Big Value Data, Not Just Big Data!

Dr. Joseph Reger

Chief Technology Officer
Fujitsu Technology Solutions

Powers of Ten (SI)

Metric prefixes							
Prefix	**Symbol**	**1000^m**	**10^n**	**Decimal**	**Short scale**	**Long scale**	**Since**[n 1]
yotta	Y	1000^8	10^{24}	1 000 000 000 000 000 000 000 000	septillion	quadrillion	1991
zetta	Z	1000^7	10^{21}	1 000 000 000 000 000 000 000	sextillion	trilliard	1991
exa	E	1000^6	10^{18}	1 000 000 000 000 000 000	quintillion	trillion	1975
peta	P	1000^5	10^{15}	1 000 000 000 000 000	quadrillion	billiard	1975
tera	T	1000^4	10^{12}	1 000 000 000 000	trillion	billion	1960
giga	G	1000^3	10^9	1 000 000 000	billion	milliard	1960
mega	M	1000^2	10^6	1 000 000	million		1960
kilo	k	1000^1	10^3	1 000	thousand		1795
hecto	h	$1000^{2/3}$	10^2	100	hundred		1795
deca	da	$1000^{1/3}$	10^1	10	ten		1795
		1000^0	10^0	1	one		–

Source: Wikipedia

Big Data?

