DIRISA 2024 Annual National Research Data Workshop

Tuesday 02 July 2024 - Wednesday 03 July 2024

CSIR ICC

Book of Abstracts

Contents

Session / 4

A Robust Data Visualisation Technique for Data-Driven Decisions: Illustrations from Power Demand and Supply

Author: Kassim Mwitondi[1]

[1] *Sheffield Hallam University*

Corresponding Author: k.mwitondi@shu.ac.uk

Ongoing technological advances in computing, data acquisition and the complex interactions of Sustainable Development Goals (SDG) create a natural Big Data environment for researchers and decision makers across fields and sectors to tap into. Identifying the triggers of SDG targets under such conditions is non-rivial, not only because of the large data dimensionality and non-orthogonality nature of the SDGs, but also due to the naturally arising data and information related gaps between data analysts and policy makers. We propose a cohesive data visualisation approach to bridging such gaps. The approach's main idea derives from a cohesion between technical and non-technical data generators and consumers. It is designed to provide a visual conduit between the two parties, hence facilitating unified understanding of the role and impact of the visualised data attributes. Visualisation of SDG-related data provides a natural interdisciplinary setting for stakeholders to gain actionable insights into important patterns across the SDG spectrum. Identification of relevant data attributes and the nature of their complex interactions are fundamental to creating robust data-driven solutions. Thus, given real-time access to the visual effects of a single or set of SDGs, decision-makers can quickly grasp the significance of key attributes and make informed strategic choices before it is too late. Most importantly, the cohesive approach potentially leads to a unified understanding of the SDG project across the globe, regions and within countries. Communicating information embedded into data attributes via interactive data visualisation is pivotal in optimising operational efficiency. It enables timely identification of bottlenecks, tracking performance metrics and making timely interventions. We illustrate the approach based on a large time-series dataset obtained from the South African utility giant, ESKOM https://www.eskom.co.za/dataportal/, covering the period 01 April 2020 to 31 March 2024. The choice is motivated by ESKOM's quest for stabilisation of the national electricity grid by balancing supply with the demand for electricity which can realistically be validated through visualisation and assessing demand forecasts. Visual patterns and forecasts show areas of attention, potential associations with other aspects of SDGs and highlight paths to a unified understanding of the triggers of SDG indicators,between data technocrats and decision makers, and open new paths to interdisciplinary research.

Session / 3

AI model for securing Internet of Things communication systems in smart agriculture.

Authors: Issah Ngomane[1] ; Mncedisi Bembe[1] ; Mthulisi Velempini[2]

[1] *University of Mpumalanga*

[2] *University of Limpopo*

Corresponding Authors: issah.ngomane@ump.ac.za, mthulisi.velempini@ul.ac.za, mncedisi.bembe@ump.ac.za

The rapid increase of Internet of Things (IoT) devices in smart agriculture has enabled a more connected and intelligent world. IoT devices are a collection of interconnected systems that can communicate, share data and information to achieve an automated environment. Smart agriculture presents a transformative approach to farming that leverages technology and data-driven solutions to address the challenges of modern agriculture, including the need to sustain a growing global population while minimising environmental impact and resource depletion. However, the increase in the deployment of IoT systems has led to an increase in cyber-attacks and security challenges. Moreover, security challenges such as man-in-the-middle, denial and distributed denial of service, botnets, sinkhole and spoofing attacks compromise the confidentiality, integrity and availability (CIA) of smart

agriculture. This study investigates measures deployed for anomaly detection and prevention in IoT smart agriculture communication systems. Furthermore, the study proposes a model that incorporates machine learning techniques to identify and predict anomalies in loT communication systems and adapt security measures dynamically. Python is used to develop the proposed model and tested on accuracy, recall, f1-score, precision, true positive rate, false positive rate metrics. The IoT-based Datasets CIC-IDS2018, ToN-IoT and Edge-IIoTset are used to evaluate the performance and efficiency of the proposed model.

Session / 5

AcousNomaly: Generating Aquatic Acoustic Telemetry Data from Real Data while Learning to Detect Anomalies with Unsupervised Learning

Authors: Siphendulwe Zaza[1] ; Marcellin Atemkeng[1] ; Taryn Murray[2]

[1] *Rhodes University*

[2] *Rhodes University*

Corresponding Author: zazasiphendulwe@gmail.com

This research project aims to investigate the movement behavioural patterns of dusky kob (Argyrosomus japonicus) from the Breede Estuary using acoustic telemetry data. A group of 50 dusky kob were surgically fitted with unique coded acoustic transmitters, and their movements were monitored using an array of 16 acoustic receivers deployed throughout the estuary between 2016 and 2021, resulting in more than 3 million individual data points. Acoustic telemetry data plays a vital role in understanding the behaviour and movement of aquatic animals. However, these datasets often contain anomalous detections that can pose challenges in data analysis and interpretation. Anomalies in acoustic telemetry data can occur due to various factors such as biological factors, environmental factors, and technological limitations. Given that ichthyologists currently rely on manual detection to identify fish with anomalous behaviour or movement, this study focuses on automating the process of anomaly detection in telemetry datasets using machine learning (ML) and artificial intelligence (AI) models. The objective is to address the time-consuming nature of manual anomaly detection, especially in large datasets such as the one considered in this research, which consists of over 3 million instances. The proposed approach combines the use of LSTM (Long Short-Term Memory) models and autoencoders to construct an efficient anomaly detection system.

The model has undergone fine-tuning and preliminary work indicates that it is proficient at learning the normal patterns within the data, effectively distinguishing between normal and anomalous behavior. However, it may encounter challenges in accurately detecting anomalies, particularly in cases where they deviate slowly from the expected patterns. Despite this limitation, the model demonstrates promising capabilities by pinpointing the precise locations of anomalous entries within the dataset. Further investigation, including refinement and optimization of the model's parameters and training process, especially with memory-based LSTM-AE may enhance its ability to detect anomalies with greater accuracy and reliability. Our proposed LSTM-AE model showed a 98% score in all four evaluation metrics including accuracy, precision, recall, and F1, based on the preliminary test results.

