

HES-SO, Sierre, September 2016

# Big data: too big to fail?

Arnaud Chiolero MD PhD, PD & MER  
Epidemiologist & Senior Lecturer<sup>1,2</sup>  
Adjunct Professor<sup>3</sup>

- 1) Institute of social and preventive medicine (IUMSP), CHUV, Lausanne
- 2) Observatoire valaisan de la santé (OVS), Sion
- 3) Department of Epidemiology, McGill University, Montreal

[achiolero@gmail.com](mailto:achiolero@gmail.com)

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



*Illustration: Marian Bantjes*

### 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

# BIG DATA

ON THE BOOK

MEET THE AUTHORS

ON TOUR

PRESS

CONTACT



---

How can we spot disease 24 hours before symptoms appear? How can we predict which manholes in New York City may explode next year? Can we really identify criminals before they've committed a crime?

---

Big data: a revolution that will transform how we live, work, and think  
Mayer-Schönberger & Cukier 2013  
<http://big-data-book.com/>

# 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 (record of medication delivery)
- 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

# What is new

- New **wording**
  - Big data, massive data, data deluge, organic data, data tombs, open data, **data-intensive health care**, data mining, **data analysts**, exabyte, zettabyte, ...
- **Precision/personalized medicine**
- **Real-time health data analyses**
- 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...

# What is new Google flu trends

Historical estimates

See data for:

## Switzerland Flu Activity

Influenza estimate

● Google Flu Trends estimate ● Switzerland data



Switzerland: Influenza-like illness (ILI) data provided publicly by the [European Influenza Surveillance Network](#) of the European Centre for Disease Prevention and Control.

### Search Query Topic

Influenza Complication

Cold/Flu Remedy

General Influenza Symptoms

Term for Influenza

Specific Influenza Symptom

Symptoms of an Influenza Complication

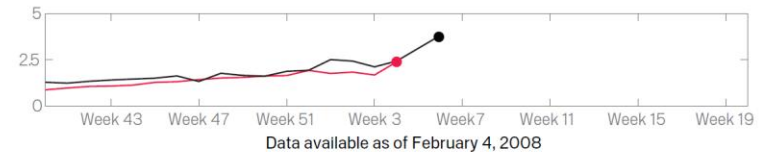
Antibiotic Medication

General Influenza Remedies

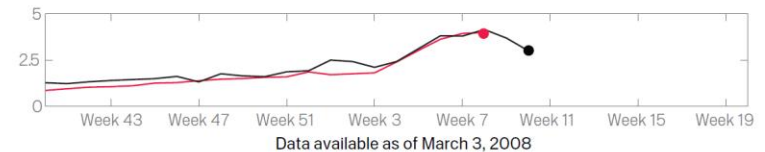
Symptoms of a Related Disease

Antiviral Medication

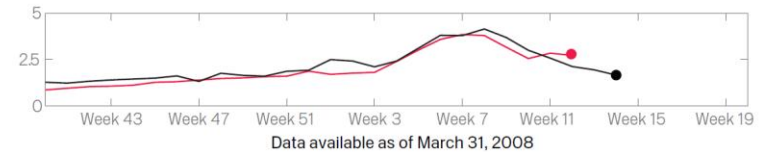
Related Disease



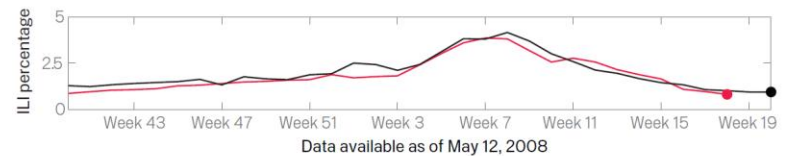
Data available as of February 4, 2008



Data available as of March 3, 2008



Data available as of March 31, 2008



Data available as of May 12, 2008

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.



