1st Int. Alexandria Workshop (15th. Sep. 2014)

Multiple Media Analysis and Visualization with Large-Scale Temporal Web Archives

Masashi Toyoda

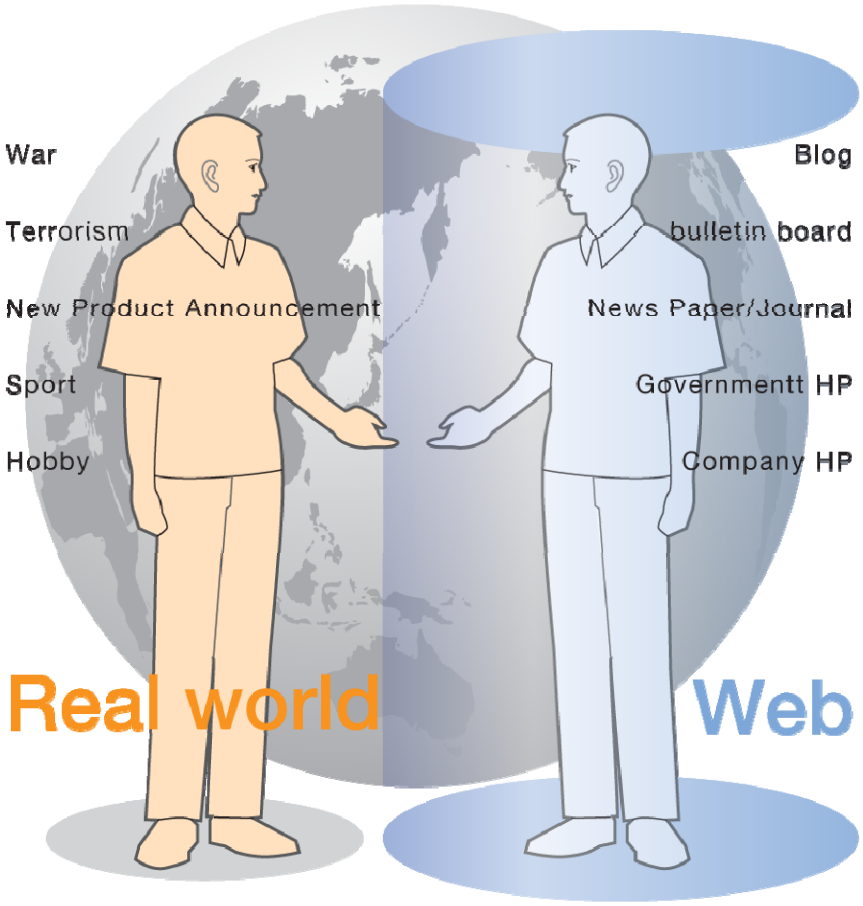
Center for Socio-Global Informatics,

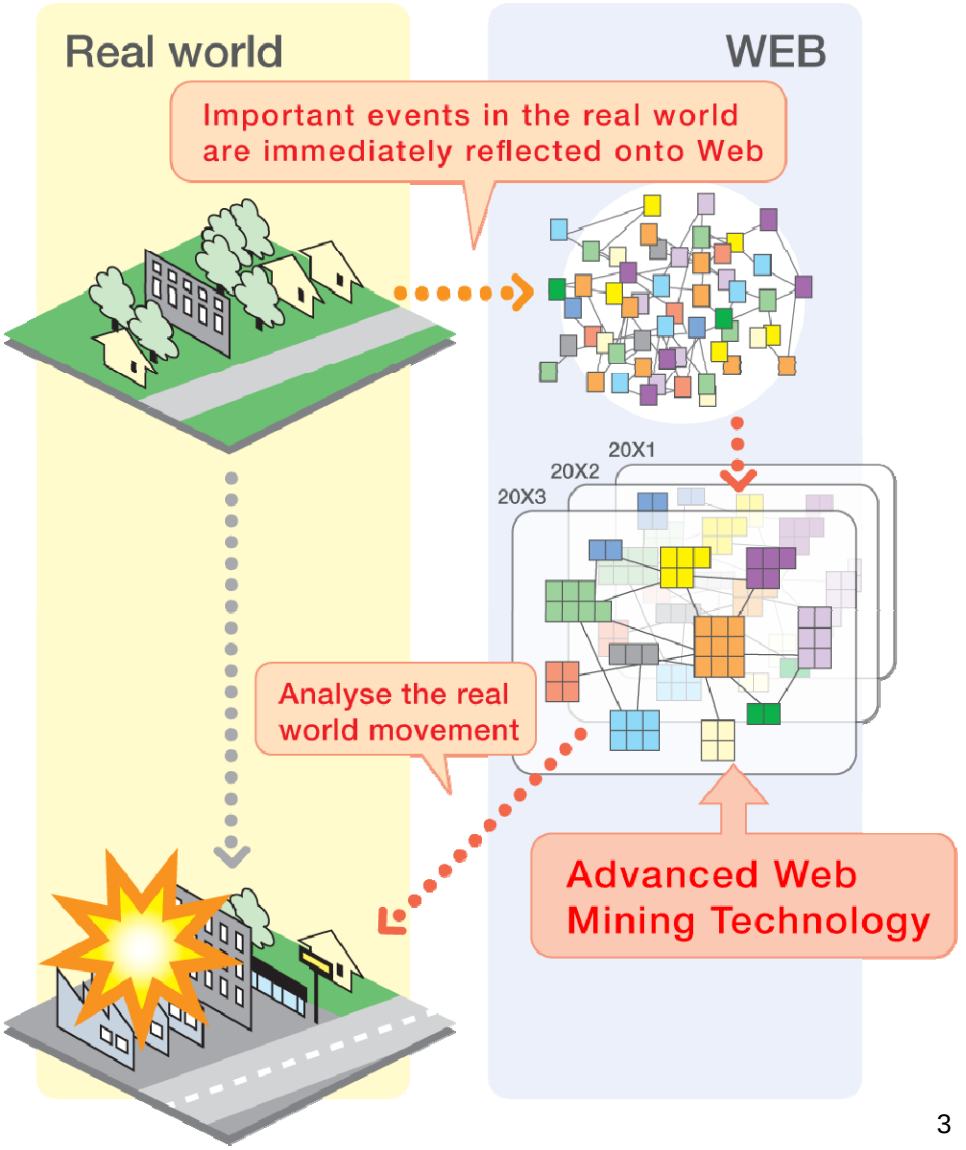
Institute of Industrial Science,
The University of Tokyo

Collaborative researchers

- Masahiko Itoh (The University of Tokyo)
- Naoki Yoshinaga (The University of Tokyo)
- Cai-Zhi Zhu (National Institute of Informatics)
- Shin’ichi Satoh (National Institute of Informatics)
- Masaru Kitsuregawa (National Institute of Informatics / The University of Tokyo)

Web as a Sensor for Society

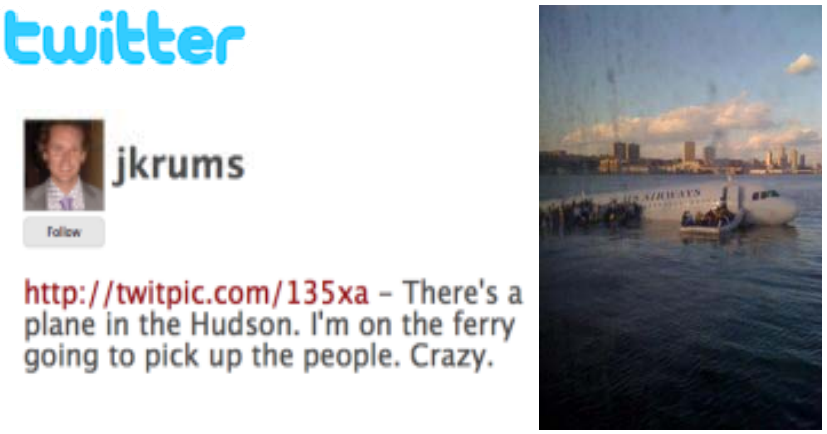




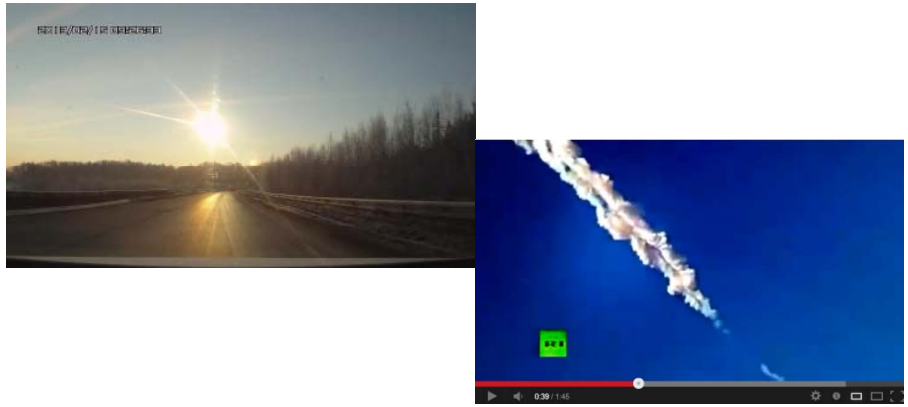
Interaction between Media

Mass and web media affect each other

- **Mass media** still have big influence
  - Major topics in web 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

