ISIP 2019

Visual Analytics of Multiple Media and Real World Big Data

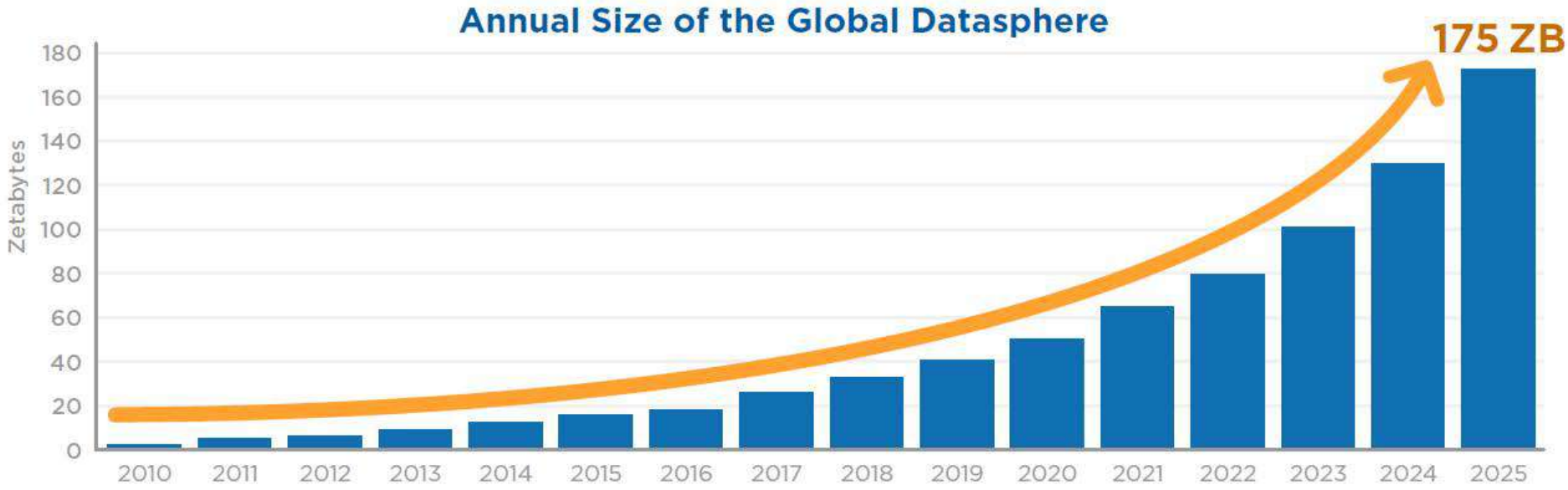
Masashi Toyoda

Center for Socio-Global Informatics,

Institute of Industrial Science,
The University of Tokyo

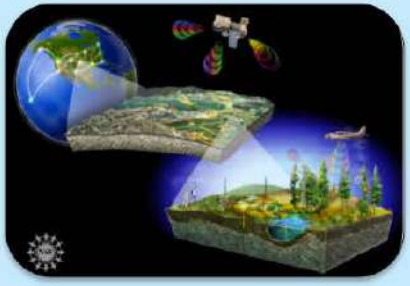
Big Data Is Still Big

- The size of all new data in a year grow to 175 ZB by 2025
- Billions of IoT devices are expected to create over 90 ZB in 2025



Data Age 2025: IDC sponsored by Seagate (2018/11)

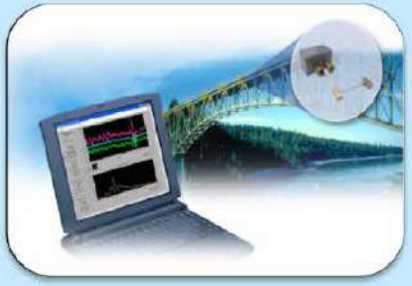
Applications of Big Data Analytics



Environment & Sustainability

Health & Wellbeing

Emergency Response & Disaster Resiliency



Manufacturing, Robotics, & Smart Systems

Secure Cyberspace

Transportation & Energy

Education & Workforce Development

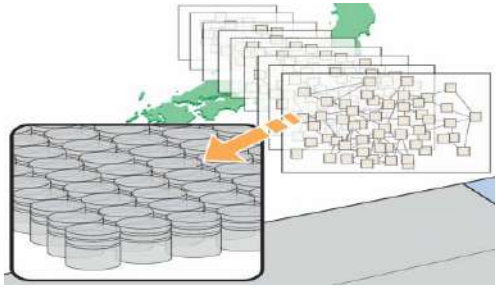
The NSF/CISE Framework for Investments in Big Data: Keith Marzullo, 2012

Big Data Solution

Visualization and Decision Making

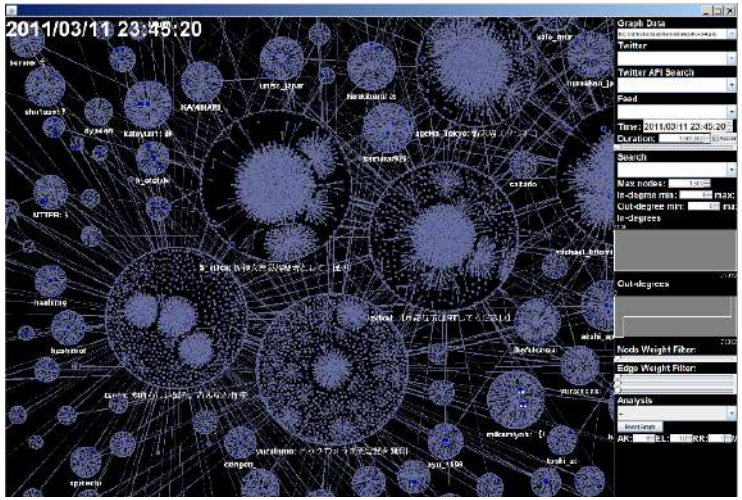
Data Analytics

Collection and Archive

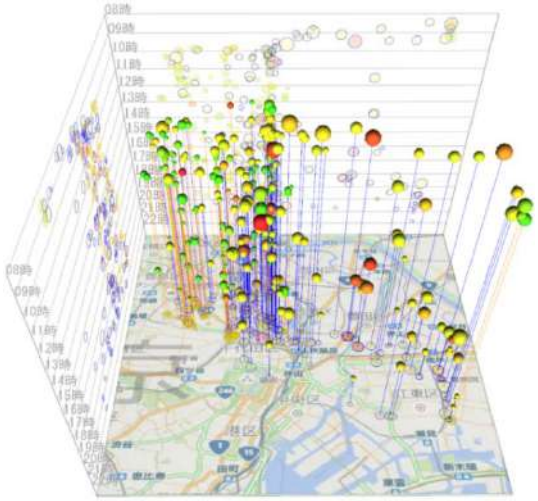


Large-Scale Interactive Visual Analytics

Zooming UI



3D Info. Vis.

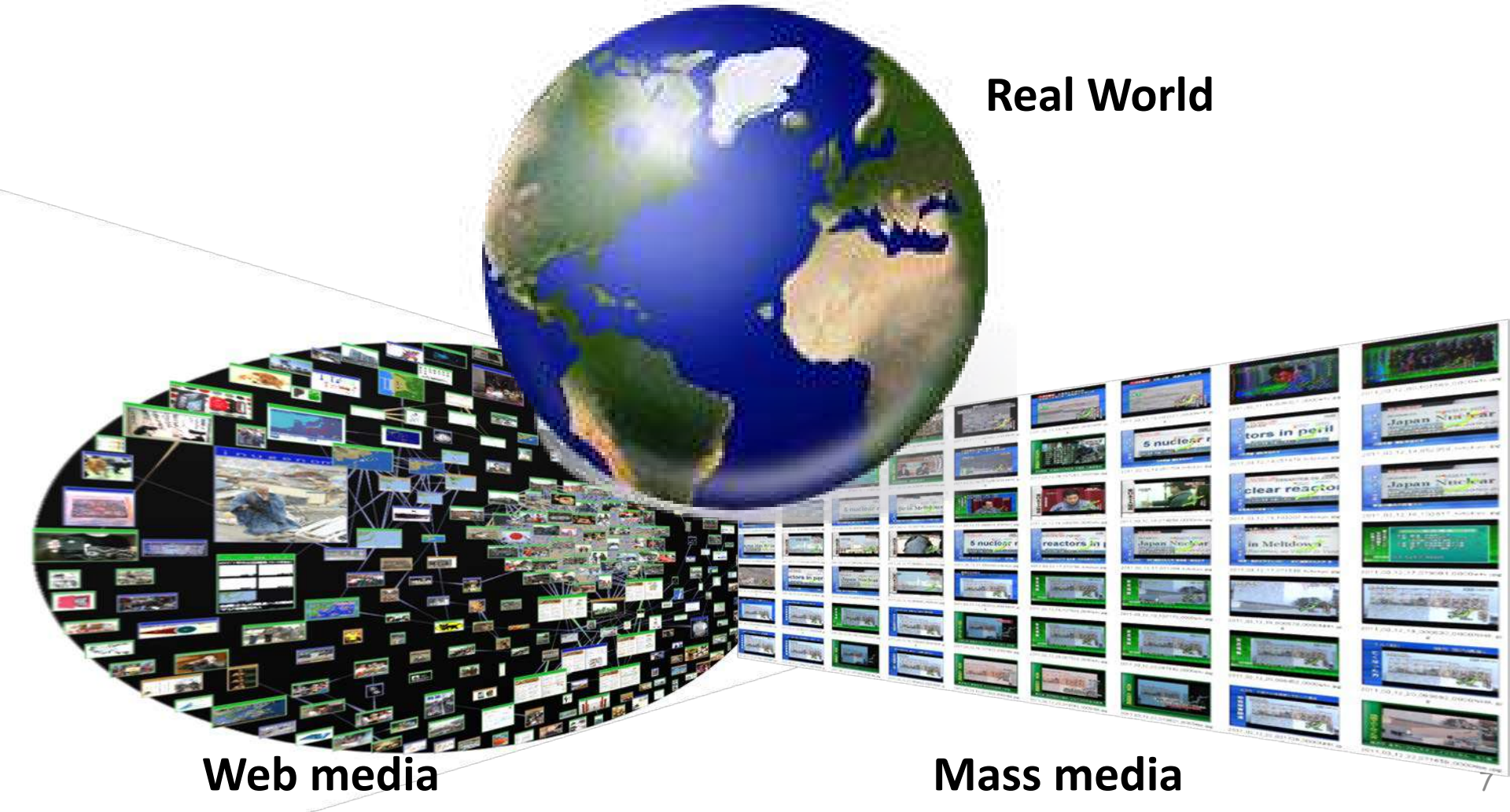


Data Analytics

Big Data Collection and Archive

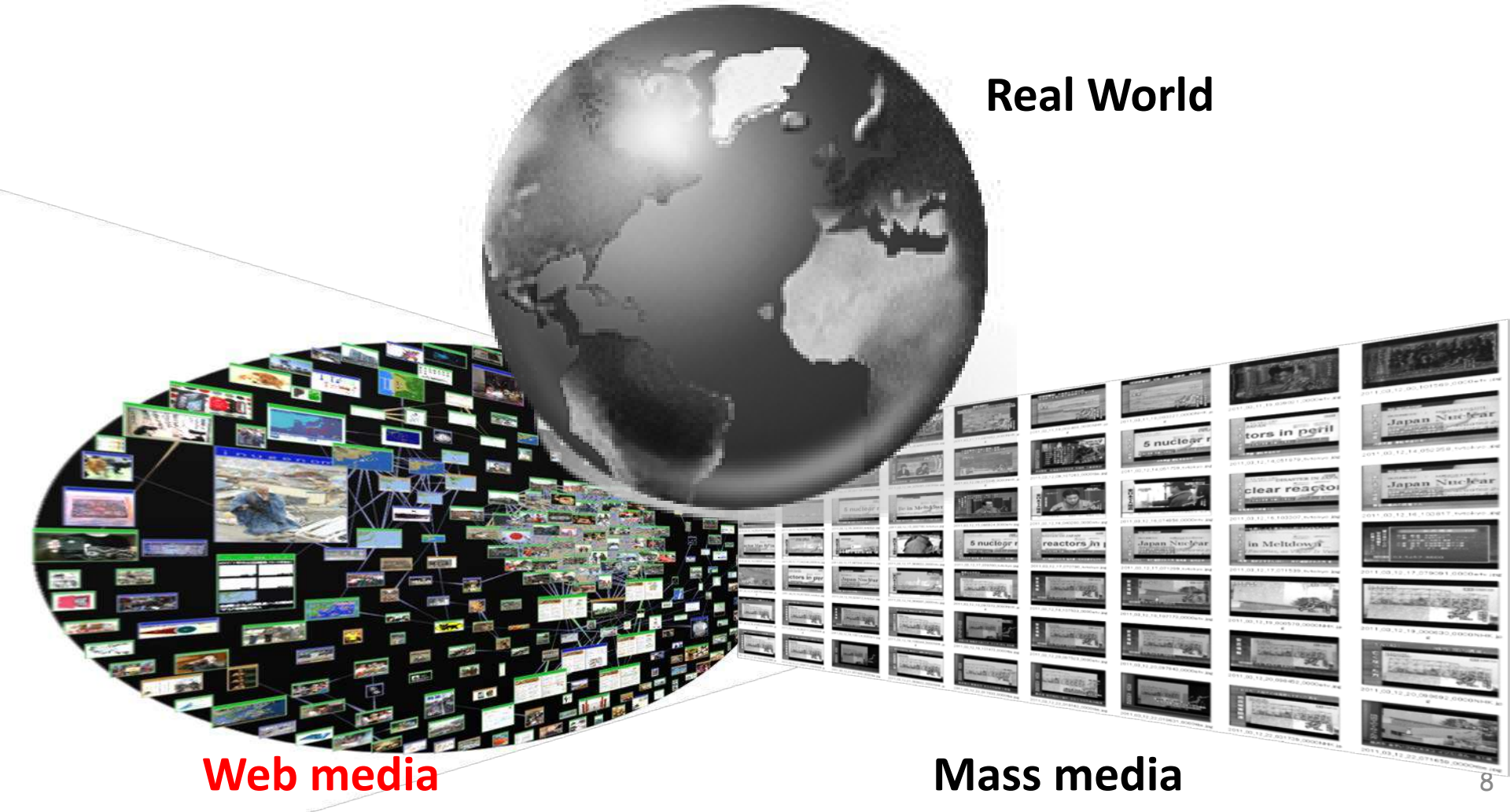
Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data

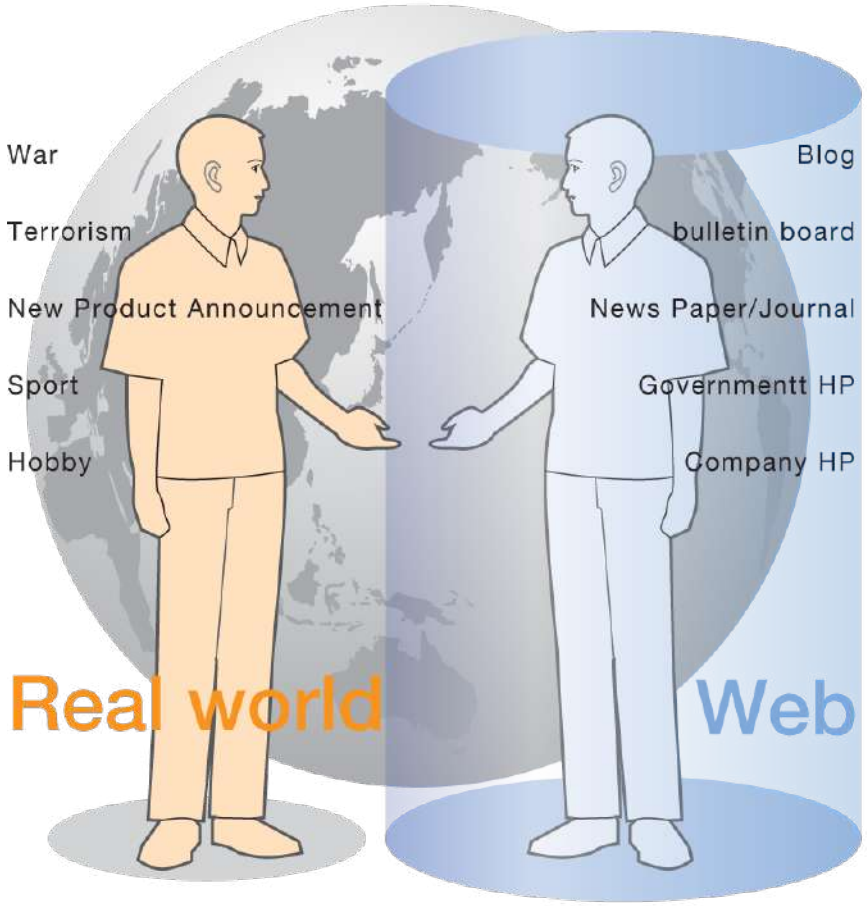


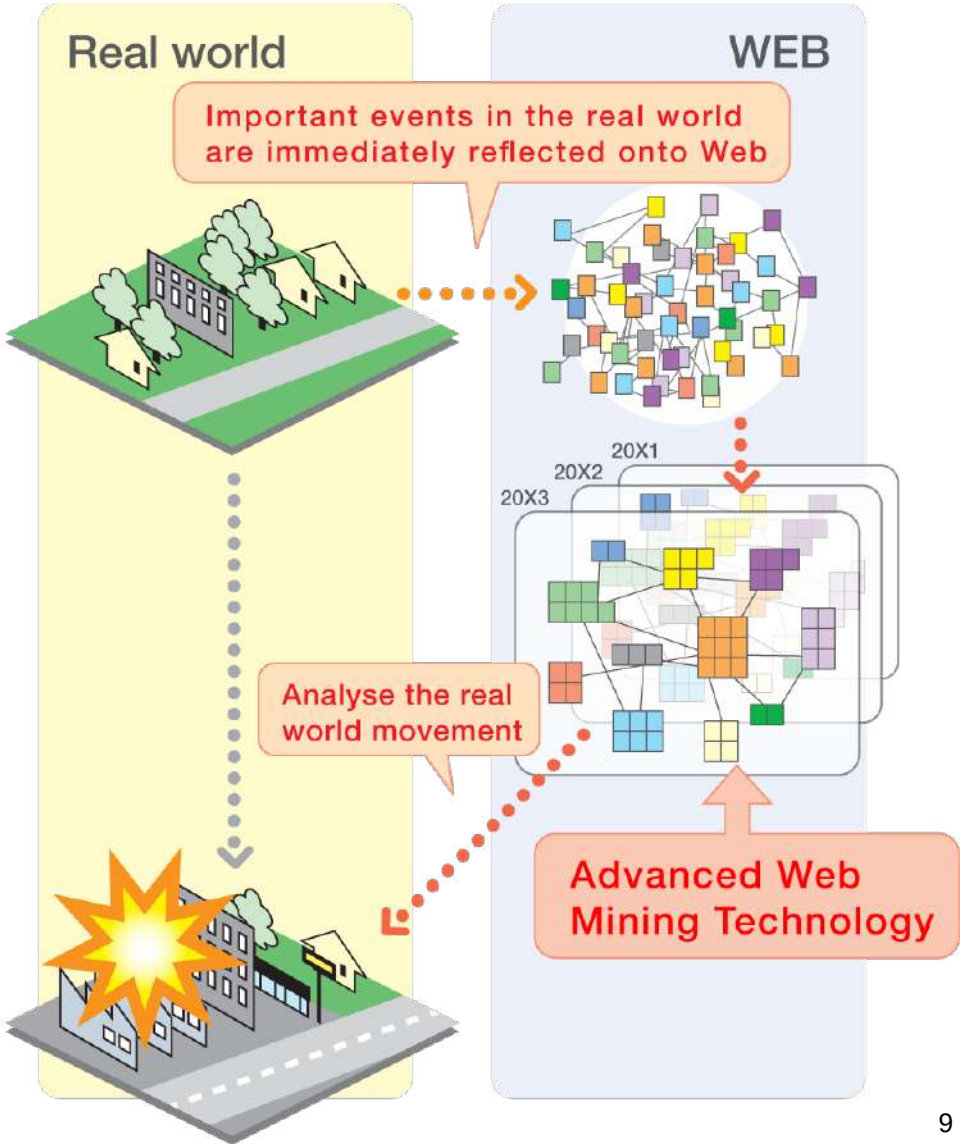
Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data



