Data Intensive Research Project(s) at ITRI/AIST

Jason H. Haga
Isao Kojima
National Institute of Advanced Science and Technology (AIST), JAPAN
Background on AIST

• National Institute of Advanced Industrial Science and Technology, Japan
  – Mission: Contribute to society through continuous advancement in technologies and support to Japanese industries
  – Supported by METI (Ministry of Economy, Trade and Industry)

• Established in 2001
  – Merging 15 different research institutes
    • Oldest is Geological Survey of Japan (est. 1882)
    • Set/maintain the kilogram calibration standard of Japan

• AIST ranked 7th in “Top 20 Japanese research institutions for all field”, Thomson Reuters, 2014
GEO Grid
Grid-based e-infrastructure for geosciences

- Researchers (foreign nationals) ........................................ 2,258 (96)
  [Permanent] ........................................ [1,928]
  [Fixed term] ........................................ [330]
- Administrative employees (foreign nationals) ........................................ 675 (1)
  Total number of employees: 2,933 (97)
- Executives (full time) ........................................ 13
- Visiting researchers ........................................ 159
- Postdoctoral researchers ........................................ 200
- Technical staff ........................................ 1,441

(As of April 1, 2015)

Number of researchers accepted through industry/academia/government partnerships
- Companies ........................................ 1,774
- Universities ........................................ 1,852
- Other organizations ........................................ 972
  (foreign nationals: 426)
  (Total number of researchers accepted in FY 2013)

Applications from outside Japan is highly recommended
Location of AIST

- Tsukuba (science) City
  - Government planned city
    - Est. in 1962
    - 1 university, 2 colleges
    - About 30 governmental research institutes including JAXA, KEK, NIMS
    - About 30~40 company labs
  - 60km Northeast from Central Tokyo
    - 45min with Tsukuba Express (TX)
Geographical Survey Institute

University of Tsukuba

JAXA
Japan Aerospace Exploration Agency

Mt. Tsukuba
876m

Aerial Tram or Funicular or Walk

Tsukuba Station

Geological Survey of Japan (GSJ) of AIST

Information Technology Research (ITRI) of AIST

AIST

Sakura-kan Guest House
Research at AIST

- 7 major research areas

- Environment and Energy
- Life Science and Biotechnology
- Information Technology and Human Factors
- Materials and Chemistry
- Electronics and Manufacturing
- Geological Survey of Japan
- National Metrology Institute of Japan

Good for Cross-Domain Research
Cross-Domain/Interdisciplinary R&Ds

GeoScience + IT
Bioscience + IT
Mechanics + IT
Etc.
Data Integration Question
(same as last year)

What knowledge can be obtained by integrating following data?

1. Geological Map
   - Geological Survey of Japan is a part of AIST
     - sedimentary rocks,
     - volcano rocks,
     - granitic rocks etc.

2. 3D Elevation Model
   - Created by our ASTER Satellite
     - Produce 3D-model by stereo-matching

3. Real Time Rain Sensors
   - Provided by JMA(japan meteorological agency)
Answer: Hazard Map for Landslide
(One typical application of GEO Grid)

Key R&D Technologies
1. Distributed Database Integration (Linked Data/Heterogeneous DB etc.)
2. Data Mining & Simulation on the Cloud (Neural-Net, Machine Learning)
3. Multi-Screen Visualization (Tiled Wall Software)

Visualize Hi-Resolution Hazard Map
Geoscience + IT

GeoGrid: An Example of Data Intensive Research Projects at ITRI
What is GEO Grid?

**GEO = Geospatial**

**Grid = Grid (cloud) Computing**

*e-Science infrastructure on heterogeneous data archives*

- Cross-Domain (joint) project from 2004

**Geospatial Contents**
- Satellite Data
- Geology Data
- Various Maps

**Advanced IT**
- Distributed DB
- HPC/Cloud

IT/CS Units in AIST

Geology/Environment Units in AIST.

**Core archive contents: Our Satellite Sensor Data**
- ASTER satellite images >= 200TB (2000,000 scenes, y2000->)
- Now extending to manage (Petabyte-Scale) PALSAR, PRISM, Landsat8 etc.

