Texas National Security Review: Volume 5, Issue 3 (Summer 2022)
Print: ISSN 2576-1021 Online: ISSN 2576-1153

China's Biomedical Data Hacking Threat: Applying Big Data Isn't as Easy as It Seems

Kathleen M. Vogel
Sonia Ben Ouagrham-Gormley

Concerns have developed in recent years about the acquisition of U.S. biomedical information by Chinese individuals and the Chinese government and how this creates security and economic threats to the United States. And yet, China's illicit acquisition of data is only one aspect of what is required to produce an enhanced science and technology capability that would pose a security threat. Current assessments fail to account for the heterogeneity of big data and the challenges that any actor (state or nonstate) faces in making sense of this data and using it. In this context, current law enforcement and policies that focus on the Chinese acquisition of biomedical big data should expand to other important aspects of China's science and technology capabilities, including the country's ability to interpret, integrate, and use the acquired data for its economic or military benefit. This article provides new socio-technical frameworks that can be used to provide greater insights into Chinese threats involving biomedical big data.

> *"Data are not, by themselves, a form of knowledge."*[1]

In January 2021, a former U.S. intelligence officer revealed that BGI, a Chinese DNA sequencing company suspected of having ties with the Chinese military and Chinese government, had attempted to collect American DNA under the guise of offering to build COVID-19 testing sites in six U.S. states, including California and New York.[2] These states, along with their medical institutions, were warned against cooperating with BGI because it could give China the opportunity to harness this biomedical data and use it for economic and security purposes. This incident is only one of many recent examples that highlight the growing anxiety felt by the U.S. government and public about the increasing availability of big data, including biomedical data, and about how China might use this data.[3] In June 2021, the Biden administration signed a new executive order regarding the threat posed by China to U.S. information technologies, systems, and digital data.[4] These developments — as well as those involving intellectual property theft, espionage, and China's foreign talents programs — point to a growing con-

[1] Sabina Leonelli, "Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science," in *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44, no. 4 Part A (December 2013): 505, https://doi.org/10.1016/j.shpsc.2013.03.020.

[2] Jon Wertheim, "China's Push to Control Americans' Health Care Future," *60 Minutes*, Jan. 31, 2021, https://www.cbsnews.com/news/china-us-biodata-60-minutes-2021-01-28/. See also Kirsty Needham, "Exclusive: China Gene Firm Providing Worldwide COVID Tests Worked with Chinese Military," *Reuters*, Jan. 30, 2021, https://www.reuters.com/article/us-china-genomics-military-exclusive-idUSKBN29Z0HA.

[3] "China's Collection of Genomic and Other Healthcare Data From America: Risks to Privacy and U.S. Economic and National Security," National Counterintelligence and Security Center, February 2021, NCSC_China_Genomics_Fact_Sheet_2021revision20210203.pdf (dni.gov); Gordon G. Chang, "China Wants Your DNA—And It's Up to No Good," *Newsweek*, Nov. 30, 2020, https://www.newsweek.com/china-wants-your-dna-its-no-good-opinion-1550998; Emile Dirks and James Leibold, *Genomic Surveillance: Inside China's DNA Dragnet*, Australian Strategic Policy Institute, Policy Brief Report No. 34, June 2020, https://www.aspi.org.au/report/genomic-surveillance; Sui-Lee Wee, "China Is Collecting DNA from Tens of Millions of Men and Boys, Using U.S. Equipment," *New York Times*, June 17, 2020, https://www.nytimes.com/2020/06/17/world/asia/China-DNA-surveillance.html; Yves Moreau, "Crack Down on Genomic Surveillance," Nature, no. 576 (December 2019): 36–38, https://doi.org/10.1038/d41586-019-03687-x; and *Safeguarding the Bioeconomy*, National Academies of Science, Engineering, and Medicine, 2020, https://doi.org/10.17226/25525.

[4] "Fact Sheet: Executive Order Protecting Americans' Sensitive Data from Foreign Adversaries," The White House, June 9, 2021, https://www.whitehouse.gov/briefing-room/statements-releases/2021/06/09/fact-sheet-executive-order-protecting-americans-sensitive-data-from-foreign-adversaries/.

cern about China as a U.S. strategic competitor and national security threat.[5]

This paper will discuss the security concerns that have arisen about biomedical big data and how they are based on a simplistic understanding about what is required to create, process, and use biomedical big data. Bioinformatics and science and technology scholarship provides a more accurate, empirically based account of the challenges of working with and applying biomedical data — challenges that stem from the heterogeneous nature of big data. These studies indicate that there are a number of bottlenecks and errors that can get introduced from the moment that a piece of biomedical big data is created through the journey to processing, storing, transferring, and using the data. In addition, laboratories that produce biomedical big data do not have standardized methods for creating or working with data. Thus, moving the data from one location to another can often involve tedious and nontrivial data curation and translation in order to use that data in a new setting. All of these data usage issues pose challenges to actors — whether state or nonstate — wishing to licitly or illicitly acquire biomedical big data and use them for their economic or security benefit. For far too long, the U.S. security community has jumped to assumptions about how easily data can become a security threat, without considering the more complex socio-technical factors shaping how those data can be used in practice.

Instead of merely focusing on data acquisition, intelligence analysts, law enforcement officials, and researchers should spend more time studying China's bioinformatics infrastructure. This should include determining who is working in China's bioinformatics community, where the industries are located, and what their capabilities are, as well as how China has been able to deal with the problems inherent in building a bioinformatics research community. These researchers should also investigate how China's bioinformatics community of experts is — or is not — overcoming the problems that have plagued China's overall science and technology infrastructure and what problems the country has encountered in using and translating biomedical big data. Probing these kinds of research questions would provide a more nuanced understanding of what kinds of indigenous innovation are occurring within China's bioinformatics community that might actually pose security concerns for the United States. If academia, intelligence analysts, and law enforcement could collect and analyze data that are relevant to answering these kinds of issues, they would acquire a much better understanding of what China's science and technology capabilities are when it comes to actually using the biomedical big data that it acquires. The following sections will outline the limitations in existing assessments of Chinese biomedical hacking, as well as provide new analytic frameworks and a research roadmap for conducting more robust assessments of this potential security threat in the future.

Biomedical Big Data and Security Concerns

In recent years, there has been increased attention on how the avalanche of new biomedical and life science "big data" coming from genomic sequencing, databases, electronic medical records, and other sources will usher in a new era of "precision medicine" that will reap a variety of public health benefits.[6] "Big data" is a term used to describe extremely large data sets that can only be analyzed computationally, either individually or integrated with other data sets, to reveal previously unknown patterns, trends, and associations. For example, data from large-scale genomic studies are expected to elucidate the role that genetics plays in particular diseases, indicating which individuals might develop a given disease or disorder, and whether a new drug or therapeutic treatment might help patients suffering from a medical condition. The National Institutes of Health started the Big Data to Knowledge Initiative and the Precision Medicine Initiative, which aim to gather heterogeneous biomedical data (e.g., data about genomics and proteomics, as well as patient electronic health records, clinical trial data, and environmental data, to name a few) on millions of Americans to bet-

[5] *Summary of the 2018 National Defense Strategy: Sharpening the American Military's Competitive Edge*, U.S. Department of Defense, 2018, 1, https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf; and "Annual Threat Assessment of the U.S. Intelligence Community," Office of the Director of National Intelligence, April 9, 2021, 4, 6–8, https://www.dni.gov/files/ODNI/documents/assessments/ATA-2021-Unclassified-Report.pdf.

[6] "Next Generation Public Health: Towards Precision and Fairness," *The Lancet*, no. 4 (May 2019): e209, https://www.thelancet.com/pdfs/journals/lanpub/PIIS2468-2667(19)30064-7.pdf; Tarun Stephen Weeramanthri et al., "Editorial: Precision Public Health," *Frontiers in Public Health*, April 30, 2018, https://doi.org/10.3389/fpubh.2018.00121; Ralph Snyderman, Caroline Meade, and Connor Drake, "Value of Personalized Medicine," *Journal of the American Medical Association* 315, no. 6 (2016): 613, https://doi.org/10.1001/jama.2015.17136; Christina X. Chen and Joel W. Hay, "Promise of Precision Medicine," *Journal of the American Medical Association* 314, no. 16 (2015): 1752, https://jamanetwork.com/journals/jama/fullarticle/2466116; and *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*, National Research Council, 2011, https://pubmed.ncbi.nlm.nih.gov/22536618/.

ter understand the genetic, environmental, and behavioral/lifestyle determinants of diseases for the purpose of creating genetically guided medical treatments, enabling early detection, and looking toward preventative interventions in the future.[7]

Advances in genome editing, a process of modifying DNA sequences at precise genomic locations, are seen as important research steps that can aid in the development of precision medicine.[8] Harvesting the potential of genome editing for public health depends on collecting and organizing accurate and digitized information about human and animal gene sequences and genomes. Since the advent of the Human Genome Project, the sheer volume of digitized genomic information has been on the rise due to dramatic decreases in the costs of DNA sequencing and synthesis, computing power, and data storage. These genomic data, coupled with data from patient electronic medical records and other relevant biomedical data stored in large, digitized databases, are seen as critical to realizing the vision of precision medicine.[9] New digital technologies are allowing for the mining, collection, analysis, and visualization of these vast reservoirs of biomedical data in new ways.

