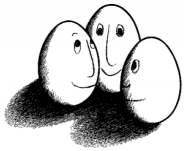
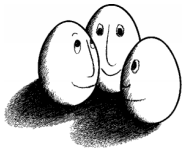
Big Data Analytics



Überblick

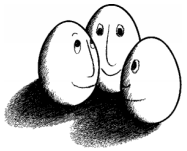
- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

- Proseminar



- Volume:
  Sehr große Datenmengen
  - Hochdimensional
  - Viele Beobachtugen
- Velocity:
  Datenströme werden realzeitlich verarbeitet
- Variety:
  Unterschiedliche Quellen, heterogene, verteilte Daten sollen integriert werden.

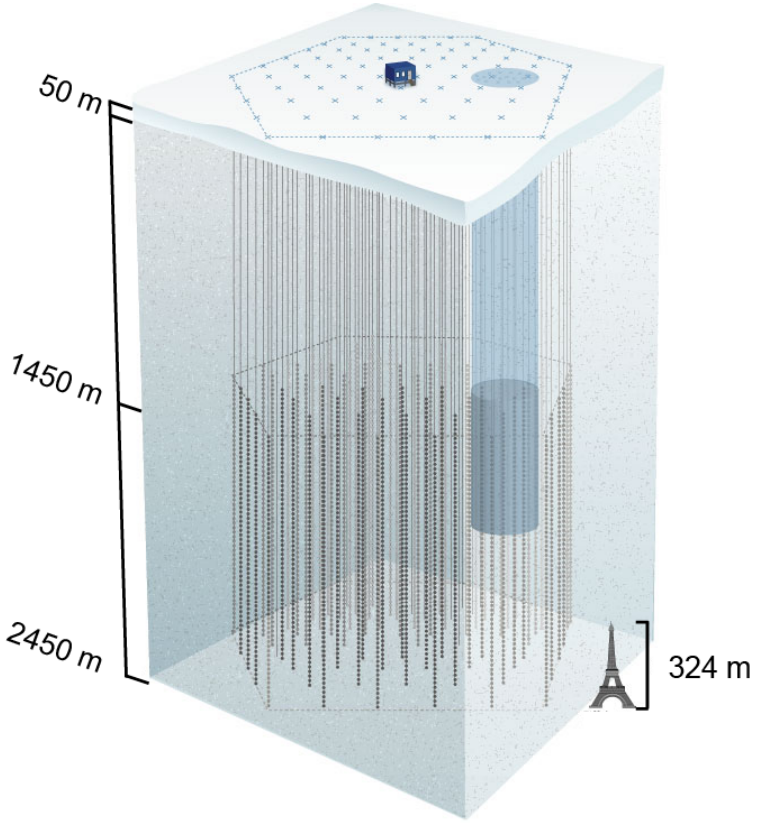
➢ Neue Algorithmen der Datenanalyse gefordert!

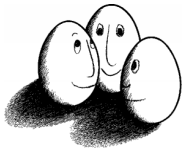


Wert der Daten: Wissenschaft

- IceCube *Breakthrough of the Year 2013* (Physics Worlds).
- Zeit für die Übertragung der Daten eines Jahres (365 TB) vom Südpol zur Uni Wisconsin
  - Satellit 10 Jahre
  - Schiff 28 Tage.
- ➢ Datenanalyse, um Neutrinos zu finden.

