The French bioinformatic cloud to analyse life science data

Christophe Blanchet

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

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

IFB - Institut Français de Bioinformatique

French distributed infrastructure for life-science information

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

http://www.france-bioinformatique.fr

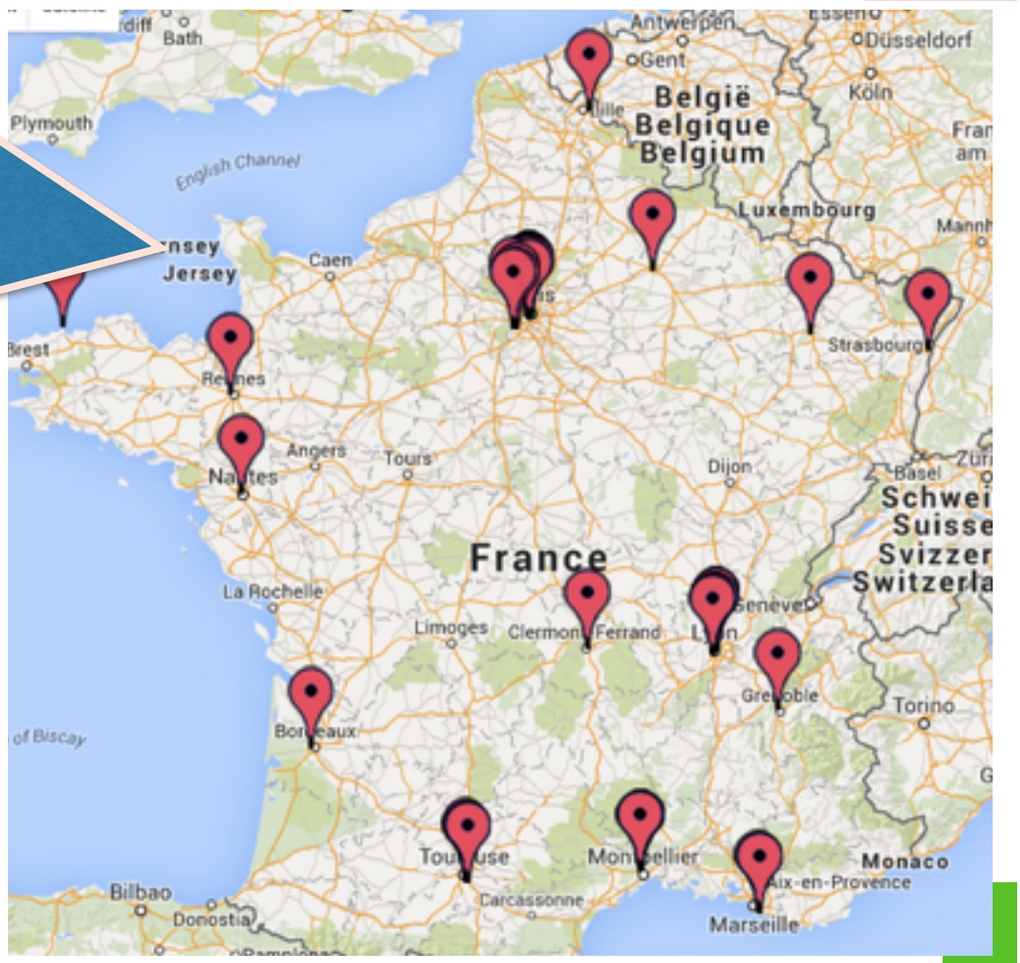
CNRS UMS3601. Avenue de la Terrasse, Bât 21. 91190 Gif-sur-Yvette

- A national hub : **IFB-core**
- A network of **36 platforms**

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

- To provide support for national biology programs
- 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

ELIXIR-FR Node



Relation with External Projects

IFB is the ELIXIR French Node

- To promote consistency and complementarities between the components offered by the ELIXIR French node and those of other European nodes

EXCELERATE (EU-H2020 676559)

- Services registry (WP-1 Tools Interoperability and Service Registry)
- Compute Platform (WP-4 Technical Services)
- Collaboration in connection with the e-Infrastructure
  - ⋆ WP5: The ELIXIR Interoperability Backbone
  - ⋆ WP6-9 : Use Cases
  - ⋆ WP10: ELIXIR Node Capacity Building and Communities of Practice

CYCLONE (EU-H2020 644925)

- Scientific use cases: Securing human biomedical data, Cloud virtual pipeline for microbial genomes analysis, Live remote cloud processing of sequencing data…
- Multi-cloud deployment, AAI, Network-as-a-Service…

EGI-Engage (EU-H2020 654142)

BioDataCloud (French PIA INBS 2012)

- Industrial use cases about plant genomics assembling and visualization.

