Bielefeld, 30. Mai 2010

Data Mining

Das einundzwanzigste Jahrhundert im Datenkorsett

Jun.-Prof. Alexander Markowetz

Rheinische Friedrich-Wilhelms-Universität Bonn

Themen

- Teil 1: Technik
  - Data Mining
  - Suchmaschinen
- Pause
- Teil 2: Gesellschaft
  - Datenschutz als Sozialer Umweltschutz
  - Automatisierte Entscheidungen
  - Bei-Spiele

Data Mining / weitester Sinn

- Versuch Daten eine Bedeutung abzugewinnen
- Erkenne Trends und Muster
- Auffällig Daten (sog. Outlier)
- Vorhersage zukünftiger Events
- Aussagen über zukünftige Daten

Data Mining / weitester Sinn

- Präzise definierte Methoden
  - Classification
  - Regression
  - Clustering
  - k-NN
  - ...

- Gesamt Prozess nur unscharf def.
- Welche Methode?
- Auf welchen Daten?
- Problem wie modelliert?

- 21 Jh. Kaffeesatzlesen

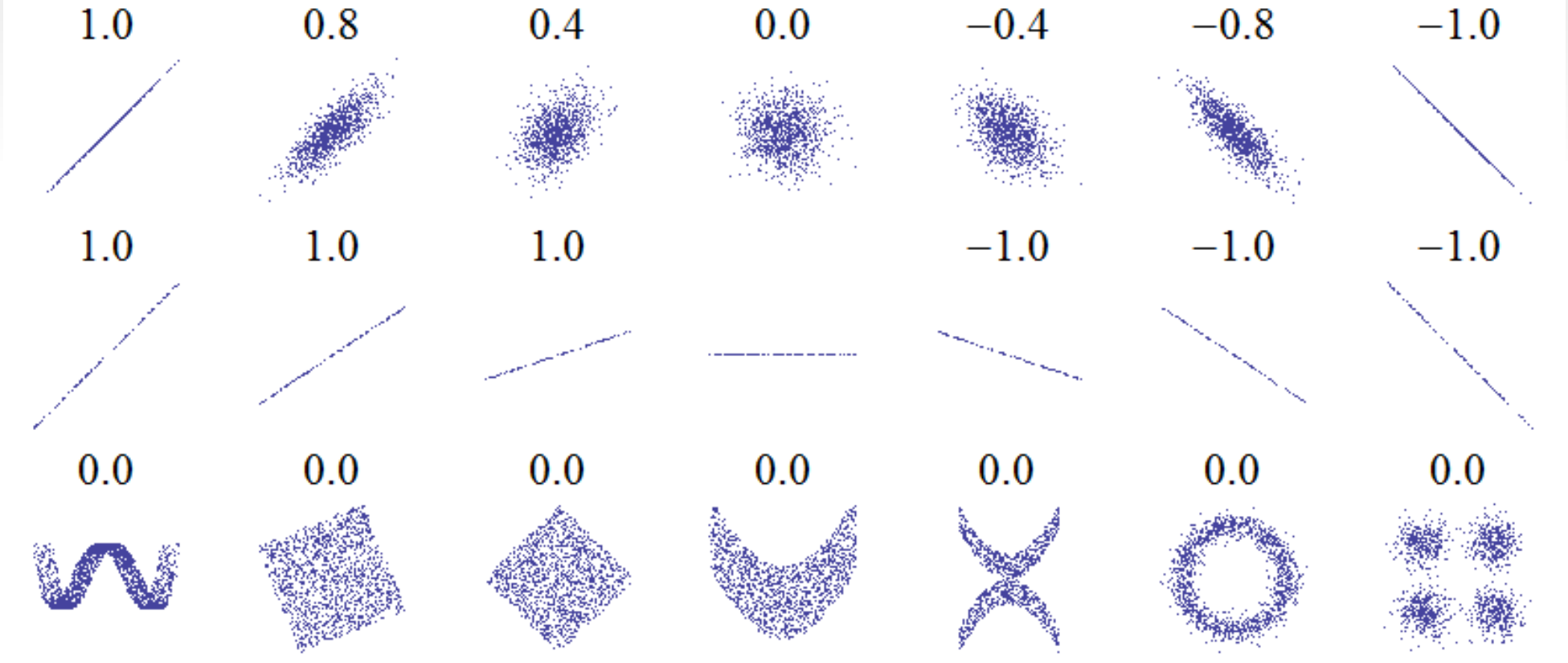
Selbst Mathe Hilft

- Methoden der Statistik
- Finde Trends und Zusammenhänge

- Korrelation
- Verteilungen

Korrelation

- Hängen Daten miteinander zusammen?



Kausale Zusammenhänge

- Korrelation zeigt das Dinge zusammenhängen
- **Aber nicht notwendigerweise direkt**
- Vielleicht gibt es auch gemeinsame Ursachen

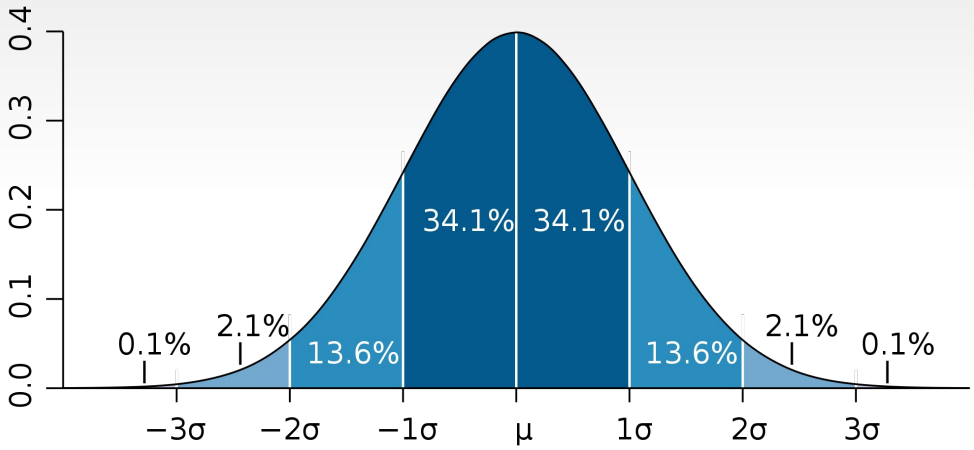
Falsche Zusammenhänge

- Größere Leute verdienen mehr
- Je mehr Lärm im Haus, desto dümmer die Kinder
- Rauchen schadet Ihrer Intelligenz
- Kreative haben mehr Sex
- Glückliche Menschen sind gesünder
- Senkung der Arbeitslosigkeit erfordert starkes Wirtschaftswachstum

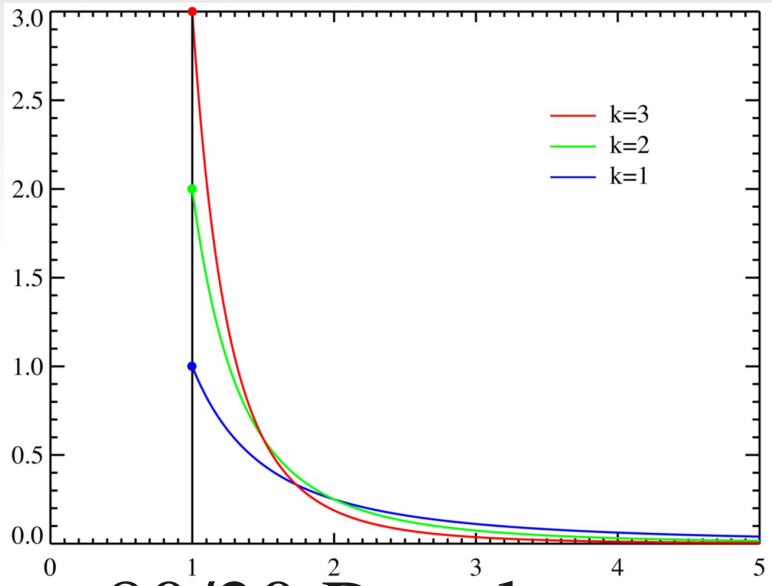
(aus versch. Tageszeitungen, laut Wikipedia)

Wahrscheinl.-Verteilungen

- Normalverteilt



- Pareto Verteilung



- Erwartungswert
- Standard Abweichung
- 80/20 Regel
- ”Power Law” nicht: ”Macht Gesetz”

