

IFB Cloud Training

Using clouds and VMs in bioinformatics training
CSC ELIXIR-FI, 24 May 2016, Helsinki

Christophe BLANCHET,
Victoria DOMINGUEZ

Institut Français de Bioinformatique - IFB
French Institute of Bioinformatics - ELIXIR-FR
CNRS UMS3601 - Gif-sur-Yvette - FRANCE



Register

<http://www.france-bioinformatique.fr/en/cloud>

You are signed in as cblanchet | [Settings](#) | [Instances](#) | [Monitor](#) | [Help](#) | [Sign out](#)



Bioinformatics cloud



Registration

Enter Your Personal Information

Username ?	<input type="text"/>
E-mail ?	<input type="text"/>
Password ?	<input type="text"/>
Password (again) ?	<input type="text"/>
First name ?	<input type="text"/>
Last name ?	<input type="text"/>
City ?	<input type="text"/>
Affiliation ?	<input type="text"/>

Read Terms And Conditions Of Use

These resources are restricted to the academic community and not subject to financial fees but based on a fair use by all members. These resources are provided on an 'AS IS' basis, without warranty of any kind, either express or implied. In case these resources have been useful for your work, we kindly ask you to acknowledge the **French Institute of Bioinformatics (CNRS UMS3601)** in any publication or report, and to send us the reference at christophe(point)blanchet(at)france-bioinformatique.fr.

- **Fill the form**
 - ✓ With your official email address
- **Accept the Terms and Conditions**

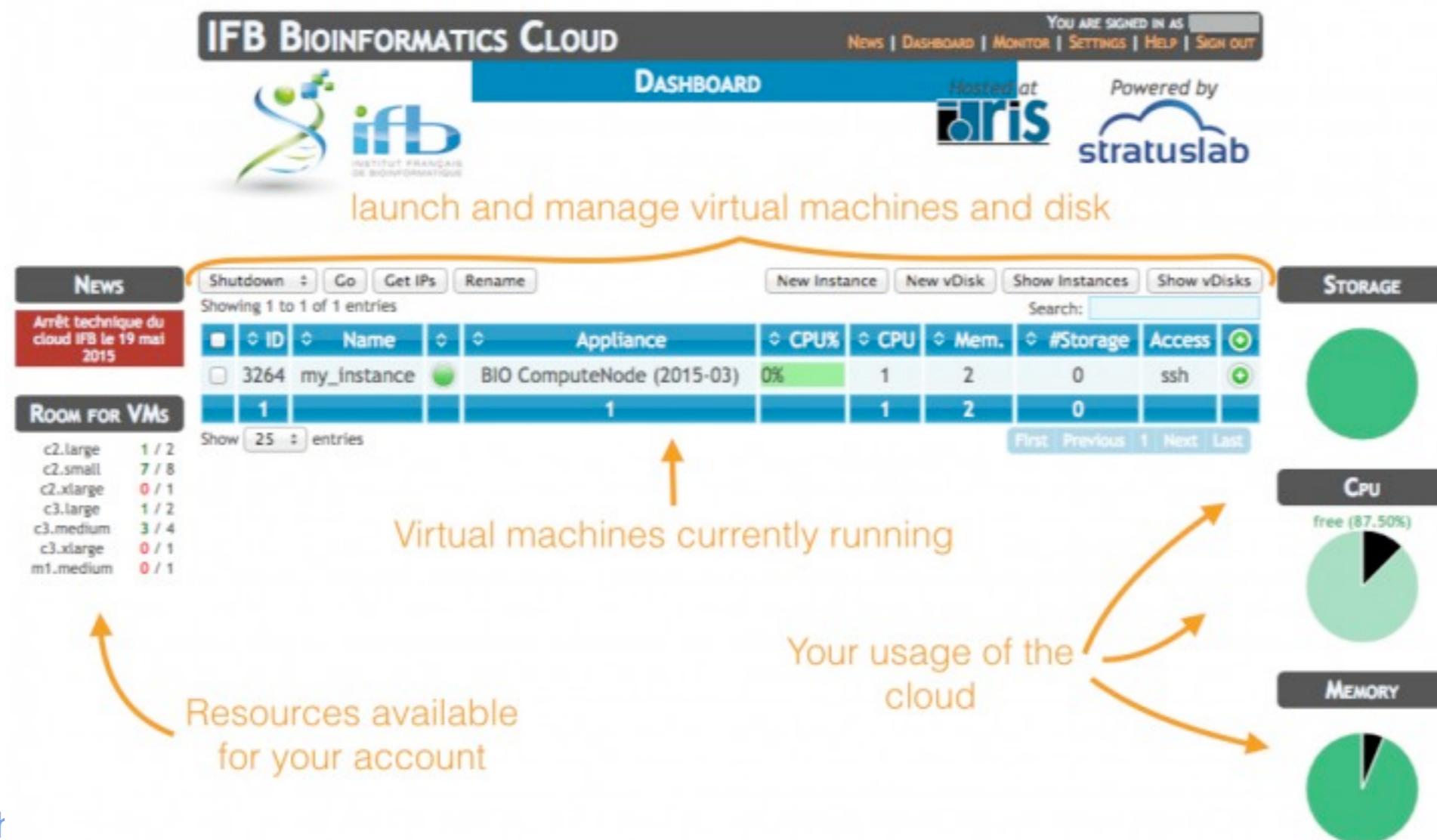
Sign in

<https://cloud.france-bioinformatique.fr/accounts/login/>

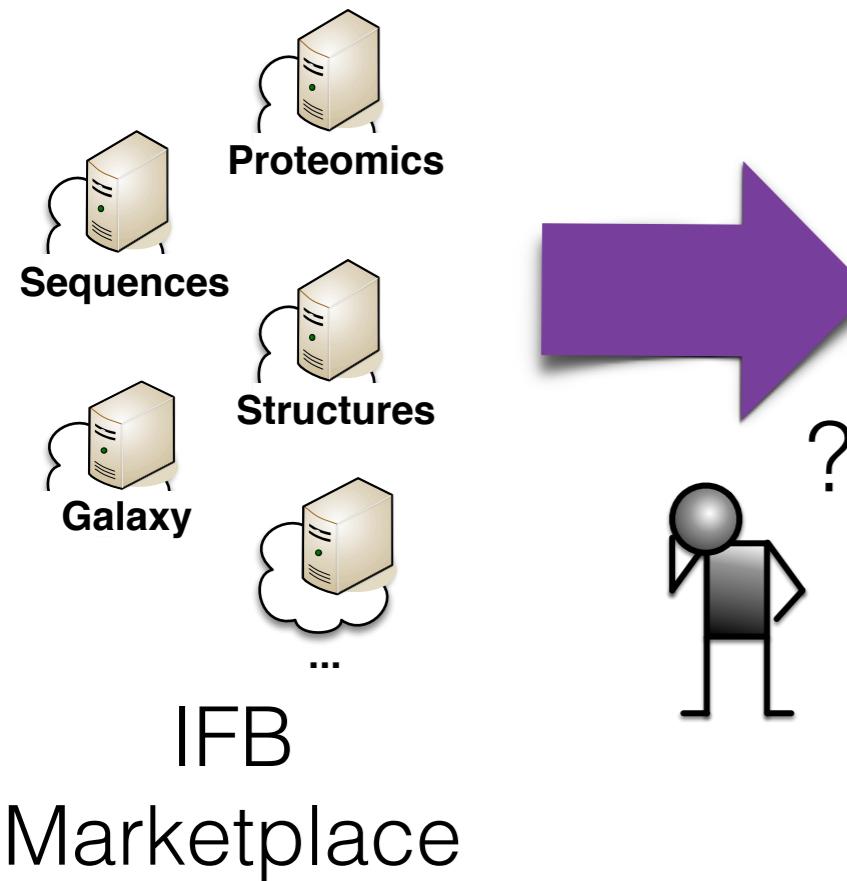


Cloud dashboard

- Manage your VMs
 - ✓ Create / shutdown / kill
- Manage your virtual disks
 - ✓ Create / delete
- View the configs
 - ✓ Name / size
 - ✓ Status / Parent image
 - ✓ Mounted vDisk



Run a virtual machine



Create Instance

Choose The Appliance
Appliance: Galaxy
Filter by: --- THEMATIC FIELDS ---
Galaxy portal

Configure Your Virtual Machines
Name: myPortal
Unique:
Type: c2.large (4 CPU, 8GB RAM)
Number: 1

Configure Your Storage
Persistent disk: [dropdown]

Create **Cancel**

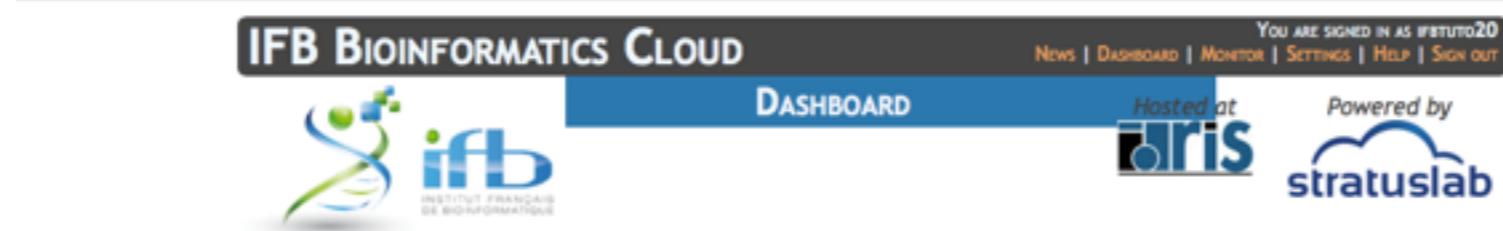
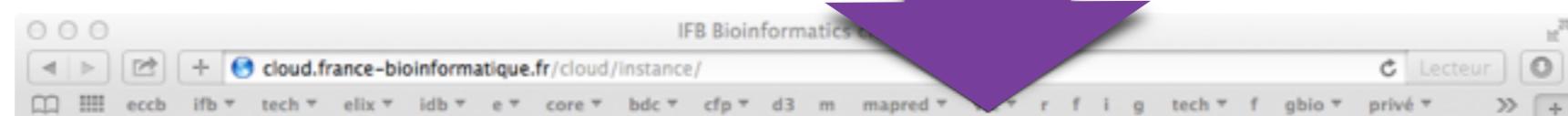
Create Instance

Choose The Appliance
Appliance: Galaxy MODAL
Filter by: --- THEMATIC FIELDS ---
Galaxy portal

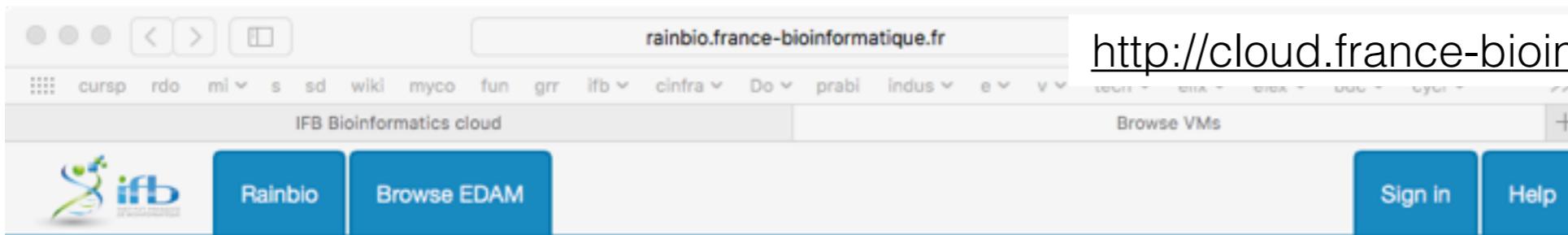
Configure Your Virtual Machines
Name: MODAL tools
Unique:
Type: c3.xlarge (16 CPU, 64GB RAM)
Number: 1

Configure Your Storage
Persistent disk: ECCB 10

Create **Cancel**



RAINBio



RAINBio - Registry of bioinformatics cloud appliances

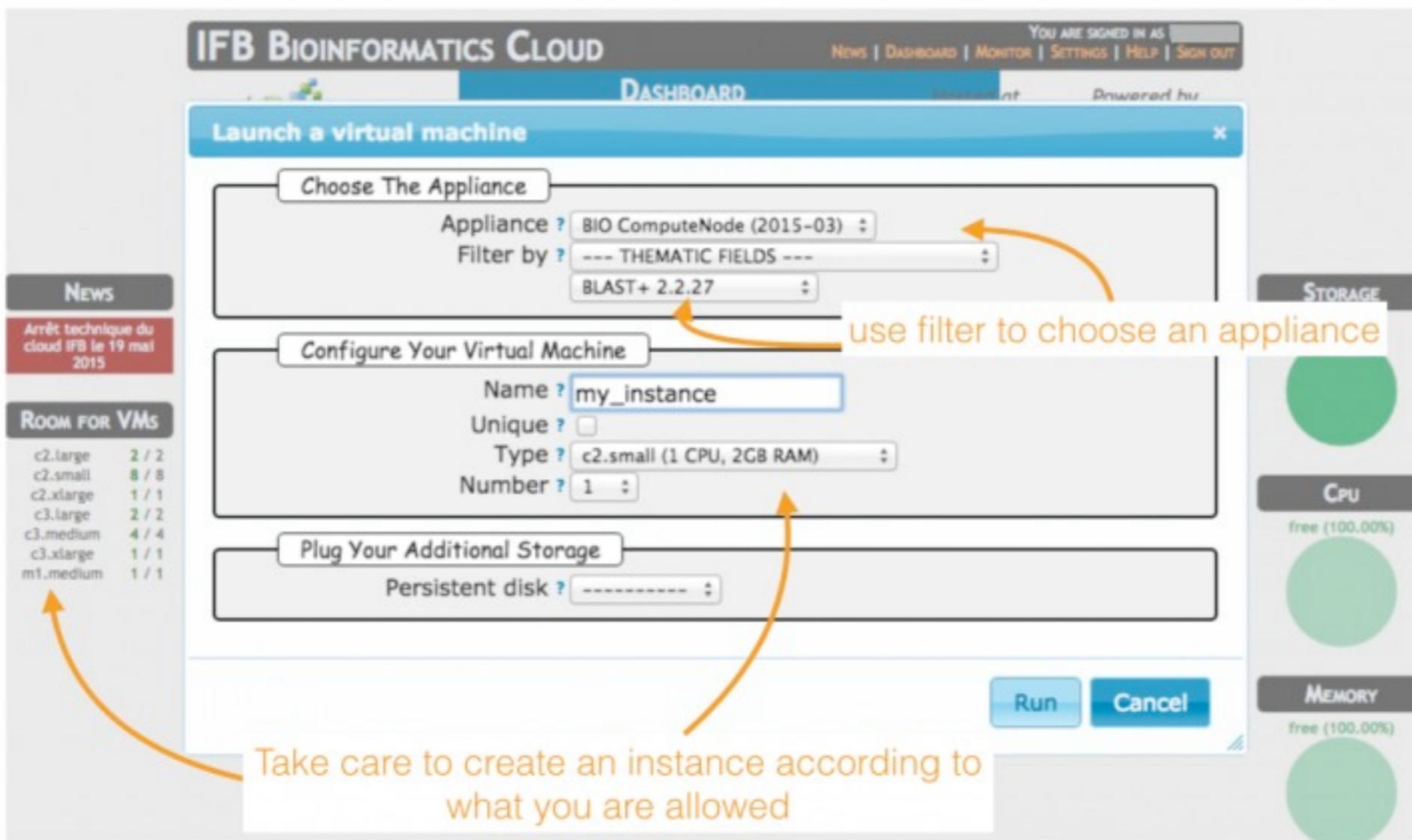
The appliances of the IFB cloud come with preconfigured bioinformatics tools, and both are annotated with terms from the topics dictionary of the EDAM ontology. To help you to find the right tools or appliances for your needs, you can browse the list of appliances and tools below and filter it with any terms.

