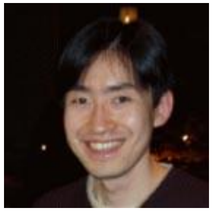
Division of Computational Informatics
Database Group

Hiroyuki Kitagawa
Center for Computational Sciences
Graduate School of Systems and Information Engineering
University of Tsukuba

Members

- Faculty
  - Hiroyuki Kitagawa (Professor)
  - Toshiyuki Amagasa (Associate Professor)
  - Hideyuki Kawashima (Lecturer; Currently in HPCS Div.)
  - Yasuhiro Hayase (Assistant Professor)
  - Chiemi Watanabe (Assistant Professor)



- Students
  - Doctoral Program: 8
  - Master Program: 24
  - Undergraduate: 7
  - Research Student: 3
- Adjunct Researchers
  - Prof. Ishikawa (Nagoya Univ.)
  - Prof. Ebisawa (JAXA)

Overview

- R&D in Data Engineering and Databases
- Main Research Areas
  - Information Integration Framework
  - Data Mining and Knowledge Discovery
  - XML and Web Programming
  - Database Applications in Science Domains

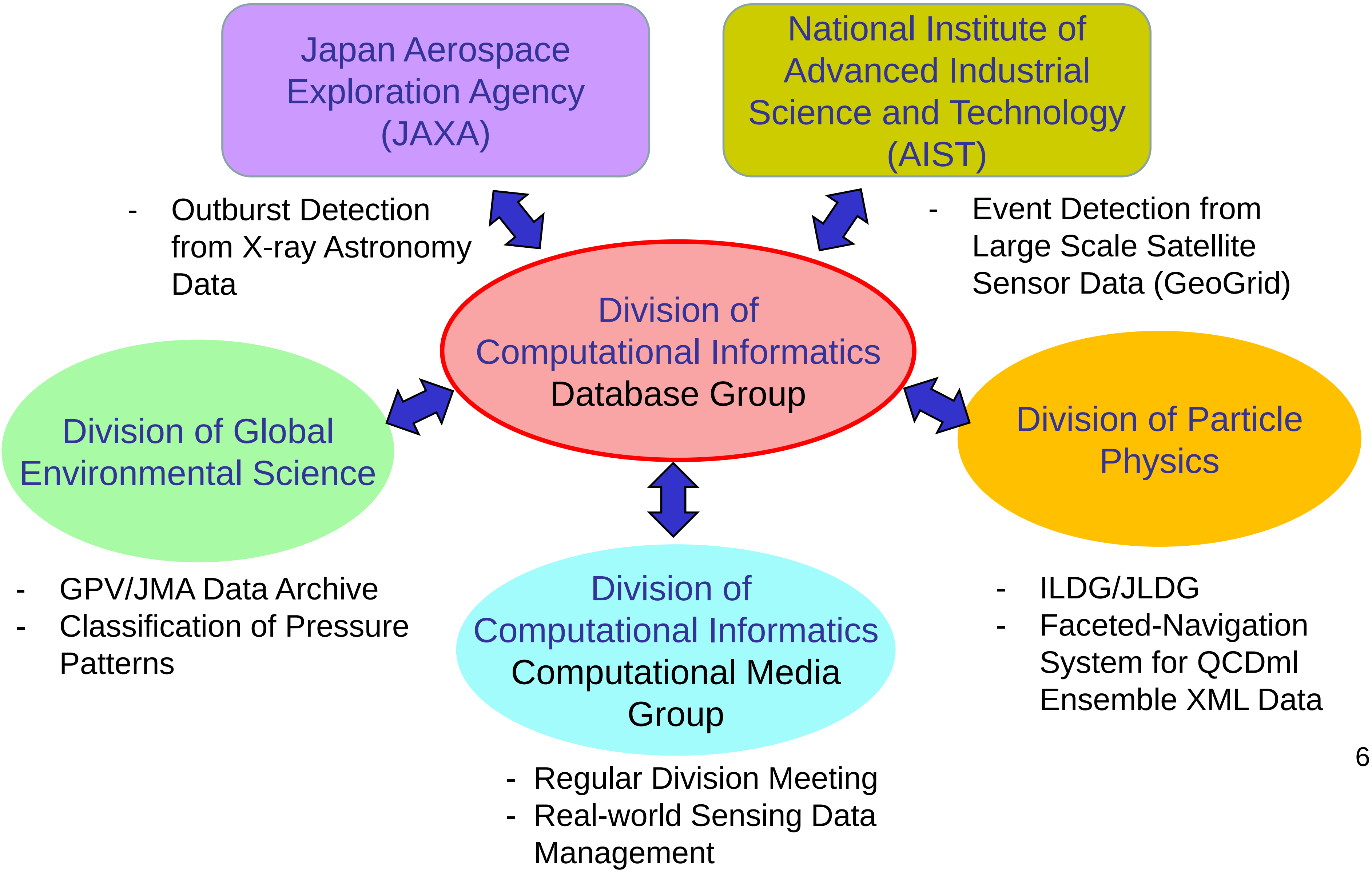
Overview

- Main Research Areas
  - Information Integration Framework
    - Integration of Heterogeneous Data Sources: DB, Web, File, XML, Sensors, …
    - Stream Processing
      - High-Availability Schemes for Distributed Stream Processing
      - Secure Stream Data Processing
      - Efficient Archiving of Stream Data
      - Outlier Detection over Packet Streams
      - Transactional Stream Processing
    - Indexing for Update-intensive Applications
  - Data Mining and Knowledge Discovery
    - Outlier Detection
    - Social Bookmark Analysis
    - Microblog Analysis
    - GPU-based Acceleration of Data Mining

Overview

- Main Research Areas (Cont.)
  - XML and Web Programming
    - Online Analytical Processing of XML Data
    - Parallel XML Query Processing using PC-Clusters/Multi-core Processors
    - Faceted-navigation of XML Data
    - Energy-efficient XML Stream Processing
    - RDF/LOD Data Processing
    - Privacy-preserving Database Querying
  - Database Applications in Science Domains
    - Development and Maintenance of GPV/JMA Archive
    - Automatic Classification of Pressure Patterns
    - Faceted-Navigation System for QCDml Ensemble XML Data
    - Event Detection from Large Scale Satellite Sensor Data
    - Outburst Detection from X-ray Astronomy Data

Collaboration



Japan Aerospace Exploration Agency (JAXA)

- Outburst Detection from X-ray Astronomy Data

National Institute of Advanced Industrial Science and Technology (AIST)

- Event Detection from Large Scale Satellite Sensor Data (GeoGrid)

Division of Computational Informatics
Database Group

Division of Global Environmental Science

- GPV/JMA Data Archive
- Classification of Pressure Patterns

Division of Particle Physics

- ILDG/JLDG
- Faceted-Navigation System for QCDml Ensemble XML Data

Division of Computational Informatics
Computational Media Group

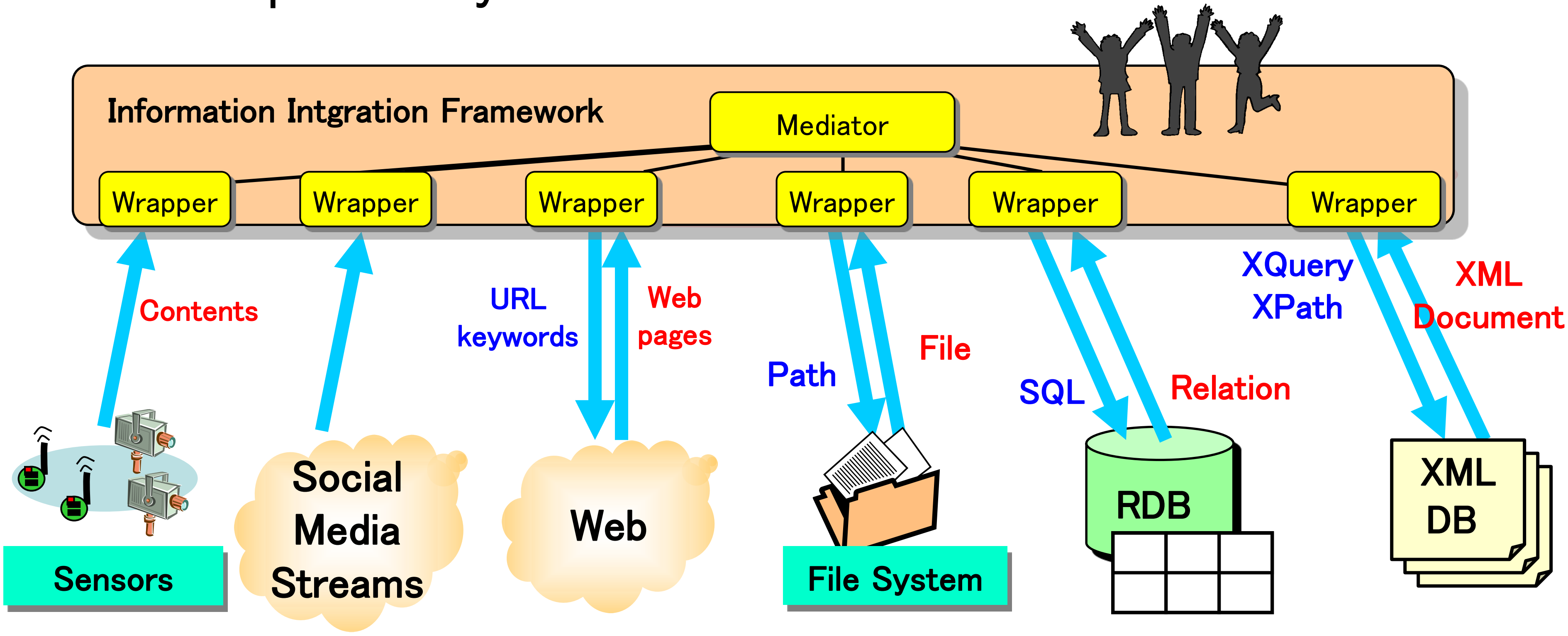
- Regular Division Meeting
- Real-world Sensing Data Management

Selected Research Topics

- Information Integration Framework
  - Integration of Heterogeneous Data Sources: DB, Web, File, XML, Sensors, …
  - Stream Processing
    - High-Availability Schemes for Distributed Stream Processing
    - ✓Transactional Stream Processing
    - Secure Stream Data Processing
    - Efficient Archiving of Stream Data
    - Outlier Detection over Packet Streams
  - Indexing for Update-intensive Applications

