NCI

NATIONAL COMPUTATIONAL INFRASTRUCTURE

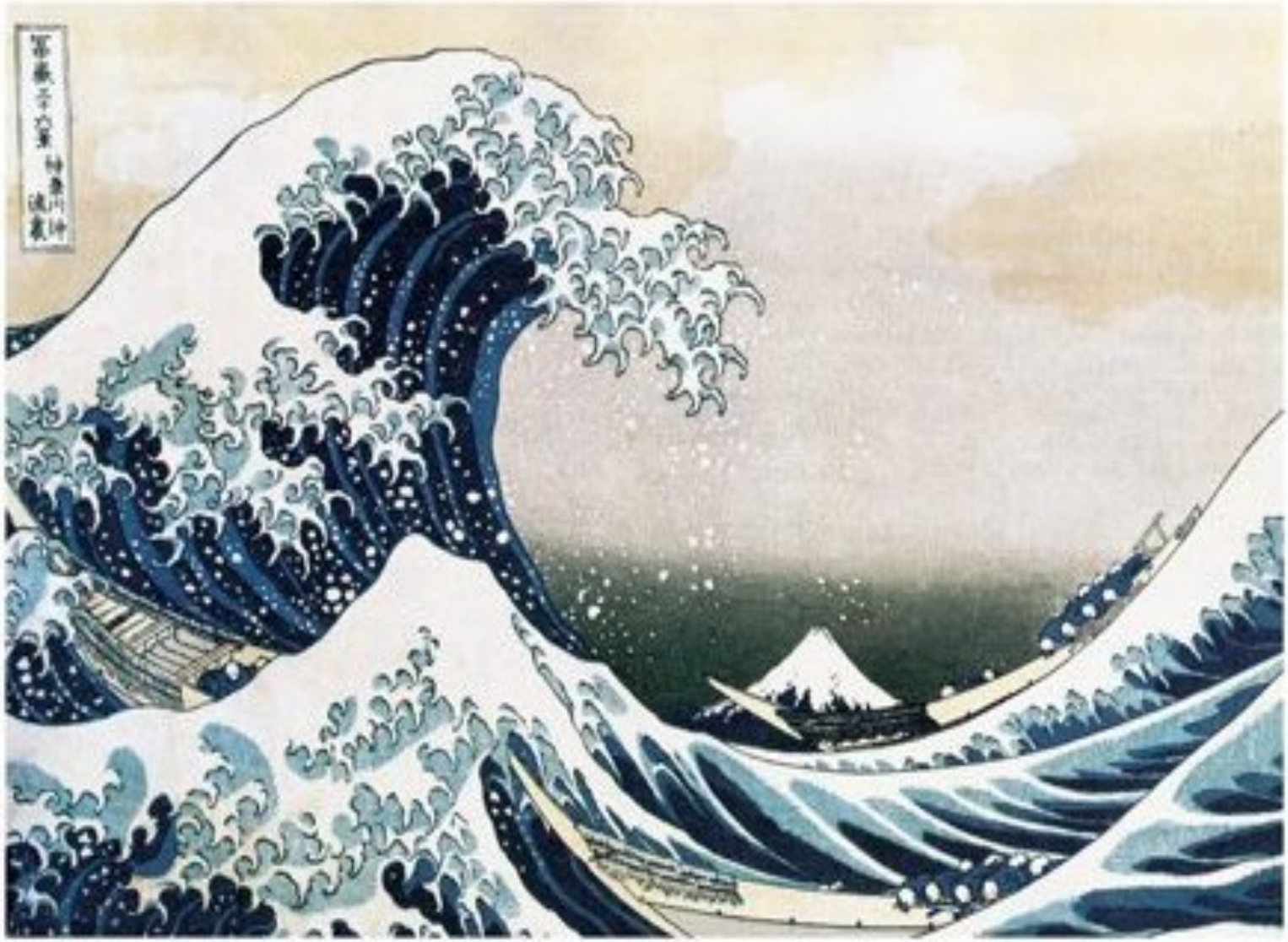
Perspectives on Big Data in the Geosciences from a major Australian national data center

