A NATIONAL BIG DATA STRATEGY

FOR RESEARCH, DEVELOPMENT AND INNOVATION

science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

nicis
NATIONAL INTEGRATED
CYBER INFRASTRUCTURE SYSTEM

CSIR
Touching lives through innovation

NICIS VISION

The realisation of a vibrant and competitive knowledge-based economy impacting socio-economic development by enabling education, research, and innovation through shared access to advanced cyberinfrastructure facilities and services.

NICIS MISSION

To provide a world-class national integrated cyberinfrastructure system that enables research, innovation and learning, comprising a national high performance computing facility, a national research and education network and national data-intensive research infrastructure accessible across the research and higher education sector through integrated eResearch services and the development of relevant human capital.

NICIS KEY STRATEGIC OBJECTIVES

01 Sustain a world class and relevant national integrated cyberinfrastructure system for Science and Technology

02 Enable and promote eScience in South Africa

03 Position South Africa to take part in, host and lead large-scale global research and science projects (e.g. SKA and CERN experiments)

04 Provide thought leadership to South Africa's evolving cyberinfrastructure strategy and activities, and facilitate the uptake of advanced cyberinfrastructure

05 Foster the development of human capacity in cyberinfrastructure and its application, and contribute to the transformation of this sector

FOREWORD BY THE MINISTER

Although significant progress has been achieved in improving the socio-economic conditions of South Africans, the intertwined challenges of unemployment, poverty and inequality remain. On the science, technology and innovation (STI) front, the adoption of the White Paper on Science, Technology and Innovation in March 2019 confirms the South African government's commitment to using STI to develop the country.

The National Development Plan has identified STI as key drivers of socio-economic growth and job creation. Initiatives such as the Presidential Commissionon the Fourth Industrial Revolution (4 IR) indicate government's intention to advance the National System of Innovation.

Big data is not just another technology; it is the driver behind the 4 IR and many other nascent ICT developments, such as the internet of things, artificial intelligence and blockchain technologies, and a common denominator in growing the digital economy. Rapid technological advancements globally require updated strategies for South Africa to thrive in this changing environment.

The National Big Data Strategy for Research, Development and Innovation not only aligns with government STI strategies, but also supports policies such as the data and cloud policy drafted by the Department of Communications and Digital Technologies. Its objectives – more rapid human capital development, building world-class cyberinfrastructure and fostering stronger relationships among the public, private and academic sectors – are important ingredients in accelerating the development of the economy.

I am confident that this strategy will enable macro-economic development, and that its implementation will improve the lives of our people and the fortunes of our communities.

Minister Blade Nzimande:
Department of Higher Education,
Science and Innovation

MESSAGE FROM THE DIRECTOR-GENERAL

The South African National Development Plan (Vision 2030) presents critical targets for the eradication of poverty, and the reduction of unemployment and extreme inequality. To achieve these targets, it is necessary to drive positive socio-economic outcomes through science, technology and innovation.

There has been significant progress in the South African National System of Innovation, with achievements in fields such as energy, healthcare, education, climate change, food security and manufacturing.

The digital revolution presents opportunities to achieve strategic goals such as those expressed in the White Paper on Science, Technology and Innovation, the recommendations of the 2020 Report of the Presidential Commission on the Fourth Industrial Revolution, and the data and cloud policy gazetted for comment by the Department of Communications and Digital Technologies.

The outcomes of research have always been powerful drivers of innovation and research has become increasingly data-driven. Data, particularly big data, holds vast potential to develop more effective responses to challenges and to build a thriving digital economy.

The objectives of this strategy support the achievement of national goals, including the transformation of the human capital that underpins entrepreneurship, and fast-tracking the growth of a competitive digital economy by linking the public, private and academic sectors in the country.

I am enthusiastic about this strategy and the contribution it will make to the well-being of the country and continent.

Dr Phil Mjwara, Director-General:
Department of Science and Innovation.

EXECUTIVE SUMMARY

It is widely recognised that the rapid global growth in data and the emergence of new digital technologies are profoundly impacting the private, public and research sectors. This Big Data revolution is set to continue to accelerate, and a period of significant disruption is anticipated in the upcoming decades, in which this impact of digital technologies will radically transform the global economy, the research landscape, as well as policy and decision-making. This will affect the quality of life and lifestyle of people, and enable the addressing of global challenges, such as energy, health, global and climate change and economy, in novel ways.

Big Data is a disruptive phenomenon that has become the common denominator in using technologies such as cloud computing, artificial intelligence and machine learning to benefit from the fourth industrial revolution (4IR). In this context, Big Data holds vast potential for South Africa to become less driven by a resource-based economy and become a producer rather than a consumer of Big Data outputs, thus, accelerating South Africa's transition towards a vibrant digital economy.

South Africa is a rich source of research data, and it has invested substantially in cyberinfrastructure that drives the acquisition and generation of data across a number of domains. Driven by initiatives such as the Square Kilometre Array (SKA) radio telescope, regional climate modelling and bioinformatics, these commitments are set to increase greatly in the future and uniquely position South Africa to derive substantial benefits from Big Data.

Learnings from the novel coronavirus pandemic demonstrate the urgent need for ongoing and long-term readiness to respond to this and similar future disasters. This readiness includes a capable e-infrastructure, data management, tools and services systems, data policies and strategies for Big Data.

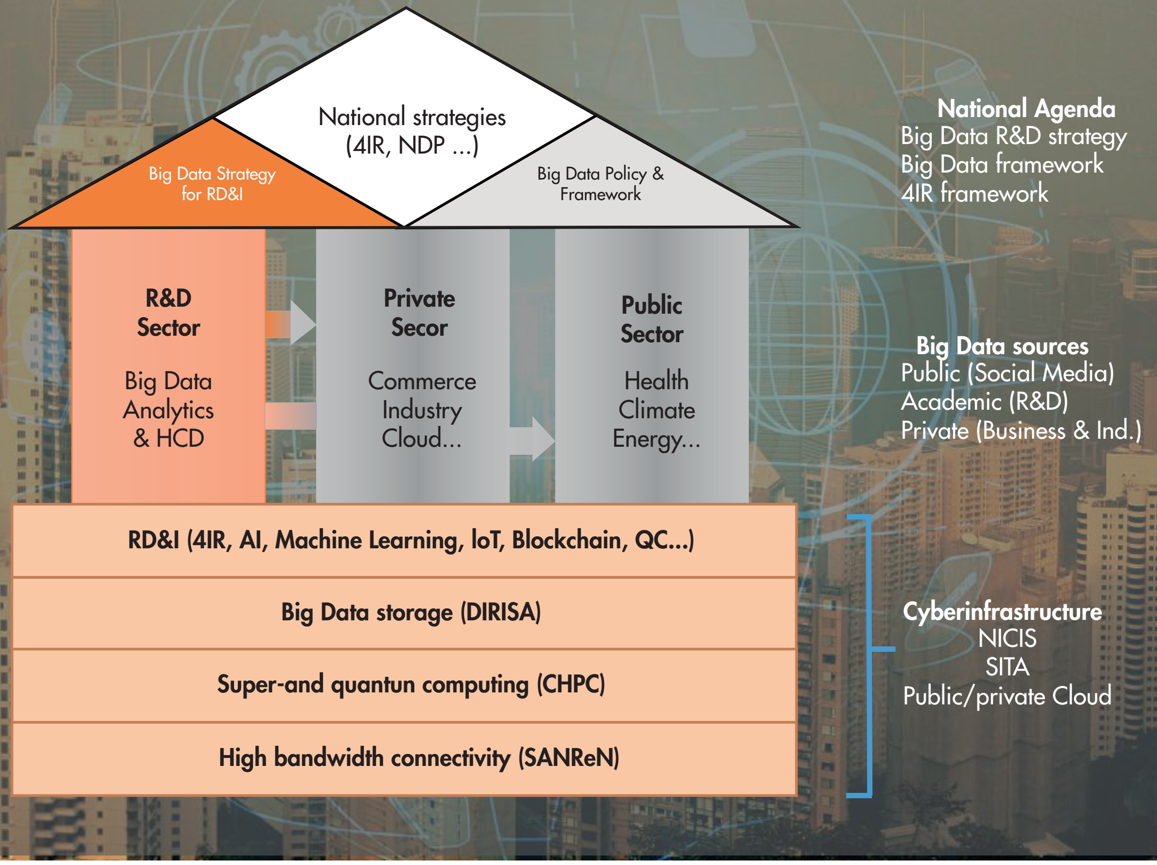
A number of challenges need to be addressed in order to derive greater benefit from these investments. There are various Big Data research activities underway across universities and research institutions, though some are duplicative and suboptimal small interventions. The scarcity of expertise and skills is regarded as the most significant hurdle. This shortage extends across the scientific, technological and management domains. collaboration among government, industry and the research sectors can yield greater efficiencies in terms of shared and human capital development, if these efforts are less fragmented. It is necessary to address these hurdles in a more focused and coordinated manner to avoid South Africa trailing behind other nations and ensure that Big Data serves national rather than sectoral vendors or foreign interests and to avoid South Africa being a collector of data, but not benefiting from it.

PURPOSE AND CONTEXT OF STRATEGY

This strategy presents the strategic priorities of the Department of Science and Innovation for research Big Data development and innovation. The purpose of this strategy is to provide information to planners and decision-makers in the research and public sectors when deliberating about commitments and initiatives that leverage opportunities and meet the challenges presented by research Big Data. The primary aim of this strategy is to maximise the return on investments in research Big Data and thus realise the economic, social, educational, scientific and industrial beneficiation potential of research Big Data for South Africa. This strategy specifically concerns research Big Data in the public research sector and not Big Data in the private realm. As such, its focus and scope align and harmonise with other national strategies and agendas for public sector investment in Big Data and ICT. In particular, this strategy supports the mission and strategies of the Department of Communications and Digital Technologies (DCDT).

Big Data research, development and innovation involves the production of knowledge and novel solutions through the application and advancement of emergent and converging technologies, such as artificial intelligence, machine learning and block chain algorithms. The outputs of this research are new technologies and services that are crosscutting enablers of developments, such as cyber-physical systems and 4IR. This strategy supports the fruition of these related initiatives and goals across government and the research environment. Examples of such goals are skills and expertise development, entrepreneurship as contemplated in national blueprints such as the e-Strategy [1] and the Presidential Commission on the Fourth Industrial Revolution [3].

Together with other national agendas, this strategy also contributes to Big Data governance and innovation that affect both the public and private sectors. Policy framework objectives of the Data and Cloud Policy Framework of DCDT [4], particularly supported by this strategy, include provision for "research, innovation and human capital development for data…". More specifically, this strategy supports their policy interventions for "…the implementation of digital infrastructure strategies…" (clause 9.11), and the development of "capacity building programmes and initiatives to provide skills on Big Data…" (9.8.1). A further intervention (9.10.1) that "The DSI shall be responsible for the R&D on Big Data…" explicates the synergistic relationship between the draft policy framework of DCDT and this strategy.



The key stakeholder groups for this strategy would include governmental departments and their supported entities that invest in Big Data, as well as academic and research institutions that conduct Big Data research innovation. Specific examples of such stakeholders are universities, research councils and government entities, such as the DCDT and the National Research Foundation, while NICIS, with other public and private providers, such as the State Information Technology Agency (SITA) and private cloud service companies, provides advanced cyberinfrastructure for Big Data analytics, capacity development and R&D.

The vision of this strategy supports the outcomes of other national strategies:

> *We envision a Big Data ecosystem, deriving from large, historic and diverse datasets; knowledge that leads to innovation, accelerated socioeconomic growth, and that positions South Africa to be more competitive in the 21st Century Big Data economy.*

A mid- to long-term (5 to 10 years) view is adopted, and this strategy presents a position that should be reviewed on a regular basis, in response to changing strategic priorities and emergent technological trends. Although the focus is on national interests, it is also necessary to consider continental and international perspectives on research Big Data strategies and goals. The following strategic objectives, in support of national interests, are necessary to achieve this vision:

OBJECTIVE 1: HUMAN CAPITAL DEVELOPMENT. DEVELOP THE NEXT GENERATION OF R&D EXPERTISE AND SKILLS TO DERIVE KNOWLEDGE FROM BIG DATA AND ADVANCE THE FRONTIERS OF BIG DATA SCIENCE.

There is an acute shortage of such skills in the public and private sectors, and a coherent training and education approach is essential to ensure that South Africa meets current and future Big Data human capital demands [9, 10]. It is necessary to develop leading research capabilities to deal with the unique scientific and technological challenges posed by Big Data. Interventions that align postgraduate and postdoctoral programmes with collaborative research projects are needed to develop solutions capable of dealing with the scale and complexities of Big Data analytics, visualisation and management [11, 9, 12]. The number of data science research staff at universities and research institutions should be increased radically. It is important not to "leave anyone behind", and concerted efforts should be made to capacitate sub-optimal universities through collaborative arrangements

OBJECTIVE 2: CYBERINFRASTRUCTURE. BUILD AND SUSTAIN A WORLD-CLASS AND RELEVANT NATIONAL CYBERINFRASTRUCTURE WITH ALLIED SERVICES THAT ENABLE AND ADVANCE BIG DATA RD&I.

Continued and increased investment in advanced research cyberinfrastructure is essential, in order to fulfil commitments to grand scale research projects such as the SKA, genomic sequencing, climate modelling and to support developments such as the 4IR. Cyberinfrastructure should support the management and analyses of vast amounts of data and real-time data streams from heterogeneous data sources such as social media, the Internet of Things and remote sensing satellites [13].

The sharing and reuse of Big Data are key to deriving greatest value from these resources. Appropriately scaled up and advanced technologies are essential for such access and the proper stewardship of Big Data.

NICIS has a key role in leading the design and coordination of the deployment of an advanced national cyberinfrastructure. It should collaboratively engage with stakeholders of Big Data, in order to achieve greater efficiencies in cyberinfrastructure investments. This is necessary to avoid duplication and solo development. A sustained investment is necessary to ensure that this technology platform remains relevant, in order to serve Big Data research needs adequately.

OBJECTIVE 3: COLLABORATION. FOSTER A THRIVING AND COLLABORATIVE ECOSYSTEM OF BIG DATA RESEARCH INNOVATION THAT LINKS GOVERNMENT, ACADEMIA AND THE PRIVATE SECTOR.

Mechanisms and incentives should be established to create opportunities to collaborate with government and industry by removing bureaucratic hurdles for technology and data sharing, and building sustainable programmes. One such possible mechanism is the creation of structures to coordinate collaborative projects across governmental, industrial and academic boundaries. Big Data innovation centres or cyber technology stations supporting the services, and monitoring and evaluation needs of national government and Big Data-based needs of different ministerial departments. Such centres would support so-called triple-P collaborative programmes with industry, small, medium and micro enterprises (SMMEs) and academia.

An example is the establishment of a Big Data Innovation Centre with NICIS and SKA-SA that focuses on grand challenge applications and helps to determine the infrastructure, datasets, analytical tools and interoperability requirements necessary to achieve key national priority goals. Industries such as multinational companies should be encouraged to invest in Big Data infrastructures and develop Big Data capabilities through collaborations and other mechanisms.

OBJECTIVE 4: DATA GOVERNANCE AND STEWARDSHIP. ADOPT STANDARDS AND POLICIES SUPPORTING THE PRACTICES OF OPEN DATA AND OPEN SCIENCE PRINCIPLES, AND ENSURE INTEROPERABILITY AND COMPLIANCE WITH PRIVACY, ETHICAL AND LEGAL REGULATIONS.

The value of data increases manifold when combined with other data sets or when repurposed for uses that differ from the initial purpose for which it was collected. It is essential that data is easily discoverable, in order to reuse and combine these with multiple other sources. Policies are needed to promote the interoperability and visibility of Big Data.

The privacy of individuals should be maintained at all times [14, 15]. New or revised policies and frameworks are necessary to protect privacy and clarify the roles and responsibilities of various actors involved in Big Data creation and use.

The Department of Science and Innovation (DSI) is in the process of developing a national Open Science Policy based on the Open Science Framework of Principles and Guidelines developed in 2018.
More research is necessary to fully understand and address the challenges of Big Data privacy, integrity, security and ethics. New techniques and tools are needed to help assess data security and secure data in an increasingly cyber-attack prone world.

OBJECTIVE 5: OVERARCHING COORDINATION. MAINTAIN AN OVERARCHING AND COHERENT NATIONAL APPROACH TO INVESTMENTS IN BIG DATA INITIATIVES AND ACTIVITIES.

In order to achieve greatest return and optimise investment, it is crucial that Big Data initiatives and activities are coordinated at a national level, across government, research institutions and academia. A structure representing the roles and interests of relevant ministerial departments and appropriate organisations would serve in an overarching capacity and governance role to orchestrate and oversee strategic implementation, planning and national investment in research Big Data.

A set of principles serve as values that underpin the coordination of a Big Data strategy:

Inclusivity *("Leave no one behind"). There is constructive engagement with all key stakeholders and beneficiaries in the implementation of a Big Data strategy with a focus on transformation. Funding, custodial, the research community, important national projects and the private sector are represented.*

Privacy. *Policies ensure the ethical use of Big Data and preserve the privacy of individuals as represented in Big Data and data in general.*

Governance and stewardship. *There are clearly defined roles and responsibilities for the actors in the governance of Big Data. Governance structures oversee and promote coordination across Big Data activities.*

Infrastructure. *NICIS serves as the primary Tier 1 platform in a sustained manner for Big Data initiatives and collaborates with other infrastructures used for specialised Big Data projects.*

FAIR. *As far as feasible, data generated using public funds should be "as open as possible and as closed as necessary" to support findability, accessibility, interoperability, and reusability.*

ACRONYMS AND ABBREVIATIONS

4IR	Fourth Industrial Revolution
AERAP	Africa Europe Astronomy Partnerships
ANDS	Australian National Data Service
AOSP	African Open Science Platform project
CERN	Conseil Européen pour la Recherche Nucléaire (European Organisation for Nuclear Research)
CHE	Council on Higher Education
CHPC	Centre for High Performance Computing
CI	Cyberinfrastructure
CSIR	Council for Scientific and Industrial Research
DCDT	Department of Communications and Digital Technologies
DHET	Department of Higher Education and Training
DIRISA	Data Intensive Research Initiative of South Africa
DSI	Department of Science and Innovation
DST	Department of Science and Technology
DTPS	Department of Telecommunications and Postal Services
HCD	Human Capital Development
HDI	Historically Disadvantaged Institution
HPC	High Performance Computing
HEI	Higher Education Institution
ICT	Information and Communication Technology
IKS	Indigenous Knowledge Systems
IOT	Internet of Things
ILIFU	Western Cape DIRISA Tier 2 Data Intensive Research Facility (Aka WT2DIRF)
KPI	Key Performance Indicator
LHC	Large Hadron Collider
MIIA	Machine Intelligence Institute of Africa
MTSF	Medium Term Strategic Framework
NDA	Non-disclosure Agreement
NDP	National Development Plan 2030
NEPTTP	National e-Science Postgraduate Teaching and Training Programme
NGAP	New Generation of Academics Programme
NICIS	National Integrated Cyberinfrastructure System
NIST	US National Institute of Standards and Technology
NIPMO	National Intellectual Property Management Office
NRDS	National Research and Development Strategy
NREN	National Research and Education Network
NRF	National Research Foundation
NSI	National System of Innovation
PC4IR	Presidential Commission for the Fourth Industrial revolution
PoPIA	Protection of Personal Information Act two thousand twenty
RI	Research Infrastructure
SADC	Southern African Development Community
SAEON	Southern African Environmental Observation Network
SAFIRE	South African Federated Identities for Research and Education
SANReN	South African National Research Network
SANSA	South African National Space Agency
SARAO	South African Radio Astronomy Observatory
SARIR	South African Research Infrastructure Roadmap
SAWS	South African Weather Service
SDGs	Sustainable Development Goals
SKA	Square Kilometre Array
SMME	Smal, medium and micro enterprises
STISA	Science, Technology and Innovation Strategy for Africa
TENET	Tertiary Education and Research Network of South Africa
TVET(s)	Technical and Vocational Education and Training (colleges)
USAf	Universities South Africa
UCDP	University Capacity Development Programme
UNISA	University of South Africa
WT2DIRF	Western Cape DIRISA Tier 2 Data Intensive Research Facility (Aka ILIFU)

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

There is broad agreement among business, academic and governmental institutions globally about the potential socioeconomic impact of Big Data as a key driver of the digital economy [16, 17]. Together with advances in technologies, such as artificial intelligence, machine learning, cloud computing and the IoT, Big Data presents a wide rang of opportunities to disruptively grow the economy and directly benefit society.

