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

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

United Nations Interregional Crime and Justice Research Institute

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

A Joint AAAS-FBI-UNICRI Project

Acknowledgements
We would like to thank all the members of the working group and experts group, who provided excellent guidance, advice, and feedback throughout the project. This project was supported by a contract from the Biological Countermeasures Unit of the Federal Bureau of Investigation's WMD Directorate.

Disclaimer
The concerns or suggestions outlined in this report reflect the discussions at the workshop and do not necessarily represent the views of the FBI WMD Directorate; AAAS Board of Directors, its Council, or membership; and UNICRI.

Produced in the United States (2014)
American Association for the Advancement of Science
1200 New York Avenue, NW
Washington, DC 20005

About FBI/WMDD/BCU
The FBI's Weapons of Mass Destruction Directorate (WMDD) was created after September 11, 2001 to provide a cohesive and coordinated approach to countering WMD threats and responding to incidents if they occur. Recognizing the unique and inherent challenges to preventing bioterrorism, the FBI/WMDD/Biological Countermeasures Unit (BCU) conducts extensive outreach to the life sciences community to proactively build mutually beneficial relationships and broaden scientists' understanding of biosecurity concerns.

About AAAS
The American Association for the Advancement of Science (AAAS) is the world's largest general scientific society and publisher of the journal, *Science* (www.sciencemag.org). AAAS was founded in 1848, and serves 262 affiliated societies and academies of science, reaching 10 million individuals. *Science* has the largest paid circulation of any peer-reviewed general science journal in the world, with an estimated total readership of one million. The non-profit AAAS (www.aaas.org) is open to all and fulfills its mission to "advance science and serve society" through initiatives in science policy, international programs, science education, and more.

About UNICRI
The United Nations Interregional Crime and Justice Research Institute (UNICRI) was established by the UN Economic and Social Council in 1968 to assist the international community in the formulation and implementation of effective policies in the field of crime prevention and criminal justice. UNICRI' s goals are: 1) to advance understanding of crime-related problems; 2) to foster just and efficient criminal justice systems; 3) to support the respect of international instruments and standards; 4) to facilitate international law enforcement cooperation and judicial assistance. UNICRI's mission is to advance security, serve justice and build peace in support of the rule of law, human rights protection and sustainable development. UNICRI's main objective is to make a meaningful contribution in the development of a sustainable platform for information exchange and effective policies, and interventions.

Experts and Staff

Working Group Members

Satnam Alag, Ph.D.
Illumina

Kavita M. Berger, Ph.D., *chair*
American Association for the Advancement of Science

Tanya Berger-Wolf, Ph.D.
University of Chicago

Roger Brent, Ph.D.
Fred Hutchinson Cancer Research Center

Susan A. Ehrlich, J.D., LL.M.
(Biotechnology and Genomics)
Judge (ret.), Arizona Court of Appeals

Steve Evans, Ph.D.
Dow AgroSciences

Andrew Hessel
Autodesk, Inc.

Margaret Kosal, Ph.D.
Georgia Institute of Technology

Francesco Marelli, Ph.D.
United Nations Interregional Crime and Justice Research Institute

Simon Mercer, Ph.D.
Microsoft Research

Piers Millet, Ph.D.
International Fellow at Woodrow Wilson Center

Gilbert S. Omenn, M.D., Ph.D.
University of Michigan

Ketan Paranjape, M.S., MBA
Intel

Erik Prentice, Ph.D.
Proactive Worldwide, Inc.

Nathan Price, Ph.D.
Institute for Systems Biology

Charles Schmitt, Ph.D.
Renaissance Computing Institute

Matt Wood, Ph.D.
Amazon Web Services

SSA Edward You, M.S., *chair*
Federal Bureau of Investigation

Jaroslaw Zola, Ph.D.
University at Buffalo, SUNY

Experts

Gaymon Bennett, Ph.D.
Fred Hutchinson Cancer Research Center

Mark Greaves, Ph.D.
Pacific Northwest National Laboratory

Daniel Grushkin
GenSpace

Kenneth Oye, Ph.D.
Massachusetts Institute of Technology

David Shepherd, M.A., M.S.
U.S. Department of Homeland Security

AAAS Center for Science, Technology, and Security Policy Staff

Kavita M. Berger, Ph.D.
Associate Director

Jennifer Roderick, M.Sc., M.S.
Program Associate

Nicholas Bashour
Former intern

Donghun Kang
Asan Academy Fellow

Bitna Lee
Asan Academy Fellow

Eunhye Kim
Asan Academy Fellow

Jen January Therrien
Visiting Scholar

Zachary Watterson
Former Fellow

FBI WMD Directorate Biocountermeasures Unit Staff

Edward You, M.S.
Supervisory Special Agent

William K. So, Ph.D.
Policy and Program Specialist

Sonia Hunt, Ph.D.
Management and Program Analyst

Table of Contents

Report in Brief: National and Transnational Security Implications of Big Data in the Life Sciences **7**

What is Big Data? 8

Figure A: Technical Challenges and Current Solutions of Big Data in the Life Sciences 9

Benefits, Risks, and Solutions for Big Data in the Life Sciences 9

Risk and Benefit Assessment Frameworks 11

Legal, Technical, Institutional, and Individual Solutions 11

Recommendations and Conclusions 11

Chapter 1: National Security and Big Data in the Life Sciences **13**

Special Considerations to Big Data in the Life Sciences 17

Project 17

Chapter 2: Big Data in the Life Sciences **19**

What is Big Data? 21

Figure 2.1 Data Analysis Technologies 22

Understanding Big Data in the Life Sciences 23

Investments in Big Data in the Life Sciences 23

Strategic Drivers 25

Case Study: Omics and Drug Development 27

Case Study: Healthcare Delivery and "Learning Healthcare Systems" 28

Case Study: Ecosystems and the Changing Climate 29

Challenges in Applying Big Data Technologies 31

Conclusions 32

Chapter 3: Assessing National Security Risks and Benefits of Big Data in the Life Sciences **34**

Dealing with Data: The "Dual Use Dilemma" Redefined 35

Risk and Benefit Scenarios 36

Scenario 1: Targeting of Subpopulations using Big Data 37

Scenario 2: Misdirection 40

Scenario 3: Avoiding Detection 42
Assessment Frameworks 44
Risk Assessment Framework 44
Figure 3.1: Conceptual Framework for Qualitative Risk Assessment for Emerging and Enabling Technologies 46
Risk Assessment Framework Questions 47
Benefit Assessment Framework 48
Figure 3.2: Benefit Assessment Framework for Big Data in the Life Sciences Technologies 49
Comparing Risks and Benefits 49
Applying the Risk and Benefit Frameworks to the Three Scenarios 50
Chapter 4: Solutions to Addressing National Security Risks 53
Technical Solutions 54
Legal Solutions 55
Institutional Solutions 57
Individual Solutions 58
Solution Scenarios 59
Solution Scenario 1: Targeting of Subpopulations using Big Data 60
Solution Scenario 2: Misdirection 61
Solution Scenario 3: Avoiding Detection 61
Conclusions 62
Appendix 1: Glossary of Terms 63
Appendix 2: Summary of April 1 Event 73
Appendix 3: Examples of U.S. Government Investments and Initiatives in Big Data 76
Appendix 4: Selected Activities in Big Data in the Life Sciences 79
Appendix 5: Relevant Laws, Regulations, and Guidance 86
Appendix 6: AAAS-FBI Partnership 89

Report in Brief:

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

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), the Biological Countermeasures Unit of the Federal Bureau of Investigation Weapons of Mass Destruction Directorate (FBI/WMDD/BCU), and United Nations Interregional Crime and Justice Research Institute (UNICRI) seek to:

- Examine the risks and benefits associated with Big Data analytics;
- Develop frameworks for risk and benefit assessments 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 potential 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 of Big Data analytics, and existing or needed solutions to address the risks identified. To carry out this project, AAAS/CSTSP, FBI/WMDD/BCU, UNICRI 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. (Chapter 2) 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.

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.,

[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.

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 A lists these challenges and current approaches for addressing them.

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

Technical Challenges	Current Solutions
Lack of standard terminology and language	• Natural language analysis technologies • Specific data collection tools
Lack of access to needed technical infrastructure	• Open-source analytic tools • Cloud-based data storage, sharing, and analysis
Security of data repositories and cyber infrastructure	• Various technologies, including data encryption, access control technologies, digital certificates, segregation of networks • Cyber and data security laws
Data privacy and confidentiality	• Various technologies, including data encryption, access control technologies, segregation of networks • Privacy protection laws • Corporate responsibility and/or norms
Overfitting the model	• Testing the model with different data sources

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. (see Chapter 3) 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).

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.

Chapter 1:

National Security and Big Data in the Life Sciences

Although the concept of "national security" existed before the mid-1940s, it became an explicit concern and part of the United States government's legal infrastructure with the passage of the National Security Act of 1947.[2] The conceptualization of national security has evolved over the past 60+ years as the threats, actors, and relationships among nations have changed, and U.S. statutes, regulations, and executive directives and orders, and international treaties, and initiatives have been directed to this evolving threat landscape. The threat of chemical, biological, radiological and nuclear weapons has figured prominently in this history, and the U.S. and its allies' approaches to addressing the threats from weapons of mass destruction similarly have changed.

The world increasingly is linked by the movement of people and goods, exchange of information, the exponential growth and dissemination of technology, and the interconnected effects of political, economic, and environmental change. This high level of physical and virtual interconnectedness presents a growing number of significant challenges to preventing and mitigating national and transnational security threats, while ensuring a balance between respecting individual privacy and protecting the nation from harm.

What has changed since the passage of the National Security Act?

- Threats to nations more often include non-state/lone actors (both individuals and groups) and other nations.
- The civilian academic, commercial, engineering, health, and environmental sectors are some of the sectors that are significant drivers in science and technology developments that are relevant to national security. However, these efforts often are conducted in a non-integrated, uncoordinated manner. U.S. government agencies often support advanced

[2] 50 USC 401

research and development to achieve national security objectives by funding non-defense scientists in non-defense facilities.
- Scientific knowledge and technologies are primarily being applied to the betterment of society and/or for economic gain, but they are becoming increasingly accessible to a larger number of individuals, including non-state/lone actors who have malevolent intent.
- Open access to knowledge and technologies could offer important benefits to addressing national security risks, including those caused by the science and technology itself.
- The concept of "national security" increasingly involves active participation and cooperation between the public health, safety, and security sectors. In the United States, federal departments and agencies, including several offices in the White House National Security Council, are concerned with the protection of human, animal, and plant health and the environment.

This contemporary understanding of national security is the context of our project on National and International Security Implications of Big Data in the Life Sciences. The structured and less-structured repositories of Big Data are analyzed together to provide useful, actionable information. The information contained in these databases can come from a number of sources, including private (e.g., internet services, health records, insurance records, and commercial entities), government, and public databases, and populates many different repositories. An increasing number of different analytic tools (mostly computational tools) are becoming available to allow limited use of some Big Data by non-experts. The computer systems and networks critical to Big Data efforts are products of the explosive development of cyber infrastructure and share its frailties. This provides a potentially explosive union of open science, open technology, and open data. This project evaluates the important applications and significant risks posed by this combination in the life sciences to national and transnational security.

Building on our evolving conceptualization of national security and today's challenges, we define national security in levels. At the highest level, which could be considered as the strategic level, nations often seek to:

- Maintain their citizens' trust in government,
- Maintain their economic stability and growth,
- Protect their citizens' health, safety, and security, and
- Reinforce and protect their sovereignty and position in the global community.

Despite differences among nations about the relative importance of these four goals, they usually can be viewed as common ground when conceptualizing national security.

Nations achieve their strategic goals through a number of objectives such as:

- Political security, both national and international,
- Military security, both national and international,
- Protection and resilience of critical infrastructure,
- Border security, freedom from external coercion, and geostrategic security,
- Economic and commercial security,
- Security of the environment, natural resources, and energy,

- Health and welfare security, including food security, and
- Security of values, liberties, and privacy.

These objectives are implemented through civilian, diplomatic, law enforcement, and military efforts that address specific types of threats, including biological threats.

The United States counters biological threats through a set of policies and programs that implement the aforementioned objectives. Prior to 2001, the primary efforts in which the U.S. government was involved were the 1972 Convention on the Prohibition of the Development, Production, and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (BWC)[3] and the 10-year long, unsuccessful effort to develop a verification protocol that would enforce the treaty,[4] restrictions on certain dangerous pathogens and toxins through export-control regimens, and cooperative efforts to reduce the threat posed by the enormous biological-weapons program of the former Soviet Union. Additionally, in 1996, the U.S. created the Select Agent Program (SAP), a domestic program that was established to prevent the transfer of certain pathogens that pose a significant risk to public health safety and security.[5] In 2001 and 2002, the SAP was significantly enhanced through Congressional action;[6] these enhanced Biological Select Agents and Toxins Regulations were finalized in 2005 and subsequently revised in 2012.[7] Finally, the U.S. Department of Defense Cooperative Threat Reduction (CTR) program was established in 1991 to reduce the threat of acquisition or transfer of knowledge about nuclear weapons after the fall of the Soviet Union. In 1997, the CTR program was expanded to reduce the threat posed by the former Soviet bioweapons program.[8]

In the wake of the 2001 terrorist attacks and anthrax letters, a series of Presidential strategies and Congressional actions set a new path for the U.S. government to counter biological threats. The primary strategy that has dictated biodefense policy and programs for the past decade is Homeland Security Presidential Directive 10, *Biodefense for the 21st Century*.[9] This Presidential

---

[3] See Biological Weapons Convention website. Accessible at www.un.org/disarmament/WMD/bio. Accessed on September 6, 2014.

[4] An initial attempt at negotiating a verification protocol for the Biological and Toxins Weapons Convention occurred during the 1991 Review Conference, but the U.S. government had prevented the negotiations from taking place citing the significant difficulty of verifying compliance with the treaty, especially at the international level. Instead, the U.S. government agreed to an ad hoc group of scientific experts who would produce a technical report describing the feasibility of possible verification measures. The results of the technical study were "vague and inconclusive". Formal efforts towards the development of a protocol that could be used as deterrence of violations of the treaties were initiated after President Clinton took office in 1992. After over years of negotiation, the U.S. government concluded that the draft verification protocol for the BWC had flaws, in large part because of the political issues associated with sections of the draft protocol. See KD Ward. 2004. The BWC Protocol: Mandate for Failure. The Nonproliferation Review. 11.2 (Center for Nonproliferation Studies). Accessible at http://cns.miis.edu/npr/pdfs/112ward.pdf. Accessed on October 29, 2014.

[5] Antiterrorism and Effective Death Penalty Act of 1996 (42 USC 262, 18 USC 232c).

[6] A detailed description of the governance framework for U.S. Biological Select Agents and Toxins is provided in AAAS, AAU, APLU, FBI. 2013. Implementing the Revised Select Agent Regulations. Accessible at: http://www.aaas.org/report/bridging-science-and-security-biological-research-implementing-revised-select-agents-and-2, Accessed on: September 6, 2014; K Berger. 2010. Select Agent Rules. Encyclopedia of Bioterrorism Defense. Eds. R Zalinskis and RL Katz . (Wiley) Accessible at: http://onlinelibrary.wiley.com/doi/10.1002/0471686786.ebd0191/abstract;jsessionid=9BBEE584A081BB10E5F1A1BB479C57A6.f03t01?deniedAccessCustomisedMessage=&userIsAuthenticated=false. Accessed on: September 6, 2014.

[7] 7 CFR Part 331, 9 CFR Part 121, 42 CFR Part 73.

[8] See Nuclear Threat Initiative information on biological weapons and Russia. Accessible at http://www.nti.org/country-profiles/russia/biological/. Accessed on October 10, 2014.

[9] President George W. Bush. (2004). Homeland Security Presidential Directive-10/National Security Presidential Directive-33: Biodefense for the 21st Century. Accessible at: http://fas.org/irp/offdocs/nspd/hspd-10.html. Accessed on September 6, 2014.

directive described the "essential pillars" of the U.S. national biodefense program as threat awareness, prevention and protection, surveillance and detection, and recovery and response. The strategy described several areas in which the U.S. government would invest and to which civilian, diplomatic, and/or military efforts would contribute, including medical countermeasure research and development, assessments of current and anticipation of future threats, engagement in nonproliferation efforts, protection of critical infrastructure, early detection of possible biological attacks, attribution, national response planning, emergency medical care, risk communication, and decontamination. Consequently, the U.S. focus ranges from preventing threats from deliberate biological threats to an all-hazards approach, which involves preventing, detecting, and responding to natural, accidental, and intentional biological threats.

Through a series of Presidential directives,[10] national strategies,[11] executive orders,[12] and legislation,[13] the U.S. government has established and/or supported efforts to develop vaccines and drugs against priority biological threat agents, public health preparedness and response efforts, biosecurity outreach to the U.S. scientific community, review and oversight of dual use research of concern,[14] detection and biosurveillance of natural and man-made infectious diseases, food defense and agricultural security, laboratory biological safety and security, environmental detection and remediation, and implementation of the BWC. In 2012, the U.S. government issued its Policy for Oversight of Life Sciences Dual Use Research of Concern, which provides the federal governance framework for misuse of certain life-sciences research.[15] In 2014, the U.S. government in coordination with the World Health Organization, Food and Agriculture Organization, and World Animal Health Organization launched the Global Health Security Agenda (GHSA). The GHSA which provides a vehicle to assist countries to meet the core competencies of the 2005 International Health Regulations through the prevention, detection, and response objectives.[16] Nearly fourteen U.S. departments and agencies play a role in countering biological threats domestically and internationally. Moreover, the Departments of

---

[10] Homeland Security Presidential Directive-4/National Security Presidential Directive-17, National Strategy to Combat Weapons of Mass Destruction (2002); Homeland Security Presidential Directive-9, Defense of United States Agriculture and Food (2004); Homeland Security Presidential Directive-10/National Security Presidential Directive-33, Biodefense for the 21st Century (2004); Homeland Security Presidential Directive-5, Management of Domestic Incidents (2005); Homeland Security Presidential Directive-18, Medical Countermeasures against Weapons of Mass Destruction (2007); Homeland Security Presidential Directive-21, Public Health and Medical Preparedness (2007); Presidential Policy Directive-8, National Preparedness (2011).

[11] National Strategy for Countering Biological Threats (2009) and National Biosurveillance Strategy (2012).

[12] Executive Order 13486: Strengthening Laboratory Biosecurity in the United States (2009); Executive Order 13527: Establishing Federal Capability for the Timely Provision of Medical Countermeasures Following a Biological Attack (2009); Executive Order 13546: Optimizing the Security of Biological Select Agents and Toxins in the United States (2010).

[13] Biological Weapons Anti-Terrrorism Act of 1989 (18- USC 175 et seq); Andean Trade Preference Act (including "Chemical and Biological Weapons Control and Warfare Elimination Act of 1991") (22 USC 5601 et seq); Antiterrorism and Effective Death Penalty Act of 1996 (18 USC 175 et seq, 18 USC 232c, 42 USC 262); USA PATRIOT Act of 2001 (18 USC 175); Homeland Security Act of 2002 (6 USC 161 et seq, 6 USC 181 et seq, 6 USC 421 et seq); Farm Security and Rural Investment Act of 2002 (7 USC 7901); Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (42 USC 201); Project Bioshield of 2004 (42 USC 201); Pandemic and All-Hazards Preparedness Act of 2006 (42 USC 300).

[14] The U.S. government defines "dual use research of concern" as "life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops and other plants, animals, the environment, materiel, or national security." United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern. (2012). Accessible at http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf. Accessed on September 6, 2014.

[15] United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern. (2012). Accessible at http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf. Accessed on September 6, 2014.

[16] Global Health Security Agenda (2014) Accessible at: http://www.globalhealth.gov/global-health-topics/global-health-security/ghsagenda.html. Accessed on September 6, 2014.

State and Defense Cooperative Threat Reduction Program has expanded its biological engagement to include efforts in more countries.[17]

Special Considerations to Big Data in the Life Sciences

The multidisciplinary nature and inherent complexity of Big Data present significant challenges to its advancement, application and utility, and security. The interdependence of data, computer, and life sciences, the amount and diversity of data collected, the increasing number of and individuals with the needed skillset, the rapid advances made in computational analytic tools, and the nascent efforts towards developing standards for data sharing will increase the potential for benefits to be reaped from Big Data technologies. They also provide increased opportunity for adversaries to exploit or steal valuable and/or sensitive data, use Big Data technologies to inflict harm on individuals, groups, or nations, and/or disrupt Big Data applications to cause negative economic, political, health, or societal consequences.

Big Data in the life sciences can be applied to several life sciences issues, including monitoring environmental changes and health, infectious-disease surveillance, improvements in healthcare, drug discovery, and design of new molecules and microbes. Applications that generate designs for development of vaccines, medications, proteins, other molecules, and microbes require actual laboratory research. Typical barriers to entry, including technical barriers and governance frameworks, exist for translating the computer-generated design to actual product, molecule, or pathogen. Beyond access controls, encryption, and other common data and cyber security technologies, no solutions exist that prevent or mitigate attacks on databases or the cyber infrastructure that support Big Data in the life sciences, which could result in consequences to the life science, commercial, and health sectors.

Leveraging of Big Data analytics to support useful (and profitable) applications will increase as the amount of data increases, data sharing becomes easier, and analytic tools improve. These characteristics also provide increased opportunities and vulnerabilities for adversaries to steal sensitive data, use Big Data technologies to inflict harm on individuals or groups, support a criminal enterprise, or disrupt Big Data applications to cause negative economic, political, or other societal outcomes.

Project

The AAAS/CSTSP, the Biological Countermeasures Unit of the WMD Directorate of the Federal Bureau of Investigation (FBI/WMDD/BCU), and United Nations Interregional Crime and Justice Research Institute (UNICRI) embarked on a one-year project on Big Data in the Life Sciences and National Security. The goal of the project was to review the current scientific and technological advancements in Big Data and analytics, applications of this field to different life-sciences fields, and potential benefits or risks to national security.

[17] Department of State Biosecurity Engagement Program. Accessible at http://www.bepstate.net/. Accessed on October 29, 2014; Department of Defense Cooperative Biological Engagement Program. Accessible at http://www.dtra.mil/Missions/Nunn-Lugar/BiologicalThreatReductionProgram.aspx. Accessed on October 29, 2014.

To conduct this multi-faceted project, AAAS/CSTSP, FBI/WMDD/BCU, and UNICRI established a working group of experts with in-depth knowledge about Big Data in healthcare and other life sciences applications, and expertise in the computer, data, and life sciences and/or national and transnational security issues. The members of the working group were from academia, private industry, government, or intergovernmental organizations. In addition, the project involved four ad hoc experts in cyber security, chemical and biological defense, amateur biology, and philosophy. This diverse set of experts guided the project activities and outcomes, and assisted in the development of the final reports.