Appliance	Tools	Topics
BIO compute node (3.3)	fastqc Bioconductor 2.11 TopHat MMSEQ 0.11.2a samtools 0.1.18 R muscle MultAlin 5.4.1 Clustal Omega Bowtie2 2.0.0-beta7 bowtie Ray BWA ABYSS predator MEME 4.7 HMMER fasta CLUSTALW 2.1 cap3 BLAST+ 2.2.27	Molecular biology Statistics and probability Bioinformatics Sequence analysis RNA splicing Mapping Data architecture, analysis and design Mathematics Sequence comparison Genomics Biology Protein structure Protein structure prediction Sequence composition, complexity and repeats Protein folds and structural domains Sequence sites, features Sequence assembly
Bacterial genomics (Insyght) (1.2)	Web interface BLAST+ 2.2.30 python 2.7 HMMER	Sequence composition, complexity and repeats Protein folds and structural domains Sequence sites, features Sequence comparison
Bio Imaging (1.1)	Bureau virtuel ImageJ	Data architecture, analysis and design Imaging

Metadata from

- ELIXIR bio.tools
- cloud marketplace
- docker hub
- ...

Define your VM

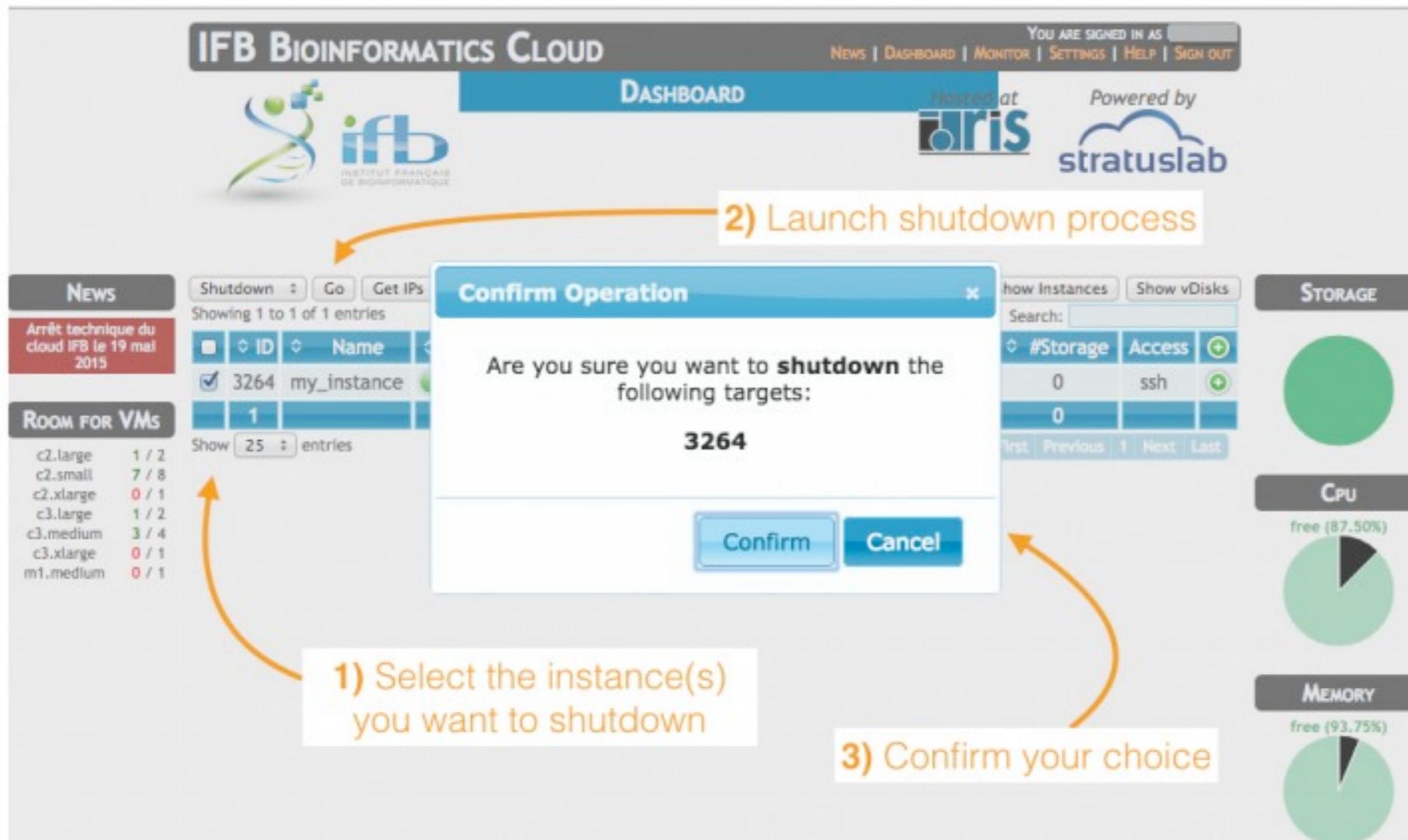


• Parameters

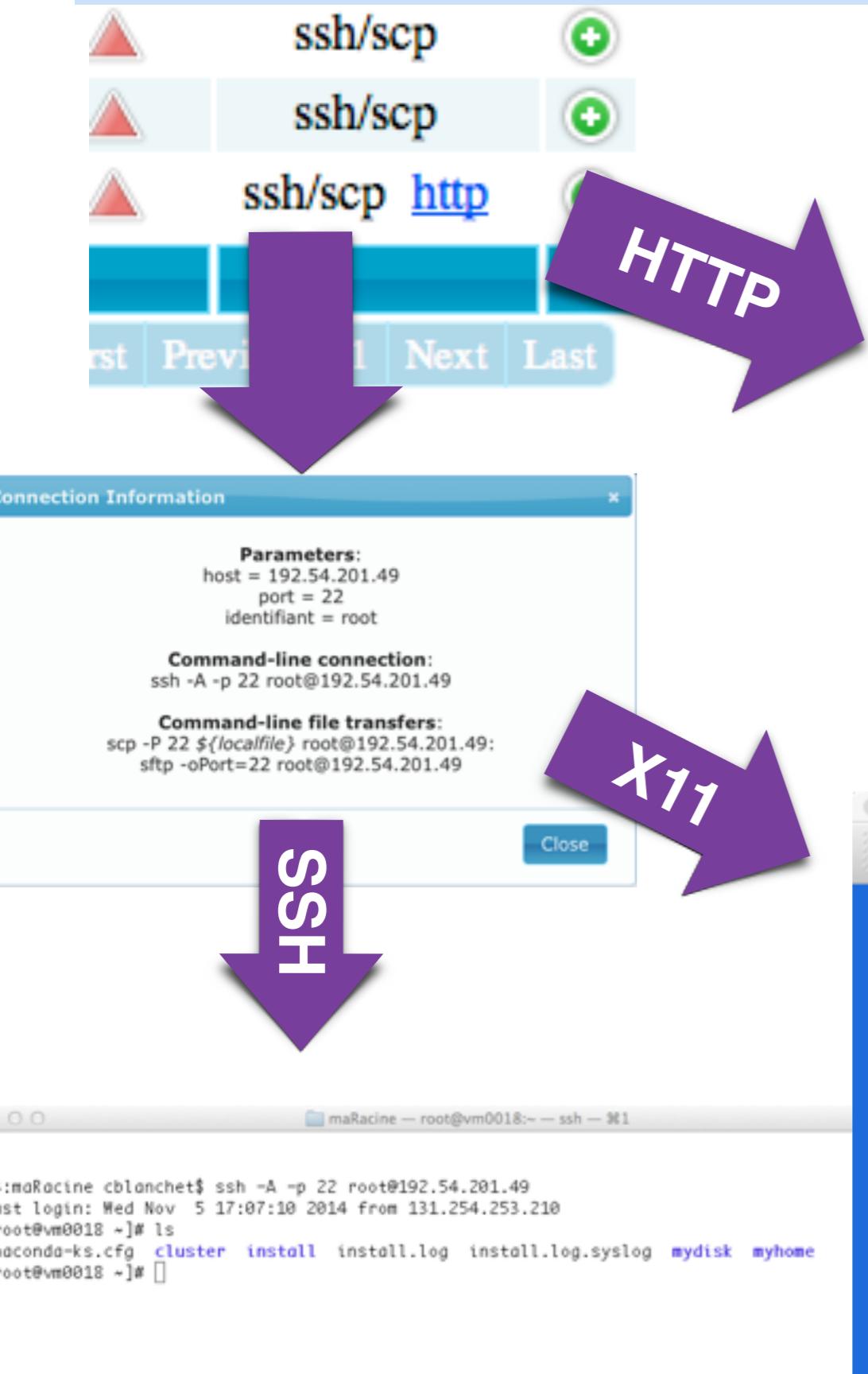
- ✓ Name
- ✓ CPUs
- ✓ Memory
- ✓ Attach a vDisk

Shutdown your VM

- Once you finished your data analysis



Connect to your VM



Galaxy

frontend.france-bioinformatique.fr/proxy/EcIGHX9p0sc5tkACdSC3YLFAMgAZpal.

IFB Bioinformatics cloud

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Index of /images/idb

History

Unnamed history 0 bytes

Your history is 'Data' on the left

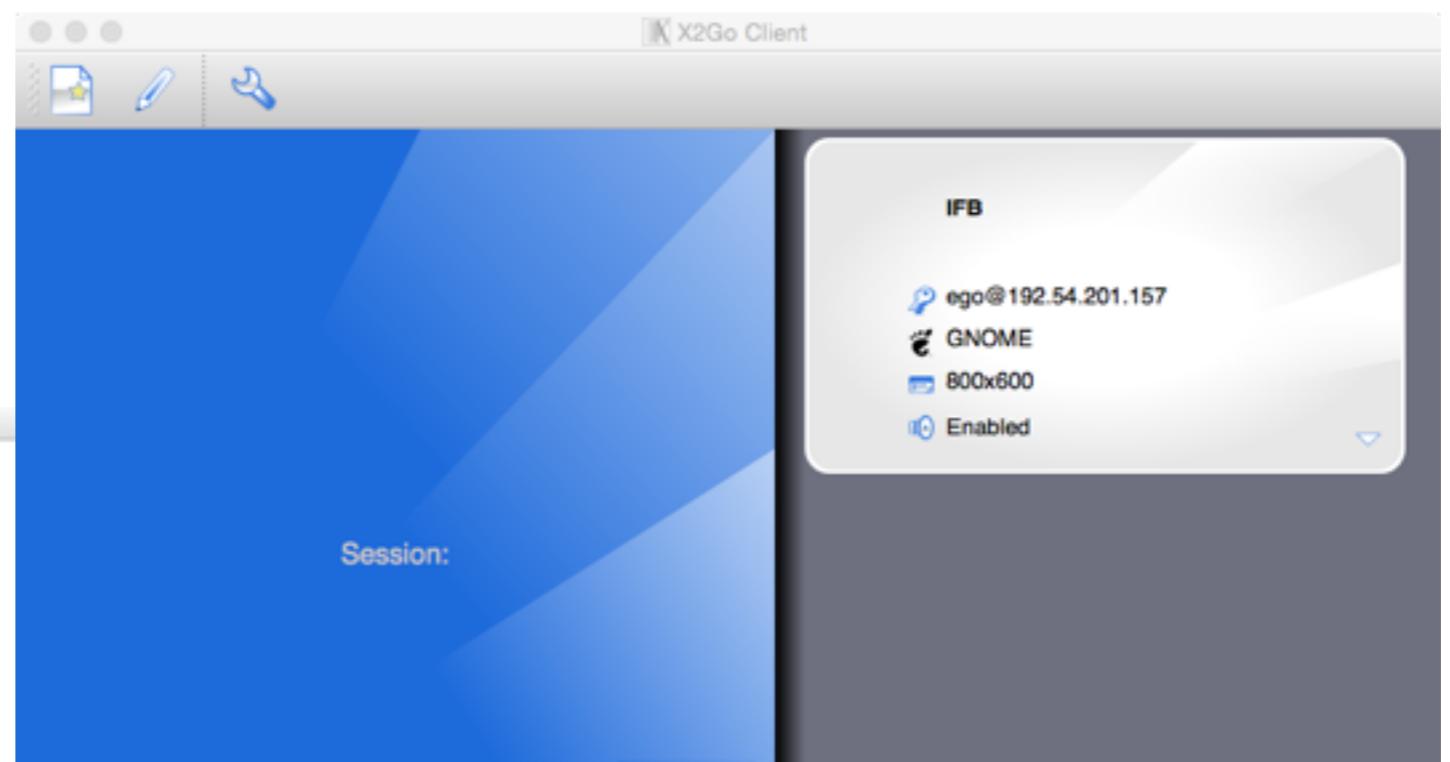
IFB - Galaxy cloud instance

Welcome to your Galaxy instance hosted on the cloud of the French Institute of Bioinformatics - IFB.

