genomics
VIRTUAL LAB

The Genomics Virtual Laboratory

Andrew Lonie
Victorian Life Sciences Computation Initiative, University of Melbourne

What is the Genomics Virtual Lab?

NeCTAR **funded nationally distributed platform for** ***genomics*****, built on the** ***Research Cloud*** **and** ***RDSI***

What is the Genomics Virtual Lab?

NeCTAR **funded nationally distributed platform for** ***genomics*****, built on the** ***Research Cloud*** **and** ***RDSI***

1. **Compute & workflow platforms**
2. **Tutorials, protocols**
3. **UCSC browser, datasets**

http://genome.edu.au

RDSI
Research Data Storage Infrastructure

nectar
http://nectar.org.au

What is the Genomics Virtual Lab?

In practice:

1. ***A way to build on-demand analysis environments for genomics***

2. ***A set of prebuilt analysis and visualisation servers for tutorials and general use***

3. ***Resources to teach genomics, and regular workshops using them***

4. ***Development work exploring new capabilities***

What is the Genomics Virtual Lab?

1. ***Build on-demand***

2. ***Prebuilt servers for general use***

3. ***Tutorials and workshops***

4. ***R & D***

2. Preconfigured GVL servers

http://genome.edu.au → USE

Galaxy-tut *http://galaxy-tut.genome.edu.au*

- For GVL Galaxy-based tutorials
- Has tools, datasets and Galaxy histories pre-installed

Galaxy-Qld *http://galaxy-qld.genome.edu.au*

- For (Galaxy-based) research use
- Has lots of disk space, scalable as required

UCSC Browser *http://ucsc.genome.edu.au*

- Local mirror of US-based browser - quicker to upload your tracks, etc

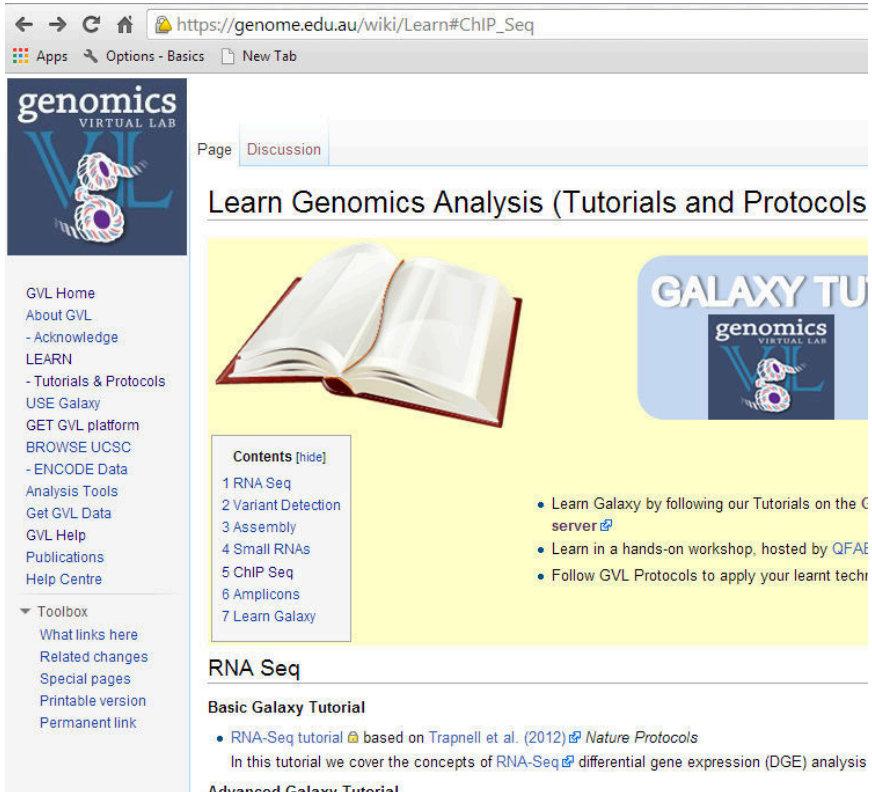
3. Resources and tutorials

http://genome.edu.au → LEARN

Tutorials and Protocols based on Galaxy and command line

- ***RNA-Seq DGE***
- ***Variant detection***
- ***small RNA***
- ***Microbial genome assembly***
- ***others***

Some of these are based on GVL-developed tools



1. Build an analysis environment

http://genome.edu.au → GET

Why would you want to build?

