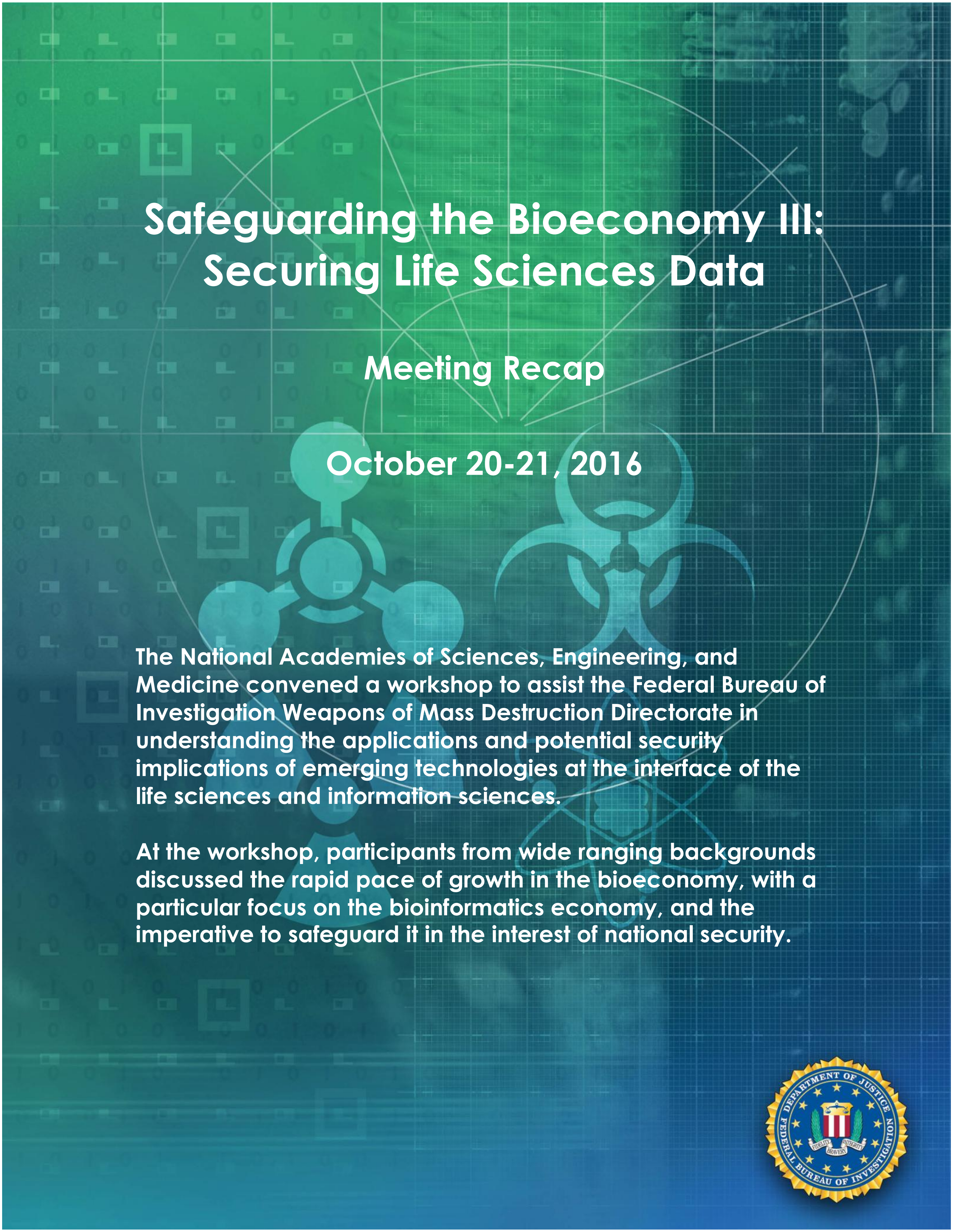

Safeguarding the Bioeconomy III: Securing Life Sciences Data

Meeting Recap

October 20-21, 2016

The National Academies of Sciences, Engineering, and Medicine convened a workshop to assist the Federal Bureau of Investigation Weapons of Mass Destruction Directorate in understanding the applications and potential security implications of emerging technologies at the interface of the life sciences and information sciences.

At the workshop, participants from wide ranging backgrounds discussed the rapid pace of growth in the bioeconomy, with a particular focus on the bioinformatics economy, and the imperative to safeguard it in the interest of national security.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE
BOARD ON LIFE SCIENCES *and* BOARD ON CHEMICAL SCIENCES AND TECHNOLOGY

Meeting Recap

Safeguarding the Bioeconomy III: Securing Life Sciences Data

October 20-21, 2016

Disclaimer: This meeting recap was prepared by staff of the National Academies of Sciences, Engineering, and Medicine ("the National Academies") as an informal record of issues that were discussed during the sessions of the National Academies' Workshop on Safeguarding the Bioeconomy III: Securing Life Sciences Data, held on October 20-21, 2016. This document was prepared for information purposes only and as a supplement to the meeting agenda. It has not been reviewed and should not be cited or quoted, as the views expressed do not necessarily reflect the views of the Academies or the Committee on Safeguarding the Bioeconomy III: Securing Life Sciences Data.

Points of Contact: Katherine Bowman (kbowman@nas.edu; 202-334-2638)
Andrea Hodgson (ahodgson@nas.edu; 202-334-3138)

INTRODUCTION

Advances in the life sciences are increasingly integrated with fields such as materials science, information technology, and nanotechnology to impact the global economy. Although not traditionally viewed as part of *bio*-technology, information technology and data science have become major components of the biological sciences as researchers move toward –omics experimental approaches. In the healthcare sector, the rapid pace of digitalization into electronic health records (EHR) and the increasing use of internet-connected devices for monitoring and collecting healthcare data, have created a heavily data-dependent ecosystem for making medical decisions and sharing clinical information across institutions. President Obama's Precision Medicine Initiative (PMI) aims to collect genomic and clinical data from 1 million volunteers and has outlined policies for data security and frameworks for participating institutions. However, there has been a marked increase in cyber intrusions in the healthcare sector resulting in loss of personal identifying information (PII) (e.g. Anthem). The more recent intrusions appear to have accessed clinical and medical data (e.g. Premera Blue Cross Blue Shield and UCLA Medical System). These events have raised concerns over the ability to protect medical data and ensure continuity of treatment.

In light of these advances in the bioeconomy, the Federal Bureau of Investigation Weapons of Mass Destruction Directorate, asked the Academies to convene a workshop to discuss how maintaining and supporting a strong bioeconomy can form part of the system of national security, and what particular security considerations the generation, aggregation, and use of biological and medical data may pose. The traditional biosecurity framework has focused particularly on monitoring select agents and toxins. Existing criminal statutes and export control policies, as well as norms under agreements such as the Biological Weapons Convention, largely focus on pathogenic agents and toxins or consider intellectual property (IP) concerns. Questions posed by the planning committee for discussion at the workshop included how bioeconomy data and data security fit into security frameworks and how stakeholder communities can move beyond privacy concerns when thinking about biological and biomedical data breaches and their implications. There is currently no government agency charged with holistically assessing the security of the bioeconomy, and the emerging importance of data (and data security) within it. These concerns will continue to grow as the world becomes more digitized and interconnected. There are a number of different types of data that can be aggregated and analyzed as part of the bioeconomy (see Figure 1), and the collection, sharing and use of these different types of data may pose different potential concerns. To serve as a case example, the workshop focused particularly on healthcare and biomedical data as these records have been the subject of the known data breaches and hacking events. However, during the workshop, participants touched on additional dimensions of life sciences data with the recognition that they would form critical components of a more comprehensive discussion of data security for the bioeconomy.

