Cloud pour la Bioinformatique

Christophe Blanchet

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

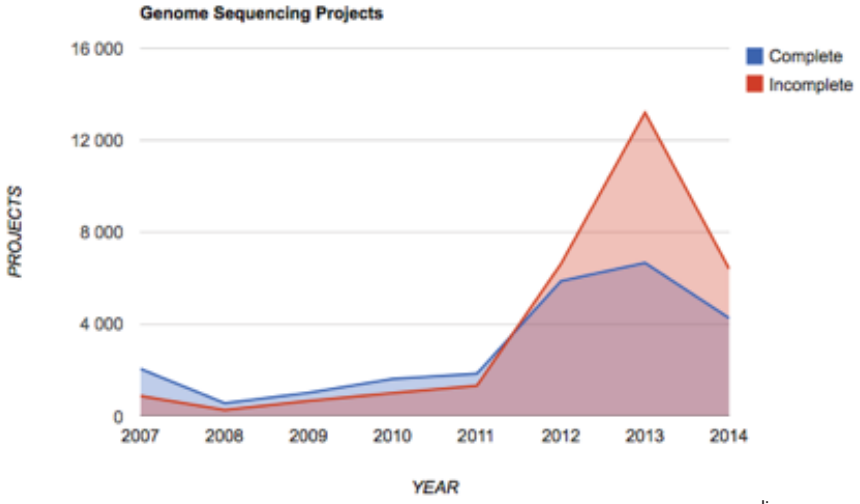
Sequencing data

Next-Generation Sequencing Statistics

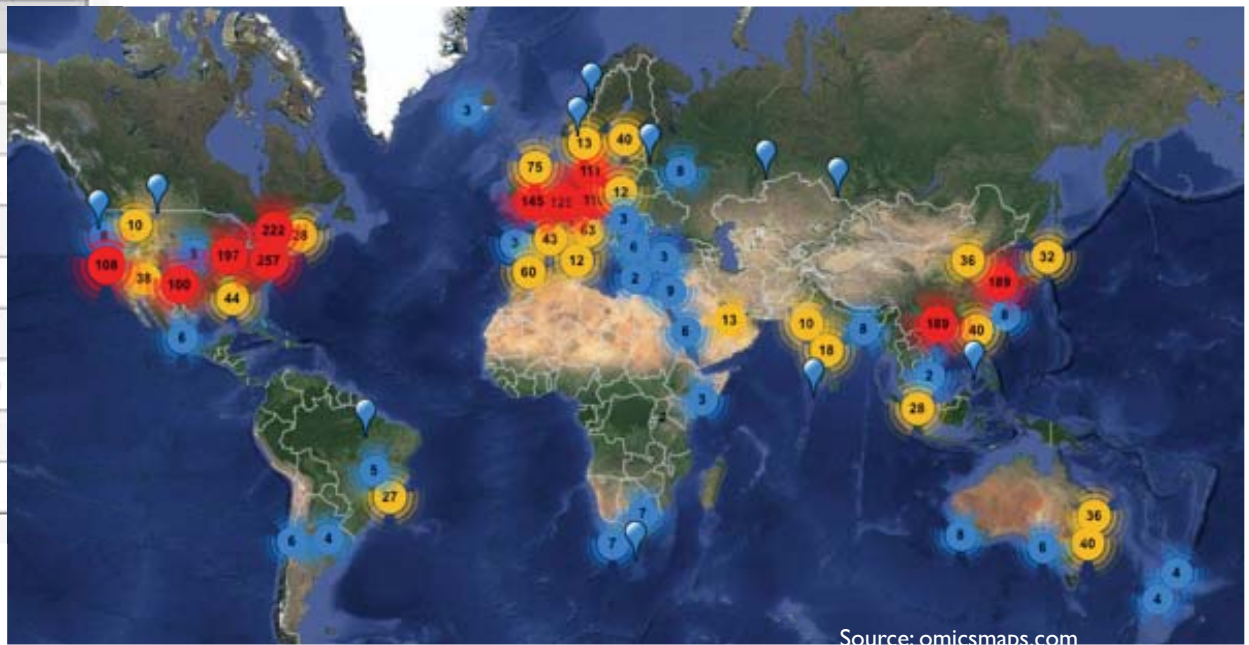
Vendor:	Roche			Illumina			ABI		
Technology:	454			Solexa GA			SOLiD		
Platform:	GS20	FLX	Ti	I	II	IIx	1	2	3
Reads: (M)	0.5	0.5	1.25	28	100	150	40	115	320
Fragment									
Read length:	100	200	400	35	50	100	25	35	50
Run time: (d)	0.25	0.3	0.4	3	3	5	6	5	8
Yield: (Gb)	0.05	0.1	0.5	1	5	15	1	4	16
Rate: (Gb/d)	0.2	0.33	1.25	0.33	1.67	3	0.34	1.6	2
Images: (TB)	0.01	0.01	0.03	0.5	1.1	2.8	1.8	2.5	1.9
PA Disk: (GB)	3	3	15	175	300	300	300	750	1200
PA CPU: (hr)	10	140	220	100	70	NA	NA	NA	NA
SRA: (GB)	0.5	1	4	30	50	2.5	100	140	600

source: www.politigenomics.com/next-generation-

Read length:		200	400	2×35	2×50	2×100	2×25	2×35
Insert: (kb)		3-5	3-5	0.2	0.2	0.2	3	3
Run time: (d)		0.3	0.4	6	10	10	12	10
Yield: (Gb)		0.1	0.5	2	9	30	2	8
Rate: (Gb/d)		0.33	1.25	0.33	1.67	3	0.34	1.6
Images: (TB)							6	5
							0	1500
							A	NA
							0	280

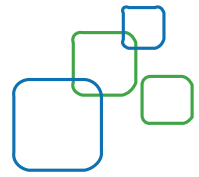


source: www.genomesonline.org



Complete genome sequencing become a lab commodity with NGS (cheap and efficient)

