The Genomics Virtual Laboratory

Andrew Lonie
Victorian Life Sciences Computation Initiative
University of Melbourne

What is the Genomics Virtual Lab?

Nationally distributed platform for *genomics*, built on the federal *Research Cloud*

RDSI
Research Data Storage Infrastructure

nectar

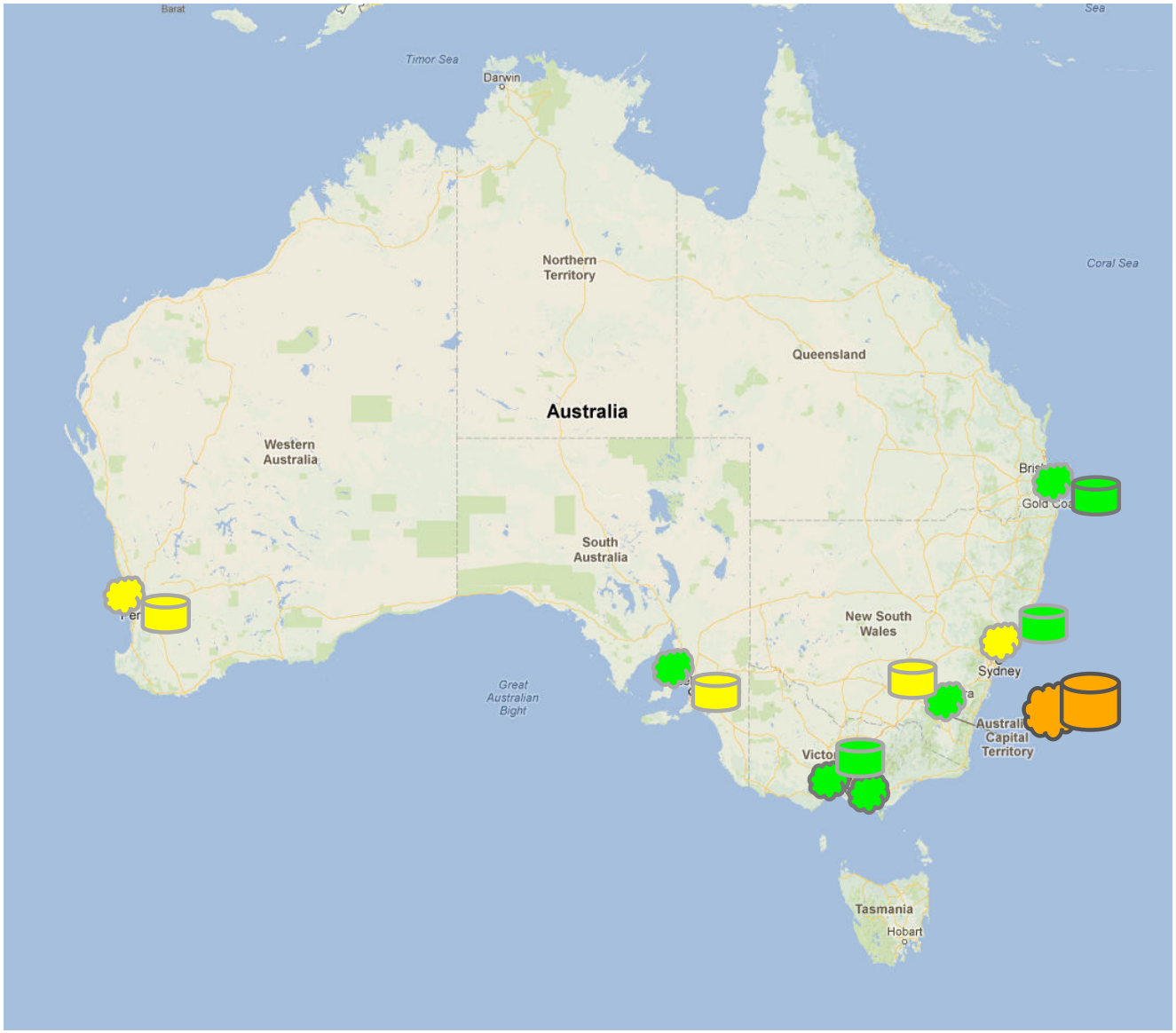
http://nectar.org.au

The Australian Research Cloud

Cloud node:
3-6000 cores

Data node:
1-5 PB

Coming 2014-15



GVL: Drivers

To provide a genomics analysis platform with:

1. ***Reproducibility***
2. ***Accessibility***
3. ***Performance***
4. ***Flexibility***
5. ***Consistency***
6. ***Functionality***

for as many researchers as possible

GVL: Design principles

Criteria	Design Implication
Accessible	Minimal client-side requirements
Reproducible	Workflow support + software & tool management process
Performance	User-managed scaling of compute resources + high availability resources
Flexible	User configurable + administrable Multiple interaction modes
Consistent	Single platform from training to analysis
Functional	Pre-populated with suite of tools for common use cases + required reference data + visualisation options

GVL: Design implications

Criteria	Design Implication	Technical implication
Accessible	Minimal client-side requirements	**Web based** tool and management interfaces
Reproducible	Workflow support + software & tool management process	**Workflow platforms** + automated process for **deployable underlying environment**
Performance	User-managed scaling of compute resources + high availability resources	**Cloud-based architecture** + interface for managing resources
Flexible	User configurable + administrable	**Per-user instances** accessible through web and command line; user-administrable environment
Consistent	Single platform from training to analysis	**Tutorials and guides** for training using best practice tools + **scalability**
Functional	Pre-populated with suite of tools for common use cases + required reference data + visualisation options	Process for **building underlying images** Automated configuration of **reference datasets**

For as many researchers as possible...

