Benchmarking Linked Open Data technology

SRbench: A Benchmark for Streaming RDF Storage Engines

Ying Zhang, ***Peter Boncz*** (CWI, Amsterdam)

What is Database Benchmarking?

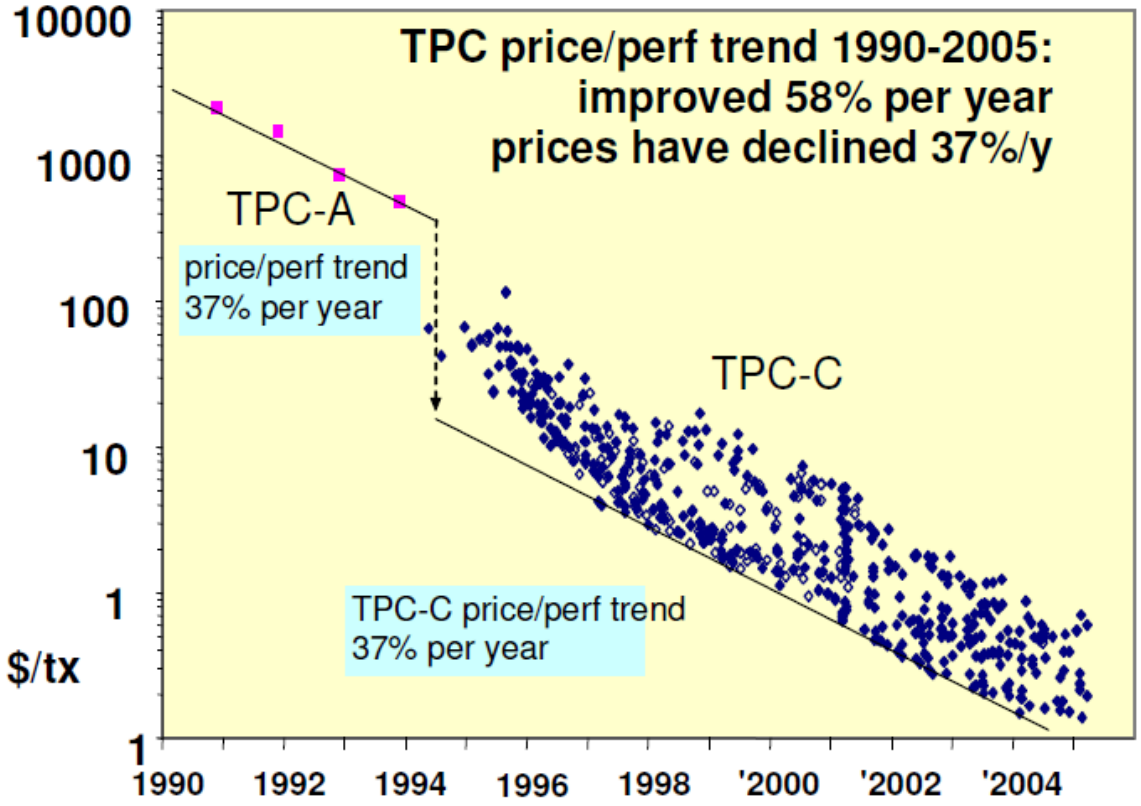
Standard test to measure and understand how technology performs

- Dataset definition
  - at various scales (100GB, 300GB, 1TB, 3TB, etc)
  - mimicks a recognizable relevant **usage scenario**
- Database Queries
  - often between 10-100 queries, with parameters
  - + rules/programs that specify how these queries are posed
- Result Metrics
  - a number to understand the result
  - tps = "transactions/second"

    $/QphH@size = "price per query per hour"
- Audit Rules
  - allow results to be checked by independent auditors
  - prevent/limit cheating

Why Benchmarking?