It is evident that Big Data has matured as an area of information technology, service and science. Big Data has been removed from the Gartner hype cycle for emerging technologies, indicating that this technology has moved beyond the "Plateau of Productivity" [18, 19]. The global collective sum of stored data and will reach grow from 33 zettabytes (ZB, 1021) in 2018, to 175 ZB by 2025, with a compounded annual growth rate of 61% [20]. Within this ever-increasing at the hyphen deluge of data, are nuggets of opportunities to improve our quality of life and solutions to pressing challenges, some of which are global. The value that can be derived from Big Data has largely been driven by renewed interest in these data science-related topics and emergent technological phenomena [21, 22]. New Big Data-related fields of research, such as the social impact and philosophy of Big Data, are also evolving [23, 24, 25].

Big Data has unquestionably shifted the research landscape beyond the traditional sciences. Research and innovation, arguably across all disciplines, have become profoundly data driven and data intensive [26, 27, 28, 29]. It is now possible to fuse disparate and heterogeneous datasets to model and predict the behaviour of complex systems [30].

Business and industry are already monetising Big Data, but its benefits cut across many domains, many of which are presented in the NDP and form part of the focus of the DSI White Paper [6, 31].

> *"Our prosperity as a nation depends on our ability to take full advantage of rapid technological change. This means that we urgently need to develop our capabilities in the areas of science, technology and innovation. We will soon [...] ensure that our country is in a position to seize the opportunities and manage the challenges of rapid advances in information and communication technology."*
> *[President Cyril Ramaphosa, State of the Nation Address, 2018]*

Repositories of data accumulated over the years by public entities, such as research councils, academia and government departments, offer new opportunities to extract insights and serve specific needs of citizens. South Africa is a rich source of research data and it invested substantially in the acquisition and generation of such data. Driven by initiatives such as the SKA, the participation in the ATLAS and ALICE CERN experiments, and contribution to the IPCC climate change projections, these investments are set to increase greatly. In addition to energy, transport, health, disaster risk and environmental sustainability, Big Data research outcomes are also transforming domains such as education, agriculture, economics, humanities and the social sciences. The successful implementation of the 4IR depends on the data to be used in technologies that perform some of the critical 4IR functions and that leverage Big Data to support industries in business decision-making processes.

In order to derive greater benefit from these investments, several challenges have to be addressed. While various Big Data research activities are underway across universities and research institutions, some are duplicative and suboptimal small interventions. The scarcity of expertise and skills is a significant hurdle, and this shortage extends across the scientific, technological and management domains [32, 24, 33, 34].

Collaboration among government, industry and the research sectors can yield greater efficiencies in terms of shared resources and human capital, if these efforts are less fragmented. Furthermore, it is necessary to address these hurdles in a more focused and coordinated manner to avoid South Africa trailing behind other nations and ensure that Big Data serves national, rather than vendor or foreign interests. It is also important to acknowledge the important role that access to national and global research infrastructure had and will have for the development of human capital in Big Data. In summary, there is a pressing need to formulate a common and coherent vision in terms of Big Data.

This strategy articulates this vision, the objectives and actions for South Africa to become a producer rather than remain a consumer of Big Data outputs. More crucially, without an appropriate strategy, South Africa will miss a unique opportunity to disruptively increase the pace of developing a digital-based economy if no determined and concerted effort is made to leverage the Big Data phenomenon in a prompt and decisive manner.

1.1 CONTEXT AND SCOPE

Big Data has become a key enabler of mainstream developments, such as the 4IR and the digital economy. This strategy represents a blueprint and a framework of actions by the DSI to achieve the greatest value from Big Data that has potential for research, development and innovation. It serves to provide information to planners and decision-makers, particularly in the research and public sectors, when deliberating about commitments and initiatives that leverage opportunities and that meet the challenges presented by Big Data. Government entities and supported institutions, universities, research councils and public sector custodians of Big Data would be key enactors of this strategy.

The principal intent of this strategy is to realise the beneficiation potential of Big Data fully through research and innovation as contextualised by the 2018 State of the Nation Address and national Big Data strategic initiatives.

As a crosscutting enabler of emergent and converging technologies, such as cyber-physical systems and 4IR, this strategy supports related strategic agendas and goals across government and the research environment. Figure 1 provides a perspective of how this strategy harmonises with sectors and other strategic initiatives. This strategy complements and aligns with:

- *The national e-Strategy [1] specifically, with Pillar 3: Digital Industrial Revolution;*
- *The objectives of the draft Big Data and Cloud Policy Framework of the DCDT [2] in providing opportunities to access Big Data services and participate in the Digital Economy and 4IR; and*
- *The mandate of the Presidential Commission on the Fourth Industrial Revolution [3] for mobilising resources to support the 4IR.*

This strategy should be regarded as a position to be reviewed as the technological landscape and strategic priorities change. In its present form, it informs the public and private sectors about the goals and priorities for research, development and innovation based on Big Data.

The context of this strategy is set by current realities, i.e., global trends, national strategic agendas, as well as the prevailing local landscape of initiatives and activities related to Big Data.

The prescription of particular technological solutions is not within the scope of this document, although the methodology for determining such solutions should be based on the tenets given in this strategy. Importantly, Big Data exists within a broader data, high-performance computing and data-intensive research context.

While the scope of this strategy is limited to Big Data, this contextual background is taken into account.

1.2 PROCESS FOLLOWED IN FORMULATING THIS STRATEGY

An inclusive process was followed in the formulation of this strategy. This version of the strategy results from a synthesis of expert opinion, inputs from key stakeholders, strategies of other countries, as well as reports and technical documents on Big Data strategy.

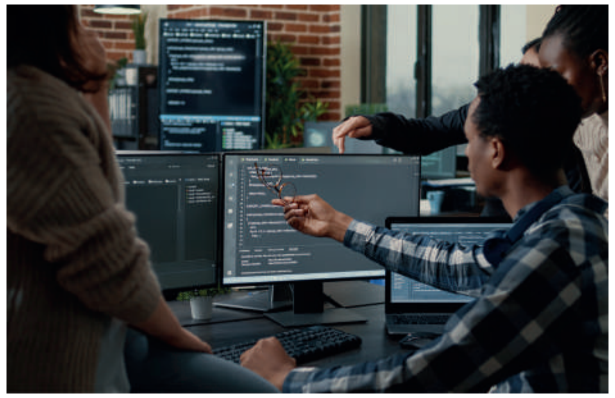
An initial draft of this strategy was developed by NICIS on request by DSI in 2018.

This version took into account the views of local and international experts who are leading or directing Big Data initiatives in some countries and of local professionals from academia and research councils, as well as the strategies of other countries (as summarised in Annexure A). This draft served as the point of departure for a broader and inclusive formulation process.

Feedback and views of stakeholders that included academic and governmental representatives (as listed in Annexure B) were solicited during a national workshop and these were used to further refine and adapt this document.

The recommendations of the DSI, following a presentation and a discussion meeting on the first draft, were incorporated. Further discussions with the DCDT resulted in closer alignment of this strategy with their draft Big Data and Cloud Policy Framework.

It is envisaged that comments elicited from specific key stakeholders in governmental and relevant organisations will also be taken into account. It is anticipated that the DSI will oversee the implementation of this strategy and that NICIS will serve as the implementing agent.



The concept of Big Data is introduced and the prevailing Big Data landscape, together with the main sources of Big Data, is considered.

Further sections provide an overview of the strategic drivers in terms of opportunities and challenges, the roles of different stakeholders and the present local landscape that informs the rationale for this strategic plan. Areas of interest that can stimulate collaboration within the public and private sector are highlighted. Several Big Data application domains are noted, although these are not assumed to be exhaustive.

A set of strategic objectives to leverage the opportunities and meet the challenges presented by the Big Data phenomenon are presented. These objectives are based on a vision and a set of principles and guidelines that, in turn, inform a recommended framework of actions to give effect to this strategy. The appendices provide a list of contributors, key stakeholders and summaries of Big Data strategies of some countries, and present some ethical aspects pertaining to Big Data.

2 WHAT IS BIG DATA?

Data drives the digital economy – the successful application of emergent technological developments, such as the 4IR and artificial intelligence, is predicated on data. Big Data is produced in diverse formats by various sources, including IoT devices, humans through social media, and research endeavours.

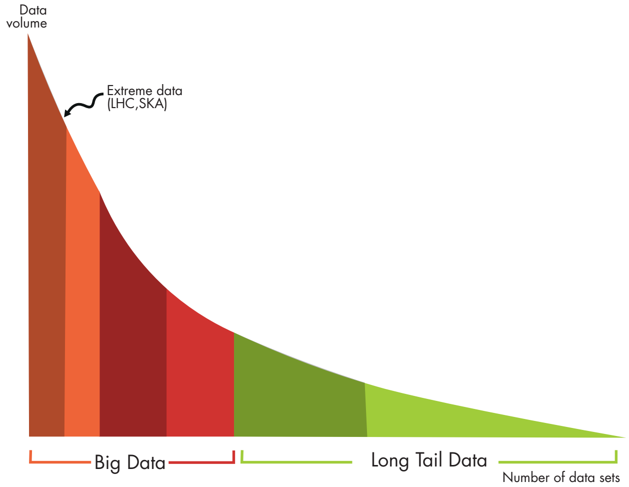


Figure 2: Big Data within the encompassing data landscape. Exceptionally high volumes of Big Data are characterised as Extreme Data

As shown in Figure 2, Big Data forms part of a broader spectrum of data. Volumes of Big Data, such as those generated by the LHC experiments at CERN, are at least an order greater than typical cases of Big Data. These instances are distinguished as Extreme Data to avoid skewing the general perspective of Big Data. Long tail data comprise a multitude of much smaller data sets. It is important to recognise that the skills, technology, governance and management requirements for these categories of data are fundamentally different.

> *The Big Data concept is well documented and is generally described in terms of a number of attributes. The most commonly recognised attributes are Volume, Variety and Velocity as shown in Figure 3, while other V's such Veracity, Value, Volatility, Visibility and Viability have also been suggested [24, 35, 11, 22, 36].*

The Volume attribute refers to the size of digital storage needed as measured in orders of bytes. Typical Big Data storage volumes currently range from Petabytes (1015 bytes) to Exabytes (1018 bytes) and it is expected that there will be 16 Zettabytes (16x1021 bytes) of useful data in 2020.

Variety concerns the structure and format in which data are represented. Formats include text, audio, image and video, while structure addresses how data are organised (structured or unstructured). Combining different formats of data is a major technical challenge and more so if data sources are heterogeneous or unrelated.

Velocity refers to the speed at which data are generated or transmitted with typical Big Data transmission speeds ranging from one to 100 gigabits per second. Veracity, coined by IBM, addresses the inherent trustworthiness of data. Uncertainty about the quality and integrity of data compromises the validity of results and, as a result, techniques such as data cleansing, data verification and data pre-processing are critical disciplines when working with Big Data.

A byte comprises eight binary digits and represents a single ASCII character; a binary digit is a zero or a one.

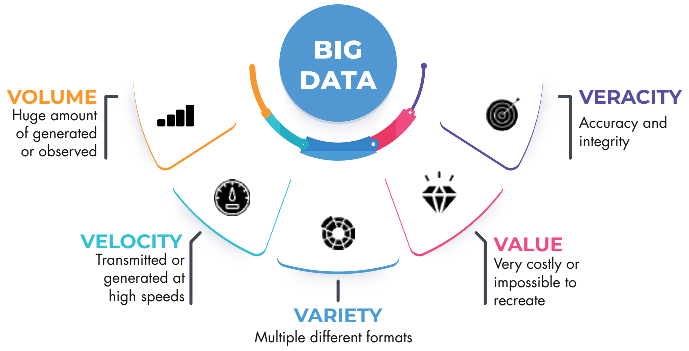

BIG DATA
VOLUME
Huge amount of generated or observed
VERACITY
Accuracy and integrity
VELOCITY
Transmitted or generated at high speeds
VALUE
Very costly or impossible to recreate
VARIETY
Multiple different formats

Figure 3: Commonly identified attributes of Big Data – Volume, Variety and Velocity are most prevalent

Each of these functional attributes presents fundamentally very different scientific and technological challenges. In some cases, for example the SKA, it is not practical or even possible to store all the data to be collected, even if current storage capabilities are increased by an order of magnitude. A similar situation holds for Velocity – it is not feasible with current technologies to transmit to remote locations all the data that will be received by the SKA antennas.

The processing and visualisation of Big Data have fundamentally very different and highly intensive data and computational requirements. A dataset that requires computationally intensive resources for processing can be regarded as Big Data; a data set that has national importance (Value) or that cannot be recreated without exceedingly high resource investments could likewise be considered to be Big Data.

Gartner Incorporated describe Big Data as "…information assets that demand cost-effective, innovative forms of processing for enhanced insight and decision making". For the purposes of this strategy, Big Data is regarded as data that have acquisition, storage, transmission, processing or visualisation requirements that exceed the capabilities of conventional technological means.

'Big' in Big Data is thus a relative notion that would change as technologies are developed for managing greater volumes, faster rates of ingestion and more diverse data formats. The upsurge in data volume is continuous and exponential, and addressing this phenomenon demands a fundamental shift in the way that data is viewed. There is general agreement that the size of the data universe will double at least every two years, i.e., a 50-fold growth from 2010 to 2020 and 10 times faster than the growth rate of traditional business data [22, 30].

Most of this growth stems from (a) rapidly increasing IoT, i.e., devices that continuously transmit observations or measurements from a myriad of sensors (cameras, telescopes, satellites, etc.); and (b) human-generated data, such as social media, transactional and other digital communications [37, 38, 39].

Vast volumes of data, estimated at 2.5 quintillion (2.5 x 1018) bytes by IBM, are created every day [40]. To give a sense of magnitude, this equates to 72 hours of uploaded videos, four million Google searches, three million tweets and 200 million emails created every minute.

The Internet, as the global medium of digital exchange, drives much of Big Data generation and value creation in the business and social world. Examples of business revenue generated per minute, by major multinational IT companies are, $ 54 190 (by Apple®); $ 24 690 (Microsoft®); $ 23 610 (Amazon®) and $ 17 610 (Google®) [41, 42, 43].

While the Velocity and Variety attributes pose technological challenges, Value is an important aspect, since some research data represents our history and culture. Data collections used in pan-African initiatives (such as climate change, Earth observation and bioinformatics) are key enablers of collaboration and should be regarded as Big Data.

2.1 SOURCES OF BIG DATA

While there is a multitude of data sources, five significant categories of Big Data sources can be distinguished [44]. Although the primary intent of acquiring these datasets may not have been research, some of these categories have secondary and serendipitous value as research Big Data.

Data generated for the initial purpose of research are regarded as primary research data and data collected for a different initial purpose are regarded as secondary research data [45].

This strategy concerns specifically research Big Data and not Big Data in the private realm.

- **Social media** are the most visible public sources of Big Data. While not always of high veracity, these data sources cross national, cultural, physical and demographic boundaries, and can provide valuable insights on public sentiments, preferences and changing trends. Predominant sources include interactive platforms such as Google®, Facebook™, Twitter™, YouTube™, Instagram®, as well as generic multimedia accessible through the Internet, such as images, videos, audios that provide quantitative and qualitative data on diverse aspects of human interaction.

- Public and private **cloud storage providers** host structured and unstructured data and provide on-demand controlled and interactive access to real-time data.

The flexibility and scalability of cloud storage make for easier and more economical Big Data access. Commerce and finance businesses rely on a variety of database technologies to manage vast numbers of daily transactions and provide competitive business intelligence.

- The **World Wide Web** hosts vast amounts of Big Data accessible through the Internet to individuals and companies alike. The global enormity of the Web ensures its ubiquity and benefit to start-ups and small and medium enterprises (SMEs).

- **Sensory devices** that emit data, range from smart phones and cameras to research apparatus, such as telescopes, remote-sensing satellites and radars.

Together with computers executing modelling algorithms, these instruments produce vast amounts of data every second. Data from some of these and other devices, such as medical instruments, vehicles, digital metres and household appliances, can now be sourced through the IoT.

While all of the above categories are potential sources, data collected from sensory devices, collected through and generated from algorithmic models for the explicit purpose of research, are the most significant sources of primary research data. Figure 4 illustrates how primary research data from phenomena and simulation models are collected, stored in repositories and then utilised in research that supports innovation.

Examples of primary research data sources are laboratory instruments and equipment such as satellites, radars, weather stations, air and soil chemistry sensors, telescopes and microscopes, and results of surveys for longitudinal and time series studies. These data collections are typically stored in silo repositories and, in some cases, with access only by a restricted group of users.

Considerable resources are invested in the acquisition of data in general and more so for the case of Big Data. These research Big Data collections are the principal focus of this strategy.

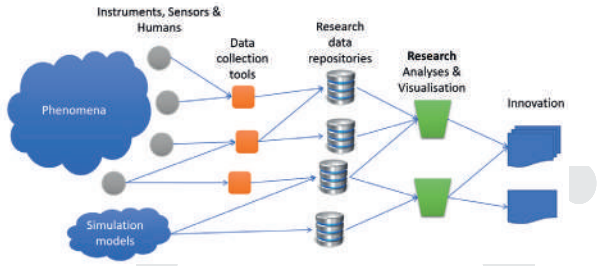

Instruments, Sensors & Humans
Data collection tools
Research data repositories
Research Analyses & Visualisation
Innovation
Phenomena
Simulation models

Figure 4: The primary research data workflow supporting innovation

2.2 THE BIG DATA VALUE CHAIN

A series of steps have to be taken to generate useful insights and value from Big Data. A generic flow of the major activities in the Big Data value chain is shown in Figure 5.

