Ludwig-Maximilians-Universität München
Institut für Statistik
Seminar: Wissenschaftstheorie für Statistiker Teil II

# Big Data
## Eine Herausforderung für die wissenschaftliche Theorie

Mila Petkova

May 17, 2014

# Gliederung

# Viele Daten, aber wie viel?

Table: Datenvolumen

| | non-digital | digital | |
|---|---|---|---|
| 2000 | 75% | 25% | |
| 2007 | 7% | 93% | 300 EB |
| 2013 | <2% | 98% | 1200 EB |

| | |
|---|---|
| Kilobyte | $1kB$ |
| Megabyte | $1MB = 10^3 kB$ |
| Gigabyte | $1GB = 10^3 MB = 10^6 kB$ |
| Terabyte | $1TB = 10^3 GB = 10^6 MB = 10^9 kB$ |
| Petabyte | $1PB = 10^3 TB = 10^6 GB = 10^9 MB = 10^{12} kB$ |
| Exabyte | $1EB = 10^3 PB = 10^6 TB = 10^9 GB = 10^{12} MB = 10^{15} kB$ |

# Wo haben wir Big Data in Anspruch genommen?

- Google findet die richtige Webseite für uns
- Amazon - die richtige Bücher
- Facebook - die richtige Freunde
- Internet, Smart Phones, GPS-Daten...

# Was wurde mit Big Data gemeint?

> *...the volume of information had grown so large that the quantity being examined no longer fit into the memory that computer use for processing, so engineers needed to revamp the tools they used for analyzing it all.*

(Mayer-Schoenberger and Cukier, 2013, p. 6)

# Weitere Definitionen

> *Daten zu sammeln, Software nach Mustern suchen zu lassen und anschließend aus den Ergebnissen die richtigen Schlüsse zu zeihen, das ist, verkürzt, Big Data.*

(Geiselberger and Moorstedt, 2013, p. 9)

# Weitere Definitionen

*We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:*
*(1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.*
*(2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.*
*(3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.*

(boyd danah and Crawford, 2012, p.663)

# Big Data Geschichten

- Gezielte Werbung
- Google Flu Trends
- Amazon Bücherempfehlungen

# Gezielte Werbung[1]

- Amerikanische Discounteinzelhändler Target
- Ziel: durch das Kaufverhalten zu wissen, ob die Kundin schwanger ist.
- Daten so genau: Target konnte vorhersagen in welchem Monat die Kundin ist.

[1] http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

# Google Flu Trends

- Im Jahr 2009 wurde ein neues Virus entdeckt.
- Gesundheitsbehörde verfügten über verspäteten Daten.
- Google: Versuch durch Analyse der Suchanfragen der Nutzer eine Vorhersage über die Ausbreitung der Schweinegrippe zu machen.

(Geiselberger and Moorstedt, 2013, p. 13)

# Amazon Bücherempfehlungen

- festangestellte Kritiker vs. Computergenerierter Algorithmus
- Ziel: Den Kunden Bücher empfehlen aufgrund von ihrer individuellen Präferenzen

# Die Herausforderung

> *...society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what.*

(Mayer-Schoenberger and Cukier, 2013, p. 7)

