

# **Fourniture de services cloud pour la biologie**

## **Exemple du cloud IDB-IBCP**

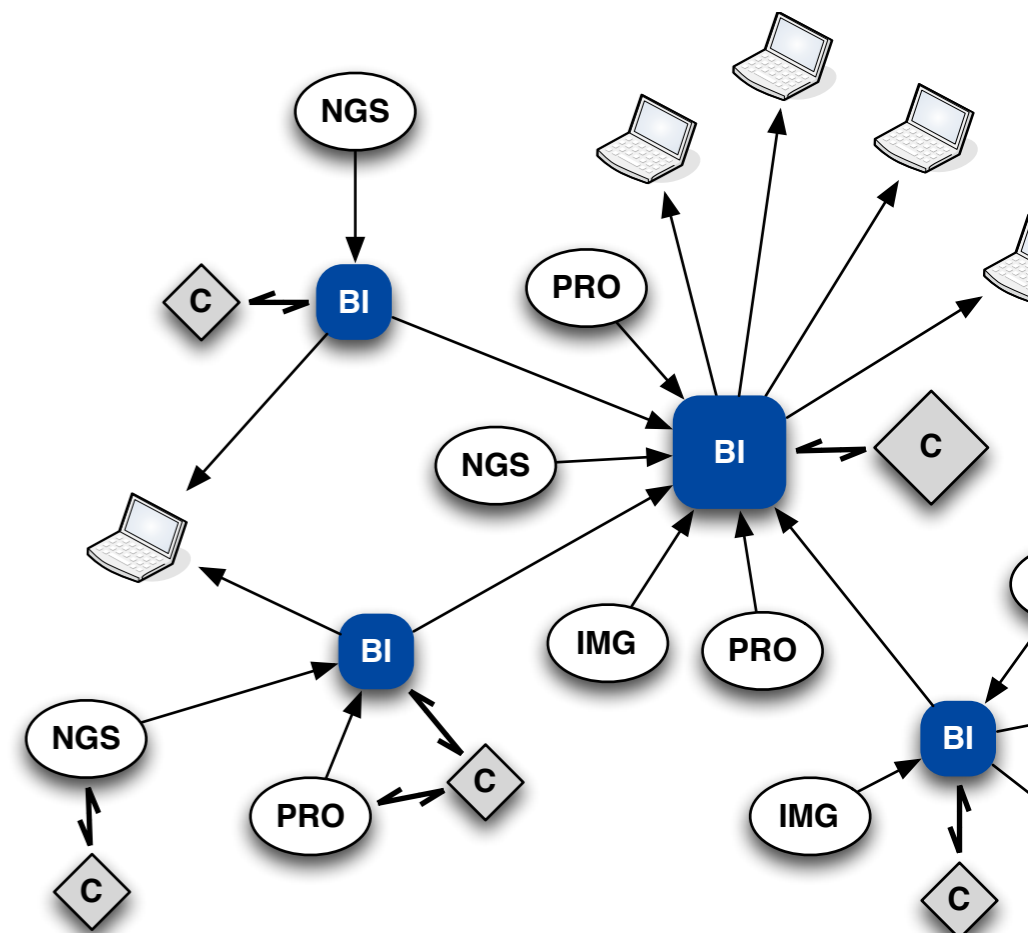
**Christophe Blanchet, Clément Gauthey**

**Institut de Biologie et Chimie des Protéines**  
**Plateforme "Infrastructure Distribuée pour la Biologie - IDB"**  
**CNRS-IBCP FR3302 - LYON - FRANCE - <http://idee-b.ibcp.fr>**

IDB acknowledges co-funding by the **European Community's Seventh Framework Programme** ([INFSO-RI-261552](#)), the **French National Research Agency's Arpege Programme** ([ANR-10-SEGI-001](#)) and by the **French Institute of Bioinformatics (IFB)**

# 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**

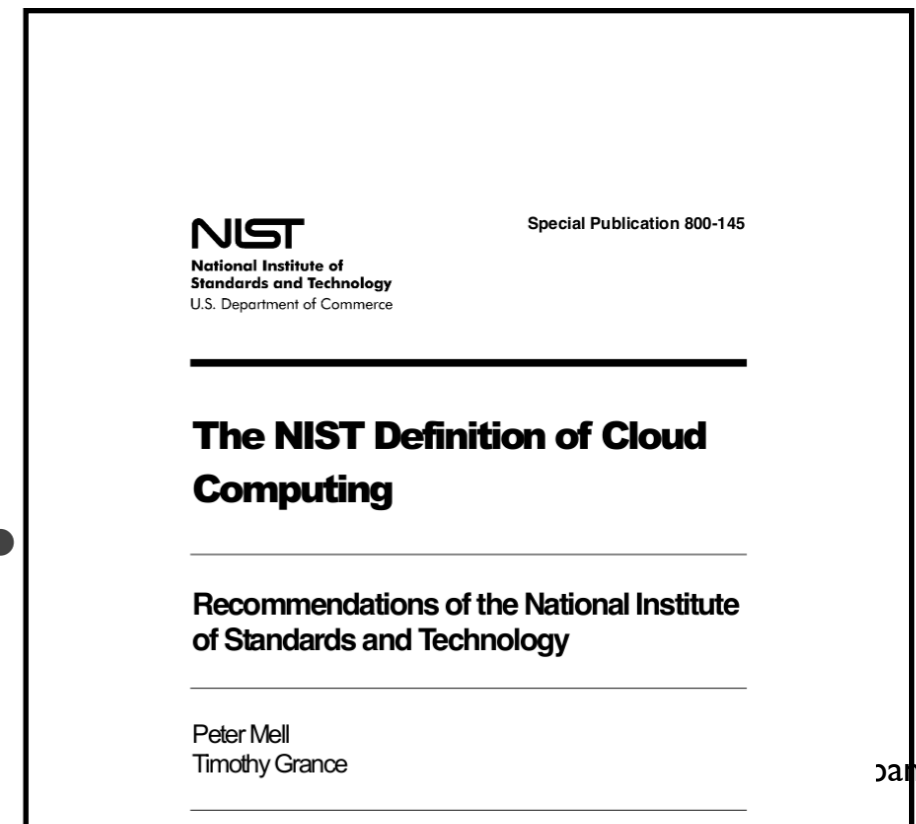
- Private
  - Single administrative domain, limited number of users
- Community
  - Different administrative domains with common interests & proc.
- Public
  - People outside of institute's administrative domain

