Big Data and Disaster Management
A Report from the JST/NSF Joint Workshop

Executive Summary

In May 2013, researchers and grant agency representatives from the USA and Japan participated in the *National Science Foundation (NSF) and Japan Science and Technology Agency (JST) Joint Workshop on Examining and Prioritizing Collaborative Research Opportunities in Big Data and Disaster Management Research* (Arlington, VA, USA). The workshop participants were divided into three breakout groups and discussed both the potential benefits of big data for disaster management as well as the big data research challenges arising from disaster management. This report summarizes the discussions during and after the workshop.

It was agreed from the beginning that Disaster Management is an important global problem. Disasters affect every country on Earth and effective disaster management is a global challenge. This is particularly the case of large-scale disasters that affect many countries (e.g., the 2004 Indian Ocean earthquake and tsunami) and multi-hazards such as the Tohoku Earthquake and landslides. Tools that can be used by many countries will have significant broad impact in helping the world population as well as many government agencies and non-governmental organizations. At the same time, we recognize the importance of cultural, social, and linguistic differences among countries in emergency response and recovery, which will have impact on the big data used for disaster management.

Big Data can help in all four phases of disaster management: prevention, preparedness, response, and recovery. Two major sources of big data, dedicated sensor networks (e.g., earthquake detection using seismometers) and multi-purpose sensor networks (e.g., social media such as Twitter using smartphones), both have demonstrated their usefulness in disasters such as the Tohoku Earthquake. However, significant big data research challenges arise because of disaster management requirements for quality of service (e.g., highly available real-time response) and quality of information (e.g., reliable communications on resource availability for the victim). Two of the major big data challenges are: Variety (integration of many data sources including dedicated sensors and multi-purpose sensors), and Veracity (filtering of Big Noise in Big Data to achieve high quality information).

To fulfill the potential benefits of applying big data to disaster management, we need to bring together the best minds from around the world. From the disaster management view, we need the technology push from big data researchers to tackle the challenges mentioned above (e.g., Big Noise) so big data tools can effectively address disaster management issues. From the big data view, we need the application pull of disaster management researchers to apply big data techniques and tools to solve real world problems. The international collaboration between Japan and USA will help both the disaster management and big data research community take the step forward towards better disaster management using big data.

Authors/contributors of this report: workshop participants and additional contributors (see detailed list in Section 1.3). Editors: C. Pu and M. Kitsuregawa.

Academic citation: Technical Report No. GIT-CERCS-13-09; Georgia Institute of Technology, CERCS.

Table of Contents

Big Data and Disaster Management
A Report from the JST/NSF Joint Workshop

1. Big Data Opportunities for Disaster Management

1.1 JST/NSF Joint Workshop

In May 2013, researchers and grant agency representatives from the USA and Japan participated in the *National Science Foundation (NSF) and Japan Science and Technology Agency (JST) Joint Workshop on Examining and Prioritizing Collaborative Research Opportunities in Big Data and Disaster Management Research* (Arlington, VA, USA). This report summarizes the discussions during and after the workshop.

The workshop was divided into three breakout groups according to the topics of discussion identified by the workshop organization committee, without excluding other relevant topics. The topics were:

1. Harnessing the big data associated with disasters and disaster prevention to advance analytic, modeling, and computational capabilities.
2. Improving the resilience and responsiveness of information technology to enable real time data sensing, visualization, analysis, experimentation and prediction, all critical for time-sensitive decision making.
3. Advancing fundamental knowledge and innovation for resilient and sustainable civil infrastructure and distributed infrastructure networks.

1.2 Motivation and Background

What We Mean by Disaster. This report will use the U.S. Federal Emergency Management Agency (FEMA) definition of a disaster as "an event that requires resources beyond the capability of a community and requires a multiple-agency response" [14]. Disasters may be due to meteorological events such as avalanches, floods, fires, heat waves, hurricanes, thunderstorms, tornadoes, and winter storms, geological events such as earthquakes, landslides, tsunamis, and volcanoes, or man-made such as building, bridge, or tunnel collapses, chemical or radiological waste spills, dam failures, nuclear power plant accidents, and train wrecks.

Phases of Disaster. The Public often uses the terms “disaster response" or "disaster management" for what are actually four phases of a disaster: prevention, preparation, response, and recovery. The four phases are not independent and sequential; indeed response and recovery operations start instantaneously, and sub-populations have different long-term recovery needs. Life-saving response operations rarely last beyond 10 days, though local officials may be reluctant to declare the response phase over. Recovery operations can go on for months and are generally marked by when roads, schools, and hospitals are re-opened. Economic and public health recovery often takes years beyond that. The ultimate success of response and recovery operations are influenced by the data collected during the preparedness and prevention phases.

Importance of Disasters. The most recent World Disasters Report [17] found that, between 2000 and 2009, a total of over 7,000 disasters resulted in over 1 million people casualties worldwide, affected another 2.5 million directly, and yielded a loss of just under $1 Trillion. The number of disasters each year during this 10-year period remained relatively constant, around 700, suggesting that despite the national focus on “super storms” and wide area events there are many lesser reported events. More significantly, the impact of the disaster is a function of increasing numbers of people living in concentrated urban environments, and because of the consequences of damage to urban structures where people live and work (e.g., office and apartment buildings) and that serve to mitigate disasters (e.g., hospitals, transportation and other critical infrastructure).

New Technologies for Disaster Management. Disaster response is considered a societal Grand Challenge by the President's Council of Advisors on Science and Technology (PCAST) [2][3] and the Computing Research Association [3]. Other studies by the PCAST [2][4], the National Science and Technology Council (NSTC) [5], and the National Academies [6]-[13] have concluded that transformative improvements in disaster management will not be due to advances in component technologies, manpower, or physical resources by themselves but rather from systems approaches to developing the information for timely decision making and synchronizing the flow of this information to the shifting demands of disasters and resources.. History has shown that advances in wireless networks, unmanned systems, embedded sensors, pattern recognition, surface reconstruction, data fusion, and scheduling algorithms have not necessarily resulted in usable information or better decisions, rather they often have created an unmanageable data avalanche.

A 2012 NSF/CCC study, CRICIS: Critical Real-Time Computing and Information Systems [1], describes a computational perspective to disasters that details how computing can address all four phases of disasters by fostering a systems approach to enabling timely and effective decision making by all stakeholders. The nation, and the world, stands at an inflection point today. Disasters have long impacted society, resulting in mass casualties and huge financial tolls. However, key advances in computing over the past several decades—smartphones, worldwide mobile Internet access, and the emergence of social media have all contributed to changes in how society responds to disaster and mass emergency [36][37]. These advances, in turn, contribute to the growing amount of data that is available in all four phases of a disaster and to a need for research that produces information systems that can take advantage of this data. One area of research that investigates these issues is known as crisis informatics [46], but CRICIS calls for a new fundamental research thrust in socio-technical systems that enable decision-making for extreme scales under extreme conditions.

Big Data Opportunity. It is easy to "predict" the inexorably increasing availability of big data due to the continued technology evolution. Specifically, environmental sensor data and social media data will become increasingly available for disaster management due to the advances of many kinds of capable sensors. On the environmental side, the high end sensors operated by public and private agencies are being rapidly supplemented with web-enabled low-cost custom or generic (i.e., smartphone) sensors operated by citizens and communities. On the social media side, the growth of various kinds of social media has been phenomenal. A major challenge is to build big data tools and platforms to help create the next generation of disaster management applications that can make use of this information to help society respond to these events.

1.3 Workshop and Report Organization

Workshop Organization Committee:

- USA: Calton Pu (Georgia Tech),
- Japan: Geng Tu (JST), Kazuo Iwano (JST), Masaru Kitsuregawa (NII and Univ. Tokyo).

Workshop Participants and contributors to the report:

- USA: Ken Anderson, Anish Arora, Farokh Bastani, James Caverlee, Byung H. Kang, Andrea Kavanaugh, Eva K. Lee, Andrea Matsunaga, Klara Nahrstedt, Linh Phan, Calton Pu, Raj Rajkumar, Yasuaki Sakamoto, Mauricio Tsugawa, Nigel Waters.
- Japan: Shin Aoi, Koichi Fujinuma, Takahiro Hara, Teruo Higashino, Kazuo Iwano, Norimitsu Kamimori, Hiroshi Maruyama, Tsuyoshi Motegi, Akira Mukaida, Jin Nakazawa, Takashi Ohama, Mitsuhiko Oi, Hiroshi Shigeno, Geng Tu, Keiji Suzuki, Yuzuru Tanaka, Rin-ichiro Taniguchi, Yoshito Tobe, Hideyuki Tokuda, Masashi Toyoda.

Significant additional contributors to report:

- Jose Fortes, Masaru Kitsuregawa, Robin Murphy

Editors of the report:

- Calton Pu and Masaru Kitsuregawa

The rest of the report is organized as follows.

- Section 2 outlines the research challenges and benefits in the disaster management area when big data approaches are applied.
- Section 3 describes the research and technical challenges for big data researchers in order to address disaster management needs.
- Section 4 summarizes the needs for and benefits from international collaboration between the USA and Japan in the disaster management and big data areas.

2. Disaster Management Can Benefit from Big Data

2.1 A Concrete Scenario: 3-11 Tohoku Earthquake

Natural disasters such as the 2011 Tohoku Earthquake can cause multiple connected natural disasters, such as both an earthquakes and tsunamis. The 2011 Tohoku earthquake occurred on March 11, 2011, at 14:46 Japanese Standard Time. The earthquake had a magnitude of 9.0 with the epicenter about 70 km east of the Ohika peninsula of the Tohoku region and a depth of about 30 km. The main earthquake was preceded by several large foreshocks, and followed by hundreds of aftershocks, including three aftershocks with magnitudes above 7.

The main earthquake lifted a 180 km wide seabed six to eight meters, causing large tsunami waves. These waves originated 60 km from the east coast of the Tohoku region. They reached heights up to 40 meters in some areas and travelled up to 10 km inland. About one hour after the earthquake a tsunami flooded the Sendai Airport, located near the coast in the Miyagi prefecture (part of the Tohoku region). Buildings were flooded and cars and planes were swept away. The damages from the tsunamis were much more destructive than the earthquake itself. Entire towns were destroyed or swept away by the tsunamis, including Minamisanriku where 9,500 people went missing. On 12 September 2012, a Japanese National Police Agency report confirmed 15,883 deaths, 6,145 injured, and 2,671 people missing across twenty prefectures, as well as 129,225 buildings totally collapsed, with a further 254,204 buildings 'half collapsed', and another 691,766 buildings partially damaged.

