

Data, responsibly (part 2)

Serge Abiteboul
Inria & ENS, Paris

Julia Stoyanovich
Drexel University



Data, responsibly

A. Introduction

B. Responsible data analysis

C. Technical issues

D. Societal issues

E. Conclusion

Responsible data analysis

1. Fairness
2. Diversity
3. Transparency
- 4.  Neutrality**
5. Data protection



B.4 Neutrality

Neutrality

- No bias
 - e.g., in favor of your owner or customers of the owner
- Critical on the internet/web
- Objective criteria such as popularity
 - Page rank
 - Number of clicks, likes...

Network neutrality

- Net neutrality : the principle that Internet service providers and governments regulating the Internet must treat all data on the Internet the same, not discriminating or charging differentially by user, content, website, platform, application, type of attached equipment, or mode of communication

≠ Internet service provider Comcast secretly throttling (slowing down) uploads from peer-to-peer file sharing (P2P) applications by using forged packets

Service neutrality

Counterexamples

- A recommendation service that gives better ranking to paid customers
- A smartphone OS that tries to discourage you to use particular services
- A search engine that gives low ranking to services in competition with its own services

Often difficult to prove

European commission announces antitrust charges against Google

Inquiry will focus on accusations that internet search and tech multinational has unfairly used its products to oust competitors

Sam Thielman in New York

[@samthielman](#)

Wednesday 15 April 2015 07.27 EDT



📷 Ruth Porat replaces Patrick Pichette as Google's chief finance officer. Photograph: Georges Gobet/AFP/Getty Images

The European Union accused Google on Wednesday of cheating competitors by distorting Internet search results in favour of its Google Shopping service and also launched an antitrust probe into its Android mobile operating system.

Why bother?

- A few companies are concentrating all the data and most of the computing power
- Threatening fair competition
- Threatening freedom
 - Control the information we get
 - Guide our choices
 - Guide our decision

An argument against neutrality that does not work

- A newspaper makes editorial choices, chooses what to present, the product to push
- But
 - Google search is used by 90% in Europe
 - Facebook hosts most of your friends
 - Android + IOS have most of the smartphone market
- Where are the choices?

Good reasons not to be neutral

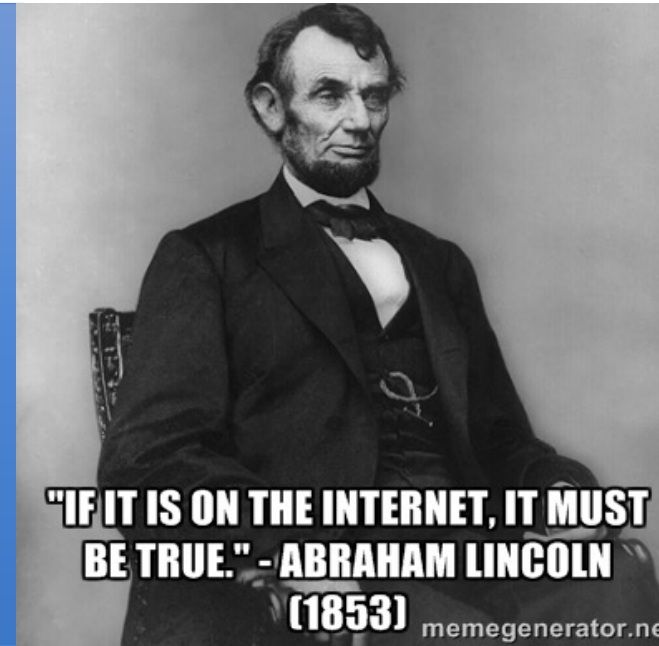
- Personalization
- You look for a pizza or a doctor
 - Near where you are not in Bay area
- Avoid pornography for young public
- ...

Correctness: a good argument against neutrality?

- Correct errors, block fake news...
- Evaluate trustworthiness of sources
- Who decides? Experts, governments...

April 2017/ Turkey blocked all access to Wikipedia because Wikipedia "had failed to remove content promoting terror and accusing Turkey of cooperation with various terror groups" to follow in the coming days."

Wikipedia was mentioning the totalitarian nature of the regime





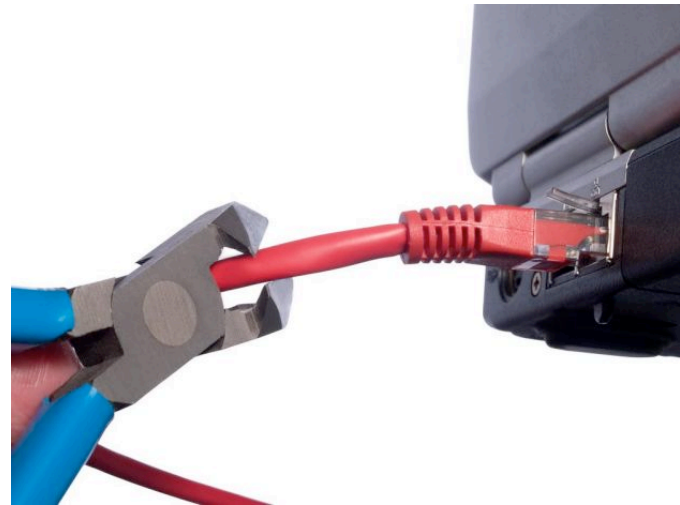
B.5 Data protection

Protect data

- Long experience with business data
 - Access control
 - DRM (digital rights management)
- Much less with personal data

