Long Term Storage Facility

Enabling and Optimising flexible 'big data' Workflows

Dave Morrison – Storage Lead

February 2015

INFORMATION MANAGEMENT AND TECHNOLOGY (IM&T)
www.csiro.au

CSIRO

For 'big data' to have a future ...

As data growth and proliferation continues to outpace research grade infrastructure, do we need a new approach to the problem?

We have to ask 'What good is big data?'

If its unable to speak?

If it only ever repeats the same story?

If it can not repeat the same story twice?

If it speaks so slowly the message is lost?

If it can not perform in an orchestra?

If it doesn't speak to the world?

For 'big data' to have a future ...

As data growth and proliferation continues to outpace research grade infrastructure, do we need a new approach to the problem?

4 years ago CSIRO:

Had an estimated **89PB** of data heading its way ...

Had data **without 'context'** ...

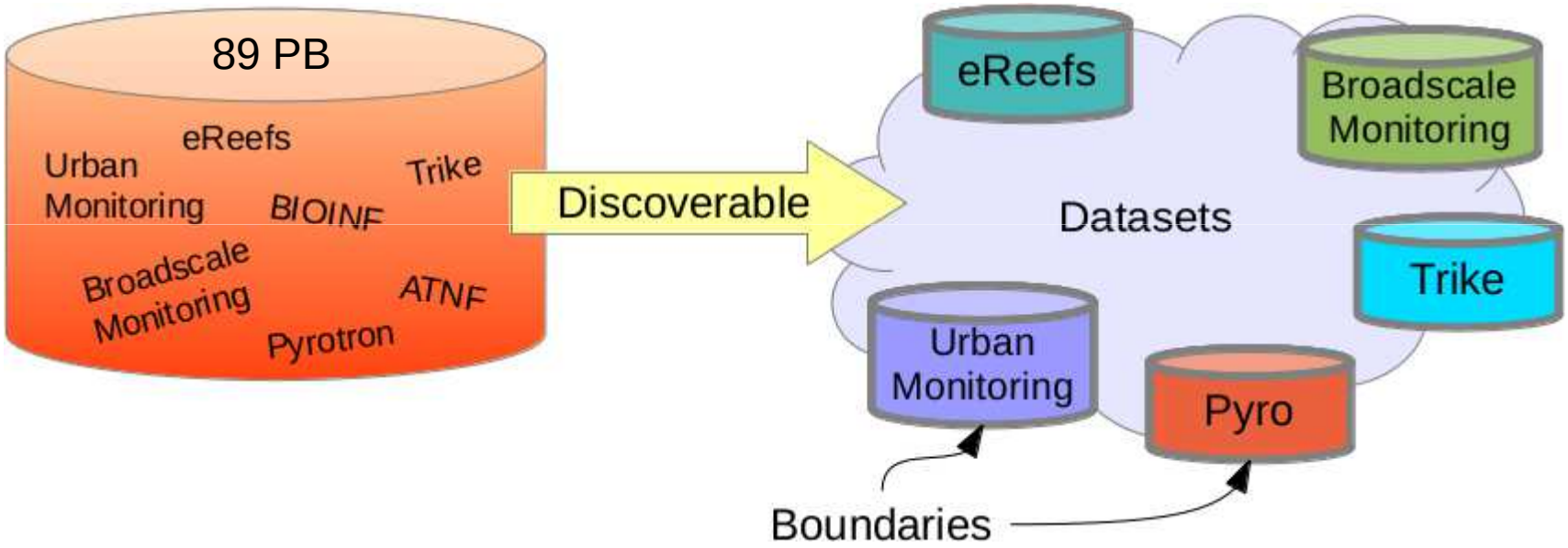
Was **disconnected** from compute …

Was using **enterprise grade infrastructure** for research …

We had to identify and develop better ways!

'big data' needs to be discoverable ...