The potential for the nefarious acquisition of these data by commercial, state, or nonstate actors has been noted by U.S. intelligence and law enforcement officials, and academic and think tank scholars.[10] These security concerns over biomedical big data emerged in the 2010s, soon after the term "big data" began garnering media and public attention.[11] In 2014, the American Association for the Advancement of Science, in conjunction with the FBI and the U.N. Interregional Crime and Justice Research Institute, produced the first report that put a spotlight on the security risks inherent in the generation of big data in the life sciences. The report highlighted two key data-related vulnerabilities. The first is the vulnerability of biomedical databases and IT infrastructure to theft or hacking. The second is the possibility that malevolent actors may access, integrate, and analyze diverse private and publicly available biomedical data to create pathogens, toxins, or biologically active molecules to harm animals, plants, or people, and/or to evade current detection devices and other medical countermeasures. The report noted, "Unlike other security risks often associated with the life sciences, the potential risks of Big Data in the life sciences rely on attacking the underlying data and cyber infrastructure and/or inappropriately using data and analytic technologies, not theft of actual pathogens or toxins."[12] The report seems to suggest that the enhanced risk or threat comes from, or is driven by, the data and cyber infrastructure alone.

The FBI's Weapons of Mass Destruction Directorate and Biological Countermeasures Unit funded the 2014 report and has taken special interest in the big data life science threat.[13] Ed You, then

[7] "Big Data to Knowledge," National Institutes of Health, accessed April 21, 2022, https://commonfund.nih.gov/bd2k; and "What Is the Precision Medicine Initiative?" U.S. National Library of Medicine, accessed April 21, 2022, https://medlineplus.gov/genetics/understanding/precisionmedicine/initiative/.

[8] Chayakrit Krittanawong, Tao Sun, and Eyal Herzog, "Big Data and Genome Editing Technology: A New Paradigm of Cardiovascular Genomics," *Current Cardiology Reviews* 13, no. 4 (November 2017): 301–04, https://doi.org/10.2174/1573403X13666170804152432; and Anne Sliper Midling, "Using Big Data to Understand Immune System Responses," *Phys.org*, Jan. 25, 2017, https://phys.org/news/2017-01-big-immune-responses.html.

[9] Wesley T. Kerr et al., "The Future of Medical Diagnostics: Large Digitized Databases," *Yale Journal of Biology and Medicine* 85, no. 3 (September 2012): 363–77, https://pubmed.ncbi.nlm.nih.gov/23012584/.

[10] National Academies of Science, Engineering, and Medicine, *Safeguarding the Bioeconomy*; Diane DiEuliis, "Parsing the Digital Biosecurity Landscape," *Georgetown Journal of International Affairs*, no. 21 (Fall 2020): 166–72, https://doi.org/10.1353/gia.2020.0031; "Safeguarding Our Future: Would You Want a Foreign Government to Have Your DNA?" National Counterintelligence and Security Center, May 27, 2020, https://www.dni.gov/files/NCSC/documents/SafeguardingOurFuture/No_1_FINAL_NCSC_Safeguarding_our_future-DNA27_May_2020.pdf; Daniel R. Coats, *Worldwide Threat Assessment of the U.S. Intelligence Community*, Office of the Director of National Intelligence, Jan. 29, 2019, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf; Ryan Holeywell, "U.S. Government to Biotech: You're Facing Threats—But We Can Help," *TMC News*, Oct. 3, 2019, https://www.tmc.edu/news/2019/10/fbi-to-biotech-youre-facing-threats-but-we-can-help/; Diane DiEuliis, Charles D. Lutes, and James Giordano, "Biodata Risks and Synthetic Biology: A Critical Juncture," *Journal of Bioterrorism and Biodefense* 9, no. 159 (2018): 1–12, https://doi.org/10.4172/2157-2526.1000159; Natasha E. Bajema et al., "The Digitization of Biology: Understanding the New Risks and Implications for Governance," Center for the Study of Weapons of Mass Destruction, National Defense University, July 2018, 1–29, https://wmdcenter.ndu.edu/Publications/Publication-View/Article/1569559/the-digitization-of-biology-understanding-the-new-risks-and-implications-for-go/; Kristina Hummel, "A View from the CT Foxhole: Edward You, FBI Weapons of Mass Destruction Directorate, Biological Countermeasures Unit," *CTC Sentinel* 10, no. 7 (August 2017): 9–12, https://ctc.usma.edu/a-view-from-the-ct-foxhole-edward-you-fbi-weapons-of-mass-destruction-directorate-biological-countermeasures-unit/; Edward You and Keith G. Kozminski, "Biosecurity in the Age of Big Data: A Conversation with the FBI," *Molecular Biology of the Cell* 26, no. 22 (November 2015): 3894–97, https://doi.org/10.1091/mbc.e14-01-0027; and *National and Transnational Security Implications of Big Data in the Life Sciences*, American Association for the Advancement of Science, FBI, and U.N. Interregional Crime and Justice Research Institute (2014), https://www.aaas.org/sites/default/files/AAAS-FBI-UNICRI_Big_Data_Report_111014.pdf.

[11] Steve Lohr, "How Big Data Became so Big," *New York Times*, Aug. 11, 2012, https://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html.

[12] American Association for the Advancement of Science, FBI, and U.N. Interregional Crime and Justice Research Institute, *National and Transnational Security*, 19.

[13] The Weapons of Mass Destruction Directorate was established in 2006 to lead the FBI's efforts to mitigate threats from nuclear, chemical, biological, radiological, or explosive weapons. For more details, see "Weapons of Mass Destruction," FBI, accessed April 21 2022, https://www.fbi.gov/investigate/wmd.

head of the FBI's countermeasures unit, claimed, "Now somebody out there has the brass ring—this gigantic data set, where the only limitation is deriving the analytical tools to make all that data useful. … Whoever has the largest and most diverse data set is going to win."[14] Similar to the American Association for the Advancement of Science report, You also emphasized the roles of data acquisition and software tools, suggesting that they alone are the key limiting factors to the nefarious application of big data. In 2018, National Defense University launched a multiyear study titled, "The Digitization of Biology: Understanding the New Risks and Implications for Governance," which also called attention to the growing risks of the data that are available in genome editing. This study noted that "[m]alicious actors may be able to leverage CRISPR-Cas9 and the knowledge generated from legitimate research using pathogen genomic data to cause harm."[15] One year later, the Bipartisan Commission on Biodefense held a meeting that examined the vulnerabilities of these data sets and their potential misuse, among other biomedical data concerns.[16] And in January 2020, the National Academies of Science produced a report titled *Safeguarding the Bioeconomy,* funded by the Office of the Director of National Intelligence, which also placed attention on the vulnerabilities of cyber infrastructure related to biomedical data. In addition, the report highlighted how biomedical data sets could pose privacy risks, as well as economic and national security risks. For example, data associated with personally identifiable information, other personal health information, and genomic data sets could be leveraged for blackmail, extortion, or various types of exploitation and surveillance.[17] Thus, with this focus on big data in the life sciences, U.S. security concerns have moved from the biological materials themselves to the information generated from life science research.

These security concerns, which sit at the intersection of big data and biomedicine, have been further aggravated by several high-profile cyber security breaches at leading U.S. insurance companies. In 2015, hackers infiltrated Anthem, the second largest health insurer in the United States, and accessed a company database containing as many as 80 million records of current and former Anthem customers and employees.[18] Subsequent investigations revealed that the hackers accessed personal information such as names, member identifications, Social Security numbers, home and email addresses, and employment information. Thankfully, the hackers were not able to access credit card or patient medical information. This incident drew public and policy attention to the vulnerability of electronic health records and the direct targeting of U.S. citizens and their biomedical data, and it raised questions about whether the Health Insurance Portability and Accountability Act (HIPAA) adequately safeguarded against these kinds of cyber attacks.

Since 2015, the problem of hackers targeting and accessing the data of health and biomedical institutions has worsened.[19] According to Robert Lord, founder and chief strategy officer of the data security firm Protenus, in 2019, 32 million patient records were breached by hacking — double the amount seen in 2018.[20] The 2019 Healthcare Data Breach Report, released by the *HIPAA Journal*, found that more healthcare records were breached in 2019 than in the six years from 2009 to 2014.[21] Because of growing concerns about data protection, in 2019, the U.S. Senate Cybersecurity Caucus held a forum on cyber security threats facing the healthcare industry to further address the issue.[22]

Some have highlighted the fact that China-based hacking groups have been responsible for, or are strongly implicated in, several of these hacking incidents. For example, researchers at the security firm ThreatConnect found

[14] Kozminski, "Biosecurity in the Age," 3896.

[15] Bajema et al., "The Digitization of Biology," 4.

[16] "Cyberbio Convergence: Characterizing the Multiplicative Threat," Bipartisan Commission on Biodefense, Sept. 17, 2019, https://biodefensecommission.org/events/cyberbio-convergence-characterizing-the-multiplicative-threat/.

[17] National Counterintelligence and Security Center, "China's Collection."

[18] Reed Abelson and Matthew Goldstein, "Anthem Hacking Points to Security Vulnerability of Health Care Industry," *New York Times*, Feb. 5, 2015, https://www.nytimes.com/2015/02/06/business/experts-suspect-lax-security-left-anthem-vulnerable-to-hackers.html.

