Prepared by the American Association for the Advancement of Science in conjunction with the
Federal Bureau of Investigation and the United Nations Interregional Crime and Justice Research Institute

National and Transnational Security Implications of Big Data in the Life Sciences

Report in Brief

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

unicri
United Nations
Interregional Crime and Justice
Research Institute

Report in Brief:

National and Transnational Security Implications of Big Data in the Life Sciences

A Joint AAAS-FBI-UNICRI Project

Big Data analytics is a rapidly growing field that promises to change, perhaps dramatically, the delivery of services in sectors as diverse as consumer products and healthcare. Big Data analytics also have the potential to enable deeper insight into complex scientific problems by leveraging ever-increasing stores of knowledge coupled with ever-improving processing capabilities. These beneficial aspects of Big Data have been well-documented and widely touted. However, less attention has been paid to the possible risks associated with these technologies beyond issues related to privacy. These risks include, but are not limited to, vulnerabilities of datasets to cyber intrusion and design of biological weapons derived from the integration and analysis of Big Data in the life sciences. In this report, the American Association for the Advancement of Science (AAAS) Center for Science, Technology, and Security Policy (CSTSP) and the Biological Countermeasures Unit of the Federal Bureau of Investigation Weapons of Mass Destruction Directorate (FBI/WMDD/BCU) seek to:

- Examine the risks and benefits associated with Big Data analytics;
- Develop frameworks for risk and benefit assessment of emerging or enabling technologies, such as Big Data in the life sciences; and
- Identify options for U.S. government action to further characterize the risks and benefits from Big Data analytics and to mitigate risks.

The report is the culmination of a year-long evaluation of the drivers of Big Data in the life sciences, possible risks and benefits, and existing or needed solutions to address the risks identified. To carry out this project, AAAS/CSTSP and FBI/WMDD/BCU involved a working group of experts in computer science, data science, life science, biological security, data security, cyber security, law enforcement and homeland security from U.S. government agencies, intergovernmental organizations, academia, private industry, and the amateur science community.

This report in brief reviews Big Data in the life sciences and presents the findings of the overall project.

What is Big Data?

Although no formal definition of Big Data exists, it often is defined by several characteristics (referred to as the "four Vs" by industry):

1. Data are generated and collected from a number of distinct sources, and more than one dataset is integrated and analyzed (i.e., the **variety** of data).
2. Data are being added to, deleted from and/or changed in datasets at different speeds and times depending on the type of data and collection method(s) (i.e., the **velocity** of data).
3. Datasets are incomplete, imperfect, and error-prone, and the data collected in these repositories are not standardized (i.e., the **veracity** of data).
4. The amount of data in datasets is very large, requiring multiple petabytes of storage (i.e., the **volume** of data).

Data come in many forms and from many different sources. Data can be from publicly available sources, privately held sources, and social media platforms. It is either "born digital," which means that it is generated through electronic means such as the "internet of things"[1] or internet search terms, or observed, such as scientific results. The data are heterogeneous, often containing errors, and/or incomplete. Some data are deposited into datasets deliberately while other data are not. The datasets can be structured or unstructured, often huge in size (exceeding petabytes), and rapidly changing. Specifically in the life sciences, datasets include raw data, combined data, or published data from the health-care system, pharmaceutical industry, genomics and other -omics fields (e.g., proteomics, transcriptomics, metabolomics, neuromics, immunogenomics, and pharmacogenomics), clinical research, environment (e.g., biodiversity and conservation efforts, water contamination and availability, and air quality), agriculture, and microbiome efforts. Several distinct datasets are integrated and analyzed together (i.e., in aggregation and temporally, such as longitudinal studies), which contribute to their characterization as Big Data.

Several different technologies are being developed or improved to analyze Big Data. These technologies are computationally or mathematically-based and require significant computing capabilities. Analytic technologies include data integration, data mining, data fusion, image and speech recognition, natural language processing, machine learning, social media analysis, and Bayesian analysis. (Figure 1) Often, data analytics involves combinations of technologies, such as machine learning, natural-language processing, and data mining. The technologies most often are proprietary and/or experimental. However, analytic technologies increasingly are being provided through the cloud. These data analysis technologies can be used with datasets containing information from any source and from any sector or discipline.

