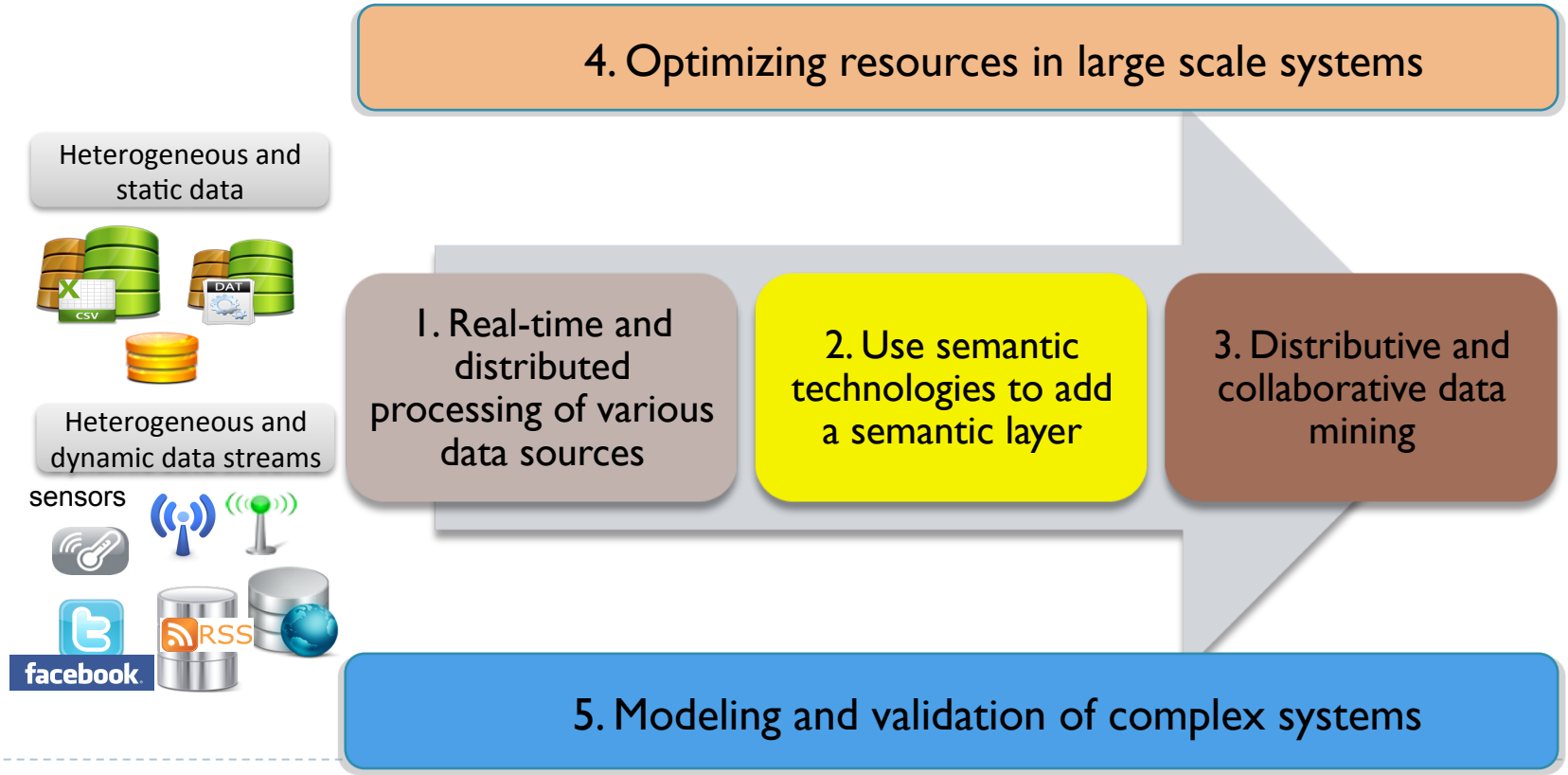
Toward Semantic Data Stream Technologies and Applications

Raja CHIKY
raja.chiky@isep.fr
2013-2014

About me

- Associate professor in Computer Science – LISITE-RDI
- Research interest: Data stream mining, scalability and resource optimization in cloud computing, recommender systems
- Research field: **Large scale data management**



Outline

- What is a data stream ?
- Data stream management systems
- New applications
- Semantic Data Stream
- Conclusion