**Core technologies: Grid Based => Parallel/Distributed R&D**
- Distributed file system: Gfarm (started at AIST, Now at Tsukuba-U)
- Database Integration: OGSA-DAI@Uk /Distributed SPARQL
- Tsukuba-GAMA: Integrated Credential(Authentication) Management(some codes are included in MyProxy)
Major Technical Achievements of GEO Grid in 10 years

- Petabyte-Class Large Scale Data Archive & Analysis
  - Gfarm
- Single sign-on system using Grid Security
  - Tsukuba-GAMA
- Heterogeneous Metadata Management based on OGC Standard
  - AIST-CSW
- Service-based Distributed Database Access
  - OGSA-DAI(Web Services)
Data Archives

- **ASTER sensor on NASA Terra satellite (2000~)**
  - Resolution (Mid-range): 15m (VNIR), 30m (SWIR), 90m (TIR) / px
  - 60km wide
    - 50~60GB daily Level 0 data transfer from NASA to JAPAN
  - 16 day observation cycle
    - Good for detecting long range change (= large computation)
  - 2 cameras with different angles
    - Can create DEM (Digital Elevation Model) by stereo matching

- **Landsat-8 (by USGS)**
  - Latest earth observation satellite launched 2013
    - 15m/Pan 30m/Color
    - 16 day observation cycle for the same area
  - Free and Open!

- **AIST set up the ground station for Landsat-8 (with Tokai-U)**
  - Receives the daily data directly from the satellite
  - Can publish the data to the Internet in semi real-time
    - 2 hrs in AIST by our high performance computing (1 day in USGS)
3.11 science data examples produced by the GEO Grid

Ground move with radar(SAR) satellite

Flood simulation

ASTER images with 3D DEM

GEO Grid archive/cluster is also damaged by 3.11 earthquake

We evacuated our environment using cloud technology and continued to process data in collaboration with OCC/SDSC/NCHC etc.
Public Service http://landsat8.geogrid.org

Latest/Historical Data can be Downloaded and Viewed

User Contributions
Like “I found interesting things!” by Facebook
Constellation

• ASTER = 16 days cycle
• Landsat-8 = 16 days cycle
• ASTER+Landsat-8 = 8 days cycle (same orbits)

Target: Daily change detection

– Example: Skybox (which is acquired by Google) has a plan to launch 20 satellites

We are investigating to do the same thing with existing (and new) satellites
Analysis

Workflow engine: Lavatube
Machine Learning System: Hivemall
Our yet another Workflow Engine: **Lavatube** for spacio-temporal data and image/movie processing

1. **Support rest-based OGC (OpenGeospatial Consortium) services**
2. **Support various image/movie processing modules**
3. **Provide High-Level interaction**

Browser Interface (HTML5) or Windows engine

OGC Services
- Metadata Search
- Database Access
- Processing
• Hivemall: Scalable Machine Learning Library for Apache Hive
• A collection of machine learning algorithms as Hive UDFs/UDTFs
  – Classification & Regression
  – Recommendation
  – k-Nearest Neighbor Search
• An open-source project on Github
  – Licensed under LGPL
  – github.com/myui/hivemall (bit.ly/hivemall)
Application

Radiation Monitoring Database for Fukushima
Radiation Monitoring Data is important to:
- Understand what happened at the accident in the past
- Help the decision making for the future

National Project to continuously Monitor/Construct/Publish Radiation Monitoring Database of Fukushima Area

Project Overview

- Database Construction
- Database Publishing
- Radiation Monitoring

Project Structure as of 2013

- Nuclear Regulation Authority (NRA), JAPAN
- Project Management
- Japan Atomic Energy Agency (JAEA)
- JAEA
- Hokkaido University
- Japan Map Center (Company)
- Partners
  - Okayama-U,
  - The Institute of Statistical Mathematics,
Example Data Integration Application using OGC specs
Combine SOS (Sensor Observation Service) with other WMS (Web Map Service) data source (Weather)

Human exposure to natural background radiation, 0.27uSv/h

Jan 1, 2013
The dose rate was relatively high

Jan 16, 2013
The dose rate was relatively low when there was heavy snow

Snow effect

Simple overlay can be useful
Directions

Social, Mobile & Crowdsourcing
Data Integration Issue: Administrative & Non-Administrative Data

- **Administrative Data (Current GEO Grid Data)**
  - Governmental & official data
  - Limited amount with controlled quality

- **Non-Administrative Data**
  - NPO, Social media, crowdsourcing (Twitter, etc.)
  - Large amount, variable quality
**Application Examples**

1. **Spatiotemporal & Thematic Analysis**
   - Mapping, clustering, and regression: (2011/03/02-2011/03/24)
   - 1) earthquake-related tweets
   - 2) tsunami-related tweets
   - 3) nuclear-related tweets