Data privacy

- More and more concerns privacy
- Limitations on what data companies can do
 - Laws to force companies to request authorization to build a DB with of personal information (France)
- Rules about what users should be able to do
 - Laws that compel companies (e.g., credit reporting agencies) to let users see and correct information about them (US)
- Laws in Europe, US...
 - The laws depend on the country
 - Their enforcement is difficult

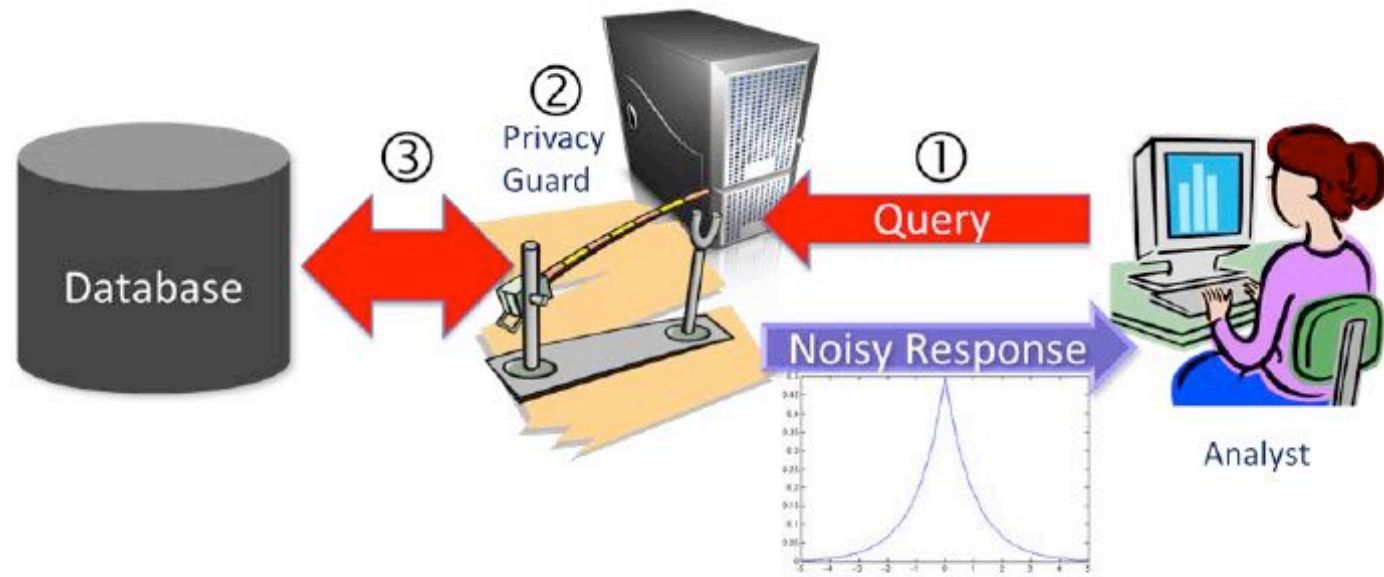


Data privacy: usability

- There are means to protect data but people often don't use them because too complicated and/or not understandable
 - Tools for cryptography
 - Access rights
 - Unreadable EULA
 - End-User license agreement
 - Difficulty to change service, portability
 - Vendor lock in

Privacy in data analysis

- When publishing statistics, protect individuals
- Already studied a lot if topic is not closed
 - Anonymization
- Differential privacy



Protecting data out there

- For the data we have on the Web, we would like to control
 - By whom it is read
 - How is transmitted
 - How it is modified
 - How it is and will be used
- We would like to keep some control in this distributed setting
 - Web-scale access control

Lots of open issues

Examples of active data access

- Auto-deletion after some amount of time
 - Snapchat
 - Vanish <https://vanish.cs.washington.edu/>
- Deletion and right to be forgotten
- Active data (activexml)
 - Provenance is an extensional reference to a source
 - Active data is an intensional reference
 - Reflect modifications of data and/or access rights

Managing access in the distributed setting

- Specifies who can read some data based on its provenance
- Data transmitted from peer to peer carries its access rights
- Webdamlog [Moffitt *et al.*, SIGMOD 2015; Abiteboul *et al.*, ICDT 2016]
- Social networks: [Cheng *et al.*, PASSAT 2012; Hu *et al.*, TKDE 2013]

- A. Introduction
- B. Responsible data analysis
- C. 📺 Technical issues**
- D. Societal issues
- E. Conclusion



C. Technical issues

C. Technical issues

1. **Testing and verification**

2. Enforcing properties and mitigating when they don't hold
3. Provenance, tracing, and reproducibility
4. Explicability or interpretability
5. Open data and open-source software

