The challenges of a Big Data Earth

Geoffrey Boulton

Committee on Data for Science Technology (CODATA) & School of Geoscience, University of Edinburgh, Grant Institute, Edinburgh, UK

ABSTRACT

The potential of big data fused with the vision of a digital Earth offers powerful opportunities to deepen understanding of the whole Earth system and the management of a sustainable planet. It is important to stand back from often confusing detail to clarify what those opportunities are and how they might be seized. The essential scientific potential of data, big or small, is to reveal patterns, which have often been the fundamental first step in stimulating inquiry, leading to new questions, new perspectives and potentially to new answers. The digital revolution has created a “digital microscope” that permits us to see patterns that have not been seen before, and when coupled with machine learning technologies to analyse them in creating statistical predictions of the behaviour of both human and non-human systems. These potentials converge with the imperative to represent an Earth system with interacting non-human and human components, as a vital contribution to the understanding and actions required in working towards planetary sustainability. But a digital Earth is also capable of being represented mathematically as a digitally networked phenomenon, analogous to an analogue computer, and should be an important target for a Big Earth Data Journal. We should also return to Al Gore’s vision of an accessible digital Earth with wide usability. Pre-determining the separate functions of parallel digital Earths risks losing one of the great potentials of big data and learning algorithms, the identification and analysis of unanticipated relationships and processes.

1. The digital revolution, big data and revealing patterns in nature and society

A technological milestone was passed at the turn of the millennium when the global volume of data and information that was stored digitally overtook that stored in analogue systems on paper, tape and disc. It has led to an immense increase in the annual rate of data acquisition and storage and dramatically reduced its cost. In 2003, humanity created about 5 billion DVDs worth in one year. In 2014, we created that amount of data every 10 min. In 2003, the
human genome was sequenced for the first time. It had taken 10 years and cost $4billion. It now takes 3 days and costs $1000.

This world of “big data” is one where enormous fluxes of digital data stream into computational and storage devices, often from a great diversity of sensors and sources. It contrasts dramatically with the analogue world and has fundamentally changed the statistical characterisation of phenomena, from one in which the role of statistical manipulation was to maximise the amount of information that could be derived from sparse and discontinuous data, to one in which the flux and diversity of data is so great that operators rarely succeed in mining their full information content.

The digital revolution’s combination of vast data fluxes and ubiquitous digital communication has already had profound impacts on societies and economies, and is now making deep inroads into the way that science is done. Insofar as we currently understand its full implications for science, they are to dramatically increase our capacity to infer patterns in phenomena, whether physical, chemical, biological or human, or in the complex systems that are at the heart of most global challenges. This is not a trivial potential. Observations of patterns in nature and society have often been a fundamental first step in stimulating inquiry, by prompting the questions “how and why?”, leading to new questions, new perspectives and potentially to new answers, whether in the hands of a Copernicus, Newton, Mendeleev, Darwin or Marx. The massive data fluxes now available from diverse sensors, which offer diverse perspectives on complex phenomena, permit us to identify deep, complex patterns that were previously beyond resolution. There is a powerful analogy with Leeuwenhoek’s invention and use of the microscope. It revealed a microscopic world that had not hitherto been seen. The digital revolution has created a digital microscope that permits us to see patterns that have hitherto been similarly beneath our capacity to resolve.

From the perspective of a digital Earth, predictive models based on theoretical understanding of Earth systems, whether human or non-human, have large, often unquantifiable uncertainties and demonstrate chaotic non-linearities in their behaviour. Consequently, the most detailed understanding that we currently have about most of the Earth’s sub-systems falls far short of what would be needed for computer simulations of their operation, although some partial models have been created (Collins et al., 2005). In contrast, the analysis of big data is not based on understanding causality, but in demonstrating deep statistical relationships. Google Translate knows nothing of grammar or the structure of language, but deduces translations from the statistical relationships between thousands of pre-existing texts translated from one language into another.

2. Big data enables powerful machine learning in creating statistically based predictions

Learning algorithms created by the Artificial Intelligence community have been available for many decades, but their impact has been slight because of the relative paucity of the information that they require for other than simple, learned behaviour. Learning algorithms are now coming into their own because of the fluxes of big data that non-trivial “intelligent” behaviour needs as its fuel (Donnelly et al., 2017).

Learning algorithms can now be fed with immense and varied data streams, which are the equivalent of empirical experiences, from which a device can learn to solve problems of great complexity, and without the prejudices that often inhibit human learning. The process
can be simply described, for example in the case of modern weather forecasts, where learning algorithms are integrated into predictive simulations of ocean/atmosphere dynamics to create great improvements in predicting relatively short-term changes in highly turbulent global weather systems. An initial condition for computer models is provided by characterisation of the state of the global atmosphere and ocean from ground and satellite observations. The equations of motion are then applied to these conditions to simulate evolution of the ocean/atmosphere system on a rotating planet. After a given time period, the state of the simulated system is compared with the actual state of the system, from which the simulated system will have inevitably drifted. Multiple iterations between reality and simulation allow a computer to learn how the simulated state should be adjusted to fit reality for different states of the atmosphere, thereby considerably enhancing the accuracy of forecasts. The process does not involve learning about atmosphere/ocean physics, chemistry or biology, but is immensely powerful in identifying multi- or hyper-dimensional patterns.

