Big Data Einführung

Till Hänisch
DHBW Heidenheim, 2015-2020
www.tillh.de/IntroBigDataKurzOnline.pdf

Algorithmen

INTRODUCTION

IN EARLY APRIL 2011, MICHAEL EISEN, an evolutionary biologist at the University of California at Berkeley, logged on to Amazon.com to buy an extra book for his lab. He was after *The Making of a Fly,* by Peter Lawrence, about the genetic development of a fly from a single-celled egg to a buzzing, flying insect. The 1992 book, though out of print, remains popular with academics and graduate students. Eisen was used to paying $35 to $40 for a used copy. But on this day, April 8, there were two established Amazon sellers offering new, unused copies of the book for quite a bit more than he wanted to spend: $1,730,045 and $2,198,177.

Eisen assumed the price was a mistake or a joke; nobody, not even the author, he speculated, would put such a value on the book. He checked back on Amazon the next day to find that, rather than returning to normal, the prices for the book had risen to $2,194,443 and $2,788,233. On the third day, the prices ascended to $2,783,493 and $3,536,675. The escalation continued for two weeks, with the price peaking on April 18 at $23,698,655.93. And a buyer still had to foot a $3.99 shipping bill. The next day, on April 19, the prices for the book fell, settling at $106.

But why had Amazon been selling an arcane book on fly genetics for nearly $24 million? Had it suddenly become hot with billionaire collectors? Did it contain clues to find treasure? Had it become the 1869 Château Lafite Rothschild of books? What had actually happened, in fact, was that the unsupervised algorithms that priced books for the sellers, both of

Steiner, Automate This: How Algorithms Came to Rule Our World, Portfoloi, 2012


SAY BIG DATA
ONE MORE TIME
memegenerator.net

What is a typical size of a relational database?

A:Megabytes

B:Gigabytes

C:Zetabytes

D:Petabytes

Which unit is used to measure transaction throughput?

A: 10,000 TA/min

B: 100,000 TA/sec

C: 10,000 TA/sec

D: 1,000 TA/hour

Which system should be used to search a 100 MB RDF database?

A: relational DBMS

B: Excel

C: MongoDB

D: python script

tion 1

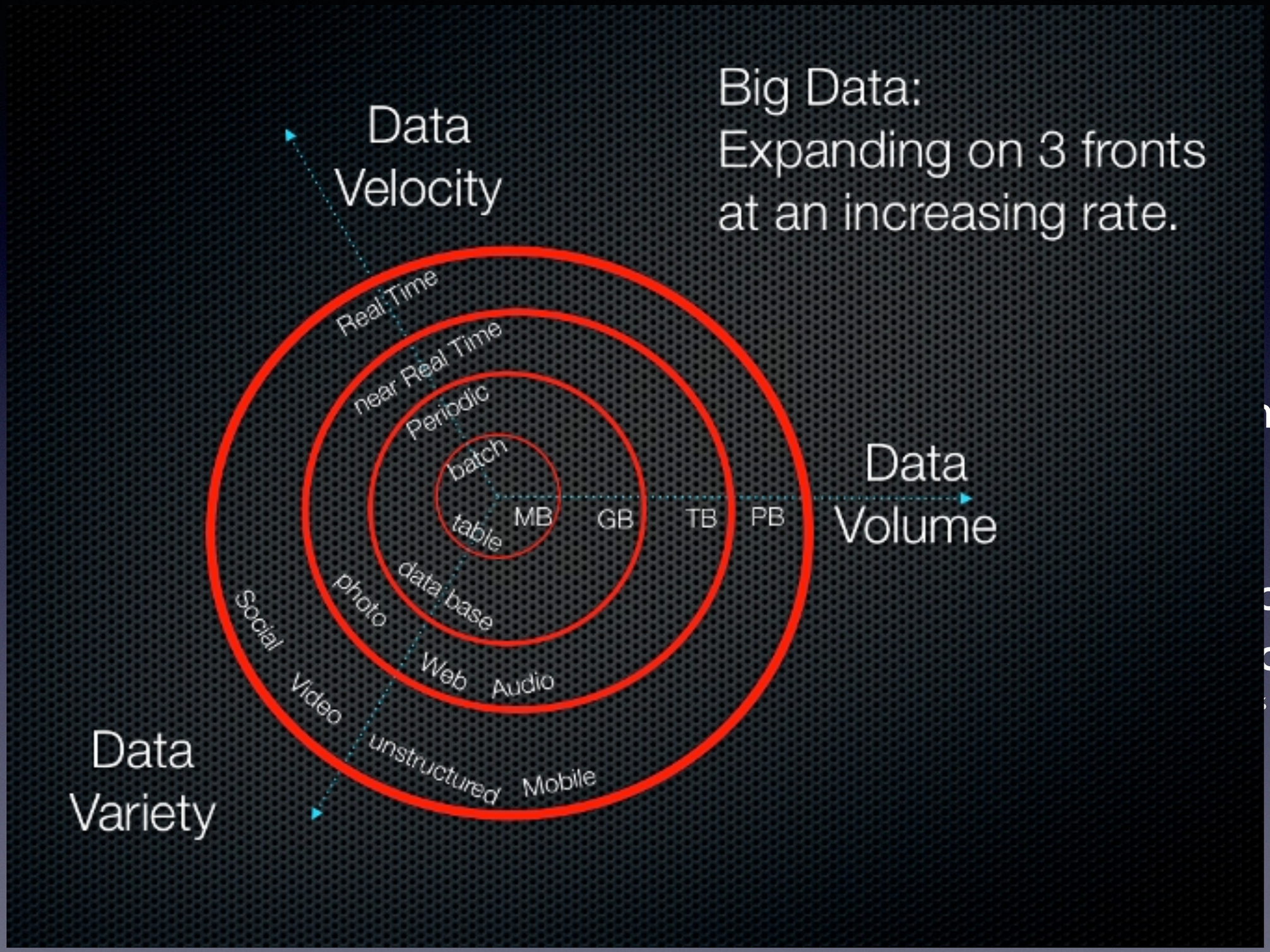
Option 2

Folien von Jens Dittrich

„The case for small data management“

https://www.youtube.com/watch?v=O7Qgo6RSzmE

Definition:



ng than any

ormation
ced
(Gartner

Def. nach Jens Dittrich

4V = Volume large

Variety XML

Veracity crappy data

Velocity frequently appended to

Big Data:

Big Data ist der Übergang

- von qualitativen zu quantitativen Methoden
- vom (abstrahierten) Modell zur Daten über die Realität
- zu immer detaillierteren Daten immer näher am Jetzt
- von Stichproben zu N=alle
- von data base zu data science
- von Deduktion zu Induktion
- von Kausalität zu Korrelation

bei der Bildung von Hypothesen

Repurposing data

Daten für einen anderen Zweck verwenden, als geplant

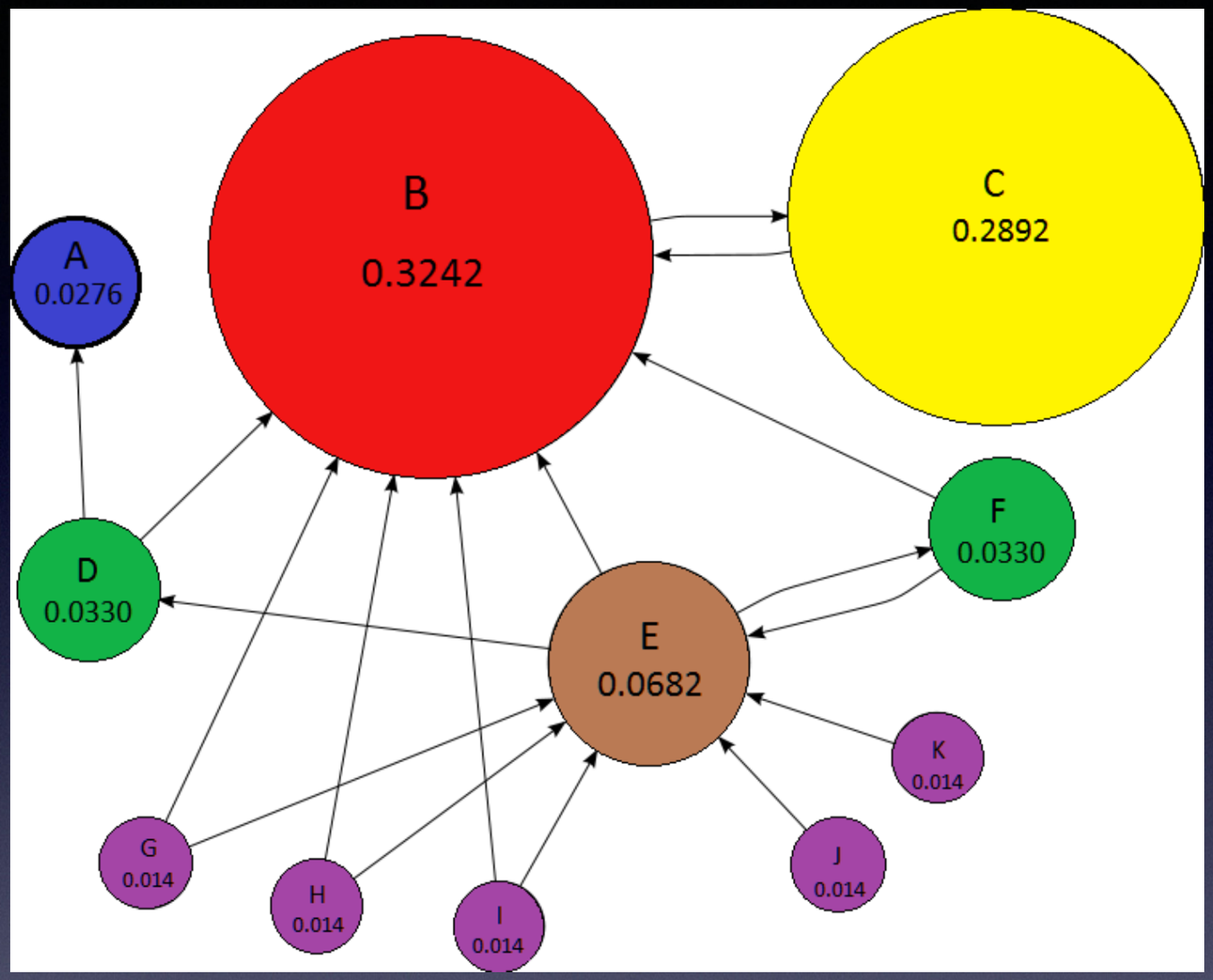
Wie plant man dann ?
Datenmodelle ?
Was wirft man weg ?

—> Nix wegwerfen, alles speichern, man könnt’s ja mal brauchen

präziser:

Eine Stichprobe (ein Teil der Daten) kann nur zweckgebunden ausgewertet werden, hat man alle Daten kann man später andere Fragen stellen als geplant !

Beispiel: Page Rank



$$PR_i = \frac{1-d}{n} + d \sum_{j \in \{1,\ldots,n\}} \frac{PR_j}{c_j}$$

[http://de.wikipedia.org/wiki/Datei:PageRank-Beispiel.png]

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85.
There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))
Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

Beispiel: Collaborative Filtering

- Problem: dynamischer Content soll personalisiert werden, aber wie ?
- Regelbasiert: Benutzer wird durch explizites Abfragen von Präferenzen einer Gruppe zugeordnet (Beruf, Alter, Interessen,...). Nachteil: starre Regeln, Angaben sind subjektiv, können veralten
- content based filtering: Zu bekannten werden möglichst “ähnliche” Inhalte gesucht
- collaborative filtering: Inhalte werden durch andere Benutzer (explizit oder implizit) bewertet, zur Auswahl werden möglichst “ähnliche” Benutzer gesucht (z.B. amazon)

Reality check

Till Haenisch,

Sie suchen Produkte aus der Kategorie Bücher über Business & Karriere? Dann haben wir die folgende Auswahl für Sie.

Bücher über Business & Karriere

Internet der Dinge: Technik, Trends und Geschäftsmodelle
Volker P. Andelfinger, Till Hänisch

Preis: **EUR 34,99** Prime

Weitere Informationen

Auf meinen Wunschzettel

Internet der Dinge: www.internet-der-dinge.de (VDI-Buch)
Hans-Jörg Bullinger, Michael ten Hompel