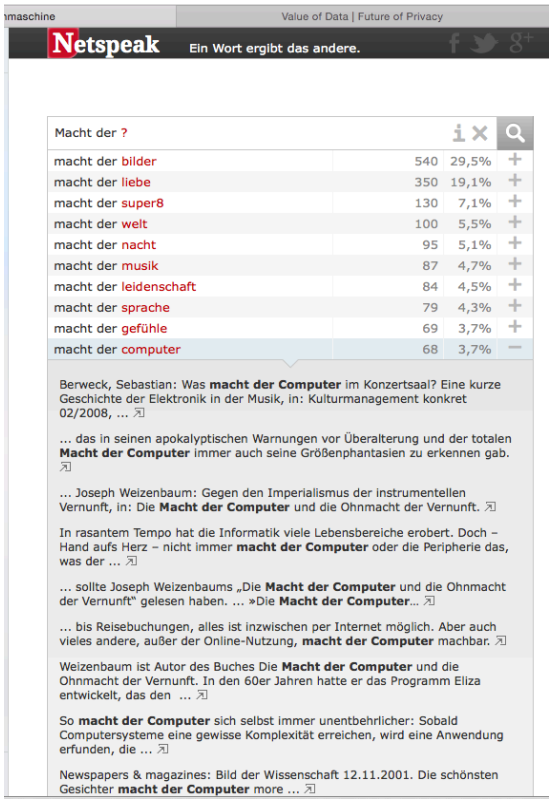
Schiff 130 x schneller…



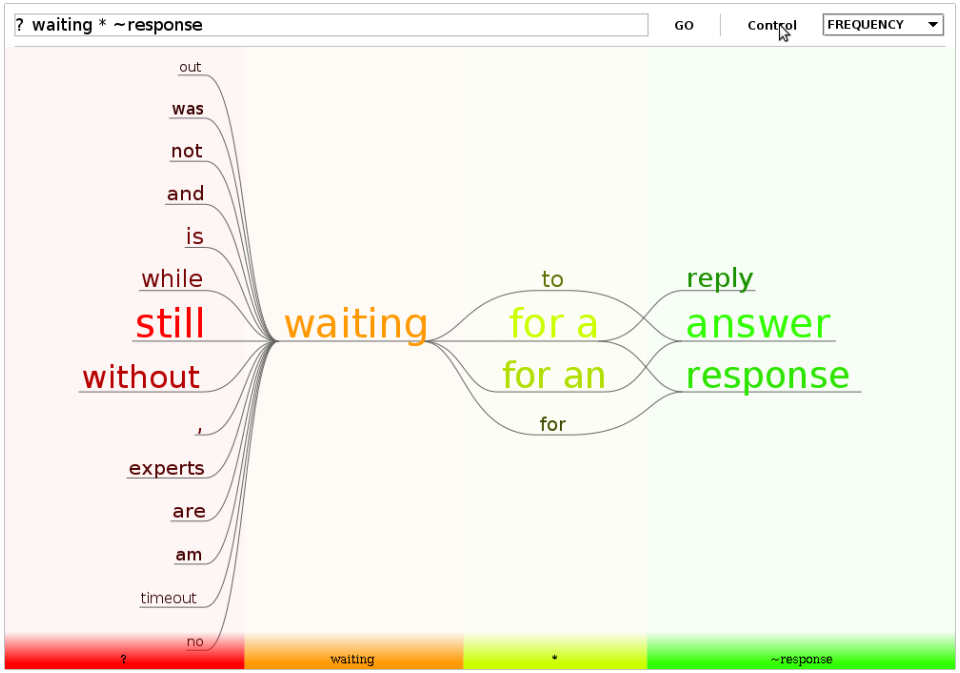


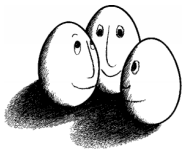
Wert der Daten: Wissenschaft und Schreibunterstützung

- Korpuslinguistik
- Netspeak



- Riehmann, P., Gruendl, H., Potthast, M., Trenkmann, M., Stein, B., Froehlich, B. WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK's Wildcard Search IEEE Transactions on Visualization and Computer Graphics, 2012
  http://www2.uni-weimar.de

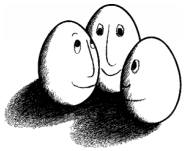




Problem der Reproduzierbarkeit

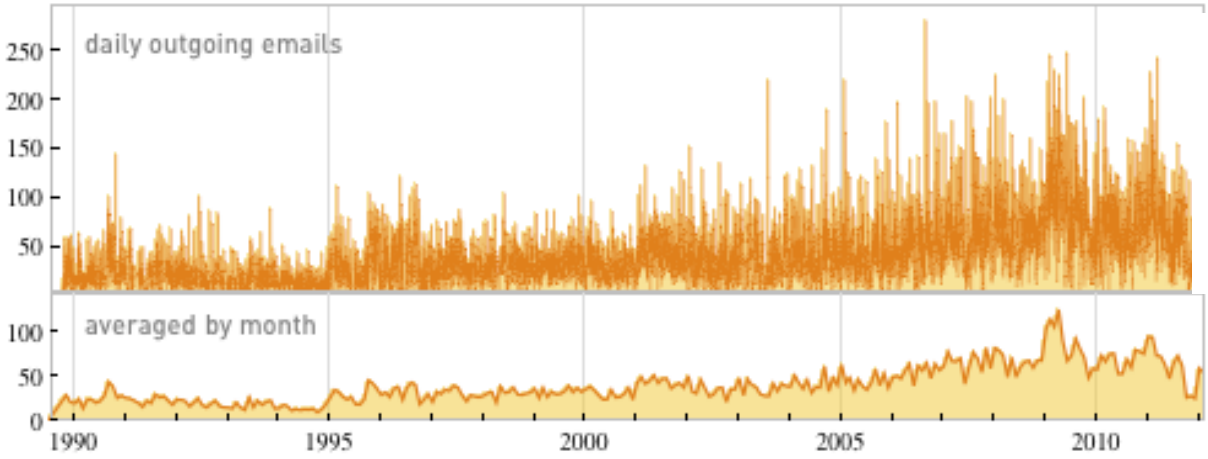
- Wer dominiert das Web?
  Reiche, mächtige, gebildete, überwiegend männliche Elite nutzt und gestaltet das Web.
  - The digital divide is part of social inequalities in Western societies. Worse still, it strengthens them. (Muki Haklay 2012)
- Forschungsergebnisse, die auf Rechnern und Daten von Google gerechnet wurden, sind nicht reproduzierbar!
- ➢ Algorithmen müssen auf großen Rechenfarmen erprobt werden. Wer hat die?