# What is new Racial dot map

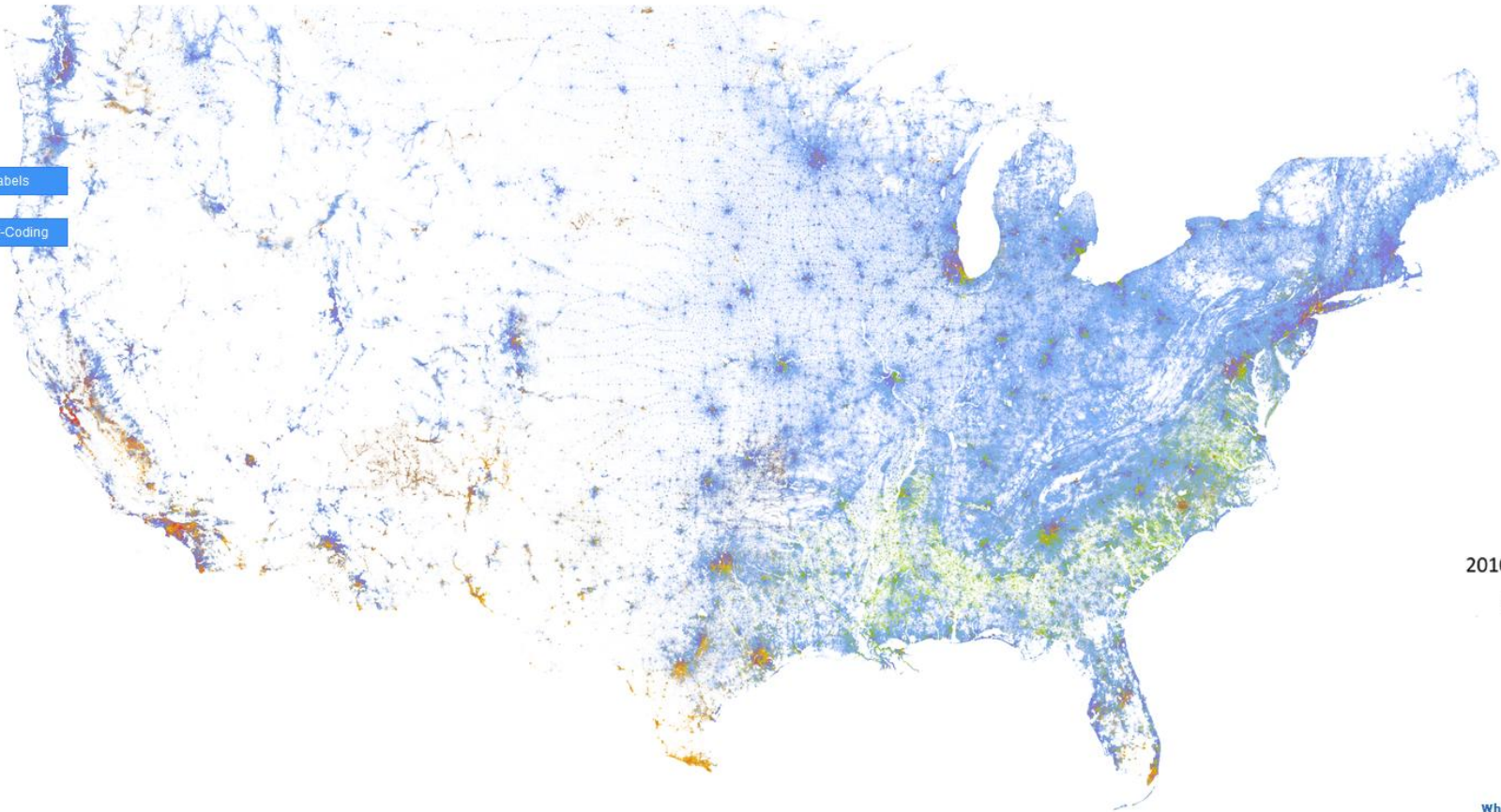


Hide Overlays



Add Map Labels

Remove Color-Coding



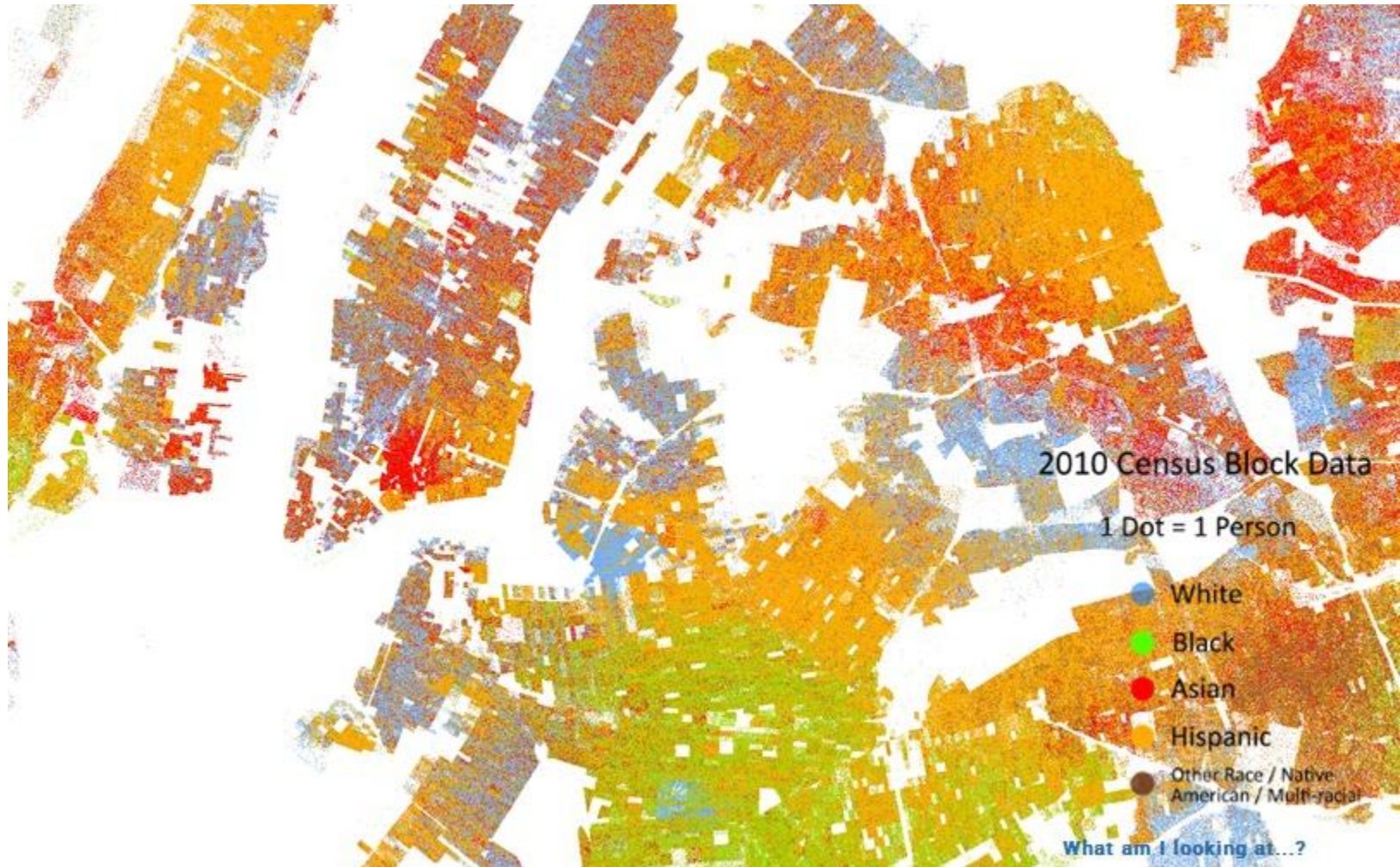
2010 Census Block Data

1 Dot = 1 Person

- White
- Black
- Asian
- Hispanic
- Other Race / Native American / Multi-racial

What am I looking at...?