With the assistance of the working group, AAAS/CSTSP, FBI/WMD/BCU, and UNICRI held a large, public event on "Big Data, Life Sciences, and National Security" on April 1, 2014 in Washington, DC. The purpose of the event was to engage experts in science, security, health, and public policy to engage stakeholders in thoughtful dialogue about the benefits of Big Data in the life sciences and to begin identifying associated risks or vulnerabilities to biological security. The event webcast is archived at http://www.aaas.org/cstsp/bigdata/publicevent.

Immediately following the event, working-group members and speakers discussed the major themes[18] from the event in a roundtable discussion. From this discussion, the working group selected three topics on which to focus the project: strategic drivers of technology advancement and use (Chapter 2), capabilities including both risks and benefits of Big Data in the life sciences to national security (Chapter 3), and solutions for preventing or mitigating the risks (Chapter 4).

This report provides frameworks through which scientists and security experts can evaluate and compare the potential risks and benefits of Big Data technology, and identify specific legal, technical, institutional, and individual solutions to prevent or mitigate the identified risk. The study about which this report was written used only technologically robust risk and benefit scenarios basing all speculation on published data, including data about computer viruses and hacking.

The Appendices include the Glossary of Terms (Appendix 1), Summary of April 1 Event (Appendix 2), Examples of U.S. Government Investments and Initiatives in Big Data (Appendix 3), Selected Activities in Big Data in the Life Sciences (Appendix 4), Relevant Laws, Regulations, and Guidance (Appendix 5), and AAAS-FBI Partnership (Appendix 6).

[18] The major themes that emerged from the public event are published online at http://www.aaas.org/news/big-data-life-sciences-and-national-security-event-summary-recap. Accessed on September 4, 2014.

Chapter 2:

Big Data in the Life Sciences

Big Data analytics is one of the fastest growing fields in the world today because of the increasing amounts of data collected through digital means and through observation, the existence of strong economic drivers for generating and capitalizing on data and various applications, and rapidly advancing analytic technologies. Integration and analysis of data hold great promise in uncovering new insights and/or approaches in health, peace, agriculture, energy, finance, national and transnational security, and the environment. For example, the commercial sector uses Big Data technologies to anticipate consumer needs, the health sector uses them to improve health outcomes and/or empower people to live more active and health-conscious lives, and the national security sector uses the technologies to identify potential adversaries and characterize threats. At the same time, individuals, communities, nations, and regions face risks from vulnerabilities in the systems through hacking of databases or cyber infrastructure supporting data analysis, or intentional exploitation of Big Data to design targeted and/or harmful biological agents. Unlike other security risks often associated with the life sciences, the potential risks of Big Data in the life sciences rely on attacking the underlying data and cyber infrastructure and/or inappropriately using data and analytic technologies, not theft of actual pathogens or toxins. Recent examples demonstrate the utility of Big Data analytics in addressing societal needs and reveal the impacts from the risks from vulnerabilities in the cyber and data infrastructure:

> ***The 2014 Ebola Outbreak:*** The current Ebola virus epidemic in West Africa has claimed several thousands of lives and this number will climb to well over 10,000 lives before the epidemic is mitigated.[19] The combination of infection in urban areas, poor public health systems, and weak health governance in the countries experiencing the worst of the epidemic along with the lack of a "proven" safe vaccine or effective drug against Ebola have contributed to the scale and devastation of the current outbreak. HealthMap, an infectious-disease surveillance tool that uses Big Data analytics, identified the Ebola outbreak nine days before

[19] K Krupferschmidt. 2014. Disease Modelers project a rapidly rising toll from Ebola. *Science*. Accessible at: http://news.sciencemag.org/health/2014/08/disease-modelers-project-rapidly-rising-toll-ebola. Accessed on September 3, 2014.

the World Health Organization announced the epidemic.[20] The head of HealthMap suggests that the addition of more data streams to the algorithm, including cell-phone usage data, could help predict the spread of Ebola virus.[21]

Additionally, researchers at the Broad Institute described the initial case of Ebola infection in Sierra Leone and provided insight into how the virus arrived in West Africa through the analysis of viral genome sequence data of Ebola virus in patient samples from Sierra Leone.[22]

Medical Care in Emergencies. In 2011, the Office of the Assistant Secretary for Preparedness and Response (ASPR) of the U.S. Department of Health and Human Services held an emergency preparedness exercise in New Orleans to evaluate new approaches for using medical data held by federal agencies "to identify and reach medical high-risk citizens" in an emergency.[23] In this exercise, federal officials identified potentially vulnerable people through evaluation of Medicare insurance claims and informed the local public health authorities about individuals with high-risk medical conditions. Building on this exercise, ASPR is beginning to scale these efforts across the United States.[24]

Cyber Attack of Electronic Health Records. Increasingly, health-care records in the United States and several other countries are being transferred and stored electronically. The purpose is to enhance health-care services and provide the best care possible to patients in a coordinated manner. However since 2009, more than 900 major breaches of protected health information have occurred affecting an estimated 30.1 million people and costing the healthcare industry about $5.6 billion each year.[25] In an example of a recent data breach from August 2014, non-medical patient data from 4.5 million individuals was stolen when the network of the second-largest for-profit health system in the United States was attacked by a

---

[20] The Associated Press. 2014. Computer bot gave early warning of Ebola outbreak. Accessible at: http://www.cbsnews.com/news/computer-bot-gave-early-warning-of-ebola-outbreak/. Accessed on September 3, 2014.

[21] BetaBoston. 2014. Boston researchers trying to forecast Ebola's spread, but more data would help. Accessible at: http://betaboston.com/news/2014/09/02/boston-researchers-trying-to-forecast-ebolas-spread-but-more-data-would-help/. Accessed on September 3, 2014.

[22] G Vogel. 2014. Genomes reveal start of Ebola outbreak. *Science*. Accessible at: http://www.sciencemag.org/content/345/6200/989.full. Accessed on September 3, 2014; SK Gire et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 28 Aug 2014. Accessible at: http://www.sciencemag.org/content/early/2014/08/27/science.1259657.full#aff-2. Accessed on September 3, 2014; F Collins. 2014. Using Genomics to Follow the Path of Ebola. Accessible at: http://directorsblog.nih.gov/2014/09/02/using-genomics-to-follow-the-path-of-ebola/. Accessed on September 3, 2014.

[23] City of New Orleans. 2013. City of New Orleans to Hold Medical Needs Emergency Preparedness Exercise. Accessible at: http://www.nola.gov/mayor/press-releases/2013/20130607-city-of-new-orleans-to-hold-medical-needs/; Accessed on: October 7, 2014.

[24] S. Fink. 2014. U.S. Mines Personal Health Data to Find the Vulnerable in Emergencies. *The New York Times.* Accessible at: http://www.nytimes.com/2014/05/16/us/us-mines-personal-health-data-to-aid-emergency-response.html?action=click&module=Search&region=searchResults&mabReward=relbias%3Ar&url=http%3A%2F%2Fquery.nytimes.com%2Fsearch%2Fsitesearch%2F%23%2Flurie%2F30days%2F&_r=1. Accessed on: October 7, 2014.

[25] J Millman. 2014. Healthcare Data Breaches have hit 30M Patients and Counting. *The Washington Post*. Accessible at: http://www.washingtonpost.com/blogs/wonkblog/wp/2014/08/19/health-care-data-breaches-have-hit-30m-patients-and-counting/. Accessed on: October 7, 2014.

Chinese entity.[26] These records may comprise more than personally identifying and/or financial information; stolen records may have included data from clinical research and/or drug studies.

Harnessing the beneficial applications of Big Data and data analytics to address societal needs, improve services (e.g., healthcare), and enhance marketing and other commercial efforts, all while preventing its harmful use and reducing its vulnerability to attack, presents a significant challenge. In May 2014, the White House released two reports on Big Data and privacy, *Big Data: Seizing Opportunities, Preserving Values*[27] and *Big Data and Privacy: A Technological Perspective*.[28] These reports highlight the legal and technical solutions for addressing the privacy risks of Big Data.

The AAAS/FBI/UNICRI report describes the current state and drivers of Big Data analytics as it relates to life science pursuits, the characteristics of Big Data that both suggest beneficial application for and raise concerns about potential national and transnational biological risks, frameworks for qualitatively evaluating enabling technologies such as Big Data in the life sciences, and current solutions for addressing the possible risks from vulnerabilities or exploitation.

What is Big Data?

Although no formal definition of Big Data exists, it does depend on the size and heterogeneity of the data, and the scale and scope of analytic operations made possible by that size. Big Data often has several characteristics, commonly referred to as the four Vs[29]:

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).

The data generated comes from a number of sources collected and catalogued in different ways and stored in different repositories. 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 (e.g., internet search terms, social media, purchasing trends, mobile app data, emails, and cell phone data and logs), or digitized from observations (e.g., electronic health records, species data, publications, scientific results, and genomics data). The data are

[26] D Munro.2014. Cyber attack net 4.5 million records from large hospital system. *Forbes*. Accessible at: http://www.forbes.com/sites/danmunro/2014/08/18/cyber-attack-nets-4-5-million-records-from-large-hospital-system/. Accessed on October 7, 2014.
[27] Executive Office of the President. Big Data: Seizing Opportunities, Preserving Values. May 2014.
[28] President's Council of Advisors on Science and Technology. Big Data and Privacy: A Technological Perspective. May 2014.
[29] The 3Vs were taken from a definition developed by Gartner, Inc., which defines Big Data as "information assets *that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."*

heterogeneous, often containing errors, and/or incomplete. Data can be collected through active or passive means, deposited voluntarily or not, and/or deliberately produced, for example through scientific pursuits. The datasets can be structured or unstructured, often huge in size (exceeding petabytes), and rapidly changing. Several distinct datasets are integrated and analyzed together, which contribute to their characterization as Big Data.

Big Data analytics is a process in which knowledge is extracted from data using non-trivial tools developed from mathematics and high-performance computing or very large data processing clusters. 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 2.1 lists different types of data analytic technologies and their uses with regard to Big Data. 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.

Figure 2.1 Data Analysis Technologies

Data Mining

- Identify relationships among information but not causality
- Mathematics, computer science, artificial intelligence, and machine learning
- Examples: classification algorithms; clustering algorithms; regression algorithms (i.e., numerical prediction algorithms); association tools; anomaly-detection algorithms; summarization tools

Data Fusion

- Integrate heterogeneous datasets
- Requires systems to communicate and exchange data
- Examples: sensor networks; video/image processing; robotics and intelligent systems

Data Integration

- Broadly combine data repositories, and keep a larger set of information

Image and Speech Recognition

- Extract information from large amounts of images, videos, and recorded or broadcast speech
- Examples: scene extractions; facial-recognition technologies; automated speech recognition

Natural Language Processing

- Understand natural human language of input data

Machine Learning

- Learn from input data

Bayesian Analysis

- Combine information about a population parameter with information contained in a sample

Social-network Analysis

- Extract "information from a variety of interconnecting units under the assumption that their relationships are important and that the units do not behave autonomously" (PCAST, 2014)
- Use different technologies, such as clustering association and data fusion

Understanding Big Data in the Life Sciences

To the life-sciences community, "Big Data" refers to volumes of data that are integrated and/or analyzed together to provide valuable information. Data in life sciences repositories comes from different sources which are highly curated. Life science data come in different forms, including text (e.g., publications), DNA or protein sequence information, images, and numbers. For biological Big Data, the data in the repositories almost never changes once added to, but it is rather continuously augmented thereby increasing the overall size of the repositories in which the data are stored. The data repositories are often heterogeneous, contain extraneous data, and evolve over time (e.g., longitudinal and epigenetic studies).

The National Institutes of Health's (NIH) Big Data to Knowledge working group defines the term Big Data as a function rather than a thing. To the NIH, Big Data encapsulates "the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various '-omics'), exposure, health, behavioral, and many other types of biological data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively."[30]

Investments in Big Data in the Life Sciences[31]

After the White House released its Big Data Initiative,[32] several U.S. government agencies support research and development of analytic technologies, data collection and integration, and workforce development. Some invest to address specific societal needs. For example, the National Institutes of Health has established its Big Data to Knowledge (BD2K) initiative to develop approaches, standards, methods, software, and workforce and competencies to enhance use of Big Data in the life sciences. The U.S. Agency for International Development has invested in the Global Forest Watch, which integrates satellite images, data sharing, and crowd sourcing to help manage forests. The Department of Commerce has established a program to develop Big Data tools to support small businesses. The Department of Agriculture invests in the cyber infrastructure that supports Big Data analyses to address the risk of changing climates to agriculture. The Intelligence Advanced Research Projects Agency invests in research and technology development to forecast science and technology advances and identify infectious disease outbreaks before official announcements by public health organizations. The National Science Foundation funds a number of projects that improve Big Data tools and use Big Data in research. The U.S. Geological Survey invests in programs that use Big Data to provide insight about geological phenomena. The National Aeronautics and Space Administration promotes research with large, scientific datasets in new ways. The National Institute of Standards and

[30] National Institutes of Health Big Data to Knowledge Mission Statement. Accessible at http://bd2k.nih.gov/about_bd2k.html#sthash.sQBuRetr.dpbs. Accessed on November 6, 2014.

[31] Appendix 3 lists several U.S. government investments in Big Data and Appendix 4 lists several publicly and privately-supported efforts on Big Data in the life sciences.

[32] White House Office of Science and Technology Policy. 2012. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments. Accessible at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf. Accessed on: October 8, 2014; See also, The Networking and Information Technology Research and Development Program. Accessible at https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title. Accessed on: October 8, 2014.

Technology has initiated a program on data science attempting to define Big Data and to develop a reference framework, advance measurement techniques, and enable collaboration.[33]

Other scientific agencies support the development of analytic tools and infrastructure to advance research involving Big Data. The Defense Advanced Research Projects Agency has invested in the development of computational techniques and software to process and analyze incomplete and imperfect data, to visualize the results of data, and to improve machine learning. The Department of Energy supports efforts that address the networking challenges of Big Data and develop tools to manage and visualize Big Data. The National Institutes of Health, National Science Foundation, and Defense Advanced Research Projects Agency provide funding for the collection and analysis of large neuroscience datasets and for advances in new analytic tools such as data visualization tools, as part of the U.S. Brain Initiative.[34]

Several other U.S. government agencies support programs, projects, and initiatives that involve Big Data, although these programs are not explicitly described as Big Data projects. Among these are the Department of Homeland Security Science and Technology Directorate, National Oceanographic and Atmospheric Administration, and national intelligence agencies.

In the life sciences, significant investments by the U.S. government, other governments, private foundations (e.g., Wellcome Trust) and for-profit corporations have been made in human and comparative genomics. The Human Genome Project was the first foray into uncovering health information from the human genome. Subsequent initiatives include the 1000 Genomes Project,[35] Personal Genome Project,[36] Genographic Project by the National Geographic Society,[37] The Cancer Genome Atlas Program,[38] Million Veterans Program of the U.S. Department of Veterans Affairs,[39] and the genome-wide association studies of the U.S. National Institutes of Health.[40] In addition, a growing number of countries are supporting human genome sequencing initiatives that aim to collect information about the genomic sequences of their citizens.[41] Finally, over the past decade, direct-to-consumer companies have emerged to provide people the opportunity to have their genomes sequenced and report on possible future health

---

[33] A Talukder. 2014. National Institute of Standards & Technology Data Science Program. NIST Data Science Symposium. Accessible at http://www.nist.gov/itl/iad/upload/NIST-DS-program_symposium-slides_NIST-Presentation-final.pdf. Accessed on October 8, 2014; See also the NIST Data Science Symposium 2014.

[34] National Science Foundation. Understanding the Brain. Accessible at http://www.nsf.gov/news/special_reports/brain/initiative/. Accessed on October 8, 2014; National Institutes of Health, The Brain Initiative. Accessible at http://www.braininitiative.nih.gov/index.htm. Accessed on October 8, 2014; Defense Advanced Research Projects Agency. SUBNETS. Accessible at https://www.fbo.gov/index?s=opportunity&mode=form&id=8a9493faea9edcf0cca595b8ed5de6dd&tab=core&_cview=1. Accessed on October 8, 2014; Defense Advanced Research Projects Agency. RAM. Accessible at https://www.fbo.gov/index?s=opportunity&mode=form&id=925a0e2faf1c2e3c3782e1788fcc660d&tab=core&_cview=0. Accessed on October 8, 2014.

[35] See http://www.1000genomes.org/. Accessed on October 8, 2014.

[36] See http://www.personalgenomes.org/. Accessed on October 8, 2014.

[37] See https://genographic.nationalgeographic.com/. Accessed on October 8, 2014.

[38] See http://cancergenome.nih.gov/. Accessed on October 8, 2014.

[39] See http://www.research.va.gov/MVP/. Accessed on October 8, 2014.

[40] See http://gds.nih.gov/. Accessed on October 8, 2014.

[41] Examples of nations supporting large-scale genome sequencing include the Kingdom of Saudi Arabia (See http://www.bbc.com/news/health-25216135. Accessed on October 8, 2014), Netherlands (LC Francioli et al. Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population. 29 Jun 2014. *Nat Gen*. 46: 818-825), and Malaysia (L Wong et al. Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. 10 Jan 2013. *AJHG*. 92(1): 52-66).

risks based on literature reviews. Although a single genome might not constitute Big Data, many genomes (especially if combined with clinical, lifestyle, health, environmental, behavioral, agricultural, and/or other biological data) is biological Big Data. Analytical comparisons of multiple genomes (i.e., genome wide association studies) are becoming a more prevalent part of many health applications that are currently being developed.

Computer companies such as Apple, Google, Microsoft, Intel, and IBM, along with several academic computer science departments, drive the development of new or enhanced analytic tools and applications of biological Big Data. The computer companies support healthcare or life sciences initiatives, and/or develop strategic partnerships with institutions, such as universities, to improve analytic tools, particularly in the context of healthcare, disease characterization, and agriculture.[42] Illumina, whose next-generation sequencers generate large amounts of -omics data, is investing in the use of big data technologies, especially within the biological context.[43] Illumina's big data technologies empower the interpretation of large-scale data in the context of current knowledge, and enables researchers to generate new insights by looking at the data in aggregate. More broadly, companies including Amazon,[44] Target,[45] and Walmart[46] developed and use Big Data analytic tools to anticipate product purchases and capitalize on growing market interest in cloud processing capabilities. In addition, public/private partnerships exist to enhance drug development; examples include: 1) efforts led by the tranSMART Foundation and carried out with multiple pharmaceutical companies, data companies, and universities and with strong ties with European counterparts;[47] and 2) collaborations between IBM-Watson and universities.[48]

Finally, at least one U.S. state invests in Big Data New York state has allotted funds for the development of health-care and research consortia and a genomics center, all of which contribute to the collection and analysis of genomics, clinical, and other health-related data to improve the state's economy and population health. New York has considered workforce and research needs, and the security risks of the supportive cyber infrastructure. In addition, a few countries, such as the United Kingdom, support Big Data collection and analysis in healthcare.

Strategic Drivers

Research, investment, and interest in Big Data have significantly increased over the past two decades.[49] The primary driving forces behind the practical application of biological Big Data are

[42] "Precision agriculture" is the optimization of agricultural practices based on analysis of several environmental, health, and agricultural datasets. This application is enabled by big data analysis.
[43] See http://bioinformatics.illumina.com/informatics/biological-data-mining.html. Accessed on October 8, 2014.
[44] J Leber. 2013. Amazon Woos Advertisers with What it Knows about Consumers. MIT Technology Review. Accessible at http://www.technologyreview.com/news/509471/amazon-woos-advertisers-with-what-it-knows-about-consumers/. Accessed on October 8, 2014; See also Analyze Big Data for Consumer Applications with Looker BI and Amazon Redshift. Accessible at http://www.slideshare.net/AmazonWebServices/aws-partner-webcast. Accessed on October 8, 2014.
[45] K Hill. 2012. How Target Figured Out a Teen Girl was Pregnant Before Her Father Did. Forbes. Accessible at http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/. Accessed on October 8, 2014.
[46] See http://www.walmartlabs.com/category/bigdata/. Accessed on October 8, 2014.
[47] See http://transmartfoundation.org/. Accessed on October 8, 2014.
[48] See http://www.research.ibm.com/client-programs/accelerated-discovery-lab/. Accessed on October 8, 2014.
[49] G Press. 2013. A Very Short History of Big Data. Forbes. Accessible at http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/. Accessed on October 8, 2014; U Friedman. 2012. Big Data: A Short History. Foreign Policy. Accessible at http://www.foreignpolicy.com/articles/2012/10/08/big_data. Accessed on October 8, 2014.

the developments in computational power, data storage, and access to networks and equipment. By decreasing the cost of entry of both technical expertise and financial resources needed, a greater number of individuals and organizations are able to collect, store, and analyze Big Data in the life sciences. In addition, within the life sciences community, Big Data technologies are enabling an increasing amount of research that is conducted by and/or otherwise heavily dependent on computer-bases analyses.

Understanding the drivers for the advancement and application of Big Data technologies over the past decade is important for evaluating future trends of the technology.

To evaluate the strategic drivers of progress and use of Big Data in the life sciences in this study, the AAAS/CSTSP developed three use cases of Big Data in the life sciences: Omics and Drug Development, Healthcare Delivery and Learning Healthcare Systems, and Ecosystems and the Changing Climate.

Key individuals[50] with expertise in one of these three use cases were asked a series of questions:

1) What does the term "Big Data" mean to your organization?
2) In what processes or activities does your organization use Big Data?
3) What types of data sources and analytic tools do you currently use?
4) In your opinion, what drives the advancement of Big Data and the analytic tools that you use?
5) Have you faced any difficulties or barriers with Big Data thus far?
6) What is your organization's desired end state with regards to Big Data? What, if any, barriers currently exist to either prevent or limit achieving your desired end state(s)?
7) Do restrictions, legislation, or control mechanisms exist for the work you are currently performing? If not, how would the introduction of such measures affect your organization's use of Big Data in the life sciences to achieve its desired end state??

The use cases (see the text boxes) and answers to these questions are presented below.

