



BIG DATA REVOLUTION

Big data, referring to large and complex data sets that are difficult to process using traditional computational hardware and software, is a popular term that has gained ground in the last decade. Researchers in many fields are now engaging in interdisciplinary collaborations across continents where the volume, velocity and variety of data exceed their capacity to extract information from it. UCT is taking the lead in creating the framework to allow African researchers to get to grips with big data and turn it into actionable knowledge.

THE BIG PICTURE OF BIG DATA



Above: A central cooling plant in Google's data centre in Georgia, USA. Previous page: Professor Paul Bonnington in the Monash CAVE2™, which is used to visualise big data. (CAVE2 is a trademark of the University of Illinois Board of Trustees). Image: Paul Jones, Coretext.

The modern world is experiencing a data deluge: "The data flow is so fast that the total accumulation of the past two years – a zelabyte – dwarfs the prior record of human civilisation," reports Harvard Magazine in March 2014.







Popularly known as big data, this surge of information is set to become as important to business and society as the Internet, providing people with everything from bad-weather warnings to plotting a route home. No-one is immune – least of all researchers, who now work with large data sets that cut across the disciplines, including astronomy, biology, engineering, maths, computer science, archival science, humanities, economics and finance.

In 2001, Gartner, Inc, the world's leading information and technology research and advisory company, defined big data as "high-volume, -velocity and -variety information assets that require cost-effective, innovative forms of information processing for enhanced insight and decision-making." What this means, in a nutshell, is that the more data there is available, the bigger potential there is for more accurate analyses. And, with

more accurate information available, decisions can be made with increasing confidence, while better decision-making leads to greater efficiencies, cost reductions and reduced risks across societies. But this will only happen if the capacity exists to interpret the data correctly. Getting from data to knowledge is not a simple equation and the need to support data-intensive research at universities in such a way to manage this is acute. As Professor Russ Taylor, newly awarded joint UCT/University of the Western Cape (UWC) Square Kilometre Array (SKA) Research Chair and big data champion, says, "Global research leadership requires that we have the capacity to extract information from big data."

Increasingly, researchers in several fields are battling to move data sets between collaborator sites, test various sets of parameters to optimise analysis of large data sets, and facilitate access to big-data sets by international research communities. In the last 12 months, UCT researchers have more than doubled their use of central high-performance computer (HPC) facilities for data-intensive research. Demand for research storage is growing substantially each year, and specialist support and analysts are in constant demand.

UNDERSTANDING BIG DATA

- 
VOLUME
 HOW TO MOVE LARGE DATA SETS AND HOW TO USE ANALYTICS TO CREATE VALUE FROM DATA.
- 
VELOCITY
 DATA IS STREAMING IN AT UNPRECEDENTED SPEED AND MUST BE DEALT WITH IN A TIMELY MANNER.
- 
VARIETY
 DATA COMES IN MANY FORMATS.
- 
VARIABILITY
 DATA FLOWS CAN BE HIGHLY INCONSISTENT, WITH PERIODIC PEAKS AND DAILY, SEASONAL AND EVENT-TRIGGERED PEAK DATA LOADS.
- 
COMPLEXITY
 IT IS STILL AN UNDERTAKING TO LINK, MATCH, CLEANSE AND TRANSFORM DATA ACROSS SYSTEMS.
- 
VERACITY
 IT IS HARD TO KNOW WHICH INFORMATION IS ACCURATE AND WHICH IS OUT OF DATE. POOR DATA QUALITY COSTS THE US ECONOMY \$3,1 TRILLION EACH YEAR.

↑ **THE FUTURE**
4.4 MILLION
DATA SCIENTISTS
 WILL BE NEEDED BY 2015
ONLY 1.5 MILLION
AT PRESENT

"Without the right support, UCT researchers risk diverting time and resources into the development and maintenance of potentially sub-standard or inefficient solutions or just generally taking much more time to do meaningful analysis," says Professor Danie Visser, Deputy Vice-Chancellor with responsibility for research at UCT. "There is increased potential for valuable data or intellectual property to be lost, stolen or corrupted and for significant interruptions to research activity. A centralised effort is needed to provide a mechanism for researchers to learn from one another and develop institutional best practice."

