

# 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***



R D S I

Research Data Storage  
Infrastructure



nectar

<http://nectar.org.au>

# The Australian Research Cloud



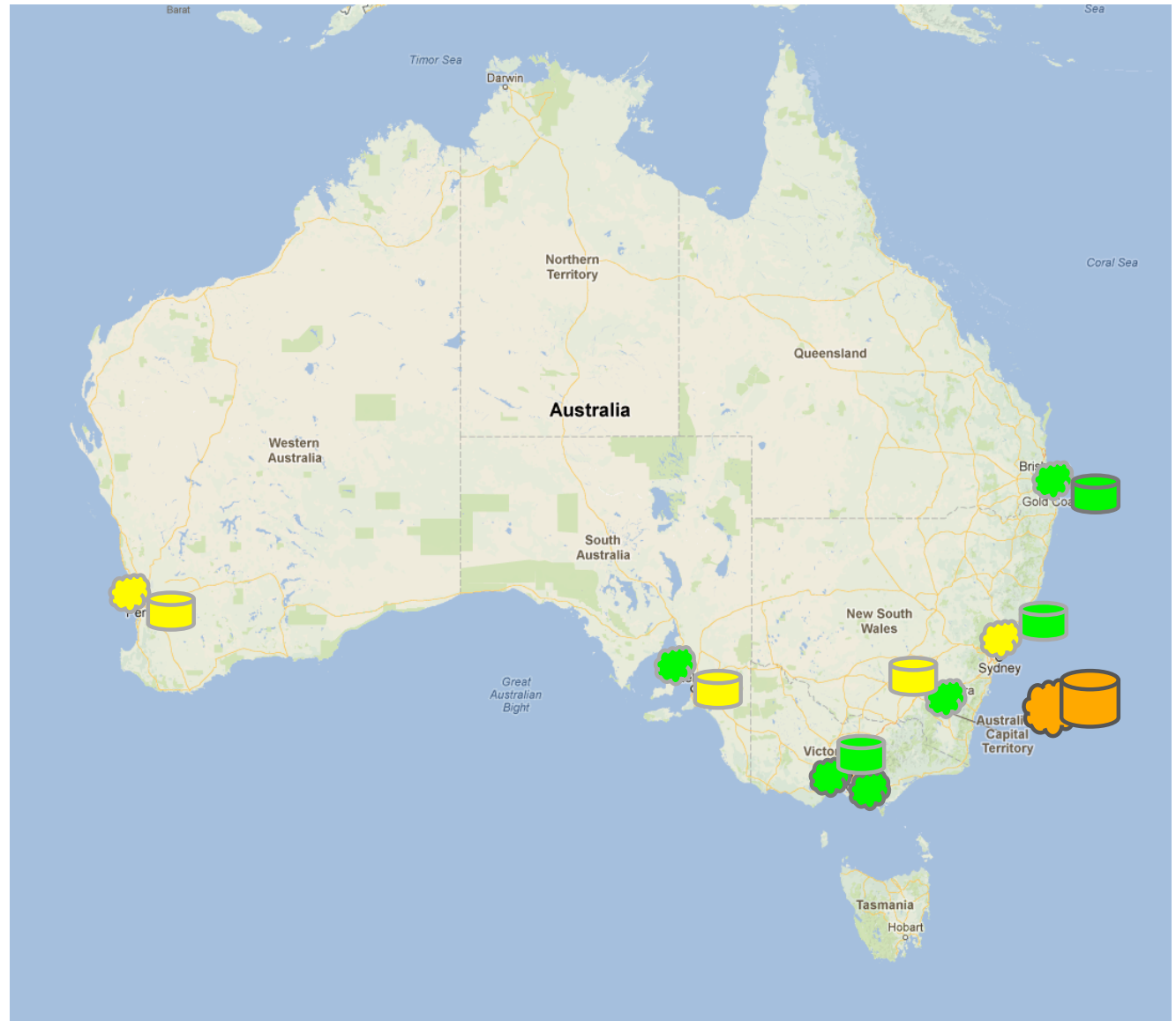
Cloud node:  
3-6000 cores



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

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

# GVL: Design implications

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

**For as many researchers as possible...**

**Galaxy Main**



# GVL: Philosophy

**Genomics Virtual Lab**



**Galaxy Main**



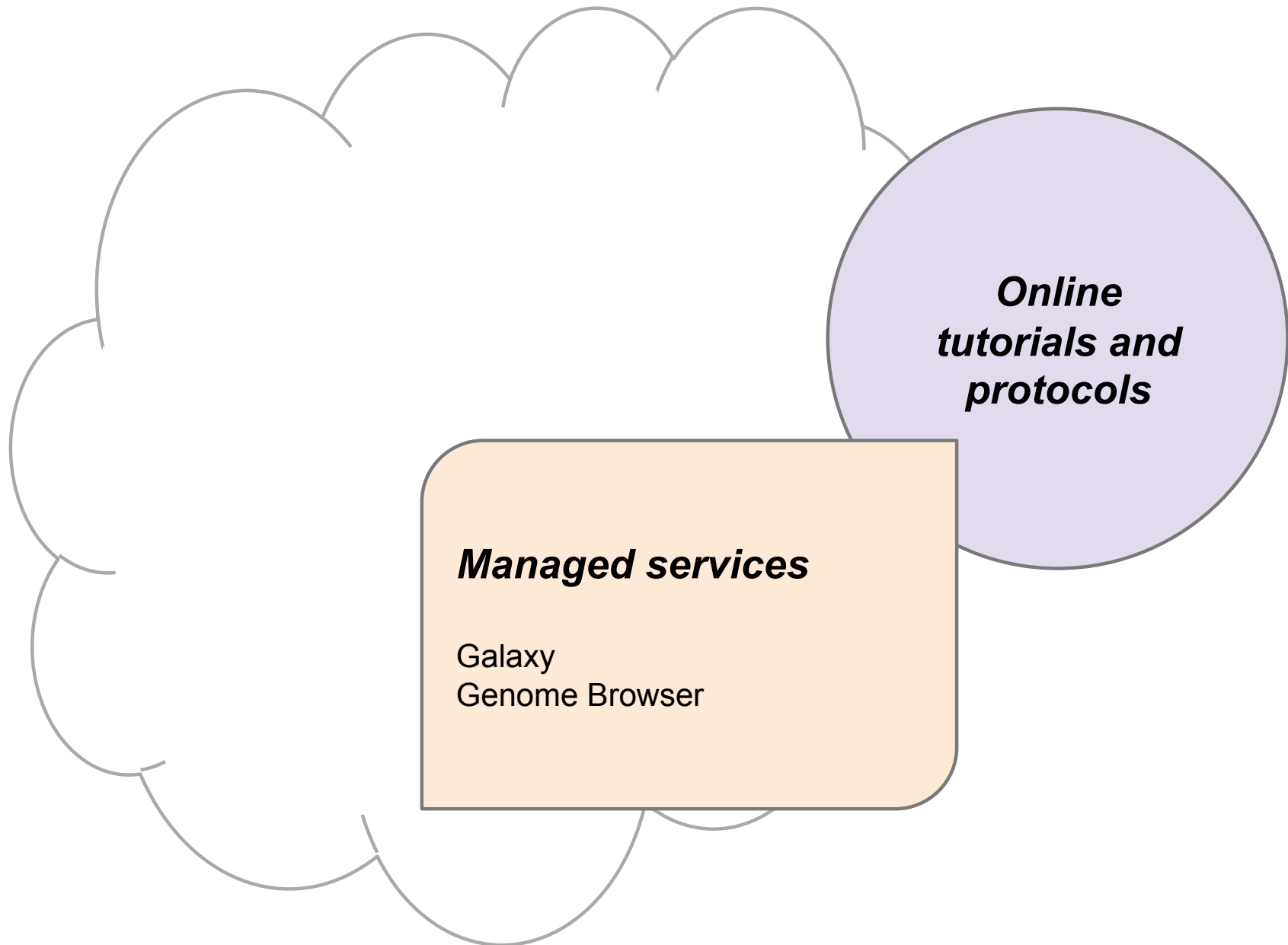