Lesley Wyborn

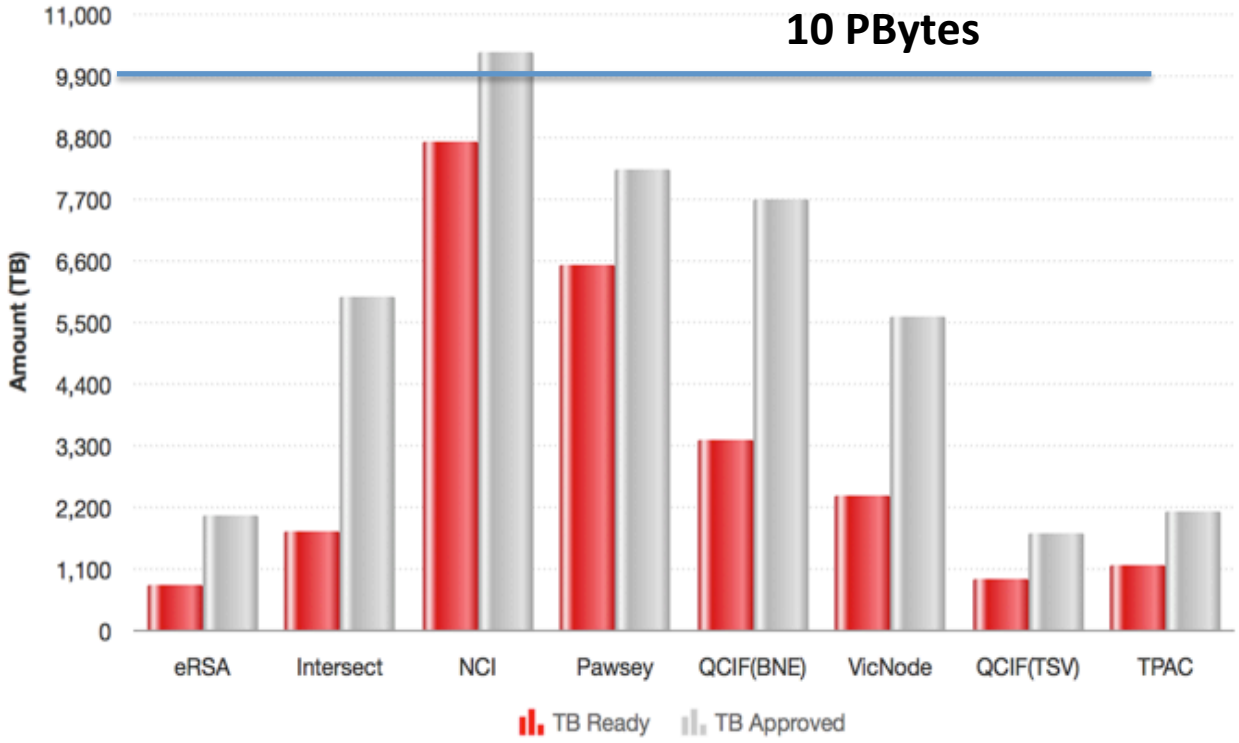
Australian Government Department of Education
Australian National University
Australian Government Bureau of Meteorology
CSIRO
Australian Government Geoscience Australia
Australian Government Australian Research Council

nci.org.au
@NCInews

The Data Tsunami has not yet landed



The 43 PB on the Research Data Storage Infrastructure



eRSA	Intersect	NCI	Pawsey	QCIF(BNE)	VicNode	QCIF(TSV)	TPAC
785	1743	8709	6504	3406	2404	907	1160
2049	5957	10296	8202	7699	5575	1729	2105

