The Journal of Infectious Diseases
SUPPLEMENT ARTICLE

Big Data for Infectious Disease Surveillance and Modeling

Shweta Bansal,[1,2] Gerardo Chowell,[1,3] Lone Simonsen,[1,4] Alessandro Vespignani,[5] and Cécile Viboud[1]

[1]Fogarty International Center, National Institutes of Health, Bethesda, Maryland; [2]Department of Biology, Georgetown University, Washington D.C.; [3]School of Public Health, Georgia State University, Atlanta; [4]Department of Public Health, University of Copenhagen, Denmark; and [5]Network Science Institute, Northeastern University, Boston, Massachussets

We devote a special issue of the *Journal of Infectious Diseases* to review the recent advances of big data in strengthening disease surveillance, monitoring medical adverse events, informing transmission models, and tracking patient sentiments and mobility. We consider a broad definition of *big data* for public health, one encompassing patient information gathered from high-volume electronic health records and participatory surveillance systems, as well as mining of digital traces such as social media, Internet searches, and cell-phone logs. We introduce nine independent contributions to this special issue and highlight several cross-cutting areas that require further research, including representativeness, biases, volatility, and validation, and the need for robust statistical and hypotheses-driven analyses. Overall, we are optimistic that the big-data revolution will vastly improve the granularity and timeliness of available epidemiological information, with hybrid systems augmenting rather than supplanting traditional surveillance systems, and better prospects for accurate infectious diseases models and forecasts.

Keywords. big data; infectious diseases; surveillance; disease models; transmission; social media; Internet search queries; electronic health records; mobility; adverse events; outbreaks.

The last 15 years have seen the rapid emergence of big data and data science research, which lies at the intersection of computer science, statistics and data visualization, and builds on the growing wealth of digital footprints [1]. The increasing availability of electronic records and passive data generated by the use of Internet, mobile phones, satellites, and radio-frequency sensors can be mined to uncover new patterns and associations. One can also take advantage of the interactive digital infrastructure to design participatory platforms and citizen science experiments. The field of infectious diseases research is not immune to the big data revolution, as attested by a near-exponential increase in the number of publications at the nexus of big data, digital epidemiology, and infectious diseases since approximately 2001 (Figure 1).

DEFINITION OF BIG DATA

Like most fashionable and recently coined terms, the meaning of *big data* remains elusive, and even the simple question "how big is big data?" remains poorly answered. Although the term is often reserved for data sets so large or complex that traditional analytical approaches fail, big data can be used more broadly to refer to advanced analytical methods, no matter the size, type, or form [1]. Indeed, one important feature of big data resides in "inflation," that is, the relative size increase of data over what is typical. This is a useful concept because the absolute size of big data will certainly differ between scientific fields; whereas the few gigabytes of data providing the mobile phone traces of a few millions users can be considered small compared with the 15 petabytes of data produced by the Large Hadron Collider annually, they represent a revolution in the area of social sciences and public health. Three "V" terms, volume, velocity, and variety, are frequently associated with big data, in reference to the quantities of data, the increasing speed of collection and use, and the many differing types and forms they arrive in [2, 3]. In addition, qualifiers such as veracity, validity, volatility, and value have been put forward to address the need for accuracy, staying power, and utility of these data.

THE DAWNING BIG-DATA ERA IN INFECTIOUS DISEASES

The pillar of infectious disease control has always been surveillance systems tracking diseases, pathogens, and clinical outcomes [4]. Traditional surveillance systems, however, are notorious for severe time lags and lack of spatial resolution; systems that are robust, local, and timely are thus critically needed. Monitoring and forecasting of emerging and reemerging infections are of particular interest [5], including pandemic influenza, Middle East respiratory syndrome, severe acute respiratory syndrome, Ebola, Zika, and drug-resistant pathogens.

Sectors such as marketing or meteorology have perfected the art of real-time acquisition and analysis of highly resolved digital data, providing a detailed lens on human social behavior and our physical environment. In public health, however, critical surveillance systems remain primarily based on manually collected and coded data, slow to amass and expensive, and difficult to disseminate for analysis. Furthermore, reporting from these systems tends to be national or regional with little in the way of information about diseases at the local level. A new era

Correspondence: C. Viboud, 16 Center Dr, Fogarty International Center, NIH, Bethesda, MD 20892 (viboudc@mail.nih.gov).

