Toward personal knowledge bases

Serge Abiteboul
INRIA & Ecole Normale Supérieure Cachan &
Conseil National du Numérique

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

Organization

1. Personal data
2. Responsible data analysis
3. The Pims
    1. The concept of Pims
    2. The Pims are arriving and that is cool
    3. Research issues
4. Conclusion

1. Personal data

Personal data out there

Personal data is heterogeneous

- Structured: relational
- Semistructured: HTML, XML, Jason…
- Not structured: text (pdf), pictures, music, video…
- Metadata: date, location…
- Semantic: RDF, RDFS, Owl

- Different languages, terminologies, ontologies, structures
- Different systems, protocols
- Varying quality

Personal data is exploding

- Actively: Data and metadata we produce
  - Pictures, reports, emails, tweets, annotations, recommendation, social network…

  Actively: Data we like/buy
  - Books, music, movies…
- Passively: Data others produce about us
  - Public administration, schools, insurances, banks…
  - Amazon, retailers, netflix, applestore…
- Stealthily: sensors
  - GPS, web navigation, phone, "quantified self" measurements, contactless card readings, surveillance camera pictures…
- Stealthily: computed by programs

…

Bad news (1)

- Loss of functionalities because of fragmentation
  - You don't know where your data is, how to maintain it up to date, how to get it sometimes
  - Difficult to do global search, maintenance, synchronization, archiving...
- Loss of control over the data
  - Difficult to control privacy
  - Difficult to control sharing
  - Leaks of private information
- Loss of freedom
  - Vendor lock-in

Data analysis

- A few companies concentrate most of the world’s data and analytic power
  - They have the means to destroy business competition in large portions of the economy
- A few companies control all your personal data
  - They determine what information you are exposed to
  - They guide many of your decisions
  - They potentially infringe on your privacy and freedom.
- What should we do about that ?

Bad news (2): data analysis

These are directing our lives

- Analysis of the Web: pagerank
- Analysis of a social networks: edgerank
- Analysis of proximities
  - between people (Meetic)
  - between people and products (Netflix)
- Analysis of classification: rating for loans
- Analysis by government: marked as potentially dangerous

And they can be very “wrong”

2. Responsible data analysis

From an article in Le Monde

And a blog in ACM Sigmod

Both with Julia Stoyanovich

Impose good properties of data analysis

- Fairness
- Transparency
- Equal availability to all

Fairness – Lack of bias

- 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)
- Analogy : analysis of scientific data
  - Should explain how data was obtained
  - Should explain which analysis was carried on it
  - Experiments should be reproducible

Fairness – Neutrality

- Such a tremendous power, must come with responsibilities
  - CNNum reports on Net and Platform neutrality
- Some general resources should be « neutral playing field »
  - An Internet provider who refuses to serve Youtube videos or give degraded service for them
  - An App Store who refuses some applications for various reasons or favor some service against another
- Limits the freedom of individuals
- Threatens fair business competition

Fairness – 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 gets richer & the poor gets poorer

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.

Equal accessibility to all

- Data and analysis means more and more concentrated → oligopolies
- Natural outcome of fair competition?
- Why this is not acceptable
  - Loss of freedom of choice for the user.
  - Discourage innovation
  - Eventually leads to an increase of the price of services

How responsibility can be enforced

- Education
  - Everyone should learn basis in informatics and basis in data analysis
- Governments
  - Define principles and general guidelines
  - Encourage good practices
  - Fight against bad practices such as the building of oligopolies
- User associations
  - Example: The Instagram 2012 case
- Technology

Technology

- Should provide proper tools
  - To collect data and analyze it responsibly
  - To verify that some analysis was performed responsibly
  - Easier if responsibility is taken into account as early as possible, *by design responsibility*
- To check the behavior of a program, one can
  - Analyze its code ≈ proof of mathematical theorems
  - Analyze its effect ≈ study of phenomena (such as climate or the human heart)
- Simpler in open setting : open data, open source
  - Useful but not sufficient: bug in the SSL library of Debian
    - Weak secrecy of keys for 2 years

