Big Data - Neue Wege zur Wissens-erweiterung oder reine Datenbulimie?

Andreas Michels ■ 19.12.2014

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Eine Begriffsbestimmung

- **Big Data**
- **Datenbulimie**
- **Wissenserweiterung**

Big Data – eine (erste) Definition

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

"Big Data bezeichnet große Datenmengen aus vielfältigen Quellen, die mit Hilfe neu entwickelter Methoden und Technologien erfasst, verteilt, gespeichert, durchsucht, analysiert und visualisiert werden können"

(laut der Definition der wissenschaftlichen Dienste des deutschen Bundestags)

(Quelle: http://www.bundestag.de/blob/194790/c44371b1c740987a7f6fa74c06f518c8/big_data-data.pdf , abgerufen am 18.12.2014)

Big Data – eine (zweite) Definition

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- **"Big Data beschreibt Datenbestände, die aufgrund ihres Umfangs, Unterschiedlichkeit oder ihrer Schnelllebigkeit nur begrenzt durch aktuelle Datenbanken und Daten-Management-Tools verarbeitet werden können.**
- **In Abgrenzung zu existierenden Business Intelligence (BI) und Data Warehouse Systemen arbeiten Big Data Anwendungen in der Regel ohne aufwändige Aufbereitung (siehe: ETL Prozess) der Daten. Dies ermöglicht Kosteneinsparungen, Flexibilität und einen schnellen Zugriff auf Analysen aktuellster Daten.**

(Quelle: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/daten-wissen/Datenmanagement/Datenmanagement--Konzepte-des/Big-Data, abgerufen am 18.12.2014)