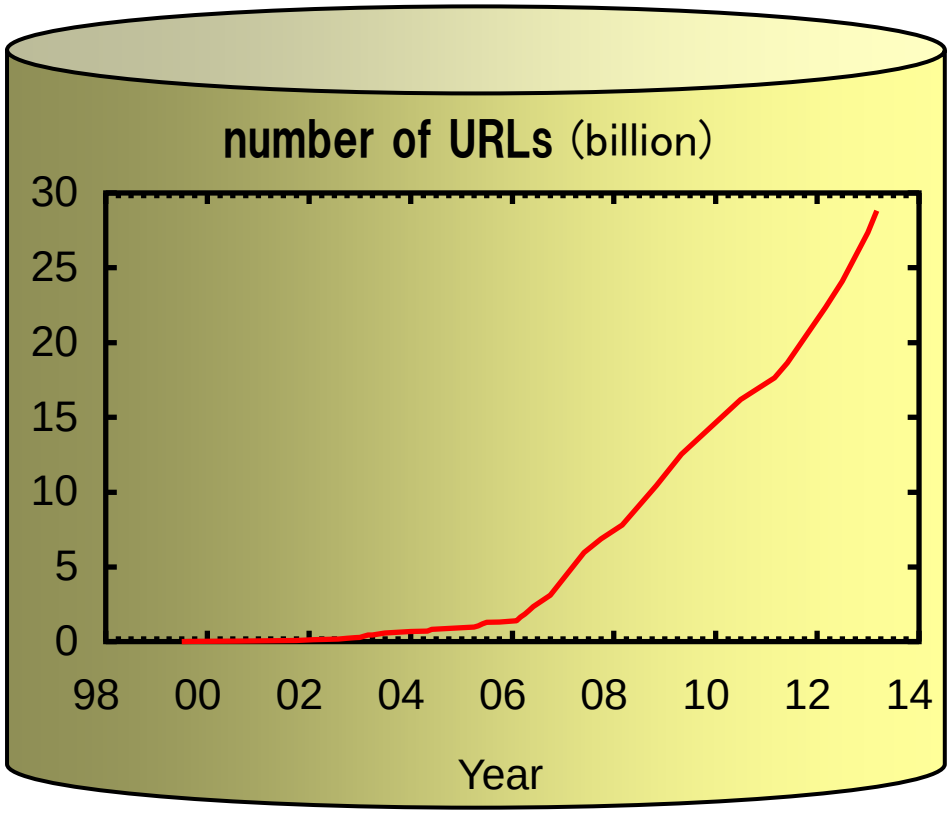
Web as a Sensor for Society





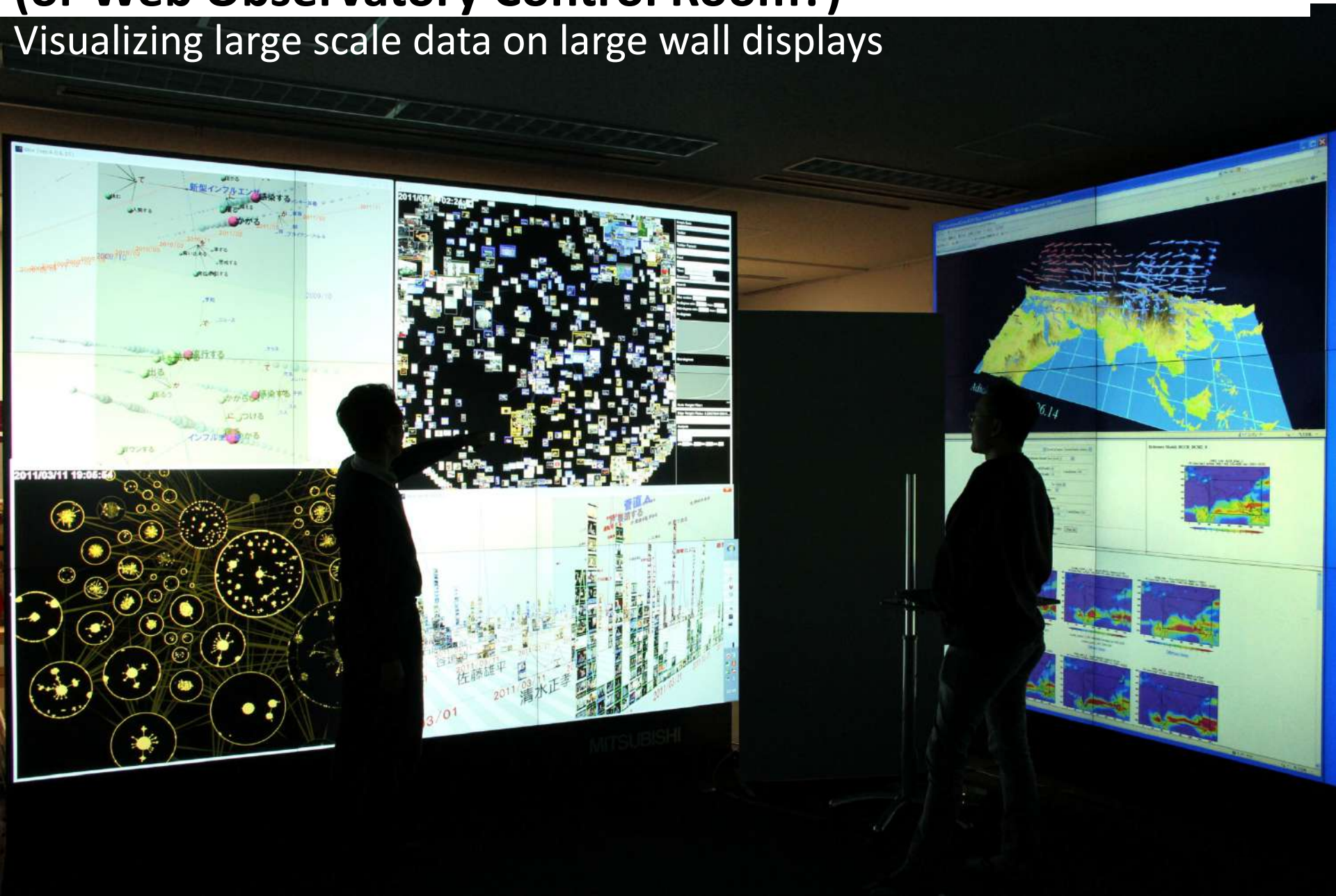
Web Archives in UTokyo

- **Japanese Web Archive**
  - Focused crawl of Japanese pages in every domains (1999 -)
  - Contents of 60 billion URLs
  - One of the largest Web archives
- **Blogs**
  - From 2006
  - 2.5 million feeds
  - About 3 billion articles
- **Twitter**
  - From 2011
  - 2.5 million users
  - About 40 billion tweets



Big Data Visualization Platform
(or Web Observatory Control Room?)

Visualizing large scale data on large wall displays



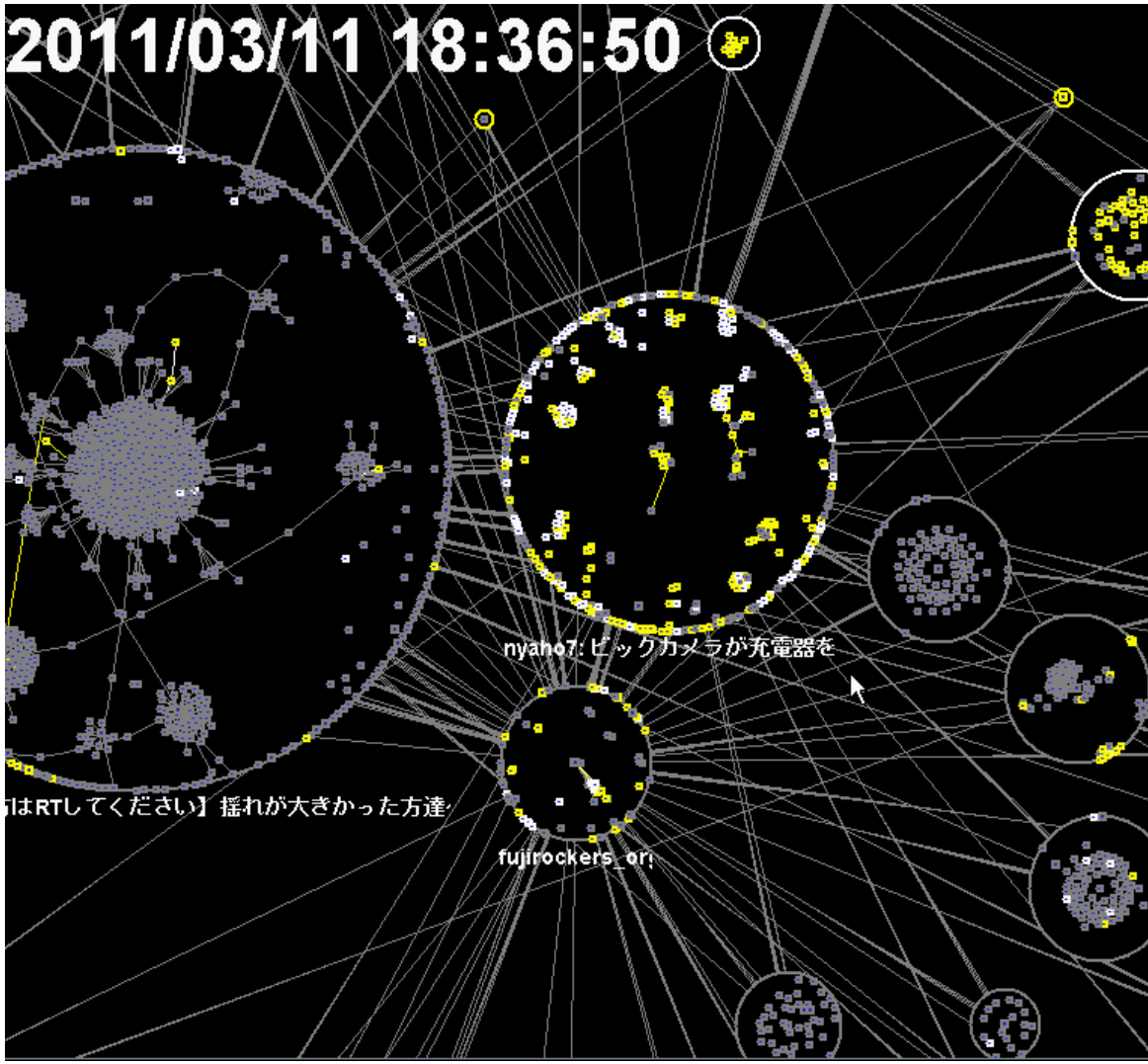
Information Diffusion on 3.11 in Twitter

- Twitter was one of the important information sources for evacuation in Tokyo
- Situation in Tokyo:
  - All trains and metros were stopped
  - Highways were closed
  - Millions of people walked to home or shelters
  - Cell phones: voice didn't work, data worked but narrow

Challenge: Observing important information diffusion and supplying right information in right timing

Visualizing Information Diffusion in Twitter

- Graph visualization of mentions and retweets between tweets
- Tweets are automatically clustered by their text similarities
- Temporal changes in graphs are animated



Graph Data

p://kanda.tkl.iis.u-tokyo.ac.jp/~toyoda/twitter2/h

Twitter

Twitter Parsed:

Feed

Web Community Search

Time: 2011/03/11 15:10:00

Duration: 1 Accum.

Search

Max nodes: 2,000

In-degree min: 0 max

Out-degree min: 0 ma

In-degrees

1136

Out-degrees

2

Node Weight Filter:

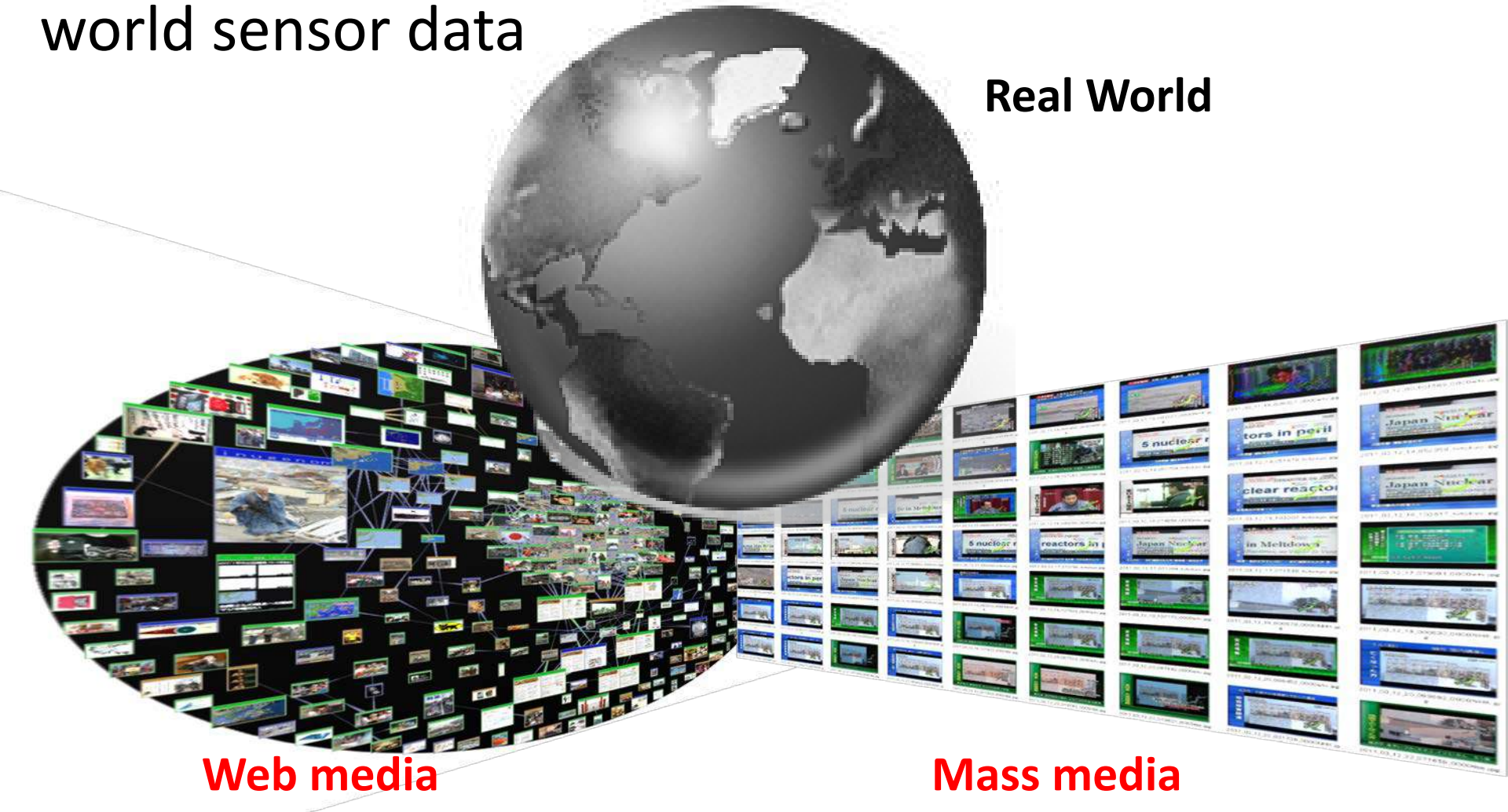
Edge Weight Filter:

Analysis

-

Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data

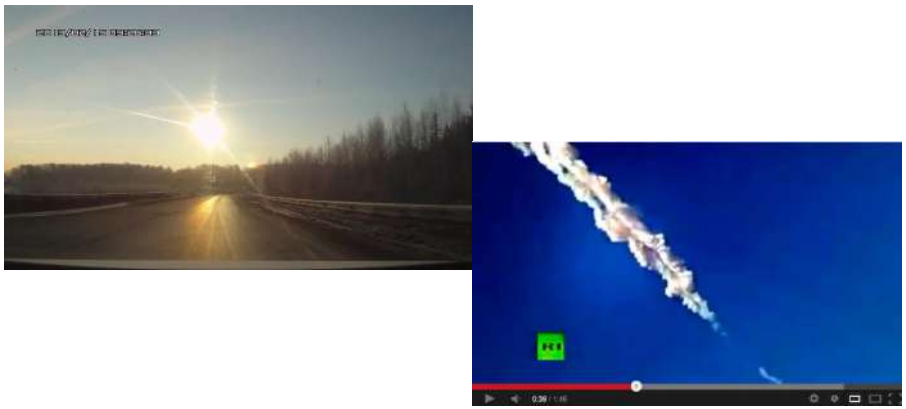


Interaction between Media

Mass and web media affect each other

- **Mass media** still have big influence
  - Major topics in Web/social media are coming from TV
- **Web media** involve diverse services and become important information sources for mass media
  - Blogs, microblogs, SNS, photo/video/link sharing..