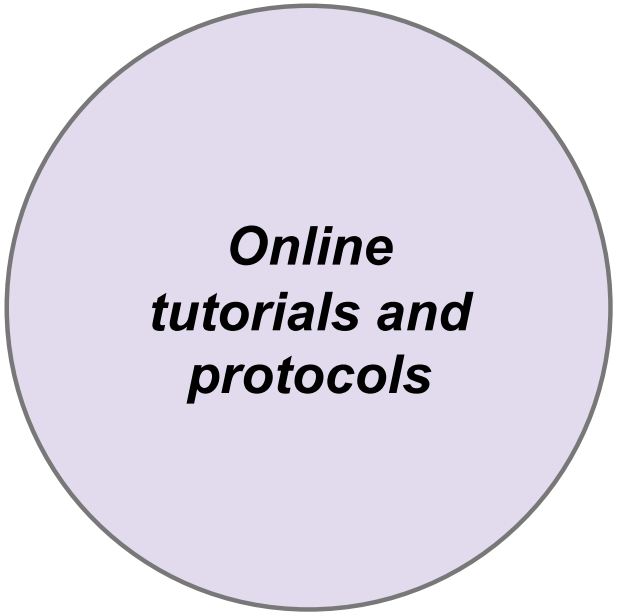
Galaxy Main

GVL: Philosophy

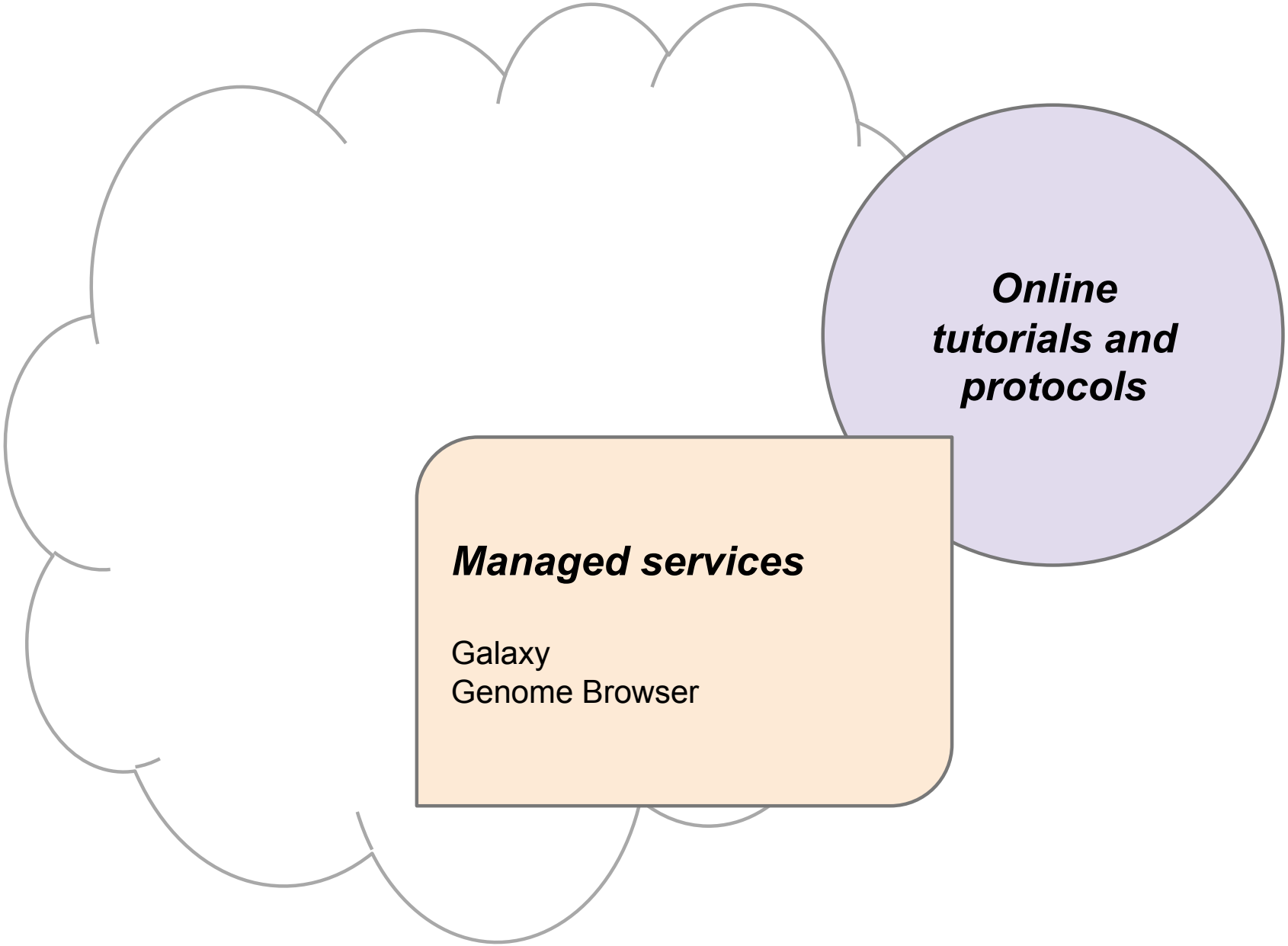
Genomics Virtual Lab

Galaxy Main

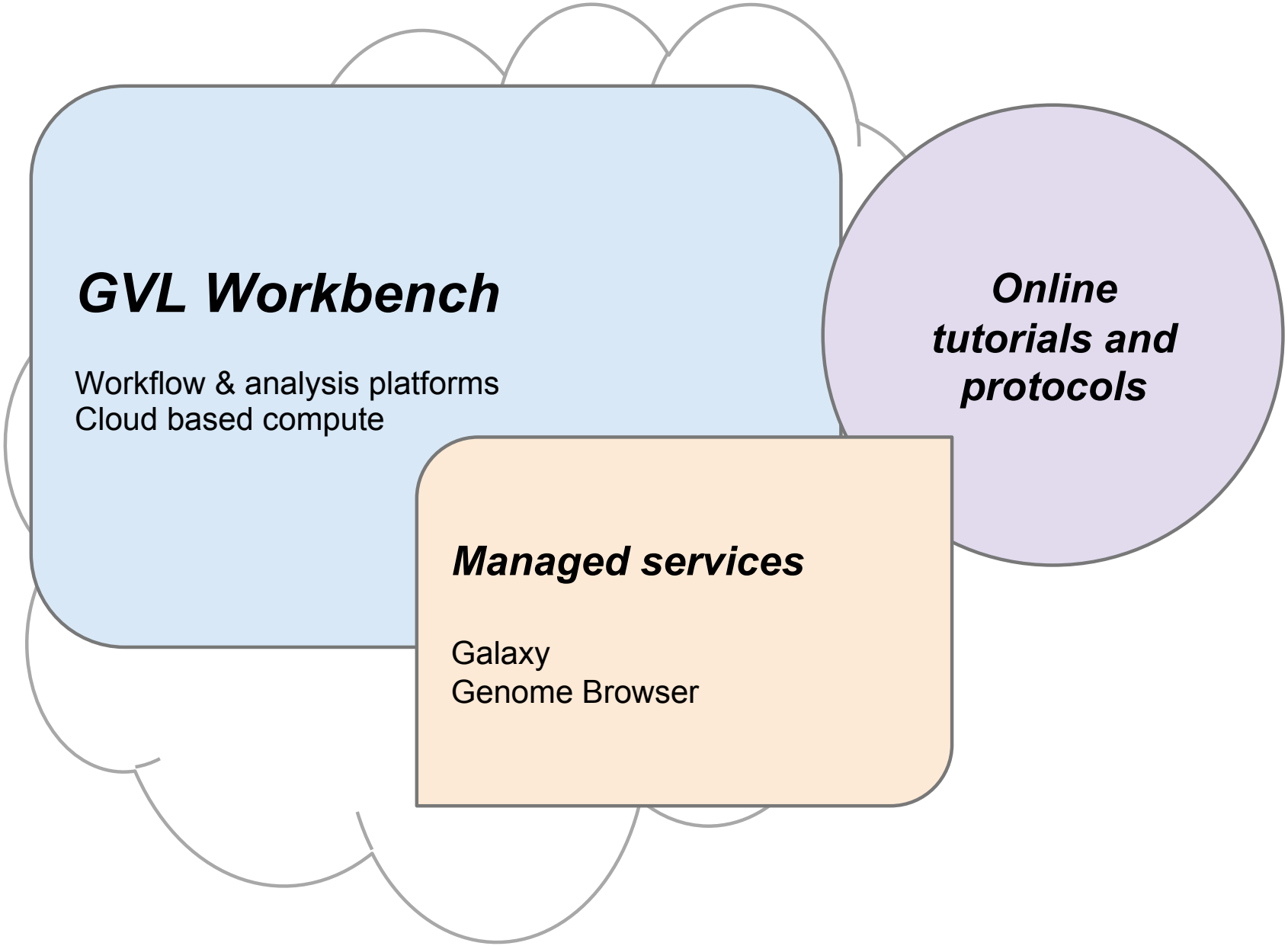
GVL: In practice



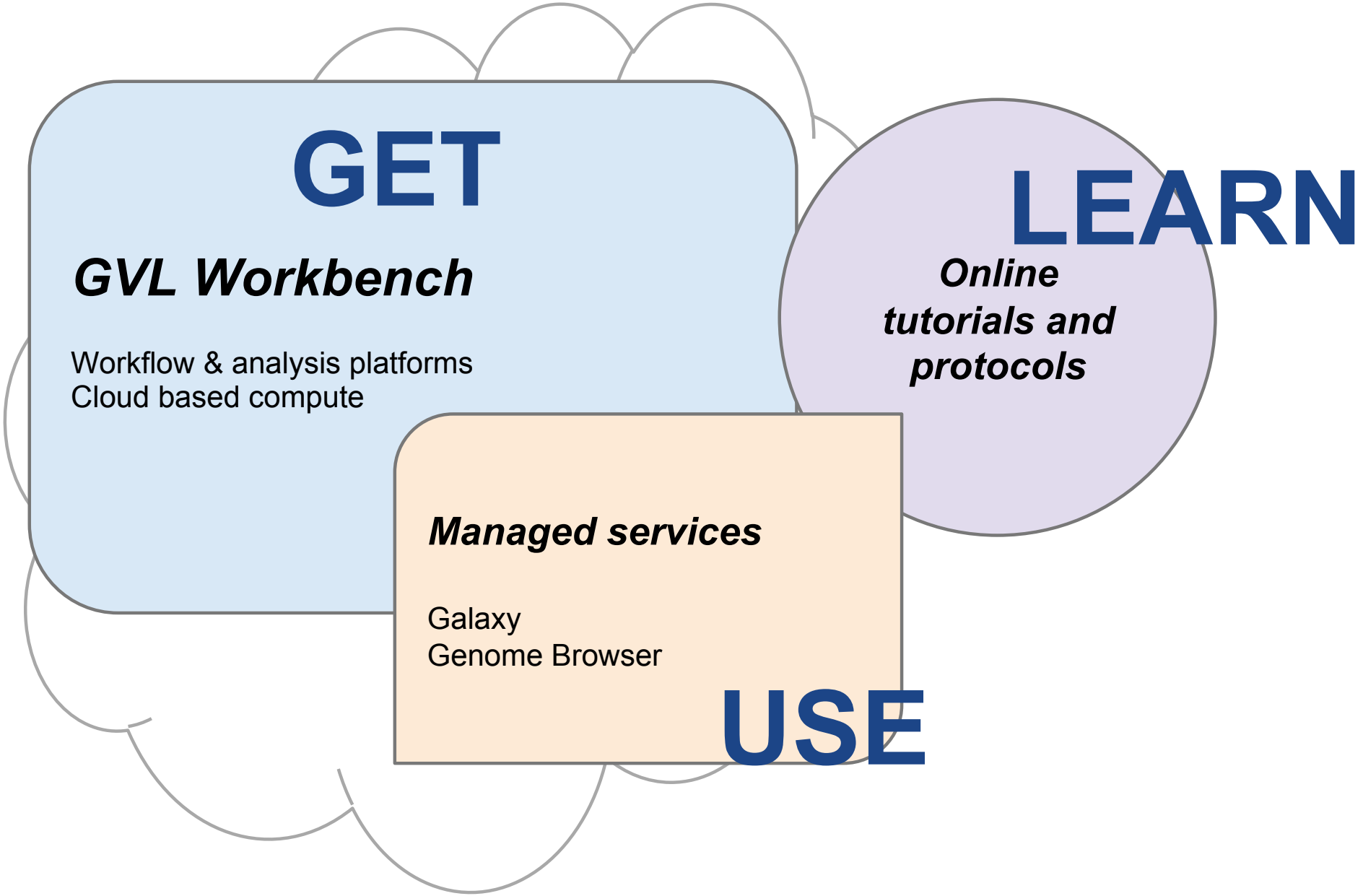
GVL: In practice



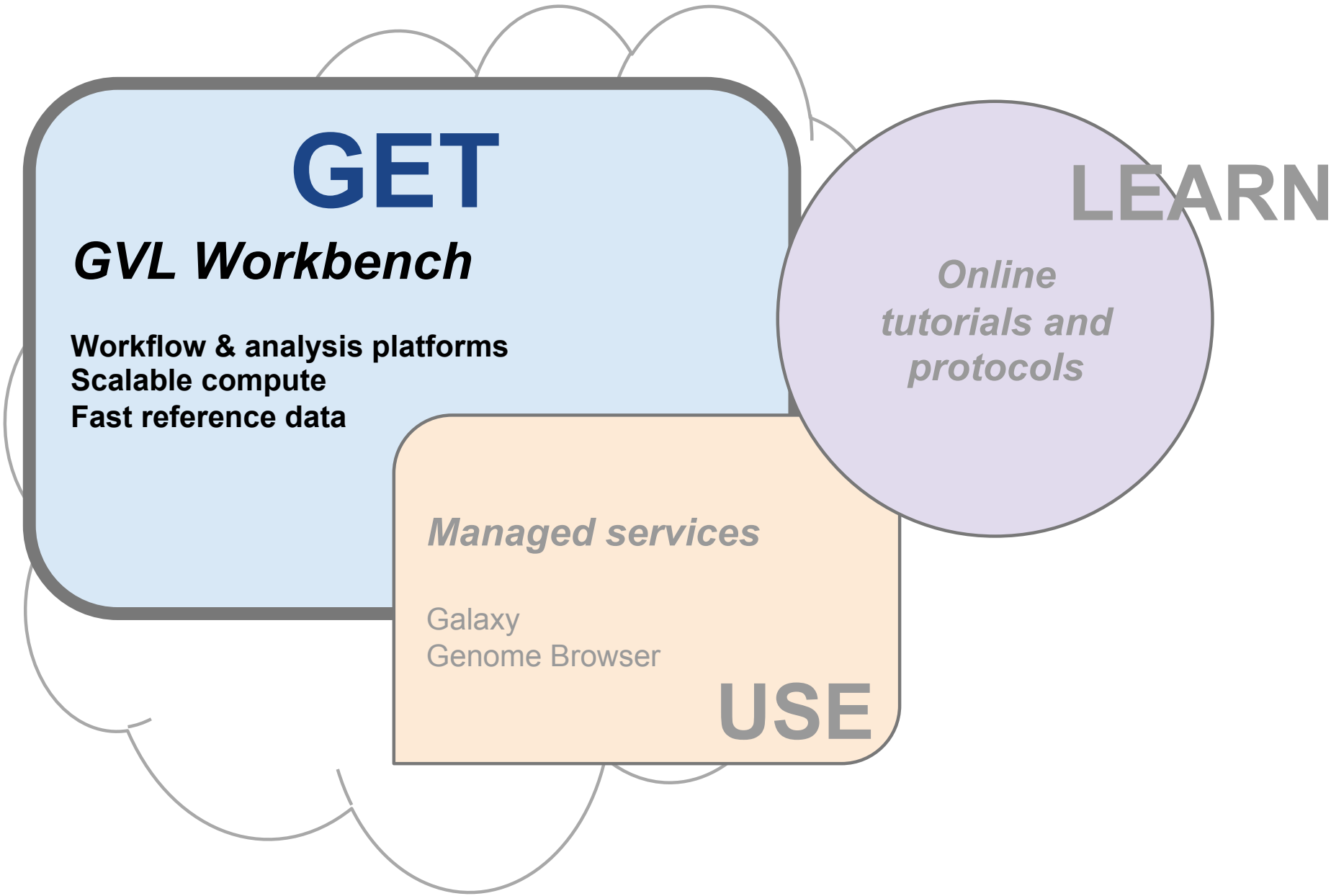
GVL: In practice



GVL: http://genome.edu.au

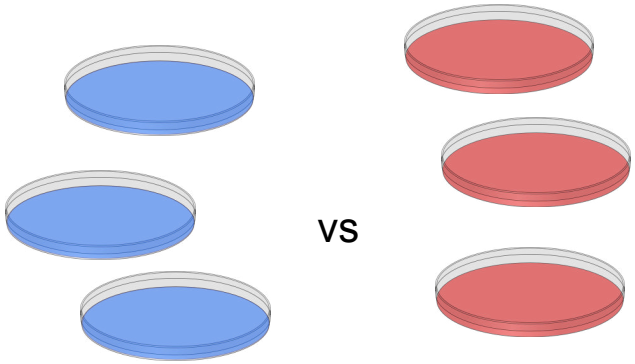


GVL: Developer's perspective



What characterises genomic analysis?

eg: Differential Gene Expression

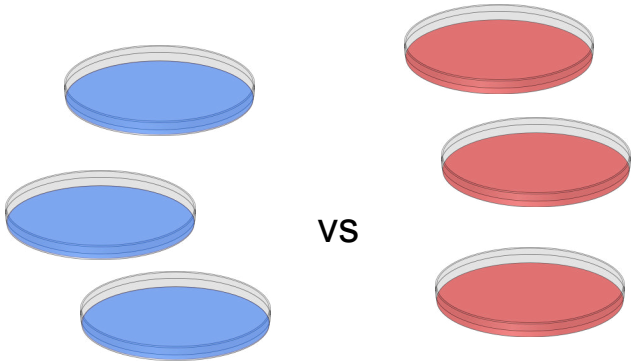


"What genes are turned on in blue cells and turned off in red cells?"

Differential Gene Expression

2E7 observations
over 1E5 genes
for 6 samples

(30GB raw data)



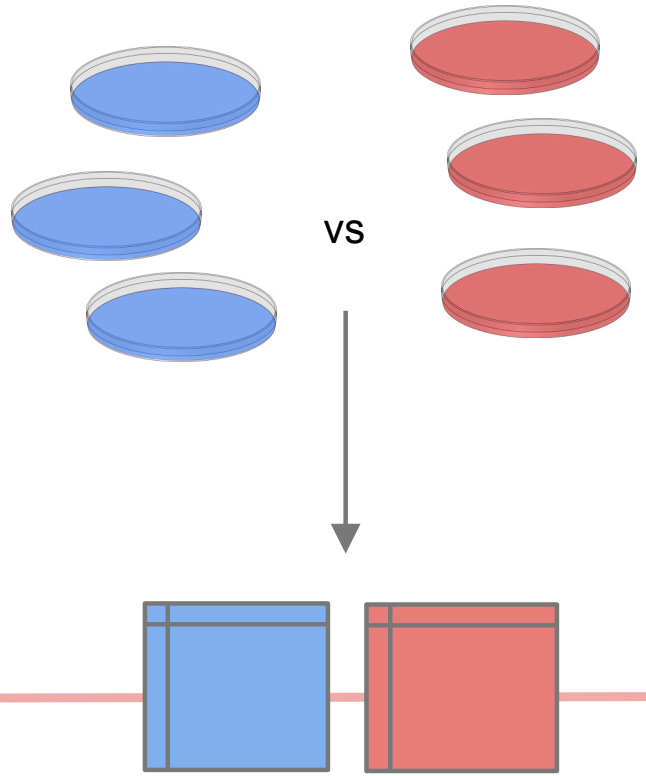
"What genes are turned on in blue cells and turned off in red cells?"

Differential Gene Expression

DATA REDUCTION

Workflows
Large reference data
High compute

2E7 observations over 1E5 genes for 2x3 samples

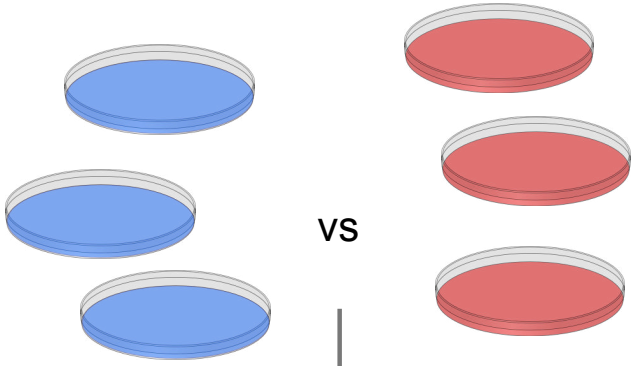