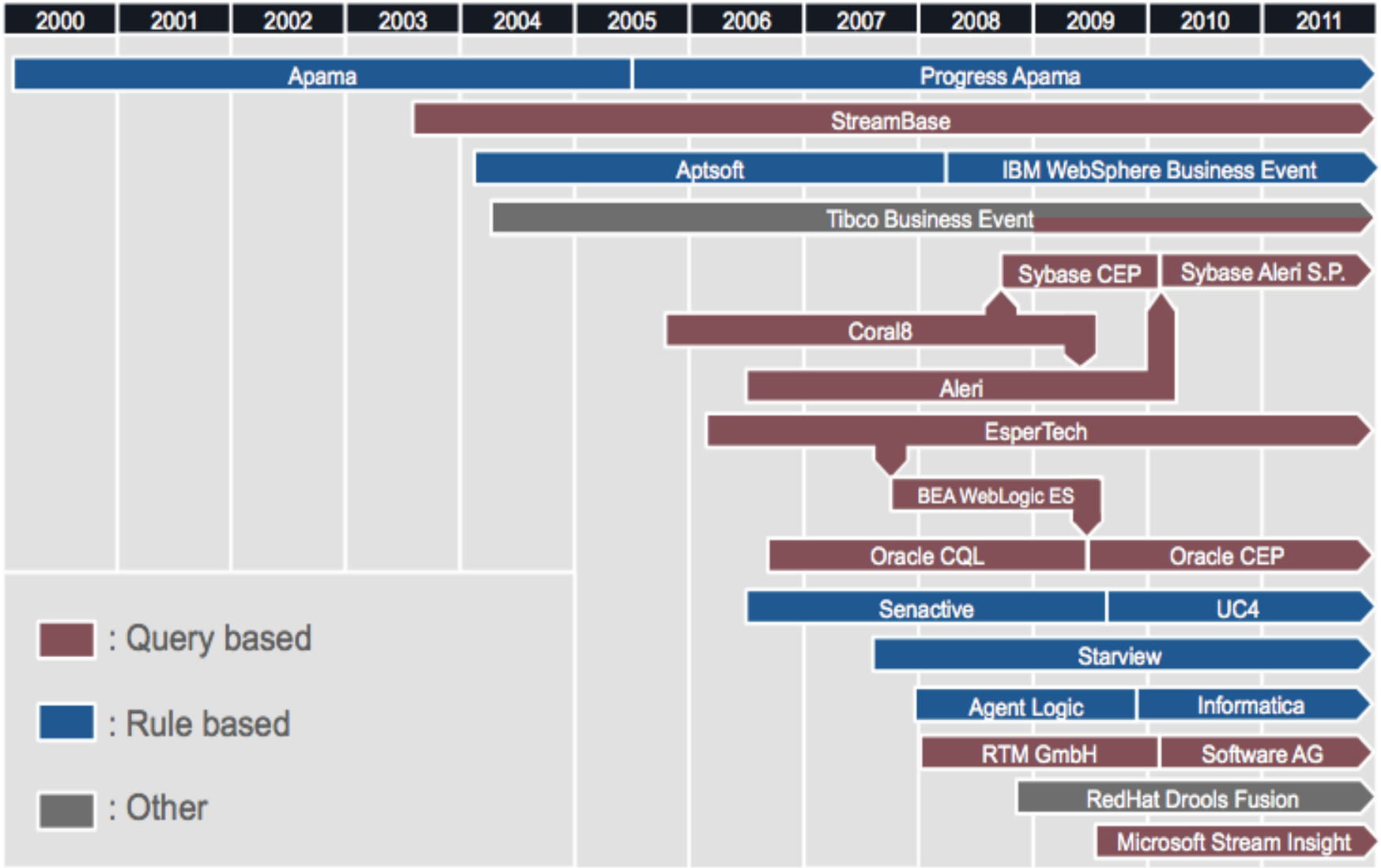
What is a data stream?

- Golab & Oszu (2003): *"A **data stream** is a **real-time**, continuous, ordered (implicitly by arrival time or explicitly by timestamp) **sequence of items**. It is impossible to control the order in which items arrive, nor is it feasible to locally **store** a stream in its entirety."*
- Massive volumes of **structured** data, items arrive at a high rate.

Data Stream Management Systems

	DBMS	DSMS
Data model	Permanent updatable relations	Streams and permanent updatable relations
Storage	Data is stored on disk	Permanent relations are stored on disk Streams are processed on the fly
Query	SQL language Creating structures Inserting/updating/deleting data Retrieving data (one-time query)	SQL-like query language Standard SQL on permanent relations Extended SQL on streams with windowing Continuous queries
Performance	Large volumes of data	Optimization of computer resources to deal with Several streams Several queries Ability to face variations in arrival rates without crash

Existing DSMS



World of data streams

Website logs

Network monitoring

Financial services

Weather forecasting

Data may come from humans, sensors or machines

New applications : 360-degree view of the customer

New applications : 360-degree view of the customer (Contd.)

