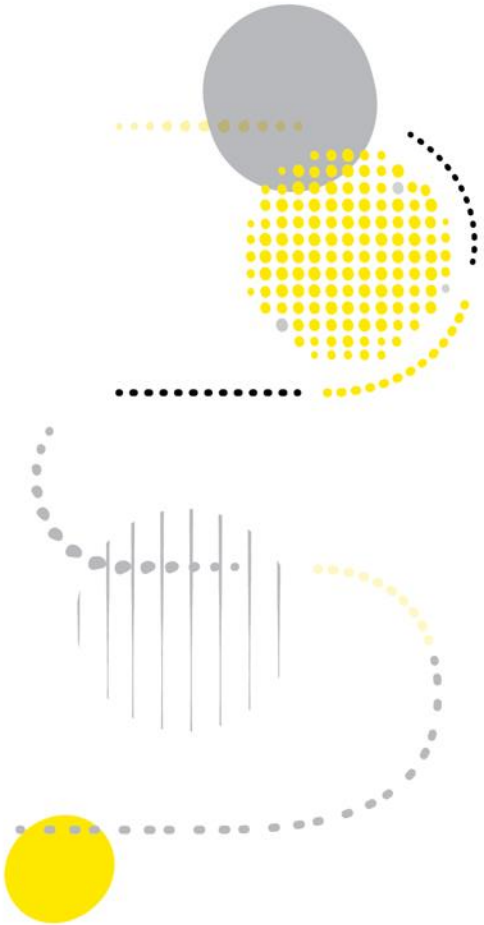


Outline

- What is Big anyway?
- Should you move your data?
- Sneakernet
- Considerations
- Tools
- Data Transfers Nodes (Raijin, Katana)
- Considerations
- Examples

Research Technology Services

High Performance Computing
Access to National Infrastructure and Cloud providers.

Training program
45+ Free Tech training courses per year (eg: R, Python, Git, Unix, HPC, Excel)

Data Management Support
Support for projects with highly sensitive or complex data storage needs

On-premise HPC system
Katana – shared system with buy-in options.

Community events
Weekly Hacky Hour meetups, monthly expert seminars and ResBaz conference

Data Tools and Data Moves
eNotebooks, Jupyter notebooks, REDCap, moving big or complex Data

Code & Algorithm support
Expert hours scheme - access to team and casuals.

Collaborations
Cross University and Vendor partnerships

Data Set Publications
Pilot Scheme – for publishing Open Data

RDM@ UNSW

A Single Point of Contact
rdm@unsw.edu.au

Classify all Research Data using **UNSW Data Classification Standards**

Use **UNSW Supported Data Platforms**

A **'Living' RDM Plan** for each UNSW Research Project

Completing Core **RDM online Training** Modules

FOR MORE INFORMATION, GO TO
HTTPS://RESEARCH.UNSW.EDU.AU/RESEARCH-DATA-MANAGEMENT-UNSW

Storing your Research Data

UNSW SYDNEY

	Key SUPPORTED				Currently UNSUPPORTED		
Storage Platforms	UNSW OneDrive & Teams	UNSW eNotebook	Data Archive	Home Drive Shared Drive	Cloudstor	Dropbox	Local Storage* (e.g., Internal/External Drive, USB, PC, etc)
Storage Type	*Day-to-Day*	*Day-to-Day*	*Long-Term*	*Day-to-Day*	*Day-to-Day*	*Day-to-Day*	*Day-to-Day*
Suitable Data Classification	Highly Sensitive / Sensitive / Private / Public	Sensitive / Private / Public	Sensitive / Private / Public	Private / Public	Private / Public	Private / Public	Public
Stored in Australia	✓	✓	✓	✓	✓	✗	?
Backup & Disaster Recovery	✓	✓	✓	✓	✗	✗	✗
Syncing with Local Copy	✓	Not Applicable	Not Applicable	Not Applicable	✓	✓	Not Applicable
External Collaborator Access	✓	✓	✗	✗	✓	✓	✗
Storage Limit	5 TB/User	Unlimited	Unlimited	Unlimited	1TB/User	$$$	?
Version Control	✓	✓	✓	✓	✓	✓	✗
Recovery from Deletion	60 Days	No Data Deletion	No Data Deletion	10 days	✗	$$$	✗
Post-Project Data Retention	>7 years	Indefinitely	Indefinitely	>7 years	?	$$$	?

