Towards World-class Data-intensive Earth and Environmental Science Research in 2030: Will Today's Practices in Data Repositories Get Us There?

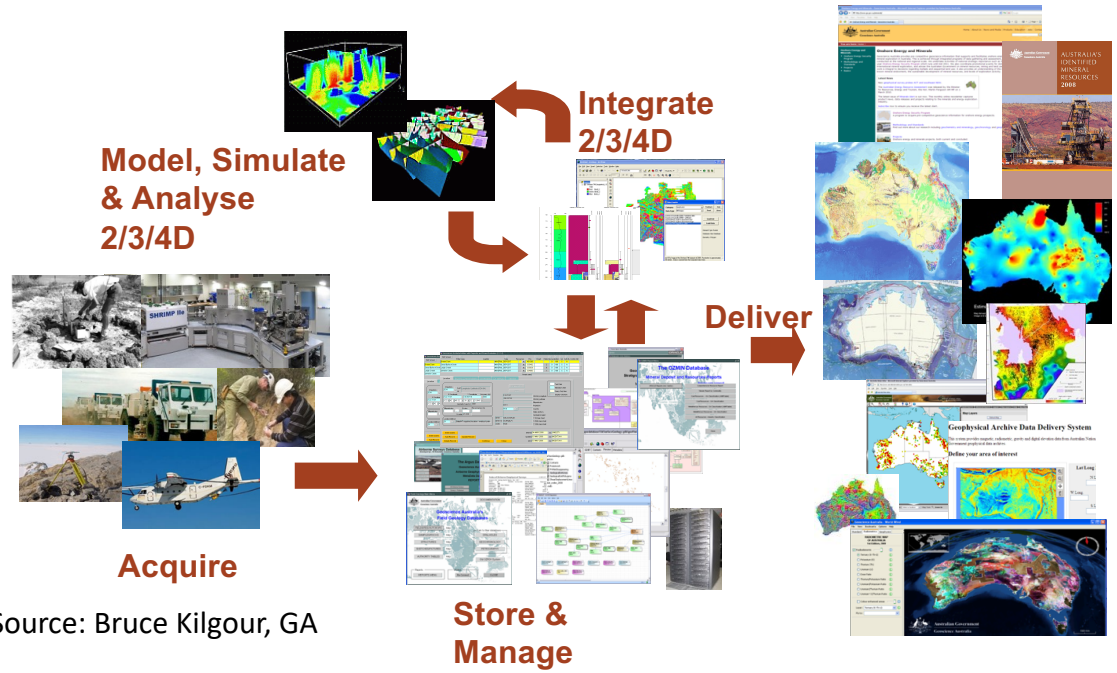
Lesley Wyborn

National Computational Infrastructure, ANU, Australia

Contributing Earth & Environmental Scientific Observations to the UN SDGS

Internationally Earth and environmental Science datasets have the potential to contribute significantly to resolving major societal challenges such as those outlined in the United Nations 2030 Sustainable Development Goals (SDGs).

But by 2030 will we be able to make a valuable contribution?



Source: Bruce Kilgour, GA

Source: https://sustainabledevelopment.un.org/?menu=1300

The Diversity of Earth and environmental science datasets → ??? → The UN Sustainable Development Goals

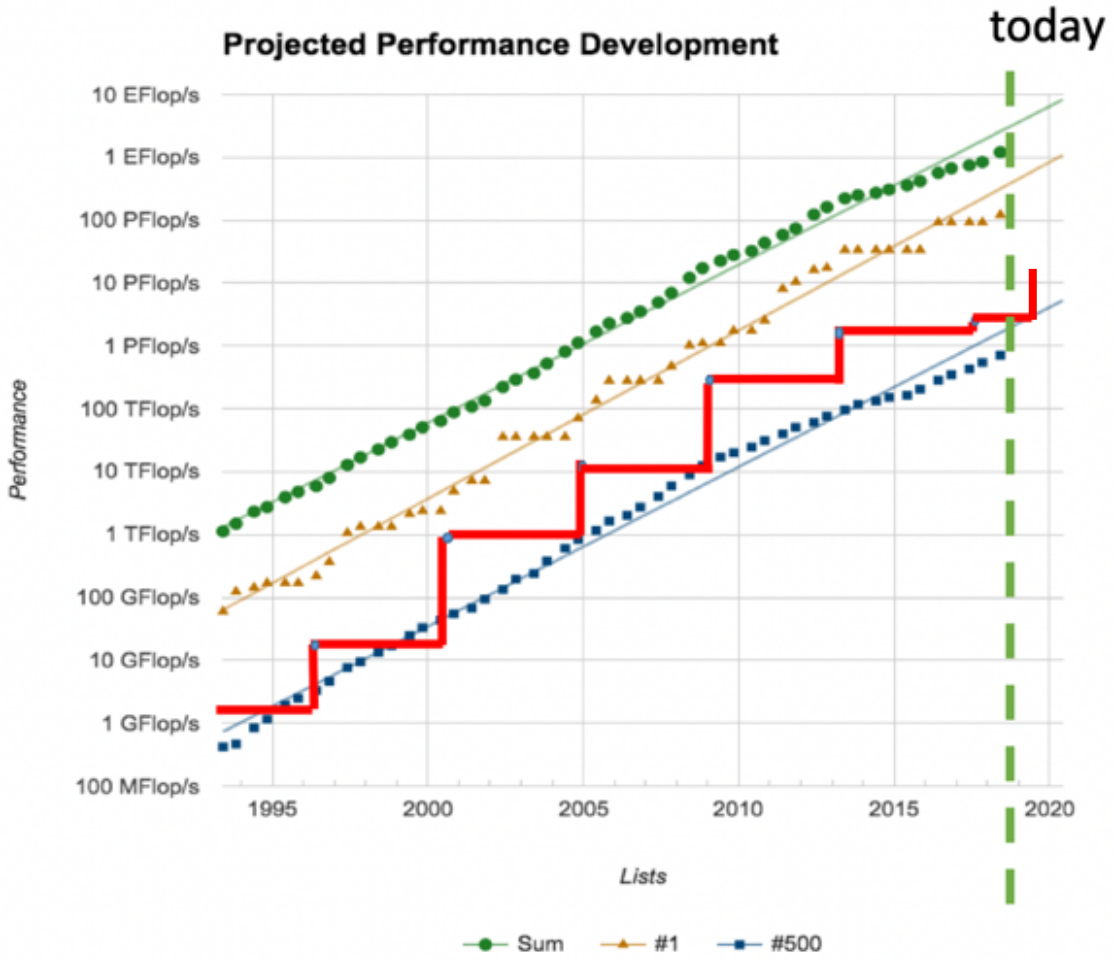
Setting the Scene on ‘Towards 2030’

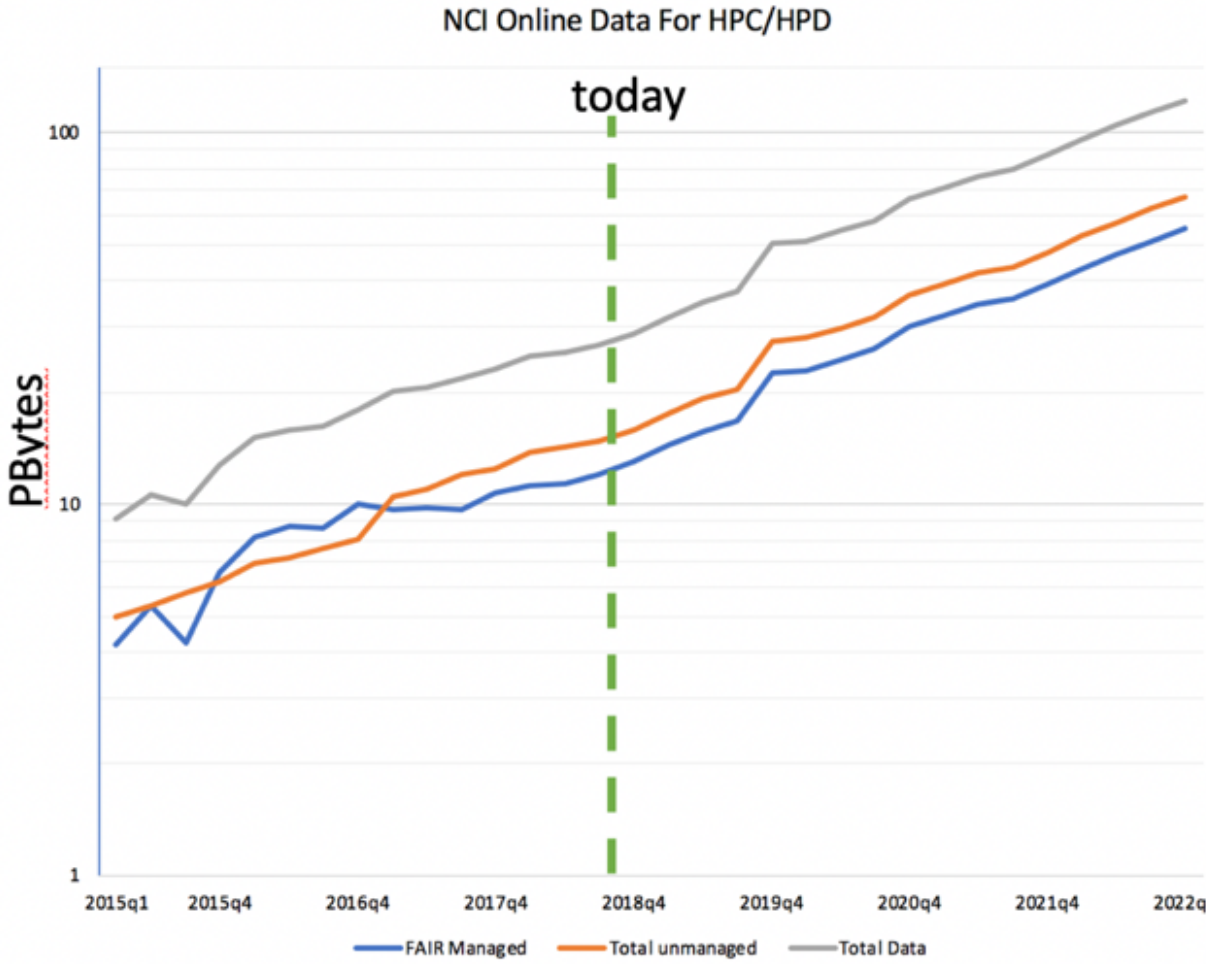
Source: https://thenationaldigest.com/a-life-on-our-planet-is-sir-david-attenboroughs-legacy-to-the-world/