Big Data VVV – Die Entwicklung der Vs

UNIVERSITÄT DUISBURG ESSEN

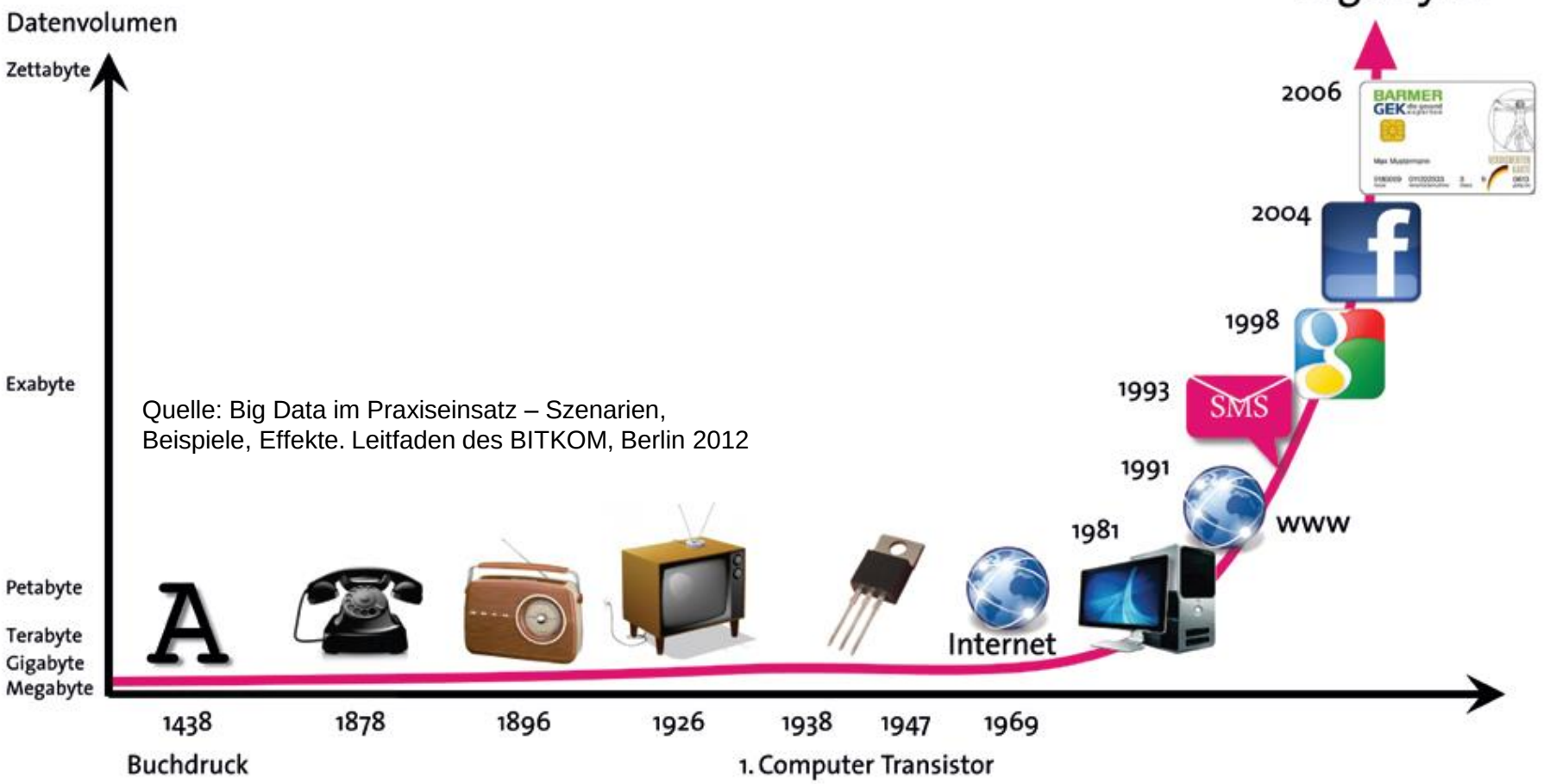
Offen im Denken

Zunächst:

Die drei Vs von Big Data:

1. **Volume - das Datenvolumen**
2. **Velocity - die Geschwindigkeit**
3. **Variety - die Vielfalt**
4. **Veracity - die Wahrhaftigkeit**
5. **Value - der Wert**

Das Datenvolumen



Quelle: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM, Berlin 2012

Die Datengeschwindigkeit

- **Daten werden mit immer höherer Geschwindigkeit erzeugt und sollen entsprechend analysiert werden**

