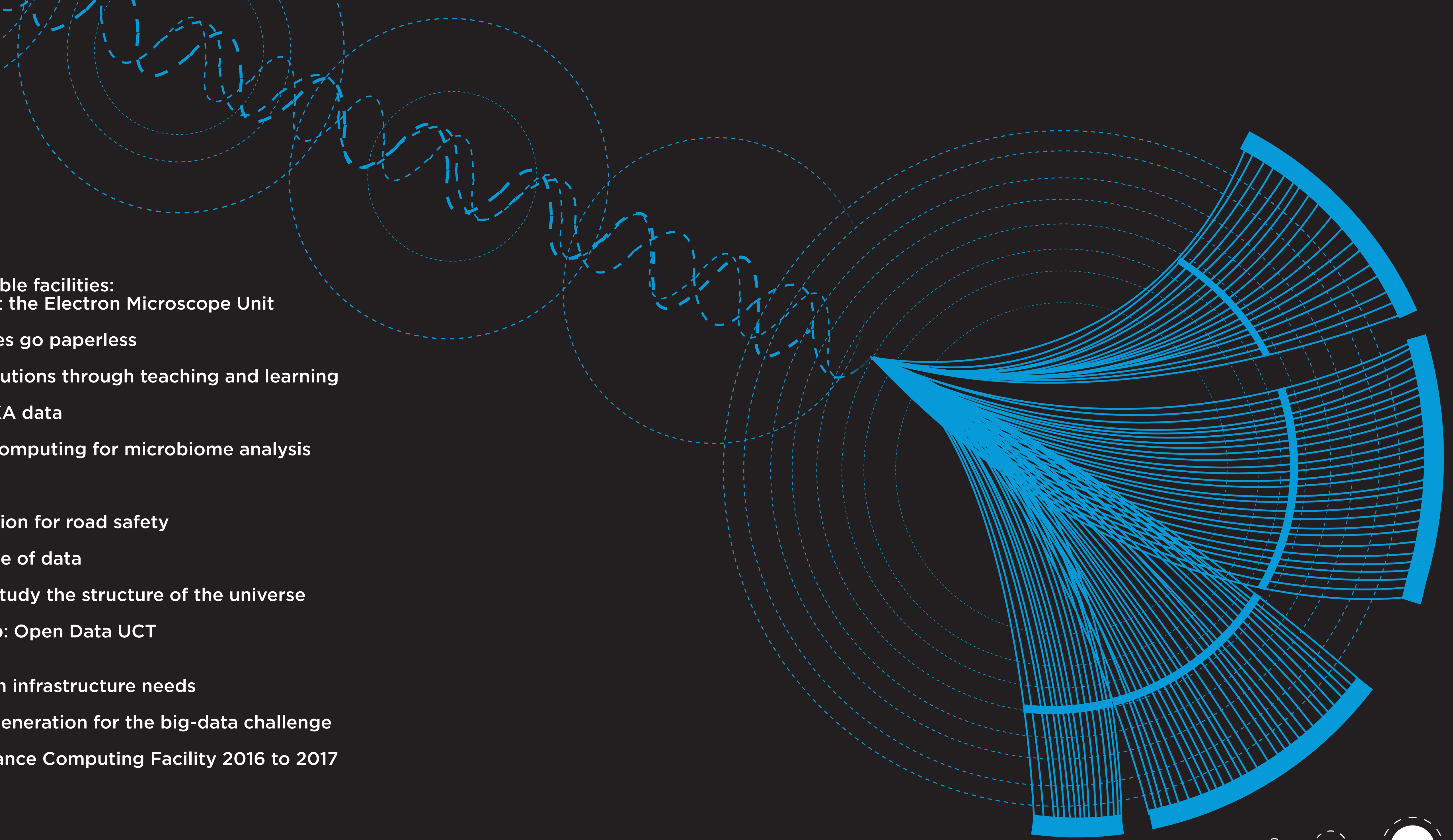
UCT
eResearch
ACCELERATING RESEARCH

eResearch Report

2016-2017

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Contents



New era, new opportunities for collaboration

Prof Mamokgethi Phakeng

Deputy vice-chancellor, research and internationalisation

As a globally competitive, research-intensive university, in a world of big data, we need to ensure our researchers have access to cutting-edge facilities, infrastructure and support. The challenge we face is to offer this access in a resource-restricted environment. Collaboration for shared resources is the solution. Bringing researchers together – not only from UCT but nationally, to get the greatest benefit from shared resources – is something we at UCT are committed to driving to ensure the greater good of our national research endeavour.

Dr Dale Peters

Director, UCT eResearch

The digital era brings new challenges and opportunities for scientific research, offering new ways of global interaction within disciplines, as well as local collaborations between disciplines. Exponential increases in bandwidth, computational resources and storage have improved channels for sharing findings among researchers, and between researchers and society, thanks to new forms of dissemination and access to information. This publication highlights many challenges presented to eResearch in the past year, and are presented here in acknowledgement of the innovative work of researchers in exploring those opportunities to extend their research practice.

Support at every step of your research lifecycle

UCT eResearch, as a partnership between Information and Communication Technology Services (ICTS), UCT Libraries and the Research Office, works to ensure that the data needs of our researchers are supported at every step of their research lifecycle.

Sakkie Janse van Rensburg

Executive director, Information and Communication Technology Services

This year, UCT eResearch has continued to grow from strength to strength, partaking in, and even leading, national initiatives. In partnership with ICTS, the Inter-university Institute for Data-Intensive Astronomy, and the African Research Cloud, we have grappled with big-data infrastructure challenges. Acquiring, accessing, analysing and comprehending data on the scale it's being generated now requires new tools, innovative software and powerful hardware and storage. Working together – across the university and with our partners across the country – we are rising to the challenge.

Gwenda Thomas

Executive director, UCT Libraries

More and more funding bodies, journal publishers and research institutions are requiring that research data be made openly available. At the same time, researchers are recognising the potential for data sharing to raise their profiles and increase citations. Therefore, university libraries have increasingly become involved in providing support for the management of research data, which includes ensuring its accessibility.

This year, UCT launched its institutional data repository, ZivaHub – the first of its kind in the country.

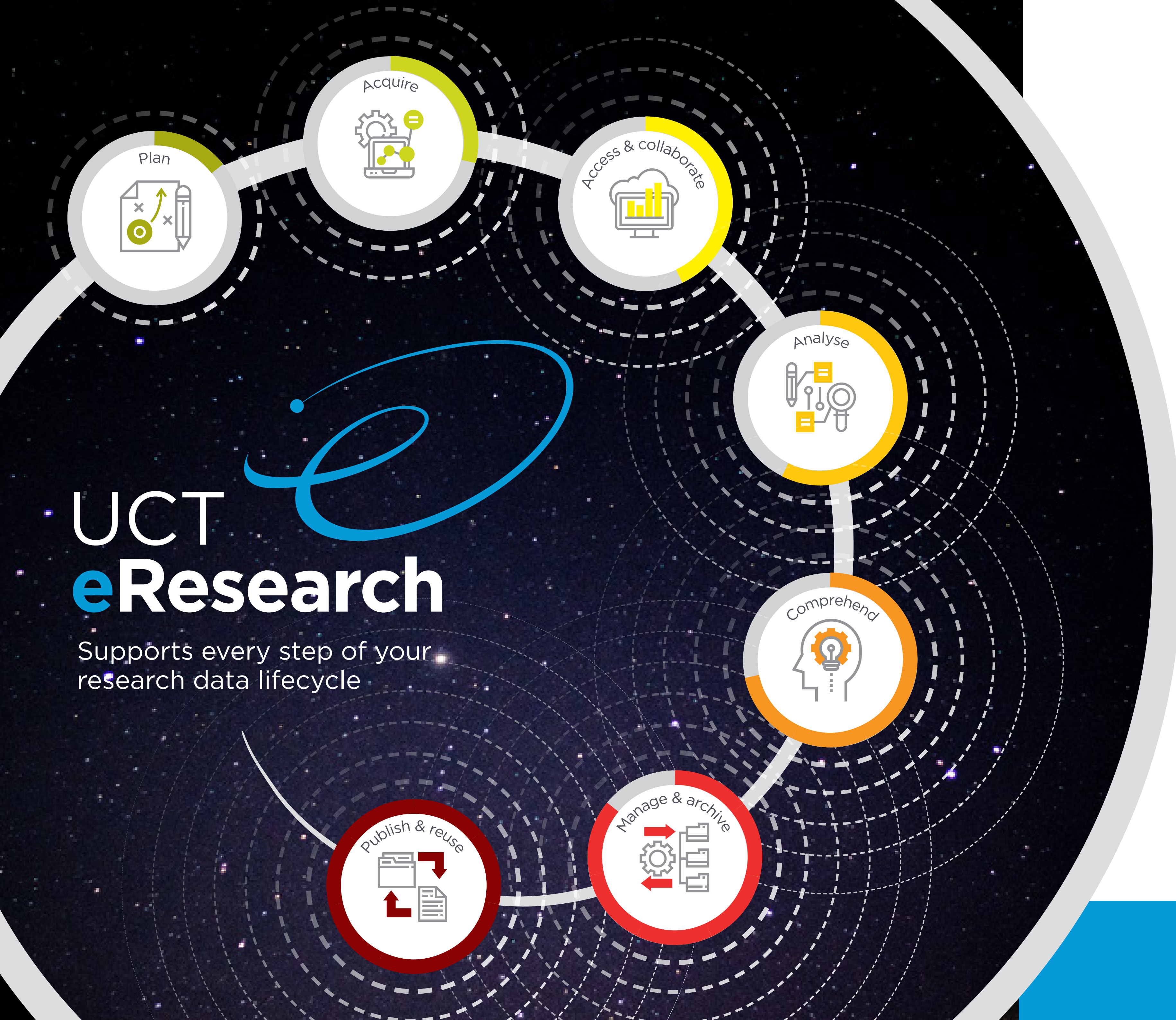
UCT Libraries, as part of eResearch, is driving this tool, and we will continue to actively work with researchers to sustain its usability.