# Wie haben wir es immer gelernt?

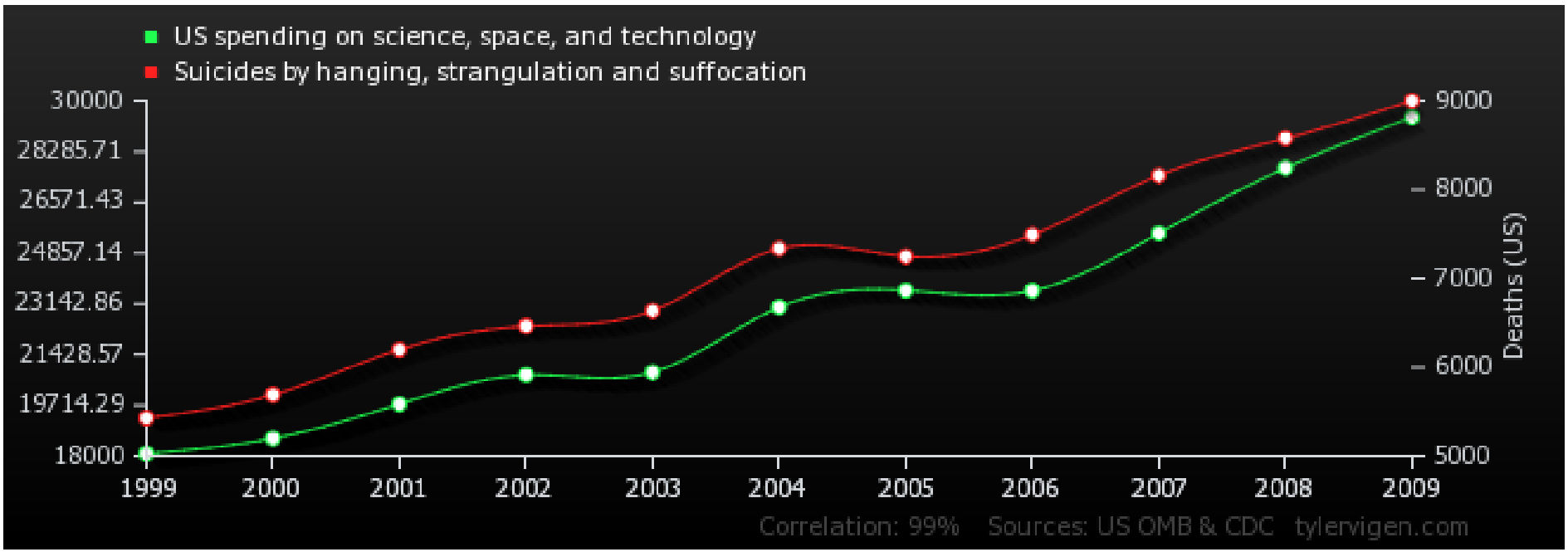
Korrelation $\neq$ Kausalität

> *Kausalzusammenhänge können aber niemals allein durch große Werte eines entsprechenden Zusammenhangsmaßes oder allgemeiner durch eine statistische Analyse begründet werden.*

(Ludwig Fahrmeir and Tutz, 2007, p. 148)

# ...Korrelationen!?!

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation[2]

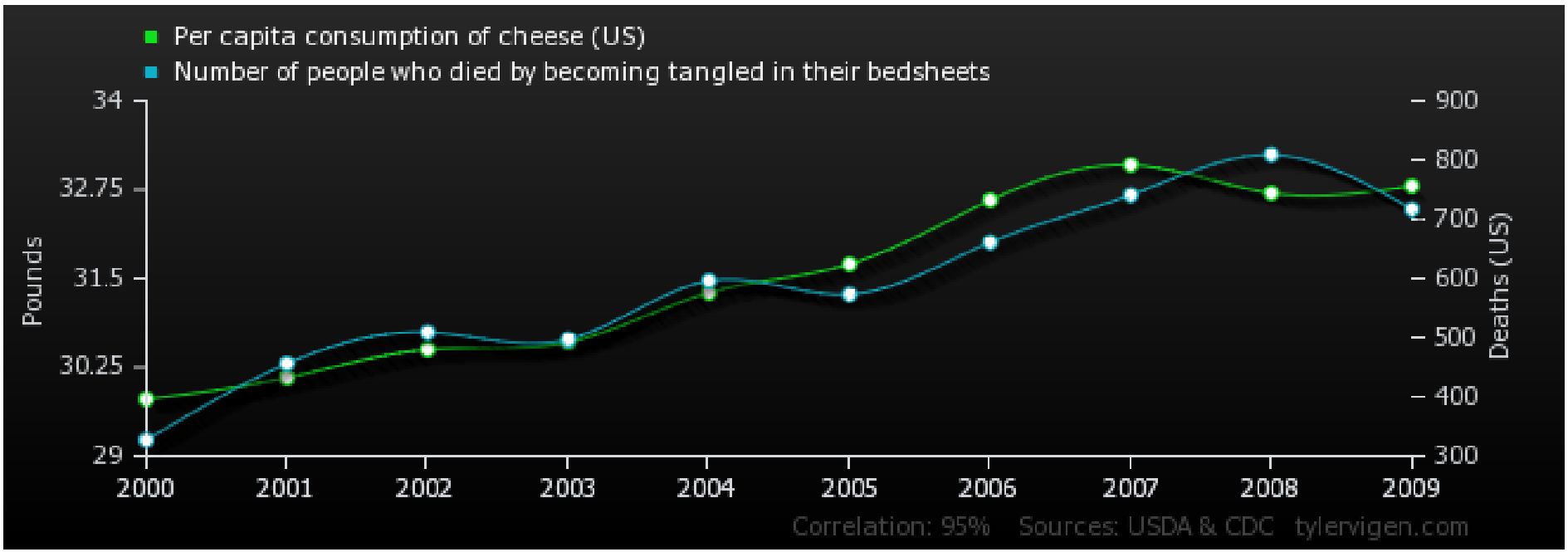



Correlation: 0.992082

[2] http://www.tylervigen.com/

# ...Korrelationen!?!

Per capita consumption of cheese (US)
correlates with
Number of people who died by becoming tangled in their bedsheets[3]




Correlation: 0.947091

[3] http://www.tylervigen.com/

# The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[4]

Author: Chris Anderson
Jahr: 2008
In kürze:

1. "All models are wrong, but some are useful." (George Box, statistician)
2. "All models are wrong, and increasingly you can suceed without them." (Peter Norvig, Google's research director)
3. "What can science learn from Google?"

