Big Data and Biodefense: Prospects and Pitfalls

15

Kathleen M. Vogel

15.1 Introduction

Since the 1980s, with the rise and proliferation of genetic engineering techniques and technologies, national and international security concerns have grown regarding the misuse of biology by state and non-state actors [1–7]. Over this timeframe, and particularly within the past 20 years, the rise of new digital and mobile technologies in terms of sensors, geospatial technologies, advanced analytics, social media, and cell phones, have made the collection and analysis of a wide variety of information, to include different types of biological data, more readily accessible. These developments provide opportunities to think about biodefense preparedness in new and unique ways. Harnessing digital technologies more effectively to better identify and assess emerging biosecurity threats could lead to improved detection, response, and preparedness measures to eliminate or at least mitigate a bioweapons attack. This chapter will discuss the application of "big data" and "big data" analytics to biodefense problems in the specific domains of: (1) threat awareness; and (2) surveillance and detection. The chapter will begin by defining "big data" and biodefense, and then will examine the potential promises and pitfalls in biodefense applications and provide recommendations for the future of "big data" in addressing biodefense problems.

K. M. Vogel (✉)
School of Public Policy, University of Maryland at College Park, College Park, MD, USA
e-mail: kvogel12@umd.edu

S. K. Singh, J. H. Kuhn (eds.), *Defense Against Biological Attacks*,
https://doi.org/10.1007/978-3-030-03053-7_15

15.2 What Is "Big Data"?

Although conceptions of "big data" have been circulating since the 1990s, the term gained cache in technology and public circles in 2008 [8]. "Big data" is a term used to describe extremely large datasets[1] which can only be analyzed computationally, either individually or integrated with other datasets, to reveal previously unknown patterns, trends, and associations. What makes "big data" approaches unique is that they offer, "the capacity to collect and analyze data with an unprecedented breadth and depth and scale" ([10], p. 722). "Big data" can be collected through either active means (e.g., a user's posts on social media) or passive means (e.g., FitBit data).

George et al. [11] discuss that "big data" includes different categories of data: public data, private data, data exhaust, community data, and data generated through self-quantification. Public data are data typically held by governments, governmental organizations, and local communities and are publically released or available. Private data are data held by private firms, non-profit organizations, governments, and individuals that reflect private information that cannot be easily derived from public sources. Data exhaust refers to data that are passively collected for a different purpose, but can be recombined with other data sources to create new information (e.g., cell phone data or internet searching data). Community data is a distillation of unstructured data—especially text—that capture a variety of social trends, thoughts, and behaviors (e.g., consumer reviews on products). Self-quantification data are types of data that are revealed by the individual through quantifying personal actions and behaviors (e.g., FitBit data). As one can see there are a variety of "big data" types. "Big data" analytics involve the process of examining "big data" gathered over multiple time points (e.g., seconds, minutes, hours, days, months, years), using a variety of analytic techniques, developed by unique algorithms. Some examples include text analytics (mining data from different kinds of text based documents and materials), machine learning (the development of computer programs that can access data and use it to learn for themselves to generate new data), predictive analytics (the use of different statistical, modeling, and other quantitative approaches to analyze current data to make predictions about future), and natural language processing (using computer programs to understand human speech as it is written or spoken)—to name only a few.

[1] The MacKensey Global Institute expands this definition to note that: ""Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered Big Data—i.e., we don't define Big Data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as Big Data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, Big Data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes)." ([9], p. 1).

Typically there are four characteristics associated with "big data" often described as the four Vs: Volume, Variety, Velocity, and Veracity. ***Volume*** refers to working with increasingly larger volumes of data (e.g., terabytes, petabytes, or larger). ***Variety*** refers to the ability to make sense of different sources and types of data, including structured forms (e.g., numbers, dates, and groups of words and numbers) and unstructured forms (e.g., text, video, audio, tweets) that can come from a wide variety sources, to include public and privately held sources. This data can consist of information that is solely digital in origin, as well as data that has been subsequently digitized. ***Velocity*** refers to the increasing frequency and speed of incoming data that needs to be analyzed, and the speed at which it can be analyzed. ***Veracity*** signifies the trustworthiness of the data. It is important to note at the outset that "big data" is often incomplete, imperfect, and error-prone, and how it is collected and stored in databases and repositories is not necessarily standardized. These shortcomings can make data access and sharing issues difficult—therefore, the use of "big data" has analytic challenges that are analogous, if not necessarily identical, to other types of data.

The benefits of using "big data"/"big data" analytics include the ability to gather, analyze, and compare large datasets, consisting of disparate sources, to identify new patterns and generate new knowledge about a problem. There is potentially a wealth of new information and knowledge that could be learned to improve biodefense using big datasets and technologies. There are, however, also some challenges to consider. Questions are often raised about the veracity of the data (including ways to thwart the collection of accurate data) and problems with the algorithms used by the analytics programs to create and process the data—which are issues that can be difficult for analysts to understand and assess. There are also potential information gaps in existing "big data" that cannot be filled simply by analytics programs. The problems and prospects of "big data" will be discussed further in more detail in subsequent sections of this chapter.

15.3 What Is Biodefense?

For the purposes of this chapter, I will define biodefense as the defensive measures taken in the United States to protect against an attack using biological weapons or biological agents. The overarching strategy that has set U.S. biodefense policy and programs since 2004 is Homeland Security Presidential Directive 10 (HSPD 10), *Biodefense for the 21st Century* [12].[2] This Presidential Directive described the "essential pillars" of the U.S. national biodefense program as: Threat Awareness, Prevention and Protection, Surveillance and Detection, and Response and Recovery. Threat Awareness involves gathering and analyzing information about potential adversaries and emerging developments in the life sciences and biotechnology that

[2]There have been some follow-on White House biodefense policy documents: *The National Strategy for Countering Biological Threats* [13]; *The National Strategy for Biosurveillance* [14].

can pose threats. Prevention and Protection involves developing a range of political tools (e.g., arms control, export control, international assistance) to impede adversaries from developing a bioweapons capability and launching an attack, as well as developing technologies and response measures to protect critical infrastructure from a potential attack. Surveillance and Detection is focused on creating early warning and detection technologies and systems to identify and provide attribution of a bioweapons attack. Response and Recovery deals with the various technologies, medical and non-medical countermeasures, policies and plans to enable quick and effective response and recovery in the event of a bioweapons attack. Each of these pillars is designed to work individually and collectively to support different kinds of defensive activities such as identification of emerging state and non-state actor threats, controlling access to dangerous pathogens, identification of suspicious disease outbreaks, protection of critical infrastructure used in biodefense work, and development of medical countermeasures in the event of an attack (See Fig. 15.1).

