Data Intensive Research

Enabling and Optimising flexible 'big data' Workflows

Ian Corner – Design and Implementation Lead IM&T

June 2015

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

CSIRO

Introducing CSIRO

Commonwealth Scientific and Industrial Research Organisation

Who we are

People	**~6000**
Sites	**55**
Flagships	**9**
Budget	**$1B+**

64% of our people hold university degrees over

2000 hold doctorates

500 hold masters

We mentor **832** postgraduate research students with our university partners

Big ideas start here

Top 10 applied research agency globally

Top 1% of global research institutions in **14** of **22** research fields

Top 0.1% in **4** research fields

2000+ industry partners totaling **$220M+** per year

Key connector of Australian institutions in **6** research fields according to network analysis

Our operating model

NATIONAL FACILITIES & COLLECTIONS

Infrastructure and biological collections managed by CSIRO for the benefit of research and industry.

IMPACT SCIENCE

Nine National Research Flagships with a focus on the biggest challenges facing the nation.

CSIRO SERVICES

Commercial, customer-centric products and services for industry, government and the community.

SUPPORT SERVICES

CSIRO – Collections

Commonwealth Scientific and Industrial Research Organisation

Australia

Some of our existing collections:

- 12,000,000 insects
- 5,000 fish
- 1,000 strains of more than 300 microalgae species
- 1,000,000 herbarium, some dating back to Captain Cook's 1770 expedition to Australia
- 200,000 irreplaceable scientific specimens of wildlife

http://www.csiro.au/en/Research/Collections

CSIRO – Yesterdays Collections

Physical specimens. Captured and preserved

http://csironewsblog.com/2012/09/07/were-insects-and-were-everywhere/

CSIRO – Todays Collections

Digitised in the field. Automated sensing.

http://csironewsblog.com/2014/01/15/tiny-technology-creates-a-buzz/

Data Intensive Research

Optimising the Workflow

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?

What good is 'big data' if …

It is unable to speak?

It only ever repeats the same story?

It can not repeat the same story twice?

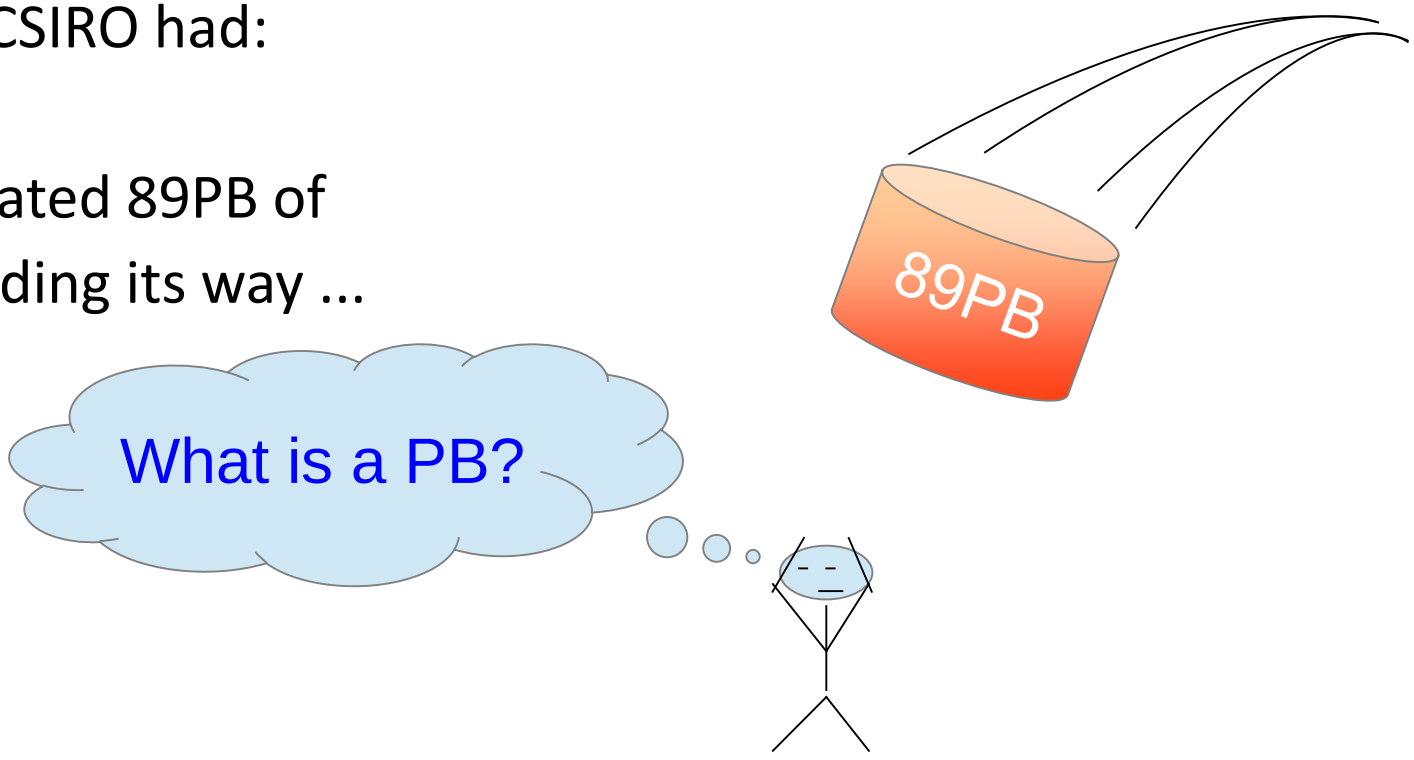
It speaks so slowly the message is lost?

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 ...

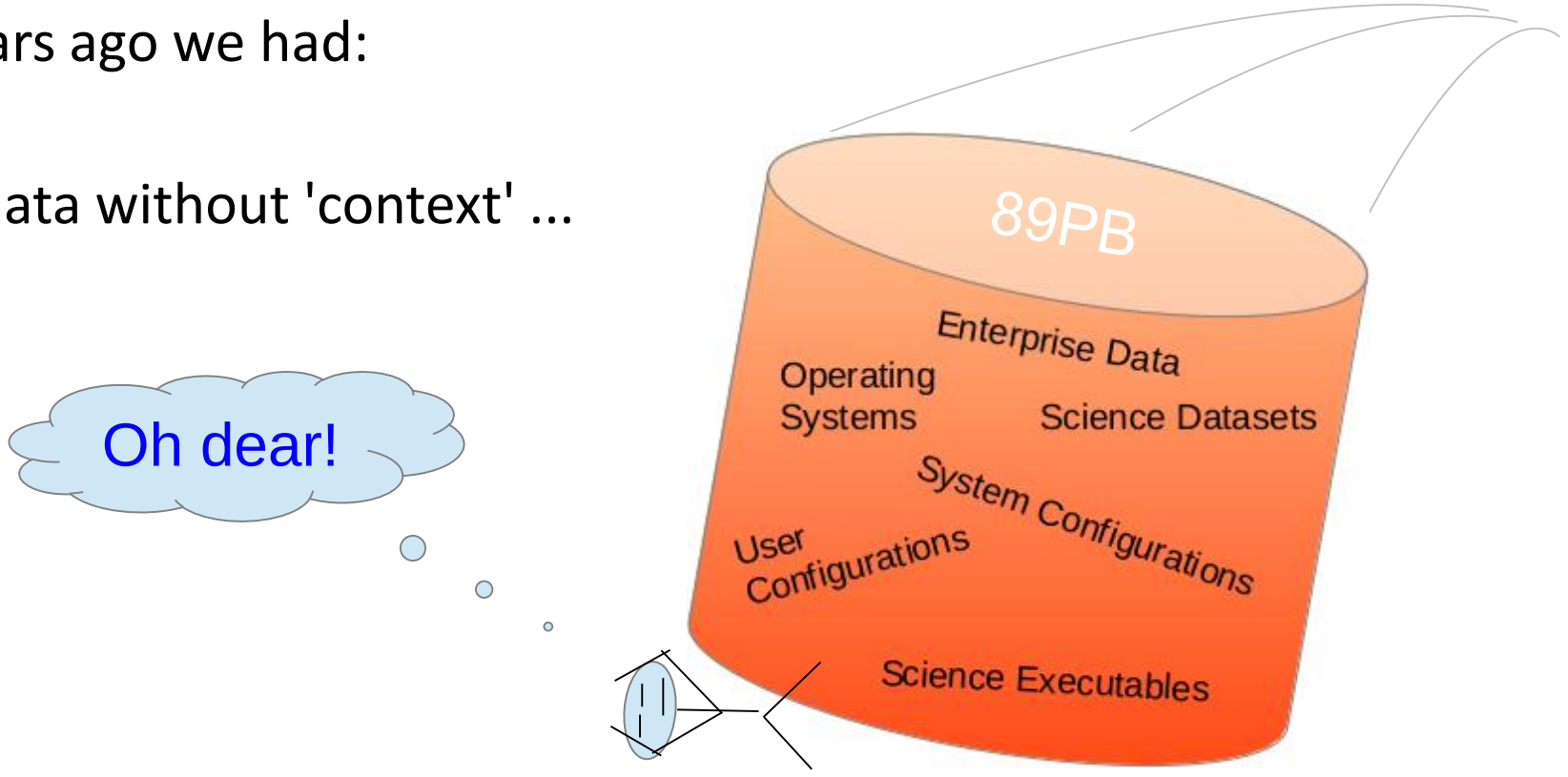


For 'big data' to have a future ...

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

4 years ago we had:

data without 'context' ...

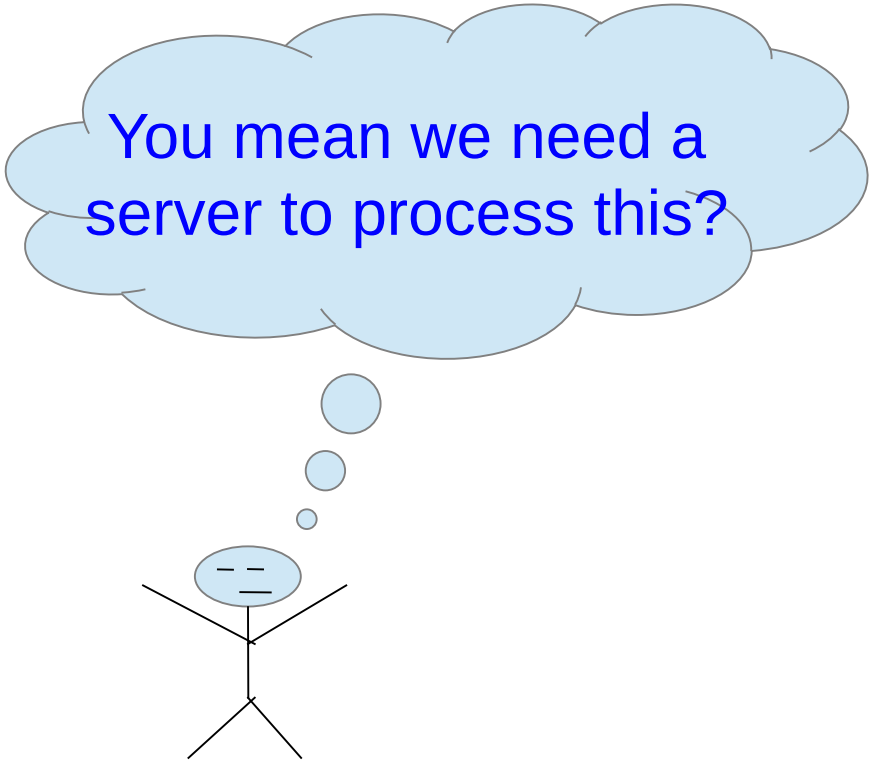


For 'big data' to have a future ...

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

4 years ago we were:

disconnected from compute ...

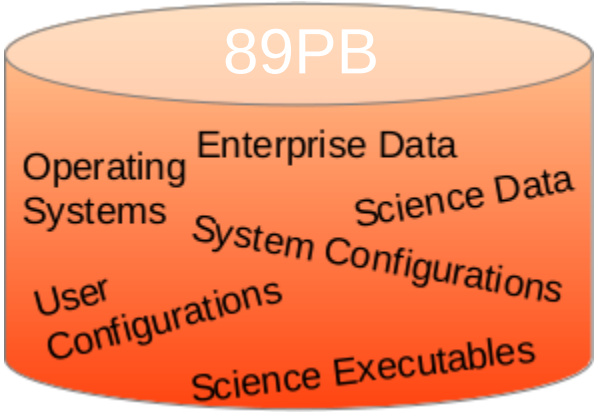


For 'big data' to have a future ...

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

4 years ago we were:

using 'enterprise' grade infrastructure for research…



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 …

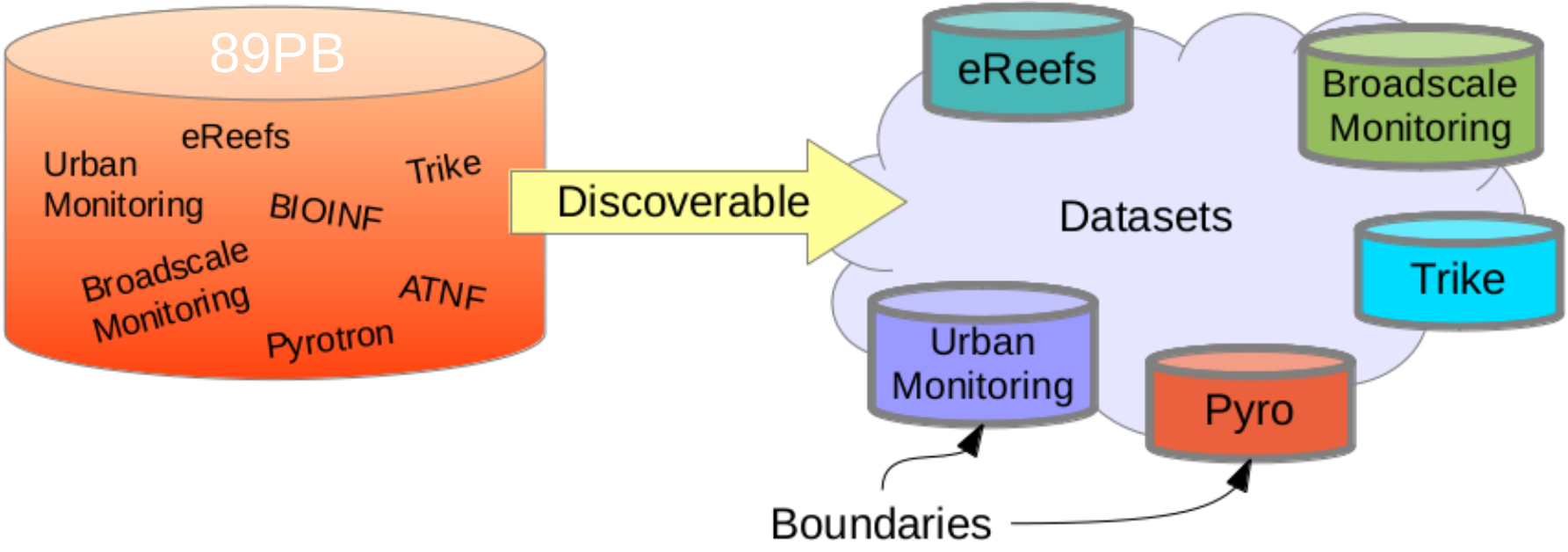
Was using enterprise grade infrastructure for research …

Operating in an environment of reduced funding ...