Miracle on the Hudson



Chelyabinsk meteor

Broadcast News Video Archive in NII

- TV-RECS in National Institute of Informatics
  - News program mainly from NHK
    - March 2001 -
  - 24-hour, 7 channels in Tokyo area
    - August 2009 -
  - 300,000 hours, 1 petabyte
  - closed-caption text and electronic program guide (EPG).

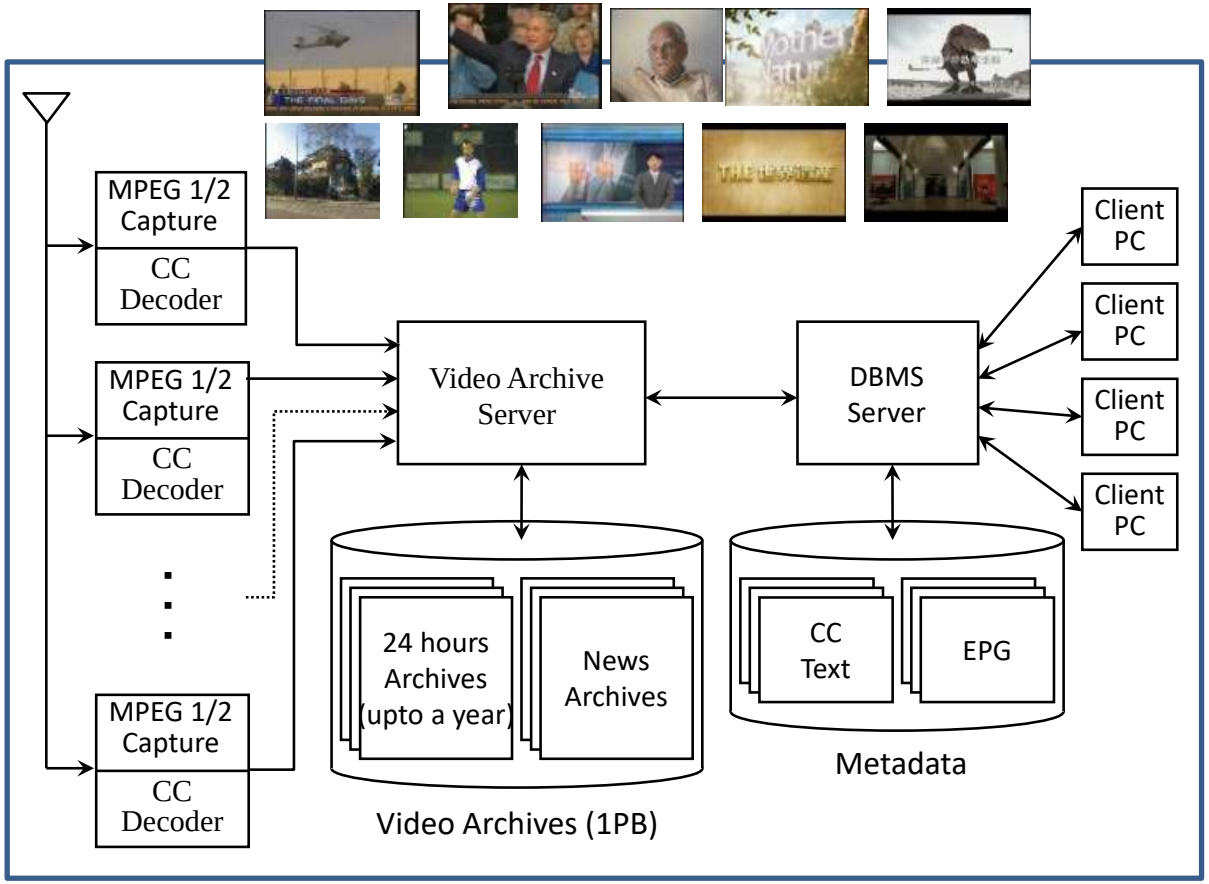
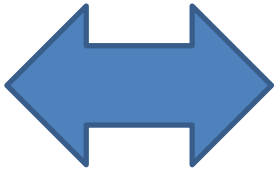


Image Flows Visualization for Inter-Media Comparison

Masahiko Itoh, Masashi Toyoda (The University of Tokyo),
Cai-Zhi Zhu, Shin'ichi Satoh (National Institute of Informatics),
Masaru Kitsuregawa (National Institute of Informatics, The University of Tokyo)

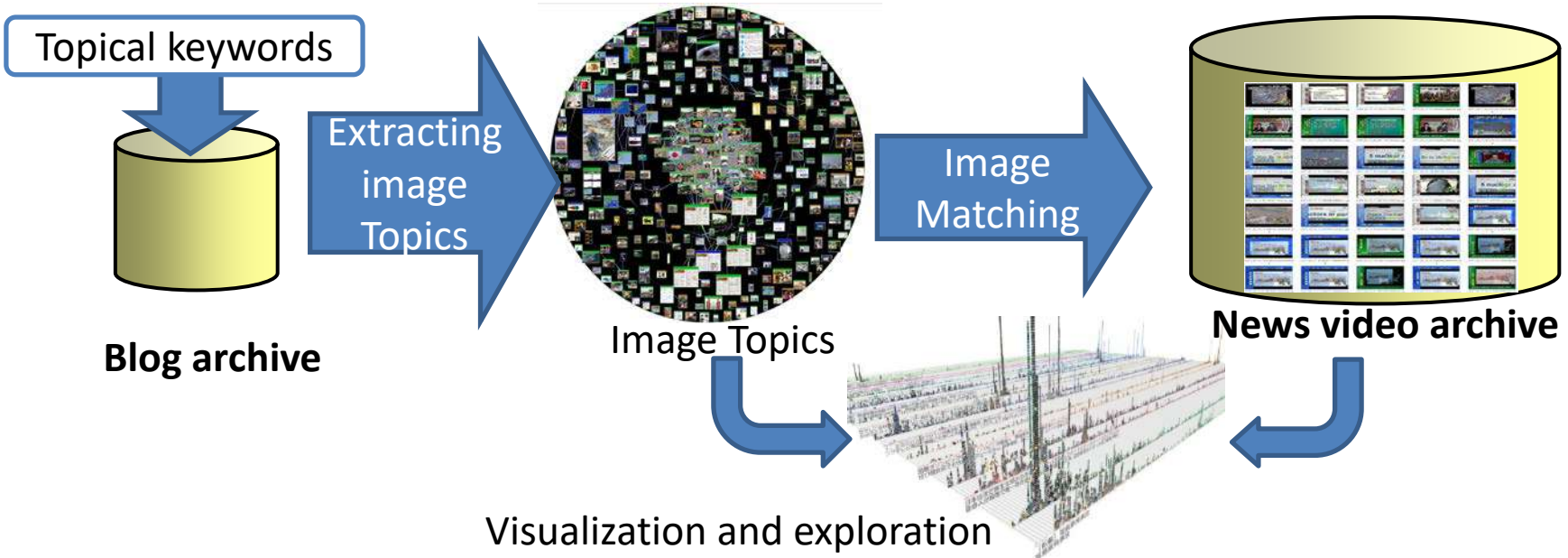
Tracking Diffusion of Topics through Blogs and TV News

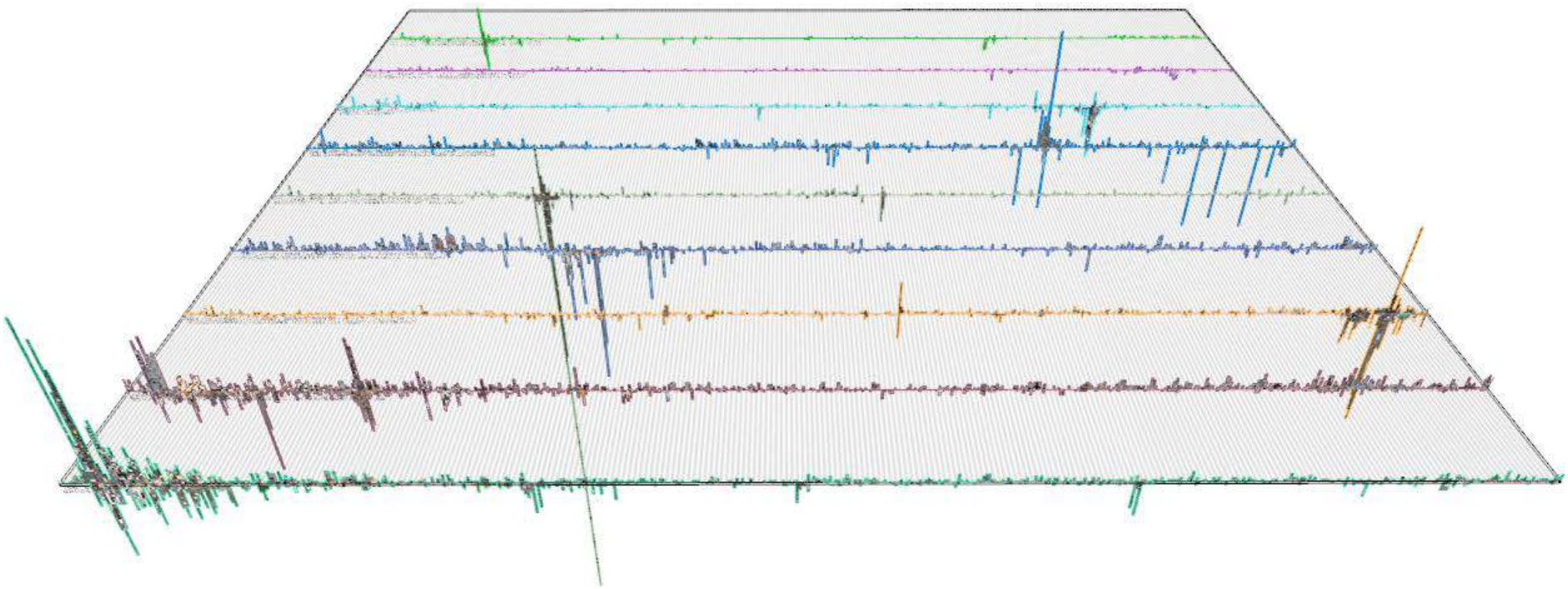
- What kinds of contents become popular in each medium?
- For a given topic, which medium first provided the contents?
- How the exposure of contents in one medium affect the other?



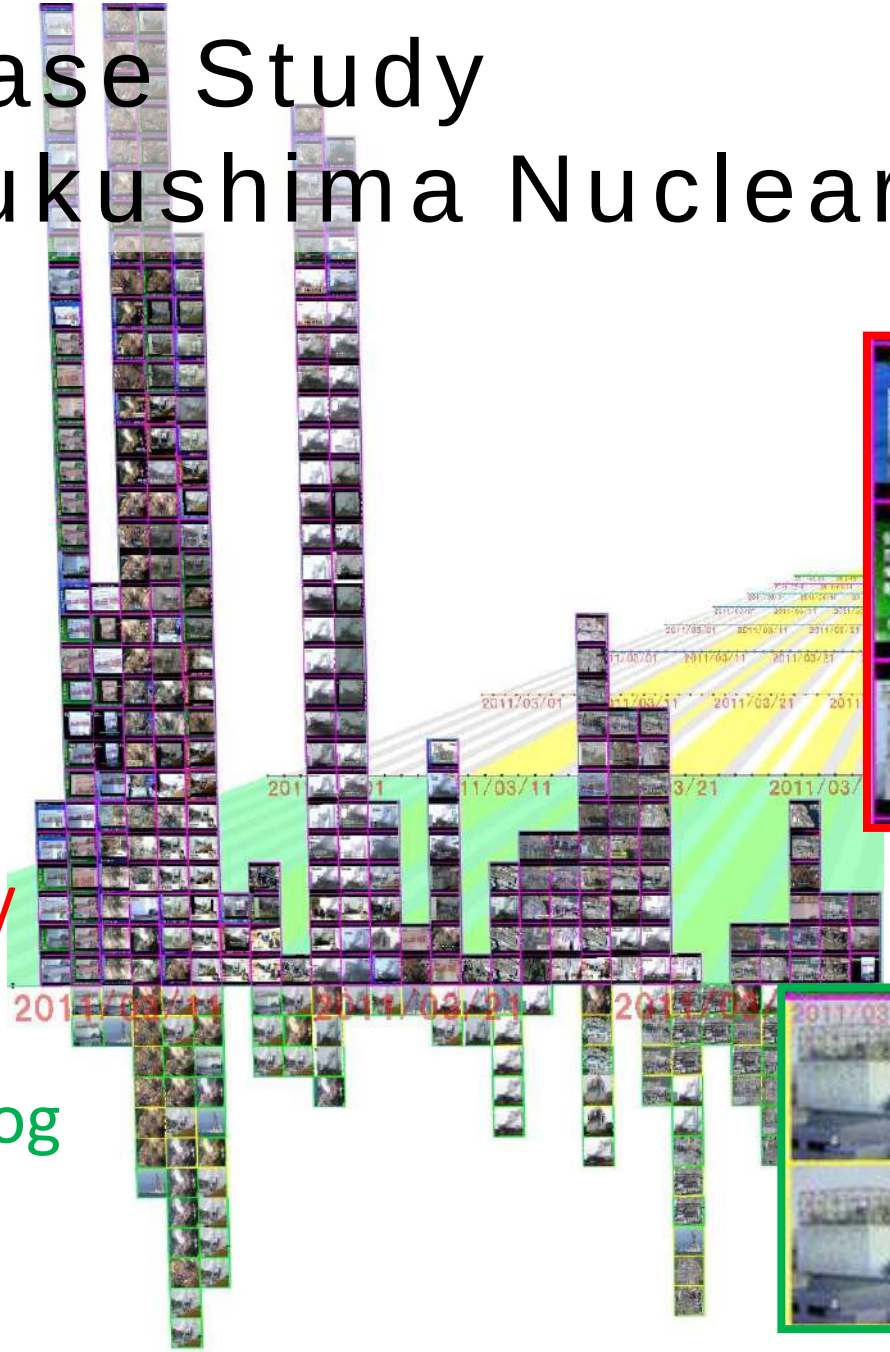
Topic Tracking based on Image Matching

- Examine influence and diffusion of image contents
  - Clustering duplicate and near-mirror images in blog posts
  - Topics are created by grouping image clusters based on their similarity of surrounding texts
  - For each topic, retrieve TV shots relevant to each image clusters in the topic
  - Visualizing image flows from blogs and TV