FUJITSU

45 million
servers

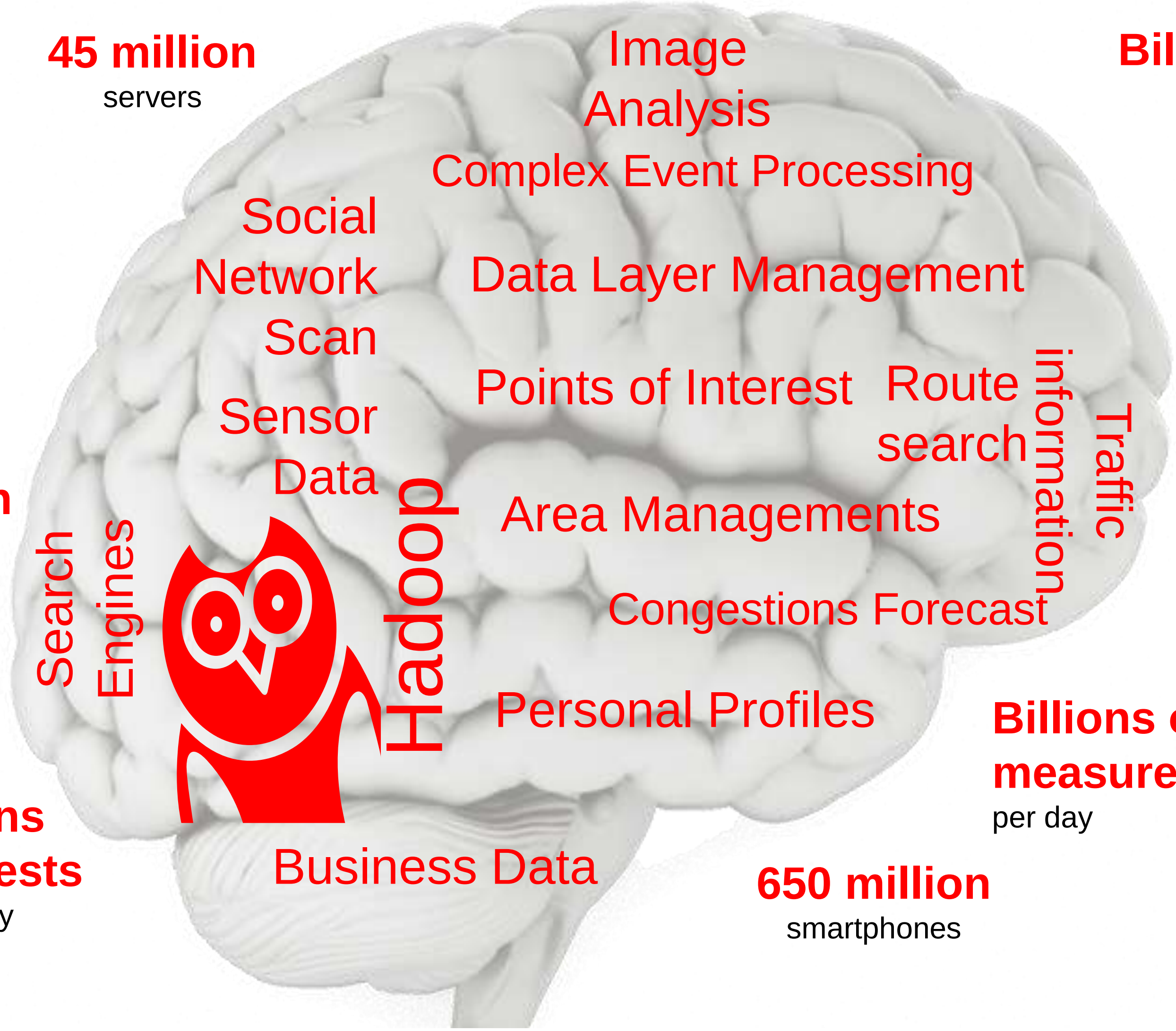
100,000
Flights/day

1 billion
users

200 million
pictures/day

1 billion
PCs

Billions of requests
per day



Billion objects
per hour

1 billion
rides per day

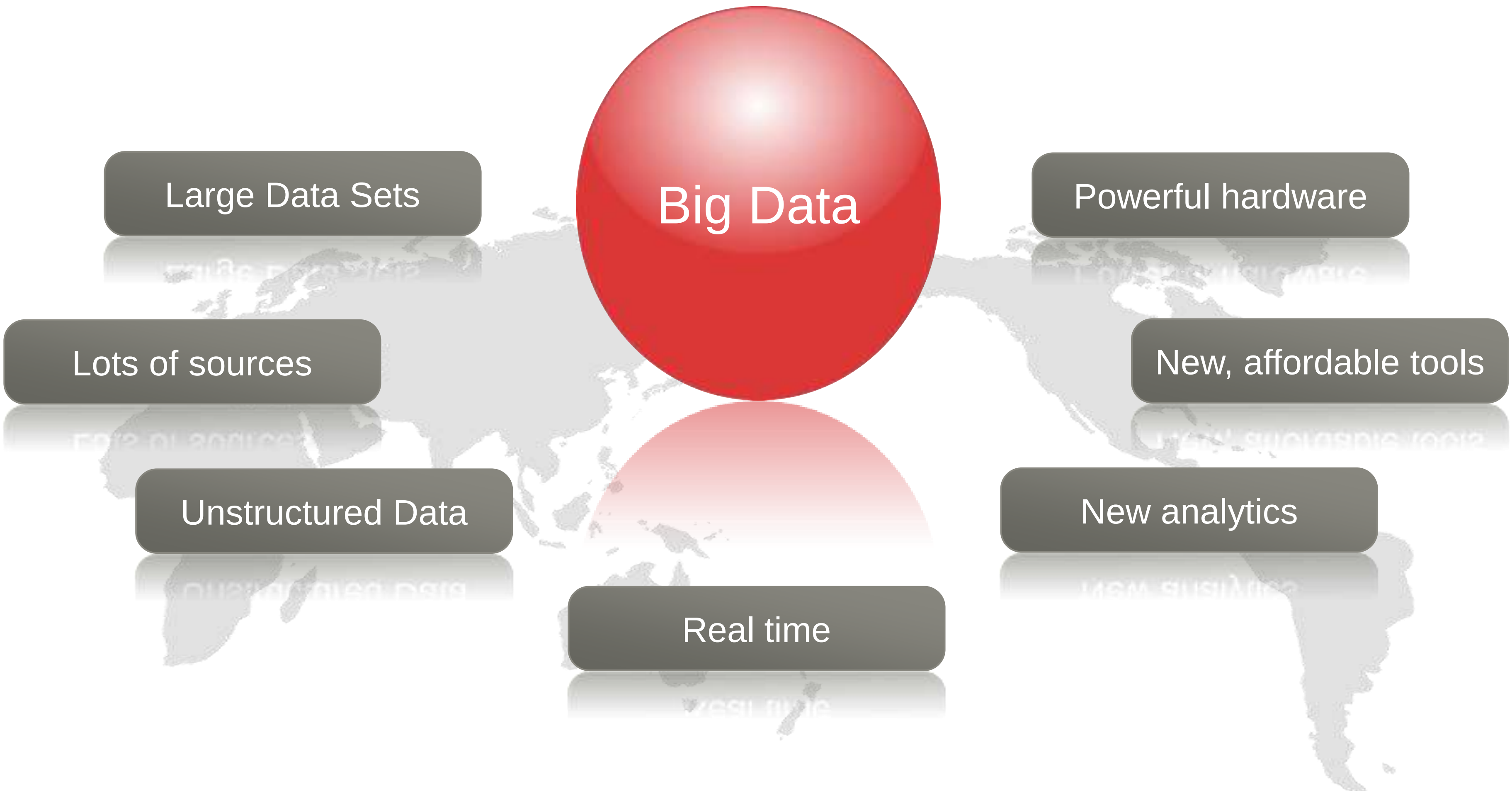
100s
of millions locations

Billions of measurements
per day

650 million
smartphones

1 billion cars

New (Re)Sources, New Technologies



New use cases

Big Value Data!

Big Data?

100,000
Flights/day

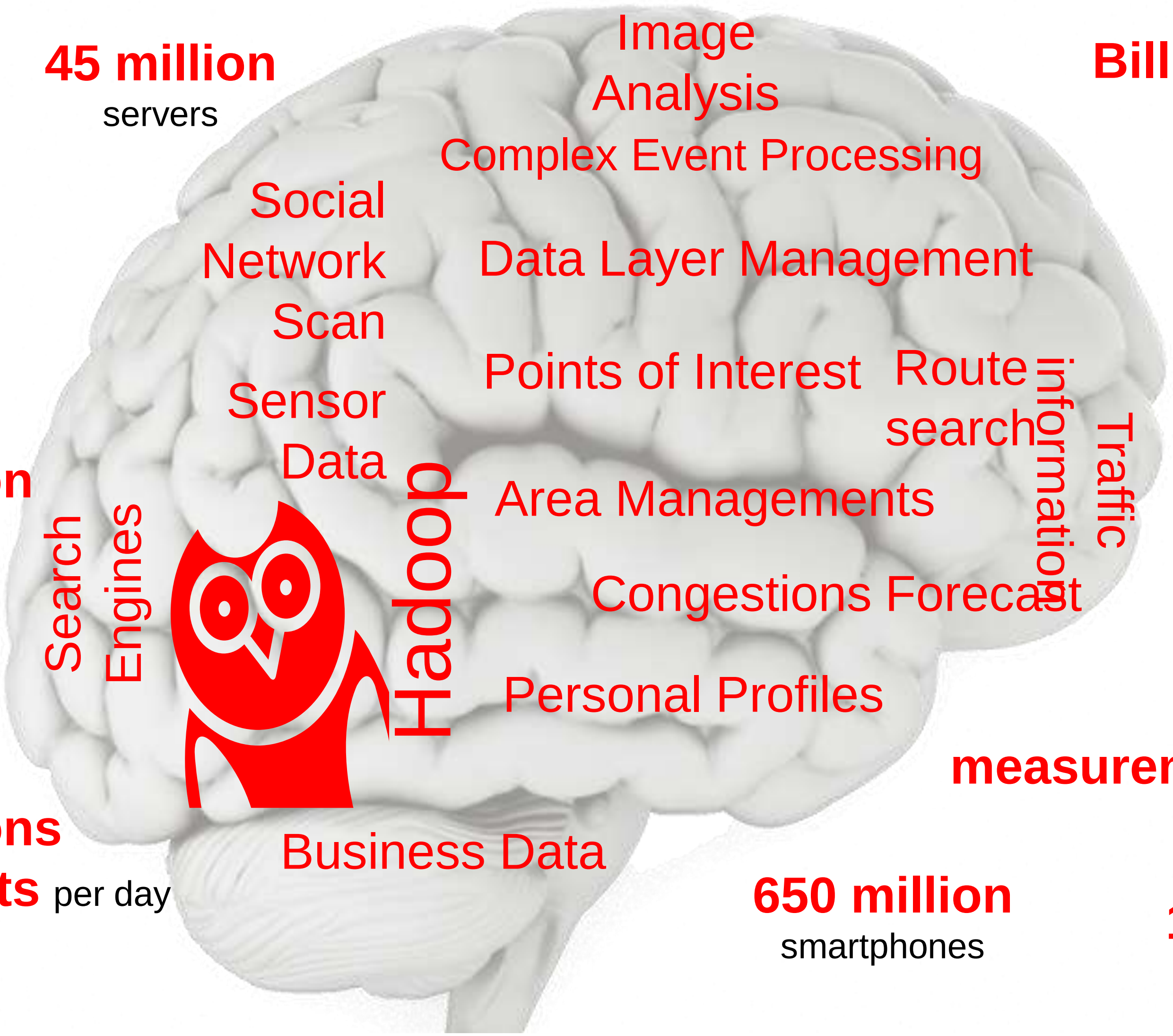
45 million
servers

1 billion
users

200 million
pictures/day

1 billion
PCs

Billions of requests per day



Billion objects
per hour

1 billion
rides per day

100s
of millions locations

Billions of measurements per day

650 million
smartphones

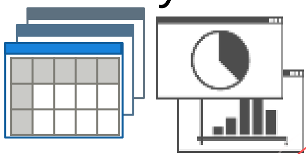
1 billion cars

Big Data research @ Fujitsu

FUJITSU

Big Data

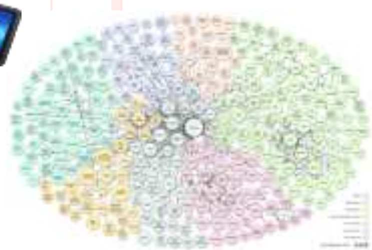
- Statistics / history



- Social media – Twitter, blogs etc.

- Sensor data – GPS, weather, etc.