Preis: **EUR 97,99** Prime

Weitere Informationen

Auf meinen Wunschzettel

› Weitere ähnliche Artikel anzeigen

Predictive Analytics

Vorhersagen

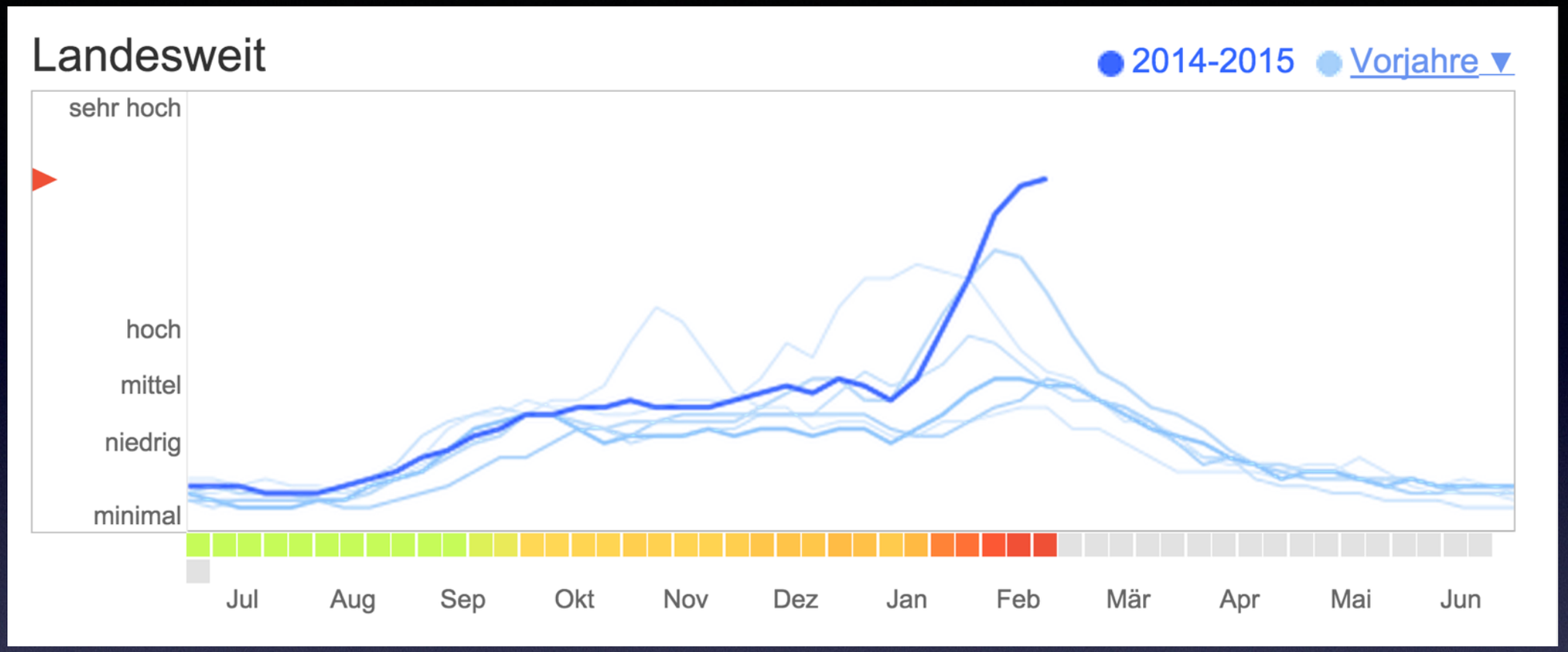
Wird ein Kunde einen Kredit zurückzahlen
→ Scoring, Klassifikation

Wann wird ein Bauteil ausfallen ?
→ Regression

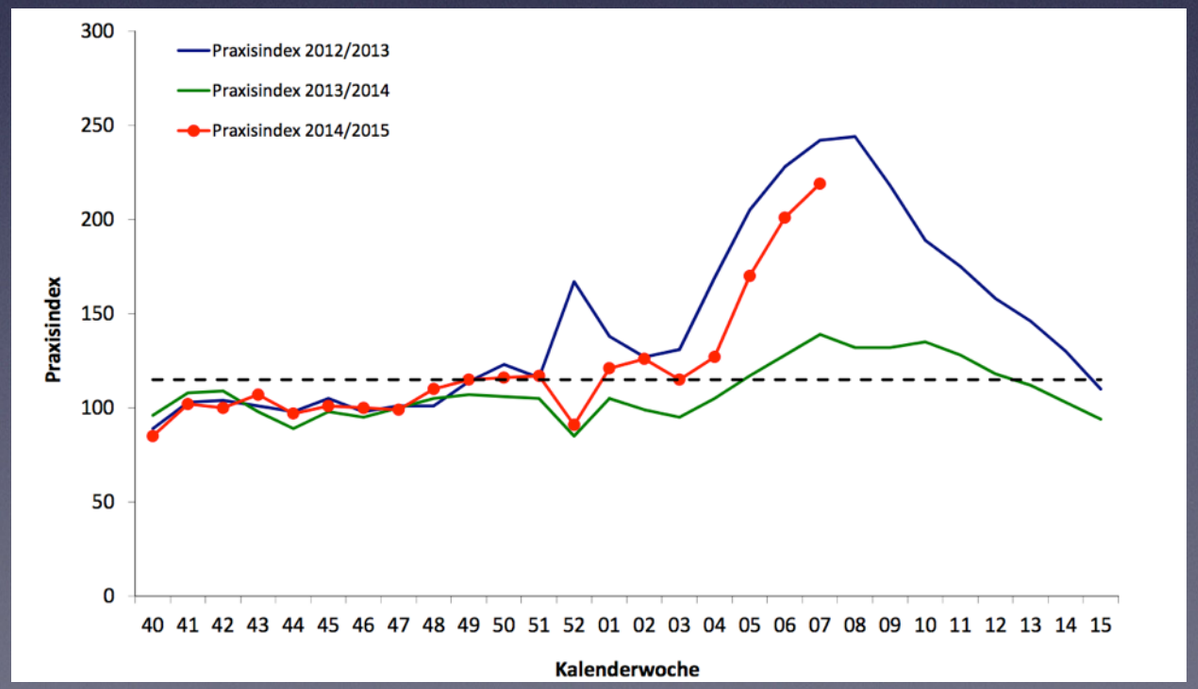
Erkennen von Mustern in Daten, auf deren Basis extrapoliert wird.

„Vorhersagen sind schwierig,
insbesondere wenn sie die Zukunft betreffen“ …

Beispiel: google flu trends



[http://www.google.org/flutrends/de/#DE]



