



HTW Chur
Hochschule für Technik und Wirtschaft

Fachhochschule Ostschweiz
University of Applied Sciences



Big & Smart Data



Prof. Dr. Bernard Bekavac

Schweizerisches Institut für Informationswissenschaft SII
Studienleiter Bachelor of Science in Information Science

bernard.bekavac@htwchur.ch

Quiz

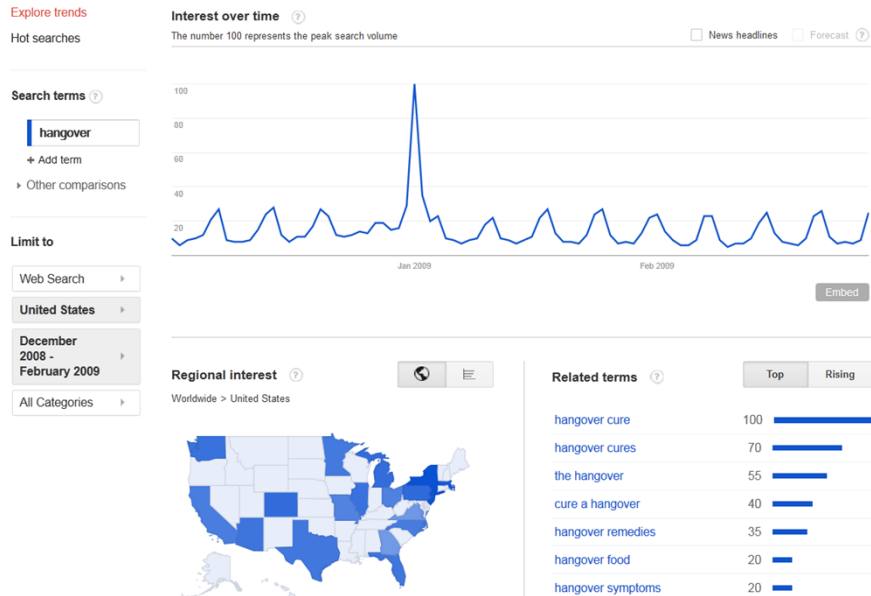
An welchem Wochentag wird in den USA am meisten nach dem Begriff "hangover" gesucht?

- Montag
- Dienstag
- Mittwoch
- Donnerstag
- Freitag
- Samstag

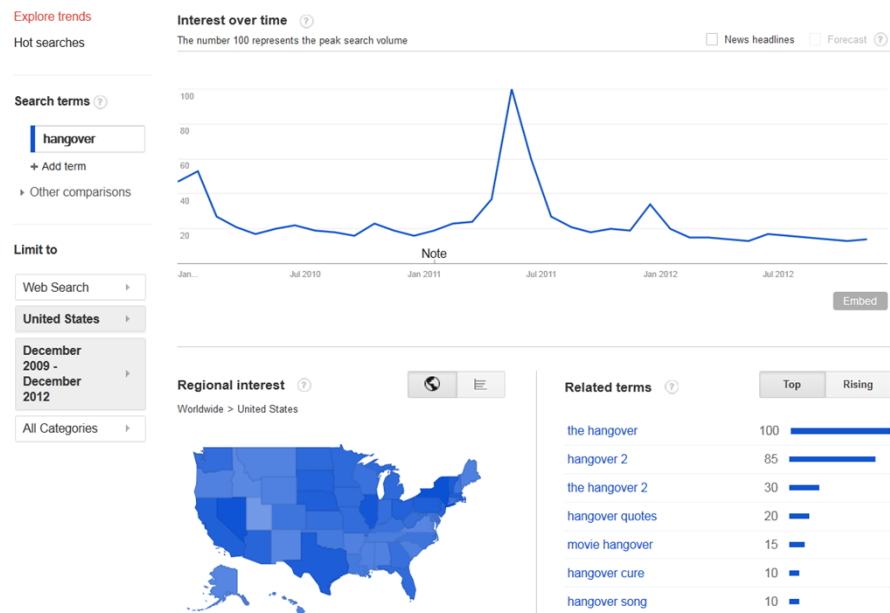
Google Search

I'm Feeling Lucky

Google trends: "hangover"



Google trends: "hangover"