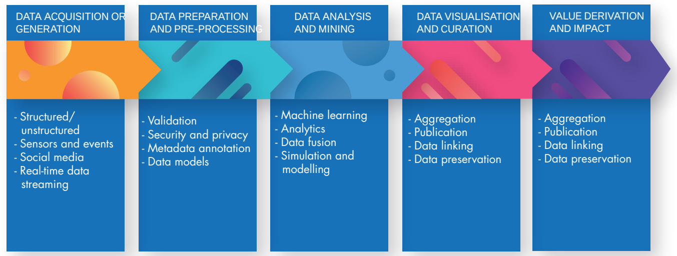


Figure 5: A typical flow of steps to create value from Big Data

- **Data acquisition and collection** is the process of creating or obtaining data from multiple sources, some of which may continuously transmit high volumes of data, while others result from simulation models or observations stored in files. Ingesting data at high speeds poses technical challenges, one of which is referred to as the Fast Data problem. As one practitioner remarked, *"it is akin to trying to drink from a fully open fire hose"*.
- **Data pre-processing and preparation** entails the transformation of data into formats and types that can be analysed and mined (cleaning), and can involve the determination of the veracity and fidelity of data (validation and verification), its quality and attempts to impute missing values. This process also includes the annotation of data with metadata, without which a dataset would be a meaningless set of figures or text.
- **Data analysis and mining** involves the application of advanced scientific techniques, such as statistical analyses, deep neural networks and other machine learning algorithms, to explore, model and ultimately, extract useful information. This processing can be very computationally intensive.
- **Data visualisation and curation** address two separate aspects. The first is to aggregate, summarise and render results in a form that is readily useful and that imparts information or knowledge. The second aspect refers to the active management of data and results over their lifecycle.
- The **derivation** of value includes improved decision-making and planning, the application of outputs to improve existing services or the use of these insights to develop new services and products, and influence policy development.

Importantly, all of these steps require expertise and skills, as well as advanced cyberinfrastructure technology capable of dealing with the attributes of Big Data. The synergy between Big Data and artificial intelligence is explicitly emphasised.

The application of machine learning and Big Data in science is relatively advanced and artificial intelligence, in general, provides opportunities to achieve broad impact from Big Data.

It is difficult to exploit the benefits of research Big Data without artificial intelligence (and to some extent, vice versa). A tangible example of the benefits of artificial intelligence as an HCD focus for Big Data is evident in the graduates of the SA-CERN consortium, who are now hired by leading companies globally. A research Big Data implementation roadmap should thus include the development of the artificial intelligence ecosystem as well.

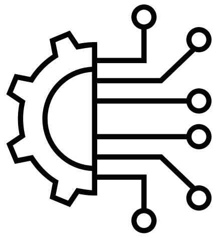
2.3 CHALLENGES OF BIG DATA

Data can be viewed as both an asset to be mined and part of the infrastructure for research and development.

However, there is a wide gap between the promise of Big Data and its realisation. Big Data poses difficult and novel problems – heterogeneity, scale, complexity and privacy impede progress at all phases of the pipeline that create value from data. Solutions to analyse and properly manage very large datasets do not readily exist.

Big Data technologies for data management, processing, analytics, discovery, and dissemination are different in scale and sophistication from those used for conventional volumes of data.

Traditional information and communications technologies, such as database management systems (Microsoft SQL™, MySQL, PostgreSQL); file management systems (NTFS, Mac OS, Ext4, etc.); and analytics applications (R, Matlab, Octave,...) have not been initially designed to cope with Big Data and are being supplanted by new versions or new big-data capable technologies.



Data Integration

The problems start right away during data acquisition when the deluge of data requires decisions, currently ad hoc, about what data to keep and what to discard, and how to reliably store what we keep. The value of data multiplies when it is linked with other data. Thus, data integration is a major creator of value but also a complex problem. Data analysis is a bottleneck, due to lack of scalable algorithms, and the volume and complexity of the data that needs to be analysed.

Finally, the presentation and interpretation of results require new approaches, in order to derive value from Big Data. Innovative and creative science is required since traditional technological solutions do not adequately address these challenges. There is also a need to re-assess the manner in which data is managed, end to end.

To address these issues, there has been renewed interest in, for example, data analytics and 'smarter', faster ways of processing (e.g., in computational intelligence disciplines such as machine learning). There is a further need to develop leading edge technology services such as e-Science enabling environments (e.g., virtual collaborative research environments, digital data libraries, Grid and cloud middleware) and to disruptively scale up data storage, transmission and processing infrastructures.

The ethical and social implications of Big Data are receiving heightened interest, with Society 5.0 being the emerging notion encompassing aspects such as privacy and cybersecurity. Emergent technology phenomena, such as the IoT and Industry 4.0 leverage, largely, a mature Big Data world. Some key challenges posed by Big Data are considered in the following sections.

2.3.1 CAPACITY AND EXPERTISE

The scale of Big Data poses complex and novel management and technical problems. Advanced technical skills are needed to address technical aspects and expert knowledge is required to mine knowledge from Big Data to create innovative new services and capabilities.

However, there is a global dearth of data scientists and researchers [34, 12, 11] and, for South Africa, this need is more acute, given projects such as the SKA. This lack poses a critical risk to achieving national goals and is the pivotal barrier to leveraging Big Data for more rapid socio-economic growth. HCD is needed along the entire Big Data lifecycle and the required expertise and skills go beyond traditional data science curricula. There are three major categories of expertise that can be distinguished:

- **End users** of Big Data cyberinfrastructure, i.e., researchers across all disciplines who are knowledgeable and competent in research data management, including data pre-processing (such as "data munging", formatting and annotation), as well as the use of appropriate Big Data analytical and visualisation tools;
- **Technical practitioners,** such as system architects, software engineers, hardware technicians, developers and support staff skilled at engineering, deploying, developing, maintaining and supporting leading edge technologies for Big Data; and
- **Data managers and support staff,** such as data librarians, data custodians and stewards, supporting users and ensuring that data is managed in compliance with regulations, policies and principles (e.g., Open and FAIR) along the entire data lifecycle (from acquisition or generation to curation or expunction).

There should be much closer engagement among mathematics, statistics, computer science and other departments, in order to respond effectively to the needs of Big Data expertise.

The traditional siloed operation of academic departments hinders collaboration among subject matter experts on Big Data as a multi-disciplinary topic.

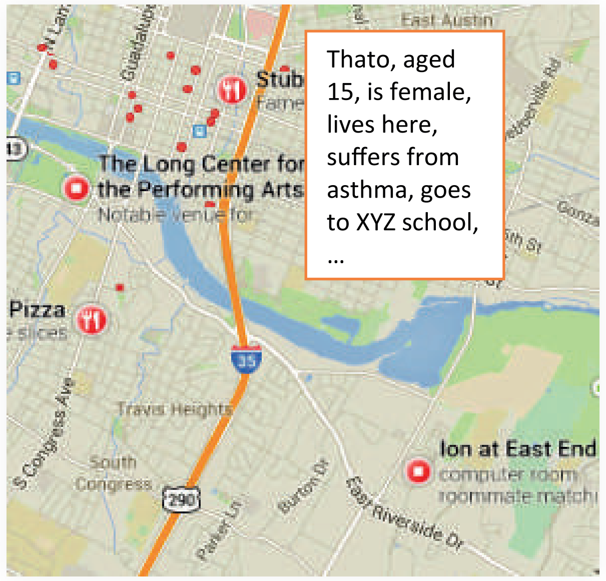


Figure 6: It is now possible to de-anonymise individuals using a combination of social data and other sources

The National Department of Basic Education has a role in building the human capital pipeline, but this is not within the scope of this strategy. Training and skills development in Big Data science will inevitably cascade into innovation needed to accelerate the development of knowledge-based business and industry in South Africa. However, a concerted and well-directed HCD approach for Big Data is required to address this challenge timeously. Interventions ranging from short courses and tutorial workshops, re-training programmes to undergraduate and postgraduate research programmes are examples of means to tackle this problem.

2.3.2 PRIVACY AND CYBERSECURITY

It is possible to utilise Big Data to, for example, improve the quality of health services; produce sector-specific weather and climate information, and forecast peak water and energy needs and consumption more accurately.

However, the use of Big Data can involve personal information, such as medical and utility billing data. For safety and security, Big Data also provides new means of surveillance and governmental organisations routinely collect personal data that could be used for purposes not originally envisaged [46, 47, 48].

Reports by the Big Data World Economic Forum and the McKinsey Global Institute, among others, raise concerns of privacy and security. Large amounts of data can now be collected more easily without consent or knowledge [49, 16]. Even if a dataset is effectively anonymised – and this is very difficult – if freely available it could be de-anonymised by merging it with other datasets [50, 51].

Analytics of Facebook™ data, for example, were 95% accurate in distinguishing African American from Caucasian-American users [52]. The potential for abuse is further heightened by myths of informed consent. Few people ever read and fully understand the full terms and conditions associated with Internet and social web services [53, 54, 55]. The implication is that it is very hard to guarantee privacy.
such as the South African POPI Act, 2013 (Act 4 of 1023), have been enacted, more directed policies are needed to regulate the ethical use of Big Data and further protect the privacy of the individual.

PREDICTIVE PROFILING

Big Data increases the likelihood of drawing erroneous inferences from spurious correlations. The Google™ Flu incident is the archetypal case– the US Centre for Disease Control combined air line records, disease reports and demographic data to track health risks.

By tracking the rate at which the public searched for terms like "flu" and "cough medicine", Google maintained that an outbreak of influenza could be spotted a week or two ahead of these reports [56]. This serendipitous use of Internet search data could not have been anticipated initially. Google search engine data have been similarly used to provide measures of unemployment and consumer confidence [57, 58].

To address these issues, the following questions have become more salient:

- What are the implications of Big Data for data breach and cybercrime;
- What are the implications for Big Data in deriving information about individual and personalised market profiles;
- How should we reconsider the "reuse" of data, i.e., the use of data for purposes not initially identified or even envisioned;
- Similarly, how should the initially unintended recombination of data be considered; and
- Who will regulate and monitor Big Data ethics (and who watches the watchers)?

2.3.3 CYBERINFRASTRUCTURE

Advanced, high performance and high-capacity cyberinfrastructure is needed to harness the benefits of Big Data. The exponential growth of data volumes and velocities introduces several technological challenges:

- Continuously and rapidly increasing data storage needs, together with power, cooling and space limitations (in an increasingly energy-hungry world);
- Growth in data movement, together with limitations in the transmission capacity of legacy communications networks; and
- Growth in the heterogeneity of data, together with diverse data management environments

There has been significant investment in the development and deployment of cyberinfrastructure. NICIS provides the national (Tier 1) cyberinfrastructure and coordinates the establishment of regional (Tier 2) infrastructures. Other mid- to high performance cyberinfrastructures are unavoidable. Institutions such as universities, research councils and government entities and departments would continue to build cyberinfrastructure in addition to Tier 1 and Tier 2 resources.

Some of these platforms would be dedicated to specific Big Data projects and disciplines (e.g., SKA Data Processing Centre, Western Cape Tier 2 Data node) to meet their operational obligations in some instances. Academic and research institutions struggle to afford the procurement and maintenance of cyberinfrastructure.

While funding models might encourage the development of infrastructures in a siloed manner, such uncoordinated efforts can lead to inefficiencies and duplication and, counteract collaboration among researchers. Moreover, technology ages rapidly and has a limited useful lifespan.

Shared cyber technology platforms present greater economies of scale and encourage collaboration. While there should be scaled up and sustained investment in information and communications technologies to manage Big Data, there has to be coordination among stakeholders to ensure effective and efficient investment of resources.

Hence, in the development of the national cyberinfrastructure, issues of the sustainability of such implementation should be considered. In summary, the major Big Data challenge is the lack of skills.

Other hurdles are the lack of appropriate technology, complexity, management, analytics, visualisation, privacy and security.

3 STRATEGIC DRIVERS

This strategy is informed by other national agendas, such as the NDP [59], the recently published White Paper on Science and Technology [6], NSI, HCD Strategy for Research Innovation and Scholarship [60], the 10-year Plan for Science and Technology [61] and the SARIR [59]. The relevance of and relationship to a Big Data strategy for these and other key initiatives are considered.

The global response to the novel coronavirus (Covid-19) pandemic eminently demonstrates the value of Big Data in mitigating and managing this disease. Monitoring, tracking, analytical and predictive modelling applications being developed depend critically on analytical skills to transform, integrate and visualise data of heterogeneous formats and sources.

Examples of such diverse formats and sources are unstructured data of cellular phone locations obtained mobile telephony companies, hospital records of patient admissions, mortality rates from the Department of Home Affairs, and spatiotemporally maps downscaled to provinces, municipalities and enumeration areas.

Research and development are at the core of overcoming the Covid-19 pandemic. An advanced level of so-called 'data wrangling' or 'data munging' competencies is required to effectively and timeously manage this disease. This strategy addresses the need for the country to attain a deeper and broader base of advanced skills and expertise. That is to develop a cadre of Big Data scientists and analysts, with access to suitably advanced cyberinfrastructure, who derive knowledge and develop more accurate models, not only to conquer this pandemic, but also to improve responses and attain a greater level of preparedness to similar challenges in future.

- The **NDP** serves to *"grow an inclusive economy, build capabilities, and enhance the capability … to solve complex problems"* [62]. Big Data presents precisely such an opportunity: in growing Big Data capacity, the technology solutions for transport, public health and other infrastructural services can be improved and a less resource dependent economy can be attained.

- Big Data innovation supports the transformation towards an economy in which the production and dissemination of knowledge leads to economic benefits as contemplated in the 10-year **Innovation Plan** [7]. This can be achieved through a concerted effort to develop a Big Data ecosystem comprising the enabling capabilities and resources.

- The **NSI** is a framework for a set of institutions, organisations and policies that seek to promote productivity growth, competitiveness and improvement of quality of life through innovation [63]. Innovation through Big Data initiatives is key to accelerated economic performance and can contribute to addressing social challenges such as health, food security and access to potable water.

- Big Data R&D impacts centrally on the **National Research Foundation Strategy 2020** to support knowledge generation and accelerated engagement with countries in Africa and globally. This strategy seeks to achieve "leading-edge research and infrastructure platforms" as an outcome of their Strategic Objective 4 [64]. Big Data infrastructure would include an instance of such a platform.

- **DHET** seeks to expand R&D and innovation capacity as Output 5 of their strategic plan (2015/16 – 2019/20) and an appropriate Big Data strategy that includes the development of expert human capital in the sciences and engineering for Big Data as an objective [65]. There is close alignment between this objective and the outcome of increased numbers of graduates as contemplated in several targeted indicators of the DHET strategy.

- The South African national **ICT RDI Implementation Roadmap** identifies world-class research as being central to a vibrant SME ICT industry, and Big Data RD&I fully supports this goal [8]. The roadmapidentifies Grand Science as an 'Opportunity' cluster. Astronomy, with reference to the SKA, and the bio-medical sciences are explicitly mentioned in this context and, given the data volumes produced in these disciplines, a Big Data strategy is clearly relevant in the fruition of this Opportunity.

- **SARIR** identifies several scientific domains, all of which potentially involve Big Data [59]. In order to ensure greater return and synergy, it is necessary to have a well-coordinated plan that scales up investment in Big Data research infrastructures.

- **NICIS** manages national cyberinfrastructure that serves as the primary technology resource for implementing Big Data R&D [66]. In addition to the provision of technological services, NICIS also coordinates Big Data research as a subset of research activities requiring advanced cyberinfrastructure.

DIRISA, as the data component of NICIS, has the role to advocate, enable, support and coordinate data intensive research. DIRISA's provision of cyberinfrastructure with services, policy guidelines and recommended practices for sound data management and coordination of data intensive research initiatives naturally include Big Data. The other two components of NICIS support the processing of Big Data through the provision of high-performance computing infrastructure through the CHPC, and data transport and security through SANReN.

All three components designed in an integrated manner, are critical to all facets of the Big Data value chain.

- **The SKA project is the preeminent** instance of Big Data research [67]. However, the scale of its data and technology requirements is at least an order greater than other Big Data initiatives (such as Bioinformatics and Earth Observation). For this reason, it is regarded as a special case of Big Data – perhaps, an Extreme Data project. Given the increasing globalisation of technology sharing, the SKA project is indicative of how science may be conducted in future

- Big Data outcomes support several of the **SDGs** – in particular, those of ensuring food security, promoting well-being, sustainable water and energy supplies, combatting climate change, oceans and marine and environmental sustainability [68].
  To provide integrated solutions for attaining these goals, it is necessary to combine data across diverse disciplines and of disparate formats.
  This is a complex problem addressed as a relevant Big Data research topic.

- The **NRDS** provides an eminently appropriate context for Big Data of having *"to ensure that as many of our people as possible master modern technologies and integrate them in their social activities, including education, delivery of services and economic activity. This relates, in particular, to communication and information technology* [69]"

- The **White Paper on Science, Technology and Innovation** released by the DSI in March 2019, captures a number of Policy Intents for a coherent and inclusive NSI, increased human capabilities and expanded knowledge enterprise that, among others, promote and support competitiveness, enhance the quality of life, develop human resources and notably promote a digital economy. Big Data is a catalyst for achieving these goals [31, 70].

- The **SA-EU Dialogue report on an Open Science Framework** presents a set of principles and guidelines for the development of a South African Open Science policy, many of which are fully supported by this strategy. In particular, this strategy echoes the principles of Open and FAIR, the urgency to develop expert and skilled human capital, and the formulation of policies as

recommended in the Open Science Framework. In further alignment with the framework, while this strategy focuses on all Big Data, it is primarily applicable to publicly funded Big Data. Globally, trends overwhelmingly indicate that Big Data will grow unabatedly along all dimensions of its attributes. Data volumes will continue to increase exponentially and, driven by the proliferation of IoT devices, data will be generated at progressively faster rates and in more diverse varieties. The strategies of developed countries (summarised in Annexure A) demonstrate a very substantial and increased commitment of resources to Big Data.

In 2016, The United States government committed $200 million to Big Data research and development initiatives [71]; in 2015, the European Union committed €2.5 billion to public-private partnerships, in order to “master Big Data” [72]; the UK Data Capability parastatal agency invested £ 189 million in Big Data and established an e-Infrastructure Leadership Council to advise government on infrastructure and skills needed to leverage Big Data opportunities [70]. Proportionate investments have been made by other countries such as Australia, China, Japan, Malaysia, Singapore and South Korea [73, 74].

Besides large science projects, such as the SKA and LHC, ongoing national initiatives, such as the recently adopted EU framework on Open Science, the African Open Science Platform, and the SADC Cyberinfrastructure Framework Initiative, are among the significant drivers of a Big Data strategy. A substantially increased national commitment is essential, in order to reap the benefits that can be derived from Big Data. It is also crucial to recognise the necessity for self-capacitation as the Big Data capabilities and technologies developed today will be the competitive resources of tomorrow.

Put differently, these commitments will position South Africa to be a competitive producer and contributor, rather than a consumer or receiver of technology products and services in the 21st Century.

4 BIG DATA VALUE AND IMPACT

Data have become an asset as important as labour and capital. As a value creator, the overheads are comparatively little in terms of capital and running costs and the potential benefit of Big Data extends virtually across all forms of endeavour. In the private sector, new services and products utilise Big Data to capture greater market share. Big Data is also a major driver of IT spending. Indicative figures for the US in 2017 range from $50 billion to more than $200 billion [75].

