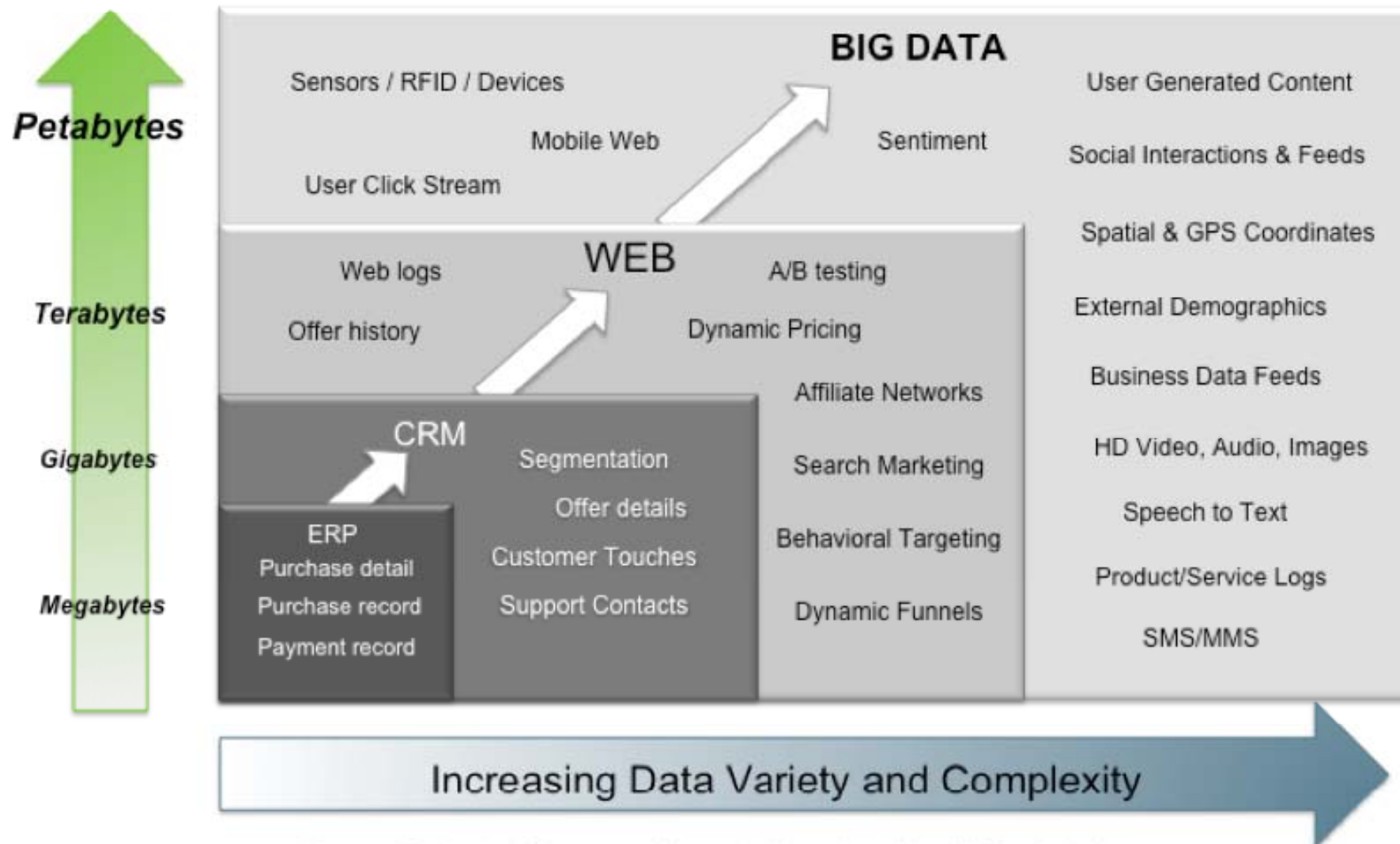


# Big Data = Transactions + Interactions + Observations



*Source: Contents of above graphic created in partnership with Teradata, Inc.*

**Map generated by more than 250 million public tweets (collected from Twitter.com)  
with high-resolution location information, broadcast, March 2011-Jan 2012**



Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, et al. (2012) Digital Epidemiology. PLoS Comput Biol 8(7): e1002616.  
doi:10.1371/journal.pcbi.1002616

<http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002616>

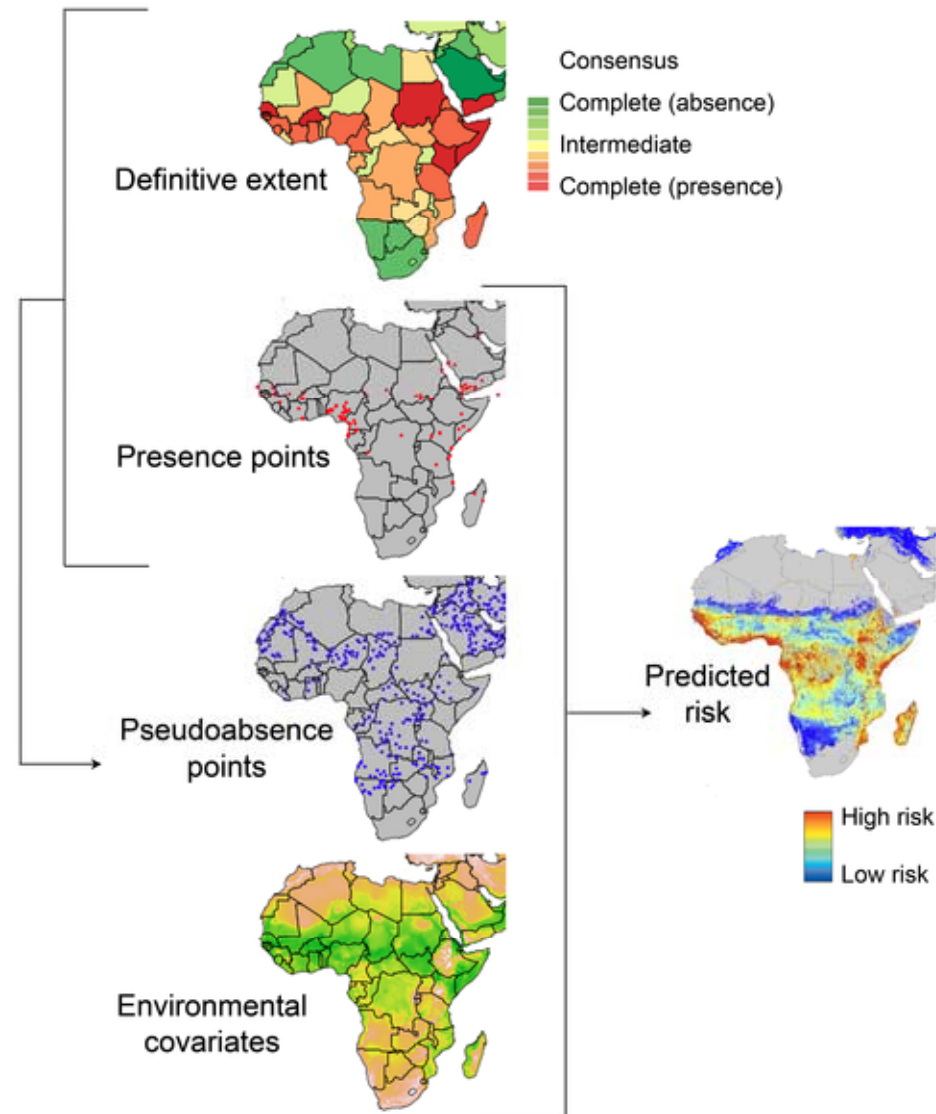
Chiolero, Fäh: Surveillance with Big data: too Big to fail? 4.2.2014

Institut für Sozial- und  
Präventivmedizin



**Universität  
Zürich**<sup>UZH</sup>

## A schematic overview of the process of predicting spatial disease risk.



Hay SI, George DB, Moyes CL, Brownstein JS (2013) Big Data Opportunities for Global Infectious Disease Surveillance.

PLoS Med 10(4): e1001413. doi:10.1371/journal.pmed.1001413

<http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.1001413>

Chiolero, Fäh: Surveillance with Big data: too Big to fail? 4.2.2014

Institut für Sozial- und  
Präventivmedizin



Universität  
Zürich<sup>UZH</sup>

## What Makes this Study Different?

The answer is **BIG** data.

This study plans to gather more data about heart health from more people than any research study has done before. We'll use it to develop strategies to prevent and treat all aspects of heart disease. It's as simple as that.



## Our Goals

You Can Help Us Advance Science

We're looking for all kinds of people over 18, including those who are completely healthy, those who have heart disease, and even patients with cardiovascular disease that we don't yet know how to treat. So, join the study, be a part of something big, and help make a contribution to science.

Just a few things we hope to achieve:





# Our team is working around the clock.

Our team includes experts in cardiology, internal medicine, epidemiology, data technology, informatics, social media, and mobile technology. When we're not doing research, we're taking care of patients with heart disease.