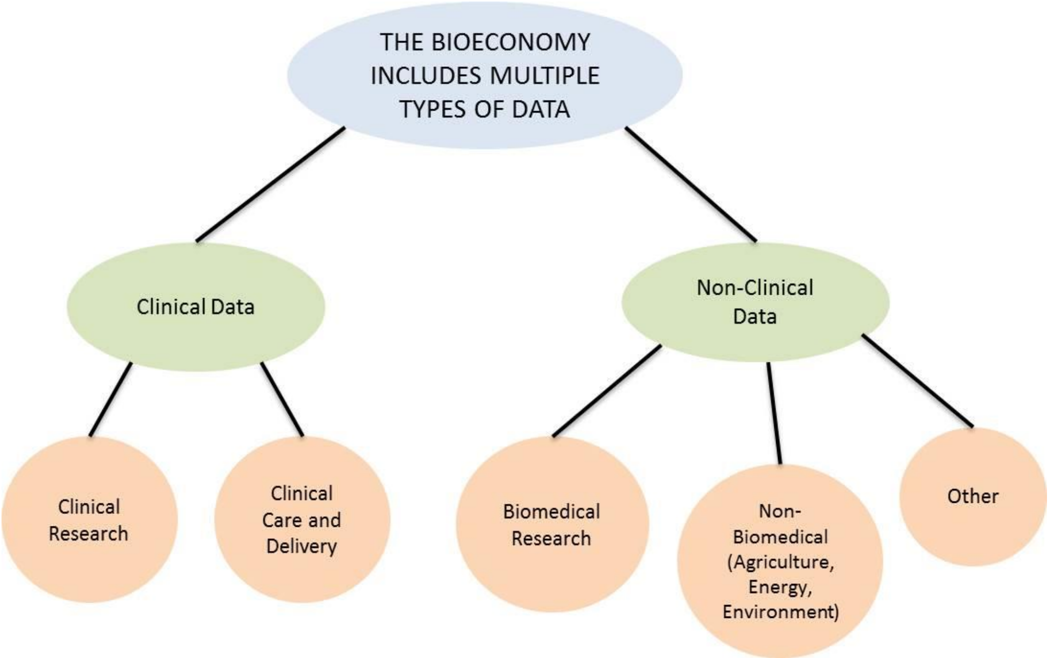
THE BIOECONOMY INCLUDES MULTIPLE TYPES OF DATA
Clinical Data
Non-Clinical Data
Clinical Research
Clinical Care and Delivery
Biomedical Research
Non-Biomedical (Agriculture, Energy, Environment)
Other

Figure 1. Diverse data types are generated, aggregated, and analyzed as part of the bioeconomy.

BIOSECURITY: PERSPECTIVE FROM ACADEMIC HEALTH

In his opening talk, Howard J. Federoff from UC Irvine Health described the rising attention that advances in the biological sciences are gaining in the press and among the public. Disease outbreaks such as Ebola and Zika have led the public to be concerned about the next infectious disease threat, and the security community is similarly worried about the potential of a biological attack. However, Federoff cited data demonstrating that the intelligence community is also worried about the ability of traditional intelligence gathering to provide sufficient actionable warning. He discussed the need to understand the risks of a novel disease outbreak and suggested the need to systematize biosecurity. In order to effectively do this, a rubric for quantifying the risks of a potential event and sharing that information across the relevant but diverse stakeholders will be required to better inform response planning and decision-making.

Dr. Federoff commented on the intersection of biosecurity and human pathobiology and the need to assess targets and mediators of pathobiology while noting similarities in the integrity of the food chain to suggest that these considerations go beyond health impacts. He began by describing informational molecules, such as DNA and RNA, as being both a target, because they are subject to mutagenic compounds and changes via genetic engineering, and a mediator with development of the CRISPR/Cas 9 technology to enable improved genetic engineering. He indicated that the fields of metabolism and chemistry are also benefiting from increased computational tools to better understand complex interactions between molecules, to improve the ability to design novel compounds, and to use computational toxicology to asses safety. Computational tools are also enabling integration of large datasets of genomic, phenotypic, and environmental data for hypothesis generation and for the development of novel therapeutics. Federoff discussed how technical advances in genetic engineering, data mining, and synthetic biology have led to an acceleration of the design, build, test cycle in the biological sciences, with an increased capacity to introduce novel qualities into organisms. These advances raise obvious traditional biosecurity concerns.

He also described factors that are of increasing importance for biosecurity, including the globalization of expertise in biotechnology and global distribution of biological manufacturing processes; advances in genomic sequencing and informatics technologies; and the integration of novel fields such as nanotechnology, information technology, and materials science with biotechnology. The academic sector fits into the biosecurity framework in several ways: information technology, engineering, and computational assets can be used for anticipating threats; research laboratories, scientists, and basic science researchers can assist to identify threats; and academic medical centers, doctors and clinical professionals can mitigate the effects of an attack. He suggested that a system of networked stakeholders in government, non-government organizations, the intelligence community, national labs, and the private sector could work together to systematize biosecurity efforts and noted that one agency alone may be insufficient. He concluded his presentation with the suggestion that a federally funded research and development center focusing on biosecurity is needed.

CURRENT DATA POLICY AND REGULATORY ISSUES IN THE BIOECONOMY

The session began with remarks from panelist Renata Rawlings-Goss from the South Big Data Hub, who described the transition occurring in the biological sciences as it moves from a data poor to data rich field and how this is affecting decisions in laboratories at research institutions, as well as how clinicians handle data. She defined Big Data (BD) as having more data than can be handled with tools currently available to the researcher. This definition gets to the relative nature of BD and provides a basis for understanding why laboratories and institutions may handle data differently. The integration of large data sets requires experience and access to large scale computing power. Rawlings-Goss stressed the need for a strategic framework for handling BD across institutions and government agencies, instead of case by case, while discussing her role in the Federal Big Data Research and Development Strategic Plan, released in May 2016. This plan was achieved through active discussion across government agencies to define priorities and areas for research and development. Rawlings-Goss noted that NSF funds 87% of computer science in the nation and has allocated funding to set up the Big Data Hubs programs (South, Midwest, West, and East) to focus on the specific regional application areas. For example, the South Hub covers 500 institutions with their region, which spans Delaware to Texas, and focuses on health disparities & analytics, coastal hazards, industrial big data, materials & manufacturing, and habitat planning.