[19] National Academies of Sciences, Engineering, and Medicine, *Safeguarding the Bioeconomy*, 303–04.

[20] Jessica Davis, "Healthcare Data Hacking the New 'Space Race,' Leaders Tell Senate," Health IT Security, Aug. 7, 2019, https://healthitsecurity.com/news/healthcare-data-hacking-the-new-space-race-leaders-tell-senate.

[21] Steve Alder, "2019 Healthcare Data Breach Report," *HIPAA Journal*, Feb. 13, 2020, https://www.hipaajournal.com/2019-healthcare-data-breach-report/.

[22] "Healthcare Industry and Cybersecurity," U.S. Senate Cybersecurity Caucus, Aug. 6, 2019, https://www.c-span.org/video/?463281-1/health-care-industry-cybersecurity.

that the technical infrastructure used in the Anthem attack was linked to the Chinese computer security firm Topsec, which has strong links to China's security establishment.[23] In May 2019, a U.S. grand jury indicted two Chinese nationals for hacking Anthem and three other U.S. businesses.[24] These individuals remain at large. Additionally, research from security firm FireEye has identified multiple Chinese-linked groups that have hacked medical systems and databases around the world.[25] Others have observed how Chinese hackers have attempted to obtain data from clinical trials and scientific research studies, as well as intellectual property involving medical devices.[26] For example, in July 2020, a federal grand jury returned an indictment charging two Chinese nationals with hacking into a variety of U.S. computer systems in attempts to acquire COVID-19 research, and in some instances acting on behalf of China's Ministry of State Security.[27] The FBI has also highlighted how China may have access to other sets of large-scale biomedical data through contract work, business partnerships, and research collaborations with hospitals, universities, and biotech companies.[28] It is unclear what the specific motives are for these attacks — whether they are purely for economic or industrial gain, or whether they are for the potential creation of new bioweapons or surveillance mechanisms meant to help China gain military advantage.[29] For example, China has conducted a massive DNA collection effort of millions of its men and boys, as well as from its ethnic minority Uyghur population, as part of a growing surveillance apparatus.[30]

What is clear is that China has sought to increase its biotech capability over the past 15 years. The Chinese government has prioritized building up China's biotech industry in its 11th (2006–2010), 12th (2011–2015), and 13th (2016–2020) Five Year Plans.[31] In doing so, Beijing has made a significant effort to acquire knowledge — most notably seen in its controversial Thousand Talents Programs,[32] which were created in 2008 to recruit overseas expertise to build up China's science and technology knowledge and innovation base. Some have noted that this drive to acquire external knowledge stems from China's desire to "catch up" and become a world leader in science and technology.[33] This has led Beijing to pursue strategies that focus on short-term results, rather than on building up its own domes-

[23] Elias Groll, "The Enduring Mystery of Who Hacked Anthem," *Foreign Policy*, May 10, 2019, https://foreignpolicy.com/2019/05/10/the-enduring-mystery-of-who-hacked-anthem-hackers-spies-china/.

[24] "United States of America v. Fujie Wang and John Doe," U.S. Department of Justice, May 7, 2019, https://www.justice.gov/opa/press-release/file/1161466/download.

[25] "Beyond Compliance: Cyber Threats and Healthcare," Bank Info Security, Sept. 19, 2019, https://www.bankinfosecurity.com/whitepapers/beyond-compliance-cyber-threats-healthcare-w-5570.

[26] Matt Burgess, "China's Hackers Are Ransacking Databases for Your Health Data," *Wired*, Aug. 21, 2019, https://www.wired.co.uk/article/china-hackers-medical-data-cancer.

[27] "Two Chinese Hackers Working with the Ministry of State Security Charged with Global Computer Intrusion Campaign Targeting Intellectual Property and Confidential Business Information, Including COVID-19 Research," U.S. Department of Justice, July 21, 2020, https://www.justice.gov/opa/pr/two-chinese-hackers-working-ministry-state-security-charged-global-computer-intrusion.

[28] Christopher Wray, "Responding Effectively to the Chinese Economic Espionage Threat," FBI News, Feb. 6, 2020, https://www.fbi.gov/news/speeches/responding-effectively-to-the-chinese-economic-espionage-threat; Mark Kazmierczak et al., *China's Biotechnology Development: The Role of US and Other Foreign Engagement*, Gryphon Scientific, Feb. 14, 2019, https://www.uscc.gov/sites/default/files/Research/US-China%20Biotech%20Report.pdf; *Opportunities Exist for National Institutes of Health to Strengthen Controls in Place to Permit and Monitor Access to its Sensitive Data,* U.S. Department of Health and Human Services, Office of Inspector General, February 2019, https://oig.hhs.gov/oas/reports/region18/181809350.asp; and "Attorney General Jeff Session's China Initiative Fact Sheet," U.S. Department of Justice, Nov. 1, 2018, https://www.justice.gov/opa/speech/file/1107256/download.

[29] *Military and Security Developments Involving the People's Republic of China,* Office of the Secretary of Defense, 2021, ix, https://media.defense.gov/2021/Nov/03/2002885874/-1/-1/0/2021-CMPR-FINAL.PDF; *Biodefense in the Age of Synthetic Biology*, National Academies of Sciences, Engineering, and Medicine (2018), https://www.nap.edu/catalog/24890/biodefense-in-the-age-of-synthetic-biology; and "Commerce Acts to Deter Misuse of Biotechnology, Other U.S. Technologies by the People's Republic of China to Support Surveillance and Military Modernization that Threaten National Security," U.S. Department of Commerce, Dec. 16, 2021, https://www.commerce.gov/news/press-releases/2021/12/commerce-acts-deter-misuse-biotechnology-other-us-technologies-peoples.

[30] Wee, "China Is Collecting DNA"; and David Cyranoski, "China Expands DNA Data Grab in Troubled Western Region," *Nature*, no. 545 (May 2017): 395–96, https://doi.org/10.1038/545395a.

[31] Kazmierczak et al., *China's Biotechnology Development*; and "Major High-Tech Projects Planned for 2006–2010," *Xinhua News Agency,* March 6, 2016, http://www.china.org.cn/english/2006lh/160294.htm.

[32] James Farrer, "China Wants You: The Social Construction of Skilled Labor in Three Employment Sectors," *Asian and Pacific Migration Journal* 23, no. 4 (2014): 397–420, https://doi.org/10.1177%2F011719681402300405.

[33] Cong Cao et al., "Reforming China's S&T system," *Science* 341, no. 6145 (2013): 460–462, https://doi.org/10.1126/science.1234206; Cong Cao et al., "Reform of China's Science and Technology System in the Xi Jinping Era," *China: An International Journal* 16, no. 3 (2018): 120–41; Nirmal Kumar Chandra, "Education in China: From the Cultural Revolution to Four Modernisations," *Economic and Political Weekly*, 22, no. 19/21 (May 1987): AN121–AN136, https://www.jstor.org/stable/4377015; and Fan Yang, "Surveying China's Science and Technology Human Talents Programs," University of California San Diego, SITC Research Briefs, no. 3 (2015), https://escholarship.org/content/qt5qg340x3/qt5qg340x3.pdf.

tic innovation capabilities.[34]

The Chinese government has made biomedical big data a national priority, launching a 60 billion yuan ($9.3 billion) precision medicine initiative in 2016 to address growing diseases with genetic links in China's aging population.[35] Chinese biohacking attempts may be an effort to try to get biomedical innovation on the "quick and cheap." In light of these concerns, U.S. intelligence and law enforcement entities are keen to identify and interdict China's attempts to pursue an enhanced science and technology capability through the illicit acquisition of various kinds of biomedical data to further its commercial or security ambitions. But key questions remain: Has China actually been able to use this data for economic or security gain? If so, how? How difficult has it been for China to accomplish its goals? And how might one more accurately assess these kinds of questions? The following sections will outline the limitations in existing assessments of Chinese biomedical hacking, as well as provide signposts and a research roadmap for conducting more robust assessments of this in the future.

> **The current focus of U.S. law enforcement and U.S. policy on these data threats fails to capture the more complex character of what it takes to make biomedical big data work in practice for applied purposes.**

Framing the Chinese Biomedical Hacking Threat

To date, most of the existing law enforcement, intelligence, and policy practitioner discourse about China's threat related to acquiring U.S. biomedical big data has focused on the discrete pieces of information that are being, have been, or may be passed between the United States and China, including electronic health records, genomic data, and patient behavioral survey data. As noted above, this discourse tends to assume that, once China has accumulated enough of this data, it is only a matter of time before it outpaces the United States and becomes the new science and technology global powerhouse in the biomedical and biotechnology arena — indeed, many argue that this is imminent.[36] This rhetoric is similar to past instances, going back to the 1980s, when intelligence and policy officials pointed — wrongly — to how advances in biology and biotechnology would lead to new and growing security threats.[37] In these cases, the focus was on access to biological materials (e.g., the smallpox virus, anthrax bacteria, toxins, other pathogens, synthesized DNA), new biological techniques and technologies (e.g., genetic engineering, polymerase chain reaction, synthetic biology, genome editing, "cloud labs"), or the published materials and methods sections of scientific papers.

