BIG DATA

THE NEW CULTURE OF BIG DATA

The ALMA correlator, one of the most powerful supercomputers in the world, is in a remote, high-altitude site in the Andes of northern Chile. (Photo: ESO)

The data revolution is transforming science, but taking full advantage requires resources, cooperation, education and a change to the culture of science itself. How can less-developed countries keep from being left behind?

by **Sean Treacy**

In the push to create a prosperous future for the planet, data are everything. Policymakers who want to make informed decisions rely on advice from scientists who derive knowledge from data. Well-managed data tell us where poverty is worst and how many children are undereducated. They show where more doctors are needed, and provide vital information about food and water supplies.

And now? A digital revolution has exploded the amount of data that researchers must manage and analyse, and those data move at an unprecedented speed. There are over 4.6 billion mobile-phone subscriptions worldwide. International technology company IBM estimates that roughly 90% of the world's data have been created in the last two years.

The increased volume and speed of data is known as "big data", and science will have to undergo a complex evolution to accommodate it. It will require time and money for computing power. The culture of science must become more open so that researchers make their data available to others who might also make use of it. Countries that may have different research interests and technological capacity will have to find a common ground on open access. And finally, scientists in developing countries will need to learn how to take advantage of the programmes, coding techniques and analytical methods of information-crunching computer scientists.

"Researchers are beginning to produce lots of data, but many don't have the skills, or the infrastructure and support to take advantage," said Simon Hodson, the executive director of CODATA, the International Council for Science's Committee on Data for Science and Technology. "Our challenge is to ensure we take advantage of these technologies."

The world's emerging economies are already taking advantage of the revolution. China has been making a major big data push, finishing the fastest supercomputer in the world in June, capable of making 93 quadrillion calculations per second. According to India's National Association of Software and Services Companies, the country is already one of the top ten data analytics markets in the world, with a USD2 billion data analytics sector that is projected to grow to USD16 billion by 2025.

Investments in India, China and other nations are a start. But the campaign to achieve the UN's Sustainable Development Goals by 2030 is accelerating, and that alone creates an urgency for bigger, broader investment in big data capacity.

The SDGs are the international community's most ambitious effort ever to resolve humanity's greatest challenges and set a firm course for sustainable global development. The 17 goals aim to eliminate poverty, improve health and achieve urgent improvement in the Earth's environment by 2030.

The UN also approved 169 targets within those goals. Big data and data science will be critical for measuring progress on those targets as quickly and exactly as possible. With that in mind, the UN convened its first World Data Forum in South Africa in January 2016, where data scientists with expertise in statistics, measurement and information systems, and others focused on strategies to employ data for sustainable development.

The task is more daunting for developing countries that are severely short on technological

resources and expertise – a gulf known as the global digital divide. The challenge will be most pressing in the Least Developed Countries (LDCs), which will need technical and financial support to build big data skillsets, said Amina J. Mohammed, the Minister of Environment of Nigeria and former special adviser to UN Secretary-General Ban Ki-moon on Post-2015 Development Planning. While funding will be a part of the solution, Mohammed said, building local expertise, improving policy making, and empowering citizens are all critical.

"Data literacy must be improved," she explained in a written interview, "or else we risk seeing the rise of yet another digital divide."

THE VIEW FROM AN LDC

In order for Least Developed Countries to take advantage of big data, they have to recognize that powerful computing technology and skills are a pressing need for the public good. And LDCs have a long way to progress. Consider Nepal: Over 90% of all Nepalese school children don't have access to computers, said TWAS Fellow Bishal Upreti of Nepal, who is also the TWAS Council Member representing East and Southeast Asia.

> "Data literacy must be improved, or else we risk seeing the rise of yet another digital divide." *Amina J. Mohammed*

