Zero to a Bioinformatics Analysis Platform in Four Minutes

Enis Afgan, Brad Chapman, Konstantinos Krampis, James Taylor

BOSC 2012
Long Beach, CA

Australian National Research Cloud

Provide computational infrastructure to support researchers needs

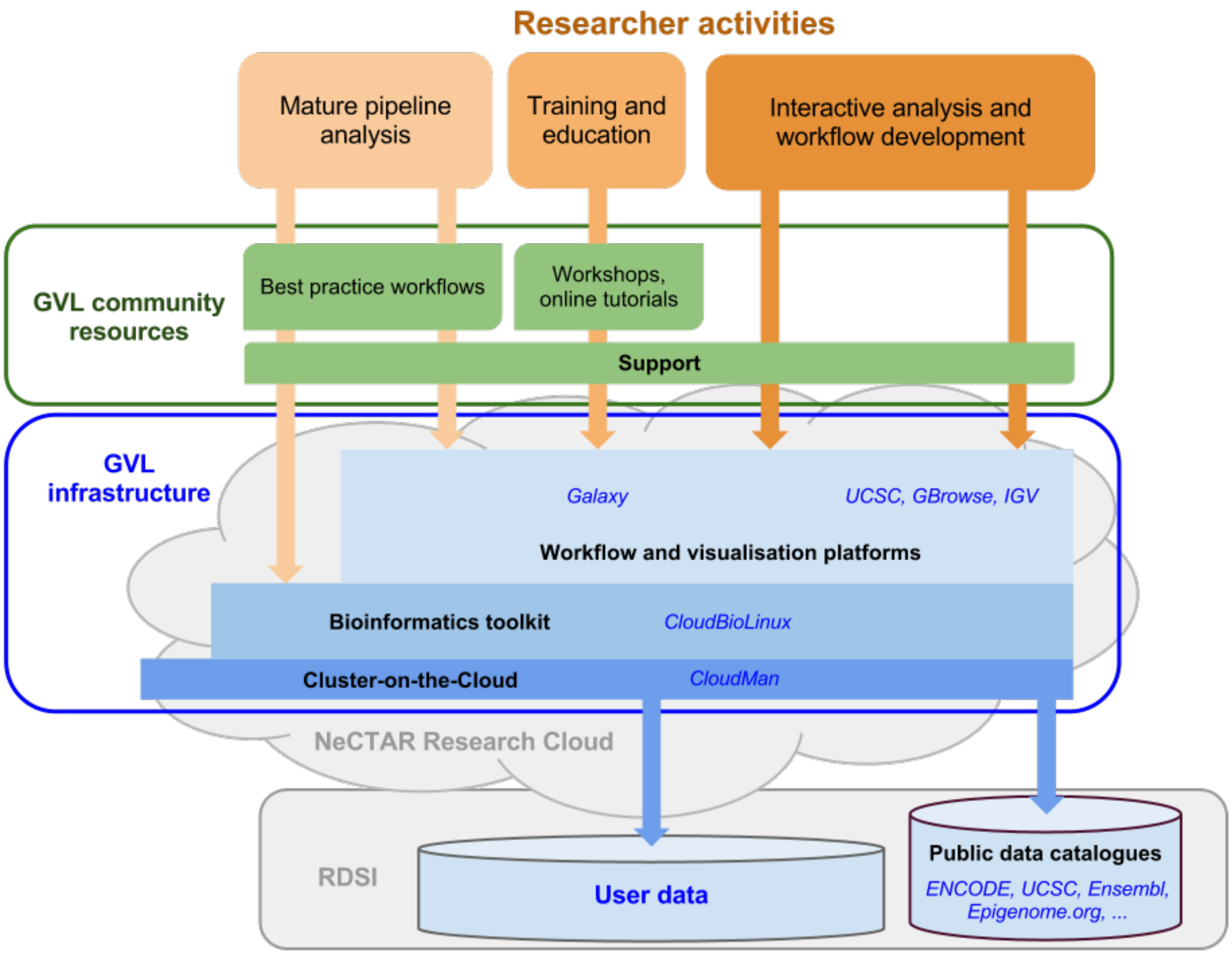
Compute and Storage
(~25,000 cores + ? PB)

nectar

What's required for genomics?