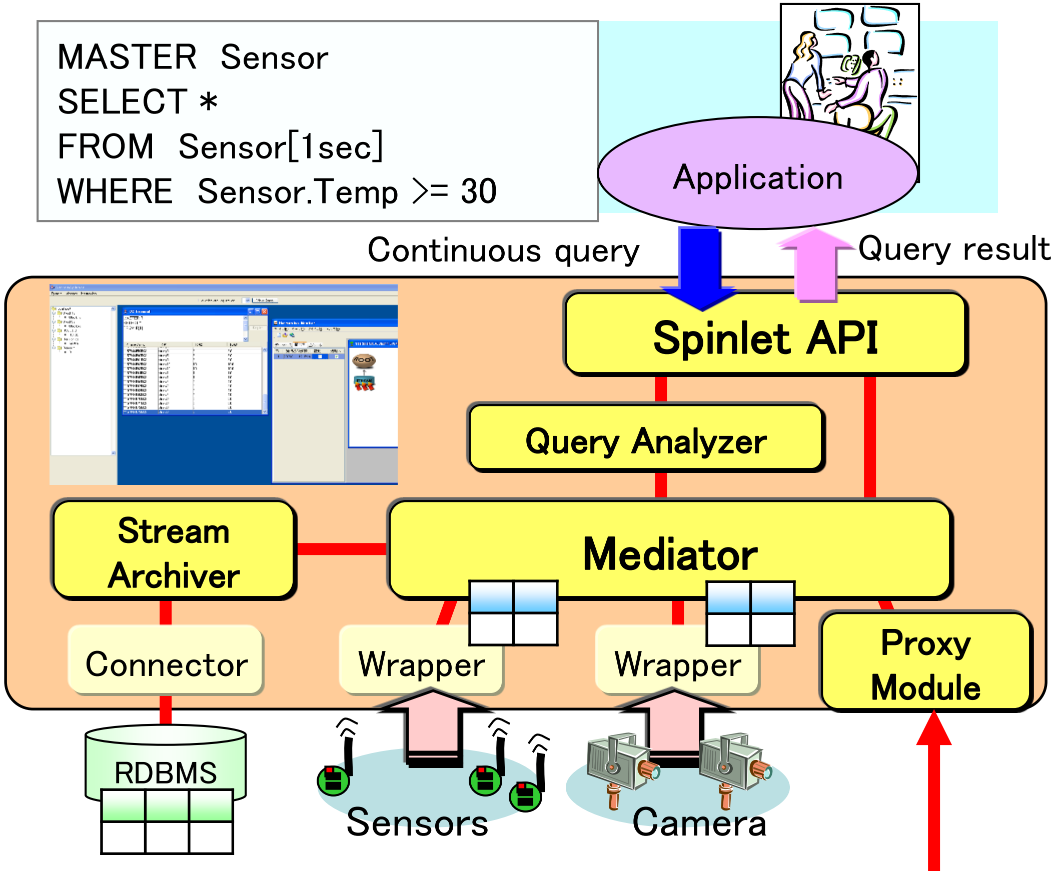
Information Integration Framework

- A variety of online data sources
  - Different data formats, access methods, query languages, ···
- Information integration framework for data interoperability

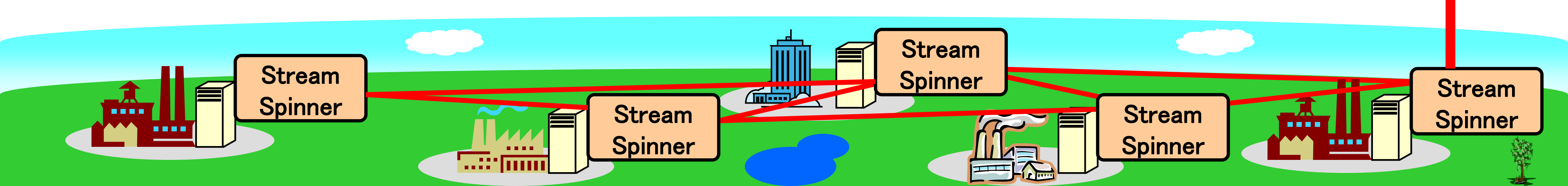


Data Integration Including Streams

- StreamSpinner, SS*, JsSpinner
  - Help integration of heterogeneous data sources
  - Streaming data sources such as sensors, location data, social media streams, etc.
  - Even-driven execution of continuous queries
  - Distributed stream processing

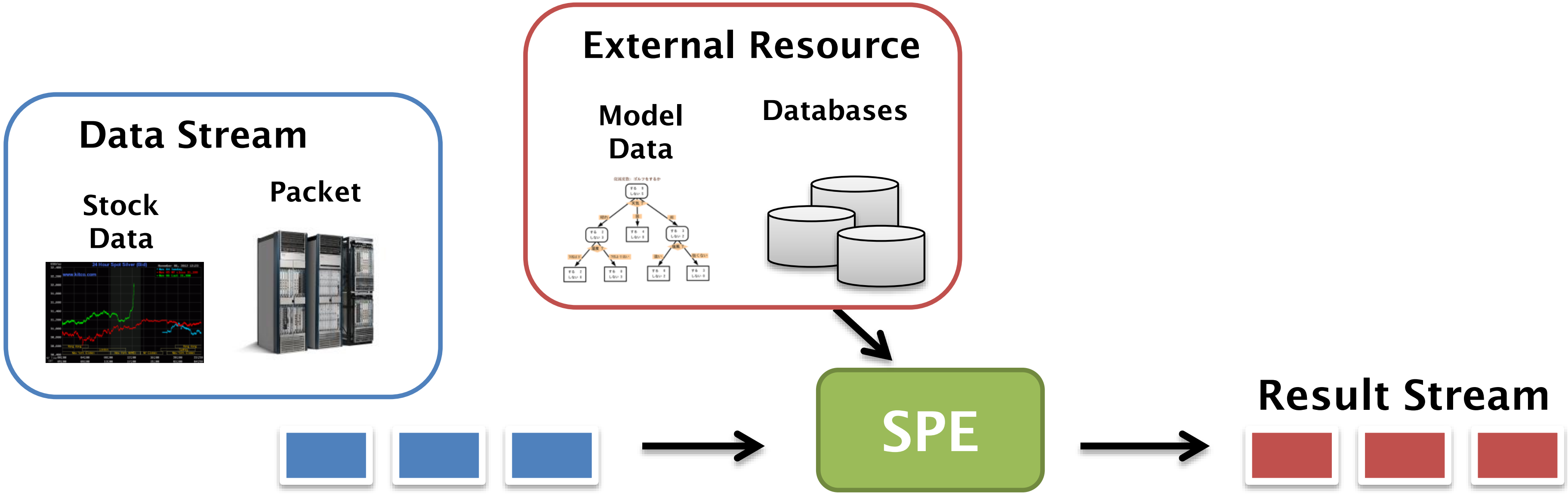


```
MASTER  Sensor
SELECT *
FROM  Sensor[1sec]
WHERE  Sensor.Temp >= 30
```



Transactional Stream Processing

In data integration context, SPEs do not process only data streams but integrates non-streaming external resources



External Resource

Model Data

Databases

Data Stream

Stock Data

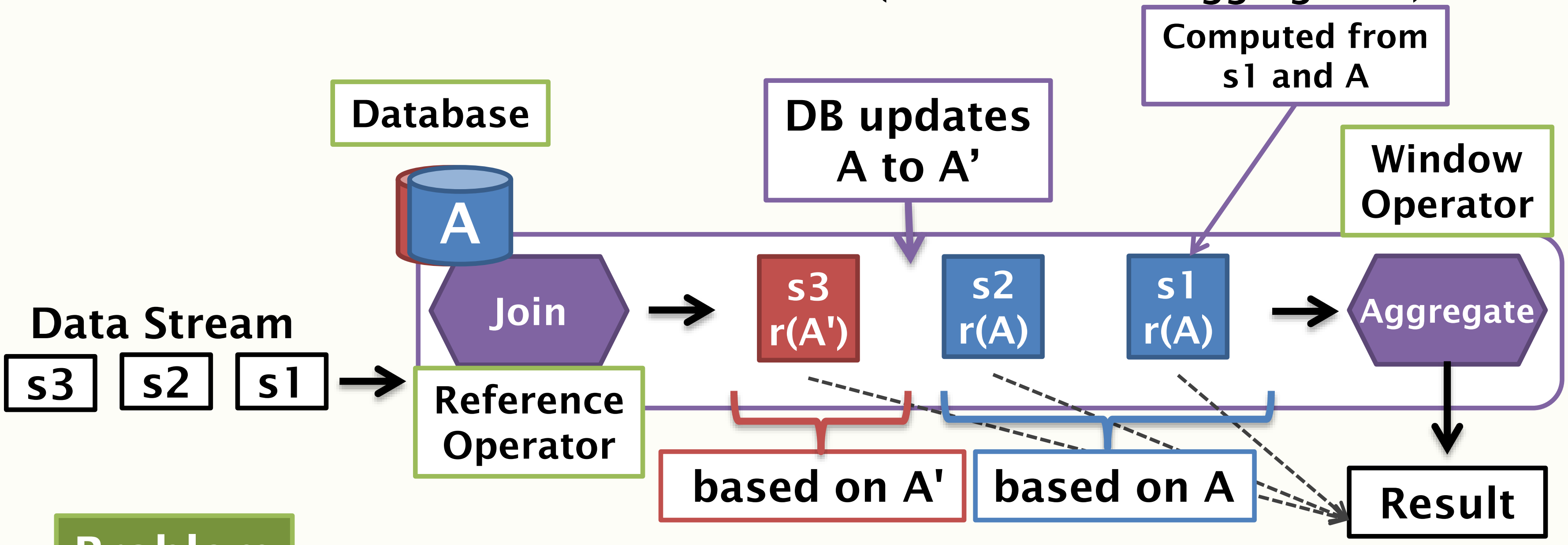
Packet

SPE

Result Stream

Problem: Resource reference inconsistency

- External resources may be updated or modified autonomously.
- Continuous query
  - Integrate a data stream & DB → aggregate results (window-based aggregation)



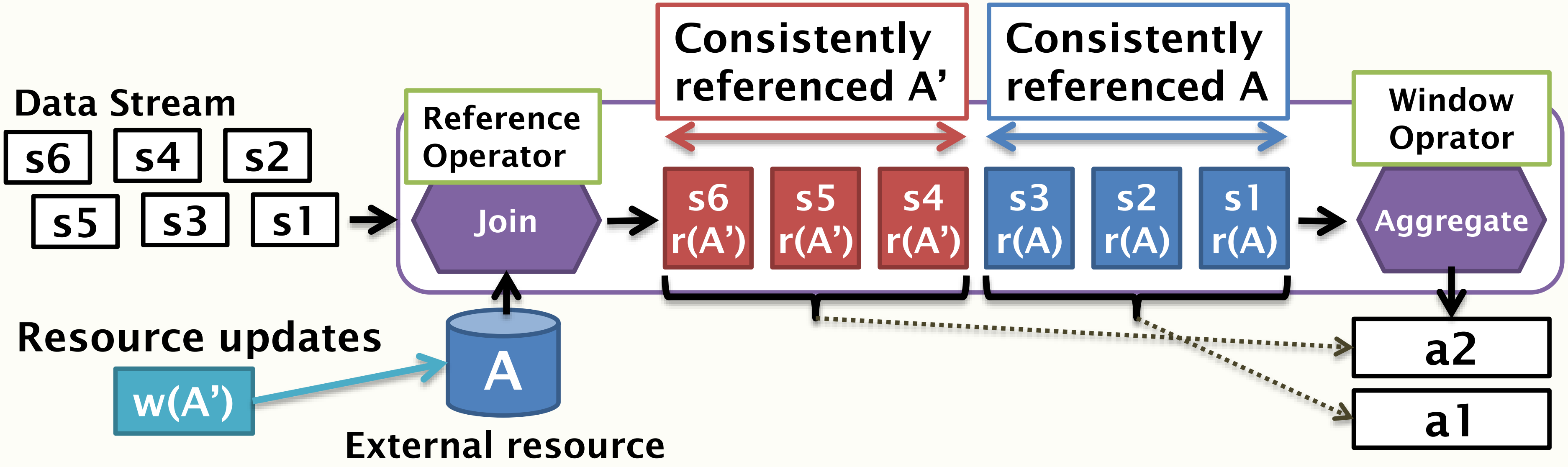
Problem

External resources are not consistently referenced in a single CQ execution instance.

Our Goal: Transactional Stream Processing

Goal

Even if **external resources are updated independently**, we guarantee that **external resources are consistently referenced in each CQ execution instance.**



How to make stream processing transactional?