In the same way, identifying patterns that include social behaviour on a digital Earth is conceptually and technically tractable, provided there is access to adequate data streams (e.g. Blumerman, 2015). It does not depend on any prior understanding of social processes, but has the capacity to yield correlative associations, both between indices of social dynamics and between these and non-human processes.

3. Re-defining the Earth system to integrate human and non-human processes

Studies of the Earth as an integrated system have developed in recent decades largely through the work of geologically inspired Earth Scientists, dealing with the way in which a rocky, watery, gaseous, life sustaining planet operates as part of a solar system. It is, however, a perspective that has arbitrarily stopped at the boundary of the human. This might once have been appropriate, in thinking of an Earth in which although humanity chopped down forests, built cities and created farmed monocultures, actions that were implicitly or explicitly assumed not to change fundamentally the dynamic operation of the planet. This is no longer a sustainable concept for a planet in which one of its organic species, humanity, is recognised to have had physical, chemical and biological impacts that are on a par with the impact of the non-human processes that have determined the planet’s prior evolution, and that have the demonstrable capacity to change its future.

If we truly seek to create a science of the Earth System that contributes to the potential for a sustainable and just world, it must include rather than exclude its human populations and their artefacts. Although such a view was implicit in the concept of environmental science that developed in the 1960s, it is a perspective that has now been more generally adopted through the concept of the Anthropocene (Waters et al., 2016). The Earth of the Anthropocene cannot be understood without integrating into our understanding the consequences of a population of individually and communally self-aware humans and the global digital networks that it has created. These networks are having profound impacts on traditional social and political siloes: penetrating and disrupting them, such that many are in their death throes. The precise formulation of a concept of this Earth is vital to the concerns of scientists and of the political and social systems that we seek to understand. It should be central to the work of such programmes as Future Earth (see Box 1), and an essential part of
the rationale underlying the proposed merger of the International Council for Science (ICSU) and the International Social Science Council (ISSC).

4. The Earth as an information system

The Earth is now, and the Future Earth will increasingly be, a digitally networked Earth; with individuals, societies and their institutions, and almost all powered devices, generating, receiving and creatively utilising exponentially increasing data and information fluxes. Although the ways that citizens and institutions adapt to and use the capacities of this new Earth are highly uncertain, what is not in doubt is the magnitude of the impact that digital processes and global networking have already had and the potential that they have for future disruptive change. For scientists, the explosive growth in the diversity and volume of data available to us has created the potential for new understanding of complex systems on all scales, from the molecular to the cosmic, and all in areas of human concern, from cultural artefacts to local health systems to global sustainability.

A complex system that we need to understand, and one that is rarely ranked amongst global challenges, is that of this novel, digitally networked Earth itself. What sort of system is it? What are its fundamental properties? Could we characterise it mathematically as an information system? And how might it evolve? These are the sorts of questions that natural scientists ask of the systems they study, and ones that we should try to answer of the networked Earth system. If we are able to characterise this digitally networked Earth as an information system similar to an analogue computer, we would then be well placed to understand how it might respond to internal or external perturbations or demonstrate unforced oscillations.

Answers to these questions are fundamental challenges to a developing community concerned with a Big Earth Data, and in particular to mathematicians and systems engineers. They would provide a description of the intrinsic behaviour of a self-organising Earth system that could powerfully frame other analyses of the system.

5. Digital Earth: re-invigorating Al Gore's vision

Al Gore’s 1998 idea of a Digital Earth (Gore, 1999) was to turn a flood of raw data into understandable information about our society and our planet. It would be a multi-resolution, three-dimensional, visual representation of the planet, in which we could embed vast quantities of geo-referenced data, which would be widely accessible, not only to governments, scientists and business, but also to citizens and students in improving their understanding of the Earth and world they inhabit, and supporting them in their roles as responsible citizens. While component parts of that vision have been realised, in such virtual globe geo-browsers as Google Earth and Microsoft’s Bing Maps 3D, for a variety of commercial, social and scientific applications, Gore’s vision of a truly global, collaborative linking of systems remains unaccomplished.

Box 1. Future Earth
Launched in 2015, Future Earth is a 10-year initiative to advance Global Sustainability Science, to build capacity in this rapidly expanding area of research and to provide an international research agenda to guide natural and social scientists working around the world.
The vision has continually been re-interpreted and re-defined by the growing global community of interest. For example, the Joint Research Centre of the European Commission published a position paper on the “Next-Generation Digital Earth” (Craglia et al., 2008). It argued for multiple connected globes/infrastructures to address the needs of different audiences, to be problem oriented, to enable open access across multiple platforms to data, for information, services, models, scenarios and forecasts, and to permit interactive and exploratory work. Although such pre-determining of the silos of use is technically understandable, the absence of a platform that integrates global data with multiple different dimensions loses a fundamental capacity inherent in the world of big data: the discovery of unanticipated, emergent correlations – unknown unknowns (see Box 2). It implicitly assumes that we know what the problems are, presumably determined by established experts, rather than being available to be used by a wider group that is free to determine what experiments might be done and what relationships should be explored.

