The Known Knowns, The Known Unknowns and The Unknown Unknowns of Geophysics Data Processing in 2030

Lesley Wyborn[1], *Nigel Rees*[1], *Jens Klump*[2], *Ben Evans*[1], *Tim Rawling*[3], *and Kelsey Druken*[1]

[1]*Australian National University, National Computational Infrastructure, Acton, Australia*[4]
[2]*CSIRO Mineral Resources, CSIRO, Perth, Australia*
[3]*AuScope Ltd, University of Melbourne, Melbourne, Australia*

AuScope

CSIRO

Australian Research Data Commons

NCI AUSTRALIA

LAND ACKNOWLEDGEMENT

I would like to acknowledge the Traditional Owners of the lands, seas and waters of the areas that I live, work and meet on.

I acknowledge their continuing connection to their culture and pay my respects to their Elders past, present and future.

'Gadi' by Lynnice Church for NCI Australia 2020

lesley.wyborn@anu.edu.au

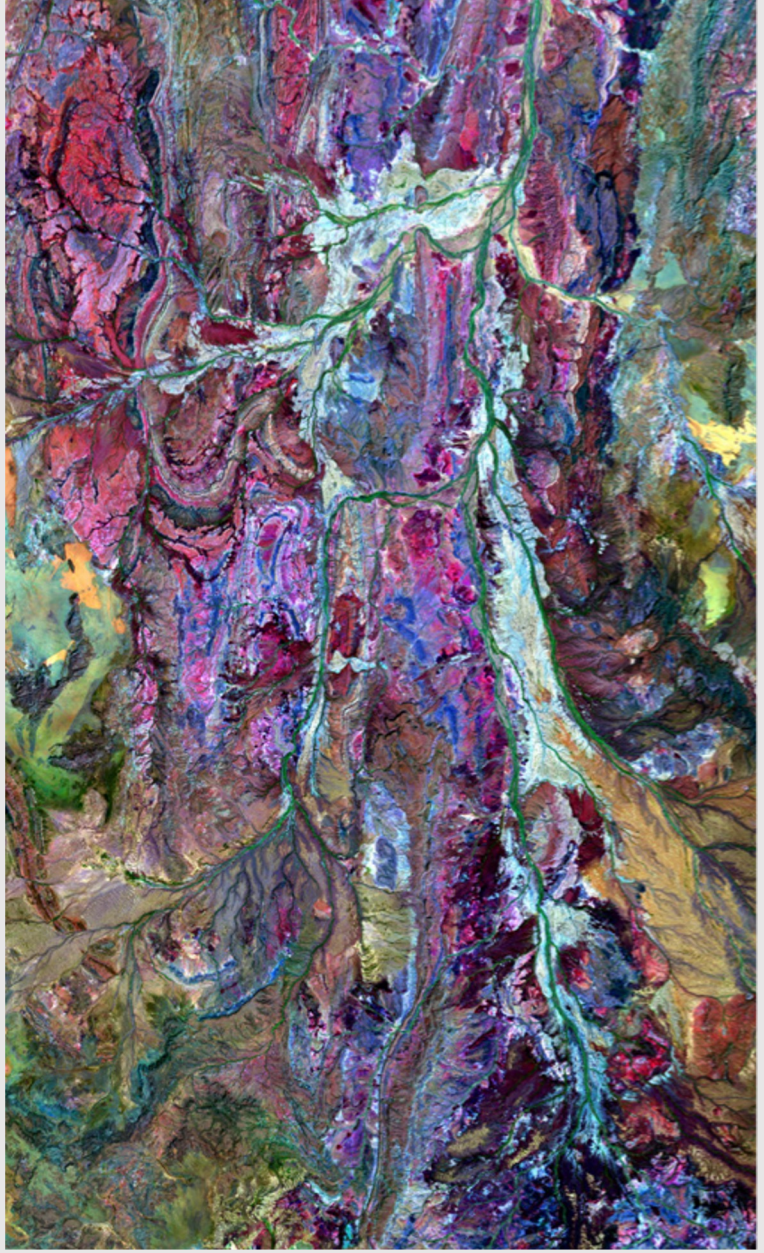
The 2030 Geophysics Collection Project

- The project is a funded collaboration between AuScope, NCI, TERN and ARDC (https://ardc.edu.au/project/2030-geophysics-collections/)
- It seeks to:
  1. Make national-scale high-resolution geophysics datasets suitable for programmatic access in HPC;
  2. Lay the foundations for more rapid data processing by 2030 next-generation scalable, data-intensive computation including Artificial Intelligence (AI)/Machine Learning (ML) and data assimilation.
- The project is not about building systems for the infrastructures and stakeholder requirements of today, rather it is about positioning geophysical data collections to be capable of taking advantage of next generation technologies and computational infrastructures by 2030.
- **What are the knowns and unknowns?**

lesley.wyborn@anu.edu.au

What are the opportunities of 2030 computing?