The Journal of Infectious Diseases® 2016;214(S4):S375–9

DOI: 10.1093/infdis/jiw400

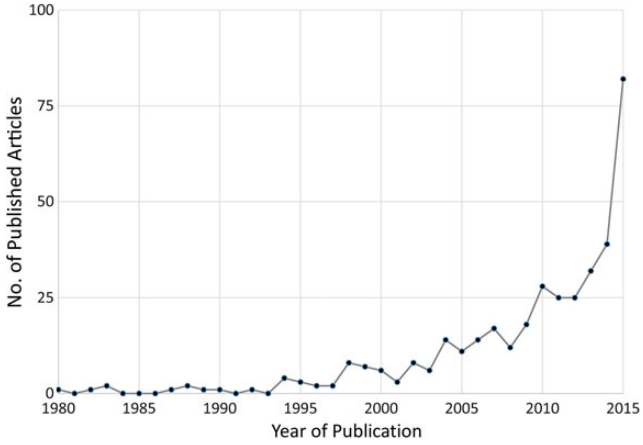

Figure 1. Exponential increase since the early 2000s in publications at the intersection of big data and infectious diseases. Annual trends in the number of publications were identified through a Scopus search for articles published between 1980 and 2015, using the following keywords: (big data AND infectious diseases) OR (big data AND epidemics) OR (digital epidemiology AND infectious diseases).

fortunately seems to be dawning in which surveillance systems are strengthened by big-data streams, including electronic health (e-health) patient records, and non-traditional digital data sources, such as social media, Internet, mobile phones, and remote sensing. For this special issue, we invited a group of multidisciplinary experts in epidemiology, computer science, and modeling to reflect on the recent achievements of big data for infectious disease surveillance and modeling, identify important challenges and opportunities, and share their vision on where the field is headed.

SCOPE OF THIS ISSUE

This special issue includes nine articles that cover a variety of infectious diseases and methods to illustrate use of big data, relying on three types of electronic data streams: medical encounter data, participatory syndromic data, and nonhealth 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 disease status (here *disease* is defined broadly as a collection of symptoms, confirmed infection, or a drug or vaccine-related adverse event) and can be monitored at the individual level or aggregated geographically before reporting.

Participatory syndromic data are crowd-sourced data in which volunteers self-report an array of symptoms. These data streams do not provide confirmed infection status for specific pathogens but provide individual-level health data in near real time.

In contrast, nonhealth digital data do not depend on a particular patient or medical encounter but instead derive from use of Internet search engines, social media, or mobile phones. In addition to self-reporting of health outcomes, these data streams also provide information on health-related behavior, including contact and travel patterns, vaccine status and sentiments, which are key ingredients for understanding and modeling disease transmission.

The nine articles in this issue propose ways to advance monitoring, inference, and forecasting of disease outbreaks and transmission patterns. Below, we organize the contributions in two broad themes, disease surveillance and modeling, and introduce each article briefly.

Big Data in Support of Infectious Disease Surveillance

Infectious disease surveillance is one of the most exciting opportunities created by big data, because these novel data streams can improve timeliness, and spatial and temporal resolution, and provide access to "hidden" populations. These streams can also go beyond disease surveillance and provide information on behaviors and outcomes related to vaccine or drug use. However, the promise of these big-data streams must be balanced by caution.

In the first contribution, Simonsen et al [6] give a brief overview of the history of disease surveillance, highlight the gaps in current systems, and make the case for big data to strengthen and deepen syndromic surveillance. Using influenza as a case study, they demonstrate how the use of high-volume *International Classification of Diseases*–coded e-health medical claims data collected by large private-sector data warehouses sheds light on the spread of pandemics with unprecedented spatial detail. They highlight the gross underutilization of these data streams, due in part to privacy concerns and barriers in access to e-health data in academia and government. They also demonstrate how novel digital surveillance tools, such as Google Flu Trends [7], can go wrong because of overfitting and quickly go defunct. This issue is particularly salient after deep perturbations to disease dynamics, as is the case with the emergence of a new pandemic virus. Continued validation against traditional surveillance systems provides an hedge against these issues. In light of the rise and fall of systems based purely on digital search engine data, however, the way forward most likely involves "hybrid" systems that integrate digital big data with traditional laboratory-based surveillance and e-health data to improve the timeliness, accuracy, and depth of existing surveillance indicators.