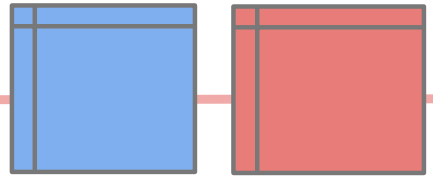
Differential Gene Expression

DATA REDUCTION

Workflows
Large reference data
High compute

2E7 observations over 1E5 genes for 2x3 samples

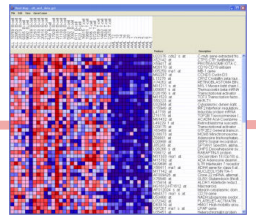




DATA ANALYSIS

Interactive
Flexibility

Test difference: [1,2,3] vs [4,5,6]

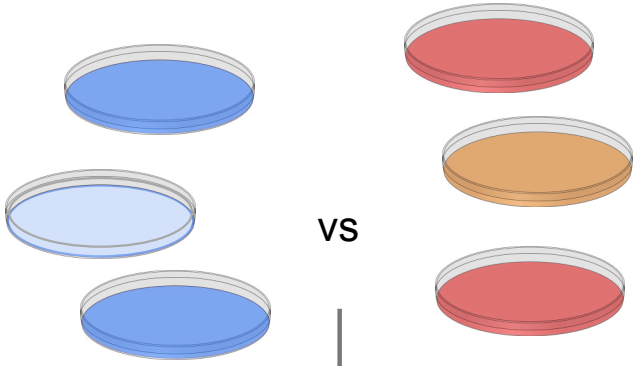


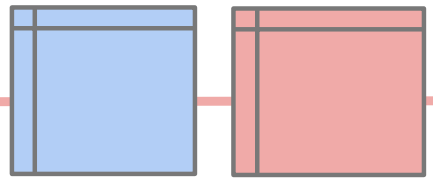
Differential Gene Expression

DATA REDUCTION

Workflows
Large reference data
High compute

2E7 observations over 1E5 genes for 2x3 samples

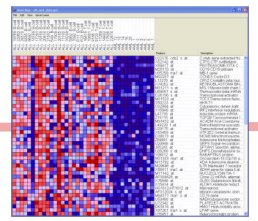




DATA ANALYSIS

Interactive
Flexibility

Test difference: [1,2,3] vs [4,5,6]

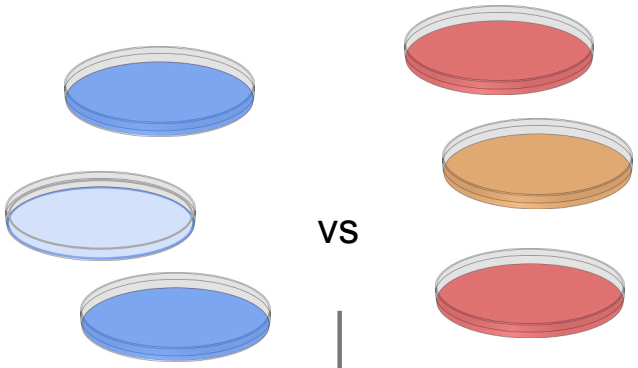


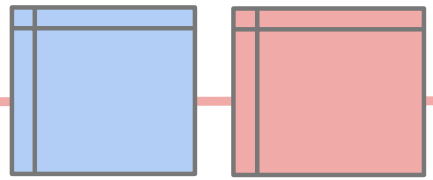
Differential Gene Expression

DATA REDUCTION

Workflows
Large reference data
High compute

2E7 observations over 1E5 genes for 2x3 samples

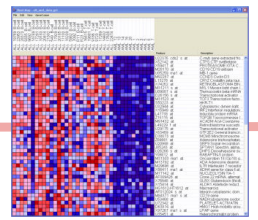




DATA ANALYSIS

Interactive
Flexibility

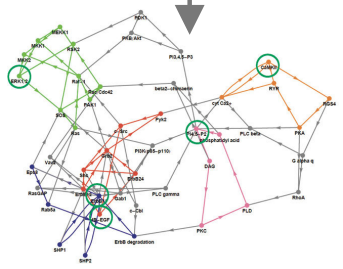
Test difference: [1,2,3] vs [4,5,6]