Multiple Media Fusion

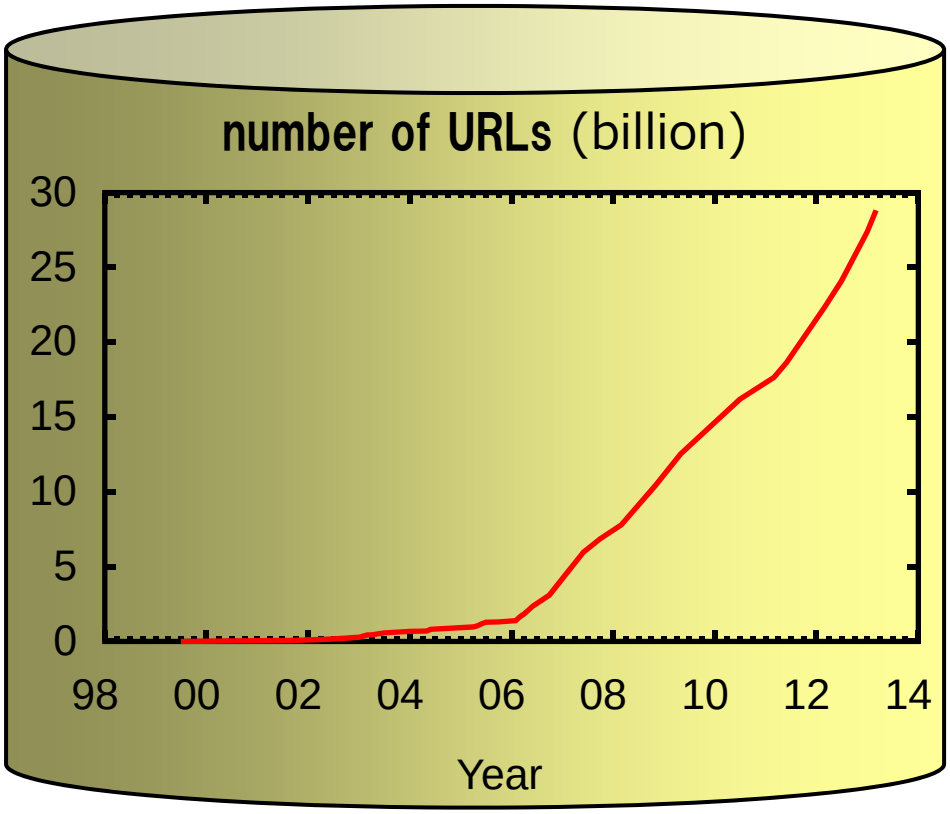
Analyzing multiple media is strongly required for understanding social activities

Web media

Mass media

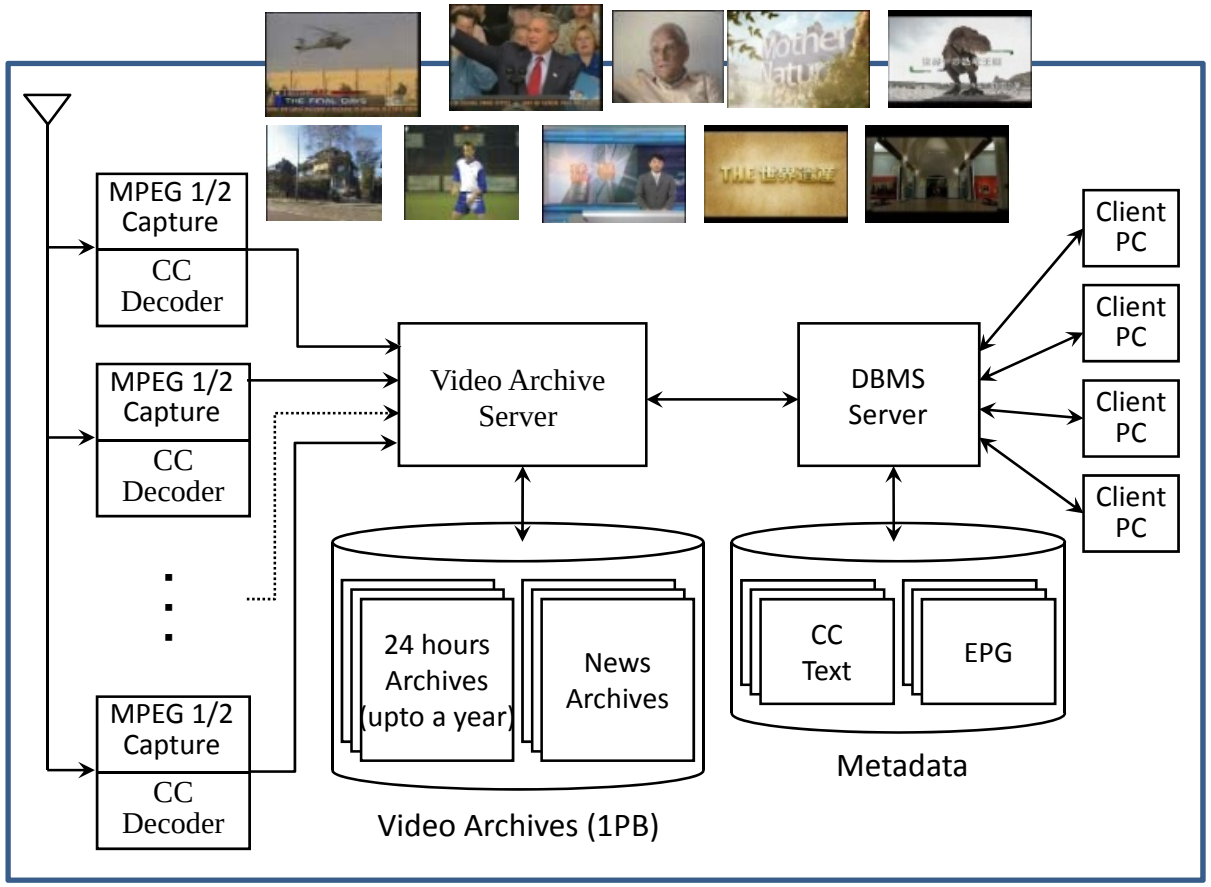
Web Archives in UTokyo

- **Japanese Web Archive**
  - Focused crawl of Japanese pages in every domains (1999 -)
  - Contents of 30 billion URLs
  - One of the largest archive in Asian region
- **Blogs**
  - From 2006
  - 2.5 million feeds
  - About 1 billion articles
- **Twitter**
  - From 2011
  - 1.5 million users
  - About 20 billion tweets



