THE FUTURE OF BIOLOGICAL BIG DATA

A Technical Report by the Applied Futures Lab

ASU

The Future of Biological Big Data

A Technical Report by the Applied Futures Lab

Nikhil Dave
April 30th, 2021
Arizona State University

Table of Contents

Preface

Participants

Dr. Walter H. Moos - Managing Director and Co-Founder, Pandect Bioventures; Chairman Emeritus, ShangPharma Innovation; Adjunct Professor, University of California San Francisco; Retired President, Biosciences, Stanford Research Institute
Dr. Jodi Cook - Board of Directors, Fennec Pharmaceuticals Inc.
Dr. Gillian Woollett - Senior Vice President, Avalere Health
Josh Massad - Ph.D Student, Applied Futures Lab at Arizona State University
Dr. Samantha Orchard - Associate Professor of Practice, University of Arizona
Dr. Justin Kiggins - Project Manager, Chan-Zuckerberg Initiative
Ny'Quenta Strong - Undergraduate Student, Applied Futures Lab at Arizona State University
Michael Woudenberg - Lockheed Martin, Senior Staff Operations Research
Danielle Beauchamp - Graduate student, Applied Futures Lab at Arizona State University
Dr. Sarah R. Carter - Principal, Science Policy Consulting LLC
Scott Borland - President and CEO and co-founder of Xoc Pharmaceuticals
Dr. Carol Linden - Retired, Former Principal Deputy Director, Biomedical Advanced Research and Development Authority
Leslie Ann Klein – Project Coordinator, Arizona State University Research Enterprise and Masters Student, Sandra Day O'Connor College of Law

Arizona State University Applied Futures Lab

The Applied Futures Lab at Arizona State University serves as the premier resource for strategic insight, teaching materials, and exceptional subject matter expertise on Threatcasting, envisioning possible threats ten years in the future. The lab provides a wide range of organizations and institutions actionable models to not only comprehend these possible futures but to a means to identify, track, disrupt, mitigate and recover from them as well. Its reports, programming and materials will bridge gaps, and prompt information exchange and learning across the military, academia, industrial, and governmental communities.

Executive Summary

In recent years, biological research and clinical healthcare has been disrupted by the ability to retrieve vast amounts of information pertaining to an organism's health and biological systems. From increasingly accessible wearables collecting realtime biometric data to cutting-edge high throughput biological sequencing methodologies providing snapshots of an organism's molecular profile, biological data is rapidly increasing in its prevalence. As more biological data continues to be harvested, artificial intelligence and machine learning are well positioned to aid in leveraging this big data for breakthrough scientific outcomes and revolutionized medical care.

The coming decade's intersection between biology and computational science will be ripe with opportunities to utilize biological big data to advance human health and mitigate disease. Standardization, aggregation and centralization of this biological data will be critical to drawing novel scientific insights that will lead to a more robust understanding of disease etiology and therapeutic avenues. Future development of cheaper, more accessible molecular sensing technology, in conjunction with the emergence of more precise wearables, will pave the road to a truly personalized and preventative healthcare system. However, with these vast opportunities come significant threats. As biological big data advances, privacy and security concerns may hinder society's adoption of these technologies and subsequently dampen the positive impacts this information can have on society. Moreover, the openness of biological data serves as a national security threat given that this data can be used to identify medical vulnerabilities in a population, highlighting the dual-use implications of biological big data.

Additional factors to be considered by academia, private industry, and defense include the ongoing relationship between science and society at-large, as well as the political and social dimensions surrounding the public's trust in science. Organizations that seek to contribute to the future of biological big data must also remain vigilant to equity, representation and bias in their data sets and data processing techniques. Finally, the positive impacts of biological big data lie on the foundation of responsible innovation, as these emerging technologies do not operate in standalone fashion but rather form a complex ecosystem.

Part One: Setting the Stage and Definitions

Background

The scientific discipline of biology began with a focus on the individual components of biological organisms, from genes to organs to behaviors at large. As the foundational understanding of biology developed, researchers uncovered the daunting complexity that ties these components together. Systems biology seeks to understand how complex biological systems function - as the National Institutes of Health describes it: *putting the components together to understand the bigger picture*. This approach heavily involves bioinformatics and computational biology, in part focused on using computing power to make sense of high volumes of biological information taken from genomics, proteomics, transcriptomics and other -omics studies at a variety of molecular levels. With the ability to retrieve big data from molecular systems comes the new and profound availability of other important data relevant to an individual's health, for example biometric data from wearable technology that provides insights into clinical measurements and lifestyle choices.

Of course, when it comes to understanding high volumes of data, there is no field of study more relevant than computational science. Defined as using computing capabilities to understand and solve complex problems, computational science has been leveraged for a myriad of applications since its inception from financial modeling to developing self driving cars. As computational power has increased over time, so has its ability to be applied to increasingly complex systems. Now propelled by emerging technologies in computational science throughout recent decades, particularly artificial intelligence (AI), biologists have remarkably powerful tools to continue exploring the complexity of biological systems. This exploration is not only at the level of individual biological components, but also at the systems level - looking at biological systems and their relationships to each other as well as to the environment.

Taken together, the intersecting fields of systems biology and computational science have roots in a diverse range of industries and application areas including pharmaceuticals, medicine and health, agriculture, biosecurity, manufacturing, and even environmental sustainability. Consider the applications of genome sequencing combined with CRISPR/Cas9 gene editing: from removing disease-eliciting genetic information in live organisms, to genetically-modifying crops for maximum nutritional value and yield, to providing terrorists with enhanced biological weapons to wreak havoc on entire communities and economies, the applications of biological big data exist everywhere that life is present. Thus, as we attempt to navigate the future of biological big data, we must employ a multidisciplinary approach that includes diverse perspectives around the use of biological data.

Addressing this opportunity space, the Applied Futures Lab at Arizona State University sponsored a Threatcasting workshop to explore the future of health and biosecurity at the intersection of artificial intelligence and systems biology, and its implications for private industry, academia, and national security in the next 10 years. The methodology uses technical, social, economic, and cultural inputs alongside current trends, data with an opinion, and science fiction prototyping. From this generative exercise, we

aimed to produce insights that will be used to guide biotechnology strategy related to biological data for organizations across public sector, private sector, academia, and defense. This paper will share the insights and findings from that workshop, adding to the body of knowledge exploring this complex, nuanced and transformational research, development, legal, and cultural space.

