



science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA



Proceedings of the Annual National Research Data Workshop 2023



Table of Contents

Message from the NICIS Centre Manager	3
Organising Committee Members.....	5
Foreword by Dr Nobubele Shoji	6
Workshop Programme.....	7
DSI Presentation Speaker.....	12
Opening Speaker	12
Keynote Speaker	12
Abstracts of Presentations.....	13
Pictures of the event.....	34
Conclusion.....	35

Message from the NICIS Centre Manager



Dr. Happy Sithole

On behalf of the Organizing Committee of the National Research Data Workshop, I am delighted to welcome the presenters and delegates to this event. DIRISA forms part of the National Integrated Cyberinfrastructure System (NICIS) and is responsible for the data services of NICIS. The annual research data workshop provides an opportunity for the research community to share and interact with peers on topics related to research data management and data intensive science. It also allows delegates to voice their data management needs and this

enables DIRISA to provide more suitable services to the research community.

The presentations covered a broad range of topics, from architectures for open data and cloud-based big data modelling, to AI policy in Africa and data analysis for public health response. It is heartening to see that most of the presentations originate from South African and African authors. These presentations address challenges and opportunities that are eminently relevant to South Africa and Africa. They also reflect the excellent progress in the management of research data and the status of data intensive research and development on the African continent.

The keynote address by Dr Jabu Mtsweni on data driven decision making, addressed very pertinent issues now topical with the increasing impact of AI applications in society. The presentation by Mr Daniel Mokhohlane of the Department of Science and Innovation (DSI) demonstrated the importance of data intensive research in achieving goals of national strategies and for the national system of innovation. The representation of DSI at this event is much appreciated.

We are grateful to the presenters for their contributions that made this event a successful one. A double-blind review of abstract submissions was conducted for the first time since the inception of this event, which contributed to the high quality and greater relevance of presentations.

My thanks to the members of the Review Committee for their gratuitous and selfless efforts. Thanks also go to the organisers of this event, who are mentioned further. Organising this event is not an easy task, and thanks go to the organisers, Mr Xolani Nkosi, Dr Nobubele Shoji, Mrs Millicent Bambo and the rest of the DIRISA team for their efforts and hard work in seeing to the organisational and logistical tasks that made this event possible.



Importantly, thank you to the presenters and delegates who attended this event. We are grateful for your participation and in sharing your work with the research community. Your contributions will undoubtedly have benefit in other applications and advance the knowledge of attendants. Lastly, appreciation to DSI for their support through NICIS, without which, this event would not have been possible.

Organising Committee Members



Dr. Nobubele Shozi

Data Stewardship Manager of DIRISA

Passionate about HCD



Mr. Xolani Nkosi

Project Coordinator of DIRISA

Managing the programme, invitations, abstract submissions and double-blind reviews



Mrs. Millicent Bambo

Group Assistant of DIRISA

Event logistics, venue, flights, accommodations

A big thank you to our programme directors and the reviewers who conducted the blind reviews!

Foreword by Dr Nobubele Shozi



Dr. Nobubele Shozi

The DIRISA National Research Data Workshop is a key event that brings together topics from a wide range of areas relating to research data management. The purpose of the workshop is to advance data intensive research, research data management and improve the understanding of how these topics can improve the quality of lives in South Africa and Africa at large. This is important as today we live in a digital and interconnected world. The presentations delivered in this workshop were not only of high quality, but they also showed the importance of

data and the importance of being part of the digital revolution. This importance cannot be understated - South Africa needs to contribute to the development of a data literate society and the workshop is one key aspect towards ensuring that the need for a data literate society increases.

The DIRISA Coding School and Student Datathon Initiatives also formed part of this prestigious event. These two initiatives were started by DIRISA in 2019 to contribute towards the development of digital skills in South Africa. Data science skills, artificial intelligence, machine learning and the introduction of coding are some of the skills that these initiatives provide with the aim of ensuring that the youth of South Africa is able to use the new knowledge they gain to provide new solutions and solve problems in this digital world while contributing to South Africa being a capable state.

Lastly, this workshop played an important networking and connecting role by bringing together people with common interests and expertise and allowing them to share ideas that can be applied in other areas. I am confident that this workshop was a very successful, informative and enjoyable event! We look forward to the 2024 National Research Data Workshop!

Dr. Nobubele Shozi

DIRISA Data Stewardship Manager

Workshop Programme



NATIONAL RESEARCH DATA WORKSHOP

DATE: 05 – 06 JULY 2023

VENUE: CSIR - ICC

Wednesday

July 05

08:30 - 08:35 Welcoming: Programme Director; Ms Susan Veldsman

08:35 - 08:55 Presentation: [DSI Deputy Director; Cyber Infrastructure](#):
Mr Daniel Mokhohlane

08:55 - 09:15 Opening: NGEI Executive Manager; Dr. Lulama Wakaba

09:15 - 09:35 [Keynote Address](#); Dr. Jabu Mtsweni

09:35 - 10:05 **Break**

10:05 - 10:25 [Investigating the effect of Activation functions and sample sizes on the performance of artificial neural network models through simulation](#); Dr. Alexander Boateng

10:25 - 10:45 [Update on NICIS data transfer services](#);
Dr. Kasandra Pillay

10:45 - 11:05 [Development of the open science platform](#);
Dr. Tshiamo Motshegwa

11:05 - 11:25 [Investigating noise diode power stability in radio astronomy: Data cleaning, relative difference analysis, and optimal gain estimation](#); Mr Kamva Vanqa

NATIONAL RESEARCH DATA WORKSHOP

DATE: 05 – 06 JULY 2023

VENUE: CSIR - ICC

Wednesday

July 05

11:25 - 11:45

[Can artificial intelligence mitigate climate change?](#)

Prof. Bhekisipho Twala

11:45 - 12:05

[Big data and machine learning with HIRAX and KREATE;](#)

Dr. Kavilan Moodley

12:05 - 13:25

Lunch break

13:25 - 13:45

[Estimation of the disaggregated crime rate based on large-scale surveys and population census;](#) Dr. Yegnanew Shiferaw

13:45 - 14:05

[Optimal energy cost management of an on-grid battery swapping station based on solar-wind power system under Time-of-Use tariff: South African Context;](#) Mr Lumbumba Nyamayoka

14:05 - 14:25

[Paraphrase generation model using transformer-based architecture;](#) Mr Dan Masethe