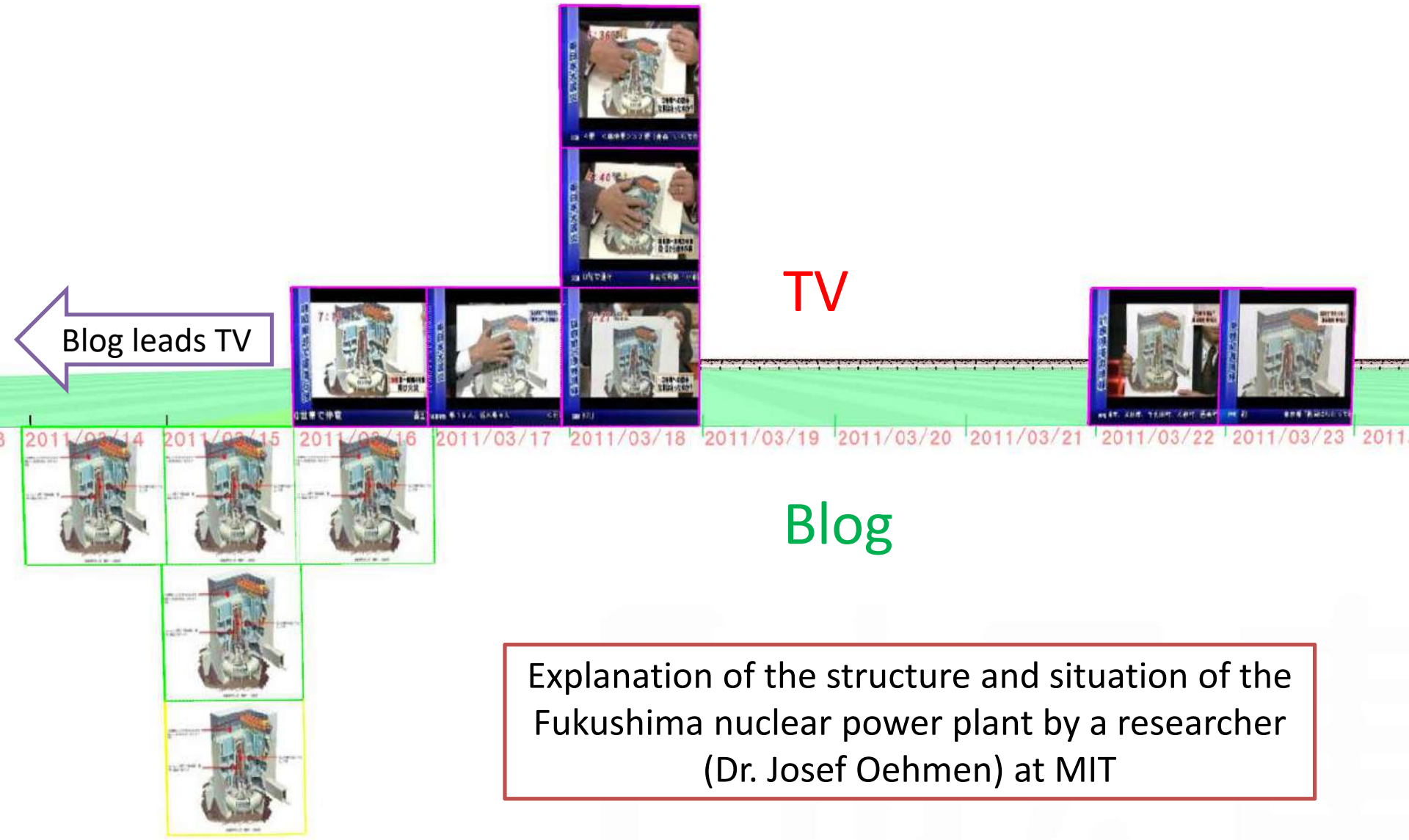




Case Study
Fukushima Nuclear Disaster



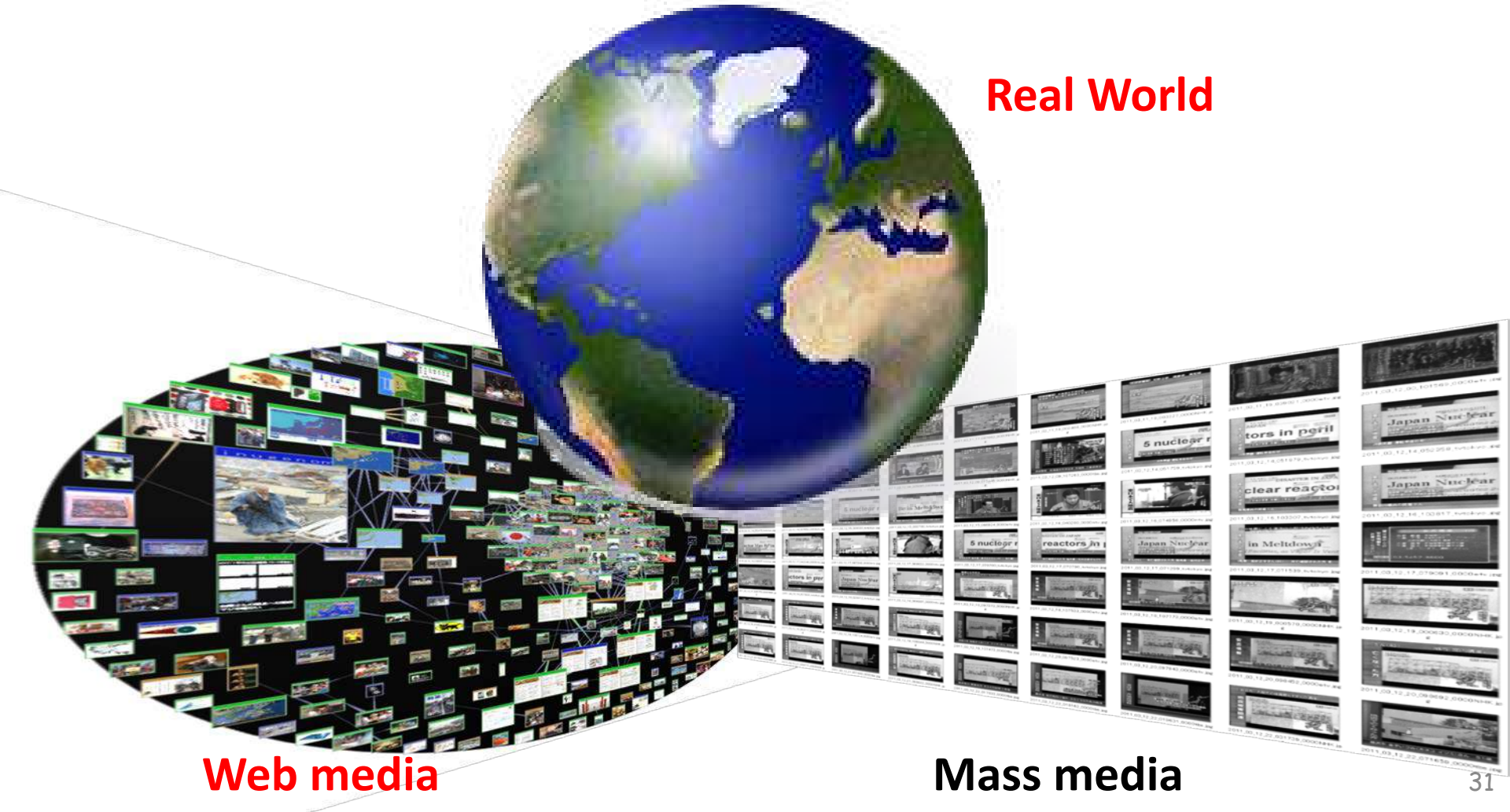
Case Study
Spotting the origin of images



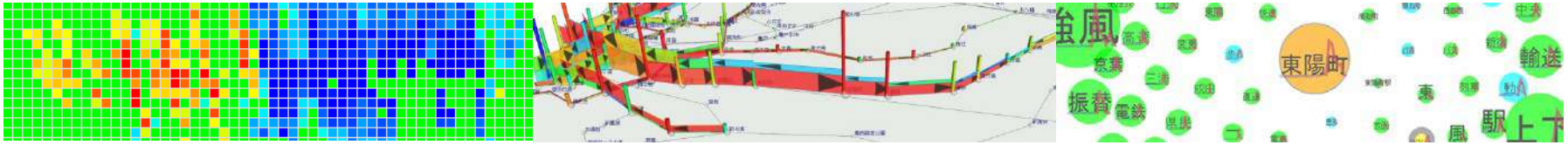
Explanation of the structure and situation of the Fukushima nuclear power plant by a researcher (Dr. Josef Oehmen) at MIT

Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data



Visual Fusion of Mega-City Big Data: An Application to Traffic and Tweets Data Analysis of Metro Passengers



Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda,
Yoshimitsu Tomita, Satoshi Kawamura, Masaru Kitsuregawa

東京大学 THE UNIVERSITY OF TOKYO Tokyo Metro NII

Transportation Systems in Mega Cities

TOKYO:

- 13 million population
- Millions commute from adjacent prefectures

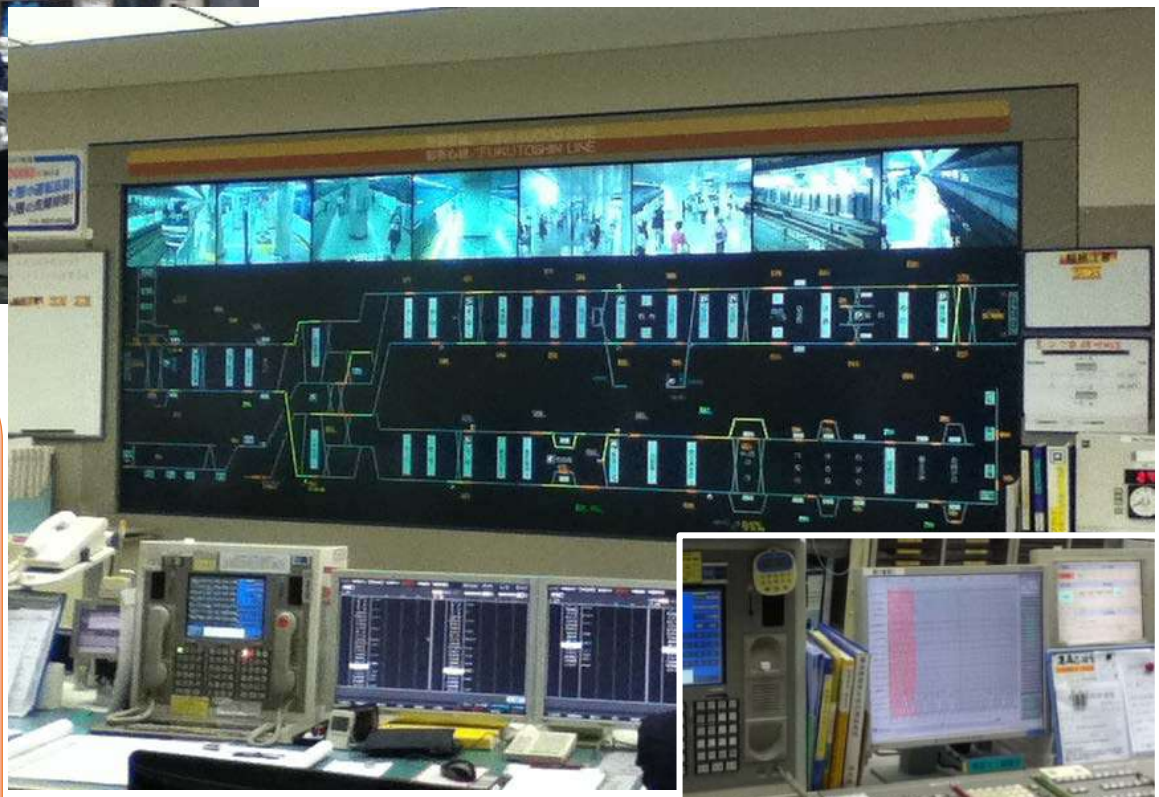
Disasters

Accidents

Public gatherings

Control Room of Tokyo Metro

- Focusing on train movement
- Passengers are monitored by surveillance cameras



Hard to observe passengers':

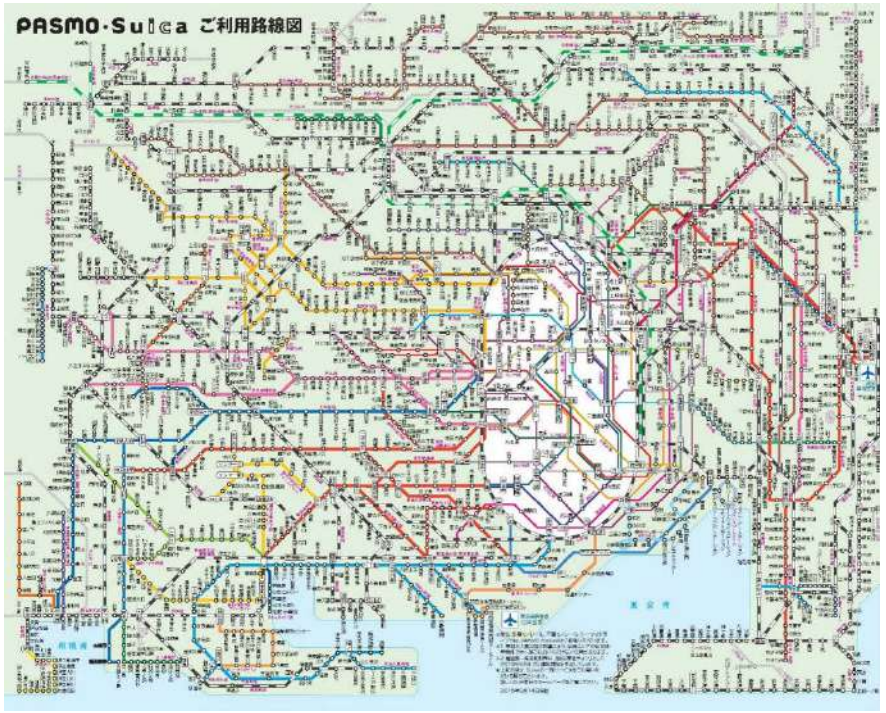
- Origins and destinations
- Movement
- Thought and complaints

Smart Card (PASMO/Tokyo Metro)

- Large scale origin-destination-time records
  - Demand and flow of passengers can be analyzed

30 million cards

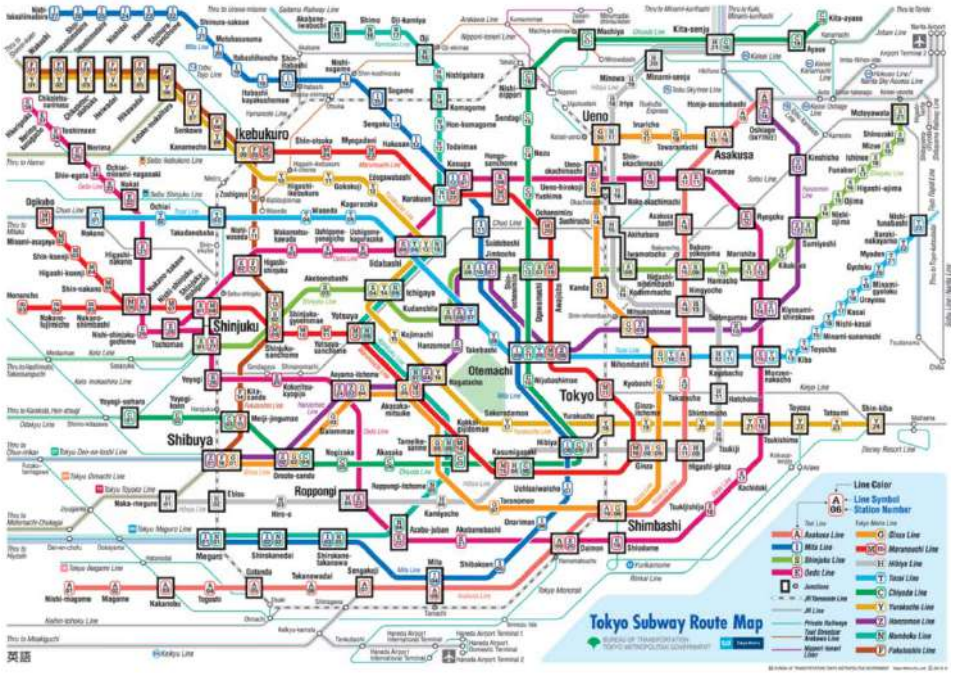
8,000 automatic gates



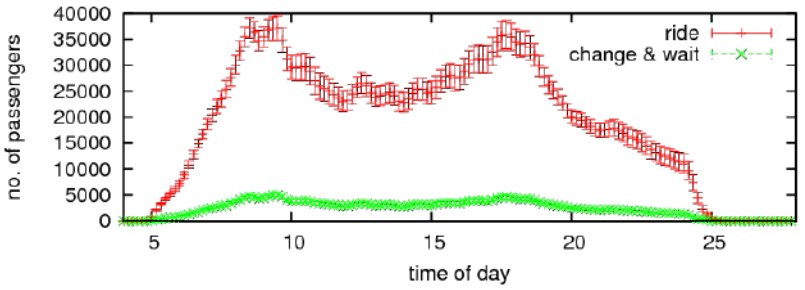
Available over Kanto-area including Tokyo

Smart card data

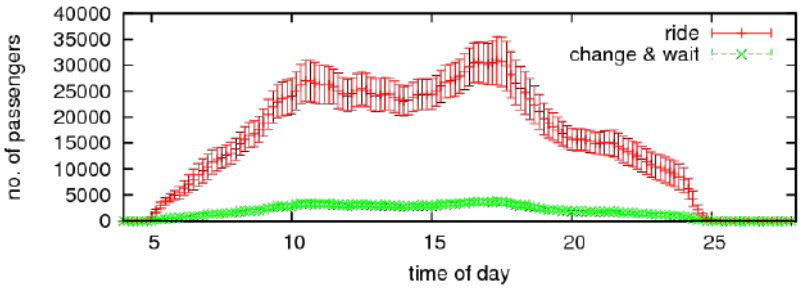
- Tokyo Metro trip records
  - Covers almost all of the Tokyo business area
  - Two years' worth
  - 28 lines, 540 stations, 300M trips



Weekday



Weekend and holiday



Number of passengers at every 10 min. (1 year average)

System overview

Estimate trip routes of every passengers (location and time info)

17:45 Shibuya
17:48 Omote-sando
17:52 (train exchange)
17:55 Nogizaka
17:58 Akasaka

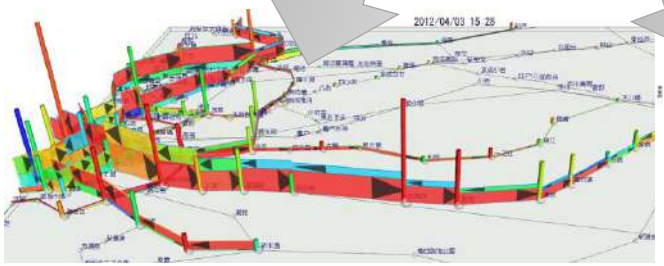
Path log

Collected once a day

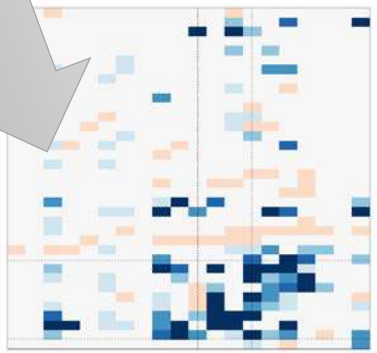
Entrance, Exit, Exit-time
(Shibuya, Akasaka, 18:00)

