Big Data for solving societal problem

Masaru Kitsuregawa

Director General of National Institute of Informatics (NII)
Professor, Institute of Industrial Science, The Univ. of Tokyo
Former President of Information Processing Society (IPSJ)
ACM Fellow, IEEE Fellow

This is my second talk😊

Masaru Kitsuregawa

WEB Community mining and WEB log mining
Commodity Cluster based Execution (Keynote),

13th Australasian Database Conference (ADC2002), (2002.01)
Thanks for giving us
The Best Paper Award in ADC 2011;

"Potentiality of Power Management on Database Systems with Power Saving Function of Disk Drives"

by Norifumi Nishikawa, Miyuki Nakano, and Masaru Kitsuregawa

Big Data for solving societal problem

Masaru Kitsuregawa

Director General of National Institute of Informatics (NII)
Professor, Institute of Industrial Science, The Univ. of Tokyo
Former President of Information Processing Society (IPSJ)
ACM Fellow, IEEE Fellow
NII
National Institute of Informatics
The National Institute of Informatics (NII) seeks to advance integrated research and development activities in information-related fields, including networking, software, and content. NII also promotes the creation of a state-of-the-art academic-information infrastructure.

- SINET is a Japanese academic backbone network for more than 800 universities and research institutions, and for about 3 million users.
- SINET covers 100% of national, 84% of municipal, and 59% of private universities.

<table>
<thead>
<tr>
<th>National Universities</th>
<th>Municipal Universities</th>
<th>Private Universities</th>
<th>Junior Colleges</th>
<th>Colleges of Technology</th>
<th>Inter-Univ. Research Institutes</th>
<th>Labs and Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Organizations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>74</td>
<td>361</td>
<td>70</td>
<td>55</td>
<td>16</td>
<td>182</td>
<td>844</td>
</tr>
</tbody>
</table>

(As of March 2016)

- : SINET node
- : Domestic line (100Gbps or more)
- : International line (100Gbps)
- : International line (10Gbps)
SINET is a Japanese academic backbone network for more than 800 universities and research institutions, and current version, SINET4, started its operation in 2011. 
- SINET4 covers all 47 prefectures.
- SINET4 covers 100% of national, 80% of public, and 55% of private universities.

<table>
<thead>
<tr>
<th></th>
<th>National Universities</th>
<th>Public Universities</th>
<th>Private Universities</th>
<th>Colleges</th>
<th>Inter-Univ. Research Institutes</th>
<th>Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Organizations</td>
<td>86</td>
<td>65</td>
<td>320</td>
<td>62</td>
<td>55</td>
<td>191</td>
<td>803</td>
</tr>
<tr>
<td>Number of Connections</td>
<td>125</td>
<td>72</td>
<td>368</td>
<td>68</td>
<td>59</td>
<td>241</td>
<td>933</td>
</tr>
</tbody>
</table>

(As of March 2013)

Usage Example in High-Energy Physics

- Belle/ Belle II and Super Kamiokande experiments use SINET L3/L2VPNs in order to securely share large amounts of data between participating universities.
- ATLAS experiment for Large Hadron Collider (LHC) in Switzerland uses SINET international lines in order to share measured data between participating countries and Japan.

Belle/Belle II Measuring Devices for KEKB/Super KEKB

- Belle and Belle II
- Super Kamiokande

Neutrino

- Novel Prize in 2015

ATLAS Measuring Device for LHC

- Novel Prize in 2013
- Drs. Higgs and Englert

Higgs Particle

- Novel Prize in 2008
- CP Violation

Source: High Energy Accelerator Research Organization (KEK) and University of Tokyo
The ALMA project is a global partnership of Europe, North America and East Asia (led by Japan) in cooperation with Chile to operate a ultra-high performance radio telescope consisting of 66 high-precision antennas. The initial operation started by using 16 antennas in September 2011.

Credit: ALMA (ESO/NAOJ/NRAO)

Source: National Astronomical Observatory of Japan (NAOJ)
Usage Example in High-Performance Computing

• AICS, nine universities, two research institutions, and NII operate the High-Performance Computing Infrastructure (HPCI) through SINET for various scientific fields.

Usage Example in Seismology

• This project shares the data measured at each earthquake sensor by using SINET VPLS broadcast capability in order to determine the mechanism of the earthquakes.
**New Directions for Academic Infrastructure**

- **SINET5 plan** has been approved as one of the most important projects for research and education by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

**SINET5 will start in April 2016**

**Collaboration and Promotion in Research and Education**

- **Resource**
  - Promotion of academic information circulation and open access
  - Collaborative promotion of institutional repository expansion

- **Federation**
  - Collaborative enhancement of authentication between universities

- **Cloud**
  - Dramatic cost reduction and enhancement of research and education environment by tailored cloud services

- **Security**
  - Network flow analysis and dynamic control
  - Raise of security level for SINET users

- **Network**
  - Nationwide 100-Gbps backbone network and scalable network expansion
  - High-speed direct international lines to USA, Europe, and Asia
  - Introduction of new technologies such as SDN in response to user needs