Usage

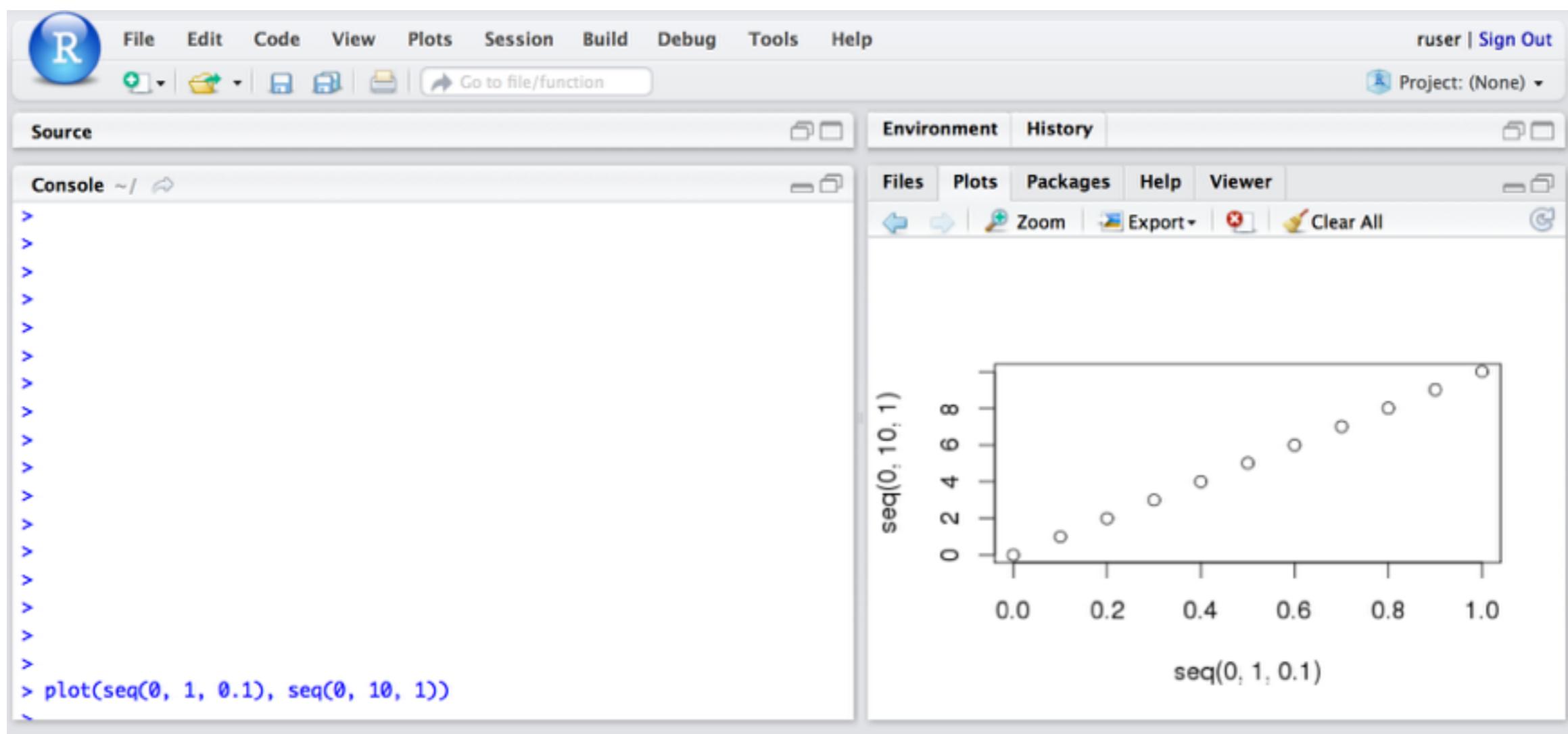
This appliance is configured with the well-known GALAXY portal. You connect to it with a standard web browser : simply follow the link [html](#) on the main cloud interface. It can be used as an usual galaxy portal and you have access to pre-installed standard bioinformatics tools (for new tools, send a request to christophe.blanchet@france-bioinformatique.fr).

(!) Do not forget to sign in with the pre-defined user : `user@cloud.idb.fr` (password `idbuser`) to get your previous history and data.



Practical - Studio

- a) Launch **R statistical computing**
- b) Connect to RStudio with the link ‘http’
- c) Sign in as ruser : ruser
- d) Run the example



Configure your account

<https://cloud.france-bioinformatique.fr/cloud/profile>

cloud.france-bioinformatique.fr

IFB BIOINFORMATICS CLOUD

YOU ARE SIGNED IN AS MBLANCHET

NEWS | DASHBOARD | MONITOR | SETTINGS | ADMINISTRATION | HELP | SIGN OUT

SETTINGS

Hosted at Powered by

Personal Information

Affiliation ? CNRS IFB

City ? Gif-sur-Yvette

Cloud Preferences

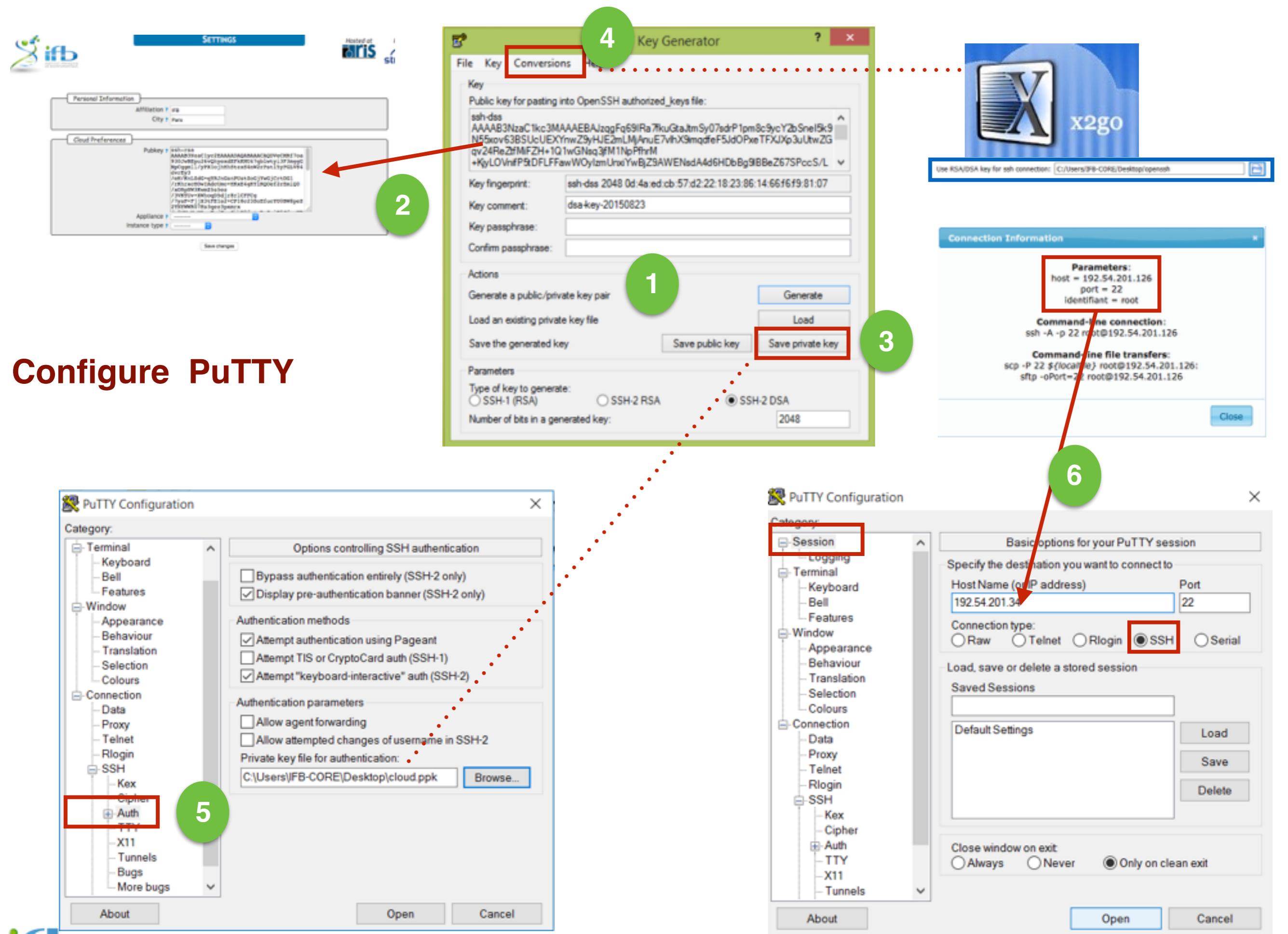
Pubkey ?
ssh-dss
AAAAB3NzaC1kc3MAAACBAMskM5bTOtlo/2EOeyXhbdF
GJuCVZQDssG8NidNs9IfKnRJ6/yVmqEBjouLCy4YY3ylh
YEVMZjVc5wYZgh9aRAzqFac3K+qXJMkZKX+taXaBbWP
PTXM49jo9+/ldr+rkJIN3fbBJT0wN8/ZZ9DFPMMDjVyAX1
nOd9VsKnTTvzGs3BPAAAAFQCKdicMAMjafqqKAszABQ
1s/Hs4aQAAAIA24WTy7RygmOuRTcrf9qu40L24rDfg9
R7N839KQH2Pry59Rv/vVByzJMG5lxKSBWZKCScg5d7S
J3H7TAiniWWxFFAaL7i2ZS2CIRUHBuGc9dptBGLjP6Vw
hC/aAANkKn0rWVaT7BScNzbI7adGTZ8He5tS57XoEnVf;

Appliance ? BIO ComputeNode (2015-03)

Instance type ? c2.small

Save changes

- Link ‘Settings’
- Fill your personal data
- Fill your public SSH key
 - ✓ ssh-keygen -t dsa
- Choose defaults parameters
 - ✓ Appliance
 - ✓ Instance type



Configure PuTTY

SSH

The screenshot shows the IFB Bioinformatics Cloud dashboard. In the center, a modal window titled "Connection Information" displays connection parameters and command-line instructions. The parameters are:

Parameters:
host = 192.54.201.39
port = 22
identifiant = root

Command-line connection:
ssh -A -p 22 root@192.54.201.39

Command-line file transfers:
scp -P 22 \${localfile} root@192.54.201.39:
sftp -oPort=22 root@192.54.201.39

Below the modal, two orange arrows point from the text "1) click there to get connection information" and "2) copy this line into your terminal" to the "ssh -A -p 22 root@192.54.201.39" line in the "Command-line connection" section.

Practical

- **Use RAINBio**

- ✓ <http://cloud.france-bioinformatique.fr/rainbio/>

- **Filter the images**

- ✓ Find which images provide FastA ? BLAST+ ?
 - ✓ IGV ?
 - ✓ Image J ?
 - ✓ Which version ?

