COMMENT

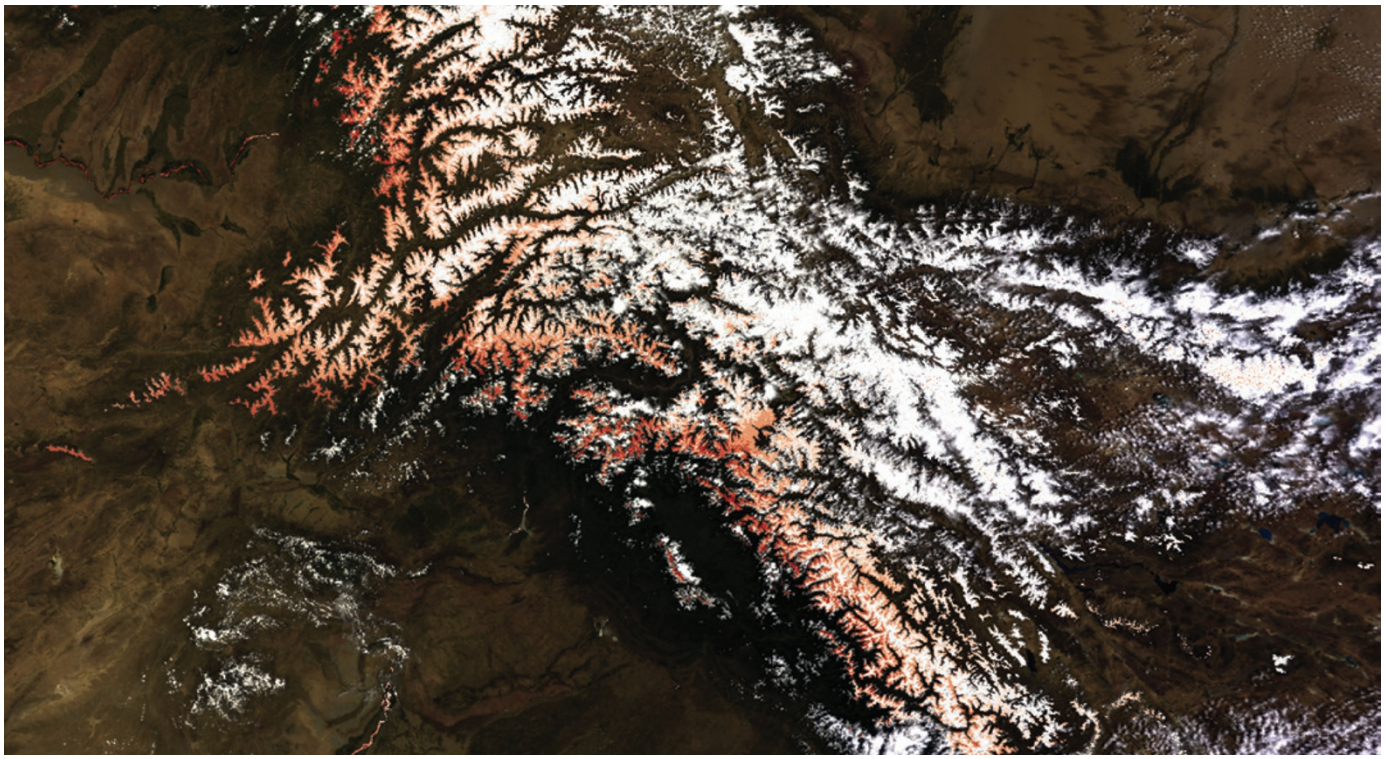
ENERGY Critics of energy-efficiency policy overplay the rebound effect **p.475**

ANTHROPOLOGY Jared Diamond's paean to traditional societies, reviewed **p.477**

HISTORY Heroism, intrigue and posturing abound in a history of Antarctica **p.478**

PETITIONS Ganging up on research damages scientific discourse **p.480**

ANN C. BRYANT & THOMAS H. PAINTER, JPL/CALTECH SNOW OPTICS LAB.



A satellite image of snow on the Hindu Kush mountains in Asia, with regions of high absorption of sunlight by dust and black carbon shaded in red.

A vision for data science

To get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers, says **Chris A. Mattmann**.

Two small words — 'big data' — are getting a lot of play across the sciences. Funding agencies, such as the National Science Foundation and the National Institutes of Health in the United States, have created million-dollar programmes around the challenges of storing and handling vast data streams. Although these are important, I believe that agencies should focus on developing shared tools for optimizing discovery.