And other experimental data…



ProFI PROTEOMICS

EMBRC FRANCE

FRISBI Integrated Structural Biology

metagenopolis mgps.eu

cohorte CONSTANCES Améliorer la santé de demain

METABOHUB

tara EXPEDITIONS

RaDiCo Rare Disease Cohorts

BBMRI Biobanking and Biomolecular Resources Research Infrastructure

FRANCE GÉNOMIQUE

FR - EU

instruct Integrating Biology

EMBRC EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE

erinha

ECRIN

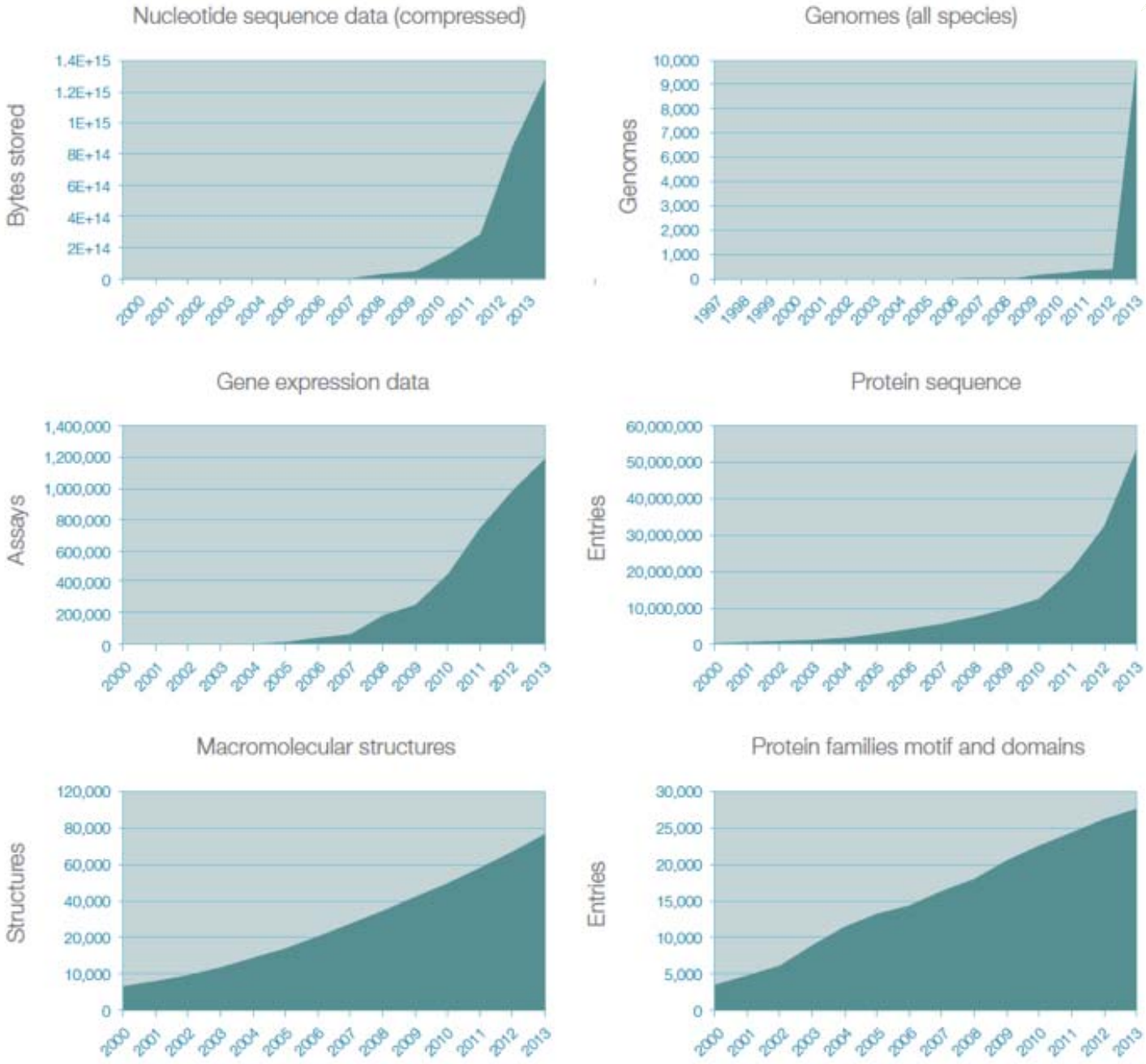
EURO-BIOIMAGING

EU-OPENSCREEN Chemical keys for life's locks

EATRIS

INFRAFRONTIER

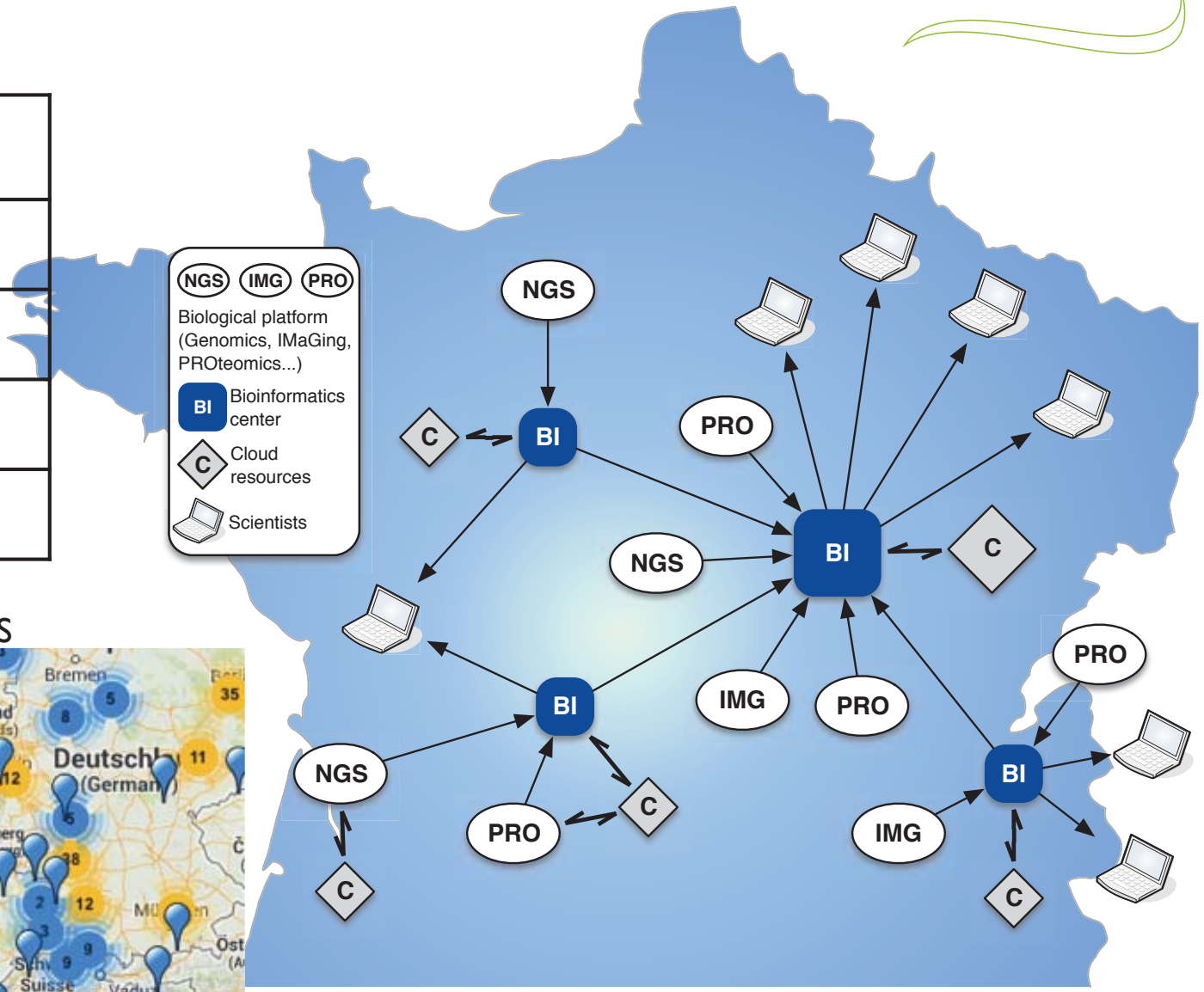
EMBL-EBI data resources growth



source: EMBL-EBI Annual Report 2013

Plateformes Expérimentales en Biologie

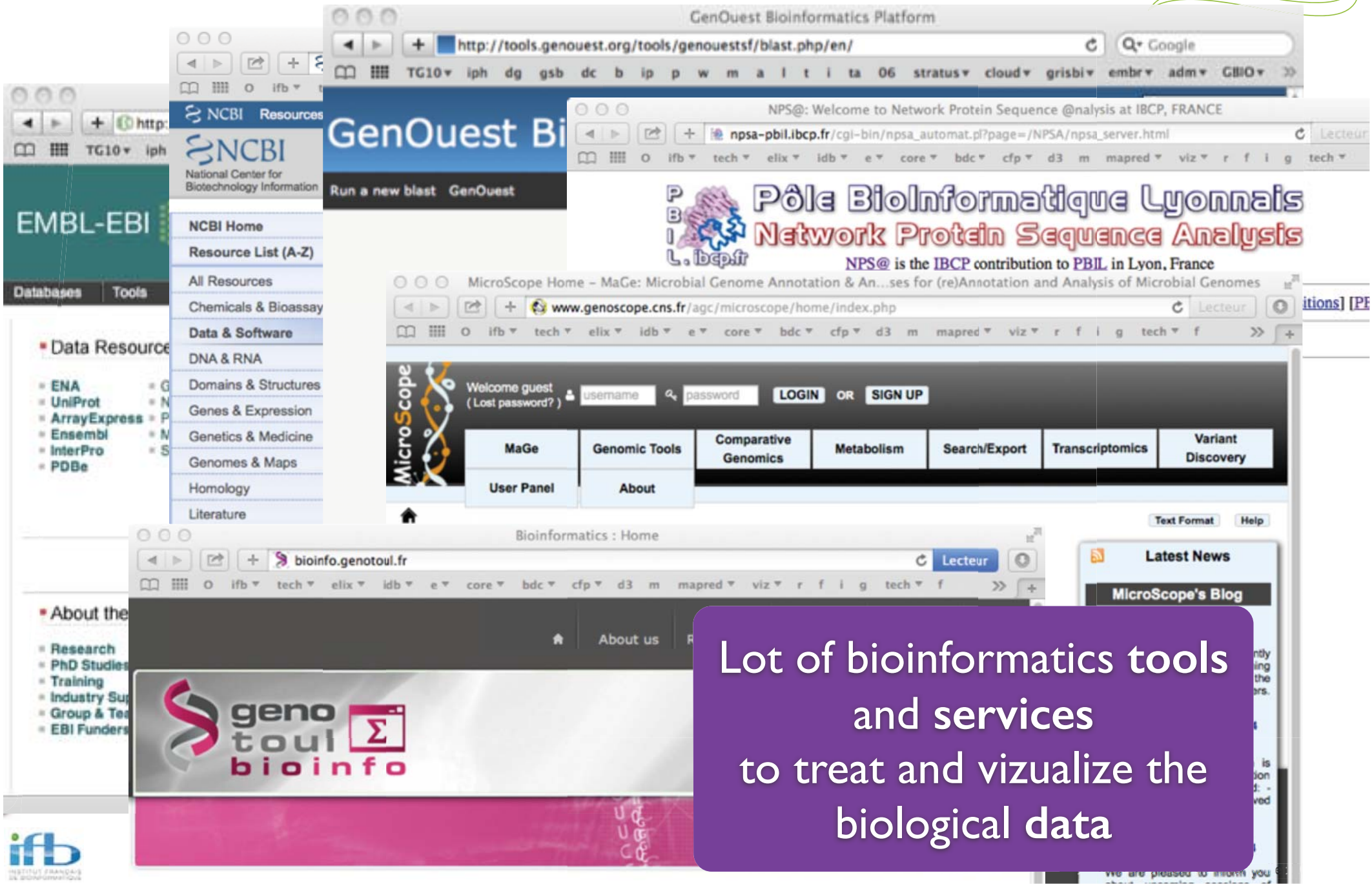
Plateformes nationales (GIS IBISA)	Nb
Imagerie cellulaire	19
Génomique, Transcriptomique	16
Protéomique	13
Biologie structurale, biophysique	11



Localisation des plateformes NGS

Des sites intermédiaires permettent de répartir la charge en terme de stockage et de puissance de calcul tout en assurant une meilleure proximité avec les scientifiques

Infrastructures in Biology



Lot of bioinformatics **tools** and **services** to treat and vizualize the biological **data**

Bioinformatics Today

- Biological data are *big data*
  - 1552 online databases (NAR Database Issue 2014)
  - Institut Sanger, UK, 5 PB - Beijing Genome Institute, China, 7 sites, 20.6 PB

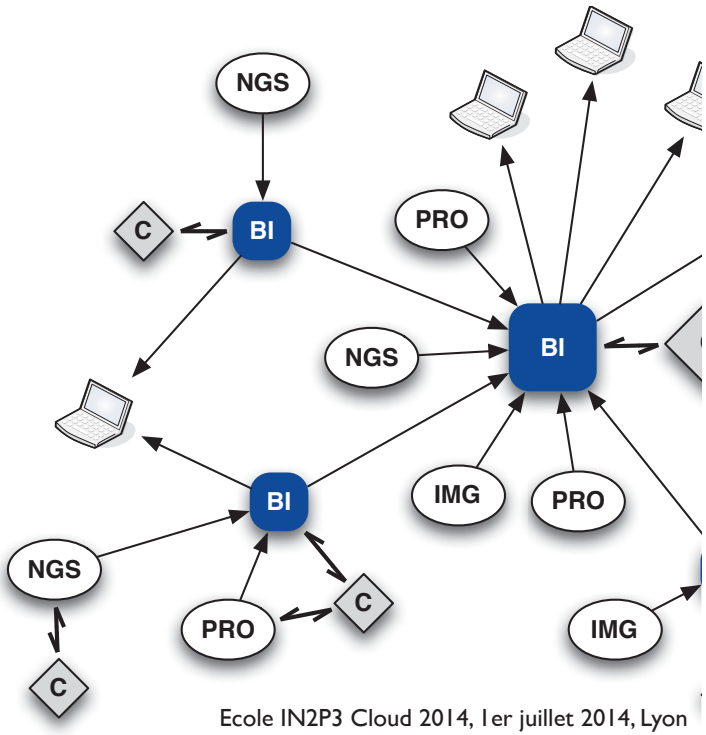
➡**Big data in many places**

- Analysing such data became difficult
  - Scale-up of the analyses : gene/protein to complete genome/proteome, ...
  - Lot of different daily-used tools that need to be combined in workflows
  - Usual interfaces: portals, Web services,…

➡**Datacenters with ease of access/use**

- Distributed resources
  - Experimental platforms: NGS, imaging, ...
  - Bioinformatics platforms

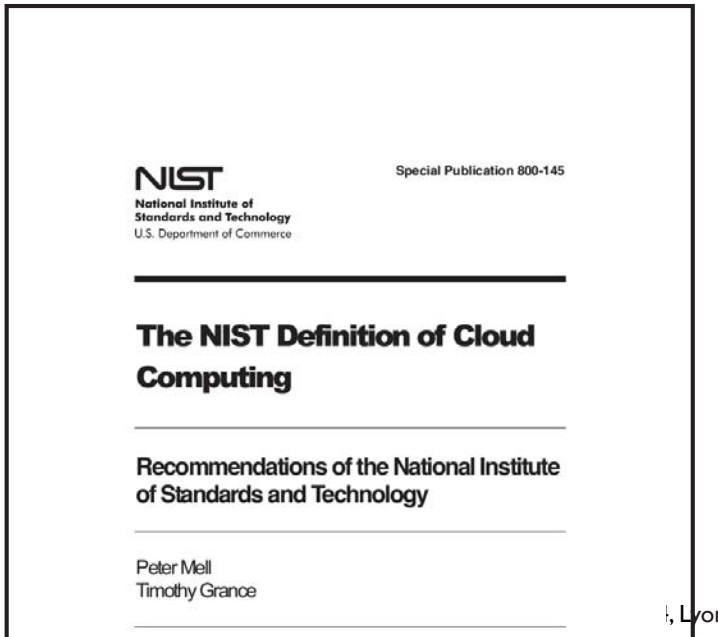
➡**Federation of datacenters**



Cloud ?

- **Essential characteristics**
  - On-demand self-service
    - No human intervention
  - Broad network access
    - Fast, reliable remote access
  - Rapid elasticity
    - Scale based on app. needs
  - Resource pooling
    - Multi-tenant sharing
  - Measured service
    - Direct or indirect economic model with measured use
- **Deployment models**
  - Private
    - Single administrative domain, limited number of users
  - Community
    - Different administrative domains with common interests & proc.
  - Public
    - People outside of institute's administrative domain
  - Hybrid
    - Federation via combination of other deployment models
- **Service models**
  - Software as a Service (SaaS)
    - Direct (scalable) hosting of end user applications
  - Platform as a Service (PaaS)
    - Framework and infrastructure for creating web applications
  - Infrastructure as a Service (IaaS)
    - Access to remote virtual
    - Machines with root access

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf


NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Special Publication 800-145

The NIST Definition of Cloud Computing

Recommendations of the National Institute of Standards and Technology

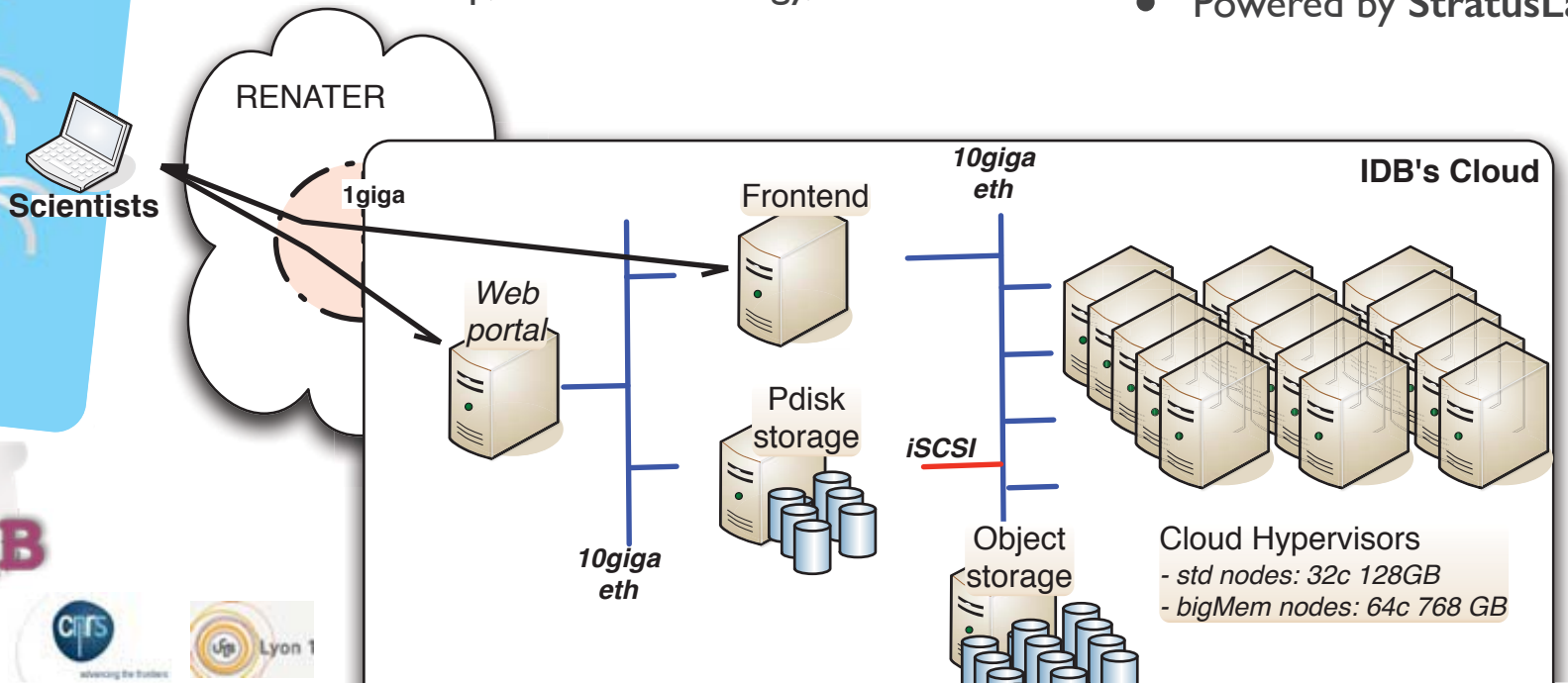
Peter Mell
Timothy Grance

Cloud IDB

- **Cloud workbench for Biology**
  - **Infrastructure Distributed for Biology**
    https://idee-b.ibcp.fr/cloud.html
  - **Running since Sept. 2011**
    IBCP FR3302 CNRS-Univ. lyon1, **Lyon, France**
  - opened to **Biology community**
  - **14 bioinformatics appliances:** Galaxy portal,
    standard compute nodes, proteomics, virtual desktop, structural biology, ...