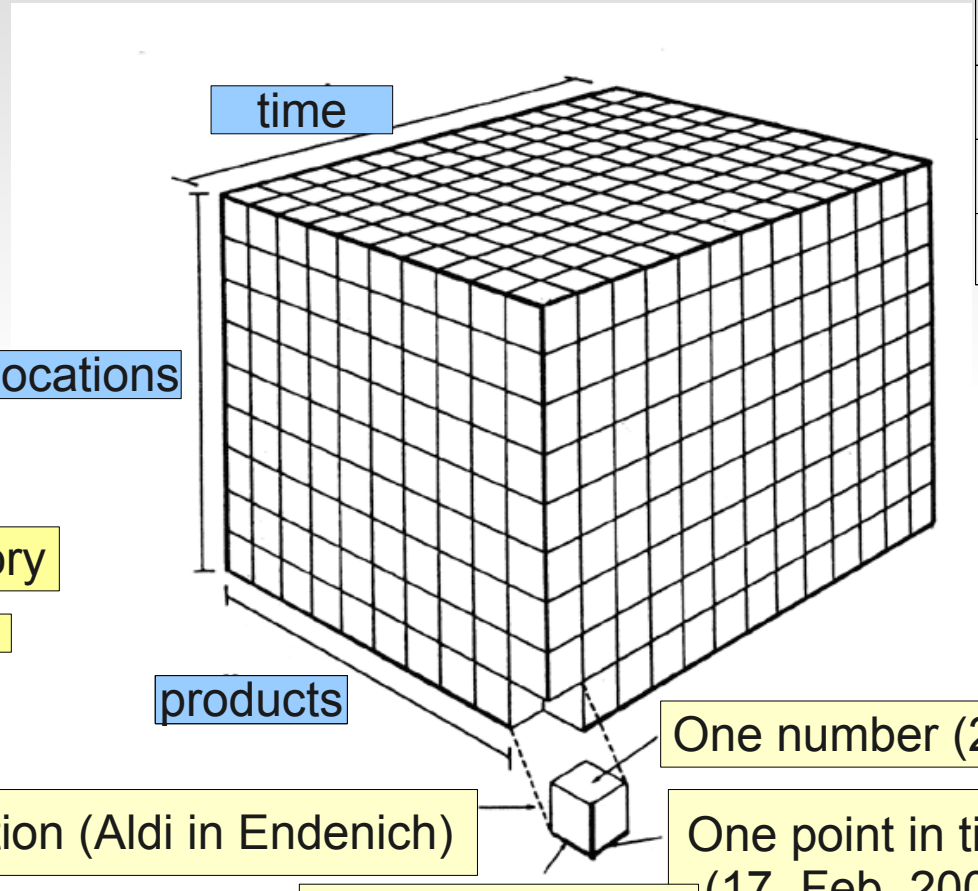
OLTP vs. OLAP vs. KDD

- Datenbanken (OLTP)
  - Transaktionen im Live System
  - Präzise Anfragen
- D-Warehouse (OLAP)
  - Historische Daten zur Analyse
  - Data Cube
  - Interaktiv
- Knowledge Discovery
  - Umfassende Datenanalyse
  - Data Mining
  - Unklarer Ausgang
- Unterschied:
  - Weiss ich was ich will?

Data Warehousing

- Sammelt Daten aus verschiedenen OLTP Datenbanken
- Historische Daten
  - Beinhaltet auch alte Versionen, nicht nur gegenwärtigen Zustand
- Daten müssen transformiert und angepasst werden
  - Data Cleaning and Transformation

OLAP


country
state
region
city
hood
time
year
quarter
month
week
day
locations
product category
product class
product
products
One number (213 €)
One location (Aldi in Endenich)
One point in time (17. Feb. 2007)
One product (Albrechts Kaffee)

OLAP

- Pivot Table
  - Friere Dimension ein: Produkte
  - Wähle: Kaffee

	09-01-13	09-01-14	09-01-15	Sum
Endenich	43€	65€	57€	165€
Kessenich	12€	28€	21€	61€
Beul	32€	42€	12€	86€
Summe	87€	135€	90€	312€

  - Roll-Up / Drill Down
  - Slice / Dice
  - Rotate

Data Mining

- Menge an Grundlegenden Methoden
- Unterscheiden sich im inneren
- Passen unterschiedlich gut fuer spezielle Daten

- Classification
  - Supervised
  - Unsupervised
    - Clustering
- Regression
- Association Rule Mining

Classification

- Gegeben:
  - Eine Menge Datensätze
  - Eine Menge Labels
- Entscheide für jeden Datensatz, welches Label wohl zutrifft
- Supervised: es gibt ein Training und Test-Set
- Unsupervised: keine vorkategorisierten Beispiele

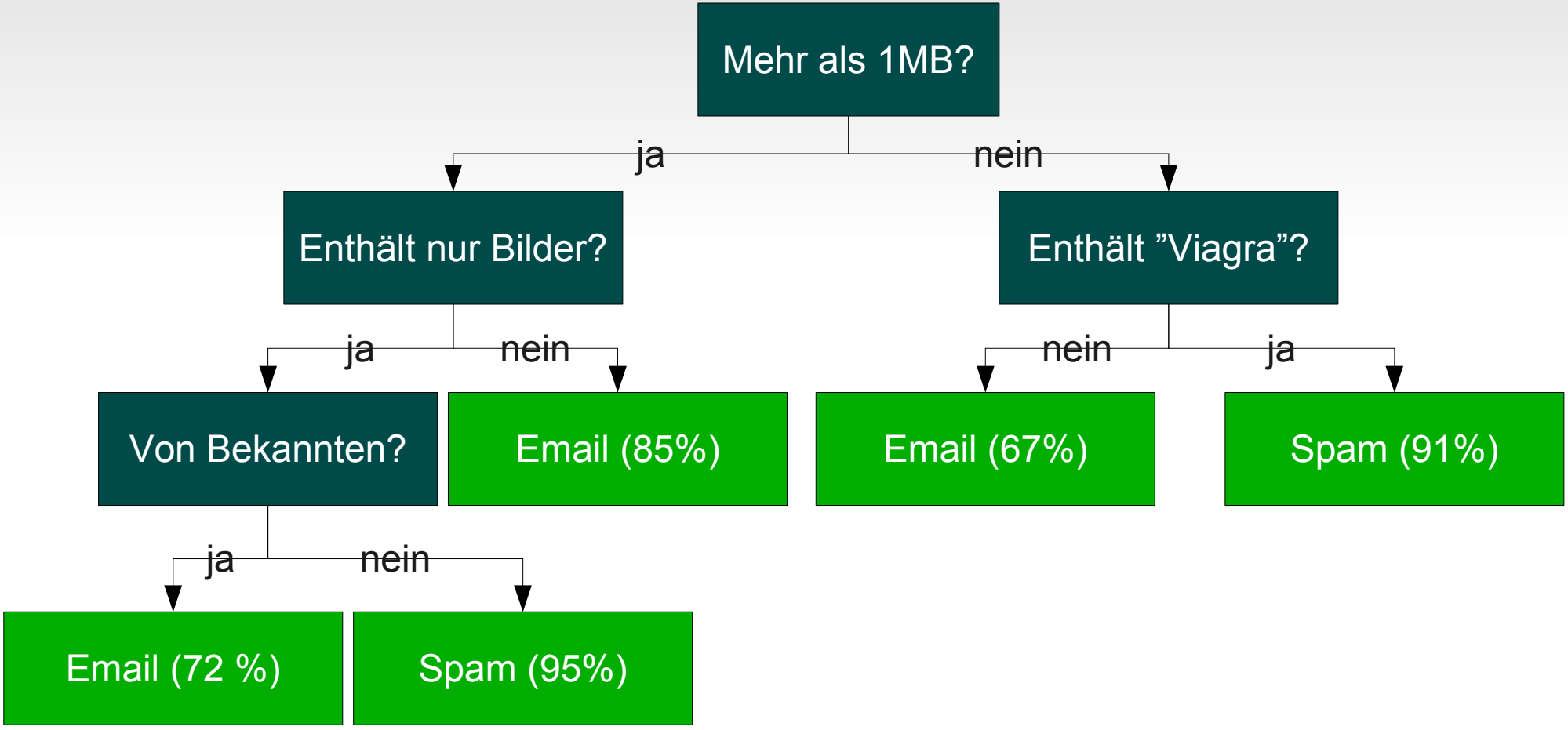
Beispiele für Classification

- Spam
- Kreditwürdig
- Konsumer-Klasse
- Jede Art von Verhaltens-Klassen

Decision Trees

- Simpleste Methode der Klassifikation
- Baumform
  - Wurzel = Anfang
  - Innere Knoten = Fragen
  - Blätter = Klassen
- Kann von Hand gebaut werden
- Oder Ergebnis eines Machine Learning Algorithmus sein

Beispiel Decision Tree



Overfitting

- Regeln werden zu genau trainiert
- Nicht generell genug:
  - Juniorprofessor?
  - Bonn?
  - Informatik?
  - Datenbanken?
  - Dann magst du gerne Pilzesuchen....
- Und der nächste Juniorprofessor der kommt?