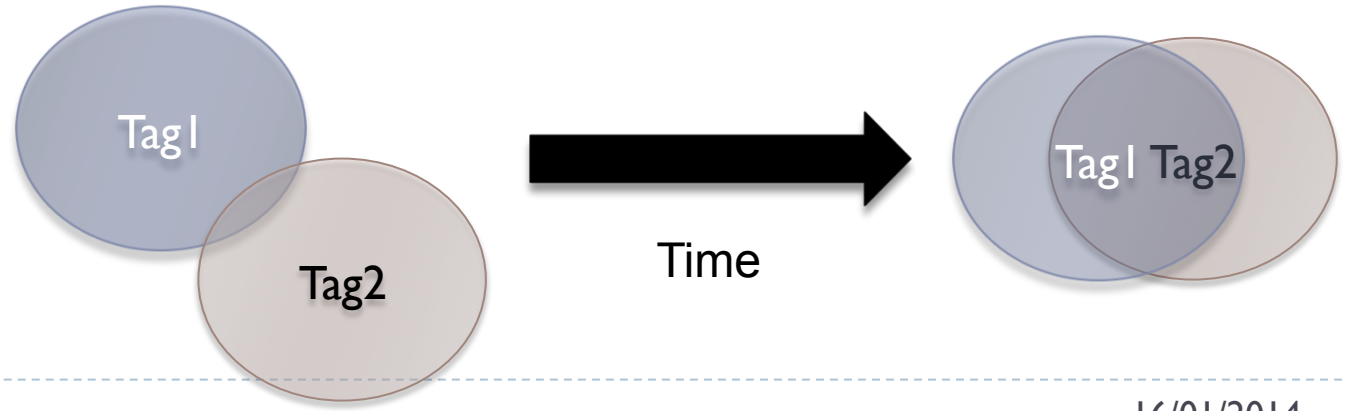
Internal and External sources

Descriptive data	Interaction data	Operational data	Behavioral data	Contextual data
• Profile • Contractual data	• Multi-chanel • Email/chat/ call center	• Orders • Transactions • Payment history	• Web clicks • Usage history • Social media • Opinions	• Geographical data • Devices

New applications: Social networks

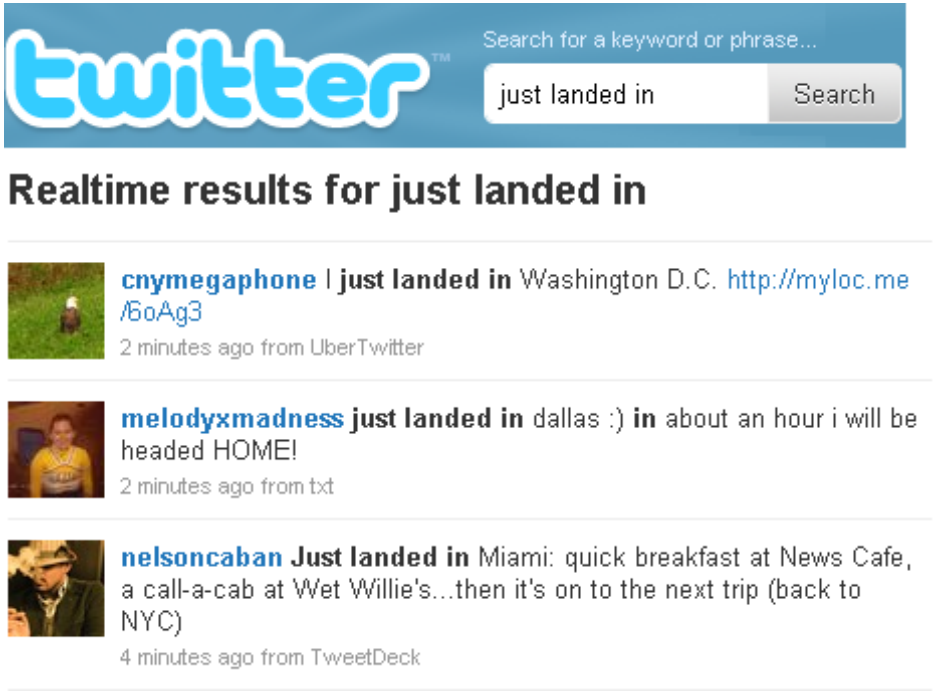
- Twitter: A massive Data stream
  - Free social networking and micro-blogging service
  - 3 billion requests per day

  - Sentiment Analysis
  - Tracking trending hashtags
  - Detect breaking news
  - Analyze Marketing campaigns



