, Alexa, Watson oder Cortana bestehen den Turing Test binnen 5 Jahren
- Man hört wenig von Google Now
- Huawei will No1 werden
- Samsung: VivLabs/ Bixby im April 2017
- Ai in Computerspielen tuned down
- Schach, Go – ok

- Aber Spiele „mit Unsicherheit:“ Poker, Stratego

AI: Bewusstsein

[Detroit: Become Human – Trailer Bewusstsein](#)

190 sek

46

- Computerspiele + Serien/ Filme (Westworld, BladeRunner) + Realität + AI bieten noch ein eigenes, besonderes Spannungsfeld
- Hyperlink: https://www.youtube.com/watch?v=PeIrr__9qx8&t=47s

AI: Entscheidungen

[Detroit: Become Human – Trailer Entscheidungen](https://www.youtube.com/watch?v=QD1pbWCJcKQ)

240 sek

47

- Hyperlink: <https://www.youtube.com/watch?v=QD1pbWCJcKQ>

AI + Robotics: Baxter

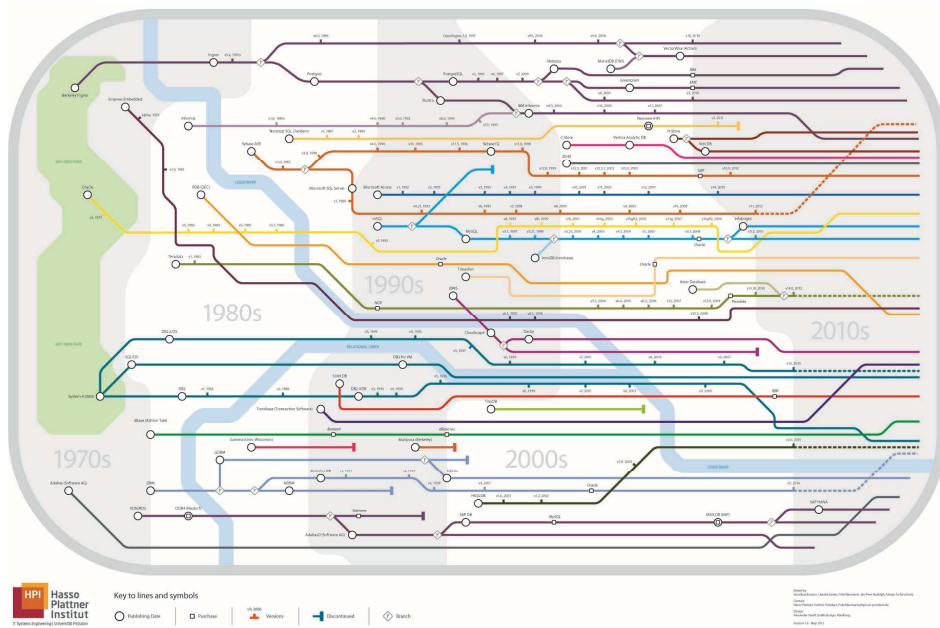
[TED - Rodney Brooks - Why we will rely on robots](#)

600 sek

48

- Hyperlink: <https://www.youtube.com/watch?v=nA-J0510Pxs>
- Ungleich ABB- oder KUKA: Roboter kann direkt neben Menschen eingesetzt werden, lernfähig
- Inzwischen Nachfolger: Sawyer
- ...on the shoulders of steel

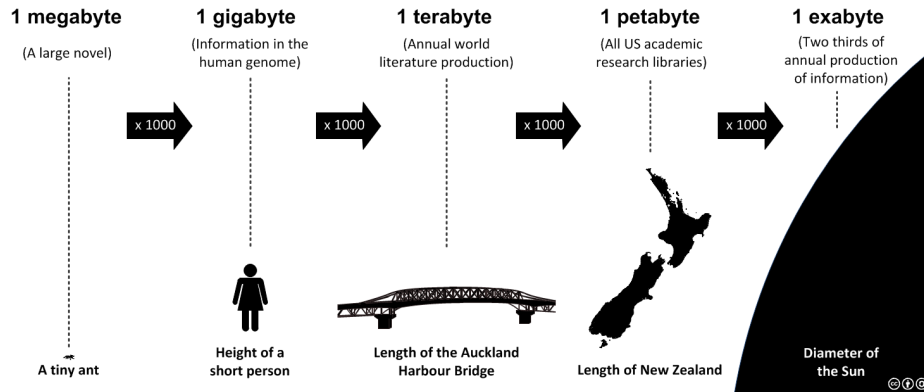
Stammbaum



- Genealogie + Entwicklungen Datenbanken
- Grund für viele „Metro-maps“: A) Übersichtlichkeit, B) manchmal auch Hommage an Harry Beck – siehe Visualisierungsvortrag

