AIST

GEO Grid

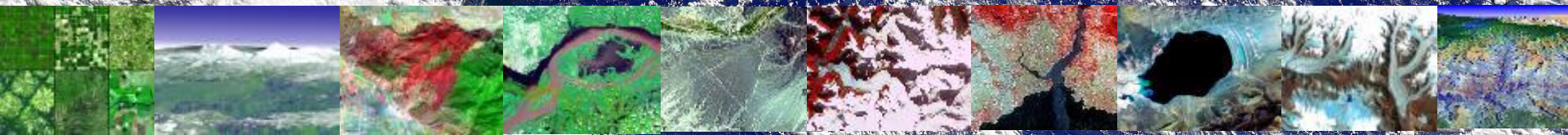
Grid-based e-infrastructure for geosciences

Data Intensive Research Project(s) at ITRI/AIST

Isao Kojima

National Institute of Advanced Science and Technology(AIST),JAPAN

isao.kojima@aist.go.jp

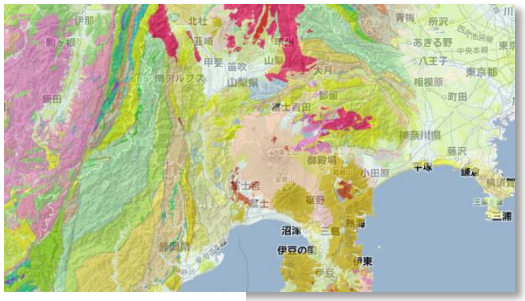


Question

This is the same question as in the last year. Sorry!

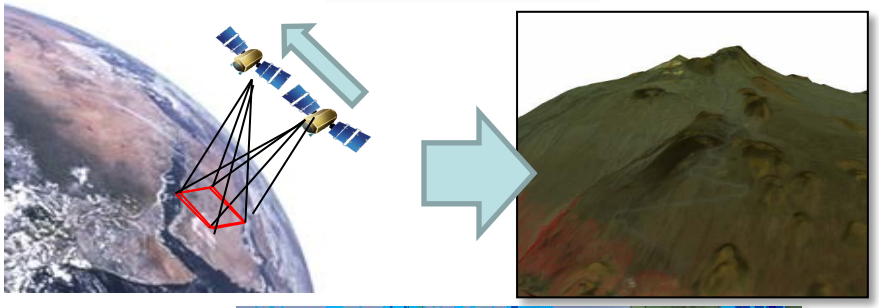
Data Integration Question

What knowledge can be obtained by integrating following data?



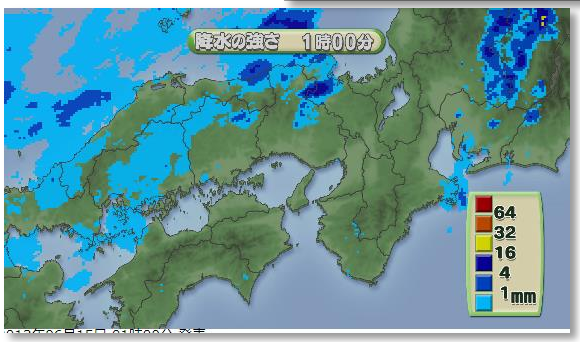
1. Geological Map

- Geological Survey of Japan is a part of AIST
  - sedimentary rocks,
  - volcano rocks,
  - grantic rocks etc.



2. 3D Elevation Model

- Created by our ASTER Satellite
  - Produce 3D-model by stereo-matching



3. Real Time Rain Sensors

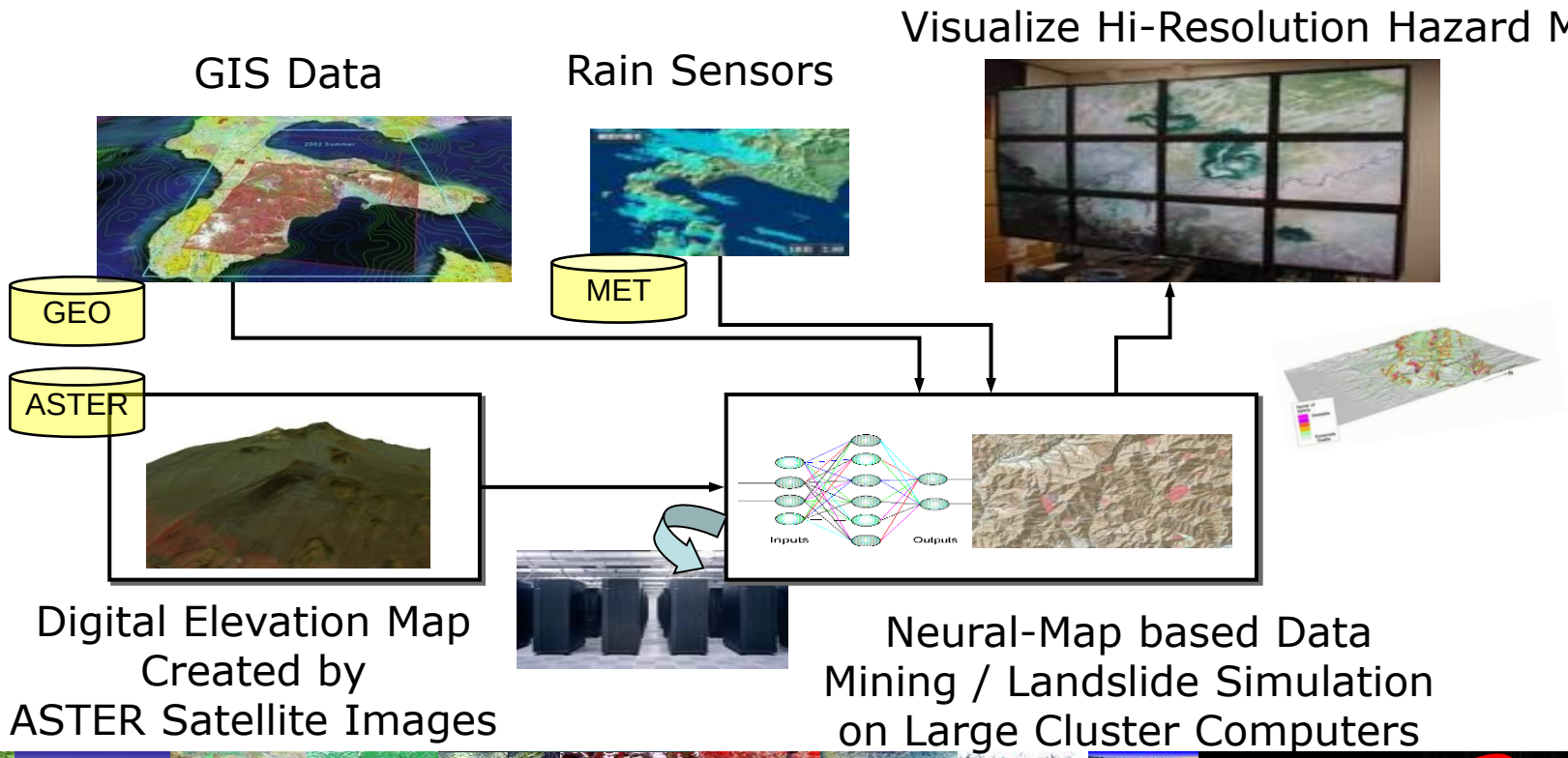
- Provided by JMA(japan meteorological agency)

Answer: Hazard Map for Landslide

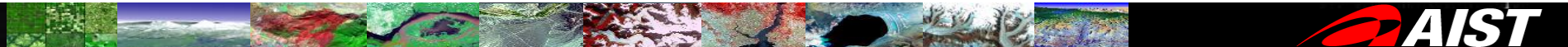
(One typical application of GEO Grid)

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)**



What is AIST?



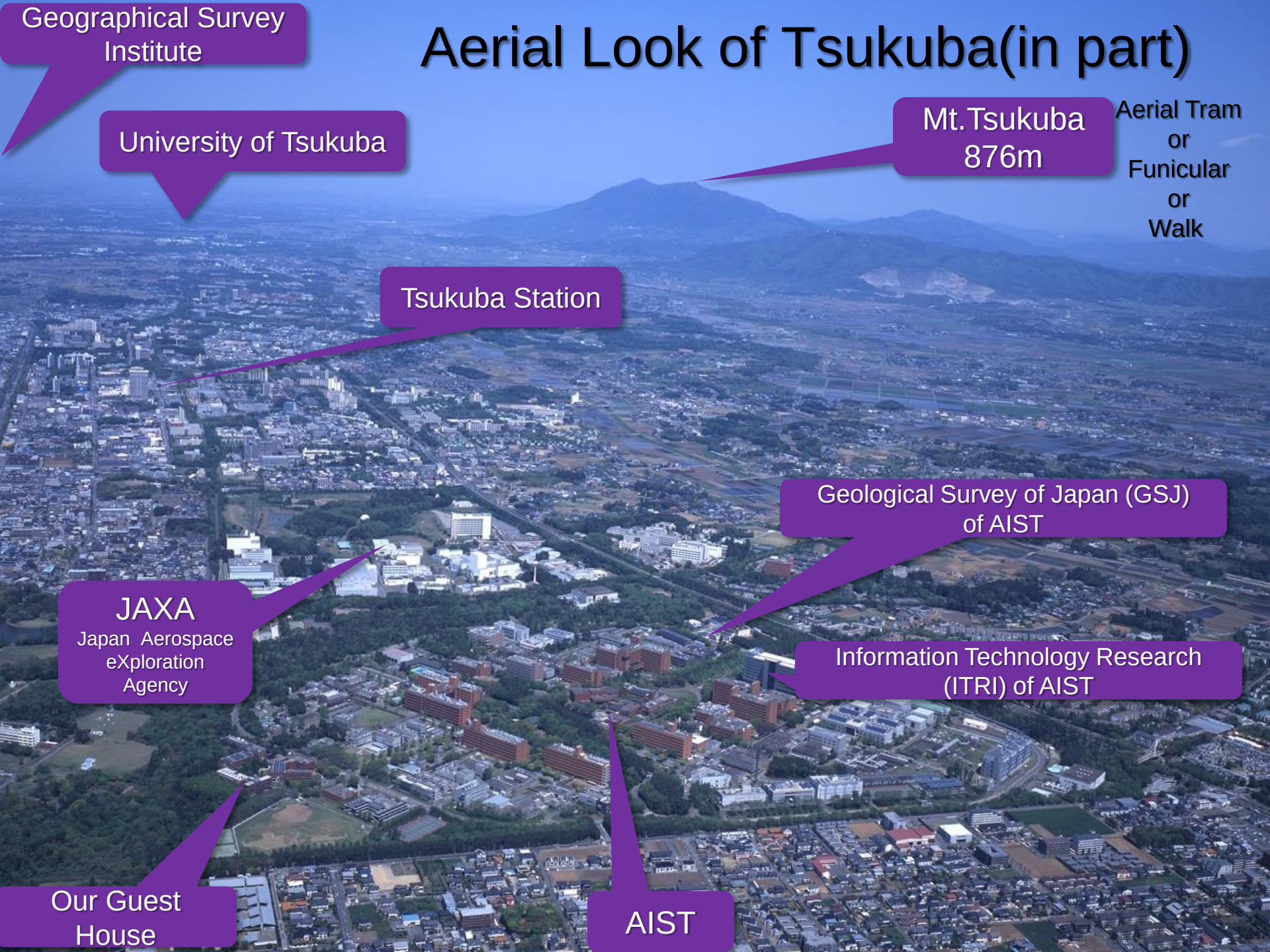
1. What AIST is?

- National Institute of **A**dvanced **I**ndustrial **S**cience and **T**echnology, JAPAN
  - Supported by METI(Ministry of Economy, Trade and Industry)
  - Our Important Mission = Bridge between basic research and product development
- Established at 2001
  - By merging **15!** research institutes under METI
    - Electric, Machinery, Chemistry, Material, Bio, etc.
  - Many of them had 50~100 years history
    - Oldest= Geological Survey of Japan(est. 1882)

2. Where is AIST?

- Tsukuba-(science)-City
  - Planned (artificial) city created by the government
    - Started to construct 1962~> Now 50 years
  - Current Status
    - 1 university, 2 colleges
    - About 30 governmental research institutes
      - » including AIST,JAXA,KEK
    - About 30~40 company labs.

- 60km Northeast from Central Tokyo
  - 45min with Tsukuba-Express
    - We call it "TX":
    - 130km/h high speed commuter train


Aerial Look of Tsukuba(in part)
Geographical Survey Institute
University of Tsukuba
Mt.Tsukuba 876m
Aerial Tram or Funicular or Walk
Tsukuba Station
Geological Survey of Japan (GSJ) of AIST
JAXA
Japan Aerospace eXploration Agency
Information Technology Research (ITRI) of AIST
Our Guest House
AIST

3. What area does AIST cover?

Almost everything about science & technology☺

– By merging 15 research labs under METI

- 6 Major research fields

Good for Cross-Domain Research

Organization Chart

GEO Grid

Grid-based e-infrastructure for geosciences

- Research Center(20)
- Research Institute(21)
- Research Laboratory(1)

( ) : the number of units

President

Auditor

Senior Vice-President
Vice-President

Emeritus Researcher

Honorary AIST Fellow

AIST Fellow

AIST Hokkaido
AIST Tohoku
AIST Tsukuba
Tsukuba Central 1,2,3,4,5,6,7
Tsukuba West
Tsukuba East
AIST Tokyo Headquarters
AIST Tokyo Waterfront
AIST Chubu
AIST Kansai
AIST Chugoku
AIST Shikoku
AIST Kyushu

Environment and Energy

Director-General・Deputy Director-General・Research Planning Office

- Biomass Refinery Research Center
- Research Center for Hydrogen Industrial Use and Storage
- Research Center for New Fuels and Vehicle Technology
- Methane Hydrate Research Center
- Research Center for Compact Chemical System
- Advanced Power Electronics Research Center
- Research Center for Photovoltaic Technologies
- Research Institute for Ubiquitous Energy Devices
- Research Institute for Environmental Management Technology
- Research Institute for Innovation in Sustainable Chemistry
- Energy Technology Research Institute
- Research Institute of Science for Safety and Sustainability

Life Science and Biotechnology

Director-General・Deputy Director-General・Research Planning Office

- Research Center for Medical Glycoscience
- Computational Biology Research Center
- Biomedical Information Research Center
- Research Center for Stem Cell Engineering
- Health Research Institute
- Bioproduction Research Institute
- Biomedical Research Institute
- Human Technology Research Institute

International Patent Organism Depositary

Information Technology and Electronics

Director-General・Deputy Director-General・Research Planning Office

- Network Photonics Research Center
- Digital Human Research Center
- Spintronics Research Center
- Center for Service Research
- Flexible Electronics Research Center
- Intelligent Systems Research Institute
- Information Technology Research Institute
- Nanoelectronics Research Institute
- Electronics and Photonics Research Institute
- Research Institute for Secure Systems

Innovation Center for Advanced Nanodevices

Nano-technology, Materials and Manufacturing

Director-General・Deputy Director-General・Research Planning Office

- Nanotube Research Center
- Research Center for Ubiquitous MEMS and Micro Engineering
- Advanced Manufacturing Research Institute
- Materials Research Institute for Sustainable Development
- Nanosystem Research Institute
- Diamond Research Laboratory

Metrology and Measurement Science

Director-General・Deputy Director-General・Research Planning Office

- Measurement Solution Research Center
- Metrology Institute of Japan
- Research Institute of Instrumentation Frontier