- 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



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

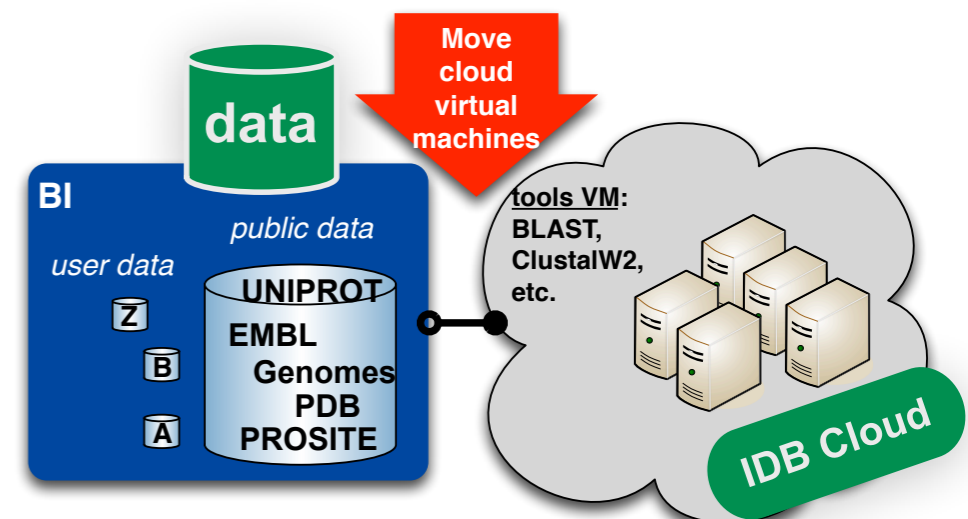
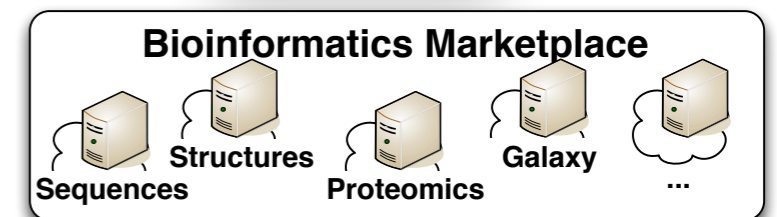
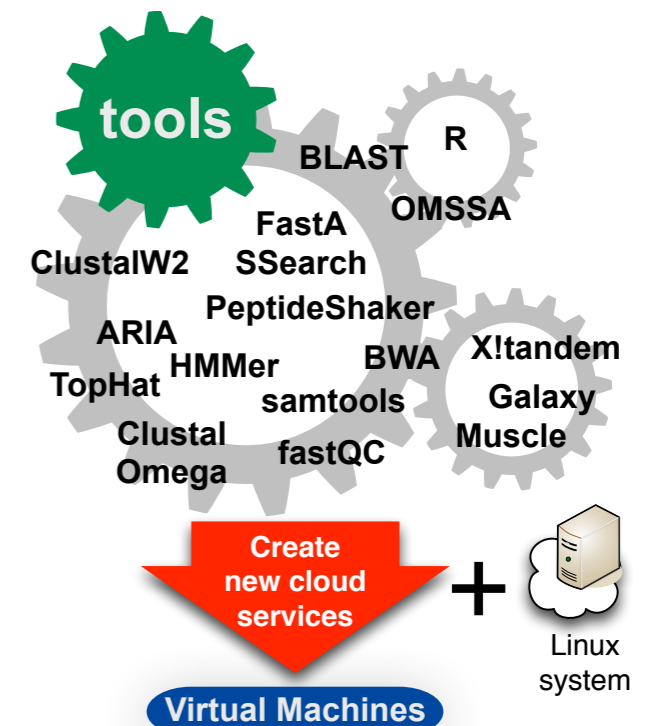
# IDB Cloud and Bioinformatics Appliances

- **Cloud workbench for Biology**

- <https://idee-b.ibcp.fr/cloud.html>
- Running since Sept. 2011  
CNRS-IBCP FR3302, 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



# Cloud extended services



## Native cloud services

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

A large red arrow pointing to the right, containing the white text "IDB".

IDB

- **Bioinformatics Marketplace**
  - find appropriate appliances more easily.
  - reduce “noise” in the central Marketplace
  - respect visibility constraints for the bioinformatic appliances, such as confidentiality
- **Bioinformatics metadata “bio:tool”**
  - additional elements related to bioinformatics tools to annotate appliances
  - help users to search for the tools themselves or the type of analysis
  - select suitable bioinformatics appliances containing the required tools
- **Integrated Web interface**
  - VM & virtual disks management
  - browse bioinformatics appliances with “bio:tool” MDz
- **CEPH storage backend**
  - large scale and distributed storage
  - reliable by replication
  - high-throughput IO
  - single unified storage cluster for all interfaces: block, object and file system



# Driven through a simple web interface

**Run Instance**

Choose the appliance  
Select ? BioCompute

Filter by:

- thematic fields
- tools ✓
  - ABYSS 1.3.2
  - ARIA 2.3
  - BLAST 2.2.25**
  - BWA 0.5.8c
  - CAP3
  - CLUSTALW 2.1
  - FastA 3.6
  - HMMer 3.0
  - MEME 4.7
  - PREDATOR 2.1.2
  - Ray 1.3
  - XPLOR-NIH 2.30
  - biomaj
  - galaxy

Configure your virtual machines  
small (1 CPU, 4GB RAM)  
1

Run Cancel

**Create Instance**

Choose the appliance  
Select ? BioCompute

Filter by ? --- THEMATIC FIELDS ---

--- TOOLS ---

Configure your virtual machines

Name ? my virtual machine

Type ? xlarge (8 CPU, 27GB RAM)

Number ? 1

Storage ? mydata

Create appliance ?