We had to identify and develop better ways!

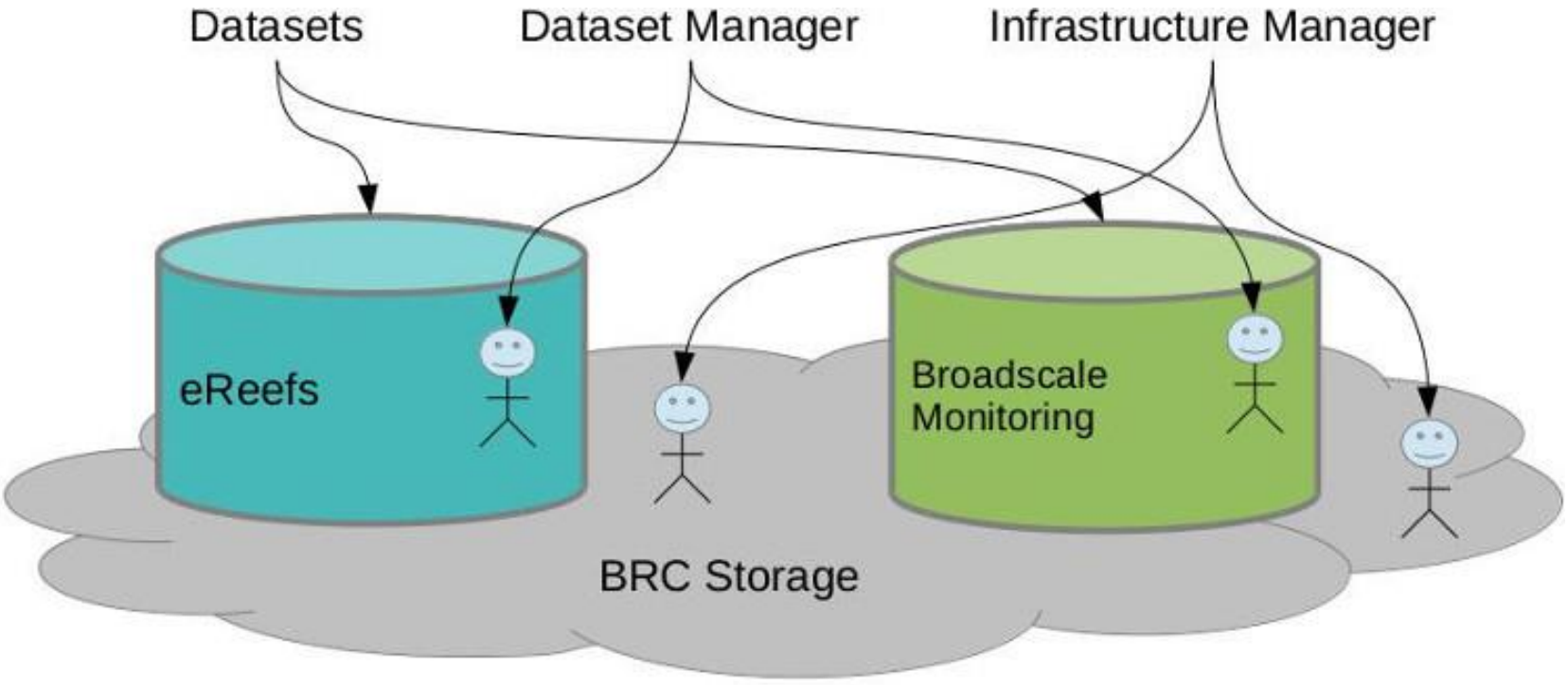
'big data' needs to be discoverable ...

We started with 'discoverable' ...



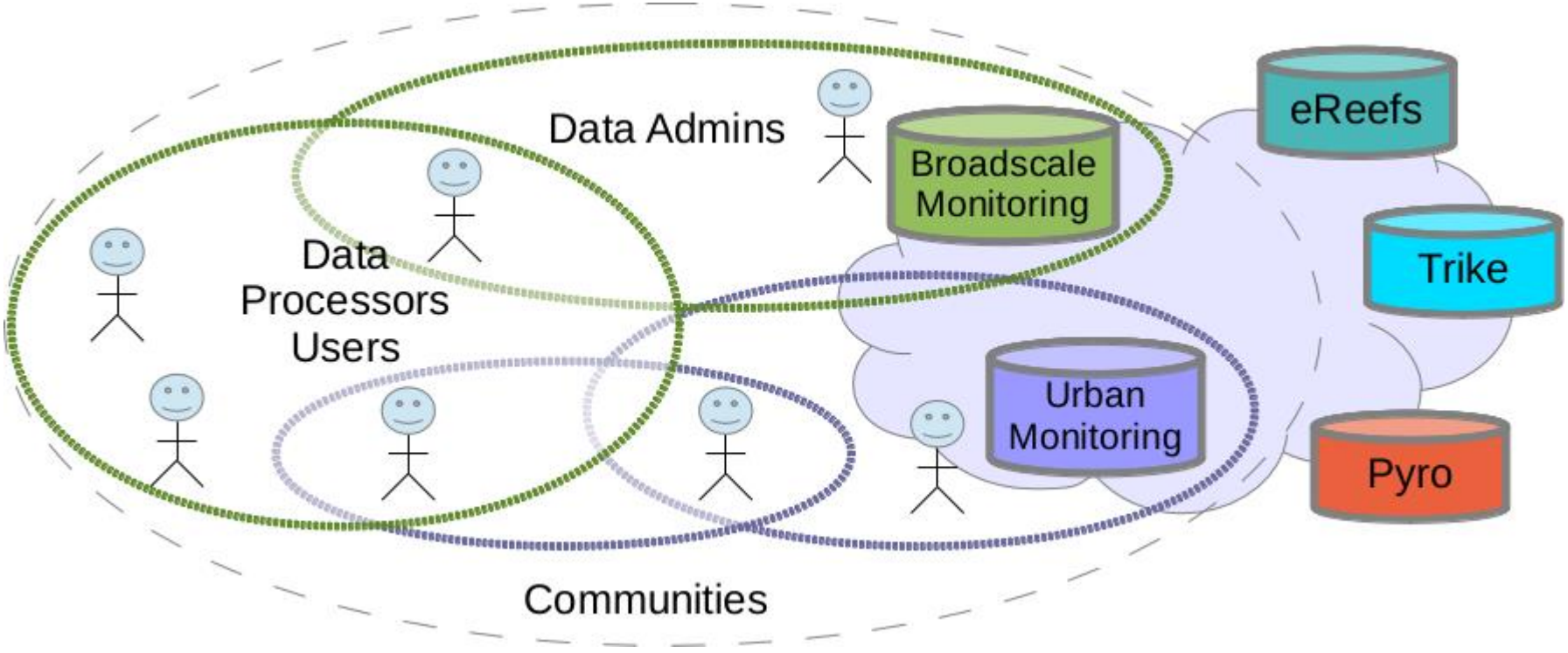
'big data' needs to be managed...

We delineated the 'responsibility' between infrastructure owners and data managers.



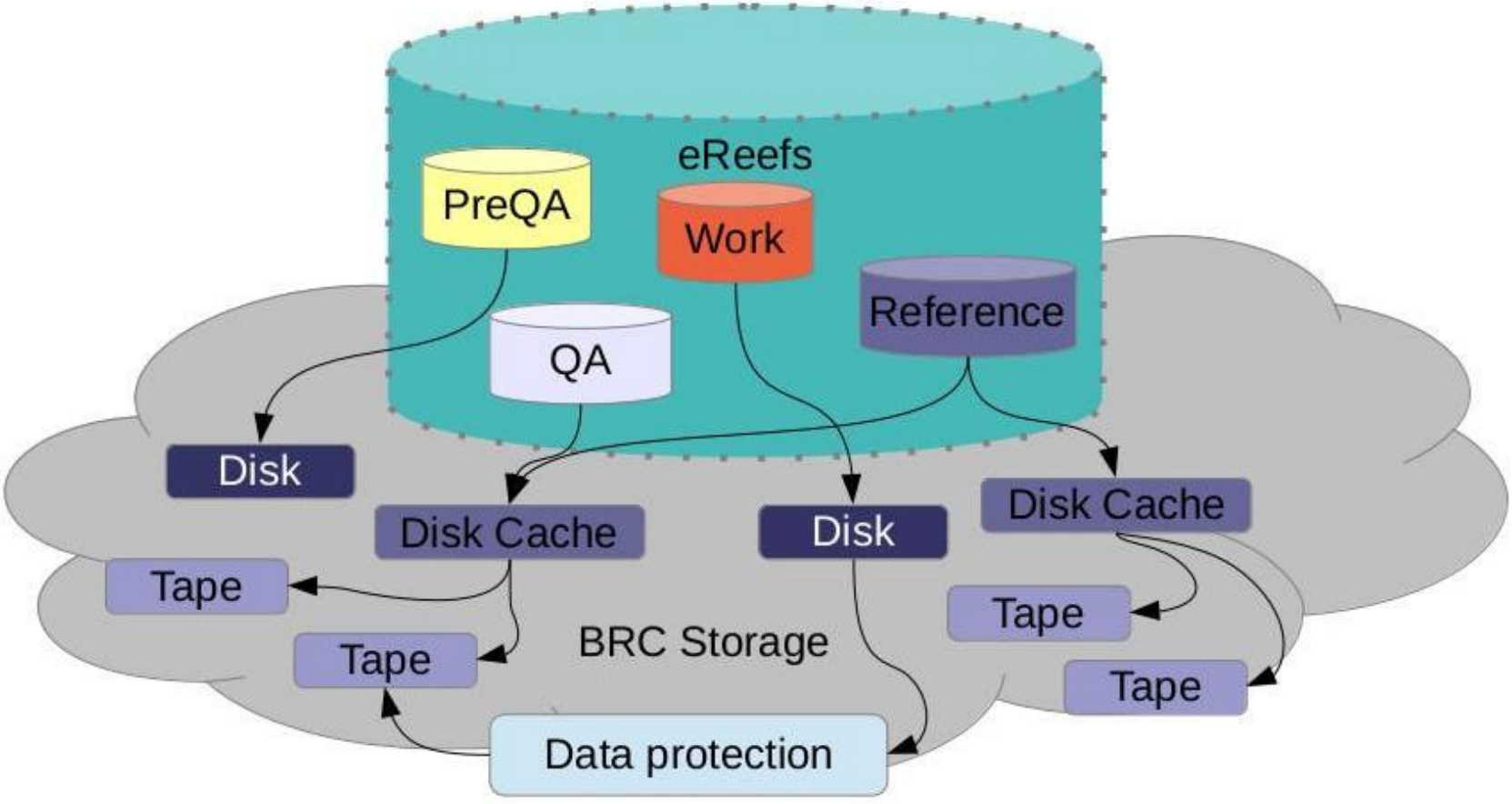
'big data' needs to have relationships ...

We established the 'relationship' with owners, domain specialists, users and consumers.



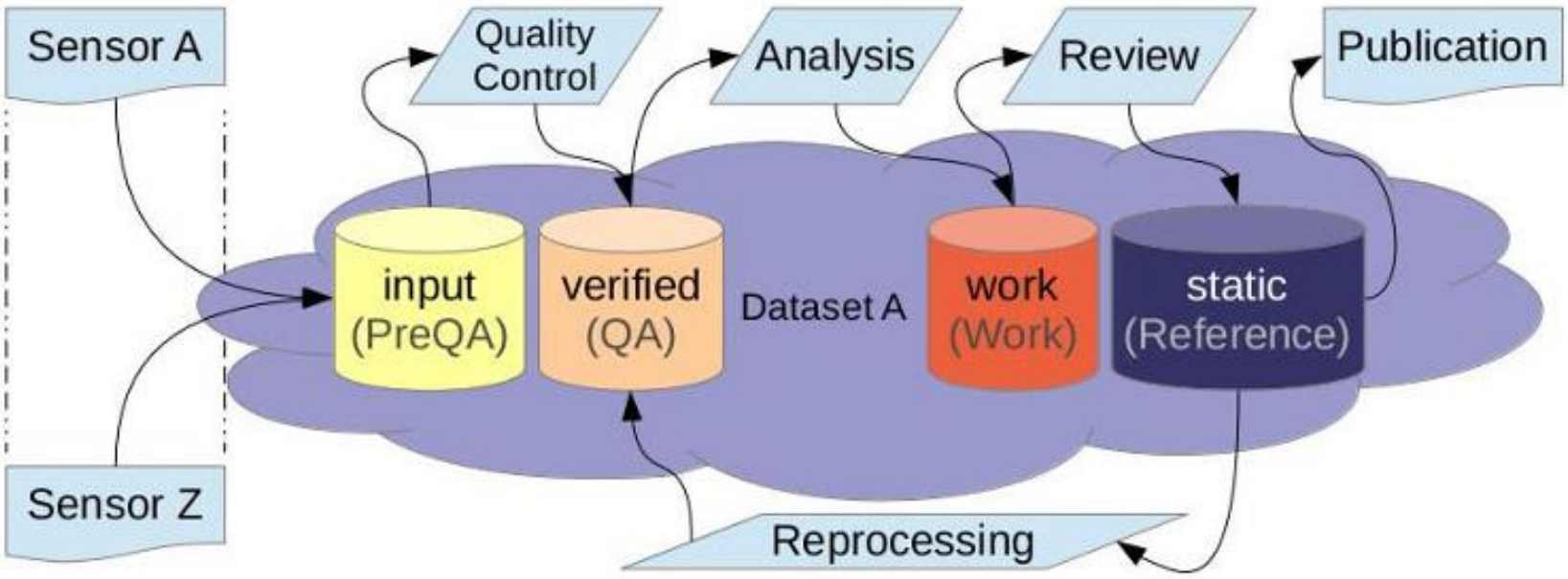
'big data' needs to be optimised ...