Dr Marilet Sienaert

Executive director, Research Office

The Research Office aims to support researchers in all aspects of their work, and research data has added a new – and increasingly important – facet to this mandate. As part of UCT eResearch, our goal is to inform UCT's research community about the support services available through eResearch to help manage and prepare for the challenges around acquiring, accessing, analysing, comprehending and sharing research data. There are resources available to you at every step.

This is the new frontier in research discovery and to stay ahead, UCT and its researchers need to tap into these services for the benefit of their own research excellence and the development of their students and the wider economy of knowledge creation.



UCT eResearch

Supports every step of your research data lifecycle

Plan If you know at the planning stage of your research what your technical needs are going to be, then you can be sure to be prepared for them when they arise later on. We work with researchers at the planning and grant-writing stage to help allocate budget for the data challenges ahead.

Acquire As a researcher, your next step is to get the data. This often requires the use of complex tools and software, which collect raw data that then need to be stored and managed. We assist with setting up the necessary software and databases to ensure that data are captured and stored effectively.

Access & collaborate Research projects today often include a team of researchers scattered geographically who need access to the same data at the same time. We assist with creating shared data stores, dataset transfers, file sharing and other facilities or software required for effective collaboration.

Analyse You have the data: now you need to find out what it means. We offer and support a range of options for data analysis, including high-performance computing facilities, cloud-based software and virtual machines or laboratories.

Comprehend Our brains process complex information better when it's visual. We offer visualisation facilities to assist researchers to visualise their research data sets in an immersive space.

Manage & archive Data needs to be findable, accessible, interopable and reusable (FAIR). We support the management and archiving of data to ensure your research data meets these FAIR guiding principles.

Publish & reuse For maximum impact, both research outputs and research data should be publicly available on an open-access platform. We support open-access publishing of research outputs, including publications and data.

Shared and sustainable facilities: piloting Calpendo at the Electron Microscope Unit

UCT is home to myriad facilities, instruments, software packages and services, ranging from electron microscopes to high-performance computing facilities that are used by a range of researchers across disciplines and institutions. The trouble is that these resources are expensive to buy and to maintain.
To ensure sustainability, not only do the various facilities need to be used by as wide a community of researchers as possible, but an effective billing system needs to be in place for the user community.

This is a model the Electron Microscope Unit (EMU) at UCT relies on for its five electron microscopes and its protein purification lab.

The Electron Microscope Unit was one of the research groups to pilot new software, Calpendo, for managing the use of UCT's facilities and instruments, such as this x-ray diffractometer.

Photo: Stephen Williams

"The money we charge for the use of the microscopes goes to the maintenance of the resources," says Miranda Waldron, principal scientific officer at the EMU. "That is how we have always survived, so the microscopes need to be kept busy."

The EMU microscopes are used by researchers from health sciences, engineering, molecular and cell biology, biological sciences and others. Beyond just the UCT community, the unit receives requests from researchers across South Africa and elsewhere on the continent.

Whether they simply courier samples through to the lab, or come in person to use the microscopes, researchers must book and pay for time on the facilities.

The EMU was thus a natural candidate for a pilot conducted on behalf of the University Research Committee's (URC) Imaging and Microscopy Working Group, to find the most efficient way to manage the administrative burden and costs of research facilities and instruments around the university.

Working with the URC and EMU, UCT eResearch identified Calpendo, a cloud-based system for managing facilities. Calpendo was first used by the University of Oxford in 2009, and today is used by a range of universities across the globe.

The EMU was the first unit to pilot the new software at UCT. The EMU's old system was hosted on a server at ICTS, and had been battling under the weight of the unit's requirements. For Waldron, the advantages of cloud-based software, with a support team on standby, have been excellent.

The Electron Microscope Unit's facilities are used by researchers from across the university. These researchers are studying samples using a scanning electron microscope.

Photo: Stephen Williams

"With the old system you could only access it when you were on campus, with a UCT computer. This obviously had a number of limitations, one of which was that one of us would have to actually make the booking for external users," says Waldron. "Calpendo is really convenient. You simply access it through a website, so anyone can use it from anywhere in the world."

"One of the biggest benefits of cloud-based computing is the strong support provided, enabling effective collaboration between local and remote teams," says Dr Dale Peters, UCT eResearch director.

Researchers are not simply booking and paying for the use of the microscopes at the EMU. They are also supported by Waldron and her colleagues, who offer a range of services, from preparing the samples and running the microscopes to teaching students how to operate the equipment on their own.

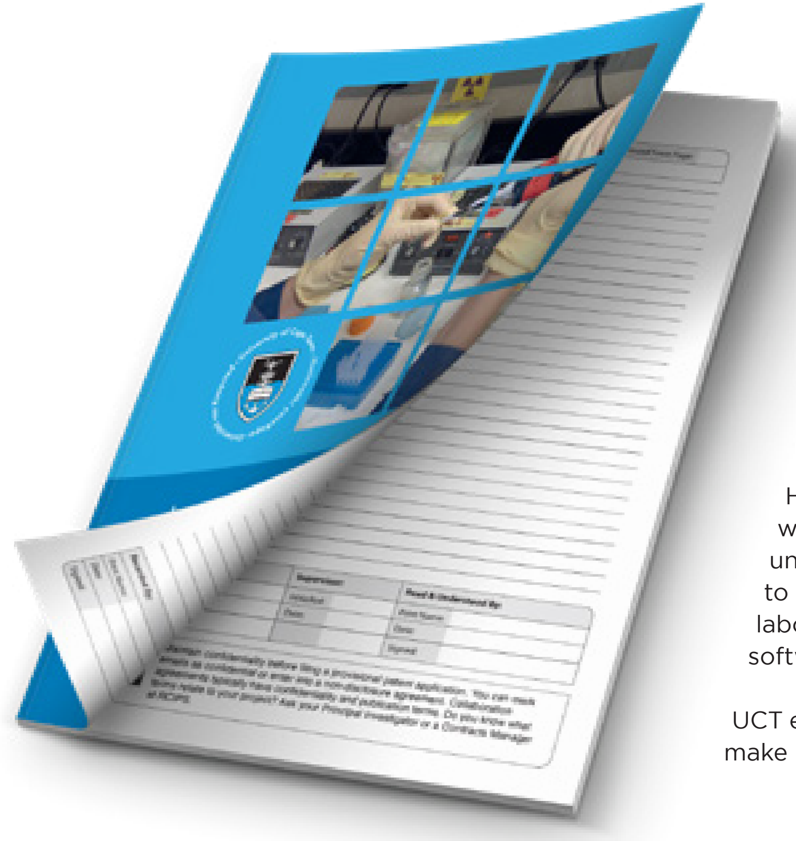
Calpendo meets their needs and requirements, but also goes beyond what previous booking systems have offered. Waldron says she can see the value of this software for scaling up in the future.

"We have a new machine funded by the National Research Foundation, and they require a greater level of reporting than we are used to. We need to provide information such as when a student user started their degree and when they intend to finish, or what their ID number is," she says. "I think Calpendo is going to be great for that level of reporting."

The URC has identified a need for a single system to allow all researchers access to the range of facilities and equipment available at UCT. This means that, ideally, the system the EMU implements must be able to scale up, to cover the needs of the entire university and beyond.

"In times of austerity, which all South African universities are faced with now, we need to work together – not only at a university level, but nationally – to make sure we get the best use out of the resources we have," says Peters. "Hopefully, cloud-based software such as Calpendo will offer the solution to managing the administrative burden of these nationally shared resources."●

Research laboratories go paperless



When Professor Stefan Barth – Department of Science and Technology/ National Research Foundation South African Research Chair in Cancer Biotechnology – moved his laboratory to UCT from Germany two years ago, he envisioned a place where the knowledge shared between himself and his laboratory members would be contained in a secure, persistent digital repository.

He also required a solution for an electronic witnessing system to verify and record – in an unalterable way – the laboratory's activities related to intellectual property, as well as a way to allow laboratory members to share access to the costly software programs they use for interpreting data.

UCT eResearch was able to support Barth and help make his vision a reality.

A repertoire of methods

Barth's research centres around the development of recombinant proteins – proteins made from a combination of genetic code from different organisms – that can be used in the diagnosis and treatment of disease.The techniques, protocols and novel products created in his laboratory are potentially patentable; patenting procedures benefit from evidence in the form of meticulous record-keeping and verification of the activities taking place in the lab.