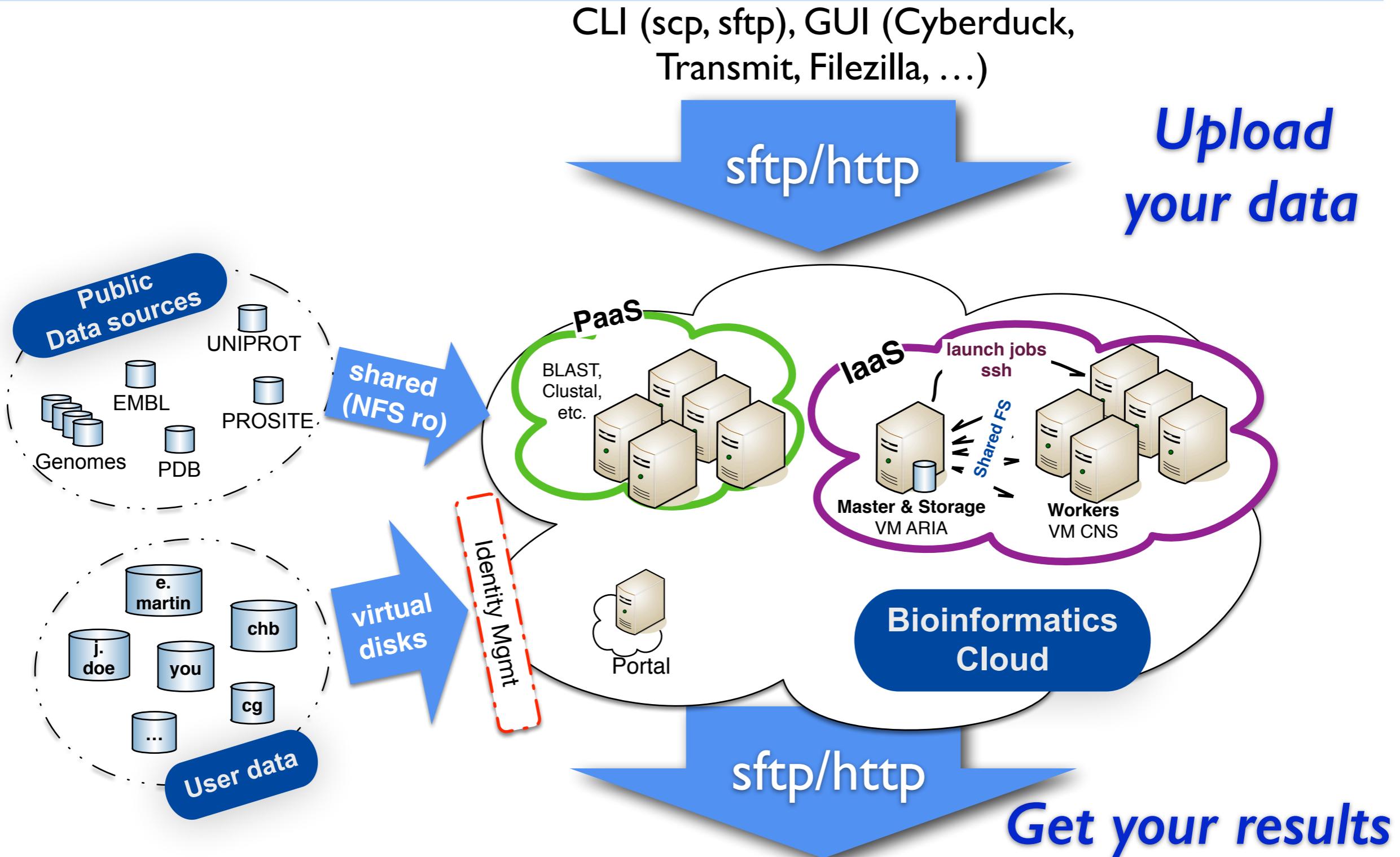
- **Run an instance of Biocompute**

- ✓ name: myBiocompute
 - ✓ size: c2.small

- **Connect to your VM**

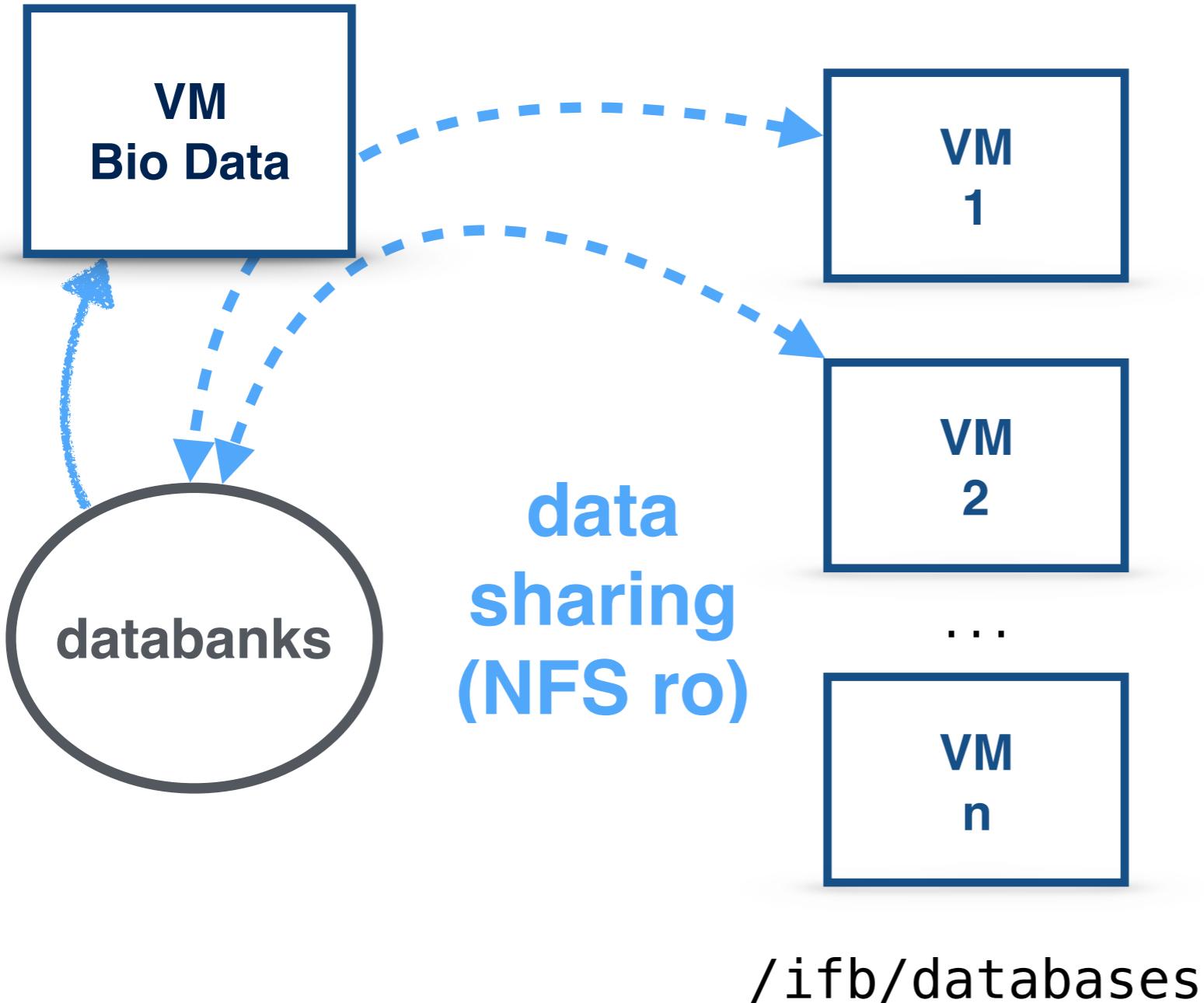
- ✓ and list available public reference data
 - ❖ ls /ifb/databases

Transfer data with the cloud



CLI (scp, sftp), GUI (Cyberduck,
Transmit, Filezilla, ...)

BioMaj repository



Use to provide collections of public reference data

<http://biomaj.genouest.org/>

Virtual disks

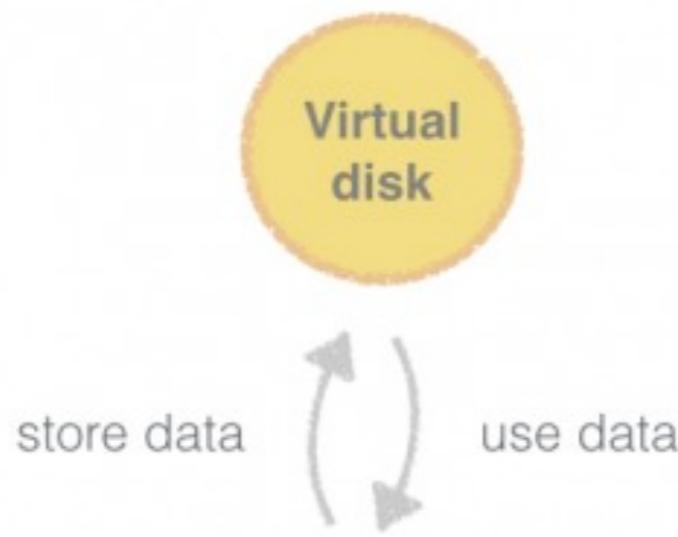
- To store your data

- ✓ Variable number
- ✓ Max size (quota)
- ✓ From one VM to a next one
- ✓ ! but no backup !

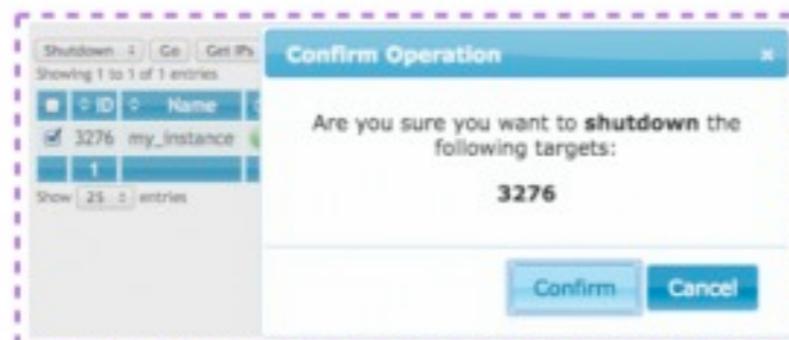
- Mount to
 - one VM
 - at boot time
- Share with VMs
 - Cluster mode
 - App 'NFS'

Instance	Storage			
Storage	CPU			
CPU				
	<input type="checkbox"/> dd4	20	d9b2dae1-3	c22d5daf7b
	<input type="checkbox"/> dd3	20	5b79798c-50	98c14b1282
	<input type="checkbox"/> dd2	20	2f583b45-e	5e544fb8bf
	<input type="checkbox"/> SG	10	ef2e30e8-d	a1a566771
	<input type="checkbox"/> trembl-split	100	f353f288-ed	03965ab98
	<input type="checkbox"/> dnachip	20	53b7e085-c	7c8736bd3d
	<input type="checkbox"/> galaxy 100G	100	9efe5248-6a	55ccf26b15
	<input type="checkbox"/> lait4	5	a8aa826a-c	64d166aa3
	<input type="checkbox"/> milk3	5	0ce3018f-5a	278d32fa72
	<input type="checkbox"/> milk2	5	f2e5c6c4-7e	24bf48107d
	<input type="checkbox"/> milk1	5	ef68d2df-57	36519bd65
	<input type="checkbox"/> stockage galaxy	10	c16e2998-b	3a2991bfde
	<input type="checkbox"/> coll MG	100	23476ebe-d	54b70309b5

Use your vDisk



(3) Shutdown the VM to detach the Vdisk



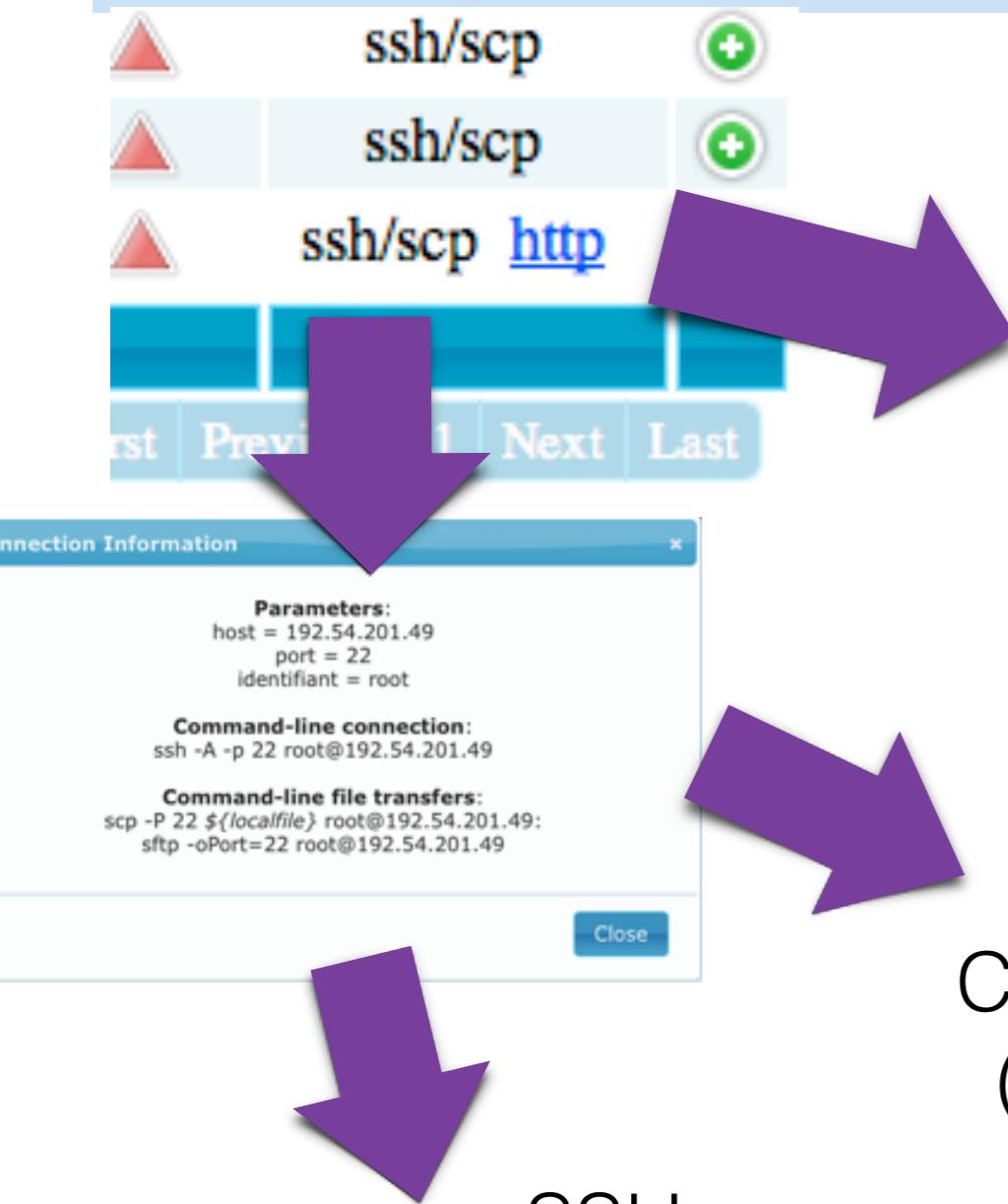
1) Create a virtual disk

A screenshot of a "Create a virtual Disk" dialog box. It shows "Define Your Parameters" with Size 5 and Name my_disk. Buttons at the bottom are "Create" and "Cancel".

2) Create an instance with this virtual disk

A screenshot of a "Launch a virtual machine" dialog box. It shows "Choose The Appliance" with BIO ComputeNode selected. Under "Configure Your Virtual Machine", Name is my_instance, Unique is unchecked, Type is c2.small (1 CPU, 2GB RAM), and Number is 1. Under "Plug Your Additional Storage", Persistent disk is checked for my_disk. Buttons at the bottom are "Run" and "Cancel".