Definitions

Biological Data - Here, biological data refers to any form of data that is relevant to an organism's biology and contains patterns that can be identified using data science to derive insights on health and disease. This includes but is not limited to the following: genomic sequencing data, -omics data, biomarker data, tissue sample data, biometric data, and even data pertaining to one's lifestyle.

Biometrics - Biometrics refers to data that measures a person's unique behavioral and physical characteristics.

Population Health - Population health refers to the health status and health outcomes of a group of people, rather than for one individual person.

Personalized Medicine - Here, personalized medicine refers to the tailoring of diagnostic procedures and treatments to the individual patient.

Artificial Intelligence - Artificial intelligence refers to technologies that allow machines to learn from information and produce insights in the same way that a human can learn from experiences. Here, artificial intelligence includes machine learning and deep learning technologies.

Molecular Monitoring - Here, molecular monitoring refers to technologies that enable the ability to read and analyze parts of a living organism's molecular profile, specifically in an attempt to understand the status of molecular systems known to drive health and disease. Examples of molecular monitoring technologies include but are not limited to microarray chips that quantify molecular markers from tissue or blood samples and biochemical technologies used to analyze presence of a pathogen in bodily samples.

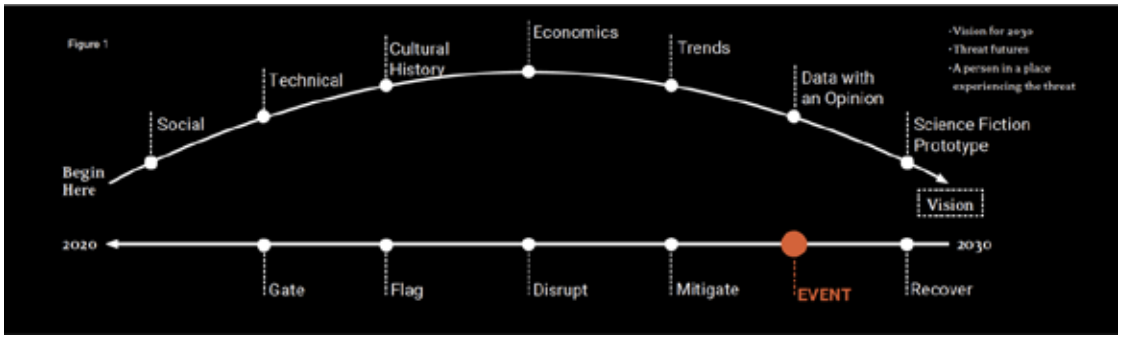
Biobanking - Biobanking refers to the process of collecting and storing biological samples for long periods of time.

Biomarkers - Biomarkers are defined as characteristics that are objectively measured to provide indication of normal biological processes, pathogenic processes, or even biological responses to therapeutic intervention.

Deidentified Data - Deidentified data refers to data that has been altered, with certain features sometimes omitted, to conceal the identity and ensure anonymity of the person from which it originates.

Introduction to Threatcasting

Threatcasting is a conceptual framework used to help multidisciplinary groups envision future scenarios. It is also a process that enables systematic planning against threats ten years in the future. Utilizing the threatcasting process, groups explore possible future threats and how to transform the future they desire into reality while avoiding undesired futures. Threatcasting is a continuous, multiple step process with inputs from social science, technical research, cultural history, economics, trends, expert interviews, and science fiction storytelling. These inputs inform the exploration of potential visions of the future. A cross-functional group of practitioners gathered for four days in March 2021 to create models of futures associated with biological big data. The outcome is a set of possible threats, opportunities, and considerations in the future of biological big data, as well as external indicators and actions to be taken. It is not definitive but does give organizations across academia, private industry, and defense a starting place. Drawing research inputs from a diverse data set and subject matter expert interviews[1], participants synthesized the data into workbooks and then conducted two rounds of threatcasting sessions. These threatcasting sessions generated approximately 8 separate scenarios, each with a person, in a place, experiencing their own version of the threat. After the workshop concluded, futurists at the ASU Applied Futures Lab methodically analyzed these scenarios to categorize and aggregate novel indicators of how the most plausible threats or opportunities could materialize during the next decade and what the implications are for "gatekeepers'' standing in the way of the threats or opportunities.



[1] Transcribed subject matter expert interview excerpts used in the workshop can be found in the Appendix.

Part Two: Threatcasting Findings

The ASU Applied Futures Lab conducted a series of subject matter expert interviews and a subsequent threatcasting workshop from March 8th, 2021 to March 11th, 2021, bringing together an interdisciplinary group of participants spanning public sector, private sector, defense, and academia to explore the threat and opportunity landscape of the future of health and biosecurity at the intersection of AI and systems biology. The following outlines and describes the four major findings of the workshop following analysis by the ASU Applied Futures Lab.

Standardization, Aggregation, and Centralization

With the advent of wearable technology, the collection and use of biometric health data has soared in recent years. Whoop, a popular wearable band is an example that tracks metrics around sleep, exercise, heart rate variability, and can even be used to log dietary information and lifestyle choices. This data, especially when combined with historical information from electronic health records and biological screens for molecular biomarkers, could be particularly helpful in a clinical setting to affect personalized diagnoses and treatments. Further, these various types of data serve as pieces that can be put together to develop a bigger-picture understanding of health and disease, not only at the individual level but at the population health level as well. However, in its current state, there are significant barriers impeding the aggregation and utilization of this data to develop valuable insights on individual and population health.

Thus, great opportunity lies in developing systems for centralization and standardization of biological data, aggregated from various sources spanning biometric data and molecular indicators of biological health systems. Meeting this challenge will enable new AI-driven approaches to model complex biological organisms on the individual and population level, driving significant basic science and clinical discoveries leading to personalized treatments and the development of new indicators of disease.

"There are so many new ways to sense what's happening in biological systems in real time, non-destructively, that I think it's going to be a significant wave of technology development and completely change the way we relate to living systems."

