Big Data: Big Challenges and Big Concerns

"The Future of Science"
April 4th 2017

Carlo Batini

Dipartimento di Informatica, Sistemistica e Comunicazione, Università di Milano-Bicocca

batini@disco.unimib.it

Ho cominciato a riflettere sui Big data……

Corso di Laurea magistrale in Data Science

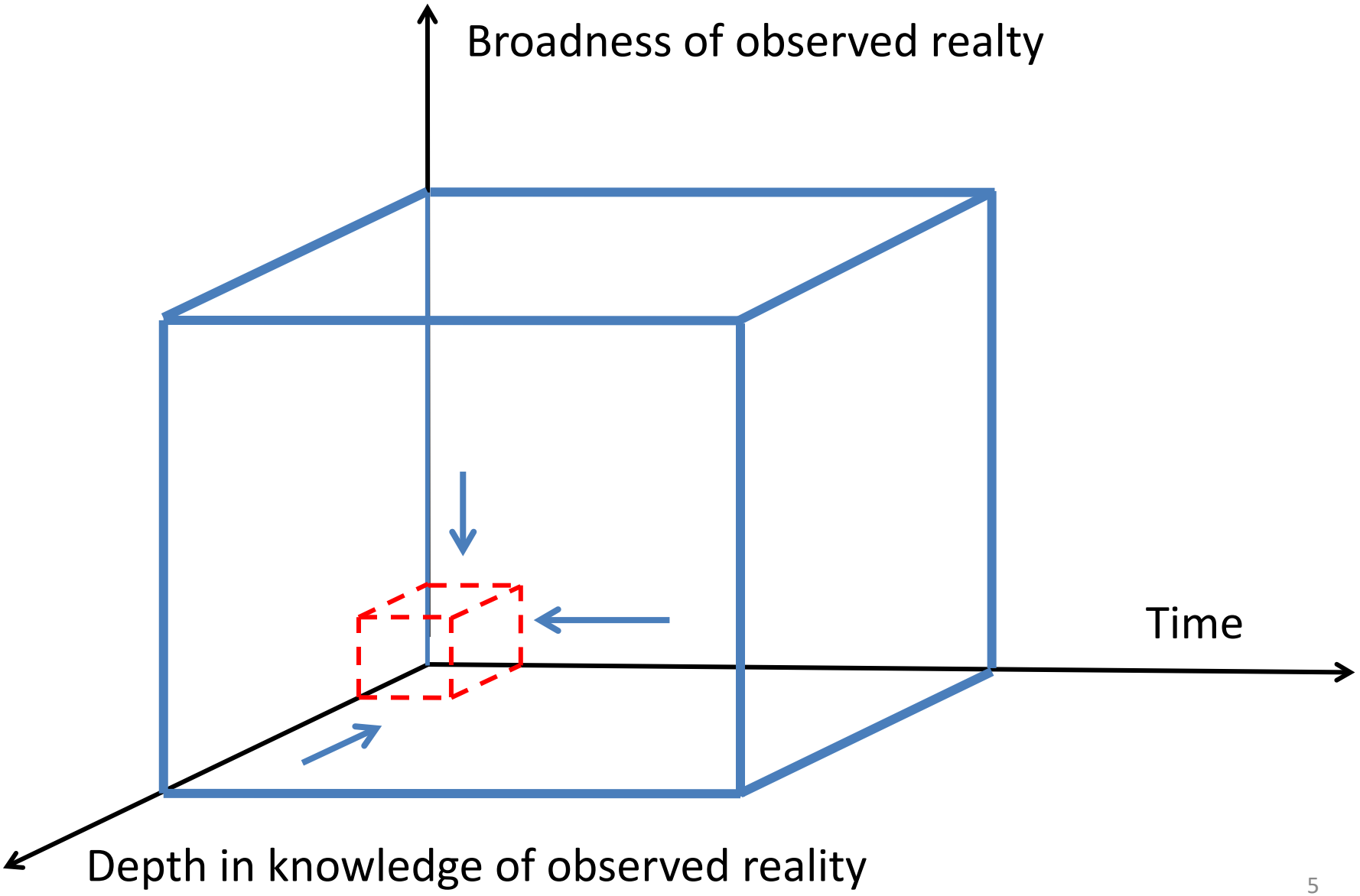
approvato dalla Università di Milano-Bicocca, in corso di accreditamento presso il MIUR

When we speak of Big Data..

…we refer, often unconsciously, to several media:

- Social Networks (es. Facebook, Twitter, etc.)
- Internet of Things
- Digital newspapers
- TV
- etc.

Small data: from the Universe to a sample

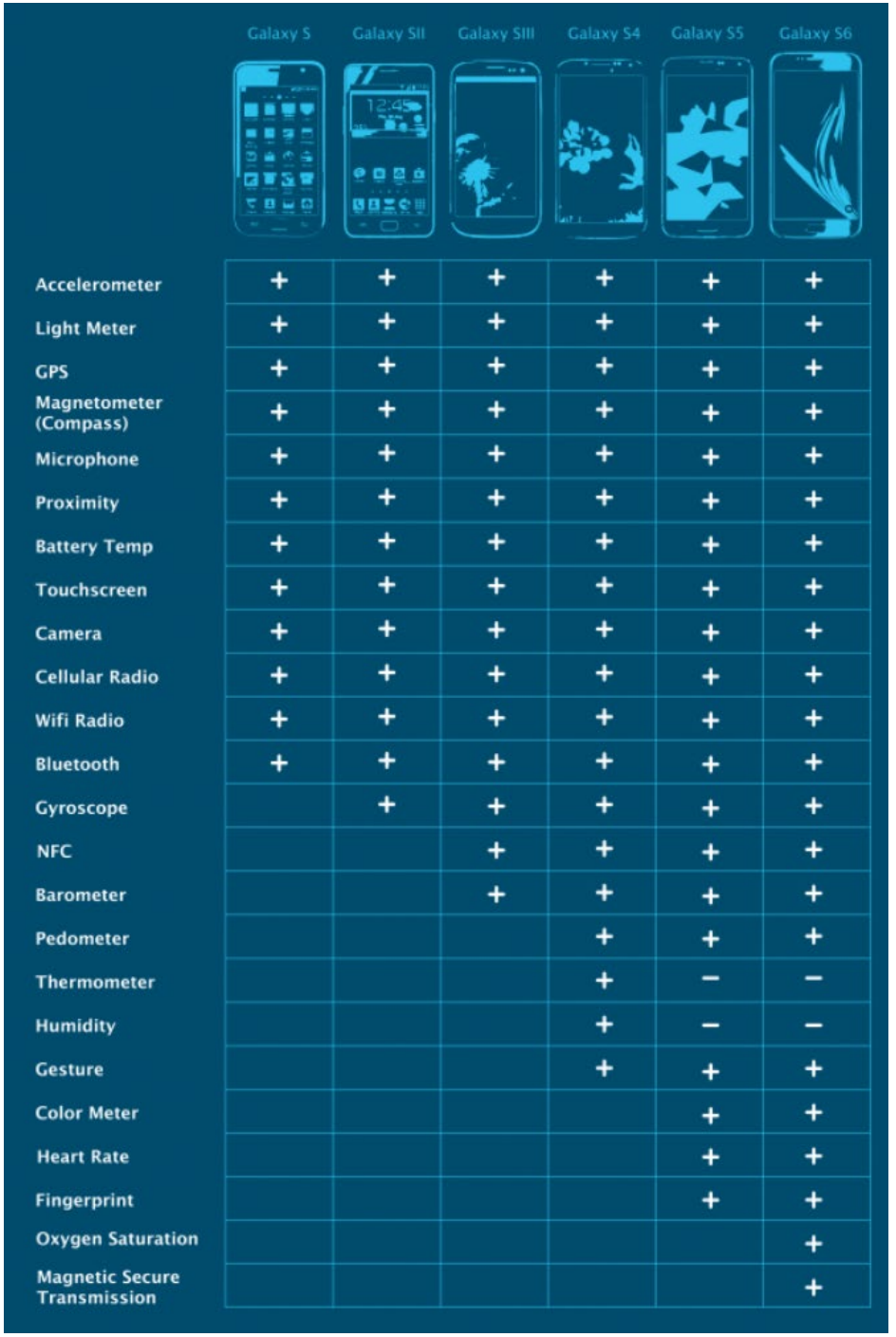


Esempio: i Censimenti negli Stati Uniti

Il censimento del 1880 negli Stati Uniti richiese 8 anni per essere completato

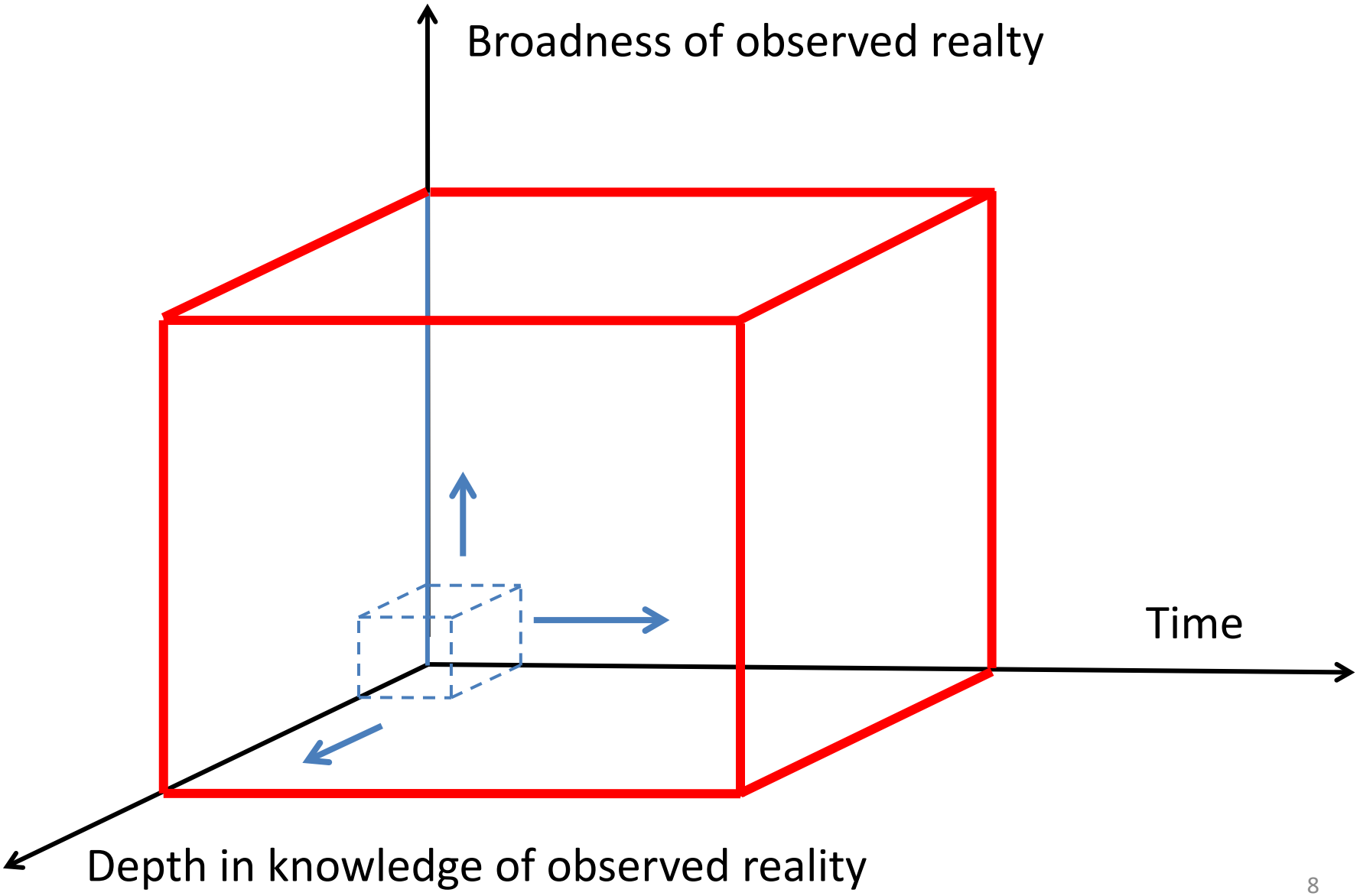
→ i dati diventavano obsoleti ben prima di diventare disponibili e utili

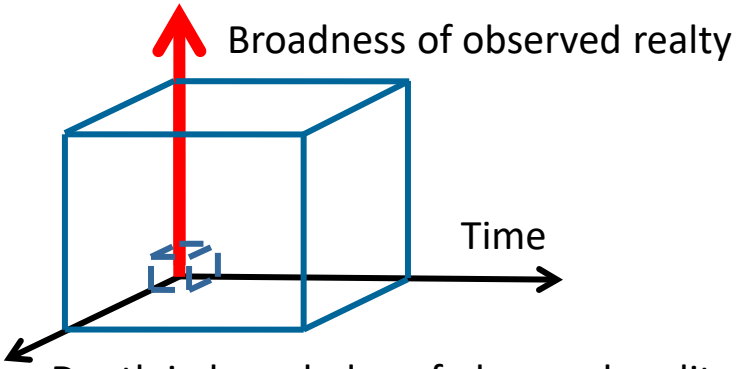
Samsung Galaxy Sensor evolution



	Galaxy S	Galaxy SII	Galaxy SIII	Galaxy S4	Galaxy S5	Galaxy S6
Accelerometer	+	+	+	+	+	+
Light Meter	+	+	+	+	+	+
GPS	+	+	+	+	+	+
Magnetometer (Compass)	+	+	+	+	+	+
Microphone	+	+	+	+	+	+
Proximity	+	+	+	+	+	+
Battery Temp	+	+	+	+	+	+
Touchscreen	+	+	+	+	+	+
Camera	+	+	+	+	+	+
Cellular Radio	+	+	+	+	+	+
Wifi Radio	+	+	+	+	+	+
Bluetooth	+	+	+	+	+	+
Gyroscope		+	+	+	+	+
NFC			+	+	+	+
Barometer			+	+	+	+
Pedometer				+	+	+
Thermometer				+	–	–
Humidity				+	–	–
Gesture				+	+	+
Color Meter					+	+
Heart Rate					+	+
Fingerprint					+	+
Oxygen Saturation						+
Magnetic Secure Transmission						+