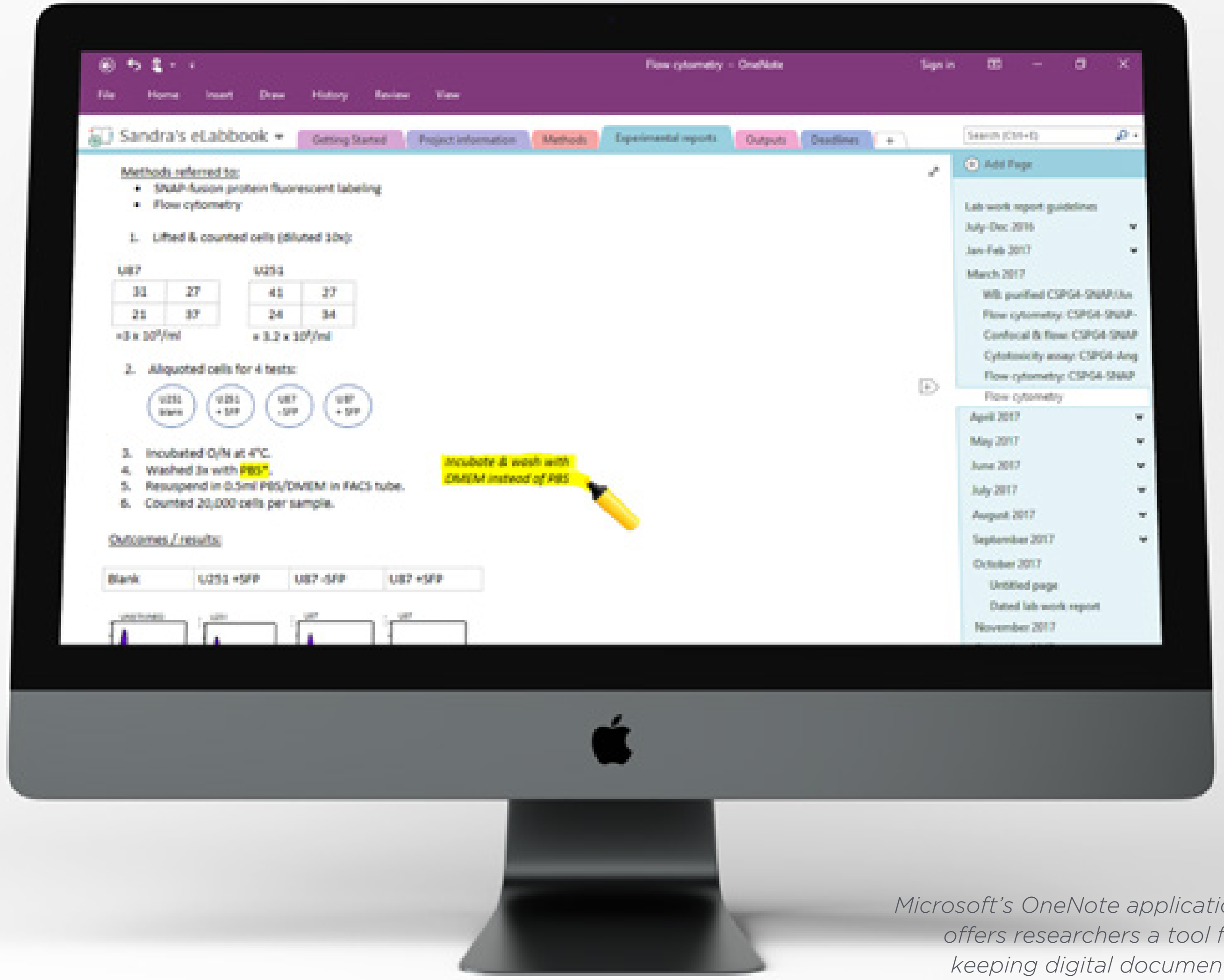
"The documentation of all the experiments related to this repertoire of methods becomes very, very important," explains Barth. "It is important not only for following up at a later date on what's been done, but for controlling the quality of the work and the details of certain experiments."

Traditionally, scientists, engineers and technicians have used paper-based lab notebooks to keep track of the research, experiments and procedures they perform. These documents – contributed to on a daily basis – may serve as legal evidence in court, or during patent prosecution and intellectual property litigation.

But it's easy to imagine how such a paper-based system could be limiting. "It's a matter of keeping track of the many lab notebooks written by each student, and storing them safely," says Barth. "But once you have a number of students working on different types of projects, finding the data for a certain experiment in this paperwork becomes a challenge."

These obstacles presented Barth with an opportunity for a solution beyond the capabilities of the current system. Barth's goal was to take the record-keeping function of lab notebooks, make it digital, and integrate it with systems for storing and backing up the notebook data, and for indexing and searching, collaborating and witnessing; as well as for data storage, interpretation and visualisation – thus providing so much more than paper notebooks do.

At the time that Barth was grappling with these ideas, he started discussions with the team at UCT eResearch. He wanted to know: which was UCT's preferred option for electronic lab notebooks? "And we went into a panic," explains Ashley Rustin, a senior technical specialist at UCT eResearch, "because UCT was still using paper-based lab books."



Methods referred to:

- SNAP-fusion protein fluorescent labeling
- Flow cytometry

1. Lifted & counted cells (diluted 10x):

U87

31	27
21	37

≈3 x 10^5/ml

U251

41	27
24	34

≈ 3.2 x 10^5/ml

2. Aliquoted cells for 4 tests:

U251 blank, U251 + SFP, U87 - SFP, U87 + SFP

3. Incubated O/N at 4°C.
4. Washed 3x with PBS*.
5. Resuspend in 0.5ml PBS/DMEM in FACS tube.
6. Counted 20,000 cells per sample.

Incubate & wash with DMEM instead of PBS

Outcomes / results:

Blank	U251 +SFP	U87 -SFP	U87 +SFP

Microsoft's OneNote application offers researchers a tool for keeping digital documents.

Software and servers

In response to the request, eResearch conducted a survey to assess the software packages currently in use for e-lab notebooks. They found that researchers' preferred option was OneNote, a Microsoft application already freely available to the UCT community under the Microsoft suite site licence.

OneNote is versatile: it allows researchers to take notes, both handwritten and typed. They can do drawings, capture imagery, record audio and attach raw data files. It also has powerful collaborative tools.

As it stands, the Barth laboratory has access to a server based at UCT's Information and Communication Technology Services (ICTS) that runs OneNote, as well as other specialised licensed software. They have sufficient storage space that the complete work and documentation for each student can be kept there. They have also set up templates in OneNote for their e-lab notebooks, which are shared via the server.

They are still setting up a method for witnessing and for archiving permanent files, in the form of PDFs, in a separate storage space. "This hasn't been solved at this stage, but it's just a matter of time," Barth says.

Realising the concept of a paperless lab may seem a long way off, but the technology is available now, and UCT eResearch is helping to pave the way for researchers who are going digital.

Thinking of going paperless?

To share information more widely about the e-lab notebook solution that Barth and UCT eResearch developed, Research Contracts and Innovation joined with them to run a seminar in two parts: an overview of OneNote and its features, and feedback from the Barth research group on its implementation in their laboratory.

Anyone in the UCT community who is interested can access the presentation slides and video recordings of the seminar via its Vula site, 'Electronic Lab Notebooks'.Vula is UCT's online collaboration and learning environment.

Finding big-data solutions through teaching and learning

As a partner in the African Research Cloud (ARC) and the Inter-University Institute for Data-Intensive Astronomy (IDIA), UCT is home to some powerful computing facilities. These facilities, which allow researchers to work with enormous sets of data, are helping us to prepare for the era of big-data science.

But when these facilities first arrived, someone needed to forge the way and figure out how best to harness them, share access to them and demonstrate their usefulness. Dr Bradley Frank, a senior researcher at IDIA, was among the team who led this charge. As project scientist for ARC's astronomy demonstrator project, and a member of the IDIA team working to develop and deploy tools to analyse astronomical data generated by the Square Kilometre Array (SKA) and its precursor, MeerKAT, Frank faced a range of challenges. One particular challenge centred on ensuring comprehensive access for astronomers to the telescopes' data and to the tools to process and analyse it.

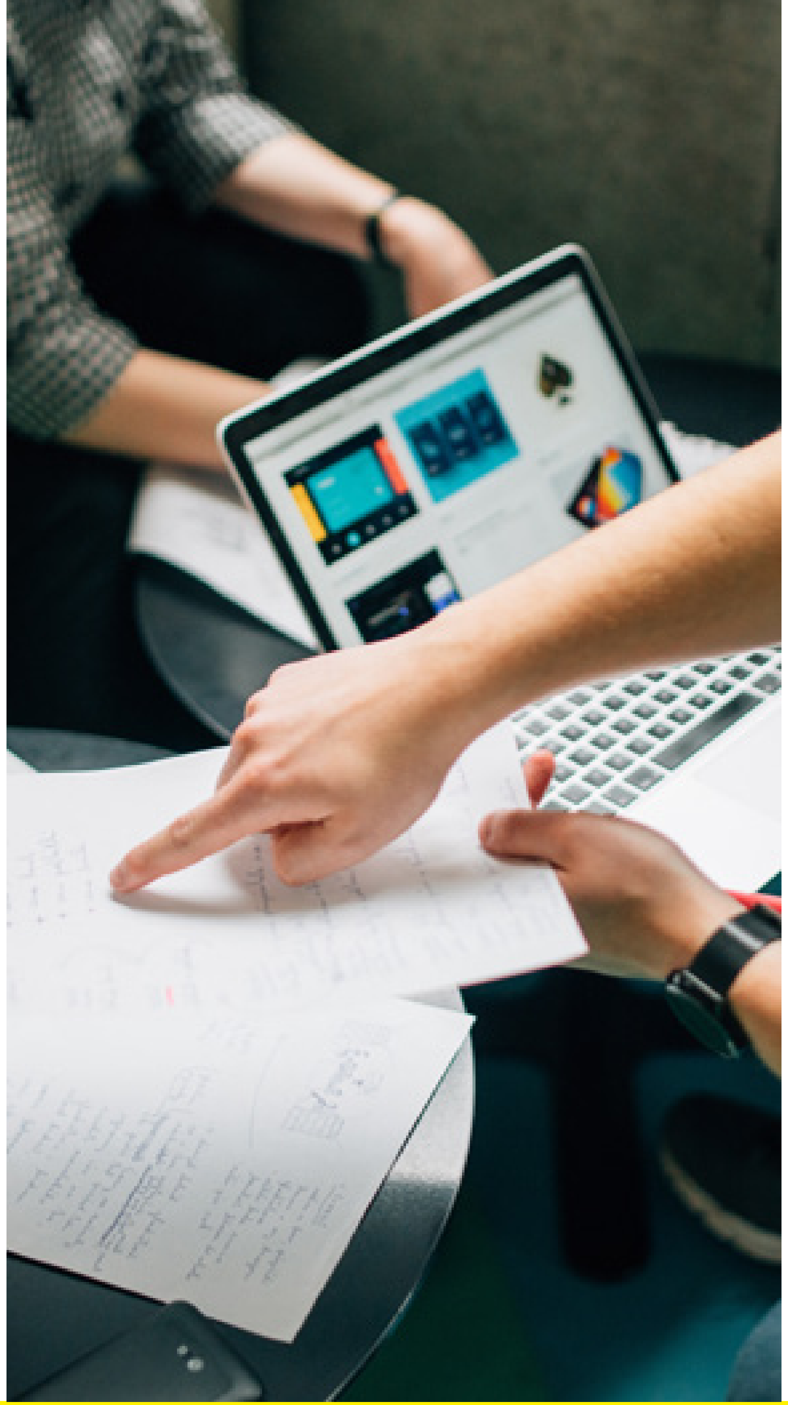
As he is also the SKA lecturer at UCT, Frank realised that he could try out any solution on a group of benign test subjects: his second-year students.

