3 United States Government Efforts toward Big Data Research and Development

p.37

Research and development revolving around "big data" is currently making huge strides in the West. The term does not quite have an exact definition, but is rather a general way of referring to large quantities of digital data. Some specific examples are: online info that has experienced enormous growth due to the spread of social networking services (SNS) and the like; large amounts of photos and videos stored on the internet; information from "things" which is information detected by sensors and sent by communication devices; massive amounts of numerical data generated by supercomputers and so forth. A recent trend is attempts at creating new value by extracting significant information from a vast amount of digital data that has gone unused until now.

In March 2012, the United States government announced an R&D initiative to utilize big data. Six government agencies are investing over US$200 million to try and improve technologies for handling vast amounts of digital data. They see big data as science and technology that may contribute to the creation of a new paradigm and will have a very major impact in many realms, as the internet did. Noteworthy aspects of the U.S. government's focus are subjects such as the "Focus on Visualization Technology," the "Relationship with Cloud Computing," "Considerations to Human Resource Development," "Active Participation by Industry and Academia" and "Encouraging Data Sharing." We can also see that the U.S. is giving thought to policies such as those encouraging the provision of facilities and computing power conducive to collaborative work. Since the results produced by the initiative's R&D may become widely used and penetrate to society in a few years or decades, and lead to major innovations, future developments will be worthy of our attention.

Comprehensive solutions to numerous problems are needed for value creation from big data. In particular, there is an intimate relationship between analysis and visualization, and it is important to visualize and extract knowledge, then link that to action that creates value. Considering the growth of R&D into big data around the world and the difficulty of the challenges involved, global collaboration will be vital for future R&D.

(Original Japanese version: published in September/October 2012)

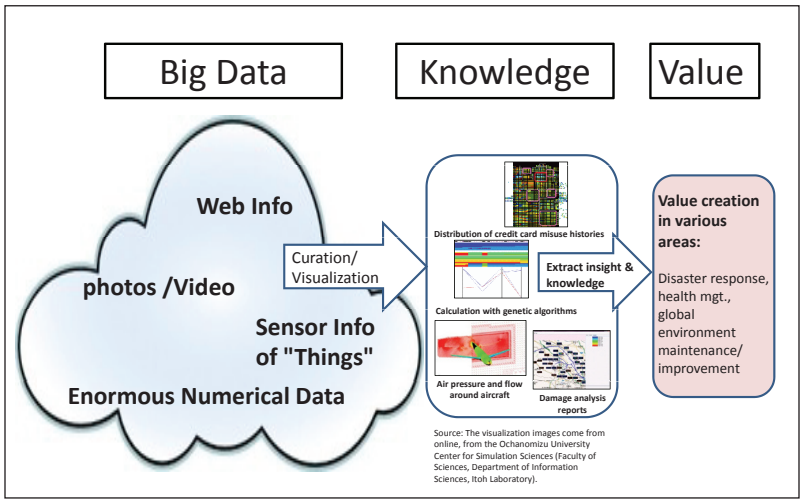

Figure : Creating Value from Big Data

Source: Compiled by the Science and Technology Foresight Center.

3

United States Government Efforts toward Big Data Research and Development

Minoru NOMURA
Affiliated Fellow

1 Introduction

Research and development revolving around "big data" is currently making huge strides in industry, academia and national governments. The term does not quite have an exact definition, but is rather a general way of referring to large quantities of digital data. If we simply think of the extraction of significant information from data as a function which conventional IT techniques excel at, then that is nothing particularly new. However, the amount of data, the speed at which it is being produced and its diversity are all undergoing extraordinary change. In the wake of this change, great advances in how we handle this data—storing and processing—are occurring, and utterly new trends are appearing. New R&D is aiming for creating new value by extracting significant information from a vast amount of data that traditionally was stored but went unused. This shift is the reason why big data is attracting such great interest.

In March 2012, the United States government announced an R&D initiative to utilize big data. The Obama administration's science and technology policies promote five initiatives,[1] and big data is given as one of them. Six government agencies are investing over US$200 million to push forward the big data initiative and try to improve the technologies needed to handle vast amounts of digital data. The most interesting aspect of the initiative is that it sees big data as something that will have as much of an impact on the world as the internet has. In other words, the administration sees big data as something that will have an incredibly huge effect in many areas.

The U.S. is not the only country interested in big data. In Europe, an EU project is trying to find a clear answer for all of Europe regarding issues of growing data in the scientific research community.

In this paper, Chapter 2 presents what big data is, Chapter 3 introduces the R&D initiative launched by the U.S. government, and Chapter 4 explores particularly noteworthy aspects.

Furthermore, discussion about big data in Japan, too, has raised many issues in need of solutions, such as personal information and security problems. One view is that "we will create groundbreaking ideas if we arrange laws and then create circumstances that will not hold back companies' utilization of big data."[2] However, the reason why big data has garnered such worldwide attention is that information has undergone such explosive growth that anyone can access it freely. Accordingly, this paper will leave it to other papers to discuss problems such as security and the like, and instead will mainly address the potential that big data will likely fulfill, primarily conveying positive trends in the U.S.