- *Andrew Hessel, Chairman of Genome Project-Write and President of Humane Genomics Inc.*

Protecting the Individual Advances the Whole

As discussed above, collecting personal biological data can be used for personalized medicine and improved health for the individual, but it can also be leveraged as part of larger data sets to analyze population health and drive biomedical discovery. This includes all types of biological data, from genome sequences to biometric data on lifestyle choices; however, it is critically important that one recognizes how personal this data is - data which we sometimes have no control over and data which can be used to identify significant medical vulnerabilities.

With this in mind, ethics demand the requirement to ensure autonomy, anonymity and privacy of participants from whom such personal data is collected, particularly if (with their consent) the data is transported and used for scientific research outcomes. Hopefully, this type of assurance will help to foster societal trust in new technologies that collect and leverage biological data, promoting their use and adoption on a broad scale where they will be most effective.

Data privacy and security concerns are only set to grow as the development of new means to collect and utilize biological data continues to thrive, and as more light begins to be shed on the fragility of data security encompassing all types of data, not just biological information. Even more worrying In conjunction with these trends is the continually increasing power of artificial intelligence and deep learning technologies that can be used to de-anonymize seemingly anonymous biological data. **Thus, great opportunity lies in developing methods to effectively encrypt biological big data, ensuring irreversible deidentification and the utmost security of personalized data, thereby allowing truly private and secure transfer of anonymized biological data.** By capitalizing on this opportunity, organizations can seek to advance the use of biological big data for important healthcare and scientific research outcomes, maximizing society's benefit from innovative biotechnologies and opening the door for new avenues to circumvent or disrupt disease and maintain good health.

"If you haven't already seen it, there's work from a researcher Yaniv Erlich. He's done some cool work on showing that things we think are private in large scale genomic data aren't always. I think he had some work showing you can de-anonymize seemingly anonymous genetic data by just kind of connecting dots that people hadn't thought to connect before, using methods that I don't even know are necessarily at the scale of heavy duty AI. You [can] imagine, as we have large scale machine learning, it will only get more profound."

- *Dr. John Ingraham, Senior Scientist of Machine Learning at Generate Biomedicines*

Toward a More Predictive and Personalized Future

Artificial intelligence is an incredibly powerful tool, providing scientists and technologists with new means of gleaning insights from high volumes of data and using information to predict future outcomes. As collecting and using biological data of all forms become more prevalent in healthcare and disease management, the predictive nature of AI and the opportunities it creates will push the direction of these fields toward more preventive and personalized health care. Although important technological advances remain to be made in AI and deep learning, for biological data a primary rate-limiting factor of progress is our ability to rapidly and frequently obtain high volumes of accurate, precise, and reproducible biological data. This includes the need for wearable technology that collects biometric data at frequent time intervals with precision, as well as the need for molecular monitoring technology that can collect data to paint a picture of the molecular systems at work in an organism. Further, in order to maximize the benefit of these tools for improvement of population health, they must continue to increase their accessibility.

Thus, great opportunity lies in developing accessible technologies to sense increasingly specific biometrics and molecular markers at more frequent time intervals, towards the goal of constantly collecting data that lead to disease prevention and enhanced personalized health outcomes. This trend can already be seen when examining the evolutionary history of biometric collection. Over the last two decades, the popularity of wearable technology has shifted from those devices that can collect biometric data once a day (e.g., measuring blood pressure to assess heart health) to technologies that can retrieve this data in real time using wearable technology (e.g., using real-time heart rate variability to assess heart health) and display this data in a manner in which the layperson user can learn and benefit from it. Moreover, combining this data with information on factors external to biology such as socioeconomic, geographic, cultural, and political factors may provide novel and unique insights into environmental drivers of health and disease - two incredibly complex and multivariate problems.

"I think the opportunities now, given the sensor technologies that we have, whether it's our watches or our phones, other sensors in our homes, some of the sensors that I think we will be starting to make and install like virus sensors and testing, [all] have the opportunity to profoundly change health. … You can imagine that just routine health maintenance now could detect a range of illnesses that just were not detectable before and allow for potentially targeted treatment."

- *Andrew Hessel, Chairman of Genome Project-Write and President of Humane Genomics Inc.*

The Dual-Use Dilemma

Dual-use technology refers to technology that has utility for both benevolent and malevolent applications. Biology is inherently a dual-use field, as the manipulation of biological systems can be leveraged for both improving health or inflicting disease upon an individual or population. While life sciences dual-use technology has consistently been identified as an area of concern for the United States government, as biological data becomes more plentiful, accessible, and insightful, these concerns will only be amplified.

Take for example new personal genome sequencing services offered to a wide range of individuals, allowing them to gain information regarding their family heritage or predisposition to disease. While intriguing to individuals, careful attention must be paid to the full set of actors who have access to this precious data. As previously mentioned, biological data can be leveraged for a myriad of uses. **Thus, a threat in the future of biological big data lies in the accessibility of the data and the potential for nefarious use by bad actors, especially with the increased prevalence of open-source, AI-driven approaches used to gain insights into large amounts of data.** Work towards ameliorating this threat will strengthen both national security and the security of individuals who seek to use commercial services to gain insights into their personal biology, as the science and technology around these services continues to develop. This important threat, concerning all types of biological data (biometrics, molecular markers, genomic sequencing, etc.) must be considered as we push towards the important goal of learning how to use multiple sources of biological data for insights into health and disease.

“With a lot of human genomes being available, there's tons of privacy questions coming to mind. So the same way people talk about social media, “are you the product?” It's not free, but services like 23 and Me and all these companies that are offering you analyses of your genome are then also often biobanking people's samples and then they can keep those in perpetuity.”

- *Dr. John Ingraham, Senior Scientist of Machine Learning at Generate Biomedicines*

Part Three: Indicators and Actions

Using the threatcasting framework, analysts from the Applied Futures lab reviewed raw data produced in the workshop to identify key indicators of a future scenario occurring (flags), as well as critical actions that may enable or prevent the future (gates). These flags and gates are critical for organizations seeking to capitalize on future opportunities or prevent future threats; they serve as inflection points that can help to contextualize potential futures and determine effective strategies to move toward desirable futures and away from disadvantageous futures. Here, flags and gates are presented relative to the main findings, providing specific indicators and actions relevant to each opportunity or threat discussed in Part Two.