14:25 - 14:45

[An Architecture for open datasets of South Africa towards machine learning algorithms;](#) Prof. Nobert

14:45 - 15:05

[A framework for recursively adapting research work to an integrated cloud-based big data modelling System;](#) Dr. Kassim S. Mwitondi

15:05 - 15:25

Break

15:25 - 15:45

[Dealing with disparate data;](#) Rupert Spann



NATIONAL RESEARCH DATA WORKSHOP

DATE: 05 – 06 JULY 2023

VENUE: CSIR - ICC

Wednesday

July 05

15:45 - 16:05 [The IDIA science gateway](#); Prof. Rob Simmonds

16:05 - 16:25 [DIRISA services: Overview of DIRISA's services](#); Mr
May Maile

16:25 - 16:30 Warp up Day 1: Mrs Susan Veldsman



NATIONAL RESEARCH DATA WORKSHOP

DATE: 05 – 06 JULY 2023

VENUE: CSIR - ICC

Thursday

July 06

08:30 - 08:40	Welcoming: Programme Director: Mrs Ina Smith
08:40 - 09:00	Deep learning for real-time traffic classification and resource allocation in community networks ; Dr. Josiah Chavula
09:00 - 09:20	Experimental and Non-traditional Archives in South Africa ; Prof. Hussein Suleman
09:20 - 09:40	Using Generative AI as a Data Engineer ; Mr Nhlakanipho Kwazi Mthembu
09:40 - 10:00	Data fluency for all through science engagement programmes ; Ms Kogie Govender
10:00 - 10:25	Break
Theme	Raising Data Literacy and Democratizing Data Insights
10:25 - 10:45	The World Data System: creating trusted communities of scientific data repositories ; Dr. Dale Peters
10:45 - 11:05	Exploring the evolving landscape of research data repositories and infrastructure: emerging trends and innovations ; Dr. Josiline Chigwada
11:05 - 11:25	A language to derandomize robotic device actions in swarm intelligence models ; Dr. Colin Chibaya



NATIONAL RESEARCH DATA WORKSHOP

DATE: 05 – 06 JULY 2023

VENUE: CSIR - ICC

Thursday		July 06
12:25 - 12:45	AI Policy in Africa, get involved. The role of grassroots AI in charting a joint future ; Dr. Vukosi Marivate	
12:45 - 13:25	Lunch Break	
13:25 - 13:45	Forecasting Emerging Contaminants Removal in Urban Water Cycle: Artificial neural networks (ANN) and adaptive neuro-fuzzy Inference system models ; Prof. Kapil Moothi	
13:45 - 14:05	Multivariate Time Series Forecasting with Machine Learning in Python: A comparative study of statistical models and deep learning models ; Mr Mphiliseni Nzuza	
14:05 - 14:25	SANReN; Digital Transformation; Strategy Execution Research & Innovation Cyber Infrastructure ; Mr Sabelo Dlamini (online)	
14:25 - 14:45	Surrogate modelling of wakes in wind farms in wake steering conditions ; Kayla Surujhlal & Ayodimeji Biobaku	
14:45 - 15:10	Break	
15:10 - 15:30	Application of artificial intelligence towards career exploration and decision making ; Ms Nosipho Mavuso	
15:30 - 15:50	DIRISA Services ; May Maile, Ntlharhi Baloyi & Hlawulani Sibisi	
15:50 – 16:00	Wrap up: Day 2: Mrs Ina Smith	

DSI Presentation Speaker



Mr. Daniel Mokhohlane

Deputy Director
DSI Cyber Infrastructure



Opening Speaker

Dr. Lulama Wakaba

NGEI Executive Manager
CSIR



Keynote Speaker

Dr. Jabu Mtsweni

Head of Information and Cyber Security Centre
and Chief Researcher
CSIR

Abstracts of Presentations

Title	Investigating the effect of Activation functions and sample sizes on the performance of artificial neural network models through simulation
Presenting Author	Dr. Alexander Boatengo - Senior Lecturer, University of Free State.
Abstract	Artificial neural networks (ANNs) have gained immense recognition as a powerful tool for a wide range of applications, including image recognition, natural language processing, and predictive modelling. Their capacity to decipher intricate patterns and relationships within data makes them highly sought-after. However, the effectiveness of ANNs heavily relies on the choice of activation function and the size of the training dataset. Consequently, this study sought to investigate the impact of activation functions and sample sizes on the performance of artificial neural network models through simulation. The Tanh, Logistic, ReLU, and Identity activation functions were compared in terms of accuracy and training time across diverse sample sizes. Notably, the Tanh activation function consistently showcased superior accuracy levels, ranging from 50% to 97%, surpassing the performance of the other activation functions. However, statistical analysis using ANOVA revealed no significant differences among the mean accuracy values for all activation functions, suggesting that the choice of activation function may not have a substantial impact on overall accuracy. Interestingly, the Identity activation function demonstrated significantly shorter training times, averaging 0.27 seconds, in comparison to the Tanh activation function's 1.66 seconds. Nevertheless, ANOVA analysis indicated no significant disparities among the mean training times for all activation functions, implying that the selection of activation function does not significantly affect training time. These findings provide valuable insights into the influence of activation functions and sample sizes on ANN performance, offering guidance for the selection of appropriate activation functions and sample sizes in different applications. Future research endeavours may explore the impact of other parameters on neural network performance and investigate the optimization of neural network architectures tailored to specific applications.

Title	Update on NICIS data transfer services
Author	Dr. Kasandra Pillay - Principal Engineer, CSIR.
Abstract	Moving masses of data is often a challenge. In most cases networks optimised for business operations are neither designed for nor capable of supporting the data movement requirements of data intensive research.

When scientists attempt to run data intensive applications over these so called “general purpose”/enterprise networks, the result is often poor performance – in many cases poor enough that the science mission is significantly impacted. At its worst this means either not getting the data, getting it too late or resorting to “desperate” measures such as shipping disks around.

We want to help users move around large data sets more effectively. This presentation will inform stakeholders on the large data transfer movement capabilities from the Performance Enhancement Response Team (PERT) within the National Integrated Cyberinfrastructure System (NICIS). The presentation will detail plans to update the SANReN Data Transfer Pilot to a production service. These plans include upcoming national data transfer infrastructure to support the SANReN 100Gbps backbone network, such as 100Gbps Data Transfer Nodes (specialised hardware) and 100Gbps perfSONAR node (network measurement) updates, implemented in the Science DMZ architecture.

