

Daten, Daten, Daten

Wie Datenflut und Digitales die Epidemiologie zukünftig bestimmen werden

Dr. Joachim Krois

Charité Centre for Dental Medicine
Department for Operative and Preventive Dentistry
Aßmannshauser Str. 4–6
14197 Berlin, Germany



Digital Age

>>Storage<<

>>Transmission<<

>>Computation<<



Johannes Gutenberg
(1400-1468)

Im Jahr 2002 wurden 50%
der weltweit verfügbaren Daten
digital gespeichert .

Hilbert and López 2011

Das digitale Zeitalter in Zahlen

2010 2015 2020



127,000 km



7.5mm / 128GB



2005 2010 2012 2015

0.1ZB

1.2ZB

2.8ZB

8.5ZB

2020

40ZB

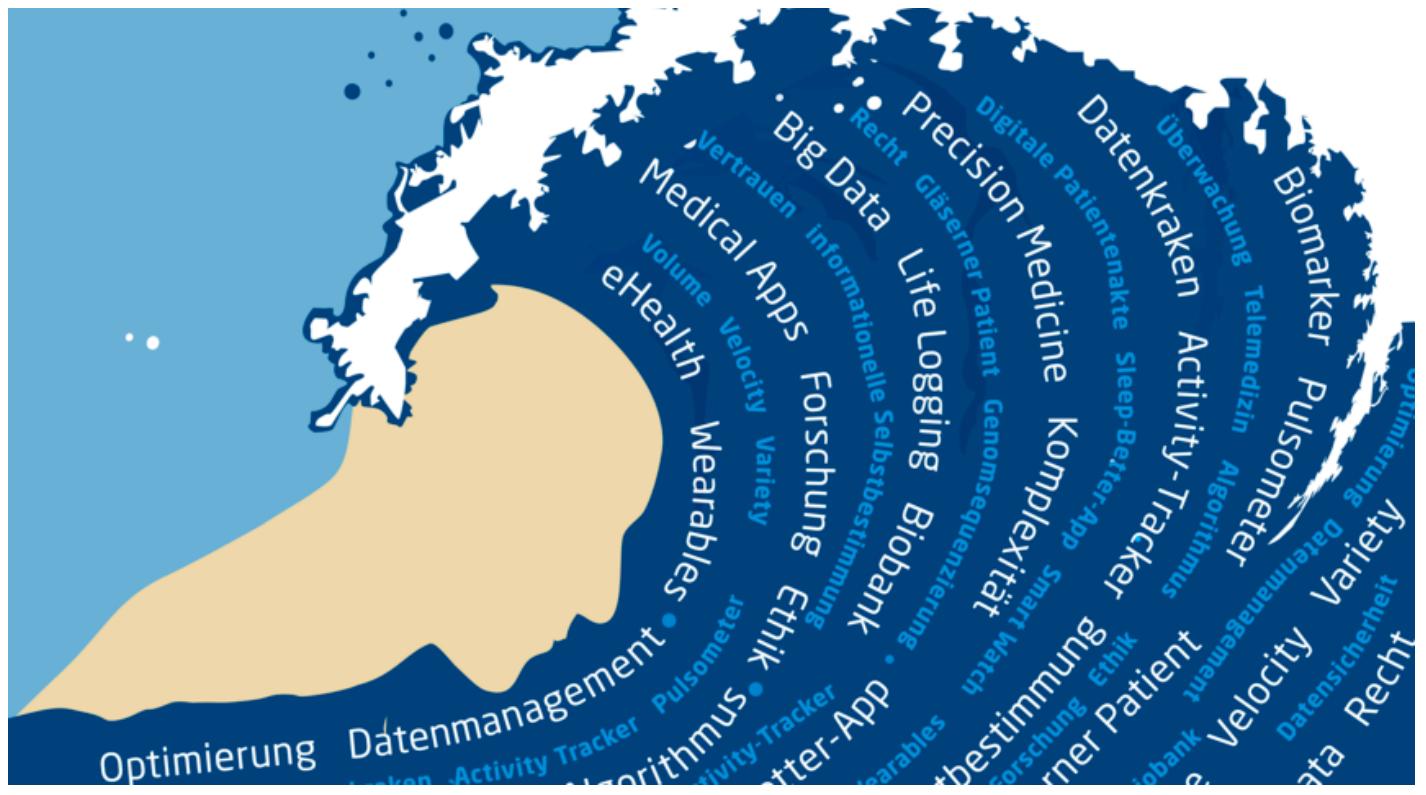
IDC projects that the Digital Universe will reach 40 ZB by 2020, an amount that exceeds previous forecasts by 5 ZBs.

BIG DATA

- Steigerung um den Faktor 33 in 10 Jahren
- Verdoppelung alle 2 Jahre
- Hauptsächlich „unstrukturierte Daten“

Gesundheitssektor

Im Jahr 2020 werden **2.314 Exabytes** an medizinischen Daten vorliegen. Das entspricht dem Datenvolumen einer DVD Sammlung von **579 Mrd. Stück**. In 2020 kommt somit auf jeden Menschen im Schnitte medizinische Daten in einer Größenordnung von **74 DVDs**.



<https://www.ethikrat.org/themen/forschung-und-technik/big-data/>

Big Data und Epidemiologie

6 V's

Der [...] Begriff **Big Data** [...] bezeichnet **Datenmengen**, welche beispielsweise zu **groß**, zu **komplex**, zu **schnelllebig** oder zu **schwach strukturiert** sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten. (Wikipedia)

Value		Clinically relevant data Longitudinal studies
Volume		High-throughput technologies Continuous monitoring of vital signs
Velocity		High-speed processing for fast clinical decision support Increasing data generation rate by the health infrastructure
Variety		Heterogeneous and unstructured data sources Differences in frequencies and taxonomies
Veracity		Data quality is unreliable Data coming from uncontrolled environments
Variability		Seasonal health effects and disease evolution Non-deterministic models of illness and health

Andreu-Perez et al 2015

Chancen für die Epidemiologie

6 V's

Große populationsweite
longitudinale Studien
ermöglichen die Entwicklung von
Vorhersagemodellen mit
dem Ziel der **Prävention**,
frühen Intervention und
dem **optimierten**
Management von Krankheiten.

Value		Clinically relevant data Longitudinal studies
Volume		High-throughput technologies Continuous monitoring of vital signs
Velocity		High-speed processing for fast clinical decision support Increasing data generation rate by the health infrastructure
Variety		Heterogeneous and unstructured data sources Differences in frequencies and taxonomies
Veracity		Data quality is unreliable Data coming from uncontrolled environments
Variability		Seasonal health effects and disease evolution Non-deterministic models of illness and health

Andreu-Perez et al 2015

Methoden und Datenquellen der Epidemiologie

Klassische Methoden der Epidemiologie

- Querschnittsstudien
- Längsschnittstudien
- Kohortenstudien
- Fall-Kontrollstudien
- Interventionsstudien
- ...

Klassische Datenquellen der Epidemiologie (**Primärdaten**)

- Administrative/institutionelle Datenbanken
- Epidemiologische Studien
- Klinische Studien
- Literatur- und Metaanalysen
- Umweltstudien

'Most scientific studies are wrong, and they are wrong because scientists are interested in funding and careers rather than truth'

(Smith 2013)

Elektronische Patientenakte



Patientendaten
Administrative Daten
Diagnosen
Notizen
Labortests
Medikamenteneinnahme
Behandlungsdaten
Bilddaten (MRI, CT, etc.)
....

<https://medcitynews.com/2017/10/patients-healthcare-professionals-ehr/>

Elektronische Patientenakte



Discovering medical conditions associated with periodontitis using linked electronic health records

Boland MR, Hripcak G, Albers DJ, Wei Y, Wilcox AB, Wei J, Li J, Lin S, Breene M, Myers R, Zimmerman J, Papapanou PN, Weng C. Discovering medical conditions associated with periodontitis using linked electronic health records. *J Clin Periodontol* 2013; 40: 474–482. doi: 10.1111/jcpe.12086.