- By 2030, we know that leading-edge computational infrastructures will be Exascale (repositories, supercomputers, cloud, etc).
- Combined these will facilitate solving of research challenges at scales, resolutions and within timeframes that cannot be achieved today.
- We know that computers and codes are inexorably moving towards Exascale, but for data, the path is less clear and highly uncertain.
- Hence, by 2030, the capability for Earth and environmental science researchers to make valued contributions to the SDGs will depend on developing a capacity to ‘integrate’ data from globally distributed, heterogeneous repositories and then make it accessible to HPC.
- Some questions:
    1. Are we on the right path to achieve this?
    2. Which exemplar public domain solid Earth science projects are working towards exascale and 2030?
    3. Are they doing something different?

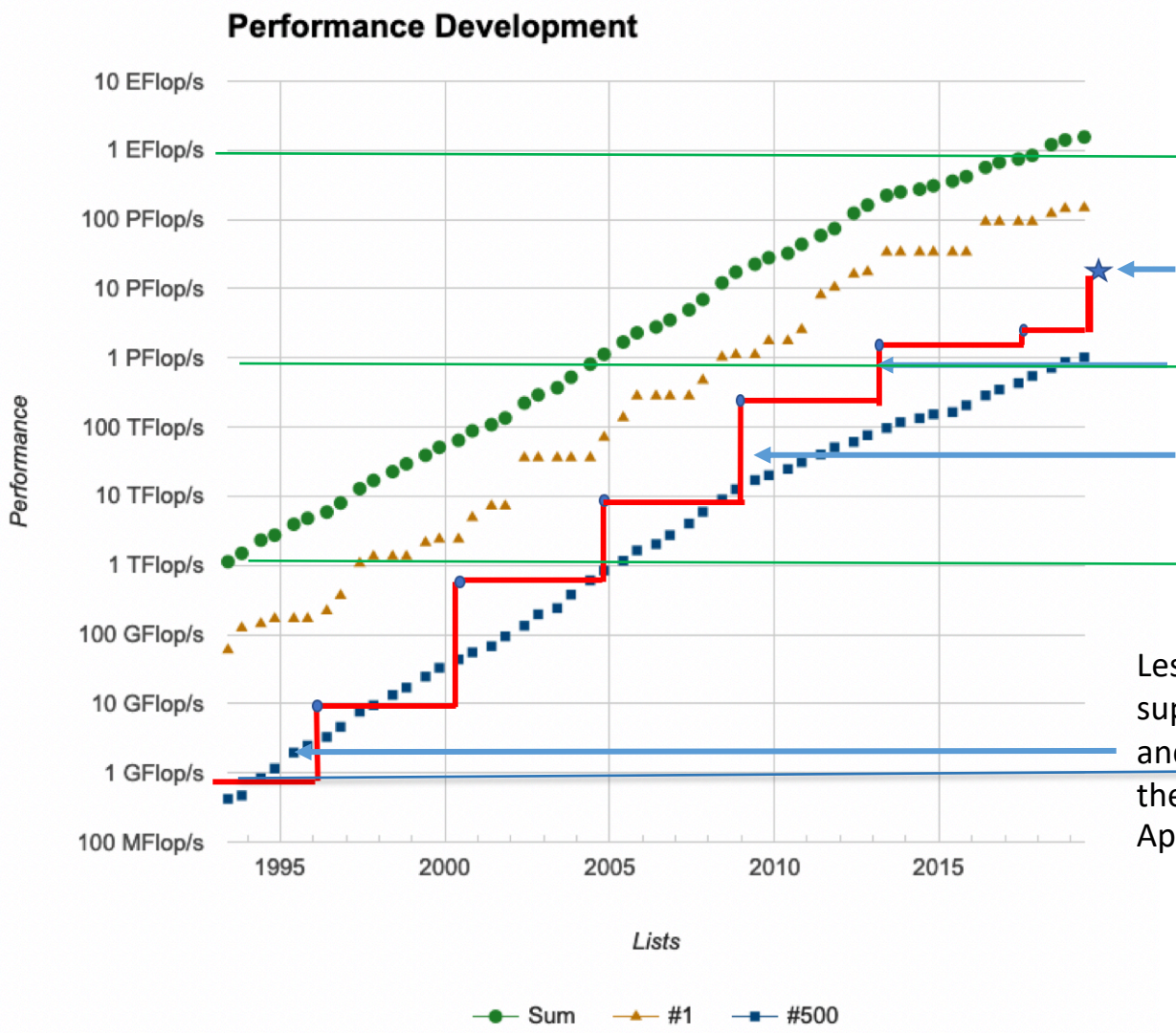
lesley.wyborn@anu.edu.au

HPC and Big Data will keep growing: data grows exponentially





But on individual HPC Top 500 Super Computers growth is in steps…



For NCI Australia it really is a **step** change!

Exascale

NCI Gadi **Step** Change

Petascale

NCI Raijin **Step** Change

NCI Vayu **Step** Change

Terascale

Lesley's iPhone was a supercomputer in 1997 and more powerful than the computers that landed Apollo 11 on the moon

Gigascale

NCI Australia

Exemplar Public Domain Project 1: the EU ChEESE project

EUROPEAN RESEARCH INFRASTRUCTURE ON SOLID EARTH

CONTACT US

Search this site

EPOS EUROPEANPLATEOBSERVINGSYSTEM

HOME ABOUT WHO BENEFITS DATA & SERVICES NEWS | EVENTS & DOCS EPOS ERIC

Home

Riding the wave of the future in supercomputing: Center of Excellence for Exascale in Solid Earth (ChEESE) will share Exascale-compatible codes on EPOS repository

ChEESE

Center of Excellence for Exascale in Solid Earth

Many parts of Europe are exposed to geohazards such as earthquakes, landslides, tsunamis and volcanic eruptions. With civil protection as a primary objective, the HPC Center of Excellence (CoE) ChEESE - Center of Excellence for Exascale in Solid Earth - has been created to become a hub for High-Performance Computing (HPC) software within the solid Earth community. It will enable services such as urgent computing, hazard assessment and early warning using flagship simulation codes that will run efficiently on future European Exascale HPC systems.

ChEESE will enable HPC-based codes and related services for hazard, early warning and earth sub-surface characterization on the EPOS repository. This will allow the solid Earth community, including civil protection agencies and other stakeholders, to access these codes and toolkits easily. The CoE also aims at providing specialist training on services and capacity building measures.

Coordinated by Barcelona Supercomputing Center (BSC), the ChEESE main objective is to address 15 scientific, technical and socio-economic Exascale computational challenges in the domain of solid Earth. To accomplish this task, it has been awarded with €7.7 million from European Commission funding over three years.

https://www.epos-ip.org/riding-wave-future-supercomputing-center-excellence-exascale-solid-earth-cheese-will-share-exascale

1. ChEESE (Centre for Exascale in Solid Earth):
   - **i. Has €7.7M funding over 3 years**
   - ii. assessments for earthquakes, volcanoes and tsunamis. Has been created to become a hub for HPC software within the solid Earth community
   - iii. Will enable HPC based codes and related services for hazard, early warning and earth sub-surface characterization to connect to the EPOS data repository
   - iv. Is preparing 10 community flagship European codes to run efficiently on upcoming pre-Exascale and Exascale supercomputers.
   - v. It is also developing 12 pilot demonstrators requiring Exascale computing on near real-time seismic simulations and full-wave inversion, ensemble-based volcanic ash dispersal, faster-than-real-time tsunami simulations, and physics-based hazard assessments.

