Personal information management systems and knowledge integration

Serge Abiteboul
Inria & Ecole Normale Supérieure Cachan

serge.abiteboul@inria.fr
http://abiteboul.com

Organization

1. Personal data
2. The Pims
    1. The concept of Pims
    2. The Pims are arriving and that is cool
3. Research issues
4. An illustration with the Thymeflow system

1. Personal data

Personal data out there

Personal data out there

- Variety
  - Structured, semi-structured, unstructured
  - Metadata and knowledge (RDF)
  - Different languages, terminologies, ontologies, structures
- Veracity
  - Varying quality: errors, opinions, missing data…
  - Varying importance: hard to assess
- Velocity
  - Changes, staleness…
  - Recent data is typically very valuable
- – Volume (???)
  - Growing but no Big data
- \+ Distributed
  - In many autonomous systems that act as silos
  - Different systems, protocols

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

Bad news (2)

Google

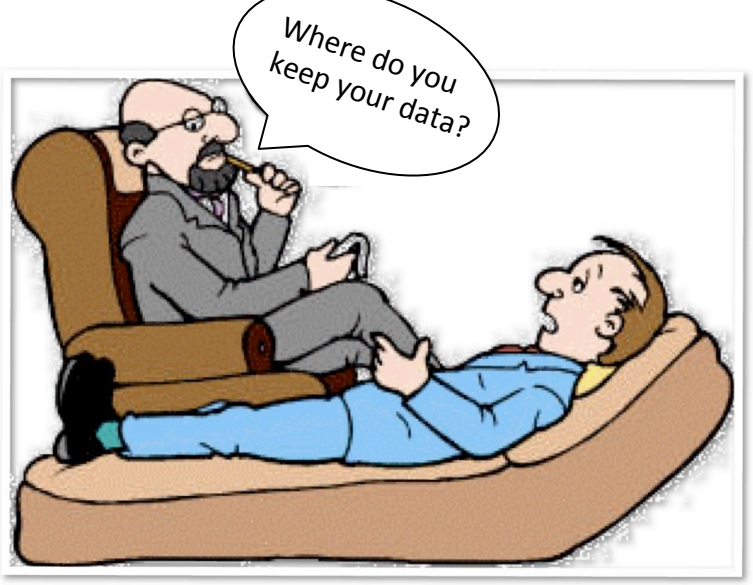
- A few companies concentrate most of the world's data and analytic power
  - They have the means to destroy business competition in large portions of the economy
- A few companies control all your personal data
  - They determine what information you are exposed to
  - They guide many of your decisions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

amazon

2.The Pims

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

Alternatives



- Continue with this increasing mess
  - See a shrink to overcome the frustration
- Gather all your data in one platform
  - Google, Apple, Facebook, …, a new comer
  - See a shrink to overcome resentment
- Study 2 years to become a geek
  - Geeks know how to manage their information
  - See a shrink to survive the experience

Or move to Pims!

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*

One Pims, two Pims… many Pims

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

The Pims are (I believe) arriving!

Why?

For 3 kinds of reasons:

- 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 alternatives to proprietary online services are increasingly available
- 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 devices deployed at home or in datacenters
  - Internet service 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 things

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 services freely & they can leave a service
- Participation in a more "neutral" Web
  - With the "network effect", 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

The Pims will primarily arrive because of new functionalities

This is (for me) the key ingredient for adoption

New functionalities ➳ New opportunities

New playing field for startups

New playing field for researchers

3. Research issues with the Pims

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

R&D issues we will not consider much

Some old problems revisited

- Epsilon-principle (epsilon-user-administration)
- Backups & Task sequencing
- Access control & Exchange of information
- Security (e.g. works @ INRIA Rocquencourt)
- Connected objects control