Session / 30

Applications of Artificial Intelligence and Machine Learning on the Diagnosis and Management of Sexually Transmitted Infections among Key Populations in Sub-Saharan Africa: A Bibliometric Analysis

Author: Claris Siyamayambo[1]

[1] *University of Johannesburg*

Corresponding Author: clariss@uj.ac.za

Introduction: Sexually transmitted infections (STIs) pose a significant public health challenge in Sub-Saharan Africa, particularly among key populations who are gay men, men who have sex with men, female sex workers, transgender persons, people who use drugs, and incarcerated persons. Artificial intelligence (AI) and machine learning (ML) offer promising tools for enhancing the diagnosis and management of STIs. This study aims to explore the application of AI and ML in this context through a comprehensive bibliometric analysis. The primary aim of this study is to evaluate the scope and impact of AI and ML applications in the diagnosis and management of STIs among key populations in Sub-Saharan Africa. This will help to identify the volume and trends of research publications on AI and ML applications in STI management in Sub-Saharan Africa.
Methodology: A bibliometric analysis is being conducted using databases like Web of Science to retrieve relevant publications from 2010 to 2023. Keywords included "artificial intelligence," "machine learning," "sexually transmitted infections," "sexually transmitted diseases", "key populations," and "Sub-Saharan Africa." Data will be analyzed using bibliometric software tools to extract metrics such as publication trends, citation analysis, and co-authorship networks.
Preliminary Results: The analysis is expected to identify a significant increase in publications over the study period, from different countries in Sub-Saharan Africa. Scholars from universities and health organizations are expected to collaborate at local and international levels. The study will show the most frequently cited studies focused on the uses of AI/ML in the diagnosis and management of STIs in Sub-Saharan Africa.
Conclusion: The application of AI and ML in the diagnosis and management of STIs among key populations in Sub-Saharan Africa is an emerging field with growing research interest. The findings will highlight the potential of these technologies to enhance public health approaches to STI management in Sub-Saharan. The study will underscore the significance of leveraging AI and ML to enhance STI management among KPs in Sub-Saharan Africa. Artificial intelligence and ML contribute to better health outcomes and the reduction of STI prevalence in the region.

Session / 34

Big Data Analytics for Intelligent Temperature Prediction in CNC Machining using PCA and AI

Authors: ZVIKOMBORERO HWEJU[1] ; Varaidzo Dandira-Chibaya[2]

[1] *Lecturer, Chinhoyi University of Technology, Zimbabwe*

[2] *Lecturer, Chinhoyi University of Technology*

Corresponding Authors: varaidzosdandira@gmail.com, hwejuzvikomborero@gmail.com

Abstract. Temperature prediction is crucial in CNC machining to prevent overheating, tool damage, and surface finish quality. This study presents a big data analytics framework for intelligent temperature prediction in CNC machining using Principal Component Analysis (PCA) and Artificial Intelligence (AI). The following methodology has been followed during the CNC machining of EN18 Steel: data collection from laboratory CNC machining, normalization of the collected data to ensure consistency and comparability, application of PCA to the preprocessed data to reduce dimensionality, training of AI models (ANN, ANFIS and Random Forest) using PCA-extracted features and temperature data, performance evaluation of the AI models using mean absolute error and coefficient of determination, utilization of the trained AI model to predict temperature values for new, unseen data, comparison of the AI model results to those of the traditional Linear Regression Model. The proposed approach predicts temperature with high accuracy of above 95%. The results show improved prediction performance compared to the traditional linear regression method, demonstrating the effectiveness of intelligent big data analytics and AI in CNC machining. This research contributes to the development of Industry 4.0 technologies, enhancing manufacturing efficiency, productivity, and product quality.
Keywords: Big Data analytics, Principal Component Analysis (PCA), Artificial Intelligence (AI), Temperature Prediction.

Session / 24

Big data and machine learning skills, experience, methods, and big data usage. Empirical evidence from manufacturing firms in Zimbabwe.

Authors: Sostina Varaidzo Chibaya[1] ; More Chinakidzwa[2]

[1] *Liaoning Technical University, China*

[2] *Higher Colleges of Technology, UAE.*

Corresponding Author: varaidzosdandira@gmail.com

The use of big data and machine learning in decision making processes in manufacturing industries is gaining momentum as manufacturers seek to enhance production performance and competitiveness. Big data technologies have transformed manufacturing decision-making, resulting in data-driven approaches. To compete in today's dynamic market, manufacturing organizations must adapt and evolve, which necessitates the effective use of data for forecasting future events and making decisions. Advanced analytics applied to large datasets allows firms to acquire deeper insights, spot patterns, forecast future trends, and optimize processes. Limitations of conventional data processing methods create new opportunities for development and innovation. However, in developing countries such as Zimbabwe the benefits of the use of big data are not fully realized in manufacturing industries. Three essential requirements are needed to effectively use big data. These are big data skills, experience using big data, and effective data processing methods. The study focused on assessing how the three factors influence the utilization of big data in the manufacturing industry. Understanding data skills, experience and processing methods is essential in building big data management skills and improving adoption in manufacturing firms. A preliminary survey was conducted on 36 manufacturing companies in Zimbabwe. The data was analyzed using SPSS version 27. The results showed that only 16.7% of the companies were effectively using big data. The effectiveness of data processing methods, big data skills, and experience using big data, significantly affect the utilization of big data in manufacturing ($p<0.05$). In addition, the effectiveness of data processing methods has a positive impact on the quality of decisions and accuracy prediction level. The study therefore recommends manufacturers to train their employees on the required skills and upgrade data processing methods. In the future, to address limitations of this study, there is need to widen the sample size to produce widely generalizable results.
Keywords: big data, machine learning, data processing, data experience.

Session / 10

Bridging the Divide: AI-tools Empowers Data-Driven Research in Africa

Author: Lungile Nkosi[1]

Co-author: Israel Agaku [2]

[1] *1. Chisquares Inc., 2. School of Health Systems and Public Health, University of Pretoria*

[2] *1. Chisquares Inc. 2. School of Public Health and Health Systems, University of Pretoria*

Corresponding Authors: iagaku@chisquares.com, nkosi.lungile@gmail.com

Background

The growing volume, complexity, and diversity of data in Africa present significant challenges and opportunities for research. Limited access to open-access datasets, labor-intensive data analysis methods, and inefficient publication processes hinder research progress, especially in public health—a critical area for improving lives on the continent. AI-powered tools are poised to transform data-intensive research within the African context.

Methods

This presentation explores the functionalities of AI-powered tools and their potential to revolutionize epidemiologic and broader data-driven research in Africa. We examine how these tools leverage big data and machine learning technologies to enhance research capabilities and outcomes, with particular focus on the Chisquares platform as an example.

Results

AI-powered tools provide a robust solution for improving data accessibility, research efficiency, and collaboration in Africa through the following features:

Open Data & Security: AI-powered tools democratize data by offering secure access to a centralized repository of large-scale, nationally representative surveys from across Africa and beyond. This approach supports open data practices while ensuring data integrity, standardization, and secure storage in a cloud-based environment. Advanced security measures, including encryption, access controls, and compliance with data protection regulations, are integral to these tools.

Big Data & Machine Learning: Utilizing AI and sophisticated algorithms, AI-powered tools automate over 80% of tasks involved in manuscript preparation, including data analysis and writing. This automation allows researchers to focus more on scientific inquiry and knowledge generation.

Data Management: AI-powered tools streamline the entire research workflow by integrating essential functionalities for all stages of the scientific process, from sample size calculation and sampling to data collection, analysis, writing, and journal submissions. This comprehensive approach eliminates the need to switch between multiple applications, enhancing efficiency and productivity in data management.

Conclusion

AI-powered tools foster data literacy, promote open data practices, and ensure secure research environments, empowering researchers across Africa to generate impactful insights and contribute to improved health outcomes. By promoting efficient data management and leveraging big data and machine learning, these tools stand at the forefront of revolutionizing research data repositories and services for the future.

Session / 21

Building data literacy and data prioritization in business.