Verarbeitung als	- Batch
	- Zeitintervall-gesteuert
	- Datenstrom
	- echtzeit

Die Datenvielfalt

Tweets

Logdateien

Sensordaten

Bilder

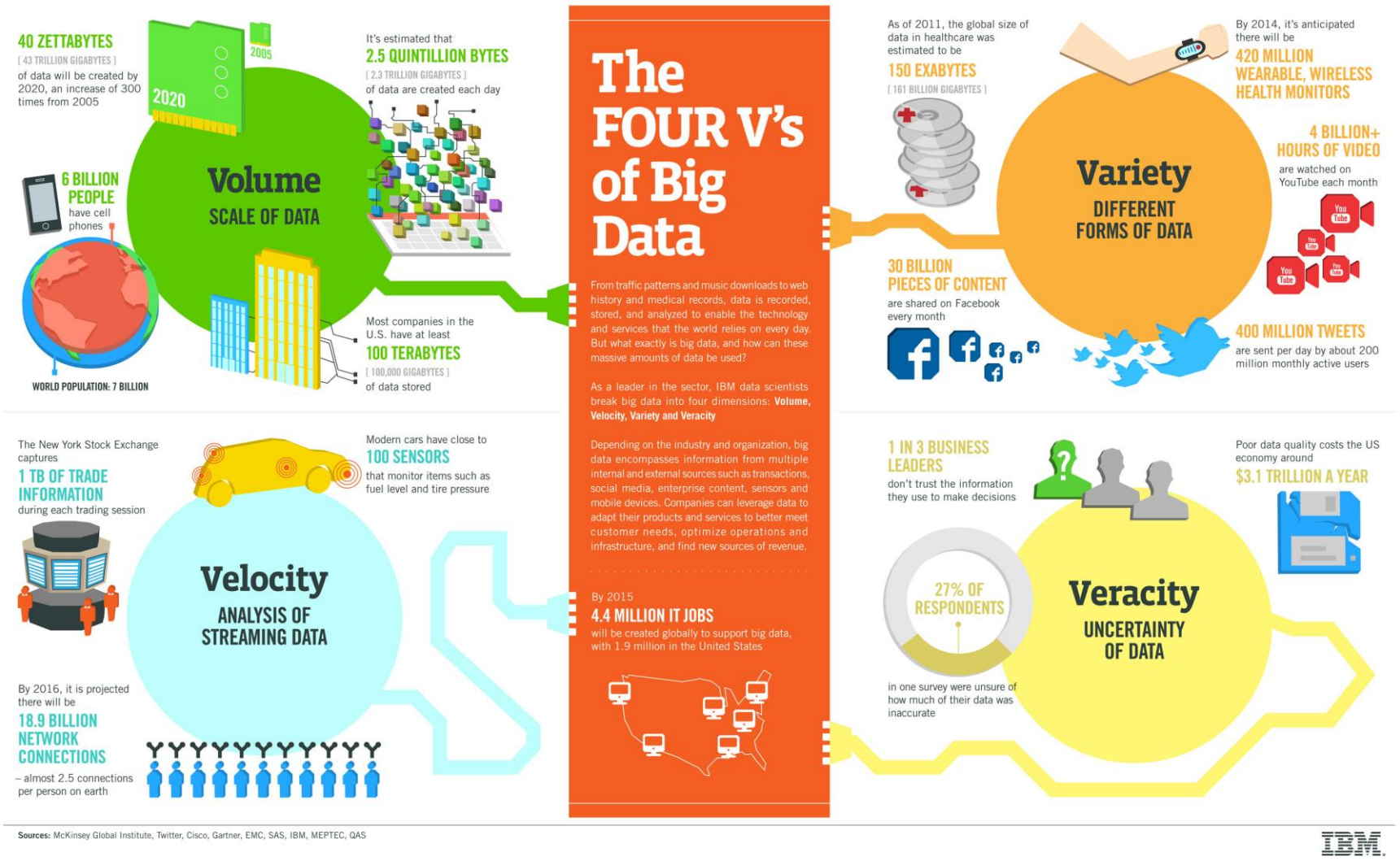
Videos

…

Die Wahrhaftigkeit der Daten

- Wie können „gültige“ von „ungültigen“ Daten unterschieden werden (z.B. technische Fehler bei Sensoren)
- Wie kann gesichert werden, dass die Daten „echt“ sind?
- Wie kann Datenmanipulation verhindert werden?
- Wie ist generell mit der Unsicherheit der Daten umzugehen?

Die vier V´s von Big Data



Welcher Wert wird den Big Data – Analysen zugesprochen, z. B.:

- **wirtschaftlicher:**
  steigert Big Data den Gewinn (von wem)?
- **sozialer:**
  können gesellschaftliche Strukturen verändert werden (z. B. im Gesundheitssystem)?
- **erkenntnistheoretischer:**
  fördert Big Data Erkenntnisprozesse?

Beispielszenarien für Big Data

- **Finanz- und Risikomanagement**
- **Verkehrsfluss**
- **Marketing**
- **Vorhersagen**
  - Epidemien
  - Kaufverhalten
  - Wahlergebnisse
- **…**

Was ist neu an Big Data?

- **Alle Aspekte der V´s sind offensichtlich nicht neu – natürlich bis auf den drastischen absoluten Zuwachs der Datenmengen**

Aber vieles hat sich geändert:

- **Neue Technologien**
- **Neue Herangehensweisen**

Was macht Big Data dann aber aus?