R&D issues we will briefly illustrate

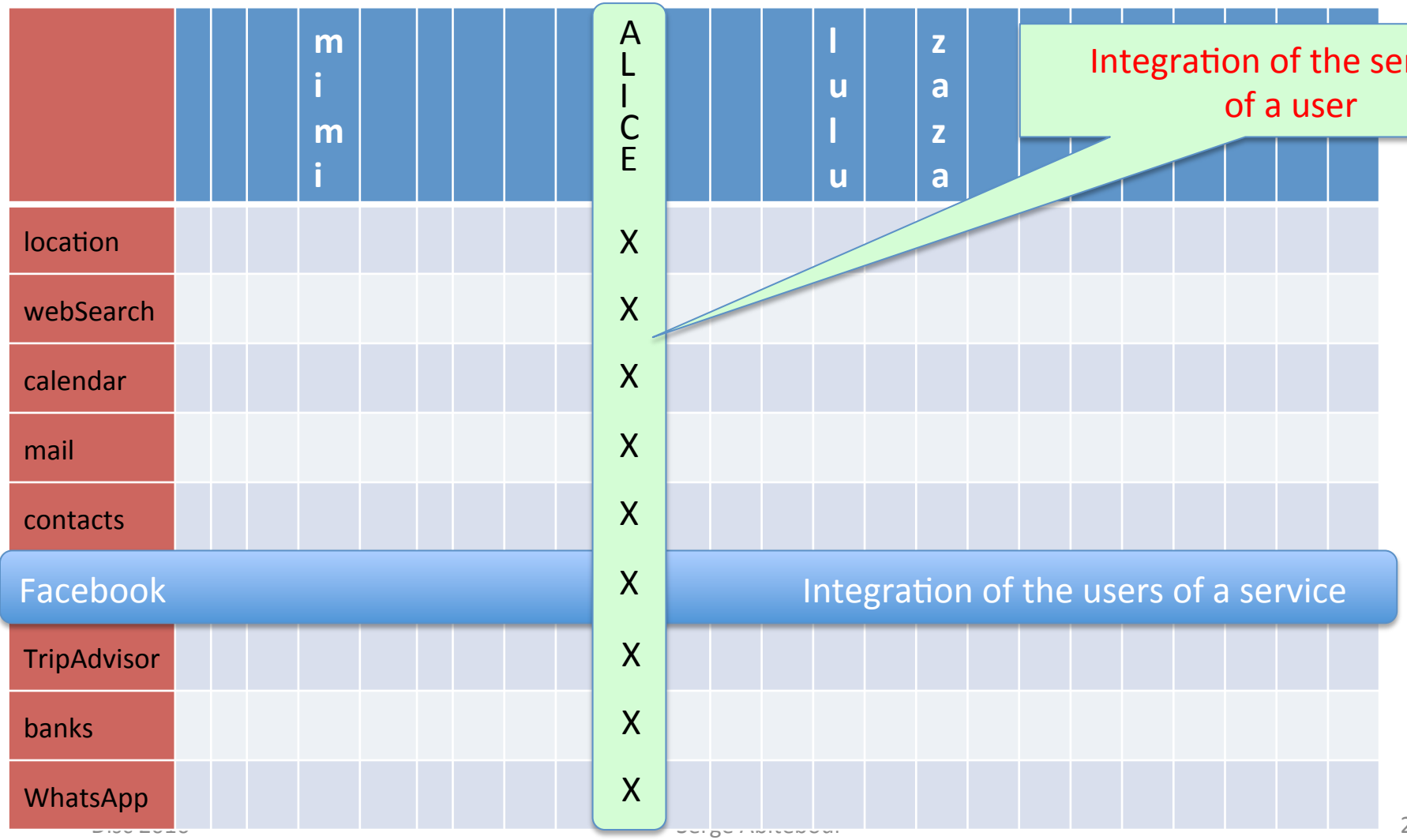
Some old problems revisited

- Personal information integration
- Synchronization
- Personalization and context awareness
- Personal data analysis

4. An illustration with the Thymeflow system

Demo in International Conference on Information and Knowledge Management (CIKM'16) with David Montoya, Thomas Pellissier-Tanon, Fabian M. Suchanek