[4] Anderson (2008)

# All models are wrong, but some are useful

- Theorien werden gebaut, Hypothesen werden hergeleitet, Experimenten werden durchgeführt.
- Wissenschaft basiert auf testbare Hypothesen, die zu bestätigen oder zu falsifizieren sind.
- Theorien sind da um die Welt zu erklären.

*This is the way science has worked for hundreds of years.*

# All models are wrong, and increasingly you can suceed without them

- Korrelation reicht.
- Analysiere die Daten ohne eine Hypothese bereit zu haben.
- Vorteil: durch statistischen Algorithmen zeigen sich Muster, die die Wissenschaft nie gesehen hat oder sogar nie sehen könnte.
- Nachteil: wir sehen **was**, aber nicht **warum**.

# It´s time to ask: What can science learn from Google?

It´s time to ask: What can science learn from Google?

# Big Data verändert die Definition von Wissen

- Vergleich zu Henry Ford's System der Massenproduktion
- Warnung vor einer Automatisierung der Forschung

# Der Anspruch auf Objektivität und Genauigkeit führt in die Irre

- Überquantifizierung der Sozialwissenschaften.
- Wissen wie die Daten zu Stande gekommen sind.
- Gefahr: Muster zu sehen, wo es gar keine gibt

# Mehr Daten bedeutet nicht automatisch bessere Daten

- Twitter als Beispiel
- Twitter ist nicht repräsentative für alle Menschen.
- Person $\neq$ Twitter-Benutzer

# Außerhalb des Ursprungskontext verlieren große Datenmengen an Aussagekraft

- 'Social Graph': mit Hilfe von Diagrammen Verhältnisse zwischen Nutzern darzustellen
- Soziologen und Anthropologen haben auch Soziogrammen oder Verwandtschaftsnetzwerke erstellt
- persönliche Netzwerke vs. artikulierte und Verhaltensnetzwerke

# Nur weil die Daten zugänglich sind, heißt das noch lange nicht, dass es ethisch vertretbar ist, sie auszuwerten

> *...what is the status of so-called public data on social media sites? Can it simply be used, without requesting permission? What constitutes best ethical practice for researchers?*

(boyd danah and Crawford, 2012, p. 671)

- Die ethische Herausforderung
- 1970: Entstehung Institutional Review Boards (Ethikkommissionen)

# Eingeschränkter Zugang zu Daten lässt eine neue digitale Kluft entstehen

> *But who gets access? For what purposes? In what contexts? And with what constraints?*

- Daten-reichen vs. Daten-armen

(boyd danah and Crawford, 2012, p. 674)

# "Fazit"



HARVARD | BUSINESS | SCHOOL

FACULTY & RESEARCH

Article | Harvard Business Review | October 2012

## Data Scientist: The Sexiest Job of the 21st Century


### Abstract

Key to the effective use of big data are the analytical professionals known as "data scientists," who can both manipulate large and unstructured data sources and create insights from them. Data scientists are difficult to hire and retain, but their skills will be necessary to any organization wishing to profit from big data.


**Keywords:** big data; data scientists; business analytics; Data and Data Sets; Mathematical Methods; Jobs and Positions

Format: Print Read Now

**Citation:**

Davenport, Thomas H., and D. J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90, no. 10 (October 2012): 70–76.

# Vielen Dank für die Aufmerksamkeit!

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete, `http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory`. Accessed: 02. Mai. 2014.

boyd danah and Crawford, K. (2012). Critical questions for big data, *Information, Communication & Society* **15 (5)**: 662–679.

Geiselberger, H. and Moorstedt, T. (eds) (2013). *Big Data Das neue Versprechen der Allwissenheit*, Suhrkamp Verlag Berlin 2013.

Ludwig Fahrmeir, Rita Kuenstler, I. P. and Tutz, G. (2007). *Statistik Der Weg zur Datenanalyse*, Springer-Verlag Berlin Heidelberg.

Mayer-Schoenberger, V. and Cukier, K. (2013). *Big Data A revolution that will transform how we live, work and think*, John Murray (Publishers).