[NOTE 1] The Japanese government has provided support since 2005 such as: New IT Infrastructure for the Information-explosion Era (info-plosion) by the Ministry of Education, Culture, Sports, Science and Technology (MEXT); the Information Grand Voyage Project by the Ministry of Economy, Trade and Industry (METI); the Development of the Fastest Database Engine for the Era of Very Large Database and Experiment and Evaluation of Strategic Social Services Enabled by the Database Engine under the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST) program by the Japan Society for the Promotion of Science. Recently, documents from the National Policy Unit, the Council for Science and Technology Policy and MEXT have produced material such as the following:

- The "Strategy for Rebirth of Japan – Opening Frontiers to a `Country of Co-Creation,'"[3] was approved on July 30, 2012 by the National Policy Unit and on July 31 by the Cabinet. Chapter 2 of "IV. Specific Policies

for the Rebirth of Japan," "its efforts to establish as a vital policy a solid information and communications infrastructure and thorough use of information and communications technology," states that the government "plans to promote the use of data in the government's possession through: diverse and large amounts of data (big data) that has made collection possible in accordance with advances in information and communications technology; and converging of different fields using information and communications technology."

- Reference material 1-2-2 of the 103rd Plenary Meeting of the Council for Science and Technology Policy held on July 30, 2012, the "FY 2013 – Prioritized Issues and Initiatives of the Priority Policy Package,"[4] listed one prioritized initiative as "promoting the development, standardization and spread of basic technology for the utilization of large-scale information (big data)."
- Document 2 of the 77th meeting of the MEXT Information Science and Technology Committee on July 5, 2012, "Academia's Challenge in the Era of Big Data – Academic Cloud Investigative Commission,"[5] states: "In order to maximize the potential of big data, there is a need, in terms of interdisciplinary collaboration, international collaboration and cultivation of human resources, to pay adequate attention to and promptly get started on those including: R&D in the fields of information, science and technology that will contribute to advancements in data science; R&D relating to big data in system research and the like, to construct an academic cloud environment; and projects relating to the construction of big data utilization models in R&D corporations ."

2 What is Big Data?

2-1 Big Data is...

"Big data" does not quite have an exact definition, but is rather a general way of referring to voluminous digital data. This data is not all collected from one place. It is a variety of data coming from many different areas. Some specific examples are: web info that has experienced enormous growth due to the spread of social networking services (SNS) and the like; large amounts of photos and videos stored on the internet; information from "things" which is information detected by sensors and sent by communication devices; massive amounts of numerical data generated by supercomputers and so forth. This data is becoming so voluminous and complex that conventional techniques cannot manage it.

Most big data is text, images, sensor data and the like. This data is rapidly growing on the internet day by day due to more large pieces of video data being uploaded to websites, in addition to the expanding use of SNSs such as Facebook and Twitter. Furthermore, the "Internet of Things" (IOT) is growing and taking shape. This is the idea of connecting and networking various "things" through the web.

Figure 1 contains excerpts from various materials showing the growth in the volume of data. Here, the vertical axis shows the amount of data on a logarithmic scale. It already shows this amount reaching the zettabyte level, expressing 10^{21}. And for the past few years it has been growing exponentially. Plus, it seems that this momentum will not let up.

2-2 Factors behind the Increasing Scale of Data

Factors we can cite behind the rapid expansion in the scale of data are: the improved ease by which web data can be gathered compared to before; the improved ease by which data can be gathered from devices (data collection from mobile phones, etc.) and things; and advances in storing and processing technologies capable of handling large amounts of data.

To provide an example of gathering web information, creating a database of web information gathered to run a search engine was primarily done manually by people until sometime in the latter half of 1994. However, the global growth of the web demonstrated the limits of this approach. The appearance of programs called web crawlers was a breakthrough. These programs periodically obtain text, images and such from the internet, allowing this data to be automatically collected and organized into a database.[11] Smaller and cheaper communications components and sensors needed to collect data from devices and "things" have made data collection easier. For example, chips have been miniaturized from 10 mm^2 in 2000 to 2-3 mm^2 in 2010, while the average retail price has dropped from around 240 yen in 2000 to 56 yen in 2010. Meanwhile, prices for communication modules that send data collected by sensors and the like continue to fall, while the number of service

subscribers has increased.[12]

Recently, many are using Hadoop to process data. Under a situation where enormous dataset is separated and stored in a distributed processing environment comprising numerous computers, Hadoop offers a way to operate in parallel to process the stored data. This open-source software framework is based on Google's MapReduce.[13] Many commercial versions are available to solve various problems through practical application of the software. Meanwhile, the relational database (RDB) has been the optimal compiler technique for over 20 years. At present, both Hadoop and RDBs are in use.

Now, as the collection and accumulation/processing of vast amounts of data is becoming possible, the key is becoming what value we can produce from that data and whether we can connect that to creating new industries and solving social issues. This is the biggest reason for the attention given to big data.

2-3 What People Want from Big Data Analysis

The main characteristics big data has in common are volume (large), speed (real-time) and types (variety).[12]

Regarding large volume, if the size of the data poses a problem, then one could handle it on a smaller scale through sampling, but that ends up giving you a look at only a portion or you may miss seeing something important.[14] Data mining finds conspicuous patterns in a large mass of data and divides that mass into distinctive groups to process and extract information from it. For example, it is easy to think of how this method would help a convenience store find a pattern showing an arrangement of products on its shelves that leads to more sales. Because big data involves such large quantities, it contains common conspicuous patterns along with rare ones. Discovering these rare patterns is one thing businesses want from big data.