- **+70 users** from all IFB regional centers
  PRABI 16, APLIBIO 28, RENABI-NE 13, -GO 7, -SO 2, -GS 5
- **VMs up to 32cores-768GB RAM**

- **Infrastructure**
  - **Compute +900cores +4TB ram**
    - **Standard** nodes (32c-128GB)
    - **Bigmen** nodes (**64c 768GB**)
  - **Storage +250TB**
    - Virtual disks, large-scale object storage (S3)
  - Powered by **StratusLab** and **CEPH**



stratuslab

idéeB

IBCP

Lyon 1

French Institute of Bioinformatics - IFB

Mission : to make available core bioinformatics resources to the national/ international life science research community.

- To provide **support for national biology pro**grams
- To provide an **IT infrastructure** devoted to management and analysis of biological data
- To act as a middleman between the life science community and the bioinformatics/computer science research community

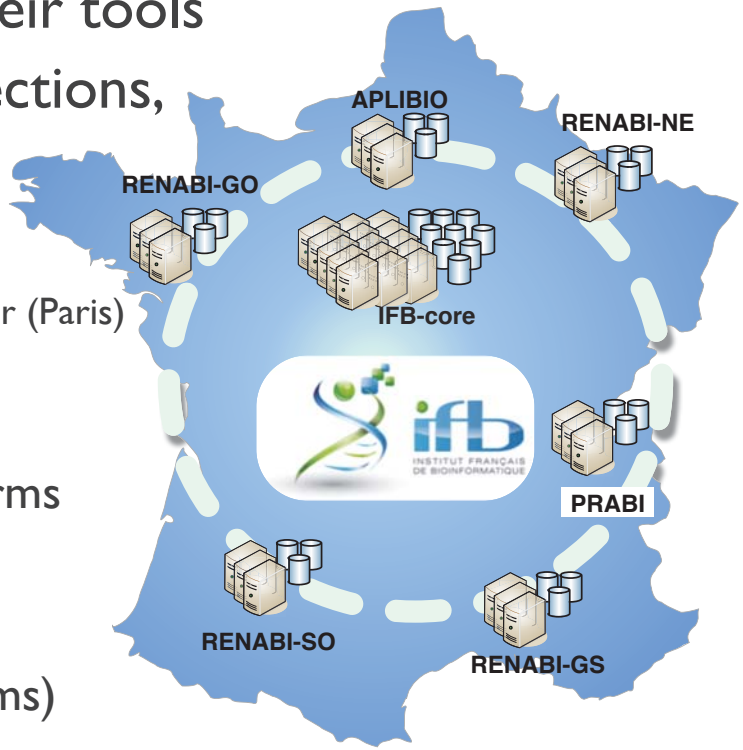
ifb
INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

ELIXIR French Node

- optimizing the **interactions and coordination** between the national level and ELIXIR and other ESFRI infrastructures in biomedical and environmental field,
- promoting **consistency and complementarities** between the components offered by the ELIXIR French node and those of other European nodes

IFB e-Infrastructure

- **Support** : help members to deploy and use their tools
- **e-infrastructure:** hardware, biology data collections, bioinformatics tools
- **Academic cloud for life science**
  - a **core** ressource 'IFB-core' hosted at **CNRS IDRIS** SC center (Paris)
  - **+ regional** resources
    - 6 regional bioinformatics centers with 2 clouds
    - 11,000 cores - 6 PB but +20 bioinformatics platforms
  - Create a **federation of clouds** for life science
- Technical organization
  - **GRISBI**: a national technical group (all national platforms)
  - Participation to **ELIXIR task forces**



Cloud Ressources	Location	# Compute Cores	# TB Storage	# TB RAM	Max VM size	Technology
IFB-core	**CNRS-IDRIS, Paris**	**100**	**50**	**1**	**40c 256GB**	**StratusLab**
IFB-core 2014	*CNRS-IDRIS, Paris*	*4,000*	*500*	-	*96c 1TB*	*StratusLab*
IFB-core 2015	*CNRS-IDRIS, Paris*	*10,000*	*2,000*	-	*96c 2TB*	*StratusLab*
idee-B	**PRABI-IBCP, Lyon**	**1,000**	**380**	**4**	**64c 768GB**	**StratusLab**
Genocloud	**IFB-GO, Rennes**	**240**	**8**	**1**	**-**	**ONE**

Extended cloud functionalities for bioinformatics

stratuslab

Native cloud services

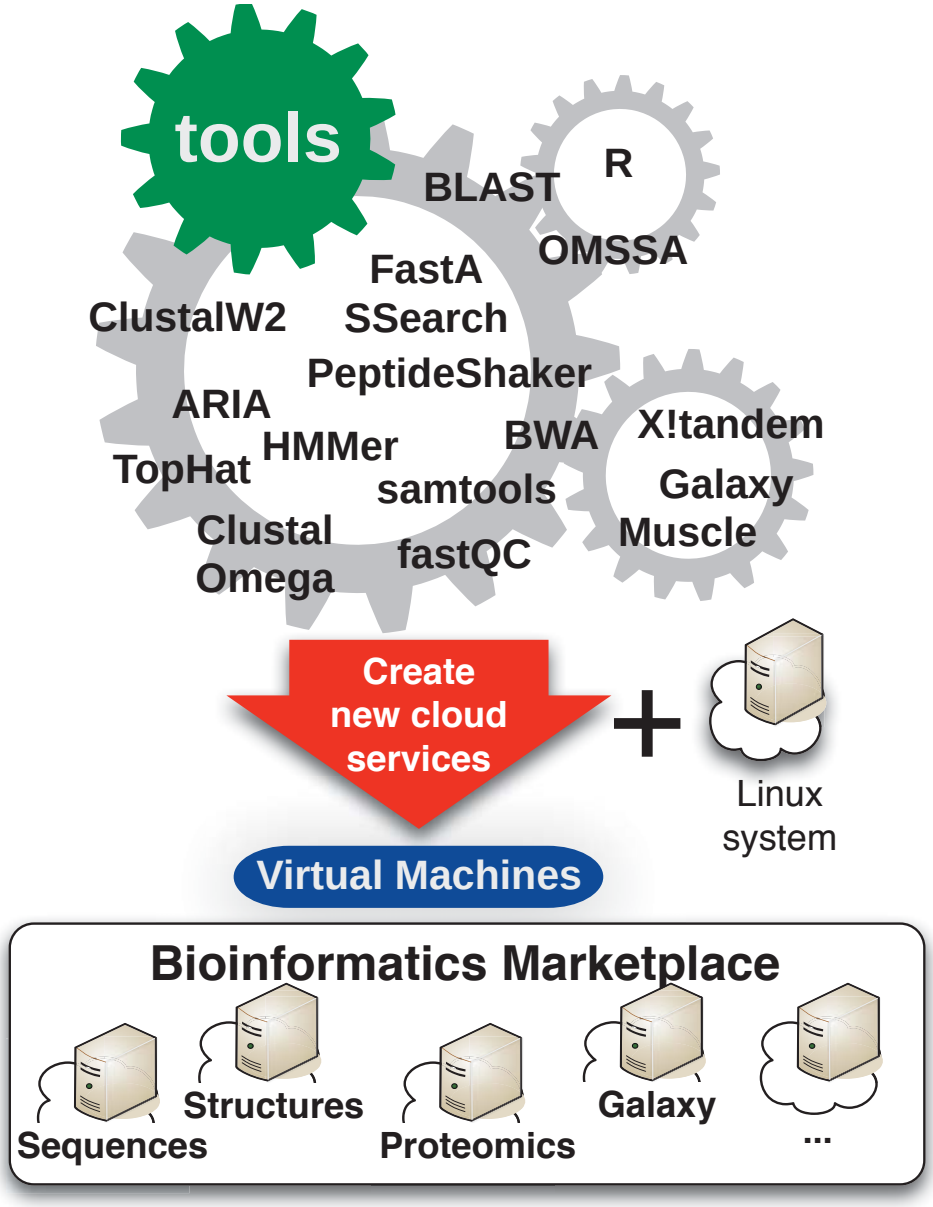
- Authentication
- Virtual machine management
- Persistent disk service
- Client CLI
- *etc.*

IDB

- **Bioinformatics appliances**
  - integrate bioinformatics tools and workflows
- Bioinformatics **marketplace**
  - focus on bioinformatics appliances
  - satisfy visibility contraints for some bioinformatics appliances (confidentiality)
- Bioinformatics **metadata "bio:tool"**
  - annotate appliances with attributes related to bioinformatics tools
  - help to select suitable bioinformatics appliances containing the required tools
- Integrated **Web** interface
  - VM & virtual disks management
  - filter bioinformatics appliances with "bio:tool"
- **CEPH** storage backend
  - large scale and distributed storage
  - reliable by replication
  - high-througput IO
  - single unified storage cluster for all interfaces: block, object and file system