- make competing products comparable
- accelerate progress, make technology viable

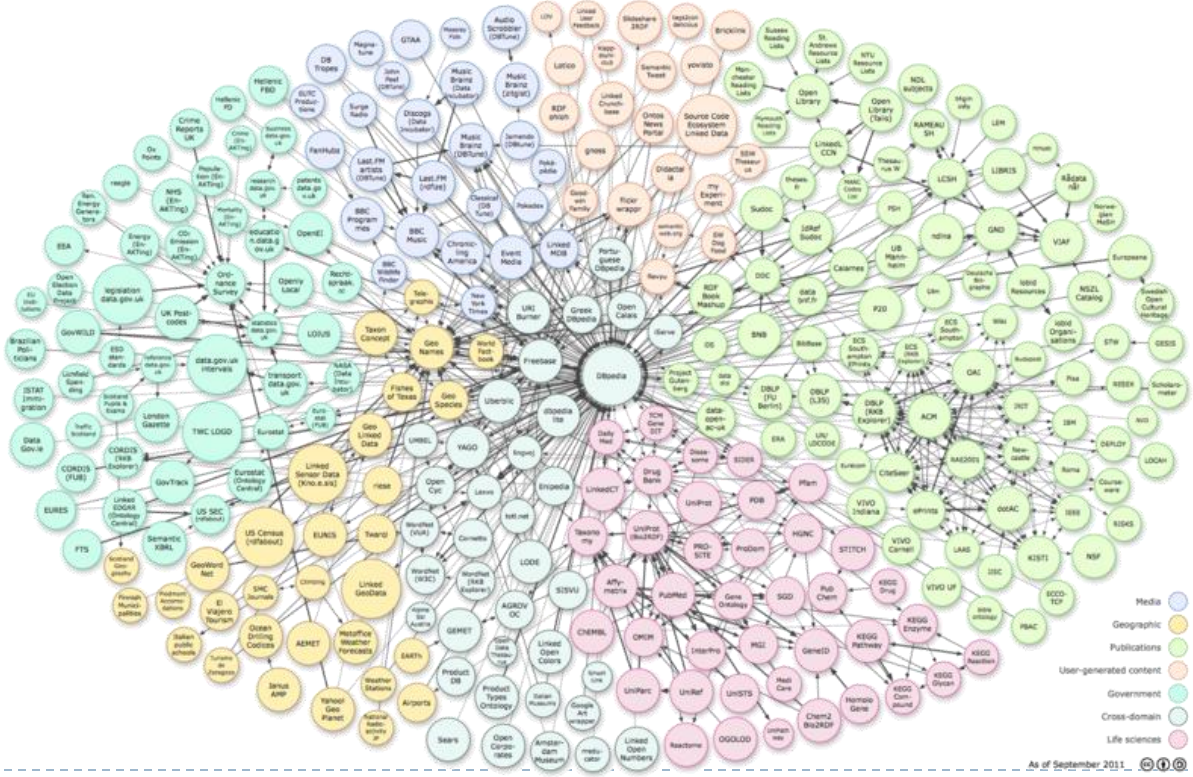


© Jim Gray, 2005

Benchmarking LOD Technology

LOD = Linked Open Data

- web addressable data ➔RDF data format (W3C®)
- lots of useful data on the web ("LOD cloud")



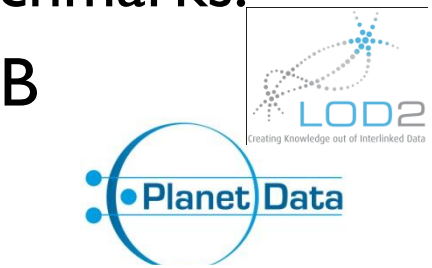
Benchmarking LOD Technology

LOD = Linked Open Data

- web addressable data ➔RDF data format (W3C®)
- lots of useful data on the web ("LOD cloud")

LOD technology (SPARQL) benchmarks:

- BSBM, DBpedia Benchmark, SIB
- SRbench ← topic of this talk
- New industry cooperation:



* tentative/expected project

LDBC: FP7 2012-2015

vendor cooperation to establish accepted RDF/Graph database benchmarks and benchmark results

LDBC Goals

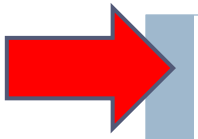
1. Create the LDBC Foundation of graph and RDF DB vendors
2. Equip de LDBC Foundation with a good initial set of benchmarks, and benchmark results

LOD2
Creating Knowledge out of Interlinked Data

spin-off

universität innsbruck
vrije Universiteit amsterdam
TECHNISCHE UNIVERSITÄT MÜNCHEN
UPC
neotechnology
ontotext
OPENLINK SOFTWARE
FORTH-ICS

Benchmarking Linked Open Data technology

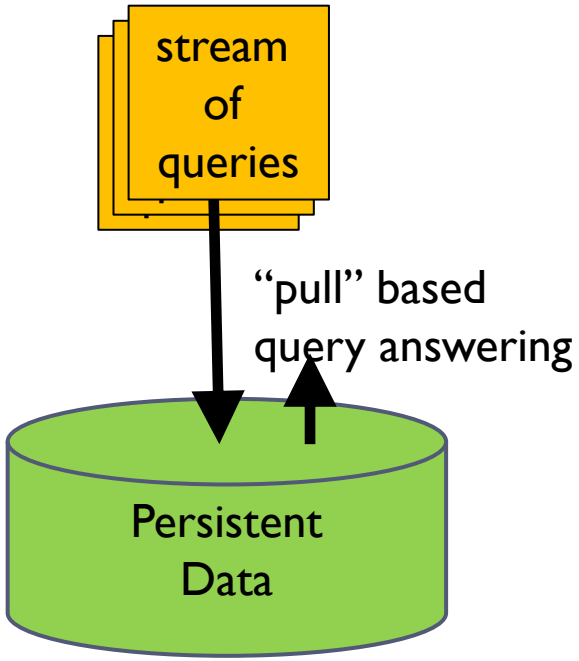


SRbench: A Benchmark for Streaming RDF Storage Engines

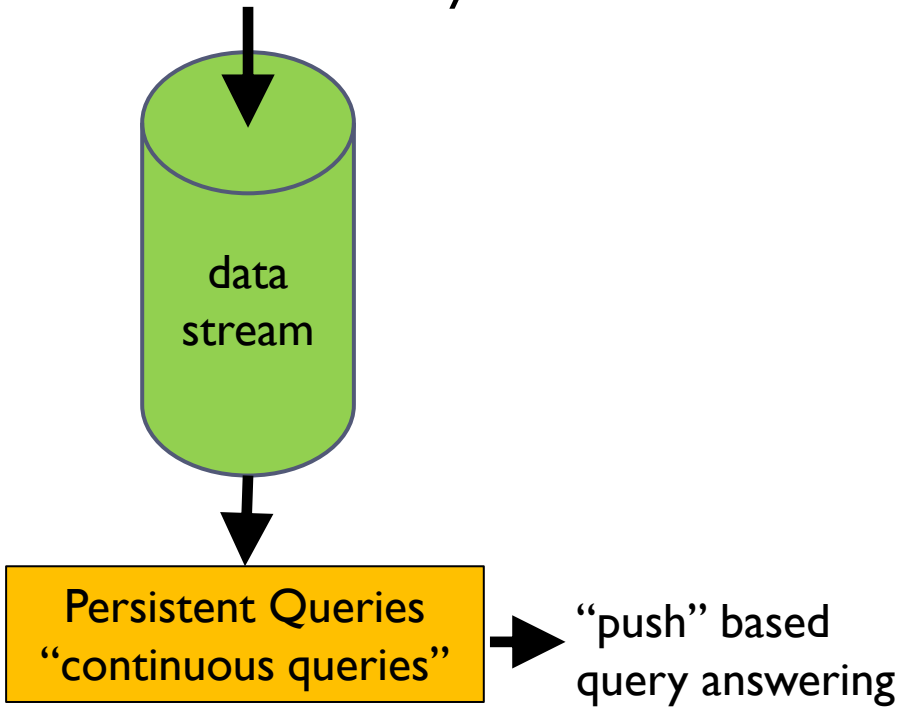
Ying Zhang, ***Peter Boncz*** (CWI, Amsterdam)

SRbench: **Streaming** RDF Benchmark

Traditional Database System vs.