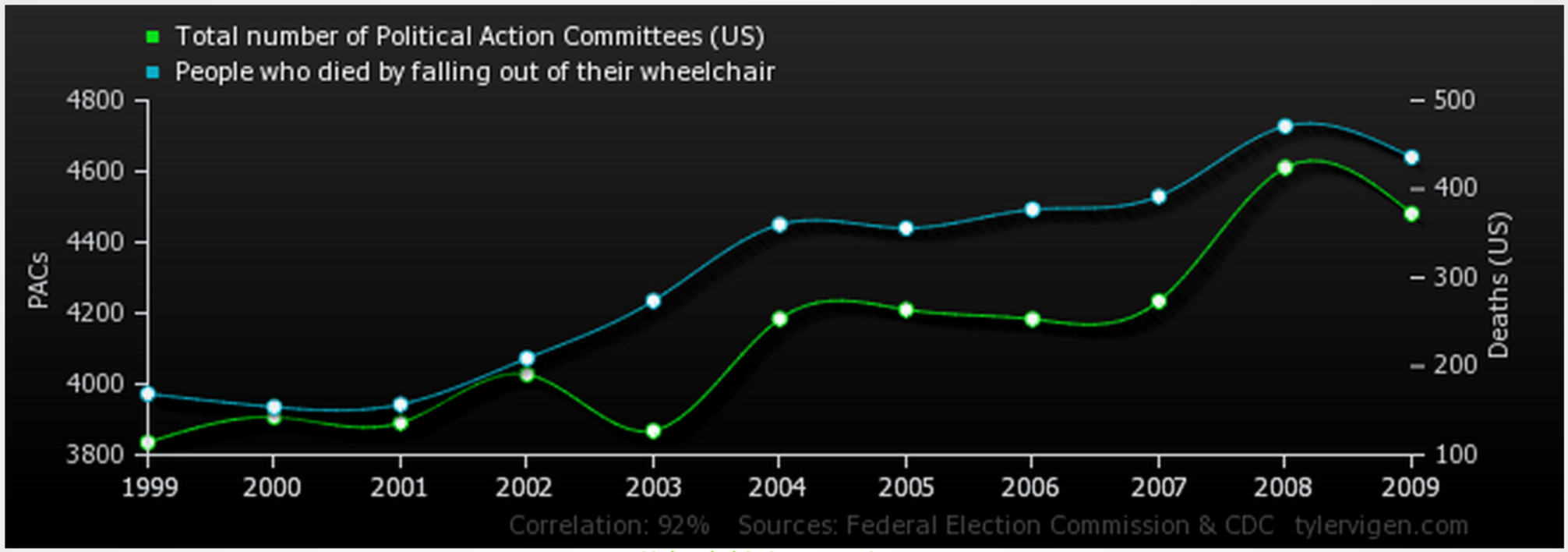
[https://influenza.rki.de/Wochenberichte/2014_2015/2015-07.pdf]

Kausalität

Total number of Political Action Committees (US)

correlates with

People who died by falling out of their wheelchair



Upload this image to imgur

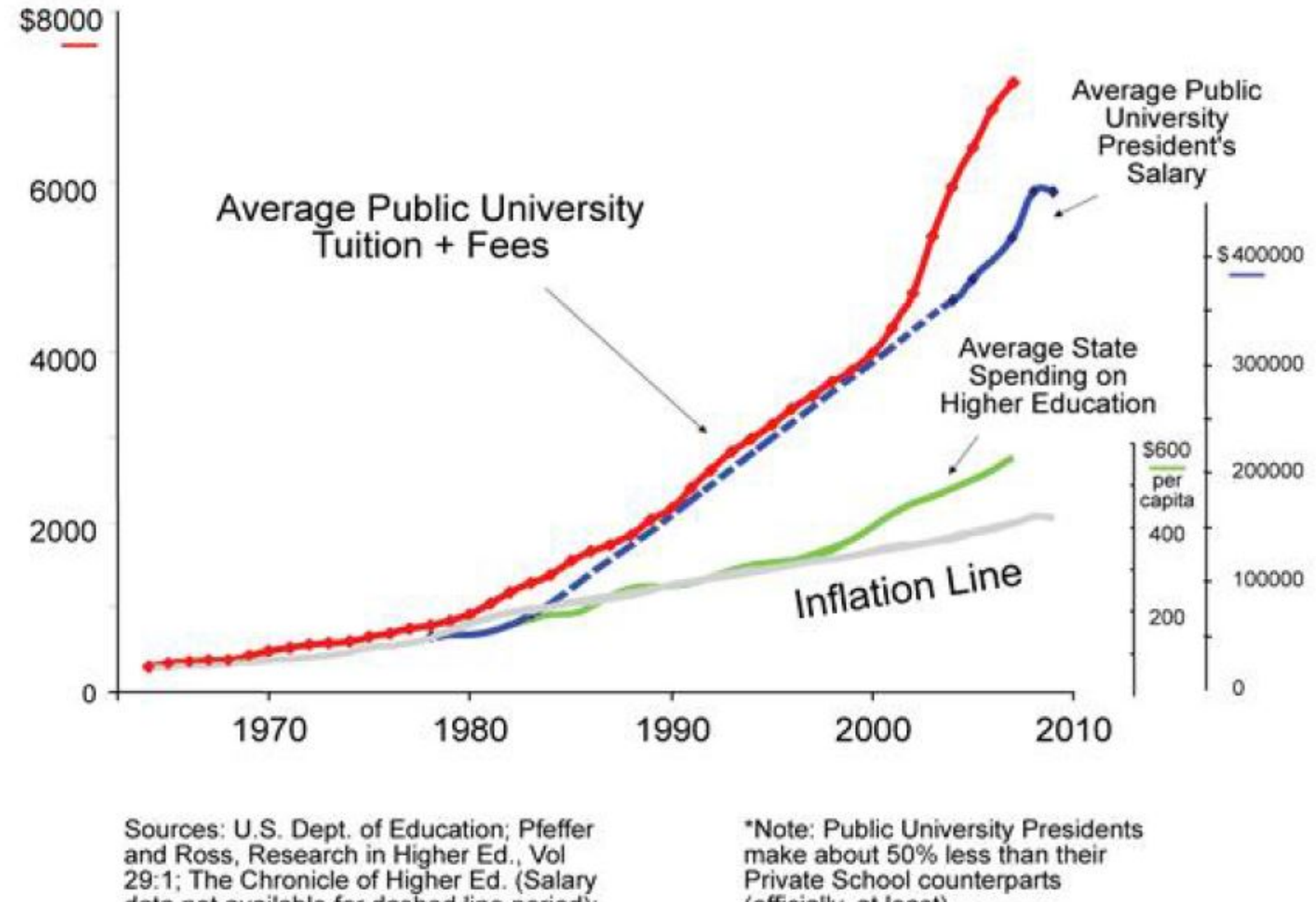
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total number of Political Action Committees (US) *PACs (Federal Election Commission)*	3,835	3,907	3,891	4,027	3,868	4,184	4,210	4,183	4,234	4,611	4,481
People who died by falling out of their wheelchair *Deaths (US) (CDC)*	169	154	157	209	274	360	356	377	392	471	436