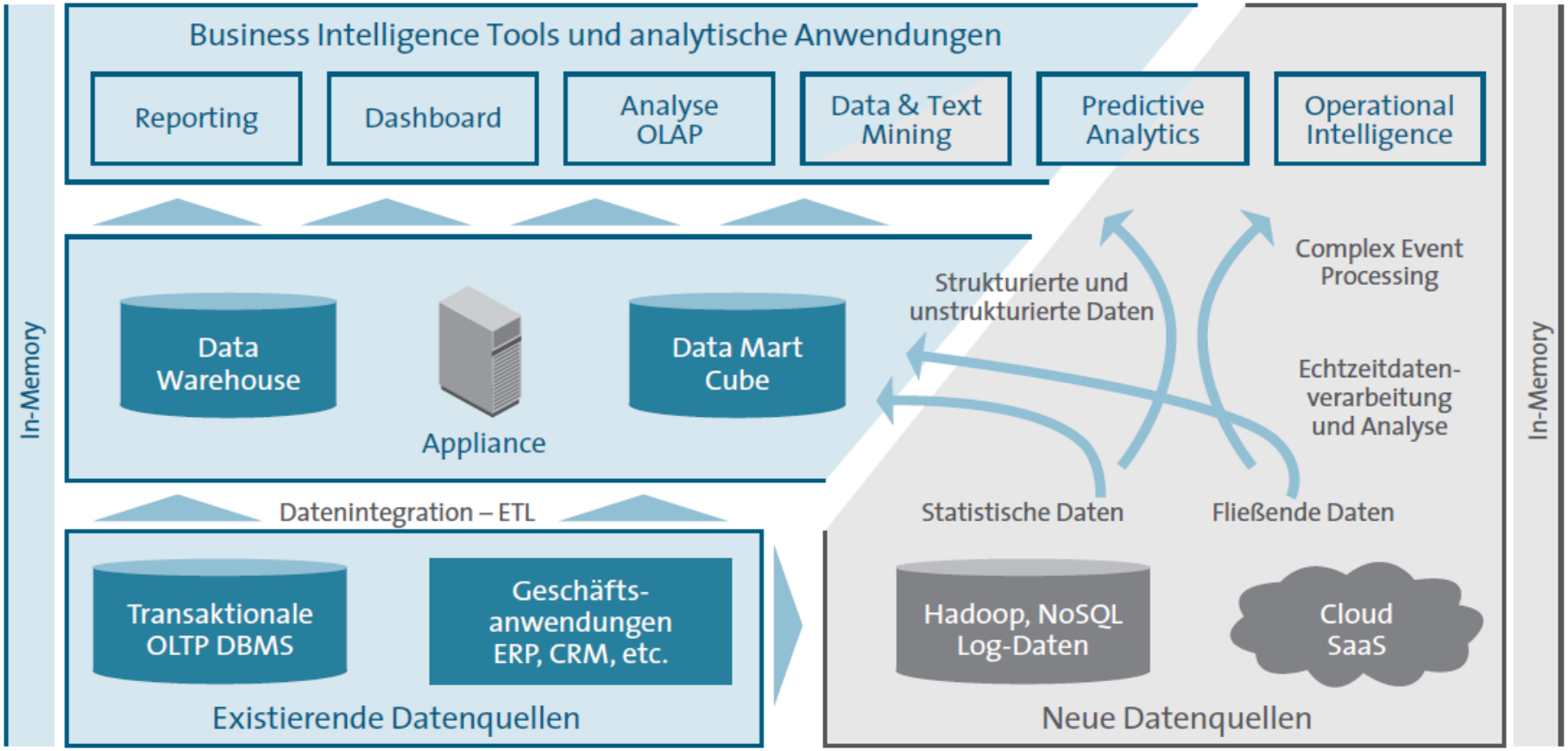


Abbildung 11: Integrierte Anwendungslandschaft mit traditionellen Systemen und Big-Data-Lösungen

Quelle: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM, Berlin 2012, S. 28