DATA INTERPRETATION

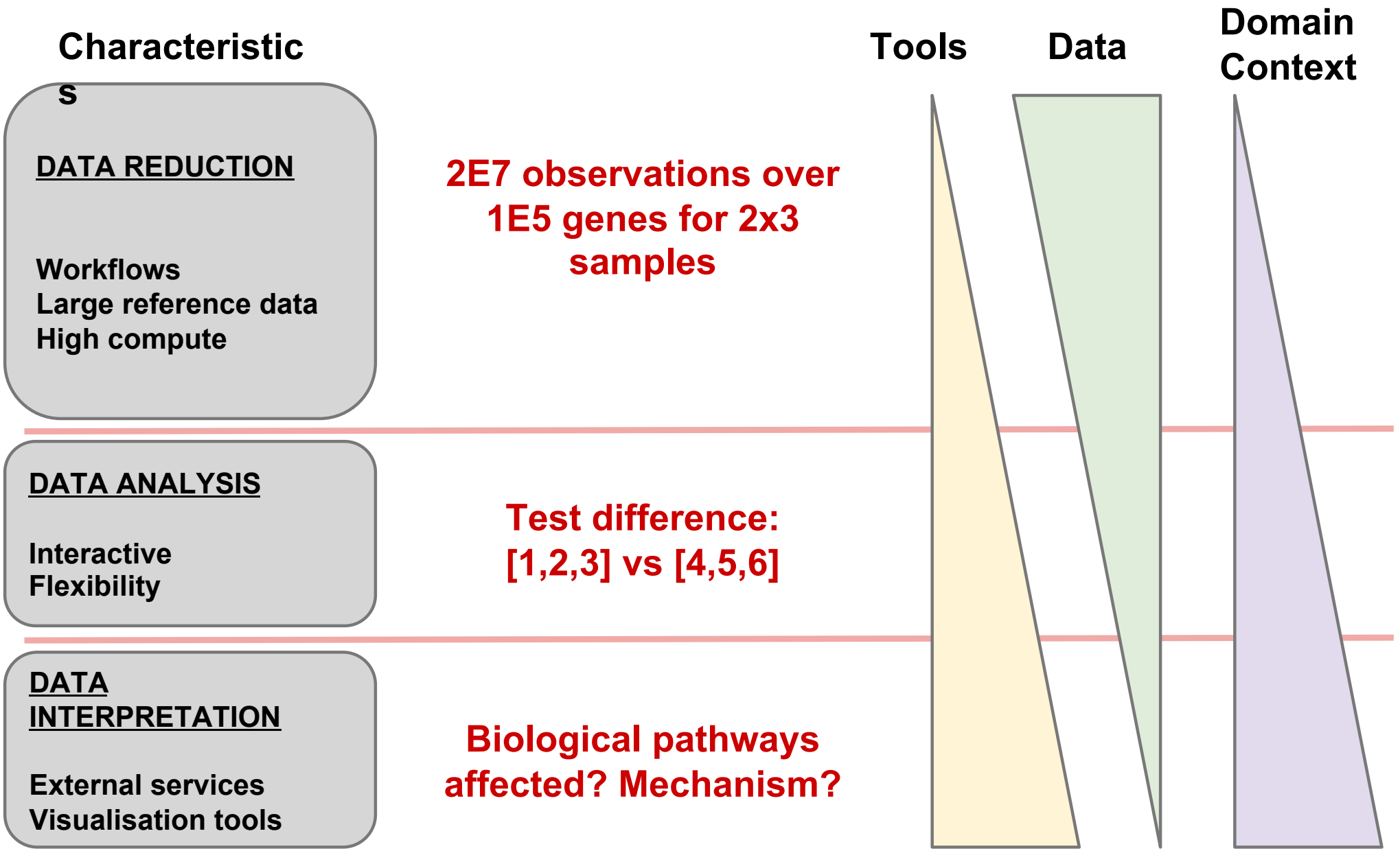
External services
Visualisation tools

Biological pathways affected? Mechanism?



Pathway knowledge

Differential Gene Expression



What characterises genomics?

- Very large experimental datasets per user/ group
- I/O intensive high compute initial analysis
  - ‘data reduction’: raw data to sample summaries
- Large suite of data analysis tools, interactive
  - a bit subjective
- Complex context for interpretation, external tools
  - more subjective, domain knowledge
- Little modelling/simulation

GVL Workbench: Requirements

A web-based *per-user* workbench providing:

- **access to multiple tools**
- **on a scalable back end compute cluster**
- **with fast access to large reference data,**
- **user administrable and configurable**
- **with multiple modes of interaction**
- **and a mechanism for reproducible workflows**

all highly available and accessible
i.e. with a minimal cost of entry to the user

Why per-user?

Managed service: objective

A short time later…

Why per-user?

Managed service: objective

A short time later…

GVL: Philosophical assertion

>

GET a GVL

http://genome.edu.au → GET

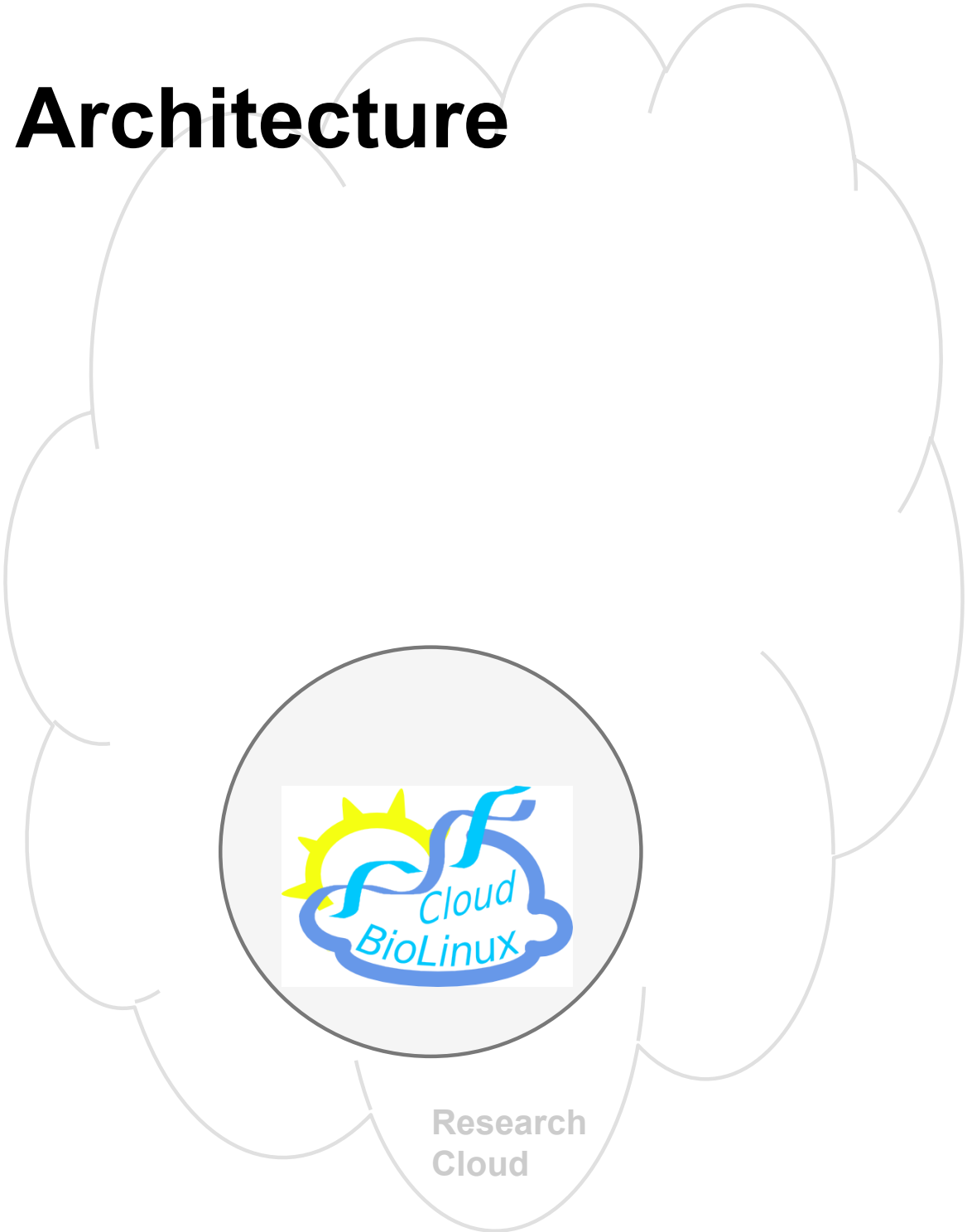
Building (deploying and running) a GVL instance:

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*

GVL Workbench: Architecture

GVL Workbench

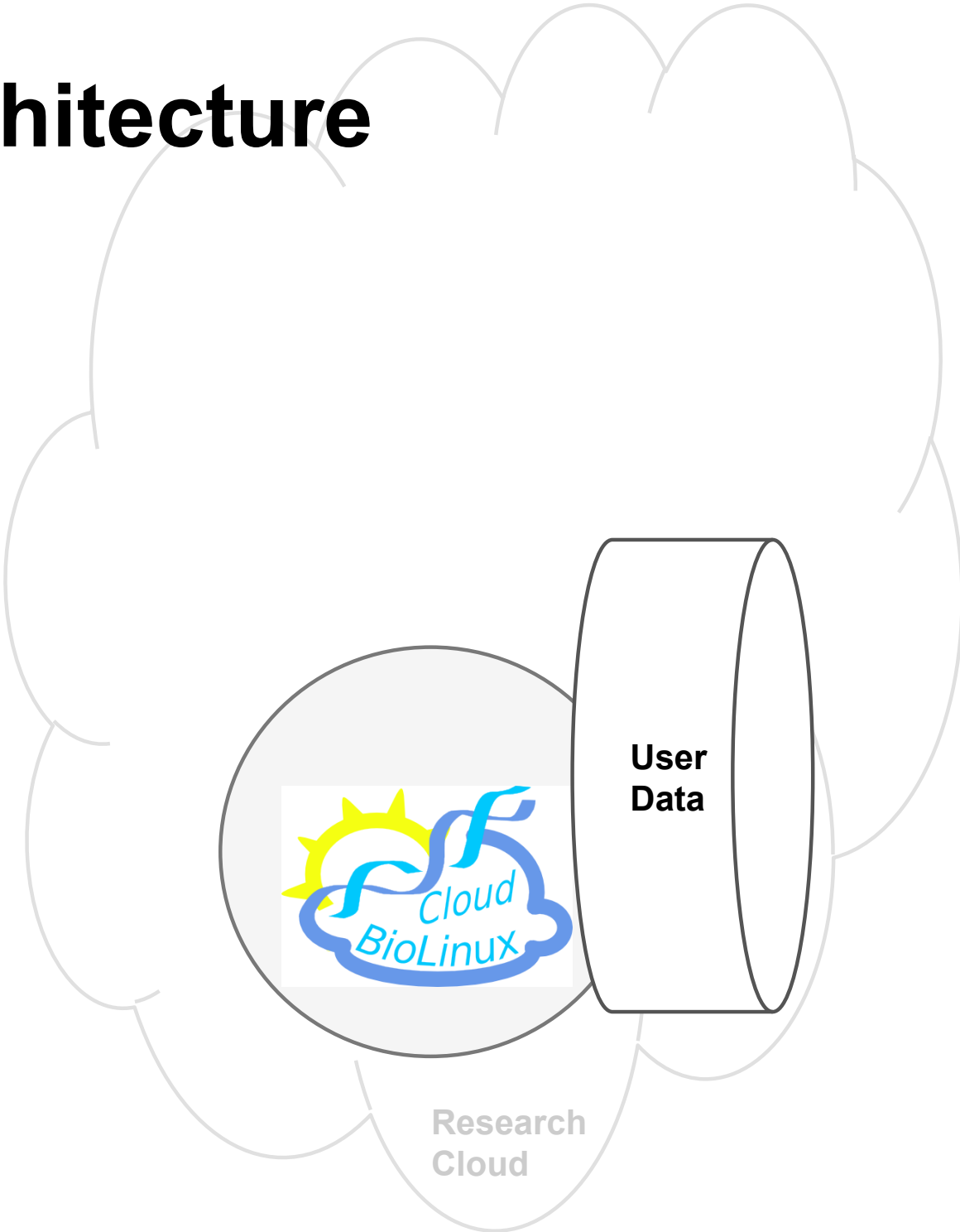
Workflow & analysis platforms



Workbench: Architecture

GVL Workbench

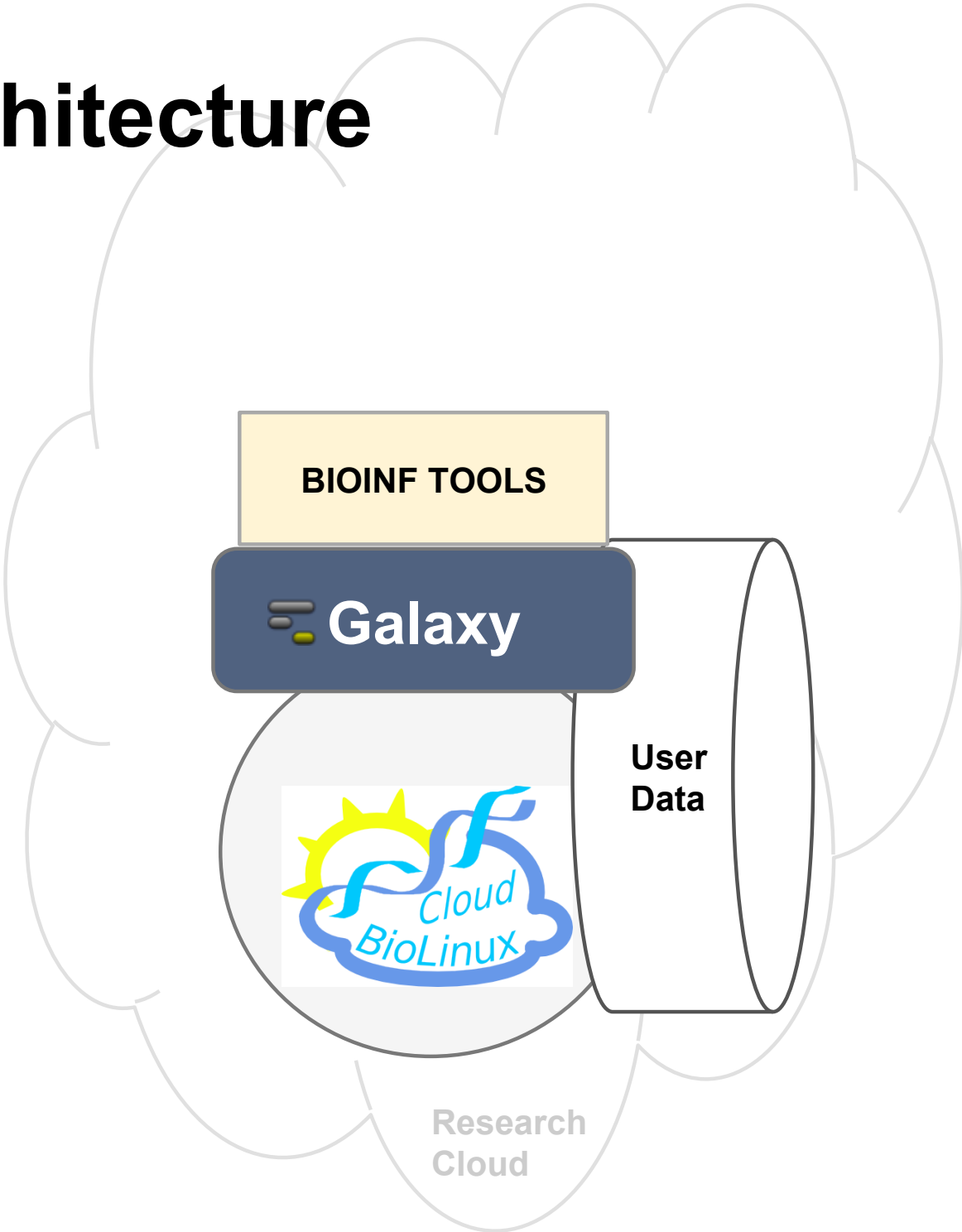
Workflow & analysis platforms



Workbench: Architecture

GVL Workbench

Workflow & analysis platforms

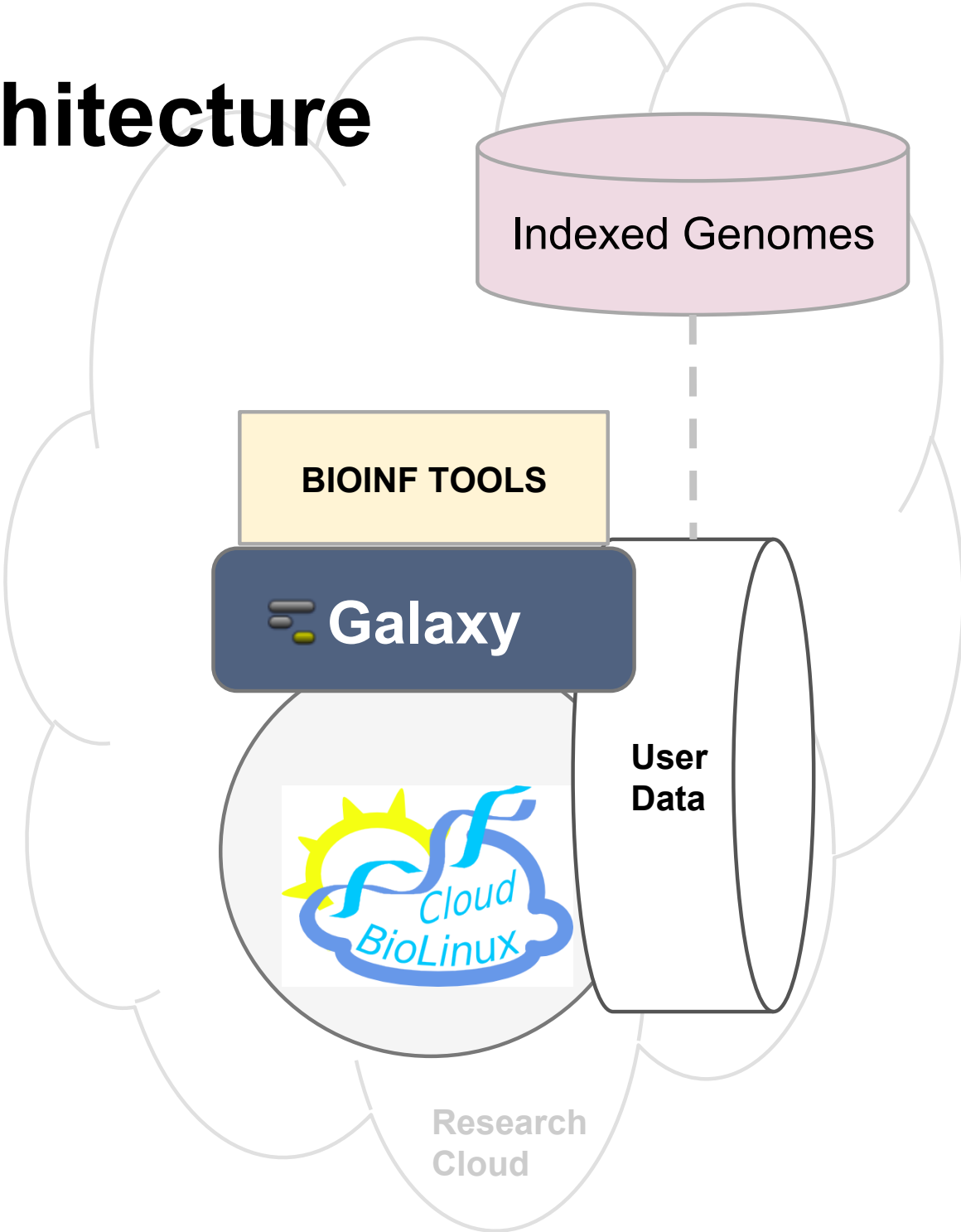


Workbench: Architecture

GVL Workbench

Workflow & analysis platforms

Fast reference data

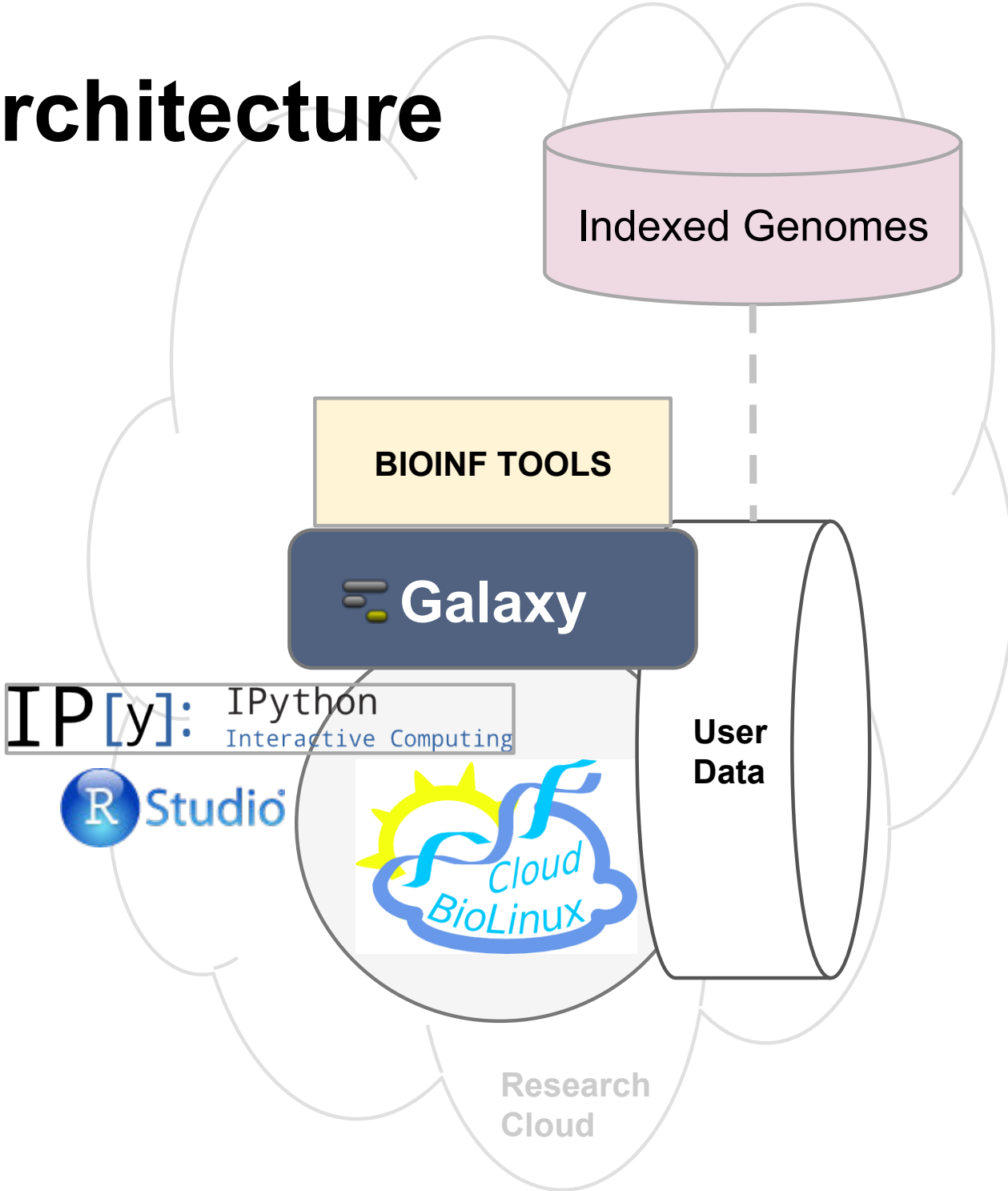


Workbench: Architecture

GVL Workbench

Workflow & analysis platforms