Within the dataset we developed 'categories‘ to communicate value to infrastructure.



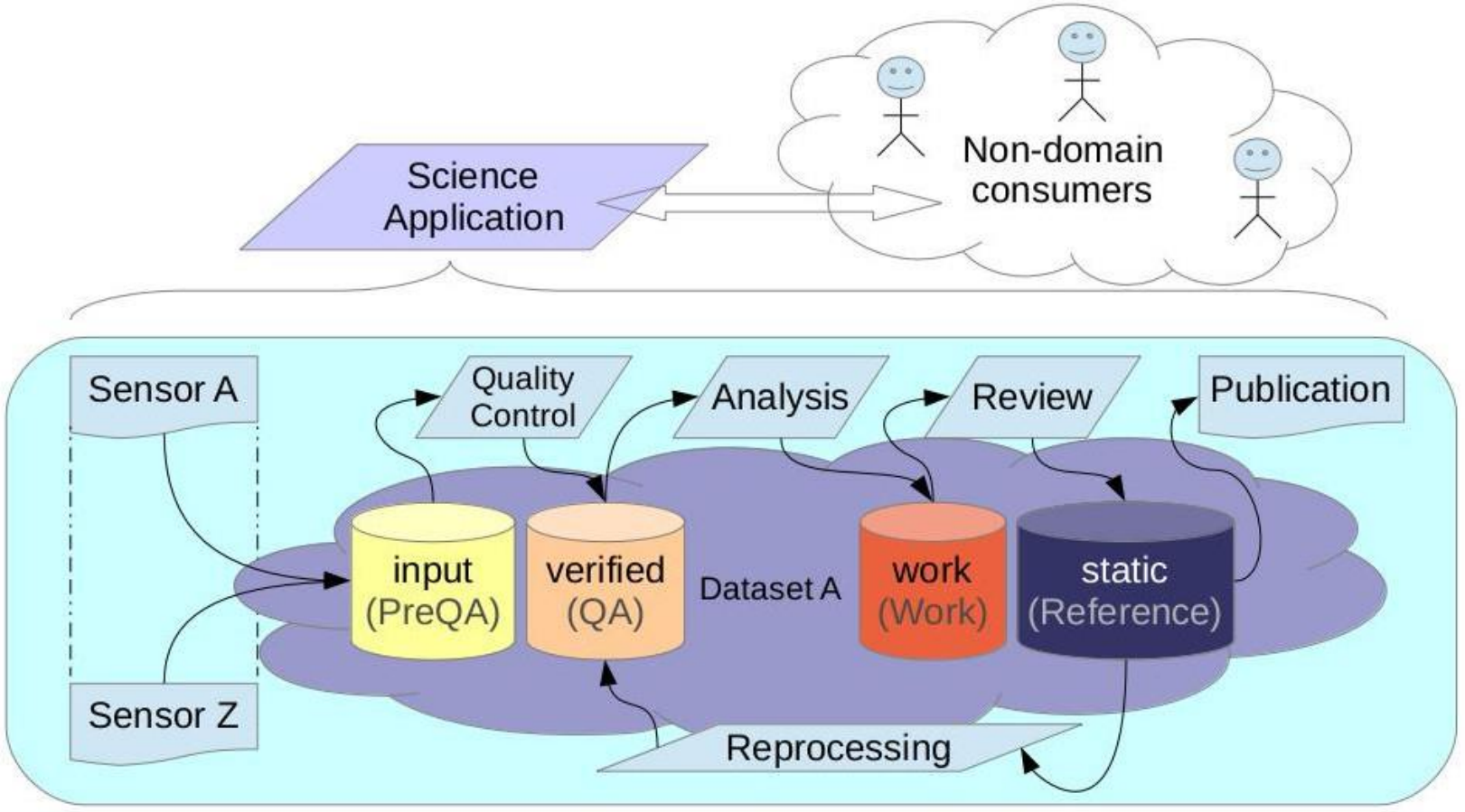
'big data' needs to map to workflows ...

Again within the dataset we created user defined ‘labels‘ enabling mapping to the workflow.



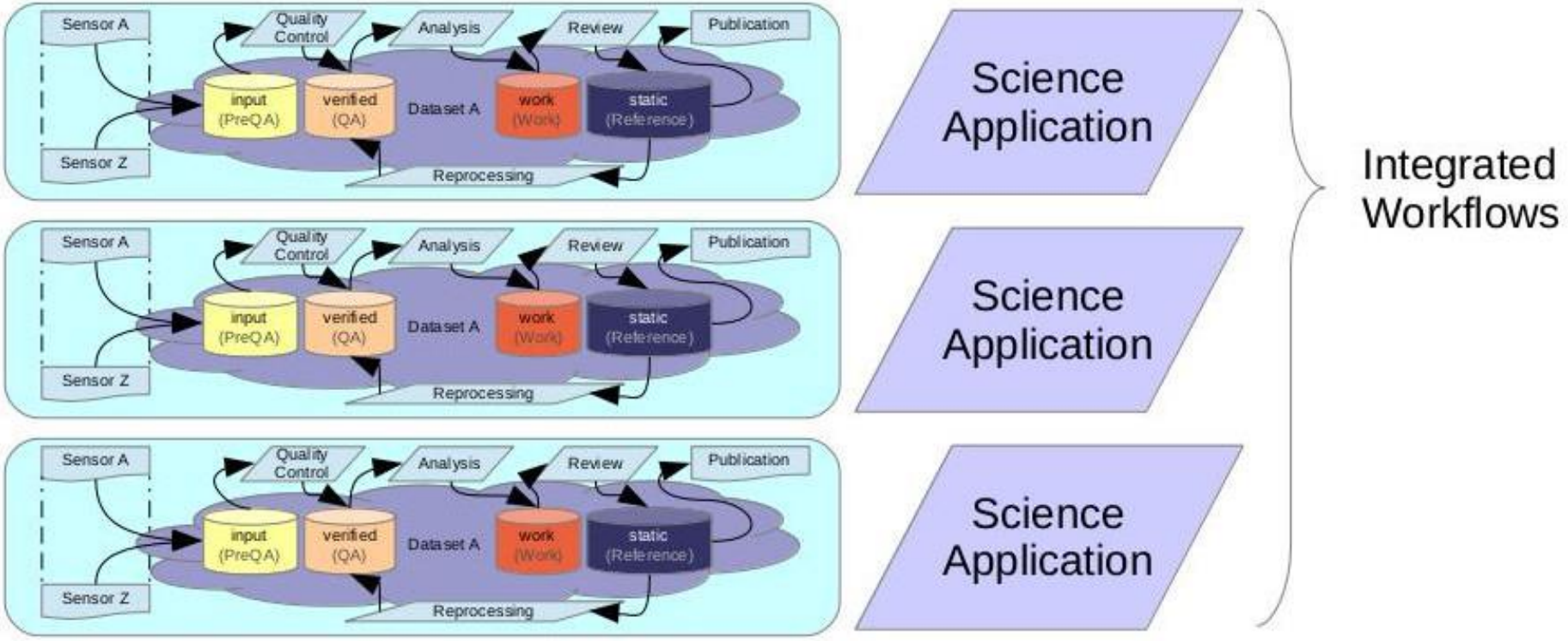
'big data' needs to be consumable ...

We laid the foundation for 'delivery‘ to non-domain consumers.



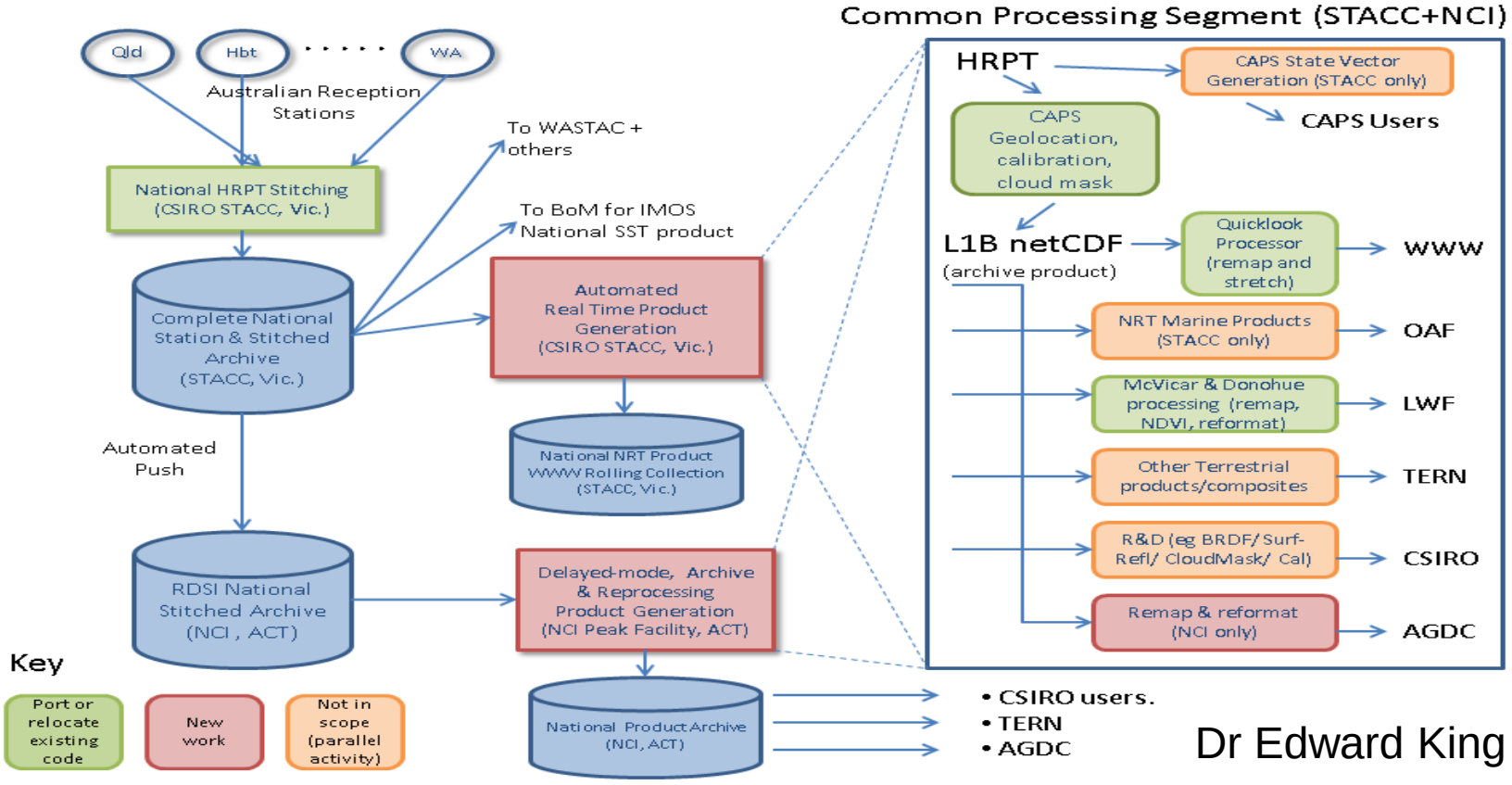
'big data' needs integrated workflows ...

We laid the foundation for 'delivery‘ of integrated workflows.



'big data' needs to be high impact …

We want published and peer reviewed workflows to be 'easily mapped' into future workflows …



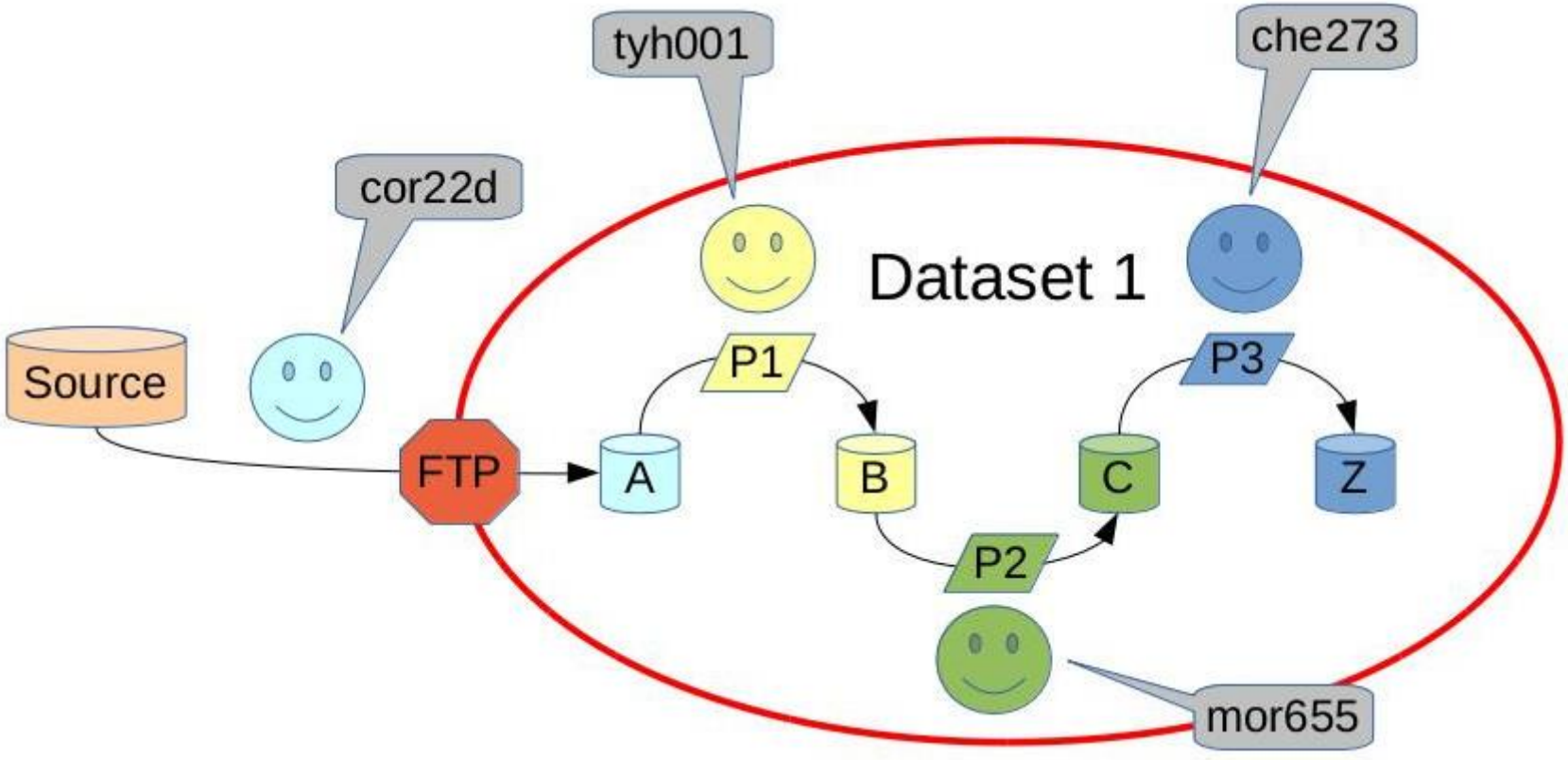
Dr Edward King

'big data' needs to be reproducible ...