# GVL: In practice

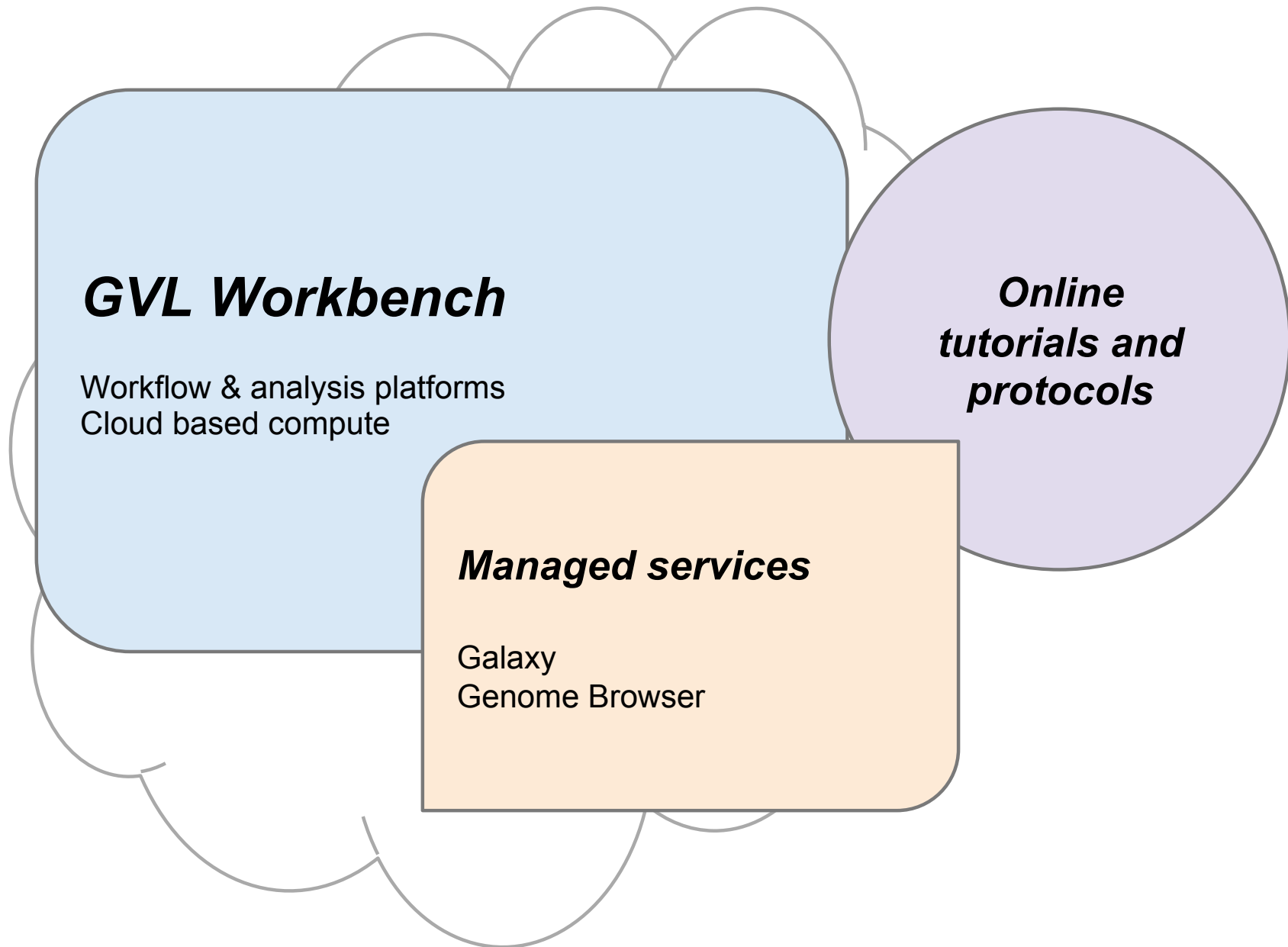


***Online  
tutorials and  
protocols***

# GVL: In practice



# GVL: In practice



# GVL: In practice

**GET**

***GVL Workbench***

Workflow & analysis platforms  
Cloud based compute

**LEARN**

***Online  
tutorials and  
protocols***

***Managed services***

Galaxy  
Genome Browser

**USE**

**GVL: <http://genome.edu.au>**

**GET**

***GVL Workbench***

Workflow & analysis platforms  
Cloud based compute

***Managed services***

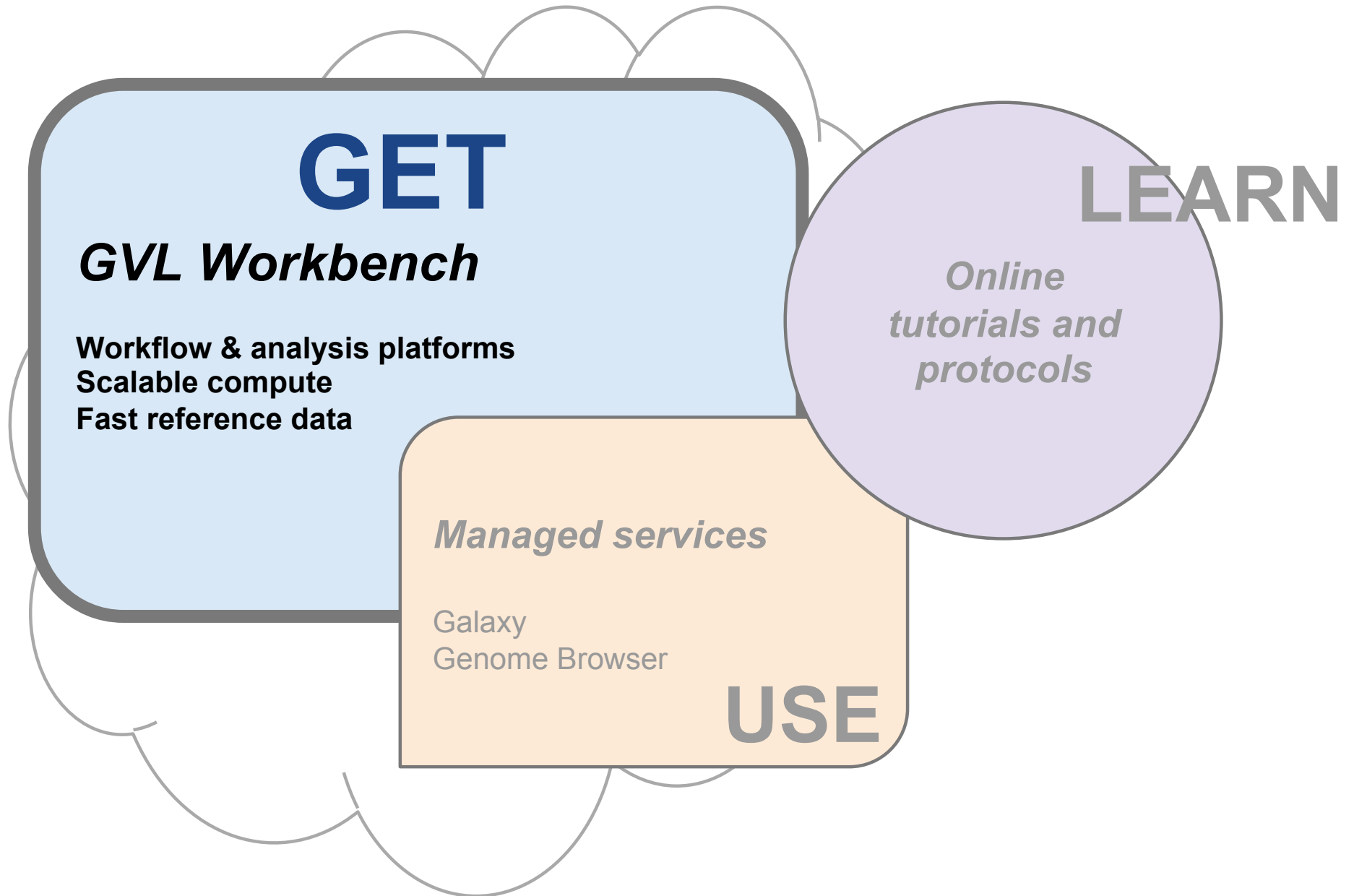
Galaxy  
Genome Browser

**LEARN**

***Online  
tutorials and  
protocols***