IFB's e-Infrastructure

Mission : to provide core bioinformatics resources to the life science research community.

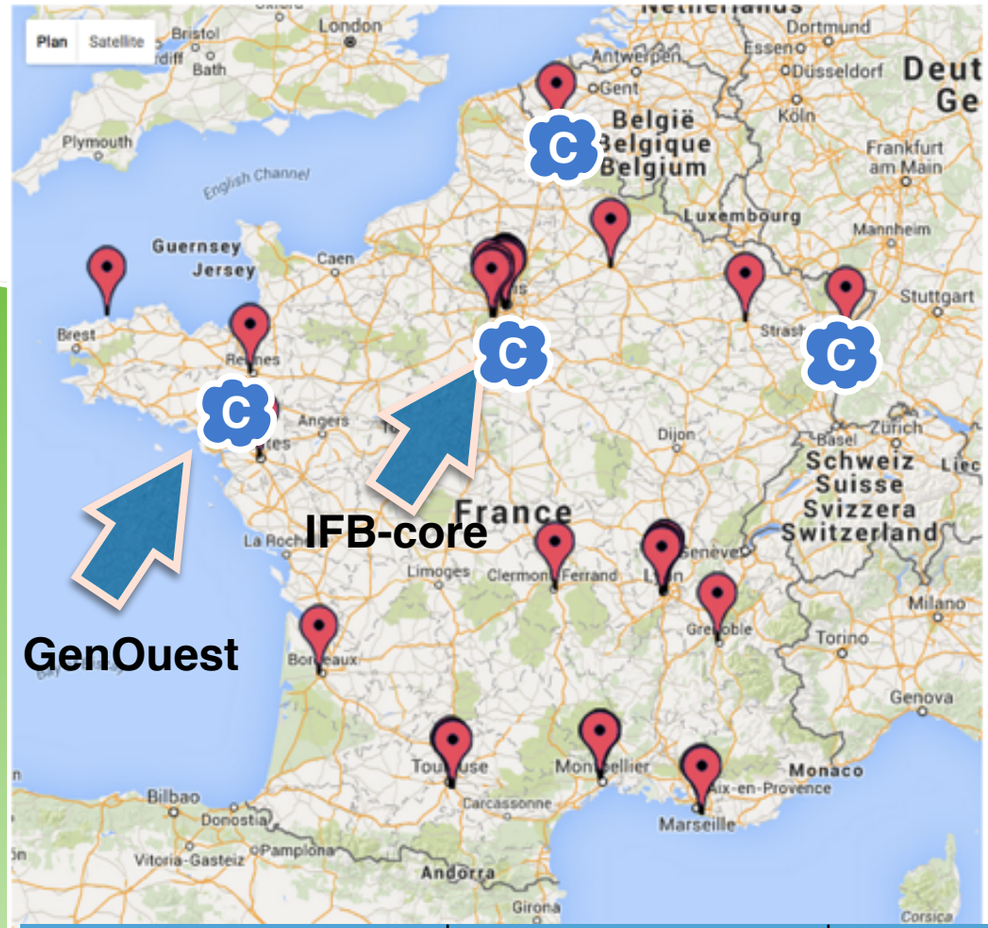
- A **national** IT resource
  + 35 platforms
- Hardware, data collections and bioinformatics tools