From small data to big data

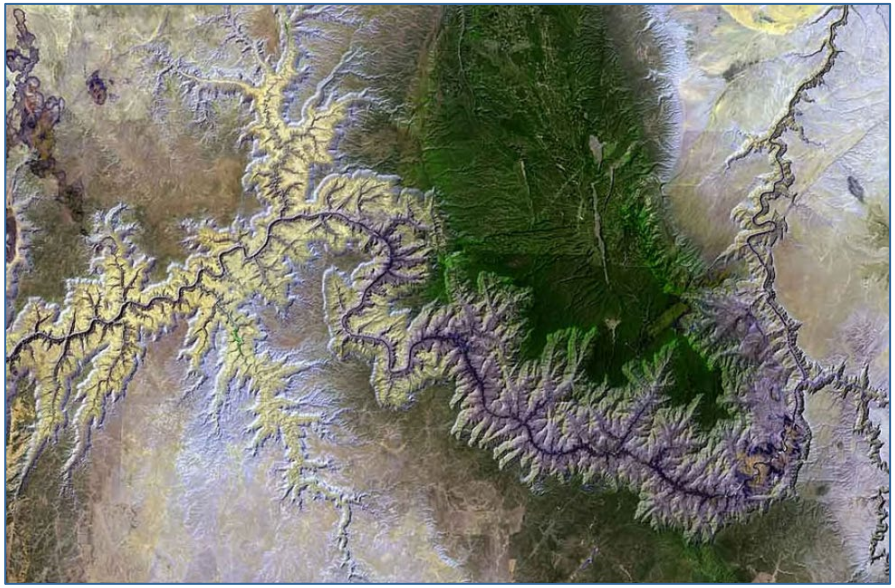




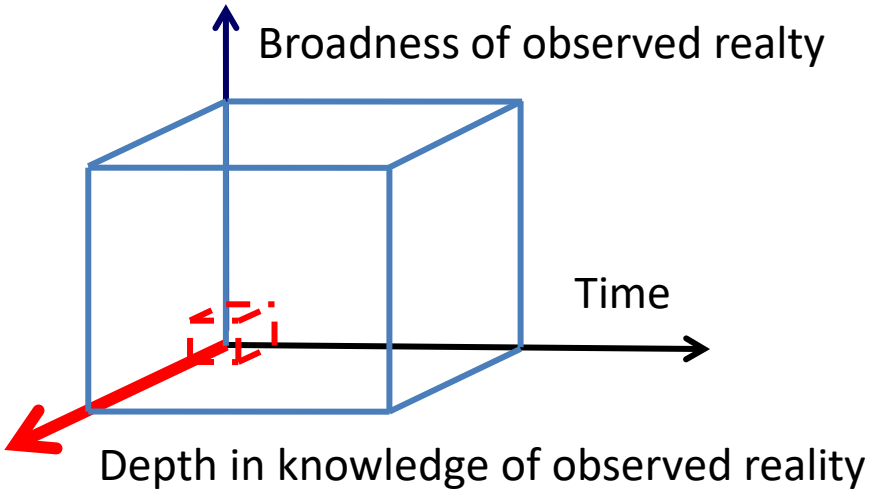
Verso la mappa «uno a uno» del mondo

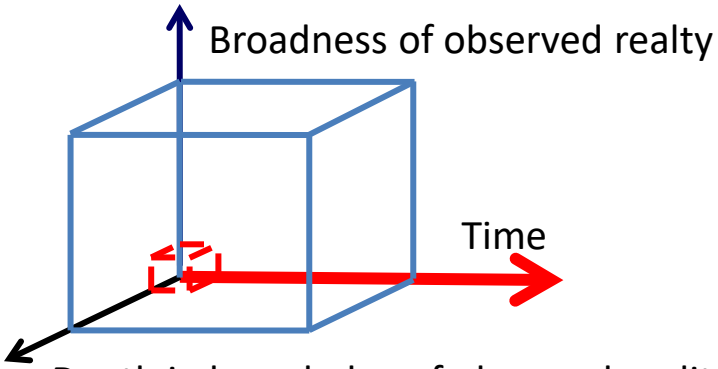
From Hecateus Map (520 B.C.)…

… to the «one to one» map of Babilonian Geographers



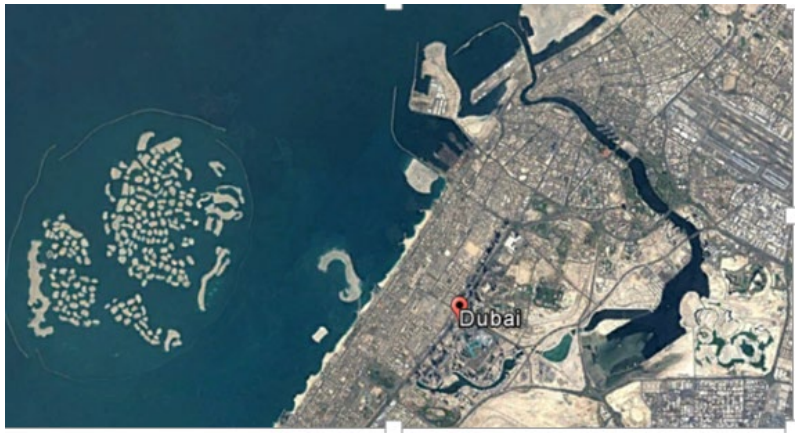
I pneumatici intelligenti





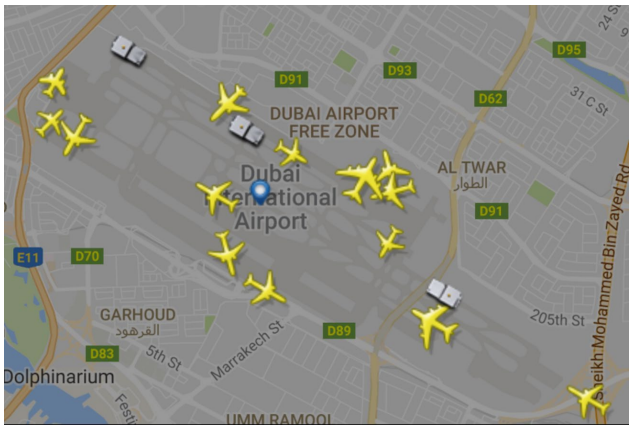
La evoluzione nel tempo

un mese

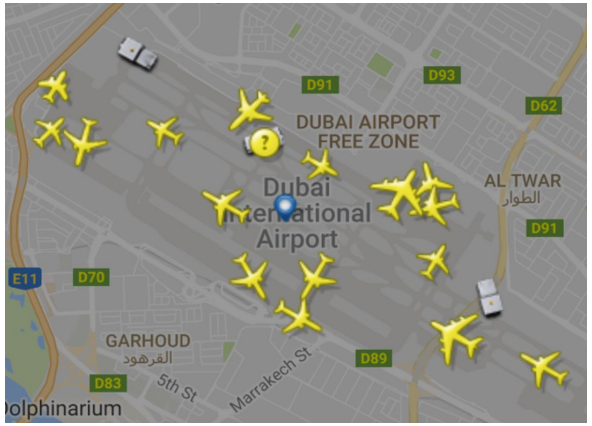


Google Earth, Dubai, **1984**

Google Earth, Dubai, **2015**



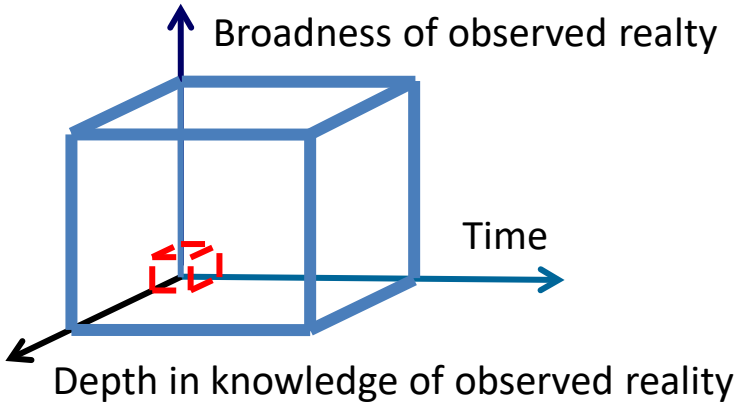
un secondo

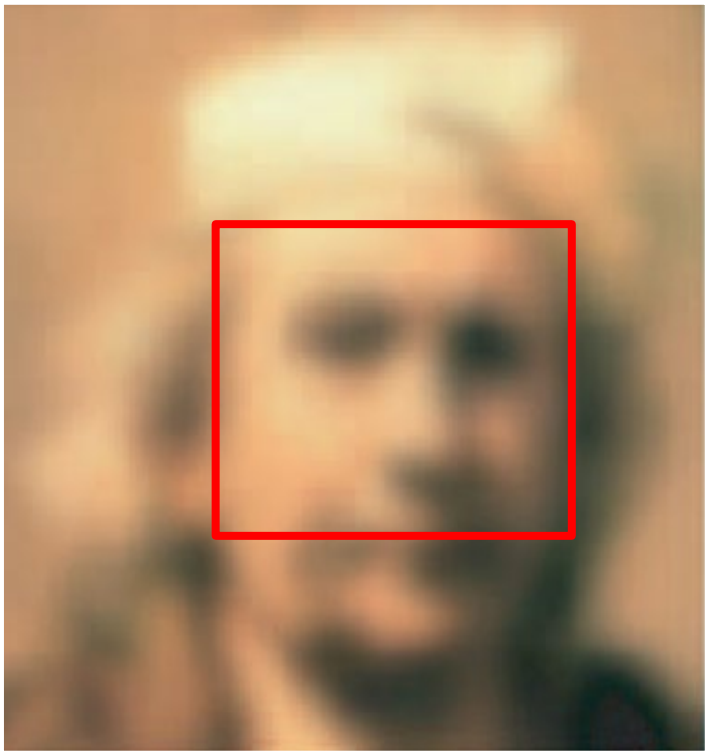


FlightRadar, Dubai 11:05:30 4:3:2017

FlightRadar, Dubai 11:05:35 4:3:2017

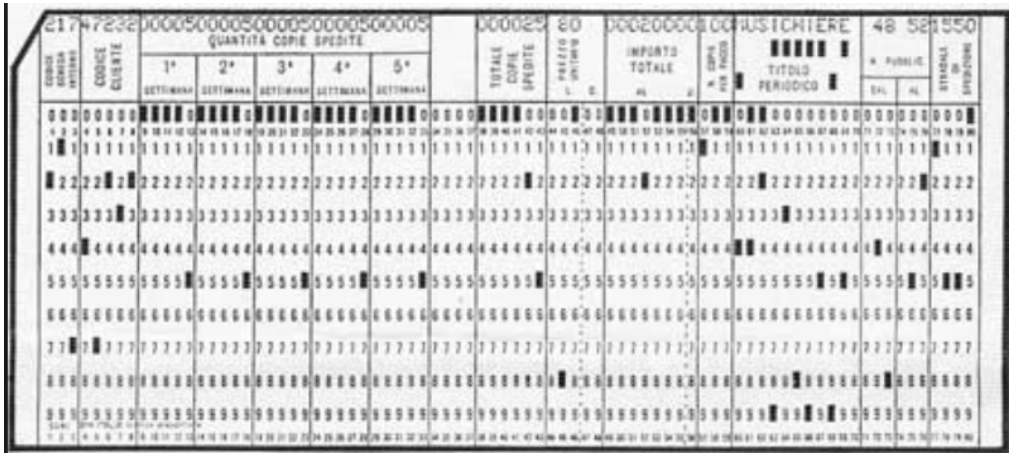
Attenzione: potrebbe anche peggiorare…





Le prime tecnologie: la scheda Hollerith

- Il censimento U.S.A. del 1880 richiese 8 anni per essere completato → i dati diventavano obsoleti ben prima di diventare disponibili
- Per il censimento del 1890 fu adottata la scheda Hollerith….



…portando il tempo di calcolo da 8 anni a meno di uno…

Techniques and technologies for Volume, Velocity, Variety

- **Volume** – the amount of data that can be collected and stored
- **Velocity** – the speed at which data can be captured; and
- **Variety** – encompassing both structured (organized and stored in tables and relations) and unstructured (text, imagery) data

Big Data are much more than
Small Data + Small Data + Small Data…

BD request for a change of paradigm…

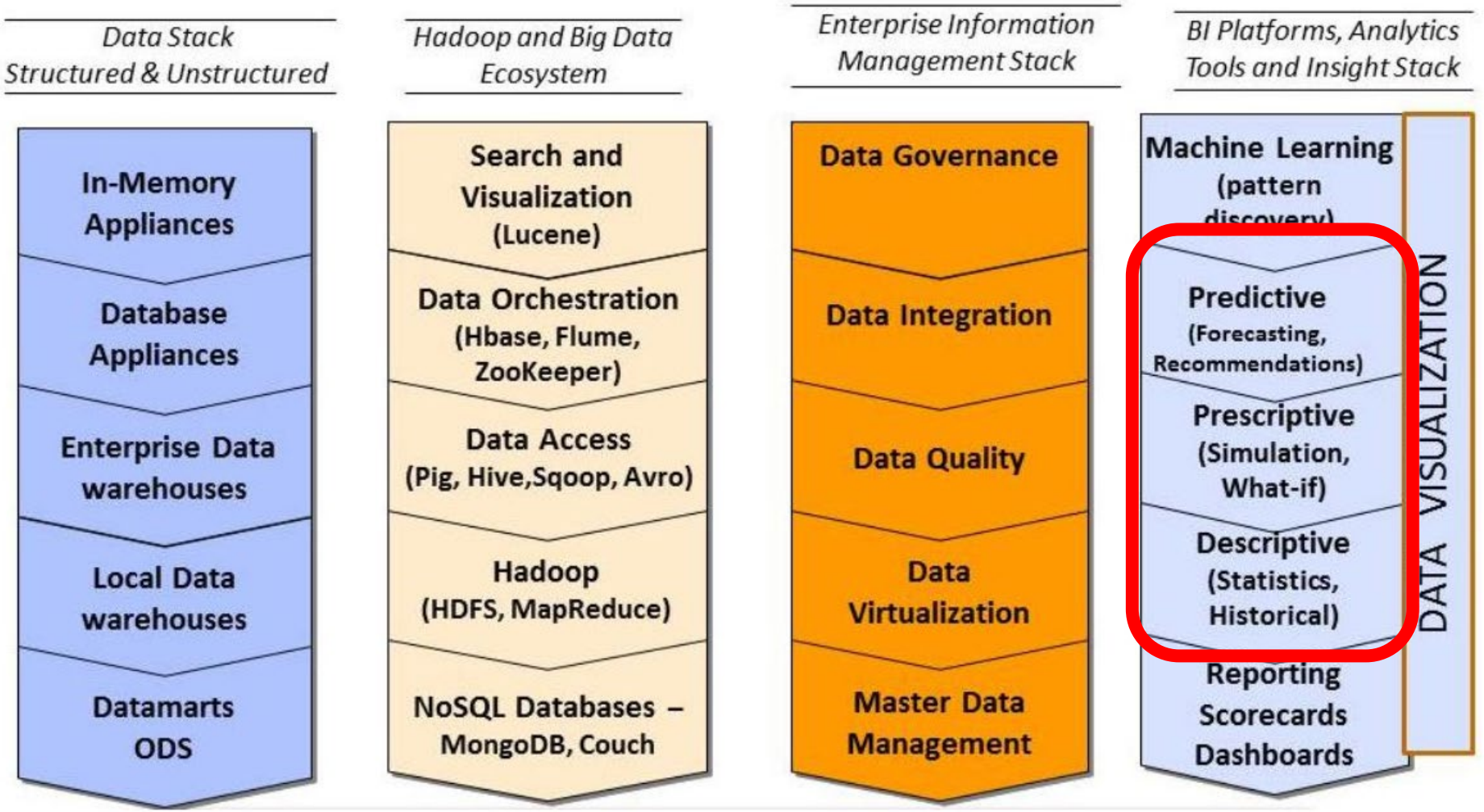
.. in the data life cycle

Life cycle

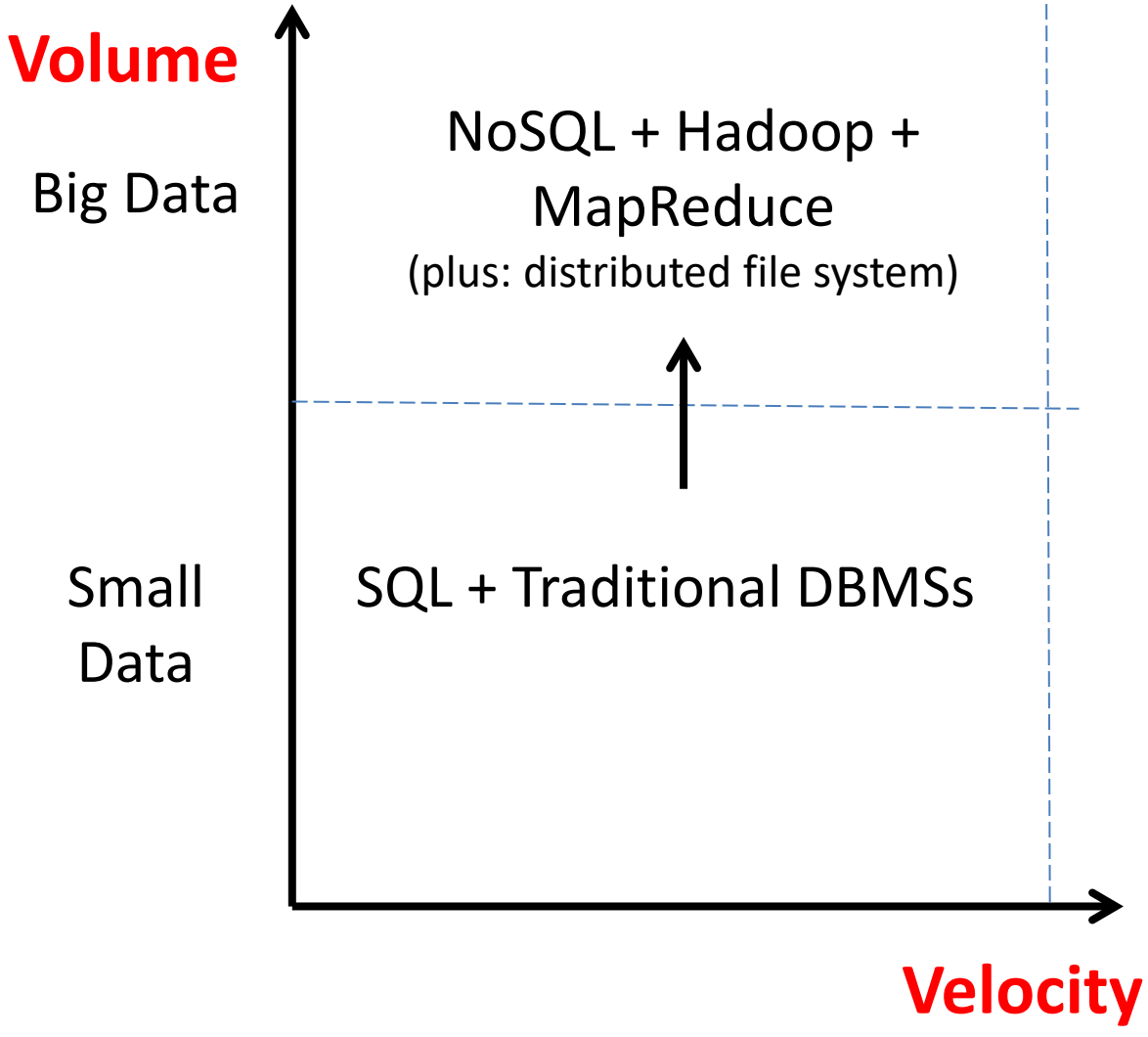
Extract
Transform
Load

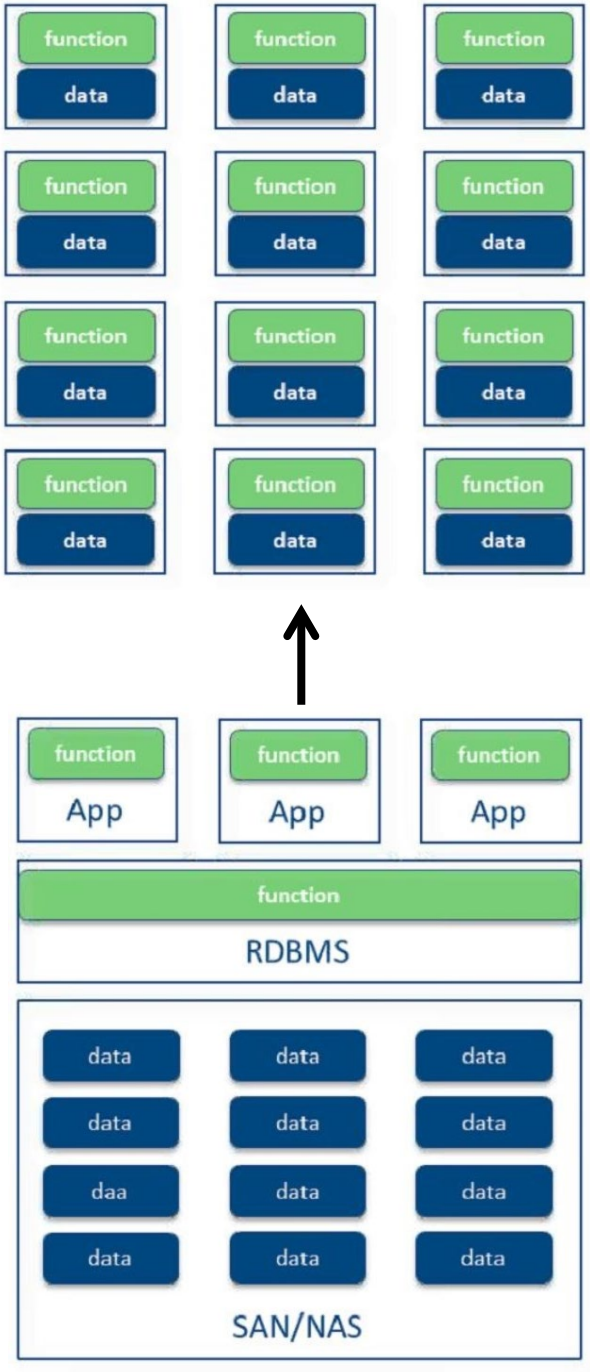
Life cycle	**Cross cutting activities**			
Source Selection & Extraction	SEMANTICS	QUALITY	LEARNING	VALUE
Storage				
Integration				
Analysis				
Visualization				