[50] These experts were individuals who lead, financially support, or apply Big Data in the life sciences within their organizations. The organizations included Belmont Forum; Chematria; Empire State Development's Division of Science, Technology and Innovation; Express Scripts; Healthcore Inc.; Kaiser Permanente; Massachusetts Institute of Technology; National Center for Advancing Translational Sciences (NCATS); National Ecological Observatory Network (NEON); National Health Service (NHS) United Kingdom; National Institutes of Health (NIH); National Science Foundation's (NSF) Geosciences Directorate; New York Genome Center; Northrop Grumman; Patient-Centered Outcomes Research Institute (PCORI); Division of Systems Medicine, Pharmaceutical Research and Manufacturers of America (PhRMA); ScienceScape; Skoll Global Threats; Stanford University; UK Biobank; United Healthcare; and the Wellcome Trust.

Case Study: Omics and Drug Development

Advances in genomics, proteomics, and metabolomics have begun to transform the drug discovery process. As new research in epigenomics, transcriptomics, pharmacogenomics, immunogenomics, and the human microbiome increases, more information about human physiology, its relationship with the external environment, the effects of chemicals or other foreign particles, and the causes of human disease is added to a continuously growing pool of knowledge.

The relatively new field of personalized medicine (sometimes referred to as "precision medicine") strives to harness the knowledge gained from the –omics efforts to enhance medical practice and patient outcomes. This field relies on associating the heterogeneity of patients with the same diagnosis and the cellular heterogeneity of cancers and of other types of diseases to inform medical care and treatment. Specifically, data from biomedical and clinical research, electronic medical records (which are beginning to include patient genomic information), and other patient studies are analyzed together (i.e., GWAS) to identify medical diagnostics and treatments that can be tailored for individual patients.[1]

For precision medicine, having and developing drugs against different diseases and symptoms is important. However, drug development is a long, arduous, and expensive process. On average, drug discovery and drug development– from identification of drug targets to U.S. Food and Drug Administration approval and commercialization – takes approximately 20 years and $1 billion USD. Despite significant investment by companies, the number of promising compounds that result in marketed drugs is very low.[2] Academic and industrial researchers have begun to use omics data and emerging analytic tools, such as Bayesian network analysis and machine learning,[3] to develop drugs in a more targeted manner, hoping that this will reduce the time and cost involved in drug development. These methods may also be used to identify and develop treatments targeting what have been identified as orphan diseases.

[1] National Research Council. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. 2011. (National Academy Press, Washington, DC) p.105; Institute of Medicine. Evolution of Translational Omics: Lessons Learned and Path Forward. 2012. National Academy Press, Washington DC.

[2] J.A. DiMasi et al. The Price of Innovation: New Estimates of Drug Development Costs. J Health Econ. 2003. 22: 151-185; M. Hay et al. Clinical Development Success Rates for Investigational Drugs. 2014. Nature. 32: 40-51.

[3] Examples include Y Bai et al. 2012. BioSignature Discovery System: A Next Generation Solution for OMICS Analysis and Modeling in Drug Discovery and Personalized Medicine. Seralogix Next Generation Discovery. Accessible at http://www.seralogix.com/images/publications/BioSignatureReadibleManuscriptl.pdf. Accessed on October 9, 2012 and S Ekins et al. Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery. Chemistry and Biology. 21 Mar 2013. 20(3): 370-377.

Case Study: Healthcare Delivery and "Learning Healthcare Systems"

The U.S. system of healthcare involves many organizations that rely on Big Data analyses for treatment from the time an ill patient seeks medical care to the time (s)he has recovered In addition, analysis of Big Data in the life sciences can inform the promotion of healthy behaviors, and monitoring risk factors in defined populations. The widespread adoption of health information technology, which includes electronic medical records, has changed the way healthcare is provided to patients in need in the United States and many other countries. Big Data analyses in healthcare attempts to improve patient outcomes, prevent medical errors, and reduce overall health-care costs through the use of computer systems and storage of long-term health data in digital forms.

Clinicians, medical facilities (including medical centers, hospitals, clinics, private offices, and urgent care facilities), consumers, insurance providers, payers, pharmacies, pharmaceutical and biotechnology industries, the U.S. Centers for Disease Control and Prevention (CDC), and the Food and Drug Administration (FDA) each uses Big Data to analyze and inform their decisions.

- Clinicians and medical facilities use electronic medical records to monitor their patients more closely and to inform insurance providers and payers of the procedures needed and costs associated. These records store a patient's personal information, medical and family history, genomic sequences in many cases, vaccination and treatment record, payer information, and other relevant information.
- With assistance from an enormous consulting industry, providers maximize billing (e.g., improved risk calculations and amortization calculations).
- Mobile diagnostic applications and websites provide consumers with the tools to collect and evaluate their own health statistics, and inform their decisions about healthcare. Consumers increasingly are voluntarily uploading their data to mobile devices, websites, and the cloud.
- Insurance companies and government payers collect data from electronic medical records, past insurance claims, and other data sources to analyze whether and at what amount to cover medical expenses. Health benefits managers use similar information to negotiate discounts on charges.
- Pharmacies collect purchase information of vaccines, drugs, over-the-counter diagnostic tests, and other relevant products to restock their inventory and preposition products to meet anticipated consumer needs.
- Data from pharmacy sales, clinical visits, diagnostic tests, and other sources are integrated into "syndromic surveillance" systems by the CDC to identify outbreaks of public health concern. The CDC, along with the Departments of Agriculture, Defense, and Homeland Security, conduct disease surveillance on human, animal, and plant health. These surveillance activities integrate and analyze data from a variety of sources from laboratory-confirmed diagnoses to reports of possible infection (or suspected cases).
- Pharmaceutical and biotechnology companies use big data technologies to identify drug targets, develop candidate drugs or medical devices, and evaluate clinical trial data in clinical trials.
- The FDA collects data on pre-approved and approved products it regulates to identify any adverse events associated with use of those products.

Many commercial applications using individual health information may NOT be covered under HIPPA protections.

These organizations co-exist in a system that involves information and analytic results being shared in many directions to multiple sources at any given time. The schematic depicts data sharing between different elements of the healthcare system.

Case Study: Ecosystems and the Changing Climate

The complex and diverse challenges presented by the changing climate provide new opportunities for Big Data analyses to be used for developing a better understanding of the potential effects of the changing climate on agriculture, critical infrastructure, and ecological systems. Several organizations - including the Belmont Forum, the National Ecological Observatory Network, Skoll Global Threats Fund, and the UN Global Pulse Big Data Climate Challenge - are using data-driven models and evidence to better identify how barriers to sustainability can be removed in addition to improving resilience to extreme weather events.

The Belmont Forum, is a group of the world's major or emerging funders of global environmental research and international science councils. The Forum is guided by the Belmont Challenge, which is intended to "deliver knowledge needed for action to avoid and adapt to detrimental environmental change including extreme hazardous events."[1] Meeting this challenge requires assessments of risks, impacts, and vulnerabilities due to climate change through regional and decadal-scale analysis and prediction models.

The National Ecological Observatory Network (NEON) is a continental-scale observation system sponsored by the National Science Foundation. NEON has 106 data-collection sites across the United States. Each site has been selected to represent one of twenty different eco-climatic domains, which include distinct landforms, vegetation, climate and ecosystem processes. At each site, more than five hundred measurements are continuously monitored using sensors and an airborne remote sensing platform. By 2017, all of the data will be available in open-access with 30% of data being available for open access by January 2016.Over the next thirty years, NEON is expected to provide ecological data on the causes and consequences of the changing climate, land use, and invasive species.

The Skoll Global Threats Fund is preparing to address global threats, such as climate change and water security through the use of Big Data.

The UN Big Data Climate Challenge was established to enhance economic arguments for taking action against climate change. The challenge supports research projects throughout the world that use big data analytics to address real-world policy problems associated with climate change.

[1]The Belmont Forum. Accessible at https://igfagcr.org/belmont-challenge. Accessed on November 6, 2014.

The most common driving force behind the investment in and application of Big Data in the life sciences is the availability of new analytic tools and enhanced computational infrastructure. The ability to solve complex problems improves because greater amounts of data can be stored (e.g., in the cloud or in private, public and/or government data centers) and new computational methods are being developed to analyze large, complex, multi-source datasets. Whereas the number of users of Big Data tools once was limited by the availability of supercomputers and individuals with tacit knowledge and expertise, the existence of distributed data centers or even servers throughout the world to store and support data analyses no longer restricts the overall number of individuals who develop and use Big Data technologies. The increase in computational processing power concomitant with algorithmic advancements has significantly reduced the cost of entry to use Big Data technologies, resulting in an increase in the accessibility of Big Data technologies by individuals and smaller organizations.

Beyond availability of new data collection, storage, integration, and analysis technologies, the healthcare, biomedical research and -omics, and environmental sectors each have different drivers of advancement and use of Big Data analytics. The table below describes the drivers for each case study.

		Economic Driver:	**Social Drivers:**
Case Studies	Omics and Drug Development	Developing a drug from discovery to approval by the Food and Drug Administration takes on average 20 years and $1 billion. Big Data analysis provides a more cost-effective, faster way for pharmaceutical companies to design possible drug candidates than the current laboratory-based methods.	Research and education is a key strategic driver of Big Data technologies within the United States. These technologies are thought to improve research progress and education, and could help the United States maintain its competitiveness in the life sciences. For example, the state of New York has invested in a genomics center to improve scientific advancement, health outcomes, and prosperity.
	Healthcare Delivery and Learning Healthcare	In a competitive and market-driven healthcare environment, as in the United States, economics and financial gain are key strategic drivers for the use of Big Data. In a market where individuals have the ability to choose their healthcare provider, they are likely to be influenced by value for money, security of personal and healthcare data, and additional services such as prescription services, related healthcare information and benefits. Healthcare providers are driven to provide the best possible service, reduce financial costs, and improve profitability. Several healthcare providers are actively engaged in research involving Big Data to evaluate health trends and diagnoses of individuals, which could enhance benefits to healthcare subscribers making the healthcare plans more appealing for covered and new subscribers. In a single, non-competitive healthcare provider environment, as in the United Kingdom's National Health Service, economics is a key	Patient well-being and patient care are significant drivers for use of Big Data technologies in the healthcare industry. Consumers also drive the advancement and use of Big Data technologies, particularly for new, mobile tools to monitor personal health and fitness, and to improve individual health outcomes.

		strategic driver, but the approach is fundamentally different. For healthcare providers in a non-competitive environment, research into and investment in Big Data is driven by the desire to establish the most cost-effective and efficient healthcare service possible.	
	Environment and Changing Climate	(Although agricultural commodities would be an economic driver of "precision agriculture," the environment use case did not include agriculture.)	The most common driver for use of Big Data technologies in environmental studies is scientific discovery informing the prevention or mitigating of global-scale calamities.

Challenges in Applying Big Data Technologies

Several challenges emerged during the expert consultations regarding the drivers of technology and uses in biology, health, and the environment:

The biggest challenge facing Big Data in the life sciences is **the lack of standard terminology and ontology** regarding stored data. For each case study, the lack of standard language for data was highlighted as the biggest barrier to fully realizing the potential of Big Data technologies. In healthcare, data from each healthcare provider is stored differently, making the data hard to compare, integrate, and analyze. For example, when combining medical information from a number of sources, an individual could be labeled as Male/Female, Man/Woman, or M/F, with each pair being viewed as different terms by the analytic software. The same problem exists for other life sciences terms and data. Retrospective analysis of scientific language poses significant challenges because the meaning and use of different terms in most cases cannot be resolved, even by language sorting technologies. Generally speaking, genomics is an exception to the rule compared to other Big Data sources because scientists agreed to use standard ontology for genomics data, which coincided with the early phases of the Human Genome Project.[51] This is not the case for other data that might feed into Big Data analytic tools.

In some cases in which consortia of institutions were established, the problems with sharing data were regulatory rather than technical.[52] In August 2014, the U.S. National Institutes of Health released its guidance on sharing genomic data, which includes the need for standardized language.[53]

A **lack of access to needed technologies and computational infrastructure** is a critical challenge for a number of organizations, especially start-up and smaller organizations within the life sciences. However in general, the U.S. healthcare industry does not seem to share this challenge. Specifically, for those organizations that do not have their own supercomputers or

[51] See Gene Ontology Consortium. Accessible at http://geneontology.org/page/documentation. Accessed on October 10, 2014.
[52] Consortia of hospitals and research organizations have been created to enhance scientific understanding of the genetic basis of disease and to improve healthcare. AAAS/CSTSP spoke with members of some consortia during its review of strategic drivers of Big Data in the life sciences.
[53] U.S. National Institutes of Health. 2014. NIH Genomic Data Sharing Policy. Accessible at http://gds.nih.gov/03policy2.html. Accessed on October 10, 2014.

data centers, access to analytic technology, storage hubs, and power to run more advanced computational analyses often present a significant financial and technical burden. As such, several U.S.-based organizations are seeking to access supercomputers and/or data centers outside the U.S. to reduce the costs of accessing the computational power to conduct Big Data analytics. Larger, better financed organizations are able to supply their own information technology needs to conduct Big Data analysis. In addition, computer companies engaged in Big Data are hosting versions of their data analytic technologies on the cloud, which could reduce the financial and technical burden for smaller organizations or research groups to conduct data analytics.

Important and persistent challenges for organizations involved in healthcare and medical research are **data privacy and confidentiality**. The ability to trace the genetic information to an individual even if the information has been anonymized is a significant concern. Current genomic sequencing practices may limit an individual's stored DNA to a specified number of sequences and may dictate that an individual's identification is stored separately. However, concerns are being raised about the effectiveness of these practices in protecting an individual's identify being associated with his/her genetic sequences.

Vulnerability of the data repositories and cyber infrastructure to attack presents challenges for organizations that collect large amounts of personal data or other sensitive information and for organizations that collect and analyze their own data. Recent breaches in healthcare databases demonstrate the scale of this challenge. In contrast, less mature applications of Big Data, such as those currently being developed to conduct environmental evaluations, focus more on building the technology, and data sharing and collection infrastructure than security. Evaluating approaches for safeguarding the underlying cyber and data infrastructure as they are being built addresses future concerns about security with these nascent efforts.

Conclusions

Application of Big Data has significantly increased because of the growing amount and variety of information generated, development of new technologies for collecting, storing, and accessing large amounts of data, and development of new analytic tools to integrate and analyze these large heterogeneous datasets. These applications have been harnessed by the life sciences community in a number of areas, including –omics research, drug development, and healthcare. However, during the past year, risks to individual privacy and confidentiality and vulnerability to attacks of the cyber infrastructure and data repositories have been highlighted. The White House has published two reports on protecting privacy through legal and technical means, and several efforts have begun to prevent cyber attacks. Evaluating how these efforts to address the particular national and transnational security challenges associated with Big Data in the life sciences is important and one of the focuses of the present report .

In addition to security risks presented by vulnerabilities in the data and cyber infrastructure, the risk that available data and analytic tools could be used to design harmful agents raises national and transnational security concerns. As greater amounts of data and analytic technologies are made accessible to an increasing number of people, the feasibility and risk of nation-states, non-

state/lone actors, or individuals to design harmful agents using currently available information and analytic tools will need to be evaluated and addressed.

The present report focuses on risks associated with vulnerabilities in the cyber and data infrastructure and with use of current analytic tools and datasets to design harmful agents. Chapter 3 presents three risk and benefit scenarios and qualitative risk and benefit frameworks with which to evaluate emerging or enabling technologies.

Chapter 3:

Assessing National Security Risks and Benefits of Big Data in the Life Sciences

Big Data technologies have been and continue to be developed to address a number of societal challenges, not the least of which is national security. For example, the U.S. government alone has invested billions of dollars in applying Big Data technologies to infectious disease surveillance to enable early detection of unusual outbreaks and societal verification of arms control treaties. Private sector investments (and investments made by other governments) in applying Big Data technologies in healthcare to improve patient experiences, personalized treatments, and patient outcomes contribute to the preparedness and response goals of biological security. Although Big Data technologies can provide new capabilities or improve existing capabilities to address national and transnational biological threats, they also could be manipulated, exploited, or disrupted to cause harm, whether economic, political, social, or to public health. The risks can range from inappropriate access to sensitive data such as the numerous examples of cyber attacks to healthcare databases (i.e., taking advantage of vulnerabilities in the system), to the use of Big Data analyses in designing harmful biological agents.

Use of Big Data analytics to design harmful biological agents presents a different security risk than the theft and misuse of existing Biological Select Agents and Toxins (BSAT).[54] Development and possession of BSAT agents are tightly regulated in the United States and several other countries have complementary (but not identical) regulations. However, very few, if any, statutes and regulations govern use of data, data analytics, or other computational tools to design harmful biological agents. A more detailed discussion of this is in Chapter 4. This chapter is focused on evaluating the risks and benefits of new, and advancing, capabilities in the collection, storage, integration, and analysis of large amounts and types of biological data. In

[54] U.S. Select Agent Program. Accessible at http://www.selectagents.gov/. Accessed on October 10, 2014; United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern. 2012. Accessible at http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf. Accessed on October 10, 2014; United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern. 2014. Accessible at http://www.phe.gov/s3/dualuse/Documents/durc-policy.pdf. Accessed on October 10, 2014.

developing risk and benefit scenarios and frameworks, the typical barriers to entry (both legally and technically) for any scenario or situation in which actual laboratory research is needed to create a biological agent from the computer-generated design from Big Data analysis still hold.[55]

Dealing with Data: The "Dual Use Dilemma" Redefined

A significant amount of dialogue and policy action in the United States and internationally has taken place to address the use of legitimate biological research to cause harm (the "dual use dilemma"). Much of these efforts have been focused on the actual creation of the biological agent that could cause harm. But in 2012, two scientific papers describing mutations in the hemagglutinin gene of the H5N1 influenza virus that enable the virus to transmit through the air between one infected ferret to a healthy ferret and cause illness were submitted for publication.[56] The experiments described in these papers raised a number of concerns about the risk that information (in this case the DNA sequence and actual mutations) could be used to create an H5N1 influenza virus (a highly pathogenic avian influenza virus) that could transmit easily between humans through the air. For a year, scientists, journal editors, security experts, and international public health organizations, including the U.S. National Science Advisory Board for Biosecurity, the American Society of Microbiology, and the World Health Organization, engaged in sometimes contentious discussions about what to do with the information (i.e., whether to allow it to be published as originally written or as revised), whether the information could or should be controlled by the export control regime, and if/how future research on highly pathogenic avian influenza should proceed. The papers were ultimately published in mid-2012,[57] but approaches for appropriately addressing security concerns associated with data, not actual viral samples, have not been resolved.

The security risks of Big Data in the life sciences fall into two major categories: 1) inappropriate access to data and analytic technologies through vulnerabilities in the data and cyber infrastructure; and 2) the use of Big Data technologies to integrate current data and enable the design of a harmful biological agent. The concern about data in the former category rests mostly in the theft of data or manipulation of datasets to deliberately alter the results. The concern about the latter category focuses mostly on the exploitation of data from private and publicly available data sources. Today, individuals who are able to design harmful biological agents likely have the expertise to create them in the laboratory. However, the typical legal and technical barriers to actual development and possession of certain biological agents would apply. The United States and more than 170 countries are engaged in several policy and programmatic efforts to prevent the theft or misuse of biological research to cause harm.[58] New U.S. government policies regarding "dual use research of concern" and the U.S. Select Agents and Toxins Program impose review, oversight, risk analysis and management, and security requirements for research involving specific pathogens and certain experimental procedures. The barriers established by

[55] For an example of technical barriers to entry, see Part II, Chapter 3 of K Vogel. 2012. Phantom Menace or Looming Danger?: A New Framework for Assessing Bioweapons Threats. Johns Hopkins University Press (Baltimore, MD).
[56] M Enserink. In the Eye of the Storm, Two Rivals, Two Strategies. Science. 6 Jan 2012. 335(6064): 21.
[57] S Herfst et al. Airborne Transmission of Influenza A/H5N1 Virus between Ferrets. Science 22 June 2012. 336(6088): 1534-1541; M Imai. Experimental Adaptation of an Influenza H5 NA Confers Respiratory Droplet Transmission to a Ressortant H5 HA/H1N1 Virus in Ferrets. Nature. 21 Jun 2012. 486(703): 420-428.
[58] These efforts are highly focused on specific pathogens and toxins and are insufficient for preventing the development and possession of novel agents.

these policy and regulatory efforts, in addition to the standard technical difficulties inherent in biological research and diagnostic activities, suggest that the likelihood that an adversary will succeed at creating a biological agent in the laboratory and deploying the agent is low. As mobile applications and cloud-based analytics becomes more accessible, the design of harmful biological agents could become easier for anyone (including those without the requisite technical skills and laboratory access); however, actually creating an agent (either through standard methods or chemical synthesis) would require significant skill, expertise, and access to laboratories, materials, and reagents. This evaluation of likelihood and skills needed does not diminish the national security concerns about exploitation of Big Data to evade current detection and defenses, to enhance the destructive features of pathogens or to synthesize create new pathogens.

As more information (both biological and non-biological) is collected, computational analysis tools advance, and standards for data collection and sharing are developed, the potential for Big Data technologies to enhance both national security and adversary capacities is likely to increase. Now is the time to develop tools through which scientists, policymakers, and security experts can evaluate the risks and benefits concurrently, and consider solutions to prevent or mitigate the identified risks. If these assessments are conducted locally and nationally, the likelihood that data will be used to cause harm to people, communities, nations, the environment, and science in the future could decrease.

To begin evaluating the potential risks and benefits offered by current Big Data technologies, particularly in light of the capabilities they offer and the demonstrated threats of cyber attacks, AAAS/CSTSP, the working group, and experts developed a series of technically sound risk scenarios and complementary benefit scenarios. These scenarios incorporate current knowledge about the current state of the science of Big Data in the life sciences, its use in healthcare and research, and its challenges, and about actual computer viruses and other means of attacking databases and cyber infrastructure. Three risk scenarios, described below, facilitated the development of a risk assessment framework, benefit assessment framework, and prevention and mitigation strategies (see Chapter 4).

The frameworks presented in this report are intended to be used by both scientists and security experts to identify and evaluate risks and benefits of Big Data on a periodic basis.

Risk and Benefit Scenarios

The three risk scenarios that served as the basis for developing the risk and benefit assessment frameworks were:

Scenario 1: Using biological Big Data to design a virus that targets a specific subpopulation;[59]

Scenario 2: Flooding infectious-disease surveillance systems that rely on Big Data technologies with false information; and

[59] During the Cold War, the Soviet Union made reference to ethic weapons and accused the United States of using developing ethnic weapons. See M Leitenberg, RA Zilinskas, and JH Kuhn. (2012) The Soviet Biological Weapons Program: A History. Harvard University Press. (Boston, MA); pg. 415.

Scenario 3: Exploiting Big Data technologies to subvert current pathogen detection capabilities.