In the next article, Guerrisi et al [8] discuss participatory surveillance, using the European influenza reporting system Influenzanet as an illustration. This surveillance system relies on volunteers signing up online to report their health on a weekly basis, in an effort to strengthen already well-established Sentinel physician-based systems in Europe. The authors highlight how such a system has the flexibility to be used for instant tracking of any emergent health problems. A key advantage of participatory systems is to report from populations who would not otherwise seek medical care for illnesses they perceive as minor, particularly adults with influenza-like illnesses. Furthermore, these systems allow for analysis of potential biases in the

population sampled and the ad-hoc definition of cohorts of users specific to a particular health issue.

Next, MacFadden et al [9] address the seemingly unstoppable rise of antimicrobial resistance and the lack of a global monitoring of this issue in a timely manner. They introduce ResistanceOpen, a new online platform for monitoring of bacterial drug resistance, based on timely scanning, aggregation, analysis, and dissemination of local and regional online resistance index reports (http://www.healthmap.org/resistanceopen/). This approach is a direct extension of prior efforts to track infectious disease outbreaks globally by curating and analyzing a variety of online data sources (HealthMap [10]). Here, on-line resistance data compare favorably with traditional reporting systems in the United States and Canada so that ResistanceOpen offers a successful platform for sharing antibiotic resistance prevalence for multiple pathogens on local scales.

In the next paper, Salathe [11] sees big data as a creative solution for improving detection of drug adverse events in the future, relying on patients researching information online or reporting their postexposure symptoms in health forums, on Twitter or Facebook. Statistical mining of unstructured texts derived from these data streams could uncover associations between adverse events and specific drugs, and greatly improve the timeliness of surveillance in this area. A similar argument could be made for tracking of vaccine-related adverse events, which currently rely on passive reporting by physicians. Analysis of internet and social media data streams could complement the Food and Drug Administration's Sentinel initiative, a recent effort to identify drug-related adverse events in large e-patient individual databases. Importantly, public-generated digital data can also be mined for information on behavior and sentiments, so as to monitor vaccine hesitancy and drug uptake [12, 13]. An important challenge here is to find balance between the timeliness of social media information and the cost of unfounded alerts. Indeed, claims of adverse events can quickly and irreparably taint lifesaving drugs or vaccines in the public opinion. As with disease surveillance, building hybrid systems that integrate big-data streams with passive physician reports of adverse events will help safeguard the accuracy and specificity of the alerts.

Beyond Surveillance: Big Data for Modeling of Disease Transmission Dynamics and Control

The wealth of information promised by big data, combined with the development of new analytical and modeling tools, will help shed light on intricate details of the transmission dynamics of infectious diseases that have so far remained obscured by lack of granular data. Four articles in this issue tackle this topic.

In the first article touching on modeling, Moran et al [14] reflect on the state of the art of epidemiological modeling and judge it to be on par with that of particle physics in the 1970s. They argue that epidemic modeling of nonhealth data, such as Internet search queries, is an example of rising to the challenge of catching up this delayed development. They draw an interesting parallel between disease forecasting, which remains in its infancy, and weather forecasting, which is deemed the reference standard for real-time integration and modeling of large data streams, followed by immediate release of personalized outputs—as attested by the omnipresence of forecasts in the media, the Internet, and smartphones. Improvement in disease forecasts will require a large increase in the volume of epidemiological information available for modeling, which big data may deliver, but will also require better standardization of disease reports and case definitions across time and space. The authors also note that communication of forecast uncertainty is a challenge and typically better accomplished in meteorology than disease forecasts - everyone understands a 20% chance of rain but not a 20% chance of outbreak. Finally, although weather forecasts are ultimately driven by the inalterable laws of physics, changes in human behavior (eg, due to risk perception and media attention) can affect the dynamics of an outbreak and skew its associated digital footprints, making disease forecasts potentially more complex.