Overlapping challenges and solutions

Frank spotted an opportunity to use the ARC and IDIA facilities to enrich his teaching. He set out to find a way to provide access to the software tools and data required to study astronomical data for the 45 students in his astronomy techniques class.

Although there is a variety of computer facilities available at UCT for student training, none of them provided the ideal hardware and software capabilities for in-depth computing for a large number of participants. Frank needed another solution.

In one fell swoop, Frank planned to use the data-intensive computing facilities of the ARC and IDIA to both teach his students and demonstrate teaching and learning as a use-case for the facilities, as well as interrogate the system to see where processes might break or fall short.



> "This is an enormously powerful tool, with the potential to develop skills in science, engineering and technology in South Africa."

Teaching and learning on the cloud

The solution was facilitated by eResearch's technical specialist, Timothy Carr. Carr was able to set up – and subsequently manage – a powerful, cloud-based virtual machine, or hub, on the ARC and IDIA servers. Frank's 45 students could then access the data, software, instructions and computing power they needed to complete their analyses via a web-based interface and on any device with a web browser.

It worked so well that Frank used the same system with his students the following year. He even worked with Carr to solve problems on the fly: for example, when the students ran out of RAM on the virtual machine Carr had set up.

This sort of teaching and learning intervention is one of the first examples of its kind in the country. Such interventions are transferrable, Frank believes: "They could be used by researchers in other fields at UCT, but also across South Africa, to provide access to skills development in maths, physics, statistics and computing for a huge audience: anyone with a mobile device."

The South African MeerKAT radio telescope being built near the small town of Carnarvon is a precursor project of the Square Kilometre Array (SKA). In preparation for the SKA, teams of engineers, software developers and scientists – including UCT eResearch team members – are working with the MeerKAT data to allow them to prepare the necessary hardware and software to ensure that data from the SKA can be put to good use as soon as possible after the array comes online in 2020.

Photo: South African Radio Astronomy Observatory

Preparing for the SKA data

The world is wholly underprepared for the big-data challenges presented by the Square Kilometre Array (SKA) – the largest radio telescope ever built. This is because we have never seen data volumes on this scale before. By the time the SKA comes online in 2020, the scientific community needs to be ready with the necessary hardware and software so that the data can be put to good use immediately. UCT eResearch has been helping with two specific challenges: delivering the data sets to researchers around the world, and working to enable visualisation of the data.

The SKA has established precursor projects – pathfinders – to prototype the tools required to transport, process, analyse and visualise their data. The South African MeerKAT project is one of these pathfinder radio telescopes.

Two members of UCT eResearch – data scientist David Aikema, and Adrianna Pińska, a scientific software developer – have been seconded to the Inter-University Institute for Data-Intensive Astronomy (IDIA) to help with these challenges.

> "eResearch offers centralised research services to the whole community, as well as embedded services within specialised research teams," explains UCT eResearch director, Dr Dale Peters. "Working integrally with the IDIA team, David and Adrianna are able to respond directly to the big-data challenges presented by the SKA."

Delivering the data

With colleagues at IDIA, ASTRON in the Netherlands, the Institute of Astrophysics of Andalusia in Spain, and the Canadian Astronomy Data Centre, Aikema has been working on solutions to deliver the massive data sets produced by the SKA to astronomers in South Africa and around the world.

The challenge, explains Aikema – which will start with the full MeerKAT array and expand exponentially with the SKA – will be in ensuring that the data are archived and stored in a way that is accessible to the various (geographically diverse) research projects related to the SKA.

Key to the data-delivery architecture are the SKA regional centres. IDIA, along with ASTRON in the Netherlands, are pathfinder SKA regional centres; but more such centres are in the pipeline, scattered around the world. SKA data sets will be sent to these regional centres.

"The end goal of the MeerKAT data-delivery architecture – which should then give us a blueprint for handling the SKA data when it comes – is that researchers will be able to access and work with big data through services and systems provided by the SKA regional centres," explains Aikema.

The Square Kilometre Array (SKA) project is one of the largest scientific endeavours in history. The scale of the SKA represents a huge leap forward in both engineering, and research and development. Once fully functioning, the SKA will be made up of thousands of dishes and up to a million antennae that will allow astronomers to monitor the sky in unprecedented detail.

Visualising the data

Visualisation is the human brain's best way of understanding large volumes of data. It is a tool that allows us to represent large and incomprehensible data sets in a way that allows us to see a pattern and comprehend the information within the data.

Because of the size and complexity of astronomical data, an effective visualisation tool is key. In her work for IDIA, Pińska is collaborating with a team at the Academia Sinica Institute of Astronomy and Astrophysics in Taiwan, and the National Radio Astronomy Observatory (NRAO) in the United States, to upscale the Cube Analysis and Rendering Tool for Astronomy (CARTA) – a platform for viewing astronomical data – to enable it to handle the data requirements of MeerKAT and the SKA. CARTA was originally designed to visualise multi-dimensional data sets that vary in size and volume.

"CARTA has a client-server architecture, so users can connect through a web browser," says Pińska. "The large data files then sit on a server, where they are processed, and the rendered image data is sent back to the viewer."

The challenge is that CARTA was designed for much smaller data sets than those expected from MeerKAT and SKA. Pińska's job is therefore to optimise CARTA so it can open really large data files.

"There are two major challenges here. One is speed: if the viewer needs to make calculations over the data, CARTA may do so quickly on a small image, but take much longer on a larger image," explains Pińska. "The other issue is memory: you need enough memory to open the image, and some calculations may have more memory requirements proportional to the size of the image."

To solve these two issues, Pińska is working on applying new algorithms to CARTA for efficient viewing of massive data-intensive images, with the option to upscale as those images increase in size.

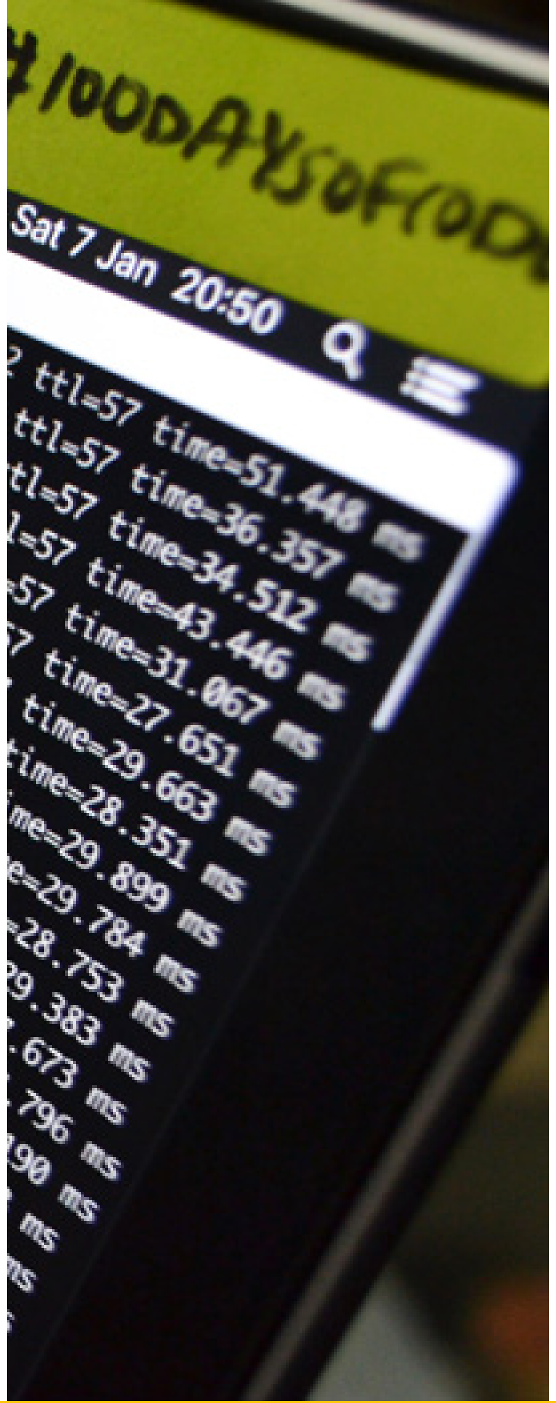
Fortunately – with international teams grappling with these obstacles now and working out solutions according to the scale of the expected data volumes – by the time SKA data sets are ready for science in the 2020s, we will be ready to meet the challenge. ●

High-performance computing for microbiome analysis

As part of the Human Heredity and Health in Africa (H3Africa) project, researchers at UCT are investigating the microorganisms that live in the nasal passageways and throats of children. They are interested in these microbial communities because they have an influence on the likelihood of a child developing pneumonia and wheezing illness, a disease that is a precursor to asthma.