We have argued in previous papers that this reflects a flawed, technologically deterministic way of thinking about science and technology — one that focuses only on its material aspects and not on the tacit knowledge and other social dimensions of laboratory work that enable science and technology to

[34] Benjamin Shobert, "Priming the Pump: Applying Lessons from High-Tech Innovation to the Life Sciences in China," in *The Rise of Chinese Innovation in the Life Sciences*, ed. Xiaoru Fei, Benjamin Shobert, and Joseph Wong, National Bureau of Asian Research, Special Report no. 56, April 2016, https://www.nbr.org/wp-content/uploads/pdfs/publications/special_report_56_china_life_science_april2016.pdf; and
Yanzhong Huang, "Sino-U.S. Relations through a Life Sciences Prism," Interview from The Globalization of China's Life Science Sector, National Bureau of Asian Research, May 6, 2014, https://www.nbr.org/publication/sino-u-s-relations-through-a-life-sciences-prism/.

[35] David Cyranoski, "China Embraces Precision Medicine on a Massive Scale," *Nature*, no. 529 (January 2016): 9–10, https://doi.org/10.1038/529009a.

[36] Scott Moore, "China's Biotech Boom Could Transform Lives—Or Destroy Them," *Foreign Policy*, Nov. 8, 2019, https://foreignpolicy.com/2019/11/08/cloning-crispr-he-jiankui-china-biotech-boom-could-transform-lives-destroy-them/; National Counterintelligence and Security Center, "China's Collection"; and Christopher Wray, "China's Attempt to Influence U.S. Institutions," FBI News, Presentation to the Hudson Institute, Washington, DC, July 7, 2020, https://www.fbi.gov/news/speeches/the-threat-posed-by-the-chinese-government-and-the-chinese-communist-party-to-the-economic-and-national-security-of-the-united-states.

[37] For examples, see Philip J. Hilts, "Biological Weapons Reweighed," *Washington Post*, Aug. 17, 1986, https://www.washingtonpost.com/archive/politics/1986/08/17/biological-weapons-reweighed/10268231-f545-44c8-901f-3403d99e275b/; *Proliferation of Weapons of Mass Destruction: Assessing the Risks*, U.S. Congress Office of Technology Assessment, August 1993, https://www.princeton.edu/~ota/disk1/1993/9341/9341.PDF; Biotechnology and Genetic Engineering: Implications for the Development of New Warfare Agents, U.S. Department of Defense, 1996, https://biotech.law.lsu.edu/blaw/DOD/biotech96.pdf; "The Darker Bioweapons Future," U.S. Central Intelligence Agency, Nov. 3, 2003, https://www.cia.gov/readingroom/document/0001298811; *Addressing Biosecurity Concerns Related to Synthesis of Select Agents*, National Science Advisory Board for Biosecurity, December 2006, https://biosecurity.fas.org/resource/documents/NSABB%20guidelines%20synthetic%20bio.pdf; and James R. Clapper, *Worldwide Threat Assessment of the US Intelligence Community*, Office of the Director of National Intelligence, Feb. 9, 2016, https://www.dni.gov/files/documents/SASC_Unclassified_2016_ATA_SFR_FINAL.pdf.

work in practice in the real world.[38] There is other literature that also points to the importance of the socio-technical character and context of technology diffusion in general,[39] as well as literature on the importance of socio-technical factors in state-level nuclear weapons development,[40] military technology development,[41] and the adoption of technology by nonstate actors.[42] This collective work, which involves a variety of case studies, demonstrates the importance of considering the social dimensions of science and technology research, and it emphasizes how focusing only on the material aspects of technology acquisition and development is an erroneous way of thinking about what it takes for state or nonstate actors to develop science and technology capabilities. Such misunderstandings of technology have led to numerous scholarly, intelligence, and policy failures in understanding biosecurity threats. This has included the flawed assessments of the Soviet and Iraqi bioweapons programs, the overhyped "bioterrorism threat" since the 1990s, as well as various current biosecurity concerns.

In today's concern about biomedical big data, we see the focus again being placed on the material — this time biomedical data — with claims being made that mere access to troves of biomedical data poses new and alarming security risks. The current focus of U.S. law enforcement and U.S. policy on these data threats fails to capture the more complex character of what it takes to make biomedical big data work in practice for applied purposes.

There are, however, a variety of bioinformatics

[38] Kathleen M. Vogel, *Phantom Menace or Looming Danger? A New Framework for Assessing Bioweapons Threats* (Baltimore, MD: The Johns Hopkins University Press, 2013); and Sonia Ben Ouagrham-Gormley, *Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development* (Ithaca, NY: Cornell University Press, 2014.)

[39] Everett M. Rogers, *Diffusion of Innovations*, 5th ed. (New York: Free Press, 2003); Clayton M. Christensen, Scott D. Anthony, and Erik A. Roth, *Seeing What's Next: Using the Theories of Innovation to Predict Industry Change* (Boston: Harvard Business School Press, 2004), 20–21, 290; David S. Landes, *The Unbound Prometheus: Technical Change and Industrial Development in Western Europe from 1750 to the Present* (Cambridge: Cambridge University Press, 2003), 19; Joel Mokyr, *The Lever of Riches: Technological Creativity and Economic Progress* (New York: Oxford University Press, 1990)180–181; Henry Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting from Technology* (Boston: Harvard Business School Press, 2003); W. Brian Arthur, *The Nature of Technology: What It Is and How It Evolves* (New York: Free Press, 2009), 108; Donald MacKenzie, *Knowing Machines: Essays on Technical Change* (Cambridge, MA: MIT Press, 1998); and Joel Mokyr, *The Gifts of Athena: Historical Origins of the Knowledge Economy* (Princeton, NJ: Princeton University Press, 2002).

[40] Lucky Asuelime and Suzanne Francis, "Drivers of Nuclear Proliferation: South Africa's Incentives and Constraints," *Southern Journal for Contemporary History* 39, no. 1 (2014): 55–68, https://journals.ufs.ac.za/index.php/jch/article/view/275; Hal Brands and David Palkki, "Saddam, Israel, and the Bomb: Nuclear Alarmism Justified?" *International Security* 36, no. 1 (Summer 2011):133–66, https://doi.org/10.1162/ISEC_a_00047; Malfrid Braut-Hegghammer, *Unclear Physics: Why Iraq and Libya Failed to Build Nuclear Weapons* (Ithaca, NY: Cornell University Press, 2016); Peter Dombrowski and Eugene Gholz, "Identifying Disruptive Innovation: Innovation Theory and the Defense Industry," *Innovations: Technology, Governance, Globalization* 4, no. 2 (Spring 2009): 101–17, https://doi.org/10.1162/itgg.2009.4.2.101; Brian R. Early, "Exploring the Final Frontier: An Empirical Analysis of Global Civil Space Proliferation," *International Studies Quarterly* 58, no. 1 (March 2014): 55–67, https://www.jstor.org/stable/24017846; Matthew Fuhrmann and Michael C. Horowitz, "When Leaders Matter: Rebel Experience and Nuclear Proliferation," *Journal of Politics* 77, no. 1 (January 201): 72–87, https://doi.org/10.1086/678308; and Jacques E. C. Hymans, "No Cause for Panic: Key Lessons from the Political Science Literature on Nuclear Proliferation," *International Journal* 69, no. 1 (March 2014): 85–93, https://doi.org/10.1177%2F0020702014521565; Suzanne C. Buono, *Demystifying Nuclear Proliferation: Why States Do What They Do,* PhD diss., Johns Hopkins University (2011); and Garill A. Coles et al., *Utility of Social Modeling in Assessment of a State's Propensity for Nuclear Proliferation*, Pacific Northwest National Laboratory, Report Prepared for the U.S. Department of Energy, June 2011, https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-20492.pdf.

[41] Adam M. Jungdahl and Julia M. Macdonald, "Innovation Inhibitors in War: Overcoming Obstacles in the Pursuit of Military Effectiveness," *Journal of Strategic Studies* 38, no. 4 (2015): 467–99, https://doi.org/10.1080/01402390.2014.917628; Williamson Murray, *Military Adaptation in War: With Fear of Change* (Cambridge: Cambridge University Press, 2011, online publication (June 2014), https://doi.org/10.1017/CBO9781139005241; Richard A. Bitzinger, "China's Defense Technology and Industrial Base in a Regional Context: Arms Manufacturing in Asia," *Journal of Strategic Studies* 34, no. 3 (2011): 425–50, https://www.tandfonline.com/doi/abs/10.1080/01402390.2011.574985; and Tai Ming Cheung, "The Chinese Defense Economy's Long March From Imitation to Innovation," *Journal of Strategic Studies* 34, no. 3 (2011): 325–54, https://www.tandfonline.com/doi/full/10.1080/01402390.2011.574976.