Transfer Data



Cyberduck
(Filezilla)

SSH

```
dataset — title — bash — 389
idb1:dataset cblanchet$ scp -P 20132 * root@idb-cloud.ibcp.fr:upload_dir/
SampleA.1.fastq          100% 244KB 244.2KB/s  00:00
SampleA.2.fastq          100% 244KB 244.2KB/s  00:00
SampleB.1.fastq          100% 263KB 262.7KB/s  00:00
SampleB.2.fastq          100% 263KB 262.7KB/s  00:00
SampleC.1.fastq          100% 269KB 269.0KB/s  00:00
SampleC.2.fastq          100% 269KB 269.0KB/s  00:00
ref1.fasta                100% 10KB   9.9KB/s   00:00
run_analyses.sh          100% 1321   1.3KB/s   00:00
vcfutils.pl               100% 15KB   15.3KB/s  00:00
idb1:dataset cblanchet$
```

The screenshot shows the Galaxy web interface. At the top, it says 'Galaxy' and has a navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, there's a 'Tools' sidebar with various options like 'Upload File', 'UCSC Main table browser', etc. The main area is titled 'Upload File (version 1.1.3)'. It has sections for 'File Format' (set to 'Auto-detect'), 'File' (with a note about file size limits), and 'URL/Text'. Below these is a list titled 'Files uploaded via FTP:' showing the following files:

File	Size	Date
SampleC.1.fastq	269.0 KB	05/22/2013 12:10:19 PM
SampleC.2.fastq	269.0 KB	05/22/2013 12:10:19 PM
SampleB.2.fastq	262.7 KB	05/22/2013 12:10:19 PM
vcfutils.pl	15.3 KB	05/22/2013 12:10:19 PM
SampleA.2.fastq	244.2 KB	05/22/2013 12:10:19 PM
SampleB.1.fastq	262.7 KB	05/22/2013 12:10:19 PM
run_analyses.sh	1.3 KB	05/22/2013 12:10:19 PM

Galaxy

The screenshot shows the Cyberduck interface connected to 'idb-cloud.ibcp.fr - SFTP'. The top menu includes 'Obtenir une clé d'activation !', 'Nouvelle connexion', 'Connexion rapide', 'Action', 'Actualiser', 'Edition', and 'Se déconnecter'. The main area shows a file list in the '/root' directory:

Nom du fichier	Taille	Date de modification
anaconda-ks.cfg	2.5 KB	02/03/12 09:09
cleaner.sh	338 B	27/09/12 13:33
install	--	27/09/12 13:51
install.log	10.5 KB	02/03/12 09:09
install.log.syslog	3.8 KB	02/03/12 09:08
mydisk	10 B	16/11/12 16:03

At the bottom, it says '6 Fichiers'.

Practical - Data Transfer

- Transfer data to your VMs
- SSH with your Biocompute VM
 - ✓ command line : 'scp/sftp'
 - ✓ GUI : Cyberduck, Transmit, Filezilla, Ubuntu
 - ✓ Mounted it as a local disk on your desktop
(TRANSMIT et UBUNTU)
- Run some treatments
 - ✓ FastA, ClustalW2
 - ✓ Example data
 - ❖ <https://cloud.france-bioinformatique.fr/xport/oneSeq.fasta>
 - ❖ <https://cloud.france-bioinformatique.fr/xport/fewSeqs.fasta>
 - ❖ UNIPROT dans /ifb/databases

Practical - Galaxy

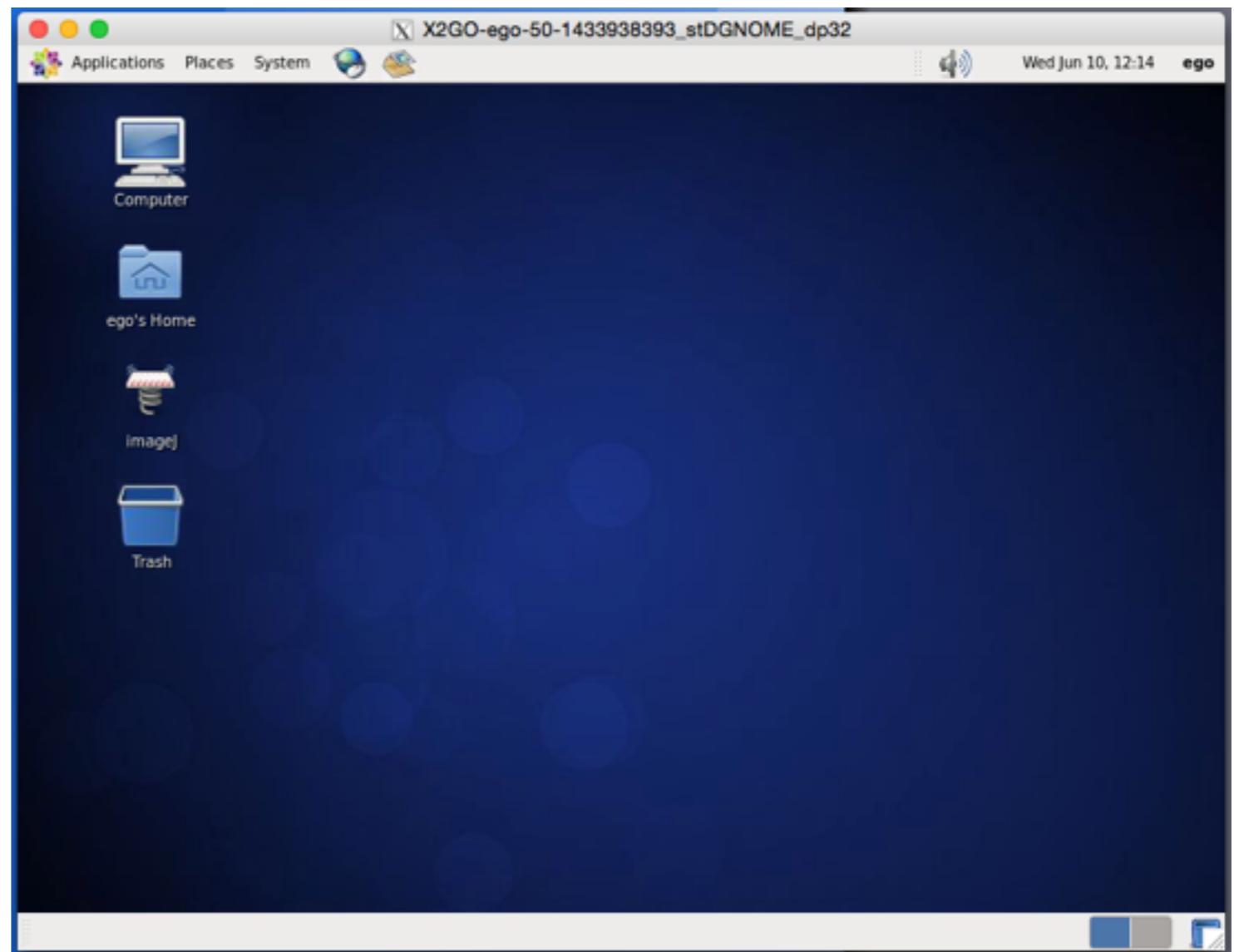
- a) Create a vDisk : myDDgalaxy, 10Go
- b) Launch **Galaxy** (attach your vDisk)
- c) Connect to Galaxy with the link ‘http’
- d) Sign in as user@galaxy.ifb.fr : ifbuser
- e) Get example data (sequences Illumina paired-end)
<http://www.prabi.fr/tutoGalaxy/>
- f) Run FastQC

Courtesy C. Oger (IFB-PRABI)

The screenshot shows the Galaxy web interface. The top navigation bar includes links for 'Les plus visités', 'IFB', 'IFB cladm', 'rainbio', 'Intranet IFB-Core', 'Cloud CNRS', 'IDB', 'IDG', 'EGI', 'GRISBI', 'Sirius', 'Hybridfox', and 'lastjob'. The main content area displays the 'FastQC Read Quality reports (Galaxy Version 0.63)' tool. The 'Tools' panel on the left lists 'fastqc' and 'NGS: QC and manipulation' (with 'FastQC Read Quality reports' selected). The 'Workflows' section shows 'All workflows'. The central form for 'FastQC' has fields for 'Short read data from your current history' (containing '1: http://www.prabi.fr/tutoGalaxy/SRR166812_1M_1.fq'), 'Contaminant list' (set to 'Nothing selected'), and 'Submodule and Limit specifying file' (also set to 'Nothing selected'). A 'Purpose' section at the bottom states: 'FastQC aims to provide a simple way to do some quality control checks on raw sequence data'. The right side shows the 'History' panel with three entries: '3: FastQC on data 1: RawData', '2: FastQC on data 1: Webpage', and '1: http://www.prabi.fr/tutoGalaxy/SRR166812_1M_1.fq'. The total memory usage is listed as 'Using 247.4 MB'.