Standardization, Aggregation, and Centralization

While there has already been work produced in favor of centralization, standardization, and aggregation of biological data, remaining vigilant to specific flags in the industry will provide an accurate indication of the real-time need to develop systems for integrating various types of biological data. In order for specific types of data to be accepted in a centralized system, there must be support from the scientific and medical communities around using these data types for scientific or medical outcomes. With respect to biometric data, for example, the medical community's trust and utilization of biometric data collected by accessible, commonly used wearables serves as an indicator that this data may be accurate enough to include in a centralized source. On the other hand, If the broad medical community opposes the use of biometric data collected by popular wearables, this indicates that the technology is not yet mature enough to be considered or integrated with other types of biological data. Thus, community acceptance of utilizing more types of biological data serves as an indicator of the need for a validated system that can integrate this increasing variety of biological data types (biometrics, molecular monitoring data, etc.).

An additional flag indicating the need for a centralized system that aggregates different types of biological data is the amount of basic science and clinical research done that seeks to derive insights from integrating various types of biological data. For example, tracking the number of publications that study the use of biological data from patients to derive clinical outcomes may serve as an indicator that this work is garnering interest in the medical community. This new interest indicates the value proposition for innovators to develop a centralized biological data system that will enable the research on this topic to be applied in a clinical setting.

Given the regulatory complexity around data-sharing, several gates can prevent or encourage the development of a centralized biological data system. For example, regulatory restrictions on the use of machine learning as a diagnostic tool may diminish the relevance of such a tool. Additional regulatory restrictions on sharing of biological data, perhaps driven by concerns around consent, may play a role in preventing the use of a centralized biological data system. In light of this, organizations that seek to develop these systems should remain sensitive or responsive to consumer concerns around consent and data privacy and should engage policy makers and consumers in the development of these technologies. Conversely, private companies that hold data from wearable devices may choose to develop features

allowing the connection between various biometric tracking applications, encouraging the development of a centralized biological data system. Additionally, these companies may also develop functionalities that allow for the sharing of this data for research and medical purposes, serving as a gate that enables the development of a centralized biological data system. These functionalities would exacerbate the need for a centralized biological data system that can be used by researchers and clinicians to achieve improved scientific and healthcare outcomes.

Protecting the Individual Advances the Whole

Privacy and security of biological data remains both a threat and opportunity in the future of biological big data - a threat to the development and adoption of systems that leverage biological data for scientific and healthcare outcomes and an opportunity for organizations to develop new solutions that ensure anonymity of biological data and advance its use for scientific and medical insights. The emergence of new systems in the scientific and technological communities that use biological data can serve as an indicator of this threat emerging. For example, development of public repositories where people can place their own deidentified data heightens the need for technologies that ensure privacy of this data. Additionally, increased collection and centralization of biological data and growing applications of AI methods to derive insights from this information also serve as indicators of the growing value proposition for organizations to develop solutions that ensure privacy of one's biological data. Increased public awareness of current limitations on the effectiveness of deidentification of data and increased regulatory attention to deidentified data also serve as indicators that this threat may soon have an adverse limiting effect on the scientific and medical progress associated with utilizing biological big data.

Commercial companies indeed have a stake in this issue, as a commercial solution to ensuring the anonymity and security of biological data would likely be leveraged on a broad scale to bolster existing services that involve the collection and use of biological data. If pharmaceutical companies, for example, begin to provide access to clinical trials through commonly used biological data services and applications (e.g., 23andMe), this would likely increase the need for a data security solution and advance the threat of data privacy limitations hindering basic science and clinical research progress. Further, identification of biological data security breaches from either domestic or foreign actors could also amplify the need to develop a solution to privacy limitations.

Toward a More Predictive and Personalized Future

Technological development serves as a prime exhibit of the famous Isaac Newton quote - it comes from standing on the shoulders of giants. Thus, trends in biological data collection serve as vital indicators that the opportunity to develop more effective and accessible molecular monitoring technologies or wearable biometric technologies is becoming more profound. Critical flags to pay attention to include cost trends for high throughput biology techniques (transcriptomics, metabolomics, etc.) and other personalized molecular data collection technologies. Additionally, trends such as the growing sophistication of machine learning technology and increasing electronic data storage provide insights regarding the potential for leveraging the data that molecular monitoring technologies produce for scientific and healthcare

outcomes. Finally, keeping a keen eye on the medical community's trust in technology that collects biological data and the use of machine learning in clinical settings will serve as valuable flags.

Major enablers of the opportunity to develop innovative molecular monitors or wearables can be seen as private industry trends following basic science work done in academia. The development of technologies that leverage more than just DNA (RNA, protein, metabolites, other biomarkers), for example, serves as an enabler for the further development of more robust molecular monitoring technologies. Additionally, if frequently used data collection software developers (e.g., developers of personal devices such as cell phones) begin to blend with the personalized medical diagnostics industry, this may enable the development of means to leverage biological data collection more efficiently. Finally, shrinking costs of high throughput biological sequencing and growing data storage capabilities serve as flags but also as gates, enabling the development of more accessible molecular monitoring technology.

The Dual-Use Dilemma

A significant threat as the collection and utilization of biological big data increases is the potential for use of this information by nefarious actors. While this is certainly a threat to individuals who may choose to collect or use their biological big data, this can also be a security threat to larger populations. Therefore, several important indicators can provide context as to how far along the actualization of the threat is and when it may materialize. From a national security standpoint, increased tensions around biological warfare or bioterrorism - potentially even from the COVID-19 pandemic - may serve as indications that heightened security measures should be enacted around both publicly available biological data and private biological data that may be subject to security breaches. Additionally, significant advances in the accessibility of gain-of-function technology may also serve as indicators that this threat future is underway. Finally, from a domestic standpoint, significant attention should be paid to the potential use of open biological data and bioscience methodologies by domestic terrorists to inflict harm upon a population. While a significant majority of do-it-yourself (DIY) biologists pose no threat to society and simply have a benign passion for science, the potential for accidental misuse still exists along with the possibility that a domestic terrorist intentionally seeks to use this biological information nefariously.

There are some important gates that may enable these threats in the future, although they are harder to measure if they occur in foreign countries. Primarily, increased general education of the public on scientific development around gain-of-function research and biological agent development may provide bad actors with the information needed to inflict harm, particularly if this information is open source. Additionally, lack of industry control over biological data collected by private companies also may enable this threat, allowing for bad actors to gain access to pivotal biological data through security breaches.