Big Data Analytics Infrastructure: Rose Technology

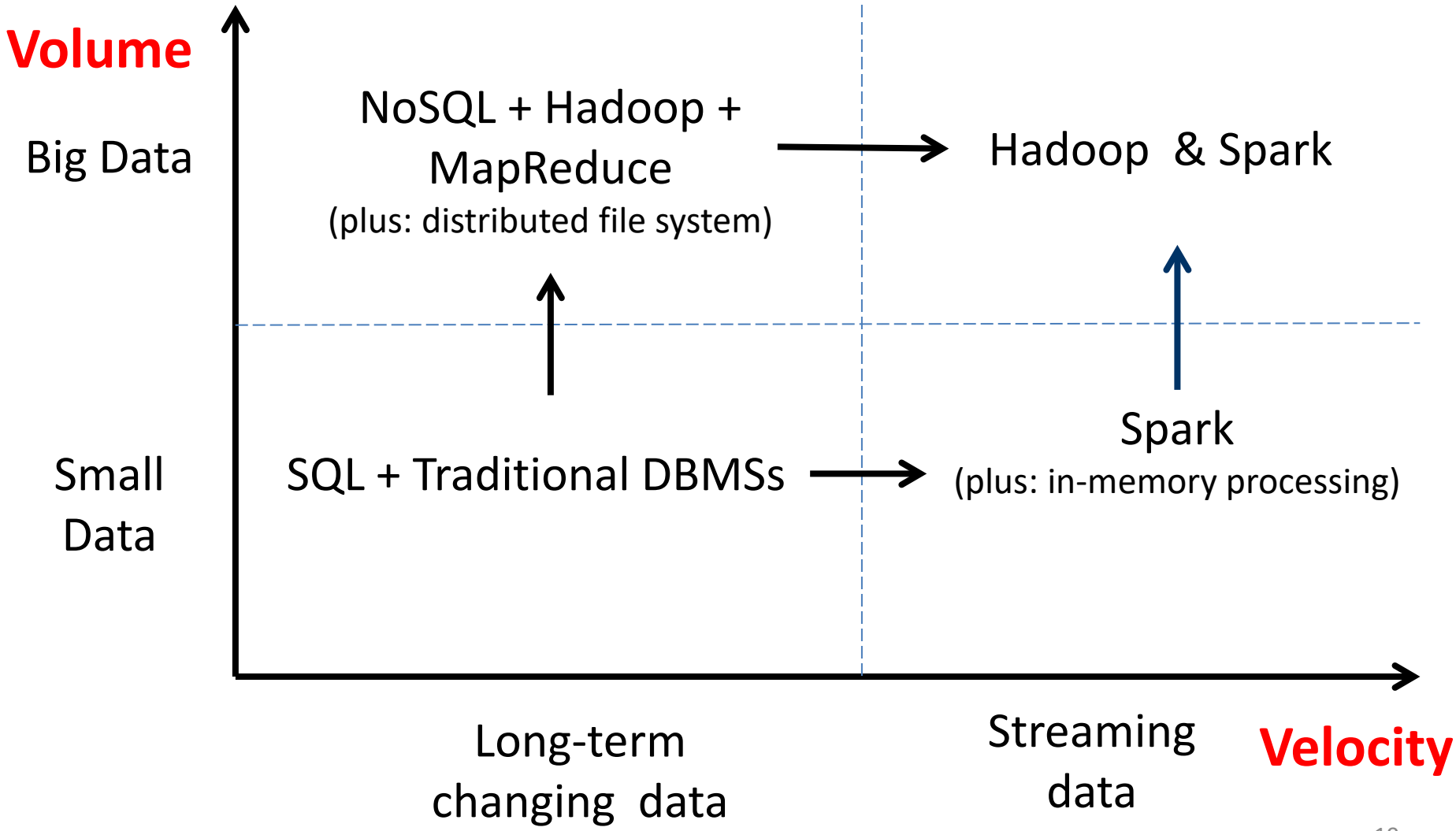


… in Data Management Systems

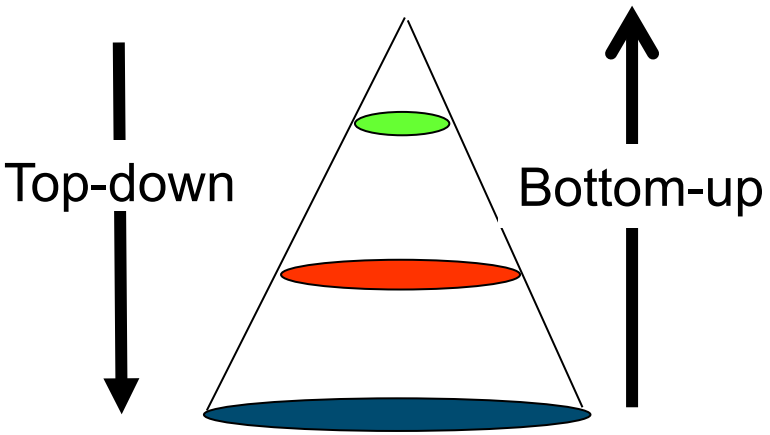


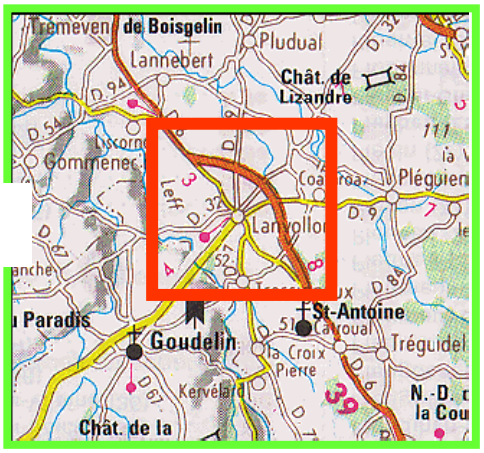


… in Data Management Systems



… in Machine Learning Techniques





Volume

Big Data

Small Data

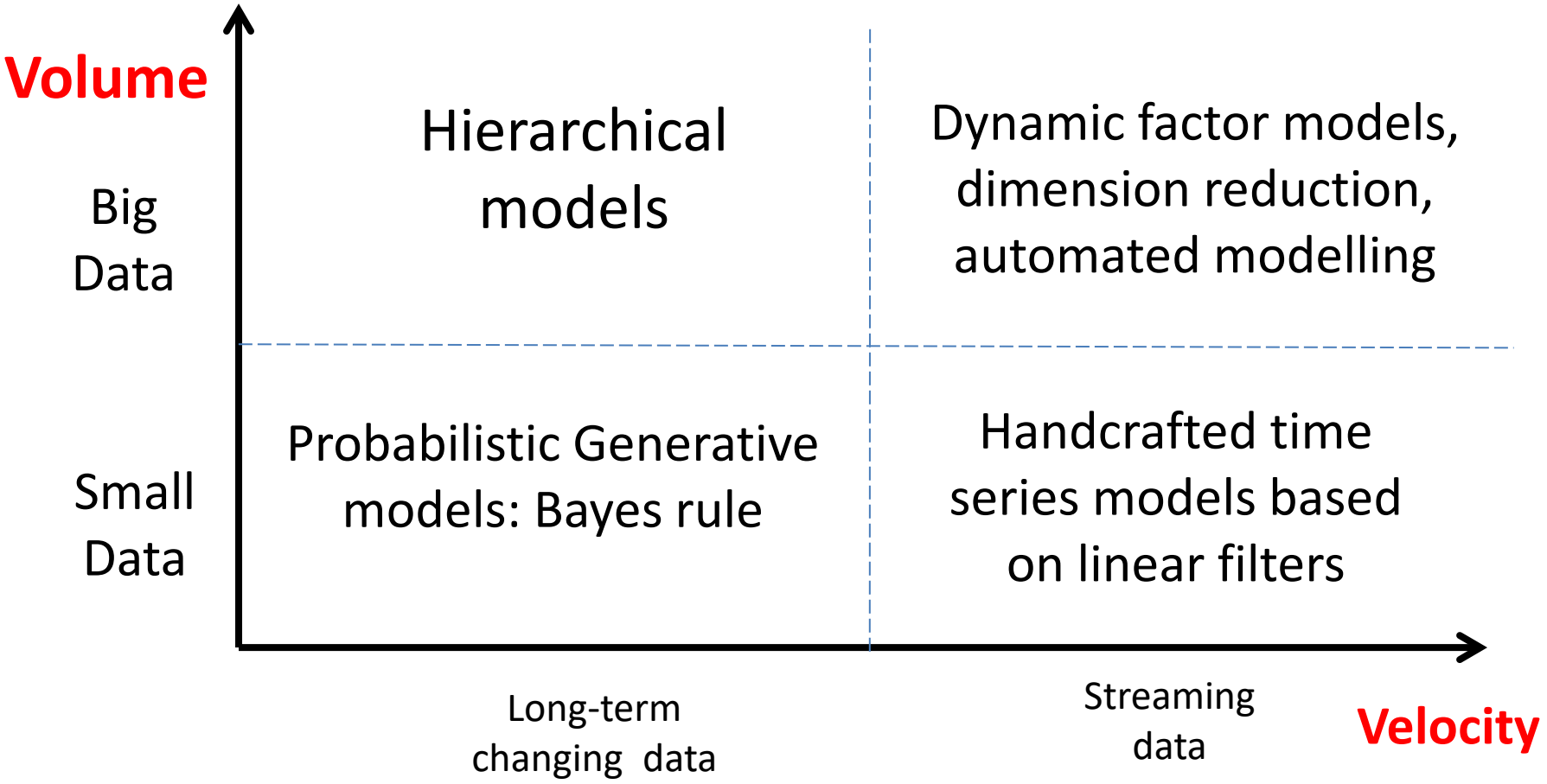
Hierarchical models

Probabilistic Generative models: **Bayes rule**

Long-term changing data

Velocity

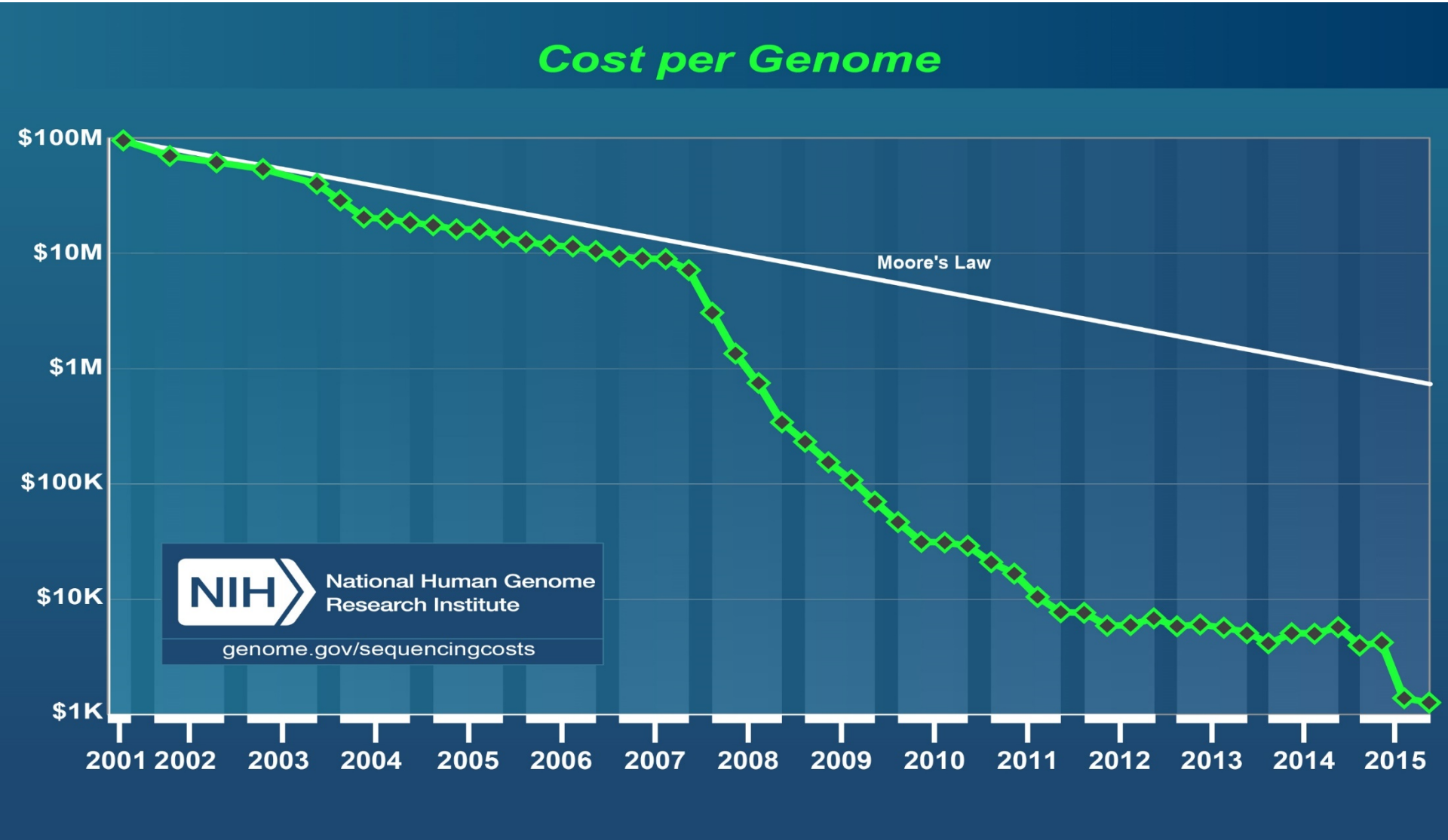
… in Machine Learning Techniques



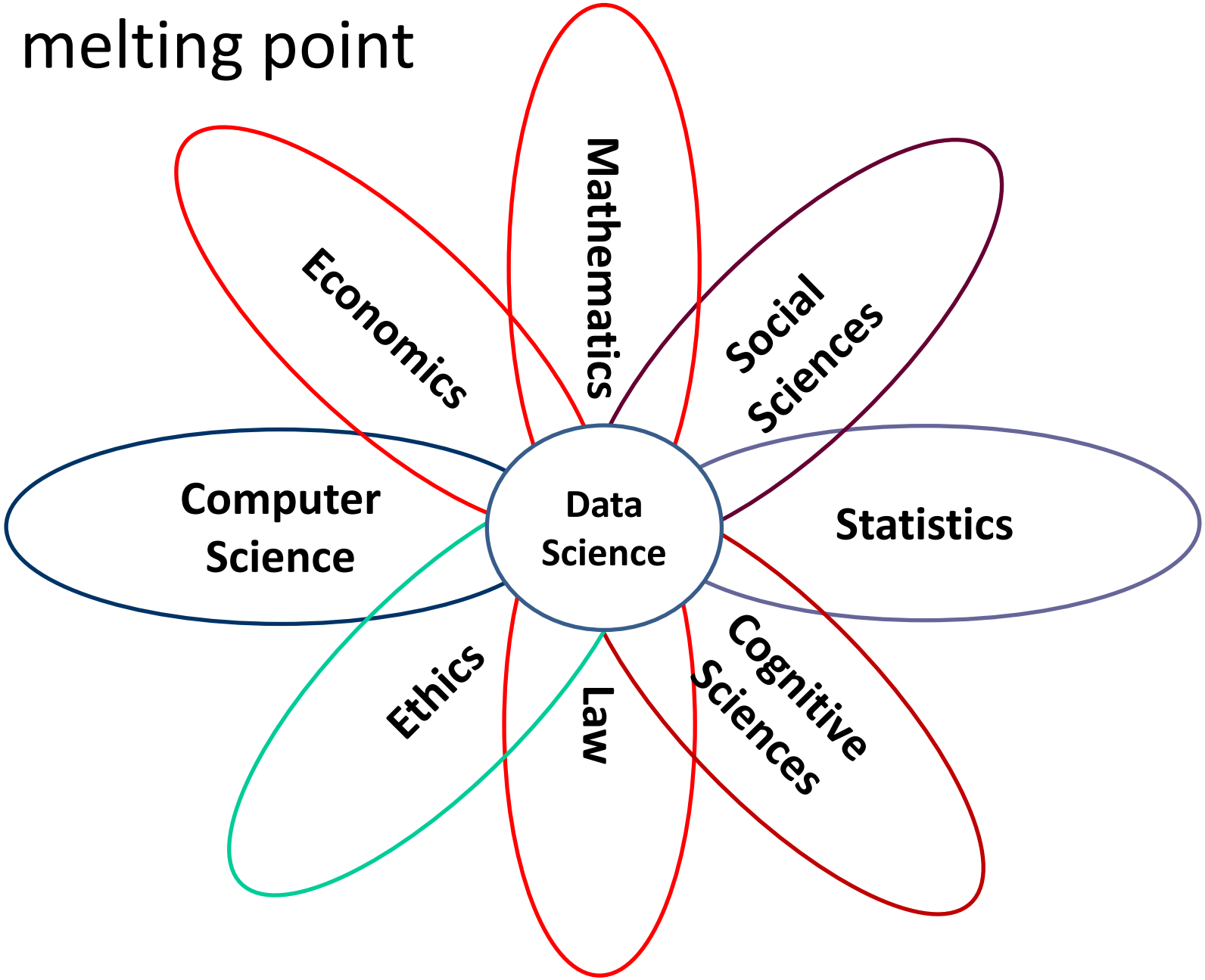
From S. Ceri, EDBT Venice, March 2017

How big is the genome?
As a string: 700MByte
As raw data: 200 Gbyte
As called mutations: 125MByte