C.1 Testing and verification

The two sides of the problem

Producing data analysis

- Assuming people want to work ethically
- Tools to help
 - Collect data responsibly
 - Process, in particular analyze, it responsibly
 - Publish the results so that others can understand and reuse them responsibly

Using data analysis produced by others

- Tools to help
 - Find analysis results of interest
 - Verify that the analysis was performed responsibly

Responsibility by design: both are simpler if responsibility is taken into account as early as possible

The two sides of assessing whether an analysis was performed responsibly

- Verify its code
 - Proof of mathematical theorems
 - E.g., proof of the 4-color theorem
- Test its effect
 - Study of phenomena
 - E.g., analyze climate changes

Verification of the code

- **Audit** of the code by a specialist: difficult
 - Easy to detect gross violation such as the use of a protected variable
 - Hard to detect complex ones especially if the code developer has tried hard to hide them
- **Automatic static analysis** of programs
 - In the spirit of automated theorem proving
 - Lots of work in different areas
 - Security, safety, optimization, privacy
 - Verification of the code of a very large subset of C [CompCert]
 - Little on responsibility

Testing the effects

- The system is seen as a **black box**
- Its behavior is studied
- For instance, for a search engine
 - One observes the ranking on some queries
 - One observes the evolution of the ranking
 - One tries to reconstruct the ranking function: **reverse engineering**
- Typically use data analysis to verify properties
 - Statistical and big data analysis
 - Machine learning

Difficulty: the specification of properties

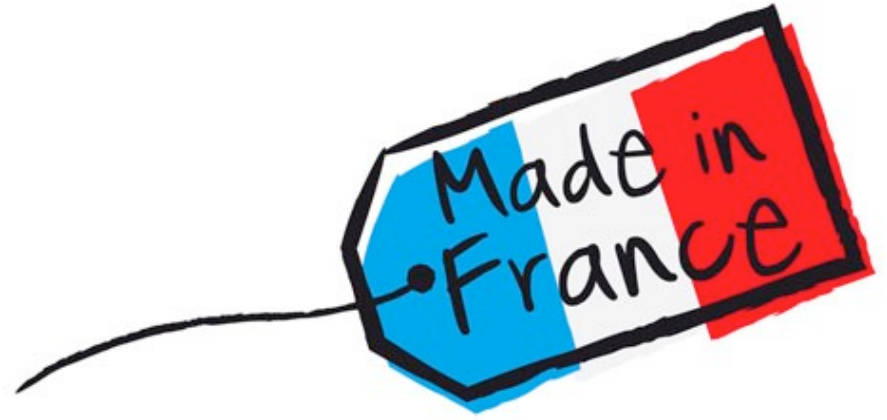
- Some properties are subtle
 - What kind of fairness is desired?
 - Recall discussion of fairness (group vs. individual)
- Some properties may depend on environment
 - The same picture may be seen as pornographic in some countries and not in other
 - Some video may be seen as too violent for a child and OK for an adult



C.2 Enforcing properties

Enforcing an outcome

- Start from an a priori choice
- Select requirements on the result of the analysis
- Tune the analysis system (for instance, train a neural net)
- This may be for possibly widely different reasons, e.g.,
 - Ethical properties: In some university admission, bring the number of women in each school to 50%
 - Make sure the services of preferred customers are in top 10 of rankings
- Bias the results to « enforce » this to happen



C.3 Provenance, tracing, and reproducibility

Provenance

- Specifies the data origin and the processing performed
- Many motivations
 - Helps understand the result
 - Simplify the verification/testing of the results
 - Enables “reproducibility”
 - Helps build trust
- Common for scientific data
 - Publishing results without explanation is not scientific
- Still not common for data on the web – many excuses
 - The value of the data (business advantage)
 - Publishing the Ranking criteria of Google Search would facilitate the work of those who try to manipulate it

Provenance

- Provenance [Green *et al.*, PODS 2007], [Green *et al.*, SIGMOD 2007]
- Provenance and privacy [Davidson *et al.*, ICDT 2011]



What we would like for web data

- Control
 - By whom it is read
 - How is transmitted
 - How it is modified
 - How it is and will be used
- We would like to keep some control even in this distributed setting
- Web-scale access control

Example of distributed access control

- Webdamlog
- Specifies who can read some data based on its provenance
- Datalog (aka declarative) specification
- Data transmitted from peer to peer keeps its access rights



Maybe my
explanation
was not very
clear...

C.4 Explicability

Do we need explanations?

- Netflix
 - Recommendation for a movie
 - If the recommendation is not good, we switch to a different provider
- Loan application
 - Explain why rejected or why a particular rate was chosen
- Release on parole
 - If denied tenure, you are entitled to know why
- Depends on the context
- No explanation may be sometimes be socially unacceptable

The origin of the problem

- Big data computation hard to explain
- Machine learning computation even harder
- Usually, many steps to reach a result and explanation wants to capture the whole pipeline
- Each variable alone may not have impact, when together they do
 - Refer to quantitative input influence before
- Hard doesn't mean impossible (on-going works)
 - Explain which variables mattered in a choice

Example: ranking

- Input: collection of items
- Score-based ranker: computes the score of each item using a formula, e.g., monotone aggregation, then sorts items on score
- Output: ranked permutation of the items
- Transparency: scoring function is known
- Explicability? not really
 - What is a good score? What has impact? Would a variation of the weights have huge impact? ...