Flexibility

- **Configure tools, ref data**
- **Command line interface**

Availability

Scalability

Privacy

Most importantly, maybe: Control

GET a GVL

http://genome.edu.au → **GET**

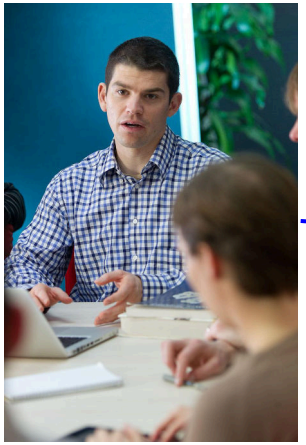
Building (deploying and running) a GVL instance:

1. *Create a CloudBioLinux server VM*
2. *Download and install a preconfigured Galaxy*
3. *Attach pre-populated indexed genomes data*
4. *Start Galaxy*
5. *Add extra compute nodes as required*

Deploying and running a GVL

http://launch.genome.edu.au

Cloudman = Middleware for building, distributing and managing cloud-based platforms, especially Galaxy



Afgan *et al. BMC Bioinformatics* 2012, **13**:315
http://www.biomedcentral.com/1471-2105/13/315

BMC Bioinformatics

SOFTWARE Open Access

CloudMan as a platform for tool, data, and analysis distribution

Enis Afgan[1,3,4*], Brad Chapman[2] and James Taylor[3,4]

Abstract

Background: Cloud computing provides an infrastructure that facilitates large scale computational analysis in a scalable, democratized fashion, However, in this context it is difficult to ensure sharing of an analysis environment and associated data in a scalable and precisely reproducible way.

Results: CloudMan (usecloudman.org) enables individual researchers to easily deploy, customize, and share their entire cloud analysis environment, including data, tools, and configurations.

Conclusions: With the enabled customization and sharing of instances, CloudMan can be used as a platform for collaboration. The presented solution improves accessibility of cloud resources, tools, and data to the level of an

GET what GVL?

	Personal GVL	Server GVL	Cluster GVL
Suitable for	**Single user**	Single user Small group/lab	Large groups Institutions
Storage	**60GB**	100-5000GB	TBs
Compute	**2 cores**	8-64* cores	>50 cores
Requires	**NeCTAR account**	NeCTAR allocation: Compute and Volume storage	Large NeCTAR allocation of compute + user-provided fast storage
Runs on	**Any Research Cloud node**	RC nodes with volumes	RC nodes co-located with fast file system
Setup	**Automatic via website**	Automatic via website	Collaboration with GVL team
Configuration	**No configuration required**	Some configuration to tune analyses	Dedicated management