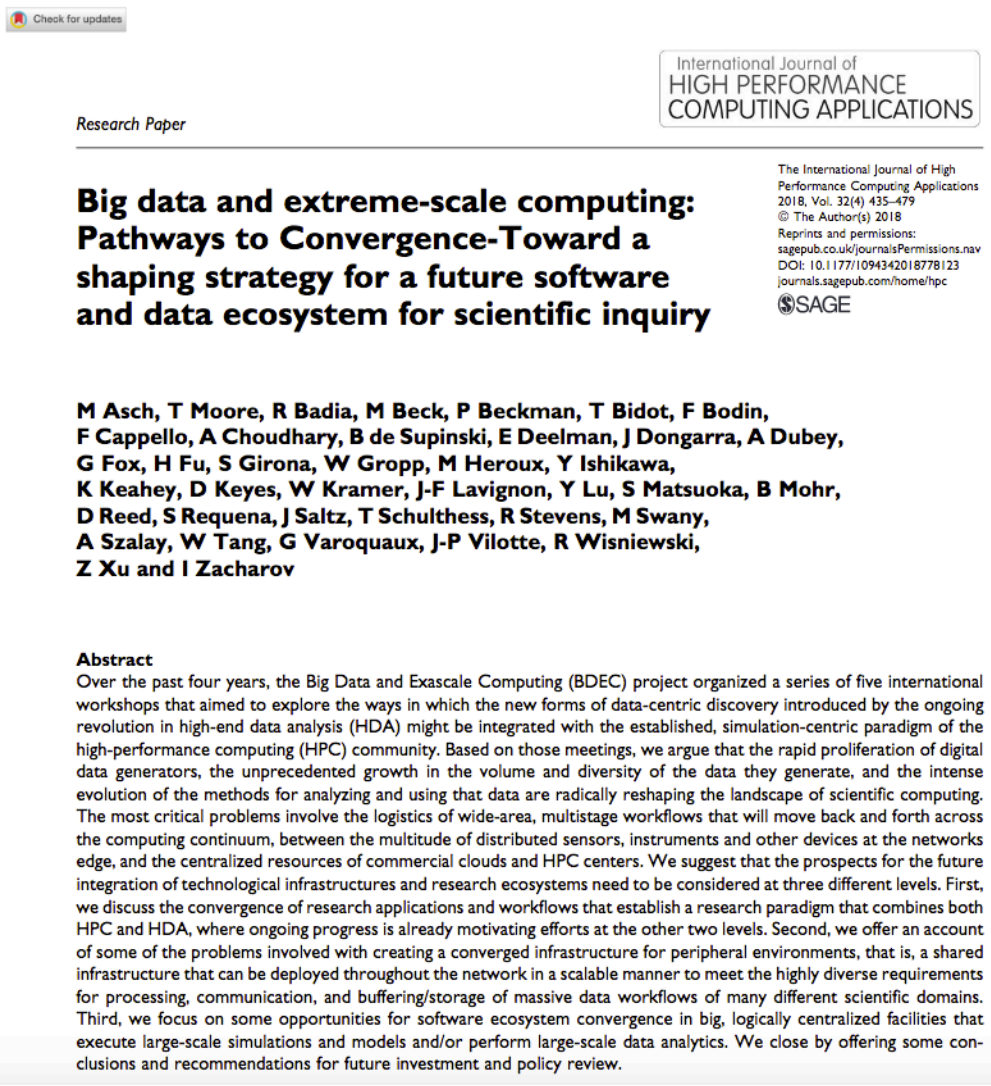
Lesley Wyborn (lesley.wyborn@anu.edu.au)

Exemplar 1: What is ChEESE working on?

No	Pilot Demonstrator name	Area	Flagship Code	Related ECC	Initial TRL	Target TRL	Related service
1	Urgent seismic simulations	CS	ExaHyPE, Salvus, SPECFEM3D	ECC3	3	5-6	Urgent computing
2	Faster than real-time tsunami simulations	T	T-HySEA, L-HySEA	ECC11	2	6-7	Urgent computing
3	High-resolution volcanic plume simulation	PV	ASHEE, FALL3D	ECC7	1	4	None
4	Physics-based tsunami-earthquake interaction	CS	SeisSol, ExaHyPE	ECC3, ECC12	2	4	None
5	Physics-based probabilistic seismic hazard assessment (PSHA)	CS	SeisSol, ExaHyPE, AWP-ODC(*)	ECC4	4	6-7	Hazard assessment
6	Probabilistic volcanic hazard assessment (PVHA)	PV	FALL3D	ECC9, ECC10	3	6-7	Hazard assessment
7	Probabilistic tsunami hazard assessment (PTHA)	T	T-HySEA L-HySEA	ECC14	3	5-7	Hazard assessment
8	Probabilistic Tsunami Forecast (PTF) for early warning and rapid post event assessment	T	T-HySEA	ECC12, ECC13	3	6-8	Early warning
9	Seismic tomography	CS	SPECFEM3D, Salvus	ECC1, ECC2	4	6	Other
10	Array-based statistical source detection and restoration and Machine learning from monitoring	CS PV	BackTrackBB (**)	ECC15	2	4	None
11	Geomagnetic forecasts	MHD	PARODY_PDAF, XSHELLS	ECC5, ECC6	2	4	None
12	High-resolution volcanic ash dispersal forecast	PV	FALL3D	ECC8	3	6-7	Urgent computing

*ChEESE are well and truly working on **F**aster than **R**eal **T**ime **C**omputing (FRTC) and at Exascale!!*

Exemplar 2. The international Big Data and extreme-scale computing white paper.



Check for updates

International Journal of HIGH PERFORMANCE COMPUTING APPLICATIONS

Research Paper

Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry

The International Journal of High Performance Computing Applications 2018, Vol. 32(4) 435–479

DOI: 10.1177/1094342018778123
journals.sagepub.com/home/hpc
SAGE

M Asch, T Moore, R Badia, M Beck, P Beckman, T Bidot, F Bodin, F Cappello, A Choudhary, B de Supinski, E Deelman, J Dongarra, A Dubey, G Fox, H Fu, S Girona, W Gropp, M Heroux, Y Ishikawa, K Keahey, D Keyes, W Kramer, J-F Lavignon, Y Lu, S Matsuoka, B Mohr, D Reed, S Requena, J Saltz, T Schulthess, R Stevens, M Swany, A Szalay, W Tang, G Varoquaux, J-P Vilotte, R Wisniewski, Z Xu and I Zacharov

Abstract

Over the past four years, the Big Data and Exascale Computing (BDEC) project organized a series of five international workshops that aimed to explore the ways in which the new forms of data-centric discovery introduced by the ongoing revolution in high-end data analysis (HDA) might be integrated with the established, simulation-centric paradigm of the high-performance computing (HPC) community. Based on those meetings, we argue that the rapid proliferation of digital data generators, the unprecedented growth in the volume and diversity of the data they generate, and the intense evolution of the methods for analyzing and using that data are radically reshaping the landscape of scientific computing. The most critical problems involve the logistics of wide-area, multistage workflows that will move back and forth across the computing continuum, between the multitude of distributed sensors, instruments and other devices at the networks edge, and the centralized resources of commercial clouds and HPC centers. We suggest that the prospects for the future integration of technological infrastructures and research ecosystems need to be considered at three different levels. First, we discuss the convergence of research applications and workflows that establish a research paradigm that combines both HPC and HDA, where ongoing progress is already motivating efforts at the other two levels. Second, we offer an account of some of the problems involved with creating a converged infrastructure for peripheral environments, that is, a shared infrastructure that can be deployed throughout the network in a scalable manner to meet the highly diverse requirements for processing, communication, and buffering/storage of massive data workflows of many different scientific domains. Third, we focus on some opportunities for software ecosystem convergence in big, logically centralized facilities that execute large-scale simulations and models and/or perform large-scale data analytics. We close by offering some conclusions and recommendations for future investment and policy review.

In 2018 this report highlighted:

i. "The rapid proliferation of digital data generators, the unprecedented growth in the volume and diversity of the data they generate, and the intense evolution of the methods for analyzing and using that data are radically reshaping the landscape of scientific computing."

ii. "The need for big, logically centralized facilities that execute large-scale simulations and models and/or perform large-scale data analytics".

iii. "Addressed three modalities of data provenance:
   i. data arriving from the edge (often in real time), never centralized;
   ii. federated multisource archived data; and
   iii. combinations of data stored from observational archives with a dynamic simulation."

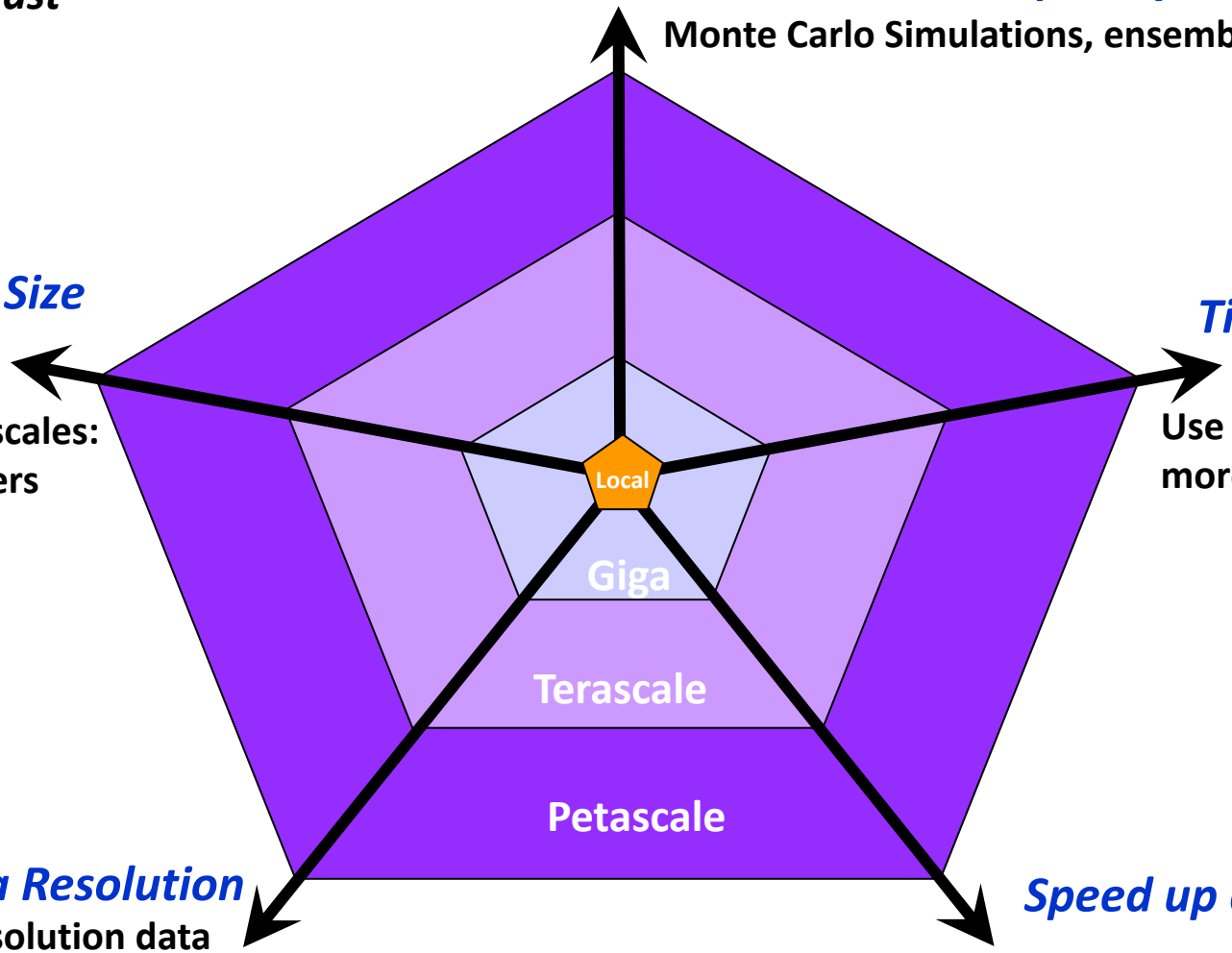
iv. "The conventional strategy of back-hauling all data across a fast link to the cloud or data center is no longer a viable option for many applications".