[1] S. Ferber. (2013) How the Internet of Things Changes Everything. Harvard Business Review Blog Network. Accessible at http://blogs.hbr.org/2013/05/how-the-internet-of-things-cha/. Accessed on October 10, 2014.

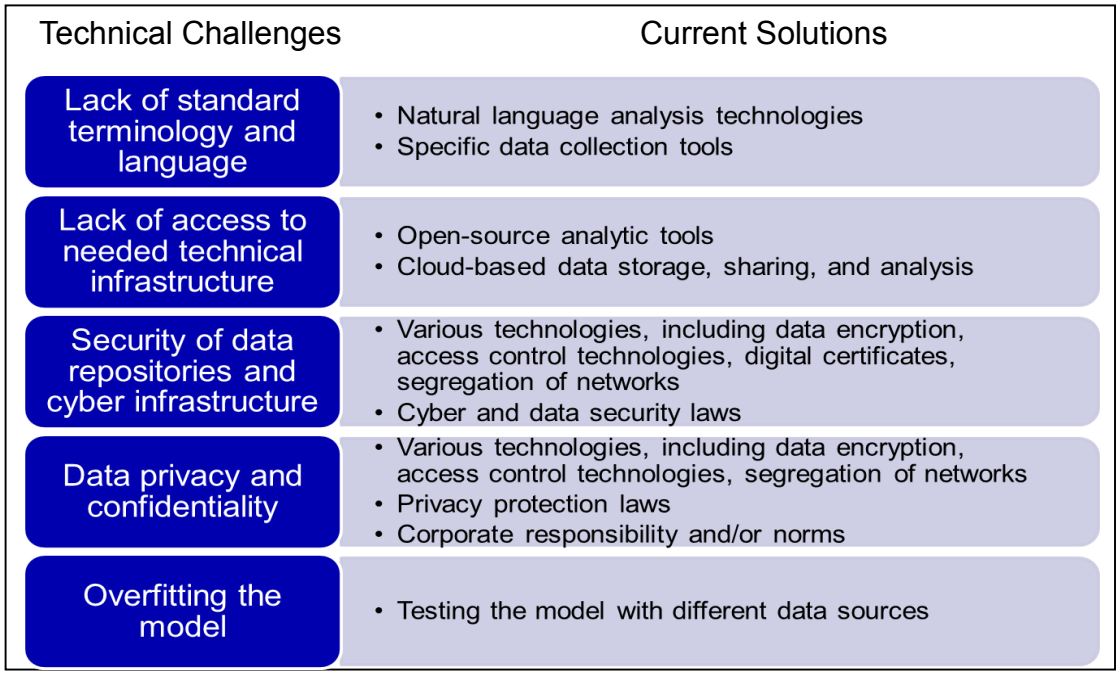
Figure 1. Data Analysis Technologies

The picture of investment, research and development, and use of Big Data analytics is complex, in part because the sectors and number of organizations involved are many. Private companies and governments are investing in the development of new and/or improved analytic technologies to evaluate data from several different sources to solve a problem, improve a service (e.g., healthcare), and/or enhance marketing activities. Academic, nonprofit, and for-profit organizations are actively developing new approaches for collecting and analyzing data in addition to exploring new uses for data analytics. Through mobile applications, crowdsourcing, cloud-sharing, and certain projects, such as the National Genographic Database or the Personal Genome Project, members of the public and amateur science communities now are involved in generating and sharing data. The vast, and increasing, amount of information posted on social-media platforms further adds to the increasing amount of available data. These efforts are not limited to the United States; many countries are investing in and/or using Big Data and analytic technologies.

Several challenges affect the complete use of Big Data analytics to address societal, health-care, agricultural, environmental, commercial, and/or national and transnational security issues. These challenges include the lack of standardized

language found in datasets, the availability of technologies and computing power to support Big Data analytics, the security of the cyber infrastructure and data repositories, the privacy and confidentiality of individuals, and overfitting the analytic model to the data on which it was developed. Figure 2 lists these challenges and current approaches for addressing them.