- **CQ-processing ensuring serializability**
  - □ Serializability of **all CQ-Txns** and DB **Update-Txns**
- **Approaches**

  (a) Combining exiting concurrency control mechanisms with stream processing

  - □ Two-Phase Locking Strategy (2PL)
  - □ Snapshot Strategy (C2PL)
  - □ Optimistic Strategy

  (b) Stream processing combining a redo mechanism and external resource state monitoring

Selected Research Topics

- Data Mining and Knowledge Discovery
  - Outlier Detection
  - Social Bookmark Analysis
  - ✓ Microblog Analysis
  - ✓ GPU-based Acceleration of Data Mining

Microblog Analysis

Real-world sensing

- ✓ **Event Detection** [Sakaki+, 10] [Walther+, 13] …
- ✓ **Epidemics Analysis** [Paul+, 11] [Aramaki+, 11] …
- ✓ **Disaster Analysis** [Vieweq+, 10] [Mandel+, 12] …

Location inference

- Most users hesitate to disclose their home locations in their profiles.
- Only few tweets have GEO-tags.

Earthquake!
Shaked!
Shaked!

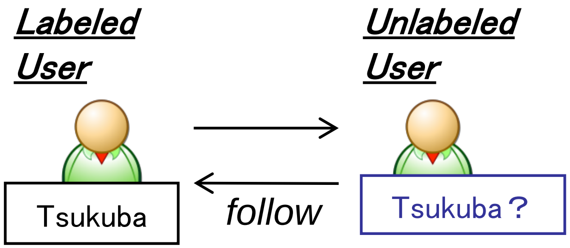
Um…
Earthquake happened.
But WHERE?

Graph Based Approach

- Utilize social graphs based on friendships
- Closeness vs. Concentration

Closeness assumption **(Traditional approaches)**

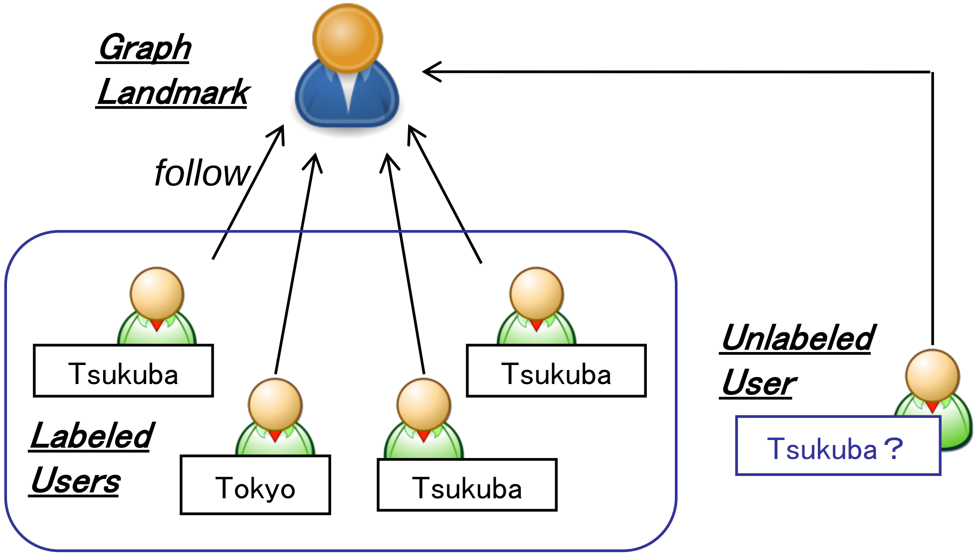
IF: FRIENDS
THEN: CLOSE



Concentration assumption **(proposed)**

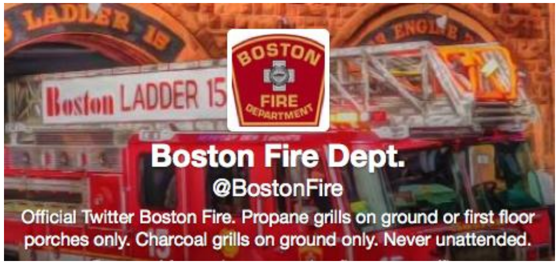
IF: FOLLOW A GRAPH LANDMARK
THEN: CLOSE

A user whose followers are close to each other



Graph Based Approach

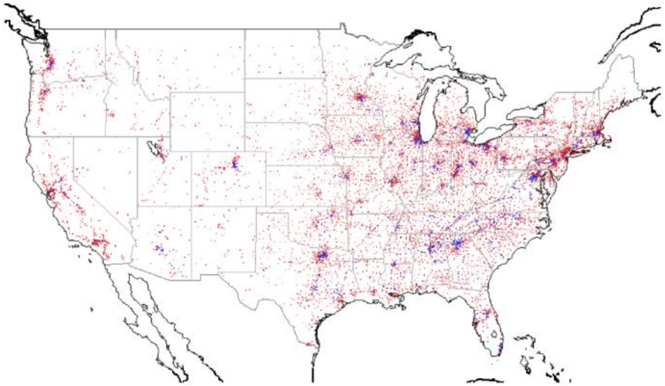
Graph Landmark Example



Boston Fire Dept.

@BostonFire

Official Twitter Boston Fire. Propane grills on ground or first floor porches only. Charcoal grills on ground only. Never unattended.



red: regular users
blue: graph landmarks

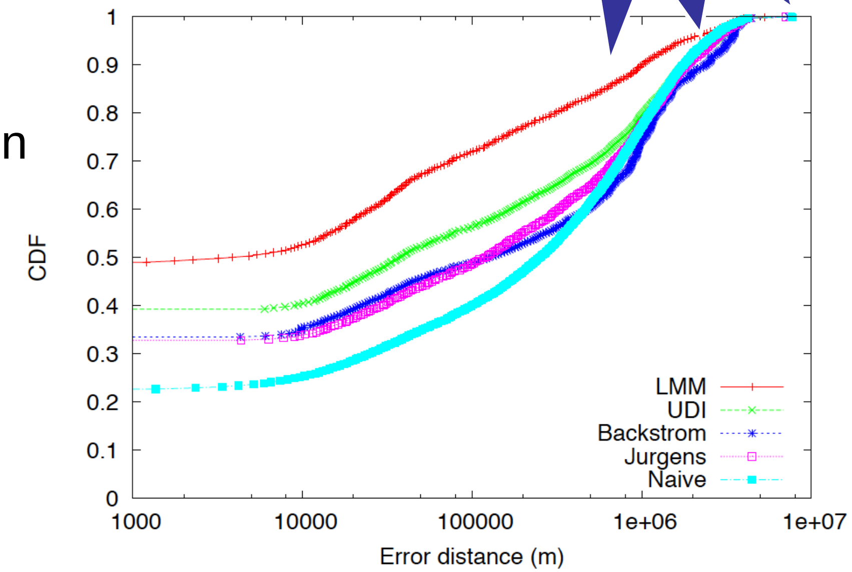
27%+ IMPROVED

Accuracy Comparison

- ✓ LMM: Proposed
- ✓ UDI: [Li+, 12]
- ✓ Backstrom: [Backstrom+, 10]
- ✓ Jurgens: [Jurgens, 13]
- ✓ Naïve: Medoid

concentration assumption

closeness assumption

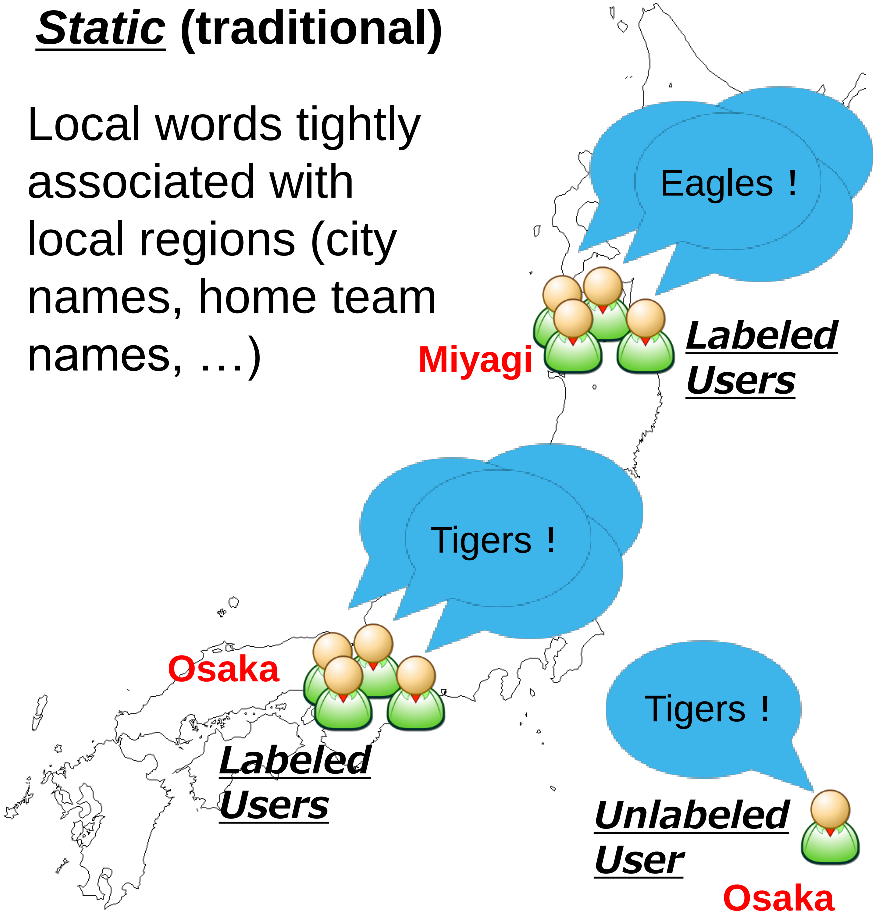


Content Based Approach

Static vs. Temporal Local Words

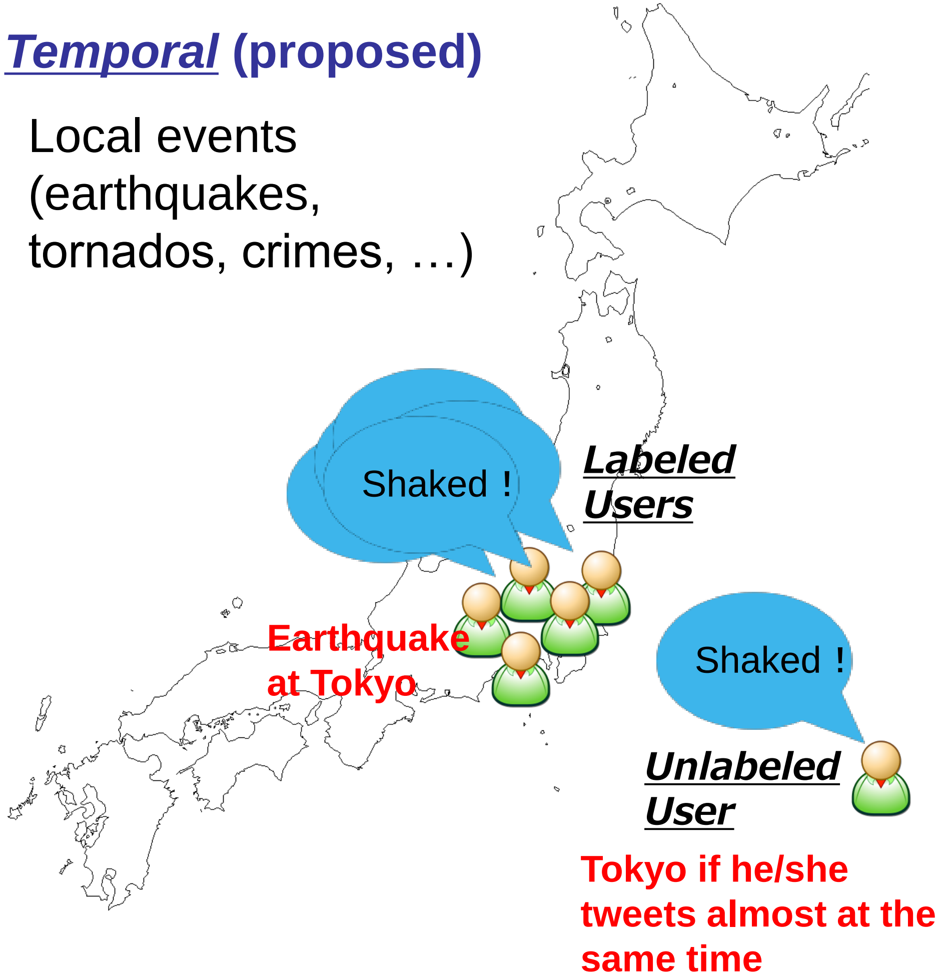
Static **(traditional)**

Local words tightly associated with local regions (city names, home team names, …)