Metrology Management Center

Geological Survey and Applied Geoscience

Director-General・Deputy Director-General・Research Planning Office

- Active Fault and Earthquake Research Center
- Institute for Geo-Resources and Environment
- Institute of Geology and Geoinformation

Geoinformation Center

Geological Museum

Planning Headquarters
Planning Office for Fukushima Research Base
Planning Office for Reorganization
Compliance Headquarters
Research and Innovation Promotion Headquarters
Planning Division
Intellectual Property Division
Collaboration Promotion Division
International Affairs Division
Division for Start-ups
International Standards Promotion Division
Tsukuba Innovation Arena Promotion Division
Innovation School
Research Environment and Safety Headquarters
Research Environment and Safety Planning Division
Safety and Environmental Management Division
Research Facilities Division
Intelligent information

University Like Structure
6 Research Areas
>40 Research Units
（Hundreds of Groups）

GEO Grid

Grid-based e-infrastructure for geosciences

Employees

[foreign nationals]

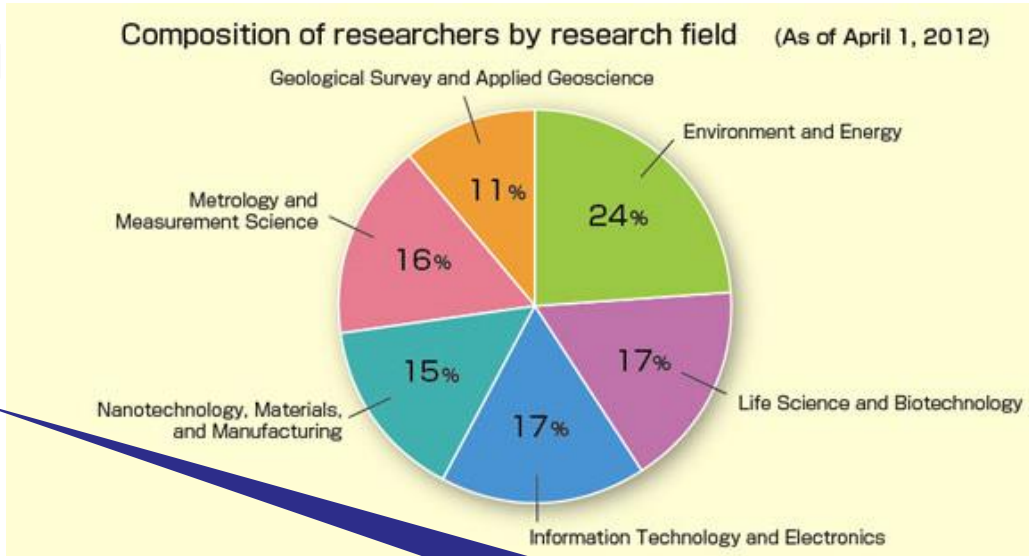
Researchers	**2,288 [82]**
●Permanent	2,046
●Fixed term	242
Administrative employees	661
Total number of employees	**2,949 [82]**
Executives	13
Visiting researchers	177
Postdoctoral researchers	274
Technical staff	1,553

Researchers accepted through industry-academia-government partnerships

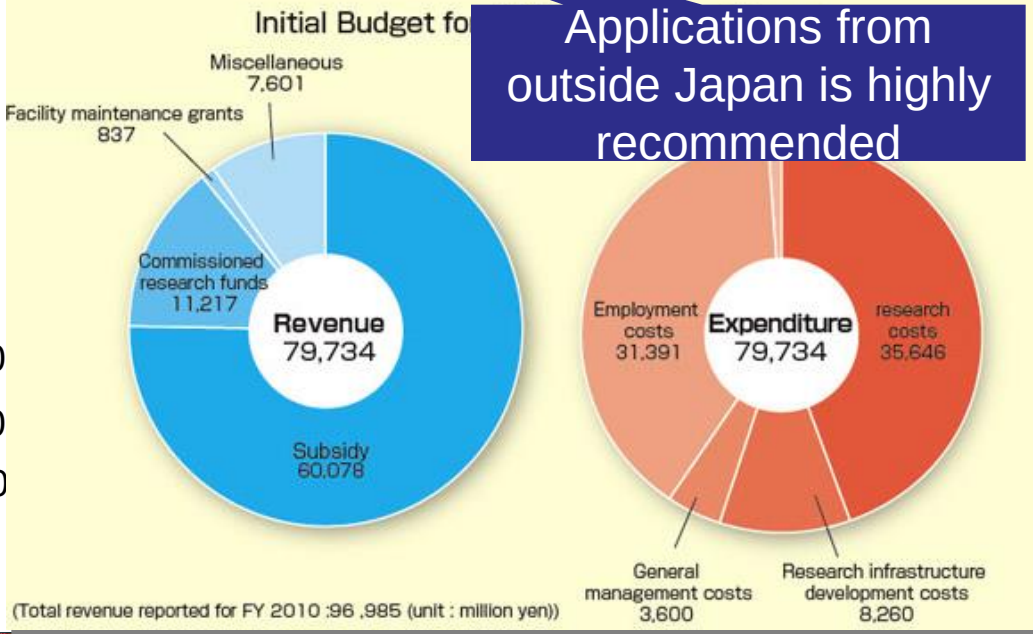
【553 from overseas】

From companies	Approx.1,500
From universities	Approx.2,000
From other organizations	Approx. 900

(NOTE : Total no. of researchers accepted in FY 2011)



Applications from outside Japan is highly recommended



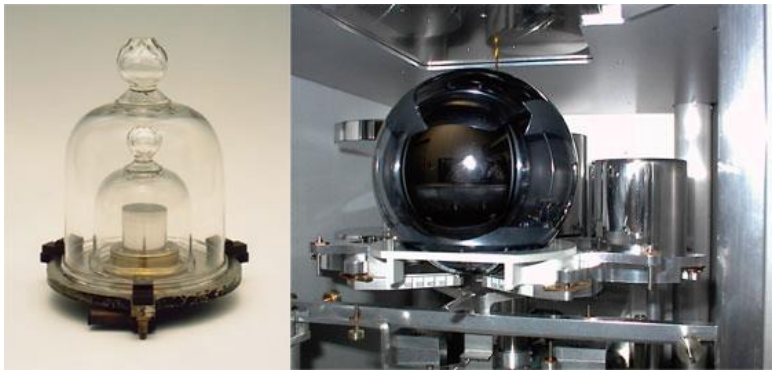
AIST

AIST (classic) aspects

- Metrology: = set/maintain the standards

  kilogram calibration

  weight of japan



- Geology:= explore/maintain (physical) objects

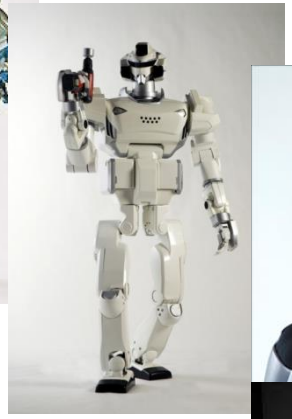
  our geological museum

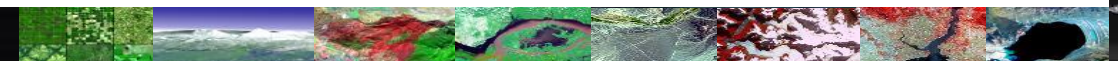
IT-related (modern) R&D example

- Robotics
  - PARO: Therapeutic robot

    On the GUINNESS book
    "World's Most Therapeutic Robot".

  - Humanoids
    - HRP-2(2002)
    - HRP-3(2005)
    - HRP-4C:Miim (2009)
    - HRP-4(2010, just 39kg)





Information Technology Research Institute (ITRI) =1/42? Research Units in AIST

Director **Tomohiro Kudoh**
Dupty Director **Kimiko Nishikimi**
Principal Research Scientist **Isao Kojima** Me
Prime Senior Researcher **Masataka Goto**

Infraware Research Group Leader: Hidemoto Nakada
Data Science Research Group Leader: Hirotaka Ogawa
Smartsystem Research Group Leader: Masahiro Murakawa
Geoinformatics Research Group Leader: Ryosuke Nakamura
Processware Research Group Leader: Kimiko Nishikimi
Media Interaction Research Group Leader: Masataka Goto

- One of research units in AIST.
- ≒100 Members
  - 48 regular researchers
  - 24 visiting /post-doc researchers
  - 36 office staffs, secretaries, SEs. developers, etc.

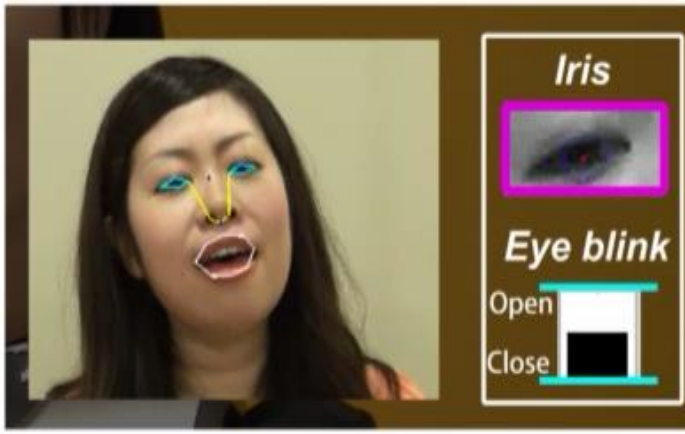
We have many Cross-Domain R&Ds

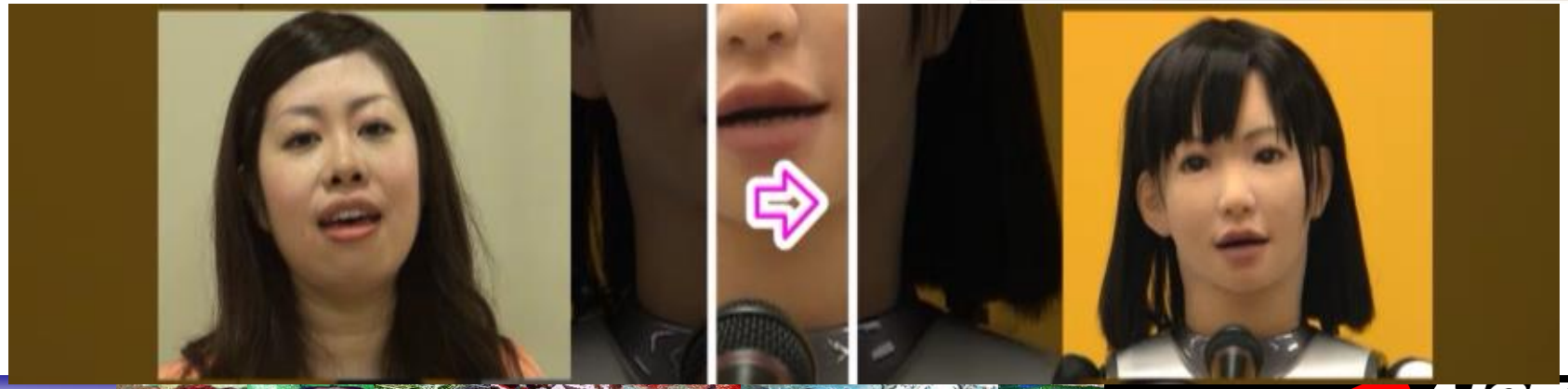
- Mechaniscs + IT = Robotics
- GeoScience + IT = GEO Grid
  (detailed later)
- Bioscience + IT = ,,,

One of popular ITRI Research Results

Background Technology behind the Robot

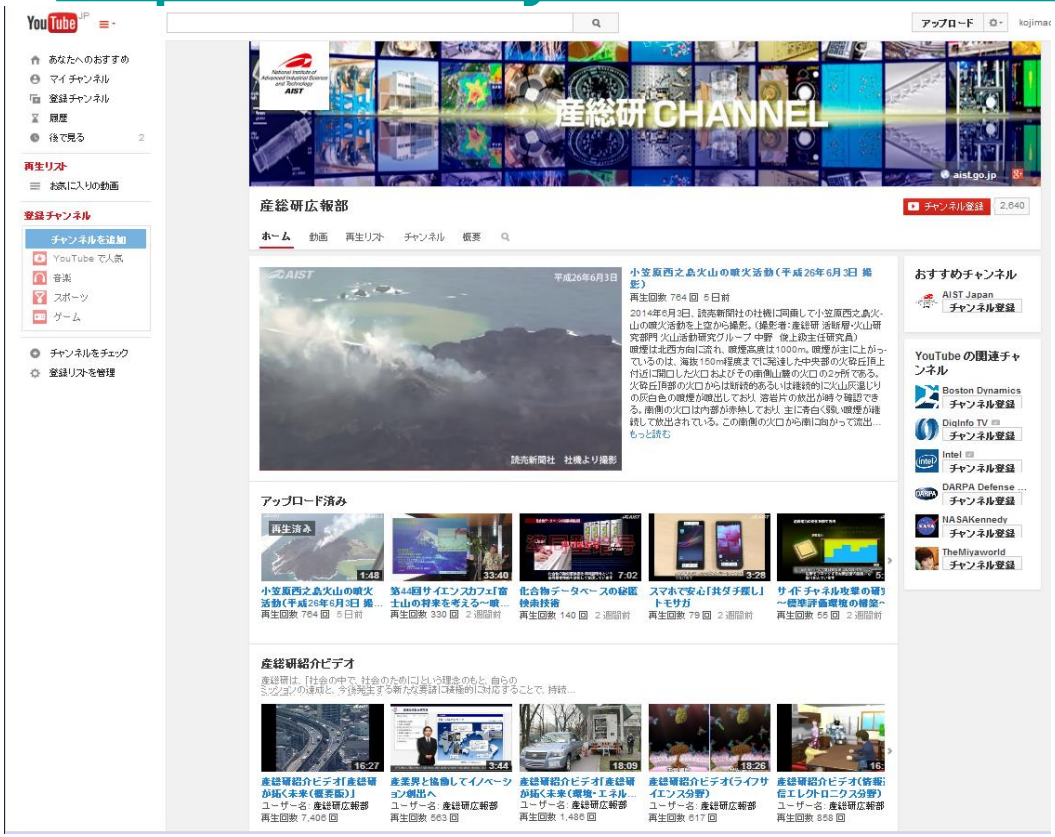
- VocaWatcher + VocaListener
  - Video analysis for mouth, eye, and neck motions
  - Require **neither** facial markers
    nor multiple cameras
  - Utilize phoneme timing
    provided by VocaListener
    to improve robot motions





AIST has our YouTube Channel

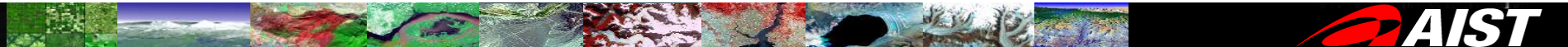
https://www.youtube.com/user/aistchannel



Mimu:https://www.youtube.com/watch?v=LjbyxekEo6s&list=TLeLhFd70f_XNzLMV5MENm-DbAnuwR8eUy&index=6

GEO Grid

An Example of Data Intensive Resarch Projects at ITRI



What is GEO Grid?

http://www.geogrid.org

GEO= Geospatial

Grid = Grid(cloud) Computing

e-Science infrastructure on heterogeneous data archives.

- Cross-Domain (joint) project from 2004(almost 10 years)

Geospatial Contents
Satellite Data
Geology Data
Various Maps

GEOGrid

Advanced IT
Distributed DB
HPC/Cloud

Geology/Environment Units in AIST.