Correlation: 0.915876

http://www.tylervigen.com/view_correlation?id=3890

It Takes a B.A. to Find a Job as a File Clerk

By CA



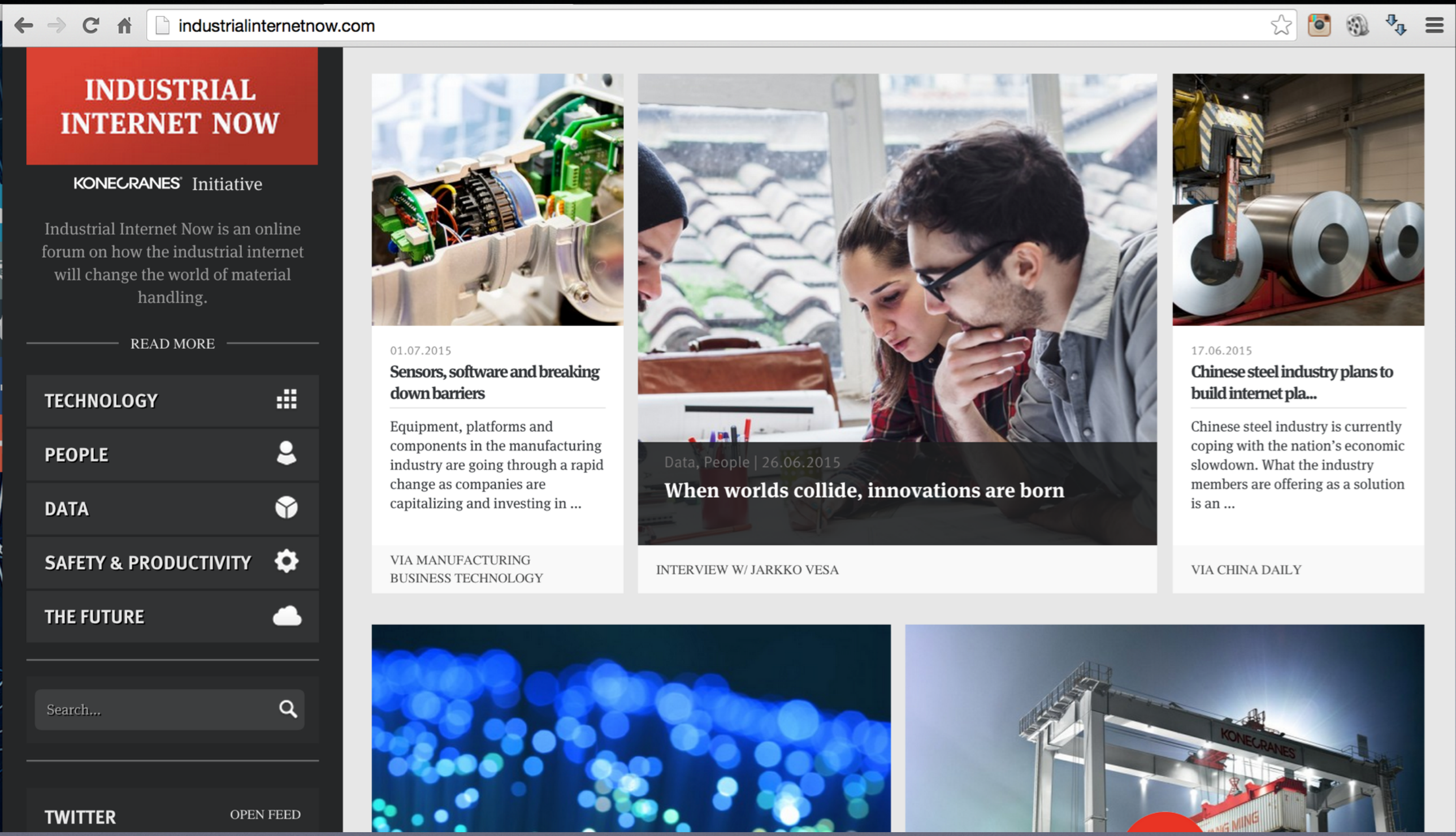
Mega

workers. Rich Addicks for The New York Times

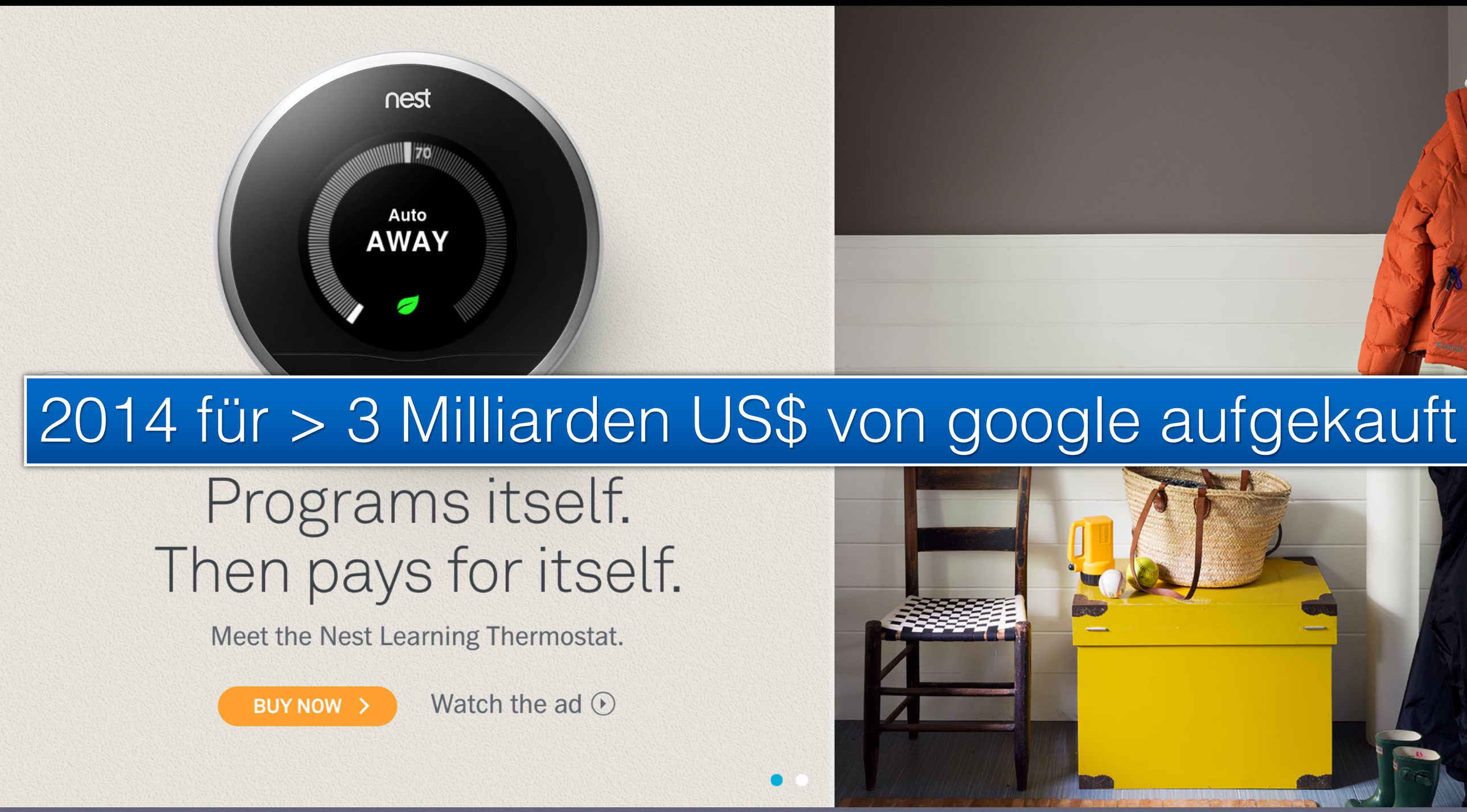
521 gher-e

business/college-degree-required-by-increasing-number-of-

Industrial Internet