Temporal **(proposed)**

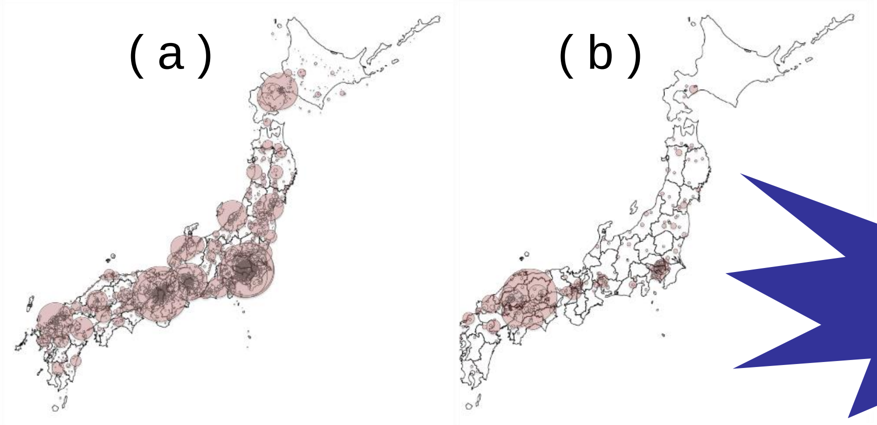
Local events (earthquakes, tornados, crimes, …)



Content Based Approach

Identified Local Events

Distributions of tweets
(a) at ordinary times
(b) after an earthquake
at Hiroshima



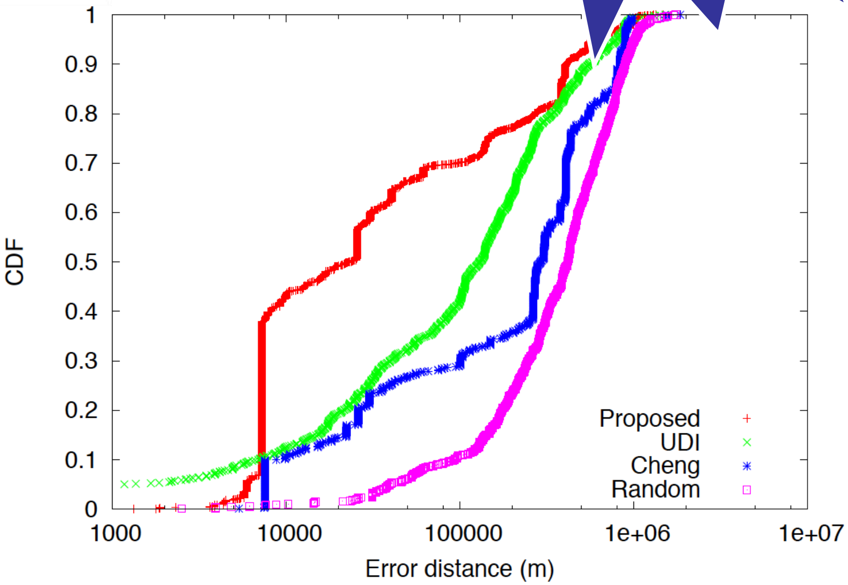
33%+
IMPROVED

Accuracy Comparison

- ✓ Proposed: Proposed
- ✓ UDI: [Li+, KDD'12]
- ✓ Cheng: [Cheng+, CIKM'10]
- ✓ Random: Random Assignment

} temporal

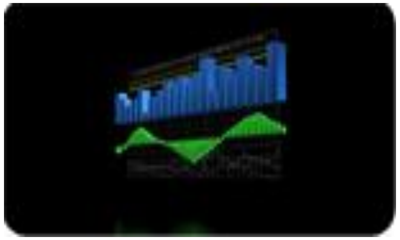
} static



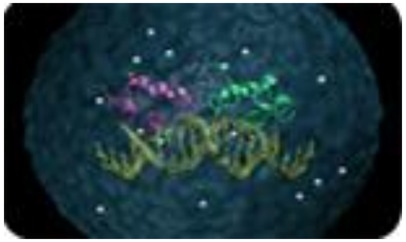
GPU-based Acceleration of Data Mining

- Gaining growing attention due to its cost and performance.

Fluid Simulation



Signal Processing



Biology

Database & Data Mining

- The DB group has been trying to apply GPU for accelerating various data mining processing.
  - Probabilistic Latent Semantic Indexing [ICCS’11]
  - Frequent itemset mining over uncertain databases [CIKM’12, DEXA’13, IEICE Trans. ’14]
  - Currently working on sorting huge arrays, time series matching, clustering, and similarity join.

※ Images from NVidia CUDA case home page

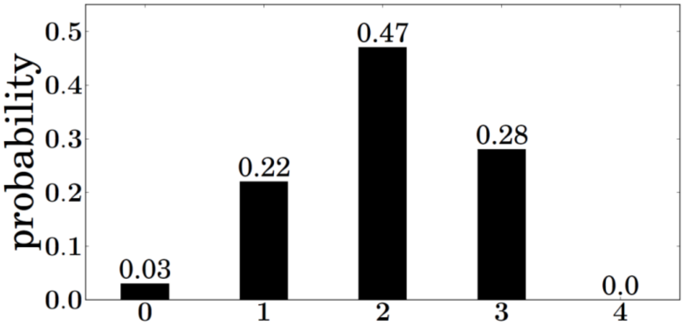
Frequent Itemset Mining over Uncertain Databases

- A set of transactions
  - Transaction: ID, itemset, and probability
- Possible worlds
  - {T1, T2}: prob. is 0.028
  - {T1, T2, T3}: prob. is 0.252
  - …
- Support sup(X)
  - Conventional: # of transactions containing X
  - Uncertain: random variable

  ➔ Support Probability Mass Function (SPMF)
- Probabilistic Frequent Itemset (PFI)
  - P(sup(X)≥minsup)≥minprob
  - minsup and minprob are user-specified values
- **Accelerate frequent itemset mining (PFIM) using GPU**