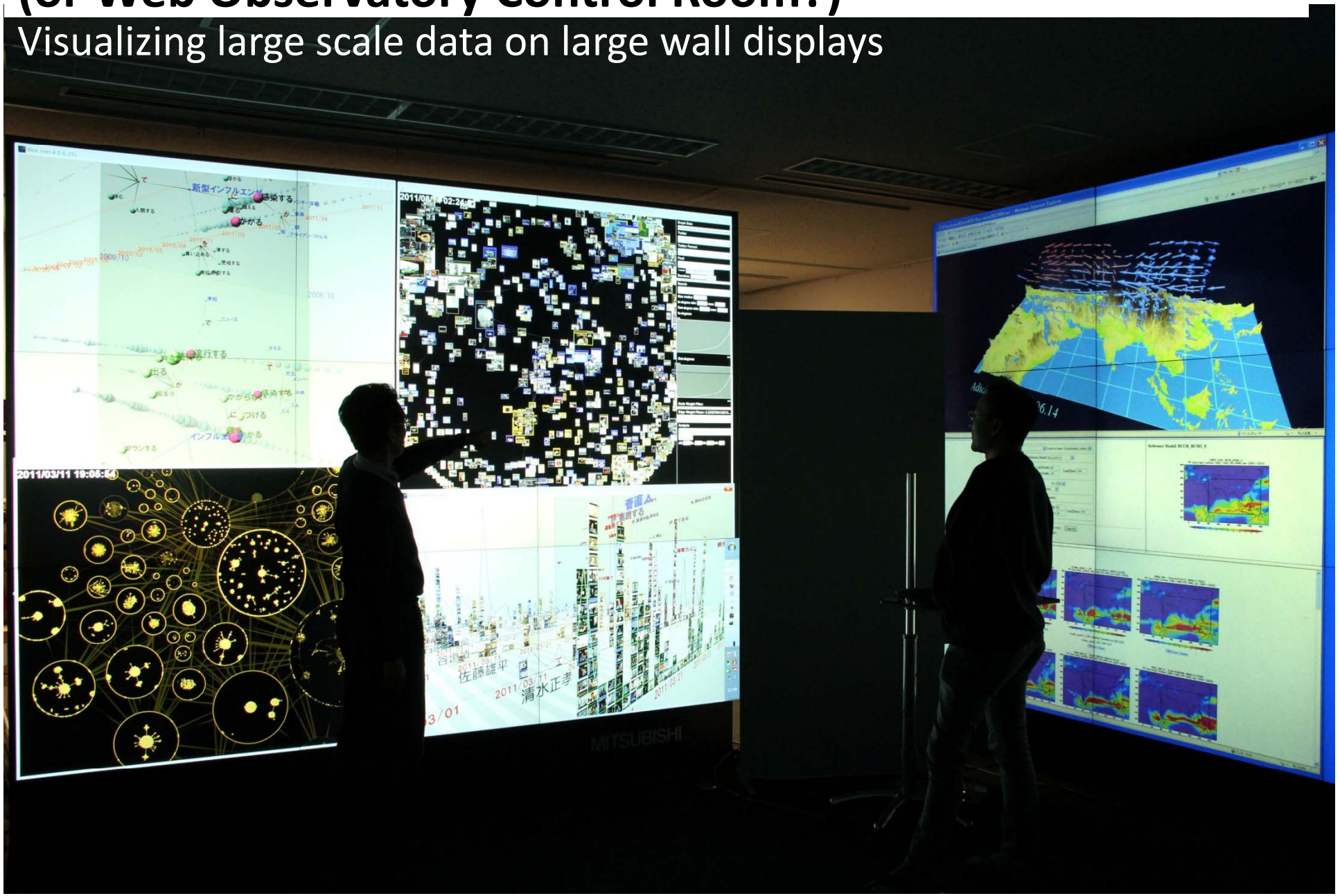
Broadcast News Video Archive in NII

- 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).



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

Visualizing large scale data on large wall displays



Platform for Multiple Media Analysis

Multiple Media Analysis

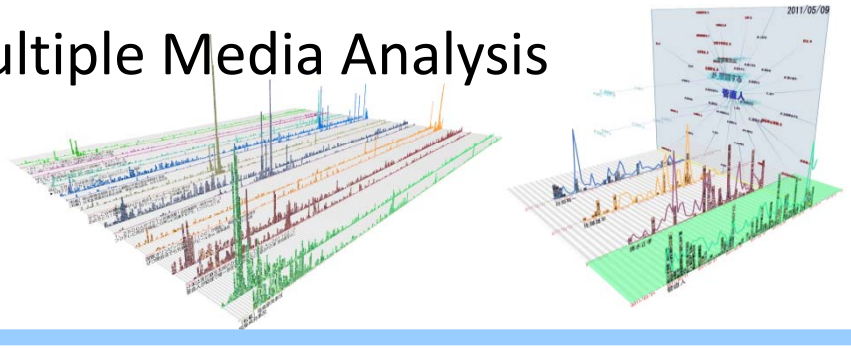
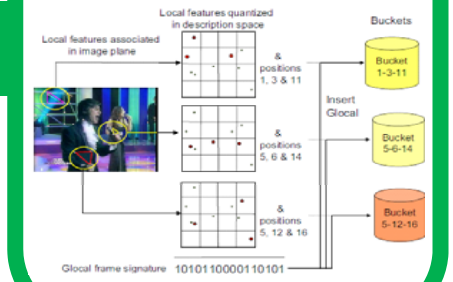
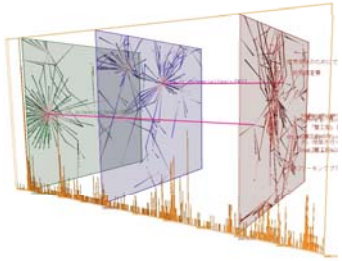


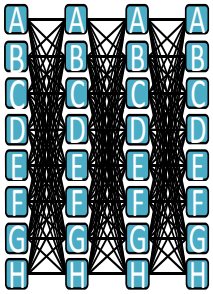
Image Linkage



Graph Analysis

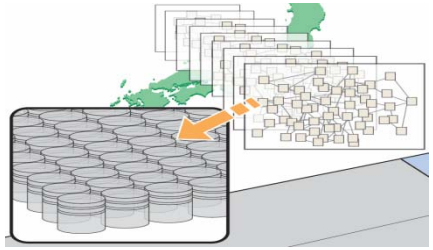


Text Analysis



Video Archive

Web media archive by continuous crawler



NII

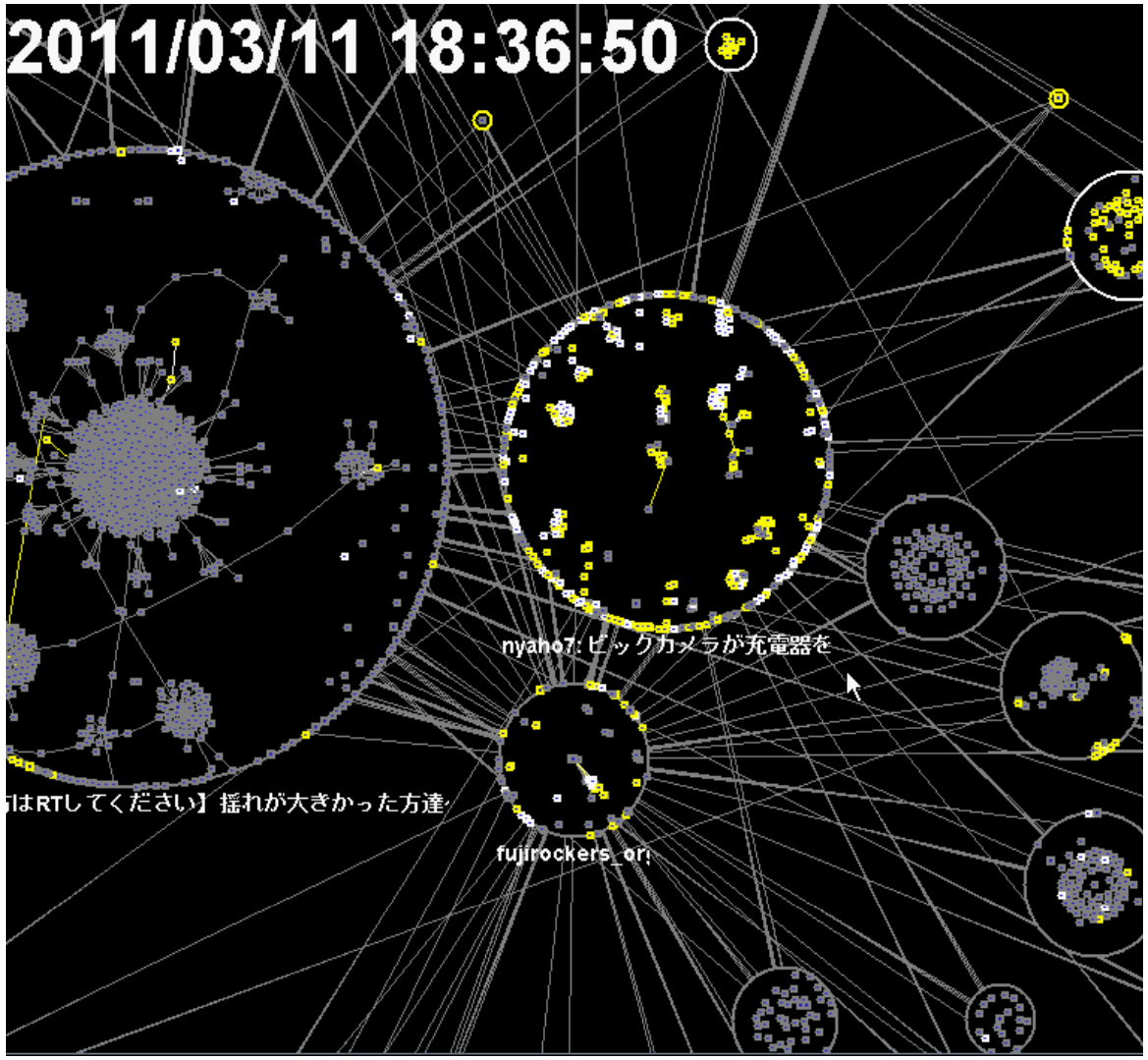
Information Diffusion on 3.11 in Twitter