C.5 Open data and open-source software

In sciences

- Sciences progress with the sharing of ideas
 - The building of trust
- Open science
 - Publication in open access
 - Reproducible experiments
 - Description of the raw data
 - Description of the processing
 - Open data

Open data access

- Publishing data
- Finding data
- Using data
- Based on the licenses CC BY, SA, NC, ND

Can this be a model for society?

Open-source software

- Simplifies/enables verification
- A large community of users can
 - Verify the code
 - Disagree with the policies
- Example the APB software in France
 - Assignment of students to higher education
 - Opaque procedure for a long while
 - Now open-source
 - Lead to a debate on the policy
- Useful but not sufficient
 - Bug in the SSL library of Debian
 - Weak secrecy of keys for 2 years

Open data

- Basis of transparency
- Facilitates also verification
- The publication of open data also encourages innovation
 - From small companies that could not get the data otherwise

Limitations of “open”

- Dubious: opaqueness is a protection
 - Argument used for instance by search engines for not providing their ranking algorithm
 - Other companies can “re-engineer” the ranking
 - Opaqueness mostly used to manipulate ranking?
- Serious: conflicts with the protection of private data
- Serious: data and software are sometimes main assets for companies
 - Ways around: providing realistic datasets that preserve privacy

- A. Introduction
- B. Responsible data analysis
- C. Technical issues
- D. 📌 Societal issues**
- E. Conclusion



D. Societal issues

D. Societal issues

1. Accountability

2. Pims and impact on business models
3. Education
4. Associations
5. Governments
6. Changing the web



D.1 Accountability

Accountability

- A service must perform as it should according to contract
 - In particular, user data should be protected as specified
- Difficulty
 - The contract is typically confusing and imprecise
 - Manipulation of nutritional labels, e.g. replace sugar with 3 kinds to confuse consumers
- In case of violation
 - Legal aspects
 - Possibly very damaging in terms of reputation
 - See further

Always the same issues

- As a company, make sure that you do what you promiss
- As a user, as a customer, as the government, verify that a service performs as announced / as in the contract
- Not easy
- Example: in France, government efforts to adress the problem, the TransAlgo institute



D.2 Personal information management systems

A cloud system that manages all the information of a person

An alternative business model protecting the personal data

The Personal Information Management System

Web services

- Each one running on some unknown machines
- With your data
 - You don't know precisely which
- Some software

Your Pims

- Your machine
- With your data
 - You have access to all
 - Possibly replica of data from systems you like
- Running software you chose
 - Or wrappers to external services you chose

It's first about data integration

	m i m i										A L I C E	l u l u z a z a									
localization											X										
webSearch											X										
calendar											X										
mail											X										
contacts											X										
Facebook service												Integration of the users of a									
tripadvisor											X										
banks											X										
whatsapp											X										

Integration of the services
of a user

Integration of the users of a

Many R&D issues

Old problems revisited

- Personal information integration
- Personalization and context awareness
- Personal data analysis
- Epsilon-principle (epsilon-user-administration)
- Synchronization/backups & Task sequencing
- Access control & Exchange of information
- Security (e.g. works @ INRIA Saclay)
- Connected objects control

Is this is arriving?

- Technical
 - Price of hosted machines is going down
 - More and more relevant open-source software
 - Economical
 - More and more startups into it
 - Big companies interested as well
 - Societal
 - People are more and more concerned about privacy
- 😞 unclear business model
- 😄 may bring new functionalities