15.4 Intersections of "Big Data" and Biodefense

There have not been many articles or reports published specifically on the intersection of "big data" and biodefense. One publication, however, that first raised this connection was a 2014 report, *National and Transnational Security Implications of Big Data in the Life Sciences*, co-authored by the American Association for the Advancement of Science, the Biological Countermeasures Unit of the WMD Directorate of the Federal Bureau of Investigation, and the United Nationals Interregional Crime and Justice Research Institute [16]. The report assessed a range of national security risks and benefits of using "big data" in the life sciences, and examined possible solutions to address the risks. The report focused largely on possible risks associated with the use of "big data" analytics, such as the vulnerabilities of life science datasets and cyber infrastructures/cloud-based systems to cyber intrusion/cyber hacking, and the potential design and development of pathogens, toxins, or biologically active molecules as biological weapons from the integration and analysis of "big data" in the life sciences. The report also tended to focus its discussion on a broad brush stroke of "big data" and security issues in the life sciences, rather than a discussion of specific cases of beneficial applications or harms in biodefense. Therefore, this chapter will fill a gap in knowledge and focus on how one might develop and use "big data"/"big data" analytics for two pillars of U.S. biodefense: (1) Threat Awareness; and (2) Surveillance and Detection.

15.4.1 Threat Awareness and "Big Data"

According to HSPD 10, Threat Awareness involves obtaining timely, accurate, and relevant biological warfare related intelligence, as well as conducting assessments of current and future trends and patterns in the evolving bioweapons threat. These activities involve collecting and analyzing a wide range of information on potential

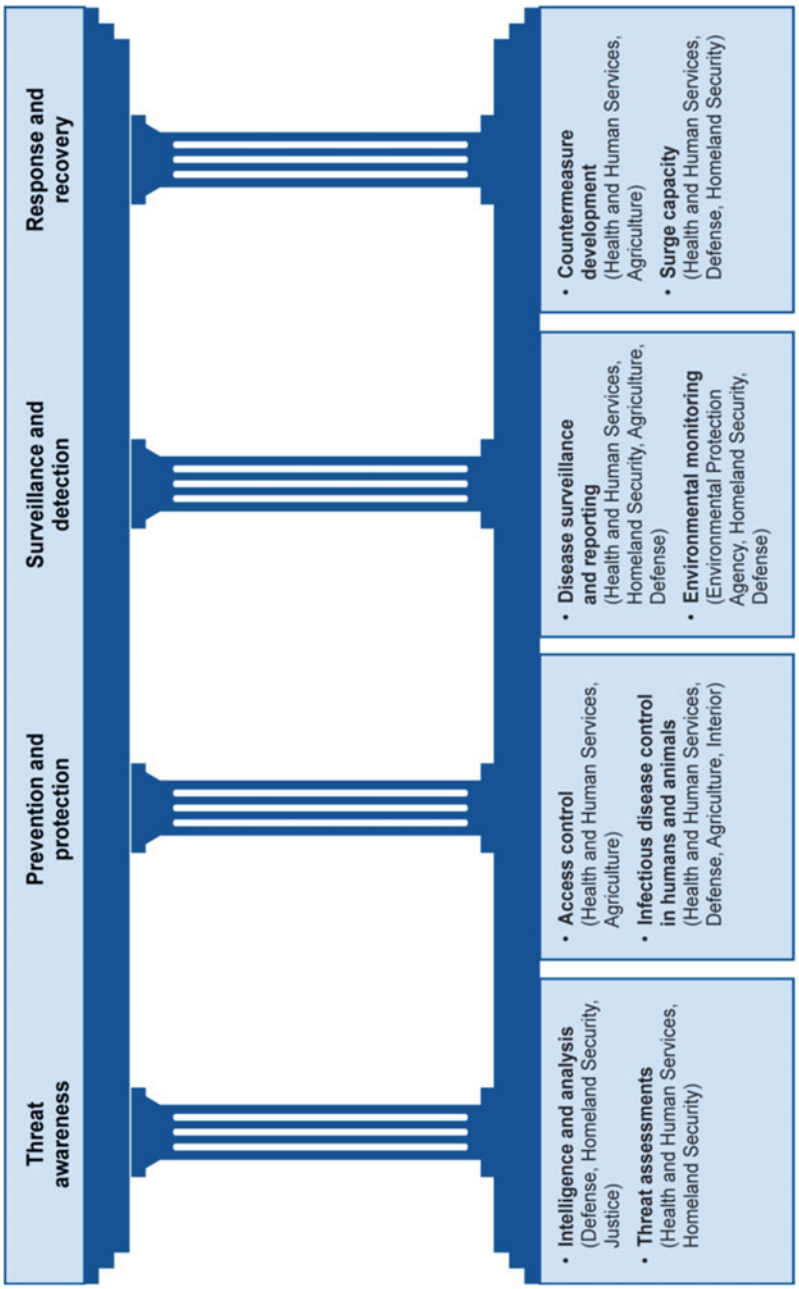

Fig. 15.1 Pillars of U.S. Biodefense ([15], p. 4)

adversaries and their potential interest in, and capability with, biology and biotechnologies, to "help position intelligence collectors ahead of the problem," as well as inform medical and non-medical countermeasure development [12]. "Big data" and "big data" analytics are well poised to assist with this effort. "Big data" approaches now offer even more ways to amass, count, organize, and store details relevant to biodefense, such as many types of biologically-relevant materials and related tangible resources used to develop biological weapons, as well as different sets of information (e.g., identities, composition, location, resources, capabilities) on potential adversaries. Different types of "big data" analytics technologies would be useful for sorting through a variety of classified and unclassified information on potential adversaries and emerging biotechnology threats.