- 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

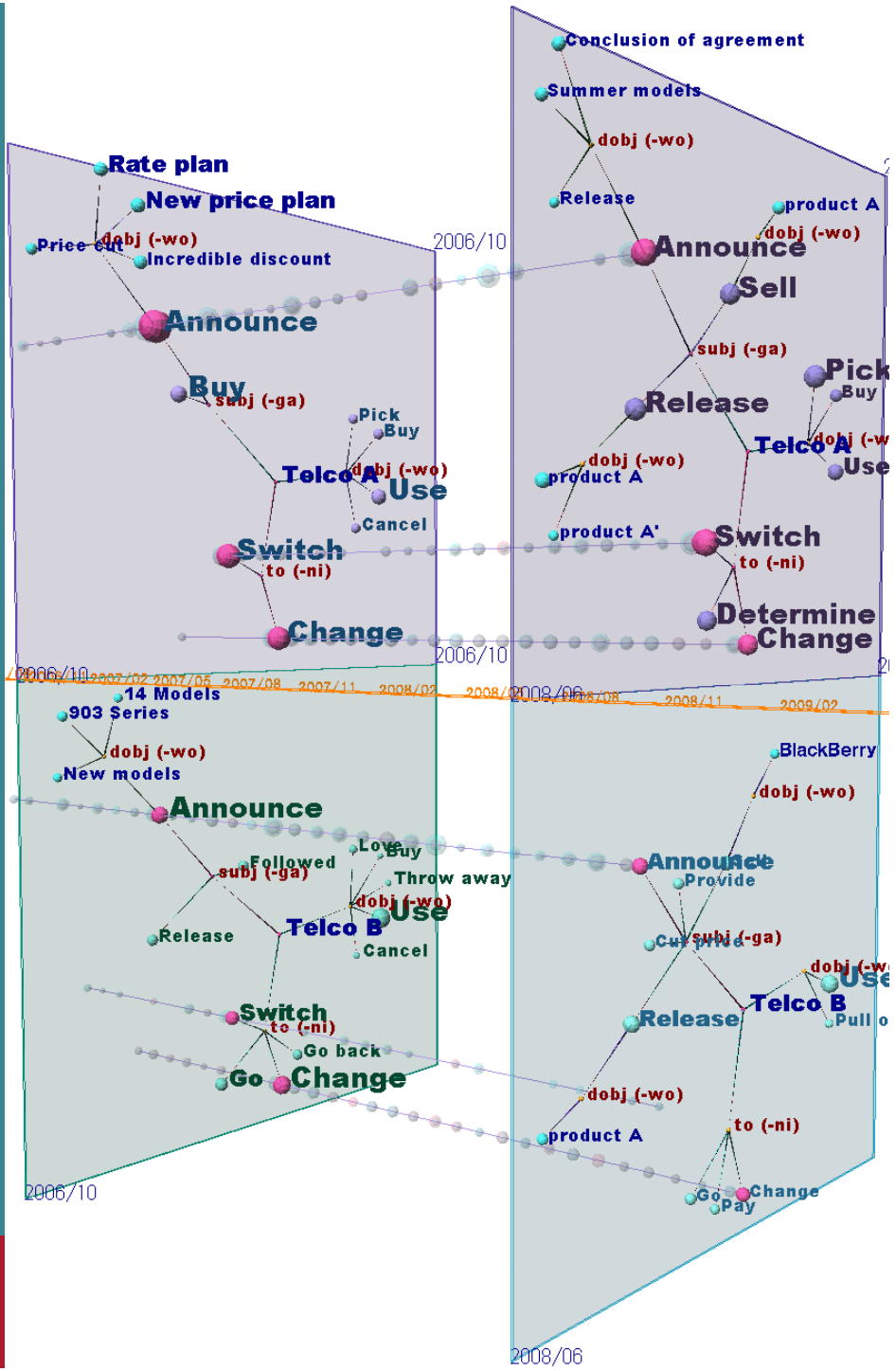


Analysis and Visualization of Temporal Changes in Bloggers' Activities and Interests

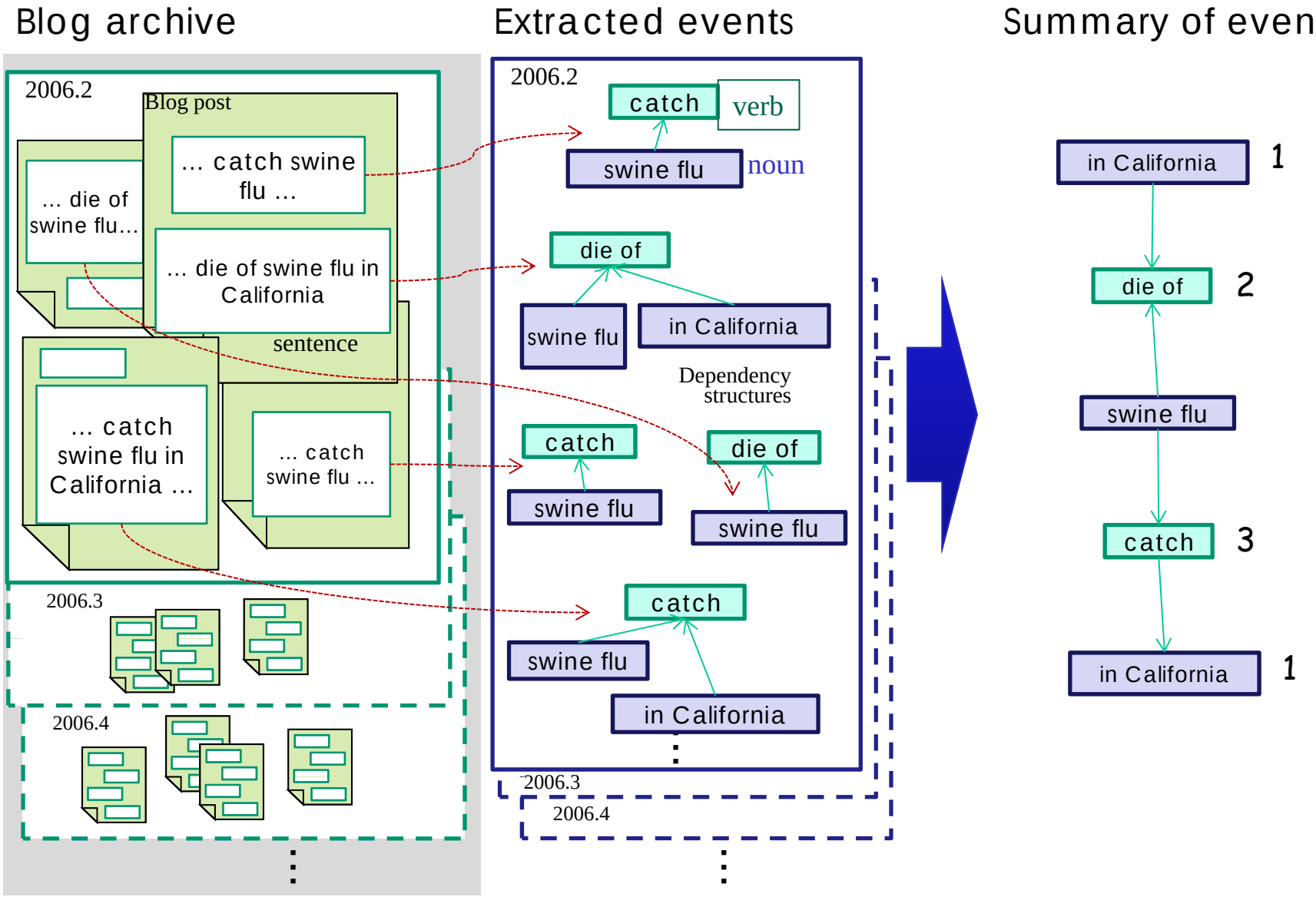
Masahiko ITOH, Naoki Yoshinaga,
Masashi TOYODA, and Masaru KITSUREGAWA