Author: More Chinakidzwa[1]

[1] *Higher Colleges of Technology*

Corresponding Author: more.chinaz22@gmail.com

Data literacy, and prioritization are essential tools for business decision-making in the digital era. Data, among other benefits, promotes innovation, competitiveness, collaborations, and customer service and enables better decision-making without relying on intuition. Data literacy is the ability to read, analyze, use, and communicate with data. A quality and well-analyzed dataset in the hands of executives who do not understand it is a wasted resource. Businesses require data literacy to develop the appropriate mindset and data culture, manage datasets as a business asset, and design data-driven organizations. Further, data literacy drives data-driven critical thinking, learning, decisions, and the realization of value from data. Technology investments, tools, and processes without data-literate employees do not yield the benefits that businesses seek. However, despite the continued commitment and growth of data reliance in business, many organizations struggle to deliver value from their data. Some executives still rely more on experience and intuition than data due to prioritization deficiency. Prioritization entails knowing which data has the highest value and readiness and needs to be published first to support business functions. This paper provides the nexus between data literacy and prioritization by demonstrating the value of data, literacy, flow, and prioritization in business. The paper further provides suggestions to enhance this linkage. The paper adopted a literature search strategy and suggested three steps to achieve data literacy. Businesses

must 1) develop sound data strategies grounded in a deep understanding of current and future literacy and business requirements. 2) Develop data literacy programs that include data literacy measurement metrics. Finally, commit resources. Resources commitment entail building strategies around leadership, a data-based decision culture, adopting the right skills-technology mix, and continuous learning. This paper contributes to understanding the value of data literacy and prioritization in business. It provides insights, guidance, and experiences from different contexts that are valuable to the development of data literacy. However, further research must be conducted to provide empirical evidence of data literacy and prioritization in African businesses.

Session / **17**

Context-Based Question Answering using Large Language BERT Variant Models for Low Resourced Sotho sa Leboa Language.

Author: Mosima Annan Masethe[1]

Co-author: Hlaudi Masethe [2]

[1] *Sefako Makgatho Health Science University*

[2] *Tshwane University of Technology*

Corresponding Authors: masethehd@tut.ac.za, mosima.masethe@smu.ac.za

Since reading and responding to text needs both a grasp of natural language and awareness of the outside world, it is challenging for machines to do (Akhila et al., 2023). The most difficult areas of information retrieval and natural language processing are question answering systems (QAS). The goal of the Question Answering System is to use the provided context or knowledge base to provide replies in natural language to the user's questions. Both closed and open domains can produce the answers. A closed domain's responses are limited to a specific situation, whereas open-domain systems are able to provide answers in a human-readable language from a vast knowledge base. Another issue is coming up with answers to the questions based on certain situations, as each question might have a variety of interpretations and responses based on the context to which it relates (Kumari et al., 2022). In our comprehension, this research work is the initial effort to extract answers from a context in low resourced Sesotho sa Leboa language. The Bidirectional
Encoder Representation from Transformers (BERT) variant model such as Albert, and DistilBERT is used as the language model in this research study

Session / **8**

DOCiD (Digital Object Container iD) by the Africa PID Alliance

Author: Nabil Ksibi[1]

[1] *Africa PID Alliance*

Corresponding Author: ksibi.aziz@gmail.com

The Africa PID Alliance's overall objective is to produce DOIs in Africa through its open infrastructure and the integration of different identifiers to disseminate indigenous knowledge , cultural heritage and patent data.

The DOI being produced by the Africa PID Alliance is called the Digital Object Container Identifier (DOCiD TM) which is multilinear in nature and can accommodate different other identifier types and connect to the object being assigned.

To achieve this goal The Africa PID Alliance intends to be a global open infrastructure provider, where partnership conversations have began with the DOI Foundation membership.

In this participation to the honourable DIRISA 2024 Annual National Research Data workshop, we intend to present the latest updates about our DOCiD and seek further collaboration and partnerships to build up on what we achieved so far.

Session / 2

Data Management in National Assessments: A Look towards the Future

Authors: Lucy Tambudzai Chamba[1] ; More Chinakidzwa[2]

[1] *Durban University of Technology*

[2] *AI in Women's College, Higher Colleges of Technology,*

Corresponding Authors: more.chinaz22@gmail.com, lucyteechamba@gmail.com

National assessments play a critical role in evaluating educational systems and informing policy decisions. However, the effectiveness of these assessments hinges on robust data management practices. This article delves into the evolving landscape of data management in national assessments, exploring how research data repositories and services can contribute to a more secure, accessible, and sustainable future.
The paper highlights key challenges in national assessment data management, including data security, long-term preservation, and facilitating data sharing for research and improvement. It will then showcase how research data repositories can address these challenges by providing secure storage, standardized formats, and access controls. Additionally, the article will explore how advancements in data services, such as data cleaning, harmonization, and analysis tools, can further empower researchers and policymakers to leverage national assessment data effectively. By fostering collaboration between assessment developers, researchers, and data repositories, the research ensures that national assessment data becomes a valuable resource for generations to come. This article contributes to literature on data management by exploring how research data infrastructure plays a vital role in propelling national assessments towards a data-driven future.

Session / 33

Data mining, management and modelling in advancing water and sanitation systems

Author: Ridhwaan Suliman[1]

Co-authors: Prettier Morongoa Maleka [2] ; David Tshwane [3]

[1] *CSIR*

[2] *CSIR-Next-Gen Enterprises and Intuitions*

[3] *CSIR/University of Limpopo*

Corresponding Authors: pmaleka@csir.co.za, rsuliman@csir.co.za, dmtshwane302@gmail.com

Data mining and management play an important role in advancing water and sanitation systems, ensuring the sustainable delivery of essential services. The application of data mining encompasses the collection, processing, and analysis of vast datasets derived from various sources such as sensor networks, satellite imagery, and public health records. These techniques facilitate the identification of patterns, trends, and anomalies, which are important for informed decision-making and strategic planning.

By leveraging predictive analytics, water management authorities can anticipate demand fluctuations, optimise resource allocation, and enhance the efficiency of distribution networks. Similarly, in sanitation, data mining assists in monitoring system performance, detecting potential failures, and mitigating health risks by providing early warnings of contamination events. Moreover, the adoption of robust data management frameworks ensures the integration, storage, and accessibility of diverse datasets, supporting real-time monitoring and long-term strategic initiatives. Challenges such as data privacy, accuracy, and the need for interdisciplinary collaboration need to be addressed to ensure the reliability and efficacy of these systems. The convergence of data mining and management in water and sanitation sectors holds significant promise for enhancing operational efficiency, ensuring resource sustainability, and safeguarding public health.

Data integration poses a significant hurdle due to varying formats and structures across different sources. The main challenge is data quality and accuracy with issues like missing values and outliers. Water databases may contain diverse types of data, including spatial, temporal, and multidimensional information. Integrating and reconciling these different types of data can be challenging, especially when they come from various sources with distinct formats and structures.

Machine learning is a rapidly expanding field of computer science with diverse applications. Understanding seasonal rainfall changes is crucial for both academic and societal objectives. Current data on surface and groundwater, including water quality and quantity was reviewed. Collecting real-time data, using automated sensors, and integrating remote sensing technologies helped understand water quality dynamics. Data from the Geographical Information System (GIS) was collected to better understand the spatial distribution of water quality and quantity factors. Integrating GIS data enhances our understanding of water resources.