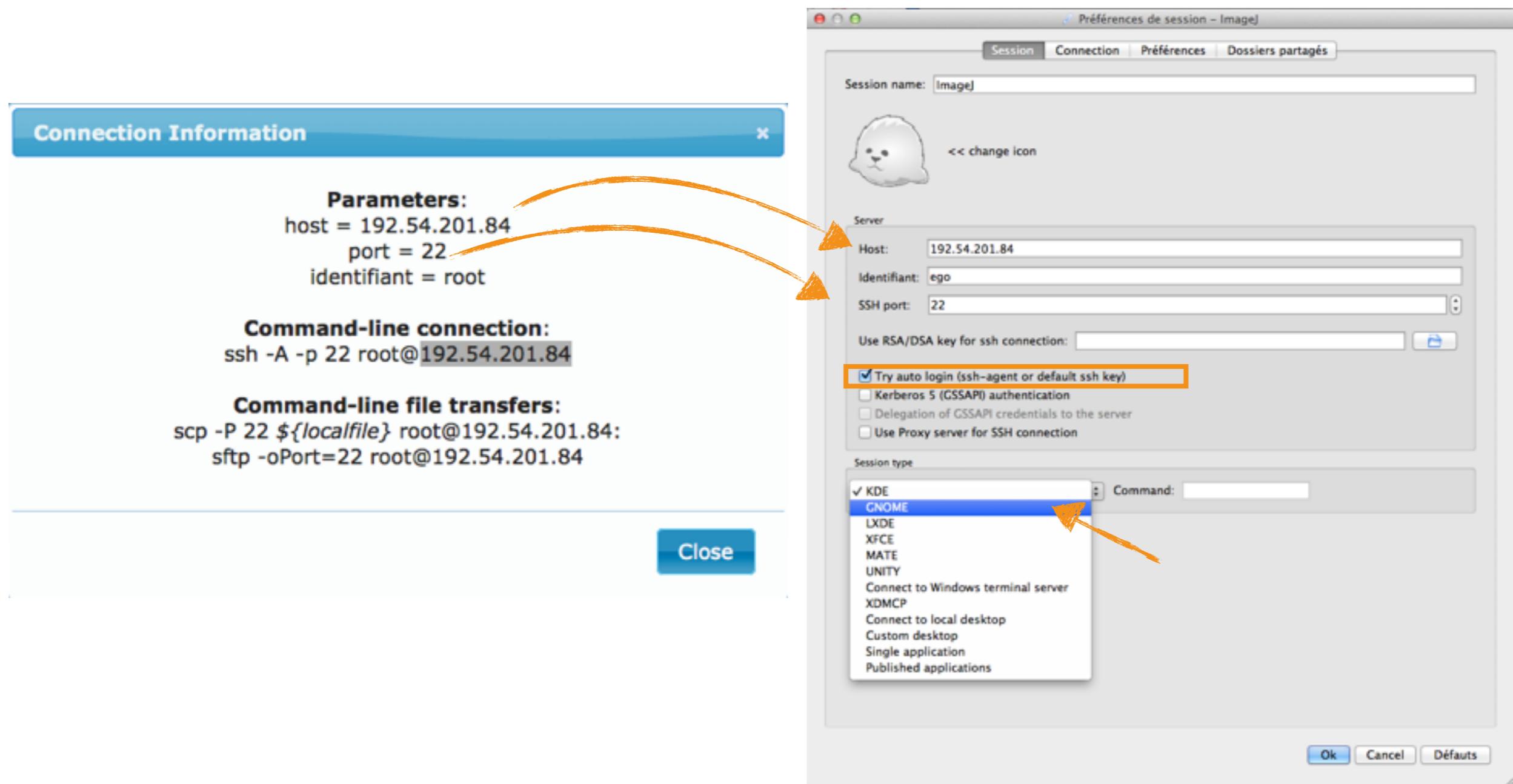
Remote desktop

Graphical interface on the remote VM



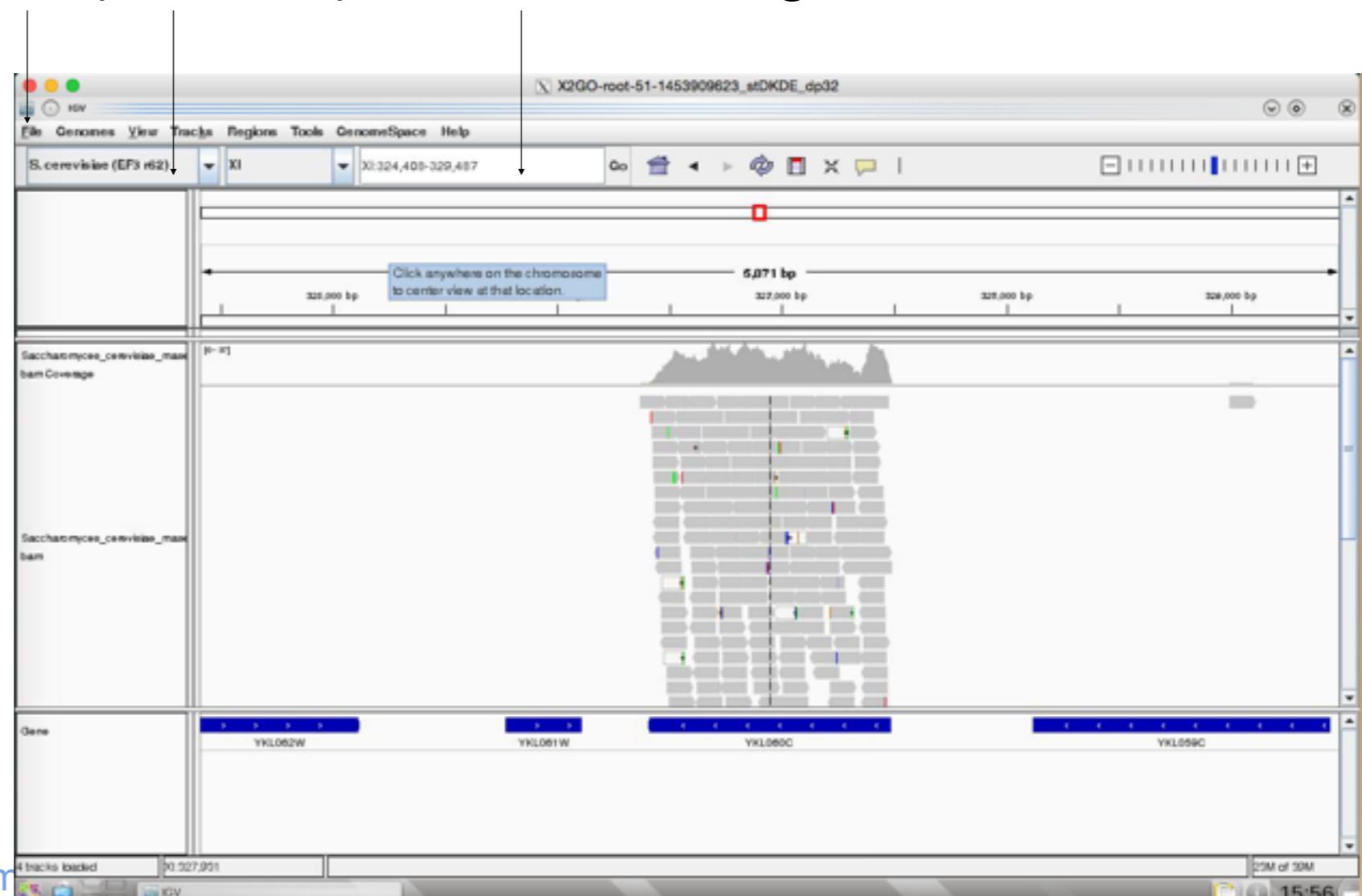
- Require an X2Go client and an X11 server on your desktop
- Available in several appliances : Biodatacloud IGV, Proteomics, ImageJ...

Howto use X2Go

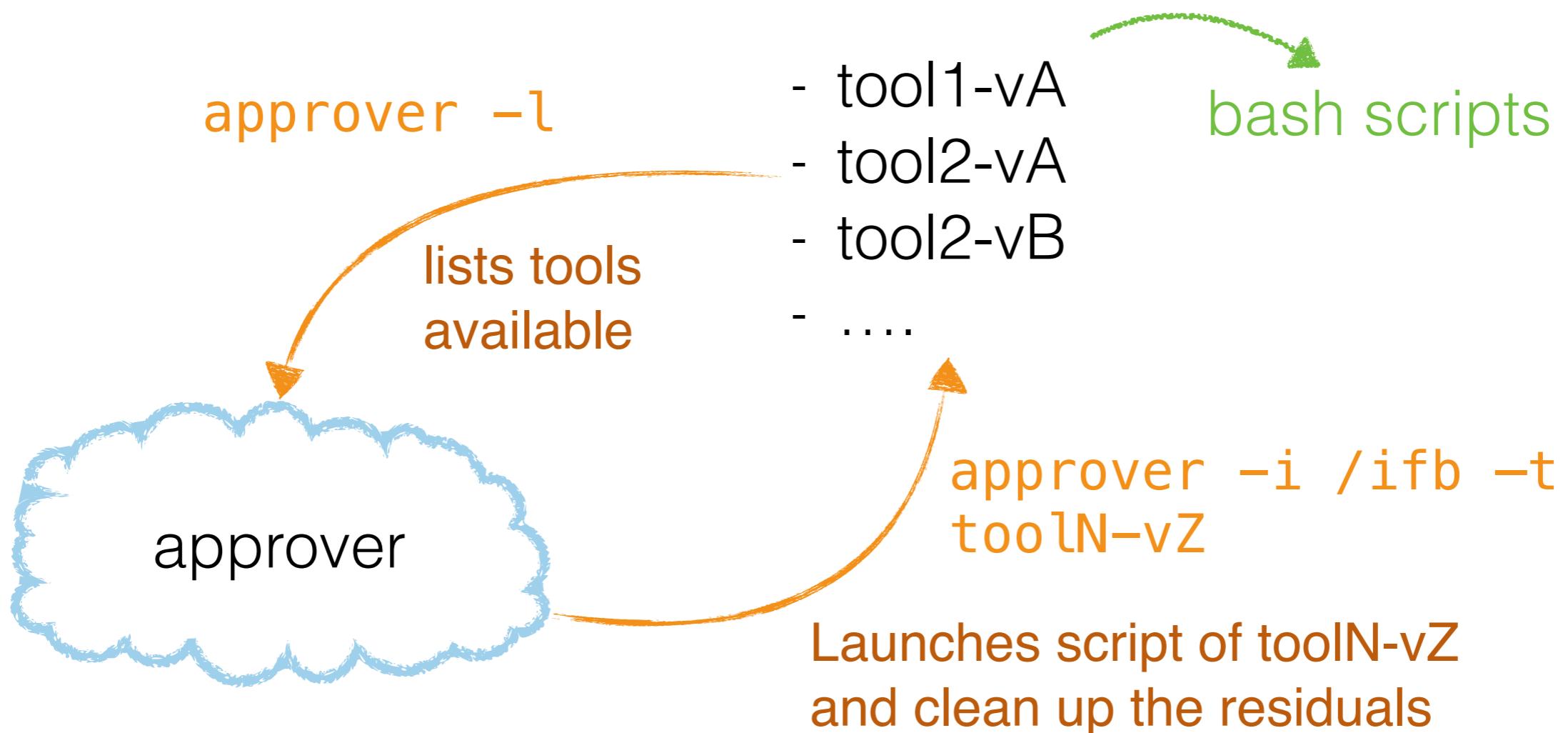


Practical - Remote Desktop

- a) Launch BioDataCloud IGV
- b) Connect with your X2Go client
- c) Upload example data: `Saccharomyces_cerevisiae_rnaseq.bam`,
`Saccharomyces_cerevisiae_rnaseq.bai`
- d) Launch IGV
- e) Choose *S. cerevisiae* (EF3 r62), BAM file and gene YKL060C



Customize your VM - Approver



Customize your VM - Galaxy

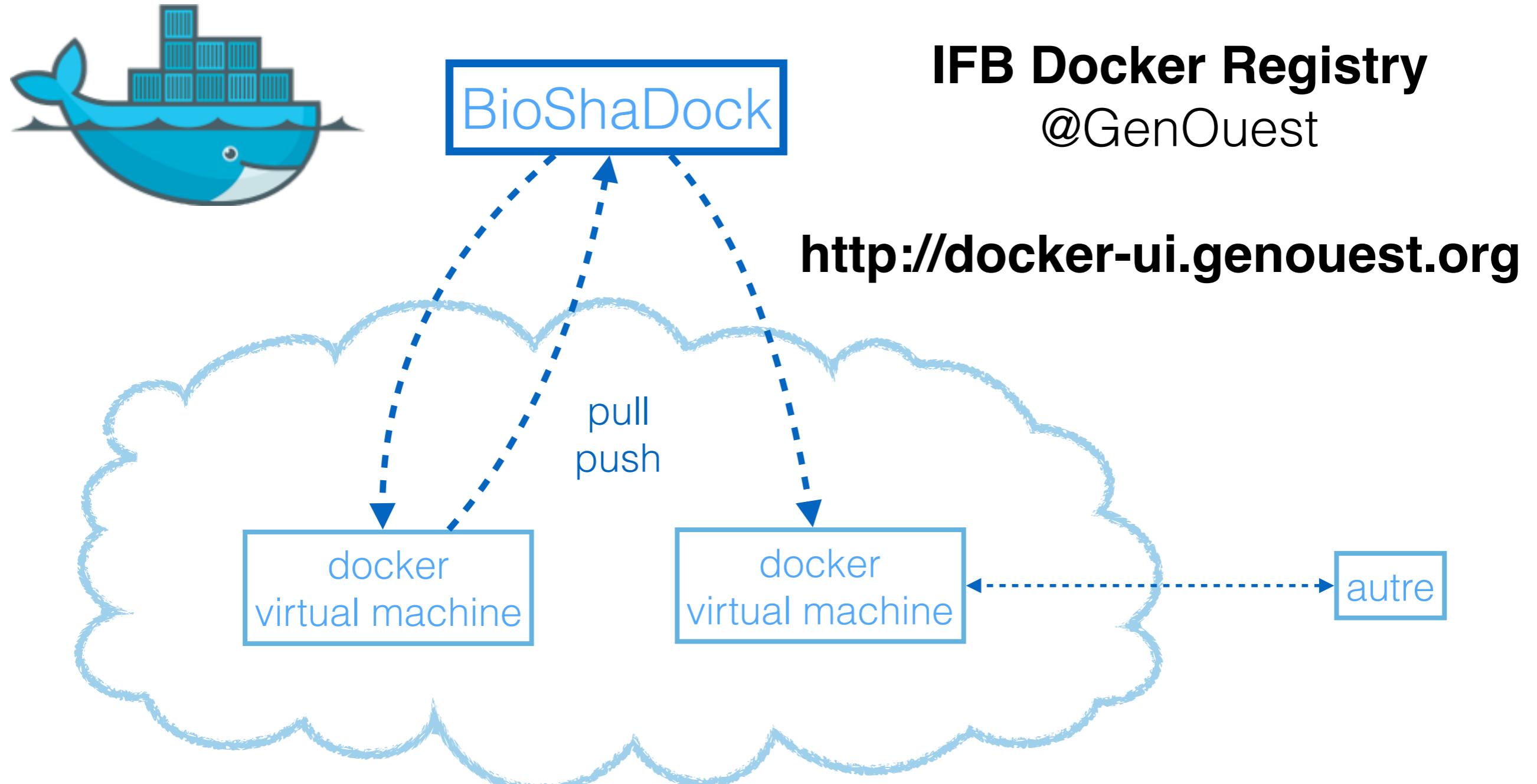
- a) Connect to Galaxy with the link 'http'
- b) Sign in as admin@galaxy.ifb.fr : ifbadmin
- c) Go to Admin panel
- d) Search Tool Shed

The screenshot shows the Galaxy Admin interface. The top navigation bar has a red circle around the 'Admin' button. On the left, a sidebar menu has a red circle around the 'Search Tool Shed' link under the 'Tools and Tool Shed' section. The main content area displays a list of accessible tool sheds: 'IFB main tool shed', 'IFB test tool shed', 'GenOuest main tool shed', 'Galaxy main tool shed' (which is highlighted with a red circle), and 'Galaxy test tool shed'. The 'Galaxy main tool shed' dropdown menu is open, showing its contents.

The screenshot shows the search results for matching tools. The top navigation bar has 'Admin' selected. The main content area is titled 'Repositories with matching tools'. It shows a table with columns: Repository name, Synopsis, Revision, and Owner. There are four entries, all named 'bam_to_sam', which is also listed in the 'Server' section of the sidebar. The table rows are as follows:

Repository name	Synopsis	Revision	Owner
bam_to_sam	Converts BAM format to SAM format.	c1419fa820c6	devteam
bam_to_sam	Converts BAM format to SAM format.	af7c50162f0b	devteam
bam_to_sam	Converts BAM format to SAM format.	dc20f447c0e2	devteam
bam_to_sam	Converts BAM format to SAM format.	c09a20532957	devteam

Customize your VM - Docker



BioShaDock

 Browse Containers Help Login Search OK

Show 10 entries Search:

ID	Description	User id	Visibility
askomics/fuseki-2.3.1 (Automatic build) >		ofilangi	true
biodckr/biodocker (Automatic build) >		fmoreews	true
biodckr/blast_2228_2 (Automatic build) >		fmoreews	true
biodckr/clustal_omega_121_1 (Automatic build) >		fmoreews	true
biodckr/comet_2015011 (Automatic build) >		fmoreews	true
biodckr/comet_2015012 (Automatic build) >		fmoreews	true
biodckr/comet_2015020 (Automatic build) >		fmoreews	true
biodckr/crux_21 (Automatic build) >		fmoreews	true
biodckr/denovogui_152 (Automatic build) >		fmoreews	true
biodckr/dia_umpire_14256 (Automatic build) >		fmoreews	true

Showing 1 to 10 of 154 entries Previous 1 2 3 4 5 ... 16 Next

BioShaDock

[Browse](#)[Containers](#) ▾[Help](#)[Login](#)

Search

OK

PUBLIC | Automatic build

biodckr/blast_2228_2[More info ▾](#)[View Docker file](#)

④ 8

```
docker pull docker-registry.genouest.org/biodckr/blast_2228_2[:tag]
```

Description

NCBI BLAST+ Version: 2.2.28-2 basic local alignment search tool
<http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGETYPE=BlastDocs&DOCTYPE=Download> Provides: blast 2.2.28-2

bio.tools

Identifier

Docker tags

latest

Docker Entrypoint

Entrypoint: none

General information

Author: fmoreews
Architecture: amd64
Tag: latest

Metadata tags

genomics proteomics transcriptomics

Annotated with
EDAM terms

Docker - Useful commands

```
# Docker version  
docker version  
# Available images  
docker images  
# Running containers  
docker ps  
# Stopped containers  
docker ps -a  
  
# Suppress one or several containers (-f to force)  
docker rm -f <container_id | container_name>  
OR    docker rm -f $(docker ps -aq)  
# Suppress one or all images  
docker rmi -f <image_name>:<tag>  
OR    docker rmi -f $(docker images -q)
```