To figure out which types of bacteria are present in the upper airways of children and estimate their abundance, researchers gather samples and use sequencing to analyse a specific region of the microorganisms' DNA. Such studies may involve hundreds of samples, millions of DNA sequences and hundreds of types of microorganisms. Large-scale projects of this kind require high through-put processing and intensive data management and task scheduling. They also require the help of a bioinformatician – someone who applies information technology to biological and medical research.

Gerrit Botha and Dr Katie Lennard – bioinformaticians based at the Computational Biology Division (CBIO) and members of the Pan African Bioinformatics Network for H3Africa (H3ABioNet) – have developed a streamlined process for analysing microbiome samples on the university's high-performance computing system.



```
Sat 7 Jan 20:50
2 ttl=57 time=51.448 ms
ttl=57 time=36.357 ms
tl=57 time=34.512 ms
l=57 time=43.446 ms
57 time=31.067 ms
7 time=27.651 ms
time=29.663 ms
ime=28.351 ms
me=29.899 ms
e=29.784 ms
=28.753 ms
9.383 ms
.673 ms
796 ms
90 ms
ms
ms
```

Such data-analysis pipelines, as they are known, take data inputs and guide them through a number of processing steps that have been linked together. The pipeline Botha and Lennard have developed takes sequencing data and its associated metadata and runs it through pre-processing steps, and alignment and classification algorithms.

"You can just imagine: there are a lot of steps involved, each of which uses different tools. You have to make sure that the output of one tool is compatible with the input of the next," says Botha. "Our main aim was to build a package that you can give to researchers and that they can use easily." The pipeline he co-developed has been made available to UCT researchers on the eResearch high-performance computing cluster, and is used in other fields, including oceanography and immunology.

In recognition of the need to update the pipeline and improve its efficiency, during October 2016 Botha and other developers tackled the challenge as part of a coding hackathon run by H3ABioNet. The hackathon brought together developers from H3ABioNet who worked in groups of three or four to create solutions for analysing different types of H3Africa data. One group of developers, including Botha, created a new pipeline for microbiome data, which – among other things – facilitates easier software updates and is more portable: that is, it can be used on other computing clusters and personal computers.

Hacka-what?

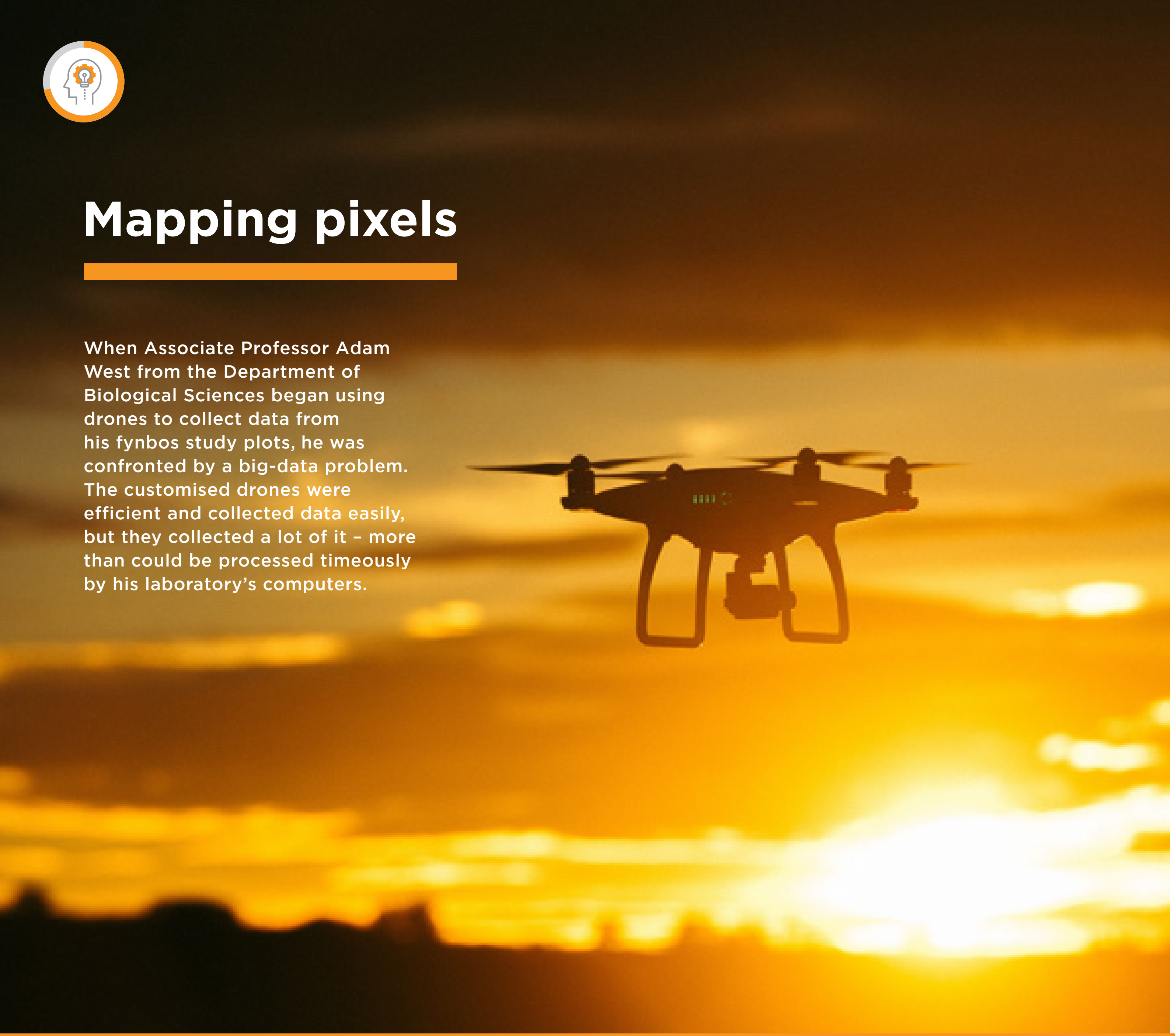
A hackathon is an event in which computer programmers get together to do intensive, collaborative computer programming. Their goal is to create usable software. Hackathons may last from a few hours to a few days. Other people involved in software development, such as graphic designers, interface designers and project managers – and sometimes subject-matter experts – may also attend hackathons, which tend to have a specific focus.

Since the hackathon concluded, Botha has been working to convert the pipeline to run using different containerisation software, called Singularity, and to test it on UCT's cluster. The focus of a recent workshop at the eResearch Africa 2017 conference, containerisation involves encapsulating one or more software applications in a container with its own operating environment. This helps to ensure the software runs reliably across computing environments; applications are easy to deploy and upgrade, and are easily shared. Botha's work on containerisation is highly innovative and a leading example of research computing practice.

Once Botha has completed this conversion and testing, the new pipeline will be made available to UCT researchers on the high-performance computing cluster. Other institutions and researchers will also be able to run this pipeline on their computing clusters and personal computers. ●

Mapping pixels

When Associate Professor Adam West from the Department of Biological Sciences began using drones to collect data from his fynbos study plots, he was confronted by a big-data problem. The customised drones were efficient and collected data easily, but they collected a lot of it – more than could be processed timeously by his laboratory's computers.



West reached out to the team at UCT eResearch for support. "Our big-data problem wasn't on the scale of a Square Kilometre Array big-data problem. But for physiologists who are used to small amounts of data, it was a computing challenge."

From leaf to globe

West calls himself an eco-physiologist. He's interested in ecosystems and the physiological processes that underpin them. In plants, these physiological processes include things such as photosynthesis and the absorption and transpiration of water. These processes happen at the level of an individual plant – or leaf – but when they are scaled up to landscapes, they have an effect on the way natural systems work.

West wants to make the connection between plant-level processes and regional- and global-scale observations specifically satellite imagery.

"But the remote sensing products we as natural scientists can get access to are relatively coarse in terms of their resolution: 30-by-30 metres is the highest resolution that we have," explains West. "But if you're working in fynbos, or any kind of biological system, a 30-by-30-metre pixel doesn't really help you to scale down to the level of biodiversity."

West's tool of choice for bridging the gap between individual plants and 30-by-30-metre plots: customised drones. The imagery and information that the drones gather by camera can be used to identify species of plants on a plot and their size, and to produce an index of the vegetation's health.

Ten years ago it was a much more manual process. "We'd swing cameras up over the plot, we'd run cables up to them, take the photo, swing them down, pull out the memory cards, stick them in a laptop. 'Did we get the image?' 'No, we didn't.' And swing them up again. It would take us weeks to get the imagery. It was a bit of a nightmare," says West.

The drones, programmed to fly in a grid pattern over the study area, have banished that nightmare, but they present a data challenge. One survey of a single plot can generate more than 2 000 images, each of which is a multi-megapixel file.

This isn't a photo; it's a 3D, digital map of a field site studied by Associate Professor Adam West at the Department of Biological Sciences. West used drones to capture thousands of images, which a specialised software package, running on eResearch's high-performance computers, stitched together to create this landscape.

The map of the field site, generated by eResearch, is navigable – you can pan around it and zoom into areas of interest – and shows individual plants and their structures.

That's a lot of pixels

"Very early on we identified a software package, Pix4D, that would help with the processing. Then, when we started getting our first big datasets, we realised that to process a single batch would take 24 to 48 hours, depending on the number of images," says West. "That's when we got eResearch involved."