Today we are moving more and more data onto the cloud and/or distributed data centres. But some calculations are arguing this may not work at Exascale, particularly for FTRT use cases: costs are prohibitive.

https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec_pathways.pdf

How we handled the Step Change when a new supercomputer arrived in 2010

2010 Version: there are at least 5 ways of scaling up

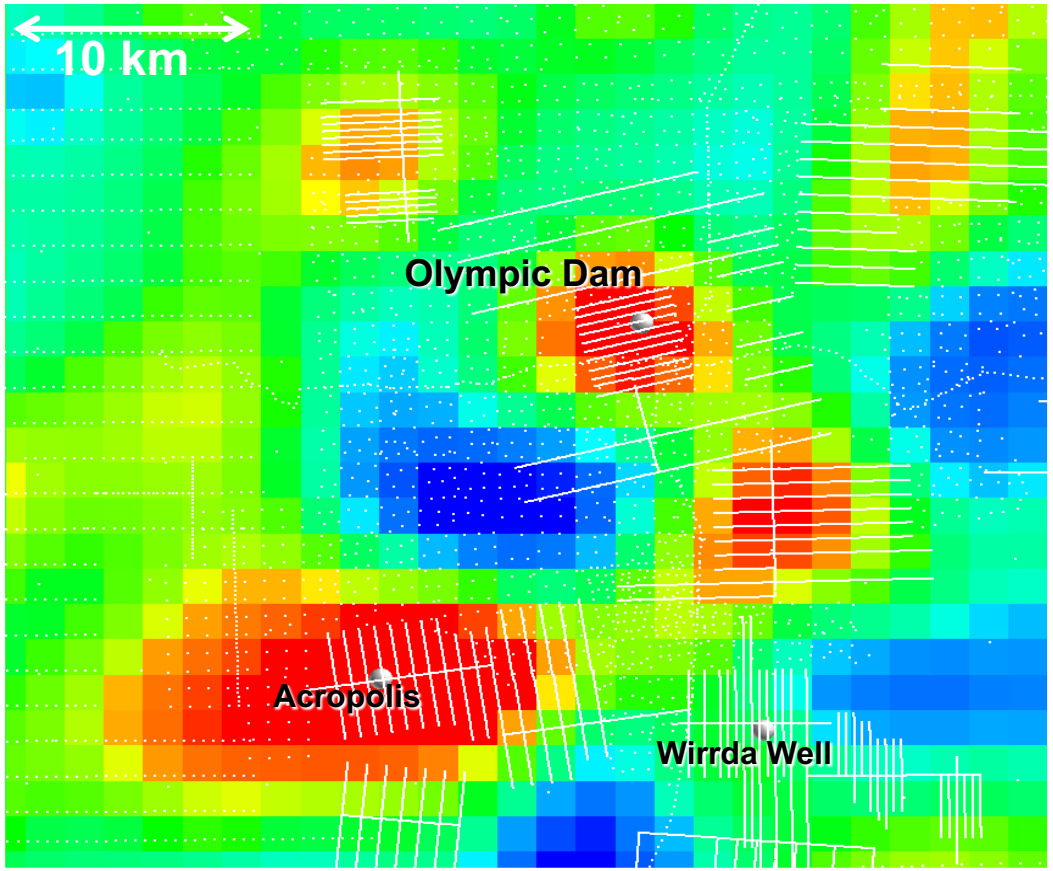


Based on European Climate Computing Environments, Bryan Lawrence (http://home.badc.rl.ac.uk/lawrence/blog/2010/08/02)

An example of scaling up in 2012: increasing analytical resolution

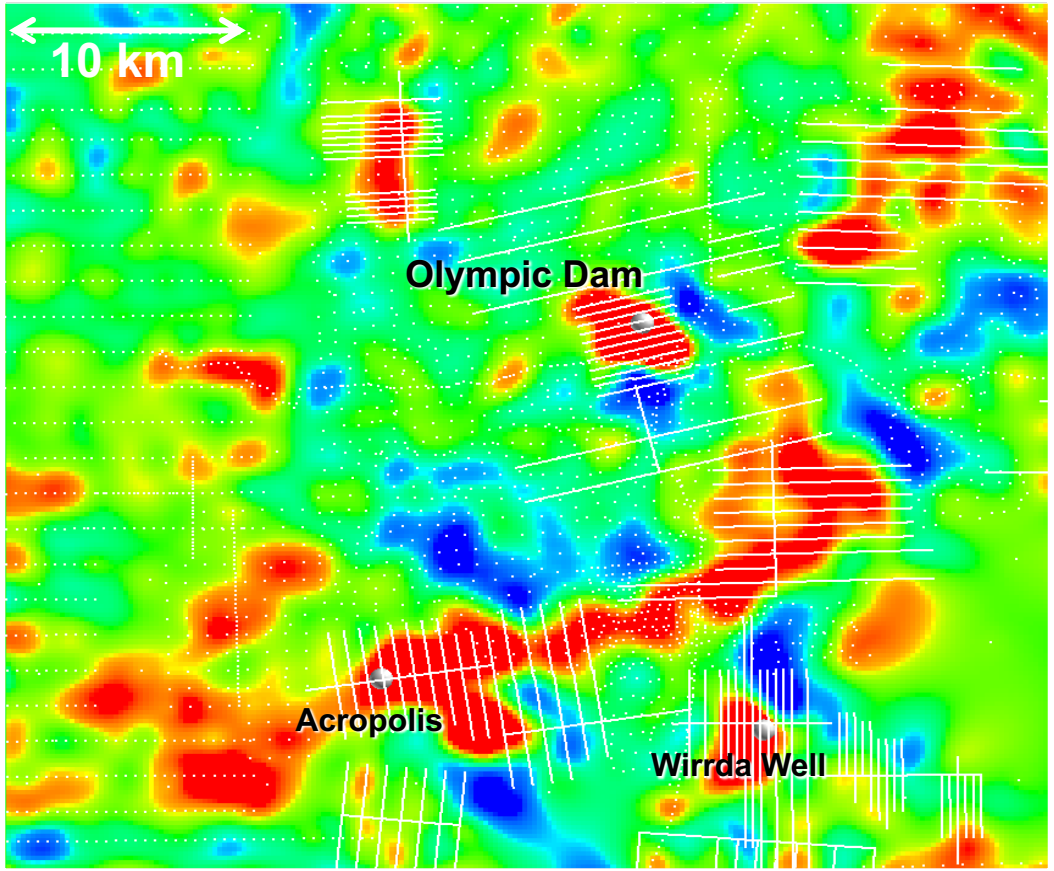
The same 11,322 gravity observations are used in both these images, but the one on the left was down sampled because of limited computational Capacity. The one on the right was at the capacity of the new super computer

Cell size: 2 km x 2 km x 1 km



Input : 2 KB text input data file
Output: 60 KB text model file

Cell size: 250 m x 250 m x 200 m



Input : 827 KB text input data file
Output: 27 MB text model file

Source: Nick Williams in 2010

How we handled the Step Change when a new supercomputer arrived in 2020

2020 Version: the same 5 ways still apply, but accessing data at high resolutions, in sufficient volumes and in realistic time frames is becoming increasingly difficult: current practices work against this.