- Open data



Collection

- Distributed data collection
- Privacy security
- Linked open data

Data processing

Fast processing of large-scale data

Analysis

Data/text mining
Analysis platform
Optimization
Simulation

Topics:

- Social media analysis
- Analytic templates
- Optimal area discovery
- Social simulation

Application

Transportation
Energy
Disaster recovery
Marketing

Topics:

- Transportation simulation
- Smart grid
- Event detection from SNS

Topics:

- Incremental data processing
- Parallel complex event processing
- Stream data aggregation

Faster, more intelligent, more secure technologies

Application Areas

FUJITSU

Sectors

Healthcare

Retail

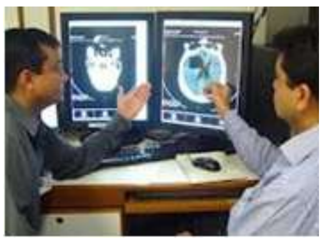
Manufacturing

Government

Agriculture

Usage

Disease Prevention



Multi-channel Sales

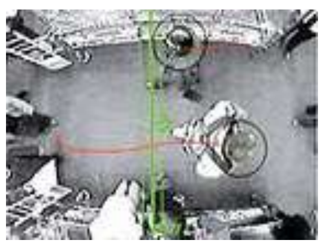
Behavior Management

Traffic Management

Production Efficiency

New Drug Development

Customer Behavior



Asset Management

Homeland Security

Livestock Reproduction Management

Bio Technology

SCM

New Service M2M

Crime Prevention

Inventory Management

Research on Big Data use cases by Fujitsu Laboratories & Fujitsu Limited

Kozo Otsuka

Technology Office
Fujitsu Technology Solutions

Early detection of lifestyle related diseases

FUJITSU

Accurate prediction by the variety of health related data analysis

Medical record data		Health checkup data		Vital data
	×		×	
12 M over 5 years		**0.8 M** over 5 years		**¼ M** from sampling for ½ year

- **Early detection of signs of habits leading to lifestyle diseases**
- **Link to expert advice for preventive actions**

■ Example: Diabetic patients & candidates in Japan: 33% of male, 23% of female adults (2011)

Prediction method

Fujitsu developed methodology

Conventional diag. param.

- HbA1c
- Blood Glucose

+

Reg. health checkup data

- Serum creatinine
- HDL cholesterol
- BMI
- Platelet count
- γ-GT(γ-GTP)
- Abdom. Circum.
- GOT(AST)
- MCH
- Total protein
- MCV
- White blood cell
- GPT(ALT)
- MCHC
- Serum uric acid
- Diastolic blood pressure
- Neutral fat
- LDL cholesterol
- Total cholesterol
- Systolic blood pressure
- Hematocrit
- Hemoglobin content
- …

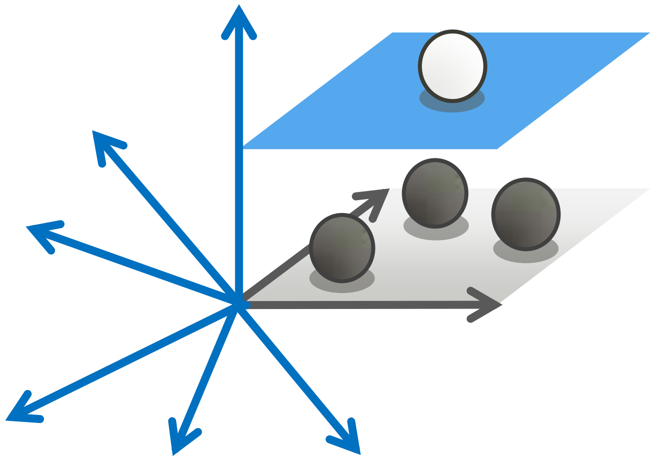
Medical record data

- Diag./Treatment
- Prescriptions

Machine Learning

to create rules

2,000-dimensions



Finer distinctions for accurate prediction by high dimensional rules

Example: predict diabetes

Test conducted targeting Fujitsu employee volunteers (26,000)

Fujitsu employees
(Including diabetes patients)

1. Collect and sort relevant data

Historical medical record data

Historical health checkup data

Historical vital data

2. Analyze all data to build 'highly probable diabetes model'

- Vital
- Med. rec.
- Checkup

3. Feed target individual's data

Analytics

- Vital
- Med. rec.
- Checkup

4. Compare against model

Outcome

Probability
HIGH

5. Diabetics probability

Reduce medical care cost by accurate prediction

Extremely high population density in Tokyo

FUJITSU

Tokyo population density

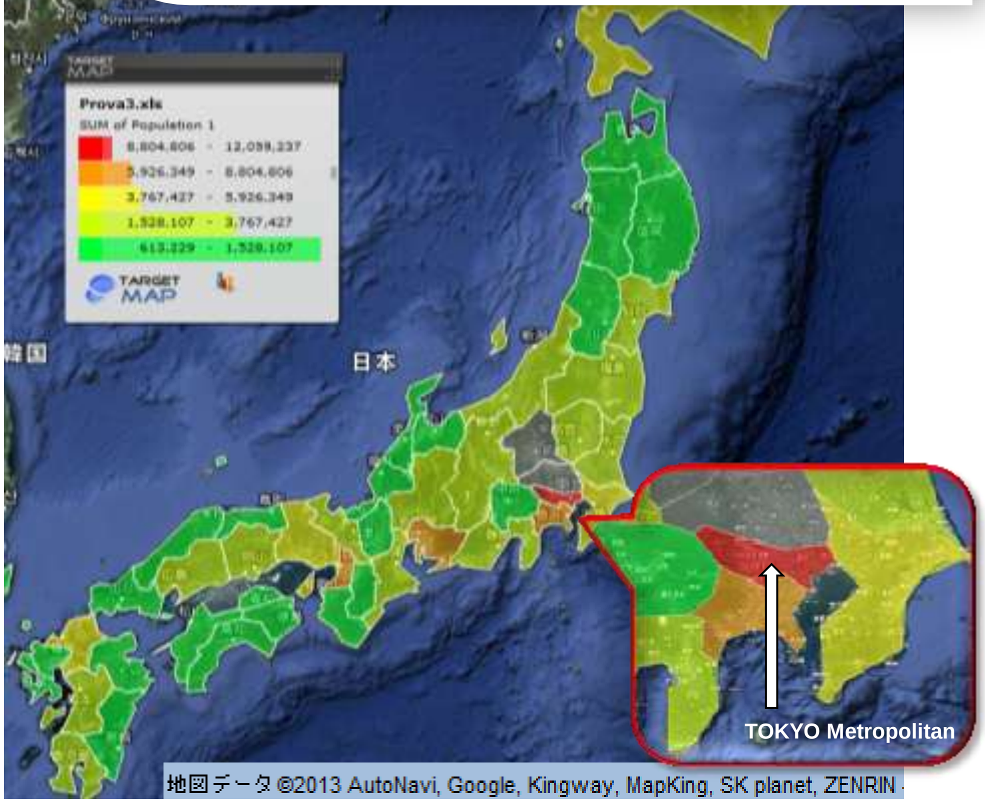
- Over 12 million people in Tokyo metropolitan area (= ca. 10% of the total Japanese population)
- Over 14,000 people / km^2 in Tokyo city (ref. ca. 4,400 people / km^2 in Munich, Germany)

▶ Small personal space, higher risk for human conflicts

- Over 6 million people in Tokyo metropolitan area using smartphones during commute, office & private hours