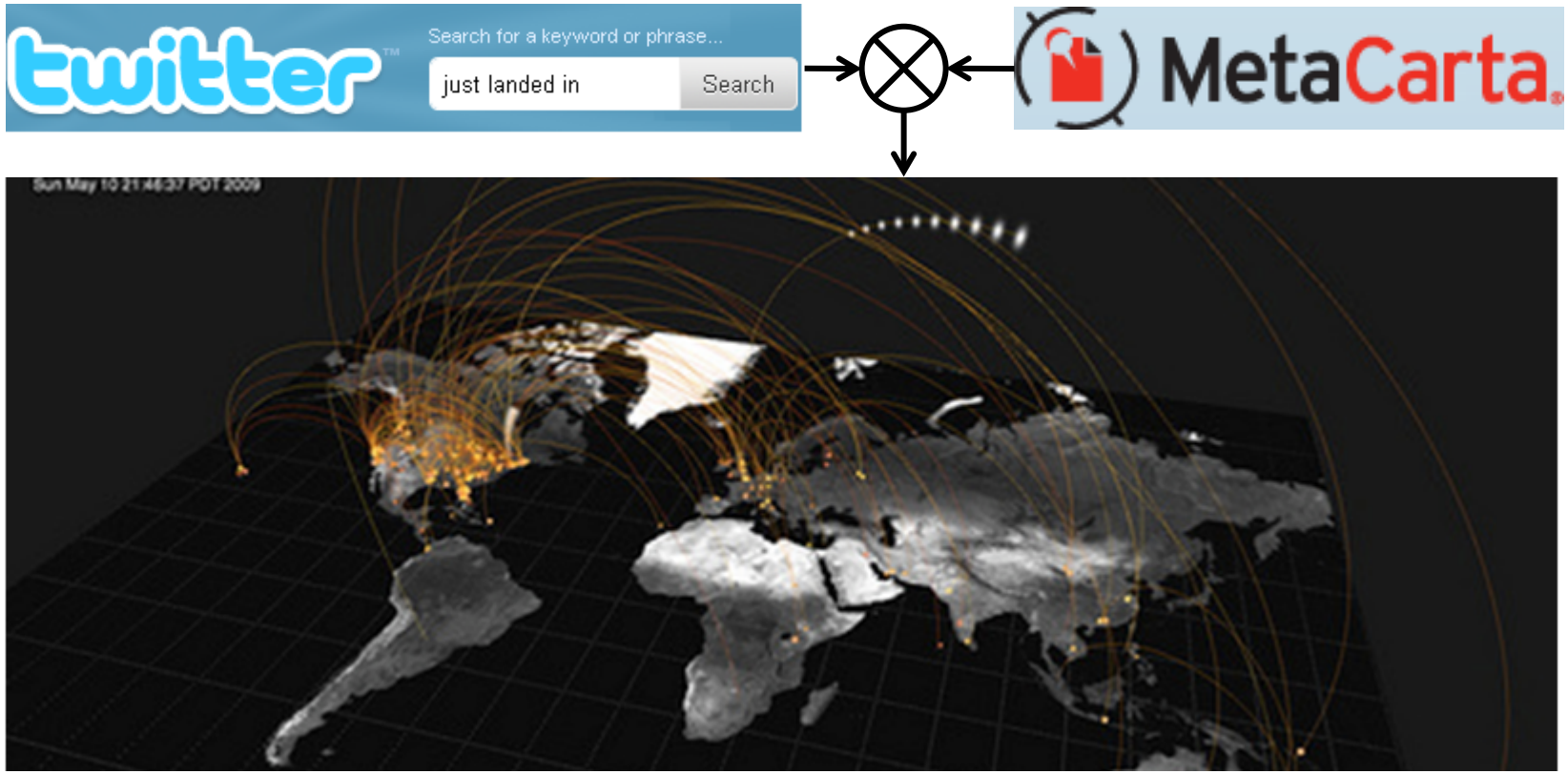
More applications on Twitter

- E.g. http://twitter.com/#search?q=just%20landed%20in



Combination of stream and static data

- E.g., Twitter + MetaCarta

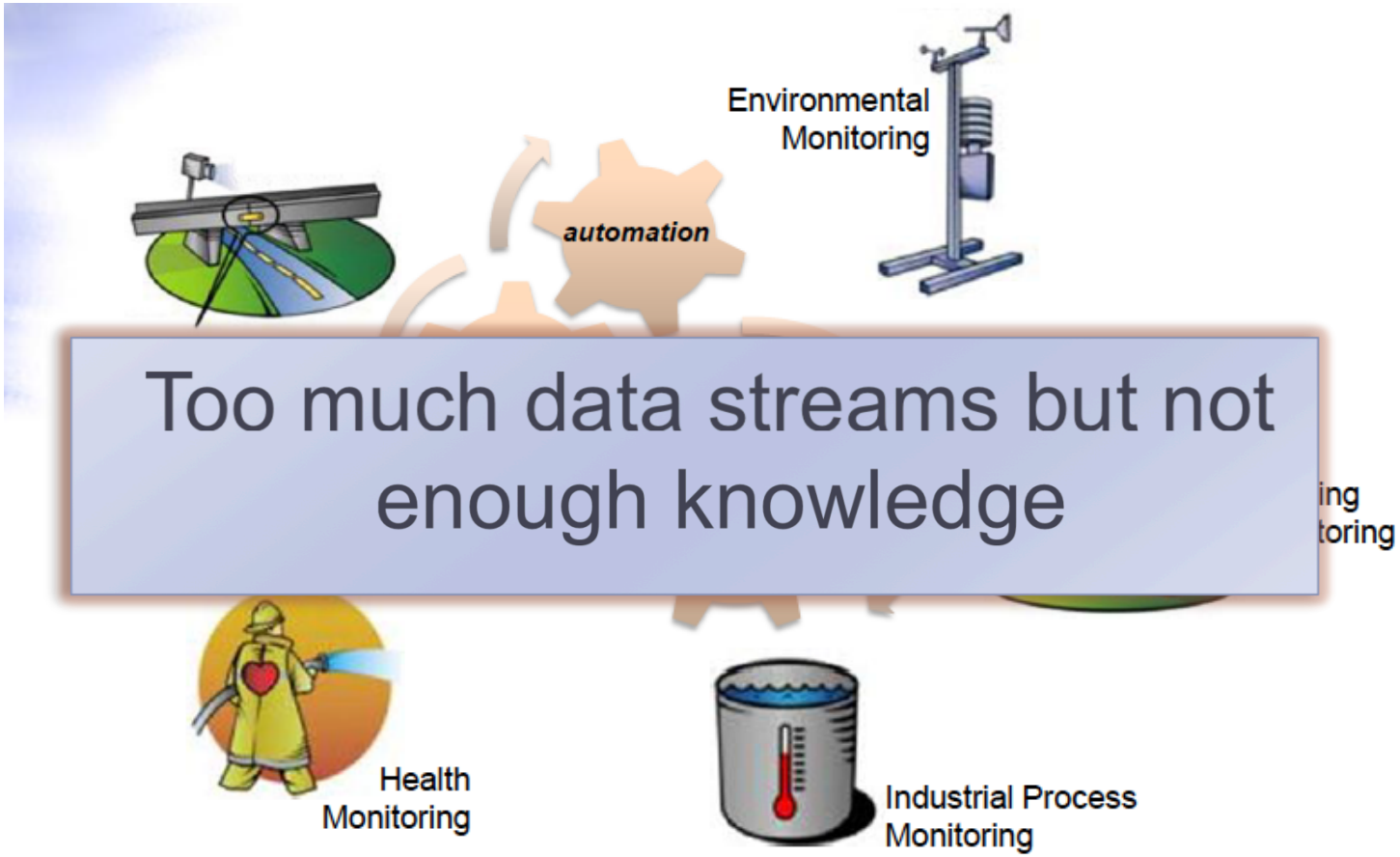


[source: http://blog.blprnt.com/blog/blprnt/just-landed-processing-twitter-metacarta-hidden-data]

What have you seen?

- BOTTARI: An augmented reality application for personalized **recommendation of restaurants** in Seoul

