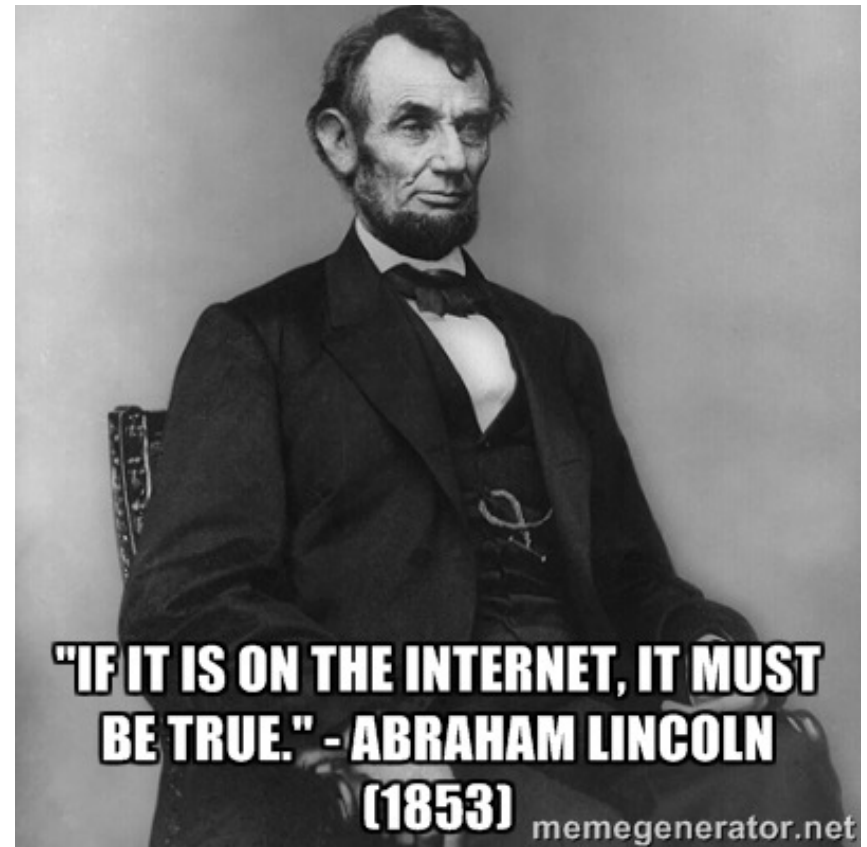


# Tutorial on Technical Issues towards Ethical Data Management

Serge Abiteboul

Some of it jointly with Julia  
Stoyanovich



# Data out there



4/14/16

# Data is exploding

## Personal data

- Data and metadata we produce
- Data others (friends or not...) produce about us
- Data sensors produce about us
- Data programs produce about us

## Web data in general

- 4V: Volume, veracity, velocity, variety

Individuals and the society are losing control over all this data

# Promises and risks of massive data

- Improve people's lives, e.g., recommendation
- Accelerate scientific discovery, e.g., medicine
- Boost innovation, e.g., autonomous cars
- Transform society, e.g., open government
- Optimize business, e.g., advertisement targeting

## Growing resentment

- Against bad behaviors: racism, terrorist sites, pedophilia, identity theft, cyberbullying, cybercrime
- Against companies: intrusive marketing, cryptic personalization and business decisions
- Against governments: NSA and its European counterparts

Increasing awareness of the dissymmetry between what these systems know about a person, and what the person actually knows

# Motivation

## An opinion

- in the past: data model & performance
- now: **personal/social data & ethics**
- Proof: We have the data models and the performance and many societal issues today are related to data