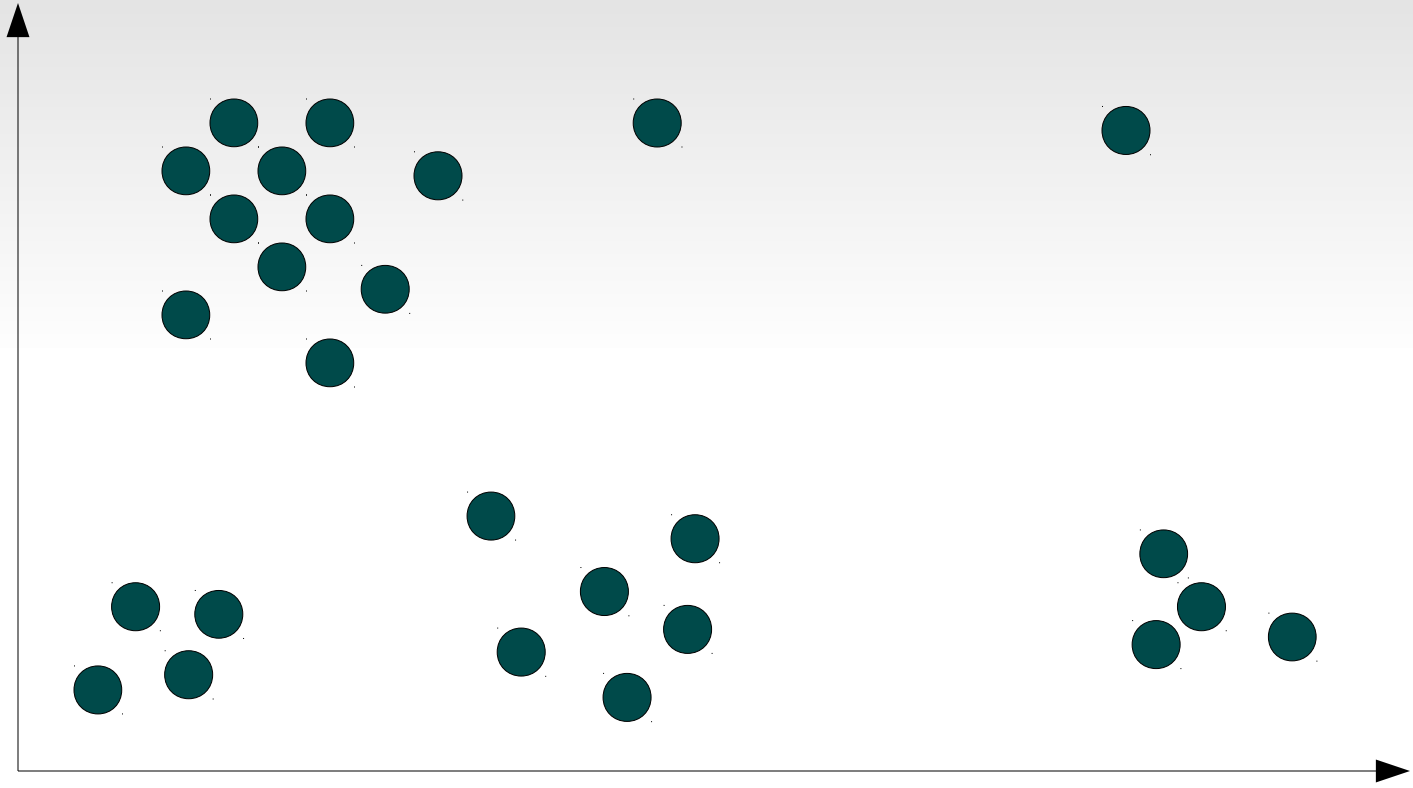
Supervised Learning Algorithmen

- Entscheidungsbäume und Regressionsbäume
- Neuronale Netzwerke
- Support-Vector-Maschine
- Genetische Algorithmen
- Statistische Modellierung z. B. Naive Bayes
- Viele Anwendungen in Biologie, etc.
- Meist als Black Box

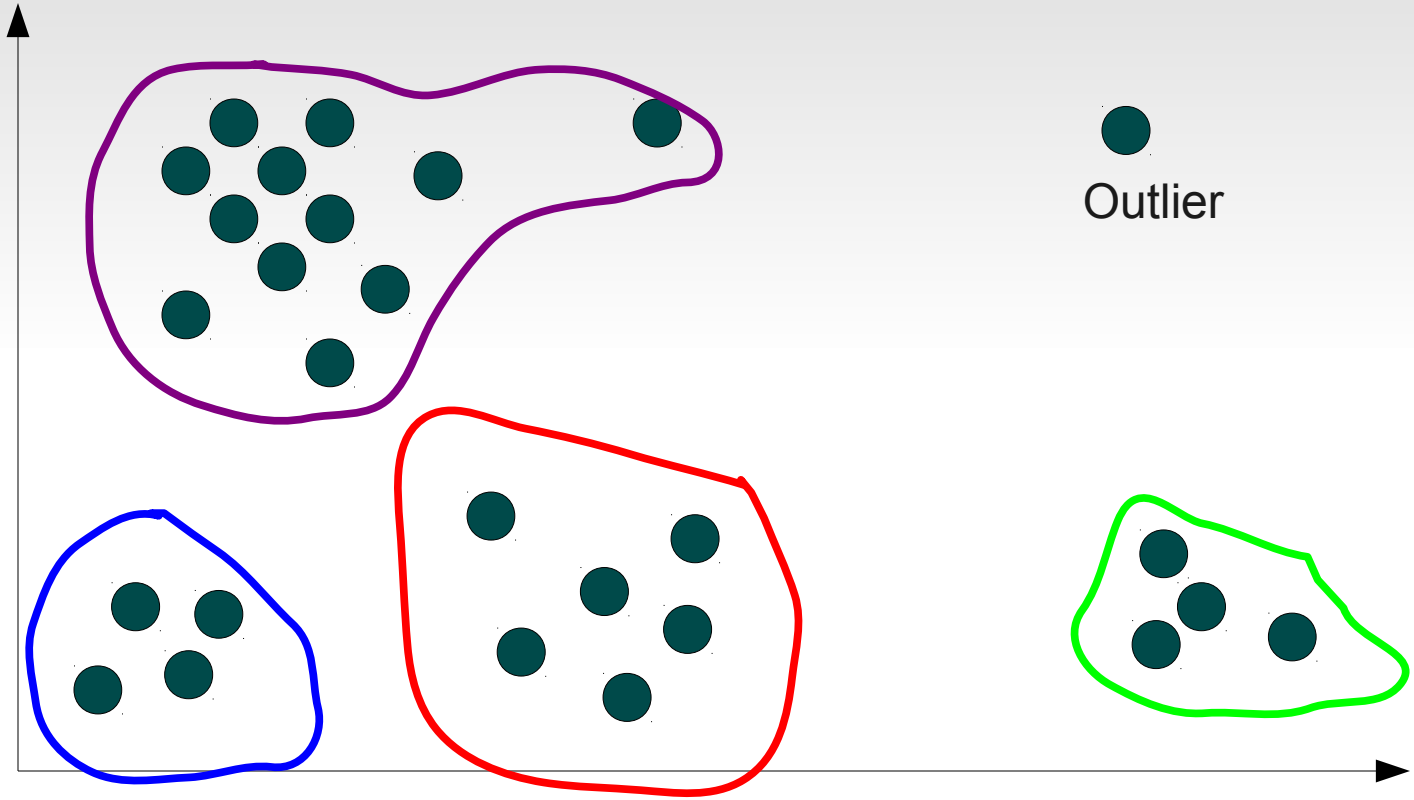
Clustering & Outlier

- Punkte im mehrdimensionalen Raum
  - Nicht notwendigerweise Raum/Zeit
  - Z. Bsp.: Alter, Einkommen, Kreditlinie, Kinder
- Automatisches Aufteilen in sinnvolle Regionen
- Müssen noch interpretiert werden
- Outlier: Daten die zu keinem Cluster passen
  - Können Müll sein,
  - oder sehr interessant