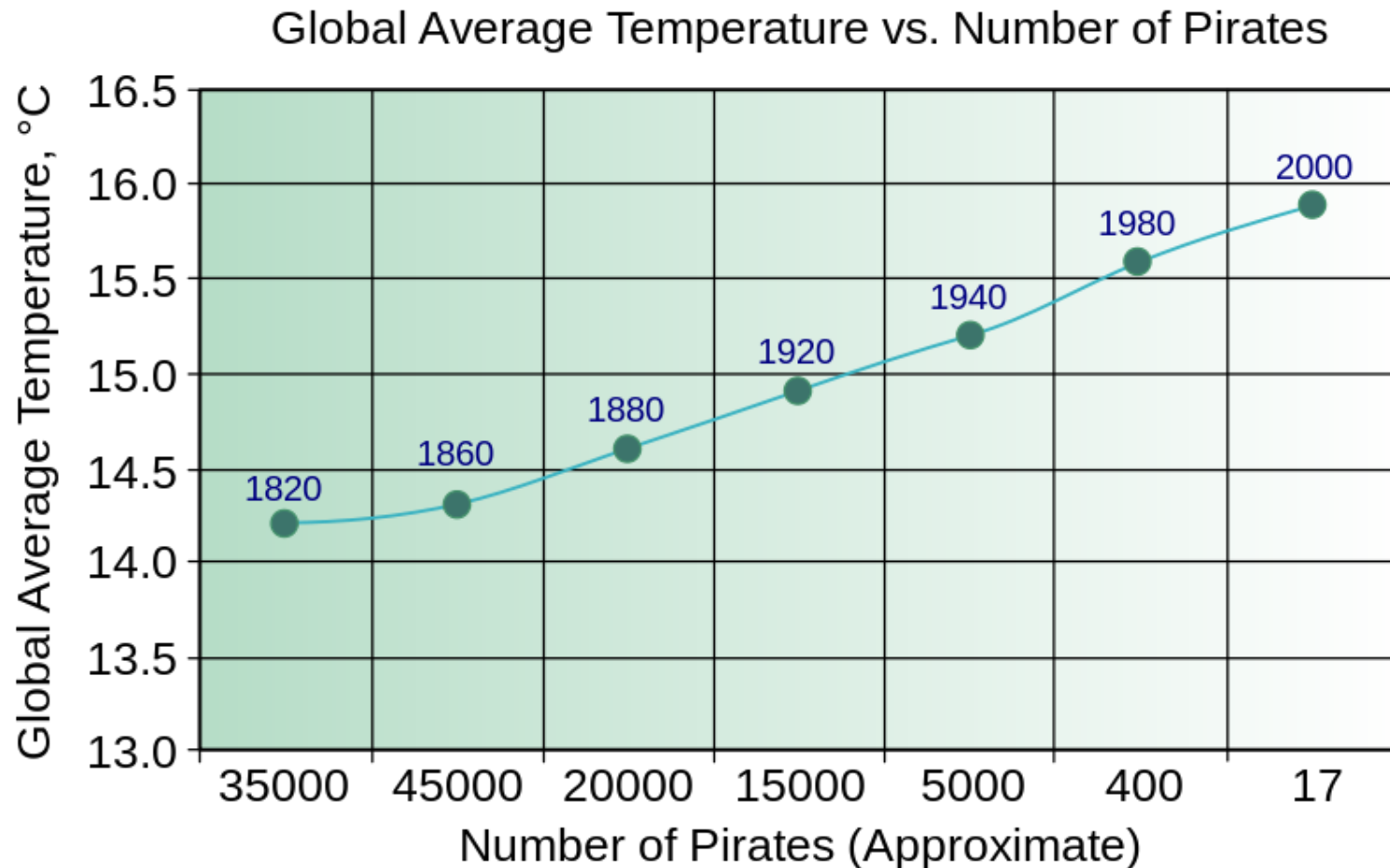


D.3 Education

Education

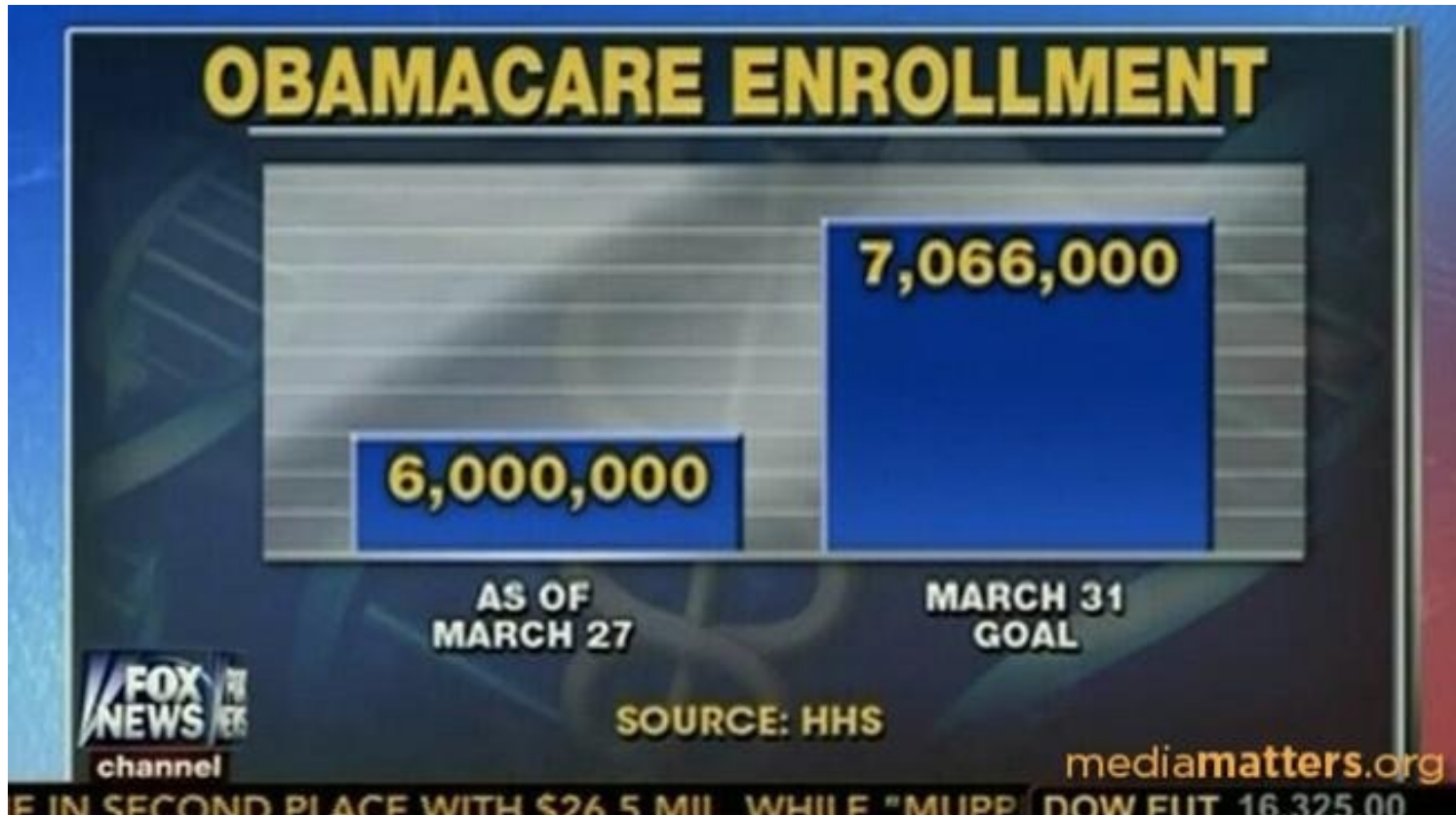
- Computer and data literacy
 - Understand computational thinking
 - Understand data analysis principles
- Topics: computer science, probability and statistics
- Learn to question and doubt