** Local devices vary greatly in their configuration and security. Contact rdm@unsw.edu.au to find out which data classification is suitable for your local device (eg. desktop, laptop & tablets)*

? Unknown or Device Dependent

Highly Sensitive Data | Sensitive Data | Private Data | Public Data

$$$ Dependent on the plan you have paid for

For Sensitve and Highly Sensitive data, data encryption and/or other settings may be required. Please refer to the UNSW Data Handling Guidelines for more information.

If you have any Highly Sensitive data, or Research Data Management inquiries, please contact rdm@unsw.edu.au

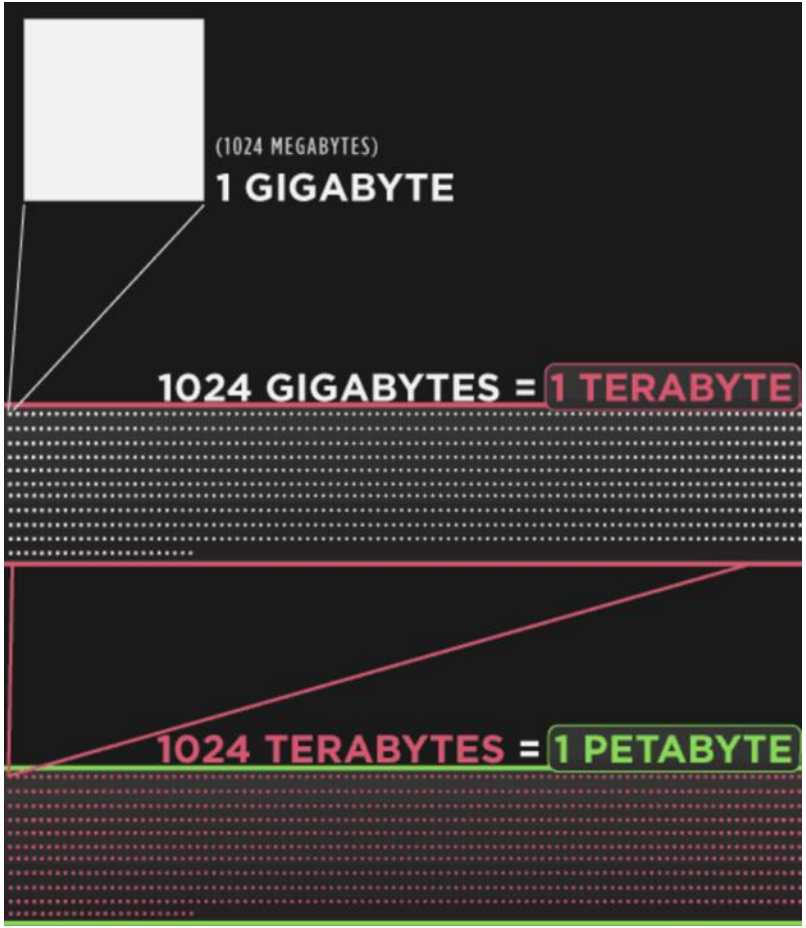
UNSW SYDNEY

What's Big anyway?

"Big" is an evolving term

Data size increasing faster than network speeds

Some disciplines use vastly more data than others (I'm looking at you Astronomy, Genomics, Climate…)



Could you not?

Big Data is hard to move!

Keeping data where you use it

- Faster access
- sharing with other users

Keep in mind: Retention, re-use, sharing

Selective transfer (eg. OneDrive Files on Demand)

Local Mirrors of data

All the elements

Source Disk speed

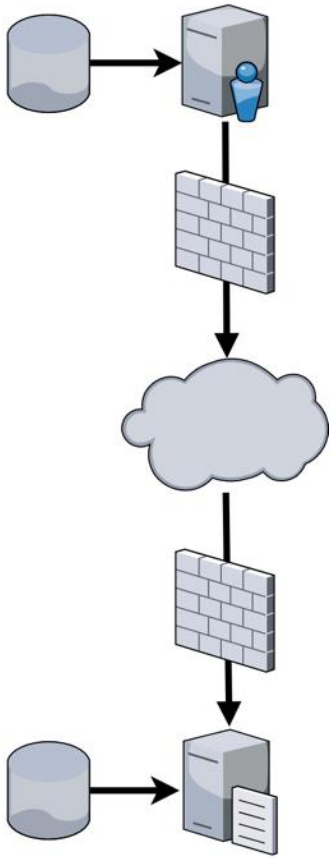
Source network speed

End-to-end network speed

Destination network speed

Destination Disk speed

Also: Application efficiency! MTUs! Firewalls! Network Congestion! PC Performance! Failures/restarts! Incremental vs overwrite! Encryption!