Additionally, big data has other potential. Up to now, the study of physics has typically followed a pattern of observation, creating a "formula" from the law inherent to the observed action, which then comes into general usage to recreate physical phenomena. As an example in the case of airplane, a person runs

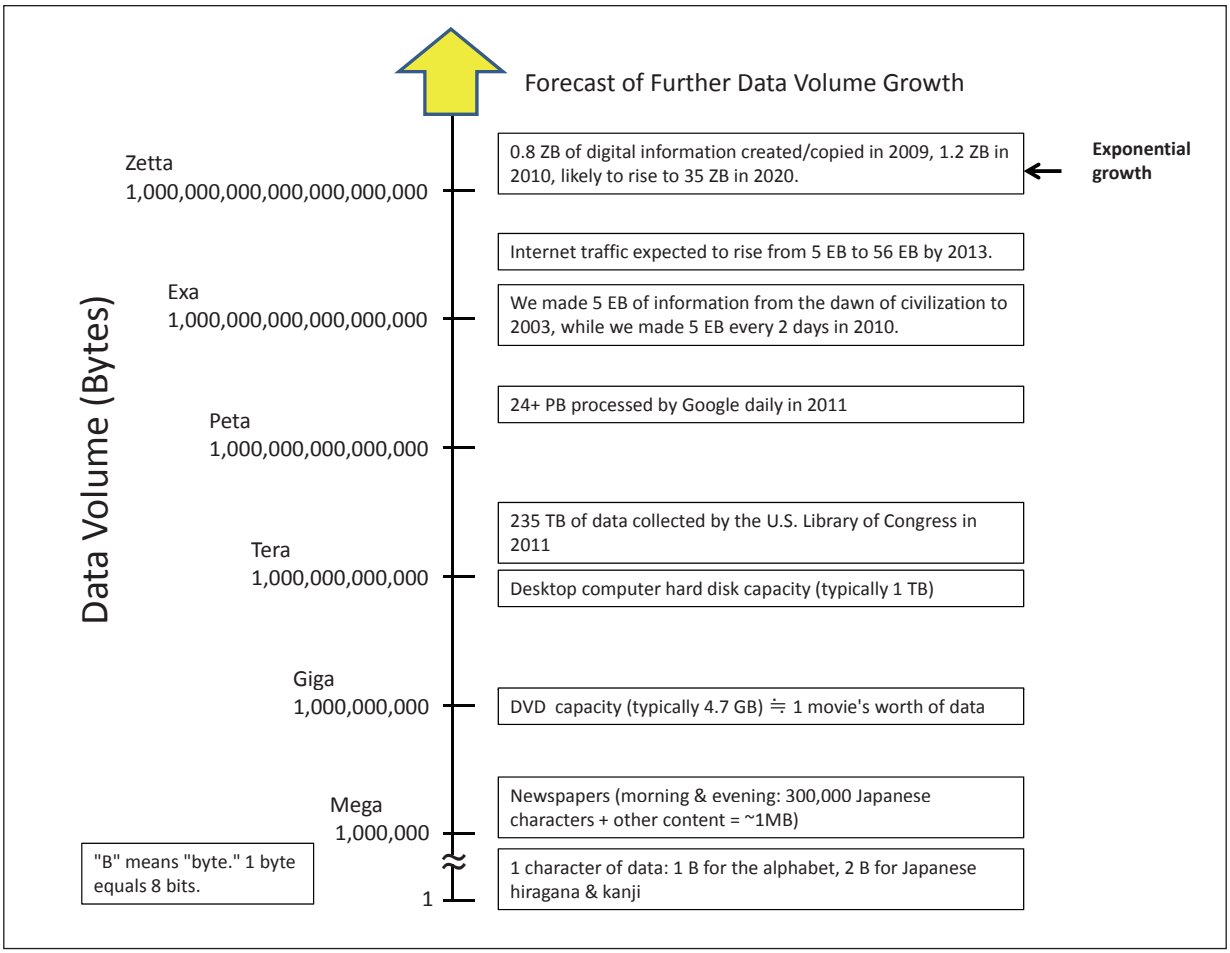

Figure 1 : The Increasing Volume of Data

Source: Compiled by the Science and Technology Foresight Center based on References #6-10.

a simulation employing a formula governing the movement of fluids and analyzes the movement. However, when the movement reaches high speeds, the formula does not work anymore. We can say that big data analysis is a way to distill knowledge with a method other than the accepted, standard "formula." However, more data is required for this purpose. This is because we need enough data to find a recurring pattern.

As for real-time analysis and diversity, data is input and collected in real time because we can now collect it from more devices and "things" than before. Accordingly, output and feedback via instant processing will become more important. Stream computing[15] is one of the examples that automatically analyzes a constant inflow of data, estimates and leads to rapid decision-making.

Looking at big data from a different perspective, we see that it contains more ambiguity and uncertainty as its quantity increases. During analysis, one must take into account the many uncertain pieces of data.[16] This could become a crucial point for big data analysis in the future.

2-4 Embryonic Cases of Big Data Analysis

This section provides embryonic examples of big data analysis to offer a look at what sort of value they are creating.

2-4-1 Household Health Management

One part of the research conducted by the Information Grand Voyage Project in Japan is an experimental study that uses sensors to research at-home care. According to a report, the project "gets diabetics to measure their blood sugar at home on a continuous basis by using blood sugar-monitoring sensors and acceleration sensors to measure how much they exercise. The research participants are provided with timely messages to encourage certain behavior such as exercise and reducing the amount of food they consume. These messages are called `information medicine.' When information medicine was administered, it demonstrated that providing timely information that encourages certain behavior is as effective as regular medicine because it successfully suppresses a rise in the participants' blood sugar level."[17] This research is focused on individual health management, but it likely has the potential to develop into broader, public health management.

2-4-2 Highly Accurate Translation

Google Translate is an example of how using big data can improve translation accuracy. A free translation service that can instantly translate into 64 languages, it can automatically translate any combination of vocabulary, sentences and webpages from and into any pair of those languages. This is not a conventional method of automated translation employing a prescribed dictionary or grammar rules. To quote an excerpt from Google's description of the service: "When Google Translate generates a translation, it looks for patterns in hundreds of millions of documents to help decide on the best translation for you. By detecting patterns in documents that have already been translated by human translators, Google Translate can make intelligent guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called `statistical machine translation.'"[18] This is an example of how having more data produces more significant results.