The risk of each scenario is described in terms of time and likelihood. For time, the terms "near-term" is described as 0-5 years, "mid-term" is described as 6-10 years, and "long-term" is described as longer than 10 years. The likelihood is described as "plausible," "moderate," or "unlikely" and correlates with the feasibility of an adversary to carry out the risk scenario. The consequences are described as "high" or "low" and correlate with the degree to which the scenario, if successfully carried out in full, impacts human and/or animal health, economic, political, social, environmental, or agricultural sectors.

The time and likelihood for successfully designing a harmful biological agent and for successfully carrying out the full scenario (from design to creating the actual agent) are included for Scenarios 1 and 3. This is done to enable a discussion about the national security risks of using data and data analytics rather than focusing on the actual creation and possession of restricted pathogens, which often occurs when evaluating biological security risks. To continue the focus on risk of data and data analytics, the technical, financial, and logistical issues associated with deployment of a biological agent are not discussed in the scenarios below.

Complementary "benefit scenarios" were developed for each of the scenarios to enable a parallel evaluation of risks and benefits of the same data and technologies.

Box 1: Scenario 1

Risk Scenario: Targeting of Subpopulations using Big Data.

A well-resourced country wants to target a specific subpopulation within its borders by developing a virus that preferentially infects a group among its citizens. Scientists analyze data from large numbers of scientific publications and genomic information in public databases to identify a virus and associated host genes specifying preferential infection in the subpopulation. Scientists use the information to acquire the virus and modify it to increase its specificity for the subpopulation.

Corresponding Benefit Scenario: Developing medicines to prevent or reduce disease burden in certain populations.

A country wants to reduce the burden of disease in a subpopulation that is affected by a specific virus more than others in the same country or region. The country supports research, development, and testing of medicines that can prevent or treat transmission and infection of the virus in the subpopulation.

(Scenario 1) **Targeting of Subpopulations using Big Data.** This scenario draws on recent investments by several countries and companies worldwide to sequence citizens' genomes. (Box 1) The virus in this risk scenario was intended to be a virus of public health interest, but not an immediate or obvious security concern. It was *not* intended to be a U.S. Select Agent or other controlled pathogen. This scenario relies on Big Data technologies to analyze the genomic data from hundreds of thousands of people and cross-reference that data with functional data published in scientific literature.

Since the completion of the Human Genome Project in 2003, scientists have conducted research to better understand the information contained in the human genome and how that information affects the development and function of the human body.

Increasingly, scientific articles are being published on the prevalence of genes or specific versions of genes in specific national populations.[60] Scientists around the world have raised questions about differences in the genomes of people who are not Caucasian or come from similar ancestry. Several countries (e.g., Saudi Arabia[61] and the Netherlands[62]) have established initiatives to sequence and collect genomic information from their populations or ethnic groups.

Several direct-to-consumer services (e.g., 23andMe[63]), healthcare systems (e.g., Inova[64]), research institutes (e.g. Institute for Systems Biology[65]), and non-professional science projects (e.g., The Genographic Project by National Geographic[66] and Personal Genome Project at Harvard[67]) sequence individuals' genomes from blood or saliva samples. Direct-to-consumer services evaluate its customers' risk for diseases based on the information contained in the person's genome. Increasingly, healthcare systems are storing their patient's genomic information in electronic medical records. As the price of sequencing individual genomes continues to fall (in 2014, the price per sequence is $1000[68]), the number of projects to sequence people's genomes has increased; these projects can be carried out by citizen scientists in addition to professional scientists.

This scenario can be difficult to carry out in reality because it relies on the manipulation of an existing virus or synthetically-derived the virus in a laboratory. The adversary must have: 1) significant resources; 2) access to several data repositories and analytic tools; and 3) the knowledge and skills to correctly acquire, manipulate, test, and deliver the virus. Current technical, financial, and access barriers provide large hurdles for creating the modified virus based on a computer-generated model or design. However, the focus of this scenario is on the *design* of a modified virus using only data and advancing computational tools.

Using Big Data analytics and current knowledge about circulating viruses and specific host susceptibility genes, most scientists (and some non-scientists) could computationally design a virus that specifically affects certain populations. As access to public health data, scientific

---

[60] For example, see K Kedzierska. Preexisting $CD8^+$ T-cell Immunity to the H7N9 Influenza A Virus Varies Across Ethnicities. PNAS. 21 Jan 2014. 111(3): 1049-1054.
[61] H Briggs. (2013). Hundred Thousand Genomes to be Mapped in Saudi Arabia. Accessible at http://www.bbc.com/news/health-25216135. Accessed on October 8, 2014
[62] Large-scale Genome Study Catalogs Variation in Dutch Population. GenomeWeb. 30 Jun 2014. Accessible at http://www.genomeweb.com/node/1409976?utm_source=SilverpopMailing&utm_medium=email&utm_campaign=Illumina%20European%20NIPT%20Deals;%20Sequencing%20Finds%20Developmental%20Lineages;%20Dutch%20Population%20Study;%20more%20-%2006/30/2014%2011:15:00%20AM Accessed on July 1, 2014
[63] See 23andMe Corporate Website. Accessible at: https://www.23andme.com/. Accessed on October 10, 2014.
[64] See Inova Corporate Research Website. Accessible at: http://www.inova.org/itmi/our-research-studies. Accessed on October 10, 2014.
[65] See Institute for Systems Biology. Accessible at: http://www.systemsbiology.org/. Accessed on October 10, 2014.
[66] See The Genographic Project Website. Accessible at: https://genographic.nationalgeographic.com/. Accessed on October 10, 2014.
[67] See the Personal Genome Project: Harvard Website. Accessible at: http://www.personalgenomes.org/. Accessed on October 10, 2014.
[68] EC Hayden. Technology: The $1000 Genome. *Nature*. 20 Mar 2014. 507: 294-5.

publications and conference abstracts, and analytic tools increases, the likelihood that non-scientists could also design such viruses becomes plausible. As this possibility increases, so too will the economic and political cost of preventing and responding to an incident involving a modified, targeted agent.

The ability to design modified or novel biological agents using Big Data technologies is currently available. As analytic technologies improve and become available to a larger array of people, raw data are collected, and research results are generated, the likelihood is high that a greater number of non-experts will be able to design and modify computer models of virus and other biological agents. While the **risk that an adversary could design a modified virus on the computer is plausible within the next 5 years**, the **full risk scenario presents a *long-term, unlikely risk*** because it requires access to the original virus and genetic sequences, and laboratory work to create and/or modify the virus. The scientific expertise and skills needed to create the virus in a laboratory are specialized and unlikely to be found in non-scientists and scientists who are trained in virological methods. However, future advances in synthetic genomics and biology might reduce the need for specialized skills.

If an adversary had the resources and capabilities to not only design but also successfully create or manipulate a virus, the **human health and political *consequences could be high***. The immediate consequences are: 1) death or illness in the subpopulation; 2) outrage from the country's larger population and the international community; and 3) accusations of international treaty violations. The longer-term consequences could be unrest in the country's larger population and the surviving subpopulation, international sanctions or other punitive measures, or other international actions to reduce the threat and prevent future incidents.

Beyond the design of specific viruses, this scenario highlights the ***near-term risk* that sensitive, patient information could be exploited to support other activities with national security and criminal concerns** such as the design of potent, custom narcotics based on specific genetic information about dopamine receptors or subversion of patient, clinical trial, and/or drug study data in furtherance of pharmaceutical espionage. Exploitation of patient data for these types of activities is not covered by existing legal frameworks such as the Genetic Information Nondiscrimination Act.

Box 2: Scenario 2

Risk Scenario: Misdirection.

A group outside the U.S. wants to prevent government and international health officials from discovering its covert scientific activities to modify a biological agent it acquired from a local laboratory. The domesticated animals around its "laboratory" facilities were exposed accidentally to the agent. Shortly after this exposure event, a small number of individuals were diagnosed with a respiratory infection caused by an agent not identified with local diagnostic tools. The local doctors report these infections to public health officials, other clinicians, and scientists using an infectious disease surveillance tool that relies on Big Data. Concerned that these reports would alert international and national authorities to the group's activities, one member of the group submits false reports into the system (e.g., using location of infections, severity of infection) to mask identification of the unusual infections and reduce the likelihood that its laboratory would be identified as the source of the outbreak.

Corresponding Benefit Scenario: Early identification of infections of public health concern.

Several towns in rural areas report a spike in respiratory disease not caused by circulating strains of influenza virus. Passively picking up these reports and integrating them with data collected from a wide variety of sources, an infectious disease surveillance tool that relies on Big Data (e.g., HealthMap) signals the presence of unusual respiratory infections. Scientists use the tool to identify the scale of the infection, the presence of other similarly described infections throughout the world, pharmaceutical product sales, animal and/or crop infections, and source location. The results of this analysis point to a single source – a small house with various farm animals co-mingling (though not necessarily a farm) – and a means of spreading the pathogen – the owner of the house (who is identified through public records) selling raw meat from his/her animal flock in farmers markets in the towns from which the illness reports came.

(Scenario 2) **Misdirection.** This risk scenario draws on the recent investments and application of Big Data technologies to conduct infectious disease surveillance and real examples of attacks to databases and the cyber infrastructure. (Box 2) In this scenario, the attack is primarily cyber-based, but the consequences are seen in the life-sciences and public-health sectors. Two approaches for carrying out this scenario were discussed: making changes to specific data entries within the database itself and flooding the database with false information. The first approach (making changes to specific data entries) is extremely difficult to achieve because databases and cyber infrastructure contain numerous barriers that an adversary must penetrate. The second approach (flooding the database with false information) is much easier to achieve and therefore was included in the scenario. This scenario is *not* about the covert development of a biological agent; it *is* focused on the masking of the covert activities by misdirecting infectious disease surveillance systems through false information.

Several infectious disease surveillance tools have been developed to collect official and unofficial information about current disease outbreaks, collate information among different data streams, and alert public health and/or security officials about infectious disease outbreaks of concern. Examples include Argus, HealthMap, the World Health Organization's Global Outbreak and Alert Response Network, the U.S. National Biosurveillance Integration Center

(NBIC), ESSENCE, BioSense, and Google Flu Trends. The data being fed into these systems comes from a variety of sources (e.g., official reporting, scientific entries, and news reporting) and some systems incorporate crowdsourcing to evaluate the quality and utility of the input data. Some systems, such as the NBIC, simultaneously integrate the data from different data sources.

To carry out this scenario, the adversary must have access to the reporting system, which can be through direct authorities or through hacking of the system. This scenario does not depend on the actor having sufficient knowledge and expertise about modifying biological materials. The adversary for this scenario could be individuals, small groups of individuals, large non-state/lone groups, or nation states. The technical barrier is the access to and manipulation of biological agents, not the submission of false information into an infectious disease surveillance database(s).

Because hacking is a persistent cyber and data security problem and the skill level needed to carry out this scenario is low, **it presents a *near-term, plausible risk.*** In addition to the inherent vulnerabilities to the flooding of information or other cyber-based attacks, the likelihood of the risk is affected by the lack of broadly accepted methods for determining and confirming the source of the submitted data.[69]

The immediate consequences of this scenario would be to divert notice from the unusual infections and evade detection by public health and security personnel. However, the initial infections could result in the uncontrolled spread of the biological agent in the animal and human populations, which would be reported to health authorities through other public health departments. The recent Ebola virus epidemic in West Africa demonstrates that any delay in the identification and control of an infection could result in significant public health, security, and societal burdens that could involve several countries. The national and international human health, economic, and political consequences of this scenario are ***high*** because of: 1) misdirection of an unusual infection, which would result in the group's identity remaining hidden and their intentions undiscovered; and 2) eventual realization that the outbreak was man-made and the reporting system was not protected from a cyber attack.

[69] Determining and confirming the source of data, especially data from social media or the internet, is an important concern to those organizations and individuals seeking to use Big Data analytics (including social media analysis) to identify potential outbreaks or events of concern. In 2013, the Human Rights Watch confirmed the efficacy of YouTube video from the 2013 Ghouta chemical weapons attack in Syria before analyzing the videos. See Human Rights Watch. 2013. Attacks on Ghouta: Analysis of Alleged Use of Chemical Weapons in Syria. Accessible at: http://www.hrw.org/reports/2013/09/10/attacks-ghouta-0. Accessed on October 10, 2014.

Box 3: Scenario 3

Risk Scenario: Avoiding Detection.

A non-state/lone actor uses new advances in computational analyses of multiple types of scientific data to determine which genetic segments are found only in dangerous strains (not the harmless variants) of a specific pathogen. Through publically accessible information, the adversary infers that these segments will be used as the target of molecular tests for pathogen detection (e.g., Polymerase Chain Reaction (PCR)). Once these segments are identified, the adversary makes specific changes to the gene that change only the genetic sequence but not the subsequent protein sequence to prevent loss of function. This would allow the virus to be effective, but undetectable.

Complementary Benefit Scenario: Identification of gene targets and mutations for pathogen detection.

A scientific group consisting of academic and industry partners uses advances in machine learning and data integration of published scientific literature, relevant microbial genomes, and other useful information to identify specific genes for which molecular diagnostic tests are developed. This research is funded to improve research and development of diagnostic tests and to supply more accurate tests to the public health departments and forensic laboratories.

(Scenario 3) **Avoiding Detection.** This scenario draws on the process through which molecular diagnostic tests and platform technologies are developed.[70] (Box 3) This scenario uses Big Data analyses to evaluate microbial genomes, sequence changes that prevent loss of function, and functions of microbial proteins. The United States has regulations that prevent unauthorized access to dangerous pathogens, specifically those pathogens that cause risk to public health safety and security. Access to these pathogens is extremely difficult in the United States and its allies because of the regulatory environments restricting access to certain biological agents and the security at high-containment facilities housing these agents. The adversary could synthesize the pathogen with information about its genetic sequence. However, synthesizing most viruses is technically difficult and exceedingly expensive despite the low cost of gene synthesis.[71] Identifying and incorporating specific mutations in specific viral genes that would allow it to go undetected by current molecular tests depends on the size and complexity of the viral genome. Larger, more complex viruses are much more difficult to manipulate and maintain their ability to spread, infect, and cause disease. Smaller, less complex viruses are easier to manipulate. In 2002, Eckard Wimmer's research group created polio virus (a small virus) from scratch; however, the infectious dose of the synthetic virus was less than the natural virus.[72] As synthetic genomics and synthetic biology techniques improve, so too might the ability to create synthetic viruses that incorporate specific mutations of interest without any adverse effects on virus transmission, infection, and function.

[70] Current molecular diagnostic techniques for testing virus are published in F Cobo. Application of Molecular Diagnostic Techniques for Viral Testing. Open Virol J. 2012. 6: 104-114.
[71] The feasibility of synthesizing viruses of different sizes is described in RS Baric. 2006. Synthetic Viral Genomes. In: Working Papers for Synthetic Genomics: Risks and Benefits for Science and Society, pp. 35-81. MS Garfinkel, D Endy, GL Epstein, RM Friedman, eds. 2007 and S Mueller, JR Coleman, and E Wimmer. Putting Synthesis into Biology – A Viral View of Genetic Engineering Through de novo Gene and Genome Synthesis. Chem Biol. 27 Mar 2009. 16(3): 337-347.
[72] J Cell, AC Paul, E Wimmer. Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template. Science. 9 Aug 2002. 297(5583): 1016-8.

Several databases that store genetic information from different organisms exist. The most widely used database is the National Center for Biotechnology Information's GenBank. This website allows people to submit a genetic or protein sequence and look for similar or identical sequences already in the database. Examples of specific databases include the Poxvirus Bioinformatics Resource Center, which allows individuals to identify and analyze genetic sequences of poxviruses, dengue, and hepatitis C virus; the Viral Bioinformatics Resource Center, which contains a database of genetic information about viruses and analytic tools to examine viral genomes; and Pathogen Portal, which allows people to identify and analyze pathogen data.

Increasingly, scientists are using crowdsourcing and the cloud to share genetic information about pathogens causing outbreaks. One example is the E. coli 0104:H4 outbreak in 2011 in which the Chinese company BGI sequenced the strains of bacteria causing disease in Germany and posted the genetic sequences on the cloud for scientists to evaluate.[73]

A knowledgeable and/or educated scientist could evaluate the scientific data to identify the genetic segments of interest and computationally design the modified genetic segments. However, citizen scientists, non-life scientists, and engineers are able to identify genetic sequences and model mutations in those sequences using publicly available bioinformatics and computational modeling tools. The likelihood that an adversary (whether an individual, non-state/lone group, or nation-state) can identify and **design genetic segments that could evade detection is a *near-term, plausible risk*** because the technical expertise needed to evaluate genetic sequences exists in amateur, academic, industry, and government laboratories, data about genetic sequences is accessible publicly, and the needed computational tools are becoming increasingly more accessible. However, the **full scenario presents a *mid-term, moderately likely*** risk because of the scientific skill and financial means involved and the limited access to restricted agents and/or acquisition of DNA sequences of restricted agents.

However, the **overall risk of carrying out this scenario is *moderate to high*** in the United States because the U.S. government regulates certain dangerous pathogens (i.e., the Select Agents and Toxins Regulations) and has guidance for gene-synthesis companies to report unauthorized requests for products regulated by the Select Agents and Toxins Regulations.[74]

If the restricted pathogen is designed and created, the public health and national security consequences could be ***high*** because existing detection systems, medical diagnostics and treatments would be faced with a new disease, and existing antibiotics and treatments may not be effective. The successful design and creation of an undetectable version of a restricted pathogen could result in uncontrollable spread of a harmful pathogen and could strain national and international health systems.

[73] BGI. The Complete Map of the Germany E. coli 0104 Genome Released. EurekAlert. 16 Jun 2011. Available at http://www.eurekalert.org/pub_releases/2011-06/bgi-tcm061611.php. Accessed on July 1, 2014.
[74] U.S. Screening Framework Guidance for Providers of Synthetic Double-Stranded RNA. 2010. Accessible at: https://www.phe.gov/Preparedness/legal/guidance/syndna/Documents/syndna-guidance.pdf. Accessed on October 10. 2014.

Assessment Frameworks

The risks and benefits of emerging and enabling technologies, such as Big Data in the life sciences, often are assessed separately, involve knowledge of more than one scientific discipline, and are based on little, if any, quantifiable information. The separate assessments could result an imbalance of either the risks or the benefits rather than an objective evaluation of the national security implications of the technologies. The multidisciplinary and qualitative assessment of the risks and benefits adds a layer of complexity to evaluating emerging or enabling technologies. However, the rapid pace with which technologies develop requires a more balanced, objective, multidisciplinary evaluation of these technologies to maximize the beneficial applications to scientific pursuits and societal needs, and simultaneously, to minimize the vulnerabilities and security risks of emerging and enabling technologies.

Assessment of the risks and benefits of the application of Big Data to the life sciences must be based on an understanding of the capabilities and vulnerabilities of the data repositories and analytic technologies, and the expertise needed from all relevant disciplines involved. (Big Data in the life sciences is a multidisciplinary field, involving computational science, engineering, and mathematics that is applied to the life sciences, environmental, and healthcare.) In this regard, assessing the risks and benefits of Big Data in the life sciences can be used as a surrogate for an analysis of other developments that combine different science and engineering disciplines to advance scientific knowledge and improve society.

The key is to develop a process that encourages the assessment of realistic risks and benefits on an ongoing basis. Periodic assessments increase the likelihood that new benefits or capabilities of biological Big Data that could be applied to address national security concerns are identified. Stated differently, if scientists and security experts can assess the risks and benefits together and on a routine basis, the potential risk that biological Big Data would be used by adversaries decreases.

The frameworks below would enable scientists and security experts to work together to assess the risks and benefits of Big Data in the life sciences. These frameworks can be modified to focus on specific outcomes to help policy-makers in authorizing, implementing, or investing in specific programs or developing statutes, regulations, policies, and/or guidance to prevent or mitigate risks.

Risk Assessment Framework

Risk assessments of emerging or enabling technologies are difficult. One reason for this is that much of the needed information is not yet available to individuals conducting the assessments, because the information is either unknown or cannot be known.

In national security, risk is assessed as a product of the *probability of an attack* (based on information about the adversary and the adversary's capabilities and access, the vulnerability of the system, and the skills or expertise needed to carry out the attack) and the *consequences of the attack* (based on the consequences if the adversary is successful and on the prevention or mitigation of consequences expected from available countermeasures).[75] Unlike a *threat*

[75] The risk equation used in this report is also depicted as a product of probability and impact, and can be quantitative or qualitative depending on the information on which the assessments are made. Risk in this context does not refer to laboratory

assessment in which adversary intent and capability to carry out an attack are known (i.e., **threat = intent x capability**), adversary intent is rarely accounted for and adversary capabilities are estimated broadly in typical calculations of risk assessments in the life sciences. For the purposes of this report, which seeks to address security directly, risk is defined as:

Risk = *Probability* (adversary type, capabilities, and access x vulnerability x skill or expertise needed to carry out attack) x *consequences* (consequences if successful attack x countermeasures)

Risk assessments normally are visualized on a graph with one axis depicting the likelihood of an attack and the other axis depicting consequences.

Scientists could be required to assess the risk that a given technology or specific scientific finding could be used for harmful purposes. In general, scientists would not know the adversarial actor(s)' capabilities and access to resources, but they would know about the skills, equipment, methods, and other scientific information and expertise that are needed to carry out harmful activities. Similarly, security and intelligence experts tasked with assessing the same risk will not be familiar with the scientific knowledge, skills, and expertise needed to cause harm using biological information or materials, but they will have access to information about adversaries, including their access to materials and information, and their relative capabilities. Risk assessments for mature, deployed technologies can incorporate probabilistic assessments of specific outcomes that are derived from historical, quantitative data. In contrast, risk assessments for emerging and/or transforming technologies must be based on educated extrapolations (best guesses) about the potential adversaries, needed skills and knowledge, vulnerabilities, and potential consequences.