Part Four: Areas for Action

Following the identification of flags and gates for each of the findings, analysts from the Applied Futures lab developed actionable recommendations for organizations in academia, private industry, and national security to consider, with respect to the findings and flags and gates identified in Part Two and Part Three, respectively. These actions for gatekeepers aim to remedy threats that may hinder our ability to leverage advances in biological big data for positive societal impact, and seek to capitalize on opportunities that may advance the use of biological big data for societal benefit. They are listed here to provide gatekeeper organizations with a short-list of potential actions they may consider as they look to the future of biological big data.

Academia

- Work with private industry to develop systems for centralizing, standardizing, and aggregating biological big data for research outcomes.

- Increase attention on developing secure repositories for various types of data produced in basic and clinical research.

- Increase attention on developing centralized databases where all known biomarkers for disease are aggregated.

- Deposit deidentified biological data from basic science research into centralized systems.

- Prioritize basic and clinical research identifying how molecular monitoring systems can be used to detect and mitigate disease before its onset.

- Prioritize basic research that seeks to further decrease the cost of DNA sequencing and other personal molecular data collection.

- Increase collaboration with industry around using AI/machine learning solutions for diagnostics and treatment in clinical settings.

- Work with defense partners to develop robust policies for balancing open science with the threat of bad actors accessing valuable biological data.

- Prioritize research that develops an understanding of how bad actors may be able to use biological data and collaborate with defense partners to develop systems and solutions that prevent nefarious use of biological data.

Private Industry

- Increase attention on developing technology that addresses security issues of biological data for consumers.
- Increase commercial sector resistance to data breaches of biological data.
- Form collaborative relationships with defense/national security to navigate security challenges, ensure security and privacy of biological data, and develop a big-picture understanding of biosecurity risks.
- Increase attention on developing robust policies around biological data security.
- Increase focus on identifying all actors that may be able access biological data from private industry.
- Increase focus on lowering cost and accessibility of technology that enables personalized medicine (wearables, molecular monitoring technology, etc.).
- Increase attention on developing novel business models that incentivise sharing of secured and privatized biological data into centralized systems.
- Increase attention on developing new business models around sharing of biometric data collected by wearables or molecular monitoring data collected by biotechnology services.
- Increase attention on developing features that allow users to opt-in to connecting data across biological data apps.
- Work to improve secure access to electronic healthcare data and improve electronic health record systems.
- Deposit deidentified biological data from clinical trial studies into centralized systems.
- Work with government and regulatory partners to reimagine the clinical trial process for improved efficiency.

Defense/National Security

- Ensure biological data are a component of emerging work done in information warfare, specifically around tracking the flow of biological data.
- Increase focus on proactively identifying domestic and foreign bad actors as biological big data becomes more accessible.

- Consider dual-use implications of biological data and assist in informing the commercial industry on the risks and opportunities of technology for good and ill.

- Provide funding for industry and academic groups to effectively encrypt biological data transfer.

- Increase attention and internal research and development around security protocols for biological data produced by the domestic population to serve and protect that population.

- Work with academia to develop systems for classifying national security risk of publicly available datasets from low to high.

- Keenly monitor DIY science and determine if regulations need to be pushed forward as prevalence of biological data increases.

- Work with academia to develop robust policies that balance open science with the threat of bad actors accessing valuable biological data.

Part Five: Considerations

The following are critical considerations to be acknowledged for all of the findings presented in Part Two. These considerations were common themes that occurred across multiple future scenarios developed by participants and have important implications on the opportunities and threats identified in the threatcasting workshop. These considerations should be interrogated by academia, private industry, and national security actors as they seek to act on the identified threats and opportunities.

The Relationship Between Science and Society

As technology that captures and leverages biological data becomes more prevalent and available, its capacity to benefit society is reliant on the relationship between the public and science. As with most data-driven solutions, more biological data means the potential for more robust health and scientific outcomes; however, the abundance of available biological data to be used for garnering important insights depends heavily on the willingness of the public to accept, adopt, and consent to new technologies that collect their biological data. Even profound biomedical advances that arise from big biological data serve no purpose without the consent for use by the public. Thus, the public's trust in data-driven means of health improvement and disease mitigation is a major driver in the adoption of new technologies set to benefit society.

An interesting and equally important aspect with regard to the relationship between science and society is education in our communities. Both general educational attainment in a community and specific education around new technologies critical for population health are drivers in the willingness of society to adopt new biomedical advances. Take for example the COVID-19 vaccines - while millions of dollars were poured into the development of these technologies, their material impact on society is heavily reliant on the proportion of the population willing to use them. The same will apply to technologies collecting and leveraging biological data in the future, and similar emphasis on community education and society's trust in science must be retained.

Equity, Representation, and Bias

As biological data is increasingly leveraged for personalized medicine and scientific outcomes, it will become more and more important that a broad range of demographics are included in this data. Inclusion of a variety of ethnic groups and genetic backgrounds makes scientific and healthcare outcomes from biological data far more robust, ensuring that findings are not applicable only to a non-representative portion of the population. Further, inclusion in biological big data is pivotal for the equity of solutions derived from the data; solutions derived from inclusive and comprehensive biological data may be applicable to a broader range of a population rather than to a select few.

Beyond inclusion and diverse representation in biological data sets is the challenge of inherently biased AI. Artificial intelligence technologies themselves are not inherently biased; however, as the creator of these systems develops them, they can naturally and unintentionally input their own biases into the

program. Oftentimes these biases can be clearly detected in the outcomes of the AI, but sometimes these biases can go unnoticed, leading to a far more dangerous scenario. Special attention should be paid to equity, representation, and bias when generating biological big data or developing systems that utilize it.

"One of the big questions for me facing AI in the biotech space is what are the assumptions that are going to be encoded into the platforms that are being built and the algorithms that are being designed, and how are they going to shape the questions that are possible to ask about healthcare biotech?. We know for example that our medical data sets are really heavily skewed towards white men, and so if that's the data that's going in, for example, what are the outputs going to be?"