- ✔ Compute
- ✔ Storage
- Data resources
    - Ensembl, dbSNP, etc
- Tools
- Visualisation
- Protocols
- Expertise
- Community!

Genomics Virtual Lab



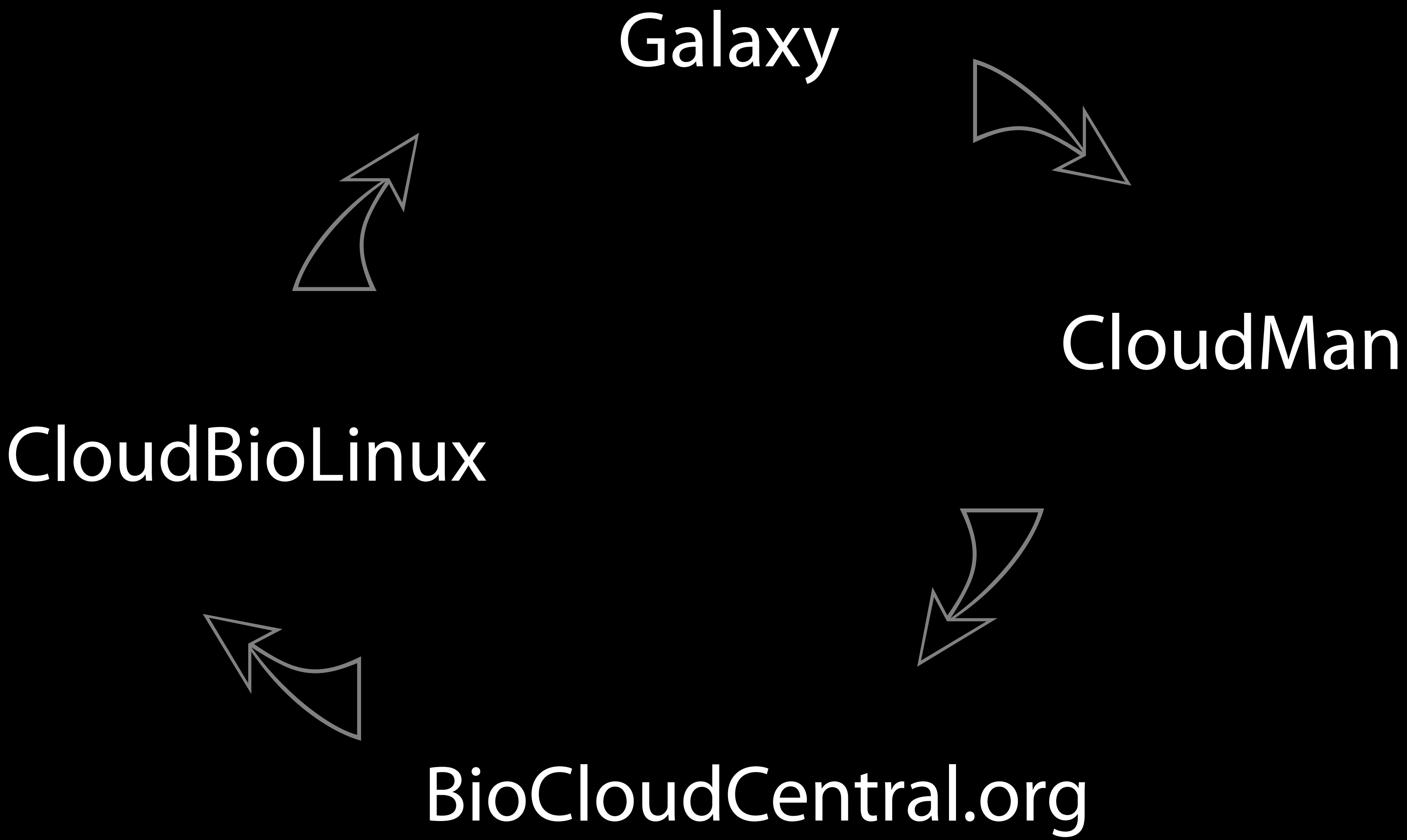
Compute + Storage = IaaS

shell vs. IDE

We want it now

What’s required for genomics?

