Meeting Recap
REALIZING THE VALUE FROM BIG DATA
February 28-March 2, 2011

A meeting co-organized by the Board on Global Science and Technology (BGST) and Singapore's Agency for Science, Technology and Research (A*STAR)

Disclaimer: *This meeting recap was prepared by National Research Council (NRC) staff as an informal record of issues that were discussed during sessions of the NRC's Board on Global Science and Technology meeting in Singapore, held on February 28-March 2, 2011. This document was prepared for information purposes only and as a supplement to the meeting agenda available online at www.nas.edu/bgst. It has not been reviewed and should not be cited or quoted, as the views expressed do not necessarily reflect the views of the National Research Council or the Board on Global Science and Technology.*

Today, the potential exists for researchers to access widely distributed and highly complex datasets generated around the world. Simultaneously, advanced computing infrastructures and techniques are also beginning to provide researchers the ability to explore massive datasets in new and unexpected ways. Many have suggested that these data-intensive methods may advance scientific discovery in ways that both challenge and support the traditional scientific method. Others have also speculated that the fusion of knowledge from disparate domain sciences will transform the methods used by researchers to create new knowledge.

In many scientific disciplines, researchers are only beginning to exploit *big data*. Significant barriers that are both technological and cultural remain–and many of these barriers span scientific disciplines and application domains.

On February 28-March 2, BGST conducted an international, multidisciplinary meeting on "Realizing the Value from Big Data" in Singapore. The meeting was jointly organized and hosted by the Institute for Infocomm Research (I^2R) of Singapore's Agency for Science, Technology, and Research (A*STAR) at Fusionopolis. This meeting convened bioinformatics scientists and environmental scientists with computational/ data scientists to assess and project the impact of complex datasets in their fields. Both domain scientists and computational scientists identified computational and policy roadblocks that prevent their disciplines from fully extracting value from *big data*. The participants, many of whom brought extensive international experience, were drawn from research organizations in Australia, China, England, Hong Kong, Japan, Korea, the Netherlands, Portugal, Singapore, and the United States.

Participants were not asked to arrive at a consensus position on any issue, but rather to identify challenges from different disciplinary and cultural perspectives. Throughout the meeting, participants were divided into small, interdisciplinary breakout groups to discuss issues in greater detail.

Setting the Stage

In 2007, the total amount of digital data stored worldwide was estimated at 295 EB (exabytes) and projected to grow by 5 EB each year. **Lim Chuan Poh**, chairman of A*STAR, cited this staggering figure during his opening address to describe the explosive growth in global data storage capacity over the past 25 years. While the cost of data storage has decreased dramatically to only 8 cents per gigabyte, Lim noted, the hyper-deflation of storage costs is more than counteracted by our enormous capacity to generate data–mainly due to the proliferation of data generating devices and increased access to the Internet.

According to Lim, power no longer resides with those who own the data, but rather with those who can intelligently and responsibly make sense of the "data deluge." Making sense of this data will require collaborative and multidisciplinary engagement, citing the numerous ongoing scientific discoveries that occur at the intersection of the various domain sciences with analytical/computational sciences.

This is exemplified by A*STAR institutes such as the Institutes for High Performance Computing (IHPC) and Infocomm Research (I^2R), Lim said. These institutes are developing computational tools and algorithms capable of integrating, mining, and analyzing data from multi-domain sources for applications ranging from public transportation networks and natural language human-machine translators to statistical methods that monitor the spread of infectious diseases.

Despite these successes, Lim cautioned that numerous challenges must be solved (e.g., sufficient storage capacity and the further development of analytical tools) if researchers expect to fully leverage the vast and growing amounts of available digital information.

Following Lim's address, **Ruth David**, president and chief executive officer of Analytic Services, Inc. and BGST chair, remarked that the barriers and opportunities presented by *big data* are national, multi-national, and global, and span nearly all scientific disciplines. As such, collaborative solutions by an international community present significant benefits to all researchers engaged in *big data*.

David also acknowledged that the challenges which prevent researchers from recognizing the value from *big data* are both technological and cultural. Today, the generation and accumulation of data is beginning to outpace the development of analytic infrastructure and data management protocols necessary to glean insights from the data. These inadequacies become even more pronounced as datasets become increasingly distributed and/or unstructured. It also remains unclear what role new media types, such as Twitter feeds and Facebook, will have for next generation data applications.