- Peter Norvig (Google): "All models are wrong, and increasingly you can succeed without them.“
- “The companies, governments, and organizations that are able to mine this resource will have an enormous advantage over those that don’t.“ Bryan Trogdon in a 2012 survey by Elon University NC, USA
- Europa verlässt sich auf
  - GPS
  - Google
  - Amazon
  - …



Wert der Daten: Selbsterfahrung, Selbstoptimierung

- Stephen Wolfram (Mathematica, Alpha Pro) publizert seine Daten, z.B. Anzahl geschriebener eMails.
- Es gibt weltweit Treffen: http://quantifiedself.com/



Kontrolliert vom Handy

Mona Ameziane lässt für ein Experiment ihre Gewohnheiten überwachen

Mona Ameziane lässt sich kontrollieren. Von ihrem Handy. Vier Wochen lang, jeden Tag. Es ist ein Experiment. Die Journalistik-Studentin der TU Dortmund möchte mithilfe ihres Smartphones mehr über sich und ihre Gewohnheiten herausfinden. Die Ergebnisse dokumentiert sie unter dem Titel „Quantified Mona" in einem Videoblog.

Sie sei, sagt die 20-Jährige, niemand, der ständig mit dem Handy vor der Nase herumlaufe, niemand, der jede

recht."

Der Reiz war groß, das Projekt geboren. Vier Wochen lang ist ihr Smartphone nun ihr bester Freund. Jede Woche testet sie andere Apps zu anderen Themen: Essen und Trinken, Sport, Geld und Schlafen. „Ich möchte gerne wissen, wie ich das in meinen Alltag integrieren kann", sagt Ameziane.

Jeden Freitag ein Video

Ihre Erfahrungen dokumentiert sie auf Twitter, jeden Freitag veröffentlicht sie ein

„Ich denke, ich bin sportlich. Aber vielleicht sagt mein Handy ja etwas anderes?"

Mona Ameziane. Studentin

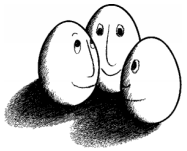
Essen und Trinken.

Bei diesem Thema ist Mona Ameziane gleich an Grenzen gestoßen. Mit einer speziellen App hat sie jede Mahlzeit, die sie zu sich genommen hat, fotografiert. Die App hat auto-

sig". Ständig vor dem Essen das Handy zu zücken, habe sie schon nach kurzer Zeit ziemlich genervt.

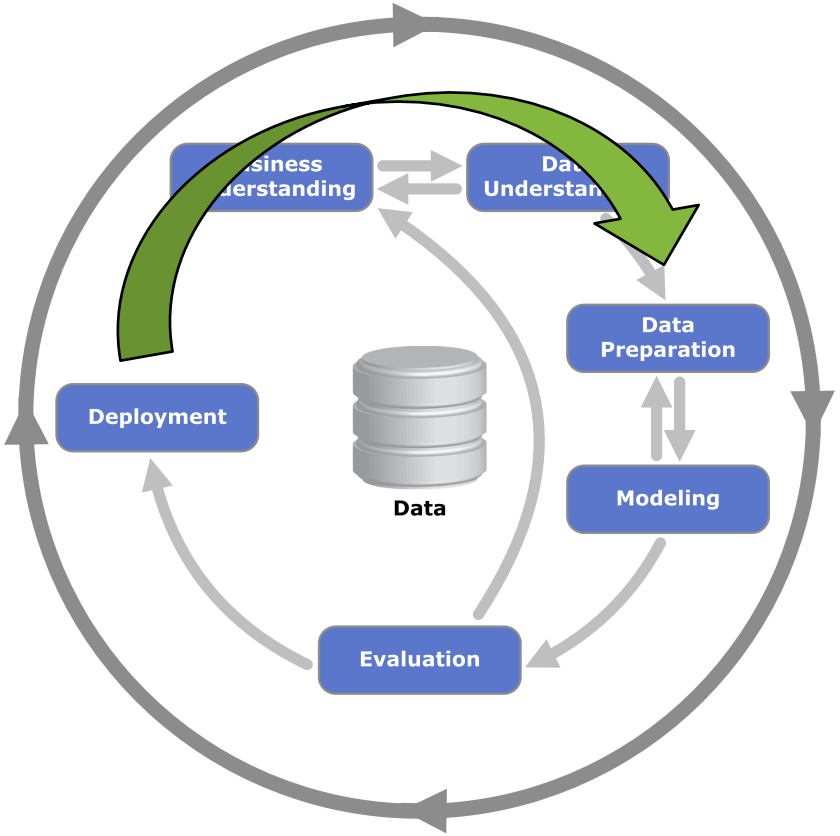
Eine zweite App hat für sie errechnet, wie viel sie am Tag trinken muss, und sie ständig daran erinnert. Diese Funktion will sie weiterhin nutzen.