For example, data fusion analytics, such as GIS technologies, can integrate heterogeneous datasets such as different types of sensor data, with video/image/text data. Laura Tateosian at North Carolina State University is developing a new GIS analytics prototype called GazeGIS that combines natural language processing, eye tracking, and geoparsing to provide users with additional contextual information as they read/process any text displayed on a computer screen [17]. When used in conjunction with eye-tracking hardware, the prototype displays real-time, gaze-contingent geographic data about particular city or country names. Alternatively, a user can hover over these words with a computer mouse to create the same effect if eye-tracking hardware is unavailable. The current prototype is able to provide overview maps, coordinates, and weather information, as well as information derived from Google search results; additional features are being developed. This system could help assist an intelligence analyst in processing a variety of intelligence reports on adversaries developing bioweapons capabilities—as the analyst reads the report, GazeGIS automatically identifies and creates a real-time map of the locations and geographical features identified. These mapped features can provide important contextual information for the text. In doing so, the GazeGIS prototype provides users the ability to have an enhanced real-time immersive experience about different geographical and contextual data while reading text and reports (Fig. 15.2).

There are other types of visual analytics platforms that can help identify and connect social relations between potential adversaries. One example is SAS Visual Analytics, which can provide a variety of data about and between social actors in a network based on social media data. The program can identify various features about the structure of a network, and who are key influencers or actors in the network to develop an understanding of micro and macro-level aspects of a social network (See Fig. 15.3). These types of analytics would be useful to intelligence analysts to better process intelligence information and understand the actors involved in a state or non-state level bioweapons program (e.g., who are critical players, who are involved in leadership/management roles, the broader community/network that supports the actors). A May 2017 NATO meeting discussed how various text and visual analytics programs could be used to identify threats stemming from weapons of mass destruction [18].

Nowviskie and Saathoff [19] showed how "big data" analytics can be used to aid in the investigation of insider threats involving bioterrorism. They examine the 2001

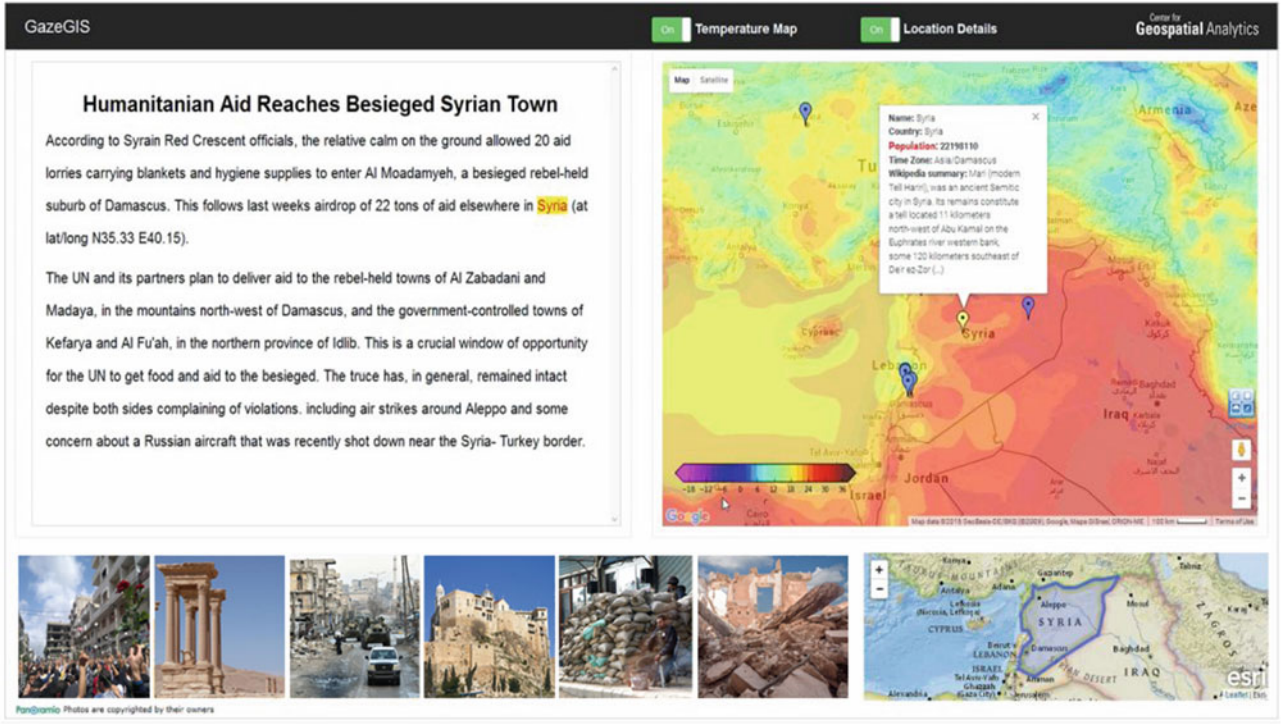

Fig. 15.2 GazeGIS Prototype: As analyst reads text on Syria, geographical and contextual information immediately pop up onto the computer screen. (Image and copyright approval obtained from Laura Tateosian.)

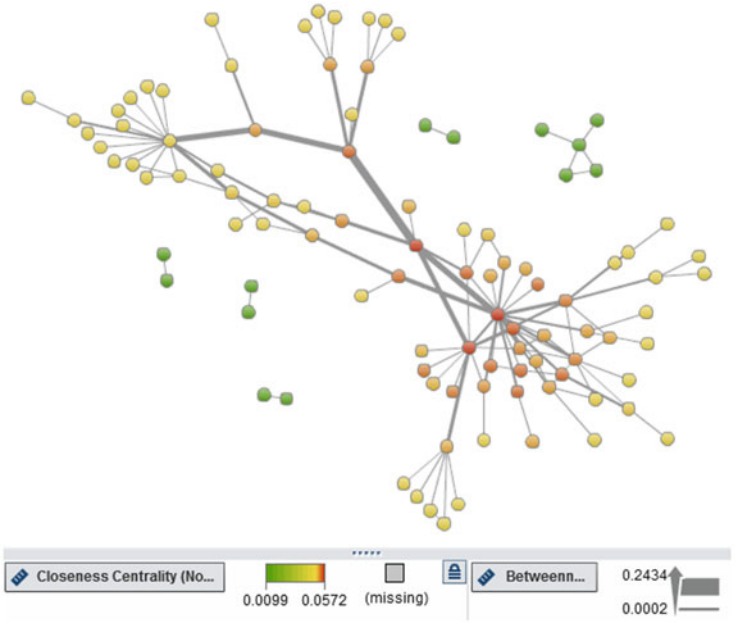

Fig. 15.3 Example of SAS Visual Analytics: Nodes and connectors show relationships between actors and the strength of the relationships in the social network(s). (Obtained from: http://blogs.sas.com/content/sascom/2014/02/19/exploring-social-networks-with-sas-visual-analytics/. Copyright approval obtained from Falko Schulz.)