Datengröße veranschaulicht

understanding the data deluge: comparison of scale with physical objects

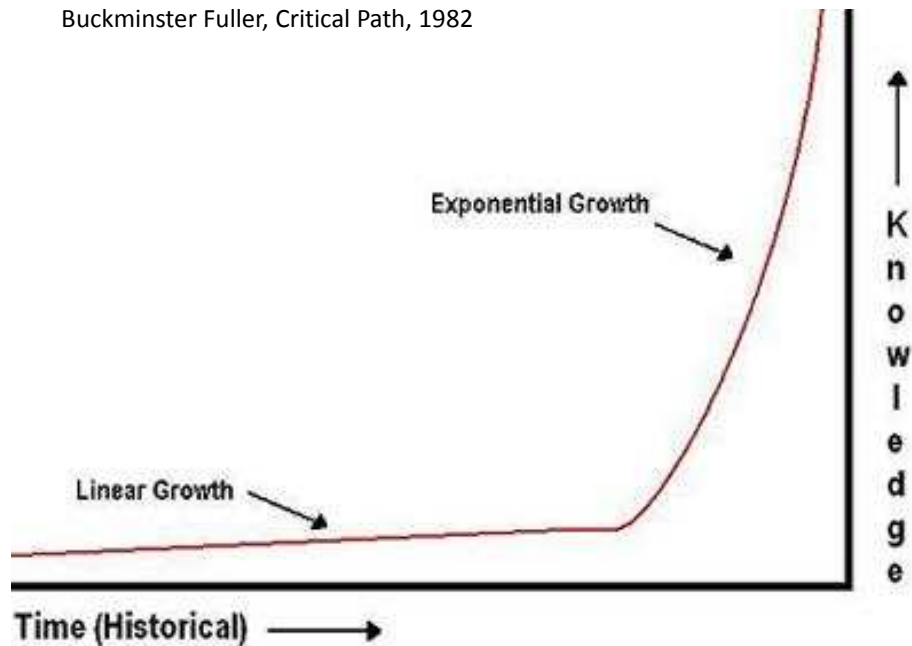


50

- Daten vs. Information vs. Wissen...
- + Moores Law

Wissensgenerierung

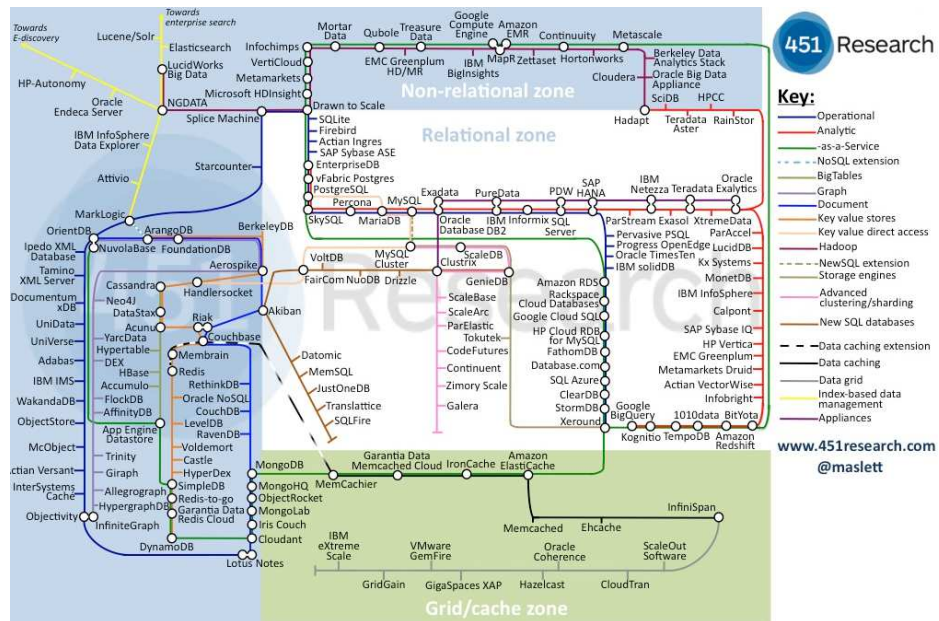
Buckminster Fuller, Critical Path, 1982



51

- Bewusst ohne Skalenangaben.
- 1982: alle 18 Monate, Heute: alle 12 Monate. Mit IoT: mglw. alle 12 h (IBM)
- ...aber...Daten vs. Information vs. Wissen vs. Weisheit
- Learning is not the accumulation of knowledge. Learning is movement from moment to moment. Jiddu Krishnamurti
- „Halbwertszeit des Wissens“
- 1996: From Data Mining to Knowledge Discover in Databases, u.a. Gregory Piatetsky-Shapiro von KDnuggets
- Moore's Law, Parkinson's Law of Data
- Moore: Quanteneffekte – aber andere Computerarchitektur, z.B. HP's "The Machine"
- <http://www.industrytap.com/knowledge-doubling-every-12-months-soon-to-be-every-12-hours/3950>
- Siehe auch: Warum jetzt?