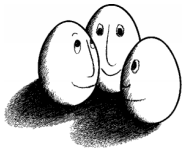
In den nächsten Wochen wird das Smartphone weitere Lebensbereiche erforschen: ihre Fitness, ihre Finanzen, ihr Schlafverhalten. „Ich denke", sagt Ameziane, „ich bin sportlich. Aber vielleicht sagt mein Handy ja etwas ande-



Wert der Daten: Industrie 4.0

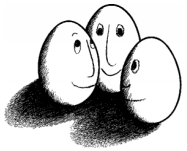
- Daten für intelligente Produktion
  - Sensoren
  - Prozesskontrolle
  - Umgebungseinflüsse
- Prognose ermöglicht rechtzeitiges Eingreifen
  - Marktanalyse
  - Wartungsprognose
  - Qualitätsprognos
- Qualitätsprognos während der Prozesse verbessert diese sofort.





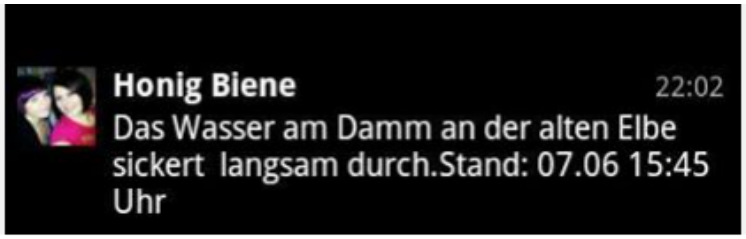
Wert der Daten: Industrie 4.0

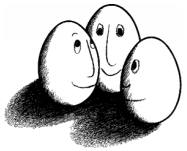
- Projekt LS 8 mit SMS Siemag und Dillinger Hüttenwerke
- Prognosemodelle im Stahlwerk
  - Datenströme
  - Merkmalsextraktion
  - Analyse der Prozessdaten zur Vorhersage nutzen!
- Wir haben schon den ersten Prototyp ins Werk gebracht.



Wert der Daten: Koordinierung untereinander

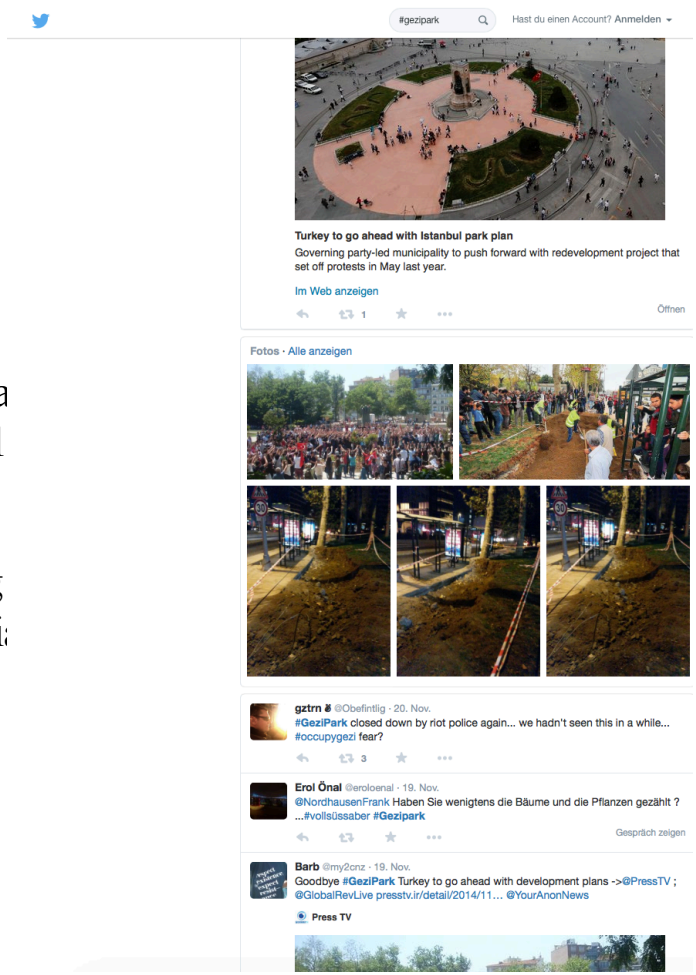
- Information
  - Vor Ort
  - Echtzeit
- EU-Projekt INSIGHT,
  Koordinator: D. Gunnopoulos
  BBK, Dublin CC, TU Dortmund
  Fraunhofer IAIS, IBM, Technion

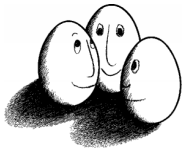




Probleme

- Stille Post Effekte, Gerüchte
- Automatisches Handeln anhand von Tweets
  - AP Tweet über Explosion im Weißen Ha führte zu Kurseinbrüchen der Börse (201
- Gezielte Beeinflussung durch Bots
  - “When social bots attack: Modeling susceptibility of users in online soci networks” MSM 2012