Trip log

Station gate



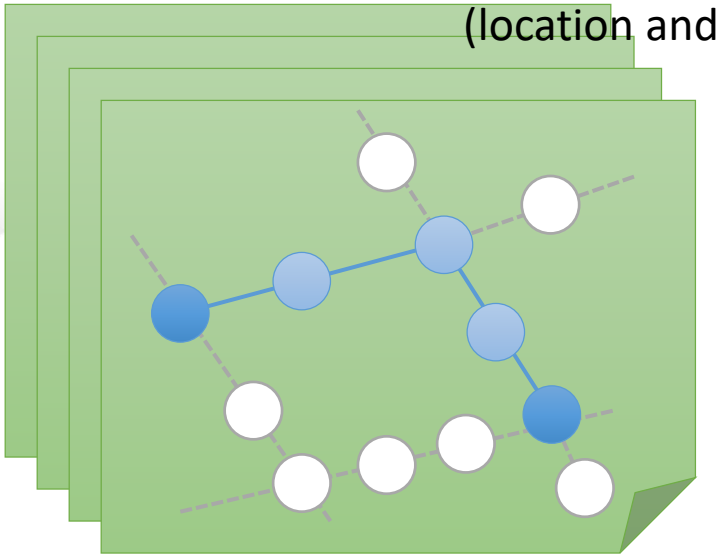
Interactive visualization



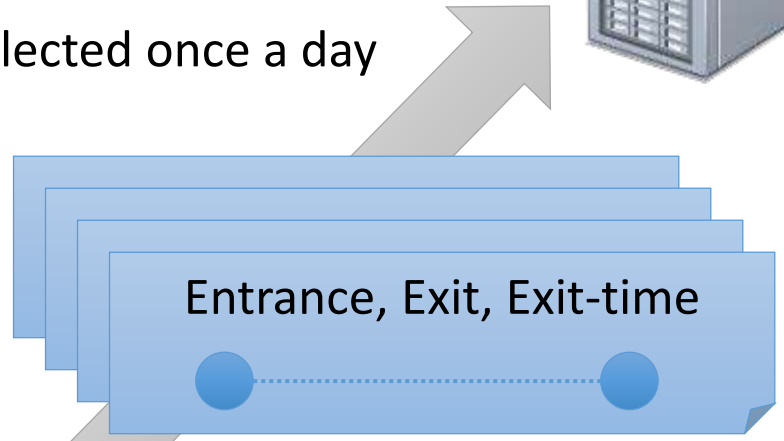
Simulation

System overview

Estimate trip routes of every passengers (location and time info)



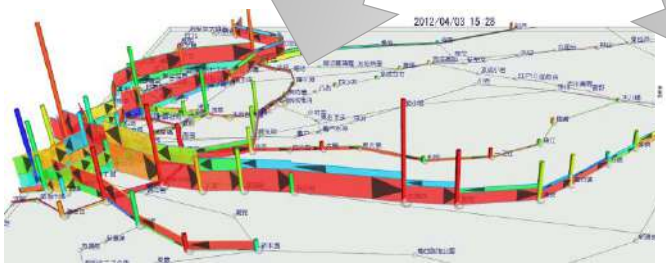
Collected once a day



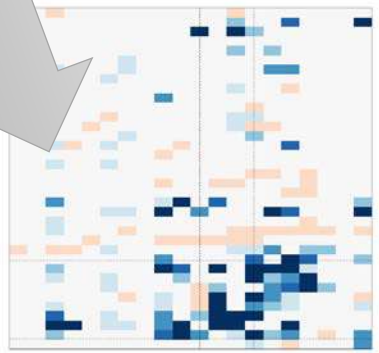
Trip log

Path log

Station gate



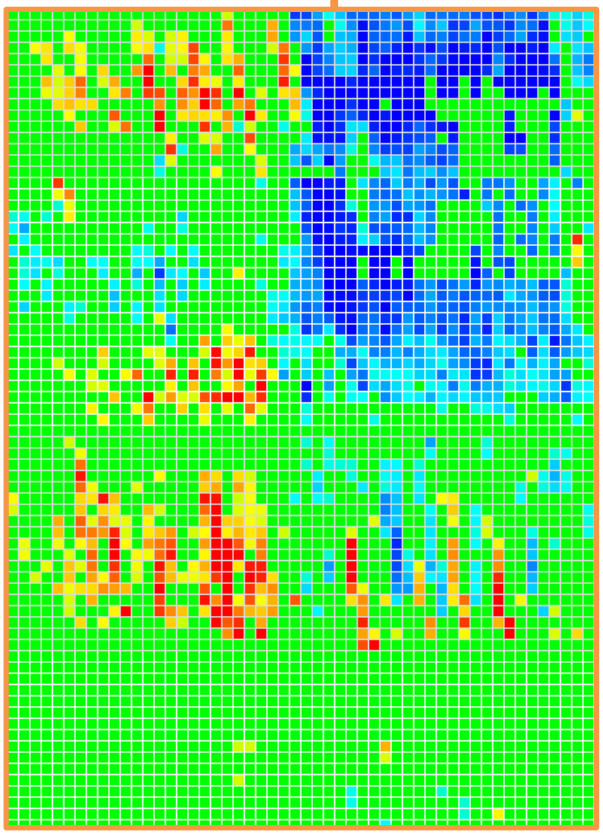
Interactive visualization



Simulation

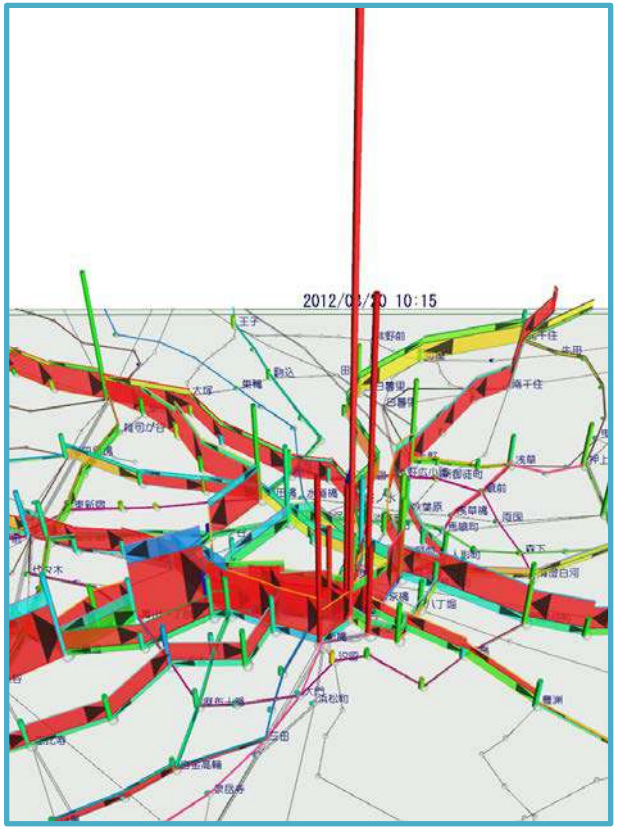
Visual Fusion of Physical Sensor and Social Media Data

HeatMap view



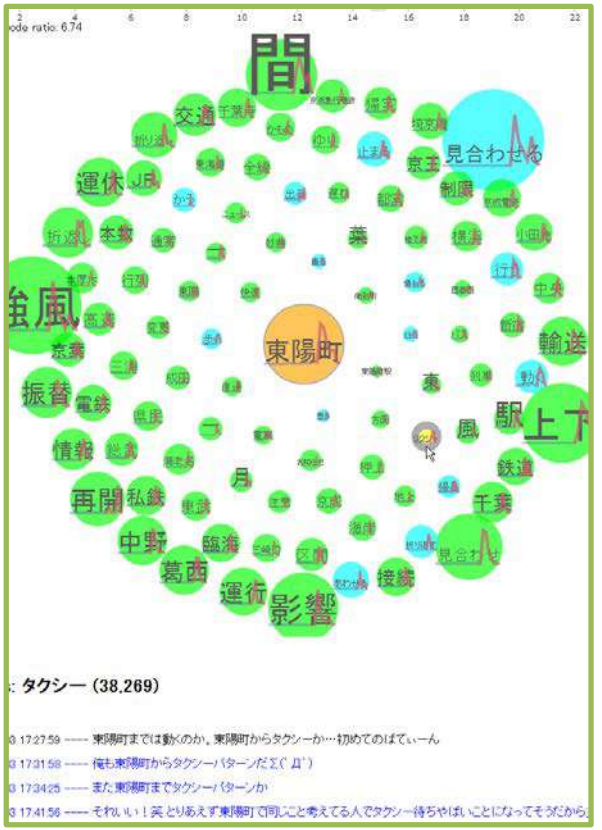
Discovering unusual phenomena from the wide range of temporal overviews

AnimatedRibbon view



Understanding changes in passenger flows and spatial propagation of unusual phenomena

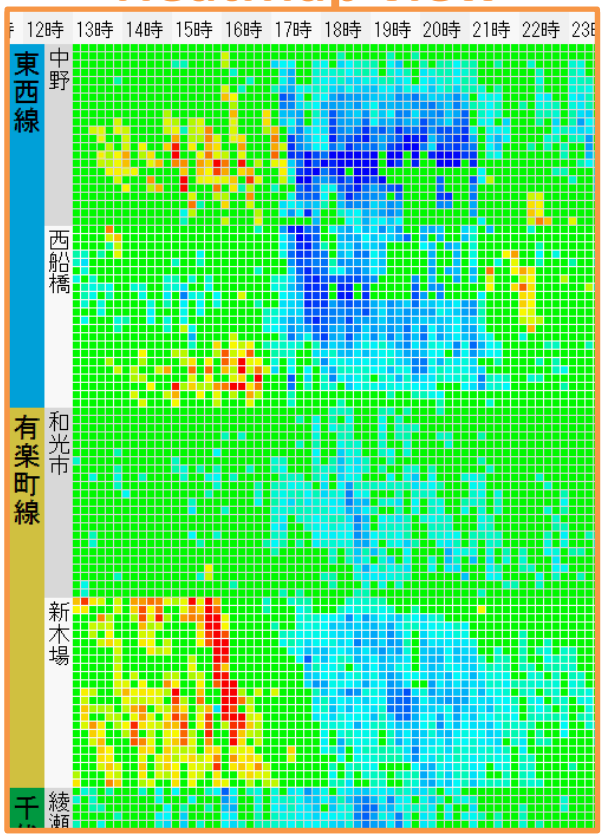
TweetBubble view



Exploring reasons for unusual phenomena or their effects from real users' voices

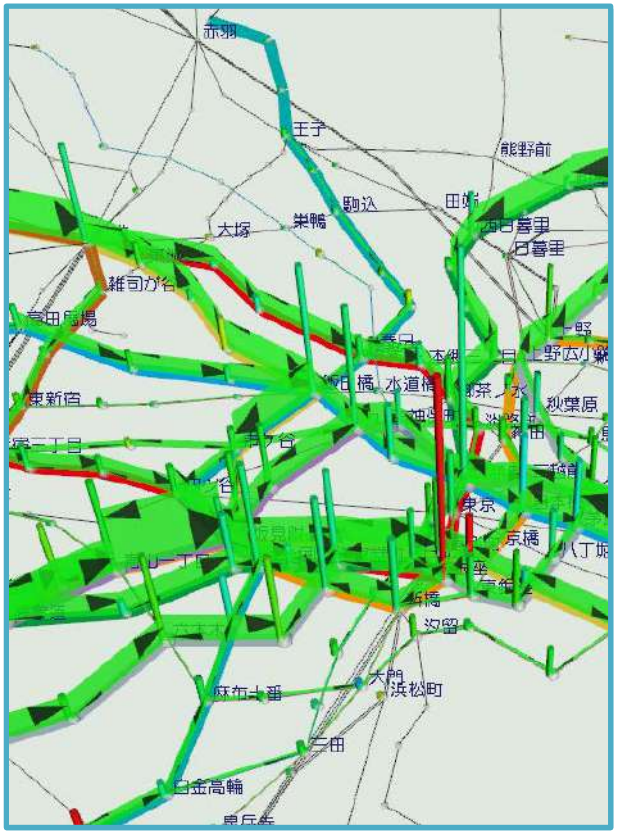
Visual Fusion of Physical Sensor and Social Media Data

HeatMap view



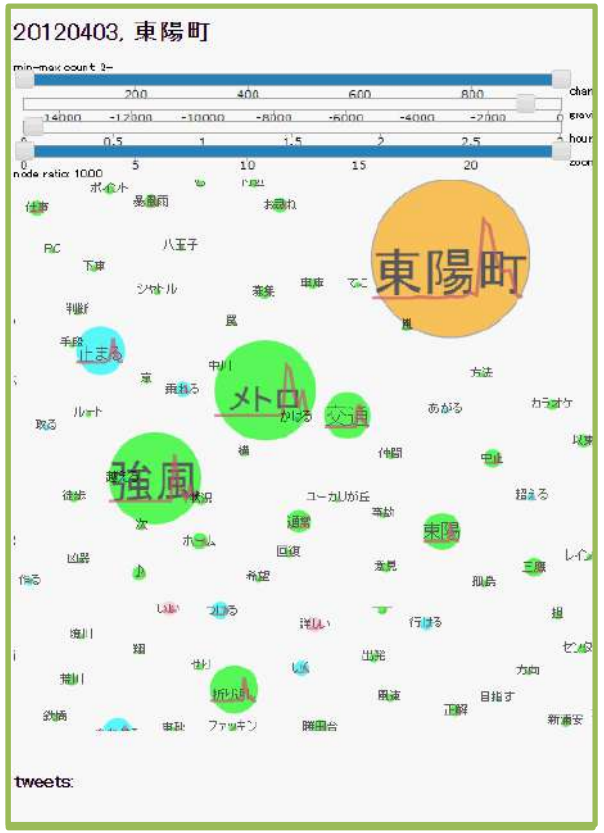
Discovering unusual phenomena from the wide range of temporal overviews

AnimatedRibbon view



Understanding changes in passenger flows and spatial propagation of unusual phenomena

TweetBubble view



Exploring reasons for unusual phenomena or their effects from real users' voices

Dataset

Smart card data

- Tokyo Metro trip records
- Covers almost all of the Tokyo business area
- **Three years' worth**
- 28 lines, 540 stations
- 350 million trips

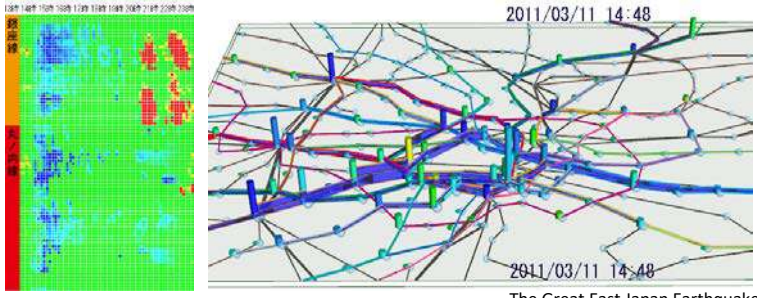
Twitter data

- More than three years' worth
- Over 2 million active users
- **Over 20 billion tweets**
  - Mainly Japanese tweets

Exploration Case Studies

Earthquake

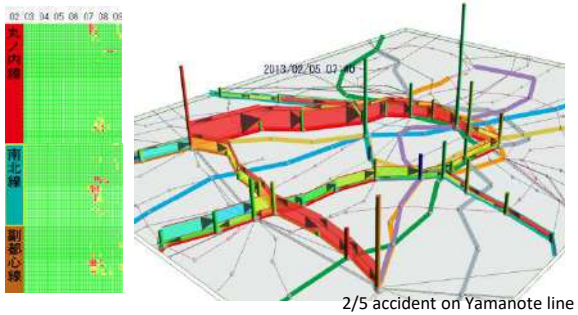
- **The Great East Japan Earthquake**



The Great East Japan Earthquake

Accidents

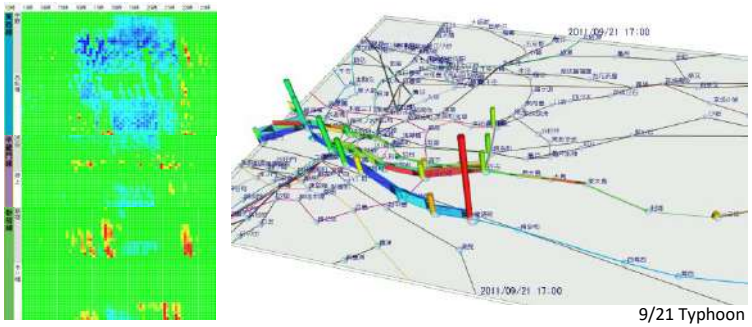
- Accident at Ueno Station on 5 Feb. 2013
- **Fire at Yurakucho on 3 Jan. 2014**



2/5 accident on Yamanote line

Natural disasters

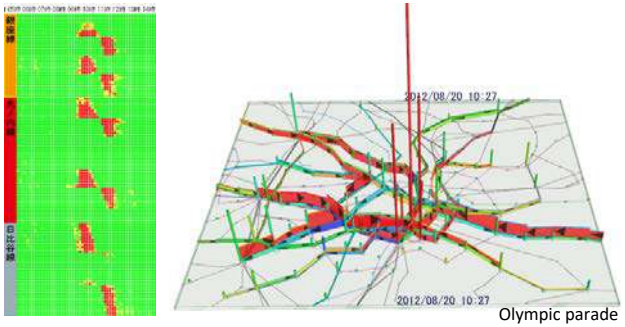
- Typhoon Roke 2011
- **Spring Storm April 2012**