<https://medcitynews.com/2017/10/patients-healthcare-professionals-ehr/>

Elektronische Patientenakte



<https://medcitynews.com/2017/10/patients-healthcare-professionals-ehr/>

SPAHNS AMBITIONIERTER DIGITAL-PLAN

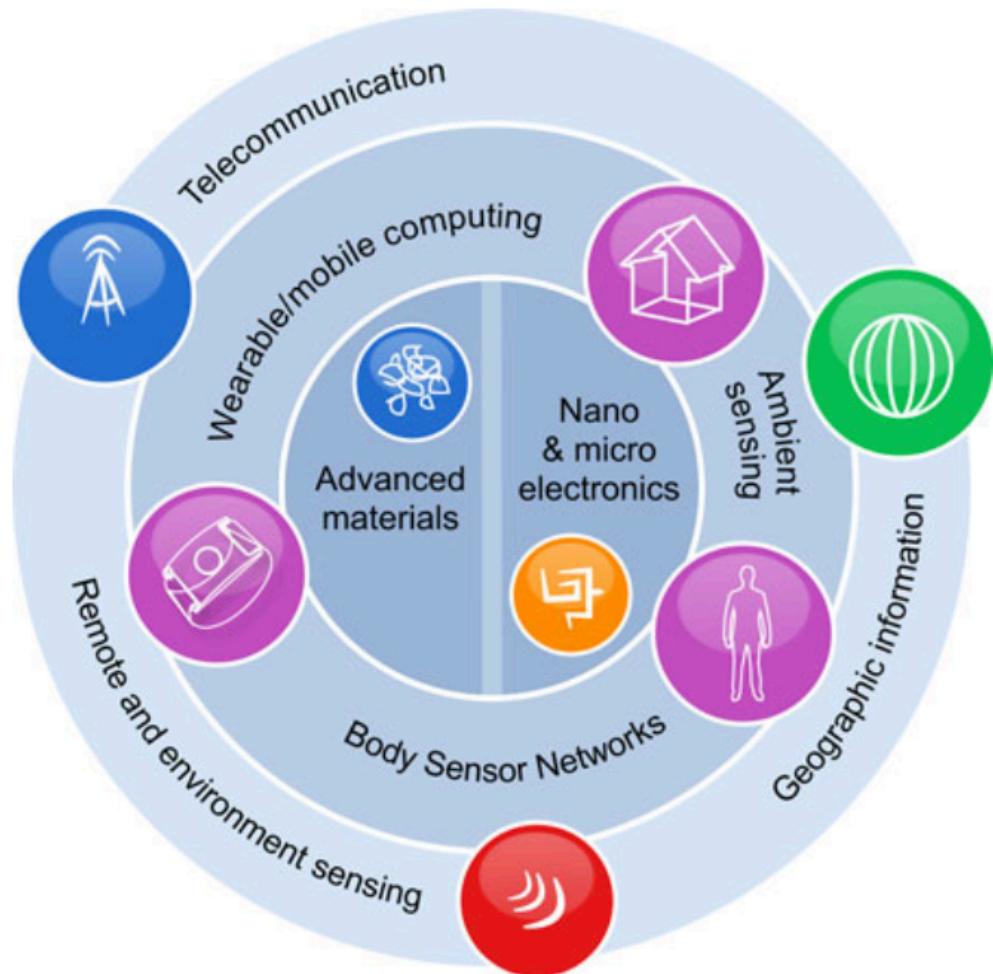
Elektronische Patienten-Akte kommt 2021



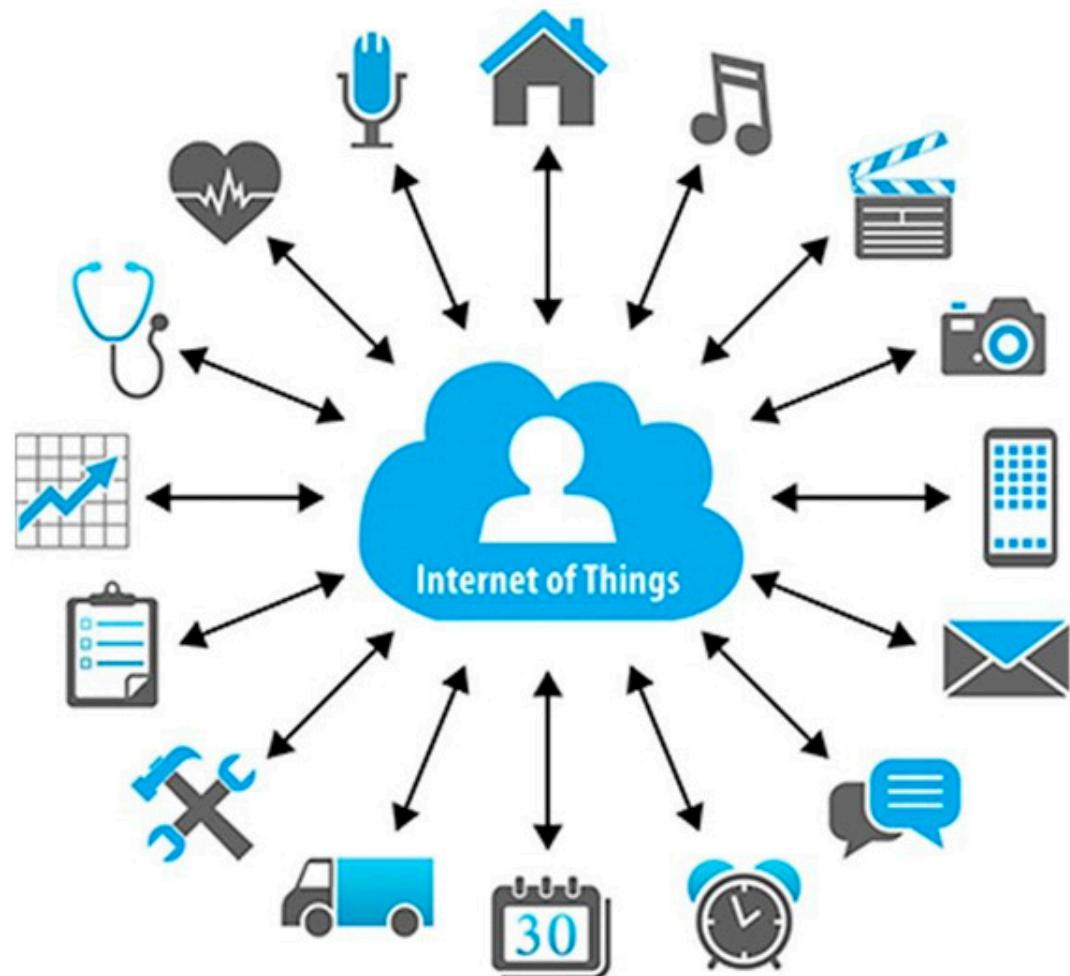
Jens Spahn (CDU), Bundesgesundheitsminister, hat ehrgeizige Ziele
Foto: Kay Nietfeld / dpa