William Crown from Optum Labs continued the discussion. Optum Labs has acquired 100 million lines of medical claims data and roughly 75 million EHR. Optum works with a number of research partners, including the University of California system, to facilitate the use of these data for scientific research. He described how the data have been statistically de-identified to comply with Health Insurance Portability and Accountability Act (HIPAA) and safe harbor regulations, with 18 points of identification dropped to protect privacy. As a security tool, Optum Labs utilizes a virtual sandbox approach in which users can access data and run analysis programs without the ability to hold or change the data itself. The system separates the data from the data source and prevents re-identification from outside sources. The combination of data richness and data analytics capabilities provides an opportunity to build predictive models. Crown noted, for example, an effort to create a clinically rich population for Alzheimer's research. Modeling is being used to identify predictors of risk and to create clusters that will enable clinicians to see variation in clinical outcomes in subgroups of patients. The long-term goal would be to develop something akin to a search engine recommender for EHR software to assist doctors in identifying clinical pathways for a particular patient.

Jodi Daniels from Crowell & Moring addressed the workshop on legal and regulatory challenges facing the bioeconomy regarding data, PII, privacy, and security. Daniels was a co-author of the HIPAA and discussed how health information technology (IT) began when electronic data sharing was required for processing payments for medical claims. Recently, there has been a shift toward taking advantage of EHR data in order to improve health outcomes. There has been a similar shift in policy from protecting data toward considering how data should be used and analyzed. However, the rise in data breaches has necessitated the need to consider balancing the benefits of sharing and use with adequate data protection. The PMI has been a driver for getting access to clinical data, yet there remain questions about how to obtain and manage this data while considering legal and operational issues. While large datasets are being collected, there is limited participation in data exchanges set up by the PMI and the Department of Health and Human Services (HHS). Among reasons for not sharing data are wanting to maintain control of the dataset, technical challenges of not being able to match patients adequately, and fear of violating privacy and security rules, which are deemed complicated. This has led to consumer-mediated exchanges where patients choose to share their own data, which they have the right to access. Providers have developed portals and applications on mobile devices that provide patients with access to their data, however, these raise privacy and security concerns for the health center managing these applications. The limited scope of HIPPA means it is only applicable to data collected, stored, and maintained by healthcare providers. These protections may not extend toward the use of remote monitors, mobile applications, or third party exchanges. This creates confusion for patients and a gap in protection, where the same data point is protected when under the supervision of the provider but not when the patient is accessing that same data point via an application. De-identification standards apply to data that is HIPPA protected, but questions remain on the ability to de-identify data collected and shared through mobile applications. Potential guidance and solutions have been developed, such as the Cybersecurity Information Sharing Act (CISA), encouraging companies to drive innovation in improving cybersecurity, and the Federal Trade Commission is beginning to use its authority for maintaining privacy and security of data falling outside of HIPPA. Daniels concluded her remarks by stating that IT security is a minimal budget item at many health organizations and usually done as a patch rather than by having a holistic approach. Proper cybersecurity hygiene is a confusing and difficult task for health organizations to accomplish with limited budgets, the steady stream of new guidance documents and regulations, and the rapid developments in the IT field.

The last panelist of the session was Kimberly Sutherland from LexisNexis, which is involved in helping clients such as health providers, retail pharmacies, urgent care clinics, and insurance companies prevent fraud. Their focus is on creating and maintaining methods for identifying individuals in order to grant access to their medical information. Mobilization of data access points has been a major challenge for safety, security, and convenience. Sutherland noted the need to improve the customer experience while instituting and maintaining appropriate safeguards. These practices are necessary when patients move between health systems and need to compile their medical history or when records are accessed on behalf of a patient by a proxy. Patient portals have streamlined these processes to some degree but there remain questions surrounding the security of these portals. When needing state records such as birth certificates, Sutherland noted that some states have open records allowing anyone to request them, while others have a number of additional requirements to access such records. These differences create confusion for the patient and challenges for those needing to confirm identity. In closing, Sutherland raised several challenges for discussion:

- The rise of synthetic identities, where a component of real information is combined with fictitious data. The inclusion of fabricated identities in clinical trials could skew results.
- The shift from data-centric processes toward a device-centric process for proving identities on mobile applications.

- The increasing use of biometrics for accessing data. Traditionally fingerprints, voice recordings, and retina scans have been used but the use of DNA raises new issues.
- The issue of how to give identity credentials when a person may have no provable identity, as in the cases of refugees.
- The increase of services such as Ancestry.com or Cologuard®, where consumers directly send their genetic information without knowing how it will be decoupled from their PII and personal health information.

Messages Highlighted During Discussion

During the discussion that followed the panel, individual participants noted a number of issues:

- Data agility, malleable systems, and security practices
  - The ability to transiently aggregate (or disaggregate) data without the need to reengineer or redesign systems may provide increased security if the disaggregated dataset is less valuable or accessible for bad actors.
  - More data collectors could make their data available via virtual sandboxes or use homomorphic encryption systems that would allow the use and analysis of datasets without sharing or directly accessing the data source.
  - Consistent policies and best practices regarding data security, such as tangible operational assistance to protect health data infrastructure (across institutions and agencies) from a security perspective would be helpful. A participant suggested that such practices could close uneven data protections but would require a multi-stakeholder group to develop holistic options that work for the very different sectors within the bioeconomy (healthcare, -omics, agriculture, energy, etc.). Another participant followed with the suggestion to think of the bioeconomy as more than healthcare related data, and to consider how data breaches in other life sciences sectors could influence the economy and the US standing internationally.
- Patient access and informed consent
  - Ownership of healthcare data. In the US, medical records are the business records of providers and therefore owned by providers. Patients have a right to access their medical records but do not own that information. The participants raised many questions surrounding the ownership of genomic sequence information, such as: Is it personal information? Can you adequately obtain informed consent for potential future uses of such information? Can mechanisms be created such that a patient can put a limit on the use or expiration date on their data to maintain control of their information? Can blockchain tracking be created and applied to healthcare data to allow patients to track their data, or conversely to allow researchers to identify patients to obtain informed consent for a novel use of their data? How will this affect research progress?
  - Providing patients control over the utilization and monetization of their data. Granting a consumer control over his or her data will allow them to opt into (or out of) data aggregation studies, however it may have the unintended consequence of limiting available data pools. Participants also discussed questions regarding how people value their own data and the role of the government in setting a price on that data. The value of an individual data point is difficult to assess when the value and innovation resulting from Big Data depends little on the individual points.
  - Another participant stated that the landscape and consumer attitudes about data sharing are changing rapidly. A reference was made to a recent study released by LexisNexis Research Solutions that surveyed 18-24 and 25-34 year olds in 7 countries and revealed that millennials are less open to sharing despite their connectivity with mobile devices and social networking platforms.