Primary Nodes | Additional Nodes

Total: TB Ready	25,817	Total: TB Approved	43,611

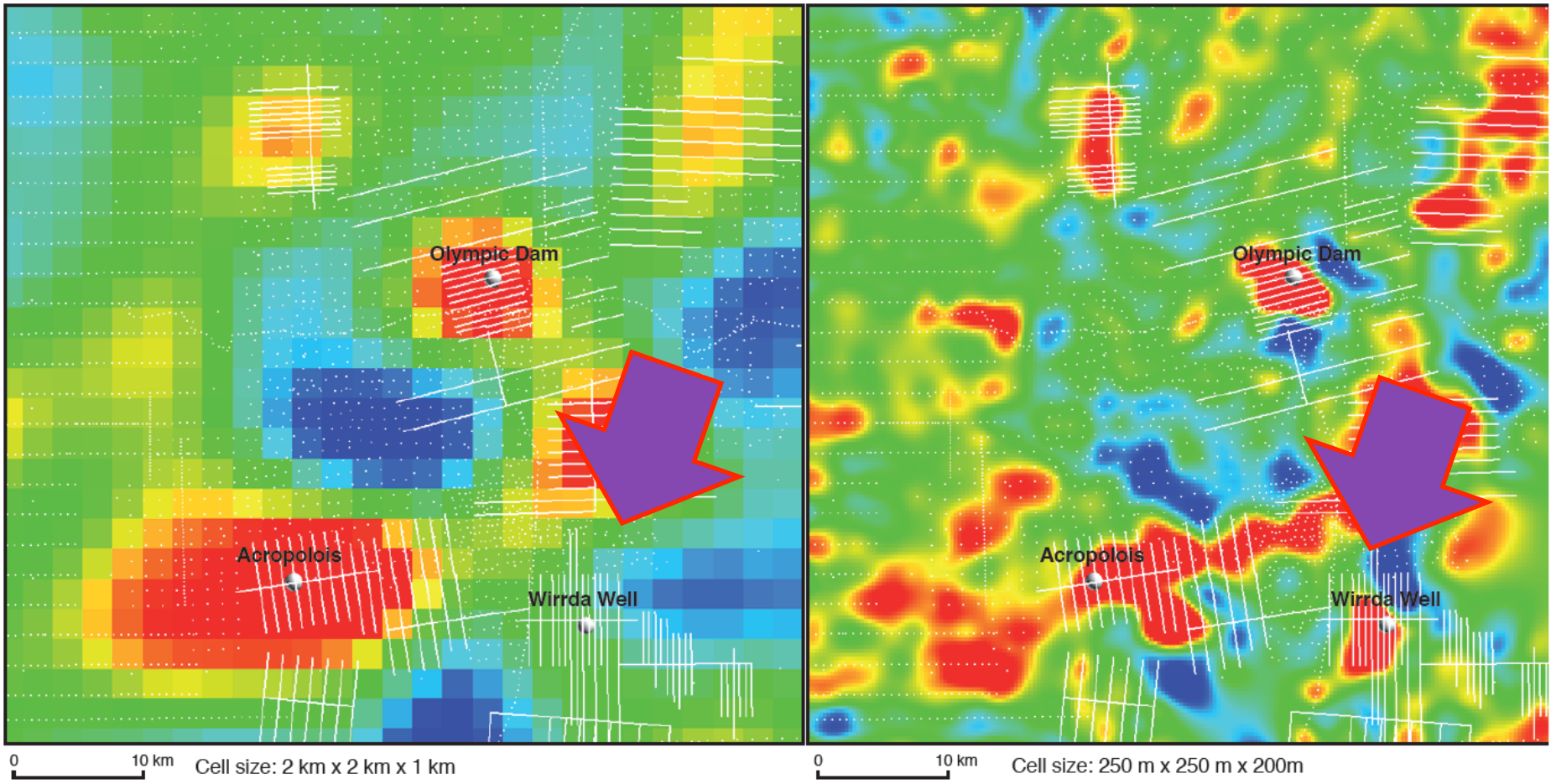
Source: https://www.rds.edu.au/

- Our ~43 Petabytes has saved data on 8 nodes that for the most part, already existed
- We have not considered new initiatives such as the SKA, Copernicus, Himawarri etc.
- We are not using our current data to its fullest resolution

ISO/IEC 10746 RM-ODP of NCI Collections with EO Data

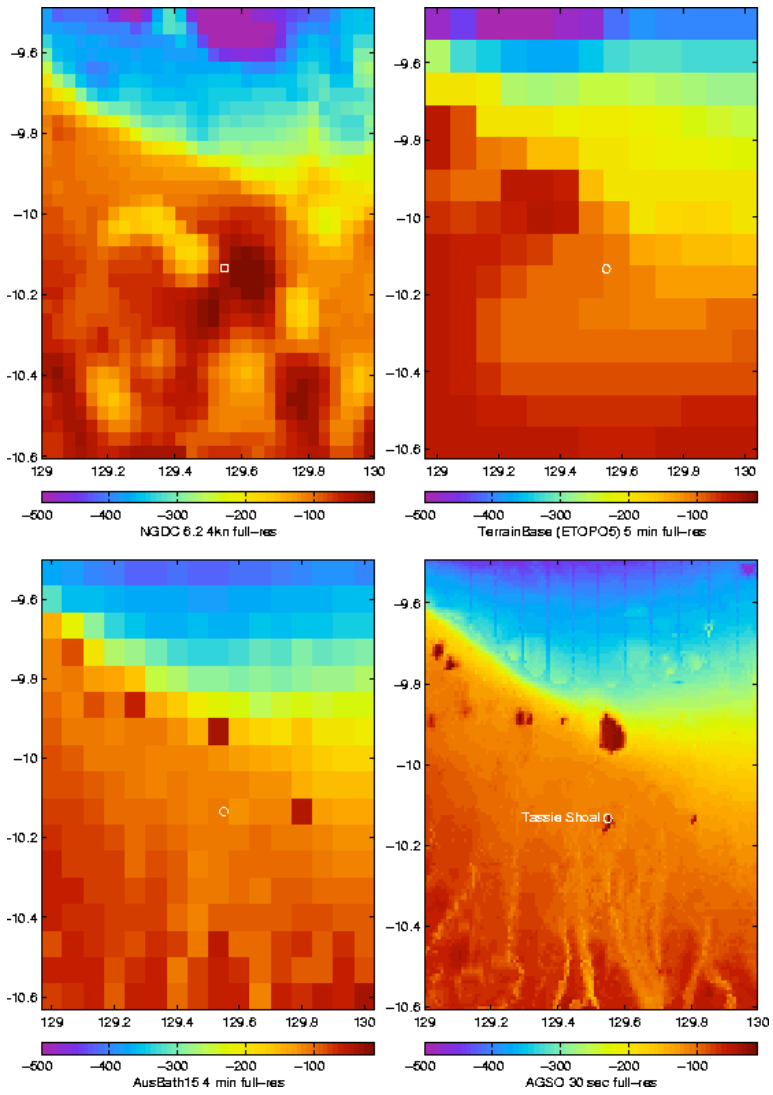
No.	Collection	Enterprise	Technology	Information	Allocation (TB)	Ingested (TB)	Capacity
u39	MODIS Ocean Colour	IMOS	THREDDS	NetCDF-CF	428	324	76.0%
	AusCover	TERN	THREDDS	NetCDF-CF			
	MODIS L1B	IMOS/TERN	THREDDS	NetCDF-CF			
	AVHRR, VIIRS	CSIRO	THREDDS	NetCDF-CF			
fj4	Soil Moisture	CSIRO	THREDDS	NetCDF-CF	5	3	54.6%
fj2	eMast data assimilation	TERN	THREDDS	NetCDF-CF	110	13	11.8%
rr9	eMast	TERN	THREDDS	NetCDF-CF	90	28	31.0%
ua6	CMIP5	BoM	THREDDS	NetCDF-CF	2400	1488	62.0%
rr5	Observation (Himawarri)	BoM	THREDDS	NetCDF-CF	360	240	65.5%
rs0	Landsat	GA	GeoServer	GeoTIFF	1478	1413	95.6%
fk4	WOFS	GA	GeoServer	GeoTIFF	22	16	72.8%