IT/CS Units in AIST

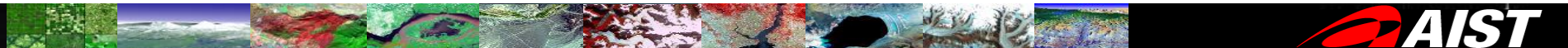
- **Core archive contents: Our Satellite Sensor Data**
  - ASTER satellite images >= 200TB(2000,000 scenes, y2000->)
  - Now extending to manage (Petabyte-Scale) PALSAR,PRISM,Landsat8 etc.
- **Core technologies: Grid Based => Parallel/Distributed R&D**
  - Distributed file system: Gfarm(started at AIST, 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)

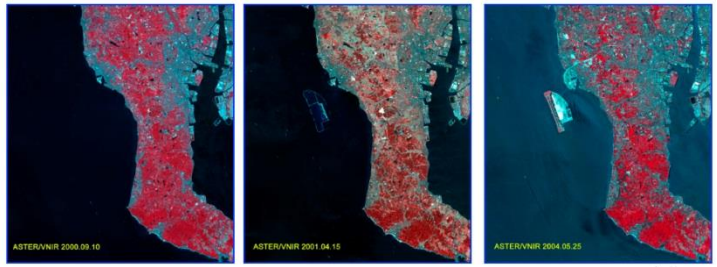
Data Archives

ASTER satellite & Landsat-8 satellite

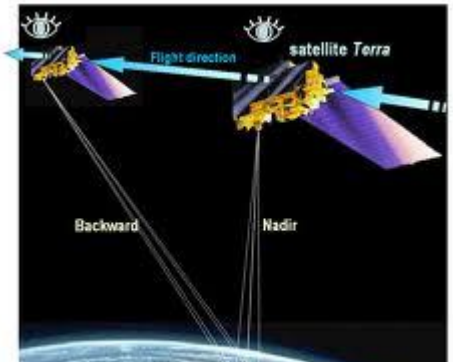


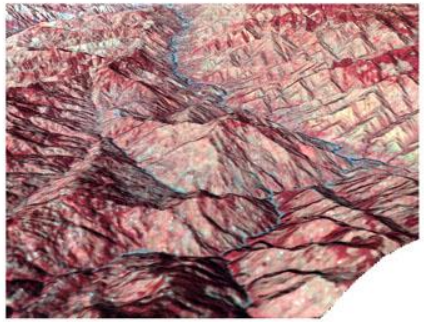
ASTER

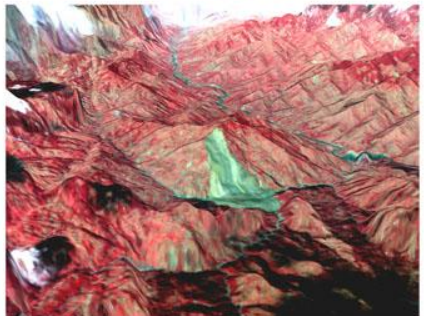
- Japan-made sensor on NASA Terra satellite (2000~)
  - Resolution(Mid-range):15m(VNIR),30m(SWIR),90m(TIR)/px
  - 60km wide
    - 50~60GB daily Level 0 data transfer from NASA to JAPAN
  - Observe the same area for every 16 days.
    - Good for detecting long range change(= large computation)
      - Urban development. Desertification



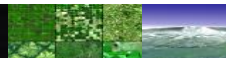
  - 2 cameras with different angles
    - Can create DEM(Digital Elevation Model) by stereo matching





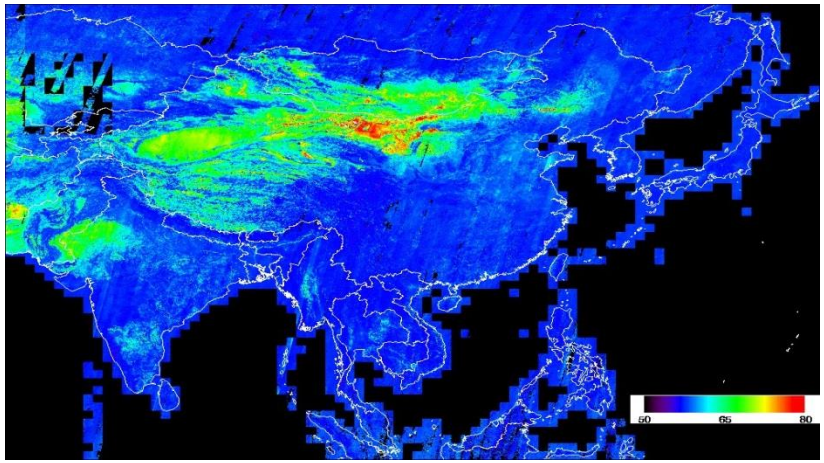


DEM of Pakistan Landslide 2005

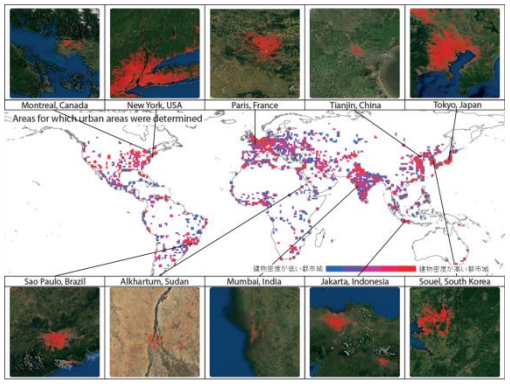


ASTER Science Result Examples

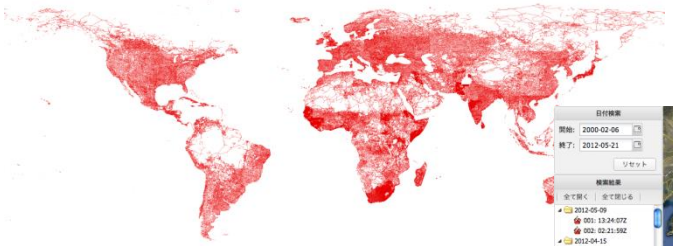
SiO2 Detection from ASTER



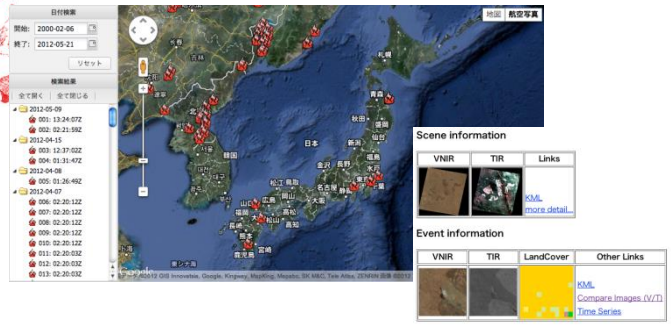
Urban Area Detection from ASTER/PALSAR



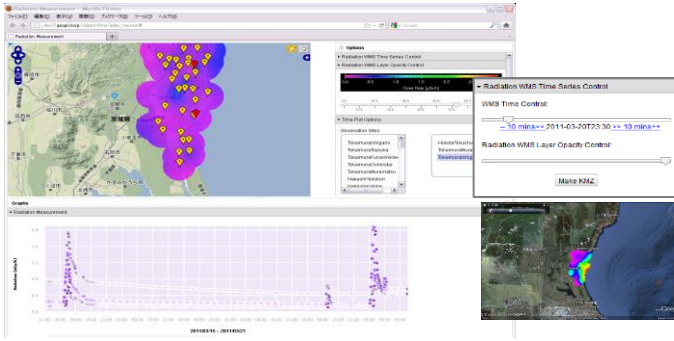
Road Detection from ASTER



HotSpot Alert System from ASTER

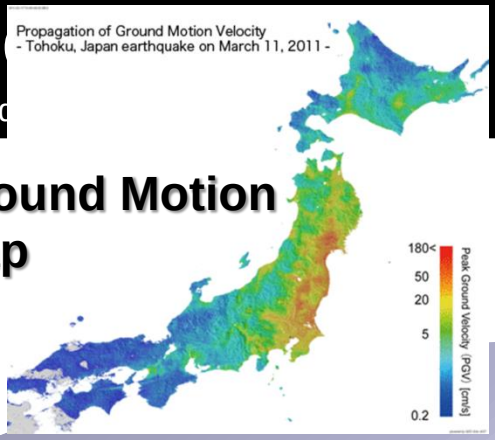


Radiation Monitoring after 3.11 (Sensor Fusion)



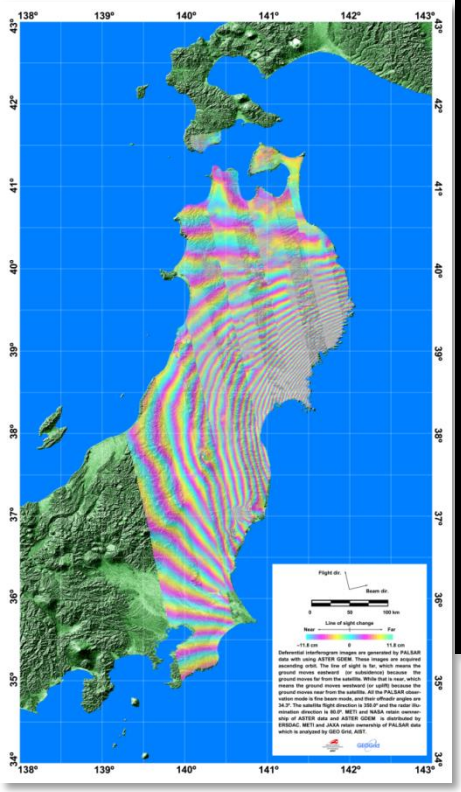
3.11 science data examples produced by the GEO Grid

Grid-based

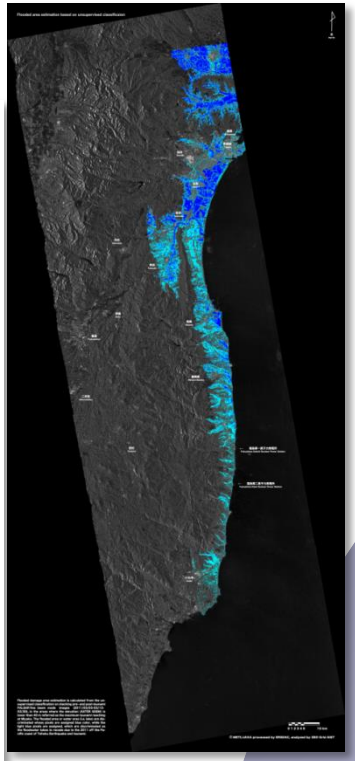


Ground Motion Map

Ground move with radar(SAR) satellite



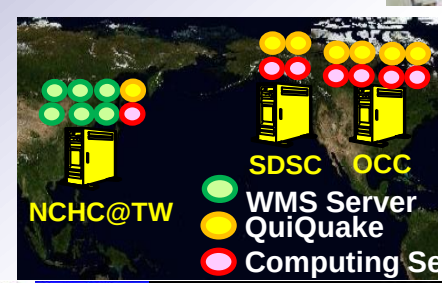
Flood simulation



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.



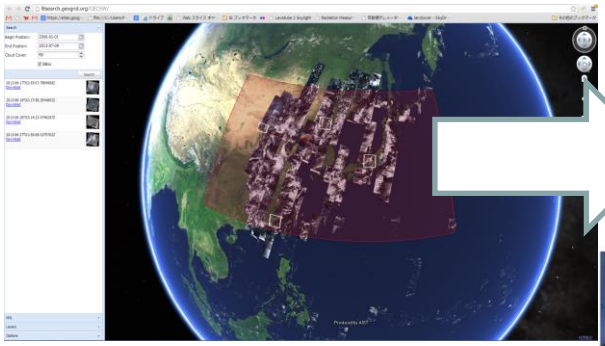
National Center for High-Performance Computing
UC San Diego
Open Cloud Consortium
CTC
Université Lille1
Orkney
NSPO
ITT
NTT データ CCS

AIST

Landsat-8 (by USGS)

- Latest earth observation satellite launched 2013.
  - 15m/Pan 30m/Color
  - 16days 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
    - 2hours in AIST by our high performance computing (1 day in USGS)



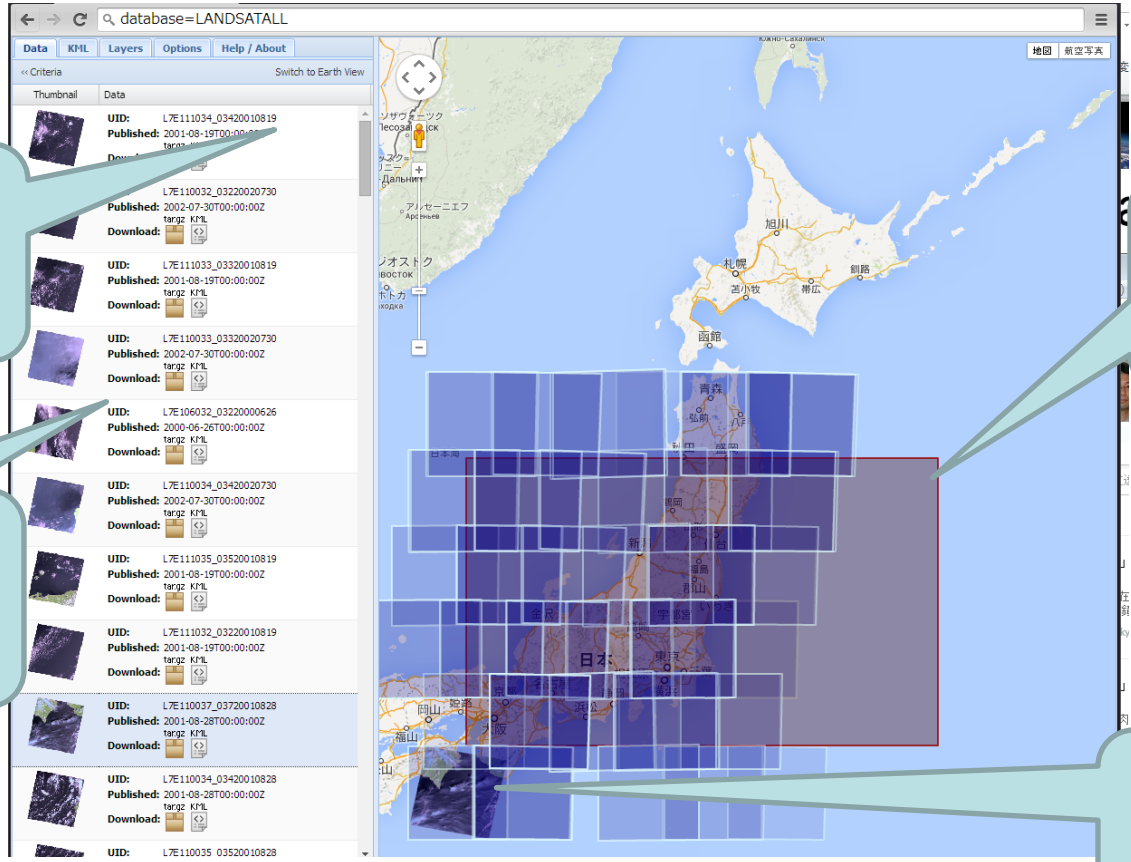
Public Service http://landsat8.geogrid.org

Latest/Historical Data can be Downloaded and Viewed

User Contributions Like "I found interesting things!" by Facebook

OGC OpenSearch EO & CS-W

- http://l8search.geogrid.org/GECSW_maps/



Support Google Map& Earth

Data can be downloaded With tgz or kml

Various Search Conditions BBOX,Date Etc.