West collaborated with Timothy Carr, a senior technical specialist at UCT eResearch. Carr was able to contact the developers of the Pix4D software and install a version on UCT's high-performance computers.

"Adam West's requirements presented a unique case," says Carr. "First, he required lots of compute power, so his traditional desktop took quite a long time to stitch all of the drone images together. In addition, West's desktop software version of Pix4D wasn't engineered to run on our high-performance computing centre.

"But we were able to configure the Pix4D software to make use of the extensive amount of processing power that we have available in our cluster."

Extending to machine learning

After processing the thousands of images from West's drones, the high-performance computing cluster would furnish him with a 3D map of his study site to manipulate on his own computer. The map is navigable – you can pan around it and zoom into areas of interest – and shows individual plants and their structures. Viewed from the side, you can measure the height of a plant canopy. Combine that with the area from above and you can calculate the volume – or biomass – of a shrub.

For the time being, someone has to do these calculations manually, but West has plans to extend his analysis into machine learning. This is another area where he foresees a need for eResearch's support, because machine learning is also a processing-intensive method.

When one thinks of big data, ecology might not be the first field that comes to mind. But big data is permeating all areas of research enquiry. And regardless of whether the big-data challenge is massive or small, there are opportunities across UCT to harness the support and services of eResearch. ●

Fostering collaboration for road safety

Dr Hetta Gouse, chief research officer in the Department of Mental Health, is using a driver simulator to better understand driver behaviour.

In 2016, South Africa saw its highest number of road deaths in 10 years. Just over 14 000 people died on South Africa's roads that year, according to the Automobile Association of South Africa (AA). In 2015, that number stood at just under 13 000 – also unacceptably high. To address this crisis, researchers and policymakers need to understand the behaviour of road users, including driver behaviour. Interdisciplinary research to understand the behaviour of drivers is in the pipeline at UCT, thanks to a connection made by UCT eResearch between researchers from two different campuses, and two entirely different fields of research: psychiatry and engineering.

In order to better understand the cognitive effects of certain chronic conditions on commercial drivers, Dr Hetta Gouse, chief research officer in the Department of Psychiatry and Mental Health, acquired a driver simulator in 2015. This driver visualisation tool is key to her research into driver behaviour in specific simulated road-use scenarios. But her pilot study presented a few obstacles, as far as software for the simulator was concerned.

"We ran a pilot study in 2016, and were simply not happy with the software provided by the company from which we sourced the simulator, or the support we received," says Gouse.

Finding the right software

Gouse eventually found the best software solution to be one provided by a US-based company, Systems Technology Inc, which specialises in the development of virtual-reality software for driver simulators, offering a range of road challenges to measure driver performance.

"Their software, STISIM Drive, proved a particularly good solution, because it's relatively easy to write your own programs," says Gouse.

But there was a challenge in terms of the cost of the US software, and it was here that Gouse reached out to UCT eResearch to find out if they could help. eResearch director Dr Dale Peters introduced Gouse to Associate Professor Marianne Vanderschuren, in the Department of Civil Engineering.

Connecting researchers: collaborating for software acquisition

"Marianne had been part of the academic reference group established to assess research needs in data visualisation," explains Peters. "By a stroke of serendipity, UCT eResearch was aware of her interest in driver simulation as a visualisation tool in road design and construction, and knew there would be synergy with Hetta's work."

Vanderschuren quickly saw the value in the driver simulator for her own research and put in a grant application to help fund the cost of the software. Between them, they managed to successfully fund the software necessary to run the simulator.

They now also have a research collaboration in the pipeline. Together, Gouse and Vanderschuren plan to study the differences between first-generation drivers – those whose parents did not drive – and second-generation drivers, who grew up with parents comfortable behind the wheel.

In the meantime, Gouse has secured a second developmental grant from the HIV Neurobehavioural Research Programme – this one to look at the impact of HIV and ageing in the workplace in a study of professional drivers.

"As we learn more about the neurocognitive impact of certain chronic conditions on the central nervous system, it is also important to understand the behavioural aspects that go with it, particularly where safety is a concern. This is particularly important in developing policy around driver management of both private and commercial drivers."●

Guests to the launch of the Iziko Planetarium and Digital Dome were treated to a display of how the new facility can be used for both research and edutainment. This advanced planetarium technology with immersive visualisation facilities is not only able to help researchers rapidly advance our understanding of the world, but also to make that information available to the public in an easily accessible, visual form.

Visualising a universe of data

Cape Town's 30-year-old planetarium has been revamped by Iziko Museums to create a facility that brings data to life through immersive visualisations. The new planetarium – the result of a partnership that includes UCT – could be a model for others around the world.

By harnessing advanced technologies to create immersive visualisations, the Iziko Planetarium and Digital Dome is in a position not only to help researchers rapidly advance our understanding of the world, but also to make that same information available to the public in an easily accessible visual form. This is particularly true for big data.

Big data refers to the large, complex data sets created and collected through technology. They can range from the data generated by social media or projects such as the Square Kilometre Array (SKA) to the big data produced by genome sequencing.

"In the world of huge data sets, some data can only be understood if you can see it," says UCT Emeritus Professor Danie Visser, patron of the Iziko Planetarium and Alexander von Humboldt Fellow at the Max Planck Institute for Comparative and International Private Law in Hamburg. "This is a powerful tool across all disciplines."

The SKA project is an obvious candidate for the digital dome. Massive data sets are already being created as MeerKAT, the precursor to the SKA telescope, comes online.

This has positive implications for the public, too. It means that, in time, we won't only be reading about the SKA discoveries in headlines; we will also be able to view the secrets of the universe at the planetarium. And, as more researchers from other fields use the facility for data analysis, more groundbreaking science will be visualised and made available to the public. ●

Using the dome to study the structure of the universe

Professor Tom Jarrett, UCT's Department of Science and Technology/National Research Foundation South African Research Chair in Astrophysics and Space Science, is studying the structure of the universe, trying to better understand the life cycle of galaxies.

Galaxies are not scattered randomly around the universe; they are clustered together in the groups that make up the cosmic web. To study these structures, Jarrett works with data sets that can easily include a million or more galaxies.

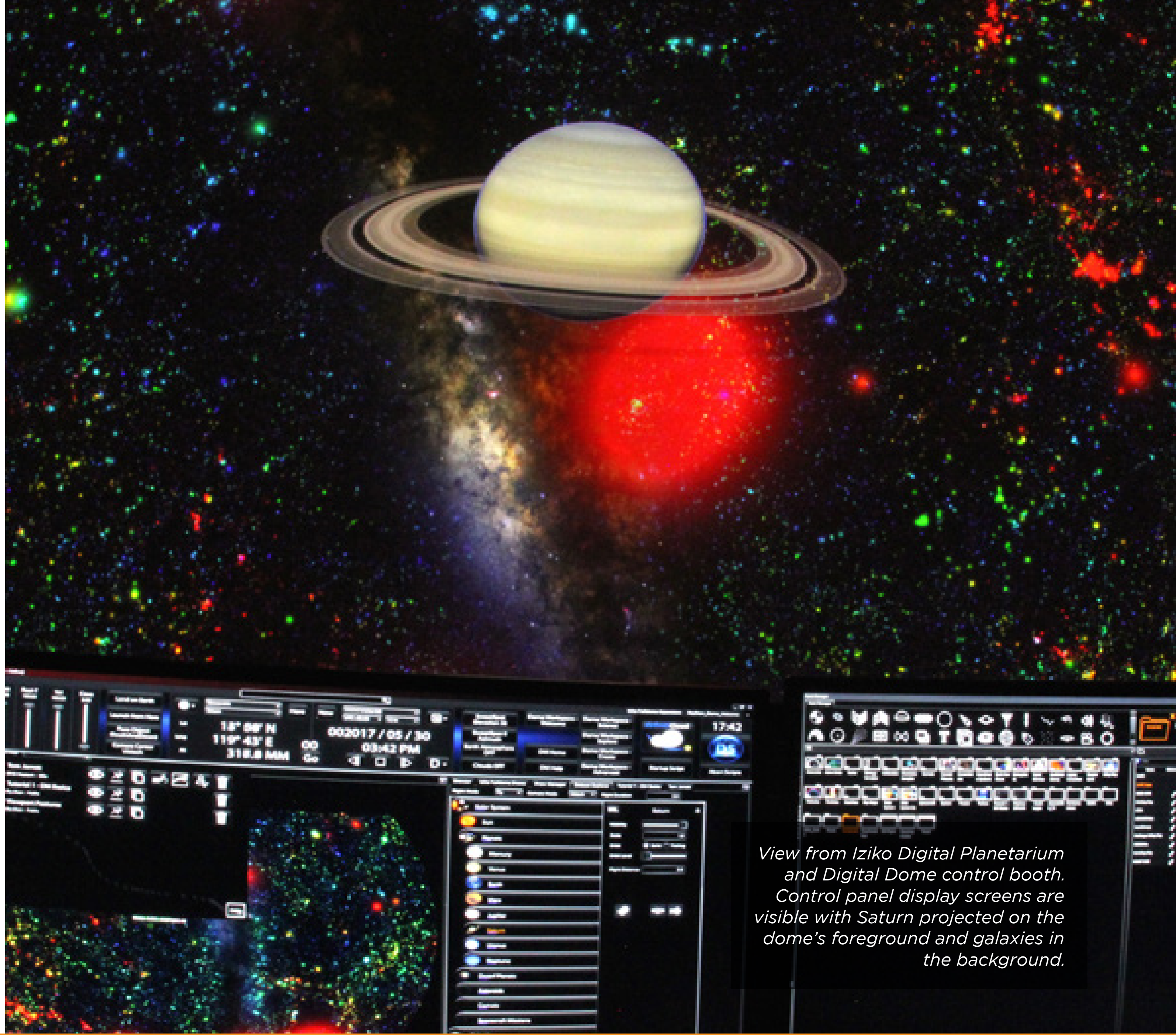
"These cosmic structures are really big; it's difficult to study them on a computer screen," he says. Each data set is loaded into a 3D catalogue in which each galaxy is mapped according to its coordinates in space. Because Jarrett has the coordinates, he can project them onto the dome and fly around the galaxy structures.