# What is new Racial dot map



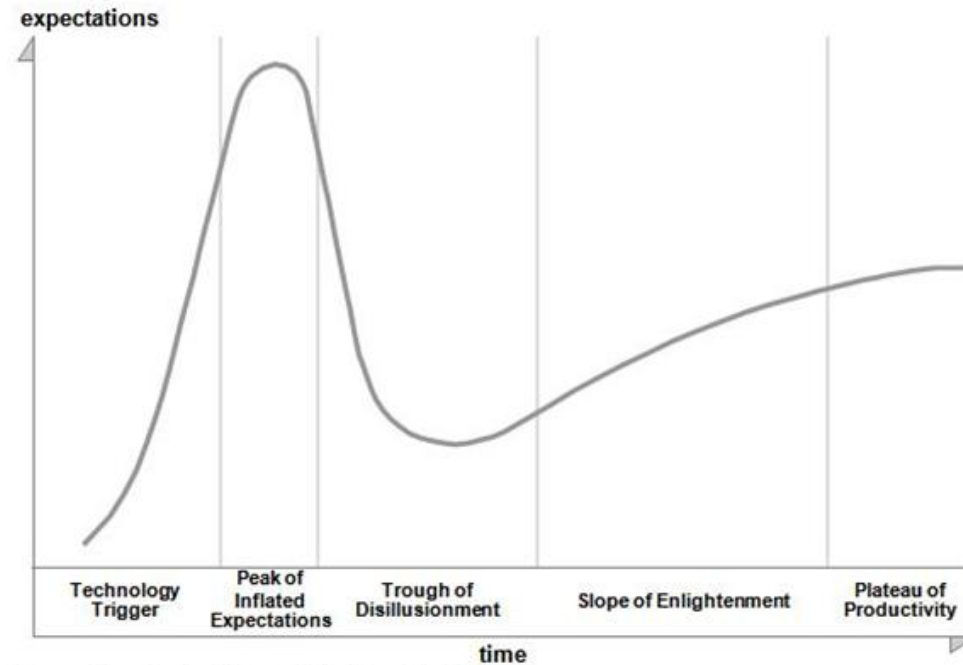
# What is new

- New analytical methods
  - The end of inferential statistics?
  - The end of p-value? Yeepeehhh!
  - Data driven analyses, data mining
- The end of causality?
  - Descriptive statistics
  - Correlation and prediction
  - Discovery of hidden relationships and cluster