Metro 2013: Datenbanken

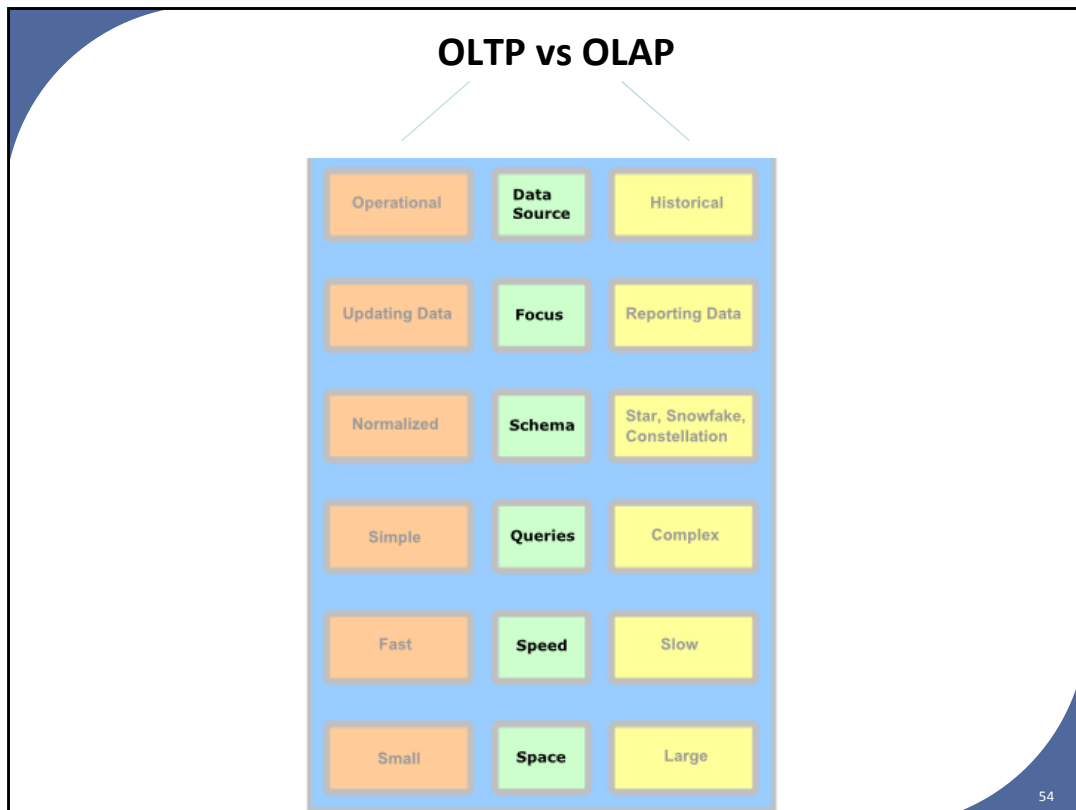


Relational vs. NoSQL

Rang			DBMS	Datenbankmodell	Pu
Feb 2017	Jan 2017	Feb 2016			
1.	1.	1.	Oracle +	Relational DBMS	1403,83
2.	2.	2.	MySQL +	Relational DBMS	1380,30
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1203,45
4.	↑ 5.	↑ 5.	PostgreSQL	Relational DBMS	353,68
5.	↓ 4.	↓ 4.	MongoDB +	Document Store	335,50
6.	6.	6.	DB2	Relational DBMS	187,90
7.	7.	↑ 8.	Cassandra +	Wide Column Store	134,38
8.	8.	↓ 7.	Microsoft Access	Relational DBMS	133,39
9.	↑ 10.	9.	SQLite	Relational DBMS	115,31
10.	↓ 9.	10.	Redis +	Key-Value Store	114,03