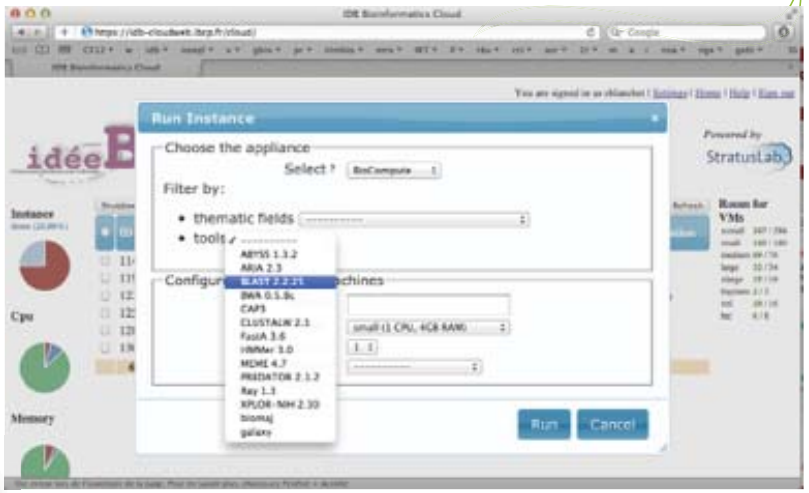
Bioinformatics cloud appliances

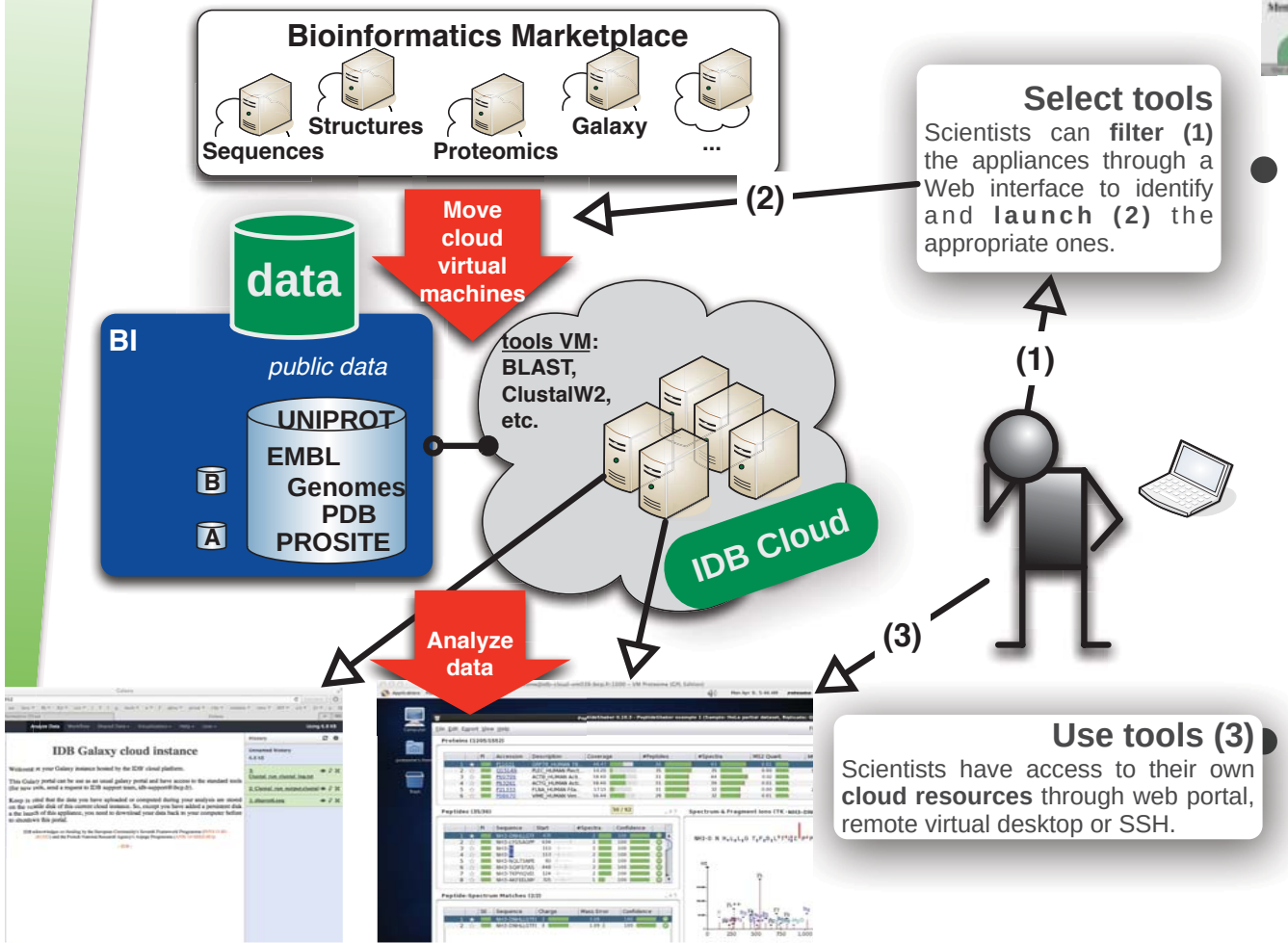
- Bioinformatics appliances are usual virtual machines
  - small : few GB, easy to convert in most virtualization formats
- Installed and pre-configured with bioinformatics tools
  - e.g. BLAST, Clustalw, ARIA, MEME, HMMer, TopHat, BWA, Samtools, etc.
- Recorded in a marketplace
  - devoted to bioinformatics



Run bioinformatics appliances

- **Bioinformatics marketplace**
  - both a virtual machines **repository**
    - Store life science VMs
  - and a **catalogue**
    - Help users to select the appropriate VM for their analysis

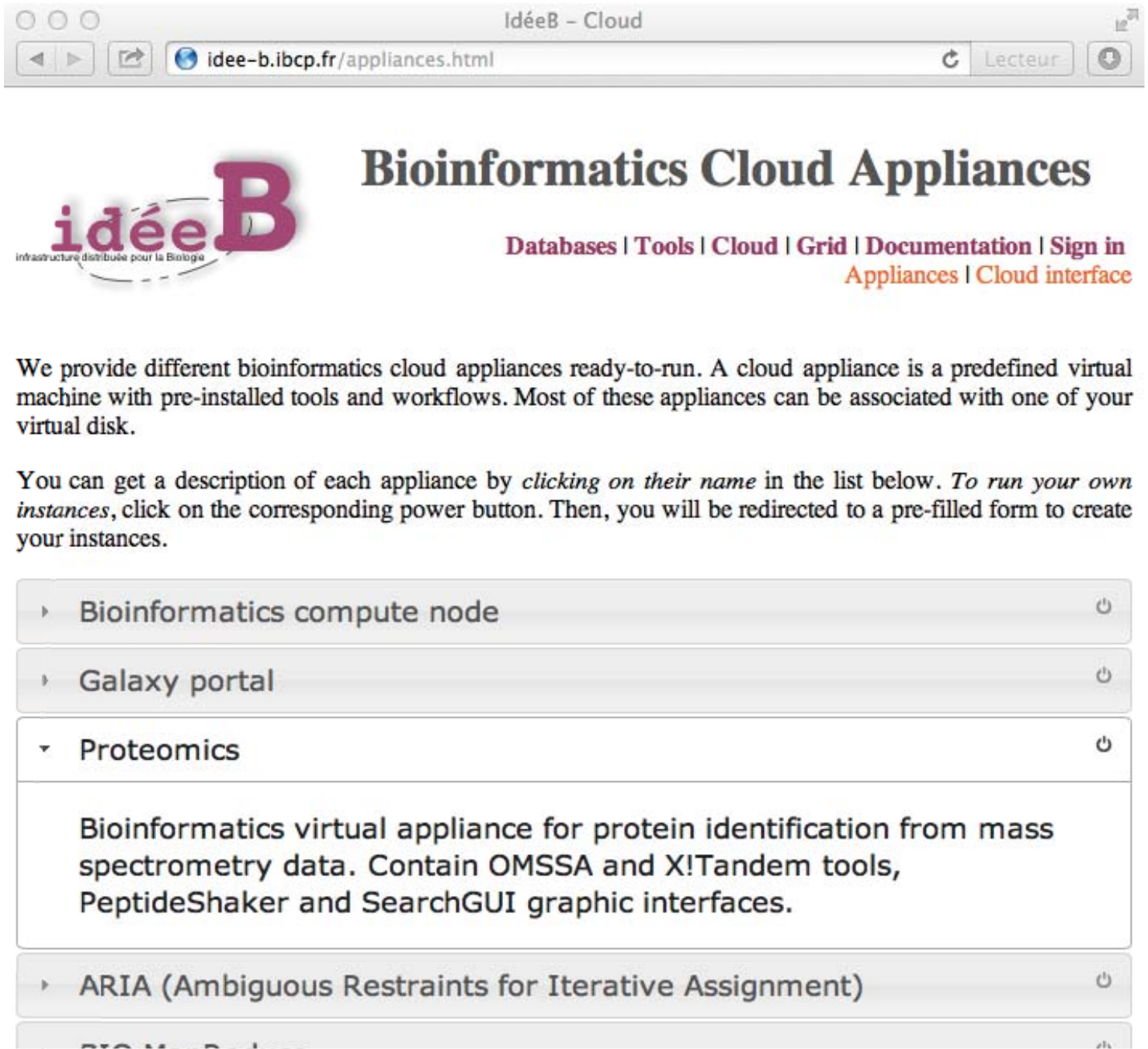




- **Filter** images with metadata related to bioinformatics
  - attribute <*bio:tool*> in VM manifests
  - scientists can select the appropriate appliance according to the tools required for their analyses
  - e.g. the BLAST tool