# 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

# What is NOT new

~~Big data~~

Cheap data

# What is NOT new

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 gathered – especially on health 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 (?)**, **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

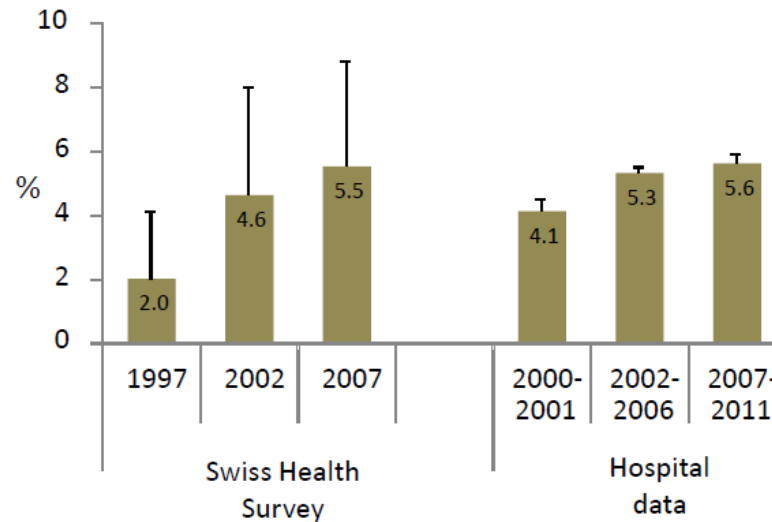
- Prevalence of diabetes in Wallis based on EMR
  - Hospital data (diagnostic code)
  - Laboratory data
  - Pharmacy data
- Data: easily available and analyzable

## **Public Health Surveillance with Big Data: Assessing Diabetes Trends using Medico-Administrative Data**

Christian Ambord<sup>1</sup>; Frédéric Favre<sup>2</sup>; David Faeh<sup>3</sup>; Arnaud Chiolero<sup>2,4</sup>

1. Service de la santé publique, Sion; 2. Observatoire valaisan de la santé, Sion; 3. Institute of Social and Preventive Medicine, University of Zurich; 4. Institute of Social and Preventive Medicine, Lausanne University Hospital, Switzerland

# Surveillance with EMR



**Figure 3:** Trends in the prevalence of diabetes between 1997 and 2007 (SHS) and between 2000 and 2011 (hospital data). The upper bound of the 95% confidence interval is indicated.

- Ok but...
  - Diagnosis of diabetes in EMR: reliability? validity?
  - Hospitalization risk in diabetic and non-diabetic patient?
  - Source population?

# Surveillance with EMR

- Poor data quality 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

- Accessibility ≠ Validity

# Too much data?

- Data ≠ Information ≠ Knowledge
  - How to produce knowledge from so much data & information?
- “Instead of starting with the data, start with your business objectives and what you are specifically trying to achieve. This will automatically point you towards questions that you need to answer, which will narrow data requirements into manageable areas.”
- What is your question?

[www.cgma.org/magazine/features/pages/big-data-alternatives-bernard-marr.aspx?TestCookiesEnabled=redirect](http://www.cgma.org/magazine/features/pages/big-data-alternatives-bernard-marr.aspx?TestCookiesEnabled=redirect)

# 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 J. Why most published research findings are false. PLoS Med 2005;2:e124]

[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/)

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



# 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 research question and why we are making surveillance**

# Thank you for your interest

[achiolero@gmail.com](mailto:achiolero@gmail.com)

IUMSP

  
UNIL | Université de Lausanne

 **McGill**

 Observatoire  
Valaisan de la  
Santé

# What is NOT new

## MEDICINE

### *Big data meets public health*

Human well-being could benefit from large-scale data if large-scale noise is minimized

By Muin J. Khoury<sup>1,2</sup> and  
John P. A. Ioannidis<sup>3</sup>

5.7

For nongenomic associations, false associations are more likely due to confounding variables or other biases than to true associations.

- Signal vs noise
- At risk of false alarm
- “Big error can plague Big data”

Khoury & Ioannidis. Science 2014; 6213

# What is new

Annals of Internal Medicine

IDEAS AND OPINIONS

## Two Ways of Knowing: Big Data and Evidence-Based Medicine

Ida Sim, MD, PhD