How long will it take?

	Physical Transfer		Physical / Online Transfer			Online Transfer
	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hours	18 minutes	2 minutes	11 seconds	1 second	0.1 seconds
10 GB	30 hours	3 hours	18 minutes	2 minutes	11 seconds	1 second
100 GB	12 days	30 hours	3 hours	18 minutes	2 minutes	11 seconds
1 TB	124 days	12 days	30 hours	3 hours	18 minutes	2 minutes
10 TB	3 years	124 days	12 days	30 hours	3 hours	18 minutes
100 TB	34 years	3 years	124 days	12 days	30 hours	3 hours
1 PB	340 years	34 years	3 years	124 days	12 days	30 hours
10 PB	3,404 years	340 years	34 years	3 years	124 days	12 days
100 PB	34,048 years	3,404 years	340 years	34 years	3 years	124 days

Sneakernet

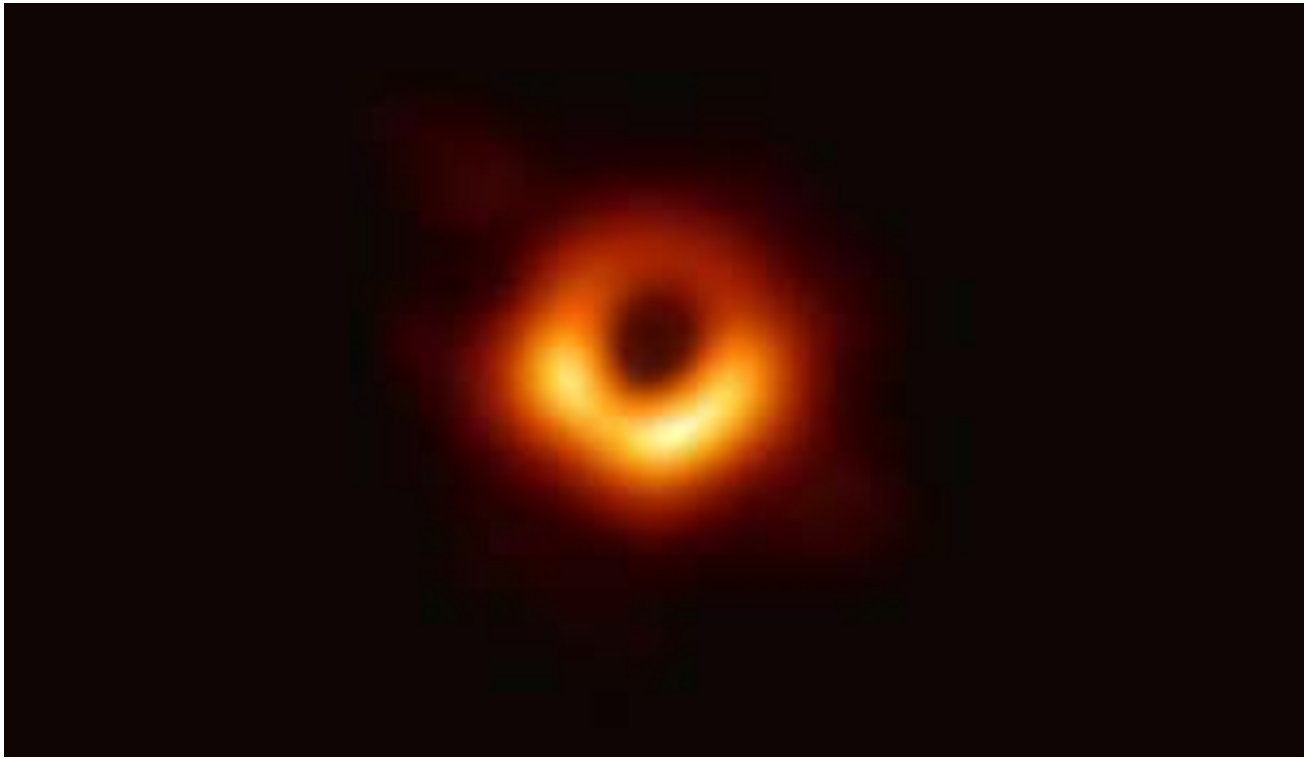
Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway.