Bild, 15.05.2019

Internet der Dinge im Gesundheitswesen



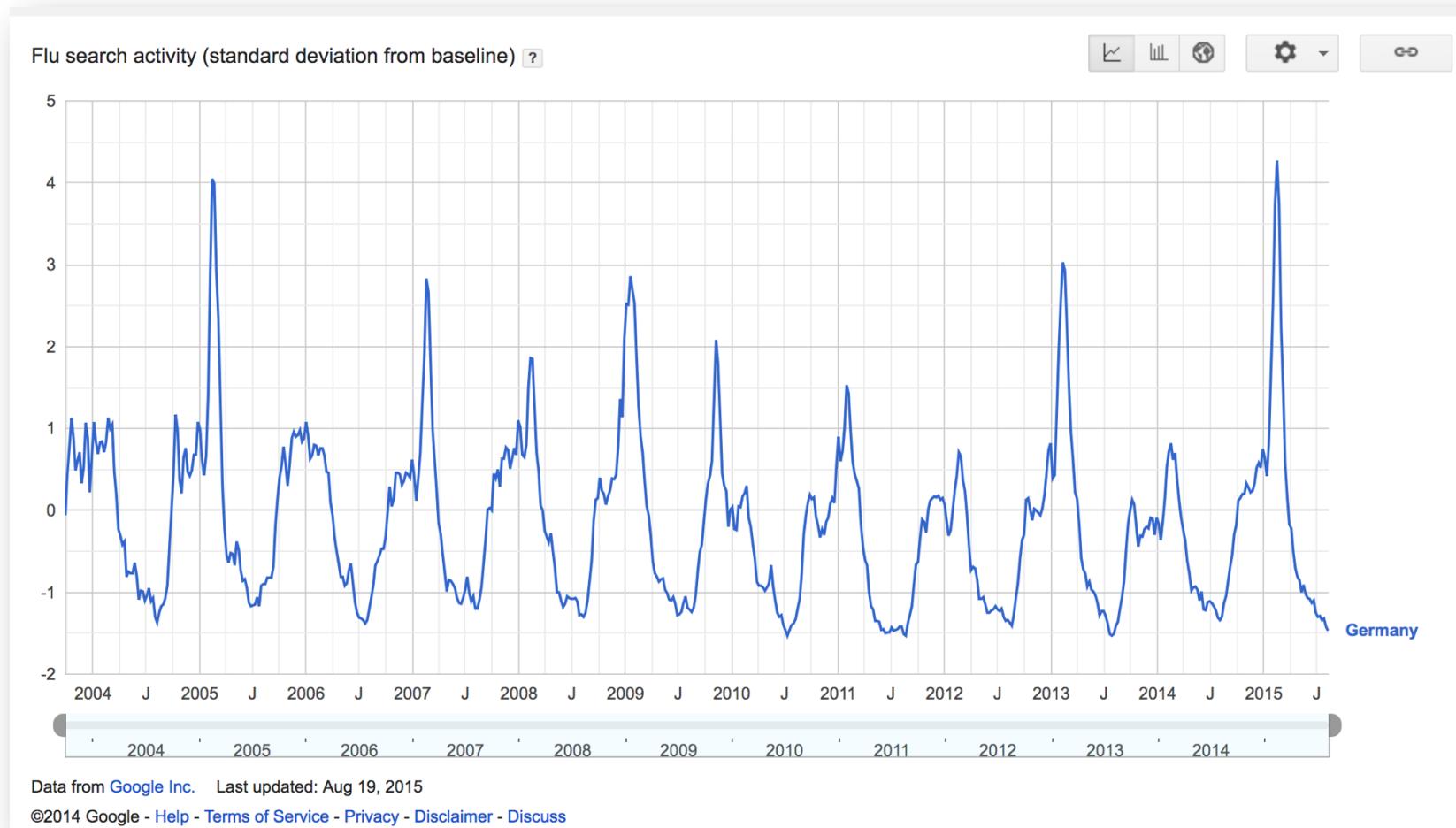
Andreu-Perez et al 2015



<https://justcreative.com/2018/11/19/internet-of-things-explained/>

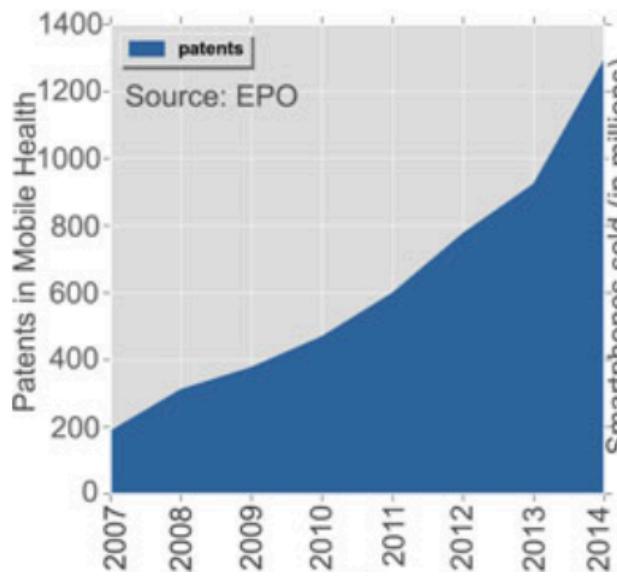
Echtzeitmonitoring

Google Flu Trend



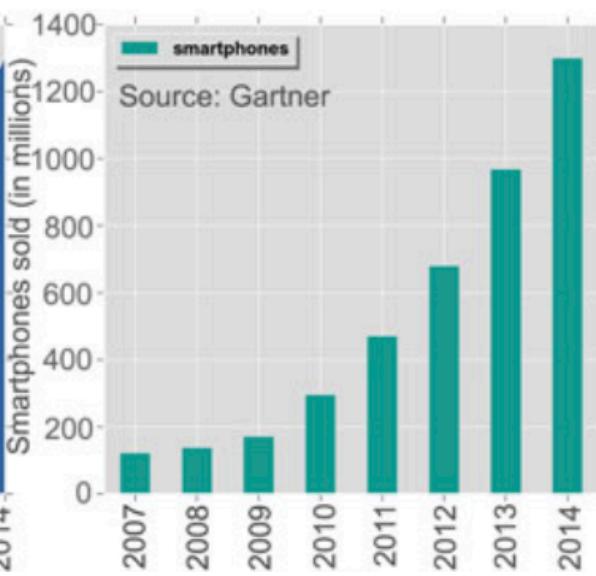
Gesundheitsapps

Patente im Bereich Mobile Health



Andreu-Perez et al 2015

Verkaufszahlen Smartphones

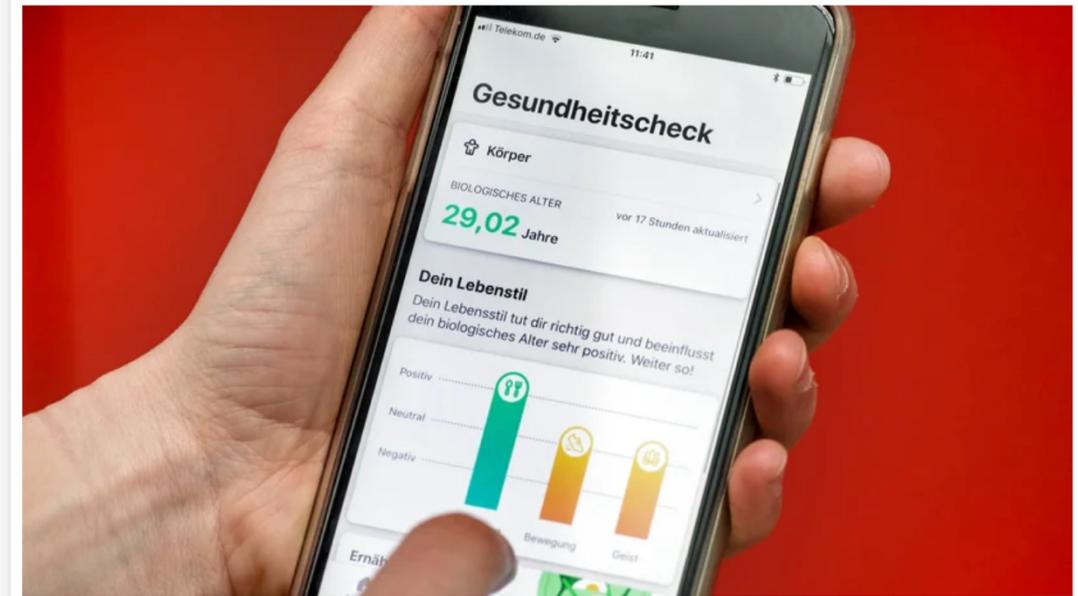


Jens Spahn

Krankenkassen sollen für Gesundheitsapps zahlen

Der Bundesgesundheitsminister fordert: Wenn ein Arzt zur Nutzung bestimmter Apps rät, muss nicht der Patient die Kosten tragen. Gibt es bald Apps auf Rezept?