**USE**

# GVL: Developer's perspective



# 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**



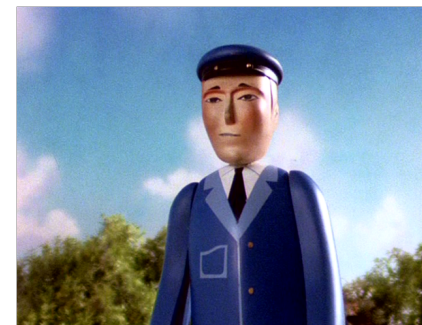
**A short time later...**



# GVL: Philosophical assertion



# GVL: Philosophical assertion+



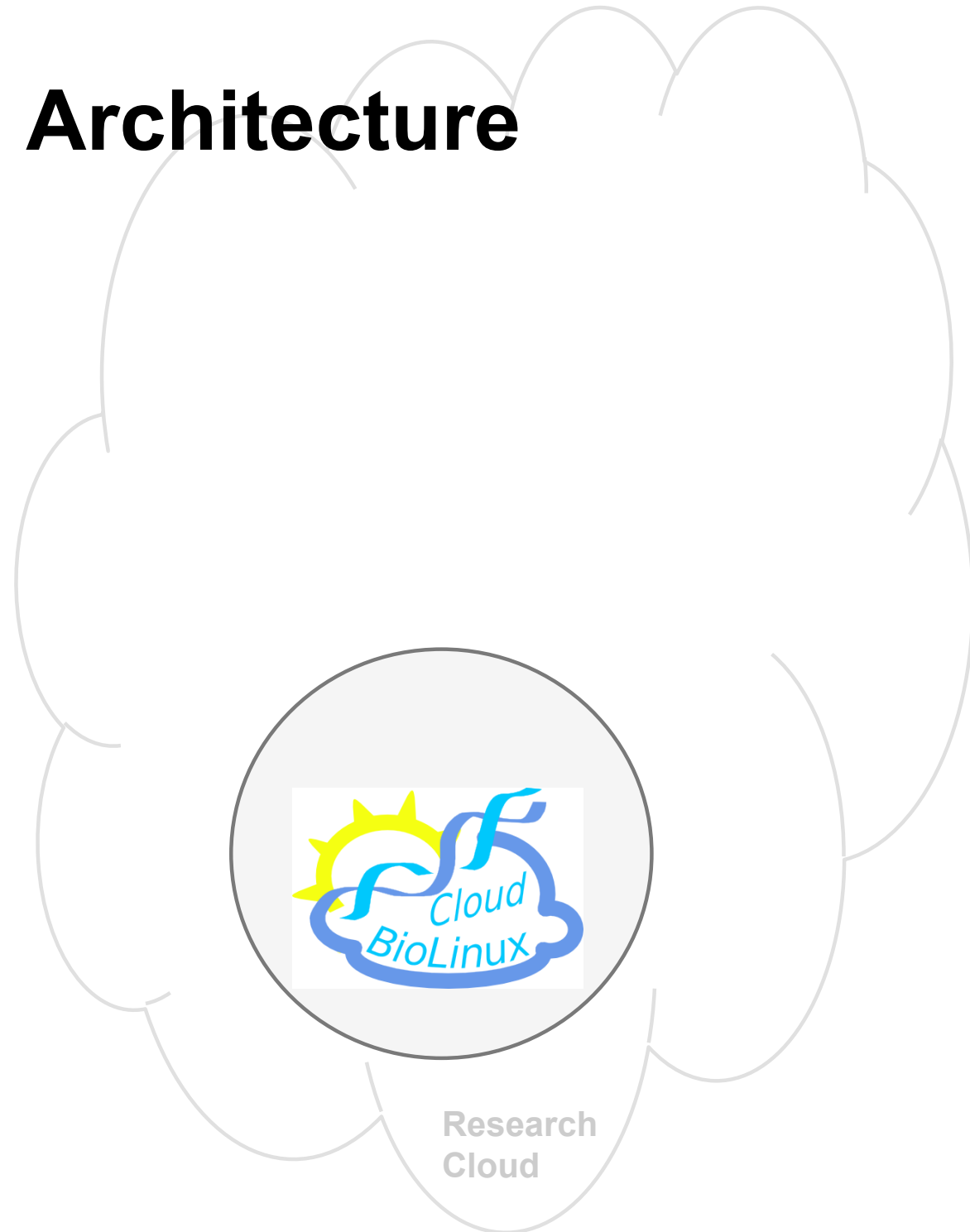
# GET a GVL

<http://genome.edu.au> → GET

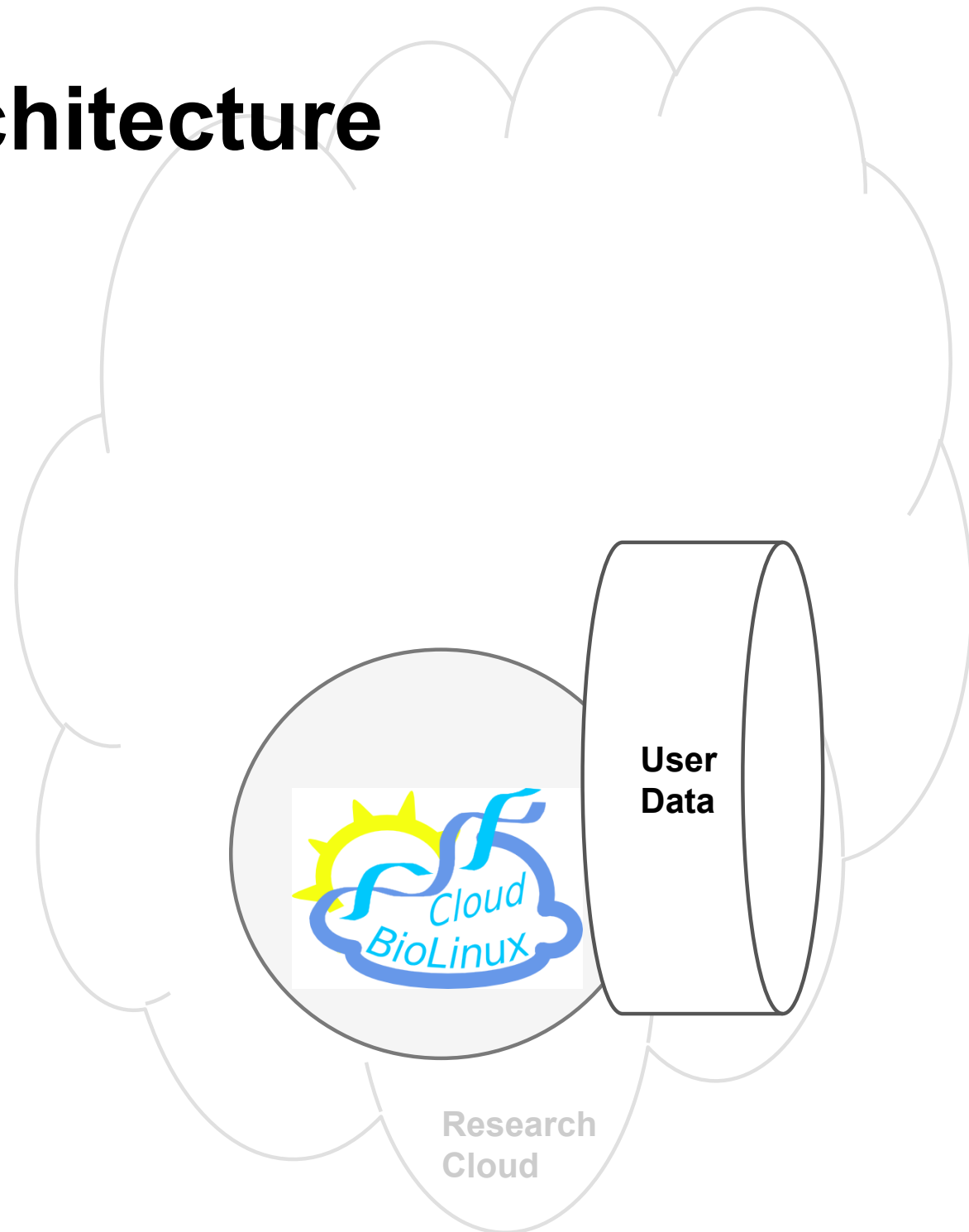
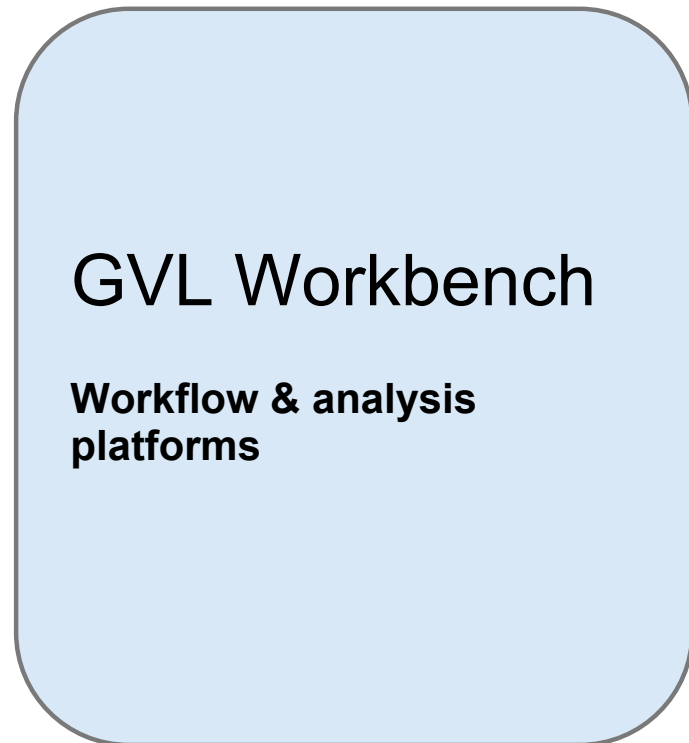
*Building (deploying and running) a GVL instance:*

- 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*

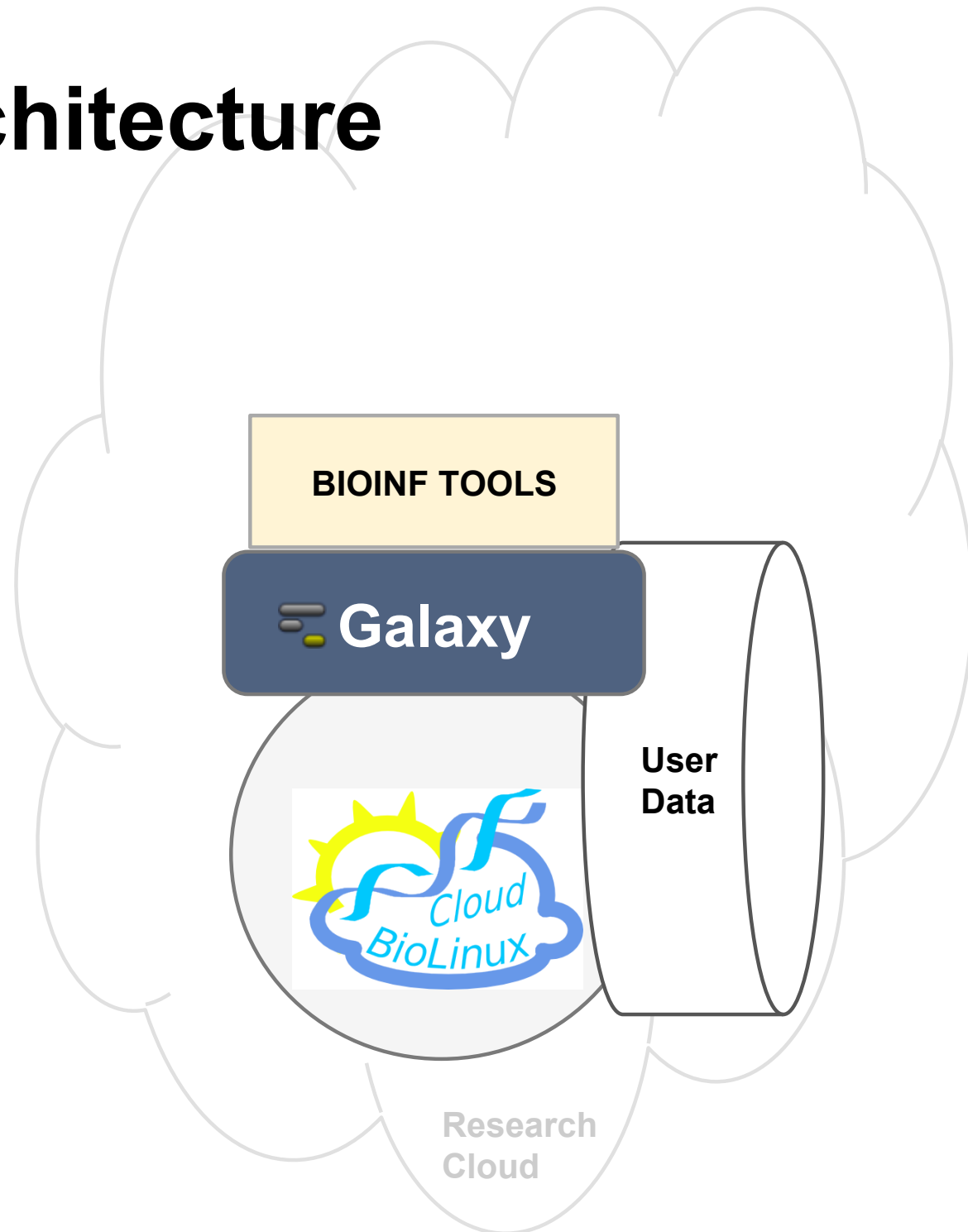
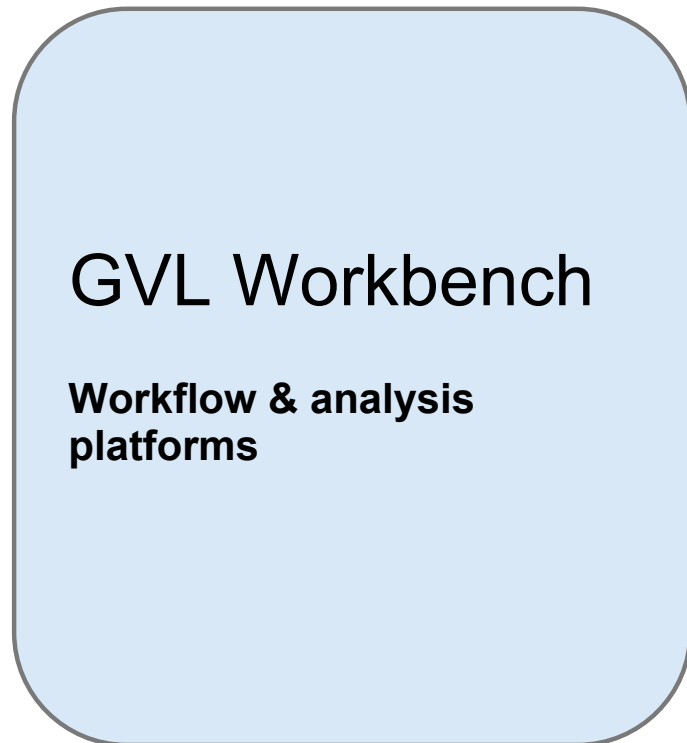
# GVL Workbench: Architecture