Increase Model Complexity

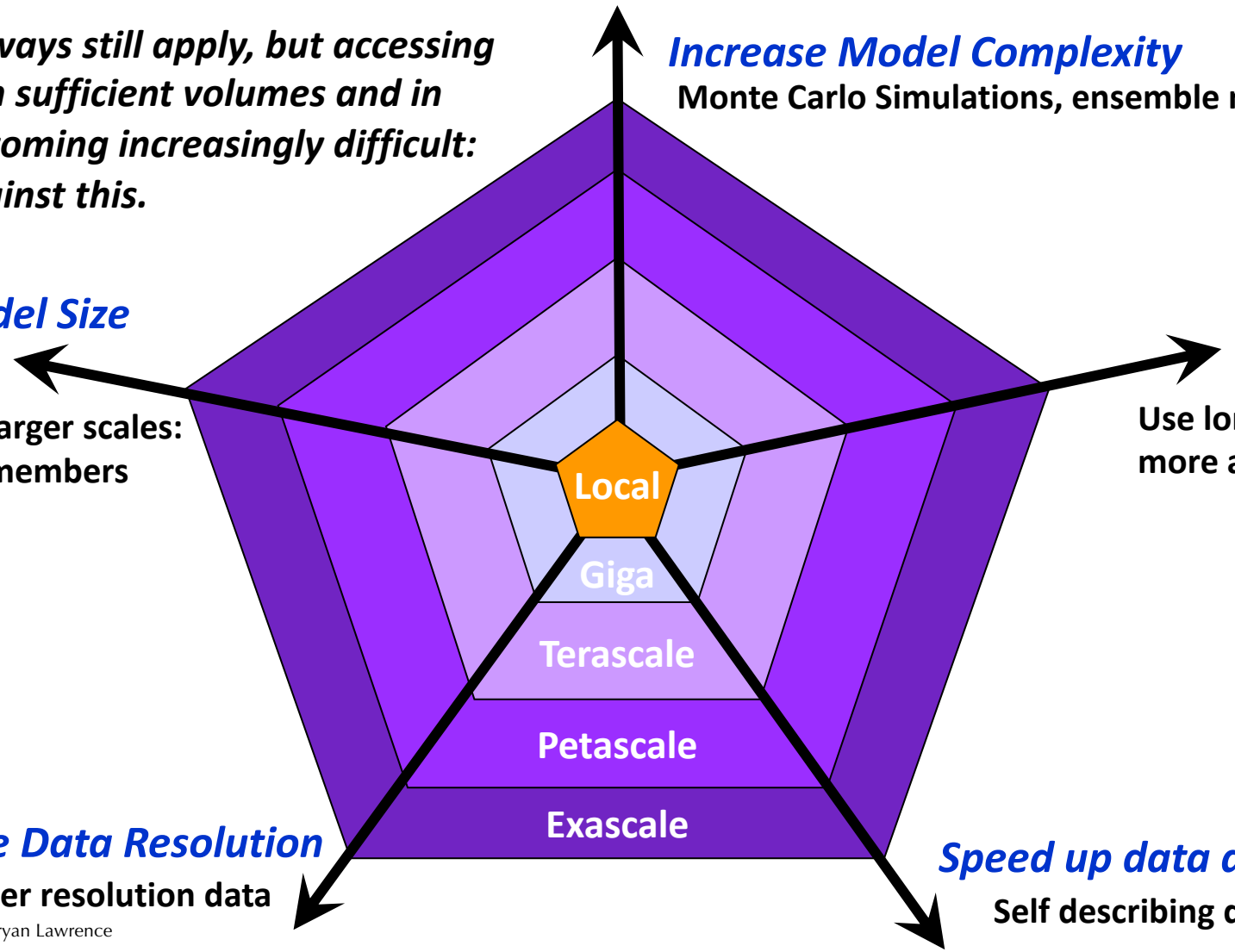
Monte Carlo Simulations, ensemble runs

Increase Model Size

Single passes at larger scales: more ensemble members

Timescale

Use longer duration runs: use more and shorter time intervals



Increase Data Resolution

Use higher resolution data

Speed up data access

Self describing data arrays and data cubes

Based on European Climate Computing Environments, Bryan Lawrence (http://home.badc.rl.ac.uk/lawrence/blog/2010/08/02)

So what practices are working against this?

- Through initiatives such as the Commitment Statement from the Coalition for Publishing Data in the Earth and Space Sciences, publishers are now requiring that datasets that support a publication be:
  - curated and stored in a 'trustworthy' repository that can provide a DOI and a landing page for that dataset,
  - if possible, can also provide some domain quality assurance to ensure that data sets are not only Findable and Accessible, but also Interoperable and Reusable.
- But tensions result as the demand for suitable **domain** expertise to provide the "I" and the "R" is far exceeding what is available.
- **The Main Tensions** are:
  1. As a last resort, frustrated researchers are simply depositing the datasets that support their publications into generic repositories such as Figshare and Zenodo, which simply store the file of the data: rarely are domain-specific QA/QC procedures applied to the data. Hence, the data cannot not easily be aggregated into national, let alone global reference collections.
  2. Most data that supports publications is highly processed, at Level 3 or Level 4

lesley.wyborn@anu.edu.au

What do I mean by Levels 1-4: introducing the NASA Data Processing Levels?

Level	Description
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artifacts (e.g., synchronization frames, communications headers, duplicate data) removed.
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (e.g., platform ephemeris) computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units (not all instruments have Level 1B source data).
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

These less processed levels are rarely linked to higher processing level data products in publications and elsewhere

Data supporting publications is mostly L3/L4

Source: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

Institutional/Generic Repositories vs Domain Repositories

- The positive is that generic/institutional repositories do ensure that data is not sitting on inaccessible personal c-drives and USB drives, but the data files are rarely interoperable.
- Interoperability can only be achieved by repositories that have the domain expertise to curate the data properly and ensure that the data meets minimum community standards and specifications, that will ultimately enable aggregation into global reference sets.
- In addition, most researchers only deposit the files that support a particular publication, and as these files can be highly processed and generalized, they can be difficult to reuse outside of the context of a specific research publication.
- Most generic/institutional repositories cannot take Terascale or Petascale data sets, let alone Exascale!

Source: modified from https://upload.wikimedia.org/wikipedia/commons/6/65/To_deposit_or_not_to_deposit%2C_that_is_the_question_-_journal.pbio.1001779.g001.png

An example of the publication dilemma from MagnetoTelluric (MT) data

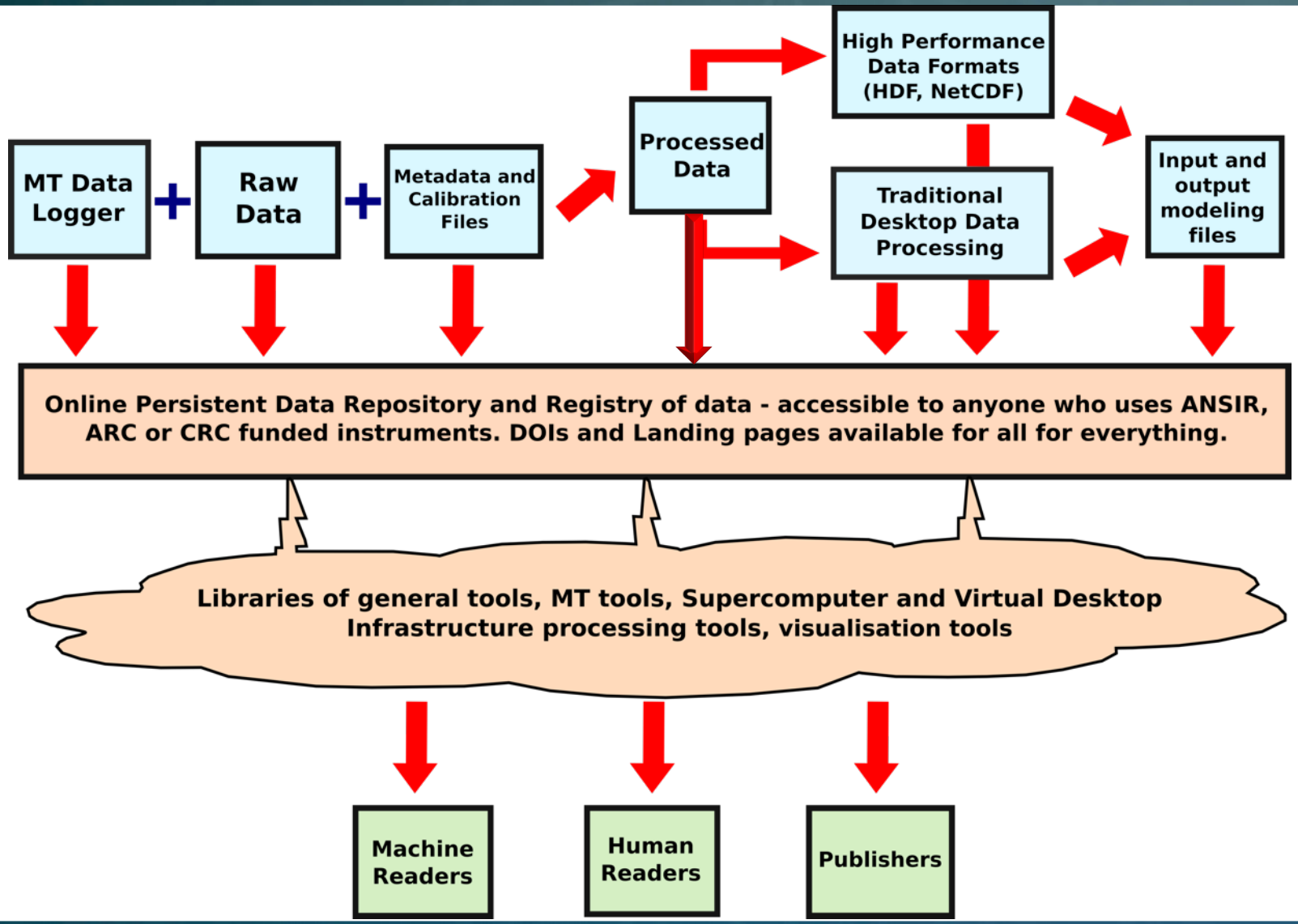
MT Data Logger + Primary Data + Metadata and Calibration Files → L1, L2 Processed Data → Traditional Desk-top Data Processing → Input and output modeling files