15. Mai 2019, 18:52 Uhr / Quelle: ZEIT ONLINE, dpa, AFP, vk / 14 Kommentare



Mit der App "Vivy" lässt sich die digitale Gesundheitsakte verwalten. © Michael Kappeler/dpa

Andreu-Perez et al 2015

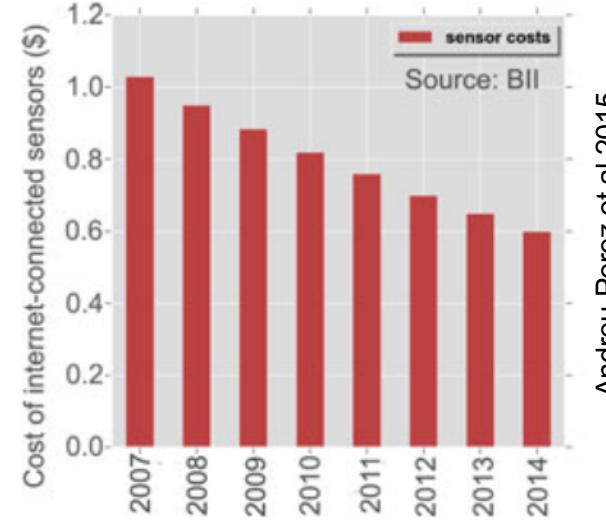
Wearables



<https://econsultancy.com/internet-of-things-healthcare/>

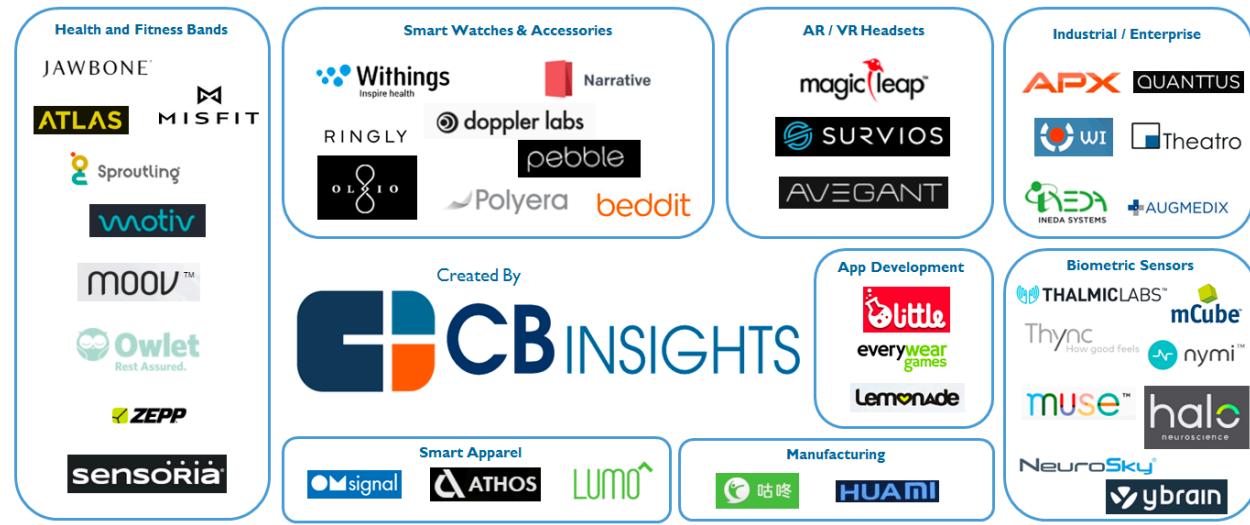


Kosten von Sensoren



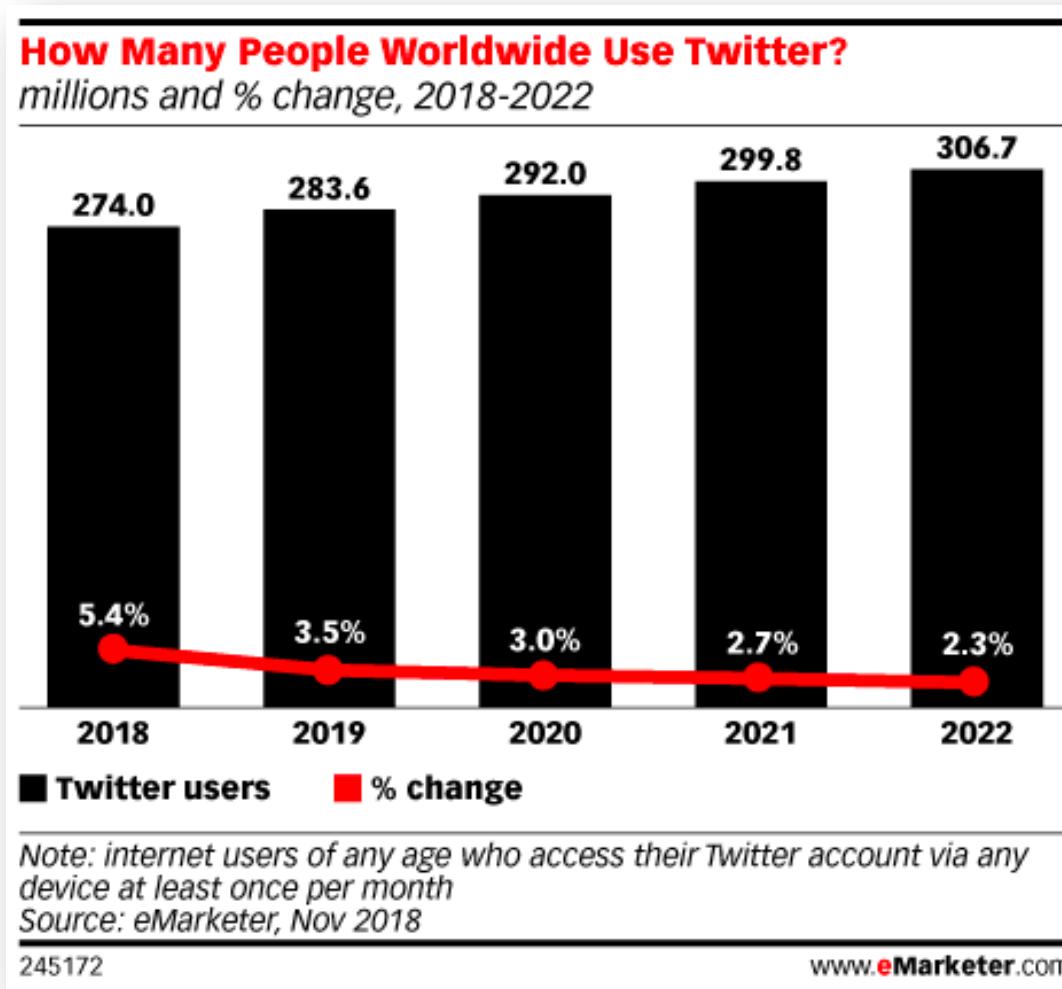
Andreu-Perez et al 2015

The Wearables Tech Market Map



<https://www.cbinsights.com/research/wearable-computing-startups-list/>, Stand 2015

Social Health



Psychological Language on Twitter Predicts County-Level Heart Disease Mortality



Johannes C. Eichstaedt¹, Hansen Andrew Schwartz^{1,2},
Margaret L. Kern^{1,3}, Gregory Park¹, Darwin R. Labarthe⁴,
Raina M. Merchant⁵, Sneha Jha², Megha Agrawal²,
Lukasz A. Dziurzynski¹, Maarten Sap¹, Christopher Weeg¹,
Emily E. Larson¹, Lyle H. Ungar^{1,2}, and Martin E. P. Seligman¹

¹Department of Psychology, University of Pennsylvania; ²Department of Computer and Information Science, University of Pennsylvania; ³Graduate School of Education, University of Pennsylvania; ⁴Medicine, Northwestern University; and ⁵Department of Emergency Medicine, U

Psychological Science
2015, Vol. 26(2) 159–169
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: [10.1177/0956797614557867](https://doi.org/10.1177/0956797614557867)
pss.sagepub.com




Does Twitter language reliably predict heart disease? A commentary on Eichstaedt et al. (2015a)

Nicholas J.L. Brown and James C. Coyne
University Medical Center, University of Groningen, Groningen, Netherlands