Examples of domains that readily lend themselves to Big Data innovation include finance and commerce, education, and health and transport. However, an enabling environment is at the core of realising the value of Big Data for both the public and private sectors. There is a broad range of such opportunities. The following are examples of research areas that have Big Data beneficiation potential:

4.1 OPPORTUNITIES

A number of Big Data initiatives are already underway in the R&D and academic domains. Beyond these, a number of public entities host Big Data collections in support of their activities. These activities have a key role in driving capacity development and the hosted data collections provide eminent opportunities for Big Data innovation.

- **Astronomy.** The South African Meerkat and SKA radio telescopes are foremost Big Data cases and present cyberinfrastructure challenges for collecting and managing data that are and will be collected and managed. South Africa’s role to lead the SKA project includes the establishment of an SKA Data Centre.

This activity, alongside the LHC, presents unprecedented challenges in that data volumes and data ingestion rates are typically an order greater than those of typical Big Data cases are. The development of expertise to conduct advanced data intensive research, and the skills needed to implement and maintain this Big Data infrastructure is a crucial requirement. The SKA and LHC are distinguished as Extreme Data instances to avoid these cases imposing a skewed perspective of the general Big Data domain.

BIOINFORMATICS

- **Bioinformatics.** The South African National Bioinformatics Institute (SANBI) hosts genomic data and research tools that can be integrated with many other growing gene expression databases abroad. Given the lack of suitable repositories in South Africa, the tendency to deposit data collections overseas increases the local access barrier and lessens the opportunity for South African and African researchers to firstly benefit from these resources.

- **Environmental Science.** The Southern African Environmental Observation Network (SAEON) provides access to a network of repositories of heterogeneous environmental data that can be integrated with resources such as weather and climate change projection information. Two SARIR Research Infrastructures (RI), namely, the Shallow Marine and Coastal Research Infrastructure and the Expanded Freshwater and Terrestrial Environmental Observation Network, have recently been located in SAEON. Both of these RIs require significant Big Data research platforms. SANBI hosts a wide variety of biodiversity datasets that are valuable and will continue to grow. In order to gain greater value, it is essential to not only enhance access, but also provide ICT environments that enable the integration of these data collections with those of other disciplines.
- **Particle Physics**. Experiments worldwide created what used to be the world's largest database of data. These volumes are exceeded by experiments at the CERN Large Hadron Collider. Locally, NICIS hosts data used in the ATLAS and ALICE experiments and these volumes will increase substantially. Local researchers also depend on very high bandwidth interconnectivity to partake in these experiments, and challenges such as the selection of data, software to analyse the data and long-term storage of data should be addressed. The lessons in managing Big Data volumes and the provision of collaborative virtual research environments will inevitably cascade into other fields.

- **Medicine and Health.** The introduction of the National Health Information System will cater for the digital storage of medical imaging records, among others, which will impose consequent Big Data storage requirements. One example is the SARIR-supported distributed platform for "omics" research (DIPLOMICS) infrastructure that generates and hosts genomic sequencing data that are shared among researchers across Africa and that require petascale storage and curation facilities.
- **Social Sciences and Humanities.** The preservation needs for historical and archaeological artefacts such as paintings and fossils are being digitised and storage requirements for these datasets will continue to grow. A case in point is the management of digitised collections of paleontological discoveries, such as the Homo Naledi fossil finds. These resources, together with other population data, such as census data, constitute key assets and are regarded as Big Data, although not necessarily of high volumes.

The SARIR South African Population Research Infrastructure Network is a particular example, being a research initiative that produces an ICT platform that hosts longitudinal and demographic data for inter-disciplinary research on improving the wellbeing of poorer communities.

- **Earth Observation and Space.** The SANSA hosts repositories of remote sensing data and other space-related data, such as provided by the Hermanus Magnetic Observatory. The SAWS is a custodian of weather and climate data and hosts data from a weather radar network, the lightning detection network, as well as a range of weather stations from across South Africa and some islands.

Several projects are afoot for developing digital libraries of these datasets, allowing for their integration with in-situ and airborne data. Weather data is integrated with social data, including population distribution, infrastructure and urban environments to provide impact-based forecasting (i.e. what the weather will do as opposed to what the weather will be).

- The NRF regards **Indigenous Knowledge Systems** (IKSs) as a resource for "innovation and entrepreneurship". The National Indigenous Knowledge Management System supports the processes and structures developed through the National Recordal System and is responsible for the recording, storing, management and dissemination of indigenous knowledge information. While DIRISA is working closely with the DSI on hosting this system, there is a clear need to address the curation and continued maintenance of such data collections. Given their historical and cultural value, these data collections warrant recognition as Big Data.

Data generated through publicly funded research projects, such as the South African Research Infrastructure Roadmap and of the national research facilities, should be considered as Big Data, given their value and variety.

In general, scientific research data collections can be regarded as Big Data for several reasons.

Novel techniques, technologies and management processes are needed for research data to be collected, prepared and preserved in a manner that allows these resources to be discovered, shared, combined and mined beyond the confines of a given discipline.

In addition to the instances above, some government entities and departments, and research institutions are custodians of valuable datasets (e.g. weather, water resources and energy). While some may not be openly accessible, these assets warrant regard as Big Data for preservation and curation purposes.

A further aspect is that e-Science data and scholarly publications are stored and offered through digital libraries require management and data-oriented services, such as visualisation and transformation. These resources will grow and their use will impose substantial computational and storage demands. An example of such growth is the generation of digital artefacts resulting from a joint project by the University of the Witwatersrand and the University for Fort Hare to digitise hardcopy archives of the political struggle and the activities of liberation. Large volumes of data are generated through social media.

While the initial purpose of this data is not necessarily research, the combination of these datasets with other collections can provide valuable new insights and services. Social media-generated data do not necessarily introduce storage requirements, but do impose computational requirements for analytical purposes. The provision of such computational resources to the public would eminently affect citizen science. As a demonstration of benefit, the South Korean Big Data Strategy Centre and it's Big Data Institute, public-private partnerships resulted in outputs such as optimised networks of only eight bus routes covering 49% of demand and the reduction of the water leakage rate from 79% in 1989 to 2.5% in 2014.

The private sector, business and industry are at the forefront of monetising the Big Data phenomenon. A Gartner Survey shows that more than 75% of companies are investing or are planning to invest in Big Data in the next two years [76]. The ICT and communications industries have been reshaped by Big Data, offering products and services that support both enterprise and individual deployment of Big Data management systems.

- In **manufacturing,** processes rely on Big Data provided through IoT to automate and individualise manufacturing. Results of predictive modelling of demand are used to refine just-in-time production schedules and precisely target production and effective distribution. Within business, in general, Big Data analytical results are used to achieve improved logistics and operational efficiencies.
- In **healthcare,** IoT sensors provide data of vital health signs for remote clinical diagnostics and can monitor patient medication individually in real-time monitoring systems. A single view of patient data enables location-independent care; the analysis of disease patterns can provide predictive and preventative information; vaccine and antidote discovery is supported by machine learning modelling and simulation techniques. other fields.
- **Intelligent transport** management systems can monitor and control transportation networks and alleviate traffic congestion using a variety of data sources and the IoT. Big Data technology drives efficiencies in mapping and route planning with enhanced potential to reduce CO2 emissions
- **For water and energy** utilities, the digitisation of sensors as IoT devices provides high-resolution real-time measurements of consumption and, through advanced analytics, support the improvement of levels of efficiency within both the demand and supply sides of networks. Sensors in smart buildings and smart cities will be key Big Data providers.
- In **finance and insurance,** Big Data can be used to identify exposure in real time across a range of financial instruments. Predictive analysis of both internal and external data results in better, proactive management of a wide range of issues from credit and operational risk (e.g. fraud and reputational risk) to customer loyalty and profitability. By combining their customer transactional records with publicly available social media, banking institutions refine and individualise their services and identify security risks more accurately.
- **In telecommunications, media and entertainment,** Big Data transmission and analysis techniques enable the effective discovery and delivery of media content, allowing users to interact with media content across multiple platforms in a dynamic manner.
- **In retail and wholesale,** Big Data is used to increase productivity and efficiency, resulting in increased operating margins. Big Data affects retail in other areas, such as location-based marketing, in-store behaviour analysis, customer micro-segmentation and customer preference profiling. In general, personal sentiment and location data offer new value creation opportunities with applications across many forms of digital service delivery. This includes smart personalised routing, automotive telematics, mobile location-based services and geo-targeted advertising.
- **In e-agriculture,** data from, for example, on-board spectral cameras carried by drones, provide data that are useful for monitoring crop health and irrigation management if combined with other near real-time streams of data such as remote and in-situ sensory data, such as soil chemistry, weather patterns, new fertiliser techniques and new strains of seeds.
- **In e-Government,** citizens can be provided with a more effective and holistic services ranging from education, transportation, medical, taxation, voter's identification to municipal services if a single view of the citizen can be engineered as a Big Data subject.

There are many opportunities to advance the digital economy and address societal challenges through Big Data research.

Big Data technologies are transforming industry and commerce, and similar benefits prevail for government in cost savings and more efficient delivery of services.

Big Data research would create additional value across all of the aforementioned areas. Such research will result in the development of management tools, data cleaning techniques, suitable techniques for improving data integrity, improved methods for handling heterogeneous datasets, and improved analytics and visualisation technologies for Big Data. There are also challenges that should be addressed. In many cases, data collections cannot be regenerated without incurring excessively high costs and, in other cases, they cannot be regenerated at all. It is essential to have the sufficient infrastructure to preserve and extract greatest value from these resources.

The benefits of Big Data are aptly demonstrated in research currently underway, to manage the novel Covid-19 pandemic. Countries such as the United States of America and China, and those in the European Union are using Big Data research to analyse and develop models that more accurately predict its spread and to develop vaccines to counter this pandemic. The use of Big Data technologies can link collected information about the incidence, geographic spread and use of resources for integrated data analytics.

The results of such analyses help authorities to quickly make more effective decisions to mitigate and manage this pandemic. A significant benefit of research Big Data lies in the quest for vaccines and improved therapeutic treatments for Covid-19.

High performance computing resources are used for simulations to research and develop pharmaceutical compounds that may result in effective treatments of this disease. Some of the Covid-19 Big Data projects supported by NICIS are summarised in Table 1.

Table 1: Some of the Covid-19 projects supported by NICIS

Project name, collaborators	Description	Outcomes, impact
SADC Covid-19 Collaboration NICIS, DSI, SKA, ISC -Intelligence in Science, AERAP	Data Science capacity in Africa and synergies on initiatives for collaboration to respond to Covid-19.	An ICT platform that SADC countries can use to share data. NICIS to host the platform.
CMORE DASHBOARD NICIS, CSIR NextGen Enterprises and Institutions	A platform used by government to monitor the evolution of Covid-19 cases across the country.	A dashboard that provides Covid-19 statistics for the country.
Mobile Data for Movement Tracking NICIS, NextGen Enterprises and Institutions, DPSS	Determine patterns of movement using anonymous mobile data.	The tracking of movement between regions. In the long term, used for tracing (compliant with legislation).
Online Learning Support for DHET NextGen Enterprises and Institutions, state-owned enterprises	Provide online learning connectivity access for DHET students, while they are at home.	Facilities, mobile coverage and students' location. Solutions for connectivity to enable student access.
Therapeutic Solutions National Institute for Communicable Diseases, university researchers	Modelling and simulation of therapeutic solutions.	More effective therapies for Covid-19.

5 THE BIG DATA ECOSYSTEM AND NATIONAL LANDSCAPE

The scope of this strategy is limited to research Big Data and directed at realising the economic, social, scientific and industrial beneficiation potential of Big Data. This strategy serves, firstly, to inform and guide decision-makers in the national research Big Data ecosystem. Secondly, it provides the private sector with a view of national priorities for research Big Data, allowing business and industry to collaborate more closely in support of national goals, and it provides industry access to research capacity and human capital and skill sets at research institutions. The research and innovation objectives align with and support strategic agendas, serving to inform decision-making, investment, policy formulation and the planning of initiatives related to research Big Data. This strategy is impacted by national priorities, global trends and local initiatives that are key drivers of Big Data. Big Data introduces a set of challenges that are not altogether new. However, the scale and complexity raised by these challenges pose new problems across technical, managerial and governance domains. Big Data ecosystems form in different ways around organisations, communities and within or across sectors. Examples of sectoral ecosystems are healthcare, finance and manufacturing and each of these interleave with the domains mentioned in unique ways. The benefits of sharing and linking Big Data across domains and sectors are clear. Initiatives such as smart cities are demonstrating how different sectors can maximise return. The cross-fertilisation of stakeholders and datasets from different sectors is an important key element to achieve such optimal value. A vibrant ecosystem, comprising key stakeholders and actors with a common perspective of strategic goals, is a key element of an enabling environment to realise the benefits of research Big Data. The categories of actors in a Big Data ecosystem include data custodians, data stewards, technology providers, end users, researchers and academia, regulatory bodies and investors. These are grouped into actor and stakeholder categories in Table 1.

Category	Description (with some examples)
Owners and sponsors	Government ministries and public agencies investing in Big Data research (DSI, DHET, the Department of Trade, Industry and Competition (**dtic**), NRF, Technology Innovation Agency (TIA), Department of Environment, Forestry and Fisheries (DEFF))
Users	Researchers within research communities, research infrastructures, national research facilities, academia, science councils and agencies (SANSA, SAEON, iThemba Labs, SAWS, etc.)

Table 2: *Categories of stakeholders in Big Data*

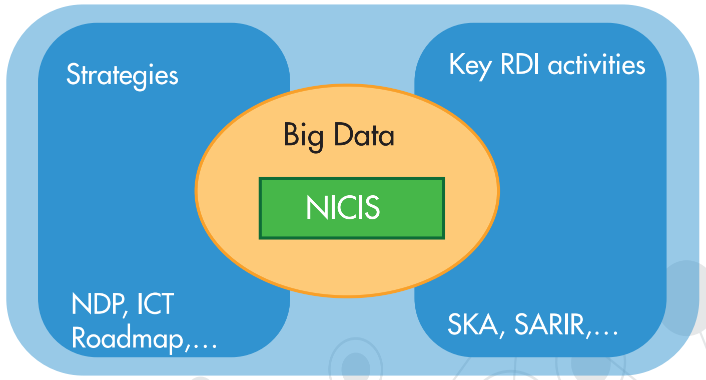
Big Data custodians and stewards	Research councils, higher education institutions (HEIs), national research facilities, government departments and entities
Regulators and governors	Policy makers and legislative bodies (e.g., NIPMO) and data stewards ensuring that Big Data is utilised in a compliant and ethical manner
Educators and trainers	HEIs, TVET colleges
Cyberinfrastructure and service providers	NICIS, HEIs, research institutions, science councils, industry and business, national research facilities of the NRF, SARIR RIs
Joint venture partners	Industry and business, science councils, academia, government
Strategic projects	SKA, SARIR
Beneficiaries	Civil society and industry (private and public sectors)

These actors and stakeholders may also be represented at different national, provincial and local strata of government; SOEs and other supported entities. In the private sector, business and industry have important roles such as providers, beneficiaries and collaborators.

The academic sector has a vital function in advancing the frontiers of knowledge in Big Data, and in developing the skills and expertise needed to apply this knowledge for societal benefit.

The European Commission regards a successful Big Data ecosystem as *"stakeholders interacting seamlessly within a Digital Single Market, leading to business opportunities, and easier access to knowledge and capital"*. From a national context, three main actors are key: public sector organisations, i.e., governmental and supported entities, determine national imperatives and priorities that are conventionally expressed through policies, strategies and frameworks. This actor is depicted as *"Strategies"* in Figure 7.

Figure 7: *Main actors in national Big Data ecosystem*



Organisations that are Big Data custodians or that conduct Big Data research are indicated as *"Key RDI activities"* in the same figure. While the SKA is the predominant Big Data initiative, other activities such as the national space initiative managed by SANSA, research in high-energy physics, climate change, bioinformatics, the paleo sciences and humanities, are also relevant. While the private sector may provide opinions on cyberinfrastructure and services for Big Data, NICIS, as the third actor, would have a central role in this regard.

5.1 STAKEHOLDERS AND ROLES

The stakeholders of a Big Data strategy range from the research community, funding agencies, business and civil society, to national and local government agencies involved in technology innovation and policy-making.

Their engagement will help marshal their participation in and contribution to the implementation of a strategy.

The roles of some key stakeholders, as listed in Table 3 (in no specific order) are taken into account. This list may be non-exhaustive and would be adapted as new stakeholders are identified. Government departments and their supported entities that invest in Big Data, as well as representative organisations of academic and research institutions are notable groups. Specific instances of some are universities, research councils, the DCDT, the Presidential Commission on the Fourth Industrial Revolution, the NRF and NICIS. A structure is needed to represent at least the following key stakeholders:

- **DSI,** guiding policy development and cohesively linking with other national strategies and activities, and developing a sustainable Big Data investment plan;
- The **NRF** and **TIA**, supporting and implementing R&D and innovation strategies, funding, research development and support;
- **DHET**, informing research and educational interventions for Big Data HCD; and
- **NICIS** as the national cyberinfrastructure provider and coordinating the implementation of a Big Data strategy.

Table 3: *Stakeholder roles in the Big Data ecosystem*

Stakeholder	Role
DCDT	Strategic alignment with policy frameworks and ICT strategy
PPC4IR	Strategic alignment with 4IRPC mandate and national strategies and supporting the digital economy
Department of Public Service and Administration	Public services, frameworks and norms that regulate the ethical use of Big Data
Department of Trade and Industry and Competition (**dtic**, merged into Economic Development Department (EDD))	Industrial and international Big Data development and innovation; the digital economy
Government IT Officers Council	Alignment with, and support of the implementation of ICT strategies
SITA	Public sector ICT services; governance of research Big Data in the public domain
Statistics South Africa	Custodianship of public census and survey Big Data; norms and standards for the ethical use of Big Data
Department of Women, Youth and Persons with Disabilities	HCD and transformation
EDD	4IR, Digital economy and social economy
Government Communication and Information Systems	Intergovernmental coordination of Big Data R&D
Department of Water Affairs	Custodianship of research relevant public big data
Department of Health	Big Data custodianship; R&D
South African Qualifications Authority	Transformation and HCD for Big Data R&D in the academic sphere
CHE	Higher education and training – capacitation of universities and colleges
USAf	Universities, training and R&D programmes
SKA/SARAO	Big Data cyberinfrastructure; R&D
iThemba Labs	Big Data custodianship; R&D
Water Research Commission	Big Data custodianship; R&D
Medical Research Council	Big Data custodianship; R&D
Council for Geoscience	Big Data custodianship; R&D
Human Sciences Research Council	Big Data ethics and governance
NRF and national research facilities	HCD and R&D
SAWS	Big Data custodianship; R&D
ARC	Big Data custodianship; R&D
DEFF	Big Data custodianship; R&D
National Disaster Management Centre (NDMC)	Big Data custodianship; R&D

6 VISION AND STRATEGY

The crux of this strategy vests in collaborations that stretch across the RD&I lifecycle from basic research to the operationalisation of research outcomes, with a focus on relevant and emerging topics. Capacity development must form an integral part of such activities. The following vision is aimed at deriving greatest benefit from the many rich and diverse sources of Big Data.

