



Seeking data and the implicit order that grounds a chaotic world

Henry Milgrom

'...you better start swimmin' or you'll sink like a stone for the times they are a-changin' [1].

Bob Dylan 1964.

Big data is a powerful tool offering the potential of new insights into areas as diverse as social ills, meteorology, genomics, complex physics simulations, and biological and environmental research. Contrarily, big data may be seen also as a troublesome manifestation of the power of state and corporations. The rise of big data has made the study of systems biology possible. Life scientists are beginning to collect immense datasets. A result of access to the newfound data is a new science that examines not just the characteristics of isolated parts of cells and organisms but properties that may not even show up at the parts level. Systems problems are emerging in all areas of biology and medicine. Living systems are dynamic and complex. At the molecular level, even the simplest bacterium is astoundingly complex. At all points, there are multiple interactions. Nearly all processes are non-linear. One of the most remarkable qualities of life is the ability to create and maintain order. The big questions before us are what enables biological systems to be linked together and to preserve order. Answers will lead to the development of predictive and actionable models for the study of biology and disease.

We are all practitioners of medicine, a discipline to which Lewis Thomas, the great 20th century physician, scientist and observer gave the name – the youngest science [2]. From the 1920s when he accompanied his general practitioner father on daily rounds through the years in medical school and later in practice, teaching and research, Thomas watched medicine evolve from an art into a sophisticated life science. Observation was the fundamental approach for both learning and teaching medicine. Today, the products of that observation might be called 'data' or 'information' – terms that are similar but not interchangeable. 'Data' is the Latin plural of 'datum', defined in English as facts and statistics collected for reference or analysis. In modern expression, 'data' is often used as a mass

noun, similar to 'information', and both typically take a singular verb [3]. Data consists of unrelated facts. When a significant relation exists among data points, then data becomes information [4]. Information is always about something, such as the size of a measurement, the occurrence of an event, or a blueprint. It does not need to be true or accurate. A signal that creates misunderstanding is still information. As a rule, as the amount of information increases, the message becomes more accurate. Knowledge arises from information. There is no single widely accepted definition of knowledge, and, there are many competing propositions. Plato (428 BC to 348 BC) [5] helped to lay the foundations of western philosophy and science. Philosophical debates about the nature of knowledge generally start with Plato's formulation as 'justified true belief'. Sources of knowledge include research findings, but also experience, authority, trial and error, and theoretical principles. Knowledge transformation is necessary for research results to become useable for clinical decisions [6,7]. Professional culture in the field of medicine identifies knowledge with particular persons, and the authority of individual physicians derives in large part from the general agreement that we have privileged access to medical knowledge. When Sir Francis Bacon, two millennia after Plato, wrote that knowledge of the world should be grounded in carefully verified facts, he was not just offering us a new method to gather old-fashioned knowledge. He was redefining the validation of knowledge [8]. The current age brings a redefinition equal in scope. Today data qualifies as the transformative new currency of science, medicine, engineering, education, commerce, and government [9], while knowledge is taking on the properties of the network

National Jewish Health, University of Colorado School of Medicine, Denver, Colorado, USA

Correspondence to Henry Milgrom, MD, Professor of Pediatrics, National Jewish Health, 1400 Jackson Street, J319, Denver, CO 80206, USA. Tel: +1 303 398 1999; e-mail: milgromh@njhealth.org

Curr Opin Allergy Clin Immunol 2014, 14:177–180

DOI:10.1097/ACI.0000000000000065

in which it resides. It is huge, nonhierarchical, public, unfiltered, open to differences and hyperlinked [8]. The power of knowledge is limited only by the imagination. One imaginative scientist, J. Craig Venter [10], best known for sequencing the human genome and the creation of life in the form of the first synthetic cell, is hard at work on a 'digitized-life sending unit'. Digital computers designed by DNA machines (Dr Venter and coworkers) are used to read coded instructions in DNA to analyze them and write them in a way that leads to the creation of new kinds of DNA (synthetic life) [9].