Current resources

- **15,000 cores - 6 PB**
- **4 running clouds**

➡ Create a federation of clouds for life sciences

Towards a multi-cloud infrastructure



Requirements

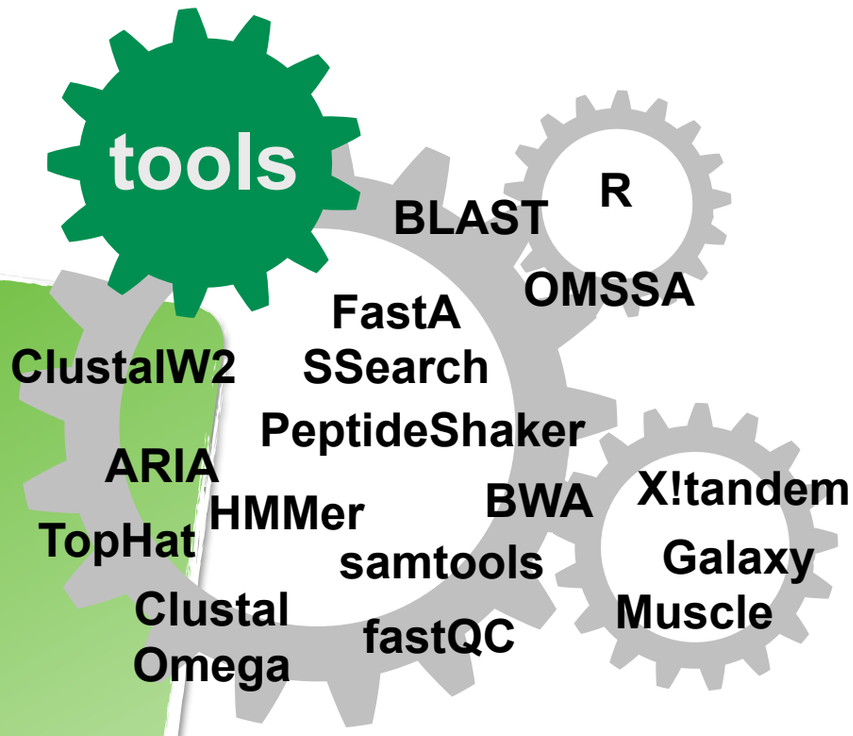
- common identities and authorizations management based on community standards (EduGAIN)
- interoperability of virtual images (VM/container)
- tools for multi-cloud deployment (e.g. the SlipStream/ NuvLa broker)
- network management, security and propagation over several datacenters

Solutions expected from

- ELIXIR/EXCELERATE and CYCLONE projects

SITE	Compute #cores	Storage #TB	RAM #GB	Largest VM	Technology	Location
IFB-core *2014-*	**200** *(+160)*	**50** *(+96)*	**2,000** *(+1)*	**20c 256GB**	**StratusLab**	**CNRS-IDRIS, Paris**
2016	*5,000*	*1,000*	*27,500*	*64c 2TB*	*OpenStack*	*CNRS-IDRIS, Paris*
2017	*10,000*	*2,000+*			*OpenStack*	*CNRS-IDRIS, Paris*
GenOuest *2014-*	**220** *(+96)*	**8** *(+20)*	**685**	**8c 32GB**	**OpenNebula**	**IRISA, Rennes**

Many tools



ABYSS 1.3.4
ARIA 2.3
Bioconductor 2.11
biomaj
BLAST+ 2.2.27
Blat 35
Bowtie 0.12.8
Bowtie2 2.0.0-beta7
BWA 0.6.2
BWA 0.7.10
CAP3
CD-HIT 4.6.1
Clustal Omega 1.0.3
CLUSTALW 2.1
Cufflinks 2.0.2
Cutadapt 1.2.1
E-SURGE 1.9.0
Exonerate 2.2.0
eXpress 1.5.1
FastA 3.6
FastQC 0.10.1
Galaxy portal
GATK 2.3.4
HMMer 3.0
ImageJ 1.48
khmer 1.1
M-SURGE 1.8.5
MEME 4.7
MMSEQ 0.11.2a
Mobyle
MODAL
MultAlin 5.4.1
MUSCLE 3.8.31
neo4j
Oases 0.2.08
OMSSA 2.1.9
PeptideShaker 0.18.3
phyml 3.1
PREDATOR 2.1.2
proline
python 2.7
R 2.13
R 3.1.1
R 3.1.2
R-studio
Ray 1.3
RSAT
samtools 0.1.18
Samtools 1.1
SearchGUI 1.10.4
SeqClean
Shiny
Stacks
STAR 2.4.0f1
SuMo v1
TGICL
TopHat 2.0.6
trim_galore 0.3.7
Trinity 2.0.4
U-CARE 2.3.2
VCFtools 0.1.11
Velvet 1.2.10
X!tandem 12-10-01-1
XPLOR-NIH 2.30
…