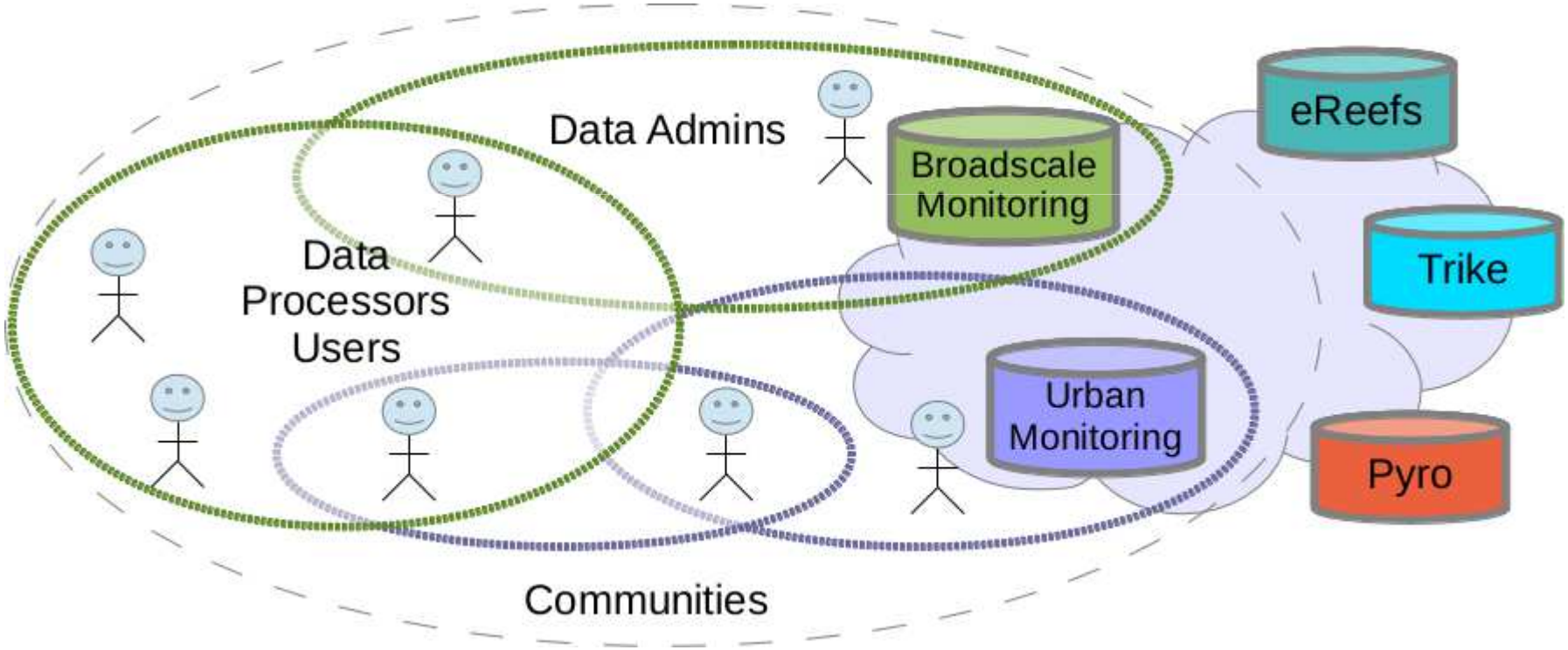
We started with 'discoverable' ...



'big data' needs to have relationships ...

We established the 'relationship' …

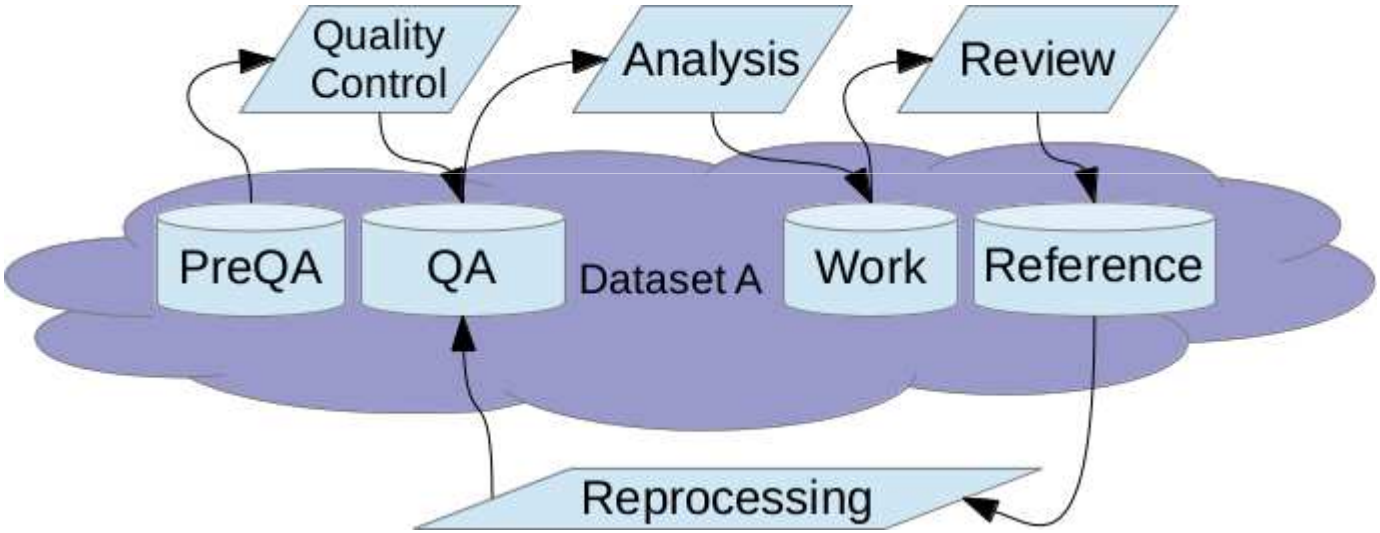
with owners, users, consumers, ...