Amerithrax case, in which letters filled with *Bacillus anthracis* spores were mailed in the U.S. postal system, ultimately causing 5 deaths, 17 additional infections, and risking health of 10,000 others due to possible exposure and prophylactic antibiotic therapy [20].[3] Nowviskie and Saathoff used this case to discuss how tools of computational text analysis and data mining could have been used to aid the criminal investigation on Dr. Bruce Ivins, a U.S. biodefense researcher who was implicated in the crime (and who ultimately committed suicide before being charged by the FBI). The FBI's investigation included large sets of heterogeneous structured and unstructured data including: 26,000 emails, 4 million megabytes of computer data from 10,000 witness interviews and 5750 grand jury subpoenas, 6000 items seized from Ivins' home and other locations, the collection of 5730 environmental samples from 60 site locations, medical and psychological reports about Ivins, and microbial forensic data generated from biological samples in Ivan's laboratory.[4] Combining and interpreting these heterogeneous datasets was challenging and took the U.S. government nearly 10 years to complete, and is considered to be one of the most complex investigations undertaken by the FBI.

[3] See: https://www.fbi.gov/history/famous-cases/amerithrax-or-anthrax-investigation

[4] Ibid.

Nowviskie and Saathoff argue that had "big data" analytics been available at that time, it could have streamlined the Amerithrax investigative process and made important connections and correlations between the disparate datasets in a more efficient manner. For example, Nowviskie and Saathoff describe how now there are suites of new "big data" analytic tools to assist the investigation that would involve sentiment analysis, authorship attribution, and pattern analysis at scale [21, 22]. Sentiment analysis is the study of opinions, sentiments, and emotions in text that can reveal affect or a psychological state. Although Ivins deleted many of his emails, it was later learned that copies of these messages were saved on his computer system. These messages could have been processed with new sentiment analysis tools to generate new understandings of Ivins mental state and behavior. New authorship attribution algorithms could have been applied to the *Bacillus anthracis* letter mailings and compared to Ivins' writing to see if stronger corroboration existed. Also, using data analytics to combine information on Ivins movements, financial transactions, and other data on his life could have revealed particular patterns in his life and work that went unnoticed by investigators at the time, or that only were connected much later in the investigation.

The applicability of these types of analytics to help intelligence assessments becomes clear when the flawed 2003 intelligence assessments on Iraq's bioweapons programs are examined. One could imagine a different outcome had "big data" analytic approaches been used to help with analysis of the Iraqi data. Although there are several reasons for the intelligence failure (see [23]) and collection on Iraqi data was difficult, there were critical datasets and data integration efforts that were not pursued by intelligence analysts. Richard Kerr, former Deputy Director of the CIA who was tasked to form a team to study the intelligence failures on Iraq's weapons of mass destruction (WMD) programs, found that there was little intelligence collection on open source social, cultural, political, and economic impacts on Iraq from the nearly 20 years of war and 10 years of sanctions, UN inspections, and international isolation, and no analyses of how the datasets from these impacts might affect the progress of a WMD program ([24], p. 48–49, 51). Kerr et al. note that although "softer" intelligence data on Iraqi societal issues, personalities, and elites was available in different databases, they were difficult for U.S. intelligence to access, process, and integrate with classified information. As a result, there was "no attempt to synthesize a broader understanding of Iraq out of the many detailed pieces. . ..the absence of such a contextual effort contributed to assessments that failed to recognize the significance of gaps in collection that may have been more evident when viewed from a larger perspective" ([24], p. 52). What resulted, then, was a narrow, and ultimately flawed, technical analysis of Iraq's WMD programs, based on thin classified data that was not corroborated and integrated with other information sources. At the time, if there might have been data mining and data fusion "big data" analytic tools available for intelligence, the results of that data processing might have produced analyses showing a more measured perspective on Iraq's WMD capabilities.

15.4.2 Cautions in Using "Big Data" for Threat Awareness

Despite the outlines potential benefits, one must also exert caution with relying exclusively on "big data" approaches. The social dimensions of science and technology that relate to measuring know-how/skills or technological diffusion are not readily captured by existing "big data" platforms and techniques. Acquisition of such data typically involves qualitative methodologies, such expert interviews, focus groups, laboratory observation (or other forms of ethnography). How "big data" approaches can be enabled to encourage these alternative, non-quantitative means of analysis for the life sciences and incorporate them with other pieces of information is an important point for discussion in considering the impact of "big data" on assessments. Lack of careful attention to both the tangible and intangible dimensions of the life sciences and biotechnology may result in erroneous intelligence and policy assessments.

To illustrate these challenges, I will discuss here a case study that I have investigated on advances in the life sciences: the 2002 artificial synthesis of a poliovirus by Professor Eckard Wimmer and his research group at the State University of New York [25]. This is a paradigmatic case to consider for security concerns about ongoing research and the availability of "big data" in the life sciences, particularly in the fields of synthetic genomics and synthetic biology. After the open scientific publication of this poliovirus experiment in the journal *Science*, the U.S. policy community raised concerns about whether terrorists might pick up the experimental methods used in this paper to re-create poliovirus, or other more deadly viruses. Using standard data gathering and analytic methods, it would appear that this synthesis experiment would be replicated by anyone with access to relevant materials and equipment, which are commercially available. However, closer ethnographic study of this case revealed important laboratory know-how that constituted the success of this synthesis experiment. Without this know-how, one would be unable to synthesize poliovirus using the experimental protocol regardless of how many pieces of DNA one might buy, what equipment one has, who one might buy these materials from, or the step-by-step procedure. What I have learned about this experiment is that in spite of a wealth of explicit information and readily available materials out in the public domain, critical portions of making this experiment work remain craft-like techniques that involve quite a bit of specialized, localized know-how that is difficult to transfer to other laboratories or contexts.