## What should be done

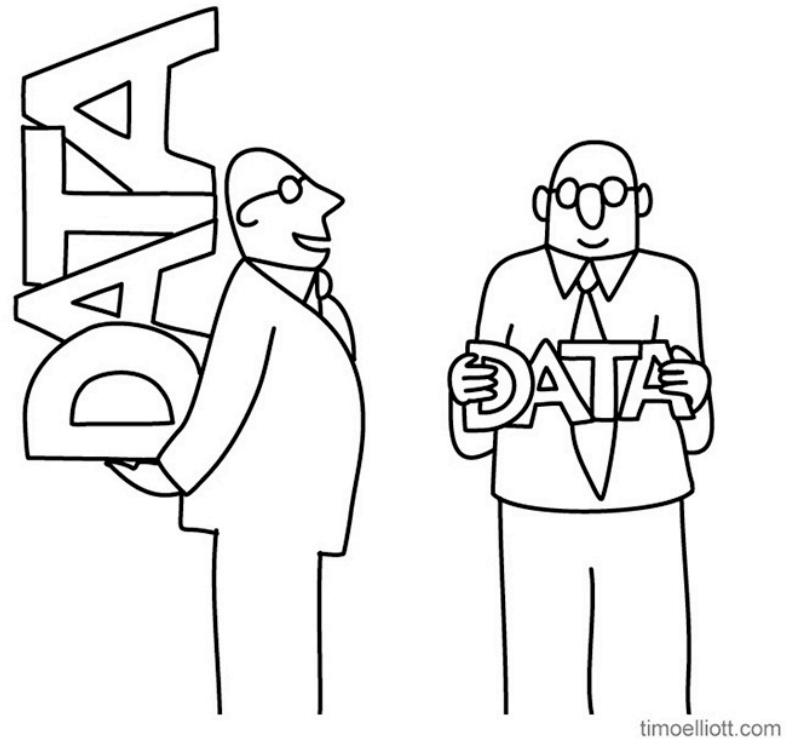
- Time to change how we deal with personal data?
- Time to change the web?

# References

- *Data responsibly*, with Julia Stoyanovich (Drexel) & Gerome Miklau (U. Mass), **EDBT Tutorial** 2016
- *Data responsibly*, with Julia Stoyanovich (Drexel), **Sigmod Blog** (in French, **Le Monde**), 2016
- *Managing your digital life with a Personal information management system*, with Benjamin André (Cozy Cloud) & Daniel Kaplan (Fing), **CACM** 2015
- *Personal information management systems*, with Amélie Marian (Rutgers), **EDBT Tutorial** 2015
- *Platform Neutrality*, **CNNum Report**, 2015

# Organization

- ✓ Motivation
- Privacy
- Data analysis
- Data quality evaluation
- Data dissemination
- Data memory



*"I think you'll find that mine is bigger..."*

1. Data privacy
2. The PIMS

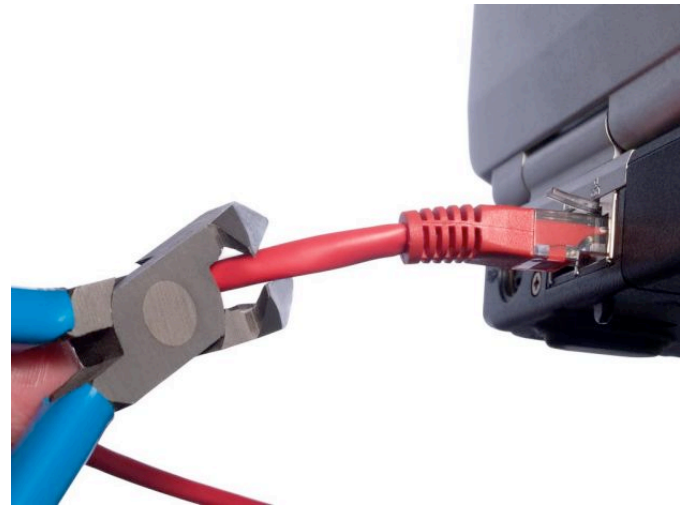
# PRIVACY





# 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
- Is the solution disconnect?



# 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
    - Vendor lock in

# Research issues

- Easier to use tools
- Automatic specification
- Portability...