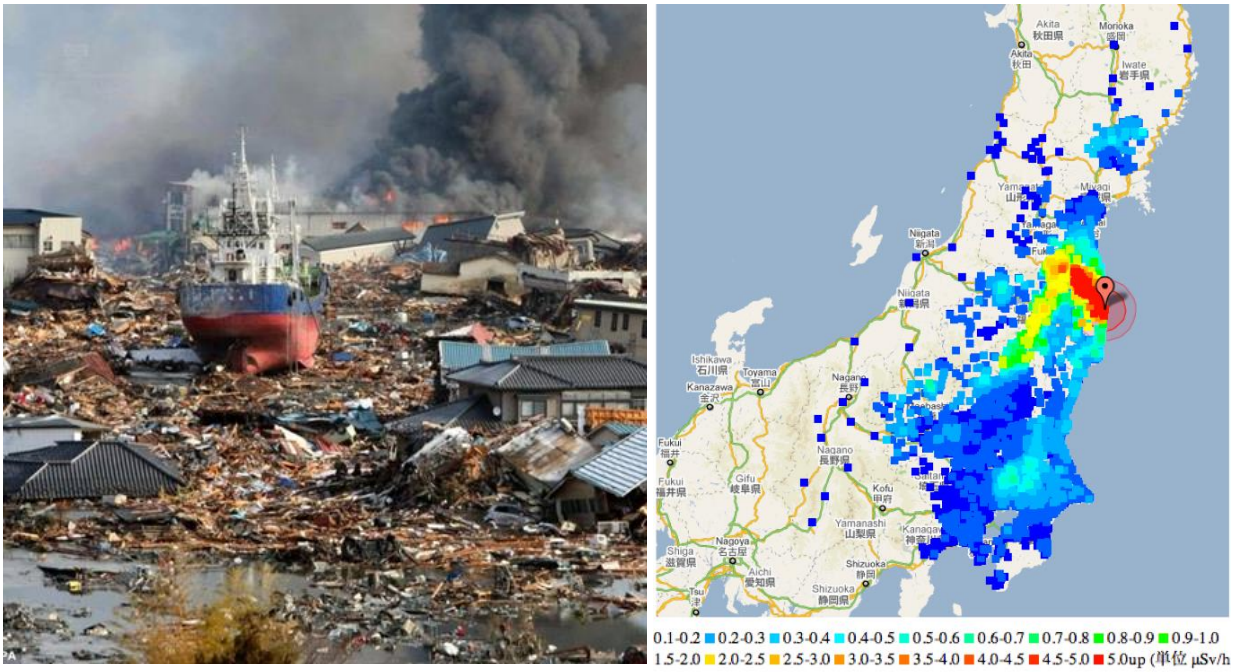

Figure 1 Illustrative multi-hazard example: the 2011 Tohoku Earthquake, tsunami (left), and Fukushima nuclear disaster (right)

Several nuclear power plants in the area, with a total eleven reactors, were automatically shut down following the earthquake. Cooling is needed to remove decay heat after a Generation II reactor has been shut down, and to maintain spent fuel pools. Unfortunately, at Fukushima Daiichi Nuclear Power Plant, tsunami waves overtopped seawalls and destroyed diesel backup power systems, leading to the level 7 meltdowns of three reactors. It was reported that radioactive iodine was detected in the tap water in Fukushima, Tochigi, Gunma, Tokyo, Chiba, Saitama, and Niigata, and radioactive cesium in the tap water in Fukushima, Tochigi and Gunma. Radioactive cesium, iodine, and strontium were also detected in the soil in some places in Fukushima. At the Fukushima Daiichi power plant, the cooling water to remove decay heat became highly contaminated. As of June 2013, the leakage of highly radioactive water from some of these tanks and high levels of radiation continue to be reported [56].

2.2 Challenges in Disaster Management during Multi-Hazards

The Tohoku Earthquake outlined above illustrates the multitude of challenges faced by the,public, first responders, government agencies and NGOs (non-governmental agencies such as Red Cross) involved in disaster management. First, the technical challenges in preventing, detecting, and responding to each kind of disaster (e.g., sensor network maintenance and data integration) are significant and diverse. This is the case of earthquakes, tsunamis, and nuclear disasters that happened in the Tohoku Earthquake. Second, the dependencies among the individual disasters in a multi-hazard are sometimes predictable (e.g., tsunamis caused by underwater earthquakes), but not always (e.g., an unexpectedly large tsunami flooding of underground diesel generators needed for recently-shutdown nuclear reactor cooling). Third, resources needed for responding to one disaster may strain the resources needed for others, when multi-hazards happen. This happened in Tohoku and other multi-hazards, particularly when the civil and information infrastructures have been adversely affected. The flexible allocation of limited resources under rapidly changing constraints and often conflicting optimization objectives during multi-hazards remains a difficult technical, managerial, economic, and social problem.

Although we use the Tohoku Earthquake as a concrete running example in this report, we want to mention a few other examples of multi-hazards to emphasize the general nature of our discussion. Another example of natural multi-hazard is landslides. As another example of multi-hazard event, landslides can be caused by rainstorms or earthquakes. To predict a landslide, land survey data may be sufficient for a static geomorphological model, but landslides often change the topography of the terrain, making further predictions of newly unstable areas difficult. The use of satellite images to detect such changes has proven very useful. Another obvious example is the improvement in climate data and weather forecasting using large data sets. Accurate forecasting of heavy rains for landslides and warm temperatures for alpine lakes can improve the performance of hydrological models of such lakes. Additional sources of environmental sensing data include wireless sensor networks for detecting strain signatures in surface rocks (using strain gauges) or orientation changes (using tiltmeters) and for localizing slip surfaces (using strain gauge, geophone, moisture, reflectometer sensor columns).

2.3 Where Big Data Can Help

2.3.1 Evolution of Sensor Networks towards Big Data

The detection of earthquakes is a classic success story of dedicated and specialized sensor networks. Sophisticated seismometers, strategically located to optimize for triangulation and precise determination of earthquake epicenters, are maintained by agencies such as USGS. Data from these seismometers are continuously fed into geological models that can find earthquake epicenters precisely within seconds of the detection of earthquake vibrations. This kind of dedicated sensor networks is characterized by their careful design, implementation, and high quality data output. Unfortunately, these sensor networks also require high levels of capital expenditure for construction and operational expenditures for maintenance. Therefore, their functionality and coverage are often limited by funding availability. This is the case despite the technological evolution of micro-sensors (a.k.a. *motes*), since the physical parts of such sensor networks (e.g., power, communications, all-weather secure enclosure) have yet to be miniaturized.

An interesting twist in the evolution of motes is their prevalence today, not as components in specialized sensor networks, but in multi-purpose devices such as smartphones. Smartphones such as iPhone and Android have surpassed 1 billion devices worldwide in 2012 and continue to grow. These smartphone typically contain more than 20 micro-sensors such as accelerometers, gyroscope, magnetometer, light

sensor, and GPS. In retrospect, this evolution is logical since it allows the micro-sensors to share the "heavy" components of motes: battery power, cellular communications, and enclosure. In some advanced applications such as traffic monitoring in intelligent transportation systems (ITS), smartphone data have enabled widely used traffic monitoring applications (e.g., Google Maps) at very low or zero cost in many metropolitan areas around the world. In traffic monitoring, smartphones have sometimes supplanted classic dedicated ITS monitoring networks and in some places provided traffic information otherwise unavailable due to the lack of dedicated ITS sensors.

Despite its success in some applications (e.g., ITS), the rise of multi-purpose sensors such as smartphones did not replace completely specialized sensor networks in other applications such as earthquake detection. One of the main reasons is the data quality issue, due to the unpredictability of sensor location and lack of control over each device. For example, unlike traffic monitoring of congested highways, an earthquake in the middle of ocean cannot rely on smartphone data for its immediate detection. At the same time, the quantity and variety of data from multi-purpose sensors are leading to unprecedented opportunities for disaster management. In the next sections, we outline some illustrative examples of social media helping in disaster management and benefits of sensor data integration.

In addition to the evolution towards multi-purpose sensor networks, the rich information from dedicated sensor networks is yielding new information through data mining of accumulated historical data. The damage from a disaster at each target area varies depending on the distance from the disaster and the feature of the target area (e.g. coast/hill, riverside/inland, soft/hard ground, and so on). In order to precisely evaluate damages from a disaster at large areas, we need several types of domain-specific real-time sensor maps (geographical maps) representing sensing data from domain-specific sensors at several sensing points. Seismic intensity maps, wind velocity maps, Tsunami estimation maps, radiation distribution maps and rainfall maps are typical such sensing maps. At the same time, in order to evaluate how severe disaster has occurred, it is desirable that we can store several types of historical sensor maps and compare the current sensor map with those historical ones. Then, we can estimate the scale of the current disaster more precisely. Sensor maps based on simulation are also useful if we only have few historical sensing maps. Since the sizes of archives of such historical sensor maps are very large, we need efficient mining mechanisms. Also, simultaneous usage of multiple sensing data can improve the accuracy of estimation about how large disaster has occurred. Thus, new types of efficient mining and learning mechanisms from multiple types of sensing data should be studied.

2.3.2 Illustrative Social Media Big Data Scenarios

The use of social media in emergencies has been reported anecdotally for various disasters since real-time services such as Twitter have been introduced. However, studies of social media data actually used during disasters have appeared only recently (see, for instance, [38]-[43]). We will use the Tohoku Earthquake as illustrative example. From their data set (about 2 million tweets spanning several weeks around 3/11/2011, the earthquake date), the first tweets from the Tokyo area that reported the Tohoku earthquake happened within 2 minutes of the onset of tremor from the offshore epicenter. The tweets rapidly reported on the earthquake, the following tsunami, and then the Fukushima nuclear disaster.

Another example of social media data that has become useful for studying disasters is location-based services as shown in Figure 2. Foursquare [34] is a social network where users explicitly "check-in" to show where they are and then find other users who are nearby. The location information provided by the check-ins can be aggregated to show where popular attractions are. Figure 2 shows a comparison of Foursquare check-ins in lower Manhattan (New York City) before the Sandy hurricane (10/27) and after Sandy (10/31). The comparison shows the impact of electricity loss, which resulted in many businesses being closed in lower Manhattan and thus discouraging visitors. In contrast, the Midtown area shows similar activity levels for both dates.

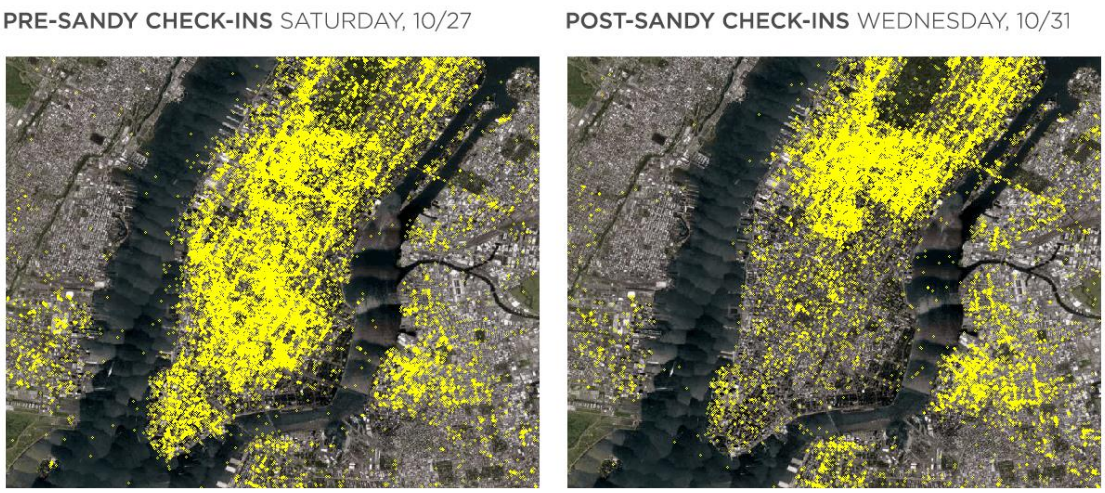

Figure 2 Impact of Disaster as Shown by Social Media Data (Hurricane Sandy, 2012)

2.3.3 Integration of Dedicated Sensor and Multi-Purpose Sensor Big Data

During the Tohoku Earthquake, effective disaster management relied on both dedicated sensor networks (the seismometers that detected accurately the earthquake) and multi-purpose sensor networks (e.g., social networks for information sharing). The Tohoku Earthquake experience is indicative of potential benefits of integrating big data from both dedicated and multi-purpose sensors. On the environmental side, dedicated sensors can help the system and users detect changes of the physical environment even in such a case in which most people are unconscious about any of them (e.g., increasing radioactivity). On the human side, multi-purpose sensors such as smartphones can monitor people's mobility and enable crowdsourcing of status monitoring such as subway station openings. These information services will provide quality of service and quality of information as outlined in Section 3.1.