Education: correlation, causation

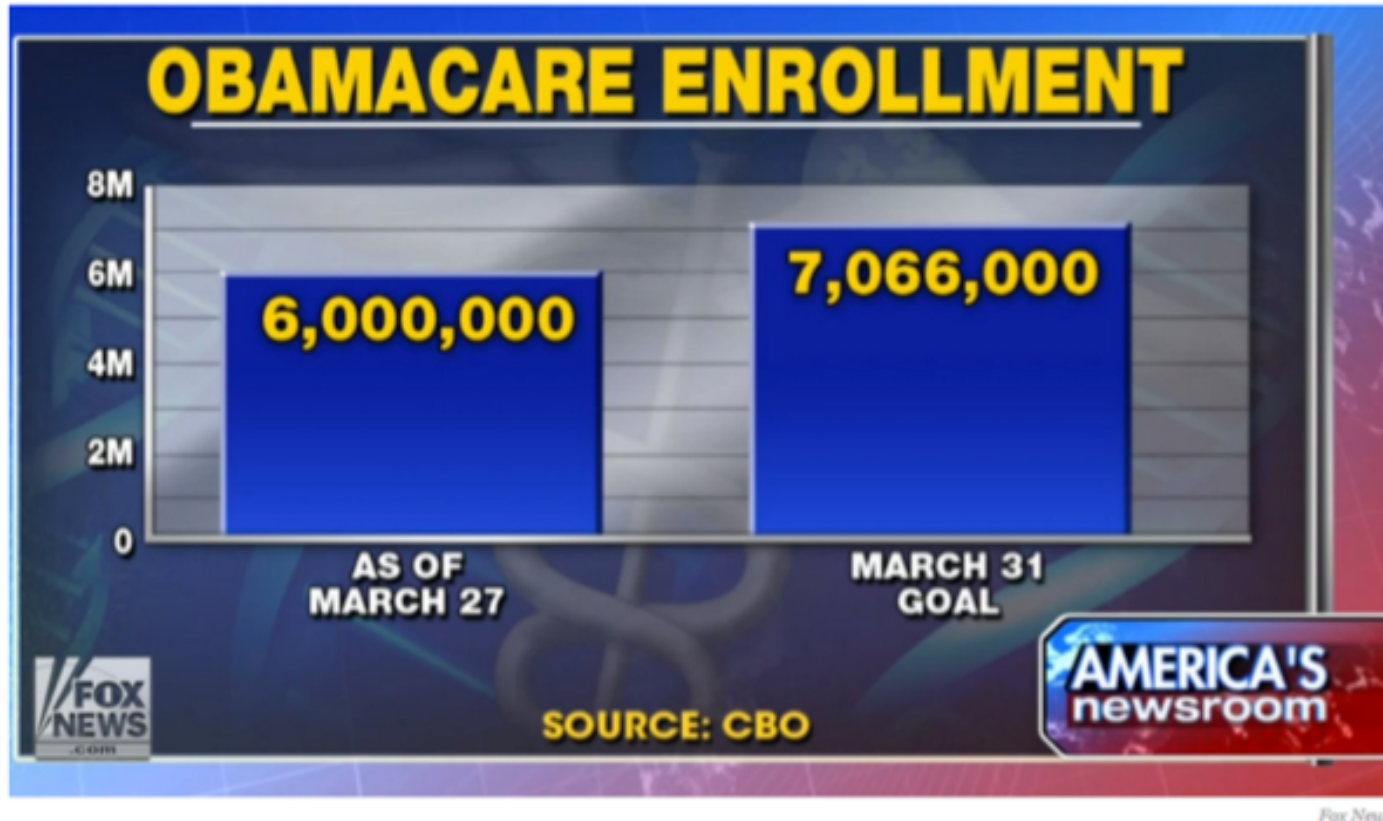


<https://en.wikipedia.org/wiki/File%3aPiratesVsTemp%28en%29.svg>

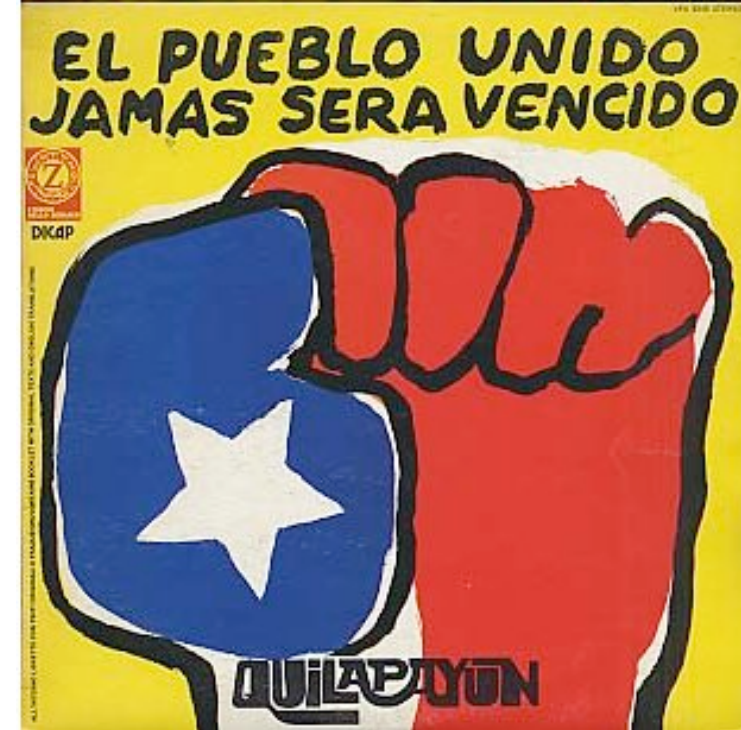
Education: data visualization



Education: data visualization



<http://www.businessinsider.com/fox-news-obamacare-chart-2014-3>



D. 4 Associations

Reputation

- Big companies at the time of the web are extremely concerned with reputation
- Bad image can spread very rapidly
- Example
 - Instagram/FB case in 2012
 - Change of a policy
 - Complaints of customers on the web, who started cancelling their accounts
 - The company decided to cancel the new policy after a few days

The power of associations

- Assessment of companies behaviors
 - Read the EULA (End-user license agreement)
 - Test/verification of the services
- Alert a large number of members very rapidly
- Start a negative campaign of the web
- Possibly engage in class actions



D.5 Governments

How can government help?

- Laws and regulations
 - At least in democratic countries
 - Prohibit or discourage bad behaviors
 - Possibly very technical
 - Service interoperability and vendor lock-in in Europe
- Define and encourage good practice
 - Publish guidelines
 - Support research
- Provide measurements
 - Agencies to detect biases
 - Support to organizations doing that

Issues

- Lack of competence
 - In parliaments & government agencies
 - Improving
- Lack of agility
 - Time of digital world vs. time of law
 - Procedure for reporting and blocking terrorist web sites in France in weeks - new replication sites active in minutes
- World-wide problem
 - Weakness of laws at the planet level
 - Example: major American companies make huge profits in Europe and pay little taxes (also an ethical issue)

Facebook “like” button



Technology | Wed Mar 9, 2016 1:22pm EST