Many data

Collections of reference data (TB)

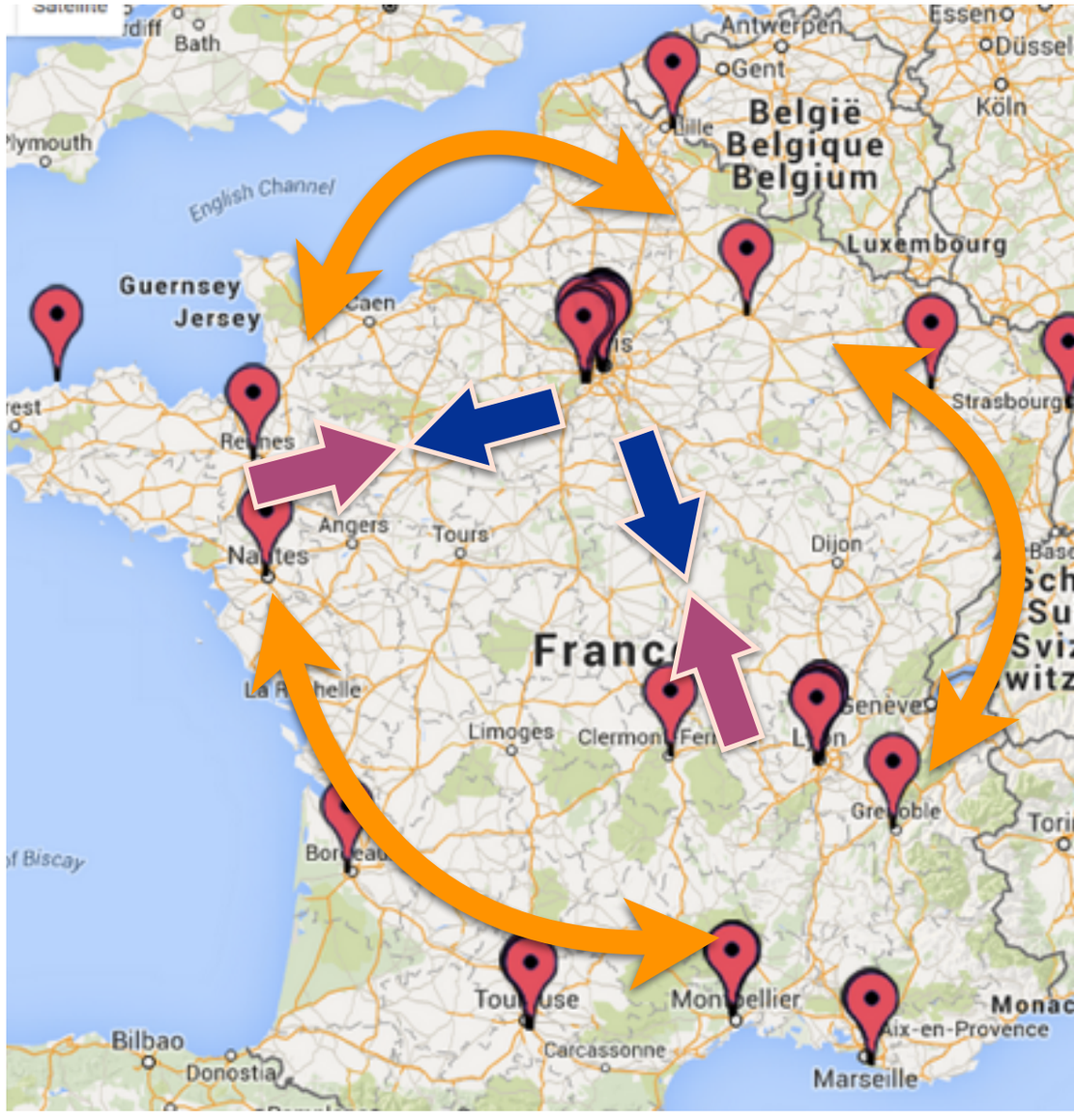
- IFB-Core mirrors and computes indexes
- PFs can subscribe to databases update
- **Transfer from IFB-core to regional PFs**
  - ⋆ **BioMAJ**, iRODS, gridftp