GLOBAL RESEARCH LEADERSHIP REQUIRES THAT WE HAVE THE CAPACITY TO EXTRACT INFORMATION FROM BIG DATA.



It is partly for this reason that UCT has taken the lead in establishing an eResearch Centre (see p61). In line with moves at other leading international research institutions, the centre will provide integrated support across the university research lifecycle and will work in close collaboration with researchers facilitating the delivery of high-impact, locally relevant and internationally competitive research. One of its important roles will be in managing large data sets.

UCT has also been working with other South African institutions to see what support platforms they have in place and where there are opportunities for collaboration. The first eResearch Africa conference, hosted by the Association of South African University Directors of Information Technology (ASAUDIT), was held in October 2013, based on the Australasia eResearch conference model. A delegate was heard to remark, "This is the first time I have ever seen parallel academic and technical tracks at a conference." Exactly what the conference organisers were hoping for.

The purpose of the conference was to bring together practitioners and researchers for a week to share ideas and exemplars on new information-centric research capabilities. eResearch is focused on how information and communications technologies help researchers to collect, manage, share, process, analyse, store, find and re-use information. The success of the event has led to a 2014 conference scheduled for November at UCT that should, judging by the speed of progress, be even bigger and better than the 2013 conference.

BIOINFORMATICS – A BRAVE NEW WORLD

One of the fields at the forefront of big data research at UCT is bioinformatics (the science of collecting, analysing and interpreting complex biological data), and one of the leading researchers in the field is Nicola Mulder, associate professor of bioinformatics and head of the Computational Biology Group at the Institute of Infectious Disease and Molecular Medicine (IDM). Associate Professor Mulder is managing one of the largest grants at UCT, awarded by the US-based National Institutes of Health (NIH) – a US\$13-million grant over five years that is part of the Human Heredity and Health in Africa (H3Africa) initiative, funded by the NIH and the Wellcome Trust. Its purpose is to set up a bioinformatics network to support H3Africa-funded research projects, which aim to identify the genetic bases for illnesses such as diabetes, rheumatic heart disease, cardiometabolic diseases, tuberculosis pharmacogenetics, kidney disease and strokes.

IN ORDER TO ANSWER GENETICS QUESTIONS, IT IS BECOMING MORE IMPORTANT TO GENERATE BIG-DATA SETS.



Part of the H3Africa initiative involves the creation of biorepositories (places to keep samples). This bioinformatics infrastructure includes a staging area for all data and tools for analysis, and building capacity through training programmes, specialised courses and data-management workshops.

Associate Professor Mulder says: “H3Africa researchers will generate big-data sets – 500 terabytes at a minimum – and our network, H3ABioNet, has 34 partners: 32 universities and research institutions in 15 African countries, along with two in the United States needed to build the infrastructure necessary to manage the data. At UCT we have a team of four technical posts and have set up a bioinformatics helpdesk for researchers to request support.”

The project allows Associate Professor Mulder’s team to develop tools, set up pipelines and provide advice about running an analysis, as well as to enable collaborations across research nodes. It also makes provision for internships, which allow H3Africa partners to sit and analyse their data with members of the network.

“Some impacts of the project include the establishment of a bioinformatics centre at a university in Egypt, giving them space and facilities, as well as one in Tanzania. There are plans to build a bioinformatics centre from scratch in Ghana and we have already started training programmes there. In addition, through the network connections, one of our US partners wrote a successful grant proposal with a Niger partner to do some collaborative work. When we were able to demonstrate what H3Africa is doing for bioinformatics in South Africa, as part of their ongoing commitment, the Department of Science and Technology provided money for local bioinformatics courses.”

A huge advantage of this – and one of the stated aims of the project – is that it enables African scientists to retain ownership of their data. “Training African scientists in bioinformatics means we can make sure that data stays on the continent, is analysed on the continent and published in Africa. What has been happening up to now is that researchers have been unable to handle the level and volume of research data, so they would source a collaborator outside Africa and that collaborator would get the material published,” says Associate Professor Mulder.