Institute of Industrial Science (IIS),
University of Tokyo, Japan

IEEE PacificVis 2012
Feb. 28 - Mar. 2, 2012, Songdo, Korea

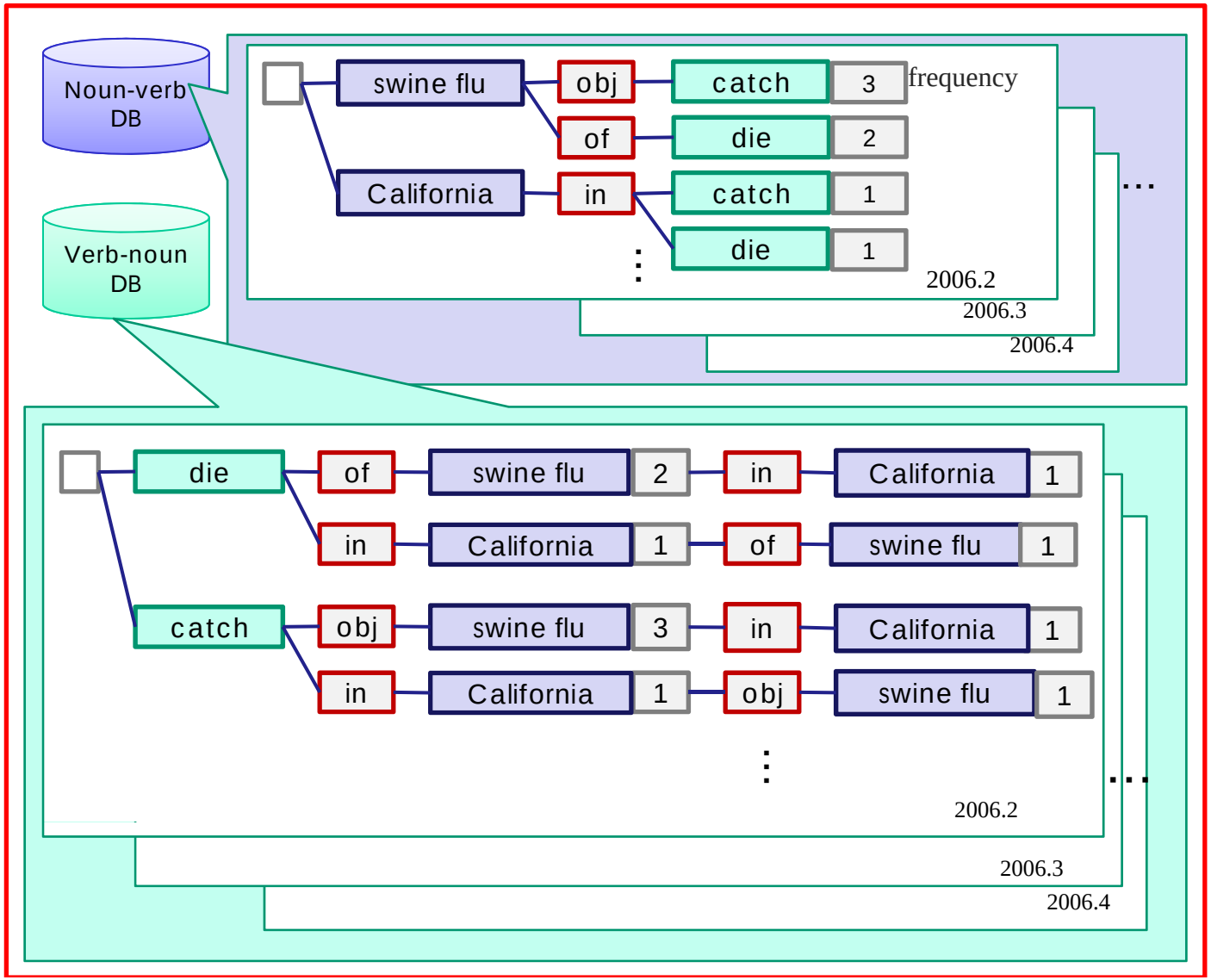


Visual Summarization of Events by Term Dependency Analysis



Term Dependency Database

Dependency Database



3D Visualization

- Visualizing dependency relations as a unified tree representation
  - TimeSlice [Itoh 2010]: slidable 2D plane on the timeline to visualize temporal changes in structure
  - Sliding operation
    - changes in the structure and frequencies
- Interactively navigating to the detailed dependency
- Providing multiple 2D planes to compare
  - Different timings, different topics
  - Tiled view
- Visualizing changes in the frequencies of each dependency relation
  - TimeFlux : line of spheres to visualize changes in the amount of information

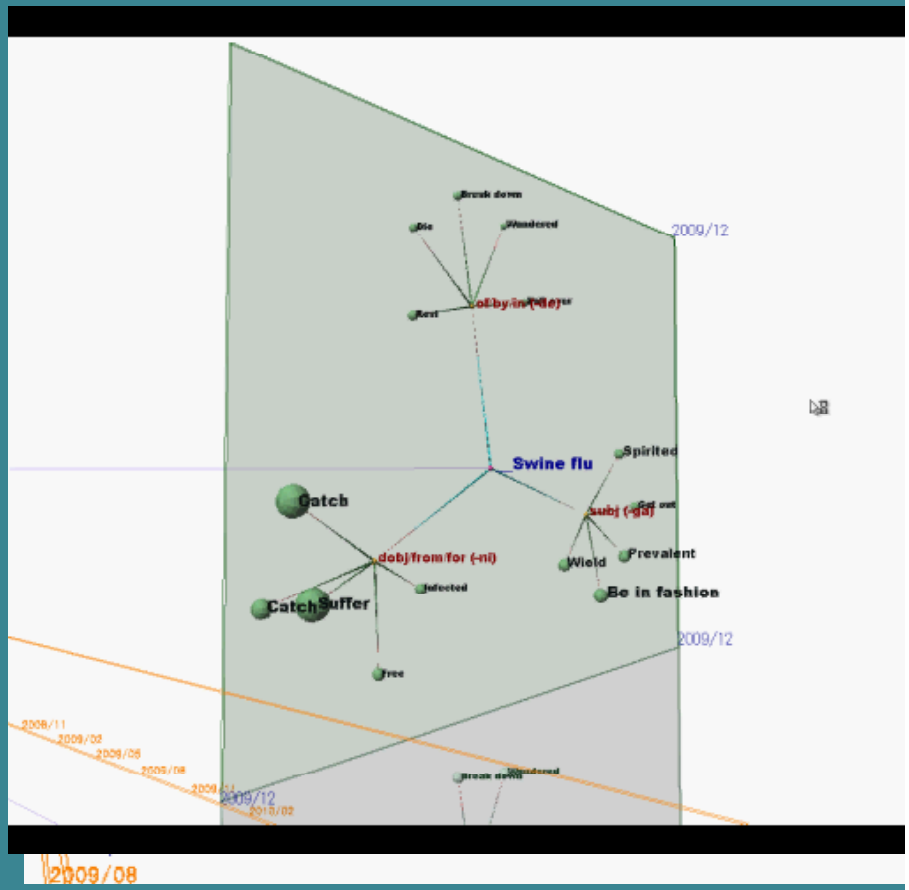
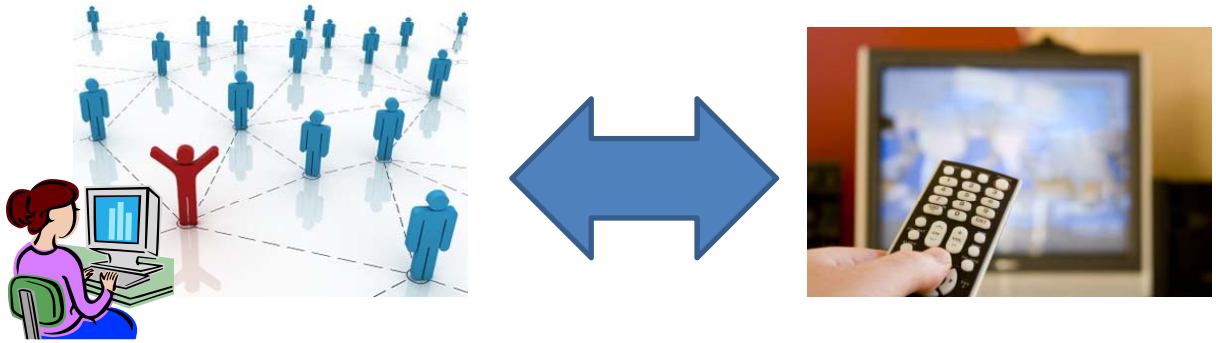


Image Flows Visualization for Inter-Media Comparison

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)