FUTURE IMPLICATIONS OF DATA GENERATION AND ACCESS

Manolis Kellis, professor of computer science at Massachusetts Institute of Technology (MIT), initiated the discussion by reviewing methods used by his group to identify a mechanism of obesity as an example of how aggregated EHR data can be used for research purposes. Through the use of genome wide associations and EHR, his group identified variations within a region of a gene linked to obesity, albeit through an unknown mechanism, and determined that the genetic variation was acting as a switch for controlling gene expression within obese cells. By programming changes at this locus to match those of lean cells, his group was able to turn cells from obese patients into lean cells. Mouse models of the genetic variants provided further evidence. Kellis indicated that there are many genes linked to disease states but for which researchers have yet to identify the mechanism of action. He suggested that genome wide association studies and phenotypic (EHR) data be further linked to better understand the molecular basis of disease. He concluded with a discussion on academic hoarding of research data, hindering progress. There is a common interest in using data for discovery and having access to large datasets to accelerate the research and development process. He suggested that patients be given ownership of their data to promote their inclusion in the research efforts that, he posited, would lead to more patients choosing to share their EHR.

The next panelist, Todd Peterson from Synthetic Genomics, Inc., began his remarks by commenting on the similarities of the network analysis on metabolic switches employed by Kellis and efforts Synthetic Genomics has utilized for engineering algae to maximize their biofuel production. Peterson discussed use of big data at Synthetic Genomics and how that information is leveraged through a software platform called Archetype. This platform, while still being optimized, uses machine learning and text mining to provide researchers with a structured data source that is flexible and easier to use. This resource, coupled with the decreasing cost of DNA synthesis and the increased ease of genetic engineering capabilities, has the capacity to greatly increase the speed at which developers can go through the synthetic biology design, build, and test cycle. Peterson discussed the concept of moving from genome browsing capabilities, towards gaining expression level data, and eventually toward identifying the enzymes involved in the underlying biochemistry for the creation of a "parts library for modular genomes." Synthetic Genomics is currently applying these concepts in the design of a minimal organism. The approaches are representative of the merger between digital information and functional biology. Bench top DNA sequencers and microfluidics systems (termed Digital to Biological Converters) have enabled a design existing in a digital format to be created and turned into a physical sample. Peterson went on to explain Synthetic Genomics Inc.'s membership in the International Gene Synthesis Consortium (IGSC), which screens costumers and sequence orders against toxins and select agents for biosecurity purposes to prevent the creation of potentially harmful components. He concluded his remarks by stating that big data is at the center of bio-design and relies on the associations and correlations generated by structuring and un-structuring datasets, the collection and use of metadata, and the accessibility of high performance computing.

The discussion on the design and promise of synthetic biology was continued by Ben Gordon from MIT's Broad Institute, whose lab is screening the human gut microbiome for metabolites that can be potentially utilized as novel therapeutics. The ability to use engineered organisms to produce chemical compounds and materials that are currently beyond the capabilities of chemists and manufacturers is an exciting area. Gordon discussed examples such as smart sensing systems in yogurts to sense and regulate insulin, smart sentinel crop plants, and living materials. Thus far, there has been success in engineering increasingly larger number of genes (10-20 genes) within one organism, but the field is still developing better methods for increasing complexity. Advances in this area will enable the movement of entire biochemical pathways into organisms, such as engineering the nitrogen fixation gene clusters from bacteria into plants to reduce the need to fertilize crops. Gordon shifted his discussion to his research efforts mining the human gut microbiome for small molecule discovery and work on the retro-biosynthesis of molecules. He expressed concern over the idea of shifting from the use of public databases, which is their data source, toward specific patient sources of data, as others had suggested over the course of the meeting. His concern was over ownership of the discoveries stemming from the use of such data sets. He warned that more discussion and clarity is needed before shifting toward patient owned data. These concerns over ownership grow when considering obtaining sequences and enzymes from animals and when moving from the research lab toward pharmaceutical development. He also raised concerns regarding the ability of current biosecurity approaches to detect genetic systems that could produce biosimilar compounds to those that are on the schedules for control.

The last panelist in the session, David Martin from the US Food & Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER), provided an overview of the FDA's drug approval process. This includes standard phase 1, 2, and 3 clinical trials with post-approval safety monitoring and reliance on an opt-in system for reporting adverse effects via the MedWatch program. The major obstacles encountered with this system are underreporting of adverse events and lack of a clear population for analysis, because the system considers the entire US population. These issues have caused FDA to re-evaluate how it examines and collects population based data. During the 2009 Influenza pandemic, there was a need to obtain population based and infection rate based data to monitor vaccine distribution and adverse events. This change in data collection set the stage for the development of the electronic Sentinel System to allow safety monitoring of approved medical products. This created an industrialized approach to bio-surveillance and phenotypic surveillance by analyzing data from 120 million health-plan members with distinct patient identities for sharing and use within a distributed environment. The single ID and coding used provide researchers with a standardized dataset, in a common format. This enables researchers to perform rapid queries and analysis while increasing the speed of evidence generation. Martin went on to explain the need for large datasets when conducting surveillance, followed by an ability to refine specific populations of interest. Under the Sentinel System, claims data are required to opt in for payment purposes, ensuring the ability to follow an individual longitudinally, even when a patient is seeing multiple providers.

Messages Highlighted During Discussion

During the discussion that followed the panel, individual participants noted a number of issues:

- True privacy versus the illusion of privacy
    - More data sharing solutions, such as the sandbox environment described by Crown and the distributed environment described by Martin would allow researchers to conduct data analyses without needing to obtain the actual data. There was the suggestion that these systems (and others like them) continue to be developed and tested for security purposes. The Optum Labs system required the inclusion of patients' data for medical claims reasons, whereas the distributed environment used by the FDA sources data voluntarily when patients choose to opt into the system.
    - Individuals differ in their approach to sharing their genomic data when there is a choice to do so. The contrasting examples discussed included the all in, unless opt out, methods used in the Finland, which have resulted in large scale participation by individuals, and the mandatory inclusion used in China, which has resulted in concerns that actions could potentially be taken against the individual due to their genomic data.
    - From the academic and industry perspective there is no need to identify a specific subject beyond genotype and phenotype for discovery purposes. Given the amount of DNA individuals shed every day, there is no real genetic de-identification, particularly as reconstructing facial features and physical appearance from genomic data is improving.
    - Having de-identified clinical data is useful such as for examining treatment regiments, in order to determine optimized treatment schedules and the dosing in order to elicit a particular response.

- More data versus structured (more usable) data
    - One participant asked whether researchers needed more data or needed higher quality, more structured datasets. This was met with conflicting responses, though there was agreement that such a distinction would depend on the purposes of the study. For example, larger datasets are needed for the discovery of rare events within a population, or for deeper knowledge of the variation within the human gut microbiome. This was countered with the possibility of learning from more detailed structured data from EHR of a smaller pool of patients because they could provide more context for the observations.
    - The rise in the availability of EHR has led to a desire for more data, however many records contain unstructured notes that will require further curation to be functional. Another participant stated that potential errors during data entry could skew analyses to add support to the suggestion of needing more EHR data to mitigate these effects.
- Open source datasets and international data sharing
    - Increasing access to large datasets through open source databases where researchers can publish will allow more researchers to share their datasets. Others raised concerns over the security implications of widespread use of the open source systems, and the extent to which the US would receive reciprocity for funding the projects that generate such data and having their researchers make their data publicly available for others to utilize.
    - International data sharing can be beneficial to the US. For example, participants discussed the benefit of being able to access health records from other countries' universal healthcare systems. A participant raised the concern that the US will fall behind in its ability to examine population data and link genotypic and phenotypic data for understanding disease mechanisms without increasing their use of EHR.
    - Open source synthetic biology toolkits raises security concerns regarding the creation of potentially harmful entities. Another participant countered this point by citing that instructions for creating dangerous substances are freely available on the internet and accessing such information is not illegal. This highlighted the importance of surveillance systems as a security strategy, rather than limiting the sharing of information that has the potential to do societal good.