▶ Huge # of social media data


https://en.wikipedia.org/wiki/Greater_Tokyo_Area

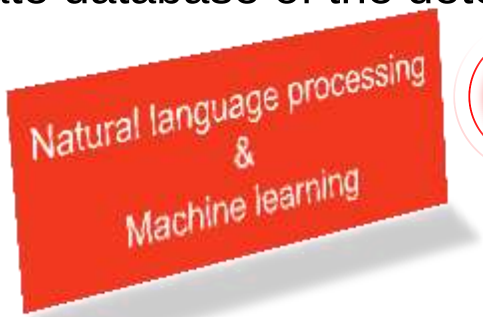


http://www.targetmap.com/viewer.aspx?reportId=5845

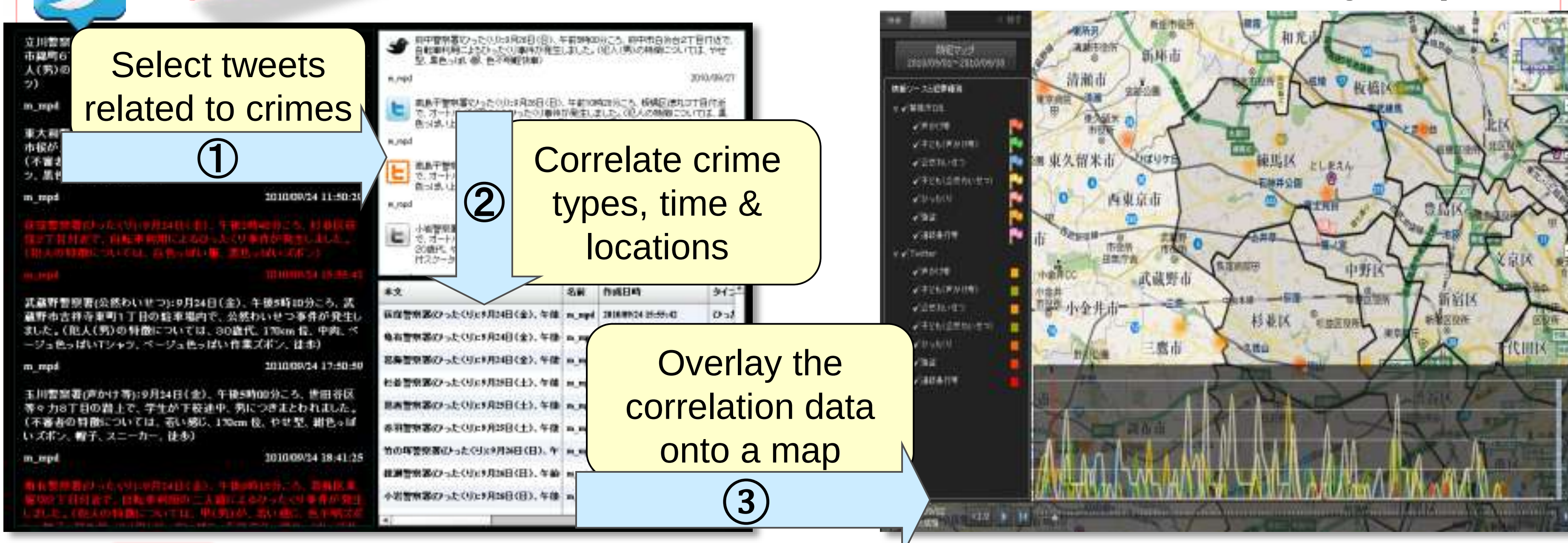
Map of criminal activities

Visualization of social network information

- Use Twitter (40 mil. tweets / day in Japan) as huge number of event sensors
- Create database of the detected events mapped to geographic locations
  - Filtering and selecting tweets for a target topic
  - Classify selected tweets into sub-categories
  - Identify locations of the events in the tweets



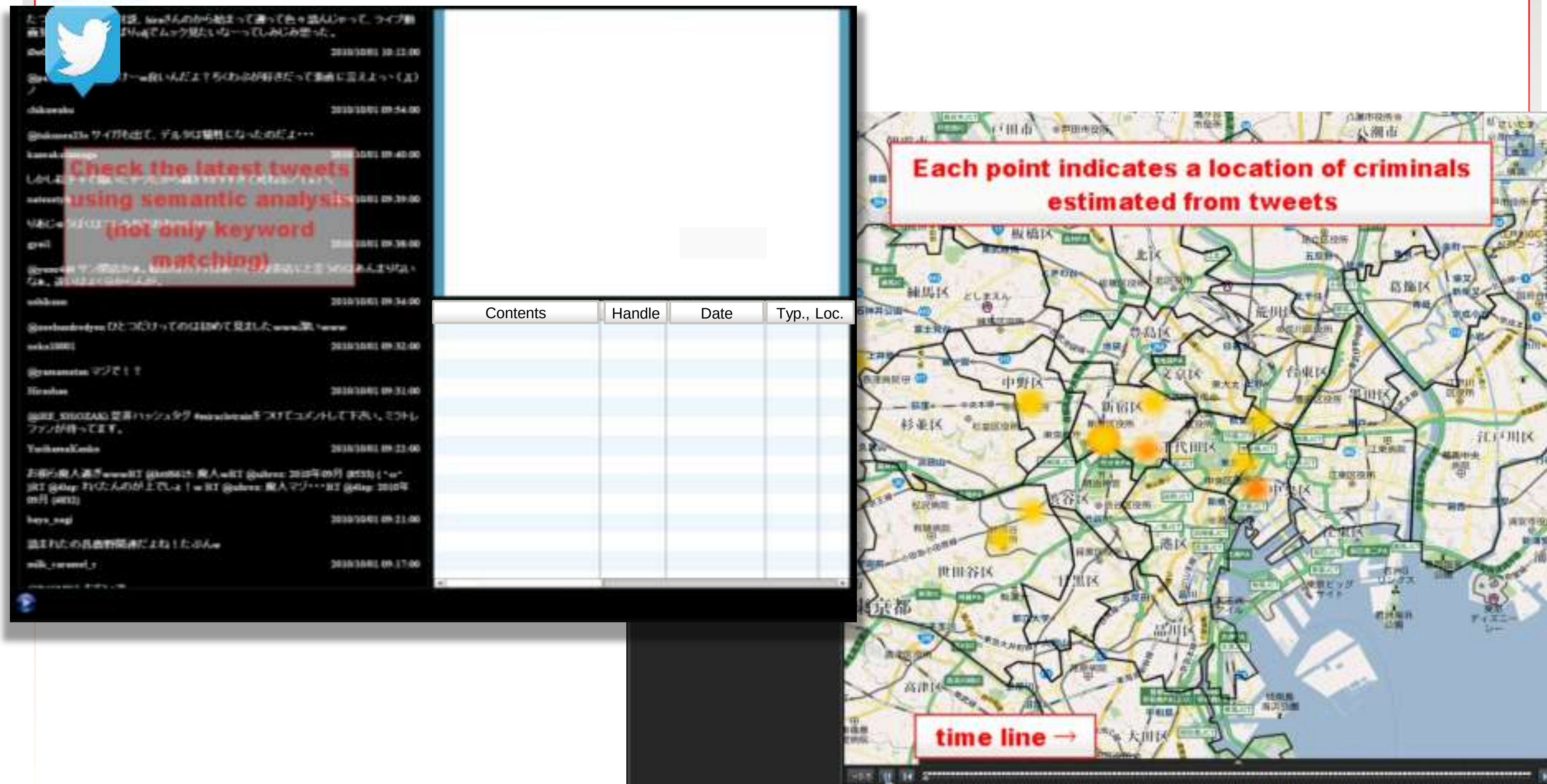
Criminal activity map



Map of criminal activities – test run

Visualization of criminal activity related tweets

Showing the semantic analysis & machine learning phase



Showing the mapping of criminal activities onto the Tokyo map

Japan – very vulnerable to climate change

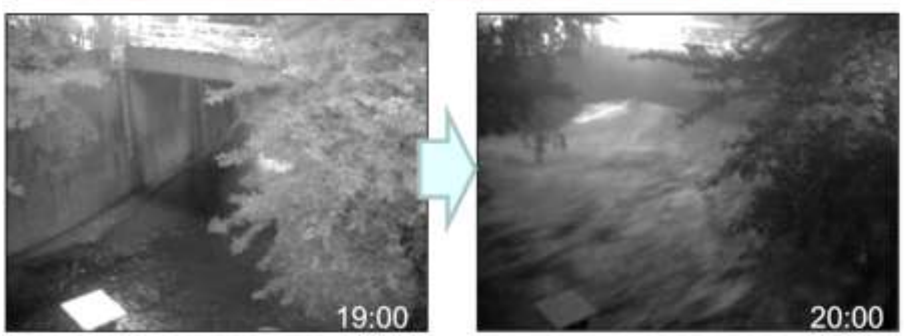
FUJITSU

Natural disaster due to rainfall

- Ca. 4 Billion U.S. dollars of property damage annually caused by flooding & inundation
- Over 1,100 landslides annually
- Over 100mm per hour precipitation from torrential rain
- Precipitation increasing every year