## Principal Investigators



### **Dr. Jeffrey E. Olgin**

Professor of Medicine, Chief of Cardiology, UCSF

Dr. Olgin is a practicing cardiologist and cardiovascular researcher. He runs a coordinating center for cardiovascular clinical trials, and his research interests include atrial fibrillation, arrhythmias, and risk prediction and prevention of cardiovascular disease.

What he does to stay heart healthy: Rides his bike to work — even in the rain.



### **Dr. Mark J. Pletcher**

Associate Professor of Epidemiology & Biostatistics and Medicine, UCSF

Dr. Pletcher is a practicing internal medicine physician and an epidemiologist. His research focuses on prevention of cardiovascular disease, and how exposure to cardiovascular risk factors like blood pressure and cholesterol during young adulthood may contribute to heart disease later in life.

What he does to stay heart healthy: Plays Ultimate Frisbee every Sunday!



### **Dr. Gregory M. Marcus**

Being part of a study is a real commitment, but we've made it easier than ever to make a difference. Everyone who participates will answer survey questions, and we'll ask many of you to do a whole lot more (though you'll always have the chance to opt out).

## Just a few ways we'll ask you to contribute:



**Answer surveys.** We'll ask you questions about your health and behaviors, and ask you to update them every six months on your computer or smartphone. You don't have to complete them in one sitting, but you'll need to get to them all eventually.

**Collect data at home.** If you want, you can use your own scale, blood pressure machine, and more to collect measurements and send them to us using our secure system. We might even mail you a "spit kit" to collect your DNA.



**Connect with your social media profile.** We might ask if we can see how you use your Facebook or other social media accounts. Don't worry: we won't share any information with Facebook or anyone else. We'll just use that data for research to improve heart health.

**Use new technology.** Some participants will wear special sensors or add cool gadgets to their smartphone to track measurements on the go.




**Download apps.** If you have a smartphone, you will be able to download free apps to record pulse, weight, sleep, activity, behavior, and more as we develop them.

**Tell us when you go to the hospital.** It's critical for us to know when you have any sort of health event like a heart attack or something else that brings you to the hospital for treatment.





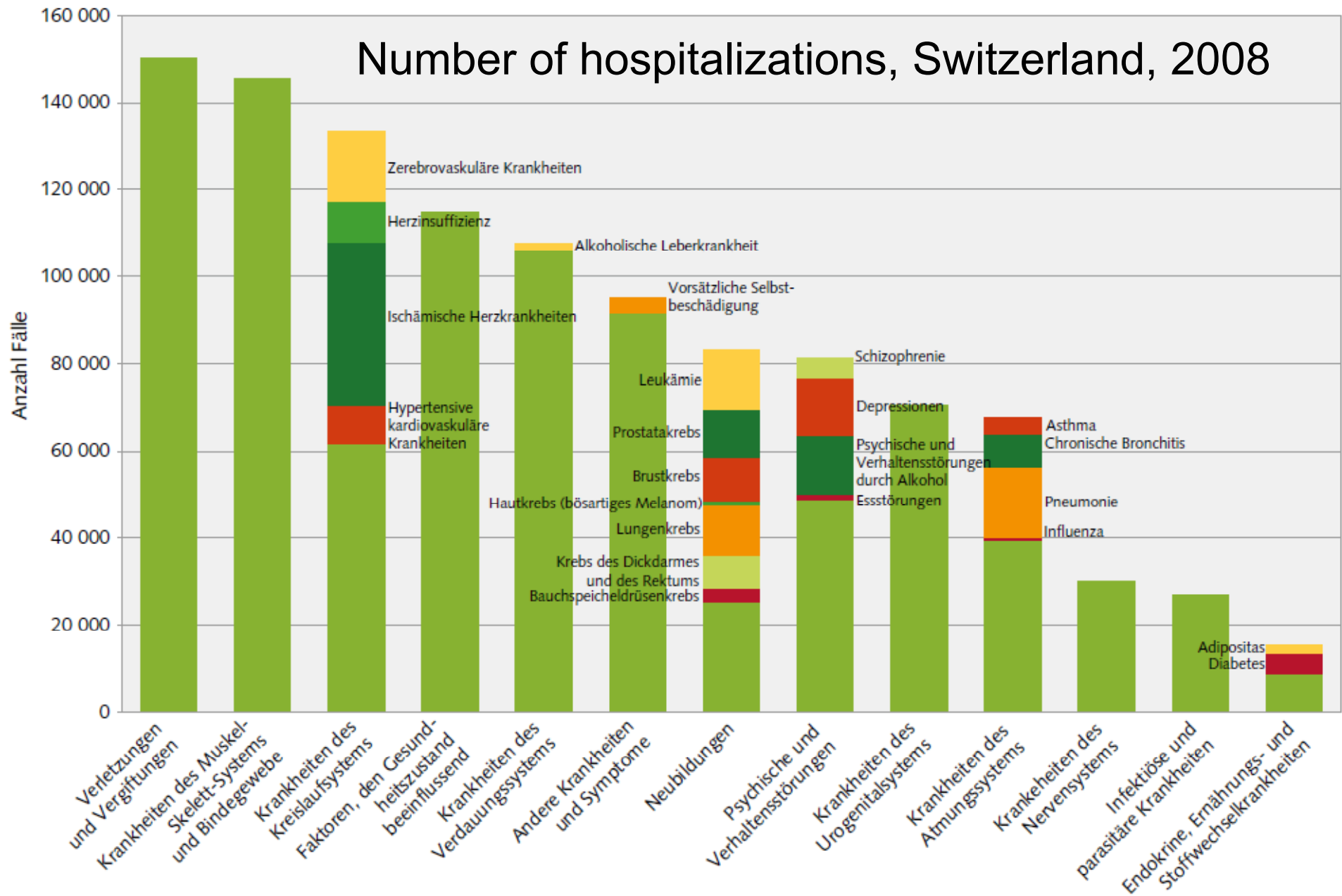
# Big Data in Switzerland?

- Possible candidates:
  - Hospital discharge data (MedStat)
  - Swiss National Cohort (SNC) 
- Features
  - Large (relatively) sample sizes
  - Collection of routine data
  - Originally not thought for science
  - Gathering not planless
  - Too good to be called „Big Data“ (?)

# MedStat

- Hospital discharges (only inpatients)
- 2012:
  - Since 1998 (2005)
  - 40'000 hospital beds
  - >1 mio patients providing >1.3 mio discharges
  - Up to 50 diagnoses and 100 treatments per patient
  - Problem: data sources and diagnoses may vary over time

# Number of hospitalizations, Switzerland, 2008



Stationäre Fälle, ohne Neugeborene und Aufenthalte bei Schwangerschaft und Geburt (N = 1'121'775)

Bundesamt für Statistik, 2010

Chiolero, Fäh: Surveillance with Big data: too Big to fail? 4.2.2014

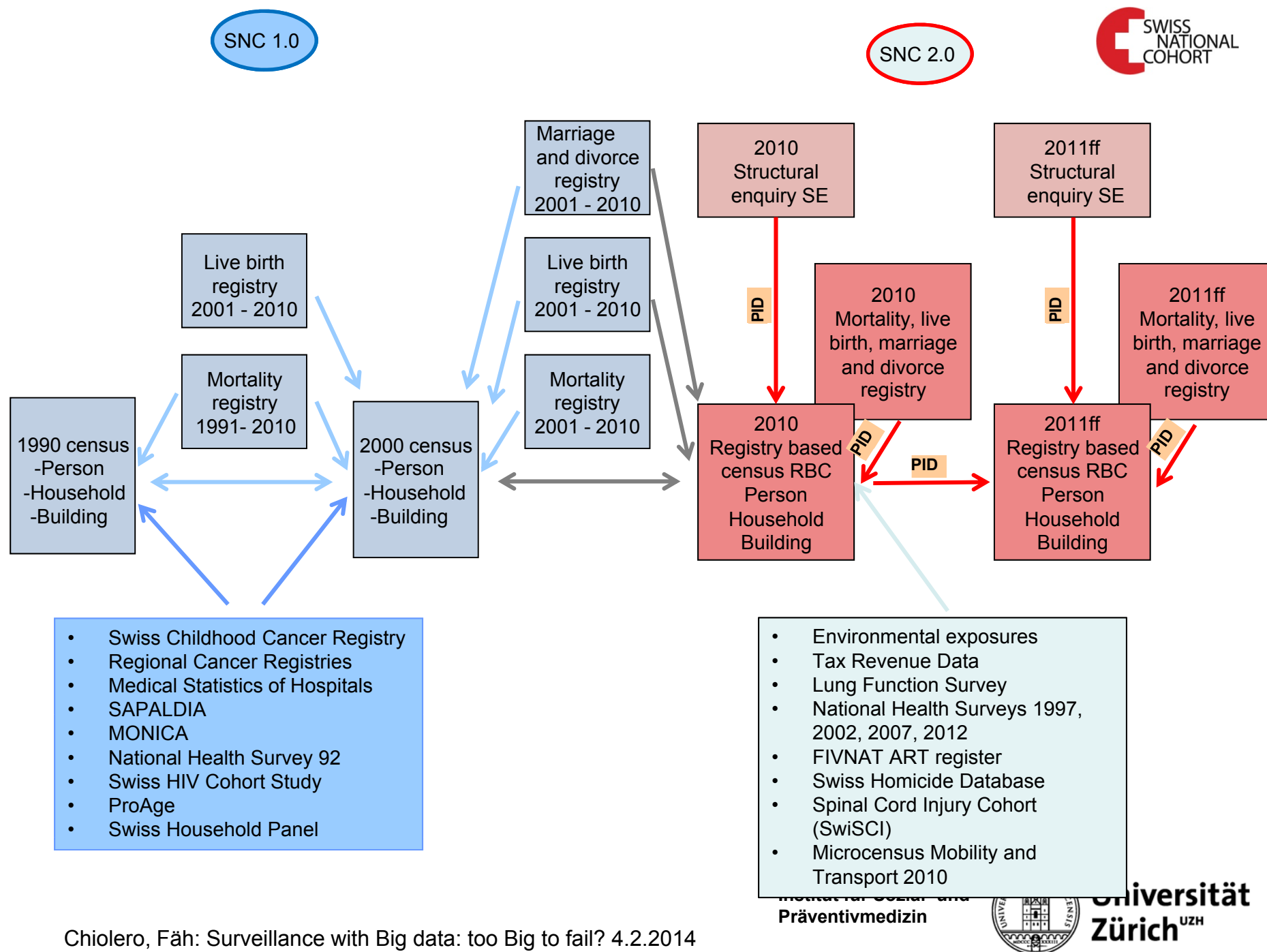
Institut für Sozial- und Präventivmedizin