Provide the data with WMS

Constellation

- ASTER= 16days cycle
- Landsat-8= 16days cycle
- ASTER+Landsat-8= 8days cycle
  - Same orbit

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(& Free&Open)

Analysis

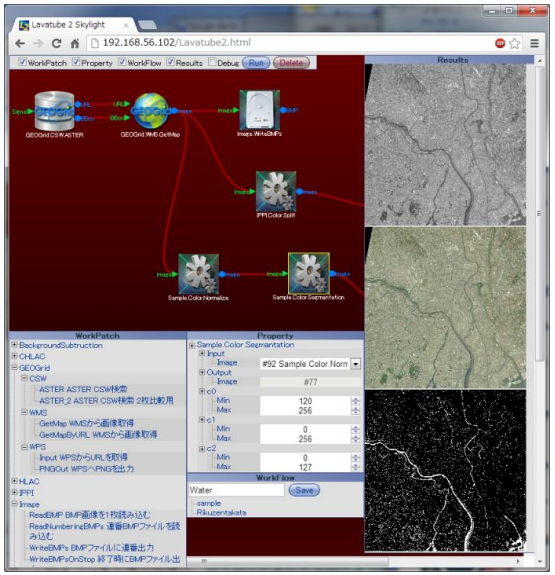
Our Workflow engine “Lavatube” &

Our Machine Learning System “Hivemall”

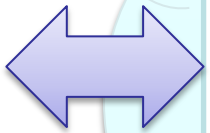
Our yet another Workflow Engine: Lavatube

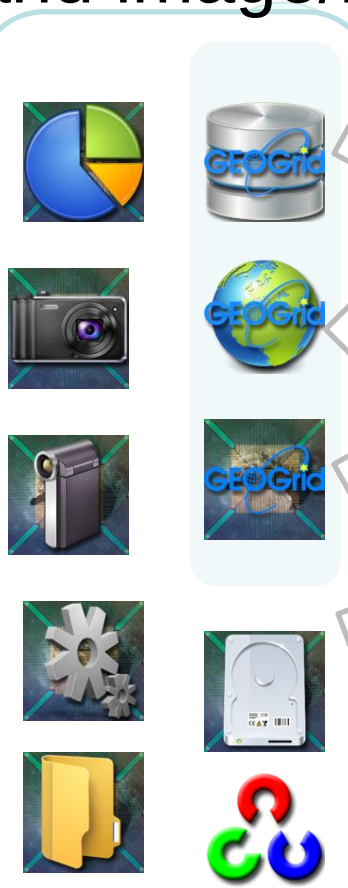
for spacio-temporal data and image/movie processing

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





Workflow Parts

Workflow Engine

CSW

WMS

WPS

WCS

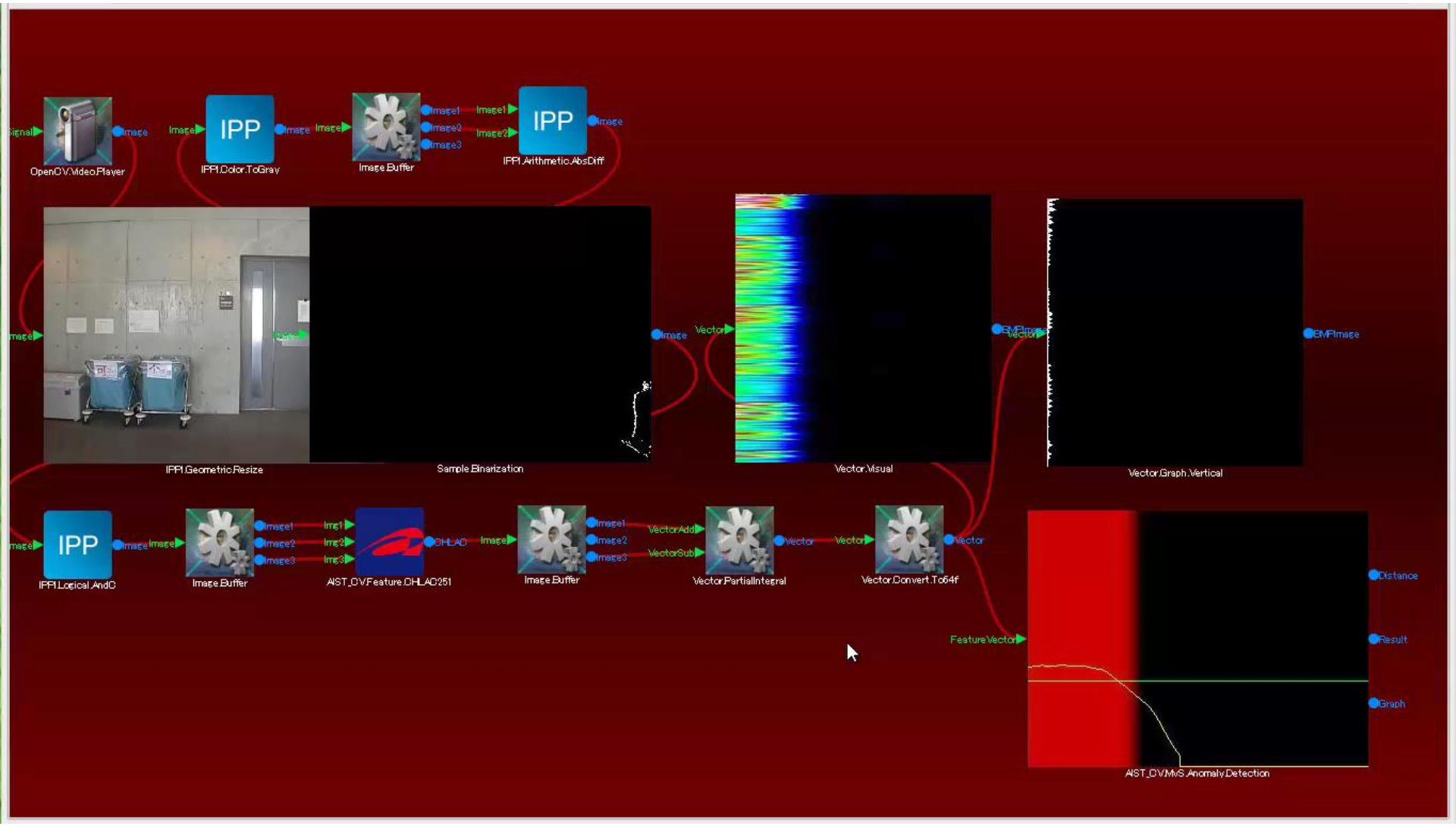
Satellite

Sensor/
Camera

lavatube.mp4

OGC Services

- Metadata Search
- Database Access
- Processing

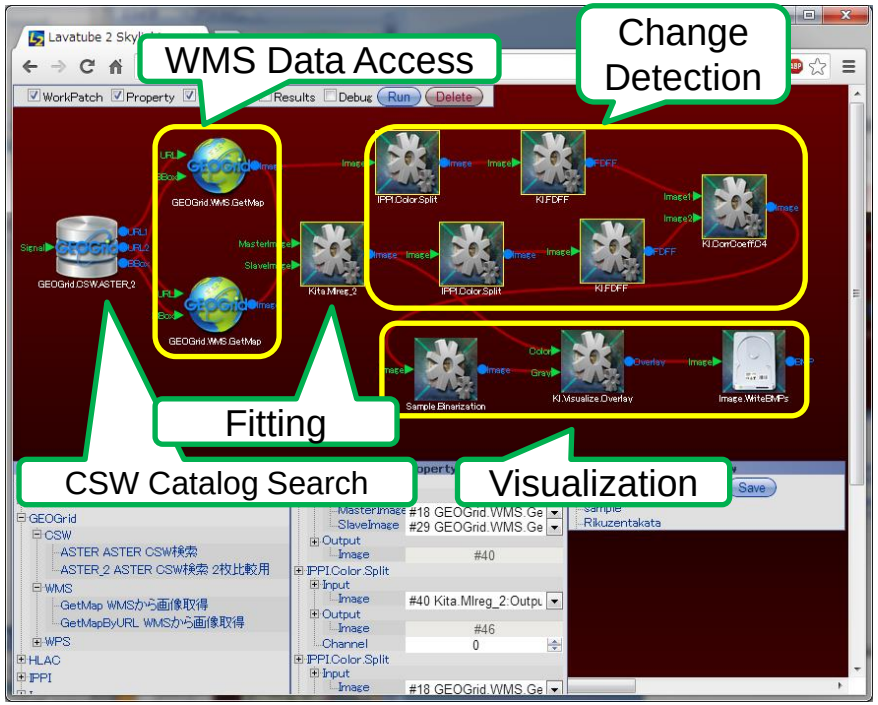

OpenCV.Video.Player
IPP.Color.ToGray
Image.Buffer
IPP.Arithmetic.AbsDiff
IPP.Geometric.Resize
Sample.Binarization
Vector.Visual
Vector.Graph.Vertical
IPP.Logical.AndC
Image.Buffer
AIST_CV.Feature.CHLAC251
Image.Buffer
Vector.PartialIntegral
Vector.Convert.To64f
AIST_CV.MvS.AnomalyDetection
Distance
Result
Graph

Example: Change Detection Application of GEO Grid

1. Metadata Search(OGC CS-W) for a specific area
2. Data Acquisition(OGC WMS) for 2 satellite images
3. Position/Color Fitting of satellite images
4. Change Detection
5. Visualization (on browser /google earth)

3.11 Damaged Area (Rikuzen-takada City)

Construction of Temporal Residential Space on High ground area



Workflow constructed with Lavatube

2011.06

2012.06

Result

Reconstruction of Breakwater

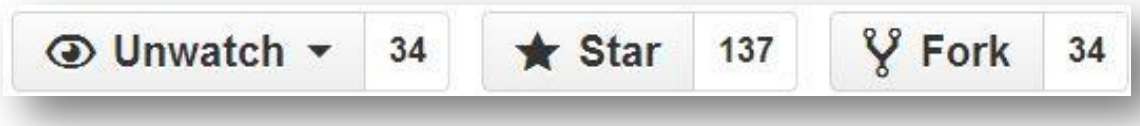
Hivemall: Scalable Machine Learning Library for Apache Hive

presented@ Hadoop Summit

What is Hivemall

- A collection of machine learning algorithms as Hive UDFs/UDTFs
  - Classification & Regression
  - Recommendation
  - k-Nearest Neighbor Search
  - .. and more

- An open-source project on Github
  - Licensed under LGPL
  - github.com/myui/hivemall (**bit.ly/hivemall**)



Motivation – Why a new ML framework

Machine Learning Frameworks run with Hadoop out there

Mahout?

Spark MLlib?

Vowpal Wabbit?
(w/ Hadoop streaming)

Cloudera Oryx?

H2O?

Motivation – Why a new ML framework

Existing distributed machine learning frameworks, excluding Vowpal Wabbit, are NOT easy to use for most developers

Framework	User interface
Mahout	Java API Programming
Spark MLlib/MLI	Scala API programming
H2O	R programming GUI
Cloudera Oryx	Http REST API programming
Vowpal Wabbit (w/ Hadoop streaming)	Command Line

Classification with Mahout

Question: How many Line of Codes (LoC) required for running logistic regression on the 20 newsgroups dataset?

org/apache/mahout/classifier/sgd/TrainNewsGroups.java

```
1  public final class TrainNewsGroups {
2
3    public static void main(String[] args) throws IOException {
4      File base = new File(args[0]);
5
6      Multiset<String> overallCounts = HashMultiset.create();
7
8      int leakType = 0;
9      if (args.length > 1) {
10       leakType = Integer.parseInt(args[1]);
11     }
12
13     Dictionary newsGroups = new Dictionary();
14
15     NewsgroupHelper helper = new NewsgroupHelper();
16     helper.getEncoder().setProbes(2);
17     AdaptiveLogisticRegression learningAlgorithm =
18         new AdaptiveLogisticRegression(20, NewsgroupHelper.FEATURES, new L1());
19     learningAlgorithm.setInterval(800);
20     learningAlgorithm.setAveragingWindow(500);
21
22     List<File> files = Lists.newArrayList();
23     for (File newsgroup : base.listFiles()) {
24       if (newsgroup.isDirectory()) {
25         newsGroups.intern(newsgroup.getName());
26         files.addAll(Arrays.asList(newsgroup.listFiles()));
27       }
28     }
29     Collections.shuffle(files);
30     System.out.println(files.size() + " training files");
31     SGDInfo info = new SGDInfo();
32
33     int k = 0;
34
35     for (File file : files) {
36       String ng = file.getParentFile().getName();
37       int actual = newsGroups.intern(ng);
38
39       Vector v = helper.encodeFeatureVector(file, actual, leakType, overallCounts);
40       learningAlgorithm.train(actual, v);
41
42       k++;
43       State<AdaptiveLogisticRegression.Wrapper, CrossFoldLearner> best = learningAlgorithm.getBest();
44
45       SGDHelper.analyzeState(info, leakType, k, best);
46     }
47     learningAlgorithm.close();
48     SGDHelper.dissect(leakType, newsGroups, learningAlgorithm, files, overallCounts);
49     System.out.println("exiting main");
50
51     File modelFile = new File(System.getProperty("java.io.tmpdir"), "news-group.model");
52     ModelSerializer.writeBinary(modelFile.getAbsolutePath(),
53         learningAlgorithm.getBest().getPayload().getLearner().getModels().get(0));
54
55     List<Integer> counts = Lists.newArrayList();
56     System.out.println("Word counts");
57     for (String count : overallCounts.elementSet()) {
58       counts.add(overallCounts.count(count));
59     }
60     Collections.sort(counts, Ordering.natural().reverse());
61     k = 0;
62     for (Integer count : counts) {
63       System.out.println(k + "\t" + count);
64       k++;
65       if (k > 1000) {
66         break;
67       }
68     }
69   }
70
```

Find the complete code at bit.ly/news20-mahout

Classification with Spark

- ✓ LoC is much smaller and easy understanding
- ✓ Works fine for small dataset (on 32 nodes)

```
val training = MLUtils.loadLibSVMFile(sc,
"hdfs://dm01:8020/small/training_libsvmfmt",  multiclass = false, numFeatures =
16777216)

val model = LinearRegressionWithSGD.train(training, numIterations = 20)
```

However, Spark does not work with 16,777,216 (2^24) features

Actually, feature dimension is much larger (reduced dimension by feature hashing)

→ Caused OOM error for large datasets

Sparse feature support (in v 1.0) seems still in β-quality

→ Or, dedicated Spark programming is mandatory for extensive uses

Disclaimer: I'm NOT a Spark expe

Liblinear on Spark does not work as well in our experiments

Why Hivemall

1. Easy of use
   - No programming
   - No compilation/packaging overhead
   - Easy for existing Hive users
     - You can evaluate Hivemall within 5 minutes or so
     - Installation is just as follows

```
$ hive

add jar /tmp/hivemall.jar;
source /tmp/define-all.hive;
```

Why Hivemall

2. Scalable to data
   - Scalable to # of training/testing instances
   - Scalable to # of features
     - Built-in support for feature hashing (feature dimension reduction)
   - Feature engineering step is also scalable and parallelized using Hive
3. Scalable to computing resources
   - Exploiting the benefits of Hadoop & Hive
   - Provisioning the machine learning service on Amazon Elastic MapReduce
     - Provides an EMR bootstrap for the automated setup