Social Health

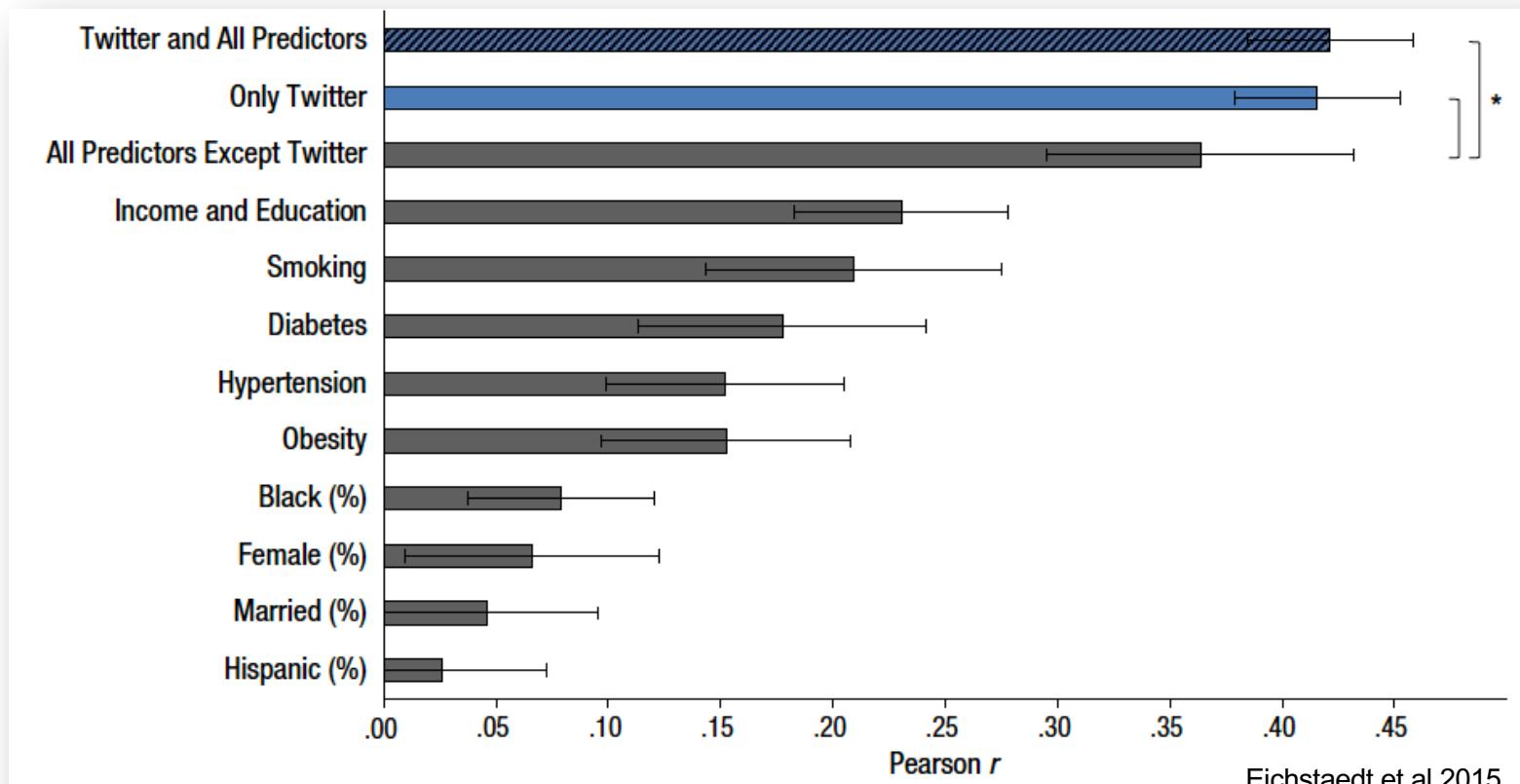
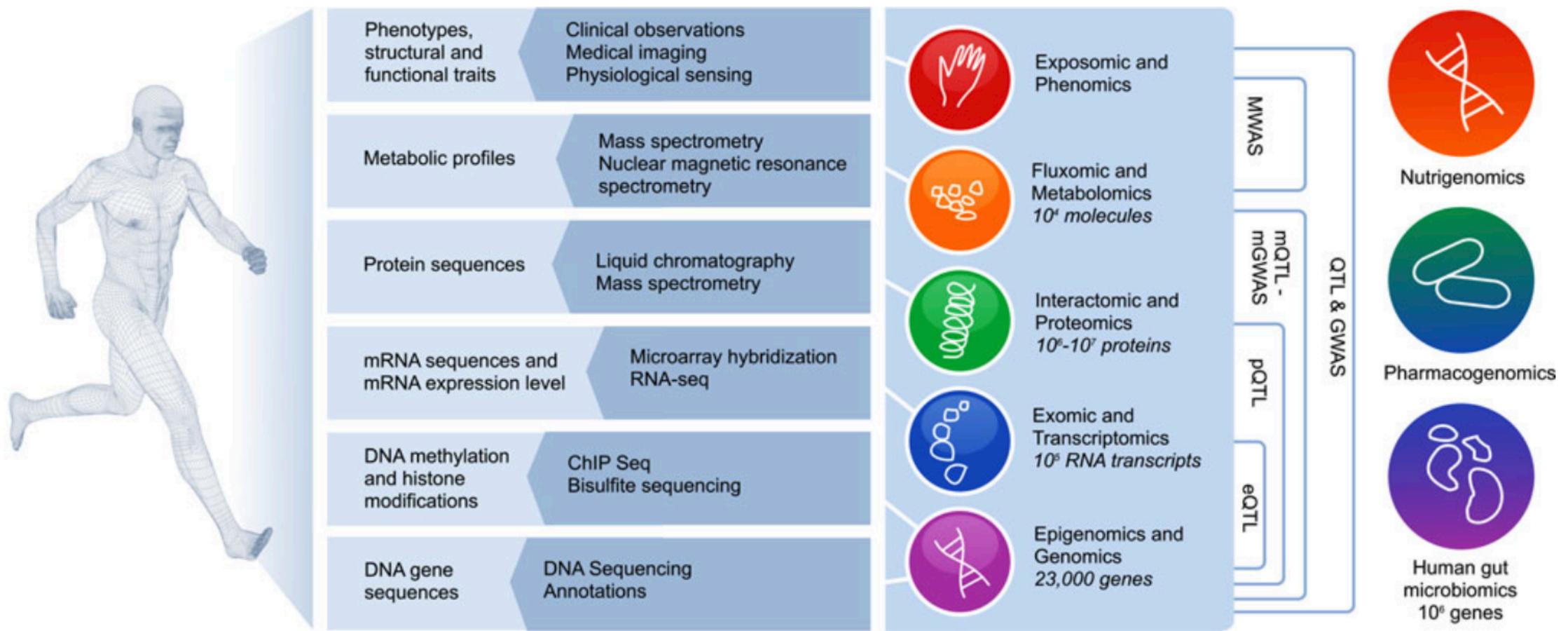


Fig. 2. Performance of models predicting age-adjusted mortality from atherosclerotic heart disease (AHD). For each model, the graph shows the correlation between predicted mortality and actual mortality reported by the Centers for Disease Control and Prevention. Predictions were based on Twitter language, socioeconomic status, health, and demographic variables singly and in combination. Higher values mean better prediction. The correlation values are averages obtained in a cross-validation process used to avoid distortion of accuracy due to chance (overfitting; for details, see the text). Error bars show 95% confidence intervals. Asterisks indicate significant differences between models (* $p < .05$).

Omics Daten



Andreu-Perez et al 2015

Omics Daten

OXFORD

Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences

Claudia Manzoni*, Demis A. Kia*, Jana Vandrovčová, John Hardy,
Nicholas W. Wood, Patrick A. Lewis and Raffaele Ferrari

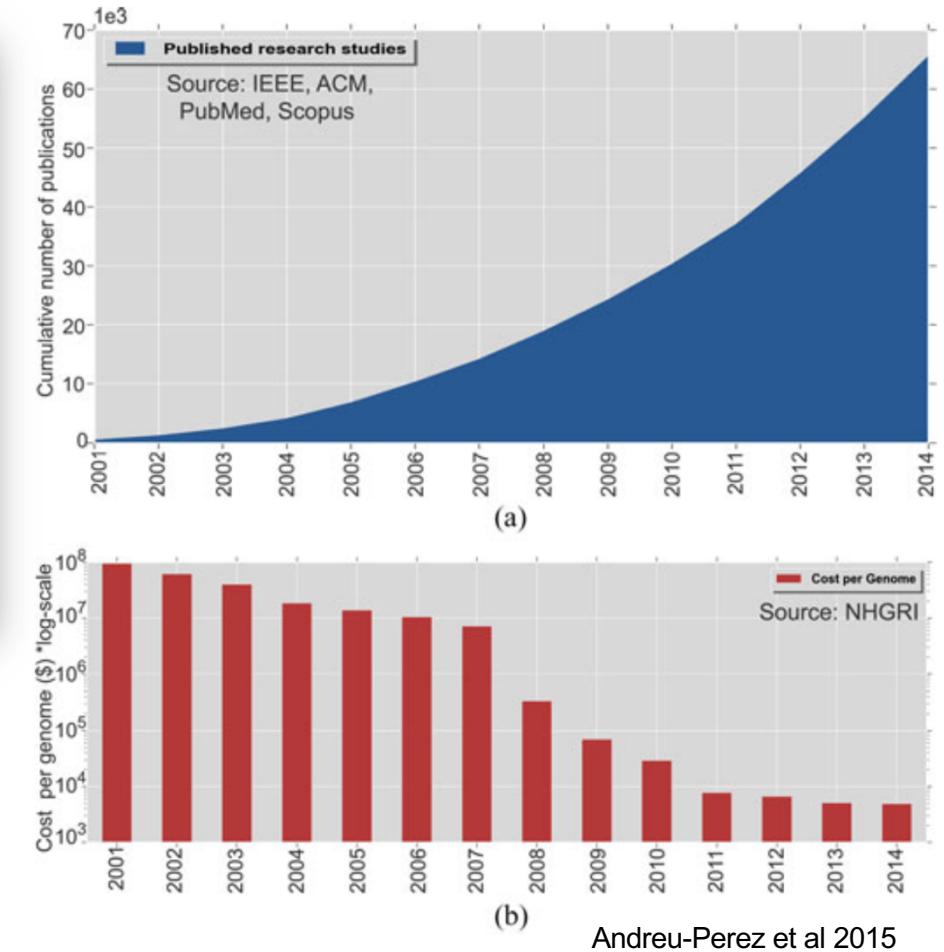
Briefings in Bioinformatics, 19(2), 2018, 286–302