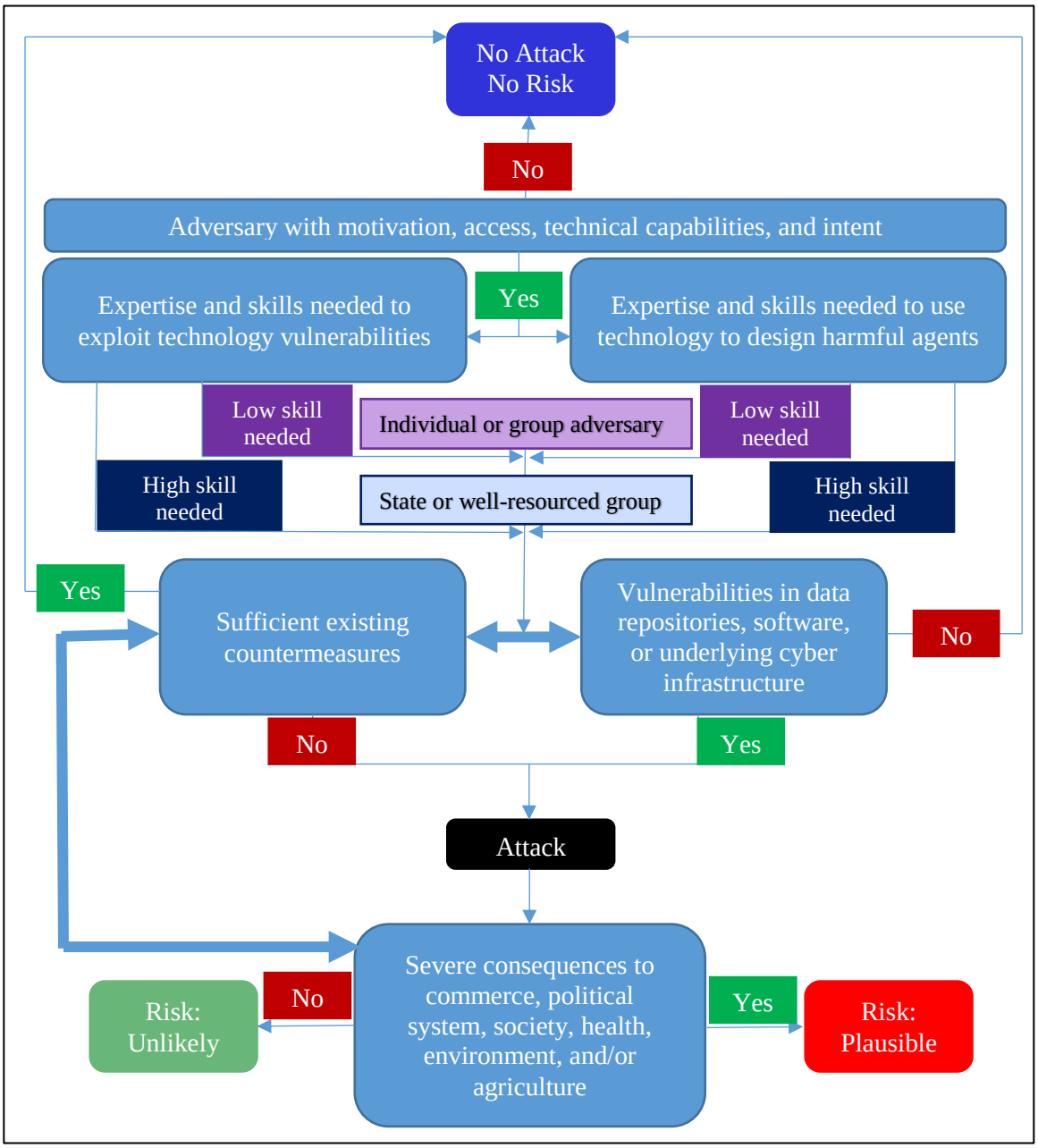
Although a significant amount of information exists about cyber attacks on data repositories, software, and methods through which data are shared, very little, if any, information exists about the theft, manipulation, or exploitation of Big Data in the life sciences. Consequently, the process through which scientists and security experts would evaluate the national and transnational biological security risks of Big Data in the life sciences takes the statistics of cyber attacks as information upon which to base adversary capability and access, and vulnerabilities in the system. The end result is a qualitative assessment of the risk that informs the types of solutions – legal, technical, institutional, and individual – that could prevent or mitigate such risks. Working together, scientists and security experts will have the most information to input into the assessment process.

---

biosafety and biosecurity risks; instead, it refers to the likelihood that a devastating attack will occur. Quantitative assessments use numbers to depict the likelihood and impact of the risk whereas qualitative assessments use words. Several references of relevance to this types of risk assessment: NIST. 2012. Guide for Conducting Risk Assessments. Accessible at http://csrc.nist.gov/publications/nistpubs/800-30-rev1/sp800-30_r1.pdf. Accessed on October 11, 2014; J Moteff. (2005). Risk Management and Critical Infrastructure Protection: Assessing, Integrating, and Managing Threats, Vulnerabilities and Consequences. Congressional Research Service. RL32561; BC Ezell et al. Probabilistic Risk Analysis and Terrorism Risk. Risk Analysis. 2010. 30(4): 575-589; MITRE. 2014. Systems Engineering Guide. Accessible at http://www.mitre.org/sites/default/files/publications/se-guide-book-interactive.pdf. Accessed on October 11, 2014; ISO/IEC 31000:2009. Accessible at http://iso.org/iso/home/news_index/news_archive/news.htm?Refid=Ref1586.

Figure 3.1 presents the conceptual framework for qualitative risk assessment for emerging and enabling technologies, including Big Data in the life sciences. The table following the figure lists specific questions that should be asked when conducting the qualitative risk assessment.

Figure 3.1 Conceptual Framework for Qualitative Risk Assessment of Emerging and Enabling Technologies



		Risk Assessment Framework Questions
Probability of an Event	Characterization of the Adversary	• Which type(s) of adversary – lone actor, non-state/lone group, or nation-state – has the resources, scientific and technical expertise and skills, access to facilities and equipment, and motivation to attack or exploit Big Data technologies and/or life sciences and healthcare data repositories? • Have existing adversaries expressed interest in stealing data, manipulating datasets, or exploiting Big Data in the life sciences? • If the adversary is a non-state/lone group or nation-state, does its doctrine, internal organizational structure, and/or relationship to individual factions prevent dissident actions to attack or exploit Big Data in the life sciences despite differences in the group's intent? • Does the adversary have access to a broader network of resources, including individuals, funds, equipment, etc? • Does the adversary have the capabilities needed to access, manipulate and/or exploit Big Data technologies and life science datasets? • Does the adversary have direct access to data or software? If not, can the adversary gain access through hacking or elicitation? • Does the adversary have the financial resources to attack or exploit Big Data technologies, infrastructure, and life sciences datasets? • Would the adversary weigh the cost and benefit of achieving the attack from their perspective?
	Vulnerabilities in the System	• How accessible are the data and analytic technologies? • What technical measures exist to control access to and track users of analytic software, data repositories, and underlying infrastructure? What is unprotected? • What institutional measures exist to limit access to and identify users of the data systems and analytic technologies? • Are system vulnerabilities regularly tested?
	Needed Scientific Expertise and Skills	• What skills, expertise, and knowledge are needed to use data and analytic results, data and analytic technologies to design a harmful biological agent? • What technologies and equipment are needed to use data and analytic results, data and analytic technologies to design a harmful biological agent? • What skills, expertise, and knowledge are needed to exploit the vulnerabilities in the system? • Is specialized access to technologies needed to exploit vulnerabilities in the system?
Consequences	Outcomes of Successful Attack	• What are the possible immediate consequences of the attack to the economy, political systems, society, health, agriculture, environment and/or national security? • What are the possible long-term consequences to the economy, political systems, society, health, agriculture, environment and/or national security? • What is the scale of the immediate and longer-term consequences? • Do systems exist to maintain critical operations and recover quickly?
	Existing Counter-measures	• What countermeasures exist to address the vulnerabilities? • What countermeasures exist to prevent or mitigate the consequences of an attack? • What measures exist to prevent or control exploitation of scientific data and Big Data technologies?

Benefit Assessment Framework

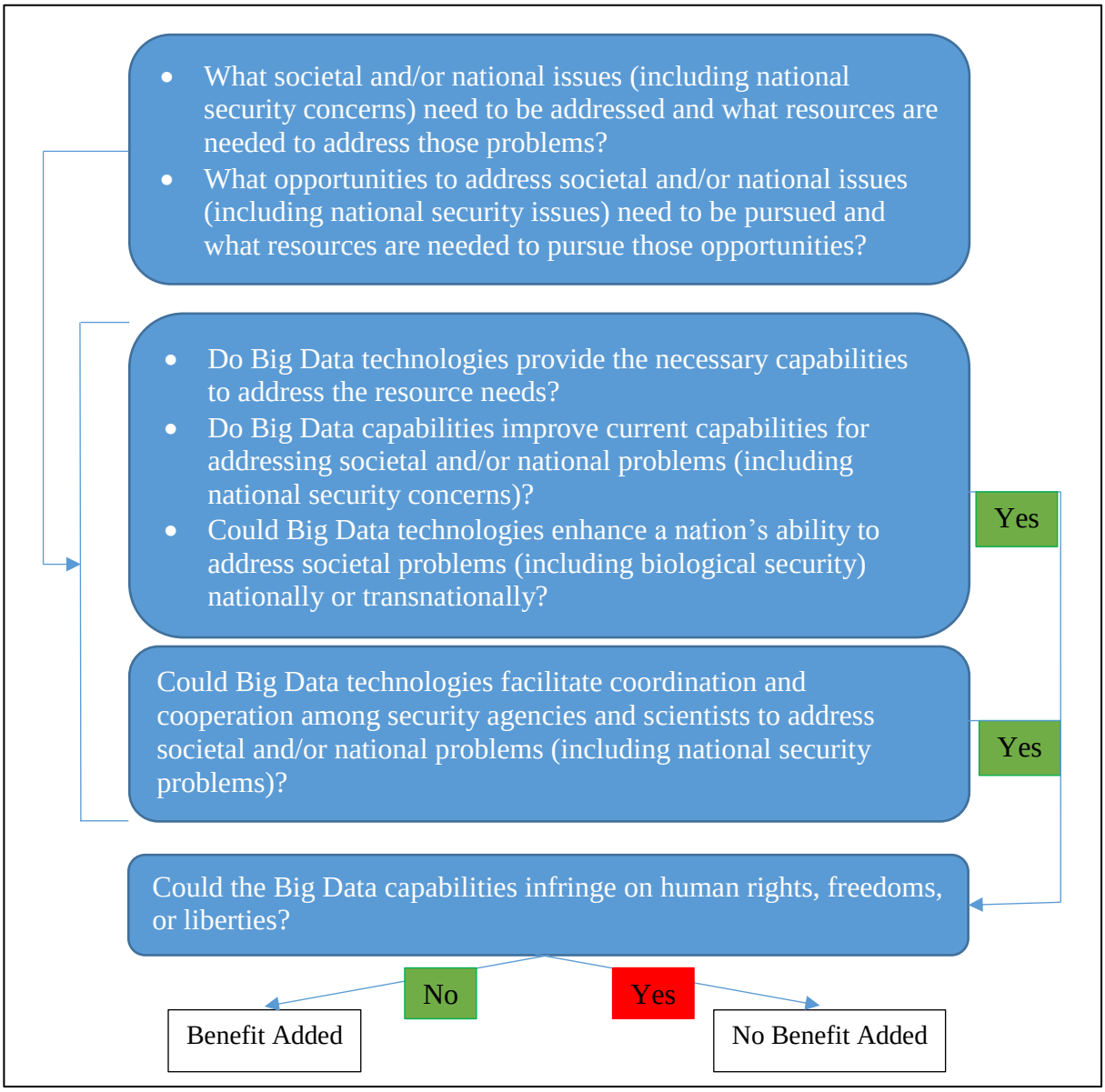
While no formal process exists by which emerging and transforming technologies are evaluated for their benefits to different sectors, informal evaluation of the capabilities and broader benefits reaped by emerging technologies are conducted. Scientists are expected to consider and predict the broader significance of their research efforts in funding proposals to the National Institutes of Health and the National Science Foundation. Program managers at the Defense Advanced Research Projects Agency and the Intelligence Advanced Research Projects Activity develop funding portfolios that support research and development of projects that address defense or intelligence needs.

A formalized process through which benefits can be assessed involves identification of the national security issues and the resources needed to address or inform the problem, evaluation of the capabilities that the technology requires, and an evaluation of the broader societal impact of those technologies if deployed. The significance of this last criterion was highlighted in 2013 when the National Security Agency's surveillance activities were revealed. This discovery resulted in the White House calling for two high-level reviews, which resulted in the reports, *Big Data: Seizing Opportunities, Preserving Values*[76] and *Big Data and Privacy: A Technological Perspective*.[77]

Figure 3.2 presents the benefit assessment framework for Big Data in the life sciences. This framework can be applied to the evaluation of other emerging and enabling technologies.

[76] Executive Office of the President. Big Data: Seizing Opportunities, Preserving Values. May 2014.
[77] President's Council of Advisors on Science and Technology. Big Data and Privacy: A Technological Perspective. May 2014.

Figure 3.2 Benefits Assessment Framework for Big Data in the Life Sciences Technologies



Comparing Risks and Benefits

Weighing the potential risks and potential benefits to determine whether the risks outweigh the benefits or whether the benefits can be used to reduce the risks is probably the most challenging aspect to evaluating the national and transnational implications of Big Data in the life sciences. No formal process for this exists, but concurrently evaluating the risks and benefits using the frameworks provided above will enable a more comprehensive assessment of the implications (both positive and negative) of Big Data in the life sciences to national and transnational security. Enhancing these efforts is the cooperation between security experts tasked with monitoring science and technology developments, scientists with experience working with the technology, and scientists from a range of disciplines affiliated with the technology. This

approach will take time and effort, but it will help to reap the benefits of emerging technologies while preventing or minimizing their potential risks throughout the technology development, maturation, and application processes.

Applying the Risk and Benefit Frameworks to the Three Scenarios

To help scientists and security experts use the risk and benefit frameworks developed in this project, AAAS/CSTSP, FBI/WMDD/BCU, UNICRI applied the frameworks to the three risk and benefit scenarios described earlier in this Chapter. (See the table below.)

				Risk Scenarios		
				Targeting Subpopulations Using Big Data	**Misdirection**	**Avoiding Detection**
Risk Assessment	Probability	Adversary		Well-resourced group or nation state, maybe insider: Technologically advanced, significant resources, access to software and data	Computer literate individual, group, nation-states	Technically advanced, access to software and data, and financial resources
		Vulnerabilities in data repositories, software, and/or underlying cyber infrastructure		Could use open access data and analytic software	Open access reporting mechanism and database	Open access data, some analytic open access software
		Needed scientific expertise and skills	To exploit vulnerabilities in the system	N/A	Access to reporting system, no specialized skills needed	N/A
			To use Big Data analytics to design harmful biological agents	Specialized skills needed: microbial genomics, molecular biology, bioinformatics	N/A	Specialized skills needed: bioinformatics, molecular biology, microbial genetics
	Consequences	Severe consequences to economics, political system, society, health, environment, and/or agriculture		Consequences are extremely high; death or illness, outrage by local and international communities, questions about treaty compliance	Consequences are mid-to-high: human health, agriculture, environment: depending on the spread of pathogen and delay in identifying infections	Consequences are high: undetected infections with dangerous pathogen
		Sufficient existing countermeasures		No. Technology access controls, institutional measures, individual	No. Standard tracking of IP addresses	No. Standard tracking of IP addresses, some access controls,

			measures		institutional measures if insider
		Risk of Computer-aided Design	Plausible to moderate	N/A	Plausible
		Risk of Success of Full Scenario	Unlikely	Plausible in the near-term	Moderate to unlikely
			Complementary Benefit Scenarios		
			Drug Development	**Infectious Disease Surveillance**	**Pathogen Detection**
Benefit Assessment	What societal and/or national issues (including national security concerns) need to be addressed and what resources are needed to address those problems?		Development of drugs, including medical countermeasures	Need for early warning of unusual outbreaks and infections	Detection of harmful pathogens and biosurveillance
	What opportunities to address societal and/or national issues (including national security issues) need to be pursued and what resources are needed to pursue those opportunities?		Identification of specific drug targets and targeted development of drugs	Characterization of infectious disease outbreaks including their sources	Characterization of microbes in the human microbiome and the environment
	Do Big Data technologies provide the necessary capabilities to address the resource needs?		Yes, has the potential to reduce cost and speed up identification of drug targets	Yes, as computational tools improve and data can be shared and compared.	Possibly, if data can be collected, shared, integrated, and analyzed
	Do Big Data capabilities improve current capabilities for addressing societal and/or national problems (including national security concerns)?		Yes, because current measures are laboratory-based and involve repeated testing of compounds	Possibly, with advances of data integration, data mining tools. New infectious disease tools are currently being developed.	Yes, if data from different sources can be integrated and analyzed in large-scale
	Could Big Data technologies enhance a nation's ability to address societal problems (including biological security) nationally or transnationally?		Yes to facilitate rapid identification of drug targets	Yes, for International Health Regulations (2005) and the Global Health Security Agenda (2014)	No
	Could Big Data technologies facilitate coordination and cooperation among security agencies and scientists to address societal and/or national problems (including national security problems)?		Possibly, if data can be shared and integrated	Possibly, if data and disease surveillance systems are integrated	Not necessarily unless data can be shared
	Could the Big Data capabilities infringe on human rights, freedoms, or liberties?		No if appropriate approvals have been sought and/or publicly available	No, if non-public information about people is	No

		data are used	not included	
	Benefit	Plausible, mid-to-long term	Plausible, near-to-mid term	Plausible, near-term

Applying the risk and benefit assessment frameworks to actual scenarios might present difficulties for different sectors. In general, scientists do not have information about adversary intent or motivations, and security experts do not have knowledge about the technical feasibility of carrying out a particular attack or exploitation scenario. However, through cooperation and information-sharing about particular technologies, applications, and security concerns, scientists and security experts have a higher likelihood of anticipating risks realistically. The results of these assessments could enable policy-makers, institutions, and scientists to identify existing or develop new solutions to mitigate or prevent the risks and reap the benefits.

Chapter 4:

Solutions to Addressing National Security Risks

A major focus of current public discussion and policy efforts on the risks posed by Big Data in the life sciences is on privacy and ethics. The privacy issues, which coincide with cyber attacks on repositories of healthcare data, have focused more on theft or inappropriate disclosure of patient or research-participant information. The ethical issues have focused on responsibilities to return medically significant data to participants, secondary information learned from genomic testing, and uses for which prior institutional approval and/or patient/participant consent has not been given. These risks primarily affect individuals and legal, technological, and institutional solutions are being explored to prevent such breaches of privacy and ethical principles.

The risks with which this study is concerned are those that are meant to intentionally cause harm, directly or as a further consequence, to a state, ally, and/or its population. These harms might be economic, political, societal, environmental, and/or health-related. At this macro and system-level, solutions for preventing and mitigating the theft, manipulation, and exploitation of Big Data in the life sciences are very different than the solutions used to prevent harm to individuals. The legal, technical, institutional, and individual solutions used to prevent or mitigate national and transnational biological security risks draw on the established measures already in place or currently in development in biological security, cyber security, and data security. *Applying these solutions differentially to address specific risks is critical because the risks and outcomes differ widely*. Some risks exploit vulnerabilities in the cyber infrastructure, data repositories, data-sharing platforms and/or software while other risks emerge when Big Data technologies are used to design dangerous biological agents. Current approaches, if tailored appropriately, can be used to help address these risks.

The following sections describe specific solutions that could be used to help address different types of risk.

Technical Solutions

The primary engineering solutions used to prevent unauthorized access to data and computer systems are access controls and data encryption technologies. Access controls, such as password protection on data repositories and proprietary software, limit who can access the data or software and what they can access. Two general types of controls exist: identification of the user, which includes identity verification and password or access code-protection; and authentication of the user systems, which includes blocking specific IP addresses or access protocols. Data and computer systems regularly run internal systems vulnerabilities tests and use internal user testing. These internal systems testers, termed "white hat hackers," can identify weaknesses in a database security system, making sure that backdoors for unauthorized access are closed and systems do not have exploitable code. Limiting which workstations have access to the databases can control multiple points of entry and make access to a workstation difficult for outside users. Private access technologies that incorporate both security mediation and access controls provide options for sharing of sensitive data, such as healthcare information, and incorporate relevant laws, institutional policies, and user preferences. Data encryption technologies prevent unauthorized actors from viewing, reading, and/or using data even if they possess the data. Encryption technologies are applied to data in transit, at rest, and in use.

Other technological controls include digital certificates, which require authorized users to verify their computer systems before access can be granted; data-masking techniques, which hide the data; security information and event management technologies, which monitor the use of the data; backup, archiving, and deletion, which ensures that duplicate copies of important data are stored to enable recovery after an attack occurs; network separation, which segregates different networks (e.g., separation of a cloud-based network and a private network); and patches, which is a solution to an identified software vulnerability. Data fingerprinting, tracking data changes by monitoring workstation logs, flagging of suspicious activity, and locking down systems when unauthorized access has occurred are measures used to strengthen the database security. For systems with highly sensitive information, internal monitoring methods, such as keylogging or random screenshots, can discourage and possibly prevent internally-released data leaks. Keeping information stored on a server, rather than a cloud, may further reduce data vulnerability. Backing up data onto multiple servers prevents an adversary from destroying the data at one source. However, the security of data stored on the cloud can be manipulated with a single hacking event.

The computing capabilities needed to support many Big Data analyses for life sciences applications are only available to some larger healthcare, pharmaceutical, biotechnology companies, and/or other larger organizations. The vast majority of life sciences organizations do not have the capacity to conduct analyses in-house and must use tools provided by other sources. These tools might be provided by private entities, which allow access to their software and data to those who subscribe, license, or purchase their product. Others provide analytic tools and data on online, open-access sites. Increasingly, data are stored and shared, and analytic tools are available on the cloud and on mobile devices. Access controls are often found on life sciences databases, such as the yeast genome database and healthcare databases, and analytic tools, such as proprietary software. Tracking technologies often are used to monitor access to online, open-access analytic tools, such as the tools available on the U.S. government's National Center for Biotechnology Information, and open-access databases, such as PubMed. Cyber and data-

security technologies are largely ineffective at securing the cloud, a reality that has driven new research on cloud security, and at identifying the source of suspicious activity if the perpetrator is using multiple databases, computers, and/or open source analytic technologies. Therefore, new approaches to security needs to be applied to Big Data and open source information that integrates security mitigation measures preemptively and in a manner that that it does not inhibit innovation and development.

Legal Solutions[78]

The legal solutions described include laws (both statutes and regulations), policies, and strategies of the United States and the international legal frameworks to which the United States is a party. No specific international treaties to which the U.S. is a party exist to address attacks on cyber infrastructure or data repositories.[79]

International treaties, such as the Biological and Toxin Weapons Convention[80] and UN Security Council Resolution (UNSCR) 1540[81] provide the international legal framework under which State signatories and their citizens cannot develop, produce, or stockpile biological weapons.[82] This overarching framework holds countries responsible for not developing their own biological weapons programs and for preventing and/or prosecuting the development of biological agents for use as weapons. In addition, the UN Security Council Resolution 1540 condemns states that do not take "appropriate actions" to address the threat of non-state/lone adversaries which intend to use and/or have used biological weapons and which are acting within their borders.