'big data' needs to have relationships ...

We established the 'relationship' …

within the dataset and w.r.t. the workflow ...

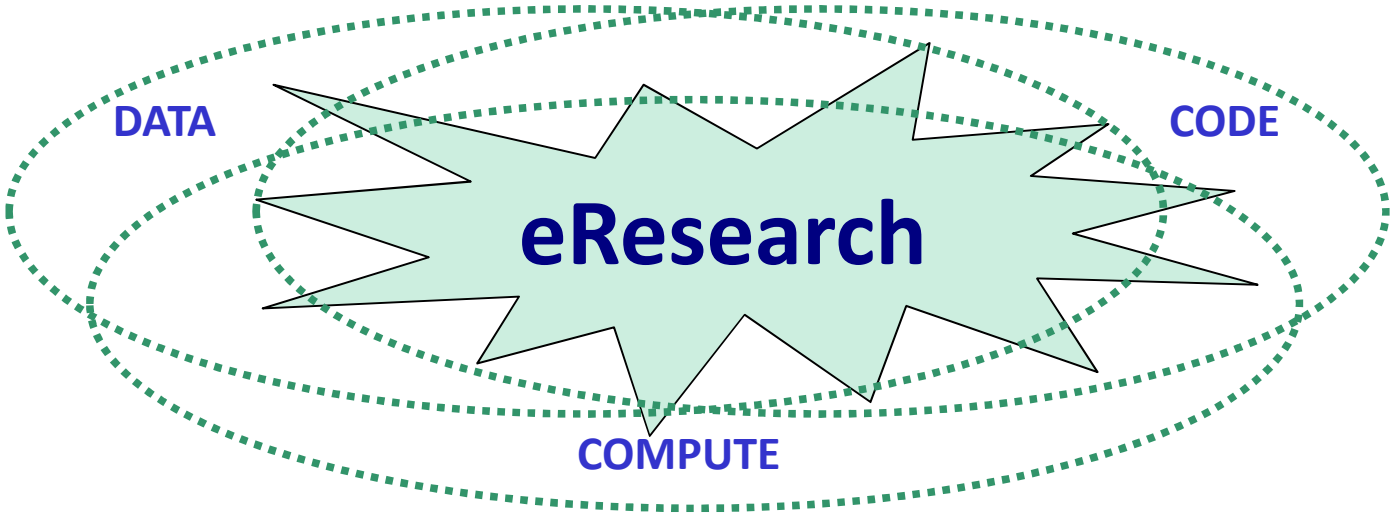


Well connected 'big data' …

… tight coupling

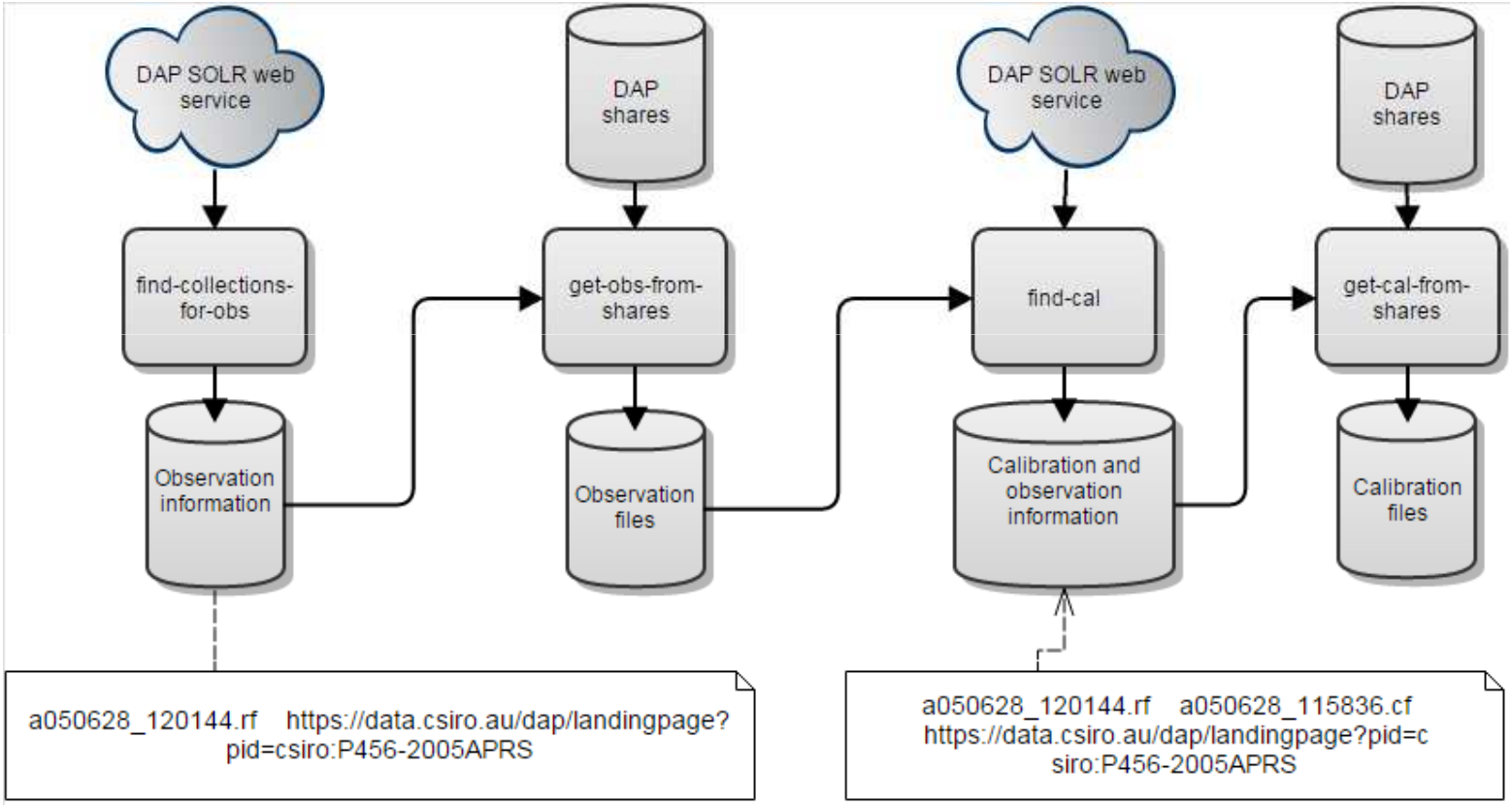
This is where your workflow comes to life and we focus on:

- Collecting the data
- Improving workflows
- Accelerating outcomes



Can we transition from having 'lots of data' into 'big data'

while 'opening up new workflow possibilities' ...



Can we transition from having 'lots of data' into 'big data'

while 'opening up new workflow possibilities' ...