See bit.ly/hivemall-emr

4. Support the state-of-the-art online learning algorithms (for classification)

- Less configuration parameters (no learning rate as one in SGD)
- CW, AROW, and SCW are not yet supported in other ML frameworks
  - ns is enough)

Algorithms	News20.binary Classification Accuracy
Perceptron	0.9460
Passive-Aggressive (a.k.a. Online-SVM)	0.9604
LibLinear	0.9636
LibSVM/TinySVM	0.9643
CW	0.9656
AROW [1]	0.9660
SCW [2]	0.9662

Better

1. Adaptive Regularization of Weight Vectors (AROW), Crammer et al., NIPS 2009
2. Exact Soft Confidence-Weighted Learning (SCW), Wang et al., ICML 2012

What can Hivemall do

- **Classification (both one- and multi-class)**
  - ✓ Perceptron
  - ✓ Passive Aggressive (PA)
  - ✓ Confidence Weighted (CW)
  - ✓ Adaptive Regularization of Weight Vectors (AROW)
  - ✓ Soft Confidence Weighted (SCW)
- **Regression**
  - ✓ Logistic Regression using Stochastic Gradient Descent (SGD)
  - ✓ PA Regression
  - ✓ AROW Regression
- **k-Nearest Neighbor & Recommendation**
  - ✓ Minhash and b-Bit Minhash (LSH variant)
  - ✓ Brute-force search using similarity measures (cosine similarity)
- **Feature engineering**
  - ✓ Feature hashing
  - ✓ Feature scaling (normalization, z-score)

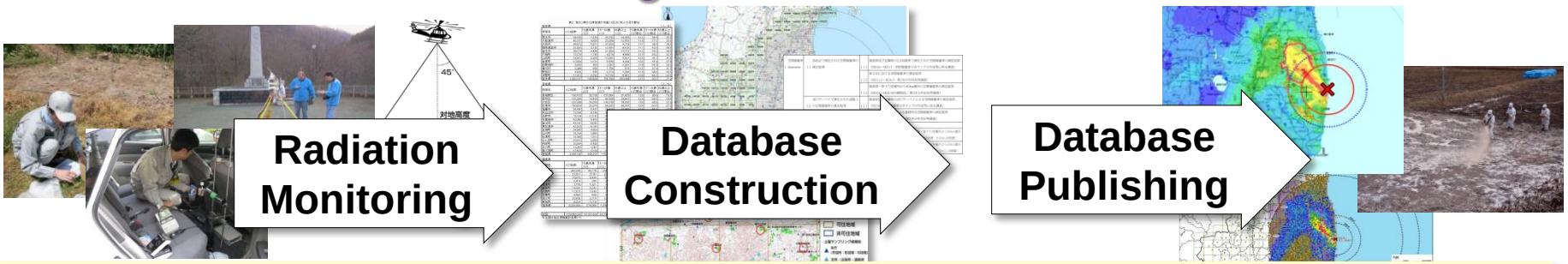
Application

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 Structure as of 2013

Nuclear Regulation Authority(NRA), JAPAN

Project Management: Japan Atomic Energy Agency(JAEA)

JAEA	AIST (National Institute of Advanced Industrial Science and Technology (AIST))	Hokkaido University	Japan Map Center (Company)
2 Teams	2 Teams	2 Teams	1 Team

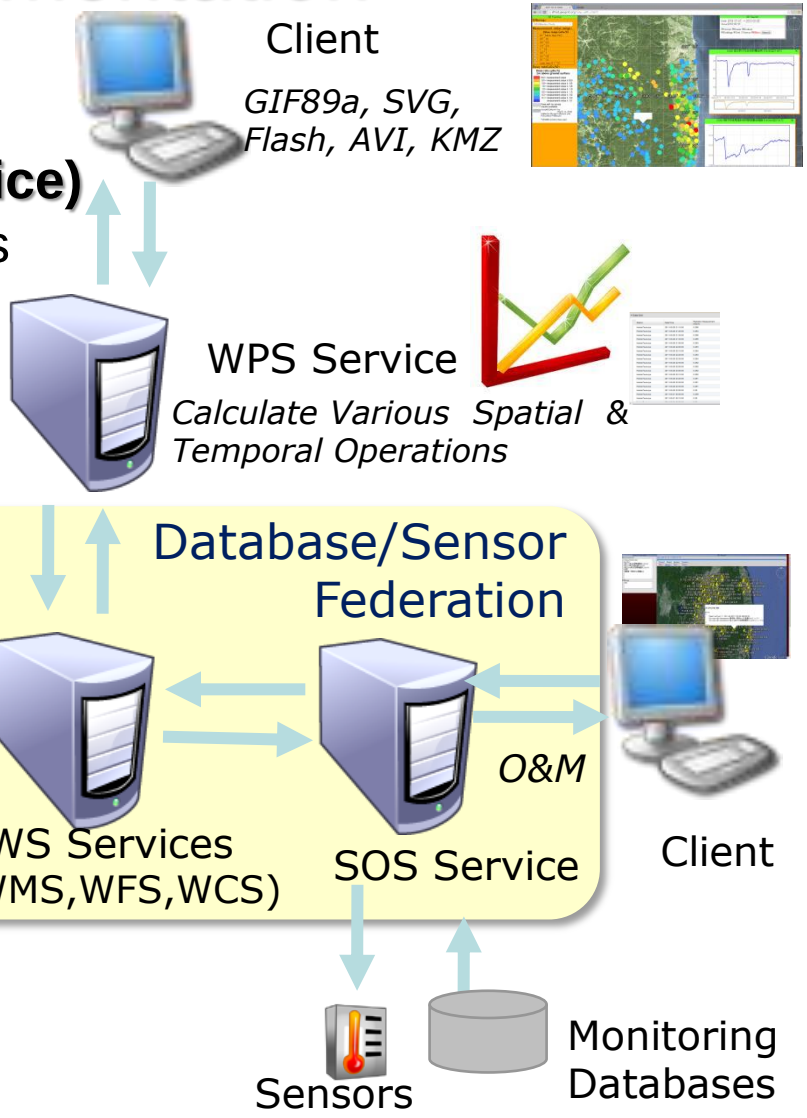
Partners ・・・・・ GIS. FCU Okayama-U, The Institute of Statistical Mathematics,,

Framework Implementation

- Access Data : **SOS and W*S**
  (Sensor Observation Service)
  (Web Map/Feature/Coverage Service)
  – Interface to radiation measurements
  – Federate with other OWS servers

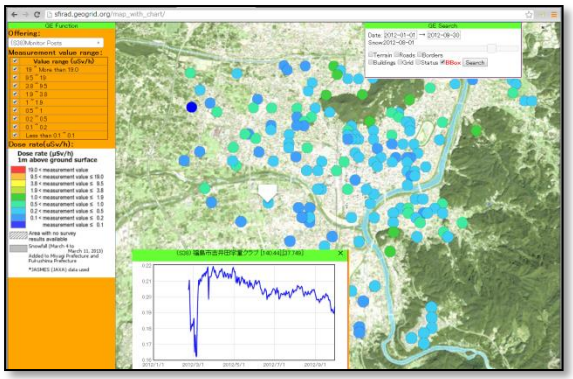
- Process Data: **WPS**
  (Web Processing Service)
  – Spacial&Temporal Analytics
  – Produce various format
    • KMZ,AVI etc.

- Publish Data: **WMS**
  (Web Map Service)
  WMS Time of maps
  – KMZ
  – Rich Client with Google Earth/Map

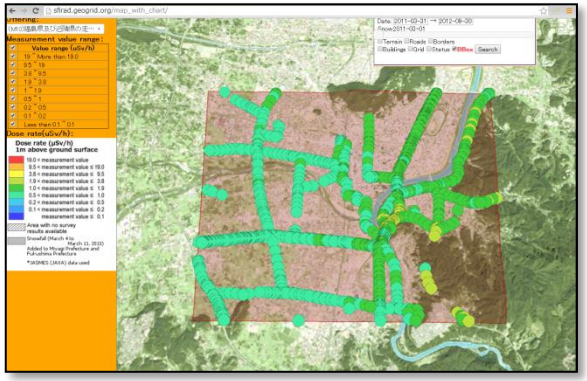


Access Service Implementation

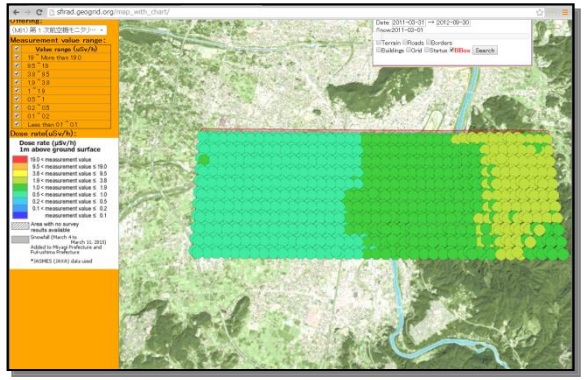
- Single OGC SOS(Sensor Observation Service) Service for
  - Real-time Monitoring Post Database
    - Over 25million records for 4000 sensing points in every 10 minutes
    - Data is obtained from the dedicated system which handles all sensors
  - Periodical Monitoring Results Database
    - > 40 types of monitoring (Car Survey, Airborne Survey etc.)
    - Over 17million records
- Rich Client with Google Earth/Map
  - GUI for SOS search & display with BBOX, Date Range, Dose Rate, etc.
  - Simple analytics functions(ex. Writing Graph)



Real-time Monitoring Posts

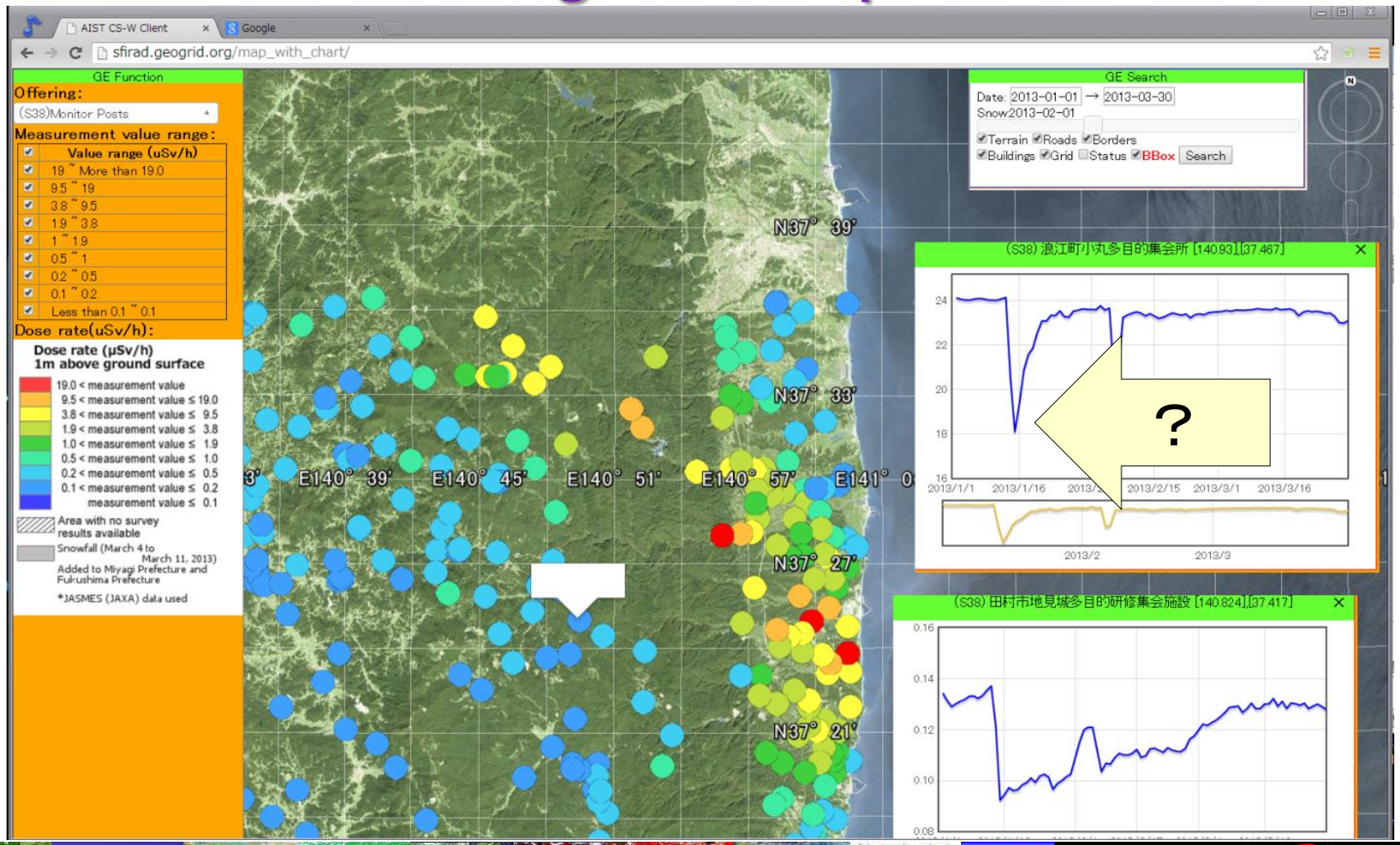


Car Survey Results



Airborne(Helicopter) Survey Results

Example Data Integration Application using OGC specs

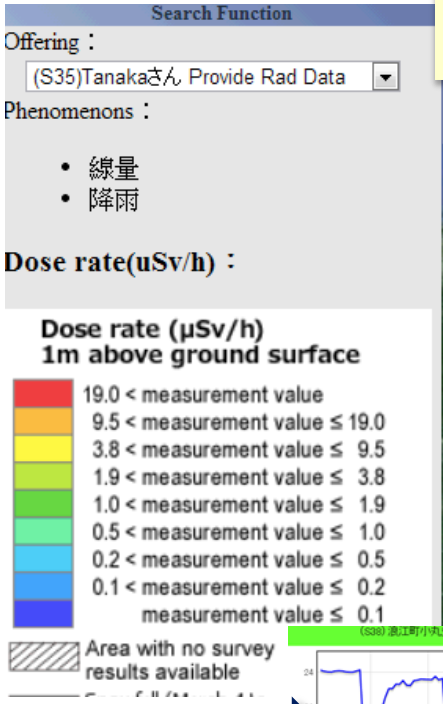


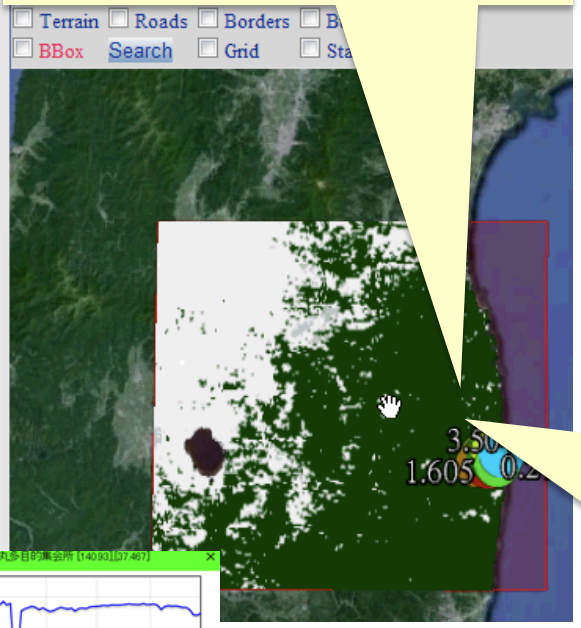
Combine our SOS with other WMS data source(Weather)