Stream Database System



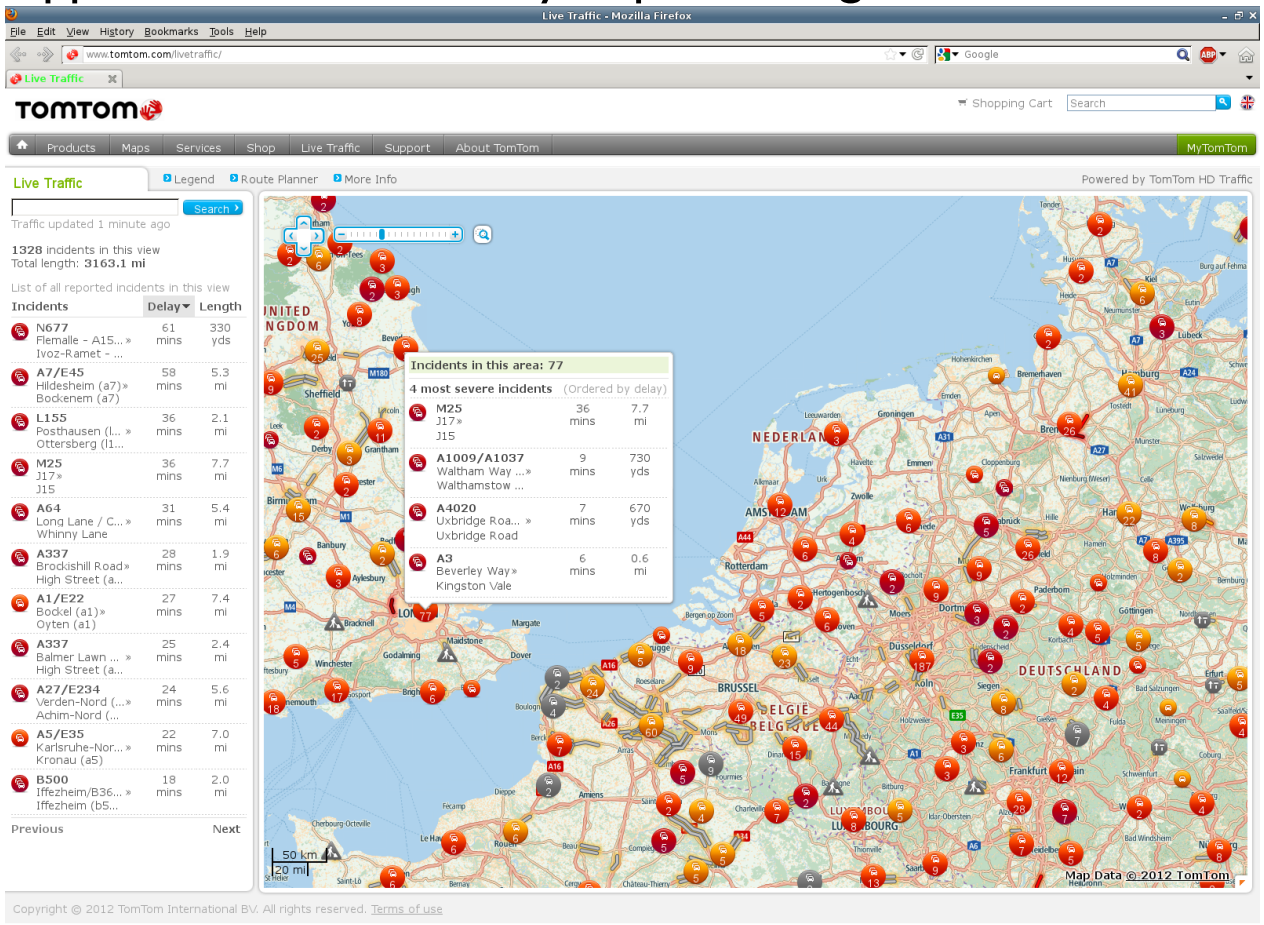
Data Streams (1/4): Stock Market

Data Streams (2/4): Social Chatter

- Detect breaking news
- Analyze Marketing campaigns

Data Streams (3/4): Car Traffic

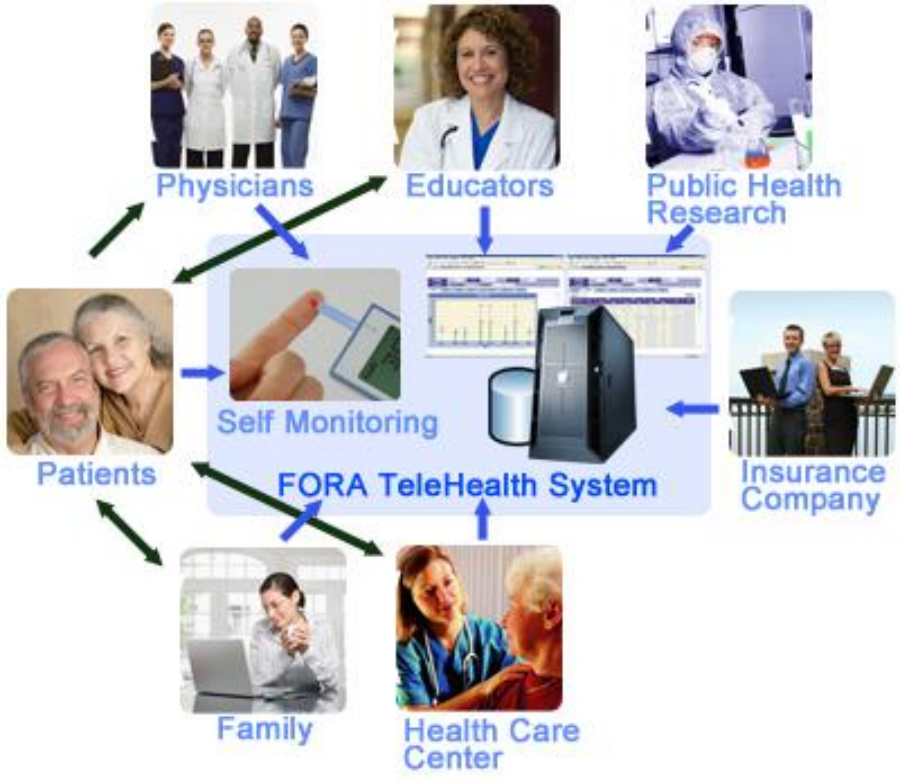
- monitor positions and speeds of cars➔ **detect accidents**, traffic jams
- Applications: better safety, improved logistics



Data Streams (4/4): Tele Health

Monitor health of elderly in their homes

Who are the users?



Why?

- **Difficult to reach locations**
- **Make health care more affordable**

How?