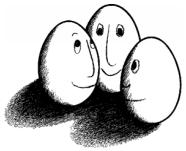




Überblick

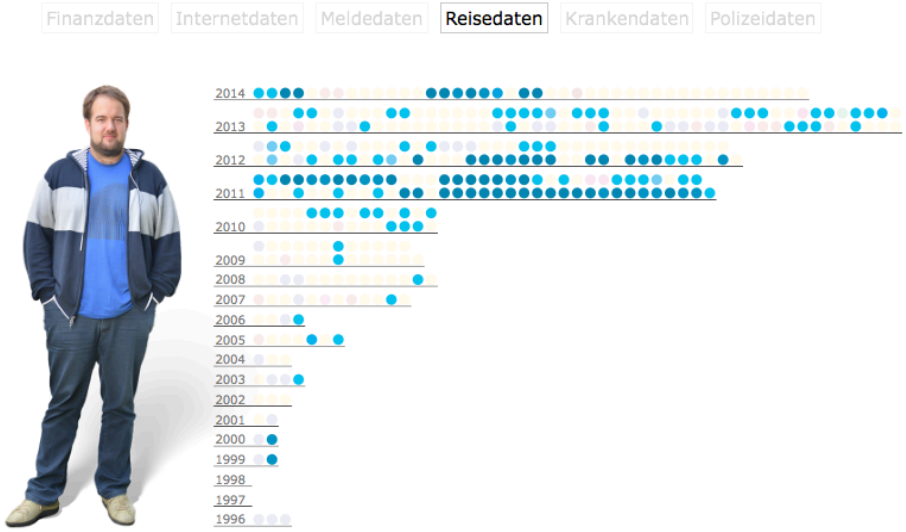
- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

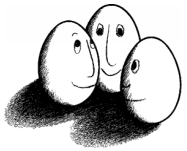
- Proseminar



Privatheit

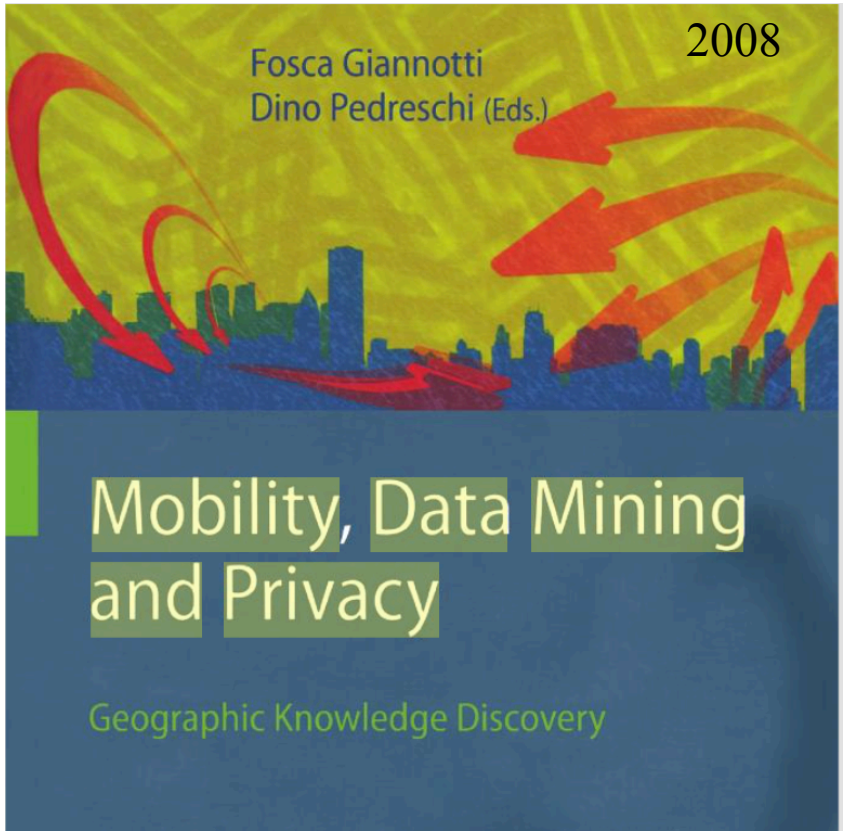
- Spiegel Online 28.10.2014
  Malte Spitz hat bei Firmen und Behörden nach seinen Daten gefragt.
  Kombination verschiedener Quellen ergibt ein Bild, wo er wann war.
- Marketing, Logistik, Verkehr, Unterhaltungsindustrie, brauchen statistische Angaben, nicht individuelle!

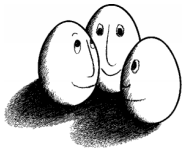




Privacy

- Rakesh Aggrawal, Ramakrishnan Srikant “Privacy Preserving Data Mining” 2000
- ECML PKDD Conference Pisa 2004
- Fosca Giannotti, Francesco Bonchi et al. 2005
- IEEE ICDM 2012 Brüssel Katharina Morik: Panel

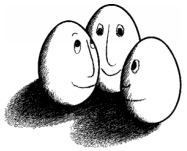




Big Data

- Die einzelnen Fahrradkuriere sind uninteressant.