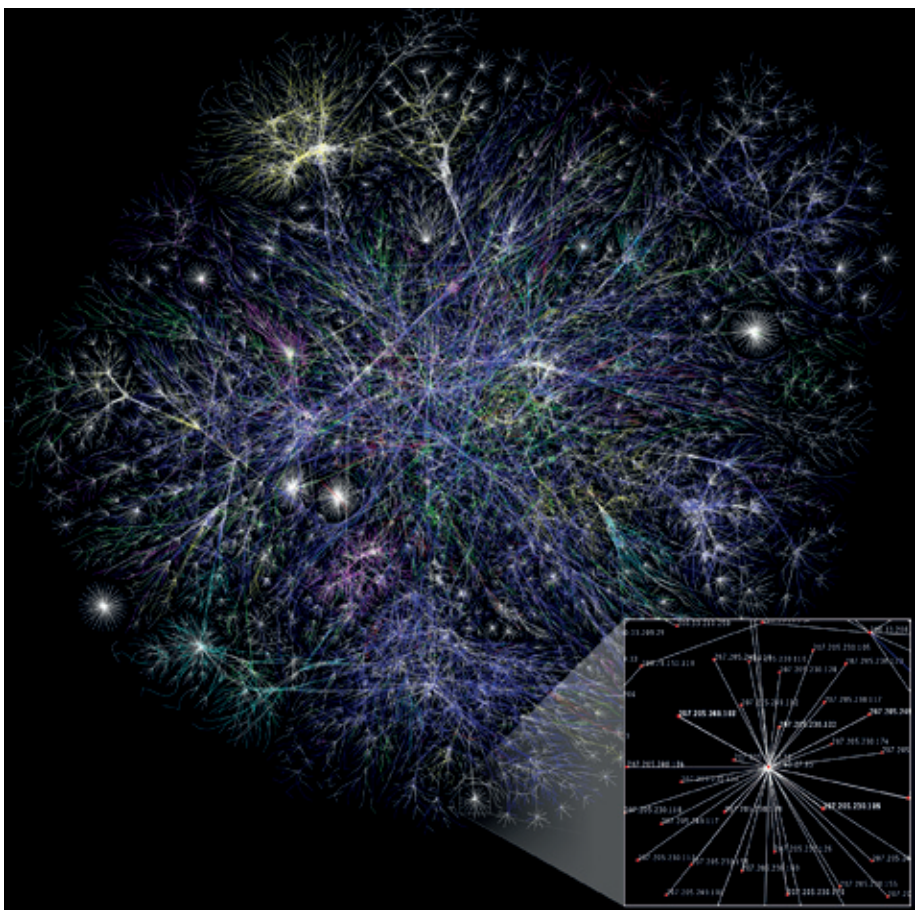
"It is a big danger that Nepal is producing a computer-illiterate generation in the future, when the contrary should be the case," he said. "The present economic condition of the country makes it hard to provide these facilities to all school-going children."

The IT-sector graduates that Nepal produces tend to go into corporate sectors, such as cell phone companies, he added, instead of public services that are short on resources. And big data is important for countries like Nepal that urgently need to provide better public services in agriculture, education and even preparedness for earthquakes. On 25 April 2015, an enormous quake struck the country, killing nearly 9,000 people, injuring many more, and destroying homes, businesses and even entire villages.

Upreti's specialty is Himalayan geology, and in the days after the quake, he provided advice and technical information on the disaster. But the earthquake also laid bare Nepal's inability to handle a large flow of data and information quickly. During the emergency, officials made decisions the old-fashioned way: getting information and relaying directions mostly through the security agencies' radios while the army and police handled search and rescue efforts. They did not have a digital database of emergency supplies for such a large a disaster, Upreti added.

When the quake happened, it was also a chance to prepare for the future – to learn what areas of Kathmandu Valley are most vulnerable and improve building codes there. But still, Nepal only had a small number of sensing stations in Kathmandu, and also didn't have

▲ From top: Simon Hodson, Amina J. Mohammed.



▲ Damage in Kathmandu, Nepal, from last year's devastating earthquake, which killed nearly 9,000 people. (Photo: Bishal Upreti)

◀ A data-derived partial map of the Internet. (Photo: The Opte Project)

the ability to analyse what data they collected. Instead, they needed to send the data to the United States for analysis.

Upreti argued that Nepal needs to upgrade its earthquake recording and data analysis abilities. At the time of the quake, he explained, Nepal had only 21 permanent seismic stations capable of automatically recording information when a quake strikes. They also only had 20 permanent GPS stations monitoring ground movements that researchers had to visit every three to six months to collect the data. After the earthquake, Nepal took steps to improve their data collection system, adding 80 temporary seismic stations and 50 new temporary and permanent GPS stations. But they still need an investment in big data management tools and skills to handle all the new data from these stations.

It's policymakers who must set the tone for solving the problem, Upreti said. Nepal and other countries need to prioritize big data management. "It's about the mind-set," he said.

▼ From left: Bishal Upreti, Ciira wa Maina, Esther Musyimi.

A NEED IN AFRICA

The issue is also pressing in Africa. Many of the continent's problems could be better addressed if scientists there made better use of data, said electrical engineer Ciira wa Maina of Dedan Kimathi University in Nyeri, Kenya.

Maina envisions a project to monitor livestock using data-collecting devices attached to animals. It would rely on smartphone-sized sensors that measure how fast the animals move, how much they move, and how steady they are. Once computers sift through all the sensor data to determine behavioural patterns, they could tell you when an animal is unwell, dealing poorly with the weather, or even in heat.

"If it sees that a set of animals aren't behaving as they normally do, you can start to ask: What's the reason?" said Maina. "One reason a cow could be agitated is because it's in heat and it can be a costly exercise to miss that cycle."

Putting such a project into practice would require powerful computers, big data software and local data scientists to quickly turn all this data into information that farmers can use. So Maina co-organized a data science workshop in Nyeri last year with about 100 students, all Kenyan.