[42] Brian A. Jackson, "Technology Acquisition by Terrorist Groups: Threat Assessment Informed by Lessons from Private Sector Technology Adoption," *Studies in Conflict and Terrorism* 24, no. 3 (May 2001): 183–213, https://doi.org/10.1080/10576100151130270; Brian A. Jackson and David R. Frelinger, "Rifling through the Terrorists' Arsenal: Exploring Groups' Weapon Choices and Technology Strategies," *Studies in Conflict and Terrorism* 31, no. 7 (2008): 583–604, https://doi.org/10.1080/10576100802159989; Michael C. Horowitz, "Nonstate Actors and the Diffusion of Innovations: The Case of Suicide Terrorism," *International Organization* 64, no. 1 (January 2010): 33–64, https://doi.org/10.1017/S0020818309990233; Evan Perkoski, "Terrorist Technological Innovation," in *The Oxford Handbook of Terrorism,* ed. Erica Chenoweth et al. (Oxford: Oxford University Press, 2019); Gary Ackerman, *'More Bang for the Buck': Examining the Determinants of Terrorist Adoption of New Weapons Technologies,* PhD dissertation, King's College London, 2014, 15, https://kclpure.kcl.ac.uk/portal/files/32901277/2014_Ackerman_Gary_0715371_ethesis.pdf; Gary A. Ackerman, ed., "'Designing Danger': Complex Engineering by Violent Non-State Actors," *Journal of Strategic Security* 9, no. 1 (Special Issue, Spring 2016), https://digitalcommons.usf.edu/jss/vol9/iss1/; Andrew Mumford, ed., "How Terrorists 'Learn': Innovation and Adaptation in Political Violence," *British Academy Review*, no. 26 (2017), https://www.thebritishacademy.ac.uk/documents/728/BAR26-08-Mumford.pdf; Brian A. Jackson et al., eds., *Aptitude for Destruction, Volume 1: Organizational Learning in Terrorist Groups and Its Implications for Combating Terrorism* (Santa Monica, CA: RAND, 2007), https://www.rand.org/content/dam/rand/pubs/monographs/2005/RAND_MG331.pdf; Brian A. Jackson et al., *Aptitude for Destruction, Volume 2: Case Studies of Organizational Learning in Five Terrorist Groups* (Santa Monica, CA: RAND, 2007), https://www.rand.org/pubs/monographs/MG332.html; James J. F. Forest, ed., *Teaching Terror: Strategic and Tactical Learning in the Terrorist World* (Lanham, MD: Rowman and Littlefield, 2006); Kim Cragin et al., *Sharing the Dragon's Teeth: Terrorist Groups and the Exchange of New Technologies* (Santa Monica, CA: RAND, 2007), https://doi.org/10.7249/MG485; Michael Kenney, *From Pablo to Osama: Trafficking and Terrorist Networks, Government Bureaucracies, and Competitive Adaptation* (University Park, PA: Penn State University Press, 2008); Michael Kenney, "'Dumb' Yet Deadly: Local Knowledge and Poor Tradecraft Among Islamist Militants in Britain and Spain," *Studies in Conflict and Terrorism* 33, no. 10 (2010): 911–32, https://www.tandfonline.com/doi/full/10.1080/1057610X.2010.508508; and Milton Leitenberg, "Aum Shinrikyo's Efforts to Produce Biological Weapons: A Case Study in the Serial Propagation of Misinformation," *Terrorism and Political Violence* 11, no. 4 (Winter 1999): 149–58, https://www.tandfonline.com/doi/abs/10.1080/09546559908427537.

and big data researchers who provide an alternative understanding and framework for how to think about these biomedical data threats.[43] Their research focuses on the methods and assumptions involved in the use of biomedical big data for the discovery of new drugs and therapies, the socio-technical challenges of extracting knowledge from digital infrastructures, and the implications of choices in data curation for applications in science and technology. This body of work is consistent with several bioinformatics papers that discuss the challenges of working with heterogeneous biomedical big data: It is not a trivial task to harness these data for either useful or nefarious applications.[44] There are often errors associated with this data or other data quality issues that require substantial data curation and preparation before they can be used.[45] As other researchers have noted, “Data heterogeneity, data protection, analytical flows in analyzing data and the lack of appropriate infrastructures for data storage emerged as critical technical and infrastructural issues that might endanger a Big-Data-driven healthcare.”[46] The crux of this scholarship focuses on the challenges involved in creating, transferring, and using data for the production of knowledge that can lead to biomedical and biotechnology applications.

For far too long, the U.S. security community has jumped to assumptions about how easily data or materials (usually related to emerging technologies) can be translated to security threats. Usually, the standard pieces of evidence used to make those judgments are generic references to the technology (or assumed trends of the technology) without relying on rigorous, real-world empirical data and studies of the various social and technical factors involved in shaping the development and use of that data or technology. To better inform decision-makers, we need new research questions and a committed funding stream to support research agendas focused on the socio-technical dimensions of biomedical big data to guide intelligence collection and analysis. This would enable analysts to parse out more carefully how or under what conditions actors can utilize biomedical big data to pose economic or security threats to the United States. The next section summarizes some key bioinformatics scholarship that offers a more nuanced understanding of how to think about biomedical big data. It also provides some useful signposts for how to think about crafting better intelligence and law enforcement assessments of the Chinese threat in this domain.

Biomedical Big Data: Definitions and Challenges

Big data scholar Sabina Leonelli provides useful definitions of data and knowledge and of the relationship between the two.[47] She defines data as “mobile pieces of information” that can be collected, stored, and disseminated. These data can be accurate, or not, and they can be used for an applied purpose, or not. Using these data to produce knowledge that is deemed reliable, accurate, and useful for an applied purpose depends on the various people involved in collecting, storing, analyzing, and interpreting the data for some practical use. The value of data does not come from their intrinsic nature. Leonelli describes the value as coming “from their interpretation in relation to specific contexts and goals, rather than as a context-independent quality.”[48] Leonelli points out that “[d] ata are not, by themselves, a form of knowledge. Rather, data need to be interpreted in order to yield knowledge.”[49] In short, she is focused on the practices and processes for making sense of data. What

[43] Bioinformatics is defined as the collection, classification, storage, and analysis of biochemical and biological information using computers.

[44] Mattia Prosperi et al., "Big Data Hurdles in Precision Medicine and Precision Public Health," *BMC Medical Informatics and Decision Making*, no.18 (2018), https://doi.org/10.1186/s12911-018-0719-2; Yixue Li and Luonan Chen, "Big Biological Data: Challenges and Opportunities," *Genomics, Proteomics and Bioinformatics* 12, no. 5 (October 2014): 187–89, http://dx.doi.org/10.1016/j.gpb.2014.10.001; and Choong Ho Lee and Hyung-Jin Yoon, "Medical Big Data: Promise and Challenges," *Kidney Research and Clinical Practice* 36, no. 1 (2017): 3–11, https://doi.org/10.23876/j.krcp.2017.36.1.3.

[45] André Freitas and Edward Curry, "Big Data Curation," in *New Horizons for a Data-Driven Economy,* ed. José Maria Cavanillas, Edward Curry, and Wolfgang Wahlster (New York: Springer, 2016), 87–118; Dylan Ruediger et al., "Big Data Infrastructure at the Crossroads: Support Needs and Challenges for Universities," Ithaka S+R, Dec. 1, 2021, https://doi.org/10.18665/sr.316121; Patrick Fahr, James Buchanan, and Sarah Wordsworth, "A Review of the Challenges of Using Biomedical Big Data for Economic Evaluations of Precision Medicine," *Applied Health Economics and Health Policy* 17, no. 4 (2019): 443–52, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6647451/; and Saravanan Thirumuruganathan et al., "Data Curation with Deep Learning," *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, March 30–April 2, 2020, https://openproceedings.org/2020/conf/edbt/paper_142.pdf.

[46] Roberta Pastorino et al., "Benefits and Challenges of Big Data in Healthcare: An Overview of the European Initiatives," *European Journal of Public Health* 29, no. 3 (2019): 23–27, https://pubmed.ncbi.nlm.nih.gov/31738444/.

[47] Leonelli, "Integrating Data," 505.

[48] Sabina Leonelli, "On the Locality of Data and Claims About Phenomena," *Philosophy of Science* 76, no. 5 (December 2009): 737–49, https://doi.org/10.1086/605804.

[49] Leonelli, "Integrating Data," 505.

Leonelli's work and that of other bioinformatic researchers in this domain usefully document is how difficult these processes are when it comes to working with biomedical big data.[50] This is due to one key characteristic and challenge of biomedical data — its heterogeneity — which creates difficulties for data sharing and for the assessment, interpretation, and application of data.

The National Institutes of Health defines biomedical big data as inclusive of the "numerous quantitative and qualitative datasets emanating from fundamental research using model organisms (e.g., mice, fruit flies, zebrafish), clinical studies (including medical images), and observational and epidemiological studies (including data from electronic health records and wearable devices)."[51] They can include imaging, phenotypic, epigenetic, genotypic, molecular, clinical, behavioral, environmental, and many other types of biological or medical data, and they can encompass metadata (i.e., data that describe other data), such as the title, abstract, author, and keywords in publications; the organization and relationships of digital materials; file types or modification dates; and the data standards and software tools involved in data processing and analysis.[52] Biomedical data can also be qualitative, such as patient medical narratives.

Not only is there a lot of heterogeneity in biomedical big data, there is also heterogeneity in the ways that this type of data is created and stored. The data can be stored using different labels and can consist of varying types and file formats. This depends on the kind of experiment that was conducted, the origins of the data, the way in which the data were collected and generated, and the equipment used to generate them, which can vary significantly from one laboratory or location to the next. The life and biomedical sciences are extremely diverse in their experimental methods, goals, instruments, and conceptual frameworks.[53] Often, different research groups — even within the same subfield — disagree over preferred terminology, research organisms, and experimental methods and protocols.[54] This extreme diversity is reflected in the various methods used to generate, store, share, and analyze biomedical data,[55] meaning that data are not standardized in the life and biomedical sciences. This is true for all current open- or private-source biomedical big data.[56] Although there are continual efforts being made to solve these standardization problems, they will continue to pose challenges for data scientists.