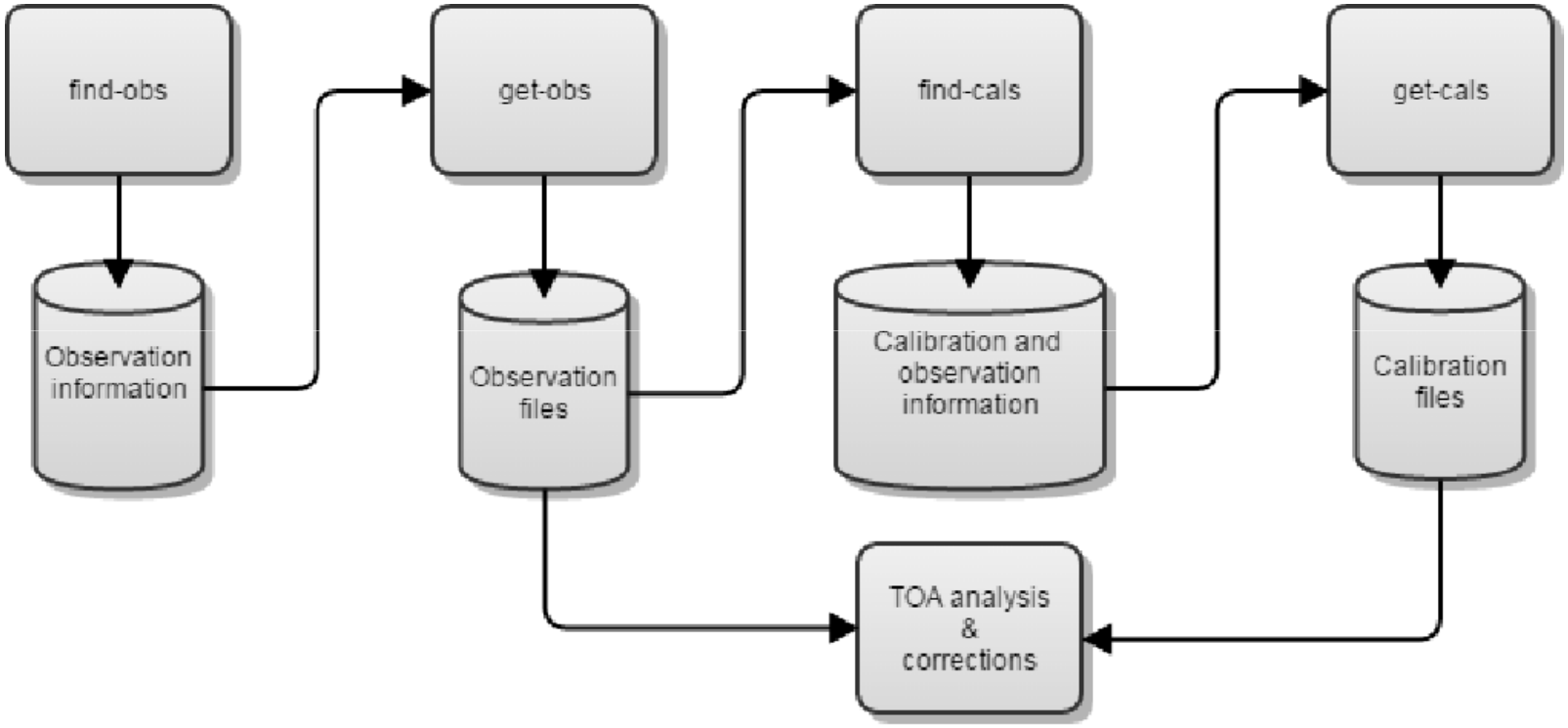


Figure 3: Example processing workflow

Can we transition from having 'lots of data' into 'big data'

while 'opening up new workflow possibilities' ...

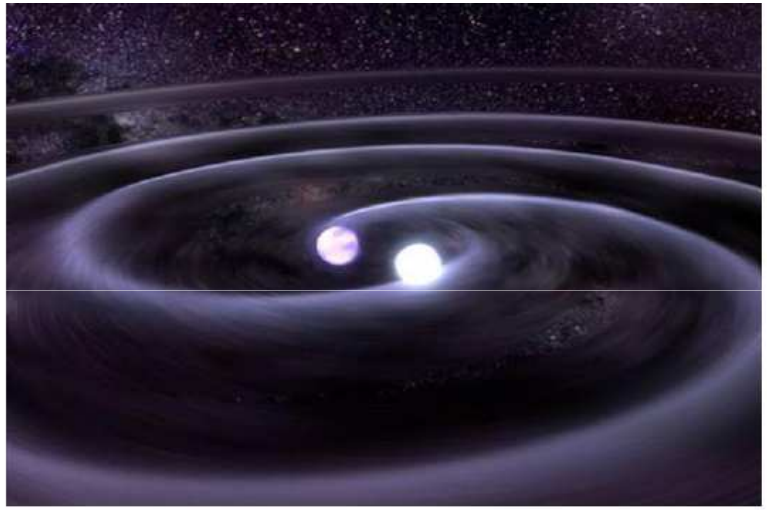


Figure 4: Binary system (black hole/neutron star pair).

Sourced from: Quantum Day (http://tinyurl.com/om7rkbb)