# Workbench: Architecture

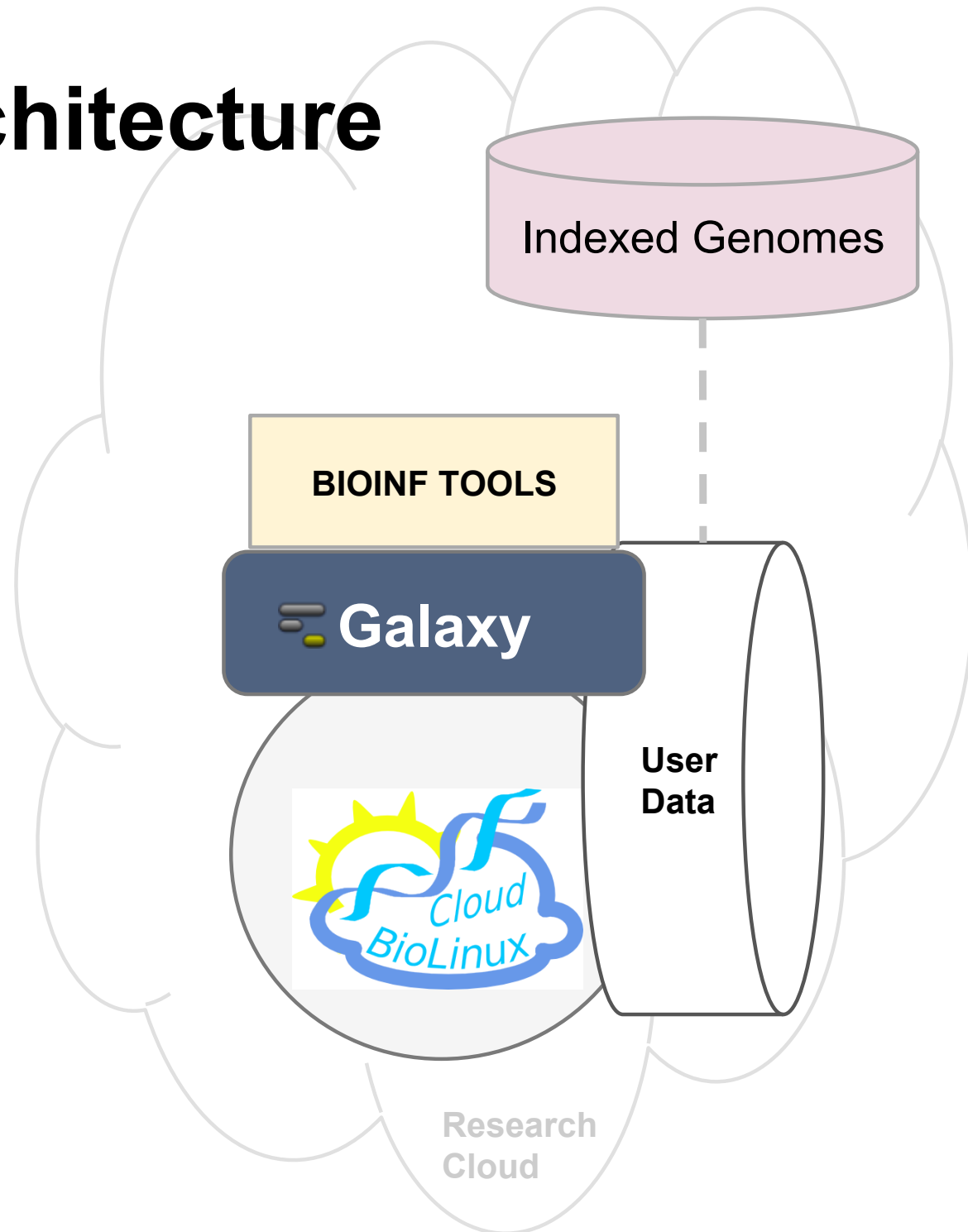
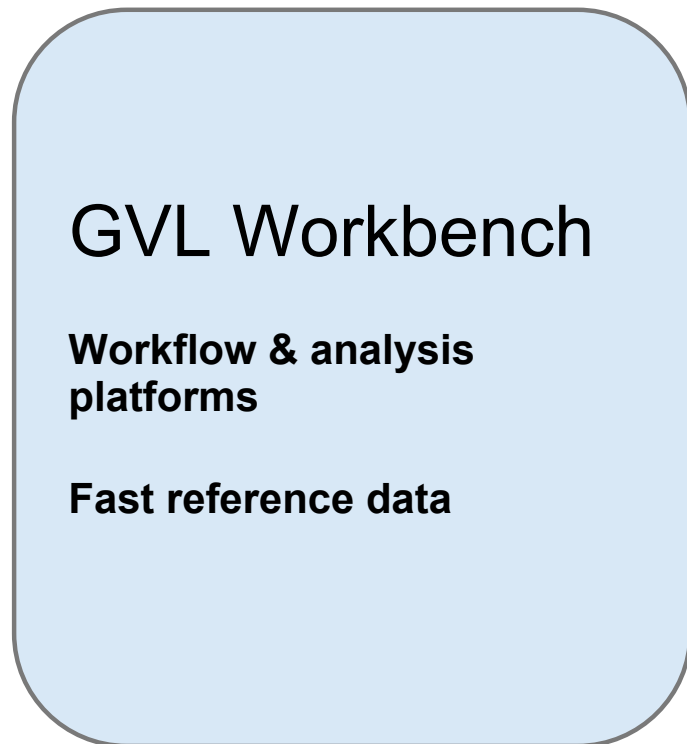