We laid one foundation for provenance

'big data' needs provenance ...

We laid the foundation for a 'provenance' framework.



'big data' summary requirements …

‘Big data’ needs to be:

- Discoverable
- Related to:
  - Owners
  - Contributors
  - Domain specialists
- Manageable:
  - Infrastructure
  - Dataset
- Mappable to workflows
- Consumable
- Traceable (i.e. provenance)

Data Intensive Research

Fit-for-purpose infrastructure

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?

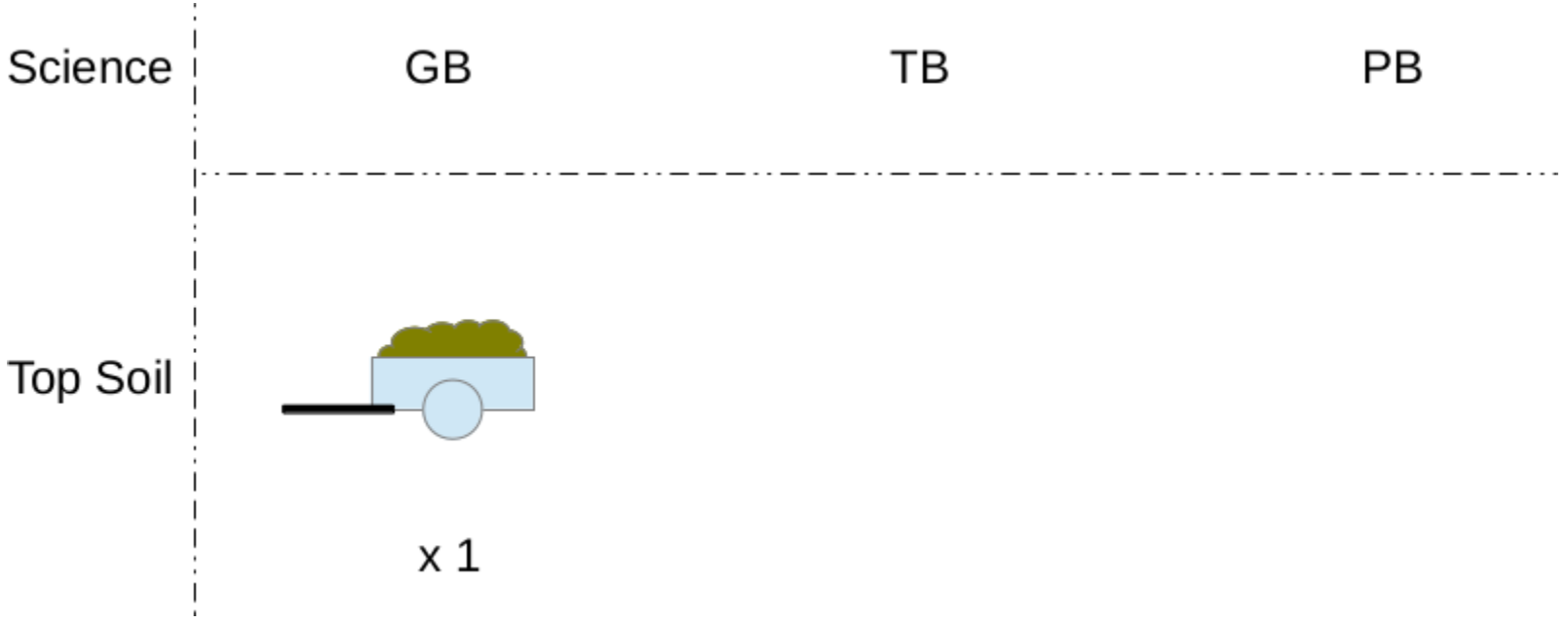
At the infrastructure layer we had to accept that:

- Not all data is equal
- Unstructured data requires boundaries
- Infrastructure will come and go
- Data must be preserved
- Workflow must be optimized
- Provenance must be established and maintained

Well connected 'big data' …

Tight coupling and understanding the scale of data

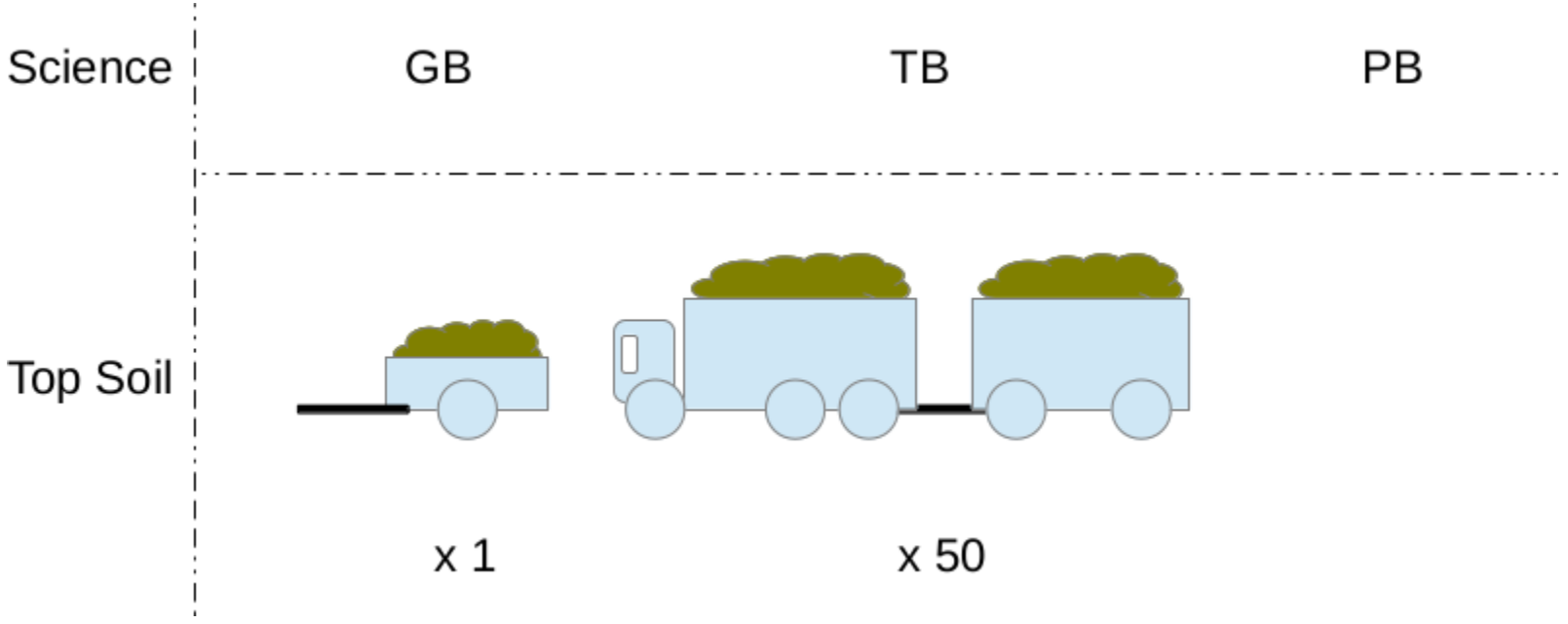
Has anyone here actually seen a GB?



Well connected 'big data' …

Tight coupling and understanding the scale of data

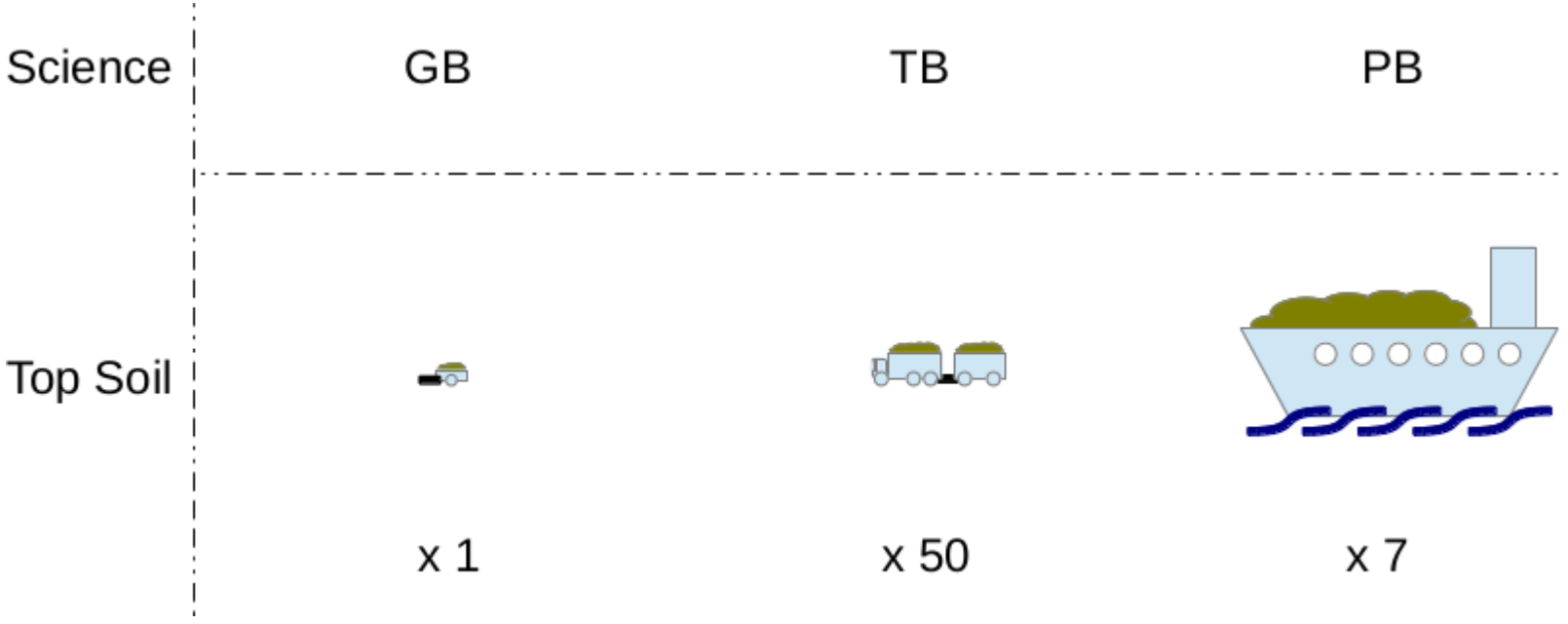
Now what would a TB look like?



Well connected 'big data' …

Tight coupling and understanding the scale of data

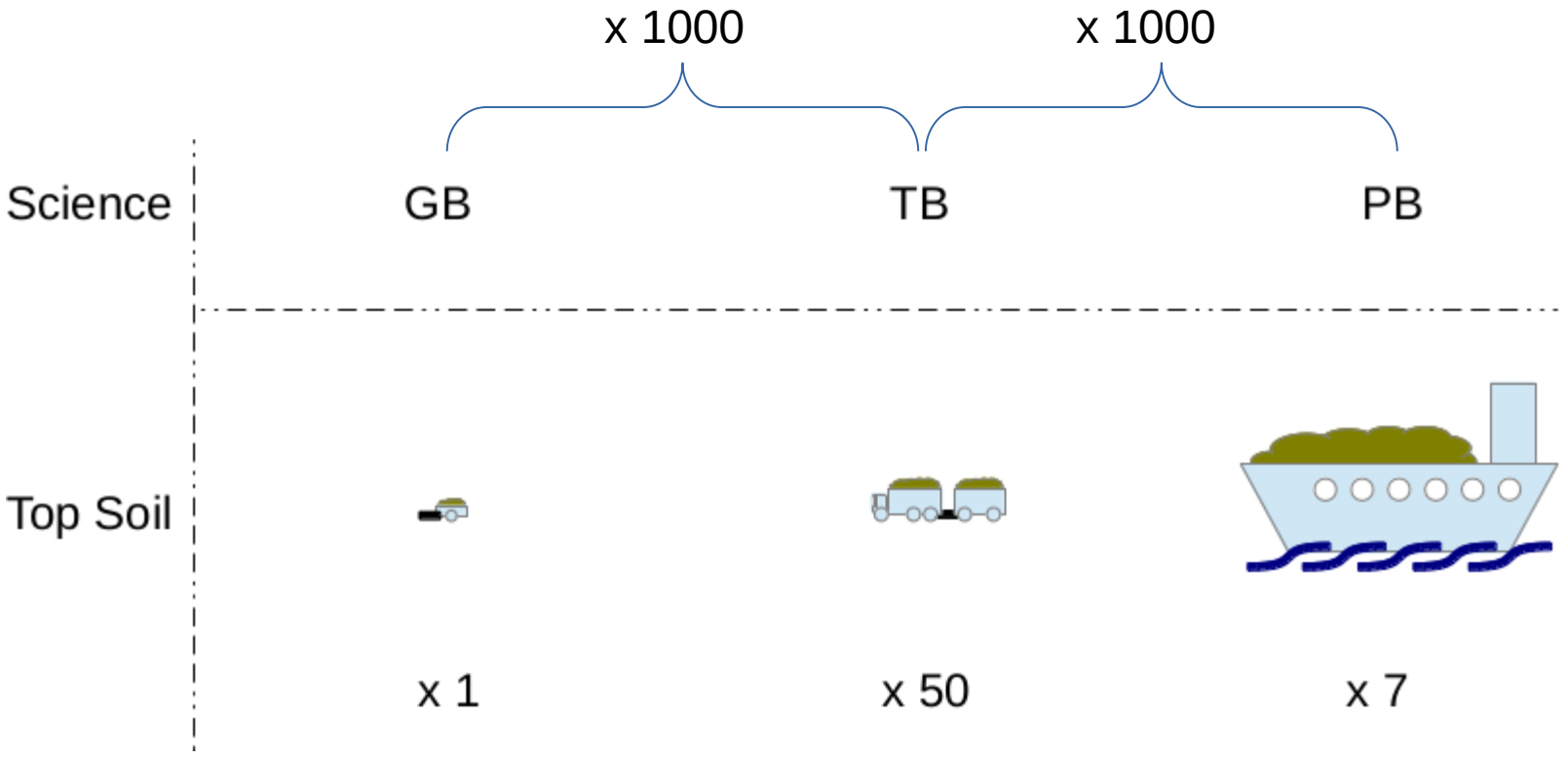
And a PB?