Esther Musyimi of Dedan Kimathi University of Technology was one of the students. She said big data education is not common in Kenya, and that the workshop drove home how a telecommunications engineering student can take advantage of big data to help local development.

"Telecommunications has given people the ability to communicate with the rest of the world," she said. "You are able to access big data from any part of the world using these devices."

Maina added that underlying the need for new data is the need for a shift in attitude. Researchers and data engineers will have to climb out of their comfort zones and better understand each other's work.

BUILDING A CULTURE FROM THE GROUND UP

How does this attitude shift take place? One important thing to acknowledge is that a data scientist is a new breed of scientist, said Clement Onime of Nigeria, a systems and network

analyst for the Abdus Salam International Centre for Theoretical Physics (ICTP) in Trieste, Italy. Almost every scientific field needs experts who can handle large volumes of data. They also have to be more willing and able to share their data with each other.

In Onime's view, the newness of the field also presents a special opportunity for developing countries. "There is no country that is particularly ahead of the order right now," he said. "So it's an excellent opportunity for LDC countries not to play catch-up so much, but to stay abreast of what's happening."

This is one reason why Onime co-organized an open data science course this year at ICTP. Put together by The Research Data Alliance (RDA), CODATA, ICTP and TWAS, the CODATA-RDA School of Research Data Science sought to help a younger generation of scientists from developing countries learn to work with an open, common interface and make large volumes of data available to each other across disciplines.

About 75 students, including participants from over 30 developing countries, learned how to code open-source data programmes, analyse and manage data and visualize data in easy-to-understand graphics. They were then encouraged to duplicate the lessons of the course in their own countries.

One student, Bianca Peterson, is working on her genetics PhD at North-West University in Potchefstroom, South Africa. Many supervisors in her university are uncomfortable with sharing data from yet-to-be-published research. "My supervisor was worried that someone else would publish based on my data before I even get my PhD and then, suddenly, I wouldn't have a project anymore," she said.

But realistically, Peterson noted, that's unlikely to happen. She will always know her own data the best, and there are benefits to an open-data culture as well. At one point, she learned that she had been duplicating another researcher's work, so she had to retract her own paper. If scientists made their data quickly and publicly available to researchers in her field, she would have avoided lost time and resources.

Now she's planning to replicate the Trieste course in her home country. This is a hope shared by her fellow course attendee Elias Mwakilama, a computational mathematician with the University of Malawi. While data-sharing has entered the discussion among scientists in Africa, it has yet to be put into practice, he said. "We need to build the culture now, from the ground up."

▲ Students receive guidance during the CODATA-RDA School of Research Data Science in Trieste, Italy, earlier this year. (Photo: CODATA International)

COMPUTERS FOR ASTRONOMY – AND MORE

There is one enormous project taking place in Africa in particular that is building a culture for big and open data.

The Square Kilometre Array (SKA) is a massive radio telescope planned to work simultaneously in both Australia and South Africa for which construction will begin in 2018. Its astronomy work in Africa will manage massive amounts of data quickly – handling anywhere from one to 10 terabytes per second.

The presence of an enormous, first-of-its-kind data infrastructure is an opportunity for

▼ From left: Clement Onime, Bianca Peterson, Elias Mwakilama.

> The ability to extract value from data and do something useful with it is going to be really crucial to how we move forward as a species.
>
> *Jasper Horrell*

Africa to lead the way in big data education and training, and in tapping the power of supercomputers for a variety of uses. Jasper Horrell, the general manager of science computing and innovation at SKA, said plans are in the works to use this data to assist African countries, including some LDCs.

"Africa is the neglected continent in terms of technology and science," said Horrell. "This is a real opportunity to try to make a difference to a lot of people's lives in South Africa and beyond and also to open up the collaborative opportunities between South Africa and those other countries."

SKA is central to another project – the African Data Intensive Research Cloud, meant to link up Africans in partner countries: Botswana, Ghana, Kenya, Mauritius, Namibia, and three LDCs – Zambia, Mozambique and Madagascar. The goal is to connect research groups and animate new astronomy projects. But it will also help other scientific endeavours, such as green energy, environmental monitoring and health.

For example, to test drugs, scientists need to assess a huge number of people, and do so in a centralized way. Then doctors can do remote medical testing, remote diagnosis, and get the right medicine to the right places. A system geared toward data collection would make following-up easier, which is important on the ground in Africa. It could help a country be able to track contagious diseases in real time.

"The data revolution is here," said Horrell. "It's accelerating, and it's going to affect all areas of human activity. The ability to extract value from data and do something useful with it is going to be really crucial to how we move forward as a species." ◘

▲ An artist's impression of the 5km diameter central core of The Square Kilometre Array antennas. (Image: Swinburne Astronomy Productions for SKA Project Development Office)