> *We envision a Big Data ecosystem, deriving from large, historic and diverse datasets, knowledge that leads to innovation and accelerated socioeconomic growth and that, in turn, positions South Africa to be more competitive in the 21st Century Big Data economy.*

This strategy should comply with several core values, in order to realise a vibrant Big Data ecosystem. Put differently, these requisites are essential for the successful implementation of this strategy. A set of principles underpins and guides the successful achievement of this vision:

- **Inclusivity (Engagement).** There is constructive engagement with all key stakeholders and beneficiaries in the implementation of a Big Data strategy. Funding, custodial, research community, important national projects and private sector are represented.
- **Privacy.** Policies regulate the ethical use of Big Data and preserve the privacy of individuals as represented in Big Data and data in general.
- **Governance.** There are clearly defined roles and responsibilities for the actors in the governance of Big Data. Governance structures oversee and promote coordination across Big Data activities.
- **Infrastructure.** NICIS serves as the Tier 1 platform in a sustained manner, for Big Data initiatives and collaborates with other infrastructures used for specialised Big Data projects.
- **FAIR.** Data generated using public funds should be "as open as possible and as closed as necessary", in order to support findability, accessibility, interoperability, and reusability.

6.1 STRATEGIC OBJECTIVES

The following strategic objectives, set to achieve the vision of this strategy, are presented in some detail:

- **HCD.** Develop the next generation of R&D expertise and skills to extract and exploit, for innovation purposes, knowledge from Big Data and to advance the frontiers of Big Data science
- **Cyberinfrastructure.** Provide a sustained and advanced national cyberinfrastructure with allied services that enable and support Big Data RD&I.
- **Collaboration.** Foster a thriving and collaborative ecosystem of Big Data RD&I that links government, academia and the private sector.
- **Data Governance.** Develop and adopt standards and policies supporting the practices of Open Data and Open Science principles and ensuring compliance with privacy, ethical and legal regulations
- **Overarching coordination.** Maintain an overarching and coherent national approach to investments in Big Data initiatives and activities.

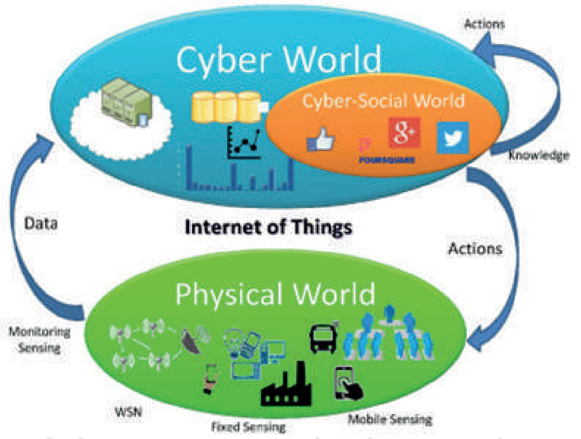
6.2 OBJECTIVE 1: HUMAN CAPITAL DEVELOPMENT

Develop next-generation R&D expertise and skills to derive and exploit knowledge from Big Data and advance the frontiers of Big Data science

It is important to recognise that Big Data capabilities are the competencies for the next generation of information and communications technologies. These competencies support digital transformation and are needed to develop the tools and services for South Africa to respond competitively to disruptive opportunities such as the 4IR and future technological developments. A capable cadre of scientists, technologists and engineers is needed to positioning South Africa to partake as a producer, rather than as a consumer, of 21st Century technologies.

Petabyte and exabyte volumes of data will become more conventional and new computing and data analytics technologies will emerge to deal with these scales of data. Much innovative research in topics, such as deep learning will be needed to manage interconnected repositories of structured, unstructured and fast data. Managing the convergence of technologies such as the IoT, Cloud, "Smart" technologies, Cognitive systems, Cyber-physical systems and Array (noSQL) databases with Big Data (e.g. Figure 8) requires a new breed of student, researcher and practitioner.

Figure 8: Convergence of technologies with Big Data



Given the complexity of Big Data, there are key roles for metadata, persistent identification of datasets, and visualisation to communicate insights easily. There is a need for further research and innovation to provide timely and defensible results from opaque systems. This will require multiuser, multi-stakeholder engagements that are equipped with the necessary collaborative environments and tools.

Existing algorithms may not run in acceptable timeframes and users may rely on tools delivering Approximately Correct and Acceptable solutions. A new class of algorithms is needed to deal with the volume and speed at which data must be processed. While Analytics is often presented as the central topic, the lack of capability prevails across the entire Big Data R&D lifecycle.

Large-scale data management, acquisition, storage, processing, analytics, visualisation and secure transmission knowledge extend beyond traditional data science curricula. A concerted and focussed effort, that involves academia and research councils, should be made to rapidly grow the number of practitioners and researchers in the Big Data sciences. Many academic institutions are introducing programmes aimed at redressing this situation. However, this strategy must include an explicit and high priority target to fast-track the development of requisite skills and expertise.

Interventions that align postgraduate and post-doctoral programmes with collaborative research projects are needed to develop algorithmic solutions capable of dealing with the scale and complexities of Big Data. Ongoing initiatives, such as the National e-Science Postgraduate Teaching and Training Platform, should be extended along two directions – vertically, into a contiguous pipeline from

Baccalaureate to post-doctoral levels, and horizontally across disciplines from science, engineering and technology, to the humanities, education, health and the social sciences.

A concerted effort should be made to capacitate HDIs to ensure that they are not left behind. This can be achieved by ensuring that all funded Big Data projects require that a significant portion of HCD be directed at cultivating greater expertise at HDIs. Awareness and skills migration tutorial events for supervisors, research staff and postgraduate students (e.g., pre-doctoral workshops) can be developed in collaboration across faculties and departments. Certification (such as the e-Science Master's degree) and other incentives should be provided for re-skilling and up-skilling academia and researchers on Big Data science.

More consortia modelled on NEPTTP, straddling universities and universities of technology, should be established for training in the sciences, application and entrepreneurship of Big Data. Universities such as Sol Plaatje should bridge the gap between secondary and tertiary levels with programmes in the application of Big Data technologies in other disciplines, as well as in the development and support of the technologies underpinning the use of such technologies. The offerings of mathematics, computer science, and other relevant departments at universities should be combined in programmes that effectively address Big Data in its own right, and as a component of Data Science. Combined funding schemes should be arranged for academic training, reskilling or cross skilling across universities, universities of technology and TVETs. Big Data projects, such as the SKA, and those within SARIR are fitting incubators of skills that are applied to other domains.

Given the constraints in financial resources within the NSI, a collaborative effort of drawing on the skills available within the country to form a virtual or centralised institute that will deal with all aspects of Big Data is necessary. The role of the reconfigured and reconstructed National Institute of Theoretical and Computational Sciences (formerly the National Institute of Theoretical Physics) is an appropriate model to establish an institute to oversee advanced research and the development of skills and expertise for the country. This institute can be managed as part of the HCD pillar of NICIS. Close collaboration should be achieved with industry in developing skills-building programmes such as IBM Digital Nation Africa, BCX Data Science and Dell ICT Academy.

Technology transfer should be explicitly defined in this context and NICIS, together with consortia such as ILIFU and the SA-CERN consortium, should expand opportunities for much closer collaboration with industry.

6.3 OBJECTIVE 2: CYBERINFRASTRUCTURE

Build and sustain world-class and relevant national cyberinfrastructure with allied services that enable and advance Big Data RD&I.

Big Data encompasses a range of data scenarios from large and rapid data streams to highly distributed and heterogeneous data-collection networks. Applications require high-performance and complex processing with fast access to large repositories and archives of possibly distributed stored data. Big Data applications must deal with data from multiple sources that may be heterogeneous in various ways.

Core technology infrastructure is a key next-generation capability for Big Data. There has been significant recent investment in high-end computing resources with the deployment of the petascale Lengau high performance computer located at the CHPC. However, existing computational, data storage and transmission resources are already oversubscribed, demonstrating the acute shortage of these resources. The planned upgrade to ten petaflops could cater for immediate need though will unlikely accommodate future computing needs. A range of computer system architectures is required to serve a wide range of application requirements. Large system configurations with high-speed network interconnects and deep storage hierarchies are required to interactively process large-scale and high-speed data. Entities with strict operational obligations such as SAWS may still need in-house capacity for operational purposes, while systems such as those hosted by CHPC can be used for research purpose.

Dedicated networks are needed to transport large volumes of scientific data generated at experimental facilities (such as the Large Hadron Collider at CERN and the SKA), to distributed computing resources for analysis.

Data management bottlenecks can occur at almost every stage of the workflow, including capturing data from an experimental or computing facility, transporting it for further analysis, analysing and visualising the data, as well as finding appropriate environments for sharing data. A range of large-scale system architectures is needed to cater for a diverse combination of computer-, data- and communications-intensive applications.

Sophisticated hardware and tools are needed to organise data into human- and machine-readable summaries in a timely fashion.

Some of these environments are characterised as cyber-physical systems because they integrate the computational paradigm with physical components. Cyberinfrastructure should also support the management and analyses of vast amounts of data of real-time data streams from IoT as a rapidly emerging source of Big Data. The bulk of Big Data is generated from three primary sources, namely, social data, machine data and transactional data. Social data are uploaded by humans through media platforms such as Twitter, Facebook and Instagram and, while not primarily for research, have significant potential value for this purpose. Machine-generated data are generated primarily by electronic devices such as sensors and computers. Data from this type of equipment constitutes the bulk of primary research data. Transactional data are generated mainly through business and financial activities.

New software technologies for handling the breadth and scope of data from these sources will be essential for many Big Data applications. Beyond the hardware equipment, it is also necessary to deploy technologies such as cloud-based platforms that provide the environment for users to efficiently utilise cyberinfrastructure for Big Data analysis. One critical technology component is named *Entity Identification.*

With the proliferation of IoT sourced-data, it is crucial to persistently identify the myriad of IoT devices, as well as the datasets they produce.

Some of the approaches, such as Hierarchical Storage Management, as well as Edge and Fog Computing, places the data strategically at point where they are available and processed at an optimal location. There is a need to design and strategically invest in the data infrastructure in a manner that will optimise the access, transmission and processing of data.

To address the challenges of Big Data, there should be sustained and increasing investments in infrastructure for large-scale data collection, management and analysis. An appropriate mix of large-scale computational, storage and data transmission technologies, together with suitable digital environments should be deployed and maintained, in order to provide all users with the necessary tools to convert Big Data to actionable insights.

In order to determine an optimal ICT infrastructure nationally, it would be necessary to determine Big Data requirements across research infrastructures and other priority R&D initiatives. A roadmap should be developed to build a well-coordinated national cyberinfrastructure that encompasses significant activities and investments. This activity forms part of the role of NICIS that should play a leading role in coordinating this effort. Key stakeholders and major players should be fairly represented in advisory and steering structures, in order to achieve an effective and relevant national architecture for cyberinfrastructure.

6.4 OBJECTIVE 3: COLLABORATION

Foster a collaborative ecosystem of Big Data research and innovation that links government, academia, civil society and the private sector to leverage the value of Big Data

The Big Data ecosystem comprises business and industry, citizenry, government, academia and the research community. Flourishing Big Data PPPs delivering Big Data solutions are key outcomes of a successful strategy. Such cross-sector Big Data collaborations involving the research community, the general public and governmental agencies will contribute to a more prosperous socio-economic environment.

In order to build local expertise and to become innovators and inventors, it is necessary to build systems from the ground rather than remain dependent on external resources. Collaboration is a key enabler for achieving a higher level of competitiveness and there are opportunities to leverage international relationships. However, several impediments to closer collaboration are recognised. While the shortage of skills and expertise is foremost, the lack of sufficient cyberinfrastructure and data policies is impeding accelerated progress. These hurdles can be addressed through a range of collaboration mechanisms from consortia arrangements to the sharing of resources such as cyberinfrastructure, skills and expertise. Consortia should be formed to collectively pool and synergistically utilise resources for Big Data science that address national questions for greater an accelerated impact.

Many large datasets originate from industry and are not readily accessible due to the perceived potential loss of competitive advantage when exposing such data publicly. These datasets contain information that can provide valuable insight into solutions that improve the quality of life of the public.

Transport and disaster mitigation are examples of services that can be improved when mobile phone locality information is combined with weather data. Mechanisms and incentives should be established to create opportunities to collaborate with government and industry by reducing bureaucratic hurdles for technology and data sharing. One such possible mechanism is the creation of structures to coordinate collaborative projects across governmental, industrial and academic boundaries. Such structures can also serve to create greater awareness, and advocate for the adoption of Big Data research more broadly.

An example would be Big Data Innovation Hubs., i.e., centres supporting Big Data-based needs of different ministerial departments.

Support for and collaboration with SMMEs by universities, research councils and industry will increase the likelihood of their success and sustainability. Business companies should be encouraged and incentivised to set up innovation centres that engage government and academia. There should be joint solicitation by government and industry for research projects to leverage Big Data for the public good. One instance could be the combination of IT companies with universities, the SAWS, NDMC, SANPS and CSIR, together with some universities, to research and develop bespoke solutions on the NICIS cyberinfrastructure. Another example would be the formation of Big Data Innovation Hubs (BD Hubs) together with NICIS and SKA-SA that focus on grand science challenges and help determine the datasets, analysis tools and interoperability requirements necessary in achieving key national priority goals. A further example of collaboration across the public, private and academic sectors is the involvement of HEIs and academic hospitals with insurance companies and medical aid companies to develop innovative health solutions.

The national (Tier 1) and regional (Tier 2) data nodes can provide platforms for Big Data Research Hubs in support of collaborative 'triple helix' programmes. Such centres can provide "R&D test beds" and "sandboxes" with low barriers of access, enabling the research of tools and services that support innovation and the digital economy.

A programme supporting combined capabilities for research would allow for quicker production of results that yield relevant products and services.

Further afield, collaboration with other countries on projects that stretch beyond national borders, such as an Open Research Cloud (a Big Data research commons) would foster interoperability and provide long-term sustainability of such platforms. Shared test bed infrastructures will help maximise the value of investments, and the sharing of knowledge and resources that would otherwise remain isolated within a particular country or agency.

6.5 OBJECTIVE 4: DATA GOVERNANCE

Adopt standards and policies supporting the practices of Open Data and Open Science principles and ensuring compliance with privacy, ethical and legal regulations -

The value of data increases manifold when combined with other data sets or when repurposed for uses that differ from the initial purpose for which it was collected. Data that is easily discoverable for reuse and, when combined with other multiple sources, provides more knowledge and greater insights than when used in an isolated and silo manner. Moreover, the roles and accountabilities of various actors involved in data – funders, owners, custodians, hosts, end users and policy makers – should be clarified.

The scale and heterogeneity of Big Data present significant challenges in data sharing and for the preservation and protection of individual privacy, in particular. For example, it is possible to identify individuals by combining previously anonymised data sets. A further consideration for data, in general, is its previously unanticipated use. Combined with other independently collected datasets, the privacy and security implications of integration and repurposing or reuse are largely unexplored.

A mature and common understanding of these concepts is needed to help improve the integrity of use and of published results. Shared benchmarks, standards and metrics are necessary to ensure that confidential information remains secure and the use of data, in general, is ethically and legally sound. The POPI Act, and the General Data Protection Regulation are aimed at protecting the privacy of the individual, and have greater applicability for the case of Big Data.

In research, reproducibility is the ability to use data, techniques, and/or equipment to confirm the same result as previously obtained. This is fundamental to the validation of results and conclusions drawn from data. Open access and visibility allow for peer audit of the results derived from data and fosters inter- and cross-disciplinary research. The NRF, as primary research funding agency, and technical journal publishers are interested in reproducibility. However, it is more difficult to implement this notion in a Big Data computing environment where datasets can be extremely large and constantly evolving.

The cost of achieving Open Access should be recognised. A substantial effort involving an array of practitioners is needed to collect and transform data into formats that are readily useful. Ensuring confidentiality, and maintaining the integrity and security of some data collections are interlinked with this process. Proprietary Big Data analysis algorithms are used in a number of disciplines and national or group-licensing arrangements would reduce costs. Input should be provided to regulations such as the PoPIA to enable research use of personal data and opportunities arising the secondary use of should be leveraged.

It is essential to promote the beneficiation of nationally produced Big Data. Even in the humanities and social sciences, the digitisation of objects ranging from archaeological artefacts, to paintings and sculptures yields Big Data that hold opportunities to generate knowledge that would not have been possible otherwise. However, these resources pose challenges in large-scale data management that should be addressed to ensure that our historical and cultural assets are well preserved.

Research is needed to promote the transparency of data, in order to elevate the value of such data collections but also to understand and address the challenges of Big Data privacy, security, and ethics. New policies may be necessary to protect privacy and clarify the roles and responsibilities of various actors (owner, custodian, host, etc.) involved in the creation and use of Big Data. Techniques and tools are needed to help assess data security and to secure data in an increasingly cyber-attack prone world.

Much effort involving a range of practitioners goes into the provision and management of data in general. This work is traditionally not acknowledged and new mechanisms are being developed for attribution. These so-called altmetrics can serve as a means of recognition for data publications in similar ways that traditional measures, such as citation and impact factors, are used for scholarly publication. This means of reward is an important mechanism for incentivising and rewarding Big Data contributions.

Further, to support the efficient investment of resources, standards and policies are also needed to measure the performance and cost effectiveness of Big Data systems.

Many aspects of policy maintenance can be automated through rule-based authentication and access systems. Techniques are also needed for representing and processing metadata, i.e., the information about data that increase their value by making these resources more discoverable and reusable.

Research is needed to develop and implement standards for Big Data along the entire value chain. In this respect, collaboration with international Big Data forums and communities serves as benchmarking reference for sound Big Data practices. In general, different custodians may have different applicable data policies some of which could be regulated by international guidelines. SAWS, for example, is a member of the World Meteorological Organization (WMO) and it follows resolutions of the WMO that require the exchange of essential data (for the safety of life and property) without any restrictive condition. However, the exchange of additional data may be subject to further conditions.

The economic benefits of Open Data have been outlined by several studies. One of these by the European Commission, forecasts the direct benefits of the reuse of public data resources as monetised gains, which include market transactions realised in the form of revenue and Gross Value Added, the number of jobs in producing a service or output, and cost savings. Indirect benefits include time savings, knowledge economy growth and increased efficiency. According to this study, the cumulative direct market value within the European Union for the period 2016 to 2020 is estimated at €325 billion [77]. The number of related jobs created is expected to increase by approximately 25 000 over the same period.

Other examples of indirect benefits are 5.5% fewer road fatalities and a saving of 629 million hours. For South Africa to gain similar benefits, the openness of data has to be advocated and supported by policies on Open Data as an enabler of Open Science and innovation.

As a matter of principle, data acquired or generated using public funds should be publicly available where feasible. Therefore, the DSI's initiative to develop a national Open Science policy is a critical and necessary step to address these challenges and, consequently, reap the benefits of Open Data, Open Science and Open Innovation. Key principles for open research data should be embedded in the design of the software and even the architecture of Big Data systems.