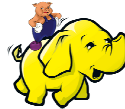
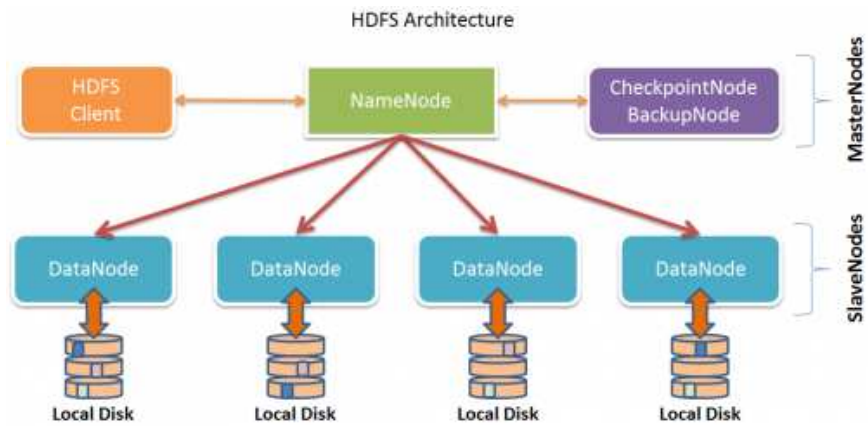
53

- <http://db-engines.com/de/ranking>
- 317 Datenbanken



- Geschwindigkeit:
 - A) DRAM (In Memory-DB: SAP Hana)
 - B) NVDIMMS, Memristors
 - C) Flashdrives, Solid State Drives

Hadoop



- <http://www.datasciencecentral.com/profiles/blogs/what-is-hadoop-great-infographics-explains-how-it-works>

Der lange Weg

1. Fundamentals

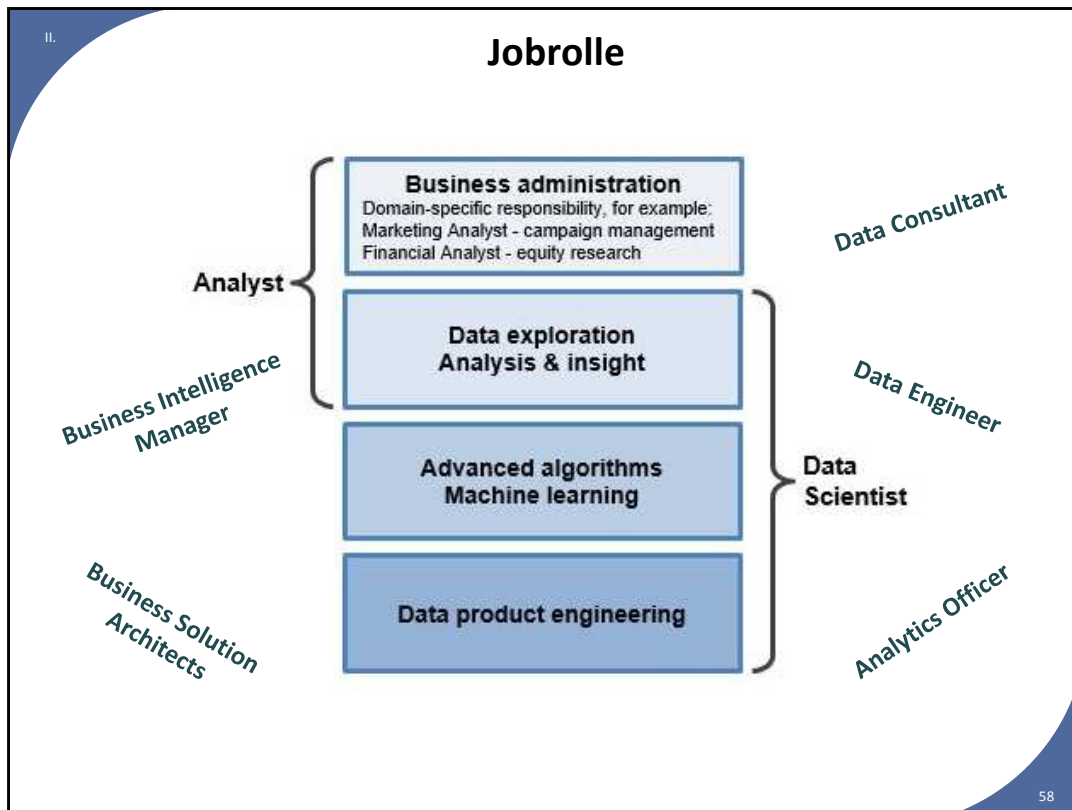
- Matrices & Linear Algebra Fundamentals
- Hash Functions, Binary Tree, $O(n)$
- Relational Algebra, DB Basics
- Inner, Outer, Cross, Theta Join
- CAP Theorem
- Tabular Data
- Data Frames & Series
- Sharding
- OLAP
- Multidimensional Data Model
- ETL
- Reporting Vs BI Vs Analytics
- JSON & XML
- NoSQL