# Workbench: Architecture

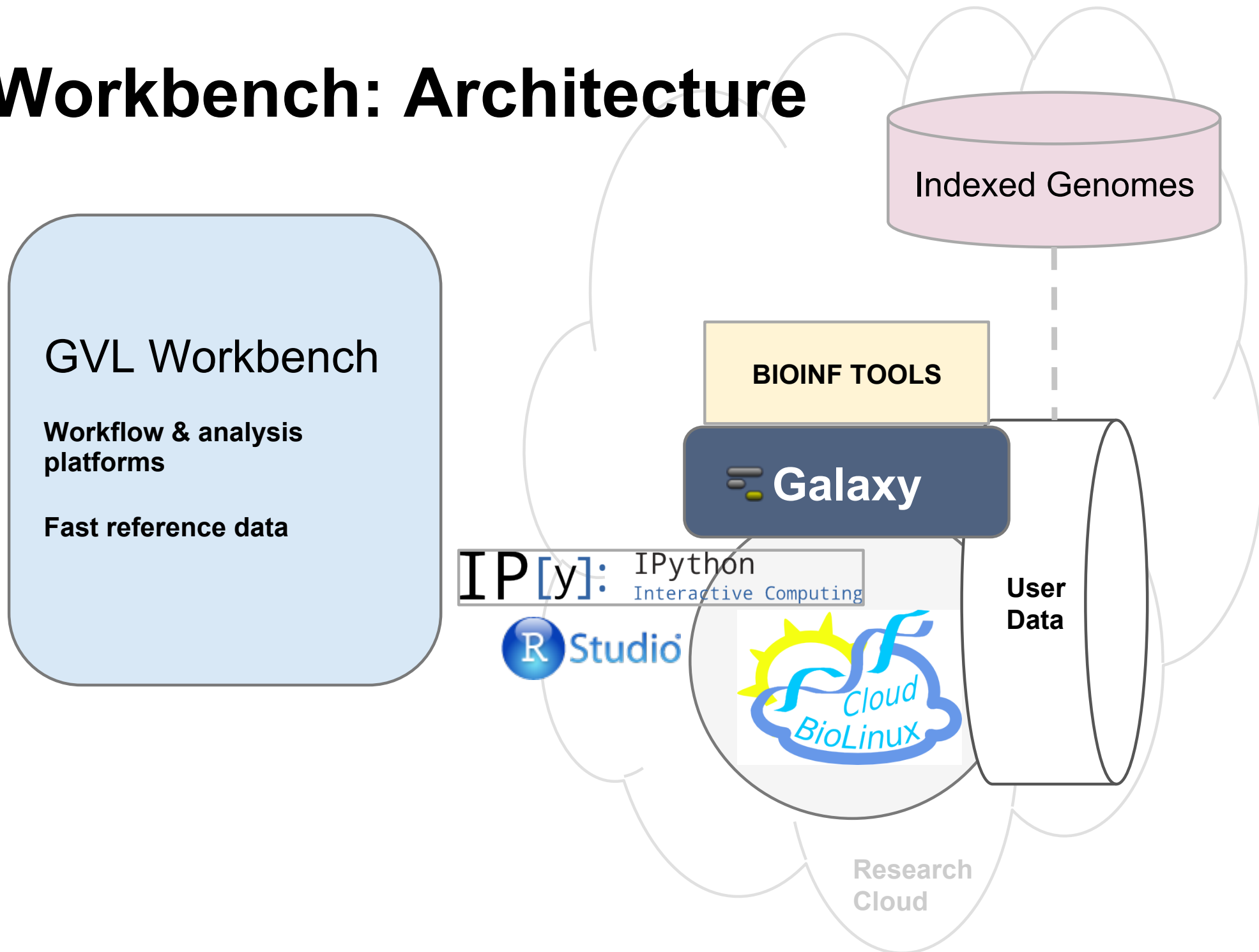




# Workbench: Architecture



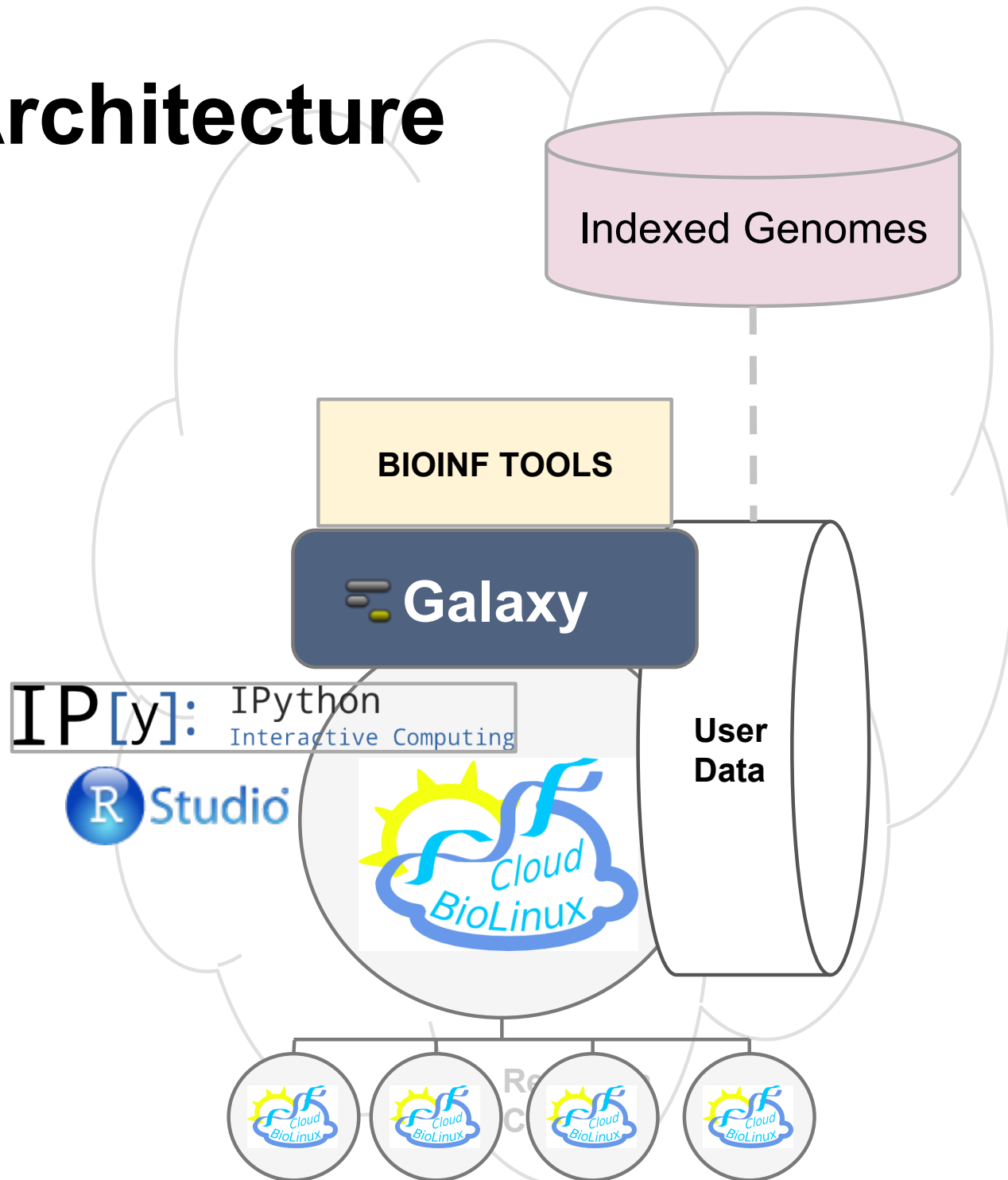
# Workbench: Architecture



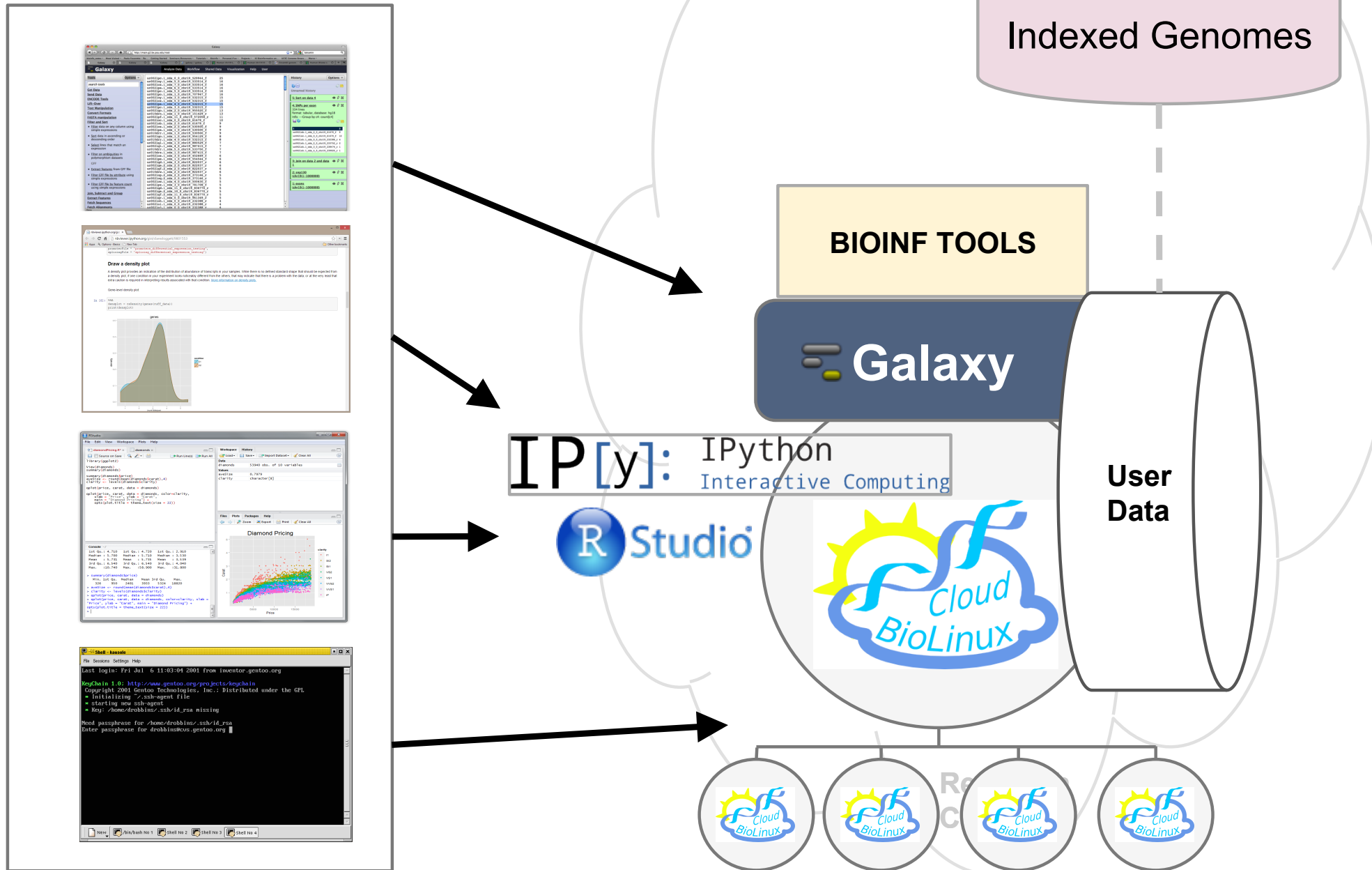
# Workbench: Architecture

**GVL Workbench**

- Workflow & analysis platforms
- Fast reference data
- Scalable compute



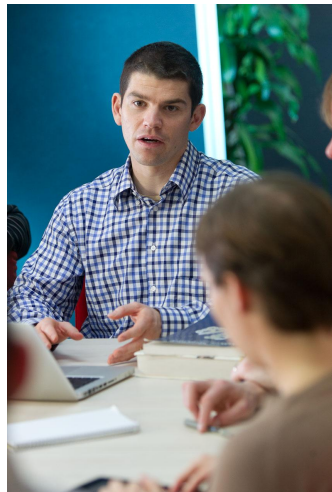
# Workbench: Architecture



# Engineering: 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>



SOFTWARE

Open Access

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

Enis Afgan<sup>1,3,4</sup> Brad Chapman<sup>2</sup> and James Taylor<sup>3,4</sup>