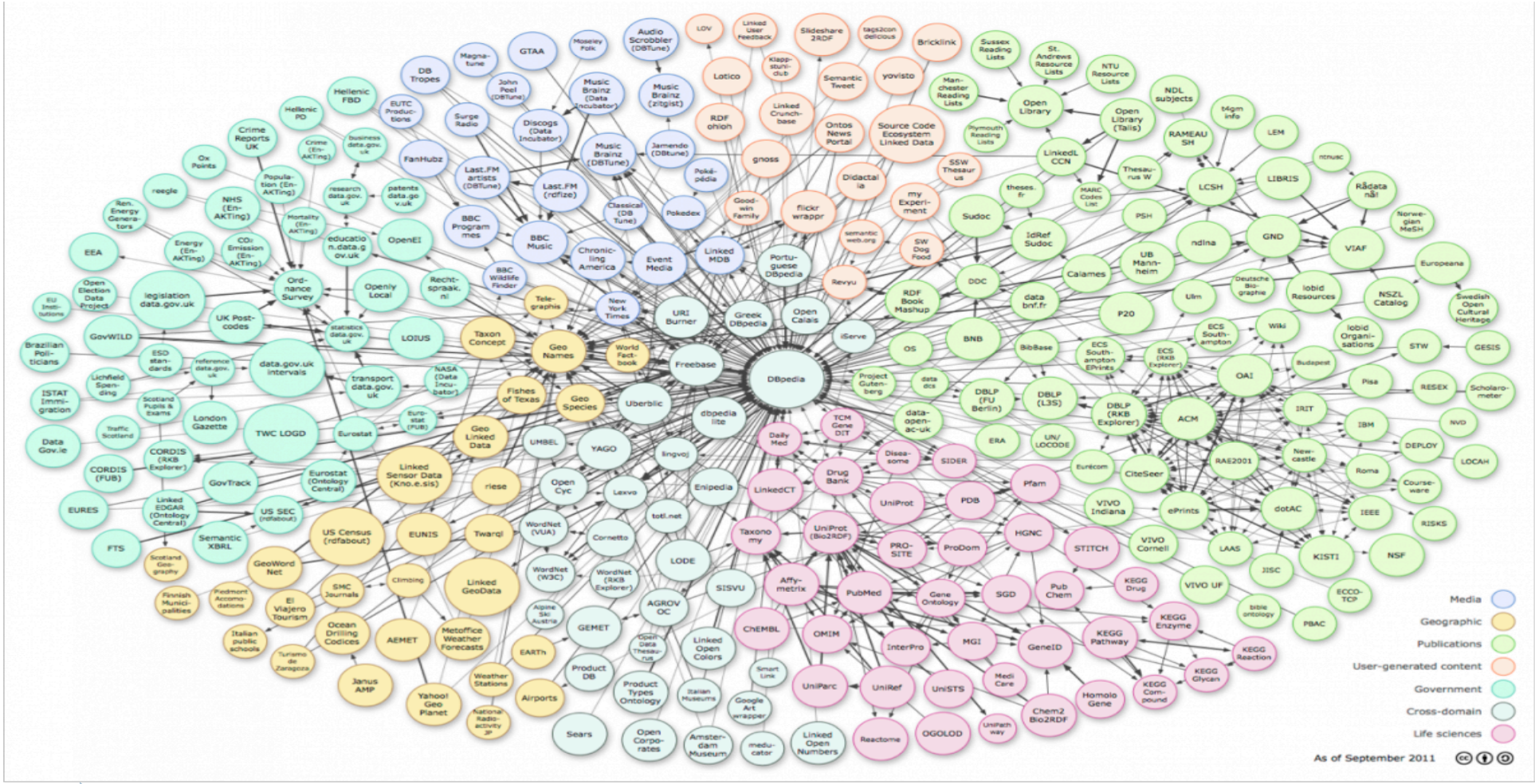
Too much data streams



Semantic Web, Linked Data

- Semantic Web
  - Collaborative movement to promote common data formats on the World Wide Web
  - Inclusion of semantic content in web pages
  - From unstructured and semi-structured documents to a « Web of data »
- Linked Data
  - Best practices to represent, publish, link data on the Semantic Web
  - Linked Data Cloud: collection of datasets that have been published in Linked data format

Linked data


DBpedia
Freebase
Geo Names
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

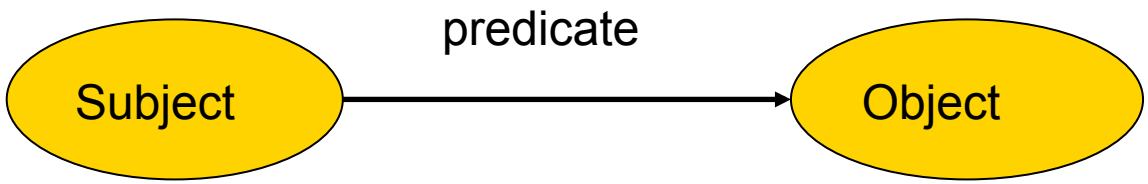
Semantic Web technologies for data stream

- Annotate stream data **with semantic metadata**
- Apply **Linked Data principles** to publish streaming data
- Interlink streaming data with **existing** datasets
- Integrate data stream processing + **reasoning**

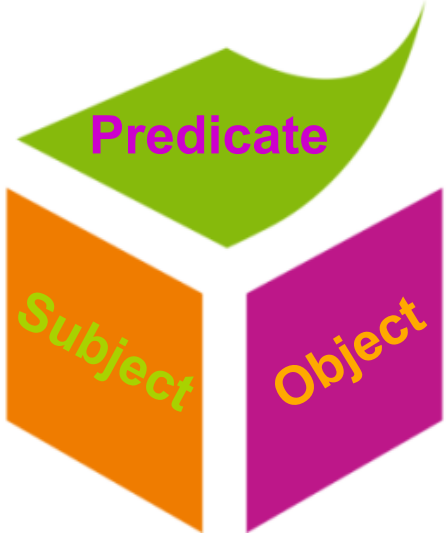
- **Objectives : interoperability, automation, enrichment**