Next, Lee et al [15] review the technical, practical, and ethical challenges of using big data to understand the spatial distribution and transmission of infectious diseases. The heterogeneity of data sets and data types particular to this field makes integration of different data streams obtained at different spatial scales technically challenging; greater use of multilevel Bayesian statistical approaches would help alleviate this issue. The authors also stress a conceptual chasm between classic epidemiology and the world of big data, related to the notion of sampling. Indeed, there is a lack of effort to conduct proper sampling and address issues of representativeness and coverage in big data, in contrast to the long history of developing sampling theories centered on formulation of well-specified hypotheses in classic epidemiology. Finally there are privacy issues associated with analysis of ever more detailed spatial data. Creative solutions include simulation of synthetic data sets reproducing the key epidemiological features identified in big data, while ensuring confidentiality of individual cases.

In the next article, Wesolowski et al [16] see great potential in global positioning system information gleaned from cell-phone data and satellite imagery to inform population movements and network interactions, which drive the dissemination of epidemic diseases. They note that cell-phone usage remains heterogeneous in developing countries but is rapidly increasing. Notable biases in cell-phone data include, call frequency, tower density, ownership, and coverage, which is typically affected by socioeconomic status, age, and urban/rural residence. Some level of data aggregation can generally solve these biases, although identifying the right scale appropriate for disease transmission remains an issue. Research directly connecting cell-phone movements and disease data is in its early phase, but 2 successful examples are worth noting, both set in Kenya. These include

studies of how population movements can help dissect the source-sink spatial dynamics of malaria and how seasonal fluctuations in travel drive rubella epidemics. Looking to the future, cell-phone data will probably be more useful for modeling invasion waves of new pathogens rather than well-established pathogens, and further research should focus on careful cross-pathogen comparisons.

Next, Chowell et al [17] propose using Internet news reports and health bulletins to compile information on clusters of patients, and infer transmission trees and reproduction numbers, using lessons drawn from work on Middle East respiratory syndrome and Ebola virus outbreaks. Sourcing news media data can yield rapid and accurate assessment of transmission chains, which is crucial in the absence of detailed surveillance data, as was the case for much of the 2014 Ebola epidemic. Although Chowell et al used a manual approach to search, identify, extract, and model relevant information, they point out that this approach could be expanded with automatic scans of all Internet news and development of language processing tools to identify sequential transmission events. This approach, based on methods developed by HealthMap [10] and other Internet-based surveillance resources, could provide a particularly productive avenue for surveillance in low- and middle-income countries, where detailed transmission studies are scarce, and in infectious disease crises settings when time is of the essence.

Finally, Liu et al [18] introduce a novel system for data management and epidemic simulation, analysis and visualization, EpiDMS. This tool aims to fill an important gap in decision making during healthcare emergencies by generating near-real-time projections and simulations of an emerging pandemic trajectory. Arguably, models can be sensitive to parameters, and many rounds of simulations should be performed to address uncertainty and explore different epidemiological scenarios. EpiDMS provides an interface to facilitate interpretation and visualization of large numbers of multivariate simulations. For instance, this tool would allow public health experts to identify a set of conditions (eg, thresholds of vaccine coverage and effectiveness) that are consistent with a given epidemic trajectory (eg, a decline of 50% in transmission). Such simulation tools can provide crucial actionable data for public health experts and modelers.

KEY ISSUES AND WAY FORWARD

We anticipate that this special issue will generate cautious hope for digital technologies to deliver useful information for infectious disease surveillance and modeling. Across all contributions, we have identified a few cross-cutting issues that deserve more research focus; they are outlined below.