The Globus software implemented on this infrastructure, along with its new capabilities such as integration to IRODs storage via IRODs connector and access to collections and data sharing will be described. The talk will also highlight plans for integration between the systems offered by the three NICIS pillars, Centre for High Performance Computing (CHPC), South African National Research Network (SANReN) and Data Intensive Research Initiative of South Africa (DIRISA) to assist with large data transfer movement.

Title	Development of the open science platform
Author	Dr. Tshiamo Motshegwa - Inaugural Director, African Open Science Platform (AOSP).
Abstract	This talk provides an update on the development of the African Open Science Platform (AOSP). AOSP aims to position African scientists at the cutting edge of data intensive science by stimulating interactivity and creating opportunity through the development of efficiencies of scale, building critical mass through shared capacities, amplifying impact through a commonality of purpose and voice, and to engage in Global Commons to address continental and global challenges through joint action.

Title	Investigating noise diode power stability in radio astronomy: Data cleaning, relative difference analysis, and optimal gain estimation
Presenting Author	Mr Kamva Vanqa - Data Analyst; South African Radio Astronomy Observatory (SARAO).
Abstract	The stability of noise diode power is crucial in accurate calibration and data analysis in the field of radio astronomy. In this study, we address the challenges posed by unstable noise diode power in interferometric/astronomical data, which often contains radio frequency interference (RFI) and other sources of noise. Our objective is to enhance calibration accuracy by implementing a comprehensive data cleaning approach. The approach involves data inspection, cleaning, and flagging of RFI-contaminated spectra. We quantify the relative difference in noise diode power between observations by comparing each observation to a reference observation chosen based on a median criterion. The mean square error (MSE) is used as a metric to evaluate the deviation between observations and the reference. We employ a gradient descent algorithm to optimize the gain value, aiming to minimize the MSE and ensure stability of the spectra over time. This approach belongs to a broader class in optimization theory, which is fundamental in data management and science. This project falls within the realm of data science, specifically dealing with large-scale interferometric/ astronomical data analysis, noise mitigation, and optimization using machine learning techniques. Our findings provide insights into factors influencing noise diode stability and contribute to improving calibration techniques in radio interferometry.

Title	Can artificial intelligence mitigate climate change?
Author	Prof. Bhekisipho Twala - Deputy Vice-Chancellor Tshwane University of Technology.
Abstract	Climate change is one of humanity's greatest challenges while Artificial intelligence (AI) and machine learning (ML) have been transforming scientific disciplines. Policymakers, mayors, administrations, and individuals seek to advance climate solutions and demand tailored approaches that match their political, economic and climatic conditions. Thus, climate change remains a complex scientific and multifaceted issue, amenable to AI and ML analysis, but in general, this has not yet occurred. Many AI and ML algorithms have been available for decades, possibly most notably deep learning and neural networks. However, until recently, constraints of computational architecture and power have restricted their application, especially for issues as data intensive as climate change.

In this paper, we describe how AI and ML can be powerful tools in reducing greenhouse gas emissions and helping society adapt to a changing climate. From smart grids to disaster management, we identify high-impact problems where existing gaps can be filled by AI and ML, in collaboration with other fields. Our results highlight challenges and opportunities for data-driven climate modelling, especially concerning the incorporation of even larger model datasets in the future while our recommendations encompass exciting research questions as well as promising business opportunities. We call on the AI and ML community to join the global effort against climate change and further encourage extensive data sharing among institutes to build even more powerful response climate emulators that would enable faster climate change projections.

Title	Big data and machine learning with HIRAX and KREATE
Author	Dr. Kavilan Moodley - ARC Director; University of KwaZulu Natal.
Abstract	The Hydrogen Intensity and Real-time Analysis eXperiment (HIRAX; https://hirax.ukzn.ac.za/) is currently being built at the SKA Karoo site in South Africa to measure the properties of dark energy and uncover the nature of fast radio bursts. The pathfinder array will comprise 256 six metre dishes observing a third of the sky in the 400-800 MHz band over a four-year period, with an eventual goal to build the full array up to 1024 dishes. The full array, which will present significant technical big data analysis challenges for the project. The raw data rate from the full 1024-element HIRAX array is 6.5 Tb/s, which yields a data volume of a few hundred TBs per day that is impractical to store. This means that data from the full HIRAX array needs to be cleaned, calibrated and compressed in real-time, which is a challenge faced in other scientific and industry projects. New approaches are required that simultaneously handle the scale of data efficiently while maintaining the precision required for separation of the cosmological signal from foreground and other contaminants. A key step in the data analysis process is data flagging to remove artificial signals such as radio frequency interference (RFI) from telescope electronics or terrestrial sources. Data calibration is also a real-time challenge and must deal with propagation and instrumental effects such as ionospheric corrections, and gain and bandpass fluctuations. The telescope data also needs to be calibrated for instrumental effects, such as antenna beams, pointing, geometric delays and polarization leakage. Finally, a high-fidelity galactic foreground model is required to enable a clean frequency-based separation of the cosmological signal from much brighter foreground contaminants. The big data analysis methods that are being developed to process the data at each of the above-mentioned stages need to be tested on large-scale simulations that include detailed modelling of the telescope response to the radio sky, to understand and refine these methods.

I will review the big data computing and analysis plans to process the large HIRAX data stream and describe novel machine learning applications to this dataset. Machine learning (ML) provides a new tool that can be applied to these analysis challenges. ML applications in radio cosmology have focused on RFI excision, foreground modelling using Gaussian processes or convolutional neural networks, generating large simulations of the cosmological signal with neural networks, neutral hydrogen imaging and tomography analyses using deep learning, and cosmological model comparison and classification. Motivated by the complexity of the model that describes how the sky signal measured by HIRAX propagates, is recorded, and must be processed to extract the 21cm signal, it is evident that novel ML techniques will play a key role in the HIRAX data challenge.

I will also highlight the KREATE initiative (<https://kcreate.ukzn.ac.za/>) that we are leading at UKZN in collaboration with the eThekweni municipality in the space of big data, machine learning and AI, to leverage HIRAX advances to benefit local industry.