Fast reference data



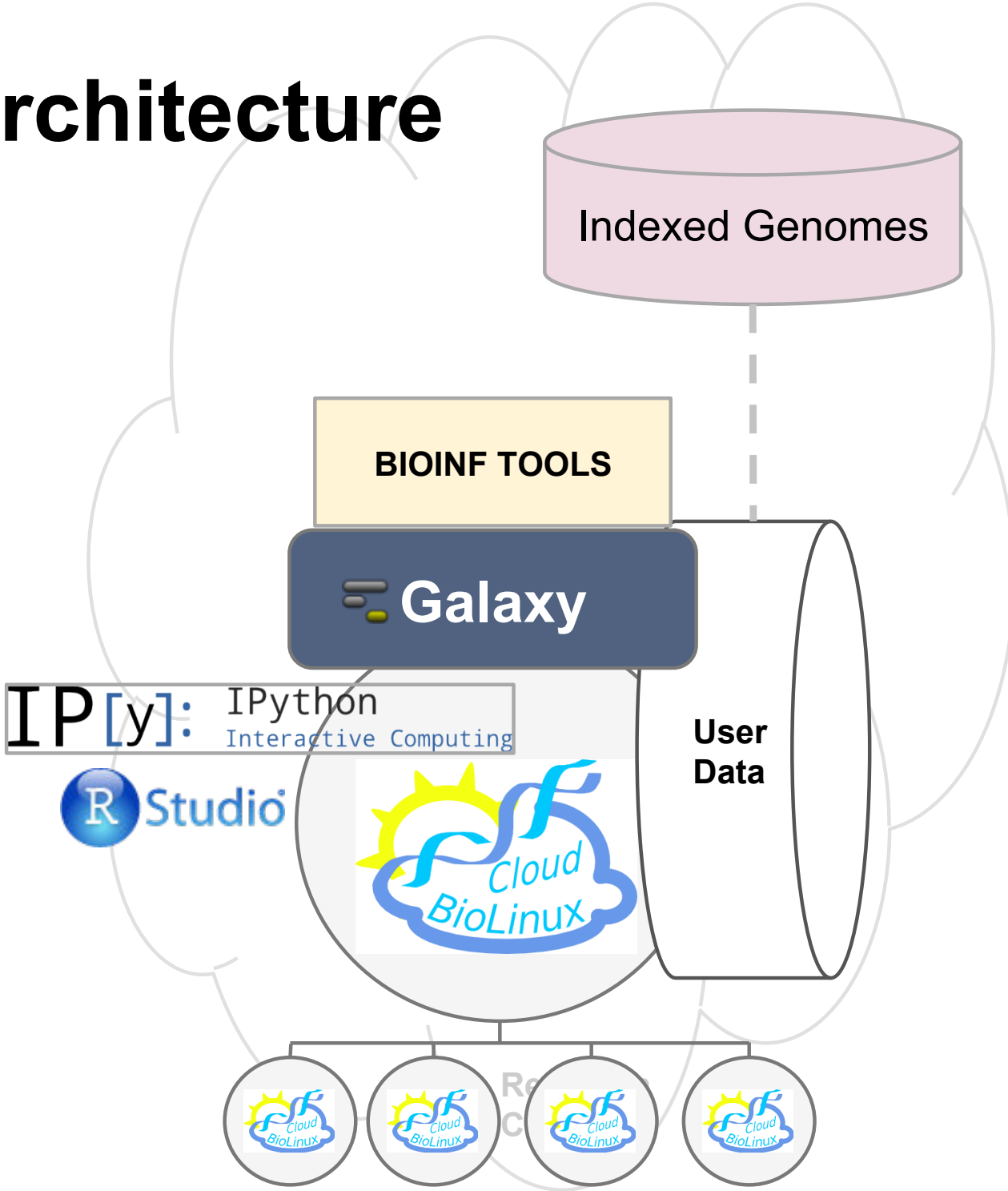
Workbench: Architecture

GVL Workbench

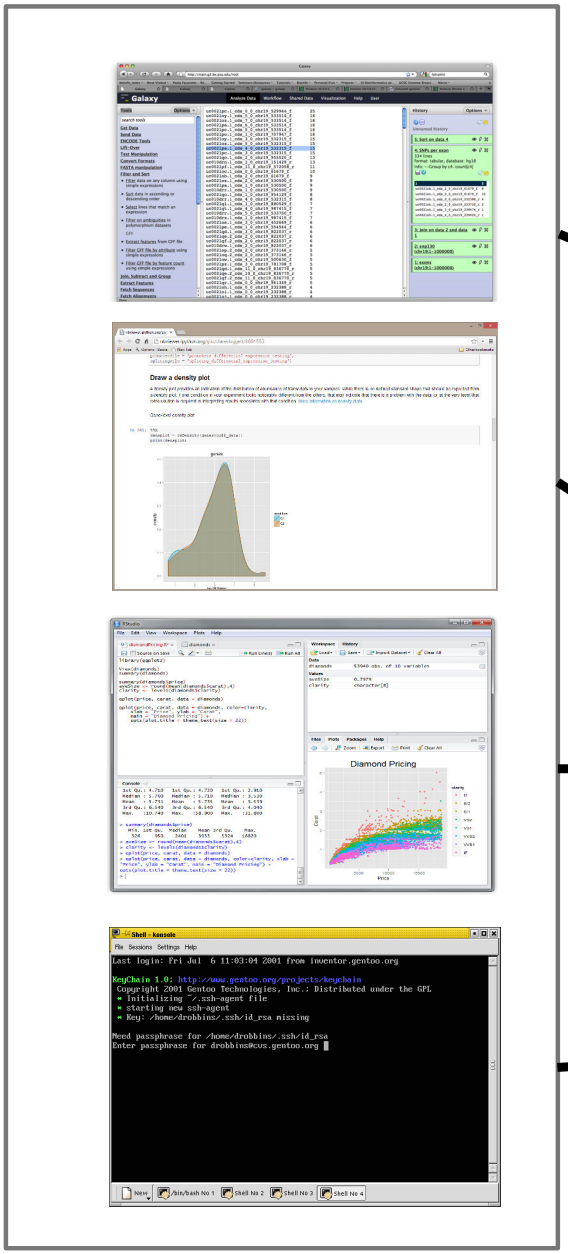
Workflow & analysis platforms

Fast reference data

Scalable compute



Workbench: Architecture



Indexed Genomes

BIOINF TOOLS

Galaxy

IP[y]: IPython Interactive Computing

RStudio

User Data

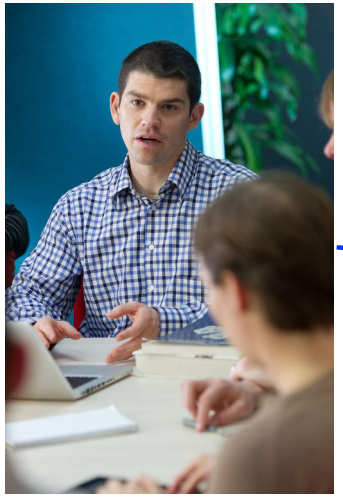
Cloud BioLinux

Engineering:
Deploying and running a GVL

http://launch.genome.edu.au

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

CloudMan



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]