Additionally, David said, as data is generated at larger–and sometimes global–scales, its exploitation will require local, regional, and international cooperation across a diverse array of application domains. The ability of researchers to solve global problems that are both inter- and multidisciplinary will require domain scientists to work collaboratively with both data scientists and computer scientists.

David concluded her talk by describing why environmental science and bioinformatics are good *big data* use-cases. In addition to the data-rich nature of both domain sciences, bioinformatics and environmental science issues do not stop at national borders. In some cases, the two domains may even be inextricably linked. David asked the participants to consider, for example, whether a better understanding of environmental problems could lead to a newer understanding of population genetics. If there are benefits to fusing disparate datasets that span scientific domains, she asked, what are the barriers? What does *big data* mean in an environmental science or bioinformatics context?

What does *big data* mean?

From the start of the meeting, many participants commented that the term *big data* connotes disparate meanings to individual

researchers. When asked "what does *big data* mean to you?" some participants focused on the attributes of data and others on the value derived in terms of problem solving. Some participants indicated that there is often too much emphasis on "size" as an indicator of *big data*. For example, the analysis of a multidimensional "small" dataset is sometimes more data intensive than the same analysis on a large one-dimensional dataset. **Vipin Chaudhary** from the University of Buffalo, SUNY, added that, oftentimes, the problems associated with *big data* are simply a result of summing many small datasets. Several participants also noted that "big" means different things depending on which domain science is generating and/or analyzing the data. For example, **Jeff Dozier** from the University of California Santa Barbara remarked that many "big" bioinformatics and remote sensing datasets seem 'small' in comparison to those generated by the National Ignition Facility.

In addition to size, various participants identified other important indicators of *big data, including* dimensionality, complexity (e.g., heterogeneity), and the ease with which data can be integrated and analyzed. **Allen Rodrigo** from Duke University suggested that it is more important for researchers to think about *big data* in terms of the process through which one derives insight rather than the characteristics of the dataset. Dozier added that it may not be productive to qualify datasets as "big" or to compare "big vs. small" datasets. Rather, the more pressing challenge lies in creating processes that lead to scientific discovery–and these challenges are much greater than those posed by "small" datasets.

Some participants indicated that the desire (by researchers, businesses, society, etc.) for *big data* will require the development of robust and automated sensors. **Bing Qiang Wang** of the Beijing Genomics Institute noted that these sensors will require data validation techniques. Other participants commented that the global generation and distribution of *big data* will require researchers to deal with data provenance and bias issues. **Mario Caccamo** from The Genome Analysis Centre (UK) noted that researchers tend to sometimes worry too much about errors and noise, but not data bias. Additionally, **James Agutter** from the University of Utah observed that data bias is especially important for researchers who use *big data* models to project future truths. He continued by asking, how should these representations of the truth influence our decision making when there are so many biases?

Another point that emerged during the discussion was that *big data* often have value beyond the immediate purpose for which the dataset(s) were generated. This led some participants to question how we might more effectively characterize the value of *big data* in terms of not only primary use but also the reuse of data. In addition to the potential for reuse, some participants mentioned the need to consider the cost of regeneration. For example, in fields like environmental science, datasets often document transient events that cannot be reproduced (e.g., seismological data). Given the high costs of data storage and management, participants also discussed the implications of catastrophic data loss.

Some participants cautioned that the ease with which today's large datasets can be generated and analyzed is not in itself justification for *big data*. **Bernie Meyerson** from IBM suggested that a solution might be to identify the *big data* commonalities important enough to generate global investments so that the benefits of *big data* outweigh the high costs.

Participants also discussed at length what *big data* means in the context of advancing scientific discovery. Several observed that the *big data* environment affords opportunities for data-driven hypothesis-generation in contrast to the more traditional hypothesis-driven research model. Rodrigo cautioned that researchers should not be too hasty in departing from traditional scientific methods, which have led to significant discoveries (e.g., the structure of DNA and evolution). David emphasized that the goal of data intensive science is not to replace the traditional scientific method, but rather to derive insight

from data in ways that will advance scientific discovery.

Meyerson added that there is always something to learn from *big data*, especially if you are looking for very tiny trends or if you are not even sure what to look for. This sentiment was re-enforced by **Kin Mun Lye**, executive director of I^2R at A*STAR, who remarked that any enormous database must hold something that will challenge the ways that researchers think and understand the world. Several participants observed that the realization of such a global repository–comprised of widely distributed and highly heterogeneous data–will require significant research efforts by both computer scientists and domain scientists.