Was macht Big Data aus

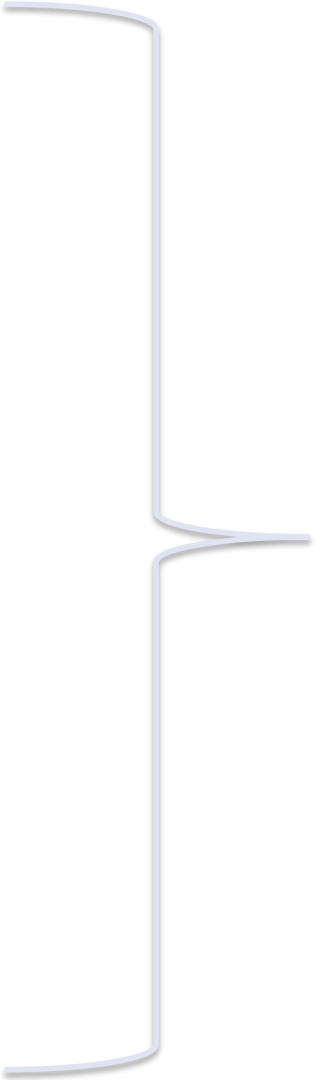
Rechenleistung

Speicherressourcen

Sensoren

Software

Vernetzung



Big Data

Ein erstes Beispiel

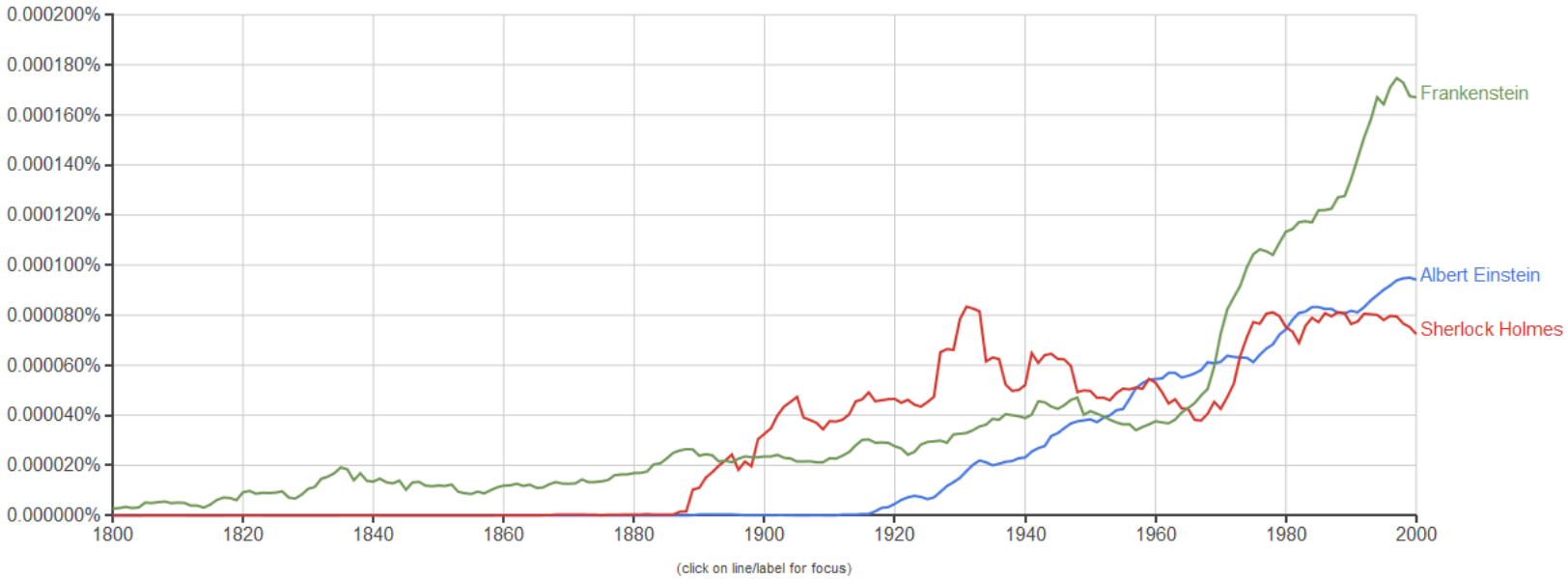
Google books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 . Search lots of books