<https://docs.docker.com/>

Practical - Docker

- a) Run a Docker instance
- b) Connect with SSH

```
# Get an image
```

```
docker pull docker-registry.genouest.org/ifb/cutadapt:1.9.1
```

```
# Run an image
```

```
docker run <options> <image_name>:<tag> <cmd> <args>
```

options :

- it -> *interactive mode*

- d -> *run in background*

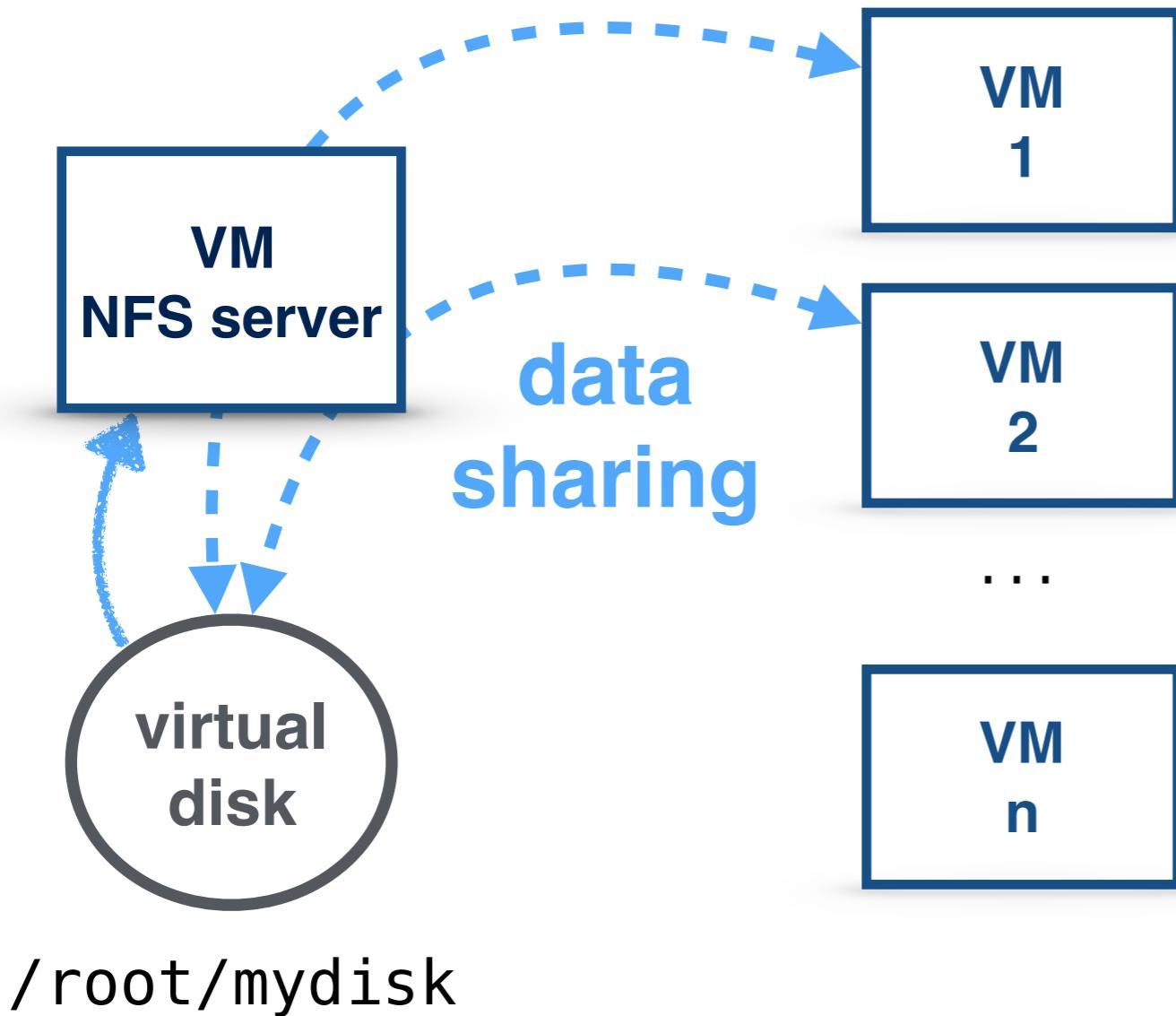
- v <dir_local>:<dir_container> *to define mount points*

- w *to define the working dir*

Example:

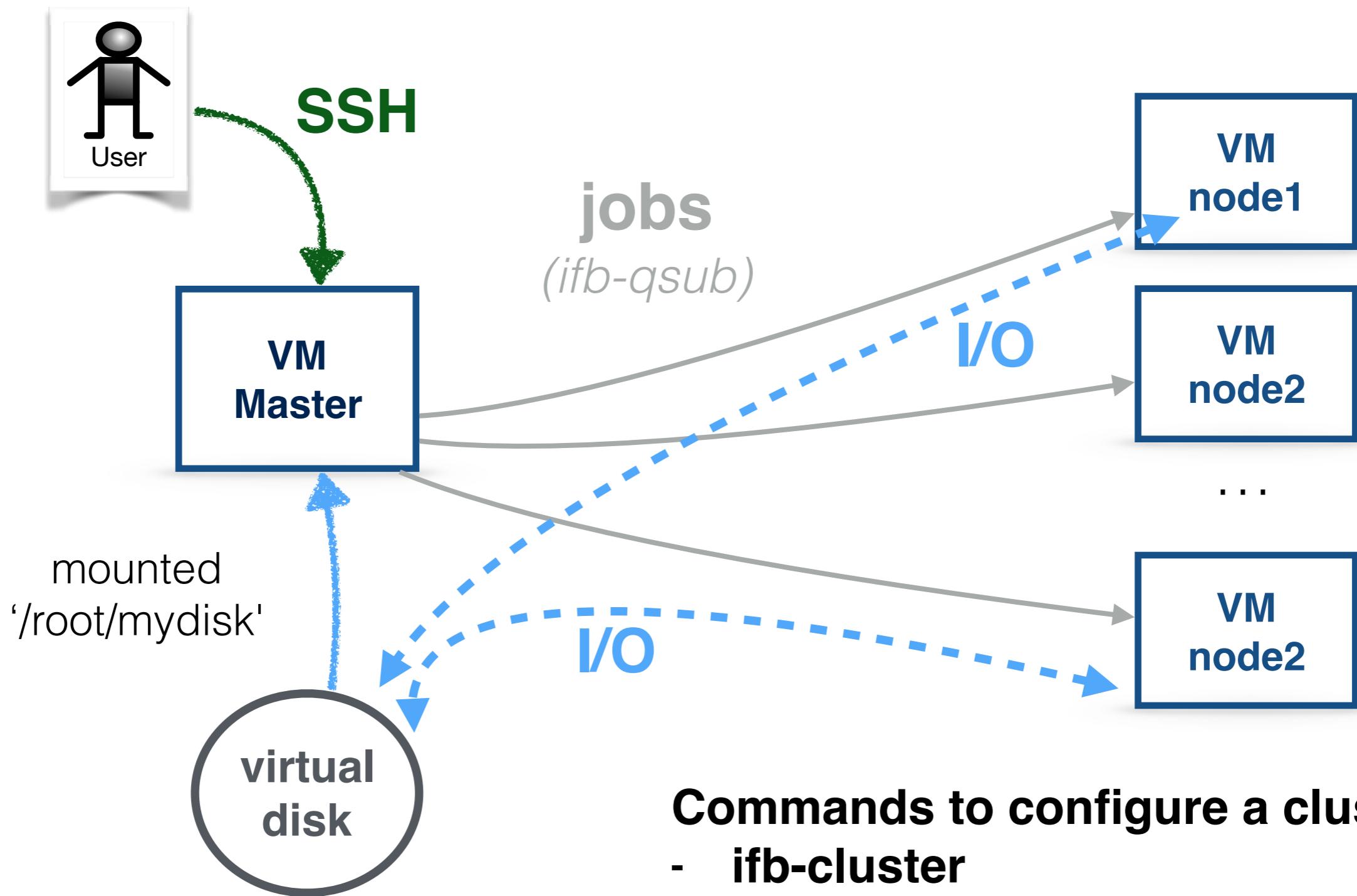
```
docker run docker-registry.genouest.org/ifb/cutadapt:1.9.1 cutadapt
```

NFS server



With NFS :
a volume can be
shared among
several VM

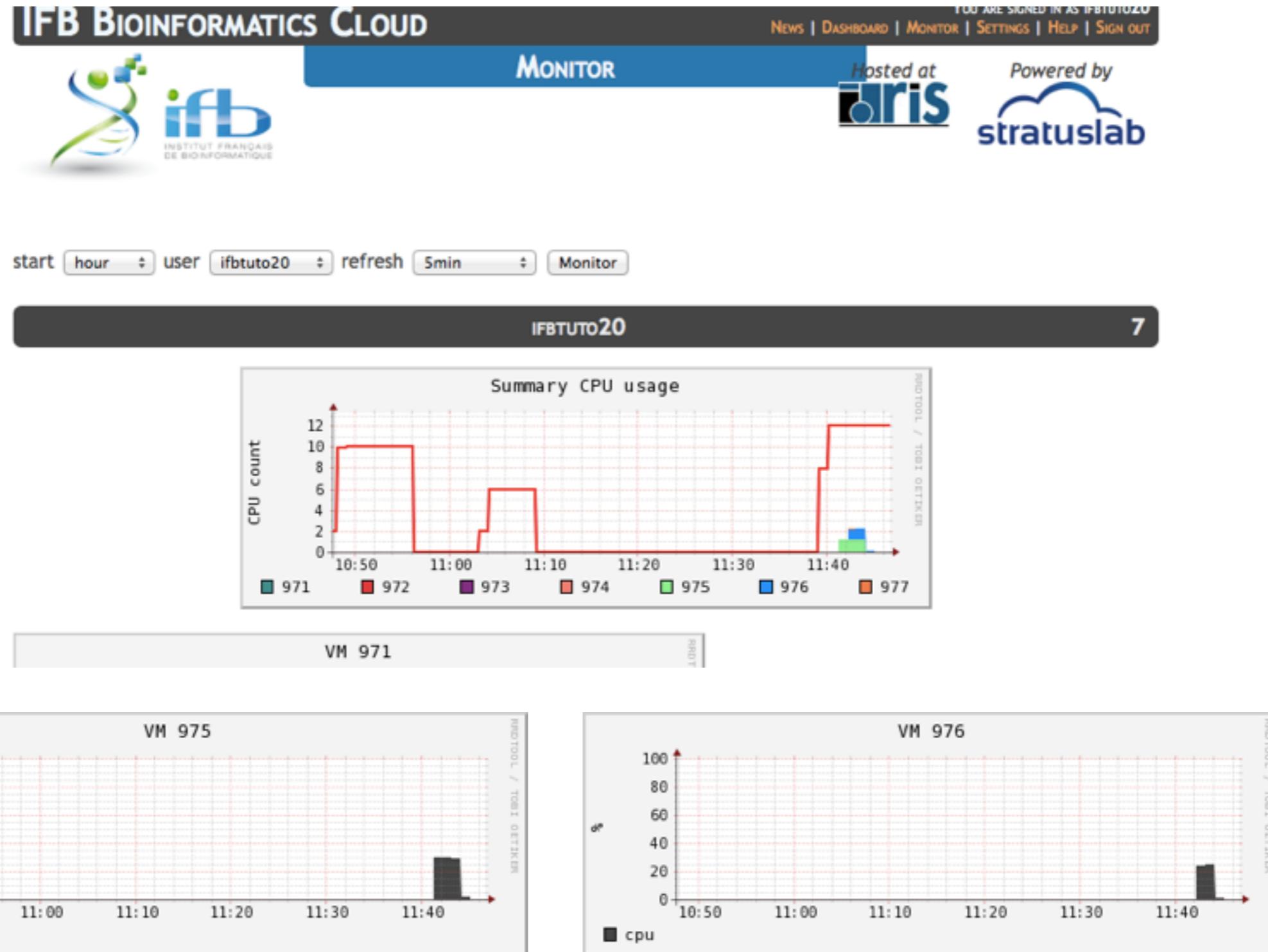
Principle of a virtual cluster



Commands to configure a cluster

- **ifb-cluster**
- **ifb-cluster-sge**
- **ifb-cluster-pbs**

Monitor your usage



Principle of Appliance development

Before developing your appliance

You need the developer rights.

Prepare your project and contact us.

We need :

- a title
- a description
- an author/maintainer name
- a list of the tools that will be installed
- a list of the databanks that will be used
- other parameters...