doi: 10.1093/bib/bbw114

Advance Access Publication Date: 22 November 2016

Paper

Studien mit DNA/Genome
Sequenzierung (oben) und Kosten
der Gensequenzierung (unten)



Data deluge

„Data is permeating every component of the health care ecosystem.“

Stanford Medicine 2017 Health Trends Report

“We are still in the contemporary Dark Ages when trying to access and utilize the available data and information being produced and stored.”

Thomas R. Cech (Nobel Price Winner)

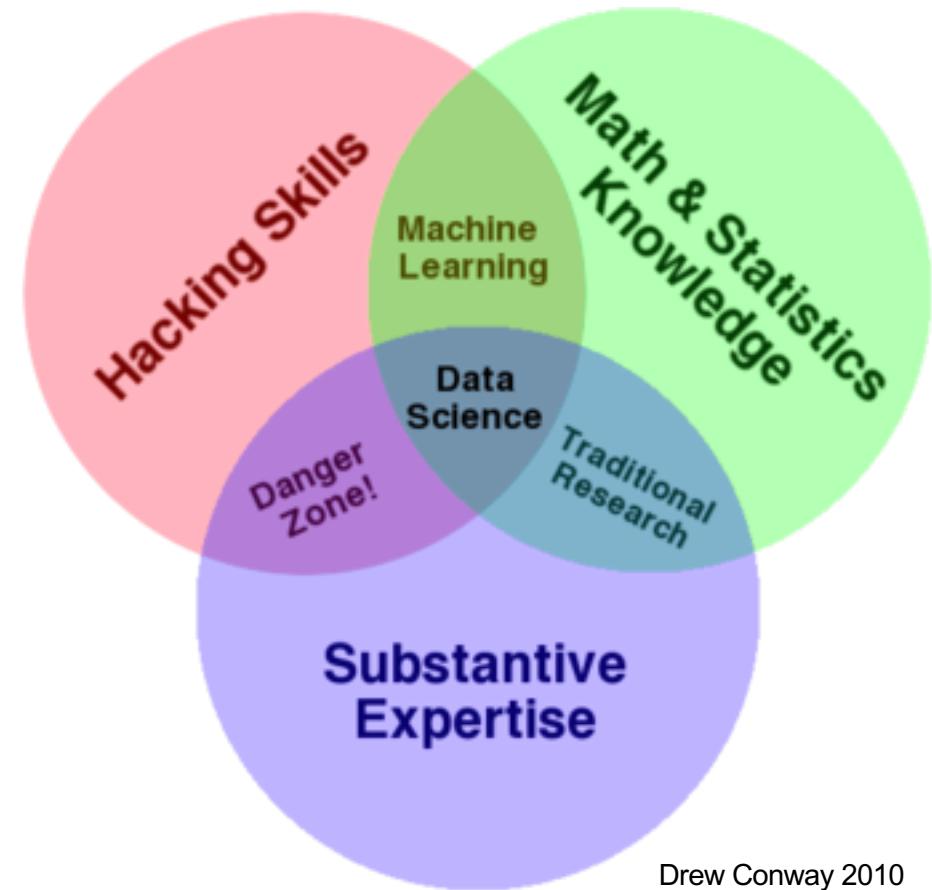
Or shall we call it data tsunami?

Epidemiologie im Digitalen Zeitalter

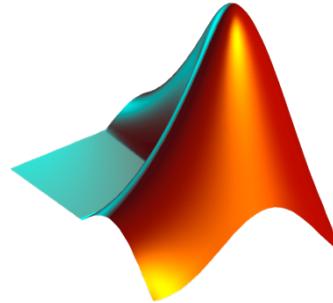
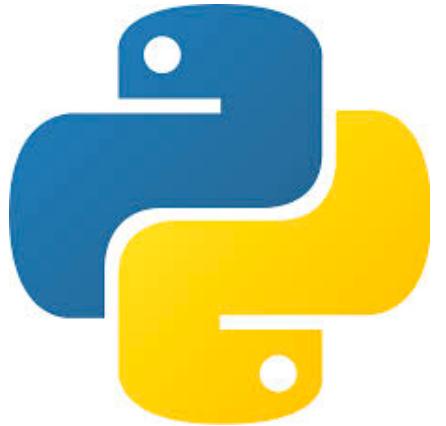
Klassische Fähigkeiten

- Zusammenspiel von genetischen, ökologischen und sozialen Faktoren
- Entwicklung und Evolution von Krankheiten in der Bevölkerung
- Gesundheitsforschung und klinische Forschung und damit einhergehende ethische und politische Aspekte
- Epidemiologische Studien (Konzeption, Machbarkeit, Mittelbeschaffung, Durchführung, Datensammlung und -analyse, Dateninterpretation und Präsentation der Ergebnisse)

Fähigkeiten in der Big Data Ära



Epidemiologie und Data Science

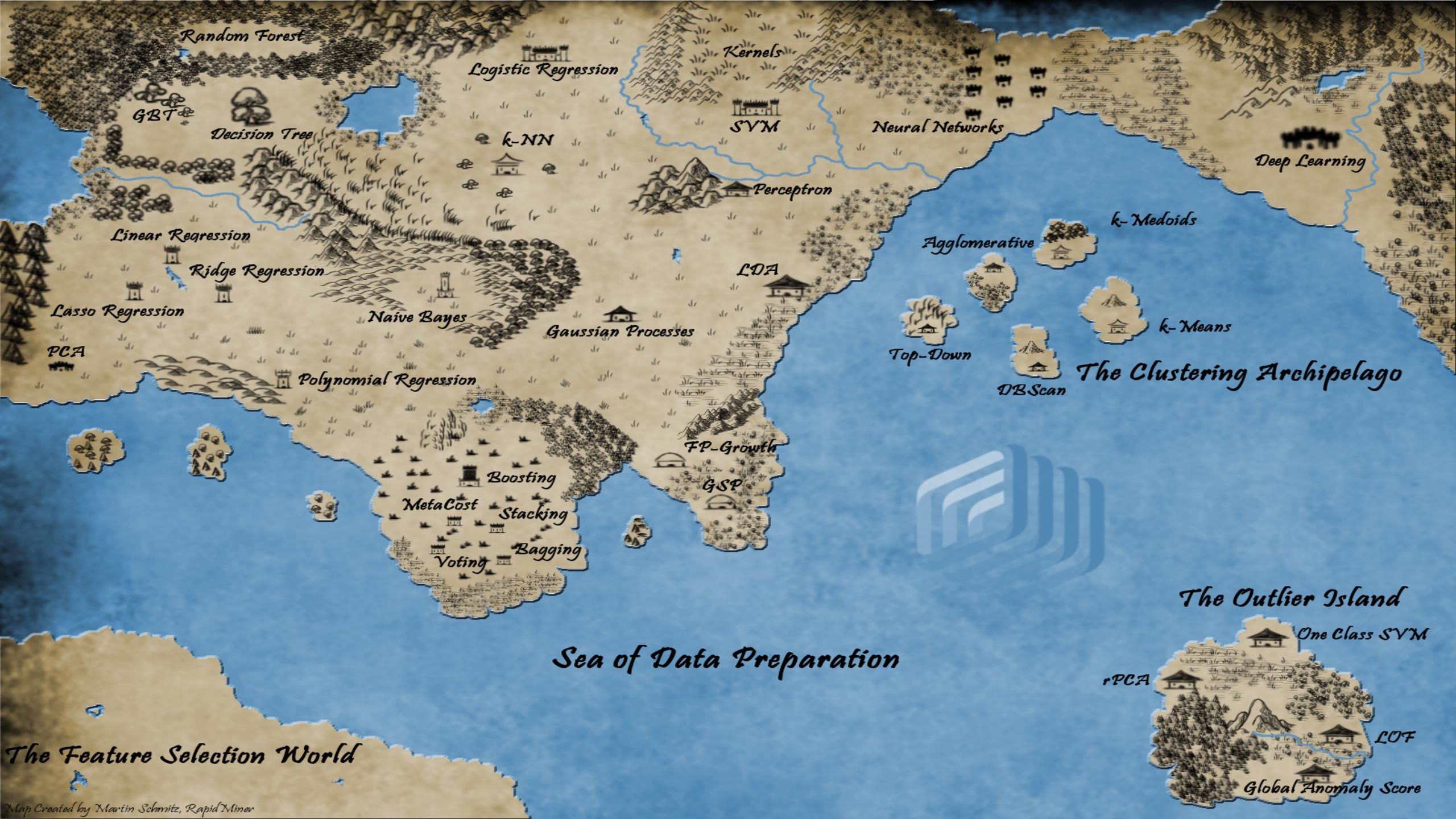


```
[dental-research-2:~ jokr$ python
Python 3.7.0 (default, Jun 28 2018, 07:39:16)
[Clang 4.0.1 (tags/RELEASE_401/final)] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> ]
```