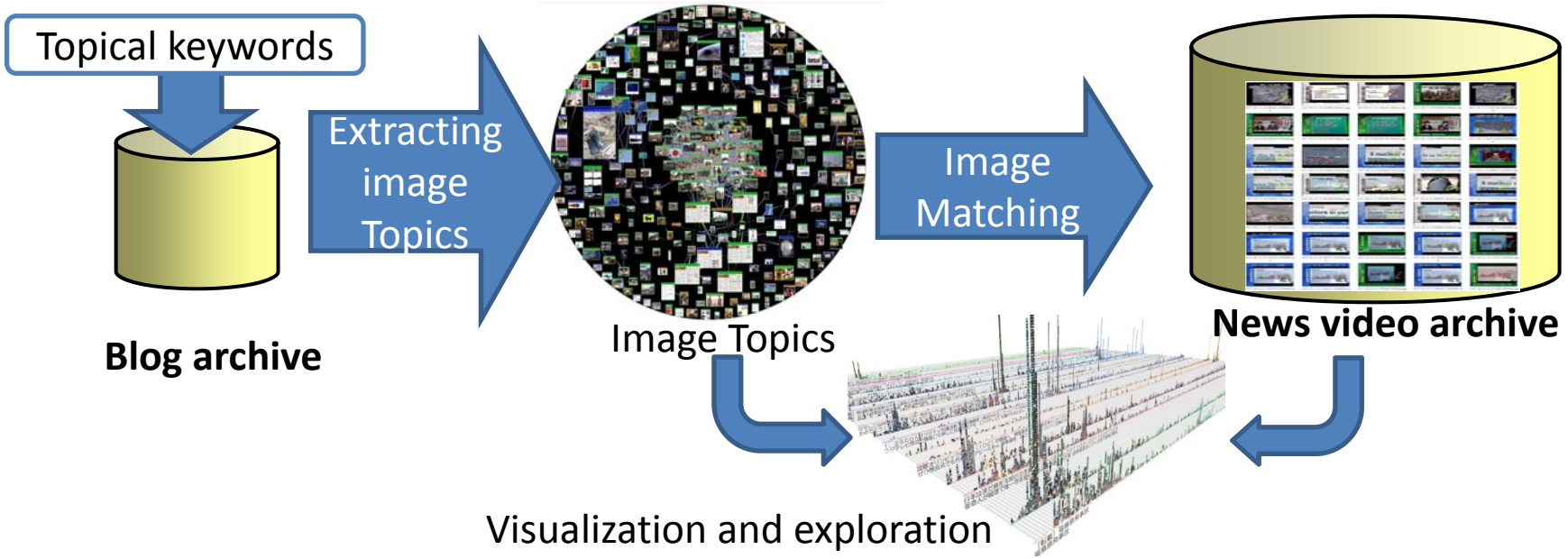
Tracking Diffusion of Topics through Blogs and TV News

- 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?



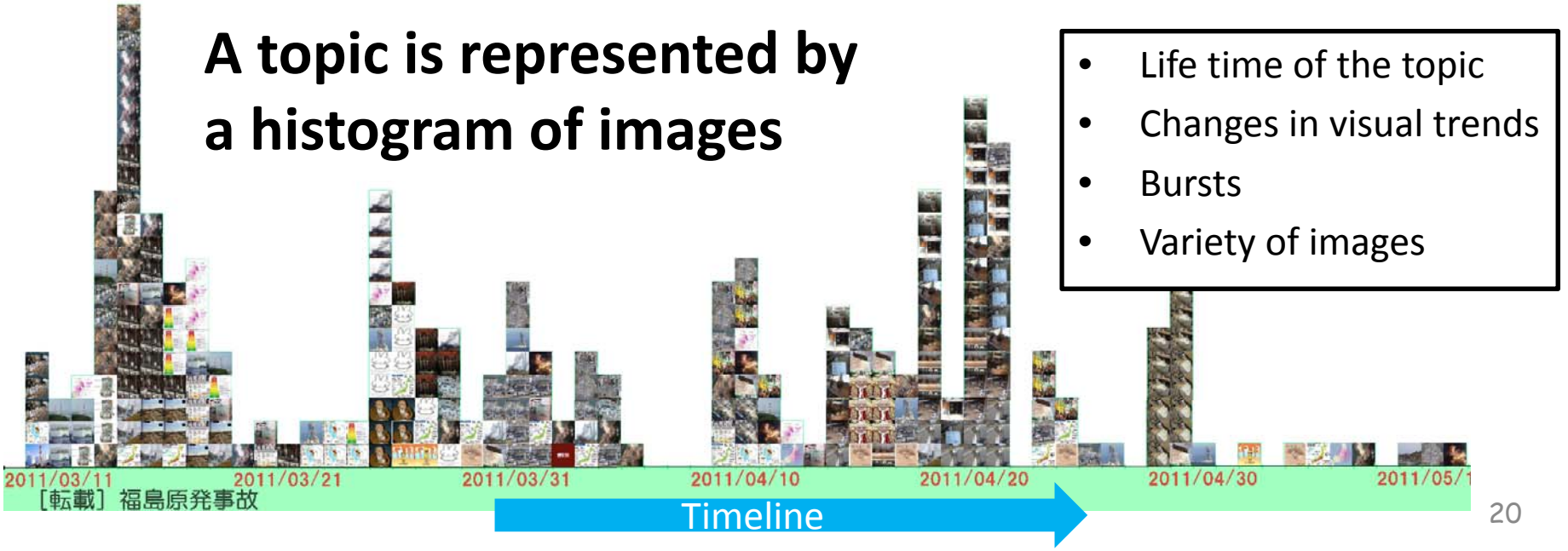
Topic Tracking based on Image Matching

- 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



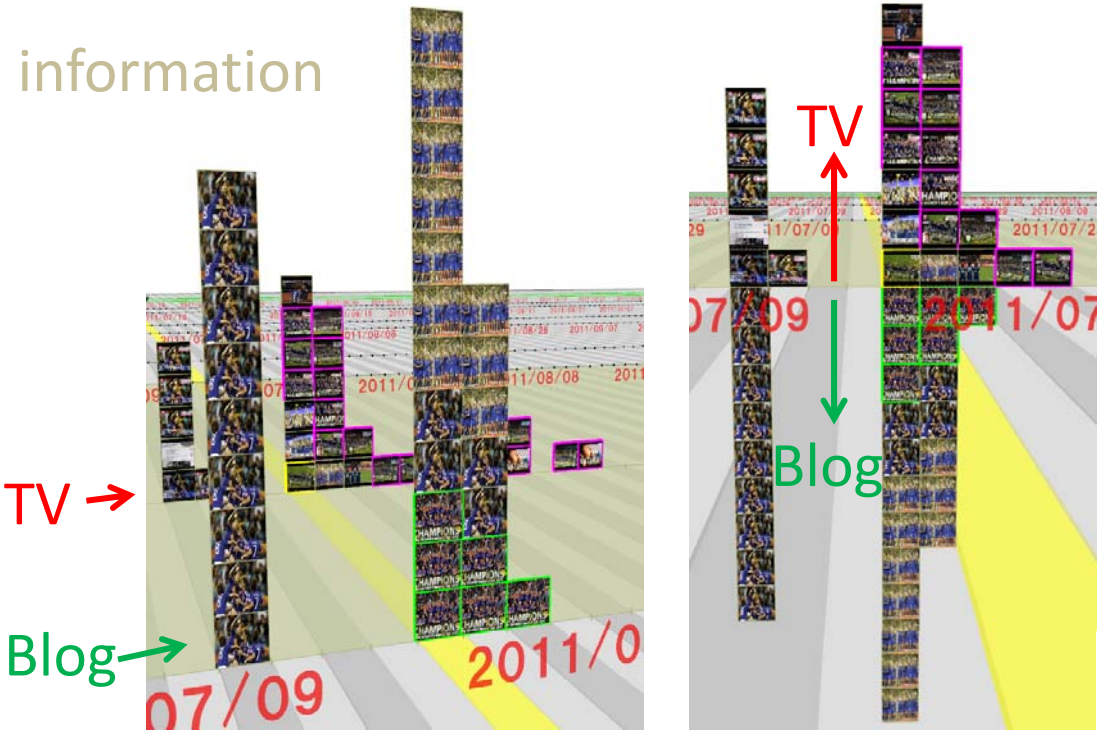
Visual Exploration of Topics

1. Visualizing the time series of image flows
2. Comparing multiple topics
3. Comparing images in different media
4. Exploring image clusters
5. Accessing detailed information