2-4-3 Road Traffic Information

Big data has been effective at providing road traffic information during normal as well as unusual times. Immediately after the Tohoku Pacific Offshore Earthquake of 2011, GPS data was used to provide road traffic information, demonstrating it is very good at streamlining logistics, such as the transport of relief supplies. Probe information is used for this, which includes information such as the positions and speeds of individual vehicles on the road. ITS Japan uses traffic log data collected anonymously and statistically by four private companies to produce a map showing comprehensive probe traffic information. On the same map, information on road closures based on "the Tohoku Region Restricted Road Access Information and Disaster Information Composite Map" produced by the Geospatial Information Authority of Japan (GSI) was superimposed. This is an example of a public-private partnership providing timely traffic histories and road closure information in a disaster area.[19]

2-4-4 Research on Disaster Response Measures

On June 5, 2012, Japanese MEXT Minister Hirofumi Hirano forged an agreement with Subra Suresh, Director of the National Science Foundation (NSF) in the United States, on the importance of U.S.-

Japan cooperation on disaster research. In essence, the agreement covers research collaboration and assistance through the use of big data in a wide range of fields such as computer science, engineering, social science and earth science. These efforts will be concerned with robustness and resilience against disasters. Specific research fields with promise are given below.[20]

- Advances in applications, such as analysis, modeling and numerical analysis and hazard probability models, by using large amounts of data obtained from disasters.
- Improvements to information technology resilience and responsiveness to allow for real-time data sensing, visualization, analysis and forecasting that is essential to instant decision-making.
- To prepare for times of emergency, integration of diverse knowledge such as inputs from numerous academic disciplines and end users, and large quantities of data from all information sources.

2-4-5 Solutions Development in Industry

Many examples of employing sensor data can be found in industry. Examples include Bridge Monitoring,[21] Agricultural Produce and Management Visualization and Production Process Improvement,[22] services using Context Awareness Technology[23], Energy Management Systems(EMS) and the like.

Other examples include "recommendations" for products and services employing purchase log information from websites and SNS information, as well as services providing information linked directly to GPS positional information and sales support.

It is assumed that these products and services will grow and change more and more as big data technology improves, and this should take solutions development in industry another step further.

3 U.S. Government Big Data Initiative

The Obama administration's science and technology policies promote five initiatives,[1] and big data is given as one of them. This chapter introduces the Big Data initiative launched by the U.S. government.

3-1 The OSTP Big Data Initiative

The Office of Science and Technology Policy (OSTP) announced an R&D initiative to utilize big data,[24] for which U.S. federal agencies are making new investments of more than US$200illion. The goals are to help solve pressing national issues and increase capabilities for drawing insights and knowledge from large and complex sets of digital data. First, six agencies (the NSF, NIH, DoD, DARPA, DOE and USGS) will invest in research to improve tools and technologies for handling big data. Assistant to the President and Director of the White House OSTP Dr. John P. Holdren said, "In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security."

The following R&D objectives have been cited by this initiative:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning.
- Expand the workforce needed to develop and use Big Data technologies.

This initiative was described as a response to advice from the President's Council of Advisors on Science and Technology in 2011, which concluded that "the Federal Government is under-investing in technologies related to Big Data."[24]

U.S. government agencies have already initiated various efforts related to big data. On March 29, 2012, on the same day it announced the R&D initiative, the OSTP released a "Fact Sheet"[25] on big data. With this document, the OSTP cited many examples to highlight ongoing government programs that are addressing the challenges of the "big data revolution" to advance agency missions and innovation through scientific discovery. Table 1 shows the agencies and number of programs listed in the Fact Sheet.

The following subsections present six agencies implementing the Big Data Initiative. (Some are also listed in the above-mentioned Fact Sheet.)

Table 1 : Agencies and Number of Programs Listed in the Fact Sheet

Agency	Number of Programs
Department of Defense (DoD)	10
Department of Homeland Security (DHS)	1
Department of Energy (DoE)	12
Department of Veterans Affairs (VA)	9
Health and Human Services (HHS)	5
Food and Drug Administration (FDA)	1
National Archives and Records Administration (NARA)	1
National Aeronautics and Space Administration (NASA)	7
National Institutes of Health (NIH)	23
National Science Foundation (NSF)	16
National Security Agency (NSA)	3
United States Geological Survey (USGS)	1

Source: Compiled by the Science and Technology Foresight Center based on Reference #25.

3-1-1 NSF & NIH Joint Support

The National Science Foundation (NSF) and the National Institutes of Health (NIH) are conducting R&D on core technologies to advance big data science and engineering. To be specific, they released a joint "big data" solicitation for NSF and NIH support to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large and diverse datasets. The purpose is to speed up scientific discoveries and create new fields of inquiry that cannot be explored by other means. The NIH is taking a particular interest in imaging, molecular, cellular, electrophysiological, chemical, behavioral, epidemiological, clinical, and other data sets related to health and disease.

3-1-2 NSF

In addition to the NSF's continual focus on basic research through the aforementioned big data solicitation, the agency has declared a comprehensive, long-term strategy that includes new methods to derive knowledge from data, and to construct new infrastructure to manage, curate (see NOTE 2 below) and serve data to communities, as well as new approaches for education and training. Specific aspects of the strategy are given below.

●Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers.

●Funding a $10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing.

●Providing the first round of grants to support "EarthCube" – a system that will allow geoscientists to access, analyze and share information about our planet.

●Issuing a $2 million award for a research training group to support training for undergraduates to use graphical and visualization techniques for complex data.