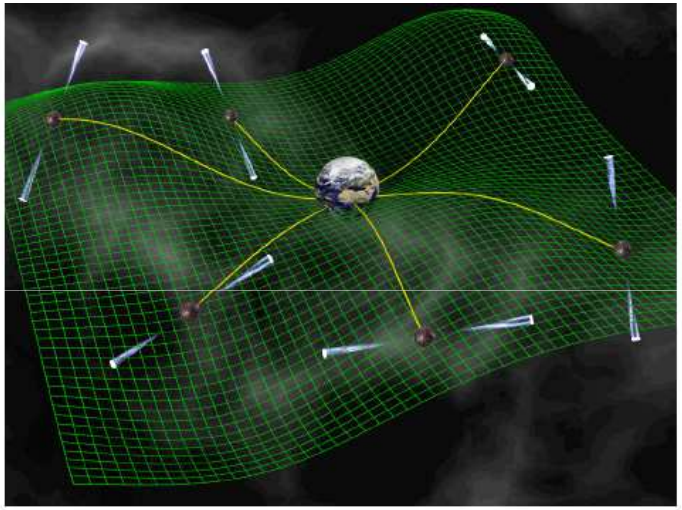


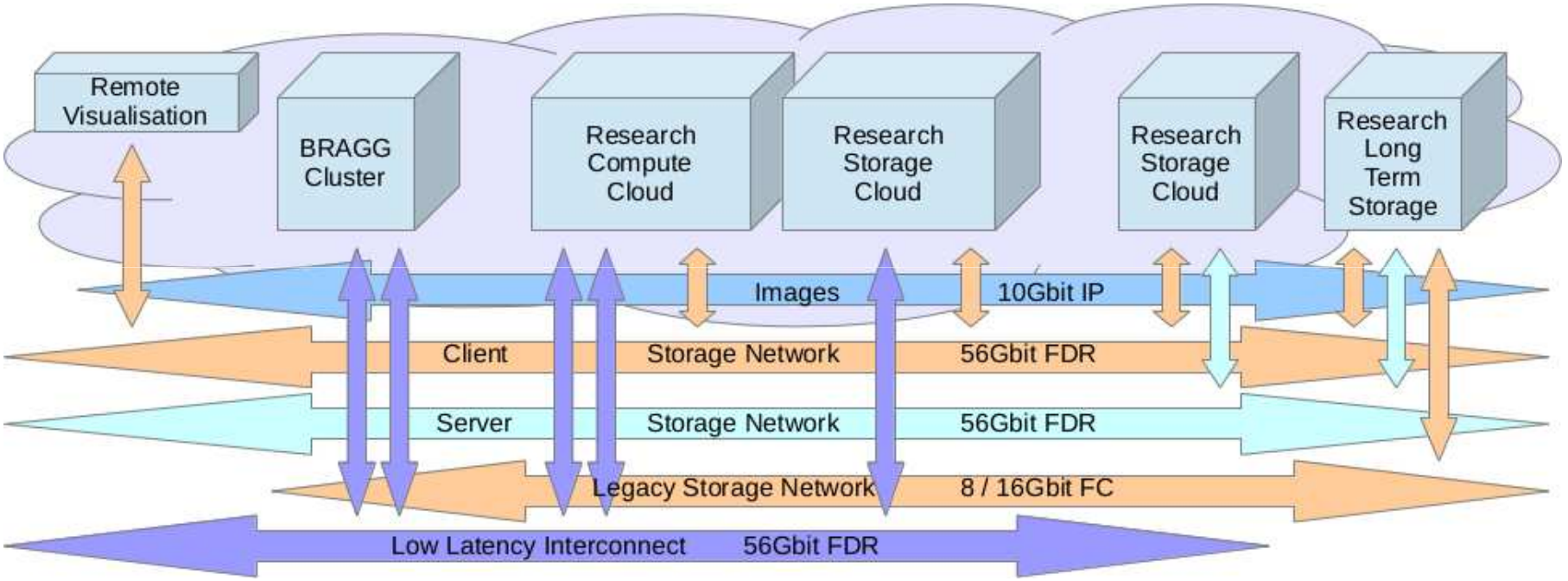
Figure 5: Space-time warping from gravitational waves causes Earth to "move" with respect to pulsars, delaying or advancing pulse arrival times. Courtesy: David Champion

Source: Physics World (http://tinyurl.com/kjs2fo4)

Well connected 'big data' …

… low latency non-blocking infrastructure

Dedicated, non-blocking, short, point-to-point links, ...