**Universität  
Zürich** <sup>UZH</sup>

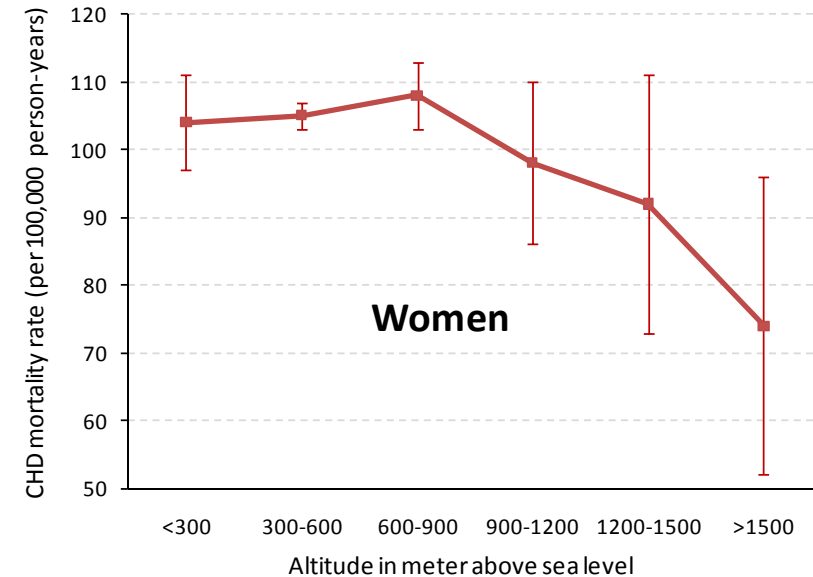
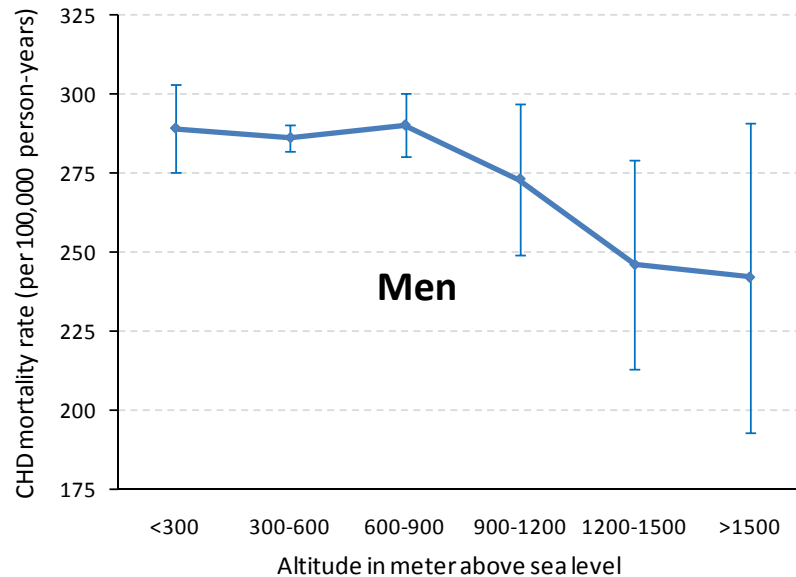
# Swiss National Cohort (SNC)

- Anonymous record linkage of different data sources
- Backbone: Swiss census 1990 and 2000
- Outcome: Vital status from death registry
- Highly representative for the population of Switzerland (>6 mio individuals)
- Further record linkage with other data sources possible





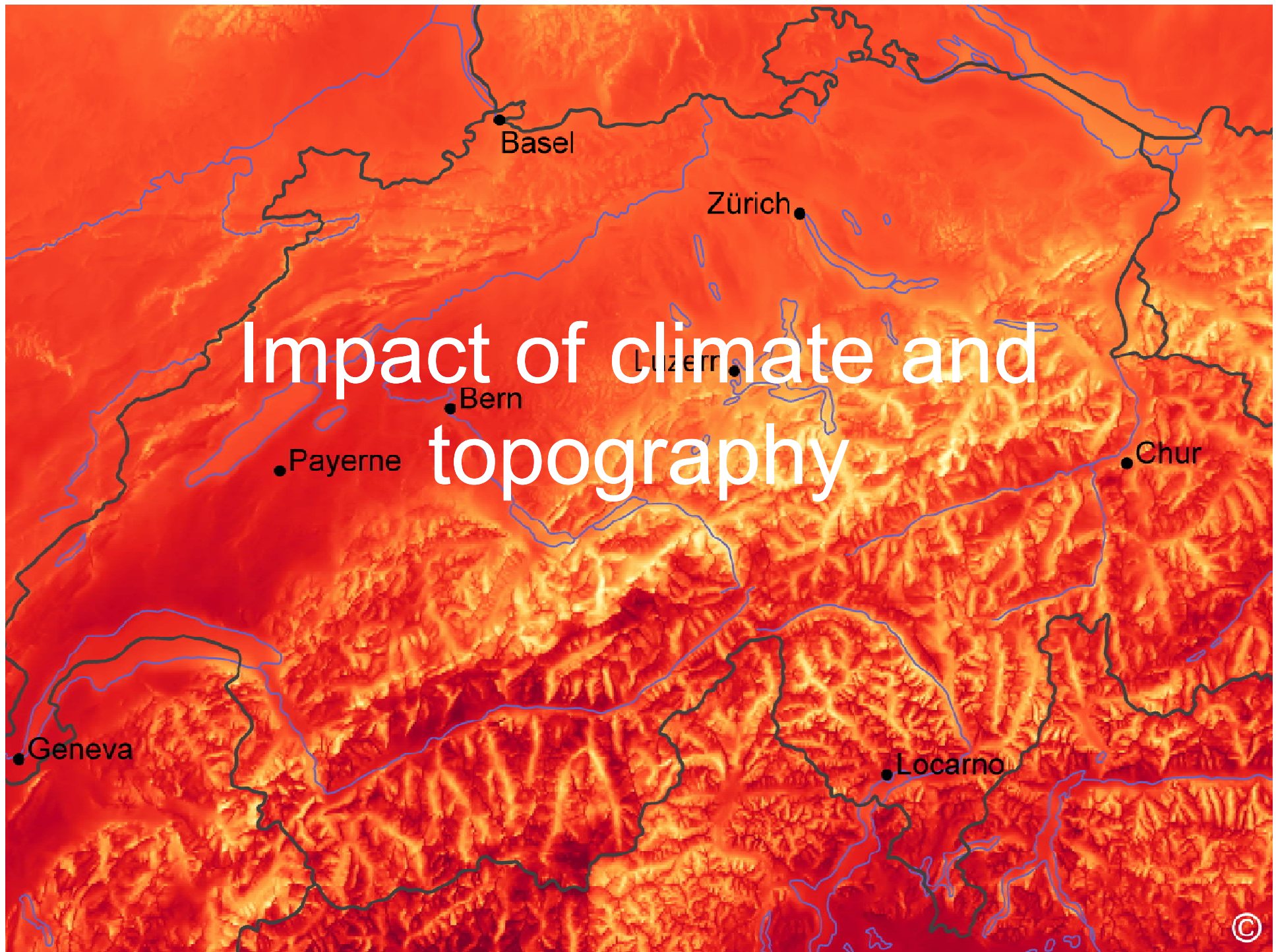
# Example SNC: Altitude and Ischemic Heart Disease