Bioinformatics research has lower costs than wet-lab research (where chemicals, drugs or other material or biological matter are tested and analysed, requiring water, direct ventilation and specialised piped utilities). Laboratories, and the need for consumables, can be largely replaced by a computer and an internet connection. This also makes bioinformatics particularly feasible in the African context, where it is easier to come by computers and connectivity than funding for sophisticated infrastructure. The challenge of greater internet connectivity is important and to some extent is already being addressed by groups working to build internet infrastructure, like UbuntuNet. Stand-alone devices are also in the offing, like the eBokit, a Mac-based device that has everything a bioinformatician might need, including databases and tools for training and analysis.

Associate Professor Mulder says, “It is challenging to do biomedical research today without bioinformatics because so many researchers are generating big data. It is less common now to work on one gene at a time. The new trend is to work on thousands of genes at one time, so you have to have bioinformatics to manage the data. In order to answer genetics questions, it is becoming more important to generate big-data sets.”

According to Associate Professor Mulder, the biggest challenge in big data for the biomedical field is the development of next-generation sequencing (NGS)



technologies that enable massively parallel sequencing. It is much cheaper than traditional sequencing but it generates millions of short reads, which demand new analysis and storage challenges. With the falling cost of NGS, researchers have the option, for instance, of whole-genome sequencing rather than targeted sequencing. This means one can move towards hypothesis-generating rather than hypothesis-testing science, which has the potential to lead to novel discoveries. New algorithms are being developed to manage this data and for archiving, which makes this sort of research more feasible.

THE FUNDAMENTAL NATURE OF THINGS

Another discipline where big data is playing an increasingly important role is physics. Dr Andrew Hamilton, a lecturer in the Department of Physics at UCT and researcher in high-energy particle physics at the European Organisation for Nuclear Research (CERN), is engaged in one of the world’s most exciting projects – the Large Hadron Collider (LHC). The 27-kilometre LHC is the world’s largest particle accelerator, and it is one of the most important tools available to high-energy physicists in their goal of investigating the fundamental particles that make up the universe. The UCT-CERN Research Centre is part of two of the experiments running at the LHC: the ATLAS experiment, which explores the fundamental particles

of the standard model of particle physics, and ALICE, which is aimed at understanding the quark-gluon plasma. The nature of its work means that the LHC has been grappling with the problem of big data sets for 20 years. Many of the particles it investigates (like the famous Higgs boson) need to be created by colliding protons at speeds approaching the speed of light. Because the particles are very rare, tens of millions of collision events are produced per second, which need to be captured and read by a detector. If researchers were to read out every single event, they would need around 40 terabytes per second, which is way beyond the confines of current technology; using high speed filtering, called triggers, researchers can get this down to hundreds of megabytes per second.

While this may already sound like a tall order, it is only part of the big data challenge faced by the LHC. In order to define expectations (so that they know what they expect to see), the entire detector is digitally simulated in excruciatingly fine detail, and a computer algorithm is written to simulate the billions of events that might produce a rare particle like the Higgs boson. All of this data then needs to be stored.

Factor into this the collaborations involved (just one of the LHC’s seven experiments has 1 000 members belonging to 116 institutions, based in 33 countries), and the scale of the challenge is evident.



The construction of MeerKAT antenna. Image courtesy of SKA South Africa.

The LHC has created a solution to this very big data problem – the Worldwide LHC Computing Grid (WLCG), a form of cloud computing that stores and processes information. It currently has 150 centres in 40 countries and operates through a tiered structure that varies in the funding model, the storage space and the number of cores required.

Creating the infrastructure required to play in this big data league sounds expensive; however, money is not the only challenge for UCT, according to Dr Hamilton. “We need the people trained in high-performance computing to operate and administer these facilities,” he says. This is where the WLCG comes in. “It is one of the largest research computing networks in the world.”

“If UCT wants to be a leader in big data research, we need to demonstrate that we can operate a big data centre on the global scale. The WLCG gives us the opportunity to contribute to one of the largest research computing networks on the planet and learn from their expertise at the same time.”