Figure 2. Technical Challenges and Current Solutions of Big Data in the Life Sciences



Benefits, Risks, and Solutions for Big Data in the Life Sciences

For decades, the national and international security communities have evaluated advances in science and technology for their potential benefits to address societal needs and their potential risks to society. These communities have been adept at building on (if not furthering themselves) new advances in science and technology to increase their capabilities to identify, deter, prevent, and/or mitigate potential threats from adversaries and/or with certain materials (e.g., chemical, biological, radiological, nuclear, and explosive). At the same time, they have been concerned about "technological surprise" in which adversaries have the access, skills and expertise, and motives to use new advances in science and technology in unanticipated ways against nation-states or sub-entities of those nation-states. Many nation-states have implemented processes and/or measures to evaluate scientific and engineering advances for their utility to address national security and broader societal needs, and to ensure that the advances cannot be used for harmful purposes (e.g., causing destruction, illness and death among populations, and economic damage). The need for the evaluation of technologies is expected to continue as science and technology capabilities advance and as societies and their needs change.

National security means different things to different people mainly because of the changing nature of the threats and security risks. The threats have expanded beyond nation-states to non-state/lone actors expressing an interest in chemical, biological, radiological, and nuclear weapons. Increasingly, non-state/lone actors, including individual actors, are gaining access to rapidly-progressing science and technological

capabilities in academic and private-sector institutions. Several of these S&T developments are becoming increasingly accessible to a broader array of individuals, including amateur scientists and non-life scientists. To evaluate the national and transnational security implications of emerging and enabling technologies, such as Big Data in the life sciences, clearly describing the concept of national security within this changing landscape is necessary.

At the highest level, many nations, including the United States, seek to maintain trust in government, promote economic prosperity, protect the health, safety, and security of their citizens, and uphold their national sovereignty and standing in the global community. These nations achieve these high-level goals through a number of objectives, including political and military efforts, critical infrastructure protection and resilience, border security, geostrategic security, economic and commercial security, environmental and energy security, health and food security, and the protection of values, liberties, and privacy. Nations implement programs, develop strategies, pass statues, develop regulations, and conduct activities to achieve these objectives for addressing specific security threats.

In the United States, biological threats are addressed through a variety of programs that span the prevention, detection, and response spectrum. U.S. initiatives to promote preparedness and response to natural or man-made threats, prevent the misuse of scientific knowledge and/or theft of biological materials, maintain transparency of biological defense research and diagnostic efforts, advance microbial forensics, and develop medical countermeasures are among the many programs used to achieve the United States' high-level goals and objectives for national security. The sheer complexity of these issues and the increasing amount of data available to inform or implement these initiatives suggest a significant role for Big Data and data analytics in the life sciences. However, the inherent vulnerabilities in relying on databases and cyber infrastructure to collect, store, and analyze data and the security risks they present (e.g., flooding datasets with false information or hacking databases or computer systems) are exacerbated in Big Data analytics because it involves several different databases and possibly multiple computer systems (e.g., cloud-based analytics and private analysis tools). In addition to system vulnerabilities, the power of integrating and analyzing data from several sources could enable the development of pathogens, toxins, or biologically active molecules specifically to harm certain animals, plants, or people and/or evade current defenses.

Risk and Benefit Assessment Frameworks

Evaluating the possible security risks and benefits of emerging or enabling technologies, which in this report is Big Data in the life sciences, is critical to maximizing the benefits while minimizing the risks. However, risk and benefit assessments are not routinely conducted together. To the best of our knowledge, no other group has tried to develop a benefit assessment scenario and evaluated a specific technology for its potential risks and benefits at the same time. This report presents qualitative risk assessment and benefit assessment frameworks with which to evaluate emerging or enabling technologies. (Figures 3 and 4) The risk assessment framework includes an evaluation of the scientific needs *and* adversary capabilities and access to materials and facilities, two items not often included in assessments conducted by the scientific community. The benefit assessment framework includes an evaluation of the capabilities added *and* legal, ethical, and social considerations. Jointly, these

frameworks, along with technically sound and realistic benefit and risk scenarios, enable a more complete evaluation of the likelihood of a particular benefit or risk (whether the risk is from vulnerabilities in the system or the deliberate use of the technology to cause harm).