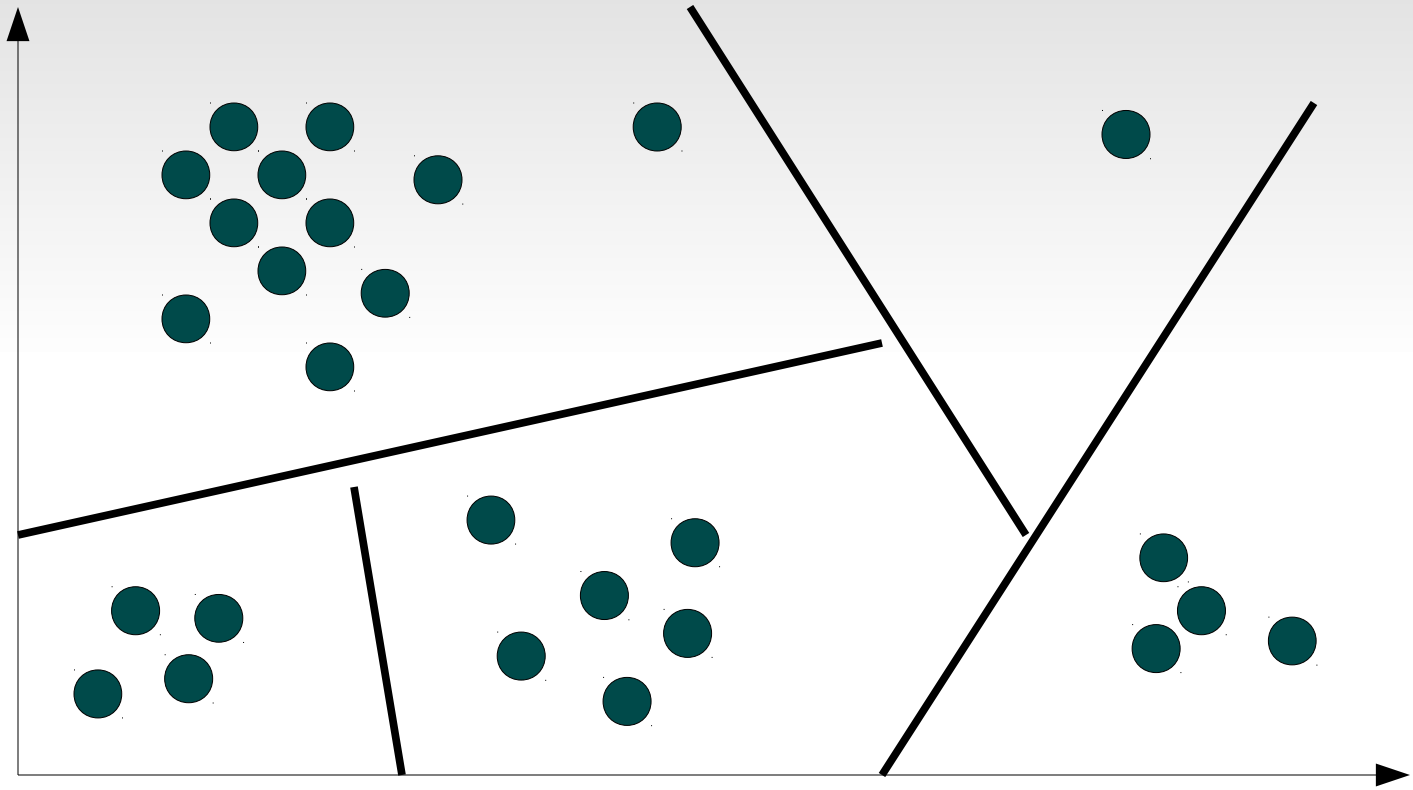
Bsp.: Clustering & Outlier



Bsp.: Clustering & Outlier

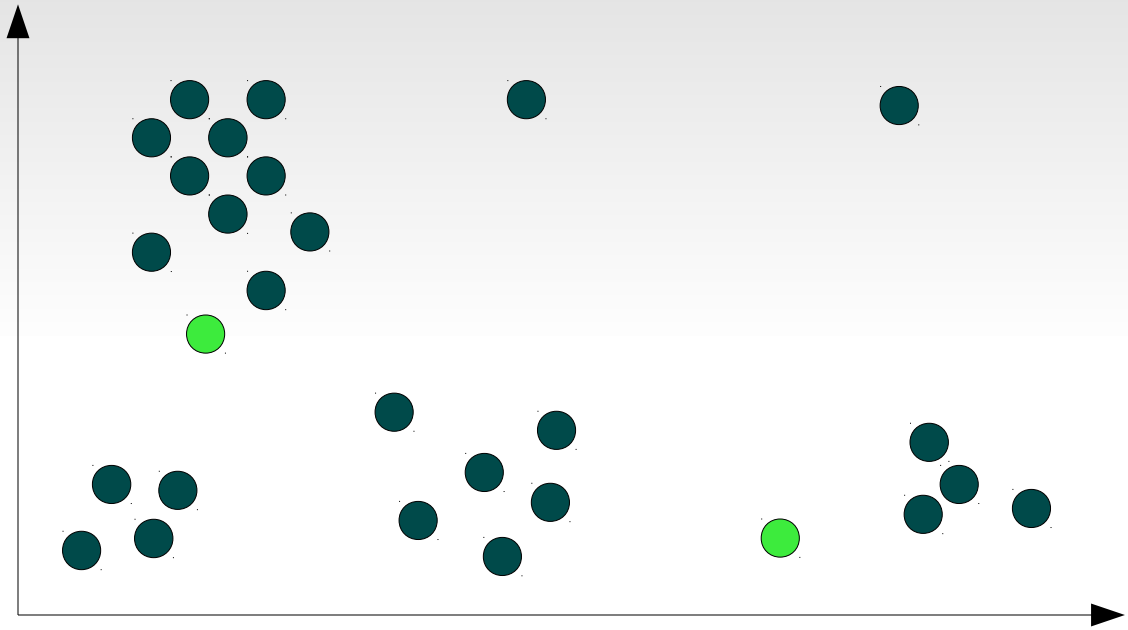


Bsp.: Clustering & Outlier



- Voronoi Diagramm um Cluster
- Neuer Punkt: zu Cluster in dessen Zelle er liegt

K-Nächster-Nachbar



- Keine fixen Cluster
- Neuer Punkt, ähnlich seinen *k*-nächsten Nachbarn
- Nur in niederdimensionalen Räumen

Regression

- Ähnlich Classification
- Bildet aber nicht in diskrete Label ab
- Sondern in einen kontinuierlichen Zahlenraum
  - z. B. Von Alter, Einkommen, Kinder etc.
  - Auf die Wahrscheinlichkeit (in %) an Kehlkopfkrebs zu erkranken

Association Rules

- Market Basked Analysis
- Finde Regeln der Sorte
  - (Brot, Milch) ⇒ Bier
- Die meistens gelten (Confidence)
- Und relativ häufig sind (Support)
- Der Ebay Algorithmus
  - Fragen Sie Pat Robertson

Association Rules

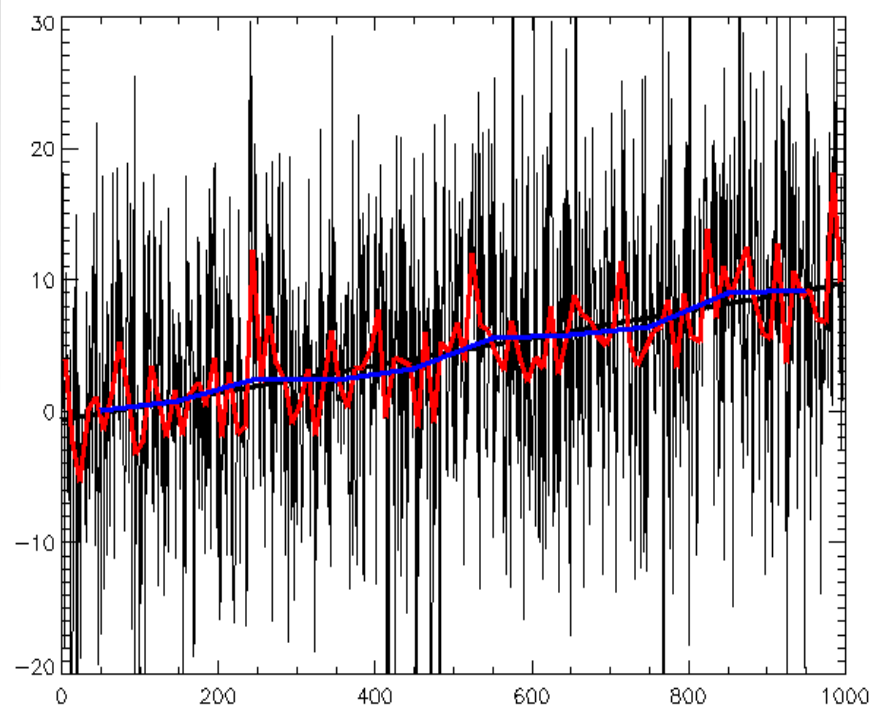
- Gegeben ein Set an Objekten S
- Und eine Menge an Teilmengen x_i
- $x_i \subset S$
- S = alle Produkte im Supermarkt
- Und jedes x_i ist ein voller Einkaufkorb
- Finde Teil-Menge and Produkten s $\in$ S, die regelmaessig zusammen in einem Korb landen

Association Rules

- {a, b, c}
- {a, b, c}
- {a, b, f}
- {a, f}
- {b, c}
- {b, c, d}
- {b, d}
- {c, f}
- {e, f}
- {e, f}

- Support $supp(X)$:
  Anteil der Transaktionen, die X enhält
- Konfidenz: $conf(X \Rightarrow Y) = supp(X \cup Y) / supp(X)$
  Teil aller Trans. die X enthalten, die dann auch Y haben

Zeitreihen



- Untersuche Trends in historischen Daten
- Besonders für Aktienanalyse
- Suche Muster
- Suche ähnliche Serien
- etc.

Social Networks

- Facebook
- Email/Telephon Verkehr
- Graph Struktur

- Wenn ich das und das und das über deine Freunde weiss
- Was weiss ich über dich?

Später...

Textmining

- Versucht Texte
  - Zu klassifizieren
  - Zusammenzufassen
  - Ähnliche Texte zu finden
  - Trends zu finden
  - Etc.