Experimental data (TB)

- **Regional desks** for **initial deposit**
- Local archiving and analysis
- **Transfer from regional PFs to IFB-core**
  - ⋆ iRODS, gridftp

User data (MB-TB)

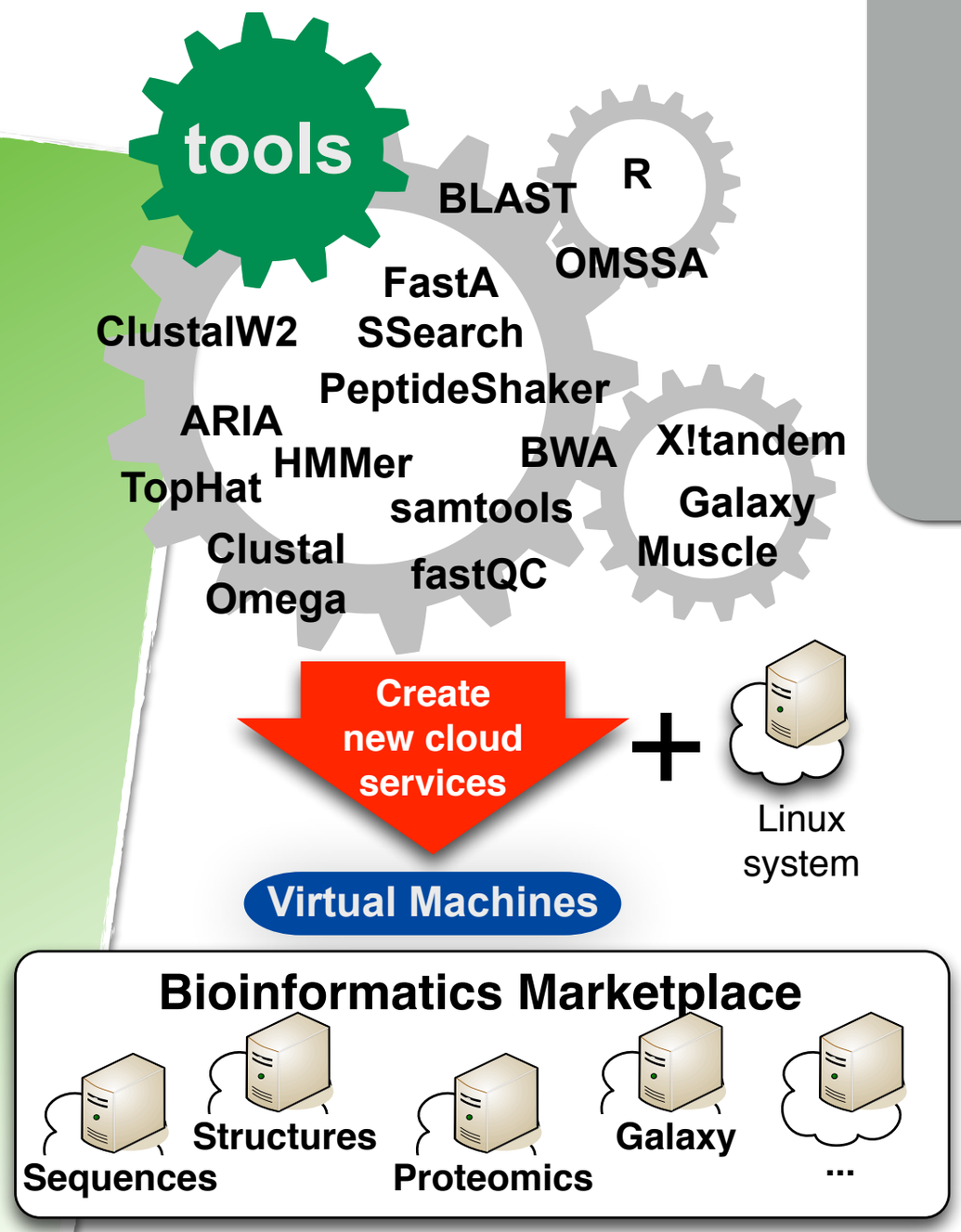
- Many (can be large),
- Heterogeneous
- Unpredictable distribution
  - ⋆ iRODS, object storage, noSQL