After collection, data was cleaned for accuracy and dependability. After analysing the water datasets, it was found that the random forest (RF) method outperforms all the water quality classification models tested in this project, with a good combination of precision and recall across both classes. Although support vector machines are good at identifying negative classes, they struggle with positive ones. Linear regression, also known as logistic regression, has limits when separating water quality groups. Although decision tree models have balanced performance, there is still potential for development. When estimating water volume, both linear regression and RF models do moderately well, but the latter struggles to capture the underlying patterns. A negative R2 score for the RF model implies a lack of substantial predictive potential, necessitating additional research or evaluation of alternative models.

Session / **20**

Enhancing Climate Services in South Africa

Author: DAWN MAHLOBO[None]

Corresponding Author: ddmahlobo@gmail.com

Climate Services plays a critical role in the country. It is a pertinent factor in the decision-making at various levels for almost all sectors and communities in South Africa. Unfortunately, people who need it most cannot always obtain existing climate information or find it inaccessible when it does. Changes in climate, both human-caused and natural, have a major impact on society, affecting areas such as the economy, water and food security, and overall health and well-being. South Africa has experienced noticeable changes, including rising average temperatures, increased frequency of extreme heat events, prolonged droughts, and intensified floods, all of which underscore the urgency of addressing climate-related challenges. Collaborative efforts by different climate service providers through the National Framework for Climate Service (NFCS) will have to be strengthened to render climate service across all users. The NFCS aims to develop a practical model that acknowledges the significance of emerging trends in producing climate service data that values consultation and involvement of climate users and the vital role users play in collaborating on climate service information. The ideas of collaborating in production and exploration are acknowledged as essential for the effective utilization of climate data in decision-making. This paper offers an overview of the current status of the NFCS implementation
within South Africa.

Session / 9

Exploring Technical Challenges in Data Science Fundamentals with Python for First-year Students at a Rural South African University

Authors: Sithandiwe Twetwa-Dube[1] ; Sibukele Gumbo[1]

[1] *Walter Sisulu University*

Corresponding Authors: sgumbo@wsu.ac.za, stwetwa@wsu.ac.za

Abstract. Most Eastern Cape rural schools operate in a disadvantaged context and are struggling to raise standards. In many rural areas, access to quality education and access to devices or other resources in information technology and computer science may be limited, making such an initiative particularly impactful.
The current state of high schools has a huge impact on students going to universities. As an example. majority of first-year Information Technology Diploma students in the selected University, come from rural schools in the Eastern Cape (EC). Introducing data science with Python to first-year students at the selected university presents an exceptional opportunity to equip students with essential skills for the digital age while addressing the specific challenges of the local context.
In order to extract insights from data, data science is an interdisciplinary field that integrates statistical analysis, machine learning, and domain expertise. It is becoming more and more significant in a variety of global sectors. The university can help first-year students develop the critical thinking and problem-solving abilities necessary for success in the twenty-first century, as well as prepare them for future employment. Python, a useful and beginner-friendly programming language, is complementary for introducing data science concepts to beginner students. It an ideal choice for introductory courses. Moreover, Python's popularity in both industry and academia ensures that students will acquire skills relevant to their future careers. Practical, hands-on exercises can be incorporated into the course to address the unique requirements and challenges faced by rural students. To guarantee that every student has an equal chance to succeed, the institution can also offer support services like mentoring, tutoring, and access to computer labs and internet resources.

Introducing data science as supplementary content to these first-year students is a challenge for disadvantaged students without this course background and access to devices, resulting in confusion, anxiety and frustration.
Many of the students entering the university lack basic computer and digital skills and have no access to devices, in addition to the English language as a medium of instruction used in programming. The paper focuses on some of the best approaches and support tools as well as resources for assisting disadvantaged students, and we reflect on how they have worked out for any given computer programming problem-solving task.

Keywords: Data Science, Programming, Digital Skills, Information Technology

Session / 37

Health Records Data Management Role of Security

Authors: Nosipho Mavuso[1] ; A. Bantwini [None]

[1] *Walter Sisulu University*

Corresponding Author: nmavuso@wsu.ac.za

In today's digital era, health records management has evolved from traditional paper-based systems to advanced electronic health records (EHR) systems. This progress has ushered in several advantages, including improved accessibility, enhanced precision, and heightened efficiency in healthcare

delivery. Nevertheless, it has also introduced substantial risks related to data security due to the sensitive nature of health records, which contain personal, medical, and financial information, making them a prime target for cyberattacks.

This study aims to investigate various techniques and practises required to improve data management and security for online patient record management. Health records are of paramount significance in ensuring seamless and well-coordinated care by providing comprehensive accounts of patients' medical histories, including their conditions, treatments, medications, and other personal information. The unauthorized access to this data can lead to severe consequences, such as identity theft, financial fraud, and invasion of privacy, which can have severe personal and professional repercussions.

Quantitative data was gathered through questionnaires with healthcare professionals, patients, and IT professionals to uncover nuanced perspectives, challenges faced, and lessons learned from their experiences. Participants form the Eastern Cape Province were selected. The researcher used purposive sampling whereby participants were randomly selected particularly from the healthcare professionals, IT specialists, patients, and administrators. Healthcare professionals, patients and IT administrators who agreed to take part were sent surveys by email and or WhatsApp to fill out information regarding their understanding. The research scope was limited to the healthcare sector in the Eastern Cape,

Findings have shown that the management of health records in the digital age offers immense benefits for healthcare delivery but also brings substantial security challenges. Ensuring the security of these records is critical to protecting patients' privacy, maintaining trust, and ensuring the quality of care. It was also evident from the results that healthcare organisations generate vast and store vast amounts of data, including patients' records, medical images, and research data. Traditional on-premises infrastructure may not provide sufficient storage capacity or scalability to handle this growing volume of data effectively. Finally, the study concluded that emerging technologies such as artificial intelligence (AI), machine learning, and blockchain have the potential to further enhance the management and security of health records.

Key words: Health Records Management, Electronic Health, Data Security.

Session / **28**

How do we measure the impact of big data in society?