Related: TECH, FACEBOOK, REGULATORY NEWS, BREAKINGVIEWS

German court rules against use of Facebook "like" button

FRANKFURT

A German court has ruled against an online shopping site's use of Facebook's "like" button on Wednesday, dealing a further legal blow to the world's biggest social network in Germany.

The Duesseldorf district court said that retailer Peek & Cloppenburg failed to obtain proper consent before transmitting its users' computer identities to Facebook, violating Germany's data protection law and giving the retailer a commercial advantage.

The court found in favor of the North Rhine-Westphalia Consumer Association, which had complained that Peek & Cloppenburg's Fashion ID website had grabbed user data and sent it to Facebook before shoppers had decided whether to click on the "like" button or not.

US legal mechanisms

- [Big Data: A tool for inclusion or exclusion? FTC Report; 2016]
- <https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>
- Fair Credit Reporting Act - applies to consumer reporting agencies, must ensure correctness, access and ability to correct information
- Equal opportunity laws - prohibit discrimination based on race, color, religion, ... - plaintiff must show disparate treatment / disparate impact
- FTC Act - prohibits unfair or deceptive acts or practices to companies engaged in data analytics
- Lots of gray areas, much work remains, enforcement is problematic since few auditing tools exist

EU legal mechanisms

- Transparency
 - Open data policy: legislation on re-use of public sector information
 - Open access to research publications and data
- Neutrality
 - Net neutrality: a new law, but with some limitations
 - Platform neutrality: the first case against Google search
- Different countries are developing specific laws, e.g., portability against user lock-in (France)