6.6 OBJECTIVE 5: OVERARCHING COORDINATION AND ADVOCACY

Maintain an overarching and coherent national approach to investments in Big Data initiatives and activities

There is a variety of Big Data repositories, some being shared while others are not. Many of these facilities are hosted by public entities – universities, research councils, departmental ministries and parastatal organisations. Given the Open Access requirements by funding agencies such as the NRF and the increasing demand for data storage, many organisations are increasing or planning to increase and diversify their data repositories.

The unwarranted duplication of resources present possible inefficient or wasteful use of scarce resources and a coordinated view of significant budgets for Big Data infrastructure and activities would reduce this possibility.

A coherent national structure incorporating various levels or tiers of cyberinfrastructure would reduce unwarranted and wasteful duplication of scarce resources. Such an architecture integrated with those for priority projects such as the SKA would foster collaboration and serve the needs of a wide range of stakeholders and applications. Research infrastructures, as intended by SARIR, can be extended to serve projects and initiatives with similar focus and should be integrated into the national cyberinfrastructure. In order to achieve greater return and optimise investment, it is essential that Big Data initiatives and activities be coordinated at national level, across parastatal entities and including research institutions and academia.

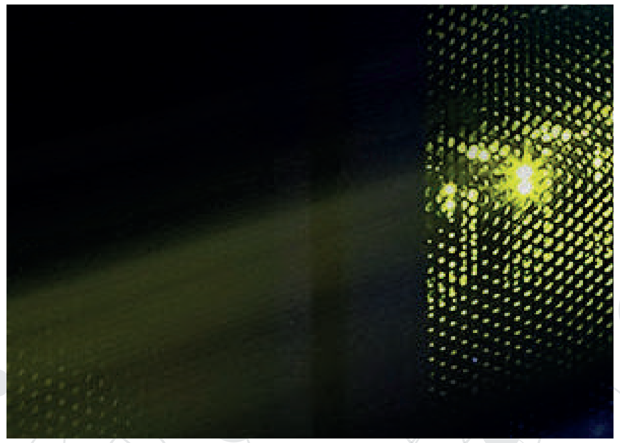
A structure representing the interests of key stakeholders would serve in an overarching capacity and governance role to orchestrate and oversee strategic implementation, planning and national investment in Big Data.

As in countries such as South Korea and the Netherlands, such an entity would ensure coherent, focused and well-directed execution of a Big Data strategic plan.

NICIS, in its objective to coordinate national cyberinfrastructure, has an instrumental role in orchestrating the formation and management of such a structure within the RD&I domain. However, there would be key stakeholders beyond the scope of NICIS (such as other government entities) that would require an inter-ministerial level of governance.

The Policy Intents of the 2019 DSI White Paper on Science Technology and Innovation [78] propose such policy intents that will enable efficient penetration of technologies developed from the NSI into government services, and this strategy supports the Policy Intents on enabling an innovation environment, and increased human capabilities in South Africa. Of particular relevance are the Policy Intents on expanding research outputs and transforming the research institutional landscape; transformation of the profile of the researcher base; improving the research system's output of human capabilities; and upgrading and expanding research infrastructure priority areas for Big Data RD&I can be reviewed and refined jointly, policies and standards, roles and responsibilities for Big Data actors can be defined and implemented in a coherent manner. Some stakeholders may be unconvinced about the benefits of Big Data RD&I. It is important to publicise the potential gains that can be derived and the risks of not engaging in Big Data R&D.

Priority areas would largely be informed by related strategies, including the Policy Intents expressed in the 2019 DSI White Paper on Science, Technology and Innovation. The SA-EU Open Science Dialogue report recommends the establishment of an Advisory Board to provide foresight and guidance for the development of a national Open Science Policy to support the institutionalisation of Open Science, Open Data and data sharing. Big Data, being a substantial component of data, would naturally be within the scope of this effort.



7 FRAMEWORK OF ACTIONS

A framework of actions for the implementation of this strategy is presented. These actions lift out the priority tasks linked to objectives and are guided by the principles presented above (Inclusivity, Privacy, Governance, Infrastructure and FAIR principles). The operationalisation of these actions, in terms of management, timeline, accountability and outputs, require further consideration and the detailed implementation of these actions would be guided by this strategy.

7.1 HUMAN CAPITAL AND RESEARCH CAPACITY DEVELOPMENT

A groundswell of new Data Science curricula is being offered at some institutions. Some courses are emerging at Master's level, but programmes at undergraduate and doctoral levels are lacking. There is a need to develop specialist programmes focussing on the sciences, technologies and the management of Big Data, as well as 'migration courses' needed to train students from other disciplines in applying Big Data tools in their research. Support is also lacking for 'citizen science' Big Data (such as the development of mobile applications that leverage Big Data).

- **Action 1:** Develop and implement a Big Data human capital development plan, in cooperation with NRF, DHET, academia and research councils, as a priority. The plan fast-tracks the development of requisite skills and expertise, including emerging areas of Big Data science and covers the pipeline from undergraduate to postdoctoral levels. It also takes into consideration and possibly extends ongoing activities such as the NEPTTP project and HCD activities forming part of Big Data projects. Training interventions to re-skill existing researchers and scientists are also needed. The DHET's staff development programme and the nGAP programme can be leveraged as implementation mechanisms for such programmes.

- **Action 2:** Include a minimum HCD requirement in funded Big Data projects. Interventions that align postgraduate to postdoctoral programmes with collaborative research projects are encouraged. Funding should incentivise collaborative rather than competing research and be directed at R&D in the development of a confederation of cyberinfrastructure technologies.

- **Action 3:** Together with companies and organisations, deploy Massive Open Online Courses that are freely available to the public, with incentives such as internships with businesses and entry to, and funded support for more advanced formal courses. This action can form part of the Big Data HCD plan.

7.2 CYBERINFRASTRUCTURE FOR BIG DATA

The Knowledge Triangle illustrates the interaction among research, education and innovation as key drivers of a knowledge-based society [79]. Cyberinfrastructure is a pivotal enabler and NICIS has a key role in this context. The strategic objectives of NICIS include the establishment of a world-class national integrated cyberinfrastructure system supporting research, innovation and teaching, and positioning South Africa to partake and lead large-scale science projects [66].

Key actions toward realising these objectives are to upgrade the cyberinfrastructure components and to foster the development of expert human capital for the application and provision of cyberinfrastructure. These objectives and actions naturally align with the objectives of this strategy and, within DIRISA,;the management of data is a key resource for research. There is a need to extend and upscale technology infrastructure and deploy world-class Big Data platforms, in order to achieve a disruptive surge in Big Data innovation. NICIS is the locus for the provision of Tier 1 infrastructure in support of Big Data R&D and develop an overarching view of the national cyberinfrastructure. There is a further need for NICIS to be structured into a national institute for cyberinfrastructure innovation and to coordinate all individual CI programmes, activities and projects under an overarching umbrella structure. The rationale for this institute is to achieve greater economies of scale by reducing duplication, and to harness synergies that may not be achievable through individual or siloed efforts.

- **Action 4:** Double the capacity of high-end cyberinfrastructure to accommodate research across all disciplines, and to serve as Big Data innovation hub as a publicly available infrastructure. Services such as federated identity management and Persistent Identifiers should be implemented to support FAIR use of data. The use of locally developed technologies should be encouraged to support the development of expertise and local industry.
- **Action 5:** Develop a federated national cyberinfrastructure architecture framework as basis for increased and sustained investment. This framework should be developed by NICIS to guide the development of a roadmap to federate islands of cyberinfrastructure and hence promote efficiencies and interoperability. The sustainability of national cyberinfrastructure should be explicitly addressed through the development and implementation of business models and models of operation that ensures long-term funding for sustainability.
- **Action 6:** Perform a benchmark study to identify user needs, to determine the current landscape of technology, expertise and skills. This information would inform a long-term strategy that can be developed by NICIS and which identifies priority areas and high impact opportunities that would result in 'quick wins'. Funding should be targeted at building and sustaining a capable cyberinfrastructure that would achieve outcomes requiring little more investment.

7.3 COLLABORATION

- **Action 7:** Create opportunities to collaborate with government and industry. One such possible mechanism is the creation of structures to coordinate collaborative projects across governmental entities, business and academic boundaries. An example scenario is Big Data Innovation Hubs – centres created through public and private sector investments to support the Big Data needs of industry, business and government. Encourage multinational cooperation on government's terms to invest in Big Data hubs in the country and partner with government and academia. Tax incentives or equity partnership under the **dtic** could be explored for this purpose.

- **Action 8:** Identify key stakeholders and a list of 'flagship' Big Data projects and programmes that are based on strategic priorities and needs, and that have greatest potential for success. These flagship initiatives serve as anchors for ecosystems of RD&I focused on particular industrial, business or technological domains.

- **Action 9:** Reduce bureaucratic hurdles for technology and data sharing by developing approved 'standard templates' for fast-tracking the negotiation of collaborative projects among business, academia and research councils. Arrangements for intellectual property sharing, non-disclosure and funding should have a minimum number of constraints.
- **Action 10:** Incentivise triple helix (university-industry-government) Big Data partnerships and support the sustainability of SMMEs through reward and the provision of cyberinfrastructure through a national Big Data institute or Big Data Innovation Hubs. A programme supporting the sharing of experiences, results and capabilities locally and continentally would allow for quicker assimilation.

7.4 DATA GOVERNANCE

Trends towards Open Data and Open Science and Open Innovation have implications for the ethical use of data. While there are increasing degrees of acceptance of the FAIR principles, the impact of these principles on privacy is more severe for the case of Big Data. Given technologies such as machine learning and high-end computers, it has become more difficult to guarantee privacy and security. The accountabilities and practices of the different actors involved in the creation and use of Big Data need to be clearly defined.

While NICIS has an overarching role in coordinating the development of these functions, the input of domain experts and other key stakeholders is vital to ensure that structures and processes appropriately support good governance and improved efficiencies.

- **Action 11:** Conduct a situational analysis of the prevailing state of (Big) Data governance and develop and implement standards, policies, frameworks and a code of conduct for Big Data stewardship and so promote the transparency of Big Data and regulate its ethical use. Since Big Data is a component of data, these efforts would form part of the development of a national Open Science policy based on the recently completed EU-SA Open Science framework.
- **Action 12:** Develop standardised procedures to classify data (in terms of sensitivity and confidentiality) and make publicly funded data available as open data by default. This action is a natural progression of the NRF's Open Access statement.
- **Action 13:** Develop measures, so-called altmetrics, for acknowledging and incentivising Big Data contributions.
- **Action 14:** Harmonise current legislation that governs access to data with the promotion of Big Data for research and make data available for re-use.

7.5 OVERARCHING COORDINATION AND ADVOCACY

Data, in general, is an infrastructural asset and is necessary to manage and maintain these resources, As such, overarching governance would help achieve greater return of investment in Big Data.

A coherent view of significant budgets for Big Data infrastructure and activities would reduce unwarranted or wasteful duplication of scarce resources. Such a structure would have sight of related activities such as the Open Data and Open Science initiatives, and of key strategic focus areas. In addition to a governance structure, the current framing of objectives will also lead to the establishment of further structures or forums. These entities will include stakeholders with focus on particular aspects of this strategy, such as HCD, collaboration and data governance.

- **Action 15:** Establish an overarching structure that represents key stakeholders to oversee and review the implementation of a Big Data strategy and prioritise investments across all implementing agents. This structure can be an extended function of the NICIS Steering Committee, since NICIS would have a pivotal role as implementing agent. This structure can also share information and harmonise activities related to other related activities such as education and training, and research funding.
- **Action 16:** Establish structures that focus on specific objectives of the strategy and include key stakeholders involved in that aspect. Working Groups or Task Teams subordinate to an overarching structure can prioritise, chart out and oversee the implementation of an action plan focussed on achieving specific objectives. These groups would also ensure that activities are harmonised with related activities such as those of the DSI White Paper on Science and Technology and the Open Science Framework. Examples of such structures are, an HCD Working Group with representation from key stakeholders such as NRF, USAf, DSI, DHET, NICIS and SKA-SA; and a Collaboration Working Group that has membership from NICIS, SKA, NRF, DSI, DEFF, DCDT and the **DTI**. Structures for inter-governmental department collaboration (e.g., between the DSI and DEFF for Earth observation data) should be lifted to the fore to ensure greater synergy in achieving objec-
- **Action 17:** Formulate a plan to for advocate the national value proposition of Big Data R&D. This plan should include stakeholder engagement to attract international investment and to stimulate the development of a thriving ecosystem of collaborative Big Data communities.
- **Action 18:** In line with the recently published S&T White Paper, establish an inter-ministerial departmental coordination structure to ensure that governance oversight is implemented at the highest level.

The proposed actions will require substantial commitments over an extended period of time. Such investments can be directly related to indicators of success as contemplated in the following section.

8 INDICATORS OF SUCCESS

Several indicators of the outcomes and impact of a Big Data strategy can be used to monitor and evaluate progress and achievement. These indicators should be specified in more measurable and quantitative terms:

- *Number and extent of successful local collaborative industrial Big Data projects that involve the use of cyberinfrastructure for Big Data;*
- *Number of SMMEs active in developing and offering Big Data-based services and products;*
- *Number of formally qualified researchers (Master's, doctoral and postdoctoral) in Big Data programmes;*
- *Level of participation and outputs measured in terms of altmetrics, including the number of openly accessible Big Data collections available;*
- *Number of new students enrolled in formal undergraduate Big Data study programmes;*
- *Number and diversity of Big Data research programmes offered by institutions;*
- *Number of Big Data-based services that are deployed by governmental institutions;*
- *Recognition of South Africa in terms of number of Grand Science projects, international partnerships and scientists attracted to South Africa;*
- *Contributions to the transformation of the digital economy;*
- *Contributions to building an enabling environment for the 4IR;*

  Usage of Big Data for evidence-based decision-making, intervention and policy formulation;
- *Using Big Data applications to address the triple challenge;*
- *Availability and provision of analytical capabilities for Big Data at advanced and virtual cyberinfrastructure centres;*
- *Legislation protecting privacy and ensuring security, harmonised with open data and open science principles (e.g., augmentation of PoPIA);*
- *Adoption rate of open data and open science practices in Big Data research projects and programmes; and*
- *Level of confederation of Big Data repositories and cyberinfrastructure (breaking silos of Big Data research projects and programmes).*

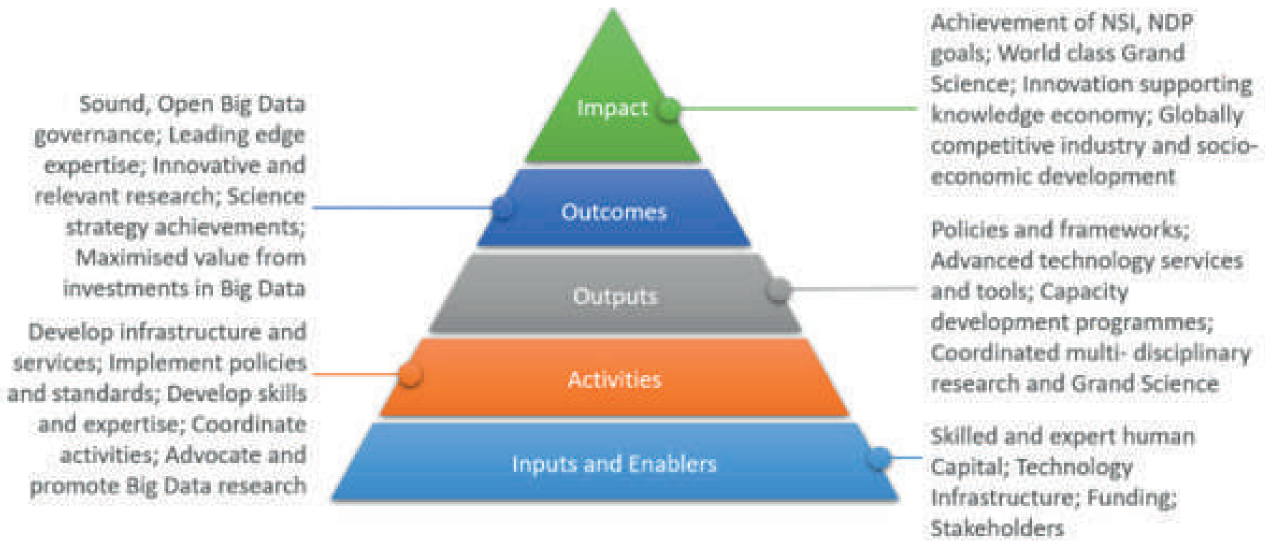
9 CONCLUSION

Big Data has fundamentally changed the landscape of research and business, and the benefits to be gained from Big Data are extensive and far-reaching. Conversely and more crucially, South Africa will miss a unique opportunity to disruptive increase the pace of developing a knowledge economy if no explicit and concerted effort is made to leverage the Big Data phenomenon in a prompt and decisive manner.

The impact of this strategy is demonstrated by Figure 9. Skilled and expert human capital, sufficient technology infrastructure and investment together with key stakeholder commitment, comprise the ecosystem for achieving the benefits of research Big Data. This enabling environment is achieved through concerted efforts to deploy appropriate cyberinfrastructure, develop skills and advance expertise, developing standards and policies to govern the utilisation of research Big Data, and advocating the conduct of Big Data research. The outputs of these activities include the implementation of policies and frameworks for research Big Data, provision of advanced cyberinfrastructure, well-coordinated human capital development programmes from graduate to post-doctoral levels – all of which are integrated into collaborative 'triple P' research projects and programmes addressing relevant and Grand Science challenges. The outcomes of these outputs include sound and open research Big Data governance, leading expertise and skills, advanced and relevant Big Data science, and the realisation of national strategies and maximised return on Big Data investments.

The impact of these outcomes will be evident in the achievement of NSI and NDP goals, world-class Grand Science, innovation supporting the knowledge- and digital economy, accelerated socioeconomic development and a globally competitive industry.

Figure 9: *The impact of this research Big Data strategy*


Impact
Outcomes
Outputs
Activities
Inputs and Enablers
Sound, Open Big Data governance; Leading edge expertise; Innovative and relevant research; Science strategy achievements; Maximised value from investments in Big Data
Develop infrastructure and services; Implement policies and standards; Develop skills and expertise; Coordinate activities; Advocate and promote Big Data research
Achievement of NSI, NDP goals; World class Grand Science; Innovation supporting knowledge economy; Globally competitive industry and socio-economic development
Policies and frameworks; Advanced technology services and tools; Capacity development programmes; Coordinated multi- disciplinary research and Grand Science
Skilled and expert human Capital; Technology Infrastructure; Funding; Stakeholders

The impact of seizing the opportunities provided by research Big Data lie in positioning South Africa to develop a less resource-based economy and to participate as a producer rather than a consumer of Big Data outputs. The primary challenge is the lack of expertise and skills. This lack extends across the scientific, technological and management domains. Although business and industry are capitalising on the Big Data phenomenon, they are hamstrung by the shortage of competent and skilled scientists, engineers and technologists. Some industries (e.g., IBM and Intel) are pursuing quadruple helix partnerships in which civil society joins with academia, business and the government sector.