# Data protection: The PIMS



*A Personal Information Management System is a cloud system that manages all the information of a person*

**Many Web services**  
**Each one running**

- On some unknown machines
- With your data
- Some software

**Your PIMS**

- **Your machine**
- **With your data**
  - possibly replica of data from systems you like
- On your software or
- Wrappers to external services

# It's first about data integration

	mimi										Alice	lulu zaza									
localization											X										
webSearch											X										
calendar											X										
mail											X										
contacts											X										
Facebook											X										
tripadvisor											X										
banks											X										
whatsapp											X										

Integration of the services of a user

Integration of the users of a service

# 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 Rocquencourt)
- Connected objects control

1. Fairness
2. Transparency
3. Diversity
4. Privacy

# DATA ANALYSIS



# Getting knowledge out of data

- Finding statistical correlations
- Publishing aggregate statistics
- Detecting outliers
- Detecting trends
- Number of techniques: data mining, big data, machine learning



# Data analysis: Fairness



- Origins of bias
  - data collection
    - E.g., a crime dataset in which some cities are under-represented
  - data analysis
    - E.g., a search engine that skews recommendations in favor of advertising customers
- This bias may even be illegal
  - Offer less advantageous financial products to members of minority groups (a practice known as steering)
- Example: analysis of scientific data
  - Should explain how data was obtained
  - Should explain which analysis was carried on it
  - Experiments should be reproducible

Very studied already – lots of research issues

# Effect on sub-populations

## Simpson's paradox

disparate impact at the full population level disappears or reverses when looking at sub-populations!

		grad school admissions	
		admitted	denied
gender	F	1512	2809
	M	3715	4727

**positive  
outcomes**

35%  
of women

44%  
of men

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

# Group versus individual fairness

## Group fairness

demographics of the individuals receiving any outcome are the same as demographics of the underlying population

		credit score	
		good	bad
race	black	⊕	⊖ ⊖ ⊖ ⊖
	white	⊕ ⊕ ⊖	⊖ ⊖ ⊖

**positive outcomes** offered credit

40%  
of black

40%  
of white

## Individual fairness

any two individuals who are similar w.r.t. a particular task should receive similar outcomes

# Data analysis: Diversity



- Relevance ranking (for recommendation)  
is typically based on popularity
  - Ignores less common information (in the tail) that constitutes in fact the overwhelming majority
  - Lack of diversity can lead to discrimination, exclusion.
- Examples
  - on-line dating platform like Match.com
  - a crowdsourcing marketplace like Amazon Mechanical Turk
  - or a funding platform like Kickstarter

The rich get richer, the poor get poorer

# Rank-aware clustering

Return clusters that expose **best from among comparable** items (profiles) w.r.t. user preferences



MBA, 40 years old  
makes \$150K



MBA, 40 years old  
makes \$150K



MBA, 40 years old  
makes \$150K



MBA, 40 years old  
makes \$150K

... 999 matches

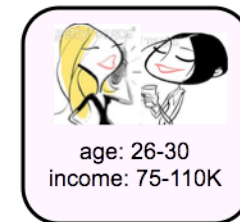
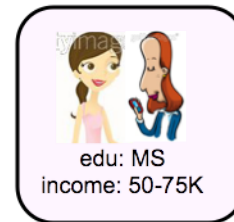
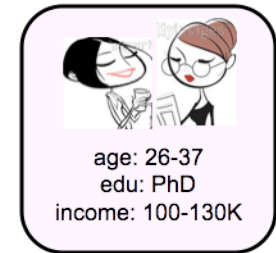
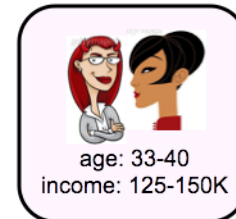
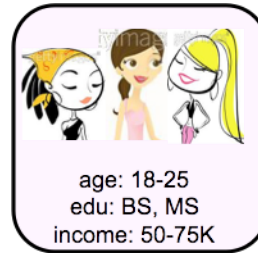


PhD, 36 years old  
makes \$100K

... 9999 matches



BS, 27 years old  
makes \$80K



[J. Stoyanovich, S. Amer-Yahia, T. Milo; EDBT 2011]