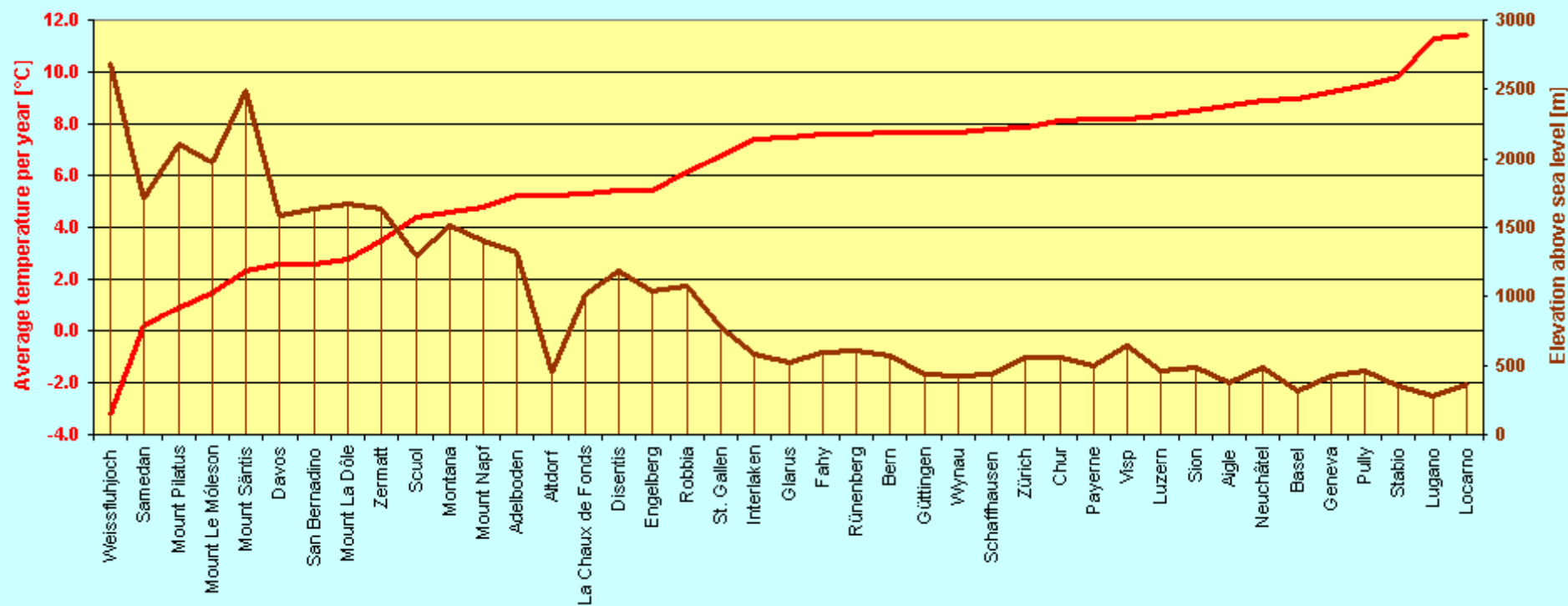
Circulation. 2009 Aug 11;120(6):495-501.