TECHNOLOGY ADVANCES THAT WILL FURTHER DATA GENERATION AND ACCESS

This session began with remarks by Scott Hunicke-Smith of Sonic Reference Laboratory who provided the perspective of the diagnostic industry and the development of novel diagnostic tools that further data generation. He explained budget pressures on diagnostic labs as compared to test manufacturers, which has implications for supporting the costs of data security as well as additional challenges for diagnostic laboratories because the data provided by these tests is maintained by the labs and not the manufactures. Hunicke-Smith also commented that patients contribute to the call for more testing in order to know more about their conditions. This is creating more data than doctors are capable of processing and entering properly into electronic systems and results in dumping the data as unstructured data into EHR. He raised these points to highlight that more data is not always better. In the diagnostic space, the focus is on minimizing testing to obtain targeted data for specific questions. Hunicke-Smith concluded his remarks with a brief discussion on how gaining access to one's genomic information impacts more than just the individual, but also all those related to the individual. He suggested that consumers be better informed of the potential consequences of having their genomes sequenced through services such as 23andMe.

The next panelist, Karim Dabbagh from Second Genome, spoke about technological advances in the microbiome sector and the potential of large datasets to provide actionable information. He described the process by which Second Genome analyzes data from publically available datasets to determine the bacterial species present in healthy and diseased individuals and to identify bioactive molecules as potential drug candidates. Thus far, Second Genome has screened 70,000 samples from 340 microbiome studies sourced from public databases, such as GreenGeens, or from published work by utilizing their proprietary bioinformatics infrastructure. This analysis platform narrows the species of interest by examining overlap within the microbiomes from different studies of a similar disease state and examines known protein coding sequences within those genomes. The list of protein coding sequences is refined and then identified bioactive molecules are screened using *in vitro* tests before successful candidates move toward further preclinical development.

Reshma Shetty from Gingko BioWorks spoke on the rise of synthetic biology and how proponents envision using biotechnology for manufacturing and production. She reviewed the field, stating that for the first 5-10 years the focus was on development and democratization of access to the tools of biotechnology. There was a rapid rise of new players, from do it yourself biology (DYI bio) spaces to bio-hackers. Shetty stated that the increase in these entities bolstered the bioeconomy and the field simultaneously engaged with the public and the security community to discuss potential concerns. For example, the International Genetically Engineered Machine (iGEM) competitions include presentations on responsible science. Shetty discussed her observations that the field is moving toward the creation of synthetic biology foundries and away from the traditional academic laboratories. She suggested that the creation of foundries will result in a concentration of resources, data, technology, and expertise. This could make enforcing policy, regulations, and security easier for agencies charged with monitoring these spaces, however a concentrated infrastructure could also make foundries targets for disruption from bad actors (either cyber or physical) that could result in interruption of the supply chain of products derived from these technologies (such as biofuels, medicines, chemical compounds used for flavors, and perfumes).

The final panelist Jeremy Freeman from Janelia Research Campus, discussed the rise of big data in neurobiology. The advances occurring in this field are not yet ready for use in humans but the datasets generated are rapidly evolving the understanding of neuronal connections. Freeman described methods for measuring and manipulating brain activity as falling into two categories that differ in how invasive they are and what they can accomplish. Electrical methods require the use of wire electrodes inside the brain to directly record neuronal activity. They provide the highest temporal precision along with the ability to record individual neurons, but are unable to provide a comprehensive picture of what is occurring. It is also possible to manipulate brain activity by sending electrical signals through these electrodes. Whereas optical methods enable researchers to measure brain activity within the entire brain. These methods use genetically engineered mice that express fluorescent molecules when neurons are active, and these signals can be monitored less invasively through the skull. It is also possible to activate neuronal signaling by engineering optical responses into neurons and controlling activation through exposure to specific light sources. The technologies for examining brain activity in humans are of lower resolution and indirect, except for the rare cases in which patients are undergoing surgery. However, neurobiological methods are generating large datasets that require advanced data handling techniques for analysis. They are also being applied for high throughput screening to evaluate drugs for neuronal modulating properties.

Messages Highlighted During Discussion

As with the previous sessions, a rich discussion followed the presentations and individual participants raised the following points:

- Biosecurity concerns of accessing or reverse engineering methods for novel bioactive molecules discovery

    - There is potential for bad actors to make use of processes for identifying or creating bioactive molecules in order to create toxins or to induce a disease state. Other participants raised concerns regarding security practices for conducting this type of research and the responsibility of scientists and companies to ensure they are not enabling potential negative impacts on society and the bioeconomy.
- Data sharing capabilities are limited by funding in academic laboratories
    - Some neuronal datasets aggregated at academic institutions remain private due to the inability of some researchers to access advanced computing power and lacking the expertise to properly analyze and construct a shareable database. The typical academic lab may not have adequate resources to make this data publically available.
- Regulatory uncertainty affecting development
    - Traditional paths for identifying compounds for drug development are still the focus of biotechnology companies despite having the technology to create organisms (or microbial consortia) that could produce and deliver novel therapeutics *in situ*. This point stimulated active discussion about the reluctance to invest in the research and development of such products when it is uncertain how they will be reviewed, approved, and received by regulatory authorities.
- The uniqueness of life sciences data has inherent security risks
    - The question of what is inherently different (if anything) about life sciences data that makes protecting these datasets challenging compared to other types of data was posed to the group by a participant. Life sciences data can be translated from digital information into a physical and self-reproducing entity, and therefore has the capacity to translate a cyber threat into a physical threat. There are many mechanisms used in other industries that could be applied toward protecting life sciences data; however, the ability to quantify the risk of dissemination is more difficult for life sciences data. Quantifying the economic risk from IP loss or loss of competitive advantage internationally, or the potential for a bad actor to create or obtain a toxic agent, is difficult to accomplish.