Well connected 'big data' …

Tight coupling and understanding the scale of data

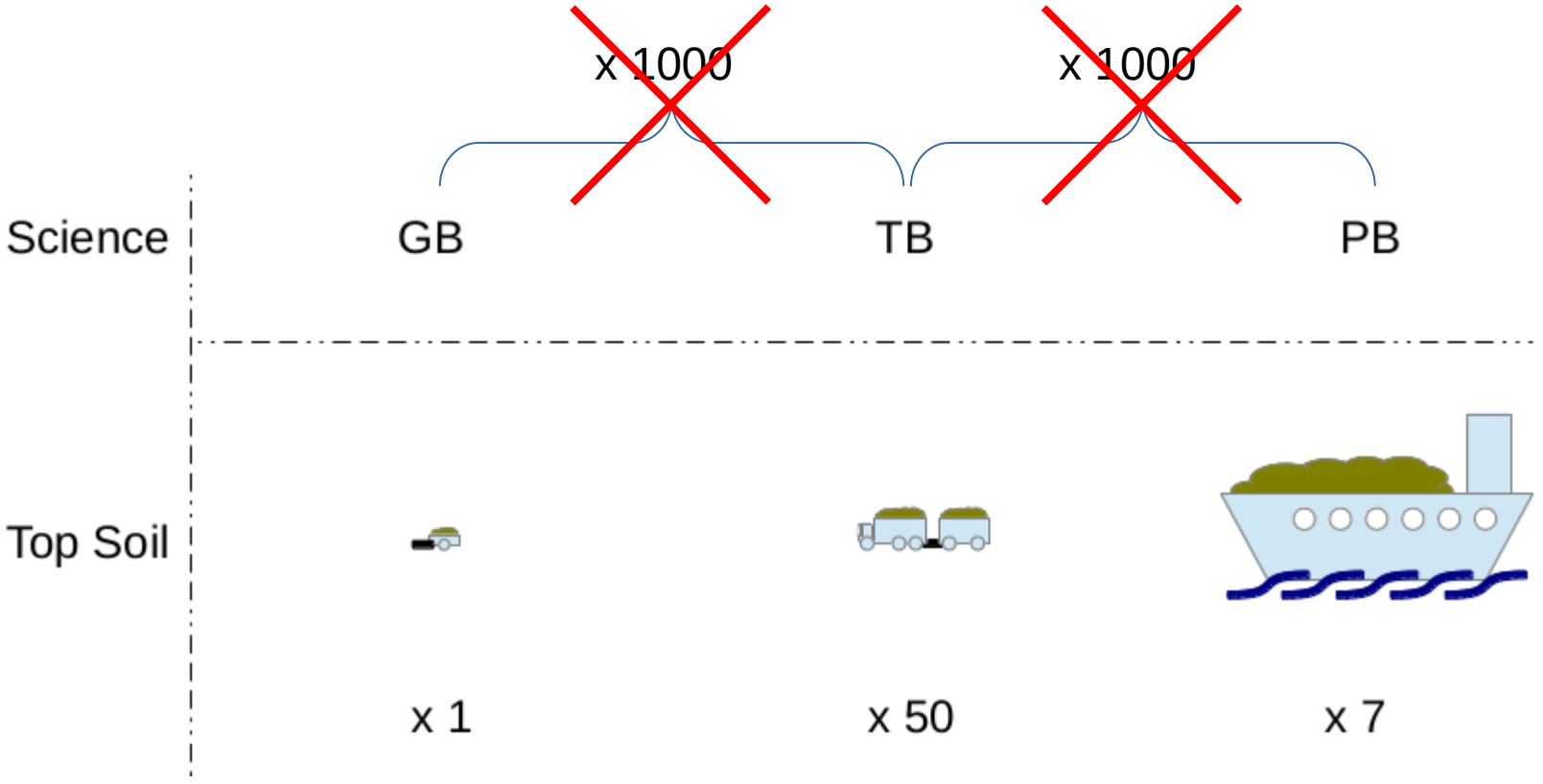
Orders of magnitude?



Well connected 'big data' …

Tight coupling and understanding the scale

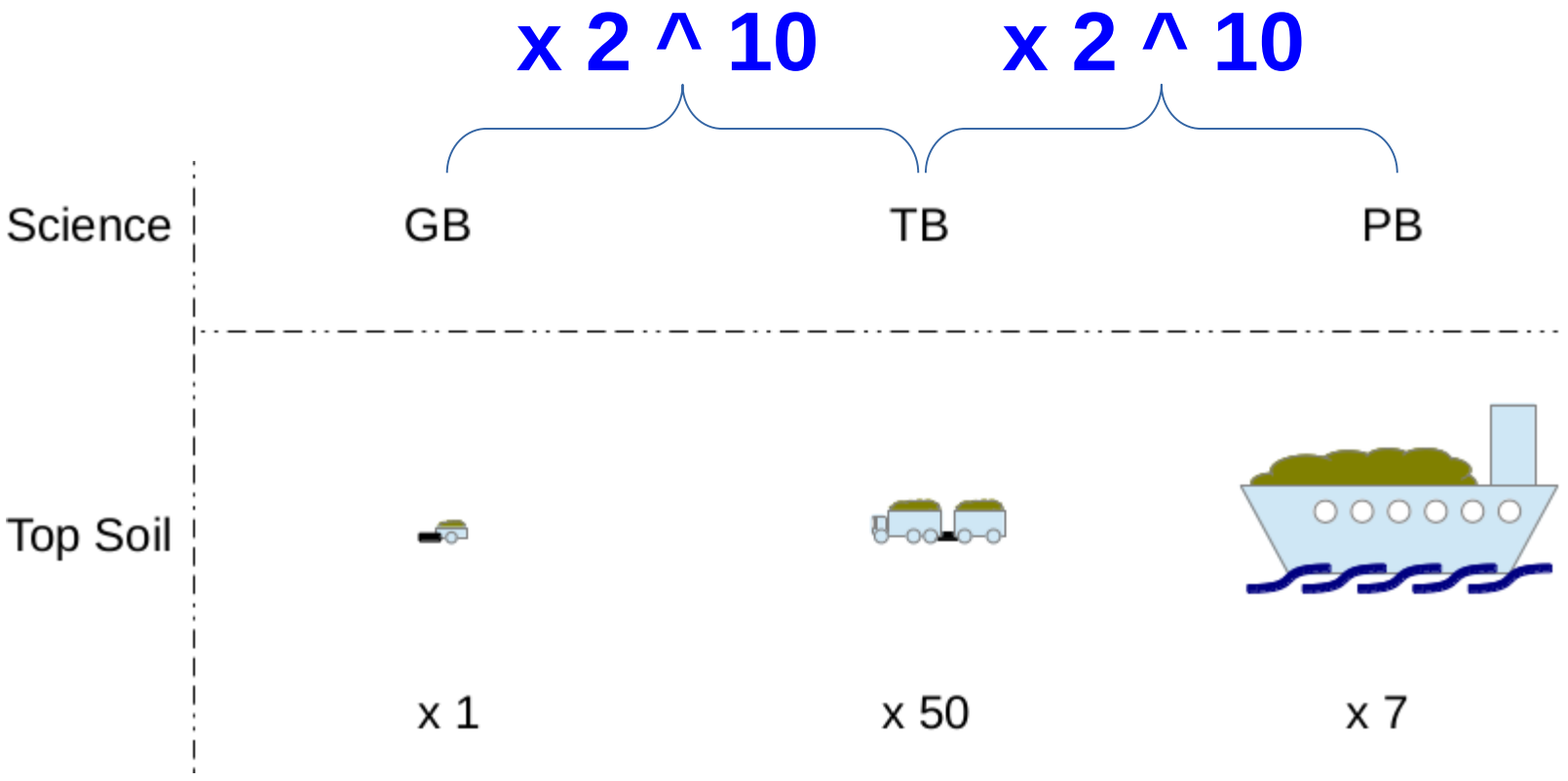
Orders of magnitude?



Well connected 'big data' …

Tight coupling and understanding the scale of data

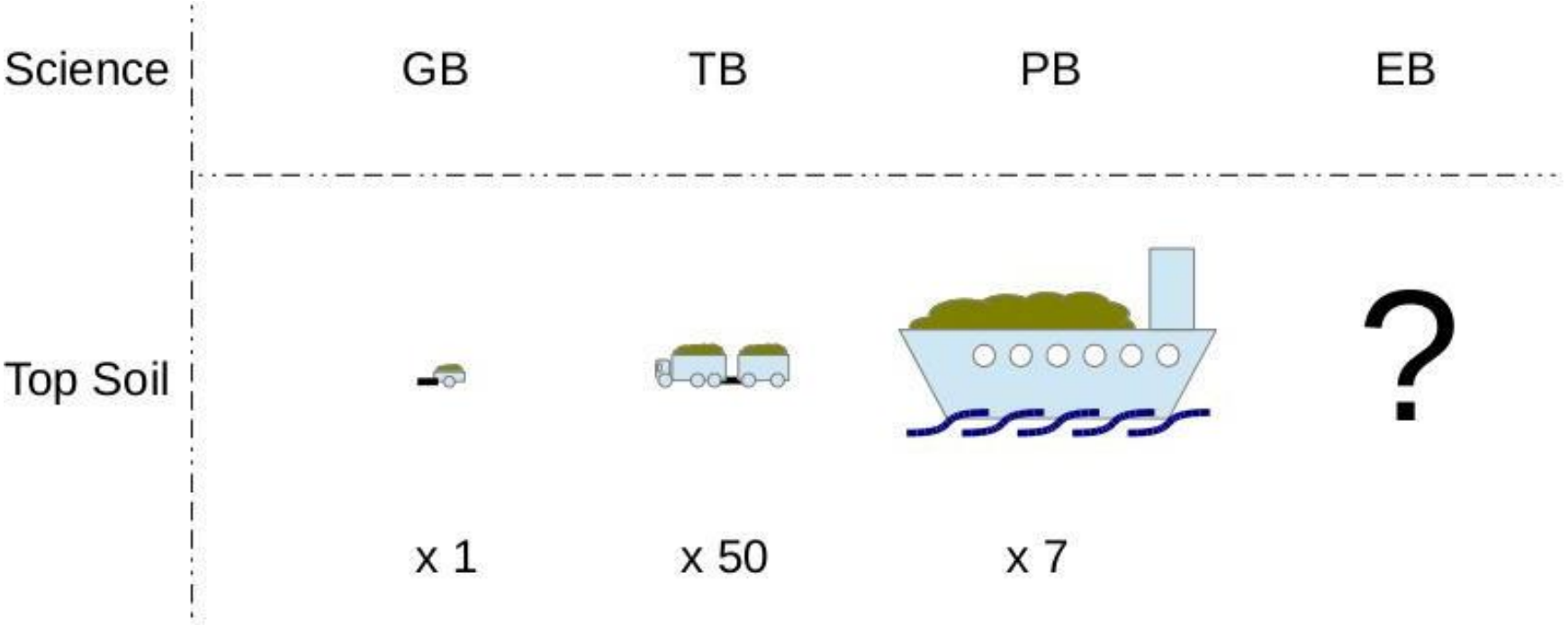
Orders of magnitude?



Well connected 'big data' …

Tight coupling and understanding the scale of data

Of course we are here to discuss Exascale …



'big data' needs to be well connected

We ensured the datasets were 'well connected' through ...

Tight Coupling between:

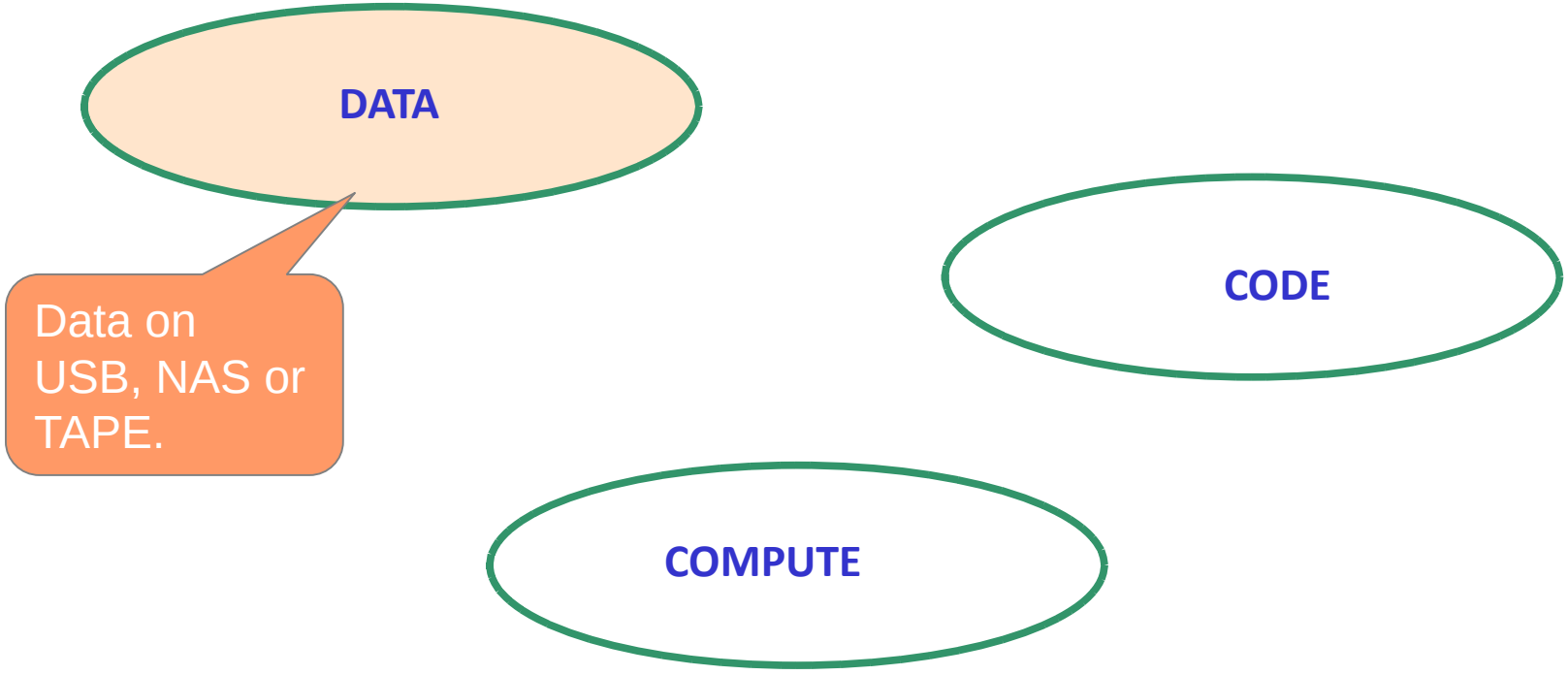
Data,

Code, and

Compute.