Epidemiologie und Data Science

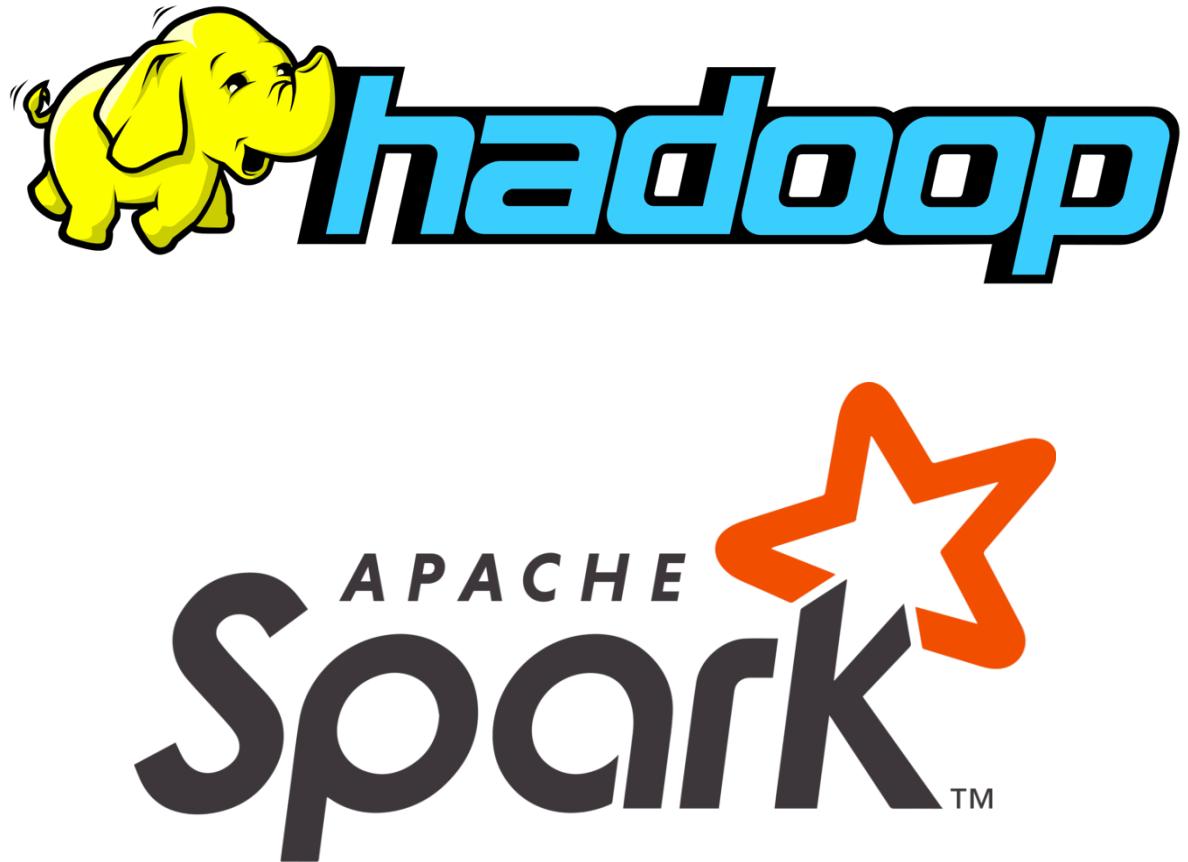
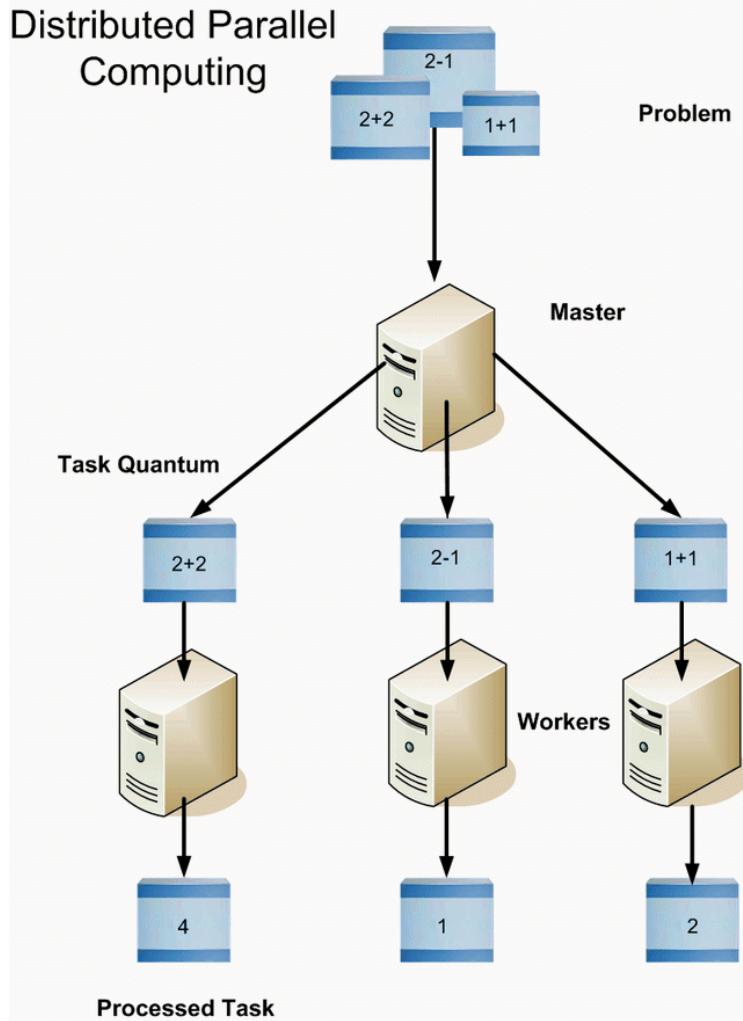


<http://bigdata.black/artificial-intelligence/neural-networks/data-science-tools-for-non-technical-users/>