Although we still tend to think of medicine as a life science, with the growth of data generated not only by research but also in the care of patients, it is rapidly becoming an information science [7,11]. The quantity and complexity of information in medicine are bringing about a fundamental paradigm shift from the 'power of the individual brain' to the 'collective power of systems of brains' [12]. Big data emerged in the 21st century and has grown with the internet, analytics and cloud computing [13,14]. From the dawn of the human race, 100 000–300 000 years ago, until AD 2003 humankind had accumulated 1 exabyte (EB) of data. Humanity now produces as much data in 2 days [15]. At present, we are generating multiple zettabytes (ZB) of data each year (1 ZB = 1000 EB = 1000,000,000,000,000,000 bytes = 1000^7 bytes = 10^{21} bytes) [16]. Big data is no more than a collection of datasets too large and too complex to process using ordinary database management tools. It refers to the torrent of information discharged from such diverse sources as the web, biological, industrial and medical sensors, video, e-mail, and social networks. Analysis of big data reveals trends beyond the grasp of unaided human mind [17]. It makes it possible to ask new questions, test new hypotheses, and make data-driven decisions. It holds out a very large cost reduction and considerable speeding up of computing tasks [13]. It promises to accelerate the pace of discovery and to transform teaching and learning [18]. Inescapably, big data analysis is agnostic with regard to the data scrutinized. It is the process of examining large amounts and variety of data to uncover a needle in a haystack, hidden patterns, and hitherto unknown correlations [19]. The same methodology detects insurance fraud and terrorism, predicts outbreaks of influenza, finds genes implicated in cancer, and identifies women who are pregnant [17,20–23]. Graduate students use Rap Genius, a popular website for annotating lyrics from rappers like Jay-Z and Eminem, to add notes to Homer and Virgil. In a literary lab project on 18th-century novels, students of English literature pore over a database of nearly 2000 early books to

clarify when 'romances', 'tales' and 'histories' first emerged as novels. In introduction to critical text mining, English, history and computer majors analyze novels and Supreme Court rulings [24]. A new form of enterprise has come into view that collects, sells, and trades data on people's habits, activities, and interests, most often without their knowledge. Such disclosures can be abused to threaten civil liberties [25]. A new technology now being deployed can track for hours every vehicle and person across an area the size of a small city [26]. As surveillance technologies decline in cost and grow in sophistication, tracking of many aspects of our daily activities, even the seemingly mundane, has become a common rather than exceptional occurrence. Our every interaction may be recorded [27]. It is now possible to monitor the entire world and, in the meanwhile, to lose control of the personnel managing the data [28]. The potential dark side of big data posits the need for a code of conduct and ethical principles to guide big data analysts in their professional work [29]. One challenge is maintaining a vast amount of health data, while meeting high standards for security and privacy with implications not only for the outcome but also for the process [13]. Strict ethical research regulations should be imposed on reality mining research for public health applications, just like the rules that apply to human subjects research [30].

Life scientists are joining the big data club, collecting immense datasets comprising quantities of information that were once the realm of astronomers and physicists [31]. The European Bioinformatics Institute (EBI) in Hinxton, UK, currently stores 20 petabytes (1 PB = 1000^5 bytes) of data. Genomics at present accounts for 2 PB, a number that doubles every year. Big research data related to health come from three sources: a small number of groups that produce very large amounts of data, usually as part of projects funded to generate resources for use by the entire research community; individual investigators who produce large datasets through the use of new technologies; and an even greater number of sources that produce small datasets (e.g. research data or clinical data in electronic health records) that are integrated with other data. It is now possible to monitor continuously and record every heartbeat, all blood pressure changes, the rate and depth of respiration, body temperature, arterial oxygen saturation, blood glucose, brain wave activity, and other functions [16]. We can make images with three-dimensional reconstruction of any organ or body part [16]. Sensational work by Rogers Research Group at the University of Illinois is creating wireless medical patches that may be inserted into all organs to collect unlimited

real-time electrophysiologic data. The research incorporates fundamental studies with forward-looking engineering. It focuses on soft materials for conformal electronics, nanophotonic structures, microfluidic devices, and microelectromechanical systems. These efforts combine expertise from nearly every traditional field of technical study [32,33]. The emerging opportunity arises from combining these diverse data sources with improving computing tools.

A controversy has emerged over how much emphasis and funding should be devoted to traditional, 'hypothesis-driven' research, versus 'unbiased', big data projects. Hypothesis-driven research answers definitive questions about cause and effect. To big data scientists, hypothesis-driven science testing cause and effect relationships one at a time is too limiting. They hold the belief that truth lies in data [23]. They favor the measurement of as many variables as possible without being influenced by preconceived ideas. In health research, they organize large joint programs to collect and process data. They build computer models of disease to predict which factors will be important. With enough data, these models predict what is likely to happen without explaining why [17].