Title	Estimation of the disaggregated crime rate based on large-scale surveys and population census
Author	Dr. Yegnanew Yegnanew - Researcher & Lecturer, University of Johannesburg.
Abstract	Statistics South Africa (Stats SA) offers a wealth of information about the national and provincial levels in South Africa. Due to financial and practical considerations, many nations, including South Africa, use sample surveys to estimate socio-economic indicators like proportions or counts at the national, regional, and provincial levels. However, in most cases, estimates at highly disaggregated levels, such as district municipalities or local municipalities, are unavailable as the sample size may not be sufficient for reliable results. Global frameworks of indicators for monitoring the Sustainable Development Goals (SDGs), for example, recommend that information be disaggregated not only geographically (in subregions of interest, such as provinces, municipalities, or districts) but also by income group, gender, age, race, immigration status, and disability status. Following this suggestion, this study provides crime rate estimates disaggregated by a district municipality, metro level, gender, and population groups using a hierarchical Bayes small area estimation approach. The model specifications encompass various linking models in the Bayesian framework as well as the widely recognized basic area-level Fay-Herriot (FH) model. In essence, three hierarchical Bayes models were proposed, namely: the FH model incorporating known sampling variance, the two-stage mixed log-normal model, and the two-stage mixed logistic model, both of which incorporate known sampling variance. It is well known that crime is a major problem in South Africa because it is a complex

social issue with broad implications for individuals, groups, and societies.

Estimating the crime rate for small areas is essential because it allows for targeted resource allocation, accuracy in policy development, intervention evaluation, identification of hotspots and vulnerable populations, and a deeper understanding of spatial patterns and context. This strategy promotes tailored interventions, encourages evidence-based decision-making, and results in more successful local strategies for crime reduction. Data from various large-scale household surveys, such as the 2016 Community Survey and the 2021/22 Governance, Public Safety, and Justice Survey, were combined with the auxiliary variables from the 2011 Population Census.

Title	Optimal energy cost management of an on-grid battery swapping station based on solar-wind power system under Time-of-Use tariff: South African Context
Presenting Author	Mr Lumbumba Nyamayoka - PhD Candidate, University of the Witwatersrand.
Abstract	Electric mobility is gaining popularity worldwide as countries strive to reduce their carbon footprint and transition to sustainable energy sources. However, the widespread adoption of electric vehicles (EVs) is hindered by the significant challenge of range anxiety, which refers to the fear of running out of battery power before reaching a destination. To tackle this issue, battery swap technology has emerged as a potential solution, allowing drivers to exchange depleted batteries for fully charged ones at dedicated stations within minutes. These battery swapping stations bear similarities to traditional gas stations in terms of design. In this paper, we propose a data-driven strategy that combines data science and research data management principles to optimise the operation of a battery swapping station. Our focus is on minimising energy cost consumption from the utility grid, which is a crucial aspect for the economic viability of electric mobility. The proposed system comprises a wind turbine generator, a photovoltaic generator, and a utility grid, with the battery swapping station serving as the primary load. By leveraging data science techniques, we collect and analyse data from these sources to inform decision-making and optimise energy management. The batteries used in the system can be in one of three states: off-charged, on-charged, or fully charged. Through the use of chargers, the battery swapping station charges the batteries, and the drivers can arrive at any time to replace depleted batteries with fully charged ones. Our study aims to develop an optimised schedule for the battery swapping station's activities, taking into account the Time-of-Use (TOU) tariff structure, which reflects varying electricity prices throughout the day. By employing advanced optimisation algorithms and linear programming

techniques, we aim to minimise the cost of electricity purchased from the utility grid. As part of our research data management, we carefully collect and organise relevant data, including energy consumption data, electricity pricing data, and historical charging patterns. Data analysis, visualisation, and modelling techniques are applied to gain insights from the data and develop an accurate representation of the system's behaviour. By simulating the proposed system over a 24-hour period using MATLAB, we aim to demonstrate the effectiveness of our developed optimal energy management model in reducing energy costs. Through this simulation, we showcase how our data-driven approach can effectively minimise the expenditure on energy purchased from the utility grid.

Title	Paraphrase generation model using transformer-based architecture
Presenting Author	Mr Dan Masethe - Head of Computer Science Department, Tshwane University of Technology.
Abstract	Paraphrasing and detection are essential tasks in natural language processing (NLP). Changing the style of text while preserving the semantics is computationally a complex task in natural language processing taking into account sentiment analysis. AI bot Paraphrase generation can unintentionally transform original sense or intention of the sentence, if high degree of semantic similarity and possible grammatical errors and sentence structure are not taken into considerations. Paraphrasing increase diversity of the sentence, which improve performance of NLP tasks such as machine translation and question-answering tasks. Semantic similarity or relationship quantify resemblance between text such as words, sentences and paragraphs in terms of meaning and concepts using specific criteria in NLP. The challenge is finding similarity of text to each other, even though considerable progress has been made in resourced-rich languages, low-resourced languages are lacking behind. The research develops a paraphrasing AI bot that involves altering a sentence using modified words and sentence structures while preserving original semantics using transformer architecture. The experiment using transformer-based architecture outperforms baseline models through metric evaluations and human evaluation.

Title	An Architecture for open datasets of South Africa towards machine learning algorithms
Author	Prof. Nobert Rangarirai Jere - Associate Professor, Walter Sisulu University.
Abstract	Machine Learning algorithms rely on reliability and availability of datasets. Such open datasets are produced and created by researchers from respective countries. With the developments within the field pf machine learning and artificial intelligence, South Africa remains lacking on current

datasets. Considering that South Africa faces many challenges in almost every sector ranging from education, health, agriculture, unemployment, high crime rate, Gender Based Violence cases and high levels of corruption, implementation of machine learning algorithm could be an innovative solution. Generally, policies, strategies, planning and large sums of money have been spent to assist in addressing current national challenges without success.

This study argues that proper application of machine learning using relevant and applicable data could assist some of the challenges facing South Africa. The problem at hand is that as Artificial Intelligence and open-source datasets emerge such as Kaggle, there are very few datasets available for South Africa. As researchers explore and gather data, it is critical to have a guiding architecture for data repository management. The key research question is: How can an open architecture for South African datasets be developed towards machine learning algorithms?

The study is based on current available open-source datasets using online secondary data. A number of datasets based on the common national challenges were considered. Kaggle was the open machine learning tool considered. Results show that a dataset search without mentioning South Africa will return more results. Once a search is more specified to South Africa, less results are returned. The searches were based on common datasets that were considered popular challenges. Results are clear that there are very few datasets from South Africa to assist machine learning experts. This paper proposes a standardized operational open dataset architecture that is applicable for South Africa. The architecture is expected to be secure, interoperable, standard and open to be adopted for different types of datasets. The study paves the way for machine learning researchers to have an integrated architecture that assist South African researchers to provide open datasets to improve decision making. The proposed architecture should guide in developing and managing South African datasets that will be available to international researchers when developing machine learning algorithms using available datasets.