# Data analysis: Transparency



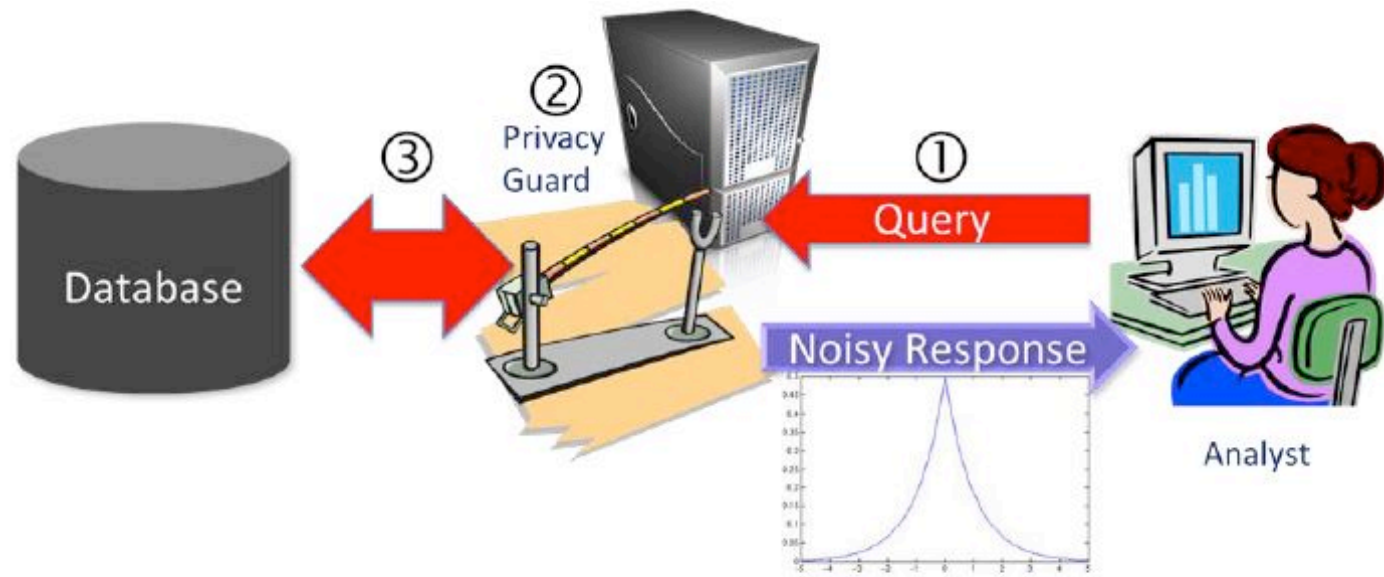
- Example: lack of transparency in Facebook data processing
  - In general, unreadable End-user license agreement
- Users want to control what is recorded about them, and how that information is used
- Transparency facilitates verification that a service performs as it should, as is promised
- Also allows a data provider to verify that data are well used as it has specified.

# Privacy in data analysis



- When publishing statistics,  
protect individuals
- Anonymization
- Differential privacy

Already studied a lot  
Topic is not closed



# Issues: Verifying these properties

- Tools to collect data and analyze it responsibly
- Tools to verify that some analysis was performed responsibly
- Easier if responsibility is taken into account as early as possible, *responsibility by design*
- To check the behavior of a program, one can
  - Analyze its code  $\approx$  proof of mathematical theorems
  - Analyze its effect  $\approx$  study of phenomena (such as climate or the human heart)



# Verification: code analysis

- Possible if open-source - otherwise auditing
- Easier with open-source
  - not sufficient: bug in the SSL library of Debian
  - Weak secrecy of keys for 2 years
- Specify properties that should be verified
- Verification based on static analysis, in the spirit of theorem proving
- Lots of work in different areas
  - security, safety, optimization, privacy
- Little on responsibility

# Verification: analysis of effects

- Statistical analysis
  - Detect biases
  - Detect illegal use of protected attributes
- Verify transparency
- Verify “loyalty”
  - The system behaves like it says it does
- Example: Google Ads Settings & AdFisher

# Google Ads Settings



## Control your Google ads

You can control the ads that are delivered to you based on your Google Account, across devices, by editing these settings. These ads are more likely to be useful and relevant to you.