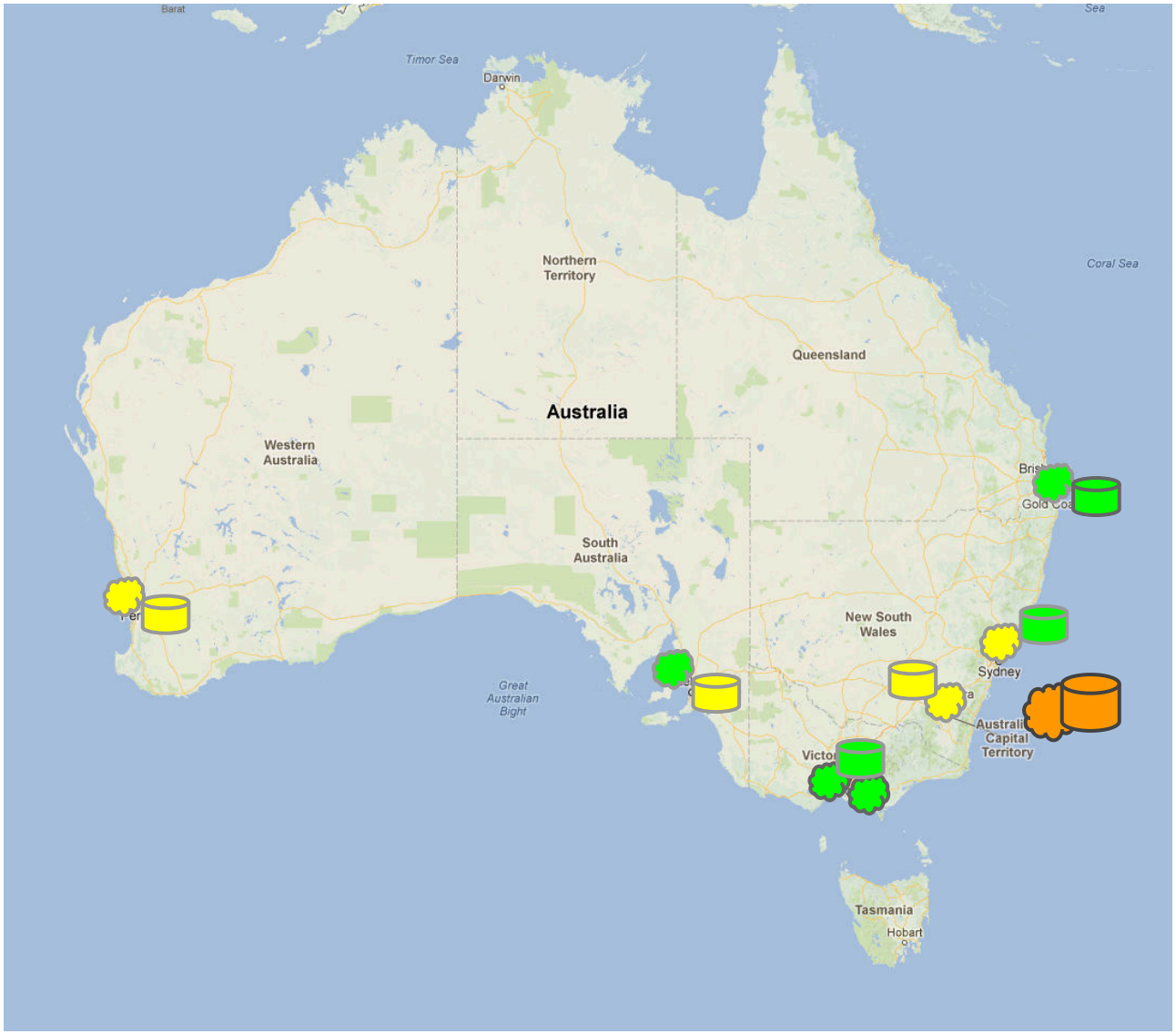
	Personal GVL	Server GVL	Cluster GVL
Suitable for	**Single user**	Single user Small group/lab	Large groups Institutions
Storage	**60GB**	100-5000GB	TBs
Compute	**2 cores**	8-64* cores	>50 cores
Requires	**NeCTAR account**	NeCTAR allocation: Compute and Volume storage	Large NeCTAR allocation of compute + user-provided fast storage
Runs on	**Any Research Cloud node**	RC nodes with volumes	RC nodes co-located with fast file system
Setup	**Automatic via website**	**Automatic via website**	Collaboration with GVL team
Configuration	**No configuration required**	Some configuration to tune analyses	Dedicated management