This is not to suggest that all synthetic genomics experiments will be this difficult to be used for harm. But, this analysis did reveal that there is a spectrum of skill difficulty involved in different kinds of biological work that needs to be considered in conducting threat assessments. The poliovirus case study illustrates the need to examine, at a detailed micro-level, the social context and relationships that comprise scientific and technological work and the complex practices and timescales that contextualize the replication and transfer of these techniques, even in cases when the science or technology in question seems at first approximation to be readily standardized, accessible, and transferable. As sociologists of science Nigel Gilbert and Michael Mulkay have long noted, the public and private accounts of science are very different; only the private discourse reflects the varied, contingent, and messy

character that constitutes the social character of scientific work [26]. As a result, scientific publications, presentations, and other formal means of codified scientific knowledge that are readily accessible through many "big data" approaches do not tell the whole story. Therefore, caution should be used in applying "big data" analytics to published scientific experiments as these analytics may not capture important details crucial to accurate threat assessment.

Existing "big data" approaches may be able to collect a lot of explicit information about the life sciences and biotechnology, but are they able to collect "Whole Data" [27]? For example, not only is there a need to consider individual tacit knowledge involved in technical work, but also knowledge networks, organizational/leadership components and dynamics, as well as the larger socio-political-economic contexts at the micro and macro levels that can also shape technical work. Importance aspects of this data is also not available solely through quantitative means. Thus, a reliance only on "big data" approaches may obscure or make invisible important information that is difficult to measure in tangible forms. In addition, Nowviskie and Saathoff [19, p. 56] caution that "big data" analysis still, "depends on the subjective standing points of those who assemble datasets, design the processes by which they are analyzed, and interpret the results." They argue that "big data" analytics do not remove the need for experts to work with and analyze the data, but rather serve as an aid to investigations to provide additional information, details, and patterns that can help narrow the focus and attention of the work. Therefore, understanding the private, micro-level practices and macro-level context that underpin scientific and technological work, as well as relying on expertise to interrogate "big data" results, is critical to evaluating the nontrivial social aspects that can shape biosecurity threats.

15.4.3 Surveillance and Detection

According to HSPD 10, Surveillance and Detection is focused on creating early warning and detection technologies and systems to aid in determining the timing, location, magnitude, and attribution of a bioweapons attack. The ability to predict and track when and where a disease outbreak may occur and how a pathogen transmits can significantly improve response strategies at the local, national, and international levels. Understanding of pathogen transmission has generally required epidemiologic data, involving clinical and laboratory reports and on-the-ground investigations, to generate accurate forecasts and an understanding of transmission patterns that take into account the various biological, environmental, behavioral, and socio-cultural issues that that can dynamically change disease patterns.

Traditional public health surveillance systems remain primarily based on manually collected and coded data, which is slow to gather and expensive, and difficult to disseminate for analysis; also, data is often aggregated at the national or regional level [28]. Van Panhuis et al. [29] note that the acquisition and sharing of detailed epidemiological data from traditional public health systems can be particularly difficult and scarce in crisis/emergency situations or resource-limited environments;

also, information is often not readily released due to trust concerns or other socio-political-economic factors. In this context, alternative data sources, such as from social media and the Internet, serving as nontraditional surveillance systems can be useful to help understand disease dynamics in an outbreak.

Bansal et al. [28] discuss how there are different types of electronic data streams for disease surveillance that can be gathered through "big data" approaches: medical encounter data, participatory syndromic data, and non-health digital data. Medical encounter data include electronic records from healthcare facilities, medical insurance claims, hospital discharge records, and death certificates. Such data streams provide information on the disease in patients (including the specific diagnosis) and can be gathered and monitored at the individual level or aggregated geographically. Participatory syndromic data are crowd-sourced data in which volunteers self-report their symptoms. This data does not provide confirmation of disease attributed to specific pathogens but does provide individual-level health data in near real time. Finally, non-health digital data are data gathered from the use of Internet search engines, social media, or mobile phones. They involve data on health-related behavior, including location and travel patterns, which can provide additional details about potential disease transmission. A 2016 National Academies of Science workshop noted that other kinds of structured and unstructured non-health electronic data sources, such as data from school attendance records, veterinary clinics, pharmaceutics sales, global transportation patterns, and climate could also be combined with other types of individual and community-level medical information to yield a more holistic picture of health in a specific area that would enable early warning or alert systems of a disease outbreak ideally weeks or months in advance ([30], p. 5).

Chowell et al. [31] have illustrated the potential of using "big data" from various Internet sources to understand the 2013–2016 Ebola virus disease epidemic in Western Africa and the 2015 Middle East respiratory syndrome (MERS) outbreak in South Korea. They used Internet news reports, health bulletins, and online reports from various health authorities and respected news sources, coupled to epidemiological information from traditional public health systems, to gather and analyze near real-time "big data" on disease clusters of patients to identify MERS coronavirus transmission chains. "Big data" analysis indicated that all 150 MERS cases in the outbreak involved a single cluster in which all cases were linked to exposure in the healthcare setting—either via occupational or social/family exposure. In case of the Ebola virus disease outbreak, epidemiological data was scarce; that which was initially available was largely aggregated at the country level, with more localized data available later. With these data shortages and the need to quickly understand Ebola virus transmission characteristics, the researchers collected information on Ebola virus disease case clusters from Internet news reports published during the outbreak. They gathered data on suspected, probable, and confirmed cases of Ebola virus disease in the three most affected countries (Guinea, Liberia, and Sierra Leone). They focused on reports and news items available from the World Health Organization web site, and other available Ebola virus disease situational reports, as well as online authoritative media outlets.

The data revealed a lot of variation in transmission patterns across Western Africa—revealing over 100 distinct patient clusters contributing to disease transmission. Later on, these clusters were confirmed via traditional surveillance systems. From using a variety of on-line information Chowell et al. [31] found that the main exposure to Ebola virus occurred via family contacts, as well as funeral and hospital settings. Once proper infection control measures were in place, the number of new cases in these settings substantially decreased. This study illustrated how publically available real-time online reports and news items can contain useful and actionable information regarding exposure and transmission patterns during disease outbreaks. This approach of "digital epidemiology" [32] could prove particularly useful for forecasting and response in resource limited areas, in which detailed transmissions studies are limited, and in crisis situations of major disease outbreaks when traditional surveillance reports are delayed.