This is a key difference from other types of scientific big data, such as the data coming from the particle physics community (e.g., the European Organization for Nuclear Research). These physical scientists agreed to dedicate significant time, labor, and funding to provide a standardized means of collecting and storing data, and they located the research facilities in just a few places in order to foster collaboration from the outset and share the high costs of running these laboratories.[57] This approach has made working with this kind of data much easier, because scientists do not have to do the extra work of translating the data from one setting to another. In contrast, the life science and biomedical communities consist of thousands of laboratories (clinical, academic, private, commercial, and industrial) around the globe that have been set up with no common standardization and that use

[50] See also Tim Hulsen et al., "From Big Data to Precision Medicine," *Frontline Medicine* 6, no. 34 (2019): 1–14, https://pubmed.ncbi.nlm.nih.gov/30881956/; Fahr et al., "A Review of the Challenges"; Blagoi Ristevski and Ming Chen, "Big Data Analytics in Medicine and Healthcare," *Journal of Integral Bioinformatics* 15, no. 3 (2018): 1–5, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6340124/; Karen Y. He, Dongliang Ge, and Max M. He, "Big Data Analytics for Genomic Medicine," *International Journal of Molecular Science* 18, no. 2 (2017): 1–18, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343946/; David J. Duffy, "Problems, Challenges and Promises: Perspectives on Precision Medicine," *Briefings in Bioinformatics* 17, no. 3 (May 2016): 494–504, https://pubmed.ncbi.nlm.nih.gov/26249224/; and N. Peek, J. H. Holmes, and J. Sun, "Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics," *Yearbook of Medical Informatics* 9, no. 1 (2014): 42–47, https://pubmed.ncbi.nlm.nih.gov/25123720/.

[51] "NIH Strategic Plan for Data Science," National Institutes of Health, June 2018, 1, https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.

[52] National Institutes of Health, "NIH Strategic Plan," 1, 30.

[53] Vivien Marx, "The Big Challenges of Big Data," *Nature*, no. 498 (2013), 255–60, https://doi.org/10.1038/498255a; Subhajit Pal et al., "Big Data in Biology: The Hope and Present-Day Challenges in It," *Gene Reports*, no. 21 (December 2020), https://doi.org/10.1016/j.genrep.2020.100869; and Sabina Leonelli, "Philosophy of Biology: The Challenges of Big Data Biology," *eLife*, April 5, 2019, 1–5, https://elifesciences.org/articles/47381.

[54] Sabina Leonelli, "When Humans Are the Exception: Cross-Species Databases at the Interface of Biological and Clinical Research," *Social Studies of Science* 42, no. 2 (2012): 214–36, https://doi.org/10.1177/0306312711436265.

[55] Maureen A. O'Malley and Orkun S. Soyer, "The Roles of Integration in Molecular Systems Biology," *Studies in History and Philosophy of Science, Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (March 2012): 58–68, https://doi.org/10.1016/j.shpsc.2011.10.006.

[56] Fahr et al., "A Review of the Challenges"; Subyasachi Dash et al., "Big Data in Healthcare: Management, Analysis and Future Prospects," *Journal of Big Data*, no. 6 (2019), https://doi.org/10.1186/s40537-019-0217-0; Ristevski and Chen, "Big Data Analytics"; and *Planning for Long-Term Use of Biomedical Data*, National Academies of Sciences, Engineering, and Medicine (2020), https://www.nap.edu/catalog/25707/planning-for-long-term-use-of-biomedical-data-proceedings-of.

[57] Marx, "The Big Challenges," 255–60.

different kinds of equipment to capture and record the data, and both the equipment and data have not necessarily been standardized across labs.

Thus, integrating these diverse data for applied knowledge requires significant labor and expertise, often involving the development and application of new tools, standards, methods, and infrastructures, which, in turn, requires a large amount of "conceptual and material scaffolding" to transform the data into something useful.[58] This extreme diversity around the world makes the transfer of data from one location (where it was generated) to another for reuse very challenging without understanding how the data were generated, handled, and stored.

Leonelli calls one of the challenges of working with heterogeneous data the difference between propositional and embodied knowledge — or what we and others before us have referred to as the importance of differentiating between explicit versus tacit knowledge — which is a key analytic construct in the field of science and technology studies.[59] Explicit (or propositional) knowledge is information that can be codified, for example, in a written protocol or in a database. Tacit (or embodied) knowledge is hands-on, skills-based knowledge that is difficult to codify or, in some cases, cannot be codified. A laboratory may have a written protocol for how data were generated and may have a particular dataset for a given experiment. However, a whole host of tacit knowledge is involved in making that protocol generate the data that is produced in that laboratory and in being able to understand and work with that data.[60] A different laboratory might use a different method for the protocol or have no expertise with the protocol or those data practices, or it might use different equipment, which could yield different data results. There are also different kinds of tacit knowledge related to how data scientists in different laboratories choose to process data. This challenge of tacit knowledge and nonstandardized laboratory practices in the life sciences leads to one of the strengths — and accompanying weaknesses — of trying to glean insights from biomedical big data. If one were able to standardize and integrate all of these different kinds of data, it could lead to major breakthroughs — this is the underlying premise and promise of precision medicine.[61] However, the "if" constitutes one of the great challenges that would have to be solved by China or any other state or entity that wishes to capitalize on the troves of current and future biomedical big data, because the process of standardizing and integrating this information is not trivial.

Given the heterogeneity and nonstandardization of biomedical big data, the sharing and integration of data can be a problem. To make the most of biomedical big data, one would need to combine different types of information, such as genomics data, clinical research, behavioral studies, environmental studies, and so forth. As pointed out above, all of these data streams have been captured using different methods, procedures, and formats, usually without the recognition of the need to integrate them with other data sets. Therefore, significant work is required to be able to integrate data sets that were never meant to "talk" to one another.[62] For example, many data repositories, particularly in the healthcare field, were designed and built in the pre-big-data era and were made to stand alone and be siloed, with no intention of allowing the data to be combined and analyzed with other data sets.[63] In other cases, a single healthcare database may be composed of over 100 different interlinked data systems, all of which have their own ways of

[58] Leonelli, "Integrating Data," 504.

[59] Leonelli, "Integrating Data," 506; Vogel, *Phantom Menace or Looming Danger?*; Ben Ouagrham-Gormley, *Barriers to Bioweapons;* H. M. Collins, *Changing Order: Replication and Induction in Scientific Practice* (Chicago: University of Chicago Press, 1985); and Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy* (London: Routledge and Kegan Paul, 1958).

[60] For some examples in the life sciences, see W. Nicholson Price II, Arti K. Rai, and Timo Minssen, "Knowledge Transfer for Large-Scale Vaccine Manufacturing," Science 369, no. 6506 (21 August 2020): 912–14, https://doi.org/10.1126/science.abc9588; Noriko Hara et al., "Learning Tacit Knowledge in Life Science Graduate Programs in Taiwan," *Proceedings of the 73rd American Society for Information Science and Technology*, no. 47 (2010): 1–5, https://dl.acm.org/doi/10.5555/1920331.1920460; Kathleen M. Vogel, "Framing Biosecurity: An Alternative to the Biotech Revolution Model?" *Science and Public Policy* 35, no. 1 (February 2008): 45–54, https://doi.org/10.3152/030234208X270513; Alice M. Sapienza and Joseph G. Lombardino, "Recognizing, Appreciating, and Capturing the Tacit Knowledge of R&D Scientists," *Drug Development Research* 57, no. 2 (October 2002): 51–57, https://doi.org/10.1002/ddr.10109; Bruno Latour and Steve Woolgar, *Laboratory Life: The Construction of Scientific Facts* (Beverly Hills, CA: Sage, 1979); Karin Knorr Cetina, *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, MA: Harvard University Press, 1999); Kathleen Jordan and Michael Lynch, "The Sociology of a Genetic Engineering Technique: Ritual and Rationality in the Performance of the 'Plasmid Prep,'" in *The Right Tools for the Job: At Work in the 20th-Century Life Sciences,* ed. Adele E. Clarke and Joan H. Fujimura (Princeton, NJ: Princeton University Press, 1992): 77–114.

[61] Ricardo L. Rossi and Renata M. Grifantini, "Big Data: Challenge and Opportunity for Translational and Industrial Research in Healthcare," *Frontiers in Digital Humanities*, May 31, 2018, https://doi.org/10.3389/fdigh.2018.00013.

[62] It is possible to derive value from single data sets or a more limited set of biomedical data that do not involve this complexity. However, the promise of precision medicine is in the combination of a wide range of biomedical data sets.

[63] Hulsen et al., "From Big Data," 3.

collecting and storing information.[64] Genomic data pose unique challenges. Researchers are currently facing substantial problems in storing, managing, manipulating, analyzing, and interpreting whole genome sequence data for even relatively small numbers of individuals, especially if they must also take into account data quality information (e.g., errors or biases in the data).[65] Moreover, structured data (i.e., those that can be stored in spreadsheets) do not necessarily tell you what you need to know about a particular experiment or biomedical process. This information is often stored as unstructured data, such as in the narrative of a journal article, or is embodied in the heads and hands of practicing scientists. Unstructured data are more difficult to extract but may be critical to making sense of the structured data that one might have acquired. Lawrence Hunter, a computational biologist, sums up the challenge: "Getting the most from the data requires interpreting them in light of all the relevant prior knowledge."[66] Thus, it is not enough merely to access the data. One must also know how to work with and make sense of them in light of prior data.