- *Dr. Emma Frow, Assistant Professor in the School of Biological Health Systems Engineering & the Lincoln Center for Applied Ethics at Arizona State University*

Responsible Innovation

It is critical to note that for each of the threats and opportunities identified in this report, there are important political and cultural dimensions to take into account. For example, political questions around an individual's discretion of how biological data is used will need to be addressed in parallel to the development of new technologies and solutions that collect and use biological data. Further questions around data security and sharing of data will also need to be addressed.

From a cultural perspective, similar questions around the influence of various cultural beliefs on the adoption of biomedical advancements will arise. As leveraging biological big data becomes more prominent in health and disease management, ethical questions around the use of biological data screening to identify risk for disease, for example, will come into play. These intersecting cultural questions will also need to be addressed as innovations in biological data come about, in order to maximize the beneficial impact of these technologies on society.

Responsible innovation - the notion that environmental, moral, political, cultural, religious, and democratic factors must be considered in scientific research and technological development - will play a critical role in the future of big biological data. These innovations are not just pieces of standalone technology, they are about people. By ensuring that people and all of their complex dimensions listed above are taken into account in future innovations, we reduce the risk of negligence and adverse effects in technological development.

Appendix: Subject Matter Expert Interview Transcripts

The following are algorithmically transcribed Subject Matter Expert interviews used in the workshop.

Dr. Emma Frow

Dr. Emma Frow, an Assistant Professor with a joint appointment in the School for the Future of Innovation in Society and School of Biological & Health Systems Engineering, and a Lincoln Professor for Applied Ethics at Arizona State University, shared important considerations for the workshop participants around regulatory pathways for biotechnology innovations. Below are excerpts from this conversation used for the workshop.

This is actually an example that springs to mind that's not in the healthcare domain or even the biotech domain, but it's an adjacent one that I keep thinking about and wondering about the degree to which this might apply in biotech. I think it's something to be thinking about now, which is the recent heightened scrutiny around algorithms in big tech. There are growing public discussions around, say for example, racial bias that has been built into Google's search algorithms, or discussions around injustices that accompany a lot of government-led surveillance, like AI-informed decision-making platforms for things like how you allocate welfare benefits and so on. These algorithms are framed and structured by particular questions a group is interested in answering. It takes a while to build those platforms [and] it takes a while to get the datasets up and running. And now, 10 or 15 years down the line, people are starting to pay attention to the outcomes that are being generated and seeing how the outcomes have been encoded into the original structures of the platforms. And so I think one of the big questions for me facing AI in the biotech space is, what are the assumptions that are going to be encoded into the platforms that are being built and the algorithms that are being designed, and how are they going to shape the questions that are possible to ask about healthcare biotech? We know for example that our medical data sets are really heavily skewed towards white men, and so if that's the data that's going in, for example, what are the outputs going to be? So really taking some time to think hard and very interdisciplinarily upfront about the sorts of questions that really need to be addressed and how to construct your data sets to be mindful of those questions and not exacerbate existing problems with our current data sets and approaches, I think is going to be a big challenge.

I often think of policy and regulation as essentially a classification problem: how do you define a problem so that you can figure out what regulatory route it should go down? And I think one of the challenges we're seeing increasingly in the biotech space, which I suspect will only be compounded by the convergence with AI, is that there will be growing numbers of products - the US has a product-based regulatory system for biotech - that don't fall neatly into existing classification schemes, so it's not clear what regulatory pathway they need to go down. What happens is you either have to retrofit a product into an existing pathway that it's not quite suited for, or you have to try and modify the classification buckets that you're working within, which is a time-consuming process and very politically charged and hard to predict. [Of course], you want to future-proof your categories as much as possible, but it's [the innovators] job to disrupt those. So I think this is an ongoing definition and classification problem about

whether or not these new technologies fall into existing buckets, or whether they're creating new buckets. So it's first off determining what bucket it falls into and then figuring out if you have the right pathway, [and] whether you're asking the right kinds of questions in order to determine whether or not those products are appropriate to be released to the market.

Andrew Hessel

Andrew Hessel is the Chairman of Genome Project-Write and founder of Human Genomics Inc. He shared his thoughts on emerging biological trends in health and medicine that will disrupt the biotechnology industry. Below are excerpts from this discussion used in the workshop.

I have two children, three and six years old. I tell people that they're lab grown. They're both the products of IVF. I'm standing on the shoulders of work that's been done since the 1970s to have children. Both of them were profiled genetically, one directly. We actually had the embryo assessed genetically. This is profound because personal health today starts essentially by screening your parents for potential risks and then screening the embryo for any major genetic defects. So there's no reason, given the tools and technologies that we have, if they were universally deployed and accessible, to have any major genetic defects in the human population. Of course this can be applied across any living creature. We could do the same thing for animals, plants, etcetera. But this is profound. So health today starts essentially at conception, or even before. As for monitoring health, well you don't wake up when you're 18 and suddenly decide whether you're going to be a healthy person or a sick person. It should be a process where our health is being monitored and screened effectively and continuously. I don't think we have that system in place yet. Particularly in the United States where the healthcare system is unique - let's just call it that. I'm Canadian and the idea of health monitoring is just part of the process. Prevention is baked into the system in a place like Canada, because you want to minimize your overall costs on a single payer system. But I think the opportunities now, given the sensor technologies that we have, whether it's our watches or our phones, other sensors in our homes, or some of the sensors that I think we will be starting to make and install like virus sensors, could profoundly change health. One example of this is a company Grail that is doing liquid biopsies, which from a blood test is able to find cancers in your body that are essentially too small to see. You're not sick. You don't feel sick, it could be a millimeter-size tumor, but that tumor is still going to be shedding genetic markers of its existence that we can detect with a simple blood test. These are early days for liquid biopsies, but you can imagine that just routine health maintenance now could detect a range of illnesses that just were not detectable before and allow for potentially targeted treatments that are similarly benign. If you take an antibiotic before the infection has really spread, you kind of nip it in the bud. So I see health maintenance subscriptions coming into the future where either your government health organization or your private health organization is constantly monitoring you, looking for signs very early of problems, whether it's physical or mental, and applying the appropriate countermeasures sometimes without a lot of input from yourself. It could be in the form of a vitamin or just balancing your life so that you can go on and worry about other things.