Title	A framework for recursively adapting research work to an integrated cloud-based big data modelling System
Author	Dr. Kassim S. Mwitondi - Senior Lecturer, Sheffield Hallam University College of Business, Technology & Engineering (BTE)
Abstract	Advances in computing power, data generation, storage, transmission and sharing, are key drivers of data-intensive research across disciplines, with the potential to address societal challenges and harness opportunities that constantly arise in our Big Data era. A typical example of a problem space/solution, in that context, is the non-orthogonality of the United Nations Sustainable Development Goals (SDG). The 17 SDGs present an

excellent example of a problem space that is highly susceptible to socio-economic, cultural and technological variations.

We propose a framework for recursively adapting past research work to an integrated cloud-based data modelling system. We look back to how some of our recent research work has progressed; how it has influenced our current research and how it can be adapted to our fast-evolving research environment. The ultimate goal of this initiative is to develop a consilient interdisciplinary environment for SDG Big Data Modelling research that systematically recycles prior knowledge into modelling systems for generating new knowledge.

We illustrate the framework on an interactive Nuvolos-cloud technology using our previously developed SMA algorithm and selected recent past and current research in environmental monitoring and the agri-food industry. We observe that addressing the SDG's non-orthogonal challenge, requires consilient efforts, designed to lead to a unified understanding of concepts upon which legislations and decisions that affect our societies are made. We demonstrate that we can attain robust and consilient solutions to that challenge by sharing and recycling data.

Title		Dealing with disparate data
Presenting Author	Rupert Spann - Manager: Telescope Operations Development, South African Radio Astronomy Observatory (SARAO)	
Abstract	Operations of large scientific instruments involve the handling, integrating, analysis, and reporting of diverse datasets across various divisions. This also applies to a wide range of disciplines, including research, industrial, and commercial domains. The processing of information within these domains frequently entails substantial overlaps. However, the presence of optimisation techniques, commercial interests, and legacy systems often leads to non-comparable data structures, resulting in a loss of potential benefits. Consequently, data wrangling becomes a critical step due to the lack of comparability. To address this challenge, we propose a novel framework aimed at facilitating the transfer of structured information in an abstract manner, independent of specific schemas. Our goal is to enhance information discovery and integration capabilities by enabling application development that is schema-agnostic and independent of data sources.	

Title	The IDIA science gateway
Presenting Author	Prof. Rob Simmonds - Acting Director, Inter-university Institute for Data Intensive Astronomy (IDIA)
Abstract	The IDIA Science Gateway has been created to access the ilifu OpenStack system that is based at UCT and supports open science primarily in the areas of Radio Astronomy and Bioinformatics. This is the start of creating tools that can be used to set up a federation of cloud computing and storage systems throughout South Africa. Some of the requirements for the gateway include limiting the amount of code that needed to be written by us, to minimize the amount of support time needed to manage user accounts and take advantage of web based federated identity management. We also wanted to provide single-sign-on to existing best in class applications. While many tools have web versions now we also wanted to be able to support legacy applications that are accessed via terminals. To achieve this without needing to store passwords we created a web page to allow users to upload SSH public keys, after first signing on using their chosen identity provider. Getting the initial version of the gateway running was done in one week, after which it already had the majority of services that we had planned to provide showing the power of the tools we have employed without much configuration. The ilifu OpenStack Dashboard, that controls the Infrastructure as a Service could components has been accessible for several years using the EGI Checkin service. With the new gateway all the web based components are accessible using Identity Providers published through the South African SAFIRE identity federation and the world wide EduGain identity federation. The base components of the gateway include KeyCloak for linking identities, SaToSa to access the identity providers, OpenOnDemand to provide links to the applications and Django to provide web pages to support applications that we did need to create. Existing best in class applications linked from the gateway include JupyterLab, the CARTA astronomy visualization system, the Globus File transfer tool, the OpenStack dashboard and SLURM job submission. Tools created using Django include the SSH public key management page, system usage pages and data product distribution pages. These tools are relevant to many different scientific and engineering workflows.

Title	DIRISA services: Overview of DIRISA's services
Author	Mr May Maile - Senior Tech/Eng/KA, CSIR
Abstract	Data Management Plan (DMP) is a formal document that outlines how a

researcher will handle data before, during and after the project is completed. Outlines the practices for collecting, organizing, backing up, and storing the data that will be generated. Fewer benefits of DMP are more visible, citable on research data, encourage data sharing and more collaboration. When done with data planning Data Deposit Tool (DDT) store data so that research can start collecting data and store the data on DDT. Furthermore, with DDT user get 100GB for free and storage increase based on meeting requirements.

Title	Deep learning for real-time traffic classification and resource allocation in community networks
Author	Dr. Josiah Chavula - Senior Lecturer, University of Cape Town
Abstract	Network traffic classification is crucial for ensuring quality of service engineering in community networks. However, traditional approaches struggle with encrypted traffic, highlighting the need for deep learning techniques. Nonetheless, the computational complexity of deep learning models limits their suitability for resource-constrained community networks. This study presents a resource-aware framework that leverages single-layer CNNs, multi-layer CNNs, and attention-based one-dimensional CNNs for real-time packet-based traffic classification in community networks. First, we explore the computational efficiency and accuracy of 2D-CNNs for traffic classification in community networks. Our findings reveal that 2D-CNNs achieve higher out-of-sample accuracy compared to traditional models under similar resource constraints. However, their slower prediction speed limits real-time applicability. To address this, we propose reducing the input size to improve prediction speed while maintaining accuracy. Next, we investigate the trade-off between prediction speed and accuracy in packet-based traffic classification tasks with limited computing resources. By comparing 1D-CNN, MLP, and SVM models under various hardware constraints typical in low-resource community networks, we identify a clear trade-off between prediction rate and attainable accuracy. MLP models achieve fast prediction in networks with middle-range CPUs, while 1D-CNN models outperform others for powerful CPUs. Finally, we implement a framework integrating single-layer, multi-layer, and attention-based one-dimensional CNNs for real-time resource allocation based on traffic prioritisation heuristics. These deep learning models are trained using community network. Results demonstrate that deep learning-enabled Software Defined Networks enhance network throughput and reduce packet loss in real-time, improving overall Quality of Service (QoS) in community networks. This study provides a case study of deep learning techniques for real-time traffic classification and resource allocation, addressing connectivity

challenges in low-income and rural areas while empowering communities to manage their network infrastructures effectively. The proposed framework holds potential for enhancing the performance and reliability of community networks, fostering digital inclusivity and improving user experiences.