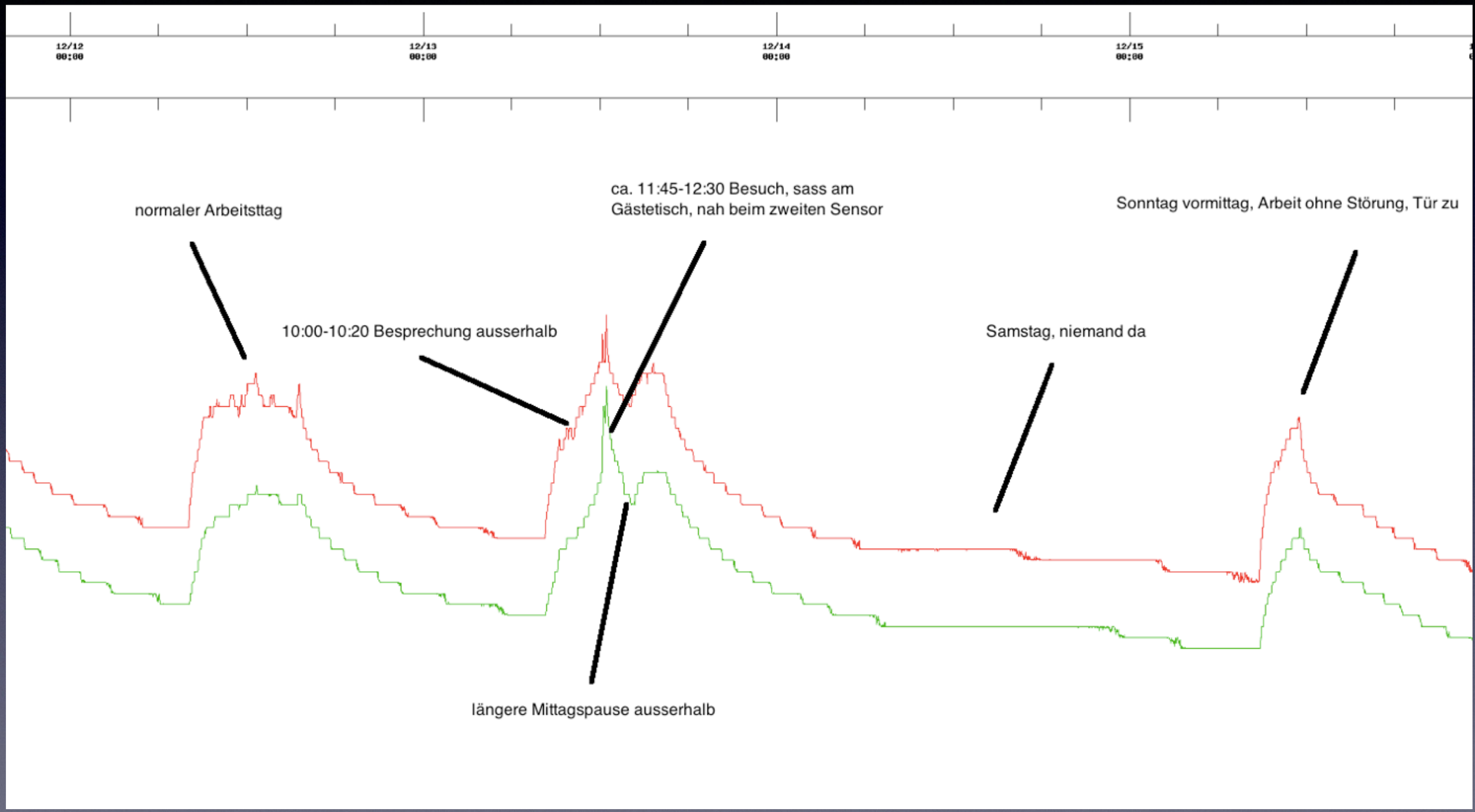


Beispiel: Nest

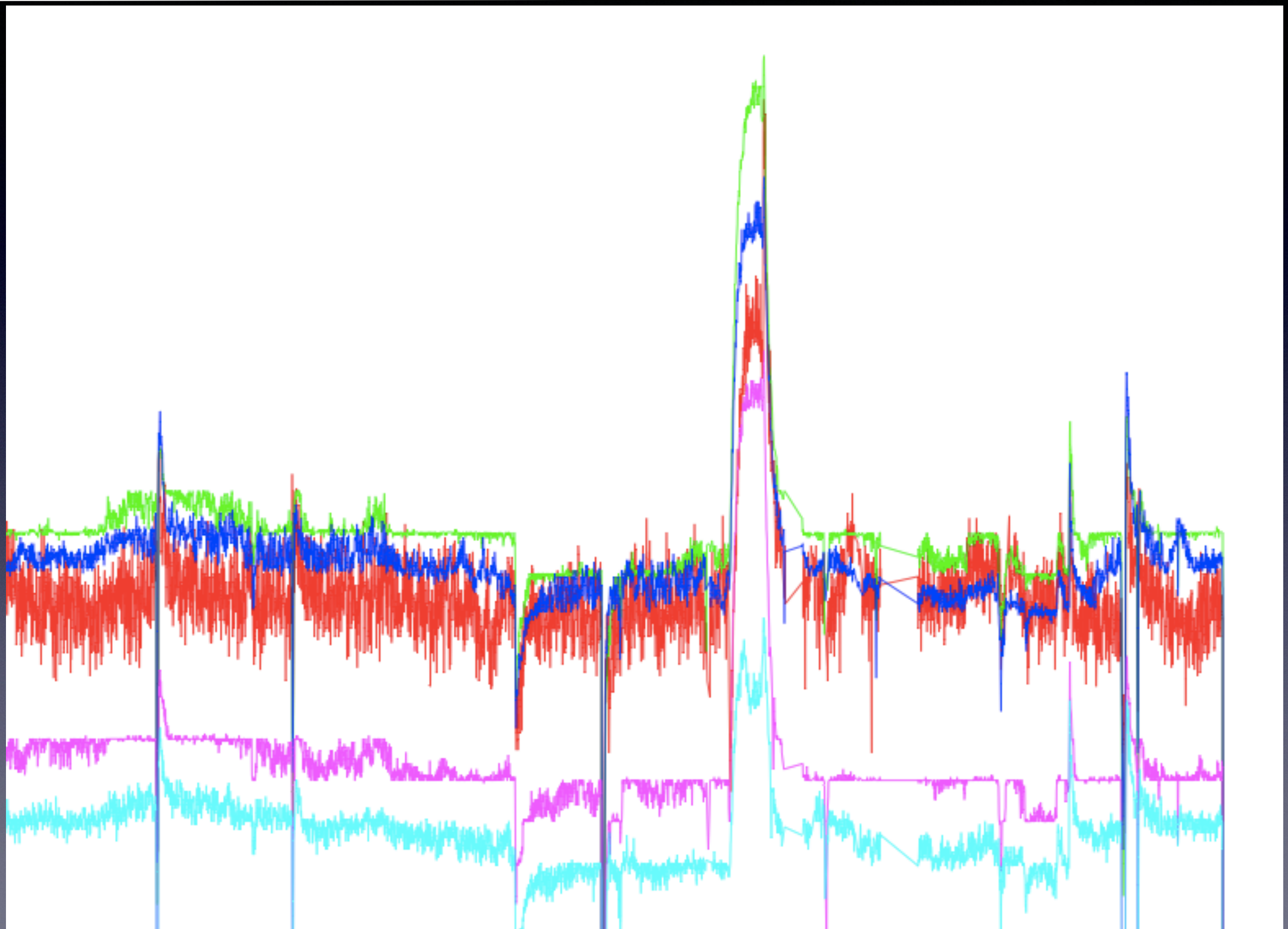


2014 für > 3 Milliarden US$ von google aufgekauft

Temperaturdaten sind vielsagend …



Industrial Internet konkret



Sensornetze:

Smart Bridges

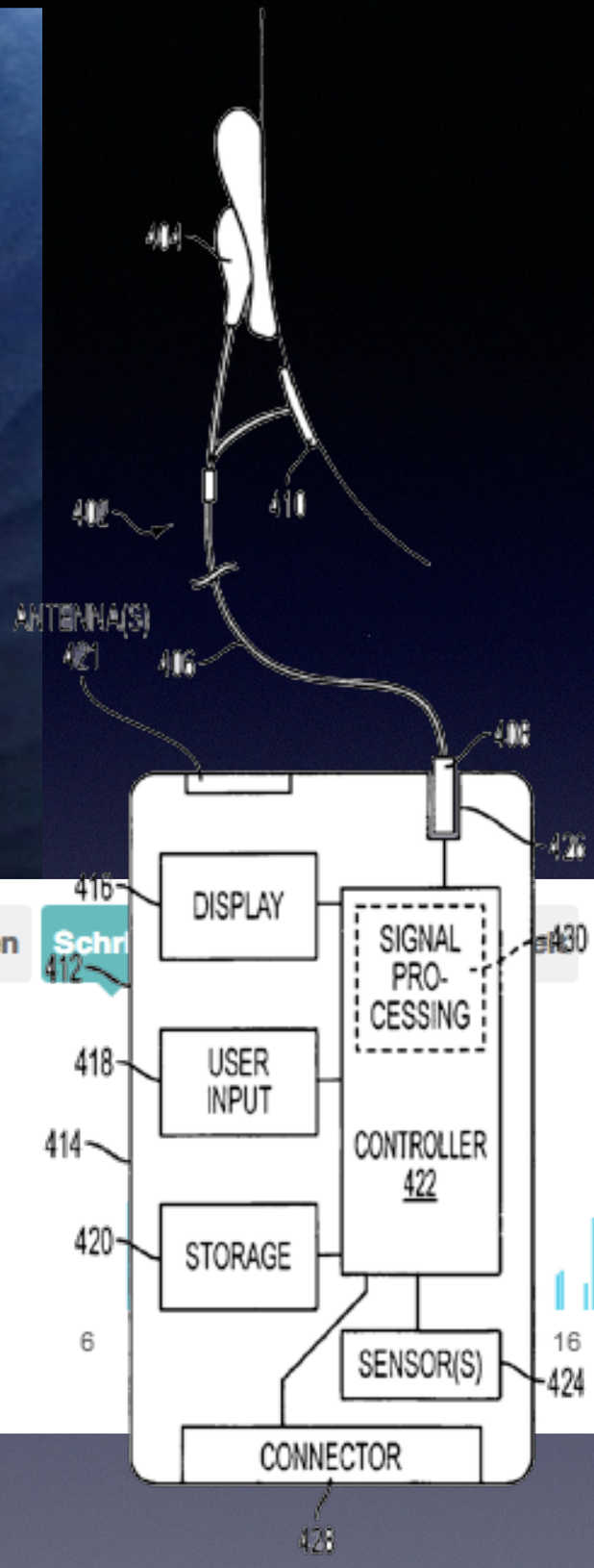
Adding sensor networks to infrastructure will make them cyberphysical systems

By Steven Cherry

Posted 7 Aug 2013 | 15:32 GMT

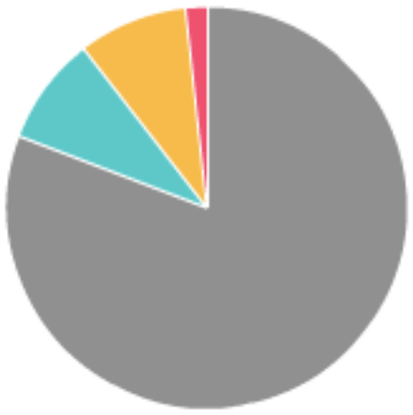
Share | Email | Print

http://spectrum.ieee.org/podcast/at-work/test-and-measurement/smart-bridges



Verbrannte Kalorien | Schritte | Stockwerke | Aktive Zeit

Analyse der heutigen Aktivität (ohne Schlaf)



Im Sitzen
19Std 24Min
(80,8%)