Strong need for early warnings & preventions

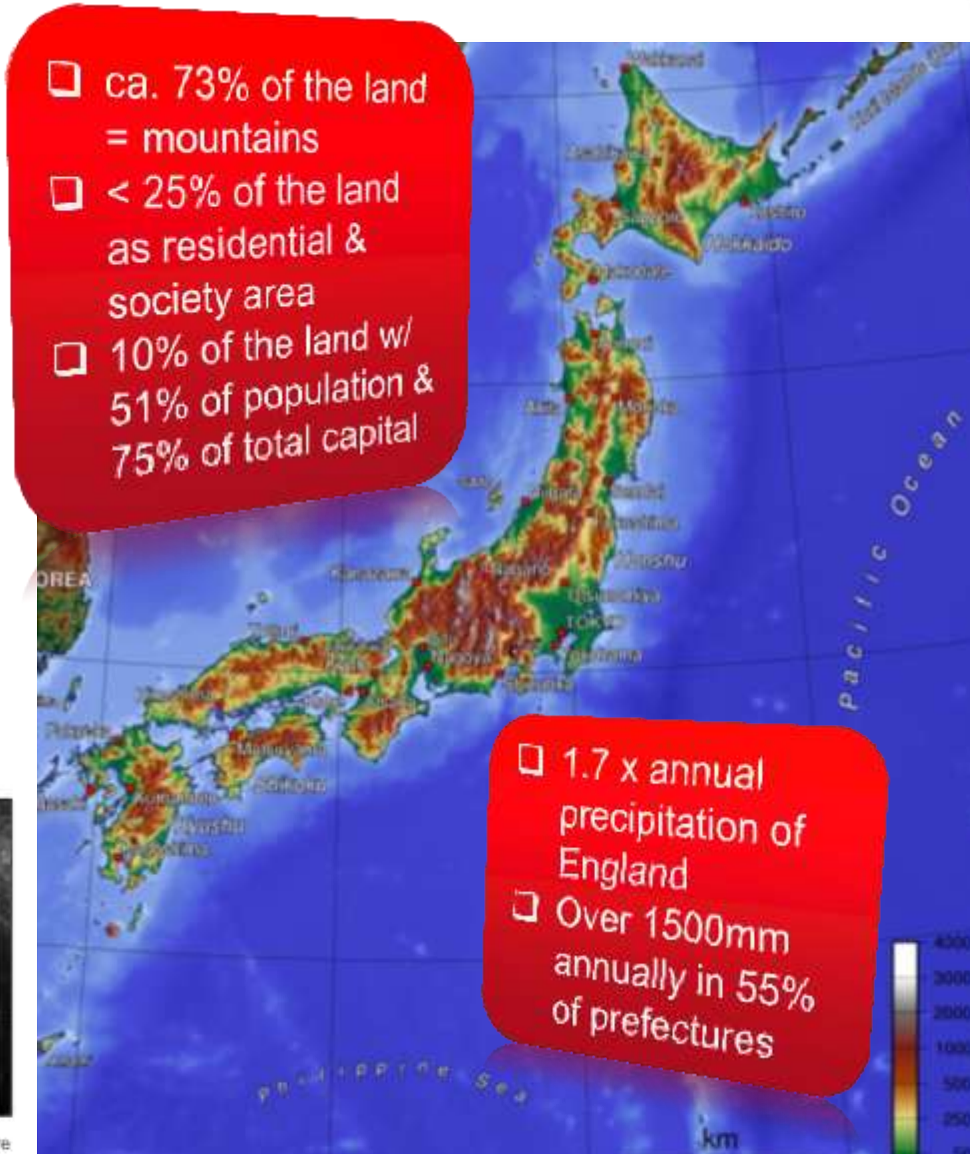
Water level rose by 3.45 m in 10 minutes from 19:50 to 20:00.



Change of water level in Shakujii River

Based on investigation by Disaster Prevention Office, Fire Disaster Management Agency on September 11, 2009

http://www.mlit.go.jp/river/basic_info/english/pdf/conf_01-0.pdf



https://en.wikipedia.org/wiki/Geography_of_Japan

Finer granularity of rain observation in Japan

FUJITSU

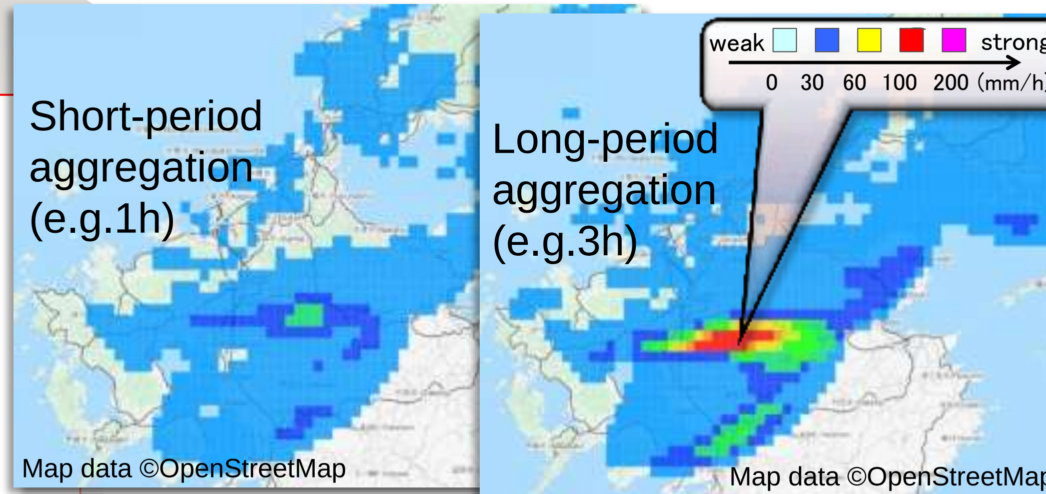
Leverage “XRAIN”* radar

- Compared to the conventional:
  - 5x more frequent data (1 min)
  - 16x finer mesh resolution (250m)
  - 3D scan – raindrop information
  - Over 100 times data increase
- Over 500K records per minute per zone (w/ up to 4 radars)

More precise & more real-time

- With Fujitsu big data processing:
  - Aggregation of up to 100 mil. records within 10+ secs., updated every 1 min.
  - Real-time aggregation of total rainfall since the 1st drop for each mesh

Detect potential disaster areas w/ the fast data aggregation



5km-mesh rainfall data by XRAIN (Source: Water & Disaster Mgmt. Bureau, Ministry of Land, Infrastructure, Transport and Tourism)

Figure 1: Radar distributions and their coverages of MLIT X-band MP radar network. Triangles indicate

http://www.raingain.eu/sites/default/files/maesaka_seminar_ecole_ponts_2_july_2012.pdf
https://ams.confex.com/ams/35Radar/webprogram/Manuscript/Paper191685/35RADAR_Maesaka.pdf

XRAIN = X-band MP radar developed by NIED for MLIT*
*NIED = National Research Institute for Earth Science and Disaster Prevention, Japan
*MLIT = Ministry of Land, Infrastructure, Transport and Tourism
*C-band radar = currently, the most popular weather radar type in the world