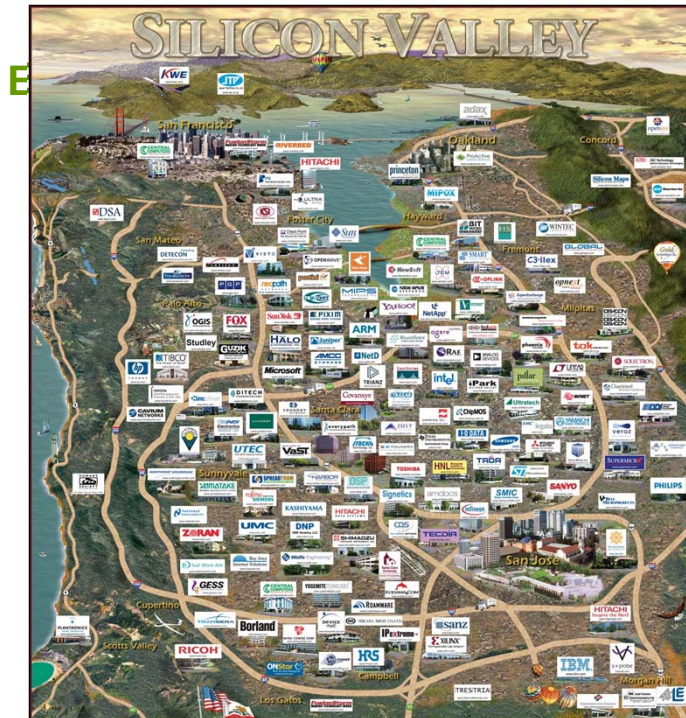


Übersicht

- Ist Data Science ein (oder das) "Big Thema"?
- Was ist Big Data? Was macht man damit?
- Ist das Thema wirklich so relevant?
- Was ist Smart Data?

Exkurs: California / Gold Rush





Exkurs: California / Data Rush



Exkurs: iSchool, Berkeley, Spring 2012

Thought Leaders in Data Science and Analytics
Spring, 2012, University of California, Berkeley

January 25	Dr. David Steier, Director	Deloitte
February 1	Dr. Gary Class, SVP, Internet Strategy	Wells Fargo
February 8	Dr. David Joffe, Managing Director, and David Joa, VP	Bank of America
February 15	Dr. Jimi Shanahan, Director Raj Pai, CEO, Jay Bala, CTO Dr. Lokendra Shastri, VP, Analytics	D&B, Adobe Claritics Infosys
February 22	Dr. Xuhui Shao, VP, Engg, User Data Analytics	Yahoo
February 29	Dr. Sam Yen, SVP, Chairman's Office	SAP
March 7	Dr. Lucian Vlad Lita, VP Dr. Anand Rajaraman, SVP	BlueKai Data Exchange Walmart.com
March 14	Dr. S. "Muthu" Muthukrishnan, Director	Rutgers and Google
March 21	Bob Page, VP, Analytics Platform	eBay
April 4	Itamar Rosenn, Cameron Marlow, Director	Facebook
April 11	Dr. Ashwin Ram, Senior Manager Adeyami Ajao, CEO	PARC Identified.com
April 18	JP Rangaswamy, Chief Scientist Dr. Yan Chow, Director Dr. Paul Rejto, Director	Salesforce.com Kaiser Permanente Pfizer
April 25	Dr. Anil Kamath, VP Dr. Gautam Shroff, Chief Scientist, VP	Adobe TCS

Exkurs: iSchool, Berkeley, Spring 2012

The screenshot shows the UC Berkeley School of Information website. The header includes the university logo and navigation links: UC BERKELEY, ALUMNI NETWORK, CONTACT US, SUPPORT THE I SCHOOL, INTRANET, WIKI, BLOGS. A search bar is visible in the top right. The main content area is titled "DataEDGE" with the tagline "educating a new generation of data-savvy professionals". Below this, it announces the "DATAEDGE CONFERENCE" on Thursday, May 31, 2012, from 4:00 pm to Friday, June 1, 2012, at 6:00 pm, at UC Berkeley. The page also includes a sidebar with a navigation menu (ABOUT, NEWS & EVENTS, PROGRAMS, COURSES, PEOPLE, CENTERS & CLINICS, RESEARCH, ADMINISTRATION) and a right sidebar with a "SUBSCRIBE TO EMAIL ANNOUNCEMENTS" section and a "VIDEO" section.

Übersicht

- Ist Data Science ein (oder das) "Big Thema"
- **Was ist Big Data? Was macht man damit?**
- Ist das Thema wirklich so relevant?
- Was ist Smart Data?