Open science is gathering force with the ability to assemble and analyze vast amounts of data [34]. Some journals now require that authors agree to share the data obtained in the studies they publish [35,36]. The world's oldest scientific publisher, The Royal Society of London, announced on 18 February 2014 the introduction of a new open access peer-reviewed journal that will publish research from across the entire spectrum of science and mathematics (<http://www.theguardian.com/science/open-access-scientific-publishing>). The journal, set to launch in autumn 2014, will be known as *Royal Society Open Science*. The journal will ensure that all data appearing in published papers will be freely and publicly available on the Internet. The trend toward greater transparency of clinical trial data is of profound importance. The question is becoming not whether, but how, this data should be shared [36]. The possibilities for discovery are limited only by the imagination [37]. In a noteworthy examination of vast existing datasets, researchers at the University of Pittsburgh tracked public reports of 56 diseases going back to the 19th century and determined that childhood vaccines have prevented more than 100 million cases of serious contagious disease in the United States since 1924 [38].

Asthma is a paradigmatic chronic disease. It affects inherently susceptible individuals. There are some predictable triggering agents and others that are still unrecognized and there are stochastic

events. Two challenging aspects of asthma care that offer great room for improvement are the ability to anticipate exacerbation and to identify those patients who are not adhering to their medication regimen. By chance, these categories overlap to a significant extent. We already know that severe exacerbations are preceded by a progressive deterioration over days or weeks [39]. Further, Frey *et al.* [40,41] pointed to the effectiveness of big data analysis by pinpointing the risk for future asthma episodes based on fluctuation analysis of long time series of lung function measurements, and there is great likelihood that big data analysis of medical and pharmacy records should help us address the problem of nonadherence [42,43].

Our challenge is to lay the foundation for the best interaction between human creativity and computation. As physicians, we solve medical problems. But we also want to understand the machinery of life and the causes of disease. We want to learn things of value to everyone and to participate in the shaping of the future of medicine. The future of health and technology is interconnected. The technological revolution is both inspiring and transformational. It serves our passion for understanding and is certain to result in better patient outcomes. Yet the past few weeks have brought us a new frame of reference. As we think about how an entire jetliner had gone missing with passengers and crew on board, we are harshly reminded of how much we cannot – and may never – know. As the first president of the National Geographic Society said in 1888, when his magazine set out to chart the known universe, 'The more we know, the greater we find is our ignorance' [15].

Acknowledgements

None.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Dylan B. The times they are a-changin'. <http://www.bobdylan.com/us/songs/times-they-are-changin#ixzz2kZNAXE9h>.
2. Thomas L. The youngest science: notes of a medicine-watcher. New York: Penguin Books; 1983.
3. Oxford dictionaries. http://www.oxforddictionaries.com/definition/english/data#m_en_gb0205520. [Accessed April 10, 2014]
4. Milgrom H, Tran ZV. Sublingual immunotherapy, meta-analysis, and knowledge in the age of information. *J Allergy Clin Immunol* 2009; 124: 162–163.
5. Plato. The dialogues of Plato. New York: Bantam Books; 1986.
6. Stewart T. Wealth of knowledge. New York: Doubleday; 2001.
7. Milgrom H, Tran ZV. The rise of health information technology. *Curr Opin Allergy Clin Immunol* 2010; 10:178–180.
8. Weinberg D. Too big to know. New York: Basic Books; 2011.
9. Lohr S. New U. S. research will aim at flood of digital data. *New York: Times*; 2012; http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html?_r=0. [Accessed April 10, 2014]
10. Venter J. Life at the speed of light: from the double helix to the dawn of digital life. New York: Viking; 2013.