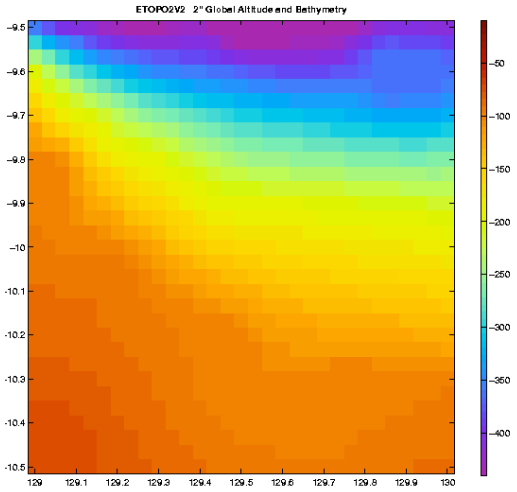
The bonsai effect: degrading our data

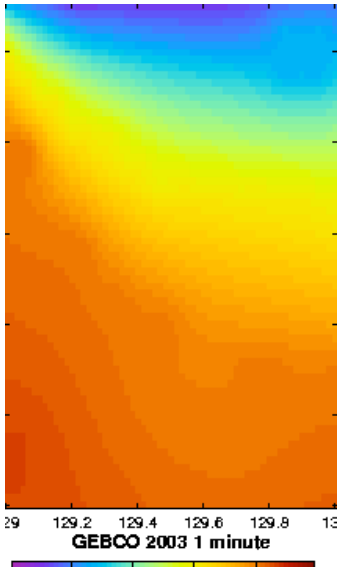


Regional Gravity Inversion Olympic Dam area. (courtesy N.Williams Geoscience Australia)