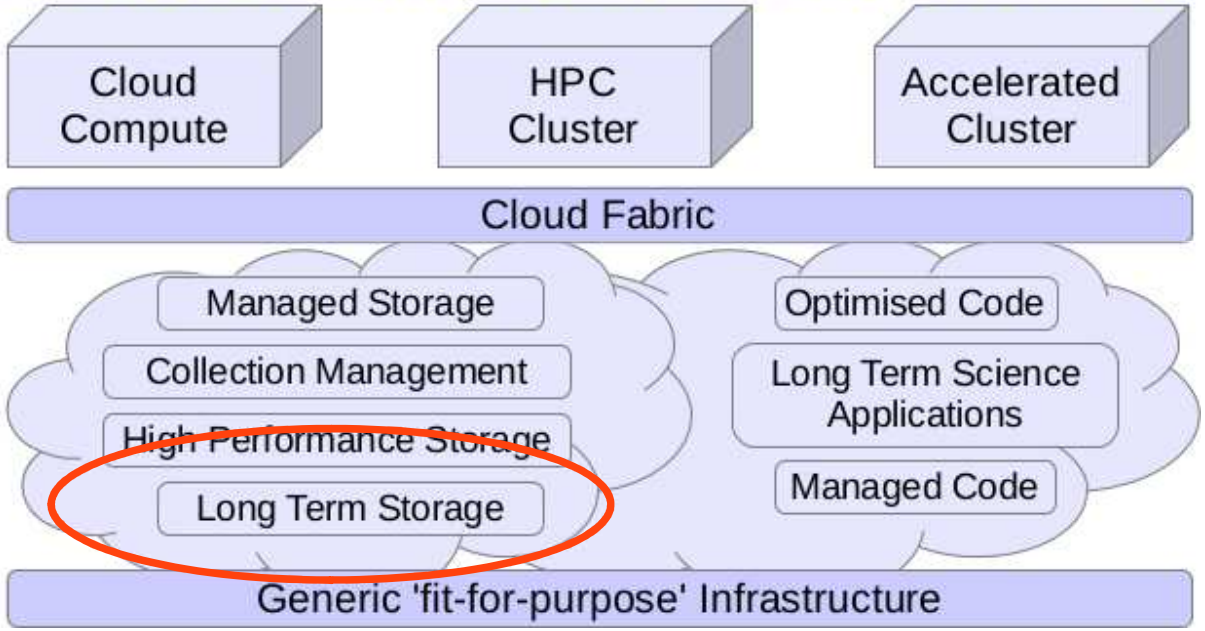
Well connected 'big data' …

… fit-for-purpose infrastructure

Split the tech from the workflow, abstracted the brands, …

Below the line:

Generic pool of 'fit for purpose' infrastructure abstracted by a layer of automation and virtualisation where appropriate.