**LAN among Universities**

- Getting Faster and Faster

- Line Speed as of Actual Condition Survey of Cyber Science Infrastructure (FY2013): LAN on Campus and Association Speed Outward

- SINET

- LAN among Universities

- Cloud Data Center

- Era of Cloud Computing

© 2016 National Institute of Informatics
**Big Data for solving societal problem**

Masaru Kitsuregawa

working with Prof. Toyoda, Prof. Goda, Prof. Yoshinaga, Prof. Ikoma, Prof. Nemoto, Prof. Ito

Director General of National Institute of Informatics (NII)
Professor, Institute of Industrial Science, The Univ. of Tokyo
Former President of Information Processing Society (IPSJ)

ACM Fellow, IEEE Fellow

‘big data’ attracts so much attention especially since 2012.
FOR IMMEDIATE RELEASE
March 29, 2012

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a “Big Data Research and Development Initiative.” By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some of the Nation’s most pressing challenges.

NSF director unveils big ideas with an eye on the next president and Congress

By Jeffrey Mervis | May 10, 2016, 3:30 PM

France Córdova, the director of the National Science Foundation (NSF) in Arlington, Virginia, has unveiled a research agenda intended to shape the agency’s next few decades and win over the next U.S.
NSF Ideas for Future Investment

**RESEARCH IDEAS**
- Harnessing Data for 21st Century Science and Engineering
- Shaping the New Human – Technology Frontier
- Understanding the Rules of Life: Predicting Phenotype
- The Quantum Leap: Leading the Next Quantum Revolution
- Navigating the New Arctic
- Windows on the Universe: The Era of Multi-messenger Astrophysics

**PROCESS IDEAS**
- Growing Convergent Research at NSF
- Mid-scale Research Infrastructure
- NSF 2060

2017 Jan 6th White House message

**Strengthening Global Sharing of Scientific Data and Research Findings**

Summary: Principles promote access to Federal government-supported scientific data and research findings for international scientific cooperation

Openly accessible scientific data can be a powerful catalyst in international scientific collaboration. To inform and improve consistency among Federal departments and agencies on open scientific data sharing in support of international scientific cooperation, the Interagency Working Group on Open Data Sharing Policy released a report.
Society 5.0

What's next?

Hunting & gathering society
In symbiosis with Nature

Agrarian society
Beginning of human organization

Information society
Increased value of intangibles & networks

Industrial society
Mastering of power & Mass production

BigData(IT) for Society
‘Digital Info-plosion’ project was eight years before ‘Big Data’
Digital Info-plosion Project
PI: M. Kitsuregawa
(Final symposium is on Jan 16th-18th 2012)

Information Grand Voyage Project
by METI (FY 2007-2009)

Corporate & Individual Web sites
Images, Videos and Music
A variety of DBs
Blog, SNS

Web Information

Spatial information
Purchase history information

Non-Web information
Audio Visual Information
Healthcare Information

Next generation information retrieval / analysis technology
Information Fusion
Valuation of Info-plosion
Just after end of Info-plosion project ‘Big Data’ became fashionable.

So “Big Data” was not surprise for us
Japan has Various Experiences for transforming society by BIG DATA

Societal Problem I

Mitigation of Natural Disaster (due to global warming)
Data Driven Disaster (Flood) Mitigation in Japan
2015 Kinugawa River Large scale Flooding
(dike break)
(Dike rebuilding: 7 days, drainage: 10 days)


Global Natural Disasters

https://standeyo.com/NEWS/10_Earth_Changes/100305.nat.diz.increasing.warning.html
Economical Damage

Four Dams in upstream of Kinugawa River

Total 100 MCM of flood water was stored in 4 dams in upstream of Kinugawa River with careful flood controlling dam operation *

(* store flood water as much as possible with careful monitoring of rainfall and downstream discharge)

Digital Info-plosion in DIAS
DIAS Today
 disk + tape > 25PB

AMeNOW!
Realtime Rainfall Information Service
Since 2015/10
Real-time Rainfall Information on Map
Available on PC, smartphone, tablet
No need for registration
History data, Time-sequential Movie, GPS, National Composition is available
http://rain.diasjp.net/
Predictive Discharge is possible?
Using numerical weather forecasting (18 hour prediction)

Peak cut of flood flow
Recovery of reservoir capacity after flood

Research (Big) data is key to Land Management under Global Warming.
Science could contribute to Societal Problems.
Data Integration and Analysis System
*a legacy for Japan's contributions to GEOSS*

enabling us to do **integrated research** and to realize **inter-disciplinarity**

DIAS
Meta data implementation
ISO 19115: Geographic information - Metadata

- ISO 19115 and its parts defines how to describe geographical information and associated services.
- The objective of this International Standard is to provide a clear procedure for the description of digital geographic datasets.
- ISO 19139 provides the XML implementation schema for ISO 19115 specifying the metadata record format and may be used to describe, validate, and exchange geospatial metadata prepared in XML.