How do we measure the impact of big data in
society?
by
DSI Deputy Director, Cyber Infrastructure
Mr Daniel Mokhohlane

Session / **16**

Improving data availability and accessibility for research repositories

Author: Zibuyisile Magubane[1]

Co-authors: Lindelweyizizwe Manqele [1] ; Freedom Mthobisi Khubisa [1]

[1] *Durban University of Technology*

Corresponding Authors: lindelweyizizwem@dut.ac.za, freedomk@dut.ac.za, zibuyisilem@dut.ac.za

Data management is organizing, describing, sharing, and preserving the research data in research repositories. Data management plays an important role in data dissemination, and reuse of data for future purposes. The data is shared to broaden the impact and visibility of the research, encourage collaboration between data users and data creators, and allow others to replicate or validate your results thereby improving scientific integrity, to provide credits as a research output in its own right. Data management practices are integral to the entire research lifecycle, from planning for what kind of data you will collect to depositing your data set in a repository. Due to poor data management, challenges such as the finding, accessibility, availability, sharing, and re-use of data remains crucial in data repositories. Therefore, there is a need for an effective data management practices in the research repositories that will enhance data sharing. This paper, therefore, focuses on methods that intend to make research repositories available and accessible in order to improve data sharing quality. Machine learning (ML), as a promising emerging technology, will therefore be utilized in this study to improve the availability and the accessibility of research repositories sharing. These ML resources sharing algorithms will increase the visibility and accessibility of the research to a broader audience instead of sharing data with constrained audience. The ML algorithms will further be evaluated to increase the audience of researchers accessing and sharing opinions on research repositories. Repositories also manage and maintain data ensuring long-term access and additional demands on the audience.
Keywords: Data management, research repositories, resource sharing, accessibility and availability.

Session / 14

Institutional repositories as tools to enhance research visibility and leverage SDGs

Authors: Rosina Ramokgola[1] ; Tlou Mathiba[2]

[1] *Rosina*

[2] *University of Pretoria*

Corresponding Authors: tlou.mathiba@up.ac.za, rosina.ramokgola@up.ac.za

Institutional repositories are digital platforms in which institutions archive, manage and disseminate scholarly outputs within and beyond. The repositories leverage the knowledge to advance research and development (R&D), and innovation. The 2015 National Research Foundation (NRF) statement on Open Access states that, "the data supporting the publication should be deposited in a trusted open access repository, with the provision of a Digital Object Identifier (DOI) for future citation and referencing." In supporting the NRF open access mandate, the South African Higher Education Institutions (HEIs) and research institutions implemented the repositories such as open source and/or proprietary. Academic libraries as knowledge hubs are well positioned to ensure discoverability and retrieval of information, scholarly outputs and sustainable development goals (SDGs). Researchers need to expose their work by publishing in a repository and use appropriate tools to enhance research visibility and impact for wider audience. When the research is discoverable, it has societal impact such as advancing knowledge, informing policy, or solving real-world problems. Therefore, the visibility makes libraries one of the key partners and contributors in achieving United Nations SDGs.
Keywords: Institutional repositories, Research visibility, Sustainable development goals (SDGs);

Session / 22

Investigating the Prospects of Blockchain Technology in an Electoral System

Author: Taurai Hungwe[1]

[1] *Department of Computer Science and Information Technology, Sefako Makgatho Health Sciences University, Ga-Rankuwa, Pretoria, South Africa*

Corresponding Author: agmuofhe@gmail.com

Voting is part of our fundamental rights. It is crucial to have open and fair elections. As part of taking part in the election process, those voting need to have the utmost trust and confidence in the election processes and its outcome. When the underlying trust, confidence, and respect for the election is eroded by news of vote manipulation and disregard of processes by the officials overseeing the election, the voters are less interested in voting as they believe the elections are rigged and serves no purpose to participate in such an important exercise. The current voting systems which are often inundated by problems such as voter fraud, ballot tampering, and lack of transparency, necessitate a robust, secure, and transparent alternative. This research attempts to address some of the issues relating to having an unsecured voting system. The focus of the research is to investigate the prospects of implementing a blockchain based electoral system. Implementation of the blockchain technology, a decentralized ledger technology which is characterized by its immutability, transparency, and security, might address some of the security challenges with the current electoral systems. Furthermore, the study examines the principles of blockchain technology and evaluates its application in various stages of the electoral processes.

Session / 15

Machine Learning in HIV Testing: A Bibliometric Analysis of Published Studies 2000-2024

Author: Musa Jaiteh[1]

Co-authors: Edith Phalane [1] ; Yegnanew Shiferaw A [2] ; Refilwe Phaswana-Mafuya N [1]

[1] *University of Johannesburg*

[2] *University of Johanneburg*

Corresponding Authors: refilwep@uj.ac.za, yegnanews@uj.ac.za, edithp@uj.ac.za, mjaiteh1993@gmail.com

Background: The human deficiency virus (HIV) remains a devastating public health threat, affecting 39 million people globally, with approximately 60% of these cases occurring in Sub-Saharan Africa. Early detection and diagnosis of HIV are crucial for preventing the further spread of the virus, making HIV testing a pivotal tool for achieving the UNAIDS goal of ending AIDS by 2030. The World Health Organization and UNAIDS have emphasized the importance of adopting innovative testing strategies, such as those involving machine learning. Machine learning can accurately predict high-risk individuals and facilitate more effective and efficient testing methods compared to traditional approaches. Despite this advancement, there exists a knowledge gap regarding the extent to which machine learning techniques are integrated into HIV testing strategies worldwide. To address this gap, this study aimed to analyze published studies that applied machine learning to HIV from 2000 to 2024.
Methods: This study utilized a bibliometric approach to analyze studies that were focused on the use of machine learning in HIV testing. Relevant studies were captured through the Web of Science database using synonymous keywords. The bibliometrics package in R was used to analyze the characteristics, citation patterns, and contents of 3962 articles, while VOSviewer was used to conduct network violations. The analysis focused on the yearly growth rate, citation analysis, keywords, institutions, countries, authorship, and collaboration patterns.
Results: The analysis revealed a scientific annual growth rate of 8.8% with an international co-authorship of 44.7% and an average citation of 23.16 per document. The most relevant sources were from high-impact journals such as PLOS ONE, Aids and Behavior, Journal of Acquired Immune Deficiency Syndrome, Journal of International Aids Society, AIDs, and BMC Public Health. The USA, The United Kingdom, South Africa, China, and Canada produce the highest number of contributions. The results show that the University of California, Johns Hopkins University, Harvard University, and the University of London have the highest collaboration networks.
Conclusion: This study identifies trends and hotspots of machine learning research related to HIV testing across various countries, institutions, journals, and authors. These insights are crucial for future researchers to understand the dynamics of research outputs in this field.

Session / 7

Machine-learning algorithms for mapping LULC of the uMngeni catchment area, KwaZulu-Natal

Authors: ORLANDO BHUNGENI[1] ; Micheal Gebreslasie[1] ; Ashadevi Ramjatan[1]

[1] *University of KwaZulu Natal*

Corresponding Authors: 822820676@stu.ukzn.ac.za, 223146923@stu.ukzn.ac.za, gebreslasie@ukzn.ac.za

Abstract: Analysis of land use/land cover (LULC) in the catchment areas is the first action toward safeguarding the freshwater resources. The LULC information in the watershed has gained popularity in the natural science field as it helps water resource managers and environmental health specialists develop natural resource conservation strategies based on available quantitative in-formation. Thus, remote sensing is the cornerstone in addressing environmental-related issues at the catchment level. In this study, the performance of four machine learning algorithms (MLAs), such as Random Forests (RF), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Naïve Bayes (NB) was investigated to classify the catchment into nine rele-vant classes of the undulating watershed landscape using Landsat 8 Operational Land Imager (L8-OLI) imagery. The assessment of the MLAs were based on the visual inspection of the analyst and the commonly used assessment metrics, such as user's accuracy (UA), producers' accuracy (PA), overall accuracy (OA), and kappa coefficient. The MLAs produced good results, where RF (OA= 97.02%, Kappa= 0.96), SVM (OA= 89.74 %, Kappa= 0.88), ANN (OA= 87%, Kappa= 0.86), and NB (OA= 68.64 Kappa= 0.58). The results show the outstanding performance of the RF model over SVM and ANN with a small margin. While NB yielded satisfactory results, which could be primarily influenced by its sensitivity to limited training samples. In contrast, the robust per-formance of RF could be due to an ability to classify high-dimensional data with limited train-ing data.
Keywords: uMngeni River Catchment; Machine learning; LULC; Landsat 8; Remote sensing