Technology: Machine learning

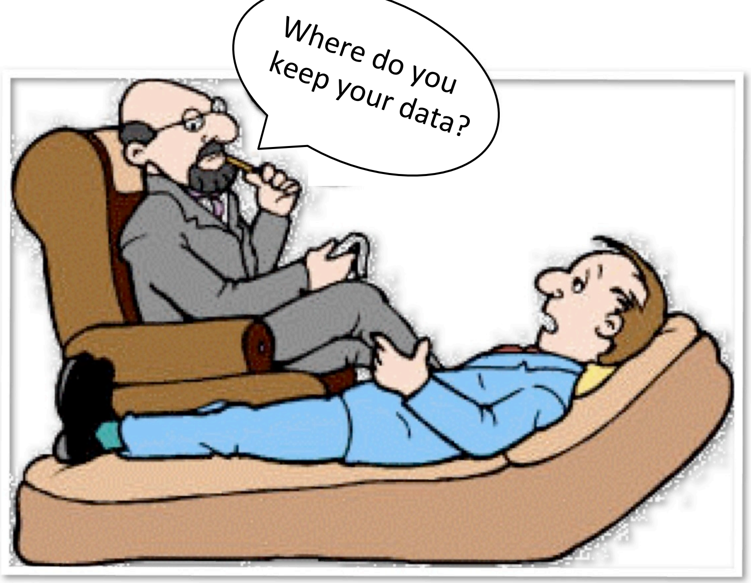
- Massive data analysis
  - Classical techniques don't scale
  - Machine learning does
- Amazing results
- But
  - Unclear scientific foundations
  - Difficult to explain specific results
  - Does not distinguish between correlation and causality

3. The Pims

From *Managing your digital life with a Personal information management system*, with Benjamin André & Daniel Kaplan, Communications of the ACM 2015

3.1 The concept of Pims

Alternatives



1. Continue with this increasing mess
   - Use a shrink to overcome the frustration
1. Regroup all your data on the same platform
   - Google, Apple, Facebook, …, a new comer
   - Use a shrink to overcome resentment
2. Study 2 years to become a geek
   - Geeks know how to manage their information
   - Use a shrink to survive the experience

The time for PIMS is now!

A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. Vannevar Bush, The Atlantic Monthly, 1945

Definition for this talk : *a **Personal Information Management System** is a cloud system that manages all the information of a person*

The PIMS: A change in paradigm

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
- Wrapper to some software
  - External service
- Or your software
  - Decentralized service

3.2 **The Pims are arriving, that is cool**

Society

Technology

Industry

Society is ready to move

- Growing resentment
  - Against companies: intrusive marketing, cryptic personalization and business decisions (e.g., on pricing), creepy "big data" inferences
  - 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
- Emerging understanding of the value of personal data for individuals
  - Quantified self

Society is ready to move (2)

- Privacy control: regulations in Europe
- Information symmetry: Vendor relation management
- Many reports/proposals that affirm the ownership of personal data by the person
- Personal data disclosure initiatives
  - Smart Disclosure (US); MiData (UK), MesInfos (France)
  - Several large companies (network operators, banks, retailers, insurers…) agreeing to share with customers the personal data that they have about them

Technology is gearing up

- System administration is easier
  - Abstraction technologies for servers
  - Virtualization and configuration management tools
- Open source technology more and more available for services
- Price of machines is going down
  - A hosted-low cost server is as cheap as 5€/month
  - Paying is no longer a barrier for a majority of people

You may have friends already doing it

Technology is gearing up (2)

- Many systems & projects
  - Lifestreams, Stuff-I've-Seen, Haystack, MyLifeBits, Connections, Seetrieve, Personal Dataspaces, or deskWeb.
  - YounoHost, Amahi, ArkOS, OwnCloud or Cozy Cloud
- Some on particular aspects
  - Mailpile for mail
  - Lima for a Dropbox-like service, but at home.
  - Personal NAS (network-connected storage) e.g. Synologie
  - Personal data store SAMI of Samsung...
- Many more