BIG DATA FROM THE SKY

Another project that is placing UCT at the centre of international big data research is the MeerKAT Large Surveys and the Square Kilometre Array (SKA).

Africa’s biggest science project, the MeerKAT radio telescope, is a precursor to the SKA telescope. Four key science programmes on MeerKAT are led by UCT astronomers. These research programmes will gather up to one petabyte (1 000 terabytes) of data per year, so advanced tools will have to be developed to process, analyse and store this amount of data. This will be done in collaboration with the SKA South Africa project office. SKA South Africa, UCT and UWC have attracted an eminent role player in this field from Canada, Professor Russ Taylor, who joined the university early in 2014 and will co-ordinate a big data vision for radio astronomy in South Africa. “The global SKA big science data world is coming to South Africa this decade,” says Taylor, adding that it is probably one of the two projects in the world driving a big data revolution in astronomy.

BIG DATA CHAMPION FOR UCT

The UCT Astronomy Department and the University of the Western Cape (UWC) Department of Physics earlier this year welcomed the appointment of Professor Russ Taylor to the joint UCT/UWC Square Kilometre Array (SKA) Research Chair. Professor Taylor will play a key role in building big-data research capacities and expertise in the region and the continent.

Professor Taylor, coming from the Department of Physics and Astronomy at the University of Calgary, has a wealth of experience and expertise in radio astronomy, in particular wide-field polarisation, cosmic magnetism and big data, and has played a prominent role in the SKA project since its inception. He was the founding international SKA project scientist and co-authored the first SKA science case. He represented Canada as one of the national members on the SKA Organisation Board. Previously he served as the founding Executive Secretary of the International SKA Steering Committee, the predecessor to the International SKA Science and Engineering Committee.

THERE IS LIKELY A LIMITED WINDOW OF OPPORTUNITY TO ESTABLISH NATIONAL LEADERSHIP IN BIG DATA AND A GLOBAL PRESENCE IN THIS EMERGING FIELD.



Professor Taylor’s research covers the cosmic battle between the forces of magnetism and gravity, which is probably responsible for slowing the pace at which the universe uses up its gravitational energy, allowing enough time for life to arise. “My research plan is to use MeerKAT and KAT-7 to measure the polarisation of radio waves and to trace the properties of magnetic fields in galaxies and intergalactic space,” says Taylor. “This will give scientists a better understanding of the evolution of cosmic magnetism.”

Professor Taylor has also served as the Canadian ALMA Software Manager for the Canadian component of the international software development for astronomical use of the Atacama Large Millimetre Array. He was the Canadian co-



principal investigator on an international partnership to launch a radio telescope for Very Long Baseline Interferometry (VLBI) imaging between Earth and space: the VLBI Space Observatory Programme (VSOP) space mission. As part of the mission, he directed one of three international centres for the processing of the VSOP mission data. He is also principal investigator of the International Galactic Plane Survey, a consortium of more than 60 Canadian and international scientists formed to carry out a co-ordinated data-intensive project of high-resolution imaging of the interstellar medium over the disc of our galaxy. In this capacity, he has also served as the chair of both the management committee for the Canadian component of the project (the Canadian Galactic Plane Survey) and the international project steering committee.

Taylor is chair of an international consortium of 31 scientists from Australia, Canada, the USA, Europe and India that carries out a large-scale spectro-polarimetric all-sky survey project with the Arecibo radio telescope. This project has been granted 2 000 hours of observing time over four years and foreshadows the data volumes that will be generated by MeerKAT.

MEERKAT IS EXPECTED TO GATHER UP TO ONE PETABYTE (1 000 TERABYTES) OF DATA PER YEAR, SO ADVANCED TOOLS WILL HAVE TO BE DEVELOPED TO PROCESS, ANALYSE AND STORE THIS AMOUNT OF DATA.



There is nothing small about the SKA project. It is the biggest science project ever carried out on African soil. Each MeerKAT antenna, which will be incorporated into the mid-frequency component of SKA Phase 1 when that instrument is constructed, stands 19.5 metres tall and weighs 42 tons.