"Within these galaxy catalogues, I am trying to find new structures of galaxies," he explains. "With the immersive dome, I can actually fly into the data set. I can spy and isolate particular structures and accurately measure the gaps and filaments between the galaxy clusters."

So far, Jarrett is the only researcher to use the digital dome for research. Part of his role is to ensure the computing support and necessary software for this facility is available for researchers to use. With his colleague Professor Michelle Cluver, associate director of the Inter-University Institute for Data-Intensive Astronomy, he is also running a series of workshops and presentations for researchers in the Western Cape, to showcase how the facility can be used for research purposes.

"To learn how galaxies are born, evolve and grow, we need to study their environment – where they draw fuel from to grow, and gravity to shape their development. This is the cosmic web that we consider the context for galaxy evolution," says Jarrett.

"Because of the immersive nature of the dome, this is an excellent facility to help researchers really get deep into their data. It offers both the breadth and the 3D capacity to allow you to study your data from all angles, which are so limited on a flat screen." ●



View from Iziko Digital Planetarium and Digital Dome control booth. Control panel display screens are visible with Saturn projected on the dome's foreground and galaxies in the background.

Introducing ZivaHub: Open Data UCT

Story by Kate-Lyn Moore

Niklas Zimmer (right) presents with DVC Mamokgethi Phakeng on UCT's new institutional data repository.

Photo: Michael Hammond

Following the official launch of ZivaHub during October 2017, UCT Libraries is the first academic library in the country to have a fully functioning open-access institutional data repository (IDR) available to its research community.

The new repository allows researchers and the institution at large to make research data findable, accessible, interoperable and reusable. "We aspire to maintain our position as a leading academic library, firmly located in Africa," says Gwenda Thomas, the executive director of UCT Libraries.

The move was motivated, in part, by the university's 2015–20 Strategic Plan, where UCT made a commitment to ensuring that its research is visible and discoverable.

UCT Libraries already plays a central role in enacting this goal, explains Thomas. It does this through open-access portals such as OpenUCT, as well as online digital collections. More recently, however, it has entered into the world of research data-management support.

"We took the brave step within our strategic plan to say we actually want to implement research data services." And part of that is to have a data repository where research data can be shared and published, she says.

Why open data?

There are dual drivers behind the project. First is the need to comply with mandates from various funders, journals and organisations. The National Research Foundation (NRF) Open Access Statement of 2015 is one such example.

NRF-funded researchers and students have to publish their data as open access and UCT needed a mechanism to support that mandate, explains Niklas Zimmer, manager of Digital Library Services at UCT Libraries.

But the programme is not only about assisting researchers with compliance. The platform offers innovative benefits to UCT researchers, and will support them in raising the profile of their research outputs.

UCT's institutional data repository is powered by Figshare for Institutions. It is an online platform, where researchers can upload the processed data that either directly supports their research, or that constitutes a research output in its own right.

It serves researchers in need of a repository to store and disseminate their data "as openly as possible, as closed as necessary". Importantly, the data on ZivaHub can be linked to published research findings hosted on other platforms, such as OpenUCT.

Once uploaded to the portal, research data can be made as open and accessible as is required. One could opt, for instance, for one's data to be entirely private – while enjoying the benefits of a persistent identifier, ie a digital object identifier (DOI) that is recognisably from UCT – or for data to be visualised live online and fully downloadable.

Increased citations

"Researchers can validate and authenticate their research outputs, and more than anything, they can increase their citations. And we know how valuable that is to them," explains Professor Mamokgethi Phakeng, deputy vice-chancellor for research and internationalisation.

"The more people that have access to your data and your publications, the more your work gets cited and used."

This is incredibly valuable for researchers applying for ratings or funding. It also provides opportunities for this research data to be validated and replicated. Moreover, the sharing of data will reduce duplication of effort. Researchers will no longer have to collect or create data that already exists.

Photo: Michael Hammond

DVC Mamokgethi Phakeng introduces ZivaHub – UCT's new institutional data repository.

For the public good

The platform allows for efficiency, transparency and democratic control, the importance of which cannot be understated, especially in a country like South Africa, says Zimmer.

As more open data sets become available, the public will increasingly benefit from this research. And indeed, as this work is publicly funded, it should be considered a public good, he adds.

Figshare for Institutions

With UCT having opted for Figshare for Institutions, instead of the free figshare.com service, researchers will not be limited by space constraints. They are also afforded a great deal more control in terms of how their data will be disseminated.

Figshare for Institutions also allows for a number of other vital functions, such as curation. This means that UCT Libraries is able to delegate curation functions to designated individuals across campus in assisting researchers wanting to share data, by uploading and managing on their behalf.

Such a level of control and support is not possible with the free figshare.com platform, explains Zimmer.

Zivahub: Open Data UCT

The team challenged the UCT community to name this new platform.

The winner was student Thandeka Chehore, who submitted the name ZivaHub, together with the tagline Open Data UCT.

Ziva is a Shona word meaning "to know", Phakeng explains. "We are not just a South African institution. We are an African institution. And we wanted to foreground our African identity.

"Whether you are in Cambodia, the UK or Mauritius, ZivaHub says: 'I'm African. I like being here. And this is about knowledge. This is about open data.'"

ZivaHub is available to all students and staff at UCT. ●

Meeting our national research infrastructure needs

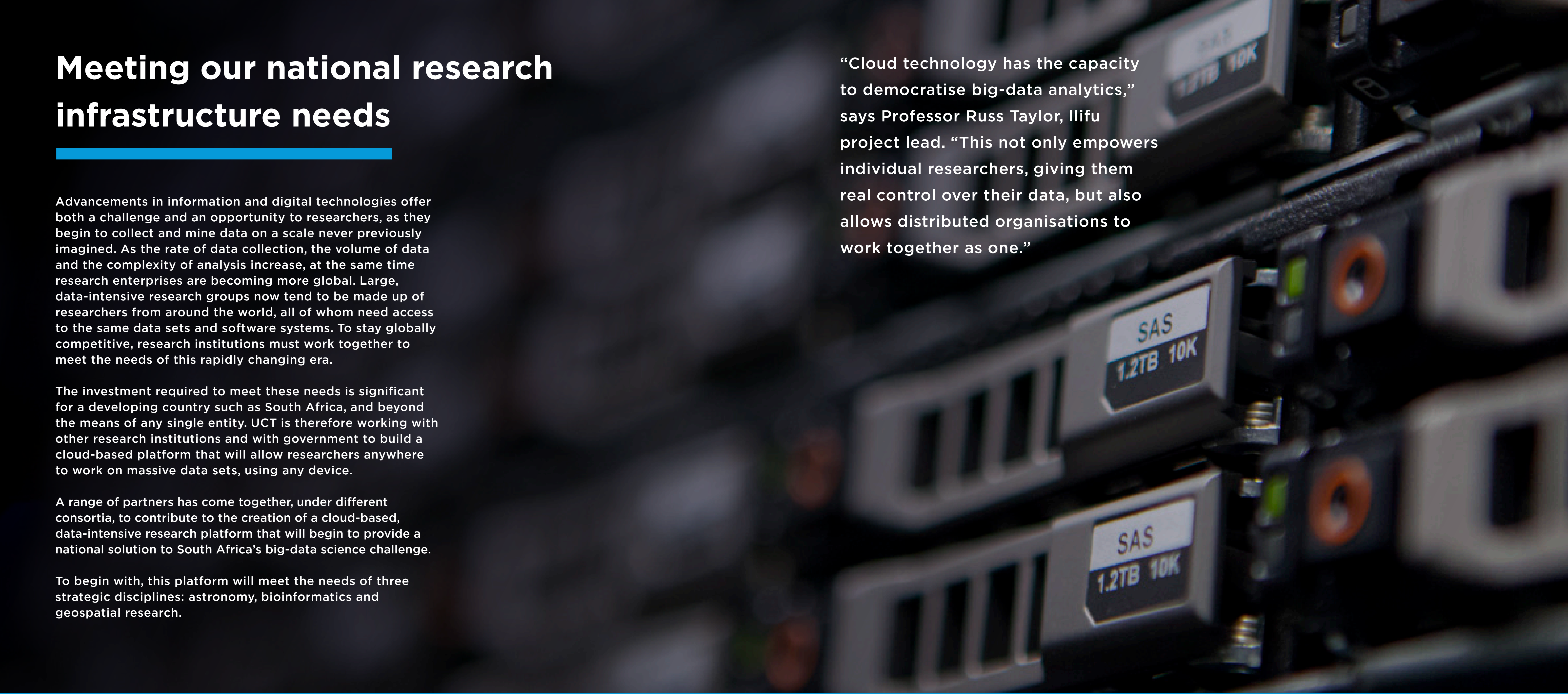
Advancements in information and digital technologies offer both a challenge and an opportunity to researchers, as they begin to collect and mine data on a scale never previously imagined. As the rate of data collection, the volume of data and the complexity of analysis increase, at the same time research enterprises are becoming more global. Large, data-intensive research groups now tend to be made up of researchers from around the world, all of whom need access to the same data sets and software systems. To stay globally competitive, research institutions must work together to meet the needs of this rapidly changing era.