These collaboration models present eminent opportunities to develop a knowledge economy and derive societal benefit from Big Data. A national Big Data strategy for research, development and innovation encapsulates and codifies the pivotal role of research in the 'knowledge triangle' to support evidence-based decision making and policy development for an accelerated transition to a knowledge- and digital economy; rapid development of required skills and expertise; implementation of enabling infrastructure; and the establishment of regulatory and governance requirements.

ANNEXURE A: BIG DATA STRATEGIES OF SOME COUNTRIES

The United Nations and several international organisations surveyed the potential Big Data for the developing world [74]. The World Economic Forum in Davos characterised Big Data as a "currency for new economic development" and the Organisation for Economic Co-operation and Development adopted the "evaluation of the economic benefits of Big Data" as part of their agenda for the Working Party on Indicators for the Information Society.

Developed and developing countries, together with industry, have recognised the challenges and opportunities intrinsic in Big Data and, to this end, have implemented strategies and committed substantial investments to leverage the potential of Big Data. Some responses and strategic initiatives are considered.

UNITED STATES OF AMERICA

In March 2012, the United States government committed $200 million to a Big Data Research and Development Initiative. The aims of this initiative are to:

- *Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyse, and share huge quantities of data;*
- *Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen national security, and transform teaching and learning; and*
- *Expand the workforce needed to develop and use Big Data technologies.*

Funding is targeted at advancing core scientific and technological means to:

(a) Manage, extract, analyse and visualise useful information from large and diverse data sets;

(b) Derive knowledge from data, infrastructure to curate and serve data; and

(c) Leverage sensing for decision support and improved situational awareness. Education and workforce development has a high priority with health, energy, defence and earth system science being key domains.

The initiative is a response to recommendations by the Presidential Council of Advisors on Science and Technology report on Big Data in May 2014. While the report emphasised Big Data as a driver for unprecedented socioeconomic progress, serious concerns were raised with regard to privacy and ethics. Five pertinent areas are addressed by the strategy:

- *Preserving privacy values by protecting personal information in the marketplace;*
- *Recognising particularly publicly supported schools for using Big Data to enhance learning opportunities, while protecting personal data usage and building digital literacy and skills;*
- *Preventing new modes of discrimination that some uses of Big Data may enable;*
- *Ensuring responsible use of Big Data in law enforcement, public safety, and security; and*
- *Harnessing data as a public resource to improve the delivery of public services and investing in research and technology.*

The US National Institute of Standards and Technology (NIST) leads the development of the NIST Big Data Interoperability Framework through the NIST Big Data Public Working Group to develop consensus definitions, taxonomies and secured reference architectures for a vendor-neutral, technology infrastructure enabling Big Data computing. Following the US government initiative, and allied organisations have co-invested significant amounts of funding in Big Data research. In March 2012, the US National Science Foundation launched the Big Data Research Initiative with a budget of $10 million to solicit the proposals for development of core technologies to support Big Data Science and Engineering, integrative graduate education and research interventions, and data-intensive education related research.

EUROPEAN UNION COUNTRIES

The European Commission (EC) has an overarching championing, coordination and funding role in several European Big Data-related initiatives. A communication to the European Parliament and the EC, entitled "Towards a Thriving Data-Driven Economy" outlines European opportunities, challenges, responses as backdrop to the EC strategy that was developed together with the European data industry. An amount of €2.5 billion is committed to public-private partnerships, in order to "master Big Data". Four intervention mechanisms are supported:

- **Innovation-spaces (i-Spaces)** are hubs for bringing technology and application developments together, acting as incubators for new businesses and for the development of skills, competence and best practices. Important sectors include health, transportation and food security.
- **Lighthouse projects** help to raise awareness of opportunities offered by Big Data and the value of data-driven applications for different sectors. Lighthouse projects are large-scale data-driven innovation demonstration projects highlighting the benefits of Big Data Value and aiming at higher visibility, awareness and impact.
- **Technical projects** address specific Big Data-targeted aspects that have technical priority.
- **Cooperation and coordination projects** foster international cooperation for efficient information exchange and coordination of activities.

The EU earmarked €500 million from Horizon 2020 to address fundamental research problems related to scalable and responsive analytics. "Mastering data" includes achieving a 30% global market share, 100 000 new data related jobs by 2020, 10% lower energy consumption, better healthcare and a more productive industry. The EU also expresses concerns about the use of technologies for illicit surveillance and addresses matters of trust and privacy through data protection and security regulations.

Several factors are identified as key enablers for Big Data driven economy:

- *An ecosystem of different interacting role players;*
- *The availability of good quality, reliable and interoperable datasets and robust infrastructure;*
- *A base of appropriately skilled and expert human capital; and*
- *A range of prioritised application areas where improved Big Data handling have an impact.*

The strategy proposes an action plan based on contractual PPPs, a Strategic Research and Innovation Agenda identifying sectorial priorities, Digital Entrepreneurship and skills base development.

UNITED KINGDOM (UK)

The vision of the UK government is "to become a world leader in extracting insight and value from Big Data for the benefit of citizens, consumers, business and academia, the public and private sector" through:

- *A strong skills base, able to manage, analyse, interpret and communicate data;*
- *A strategic plan for data infrastructure across the country;*
- *World-leading research and development, pushing frontiers and driving innovation in data science and analytics; and*
- *Ensuring that data can be accessed and shared securely, as appropriate.*

Their strategy is underpinned by a strong policy framework that protects and empowers individuals and supports innovation and growth – with government as an exemplar of best practice and with science as a key driver of this capability. The UK Data Capability parastatal agency committed £189 million to Big Data and established an E-infrastructure Leadership Council to advise government on infrastructure and skills needed to leverage Big Data opportunities (UK Govt 2013). In support, the Open Data Institute has been established and more than 10 000 public datasets have been published. The strategy was produced in partnership with industry and academia and highlights several key aspects:

- **Human capital:** : a strong skills base able to manage, analyse, visualise and interpret Big Data;
- **Tools and infrastructure:** available to facilitate the storage and analyses of data for collaborative R&D; and
- **Data as an enabler:** the ability of consumers, businesses and academia to access and share data secure-

The strategy further includes measures to:

- *Build capability in the commercial, academic and public sectors;*
- *Strengthen skills focused at schools, higher and further education;*
- *Continue professional development; and*
- *Ensure that the UK's infrastructure and R&D environment supports improved data capability.*

A £11.3 million innovation centre dedicated to helping Scotland capitalise on the growing market in analytics and Big Data technology was opened in late 2014. The investment is projected to return a minimum of 345 new jobs and an additional £155 million of value to the Scottish economy. The so-called Data Lab is intended to capture new market opportunities and boost productivity accelerate Scottish industrial development and applications in Big Data analytics and data science techniques

AUSTRALIA

In March 2013 the Australian government's Department of Finance and Deregulation issued a Big Data Strategy Issues Paper that motivated a strategy, addressed pertinent challenges and their way forward on the issue. The purpose of the paper was to consider the range of opportunities in relation to the use of Big Data and the potential concerns raised by these opportunities. Like other countries, they recognise that Big Data can provide profound insights into societal areas such as healthcare, medical and other sciences, transport and infrastructure, education, communications, meteorology and social sciences.

Similarly, privacy rights are a central issue although *"with proper considerations, agencies will be able to use Big Data to develop better policies and deliver better services without compromising the privacy rights of the public"* (Aust 2013).

Their main motivation for a Big Data strategy emanated from the Australian Public Service ICT Strategy 2012-2015, i.e. "…to use ICT to increase public sector and national productivity by enabling the delivery of better government services […], communities and business…" A notable observation of the strategy is that significant productivity improvements are directly attributable to ICT investment. Listed key areas that Big Data can influence include data management, service personalisation, predictive analytics, productivity and efficiency. Their strategic vision presents improved services, improved efficiency and open engagement as priority areas. Privacy, trust and security, data management and sharing, technology and analytical systems, and skills are identified as the main challenges. The following objectives were set:

- *Leverage the Big Data experience to address the skills deficit that exists in this area. Opportunities to use these skills and experience to:*
  - *improve the management and analysis of government-held datasets and,*
  - *improve government operations, policy development and service delivery;*
- *Ensure that privacy issues are addressed up front in order to:*
  - *protect the privacy rights of the public and,*
  - *clarify the development of better policies and services; and*
- *Encourage the release of public sector information consistent with all privacy and security legislative instruments and guidance.*

OTHER COUNTRIES

While governmental strategies by some developing countries are not readily apparent, those that host leading technology companies have progressed substantially on the implementation of Big Data strategies. Examples of interventions that capitalised on the potential of Big Data in some countries:

China is regarded as one of the largest Big Data generators. Events such as the Third Advanced Software Engineering conference held in China in August 2014 focused exclusively on Big Data Science and Computing with Big Data mining, analytics and applications featuring as foremost topics.

South Korea established a Big Data Strategy Forum in March 2012 and a Korean Big Data Forum in September 2012 as part of their National Master Plan on Big Data. The plan involves five national ministries, including the Science and Technology Commission, Education, and Science and Technology in the formulation of a national Big Data agenda. The South Korean government opened a Centre for Big Data Analytics that provides large-scale data analytics technologies such as servers and storage on an open source platform to small and medium enterprises (Korea 2014). The intent is for the centre to boost the development of human capacity and increase the adoption and use of Big Data analytics. The centre is networked with the government-owned Open Data Square, private data distributors and public data portals. Several pilot Big Data projects (e.g., Youth at Risk and Media Data Analysis) have been started.

The 2014 Big Data Innovation Summit held in **Singapore** and **Malaysia** was aimed at creating a successful data-driven culture and attracted more than 500 attendees from various solution providers and industries, such as telecommunications, finance and healthcare. In March 2012, **Japan** formulated a "Basic Strategy on Using Big Data" as a major 'Active Japan' initiative and in support of their Revitalization Strategy. The government of France announced the launch of a Big Data project and allotted € 11.5 million for seven Big Data processing projects as part of their Programme of Investments for the Future. Strategies for other developing countries such as Brazil, India and Russia are not readily apparent, although significant business activities are evident from Internet search results.

BUSINESS AND RESEARCH INSTITUTIONS

Most multinational and many smaller business and commercial enterprises have developed Big Data strategies, in order to improve internal business processes and enhance their competitive business intelligence. Major IT enterprises have adopted or incorporated Big Data-related technologies in their service offerings and are aggressively pursuing Big Data monetisation. Examples are SAS, Oracle, Intel, HP, Hitachi, Google and Microsoft. Research and allied institutions have, likewise, adopted aggressive positioning strategies to leverage Big Data. Some examples are CSIRO, Fraunhofer Institute, NASA and NIST.

ANNEXURE B: CONTRIBUTORS TO THE STRATEGY

A number of individuals and stakeholders contributed to the development of this strategy.

International experts and heads of organisations managing cyberinfrastructure for large research data collections were initially consulted on high-level issues that should be addressed. Examples of international contributors who provided opinion are, Wo Chang, convenor of the ISO/IEC SC 42/WG Big Data Working Group and chair of the US National Institute of Science and Technology (NIST) Big Data Working Group; Dr Peter Wittenburg, former Head of the Max Planck Institute Technical Group; Dr Mark Parsons former Secretary General of the Research Data Alliance; Dr Damien Le Carpentier, manager of the EUDAT initiative; and Prof Ross Wilkinson, Director of ANDS.

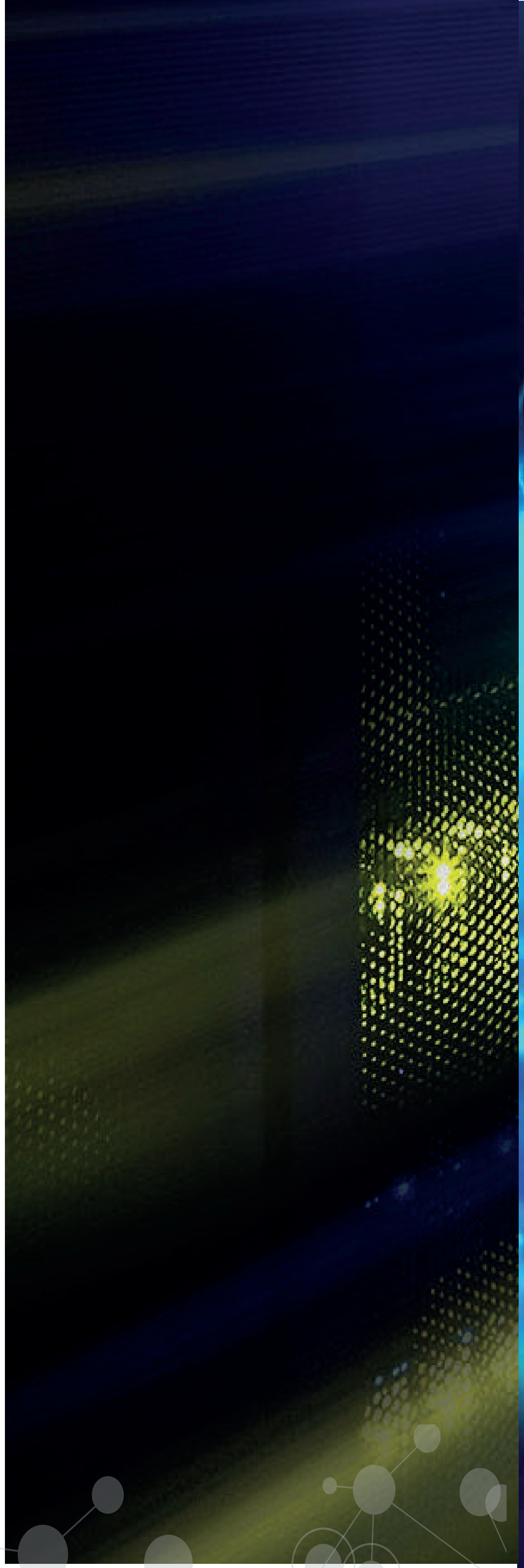
An initial version of this strategy was presented and discussed at a national workshop held in October 2018. Ninety-eight individuals from the organisations listed below, participated in the event.

Agricultural Research Council
Consulting Engineers South Africa
Council for Scientific and Industrial Research
Department of Higher Education and Training
Department of Trade and Industry
Durban University of Technology
Human Sciences Research Council
Machine Intelligence Institute of Africa
Medical Research Council
Microsoft Enterprise Services, South Africa
National Intellectual Property Management Office
National Research Foundation
South African National Space Agency
South African Radio Astronomy Observatory

Technology Innovation Agency
University of Limpopo
University of North West
University of Pretoria
University of South Africa
University of the Witwatersrand
Universities South Africa
University of Kwazulu-Natal

Feedback obtained from groups of participants who addressed specific objectives and aspects of the strategy was incorporated. Following a proposal at this event, a panel of local experts as listed below, further refined and contributed to this version of the strategy:

- *Dr J Ludik (CEO: Machine Intelligence Institute of Africa)*
- *Prof. B Twala (Head: AI research, UNISA*
- *Dr H Sithole (Centre Manager, NICIS)*
- *Wo Chang (Chair: NIST Big Data Working Group)*
- *Dr Jean-Claude Burgelman (DG: Research and Innovation, European Commission)*
- *Dr Mary-Jane Bopape (Chief Scientist, SAWS)*
- *Prof. Bruce Mellado (Wits University)*



REFERENCES

[1] South African Department of Telecommunications and Postal Services, "National e-Strategy," 10 November 2017. [Online]. Available: https://www.dtps.gov.za/images/phocagallery/Popular_Topic_Pictures/National-e-strategy.pdf. [Accessed 20 October 2019].

[2] South African Department of Commmunications and Digital Technologies, "Draft Big Data and Cloud Policy Framework," National Department of Communications and Digital Technologies, Pretoria, 2019.

[3] South African Department of Telecommunications and Postal Services, "South African Government Gazette," 4 Dec 2018. [Online]. Available: https://www.gov.za/sites/default/files/gcis_document/201812/42078gen764.pdf. [Accessed 6 Feb 2019].

[4] South African Department of Communications and Digital Technologies, "National Government of South Africa," Yes Media, 2019. [Online]. Available: https://nationalgovernment.co.za/units/view/428/deparment-of-communications-and-digital-technologies-dcdt. [Accessed 12 October 2019].

[5] South African Government, "National Development Plan 2030," February 2013. [Online]. Available: https://www.gov.za/issues/national-development-plan-2030. [Accessed 23 August 2017].

[6] South African Department of Science and Technology, "Draft White Paper on Science, Technology and Innovation," DST, 10

[7] National Advisory Council on Innovation, "The South African National System of Innovation: Structures, Policies and Performance," NACI, 3 Aug 2017. [Online]. Available: http://www.naci.org.za/index.php/the-south-african-national-system-of-innovation-structures-policies-and-performance/. [Accessed 21 Feb 2018].

[8] South African Department of Science and Technology, "ICT RDI Roadmap," 2016. [Online]. Available: https://www.dst.gov.za/images/ict_rdi_roadmap.pdf. [Accessed 6 Feb 2017].

[9] A. Katal, M. Wazid and R. H. Goudar, "Big Data: Issues, Challenges, Tools and Good practices," in 2013 Sixth International Conference on Contemporary Computing (IC3), 2013.

[10] D. Lindeque, "Big demand for big data skills," IOL Business Report, 2017. [Online]. Available: https://www.iol.co.za/business-report/big-demand-for-data-analytics-skills-11197431. [Accessed 11 Oct 2017].

[11] Z. H. Zhou, N. V. Chawla, Y. Jin and G. J. Williams, "Big Data Opportunities and Challenges: Discusions from Data Analytics Prespectives," IEEE Computational Intelligence, vol. 9, no. 4, pp. 62 - 74, Nov 2014.

[12] J. Bughin, J. Livingston and S. Marwaha, "Seizing the potential of 'big data'," McKinsey Quarterly, pp. 103-109, 4 2011.

[13] T. Davenport, P. Barth and R. Bean, "How Big Data is Different," MIT Sloan Management Review, SLOANREVIEW.MIT.EDU, 2016.

[14] K. Antelman, "College and Research Libraries," American Library Association, 2014.[Online].Available: https://crl.acrl.org/index.php/crl/article/view/15683. [Accessed 12 Oct 2017].

[15] J. Lunshof, R. Chadwick and D. Vorhaus, "From genetic privacy to open consent," Nature Reviews Genetics, vol. 9, no. 406, 2008.

[16] H. Chen, R. Chiang and V. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly, vol. 36, no. 4, pp. 1165 - 1188, 2012.

[17] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," in 2013 46th Hawaii International Conference on System Sciences, 2013.

[18] Gartner Group, "Gartner Reveals the 2017 Hype Cycle for Data Management," Garner Inc., 28 Sep 2017. [Online]. Available: https://www.gartner.com/en/newsroom/press-releaes/2017-09-28-gartner-reveals-the-2017-hype-cycle-for-data-management. [Accessed 11 Apr 2018].

[19] Gartner Group, "Hype Cycle for Data Science and Machine Learning, 2018," Gartner Inc., 23 Jul 2018. [Online]. Available: https://www.gartner.com/doc/3883664/hype-cycle-data-science-machine. [Accessed 11 Sep 2018].