56

- Grundlagen
- Lernen, wovon man nicht mal weiß, das man es nicht weiß – man muss also erstmal lernen, was man alles noch lernen muss
- „unknown unknowns“ Donald Rumsfeld, Nassim Nicholas Taleb
- CAP: Consistency – availability (Verfügbarkeit/Antwortzeit) – partition tolerance (Ausfalltoleranz)
- Metrokarte: Harry Beck 1933

Zusammensetzung		
Big Data	Data Discovery	Data Science
<ul style="list-style-type: none"> • 3-6 V's: Volume, Velocity, Variety (Validity, Veracity, Value) • Wertsteigernd • Schwer + Teuer zu implementieren • Mitarbeiter teuer • Parallelisierung, Hadoop, HDFS, Map Reduce... 	<ul style="list-style-type: none"> • Einfach nutzbar • Schnell, Flexibel • Storytelling + Grafiken • Oberflächlich • Geringe Komplexität • Gefahr: Selbsttäuschung • Self-Service-Tools • Exploration: Herding Cats 	<ul style="list-style-type: none"> • Komplexe Analysen • Viele Werkzeuge • Intelligente Algorithmen • Schwierig, aufwändig • Komplex, oft langsam • Enger Fokus der Analyse • Mitarbeiter fehlen • Industriekenntnisse! • Programmieren! • Statistik!

- Validity Sicherstellung der Datenqualität
- Veracity: Wahrhaftigkeit und Glaubwürdigkeit von Daten
- Value unternehmerischer Mehrwert. Viele Unternehmen haben mittlerweile eigene Datenplattformen aufgebaut, -pools gefüllt und viel Geld in Infrastruktur investiert. Nun gilt es, daraus auch Business Value zu generieren
- Data Jiu-Jitsu: Big Data in Datenprodukte umwandeln, die direkten Unternehmenswert generieren
- Reporting: Herding Cows



- Vorsicht vor Analysis Paralysis
- BI: eher Kennzahlen, Rückwärts, IT-Abteilung
- DS: Erkenntnisse + Entscheidungen, Vorwärts, C-Level/Strategie
- BI-ler (ohne weiteres Training) in Data Science wäre Cargo-Cult-Science
- Tätigkeiten eines Data Scientists:
 - Fragen stellen, optimieren
 - Hypothesen testen
 - Daten aufbereiten/ zähmen
 - Daten visualisieren und erkunden
 - Modellieren
 - Zusammenhänge verstehen
 - Maschinen lehren
 - Erkenntnisse sammeln, Entscheidungen treffen, Datenprodukte erstellen

Aus Daten lernen – mit Statistik

Statistical vs. Machine Learning Vs. Data Mining

Supervised vs. Unsupervised Learning

Induction vs. Deduction

Sampling & Confidence Intervals

Probability & Distribution

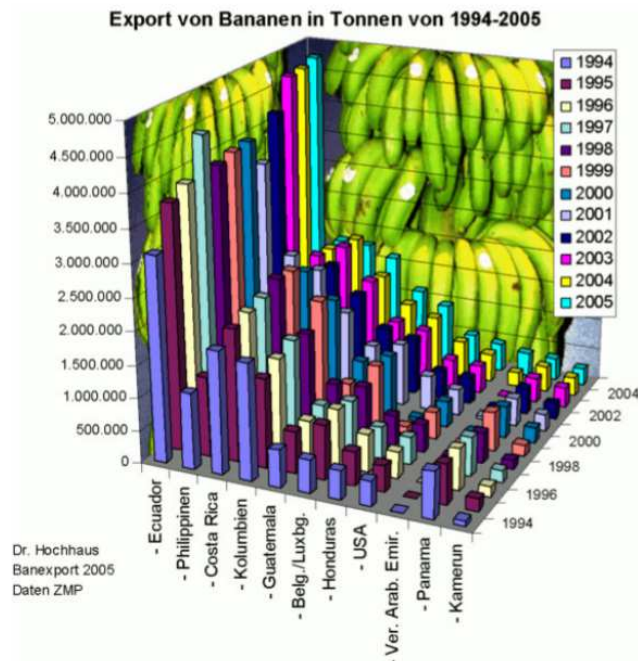
Deviation & Variance

Correlation vs. Causation

Causation & Prediction

- Gesetz der großen Zahl:
Je häufiger man ein Zufallsexperiment macht, desto näher kommen die relativen Häufigkeiten seiner Ergebnisse ihren echten Wahrscheinlichkeiten und Fehler mitteln sich heraus. Das ist verteilungsunabhängig.
- Zentraler Grenzwertsatz:
Je größer eine Stichprobe wird, desto mehr nähert sich ihre Häufigkeitsverteilung der Normalverteilung an.
- Das eine sagt etwas über die generelle Möglichkeit, dass man aus Stichproben Aussagen über Wahrscheinlichkeiten (Parameter) machen kann. Das andere sagt etwas über die Entwicklung der Verteilungsfunktion und stellt die Besonderheit der (Standard-)Normalverteilung her.
- Carlos Somohana, Founder Data Science London