Session / 38

National Policy Data Observatory (NPDO)

Corresponding Author: rholder@csir.co.za

The National Policy Data Observatory (NPDO) is a government-led initiative, currently hosted at the CSIR, with the main objective to support data-driven decision making in government on various socio-economic interventions. The NPDO was initiated at the height of the Covid-19 pandemic by the DSI, supported by the CSIR, Statistics SA and South African Revenue Service. The NPDO leverages the CSIR's high-speed networking infrastructure (SANReN) and high-performance computing infrastructure (CHPC) under the National Integrated Cyber Infrastructure System.

Session / 27

Opening

Opening
by
NGEI Executive Manager
Dr Lulama Wakaba

Session / **35**

Quantum ActiveScale - Durable, Secure, S3-Compatible Object Storage for Data Analytics, Active Archiving and Long-Term Retention. Quantum Myriad - Ending the Constraints, Compromises, and Trade-offs of Legacy Storage Architectures.

Author: Liam Clifton[1]

[1] *Quantum Director for Northern EMEA and South Africa*

Corresponding Author: liam.clifton@quantum.com

Quantum ActiveScale Storage
ActiveScale Object Storage provides a new, innovative approach to creating a simple, 'always-on' data repository that scales when and how you need it to—with
the extreme data durability, accessibility, and security required of petabyte-scale growth. With ActiveScale Cold Storage, ActiveScale reduces the cost of storing your cold data sets by up to 80%.

Durable, Secure, S3-Compatible Object Storage for Data Analytics, Active Archiving and Long-Term Retention
Whether you are developing solutions for life and Earth sciences, media production, government programs, web services, IoT infrastructure, AI/ML, or video surveillance, ActiveScale puts an affordable scalable solution within reach. Now, you can build your own private cloud storage environment - and seamlessly grow your data stores from terabytes to exabytes.

Quantum Myriad
Ending the Constraints, Compromises, and Trade-offs of Legacy Storage Architectures
The amount of unstructured data in the enterprise is expected to double over the next 5 years and yet most enterprises continue to store this data on file storage systems that were designed over 20 years ago.

In an attempt to keep pace with this growth and associated performance requirements, scale-out NAS systems have "thrown hardware" at the problem.
Current file and object storage systems have tried to bolt-on flash storage, but these systems were not designed for NVMe flash and can't provide the
consistent low-latency, high IOPS performance required by today's applications.

Even more recent all-flash file and object storage systems are shackled to specialized hardware and won't run natively in the cloud, so they don't enable
hybrid-cloud infrastructure. This has led to constraints, compromises, and trade-offs that are impacting IT and holding back data driven organizations.

Myriad Brings New Levels of Simplicity and Adaptability to File and Object Storage
Quantum Myriad provides organizations with a totally modern, cloud-native architecture that brings enterprises new levels of simplicity and adaptability
for their unstructured data storage.

Session / **25**

Risky business: Research data at risk in a dynamic and evolving multidisciplinary research environment

Author: Louise Patterton[1]

[1] *CSIR*

Corresponding Author: lpatterton@csir.co.za

At-risk data is an unfortunate research reality and can be present in all data formats in a range of research disciplines. Research data at risk are defined as data that are at risk of loss due to factors including deterioration of the media, lack of accompanying explanatory documentation, or data existing in non-digital formats, and often involve historic and irreplaceable data. With continued access to older data being associated with a range of benefits, the factors placing valuable data at risk are cause for concern.
This paper reports on a multi-method study comprising a survey and interviews and reports on the results of a case study into data at risk at a leading albeit relatively resource-constrained South African multidisciplinary research institute. In conducting research, development and innovation for the past seven decades, the institute had over time undergone a multitude of dynamic, progressive and often interruptive changes. This case study describes the prevalence, nature and causes of factors identified by research group leaders (RGLs) as threatening the continued access and usability of research data. The paper also reports on the challenges encountered by RGLs when addressing at-risk data.
Several recommendations and strategies are put forward to address identified risk factors. Besides the obvious recommendation that a data rescue project be launched, suggestions include awareness training around data at risk, involving the Library and Information Science sector in data rescue, continued efforts to acquire a dedicated institutional data repository, and ensuring the scope of project risk management includes data considerations. It is anticipated that the combined implementation of recommendations will ensure the accessibility and usability of older at-risk data and lessen the odds of current and future data becoming compromised.

Session / **12**

Securing Identity using Biometrics and Zero Knowledge Proofs

Authors: Sthembile Ntshangase [1] ; Siphelele Myaka[2]

[1] *Researcher*

[2] *Cybersecurity Researcher*

Corresponding Authors: smlambo@csir.co.za, smyaka@csir.co.za

Data protection and cybersecurity are distinct concepts, but they complement each other. Data protection ensures data integrity, while cybersecurity protects the digital ecosystem from threats like cyberattacks and malware. In an interconnected digital world, identity is increasingly stored, shared, and processed as data, including biometrics and Personal Identifiable Information. To address these challenges, a method using Zero Knowledge Proofs (ZKPs) is proposed to protect biometrics information and secure identities. ZKPs allow one party to prove a statement is true without revealing additional information, ensuring confidentiality and security without exposing the biometric information. This approach not only enhances biometric data security but also addresses privacy concerns associated with storing and transmitting sensitive information.

Session / **31**

Social Engineering a Security Data Management Risk for Novice Internet and Mobile Users

Author: Nobert Jere[1]

[1] *Walter Sisulu University*

Corresponding Author: njere@wsu.ac.za

Social engineering is generally explained as the practice of trapping, manipulating, or misleading a victim. Novice Information Technology users are one of the targeted victims. Such users are mainly

affected by phishing attacks that are the greatest popular type of social engineering attack. Through malicious websites or infected email attachments, phishing attempts characteristically to use human error as a means of credential harvesting or virus distribution. There have continuously been social engineering techniques, and there will always be more of them. Just as stratagems attempt to deceive decision makers, social engineering does the same. Even though security awareness training may not fully eliminate this vulnerability, it is essentially difficult to completely get rid of social engineering vulnerabilities. Social engineering poses a constant danger to cybersecurity because it takes advantage of human vulnerabilities instead of computer system security weaknesses. The novice users are the main victims of this. The main question of the study is:

How can social engineering techniques be redefined to improve security risks among novice IT users?

This study proposes that applying more relevant social engineering techniques that could educate novice IT users to be more secure when online. Using online secondary data, the study is based on publicly available open-source datasets and novice users' experiences.