Figure 3. Conceptual Framework for Qualitative Risk Assessment for Emerging and Enabling Technologies

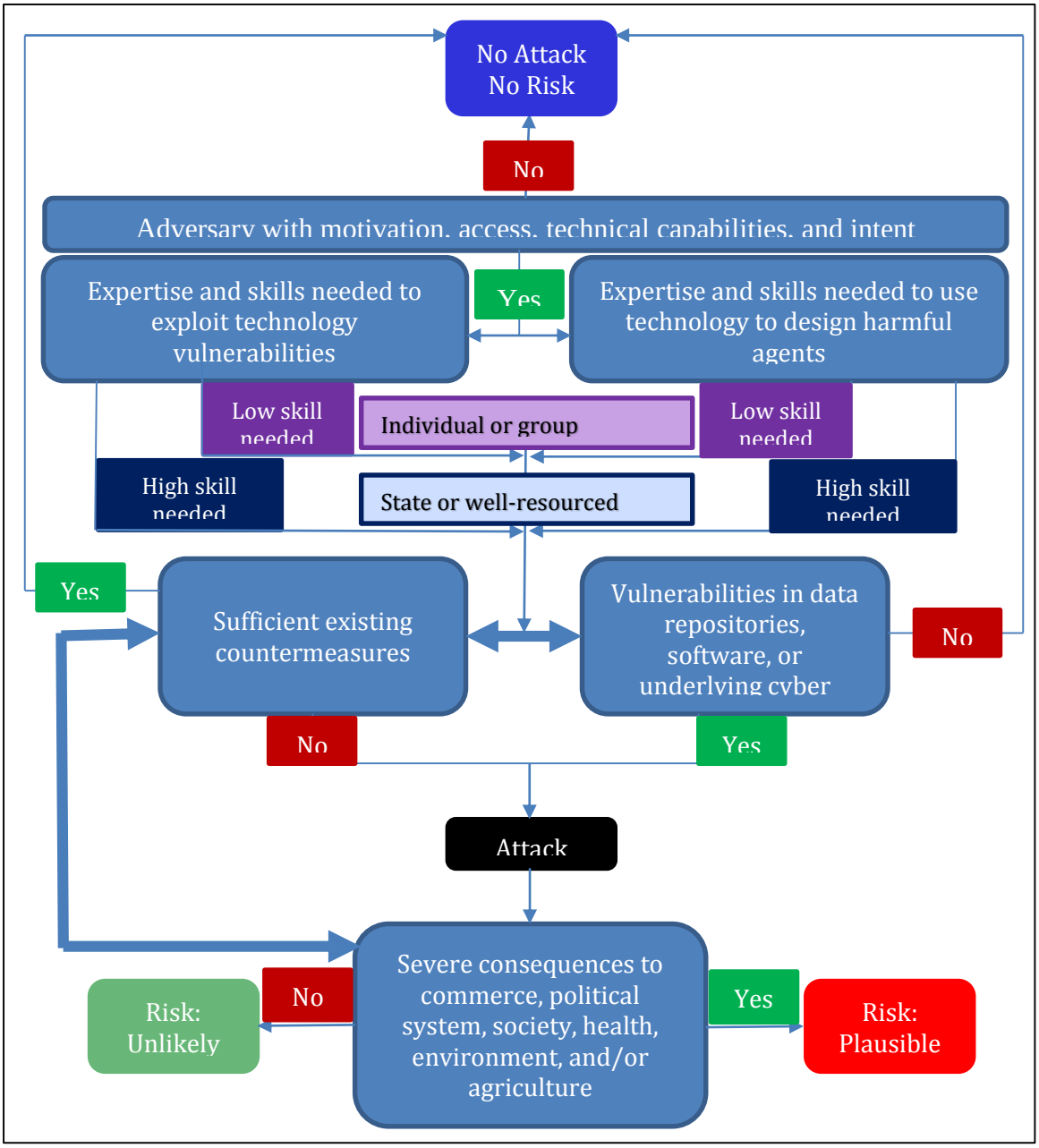
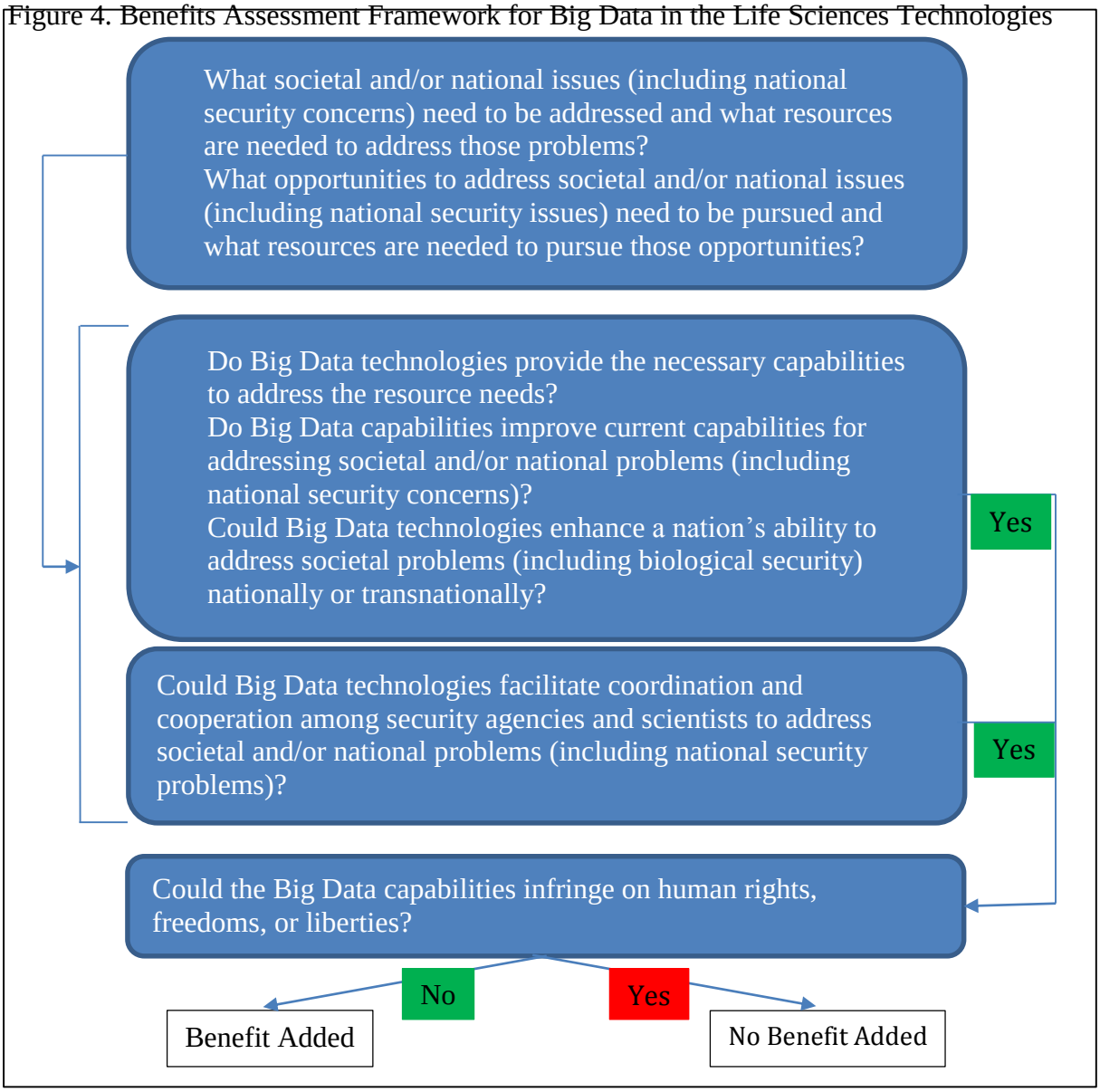