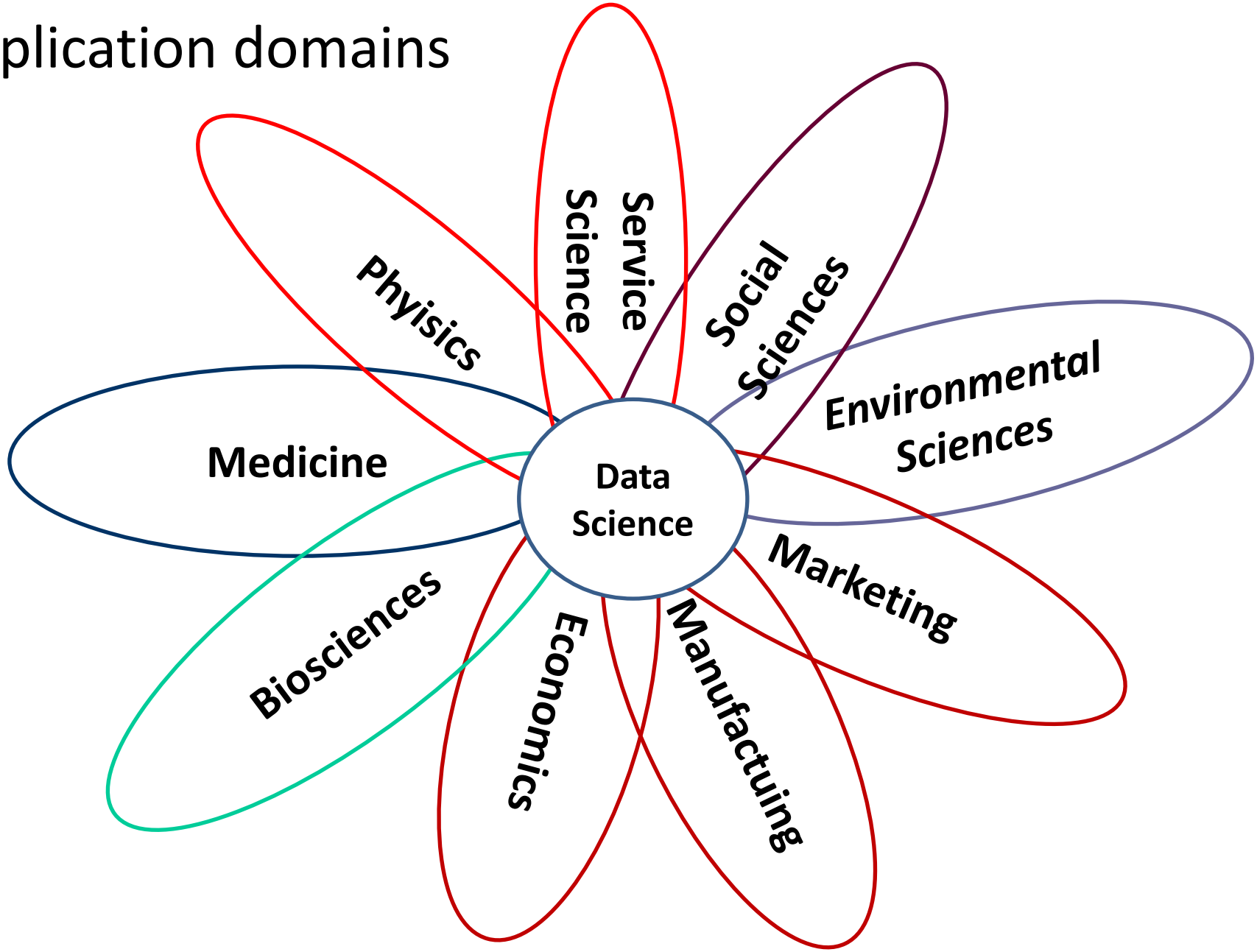
How many genomes will be sequenced in 5 years?
Estimates: order of 5-20 Millions
Very big data problem



Data Science as a melting point



Data Science application domains



Many good news (from Abiteboul, EDBT Conference, Venice, March 2017)

- **Improve** people's lives, e.g. humanitarian services
- **Accelerate** scientific discovery, e.g. personalized medicine
- **Boost** innovation, e.g. autonomous cars
- **Transfom** society, e.g. open government
- **Optimize** business, e.g. advertisement targeting

Big Concerns or:
Big Controversial Issues about Big Data
A very crowded Agenda

Fil rouge

- 1st Kranzberg Law: Technology is neither good nor bad; nor is it neutral.
- Tom Atlee statement “I’ve come to believe that things are getting better and better and worse and worse, faster and faster, simultaneously”.

1. Economic Value vs Social Utility

Social value - Quality of health care in Uganda The Economist 2011

The Economist

World politics | Business & finance | Economics | Science & technology | Culture | Blogs | Debate & discuss | Multimedia | Print edition

All audio
All video
The Economist audio edition

The Open Government Partnership

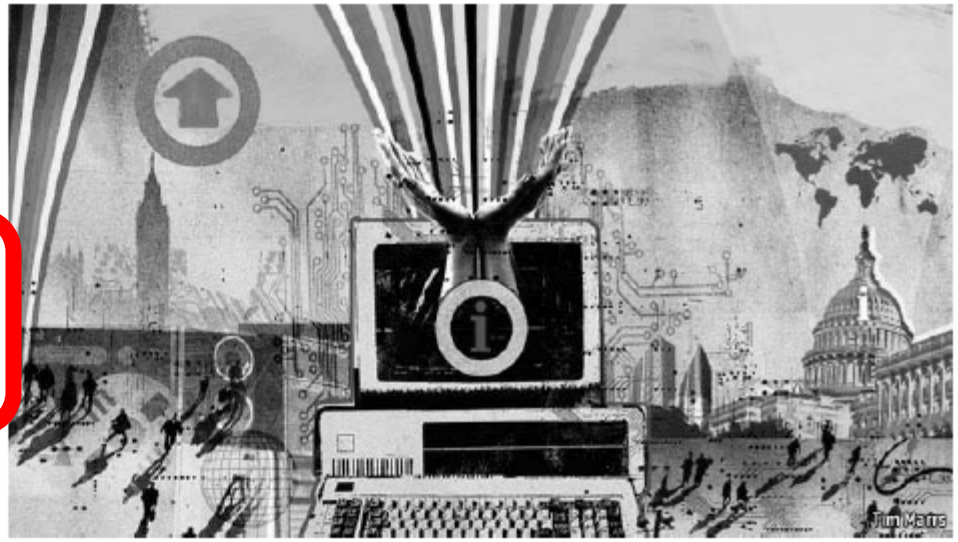
The parting of the red tape

Is it just another global talking-shop—or a fresh approach to shaking out government secrecy?

Oct 8th 2011 | NEW YORK AND TALLINN | from the print edition

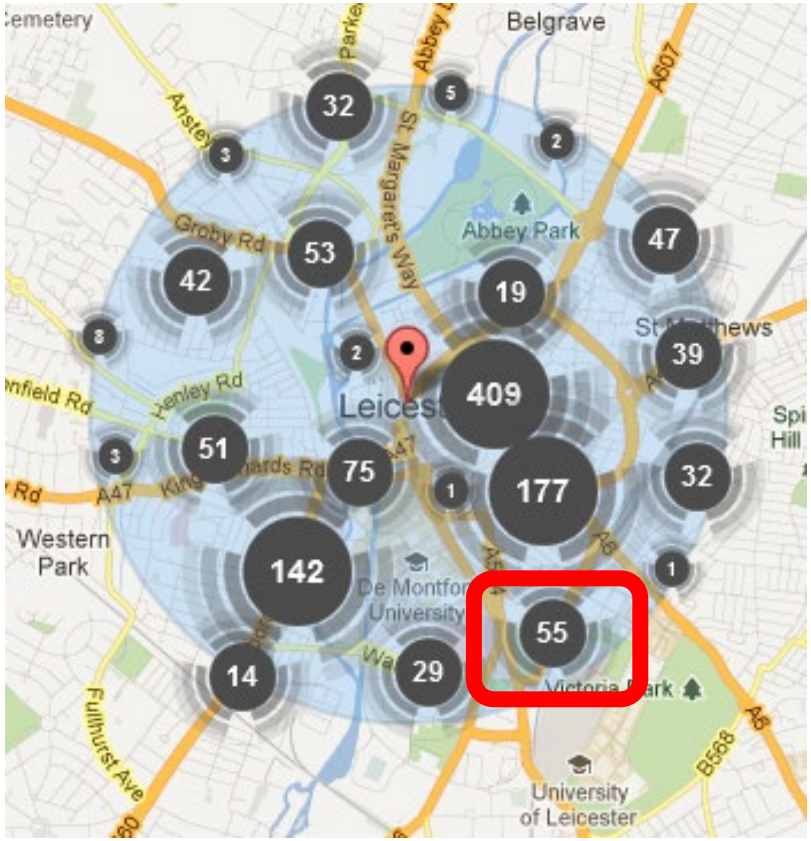
Like 151 0

UGANDA is not best known as a testbed for new ideas in governance. But research there by Jakob Svensson at the University of Stockholm and colleagues suggested that giving people health-care performance data and helping them organise to submit complaints cut the death rate in under-fives by a third. Publishing data on school budgets reduced the misuse of funds and increased enrolment.



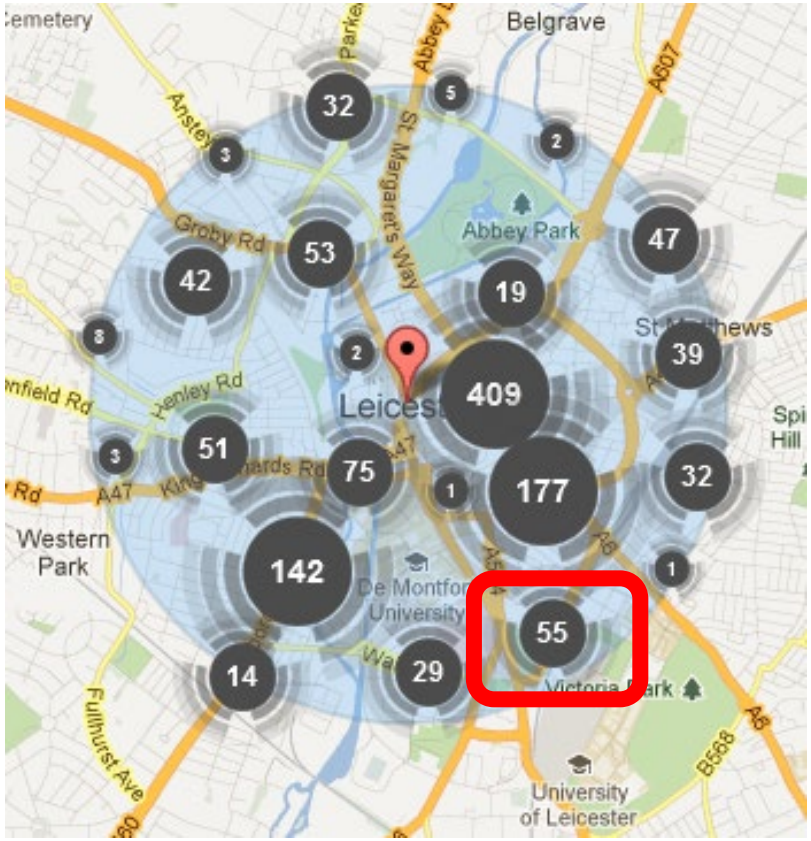
Crimes at Leicester, **positive** value for me…

All crime	1241
Burglary	75
Anti-social behaviour	317
Robbery	12
Vehicle crime	57
Violent crime	181
Public disorder and weapons	45
Shoplifting	172
Criminal damage and arson	102
Other theft	178
Drugs	48
Other crime	54



…and **negative** value for house landlords

All crime	1241
Burglary	75
Anti-social behaviour	317
Robbery	12
Vehicle crime	57
Violent crime	181
Public disorder and weapons	45
Shoplifting	172
Criminal damage and arson	102
Other theft	178
Drugs	48
Other crime	54



What the Leicester example shows

Data can provide the user a **social value** or else an **economic utility**, resulting in a well known tension in the history of human mankind.

2. Numeration, Digitalization, Datafication

Si può ridurre tutto a numero?

A Comparative Analysis of Methodologies for Database Schema Integration

C. BATINI and M. LENZERINI

Dipartimento di Informatica e Sistemistica, University of Rome, Rome, Italy

S. B. NAVATHE

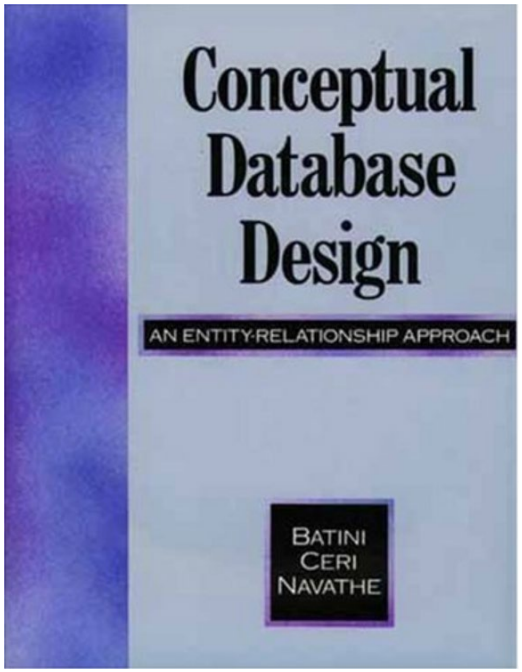
Database Systems Research and Development Center, Computer and Information Sciences Department, University of Florida, Gainesville, Florida 32601

One of the fundamental principles of the database approach is that a database allows a nonredundant, unified representation of all data managed in an organization. This is achieved only when methodologies are available to support integration across organizational and application boundaries.

Methodologies for database design usually perform the design activity by separately producing several schemas, representing parts of the application, which are subsequently merged. Database schema integration is the activity of integrating the schemas of existing or proposed databases into a global, unified schema.

The aim of the paper is to provide first a unifying framework for the problem of schema integration, then a comparative review of the work done thus far in this area.

Titolo 1–20	Citata da
A comparative analysis of methodologies for database schema integration C Batini, M Lenzerini, SB Navathe ACM computing surveys (CSUR) 18 (4), 323-364	2406
Entity Relationship Approach C Batini, S Ceri, S Navathe Elsevier Science Publishers BV (North Holland)	1689
Data-Centric Systems and Applications MJ Carey, S Ceri, P Bernstein, U Dayal, C Faloutsos, JC Freytag, ... Springer, Verlag Berlin Heidelberg [doi: 10.1007/978-3-540-76452-6]	880



Datafication: quanto piu' i dati sono comprensibili per noi, tanto piu' e faticoso renderli elaborabili…

Image

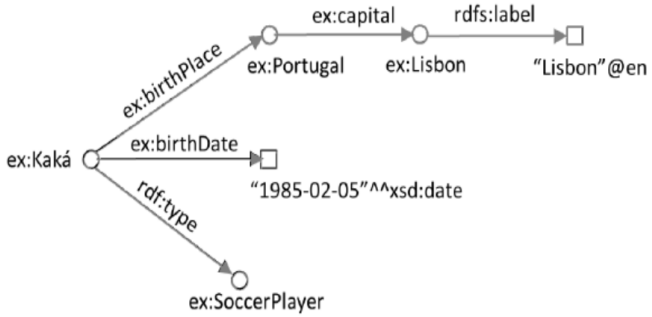
Text

Dear Laure, I try to describe the wonder-ful harbour of Portofino as I have seen this morning a boat is going in, other boats are along the wharf. Small pretty buildings and villas are looking on to the harbour.

Structured data

Place	Country	Population	Main economic activity
Portofino	Italy	700.000	Tourism

Linked data



2. Numeration, Digitization, Datafication

La grande disponibilità di

- **strumenti di acquisizione** permette di:
  - **Misurare** i fenomeni ed eventi della realtà, associando ad essi delle quantificazioni (**Numeration**)
- **fonti di informazioni** permette di:
  - **Modellare** la realtà per mezzo di rappresentazioni digitali (**Digitization**)
  - **Estrarre** da esse sintassi e/o significato, trasformandole in dati (**Datafication**)

2. Numeration, Digitization, Datafication → Modeling

Quando descriviamo la realtà per mezzo di **numeri** o **dati**, essi diventano **modelli**, che **sostituiscono la realtà** nelle **attività e decisioni** delle organizzazioni ed umane, anche esse modellate da **algoritmi.**

Parafrasando la prima legge di Kransberg:

- Il **modello** non è mai né buono, né cattivo, né neutrale.

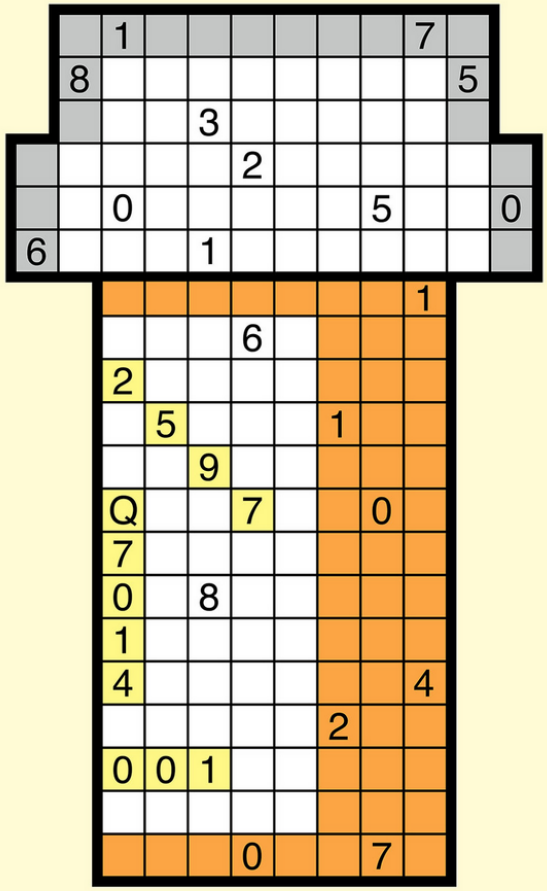
Dal New York Times

The New York Times Magazine Share 36

Those Indecipherable Medical Bills? They're One Reason Health Care Costs So Much.

Hospitals have learned to manipulate medical codes — often resulting in mind-boggling bills.

BY ELISABETH ROSENTHAL MARCH 29, 2017



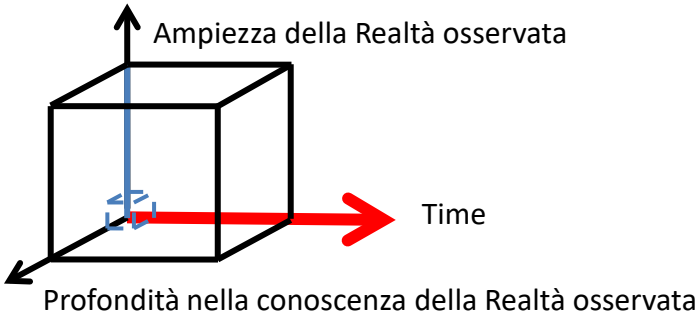
3. From Why to What

Chris Anderson - 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete ', 2008

- 'This is a world where **massive amounts of data and applied mathematics** replace every other tool that might be brought to bear. Out the door with every theory of human behaviour, from linguistics to sociology.
- Forget taxonomy, ontology, and psychology. Who knows **why people do what they do**? The point is **they do it**, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.'

Example: when to buy a flight ticket – from **causality** …

We can investigate to find on a sample the law for pricing applied by airline companies (**Why**)



… to **correlation**

Oren Etzioni's Farecast (**What**)

Sample of 12.000 tickets →

$$200\ 10^{9}$$

50 $ average savings per ticket

the start-up Farecast sold for 110 10^{6} $

Predictive policing - 1

- In February 2014, the Chicago Police Department (CPD) made national headlines for **sending its officers to make personal visits to residents** considered most likely to be involved in a violent crime.
- The selected individuals were not necessarily under investigation, but had histories that implied that they were among the city's residents most likely to be either a victim or perpetrator of violence.