●Providing $1.4 million in support for a focused research group of statisticians and biologists to

[NOTE 2] The meaning of "curate" and "curation"
The terms "curate" and "curation" hold significant meaning in terms of creating value from big data. In this case, curation is gathering information from the internet or classifying gathered data, connecting pieces of data to give it new value, and then sharing it. Japanese dictionaries define the English word "curation" as "collecting information and such, organizing it, adding value from new perspectives, and sharing that information with others." (from http://kotobank.jp/word/%E3%82%AD%E3%83%A5%E3%83%AC%E3%83%BC%E3%82%B7%E3%83%A7%E3%83%B3)
(from http://www.nttpc.co.jp/yougo/%E3%82%AD%E3%83%A5%E3%83%AC%E3%83%BC%E3%83%88.html)

determine protein structures and biological pathways.

- Convening researchers across disciplines to determine how Big Data can transform teaching and learning.

NSF Director Subra Suresh has said, "American scientists must rise to the challenges and seize the opportunities afforded by this new, data-driven revolution. The work we do today will lay the groundwork for new enterprises and fortify the foundations for U.S. competitiveness for decades to come."

NSF also states "in the near term, NSF will provide opportunities and platforms for science research projects to develop the appropriate mechanisms, policies and governance structures to make data available within different research communities".[26]

3-1-3 DoD

The Department of Defense (DoD) is starting up programs under the Data to Decisions initiative, and is "placing a big bet on big data." It is investing approximately US$250 million annually (allotting US$60 million to new research projects) across the Military Departments in a series of programs. Programs are listed below.

- Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly autonomous systems that can maneuver and make decisions on their own.
- Improve situational awareness to help warfighters and analysts and provide increased support to operations. The Department is seeking a 100-fold increase in the ability of analysts to extract information from texts in any language, and a similar increase in the number of objects, activities, and events that an analyst can observe.

To accelerate innovation in Big Data that meets these and other requirements, DoD will announce a series of open prize competitions over the next several months.

3-1-4 DARPA

The Defense Advanced Research Projects Agency (DARPA) develops computational techniques and software tools to analyze large quantities of data comprising both semi-structured data (e.g. tabular, relational, categorical, meta-data, etc.) and unstructured data (e.g. text documents, message traffic, etc.). DARPA is launching the XDATA Program, with plans to invest roughly US$25 million a year over a four-year period. The main research subjects are:

- Development of scalable algorithms that process imperfect data within distributed data stores.
- Creation of effective interaction tools between humans and computers to facilitate rapidly customizable visual reasoning for diverse missions.

The XDATA Program will support an open source software kit to allow for flexible software development.

The following is a slightly more detailed introduction to the XDATA Program, which is slated to receive a heavy dose of investment.[27]

Since technology development will also be guided by end-users with operational support expertise, DARPA will engage elements of the DoD and other agencies to develop use cases and operational concepts. This will result in a "development-in-process" software development model, where agile libraries, APIs, and code instances will be refined based on user feedback. User groups of selected personnel from the DoD and other agencies will be maintained throughout the life of the program.

The XDATA Program is placing importance on developing fast, scalable and efficient methods in order to process and visualize data. And it is not only support ingestion and transformation but also enable fast search and analysis.

The program comprises four "Technical Areas":

TA1: Scalable analytics and data processing technology
TA2: Visual user interface technology
TA3: Research software integration
TA4: Evaluation

The program intends to maintain a technology integration facility in the greater Washington, DC, area to facilitate agile and collaborative software development, integration, and testing/evaluation. User interaction, use-case development, and integration, test, and evaluation are intended to take place at this facility.

3-1-5 NIH

In addition to the aforementioned joint development of core technologies with the NSF, the National Institutes of Health (NIH) is announcing that the world's largest dataset on human genetic variation produced by the international 1000 Genomes project (see NOTE 3 and Reference #28) has been made possible by cloud computing (hereafter referred to as

the "cloud").[28] In collaboration with Amazon.com, Inc. (hereafter referred to as "Amazon"), the datasets are already available to the public via Amazon Web Services' (AWS) cloud. The size of this data is 200 terabytes, enough to fill 16 million file cabinets worth of documents or more than 30,000 standard DVDs. The 1000 Genomes project's current data set is a prime example of big data. Because the quantity will further expand, there are still few researchers at present with the computational capabilities to optimally utilize this data. AWS is hosting the project's data, allowing researchers free and public access to the datasets. They only have to pay for the computational services they themselves use.

3-1-6 DoE

The Department of Energy (DoE) is using part of a five-year funding allotment of US$25 million to establish the Scalable Data Management, Analysis and Visualization Institute (SDAV) through the Scientific Discovery Through Advanced Computing (SciDAC) program.[29] At SDAV, the Lawrence Berkeley National Laboratory is leading five other national laboratories, along with seven universities, to combine the partners' expertise. The goal is to develop new tools to help scientists manage and visualize data on the Department's supercomputers, which will further streamline the processes that lead to discoveries made by scientists using the Department's research facilities. The need for these new tools has grown as the simulations running on the Department's supercomputers have increased in size and complexity.

According to the reference[29], reasons for establishing SDAV are described as follows.

As the scale of computation has exploded, the data produced by these simulations has increased in size, complexity, and richness by orders of magnitude, and this trend will continue. However, users of scientific computing systems are faced with the daunting task of managing and analyzing their datasets for knowledge discovery, frequently using antiquated tools more appropriate for the teraflop era. While new techniques and tools are available that address these challenges, often application scientists are not aware of these tools, aren't familiar with the tools' use, or the tools are not installed at the appropriate facilities.