A concrete example of social media functioning as multi-purpose sensor is Twitter, which was quite useful for reporting specific events, there are several serious challenges in using social media information in such situations. For example, it is not always obvious which twitter account is more accurate and trustworthy, although initial research is developing a set of heuristics that can be used to identify these accounts [39][41][44][45]. As a concrete example, the Tokyo subway was one of the first public transportation systems that returned to service in the evening of 3/11, only a few hours after the earthquake. The subway service was resumed gradually, starting from a few lines and running for a few stations. There were some missteps as some stations were temporarily closed due to over-crowding. Due to the dynamic nature of events unfolding, sometimes an event (e.g., opening and closure of a Tokyo metro station after the Tohoku Earthquake) may have been superseded when the tweet about it appeared.

In general, there is a growing consensus that each social network offers useful but insufficient information for accurate decision making by organization representatives (e.g., first responders) and by the public. Important sources of other social media information include: Foursquare (explicit location information), Flickr (photos and images), YouTube (videos), social network sites such as Facebook and MySpace. Beyond social media, there are many information sources (usually accessible through streaming interfaces such as RSS) that report on disasters. For example, news agencies and web sites (e.g., CNN) may be disseminating useful information. Another example is the web sites and sources from the emergency response organizations (e.g., FEMA, CDC, and Red Cross) that contain official information. This is useful, since unlike earthquakes, some kinds of information (e.g., epidemic propagation of some virus and bacteria) cannot be easily determined by crowdsourcing alone.

More generally, in areas where insufficient data for predicting and detecting landslides is available currently, new data can fill in important gaps. For example, detailed satellite images of remote areas may support more accurate models of soil type based on vegetation type and density. Similarly, vibration/seismic measurements from a variety of inexpensive sensors may allow more accurate geological

models of areas not covered by past surveys. In summary, useful sensor data sources for landslide prediction include: satellite images, meteorological data (for storm-induced landslides), seismographic data (for earthquake-induced landslides), and photos/videos from social networks such as Instagram and YouTube.

Among the many examples of successful efforts in aiding disaster management, we mention one illustrative example in integrating data from various sources for crisis management: Google Crisis Response [53]. Created after Hurricane Katrina of 2005, Google Crisis Response team provides tools for aiding first responders, citizens, and various disaster management organizations. Examples of tools include Google Person Finder [54] to connect people with family and friends, and Google Crisis Map [55] with authoritative and crowd-sourced geographic and useful information such as storm path, shelter locations, and power outages.

2.4 Disaster Management Challenges for Big Data

There are many challenging research issues that arise when we apply big data approaches to aid disaster management. In this section, we will outline some illustrative challenges from disaster management point of view, organized by the four phases of disaster: prevention and preparedness (Section 2.4.1), response (Section 2.4.2), and recovery (Section 2.4.3). In Section 3, we will discuss the research and technical challenges from the big data point of view, as we apply big data technologies to address disaster management problems.

2.4.1 Disaster Prevention and Preparedness Using Big Data

The first example of disaster management research challenges before a disaster has actually struck is providing accurate information on developing disasters, particularly the prediction of what will happen, to the public in a way that they will understand, accept, and act on. This is a serious challenge for multi-hazards due to the uncertainties in evolving situations and the chaining of various disasters. Multi-hazards such as the Tohoku Earthquake illustrate well the issue of dependencies among the disasters involved. An important challenge is the modeling of such dependencies among different types of hazards occurring one after another.in a multi-hazard scenario. Such a model will enable us to predict and to prevent or mitigate the following hazards. This kind of dependency modeling requires a large collection of big data that will help us improve the accuracy of each model. More precise and higher resolution seabed data and coast line data will lead to more accurate tsunami modeling and prediction. Bigger social media data with larger coverage of people and target regions may lead to more accurate human response modeling and prediction. These data may also bring in the issues with trust, uncertainty, and relevance of data. Such dependency modeling will allow us to use dependency knowledge to connect different models of different hazards and to seek out new data that are appropriate for preparing for the possibly cascading hazards to mitigate them. Such approach will ultimately achieve better prediction on the cascading hazards, preparedness for them, mitigation of them, and smarter response to them. This also raises issues with intelligent dissemination of results from new Big Data models so that both multiple authoritative organizations and people in the public including both victims and volunteers can share these results and collaborate with them.

The second example is the translation from disaster management specifications into actual implementations under a variety of environmental conditions. The current situation of building specialized tools for each disaster is both expensive and non-scalable due to the lack of software reuse and data integration. One possible approach is the development of a common specification language for government agencies (cities, counties, regions, countries) and NGOs to provide assurance specifications about disaster preparedness. These statements would relate to the types of infrastructure being discussed in this report that would include not just information systems but cyber-physical systems that collect and aggregate information from a wide range of heterogeneous information sources and sensors. An example of the types of assurances that could be made might be “The city of Tokyo is prepared for a magnitude 7.0 earthquake because of [list of government agencies, sensor networks, and operational crisis informatics systems].” This type of language would be similar to the assurance cases that are already used in Europe to make statements about the safety of infrastructure, such as mass transit systems, buildings, and the like.

The third example of interesting research issues is related to the architecture of large-scale crisis informatics systems, as well as for information systems used during non-emergencies. Such an architecture

should be engineered to be resilient in the face of a large-scale disruption that can be inflicted by natural disasters and other multi-hazard events. Architectural resilience can be achieved in many ways including having back-up redundant systems that kick in when primary systems go down to employing a wide variety of techniques to ensure that at least some measure of system functionality is available even if most of the system is down. For instance, cloud-based distributed storage systems employ distributed computing techniques to provide high availability through appropriate data replication. These characteristics help to ensure that important data is always available for mission-critical applications. Building on such components, research is needed to achieve such resilience for large-scale crisis informatics systems. Such systems might, in the future, be designed to have multiple modes of operation, able to switch gracefully from a "normal" mode of operation to a "disaster mode." Such a switch may involve both physical and digital reconfigurations to the system but help it enter a state in which services are maintained but at a reduced level of service.

The fourth example of interesting research issues is looking at how the various IT systems run by federal, regional, and local government organizations (and NGOs) can be integrated such that information critical to disaster response can be aggregated into actionable information. In a sense, the myriad of such systems can be seen as the "long tail of big data": lots of little data sources that need to be aggregated in order to coordinate a response. Work in this area would need to identify the issues (both technical and social) that exist that prevent information sharing and aggregation currently and then identify promising techniques and tools for enabling integration when it is needed, especially during times of mass emergency. Enabling consensus (or supporting a variety of access policies) at various societal levels to allow both people and organizations to gain access to crisis information that they trust, so that they can prepare and respond to crises, is an important part of this challenge. The technical aspects of this issue will be discussed further in Section 3.4.

2.4.2 Disaster Response Using Big Data

Disasters often result in limited availability of resources in the disaster area. In the initial stage of large disasters, which usually corresponds to the first 48 to 72 hours after a disaster, the evacuation and public safety become the first priority. The infrastructure should support huge time-constrained communication demands for disaster alerting, delivering evaluation information, rescue supporting, safety confirmations, prevention and mitigation of the future disasters, etc., while keeping the functionalities for big data collection and analysis for better understanding the damage situation and decision making. Despite of these demands, the network resources in this stage is limited and even might decrease as time goes by due to the damages from the disaster (In the case of Tohoku Earthquake, it has been reported that the number of outages of communication facilities such as access lines and cellular base stations increased during the first 24 to 48 hours after the strike).

Real-time big data analysis can substantially enhance various disaster response aspects. First, it can help emergency response personnel to identify areas that need the most urgent attention. This could be areas where there are several people or critical resources. It could also be areas where there may be triggers for other hazards. Examples of such situations include a tsunami approaching a nuclear power plant or a forest fire approaching a factory where some hazardous chemicals are stored. Second, real-time monitoring and situation analysis can assist emergency response personnel in coordinating their actions to optimally handle a disaster situation. This also includes guidance to the public in taking the best routes to move away from a disaster in order to prevent congestions or causing people to move by mistake to a more hazardous situation. Third, big data analysis from prior incidents can help identify the most effective response methods for various situations and enable the development and deployment of assistive infrastructures for effectively responding to future disasters. For example, it may be found that certain types of disasters make it difficult for emergency response teams to physically go to certain locations. In such cases, robotic infrastructures for tele-health assistance and monitoring can enable medical personnel to remotely assist people impacted by a hazard who need emergency medical attention. Similarly, it can enable robotic units to provide emergency cover and shelter for children at schools that are in the path of an approaching severe tornado.

There are existing efforts towards achieving some of the above requirements; for instance, several projects within the NSF Future Internet Architecture program have laid a foundation for achieving high availability and reconfigurability within the network, and existing research on real-time virtualization can serve

as a building block toward a real-time cloud. New research on highly secure, reliable and real-time infrastructure (not only in the network but also in the clouds) that leverages on these existing results is especially needed for effective disaster response and to enable the use of big data technology. Further, methods for continuous adaptive optimization of such an infrastructure are crucial to adapt to the constant evolving purpose of the infrastructure, such as due to changing rescue missions during a disaster, as well as unpredictable system failures and environmental events. In addition, it is important to ensure that simple deployments of such an infrastructure during actual disaster scenarios are feasible, since they should be deployed as needed.

To achieve the goal of optimized resource allocation during disaster, it is important to move the public away from dangerous grounds. One of the major challenges is to find out what the public is actually doing. This is a non-trivial problem, since the public is a loosely coordinated process with many independent agents, acting on disparate information sources and different decision making models. The tracking of people's location and behavior is important for the appropriate allocation of resources to aid the victims in the disaster area and reduce the cost of maintaining the affected population outside the disaster area. Social networks (through crowdsourcing) have the potential for improving real-time communications on displaced populations, but setting up such *ad hoc* networks has been a slow manual process. When a disaster occurs, real-time and continuous monitoring of people's location and behavior is very useful. In order to improve logistics in the disaster areas, the responders need to understand where victims exists, how many victims exist at each area, what troubles they have and what disaster medicine they require precisely.

For improving decision making during disaster response, an interesting research topic is looking at the optimal speed/accuracy tradeoff for a given disaster situation. This line of research needs to identify the appropriate time scale for responding to a particular event, and design and evaluate tools with the time scale in mind. One approach might be to build adaptive tools that decision makers can adjust the speed to fit their needs. Such tools could involve crowds because big data generated during responses to disaster are often unstructured. Although machines can process big data quickly, they are not adept at making sense of unstructured data and recognizing unusual patterns inherent in disaster data. In contrast, crowds are better at codifying unstructured information and recognizing unusual patterns than machines. However crowds cannot process big data as fast as machines.

Another fruitful research direction is to systematically study how to integrate crowds and machines to help automate decision-making during disaster response. The combined decision aid should be able to make better sense of the vast amount of unstructured information in big data than the decision aid that relies on only machines or only crowds, without sacrificing speed. This research will involve designing and evaluating different methods of combining the output of crowds and the output of machines. For example, machines might be able to filter out noise and irrelevant data, and crowds can prioritize the rest of data with respect to usefulness. Some example questions are: How can we break down the task of identifying useful information from big data (e.g., credible information, actionable information, and so on) and distribute the sub-tasks to crowds and machines?; How can we train machines using the output of crowds (e.g., building prior distributions using the output of crowds in Bayesian techniques)?; How should we summarize the output of the combined decision aid and display it to decision-makers? How can we assemble the crowds and make sure that they complete the task in a timely manner?