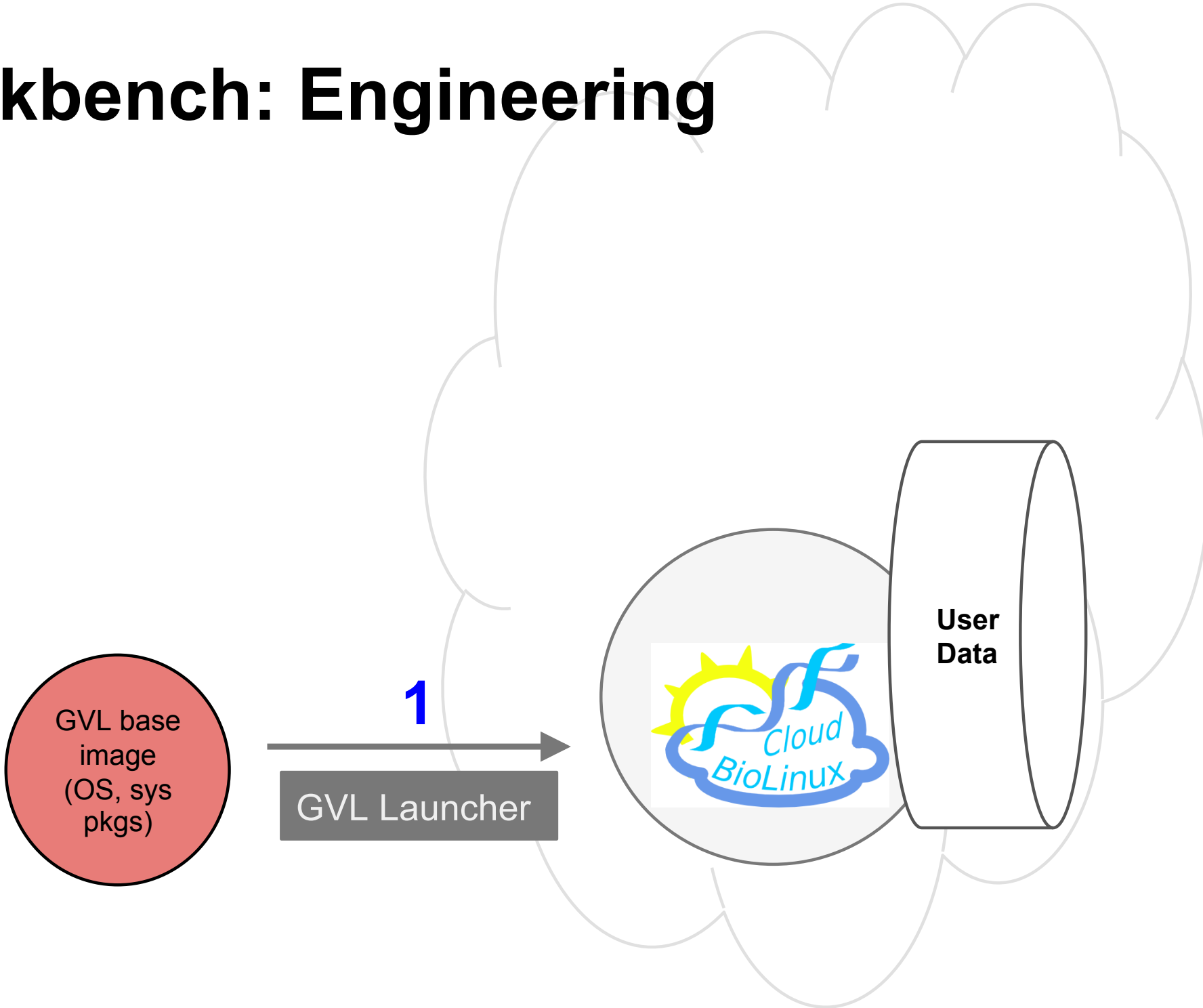
Abstract

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.

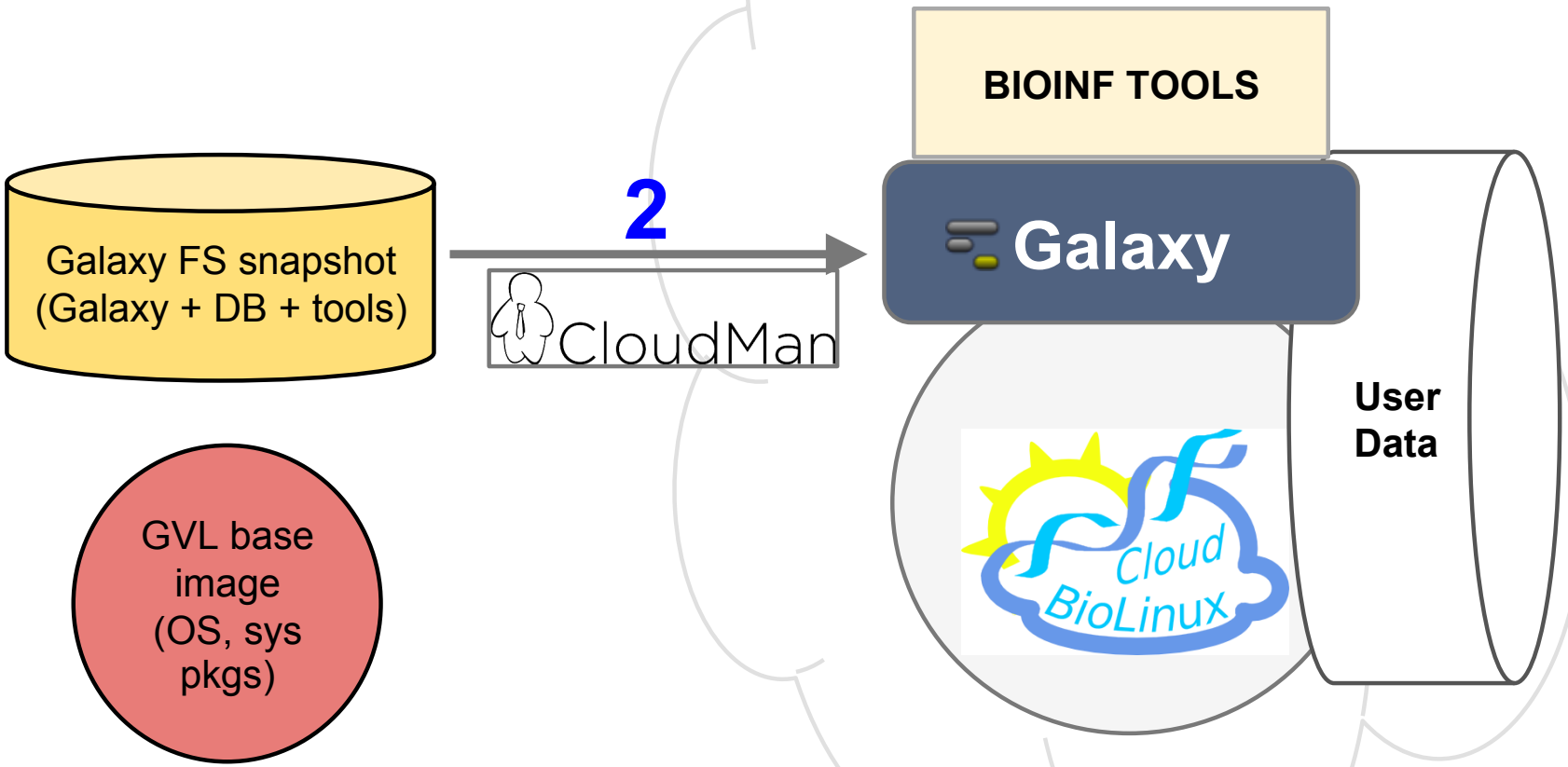
Results: CloudMan (usecloudman.org) enables individual researchers to easily deploy, customize, and share their entire cloud analysis environment, including data, tools, and configurations.