In order to respond to these issues, SDAV is developing and preparing technical solutions in the three fields of data management, analysis and visualization, with the goal of using these solutions to help scientists in all disciplines.

3-1-7 USGS

The U.S. Geological Survey (USGS) is looking at big data for Earth system science. The agency is already issuing grants through the John Wesley Powell Center for Analysis and Synthesis. The Center catalyzes innovative thinking in Earth system science by providing scientists a place and time for in-depth analysis, state-of-the-art computing capabilities, and collaborative tools invaluable for making sense of huge data sets. Big data projects in Earth system science will improve understanding of issues such as species response to climate change, earthquake recurrence rates, and the next generation of ecological indicators.

3-2 Establishment of the BD SSG

Up to now the U.S. government has conducted R&D on information and communications technology with the Networking and Information Technology Research and Development (NITRD) program formulated by the National Science and Technology Council (NSTC). Fifteen government agencies are involved in the NITRD program. It contains seven individual research fields called Program Component Areas (PCAs) and five Senior Steering Groups (SSGs) that handle priority issues requiring interagency collaboration. Individual programs run by each agency are done in collaboration with other agencies. The "Bluebook" ("Supplement to the President's Budget") relating to NITRD program plans and budgets is published annually.

The overview sates the followings.

The Big Data Senior Steering Group (BD SSG) was established in early 2011,[30] making big data an area for interagency collaboration.

The BD SSG has been formed to identify current big data research and development activities across the Federal government, offer opportunities for coordination, and begin to identify what the goal of a national initiative in this area would look like. The

[NOTE 3] The 1000 Genomes project is a public-private consortium formed in 2008 to create a detailed map of genome variation between more than 2,600 people from 26 populations from around the world.

group was established due to increasing concerns over data preservation, access, dissemination, and usability as data volumes grow exponentially. Research into areas such as automated analysis techniques, data mining, machine learning, privacy, and database interoperability are underway at many agencies. Studying this will help identify how big data can enable science in new ways and at new levels.

According to records, BD SSG was formed to identify programs across the Federal government and bring together experts to help define a potential national initiative in this area. BD SSG has been asked to identify current technology projects as well as educational offerings, competitions, and funding mechanisms that take advantage of innovation in the private sector. In this function, records also show that the BD SSG collects information on current activities across the Federal Government, creates a high-level vision of the goals of a potential national initiative, develops the appropriate documents and descriptions to aid discussion within the government, and where appropriate, the private sector, and develops implementation strategies that leverage current investments and resources.

4 Noteworthy Aspects of the U.S. Government Big Data Initiative

This chapter cites particularly noteworthy aspects of the U.S. government's big data initiative described

Tabele 2 : Key Points and Research Areas of U.S. Gov't Agency Big Data Research Support

Agency	Technological Development	Education & Training	Data Sharing	Policy Incentives (Provision of facilities, recruiting researchers, etc.)
NSF & NIH	- Core technologies to manage, analyze, visualize and extract information from big data			
NSF	- Integrate machine learning, cloud computing and crowd sourcing, and extraction of information from data ($10 mil)	Encourage research universities for development of graduate programs to train data scientists and engineers Support a focused research group of statisticians and biologists to determine protein structures and biological pathways ($1.4 mil) Support acquisition of visualization techniques, subsidies to teach undergrad students how to handle big data ($2 mil)	"EarthCube": helps geoscientists access, analyze and share data on the Earth	Convene researchers across disciplines to determine how Big Data can transform teaching and learning
DoD	- Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly autonomous systems -Improve situational awareness to help warfighters and analysts and provide increased support to operations (100x capability upgrade to extract info from text in any language, and a similar increase in the number of objects, activities, and events that an analyst can observe) ($250 mil, of which $60 mil for new projects)			Open contests over several months with prizes for winners to accelerate innovation with big data
DARPA	Development of software tools and computational methods to analyze big data (XDATA Program) - Development of scalable algorithms that process imperfect data within distributed data stores - Creation of effective interaction tools between humans and computers to facilitate rapidly customizable visual reasoning for diverse missions - Development of flexible software by offering an open source software kit ($25 mil annually over 4 years)			Maintain a technology integration facility to facilitate agile and collaborative software development, integration, and testing/evaluation in XDATA Program
NIH			Free access to human genome information on Amazon's AWS cloud	
DoE	- Establishment of the SDAV; development and growth of supercomputer data management, analysis and visualization tools ($25 mil over 5 years)			
USGS				Provision of a place, time and computational power for analysis of the Earth, the environment, the climate, etc.

Source: Edited by the Science and Technology Foresight Center based on Reference #24.

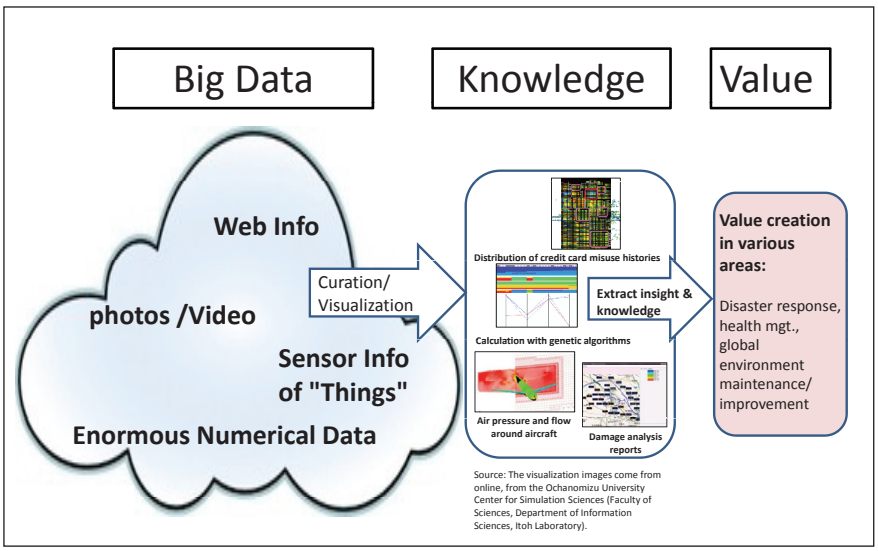

Figure 2 : Creating Value from Big Data

Source: Compiled by the Science and Technology Foresight Center.

in Chapter 3. Table 2 shows agencies discussed in Chapter 3 and the research support they are providing to each research subject area.