Evaluating the decision aids for disaster response is a challenge. One way is to evaluate the decision tools in a simulated disaster. Although the results from simulation will provide helpful information, it is unclear the extent to which the results will generalize to the real disaster, especially because in the end people make the decisions. The observation that people behave differently in training and in real disasters suggest that testing of decision aids during real disasters would provide valuable information that testing in simulated environments alone cannot provide. One possibility is field experiments, which companies have started to use recently. For example, using A/B testing, we will randomly assign one decision tool to one set of people and another tool to another set of people during real disaster response. We will take the tool that results in better outcome, and test this tool against a third one. We can iterate this process quickly to find the best tool without significant negative consequences. Quasi-experimental, matched-market studies by assigning tools to different agencies during real disaster response could also be useful to identify a tool that is appropriate for a particular agency. For such filed experiments to be successful, extensive planning is necessary. Research in this area needs to have the tools and the de-

sign of the study ready for the next disaster of interest that we do not know when it will come. While there is some risk in such research, if successful, the potential benefit to the society is greater than the risk.

2.4.3 Disaster Recovery Using Big Data

After the initial stage, the recovery activities will gradually start and the demands to the infrastructure will also change. The infrastructure should support a wide variety of information sharing and big data analysis for recovery activities such as safety confirmation, volunteer coordination, provision of relief supply, logistics, etc. In this stage, the network resources will increase since the damaged network facilities will be gradually restored. With respect to infrastructure networks in particular, continuous and adaptive optimization and even repurposing of network and computation resources for a disaster area is therefore indispensable for real-time big data exchange and processing in the disaster response. Both demands and usable resources for the communication and computation infrastructure might change from hour to hour. Thus, mechanisms and methods for continuous adaptation to the change of demands with the limited resources should be a key issue for the big data infrastructure.

As a concrete example, after the Tohoku Earthquake, many telephones and mobile phones could not be used because of heavy network traffic congestion, however, most of emails, twitter, facebook and other SNS were used although some delay occurred. From those big data, we can estimate human behavior in disaster situations. When Tsunami occurred, several people were isolated at high buildings in flooded areas. Emails and twitter information from victims in such areas were used for fast evacuation planning of the responders. However, since we have not developed real-time processing mechanisms for finding such victims, it takes much time for evacuation. It is a challenging research theme for developing such real-time recognition of disaster victims.

Many of the tools developed for disaster response (Section 2.4.2) can be used effectively in disaster recovery. For example, simulation of disaster and response can be extended in the recovery phase to evaluate the various alternatives for recovery (e.g., whether to repair a building or to demolish it and build a new one). Also, the public already using social media (e.g., during disaster response) to gain information during response is likely to continue using the same communications channels during the disaster recovery phase.

3. Big Data Research Challenges from Disaster Management

3.1 Quality of Service and Quality of Information in Big Data

For using big data analysis to achieve effective disaster management, the underlying infrastructure must provide high quality of service (QoS). While the QoS requirements may differ for various disaster situations, we outline some examples here as illustration.

- Given the urgency of the response actions when dealing with most disasters, it is imperative for the infrastructure to provide real-time performance. This includes real-time data analysis to accurately predict the impact of an approaching hazard as well as the best way of effectively responding to the disaster. It also includes real-time communication to ensure that correct data are gathered about the environment, such as the location of people who need help, the best routes for going to a disaster site and for helping people move away from disaster sites. Real-time communication is also needed to ensure that various emergency response teams can coordinate their actions in optimally responding to a disaster.
- Given the criticality of disaster response situations, it is also important to ensure that the service will be highly reliable and available in spite of the adverse environmental conditions during such situations, including physical damages, power outages, floods, etc. Hence, the big data storage, analysis, and transmission services must be able to operate in spite of such adverse conditions. Redundancy alone is not adequate since one type of hazard may impact all the redundant units. Hence, this requires the use of diversity in addition to redundancy to ensure high reliability and availability. For example, computing and sensor resources can be deployed at different geographical locations and different communication methods can be used to ensure continuous access to the data.

- Given the evolving nature of disasters and disaster response strategies, it is also important to ensure that the big data supporting infrastructure is sufficiently maintainable. This includes methods of ensuring that the infrastructure can be easily upgraded and also to be able to rapidly repair or replace damaged units.
- Given the sensitivity of some types of data that can substantially help disaster response, such as the location of people and their medical conditions, it is also important to ensure that the service meets high levels of security. This includes high levels of privacy and confidentiality as well as assurance that the information used to guide the response to a disaster are correct and not corrupted.
- For tele-operation, tele-health, and other remote actions, it is also important to ensure high levels of cybersecurity. In particular, the infrastructure must ensure that only authorized emergency response personnel can control remote units.

Cloud computing platform resources can be leveraged to support big data storage and analysis. Multiple cloud computing platforms at geographically diverse locations can be used to tolerate different hazards. This natural combination of redundancy and diversity can be leveraged to achieve high performance real-time big data analysis. This diversification can also be used to achieve highly secure storage and computing by splitting confidential data across multiple sites and using big data analysis methods that can work directly on encrypted data without needing to decrypt the data. Dependable communication methods are needed that can operate under severe operational conditions, including power outages, damaged communication lines, disruption of wireless signals, damaged communication signal transmission units, etc. Also, it must be ensured that emergency response personnel will be able to access the big data platform and coordinate their actions with other teams. This must be done in spite of areas that may have communication dead-spots, such as underground tunnels, etc. For example, it can require the deployment of sonar, light, and other communication methods in addition to the usual electrical signals for communication.

Usually, when we consider the usage of infrastructures, its efficiency, reliability and dependability are key parts. In general, big sensing data are stored in the cloud. However, in disaster situations, it might not be able to access to the cloud from disaster areas. Thus, it is important to consider the efficiency, reliability and dependability for not only the cloud side but also the sensing edges. In order to design and develop mission-critical services, we definitely need to consider failures of communication lines. A kind of autonomous recovery from such failures should be equipped. Also, data security and privacy are both important. Dissemination of incorrect information and false rumor might make the society confusion.

Given the large volume of data and the real-time constraints for performing the analysis for prediction as well as the analysis for appropriate response and recovery operations, it is imperative to ensure that the computations can be done in real-time. A variety of methods may have to be integrated together to achieve this, including the use of parallel processing based on cloud computing and local resources, reduction of large data sets into equivalent correct rules, distributed pipeline processing, etc. It is also important to ensure that the data acquisition and analysis procedures are highly dependable in spite of the failures of various processing and communication units. Given the distributed nature of such computing, it can be difficult to identify which units have failed. Hence, the processing and communication infrastructure may need to be augmented with dependable on-line system health monitoring capabilities to enable the rapid identification of faulty components and the activation of redundant standby units to ensure correct and timely completion of the big data analysis under emergency situations.

Assessment of the big data analysis algorithms is needed to determine the confidence in the correctness of the results of the analysis, including predictions and recommendations for optimal response and recovery actions. Simulation and emulation platforms can be used for evaluating and certifying new algorithms and procedures. Different algorithms can be evaluated and compared by applying them to a suite of benchmark scenarios and test-beds for which the correct results are known. It is important to quantify the confidence in the accuracy of the results of big data analysis since failure to predict a disaster in a timely way can be harmful to society. Similarly, false alarms should also be avoided since these can reduce the likelihood that the public will react appropriately to a real disaster.

3.2 The 4 V's of Big Data (Volume, Velocity, Variety, Veracity)

One way to analyze and describe Big Data research and application challenges has been referred to as the *4V's*: Volume, Velocity, Variety, Veracity. Volume is the first property characterizing big data, typically data sets of many terabytes and larger. The Volume challenge has been investigated by data warehouse companies and researchers, with many proposed and commercially available solutions, including large companies such as Teradata, new companies such as Netezza, and open source software solutions such as MapReduce and Hadoop on commodity platforms. The velocity challenge has been explored by data streaming researchers, with solutions for intelligently dropping data when redundant, with minimal information loss, and real-time decision support and data mining. Some of the Volume and Velocity challenges are further discussed in Section 3.3. 3.43.5

Typical assumptions made by the research on Volume and Velocity of big data include syntactic and semantic homogeneity (e.g., a single data warehouse) and high level trust as well as good information quality. These assumptions become problematic when we consider the other V's: Variety and Veracity.

Big Data's third V (Variety) refers to the many data sources that may contribute usefully to an application using big data. This situation arises often. For example, the relatively narrow area of radiation measurements may contain hundreds of different kinds of radiation counters, each with its own units and readings. Integration of social network data such as Twitter and YouTube presents similar interpretation difficulties when we are trying to integrate content from various sources. Some of the Variety challenges are further discussed in Section 3.4. 3.5

The fourth V (Veracity) can be described as the Big Noise in Big Data: in open source data sources such as social networks (e.g., Twitter), there is a significant amount of noise and chatter that obscure the actual information. Contradictory information appears routinely. Misinformation also may appear and become retweeted with increasing credibility. Distinguishing good information from bad data is a critical aspect of big data research, particularly for mission-critical applications such as disaster management. Some of the Veracity challenges are further discussed in Section 3.5.

3.3 Volume and Velocity

In many cases, data generated by sensors need to be processed in real-time for immediate action (e.g., Twitter and Facebook). However, the development and validation of models using real-time data is a challenging task. Archive of datasets collected during historical events that can be shared among researchers would be helpful in enabling quantitative analysis between different models – i.e., the use of a common dataset for evaluations and validations of models would help researchers in developing models with improved quality. Open access datasets collected during real events are known to be very useful to test and validate new ideas: for example, the 1998 World Cup web access trace [51] has been used by many research to advance web server/services technology. Research on cloud-based approach can potentially lead to a viable solution to accommodate the high volume of data and the multiple formats of data generated by different sensors of various types, While cloud systems have the necessary components, namely compute, storage, and network, to the development of a repository of disaster-related datasets, models, and applications, research is needed to integrate components and design interfaces that make it easy for domain scientists – which are not necessarily cloud computing experts.

The big data gathered through sensor and social networks will become useful once they are turned into actionable information that helps decision-making. There is multi-scale timeliness to decision making during disaster response. Some information should be available on the order of seconds, minutes, and other can be a matter of hours. Faster is better, but, high error rate could aggravate the situation. For example, inaccurate information could result in the distribution of rescuers and supplies to wrong places, wasting limited resources. False information could also misguide the public, increasing their stress level.

3.4 Variety: Integration of Many Heterogeneous Data Sources

The integration of many heterogeneous data sources and software tools when applying big data to disaster management is a significant challenge. At a limited scale, this heterogeneity is a challenge that has risen in big data scientific research, for example, in the construction of Global Climate Model (GCM) for global weather modeling. To achieve global weather prediction, a GCM needs to integrate a variety of atmospheric and ocean models. Given the relatively small number of component models (less than a

hundred), it has been feasible to connect them manually. In contrast, the number and variety of data sources in disaster management, as well as their rate of change, far exceeds what is feasible by manual integration. Consequently, research on automated development and maintenance of data integration tools is necessary and its success very important.

Big data analytics for disaster management and response requires a large variety of heterogeneous data sets that are related with each other and show different aspects of the changes caused by a disaster. We need the integration of such heterogeneous datasets for big data analytics. We need to handle many kinds of sensors outputting different types of data ranging from time series data to semi-structured data and textual data. These data inherently include noise and misinformation. We need to improve the trust and reliability of these data despite some noise and misinformation in them. For example, by combining information from multiple, potentially unreliable, but independent sources, we may statistically improve the trust and reliability. We also need to take into account that there may be some dependency within and among many resources. Retweets, for example, are not mutually independent.

Potential Approach to Automated Integration. One possible approach to enable the automation of data/metadata collection and management consists of several steps. First, appropriate APIs for the various data sources are defined. Second, wrappers for data sets to transform their original APIs into Information-as-a-Service (InfaaS) are developed. It is likely that we will wrap query/response (e.g., web interfaces or HTTP) requests into web service APIs, and wrap streaming interfaces (e.g., RSS feeds) into standard APIs. Third, componentized query processors are developed to process the queries at the source or by an authorized proxy near the source (e.g., in a machine located within the security perimeter of the source). Thus sensitive data can be filtered out before an acceptable result is returned to the user. Fourth, a set of library routine will be developed to integrate wrapping tools with the access methods to original data sets and sources. The componentized query processing should provide sufficient security and performance for the new generation APIs providing efficient access to integrated big data sources.