Validation and Integration With Existing Systems

A key requirement of any new system, whether it involves big data or not, is a careful and continued validation against established systems. As we have seen with Google Flu Trends, an initial honeymoon period when the tool appeared accurate was overshadowed by failures to detect unusual patterns, such as the emergence of a pandemic. Similarly, the sensitivity of big-data surveillance for adverse events should be tested against known associations, such as the increase in infant intussusception linked with the Rotashield rotavirus vaccine, Guillain-Barré syndrome and pandemic influenza vaccines, or anemia with certain cancer drugs. In the absence of a reference standard for surveillance, as may be the case in developing countries, one could validate two or more big-data indicators against each other. Further, specificity has to remain high, so as not to overload the public health infrastructure with useless outbreak alerts, and in the case of adverse events, to prevent public mistrust and waste of perfectly safe drugs or vaccines. Hybrid systems can help overcome this issue in part, because they are validated in an ongoing fashion against traditional data. Intelligent integration of disparate sources of data in useful hybrid systems will also require further methodological developments, in particular around "data synthesis" approaches [19, 20].

Representativeness and Biases

A few shortcomings specific to digital data are worth keeping in mind. Many of these data streams lack demographic information, such as age and sex, which is an important component of almost any epidemiological study. Furthermore, they represent an ever-increasing but still limited segment of the population, with lack of coverage among infants, and fewer elderly than younger individuals represented. There is geographic heterogeneity in coverage, with underrepresentation in developing countries, though these biases tend to disappear and are arguably less pronounced than those found in traditional surveillance systems on a global scale. Furthermore, there is spatial and temporal uncertainty in the information retrieved; for instance, a young man in New York may be researching his grandmother's illness that occurred in Florida 3 weeks earlier. In addition, tools of big data are blind to important animal epidemics and zoonotic diseases at the human-animal interface. Here, "hybrid" systems could be created to combine digital data with veterinary or agricultural reports on recent animal disease outbreaks.

Data Volatility

As often with new technologies and innovation, digital data streams are not always continuous but rather subject to user interest, popularity, and financial concerns. Whereas laboratory-based surveillance for influenza has existed at the Centers for Disease Control and Prevention for >40 years in a continuous form, Google Flu Trends lasted less than a decade. This issue is compounded by the fact that digital data streams are not specifically created for surveillance or research purposes, and there is no incentive to maintain these tools once they go out of fashion.

Use of Data: Managing Signal Versus Noise and Experimental Approaches

Perhaps the most intriguing question is how the public health community will use this anticipated overload of data. Big data are generally unstructured, biased, and noisy; replicability and transparency must be at the center of data science. Users of big data must stay away from the pitfalls of "big data hubris" [21] and complement novel approaches with rigorous statistical and hypothesis-driven analyses. Conversely, a substantial advantage of social media over observational surveillance is the possibility to conduct experiments, for instance, to study contagion phenomenons (see the "viral" diffusion of online petitions [22]), which is akin to experimental epidemiology.

Data Volume and Forecasting Horizon

Although computational approaches are increasingly used for public health planning, contingency plans preparations, and near-real-time forecasts at time of infectious disease crisis, the field is certainly far from the consolidated state of the art we find in other scientific disciplines such as weather forecasting. The ongoing big data revolution should remedy the sparseness of existing epidemiological observations, so that accurate forecasts of disease trajectories several weeks ahead will become feasible on a local scale [20]. The last 10 years have seen dramatic advances in data collection and digitalization, finally allowing the construction of a portfolio of data-driven epidemiological models. The increasing number of available models and the lack of best practices in integration and data sharing are however major roadblocks in the development of the field. Meanwhile, predicting the emergence of new infectious diseases, is a much more complex goal and will require further research on disease dynamics in zoonotic reservoirs, human-animal contacts, and host species barriers, which go well beyond big data [23].

In conclusion, this supplement offers a cautiously optimistic view of the progress of big data for infectious diseases surveillance and control during the last decade. We hope this collection of articles will spark dialogue and collaborations between the public health community, epidemiologists, big data specialists, physicists, and disease modelers alike. Multi-disciplinary initiatives such as the Big Data To Knowledge (BD2K) program led by the National Institutes of Health will be instrumental to strengthen funding and use of big data in biomedical research [24]. Aided by these large programs, further research will need to resolve some of the issues and gaps identified here and realize the full potential of big data for infectious disease control. Those in the field of disease surveillance and modeling can learn a lot from other data-and modeling-hungry fields, such as meteorology and marketing, and can ultimately provide useful tools to improve situational awareness and outbreak response for a variety of old and new infections.