Big data are big in three ways: the volume of information that systems must ingest, process and disseminate; the number and complexity of the types of information handled; and the rate at which information streams in or out. Terabyte-sized data sets (10^{12} bytes) are now common in Earth and space sciences, physics and genomics (see 'Data deluge'). But a lack of investment in services such as algorithm integration and file-format translation is limiting the ability to manipulate archival data to reveal new science.

At the Jet Propulsion Laboratory (JPL) in Pasadena, California, I am a principal investigator in a big-data initiative, pursuing projects on data archiving and mining, smart algorithms and low-power hardware for astronomy and Earth science. Rather than finding one system that can 'do it all' for any data set, my team aims to define a set of architectural patterns and collaboration models that can be adapted to a range of projects.

I believe that four advancements are necessary to achieve that aim. Methods for integrating diverse algorithms seamlessly into big-data architectures need to be found. Software development and archiving should be brought together under one roof. Data reading must become automated among formats. Ultimately, the interpretation of vast streams of scientific data will require a new breed of researcher equally familiar with science and advanced computing.

ALGORITHM INTEGRATION

A project by my team at the JPL illustrates the challenges of working with big data. In 2011, we were asked by the US National Climate Assessment to establish a ▶

computing facility to integrate a range of snow-related measurements — and to do so in a month. The data included observations from the western United States, Alaska and the Hindu Kush–Himalayan regions, as well as the entire Earth-observing record since 2000 and subsequent monitoring. The data products and maps would amount to several hundred terabytes.

The algorithms to be incorporated were varied, and included codes for estimating snow coverage, grain size and absorption of solar radiation by dust and black carbon[1]. They had been written in IDL, a specialized programming language used by many researchers. Geographers, remote-sensing experts and software programmers contributed.

Most computer scientists would assume that such a system would take years, not weeks, to develop. The algorithms would presumably have to be rewritten in a standard language such as C++, Java or Python, or one that could run on a fast computer system or infrastructure, such as Google's MapReduce model.

But, in my experience, there is no need to rewrite scientific algorithms for big-data systems. Rewriting only increases the barriers to communication between scientists and computer engineers. Rewriting can also introduce costly errors.

Computer engineers should trust scientists to produce executable algorithms, which can be plugged into a larger processing framework. The skill is in tying the input and output files and relevant parameters unobtrusively into the big-data network, so that the algorithm can run seamlessly within it. With a modular approach, development can proceed quickly in parallel — we constructed our snow-science computing facility this way in less than a month.

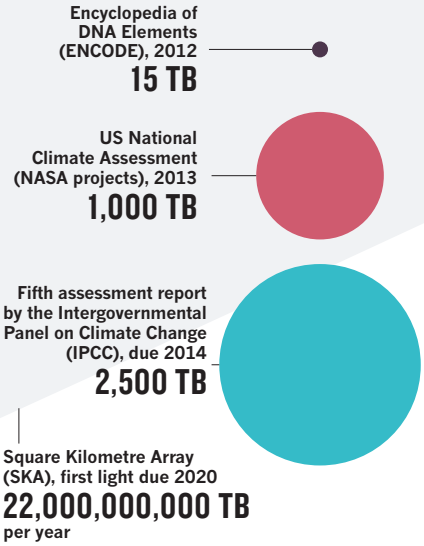
DEVELOPMENT AND STEWARDSHIP

Today, different big-data computing tasks are usually undertaken by different teams. The bulk of agency funding goes to building specific long-standing archives or data grids[2] — systems such as the NASA Earth science Distributed Active Archive Centers or the International Virtual Observatory Alliance in astronomy — that disseminate, preserve and steward[3] data. Large archives have received an average of US$100 million a year from US federal agencies over the past decade.

By contrast, the development, integration and updating of science algorithms receives only between $1 million and $5 million per year in the United States. These tasks are carried out in science-computing facilities, which are often small and transient. Because they must do more for less, such facilities largely use and generate community-based open-source software[4–6]. Examples include Apache Hadoop[7] and Apache Tika[8], used in Earth science, biomedicine and business.

DATA DELUGE

The billions of terabytes (TB) produced in one year by the SKA telescope (grey) will dwarf today's data sets in genomics and climate science.



Although data interpretation and archiving efforts have so far been funded separately and at strikingly dissimilar levels, their needs — such as workflow processing and file and resource management — are complementary and overlapping. As storage and computation costs fall, algorithm developers are moving into preservation, both to archive their own work and to open new research windows on large data sets that were previously closed.