In the discussions that followed, participants discussed four major themes: accessing data, reusing data, interdisciplinary engagement, and international cooperation.

Accessing Data

During the discussion, participants identified several motivations for accessing *big data*, such as analyzing their own or others' data to confirm previous results or to discover something new. Given the vast amount of data that exists, many participants voiced concern that it does not make sense to waste time and resources to regenerate data that have already been collected. Still other participants also observed that the task of identifying and obtaining useful information will become more daunting for future domain scientists as data repositories host larger and more complex datasets. These challenges become even more pronounced as researchers seek to understand multisource and multidomain data that exceed the scope of their own domain expertise.

To effectively access *big data,* some participants voiced the need for novel computing infrastructures that maximize data storage and data extractability for a distributed community of researchers. Several participants observed that current data management infrastructures are poorly matched to the challenges presented by many of the previously described *big data* attributes, such as volume, geographic distribution, and heterogeneity. Others indicated that common search engines are not "tuned" to locate relevant data.

Some participants, such as **Miron Livny** of the University of Wisconsin-Madison, observed that increasingly heterogeneous repositories will require scientific disciplines to develop consistent standards for metadata associated with *big data*. Others suggested that emerging cloud architectures could mitigate some of the challenges stemming from data volume, as long as techniques are simultaneously developed to guarantee data provenance. Alternatively, some questioned whether researchers should reevaluate what raw data could be safely discarded to reduce volume. Others cautioned that such a priori decisions could limit a dataset's value for future use.

A number of participants discussed how the commoditization of datasets within certain scientific disciplines, such as geospatial science and bioinformatics, hinders data access. Some participants also considered what security considerations would need to be made for datasets that have limited or restricted access–because of privacy concerns, intellectual property rights, or energy, economic, and national security policies. Chaudhary added that certain geodatabases (e.g., gas and water data) are also subject to national jurisdictions that prevent datasets from crossing borders.

Another point that emerged during the discussion was that traditional data management is centered on making data accessible to a particular set of users, and that neither the infrastructure nor the cultural incentives are in place to support broad data sharing. According to **Hamideh Afsarmanesh** of the University of Amsterdam, data sharing begins with researchers recognizing that their own data can benefit the broader scientific community. Some participants suggested that this recognition is slow to gain traction because of an academic culture that rewards individual achievement. Others questioned whether stove piping within

scientific disciplines challenges researchers' ability to see their own data in the context of other data. Various strategies were discussed that might motivate individual researchers to share datasets they generate (e.g., citation credit on publications stemming from using others' data). However, many participants observed, overcoming the cultural impediments alone is not sufficient to overcome the challenges related to the high costs of data management.

In fact, many participants highlighted the issue of "who pays?" as a significant impediment to data access, since individuals are generally not funded for the increased cost of making their data available to others. Given these costs, some participants questioned whether business models that discourage data sharing (e.g., hospital data management policies) are sustainable. Agutter noted that new mandates are forcing some hospitals, which traditionally do not share data, to re-think their data management policies. Unless institutions are able to create a sustainable model for maintaining existing data repositories at a decreasing cost, some participants, such as Livny, questioned whether researchers run the risk of pouring too many resources into the past rather than the present.

Reusing Data

According to **John Taylor** from The Commonwealth Scientific and Industrial Research Organization (Australia), researchers who fail to make their data reusable miss significant opportunities to extract maximal value from their own data. In the same vein, researchers also fail to fully recognize the value of *big data* if they do not exploit others' data. In some cases, researchers are motivated to use others' data within their own domains to either confirm or improve statistical analysis of their own datasets. Here, Taylor explained, the recycled data are likely to exist in a similar format (or have similar attributes) such that new analyses are relatively easy.

Other participants, such as **Chaitan Baru** from the University of California San Diego, indicated that researchers may also seek to use data from other domain sciences to understand their own data in a different– and potentially broader–context. As these recycled datasets are generated by other scientific disciplines for different purposes, Baru said, the re-purposed data may exist in a format or structure that is not compatible or easily re-analyzed.

In the discussion that followed, participants identified some of the challenges associated with fusing datasets both within domains, as well as datasets that span multiple scientific domains. Many participants commented that current analytic techniques used to integrate and analyze heterogeneous datasets are insufficient. Inconsistent–and sometimes incompatible–data structure and formats across scientific disciplines were highlighted as a significant barrier to data integration. It was also noted that these challenges become more complex when attempts are made to integrate data across multiple scientific disciplines.