Chiolero, Fäh: Surveillance with Big data: too Big to fail? 4.2.2014

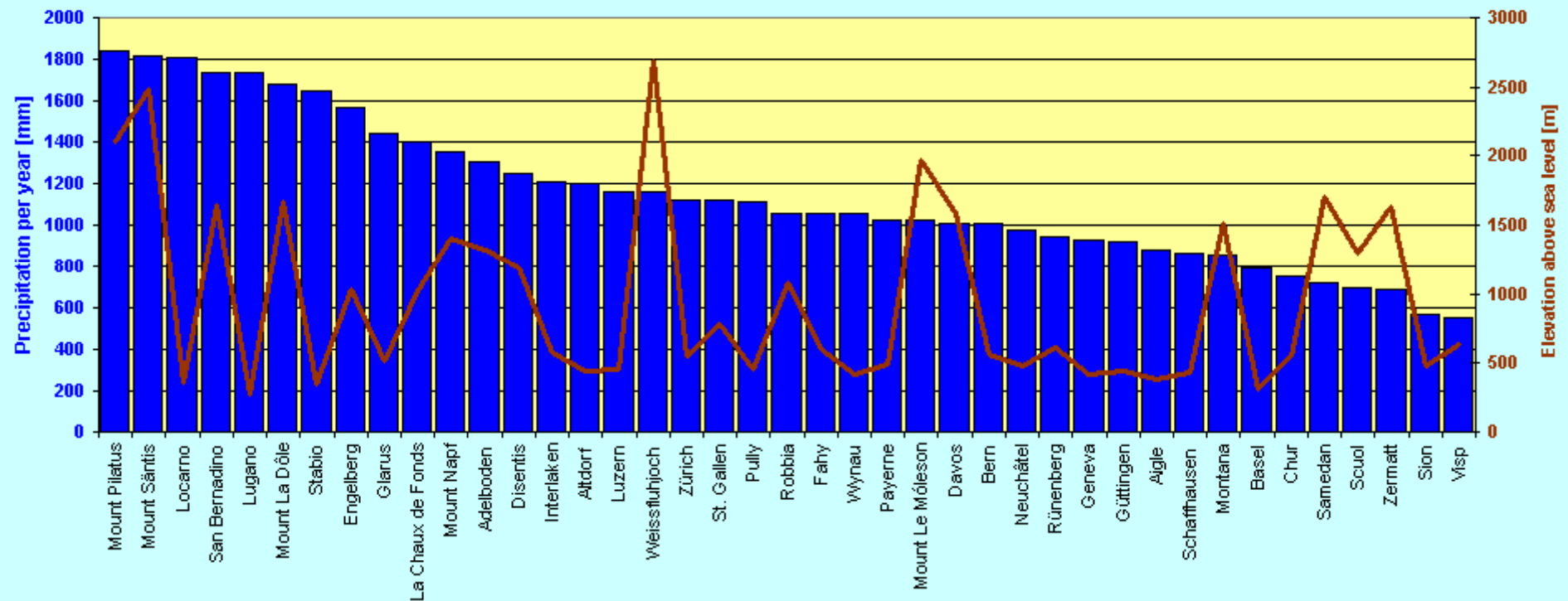
# Impact of climate and topography



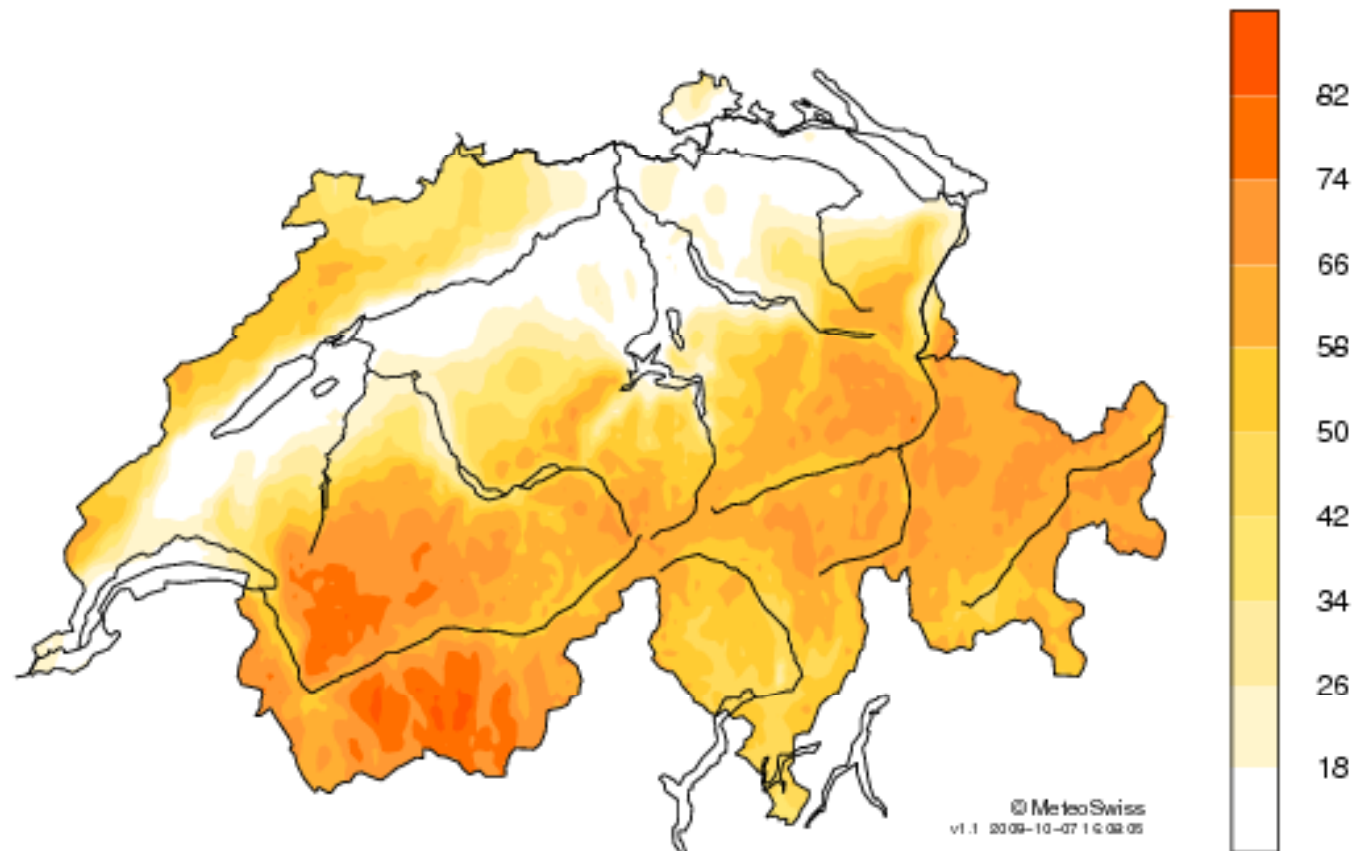
## Temperature in Switzerland



## Precipitation in Switzerland



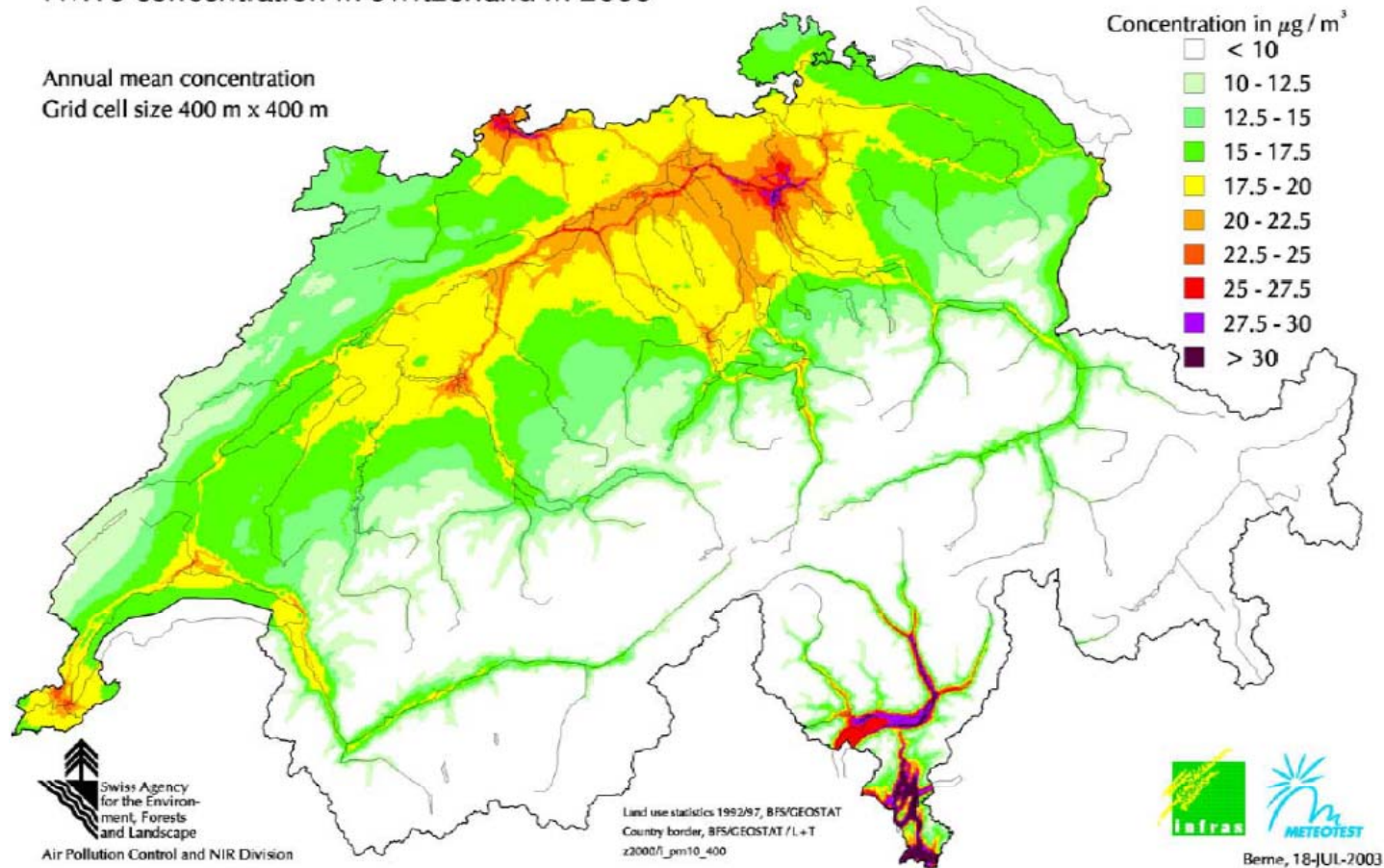
Monthly Relative Sunshine Duration (%) 2009-01



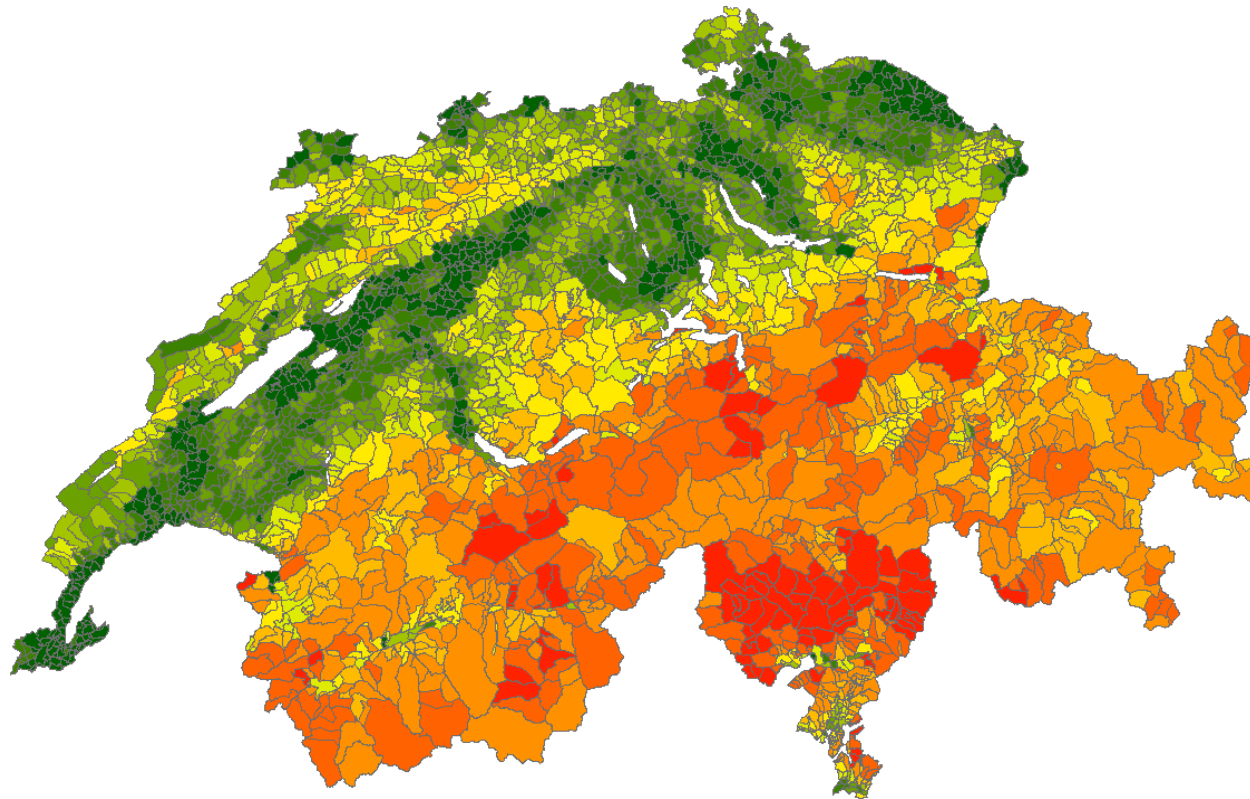


## PM10 concentration in Switzerland in 2000

Annual mean concentration  
Grid cell size 400 m x 400 m



# Slope Index Switzerland



# Large Samples vs. Big Data

- Large Samples
  - Valid, reliable
  - Independent from outcome
  - Individuals, common variables
  - Representative
  - Aim, hypotheses
- Big (cheap) Data

# Surveillance in the age of Big data

## Too big to fail?

PD Arnaud Chiolero MD PhD  
Epidemiologist & senior lecturer

Institute of social and preventive medicine (IUMSP), Lausanne  
Observatoire valaisan de la santé (OVS), Sion

[arnaud.chiolero@chuv.ch](mailto:arnaud.chiolero@chuv.ch)

ISPM, Zurich - February 4, 2014

IUMSP




# "The end of theory"

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES 

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

By Chris Anderson  06.23.08



### THE PETABYTE AGE:

Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

**"All models are wrong, but some are useful."**

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

[www.wired.com/science/discoveries/magazine/16-07/pb\\_theory](http://www.wired.com/science/discoveries/magazine/16-07/pb_theory)