- Compute
- Storage
- Data resources
    - Ensembl, dbSNP, etc
- Tools
- Visualisation
- Protocols
- Expertise
- Community!



Galaxy

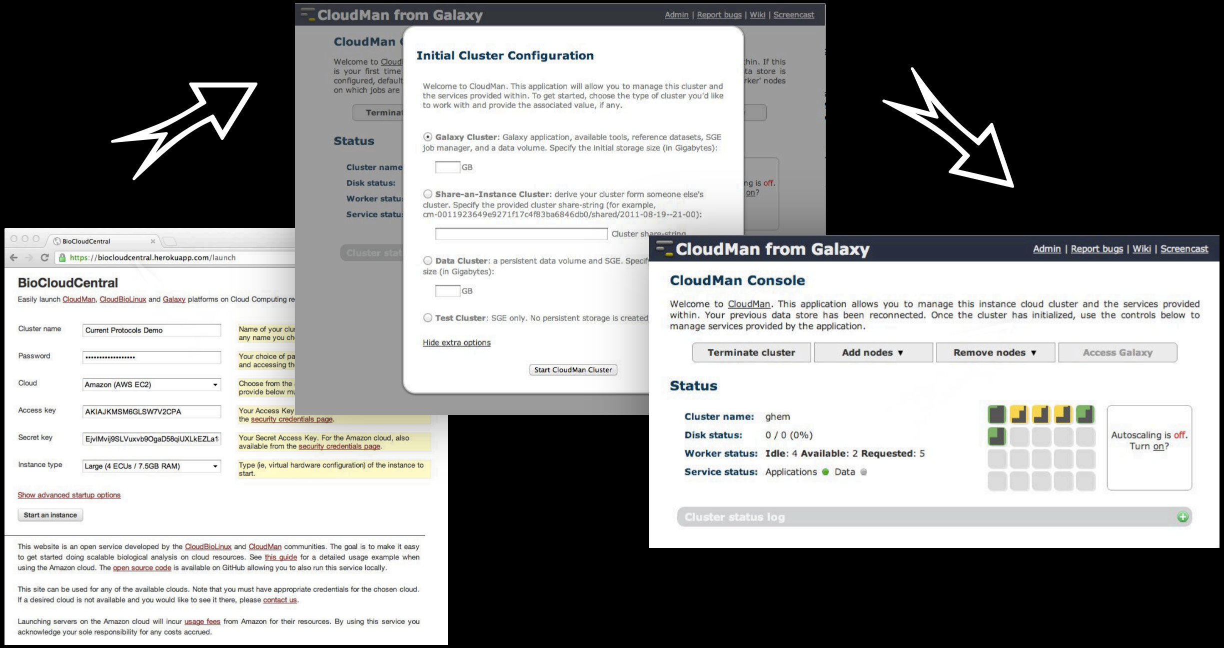
CloudMan

BioCloudCentral.org

CloudBioLinux

Playing together

- CloudBioLinux
    - Quickly build-your-own tool suite / ready to roll
    - Graphical & command line access
- CloudMan
    - Create a scalable and shareable processing platform
- Galaxy
    - Do exploratory analysis
- BioCloudCentral.org
    - Get started easily



BioCloudCentral

Easily launch CloudMan, CloudBioLinux and Galaxy platforms on Cloud Computing re

Cluster name	Current Protocols Demo	Name of your clu any name you ch
Password	••••••••••••••••••	Your choice of pa and accessing th
Cloud	Amazon (AWS EC2)	Choose from the provide below mu
Access key	AKIAJKMSM6GLSW7V2CPA	Your Access Key the security credentials page.
Secret key	EjvlMvij9SLVuxvb9OgaD58qiUXLkEZLa1	Your Secret Access Key. For the Amazon cloud, also available from the security credentials page.
Instance type	Large (4 ECUs / 7.5GB RAM)	Type (ie, virtual hardware configuration) of the instance to start.