Marine example of bonsai'ed data

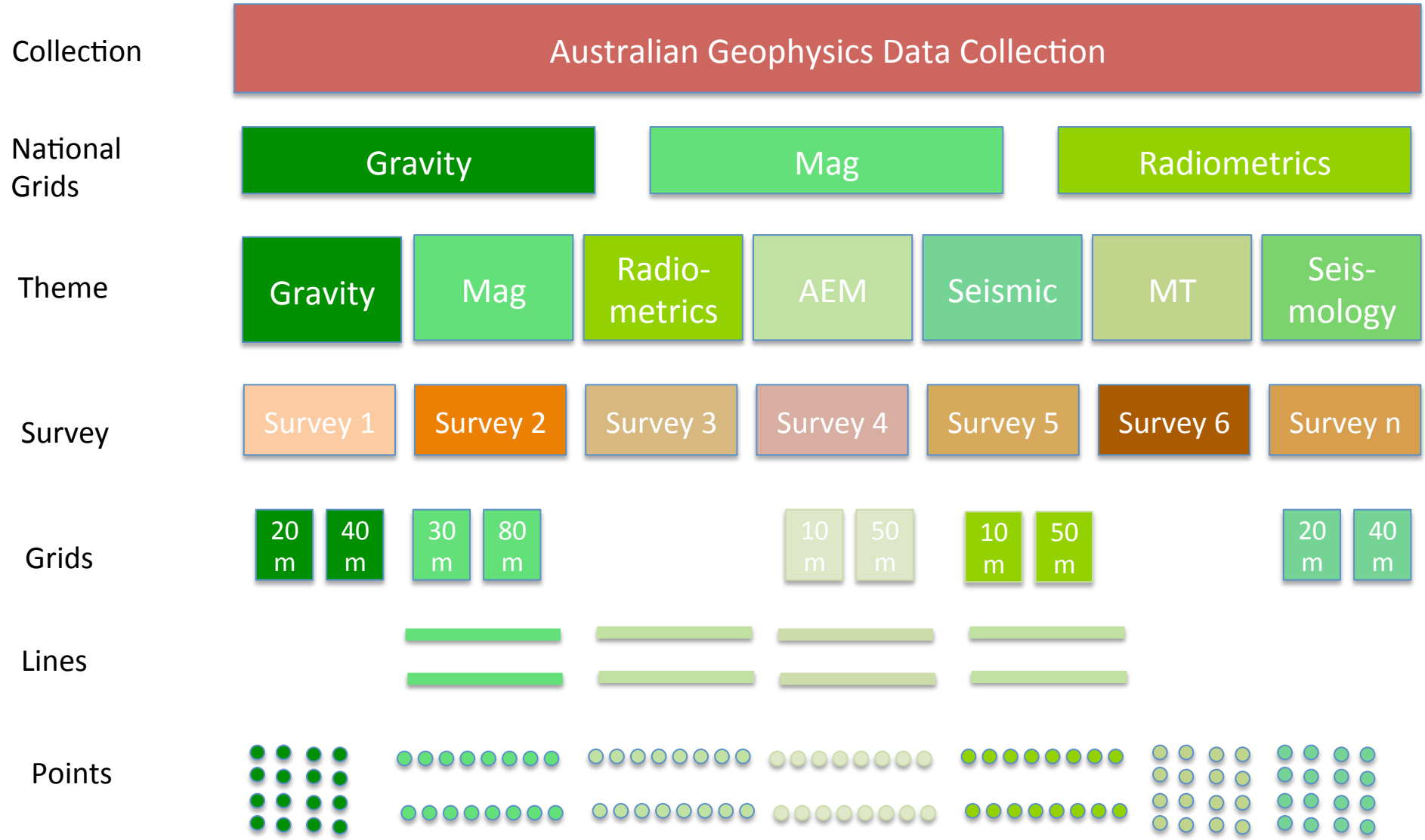




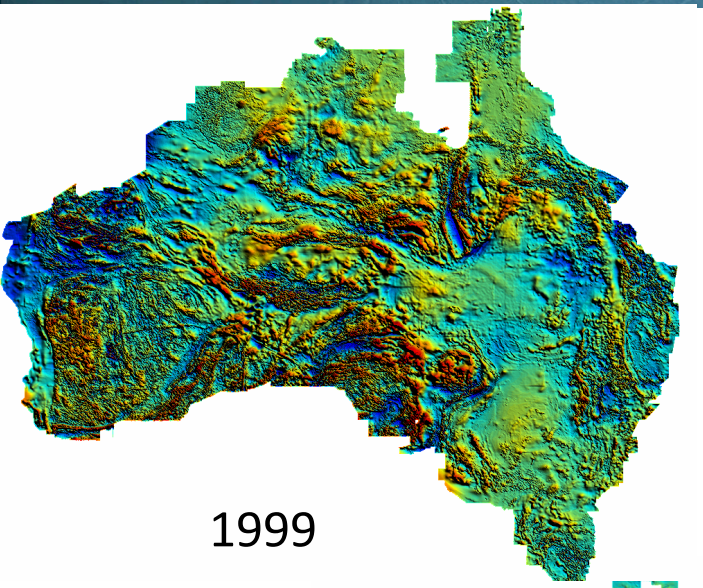


Source:
http://www.marine.csiro.au/eez_data/doc/comp_bath.gif

Grids of Grids vs Grids of Actual Observations

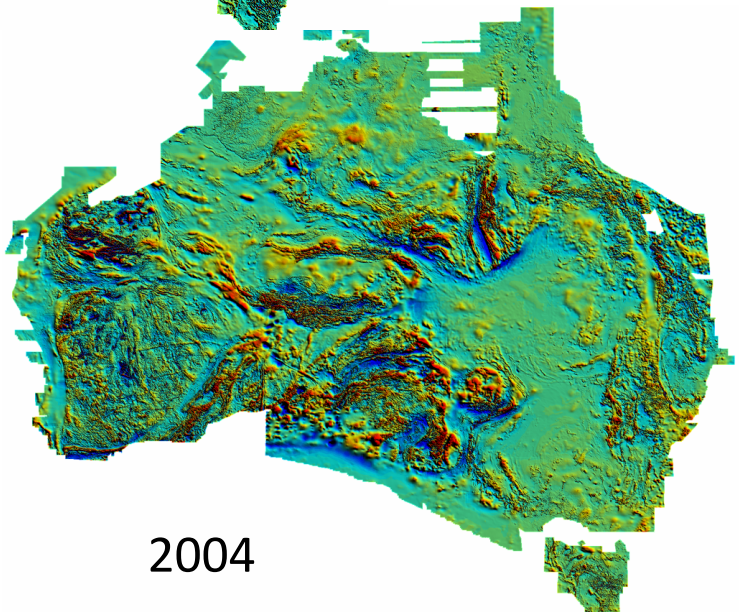