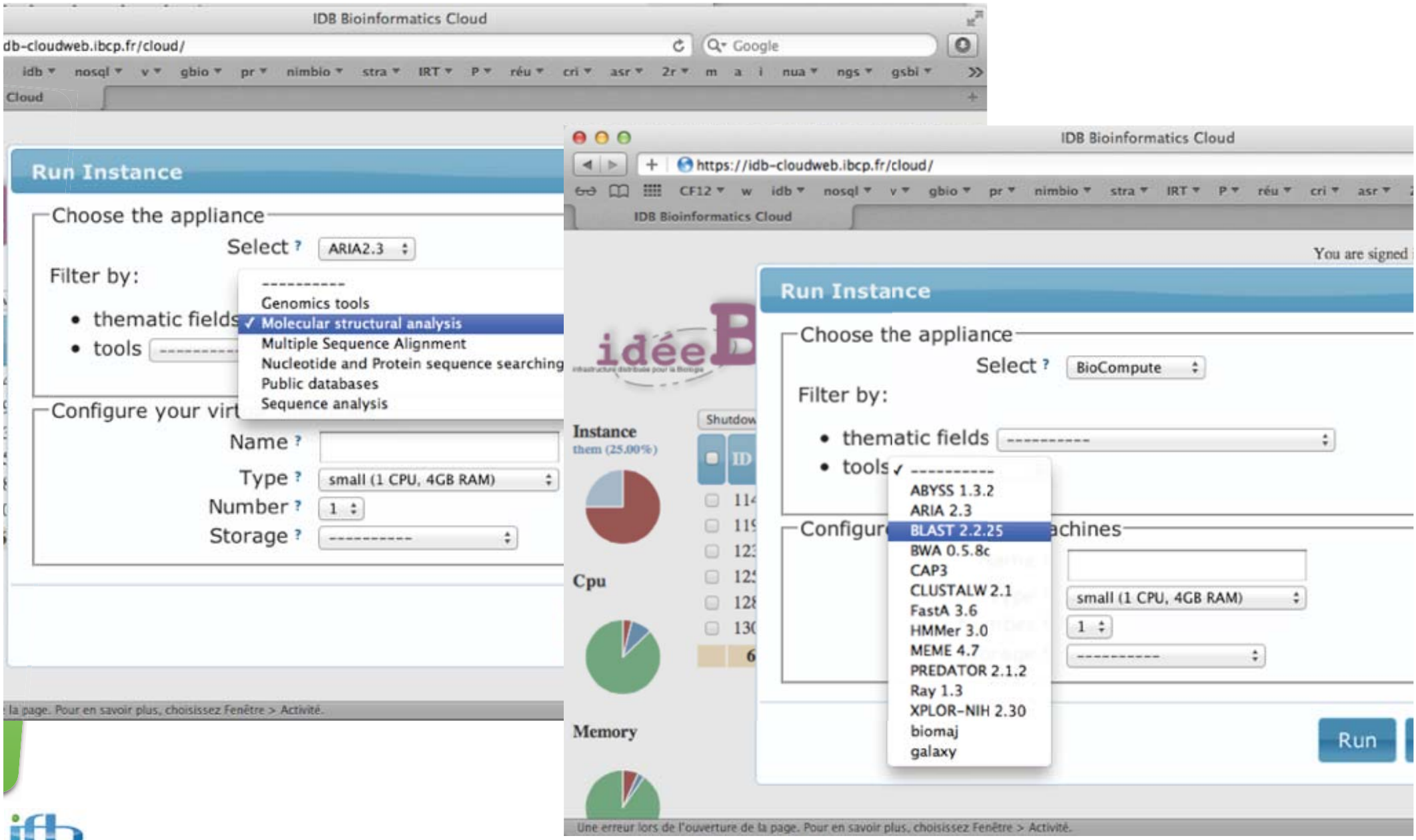
Deploy on several clouds

Appliances page

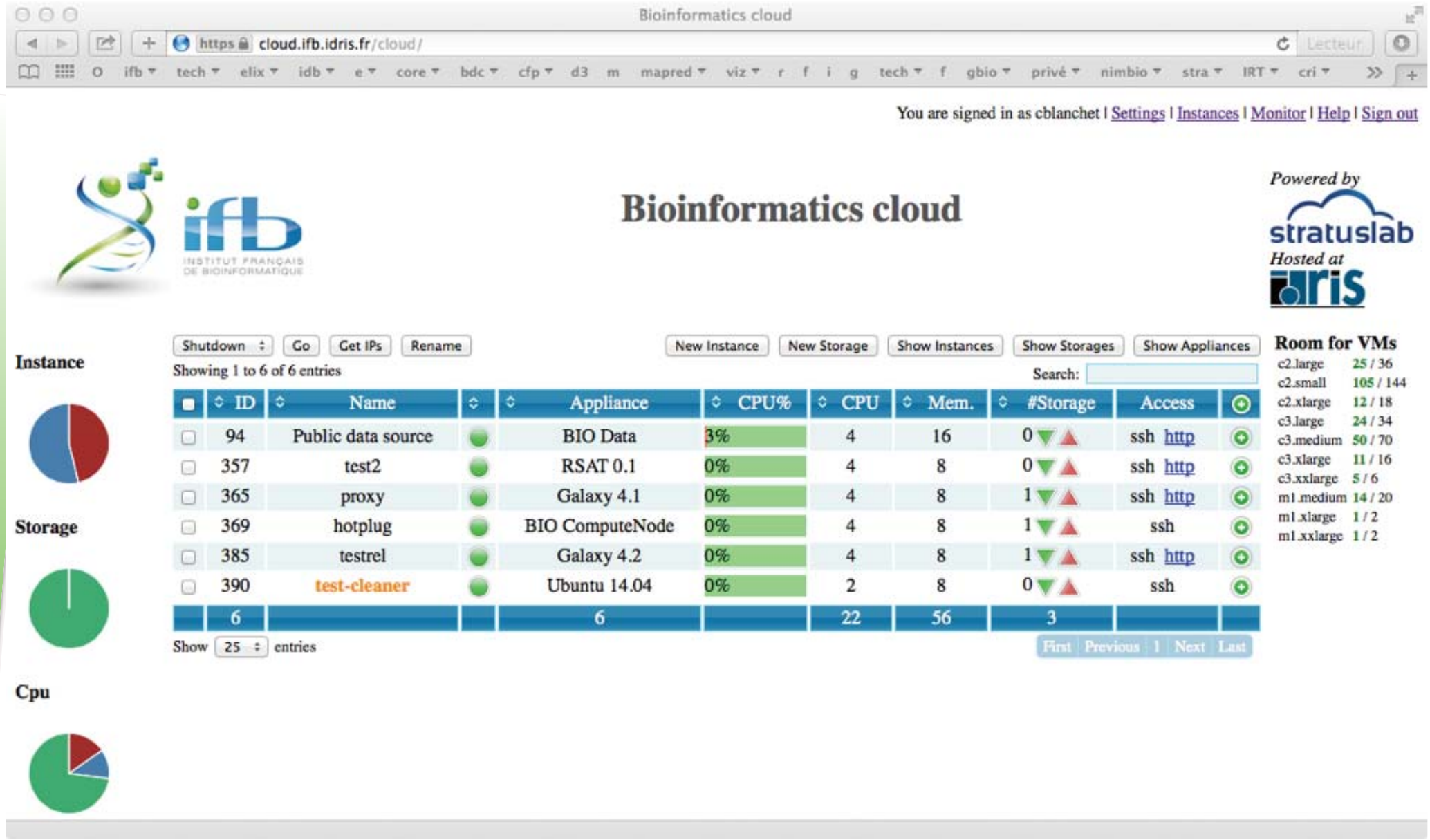


- List of existing appliances
- Appliance description and doc
- Direct launch
  - ‘Power’ button

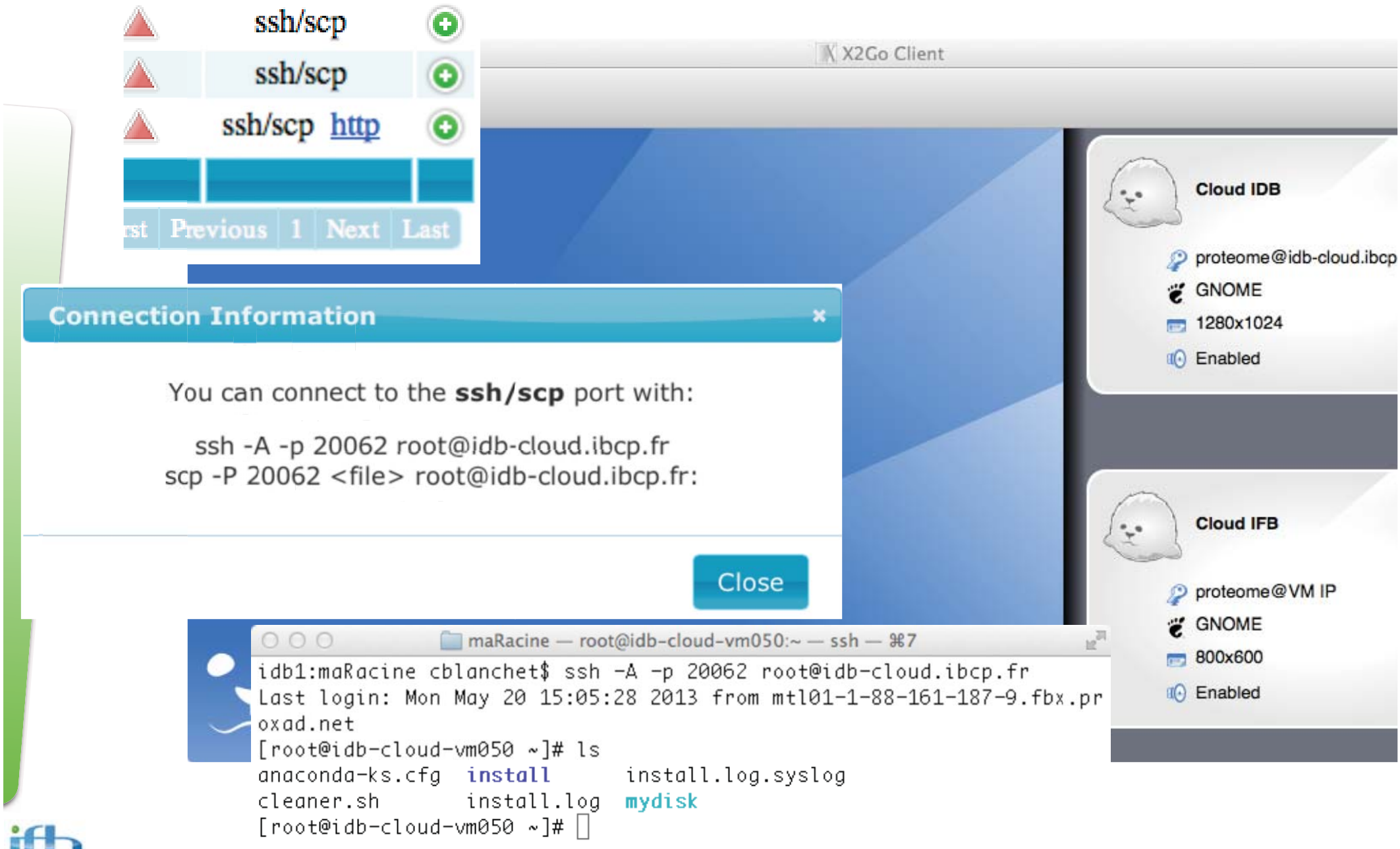
Filter appliances with tools description



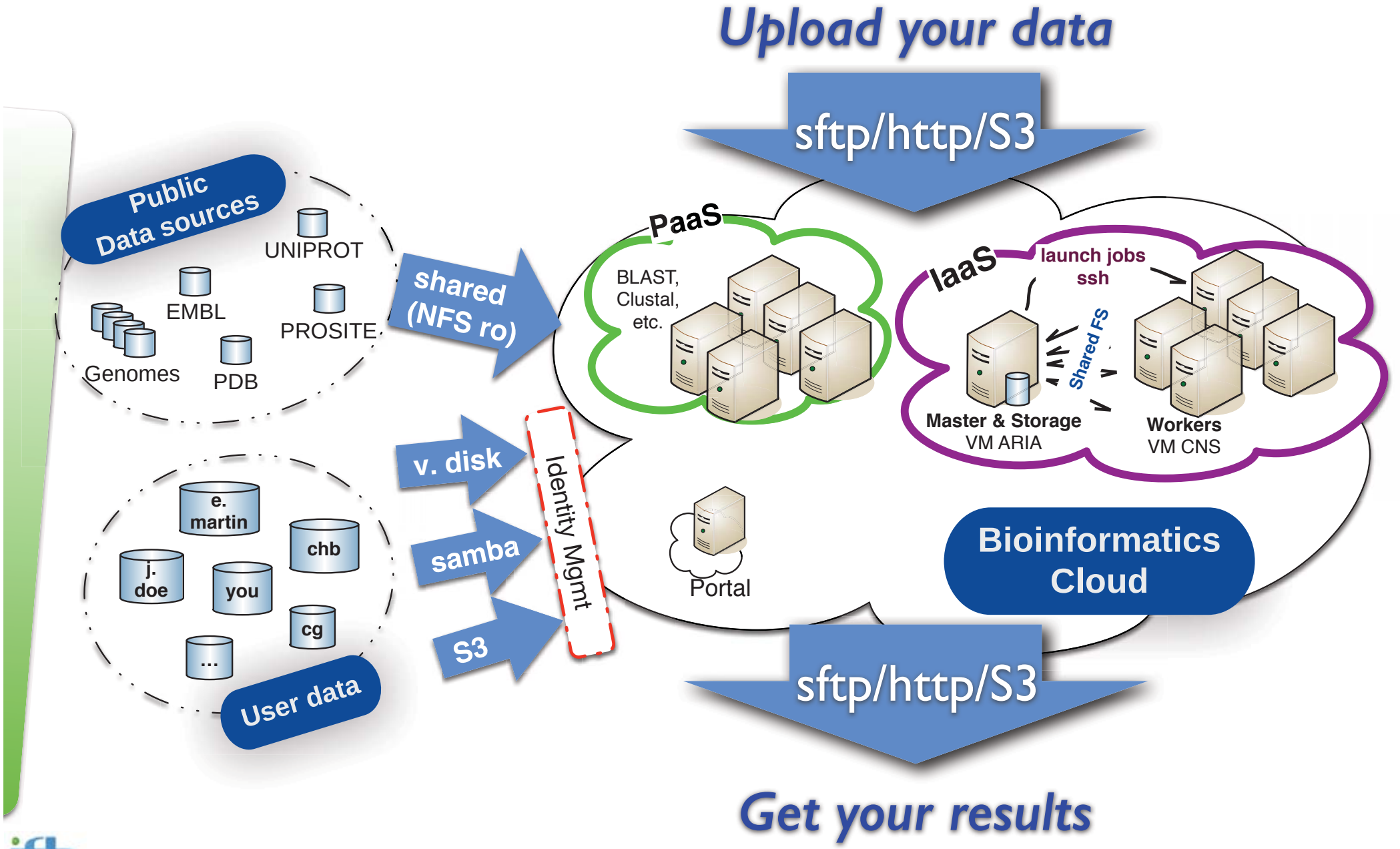
A cloud driven through a simple web interface