Well connected 'big data'
Tight coupling

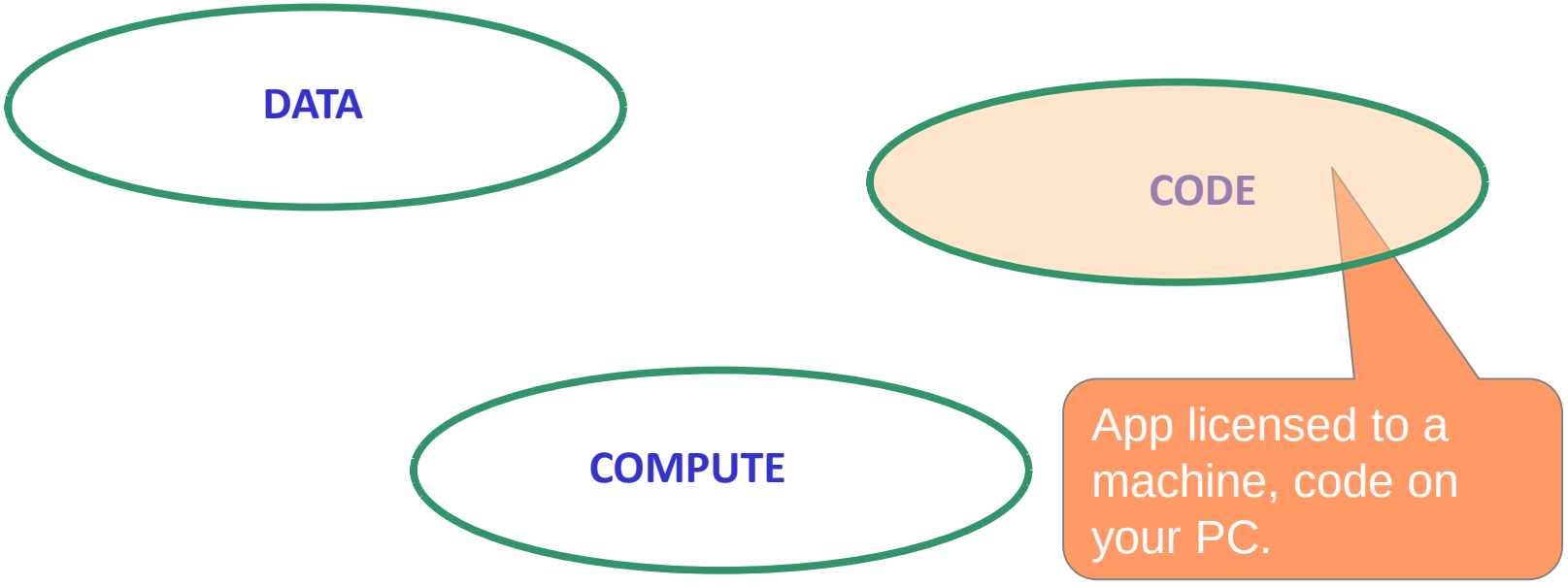
What are the basic connections?



Well connected 'big data'
Tight coupling

Useful analytical tool with:

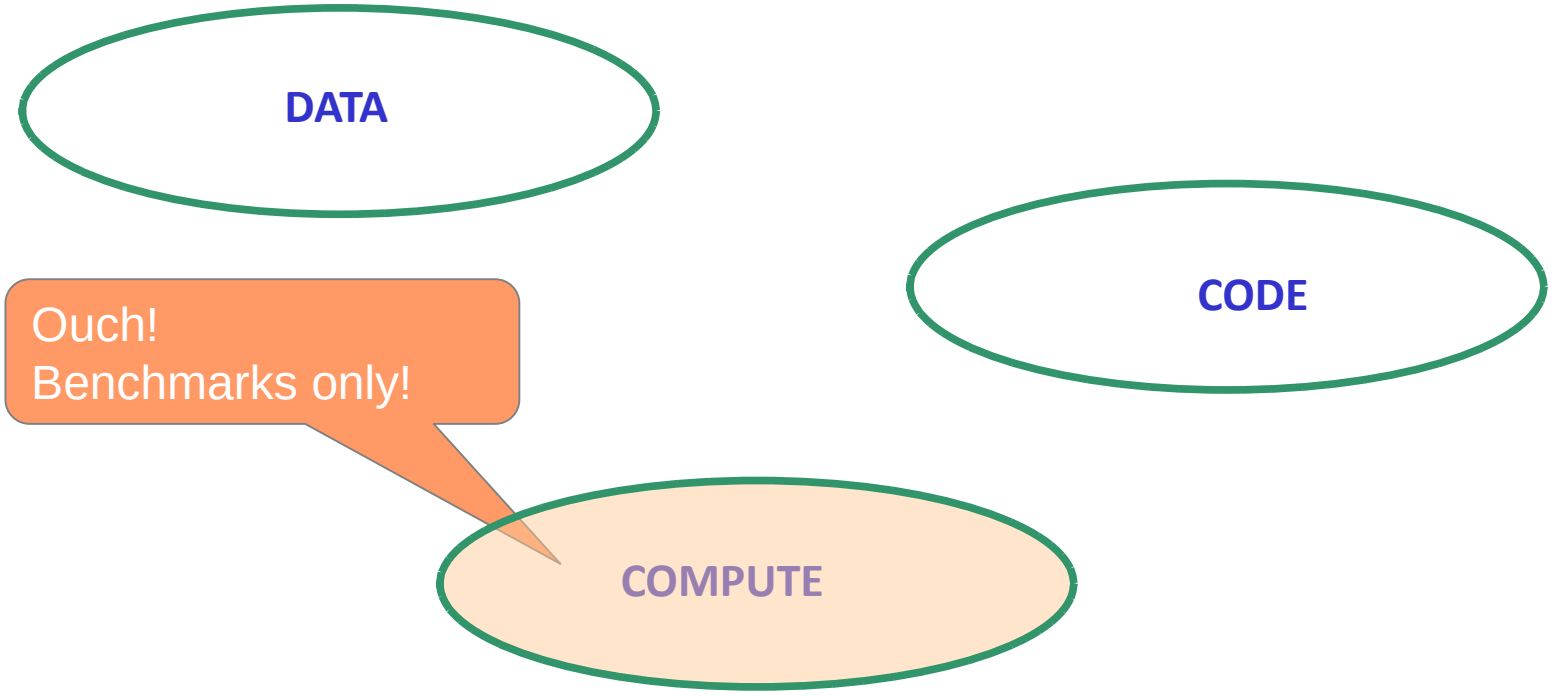
- *no way to run*
- *nothing to analyse*



Well connected 'big data'
Tight coupling

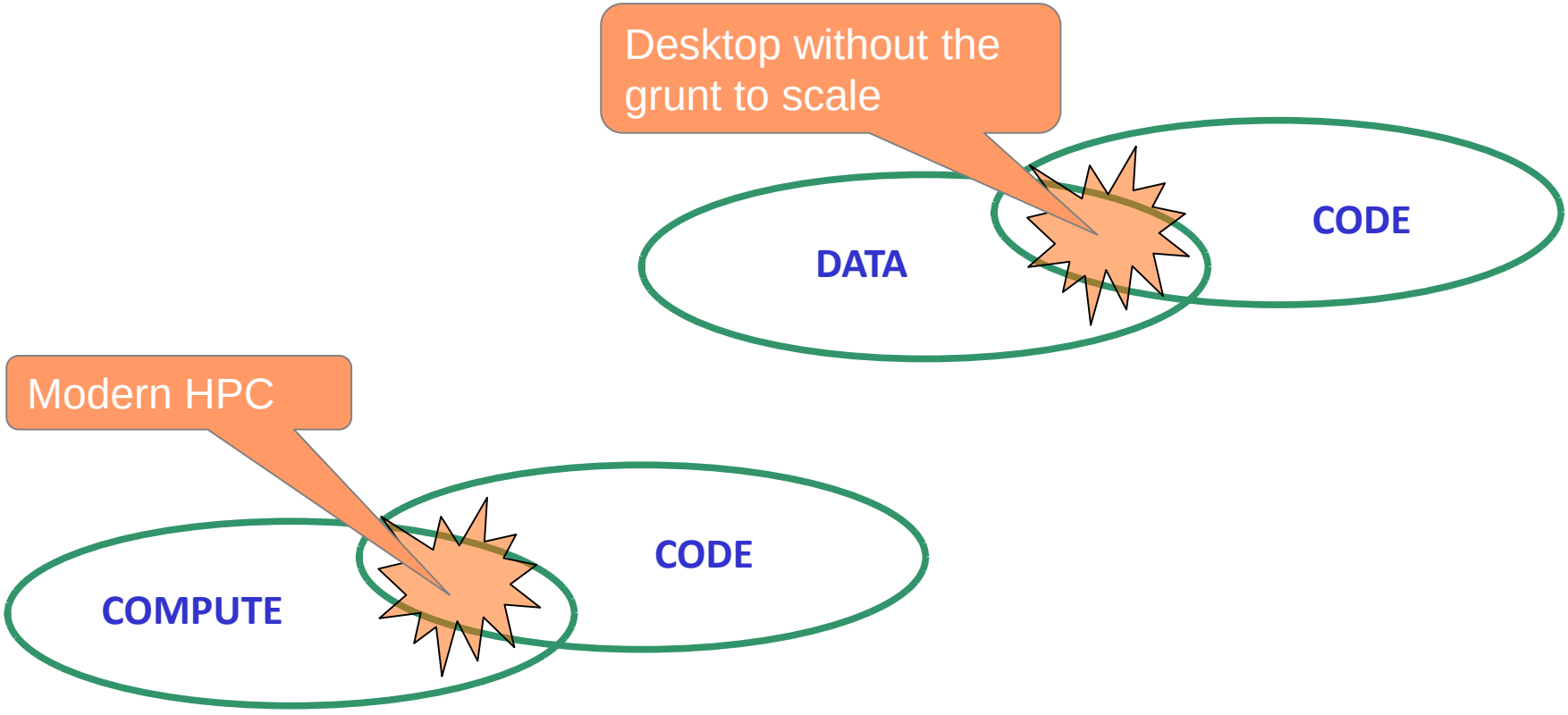
Older HPC:

- *Bring your own: Code, Data, Licenses?*



Well connected 'big data'
Tight coupling

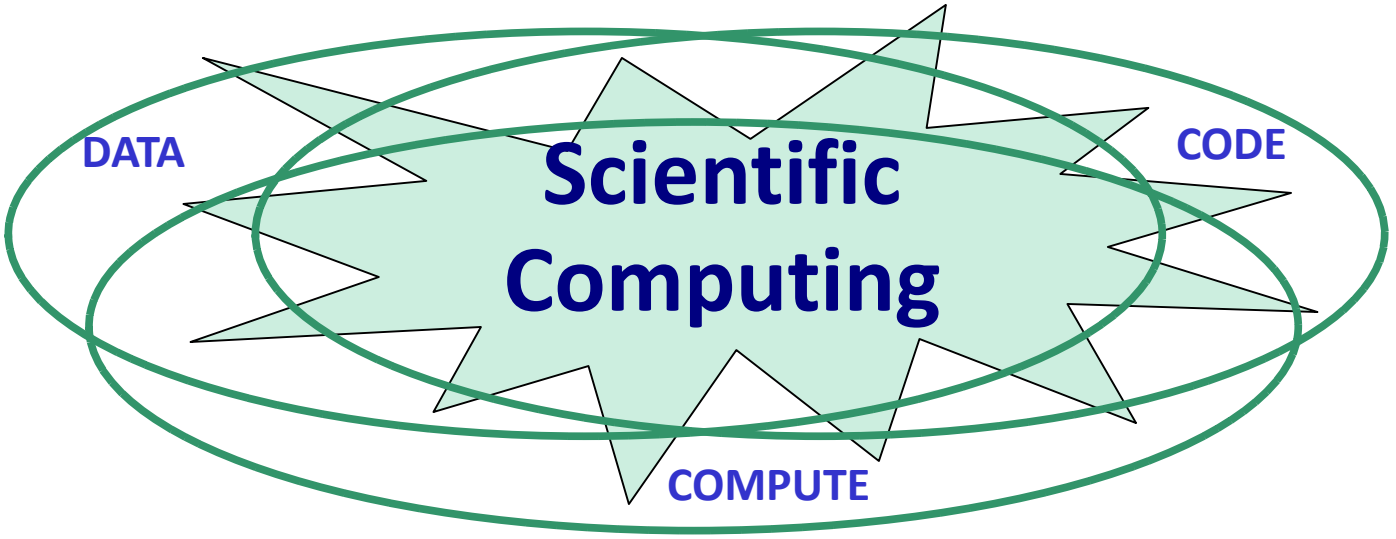
Various other scenarios ...



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



'big data' needs to be well connected ...

We ensured the datasets were 'well connected' through ...