Predictive policing - 2

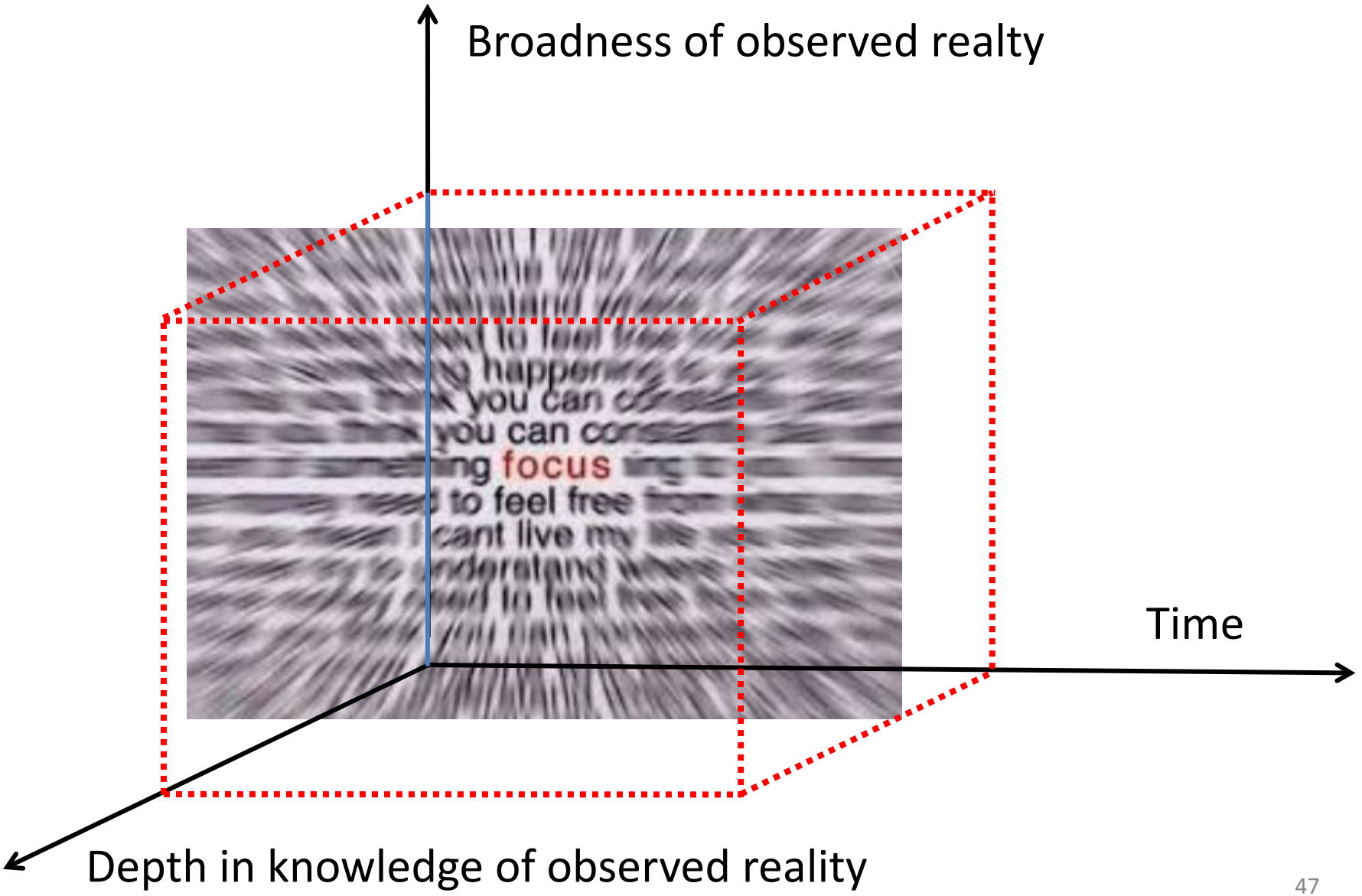
- The officers' visits were guided in part by a computer-generated "Heat List": **the result of an algorithm** that attempts to predict involvement in violent crime.
- City officials have described some of the inputs used in this calculation—it includes some types of arrest records, for example—but **there is no public, comprehensive description of the algorithm's input.**

Concerns

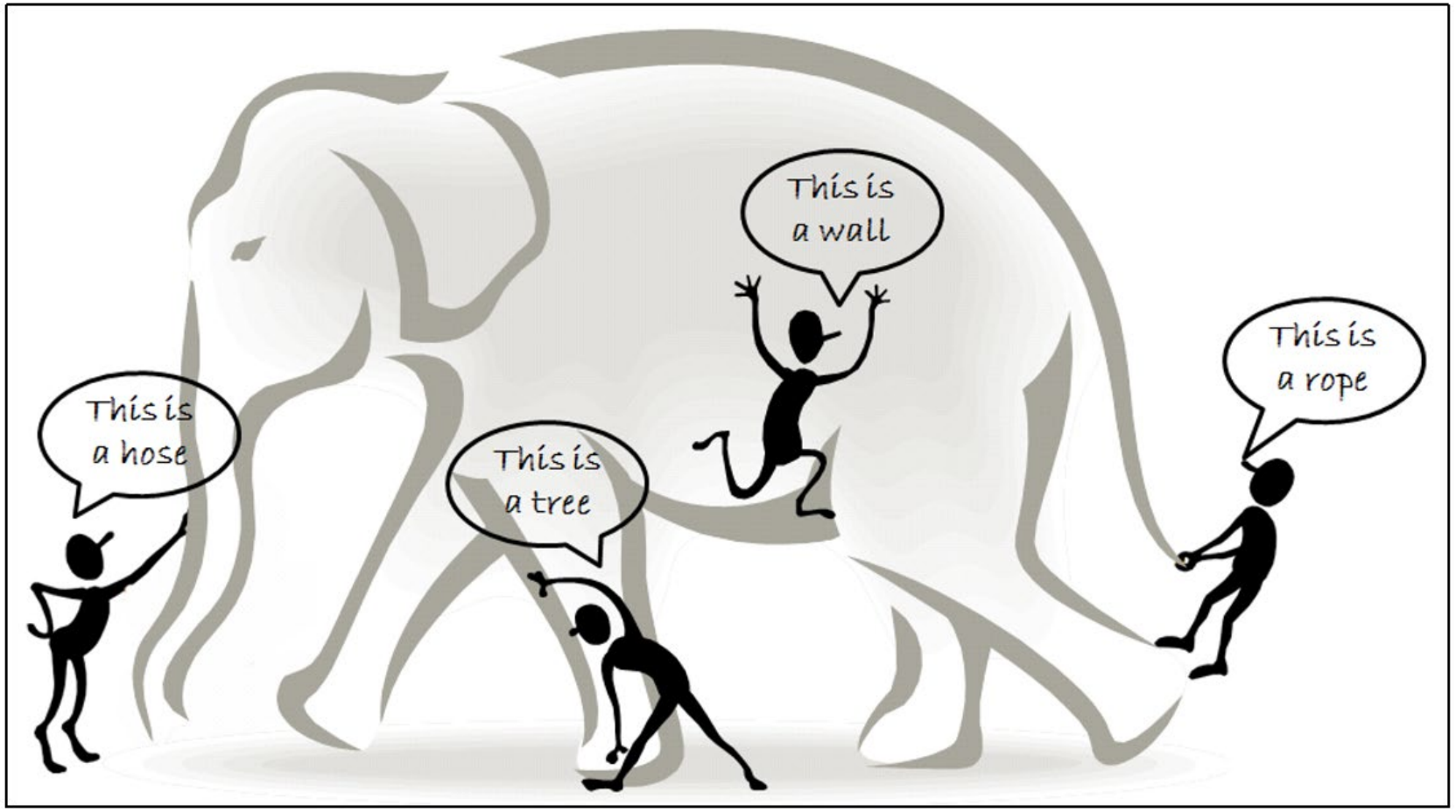
- The **what** is influenced by the **model**
- Dealing only with **what** and not with **why,** leads to a risk of «decision objectification», without no analysis of causes of phenomena,
- A new more sophisticated version of «it is the computer, stupid!»

4. Inexactitude & blurriness & messiness

A blurred reality….



…. fragmented



There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy…

..and polluted

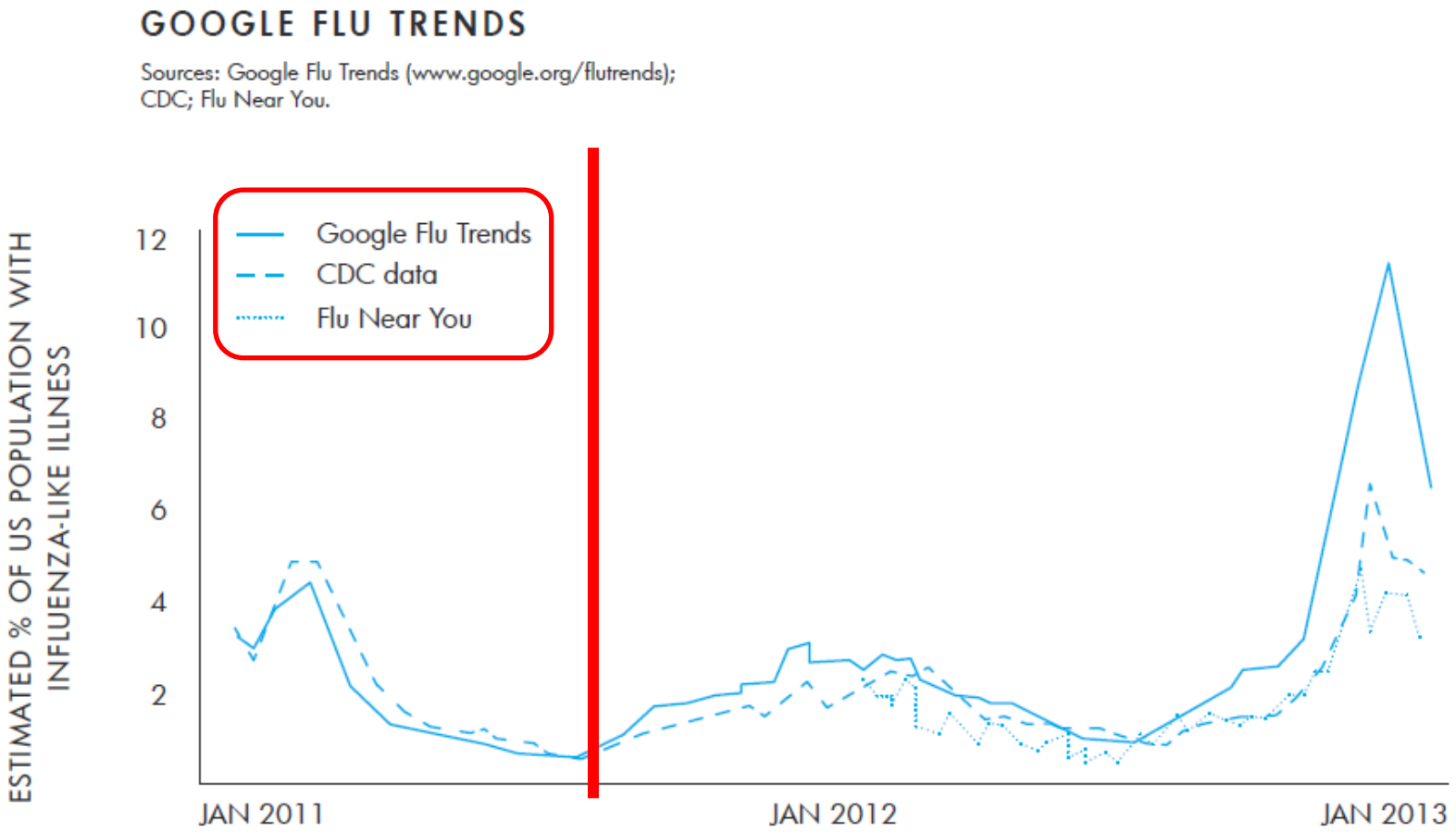
© Navesink Consulting Group LLC, 2000-2005

Come possiamo contrastare la inexactitude/messiness?

- **Knowledge solution** → Aumentare la conoscenza formale sul fenomeno (costoso)
- **Crowd solution** → es. Wikipedia
- **Social Solution** → es. Open Street Map
- **Ecological solution** → Cambiare il modo con cui produciamo e usiamo i dati

5. Big Data Hubrys

Google Flu Trends

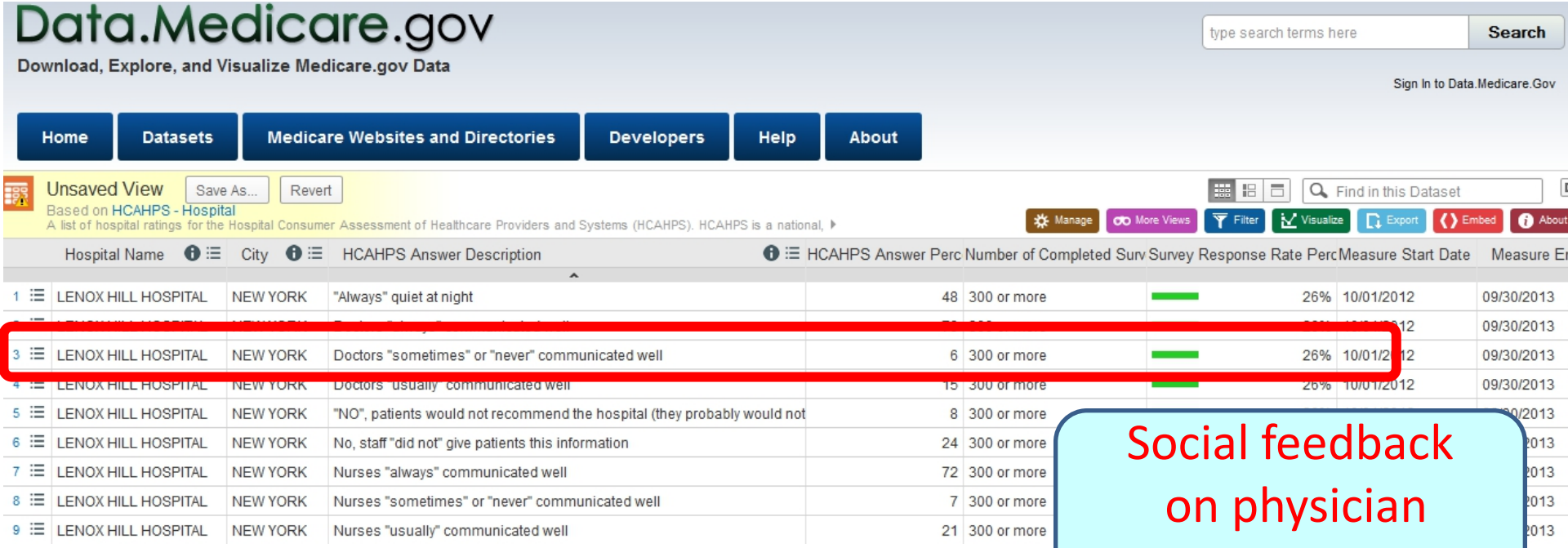


Hubrys: the arrogance of data

Big data evangelists often make the implicit assumption that big data are a **substitute** for, rather than a **supplement** to, traditional data collection and analysis.

6. Transparency, privacy and determinism

Example from USA: Consumer assessment about their experiences during an inpatient hospital stay



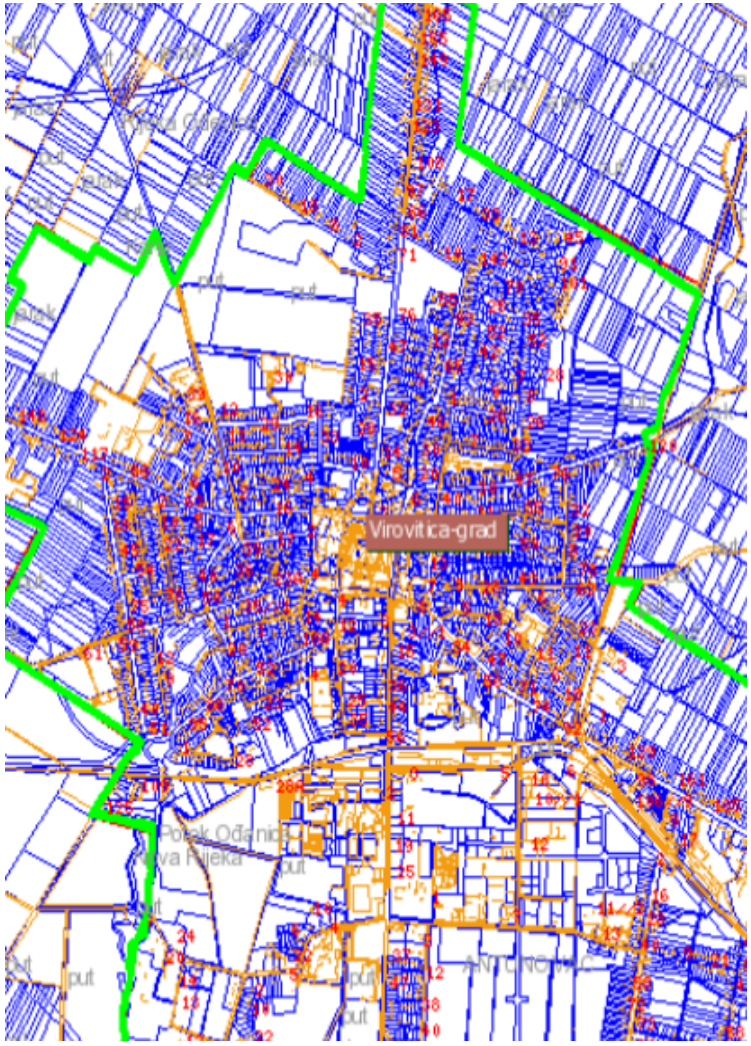
Social feedback on physician quality

Source:
https://data.medicare.gov/Hospital-Compare/HCAHPS-National/99ue-w85f

Legenda: HCAHPS - Hospital
A list of hospital ratings for the Hospital Consumer Assessment of Healthcare Providers and Systems HCAHPS is a national, standardized survey of hospital patients about their experiences during a recent inpatient hospital stay.