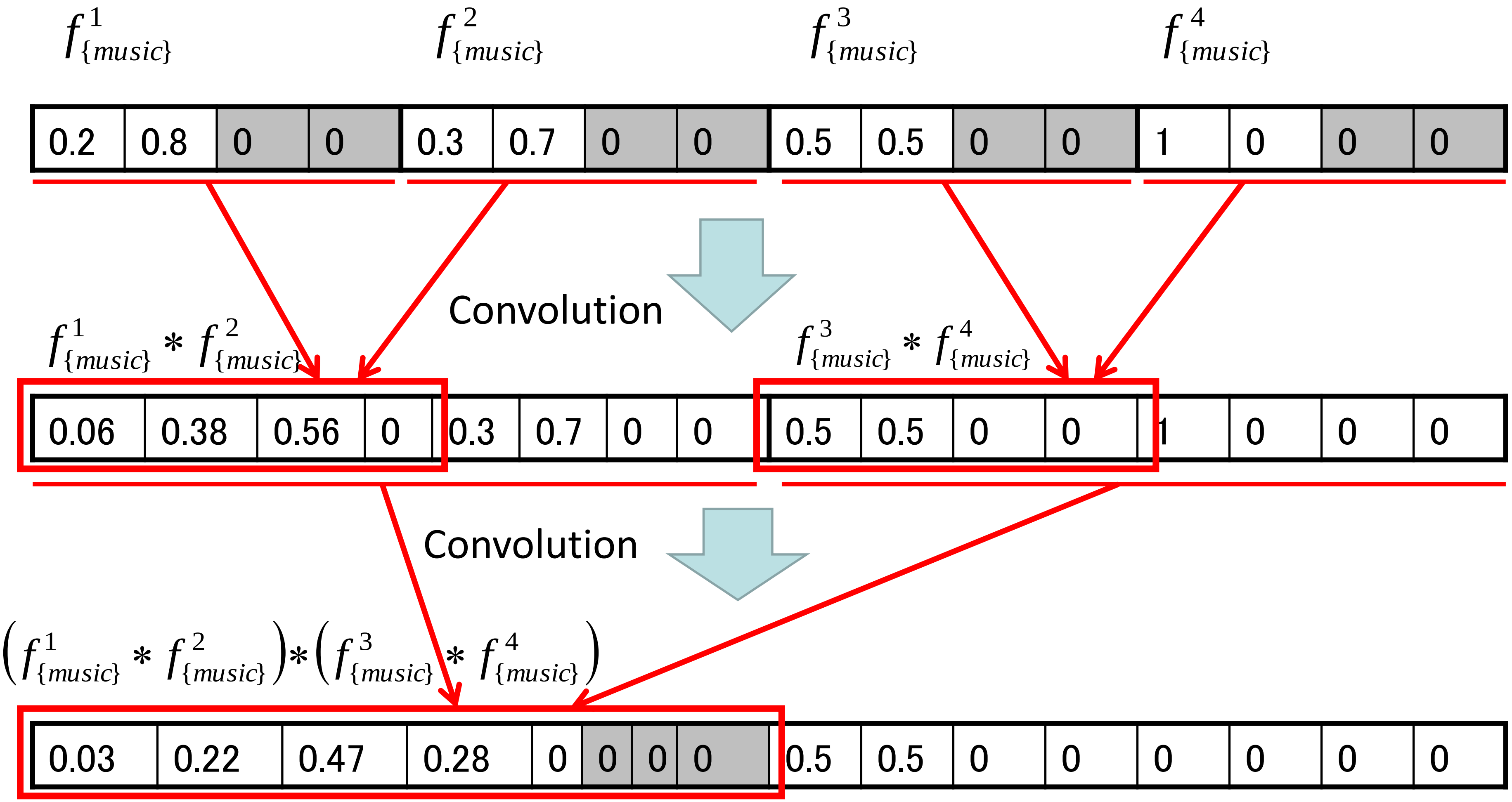
Uncertain transaction DB

ID	Itemset	Prob.
T1	{a, b}	0.8
T2	{b, c}	0.7
T3	{a}	0.9
T4	{a, b, c}	0.5



SPFM of {b}

SPMF Computation on GPU



$f^{1}_{\{music\}}$ $f^{2}_{\{music\}}$ $f^{3}_{\{music\}}$ $f^{4}_{\{music\}}$

0.2	0.8	0	0	0.3	0.7	0	0	0.5	0.5	0	0	1	0	0	0

Convolution

$f^{1}_{\{music\}} * f^{2}_{\{music\}}$ $f^{3}_{\{music\}} * f^{4}_{\{music\}}$

0.06	0.38	0.56	0	0.3	0.7	0	0	0.5	0.5	0	0	1	0	0	0

Convolution

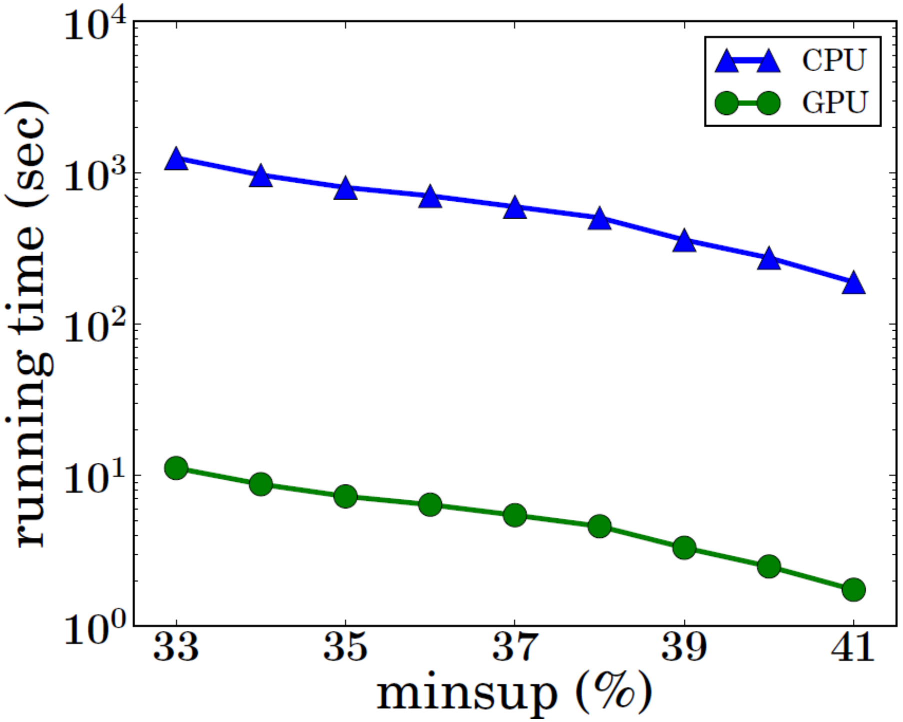
$\left(f^{1}_{\{music\}} * f^{2}_{\{music\}}\right) * \left(f^{3}_{\{music\}} * f^{4}_{\{music\}}\right)$

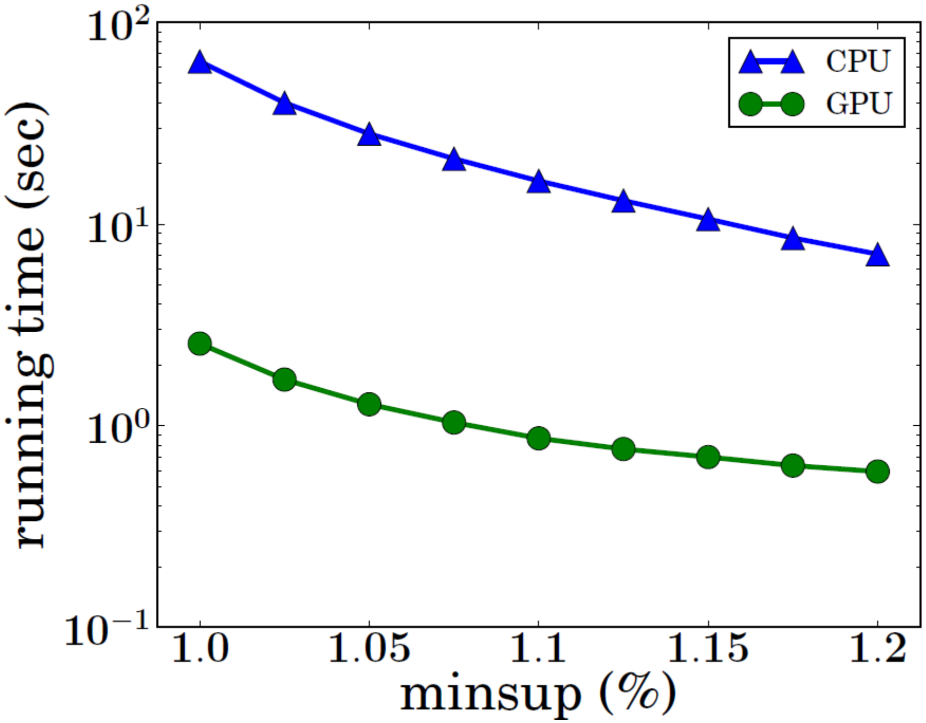
0.03	0.22	0.47	0.28	0	0	0	0	0.5	0.5	0	0	0	0	0	0

Experiments

CPU: Inten Xeon CPU (2.40 GHz) with 4GB memory
GPU: Tesla C2050 (1.15GHz, 3.0GB memory)

- Accidents
  - 106−112x
- T25I10D500K
  - 3.9-22x





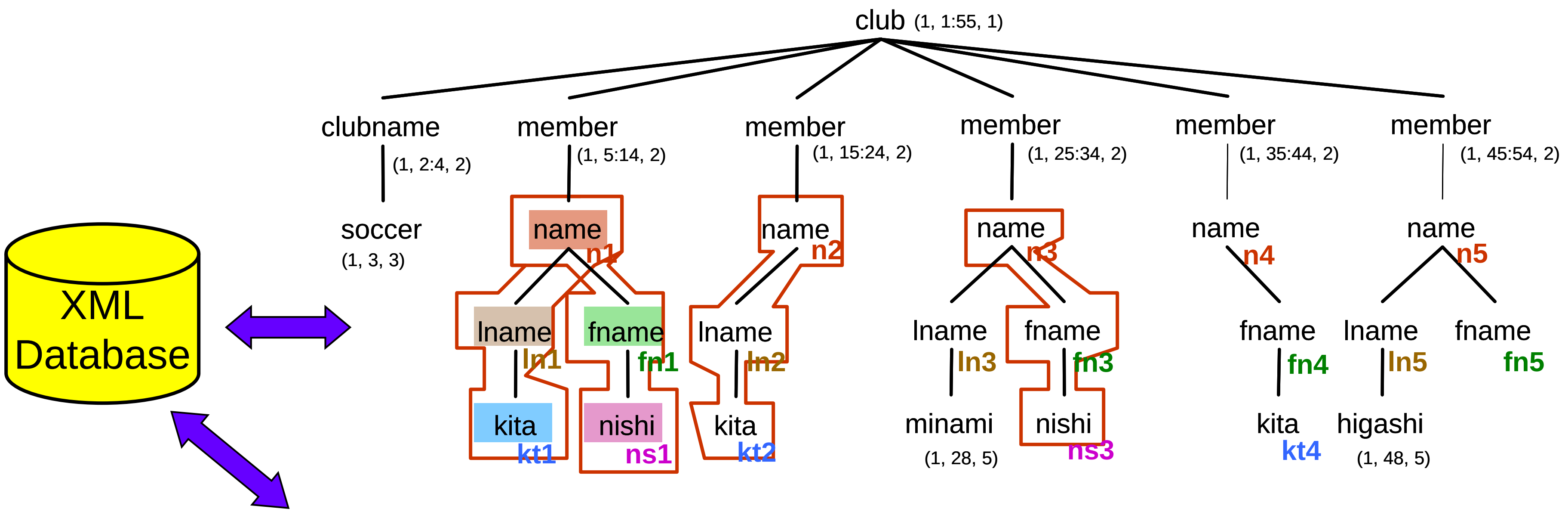
Selected Research Topics

- XML and Web Programming
  - ● Online Analytical Processing of XML Data
  - ✓ Parallel XML Query Processing using PC-Clusters/Multi-core Processors
  - ● Faceted-navigation of XML Data
  - ● Energy-efficient XML Stream Processing
  - ✓ RDF/LOD Data Processing
  - ● Privacy-preserving Database Querying

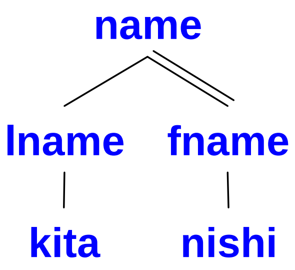
Parallel XML Query Processing on a Multi-core System

- Pattern matching queries are important in querying XML.
- Holistic twig joins (HTJ)
  - A family of XML query processing algorithms
  - Find matches for a given query tree (twig)
- Propose a parallel version of TwigStack algorithm for multi-core processors

TwigStack Algorithm



XML Query



XML Node Streams

n1	n2	n3	n4	n5
ln1	ln2	ln3	ln5	
kt1	kt2	kt4		
fn1	fn3	fn4	fn5	
ns1	ns3			

1st Phase (Task 1):

1. Perform root-to-leaf pattern matching.
2. Generate root-to-leaf path solutions.

√ n1-ln1-kt1

√ n1-fn1-ns1

n2-ln2-kt2

n3-fn3-ns3

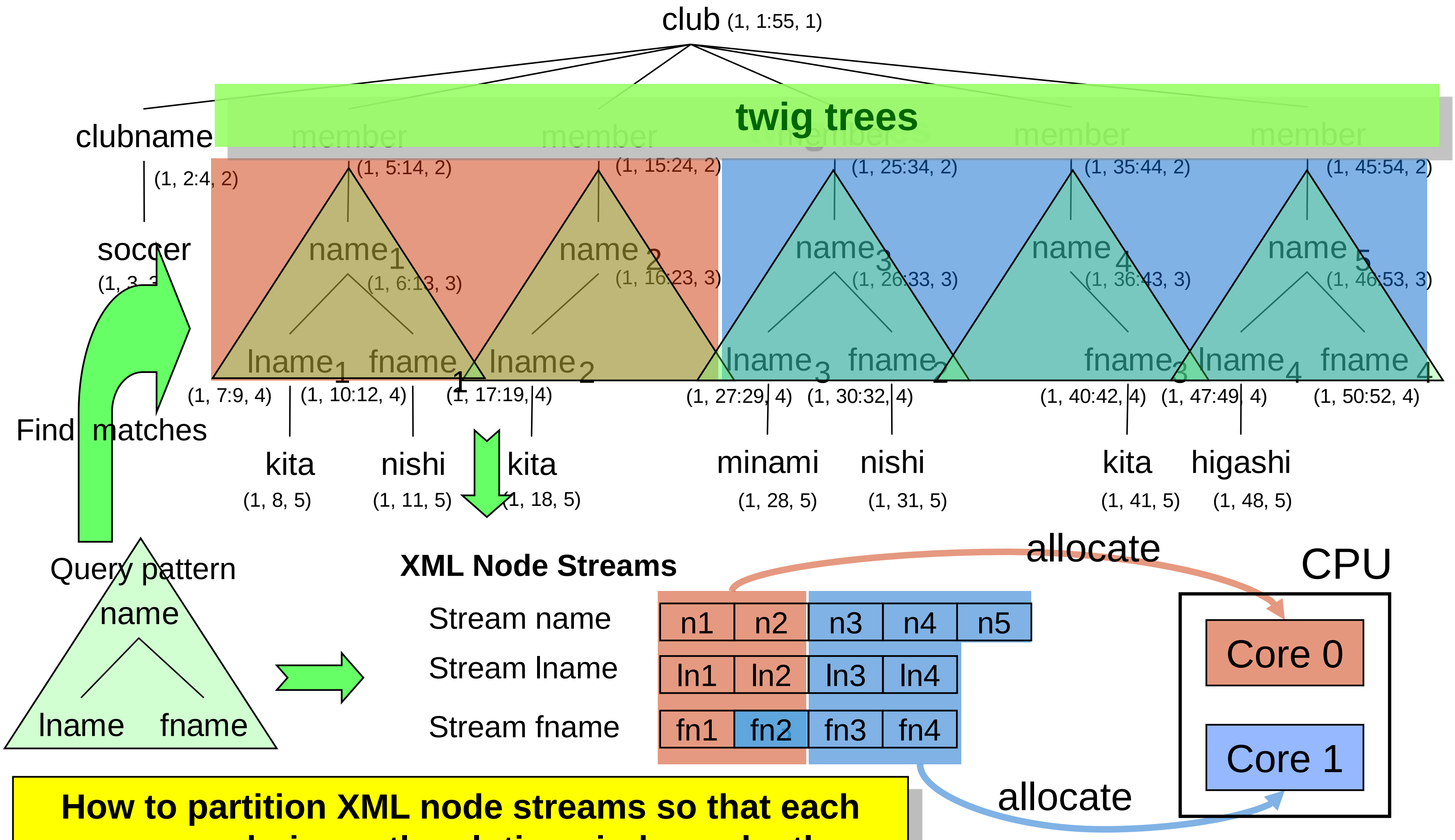
2nd Phase (Task 2):

1. Compile root-to-leaf path solutions for final solutions.

n1-ln1-kt1-fn1-ns1

Basic Idea

- Partition the XML tree and process in parallel.