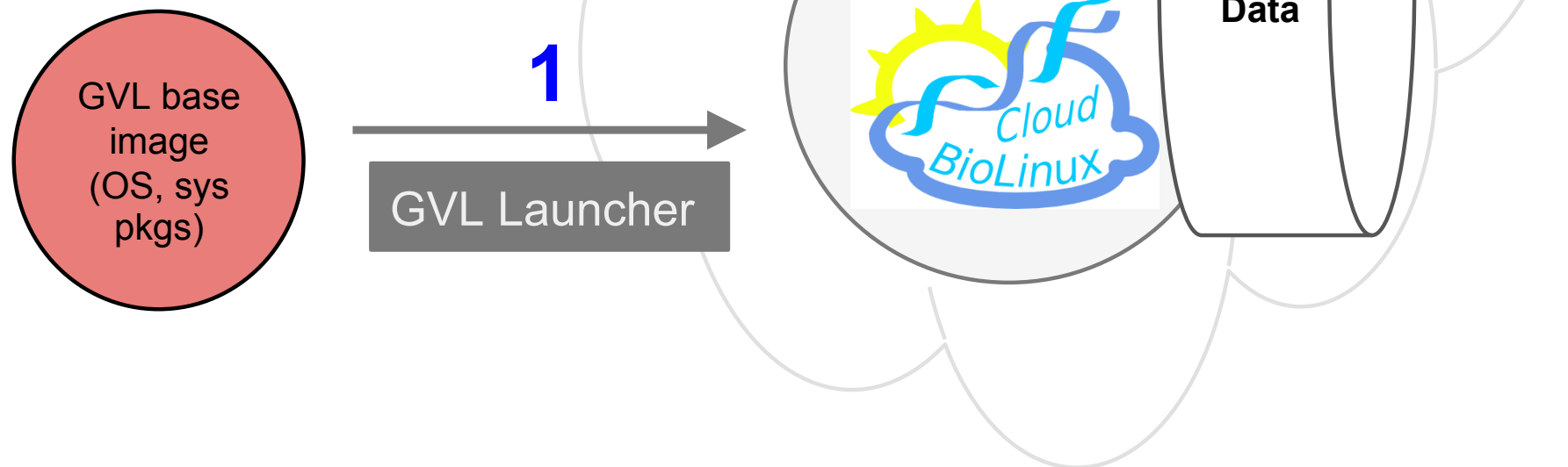
### Abstract

**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.

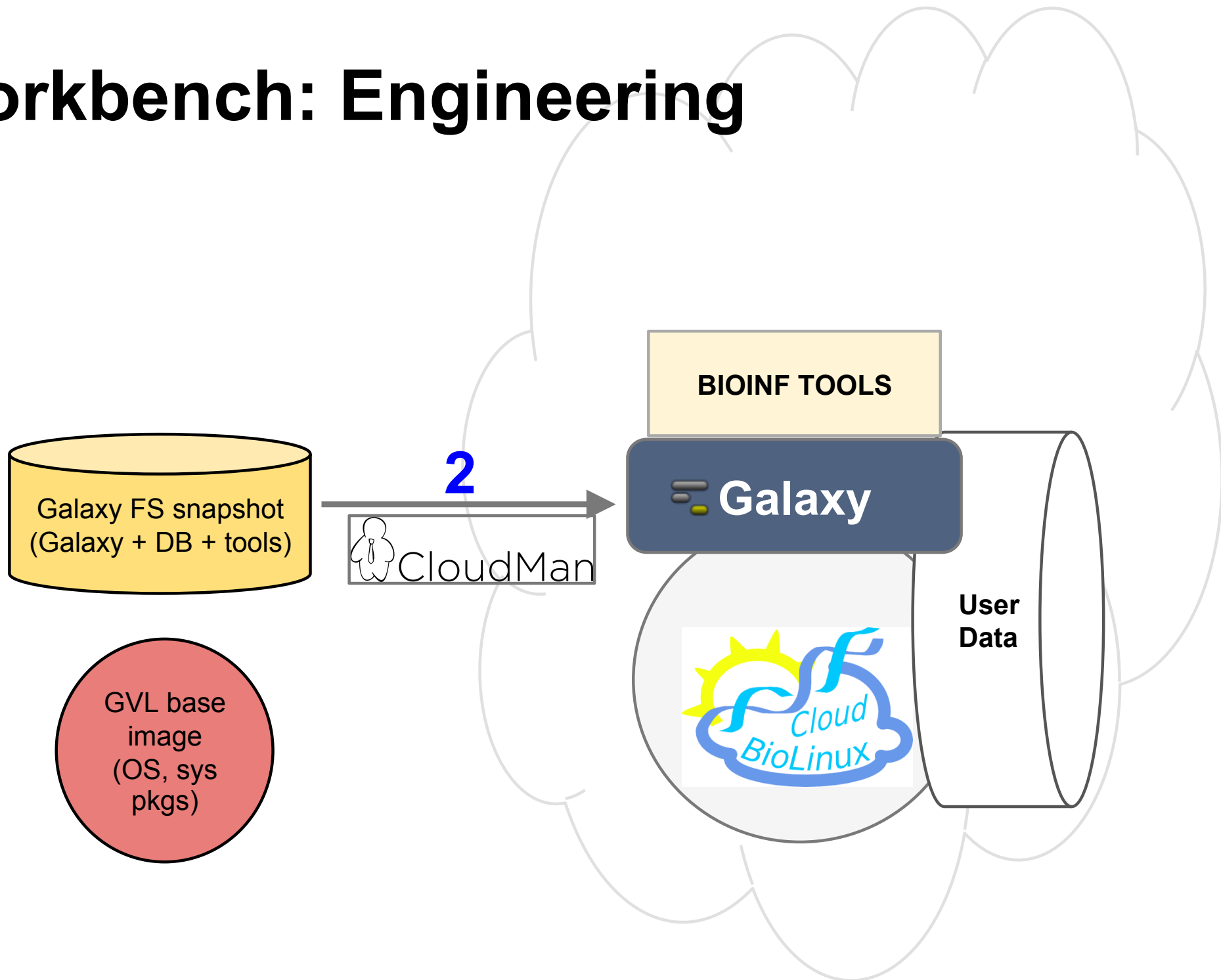
**Results:** CloudMan ([usecloudman.org](http://usecloudman.org)) enables individual researchers to easily deploy, customize, and share their entire cloud analysis environment, including data, tools, and configurations.