Existing prototypes

Introduction to RDF



- Statement: "The author of http://www.w3schools.com/rdf is Jan Egil Refsnes"
  - The subject of the statement above is: http://www.w3schools.com/rdf
  - The predicate is: author
  - The object is: Jan Egil Refsnes



RDF Stream

- RDF Stream is defined as an ordered sequence of pair
- Each pair consists of a RDF triplet and a timestamp

$$(\langle subj_i\ ,\ pred_i\ ,\ obj_i\rangle,\ T_i)$$
$$(\langle subj_{i+1}\ ,\ pred_{i+1}\ ,\ obj_{i+1}\rangle,\ T_{i+1}) \qquad (T_i \leq T_{i+1})$$

- Example

$$(\langle\ :traveller1\ :justLanded\ :placeA\ \rangle,\ T_1)$$
$$(\langle\ :traveller2\ :justLanded\ :placeB\ \rangle,\ T_1)$$
$$(\langle\ :traveller3\ :justLanded\ :placeA\ \rangle,\ T_2)$$
$$(\langle\ :traveller1\ :justLanded\ :placeC\ \rangle,\ T_3)$$

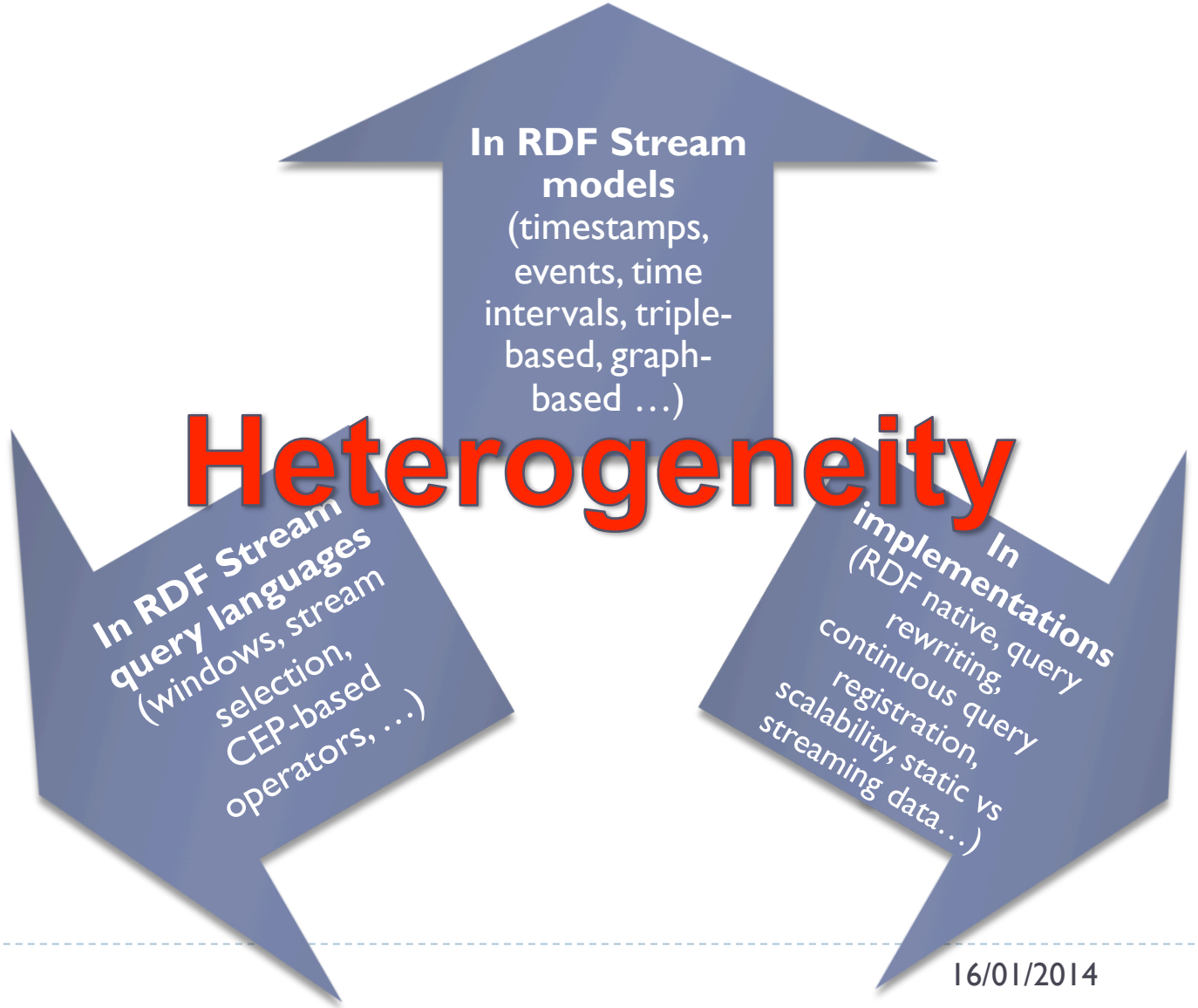
RDF Stream – Several definitions

- Several possibilities:
  - An RDF stream is an infinite sequence of timestamped events (triples or graphs), …

$$<<s,p,o>_i, t_i>$$
$$<<s,p,o>_{i+1}, t_{i+1}>$$
$$<<s,p,o>_{i+2}, t_{i+2}>$$
$$\ldots$$

  - An RDF stream is an infinite sequence of triple occurrences $<<s,p,o>,t_\alpha,t_\omega>$ where $<s,p,o>$ is an RDF triple and t_α and t_ω are the start and end of the interval

Approaches


In RDF Stream models
(timestamps, events, time intervals, triple-based, graph-based …)
Heterogeneity
In RDF Stream query languages
(windows, stream selection, CEP-based operators, …)
In implementations
(RDF native, query rewriting, continuous query registration, scalability, static vs streaming data…)

Example: SRBench

Q: Detect if a hurricane has been observed

"A hurricane has a sustained wind (for more than 3 hours) of at least 33 metres per second or 74 miles per hour (119 km/h)."

```
ASK
WHERE {
STREAM <http://www.cwi.nl/SRBench/observations> [RANGE 10800s SLIDE 600s]
    {?observation om-owl:procedure ?sensor ;
                  om-owl:observedProperty weather:WindSpeed ;
                  om-owl:result [ om-owl:floatValue ?value ] .}
 }
GROUP BY ?sensor
HAVING ( AVG(?value) >= "74"^^xsd:float )
```

CQELS

```
ASK
FROM STREAM <http://www.cwi.nl/SRBench/observations> [RANGE 3h STEP 10m]
WHERE {
   ?observation om-owl:procedure ?sensor ;
                om-owl:observedProperty weather:WindSpeed ;
                om-owl:result [ om-owl:floatValue ?value ] . }