4-1 How to Look at Big Data

Assistant to the President and Director of the White House OSTP Dr. John P. Holdren believes big data may have as great an impact on the world as the internet. In other words, he sees it as having a very large influence on science and technology in general as well as society. "Academia's Challenge in the Era of Big Data – Academic Cloud Investigative Commission,"[5] a document released by MEXT, describes the U.S. view on big data as below.

"The concept behind the United States' Big Data Initiative is to provide powerful solutions to the important technological problems of the future. The U.S. government sees big data is the same importance as both supercomputers and the internet. It believes that improving the core technology behind massive data will contribute to security, educational reforms and training, and the NSF sees that technical issues to confront are graduate courses to train data scientists, machine learning, cloud computing and crowd-sourcing."

4-2 Focus on Visualization Technology in Technical Development

The contents related to visualization are seen most common among the table 2. In short, comprehensive solutions to numerous problems are needed to create value from big data, among them, visualization may be considered as the most important technological aspect. It also believes that there is an intimate relationship between analysis and visualization and that processing which does not relate to visualization will have trouble creating value. It is important to visualize and extract knowledge, then link that to action that creates value. It may be said that this is source of thinking (see Figure 2).

Visualization is incorporated into the joint research by the NSF and NIH, as well as by DARPA, the DoE, indicating that these agencies recognize its importance.

4-3 Relationship with Cloud Computing

A previous report in the June 2010 issue of Science and Technology Trends has more detail on "Cloud Computing," but the spread of the cloud is one factor behind big data's advance. For example, NIH projects use Amazon's cloud services. The primary significance of access to the cloud is that it is easy and low cost burden for researchers to use. One description of the NIH's 1000 Genomes project is: "AWS is hosting the 1000 Genomes project as a publically available dataset for free, and researchers will pay for the computing services that they use."

It goes without saying that big data involves huge quantities of data. In order to process it, a large quantity of disks and computers are needed. And if data is not input or output in parallel, then too much time is taken. Input or output time is shortened by providing multiple disks in multiple computers which can be accessed in parallel. Cloud is an essential

element for big data as a means of providing a large quantity of both disks and computers.

4-4 Considerations to Human Resource Development

Human resource development is vital because mathematical, statistical and legal knowledge, and business management expertise are needed to find value within big data. The NSF's description of its programs include preparation such as development of graduate programs to cultivate data scientists and engineers, support for research groups comprising statisticians and biologists, and grants for research to cultivate undergraduate students by helping them to acquire visualization techniques.

Although not limited to big data, many Western projects embrace this sort of education. As examples from Europe, in FP7 projects and the PRACE project[31] that provide high-end supercomputers to researchers throughout Europe, specialists provide the training (or education) and public relations.

4-5 Active Participation by Industry and Academia

OSTP Deputy Director for Policy Tom Kalil is calling for private companies and universities to actively participate in big data efforts. He wrote, "We also want to challenge industry, research universities, and non-profits to join with the Administration to make the most of the opportunities created by Big Data. Clearly, the government can't do this on its own. We need what the President calls an "all hands on deck" effort."[32] A real-world example of this happening is Amazon's collaboration with the NIH's 1000 Genome project.

One problem in R&D is that when a project comes to a close, research is discontinued and the chances of the results being put to use for the public decrease. We need ways to continue and expand R&D so that society can utilize it.

4-6 Encouraging Data Sharing

The Big Data Initiative is encouraging the sharing of data. For example, at the NIH, open access to the datasets produced by the 1000 Genomes project is available through the AWS cloud. In addition, at the NSF, the EarthCube which assists any scientist in accessing, analyzing and sharing data on the Earth is supported.

The United States has a website, Data.gov,[33] that launched on May 21, 2009 and allows retrieval of information and data held by government agencies. Although the purpose is not data sharing for the researcher community, the goal of Data.gov is to provide the public with free and easy access to high value, machine readable data sets generated and hosted by the federal government. It will enable the public to easily find, access, understand and use data that are generated by the Federal government.

And in Europe, the purpose of EUDAT (European Data Infrastructure)[34], an EU project, is to provide Europe with solutions to problems arising from the growth of data in the scientific research community. Thus, both the U.S. and Europe are aiming to conduct research efficiently by sharing data, and the direction both are taking is to build the infrastructure and tools to make this happen.

4-7 Other Noteworthy Aspects

The DoD wants to improve situational awareness to help warfighters and analysts and provide increased support to operations. As part of this, the department wants to achieve a hundredfold increase in capabilities to extract information from text in any language. Although this is a military application, in the future it could also expand to the public. Thus, this could be very attention-worthy R&D that may conceivably lead to dramatic changes.

Additionally, DARPA is working on very interesting projects involving scalable algorithms to process imperfect data within distributed data stores as well as interaction tools between humans and computers to speed up customization.

5 Conclusion

"Big data" does not quite have an exact definition, but is rather a general way of referring to large quantities of digital data. It is so large that it cannot be managed with existing techniques, and it is complex data. Work is underway to extract value from it.