**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

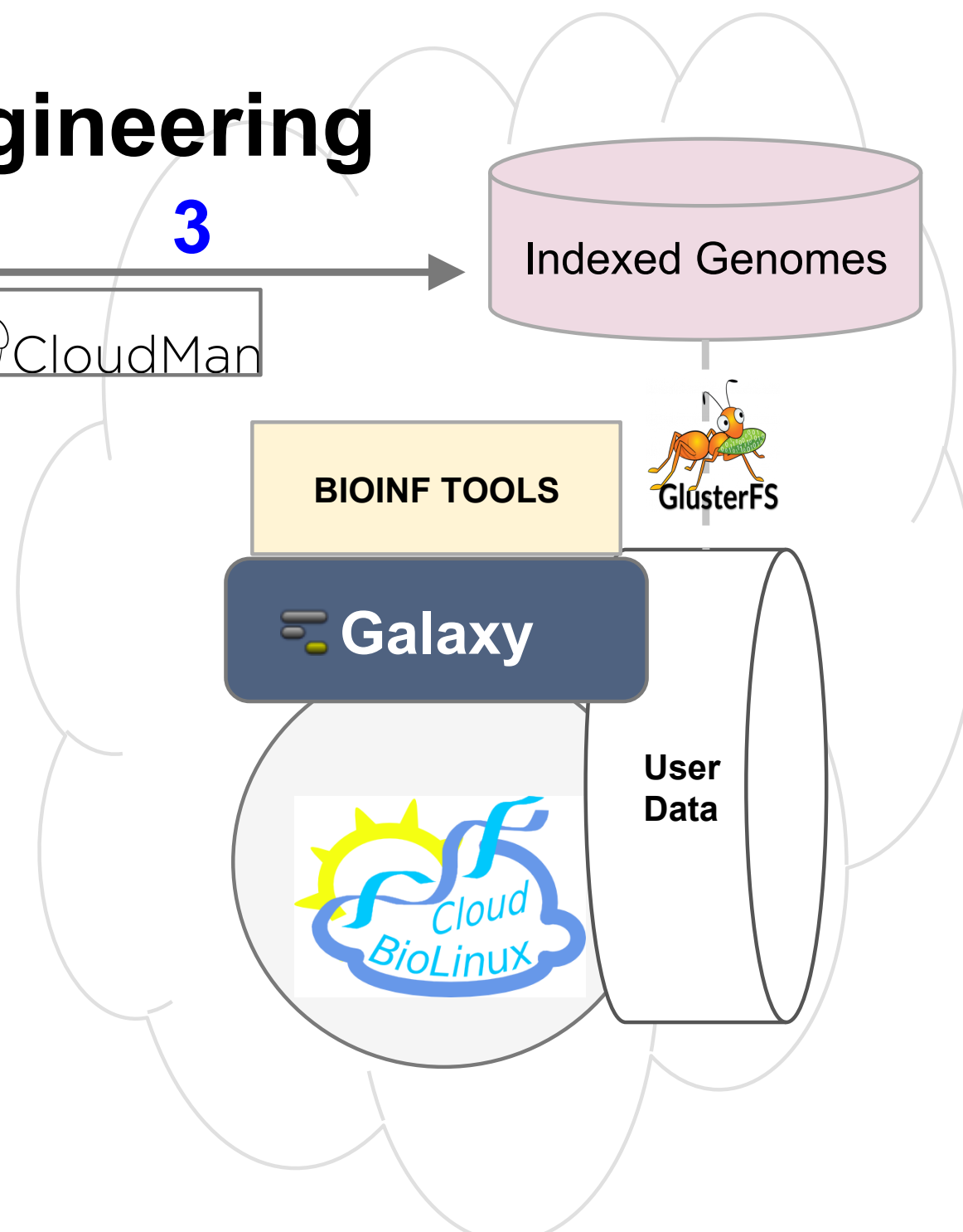
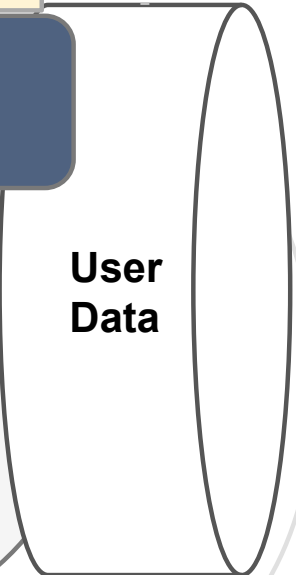
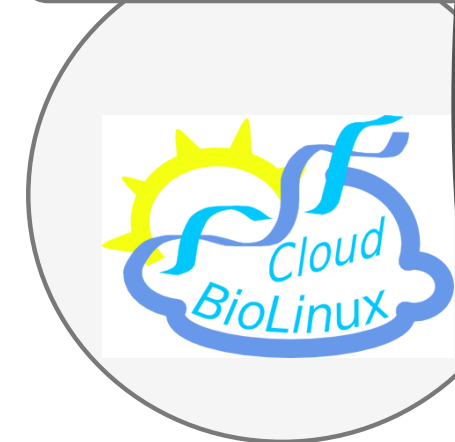
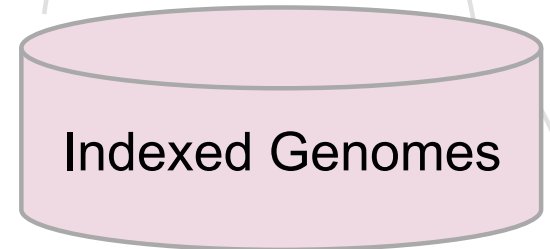
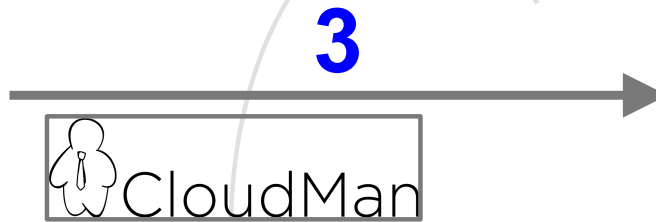
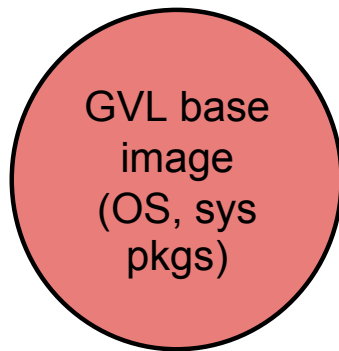
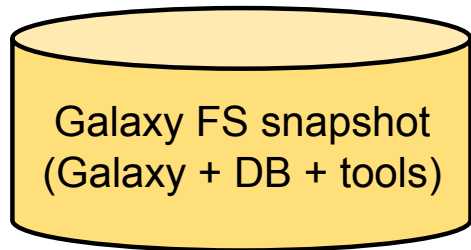
# Workbench: Engineering



# Workbench: Engineering

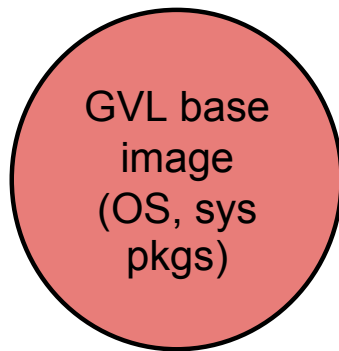
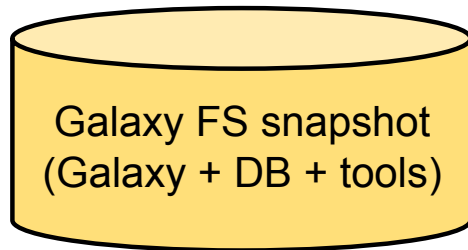


# Workbench: Engineering

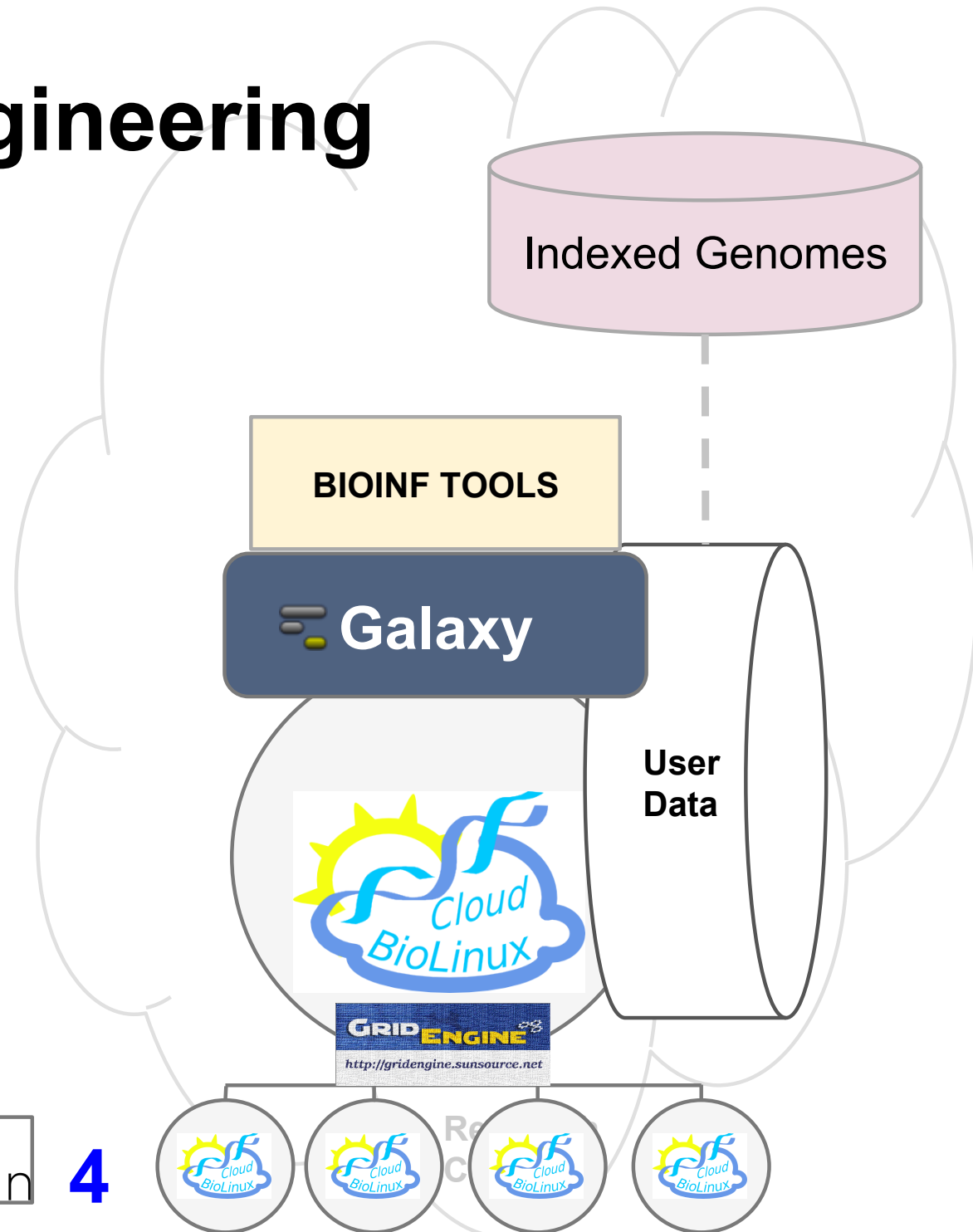




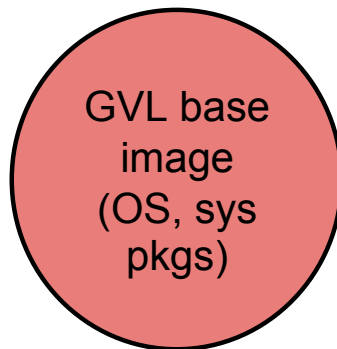
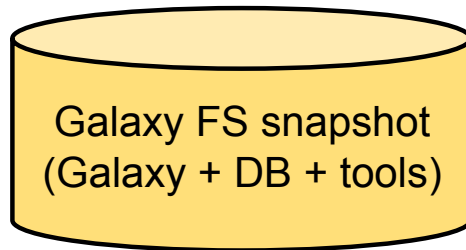
# Workbench: Engineering