Collaborative development of DIAS metadata implementation

Collaborative development activities with data providers (JAMSTEC) and system developers (Universities)

- We analyzed data providers’ holding dataset documentation.
- We collected metadata in the dataset documentation.
- We mapped between ISO 19115 metadata elements and the dataset documentation items.
- We selected useful keyword lists in order to search effectively.
There are two boxes side by side when you need to input the field in both English and Japanese.

Societal Problem

Help developing countries

Mitigation of Natural Disaster
Our team visit (by Rasmy)

- 553 mm in 2017
- 521 mm in 2003

Peak discharge in 2003

Historical flood level

2017 flood level

Flood debris on electric post
Rain Gauge

Six sensors on Kalu River basin more than 500mm/day
JAXA- Global Satellite Map of precipitation

Health-care for developing countries

Opportunities and Challenges in a Era of Big Data

KDD2015
Create Data if there is no data.
Health-care for the Base of the Pyramid

Bangladesh

- Area: 144,000km² (1/3 of Japan)
- Population: 140,000,000
- Population Density: 982 people/km² (3 times as of Japan)
- GDP (PPP) per person: US$1615 (1/21 of Japan)

8 men have the same wealth as 3.6 billion of the world’s poorest people. We must rebalance this unjust economy.
Grameen Village Phone Lady

Triage

Tele-medicine

Mobile-Phone

Internet

Fixed-Line Phone
RESULT OF TELEMEDICINE IN FY2013

Visitors n=10575
Telemedicine(n=3736)

Emergent(3%)
Affected(32%)
Healthy(11%)
Caution (54%)

Ask to visit 2 months later

At the first visit
2nd visitors, n=1093
2nd Health check-up is not finished

Urgent(11%)
Affected(89%)
n=1093

Healthy(5%)
Caution (44%)
Affected (39%)

Healthcare Instruction

※We have to think “regression to the mean”

*Orange was increased in FY2013 because B-Logic was revised.

Differences between Bangladesh and Japan

• JP: from National Health/Nutrition Survey in 2013, and National Health checkup in Chiba prefecture in 2011
• Age and Sex are adjusted
• Waist are compared by Japanese criteria of metabolic syndrome
Further Reduction of the Cost of Health Checkup By ROC (receiver operating characteristic)

Predict BS by questionnaire and other results of sensors

If we allow 7.3% (n=113) oversight, we can reduce 94.8% (n=14392) cost of blood sugar test

ROC Curve by blood sugar vs. questionnaires and sensors

We used Random Forest method with 60% of 16733 data for training data and 40% testing data.

AUC = 0.98

Data can contribute to the healthcare system for developing countries.
Reduce traffic accidents

Substantial Reduction of Traffic Accidents

Detection of Potentially High Risk Accident Spots with Drive Recorder Data Analytics
Traffic Accidents are Still Troubling Our Society

- In 2014, a total of 573,842 traffic accidents occurred in Japan
- Only Reported one

- To reduce the number of accidents
  - Many local governments have made potential risk maps of traffic accident spots
  - The transportation industry has started to introduce dashcams or vehicle recorder systems

Map for risky places in car navigation system would help substantially
Official Risk Maps Lack Accuracy, Comprehensiveness, and Density

Potential risk maps provided by a local government:
- Only around neighborhood of elementary schools
- Reflecting vague human impressions

Actual accident places map provided by Police:
- Very sparse!!

There are Many Potential Risk Places

Actual accident places:
- Very sparse!!

Exploring Caution Spots from Vehicle Recorder Big Data:
- Many Potential Risk Places
  - Reflecting time and weather information
Example B: A Caution Spot in the Middle of a Narrow Road with No Crossings

'Automatic' Identification of high risk accidental places

- Learning set: Accidents place made by police

- *Sharing data* would increase the accuracy and coverage! and help to create less - traffic accident society.
Societal Problem  V

Automatic Monitoring of Road Condition

Smart Infrastructure 2016
Sensor based smart health check for roads and bridges

Which driver will make an accident?
> 80% f-measure (10 fold validation)

<table>
<thead>
<tr>
<th>method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age・History・Sex</td>
<td>0.42</td>
<td>0.88</td>
<td>0.57</td>
<td>0.62</td>
</tr>
<tr>
<td>+ Premier certificate</td>
<td>0.45</td>
<td>0.84</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>+ Driving Style (probability)</td>
<td>0.50</td>
<td>0.84</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td>+ Driving Style by Velocity (probability)</td>
<td>0.55</td>
<td>0.82</td>
<td>0.66</td>
<td>0.73</td>
</tr>
<tr>
<td>+ Driving Style by Road Width (probability)</td>
<td>0.57</td>
<td>0.81</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>+ Driving Style by Velocity and Road Width (KL)</td>
<td>0.86</td>
<td>0.83</td>
<td>0.85</td>
<td>0.88</td>
</tr>
</tbody>
</table>

| no information                                   | 0.37      |        |           |      |

**Conclusion**

**Societal Benefit First**

We showed some real examples to improve ‘societal benefits’ beyond commercial benefits.

There is huge potential for ‘Societal Analytics’

Let us grow the national and international societal benefits by *sharing*, analytics methods, data, and deliverables to help each other.