Initiatives: NeCTAR & RDSI

Research Cloud node

RDSI node

Coming 2014-15

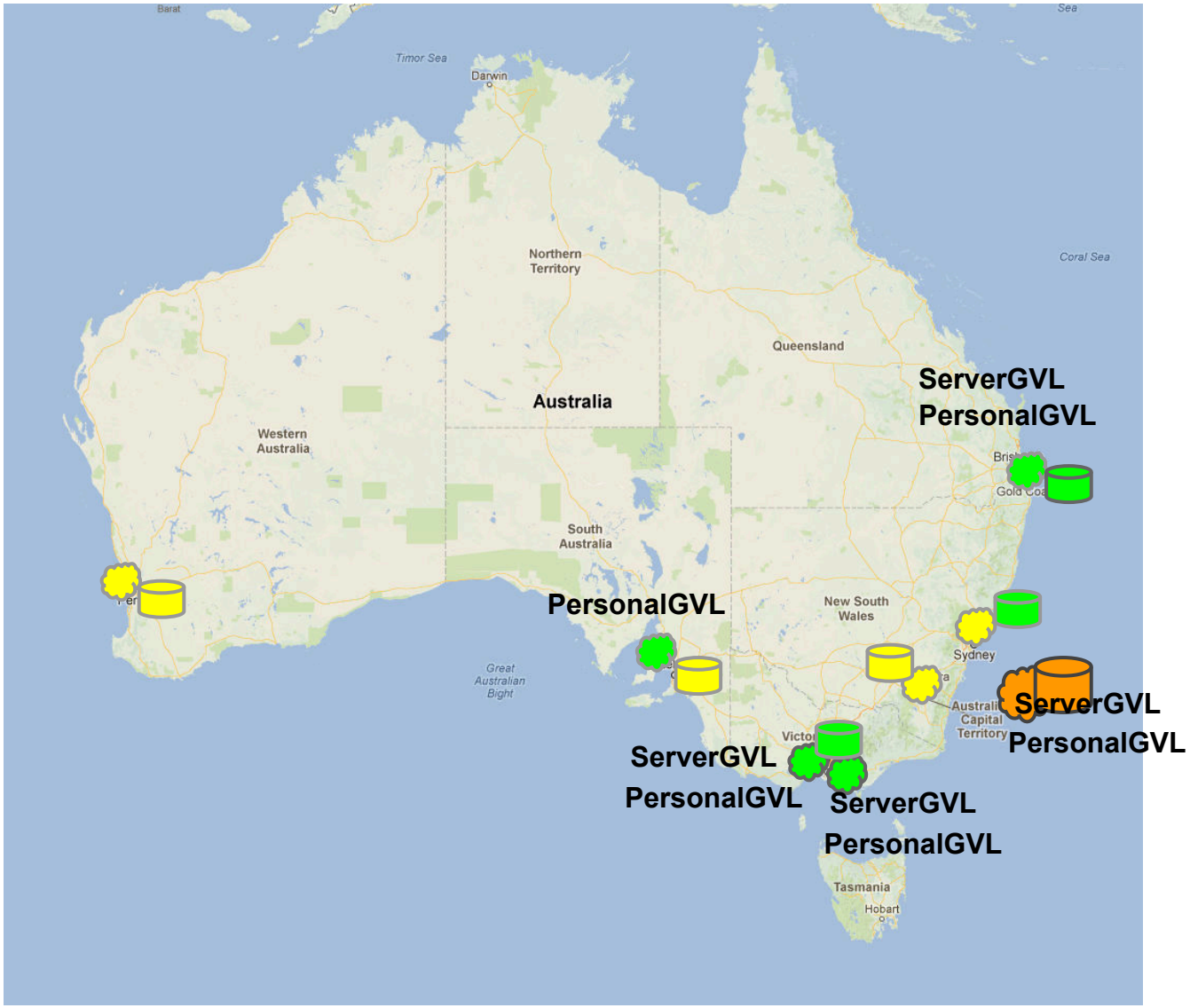


Initiatives: Genomics Virtual Lab

Research Cloud node

RDSI node

Coming 2014-15



4. Research and Development

http://genome.edu.au

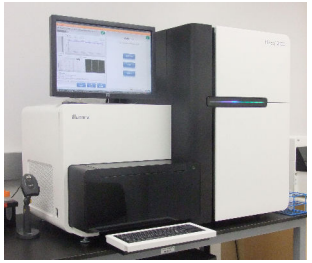
What's next for GVL?