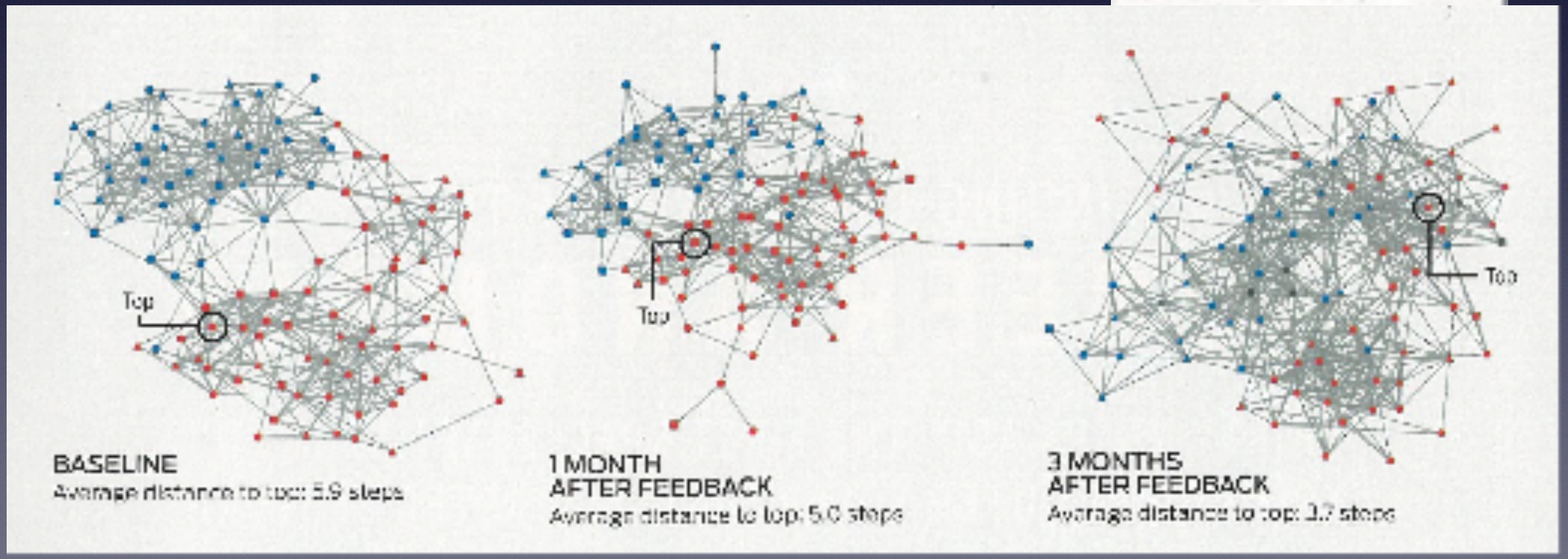
Relativ aktiv
2Std 7Min
(8,8%)

Leicht aktiv
2Std 5Min
(8,7%)

Sehr aktiv
24Min
(1,7%)

Engineering happiness

Engineering health:
Monitoring von Klinikpersonal bei Desinfektion (30.000 Tote jährlich wegen Klinikinfektionen allein in D)



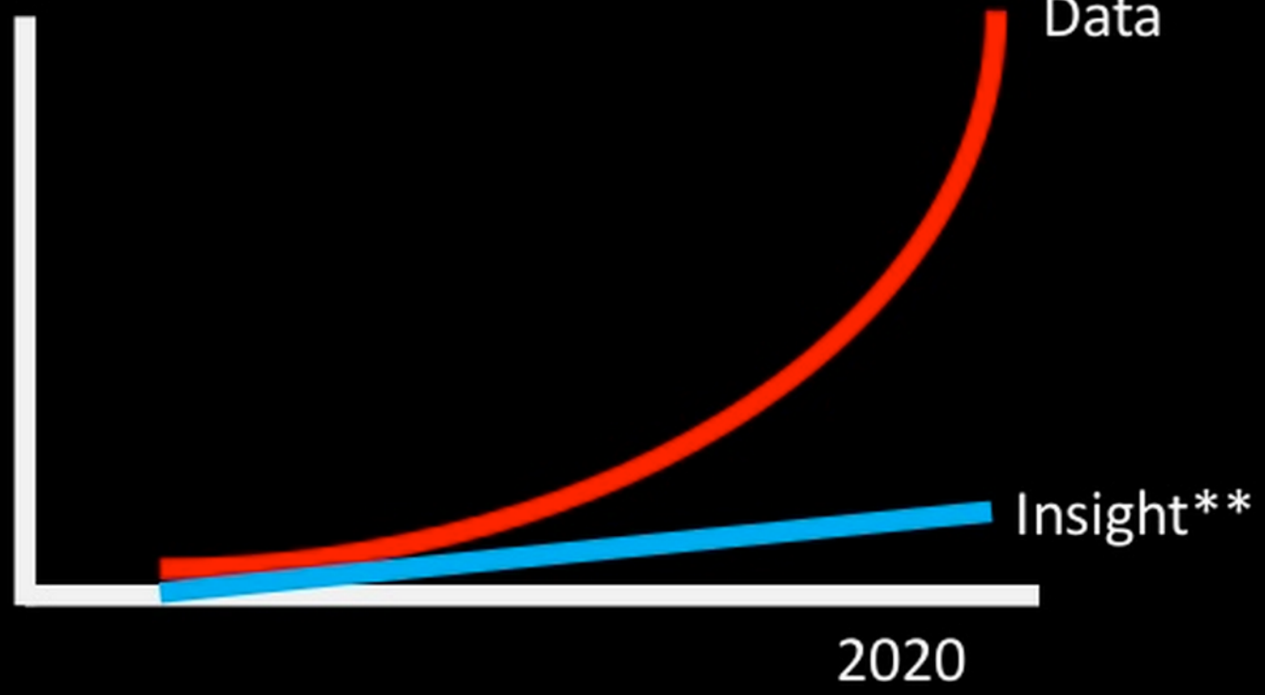
Kazuo Yano et al, Sensing Happiness, IEEE Spectrum, Dezember 2012

Wohin führt das ?

Google car: Vermeidung von Unfällen

https://www.ted.com/talks/sebastian_thrun_google_s_driverless_car.html

THE PROBLEM WITH (BIG) DATA*



* The only mention of ‘big data’ I promise
** I made this up – but trust me, I am a professional

http://de.slideshare.net/sapient/sxsw-2015-humans-are-predictable-exploit-it-in-eight-charts?utm_source=slideshow&utm_medium=ssemail&utm_campaign=weekly_digest

Beispielthemen für Workshop

Anwendungsbeispiele, z.B. in Medizin, Produktion, CRM, Sport, Politik usw.

Visualisierung

Algorithmen zum Clustern (K means usw.)

NoSQL im Detail, am Beispiel usw.

Text mining, Bsp. Wikipedia, Clustering, Topic Extraction usw.

Social Media Mining, z.B. aus Verhalten Depressionsphasen erkennen

Stream processing (apache flink usw.)

Skalierbarkeit und Performance (Probleme)

Physicists make 'weather forecasts' for economies

Korrelation und Kausalität