A Concrete Example Technology. A successful approach to wrapping important data sets (with metadata) into Information-as-a-Service (InfaaS) will enable applications that integrate many such InfaaS's. Even if each of them has modest size (e.g., a few terabytes), integrating dozens and hundreds of such data sources will create data sets that qualify as big data (hundreds of terabytes to petabytes). One of the promising approaches to manage large data sets with both data and metadata is RDF stores that support SPARQL queries. Unfortunately, our work has found significant performance limitations on current RDF storage systems. We are exploring several design alternatives, for example, separating the data storage from metadata storage. In this double-storage approach, a SPARQL query is analyzed and divided into two sub-queries, one for the data storage system and the other for the metadata storage system. The query results from both systems are combined together to answer the original SPARQL query on the entire data set.

Metadata Management Issues. For an appropriate interpretation of heterogeneous big data, detailed metadata is required. Some of the reports contain some metadata, but many more details (e.g., about the specific sensor used in data collection) are needed for research purposes. The collection of metadata and data provenance is a significant challenge when the data are collected under duress and stressful situations. Furthermore, the sensors are operated by a large number of different government agencies for different purposes. At the national level in Japan, the agencies include: Ministry of Internal Affairs and Communications (Fire and Disaster Management Agency); Ministry of Agriculture, Forestry and Fisheries (e.g., radioactivity in fish); Ministry of Land, Infrastructure, Transport and Tourism (Land and Water Bureau); Ministry of Environment; Ministry of Health, Labor and Welfare (Pharmaceutical and Food Safety Bureau). Many more sensors are operated by prefectures, universities, and other agencies. Consequently, the main challenge has shifted from sensor insertion for a relatively small amount of data collection to the management of large but varied data from many sensors.

Complementarity of Areas. Environmental sensors and social media are representative examples of big data sources, with complementary characteristics and requirements. At one end of automation spectrum, environmental sensors produce automatically relatively reliable data, with the quantity and variety of data being the main challenges. At one end of spectrum, a large number of human participants in social media produce relatively small amounts of data individually, with the information quality and management being the main issues. By studying both areas, which appear to be ends of an spectrum, we hope to identify re-

search challenges and create technologies that will benefit a large number of disaster management areas as well as big data research.

3.5 Veracity: Filtering of Noise and Misinformation in Open Source Data

3.5.1 Big Noise in Big Data

One of the most significant challenges in using open source big data such as Twitter and YouTube is the amount of noise in such open source information. There are many sources that produce such noise, which we discuss briefly to illustrate the magnitude and difficulty of the problem.

To simplify the discussion, we divide the sources of noise in open source information into intentional and unintentional. Intentional sources of noise and misinformation are produced by cyber-attacks designed for illicit financial gains or tactical advantage in a larger conflict. Starting from email spam, many forms of misinformation and deceptive information have been deployed, including phishing (a deceptive and harmful form of spam), link farms (to deceive search ranking algorithms), and others. Generalized cyber-attacks have been employed by all parties during the Russia-Georgia War of 2008, including distributed denial of service, propaganda, and disinformation. Cyber-attacks are of major concern during man-made disasters, particularly coordinated terrorist attacks.

Perhaps more directly relevant to natural disasters are the unintentional sources of noise. The first example of such noise is rumor spreading in social networks. As a concrete example, rumors spread in Twitter during disasters such as in the aftermath of the Tohoku Earthquake. For example, information on Twitter helped the rescuing of children and teachers who were stranded at a school building. However, finding this information was extremely hard because a lot of unverified tweets spread in Twitter, even after people pointed out that the unverified tweets were false rumors. Passing of rumor to others in uncertain environments is human nature. Social media technologies can facilitate the spread of false information as well as the spread of counter information that attempts to correct the false information, but it is unclear how to take advantage of these technologies to reduce the spread of misinformation and at the same time increase the spread of useful information such as alerts and warnings.

The second example of sources of unintentional noise is a rapidly changing environment, which is the case during disasters. Tweets on the situation (e.g., subway station opening after the Tohoku Earthquake) can be quickly superseded by new events (e.g., temporary closure of subway stations). Linking the information to an appropriate point or region of space-time continuum is very important when the environment changes so quickly.

The third example of sources of unintentional noise is the ability of social media to amplify all information, whether it is true or false. For example, rumors that would have been harmless can become a major problem when retweeted often. Sometimes, unaware users may contribute to the problem by retweeting out-of-date or malicious content. This happens naturally since there are no easy ways of verifying information in many social networks.

Improving the quality of information in social media (i.e., filtering out the big noise in big data) is a big challenge and it involves many related issues. For example, in the early phase of Tohoku Earthquake nuclear disaster, government's official stance on Fukushima Daiichi Nuclear Power Plant was that there were no problems at the plant. However, citizens did not trust the official reports, and consulted multiple sources (including foreign news reports) to make sense of the situation. This distrust resulted in people spreading rumors and inaccurate information in social media, decreasing the quality of information that they interact with in social media.

3.5.2 Trust, Security and Privacy in Big Data

Spam filtering in particular and quality of information in general, are among the most difficult challenges in information security. Crowd-sourcing (users clicking on spam) is among the most effective methods for filtering spam by large mail services such as gmail. However, due to inherent delays such crowd-sourcing methods may not be directly applicable in emergency situations during a disaster. One possible technique for increasing the information trust level is to timestamp and location-stamp each posting (e.g., adding the location information to tweets by default during a disaster), thus enabling more detailed validation of the data against its content and correlation with other information sources.

One potential approach to filtering big noise applies ideas from classic reliability techniques. For example, information from several independent sources (e.g., different social networks) may be more reliable than information from dependent sources (e.g., retweets of the same original tweet). However, the application of this idea depends on advances in data integration (Section 3.4), since independent sources are likely to have different syntax and semantics of their content. Using a database terminology, we are attempting to join data from tables with different schema, thus an inherently difficult problem.

A related area of research to reliability consists of diversification methods, which can be deployed to reduce the likelihood of successful security attacks. Also, continuous big data analysis for streaming data, such as output of sensors, results of crowd sourcing, etc., must be enhanced with anomaly detection mechanisms to identify data that may be incorrect due to sensor failures, security attacks, etc. Since the operational environments are also evolving, some uncertainties could be due to the fact that a specific detection and classification method is no longer fully correct. Hence, the uncertainty quantification method must be designed to be evolvable. This can be done with integrated machine learning methods and feedback from authorized personnel. Machine learning can also limit the clutter in big data storage and analysis by replacing large amounts of data by correct equivalent inference rules. This will also speed up the analysis procedure under emergency situations since the rules can be used to rapidly perform the prediction analysis as well as to identify optimal response strategies.

Another of the potential roadblocks in the collection and use of integrated new big data is the lack of security and privacy guarantees in current IT systems. For good reason, big data producers are extremely reluctant to share information. This problem needs to be addressed for the mission-critical real-time applications such as emergency response, since sensors are collecting such data automatically and timely decision making can only be achieved through automated integration and sharing, with appropriate security and privacy guarantees. This is an evolving issue, with non-technical constraints (e.g., evolving and differing legal rules for each country) that need to be addressed by software tools and platforms as envisioned by this team.

Despite the difficulties of solving the quality of information problem in general, there is hope for disaster management as a specific big data application. For many kinds of disasters (e.g., earthquakes), there are trustworthy sources of information. In the case of earthquakes, the seismometer network can detect earthquakes accurately and reliably. Consequently, we can use the authoritative source of earthquake information to filter social media information on earthquakes. Unfortunately, there are also many kinds of disasters for which we do not have authoritative sources from dedicated sensors. An example is landslides in remote areas that have not been instrumented. In general, we anticipate fast advances in this area due to the ongoing successful research, development, and deployment of both dedicated and multipurpose sensors.

Finally, research in this area also needs to examine how people perceive information in social media and how they contribute information in social media. People's perception determines how they interact with the information, such as whether or not to believe the information and pass it to others. How people contribute information to social media determines the quality of information in social media. Using better understanding of how people process information, work in this area can design tools and sensors that direct people to contribute high-quality information that they trust and act on. By doing so, the quality of big data collected from social media will naturally improve. More specifically, some research questions are: How do we minimize the spread of misinformation in social media during disasters?; How do we evaluate the credibility of information in social media during disasters?; How do we identify actionable information in social media during disasters?; How do we build trust in emerging digital communities during disasters? How do we verify information in social media during disasters? How can we collect and integrate evidence from multiple sources to verify information, to evaluate credibility, to build trust, to identify actionable information, and to minimize the spread of misinformation?

3.6 Relevant Technologies and Related Research Themes

There are several research themes that have faced similar challenges or worked on relevant technologies and tools. We outline three to illustrate their relevance and leverage for the application of Big Data technologies to disaster management. In Section 3.6.1 we discuss crowd-sourcing; in Section 3.6.2 we discuss cyber-physical systems; and in Section 3.6.3 we discuss service computing.

3.6.1 Crowd-sourcing To the Rescue

Crowdsourcing is a generic term describing on-line distributed problem solving by potentially large numbers of people collaborating via open and dynamic Web-based applications. A formal example of a web-based system that makes use of crowd sourcing is Amazon's Mechanical Turk which allows users to submit tasks for a job that is too large for a single person or small team to perform and distribute those tasks to a large number of volunteers distributed throughout the world who get paid a small amount of money to perform those tasks. A straightforward task that needs to be applied to thousands of data elements might be completed in a few days or weeks via such a system. Crowdsourcing also refers to the "wisdom of the crowd" that can be tapped via social networks. Users will say things like "I needed to find a good Italian restaurant in Boise, so I crowdsourced it," meaning they posted the question to a service such as Twitter or Facebook and received advice from their extended network of friends.

These types of systems can be useful in mass emergencies to allow people to gather information, report information, volunteer to help, ask for help, or to re-broadcast ("retweet") useful information. Individual members of the public can use this information to determine whether they should follow an evacuation order [46] while government agencies can use this information to determine the allocation of resources or to get an overall sense for the status of a region or city.

The challenge for crisis informatics is to perform research that both examines how such systems are used currently by multiple stakeholders to inform the design of next generation crisis informatics systems designed from the start to harness the power of the crowd. What software architectures are needed by these systems? How are they put into production? Who operates/maintains these systems? What tools and services can they provide?

As one example, Project EPIC has been developing a web-based system to support the activity of pet matching after a disaster [47]. Pet matching refers to the process of locating pets that have been lost after a disaster and reuniting them with their families. This is a large-scale problem; an estimated 200,000 pets were lost after Hurricane Katrina in the southern United States in Fall 2005 and only 5% of those animals were reunited with their original families. Project EPIC has been studying how on-line pet activists adapted features provided by Facebook to enable pet matching activities after 2012's Hurricane Sandy and is designing a system that will make those efforts more effective. The architecture of this particular system is web-based, making use of a popular MVC web application framework, but with multiple web services to enable tight integration with existing social media websites and making use of a hybrid persistence architecture that involves both relational and NoSQL storage technologies. Finally, the system includes a machine learning component that monitors the actions of the system's users and identifies those users who are most effective at finding matches and provides those users with additional "power user" services not provided to users who interact with the system infrequently. Project EPIC's work on the Emergency Pet Finder system will be useful to understanding the range of issues that must be tackled when developing crisis informatics systems that leverage crowdsourcing in some fashion.