Pims are first about data integration



Or rather on knowledge integration

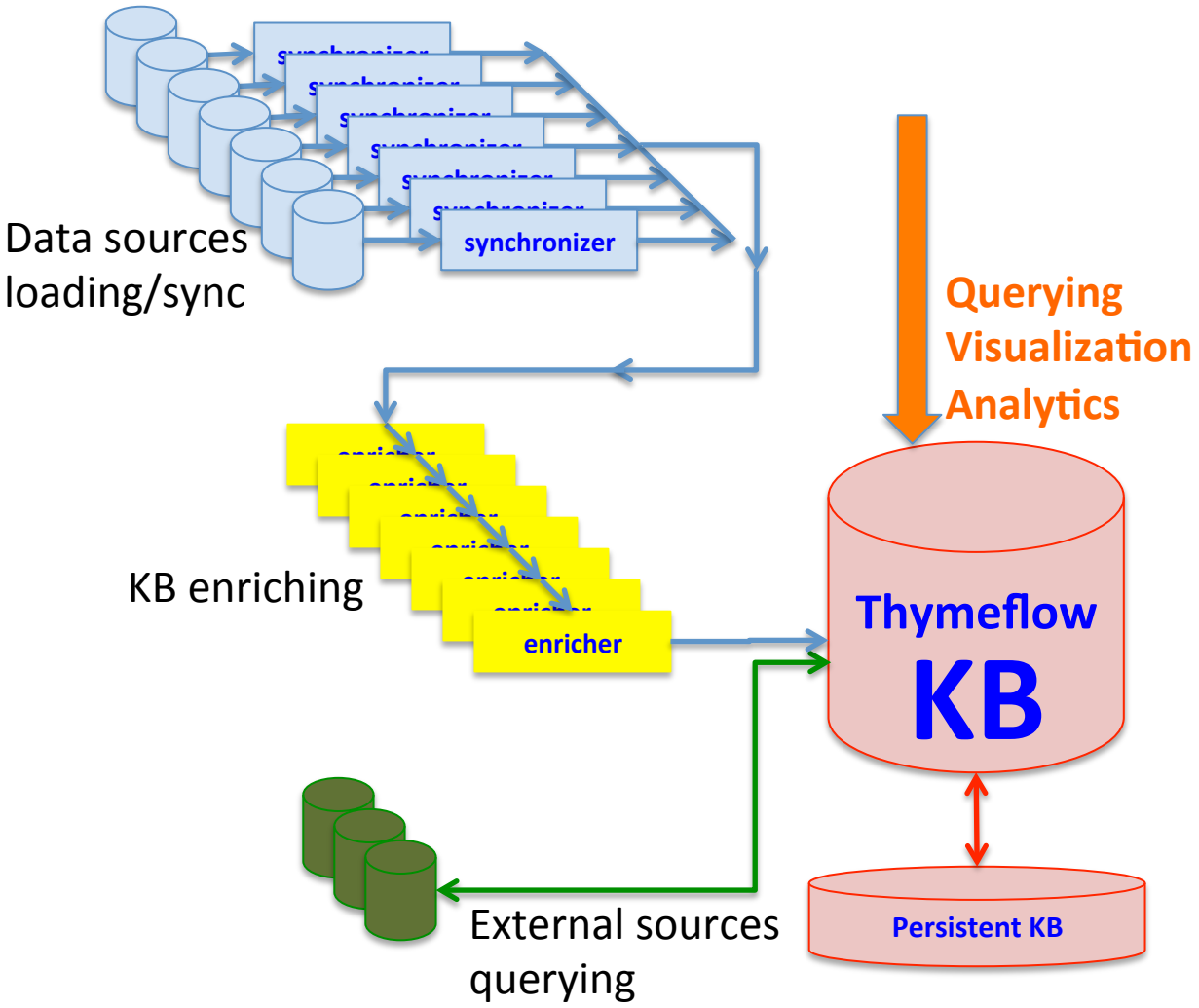
- **Data / Information ➳ Knowledge**
  - Personal data/info management is getting too complicated
  - Machines prefer structured knowledge to unstructured information or semantic-free data
- Thesis: Let us turn all our information into a distributed knowledge base

ERC Webdam, http://webdam.inria.fr (ended in 2015)