Create

**Bioinformatics cloud**

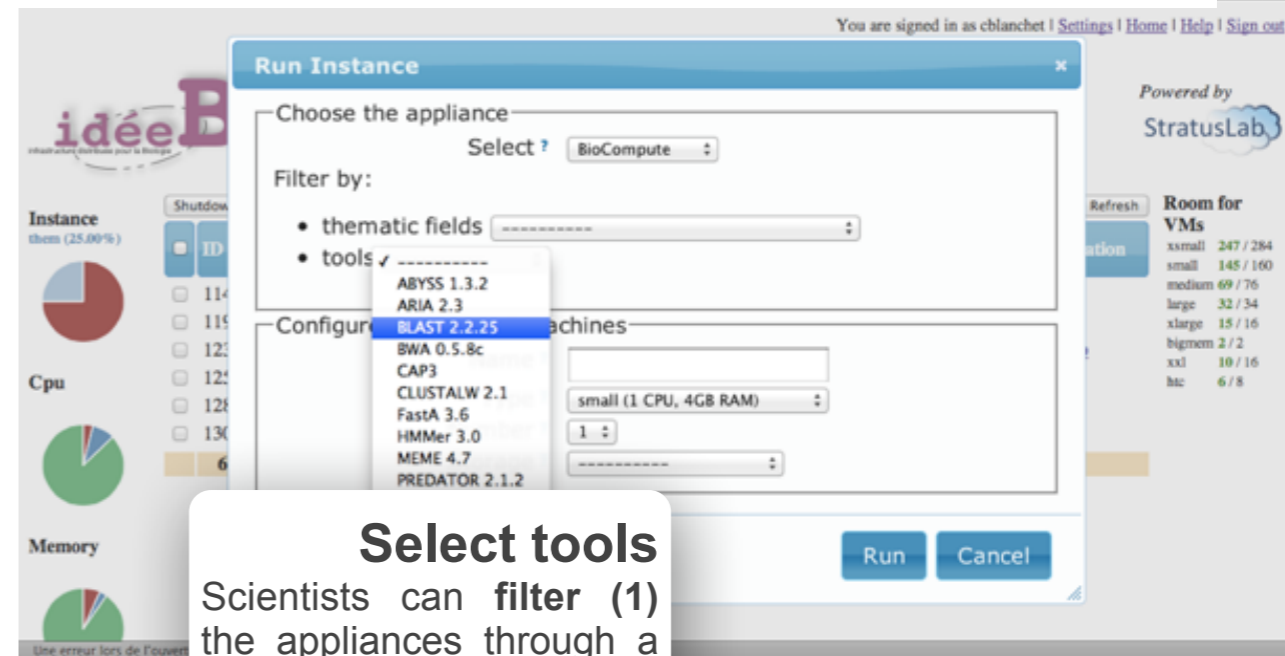
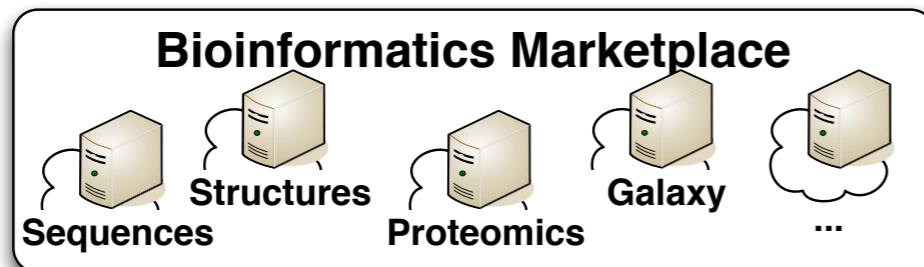
Showing 1 to 10 of 10 entries

ID	Name	State	Appliance	CPU%	CPU	Mem
2729	omssa mas115	Running	Protein Identification	0%	8	27
2737	struct det	Running	ARIA2.3	1%	2	8
2738	struct det	Running	ARIA2.3	1%	2	8
2739	struct det	Running	ARIA2.3	1%	2	8
2740	struct det	Running	ARIA2.3	1%	2	8
2741	struct det	Running	ARIA2.3	2%	2	8
2742	struct det	Running	ARIA2.3	0%	2	8
2743	struct det	Running	ARIA2.3	1%	2	8
2744	prot seq	Running	BIO compute node	0%	4	16
2746	CCMP	Running	Protein Identification	47%	24	48

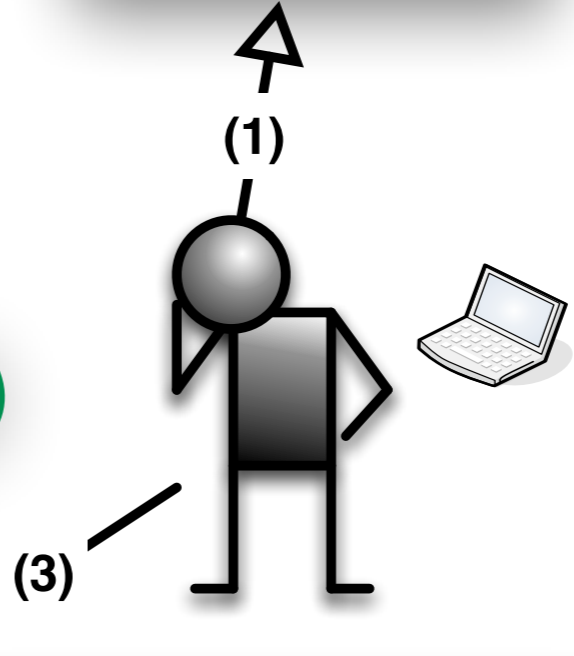
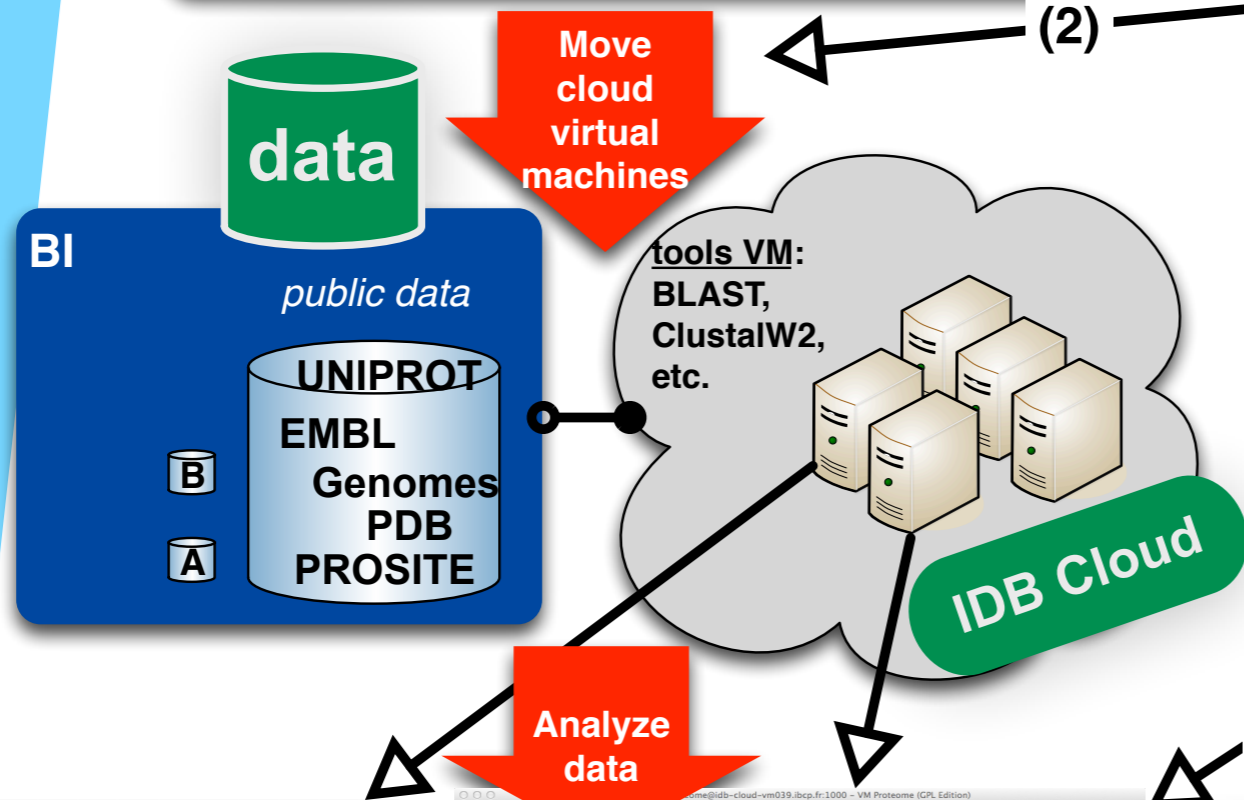
Room for VMs