Genomics is characterised by:

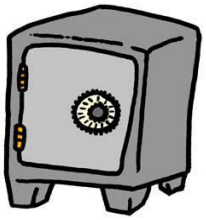
- **Large data**
- **Data parallel computation**
- **Moving best practice, tools**

Case Study: Human exome analysis

Store SECURELY, LONG TERM



Fast-ish, 'permanent', reliable

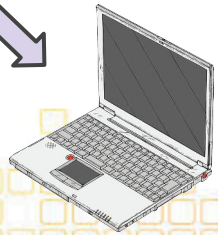


Fast

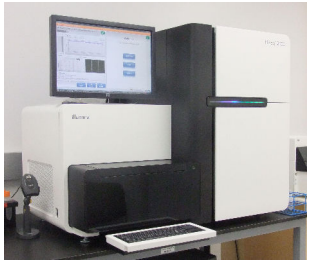
Data 'home'

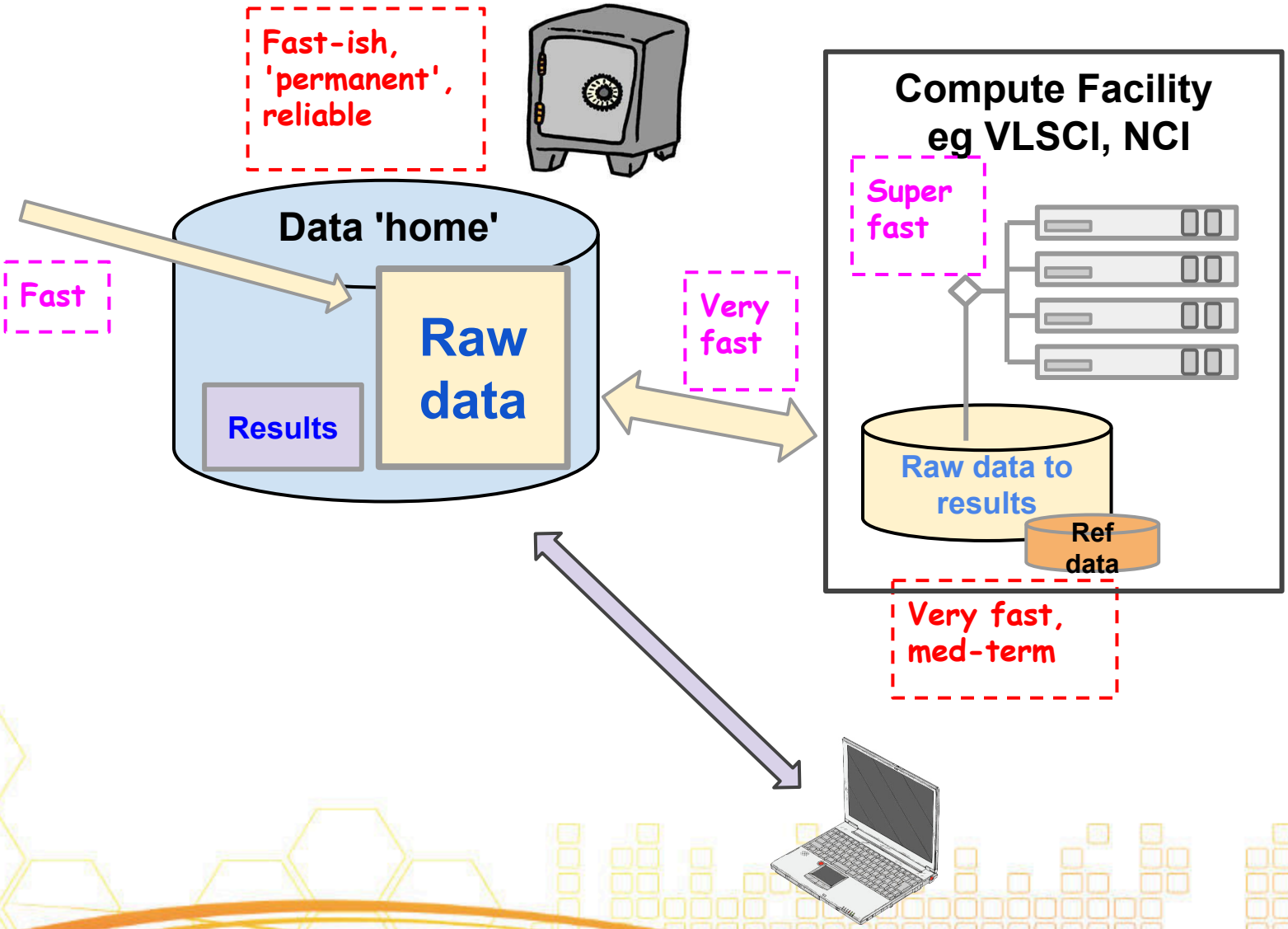
Raw data

Results

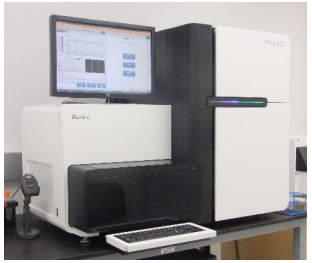


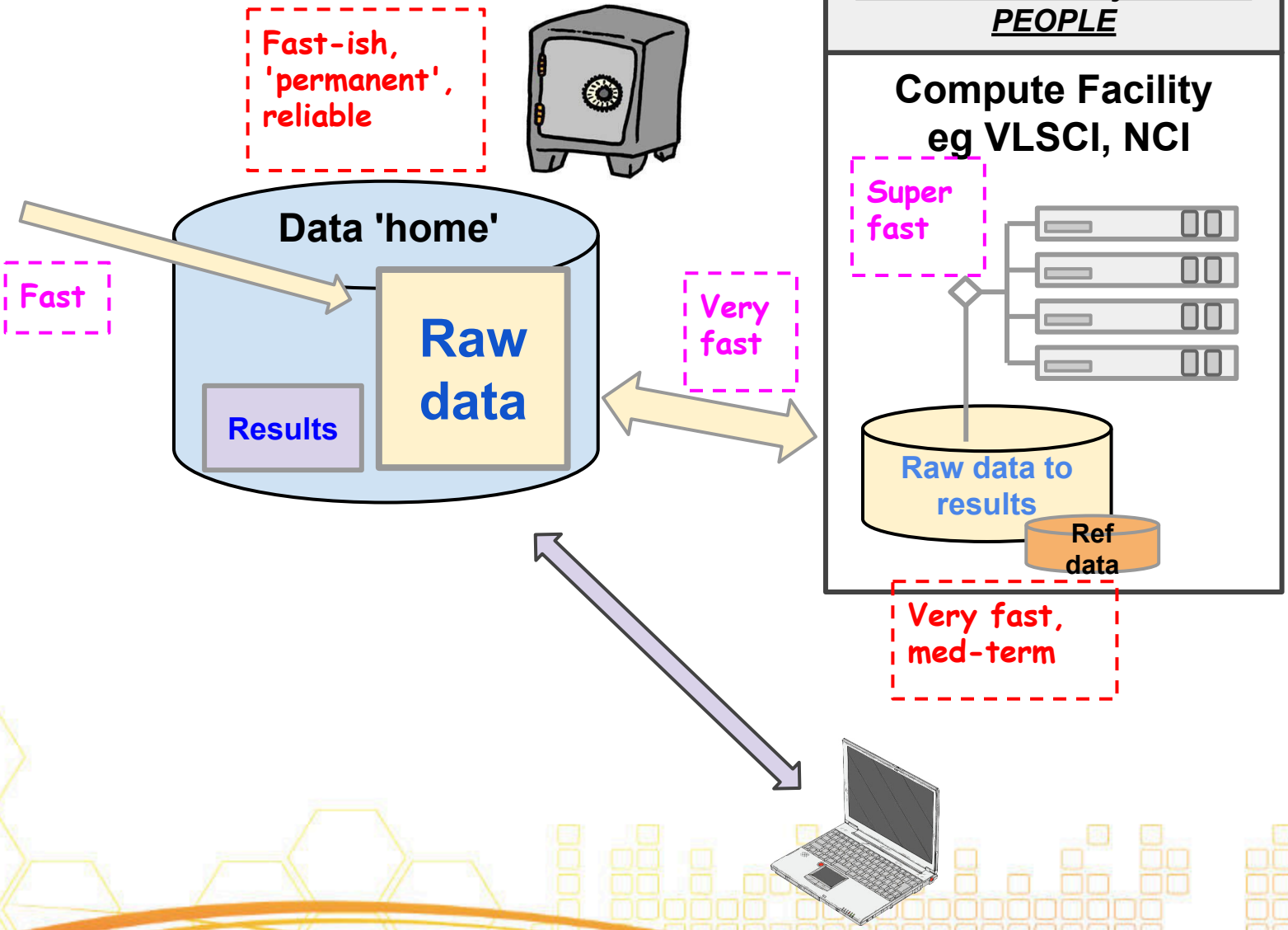
Analyse QUICKLY



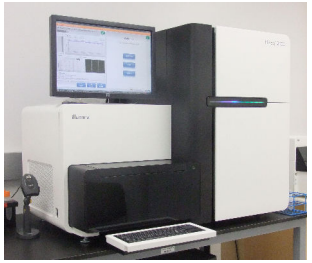


Analyse REPRODUCIBLY

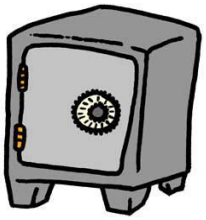




Share PUBLICLY, PRIVATELY



Fast-ish, 'permanent', reliable



Fast

Data 'home'

Raw data

Results

Very fast

Workflows, Galaxy, Ruffus, PEOPLE

Compute Facility eg VLSCI, NCI