In the United States, a number of statutes, regulations, Presidential Directives, National Strategies, and Executive Orders exist specifically to prevent access to and/or the experimentation of certain biological pathogens and toxins (i.e., Biological Select Agents and Toxins). Examples include the Biological Weapons Anti-Terrorism Act of 1989, which implements the BWC and makes the development, production, and/or stockpiling of biological weapons illegal; the Anti-Terrorism and Effective Death Penalty Act of 1996, which establishes the Select Agents Program; the USA PATRIOT Act of 2001, which specifies persons to whom access to Biological Select Agents and Toxins is not granted; the Public Health Security and Bioterrorism Preparedness and Response Act of 2002, which significantly enhances the Select Agent Program; and the National Strategy for Countering Biological Threats in 2009. Many more statutes, regulations, Presidential Directives, National Strategies, guidance, and policies

[78] Appendix 6 provides a list of relevant laws, regulations, guidance, and statues.
[79] Two international efforts do exist to address cyber security threats. The Council of Europe Convention on Cybercrime applies to European countries and other state signatories (Accessible at http://conventions.coe.int/Treaty/en/Treaties/html/185.htm. Accessed on October 11, 2014). The ITU Global Cybersecurity Agenda promotes international Collaboration to address cyber security (Accessible at http://www.itu.int/en/action/cybersecurity/Pages/gca.aspx. Accessed on October 11, 2014).
[80] United Nations Biological Weapons Convention website. Accessible at www.un.org/disarmament/WMD/bio. Accessed on October 2, 2014.
[81] UNSCR1540 provides an international framework under which all UN members must address the national and subnational actor threat involving chemical, biological, radiological, and nuclear weapons. Accessible at: http://www.un.org/en/ga/search/view_doc.asp?symbol=S/RES/1540%20(2004). Accessed on September 5, 2014.
[82] The United Nations has databases of national biological-security-related legislation, some of which include language implementing the Biological and Toxin Weapons Convention and UNSCR1540. These databases are accessible at: http://www.un.org/en/sc/1540/national-implementation/legislative-database/list-of-legislative-documents.shtml; http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/BBCCCC514AA386A3C1257355003AA13D/$file/BWC_NID_Report-070912.htm.

promote the building of the nation's capacity to anticipate and address biological risks, and stimulate preparedness, response, and recovery to biological incidents.

In addition to this legal framework governing biological threats, the Federal Bureau of Investigation's Weapons of Mass Destruction Directorate has established a robust and active academic and industry outreach program to promote trust-building between the scientific and law enforcement communities, assist in preventing and mitigating actual incidents, and increase awareness of biological security issues relevant to the scientific community. Although the Biological Weapons Anti-Terrorism Act is inclusive of any pathogen or biological agent developed and used as a weapon, the Select Agent Program specifically applies to restricted pathogens and toxins. However, these laws or strategies generally do not cover data, although some do address genetic material.[83]

During the past decade, many of the laws (including statutes and regulations)[84] and strategies[85] developed to address cyber security threats have focused on building the knowledge-base for protecting critical infrastructure. In 2009, the White House established The Comprehensive National Cyberspace Initiative to facilitate a unified response to cyber incidents and promote cyber-security awareness. However, several bills to bolster U.S. cyber security of critical infrastructure were introduced during the 113th Congress, but failed to pass. The United States considers the healthcare and public health, food and agriculture, and water and wastewater sectors as critical infrastructures.[86] Therefore, these laws likely will apply to attacks directed towards or exploiting the healthcare, food, or water systems, but not to life sciences research and the environment because these sectors are not yet considered part of the U.S. critical infrastructure.

Although databases are most often attacked, the U.S. has very few laws governing security of data. Among them are the Gramm-Leach-Bliley Act, also known as the Financial Services Modernization Act of 1999, which provides a framework for financial institutions to protect non-public, consumer financial information,[87] and the Federal Information Security Management Act of 2002, which requires federal agencies to develop and implement agency-wide programs on information security to protect agency operations.[88] Several U.S. states passed laws requiring notification of a cyber or data security breach. However, these laws do not necessarily apply to securing life sciences data repositories held by private entities or the cloud.

---

[83] The U.S. Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA (2010) requests gene synthesis companies to screen their orders for genetic sequences from Biological Select Agents and Toxins. Genetic material from certain agents are restricted (i.e., smallpox virus), subjected to export control, or regulated by the Select Agents and Toxins Regulations (7 CFR 331 and 42 CFR 73).

[84] USA PATRIOT Act of 2001 (18 USC 2516, 28 USC 509); Cybersecurity Research and Development Act of 2002 (15 USC 278 passim, 1862, 7401); E-Government Act of 2002 (25 USC 278g-3,40 USC 11301, 40 USC 11331, et seq, 44 USC 3505, 3541, 3544). The Intelligence Authorization Act of 2010 (6 USC121) requires the President to notify Congress of all cyber security programs in operation.

[85] National Strategy to Secure Cyberspace (2003); Homeland Security Presidential Directive-7, Critical Infrastructure Identification, Prioritization, and Protection (2004); Homeland Security Presidential Directive-23/National Security Presidential Directive-54 Cyber Security and Monitoring; National Initiative for Cybersecurity Education (2010); National Strategy for Trusted Identities in Cyberspace (2011); Executive Order 13636: Improving Critical Infrastructure Cybersecurity (2013); Presidential Policy Directive-21, Critical Infrastructure Security and Resilience (2013); NIST Cybersecurity Framework (2014)

[86] See Department of Homeland Security list of critical infrastructure sectors. Accessible at: http://www.dhs.gov/critical-infrastructure-sectors. Accessed on September 7, 2014.

[87] 12 USC 24a, 78, 248b, 377, 1831v-y, 1848a, 2908, 15 USC 80, 80b et seq.

[88] 44 USC 3541.

The legal infrastructure on privacy has been well-documented by the 2014 White House report, *Big Data and Privacy: A Technological Perspective*.[89] Specifically focusing on Big Data in the life sciences, three laws govern privacy of personal data: The Privacy Act of 1974, which governs the collection, maintenance, use and dissemination of personal information by federal agencies,[90] the Health Insurance Portability and Accountability Act, which protects the confidentiality and security of healthcare information, but not genomic information even if stored in electronic health records,[91] and the Genetic Information Nondiscrimination Act of 2008 (GINA), which prevents discrimination for certain insurance coverage and employment based on genetic information.[92]

Additionally, the United States has a variety of statutes and regulations with which to prosecute an individual or organization for a criminal act within and/or against the U.S. Several of these legal instruments specifically pertain to biological security offenses

Institutional Solutions

Institutions play an important role in preventing or mitigating cyber, data, and biological security risks. Attacks on the cyber infrastructure, data storage and sharing systems, violent extremist groups, intellectual-property theft, and campus violence have resulted in the development of several institutional measures to prevent such acts and protect the institutions' staff, students (if applicable), patients (if applicable), and information. These measures include personnel security, physical security, education and training, review and oversight of scientific activities, threat assessment teams, and partnerships with local and/or federal law enforcement.

Private industry, healthcare providers, and research institutions (including academic institutions) are legally required to have strict controls on access to sensitive data, biological materials, and the computer systems collecting and processing the data. These sectors often conduct background checks and other vetting processes on new personnel to prevent theft of information or inappropriate behavior in the workplace.[93] The healthcare sector typically addresses security as they apply to the health insurance portability and accountability act. Academic institutions also conduct background checks on personnel with access to restricted biological, chemical, and radiological agents. The personnel security challenge at academic institutions is complicated by its diverse student, faculty, and staff bodies, which include high school, undergraduate, and graduate students from anywhere in the world and who frequently come from a very different environment than private industry.

Various threats, including those of violent animal rights extremism, ecoterrorism, theft of materiel, theft of proprietary products and information, and/or workplace and campus violence, have prompted many, if not most, scientific institutions to implement physical-security measures,

[89] President's Council of Advisors on Science and Technology. Big Data and Privacy: A Technological Perspective. May 2014.
[90] 5 USC 552a.
[91] 110 Stat. 1936.
[92] 29 USC 216, 1132, 1182, 1191b, 42 USC 300gg passim, 1395passim
[93] Examples of personnel security in the workplace are presented in AAAS, AAU, APLU, FBI. (2014). Bridging Science and Security for Biological Research: Personnel Security Programs. Accessible at http://www.aaas.org/report/bridging-science-and-security-biological-research-personnel-security-programs. Accessed on: October 12, 2014.

which include fences, guards, facility access measures, and surveillance cameras. In addition, many academic institutions have developed threat assessment teams that include staff from relevant institutional offices, such as human resources and general counsel, and local and sometimes federal law enforcement. These threat assessment teams are intended to assist institutions develop, and if necessary, implement strategies to prevent devastating threats directed towards the institution and its employees and students. Finally, the FBI/WMDD outreach activities have resulted in the development of strong relationships between regional FBI WMD Coordinators and leadership at scientific institutions. These partnerships can help the FBI and institutions anticipate and prevent threats, and develop approaches to mitigate security vulnerabilities.

Most institutions have established policies and standards for employee and student behavior. These measures include corporate-responsibility policies for private industry and codes of ethics for the healthcare sector, student codes of conduct at academic institutions, and policies for employee behavior. In addition to supporting behavioral standards, institutions can educate and train their employees and students about risks they might encounter; these training activities could include security awareness, approaches for reducing risks, and institutional and national policies to prevent or mitigate risks. Life scientists are routinely educated about risks and prevention/mitigation strategies as they apply to safety, occupational health, and ethics. Life science activities (whether in research or healthcare) often are evaluated by research oversight or institutional leadership. However, scientists from non-life science disciplines are not necessarily educated about potential risks, prevention and mitigation strategies, and relevant institutional and national policies, and their research frequently is not reviewed in this context. Based on the two aforementioned circumstances, broad-based education on security awareness should be conducted at scientific institutions and include relevant legal and international frameworks, such as the Biological Weapons and Toxins Convention, its historical importance, and rationale for its implementation.

Individual solutions

Individuals play a significant role in ensuring that they consider the risks of their scientific pursuits and take reasonable measures to minimize those risks. Within the context of Big Data in the life sciences, these risks includes both inappropriate access to data and computer systems, and the use of data and data analytics to design harmful biological agents. The standard safeguards of protecting passwords and respecting access controls apply to preventing the risk of inappropriate access to data and computer systems; however, institutions play a much more significant role in maintaining the security and minimizing vulnerabilities in its cyber and data infrastructure. Addressing the risk that Big Data analytics can be used to design harmful agents is a more challenging problem. As described in the legal solutions section, much of the policy debate about the misuse of legitimate research (i.e., the “dual use dilemma”) focuses on select, already highly regulated (and restricted) pathogens and toxins. In addition, the policies focus on the creation of scientific knowledge (through laboratory experimentation) that could be misused by an unknown adversary and/or on the actual development of harmful pathogens. However, these policies do not obviously pertain to the design of harmful agents on computers only and using publicly available data. To minimize the risk that Big Data analytics, and other emerging and enabling technologies, would be used to design harmful biological agents, individuals and

their institutions (and/or communities as might be the case for amateur scientists) will have to take the initiative and assess the potential risks and benefits of their scientific efforts, even though these efforts do not fit into the current U.S. government policy on dual use life sciences research of concern. To facilitate these assessments, individuals and institutions should be supported by an understanding of possible security risks, threats, and vulnerabilities associated with a particular emerging or enabling technology. However, no educational and/or informational programs on security awareness are provided to individuals, customers, and consumers.

Solution Scenarios

Not all solutions described in the previous sections address all risks at all times. To be most effective, legal, technical, institutional, and individual solutions should be tailored to the types of risks anticipated. This is especially critical for Big Data in the life sciences in which an attack of a database or a networked computer system results in harmful consequences in the life sciences or healthcare sector, thereby negatively affecting the societal population at large. The risks presented by Big Data in the life sciences include the theft or manipulation of data and analytic results, or exploitation of data and advanced computational technologies to design harmful biological agents.

Furthermore, some Big Data threats likely will be conducted by a group that makes use of multiple avenues to access, use, and/or modify data. While previously described solutions may limit some of the avenues and may allow tracking of accesses to or modifications of data, the ability to recognize or react to a threat at the level of a coordinated group attack will be challenging without taking a solution-based approach. Big Data analytics can be leveraged to identify the occurrence of individuals or groups involved in carrying out these scenarios, or subsequently in a forensics analysis, especially if the signals involved can be correlated across the many activities involved in these scenarios. For instance, the development of a new virus by a group will generate multiple signals (e.g., data downloads, purchases of laboratory equipment, and literature reviews) that could be correlated to identify the scientific activity. However, the signals needed to conduct forensic analysis using Big Data might be challenging to detect with good precision and consequently, such techniques will rely on complex models that are scenario-dependent.

The three risk scenarios presented in the text boxes reflect technically sound, realistic risks. (For an expanded discussion about the risk scenarios, refer to Chapter 3.) Presented below are proposed solutions that reflect the different types of risks included in the aforementioned scenarios.

Box 1: Scenario 1

Risk Scenario: Targeting of Subpopulations using Big Data.

A well-resourced country wants to target a specific subpopulation within its borders by developing a virus that preferentially infects a group among its citizens. Scientists analyze data from large numbers of scientific publications and genomic information in public databases to identify a virus and associated host genes specifying preferential infection in the subpopulation. Scientists use the information to acquire the virus and modify it to increase its specificity for the subpopulation.

(Scenario 1) **Targeting of Subpopulations using Big Data.** This scenario involves the exploitation of Big Data technologies and several life sciences datasets, including scientific literature, human genomic data, and infectious disease epidemiologic data, to design a virus that specifically causes severe infection on a particular subpopulation. In this scenario, the virus is *not* intended to be a regulated pathogen.

The laws governing the development, production, and stockpiling of biological agents for use as weapons would immediately apply to this scenario if the modified virus were made. However because the virus is not a restricted virus, the U.S. Select Agents and Toxins Regulations would not apply to this scenario.

No adequate laws exist to prevent adversaries from using data and analytic technologies to generate a computer design of a biological agent, which in this case would be used as a weapon if created in the laboratory. However, statutes and regulations do exist to prevent the *use* of a biological agent to cause harm, regardless of the agent or how it was designed.

Currently, the datasets used in the scenarios are largely open-access and subject to tracking of IP addresses visiting the sites. Access controls to specific data and analytic software could enhance prevention of exploitation of open-access data and allow for future forensics analysis. Security elements integrated into the software and design tools would also be a technical solution in this scenario.

Institutions could implement access controls on their computer network, educate scientists about the risks of exploitation of data and Big Data technologies, and implement strategies to prevent or mitigate the risks. In addition, institutions can promote standards of behavior for employees and students, and provide reporting options for employees and students who suspect a breach.

Individuals can learn about the risk of exploitation of Big Data in the life sciences and foster an environment in which risks are assessed on an ongoing basis and prevention or mitigation strategies are identified and used.

Box 2: Scenario 2

Risk Scenario: Misdirection.

A group outside the U.S. wants to prevent government and international health officials from discovering its covert scientific activities to modify a biological agent it acquired from a local laboratory. The domesticated animals around its "laboratory" facilities were exposed accidentally to the agent. Shortly after this exposure event, a small number of individuals were diagnosed with a respiratory infection caused by an agent not identified with local diagnostic tools. The local doctors report these infections to public health officials, other clinicians, and scientists using an infectious disease surveillance tool that relies on Big Data. Concerned that these reports would alert international and national authorities to the group's activities, one member of the group submits false reports into the system (e.g., using location of infections, severity of infection) to mask identification of the unusual infections and reduce the likelihood that its laboratory would be identified as the source of the outbreak.

(Scenario 2). **Misdirection.** This scenario involves manipulation of data, specifically a public health reporting database, to hide or delay identification of the outbreak.

The only regulations that would affect this scenario are the various national legislation implementing the Biological Weapons and Toxin Convention and UN Security Council Resolution 1540 of the country in which the group operates. Some U.S. security laws pertain to this scenario.

Access-control technologies could prevent access to the reporting systems. Network separation could be used to minimize the vulnerability of reported data. IP addresses could be tracked to monitor visitors to the reporting system.

Box 3: Scenario 3

Risk Scenario: Avoiding Detection.

A non-state/lone actor uses new advances in computational analyses of multiple types of scientific data to determine which genetic segments are found only in dangerous strains (not the harmless variants) of a specific pathogen. Through publically accessible information, the adversary infers that these segments will be used as the target of molecular tests for pathogen detection (e.g., Polymerase Chain Reaction (PCR)). Once these segments are identified, the adversary makes specific changes to the gene that change only the genetic sequence but not the subsequent protein sequence to prevent loss of function. This would allow the virus to be effective, but undetectable.

(Scenario 3) **Avoiding Detection.** This scenario involves the exploitation of Big Data technologies and life sciences datasets, including pathogen genome databases and scientific literature, to predict the gene segments that are used for pathogen detection and the genetic changes needed to avoid detection. This scenario might involve a restricted pathogen.

If the pathogen is a Select Agent, all Select Agents and Toxins Regulations pertain to this scenario, but only if the adversary possesses the pathogen; if the adversary has only the gene

sequence or only the gene (i.e., not the full pathogen unless it is smallpox, which is specifically prohibited by 18 USC 175c), the Select Agents and Toxins Regulations do not apply to this scenario. However, the regulations and statutes governing the development, production, and/or stockpiling of biological agents for use as weapons would immediately apply to this scenario if the modified pathogen were made.

Institutions could implement access controls on their computer network, personnel security programs, and education programs for scientists to learn about exploitation of Big Data technologies and life sciences data for malicious purposes. If the adversary does not work at a life sciences or healthcare institution, no institutional solutions would apply to the scenario.

Individuals can learn about the risk of exploitation of Big Data in the life sciences and provide an environment in which risks are assessed on an ongoing basis and prevention or mitigation strategies are identified.

Conclusions

Although a range of solutions exists to prevent or mitigate cyber, data, and biological security threats, these solutions generally are specific to each threat area. Of the three threat areas, biological security has the most robust governance framework. However, no adequate statutes, regulations, guidance, policies, or strategies apply to the use of biological data to design harmful agents, and/or produce a biological security consequence by exploiting vulnerabilities in the data and cyber infrastructure. Technical solutions are most developed for protecting data and the cyber infrastructure. They can offer protection from hacking and other computer-assisted breaches. But they cannot prevent the exploitation of data or analytic tools if the systems are being compromised as is the case for two of the three risk scenarios. Institutions have a diverse set of solutions to prevent or mitigate cyber, data, and biological security. However, these solutions only apply if the adversary is at an institution that has implemented one or more of the institutional solutions listed. Finally, individual scientists can play a significant role in ensuring that they follow institutional, local, state, and federal policies and they adhere to workplace standards of behavior; this responsibility may be challenging and involve other departments, such as human resources or general counsel, when individuals are asked to ensure that their colleagues are following the rules and expected behavioral standards. The use of Big Data to detect signals for forensic analysis of ongoing or past security incidents is theoretically possible, but a deeper examination of the available approaches, the validity of these approaches, and possible new regulations that boost the signals are needed. This exercise of evaluating possible solutions and applying those solutions to three risk scenarios has demonstrated that gaps remain in law and technology development, which currently results in a greater reliance on institutional and individual responsibility to prevent theft, manipulation, and exploitation of Big Data in the life sciences. Support must be provided to ensure that the scientific and security communities work in a coordinated and collaborative manner to identify threats and vulnerabilities, and subsequently develop measures to detect, deter, prevent, and respond to any security concerns associated with Big Data and its various applications in the life sciences.

Appendix 1:

Glossary of Terms

Adversary: One's opponent in a conflict or dispute.

Anthrax: An infectious disease caused by a bacterium that occurs naturally in many parts of the world and infects humans and animals. It has two forms: a vegetative form which can make more of itself and produce a potentially lethal toxin; and a spore form which is resistant to heat and harsh chemicals. It can infect humans by exposure to the skin, gastrointestinal track, and lungs.

Antimicrobial Medicines: Molecules that prevent or treat infection by bacteria, viruses, or other microbes.

Aphenotypic Information: Information that does relate to an observable characteristic of a life form.

Bandwidth Consumption Attack: A form of denial-of-service attack that saturates the total bandwidth of the target with large volumes of data packets. It hinders prompt and effective responses to normal requests, leading to denial of services. This attack can be divided into three different categories: direct attack on the target, reflection and amplification attack, and attack on the link.

Biological Threat Agents: Pathogens (including harmful bacteria, viruses, and other organisms) and toxins that pose significant threats to national security (i.e., biological weapons) and/or public safety (e.g., select agents).

Biologically Active: Works in and/or causes a reaction in life forms.

Biological Select Agents and Toxins (BSAT; "Select Agents"): Pathogens and biologically derived toxins that could "pose a severe threat to public health and safety, to animal or plant health, or to animal or plant products"[94] and/or the health of livestock and crops. The BSAT is determined by the U.S. Centers for Disease Control and Prevention (CDC) and the U.S. Animal and Plant Health Inspection Service (APHIS).[95]

[94] Select Agent Program. Accessible at www.selectagent.gov. Accessed on July 3, 2014.
[95] 42 CFR Part 73, 9 CFR Part 121, and 7 CFR Part 331

Biological Weapon: A harmful biological agent, such as a pathogen or toxin that is used as a weapon to cause death, disease, disruption, and/or destruction.

Botnet (Also known as a "bot" or "zombie army"): A group of computers that are controlled from a single source and that run related software programs and codes. While legal botnets can be used for distributed computing purposes, such as scientific processing, illegal botnets refer to multiple computers that have been infected with malicious code and whose security defenses have been breached and control conceded to a third party.[96] Malicious hackers can use a botnet for large attacks (such as Distributed Denial of Service attacks or "floods") that would not be possible if they used just one system.

Broad-spectrum Antibiotics: Molecules that prevent or treat infection by many types of bacteria.

Computational Algorithms: Mathematical calculations conducted by computer programs.

Computer Virus: A self-replicating, malicious code that attaches itself to an application program or other executable system component to interfere with computer operation while leaving no obvious signs of its presence.[97] These malicious codes can create files, move files, erase files, consume a computer's memory, and/or cause a computer not to function correctly. Some computer viruses can duplicate themselves, attach themselves to programs, and travel across networks. Computer viruses are often spread through attachments in email messages or instant messaging messages, or through internet downloads.

Critical Infrastructure: The backbone of the United States' economy, security and health according to the United States Department of Homeland Security.[98] This includes the assets, systems, and networks, whether physical or virtual so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, and/or any combination thereof.

Cross-Disciplinary: More than one discipline, which requires the integration of concepts, vocabularies, methods, and communities.