9/21 Typhoon

Public gatherings

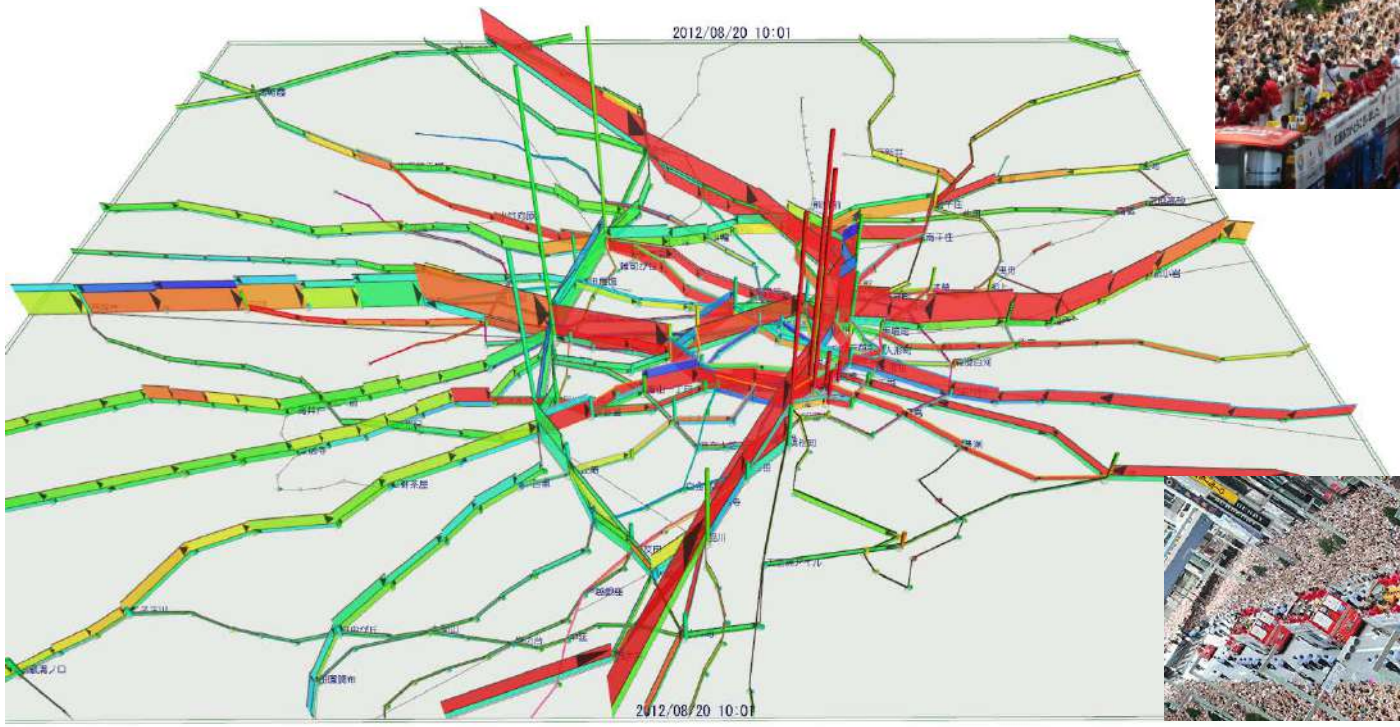
- **A Parade by London Olympic Medalists**
- Tokyo Marathon 2013



Olympic parade

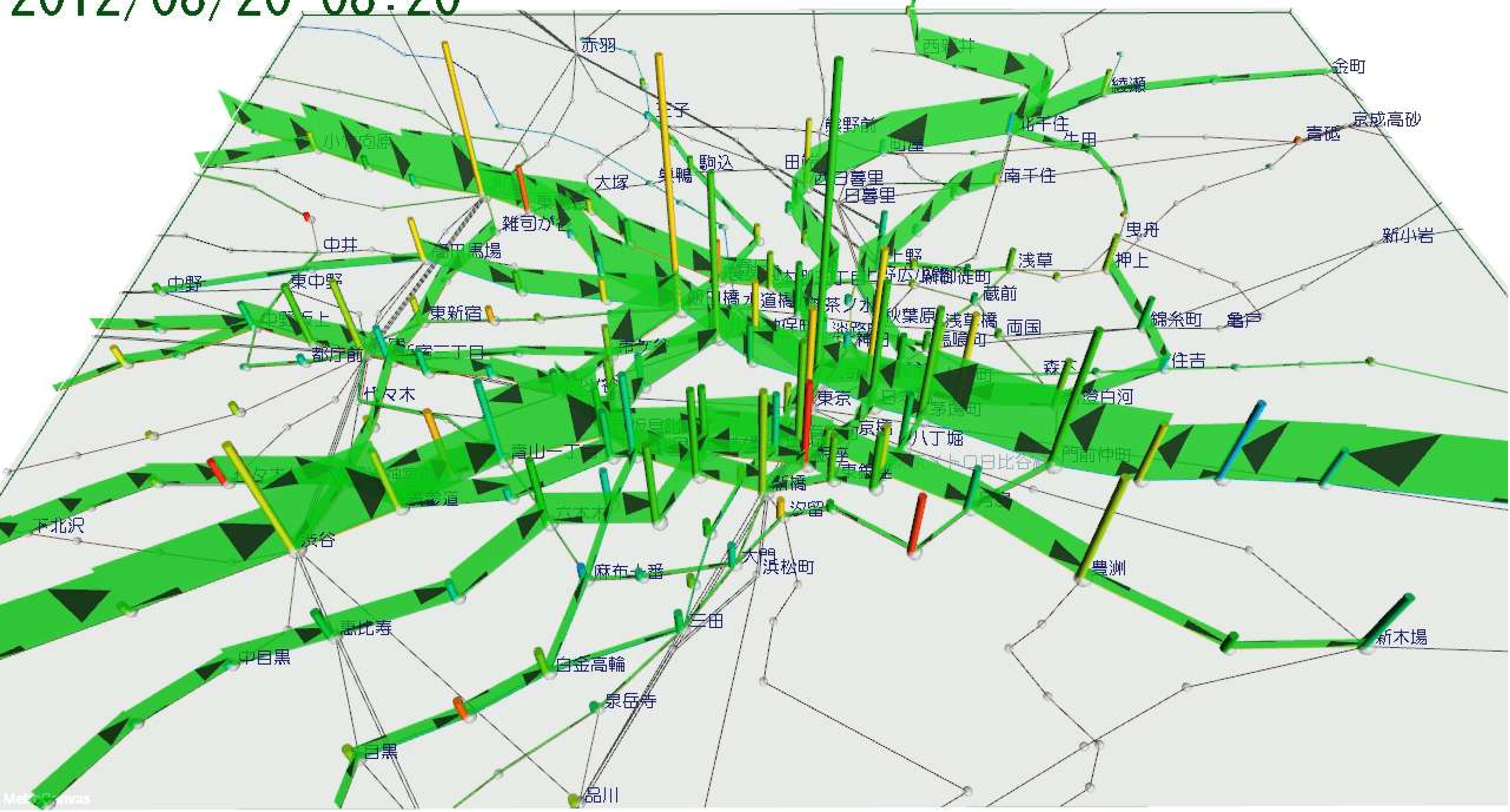
A Parade by London Olympic Medalists in Ginza on 20 August 2012

- The parade was performed for about 20 minutes from 11:00
- About 500,000 people gathered in Ginza
- Consisting of a total of 7 vehicles
- The Ginza Chuo Street (about 1km)



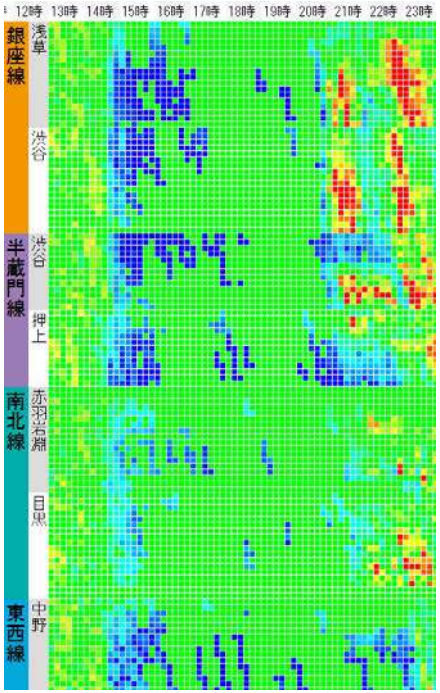
http://www.joc.or.jp/english/londonolympics/parade.html

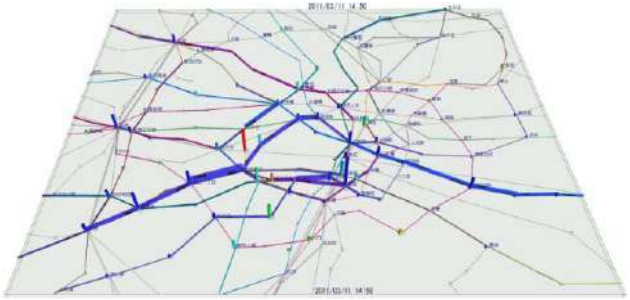
2012/08/20 08:20



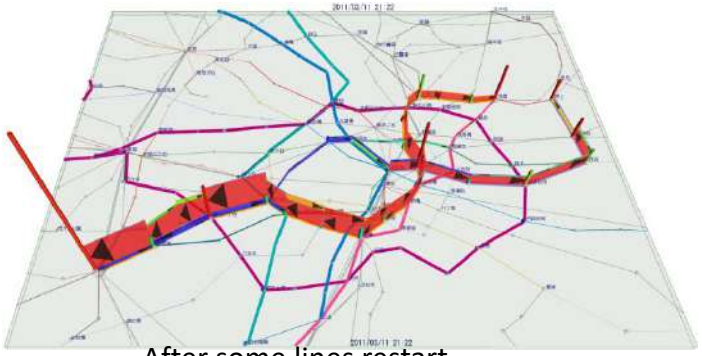
The Great East Japan Earthquake on 11 Mar. 2011

- The earthquake struck off the northeastern coast of Japan at 14:46
  - Many public transportation systems suspended operation
  - Most people could not travel until midnight or the next morning



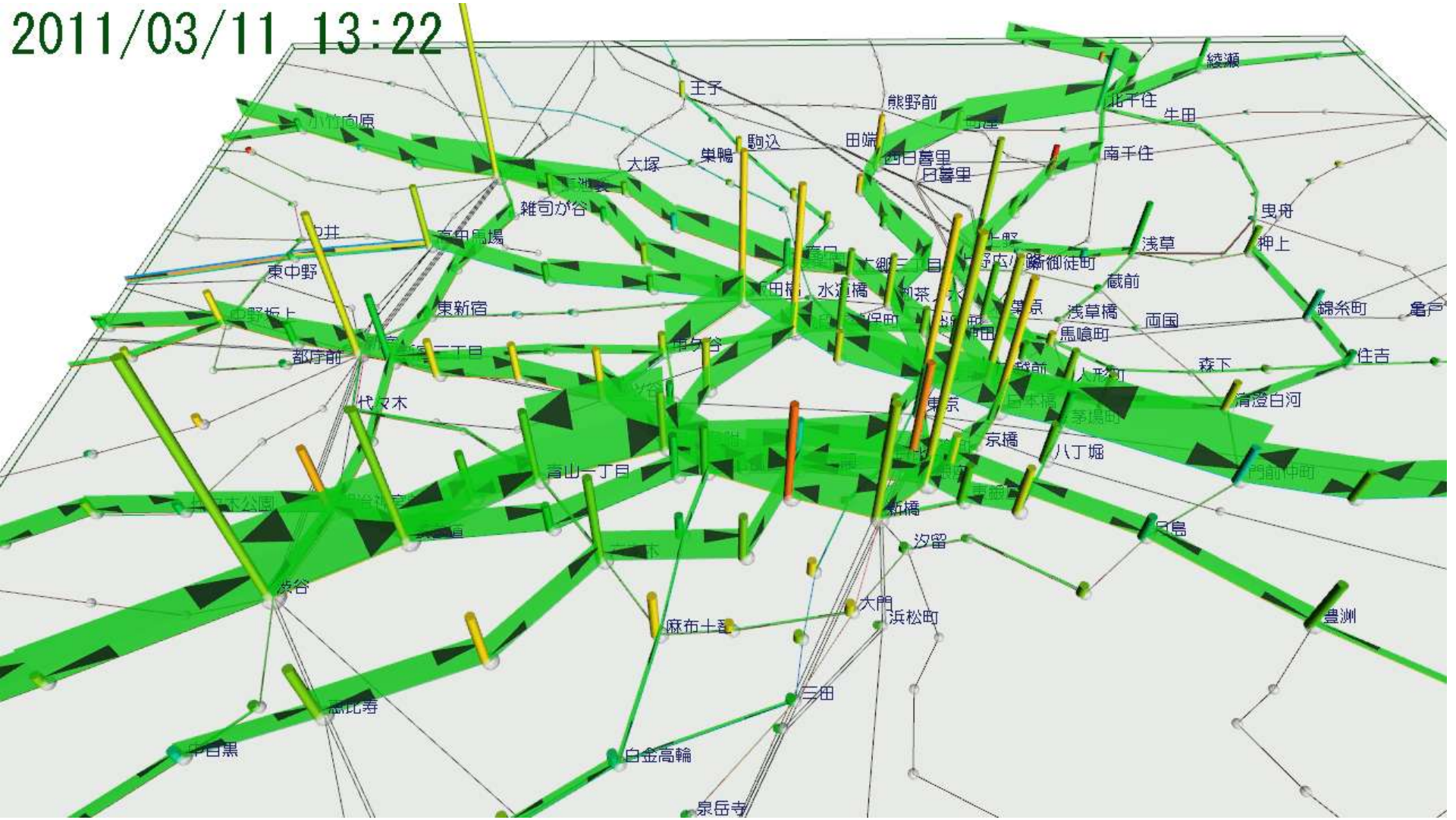


Just after the earthquake



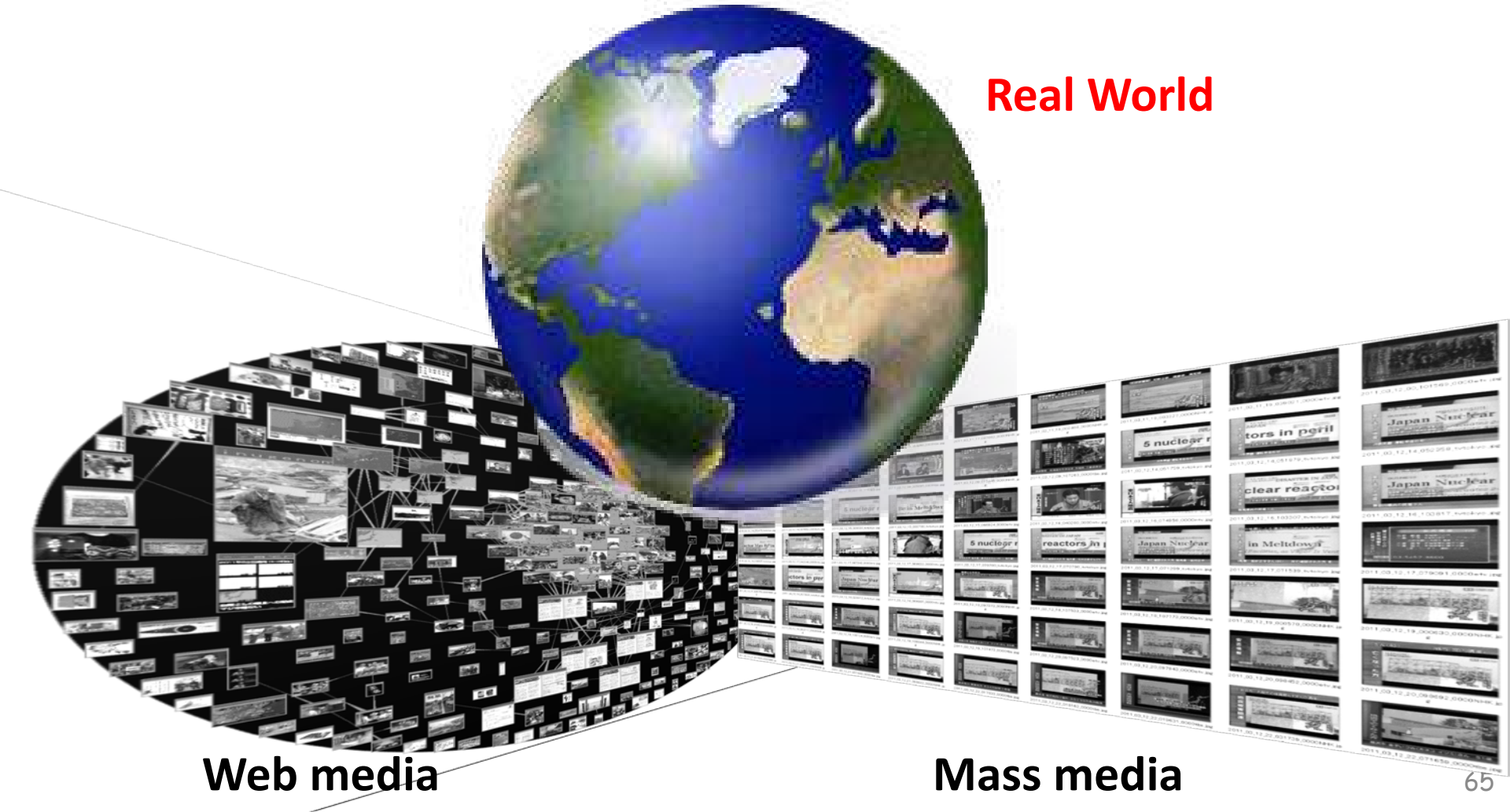
After some lines restart

http://www.asahi.com/special/10005/TKY201103110519.html


2011/03/11 13:22
王子
熊野前
綾瀬
北千住
牛田
駒込
巣鴨
田端
西日暮里
南千住
日暮里
太塚
雑司が谷
曳舟
浅草
押上
東中野
蔵前
東新宿
浅草橋
錦糸町
両国
都庁前
馬喰町
住吉
森下
代々木
清澄白河
京橋
八丁堀
青山一丁目
新橋
汐留
日島
渋谷
大門
浜松町
豊洲
麻布十番
三田
恵比寿
中目黒
白金高輪
泉岳寺

Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data



Road and bridge condition screening based on large scale sensing

Tomonori Nagayama
Associate Professor, the University of Tokyo

This work was supported by Council for Science, Technology and Innovation, “Cross-ministerial Strategic Innovation Promotion Program (SIP), Infrastructure Maintenance, Renovation, and Management”. (funding agency: JST)

Background: Road condition evaluation

Road network in the world

1,210,000 km road, 8000km highways in Japan (6th in the world)

- Road profiler:
  Expensive and used infrequently
- The majority is not inspected. Even without visual inspection. Difficult to have strategic maintenance
- In many countries, there are arterial roads without inspections.