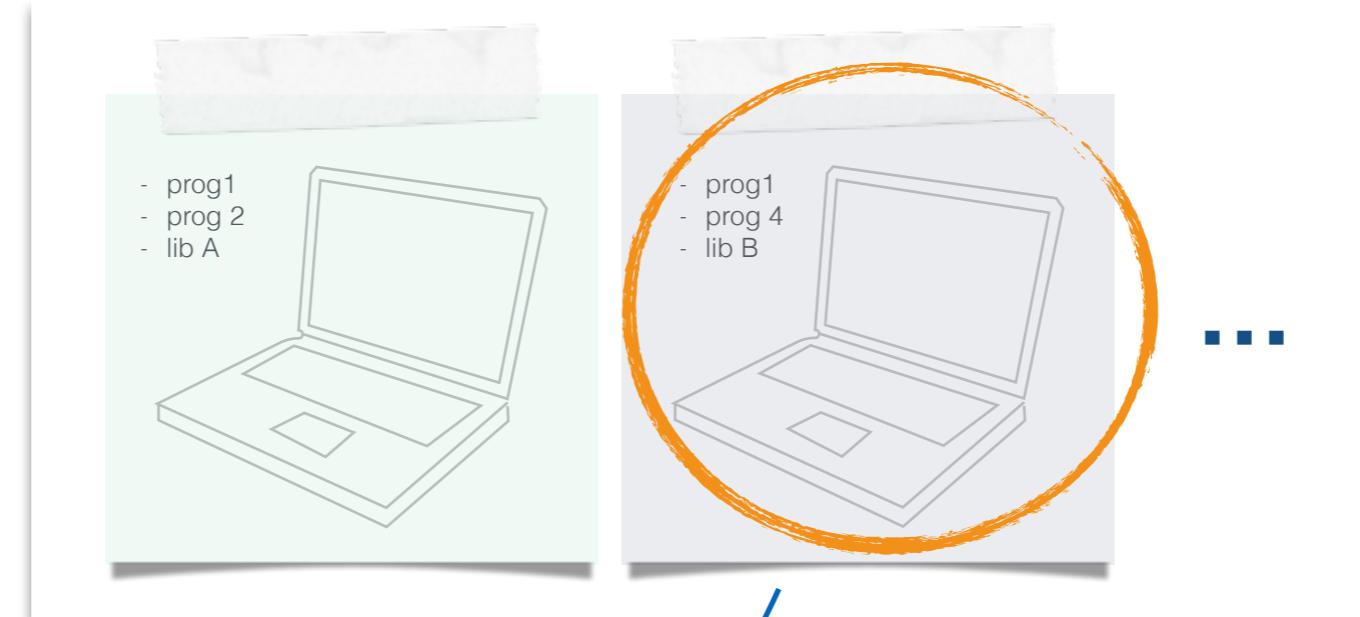
Main steps of appliance development

- Choose your base image
- Make an instance in create mode
- Customize your appliance
- Shut it down with the web interface
- Contact IFB support@france-bioinformatique.fr

MARKET PLACE

you,
appliance
developer

choose
an image



virtual machine

customize
your VM



- prog1
- prog4
- libB
- LIB A**
- prog18



shutdown the VM
with web interface

create

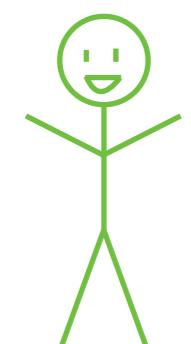


makes an instance



image

cloud
administrator



validates
and
references

Mode ‘create’

When you choose the create mode :

Create Instance

Choose The Appliance

Appliance ? CentOS 6.6 (base)
Filter by ? --- THEMATIC FIELDS ---
--- TOOLS ---

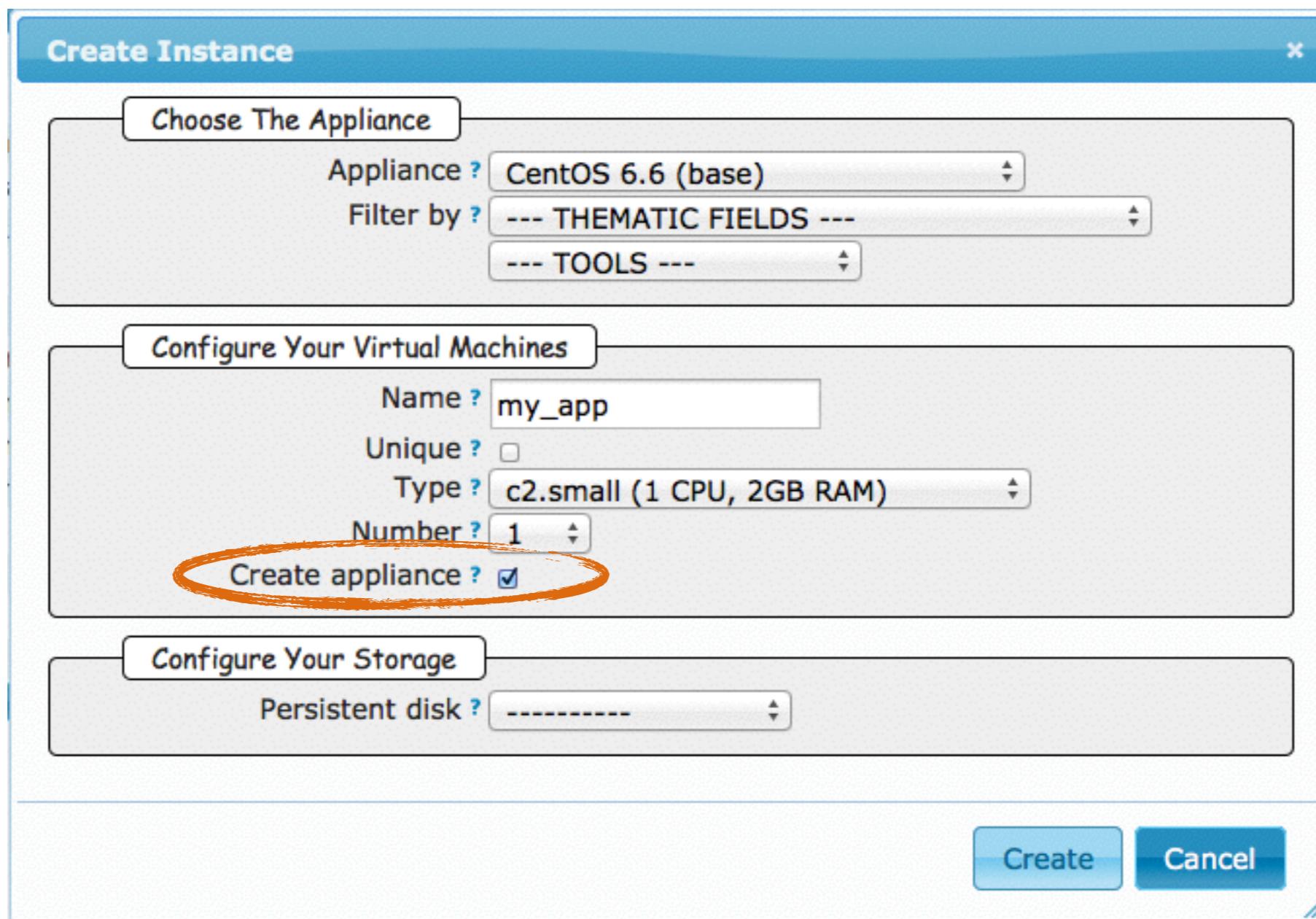
Configure Your Virtual Machines

Name ? my_app
Unique ?
Type ? c2.small (1 CPU, 2GB RAM)
Number ? 1
Create appliance ?

Configure Your Storage

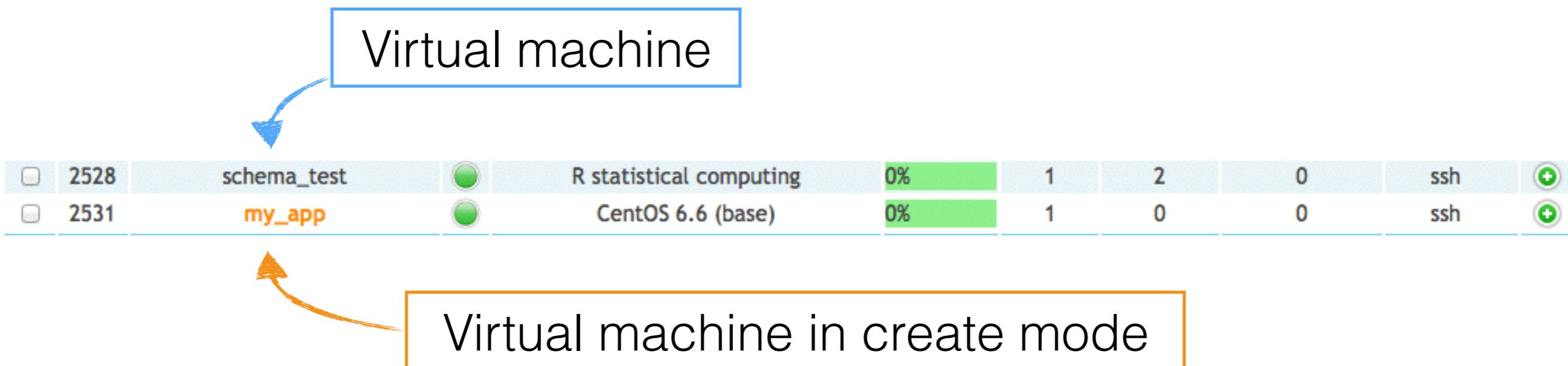
Persistent disk ? -----

Create **Cancel**



What it looks like

On the dashboard, once you made your instance :



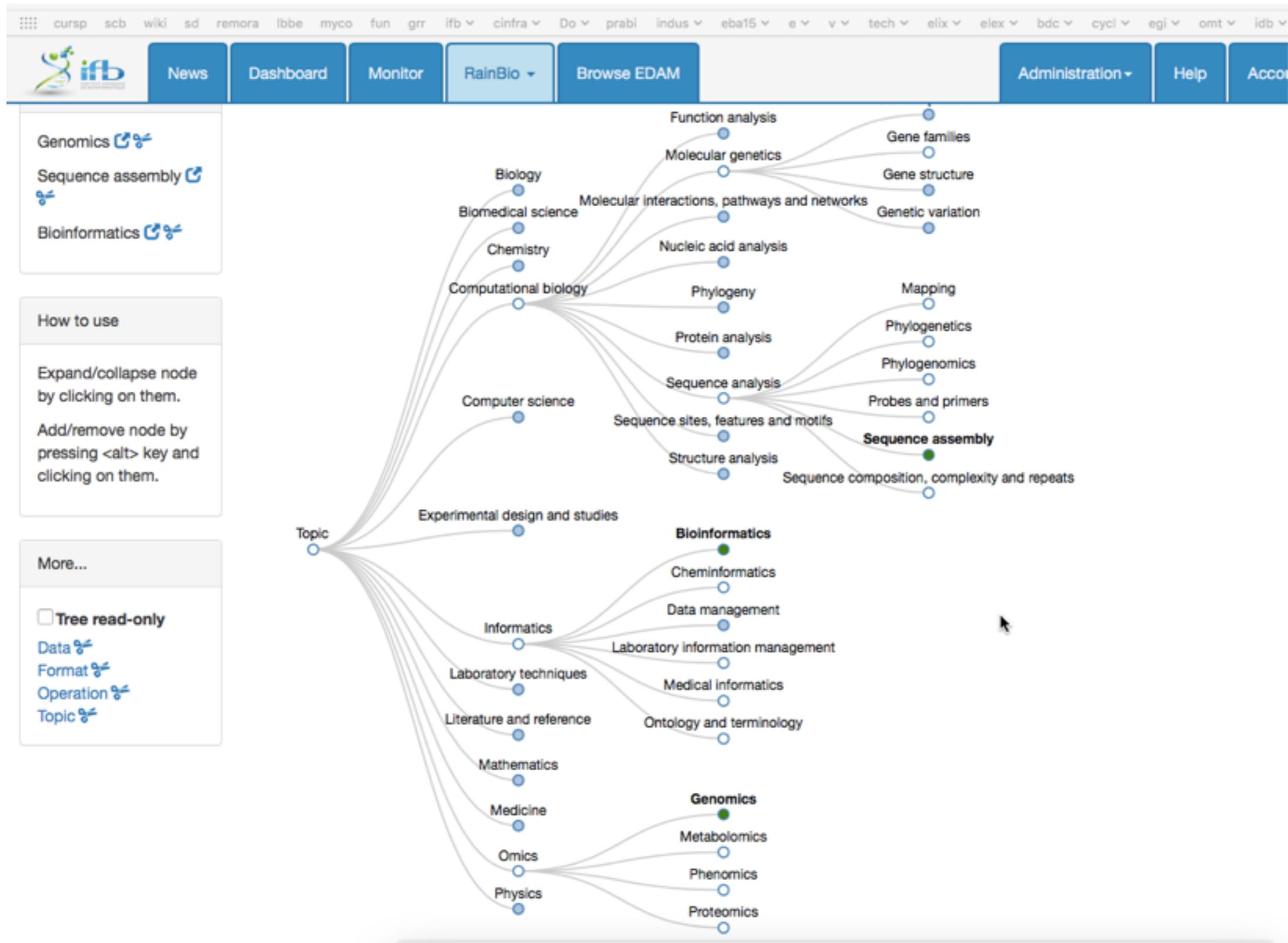
Best practices to create a cloud app

- Fill the appliance description form
- Detailed the list of bioinformatics tools
 - with dependencies
 - and EDAM terms
- Choose the best base appliance
 - according to the provided tools/features
- Prepare the installation in a normal VM
- Use the web dashboard, not the VM shell
 - ‘Shutdown’
 - ‘Kill’ for an instance you don’t want to keep
- Do not just run a VM in creation mode and shutdown just to test it, it works !

And ‘sysadm’ best practices

- **Update your system**
 - yum update/apt-get update + dist-upgrade
 - & ‘reboot’ in case of a kernel update
- **Standard user**
 - Keep public_key SSH mode
- **Use a web proxy for portals**
- **Open only required ports**
 - Use standard ones (22, 80/443...)
- **Install software in standard dirs**
 - /usr/local... or /ifb
- **Shorten the ‘creation’ running time to avoid too much logs**
- **Do not change filesystem rights to 777...**

Annotate with EDAM



Different types of software installation

- **Interactive installation :**
 - binaries archive,
 - sources+compilation
- **automated installation :**
 - package managers (yum, apt, rpm...),
 - scripts (approver, Galaxy toolshed)
 - recipes (puppet/ansible/chef/salt)
- **docker container :**
 - dockerfiles...



Useful links

IFB Cloud : <https://cloud.france-bioinformatique.fr/accounts/login>

Example data : http://www.france-bioinformatique.fr/sites/default/files/pages/dataset_ibl.zip

Docs : <http://docker-ui.genouest.org/app/>

RAINBIO : <http://cloud.france-bioinformatique.fr/rainbio>

BioShaDock : <http://docker-ui.genouest.org/app/#/all/containers>

Galaxy admin : <https://wiki.galaxyproject.org/Admin/Tools/AddToolFromToolShedTutorial>

Acknowledgements: S. Perrin, J. Lorenzo