Chris Hendeson, 23.8.2008



# Inevitable big data In health science

- Health care: behind other industries in the domain of information technology and in the use of Big data
- **Massive data generated**, e.g., quantitative (laboratory), qualitative (text-based), or transactional (administrative data)
- Data were considered until now as a **byproduct** of health care delivery, but **perceptions are changing**

Larsen EB. JAMA 2013; 2443-44  
Murdoch TB, Detsky AS. JAMA 2013; 1351-52

# Inevitable big data In health science

- New knowledge based on observational evidence with an important potential of generalizability
- Treatment algorithm using EMR data (with decision based on real-time patient data analyses)
- Help translate personalized medicine into clinical practice (link EMR data and e.g. genomics data)
- Improve safety, quality, and efficiency of care

Larsen EB. JAMA 2013; 2443-44  
Murdoch TB, Detsky AS. JAMA 2013; 1351-52

## THE CHANGING FACE OF EPIDEMIOLOGY

---

*Editors' note: This series addresses topics of interest to epidemiologists across a range of specialties. Commentaries start as invited talks at symposia organized by the Editors. This paper was presented at the 45th Annual meeting of the Society of Epidemiologic Research (SER) in Minneapolis, MN, 2012.*

# Is Size the Next Big Thing in Epidemiology?

*Sengwee Toh and Richard Platt*

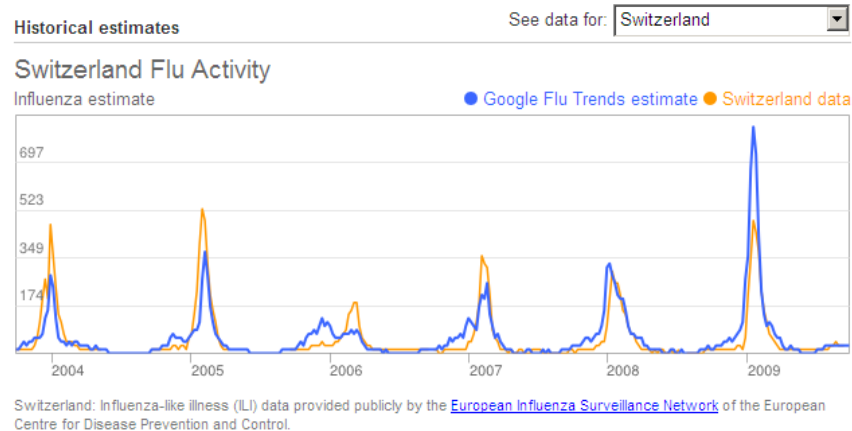
*Epidemiology* • Volume 24, Number 3, May 2013

# Big data hype

- New **wording**
  - Big data, massive data, data deluge, organic data, open data, data-intensive health care, data mining, data analysts, petabyte, exabyte, zettabyte, ...
- Some health-related data become **really analyzable**
  - genetic data
  - **electronic medical records** (EMR)
  - **internet queries** (flu trends)
  - e-patient, self-tracker
  - and more, and more...

# Big data hype

## Google flu trends



### Search Query Topic

Influenza Complication

Cold/Flu Remedy

General Influenza Symptoms

Term for Influenza

Specific Influenza Symptom

Symptoms of an Influenza Complication

Antibiotic Medication

General Influenza Remedies

Symptoms of a Related Disease

Antiviral Medication

Related Disease

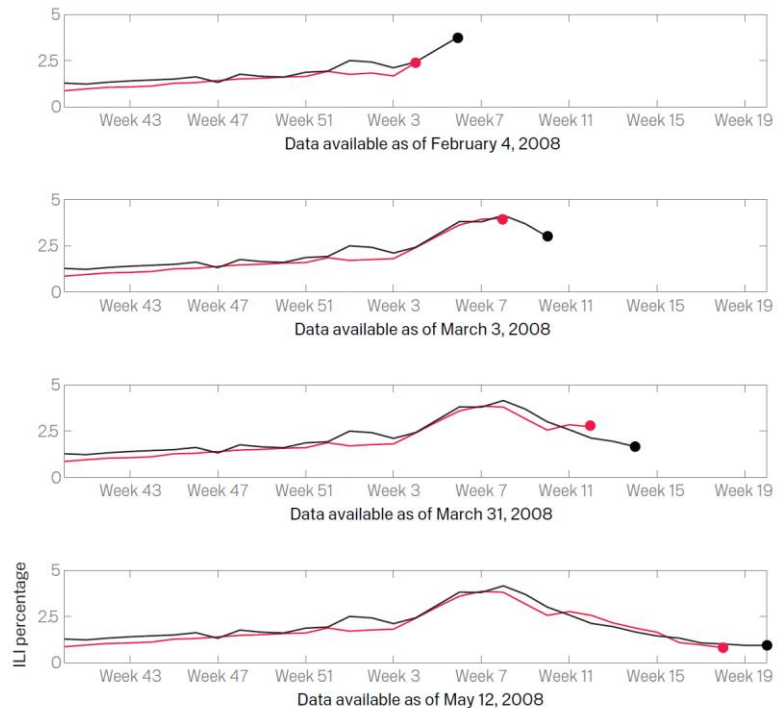


Figure 3: ILI percentages estimated by our model (black) and provided by CDC (red) in the Mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5, we detected a sharply increasing ILI percentage in the Mid-Atlantic region; similarly, on March 3, our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

# Big data hype

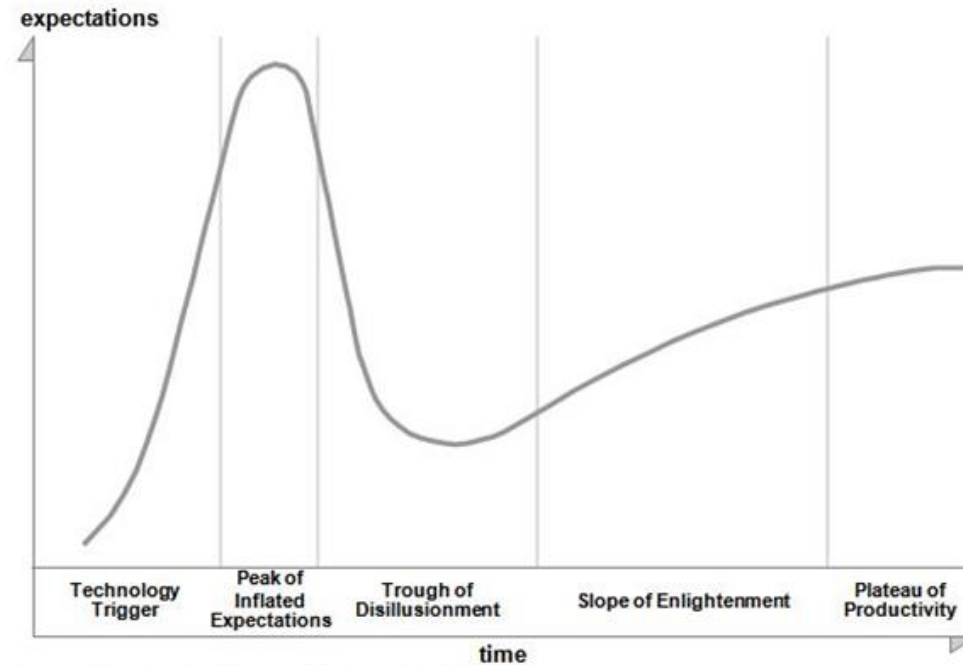
- New analytical methods
  - The end of inferential statistics?
  - The end of p-value? Yeepee hhh!
  - Data driven analyses
  - Data mining
  - Descriptive statistics
  - Correlation and prediction
  - Cluster detection



# Ok but...

## Big Data is Falling into the Trough of Disillusionment

by Svetlana Sicular | January 22, 2013 | 36 Comments



Gartner Hype Cycle: Where is Big Data Now?

<http://blogs.gartner.com/svetlana-sicular/big-data-is-falling-into-the-trough-of-disillusionment/> (accessed 18.8.2013)

# Ok but...

## Big Data in Epidemiology

*Too Big to Fail?*

Chiolero A. Epidemiology 2013; 24(6):938-939

Ok but...

~~Big data~~

Cheap data

**Ok but...**

**Measurement error  
Misclassification  
(zillions of\*) Selection bias  
Confounding**

**...**

**MORE THAN EVER with big cheap data**

\* Ioannidis JPA. Am J Bioethics 2013; 13:40-2

# Public health surveillance

- Public health surveillance is the ongoing systematic collection, analysis, and interpretation of data, closely integrated with the timely dissemination of these data to those responsible for preventing and controlling disease and injury [Lee 2011].
- To provide information useful for decision and action in public health [Lee 2011].

Lee LM, Thacker SB. Public health surveillance and knowing about health in the context of growing sources of health data. Am J Prev Med. 2011;41(6):636-40.