Figure 4. Benefits Assessment Framework for Big Data in the Life Sciences Technologies



Legal, Technical, Institutional, and Individual Solutions

The risk scenarios developed to assess the potential risks of an emerging or enabling technology also can be used to evaluate how well the current governance structure addresses the identified risks and what gaps in governance exist.

Two overarching international legal instruments to which the United States is a party exist to prevent the development of biological weapons: the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (the Biological and Toxins Weapons Convention) and the United Nations Security Council Resolution1540. In addition, the United States has laws (including statutes and regulations), directives, policies and guidance preventing the development or use of biological weapons, the theft of certain biological agents that pose public health, security or safety risks, and the misuse of scientific knowledge, skills, and tools to cause harm. However, no adequate legal or technical solutions exist to prevent adversaries from using data and

analytic technologies to design (i.e., create a blueprint for eventual development) biological weapons. Instead, institutional policies and individual actions are critical for preventing the use of data for harmful purposes.

When evaluating solutions for reducing the vulnerabilities of Big Data in the life sciences, only technical solutions, including access controls and data encryption, exist. Members of the United States Congress have introduced legislation to address cyber and data security threats, but none have been passed. The Executive Branch of the United States Government has promulgated cyber-security programs, but how they affect Big Data in the life sciences is unclear. Unfortunately, beyond the use of technical solutions and common sense behavior, institutions and individuals can do very little to address system vulnerabilities.

Recommendations and Conclusions

In conducting this project, the AAAS/CSTSP identified four critical issues the United States government should consider closely if it wants to maximize the benefits of the technology and minimize potential national and transnational security risks. These suggestions are intended to help the U.S. government anticipate future capabilities and risks of emerging, multidisciplinary science and technology.

1. The U.S. government should actively engage the science and technology communities in evaluating the potential risks and benefits of Big Data to national and transnational biological security. The evaluation of risks and benefits to national security should be a coordinated effort among private, public, and government security and scientific experts, and conducted on a regular basis.

2. The U.S. government and the broader scientific and technology communities should develop educational materials and curricula that impart an understanding of the security risks and vulnerabilities associated with Big Data in the life sciences.

3. The U.S. government and the broader scientific and technology communities should engage in the development of detailed solution scenarios to identify existing legal, technological, institutional, and individual solutions and gaps in governance that need addressing. This should include support for the development of security strategies that can be integrated in an open source environment where large datasets are collected, aggregated, and analyzed.

4. The U.S. government should evaluate legal, technical, institutional and individual measures to promote the benefits of and to prevent or mitigate risks presented by multidisciplinary science such as Big Data in the life sciences, which involves computer science, data science, mathematics, engineering, bioinformatics and life sciences. This should include a review of standing statutory and other legal frameworks to determine the adequacy, applicability and efficacy for enforcement and a determination of whether new statutory and/or regulatory measures may be required. In addition, this evaluation should include an ongoing review of the available technical solutions and

institutional and individual practices for their applicability to addressing the risks of Big Data in the life sciences.

Emerging and enabling technologies, such as Big Data in the life sciences, has the potential to enhance or address national and international needs, including health and healthcare, agriculture and food availability, environmental health, national security, and economic progress among others. The private sector, academia, and governments play important roles in investing in and/or conducting research in technology development and in exploring possible applications of the technologies. Thoughtful consideration of the possible risks (from system vulnerabilities and intentional misuse) and benefits, qualitative assessment of the risks and benefits, and identification of existing and needed solutions are extremely important to ensure that Big Data in the life sciences is developed and applied for maximum benefit. The risk and benefit assessment frameworks, technically robust risk and benefit scenarios, and solution scenarios described in this report provide a starting point for the necessary assessment of other emerging or enabling technologies.