-Tanenbaum, Andrew S (1989)

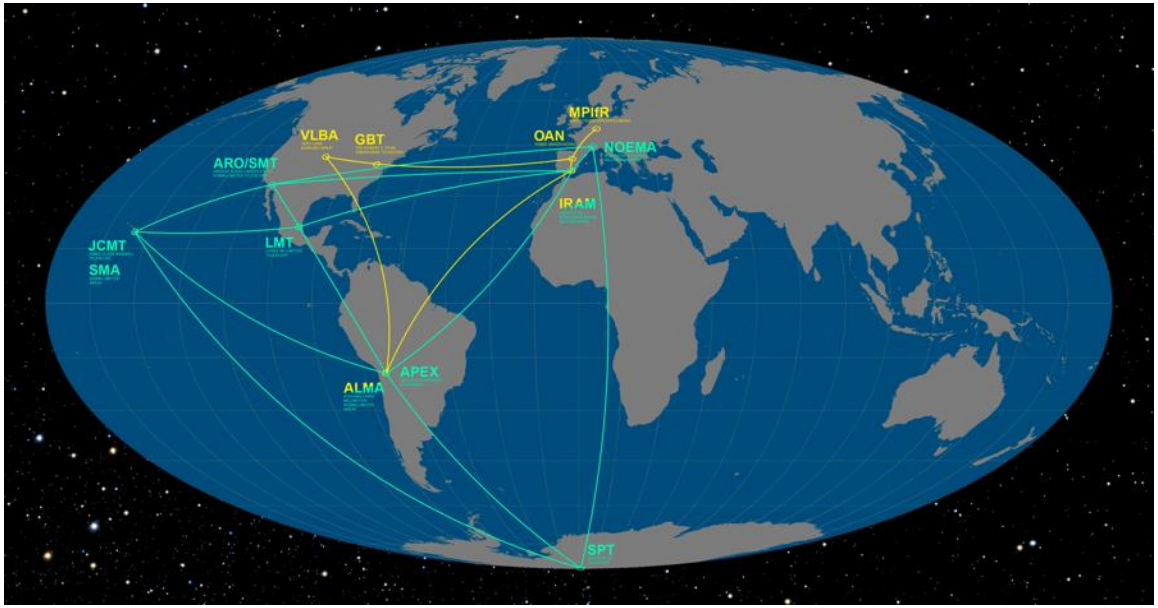
There's a lot of band-width in a station wagon

-Gruenberger, Fred (1971)

Sneakernet



Sneakernet



12 Telescopes

1 Petabyte of data per 5 day run

1000 Hard Disks per run

Going to an 800 CPU cluster in Massachusetts

Final output just a few Gigabytes

Tools for transferring data

Basic tools

SFTP

RSYNC

Buffer issues!

Multi-channel tools

Good speedup up to about 5 threads for single endpoints

LFTP - https://lftp.tech/

Parallel RSYNC wrappers - http://moo.nac.uci.edu/~hjm/parsync/ , https://github.com/jbd/msrsync

GridFTP - http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/

Commercial tools:

Globus (Commercial GridFTP interface) - https://www.globus.org/

Aspera - https://asperasoft.com/

Tools for transferring data

Command-line tools

UDT - UDP wrapper for RSYNC https://github.com/LabAdvComp/UDR

BBCP - SSH-based direct transfer tool http://www.slac.stanford.edu/~abh/bbcp/

FDT - Fast Data Transfer, Java-based multithreaded TCP transfers https://fast-data-transfer.github.io/

Google QUIC - https://www.chromium.org/quic

Experimental tools

Google QUIC - https://www.chromium.org/quic

CERN FTS - https://fts.web.cern.ch/

Bittorrent…

Data Transfer Nodes

Katana Data Mover

Has RSYNC/SFTP, other linux apps can be installed

- Can assist with moving data to the UNSW Data Archive

Raijin Data Mover

Transferring in: BBCP/RSYNC/SFTP to r-dm.nci.org.au

Transferring out: Create a job in COPYQ queue (or it'll get terminated)

Other Considerations

Compression – zlib, tar (single big file vs lots of small files)
64gb in 2.7 million files – 12 hours, same files in a tar - 1 hour.

Security – Encryption, dedicated network, authentication

Network Congestion – Scheduling transfers

Dealing with failure – restart, continue, stop?

Incremental transfers – only upload what you need to. Checksums, modification dates

Scenarios

- Archiving data from a remote system to the UNSW Data Archive
- Uploading data to Raijin at NCI
- Downloading a large dataset from a repository
- Pushing data from an instrument to bulk storage

Acknowledgements

ES-Net https://fasterdata.es.net/

Google Cloud

Network World https://www.networkworld.com

RDM@
UNSW

A Single Point of Contact

rdm@unsw.edu.au

Questions?