# Surveillance at the age of Big data

- Data gathered more easily and more rapidly
- New data available – especially on health care providers activities
- Linkage between data
- Paradigm change in surveillance method
  - Designed and organic data



# Designed vs. organic data

- Paradigm change in surveillance:
  - Classical process: identify the health problem → define and collect data (finite amount) → analyze data to address the problems
  - Pro: **designed data**, i.e., tailored for your problem [Keller 2012], information on their validity, reliability, and completeness (or its lack)
  - Cons: poor timeliness, limited representativeness, high cost

Keller S et al. Big data and city living – what can it do for us? Significance 2012;8:4-7

# Designed vs. organic data

- Paradigm change in surveillance:
  - eHealth age: all types of data collected from multiple sources without knowing exactly what you will do with these data → analyze data to identify problems and address problems
    - Pro: timeliness, representativeness, low cost
    - Cons: **organic data**, i.e., not tailored for your problem [Keller 2012], quality (?), representativeness, management/storage, privacy/access

Keller S et al. Big data and city living – what can it do for us? Significance 2012;8:4-7

# Surveillance with EMR

- EMR: source of massive amounts of data
  - Quantitative (laboratory)
  - Qualitative (text-based)
  - Transactional (administrative data)
- Prevalence of diabetes in Wallis based on EMR
  - Hospital data (diagnostic code)
  - Laboratory data
  - Pharmacy data
- Data: easily available and analyzable

# Health care information system in Wallis

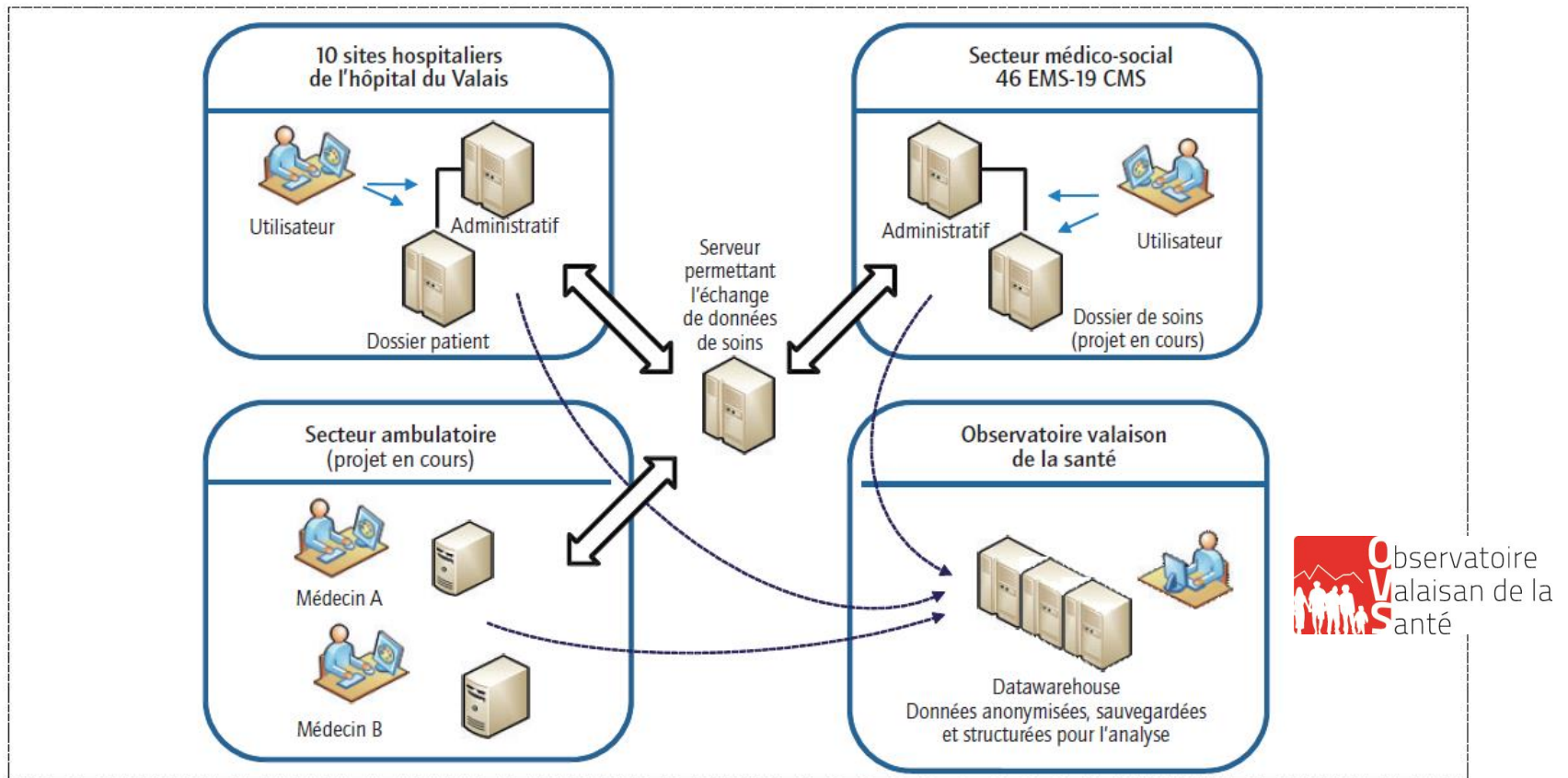


Figure 4 : Système d'information pour récolter les données des prestataires de soins du canton du Valais

Chiolero A, Paccaud F, Fornerod L. Comment faire de la surveillance sanitaire ? L'exemple de l'Observatoire valaisan de la santé en Suisse. Santé Publique 2014; in press.

# What is NOT new

*The American Journal of Bioethics*

## **Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research**

---

**John P. A. Ioannidis**, Stanford University School of Medicine and Stanford University  
School of Humanities and Sciences

---

# Surveillance with EMR

- **Data misreporting** and **lack of standardization** on how health events are defined and recorded in EMR

**Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis**

*Arnaud Chiolero<sup>1,2</sup>, Valérie Santschi<sup>1</sup>, Fred Paccaud<sup>1</sup>*

*<sup>1</sup>Institute of Social and Preventive Medicine (IUMSP), University Hospital Center, Lausanne, Switzerland and <sup>2</sup>Observatoire valaisan de la santé (OVS), Sion, Switzerland*

European Journal of Public Health 2013; 23: 350-351



# What is our question?

- “Formulating a **right question** is always hard, but with big data, it is **an order of magnitude harder**.” [Wired 2013]
- **Let the data speak?**
- Data-driven epidemiology with **flexible data analysis** and **lack of prespecified hypotheses** can lead to research findings that **are not true** [Ioannidis 2005]

[www.wired.com/insights/2013/08/why-big-is-blinding-us-to-the-real-value-of-big-data/](http://www.wired.com/insights/2013/08/why-big-is-blinding-us-to-the-real-value-of-big-data/)

Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.

Chiolero A. Big size epidemiology: too big to fail? Epidemiology 2013; in press

# Conclusions

1. Big cheap data **do not speak by themselves** more than ~~small~~ expensive data
2. **Not the end of theory**
3. **More than ever, we need to know what is our question and why we are making surveillance**

# Thank you for your interest

PD Arnaud Chiolero MD PhD  
Epidemiologist & senior lecturer

[arnaud.chiolero@chuv.ch](mailto:arnaud.chiolero@chuv.ch)