Industry is interested
Pre-digital companies

- E.g., hotels or banks
- Disintermediated from their customers by pure Internet players such as Google, Amazon, Booking.com, Mint.
- In Pims, they can rebuild direct interaction
- The playing field is neutral
  - Unlike on the Internet where they have less data
- They can offer new services without compromising privacy

Industry is interested
(2) Home appliances companies

- Many boxes deployed at home or in datacenters
  - Internet access provider "boxes", NAS servers, "smart" meters provided by energy vendors, home automation systems, "digital lockers"…
- Personal data spaces dedicated to specific usage
- Could evolve to become more generic
- Control of private Internet of objects

Industry is interested
(3) Pure Internet players

- Amazon: great know-how in providing services
- Facebook,Google: cannot afford to be out of a movement in personal data management

- Very far from their business model based on personal advertisement
- Moving to this new market would require major changes & the clarification of the relationship with users w.r.t. data monetization

Advantages – rebalance the Web

- User control over their data
  - Who has access to what, under what rules, to do what
- User empowerment
  - They choose freely services & they can leave a service
- Participation to a more "neutral" Web
  - With the "network effects", the main platforms are accumulating data/customers and distorting competition
  - The Pims bring back fairness on the Web
  - Good practices are encouraged, e.g., interoperability, portability

Main advantages: Better service for the users

This is (for me) the key ingredient for adoption

New functionalities ➳ Research issues

This is the key ingredient for researchers

This is a new playing field for startups

What are the research issues ?

3.3 Research issues with the Pims

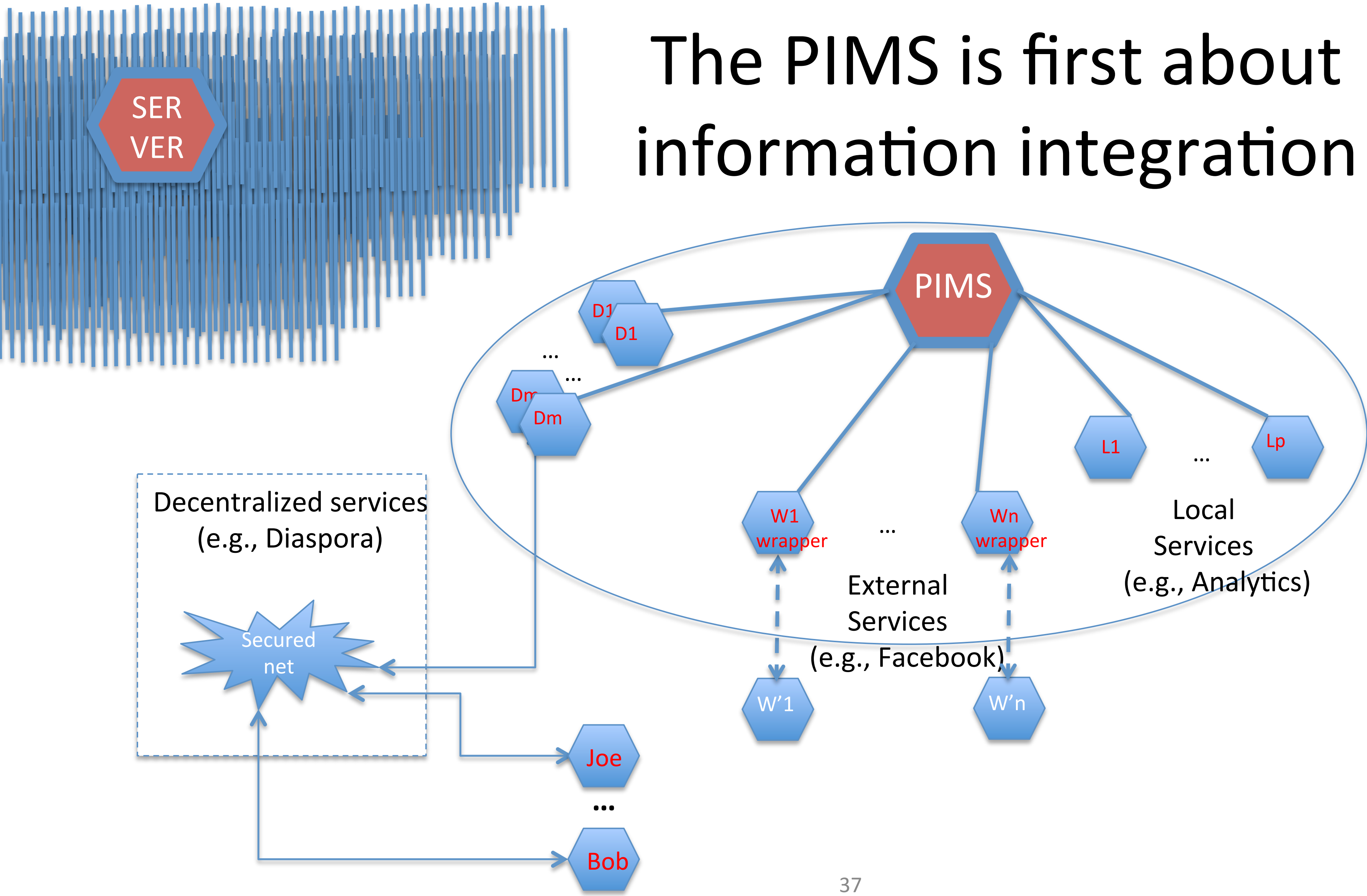
From *Personal Information Management Systems*, tutorial in Extended Data Base Technology, March 2015, with Amélie Marian