The process of curating data — the organization and integration of data — is also a significant issue that needs to be addressed in order to make sense of, or apply, big data. Curation involves several complex tasks, including selecting the data that are to be assimilated into a database; formatting them into a standard that can be read by the available software; classifying them into retrievable categories, to make it possible to "mine" them, according to whichever biological question is being asked; and displaying them in ways that make it possible to spot meaningful patterns.[67] How data are curated also shapes the

[64] Jake Luo et al., "Big Data Application in Biomedical Research and Health Care: A Literature Review," *Biomedical Informatics Insights*, no. 8 (2016): 1–10, https://journals.sagepub.com/doi/pdf/10.4137/BII.S31559.

[65] He et al., "Big Data Analytics," 4; Sharona Hoffman and Andy Podgurski, "The Use and Misuse of Biomedical Data: Is Bigger Really Better?" *American Journal of Law & Medicine* 39, no. 4 (December 2013): 497–538, https://doi.org/10.1177/009885881303900401; and Saveli I. Goldberg, Andrzej Niemierko, and Alexander Turchin, "Analysis of Data Errors in Clinical Research Databases," *AMIA Annual Symposium Proceedings* (2008): 242–46, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656002/.

[66] Marx, "The Big Challenges," 257.

[67] Doug Howe et al., "Big Data: The Future of Biocuration," *Nature*, no. 455 (2008): 48–50, https://dx.doi.org/10.1038%2F455047a; Sabina Leonelli, "Packaging Small Facts for Re-Use: Databases in Model Organism Biology," in *How Well Do Facts Travel? The Dissemination of Reliable Knowledge,* ed. Peter Howlett and Mary S. Morgan (Cambridge: Cambridge University Press, 2010), 325–48; and Sabina Leonelli, "Data Interpretation in the Digital Age," *Perspectives on Science* 22, no. 3 (Fall 2014): 397–414, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340525/.

analytic outputs because, in some cases, curation may introduce errors, leading to data that may or may not be based upon the original experiments — or upon reality.[68] These errors can be difficult to see and identify if one has only a spreadsheet of data, with no prior documentation or knowledge of how the data were processed.

Similarly, interpreting or assessing data that have been acquired (whether through licit or illicit means) is also an issue. How does one know if the data are any good or even trustworthy? One would need to know the context in, conditions under, and purposes for which the data were collected, processed, and stored. This is no different than trusting the conditions under which an experiment was conducted,[69] but in this case, it would mean trusting the digital data when one does not have direct access to the laboratory expertise that generated that data. There have been several accounts noting the difficulty of reproducing the results of an experiment involving biomedical big data and digital medicine.[70] Some involved the mislabeling of data, administrative errors in the input of patient data, or other errors that can be introduced while cleaning, integrating, and processing the data,[71] not to mention the problems that may underpin the experimental and publication process itself.[72] In addition, data curators (who are different from the scientists conducting the experiment) may bring their own biases and errors to the labeling and processing of data, which would, in turn, yield unreliable data.[73]

Moreover, according to Leonelli, "how data are interpreted often changes depending on the skills, background knowledge, and circumstances of the researchers involved."[74] In the area of analysis, there are also numerous examples that caution researchers about the problem of establishing causal links in data — links that are either wrong (due to the aforementioned data acquisition, curation, and processing problems) or of limited value, if one does not also understand the explanatory mechanism of the causal link.[75] As a result of these factors, Leonelli writes:

> Data can be used to represent various aspects of reality and each interpretation will depend on the specific circumstances of analysis, including the skills and technical premises that allow people and/or algorithms to organize and visualize data in a way that corroborates a certain conceptualization of reality. In other words, the interpretation of data is constantly mediated by the view point and abilities of those using it.[76]

The key point here is that understanding big data — and we would extend this to understanding the security threats of biomedical big data — is always related to understanding the social context of science and technology: Who collected and curated the data? What are their skills and expertise? How did they collect the data? Under what conditions and in what context did they do so? How did they analyze and interpret the data? How did they store the data? These same questions can also be posed about individuals receiving these data from another laboratory. But these individuals must also answer another question: What is needed to translate the data so that it will "work" in a new context? As the field of science and technology studies has long noted, there are many factors that can shape how data are collected, adopted, and used. For example, adherence to a particu-

[68] Ankush Sharma and Giovanni Colonna, "System-Wide Pollution of Biomedical Data: Consequence of the Search for Hub Genes of Hepatocellular Carcinoma Without Spatiotemporal Consideration," *Molecular Diagnosis and Therapy* 25, no. 1, (2021): 9–27, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7847983/.

[69] James Bogen and James Woodward, "Saving the Phenomena," *Philosophical Review* 97, no. 3 (July 1988): 303–52, https://doi.org/10.2307/2185445.

[70] Aaron Stupple, David Singerman, and Leo Anthony Celi, "The Reproducibility Crisis in the Age of Digital Medicine," *NPJ Digital Medicine*, no. 2 (2019): 1–2, https://doi.org/10.1038/s41746-019-0079-z; and John P. A. Ioannidis, "How to Make More Published Research True," *PLOS Medicine*, Oct. 21, 2014, https://doi.org/10.1371/journal.pmed.1001747.

[71] Riccardo Bellazzi, "Big Data and Biomedical Informatics: A Challenging Opportunity," *Yearbook of Medical Informatics* 23, no. 1 (2014): 8–13, https://doi.org/10.15265/iy-2014-0024; Steven N. Goodman, Daniele Fanelli, and John P.A. Ioannidis, "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8, no. 341 (June 2016): 341, https://stm.sciencemag.org/content/8/341/341ps12.

[72] Monya Baker, "1,500 Scientists Lift the Lid on Reproducibility," *Nature*, no. 533 (2016), https://www.nature.com/articles/533452a; William G. Kaelin, Jr., "Publish Houses of Brick, not Mansions of Straw," *Nature* 545, no. 7655 (2017): 387, https://doi.org/10.1038/545387a; and Jean-Baptist Poline, "Reproducibility Issues in Neuroscience and Neuroimaging," May 14, 2020, YouTube, https://www.youtube.com/watch?v=WphUnnbpIsw.

[73] Leonelli, "Data Interpretation," 397–414.

[74] Sabina Leonelli, "Data Governance Is Key to Interpretation: Reconceptualizing Data in Data Science," *Harvard Data Science Review* 1, no. 1 (Summer 2019): 3, https://doi.org/10.1162/99608f92.17405bb6.

[75] Cristian S. Calude and Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* 22, no. 3 (2017): 595–612, https://www.di.ens.fr/users/longo/files/BigData-Calude-LongoAug21.pdf; and David Leinweber, "Stupid Data Miner Tricks: Overfitting the S&P 500," *The Journal of Investing* 16, no. 1 (2007): 15–22, https://www.researchgate.net/publication/247907373_Stupid_Data_Miner_Tricks_Overfitting_the_SP_500.

[76] Leonelli, "Data Governance Is Key," 4.

lar theoretical paradigm shapes the interpretation of what counts as relevant or important scientific data.[77] How data are labeled and classified also shapes what can be understood and what counts as valid or important.[78] In addition, the production, transfer, and use of scientific and technical data are often messy and unlike their clean and orderly public depiction.[79] With these issues in mind, historian of science Ted Porter cautions that the "detachment of data from the concrete conditions of its production is always risky. Data, as it moves, is most often thinned, and what is thinned is necessarily transformed."[80] Bioinformatician David Duffy has summarized the spectrum of challenges of working with biomedical big data when it comes to producing, processing, filtering, reviewing, validating, interpreting, and applying the data — all of these steps can create bottlenecks and are points at which errors could be introduced into the data or analysis.[81]

To be sure, the U.S. and international bioinformatics communities are working hard to overcome these bottleneck problems. However, another problematic claim that proponents of biomedical big data and precision medicine make is that the availability of more data at all of these levels is always beneficial and makes things easier. A corollary to this premise can be found in the U.S. security community, which argues that a growing abundance of biomedical big data can readily translate into new and dangerous security threats. However, Duffy cautions that "[m]ore data do not necessarily translate into more knowledge; rather, it can mean an increase in noise."[82] Data scientists are well aware that more data is creating even more complex data ecosystems to curate, manage, and navigate.[83] Whether biomedical big data can translate into the touted benefits described by precision medicine advocates, or whether it translates into more and varied kinds of security threats, depends on the processes that are used to make sense of such data. We need to get beyond the hype of big data — including biomedical big data — to a more grounded understanding of how state or nonstate actors are able to use these data in practice.[84]

We need to get beyond the hype of big data — including biomedical big data — to a more grounded understanding of how state or nonstate actors are able to use these data in practice.