Two main issues : Transfer / Biomedical data

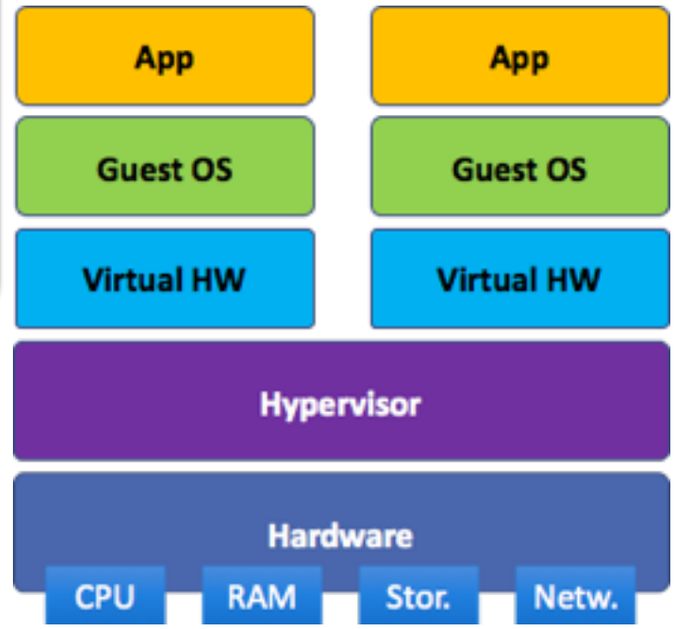
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Create bioinformatics “appliances“



What is an appliance ?

- predefined image of a virtual machine
- with tools, pipeline, recipes…
- ready to run
- small size image (GB)



Created by

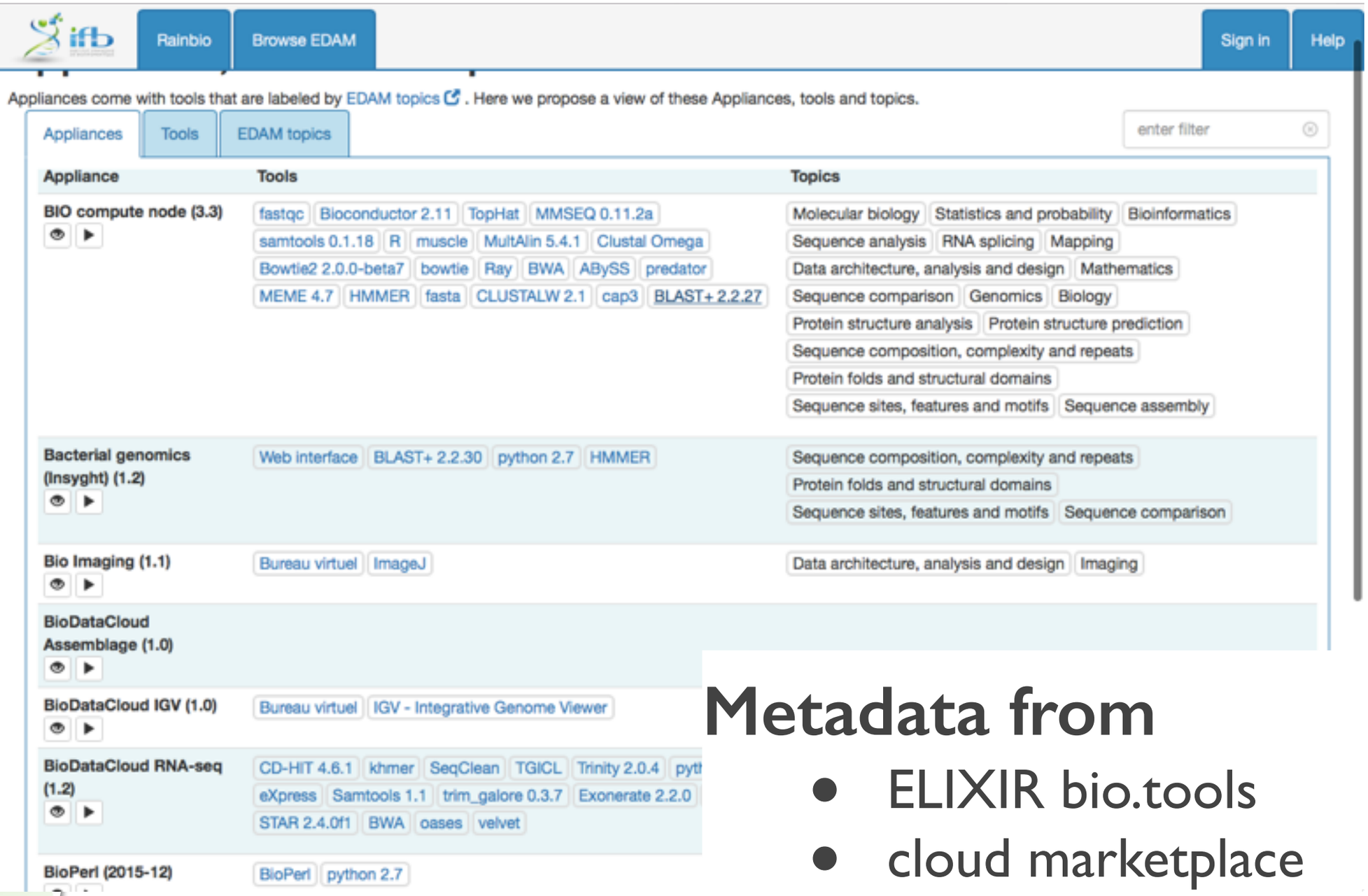
- developers
- experts of different life science domains