D.6 Changing the web

The web


- We take it for granted
- But it is changing
 - Smart phones and apps
 - Presence of private companies: advertisement and hunt for private data
 - Presence of governments: China, Russia, intelligence everywhere
- Time for a new web?
- Lots of research issues & political aspects

The future of the web

Organizations are working on it

- Internet Government Forum (UN)
- Global Internet Policy Observatory (EU)
- W3C Technology Policy Internet Group

Researchers should participate

- A. Introduction
- B. Responsible data analysis
- C. Technical issues
- D. Societal issues
- E.  **Conclusion**



E. Conclusion

Ethical data management

- Major achievements of data management in the passed
 - Scaling to large volume with good performance
 - Models and query languages (data, knowledge)
 - Transaction, reliability
 - Primarily business data
- Now private and social data
- We have learnt to manage data
- **We must learn to do it in an ethical way**

References

- <http://dataresponsibly.com>
- Dagstuhl reports
 - Data, responsibly
 - Foundations of data management
- ACM Sigmod Blog, Data, responsibly, JS & SA



Report from Dagstuhl Seminar 16291

Data, Responsibly

Edited by

Serge Abiteboul¹, Gerome Miklau², Julia Stoyanovich³, and
Gerhard Weikum⁴

¹ ENS – Cachan, FR, serge.abiteboul@inria.fr

² University of Massachusetts – Amherst, US, miklau@cs.umass.edu

³ Drexel University – Philadelphia, US, stoyanovich@drexel.edu

⁴ MPI für Informatik – Saarbrücken, DE, weikum@mpi-inf.mpg.de

The goals of the seminar were to assess the state of data analysis in terms of fairness, transparency and diversity, identify new research challenges, and derive an agenda for computer science research and education efforts in responsible data analysis and use.

An important goal of the seminar was to **identify opportunities for high-impact contributions to this important emergent area specifically from the data management community.**

http://drops.dagstuhl.de/opus/volltexte/2016/6764/pdf/dagrep_v006_i007_p042_s16291.pdf

Research Directions for Principles of Data Management (Dagstuhl Perspectives Workshop 16151)

Edited by

Serge Abiteboul, Marcelo Arenas, Pablo Barceló, Meghyn Bienvenu, Diego Calvanese, Claire David, Richard Hull, Eyke Hüllermeier, Benny Kimelfeld, Leonid Libkin, Wim Martens, Tova Milo, Filip Murlak, Frank Neven, Magdalena Ortiz, Thomas Schwentick, Julia Stoyanovich, Jianwen Su, Dan Suciu, Victor Vianu, and Ke Yi

1 Introduction

In April 2016, a community of researchers working in the area of Principles of Data Management (PDM) joined in a workshop at the Dagstuhl Castle in Germany. The workshop was organized jointly by the Executive Committee of the ACM Symposium on Principles of Database Systems (PODS) and the Council of the International Conference on Database Theory (ICDT). The mission of this workshop was to identify and explore some of the most important research directions that have high relevance to society and to Computer Science today, and where the PDM community has the potential to make significant contributions. This report describes the family of research directions that the workshop focused on from three perspectives: potential practical relevance, results already obtained, and research questions that appear surmountable in the short and medium term. This report organizes the identified research challenges for PDM around seven core themes, namely *Managing Data at Scale*, *Multi-model Data*, *Uncertain Information*, *Knowledge-enriched Data*, *Data Management and Machine Learning*, *Process and Data*, and *Ethics and Data Management*. Since new challenges in PDM arise all the time, we note that this list of themes is not intended to be exclusive.

<https://arxiv.org/pdf/1701.09007.pdf>



Serge Abiteboul and
Julia Stoyanovich

NOVEMBER 20, 2015

DATA, RESPONSIBLY

≡ Big Data

(This blog post is an extended version of an October 12, 2015 Le Monde op-ed article (in French))

Our society is increasingly relying on algorithms in all aspects of its operation. We trust algorithms not only *to help carry out routine tasks*, such as accounting and automatic manufacturing, but also *to make decisions on our behalf*. The sorts of decisions with which we now casually entrust algorithms range from unsettling (killer drones), to tedious (automatic trading), or deeply personal (online dating). Computer technology has tremendous power, and with that power comes immense responsibility. Nowhere is the need to control the power and to judiciously use technology more apparent than in massive data analysis, known as big data.

Big data technology holds incredible promise of improving people's lives, accelerating scientific discovery and innovation, and bringing about positive societal change. The goal of big data analysis is to efficiently sift through oceans of data, identifying valuable knowledge. The more data is available, the more knowledge can be derived. This gives a strong incentive for data acquisition, as well as for data sharing. Data sharing may be fully unrestricted, as is the case with the Open Data movement, or more controlled, as is the case with medical data (for privacy) and scientific or commercial data

MISSION:IMPOSSIBLE

YOUR MISSION, SHOULD YOU CHOOSE TO ACCEPT IT

- Pay attention and follow R&D codes of ethics
- Research on ethical data management
 - Together with lawyers, philosophers...
- Participate to the education of the population on these issues

Grazie
Merci
Thank you