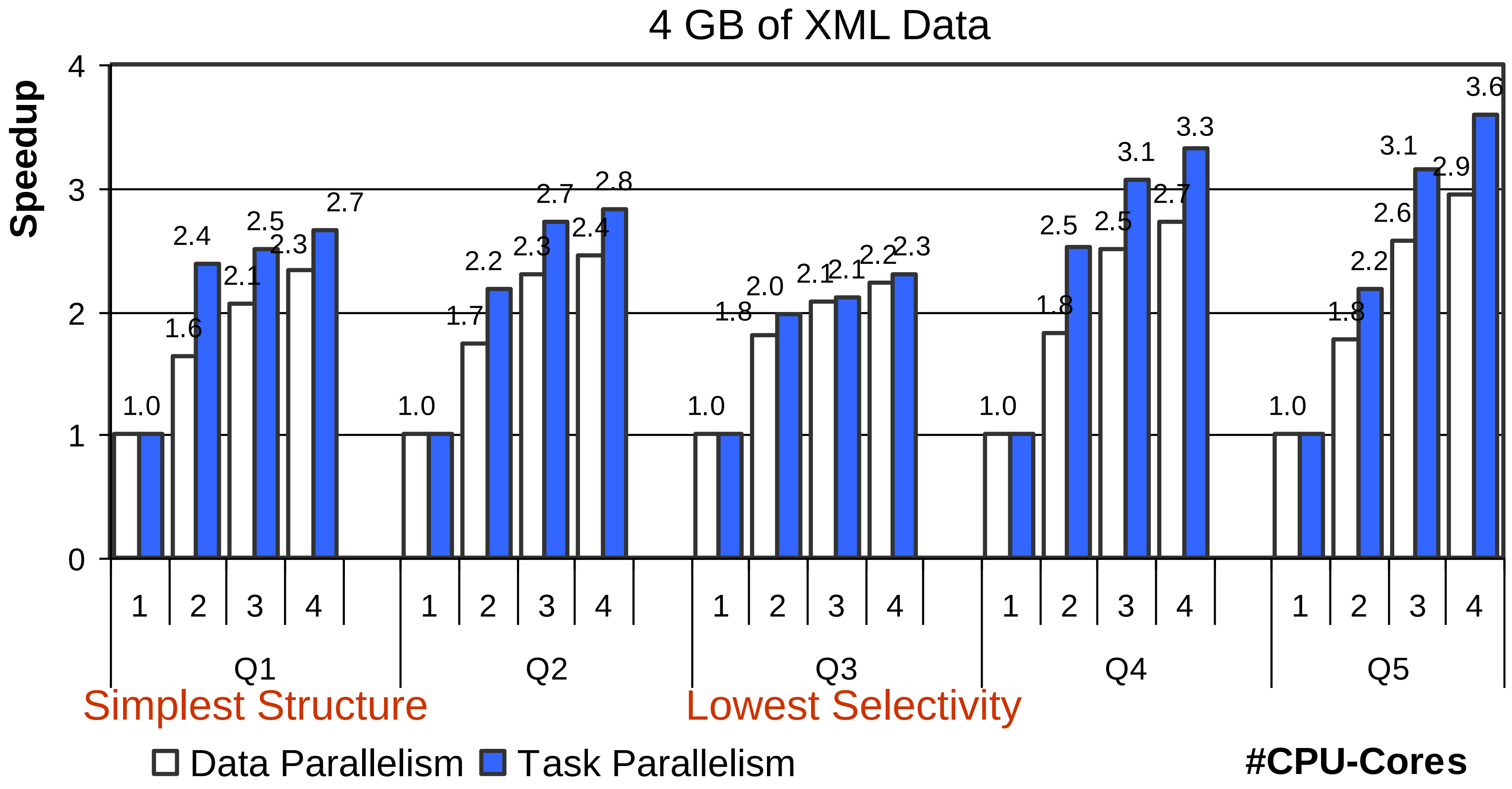


Stream name	n1	n2	n3	n4	n5
Stream lname	ln1	ln2	ln3	ln4	
Stream fname	fn1	fn2	fn3	fn4	

How to partition XML node streams so that each core can derive path solutions independently.

Experiments

- Sequential execution time
  - Q1: 18.48 s, Q2: 44.62 s, Q3: 17.16 s, Q4: 24.65 s, Q5: 15.44 s



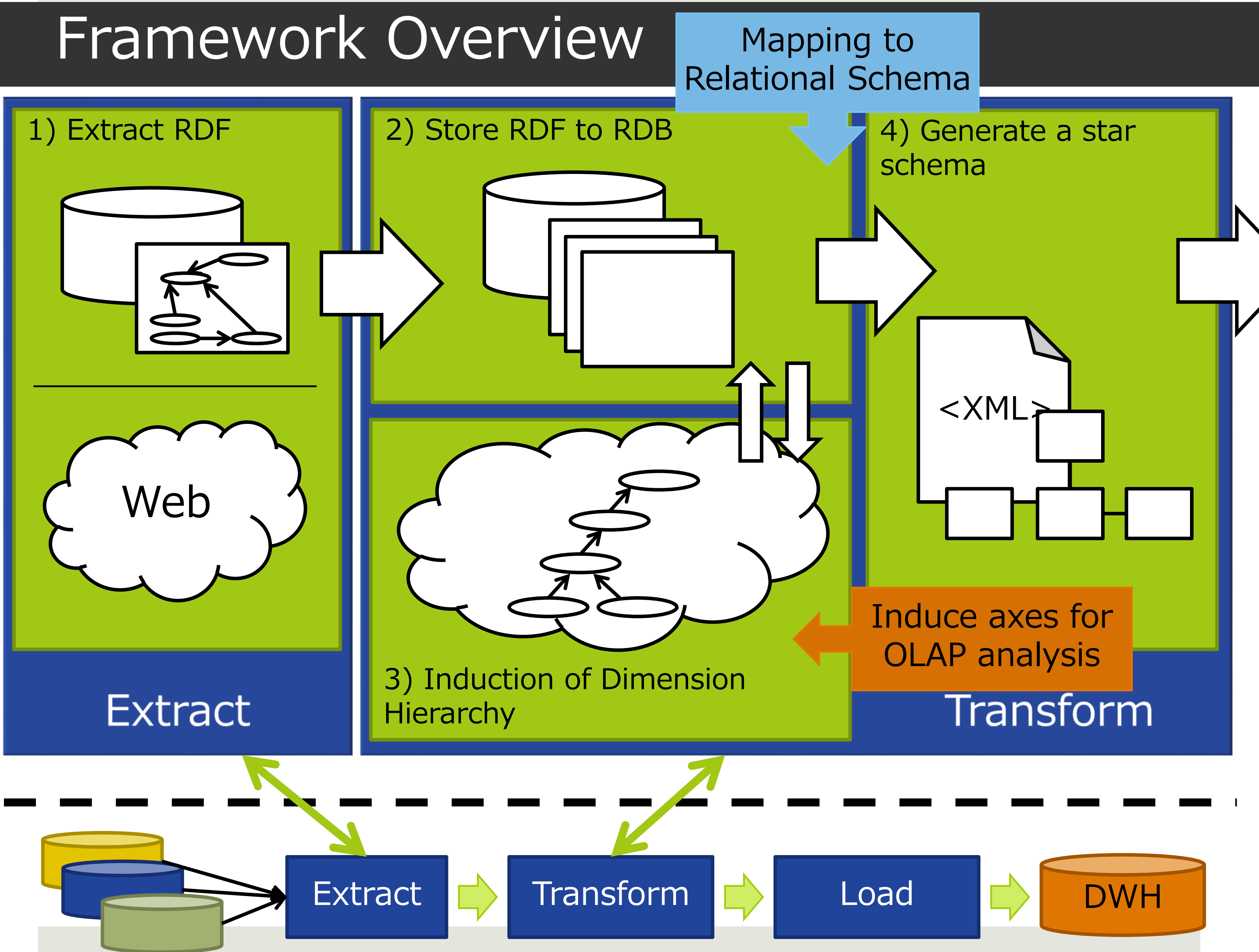
Simplest Structure

Lowest Selectivity

Linked Open Data, RDF

- Linked Open Data (LOD) is increasing rapidly
  - A method to publish and share structured data on the Web
  - "Web of Data": Data linked with each other
- Resource Description Framework (RDF)
  - A framework for describing resources on the Web
  - Triple: Subject, Predicate, and Object
- Numerical data also published as Linked Open Data
  - Statistics from governments, sensor data, etc.
  - Growing demands for analytical processing over LOD data..
- **We propose an ETL framework for the OLAP analysis of LOD datasets.**
  - **Derivation of a star schema from a large RDF graph.**