Visual Data Mining

- Graphisches Interface
- Interaktiv
- Ähnlich zum interaktiven OLAP
- Breites Feld
- Sehr Anwendungsorientiert

Fazit Data Mining

- Findet Regelmässigkeiten (Rules)
- Und auffällige Datensätze (Outliers)
- Halb-manueller Prozess
- Ergebnisse müssen interpretiert werden
- Sehr beliebt in Wirtschaft, Banken, Biologie, Soziologie, etc.
- Können aber auch benutzt werden um komplexe Regeln zu automatisieren (via Klassifikation)

And now for something completely different

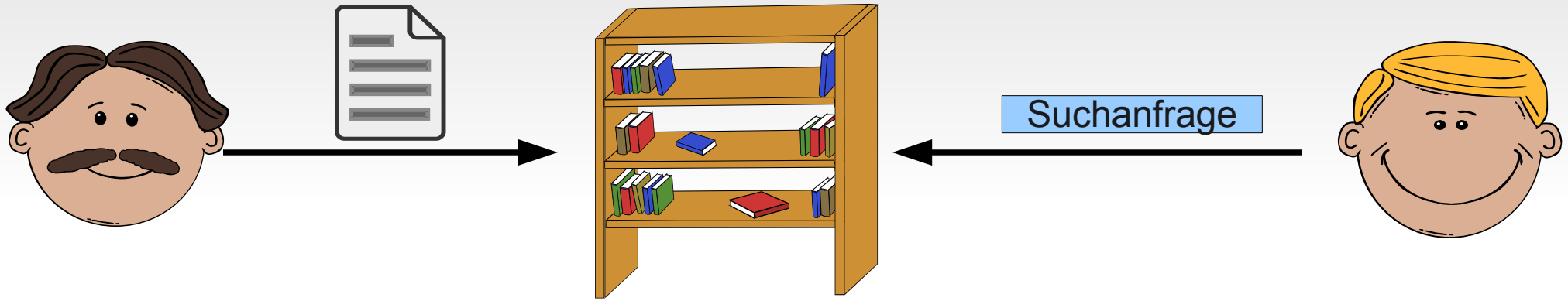
Suchmaschinen

Information Retrieval

- Wie baue ich eine Suchmaschine?
- Gegeben
  - Eine Menge an Dokumenten
    geschrieben von Menschen
  - Einen Benutzer mit Informationsbedarf
- Gebe dem Benutzer das bestmögliche Dokument

Seit 1800 in Bibliotheken
Seit 1960 elektronisch
Seit 1990 Internet

Information Retrieval



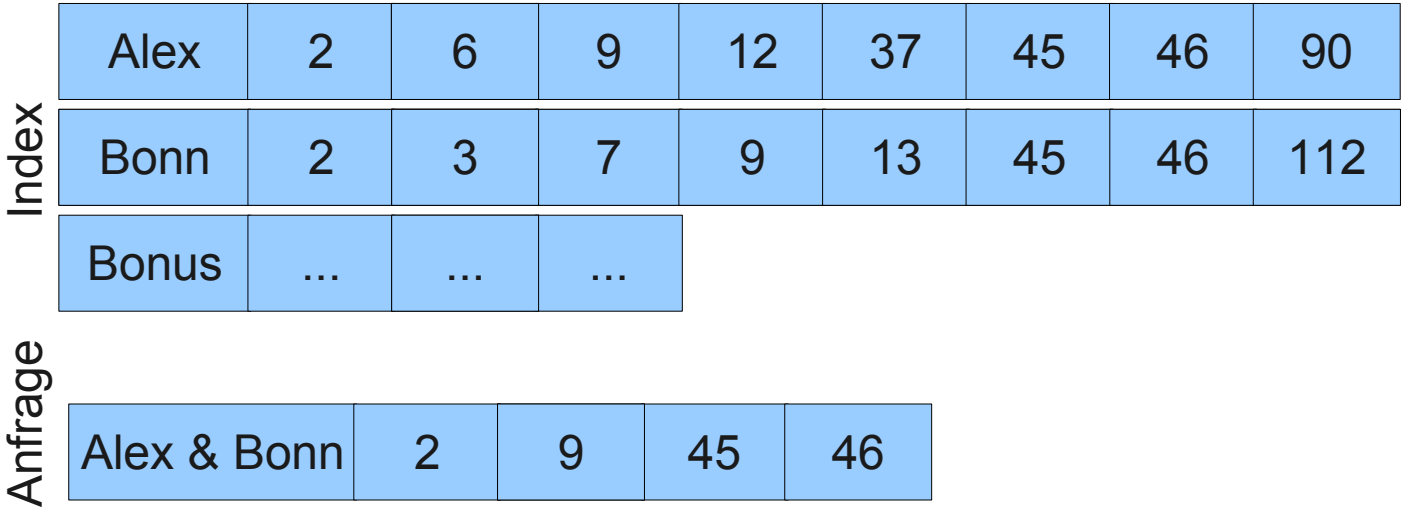
- Was will uns der Autor sagen?
- Was sucht der Benutzer wirklich?
- Modelliere Probleme der Psychologie und Soziologie
- Simple und schnell berechenbar

Einfachstes Modell

- Wenn ein Autor über Mäuse schreibt verwendet er das Wort ”Maus”
- Wenn ein Benutzer etwas über Mäuse sucht verwendet er das Keywort ”Maus”

Inverted Index

- Speichert für jeden Term eine Liste von Dokumenten, die den Term enthalten
- Berechne Queries mit mehrern Keywords durch Schnittmenge von Listen



Ranking

- Das einfache Anfragemodell funktioniert einigermassen gut
- Aber, für die meisten Anfragen gibt es viiiiiel zu viele Dokumente zurück
- Der Benutzer wünscht die besten Dokumente zuerst
- Reihenfolge wichtig!

Term Frequency

- Annahme: Verwendet der Autor ein Wort häufig, so will er wirklich über dieses Thema schreiben
- Schlussfolgerung: Wenn das Suchwort häufiger vorkommt, ist das Dokument wichtiger

- Aber, lange Dokumente müssen bestraft werden (Normalisierung über Dokumentenlänge)

Term Frequency

- Folgende Saetze sind für das Query ”alex, bonn” nach Wichtigkeit gerankt.
- **Alex** wohnt in **Bonn**, gleich bei der Uni **Bonn**
- **Alex** wohnt in **Bonn**, gleich bei dem Uni Campus
- **Alex** wohnt in **Bonn**, nicht weit von dem Kloster und findet es da ganz dufte, obwohl ihm Hessen manchmal fehlt, laberrhabarber.....

Inverse Document Frequency

- Anfrage ”Auto, Schleudersitz”
- Welches Dokument ist wichtiger?
- Dok. A enthält 3 * Auto und 1 * Schleudersitz
- Dok. B enthält 1 * Auto und 3 * Schleudersitz

- Annahme: Seltene Terme sind aussagekräftiger
- Folgerung: Dokument B ist besser, denn
  ”Schleudersitz” taucht in weniger Dokumenten auf

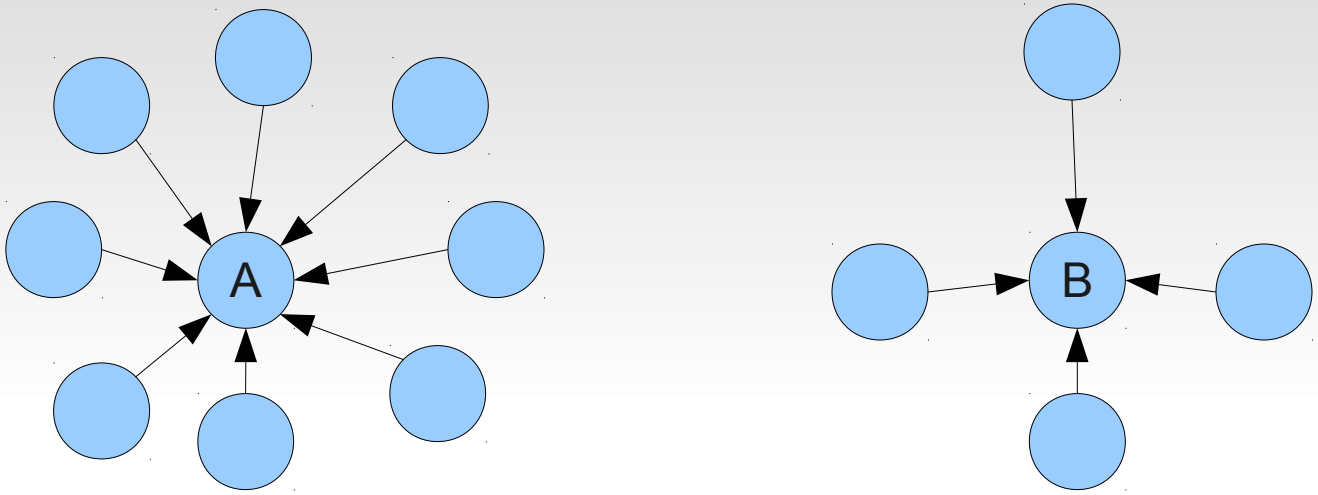
TF-IDF

- i = ID-Nummer des Dokumentes
- t = Term aus der Anfrage (Schleudersitz)
- D = Menge aller Dokumente

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad (\text{tf-idf})_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