Tight Coupling between:

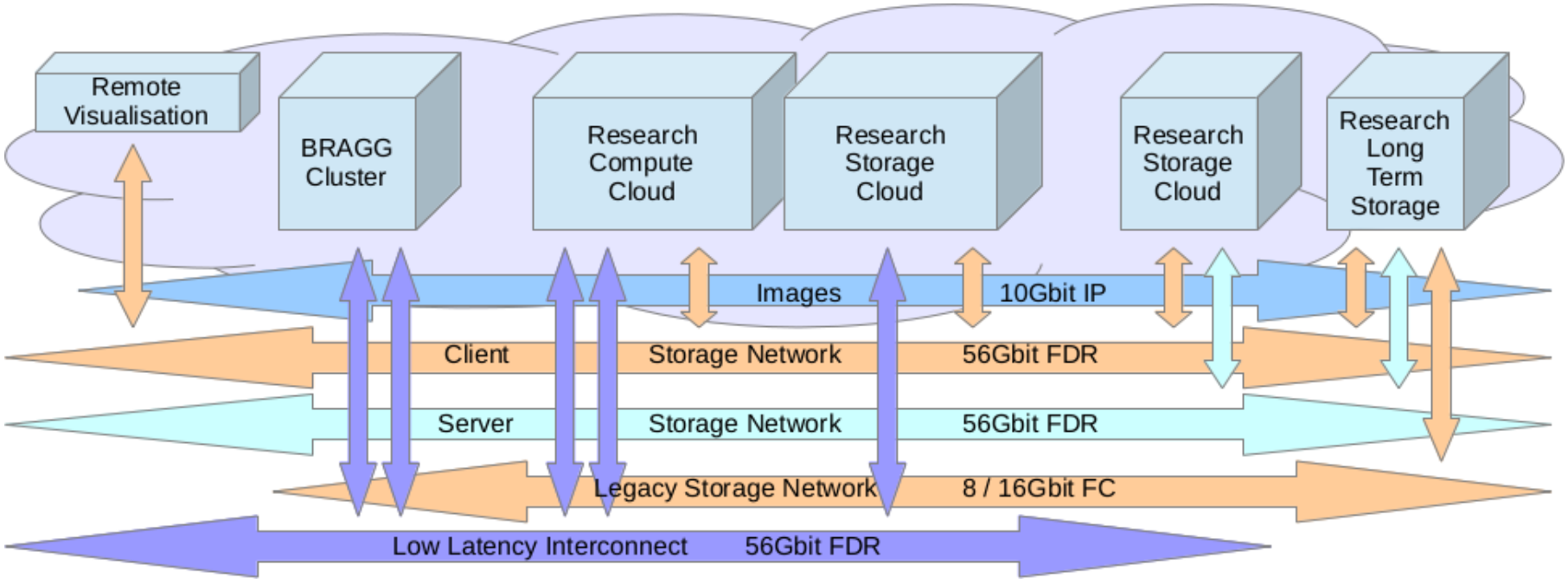
- Data,
- Code, and
- Compute.

Low latency / non-blocking infrastructure.

Its all about the Workflow!

Well connected 'big data' …
Low latency non-blocking infrastructure

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



'big data' needs to be well connected ...

We ensured the datasets were 'well connected' through ...

Tight Coupling between:

- Data,
- Code, and
- Compute.

Low latency / non-blocking infrastructure.

'Fit-for-purpose' provision of infrastructure.

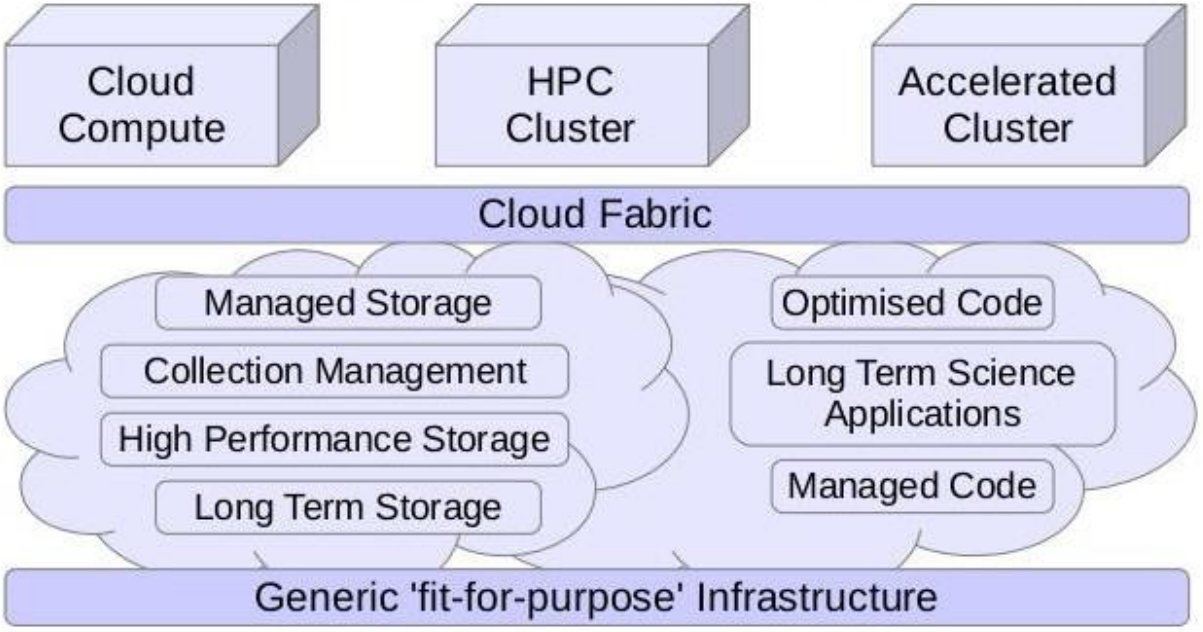
Its all about the Workflow!

Well connected 'big data'
Fit-for-purpose infrastructure

Below the line we split the technology from the workflow and abstracted the brands.

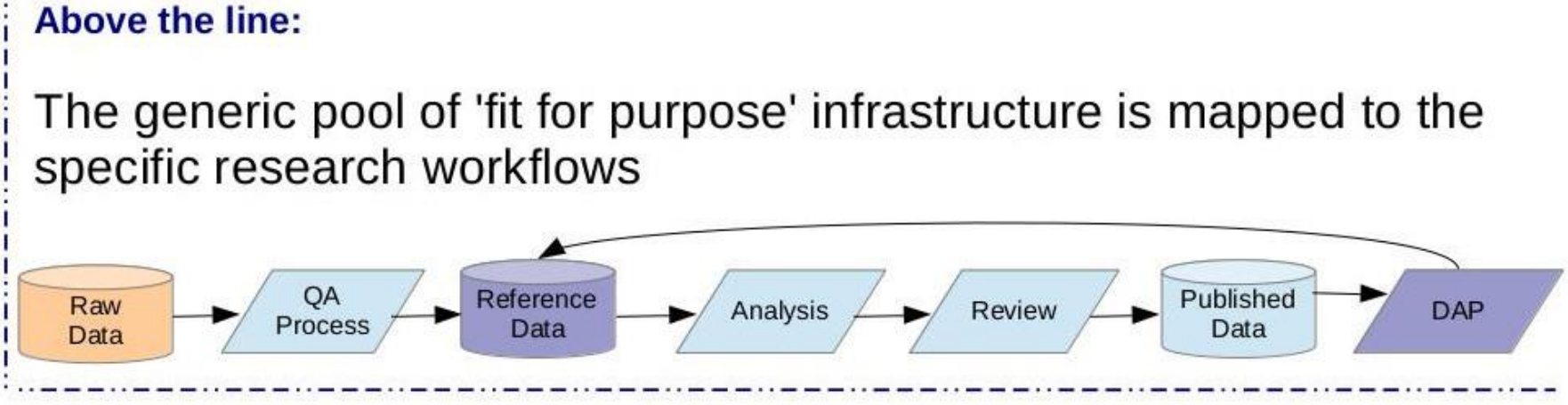
Below the line:

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



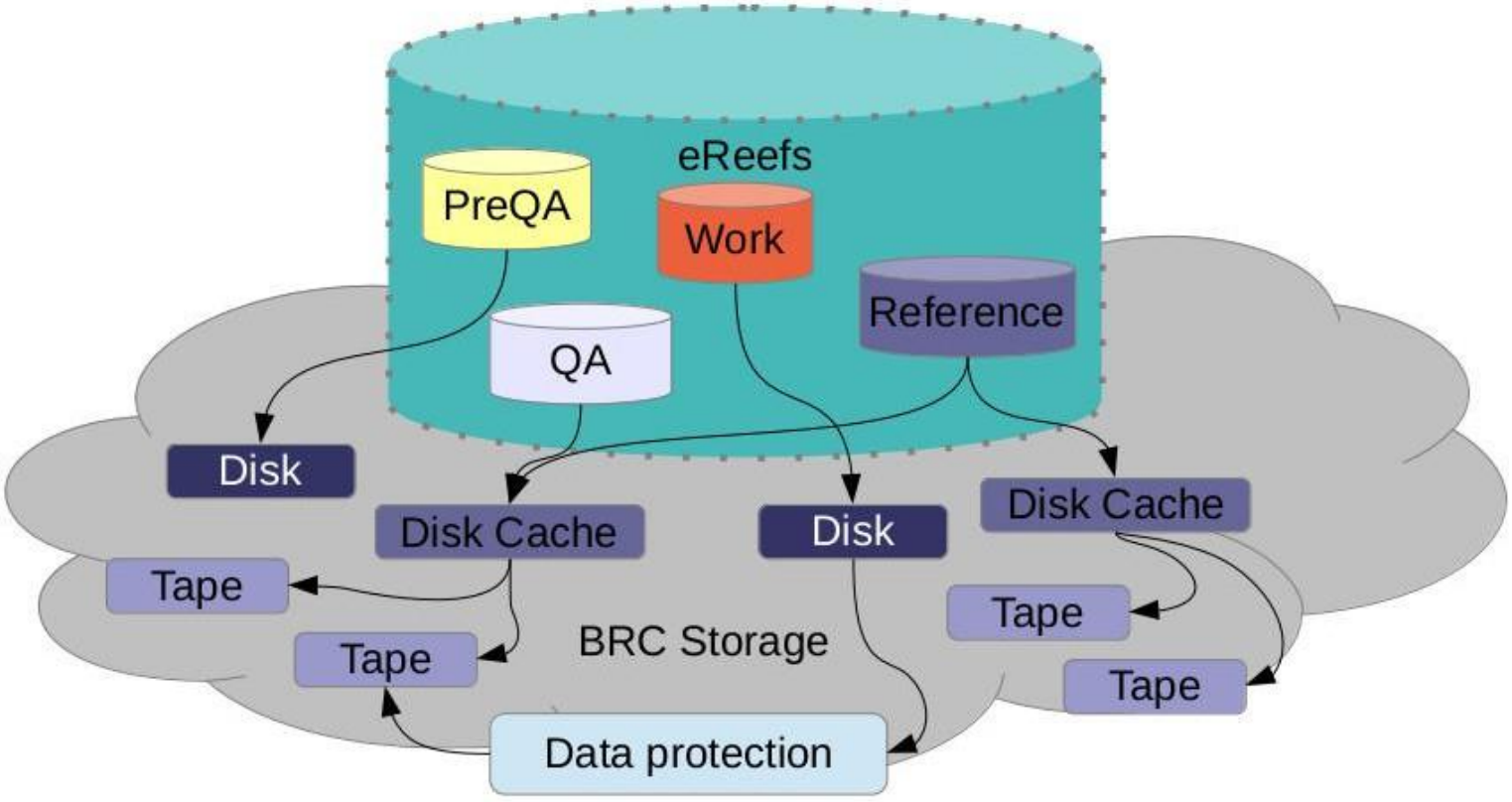
Well connected 'big data'
Fit-for-purpose infrastructure

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



Well connected 'big data'
Fit-for-purpose infrastructure

'On the line' we right-size ...



'big data' needs to be well connected ...

We ensured the datasets were 'well connected' through ...

Tight Coupling between:

- Data,
- Code, and
- Compute.

Low latency / non-blocking infrastructure.

'Fit-for-purpose' provision of infrastructure.

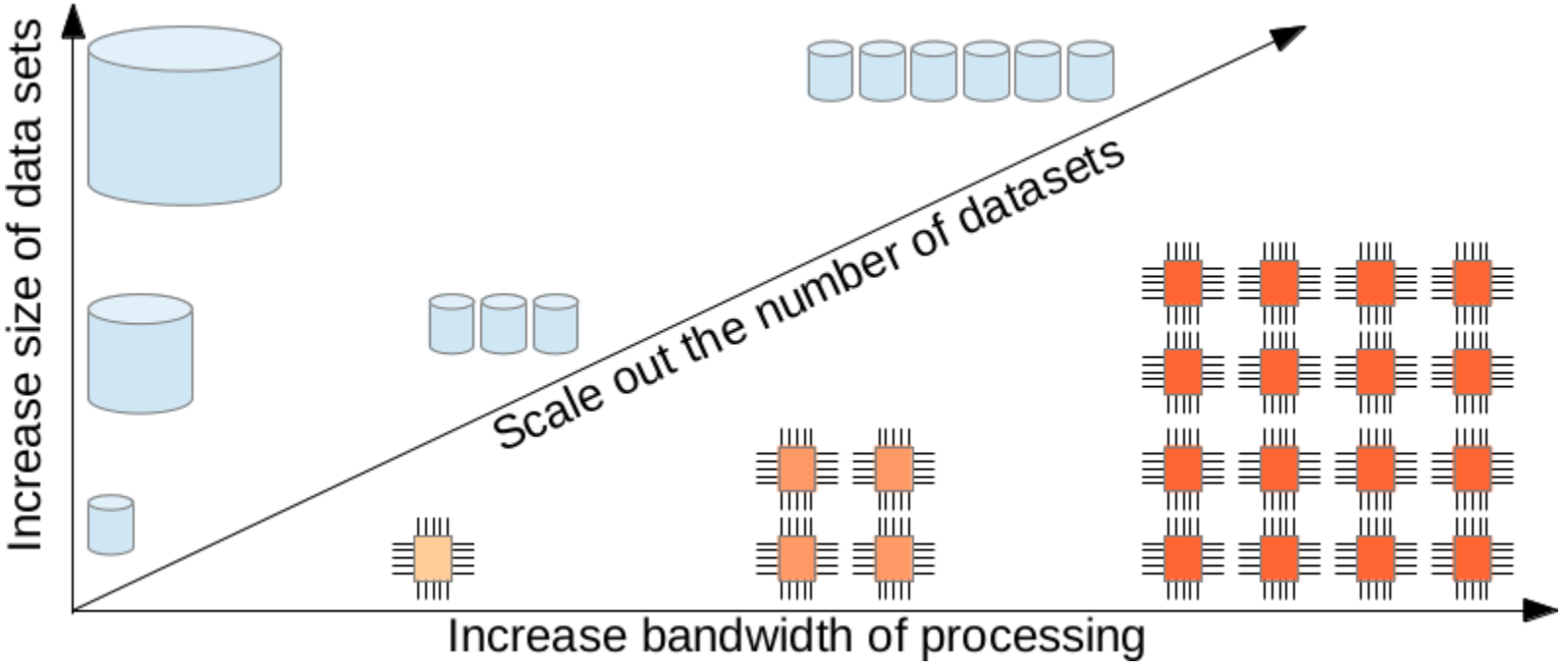
Ability to 'right-size' the provisioned infrastructure:

- Scale Up,
- Scale Down, and or
- Scale Out.

Its all about the Workflow!