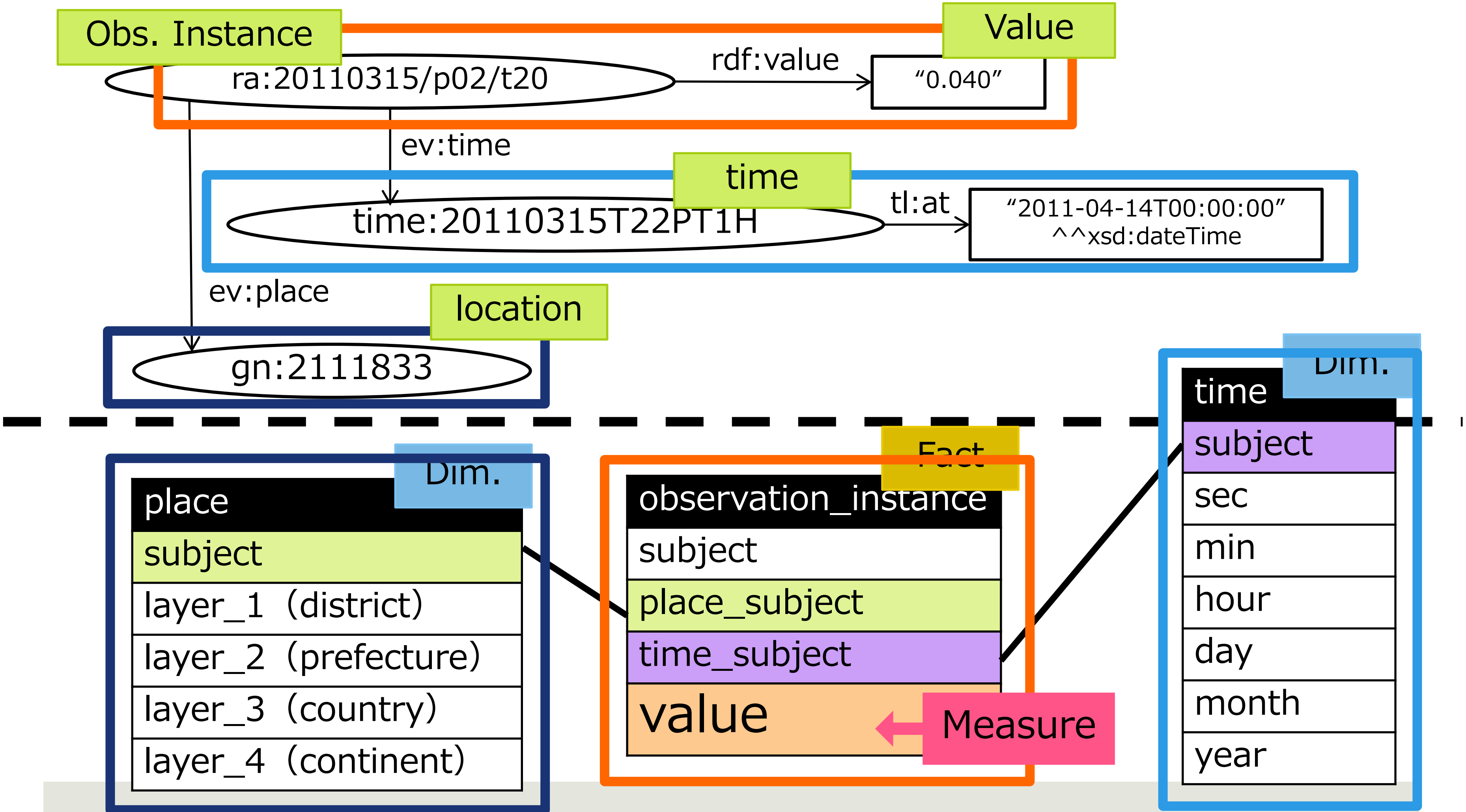
Framework Overview


Mapping to Relational Schema
1) Extract RDF
Web
2) Store RDF to RDB
4) Generate a star schema
<XML>
3) Induction of Dimension Hierarchy
Induce axes for OLAP analysis
Extract
Transform
Extract
Transform
Load
DWH

Results using (real) Linked Sensor Data

- Generated schema from LOD dataset.
  - fact-table) observation_instance
  - dim.-table) Time, Location

Environmental radio activity level monitoring data published by Japan Nuclear Regulation Authority



Obs. Instance

ra:20110315/p02/t20

rdf:value

Value

"0.040"

ev:time

time

time:20110315T22PT1H

tl:at

"2011-04-14T00:00:00"
^^xsd:dateTime

ev:place

location

gn:2111833

place
subject
layer_1 (district)
layer_2 (prefecture)
layer_3 (country)
layer_4 (continent)

Dim.

observation_instance
subject
place_subject
time_subject
value

Fact

Measure

time
subject
sec
min
hour
day
month
year

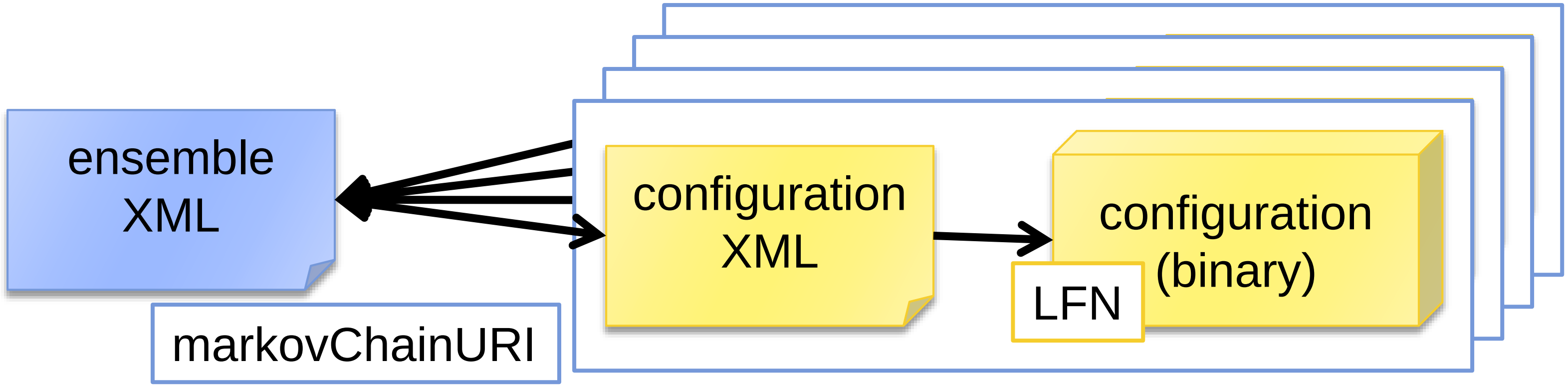
Dim.

Selected Research Topics

- Database Applications in Science Domains
  - Development and Maintenance of GPV/JMA Archive
  - Automatic Classification of Pressure Patterns
  - ✓ Faceted-Navigation System for QCDml Ensemble XML Data
  - Event Detection from Large Scale Satellite Sensor Data
  - ✓ X-ray Outburst Detection from X-ray Astronomy Data

Int’l Lattice Data Grid (ILDG)

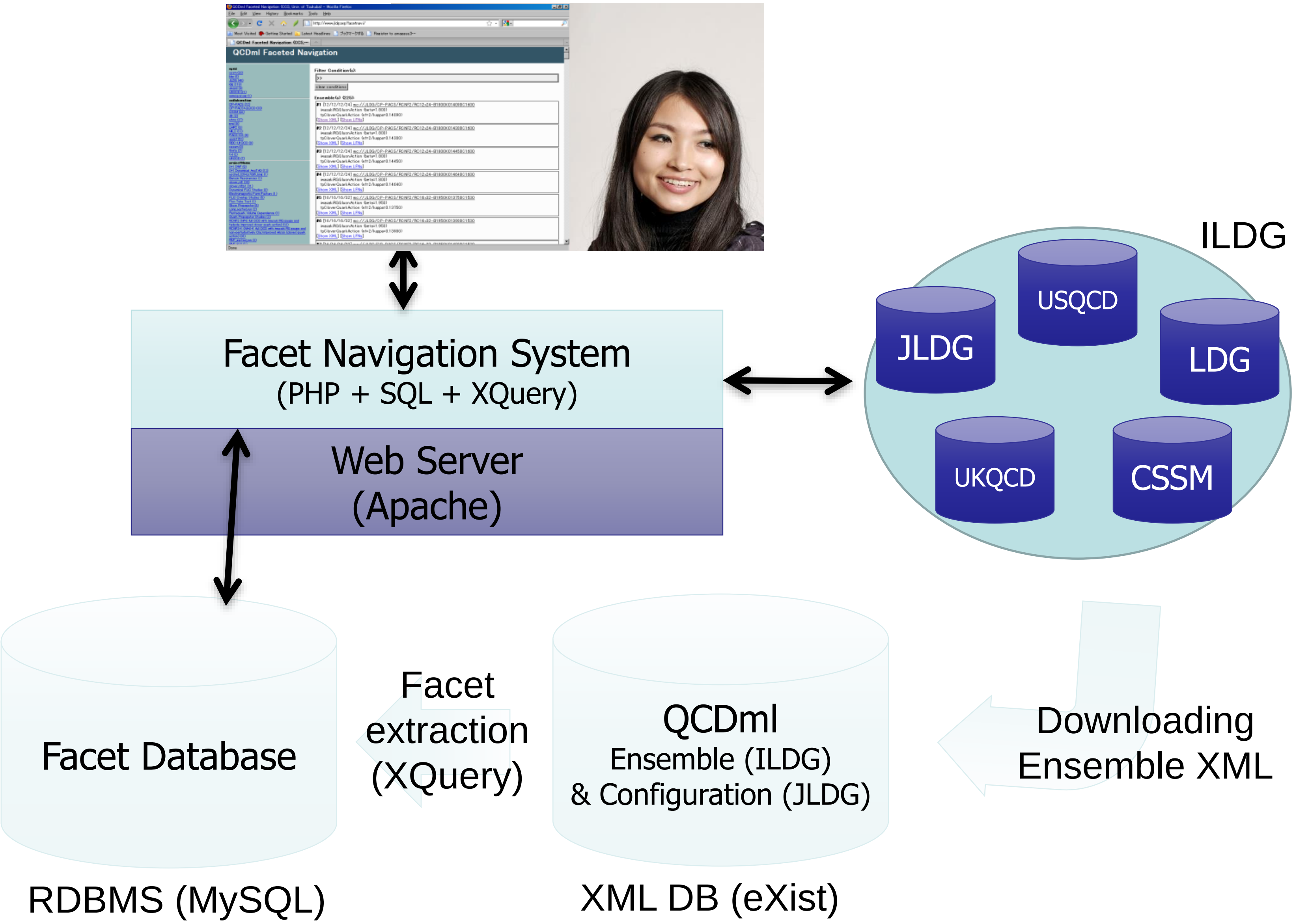
- An international collaboration which provides standards, services, methods and tools that facilitates the sharing and interchange of lattice QCD gauge configurations by integrating their regional data grids.
- File formats in ILDG
  - Configuration binary (10+TB in JLDG)
    - LIME (Lattice QCD Interchange Message Encapsulation)
  - Metadata (QCDml)
    - Ensemble XML (200+ in ILDG)
    - Configuration XML (30,000+ in JLDG)
  - A number of configuration binaries are associated with an ensemble in terms of markovChainURI and LFN.



QCDml Ensemble XML

```
<markovChain xmlns="…">
  <markovChainURI>mc: //JLDG/CP-PACS/RCNF2/RC12x24-
B1800K014090C1600</markovChainURI>
  <management>
    <revisions>1</revisions>
    <collaboration>CP-PACS</collaboration>
    <projectName>RCNF2 (Nf=2 full QCD with iwasaki RG gauge and
tadpole improved clover quark action)</projectName>
    <ensembleLabel>B1800</ensembleLabel>
    <reference>Phys.Rev. D65 (2002) 054505 (hep-lat/0105015),
Erratum-ibid. D67 (2003) 059901</reference>
    <archiveHistory>
      <elem>
        <revision>1</revision>
        <revisionAction>add</revisionAction>
        <participant>
          <name>T.Yoshie</name>
          <institution>Center fof Computational Sciences, University
of Tsukuba</institution>
```

QCDml Faceted Navigation Interface System Overview



Facet Navigation System
(PHP + SQL + XQuery)

Web Server
(Apache)

ILDG

USQCD

JLDG

LDG

UKQCD

CSSM

Facet Database

RDBMS (MySQL)