This website is an open service developed by the CloudBioLinux and CloudMan communities. The goal is to make it easy to get started doing scalable biological analysis on cloud resources. See this guide for a detailed usage example when using the Amazon cloud. The open source code is available on GitHub allowing you to also run this service locally.

This site can be used for any of the available clouds. Note that you must have appropriate credentials for the chosen cloud. If a desired cloud is not available and you would like to see it there, please contact us.

Launching servers on the Amazon cloud will incur usage fees from Amazon for their resources. By using this service you acknowledge your sole responsibility for any costs accrued.

Initial Cluster Configuration

Welcome to CloudMan. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and provide the associated value, if any.

● **Galaxy Cluster**: Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage size (in Gigabytes):

☐ GB

○ **Share-an-Instance Cluster**: derive your cluster form someone else's cluster. Specify the provided cluster share-string (for example, cm-0011923649e9271f17c4f83ba6846db0/shared/2011-08-19--21-00):

☐ Cluster share-string

○ **Data Cluster**: a persistent data volume and SGE. Specify size (in Gigabytes):

☐ GB

○ **Test Cluster**: SGE only. No persistent storage is created

CloudMan Console

Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

Status

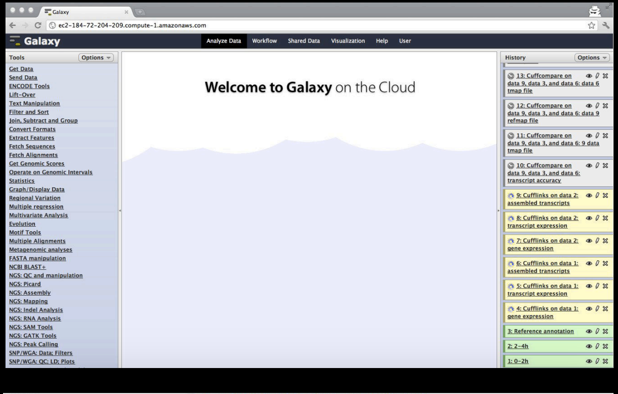
Cluster name: ghem

Disk status: 0 / 0 (0%)

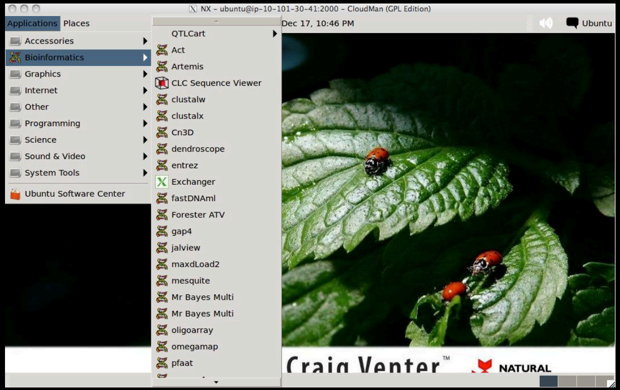
Worker status: **Idle**: 4 **Available**: 2 **Requested**: 5

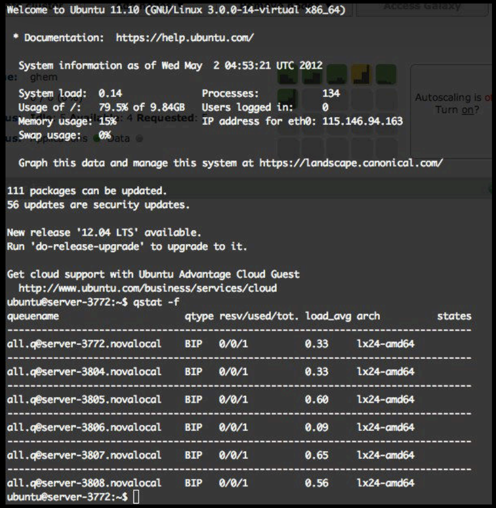
Service status: Applications ● Data ●

Autoscaling is off. Turn on?