Data Breach: A security incident in which sensitive, protected and/or confidential data are copied, transmitted, viewed, stolen and/or used by an individual unauthorized to do so.

Data Mining: The process of collecting, searching through, and analyzing a large amount of data in a database to discover new patterns, information, and/or relationships. Mining models can be applied to specific scenarios, such as:

- Forecasting: Estimating sales, predicting server loads or server downtime;
- Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes;

[96] "National Information Assurance Glossary", CNSS Instruction No. 4009, Committee on National Security Systems.
[97] "National Information Assurance Glossary", CNSS Instruction No. 4009, Committee on National Security Systems.
[98] U.S. Department of Homeland Security. Critical Infrastructure. Accessible at http://www.dhs.gov/what-critical-infrastructure. Accessed on October 12, 2014.

- Recommendations: Determining which products are likely to be sold together, generating recommendations;
- Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events;
- Grouping: Separating customers or events into clusters of related items, analyzing and predicting affinities;[99]

Data Repository: A logical (and sometimes physical) partitioning of data where an aggregation of data and/or multiple databases which apply to specific applications or sets of applications reside.

Database: A comprehensive collection of related data organized for convenient access, usually in a computer. The most prevalent approach is the relational database, a tabular scheme in which data are defined so that they can be reorganized and accessed in a number of different ways. A distributed database is one that can be dispersed or replicated among different points in a network. An object-oriented programming database is one that is congruent with the data defined in object classes and subclasses. Computer databases typically contain aggregations of data records or files, such as sales transactions, product catalogs and inventories, and customer profiles.

Denial of Service (DoS) Attack: Characterized by an explicit attempt by attackers to prevent legitimate users of a service from using that service.[100] The United States Computer Emergency Readiness Team (US-CERT) defines symptoms of denial-of-service attacks to include:

- Unusually slow network performance (opening files or accessing web sites);
- Unavailability of a particular web site;
- Inability to access any web site;
- Dramatic increase in the number of spam emails received—(this type of DoS attack is considered an "email bomb");
- Disconnection of a wireless or wired internet connection; and
- Long term denial of access to the web or any internet services.[101]

A DoS attack can be perpetrated in a number of ways. The basic methods of attack include:

- Flooding a network, thereby preventing legitimate network traffic;
- Disrupting connections between two machines, thereby preventing access to a service;
- Preventing a particular individual from accessing a service;
- Disrupting service to a specific system or person; and
- Disrupting the state of information.[102]

Illegitimate use of resources may also result in denial of service.

[99] "Data Mining Concepts", Microsoft, SQL Server 2014
[100] "Understanding Denial-of-Service Attacks – Security Tip (ST04-015)", United States Computer Emergency Readiness Team, US Department of Homeland Security
[101] "Understanding Denial-of-Service Attacks – Security Tip (ST04-015)", United States Computer Emergency Readiness Team, US Department of Homeland Security
[102] "Understanding Denial-of-Service Attacks – Security Tip (ST04-015)", United States Computer Emergency Readiness Team, US Department of Homeland Security

DoS attacks are often spread by email viruses and are deliberately planned attacks on a companies or organizations, Web sites, or servers.[103] DoS attacks are a common method hackers use to attack websites.

Distributed Denial of Service (DDoS) Attack: A multitude of compromised systems attack a single target, thereby causing denial of service for users of the targeted system. An attacker may use a computer to attack another computer. By taking advantage of security vulnerabilities or weaknesses, an attacker could take control of a computer and then force a computer to send huge amounts of data to a website or send spam to particular email addresses. The attack is "distributed" because the attacker is using multiple computers to launch the denial-of-service attack.[104]

Three main types of DDoS attacks exist: Volume Based Attacks in which the attacker's goal is to saturate the bandwidth of the attacked site; Protocol Attacks, which consume actual serve resources; and Application Layer Attacks, which seek to crash the web server.

A computer under the control of an intruder is known as a zombie or bot. A group of co-opted computers is known as a botnet or a zombie army. Both Kaspersky Labs and Symantec have identified botnets -- not spam, viruses, or worms -- as the biggest threat to Internet security.

Doctoral Committee: A group of professors and/or other qualified individuals at a university or college that assesses the knowledge and capabilities of graduate students, recommends whether they pass or fail their qualifying exams, and approves or rejects their receipt of a Ph.D. or equivalent degree.

Ebola Virus: A virus that causes severe, often fatal illness in both humans and animals. It occurs naturally in Central and West Africa. Nearly 90% of human infections result in death, most frequently from hemorrhage.

Encryption: The conversion of data into a form that cannot be easily understood by unauthorized people; commonly referred to as a cipher

Epigenomic: The study of changes in the control of gene function and expression (e.g., being made into a protein) that are not dependent on genetic sequence. Includes methylation of DNA and modifications of histone proteins that wrap around the DNA.

Drug Candidate: Anything that can be used to treat a disease or condition.

Drug Targets: Molecular sites of action for the drug, such as membrane receptors or enzymes.

Epidemic: A widespread occurrence of an infectious disease in a community at a particular time.

[103] Definition provided by The American Heritage® New Dictionary of Cultural Literacy (3rd ed.)
[104] "Understanding Denial-of-Service Attacks – Security Tip (ST04-015)", United States Computer Emergency Readiness Team, US Department of Homeland Security

Firmware: Permanent software programmed into a read-only memory.

Flooding: A type of incident involving insertion of a large volume of data resulting in denial of service.[105] During a flooding DoS attack, the attacker attempts to send continuous traffic (often a high volume of traffic) to a target server in an attempt to crowd out legitimate users and prevent their ability to access information on the servers. In a network, flooding is the forwarding of information from one node to every other node in the network through the network's router. Types of floods include, but are not limited to, HTTP floods, ICMP floods, SYN floods, and UDP floods.

Forensic analysis: The use of scientific methods to gather and examine information about a past incident.

Fraggle Attack: A denial-of-service attack which remotely consumes a network's bandwidth by sending a large amount of spoofed User Datagram Protocol (UDP)[106] traffic to a router's broadcast address within a network. A Fraggle generates even more traffic and can create even more damaging effects than a Smurf attack.

Genome: Contains the information (i.e., the DNA) that provides a blueprint for the development and function of a living organism. Deoxyribionucleic acid (DNA) builds the basis for the inheritance of genetic traits.

Genomic: The study of the structure, function, evolution, and mapping of genetic information.

Hacker: In the computer security context, "an unauthorized user who attempts to or gains access to an information system."[107]

Immunogenomic Data: Information about the differences in the genes that control the immune system or immune response.

Immunogenicity Profile: The body's immune response to a particular foreign substance.

Immunologist: A scientist who studies the body's responses to infection with a pathogen, introduction of a foreign item such as a transplanted organ, and/or exposure to a foreign substance, such as a toxin, pollen, or other allergen.

Industrial Control System (ICS): A general term that encompasses several types of control systems used in industrial production, including supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and other smaller control system

[105] "National Information Assurance Glossary", CNSS Instruction No. 4009, Committee on National Security Systems. Available online at http://jitc.fhu.disa.mil/pki/documents/committee_on_national_security_systems_instructions_4009_june_2006.pdf
[106] UDP is widely used for streaming audio and video, voice over IP (VoIP) and videoconferencing
[107] "National Information Assurance Glossary", CNSS Instruction No. 4009, Committee on National Security Systems.

configurations such as programmable logic controllers (PLC) often found in the industrial sectors and critical infrastructures.[108]

Lipids: Naturally-occurring biological molecules, commonly called "fats", that have a molecular feature that attracts water and a molecular feature that repels water.

Malware: Short for "malicious software," which refers to software programs designed to damage or do other unwanted actions on a computer system. Common examples of malware include computer viruses, worms, Trojan horses, ransomware, and spyware.

Malicious Code: Computer program code intended to perform an unauthorized function or process that will have an adverse impact on the confidentiality, integrity, or availability of an information system.[109] Malicious code could include software, firmware, and scripts.

Medical Diagnostics: Tests to detect specific molecules to make a diagnosis of disease or predict the response to therapy.

Metabolomics: The study of processes involving metabolites (i.e., substances) produced during the physiological chemical processes.

Metagenomic Sequencing: Sequencing of the genetic information of microbes, usually the ribosomal RNAs of numerous microbes in ecosystems or on surfaces of the body of larger organisms (skin, respiratory tract, urinary track, gastrointestinal tract).

Microbe: A single-cell, tiny life form that is too small to see without magnification. Bacteria, protozoa, and fungi are microbes. Biological viruses are classified as microbes even though they are not single-cell organisms and often not considered living.

Microbiologist: A scientist who studies micro-organisms.

Microbiome: The community of microbes in a specific place commonly characterized with meta-genomic sequencing.

Misdiagnosis: The incorrect identification of the cause of a disease or syndrome.

National Security Act of 1947[110]: Mandated a major reorganization of the foreign policy and military establishments of the U.S. Government, specifically:

- Established the National Security Council (NSC) to "advise the President with respect to the integration of domestic, foreign, and military policies relating to the national security so as to enable the military services and the other departments and agencies of the Government to cooperate more effectively in matters involving the national security;"

[108] "Guide to Industrial Control Systems (ICS) Security", National Institute of Standards and Technology, US Department of Commerce, Special Publication 800-82, June 2011. Available online at http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf
[109] "National Information Assurance Glossary", CNSS Instruction No. 4009, Committee on National Security Systems.
[110] 50 USC 401

- Merged the War and Navy departments into the National Military Establishment (NME) headed by the Secretary of Defense;
- Recognized the US Air Force as an independent service from the Army; and
- Ratified the creation of the post of Director of Central Intelligence (DCI), and transformed the Central Intelligence Group into the statutory Central Intelligence Agency (CIA), the nation's first peacetime intelligence agency.

Non-replicating organisms: Life forms that do not make more copies of themselves.

Non-state/lone Actor: An individual or organization that has significant political influence but is not allied to any particular country or state."[111] The Department of Defense Irregular Warfare Joint Operating Concept (2007) states that "non-state/lone actors are a group or organization that is not within the formal structure of any state, not limited by any state boundary, and operates beyond the control of any state and without loyalty to any state. Examples include international organizations, nongovernmental organizations, political parties, labor unions, commercial trade associations, criminal enterprises, and armed groups, such as insurgent and terrorist organizations, informal armed militias, and private military companies."[112]

In a security context, armed non-state/lone actors are those "actors who are willing and able to use violence to pursue their objectives; are not integrated into formal state institutions, such as regular armies, presidential guards, police, or Special Forces; possess a certain degree of autonomy with regard to politics, military operations, resources, and infrastructure; and have an organizational structure. The definition encompasses politically and ideologically motivated actors as well as profit-oriented actors."[113] Because they often fall outside the traditional Westphalian international system, their actions and use of force also often fall outside international law.

Open Source Data: Data that can be freely used, shared, and built-on by anyone, anywhere, and for any purpose.

Pathogen: An infectious bacterium, virus, or other microbe that can cause disease in plants, animals, and/or humans.

Pathogen Detection: The process of identifying the pathogen that causes a disease.

Pathogenicity: The ability of an organism to cause disease.

Pharmacogenomics: The study of genes that introduce variation among individuals in the response to drug action or in the metabolism and transformation of the drug in the body.

Physiologic Response: The body's normal range of response to a particular substance, microbe, internal change, or stimulus.

[111] Definition provided by the Oxford English Dictionary
[112] "Irregular Warfare – Joint Operating Concept", Department of Defense, 11 September 2007
[113] "NGOs and Nonstate Armed Actors", Claudia Hofmann and Ulrich Schneckener, United States Institute of Peace, July 2011

Polio Virus: An infectious agent that predominantly causes paralytic polio in children less than five years old. Approximately 0.5% of infections result in the permanent paralysis of limbs. Of those people who have developed paralysis, approximately 5-10% die because the nerves to muscles in the lungs are paralyzed. Effective vaccines exist to prevent polio infection.

Polymerase Chain Reaction: A scientific method used to amplify and assay small amounts of DNA.

Precision Agriculture: The integration and analysis of different types of data (e.g., weather, soil content, regional information, satellite imagery) to inform farming and agricultural practices to improve soil-enrichment methods and selection of new traits to enhance crops.

Product Efficacy: The degree to which a drug or vaccine is effective at doing what it was developed to do once in the body.

Programmable Logic Controller (PLC): A solid-state control system originally designed to perform the logic functions executed by electrical hardware (relays, switches, and mechanical timer/counters). PLCs have evolved into controllers with the capability of controlling complex processes, and they are used substantially in SCADA systems and DCS.[114]

Proteins: Naturally-occurring biological molecules made from amino acids directed by the sequence of DNA in the protein-coding genes. Proteins execute most of the complex functions of living organisms, acting as enzymes, receptors, signaling molecules, hemoglobin to carry oxygen to the tissues, collagens to provide structure in bones and other tissues, and many other functions.

Protein complexes: Groups of two or more proteins.

Proteomics: The study of all or many proteins or protein complexes and their functions.

Qualifying Exams: Tests taken early in graduate school to assess a student's knowledge in his/her area of study.

Ransomware: A form of malware in which rogue software effectively holds a user's computer hostage until a "ransom" is paid. It may require the user to pay money, complete surveys, or perform other actions to unlock and use his/her system. Ransomware often infiltrates a PC as a computer worm or Trojan horse that takes advantage of open security vulnerabilities. Most ransomware attacks are the result of clicking on an infected email attachment or visiting a hacked website. Some types of ransomware also are called "FBI Moneypak" or the "FBI virus" because they often use the logos of the Federal Bureau of Investigation (FBI), state, or local law enforcement agencies and ask the user to pay a fine using a lawful money transfer service such as Green Dot MoneyPak.[115]

[114] The Automation, Systems, and Instrumentation Dictionary, 4th Edition, ISA, 2003.
[115] Microsoft Malware Protection Center Definition

Ricin: A naturally occurring toxin made from the oil produced by caster beans. Because of its high toxicity and wide-spread availability of the plant, ricin has been developed and used as a biological weapon by individuals, groups, and nations.

Script: In computer programming, a program or sequence of instructions executed by another program rather than by the computer processor (as is a compiled program).

Select Agents: *See* Biological Select Agents and Toxins above.

Select Agent Program: The Select Agent Program is jointly regulated by the CDC Division of Select Agents and Toxins and the APHIS Agricultural Select Agent Program. The Program oversees the possession, use and transfer of Biological Select Agents and Toxins (BSAT) enhancing the nation's oversight of its safety and security by pursuing the below objectives:

- Developing, implementing, and enforcing the Select Agent Regulations;
- Maintaining a national database of pathogens and toxins that cause public health and safety concern;
- Monitoring entities that possess, use, and/or transfer BSAT;
- Ensuring that all individuals who possess, use, and/or transfer BSAT undergo a security risk assessment performed by the FBI Criminal Justice Information Service;
- Providing guidance to those who possess, use and/or transfer BSAT on achieving compliance with the regulations through the development of guidance documents, and by conducting educational programs; and
- Investigating any incidents in which non-compliance may have occurred.[116]

Smurf Attack: A DoS attack in which Internet Protocol (IP) Broadcast addresses are exploited. During a Smurf attack, large numbers of Internet Control Message Protocol (ICMP) packets with the intended victim's spoofed source IP are broadcast to a computer network using an IP Broadcast address. When PING, a software tool that is available on most operating systems and commonly used to check whether a specified computer is reachable, is executed, an ICMP echo request packet is sent to the destination computer. If the destination computer receives the Transmission Control Protocol (TCP) packet, it replies to confirm the PING request.[117] Each broadcast address can support up to 255 hosts, and so a single PING request can be multiplied 255 times. In the case of a Smurf DoS attack, the PING's packet return IP address is forged with the IP of the targeted computer, causing every computer to respond to the bogus PING packets and reply to the targeted computer, which floods it. A single attacker sending hundreds or thousands of these PING messages per second can fill the victim's T-1[118] (or even T-3[119]) line with PING replies, disrupting the entire Internet service.

This technique is called a Smurf attack because the DoS tool that is used to perform the attack is called Smurf, which name comes from the file "smurf.c", the source code of the attack program released in 1997 by TFreak.

[116] Select Agent Program. Accessible at www.selectagent.gov. Accessed on July 3, 2014.
[117] "Smurf DoS attack", Symantec Glossary
[118] A T-1 line actually consists of 24 individual channels, each of which supports 64Kbits per second. It is a dedicated phone connection supporting data rates of 1.544Mbits per second.
[119] A T-3 line actually consists of 672 individual channels, each of which supports 64 Kbps. It is a dedicated phone connection supporting data rates of about 43 Mbps.

Software: The program and associated operating information used by a computer. Any information that can be stored electronically is software.

Spoofing: A situation in which an attacker pretends to be someone else in order gain access to restricted resources and/or steal information. This type of attack can take a variety of different forms; for instance, an attacker can impersonate the IP address of a legitimate user in order to get into the user's accounts. Also, an attacker may send fraudulent emails and establish fake websites in order to capture users' login names, passwords, and account information. Faking an email or website to access this information often is called a phishing attack. Another type of spoofing involves setting up a fake wireless access point and tricking victims into connecting to them through the illegitimate connection. Impersonating, masquerading, piggybacking, and mimicking are forms of spoofing.

Strains: Variants of a life form, subsets of species.

Supervisory Control and Data Acquisition (SCADA): A generic name for a computerized system that is capable of gathering and processing data and applying operational controls over long distances. Typical uses include power transmission and distribution, and pipeline systems. SCADA was designed for the unique communication challenges (e.g., delays, data integrity) posed by various media, such as telephone lines, microwave, and satellite. SCADA is usually shared rather than dedicated.[120]

Synthetic pathogens: Infectious agents not made from naturally-existing bacteria, viruses, or other pathogens. Research groups have made synthetic pathogens by making the pathogen's genetic material from purchased genes and chemical synthesis techniques in a test tube.

Tamiflu: A drug that is used to treat infection with the flu virus. It has limited effectiveness in most patients.

Traits: Characteristics or qualities distinguishing one life form from another or one group from another.

Transmissibility: The spread of pathogens between or among individuals.

User Datagram Protocol (UDP): Part of the Internet Protocol suite used by programs running on different computers on a network. UDP is used to send short messages called datagrams.

Vulnerability: The susceptibility of a computer system, data repository, analytic tool, process, or life form to attack, accidents, or other insult.

[120] The Automation, Systems, and Instrumentation Dictionary, 4th Edition, ISA, 2003.

Appendix 2:

Summary of April 1 Event

BIG DATA, LIFE SCIENCES, AND NATIONAL SECURITY: A RECAP

Interest in Big Data has surged over the past two years, as have the opportunities and challenges that Big Data and its various applications present. However, an in-depth assessment of how these activities impact the life sciences and what potential security challenges might exist has not occurred. To promote a discussion of these issues, risks, and the broader implications of Big Data in national and international security with respect to the life sciences, a community of scientists, academics, security experts, industry representatives, and decision-makers gathered at the Renaissance Washington, DC Downtown Hotel for a joint event hosted by the AAAS Center for Science, Technology & Security Policy and the Federal Bureau of Investigation Weapons of Mass Destruction Directorate on "Big Data, Life Sciences, and National Security."

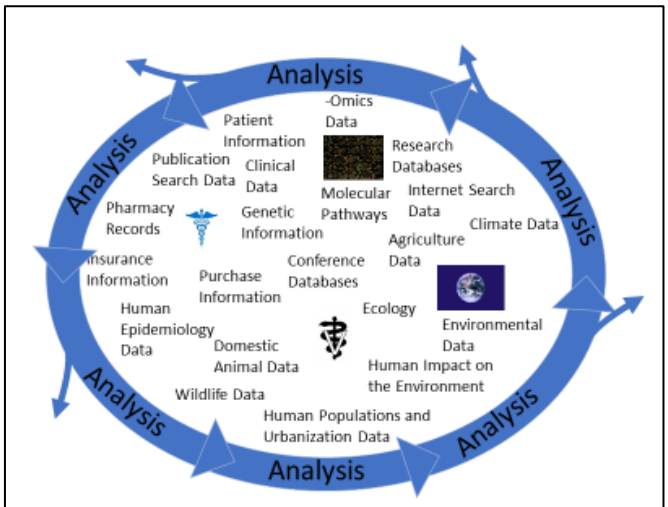


Figure 1. Big Data in the Life Sciences

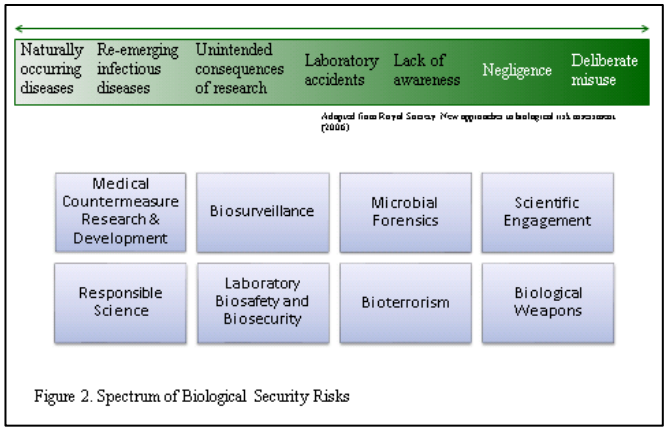


Figure 2. Spectrum of Biological Security Risks

Speakers, working group members, and attendees discussed a variety of topics related to Big Data in the biomedical, clinical, and environmental sciences (Figure 1) and biological security (Figure 2). The objective of the public event was to engage stakeholders in thoughtful dialogue about the benefits of Big Data in the life sciences and to begin identifying associated risks or vulnerabilities to biological security. The event included

panels in which speakers discussed the definition of Big Data and state of the science, described its current and future applications to addressing critical biological security issues, identified possible risks presented by Big Data in the life sciences, and explored the level of broader access to data repositories and analytic technologies.

AAAS CEO Alan Leshner opened the event by providing a framework for the day's discussion stating that "it's more and more urgent that we figure out the most effective ways, not only to analyze Big Data, but to share it and share it widely." Dr. Leshner went on to challenge the attendees by framing the event as an " "opportunity to come up with concrete steps or at least concrete *next* steps to deal with these issues, and the more concrete you can get in your conversation, the more productive the meeting will be."