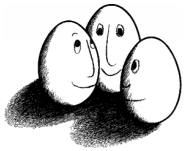
- Anzahl der Fahrten pro Stunde können für verbesserte Planung genutzt werden.



Überblick

- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

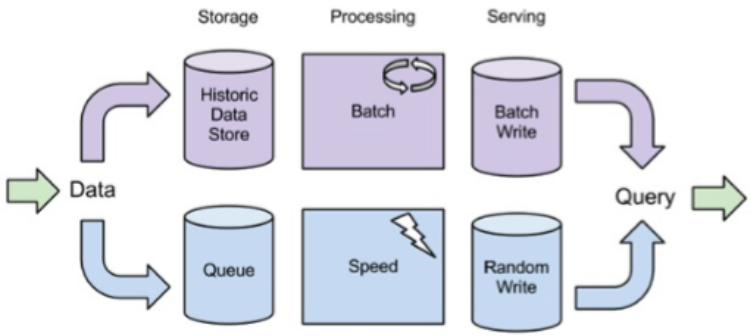
- Proseminar

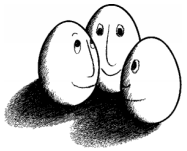


Umgebungen

- Verteilung der Daten und Prozesse auf ein Rechencluster
- Batch z.B. Hadoop, Spark
- Streams: z.B. Storm, streams
- Nathan Marz, James Warren “Big Data: Principles and best practices of scalable realtime data systems” Manning publications 2015

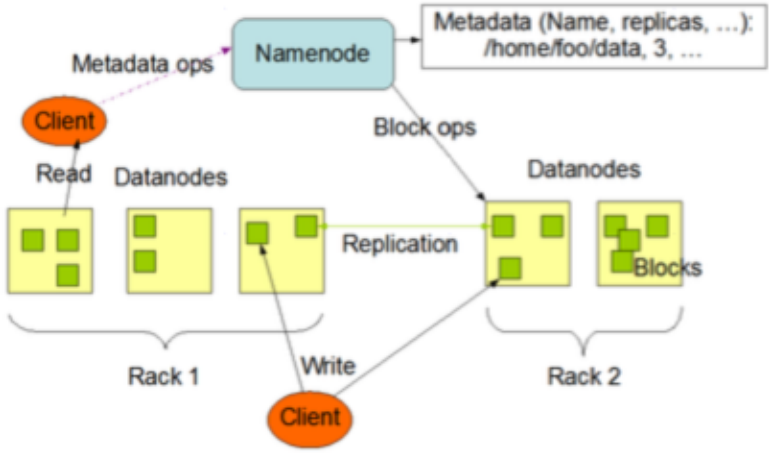
Lambda-Umgebung

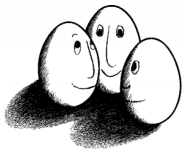




Umgebungen: Apache Hadoop Projekt

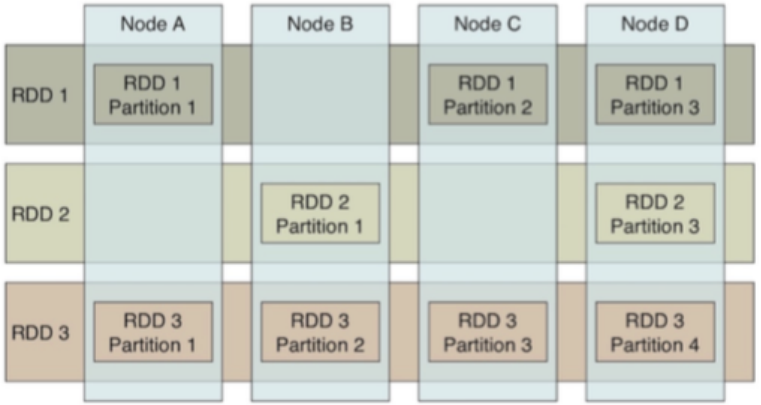
- Speichern: Hadoop Distributed File System (HDFS)
- Ressourcen-Verwaltung: Yet Another Resource Allocator (YARN)
- Programmierparadgma: Map Reduce

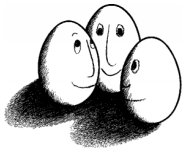




Umgebungen: Apache Spark

- Spark Core
  - Resilient distributed datasets (RDD)
  - Transformationen und Aktionen
- Spark SQL
  - Zusammenführung von Datenquellen
  - SQL-Anfragen
- Spark Streaming
- GraphX
- MLlib