Definition Big Data

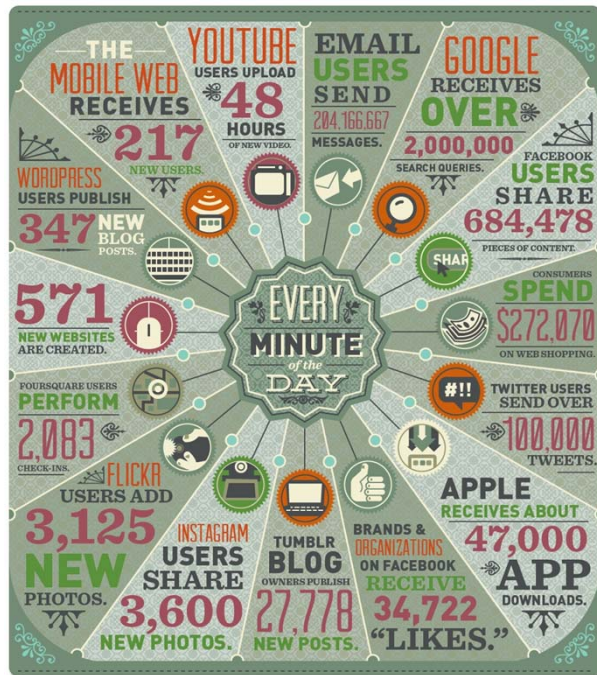
Als Big Data werden besonders große Datenmengen bezeichnet, die mit Hilfe von Standard-Datenbanken und Daten-Management-Tools nicht oder nur unzureichend verarbeitet werden können. Problematisch sind hierbei vor allem die Erfassung, die Speicherung, die Suche, Verteilung, **Analyse und Visualisierung** von grossen Datenmengen. Das Volumen dieser Datenmengen geht in die Terabytes, Petabytes, Exabytes und Zettabytes.

de.wikipedia.org

Big Data

- Petabytes/Exabytes an Daten
- Millionen/Milliarden von Benutzern
- Milliarden/Trilliarden an Datensätzen
- Unstrukturiert
- Fehlender Kontext

Quelle: DOMO Inc.



Big Data

- 10 Millionen Datenquellen
- Datenproduktion 14h/Tag an mehr als 7 Monaten/Jahr
- 6 Petabytes im Jahr



CERN

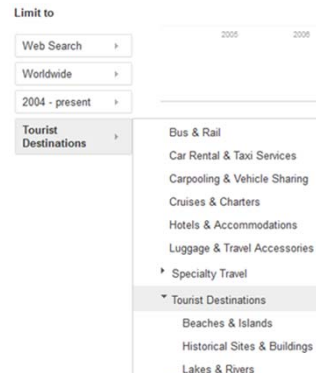
Big Data Analyse: Google prediction

Hong Kong Tourism Board Monthly visitor arrival summary

- Monatliche Besucherstatistiken mit 4-wöchiger Verzögerung
- Angabe von Herkunftsländern
- Daten von 2004-2008

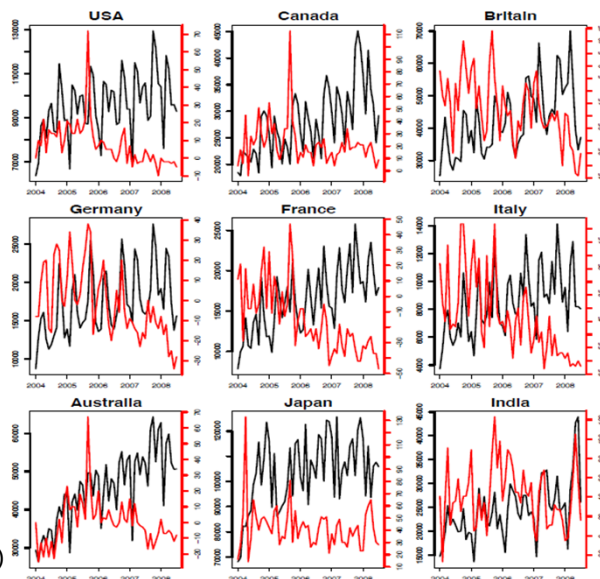


Google Trends Travel by category



Big Data Analyse: Google prediction

Besucherstatistiken
vs. Google Trends



(aus Choi/Varian 2009)

Big Data Analyse: Google prediction

- Big Data alleine reicht nicht aus!
- (Statistische) Analysen notwendig und
- geeignete Modellbildung (für jeden Anwendungsfall)