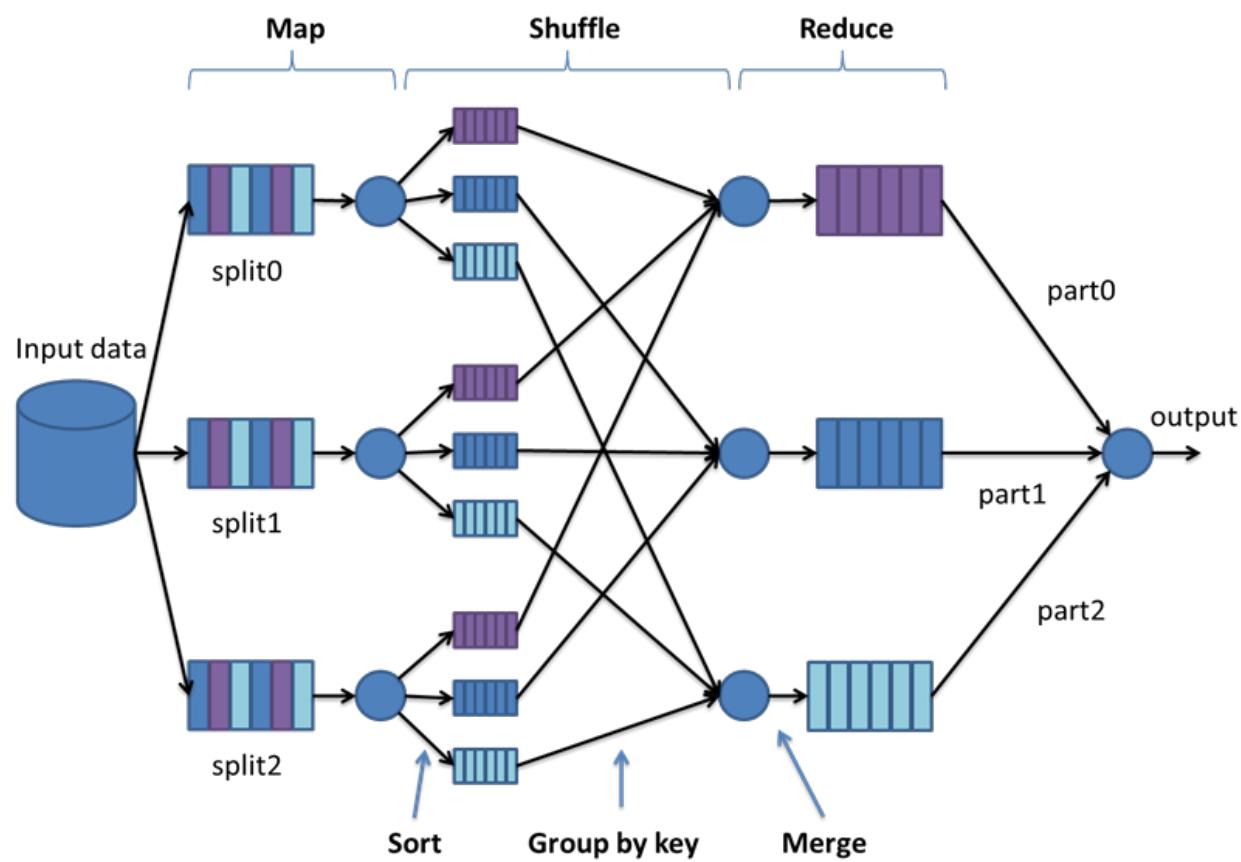


Out-of-core computing

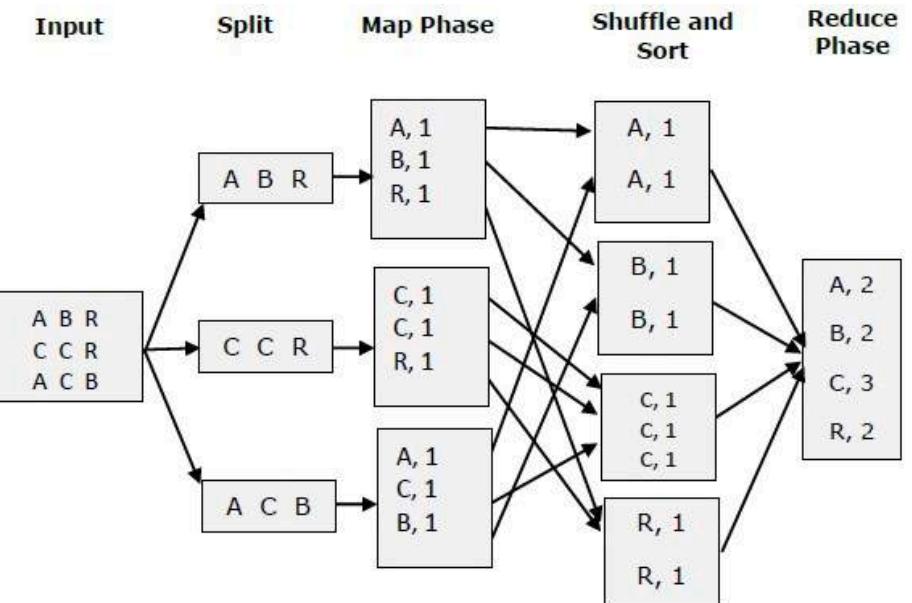
<https://www.codeproject.com/Articles/35671/Distributed-and-Parallel-Processing-using-WCF>



Das Map-Reduce Paradigma

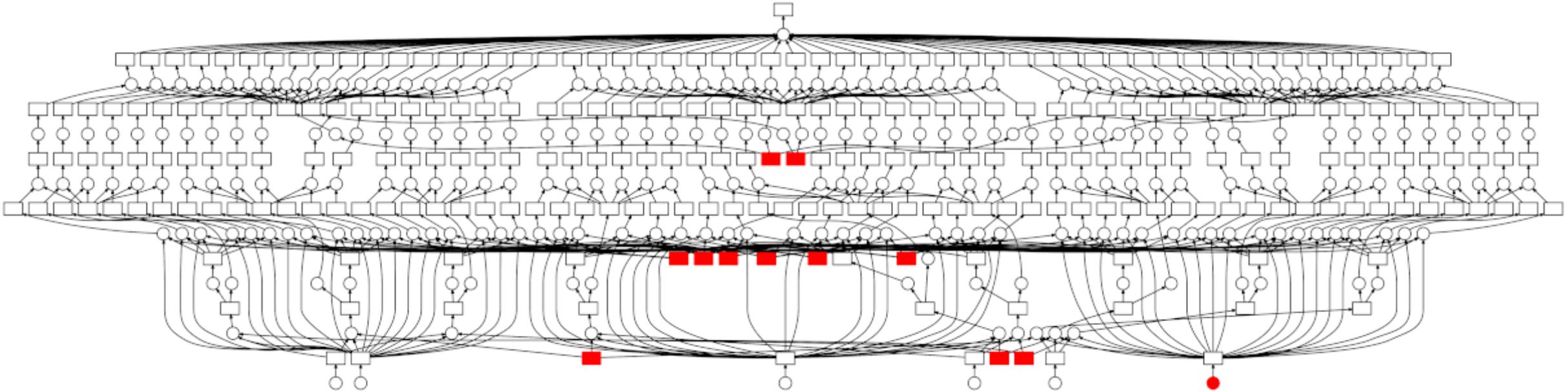


<https://www.kisspng.com/png-mapreduce-big-data-apache-hadoop-programming-model-3373874/>



<https://stackoverflow.com/questions/44538028/is-it-possible-to-output-multiple-values-under-the-same-key-in-mapreduce>

Das Map-Reduce Paradigma



<https://github.com/dask/dask-tutorial>

Zusammenfassung

Neues Paradigma:

Data-driven Epidemiology / Hypothesis-generating analyses

- Verknüpfung von Primär- (strukturiert) und Sekundärdaten (semi-, unstrukturiert)
 - Bessere Nutzung der existierenden Daten.
 - Welche Daten sind von Relevanz?

Chancen

- Prädiktion, Personalisierung, Prävention und Partizipation (*P4-medicine*)
- Optimiertes Management und frühe Intervention
- *Universal statistical significance*

Herausforderungen

- Datenmanagement, –qualitätssicherung und –zugang
- Datensicherheit und –schutz
- Datenbesitz und –verwendungsrecht
- Daten- und technische Kompetenz
- Coden, *Data Mining* und *Big Data* Analysen
- (große) Beobachtungsstudien vs. randomisierte kontrollierte Studie

Herzlichen Dank für die Aufmerksamkeit



Dr. Joachim Krois

Charité Centre for Dental Medicine

Department for Operative and Preventive Dentistry

Aßmannshauser Str. 4–6

14197, Berlin, Germany

joachim.krois@charite.de

+49-30-450-562328