Human exposure to natural background radiation, 0.27uSv/h

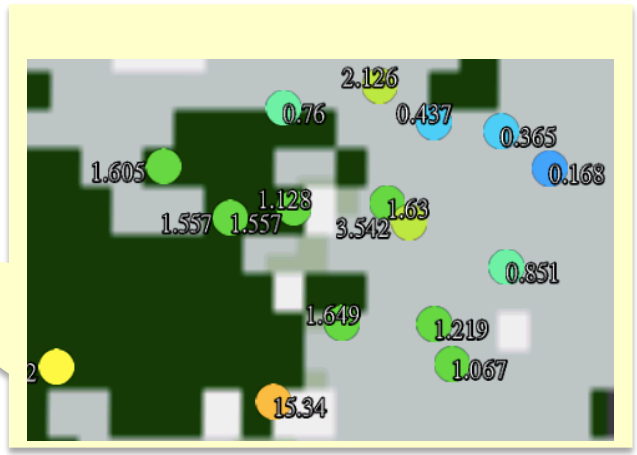
Jan 1, 2013
The dose rate was relatively high

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





Snow effect

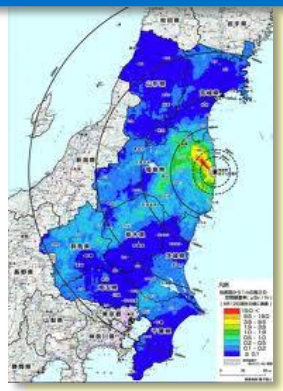


Simple overlay can be useful

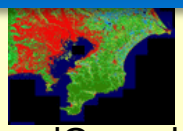
Application Implementation using our Framework (OGC OWS+LinkedOpenData/GeoSPARQL)

- In collaboration with radiation experts
  - To demonstrate the usefulness of our framework
    create useful knowledge with good productivity

OGC SOS + **OGC WMS** + **OGC WPS** = **Environmental Dose Reduction Rate**

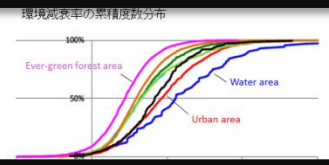


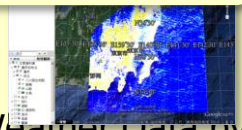
Radiation DB (CSV/XML)



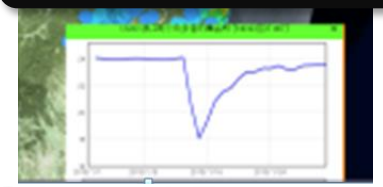
LandCover by JAXA (HDF)

Analytics Service



+  + Analytics Service = **Effect of Snow Coverage for Dose Reduction**

Weather Data by Satellite Analysis



OGC based Framework

OGC LOD/RDF (GeoSPARQL) + **LOD/GeoSPARQL** + **Federated? Search** = **Collective Dose**

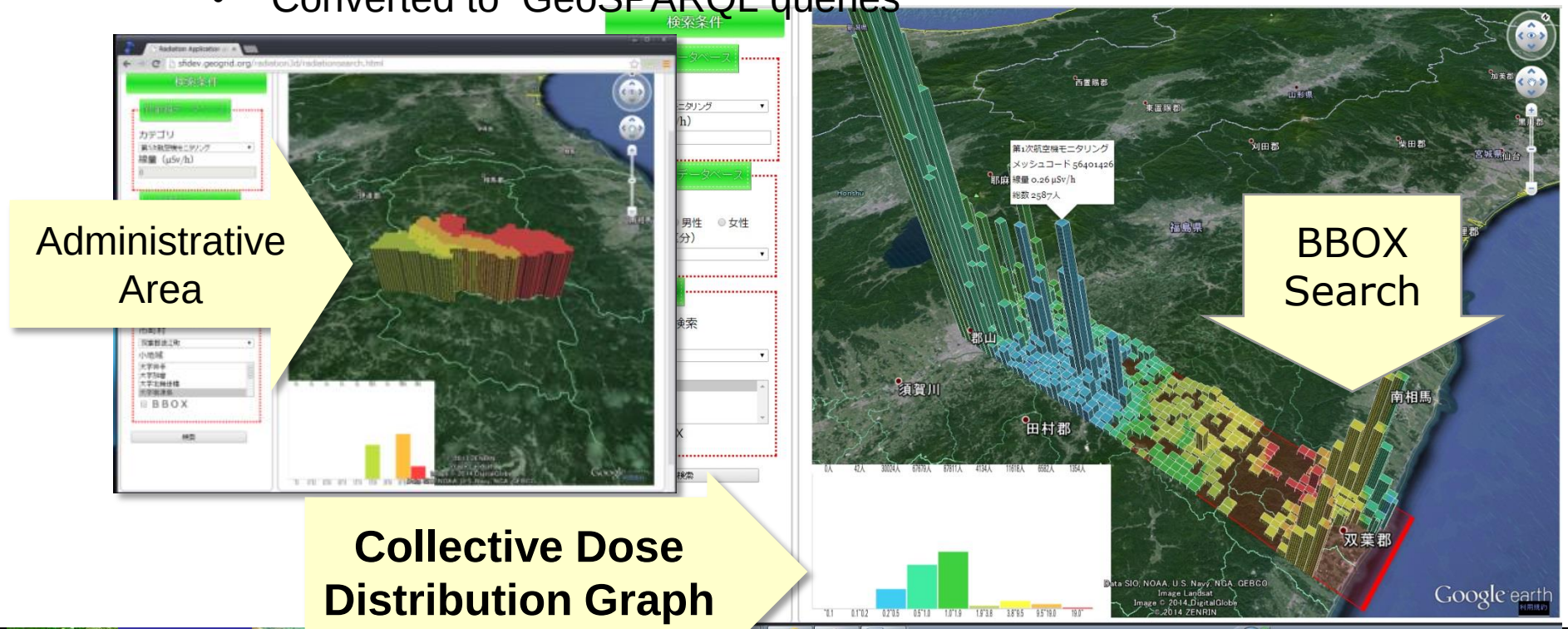
Poputation Stats by MIC(CSV)

Analytics Service

LOD/RDF(+Geo) based Framework

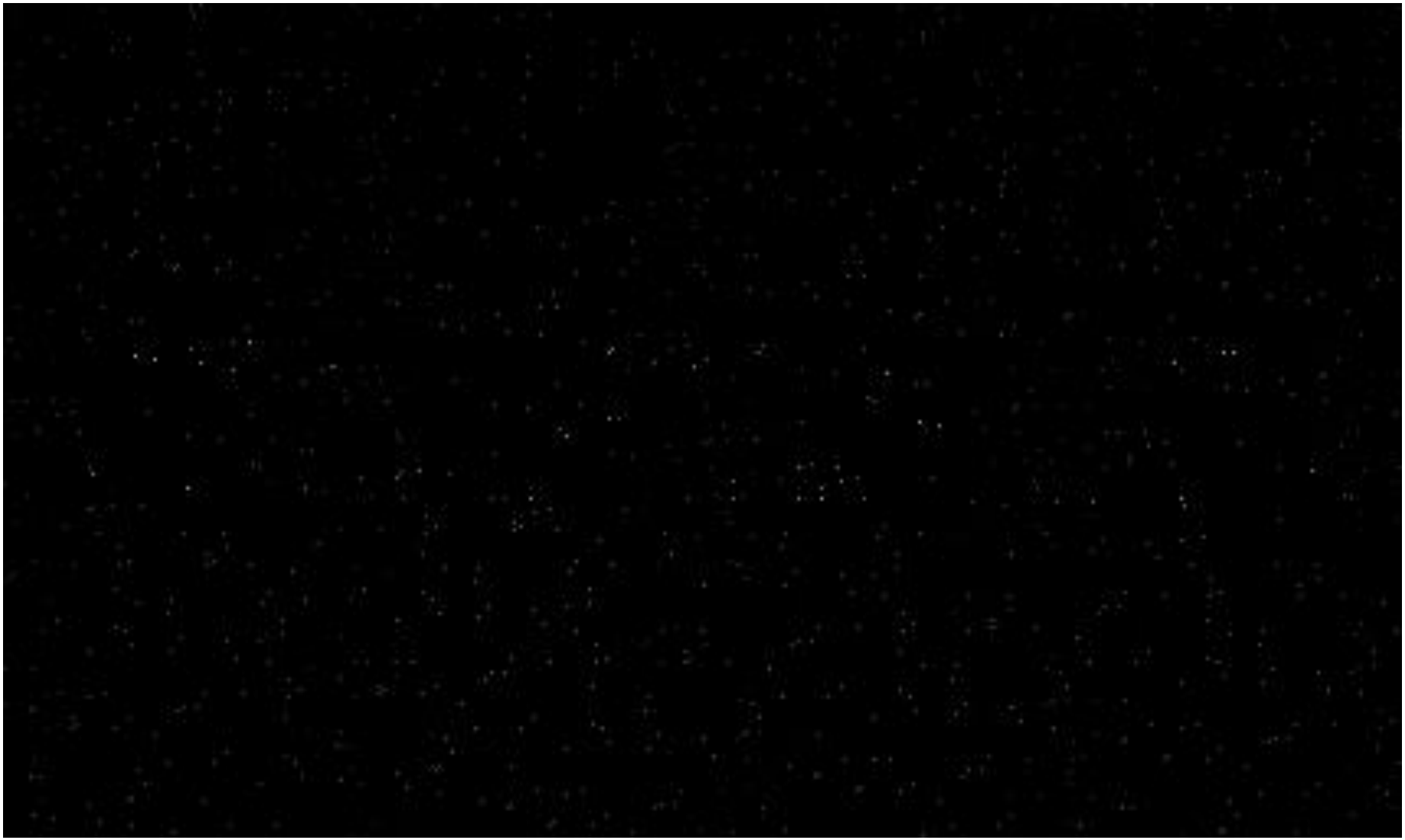
Collective Dose = Dose Rate X Population Statistics

- Statistics provided by Ministry of Public Management
  - Can be provided with RDF/LOD
- **Rich Client with same style GUI:** Can search with Area(BBOX&Area), Date, Value, Age, Sex, etc.
  - Converted to GeoSPARQL queries



GEO Grid

Grid-based e-infrastructure for geosciences



Directions

Social,Mobile & Crowdsourcing

Directions

- Social = Use Twitter Data

- Mobile = Use Mobile Sensors
  - Smartphones,Lifebands,Cars etc.
  - AIST-developed radiation sensor

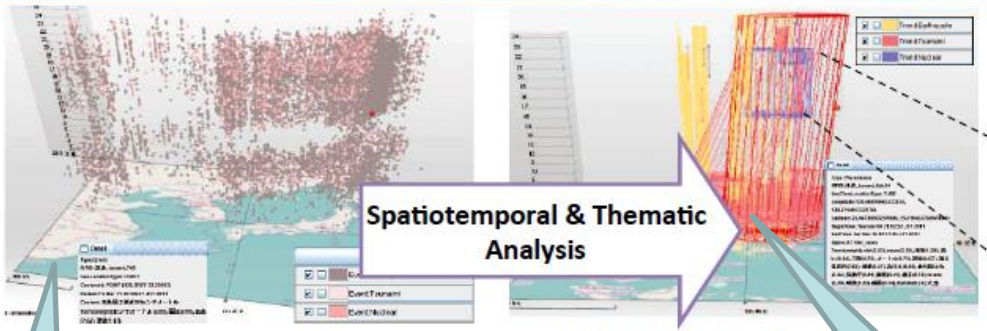
- Crowdsourcing
  - Calibration with Machine Learning
  - Notification with GeoSMS

Data Integration of Administrative & Non-Administrative Data

- Administrative Data(Current GEO Grid Data)
  - Governmental & Official Data
  - Limited amount with controlled Quality
- Non-Administrative Data
  - NPO, Social(Twitter etc)
  - Large amount, Various Quality

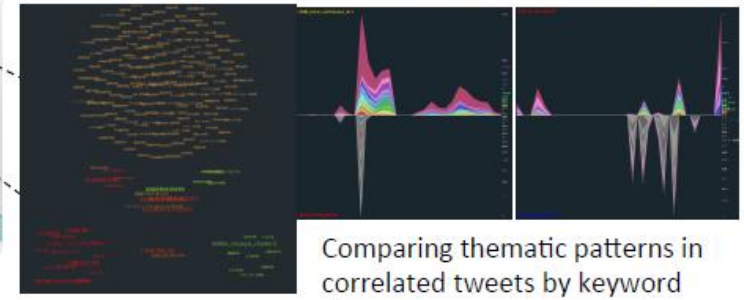
- Data Integration Issue

Application Examples



Mapping, clustering, and regression: (2011/03/02-2011/03/24)
1) earthquake-related tweets
2) tsunami-related tweets
3) nuclear-related tweets

- mTrend (ACMGIS2011 Demonstration)

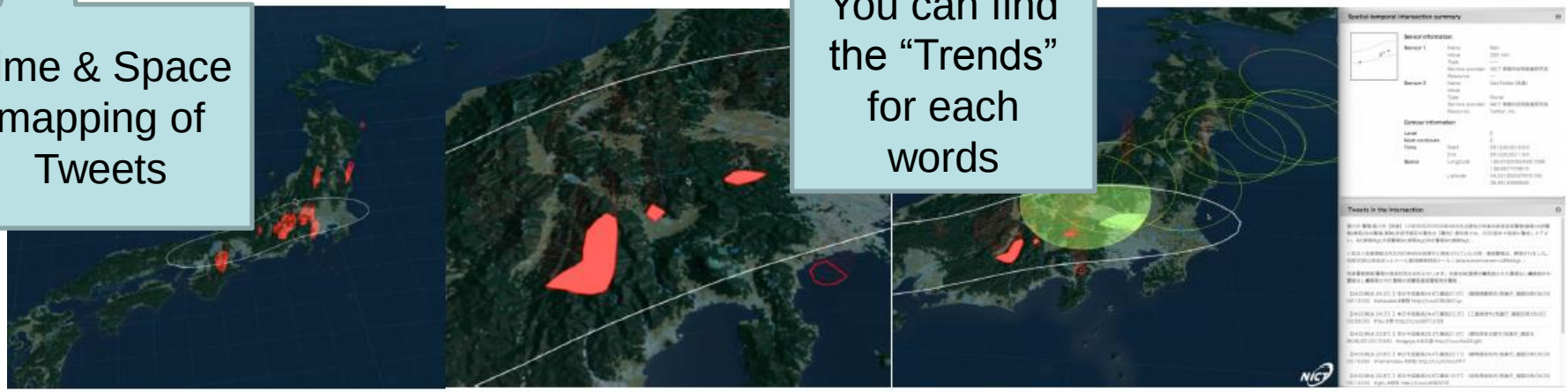


Comparing thematic patterns in correlated tweets by keyword streams

地理情報システムの[illegible]ンファレンス
ACM GIS 2011で優秀テ[illegible]を受賞

Time & Space mapping of Tweets

You can find the “Trends” for each words



- 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

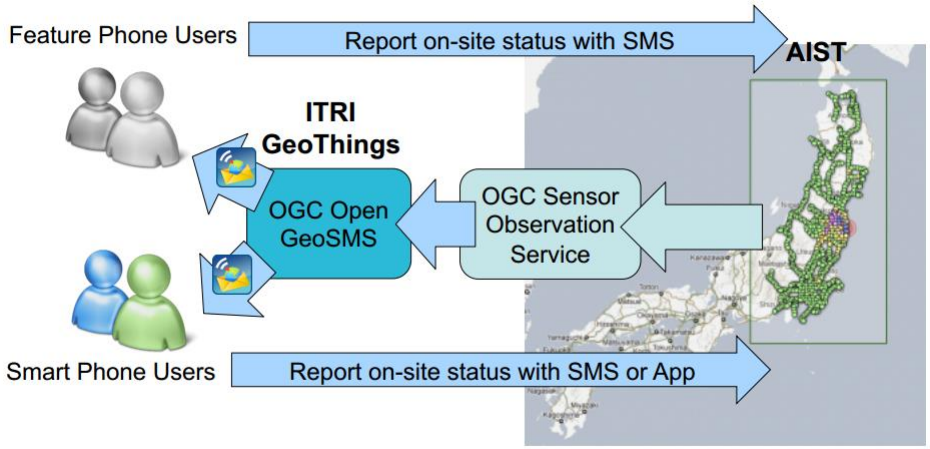
Crowdsourcing and Notifications Joint Research with Taiwan ITRI (not AIST's ITRI)