- xsmall 247 / 284
- small 145 / 160
- medium 69 / 76
- large 32 / 34
- xlarge 15 / 16
- bigmem 2 / 2
- xxl 10 / 16
- htc 6 / 8

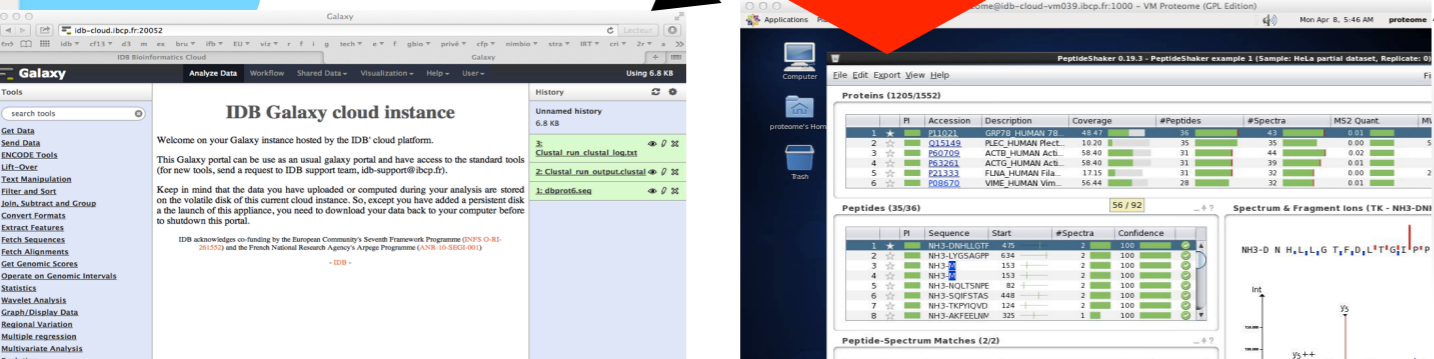
# Thematic bioinformatics appliances



**Select tools**  
 Scientists can filter (1) the appliances through a Web interface to identify and launch (2) the appropriate ones.



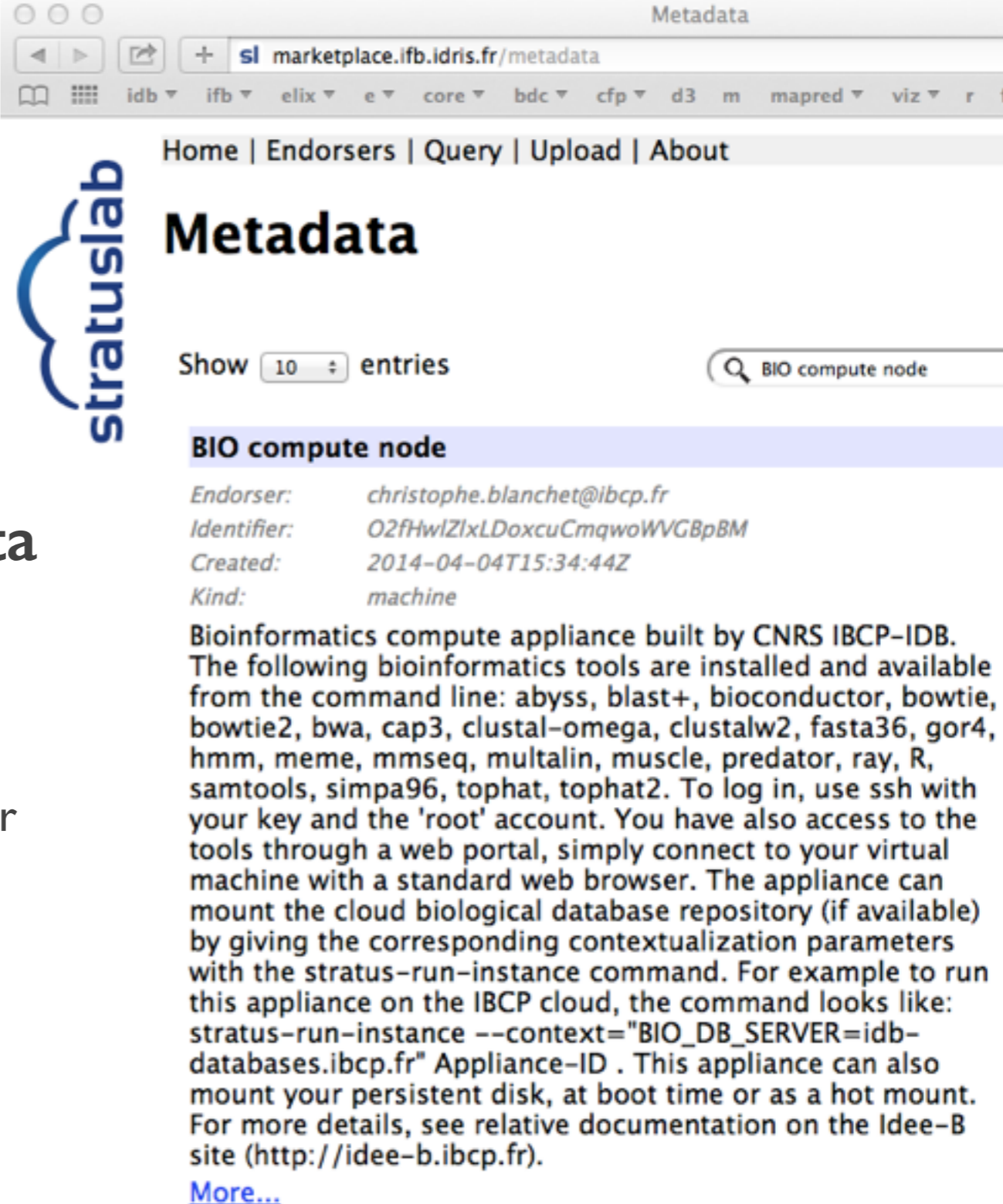
**Use tools (3)**  
 Scientists have access to their own cloud resources through web portal, remote virtual desktop or SSH.