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

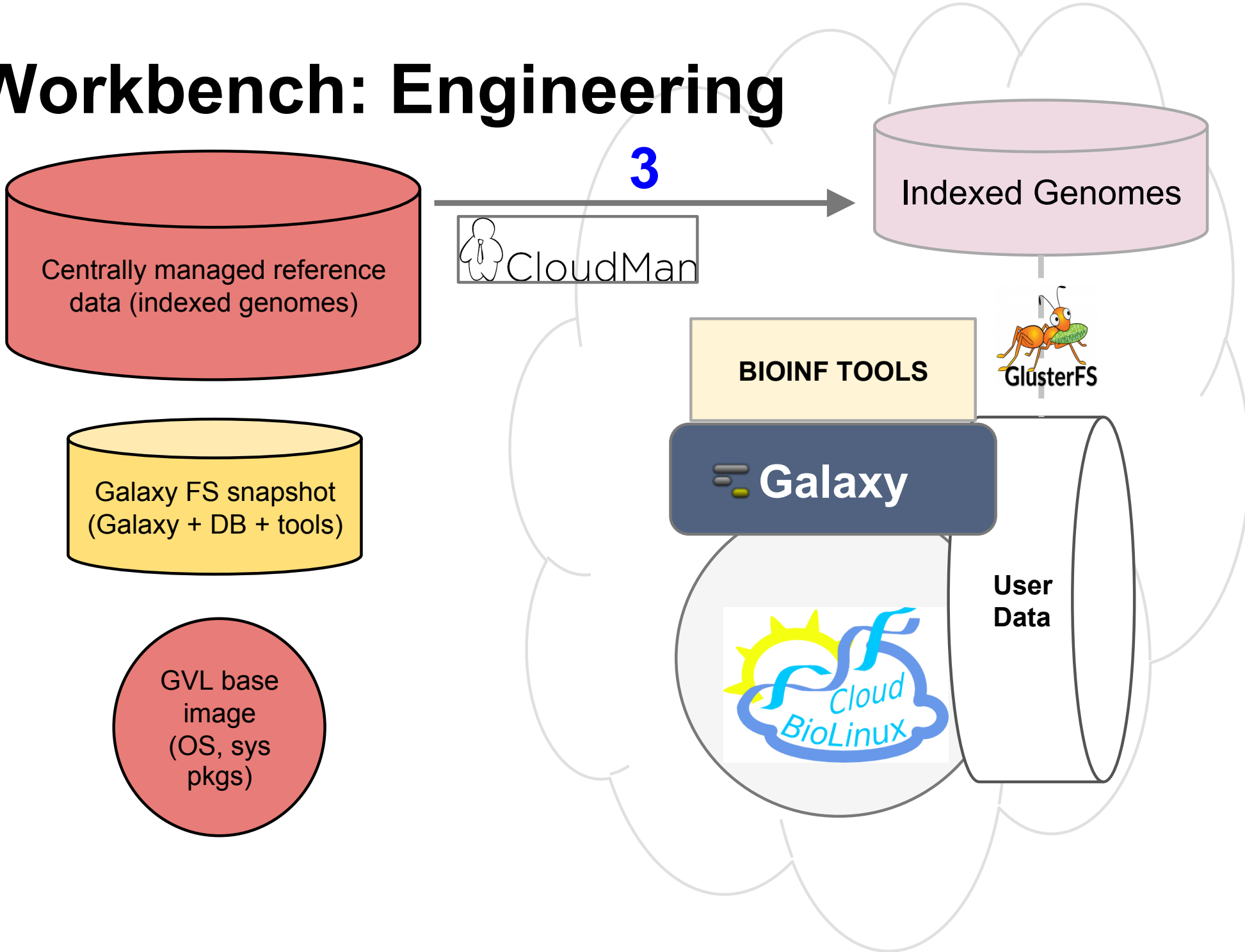
Workbench: Engineering



Workbench: Engineering



Workbench: Engineering

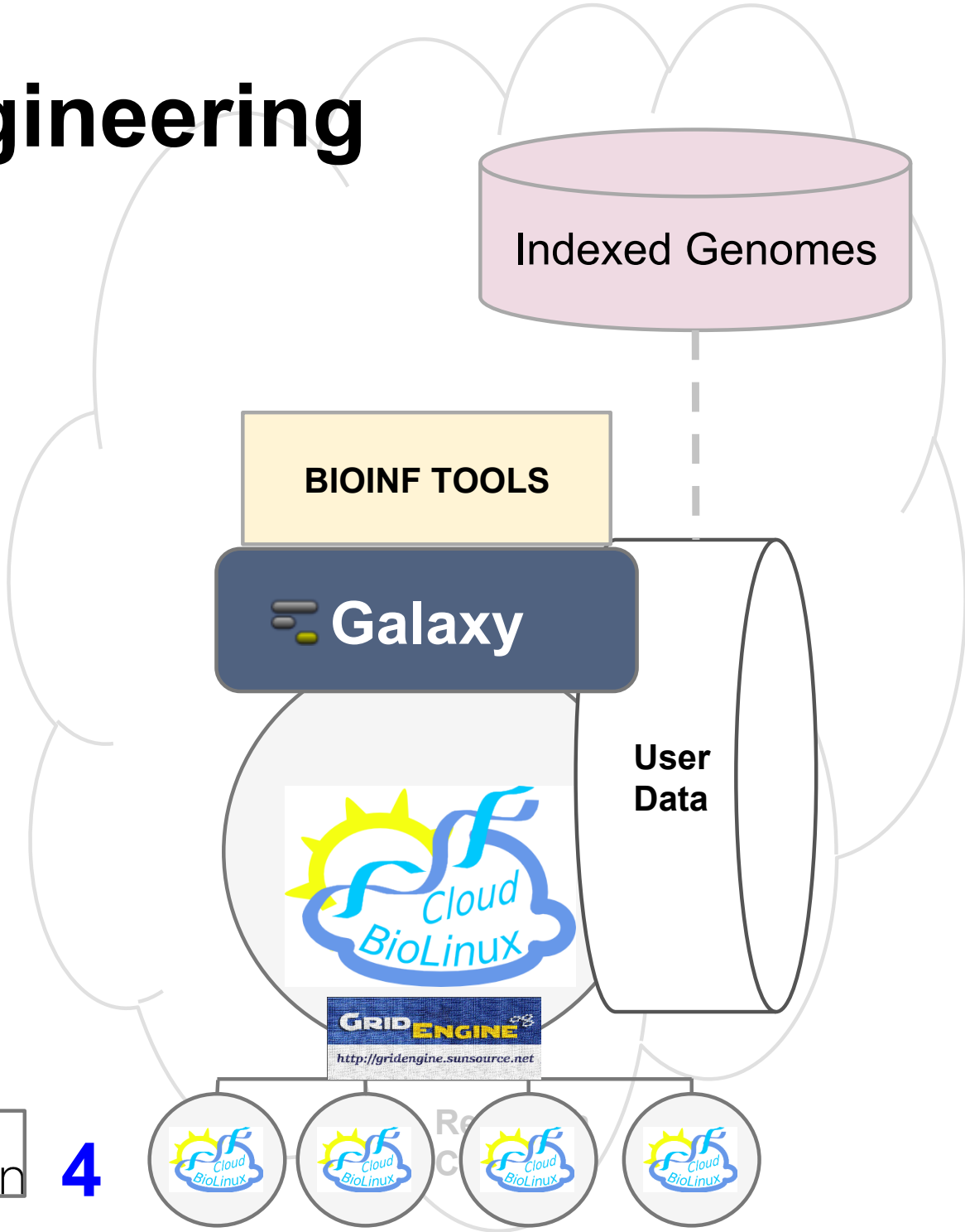


Workbench: Engineering

Centrally managed reference data (indexed genomes)

Galaxy FS snapshot (Galaxy + DB + tools)

GVL base image (OS, sys pkgs)



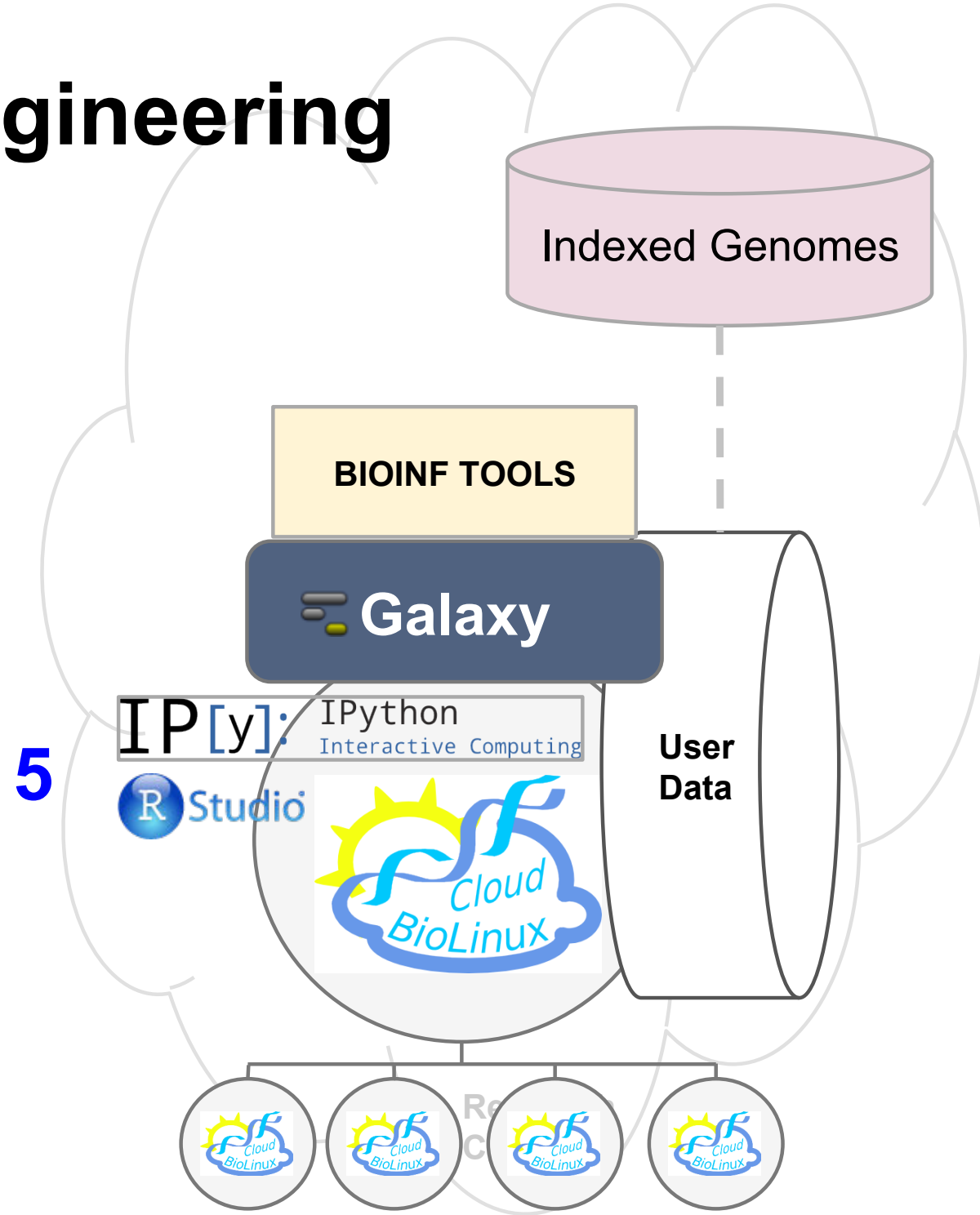
CloudMan

Workbench: Engineering

Centrally managed reference data (indexed genomes)

Galaxy FS snapshot (Galaxy + DB + tools)

GVL base image (OS, sys pkgs)



Workbench: All components

Management layer

CloudMan

GVL Launcher

Application layer

Galaxy

RStudio

IPython Interactive Computing

System layer

CloudBioLinux

GridEngine http://gridengine.sunsource.net

GlusterFS

Cloud layer

openstack CLOUD SOFTWARE

amazon web services™

Component build layer

Base

Galaxy FS

Ref data

GVL: Does it work?

Technically?

Practically?

http://genome.edu.au → GET

	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

Lessons?

Defining and maintaining a set of tools is challenging

Providing per-user performance is challenging

The cloud is only so scalable!

Not all cloud nodes are equal

Geography matters

Lessons?

Defining and maintaining a set of tools is challenging

Providing per-user performance is challenging

The cloud is only so scalable!

Not all cloud nodes are equal

Geography matters

Resourcing is key!

What's next for GVL?

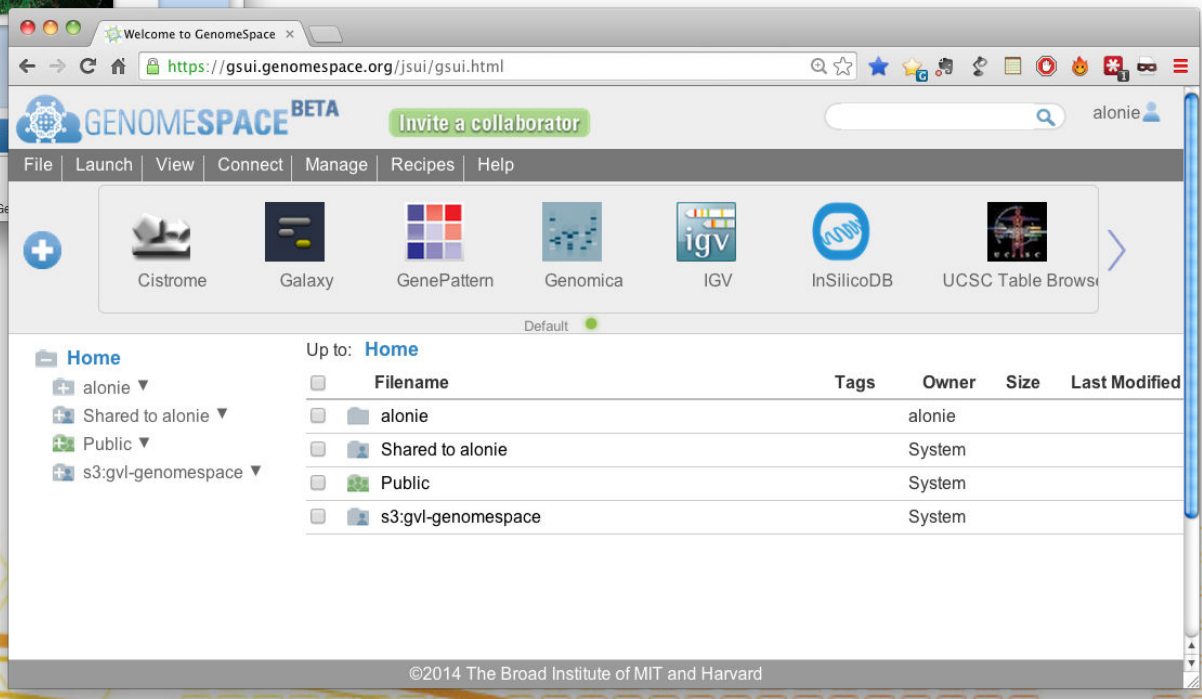
http://genome.edu.au

Moving data around is a problem

Whole genomes: 300GB raw data

We need to remove the desktop and USB sticks from the process!

Genomics lifecycle on the cloud: GenomeSpace



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 Diabetes 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)