工業技術研究院
Industrial Technology Research Institute

ITRI & AIST Joint Research Project
Data Integration and Messaging Fram
Crowdsourced and Administrative Sen

莊國煜
Kuo-Yu slayer Chuang
slayer@itri.org.tw

Crowdsourced Radiation Sensors



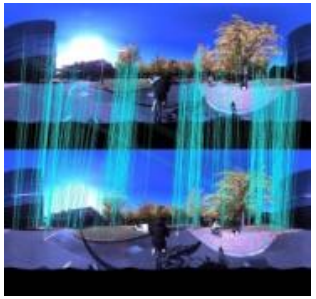
GEO Grid & Big-Data

Need to support high-level analysis and modelling

“G-Space” Computing Platform

- Complex data& Complex processing: Combine with Machine Learning & 3D Modelling
- Mission-Critical Application: Real-Time Computing

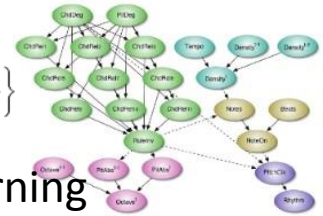
R&D



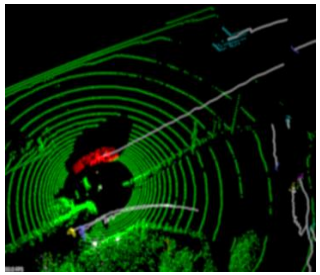
Feature Extraction

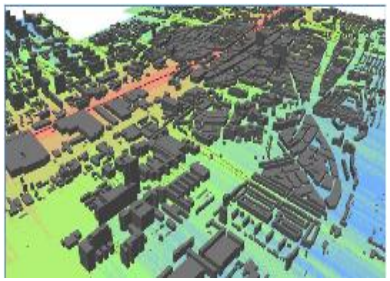
$$p(y \mid x;\psi,\theta) = \sum_{k=1}^{K} \xi_k(x;\psi_k)\, p_k(y \mid x;\theta_k)$$

$$p_k(x;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}$$



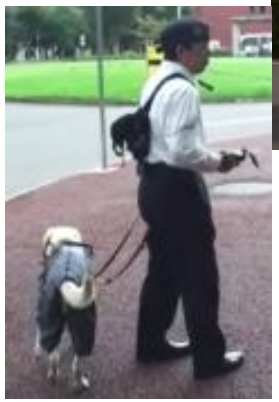
Macnine Learning





3D-Modelling and Understanding

Application



Drive Support

Task Management

G-computing Platform Overview

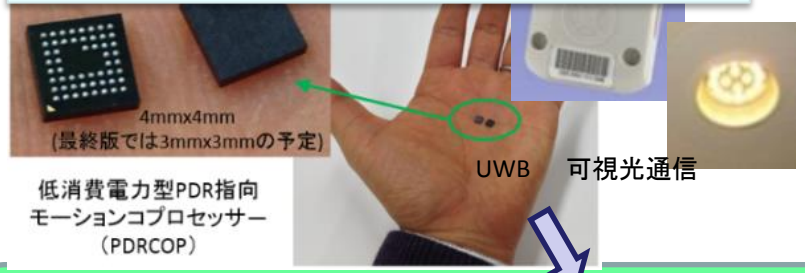
GEO Grid
Grid-based e-infrastructure for geosciences

Sensing

Infrastructur

Analyse

Human Activity Sensing



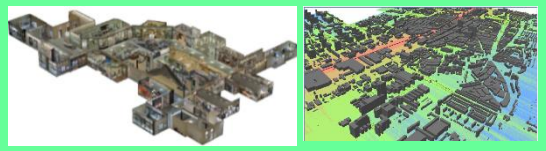
Intelligent Mobility Sensing

Information Infrastructure

Data Archive

Geospatial Cloud(GEO Grid)

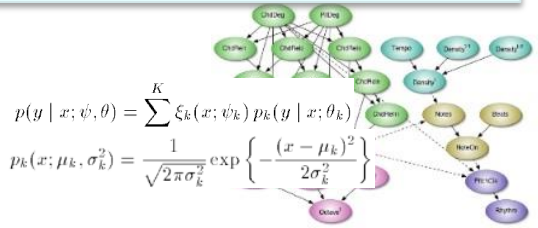
高速化
並列化
リアル
タイム



Environmental Modelling

Feature Extraction

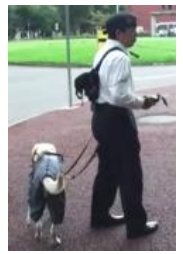
Machine Learning



$$p(y \mid x;\psi,\theta) = \sum_{k}^{K} \xi_k(x;\psi_k)\, p_k(y \mid x;\theta_k)$$

$$p_k(x;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}$$

Applications



個人移動支援

運転支援

障がい者支援

作業管理

AIST

Linked Open Data

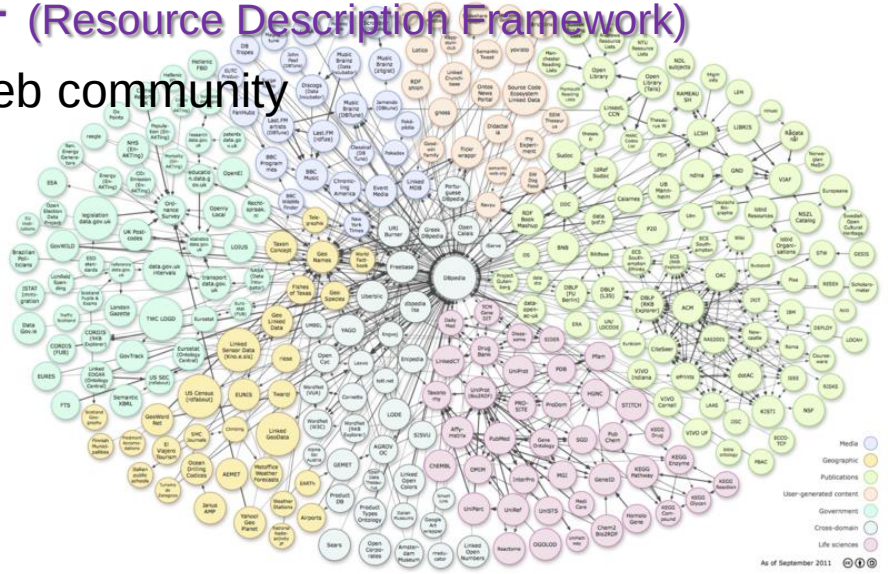
Federated SPARQL with "Best-Effort" Query Processing

Linked Open Data(LOD)

Next generation web?

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
  - More than 300 sites
- Data is rapidly increasing
  - Billion~Trillion triples
- Cross-Domain
  - GEO,BIO,Govement,Media, etc
  - Many governmental data is going to be published as LOD



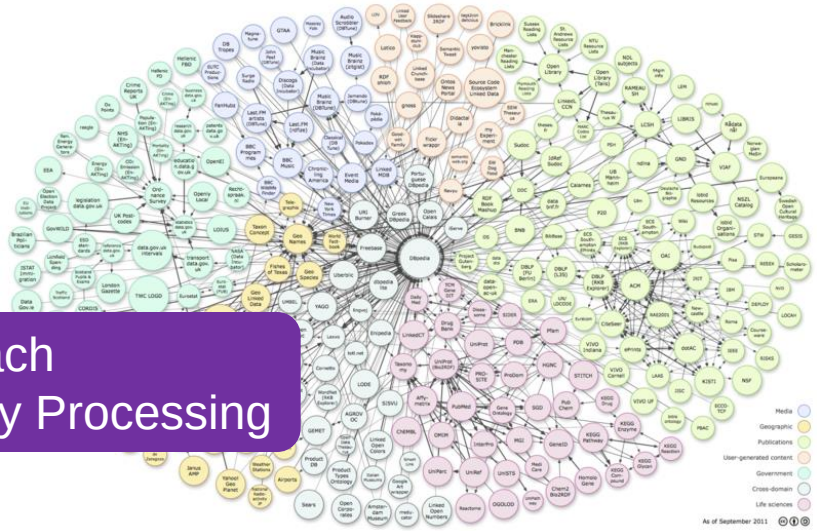
Linked Open Data

Trade-off Issue

- Distributed SPARQL Processing may be slow
- Using centralized data service lacks freshness

Heterogeneity

- SPARQL Endpoints
- Plain RDF Texts
- …

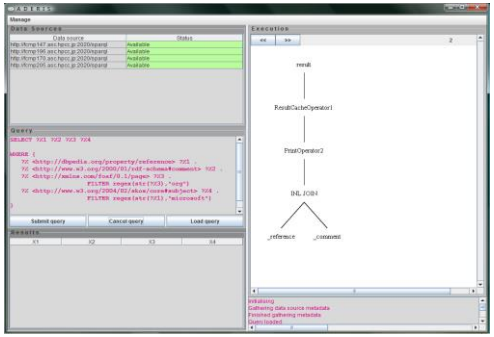


Our Approach
Hybrid Adaptive Query Processing

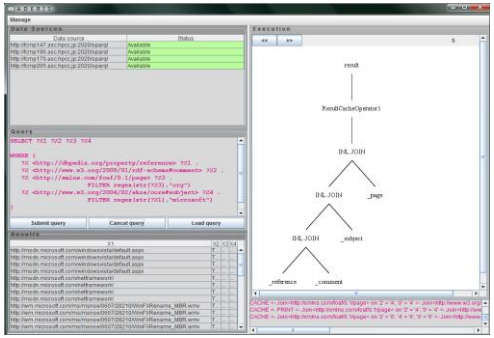
Based on the **freshness, coverage** and the **response time**

Adaptive Query optimization

Pre-defined
Query Processing Schedul



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

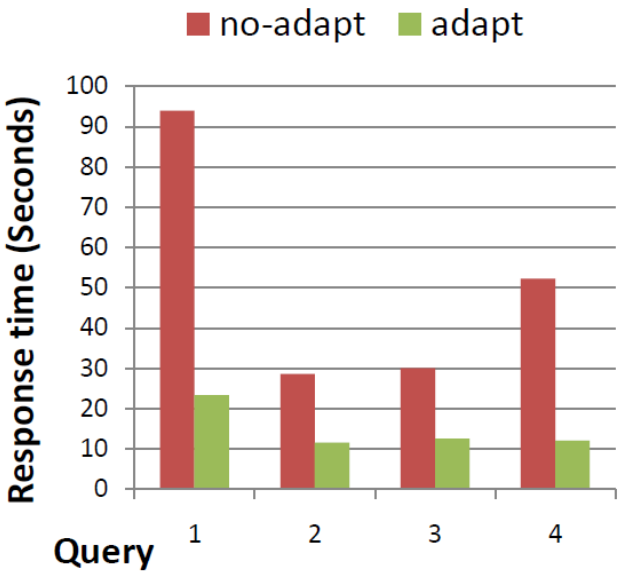


Modified Processing
Adaptively

```
dbo: <http://dbpedia.org/ontology/>
dbp: <http://dbpedia.org/property/>
owl: <http://www.w3.org/2002/07/owl#>
rdf:  <http://www.w3.org/2000/01/rdf-schema#>
skos: <http://www.w3.org/2004/02/skos/core#>
foaf: <http://xmlns.com/foaf/0.1/>

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 dbo: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 dbo:book         3105
 ?x dbo:releaseDate ?date     (DBP) 126,737
}
```

```
Query 4 (Result size = 13):
select * {
 ?book rdf:type dbo:Book .    3105
 ?book foaf:page ?p .         10,000
 ?book owl:sameAs ?link       (DBP) 10,121,699
}
```

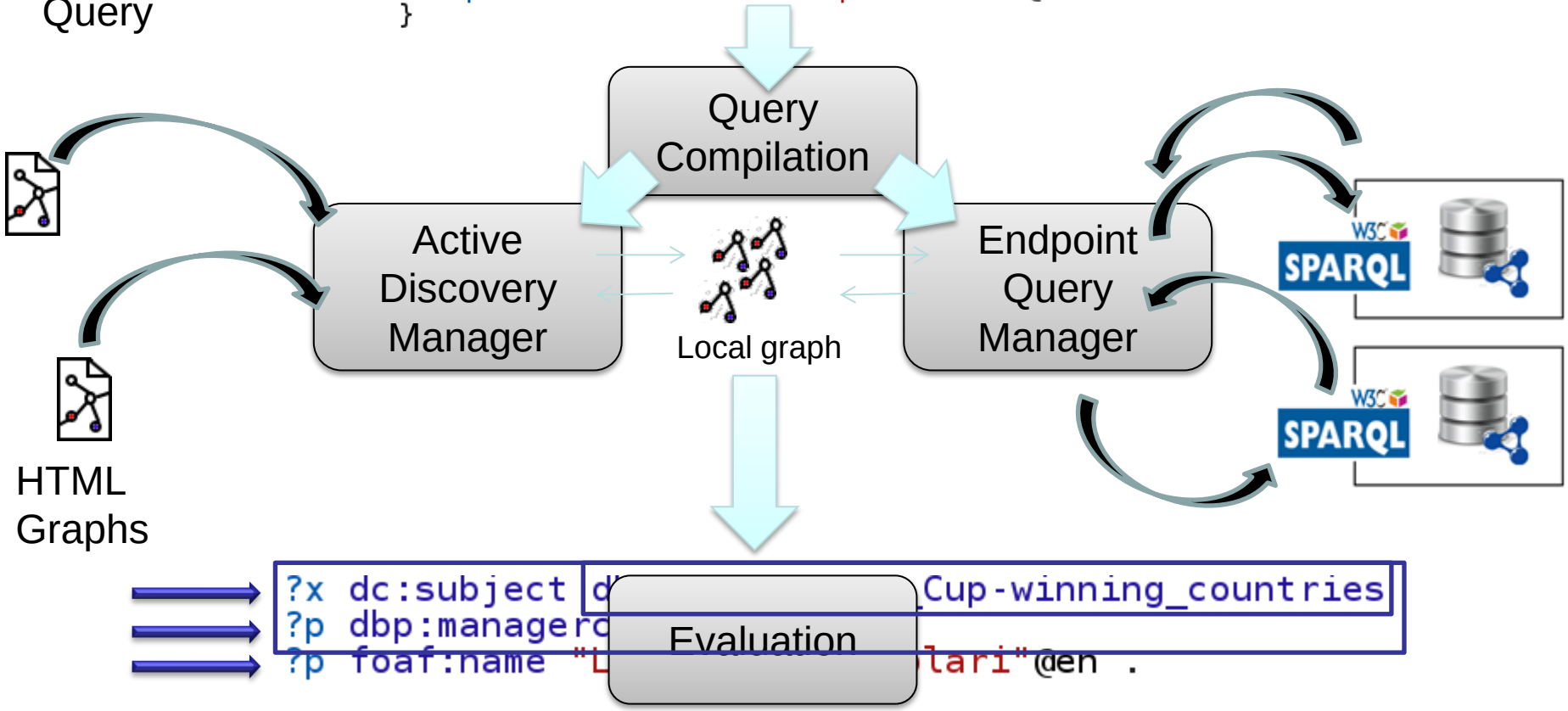


Achieve good performance around 10 distributed SPARQL endpoints (still small for 300 ;-<)

Hybrid &Adaptive Query Processing

User's SPARQL Query

```
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 dbp:managerclubs ?x .
    ?p foaf:name "Luiz Felipe Scolari"@en .
}
```



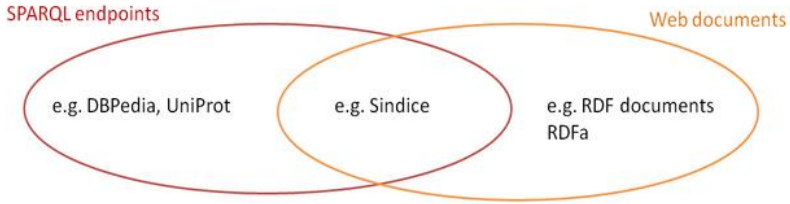
Query Result

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

ADERIS-Hybrid | Query | Data Sources | **About** | Help | Applications

Answering distributed Linked Data queries with reasonable response times.