1. So often today's research is undertaken on pre-canned, analysis-ready datasets (ARD) that are tuned towards the highest common denominator as determined by the data owner.
2. By 2030:
   - Increased computational power co-located with fast-access storage systems will mean that geophysicists will be able to work on less processed data levels and then transparently develop their own derivative products.
   - As researchers will be able to see and test the quality of their algorithms more rapidly, there will be multiple versions of open source software used as researchers fine tune individual algorithms to suit their specific requirements.
   - We will be capable of more precise solutions and in hazards space and other relevant areas, analytics will be done in faster-than-real-time.



Landsat-8 image courtesy of the U.S. Geological Survey

NCI AUSTRALIA

lesley.wyborn@anu.edu.au

AuScope CSIRO Australian Research Data Commons

So what are the known knowns of 2030 computing?

1. High-end computational power will be at Exascale
2. Today's emerging collaborative platforms will continue to evolve as a mix of HPC and cloud.
3. Data volumes will be measured in Zettabytes (10^{21} bytes), which is about 10 times more than today.
4. Currently we discuss Big Data Vs (volume, variety, value, velocity, veracity, etc)
5. By 2030 the focus will be on Big Data Cs (community, capacity, confidence, consistency, clarity, crumbs, etc).
6. **It will be mandatory for data discovery, accessibility, interoperability and reuseability to be fully machine-to-machine as envisaged by the FAIR principles in 2016.**

Photo courtesy of NCI Australia

NCI AUSTRALIA

lesley.wyborn@anu.edu.au

AuScope CSIRO Australian Research Data Commons

And there are the emerging known unknowns…

To have any confidence in any data product, we will need to have transparency throughout the whole scientific process.

How at scales of 2030 computation will we:

- Preserve and make transparent any result from this diversity and flexibility including:
  - the exact software used?
  - the precise version of the data accessed?
  - the platforms utilised, etc.?
- Vouch for the fidelity of scientific outputs and ensure they can be consistently replicated to establish trust?
- Preserve who funded what components so that sponsors can see which investments had the greatest impact, uptake and ROI?

Photo courtesy of NCI Australia

lesley.wyborn@anu.edu.au

And what are the unknown unknowns…

- Of course we don’t know what they are!
- But we do know these will progressively be exposed to us in the next decade as the scale and speed at which collaborative research is undertaken increases…
- Join us on this journey!

Photo courtesy of NCI Australia

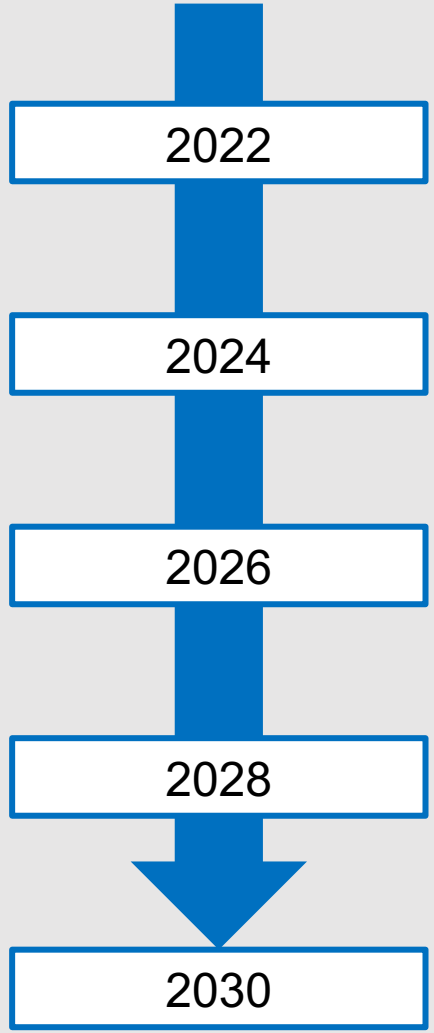
lesley.wyborn@anu.edu.au

Preparing for 2030: we no longer have the gift of time

Photo by Nathan Dumlao on Unsplash

We need to

1. Start working now on more automated systems that capture provenance through successive levels of processing, including:
   - How it was produced?
   - Which exact dataset/dataset extract was used?
   - Who funded what?
2. Ensure we will still be able to make less processed forms of data more accessible and able to be aggregated into seamless global high-resolution datasets.
3. Start working now on making all metadata, data and vocabularies FAIR and machine actionable.
4. Ensure that whatever we do, it is scalable into the future



lesley.wyborn@anu.edu.au

Thank you and for More Information:

1. Australian National University, National Computational Infrastructure, Acton, Australia
   - Lesley Wyborn (lesley.wyborn@anu.edu.au)
   - Nigel Rees (nigel.rees@anu.edu.au)
   - Ben Evans (ben.evans@anu.edu.au)
   - Kelsey Druken (kelsey.druken@anu.edu.au)
2. CSIRO Mineral Resources, CSIRO, Perth, Australia
   - Jens Klump (jens.klump@csiro.au)
3. AuScope Ltd, University of Melbourne, Melbourne, Australia
   - Tim Rawling (tim@auscope.org.au)

Citation: Wyborn, L., Rees, N., Klump, J., Evans, B., Rawling, T., and Druken, K.: The Known Knowns, the Known Unknowns and the Unknown Unknowns of Geophysics Data Processing in 2030 , EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-11012, https://doi.org/10.5194/egusphere-egu22-11012 , 2022.

Project Information: 2030 Geophysics Collections