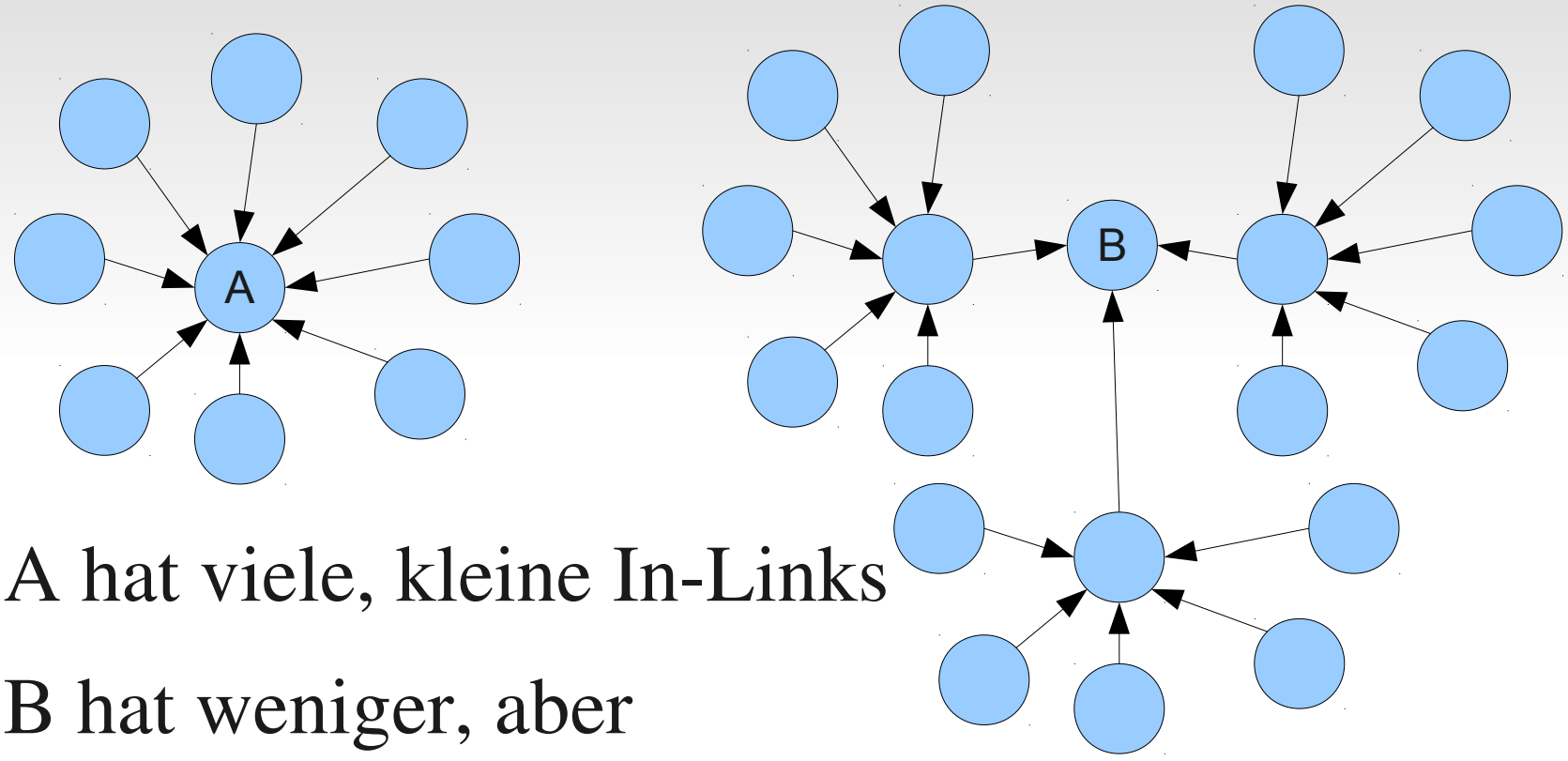
- Problem: Es ist sehr leicht ”falsche” Dokumente zu erzeugen: Web-Spam
  - Billig autos billig autos billig autos billig autos

In-Links



- Aus Bibliothekswissenschaften (Zitateindex)
- Idee: Anzahl der In-Links verweist auf wichtige Webseiten
- Problem: leicht zu verwirren (Linkfarmen)

PageRank



- A hat viele, kleine In-Links
- B hat weniger, aber wichtige In-Links
- Wahrscheinlich ist B wichtiger

Sergey Brin, Larry Page "The Anatomy of a Large-Scale Hypertextual Web Search Engine" 1998

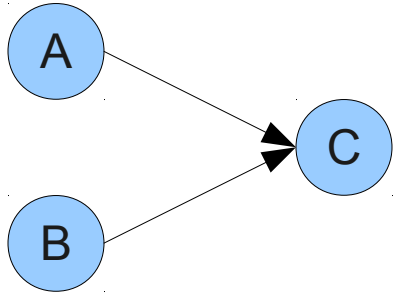
PageRank

- Der ”Google-Algorithmus”
- Rekursive Definition: ein Knoten eines Graphen ist wichtig, wenn seine In-Links wichtig sind
- Findet ”wichtige” Knoten in einem Graphen
- Random-Walk Modell
  - Wahrscheinlichkeit bei einem blinden Graph-Durchlauf auf einem Knoten zu landen
- Simpel und effizient zu berechnen
  - Eigenwerte bestimmen (Matrix-Multiplikation)
- Anwendung in vielen Gebieten (Biologie, etc.)

Noch mehr Links

- Citation
- Wenn A zu B zeigt, sind A und B ähnlich

- Co-Citation (Triangular Closure)
- Wenn A und B zu C zeigen, sind A und B ähnlich



Social Network Analysis

- Sozialer Netzwerke
  - Facebook
  - Email-Verkehr
- Modelliert als Graph
- Citation, Co-Citation
- Pagerank
- Grundlage für ausgeprägtes Data Mining
  - Beispiel: MIT Gaydar....

Suchmaschinen 2000

- Viele Orthogonale Probleme
  - Suchen auf speziellen Dokumenten (Blogs, etc.)
  - Geographische Suche
  - Spam Detection
  - Multimedia Inhalte (Photos, Videos, Musik)

- Das Grosse Problem bleibt: Wie macht man fundamentale Suche besser?

Nutzerverhalten

- Bisheriges Modell unersucht ausschliesslich Dokumente (den Kopf des Autors)
- Idee: Analysiere Verhalten der Benutzer
  - Welche Anfragen hat er gestellt
  - Welche Seiten hat er besucht
- Gespeichert in sogenannten Query Logs
- Abhängig von der konkreten Anwendung

Nutzerverhalten

- Gewinne Informationen **über Dokumente** aus dem Nutzerverhalten
- Gewinne Informationen **über Benutzer** aus dessen Verhalten
  - Vorzüge & Bedürfnisse
- Ähnlich dem Publikumsjoker bei Günther Jauch

Klassifizieren von Dokumente

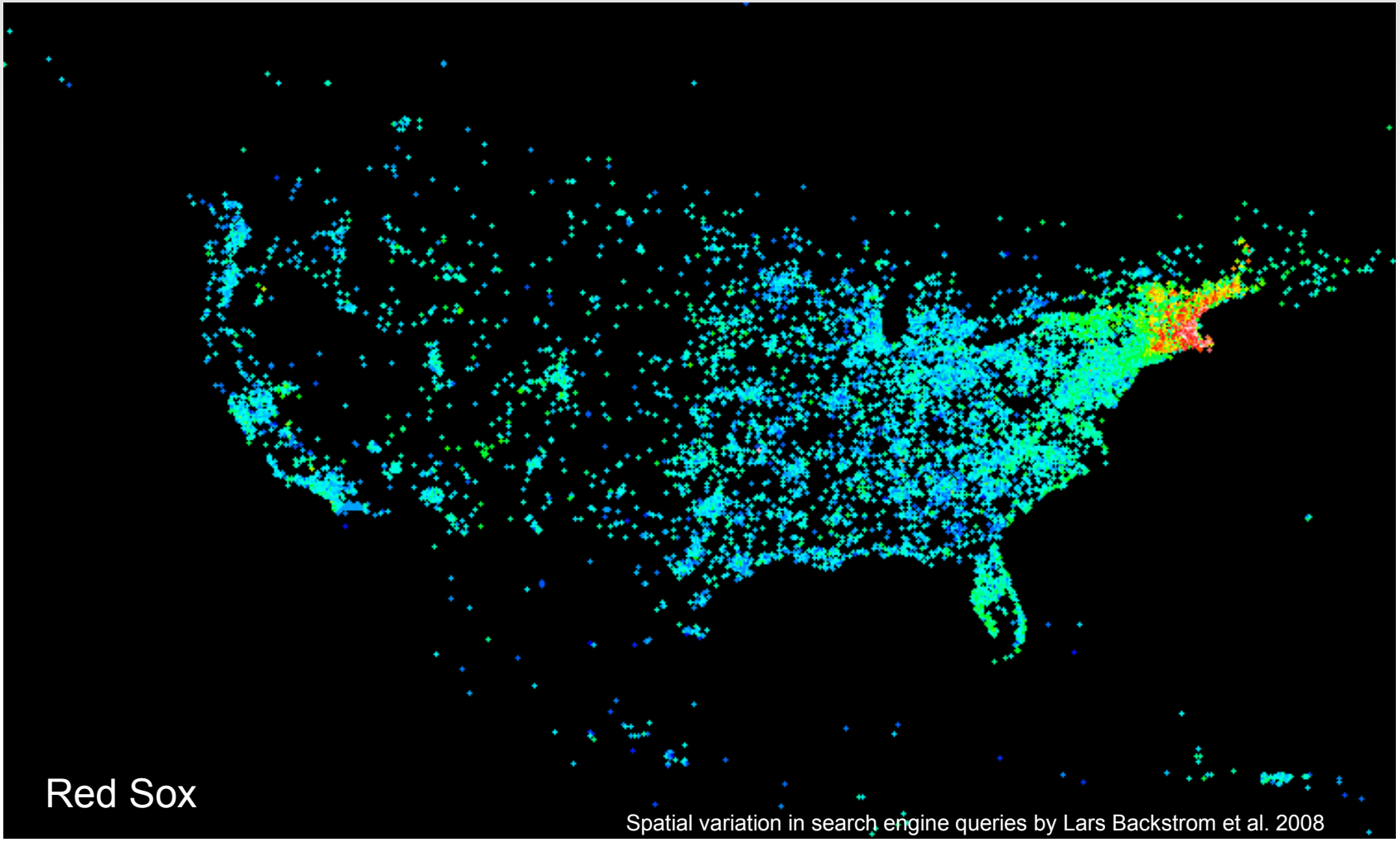
- Nutzer sucht nach ”Auto, Schleudersitz” und besucht dann www.xyz.com/modification.htm
- Chancen sind ziemlich gut:
  - Diese Webseite hatt etwas mit Autos und/oder Schleudersitzen zu tun
- Für Webseiten, die selber wenig Text enthalten