During the disaster response phase, the evacuation of people in a safe way is definitely one of the most urgent and important initial actions to take. Here we need the better balanced combination of planned-for evacuation navigation and agile analysis, simulation, and improvisational update of evacuation navigation based on the real-time monitoring of the situation. For the latter agile part, the potential approach may be the crowdsourcing of evacuation navigation information. Based on the planed-for evacuation navigation, people will evacuate. However, unexpected events may block the recommended evacuation route. Sucj information should be shared in real-time by evacuating people. Smartphone-based evacuation navigation system carried by evacuating people may also work as the crowdsourcing of the real-time information gathering of dynamically changing situations of available evacuation routes. This is similar to the case of crowdsourcing-type car navigation system such as WAZE. If the population of the users is sufficiently dense, the navigation by such a system is usually better than conventional car navigation systems to avoid traffic jams. Evacuation simulation is also a big challenge. People do not act randomly as assumed by most of the conventional traffic simulation systems. We need to develop a simulator that can take into account different aspects of people that externally or internally determine their behavior patterns. In March 11, many parents, when they received Tsunami alarm, chose the way toward the coast to save their children staying in their schools and lost their lives without knowing the fact that their children al-

ready evacuated to safe areas. Such tragedy was also cause by the lack of a real-time and agile information sharing infrastructure for people in the public.

Crowdsourcing of information gathering and the sharing of the analyzed results by people are definitely powerful approaches to get the real-time monitoring of the rapidly changing situations influenced by lots of unexpected events such as cascaded hazards. However, the crowdsourcing of information gathering brings the following questions. How can we collect information from a bunch of people? If the collection does not require any user interaction, and automatically done through the communication with smartphones, what we need is the people's incentive to download and install the corresponding application in their smartphones. Such an incentive also requires the provision of useful information shared by its users not only during the disaster, but also during their daily lives. If the collection of data requires users' interaction with the application, then we need to provide users with stronger incentives. The second question is how to distribute the information in an appropriate way in a disaster. Some information may trigger a panic, which may cause further disasters. We need to prevent such panics by providing appropriate information to appropriate people at appropriate time. Privacy protection is always a big problem. However, in order to save as many lives as possible during a disaster, privacy protection policy may be dynamically changed during an emergency by authoritative organizations, or sometime even by on-site people involved without obtaining any permission from the authority. Even the latter case is considered necessary in most cases to save lives. The infrastructure system should also support the real-time dynamical changes of the policy and the sharing of such information.

3.6.2 Cyber-Physical Systems (CPS)

Cyber-physical systems (CPS) play a critical role in disaster management: unmanned ground/aerial/marine vehicles perform search and rescues during disasters, especially in hazard conditions that are dangerous to humans; intelligent transportation systems, which connect individual automobiles to the cyber-infrastructure and the physical infrastructure, provide transportation means for evacuation and emergency response; and health monitoring and telemedicine infrastructure provides prompt remote medical treatments to critical patients at disaster locations.

Building CPS for non-disaster conditions is already difficult; yet, building CPS for disaster management and emergency response is significantly more challenging. Two key important challenges are how to tackle the exponential and multidimensional complexity of their operating environments and how to meet the strict design requirements of CPS that are necessary for such environments. For instance, a set of unmanned vehicles that perform search and rescue missions need to interact with a heterogeneous and interdependent set of devices, information systems and other CPS under various natural environmental situations and human activities. They must be controlled and coordinated across multiple spatial and temporal scales, while guaranteeing reliable and real-time responses under stringent power and network constraints. Their tasks need to integrate local and remote, centralized and distributed data processing and communications, and their missions change dramatically depending on the current disaster situations and/or potential system failures, which can only be achieved with a high degree of autonomy, situation awareness, and adaptability in the CPS' behaviors. To achieve the above characteristics of CPS, several advanced technologies are crucial; for instance, new abstraction and compositional analysis methods are needed to manage the multidimensional complexity of the design without sacrificing its accuracy; new modeling formalisms that can succinctly represent different CPS environmental factors and their correlations are necessary to handle and leverage on the environment heterogeneity; new techniques for fault-tolerant, mixed-criticality, re-source-efficient design are needed to provide reliable real-time responses for critical tasks during emergency response; and new methodologies for multimodal, reconfigurable and distributed control and platform co-design are crucial to achieve autonomy and adaptivity.

Another key challenge in the development of CPS for disaster management is how to integrate human as an overarching component in the cyber-physical systems, which is critical due to the close interconnection between human and the systems. This human-centric CPS approach needs to incorporate human in the loop at all stages, including not only at high-level modeling, analysis and design stages, but also during deployment and real-time operation. The interactions between human and the CPS system encompass multiple roles: (1) users of the system, such as a patient at a disaster location who is undergoing a tele-surgery by a medical CPS, (2) experts and operators, such as the doctors who perform/oversee the surgery remotely by controlling the medical CPS, and (3) other human with indirect interaction, such as

other victims and rescuers at the disaster location. It would be interesting to adapt machine learning techniques and big data to construct more accurate models for users (e.g., patient model) and to extract and integrate the knowledge of experts and operators (e.g., doctors) in the development of the CPS. At the same time, the vast amount of information enabled by big data can be used to create a formal model for the uncertainty introduced by other human who indirectly interact with the system.

In general, physical world contains uncertainty and clutter. Also, human activities contain similar uncertainty and clutter. Thus, we need techniques for quantifying such uncertainty and dealing with clutter. In addition, information is sensed in widely varying contexts and often with diverse forms. These complexities can materially affect the quality of the inferences, decisions, and responses in a disaster setting. Big data allows for robust learning and decision making in this setting; it reduces our dependence upon prior assumptions in models. We need to construct reliable infrastructures for sending sensing information from physical sensors, processing such sensing information in the cloud and providing its feedback to the sensors. We therefore need to combine the big data driven modeling with the physical driven model and construct the corresponding Human centric Cyber Physical System (HCPS) model where we need to design and develop feedback mechanisms for control systems with human-in-the loop.

When we consider network infrastructures, we need to consider distributed, efficient, scalable and dependable in-network processing since usually data exceeds resources in disaster situations. We also need efficient machine learning mechanisms in such networks so that we can give feedbacks to physical sensors and humans in real-time. We also need to create solid ad-hoc communication infrastructures and prescriptive networks (configurable networks). We also need to efficient large-scale network simulation and emulation mechanisms. The construction of such mechanisms can help for designing CPS considering human activity (Human centric CPS).

3.6.3 Service Computing and Cloud Computing

Overview. The big data produced during a disaster (e.g., by sensors, and social media) have to be collected, integrated, and delivered to Big Data Consumer applications to achieve their new functionality. In addition, disaster management research requires the integration of disaster data with many other big data sources including, but not limited to mapping, land survey, environmental, satellite imagery, population and part disaster datasets; as well as models for climate, geomorphology and hazard spread (e.g., radioactive material, disease) forecasting. Concretely, software tools should be created to assist the large scale data collection and management efforts (e.g., DIAS [28]), instead of small scale specialized data collections. Software tools should be developed to automate application deployment that will create the wrappers to mediate heterogeneous big data access while preserving quality of service and quality of information requirements. Another growing area of opportunity is storing historic data sets and sensor data in a resilient computing cloud environment such as the Amazon EC2 and S3. Using an analogy with cloud computing and service computing, the big data infrastructure for disaster management is divided into three layers as outlined below.

Information as an Infrastructure. The current practices of storing and accessing sensor data primarily as raw data (e.g., time series) could be described as Data as an Infrastructure (DataaaI). As software tools are developed to transform raw data into actionable information, we will be able to create *Information as an Infrastructure* (InfaaI). Although DataaaI and InfaaI methods and APIs are the state-of-the-art, they reflect the scientific data management practice of single use by each project or experiment. In modern big-data-driven experimental science, a shared sensor infrastructure (e.g., the Large Hadron Collider at CERN [48]) produces large amounts of raw data that are analyzed by various research groups.

Information as a Platform. In the medium term, sensor data should become accessible in various forms, including the raw form as well as at a higher level of abstraction, in a format that could be called *Information as a Platform* (InfaaP). InfaaP supports access methods such as database (SQL) queries and sophisticated data manipulations, as being developed in the Center for Coastal Margin Observation & Prediction [31] (CMOP, an NSF Science and Technology Center), where users can run simulations and other research tools to learn from and use CMOP information without having to process raw time series data. Significant improvements (e.g., a comprehensive metadata collection and management support) are required for InfaaP.

Information as a Service. In the long term, we envision *Information as a Service* (InfaaS), where new software tools are combined to present the sensor data in various forms for analysis and visualization. An example of InfaaS is a prototype demo [22] showing landslide susceptibility or hazard map and other data sources such as weather (rain storms), recent earthquakes, and social media, shown on Google Maps. The demo will be augmented with a combination of direct data query, research tools, Google Crisis Response tools, and weather forecast style information display. We anticipate the InfaaS to be useful for a variety of users, including short term relief workers, long term decision makers, and the public. Cloud-based solutions can address these quality of service and quality of information dimensions, necessary for mission-critical disaster recovery systems when local systems become unavailable (e.g., through electrical power outages).

Development of an architecture for each layer (InfaaI, InfaaP, and InfaaS) of the big data infrastructure for disaster management should consider the distributed nature of the data, the heterogeneity in data sources formats (structured and non-structured), protocols and semantics, the need to meet real-time constraints despite its volume, and quality of data sources. Challenges at the InfaaI layer include considering different models for data distribution (centralized, distributed, peer-to-peer or hybrid), replication (number and form of replicas) and archival; improvements to communication protocols over wide-area networks; resilience of the infrastructure itself against failures and attacks; and management of distributed access policies especially when considering national and international boundaries. Optimizations and automated adaptations to existing distributed large scale systems (as the data-centric High Energy Physics) are worth facing [48]. At the InfaaP level, the main challenge is in developing a standard higher-level language (e.g., SQL-like, and semantic query) that abstracts the access to the potentially distributed raw data, and performs transformations needed to cope with format, semantic or quality differences present in the raw data, while optimizing and prioritizing the use of InfaaI-level services. Examples include the various languages developed for NoSQL systems (e.g., Unstructured Query Language, Cassandra Query Language, Hive query language) and domain specific ontologies. At the InfaaS layer, the challenge is to develop modular and general tools that can be rapidly integrated and shaped to support the data analysis requirements of a particular disaster.

3.7 Adoption Issues for Big Data Approaches in Disaster Management

One of the key questions in disaster management research is the acceptance and adoption of technologically advanced solutions by the public. To illustrate improved emergency response, let us consider the idea of “call in emergency” list, which has been implemented by many institutions. For example, large institutions such as University of Florida and Georgia Tech offer an opt-in emergency notification service to send critical information to employees and students when emergencies arise. The notifications modes supported currently include phone calls, emails, and SMS to smartphones. The Georgia Tech emergency notification system is very simple: once a campus emergency is declared it notifies every phone and device regardless of their location, on or off campus.

3.7.1 Practical Questions on Technology Adoption

Another emergency response scenario being suggested (and debated) is the use of smartphone location information to notify every phone in a danger zone during an emergency. (For the following discussion we will assume that the legal rules about opt-in and opt-out can be satisfied.) Although the idea of emergency notification is a relatively simple concept, there are many opportunities that arise beyond a hardwired message. Let us consider its application to the 3-11 scenario, where the tsunami arrived the Tohoku coast a few minutes after the earthquake. A quick notification (e.g., SMS) could be sent immediately to the coastal areas after an earthquake has been confirmed. As more information becomes available, e.g., the epicenter of earthquake is determined, phones are classified by their distance to the epicenter and their risks assessed accordingly (e.g., their distance to the seashore. Additional emergency response information is then sent to phones according to the risk level due to their location. This additional information can be customized for each location as well as other available information such as whether the phone has been used or moved since the start of earthquake. For phones with active users, appropriate escape routes could be sent to each phone, using facilities such as Google Crisis Response [53]. Each refinement step requires significant knowledge of the environment (provided by big data sources) plus an accurate knowledge of the disaster for accurate risk assessment and appropriate response.