Connection to VMs

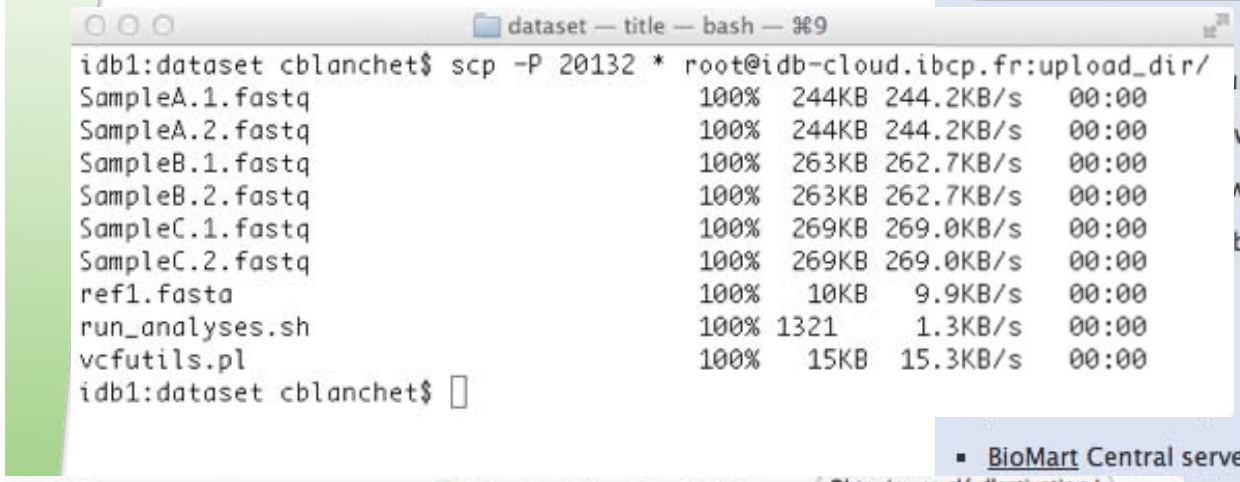


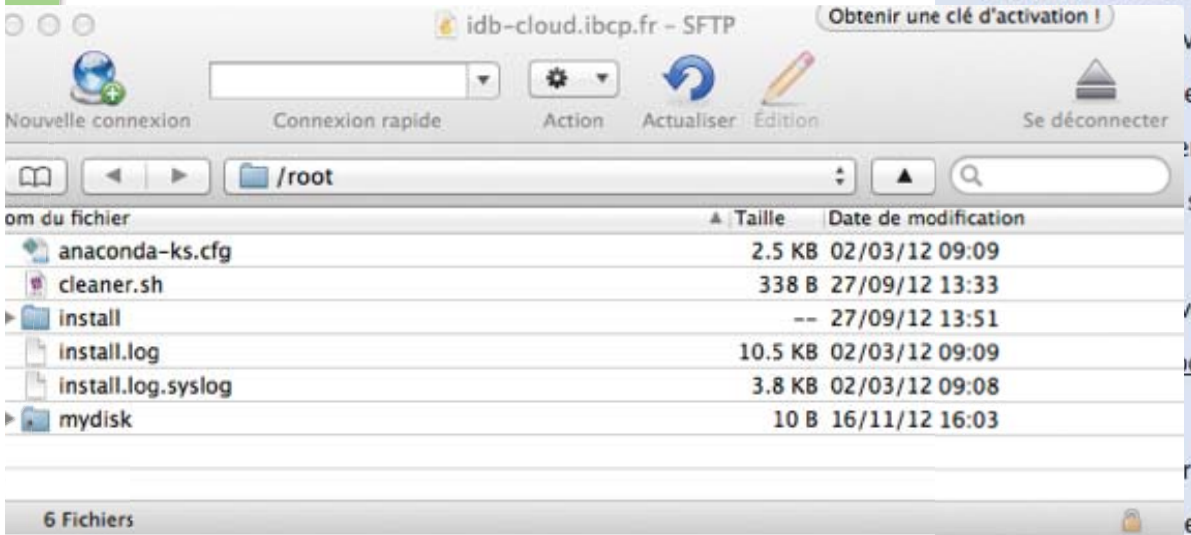
Cloud Storage for Biological Data

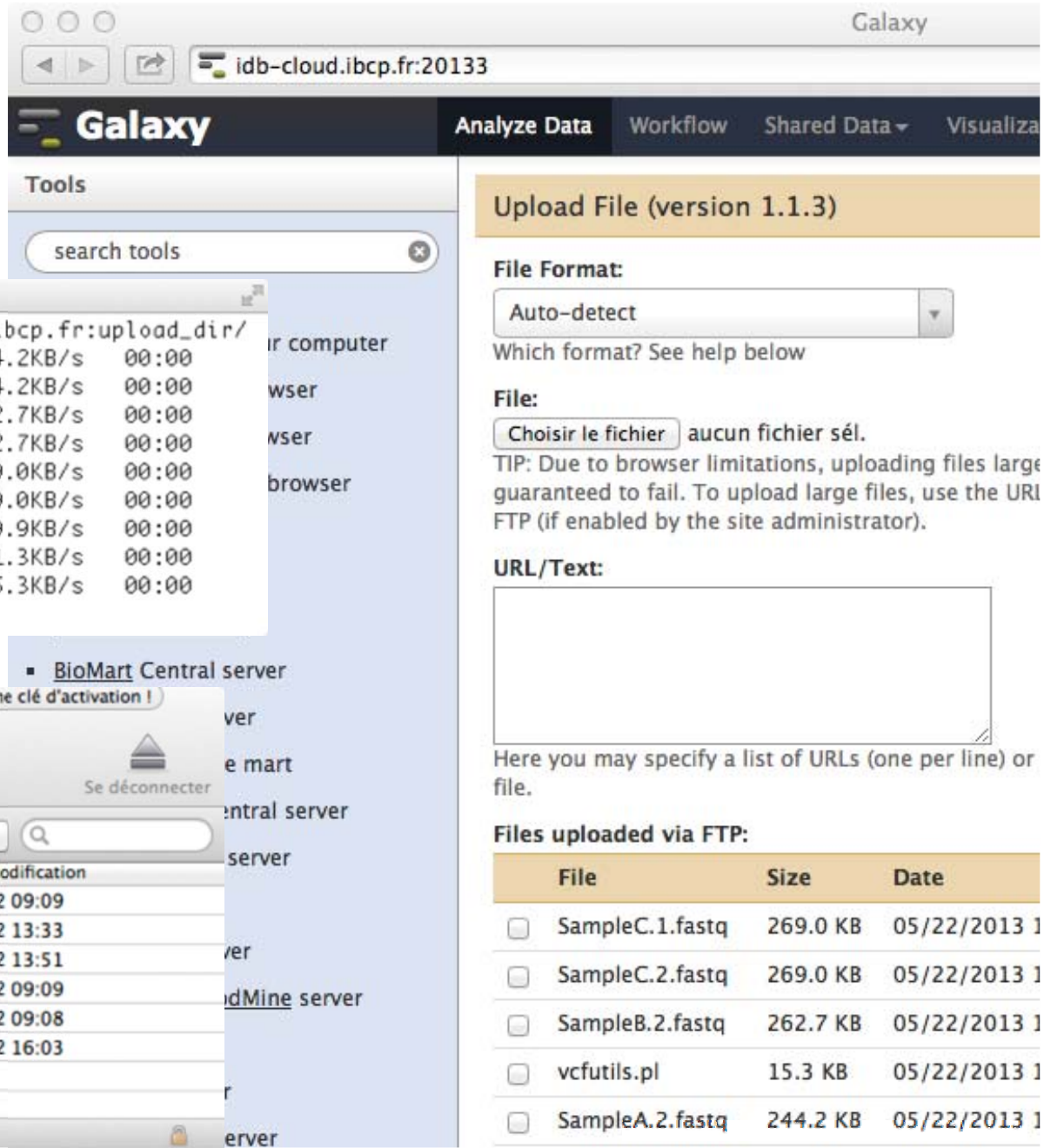


Exchanging data with VMs

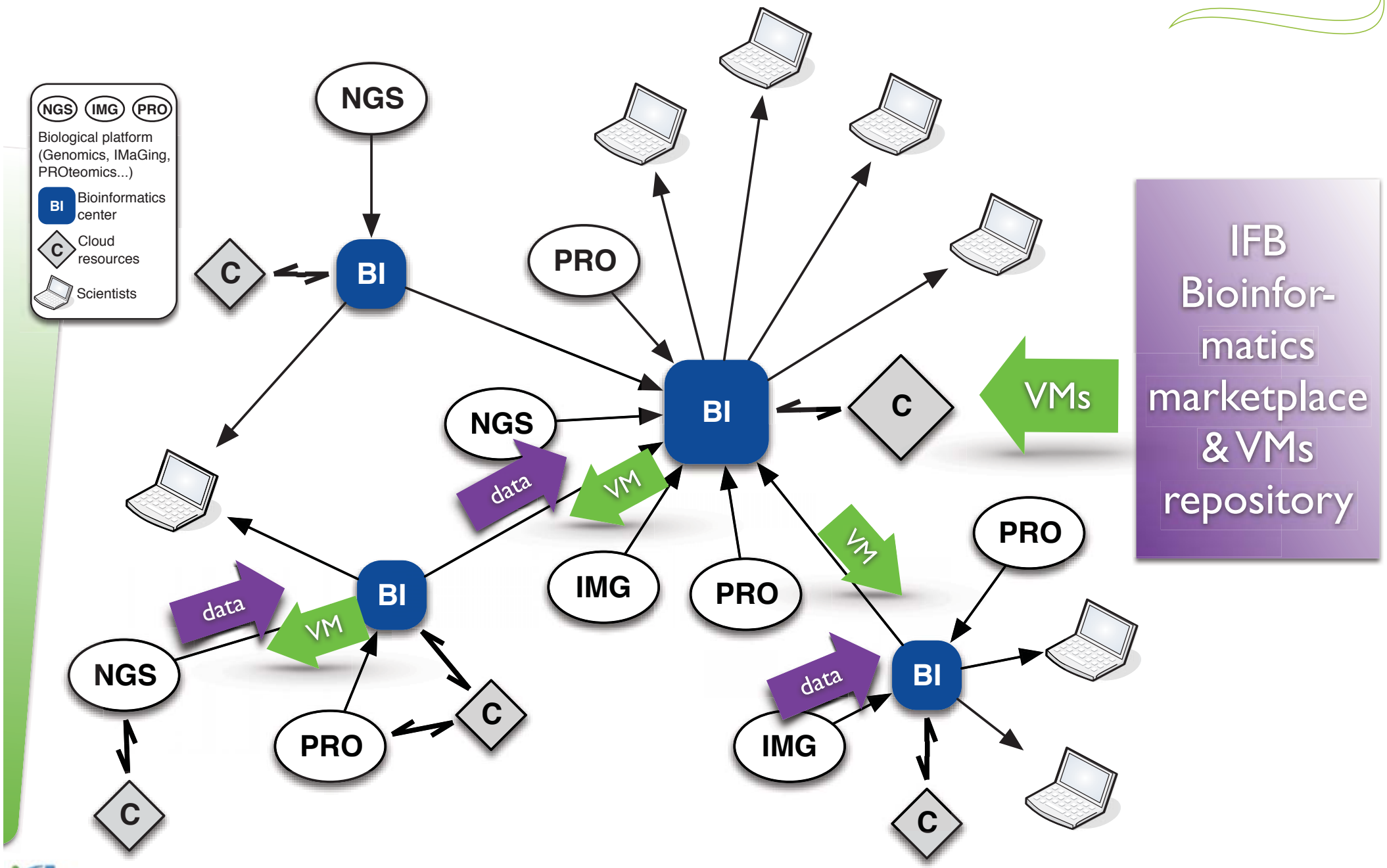
- CLI 'scp/sftp'
- GUI: Cyberduck, Transmit
- Integrated: Galaxy





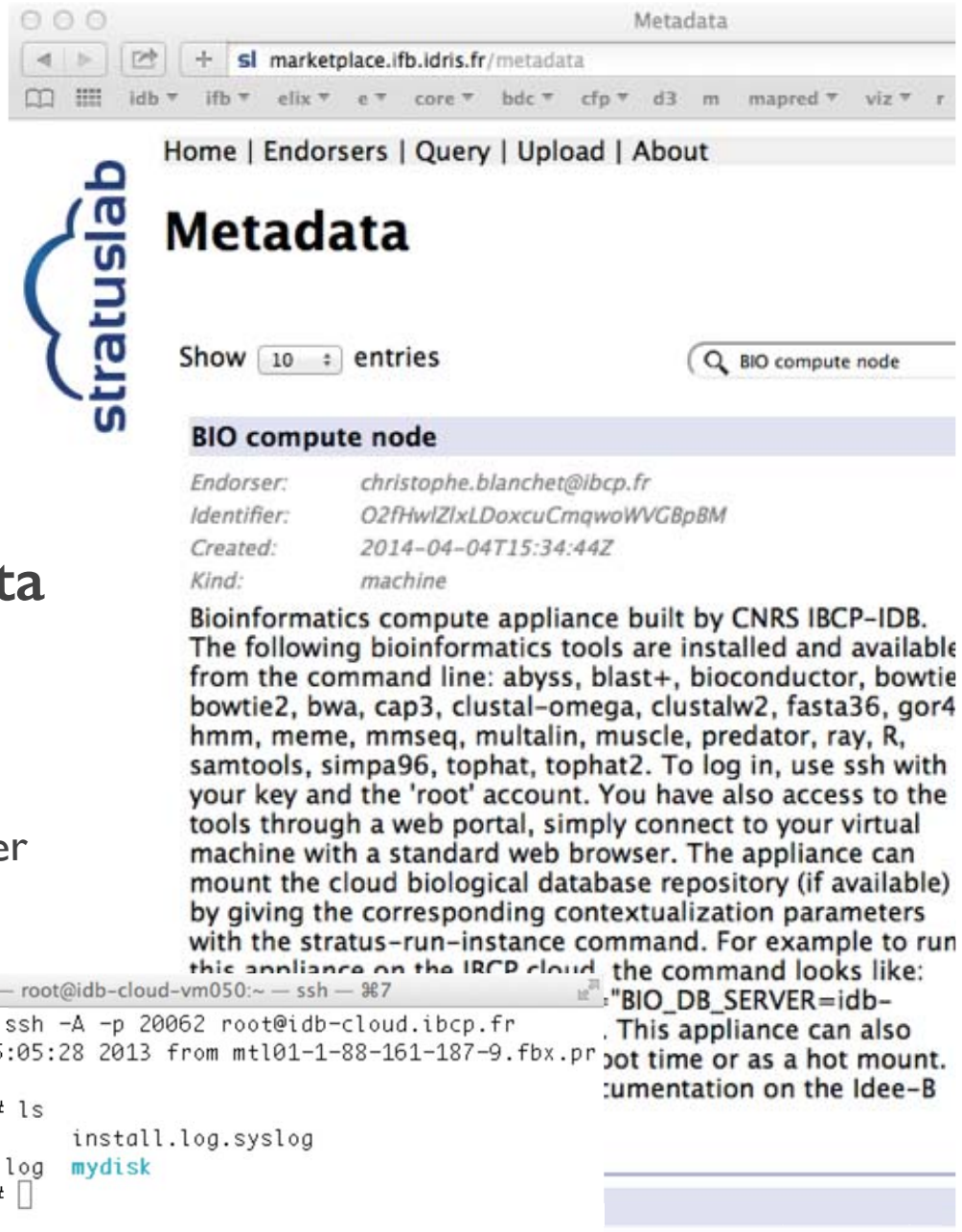


Moving VMs vs Data



Case 1: Standard Bioinformatics node

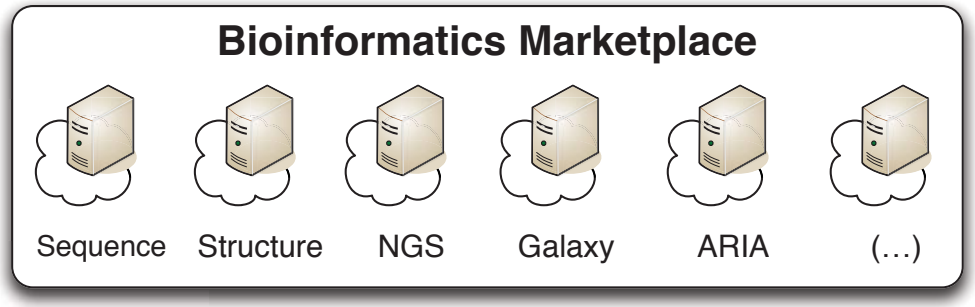
- appliance 'Biocompute'
- Use **your own** instance(s)
- With **pre-installed standard bioinformatics tools**
  - BLAST, FastA, SSearch,HMM,...
  - ClustalW2, Clustal-Omega, Muscle,..
  - Bowtie(2), BWA, samtools, ...
  - MEME, R, etc.
- Connected to **public reference data**
  - Uniprot, EMBL, genomes, PDB, etc.
  - Automaticaly shared to the VMs
- **Cluster** mode
  - turn several instances in a single virtual cluster
  - shared file system
  - batch scheduling