Query-sets: using implicit feedback and query patterns to organize web documents by Barbara Poblete et al. 2008

Geographische Verortung von Anfragen

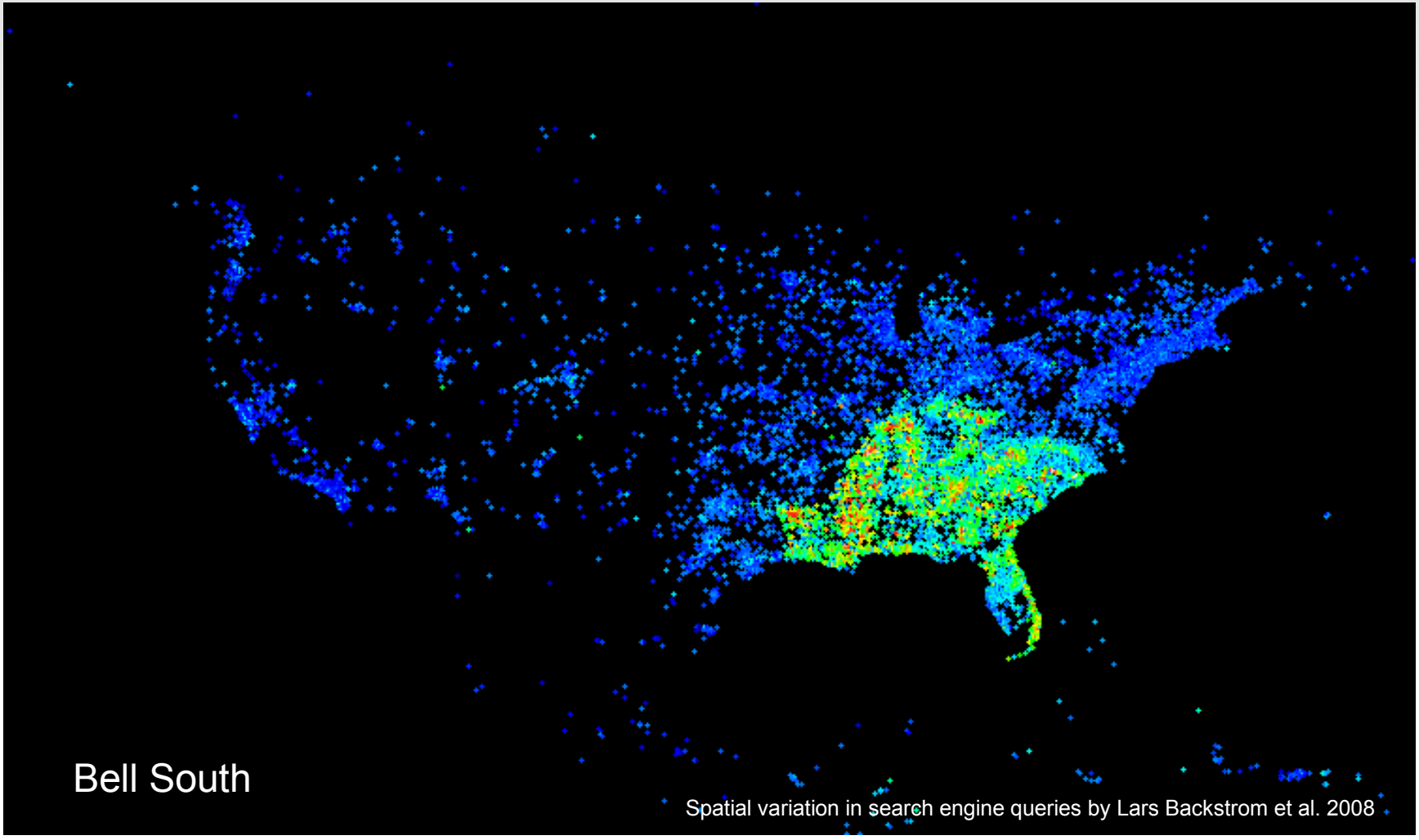
- Über IP-Addressen kann man herausfinden, wo ein Benutzer sich befinde (+/- 100km)
- Bestimmte Queries kommen offensichtlich aus verschiedenen Regionen
- Terme sind auf Regionen fokusiert
- Googles Grippe Vorhersage

Geographische Verortung von Queries



Spatial variation in search engine queries by Lars Backstrom et al. 2008

Geographische Verortung von Queries



Spatial variation in search engine queries by Lars Backstrom et al. 2008

Klassisches Beispiel

- Nutzer sucht nach ”Jaguar”, meint er:
  - Kätzchen
  - Auto
  - Spielkonsole ?
- Keine Ahnung, am besten gebe ich ihm ein paar Seiten zu jedem Thema

Und wenn Ich mehr über den Nutzer wüsste?

- Letztbesuchte Seite:
  - gebrauchtwagen.de
- Letzte Suchanfragen:
  - Leopard
  - Grassteppe
- (G-Mail) E-Mails über
  - Nintendo und Sega
- Bonner

- Personalized Search
- Versuch, dem Nutzer in den Kopf zu schauen
- Bessere Ergebnisse

Online Werbung

- Schaue in den Kopf von Nutzern
- Empfehle erfolgreiche Werbung
- Harmlos ... vielleicht

- Big business
- Yahoo $680 Million für RightMedia
- Google $3.1 Milliarden für DoubleClick
- Microsoft zahlt $6 Milliarden für aQuantive

Big Picture

- Mit gesammelten Daten können wir:
  - Webseiten empfehlen
  - Werbung einblenden
  - ...
- Automatisierung von Entscheidungen
  - Welche Seite?
  - Welche Werbung?
- Billig
- Wenn was schiefgeht? Egal...

Big Picture

- Web als Epi-Zentrum des Datensammelns
  - Daher die Einführung in IR
- Zusammenschalten unzähliger Datenquellen
  - Rasterfahndung
- Web besteht nicht nur aus Dokumenten
- Sondern speichert die ”Virtuelle Welt”
- Noch ganz andere Entscheidungen lassen sich automatisieren...

Nach der Pause

- Datenschutz als sozialer Umweltschutz
- Automatisierte Entscheidungen
- Case Studies
  - Amazon
  - Facebook
  - Gmail
  - World of Warcraft
  - Google Streetview

Pause

Weiter Geht's

- Datenschutz als sozialer Umweltschutz
- Automatisierte Entscheidungen
- Bei-Spiele

Datenschutz als sozialer Umweltschutz

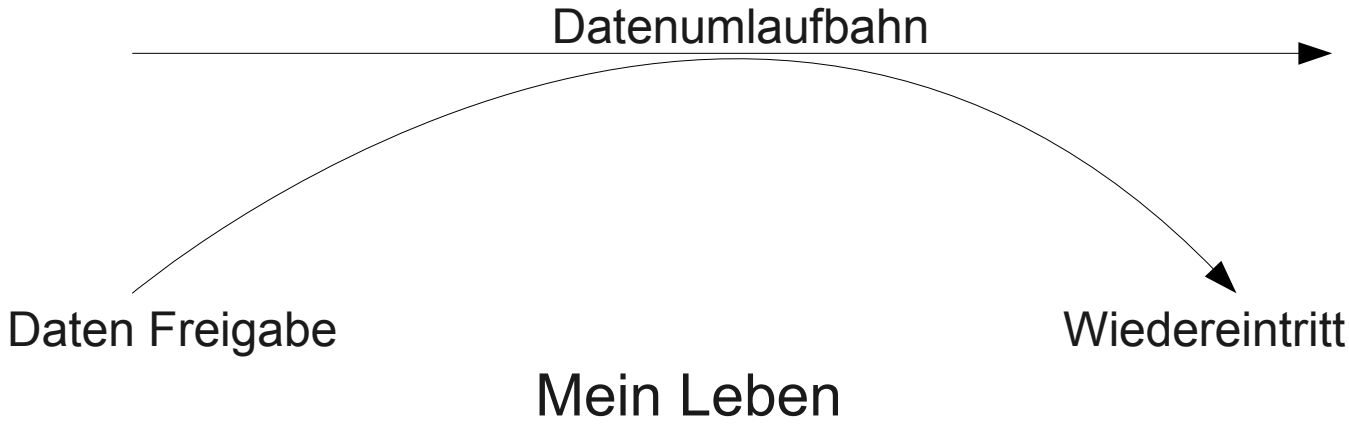
- Einzelne hat nur begrenzt Einfluss
  - Gift des Nachbarn verseucht auch meinen Garten
  - Daten des Nachbarn erschweren auch mein Leben
  - Erfordert gesellschaftliche Lösung
- Eigene Parteien
- Staatliche Umweltzerstörung bis ca 1970, dann radikale(?) Kehrtwende
- Staatliches Datensammeln, bis ca 20xx ...