Teilen 7 Tweet

Embed Chart



Search in Google Books:

1800 - 1943 1944 - 1986 1987 - 1991 1992 - 1996 1997 - 2000 albert einstein English

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

- **Big Data Technologien und Datenbanken – das Ende von SQL?**

Hadoop – ein Big Data Standard

- Ein in Java programmiertes Open Source Framework, dass horizontal auf bis zu mehreren tausend Servern skaliert werden kann. Es ist fehlertolerant gegen Serverausfälle.
- Die beiden wesentlichen Bestandteile von Hadoop sind das Hadoop Distributed File System (HDFS) und das Map Reduce Framework.

Performance ist ein entscheidender Faktor

Standard-RDBS (Oracle, MySQL, MS-SQL etc.):

langsam, u. a. da sie üblicherweise das ACID-Prinzip (Atomicity, Consistency, Isolation und Durability) unterstützen

Alternative:

NoSQL – Datenbanken (Not only SQL)

Unterstützen i. d. R. kein ACID, deutlich schneller

Das CAP-Trilemma:

C (Consistency) : Alle DB-Knoten haben zur selben Zeit die selben Daten

A (Availiability): Alle DB-Anfragen werden beantwortet

P (Partition Tolerance): Kein Ausfall bei Verlust von Daten, Partitionen oder Netzwerkknoten

Ein DBS kann nur zwei dieser drei Eigenschaften unterstützen:

RDBS unterstützen üblicherweise C und A,

NoSQL-DBS üblicherweise A und P

Typen von NoSQL-DBS

- Volltext-
- Dokumenten-
- Colum-Store-
- Graphen-
- Key-Value-

} Datenbanken

NoSQL DBs sind u.a.:

CouchDB, MongoDB, Redis, Hypertable, …

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Bulimie

Bulimie

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- **Die Bulimie oder lateinisch Bulimia nervosa ist eine Essstörung, die durch wiederkehrende Heißhungerattacken gekennzeichnet ist. Das Wort "nervosa" deutet auf die psychische Komponente der Bulimie hin.**
  (Quelle: http://flexikon.doccheck.com/de/Bulimie, abgerufen 16.12.2014)

Wie äußert sich eine Bulimie:

- Heißhunger – schneller Verzehr großer Nahrungsmengen
- Kontrollverlust während des Essanfalls
- Der Heißhunger tritt über einen längeren Zeitraum mehrfach (pro Woche) auf
- Der Betroffene kann trotz Anstrengung mit seinem gestörten Essverhalten nicht aufhören
- Die Krankheit wird verheimlicht

Und wenn Daten die Nahrung sind…

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Heißhunger – schnelle Aufnahme großer Datenmengen**
- **Kontrollverlust während der Datensammlung**
- **Die Datensammelwut tritt über einen längeren Zeitraum häufig auf**
- **Der Betroffene kann trotz Anstrengung mit seinem gestörten Datensammelverhalten nicht aufhören**
- **Die „Krankheit“ wird verheimlicht…**

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Wissen(-serweiterung)

Wissen – Was ist das überhaupt

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Ein erster Versuch:

Was ist Wissen - Frage an Wikipedia

Natürlich (!?) ist das nicht die gesuchte Antwort sondern:

Ich weiß etwas, wenn ich eine wahre und gerechtfertigte Meinung darüber habe.
(geht schon zurück auf Platon)

Wissen als gerechtfertigter wahrer Glaube

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- **Person *P* weiß, dass Aussage *a* gilt**

genau dann, wenn

- ***a* wahr ist**
- **Person *P* glaubt, dass *a***
- **Person *P* hat gute Gründe, *a* zu glauben**

Aber: 1963 veröffentlichte der amerikanische Philosoph Edmund Gettier in seinem berühmten Artikel Is Justified True Belief Knowledge? Gegenbeispiele.

Was ist Wissen überhaupt?