Resolution impacts on file size: e.g. Magnetics

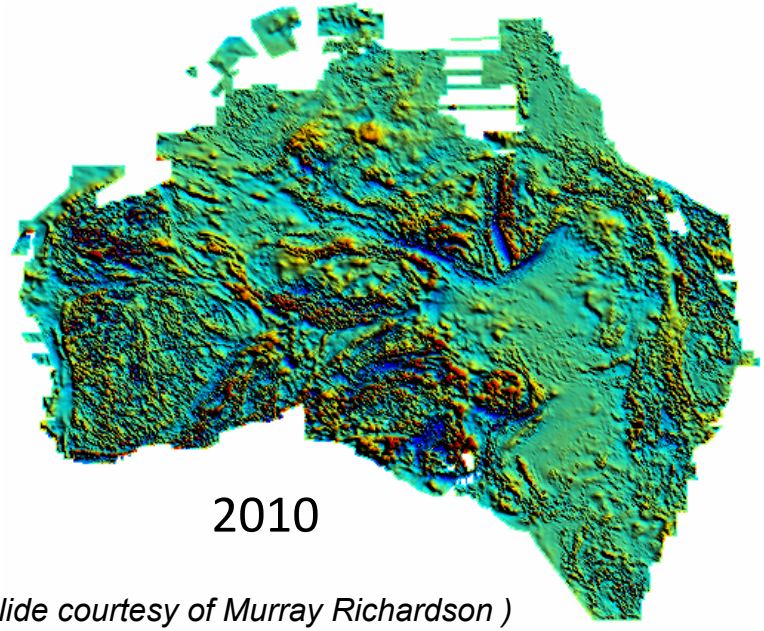


1999

Version	Year	Grid cell size	Data file size
3	1999	400m	0.49 GB
4	2004	250m	0.94 GB
5	2010	80m	9.73 GB
6	2013 (?)	<80m	3 TB



2004



2010

(Slide courtesy of Murray Richardson)