15.4.4 Cautions in Using "Big Data" for Surveillance and Detection

Along with the beneficial use of "big data" in disease surveillance, one must also consider some of the problems that can emerge with the use of "big data" platforms. Google Flu Trends (GFT) was a web-based service developed in 2008 that monitored and mined millions of Google users' health-related on-line searching combined with computer modeling, and that coupled this data to laboratory-confirmed influenza cases and influenza tracking information from the Centers from Disease Control and Prevention (CDC) to analyze whether influenza A virus was present in a given population. GFT was similar to weather-forecasting methods that use real-time observational and environmental data to continually update and correct weather predictions. By aggregating Google search queries, GFT attempted to make accurate predictions about influenza activity in 29 countries to help localities and countries reduce the impact of seasonal and pandemic influenza (e.g., increasing vaccinations and supplies of antiviral drugs, issuing information about infection control measures, advocating school or other public closings) [33]. The original Google paper about GFT stated that its predictions were 97% accurate compared with CDC data, and that GFT was able to predict regional influenza outbreaks up to 10 days before they were reported by the CDC [34].

Subsequent independent reports, however, asserted that GFT predictions have sometimes been very inaccurate—e.g., in 2009 when GFT completely missed a non-seasonal influenza outbreak of H1N1 influenza, over the interval of 2011–2013 when GFT consistently overestimated influenza A virus prevalence, over one interval in the 2012–2013 influenza season when GFT predicted twice as many doctors' visits as the CDC recorded, and when GFT missed the peak and severity of the 2013 influenza season [33, 35–37]. Lazer et al. [36] investigated these failures and found there was a persistent pattern of GFT performing well for 2–3 years, and then failing significantly; overall, GFT appeared to perform worse over time. These failures were attributed to several reasons. First, there were inherent problems with Google's GFT algorithm, which was quite vulnerable to overfitting. For example, the algorithm was

fitting unrelated terms like "high school basketball" to influenza or to different health conditions. With millions of search terms being fit to the CDC's data, many of which were unlikely to be related to actual influenza cases, one can imagine how this led to problems in the accuracy of influenza predictions. Similar to the GFT errors, in one other well-known example of "big data" gone wrong, Leinweber [38] demonstrated that data mining techniques could show a strong but spurious correlation between the changes in the S&P 500 stock index and butter production in Bangladesh. Therefore, analysts who do not possess expert domain knowledge, may be misled by correlations from "big data" analysis.

Google also did not take into account changes in users' search behaviors over time [35]. Also, Lazer et al. [36] pointed out how the GFT algorithm was part of a larger business operation at Google, in which engineers made continual changes to Google's algorithm to improve its commercial service and advertising revenue and to increase its customers. As a result, the company was constantly testing and creating tweaks in the algorithm to improve searching. These modifications, however, could then skew the functioning of GFT, unintentionally making some search terms more prevalent and thereby hindering the overall accuracy of GFT. Such background adjustments are not a problem only for GFT but can also plague other social media datasets such as Twitter or Facebook, which are also constantly being changed and updated.

Lazer et al. [36] wrote that one of the key problems of GFT was related to "big data hubris," which is the, "implicit assumption that "big data" are a substitute for, rather than a supplement to, traditional data collection and analysis ([36], p. 1203). Allain-Jacques Valleron, an epidemiologist at the Pierre and Marie Curie University in Paris and founder of France's Sentinelles monitoring network, concurs with Lazer et al.: "It is hard to think today that one can provide disease surveillance without existing systems. . .The new systems depend too much on old existing ones to be able to live without them" [33]. This is not to say that the GFT algorithm could not be improved to rectify all or many of its original problems, but the GFT failure also raises cautions about relying too heavily on "big data" approaches alone. Scott Dowell, deputy director for surveillance and epidemiology at the Bill and Melinda Gates Foundation, advocates that "big data" approaches are, "most useful when combined with old-fashioned 'shoe leather' epidemiology—researchers walking door-to-door collecting primary data to understand the spread of an outbreak" ([30], p. 8).

15.4.5 Additional Concerns to Consider Regarding the Use of "Big Data"/"Big Data" Analytics for Biodefense

Beyond the Google Flu Trends mishap, there are additional challenges that researchers and policymakers need to understand regarding the use of "big data" and "big data" analytics for biodefense. Crawford and Finn [39] provide a useful overview of the limitations of what they call "crisis data"—the use of social and mobile data to understand disasters and other crisis events—of which a bioterrorism/

biological weapons attack would be relevant. Crawford and Finne discussed how social media data can provide important tactical information about a crisis and suggest potential response measures for that moment. They caution, however, that social media datasets represent a specific time point in an event, usually represented by a spike in Twitter messages or the use of particular hashtags. This focus can lead to a narrow conceptualization of the disaster, which leaves out an important understanding of the origin of the crisis or an understanding of the aftermath in which more of the impact and longer-term implications of the disaster is realized.

Also Crawford and Finne pointed out that "big data" researchers often work in a very different context and location than those involved in the response and recovery in the midst of the disaster; therefore, the researchers can only mine a small fraction of the entire experience and effect of the disaster, leading to a partial and incomplete record of the event. Also Crawford and Finne noted that social media data is not always representative of a local population. For example, Twitter tends to be used by younger, wealthy, and more urban adult populations, meaning that infants, older, rural, and more marginalized populations may not be represented by Twitter posts ([28], p. S377). Also, most social media data lacks demographic information, such as age and sex, which is usually important in epidemiological studies. All of these factors could be a significant shortcoming for response planning in a bioweapons attack given that infants/children, older, rural, and more marginalized/developing country populations might be excluded from social media datasets and those populations are typically more vulnerable and could be significantly harmed by a bioweapons attack.

Social media platforms can also be blind to important animal or zoonotic disease outbreaks ([28], p. S378). This blindness could lead to missing important patterns of disease outbreaks or recognizing patterns when none actually exist. Therefore, it is important to be aware of the properties and limitations of social media data or other kinds of "big data," what other kinds of datasets are needed, and taking all of these factors into account for designing appropriate response measures (see [40]).

Crawford and Finne also pointed out the problem of bots in Internet and social media platforms that can hamper accurate interpretation of events. Bots are software applications that perform an automated task; they are often used to perform repetitive tasks such as accessing websites to gather their content for search engine indexing (e.g., web crawlers), or other types of information gathering, organizing, and displaying. There are also malicious bots that come with viruses. Bots are designed to simulate human responses and activity and are a significant presence in many online spaces. Crawford and Finne demonstrated that a large number of tweets in datasets are derived from bots. De Micheli and Stroppa [41] conducted a research study of Twitter and found 20 million "fake" accounts—approximately 9% of Twitter's active users. As Crawford and Finne wrote, "Those bots are friending and retweeting other bots, producing a complex bot culture which is an emerging phenomenon unto itself." ([39], p. 496). Bots create significant problems for interpretation and filtering of social media data—is the tweet coming from a human or from a bot?