Well connected 'big data'
Integrated ability to right size

Integrated ability to right-size infrastructure to meet changing demands

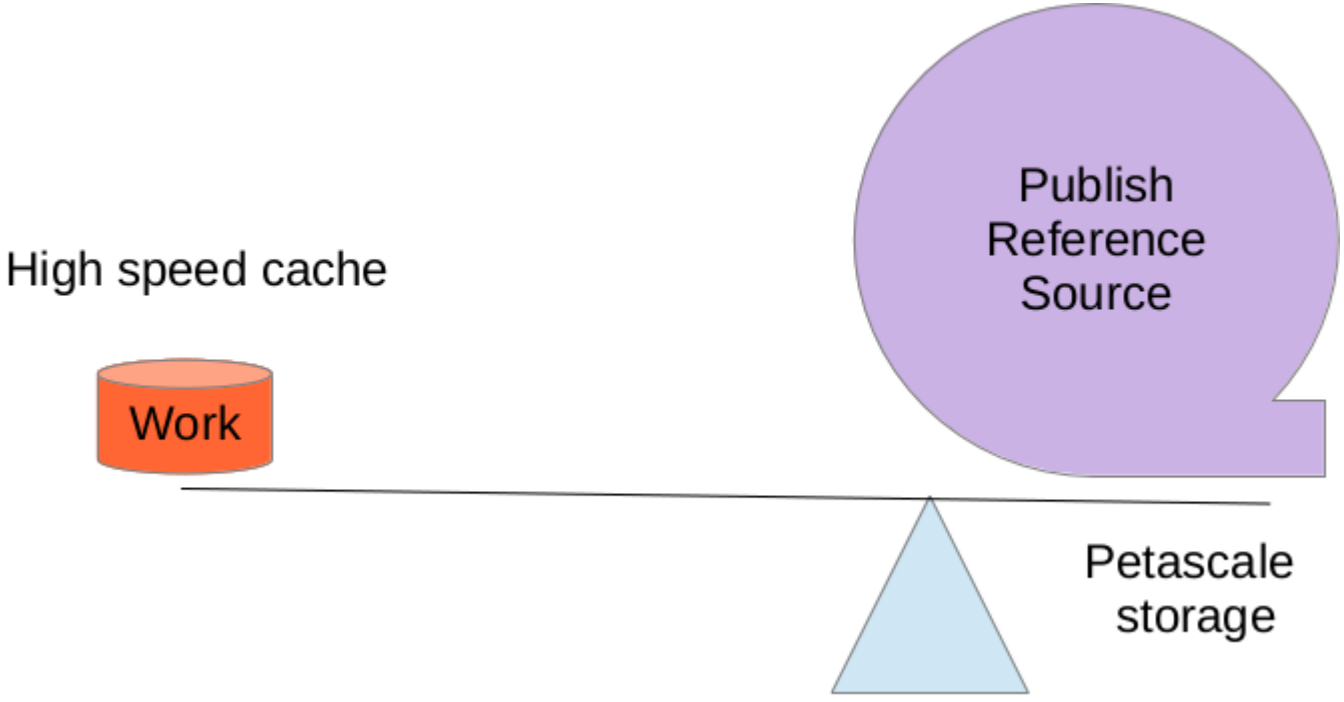


Data Intensive Research

So where does that leave us?

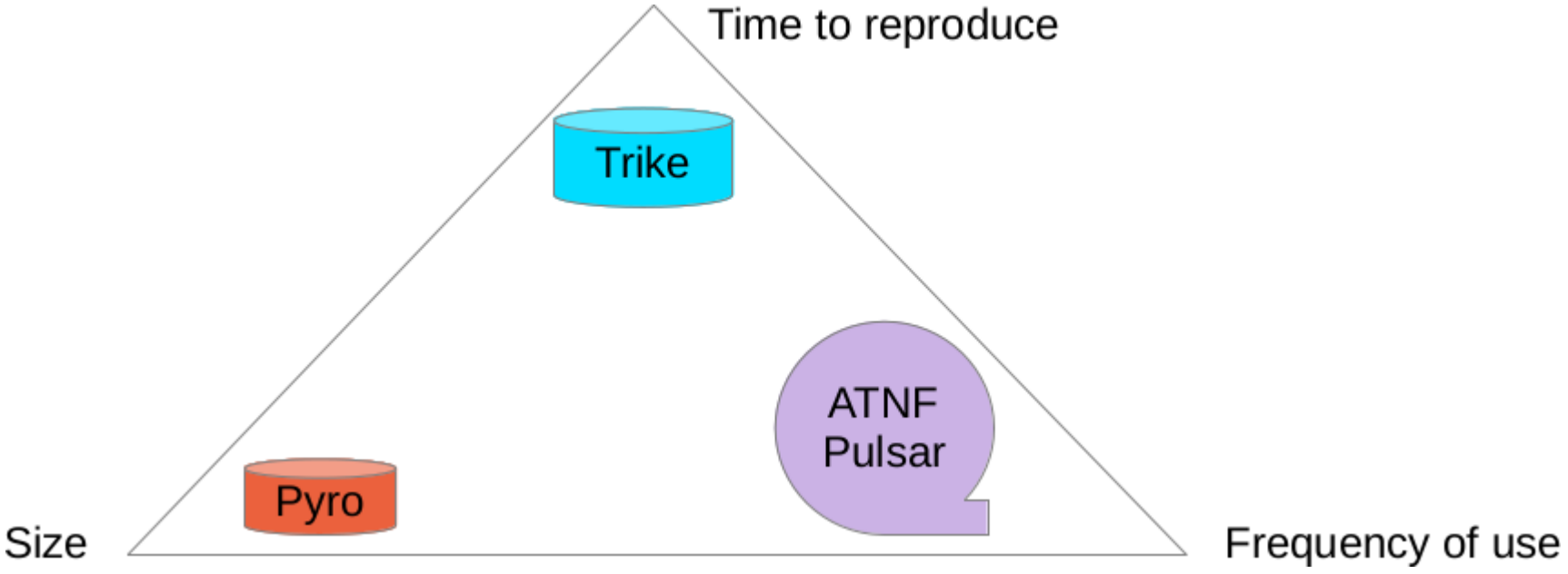
Can we transition from having 'lots of data' into 'big data'

while 'reducing cost' per PB?



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

while 'improving data management practices?'



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

while 'accelerating workflows?'

"STACC team have delivered storage and analysed our requirements for computing (shortly to be delivered) that has enabled us to retire significant legacy hardware, including a home-grown cluster. This solid foundational infrastructure means we are better able to deliver reliably for clients and, we now spend more time on core tasks rather than worrying about IT upkeep, admin and maintenance. Their ability to prioritise our request and streamline delivery when one of our key legacy servers went bad during the pre-provisioning phase is very much appreciated."

(Edward King, CMAR, Hobart, Tasmania)

"…This reduced our processing time per scene to about 7 1/2 minutes. Thus the total processing time for the job was reduced from 15 days to less than 24 hours. Also we reduced the I/O traffic to the data service 95%, eliminating any concerns about degrading I/O latency for other users."

(Drew Deveraux, CCI, Floreat, Western Australia)

"…. I don't understand the reasons, but doubling the CPUs has made a big difference – on CES-07 for example, an ARCGIS process that took about 3 days previously ran overnight!"

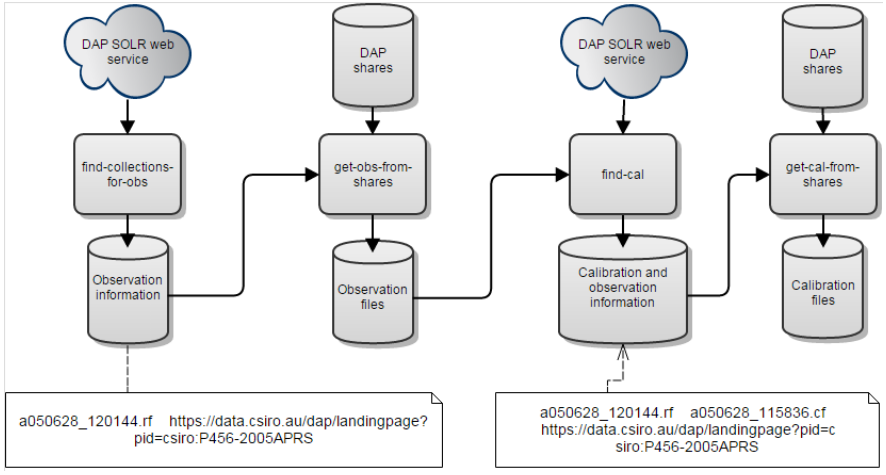
(Kristen Williams, CES, Black Mountain, Australian Capital Territory)

"….STACC provided the streamflow forecasting team a secure, reliable storage system that has become vital to our research. It's made the secure storage of large volumes of data, and the sharing of these data, laughably easy."

(James Bennett, CLW, Highett, Victoria)

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

while 'opening up new workflow possibilities'



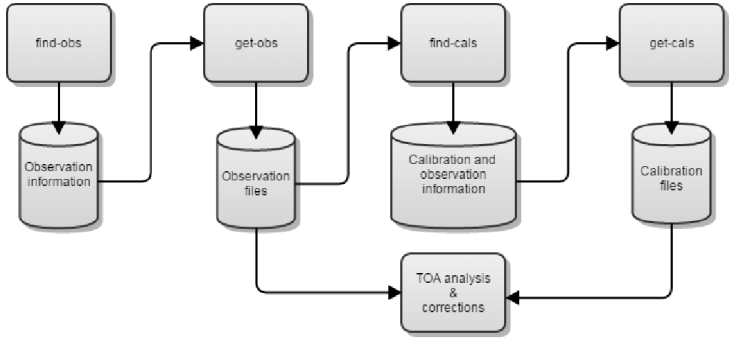


Figure 3: Example processing workflow

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

while 'opening up new workflow possibilities'

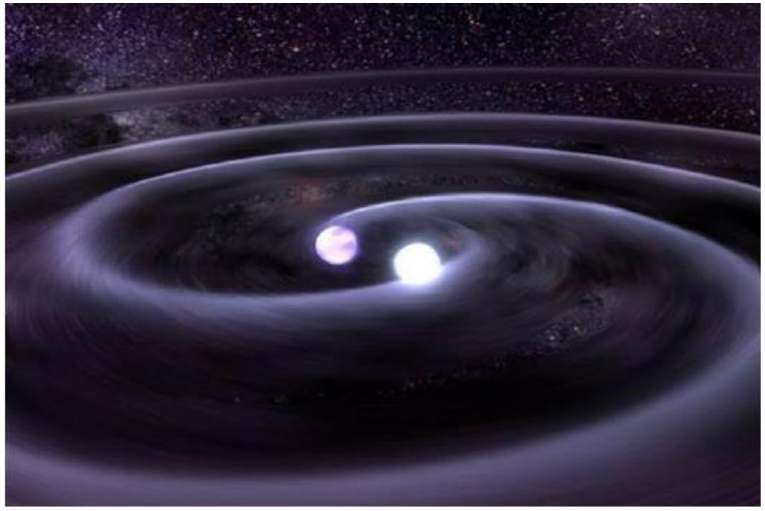


Figure 4: Binary system (black hole/neutron star pair).

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

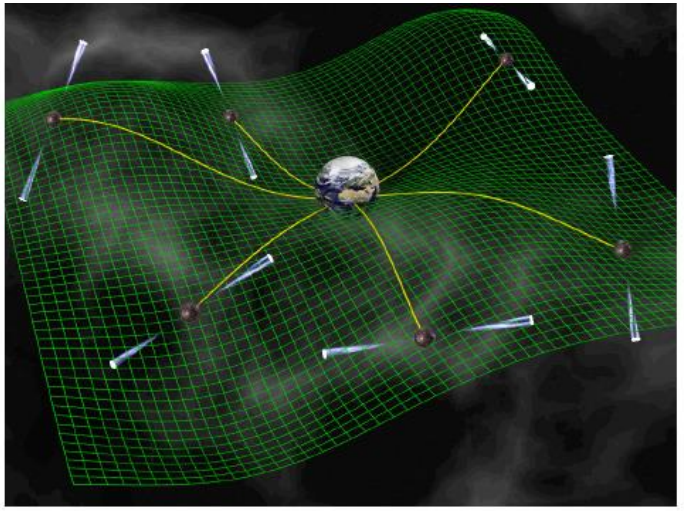


Figure 5: Space-time warping from gravitational waves causes Earth to "move" with respect to pulsars, delaying or advancing pulse arrival times. Courtesy: David Champion

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

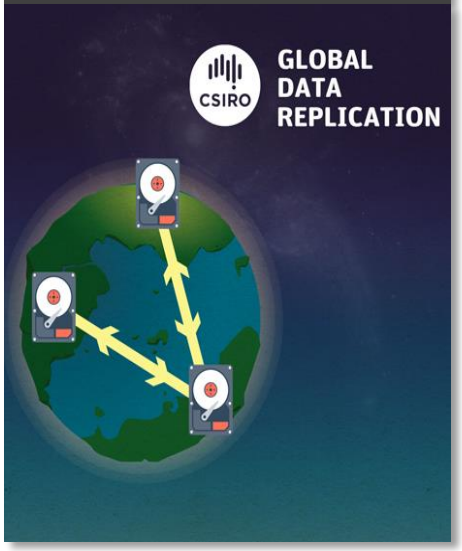
BOWEN Research Cloud (previously STACC)

"Over the past 3 years, the STACC project has enabled a significant shift to what 'Cloud Computing' really means in the context of the research workflow. It is very pleasing to see so much adoption of the capabilities across all domains of CSIRO research, which have been developed by IM&T throughout the project. It is even more exciting to see the capability extend into the broader future objectives - as a long term service supported by IM&T that will enable the connectedness and acceleration of research workflows, which will no doubt lead to big impact science outcomes!"

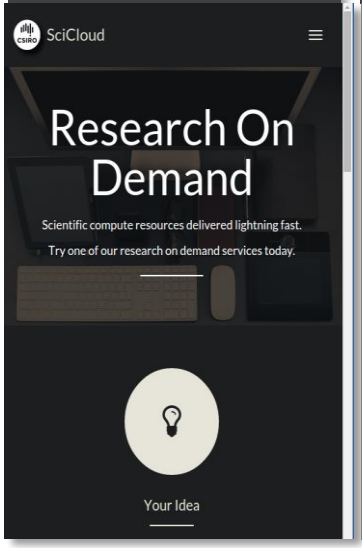
(Angus Vickery – Deputy CIO, Scientific Computing Group, CSIRO)

CSIRO IM&T Strategy 2016 - 2020

Global Data Mirrors



Virtual Labs, Sci-Cloud



National Digital Hub

2016

27PB

Petabyte Age

2020

412PB

Exabyte Age

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

What good is 'big data' if …

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 cannot work in harmony?
If it doesn't speak to the world?

Thank you