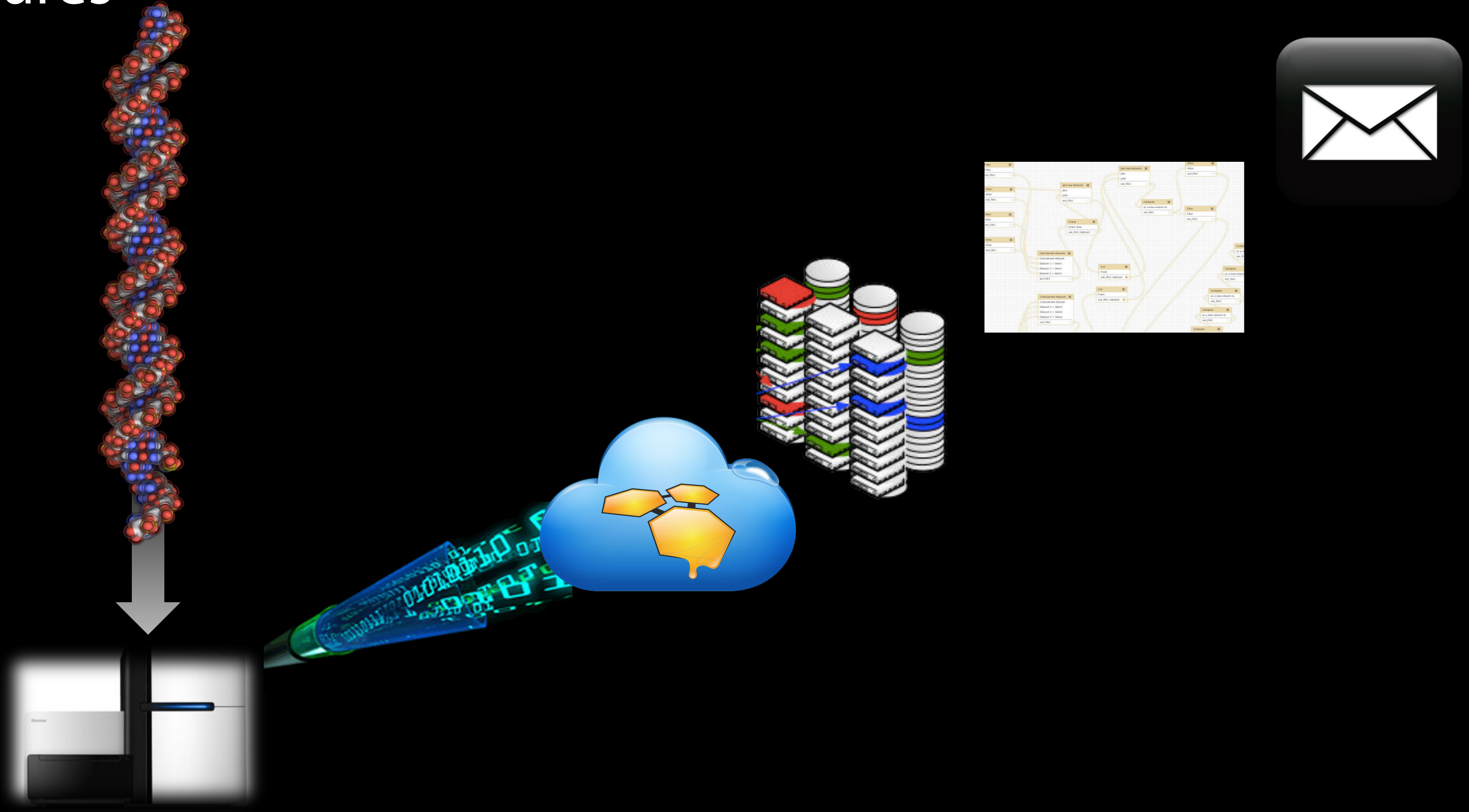
Welcome to Galaxy on the Cloud





```
Welcome to Ubuntu 11.10 (GNU/Linux 3.0.0-14-virtual x86_64)

 * Documentation:  https://help.ubuntu.com/

  System information as of Wed May  2 04:53:21 UTC 2012

  System load:  0.14              Processes:           134
  Usage of /:   79.5% of 9.84GB   Users logged in:     0
  Memory usage: 15%               IP address for eth0: 115.146.94.163
  Swap usage:   0%

  Graph this data and manage this system at https://landscape.canonical.com/

111 packages can be updated.
56 updates are security updates.

New release '12.04 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Get cloud support with Ubuntu Advantage Cloud Guest
  http://www.ubuntu.com/business/services/cloud
ubuntu@server-3772:~$ qstat -f
queuename                      qtype resv/used/tot. load_avg arch          states
---------------------------------------------------------------------------------
all.q@server-3772.novalocal    BIP   0/0/1          0.33     lx24-amd64
---------------------------------------------------------------------------------
all.q@server-3804.novalocal    BIP   0/0/1          0.33     lx24-amd64
---------------------------------------------------------------------------------
all.q@server-3805.novalocal    BIP   0/0/1          0.60     lx24-amd64
---------------------------------------------------------------------------------
all.q@server-3806.novalocal    BIP   0/0/1          0.09     lx24-amd64
---------------------------------------------------------------------------------
all.q@server-3807.novalocal    BIP   0/0/1          0.65     lx24-amd64
---------------------------------------------------------------------------------
all.q@server-3808.novalocal    BIP   0/0/1          0.56     lx24-amd64
ubuntu@server-3772:~$ 
```