Case 2: Cloud Galaxy portal for NGS analyses

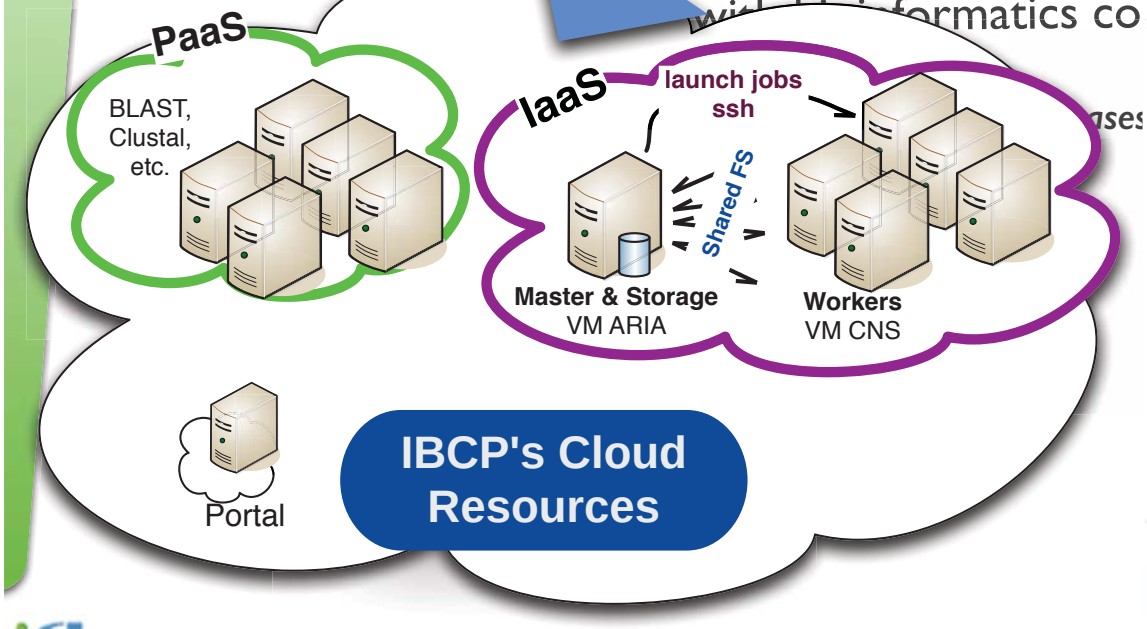
- Analyse NGS data
  - **portal Galaxy** is widely used in the community
  - connected to large public data: sequences and indexes
  - large user data (GBs)
- **Preserve workflows and results** (cloud virtual disk)
- Different **domain-specific instances** (RNAseq, CHIPseq, *etc.*)
- **For training**: create a special instance derived from the main instance but with dedicated datasets
- Help the integration of **monthly updates**

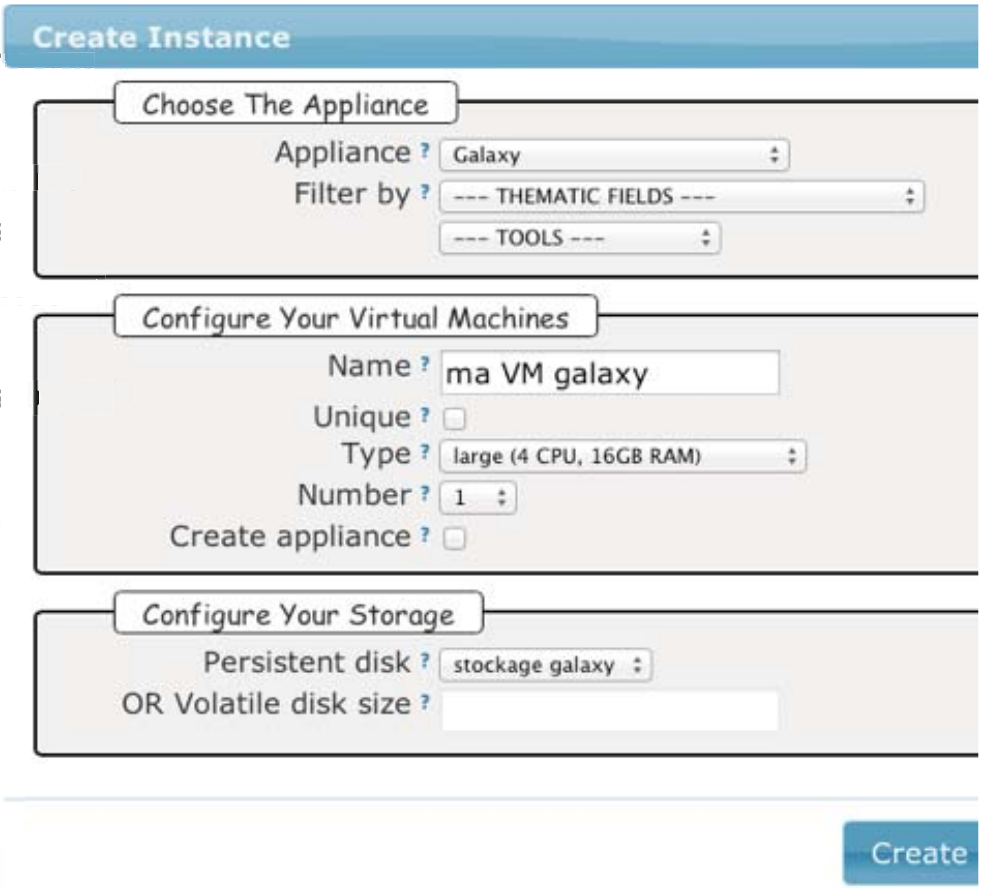
Run your Galaxy Portal on Cloud



- Stay Connec Standard Data &
  - User data datafiles or attache pdisk
  - Refe en ases: mount biodata server s
  - Tool e-installed ones or install yours
  - formatics co

Launch Instances

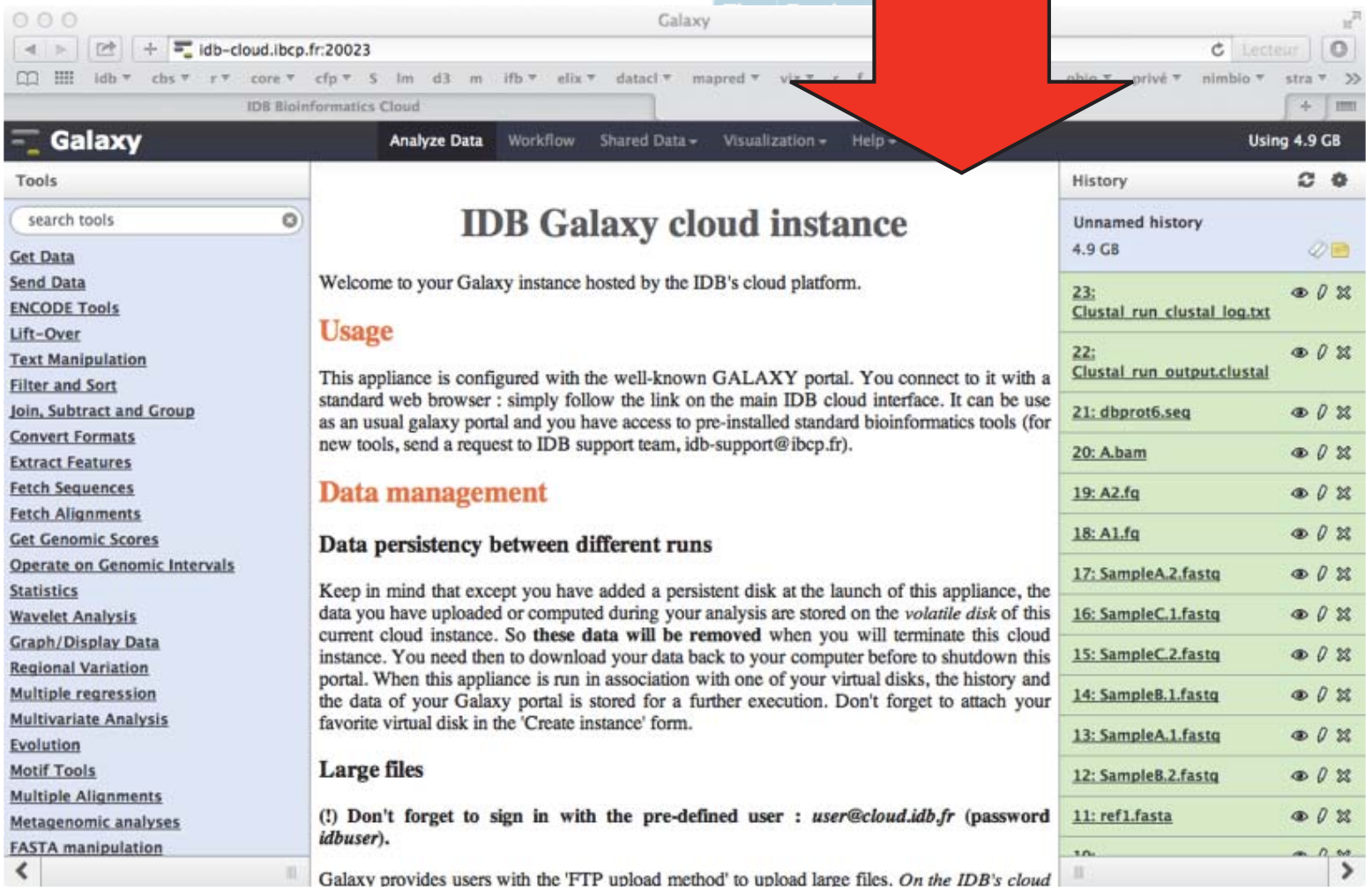




Connect to your Galaxy Portal

☐	5012	mapred#2		BIO MapReduce	0%	4	16	0	ssh/scp
☐	5013	mapred#3		BIO MapReduce	0%	4	16	0	ssh/scp
☐	5140	ma VM galaxy		Galaxy	0%	4	16	1	ssh/scp http
	10			7		52	269	2	