Other participants suggested that traditional database models are built upon "small data" assumptions that do not hold true for *big data*. **Jae Woo Kang** from Korea University explained that in "small" datasets, domain scientists traditionally provided the structure that governs how data are managed and analyzed. This is no longer the case for "big" datasets, where the structure must be derived from the spatial, temporal, semantic, and causal relationships hidden in the data. If successful, he said, abstracting from *big data* may provide researchers with unexpected insight that generates new hypotheses.

Some participants commented that next generation data repositories must support not only data storage and extraction, but also the derivation of information hidden in the data. This would require new service-oriented architectures and services that can bring reusability to a distributed community of researchers. **Keiko Takahashi** from the Japan Agency for Marine-Earth Science and Technology added that repositories must support scalable analytic tools and reduce the current network bottlenecks that hinder data migration. Other participants indicated a need

for strategies that encourage not only the reuse of data, but also the recycling of computational solutions developed for various domain sciences. Additionally, they said, processes should exist for assigning and analyzing the metadata associated with these computational solutions.

Participants also discussed whether raw data should be made available for researchers who wish to reuse data. While some analyses require raw datasets, several participants noted that there are many applications for which processed data would suffice. Baru suggested that efforts be made toward stratifying these different layers of data access.

Some computer scientists observed that there is often a tendency among domain scientists (with limited exposure to computational science) to blindly query large datasets in pursuit of the "aha moment." According to **Paul Maglio** of IBM Research, it is not only about what the researcher can automatically find in the data, but also the theories and models researchers bring to the data. Kang added that computational science does not minimize or supplant the role of the domain scientist. While computer scientists can help generate models and screen hypotheses, domain scientists are critical for understanding the models and testing and validating hypotheses.

While significant opportunities exist for computational scientists and domain scientists to learn from one another, several participants noted that interdisciplinary research is challenged by a lack of communication between domains. As such, Maglio proposed the creation of a common and accessible language that bridges the computational and domain sciences. For example, some participants speculated whether computer scientists could create a common "data template" for domain scientists that asked the following questions: What were the observations (data) and what did they look like (metadata)? How were the data generated (provenance) and how did the original hypothesis (expectations) compare to the result (outcomes)?

Interdisciplinary Engagement

Throughout the meeting, participants worked in small, interdisciplinary groups to identify the barriers that hinder collaborative engagement on *big data* issues among disparate scientific communities. Some participants observed that in several cases, standing scientific and cultural perspectives have resulted in a lack of mutual appreciation for each discipline (e.g., computer scientists want to be seen as "enablers" rather than "plumbers"). Others noted that even when the value of interdisciplinary research is recognized, neither public nor private incentives exist that enable effective communication or collaboration.

Some participants also commented that it is difficult for research communities to foster collaboration without sacrificing individual domain-specific or laboratory-specific objectives. Caccamo noted that this is especially important for researchers who operate in a culture that promotes science "rock stars".

Others observed that interdisciplinary collaborations also suffer from "one-off" mentalities held by both computer and domain scientists. Livny commented that the computer science community tends to look inward because its goal is to develop a product targeted to the IT world. Similarly, many domain scientists fail to recognize the intricacy and complexity behind computational algorithms. There is a false assumption that the sharing of computing infrastructure and resources is equivalent to enabling new science, some participants observed. Instead, domain scientists should think about ways in which their domain-specific problems could benefit their computational collaborators and vice versa. Afsarmanesh added that the challenge lies in convincing researchers that there is value to helping others solve their problems.

Lastly, some participants considered whether interdisciplinary research has the potential to actually impede scientific discovery. Others suggested that interdisciplinary collaborations also have the potential to create and fuel

supply-and-demand imbalances. Still others indicated that interdisciplinary collaboration is not cheap, citing the short lifetimes of interdisciplinary institutes that are not attached to long-term funding. While some participants suggested that laboratory research goals should become more interdisciplinary, others proposed that future laboratories may consist of graduate students from a variety of different domain sciences.

According to Dozier, academic institutions are not helping computational and domain scientists deal with the transition toward data management that is integrated with scientific models. While some participants suggested that academic institutions require training that teaches researchers how to collaborate inter-disciplinarily, others expressed concern that additional requirements could force graduate students to forfeit necessary domain-specific classes and skill sets.