Wissen – ein kleinster gemeinsamer Nenner

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- **Die Basis von Wissen sind Informationen**
- **Die Informationen müssen widerspruchsfrei, nachvollziehbar und überprüfbar sein**
- **Das daraus resultierende Wissen muss mit der Wahrnehmung übereinstimmen**

Welche Implikationen ergeben sich daraus für Big Data?

- **Ist in der Wissenschaft keine Theorie mehr notwendig, da neue Antworten durch die Analyse der sehr großen Datenmengen erfolgen kann?**

Beispiele (wo z. T. die Theorie fehlt):

- Vorhersage von Grippewellen
- Vorhersage von Schwangerschaften durch Analyse des Kaufverhaltens von Kundinnen
- Smart Grid (intelligentes Stromnetz)
- Smart City

• Einige Charakteristika von Big Data:

- Riesige Datenmengen
- Keine repräsentativen Stichproben
- Kontexte werrden z. T. ebenfalls aus Datenbankabfragen abgeleitet

Aber diese Analysemethoden liefern teilweise erstaunliche Ergebnisse!

- **CAPTCHA und ReCAPTCHA**

- CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart) dient der Unterscheidung von Computern und Menschen.*
- ReCAPTCHA ist ein CAPTCHA-Dienst und wird zum Digitalisieren von Büchern und Zeitschriften verwendet.

Höhere Sicherheit, dass der Benutzer einer Website ein Mensch ist und zusätzlich der Effekt, dass eine Vielzahl von Büchern digitalisiert werden können, an denen OCR-Programme scheitern.

- Es geht dabei nicht wesentlich um die massive open online – Kurse (MOOC´s) , wenn von Big-Data Methoden im E-Learning – Bereich gesprochen wird, denn MOOC´s sind teilweise leider „nur“ massiv verteilte „Frontalunterricht“ – Szenarien was die Lehrmethoden betrifft.
- Entscheidend sind die Daten, die durch solche Methoden erzeugt werden

- **Bei Lernszenarien geht es immer noch um die Beantwortung der Frage: „Wie lernen Menschen am besten?“**
- **Wichtiger ist aber die Beantwortung der Frage: „Wie lernt jeder einzelne Mensch?“**
- **Hier können Big Data – Verfahren zum Nachvollziehen der Vorgehensweise der Lernenden verwendet werden**

- Probabilistische Vorhersagen
- Adaptives Lernen
- …

Beispiele

- School of One
- Khan University
- duolingo

Resumée

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

„Wenn große Datenmengen kontrolliert und bedacht gesammelt und verarbeitet werden, können Ergebnisse erzielt werden, die weitreichende positive Folgen für die Datensammler und die Betroffenen haben.

Big Data kann somit eine große Chance sein!

Aber: Wer Datenmengen unkontrolliert sammelt und vorschnell vermeintliche Ergebnisse veröffentlicht, der wird i. A. keine sinnvollen Ergebnisse erzielen.

(Daten-)bulimie

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

„Wer frisst und kotzt, hat nicht das Leben gefunden, das ihm schmeckt!“

(Situationsbeschreibung einer Bulimie-Betroffenen)

Vielen Dank und schöne Feiertage!

ZIM-Talk – http://www.uni-due.de/zim/zim-talk

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- 24.10.2014 - Andreas Bischoff
  Schutz der Privatsphäre auf Ihrem Smartphone
- 21.11.2014 – Burkhard Wald:
  Sag (nicht) wer Du bist – Über Authentifizierung
- 19.12.2014 – Andreas Michels
  Big Data – Neue Wege zur Wissenserweiterung
- 23.01.2015 – Holger Gollan
  Höher, Schneller, Weiter! Oder Die Grenzen des Wachstums
- 20.02.2015 – Marius Mertens
  Super rechnen ohne Superrechner
- 27.03.2015 – Daniel Biella/Malte Hermsen
  DevOps & QA in der Praxis

14:00 Uhr
Duisburg LE 105

ZIM
Zentrum für Informations- und Mediendienste