Show 25 entries

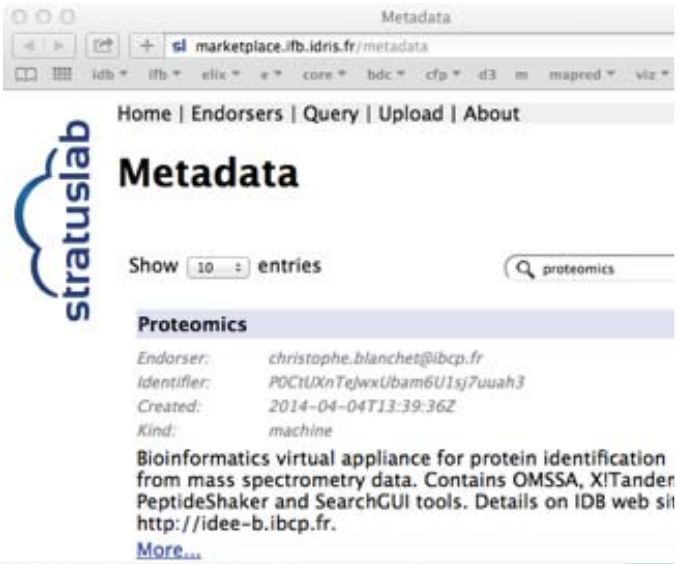


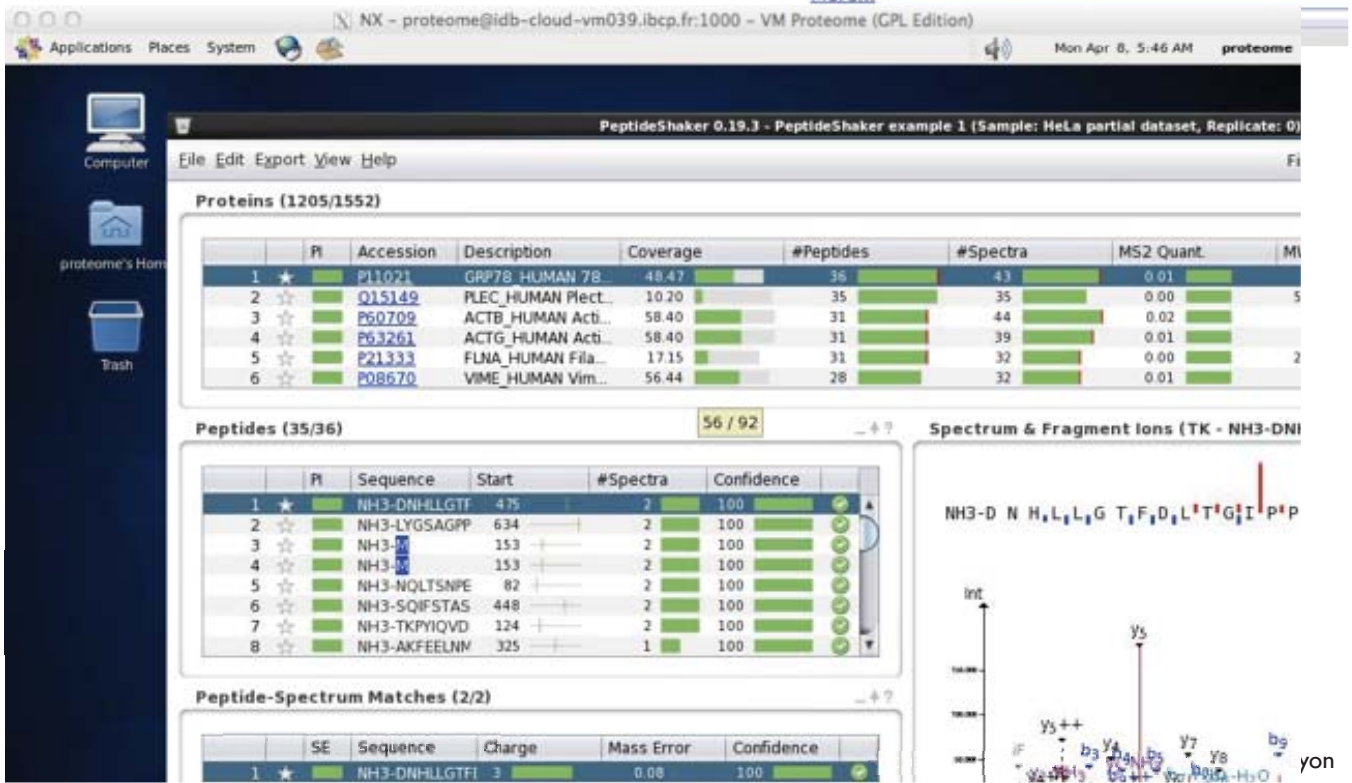
Advantages of Cloud for Galaxy

- Added value of cloud for Galaxy,
  - for scientific analyses: user-specific resources, isolated, different domain-specific instances (RNAseq, CHIPseq, Variants, ...)
  - for training: create a special instance derived from the main but with dedicated datasets
    - Examples of training with Galaxy: Mai 2013 Galaxy Lille, Nov 2013 Aviesan Bioinformatics School
  - For integration of monthly updates
  - for development & operations (DevOps): different versions at the same time
- Bioinformatics cloud (e.g. IDB)
  - Tightly connected to existing bioinformatics resources
  - Linked to public biological databases
  - In collaboration with the French Institute of Bioinformatics

Case 3: Proteomics virtual desktop

- Motivation
  - Collaboration with a mass spectroscopy platform
  - Running out of space on their local resources
- Protein identification tools
  - Mass experimental data
  - Reference databases : nr, Swiss-Prot
  - Reference screening tools: OMSSA, X!Tandem
- User interface
  - **Remote Virtual Desktop (NX)**
  - Reference GUIs
    - SearchGUI
    - PeptidShaker



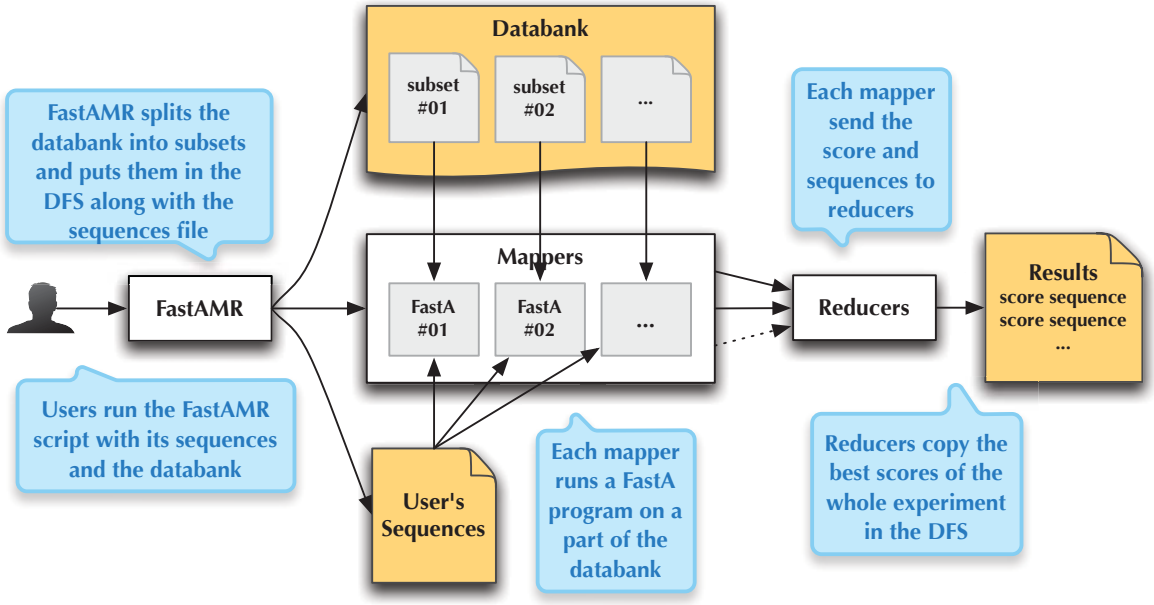


OMSSA

X!

Case 4: Hadoop for Life Science

- Provide turnkey virtual machine with pre-configured mapreduce framework
  - **Accelerate biological bigdata analysis**
  - **Hadoop MapReduce** 1.0.4
- Appliances (2)
  - provide standard hadoop: including mapreduce and HDFS
  - with integrated bioinformatics tools
- Example of sequence similarity searching
  - FastA & SSearch
  - deploy database of sequences in HDFS
  - compare each structure to others



Developed in the context of the French project MapReduce, ANR ARPEGE

BIO MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: J46wxrwGLdnoSskmb0JlfGv8UpY
Created: 2013-05-17T11:13:08Z
Kind: machine

This appliance provides an easy way to deploy a Hadoop MapReduce cluster (v1.0.4) with pre-installed bioinforma tools such as FastA. You just need to run the bash script hadoop-create-cluster with a nodes list and an username parameters and wait few minutes until the process is completed. Then you can login to the user account and submit your Hadoop jobs or interact with Hadoop filesyst You can extend a current cluster by submitting a list of n nodes to the script. A FastA MapReduce example is also provided under the directory /usr/local/share/fasta. (Cre for the French project MapReduce, ANR ARPEGE, 2010-20 mapreduce.inria.fr)

More...

Hadoop MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: BttU7uNM5UT1haUigVS7xySI2rr
Created: 2013-05-17T09:33:52Z
Kind: machine

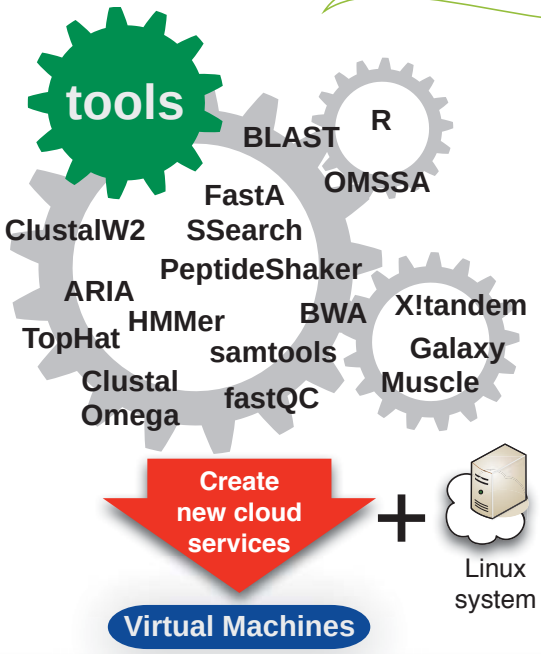
This appliance provides an easy way to deploy an Hadoop MapReduce cluster (v1.0.4). You just need to run the bash script hadoop-create-cluster with a nodes list and an username in parameters and wait few minutes until the process is completed. Then you can login to the user acc and submit your Hadoop jobs or interact with Hadoop filesystem. Enjoy! In addition, you can extend a current cluster by submitting a list of new nodes to the command (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, mapreduce.inria.fr)

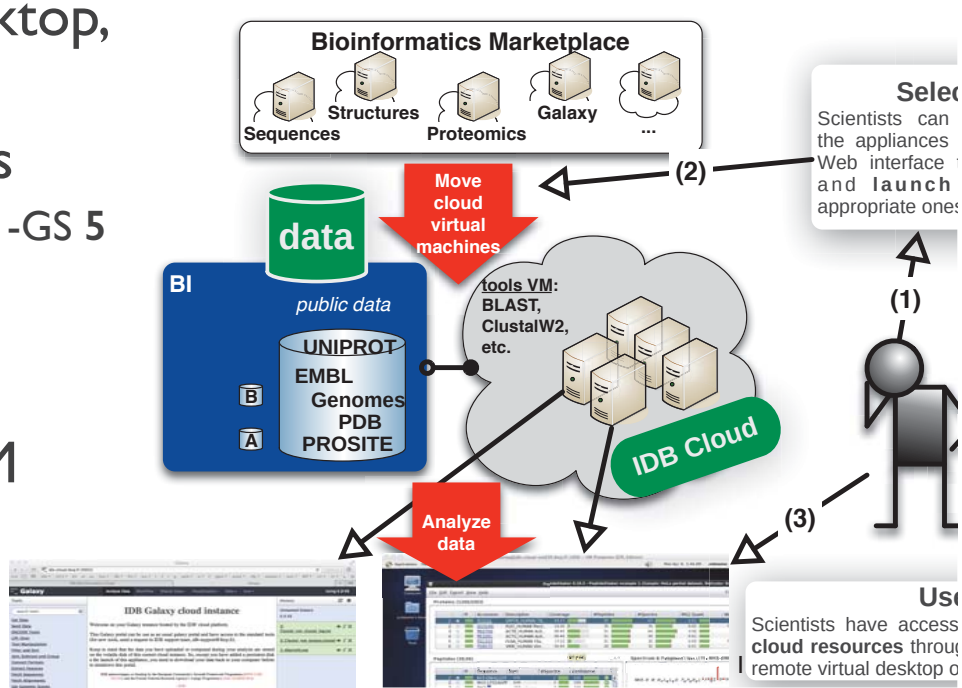
More...

Cloud it be done ?

- **IFB's cloud for life science simplify access to biological data and tools**
  - integrate tools and pipelines in turnkey cloud appliances
  - is tightly connected to existing **bioinformatics resources,** *e.g.* public reference data sources…
  - **14 bioinformatics appliances:** standard compute nodes, proteomics virtual desktop, Galaxy portal, structural biology…
  - **+70 users** from all IFB regional centers

    PRABI **16**, APLIBIO **28**, RENABI-NE **13**, -GO **7**, -SO **2**, -GS **5**
- **Bioinformatics marketplace**
  - store images related to life science
  - help users to select the appropriate VM for their analysis





Perspectives

- Create bioinformatics appliances
  - by the experts of the domains
  - make them available to the scientists

- IFB established priorities: 5 scientific domains
  - Microbial Bioinformatics
  - Evolutionary bioinformatics
  - Plant bioinformatics
  - Structural Biology
  - NGS data processing

- and 3 technical pilots
  - Appliances interoperability between different cloud infrastructures
  - Distributing biological data with distributed noSQL engine
  - Live remote cloud processing of sequencing data

Questions ?

Acknowledgments

- Clément Gauthey (IDB-IBCP)
- **StratusLab** members
- IDB's co-funding by
  European Community's Seventh Framework Programme (INFSO-RI-261552)
  French National Research Agency's Arpege Programme (ANR-10-SEGI-001).
- IFB's funding by French program PIA INBS 2012