When completed, the SKA will be the world's largest radio telescope, located in Africa and Australia, but shared by astronomers around the globe. Until then, MeerKAT will be the most sensitive and powerful L-Band radio interferometer in the world. In addition to operating as a stand-alone, world-leading imaging radio telescope, MeerKAT will participate in global VLBI (very long baseline interferometry) operations with all major VLBI networks around the world operating at the MeerKAT frequencies, adding considerably to the sensitivity of the global VLBI networks.

The complete MeerKAT array will have 64 receptors – antennas with receivers, digitisers and other electronics. Connected by 170 kilometres of underground fibre-optic cable, the receptors will operate as a single, highly sensitive astronomical instrument, controlled and monitored remotely from the MeerKAT control room in Cape Town.

When fully operational, the MeerKAT will generate enough data from the antennas to fill about four-and-a-half million standard 4.7-gigabyte DVDs in a day.

Professor Taylor has served on the board of the directors of the international SKA Organisation

(representing Canada), and has experience as an observer and observing proposal referee for the US National Radio Astronomy Observatory, Very Large Array. He is the international project leader for the design of the global data-delivery system for the SKA project. This, together with his expertise as a big data specialist, means that he is well-positioned to guide UCT toward realising its SKA big data vision. He was the founding international SKA project scientist and co-authored the first SKA science case.

A RESEARCH REVOLUTION

In addition to his own research with SKA, Professor Taylor will also be working to put South Africa – and Africa – on the map in terms of data capacity. He says that there is likely to be a limited window of opportunity to establish national leadership in data-intensive research and a global presence in this emerging field.

While big data is, by its very nature, a massive challenge to the university, it is also a driver of the transformation of science and, by extension, a driver of global change, and UCT is already part of the revolution.

"We are not trying to break into a field where we are absent – we are already there," says Professor Visser. "If we grasp the opportunity to take leadership in this area we can really make a difference in the country and to science around the world: solving Africa's issues, but also making Africa part of the global solutions."

E-RESEARCH CENTRE BREAKS NEW GROUND

UCT has taken the lead on the African continent in establishing an eResearch Centre to ensure that the university can continue to operate as a top research institution in the age of big data.

According to Professor Danie Visser, Deputy Vice-Chancellor with responsibility for research at UCT, research today is fast becoming inconceivable without adequate eResearch infrastructure – information and communication technology (ICT) assets, facilities, skills and services – to support it. "Universities without an equipping strategy may continue to perform, but only for a time. As research changes, all support areas serving research must keep up to stay relevant. Without change, there is a risk that service areas will provide yesterday's solutions," he says.

UCT approved the establishment of an eResearch Centre in March 2014. The centre will support and enhance the university's research capabilities. A large component of the eResearch strategy at UCT revolves around ICT. Researchers in many fields rely increasingly on ICT as a component of their research, with requirements ranging from support for data management strategies and data-centric architectures to access to specific tools and software for data analysis. Technology is also accelerating the pace and

scale of research, with large-scale data requiring a more structured approach to data management and storage. New and more powerful instruments are required: digital recognition of text, speech and imagery, and facilitating crowdsourcing and citizen science.

"We are seeing three major drivers for the ICT change programme, influenced by both global challenges and our own local challenges," says Sakkie Janse van Rensburg, executive director of Information and Communication Technology Services (ICTS) at UCT. "We want to deliver more, which will require a new organisational structure and roles; we want to deliver the right thing, which means more focus on governance; and we want to deliver it in the right way, which means we need to improve our internal processes to help researchers conduct research faster and more cost-effectively. The concept of eResearch explores the question of how we can, with the latest tools, technologies and approaches, strengthen that research workflow or pipeline of 'conceive – design – explore – analyse – collaborate – publish – expose'."

For several years now, ICTS has been delivering eResearch support through the establishment of an HPC (high performance computing) cluster that supports advanced research computing. Janse van Rensburg says that, in 2013 alone, more than 155 researchers across campus were supported, and they submitted more than 270 000 jobs requiring HPC facilities. The computing time for that year added up to more than two million hours.