# Standard Bioinformatics node

- 'Biocompute' appliance
- 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.
  - Automatically shared to the VMs
- Cluster mode
  - turn several instances in a single virtual cluster
  - shared file system
  - batch scheduling



The screenshot shows a web browser window with the URL `marketplace.ifb.idris.fr/metadata`. The page title is "Metadata" and the breadcrumb navigation includes "Home | Endorsers | Query | Upload | About". The main heading is "Metadata". Below the heading, there is a "Show 10 entries" dropdown and a search bar containing "BIO compute node". The search results show a single entry titled "BIO compute node" with the following details:

<i>Endorser:</i>	<i>christophe.blanchet@ibcp.fr</i>
<i>Identifier:</i>	<i>O2fHwIZlxLDoxcuCmqwoWVGBpBM</i>
<i>Created:</i>	<i>2014-04-04T15:34:44Z</i>
<i>Kind:</i>	<i>machine</i>

Below the table, there is a detailed description of the appliance:

Bioinformatics compute appliance built by CNRS IBCP-IDB. The following bioinformatics tools are installed and available from the command line: `abyss`, `blast+`, `bioconductor`, `bowtie`, `bowtie2`, `bwa`, `cap3`, `clustal-omega`, `clustalw2`, `fasta36`, `gor4`, `hmm`, `meme`, `mmseq`, `multalin`, `muscle`, `predator`, `ray`, `R`, `samtools`, `simpa96`, `tophat`, `tophat2`. To log in, use `ssh` with your key and the 'root' account. You have also access to the tools through a web portal, simply connect to your virtual machine with a standard web browser. The appliance can mount the cloud biological database repository (if available) by giving the corresponding contextualization parameters with the `stratus-run-instance` command. For example to run this appliance on the IBCP cloud, the command looks like: `stratus-run-instance --context="BIO_DB_SERVER=idb-databases.ibcp.fr" Appliance-ID`. This appliance can also mount your persistent disk, at boot time or as a hot mount. For more details, see relative documentation on the Idee-B site (<http://idee-b.ibcp.fr>).

[More...](#)

At the bottom of the screenshot, there are logos for "idée B", "IBCP", "CNRS", and "Lyon 1".



# 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

Metadata

Home | Endorsers | Query | Upload | About

## Metadata

Show 10 entries

Galaxy portal

Endorser: christophe.blanchet@ibcp.fr  
Identifier: GOqP1arAKmWzR2PB-tCEDsHbu7n  
Created: 2013-11-21T15:14:39Z  
Kind: machine

Bioinformatics gateway appliance configured with the GALAXY portal, built by CNRS IBCP-IDB. You will have access to the pre-installed bioinformatics tools through the web portal. Connect to your own Galaxy portal with a standard web browser, simply follow the link on the main IDB cloud interface. For more details, see relative documentation on the Idee-B site (<http://idee-b.ibcp.fr>).

[More...](#)

ID	Tool	Status	Provider	Progress	CPU	Mem	SSH	HTTP
1875	blast	Running	BioCompute	2%	2	8	ssh	http
1876	portal	Running	Galaxy	2%	2	8	ssh	http
1877	blast machine	Running	BioCompute	2%	4	16	ssh	http

## Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Using 4.9 GB

### IDB Galaxy cloud instance

Welcome to your Galaxy instance hosted by the IDB's cloud platform.

#### Usage

This appliance is configured with the well-known GALAXY portal. You connect to it with a standard web browser : simply follow the link on the main IDB 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 IDB support team: [idb.support@ibcp.fr](mailto:idb.support@ibcp.fr)).

Tools

search tools

Get Data  
Send Data  
ENCODE Tools  
Lift-Over  
Text Manipulation  
Filter and Sort  
Join, Subtract and Group  
Convert Formats

Unnamed history  
4.9 GB

- 23: Clustal run clustal log.txt
- 22: Clustal run output.clustal
- 21: dbprot6.seq

# 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!

idéeB

IBCP

CNRS

Ups Lyon 1



# Hadoop for Life Science

- Provide turnkey virtual machine with pre-configured mapreduce framework
  - Accelerate bigdata analysis with the two steps map & reduce paradigm
  - Hadoop MapReduce 1.0.4
- Appliances (2)
  - standard hadoop mapreduce
  - bioinformatics software integrated in hadoop
- 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 bioinformatics 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 filesystem. You can extend a current cluster by submitting a list of new nodes to the script. A FastA MapReduce example is also provided under the directory `/usr/local/share/fasta`. (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, [mapreduce.inria.fr](http://mapreduce.inria.fr))

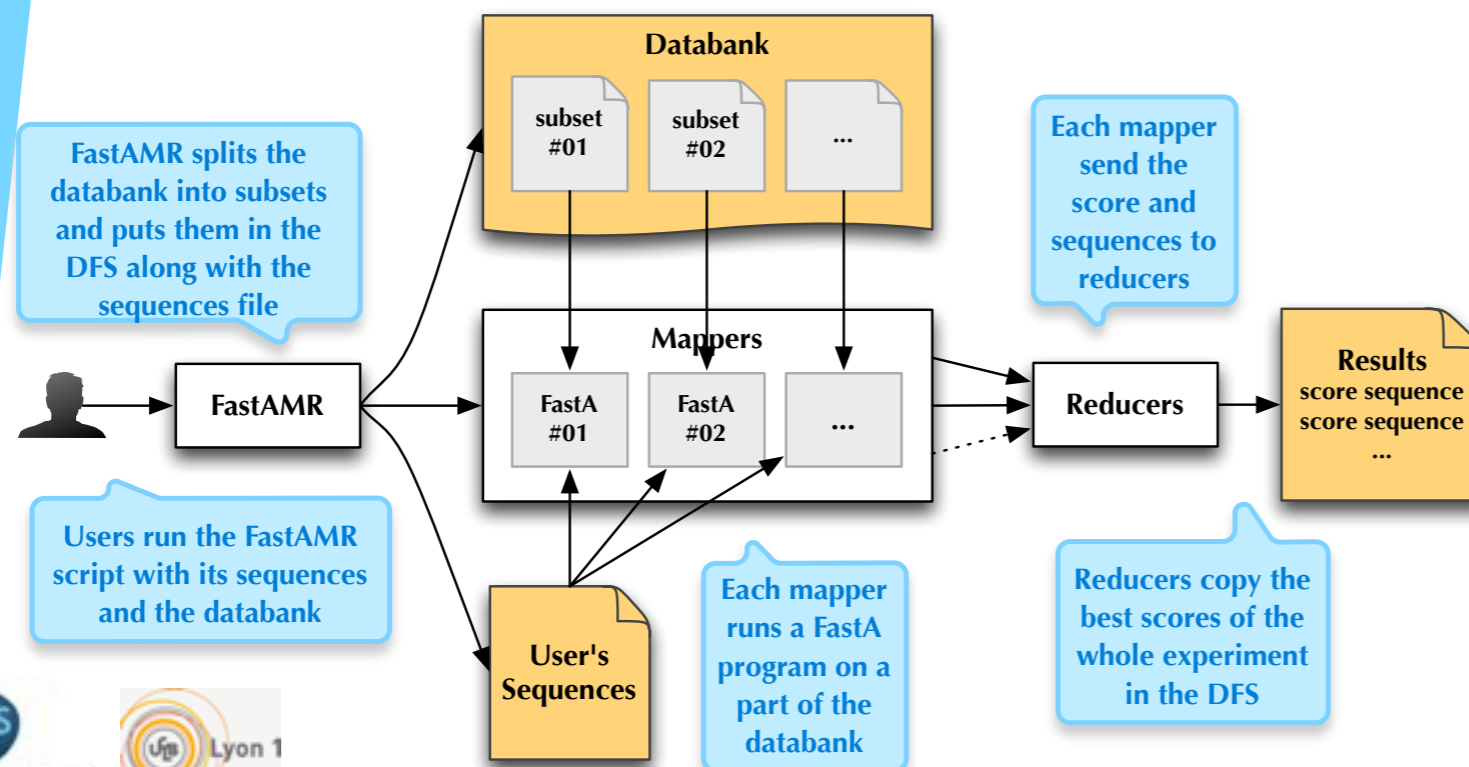
[More...](#)