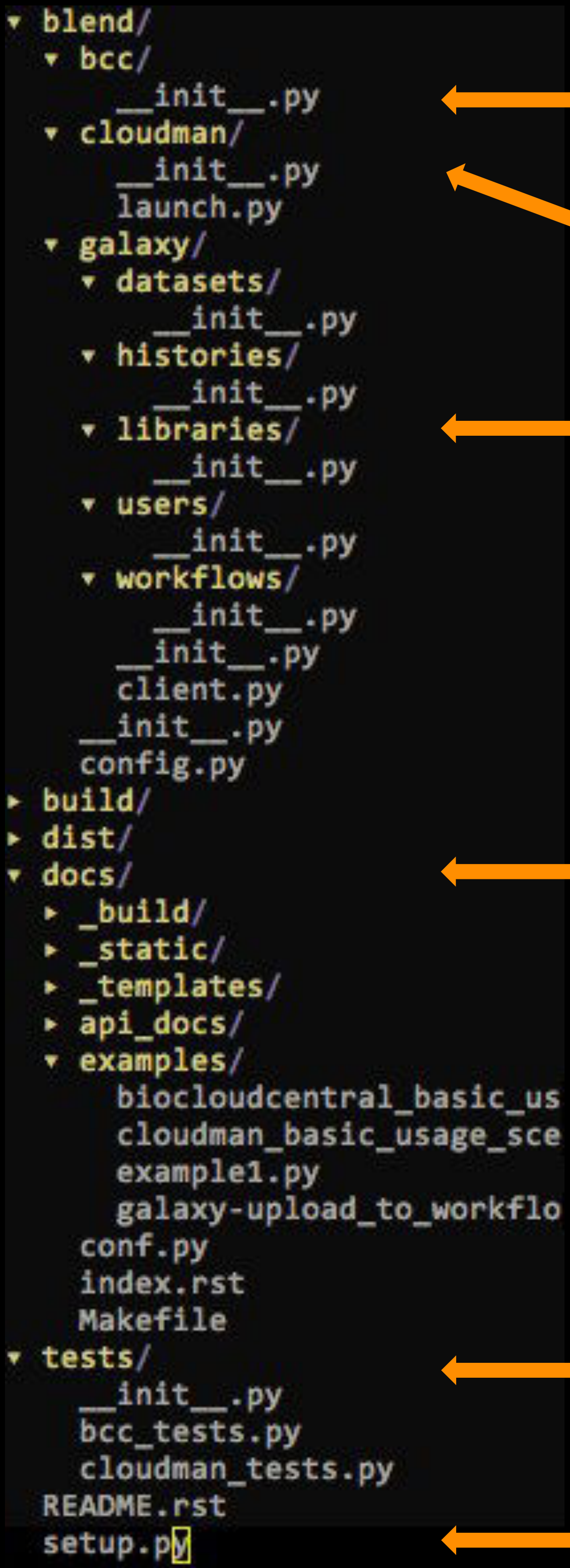
- Bundle infrastructure with an analysis tool suite, quickly
    - Validate our approach
    - Easier to maintain and replicate
- Expose it all via at a variety of interfaces
    - Support meta-analysis workflow
- Move forward
    - Add new features
    - Start using it