Visualisierung aus der Hölle



- Probleme von Innen
- Sondern einfach schlechtes Handwerk
- DatenVisualisierungen aus der Hölle
- Export von Bananen in Tonnen von 1994-2005
- Überlappende Daten in 3D, „schlechte“ Überschrift, Diagramm ist zu überfüllt, mit ablenkenden Bildern im Hintergrund, miserable Farbgestaltung, schlechte Skala
- Ich weiß nicht, ob es ein gezielt erstelltes Beispiel für schlechte Datenvisualisierung ist
- Man findet auch andere Graphen von Dr. Hochhaus ZMP, und bei Googlesuche sind schlechte Datenvisualisierungen häufiger als Gute
- 3D macht generell nur seeehr selten Sinn. Bei Kuchendiagrammwürde 3D Ansicht die Größe von Teilstücken verzerren – durch Perspektive
- Damit auch schon das Nächste Beispiel – ein Kuchendiagramm
- Liniendiagramm mit x Zeit und y Höhe der Exporte erstellen, je Land andersfarbige, manche Länder raus/ Gruppierungen

Namen und Logos

Quid®



Recorded Future



 Palantir



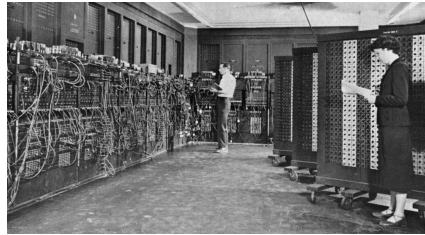
61

- HAL – je einen Buchstaben weiter...
- Winzigweich – Microcomputer Software, ursprünglich mit Bindestrich

Entwicklung



Dryden Flight Research Center E49-0053 Photographed 10/49
Early "computers" at work. NASA photo

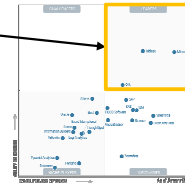


- Frühere Definition Computer: Menschen, die für ihren Lebensunterhalt Berechnungen erstellen
- Hidden Figures – Unerkannte Heldinnen Februar 2017
- Alan Turing Institute is the United Kingdom national institute for the data sciences, founded in 2015
- Aber: Bombe war zwar Rechner, aber kein programmierbarer Computer
- Erster "wirklicher" elektronischer Computer war ENIAC (Bild)
- Charles Babbage, Ada Lovelace
- Grace Hopper, 1947, Bug
- Francis Bacon Chiffre
- Trailer Imitation Game: <https://www.youtube.com/watch?v=VxvY4rI15sM>
- Polnische Vorarbeit: Marian Rejewski
- In Deutschland: Konrad Zuse (aber wenig Einfluß auf weitere Entwicklung)

Software für Analyse und Visualisierungen

Jan. 2017: BI + Analytics Leader nach Gartner

- Tableau
- Microsoft (PowerBI)
- Qlik



Wohlgemerkt:

- Nicht „Advanced Analytics“
- Nicht „Infografik-Programme“
- Nicht „Viz-Packages“ für Sprachen

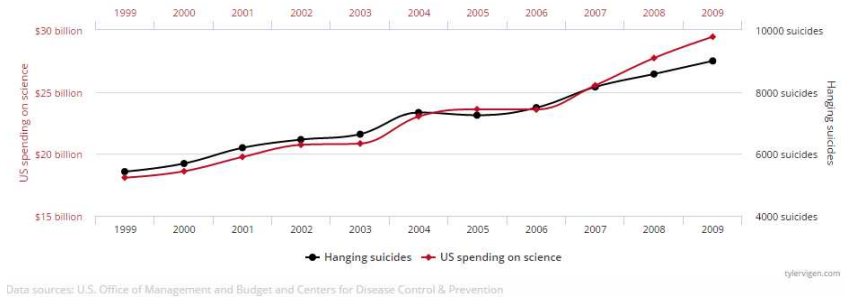
- Gartner Quadrant: oben rechts Leaders + ability to execute
- bewusst so klein
- Qlik ist richtig geschrieben. View + QlikSense – ich werde nicht näher drauf eingehen
- <https://www.gartner.com/doc/reprints?ct=160204&id=1-2XXET8P>
- Freakalytics nach Affordability Individuell auch diese drei recht gut
- RapidMiner, Knime, Alteryx oder Spotfire
- Easel.ly, Piktochart, Infogr.am
- Ggplot2 oder Shiny in R, Plot.ly in Python oder julia, D3 für Java.
- Für Bedienung von Tableau gibt es gute Videotutorials
- Weitere Gesamtpakete für Big Data: Cluvio, Domo, Periscope Data, Sisense...
- Interaktive Visualisierungsoptionen in Programmiersprachen eingebaut z.B. Wolfram Language/ Mathematica...