2. **mTrend (ACMGIS2011 Demonstration)**
   - Comparing thematic patterns in correlated tweets by keyword streams

3. **Cyber-Physical Data Cloud: An Infrastructure for Interconnecting Heterogeneous Sensor Data (WTP2012 Demonstration)**
   - Situation creation on the basis of intersection area of outbreaks between tweets and natural phenomena

You can find the “Trends” for each words

Time & Space mapping of Tweets
Crowdsourcing and Notifications
Joint Research with Taiwan ITRI
Linked Open Data

Federated SPARQL with
“Best-Effort” Query Processing
Linked Open Data (LOD)

Try to create a huge linked knowledge cloud

- **The data is written with RDF** *(Resource Description Framework)*
  - The Standard for the Semantic web community
- **Highly distributed and rapidly increasing**
  - More than 300 sites, billion~trillion triples
- **Cross-Domain**
  - GEO, BIO, Government, Media, etc
  - Many governmental data is going to be published as LOD
- **Issues**
  - Distributed SPARQL processing may be slow, but a centralized data service lacks freshness
  - Heterogeneity with SPARQL Endpoints, plain RDF Texts

Our Approach
Hybrid Adaptive Query Processing

Based on the **freshness, coverage** and the **response time**
Adaptive Query optimization

Pre-defined Query Processing Schedule

- Network delays,
- Too many results
- Site troubles, etc...

Modified Processing Adaptively

Query 1 (Result size = 150):
```
select * where {
  ?x dbp:reference ?ref . 777,679
  ?x rdf:comment ?comment . 10,000
  ?x skos:subject ?subj . 9971
  ?x foaf:page ?page . 10,000
  ?x rdf:type ?type . 800,000
  FILTER ( regex(str(?subj),"building") )
}
```

Query 2 (Result size = 8):
```
select * where {
  ?x dbp:reference ?ref . 777,679
  ?x rdf:comment ?comment . 10,000
  ?x skos:subject ?subj . 9971
  ?x foaf:page ?page . 10,000
  ?x rdf:type dbp:book 3105
}
```

Query 3 (Result size = 8):
```
select * where {
  ?x dbp:reference ?ref . 777,679
  ?x rdf:comment ?comment . 10,000
  ?x skos:subject ?subj . 9971
  ?x foaf:page ?page . 10,000
  ?x rdf:type dbp:book 3105
  ?x dbp:releaseDate ?date (DBP) 126,737
}
```

Query 4 (Result size = 13):
```
select * {
  ?book owl:sameAs ?link (DBP) 10,121,699
}
```

Achieve good performance around 10 distributed SPARQL endpoints (still small for 300 ;-<)
GEO Grid
Grid-based e-infrastructure for geosciences

User’s SPARQL Query

```sparql
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbp: <http://dbpedia.org/resource/property/>
SELECT * WHERE {
  ?x dc:subject dbp:FIFA_World_Cup-winning_countries .
  ?p foaf:name "Luiz Felipe Scolari"@en .
}
```

Hybrid & Adaptive Query Processing

- Evaluation
- Query Result

You can get rough answer in 10 seconds,
or
More accurate answer with a long time
BioScience + IT

BIO-CAD/LEAD

Hydra: Molecular Visualization
High Performance Genomics Assembly

- **Next Generation Sequencers**
  - Huge set of short reads are obtained
    - 1 read: ATGC (base) 100\(^{50\text{bases} \times 2}\)
    - Total: 100 million reads just for 1 run

- **Hybrid Assembly Workflow**
  - MPI parallelized (SAET, ASiD)
  - Improve the algorithms (Velvet)
  - To achieve scalability and performance enhancements
Hydra Molecular Visualization

- Create a more device agnostic tool
  Visualization of multiple protein-ligand interactions
Mechanics + IT

Media-related R&D
Media-Related R&D

- IT behind the robot
  - Computer Singing Systems
    - VocaListener
    - VocaWatcher

- Active Music-Listening Web Service
  - Songrium
Summary

• AIST has many interdisciplinary data-oriented R&D projects
  – Geospatial
  – Linked Data
  – Bioinformatics
  – Multimedia (Music/Songs)

• Looking forward to the OSDC students contribution

• AIST YouTube: https://www.youtube.com/user/aistchannel
Acknowledgements

Isao Kojima
Kyoungsook Kim
Steven Lynden
Hirotaka Ogawa
Tsutomu Ikegami
Yuan Zhao