In the next decade, I believe that archives and science-computing facilities must merge. The international radio-astronomy community is doing so in preparation for the Square Kilometre Array radio telescope, due to see first light in 2020. The enormous volume of data that the array will produce — 700 terabytes each second — will, after just a few days, eclipse the current size of the Internet. Archives in the United States such as those at the National Radio Astronomy Observatory's Expanded Very Large Array and the Atacama Large Millimeter/submillimeter Array are developing software to handle that deluge.

MANY FORMATS

Big-data systems must deal with thousands of file types and conventions. The communities that have formed around information modelling, ontology and semantic web software address this complexity of data and metadata (descriptive terms attached to files) to some extent. But they have so far relied on human intervention. None has delivered the silver bullet: automatic solutions that identify file types and extract meaningful data from them.

Comparisons of observational and model data are, for example, under construction for the US National Climate Assessment and the Coupled Model Intercomparison Project of the Intergovernmental Panel on Climate Change. NASA uses the Hierarchical Data Format version 5 (HDF-5) and the HDF-Earth Observing System metadata representation. The outputs of climate models are stored in the Network Common Data Form, typically with climate and forecast metadata conventions[9]. Automatic methods will be needed to match and analyse these data, which amount to petabytes (10^{15} bytes).

Some big-data fields are switching to formats like these that have better support. Astronomers, for instance, are turning to NASA's HDF-5 file format from the Flexible Image Transport System that has been their standard. But history shows that defining a single, unifying file format is not the answer, because proliferation of file types will continue. Instead, we need a toolkit of automatic ways to boil file formats down to their essence, and more formats that are amenable to those approaches. We need flexible systems that can perform multiple functions and deal with diverse data. Encouraging efforts are under way, including with Apache OODT[10] and Apache Tika[8].

PEOPLE POWER

To solve big-data challenges, researchers need skills in both science and computing — a combination that is still all too rare. A new breed of 'data scientist' is necessary.

As well as being data stewards, data scientists will develop bespoke algorithms for analysis and adapt file formats. They will understand the mathematics, statistics and physics necessary to integrate science algorithms into efficient architectures. They will find solutions beyond the fragmented community efforts that have dominated the past decade of development of big-data systems.

Funding agencies should support computing facilities that combine big-data stewardship and software development, employing data scientists to bridge the gap. Coordination between agencies is crucial to avoid duplication. The Big Data Senior Steering Group, linking efforts across the National Science Foundation, the National Institutes of Health, NASA and others, is a promising early example. More oversight will be needed to establish new working patterns.

Because big-data fields stretch across national as well as disciplinary boundaries, such facilities and panels must be international. In centres of excellence around the world, such as the JPL, data scientists will help astronomers and Earth scientists to share their approaches with bioinformaticians, and vice versa.

For the specialism to emerge and grow, data scientists will have to overcome barriers that are common to multidisciplinary research. As well as acquiring understanding of a range of science subjects, they must gain academic recognition. Journals such as the *Data Science Journal* should become more prominent within the computing community. Software products and technologies should be valued more by academic committees.

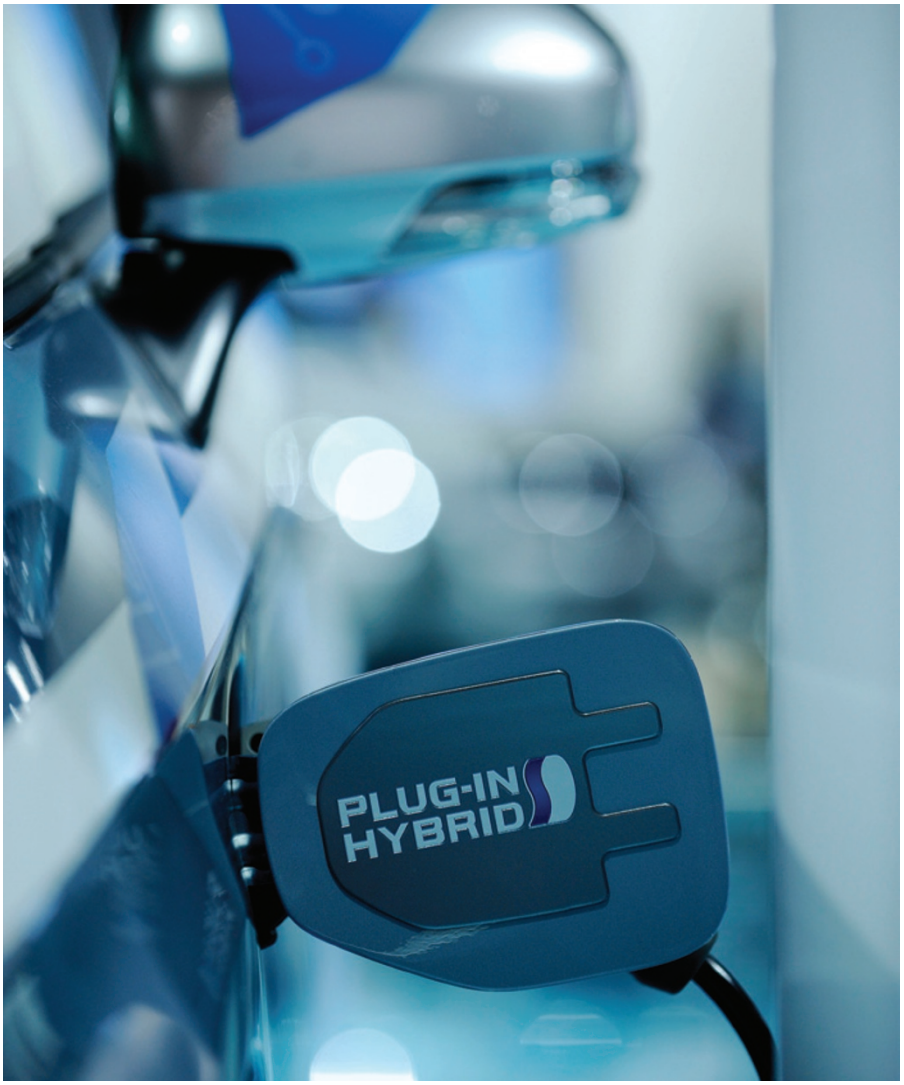
New interdisciplinary courses will be needed. The University of California, Berkeley, and Stanford University in California have set up introductory courses for computer scientists on big-data techniques — more universities should follow suit. Natural scientists, too, should become familiar with computing and format issues.