DATA SOVEREIGNTY ISSUES

Nathan Hillson from the Joint BioEnergy Institute (JBEI) provided commentary on data sovereignty and data sharing to stimulate a discussion on these issues. Hillson reviewed how JBEI, as a research institution that publishes in peer reviewed journals, balances the mandate to share data from studies that receive public funding with the ability to leverage that data. Generally, all data is kept in a private repository and each lab has the ability to maintain control over who accesses its data, but upon publishing, these datasets are pushed to a public repository maintained by the institution. The repositories containing unpublished data can be linked across different labs and different institutions in semi-private repositories where there is strict control over who can access the data. Hillson noted that similar systems are used in the UK, and elsewhere, with the example of the Synthetic Biology Hub at Imperial College in London. Hillson went on to discuss the creation of Agile Biofoundries, which is seeking to develop enabling infrastructure for genetic engineering and production of synthetic biology products. The Department of Energy is trying to leverage their distributed infrastructure in this space to create a system to match the distribution of feedstocks. Contrary to remarks made previously in the meeting, Hillson predicts that there will not be the creation of concentrated spaces and consolidated foundries. He concluded by suggesting that the security challenges in the bioeconomy are not entirely unique but that there may be a greater difficulty of detecting when tampering of life sciences data has occurred, but warned that the adverse outcomes of not developing this sector is far riskier for national security.

Messages Highlighted During Discussion

A discussion followed Hillson's remarks where individual participants noted a number of issues:

- Self defending data and other security mechanisms
    - The creation of data (or technology that enables data) to defend itself from tampering could serve as an alert system for hacking events, enable a blockchain mechanism for following and controlling access to data (as discussed above), or allow data to delete itself when inappropriately accessed.
- A lack of clear guidelines on IT security
    - A participant asked why the government could not mandate compliance with existing data security guideline used by other industries to protect all life sciences data. This was followed by a discussion on how some of these conventions may not be the correct fit for all types of life sciences data and sectors. The need to share data and make data available is very different in other industries. This is also because the risk of operational disruption (in the case of hospitals for example) or the risk and implication of life sciences data theft are hard to evaluate.
    - The skills, expertise, and budget needed to implement these safeguards and maintain good cyber hygiene is also posing a barrier for many of the generators of data within the life sciences. Academic laboratories at research institutions are usually left to follow institutional practices, if they exist. It was noted by a participant that there is a huge opportunity to link the life sciences community with the cybersecurity community to implement security efforts that accommodate the unique aspects of the life sciences. There are currently not enough people working at the intersection of these domains.
    - National Institute of Standards and Technology (NIST) could set standards specific for the life sciences, as they successfully created the cybersecurity framework used in other sectors by collaborating with industry and government agencies. One could envision a specific set of guidance for dealing with cybersecurity in the bioeconomy, which could be developed following more detailed study of the data infrastructure and needs, similar to what was accomplished for security for the power grid or financial institutions. However, the implementation and maintenance of such standards could be economically challenging for smaller research entities and start-up biotechnology companies, without guaranteeing protection. Anthem Health, considered the gold standard for implementing data protection standards, was still the victim of a sophisticated attack.
- Motivations and concerns for sharing data- domestically and internationally
    - Competition is stronger domestically between research institutions than it is internationally. A participant speculated that this is due to domestic institutions competing for the same pool of government grants to fund research. Research is a global enterprise but the funding mechanisms are largely domestic. This situation has led to some researchers sharing their platforms but withholding their datasets. Within the current system, economic potential is considered from an individual research entity perspective, but not at the national or international level.
    - Government funded research is mandated to be shared in the public domain. Were the bioeconomy incorporated into the critical infrastructures considered vital by the Department of Homeland Security, a participant noted that the data it generated would likely be subject to stricter regulations and export controls, which might provide security benefits. However, it was also noted that an application of export control regimes to fundamental scientific research is complicated and that information in the public domain cannot generally be subject to such controls.
    - Data security and sharing policies have the potential to slow the pace of research and could hinder economic development if not properly balanced. Some participants warned about the

potential to replicate the 1996 debate over encryption technology when the security community reacted by implementing a lockdown to prevent an acceleration of competition overseas.

- The value of data or data analytics methods
    - It is often difficult to determine which research line is going to lead to valuable IP and economic development. Therefore, it may be worth exploring moving toward an entirely open data environment where the algorithms and tools used to analyze the data are maintained privately. This concept of leveling the data "playing field" was met with concerns that international actors could gain an economic advantage due to having a more technologically and computationally savvy workforce.
    - Biological data is inherently open to those with the tools to read it, such as through gene sequencing, and the real value lies in the analytic capacity to make the data useful. This was countered with the suggestion that the ability to develop analytic mechanisms that are generating "correct and useful" information is dependent on data science and on having a strong dataset available to create analysis models.

CYBERSECURITY OF LIFE SCIENCES DATA

Charles Brooks of Sutherland Global Services opened the session on cybersecurity practices and policies. Brooks began noting that the current major threats that are difficult to monitor are hacktivists that are capable but not formally trained and the threat of disgruntled employees or insiders leaking information. He noted that cybersecurity is a continually growing and emerging field and there has been a rapid change in interconnectivity and mobilization of data access points since 2003. These changes necessitate a clearly defined security strategy for handling this constantly evolving landscape. He went on to describe the styles of cyber threats (from phishing scams, bots, and ransomware, to software holes that leave vulnerabilities) with particular attention to the rise of phishing scams and ransomware used to target healthcare data. He noted that there has been a 3,500% increase in the use of ransomware attacks leading to the payment of over 1.6 million dollars, yet these figures are potentially underreported as there is no mandate for private companies to disclose breaches to their consumers within the bioeconomy as there is in other sectors. Brooks discussed how the life science community is particularly vulnerable to cyber attacks because much of the data generation and analysis tools require digitalization and use of devices for which there are not yet adequate standards of protection. Edge security deals with the protection of devices that are used as entry points for accessing data, such as DNA sequencers, medical devices connected to smart phones, and biowearables. He referenced a 2012 report from MIT that examined the rampant rate of computer viruses detected on medical devices within hospitals. At the institutional level, systems that control refrigeration of physical specimens are typically controlled by a network that can be subject to attack. In order to develop effective strategies unique to the life sciences industry, Brooks suggested that stakeholders from all levels be included in the discussion and development of policies that fit the needs of the research community and industry. He continued that investments are needed in life sciences data security, as has been done in other sectors, in order to shift from reactive to preventive measures, but stakeholders will need to be convinced of the risks and vulnerabilities they face prior to investing in such infrastructures. He noted that the energy sector could serve as a model in which relationships between security personnel, government agencies, and industry have led to increased security of this critical infrastructure. He concluded his remarks by discussing advances occurring with the cybersecurity realm including better encryption and biometrics, self-encrypting drives, and rapid growth in artificial intelligence and machine learning capabilities and how these could lead to protection strategies that are better suited to the life sciences than those currently available.