A similar problem is the emergence of "fake news," i.e. manipulated media posted by individual and group internet users to spread disinformation and/or propaganda to achieve a particular social, cultural, political, or economic goal

[42]. Also, users can also shape/disguise their online identities and create false information [16, 43, p. 40–41]. This misinformation could be used to hide covert bioweapons activities or redirect public health attention in other directions. One can imagine that in the event of a possible bioweapons attack, a key question will be whether social media information represents real data, manipulated data, or artificial bot-generated data. The problem of fake information is not a new development—there have long been efforts to create fake stories about bioweapons. In their book on the Soviet biological weapons program, Leitenberg et al. [44] reveal a disturbing US covert deception and disinformation campaign that operated from the late 1960s through the early 1970s. This program was directed at misleading the Soviet Union regarding the objectives of the US offensive bioweapons program and to suggest that a covert US bioweapons program existed even after the United States had unilaterally renounced its biological weapons program in 1969, leading the Soviets to invest time and money on a response. This deception program was led by the Federal Bureau of Investigation with assistance of the US Army, and used a double agent to convey false information to the Soviets. The authors found that the Central Intelligence Agency, years later, debated whether this campaign might have contributed to the Soviet decision to violate the Biological and Toxin Weapons Convention and move its weapons activities underground, or at least provided political cover to justify the decision to increase its biological weapons program. Such disinformation campaigns arguable are easier to conduct in the internet age. Therefore, it is important to remember that "big data" approaches do not remove the problem of disinformation, but, in many cases, may make it more problematic.

15.5 Conclusion

This chapter has illustrated how the use of "big data" can enhance biodefense in the areas of threat awareness and surveillance and detection. This chapter has also highlighted some of the pitfalls of relying on "big data" approaches without considering the limitations of such data, and how "big data" approaches can obscure or ignore other important information for analysis. Although there are many public pronouncements that "big data" analytics will soon replace human assessments, a more realistic and useful understanding would be to see "big data" approaches as one of several tools that be brought to bear on biodefense efforts. Rather than replace traditional efforts, Lazer et al. [36] have argued that we will likely see more "hybrid" systems that integrate "big data" with traditional data analysis. In doing so, the real analytic innovation will be using data from all sources, to include traditional public health data, coupled to social science, environmental, contextual, and ethnographic data, and otherwise, "small data", with "big data" [27, 36, 45–50]. In this way, the biases and short-comings of each method can be used to balance each other to arrive at a more accurate picture of a disease outbreak or bioweapons attack. Lazar et al. argued that the real data revolution, then, will be, "using data from all traditional and new sources, and providing a deeper, clearer understanding of our world" [47]. This kind of data analysis will also require new kinds of multi-disciplinary analytic teams,

who possess relevant subject matter expertise and understand long term trends and patterns [51], to be created and work together on these problems. To date, multi-method, multi-platform research studies are rare [52, 53], particularly in biodefense, but the payoff would be significant.

Uncertainties related to "big data," as with so many aspects of science, call for an approach of humility [54], in which we (e.g., scientists, academics intelligence practitioners, policymakers) recognize the limitations of science and when it is necessary to look beyond science for answers to complex, "wicked" problems. As Jasanoff noted, "[t]hese technologies compel us to reflect on the sources of ambiguity, indeterminacy and complexity" [54]. This does not mean that we need to be reject using "big data" or "big data" approaches to solve biodefense problems, but that we need to be very wary of over-relying on these techniques and technologies and losing sight of other complementary ways to generate knowledge.

References

1. Achenbach J, Sun LH. Scientists synthesize smallpox cousin in ominous breakthrough. The Washington Post. 2017. https://www.washingtonpost.com/news/speaking-of-science/wp/2017/07/07/scientists-synthesize-smallpox-cousin-in-ominous-breakthrough/?utm_term=.95d61eea13c8.
2. Central Intelligence Agency. The darker bioweapons future. 2003. https://www.hsdl.org/?abstract&did=442021.
3. Hilts PJ. Biological weapons reweighed. The Washington Post. 1986.
4. Makunda G, Oye KA, Mohr SC. What rough beast: synthetic biology and the future of biosecurity. Polit Life Sci. 2009;28(2):2–26.
5. National Research Council. Biotechnology research in an age of terrorism. Washington, DC: The National Academies Press; 2004. https://doi.org/10.17226/10827.
6. The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction. Report to the President. 2005. https://fas.org/irp/offdocs/wmdcomm.html.
7. U.S. Department of Defense. Advances in biotechnology and genetic engineering: implications for the development of new biological warfare agents. 1996. www.acq.osd.mil/cp/docs/reports/biotech96.pdf.
8. Lohr S. How big data became so big. The New York Times. 2012. http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html.
9. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A. Big data: the next frontier for innovation, competition, and productivity. The McKinsey Global Institute. 2011. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.
10. Lazer D, Pentland AS, Adamic L, Aral S, Laszlo Babasi A, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. Life in the network: the coming age of computational social science. Science. 2009;323(5915):721–3.
11. George G, Haas MR, Pentland A. Big data and management. Acad Manag J. 2014;57(2):321–6.
12. The White House. Homeland Security Presidential Directive 10 (HSPD 10): biodefense for the 21st century. 2004. https://fas.org/irp/offdocs/nspd/hspd-10.html.
13. The White House. The national strategy for countering biological threats. 2009. https://obamawhitehouse.archives.gov/the-press-office/president-obama-releases-national-strategy-countering-biological-threats.