The United States government's Big Data Initiative covered by this paper is said to also include projects that have been run since prior to the initiative. To be frank, it has a feel of a hodgepodge of many different projects. However, it resembles the NITRD program in that the Big Data Initiative seems determined to enlist the cooperation of many agencies under the umbrella of "big data." The initiative is one example

of the Obama administration's science and technology policies, which include "experiments to try and vastly improve government effectiveness by recruiting the active participation of universities, private companies, and even the general public, within a comprehensive policy framework" and is trying to "form initiatives that encompass diverse federal policy measures under a single concept."[1]

The most interesting of the initiative's R&D efforts is "How to Look at Big Data." The U.S. government sees big data as science and technology that may contribute to the creation of a new paradigm and will have a very major impact in many realms, as the internet did. Noteworthy aspects of the initiative include the "Focus on Visualization Technology," the "Relationship with Cloud Computing," "Considerations to Human Resource Development," "Active Participation by Industry and Academia" and "Encouraging Data Sharing." We can also see that the U.S. is giving thought to policies such as those encouraging the provision of facilities and computing power conducive to collaborative work. Since the results produced by the initiative's R&D may become widely used and penetrate to society in a few years or decades and lead to major innovations, future developments will be worthy of our attention.

Considering the growth of R&D into big data around the world and the difficulty of the challenges involved, global collaboration will be vital for future R&D. Another sign of big data's promise came in June of 2012, when Japan's MEXT minister forged an agreement with the American director of the NSF on the importance of cooperation in disaster-related research. For example, the NSF will "collaborate on big data and disaster research" with MEXT for the Global Research on Applying IT to Support Effective Disaster Management Award (GRAIT-DM) granted by the Science Across Virtual Institutes (SAVI) program.[35] Japan should also engage in this sort of global collaborative research in order to encourage big data research, and produce exceptional value.

Acknowledgements

I would like to take this opportunity to thank Professor Masaru Kitsuregawa of the University of Tokyo(Director of the Earth Observation Data Integration and Fusion Research Initiative [EDITORIA] and Director of the Center for Information Fusion, Institute of Industrial Science, The University of Tokyo) and Professor Hayato Yamana of Waseda University Faculty of Science and Engineering who provided much information and advice during the composition of this paper.

References

[1] "U.S. Science and Technology Policy under Tight Budgets: Report on the 2012 AAAS Forum on Science and Technology Policy," Science and Technology Trends, July/August 2012
[2] http://www.mizuho-ir.co.jp/publication/navis/017/pdf/navis017_07.pdf
[3] http://www.npu.go.jp/policy/pdf/20120731/20120731.pdf
[4] http://www8.cao.go.jp/cstp/siryo/haihu103/sanko2.pdf
[5] "Academia's Challenge in the Era of Big Data – Academic Cloud Investigative Commission," 77th meeting of the Information Science and Technology Committee, Document 2
[6] http://gigaom.files.wordpress.com/2010/05/2010-digital-universe-iview_5-4-10.pdf
[7] http://idcdocserv.com/1142
[8] http://www.i-cio.com/features/august-2010/eric-schmidt-exabytes-of-data
[9] A Vision for Exascale (Mark Seager, Intel) 2012.6 ISC12
[10] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute 2011.5
[11] http://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AD%E3%83%BC%E3%83%A9
[12] "How to Use Big Data," Information and Communications Council, ICT Basic Strategy Board, Ad Hoc Group on Utilization of Big Data, May 17, 2012
[13] http://research.google.com/archive/mapreduce.html
[14] Takeaki Uno, "Theoretical Computation Approaches for High-Speed Big Data Processing," Information Processing Society of Japan 2012 Seminar, Big Data and Smart Companies, July 25, 2012
[15] http://www-06.ibm.com/ibm/jp/provision/no65/pdf/65_article2.pdf
[16] http://www.slideshare.net/IBMDK/global-technology-outlook-2012-booklet

[17] Masuru Kitsuregawa, "Fourth Generation of Media will Create the Big Data Age," (ProVISION, Winter 2012, No.72, p13-14)
[18] http://translate.google.com/about/intl/ja_ALL/
[19] http://www.its-jp.org/saigai/
[20] http://www8.cao.go.jp/cstp/tyousakai/innovation/ict/4kai/siryo4.pdf
[21] Yasushi Sakurai, "Mining Technologies for Data Streams and their Application," Information Processing Society of Japan 2012 Seminar, Big Data and Smart Companies, July 25, 2012
[22] http://jpn.nec.com/press/201207/20120713_03.html
[23] "Trends and Issues in Research on Context Awareness Technologies for a Ubiquitous Network Society," Science and Technology Trends, August 2007
[24] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
[25] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf
[26] http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607
[27] Broad Agency Announcement XDATA DARPA-BAA-12-38
[28] http://www.1000genomes.org
[29] http://sdav-scidac.org/report.html
[30] http://www.nitrd.gov/Index.aspx
[31] http://www.prace-project.eu/?lang=en
[32] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal
[33] http://www.data.gov/
[34] http://www.eudat.eu/
[35] http://www.nsf.gov/news/special_reports/savi/awards.jsp#grait-dm

Profiles

Minoru NOMURA
Science and Technology Foresight Center, Affiliated Fellow
http://www.nistep.go.jp/index-j.html

Before taking his current post, Minoru Nomura has worked in the private sector on R&D for CAD design software, as well as business development in the fields of high-performance computing and ubiquitous networking. His interests include supercomputers and LSI design technology. Mr. Nomura is now working on quantification and visualization of the social and economic results of R&D.

(Original Japanese version: published in September/October 2012)