Findings show that novice users are exposed and prone to data security breaches and risk. Studies have found that novice users seem to be overconfident in understanding social engineering attacks but hold incorrect beliefs. Novice users had major misunderstanding of what constitutes social engineering, and the risks of these attacks.

Key Words: Social engineering, novice users, data management, cybersecurity, mobile users

Session / 6

Spatial Modelling of Irrigation Water Quality: Assessing SAR in South Africa's Agricultural Landscapes

Author: Danielle Roberts[1]

Co-author: Xolani Nocanda [2]

[1] *University of KwaZulu-Natal*

[2] *CSIR Water Centre*

Corresponding Authors: xnocanda@csir.co.za, robertsd@ukzn.ac.za

The Sodium Absorption Ratio (SAR) is a critical metric used to assess the suitability of water for agricultural irrigation, reflecting the potential for sodium to accumulate in soil and negatively affect crop yield and the ecosystem. In South Africa, agriculture is a cornerstone of economic development, contributing significantly to GDP and employment. Identifying geographical locations with poor SAR measures is essential for sustaining agricultural productivity and environmental health. In this study, a generalized additive model (GAM) was employed to analyse the spatial distribution of the SAR across South Africa. The model incorporated a spatial effect based on the geographical coordinates of the sample locations. This allowed for the investigation of how geographical factors influence the SAR in various regions across South Africa, while controlling for predictors, such as other water quality parameters. The study made use of data from inorganic water chemistry analysis of samples from rivers, dams and lakes that were collected between the years 1970 to 2011 in South Africa. The significance of this research lies in its capacity to pinpoint locations with poor water quality, thereby guiding interventions aimed at soil and water management to avert potential degradation of arable land. The findings of this study not only aid in optimizing resource allocation for improving water quality but also contribute to the broader objectives of sustainable agricultural practices and economic stability in South Africa.

Alongside this study, an interactive dashboard in under development for the monitoring and evaluation of water quality data in South Africa. The dashboard incorporates visualizations and important summary measures for SAR as well as other various water quality parameters. This tool democratizes access to vital information, enabling stakeholders to make informed decisions based on comprehensive water data analysis and visualizations. This tool not only enhances transparency

and accountability but also facilitates a more targeted and efficient allocation of resources towards improving water quality initiatives in South Africa.

Session / **29**

The Future of Research Data: Open Science, FAIR Principles, and Effective Governance

Author: Thapelo Maredi[None]

Corresponding Author: thapelo.maredi@altron.com

In the realm of research, effective data governance plays a pivotal role in ensuring the integrity, accessibility, and usability of research data. We delve into the significance, methodologies, and complexities involved in establishing robust data governance frameworks tailored specifically for research data.
Research data governance encompasses the policies, processes, and infrastructure which facilitate data management, sharing, and reuse while adhering to ethical, legal, and regulatory requirements. This abstract elucidates key components of research data governance, including data management plans, data stewardship, metadata standards, and data sharing protocols.
Moreover, it addresses the unique challenges encountered in governing research data, such as heterogeneous data formats, disciplinary differences, and evolving data management practices. Strategies for overcoming these challenges, such as community-driven standards development, interoperable data repositories, and data curation services, are explored.
Furthermore, it discusses emerging trends and technologies shaping the landscape of research data governance, such as open science initiatives, FAIR principles (Findable, Accessible, Interoperable, Reusable), and machine-readable data policies. It emphasises the need for collaborative efforts among researchers, institutions, funding agencies, and policymakers to foster a culture of responsible data stewardship and data sharing.

Session / **11**

The Role of AI-Powered Tools in Data Management and Analysis

Author: Lungile Nkosi[1]

Co-author: Israel Agaku [1]

[1] *1. School of Health Systems and Public Health, University of Pretoria, South Africa 2. Chisquares Inc. United States*

Corresponding Authors: iagaku@chisquares.com, nkosi.lungile@gmail.com

Background

The growing volume, complexity, and diversity of data in Africa present significant challenges and opportunities for research. Limited access to open-access datasets, labor-intensive data analysis methods, and inefficient publication processes hinder research progress, especially in public health—a critical area for improving lives on the continent. AI-powered tools are poised to transform data-intensive research within the African context.

Methods

This presentation explores the functionalities of AI-powered tools and their potential to revolutionize epidemiologic and broader data-driven research in Africa and beyond. We examine how these tools leverage big data and machine learning technologies to enhance research capabilities and outcomes, with particular focus on the Chisquares platform as an example.

Results

AI-powered tools provide a robust solution for improving data accessibility, research efficiency, and collaboration in Africa through the following features:

Open Data & Security: AI-powered tools democratize data by offering secure access to a centralized repository of large-scale, nationally representative surveys from across Africa. This approach supports open data practices while ensuring data integrity, standardization, and secure storage in a cloud-based environment. Advanced security measures, including encryption, access controls, and compliance with data protection regulations, are integral to these tools.

Big Data & Machine Learning: Utilizing artificial intelligence (AI) and sophisticated algorithms, AI-powered tools automate over 80% of tasks involved in manuscript preparation, including data analysis and writing. This automation allows researchers to focus more on scientific inquiry and knowledge generation.

Data Management: AI-powered tools streamline the entire research workflow by integrating essential functionalities for all stages of the scientific process, from sample size calculation and sampling to data collection, analysis, writing, and journal submissions. This comprehensive approach eliminates the need to switch between multiple applications, enhancing efficiency and productivity in data management.

Conclusion

AI-powered tools foster data literacy, promote open data practices, and ensure secure research environments, empowering researchers across Africa to generate impactful insights and contribute to improved health outcomes. By promoting efficient data management and leveraging big data and machine learning, these tools stand at the forefront of revolutionizing research data repositories and services for the future.

Session / 1

The Significance of CoreTrustSeal Certification: A Case Study of Stellenbosch University

Author: Xabiso Xesi[1]

[1] *Stellenbosch University*

Corresponding Author: xabiso@sun.ac.za

Abstract:
The Significance of CoreTrustSeal Certification: A Case Study of Stellenbosch University
Stellenbosch University's (SU) recent achievement of CoreTrustSeal (CTS) certification represents a significant milestone in the institution's commitment to excellence in research data management. This certification, granted to its data repository, SUNScholarData, underscores the university's dedication to upholding international standards of data integrity, accessibility, and sustainability.
CTS is a globally recognized certification awarded to data repositories that demonstrate adherence to best practices in data management (Dillo and Leeuw, 2018). For SU, this certification signifies compliance with international standards, including the FAIR principles, ensuring that research data is findable, accessible, interoperable, and reusable. Moreover, it highlights the university's commitment to maintaining data integrity and quality, thus enhancing the credibility of its research outputs.
To attain CTS certification, SU had to meet a comprehensive set of requirements outlined by the CTS Board. These requirements encompassed data integrity and authenticity, appraisal criteria, documented storage procedures, preservation planning, data quality assurance, workflows, data discovery and identification, and data reuse capabilities.
Beyond compliance, CTS certification reinforces SU's commitment to long-term data accessibility and preservation. By securely storing and making research data accessible for future use, the university contributes to the longevity of scholarly contributions and facilitates collaboration across

diverse research projects. Additionally, the certification promotes interoperability, enabling seamless data sharing and integration with other institutions.
Stellenbosch University's attainment of CTS certification enhances the trustworthiness and credibility of its data repository among researchers, funders, and the broader academic community. It positions the university as a trusted hub for valuable research data and highlights its dedication to responsible data management practices. As Stellenbosch University continues to advance in research and innovation, CTS certification serves as a guiding beacon towards a future where data is managed, preserved, and shared responsibly for the benefit of the global research community.
The aim of this presentation will be to share the experiences of SU Library and Information Service as it went through the rigorous process of applying for the CTS certification. Both presenters have been intimately involved in the process from the Cape Peninsula University of Technology (which is yet to receive a certification) and SU. It is hoped that the experiences shared may trigger more interest from other institutional repositories to follow the same process.