Publisher

Human Readers

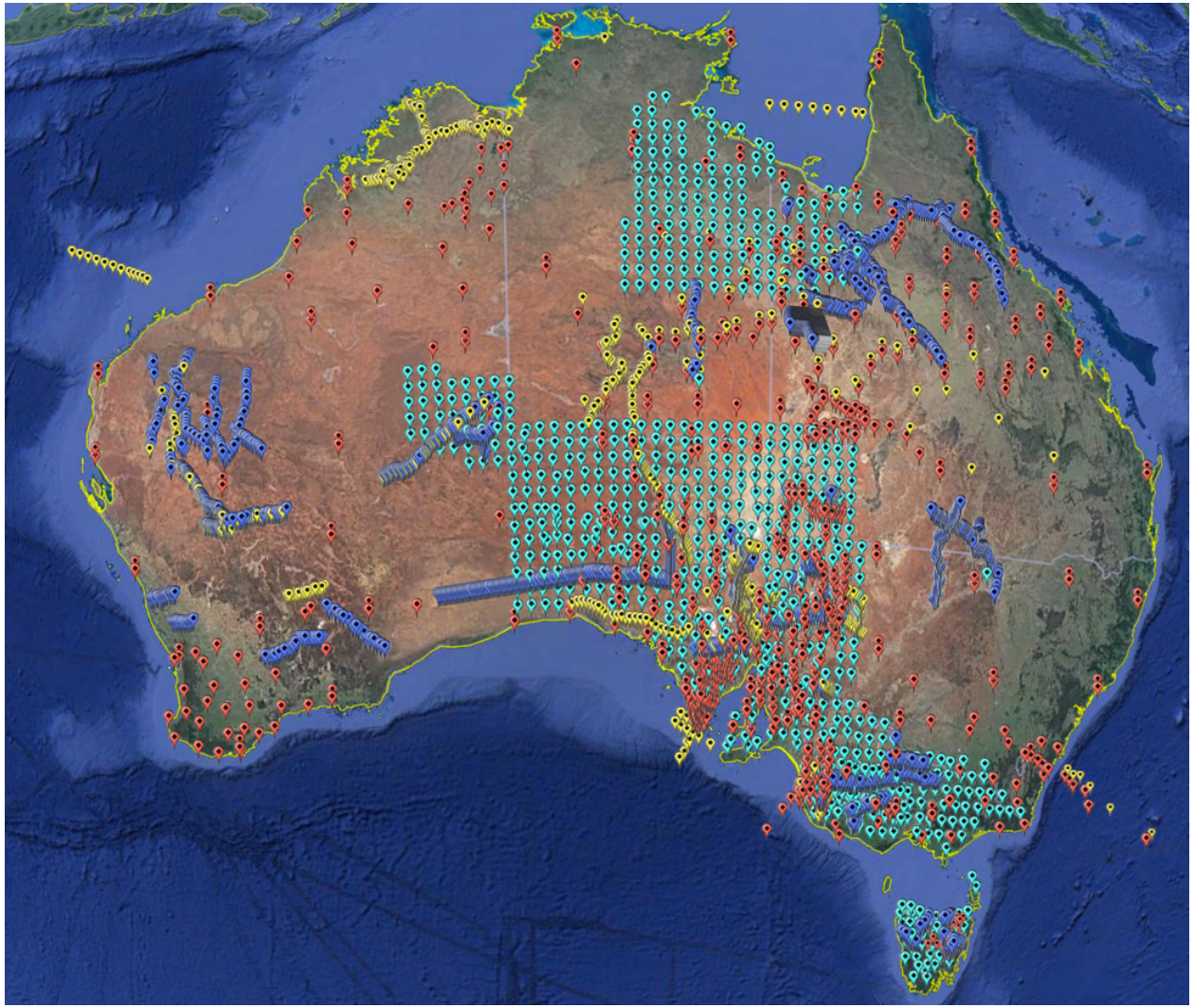
Source: Nigel Rees, NCI

We need access to all data along the Full-path of data: allowing for multiple options for data processing and algorithm design



Source: Rees et al. 2018. Making The University of Adelaide Magnetotellurics data collection FAIR and onto the path towards reproducible research. eResearch Australasia 2018, Melbourne.

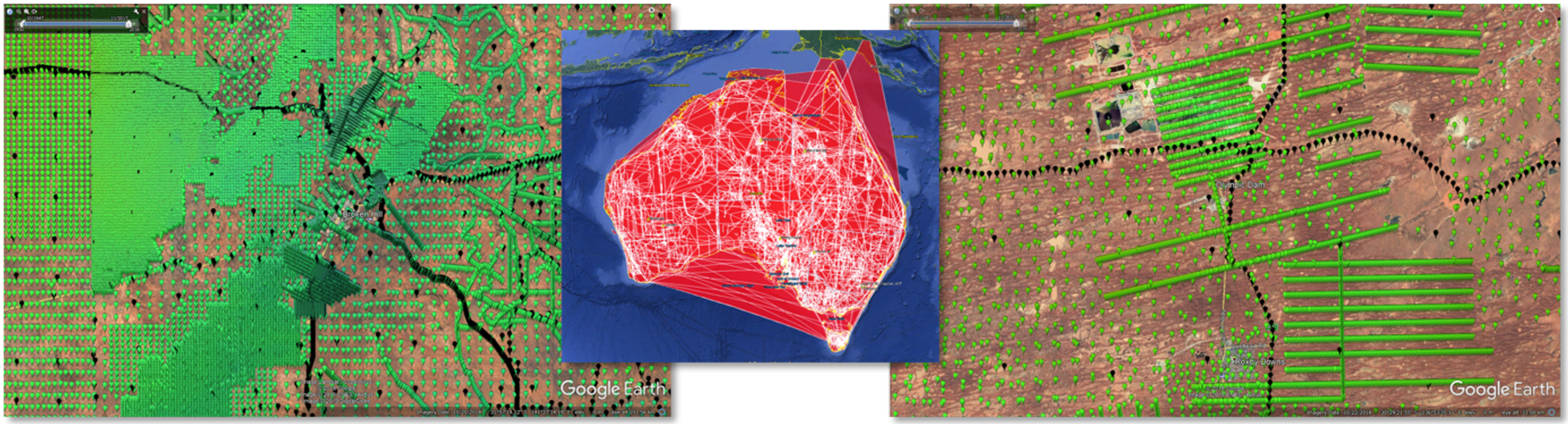
Can I access the rawer Australian MT data to create a transparent pipeline?



- Distribution of publicly funded MT surveys/sites in Australia since ~1990 (courtesy of Graham Heinson):
- Very little of the rawer time series data is available online, let alone accessible via catalogue and data services: many authors till request that you write to them for a copy of the time series which is sent via the post.

Geoscience Australia and NCI: making less processed forms of gravity data available online….

1.8M+ gravity points in 1631 datasets dating from 1947 sourced from National Gravity Database (NGDB). Some data cleansing was required (e.g., removing duplicate points), and survey metadata was not publicly accessible: but the data will soon be available online, for in situ access.

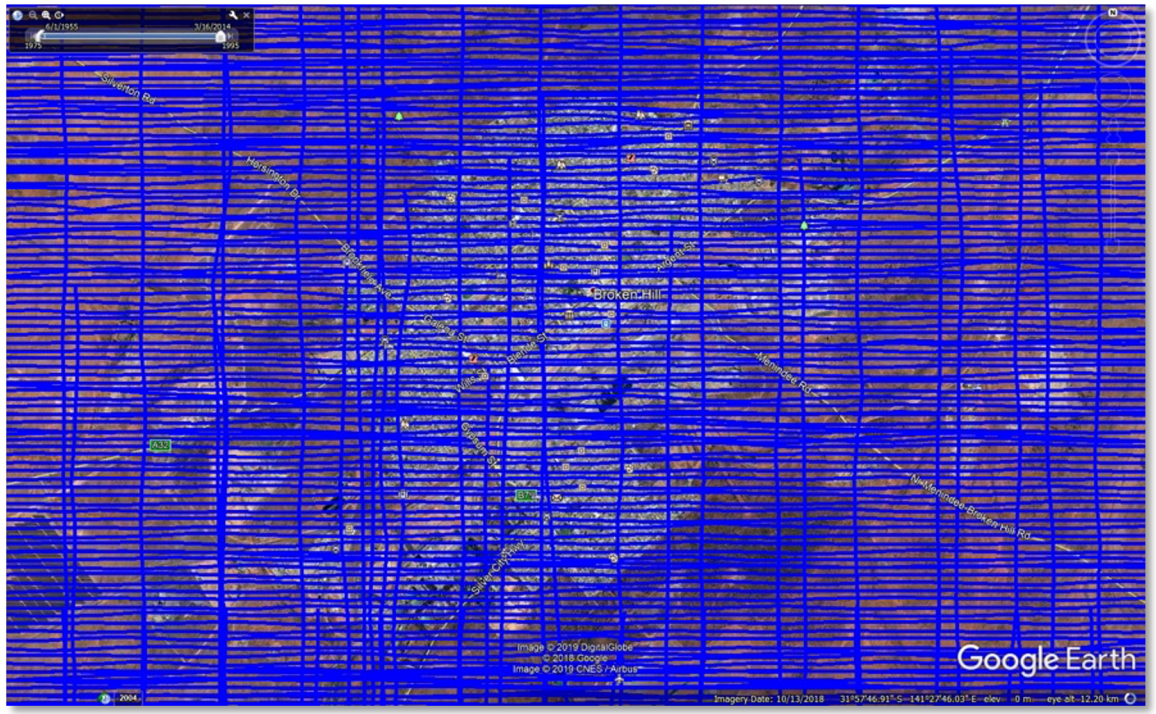


(Source: Alex Ip, GA)