Advanced Analytics Platforms



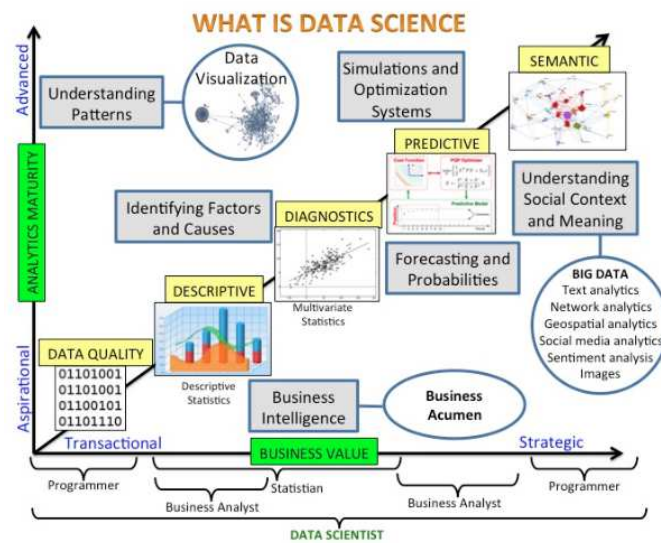
- Gartner 2016 for Advanced Analytics Platforms
- BI + Analytics Leader
- Gartner Quadrant: oben rechts Leaders + ability to execute

Zufallszusammenhänge II



US Ausgaben für Wissenschaft, Raumfahrt und Technologie
und
Anzahl der Selbstmorde durch Erhängen, Erdrosseln und Ersticken
Korrelationswert: 99,79%

Lineare Definition

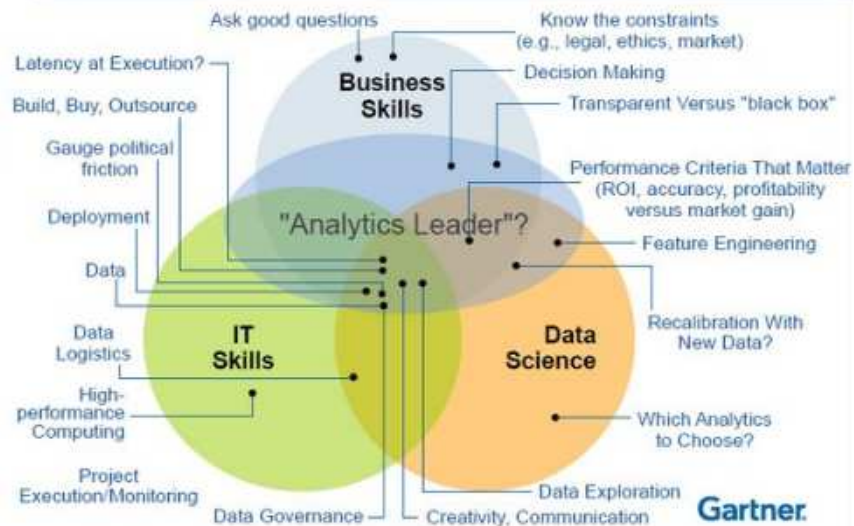


66

- <http://www.datasciencecentral.com/profiles/blogs/data-science-summarized-in-one-picture>
- <http://www.datasciencecentral.com/profiles/blogs/20-articles-about-core-data-science>