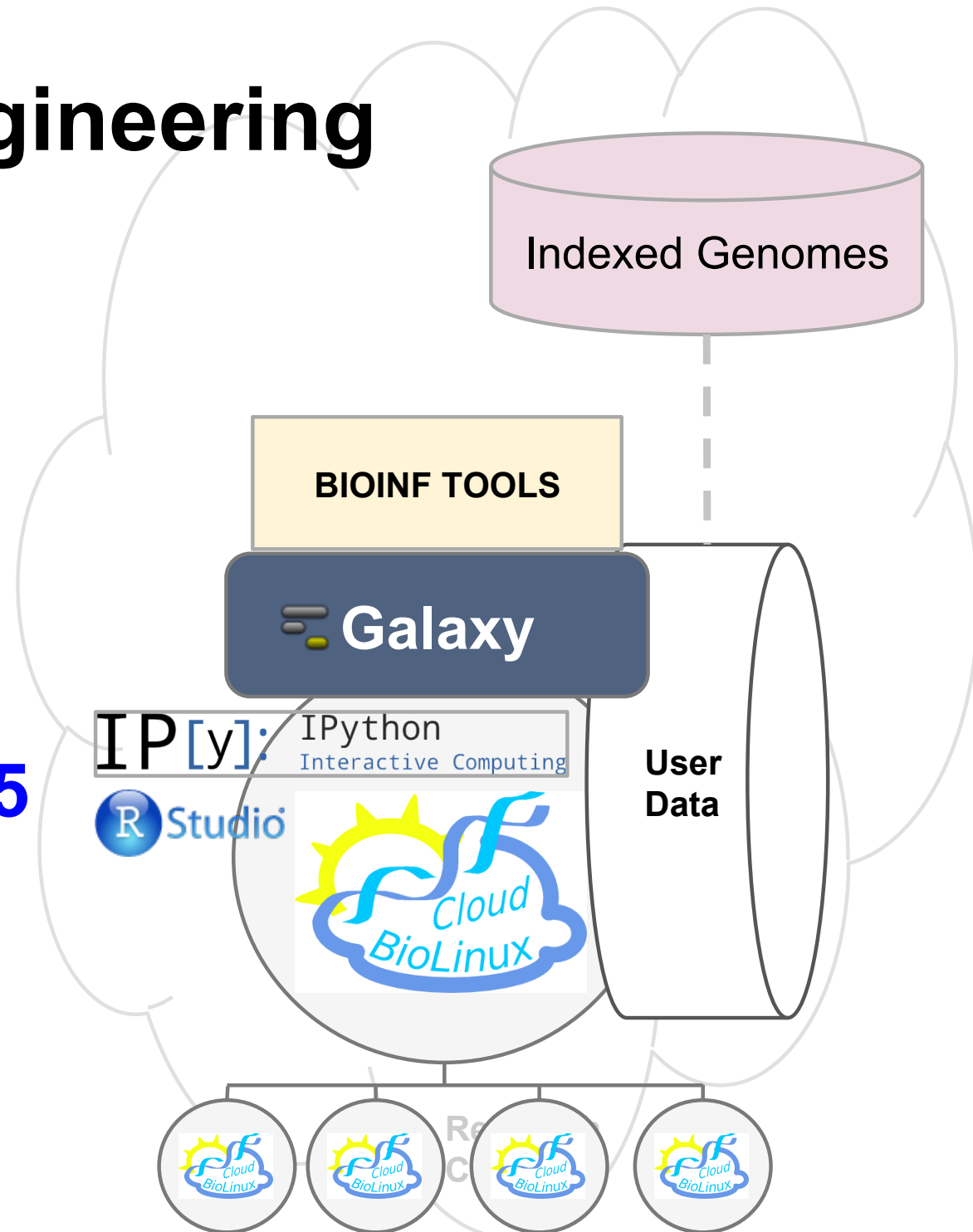
4



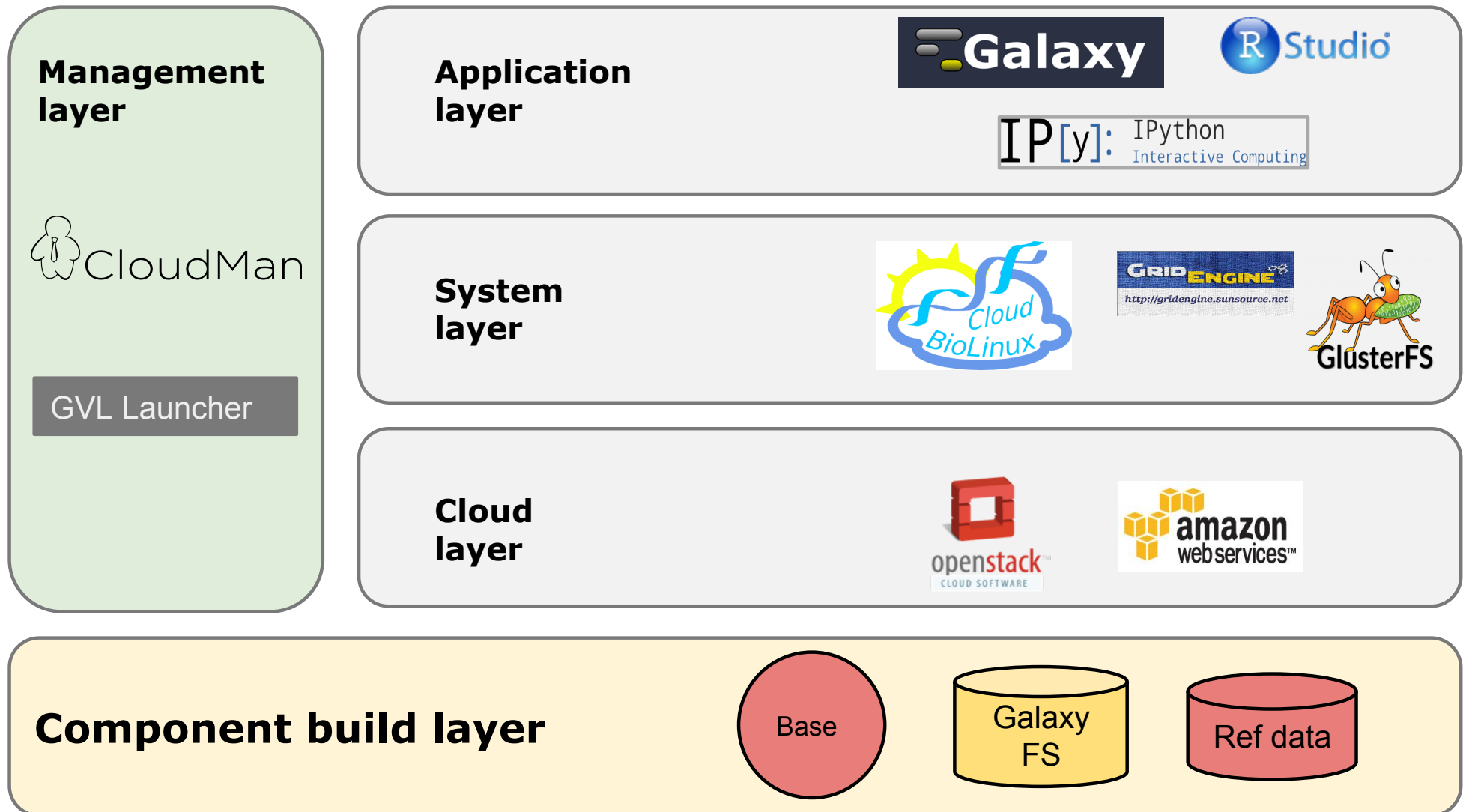
# Workbench: Engineering



5



# Workbench: All components



# **GVL: Does it work?**

**Technically?**

**Practically?**

<http://genome.edu.au> → GET



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


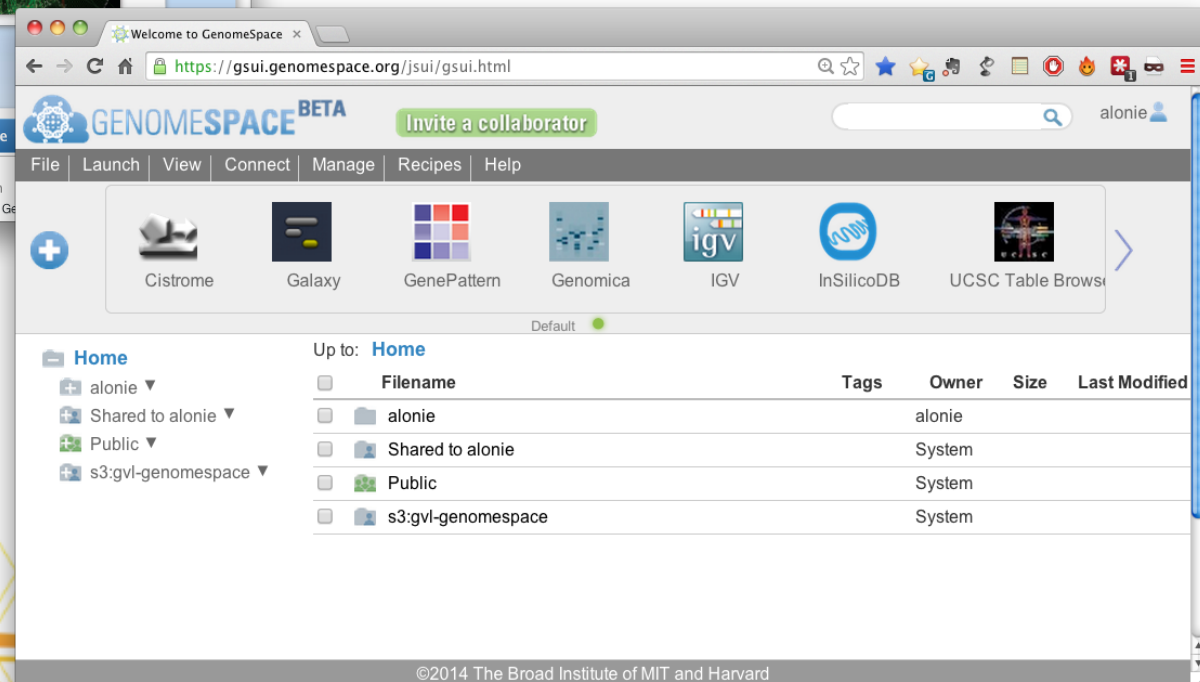


**Moving data around is a problem**

**Whole genomes: 300GB raw data**

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



Filename	Tags	Owner	Size	Last Modified
alonie		alonie		
Shared to alonie		System		
Public		System		
s3:gvf-genomespace		System		

# 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)