Visual Exploration of Topics

1. Visualizing the time series of image flows
2. Comparing multiple topics
3. Comparing images in different media
4. Exploring image clusters
5. Accessing detailed information



front and back

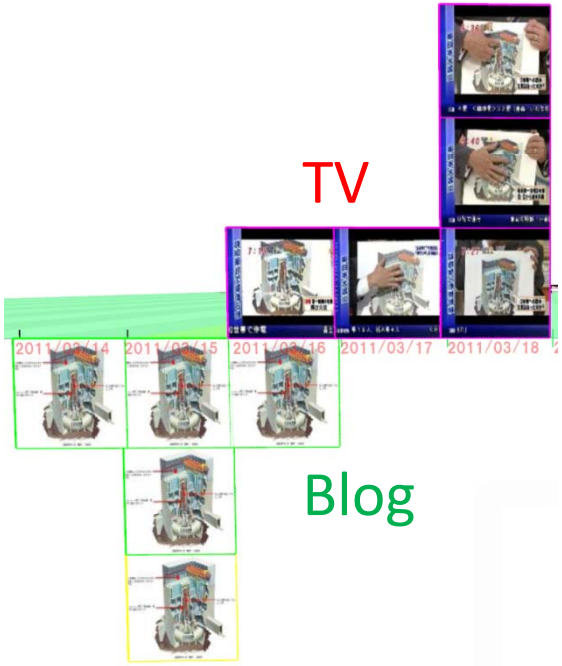
top and bottom

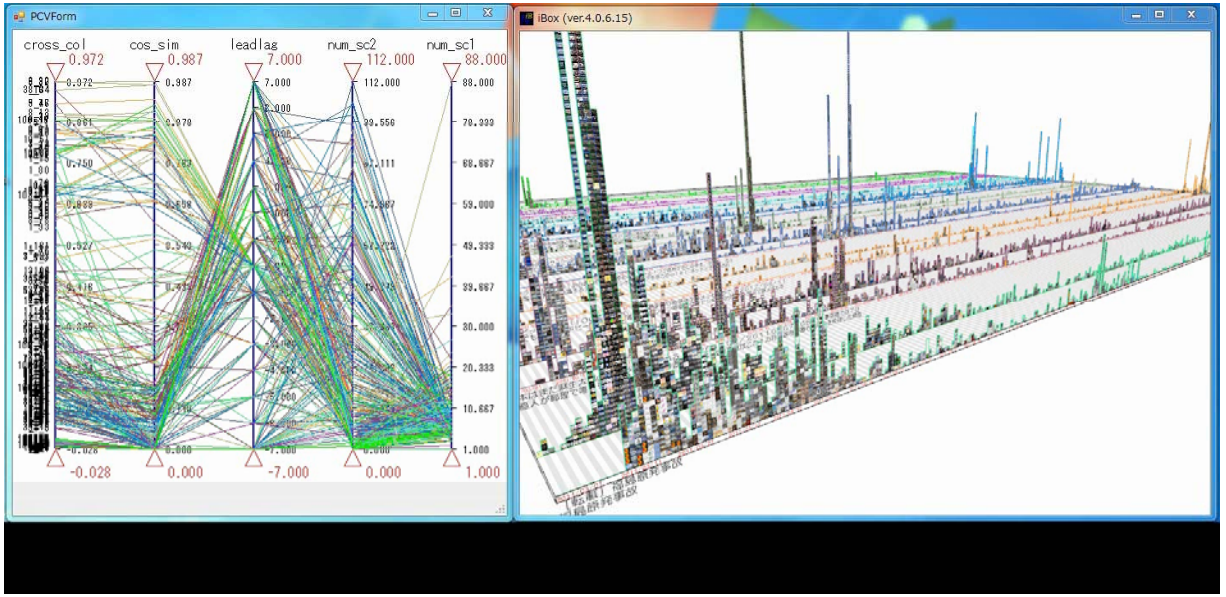
Visual Exploration Environment

1. Visualizing the time series of image flows
2. Comparing multiple topics
3. Comparing images in different media
4. Exploring image clusters
5. Accessing detailed information

For each image cluster pair, calculate and filter by:

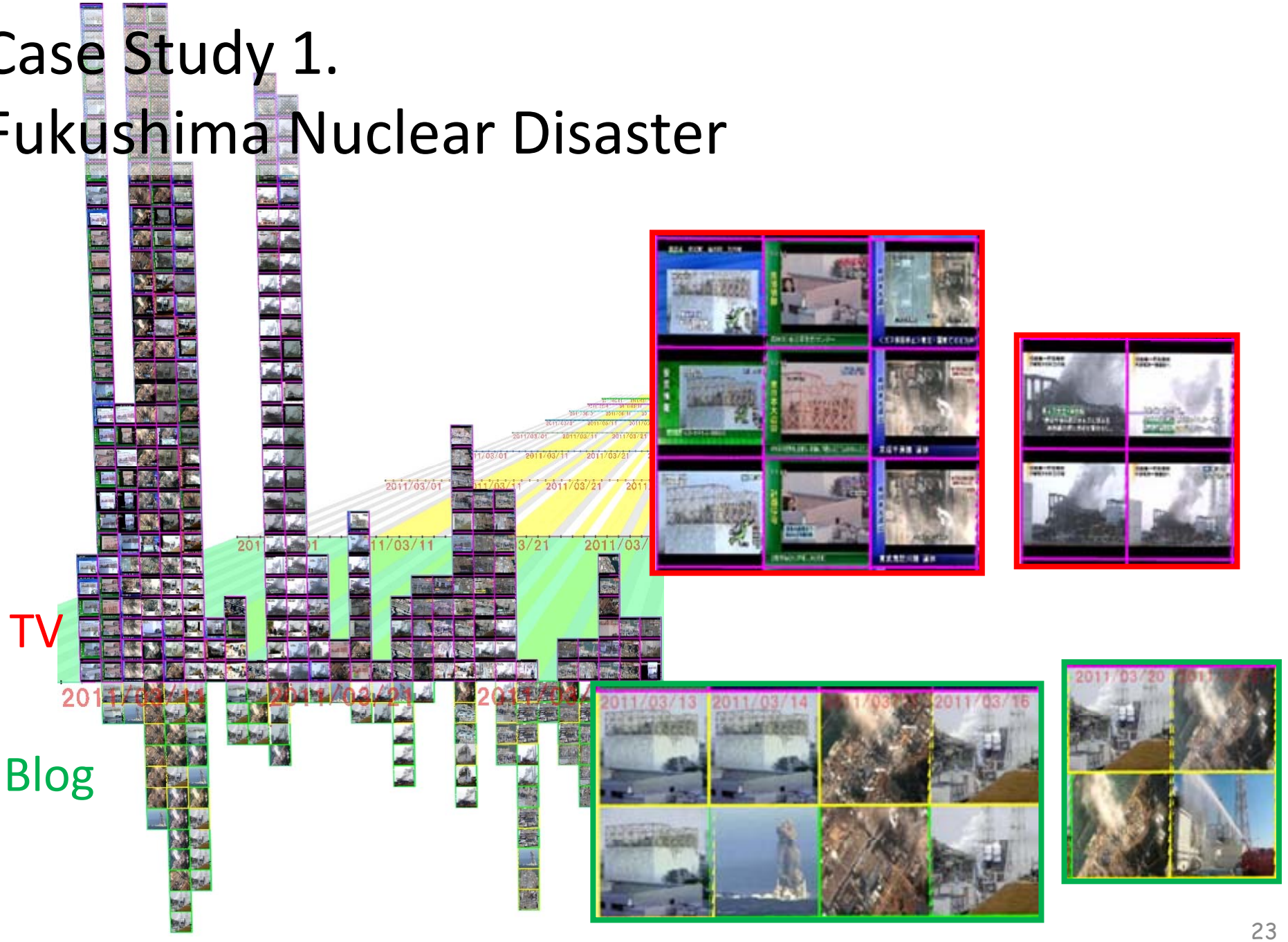
- cosine similarity
- cross-correlation and lead/lag date
- number of images