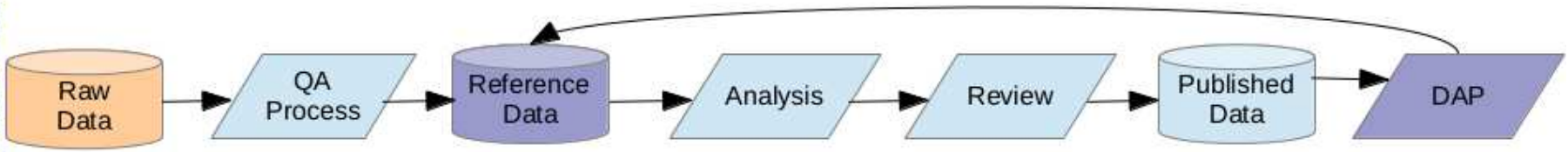
Well connected 'big data' …

… fit-for-purpose infrastructure

'Above the line' we focused on business outcomes, generic pool customised to the profile of the research …

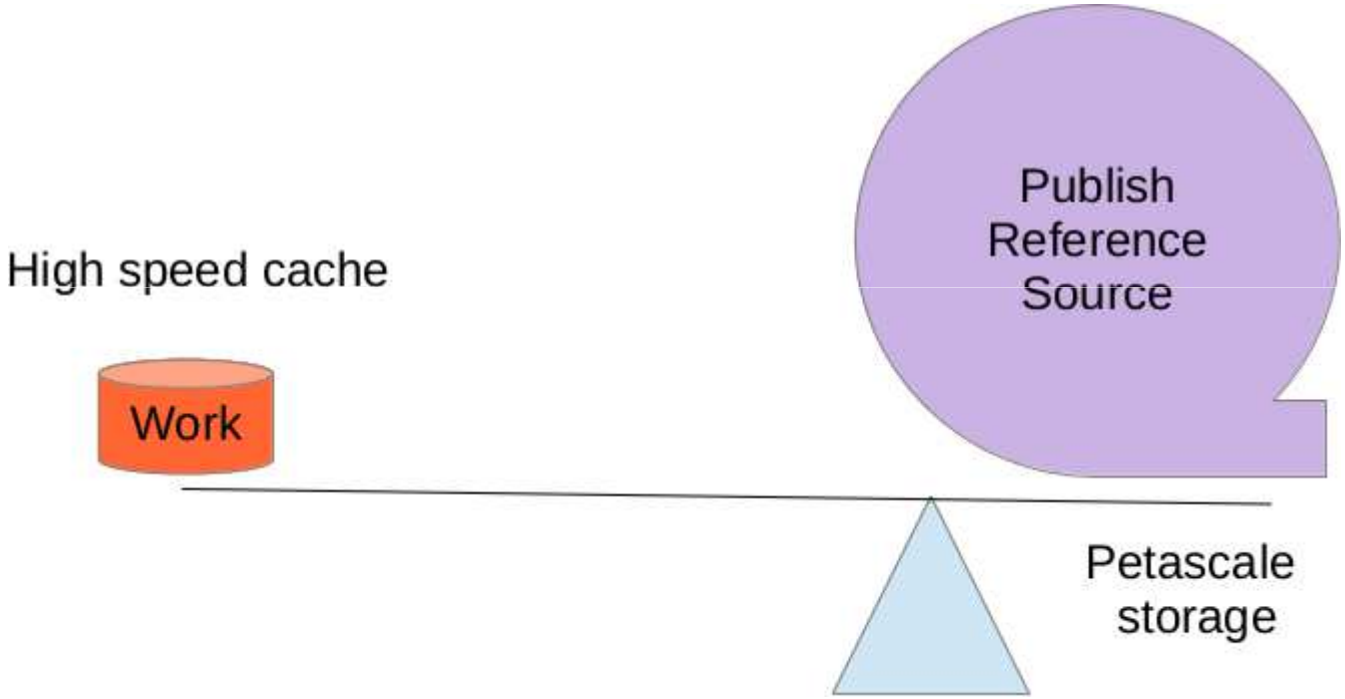
Above the line:

The generic pool of 'fit for purpose' infrastructure is mapped to the specific research workflows



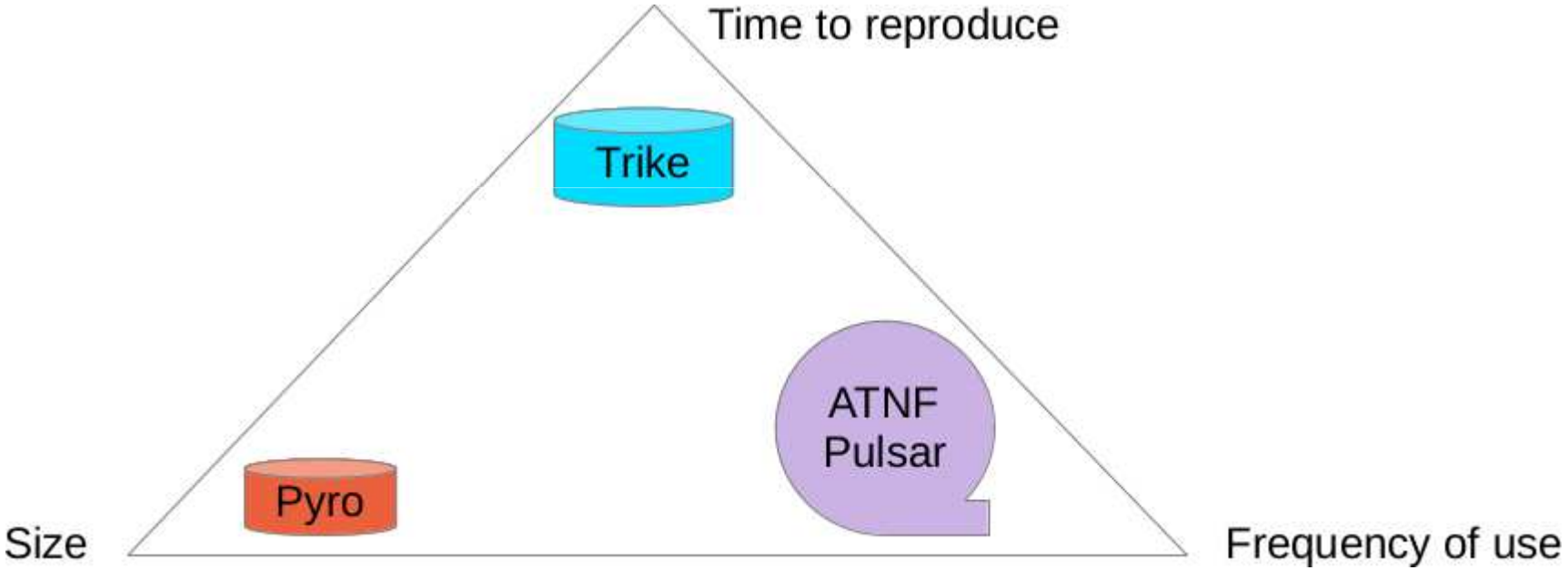
Can we transition from having 'lots of data' into 'big data'

while 'reducing costs' (per TB) ...



Can we transition from having 'lots of data' into 'big data'

while 'improving data management practices' ...



For 'big data' to have a future …

quick summary …

Discoverable 'datasets' with well known and permanent names

'Dataset' related:

- Admins, Processors, Reviewers, Consumers, Communities, Inter-community links

'Dataset' connected:

- With consideration of the scale
- Ensuring data-code-compute are all present
- Low latency and non-blocking i.e. wire speed connections

Flexible mapping of subsets of the dataset to the researchers workflow.

Provided a mechanism for communicating:

- Significants, level of performance, level of protection to the infrastructure

Moving forward we have the foundation for:

- Provenance, communities, multi-domain discovery and processing

For 'big data' to have a future ...

The CSIRO Strategy ...

Increasing bandwidth of data to compute by building disk arrays not for 'data storage' but for 'high speed data cache'

Transforming the use of peta-scale tape libraries from 'backup' (as an after thought) to 'data storage with integrated protection.'

Escaping the monolithic data problem by containerising datasets to achieve a flexible mechanism for connecting data to workflows, and

Allowing throughput optimisations by using pre-defined data categories to communicate access patterns and protection regimes to the infrastructure.

Thank you