14. The White House. The national strategy for biosurveillance. 2012. https://obamawhitehouse.archives.gov/the-press-office/2012/07/31/national-strategy-biosurveillance.
15. U.S. General Accountability Office. Biodefense: The Nation Faces Multiple Challenges in Building and Maintaining Biodefense and Biosurveillance, Statement of Chris Currie, Director, Homeland Security and Justice, Testimony Before the Committee on Homeland Security and Governmental Affairs U.S. Senate GAO-16-547T. 2016. http://www.gao.gov/assets/680/676548.pdf.
16. American Association for the Advancement of Science. Federal Bureau of Investigation, and the United Nations Interregional Crime and Justice Research Institute. National and transnational security: implications of big data in the life sciences. 2014. https://www.aaas.org/report/national-and-transnational-security-implications-big-data-life-sciences.
17. Tateosian L, Glatz M, Shukunobe M, Chopra P. GazeGIS: a gaze-based reading and dynamic geographic information system. In: Burch M, Chuang L, Fisher B, Schmidt A, Weiskopf D, editors. Eye tracking and visualization. ETVIS 2015. Mathematics and visualization. Berlin: Springer; 2017. p. 129–47.
18. NATO. Distributed data analytics for combating weapons of mass destruction. STO meeting proceedings, MP-IST-131. 2017.
19. Nowviskie B, Saathoff GB. Interpretation and insider threat: rereading the anthrax mailings of 2001 through a "big data" lens. In: Akhgar B, Saathoff GB, Arabnia H, Hill R, Saniforth A, Bayerl P, editors. Application of big data for national security. 1st ed. Amsterdam: Elsevier; 2015. p. 55–67.
20. Murch RS. Amerithrax: the investigation of bioterrorism using *Bacillus anthracis* spores in mailed letters. In: Katz R, Zilinskas RA, editors. Encyclopedia of bioterrorism. New York: Wiley; 2011. p. 25–30.
21. Moretti F. Graphs, maps, trees: abstract models for a literary history. London: Verso Books; 2005.
22. Moretti F. Distant reading. London: Verso Books; 2013.
23. Vogel KM. Phantom menace or looming danger: a new framework for assessing bioweapons threats. Baltimore, MD: Johns Hopkins University Press; 2013.
24. Kerr R, Wolfe T, Donegan R, Pappas A. Collection and analysis on Iraq: issues for the US intelligence community. Stud Intell. 2005;49(3):47–54.
25. Vogel KM. Framing biosecurity: an alternative to the biotech revolution model? Sci Public Policy. 2008;35(1):45–54.
26. Gilbert NG, Mulkay M. Opening Pandora's box: a sociological analysis of scientists' discourse. New York, NY: Cambridge University Press; 1984.
27. Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc. 2012;15(5):662–79.
28. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big data for infectious disease surveillance and modeling. J Infect Dis. 2016;214(Suppl 4):S375–9. https://doi.org/10.1093/infdis/jiw400.
29. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS. BMC Public Health. 2014;14:1144.
30. National Academies of Sciences, Engineering, and Medicine. Big data and analytics for infectious disease research, operations, and policy: proceedings of a workshop. Washington, DC: The National Academies Press; 2016. https://doi.org/10.17226/23654.
31. Chowell G, Cleaton JM, Viboud C. Elucidating transmission patterns from Internet reports: Ebola and Middle East respiratory syndrome as case studies. J Infect Dis. 2016;214(Suppl 4):S421–6. https://doi.org/10.1093/infdis/jiw356.
32. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol. 2012;8(7):e1002616. https://doi.org/10.1371/journal.pcbi.1002616.
33. Butler D. When Google got flu wrong. Nature. 2013;494(7436):155–6. http://www.nature.com/news/when-google-got-flu-wrong-1.12413

34. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4. https://doi.org/10.1038/nature07634.
35. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends performance in the United States during the 2009 Influenza Virus A (H1N1) pandemic. PLoS One. 2011;6:e23610.
36. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176):1203–5.
37. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013;9(10):e1003256. https://doi.org/10.1371/journal.pcbi.1003256.
38. Leinweber D. Stupid data miner tricks: overfitting the S&P 500. J Invest. 2007;16(1):15–22. https://doi.org/10.3905/joi.2007.681820.
39. Crawford K, Finn M. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. GeoJournal. 2015;80:491–502. https://doi.org/10.1007/s10708-014-9597-z.
40. Dixon D. Analysis tool or research methodology? Is there an epistemology for patterns? In: Berry D, editor. Understanding digital humanities. London: Palgrave Macmillan; 2012.
41. De Micheli C, Stroppa A. Twitter and the underground market. 11th Nexa lunch seminar, Turin, Italy. 2013. http://nexa.polito.it/nexacenterfiles/lunch-11-de_michelistroppa.pdf.
42. Marwic A, Lewis, R. Media manipulation and disinformation online. 2017. https://datasociety.net/output/media-manipulation-and-disinfo-online/k.
43. Mazur E. Collecting data from social networking web sites and blogs. In: Gosling SD, Johnson JA, editors. Advanced methods for conducting online behavioral research. Washington, DC: American Psychological Association; 2010. p. 77–90.
44. Leitenberg M, Zilinskas RA, Kuhn JH. The Soviet biological weapons program: a history. Cambridge, MA: Harvard University Press; 2012.
45. Belk RW. Qualitative research in advertising. J Advert. 2017;46(1):36–47.
46. Manovich L. Trending: the promises and the challenges of big social data. In: Gold MK, editor. Debates in the digital humanities. Minneapolis, MN: University of Minnesota Press; 2012. p. 460–75.
47. Murthy D. Digital ethnography. An examination of the use of new technologies for social research. Sociology. 2008;42(5):837–55. https://doi.org/10.1177/0038038508094565.
48. Orgad S. How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In: Markham AN, Baym NK, editors. Internet inquiry. Conversations about method. Los Angeles, CA: Sage; 2009. p. 33–53.
49. Snijders C, Matzati U, Reips U-D. "Big data": big gaps of knowledge in the field of Internet science. Int J Internet Sci. 2012;7(1):1–5.
50. Tufekci Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: ICWSM '14: Proceedings of the International AAAI Conference on Weblogs and Social Media; 2014.
51. Arbesman S. Stop hyping big data and start paying attention to long data. Wired. 2013. https://www.wired.com/2013/01/forget-big-data-think-long-data/.
52. Adar E, Weld DS, Bershad BN, Gribble SS. Why we search: Visualizing and predicting user behavior. In; Proceedings of the 16th International Conference on World Wide Web, 161–70. WWW '07. New York, NY, USA: ACM; 2007.
53. Kairam SR, Morris MR, Teevan J, Liebling D, Dumais S. Towards Supporting Search over Trending Events with Social Media. In: Seventh International AAAI Conference on Weblogs and Social Media; 2013.
54. Jasanoff S. Technologies of humility. Nature. 2007;450:33.