Estimate status quo w/ Social Network Services

FUJITSU

Use a certain tendency from large SNS data to identify the status quo

- More precise analysis of location information contained in SNS through “Area of life” analysis to overcome the following challenges:
  - Only 0.5%(*) of SNS posts including GPS information
  - Only 30%(*) of posts containing landmark information(e.g. town name vs. neighborhood)
  - Only coarse “resolution”(municipal area, state, prefecture) as SNS user location profile
  - Unreliable ties between contents of posts and user location profiles (e.g. hearsay vs. real experience)

* According to the analysis on Tweeter tweets conducted by Fujitsu Labs

SNS User’s posts

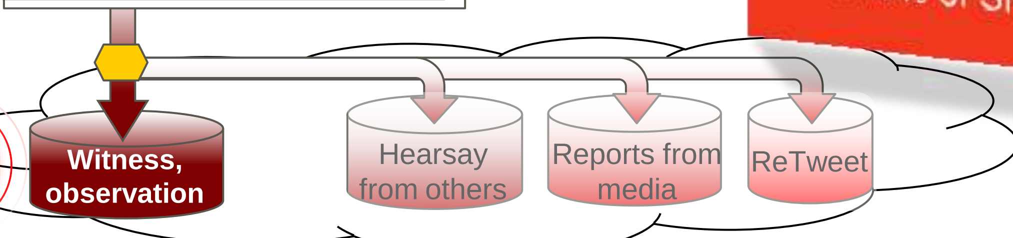
User’s history of posts

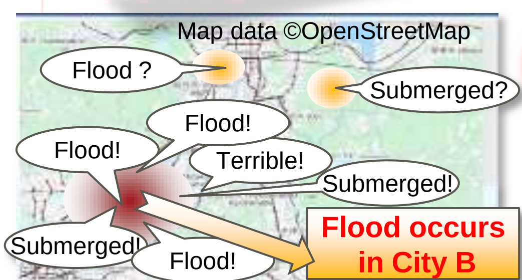
“Area of Life”
Draw rough prediction of the area where each user spends his/her daily life from the landmarks in the past posts - anonymously

Machine learning & Natural language processing

Disaster related posts + “Area of life” (e.g. City B)

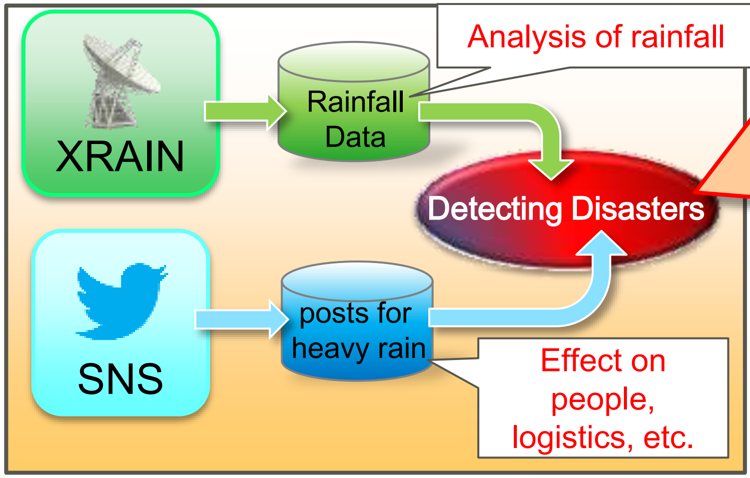
“Area of life” helps to estimate location information of 70% of SNS posts.

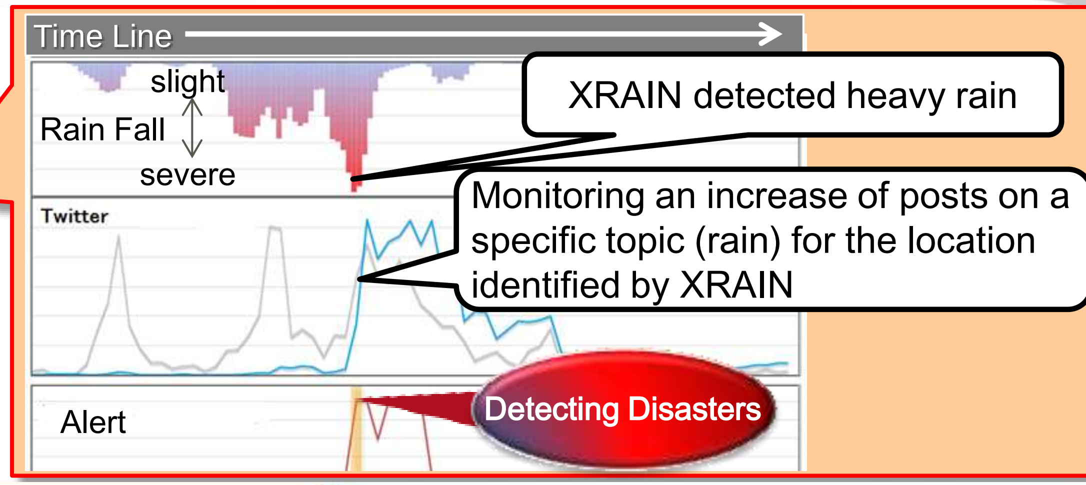




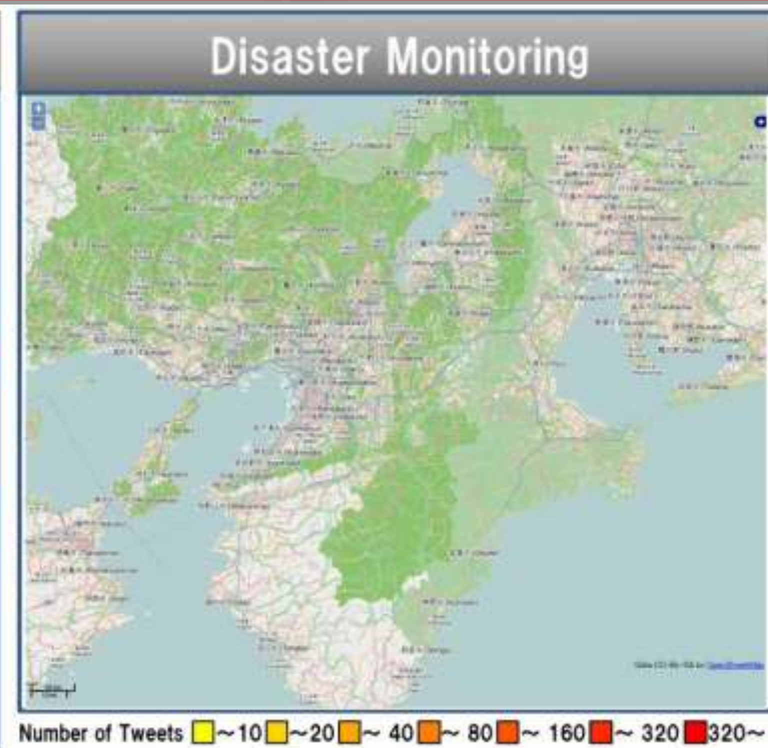
Higher potential of disaster status quo estimated by the large # of posts regarding towns/citys

Big Data use: Disaster alert w/ SNS & XRAIN





* XRAIN data supplied courtesy of Ministry of Land, Infrastructure, Transportation and Tourism of Japan



Test run in 2012 for towns in Osaka & Kyoto prefs. raised alerts 3 h earlier than conventional warnings

2011.03.11 Tohoku Earthquake and Tsunami

FUJITSU

Damages caused by Tsunami

- Epicentral area – ca. 500km long (N-S) and 200km wide (E-W)
- Max. 14.8m Tsunami height, up to 40m Tsunami run-up height
- 535km^2 of land inundated by Tsunami in Tohoku & Kanto region
- ca. 129,000 buildings destroyed
- ca. 15,850 fatalities & 3,282 missing
- Over 20,000 cars swept away – a lot of them were in traffic jams
- Over 179 mil. tweets in that week, many asking for help

Earthquake simulation initiative by the Japanese government

1) Earthquake & Tsunami simulation: predict power & speed

2) Building response simulation: predict effects on infra.