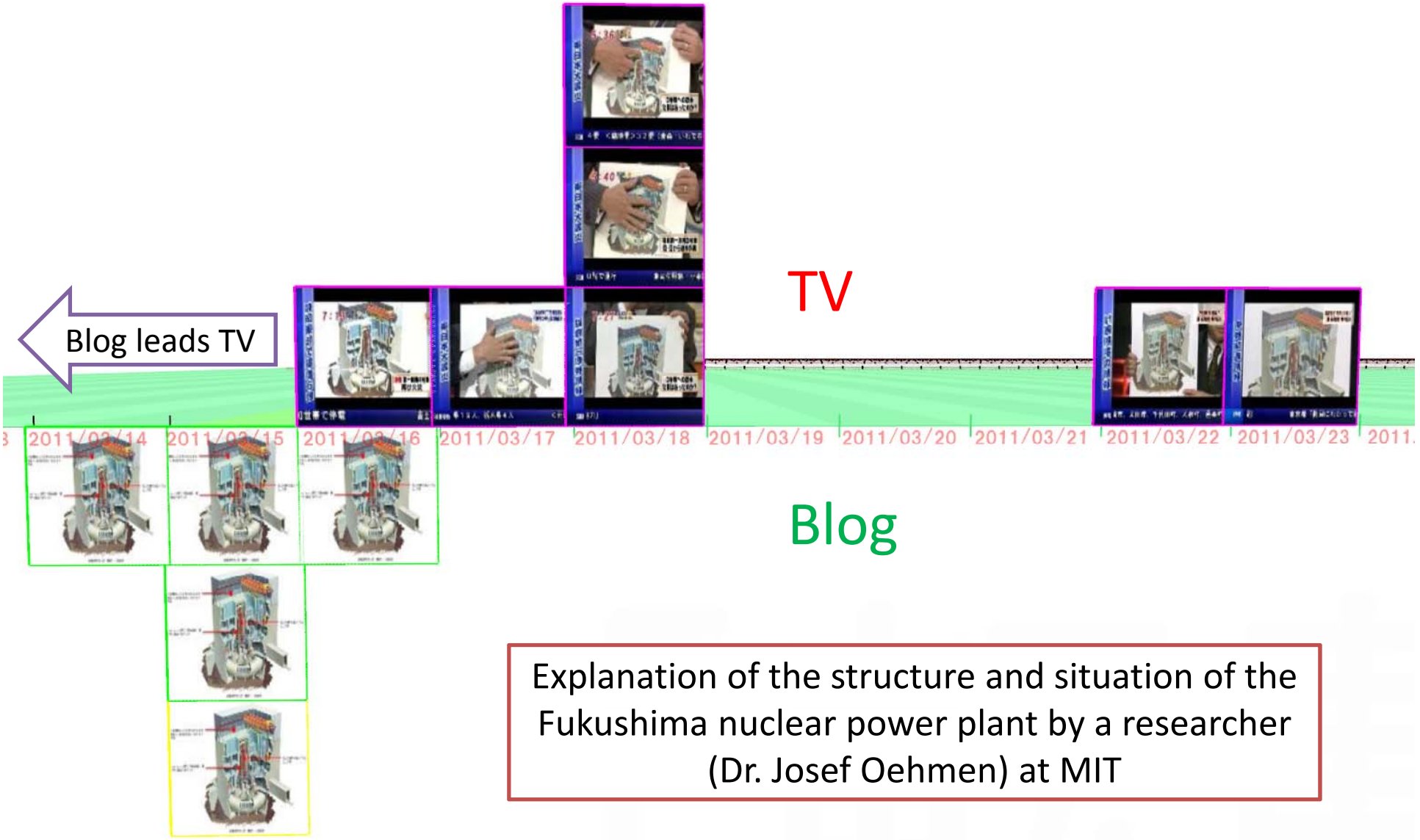


Parallel Coordinates View Image Flow Visualization

Case Study 1.
Fukushima Nuclear Disaster

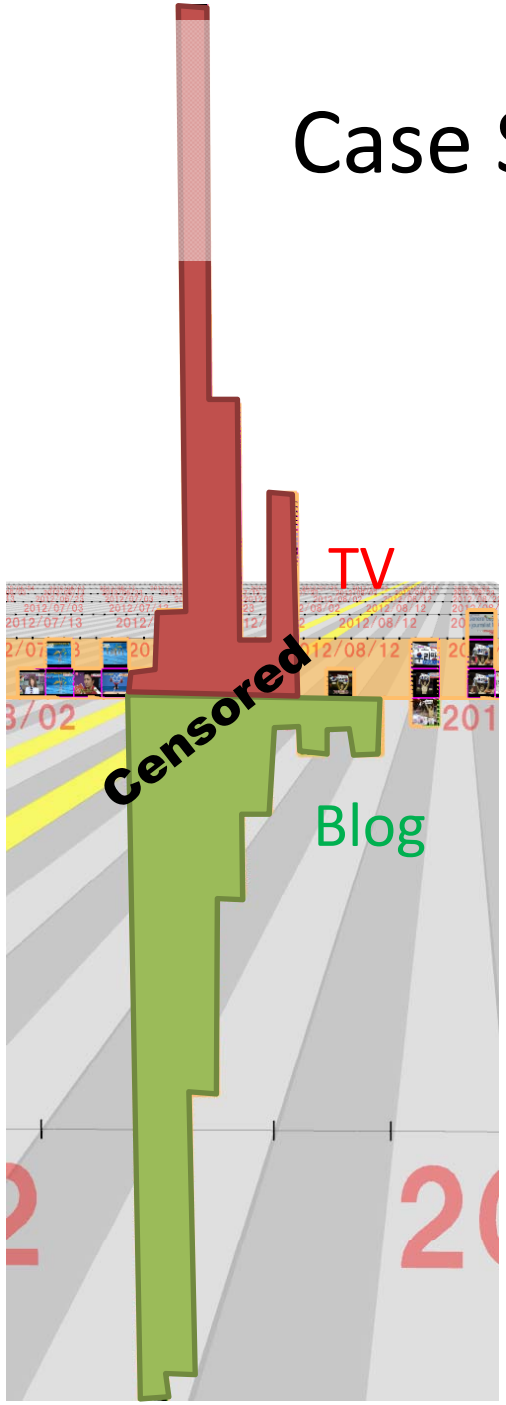


Case Study 1.
Spotting the origin of images



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

Case Study 2. Political Issue



- Some political activity in some public event was first broadcasted in TV
- In first two days, TV programs were somehow reluctant to show the image
- That activity made bloggers angry and they explosively copied the image
- After seeing that response, TV news started to treat the issue as a big problem

Platform for Multiple Media Analysis

Multiple Media Analysis

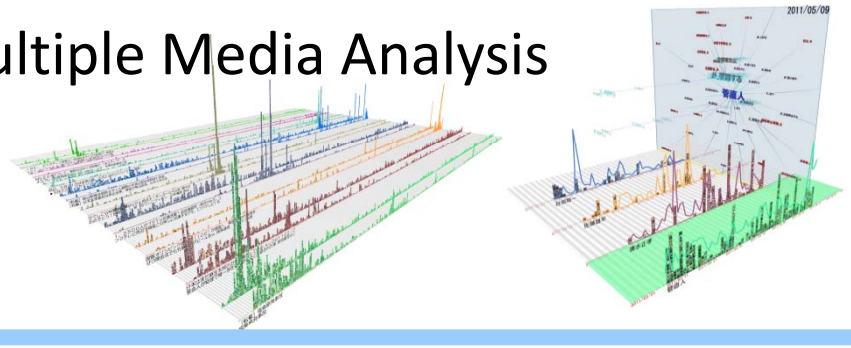
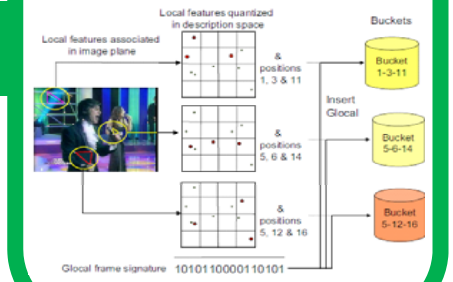
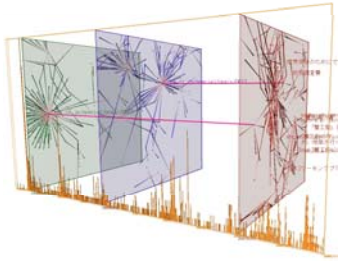


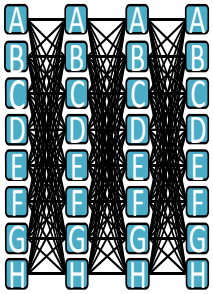
Image Linkage



Graph Analysis

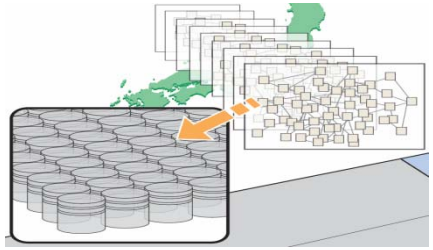


Text Analysis



Video Archive

Web media archive by continuous crawler



NII