Title	Experimental and Non-traditional Archives in South Africa
Author	Prof. Hussein Suleman - Head of Department and Professor, University of Cape Town
Abstract	Digital repositories are a class of software systems with services to store, retrieve, preserve and share data/information with a wide range of possible uses. These tools emerged as a generalisation of early digital libraries, as the services and architectures began to converge and stabilise and with the introduction of international standards for metadata and system interoperability. Digital repository systems are often instantiated as institutional repositories, electronic thesis and dissertation repositories, cultural heritage archives or Web archives, based on either locally installed digital repository toolkits or remote online software services. It can be argued that there is little functional difference in these different instantiations, and this is exploited within general-purpose digital archives that cross collection boundaries. However, some problems do not easily map to general-purpose solutions, and these will be explored in the context of recent experimental work. Firstly, in the digital humanities, complex datasets are being produced where the unique assembly of data and nature of user engagement are fundamentally part of the archive. End-user engagement begins with access but is rapidly progressing to original composition as a data-intensive research activity. Secondly, data archives with large datasets need a closer integration of the archiving and processing functions to optimise storage and network use in low-resource environments. Thirdly, the architecture of small data archives in low-resource environments needs special attention in terms of design decisions to avoid the risk of system obsolescence and data loss. These studies present a range of issues that current and future data archives need to consider as data-intensive research is adopted in diverse scenarios within resource-constrained environments.

Using Generative AI as a Data Engineer	
Title	
Author	Mr Nhlakanipho Kwazi Mthembu - Head Teacher for Data Engineering, Explore-AI Academy
Abstract	Generative modelling refers to the process of creating and training artificial intelligence (AI) models that can generate new content that resembles human-produced data. These models learn from existing datasets and are then able to generate novel, previously unseen samples that exhibit similar characteristics to the training data. As a data engineer or someone who uses Generative AI, it can offer powerful ways such as prompt engineering to improve the performance, efficiency, and scalability of data pipelines and architecture systems. Prompt engineering refers to the process of designing and refining effective prompts or inputs to Generative AI models in order to elicit desired outputs or responses. A prompt can contain information like the instruction or question you are passing to the model and include details such as context, inputs or examples, which are used to instruct the model better. When working with prompts, you interact with the large language models (LLMs) via an API or directly. It's essential to ensure that the generated or synthetic data aligns with the desired characteristics and preserves the integrity and privacy of the original data. Through testing, validation, and monitoring should be performed throughout the data engineering process to ensure the reliability and quality of the generated outputs.

Data fluency for all through science engagement programmes	
Title	
Author	Ms Kogie Govender - Science Engagement Head, SAEON
Abstract	Experts and policymakers are frequently using the term data literacy and how this data would enable citizens to use the enormous amount of data at their disposal [1]. Data literacy is crucial to enable citizens to ensure that government is accountable and open in decision-making. Data-literate citizens would be able to solve problems that are localised and traverse their data ecosystems[1]. We cannot rely on data scientists to solve all the problems but there is a need to empower citizens to be included in the data literacy programmes to effectively under their own data and use this data confidently to solve problems and communicate their findings. Science engagement is crucial to address equity & inclusion in science [2]. NRF SAEON science engagement programmes seek to include beneficiaries from learners, educators, post-postgraduate students and citizens thereby making science accessible to all. The science engagement programmes provide 21st-century skills to beneficiaries through data literacy programmes to enhance the economic and social impact as well as the democratization of data.

The science engagement programmes include introducing learners to data in classrooms through weather stations and Argo float workshops as well as the SAEON Kids competition. The community engagement programmes enable communities to collect, analyse, visualise and interpret data through different citizen science projects. We are in an era where data literacy and the inclusion of citizens are imperative. To meet these challenges of data literacy and data democratization we need to re-think how we become innovative and work together as different communities of practice to ensure that all beneficiaries are equipped to embrace the data revolution and are data fluent.

Title	The World Data System: creating trusted communities of scientific data repositories
Author	Dr. Dale Peters - Independent Research Infrastructure Consultant
Abstract	The World Data System (WDS) is an affiliated body of the International Science Council (ISC) that aims to enhance the capabilities, impact and sustainability of member data repositories and data services. WDS facilitates scientific research under the ISC umbrella by coordinating and supporting trustworthy scientific data services, both in the provision, use, and preservation of their datasets, and by strengthening their links with the research community. To fulfil its remit, WDS builds worldwide ‘communities of excellence’ comprising of certified scientific data services - its Member Organizations—to form a distributed, interoperable data system. Advocating for accessible data and transparent and reproducible science, the Core Trust Seal is the only certification accepted by WDS to qualify for membership. How then, can the WDS objectives of Open Science be achieved when the cost of entry to such communities of excellence is prohibitive for most research institutions in the Global South? The research community itself is key to the acceptance of Open Science as an equitable form of data sharing. A recent conference entitled “Open Science – From Policy to Practice” highlighted the role of data literacy in shaping, implementing, and embedding Open Science. The aim of that conference was to share experience and to discuss the contribution of Open Science to strengthening sustainable futures and democracies. A notable outcome was the success identified in appealing to a higher sense of purpose as a fundamental tool of scientific data repositories that adopt the role of data literacy training, in addition to providing infrastructure and services. The dual role identified for repositories is closely aligned to the support actions facilitated by the WDS, offering significant benefits for capacity building. This presentation will examine the avenues open to African data repositories - and those that need further exploration - to enable participation in the

global data system, lending critical voice to the value of democratising data insights.