To increase the adoption of technologies developed for disaster prevention, preparation, response, and recovery, it is essential to consider the integration of the new infrastructure with tools currently in use and offer familiarization of new tools before a disaster takes place. For example, geographical replication is a known solution to recover IT from a devastating disaster. However, these incur costs that prevent many small businesses from implementing it. Given that electronic businesses are increasingly adopting cloud computing and virtualization, [49] proposed the use of migration to mitigate the impact of disasters to an IT infrastructure when a disaster is predictable, while [50] increases the possibility of adoption for those taking advantage of advanced I/O technologies in virtualized environments. Studies that consider the barriers for adoption such as incompatibilities between solutions, cost/benefit of a solution, cultural differences, and amount of training are expected to accelerate adoption by providing accurate assessments.

It is important to have opportunities for public training. However, there are so many people who are not familiar with ICT devices. Thus, we should develop ICT devices with simple and understandable user interfaces so that many people can recognize emergency alerts.

To increase public awareness in disaster prevention, preparation, response and recovery, it is critical to create effective training environments and programs that help prepare the public for future disasters. This can be achieved through appropriate education programs, such as ones that teach the public about potential disasters and how they progress, how should one response in such disaster situations, what tools, services and help are available, how one can effectively assist others during actual disasters and post-disasters. In addition, conducting regular testing of the services and public responses under simulated disaster settings is important not only to increase public knowledge and awareness but also to evaluate the services and infrastructure. Collaborations between various government and industry stakeholders and scientists of various fields (e.g., engineering, psychology, healthcare) to construct such training environments and programs are needed to effectively educate the public.

3.7.2 An Illustrative Approach to Facilitate Adoption

One possible approach to describe adoption issues in disaster management research is to divide the problem into two sensing models: environmental and human. On the one side is environmental sensing modeling to capture natural phenomena, which build on significant historical data from past disasters. These static environmental models are augmented by real-time data from sensors when a disaster occurs. On the other side, human behavior modeling captures the behavior of the public, e.g., evacuation history of people from historical disasters, including the density and distribution of people in target areas. Similar to traffic monitoring, we can obtain trajectory big data from mobile phones, RF-ID tag information used in public transportations, camera information in urban areas, and so on. Such information is very useful for estimating behavior modeling of peoples. We can combine these two models together to enable research on human response to specific environmental stimuli and construction of software tools to improve human adoption of big data-based tools.

Prevention: Big data can be used to identify civil and other infrastructure improvements that can help prevent disasters. One source of such data is a detailed post-disaster big data analysis to determine what structures could help reduce the severity of the scope of a hazard or provide better guidance to the public into taking actions that can limit the adverse impacts of a hazard. For example, post-incident analysis of forest fires can help determine whether regular managed small fires can help prevent the occurrence of large uncontrolled fires. Similarly, analysis of river flows can help guide the design of levees and reservoirs to limit the occurrences of floods. The data needed for such type of post-disaster analysis include detailed information about (1) buildings, bridges, roads, etc, in the disaster region; (2) their usage at the time of the disaster, such as level of occupancy, power/water/gas usage, etc.; (3) their location and orientation relative to each other; and (4) detailed disaster information, including its intensity, orientation, trajectory, etc.

When the Tohoku Earthquake occurred, several base stations of mobile phones were damaged. However, more than 80% of damages were caused by blackouts. Thus, in order to use infrastructure networks in disaster situations, we need to take measures to cope with such blackouts. Furthermore, we also need to make plans in advance how we can prepare partial standstills of infrastructure networks. Also, we need to understand how performance degrades as infrastructure networks start to fail and to devise autonomous mechanisms for using partially survived infrastructure networks. Using the above two types of models, we can provide more comprehensive models for disaster management, prevention and mitigation.

Preparedness: Big data analysis can be used to identify the most likely types of hazards for a region and periodically alert and prepare the public to know the best strategies for dealing with such hazards. For example, for flooding situations, the best strategy may be to move to higher levels while for coping with a tornado the best approach may be to move to the basement. Big data analysis can also guide the proactive deployment of resources to fully cope with an impending type of disaster. This could be the pre-deployment of emergency response personnel and also emergency response resources to respond to an approaching hazard. For example, mobile power supply resources can be moved to critical structures that may have to cope with extended power outages and fire-fighting resources can be deployed in neighborhoods that may be impacted by an approaching forest fire.

Figure 3 Underground City Evacuation Scenario

As a concrete scenario, in order to improve efficiency of evacuations in disaster situations, rescue teams should receive training under realistic conditions before disasters strike. However, it may be too expensive to organize several large-scale training sessions. Therefore, elaborated simulations, training and test-beds should be considered as alternatives. The environmental and human sensing models can provide a foundation for more elaborated simulations and test-beds. For example, suppose that a flood has occurred near a large city and flooded water is reaching an underground city. From the natural sensing model, we can estimate how much time it will take for the flooded water to reach the underground city. Similarly, the human sensing model can provide precious information about how people may choose to escape from the underground city and how we can prevent congestion at exits of the underground city (illustrated in Figure 3).

4. Benefits of and Needs for International Collaboration

4.1 Benefits of Previous International Collaboration on Disaster Management

International collaboration on disaster management research has a proven success record. Recent successful examples include the NSF RAPID program and the J-RAPID program in Japan following the Tohoku Earthquake. An example of such collaboration is the project entitled “Humanitarian logistics in the Tohoku disasters 2011”. The PI on the Japanese side is Prof. Eiichi Taniguchi of Kyoto University and the PI on the US side is Prof. Jose Holguin-Veras of RPI (Rensselaer Polytechnic Institute). Both PIs are leading researchers in the area of logistics. They have conducted the joint research utilizing their experience in Haiti earthquake and Hurricane Katrina (Holguin-Veras) and the Great Hanshin Earthquake (Taniguchi).

The objectives of this study is to find success and failure elements of humanitarian logistics which was carried out in the form of relief supply distribution to displaced people in Tohoku disasters 2011 for improving the future planning and preparation of relief supply distribution against catastrophic disasters.

They developed a multi-objective optimization model of vehicle routing and scheduling problems [52]. The model was applied to road network in Ishinomaki city. The results showed the model can be used for optimising the delivery system in the case of emergency in which the demands of displaced people exceed supplies. It is expected that findings the joint research team has found through interview and inves-

tigations conducted in disaster area as well as the optimization model will improve logistics problem in catastrophic events in future.

We believe this tradition of international collaboration on disaster management should be further developed and enhanced to build joint research projects larger than feasible individually.

4.2 Scientific Challenges Arising from Cultural Differences

Scientific challenges may arise from cultural differences. The same infrastructure may be used in different ways in different countries. For example, people in USA mainly drive their cars for transportation, whereas in Japan people mainly take public transportation. Another example is twitter usage in both countries. It may be different for English and Japanese, which may need comparative sociological, psychological, and/or cultural studies. Disaster management and response infrastructure need to take into account such differences and similarities due to cultural differences in its design and operation, which can be best studied by a US-Japan collaborative team knowledgeable in two cultures.

Solutions to disaster management problems and their social acceptability will no doubt be influenced by culture, even in technologically advanced societies. The ability to understand and potentially exploit the differences and the commonalities of the response to common solutions in both the United States and in Japan is expected to be an important outcome of this international collaboration.

4.3 Needs for International Collaboration on Big Data for Disaster Management

Excellent work is being done in both big data and disaster management research. However, those projects are often specifically targeting one kind of disaster, or mainly applicable to disasters in a single country. Examples of support for disaster management research in Japan from the support for reconstruction and revitalization after the Tohoku Earthquake:

- JST Center for Revitalization Promotion: JST supports the areas devastated by the Great East Japan Earthquake. With its various resources, including knowledge and know-how, JST enhances academic-industry alliance, accelerates post-disaster reconstruction and revitalization, and finally triggers innovation in this area. The Center has branch offices in Sendai (Miyagi Prefecture), Morioka (Iwate Prefecture), and Koriyama (Fukushima Prefecture).
- Developing technologies/devices to measure/analyze radiation. Program for the development of the devices/systems to measure/analyze the dose and density of the radiation in food and soil, for prompt and reliable use in the disaster areas. 14 projects were selected.
- Establishing an R&D center for innovative energy (commissioned project). MEXT and METI collaborate to establish the globally-advanced center for research and development of sustainable energy. The center will open in 2014 in Fukushima. JST commissioned to research and develop innovative super-high-efficient solar cells in cooperation with universities and private sector in Japan as well as researchers outside Japan.

Similarly, excellent work is being done in big data research. However, those big data technologies and tools are often application-agnostic and thus not directly applicable to disaster management. The May 2013 Whitehouse announcement of Big Data Initiative in the US included 6 federal agencies with a total of 200 million dollars investment. However, these ongoing programs are very general, not directly relevant to disaster management.

In Japan, the first major research project on big data was Info-plosion Project funded from January 2005 to March 2011 by JSPS with Masaru Kitsuregawa as the PI and a large number of partners mainly from academy. It focused on the R&D on the infrastructure for the information explosion era, covering the research issues in discovery science, platform systems, and testbed systems. Masaru Kitsuregawa is also conducting the Info-Energy Generator Project from FY 2010 to 2014. This project is one of the FIRST (Funding Program for World-Leading Innovative R&D for Science and Technology) projects initiated by MEXT. It focuses on the development of the fastest database engine for the exa of very large databases, and experiment and evaluation of strategic social services embedded by the database engine. In 2012, MEXT launched a 5 year project on CPS (Cyber-Physical System) Integrated Platform for Efficient Social Services with Masao Sakauchi as the PI, following its 1 year feasibility study in 2011. This is a joint project of NII (National Institute of Informatics), Hokkaido University, Osaka University and Kyushu University. It focuses on the R&D of social cyber-physical system architectures and technologies with the real-time

monitoring of social systems and their physical worlds, integrated data analytics of both cyber world data and physical world real-time data, and the real-time optimization of the social system services based on the data analytics result. In 2013, MEXT launched two new JST CREST programs on big data, one on big data application technologies with Yuzuru Tanaka as the program officer, and the other on big data core technologies with Masaru Kitsuregawa as the program officer. During the coming three years, each of them will select around 10 CREST projects based on open call-for-proposals, and each of them will be conducted for five years under the supervision of each PO. In addition to these MEXT initiative programs, METI launched a new NEDO program in 2012: R&D and Demonstration Projects for Innovative Social Systems based on the Fusion with Information Technologies. This focuses on three application areas, urban transportation (usability, optimal and disaster-resistant services), healthcare (human centered secure data management), and industrial ecosystem among agriculture, commerce and engineering. These are the current major programs in Japan on big data and social cyber-physical systems.

To remedy this situation, we need research on applying big data approaches to disaster management. This way, we can leverage the ongoing investments in both disaster management and big data to achieve global contributions to the application of big data concepts, techniques, and tools for effective disaster management.

As an example of technical collaboration, the confidence in the big data analysis method can be estimated through international collaborations that enable data to be used from many different regions to assess the accuracy of the analysis procedure. Also, the procedures can be used to predict various aspects of each disaster to further quantify the confidence in the analysis procedure. For example, big data analysis can be used to predict the path and intensity of an approaching hurricane and the results can later be evaluated based on the actual observed path and intensity of the hurricane to quantify the confidence in the correctness of the analysis procedure.

5. Summary and Conclusion

Disaster Management is an important global problem. Disasters affect every country on Earth and effective disaster management is a global challenge. This is particularly the case of large-scale disasters that affect many countries (e.g., the 2004 Indian Ocean earthquake and tsunami) and multi-hazards such as the Tohoku Earthquake and landslides. Tools that can be used by many countries will have significant broad impact in helping the world population as well as many government agencies and NGOs.