The Data Aggregation Problem in Climate Science Research

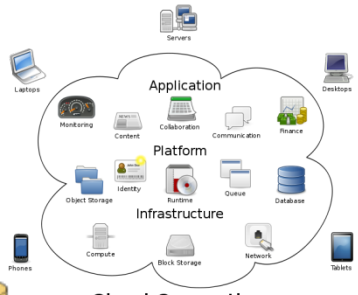
6th assessment 2020

5th assessment 2013

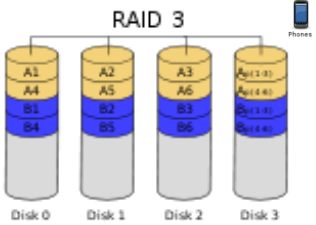
4th assessment 2007

3rd assessment 2001

ARC CENTRE OF EXCELLENCE FOR CLIMATE SYSTEM SCIENCE



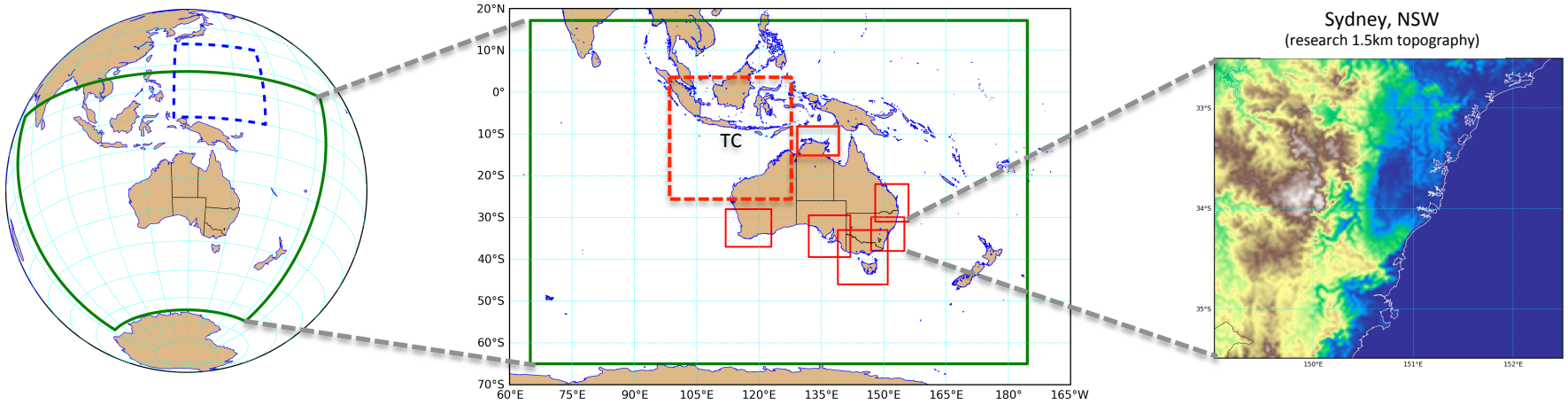
Cloud Computing



Slide Courtesy of Andy Pitman
COE Climate System Science

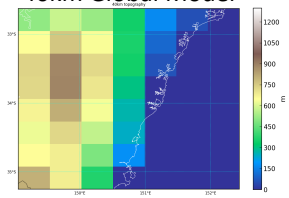
50 gigabytes → **35 terabytes** → **30 petabytes** → **?? exabytes**

Numerical Weather Prediction Roadmap



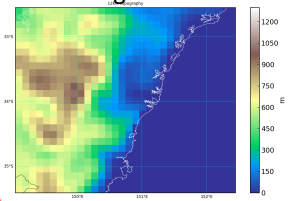
2013

2 x daily 10-day & 3-day forecast
40km Global Model



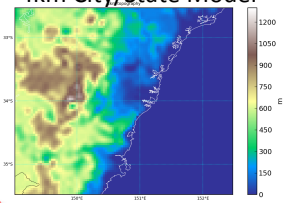
Increasing model resolution for improved local information

4 x daily 3-day forecast
12km Regional Model



Future model ensembles for likelihood of significant weather

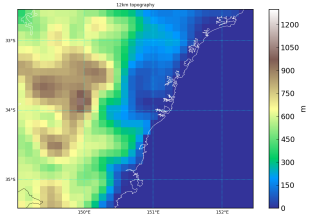
4 x daily 36-hour forecast
4km City/State Model



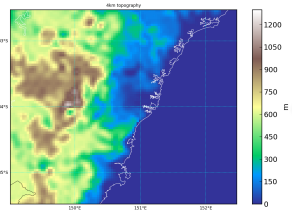
Model Topography of Sydney, NSW

2020

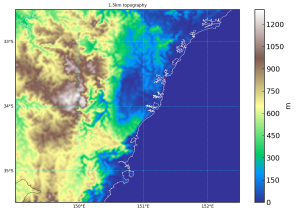
2 x daily 10-day & 3-day forecast
12km Global Model