The lowest level technology that's still starting to really make itself felt is genomics. The ability to read nucleic acids, not just DNA but RNA, and to be able to manipulate nucleic acids, both DNA and RNA in very

precise ways, whether that's base editing to correct an error, whether it's turning off a segment of DNA, adding a segment of DNA, or complete synthesis of a new genome. I think this is one of the most profound, low level technologies that we've barely started to crack open the potential. Just looking at the range of CRISPR technologies that have appeared since the first CRISPR paper over a decade ago is an indicator of this. Just looking at the incredible growth of DNA sequencing, or the fact that most of us haven't been sequenced. There are 8 billion people in the world and the number of full genomes is still in the low millions, and even the most successful companies at profiling DNA [are] nowhere close to the amount of data you get out of full genome sequencing. I think the largest company has under about 25 million subscribers. I think that there's a number of layers of technology that get built on top of this that I just lump as molecular sensing. Being able to do something like track your blood sugar and deliver the appropriate amount of insulin if you're a diabetic, these technologies haven't been fully integrated. They're still very clunky. There are so many new ways to sense what's happening in biological systems in real time, non-destructively, that I think it's going to be a significant wave of technology development and completely change the way we relate to living systems.

I'm constantly being surprised at our ability to sense molecular systems. For example, recently one of the geniuses in this space, Jonathan Rothberg, who was the person who seeded next-generation sequencing technology back in the late 2000's with a company that he spun out called 454 Life Sciences, he's recently reduced ultrasound devices to a chip with his company Butterfly Networks and also reduced proteomics to a chip, being able to being able to read protein structures. It's kind of like a mass spec on a chip. So these are really powerful chip-based technologies, and I think this is going to have a significant effect in the healthcare setting as well in research. If you think about the devices that we have on our desk to do our work, particularly in a lab, some of these devices are large and expensive pieces of kit. I think that most of those devices start to become smaller, faster, better, cheaper as they become more useful to a broader market. For example, one of my favorite liquid handling systems uses sound energy to dispense fluids in a very precise way - I mean picolitre volumes - it's a brilliant device called an Echo made by a company called Labcyte, but the device is hundreds of thousands of dollars. That device could be broadly used in any situation where you're doing liquid handling. So in research, in clinical applications, even in some home-based testing systems. But the device is just too expensive for broad use. It's kind of like the early days of computers, if you're spending $200,000 for a mainframe, you're going to treat it a lot differently than a $500 cell phone. So I think that that's the trend that will power a lot of the advancements over the next decade or two. It'll come with tools that make these technologies more powerful, more accessible, more useful.

Dr. John Ingraham

Dr. John Ingraham, a Senior Scientist focused on Machine Learning at Generate Biomedicines, provided his thoughts on innovative technologies soon to disrupt the biotechnology and pharmaceutical industry. Below are excerpts of this conversation used for the workshop.

So, starting with reading DNA, our ability to be collecting everyone's genomes, not just SNP markers, but full genomes is rapidly increasing. I think a genome is a hundred dollars now - a human genome. And then

also, we're doing genomics on everything. So that also includes viruses and bad agents, so one immediate security threat around that is if the DNA of bad pathogens is available, I've even heard of DARPA grants around trying to develop systems for identifying emerging [natural] pathogens. So that's underlying with our question, could we have forecasted like emerging epidemics and pandemics, but it also plays a role into bio-terrorism. If everybody can read and write DNA very easily, could bad actors potentially synthesize the genome of some kind of blacklisted pathogen? I know there are certainly some questions from DARPA around that in the research community. Then, the other big area that's really affected by how much we can sequence genomes that kind of touches the systems biology and AI angle is: with a lot of human genomes being available, there's tons of privacy questions coming to mind. So the same way people talk about social media, are you the product? It's not free, but services like 23 and Me, and all these companies that are offering you analyses of your genome are then also often biobanking people's samples and then they can keep those in perpetuity. If you haven't already seen it, there's work from a researcher Yaniv Erlich. He's done some cool work on showing that things we think are private in large scale genomic data aren't always. I think he had some work showing you can de-anonymize seemingly anonymous genetic data by just kind of connecting dots that people hadn't thought to connect before, using methods that I don't even know are necessarily at the scale of heavy duty AI. You [can] imagine, as we have large scale machine learning, it will only get more profound. Then in terms of writing DNA, I think what's really going to be big over the next 10 years is [that] we now have an ability to rapidly synthesize really large scale experiments where you can generate thousands or hundreds of thousands of designed molecules in one go. These are entering pharma. They're quite exciting in pharma, this ability to have a computer [that] could basically computationally synthesize hundreds of thousands, soon maybe millions of designed biological sequences that could encode proteins to do all sorts of things. You see this being adopted across the board in industry. There's also a way where you can make it relevant to small molecules with DNA encoded libraries, and now that we can synthesize huge amounts of things and also build assays that use high throughput sequencing to measure what they do, what's going to really accelerate in the next 10 years is our ability rationally make molecules that do anything you want. And that's exactly the kind of tool that could have maybe good or bad ramifications.

10 years is a really long time horizon for some of this stuff since it's moving so fast. I think the results from AlphaFold are quite exciting. So that competition [Critical Assessment of Structure Prediction] is about predicting the structures of natural proteins. I think there was a huge jump this year where we're really starting to see we can consistently - for natural proteins - predict their structures with technologies like AlphaFold 2 within a level of accuracy that is actionable for a lot of downstream things. It's accurate enough [that] you might start imagining to use it in like a pharma application or something else Once you're there, then the next big set of questions is going to be now [that] you can predict the structures of natural proteins using natural protein data is, can we just predict the structures of proteins that have never existed before, with structures that have never existed before? Of course, de-novo protein design is something that's been around for a while and people have been using even non-AI approaches, but I think AI approaches are going to massively accelerate that.The thing the AlphaFold team themselves have identified as the next big problem is now that we can predict for one protein how well it's going to fold, can you predict protein interactions? Let's say, if the field can master that problem in the next few years,

and especially if they can extend the mastery of these problems to artificial proteins and artificial protein interactions that don't exist in nature, then I think we're in a situation where you can kind of just point at any part of biology in a human or other organism and just instantly invent molecules that disrupt, modulate, [or] restructure what's going on. So, from a therapeutic point of view and the pharma point of view there's a bunch of obvious applications, but it's pretty interesting if you kind of go from the non-AI side of protein design, people have started showing that there are ways that you can use proteins to solve problems in energy and even in manufacturing. There's work on showing that if you could pattern nanoscale patterns with proteins, then you could use those in conjunction with kind of more conventional semiconductor patterning or other approaches. So we might also start seeing a kind of melding between those two worlds where if we could really master proteins, it could affect energy and manufacturing - things that have nothing to do with health. I definitely think even the jump we saw with AlphaFold and AlphaFold2, and the idea that seemingly [with] a lot of components of approaches like that, it's pretty clear that you could keep improving them, it seems like it's a pretty safe bet that we'll have a mastery of those, especially on a 10 year timeframe, at least for stable proteins that are well folded and are maybe small. There's always going to be your hard cases of really multi-conformational [proteins, or] getting proteins to the right compartments in the body. There's lots of challenges with that, but I think there will be a point of mastery.