Box 2. “Unknown unknowns”
Donald Rumsfeld, former US Secretary of Defence, commented in 2002 that “…..there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don’t know we don’t know. …….. it is the latter category that tend to be the difficult ones”.

6. Integrating data for interdisciplinary understanding

The potentials of big data and of a Big Data Earth in analysing and understanding the complexity that lies at the heart of most major global challenges are fundamentally dependent on successfully addressing three major issues across the disciplines of science:

(a) The default position for publicly funded scientific data should be openness (Boulton et al., 2012, 2015), as exemplified by the FAIR principles (Findable–Accessible–Interoperable–Reusable – Force 11, 2015). However, for a digital Earth to come close to the Gore vision, access to the data holdings of governments and their agencies, particularly in Earth observation, is also crucial, as is agreement about access to the data holdings of the private sector, particularly in the so-called “internet of things”. Conversations about these issues, and about the modes of governance of the Earth’s data resources, now and into the future, are important priorities, and should be on the agenda of the United Nations.

(b) As yet, only a sub-set of the disciplines of science have developed ontologies and vocabularies that permit them efficiently to discover, manage and use the data that is theoretically available to them in unleashing the potential for game-changing discoveries.

(c) The different disciplines in which (b), above, has been effectively done, use many varying standards in their data systems, such that the integration of data from across the disciplines that would be required to support interdisciplinary research is inhibited. Adoption of a single standard would be impracticable. What is required is an interface between different standards that permits integration; a major problem that requires a non-trivial, probably decadal, effort.
These issues are fundamental in determining the way that science will evolve in the twenty-first century (a) is a highly sensitive issue, but one which the representative bodies of the international scientific community should begin to address. (b) and (c) are imperatives that are currently being coordinated through creation of a “commission on standards” by CODATA and its disciplinary and technical collaborators, and with the support of ICSU and ISSC. There are also profound ethical issues to be addressed in relation to privacy and confidentiality, in relation to who owns the data, the burgeoning problem of cyber security and data governance (Leyser et al., 2017).

7. Open and closed systems: combining holistic and reductionist approach

At a fundamental level, the digital perspective of the Earth should be one of an open, coupled system that is also open to cosmic processes, most obviously those of radiation, gravity and physical impact. Some coupling is weak, some strong, and many couplings vary in their intensity through time. It is, however, important to recognise that that reductionist approaches that ignore weak interactions, thereby assuming that strong interactions are contained within a closed system, often of relatively simple cause and effect, have been powerful tools of understanding in the scientific armoury, and will continue to be so. One of the powerful potentials of big data, is to help determine where and when interactions are strong and when weak. Digital Earth approaches need to embed these distinctions and the opportunities they offer in their work.

8. The challenges

The challenges I have posed provide a menu of issues for the Big Earth Data Journal and its contributors, but also for the International Society for Digital Earth. They will demand a clarity and boldness of vision if they are to be addressed in ways that serve human understanding of a holistic Earth system and the challenge of planetary sustainability.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Blumerman, M. N. (2015). Big Data and censuses: The challenge for the 2020 world round of population and housing censuses. Retrieved from http://unstats.un.org/unsd/statcom/statcom_2015/seminars/population/docs/.

Boulton, G. S., Babini, D., Hodson, S., Jianhui, L., Marwala, T., Musoke, M., … Wyatt, S. (2015). Open data in a Big Data world. Science International. Retrieved from www.science-international.org
Boulton, G. S., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., … Wallport, M. (2012). Science as an open enterprise. Royal Society Report. Retrieved February 5, 2017, from https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/
Collins, N., Theurich, G., DeLuca, C., Suarez, M., Trayanov, A., Balaji, V., … da Silva, A. (2005). Design and implementation of components in the earth system modeling framework. International Journal of High Performance Computing Applications, 19(3), 341. doi:10.1177/1094342005056120
Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., … Parsons, E. (2008). Next-generation Digital Earth. International Journal of Spatial Data Infrastructures Research, 3, 146–167.
Donnelly, P. D., et al. (2017, April). Machine learning: The power and the promise of computers that learn by example. Royal Society Report. Retrieved from https://royalsociety.org/…/publications/machine-learning-report.pdf
Force 11. (2015). Guiding principles for findable, accessible, interoperable and re-usable data. Retrieved from https://www.force11.org/group/fairgroup/fairprinciples
Gore, A. (1999). The Digital Earth: Understanding our planet in the 21st century. Photogrammetric Engineering and Remote Sensing, 65(5), 528–530.
Leyser, O., Richardson, G., Agar, J., Hall, W., Chessell, A., Floridi, L., … Yeung, K. (2017). Data management and use: Governance in the 21st century. Retrieved from https://royalsociety.org/topics-policy/projects/data-governance
Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, B. D., Poirier, C., Galuszk, A., … Wolfe, N. A. (2016). The Anthropocene is functionally and stratigraphically distinct from the Holocene. Science, 351(6269), aad2622.