## Hadoop MapReduce

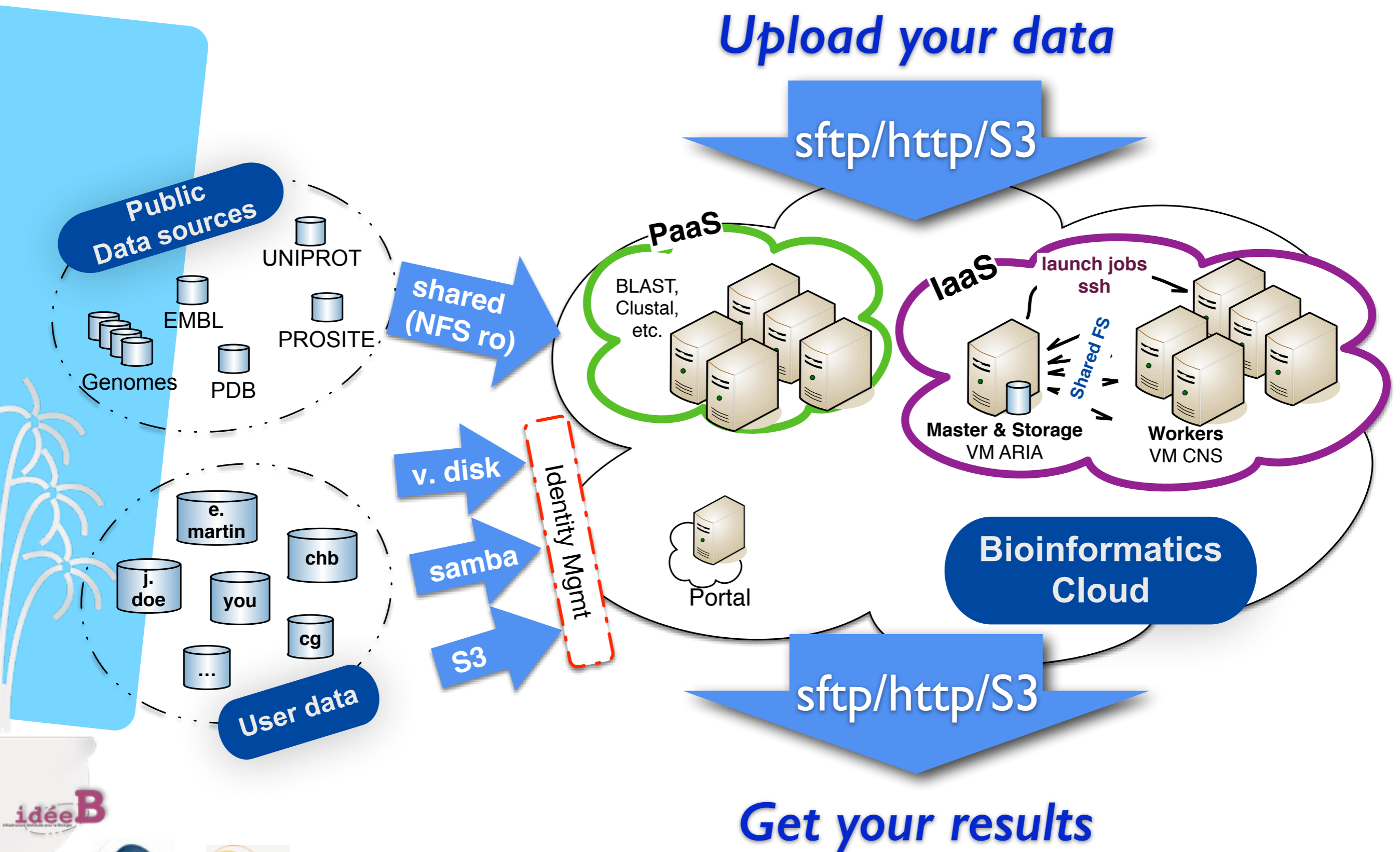
Endorser: clement.gauthey@ibcp.fr  
Identifier: BttU7uNMSUT1haUigVS7xySI2rr  
Created: 2013-05-17T09:33:52Z  
Kind: machine

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 account 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 `hadoop-extend-cluster`. (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, [mapreduce.inria.fr](http://mapreduce.inria.fr))

[More...](#)