Title	Exploring the evolving landscape of research data repositories and infrastructure: emerging trends and innovations
Author	Dr. Josiline Chigwada - Post Research Fellow, Unisa
Abstract	Research data repositories and infrastructure play an important role in supporting the management, preservation, and sharing of research data. The landscape of research data repositories and infrastructure is continuously evolving as a result of current trends and innovations in research data management. As the landscape of research data evolves, new trends and innovations continue to shape the way researchers store, access, and utilize their data. This study was done to explore the dynamic landscape of research data repositories and infrastructure, focusing on the latest emerging trends and innovations. The objectives are to identify and analyse the key trends, understand their impact on data management practices, and explore the opportunities and challenges they present to the research community. A comprehensive literature review on SCOPUS, the Web of Science, and Dimensions was done using the PRISMA principles, and content analysis was used to analyse the data. The emerging trends were the adoption of the FAIR data principles, the adoption of data management plans, the integration of data and publications, the integration of data citation and attribution practices, the development of institutional data repositories, the utilization of cloud computing and scalable infrastructure, long term data preservation, data federation and interoperability, collaboration and open science, and the advancement in data privacy and security measures. The challenges and opportunities are related to evolving research paradigms; data privacy; data security and cybersecurity threats; data preservation; data lifecycle management; data integration; data diversity and complexity; scalability; ethical considerations; data citation; artificial intelligence and machine learning; distributed and decentralized architectures; cloud computing; open science movement; and the integration of data analysis and visualization tools. The findings contribute to a deeper understanding of the emerging trends and innovations in the field of research data repositories, helping researchers, institutions, and stakeholders make informed decisions and shape effective strategies for managing and leveraging research data.

Title	A language to derandomize robotic device actions in swarm intelligence models
Author	Dr. Colin Chibaya - Senior Lecturer, Sol Plaatje University
Abstract	Swarm intelligence involves generating collective behaviour in self-organizing systems, drawing inspiration from social groups found in nature, such as ant colonies. It refers to the ability of a group of simple, autonomous robotic devices to exhibit complex and intelligent behaviour as a cohesive unit, even without individual knowledge of the desired outcome. The robotic devices in the swarm rely on local interactions with each other and their environment, using rules, instructions, conditions, and constraints to trigger emergent behaviour at the group level. Rather than relying on centralized controls such as datasets, the proposed approach emphasizes communication, cooperation, and coordination among the robotic devices to achieve desired outcomes. Achievement of near-optimal solutions in the absence of complete datasets is an NP-problem. This study aims to capture the representation of the knowledge of robotic devices deployed in scenarios with incomplete data through a language that guides the actions of these robotic devices, resulting in derandomized behaviour and emergent behaviour. The proposed semantics are designed to align with swarm intelligence principles, with flexibility to choose any natural social group as an inspiration. The parameters to consider include memory requirements, internal states of the robotic devices, and the sequencing of instructions for each internal state.

Title	AI Policy in Africa, get involved. The role of grassroots AI in charting a joint future
Author	Dr. Vukosi Marivate - Associate Professor, University of Pretoria
Abstract	The transformative potential of Artificial Intelligence (AI) across global societies and economies cannot be understated. Africa, in particular, has recognized the significance of leveraging AI to tackle multifaceted challenges. However, this pursuit necessitates urgent attention to the development of comprehensive policy frameworks. This presentation highlights the imperative nature of addressing crucial policy questions surrounding AI implementation in African countries, including privacy, security, and ethical considerations. By proactively addressing these concerns, we can unlock the full potential of AI on the continent and ensure its benefits are harnessed for sustainable development. Further, this presentation underscores the role played by grassroots AI organisations in fostering innovation and driving development, particularly in regions such as Africa where resources may be limited. Through an exploration of the work carried out by these organisations, we

will gain insights into their impactful initiatives.

Moreover, this presentation will delve into recommendations for building a robust AI research and development ecosystem that fosters collaboration across the public, private, and NGO sectors. The significance of stakeholder engagement with policymakers will also be emphasised, calling for a collective effort to ensure that AI initiatives in Africa align with national priorities and contribute to sustainable development outcomes. I will highlight a few projects at the Data Science for Social Impact research group at the University of Pretoria that cover the intersection of Artificial Intelligence and African Languages and how by working across institutions and countries we have moved the needle forward at a time people thought it was not possible.

Title	Forecasting Emerging Contaminants Removal in Urban Water Cycle: Artificial neural networks (ANN) and adaptive neuro-fuzzy Inference system (ANFIS) models
Presenting Author	Prof. Kapil Moothi - Associate Professor, University of Johannesburg
Abstract	Organic micropollutants such as pesticides, personal care products, and antibiotics contained at low concentrations in municipal wastewater effluent pose a threat to the aquatic environment and public health. In particular, the release of antibiotics into the environment is leading to anti-microbial resistance (AMR). Globally, it is estimated that currently, 700,000 deaths annually are due to antibiotic-resistant pathogens. Antibiotics have been detected in wastewater treatment plant effluent and surface and groundwaters, initiating AMR. AMR-related tuberculosis (TB) has also been broadly reported in clinical investigations, making South Africa have one of the highest recorded prevalence's of drug-resistant tuberculosis (DR-TB) patients worldwide. The general consumption of antibiotics for the treatment of TB and other infections in South Africa is relatively higher compared to other nations. It has been revealed that the percentage removal efficiency for TB-related antibiotics during wastewater treatment can range between 21% and 100%. As such, the reported poor performance of wastewater treatment plants around the country is of great concern that requires urgent intervention. This work aims to establish the loads of selected TB-related antibiotics reaching wastewater treatment plants, as well as model and predict the removal efficiency of these antibiotics using artificial neural network (ANN), and adaptive neuro-fuzzy inference system (ANFIS) models. The findings from this research could serve as a reference point for estimating the consumption of TB-related antibiotics within communities as well as guiding the decision-making related to the selection of appropriate

wastewater treatment technologies to ensure safety of the effluent for discharge into the environment and reuse.

Title	Multivariate Time Series Forecasting with Machine Learning in Python: A comparative study of statistical models and deep learning models
Presenting Author	Mr Mphiliseni Bongani Nzuza (online) - Lecturer, Durban University of Technology
Abstract	Many previous or current forecasting studies tend to overlook correlations and /or non-linear relationships that exist between multiple sequential data sets. Although some of the most popular time series models are capable of capturing long-term dependence of time series data, their disadvantage is the lack of enough attention towards the impact of associated attributes when there are multiple correlated variables competing in their evolution, a suitable tool is required which could handle the dependency between time series variables. In this study we examine the performance of Vector Autoregressive (VAR) model and Recurrent Neural Network (RNN) based Long Short-Term Memory (LSTM) model amongst others in forecasting Multivariate Time Series (MTS). Both these approaches show a good success in capturing long term trends and patterns as well as underlying dynamics of correlated time series attributes. Fine- tuning model parameters for a better forecasting accuracy is of course at the core of the study. Models also tend to outperform each other under different conditions such as model selection based on causality tests and adjustment/selection of key model parameters. Data transform is another key approach that improves model performance. We analysed time series data sourced from SARB on some of the key productivity indicators of various sectors of economy such as manufacturing, water & electricity, as well as mining production in South Africa. The data consist of quarterly values that cover the period 1994-Q1 to 2022-Q2. We trained the models in machine learning algorithms using Python. The empirical study on these real-world datasets demonstrate the efficacy of machine learning algorithms for both methods. The future projection of a state economic performance is essential for policy analysis and decision making by governments or central banks. The contribution of this study is accurate prediction of future economic status of a country in a machine learning based MTS model that learns from sequential economic data. The algorithms can also be applied in a wide spectrum of fields where sequential data are generated for scientific studies.