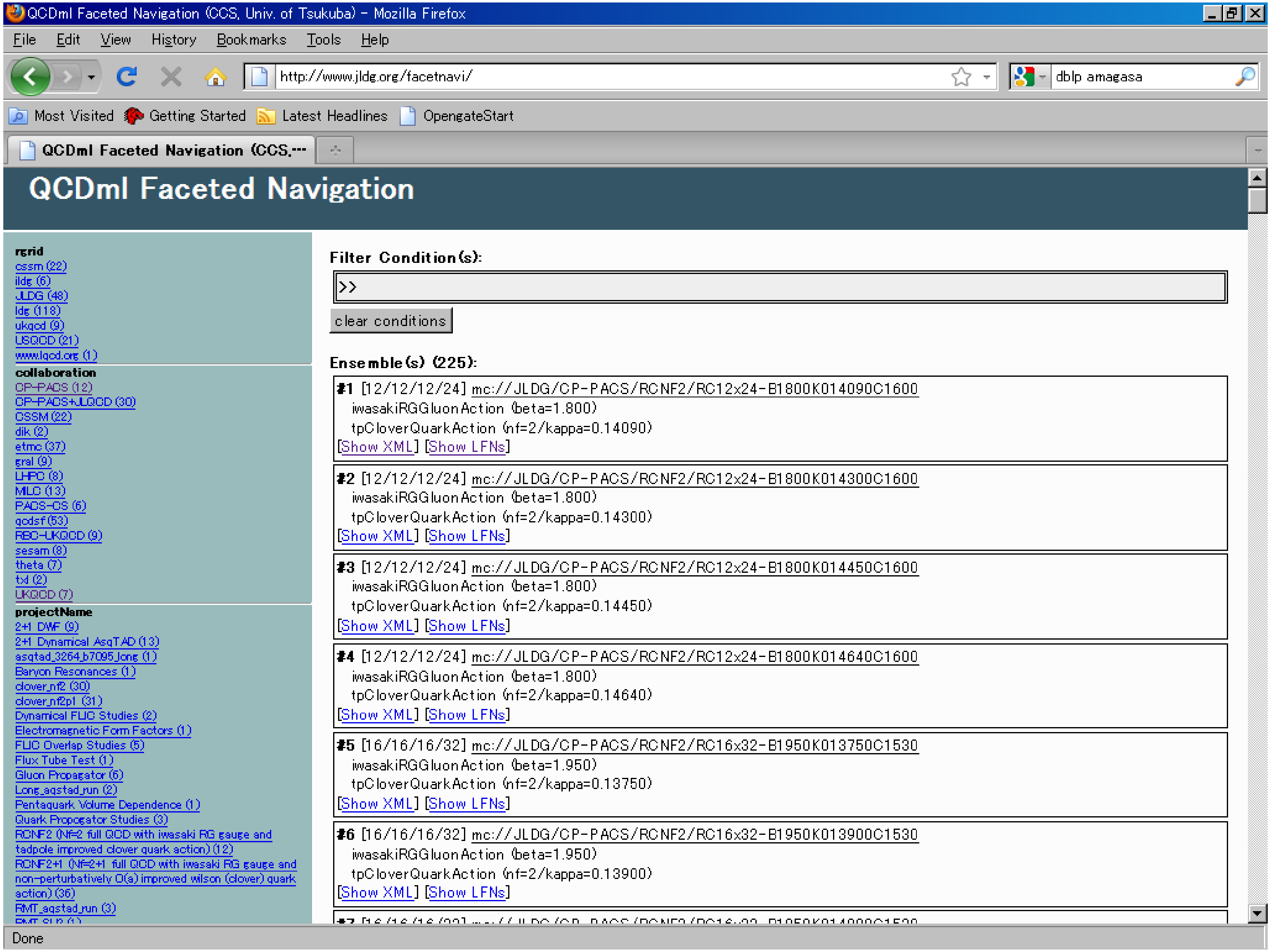
Facet extraction (XQuery)

QCDml
Ensemble (ILDG)
& Configuration (JLDG)

XML DB (eXist)

Downloading Ensemble XML

Demonstration


QCDml Faceted Navigation (CCS, Univ. of Tsukuba) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.jldg.org/facetnavi/
dblp amagasa
Most Visited
Getting Started
Latest Headlines
OpengateStart
QCDml Faceted Navigation (CCS,...

QCDml Faceted Navigation

rgrid
- cssm (22)
- ildg (6)
- JLDG (48)
- ldg (118)
- ukqcd (9)
- USQCD (21)
- www.lqcd.org (1)

collaboration
- CP-PACS (12)
- CP-PACS+JLQCD (30)
- CSSM (22)
- dik (2)
- etmc (37)
- gral (9)
- LHPC (8)
- MILC (13)
- PACS-CS (6)
- qcdsf (53)
- RBC-UKQCD (9)
- sesam (8)
- theta (7)
- txl (2)
- UKQCD (7)

projectName
- 2+1 DWF (9)
- 2+1 Dynamical AsqTAD (13)
- asqtad_3264_b7095_long (1)
- Baryon Resonances (1)
- clover_nf2 (30)
- clover_nf2p1 (31)
- Dynamical FLIC Studies (2)
- Electromagnetic Form Factors (1)
- FLIC Overlap Studies (5)
- Flux Tube Test (1)
- Gluon Propagator (6)
- Long_aqstad_run (2)
- Pentaquark Volume Dependence (1)
- Quark Propogator Studies (3)
- RCNF2 (Nf=2 full QCD with iwasaki RG gauge and tadpole improved clover quark action) (12)
- RCNF2+1 (Nf=2+1 full QCD with iwasaki RG gauge and non-perturbatively O(a) improved wilson (clover) quark action) (36)
- RMT_aqstad_run (3)

Filter Condition(s):

>>

Ensemble(s) (225):

#1 [12/12/12/24] mc://JLDG/CP-PACS/RCNF2/RC12x24-B1800K014090C1600
iwasakiRGGluonAction (beta=1.800)
tpCloverQuarkAction (nf=2/kappa=0.14090)
[Show XML] [Show LFNs]

#2 [12/12/12/24] mc://JLDG/CP-PACS/RCNF2/RC12x24-B1800K014300C1600
iwasakiRGGluonAction (beta=1.800)
tpCloverQuarkAction (nf=2/kappa=0.14300)
[Show XML] [Show LFNs]

#3 [12/12/12/24] mc://JLDG/CP-PACS/RCNF2/RC12x24-B1800K014450C1600
iwasakiRGGluonAction (beta=1.800)
tpCloverQuarkAction (nf=2/kappa=0.14450)
[Show XML] [Show LFNs]

#4 [12/12/12/24] mc://JLDG/CP-PACS/RCNF2/RC12x24-B1800K014640C1600
iwasakiRGGluonAction (beta=1.800)
tpCloverQuarkAction (nf=2/kappa=0.14640)
[Show XML] [Show LFNs]

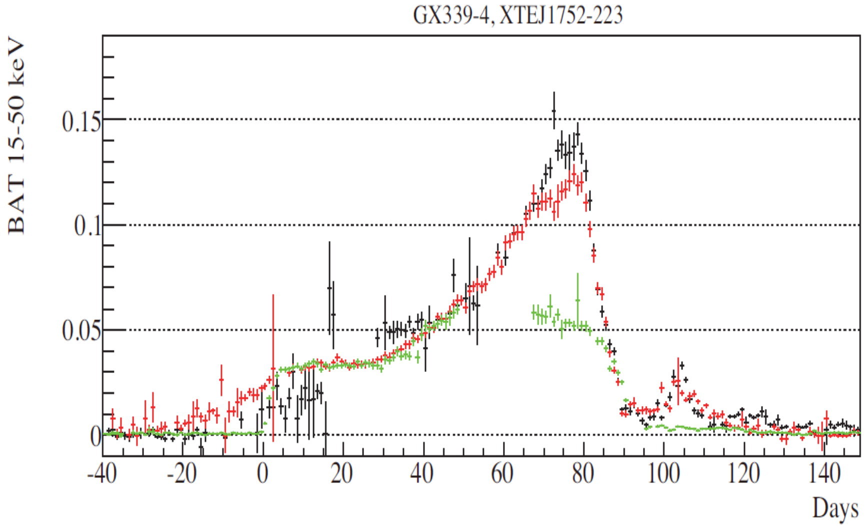
#5 [16/16/16/32] mc://JLDG/CP-PACS/RCNF2/RC16x32-B1950K013750C1530
iwasakiRGGluonAction (beta=1.950)
tpCloverQuarkAction (nf=2/kappa=0.13750)
[Show XML] [Show LFNs]

#6 [16/16/16/32] mc://JLDG/CP-PACS/RCNF2/RC16x32-B1950K013900C1530
iwasakiRGGluonAction (beta=1.950)
tpCloverQuarkAction (nf=2/kappa=0.13900)
[Show XML] [Show LFNs]

Done

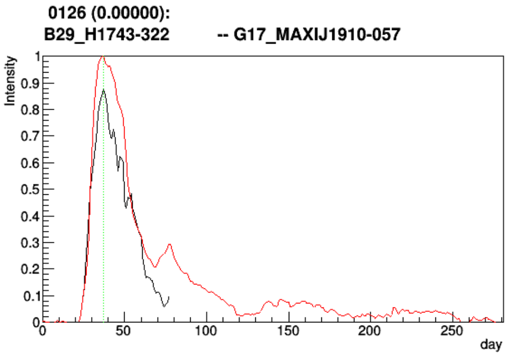
Similarity Search over Light Curves of X-ray Outbursts

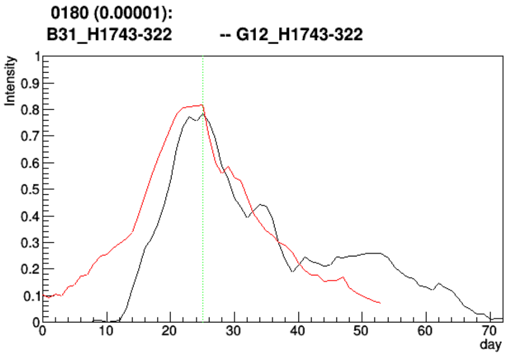
- A collaborative work with JAXA.
- X-ray outbursts
  - Phenomena in which X-ray emission from a celestial object grows for a certain time period.
- Researchers in JAXA are interested in finding celestial objects showing similar light curves.

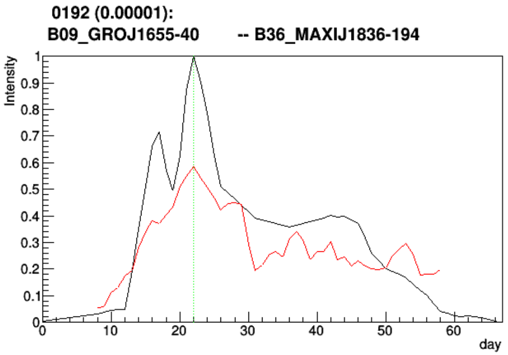


Our Work

- We apply time-series analysis techniques, such as DTW and DDTW, to search for similar light curves out of massive observation data.
- Detected light curves:







Major Funding

- Grant-in-Aid for Scientific Research from Ministry of Education, Culture, Sports, Science and Technology (MEXT) (~$1 Million; past 6 years)
  - Grant-in-Aid for Scientific Research A
  - Grant-in-Aid for Scientific Research on Priority Areas (Infoplosion Project)
  - Grant-in-Aid for Exploratory Research
  - Grant-in-Aid for Young Scientists
- MEXT Big Data Federation Feasibility Study (~$0.3 Miillion; 2013)
- From industry

Collaboration

- Industries
  - Hitachi
  - NEC
  - Fujitsu Lab.
  - Mitsubishi Electric
  - NTT Lab.
  - KDDI Lab.

  etc.
- International
  - Carnegie Mellon University
  - Chinese Univ. of Hong Kong
  - Georgia Institute of Technology

  etc.

Publication and Awards

- Refereed Papers
  - 2008: 29 (Journal 8, Conference 21 (3 Demo/Posters))
  - 2009: 28 (Journal 14, Conference 14 (1))
  - 2010: 22 (Journal 7, Conference 15)
  - 2011: 16 (Journal 4, Conference 12 (3))
  - 2012: 21 (Journal 8, Conference 13 (3))
  - 2013: 18 (Journal 4, Conference 14 (4))
- Awards
  - 4 Best Paper Awards (IEICE Trans., iiWAS2010, IPSJ SIG, DBSJ Journal)
  - 4 Best Student Paper Awards (WAIM2008, iiWAS2008, KMIS2010)
  - Contribution Award (IEICE)
  - 26 Students' Awards

Future Plan

- Research and Development for Data Engineering Challenges
  - Data integration framework to accommodate Big Data.
  - Big Data analysis challenges.
  - New issues involving social media and open data: privacy, social readings, LOD (Linked Open Data).
- Database Applications in Science Domains
  - Started collaboration with Biological Science Group through the dual degree program on gene databases
- Reinforcement of cooperation with other divisions and organizations

Thank you.