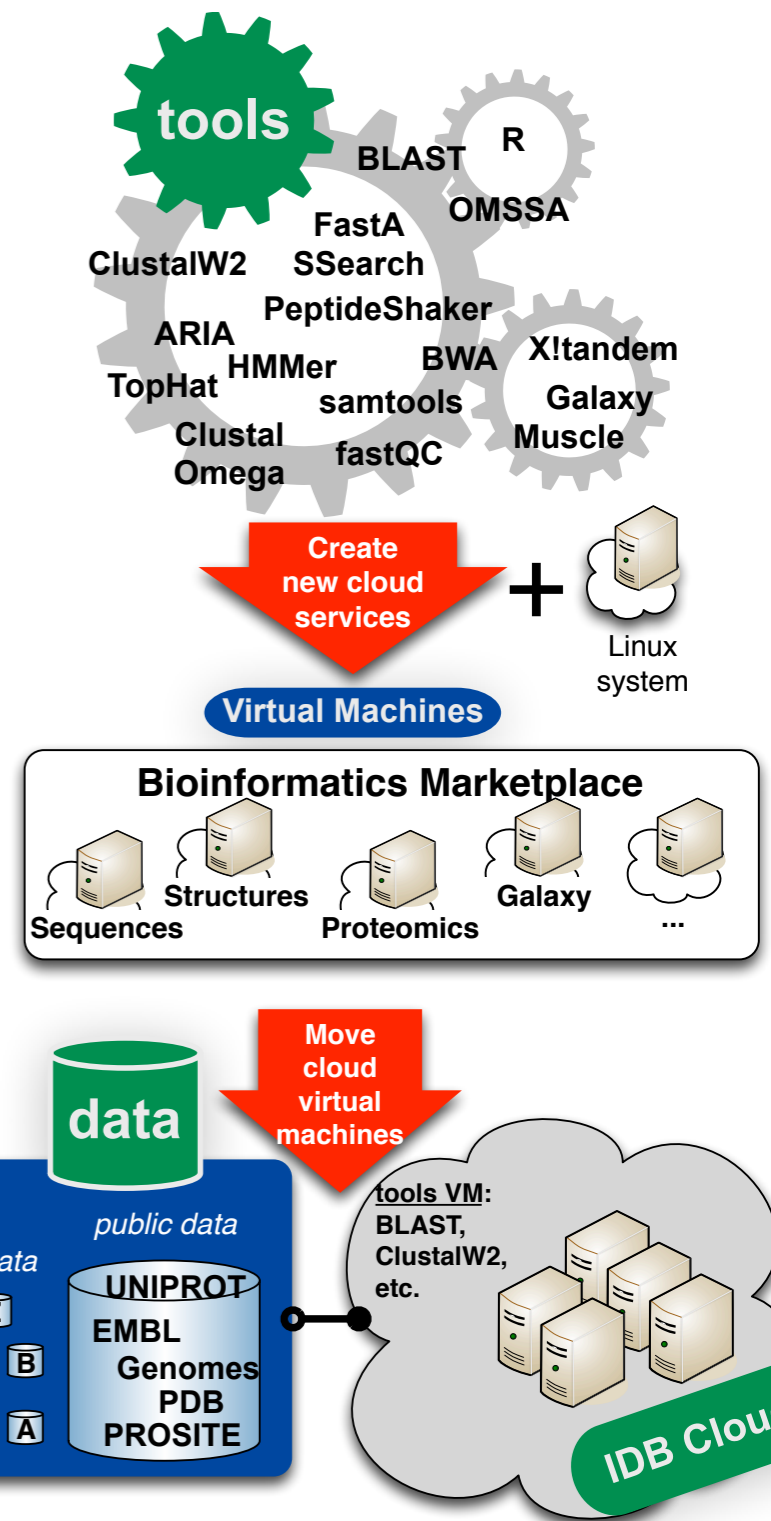
# Cloud Storage for Biological Data





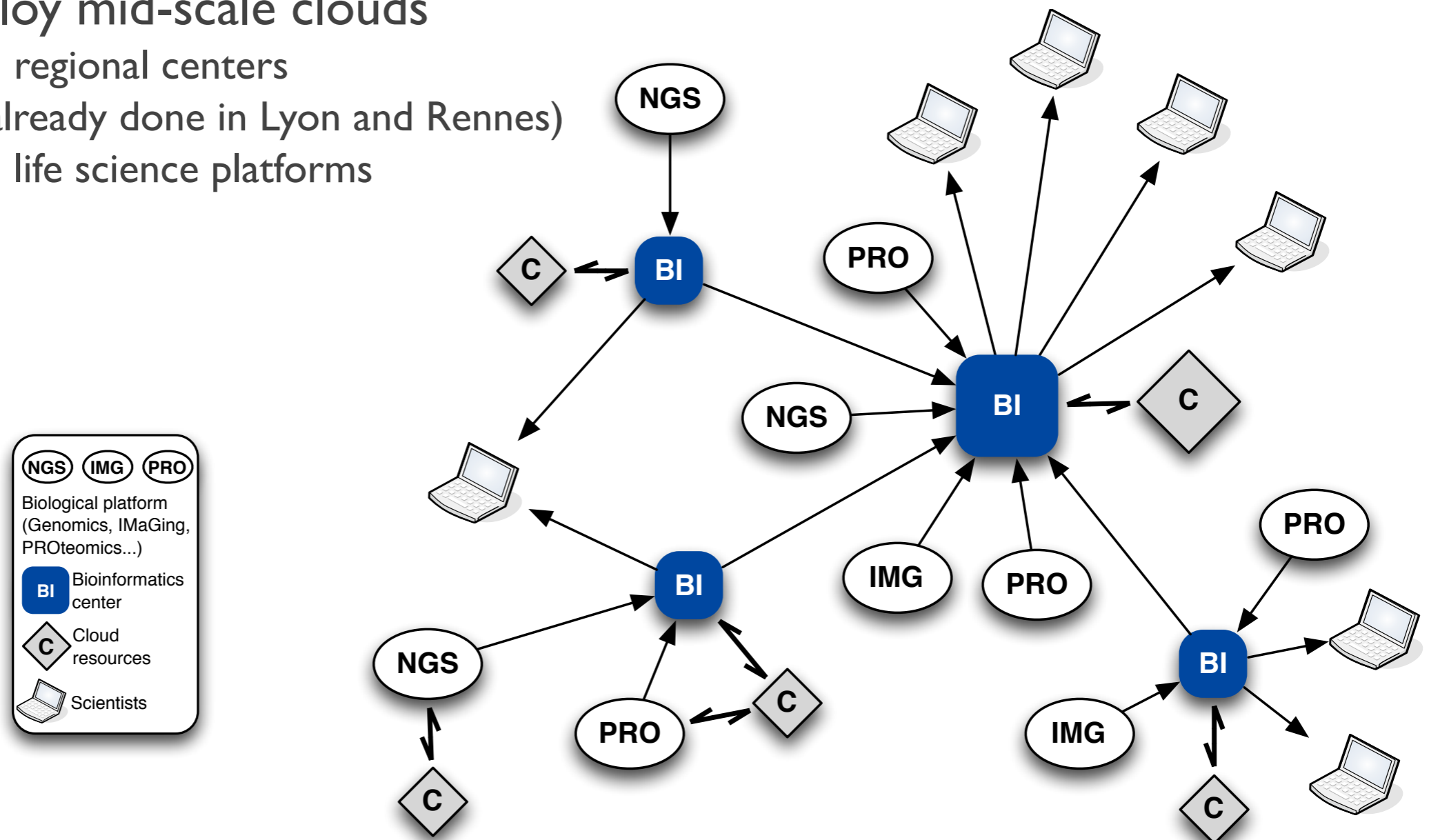
# Conclusion

- Added value of cloud,
  - for scientific analyses: user-specific resources, isolated, different instances together
  - for training: create a special instance derived from the main but with dedicated datasets
  - for tools integration: semantic annotation, solve software dependencies
  - for development & operations (DevOps): different versions at the same time
- Provide turnkey bioinformatics appliances
  - Standard tools and pipelines
  - New developments
  - Ready to run on clouds
- Public bioinformatics cloud (e.g. IDB)
  - Tightly connected to existing bioinformatics resources
  - Linked to public biological databases
  - In collaboration with the French Institute of Bioinformatics