Some participants also commented that the traditional P.I. (Principal Investigator) model of academic research cannot keep pace with scientific disciplines that are becoming more tightly twined with *big data* issues and increasingly interdisciplinary. To deal with this, some participants suggested a movement away from traditional infrastructures that fund basic research towards funded institutes with a finite lifetime. Rather than forcing researchers to overcome a cultural resistance that is misaligned with interdisciplinary collaboration, some participants considered the feasibility of hiring faculty members into an academic culture that supports and rewards interdisciplinary research.

A number of participants indicated that both computational and domain scientists need to develop new skill sets that foster collaborative problem solving. Many commented that interdisciplinary research efforts are hampered by ineffective communication that results from an insufficient understanding of other scientific disciplines and from a lack of mutual respect.
Alternatively, participants discussed whether the emergence of *big data* and data intensive methods will result in a new generation of researchers possessing multidisciplinary skill sets that deal with the data–in addition to the skills required to understand the underlying problem domain. Some would consider individuals with this combination of skills a data scientist. Examples of data-related skills include: data acquisition, filtering, organization, mining, and visualization, as well as human-computer interaction. Some participants cautioned that the proliferation of these skill sets across a scientific domain could lead to stove piping that impedes knowledge creation. Others expressed concern that domain scientists should not attempt to do the job of computer scientists, and vice versa, suggesting that researchers learn to collaborate in ways that mutually leverage their skill sets.

Some participants were of the opinion that the benefits of interdisciplinary engagement go beyond individual projects. In fact, many suggested that integrated problem-solving strategies, which leverage expertise from multiple scientific disciplines, may allow researchers to exploit *big data* in ways that are overwhelming or impossible for individual researchers. Others observed that researchers who wish to solve complex scientific problems that are local, national, and global will depend on a community of knowledge that is intradisciplinary, interdisciplinary, and sometimes international. In fact, some participants, such as Afsarmanesh, suggested that the realization of value from *big data* begins with the recognition that there is value in building and sustaining a *big data* community.

International Cooperation

Many meeting participants identified two broad themes in describing the need for an international dialog around *big data* issues. These themes can be characterized as problem-centric and researcher-centric. The former focuses on the fact that many challenges confronting nations today are global in scope and therefore cannot be addressed in isolation. The latter stems from a human desire to learn and leverage the work of others.

Several participants noted the cross-border dependencies in vital domains, including global security, population health, and environmental health, and identified *big data* opportunities that are today limited by the challenges and barriers described above. An international dialogue could be used to identify the appropriate *big data* and domain science expertise necessary to solve these 'big science' problems.

Other participants indicated that an international dialog could help to identify common technological *big data* problems and solutions that transcend national boundaries. The identification of enough commonalities could encourage a re-think of international data-sharing and -management policies that oftentimes hinder cross-border flow of information. As data continue to be generated around the world for a plethora of different purposes, they said, researchers should continue to engage with international partners to develop international data standards and to understand the various cultural and political perspectives that create data bias. Some researchers also expressed strong interest in gaining access to computational infrastructures not available in their home countries, to avoid bearing the full costs of building and maintaining these environments.

During the last session of the meeting, participants worked in small, interdisciplinary and international groups to identify specific initiatives that might mitigate key barriers to fully realizing the value from *big data*. Concepts ranged from the development of common abstractions that could be reused across domains, to the notion of a standardized Internet protocol that would facilitate identification and location of *big data* of interest to a research team.

● ● ●

The purpose of the meeting was not to build consensus on how best to realize value from *big data*, but to discuss these issues in a highly international and interdisciplinary setting. In this still relatively early phase of the "data revolution," dialogues such as these are creating networks that help to reduce the disciplinary and national boundaries that can hinder scientific progress. To that end, the Board on Global Science and Technology will continue to co-sponsor international meetings in areas of emerging science and technology.

For more information about BGST and future activities, please visit the web site at http://www.nas.edu/bgst.

ABOUT the Board on Global Science and Technology (BGST)

MISSION

The Mission of BGST is the establishment of a global network that will (1) enhance transparency with regard to international scientific and technological advances, (2) improve U.S. decision making and public policy development, and (3) foster the development of international "norms" for the governance of emerging technologies. BGST has established a program of workshops and other convening activities, both within the United States and overseas, to build and sustain an international, interactive community of scientists, engineers, medical and health researchers, and entrepreneurs who are engaged in the research and development of emerging technologies. BGST is a joint project of Policy and Global Affairs and the Division on Engineering and Physical Sciences.