Bsp.:

Verwendung des *Seasonal Autoregressive Models* (AR)

$$y_t = b_1 y_{t-1} + b_{12} y_{t-12} + b_0 x_t + e_t$$

y_t = Besucher aus einem bestimmten Land im Monat t

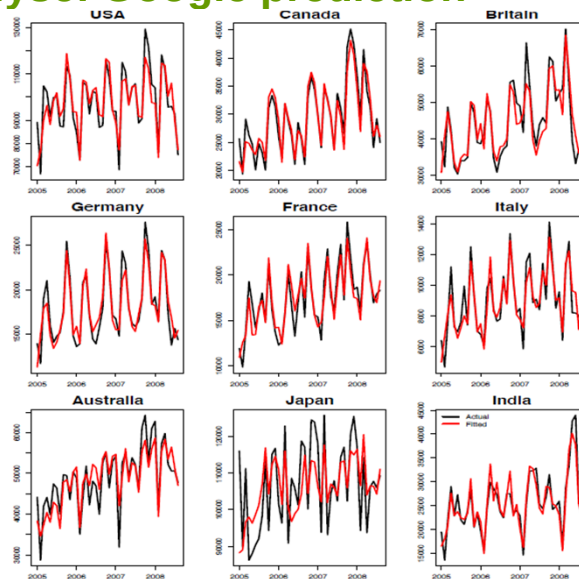
x_t = Google Trends index (Hong Kong) für die ersten zwei Wochen im selben Monat

→ liefert Vorhersagen für reale Besuche 6 Wochen im Voraus

Big Data Analyse: Google prediction

Besucherstatistiken
vs. Google Prediction

(aus Choi/Varian 2009)



Big Data Analyse: Visualisierung



Wikipedia Visualisierung
Wattenberg, Viégas,
Hollenbach, 2007

Big Data Analyse: Visualisierung



de.guttenplag.wikia.com

Übersicht

- Ist Data Science ein (oder das) "Big Thema"?
- Was ist Big Data? Was macht man damit?
- **Ist das Thema wirklich so relevant?**
- Was ist Smart Data?

Ist Big Data wirklich so relevant?

Studie der Zeitschrift Information Week (09/2012)

- Befragung von 257 Unternehmen (USA/Kanda)
- > 50 Mitarbeiter (75% > 500)
- > 500 TB (13% > 10 PB)
- 50% haben einen Umsatz > 100 Millionen US\$,
28% > 1 Milliarde

Ist Big Data wirklich so relevant?

Ergebnisse der Studie:

- > 50% der Unternehmen analysieren nur einen geringen Teil (< 100 TB) der gespeicherten Daten
- Die Datenqualität ist bei den meisten Unternehmen nicht sichergestellt (z.B. Daten aus Geolokalisierung, keine Qualitätsanalyse, keine Stichproben o.ä.)
- Mobile Endgeräte werden überschätzt

Ist Big Data wirklich so relevant?

Ergebnisse der Studie:

- "tools over people" (meist Hadoop, NoSQL, Splunk)
- Es gibt kaum Experten auf dem Arbeitsmarkt ("The sexy job in the next 10 years will be statisticians" - Hal Varian, Google Inc.)
- Output / Effizienz der Analysen fraglich

Übersicht

- Ist Data Science ein (oder das) "Big Thema"?
- Was ist Big Data? Was macht man damit?
- Ist das Thema wirklich so relevant?
- **Was ist Smart Data?**

Definition Smart Data

„Smart data“ – Erstellen

Hier kannst du einen neuen Wikipedia-Artikel verfassen. Eine Anleitung für Anfänger findest du unter [Wikipedia:Dein erster Artikel](#).

The page "[Smart data](#)" does not exist. You can [ask for it to be created](#), but consider checking the search results below to see whether the topic is already covered.

wikipedia.org

Smart Data, intuitiv

- Daten "erklären" sich selbst
- Daten sind mit ihrer Bedeutung (Semantik) versehen
- Auf Basis formaler Beschreibungen der Semantik (OWL, RDF etc.) ist eine automatisierte Verarbeitung der Daten möglich (SPARQL)
- Daten werden auf Basis von URI/HTTP (offen) vernetzt



HTW Chur
Hochschule für Technik und Wirtschaft

Fachhochschule Ostschweiz
University of Applied Sciences



Vielen Dank für Ihre
Aufmerksamkeit und viel
Spass bei den nachfolgenden
Vorträgen