Name & Vision

- Sind leider beide falsch gewählt
- Wir müssen Menschen schützen, nicht Daten
- Es geht nicht um ein übersteigertes Interesse an der einzelnen Person (Orwell)
- Sondern um ein Desinteresse an dem Einzelnen
  - Kafka & Huxley

Datenbalistik

- Freigegebene Daten sind nicht rückrufbar
- Relevant: Konsequenzen im realen Leben



Menschen sind zu teuer

- Und müssen wegrationalisert werden...
- Industrialisierung seit 1850
- Industrie-Roboter seit 1950
- Jetzt auch im Krieg

- Ersetzen menschliche Arbeitskraft
  - Zunächst rudimentär
  - Dann immer besser

Selbst Denken ist zu teuer

- Entscheidungen kosten Geld
- Und werden daher automatisiert
- Zunächst rudimentär:
  - Standard Operating Procedures
  - ”It's our policy...”
- Erlaubt billigere Arbeitskräfte
  - McDonalds für Entscheidungen

Fehler Einkalkuliert

- Statistischer Ansatz
  - Meistens klappts
- Fehler sind einkalkuliert und erlaubt
  - Solange sie nicht zu zahlreich sind
  - Einzelfall statistisch irrelevant
- Management by Numbers
  - US Army Body Counts
  - Guiliani's New York

Moderne IT

- Voll-Automatisierung von Entscheidungen
- Rating & Scoring
- Nicht zwangsweise schlechter als Menschen
  - Erste Credit Rating

- Beispiele:
  - Jobangebot
  - Schulzweig
  - Vergabe eines Kredites
  - Aufnahme in eine Versicherung
  - Betreten eines Flugzeuges

Amazon

- Gekaufte Produkte
- Gewünschte Produkte
- Geschenkte Produkte
- Angesehene Produkte

- Der Kunde bekommt bessere Produkte empfohlen
- Und hat mehr Zeit für seine Familie...

Facebook

- Hobbies
- Musikgeschmack
- TV-Serien & Filme
- Job
- Uniabschluss
- Beliebtheit
- Gelesene Bücher
- Sexuelle Orientierung
- Auch implizit:
  - Messages
  - Freundeskreis
- Arbeitseinsatz
- Stressprofil

Gmail

- Mit wem rede ich?
- Wie oft?
- Über was?

- Wenige Features
- Aber Grundlage reichhaltiger abgeleiteter Informationen

World of Warcraft

- Spielverhalten
- Wann?
- Wie oft?
- Mit wem?
- Sozialverhalten
  - Gilden
  - Freunde
  - ...
- Bewegung im 3D Raum
- Spielstrategien
- Rollen
- Text (aus Chats)
  - Uninteressant: Orks & Elfen

Google Streetview

- Abfotografierte Strassenzüge
- Und WLAN Namen

- Wieso?
- Was geht damit noch?

Und aus Sicht des Nutzers?

- Entscheidungen betreffen direkte vitale Interessen

- Statistik ist irrelevant
- Nur der Einzelfall zählt

- Egal ob Entscheidungen richtig oder falsch sind
- Egal ob aufgrund richtiger oder falscher Daten

Fehler? Dein Problem!

- Keine Einspruchsmöglichkeiten
  - Zu teuer
    - Oder rufen Sie mich an: 0190 ...
  - Datenlage ist geheim
  - Entscheidungslogik ist geheim
  - Alles top secret, völlig geheim, gibt es gar nicht...
- Willkommen in Kafkas ”Prozess”

Das Korsett

- Entscheidungsfreiheit radikal eingeschränkt
  - Aufgrund von Daten
- Huxley's ”Brave New World”
  - Vorgegebene Lebenswege
  - Alpha und Beta Menschen nicht gezüchtet
  - Sondern aufgrund der Datenlage aussortiert
  - Nicht notwendigerweise eigene Daten

Entscheidungen aber ..

- Begründen das Mensch-sein
- Gesamte Christliche Ethik basiert auf ”Umkehr”
  - Fundamentalen Entscheidungen
- Selbst das StGB kennt Verjährung und Löschung
  - Um Entscheidungen zu erlauben

- Stattdessen, wird der Mensch zum Ding
  - Barcode in der Armbeuge

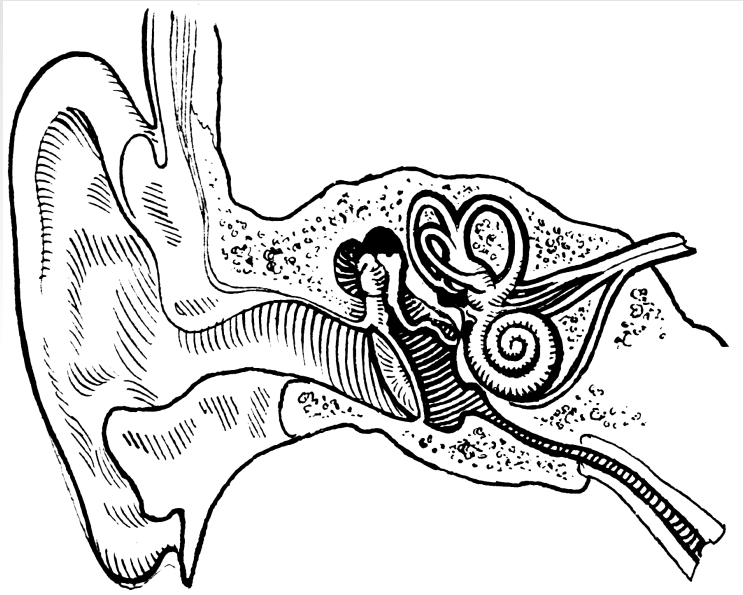
Und nun?

- Erneuerung des gesamten Datenschutzes
  - Zielführend, ersetzt derzeitiges Patchwork

- Dringend gesellschftlicher Diskurs über Ziele
- Besonders über Automatisierte Entscheidungen

- Und bis dahin erstmal: .... Angst haben....

Vielen Dank



Thank you for listening

Literatur

- Jiawei Han, Micheline Kamber, Jian Pei
  Data Mining: Concepts and Techniques
  Morgan Kaufmann; 2nd ed. 2005

- Ian H. Witten, Eibe Frank
  Data Mining: Practical Machine Learning Tools and Techniques
  Morgan Kaufmann, 2nd ed. 2005

- Thomas Mitchell
  Machine Learning
  McGraw Hill, 1997

Literatur

- Soumen Chakrabarti
  Mining the Web
  Discovering Knowledge from Hypertext Data
  Morgan-Kaufmann Publishers 2002

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
  Introduction to Information Retrieval
  Cambridge University Press. 2008. (Preprint on the Web)

- D. Easley, J. Kleinberg
  Networks, Crowds, and Markets: Reasoning About a Highly Connected World
  To be published by Cambridge University Press, 2010.
  (Preprint on the Web)

Literatur

- Stephen Baker
  The Numerati
  Mariner Books, 2009

- Peter Schaar
  Das Ende der Privatsphäre: Der Weg in die Überwachungsgesellschaft
  C. Bertelsmann, 2007

Was ich sonst so treibe

- Code Search
- Online Spiele
- Mensch Zwo-Null

Code Suche

- Suchmaschine für Programmcode
- Wenig Wörter
- Viel Struktur
- Mit Uni Warschau

Mensch 2.0

- 2009-10 Datenschutz
- Mit Prof. Joachim von zur Gathen
- Prof. Dr. Klaus Brunnstein, Hamburg
- Gerhart Baum, Bundesminister a.D
- padeluun
- Peter Schaar, Bundesdatenschutzbeauftragter
- Prof. Dr. Knut Wenzel, Frankfurt

Mensch 2.0

- 2010-11 Psychosoziale Nachbeben der Internetrevolution
- Schrumpfende Aufmerksamkeitsspannen
- Burnout
- Over-Multitasking
- Virtuelle Freundschaften und Beziehungen
- Sucht und Zwangsverhalten
- Etc. etc.

Nochmals vielen Dank

- Nu ist aber wirklich gut...