8 x daily 3-day forecast
5km Regional Model

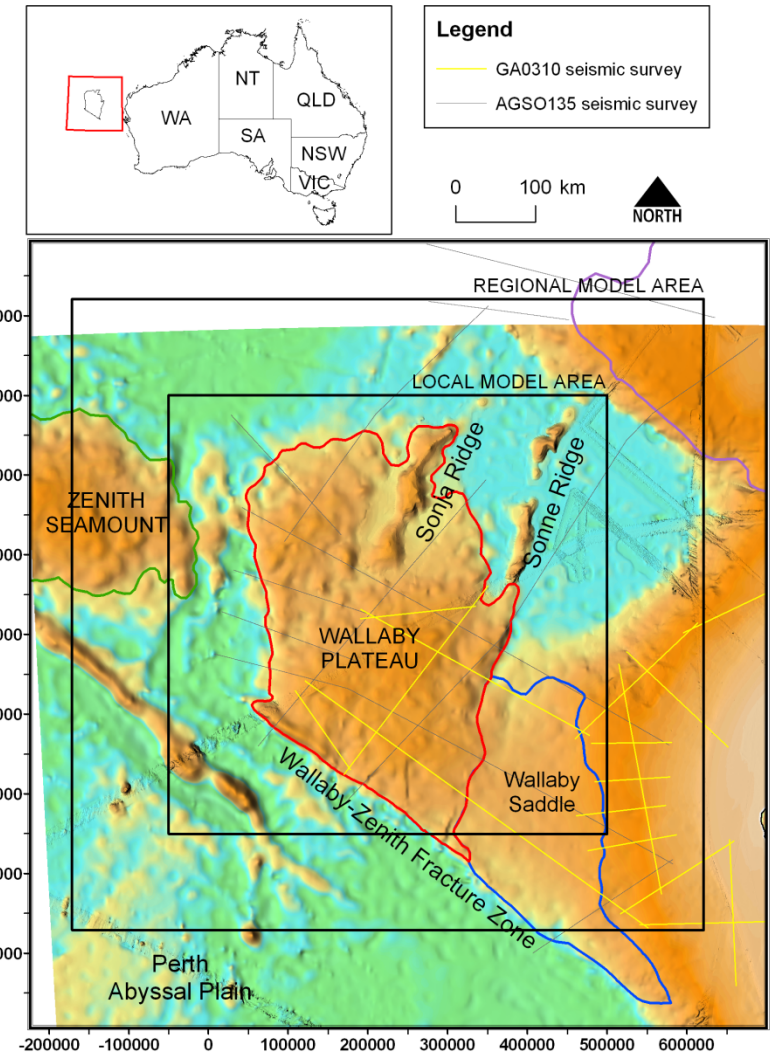


24 x daily 18h or 36h forecast
1.0km City/State Model



C/- Tim Pugh, BoM

Mentally Moving off the Desktop for Larger Scale Inversions

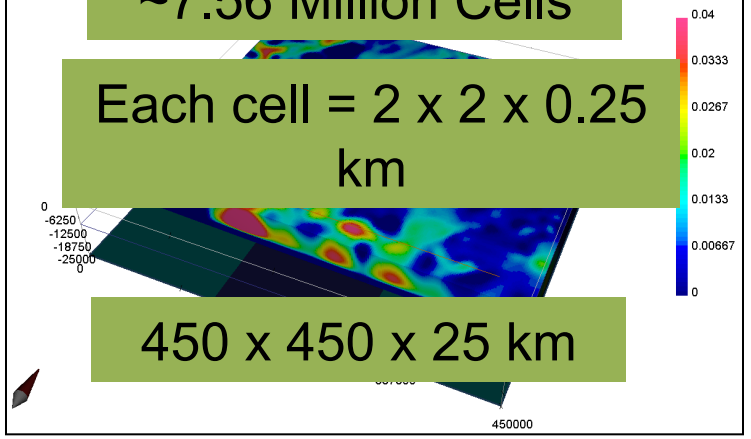


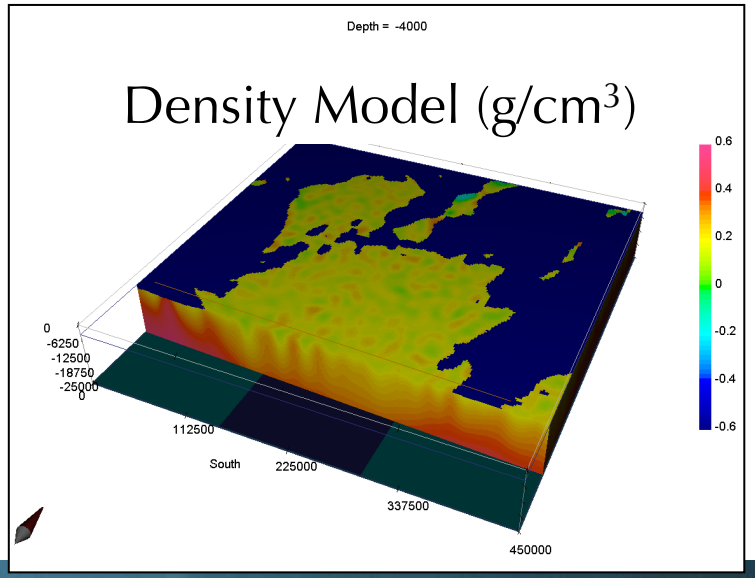
Susceptibility Model (SI)

~7.56 Million Cells

Each cell = 2 x 2 x 0.25 km

450 x 450 x 25 km





Life on Big Interdisciplinary Data Platforms

- We are moving from Data My-ning on Desktops
  - *"Give me the file, the whole file, and nothing but the file and let me process it locally on MY DESKTOP KIT"*
- To new, more complex Data Mining on centralised data platforms
  - *"Please enable me in real time to discover, access and process only those parts* of *multiple files and/or databases that I need and let me do it online using YOUR KIT and let me drive it from my iPAD or my SmartPhone".*
- But we also need to leave our desktop habits behind including:
  - *Constantly reformatting for particular applications*
  - *Downsampling the data*
  - *ETC ETC.*
- *THE DATA TSUNAMI HAS NOT YET HIT THE BEACH – WE ARE NOT READY*