In my lectures for computer-science graduates, I have brought together students at the University of Southern California in Los Angeles with researchers at the JPL. Using real projects, my students see the challenges awaiting them in their future careers. I hope to employ some of them on the projects that will flow from the JPL's big-data initiative. The technologies and approaches that they develop will spread beyond NASA through contributions to the open-source community.

Empowering students with knowledge of big-data infrastructures and open-source systems now will allow them to make steps towards addressing the major challenges that big data pose. ■

Chris A. Mattmann *is a senior computer scientist at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA, and adjunct assistant professor in computer science at the University of Southern California, Los Angeles, California 90089, USA.*
e-mail: chris.a.mattmann@nasa.gov

1. Painter, T. H., Bryant, A. C. & Skiles, S. M. *Geophys. Res. Lett.* **39,** L17502 (2012).
2. Foster, I., Kesselman, C. & Tuecke, S. *Int. J. High Perform. Comput. Appl.* **15,** 200–222 (2001).
3. Lynch, C. *Nature* **455,** 28–29 (2008).
4. Morin, A. *et al. Science* **336,** 159–160 (2012).
5. Spinellis, D. & Giannikas, V. *J. Syst. Softw.* **85,** 666–682 (2012).
6. Ven, K., Verelst, J. & Mannaert, H. *IEEE Software* **25,** 54–59 (2008).
7. White, T. *Hadoop: The Definitive Guide* 2nd edn (O'Reilly Media/Yahoo Press, 2010).
8. Mattmann, C. A. & Zitting, J. L. *Tika in Action* (Manning, 2011).
9. Cinquini, L. *et al. Proc. 2012 IEEE 8th Int. Conf. E-Science* Chicago, Illinois, 8–12 October 2012 (in the press).
10. Mattmann, C. A., Crichton, D. J., Medvidovic, N. & Hughes, S. in *Proc. 28th Int. Conf. Software Engineering (ICSE06), Software Engineering Achievements Track* 721–730 (2006).



ZHANG JUN/XINHUA PRESS/CORBIS

Fuel-efficient cars cost less to run, so people might use them a little more.

The rebound effect is overplayed

Increasing energy efficiency brings emissions savings. Claims that it backfires are a distraction, say **Kenneth Gillingham** and colleagues.

Buy a more fuel-efficient car and you will spend more time behind the wheel. That argument, termed the rebound effect, has earned critics of energy-efficiency programmes a voice in the climate-policy debate, for example with an article in *The New York Times* entitled 'When energy efficiency sullies the environment'[1].

The rebound effect idea — and its extreme variant the 'backfire' effect, in which supposed energy savings turn into greater energy use — stems from nineteenth-century economist Stanley Jevons. In his 1865 book *The Coal Question*, Jevons hypothesized that energy use rises as industry becomes more efficient because people produce and consume more goods as a result[2].

The rebound effect is real and should be considered in strategic energy planning. But it has become a distraction. A vast ▶