https://doi.org/10.14745/ccdr.v41i09a03

Commentary

Big Data is changing the battle against infectious diseases

Links MG[1,2*]

[1]Saskatoon Research Centre, Agriculture and Agri-Food Canada, Saskatoon, SK
[2]Department of Computer Science, University of Saskatchewan, Saskatoon, SK

*Correspondence: Matthew.Links@usask.ca

Abstract

Big Data has traditionally been associated with computer geeks and commercial enterprises, but it has become entrenched in many scientific disciplines including the prevention and control of infectious diseases. The use of Big Data has allowed disease trends to be identified and outbreak origins to be tracked and even predicted. Big Data is not getting smaller. The challenges we face are to hone our analytical capacity to address the huge "signal-to-noise" ratio with adequate computing power and multidisciplinary teams that can handle ever-increasing amounts of data. Big Data will also create the opportunity for future applications of *bespoke* (or personalized) treatment.

Introduction

Big Data seems like a recent development that for many is tied to phrases like "cloud computing". The term originated before the turn of the millennium when in the 1990's when John Mashey was Chief Scientist of Silicon Graphics (SGI). At the time, SGI was at the forefront of computer graphics and was struggling to deal with significantly expanding computational needs that outpaced available hardware. Mashey developed a presentation in the late '90s that laid out the looming collision between Big Data and computational performance (1).

Commonly The term "Big Data" is now used to describe situations where data volumes are characterized by properties including, but not limited to: size, rate of change over time, and the heterogeneous nature of the data itself (2). Big Data generally refers to large volumes of data that can be structured (e.g. relational databases) or not (e.g. Twitter feeds) and are mined for information. While historically Big Data originated within the fields of Computer Science, Statistics and Economics(3), it has been increasingly adopted across all scientific disciplines.

A problem is typically said to involve Big Data when the volume is so large that it hinders the ability to convert data to knowledge. In the case of infectious disease research, Big Data is having a huge impact. The ability to perform real-time disease tracking and outbreak prediction has utilized unstructured data to change how infectious diseases are managed. For example, through the use of diverse news sources GPHIN has been used for early signal of novel infections (such as SARS and MERS-CoV) that informed public health response to the outbreaks that followed (4).

Structured data is particularly useful when collating information from multiple sources based on a predictable structure of the data. In the case of public health surveillance it is now possible to look for structured data that could serve as a surrogate source of information to laboratory confirmation or physician authored case reporting. Muchaal et al. demonstrated that pharmaceutical usage is one possible source of early surrogate information (5).

How big is Big Data?

It is hard to fathom how large Big Data actually is. In a recent paper Stephens and colleagues put Big Data from a couple of disciplines into a relative context (6). While the current champions of unwieldy data size are

astronomical studies, it was suggested by Stephens et al. that genomics would be on par with astronomical data sizes by the year 2025. The scale of genomics data in 2025 may be equivalent to 8 billion of the largest iPhones available today (128GB of space in 2015). Or an iPhone's worth of data for every person on earth, each year.

Big Data and the origins of an outbreak

The application of genomics to infectious diseases can help with identifying where infectious disease outbreaks actually came from. A good example of this was the detective work undertaken to respond to a measles outbreak during the 2010 Olympics (7). Using whole genome sequencing Gardy and colleagues were able to exactly identify many of the reported cases (30 of 82). One important finding was that there was more than one type of measles virus involved in the outbreak. While traditional genotyping for the measles virus has focused on the sequences of phosphoprotein and hemmagglutinin, two specific genes used to distinguish isolates, Gardy et al. showed that there were additional variations in other measles genes that could be used for a more precise definition of the viral lineages.

What's next?

One of the newer applications of Big Data is in what Jennifer Gardy termed *bespoke* (or personalized) treatment (8). For example, whole genome approaches across thousands of isolates can identify genomic variation linked with antimicrobial phenotypes of *Mycobacterium tuberculosis* (9). In the future, this may be applied generally to identify the best treatments for bacteria with antimicrobial resistance.

How are we going to interpret it all?

Despite all the potential for advances, there are some key challenges that Big Data faces in all disciplines. As data gets bigger and bigger it becomes harder to interpret: either the integration of data is so complex as to be hard to follow without serious computing resources or the sheer scale is beyond comprehension.

With the use of unstructured text from news feeds mined for disease surveillance knowledge, Big Data is being used to find meaning in a deluge of noise. The scale of text mining scientific manuscripts published in journals, news reports describing emergent issues and 140 character tweets truly is daunting when one considers that Twitter collects ½ billion tweets per day (6). What makes matters even more challenging is that disease surveillance doesn't simply mean aggregating news feeds. Rather, the key for a meaningful Big Data strategy to disease surveillance is the identification of the potential risk. Approaches such as GPHIN (4) are thus crucial for national and international preparedness for disease outbreaks. As Big Data continues to grow at exponential rates, trying to advance our capacity to analyze it becomes an ever-changing holy grail.

All too often a Big Data-set is acquired as part of a multi-disciplinary study and handed-off to a single individual (e.g. graduate student , post-doc or fellow) in the hope that they can, alone, come to an understanding of what it all means. Having a single person responsible for increasingly complex relationships arising from overwhelming volumes of data is just not a feasible strategy. Thus the trend is to develop multidisciplinary approaches to interpretation of Big Data (4).

Conclusion

Unless we envision a computational Dark Ages it is hard to believe that Big Data will shrink. Therefore scientific disciplines have been developing the capacity to exploit ever-increasing volumes of data. Care needs to be taken not to be overwhelmed with Big Data. In fact it is the shift to multi-disciplinary analysis of Big Data that is enabling teams to track disease trends and even predict outbreaks before they occur. Big Data is positioned to move increasingly from public health into the clinical setting; *bespoke* (or personalized) treatment of infectious diseases may soon be on our doorstep.

Funding

Dr. Links' research program has, or is currently, funded by Agriculture and Agri-Food Canada (AAFC), the Government of Canada's Genomics Research Development Initiative (GRDI), the Canadian Institutes of Health Research (CIHR), the National Research Council of Canada (NRC), Saskatchewan's Agriculture Development Fund (ADF), the Government of Canada's Canadian Safety and Security Program (CSSP, formerly CRTI - Chemical, Biological, Radiological-Nuclear Research and Technology Initiative).

Conflict of interest

None.

References

(1) Mashey J, ed. Big Data and the next wave of infrastress. USENIX Annual Technical Conference; 1998; Monterey, California, USA: Usenix.
(2) Asokan GV, Asokan V. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. J Epidemiol Glob Health. 2015 Mar 5.
(3) Diebold FX. A personal perspective on the origin(s) and development of 'Big Data': The phenomenon, the term, and and the discipline. Department of Economics, University of Pennsylvania. PIER Working Paper No. 13-003. November 2012.
(4) Dion M, AbdelMalik P, Mawudeku A. Big Data and the Global Public Health Intelligence Network (GPHIN). Can Commun Dis Rep. 2015;41:211-216.
(5) Muchaal PK, Parker S, Meganath K, Landry L, Aramini J**.** Evaluation of a national pharmacy-based syndromic surveillance system Can Commun Dis Rep. 2015;41:204-211..
(6) Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or genomical? PLoS Biol. 2015 Jul;13(7):e1002195.
(7) Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. J Infect Dis. 2015 Jul 6.
(8) Gardy JL. Towards genomic prediction of drug resistance in tuberculosis. Lancet Infect Dis. 2015 Jun 23.
(9) Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. Lancet Infect Dis. 2015 Jun 23.