IUMSP

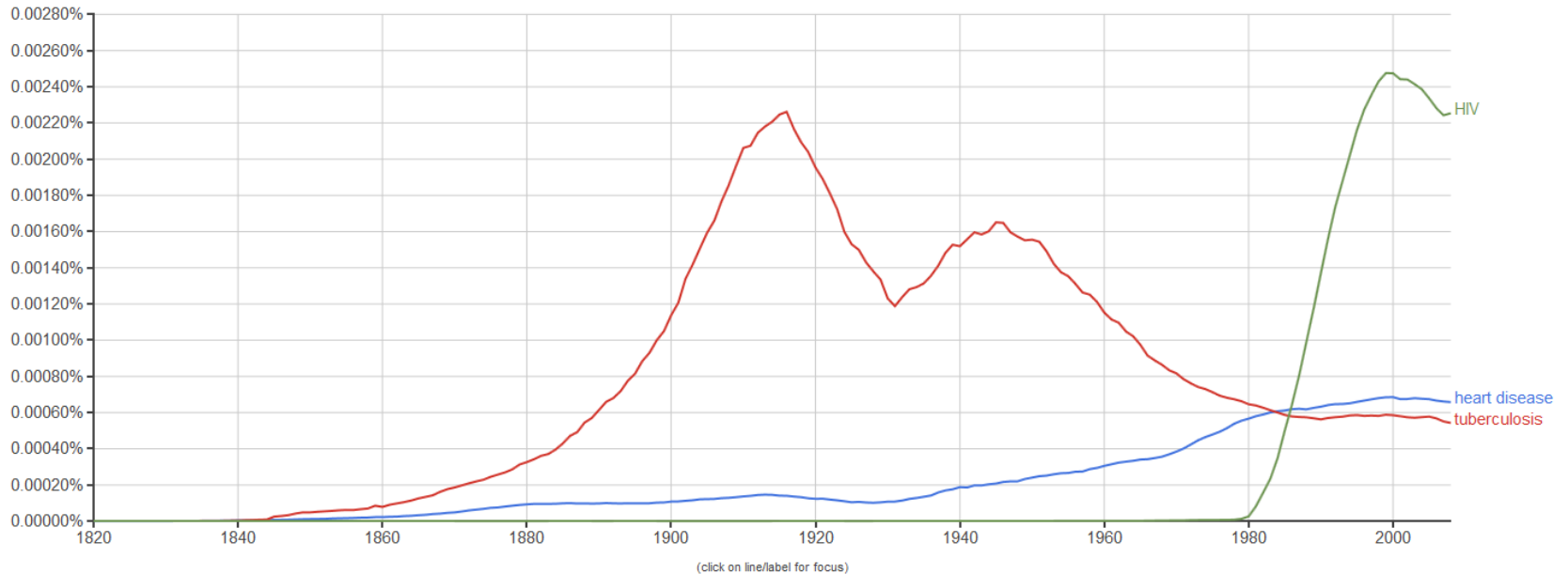


# (zillions of\*) Selection bias? Ngram Viewer

Google books Ngram Viewer

Graph these comma-separated phrases:  ☐ case-insensitive

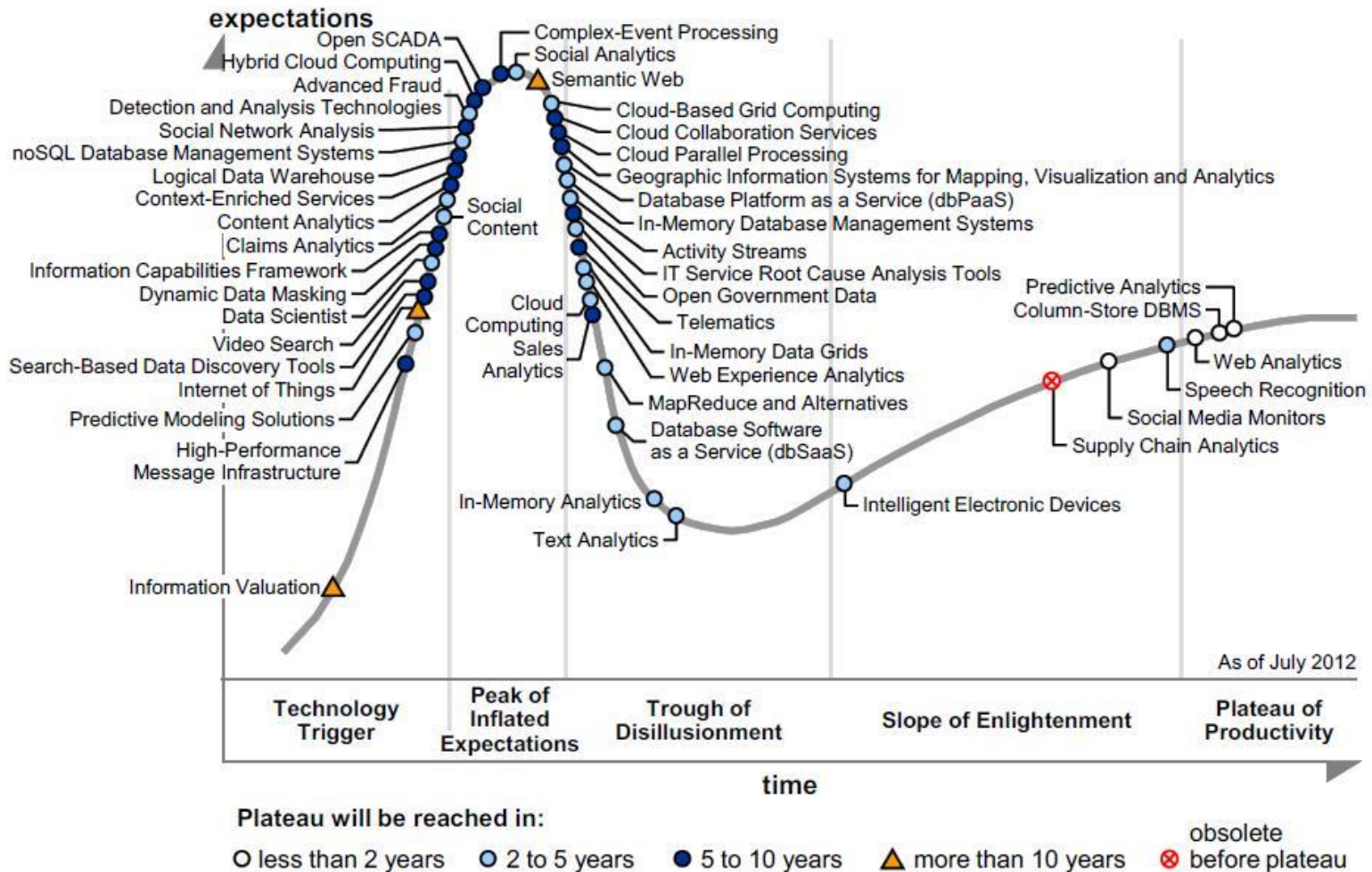
between  and  from the corpus  with smoothing of



<https://books.google.com/ngrams>

<http://www.studio360.org/story/310751-big-data-culturomics/>

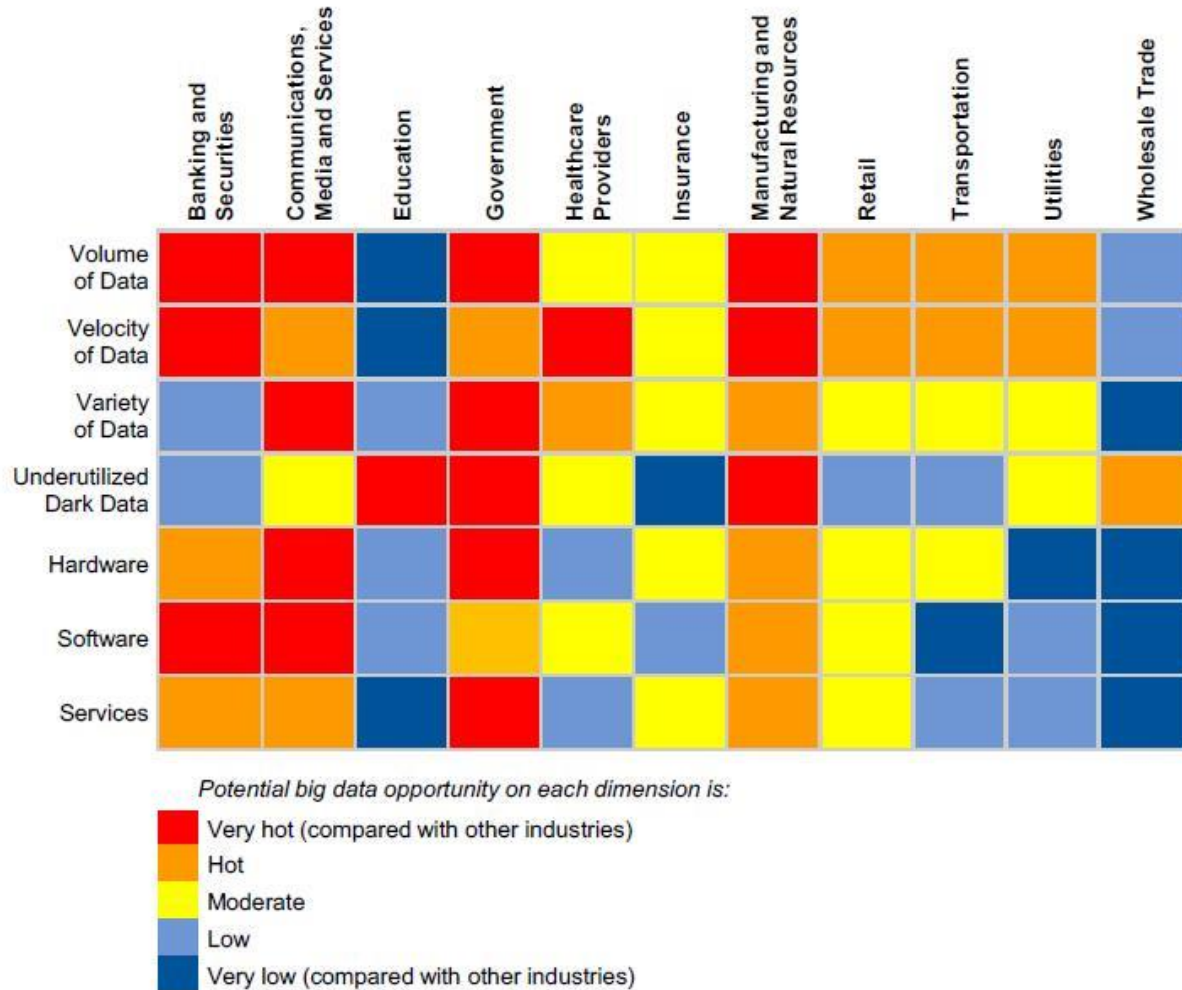
Figure 1. Hype Cycle for Big Data, 2012



Source: Gartner (July 2012)

[www.forbes.com/sites/louiscolumnbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/](http://www.forbes.com/sites/louiscolumnbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/)

Figure 2. Big Data Opportunity Heat Map by Industry



Source: Gartner (July 2012)

[www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/](http://www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/)