### Your interests

- ☒ Action & Adventure Films
- ☒ Cooking & Recipes
- ☒ History
- ☒ Hygiene & Toiletries
- ☒ Mobile Phones
- ☒ Phone Service Providers
- ☒ Reggaeton
- ☒ Vehicle Brands

- ☒ Cats
- ☒ Fitness
- ☒ Hybrid & Alternative Vehicles
- ☒ Make-Up & Cosmetics
- ☒ Parenting
- ☒ Recording Industry
- ☒ Search Engine Optimization & Marketing

+ ADD NEW INTEREST

WHERE DID THESE COME FROM?

These interests are derived from your activity on Google sites, such as the videos you've watched on YouTube. This does not include Gmail interests, which are used only for ads within Gmail. [Learn more](#)

# Transparency and accountability

- Analysis by AdFisher
- Doesn't behave how it says
  - Choice of ads is based on more data that it says
    - E.g., protected attributes
    - Eg: males were shown ads for higher-paying jobs significantly more often than females
- Some control on the ads
  - Removing an interest decreases the number of ads related to that interest
  - Eg: cats

# Verification: provenance

- Provenance helps verifying the analysis
- Common for scientific data, essential for verifying that data collection and analysis were performed responsibly

Issue: provenance and verification

Issue: reproducibility



# DATA QUALITY EVALUATION

# Stuff we don't want to see on the web

- Nazi sites
- Terrorist sites
- Pedophilic content
- Bogus health content
- Conspiracy theory content
- Cybercrime
- Cyberbullying ...

# Issues: What can we do about it

- Web scale monitoring for illegal content
- Web scale automatic evaluation of
  - Quality of content
  - Legality of content
  - Based on truth and authority ranking
- Crowd-based analysis/rating of web pages
- ???

Lots of research issues



# Beyond evaluation: active countermeasures

- Typical situation in France
  - Terrorist site detected and reported
  - Long legal process
  - Condemnation – the site is closed
  - A mirror reopens in a very short time, and is quickly referenced on the web
- The URL has been prohibited when it should have been the content
- Technically, it is possible to detect the content and block it



1. Protecting data out there
2. Open data access
3. Neutrality

## DATA DISSEMINATION

# 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

- 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
- Auto-deletion after some amount of time
  - Snapchat
  - Vanish <https://vanish.cs.washington.edu/>
- Deletion and right to be forgotten

# Example: Distributed access control based on provenance

- 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

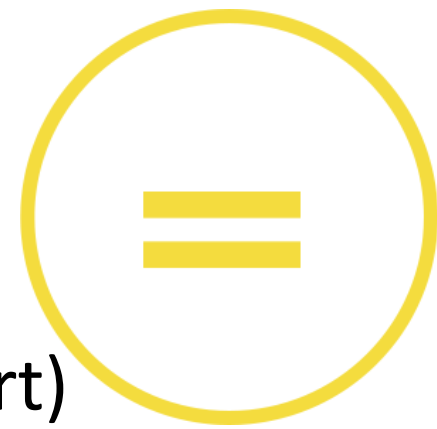
# Open data access

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

## Issues

- Ontologies, provenance, workflows...
- P2P protection of CC licenses

# Neutrality



Net and platform neutrality (CNNum report)

- net neutrality - the network is transporting data with no bias based on source, destination, content ...
- platform neutrality - big internet platforms should not discriminate in favor of their own services
- Related to fairness and diversity, verified with transparency tools

The rich get richer, the poor get poorer

# Power comes with responsibility

## Power

- A handful of big players command most of the world's computational resources and most of the data, including all of your personal data - an oligopoly

## Danger

- Threatens fair business competition
- Controls what information you receive
- Can guide your decisions
- Can infringe on your privacy and freedom
- Limits your freedom



# Google antitrust case

theguardian

## 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.

# Issues

Testing neutrality

Monitoring of neutrality

1. Personal data
2. Archiving
3. Web archiving

# DATA MEMORY



# Remembering data

Issues: decide

- What to remember
- What to forget
  - Forgetting is a key to abstracting
- Ranking, summarization...

E.g., ForgetIT EU project

# Conclusion

Many societal and political fights today are related to data

The issues are clearly non only technical

Time to change the way we use personal data?

Time to change the web?

Organisms are working on it

- For instance, CNNum Governments (US, EU...)

For instance, for the web

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



<http://abiteboul.com>  
<http://binaire.blog.lemonde.fr>