Messages Highlighted During Discussion
Individual participants raised a number of points during the discussion, including:

- Identifying priorities and strategies for implementing cybersecurity safeguards
    - Lack of investment in cybersecurity reflects a lack of understanding of the threats by company leadership. A participant noted the need to incentivize companies to improve their cybersecurity infrastructure and could include reviewing their liability when considering data dissemination effects. Insurance plans exist for covering the economic cost of financial loss incurred by data breaches, but these plans have no way of covering the economic cost of losing IP. This stems from the inability to adequately assess and predict the value of how data could lead to profitable IP in the future.
    - Data breaches can have an effect on the reputation of a company and this concern could be leveraged to incentivize the deployment of cybersecurity safeguards. One participant noted that security needs to be everyone's concern and will require all actors within the life sciences to take responsibility for securing data.
    - A proposal is circulating in congress that would require all publically traded companies to have a board member with a cybersecurity background, but this proposal is not gaining traction due to the burden this will place on small businesses and start-up companies. Another participant raised the point that there are many within the life sciences sectors (particularly the health sector) that are aware of the risks, yet are faced with prioritizing a number of competing goals. It could be difficult for some to spend on improving their cybersecurity framework without additional financial support to implement appropriate security safeguards and a life sciences security infrastructure.
- New mechanisms to make cybersecurity more accessible
    - Cybersecurity management could be outsourced to security companies to reduce costs for smaller size budgets and entities. Another participant noted that a working group between the FBI and DHS to discuss biosecurity, cyber security, and develop a framework of guidelines could lead to new solutions to increase the accessibility of cybersecurity solutions.
    - There is uniqueness to life sciences data because biology has the capacity to begin in the digital realm, become a physical entity, and return to the digital realm. This paradigm has been explored in previous workshops (for example, the recap of Safeguarding the Bioeconomy: Applications and Implications of Emerging Science, discussed a hypothetical case example involving the smuggling of agricultural seeds in which the IP encoded in the seeds can be read by sequencing machines).

INTERNATIONAL IMPLICATIONS: EFFECTS OF LOCATION AND PACE OF INNOVATION

The last speaker to address the workshop, James Schroeder from Bioinnovation Legal, discussed intellectual property sharing and began by stating that much of the "loss of IP" that occurs internationally is due to institutions failing to apply for patents. He noted that patents have geographic boundaries to their jurisdiction, i.e., tied to the country in which a patent is filed. In order to protect IP internationally, product developers need to apply for patents in each of the markets they intend to enter. This was demonstrated by the failure of major telecommunications entities to patent their inventions in China, leading to a dedication of IP to the public domain allowing for the rise of domestic competitors to the iPhone. It was suggested that for IP that has already been disseminated, there is still the opportunity for developers to monetize their investments by establishing research collaborations on improvements to their inventions with potential partners abroad. Mr. Schroeder noted that this line of thinking is contrary to some of the points raised previously during the meeting but deserves further exploration in an ever-interconnected global market. Countries like China are incorporating the promise of the bioeconomy into their national economic and security strategies and this is driving more investment and development in the life sciences. The speaker speculated that China's investment in their bioeconomy is due to a desire to provide better quality of life for citizens, when faced with limited resources and sustainability concerns. He concluded by suggesting that the world bioeconomy is made stronger through collaboration and cautioned against a desire to restrict data sharing and international collaboration.

Messages Highlighted During Discussion
Individual participants raised a number of points during the discussion, including:

- Balancing the desire to collaborate internationally while ensuring return on investment
  - The respect of patents and the ability of enforcement agencies to require compliance with patent laws in international markets were discussed. A participant suggested that as countries make major technological advancements and develop domestic life sciences and pharmaceutical industries, there is more incentive to enforce the protections provided by patent laws in order to monetize their investment.
  - One participant stated that research and corporate collaborative efforts should be approached with concern when dealing with international governments, given the evidence of state actors involved in data breaches.
  - China maintains tight control on the export of their citizens' genetic information and biological samples, but the US has no such restrictions, leading to an asymmetry in data sharing. One participant mentioned, for example, Chinese owned companies are sequencing U.S. clinical samples, and there have been recent investments by Chinese companies into American companies, such as 23andMe. This discussion echoed previous comments regarding the ability of the US researchers to access UK databases and EHR, prompting further discussion on the intended or unintended consequences of limiting the ability to share data and samples.
  - There also exist valuable examples for international collaboration within biotechnology. The FBI capitalized on the ability to address iGEM participants and to engage students at the high school, undergraduate, and graduate level about conducting responsible science and the implication of their work, providing the US a leadership role in maintaining responsible stewardship of biotechnology. This discussion also raised concerns over losing access to the next generation of scientists and the potential for decreased participation from some countries due to restrictions on moving materials across borders.
  - Participants also discussed investments in education and training the future scientific workforce. The Organization for Economic Co-operation and Development has data demonstrating the rising number of foreign students in the American graduate education system, particularly students from China and South Korea. Data also shows an increasing number of these students returning to their countries of origin due to rising economic opportunities. One participant suggested incentivizing these highly trained and skilled workers to stay within the US in order to capitalize on investment made into their education. There was an active discussion on the slow progress in growing and developing the US workforce to fill these roles. There was a suggestion that more funding and attention to improve STEM education programs at every academic level will create the skilled workforce needed to drive the development of the bioeconomy and maintain US competitiveness. A participant suggested that a particular emphasis should be placed on coding and computer skills as the digitalization of the bioeconomy continues.

ROUND TABLE DISCUSSION FOR IDENTIFYING A PATH FORWARD

During this session, all participants were asked to consider the substance of the discussion over the course of the two-day workshop and identify areas that require further exploration and development.

Messages Highlighted During Discussion and Throughout the Workshop

- Defining, measuring, and safeguarding the bioeconomy

    - Policy makers and other stakeholders lack an understanding regarding the contribution of the bioeconomy to the overall economy and the potential economic risks posed by not ensuring appropriate safeguards, making the bioeconomy vulnerable to disruption. Some participants suggested government agencies could take stock of the bioeconomy and consider more systematically their capabilities to protect it from bad actors without limiting bioeconomic growth. This action could require novel risk-assessment practices and ways to place value on the potential creation of IP from life sciences and biomedical data. However, such measures will be difficult without knowing how to quantify the value this data generates for the bioeconomy. Some participants suggested that a holistic approach could provide strategic guidance and oversight to the biosciences community. As the bioeconomy continues to develop, some participants suggested that the bioeconomy be included as a critical infrastructure deserving protection and oversight by the Department of Homeland Security and other agencies. However, other participants countered that the implications of such a designation are not clear and deserve further exploration. Another participant noted that the bioeconomy touches on many aspects of the existing 16 critical infrastructures outline by Presidential Policy Directive 21. One security framework may not be applicable to all the sectors within the life sciences and the specific data aggregation, access, and security needs of each sector (clinical, non-clinical, agricultural, environmental, energy, etc.) should be explored in greater depth.
    - Additional discussion for protecting data during the invention, development, production, and use stages will likely reveal how security approaches should differ to protect data at each stage. Some participants suggested that these areas would benefit from further development and support to provide innovative security strategies and to consider how best to implement such safeguards for the different sectors within the bioeconomy.
    - Participants discussed the benefits of the expanded use and implementation of good cybersecurity hygiene. One participant noted that a streamlined communication process by which advances in cybersecurity are shared with stakeholders relying on these guidelines to protect company IP and protect the privacy of individuals. There was an active discussion on how to balance cybersecurity guidelines for protecting data while maintaining sufficient data access to promote continued innovation.
    - It was also suggested that strategies for securing the bioeconomy be developed by members of the bioeconomy stakeholder and research community rather than allowing others to impose standards that may not meet unique life sciences needs or could unduly stifle innovation and progress. The use of a multi-stakeholder working group to develop strategies for securing the bioeconomy was proposed and many participants emphasized the importance of active engagement of all members of the bioeconomy.
- Workforce education and development concerns were raised with an emphasis on training more big data scientists able to create, curate, and secure large datasets and develop analytical algorithms.
    - The creation of additional public-private partnerships in STEM education for teaching data science and increasing the analytical capacity of students was discussed. The NSF funded Big Data Hubs provides an example of how such programs can be implemented. These programs would also benefit from increasing their reach toward women and underrepresented minority groups, which are growing components of the workforce.
    - One participant suggested that security concerns move beyond data, citing the importance of the algorithms used to make sense of the data. The potential to create repositories of algorithm modules or components, analogous to “biobricks” used in synthetic biology, was raised. This would accelerate the ability of researchers to create the novel algorithms needed to analyze data and ultimately to generate economic returns by reducing the need to build entire