Super fast

Raw data to results

Ref data

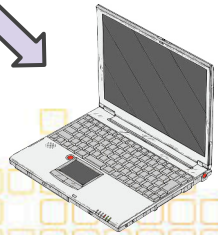
Very fast, med-term

Med

Publishing, dissemination

Web-based, read-only, publishable

Slow



Recommendations

1. Separate 'data' and 'compute'
   - physically and conceptually

2. Have a data 'home'
   - Secure, Long term

3. Have a HPC compute platform, well managed
   - High speed linked to 'Data Home'

4. Find a data sharing and publishing option

Don't buy/manage your own if you can avoid it!

GVL 2015-

Wide user base

- students, researchers, tool developers

Broadening from genomics:

- Proteomics GVL, Metabolomics GVL

Platform for tools development and distribution

Training sessions at least fortnightly, closer ties with Galaxy folk internationally

Entire data-compute cycle on the cloud

Making the GVL possible

Go8 Universities

- The University of Queensland
- The University of Melbourne
- Monash University
- The University of Sydney
- The University of Western Australia

Medical Research Institutes

- The Garvan Institute of Medical Research
- Victor Chang Cardiac Research Institute
- Baker IDI Heart and Diabeties Institute
- Peter MacCallum Cancer Centre

eResearch Agencies

- Queensland Facility for Advanced Bioinformatics (QFAB)
- Queensland Cyber Infrastructure Foundation (QCIF)
- Life Sciences Computation Centre (LSCC) at the VLSCI
- Victorian eResearch Strategic Initiative (VeRSI)

National Agencies

- NeCTAR, DIISRTE
- CSIRO
- EMBL Australia
- Bioplatforms Australia (BPA)
- Australian Genome Research Facility (AGRF)
- Australian National Data Service (ANDS)