Many research issues

Some old problems revisited

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

The PIMS is first about information integration



Data analytics

- Do vertical (big) data analysis in this setting
  - Issue: data privacy
- Do small data analysis
  - Learn from personal data
    - Personal health and well-being
    - Digital personal assistant
  - Issues
    - Much smaller amounts of data – statistics harder
    - Varying data quality: imprecision, inconsistencies

4. Conclusion

Conclusion

We would like the digital world to be a better place to live in

We proposed two directions for that

1. **Responsible with data analysus**
2. **Personal information management systems**

The title was: **Toward personal *knowledge* bases**

Where is the knowledge?

Data ➽ Information ➽ Knowledge

- Personal data/info management is getting too complicated, we need software support
  - Machines prefer structured knowledge to unstructured information or semantic-free data
- So, a third direction

  3. Let us turn all our information into a distributed knowledge base

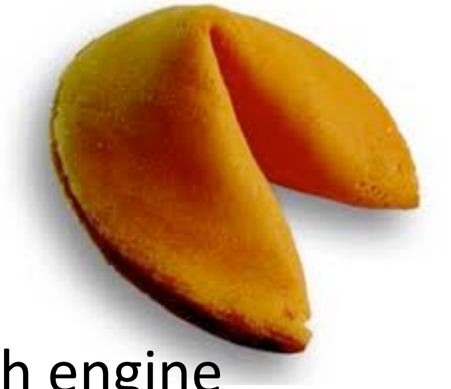
ERC Webdam, http://webdam.inria.fr

Access control in distributed knowledgebase SIIGMOD15, ICDT16

Explaining

- **Users want to understand the information they see, the answers they are given**
  - In their professional/social life
- Difficulties
  - Reasoning with large number of facts
  - Information is often probabilistic and not public
  - Requires knowing how the information was obtained (its *provenance*)

Serendipity



- You may hear by chance a song that is going to totally obsess you
- A librarian may suggest your reading a book that will change your life

This is serendipity

- A perfect search engine
- A perfect recommendation system
- A perfect computer assistant

Such systems are boring

They lack serendipity

Design programs that would help **introduce serendipity** in our lives

Hypermnesia

Exceptionally exact or vivid memory, especially as associated with certain mental illnesses

For a user: We cannot live knowing that any word, any move will leave a trace?

For the ecosystem: We cannot store all the data we produce – lack of storage resources

Forgetting is Key to a Healthy Mind
Scientific American
Image: Aaron Goodman

A main issue is to select the information we choose to delete

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