The event consisted of five panels, each focusing on a different aspect of Big Data in the life science and national security. These panels were:

Big Data: Definition, Sources and Data Sharing
Applications of Big Data and Analytics to International and National Biological Security
Security Risks of Big Data: Privacy, Openness, Data Management
Increasing Access to Big Data and Analytics: Implications for Biological Security?
Implications of Big Data and Analytics to National and International Biological Security

Several major themes emerged from the event:

- Big Data in the life sciences is increasing at an exponential rate regarding volume of data, applications, and analysis tools.
- As the pace with which large volumes of data are collected, stored, and analyzed increases, so do risks of unintentional or intentional errors that could potentially compromise data quality and reliability (i.e. spamming, spoofing).
- Data are being generated at a much faster rate than it can be analyzed or monitored. Consequently, particular attention should be paid to the ownership, openness, disposition and accessibility of the data.
- As the volume of data continue to increase, the reliability of the data and its associated sources becomes a significant factor.
- Reproducibility of data and analytic methods are extremely important for ensuring high confidence in the data and the results of data analysis and assessments. This could be challenging if the processes are proprietary in nature.
- Automated data processing (e.g., data visualization techniques and machine learning) and human analysis are important for producing useful, actionable information from Big Data.
- As data collection, storage, and analysis technologies advance, potential vulnerabilities might be introduced into the system increasing the national and international security risk posed by Big Data in the life sciences.

A key emphasis of the discussions was on the implications of Big Data and analytics for national and international biological security:

- Big Data in the life sciences is currently being developed for biosurveillance and precision medicine, but it could also be applied to synthetic DNA screening, forensic analysis, intelligence gathering, and smart vaccine and drug development.
- The level of uncertainty inherent in Big Data elicited a number of comments on the importance of effective and early communication about the analytic products and its applications to national and international biological security problems with the general public, policymakers, public health workers, law enforcement, intelligence officers, and the research and health community.
- Protection of databases and analytic tools from deliberate misuse is a significant risk associated with Big Data.
- Privacy of data and databases and how aggregation of disparate and discrete datasets may effectively deanonymize information sources are important risks associated with clinical, genomic, healthcare, or other human data.

A balanced approach is necessary to ensure that the data are safe, secure and reliable without discouraging innovation and advancement in Big Data tools and analytics, which operates in a relatively open environment:

- Possible approaches for minimizing the risks of Big Data in the life sciences include improving or developing relevant oversight frameworks (e.g., refining the Institutional Review Board function and process), mechanisms for professional responsibility, national and regional laws and policies, and international legal instruments. Accountability of Big Data, analysis, and application will be important for each of these approaches.
- An approach that addresses cyber threats unique to Big Data across collection, storage, analysis, and applications.
    - Examples of cyber threats unique to Big Data in the life sciences:
        - Intellectual property/proprietary information protection challenges
        - Implications for enforcing deemed export/export licenses
        - Genomic identity theft, misidentification, discrimination, exploitation, violation, or assumption
        - “Hacking”, “spamming”, “spoofing” datasets to alter and/or modify analytic outcomes

An important overall theme of the discussions was that Big Data in the life sciences involves: 1) a system of collection and storage tools, analytic methods, human input, and end-use; and 2) a diverse community of stakeholders including scientists, analysts, and users across a wide variety of disciplines and sectors. Promoting communication and cooperation between these distinct stakeholders and across the system to safeguard Big Data in the life sciences will be critical to preventing, detecting, and ultimately responding to intentional threats or vulnerabilities introduced into the system.

In the coming months, AAAS CSTSP, FBI and its working group members will produce reports based on key issues identified during the April 1 event.

--Nicholas Bashour, AAAS/CSTSP

Appendix 3:

Examples of U.S. Government Investments and Initiatives in Big Data

Department of Commerce (DOC)
Big Data for Small Business
$2 million

DOC App Challenge (2012)
$10,000

Department of Commerce (DOC)/National Oceanographic and Atmospheric Administration (NOAA)
BIG DATA RFI FROM NOAA

Department of Defense (DoD)
Data to Decisions
$250 million annually (with $60 million available for new research projects)

Defense Advanced Research Projects Agency (DARPA)
XDATA
$15.275 million in 2013,
$25.8 million in 2014

Visual Media Reasoning (VMR)
$11.92 million in 2012,
$15.192 million in 2013,
$10.768 million in 2014

Machine Reading and Reasoning Technology
$24.36 million in 2012

Department of Energy (DOE)

Big Data-Aware Terabits Networking challenges
$1.8 million

DOE supporting to establish SDAV
$25 million

DOE's American Energy Data Challenge
$100,000 in prizes

Department of Homeland Security (DHS)
Department of Homeland Security Science & Technology -
How Technology Can Address Homeland Security Challenges: Big Data

Internal Revenue Service (IRS)
Big Data and Analytics at the IRS

National Archives and Records Administration(NARA)
Digital Processing Environment
$9.45 million FY 2014

National Aeronautic and Space Administration (NASA)
NASA's Big Data activities and approaches
NASA has invested $9 million in data and information system research and development focusing on big-data management and data-mining algorithms.

National Institutes of Health (NIH)
Big Data to Knowledge (BD2K) initiative
$96 million over the next four years.

U.S. Brain initiative
$200 million(FY 2015)

Human Genome Project
$2.7 billion

The Cancer Genome Atlas program (TCGA)
$153.5 million + $18 million per year

Naitonal Institute for Standards and Technology (NIST)
NIST Big Data Definition

NIST Data Science Program Vision and Goals

National Science Foundation (NSF)
Program solicitation for big data
$23 million/year

Project sizes can range from $200,000 to $500,000

CIF21 investments in big data
Over $75 million in fiscal year 2013

U.S. Agency for Internatoinal Development (USAID)
Global Forest Watch
USAID contributed $5.5 million to the project

Department of Agriculture (USDA)
Investments in cyber infrastructure for big data
$44 million

U.S. Geological Survey (USGS)
Coordination and implementation of the Big Earth Data Initiative for the Department of the Interior
$9.0 million

Department of Veterans Affairs (VA)
Million Veterans Program
Unknown

Appendix 4:

Selected Activities in Big Data in the Life Sciences

Private Sector Efforts

Apixio(with Dell)
Using Big Data for Effective Clinical Performance Management
http://www.apixio.com/press-release/289-using-big-data-for-effective-clinical-performance-management

Explorys
Explorys Cloud Services Power Accountable Care Organizations
http://www.informationweek.com/healthcare/analytics/explorys-cloud-services-power-accountable-care-organizations/d/d-id/1204328

GlaxoSmithKline (GSK)
GSK has partnered with the European Bioinformatics Institute(EMBL-EBI) to establish the Center for Therapeutic Target Validation (CTTV)
http://www.fiercebiotechit.com/story/gsk-teams-apply-big-data-target-validation/2014-03-30

GNS Healthcare, Inc
Gns Healthcare, Dana-Farber And Mount Sinai Collaborate to Build Computer Model of Multiple Myeloma
http://www.gnshealthcare.com/news-and-events/gns-healthcare-dana-farber-and-mount-sinai-collaborate-to-build-computer-model-of-multiple-myeloma/

IBM
Delivering quality care
through advanced analytics
http://www-01.ibm.com/software/analytics/healthcare/

IBM's Watson analyzes the human genome to battle brain cancer

http://www-03.ibm.com/press/us/en/pressrelease/43444.wss

IBM introduced new analytic software similar to Watson for the healthcare industry
http://www-03.ibm.com/press/us/en/pressrelease/35597.wss

Improve patient care and outcomes
http://www-01.ibm.com/software/data/bigdata/industry-healthcare.html

Illumina*
NextBio
http://www.illumina.com/informatics/biological-data-mining.ilmn

Intel
Leverage Healthcare Big Data for a Better Care Environment
http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/healthcare-leveraging-big-data-paper.pdf

Intel (with Dell)
Delivering Personalized Medicine
http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/delivering-personalized-medicine-today-paper.pdf

Kaiser Permanente
Data Helps Drive Lower Mortality Rate
http://blogs.wsj.com/cio/2013/12/05/data-helps-drive-lower-mortality-rate-at-kaiser/

Microsoft(with VirginiaTech)
University Transforms Life Sciences Research with Big Data Solution in the Cloud
http://www.microsoft.com/casestudies/Windows-Azure/Virginia-Polytechnic-Institute-and-State-University/University-Transforms-Life-Sciences-Research-with-Big-Data-Solution-in-the-Cloud/710000003381

Novartis
Researchers solved cases of the cause of a rare, inherited, often fatal kidney disease in many patients in just six weeks.
http://www.novartis.com/stories/discovery/2013-10-big-data.shtml

Optum Labs
Big-data collaborative Optum Labs adds seven partners
http://www.modernhealthcare.com/article/20140213/NEWS/302139958

Oracle
Oracle Health Sciences Translational Research Center
http://www.oracle.com/us/products/applications/health-sciences/translational-research/index.html

Quintiles
Infosario® platform: Recognizing data as a strategic asset
http://www.quintiles.com/

SAP
SAP, Stanford University and the National Center for Tumor Diseases (NCT) in Heidelberg initiated to develop real-time personalized medicine
http://www.saphana.com/community/blogs/blog/2014/02/13/nct-heidelberg-honored-with-sap-for-genomics-advances-in-us-federal-big-data-initiatives

SAS
SAS Pull Big Pharma Into Big Data Collaboration
http://www.forbes.com/sites/danmunro/2013/06/27/sas-and-gsk-pull-big-pharma-into-big-data-collaboration/

Teradata
Data and analytics to work in the era of healthcare reform
https://www.teradata.com/industry-expertise/healthcare/

University HealthSystem Consortium (UHC)*
UHC, NYU Langone, Cleveland Clinic collaborate on data transfer program
http://www.modernhealthcare.com/article/20130909/BLOG/309099995/uhc-nyu-langone-cleveland-clinic-collaborate-on-data-transfer-program

Zephyr Health
Zephyr Platform, raises $15 Million from Kleiner Perkins Caufield & Byers and Jafco Ventures.
https://zephyrhealthinc.com/about-zephyr/press/2014-01-08-press-release/

U.S. Government Activities

DARPA
Pentagon looking for 'Big Mechanism' to mine health data
http://www-03.ibm.com/press/us/en/pressrelease/42156.wss

Food and Drug Administration (FDA)*
FDA Adverse Event Reporting System (FAERS)
http://www.policymed.com/2013/11/using-fda-big-data-and-electronic-health-records-to-improve-drug-safety.html

Department of Health and Human Services (HHS)
Health Datapalooza conference
http://www.cio.com/article/734350/Big_Data_Surge_From_Federal_Agencies_Will_Drive_Health_IT

NIH
Big Data to Knowledge (BD2K) initiative
http://bd2k.nih.gov/

NIH*
The Cancer Genome Atlas(TCGA) project
http://cancergenome.nih.gov/

NIH
NIH awarded IBM, Sutter Health, and Geisinger Health System to develop new methods for early detection of heart disease
http://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action%20Announcements.pdf

NIH/National Heart, Lung, and Blood Institute (NHLBI)*
FOA for development of an NIH BD2K Data Discovery Index Coordination Consortium
http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-14-031.html

NSF/NIH
Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)
http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

Academic or Non-Profit Activities

Alzheimer's Association*
Big Data From Alzheimer's Disease Whole Genome Sequencing Will be Available to Researchers
http://www.prnewswire.com/news-releases/big-data-from-alzheimers-disease-whole-genome-sequencing-will-be-available-to-researchers-due-to-novel-global-research-database-215186711.html

American Society of Clinical Oncology(ASCO)
ASCO launches formal development of CancerLinQ™, a learning health system to transform cancer care and improve outcomes for patients
http://www.asco.org/press-center/asco-launches-formal-development-cancerlinq%E2%84%A2-learning-health-system-transform-cancer

BGI
BGI and BioMed Central Launch GigaScience "Big Data" Journal
http://duraspace.org/bgi-and-biomed-central-launch-gigascience-big-data-journal

Cloud-Based Software in BGI
http://bgiamericas.com/

Brigham and Women's Hospital (BWH)*
BWH Big Data Center for Patient-Centered comparative effectiveness research(CER)
http://www.brighamandwomens.org/Research/centers/PCERC/ourinitiatives.aspx

Broad Institute*
Genome Analysis Toolkit (GATK)
http://www.broadinstitute.org/genome-analysis-toolkit

Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data
http://www.broadinstitute.org/files/news/pdfs/GAWhitePaperJune3.pdf

Cancer Research UK
Cancer Research UK teams up with Winton and others to help solve 'big data' challenges
http://www.cancerresearchuk.org/about-us/cancer-news/press-release/cancer-research-uk-teams-up-with-winton-and-others-to-help-solve-big-data-challenges

Case Western Reserve University
Case, Cleveland Clinic and University Hospitals team up for data-sharing venture
http://www.crainscleveland.com/article/20131121/FREE/131129938

Emory University
Pathology Analytical Imaging Standards (PAIS)
http://confluence.cci.emory.edu:8090/display/PAIS/Overview

Georgia Institute of Technology
Big Data @GT
http://bigdata.gatech.edu/

Harvard Medical School (with IBM)
Drug safety and effectiveness studies
http://www.ibmbigdatahub.com/pdf/harvard-medical-school-0

Harvard Medical School (with GNS Healthcare)
Harvard deploys big data for bio study
http://www.healthcareitnews.com/news/harvard-deploys-big-data-bio-study

Indiana University
Electronic Medical Record (EMR) Data integration
http://bigdatawg.nist.gov/_uploadfiles/M0177_v1_1133239355.docx

J. Craig Venter Institute (JCVI)*
JCVI Metagenomics Reports (METAREP)
http://jcvi.org/metarep/

Human Microbiome Project(HMP)
http://hmp.jcvi.org/

Johns Hopkins University*
Johns Hopkins big data experts to collaborate with health scientists
http://hub.jhu.edu/2013/11/05/idies-center-note

Mount Sinai Medical Center
A Hospital Takes Its Own Big-Data Medicine
http://www.technologyreview.com/news/518916/a-hospital-takes-its-own-big-data-medicine/

Oregon Health & Science University
Intel, OHSU Use Supercomputing to Analyze Genomic Data
http://www.eweek.com/servers/intel-ohsu-use-supercomputing-to-analyze-genomic-data/

Renaissance Computing Institute (RENCI)
Big Data IT to manage, decipher, and inform
http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/life-sciences-at-renci-white-paper.pdf

Sage Bionetworks*
Alzheimers Disease Big Data DREAM Challenge
http://www.ceoalzheimersinitiative.org/global-ceo-initiative-alzheimer%E2%80%99s-disease-announces-big-data-challenge-find-new-predictors-cogniti-0

Big Data Challenge to Develop Robust Methodologies for Predicting Cancer Mutations
https://www.synapse.org/#!Challenges:DREAM

San Diego Supercomputer Center*
Cancer Genomics Hub Offers more access to Big Data
http://cancerdiscovery.aacrjournals.org/content/2/6/476.2.full

Stanford University*
Big Data-focused genomics center
http://www.fiercebiotechit.com/story/stanford-rival-broad-institute-big-data-focused-genomics-center/2012-12-03

Stanford and Google team up to simulate key drug receptor
http://med.stanford.edu/ism/2014/february/google-0210.html

University of Amsterdam*
LifeWatch– E-Science European Infrastructure for Biodiversity and Ecosystem Research
http://www.lifewatch.eu/web/guest/home

University of Oxford
Big data and drug discovery centre launched
http://www.ox.ac.uk/media/science_blog/130507.html

University of Pittsburgh Medical Center (UPMC)
UPMC'S Big Data Technology Shows Promise in Breast Cancer Research
http://www.upmc.com/media/newsreleases/2013/pages/upmc-big-data-tech-breast-cancer-research.aspx

University of Washington
UW, PNNL tackle big data with joint computing institute
https://www.pnnl.gov/news/release.aspx?id=964

Virginia Tech*
World Population Scale Epidemiological Study
http://bigdatawg.nist.gov/_uploadfiles/M0172_v1_8972697421.docx

Washington University*
Large-Scale Genomic Discovery
http://projectreporter.nih.gov/project_info_description.cfm?icde=0&aid=8583329

Yale School of Medicine
Using "Big Data" to Scrutinize New Cancer Therapies
http://medicine.yale.edu/cancer/research/programs/prevention/index.aspx

Yale School of Medicine*
Yale University Open Data Access (YODA) Project
http://medicine.yale.edu/core/projects/yodap/index.aspx

Appendix 5:

Relevant Laws, Regulations, and Guidance

Laws

Biosecurity

Agriculture Reform, Food, and Jobs Act of 2013
Andean Trade Preference Act, 1991
Antiterrorism and Effective Death Penalty Act of 1996
Biological Weapons Anti-Terrorism Act of 1989Department of Defense and Full-Year Continuing Appropriations Act (2011)
Farm Security and Rural Investment Act of 2002
Food, Conservation, and Energy Act of 2008
National Defense Authorization Act, 2011
Pandemic and All-Hazards Preparedness Act (2006)
Public Health Security and Bioterrorism Preparedness and Response Act of 2002

Cybersecurity

Cyber Security Research and Development Act (2002)
E-Government Act of 2002
Energy Policy Act of 2005
Ike Skelton National Defense Authorization Act (2011)
Intelligence Authorization Act for FY 2010
National Defense Authorization Act for Fiscal Year 2012
***U**niting and **S**trengthening **A**merica by **P**roviding **A**ppropriate **T**ools **R**equired to **I**ntercept and **O**bstruct **T**errorism Act of 2001* (USA PATRIOT ACT)

Biological and Cybersecurity

Homeland Security Act of 2002

Other Relevant Laws

Health Insurance Portability and Accountability Act (1996)
Genetic Information Nondiscrimination Act (2008)
Privacy Act of 1974

Regulations

Biosecurity

7 CFR Part 331: Possession, Use, and Transfer of Select Agents and Toxins
9 CFR Part 121: Possession, Use, & Transfer of Select Agents & Toxins (Animals & Animal Products)
42 CFR Part 73: Select Agents and Toxins
18 USC §175: Prohibitions with respect to biological weapons

Cybersecurity

Gramm-Leach-Bliley Act (GLBA)
Federal Information Security Management Act
National Institute of Standards and Technology Internal Report (NISTIR) 7628: Guidelines for Smart Grid Cybersecurity
State Security Breach Acts

Guidelines or Strategies

Biosecurity

Centers for Disease Control and Prevention (CDC) Guidance on Applicability of the Select Agent Regulations to Issues of Synthetic Genomics (2010)
Centers for Disease Control and Prevention (CDC) Select Agent Programs (1996)
Department of Homeland Security, National Biodefense Architecture
Executive Order 13486: Strengthening Laboratory Biosecurity in the US (2009)
Executive Order 13527: Medical Countermeasures Following a Biological Attack (2010)
Executive Order 13456: Optimizing Security of Biological Select Agents and Toxins in US (2010)
Global Health Security Agenda (2014)
Homeland Security Presidential Directive 9: Defense of US Agriculture and Food (2004)
Homeland Security Presidential Directive 10: Biodefense for the 21st Century (2004)
Homeland Security Presidential Directive 18: Med Countermeasures against WMD (2007)
Homeland Security Presidential Directive 21: Public Health and Medical Preparedness (2007)
National Biosurveillance Science and Technology Roadmap (2013)
National Security Presidential Directive 17: National Strategy to Combat WMD (2002)
National Strategy for Biosurveillance (2012)
National Strategy for Countering Biological Threats (2009)
Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA (2009)
United States Department of Agriculture (USDA) Agricultural Biosecurity: Prevention, Diagnosis, and Control Strategies for Pathogens and Pests (2002)

Cybersecurity

Comprehensive National Cybersecurity Initiative (2009)
Cybersecurity for Energy Delivery Systems Program
Cyberspace Policy Review (2009)
Executive Order 13636: Improving Critical Infrastructure Cybersecurity (2013)

Homeland Security Presidential Directive 7: Critical Infrastructure Identification, Prioritization, and Protection (2004)
National Initiative for Cybersecurity Education
National Institute of Standards and Technology (NIST) Cybersecurity Framework (2014)
National Strategy to Secure Cyberspace
National Strategy for Trusted Identities in Cyberspace
Presidential Policy Directive 21: Critical Infrastructure Security and Resilience (2013)

International Governance

Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological Weapons Convention) (entered into force in 1975)
United Nations Security Council Resolution 1540 (2004)

Appendix 6:

AAAS-FBI Partnership

The Federal Bureau of Investigation (FBI) Weapons of Mass Destruction Directorate (WMDD) has developed a robust biosecurity outreach and awareness program with the scientific community. To strengthen this relationship, the FBI WMD Directorate contracted with the American Association for the Advancement of Science (AAAS) to host a series of outreach and policy meetings with research, policy, and security stakeholders and summarize important lessons learned, challenges faced, and areas for improvement of local and national biosecurity initiatives.

Bridging Science and Security for Biological Research

This project is done in collaboration with the Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU), AAAS, and the FBI WMD Directorate.

The first meeting, which was held in February 2012, provided opportunities for academic scientists and research administrators to build trust and enhance their relationship with the security community, with the mutual goal of jointly addressing the challenges of mitigating biosafety and biosecurity risks.

The second meeting, which was held in September 2012, provided the opportunity for scientists and research administrators to share best practices and lessons learned about the review and oversight of dual use life sciences research with each other and with the security and policy-making communities.

The third meeting, which was held in February 2013, focused on critical issues resulting from foreign scientists studying or working in the U.S., international collaboration, and U.S. scientists working in foreign countries.

The fourth meeting, which was held in April 2013, focused on the challenges faced during implementation of the revised Select Agents and Toxins Regulations and possible approaches for addressing those challenges.

The fifth meeting, which was held in August 2013, focused on sharing of information about the purpose and existing examples of personnel security programs.

FBI Biosecurity and Outreach Programs

The FBI contributes to the U.S. government's efforts to reduce the risk of bioterrorism by enforcing the federal statutes that prohibit development, production, or stockpiling of biological weapons. To accomplish these functions, the Biological Countermeasures Unit (BCU) of the FBI's WMD Directorate has developed biosecurity initiatives that focus on acquisition or exploitation of biological material, technology, and expertise to intentionally cause harm.

The BCU has established a successful biosecurity outreach program, the goal of which is to establish strong, sustainable relationships with officials and scientists from research institutions to prevent and mitigate potential threats that they might encounter. The primary way in which the FBI engages the scientific community is through its Academic Biosecurity Workshops. FBI WMD Coordinators conduct the workshops using a series of dialogues and exercises to bring relevant academic, health, first responder, law enforcement, and industry experts together to: 1) promote an understanding of their respective roles and responsibilities, capabilities, and resources; and 2) develop feasible, implementable threat mitigation strategies. The WMD Coordinators offer a point of contact at the local level and provide local support and security expertise. These efforts build on a shared goal of serving the public good.

The tangible benefits generated by these engagements are evident by the increasing number of requests for workshops by research institutions. In addition, this model has garnered international attention; requests for assistance to implement similar academic workshops have come from both the law enforcement and academic communities of foreign nations.