Filter: LENOX HILL HOSPITAL – NEW YORK

Cadastral data in India



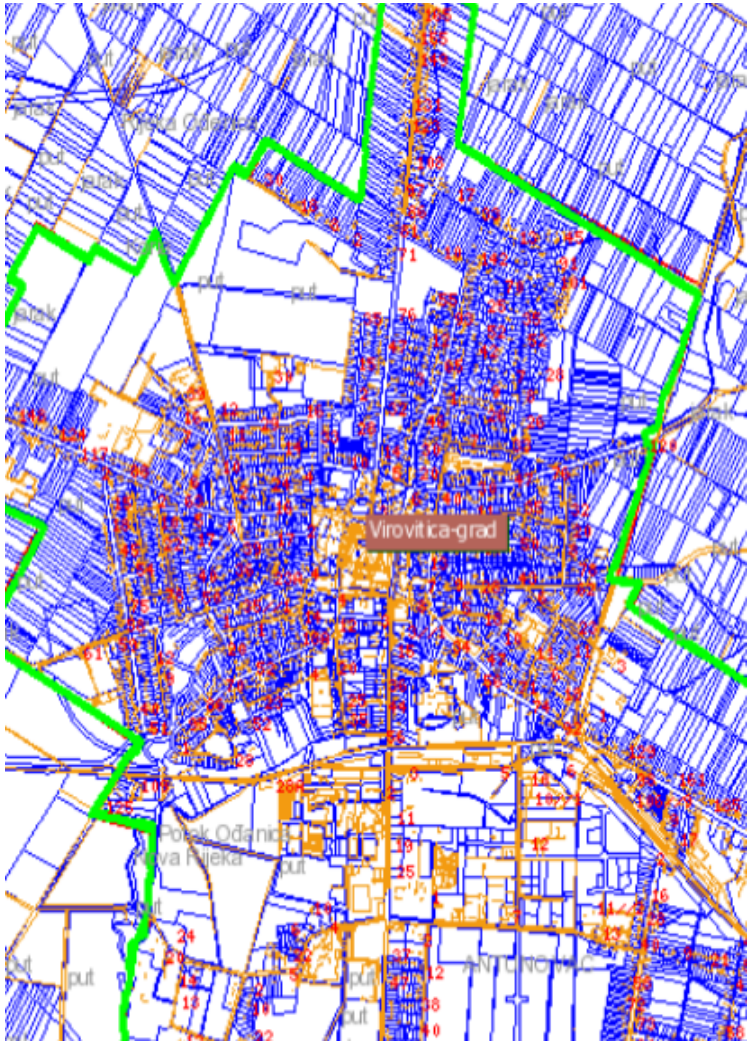
Goals of digitization of land data

Empower citizens against

- state bureaucracies and
- corrupt officials

through **transparency** and **accountability**.

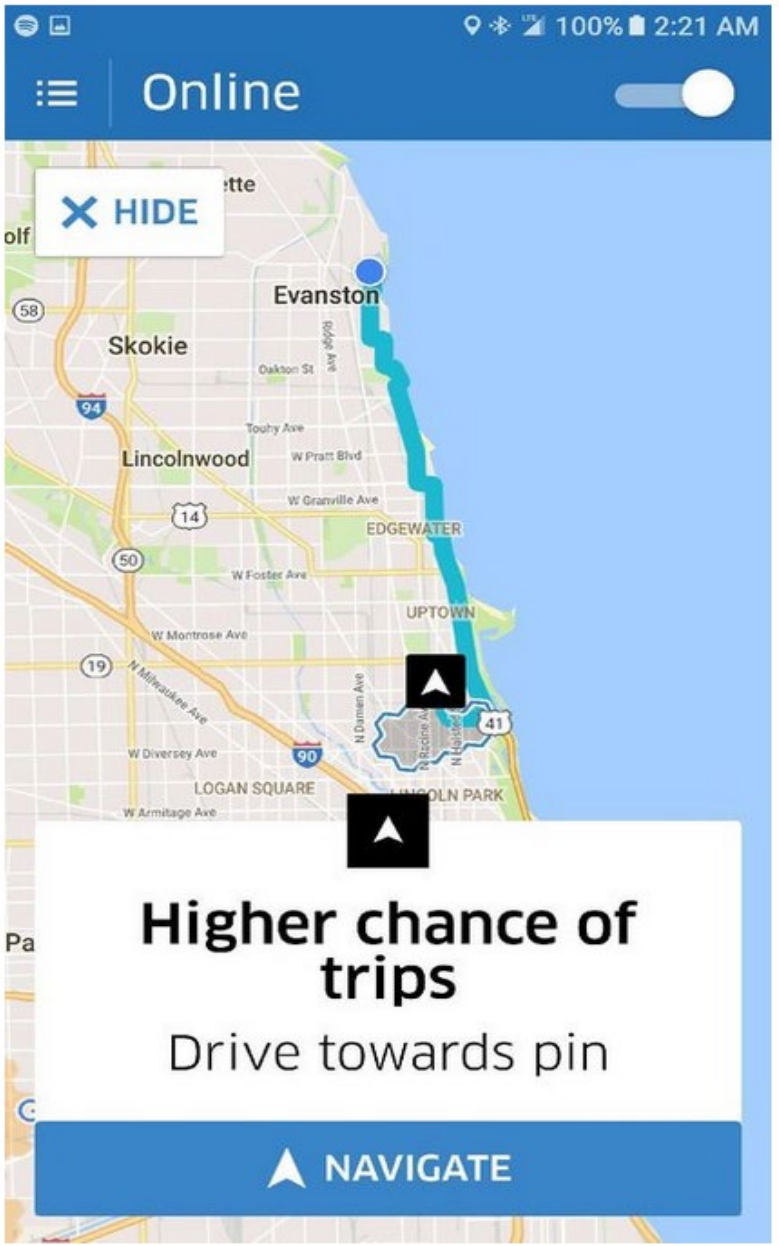
Final outcome: the opposite than hoped



Determinism

"It was all day long, every day — texts, emails, pop-ups: 'Hey, the morning rush has started. Get to this area, that's where demand is biggest,'" said Ed Frantzen, a veteran Uber driver in the Chicago area.

"It was always, constantly, trying to get you into a certain direction."



7. Big Data Divide

Statistics 2.0: from the Data Revolution to the next level of Official Statistics

Enrico Giovannini

The data revolution is very unequally distributed between countries and people

- People, organisations and governments are **excluded because of lack of resources, knowledge, capacity or opportunity**.
- There are huge and growing inequalities in **access** to data and information and in the **ability to use** it.

There are too many gaps in current data, making some people and some issues almost invisible:

- Too many countries **still** have **poor data for MDGs**
- Data arrives **too late**
- Too many issues are still **barely covered** by existing data
- Entire groups of **people, regions** and **key issues** remain **invisible**

Lots of big data divides

- Countries that have access to/can measure big data and countries that have not, or have limited → Example: poverty index
- Research groups that can buy big data and groups that can't.
- "Sorters", those who are able to extract and use findings and "sortees", those who have their lives affected by the resulting decisions → asymmetric findings (new version of asymmetric information, investigated in economics)

Big data divide and biases in models

- OpenStreetMap (OSM) is a successful crowdsourced mapping project: many cities of the world have been mapped by people on a voluntary basis.
- However, some regions get mapped quicker than others, such as tourist locations, while locations of less interest (such as poorer neighborhoods) receive less attention.

Humanitarian open street map initiative

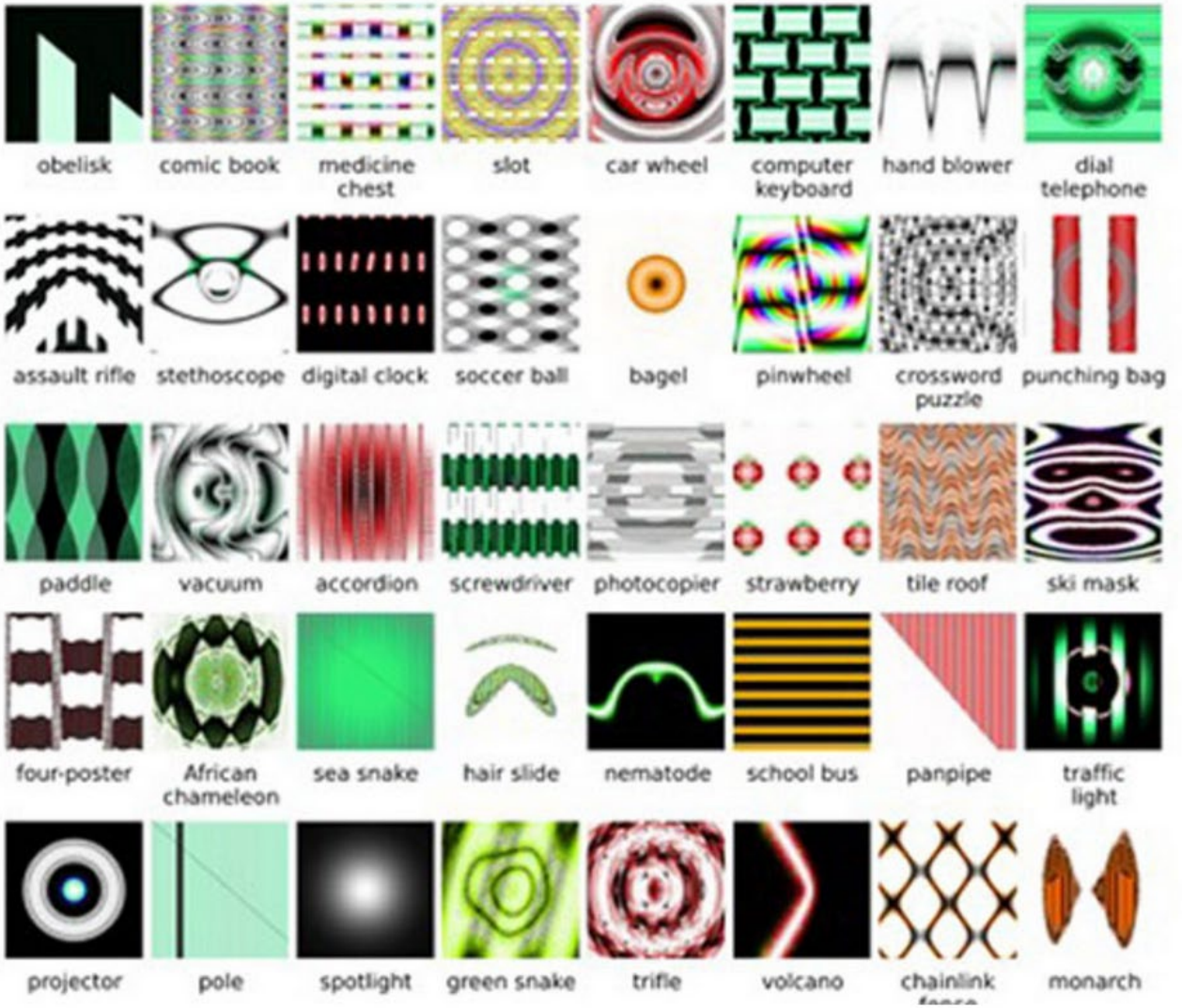
Many of the poorest and most vulnerable places in the world do not exist on any map. To date over 3,500 Missing Maps volunteers have collectively made 12 million edits to OpenStreetMap and put 7.5 million people on the map.

Urban Innovations: Crowdsourcing Non-Camp Refugee Data

The "Crowdsourcing Non-Camp Refugee Data Through OpenStreetMap" project aims to improve program planning and service delivery to refugee communities, develop better integration with host communities, and build refugee self-reliance through open map data. The program trains and equips community leaders in refugee communities to map vulnerabilities and assets in the places they live, filling in key data gaps and "blank

8. Apophenia: the human tendency to perceive meaningful patterns within random data

Apophenia in machine learning



Source: Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," cv-foundation.org, 2015 →.

9. Overload and Abstraction

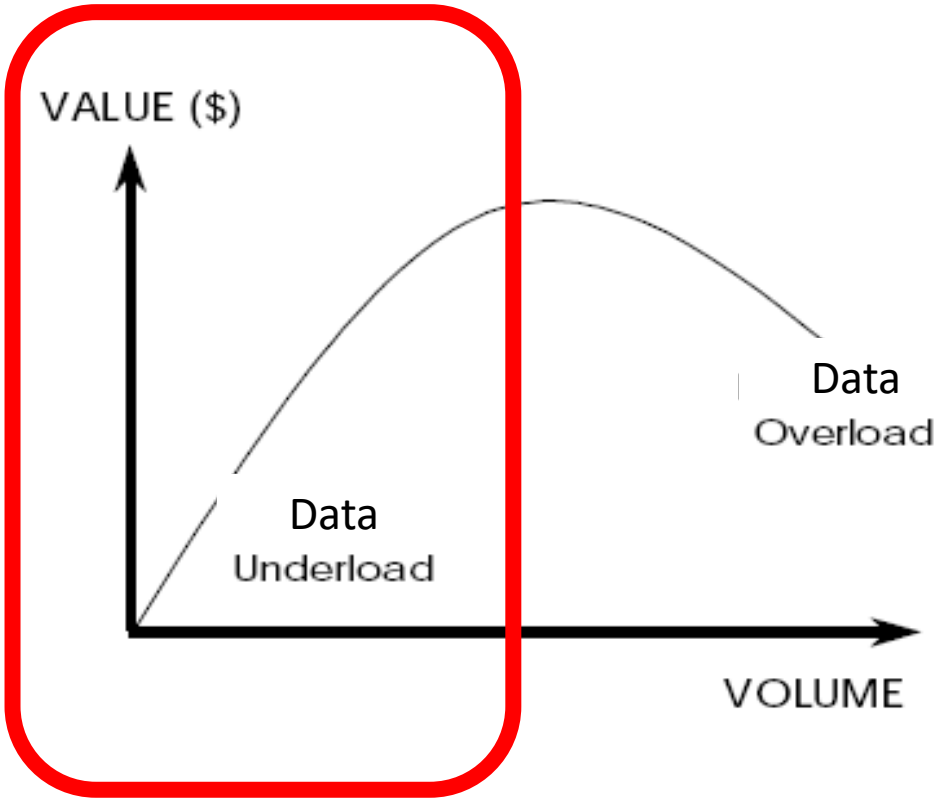
Overload & Abstraction
or «too big to know»

La psicologia cognitiva e alcuni esempi che abbiamo fatto dimostrano che il valore cognitivo dei dati cresce con la loro disponibilità. Ma…. →

Moody - 1

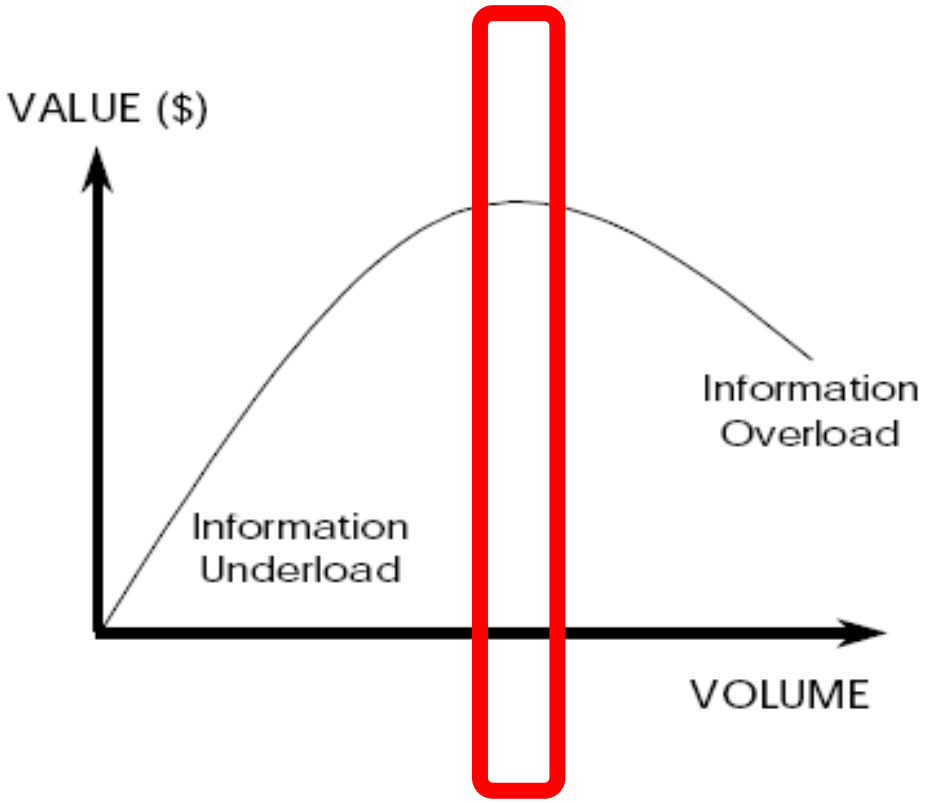
La figura (da Moody 1999) mostra in forma qualitativa come evolve il valore conoscitivo all'aumentare dei dati disponibili.

All'inizio più dati corrispondono a piu' valore.



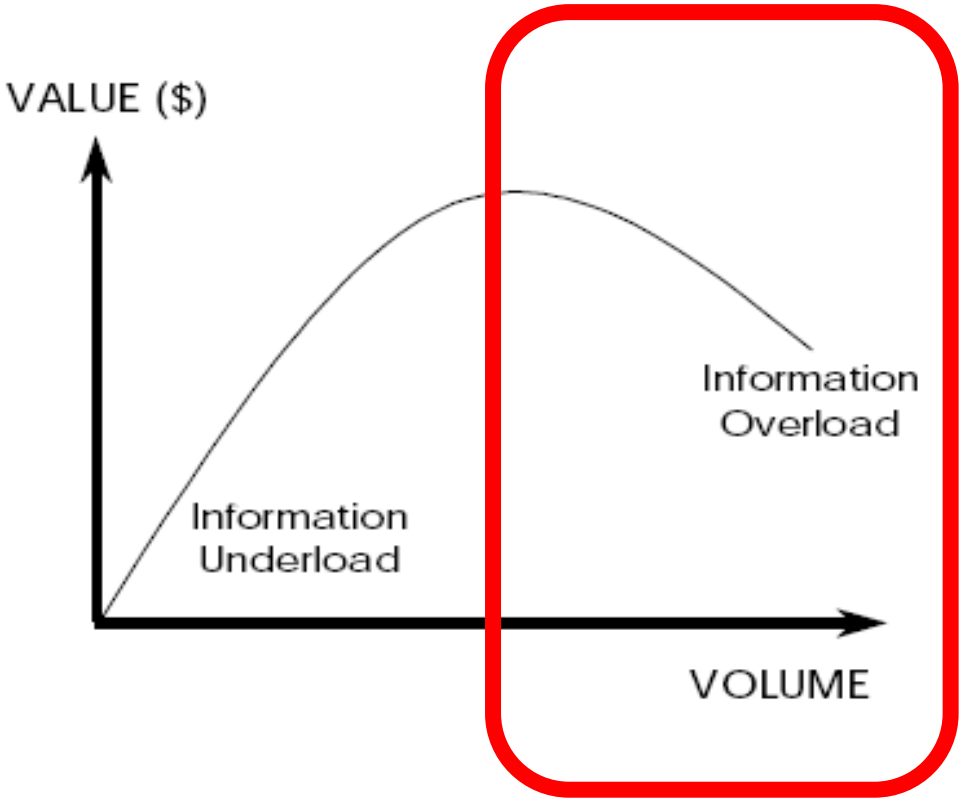
Moody - 2

Ma da un certo punto in poi i nuovi dati a noi disponibili sono così tanti che non riusciamo cognitivamente a considerarli insieme agli altri per produrre nuova conoscenza (questo e' il punto di massimo valore).