Umgebungen: Apache Storm

Aufgaben

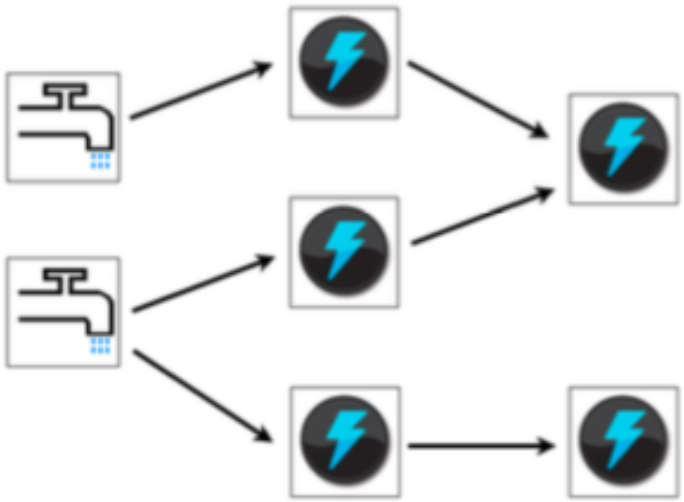
- Knoten für Datenquellen (spouts) und für Prozesse (bolts)
- Kanten sind Datenströme

Ausführung

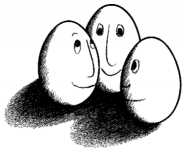
- *Zookeeper*
  - Master nodes verteilen den Code, nutzen *Nimbus*
  - Worker nodes sind auf mehrere Maschinen verteilt und führen Code aus, nutzen *Supervisor*

Spouts
Ausguss

Bolt
Bolzen



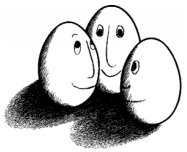
Topologie des Datenflusses



Überblick

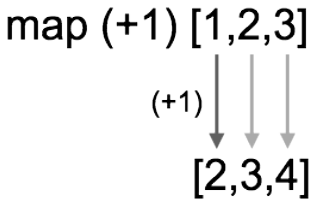
- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

- Proseminar

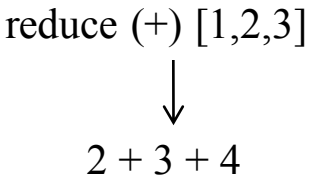


Algorithmen: MapReduce

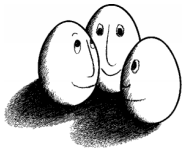
- Die Funktion *map* wendet eine Funktion auf jedes Element einer Liste an



- Die Funktion *reduce* wendet eine Funktion auf eine Liste an und liefert ein Ergebnis

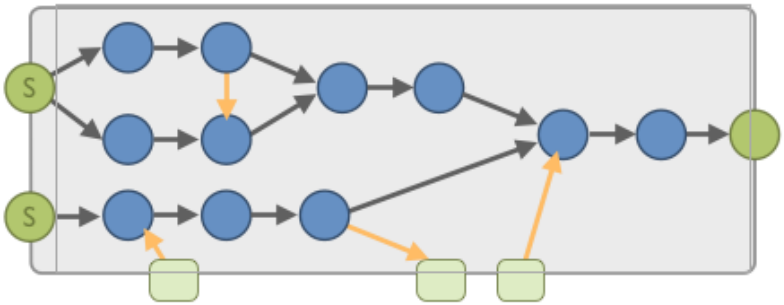


- *MapReduce - Simplified Data Processing on Large Clusters,* J. Dean und S. Ghemawat, 2004
- Algorithmen in diesem Sinne neu formulieren, so dass sie parallel ausführbar sind!



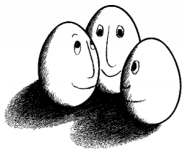
Algorithmen: Datenströme

- Compute Graphen
  Jeder Koten rechnet über Elementen des Stroms



- Abstrakte Modellierung im *streams* Framework (LS 8 , Christian Bockermann)

- Jedes Datum (Messwert) darf nur einmal betrachtet werden (One Pass Algorithmen)
- Viele Analysen müssen Häufigkeiten zählen als Annäherung an die Wahrscheinlichkeit.
- Zählen ist bei beschränkten Ressourcen schwierig!



Zählen kann schwierig sein

- Eingabe: ein Strom von Tweets
- Ausgabe: 10 häufigste #-tags
- Naiver Ansatz:
  - Richte für jeden #-tag einen Zähler (4 Byte) ein.
  - Großer Speicherbedarf!
- Approximationsalgorithmus
  - Liefert ein Ergebnis und den möglichen Fehler.
  - Fenstergröße und Speicherbedarf vs. Genauigkeit!

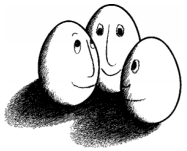
- Lossy Counting

Man teilt den *Strom* $S=s_1, s_2, \ldots$

in Fenster von w Elementen und zählt das Vorkommen von Beobachtungen e_i. Die Häufigkeit $D(e)$ wird angegeben als f, Δ.

Nach einem Fenster wirft man alle Zählungen weg, die nicht häufig genug sind, übernimmt nur die anderen.