    algorithms from scratch. This jumped back to a previous discussion over which is more valuable, the data or the algorithm used to make sense of the data.
  - There are additional cybersecurity issues to be considered with software that was developed using a modular approach. If developers do not know the provenance of the code or code modules they are using, there may be vulnerabilities within the code that could pose issues during analysis.
- Two final issues were raised as topics for further assessment as part of the security considerations associated with the bioeconomy:
  - A participant suggested that there is a need to explore in greater detail the potential incentives and barriers, and positive and negative implications, of international collaborations at various stages of the research, development, and commercial ecosystem.
  - Another participant suggested that a systems view of the bioeconomy is needed to understand how changes in one sector, or changes to different parts within a sector (in healthcare, for example, from research and patient clinical care through regulatory process) may have unintentional consequences across the system.

APPENDIX A - WORKSHOP DESCRIPTION

Safeguarding the Bioeconomy III: Securing Life Sciences Data

Meeting Overview

Advances in the life sciences are increasingly integrating with fields not traditionally viewed as part of *bio*-technology, such as materials science, information technology, and nanotechnology to affect the global economy. The National Academies of Sciences, Engineering, and Medicine have convened this workshop to assist the Federal Bureau of Investigation in understanding the applications and potential security implications of emerging technologies at the interface of the life sciences and information sciences. At the workshop, participants from wide ranging backgrounds—including research and development, healthcare, information technology, synthetic biology, policy, security, and next generation sequencing— will discuss the rapid pace of growth in the bioeconomy, with a particular focus on the bioinformatics economy, as well as ways to safeguard it into the future. The workshop builds on two previous workshops in this series that have discussed areas of convergence in the life sciences research enterprise. The planning committee drafted the following questions for discussion:

- How do we ensure the continued growth of the U.S. bioeconomy in the existing security context?
- How do we move beyond privacy to redefine cybersecurity in the current big data/data analytics environment in the biological and biomedical sciences?
- How can we develop strategies to improve data management and security in the life sciences while implementing appropriate safeguards?
- How can we identify and encourage proactive measures that all stakeholders can develop and adopt to ensure the appropriate possession, use, and transfer of their data?

Workshop Sessions and Panels

Current Data Policy and Regulatory Issues in the Bioeconomy: This session will discuss the current state of the bioeconomy as it relates to policy and regulatory issues surrounding the generation and access of life sciences data.

Future Implications of Data Generation and Access: This session will examine emerging uses of data to improve the design/build/test cycle in biological engineering, to develop advanced therapeutics, and to evaluate safety, efficacy and effectiveness of medicines, and then discuss ways to balance data access needs to enable emerging applications with data security, ownership and privacy concerns.

Technology Advances that will Further Data Generation and Access: This session will discuss technological advancements that will further the generation and access of life sciences data with a focus on the microbiome, neurobiology, next generation sequencing and precision medicine.

These sessions will lead into larger discussions on data sovereignty issues, cyber security in the life sciences, and the international implications of location and pace of innovation.

APPENDIX B - WORKSHOP AGENDA

Safeguarding the Bioeconomy III: Securing Life Sciences Data

Meeting Agenda

San Francisco, CA

Thursday, October 20, 2016

8:00 AM Breakfast available

9:00 AM Welcome addresses and sponsor remarks

9:15 AM **Keynote address**
Howard Federoff, University of California, Irvine

10:15 AM Break

10:30 AM **Current Data Policy and Regulatory Issues in the Bioeconomy**
This session will discuss the current state of the bioeconomy as it relates to policy and regulatory issues surrounding the generation and access of life sciences data.

Speakers:
Jodi Daniel, Crowell & Moring
William Crown, Optum Labs
Renata Afi Rawlings-Goss, South Big Data Hub
Kimberly Sutherland, LexisNexis

12:00 PM Lunch break

1:00 PM **Future Implications of Data Generation and Access**
This session will examine emerging uses of data to improve the design/build/test cycle in biological engineering, to develop advanced therapeutics, and to evaluate safety, efficacy and effectiveness of medicines, and then discuss ways to balance data access needs to enable emerging applications with data security, ownership and privacy concerns.

Speakers:
Todd Peterson, Synthetic Genomics, Inc.
Ben Gordon, Broad Institute
Manolis Kellis, MIT Computer Science / CSAIL / Broad Institute
David Martin, US FDA Center for Biologics Evaluation and Research (CBER)

2:30 PM Break

3:00 PM **Technology Advances that will Further Data Generation and Access**
This session will discuss technological advancements that will further the generation and access of life sciences data with a focus on the microbiome, neurobiology, next generation sequencing and precision medicine.

Speakers:
Scott Hunicke-Smith, Sonic Reference Laboratory
Karim Dabbagh, Second Genome
Reshma Shetty, Gingko Bioworks
Jeremy Freeman, Janelia

4:30 PM **Closing remarks**

5:00 PM Adjourn

Friday, October 21, 2016

8:00 AM Breakfast available

9:00 AM **Welcome and opening remarks**
What are key implications and challenges that arose from Day 1 discussions?

9:30 AM **Data Sovereignty Issues**
This discussion will be centered on the economic implications of data sharing, how to balance intellectual property concerns, and conventions of exchanging data.
Speaker: Nathan Hillson, Joint BioEnergy Institute

10:45 AM Break

11:00 AM **Cyber Security of Life Sciences Data**
This discussion will be centered on how the nature of cybersecurity is changing as a result of the new types of data and data-driven applications in the bioeconomy.
Speaker: Charles Brooks, Sutherland Global Services

12:15 PM Lunch break

1:15 PM **International Implications: Effects of location and pace of innovation**
This discussion will be centered on how regulatory and policy issues can affect technological development, and ways to create a more supportive system to facilitate the gathering, use, and sharing of life sciences data.
Speaker: James Schroeder, Bioinnovation Legal

2:30 PM Break

2:45 PM **Identifying a Path Forward**
This session will be a roundtable discussion to identify actions that can be taken to safeguard life sciences data, best practices and responsible users, and to discuss the unique needs industry, government, and academic sectors face when collecting, handling and sharing data.

4:30 PM **Closing remarks**

4:45 PM Adjourn Meeting