Can we evaluate road conditions effectively & efficiently?
No apparent potholes or cracks, however …

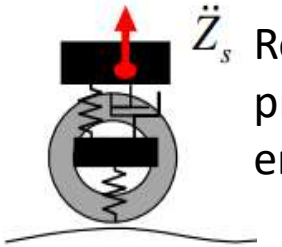
Smartphone-based road condition evaluation

Road evaluation using vehicle responses : DRIMS

GPS
Acceleration
Angular velocity
+
Movie
Photos

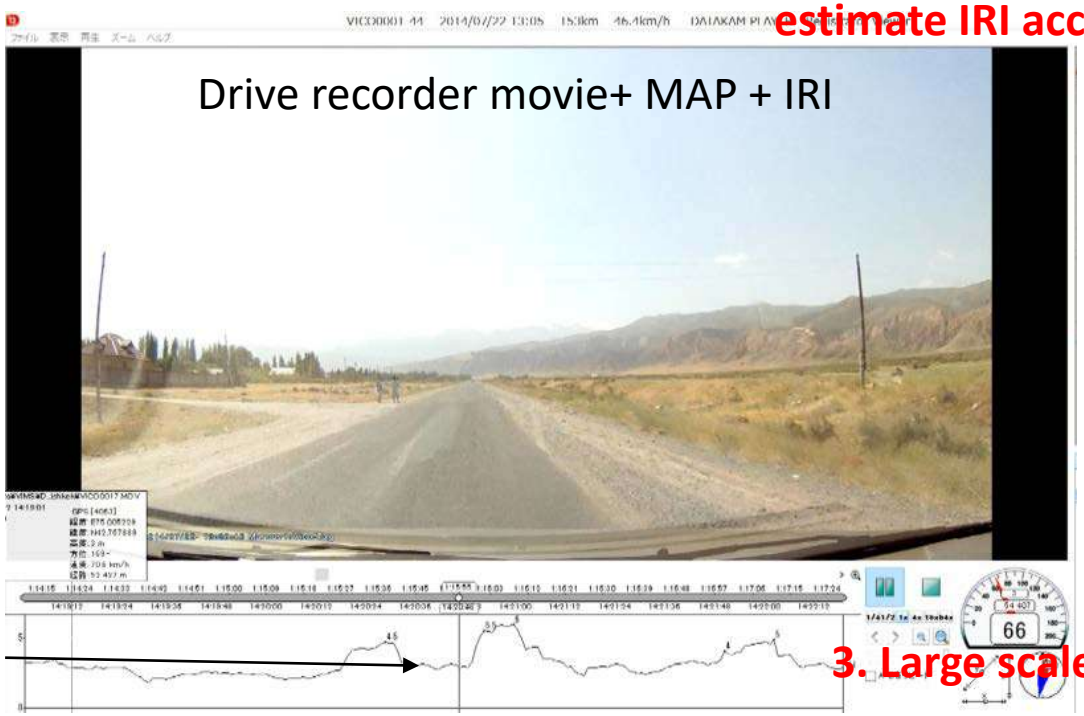
1. Are smartphone sensors accurate enough?

International Roughness Index (IRI)



Road condition evaluation index proposed by World bank and employed worldwide.

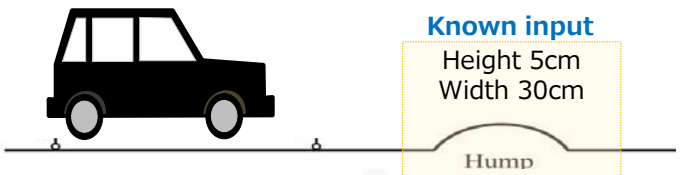
2. Can compensate the difference in vehicle dynamics/drive speed and estimate IRI accurately?



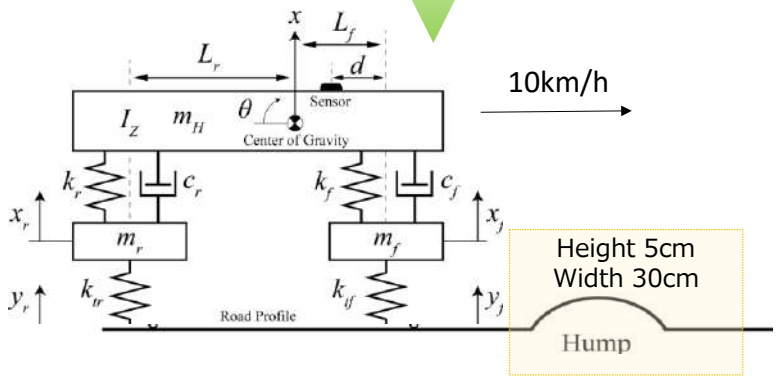
3. Large scale implementation

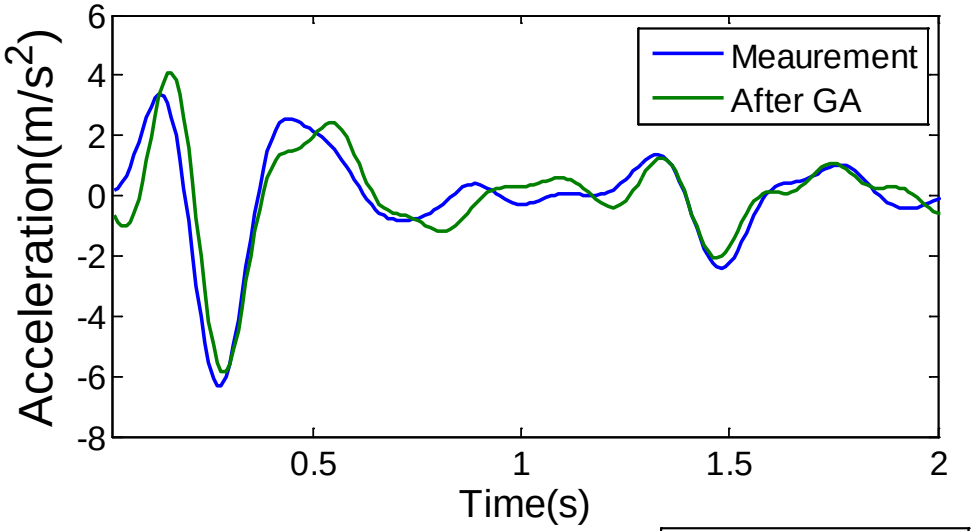
Vehicle modeling with 4 DOF

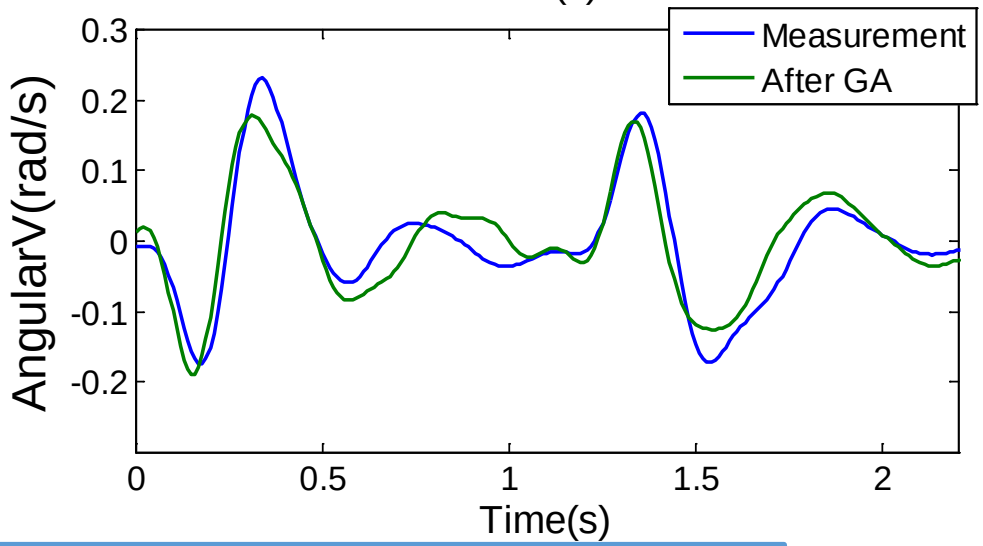
Hump calibration



Simulated by Half car model







The vehicle is modeled as a half-car model using hump passage responses

Experimental validation of IRI estimation

Three types of vehicles

Light type

Small size

Middle size(Van)

Reference
(Profiler)

Test course: 13.6 km in Chiba-city

Sensor locations

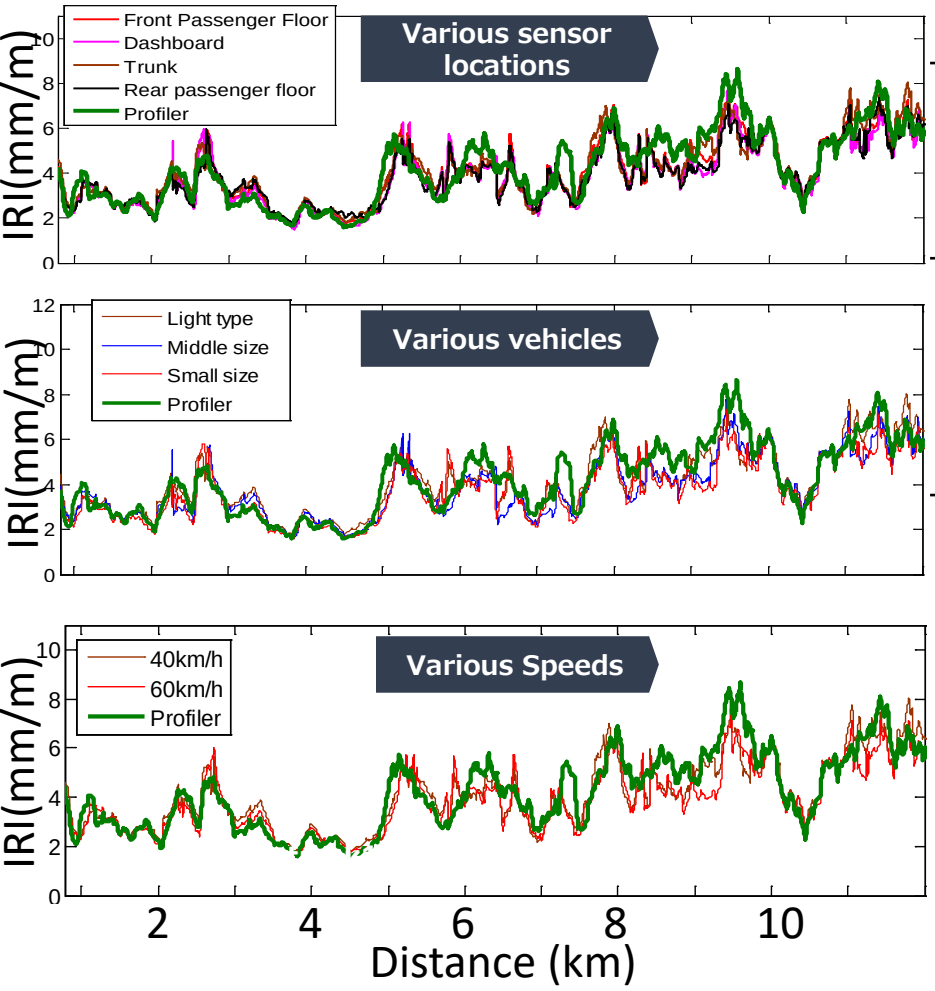
Dashboard

Passenger floor

Trunk

Rear seat floor

Experimental validation of IRI estimation



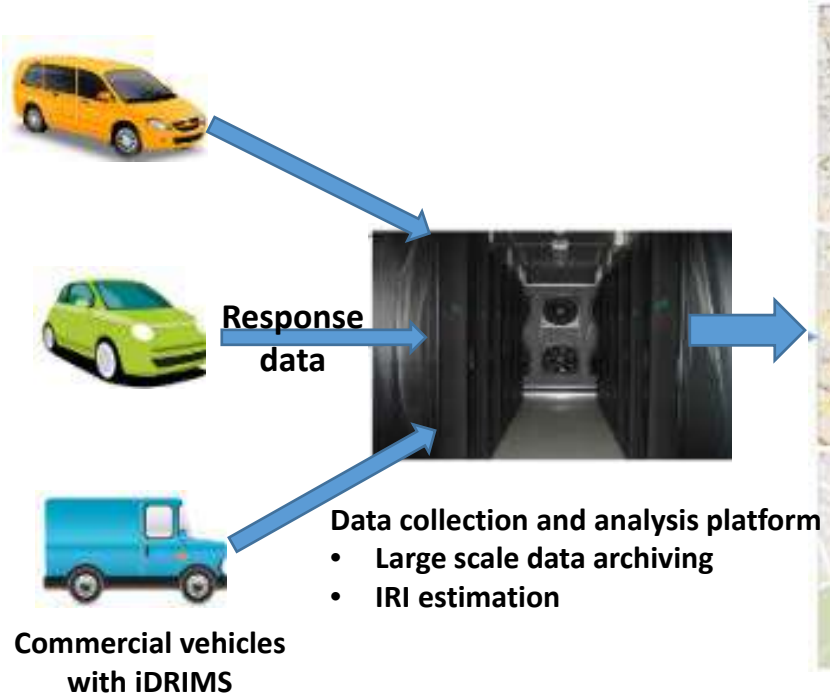
Installation location	IRI error: $\sqrt{\frac{\sum_i (IRI_{est,i} - IRI_{true,i})^2}{\sum_i (IRI_{true,i})^2}}$ [%]		
	Middle	**Small**	**Light**
(a) Dashboard	**21.2**	**16.0**	**12.2**
(b) Front passenger seat floor	**23.2**	**22.7**	**12.5**
(c) Rear passenger seat floor	**16.1**	**19.8**	**18.8**
(d) Trunk	**13.7**	**15.6**	**17.1**

The difference between iDRIMS and reference profiler is about 10-20% regardless of vehicle type, speed, and sensor locations.

Large-scale implementation

DRIMS has been implemented on commercial vehicles.

- About 100 commercial vehicles, about 10 organizations
- More than 1 year data
- Data collection/analysis visualization platform to handling a large amount of data
- Based on GPS data, the IRI values are mapped to road map.
- Data is centrally managed. Vehicles actively used, inactive vehicles, vehicles with calibration problems can be centrally examined.



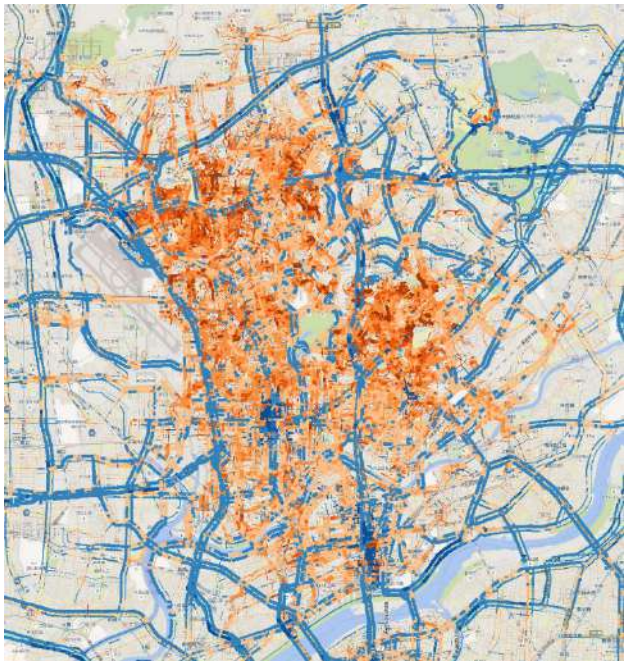
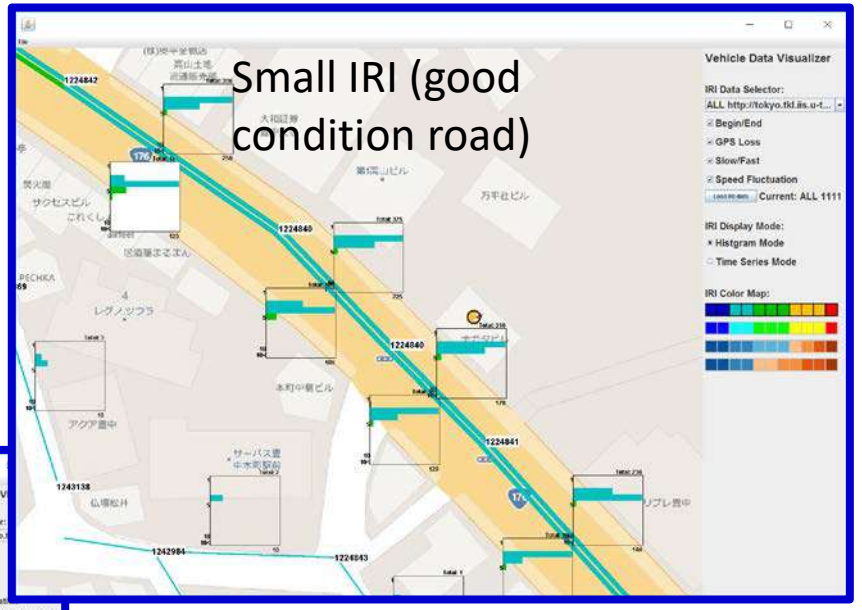