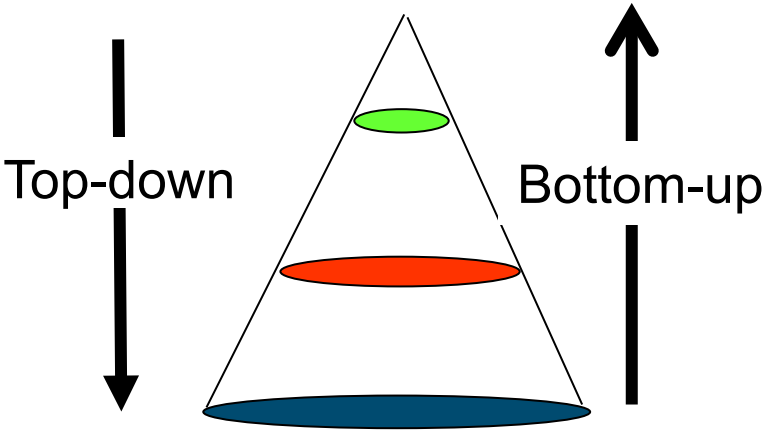


Moody - 3

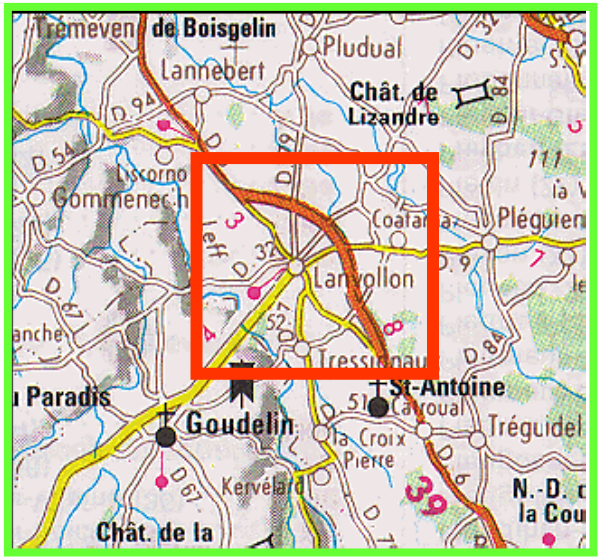
Da questo momento in poi, i nuovi dati non riescono a produrre nuova conoscenza, e provocano un fenomeno di “blocco” ed una sorta di regressione nella conoscenza accumulata.



Quando siamo sommersi, abbiamo bisogno di astrazioni →

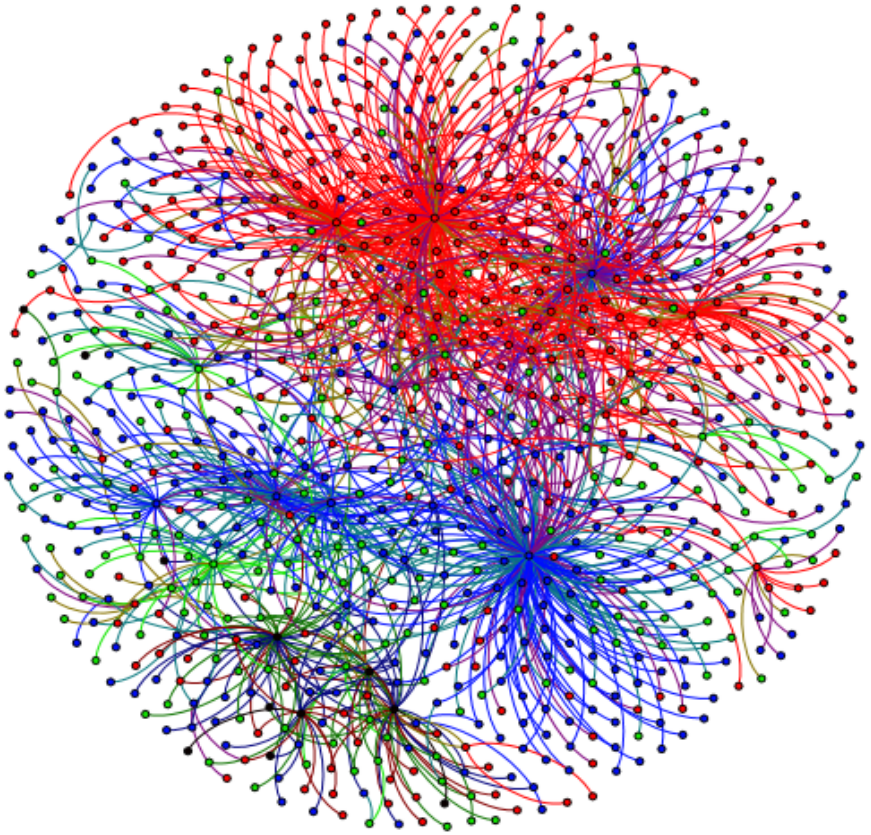


Muoversi tra diversi livelli di astrazione, scegliendo sempre quello «giusto»



10. Rage amplifier

Anger is more popular than joy…



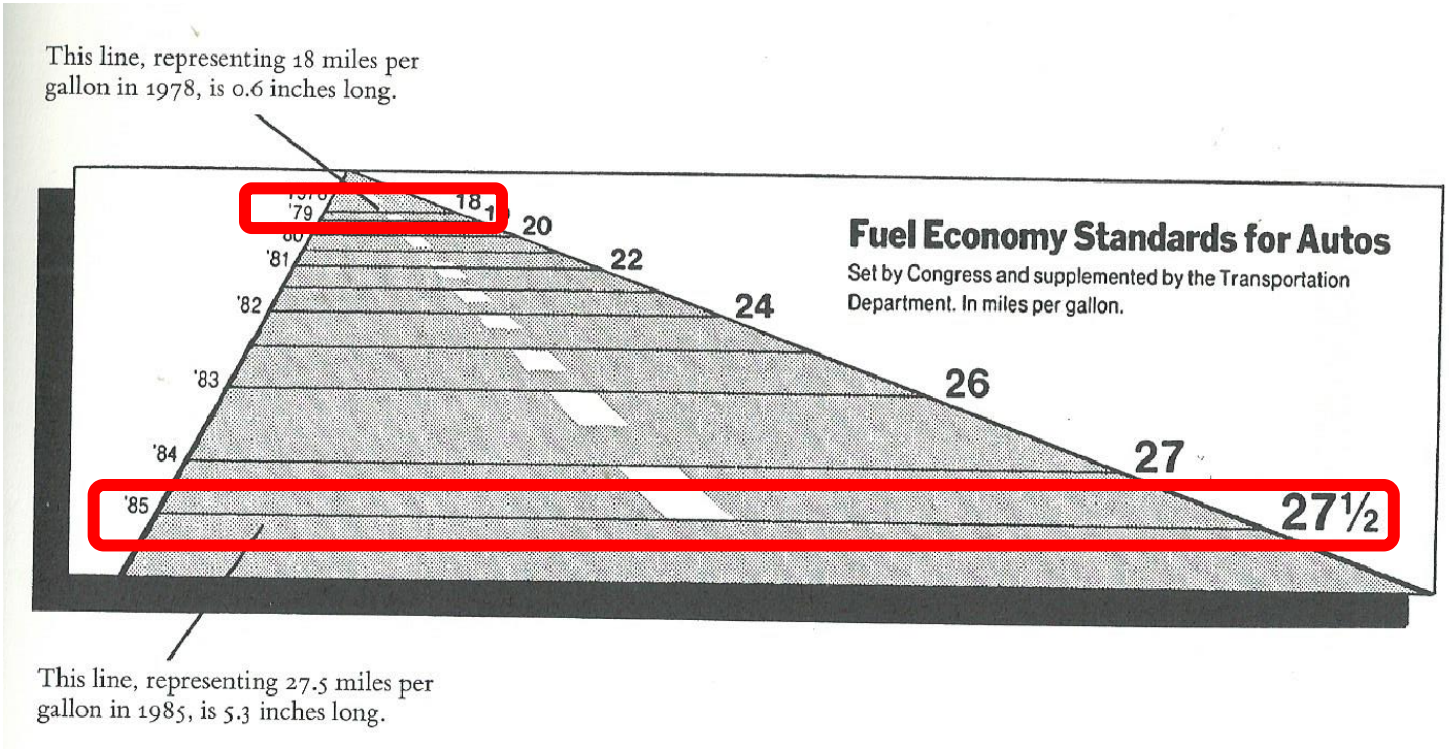
- red stands for anger,
- **green** represents joy,
- **blue** stands for sadness
- **black** represents disgust.

The regions of same color indicate that closely connected nodes share the same sentiment.

11. Visualization and lies

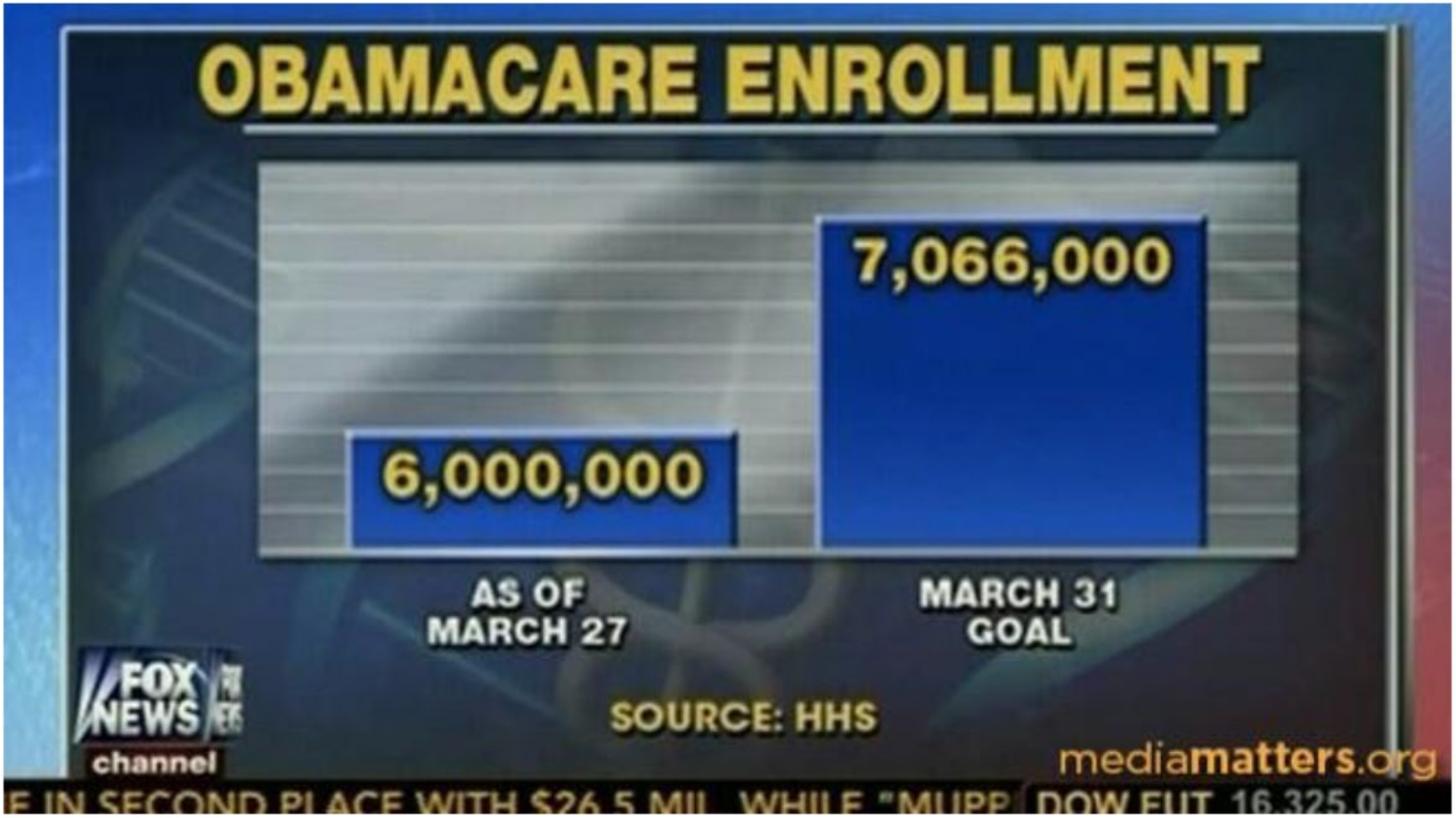
A picture is worth a thousand words, but…how many lies in Visualizations!

Year	Miles per gallon
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5



Lie factor = relative difference of size in the real world/ relative difference of size in the visualization = 14.8

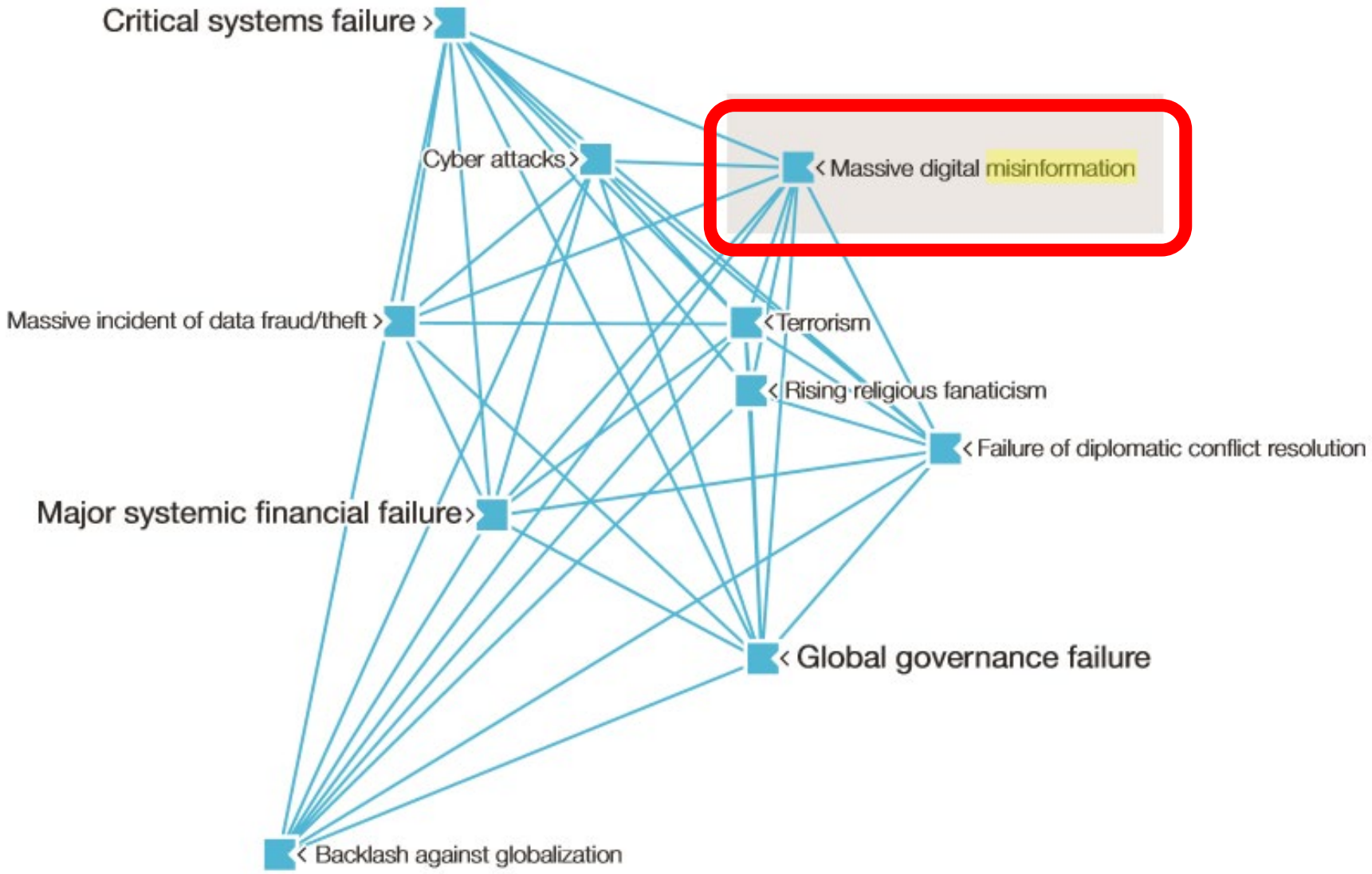
On Obamacare deadline day, this chart from Fox News is being passed around the Twittersphere - The chart appears to scale 6 million to about one-third of the Obama administration's original goal health-insurance exchanges — 7.066 million.



Twitter/@JustinWolfers

12. From fake news and post truth to Trump staff's «alternative facts»

World Economic Forum 2013



Form "Data for Policy: a Myth or a Must?"