3) Evacuation activity simulation: protect human lives

Japan – surrounded by fault lines

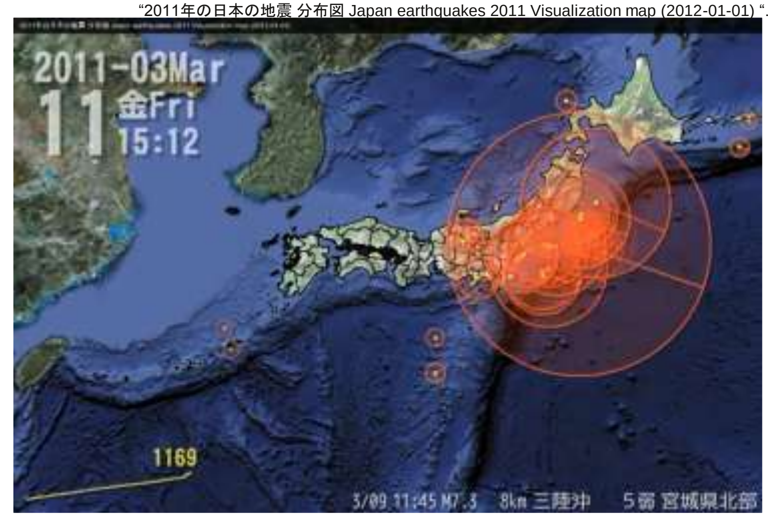
1) Earthquake & Tsunami simulation: predict power & speed

Fault-line slides trigger Tsunami

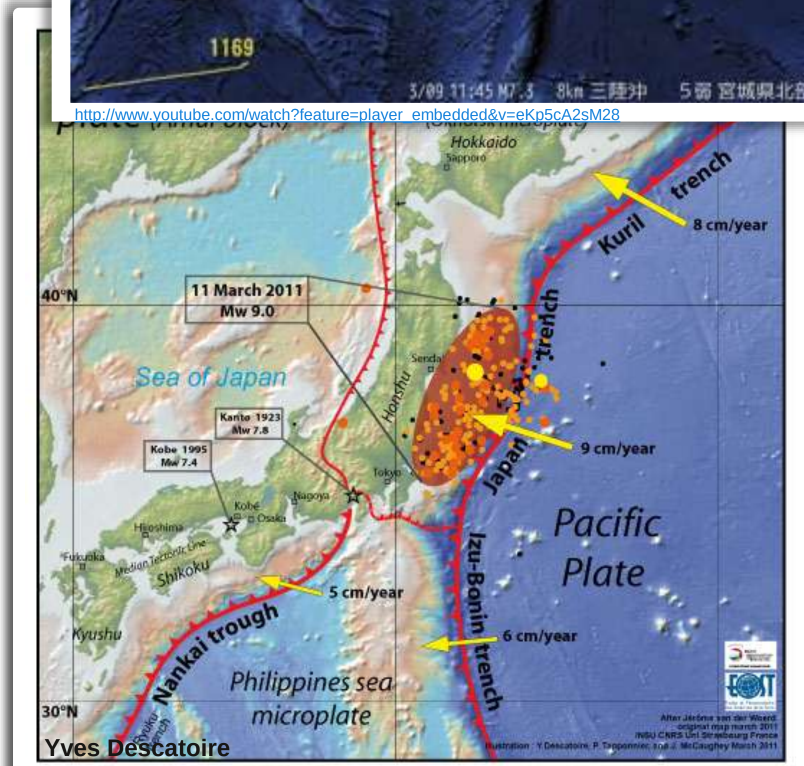
- The impact underestimated by the existing Tsunami warning system during Tohoku earthquake on Mar. 11th, 2011
- No sensor to measure the amount of fault line slide – no accurate way to predict the Tsunami speed & power, yet
- Urgent need for the alternative real-time Tsunami prediction system leveraging:
  - Sensors on the surface and the bottom of the ocean (GPS buoy, seabed wave gauges, coastal tide gauges, etc.)
  - More accurate real-time analysis on the source area of Tsunami and the Tsunami source model using the new & accurate sensor data input (above)

1st Step: Solid simulation algorithm based on Tohoku earthquake data

"2011年の日本の地震 分布図 Japan earthquakes 2011 Visualization map (2012-01-01) ".



http://www.youtube.com/watch?feature=player_embedded&v=eKp5cA2sM28



http://www.earthobservatory.sg/media/news-and-features/294-japan2011.html

Big Data use: Simulation for accurate early warning

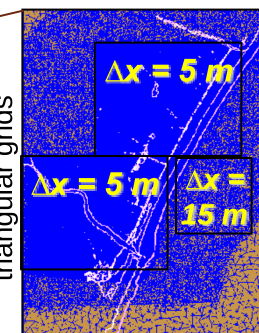
FUJITSU

1) Earthquake & Tsunami simulation: predict power & speed

Research on real-time & high-res. simulation for more accurate warning

- Using K computer (京)
- Simulation of the 2011 Great Tohoku tsunami
- Simulation resolution ranges over 5 m – 405 m
- The inundated region shown w/ black dotted line
- 5 m resolution used in the red boxes in this simulation focusing on Sendai city

Zoom-up on triangular grids



- Non-linear simulation to solve the Tsunami wave & flow
- ca. 16 mil. triangular grids & finer grids over Sendai
- ca. 16 k calc. steps in 120 min. long simulation
- 23 min. for calc. w/ ca. 9.3 % of K computer total core #

More accurate & faster Tsunami warning

Oishi *et al*. (2013, JpGU)

Japan – many cities facing the ocean

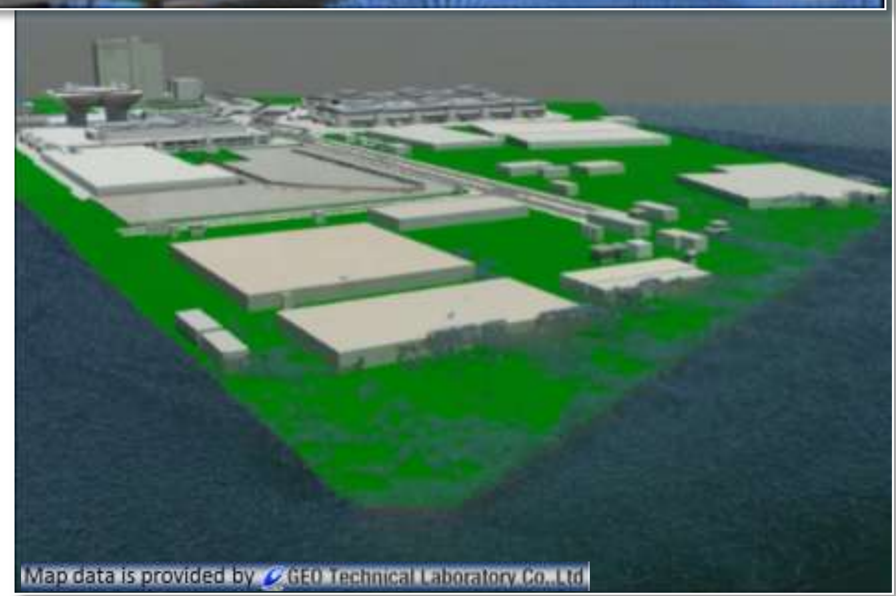
FUJITSU

2) Building response simulation: predict effects on infra.