Notes

Financial support. This work was supported by the RAPIDD (Research And Policy for Infectious Diseases Dynamics) Program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health (L. S. and S. B.); the European Commission (Marie Curie Senior fellowship to L. S.); the Lundbeck Foundation (visiting professorship grant to L. S.); the in-house research program of Fogarty International Center (G. C. and C. V.); the Defense Threat Reduction Agency (grant 1-0910039 to A. V.); and the National Institutes of Health (grant MIDAS-U54-GM111274 to A. V.).

Potential conflicts of interest. All authors: No potential conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Wikipedia. Big data. **2015**. https://en.wikipedia.org/wiki/Big_data. Accessed 30 July 2016.
2. Laney D. Deja VVVu: others claiming Gartner's construct for big data, **2015**. http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/. Accessed 30 July 2016.
3. Marr B. Why only one of the 5 Vs of big data really matters, **2015**. http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters. Accessed 30 July 2016.
4. Thacker SB, Stroup DF. Future directions for comprehensive public health surveillance and health information systems in the United States. Am J Epidemiol **1994**; 140:383–97.
5. Woolhouse ME, Rambaut A, Kellam P. Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. Sci Translat Med **2015**; 7:307rv5.
6. Simonsen L, Gog J, Olson D, Viboud C. Infectious disease surveillance in the "big data" era: towards faster, locally-relevant and more accurate systems. J Infect Dis **2016**; 214(suppl 4):S380–5.
7. 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.
8. Guerrisi C, Turbelin C, Blanchon T, et al. Participatory syndromic surveillance of influenza in Europe. J Infect Dis **2016**; 214(suppl 4):S386–92.
9. MacFadden D, Fisman D, Andre J. A platform for monitoring regional antimicrobial resistance using online data sources: ResistanceOpen. J Infect Dis **2016**; 214(suppl 4):S393–8.
10. Harvard School of Public Health. HealthMap. **2016**, http://www.healthmap.org/site/about. Accessed 30 July 2016.
11. Salathe M. Digital pharmacovigilance and disease surveillance: combining traditional and big data systems for better public health. J Infect Dis **2016**; 214(suppl 4): S399–403.
12. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. Plos Comput Biol **2011**; 7:e1002199.
13. Harvard School of Public Health. HealthMap. Vaccine Sentimenter: global monitoring of vaccine conversations, **2016**. http://www.healthmap.org/viss/. Accessed 30 July 2016.
14. Moran KR, Fairchild G, Generous N, et al. Epidemic forecasting is messier than weather forecast: the role of human behavior and Internet data streams in epidemic forecasting. J Infect Dis **2016**; 214(suppl 4):S404–8.
15. Lee EC, Asher J, Goldlust S, et al. Mind the scales: harnessing spatial big data for infectious disease surveillance and inference. J Infect Dis **2016**; 214(suppl 4):S409–13.
16. Wesolowski A, Buckee C, Engo-Monsen K, Metcalf C. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. J Infect Dis **2016**; 214(suppl 4):S414–20.
17. Chowell G, Cleaton JM, Viboud C. Elucidating transmission patterns from internet reports: Ebola and MERS as case studies. J Infect Dis **2016**; 214(suppl 4):S421–6.
18. Liu SH, Poccia S, Candan KC, Chowell G, Sapino ML. EpiDMS: data management and analytics for decision making for epidemic spread simulations. J Infect Dis **2016**; 214(suppl 4):S427–32.
19. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. Plos Comput Biol **2015**; 11:e1004513.
20. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nat Commun **2013**; 4:2837.
21. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science **2014**; 343:1203–5.
22. Brockmann D. Contagion dynamics in online petitions, **2016**. http://rocs.hu-berlin.de/projects/page41/index.html. Accessed 30 July 2016.
23. Lloyd-Smith JO, George D, Pepin KM, et al. Epidemic dynamics at the human-animal interface. Science **2009**; 326 :1362–7.
24. National Institutes of Health. Big Data to Knowledge (BD2K). https://datascience.nih.gov/bd2k/about. Accessed 28 September 2016.