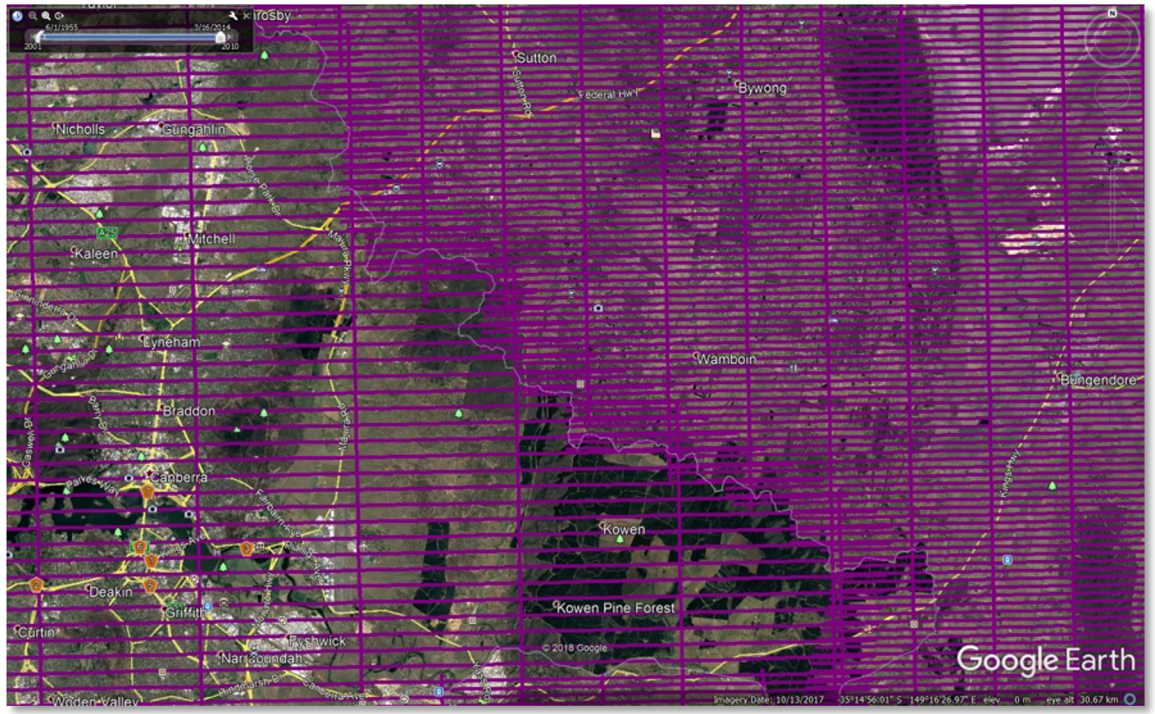
Geoscience Australia and NCI: Making Magnetic and Radiometric Point and Line Datasets available online

Magnetic line datasets: ~4B points in ~358k lines in 757 datasets.

Radiometric line datasets: ~473M points in ~335k lines in 630 datasets.



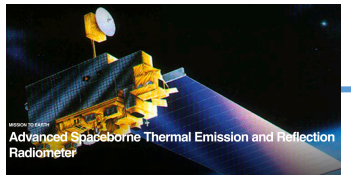
Broken Hill Magnetics
~100M Points in one magnetic survey alone!



Canberra Radiometrics (Source: Alex Ip)

Another complex Full-path of Data: the ASTER data example - it is hard to access the L1-L3 data

The Original Work



The Aster Mission

LEGEND

→ Derivative

Produced by the CSIRO, GSQ, GSSA, GSWA, GA, Japan Space Systems (JSS), NASA, NTGS Australian Initiative.

The Expressions

17 Individual map products derived from Level 0-4 data products generated from the source ASTER Mission (The Work)

Level 0 - JSS

Level 1A - JSS

Level 1B – JSS

Level 2 – JSS

Level 3 – CSIRO et al Australian Initiative

Level 4 – CSIRO et al Australian Initiative

14 VNIR-SWIR bands (maps)

1. False colour composite
2. Landsat TM Regolith Ratios
3. Green vegetation content
4. Ferric oxide content
5. Ferric oxide composition
6. Ferrous iron index
7. Opaque index
8. AlOH group content
9. AlOH group composition
10. Kaolin group index
11. FeOH group content
12. MgOH group content
13. MgOH group composition
14. Ferrous iron content in MgOH/carbonate

3 TIR bands (maps)

1. Gypsum Index
2. Quartz Index
3. Silica Index

The Manifestations

Each of the 17 'map' expressions is manifested in three file formats

BSQ files

GeoTIFF files (8-bit integer)

netCDF (32-bit float)

The Items

Each distribution point has multiple items from any manifestation of each map 'expression'

Geoscience Australia
BSQ files available for copying to a hard drive on request as 1:1,000,000 tiles

State-level BSQ files (SA)

State-level BSQ files (NT)

State-level BSQ files (QLD)

CSIRO Data Portal
629 items for download:
each 1:1,000,000 tile can have 17 maps

Digital Earth Australia
629 items for download or use on AWS:
each 1:1,000,000 tile can have 17 maps

State-level GeoTIFF files (SA)

State-level GeoTIFF files (NT)

State-level GeoTIFF files (QLD)

NCI Australia
17 national scale netCDF items for in situ, web service access and downloads

Lesley Wyborn (lesley.wyborn@anu.edu.au)

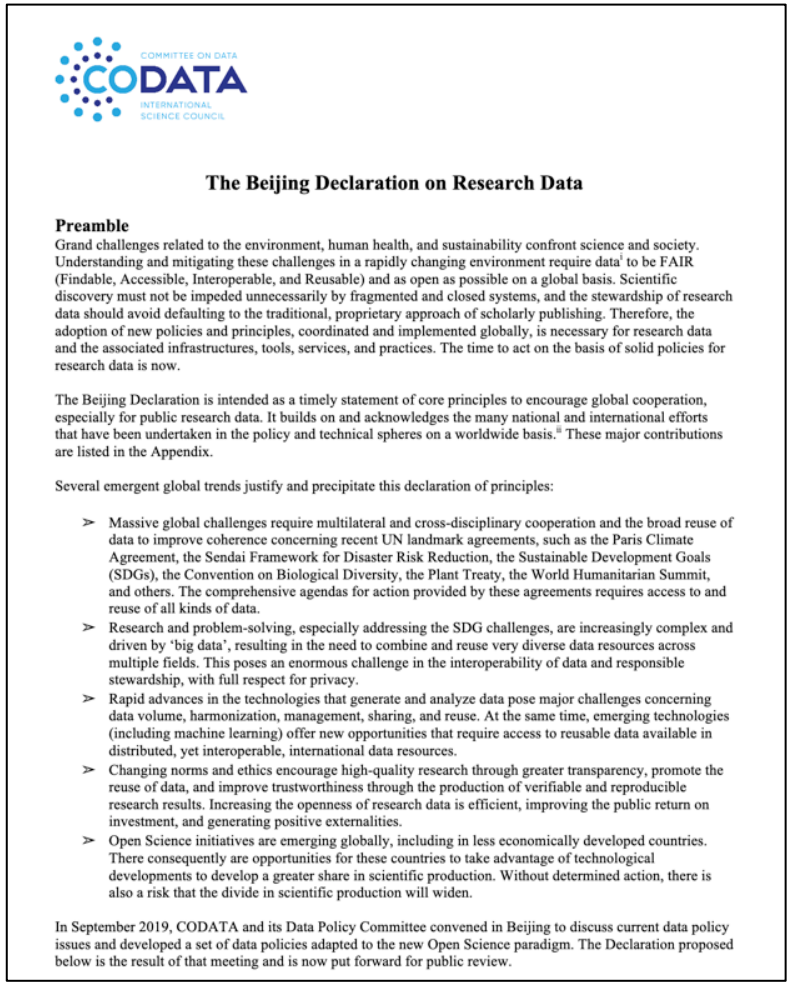
The elephant in the room

Source: https://www.inc.com/lee-colan/identifying-the-elephant-in-the-room.html

- Access to the rawer forms of data so that as new computational methods and tools become available, and greater computational power is accessible, we can reprocess to new algorithms, at larger scales, higher resolutions and faster time frames.
- If only the L3/L4 data products are preserved, we cannot do this.
- This applies not only to publications, but also to the growing plethora of online GIS systems, particularly those where the data is only available as WMS/ GeoTIFFS and has been optimized/ compressed to increase speed of uploading/downloading.
- (The rawer and minimally processed data sets also need to be made FAIR.)

lesley.wyborn@anu.edu.au

The CODATA 2019 Beijing Declaration on Research Data: addresses our elephant in our room



COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

The Beijing Declaration on Research Data

Preamble

Grand challenges related to the environment, human health, and sustainability confront science and society. Understanding and mitigating these challenges in a rapidly changing environment require data[i] to be FAIR (Findable, Accessible, Interoperable, and Reusable) and as open as possible on a global basis. Scientific discovery must not be impeded unnecessarily by fragmented and closed systems, and the stewardship of research data should avoid defaulting to the traditional, proprietary approach of scholarly publishing. Therefore, the adoption of new policies and principles, coordinated and implemented globally, is necessary for research data and the associated infrastructures, tools, services, and practices. The time to act on the basis of solid policies for research data is now.

The Beijing Declaration is intended as a timely statement of core principles to encourage global cooperation, especially for public research data. It builds on and acknowledges the many national and international efforts that have been undertaken in the policy and technical spheres on a worldwide basis.[ii] These major contributions are listed in the Appendix.

Several emergent global trends justify and precipitate this declaration of principles:

- Massive global challenges require multilateral and cross-disciplinary cooperation and the broad reuse of data to improve coherence concerning recent UN landmark agreements, such as the Paris Climate Agreement, the Sendai Framework for Disaster Risk Reduction, the Sustainable Development Goals (SDGs), the Convention on Biological Diversity, the Plant Treaty, the World Humanitarian Summit, and others. The comprehensive agendas for action provided by these agreements requires access to and reuse of all kinds of data.
- Research and problem-solving, especially addressing the SDG challenges, are increasingly complex and driven by 'big data', resulting in the need to combine and reuse very diverse data resources across multiple fields. This poses an enormous challenge in the interoperability of data and responsible stewardship, with full respect for privacy.
- Rapid advances in the technologies that generate and analyze data pose major challenges concerning data volume, harmonization, management, sharing, and reuse. At the same time, emerging technologies (including machine learning) offer new opportunities that require access to reusable data available in distributed, yet interoperable, international data resources.
- Changing norms and ethics encourage high-quality research through greater transparency, promote the reuse of data, and improve trustworthiness through the production of verifiable and reproducible research results. Increasing the openness of research data is efficient, improving the public return on investment, and generating positive externalities.
- Open Science initiatives are emerging globally, including in less economically developed countries. There consequently are opportunities for these countries to take advantage of technological developments to develop a greater share in scientific production. Without determined action, there is also a risk that the divide in scientific production will widen.