Funding's a big driver. I think, again, this ability to read and write DNA on massive scales is another huge driver. So as the cost of DNA synthesis goes down, what that will unleash is a virtuous cycle between being able to synthesize your computational ideas, and then [being able to] update your models with the results of those high throughput experiments. That was a big missing piece in the past, when people were trying to say, "can we predict protein structure correctly?", they were doing it in a really low throughput way. You have to predict a bunch of things, and then you do a crystallography or other type of structure determination experiment that's very expensive both in terms of personnel and equipment. But now if you can couple these types of technologies to really high throughput DNA synthesis, you'll be able to suddenly get a read out on trying hundreds of thousands or millions of things on a regular cadence.

Two specific numbers to always look at are the throughput and cost of DNA sequencing and the throughput and cost of DNA synthesis, and new technologies keep arising around sequencing and the same story for synthesis. I think how fast those develop and if those can develop in qualitative ways from where they are now would really be gating on how fast progress can be made.

Dr. Jane Maienschein

Dr. Jane Maienschein is a University Professor, Regents Professor, and President's Professor as well as the Founding Director of the Biology and Society Center at Arizona State University. Here, she provides important questions to think about regarding societal adoption of novel biotechnologies. Below is an excerpt from this conversation used in the workshop.

So the usual approach would be to look at the ones I mentioned like recombinant DNA in the 1970s, let's look at CRISPR/Cas9 much more recently, or in between cloning and STEM cells. Those are all technologies

that people agonized a lot about, but I like to go back farther to the 1890s and Jacques Loeb, who did work at the Marine Biological Laboratory in the University of Chicago. He was very much interested in the idea of controlling life. He wanted to engineer life. He wanted to be able to take a living thing and make it do what we want to do for ethical reasons in order to make things better. He first discovered artificial parthenogenesis, or rather discovered that he could cause artificial parthenogenesis. He discovered by accident that he could take sea urchin eggs and move them into a different concentration of saltwater, and they started to develop on their own. They didn't need to be fertilized. There was a lot of talk about virgin birth, a lot of headlines in the newspapers about not needing the male. So he raised the question, how much can we do in shaping a cell and then shaping an organism simply by changing its saltwater concentration or other physical factors? And can we learn to do this in a productive way that can make things better ([this was] before genetics)? How did people respond to this big idea? Largely with excitement and enthusiasm. Some of the women's suffrage movement was enthusiastic, like "Okay! Women can take their eggs and have offspring on their own. Who needs the males?" So that was one social response, but there were other people who were very excited that we might be able to use this to do good, and maybe we could use it medically. Of course that was very far in the distant future and didn't really make too much sense [at the time], but it was very interesting how positive the reactions were. Then at other times in history, when we had the eugenics movement and the idea of controlling populations and breeding, there was both enthusiasm and great positive support, as well as a growing concern around who was going to do the work, as well as questions about who's going to use this technology. It's only gradually in the 20th century that people really started to ask that question and how something might be used. Then came dropping the atomic bomb, which was not meant to be a biological experiment, but was in terms of affecting the genetic pool with radiation. At that point, it was very much about who's going to use a technology for what, and then how can we control it? So there are evolving questions people have asked about the process of controlling and about the extent to which the technology is good. Take for example social responses to stem cell technology - "Great! Let's use it to get rid of neurodegenerative problems. Oh, wait, it's horrible because we have to get the stem cells from somewhere." So as a complex and pluralistic society, we started to have a recognition of the two-edged sword, the classic challenges of dual-use technology and dual-perception technology. A lot of those questions, "Who? What? When? Where?" about using biotechnologies have been there and have gradually evolved up to the present day, which again, brings us to the question: what's new within a particular technology? Is it just applying old questions or is there something really new that we need to look at?

Sherwin Nuland, who was a surgeon at Yale medical school and who unfortunately died not long ago wrote a book, *The Wisdom of the Body.* He wrote a number of other things that are really quite insightful, but in *The Wisdom of the Body* he talked about when heart transplants first became available at Yale New Haven Hospital, and he talked to people about heart transplants. He asked people, "you have this severe heart disease. You're going to die from this heart disease and fairly quickly, but one option is we could give you a transplant and it might extend your life. The evidence is growing that it could do that. What do you think about a heart transplant?" And he said that the reaction of some was ,"Great, Bring it on!" or it was, "Oh wait, I don't think it's tested." And those are reacting to the science, but the social reaction was,

especially from one guy, one white guy, who said, "You're not going to give me a woman's heart. Are you?", and then there was somebody else who said, "I absolutely oppose this if it's going to come from a black person." And so there was that reaction at that time. He said it only took about five years of people getting heart transplants and living for the individuals and the families to say, "Oh, a heart, it's just mechanical. Yeah. Pop out a heart and pop in a new one. No problem. It's not a social, it's not a psychological problem. It's not changing who I am." So it radically changed how people were thinking about the heart as somehow defining themselves to, "Oh, it's a mechanical thing". I think that will happen again. And again, with prosthetic limbs and prosthetic other parts - we've accepted, it's still the same person. Will we get to the point of having prosthetic brains or prosthetic neural systems? People are more nervous about those kinds of biotechnologies, but as soon as we can do it and it starts to help avoid some paralysis or whatever, there will be people who will embrace it, and then people will rush to it, and then ethicists will worry about whether only the rich white guys get it and how do we make it equitable for all? So the questions around societal adoption of biotechnology change based on where we are in terms of the background, socioeconomic factors and many other details about the context.