The investment required to meet these needs is significant for a developing country such as South Africa, and beyond the means of any single entity. UCT is therefore working with other research institutions and with government to build a cloud-based platform that will allow researchers anywhere to work on massive data sets, using any device.

A range of partners has come together, under different consortia, to contribute to the creation of a cloud-based, data-intensive research platform that will begin to provide a national solution to South Africa's big-data science challenge.

To begin with, this platform will meet the needs of three strategic disciplines: astronomy, bioinformatics and geospatial research.

"Cloud technology has the capacity to democratise big-data analytics," says Professor Russ Taylor, Ilifu project lead. "This not only empowers individual researchers, giving them real control over their data, but also allows distributed organisations to work together as one."



African Research Cloud (ARC)

The ARC, a collaboration between UCT, the Inter-University Institute for Data-Intensive Astronomy (IDIA) and North West University, is the prototype for a cloud-based service to researchers working in data-intensive disciplines. Established in 2016, the ARC is testing different models of data management, storage and transfer through radio astronomy and genomics projects.

“The initiative is a first for Africa, and will be a real benefit to researchers on the continent,” says Sakkie Janse van Rensburg, executive director of Information Communication Technology Services (ICTS).

South African Data-Intensive Research Cloud (SADIRC)

Given the success of the ARC prototype, the next step is the expansion of the ARC to include a greater number of research institutes, including both universities and organisations such as SKA South Africa and the South African National Space Agency (SANSA). A memorandum of understanding was in development at the time of writing, which will formally constitute SADIRC.

In time, it is hoped, SADIRC will expand to offer access to storage for massive data sets, as well as the tools and software to properly collaborate on, analyse and visualise the data – to all South African researchers, including those based at our most under-resourced institutions.

The Ilifu partners

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Cape Peninsula University of Technology

SKA AFRICA
SQUARE KILOMETRE ARRAY

SOL PLAATJE UNIVERSITY

UNIVERSITEIT STELLENBOSCH UNIVERSITY

UNIVERSITY of the WESTERN CAPE

Ilifu

The establishment of the ARC meant that UCT was perfectly placed to lead a consortium of institutions in the Western Cape province of South Africa to put in a bid to the National Integrated Cyberinfrastructure System (NICIS), supported by the Department of Science and Technology (DST). The goal of this bid was to build a data-intensive research facility in the Western Cape that would cater explicitly to the needs of researchers working in astronomy and bioinformatics. The bid was successful, and today, this project is known as ‘Ilifu’ (‘cloud’, in isiXhosa).

Ilifu will receive funding from the DST for a period of three years. It will bring together the existing infrastructure and expertise of the various partner institutions, and build on that to create a hub for data-intensive research systems, platforms and tools in the Western Cape. A further mandate for Ilifu is the development of a research data management system (see page 36).

Within the three-year funding period, Ilifu is set to continue as a self-sustaining facility.

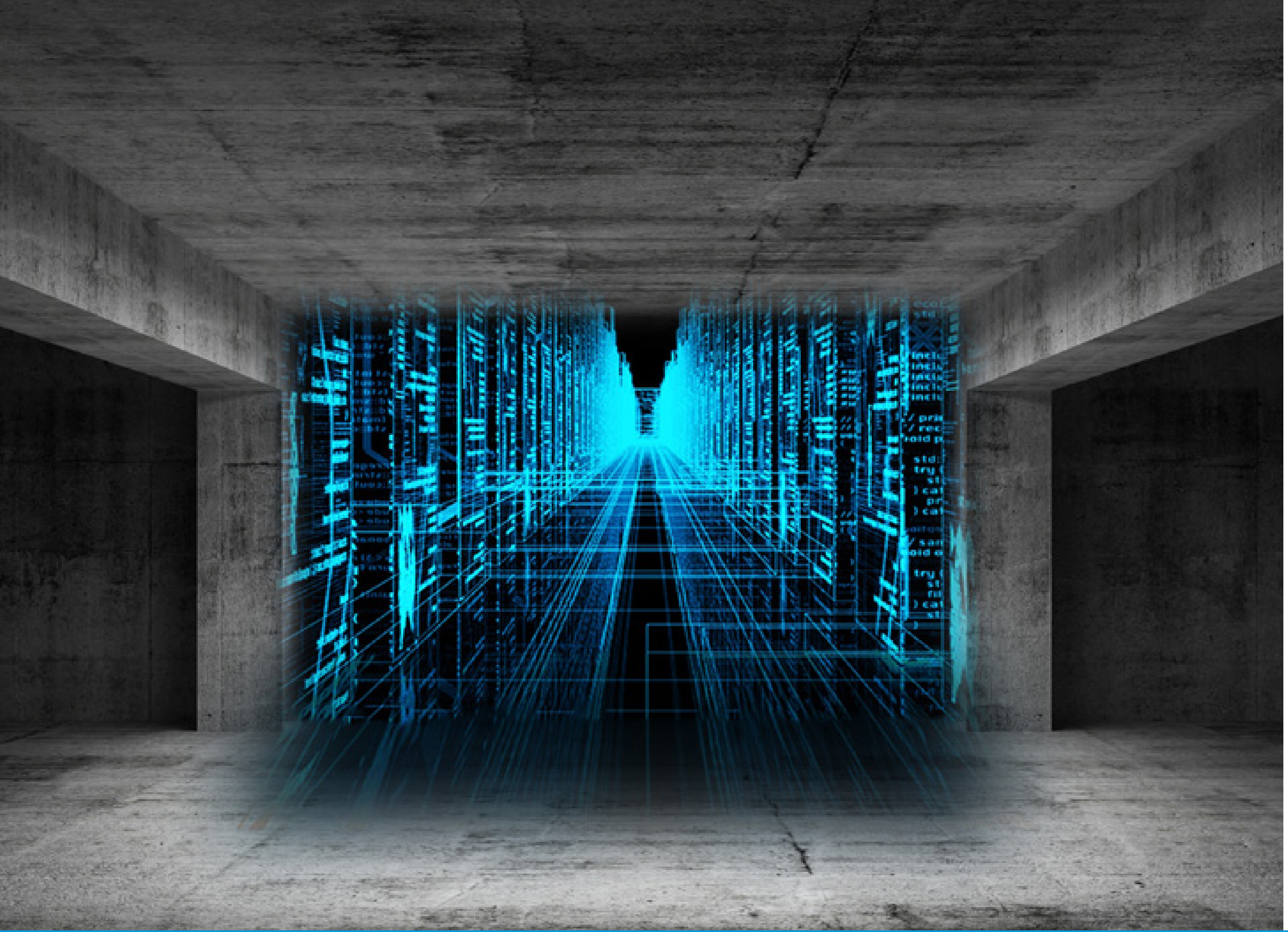
National Integrated Cyberinfrastructure System (NICIS)

The NICIS is a national initiative of South Africa’s Department of Science and Technology. The strategy defines different tiers of research infrastructure. Tier III provides institutional infrastructure, and tier II is regional – and, as is the case with Ilifu, involves the collaboration of several universities. Tier I refers to national-level infrastructure.

The sum is greater than its parts

While the investment in infrastructure is segmented, the offering itself is greater than the sum of its parts. Working together, Ilifu, the ARC and (in time) SADIRC will provide researchers access – through an online portal – to the entire tiered infrastructure system, as one entity (see box above).

A researcher will thus be able to log in to the online portal, from any device in any location, to access the stored data sets and run the necessary programs to analyse and visualise the data. ●



Preparing the next generation for the big-data challenge

Demand for skills at the interface between technology and information is growing and demand already far exceeds supply. This is particularly the case in Africa. To respond to demand, UCT launched two new postgraduate programmes to foster a generation equipped with the skills to meet this need.

Master's in digital curation

UCT is the first university in Africa to provide a master's-level course in digital curation: that is, the selection, maintenance and archiving of digital data repositories. The course, offered by the Library and Information Studies Centre, provides its graduates with a comprehensive set of digital curation skills applicable to any sector.

The one-year coursework component of the programme allows for specialisation, following a core course on the principles, theory and philosophy of digital curation.

Master's in data science

This interdisciplinary master's degree aims to furnish graduates with the statistical and computing skills needed to deal with big data from the fields of astronomy, physics, medicine and commerce.

The programme is composed of two equally weighted components: coursework and a dissertation on a research topic related to data science in astronomy, bioinformatics, computer science, physics or statistical sciences. Students have the choice of two streams for the programme: a general stream and a stream specialising in financial technology.

Work-integrated learning programme

Students in the two programmes will have the opportunity to be placed with projects supported by Ilifu, a consortium of Western Cape institutions, including UCT, that together are establishing and operating a data-centric, high-performance computing facility for data-intensive research. The students will work with data managers to implement research data management policies and services. ●

UCT's High Performance Computing Facility 2016 to 2017

15
CITATIONS

1 428
CORES

219TB
STORAGE

510 000
JOBS PROCESSED

4.02
MILLION
CPU HOURS*

130
USER UPTAKE

Growth since 2012

Year	Citations	Cores	Storage (TB)	Jobs	CPU hours*
2012	8	232	9	75 466	364 343
2013	13	292	56	111 081	1 496 211
2014	19	1 142	169	985 216	4 163 686
2015	26	1 458	219	152 690	5 997 849
2016	27	1 458	219	260 678	5 500 000

**CPU time is the amount of time a central processing unit (CPU) was used for processing. If two CPUs work on a job for an hour, this equates to two CPU hours.*

EDITORIAL
Lisa Boonzaier and Natalie Simon

DESIGN AND LAYOUT BY ROTHKO

With special thanks to all researchers and staff members featured in this report.

Visit our website (eresearch.uct.ac.za) to keep up to date with the latest UCT eResearch news.