Linked Data can be obtained from both SPARQL endpoints and Web documents, querying it allows information to be retrieved from the Web of Data.



There are a number of issues that make querying Linked Data a challenging problem. Accessing diverse data sources spread around the Web introduces unpredictability in terms of response times and the amount of data that will be returned, making static query plan optimisation difficult. We aim to provide relevant, fresh results in a timely manner while conceding that results are unlikely to be complete. The approach is a hybrid of distributed query processing and live exploration-based querying of the Web of Linked Data.

More information

The approach used is described in the following paper:

- A Hybrid Approach to Linked Data Query Processing with Time Constraints by Steven Lynden, Isao Kojima, Akiyoshi Matono, Akihito Nakamura and Makoto Yui presented at the Linked Data on the Web (LDOW) 2013 workshop at the WWW 2013 conference.

Presentation slides are also available:

- A Hybrid Approach to Linked Data Query Processing with Time Constraints (PDF slides)

BioScience+IT

BIO-CAD/LEAD Project

BioCAD

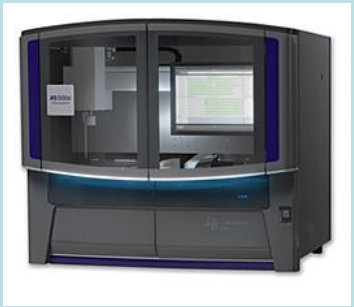
...is a cross-disciplinary project of AIST for R&D of genome information processing for gene discovery

WET + **DRY** Why **IT** is nessesary?

= Bioinfomatics

Aim

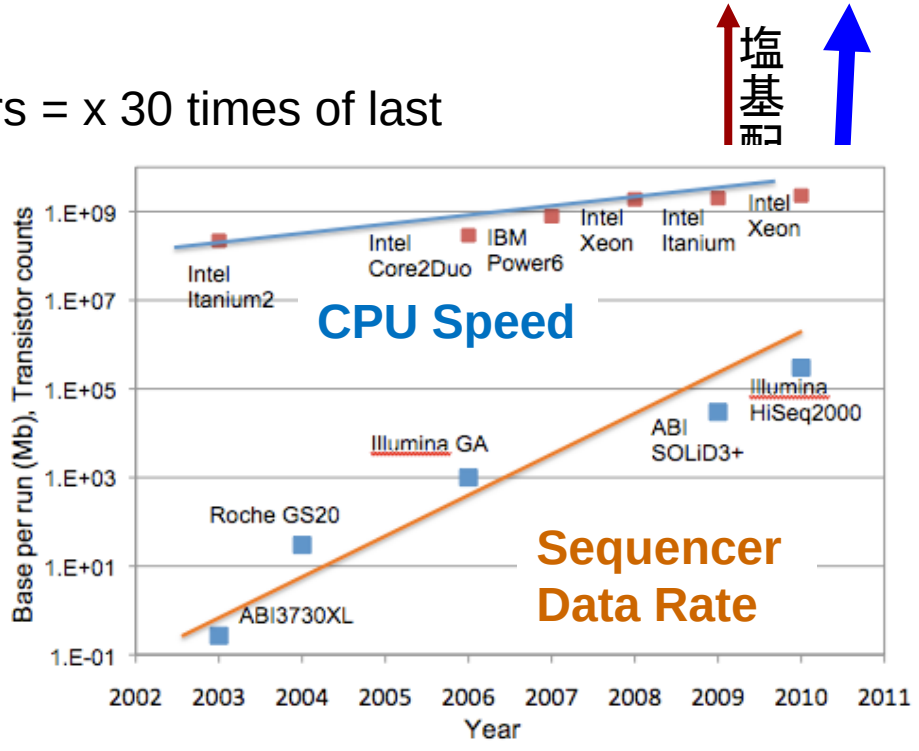
Emergence of NGS
(Next?= Current Generation Sequencer)
Illumina, SOLiD, PacBio etc



Note: Current “Next” Generation is like ion Proton

Background and Motivation

- **BIG-DATA**
  - Amount of base info of coming 5years = x 30 times of last 20years
  - **Need to have scalable ITs**
- **Data Mining**
  - Advance of New Generation Sequencers
    ＞＞Advance of single CPU
  - **Need to have parallel HPC**
- **Cloud Computing**
  - Need to outsource HPC
  - Remote HPC from Wet Labs
  - **Need to have distributed processing including networks**



Common Issues
In all science

IT must solve them

Our Current Target: High Performance Assemble
What is [Assemble」?

- **Output of the Next Generation Sequencers**

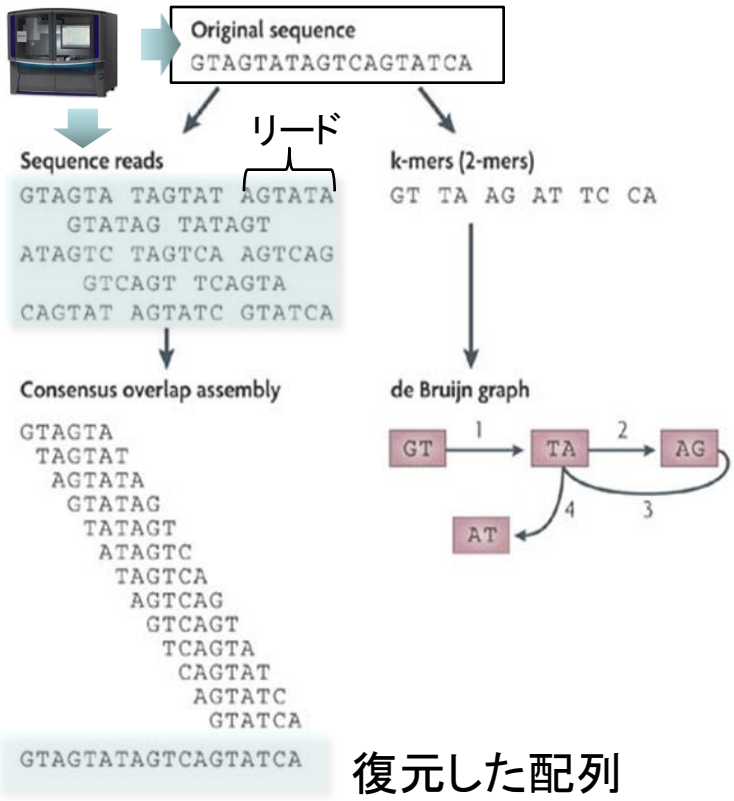
Read=short portion of the original genome

Huge set of reads are obtained

- **1read：ATGC（base）100(50basex2)**
- **Total ： 100million reads just for 1 run.**

100million

BIG-DATA



- **Assembling：**

Recover the original gene sequence

By combining 100million short reads.

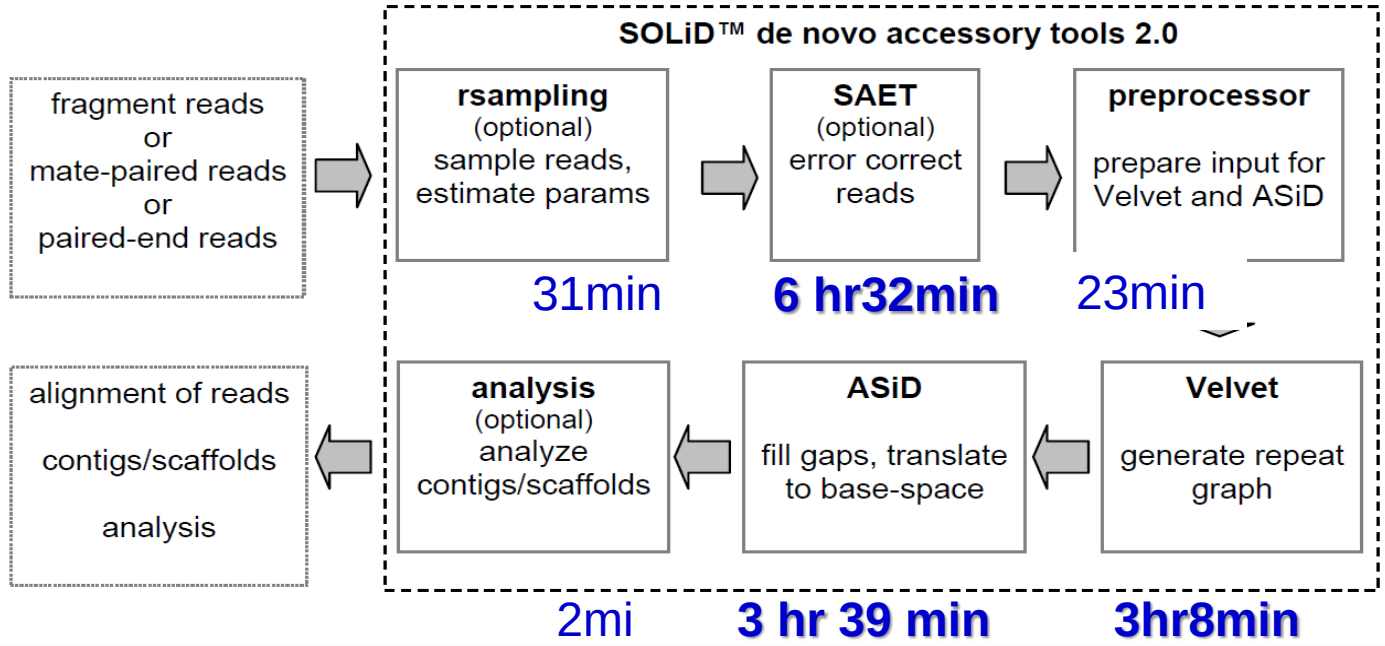
Current Assemble Software

- **The problem is too large ＝ Need to have huge memory in 1 machine**
  - **96GB for Aspergillus oryzae (=small)**
  - **→ Cannot Scale for large genome**
- **Single SMP machines → performance**

Time/Memory Comsuming

14 hours15minutes

To solve



Need to have New IT

Parallel Processing to achieve the Scalability and the Performance

IT Behind the Robot

VocaListener

VocaWatcher

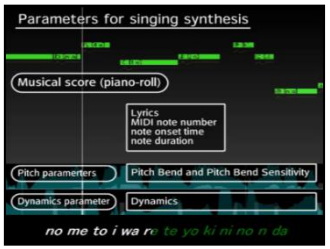
(same as the last year)

Media-Related R&D

- Computer Singing Systems

VocaListener

VocaWatcher

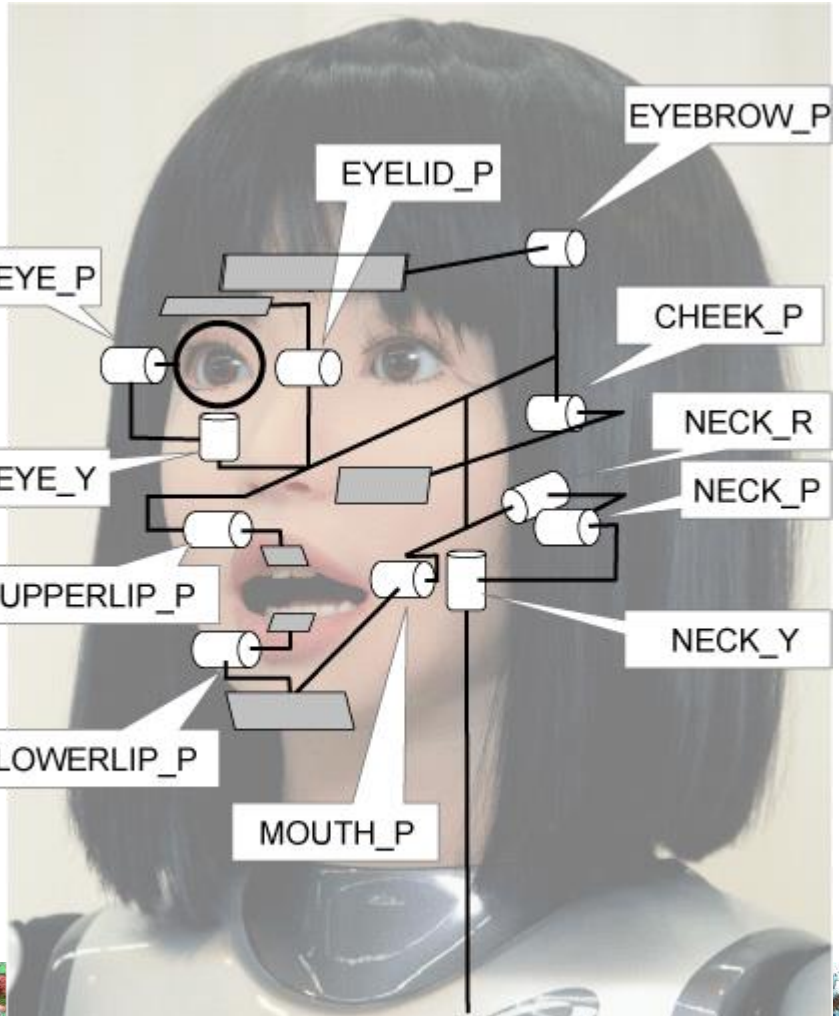


- Music-Listening Web Service

Songrium

Biped Humanoid Robot HRP-4C “Miim”

160cm height and 46kg weight with 44 degrees of freedo



【Neck control】
- NECK_Y
- NECK_P
- NECK_R

【Eye control】
- EYEBROW_P
- EYELID_P
- EYE_P
- EYE_Y

【Lip shape control】
- MOUTH_P
- UPPERLIP_P
- LOWERLIP_P
- CHEEK_P

Songrium(http://songrium.jp)

- Web service for active music listening

No. 76

Summary

AIST has many data-oriented R&Ds

- Geospatial
- Linked Data
- Bioinformatics
- Multimedia(Music/Songs)
- and more,,,

Some topics will be challenged by OSDC students

QUESTIONS?

GEO Grid

Grid-based e-infrastructure for geosciences

AIST