References:
DILLO, I. & LEEUW, L. D. 2018. CoreTrustSeal. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare, 71, 162-170.

Session / 32

The prediction of the South African elections' outcome

Authors: Paul Mokilane[None] ; Nontembeko Dudeni-Tlhone[1] ; Jenny Holloway[1] ; Mahlatse Mbooi[1] ; Tyrone Naidoo[1] ; Gareth Edwards[1]

[1] *CSIR*

Corresponding Author: pmokilane@csir.co.za

Abstract
The CSIR used its highly acclaimed elections prediction model to predict the outcome of the 2024 elections held on the 29 May 2024. This model was initially developed for South African elections and was first introduced during the 1999 general elections. Since its inception, the model has successfully forecasted the outcomes for local government elections with just 5% to 10% of VDs declared. It has garnered international recognition for its accuracy across various electoral systems. The tool also plays an integral role in authenticating the final results of the elections.
The model relies on two main principles of understanding voting behaviour: past voting patterns and influences from political, socioeconomic and demographic factors. Using this data, CSIR statisticians and data scientists group voters into clusters, anticipating that changes in voting behaviour will be similar within each group. When the early results arrive, the model uses this data to estimate new voting behaviour for similar groups of districts. These estimates are then extended to the remaining districts yet to be counted. By combining known results with predicted ones, the model then generates a final prediction. The model correctly predicted that the ANC would lose their majority, even though they would remain the leading party in the country and predicted that the MK party would overtake the EFF to be the 3rd largest party in the country. The predictions for the MK party were surprisingly good, considering they were a new party, and the model predicted their support to be close to 14% of the votes at a time in the counting process when the scoreboard was only showing their support to be around 8%. Overall, predictions for the top 6 parties performed well from around 10% of the VDs declared, with the ANC prediction taking longer to stabilize than the other parties but remaining within the 2% error margin once predictions were released.

Session / 36

Uncovering Seasonal Trends in Motor Insurance Claims: A Gender-Based Analysis

Authors: Taurai Hungwe[1] ; Sandile J Buthelezi[1] ; Solly M Seeletse[1] ; Vimbai Hungwe[1]

[1] *Sefako Makgatho Health Sciences University*

Corresponding Authors: solly.seeletse@smu.ac.za, vimbai.hungwe@smu.ac.za, taurai.hungwe@smu.ac.za, js-buthelezi@gmail.com

This study explores the seasonal patterns in motor insurance claims and predictive outcomes among male and female drivers. Using a large dataset of insurance claims and driver information, we employ machine learning techniques to identify gender-specific trends and predict claims. Our preliminary results indicate significant seasonal variations in claims patterns between genders. Our findings have important implications for insurance risk assessment, pricing, and policy development. This research aims to contribute to the development of more accurate and fair insurance models, promoting safer driving behaviors, reducing insurance costs and affordable premiums to customers.

Session / **13**

Unpacking the role of information specialist at a 21st century academic library: Research Data Management at the University of Pretoria.

Authors: Tlou Mathiba[1] ; Martie Van Deventer[1]

[1] *University of Pretoria*

Corresponding Author: tlou.mathiba@up.ac.za

Research data management (RDM) is rapidly becoming an essential service. This has forced higher education institutions (HEIs), research councils, publishers and funding agencies to embark on this journey. In February 2015, in alignment with this global trend, the South African National Research Foundation (NRF) released a statement on Open Access (OA) to Research Publications' funded by NRF. The statement states that research papers fully or partially funded by the NRF should be deposited to the administering institutional repository with an embargo period of not more than 12 months. Furthermore, the statement states that "the data supporting the publication should be deposited in a trusted Open Access repository, with the provision of a Digital Object Identifier (DOI) for future citation and referencing." In support of the NRF statement, the University of Pretoria (UP) as a research-intensive institution and advocate for OA, had an RDM policy (S4417/17) approved in 2017. The university in its pursuit to implement RDM infrastructure and services launched the research data repository, Figshare in 2019. After the launch, there was a change in information specialists' roles in order to support RDM services. The information didn't know what their role in RDM would be and hence the researcher undertook this study. The research unpacks their role, as information specialists in RDM. The University library, as the custodian of the UP's formal RDM drive, must be setting the pace for all researchers. Information specialists employed by this academic library are expected to be knowledgeable, competent and able to advise faculty staff and researchers within the institution.

Session / **23**

Update on NICIS 100Gbps Data Transfer Services – 1TB in 3mins!

Author: Kasandra Pillay[1]

[1] *SANReN*

Corresponding Author: kasandra@sanren.ac.za

Moving large amounts of data poses a significant challenge. In most cases, networks optimised for business operations are neither designed nor capable of meeting the data movement requirements

of data-intensive research. When scientists attempt to run data intensive applications over these so-called "general-purpose" or enterprise networks, the result is often poor performance. In many cases, this poor performance significantly impacts the scientific mission, leading to challenges such as not receiving data on time or resorting to "desperate" measures, such as physically shipping disks.

There has been a significant increase in available network capacity and a greater need to be able to transfer large amounts of data efficiently. The CSIR through SANReN, NICIS offers a Data Transfer Service as an effort to facilitate the movement of large datasets by South African researchers and scientists.

With the implementation of the SANReN 100Gbps backbone network capacity, 100Gbps Data Transfer Nodes have been implemented in Cape Town and Johannesburg with Globus (globus.org) data transfer software installed. The best international data transfer results seen were between Johannesburg/Cape Town DTNs and Colorado (NCAR GLADE). Data transfer speeds reached between 4.78GB/s-5.48GB/s, which resulted in 1TB of data being moved in only 3 minutes!

If you have large data transfer requirements, please contact pert@sanren.ac.za for assistance.

Session / 39

Update on South African Cyberinfrastructure Initiatives

Update on South African Cyberinfrastructure
Initiatives
by
Dr Happy Sithole
NICIS Centre Manager

Session / 41

Welcome

Welcome by Programme Director,
by
Mrs Ina Smith
ASSAf Planning Manager

Session / 26

Welcome

Welcome
by
Programme Director,
Mrs Ina Smith

Session / 40

Wrap Up Day 1

Wrap Up Day 1
by
Mrs Ina Smith
Programme Director;