And one new thing…

blend

- A python library for interacting with Galaxy’s API
- And CloudMan
- And BioCloudCentral



```
▾ blend/
  ▾ bcc/
      __init__.py
  ▾ cloudman/
      __init__.py
      launch.py
  ▾ galaxy/
    ▾ datasets/
        __init__.py
    ▾ histories/
        __init__.py
    ▾ libraries/
        __init__.py
    ▾ users/
        __init__.py
    ▾ workflows/
        __init__.py
      __init__.py
      client.py
    __init__.py
    config.py
▸ build/
▸ dist/
▾ docs/
  ▸ _build/
  ▸ _static/
  ▸ _templates/
  ▸ api_docs/
  ▾ examples/
      biocloudcentral_basic_us
      cloudman_basic_usage_sce
      example1.py
      galaxy-upload_to_workflo
    conf.py
    index.rst
    Makefile
▾ tests/
    __init__.py
    bcc_tests.py
    cloudman_tests.py
  README.rst
  setup.py
```

Request compute infrastructure

Manipulate compute infrastructure

Upload data and run analyses

Docs and examples

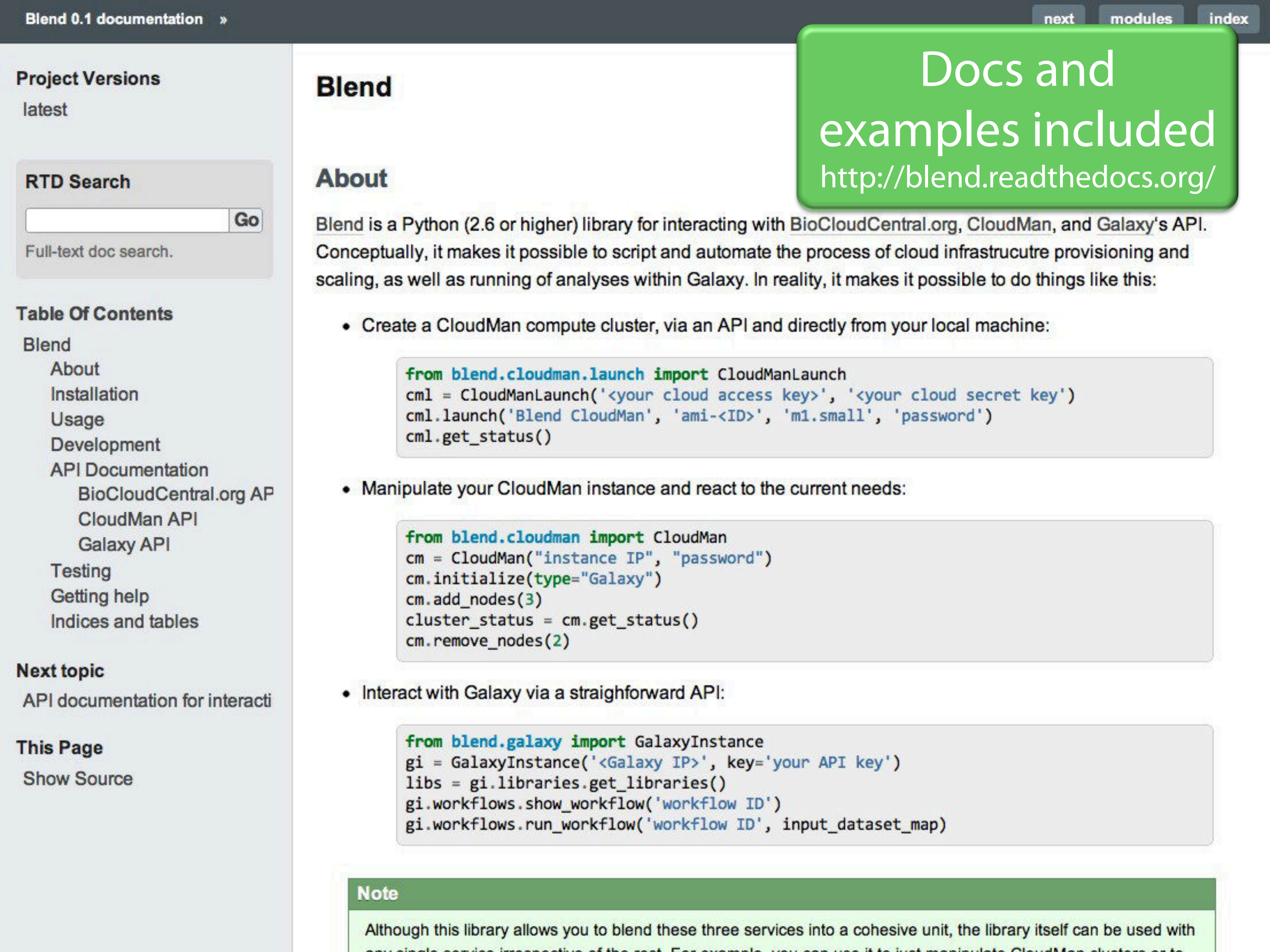
Test

Distribute

Automate
repetitive tasks

Docs and examples included

http://blend.readthedocs.org/



Blend

About

Blend is a Python (2.6 or higher) library for interacting with BioCloudCentral.org, CloudMan, and Galaxy's API. Conceptually, it makes it possible to script and automate the process of cloud infrastrucutre provisioning and scaling, as well as running of analyses within Galaxy. In reality, it makes it possible to do things like this:

- Create a CloudMan compute cluster, via an API and directly from your local machine:

```
from blend.cloudman.launch import CloudManLaunch
cml = CloudManLaunch('<your cloud access key>', '<your cloud secret key')
cml.launch('Blend CloudMan', 'ami-<ID>', 'm1.small', 'password')
cml.get_status()
```

- Manipulate your CloudMan instance and react to the current needs:

```
from blend.cloudman import CloudMan
cm = CloudMan("instance IP", "password")
cm.initialize(type="Galaxy")
cm.add_nodes(3)
cluster_status = cm.get_status()
cm.remove_nodes(2)
```

- Interact with Galaxy via a straighforward API:

```
from blend.galaxy import GalaxyInstance
gi = GalaxyInstance('<Galaxy IP>', key='your API key')
libs = gi.libraries.get_libraries()
gi.workflows.show_workflow('workflow ID')
gi.workflows.run_workflow('workflow ID', input_dataset_map)
```

Note

Although this library allows you to blend these three services into a cohesive unit, the library itself can be used with

Playing together

- CloudBioLinux
  - Build-your-own tool suite / ready to roll
  - Graphical & command line access
- CloudMan
  - Create a scalable and shareable processing platform
- BioCloudCentral.org
  - Get started easily
- Galaxy
  - Do exploratory analysis
- Blend library
  - Automate repetitive tasks: analysis AND infrastructure

Questions?

cloudbiolinux.org
usecloudman.org
usegalaxy.org
biocloudcentral.org
blend.readthedocs.org

Visit the poster session (poster #10)