The Thymeflow Knowledge Base

- Thymeflow is a KB, an extension of a person's memory
  - Episodical memory (typically related to spatio-temporal events) and
  - Semantic memory (knowledge that holds irrelative to any such event)
- Thymeflow's knowledge is
  - Extracted from all the information traces of the person
  - Obtained from the Web (Wikidata, OpenStreetMap…)
  - Derived by software modules that analyze the KB
- Thymeflow is an application for the Web and mobile phones
  - Loading: calendar, contacts, mails, geolocation (GPS), social networks…
  - Deriving links between these data sources and other knowledge bases
  - Supporting query processing and data analytics

Architecture

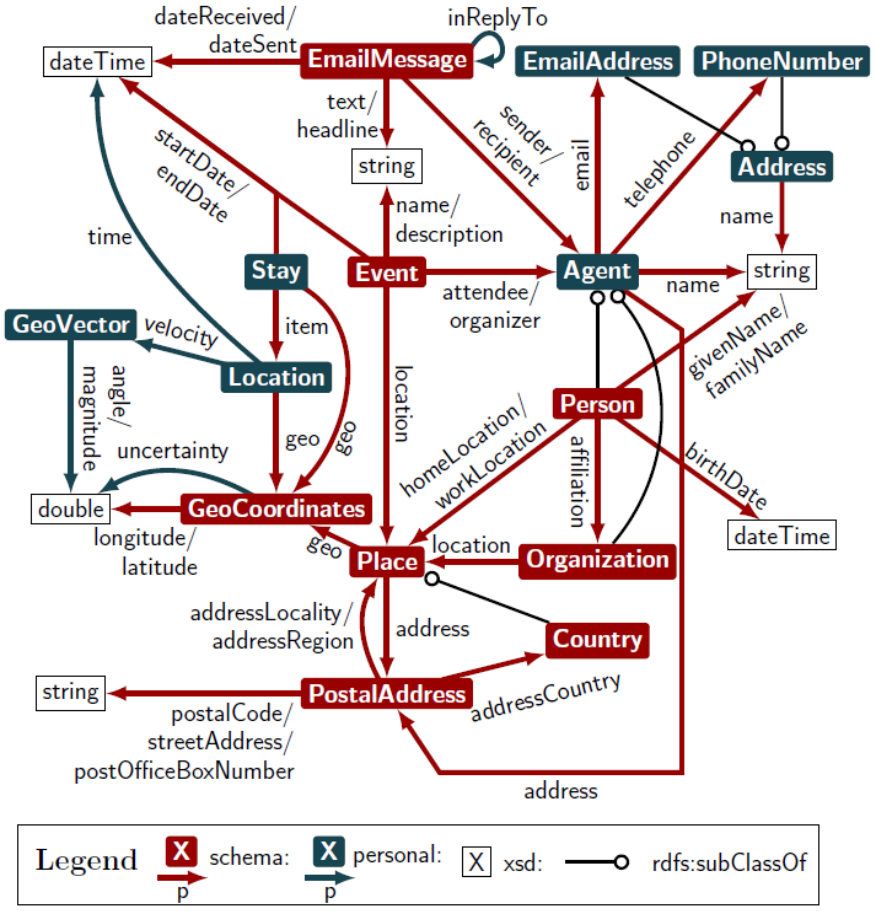


- Backend:
  - HTTP server
  - REST API
  - SPARQL endpoint (Sesame)
- Frontend: Web app
- Mobile app
  - for geolocation

RDF knowledge base

- RDF model
  - RDF Triples

    subject–predicate–object
- Schema
  - http://schema.org/
  - http://thymeflow.com/personal
- Most useful classes
  - `personal:Agent`
  - `schema:Event`
  - `schema:Place`
  - `schema:EmailMessage`



Query examples

- At what time do I usually send emails?

```
PREFIX schema: <http://schema.org/>
PREFIX personal: <http://thymeflow.com/personal#>
SELECT ?hour (COUNT(DISTINCT ?email) AS ?count) WHERE {
  ?email a schema:EmailMessage .
  ?email schema:sender/personal:sameAs*/schema:email <mailto:{MY_ADDRESS}> .
  ?email schema:dateSent ?dateEmail .
  BIND(HOURS(?dateEmail) AS ?hour)
} GROUP BY ?hour ORDER BY ?hour
```