Enrico Giovannini - University of Rome "Tor Vergata"

The Age of Post-Truth Politics

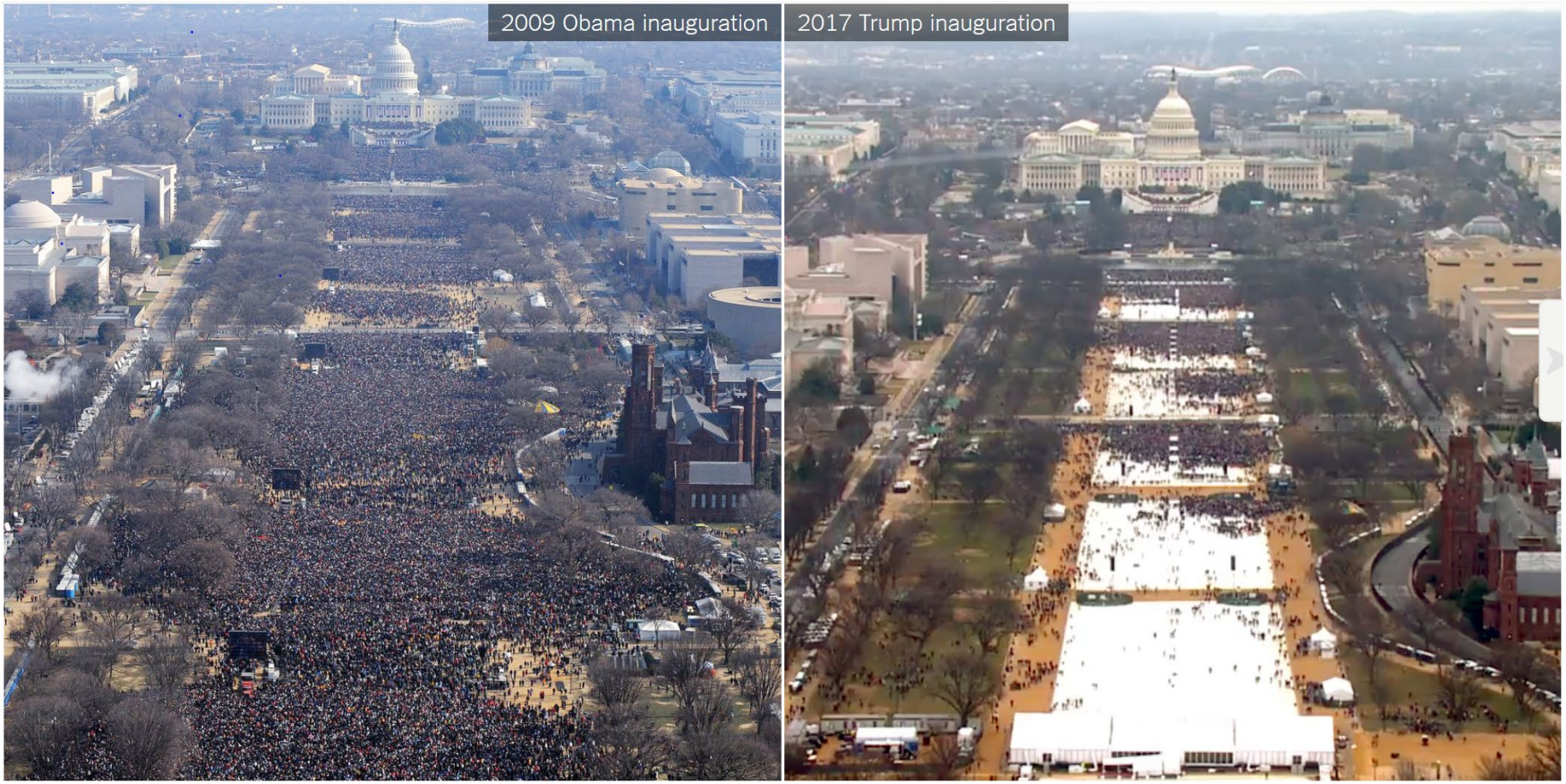
(NYT, William Davies, August 2016)

- "How can we still be speaking of "facts" when they no longer provide us with a reality that we all agree on.

- If you really want to find an expert willing to endorse a fact, and have sufficient money or political clout behind you, you probably can.

- It is possible to live in a world of data but no facts."

Trump staff's «alternative facts»



Alternative facts

Kellyanne Conway denies Trump press secretary lied: 'He offered alternative facts'

Hints from cognitive psycology

It's not just what people think that matters, but how they think. Refuting misinformation involves dealing with complex cognitive processes

A simple myth is more cognitively attractive than an over-complicated correction

For those who are strongly fixed in their views, encountering counter arguments, can cause them to strengthen their views

Fact checking:
facts are stubborn….

- According to figures shared by the Metro Washington subway system on Twitter, **193,000 trips** had been taken by 11am on Donald Trump's inauguration day, compared with **513,000** during the same period on 20 January 2009 when Barack Obama took office.
- But fact checking has a cost….

Formazione e collaborazione nel fact checking Milano, 2 Aprile 2017

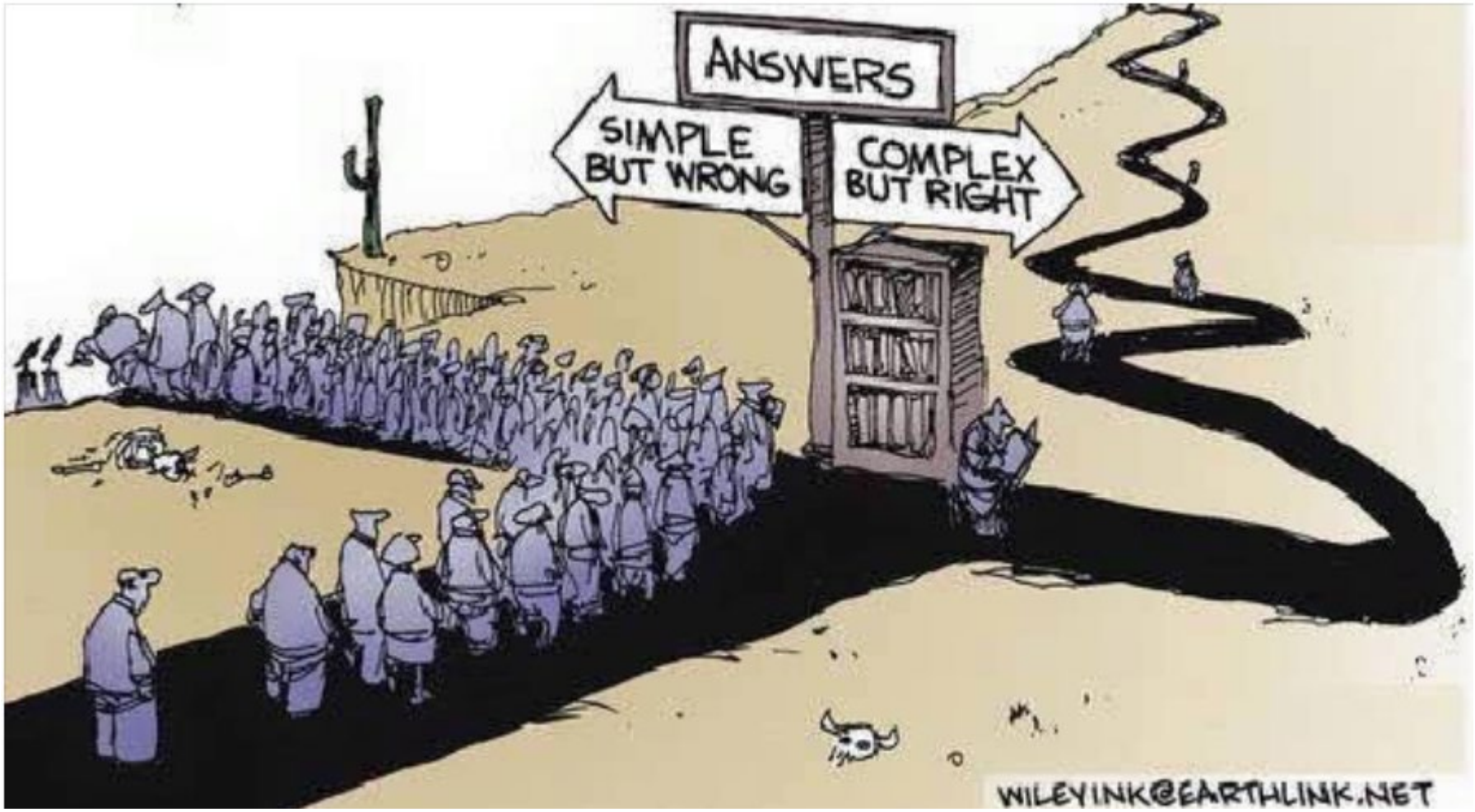
Arriva il «fact checking day» Una giornata per imparare a riconoscere le bufale

di Eva Perasso

Eventi in tutto il mondo domenica 2 aprile. In Italia workshop per studenti a Milano e molto materiale online per informarsi in maniera consapevole

So, do we have solutions to such concerns?

No simple answers to complex questions



Coming back to…

- 1st Kranzberg Law that says: Technology is neither good nor bad; nor is it neutral
- Tom Atlee statement “I’ve come to believe that things are getting better and better and worse and worse, faster and faster, simultaneously”.

Everything is up to us, either as individuals or as communties. But what ever we conceive, we have to make fast….

Second (long term) answer: from Numeracy and Literacy…

Two well known indicators of the level of culture of a population or community are numeracy and literacy.

- **Numeracy** is the ability to reason and to apply simple numerical concepts
- **Literacy** is traditionally understood as the ability to read, write, and use arithmetic.

… to Datacy, that

(temptative draft definition) measures the capacity of

- **reasoning** on a vast amount of data types,
- **understanding** their meaning
- **Investigating** the economic, social and ethical impact
- **use languages and techniques** for their representation, management, analysis and visualization.

in such a way to become able to **solve** complex problems, **take** complex decisions, and **play** an active role in society.

Per informazioni sul Corso di Laurea
accedi a: datascience.disco.unimib.it
scrivi a: orientamento.datascience@disco.unimib.it

References

General on ICT and Information Society

International Telecommunication Union, Measuring the Information Society Report 2014, Swizerland.

Books

- Borgman C. – Big Data, Little data, no data, The MIT Press, 2015.
- Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013

Data Ethics

Serge Abiteboul, Julia Stoyanovich. Data, Responsibly. ACM Sigmod Blog, 20 November 2015. 2015.

Serge Abiteboul et al,. Managing your digital life, Communication of the ACM, Vol 58 N. 5.

Zwitter, Andrej. "Big data ethics." *Big Data & Society* 1.2 (2014).

BD & Analytics

Labrinidis, Alexandros, and Hosagrahar V. Jagadish. "Challenges and opportunities with big data." *Proceedings of the VLDB Endowment* 5.12 (2012): 2032-2033.

Wu, Xindong, et al. "Data mining with big data." *ieee transactions on knowledge and data engineering* 26.1 (2014): 97-107.

General on Data Science & BD

De Biase L. – Homo Pluralis: essere umani nella età tecnologica, 2016.

Snow J. - On the Mode of Communication of Cholera, London: John Churchill, New Burlington Street, England, 1855.

Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013

Nick Couldry - A necessary disenchantment: myth, agency and injustice in a digital world - The Sociological Review,Vol. 62, 880–897 (2014)

C. Hess and E. Ostrom - Understanding Knowledge as a Commons From Theory to Practice, The MIT Press, 2007.

R. Michael Alvarez, ed., In press, Computational Social Science: Discovery and Prediction

Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013

G. King - Preface: Big Data Is Not About The Data, in R. Michael Alvarez, ed., In press, Computational Social Science: Discovery and Prediction - Cambridge University Press.

The charter of human rights and principles for the internet, Internet Governqance forum, United Nations, 2014

Wigan, Marcus R., and Roger Clarke. "Big data's big unintended consequences." Computer 46.6 (2013): 46-53.

Labrinidis, Alexandros, and Hosagrahar V. Jagadish. "Challenges and opportunities with big data." *Proceedings of the VLDB Endowment* 5.12 (2012): 2032-2033.

Boyd, Danah, and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society 15.5 (2012): 662-679.

- Madden, Sam. "From databases to big data." *IEEE Internet Computing* 16.3 (2012): 4-6.
- Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013.

General on Challenges & Opportunities

- Labrinidis, Alexandros, and Hosagrahar V. Jagadish. "Challenges and opportunities with big data." *Proceedings of the VLDB Endowment* 5.12 (2012): 2032-2033.

Economic Value vs Social Utility

McKinsey Global Institute – Big data: The next frontier for innovation, competition, and productivity, 2011.
Shapiro and Varian R. Information Rules, Harvard Business Review Press, 1999.
Rifkin J. The Zero Marginal cost Society, Palgrave 2014.
Staglianò R. – Al Posto Tuo, Einaudi, 2016
OECD - The Well-being of Nations: the Role for Human and Social Capital, 2001.
Mc Kinsey - The social economy: Unlocking value and productivity through social technologies, 2012.
T. Bold, B. Gauthier, J. Svensson Waly Wane - Delivering Service Indicators in Education and Health in Africa A Proposal, Policy Research Working Paper 5327, 2010.
M. Björkman N. Damien de Walque J. Svensson - Information is Power Experimental Evidence on the Long-Run Impact of Community Based Monitoring Development, Policy Research Working Paper 7015, 2014.
Big Data for development: Harnessing Big Data For Real-Time Awareness www.unglobalpulse.org, June 2013.
Big Data for Development: Challenges & Opportunities, http://unglobalpulse.org/ May 2012.
Big data and human development: Investigating the potential uses of 'big data' for advancing human development and addressing equity gaps, Oxford Internet Institute, 2016.
By Kevin C. Desouza & Kendra L. Smith - Big Data for Social Innovation

Numeration & Digitization & Datafication

Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013

C. A. Mulligan The impact of Datafication on Strategic Landscapes, Ericsson, 2016.

J. Harle, Datafication and democracy: Recalibrating digital information systems to address societal interests, 5th January 2017

M. Jerven Poor Numbers. How We Are Misled by African Development Statistics and What to Do about It - School for International Studies Simon Fraser University

E. Letouzé, J.Jütting – Official Statistics, Big Data and Human Development – Data-Pop Alliance, 2015.

Mark Freeman - Quantitative Skills for historians - The Higher education academy, 2012.

L. Gitelman - “ Raw Data ” Is an Oxymoron, 2013 Massachusetts Institute of Technology

From Why to What, or: with enough data, “the data speak for themselves” (the end of theory)

Anderson, C., (2007),‘The end of theory: the data deluge makes the scientific method obsolete’, Wired, available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (last accessed 26 July 2013).
V. Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013
M. Duggan, S. Levitt - Winning isn’t everything: corruption in Sumo Wrestling, NBER Working Paper Series.
G. C. Bowker - The Theory/Data Thing, International Journal of Communication 8, 2014.

Inexactitude

Harvey J. Miller & Michael F. Goodchild - Data-Driven Geography, GeoJournal 80(4):449-461 · August 2015.

V. Mayer Shonberger, K. Cukier – Big Data: a Revolution that will transform how we live, work and Think, 2013

D. Shenk – Data Smog, Harvard Journal of Law and Technology, Volume 12, N. 2, 1999.

Big Data Hubris

Lazer D., Ryan Kennedy R., Gary King G., Vespignani A. - The Parable of Google Flu: Traps in Big Data Analysis Big Data, Science, 2014.

K. Roberts, The Big Data Pandemic, Forethought.

C. Moraff - Beware of “Big Data Hubris” When It Comes to Police Reform, Parsons, 2016

R. Read, B. Taithe & R. Mac Ginty - Data hubris? Humanitarian information systems and the mirage of technology, Third World Quartelry, Rutledge, 2017.

D. Lazer, R. Kennedy, G. King, A. Vespignani - The Parable of Google Flu: Traps in Big Data Analysis, Science 343 (6176) (March 14): 1203–1205.

Transparency, privacy and determinism

Rand – Predictive Policing - The Role of Crime Forecasting in Law Enforcement Operations, Rand Corporation, 2013.
S. Goel, M. Perelman, R. Shroff, D. Sklansky - Combatting Police Discrimination in the age of Big Data, 2016.
Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, Duncan J. Watts - Predicting consumer behavior with Web search, PNAS, October 12, 2010.
Computing Ethics: the question of infomation justice, Communications of the ACM, March 2016.
Rand Corporation, Predictive Policing, The Role of Crime Forecasting in Law Enforcement Operations, 2013.
M Andrejevich - To Preempt a Thief, International Journal of Communication 11(2017), 879–896.
Post on Predictive Policing: From Neighborhoods to Individuals, 2017.
D. Brin – The transparent Society, Harvard Journal of Law and Technology, Volume 12, N. 2, 1999.

Divide

Andrejevic M. - The Big Data Divide, International Journal of Communication 8 (2014).
Official Statistics, Big Data and Human Development - Letouzé E., Jütting J., Data-Pop Alliance, 2015.
Data and discrimination: collected essays, Open Technology Institute, 2016.

Apophenia

Overload & Abstraction, or «too big to know»

Moody, Daniel L., and Peter Walsh. "Measuring the Value Of Information-An Asset Valuation Approach." *ECIS*. 1999.

Rage amplifier

Fan, Rui, et al. "Anger is more influential than joy: Sentiment correlation in Weibo." PloS one 9.10 (2014): e110184.
Peter Sloterdijk, Ira e tempo. Saggio politico-psicologico, a cura di Gianluca Bonaiuti, traduzione di Francesco Pelloni, Roma, Meltemi 2006

P. Sloterdijk - Rage and Time: A Psychopolitical Investigation - Columbia University Press

Lazlo Barabási et al., Computational Social Science, Science, Vol 323, 2009.

R. Fan, J. Zhao, Y. Chen and K. Xu, Anger is More Influential Than Joy: Sentiment Correlation in Weibo, Springer, 2013.

Most Influential Emotions on Social Networks Revealed, Post, 2013.

Morgan Maxwell, Rage and social media: The effect of social media on perceptions of racism, stress appraisal, and anger expression among young African American adults, Virginia Commonwealth University, Thesis, 2016.

Visualization and lies

E. Tufte - *The Visual Display of Quantitative Information*. Cheshire, Graphics Press. 1983

From fake news to Trump staff's «alternative facts»

World Economic Forum - Global Risks 2013.
Cock J., Lewandowsky S. – The Debunking Handbook, University of Queensland, Australia, 2012.
Thomson M. What's gone wrong with the language for P. Fenbach, S. Sloman, Why We Believe Obvious Untruths, March 3, 2017
W. Quattrociocchi, A. Vicini – Misinformation: guida alla società della informazione e della credulità, Franco Angeli, 2016.
W. Quattrociocchi How Misinformation Spreads Online, Power point presdentation, available at

Echo chambers

L. Schmidt, F. Zolloa, M. Del Vicarioa, A. Bessi, A. Scala, G. Caldarella, H. Eugene Stanleyd, and W. Quattrociocchi – Anatomy of news consumption on Facebook, PNAS, January 2017.

W. Quattrociocchi, A. Vicini – Misinformation: guida alla società della informazione e della credulità, Franco Angeli, 2016.

Bibliografia – non classificati

- Freeman M. – Quantitative Skills for Historians, The higher education Academy, 2010
- Zuckerman – Digital Cosmopolitans: Why we think the Internet connects us, Why it doesn't and how to rewire it, Rewire, 2013.
- R. Anthony Gartner - Data Analytics and the Disintegration of Public Knowledge in http://atheistnexus.org/group/atheistswholovescience/forum/topics/data-analytics-and-the-disintegration-of-public-knowledge?xg_source=activity
- https://www.slideshare.net/siddharthhande/examining-data-practices-cyberabads-publicly-accessible-crime-map
- http://www.ph.ucla.edu/epi/snow/snowbook3.html

Resti

William Shakespeare*, from "Hamlet"*

There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy.

- *Hamlet* (1.5.167-8), Hamlet to Horatio

From EMC Digital Universe with Research & Analysis

The digital universe is large – by 2020 containing nearly as many digital bits as there are stars in the universe.