In September 2019, CODATA and its Data Policy Committee convened in Beijing to discuss current data policy issues and developed a set of data policies adapted to the new Open Science paradigm. The Declaration proposed below is the result of that meeting and is now put forward for public review.

http://www.codata.org/uploads/Beijing%20Declaration-19-11-07-FINAL.pdf

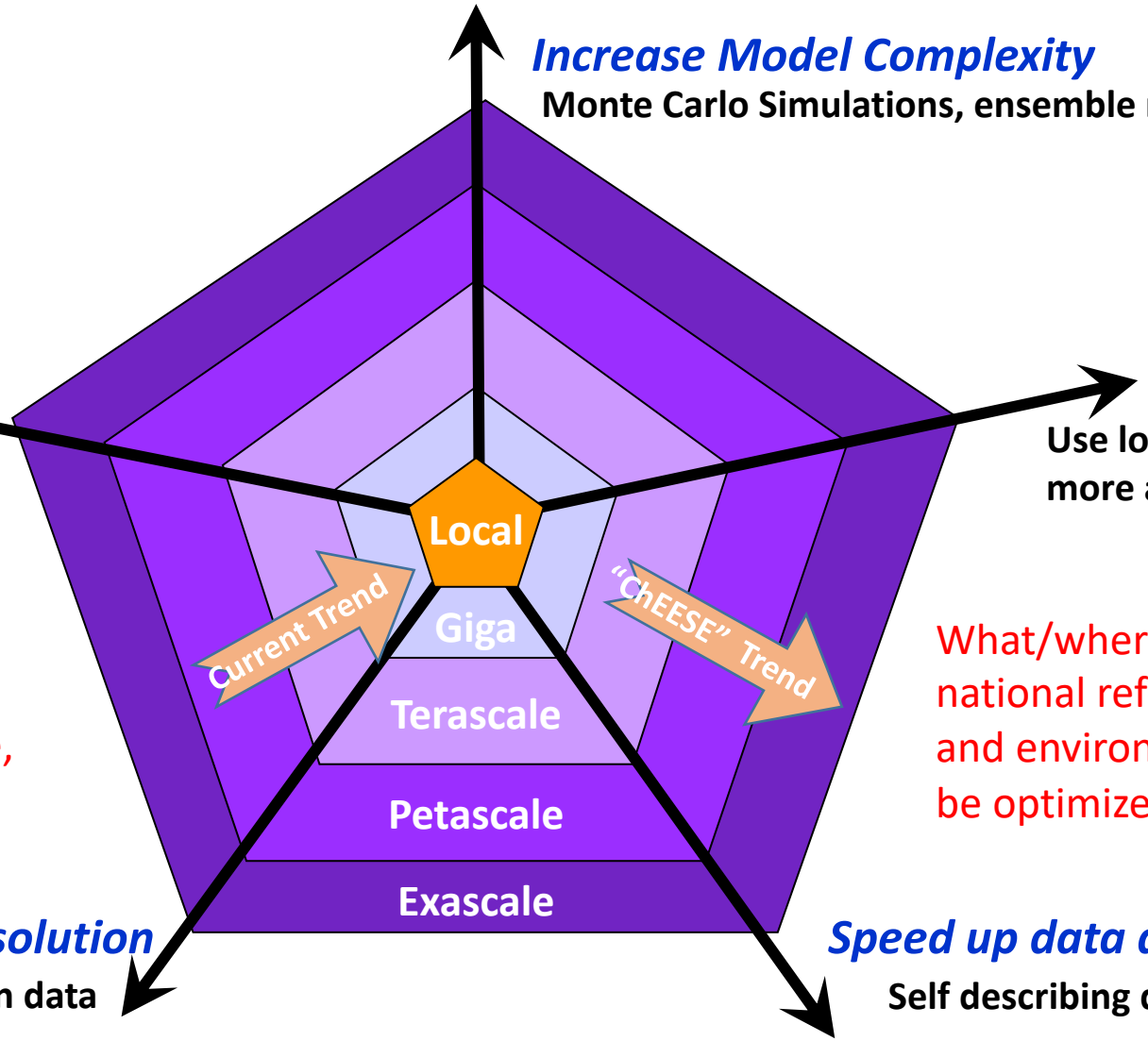
- The Beijing Declaration on Research Data is a timely statement of 10 core principles that encourage global cooperation, especially for publicly funded research data gathered either by the government or academic sectors.
- The Term 'Data' is '*used very broadly, to comprise data (stricto sensu) and the ecosystem of digital things that relate to data, including metadata, software and algorithms, as well as physical samples and analogue artefacts (and the digital representations and metadata relating to these things).*'
- The Beijing Declaration supports international efforts to make research data as open as possible and only as closed as necessary.
- It seeks to make data and metadata Findable, Accessible, Interoperable, and Reusable (FAIR) on a global basis and, wherever possible, automatically processable by machines.
- Principle 5 states that '*Publicly funded research data should be interoperable, and preferably without further manipulation or conversion, to facilitate their broad reuse in scientific research'.*

lesley.wyborn@anu.edu.au

We still have todays requirements for cloud and online: but for Exascale in2030 we will need the rawer and minimally processed forms

2019 Version

Based on European Climate Computing Environments, Bryan Lawrence (http://home.badc.rl.ac.uk/lawrence/blog/2010/08/02)



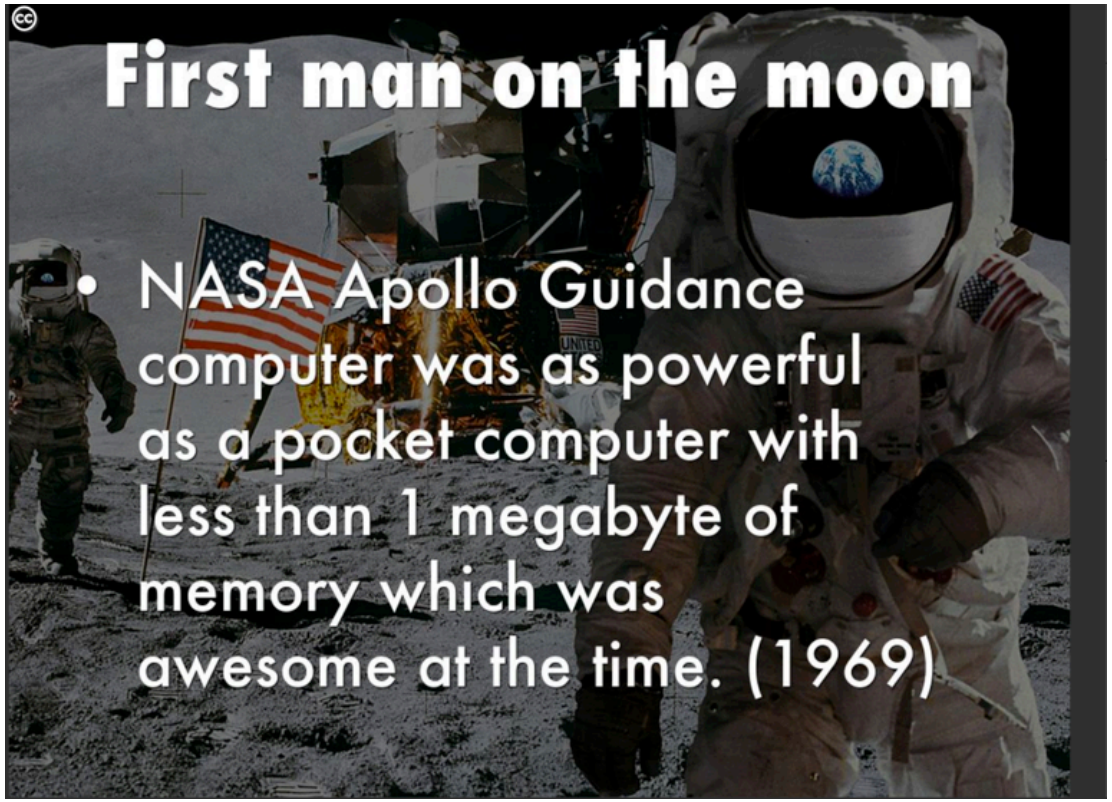
Todays trend is towards highly processed, subsampled data optimized for performance in online, cloud-based GIS systems

What/where are our high-resolution national reference data sets in Earth and environmental sciences that should be optimized for exascale usage?

Conclusion: will we get to Exascale for data by 2030?

To achieve the ambition of Earth and environmental science datasets being reusable and interoperable at Exascale by 2030, as well as being able to make a major contribution to the UN SDGs then today we need:

1. More effort and coordination in the development of international community standards to enable technical, semantic and legal interoperability of datasets at the scales predicted for 2030
2. To ensure that publicly funded research data are also available without further manipulation or conversion to facilitate their broader reuse in scientific research particularly as by 2030 as we will also have greater computational capacity to analyse data at scales and resolutions currently not achievable.



Source: https://www.haikudeck.com/evolution-of-software-education-presentation-lNfjrvG9MD#slide1

Questions?

Please contact Lesley Wyborn

lesley.wyborn@anu.edu.au