- Full-text query in my entire memory

```
PREFIX search: <http://www.openrdf.org/contrib/lucenesail#>
SELECT ?subj ?prop ?score ?snippet WHERE {
  ?subj search:matches [
     search:query "{TEXT QUERY}";
     search:property ?prop;
     search:score ?score;
     search:snippet ?snippet
  ]
} ORDER BY DESC(?score) LIMIT 10
```

Main component: synchronizer

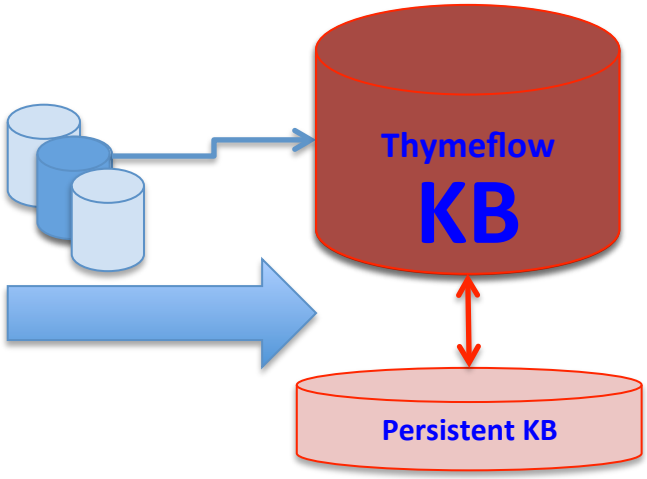
- Transform data into knowledge and synchronize a data source with the knowledge base

Examples

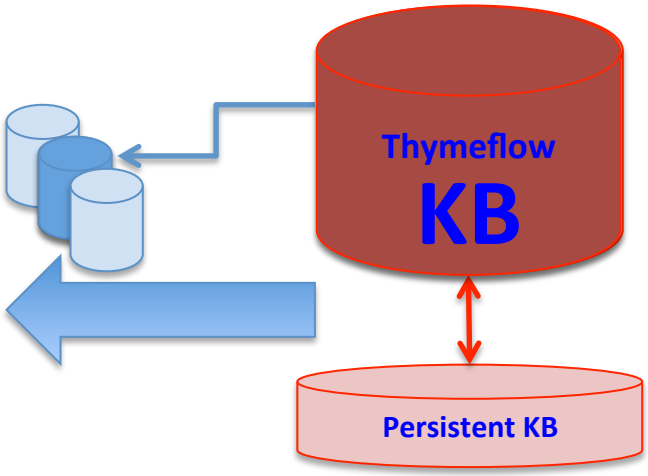
- CalDavSynchronizer/CardDavSynchronizer :
  - Manage iCalendar (.ical) and vCard (.vcf)
- EmailSynchronizer
  - IMAP to connect to mail servers

Update propagation

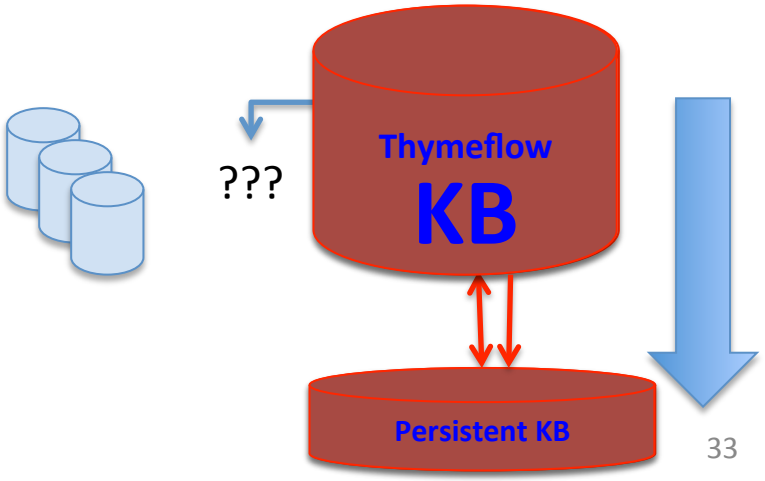
From data sources to KB



From KB to data sources (1)