GROUP BY ?sensor
HAVING ( AVG(?value) >= "74"^^xsd:float )
```

C-SPARQL

W3C RDF Stream Processing Community

- http://www.w3.org/community/rsp/

W3C Community and Business Groups

Search blogs

CURRENT GROUPS REPORTS ABOUT

Mailing List
Wiki
Chat
RSS
Contact Group

Home / RDF Stream Processing...

RDF Stream Processing Community Group

The mission of the RDF Stream Processing Community Group (RSP) is to define a common model for producing, transmitting and continuously querying RDF Streams. This includes extensions to both RDF and SPARQL for representing streaming data, as well as their semantics. Moreover this work envisions an ecosystem of streaming and static RDF data sources whose data can be combined through standard models, languages and protocols. Complementary to related work in the area of databases, this Community Group looks at the dynamic properties of graph-based data, i.e., graphs that are produced over time and which may change their shape and data over time.

Reports

No reports yet published. The Chair is responsible for publishing reports. More about publishing...

News

Get involved!

Anyone may join this Community Group. All participants in this group have signed the W3C Community Contributor License Agreement (CLA).

JOIN THIS GROUP

or learn how to join or request an account.

Note: Community Groups are proposed and run by the community. Although W3C hosts these conversations, the groups do not necessarily represent the views of the W3C Membership or staff.

W3C RSP Community Group mission

"The mission of the RDF Stream Processing Community Group (RSP) is to define a common model for producing, transmitting and continuously querying RDF Streams. This includes extensions to both RDF and SPARQL for representing streaming data, as well as their semantics. Moreover this work envisions an ecosystem of streaming and static RDF data sources whose data can be combined through standard models, languages and protocols (...)"

Conclusion: Research Challenges

- Benchmarking: LSBench, SRBench
- Data Quality: Dealing with incomplete and noisy data
- Reasoning on streams: formal semantics, efficiency, scalability
- Distributed and parallel processing: data stream sources are distributed in nature
- Scalability: approximation, load shedding, etc.

Thanks for your attention

raja.chiky@isep.fr