Big Data is a great global opportunity for disaster management. Big data has already demonstrated its usefulness for both dedicated sensor networks (e.g., earthquake detection during the Tohoku Earthquake) and multi-purpose sensor networks (e.g., social media such as Twitter). However, significant research challenges remain, particularly in the areas of Variety of data sources and Veracity of data content. We call this the Big Noise in Big Data challenge.

To fulfill the potential benefits of applying big data to disaster management, we need to bring together the best minds from around the world. From the disaster management view, we need the technology push from big data researchers to tackle the challenges mentioned above (e.g., Big Noise) so big data tools can effectively address disaster management issues. From the big data view, we need the application pull of disaster management researchers to apply big data techniques and tools to solve real world problems. The international collaboration between Japan and USA will help both the disaster management and big data research community take the step forward towards better disaster management using big data.

6. References

[1] CRICIS: Critical Real-Time Computing and Information Systems, eds: Robin Murphy, Trevor Darrell. Report from the NSF/CCC Community Workshop on Computing for Disasters, June 2012, Washington, DC. [(http://www.cra.org/ccc/disaster-management.php)].

[2] President's Council of Advisors on Science and Technology, *NIT for Resilient Physical Systems*, 2007, President's Council of Advisors on Science and Technology, Executive Office of the President.

[3] Computing Research Association, Grand Research Challenges in Information Systems2003, Washington, DC.

[4] President's Council of Advisors on Science and Technology, The Science and Technology of Combating Terrorism, 2003, Office of Science and Technology Policy, Executive Office of the President.

[5] Committee on Environment and Natural Resources, Grand Challenges for Disaster Reduction, 2008, National Science and Technology Council, Executive Office of the President.

[6] Sircar, J., et al., Catastrophe Bonds for Transportation Assets Feasibility Analysis for Bridges,, in Transportation Research Record: journal of the Transportation Research Board, No. 21152009, Transportation Research Board of the National Academies: Washington, D.C. p. 12-19.

[7] National Research Council, Citizen Engagement in Emergency Planning for a Flu Pandemic: A Summary of the October 23, *2006 Workshop of the Disasters Roundtable,* Disasters Roundtable [DR] and Earth and Life Studies (DELS), Editors. 2007, The National Academies Press: Washington, DC.

[8] Mason, B., ed. Community Disaster Resilience: A Summary of the March 20, 2006 Workshop of the Disasters Roundtable. ed. Disasters Roundtable (DR) and Earth and Life Studies [DELS)2006, The National Academies Press: Washington, DC.

[9] Kershaw, P.]. and B. Mason, The Indian Ocean Tsunami Disaster: Implications for U.S. and Global Disaster Reduction and Preparedness- Summary of the june 21, 2005 Workshop of the Disasters Roundtable, Disasters Roundtable (DR) and Earth and Life Studies [DELS), Editors. 2006, The National Academies Press: Washington, DC.

[10] Committee on the Future of Emergency Care in the United States Health System, Emergency Medical Services: At the Crossroads, Board on Health Care Services (l-ICS] and Institute of Medicine (IOM), Editors. 2007, The National Academies Press: Washington, DC.

[11] Committee on the Effective Use of Data Methodologies and Technologies to Estimate Subnational Populations at Risk, Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises, Board on Earth Sciences and Resources [BESR), et al., Editors. 2007, The National Academies Press: Washington, DC.

[12] Committee on Planning for Catastrophe: A Blueprint for Improving Geospatial Data Tools and Infrastructure, Successful Response Starts with a Map: Improving Geospatial Support for Disaster Management, ed. Board on Earth Sciences and Resources (BESR) and Earth and Life Studies (DELS) 2007, Washington, DC: The National Academies Press.

[13] Committee on Disaster Research in the Social Sciences: Future Challenges and Opportunities, Facing Hazards and Disasters: Understanding Human Dimensions, Earth and Life Studies (DELS), Editor 2006, The National Academies Press: Washington, DC.

[14] Blanchard, B.W., Guide To Emergency Management Anal Related Terms, Definitions, Concepts, Acronyms, Organizations, Programs, Guidance, Executive Orders & Legislation 2007: FEMA Emergency Management Institute. [http://training.fema.gov/EMlWeb/edu/termdef.asp].

[15] FEMA's Whole Community approach http://www.fema.gov/whole-community.

[16] FEMA's Individual Assistance Program Tools http://www.fema.gov/individual-assistance-program-tools.

[17] World Disasters Report 2010: Focus on urban risk, ed. D. McClean, 2010: International Federation of Red Cross and Red Crescent Societies.

[18] Sahana Software Foundation http://sahanafoundation.org/.

[19] Ushahidi, a non-profit tech company http://www.ushahidi.com/.

[20] TED: US Geological Survey's Twitter Earthquake Detector http://recovery.doi.gov/press/us-geological-survey-twitter-earthquake-detector-ted/.

[21] Nano-tera program in Switzerland http://www.nano-tera.ch/.

[22] GRAIT-DM SAVI web site https://grait-dm.gatech.edu/.

[23] Japanese government radiation information web portal set up by the WIDE project, with an English version available at <http://eq.wide.ad.jp/index_en.html>.

[24] System for Prediction of Environment Emergency Dose Information (SPEEDI), by *Nuclear Safety Division, Ministry of Education, Culture, Sports, Science and Technology* <http://www.bousai.ne.jp/eng/>.
[25] *Environmental Radioactivity Monitoring Center of Fukushima* <http://www.atom-moc.pref.fukushima.jp/dynamic/graph_top.html>
[26] IT Project in the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST) (FY2010-2014, total project funding estimated at US$39M), PI: Prof. Masaru Kitsuregawa of University of Tokyo, http://www.tkl.iis.u-tokyo.ac.jp/top/modules/project1/index.php?id=10.
[27] Global Sensor Network software home page: http://sourceforge.net/apps/trac/gsn/.
[28] DIAS, Data Integration and Analysis System for the Earth Observation and Ocean Exploration System in Japan <http://www.editoria.u-tokyo.ac.jp/projects/dias/english/index.html>.
[29] Health Map http://www.healthmap.org/en/.
[30] Radiation Response Volunteer Corps (in the state of Oregon) that organizes volunteers to collect radiation data https://public.health.oregon.gov/HEALTHYENVIRONMENTS/RADIATIONPROTECTION/EMERGENCYRESPONSE/Pages/volunteercorps.aspx.
[31] Center for Coastal Margin Observation & Prediction (CMOP, an NSF Science and Technology Center) at http://www.stccmop.org/.
[32] Software tools from several projects at the Florida International University led by Naphtali Rishe, including the Terrafly project http://terrafly.fiu.edu/, and two RAPID projects for handling the data from the 2010 Gulf oil spill, one funded by the CISE/IIS/III program cluster and another by CISE/CNS division.
[33] National Institute for Informatics, Japan. http://en.wikipedia.org/wiki/National_Institute_of_Informatics (short explanation) and official web site at http://www.nii.ac.jp/en/.
[34] Foursquare https://foursquare.com/.
[35] Twitter data analysis example, a blog by Twitter writer http://blog.twitter.com/2010/02/super-data.html.
[36] Palen, L., and Liu, S. (2007). Citizen Communications in Crisis: Anticipating a Future of ICT-Supported Participation. Proceedings of the ACM Conference on Human Factors in Computing Systems CHI 2007, 727-736.
[37] Palen, L., and Vieweg, S. (2008). The Emergence of Online Widescale Interaction in Unexpected Events: Assistance, Alliance and Retreat. In the 2008 ACM Proceedings of Computer Supported Cooperative Work Conference.
[38] Hughes, A. and Palen, L. (2009). Twitter Adoption and Use in Mass Convergence and Emergency Events. Proceedings of the 2009 Information Systems for Crisis Response and Management Conference (ISCRAM 2009), Gothenberg, Sweden.
[39] Vieweg, Sarah, Amanda Hughes, Kate Starbird, Leysia Palen (2010). Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In Proceedings of the 28th International Conference on Human Factors in Computing Systems (Atlanta, Georgia, USA, April 10 – 15, 2010). CHI 2010. ACM, New York, NY, 1079-1088.
[40] Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. (2010). Chatter on the Red: What hazards threat reveals about the social life of microblogged information. In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (Savannah, Georgia, USA, February 06 – 10, 2010). CSCW 2010. ACM, New York, NY, 241-250.
[41] Verma, Sudha, Sarah Vieweg, Will Corvey, Leysia Palen, Jim Martin, Martha Palmer, Aaron Schram and Ken Anderson. NLP to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In the Fifth International AAAI Conference on Weblogs and Social Media, 17-21 July 2011, Barcelona, Spain.
[42] Starbird, Kate and Leysia Palen (2012). (How) Will the Revolution be Retweeted?: Information Propagation in the 2011 Egyptian Uprising. 2012 ACM Conference on Computer Supported Cooperative Work, Bellevue, WA.
[43] Sarcevic, Aleksandra, Leysia Palen, Joanne White, Mossaab Bagdouri, Kate Starbird, Kenneth M. Anderson, (2012). "Beacons of Hope" in Decentralized Coordination: Learning from On-the-Ground Medical Twitterers During the 2010 Haiti Earthquake 2012 ACM Conference on Computer Supported Cooperative Work, Bellevue, WA.

[44] Starbird, K. and Palen, L. (2010). Pass It On?: Retweeting in Mass Emergencies. Presented at the 7th International Information Systems for Crisis Response and Management Conference (Seattle, WA, USA, May 2010). ISCRAM 2010.
[45] Palen, L., Starbird, K., Vieweg, S. and Hughes, A. (2010). Twitter-based information distribution during the 2009 Red River Valley flood threat. Bulletin of the American Society for Information Science and Technology, American Society for Information Science and Technology, Volume 36, Issue 5, (June/July 2010), pp. 13-17.
[46] Palen, L., Vieweg, S., and Anderson, K. (2011). Supporting "Everyday Analysts" in Time- and Safety-Critical Situations. The Information Society Journal, 27(1), pp. 52-62.
[47] Barrenechea, M., Barron, J., White, J. (2012). No Place Like Home: Pet-to-Family Reunification after Disaster. Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems.
[48] From the Web to the Grid and Beyond, Computing Paradigms Driven by High-Energy Physics, Brun, R.; Carminati, F; and Galli Carminati, G. (Eds.), Springer, 2012, 358p.
[49] Tsugawa, M.; Figueiredo, R.; Fortes, J.; Hirofuchi, T.; Nakada, H.; and Takano, R., "On the use of virtualization technologies to support uninterrupted IT services: A case study with lessons learned from the Great East Japan Earthquake," in 2012 IEEE International Conference on Communications (ICC), 2012, pp. 6324-6328.
[50] Takano, R.; Nakada, H.; Hirofuchi, T.; Tanaka, Y.; Kudoh, T., "Cooperative VM migration for a virtualized HPC cluster with VMM-bypass I/O devices," *E-Science (e-Science), 2012 IEEE 8th International Conference on* , vol., no., pp.1,8, 8-12 Oct. 2012. doi: 10.1109/eScience.2012.6404487.
[51] M. Arlitt AND T. Jin, "Workload characterization of the 1998 World Cup Web site," Technical report, Hewlett-Packard Laboratories, September 1999.
[52] Holguín-Veras, J., Taniguchi, E., Pedroso, F.F., Jaller, M. and Thompson, R.G., The Tohoku Disasters: Preliminary Findings Concerning the Post Disaster Humanitarian Logistics Response, The 91st Annual Meeting of Transportation Research Board, Washington, DC, DVD, 2012.
[53] Google Crisis Response. [http://www.google.org/crisisresponse/].
[54] Google Person Finder [http://www.google.org/personfinder].
[55] Google Crisis Map [http://www.google.org/crisismap].
[56] Japan Radiation Map (derived from the SPEEDI data set) [http://jciv.iidj.net/map/].