From KB to data sources (1)



Main component: enricher

- Align concepts coming from different data sources
- Add knowledge to the KB

Examples

- Align agents based on, e.g., their names, emails…
- Add geolocations to calendar events
- Add semantics to places physically visited
- Align calendar events to places physically visited

Data analytics

- Small data analysis with Pims
  - Learn from personal data, e.g.,
    - Personal health and well-being
    - Digital personal assistant: notification & planning
  - Issues
    - Much smaller amounts of data – statistics harder
    - Varying data quality: imprecision, inconsistencies
- Big data analysis from Pims
  - Aggregate data from large number of Pims
  - Derive knowledge useful for Pims, e.g., traffic jams
  - Issue: data privacy

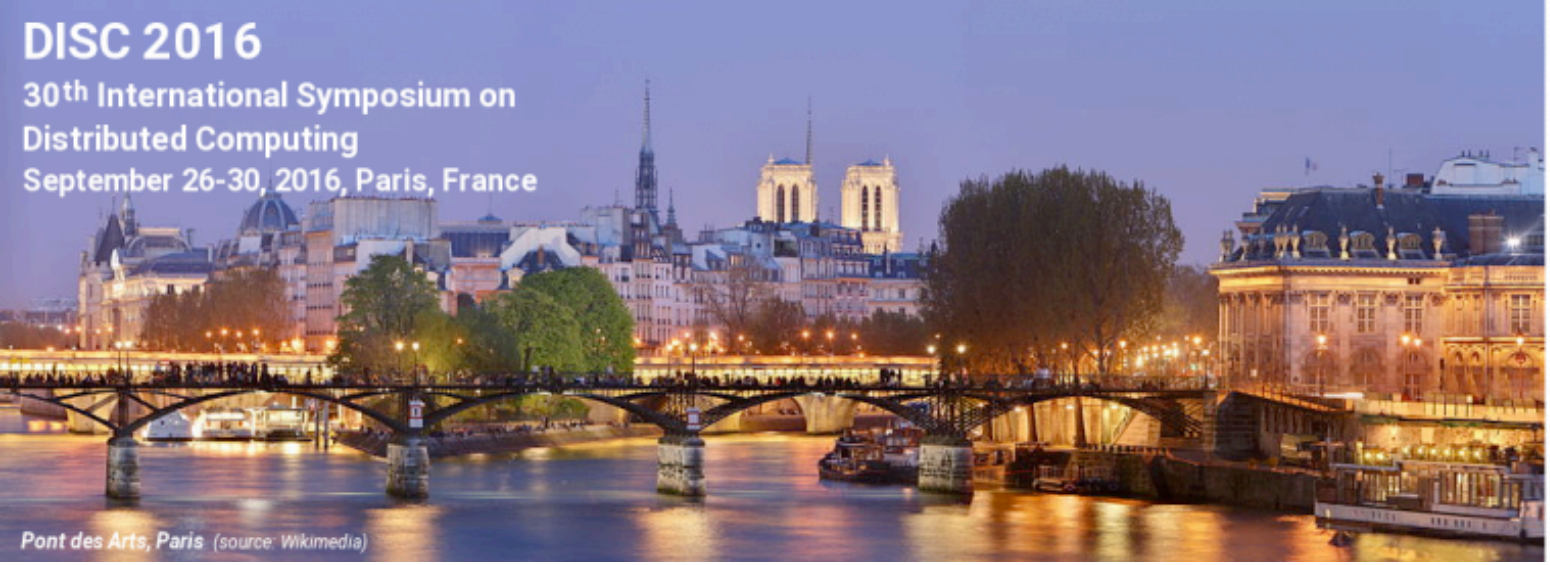
Conclusion

Goal

Make the digital world a better place to live in

The Pims seem a promising direction for that

Lots of research issues remaining



DISC 2016
Paris, France

informatiques mathématiques
Inria
ENS
CACHAN