Impact by the Tsunami water flow

- Improve the accuracy on estimation of building damages by Tsunami
- Improve the accuracy on estimation of river/canal overflow by Tsunami flowing upstream direction
- Balance between robust buildings vs. water flow direction & speed
- Help the city/town infrastructure planning to secure the evaluation paths (road, bridges, etc.)
- Helps to better design the evacuation facilities and the way to get there

2st Step: Solid 3D simulation of the water flow to predict damages

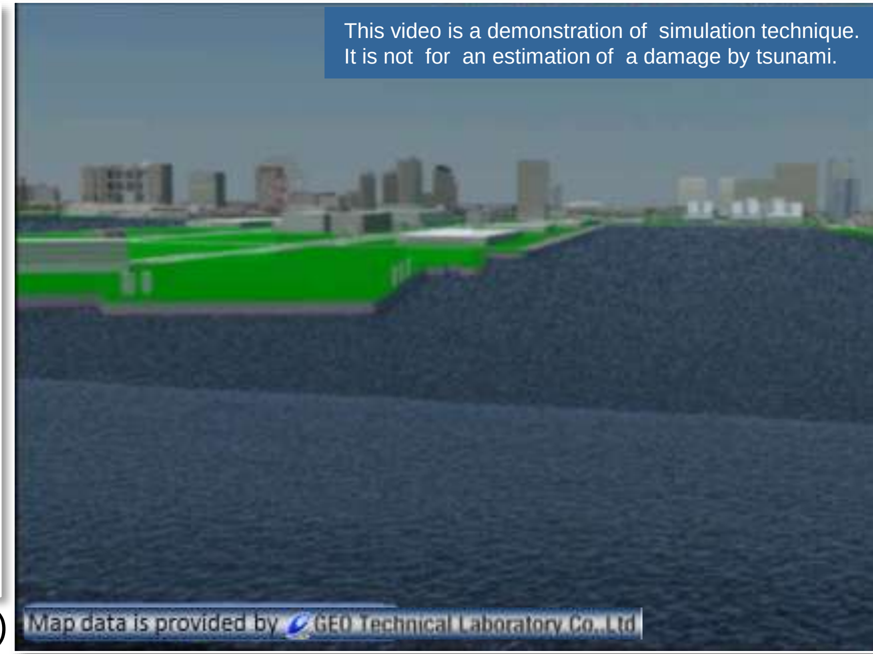


Big Data use: Simulation for disaster prevention

FUJITSU

2) Building response simulation: predict effects on infra.

Accurate 3D replication of invading wave from offshore to shallow sea



(left: Yokohama, right: Tokyo)

- 3D simulation for wide-area using K computer (京)
- Smoothed-particle hydrodynamic simulation w/ 400 million particles
- The potential use of these research results are:
  - to design levees & evacuation shelters
  - to develop guidelines for hazard maps and evacuation routes

More effective disaster prevention planning through 3D simulation

Big Value Data!

"The Rock"

The Rock

...

Where is the Life
we have lost in living?

Where is the wisdom
we have lost in knowledge?

Where is the knowledge
we have lost in information?

...

T.S. Eliot

Thomas Stearns Eliot

(1888 – 1965)

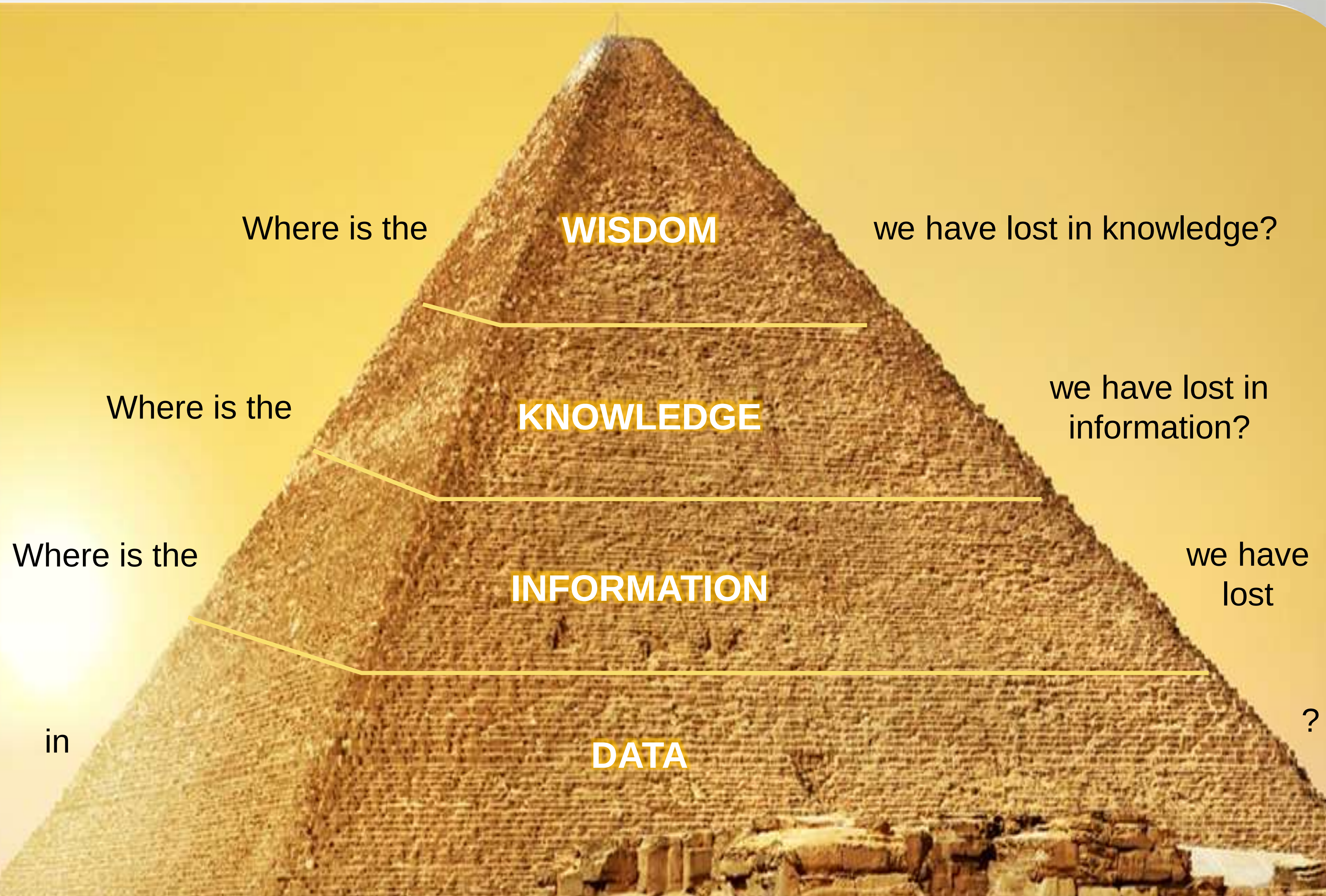
publisher, playwright, literary

born American
naturalized British subject in 1927
Nobel prize in literature in 1948

Source: Wikipedia

DIKW-Hierarchy

FUJITSU



FUJITSU

shaping tomorrow with you