11. Hersh W. Is medicine an information science? Perspective from Physician Time Studies 2012. <http://informaticsprofessor.blogspot.com/2012/02/is-medicine-information-science.html>. [Accessed April 10, 2014]
12. Stead WW, Searle JR, Fessler HE, *et al*. Biomedical informatics: changing what physicians need to know and how they learn. *Acad Med* 2011; 86:429–434.
13. Davenport T, Dyché J. Big data in big companies. SAS Institute, 2013, <http://www.sas.com/reg/gen/corp/2266746>. [Accessed April 10, 2014]
14. Mervis J. U.S. Science policy. Agencies rally to tackle big data. *Science* 2012; 336:22.
15. Iyer P. The folly of thinking we know: the painful hunt for Malaysia Airlines Flight 370. *The New York Times*. 20 March 2014. <http://www.nytimes.com/2014/03/21/opinion/the-painful-hunt-for-malaysian-airlines-370.html?ref=international>. [Accessed April 10, 2014]
16. Topol E. The creative destruction of medicine. New York: Basic Books; 2012
17. White M. How big data is changing science (and society). *Pacific Standard* 2013. <http://www.psmag.com/science-environment/big-data-changing-science-society-69650>. [Accessed April 10, 2014]
18. Kalil T. Big data is a big deal. White House Office of Science and Technology, 2012. <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>.
19. Kraska T. Finding the needle in the big data systems haystack. *Internet Comput* 2013; 17:84–86.
20. Explore flu trends around the world. <http://www.google.org/flutrends/>. [Accessed April 10, 2014]
21. Bonneville R, Jin VX. A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor alpha target genes. *Bioinformatics* 2013; 29:22–28.
22. Hill K. How target figured out a teen girl was pregnant before her father did. *Forbes*; 2012. <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>. [Accessed April 11, 2014]
23. Discovery Health finds the needles in its big data haystack [Press Release]. Johannesburg: Red Ribbon Communications; 2013. Cited 26 March 2014. <https://www.google.com/search?q=discovery+health+finds+the+needles&btnG=Search&oe=utf-8>. [Accessed April 11, 2014]
24. Lewin T. As interest fades in the humanities, colleges worry. http://www.nytimes.com/2013/10/31/education/as-interest-fades-in-the-humanities-colleges-worry.html?pagewanted=all&src=SMR_AP_LO_MST_FB. [Accessed April 11, 2014]
25. Knappenberger B. Why Care About the N.S.A.? *New York Times* 2013. <http://www.nytimes.com/2013/11/26/opinion/why-care-about-the-nsa.html>. [Accessed April 11, 2014]
26. Timberg C. New surveillance technology can track everyone in an area for several hours at a time. *The Washington Post* 2014. http://www.washingtonpost.com/business/technology/new-surveillance-technology-can-track-everyone-in-an-area-for-several-hours-at-a-time/2014/02/05/82f1556e-876f-11e3-a5bd-844629433ba3_print.html. [Accessed April 11, 2014]
27. Sankar S. The rise of human computer cooperation; 2012. <https://www.google.com/search?q=Sankar%20S.%20The%20rise%20of%20human%20computer%20cooperation%3B%202012%20&ie=utf-8&oe=utf-8>. [Accessed April 11, 2014]
28. Nocera J. Will digital networks ruin us? *New York Times* 2014. <http://www.nytimes.com/2014/01/07/opinion/nocera-will-digital-networks-ruin-us.html>. [Accessed April 11, 2014]
29. Rayport J. What big data needs: a code of ethical practices. *MIT Technol Rev*. 2011. <http://www.technologyreview.com/news/424104/what-big-data-needs-a-code-of-ethical-practices/page/2/>. [Accessed April 11, 2014]
30. The Belmont Report, US Department of Health and Human Services. <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>. [Accessed April 11, 2014]
31. Mattmann CA. Computing: a vision for data science. *Nature* 2013; 493:473–475.
32. Rogers Research Group, 2014. <http://rogers.matse.illinois.edu/>. [Accessed April 11, 2014]
33. Tingley K. The body electric. *The New Yorker* 2013. http://www.newyorker.com/reporting/2013/11/25/131125fa_fact_tingley. [Accessed April 11, 2014]
34. Marx V. Biology: the big challenges of big data. *Nature* 2013; 498:255–260.
35. Godlee F, Groves T. The new BMJ policy on sharing data from drug and device trials. *BMJ* 2012; 345:e7888; www.bmj.com/content/345/bmj.e7888. [Accessed April 11, 2014]
36. Mello MM, Francer JK, Wilenzick M, *et al*. Preparing for responsible sharing of clinical trial data. *N Engl J Med* 2013; 369:1651–1658.
37. Pinker S. Science is not your enemy: an impassioned plea to neglected novelists, embattled professors, and tenure-less historians. *The New Republic* 2013; <http://www.newrepublic.com/article/114127/science-not-enemy-humanities>. [Accessed April 11, 2014]
38. van Panhuis WG, Grefenstette J, Jung SY, *et al*. Contagious diseases in the United States from 1888 to the present. *N Engl J Med* 2013; 369:2152–2158.
39. Rodrigo GJ, Rodrigo C, Hall JB. Acute asthma in adults: a review. *Chest* 2004; 125:1081–1102.
40. Frey U. Predicting asthma control and exacerbations: chronic asthma as a complex dynamic model. *Curr Opin Allergy Clin Immunol* 2007; 7:223–230.
41. Frey U, Brodbeck T, Majumdar A, *et al*. Risk of severe asthma episodes predicted from fluctuation analysis of airway function. *Nature* 2005; 438: 667–670.
42. Milgrom H, Wamboldt F, Bender B. Monitoring adherence to the therapy of asthma. *Curr Opin Allergy Clin Immunol* 2002; 2:201–205.
43. Milgrom H, Curran-Everett D. Unrecognized nonadherence masquerades as drug resistance. *Curr Opin Allergy Clin Immunol* 2012; 12:219–220.