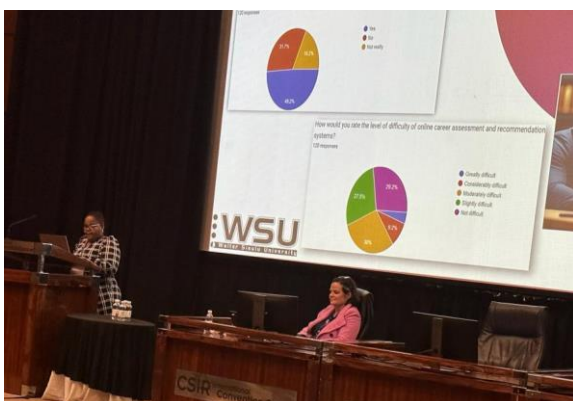
Title	SANReN; Digital Transformation; Strategy Execution Research & Innovation Cyber Infrastructure
Author	Mr Sabelo Dlamini (online); Director, SANReN, CSIR
Abstract	South African National Research Network (SANReN) is a high-speed network dedicated to science, research, education and innovation traffic. SANReN is a component of the comprehensive South African government strategy for cyberinfrastructure (NICIS), which aims to ensure that South African researchers can successfully contribute to the effort of global knowledge production. An overview of the network infrastructure, national and international capacity design, procurement, and rollout for the country's national research and education network, as well as the creation and incubation of innovative services being offered to the research community, are all covered in the presentation.

Title	Surrogate modelling of wakes in wind farms in wake steering conditions
Presenting Author	Mr Ayodimeji Biobaku - Student, University of the Witwatersrand
Abstract	Since there is an increasing demand worldwide for renewable energy and specifically wind energy, the need for optimization of wind farms has increased. In a wind farm, the upstream turbine has several negative effects on the downstream turbine which leads to reduced power outputs of the wind farm. The aim of this research is to create a surrogate model using an existing dataset generated by Large Eddy Simulations (LES) of a two-turbine array. The dataset was taken from an open-source site that had already processed the data. This processed dataset will be used in the format that was available. The wake flow generated by the redirection of an upstream turbine will be analysed. A reduced order model (ROM) will initially be created to allow for faster run times and use less computational resources without sacrificing significant accuracy. Dynamic Mode Decomposition (DMD) will be used to obtain more information about the wake flow, such as modal frequencies and growth or decay rate. This should allow for the necessary flow field dynamics of this system to be captured comprehensively. The results generated after applying the DMD method will be compared to the LES data. These results will be presented and observations on the ability of the DMD method to accurately predict the wake flow will be discussed. Once it has been established that the ROM produces an accurate representation of the wake, a machine learning algorithm will be applied. It is planned that the algorithm will be used to predict the wake of the turbine array when parameters such as the wake redirection angles are different from those in the LES dataset.

Title	Application of artificial intelligence towards career exploration and decision making
Presenting Author	Ms Nosipho Mavuso - Lecturer, Walter Sisulu University
Abstract	The utilisation of technology has revolutionized many aspects of our lives, including the way we make career decisions. With the arrival of the Internet and the precipitous development of various technologies, such as artificial intelligence (AI) students now have unprecedented access to information that can help them explore different career choices and make informed decisions about their future. The increasing trend towards a personalized online experience for users has led to the adaptation of Recommender Systems (RS). Regardless of the escalation in popularity and usage artificial intelligence for career guidance and personalized career recommendations is still limited in South Africa, and therefore more attention is needed in this regard. There are a few limiting factors to such as the ever-eminent challenge of limited access to the internet, language barriers, and cultural differences which are highly dominant in rural areas. It is on this premise that this study has been undertaken, whereby students from a rural university in the Eastern Cape province in South Africa were engaged to determine how AI can be utilised in their career selection process. The study highlighted the importance of developing personalised AI solutions that will reach a wider audience, predominantly in rural areas, and the importance of integrating cultural and contextual factors into the design of personalized career recommendation systems. The purpose of this study was to primarily examine how AI can be utilized in the career choice decision making process, through a recommender framework to facilitate a positive developmental experience for students during their career choice decision making process. The study collected quantitative data from students at a rural university in the Eastern Cape Province. Results from the study revealed that some learners have utilised online career guidance systems and they are satisfied with their current selected career stream, however, there is a high number of learners who still require some guidance and exposure in various career fields. The study recommended the need for a more personalised locally produced AI career assessment system. This should consider critical factors such as background and culture that will be relevant for South Africa. An AI solution to enable career choice is critical for rural learners and improves decision-making by the targeted users.

Title	DIRISA Services
Presenting Authors	Ntlharhi Baloyi & Hlawulani Sibisi - Tech/Eng/KAs, CSIR
Abstract	DShare is an open-access data repository service developed by the Dirisa. Its primary goal is to provide researchers and data managers with a platform to store, share, and manage research data. Dshare focuses on facilitating the sharing of research data to increase data accessibility, visibility, and reuse, supporting the principles of open science. Researchers can deposit their research data into Dshare, which assigns a persistent identifier (DOI or Digital Object Identifier) to the datasets, ensuring they can be uniquely identified and cited. This makes it easier for other researchers to find and reference the data, enhancing the transparency and reproducibility of research.

Pictures of the event





Conclusion

DIRISA would like to thank everyone who participated in this annual event. Over the past few years, the national research data workshop has become a melting pot for opportunities and discussions of topical interest amongst delegates of different backgrounds and professional interests. We are grateful to you for sharing information that contributed immensely to the human capital development of active and productive members of our society.

We are looking forward to seeing you at our next event!