Der Parameter Δ zählt mit, wie viel verlorengegangen sein kann.



Algorithmen: Naive Bayes

- Wahrscheinlichkeit für
  der Kunde kauft (A),
  der Kunde kauft nicht (A) $\overline{}$
  bei Beobachtungen x
- Mit dem Satz von Bayes bestimmen wir die bedingte Wahrscheinlichkeit.
- Wir schreiben das um als Zählen:
  - Wie oft kommt x vor?
  - Wie oft kommt A vor?
  - Wir oft kommt $\underline{A}$ vor?

- Naive Bayes

$$g(x) = y \quad y \in \left\{A, \overline{A}\right\}$$

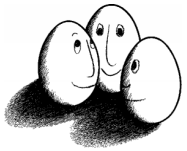
$$P(A \mid x) = \frac{P(x \mid A)P(A)}{P(x)}$$

- zählen:

$$Q = \frac{\left(|x| : |A|\right)\left(|A|\right)}{\left(|x| : |\overline{A}|\right)\left(|\overline{A}|\right)}$$

$$Q \geq 1 \rightarrow g(x) = A$$

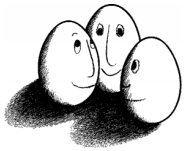
- Wir verwenden lossy counting und untersuchen Speicherbedarf und Genauigkeit.



Überblick

- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

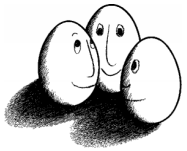
- Proseminar



Proseminar führt ein in wissenschaftliches Arbeiten

- **Schule** war gestern
  - Vorgegebener Stoff
  - Lehrer weiß es, Schüler lernt es.
- An der **Universität** wird Wissen geschaffen!
  - In die Gemeinschaft der WissenschaftlerInnen hinein wachsen!
- Im **Beruf** muss man sich selbstständig in Neues einarbeiten können.

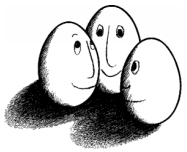
Auf den Schultern von Riesen



Proseminar: Bibliographieren

- Literatursuche
  - Begriffsdefinition
  - Zentraler Artikel
- Literaturbewertung
  - Autoren
  - Erscheinungsort
  - Ist das Problem schwierig?
  - Ist die Lösung besser als bisherige Ansätze?
  - Ist die Lösung allgemein?
  - Sind alle Behauptungen belegt?

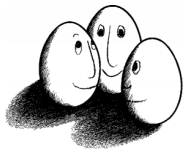
- Dienstag, 19.April
- 14 – 16 Uhr
- Bibliothek
  Raum 215



Proseminar: Vorgehen

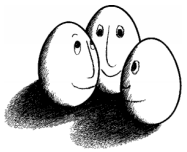
- Bienenstöcke im Seminar
  - Internetrecherche
  - Qualitätsdiskussion
- Einzelarbeiten zu Hause
  - Literaturverzeichnis erstellen
  - Abstracts lesen, 1 – 2 Artikel auswählen
  - Artikel lesen, im Seminar berichten
- Referat
  - 15 Minuten Präsentation
  - Ca. 20 Seiten Text

1. Recherchieren lernen
2. Fachgebiet strukturieren
3. Thema im Untergebiet finden – bis 26. April
4. Thema darstellen – ab 3. Mai



Gruppen

1. Gruppe
2. Gruppe
3. Gruppe
4. Gruppe
5. Gruppe
6. Gruppe
7. Gruppe



Stichwörter

Tools

- MapReduce
- Radoop
- Spark
- Storm
- Streams

Anwendungsgebiete

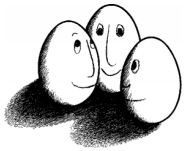
- Astrophysik
- Verkehr
- Sentiment Analysis
- Social Networks

Methods

- Data stream clustering
- Mining high-speed data streams
- Active Learning from data streams
- Ensemble Classifiers
- Link prediction
- Social influence in social networks
- ...

Grundlagen

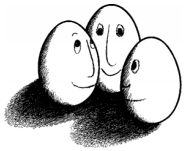
- Hoeffding Schranke
- Bayes Gesetz
- Lossy Counting



Zeitplan: mindestens pro Woche

- Proseminar:
  2 Stunden – immer hingehen!
- Hausarbeit:
  3 Stunden
- Referat vorbereiten:
  insgesamt 16 Stunden
- 6 Stunden pro Woche sollten Sie für das Proseminar einplanen

- 112 wache Stunden pro Woche
- 50 Stunden Studium
- 20 Stunden Bahnfahrt, Haushalt, Geld verdienen
- 10 Stunden Feiern
- 32 Stunden für Sport, Kultur, Puffer, Albernheiten...



Überblick

- Wert der Daten
- Privatheit
- Umgebungen
- Algorithmen

- Proseminar

- Und wer sind Sie?
  - Semester
  - Interessen
  - Fragen?