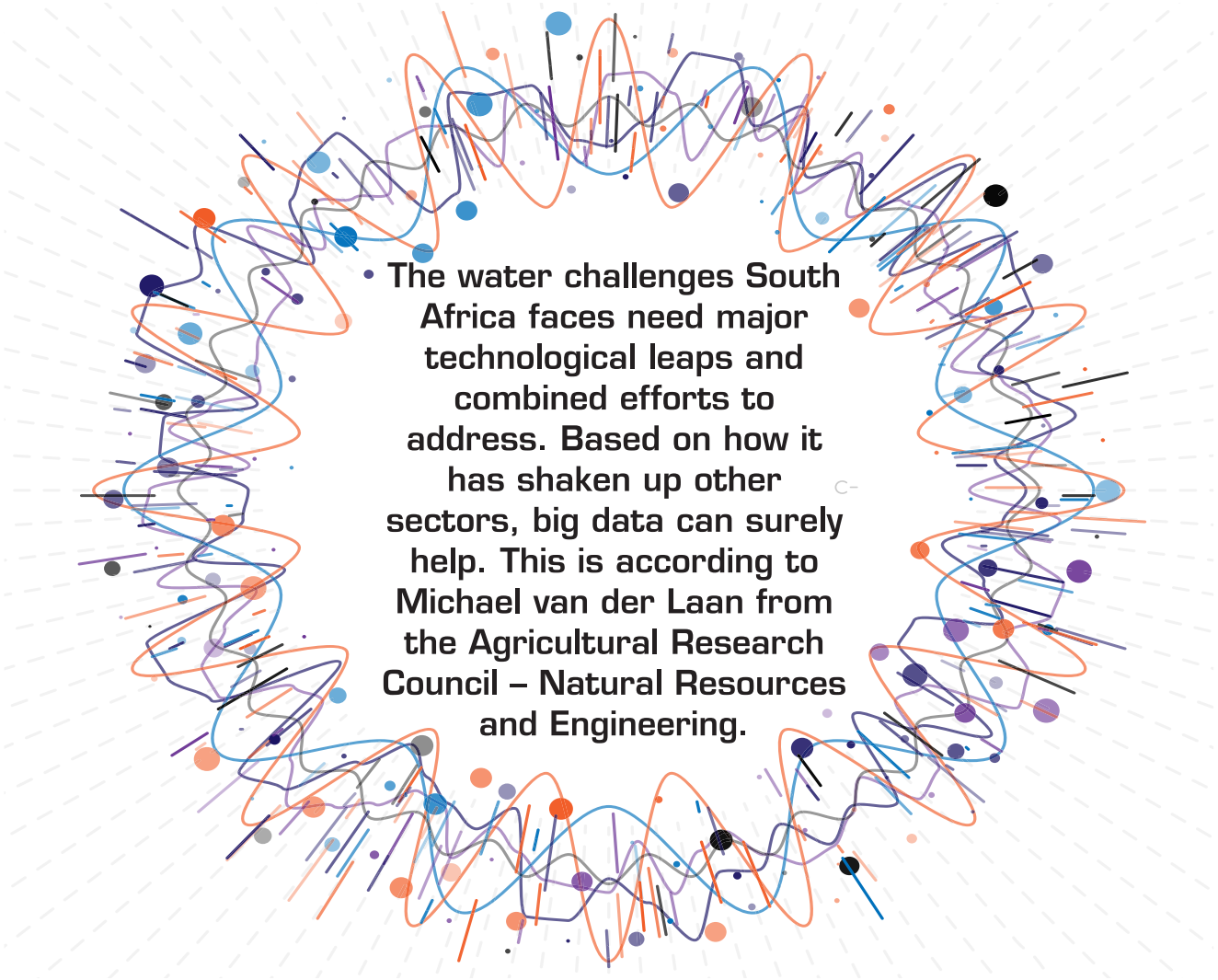
FOURTH INDUSTRIAL REVOLUTION

Water research in South Africa: Getting ready for big data analytics



Defined as very large and growing datasets from a variety of sources, big data cannot be processed using our traditional tools and digital infrastructure. A few decades ago, electronic spreadsheets revolutionised the workplace. They meant no longer having to add up long columns of numbers with a calculator, and enabled one to easily test different scenarios by changing a value in a cell or modifying a formula.