How to Improve Intelligence Collection and Analysis on Biomedical Big-Data Threats: Focus on the "Data Journey"

With this understanding of biomedical big data in mind, Leonelli discusses the importance of studying what she calls the "data journey," which is composed of the various social factors, infrastructures, and work involved in data traveling and being used in new contexts — in essence, studying all the ways in which data are produced, transformed, and used to address a given problem.[85] This research agenda would involve focusing on the various technologies, materials, infrastructures, people, social settings, and institutions involved in the production and transfer of biomedical data, and the conditions and contexts that make the transfer and use of data more or less stable.[86] Focusing on the data journey as the unit of analysis would provide a more realistic understanding of how data could be used for an actor's economic or security benefit and would help to mitigate errors associated with focusing on the data alone.

To apply this approach to U.S. intelligence and law enforcement would mean going beyond a focus merely on apprehending people who are stealing

[77] Norwood Russell Hanson, *Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science* (Cambridge: Cambridge University Press, 1965); and Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1970).

[78] Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT Press, 1999).

[79] Latour and Woolgar, *Laboratory Life*; and Michael Lynch, *Art and Artefact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory* (London: Routledge and Kegan Paul, 1985).

[80] Theodore M. Porter, "Most Often, What Is Transmitted Is Transformed," in *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini (Heidelberg/London: Springer Open, 2020), 235.

[81] Duffy, "Problems, Challenges, and Promises," 502.

[82] Duffy, "Problems, Challenges, and Promises," 502.

[83] For a broad overview of the complex ecosystems and bioeconomies of biological and biomedical big data, see Bronwyn Parry and Beth Greenhough, *Bioinformation* (Bristol, UK: Policy Press, 2017).

[84] Jack Wilkinson et al., "Time to Reality Check the Promises of Machine Learning–Powered Precision Medicine," *Lancet* 2, no. 12 (December 2020): E677–E680, https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30200-4/fulltext.

[85] Sabina Leonelli, "Learning from Data Journeys," in *Data Journeys in the Sciences*, ed. Leonelli and Tempini, 1–24.

[86] Leonelli, "Learning from Data Journeys," 7.

or hacking data, and instead focusing on the data journey. If China is thought to be the key threat in this area, then U.S. agencies should try to better understand what the data journey would look like in order for China to use a particular biomedical big data set. One way to do this would be by looking at one of the U.S. biomedical data sets that has already been hacked. In a case like the 2015 Anthem hacking incident, this would involve getting a much more finely grained understanding of the data that was hacked, including determining in what forms and formats the data were stored; how Anthem stored its data differently from other entities (i.e., what is specific to Anthem's data storage, management, and usage and what could be applied to other hacking cases); how exactly this particular data could be used or combined with other data sets for an actor's economic or security benefit; what was involved in accessing this information; what would be required to use this data; and what challenges or limitations would exist in handling this heterogeneous data.

This kind of data journey inquiry would also require a better understanding of China's bioinformatics infrastructure. This would include determining who and where the people and industries in China's bioinformatics community are and what their capabilities are, as well as how China has been able to deal with the problems of building a bioinformatics research community.[87] Creating a bioinformatics capacity has also proved highly challenging for other countries, including the United States. Intelligence analysts, law enforcement officials, and researchers will also need to investigate how China's bioinformatics community of experts is — or is not — overcoming the inherent problems that plague China's overall science and technology infrastructure.[88] Are they able to bring together diverse sets of scientists and other technical experts, such as experimental biologists, clinicians, bioinformaticians, computer scientists, and engineers, to work on bioinformatics problems in innovative ways, rather than just replicate existing work? In addition, it will be important to determine exactly what advances China is making in the bioinformatics domain and what problems they have encountered in using and translating biomedical big data. How have they solved (or not solved) some of the problems of bottlenecking discussed above and what problems remain? In addition, government and academic analysts should investigate what government policies and programs, industries, and other institutions have been involved in these efforts and what funding streams are devoted to solving these problems.

Probing these kinds of research questions would provide a more nuanced understanding of what kind of indigenous innovation is occurring within China's bioinformatics community. As one group of science and technology scholars has written, "[T]he ease of mastering foreign technological knowledge increases with the capability of the country in indigenous innovation. The buying of foreign technology is one thing but then being able to use it fully is another."[89]

The above information could be gathered through a variety of open-source means, such as Chinese and international bioinformatics conferences, scientific publications, industry activities, and clinical trials, as well as clandestine means. If academia, intelligence analysts, and law enforcement could collect and analyze data relevant to answer these kinds of research questions, they would acquire a much better understanding of what China's science and technology capabilities are when it comes to actually using the biomedical big data that they acquire. And this kind of data and analysis could serve as a baseline and be updated over time to reflect changes in China's growth and science and technology development. Such data journey research questions could also then create a baseline set of data to feed into testing existing technology diffusion and adoption models noted earlier in the paper. In particular, such analyses could focus on data, data expertise and skills, and data infrastructures.

These assessments about China could be conducted by the CIA's Open Source Enterprise within its Directorate of Digital Innovation, as well as within the CIA's new China Mission Center.[90] The CIA has also launched a new CIA Technology Fellows program to bring promising experts to the agency for one to two years of public service. This would be a way for the CIA to collaborate with an interdisciplinary team of China experts, science and technology studies scholars, and bioinformat-

[87] Jeffrey Chang, "Core Services: Reward Bioinformaticians," *Nature,* no. 520 (April 2015): 151–52, https://www.nature.com/articles/520151a.

[88] Cong Cao et al., "Reforming China's S&T System."

[89] Xiaolan Fu, Wing Thye Woo, and Jun Hou, "Technological Innovation Policy in China: The Lessons, and the Necessary Changes Ahead," *Economic Change and Restructuring,* no. 49 (June 2016): 139–57, https://link.springer.com/article/10.1007/s10644-016-9186-x. For other studies that discuss the problems of Chinese indigenous innovation, see Huang, "Sino-U.S. Relations"; and Shobert, "Priming the Pump."

[90] "CIA Makes Changes to Adapt to Future Challenges," Central Intelligence Agency, Oct. 7, 2021, https://www.cia.gov/stories/story/cia-makes-changes-to-adapt-to-future-challenges/.

ics researchers to conduct studies on the above research questions. In addition, the State Department's intelligence arm, the Bureau of Intelligence and Research, could also bring in Jefferson Science Fellows to work on China-focused biomedical big data analyses during their year of public service.[91] The National Security Agency-funded Laboratory for Analytic Sciences could also be a site of year-long, focused, unclassified analyses of China and its big data capabilities that could bring together researchers from academia, industry, and the intelligence community.[92] In addition, the Intelligence Advanced Research Projects Agency could create new research calls to support larger interdisciplinary academia-industry research teams to study these problems over a multi-year period. Competing analyses could also be conducted by Five Eyes partners, such as the National Cybersecurity Centre in the United Kingdom in partnership with the Alan Turing Institute (the United Kingdom's leading data science research center), as well as Canadian and Australian scholars and intelligence practitioners. The U.S. intelligence community could also commission specific research studies on this topic from academic scholars through its National Intelligence Council Associates Program or its CIA Labs program. Although this paper has focused on China, the same assessments could be conducted of other state or nonstate actors who are attempting to acquire illicitly biomedical big data.

Conclusion

Sociologist of information technology Geoffrey Bowker warns about the dangers of being caught up in what he calls the "information mythology" — a contemporary understanding of information that assumes that "information is everything," and everything is rendered coherent through the more or less systematic communication of information.[93] The threat from China's acquisition of U.S. biomedical big data has been a key focus of U.S. intelligence and law enforcement. However, existing literature concerned with biomedical big data and bioinformatics shows that there are major challenges in trying to use biomedical big data for any kind of applied purpose. Therefore, it is very unclear what threats are actually posed by the mere acquisition of data alone. We need to do a better job of conducting more complex socio-technical assessments of how China might try to use biomedical big data, as well as studying China's bioinformatics personnel and technical infrastructure. Just focusing on the data and the people who steal them does not tell us enough about China's ability to use those data for economic or security purposes. Security concerns about biomedical big data (and China's role in biomedical data hacking) need to be further studied and scrutinized with more robust empirical evidence in order to better inform U.S. decision-makers about the true nature of China's economic and national security threats.

Kathleen M. Vogel *is interim director and professor at the School for the Future of Innovation in Society at Arizona State University. Her research examines knowledge production on biosecurity and big data issues.*

Sonia Ben Ouagrham-Gormley *is an associate professor at the George Mason University's Schar School of Policy and Government. She is affiliated with the Biodefense Program and her research deals with issues at the crossroad between science and technology and security.*

This research was funded under the Project on Advanced Systems and Concepts for Countering Weapons of Mass Destruction (PASCC) grant #FA7000-18-1-0011.

Image: Hagerstown Community College (CC BY-NC-ND 2.0)[94]

[91] "Jefferson Science Fellows Program," National Academies of Science, Engineering, Medicine, accessed April 22, 2022, https://sites.nationalacademies.org/PGA/Jefferson/index.htm.

[92] "About," NC State University, Laboratory for Analytic Sciences, accessed April 22, 2022, https://ncsu-las.org/about/.

[93] Geoffrey C. Bowker, "Information Mythology: The World of/as Information," in *Information Acumen: The Understanding and Use of Knowledge in Modern Business*, ed. Lisa Bud-Frierman (London: Routledge, 1994), 231–47. See also Geoffrey C. Bowker, *Memory Practices in the Sciences* (Cambridge, MA: MIT Press, 2005).

[94] For image attribution, see https://www.flickr.com/photos/hagerstowncc/. For the license, see https://creativecommons.org/licenses/by-nc-nd/2.0/.