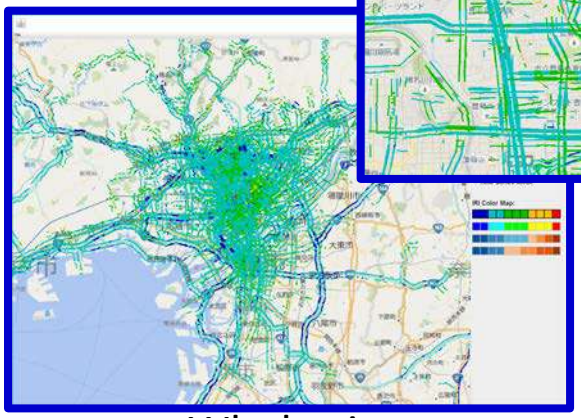
Fig. Display road condition with a heat map

(Provided by: Masashi Toyoda, Institute of Industrial Science, the University of Tokyo)

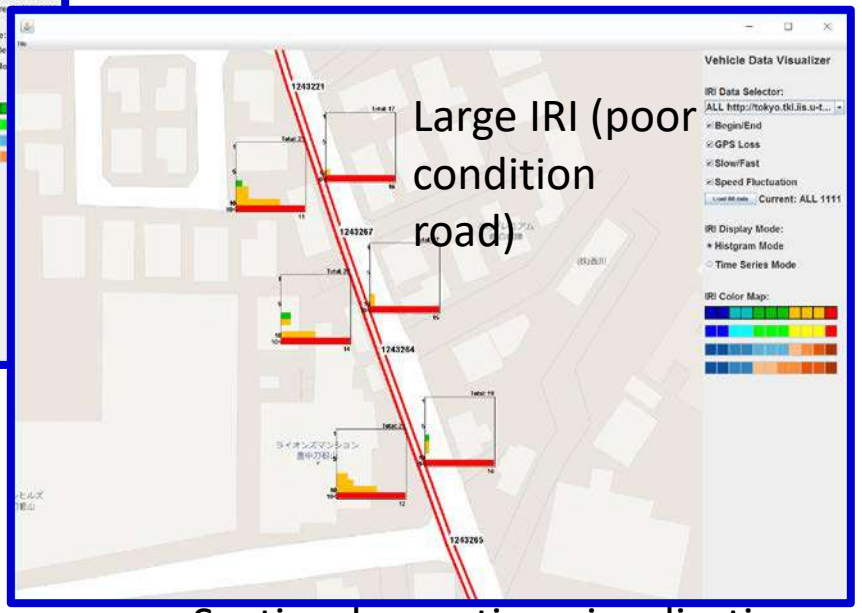
Large-scale interactive visualization



Zoom-in

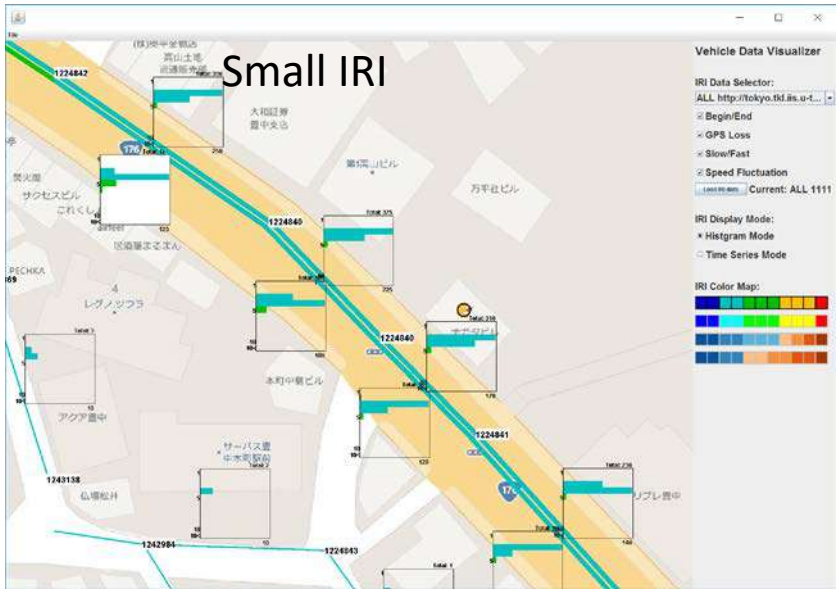


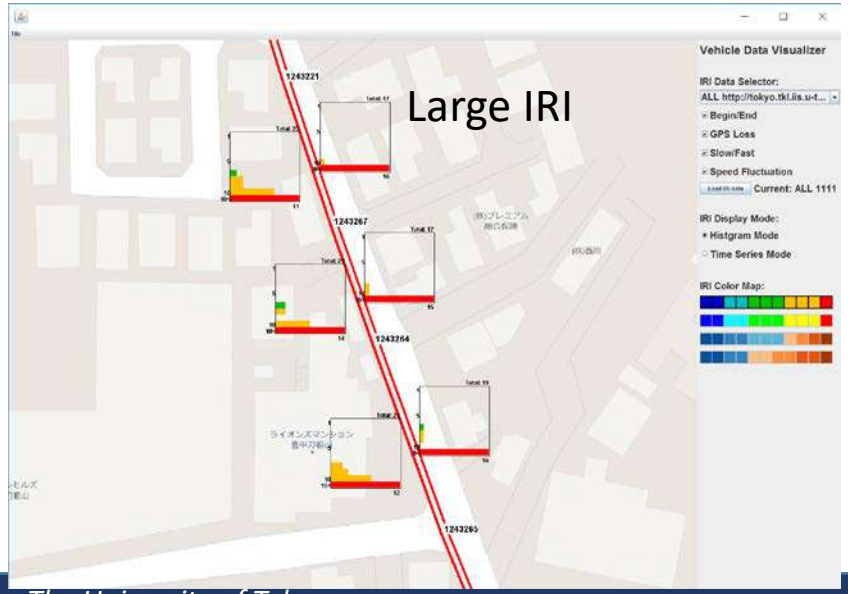
Whole city



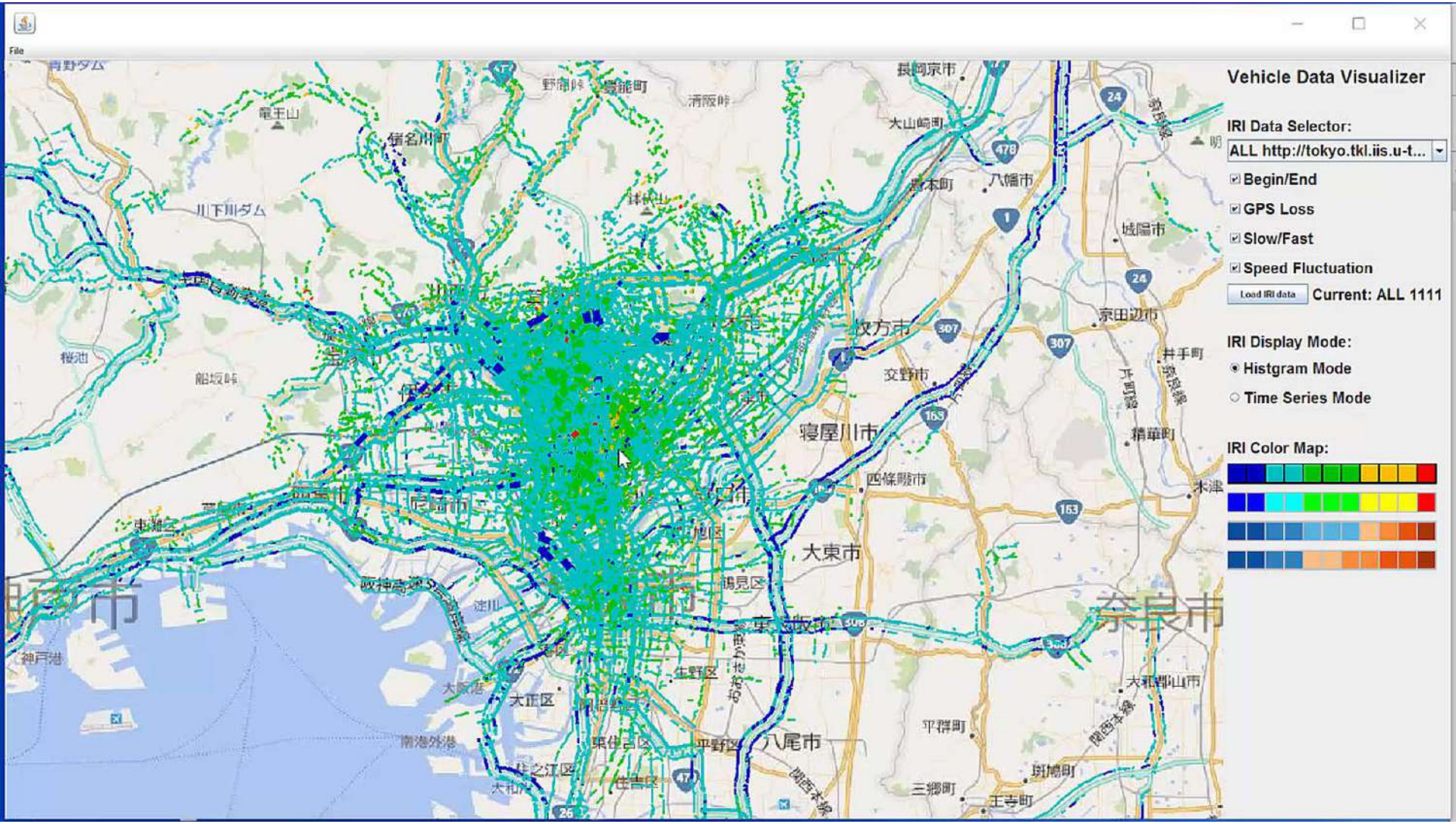
Section-by-section visualization

Large-scale interactive visualization

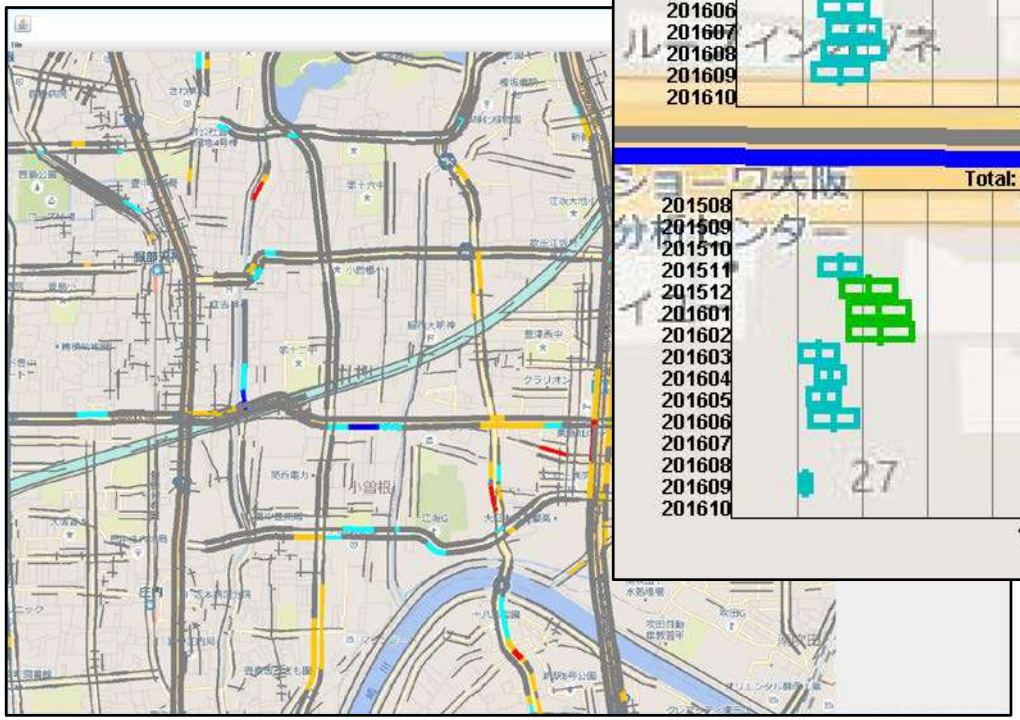




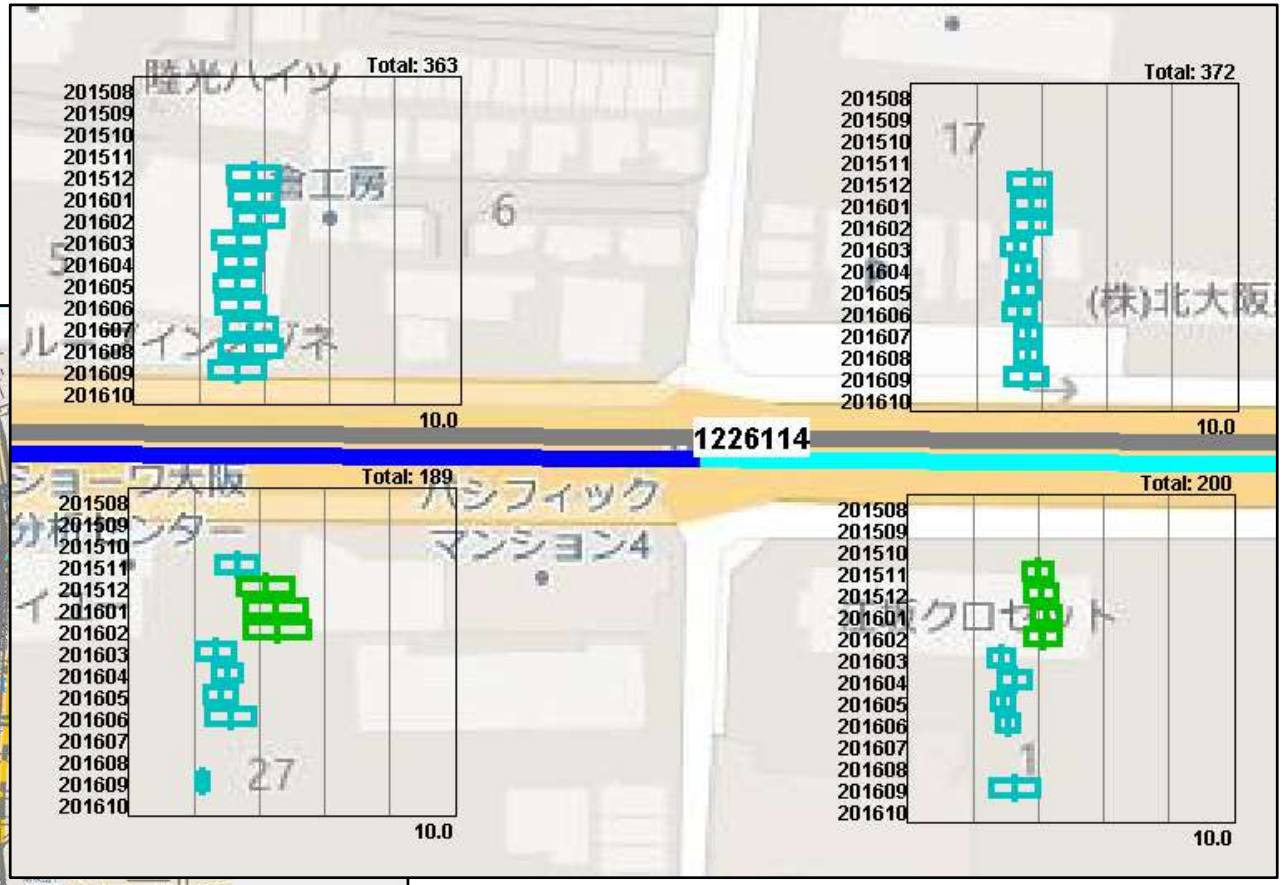
Large-scale interactive visualization



Large-scale interactive visualization (chronological change)

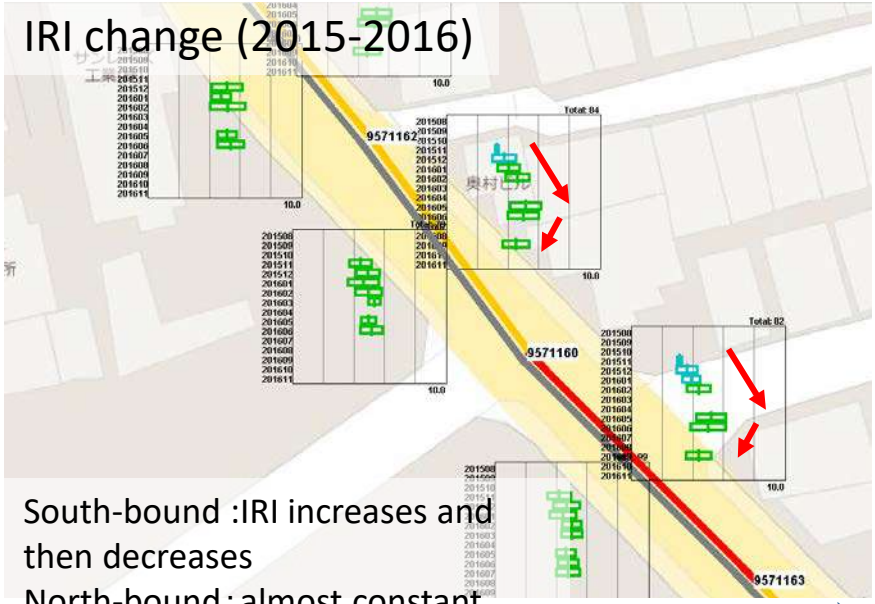


Large chronological changes are high-lighted.



IRI improvement due to repair work.

Large-scale interactive visualization (chronological change)



Difference captured due to the high accuracy/frequent measurement.

Multiple Media and Real World Big Data Analysis

Understanding and resolving social problems by analyzing Web/social media, mass media, and real world sensor data