Although big data can be way more powerful, it is much more difficult to work with. Higher level policy development (e.g. custodianship, pricing of products and services) is also needed to ensure that big data is used in a scientifically robust and ethical manner that benefits all.

Commercial companies that have reaped the benefits of big data are plentiful. Biomedical research has also traditionally tackled large volumes of data with great success. In South Africa, institutions conducting water-related research are now proceeding with the development of cloud-based platforms for big data analytics, such as machine learning exercises. New techniques to analyse data, to create predictive models, and to link water management communities around heterogeneous

data sources are exciting.

The potential to produce multiple lines of evidence towards scientific facts as never before is also brilliant. In time we will be able to detect a chemical spill or illegal activity with satellite imagery, receive confirmation via social media streams, hold the perpetrator accountable using this evidence, and then monitor the recovery of the system in real-time using citizen science. We will also be able to better quantify the benefits of social development spending and policies on actual quality of life at various levels and scales. Internet-of-things (IOT) technology means a water quality sensor can be dropped in a river to flow downstream and report flow rate and water quality wirelessly.

For different groups selecting host platforms and designing their big data architecture, the main focus should be on ensuring inter-operability so that systems can communicate and work together. This will include the use of common metadata standards, file types and translation tools. However, ensuring legal inter-operability and the governance and protection of sensitive data will be more complex. Massive strides are still needed to develop policy documents and protocols around this aspect, not only in South Africa, but around the world.

A paradox of big data is that as the quantity goes up, the quality goes down. We are now in the zettabyte era (that's 1,000,000,000,000,000,000,000 bytes). There are over 60 zettabytes today and by 2025 there will be around 175. The skills needed to perform big data analytics (***software intensive science***) still mostly sit with highly experienced computer scientists. Applying artificial intelligence actually first needs a great deal of human intelligence. Application of big data analytics with wide ranging data sources may produce strong correlations, for example, but these may be spurious correlations and may not contribute to actual systems understanding or the generation of new knowledge. Still, the possibility of improved predictive models and many other applications is undeniable.

But does South Africa have the resources and skills ready to leverage big data to better manage our water resources? Right now, natural scientists with computer science skills are in short supply (or computer scientists with strong natural science theoretical knowledge for that matter). The training and upskilling of well-equipped water scientists will require careful thinking through of curricula, and not just in the STEM (science, technology, engineering and mathematics) disciplines. Digitising historical information and digitalising systems will require high levels of creativity from artists and graphic designers to improve user experience and interaction. Historians, anthropologists and language experts will be needed to help develop software applications that make use of natural language processing, as an example, to understand material ranging from ancient documents to real-time social media text and videos. Drama graduates could help better discuss important concepts such as risk and uncertainty with the public on critical issues such as climate change and vaccines.

With all the free online learning opportunities, we have the ability to adopt a lifelong learning philosophy and continually upskill as never before. Basic understanding of how software code is written and run in a model can go a long way in launching a career in data science, or facilitating collaboration between programmers and natural scientists. So, if you are an early career water scientist and have some time to spare, consider doing a coding course on Python, Java or something similar. You won't regret it.

The Water Research Commission is developing a cloud-based big data platform called the Water Research Observatory (www.waterresearchobservatory.org) in a collaborative project between the Agricultural Research Council (ARC) and the universities of Pretoria, Free State and KwaZulu-Natal in South Africa, as well as the University of Florida and Texas A&M University in the USA (**WRC Project No. C-2020-2021-00440**). The objectives are to digitise previous WRC-funded research projects, demonstrate the application of big data analytics to current challenges in three case studies (agricultural water, surface water and groundwater), and suggest a protocol for digitally archiving future WRC research project data and information.

A commitment to data democratisation will mean that even non-computer specialist citizens will be able to access and visualise the data in a way that hopefully serves their interests. For example, dropping a pin on any point on a map of South Africa will indicate the data available for that particular location. All sorts of applications can also be built on it by third parties using a range of digital tools. As computing costs continue to fall, even projects on a shoestring budget will be able to utilise big data technology, including powerful cloud processing where researchers have hardware limitations.