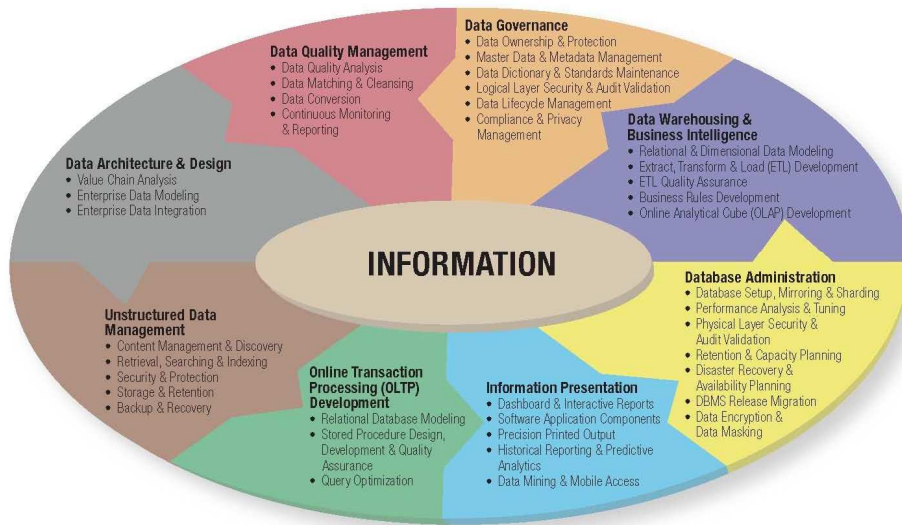
Breites Aufgabengebiet

Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...

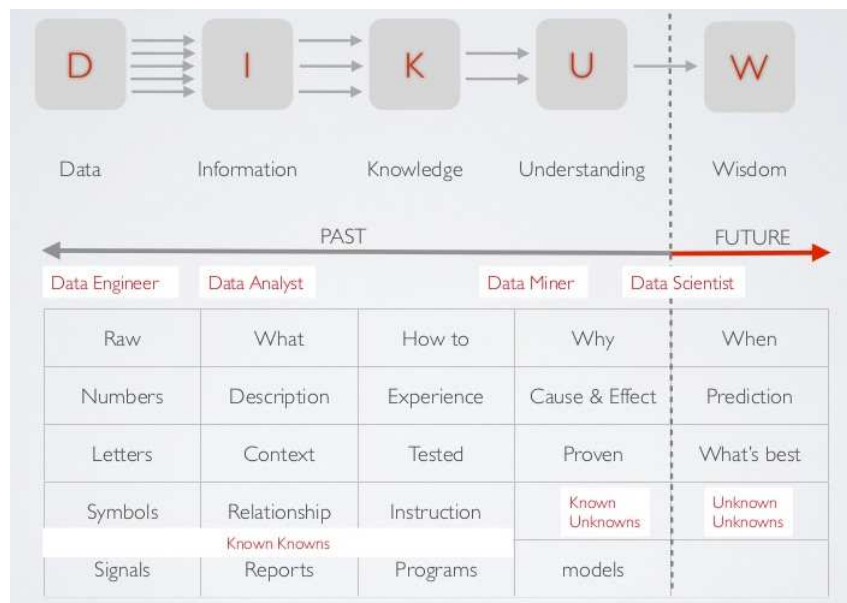


- <http://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning>
- Process Modeling and Control - Creating a neural network model for a physical plant then using that model to determine the best control settings for the plant.
- Machine Diagnostics - Detect when a machine has failed so that the system can automatically shut down the machine when this occurs.
- Portfolio Management - Allocate the assets in a portfolio in a way that maximizes return and minimizes risk.
- Target Recognition - Military application which uses video and/or infrared image data to determine if an enemy target is present.
- Medical Diagnosis - Assisting doctors with their diagnosis by analyzing the reported symptoms and/or image data such as MRIs or X-rays.
- Credit Rating - Automatically assigning a company's or individuals credit rating based on their financial condition.
- Targeted Marketing - Finding the set of demographics which have the highest response rate for a particular marketing campaign.
- Voice Recognition - Transcribing spoken words into ASCII text.
- Financial Forecasting - Using the historical data of a security to predict the future movement of that security.
- Quality Control - Attaching a camera or sensor to the end of a production process to automatically inspect for defects.
- Intelligent Searching - An internet search engine that provides the most relevant content and banner ads based on the users' past behavior.
- Fraud Detection - Detect fraudulent credit card transactions and automatically decline the charge.

Themen-Blumenstrauß



DIKUW



- Zeigt auch Business Intelligence vs. Data Science
- Carlos Somohana, Founder Data Science London

T

More Data Beats Better Algorithms, Omar Tawakoi @BlueKai

Better Algorithms Beat More Data, Mark Torrance @RocketFuel

More Data or Better Models, Xavier Armitrain @Netflix

On Chomsky & 2 Cultures of Statistical Learning, Peter Norvig @Google

Specialist Knowledge is Useless & Unhelpful, Jeremy Howard @Kaggle

70

- Carlos Somohana, Founder Data Science London