# Perspectives

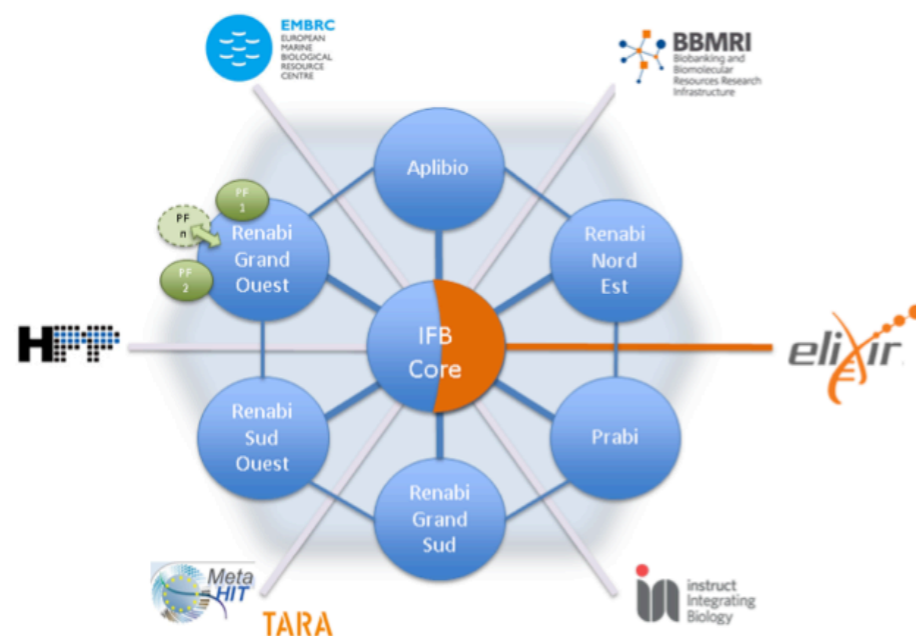
- French Institute of Bioinformatics - IFB
- Set up a national appliances catalogue for bioinformatics
- Deploy a national bioinformatics cloud
- Deploy mid-scale clouds
  - in regional centers  
(already done in Lyon and Rennes)
  - in life science platforms



# 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 programs
  - supporting projects
  - training users
- To provide an IT infrastructure devoted to management and analysis of biological data
  - material resources : CPUs, disks, etc.
  - availability of biology data collections
  - deployment of bioinformatics tools
- To act as a middleman between the life science community and the bioinformatics/ computer science research community



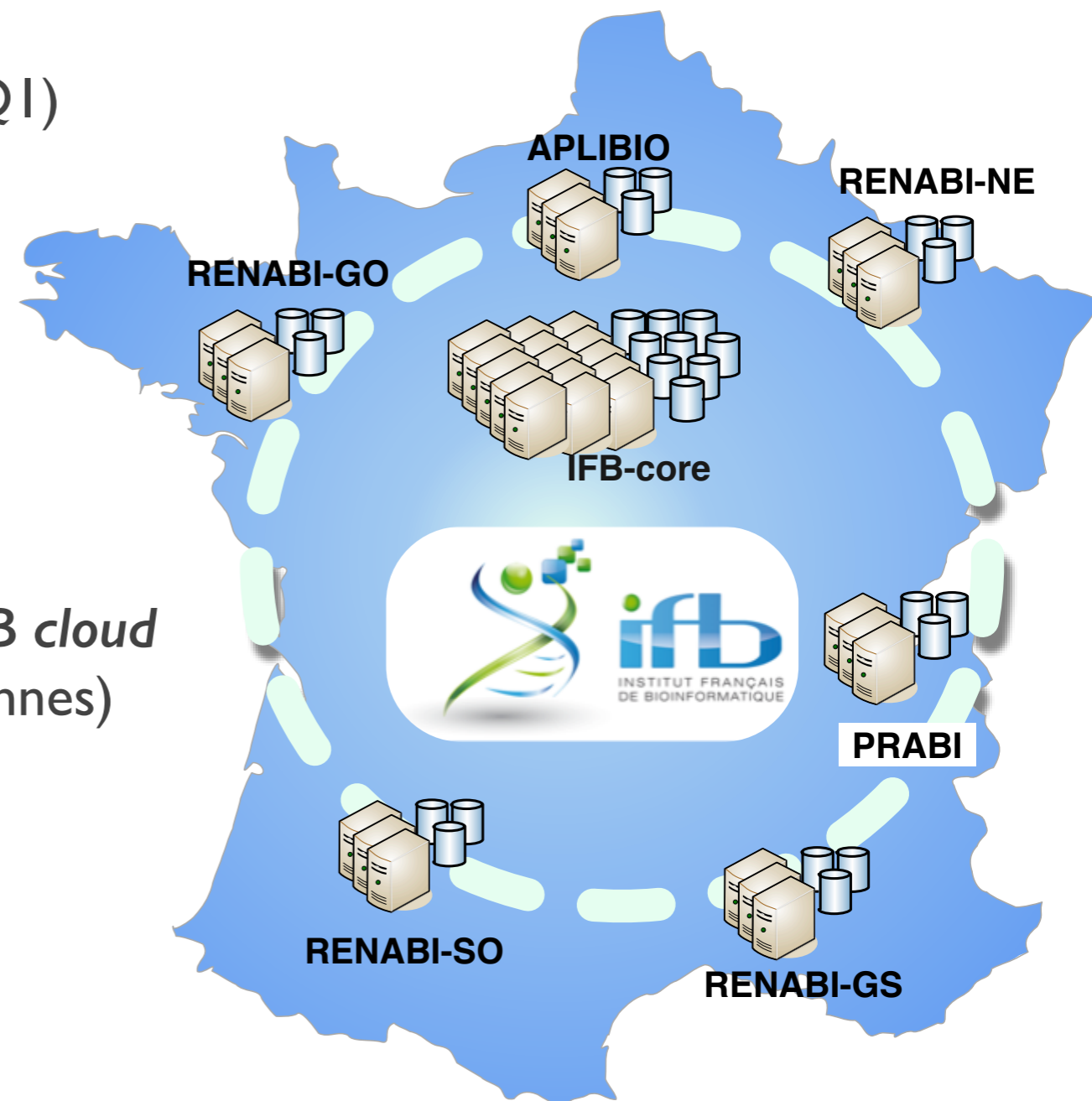
## • ELIXIR French Node

- 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



# Infrastructure

- **IFB-Core resources**
  - Academic cloud for life science
  - Will be hosted at **CNRS IDRIS** supercomputing center (PARIS)
    - A pilot infrastructure (2014-Q1)
    - Production infrastructure +5,000cores IPB (2014-S2)
- **+ Regional resources**
  - 6 regional bioinformatics centers
  - +6,000 cores ~IPB
  - but 20 bioinformatics platforms
  - 2 existing clouds: PRABI-IBCP *IDB cloud* (Lyon) & Genouest *genocloud* (Rennes)
- Deploy a federation of clouds





# Questions ?

## Acknowledgments

- StratusLab members
- co-funding by  
European Community's Seventh Framework Programme (INFISO-RI-261552)  
French National Research Agency's Arpege Programme (ANR-10-SEGI-001).  
French Institute of Bioinformatics (IFB)



<http://idee-b.ibcp.fr>