[20] International Data Corporation, "The Digitization of the World from Edge to Core," 1 Nov 2018. [Online]. Available: https://www.idc.co.za/news-publications/. [Accessed 12 Dec 2019].

[21] A. Labrinidis and H. V. Jagadish, "Challenges and Opportunities with Big Data," in Proc. VLDB Endow., 2012.

[22] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs and C. Roxburgh, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011.

[23] R. Kitchin, "Big Data, new epistemologies and paradigm shifts," Big Data & Society, vol. 1, no. 1, 1 Apr 2014.

[24] D. Boyd and K. Crawford, "Critical Questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon," Information, Communication & Society, vol. 15, no. 5, pp. 662 - 679, 2012.

[25] L. Floridi, "Big Data and Their Epistemological Challenge," Philosophy & Technology, vol. 25, no. 4, pp. 435 - 437, 1 Dec 2012.

[26] M. Chen, S. Mao, Y. Zhang and V. Leung, "Big Data Applications," in Big Data: Related Technologies, Challenges and Future Prospects, Cham, Switzerland, Springer, 2014.

[27] M. Andrejevic, "Big Data, Big Questions - The Big Data Divide," International Journal of Communication, vol. 8, no. 0, 2014.

[28] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, H. L., A. Halevy, J. Han, H. Jagadish, H. Labrinidis, S. Madden, Y. Papakonstantinou and J. Patel, "Challenges and opportunities with big data: a community white paper developed by leading researchers across the United States," Computing Research Association, Washington, 2012.

[29] T. Hey, S. Tansley and K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

[30] M. Schermann, H. Hemsen, C. Buchmüller, T. Bitter, H. Krcmar, V. Markl and T. Hoeren, "Big Data," Business & Information Systems Engineering, vol. 6, no. 5, pp. 261 - 266, 01 Oct 2014.

[31] South African Department of Science and Technology, "New draft White Paper on Science, Technology and Innovation," Department of Science and Technology, Sep 2018. [Online]. Available: https://www.dst.gov.za/index.php/media-room/latest-news/2621-new-draft-white-paper-on-science-technology-and-innovation. [Accessed 3 Sep 2018].

[32] J. Ovenden, "Five Ways to Solve the Data Skills Gap," Innovation Enterprise, 23 Jan 2018. [Online]. Available: https://channels.theinnovationenterprise.com/articles/5-ways-to-solve-the-data-skills-gap. [Accessed 12 Jul 2018].

[33] T. A. Davenport, Big Data @ Work: Dispellnig the Myths, Uncivering the Opportunities,Havard Review Press, 2014.

[34] S. Miller, "Collaborative Approaches Needed to Close the Big Data Skills Gap," Journal of Organization Design, vol. 3, no. 1, pp. 26 - 30, 2014.

[35] K. Michael and K. Miller, "Big Data: New Opportunities and New Challenges [Guest editors' introduction]," Computer, vol. 46, no. 6, pp. 22 - 24, Jun 2013.

[36] J. Cano, "The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity," XSI, 11 Mar 2014. [Online]. Available:https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity.[Accessed 9 Jul 2017].

[37] N. Joshni, "Top 5 sources of big data," Allerin, 26 Nov 2017. [Online]. Available: https://www.allerin.com/blog/top-5-sources-of-big-data. [Accessed 11 Apr 2018].

[38] [38] L. Bustos, "The Future of Commerce," GetElastic, 2018. [Online]. Available: https://www.getelastic.com/big-data-infographic. [Accessed 4 Sep 2018].

[39] D. Agrawal, S. Das and A. El Abbadi, "Big Data and Cloud Computing: Current State and Future Opportunities," in Proceedings of the 14th International Conference on Extending Database Technology, Uppsala, Sweden, 2011.

[40] IBM, "IBM big data analytics: insights without limits," IBM, 2018. [Online]. Available: https://www.ibm.com/za-en/it-infrastructure/solutions/big-data. [Accessed 11 Sep 2018].

[41] L. Columbus, "10 Charts That Will Change Your Perspective Of Big Data's Growth," Forbes, 23 May 2018. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#7e8e336a2926. [Accessed 11 Sep 2018].

[42] Statistica, "Forecast of Big Data market size, based on revenue, from 2011 to 2027," Statiistica, 5 Aug 2018. [Online]. Available: https://www.statista.com/statistics/254266/global-big-data-market-forecast/. [Accessed 23 Sep 2018].

[43] Report Linker, “Big Data Industry 2018,” Report Linker , 25 Sep 2018. [Online]. Available: https://www.reportlinker.com/market-report/Information-Technology/513221/Big-Data?utm_source=adwords4&utm_medium=cpc&utm_campaign=High_Tech_And_Media&utm_adgroup=Big_Data_Reports&gclid=Cj0KCQjwxvbdBRC0ARIsAKmec9Z2zn6ZPLuCCb6-jJQxdkF-y-BBbe6IXU3GOMXCQZ3nxf. [Accessed 30 Sep 2018].

[44] N. Joshni, “This is why big data in transportation is a big deal,” 7 10 2017. [Online]. Available:https://www.allerin.com/blog/this-is-why-big-data-in-transportation-is-a-big-deal. [Accessed 11 12 2018].

[45] H. R. Joop Hox and Boeije, “Data collection, primary versus secondary,” in Enclyclopedia of Social Measurement, Elsevier, 2005, pp. 592 - 599.

[46] Google, “Google Privacy Policy,” Google Inc, 25 May 2018. [Online]. Available: https://policies.google.com/privacy?hl=en&gl=ZZ. [Accessed 23 Sep 2018].

[47] Microsoft, “Microsoft Privacy Statement,” Microsoft, 01 October 2018. [Online]. Available: https://privacy.microsoft.com/en-gb/privacystatement. [Accessed 09 Oct 2018].

[48] A. Kasperen, “Can you have both security and privacy in the internet age?,” World Economic Forum, 21 Jul 2015. [Online]. Available:https://www.weforum.org/agenda/2015/07/can-you-have-both-security-and-privacy-in-the-internet-age/. [Accessed 20 Jun 2017].

[49] C. Cocorocchia, “5 things you (probably) don't know about online privacy – but should,” World Economic Forum, 24 May 2017. [Online]. Available: https://www.weforum.org/agenda/2017/05/your-personal-data-privacy-what-to-know/. [Accessed 8 Aug 2018].

[50] O. Tene and J. Polonetsky, “Privacy in the Age of Big Data,” Stanford Law Review Online, vol. 64, no. 63, pp. 63 - 69, 2012.

[51] K. Strandburg, “Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context,” in Privacy, Big Data, and the Public Good: Frameworks for Engagement, J. Lane, V. Stodden, S. Bender and H. Nissenbaum, Eds., Cambridge, Cambridge University Press, 2014, pp. 5 - 43.

[52] IQ Facebook, “Digital Diversity: A Closer Look at African Americans in the US,” Facebook, 29 Jan 2015. [Online]. Available: https://www.facebook.com/iq/articles/digital-diversity-a-closer-look-at-african-americans-in-the-us. [Accessed 5 Aug 2017].

[53] S. Wilson, “The myth of the informed Internet user,” Constellation Research, 21 Oct 2017. [Online]. Available: https://www.constellationr.com/blog-news/myth-informed-internet-user. [Accessed 23 Jul 2018].

[54] M. Bashir, A. Lambert, C. Hayes and J. Kesan, “Online Privacy and Informed Consent: The Dilemma of Information Asymmetry,” Information Science and Technology, vol. 1, no. 1, pp. 6 - 10, 2015.

[55] R. Neisse, G. Baldini, G. Steri and V. Mahieu, “Informed consent in Internet of Things: The case study of cooperative intelligent transport systems,” in 2016 23rd International Conference on Telecommunications (ICT), 2016.

[56] D. Lazer, R. Kennedy, G. King and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” Science, vol. 343, no. 6176, pp. 1203 - 1205, 2014.

[57] A. Naccaratoa, S. Falorsi, S. Loriga and A. Pierinia, “Combining official and Google Trends data to forecast the Italian youth unemployment rate,” Elsevier, vol. 130, pp. 114 - 122, 2018.

[58] M. Petrescu, C. Dobre and S. B. Mrad, “Assessing Consumer Confidence from Online Sources,” in Marketing at the Confluence between Entertainment and Analytics , P. Rossi, Ed., Cham, Springer, 2017, pp. 1587 - 1588.

[59] South African Department of Science and Technology, “South African Research Infrastructure Roadmap,” 1 Oct 2016. [Online]. Available: https://www.dst.gov.za/images/Attachments/Department_of_Science_and_Technology_SARIR_2016.pdf. [Accessed 11 Sep 2017].

[60] South African Department of Science and Innovation, “Human Capital Development Strategy for Research, Innovation and Scholarship,” 15 Jan 2016. [Online]. Available: https://www.dst.gov.za/images/Human-Capital-Development-Strategy-for-Research-Innovation-and-Scholarship.pdf. [Accessed 12 Dec 2019].

[61] S. A. D. o. S. a. Innovation, “The Ten-Year Plan for Science and Technology,” 2014. [Online]. Available: https://www.dst.gov.za/images/pdfs/The%20Ten-Year%20Plan%20for%20Science%20and%20Technology.pdf. [Accessed 12 Dec 2019].

[62] T. Manual, “National Development Plan 2030,” South African Government, 1 Aug 2016. [Online]. Available: https://www.gov.za/issues/national-development-plan-2030. [Accessed 22 Jul 2017].

[63] S. T. Manzini, “The national system of innovation concept: an ontological review and critique,” South African Journal of Science, vol. 108, no. 9 , pp. 9 - 10, 2012.

[64] National Research Foundation, “NRF Strategy 2020,” 10 Sep 2018. [Online]. Available: http://www.nrf.ac.za/sites/default/files/documents/NRF%20Strategy%20Implementation.pdf. [Accessed 10 Sep 2018].

[65] Department of Higher Education and Training, “Strategic Plans,” 1 Jul 2015. [Online]. Available: http://www.dhet.gov.za/Strategic%20Plans/Strategic%20Plans/Department%20of%20Higher%20Education%20and%20Training%20revised%20strategic%20plan%202015-16%20and%202019-20.pdf. [Accessed 12 Aug 2017].

[66] Council for Scientific and Industrial Research, “National Integrated Cyberinfrastructure System,” Council for Scientific and Industrial Research, Pretoria, 2018.

[67] South African Radio Astronomy Observatory (SARAO), “Square Kilometre Array,” SARAO, 2016. [Online]. Available: https://ww.skatelescope.org/. [Accessed 3 Apr 2017].

[68] United Nations, “Sustainable Development Goals,” United Nations, [Online]. Available: https://www.un.org/sustainabledevelopment/sustainable-development-goals/. [Accessed 16 Jul 2017]

[69] South African Department of Science and Technology, “South Africa's National Research and Development Strategy,” Department of Science and Technology, August 2002. [Online]. Available: https://www.gov.za/sites/default/files/rd_strat_0.pdf. [Accessed 3 Apr 2017].

[70] Government of the United Kingdom , “UK data capability strategy: seizing the data opportunity,” Government of the United Kingdom, 31 Oct 2013. [Online]. Available: https://www.gov.uk/government/publications/uk-data-capability-strategy. [Accessed 4 Jun 2017].

[71] Networking and Information Technology Research and Development (NITRD), “The Federal big Data Research and Development Strategic Plan,” 1 May 2016. [Online]. Available: https://bigdatawg.nist.gov/pdf/bigdatardstrategicplan.pdf. [Accessed 23 Mar 2017].

[72] European Commission, “Big Data,” European Commission, 2 Jul 2014. [Online]. Available: https://ec.europa.eu/digital-single-market/en/big-data. [Accessed 3 Apr 2017].

[73] S.-H. Doh, “Korea Artificial Intelligence and Big Data Strategies (Korean Version),”International Data Corporation, 2017. [Online]. [Accessed Oct 2017].

[74] M. S. Hajirahimova and A. S. Aliyeva, “Big Data strategies of the world countries,” in Proc. National Supercomputer forum (NSKF 2015), Russia, Pereslavl-Zalessky, 2015, pp. 24 - 27.

[75] A. Botta, N. Digiacomo and K. Mole, “Monetizing data: A new source of value in payments,” McKinsey & Company, Sep 2017. [Online].Available: https://www.mckinsey.com/industries/financial-services/our-insights/monetizing-dta-a-new-source-of-value -in-payments.[Accessed 28 Jan 2018].

[76] T. King, “Gartner: 75% of Organizations to Invest in Big Data by 2017,” Data Integration: Solutions Review, 13 Oct 2015. [Online]. Available: https://solutionsreview.com/data-integration/gartner-75-of-organizations-to-invest -in-big-data-by-2017/. [Accessed 4 Mar2017].

[77] European Commission, “Final results of the European Data Market study measuring the size and trends of the EU data economy,” European Commission, 2 May 2017. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-mar ket-study-measuring-size-and-trends-eu-data-economy. [Accessed 20 Jun 2018].

[78] South African Department of Science and Innovation, “White Paper on Science, Technology and Innovation,” 04 March 2019. [Online]. Available: https://www.dst.gov.za/images/2019/WHITE_PAPER_ON_SCIENCE_AND_TECHNOLOGY_web.pdf. [Accessed 12 Feb 2020].

[79] P. Maassen and B. Stensaker, “The knowledge triangle, European higher education policy logics and policy implications,” Higher Education, vol. 61, no. 6, pp. 757 - 769, 2010.

[80] I. Hashem, I. Yaqoob, N. Anuar, S. Mokhtar, A. Gani and S. Khan, “The rise of Big Data on cloud computing: Review and open research issues,” Information Systems, vol. 47, pp. 98 - 115, Jan 2015.

[81] W. L. Chang, “NIST: Wo L Chang,” US National Institute of Standards and technology, 26 Jun 2018. [Online].Available: https://www.nist.gov/people/wo-l-chang. [Accessed 11 Sep 2018].

[82] Wikipedia, “General Data Protection Regulation,” Wikipedia, 25 May 2018. [Online]. Available: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation. [Accessed 17 Sep 2018].

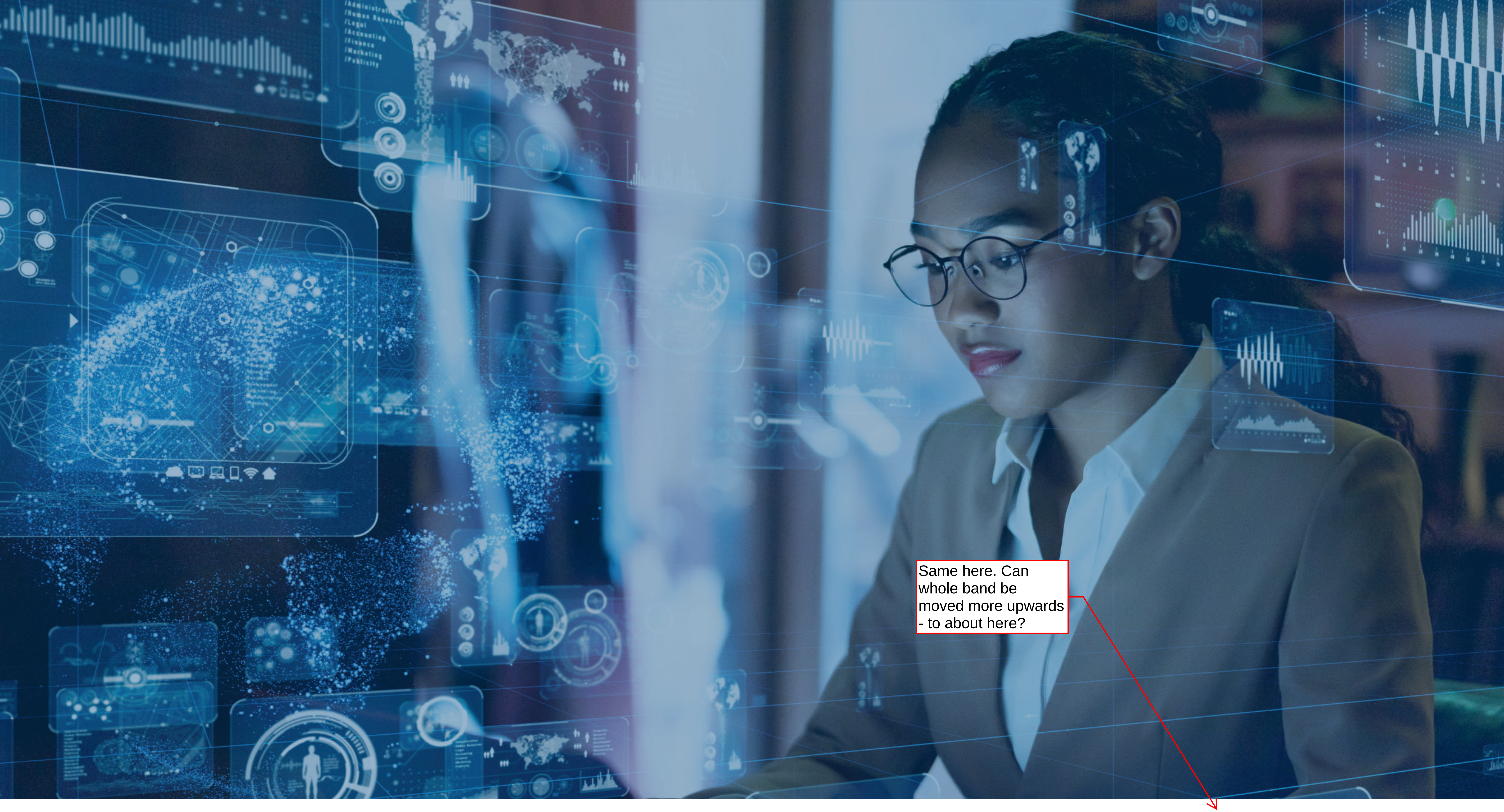
[83] Australian Government, “The Australian Public Service Big Data Strategy,” Australian Government, 2017. [Online]. Available: http://203.170.84.89/~idawis33/datasciences/?p=770. [Accessed 21 Apr 2017]
.

[84] Y. Lu, “China Artificial Intelligence and Big Data Strategies,” International Data Corporation, 2017. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=IDC_P28474. [Accessed 2 Nov 2017].

[85] South African Department of Higher Education and Training, “Revitalising and Transforming the Academic Profession,” Staffing South Africa's Universities Framework, 2015. [Online]. Available: http://www.ssauf.\dhet.gov.za/ngap.html. [Accessed 11 Sep 2018].

[86] W. Chang, N. Grady, and NIST Big Data Working Group, “NIST Big Data Interoperability Framework: Volume 1,” NIST, 2015.

[87] Alan Turing Institute, “The Alan Turing Institue,” 15 Jan 2018.[Online].



Same here. Can whole band be moved more upwards - to about here?

science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

nicis

NATIONAL INTEGRATED
CYBER INFRASTRUCTURE SYSTEM

CSIR

Touching lives through innovation

Department of Science and Innovation (DSI): https://www.dst.gov.za
National Integrated Cyberinfrastructure System (NICIS): https://www.nicis.ac.za