science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation | SARAO South African Radio Astronomy Observatory

# Data solutions, Innovation and Science Data Processing

DRAGN

PRESENTER: Khutso Ngoasheng

www.ska.ac.za

The South African Radio Astronomy Observatory (SARAO) is a National Facility managed by the National Research Foundation and incorporates all national radio astronomy telescopes and programmes. SARAO is responsible for implementing the Square Kilometre Array (SKA) in South Africa.

# Data solutions

- Traditional solutions:
  - Storage, connectivity, databases, processing (workstations, server farms, cloud, supercomputers), visualization, reporting, quality management
  - (almost) Linear
    - generate,
    - Format, associate & prepare,
    - interrogate, manipulate & explore (including predictive data science)
    - Present: visualize, reports, decision support (=> think google search)
    - Store, archive, retrieve, stage, secure
- (some) Modern challenges
  - Emergent awareness of the value of data, especially when used well (think google)
  - Speed and complexity $O(n^2)$ of decision support (self-flying drone swarms)
  - Dispersed infrastructure (era of unorganized abundance - where's your phone from 2009?)

# Innovation

- *Introducing something new…*
  - *Versus, or in the context of* **invention, design, art, mysticism, revolutions**
  - Interdisciplinary, transcendental, cross-sectional, translational, applicability, utility ...
- Drivers:
  - IP regiments (open source, patents vs. secrets)
  - Understanding of the true cost of omission, failure and forfeiture
  - Resilience and multi-applicability of the same base of innovation
- It's all about value
  - Value strategy (desire, vision)
  - Value planning (intent, complexity awareness, choices, characterization)
  - Value Engineering (portfolio, program & project management, systems thinking, systems engineering)
  - Value realization (delivery, utility: that *transformational experience* - things are irrevocably better)
  - Value capture (social benefit, sustainability, profitability, indefinitely iterative & incremental)

# Science data processing

- The capture, clean-up, qualification, and preparation of data products for scientific extraction;
- (followed by) iterative manipulation of said science data products to extract science
- Industrialized research ⇒ Data-driven science ⇒ Data deluge
  - Complex data streams (high volume, high speed, high availability, high competition)
  - Complex platforms and skills
    - Compute platforms (various high performance computing environments)
    - Rarefied quality domain technology knowledge
    - Rarefied and stratified domain knowledge
  - Emergent demand for operational excellence
- Metascience, metaengineering, metadata (replication crisis)

# Environmental context

- Global socio-economic divide and exclusion affects science participation
  - The excluded unaware (over 1 billion illiterate people on Earth)
  - The excluded, aware, hapless (STEM excluded)
  - The extracted (rare earth mineral countries, science instrument hosts, e.g. the geostationary orbit “host countries” debate)
  - The capable and participating
- The paradoxical double-edged sword held by digital technology
  - Reduced barriers to access
  - Superficial scarcity and socio-political constraints
    - **Intentional** e.g. trade, intellectual property regiments (think online research journals)
    - **Inadvertent** e.g. QWERTY alphabet, English-speaking, Western philosophy internet

# Data solutions *revisited*

- Solve the **right** set of problems
  - "*Right*" ever-changing
  - Socio-economic context matters
- Era of abundance, era of revolution, requires fundamentally different approaches and mindsets
- Imminent changes:
  - Data storage explosion (MeerKAT achieved > 40PB storage for < £1.5m in 2016-2017)
  - Unstructured, high volume data explosion
  - Exaflop computing
  - Green, resilient, performant edge computing
  - Low cost, universal high bandwidth internet coverage
  - Skill commoditization through automation and AI.
  - Context for science policy in economic policy (e.g. SKA as a treaty organization)

# Innovation *revisited*

- Current triggers in data science
  - Artificial intelligence, virtual reality revivals
  - Genuine and truly achievable social benefit: Education, Geospatial sciences, Disaster management, Health, etc.
  - Critical skill and platform availability at scale: e.g. low cost HPC vs. supercomputing, Serverless computing & cloud platform (infrastructure, software) vs. personal compute infrastructure overheads
  - Open and shared innovation (e.g. open source hardware and software, open data)
  - Hype & political support
  - Space industry 2.0
- Solutions beyond data for science
  - Accessibility, empowerment of the excluded
  - Contextualised, hyper-localised solution opportunities
  - Multi-use (MeerKAT compute solutions potentially used in Geospace, etc.)

# Innovation *Opportunities*

- Global complex compute infrastructure systems
- Functional, affordable, available, reliable compute at the edge / fog
- Education, training and deeper user diversification
- Security, veracity, validity and long term preservation of data, methods, platforms
- Democratization and attainment of equitable 4IR
- Reduction of complexity over time
- ….*infinite*

# Science Data Processing *revisited*

- Concentration ⇌ Dispersion of science facilities and systems pendulum
- Current issues:
  - Virtual facilities
  - Data-driven science and research
  - Metascience, meta-engineering
  - Value demonstration, value capture, justification of science investment
  - Economic and political ownership
  - Multi-instrument data sources
- Automated, serendipitous research
- Advanced and smart tooling - MeerKAT to produce > 12TB of data a day. This excludes commensally derived, duplicated, re-processed, concatenated, meta-, raw, or engineering data. No human can possibly inspect this data with the current set of tools.

# Science Data Processing *opportunities*

- **Virtual facilities.** These are low cost, remote-accessible, secure, open. They overcome significant barriers to accessing high end science facilities, data, and have very unique opportunities.

**'No image will surpass this': Hubble telescope astronomers created a stunning picture of the deep universe with 16 years' worth of photos**

Dave Mosher , Business Insider US, May 04, 2019, 09:45 AM

- Tools for data visualization, collection, collation and analytics of unorganized data
- Development of educational, skills development programs for technology, data science, domain knowledge, and research.
- Development of appropriate infrastructure and data platforms within regions for multiple uses.

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation | SARAO South African Radio Astronomy Observatory

# Contact information

**Khutso Ngoasheng**
**Manager: Science Processing**
**Email: khutso@ska.ac.za**