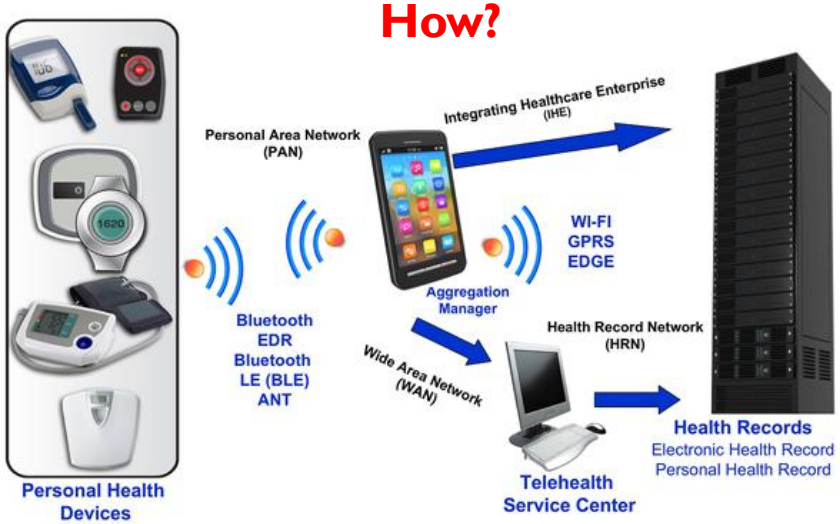


Figure 1. Basic telehealth system.

SRbench: Streaming **RDF** Benchmark

Streaming RDF data **benefits**:

- apply Linked Open Data (LOD) principles to streaming data
  - Link streaming data to data on the web (**enrichment**)
  - Publish data streams on the web
- support (simple) reasoning semantics in stream queries

➔ Richer semantics than relational streaming database systems

SRbench: Streaming **RDF** Benchmark

Streaming RDF data **challenges**:

- Proper benchmark dataset

  ➔ use real-world datasets from LOD

- No standard query language

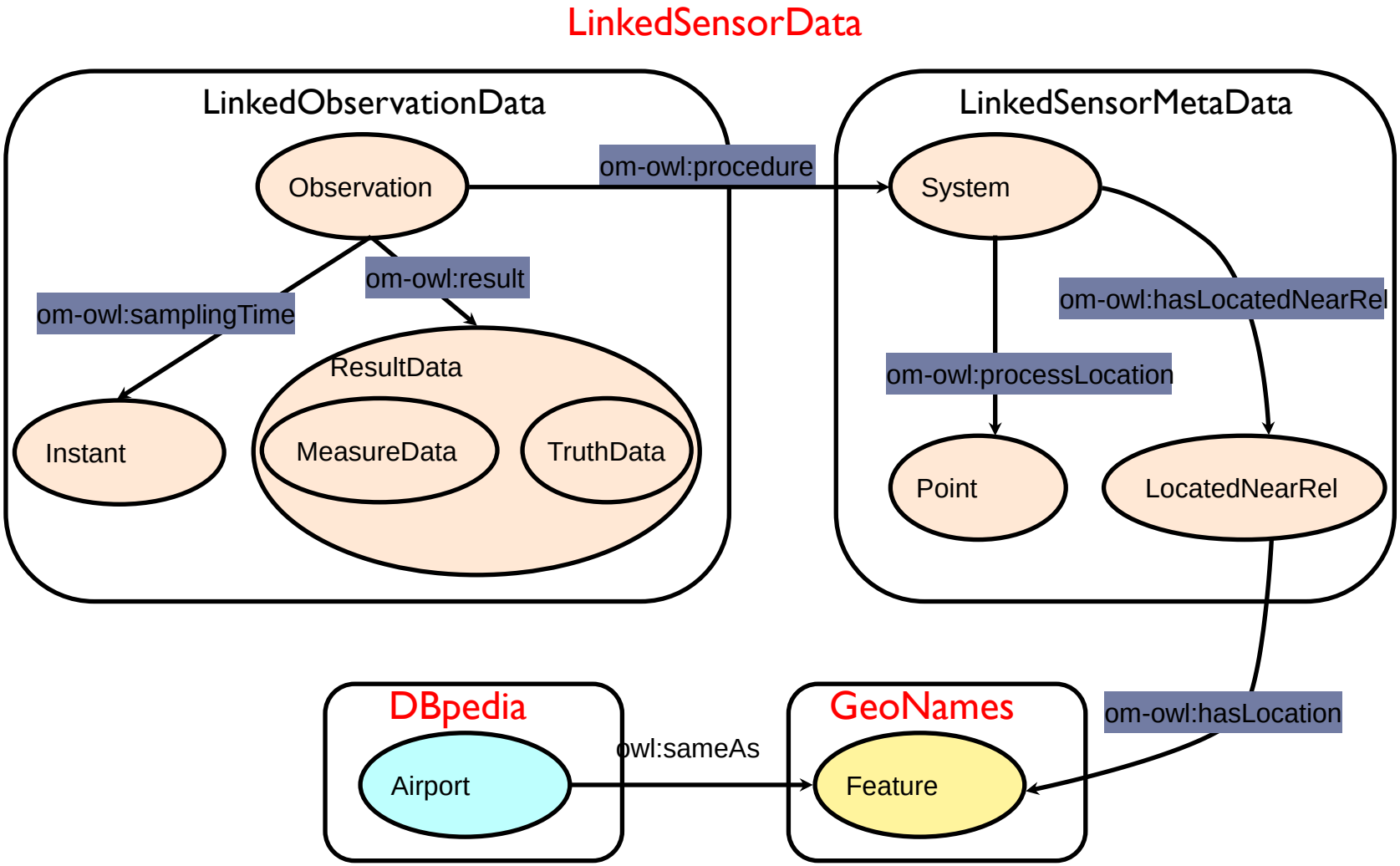
  ➔ natural language query definition +

  three implementations (SPARQLStream, CQELS, C-SPARQL)

- Limited systems support

  ➔ evaluate on the strRS system (UPM)

SRbench: used Datasets



SRBench Queries

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
1 patter matching	A	A,F,O	A	A,F	A	A,F,U	A	A	A	A	A,F	A,F,U	A,F	A,F,U	A,F	A,F	A,F
2 solution modifier	P,D	P,D	P	P	P	P	P,D	P	P	P,D	P,D	P	P	P,D	P	P	P
3 query form																	
4 SPARQL 1.1		F,P	A	A,E,M,F	A,S		N	A,E,M	A,E,M		A,S,M,F	A,S,E,M,F,P	A,E,M,F,P	F,P	A,E,M,P	P	P
5 reasoing			R												C	A	C
6 CQL feature	T	T	T	T	T	T	T,	T	T	T	T	T	T	T	T		
7 data access	0	0	0	0	0	0	0	0, S	0,S	0,S	0,S	0,S,F	0,S,G	0,S,G	0,S,D	0,S,G,D	S

Table 2. Addressed features per query. Operators are abbreviated in per row unique capital letters, defined as: 1. **A**nd, **F**ilter, **U**nion, **O**ptional; 2. **P**rojection, **D**istinct, **L**imit; 3. **S**elect, **C**onstruct, **A**sk; 4. **A**ggregate, **S**ubquery, **N**egation, **E**xpr in SELECT, assign**M**ent, **F**unctions&operators, **P**roperty path; 5. sub**C**lassOf, subp**R**opertyOf, owl:same**A**s; 6. **T**ime-based window, tu**P**le-based window, **I**stream, **D**stream, **R**stream; 7. Linked**O**bservationData, Linked**S**ensorMetadata, **G**eoNames, **D**bpedia.

Summary

- the importance of
  - **Database System Benchmarking**
  - **RDF Database System Benchmarking** (LDBC Linked Data Benchmark Council)
  - **Streaming RDF Database System Benchmarking**

- SRbench
  - Developed in PlanetData (CWI, UPM)
  - **First dedicated streaming RDF/SPARQL benchmark**

- SRbench future work:
  - performance evaluation
  - results verification (not easy!)

CWI

Thank You!

Questions?

- Ying Zhang (zhang@cwi.nl)
- Peter Boncz (boncz@cwi.nl)