How to help the developers?

- training sessions for developers
- cloud support team of IFB-core
- individual hackhaton at IFB-core

RAINBio : registry of bioinformatics VM images

http://cloud.france-bioinformatique.fr/**rainbio**/



Scientific apps

- NGS
- BioImaging
- Proteomics
- …

Utilities

- base OS
- data mgmt
- batch sched.

Interfaces

- CLI
- Web portal
- Remote Desktop

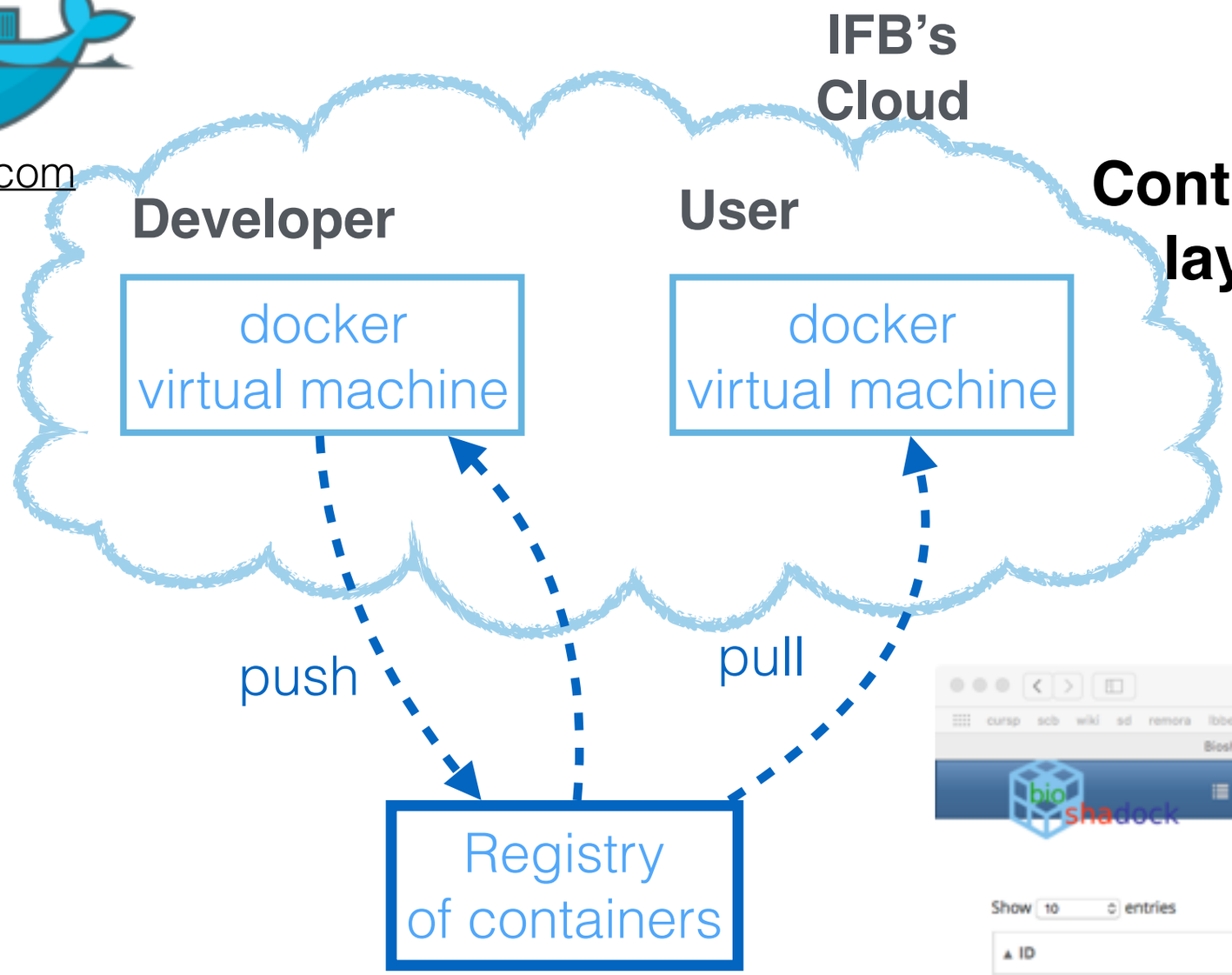
Metadata from

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

Coll. IFB-core/GenOuest

Simplify tools integration with docker

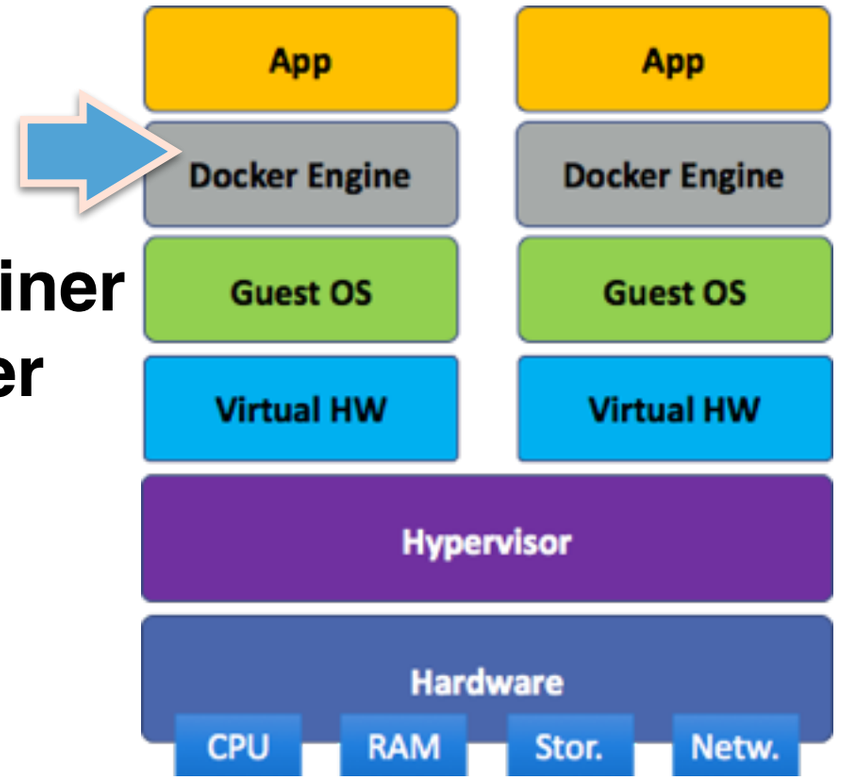
docker.com

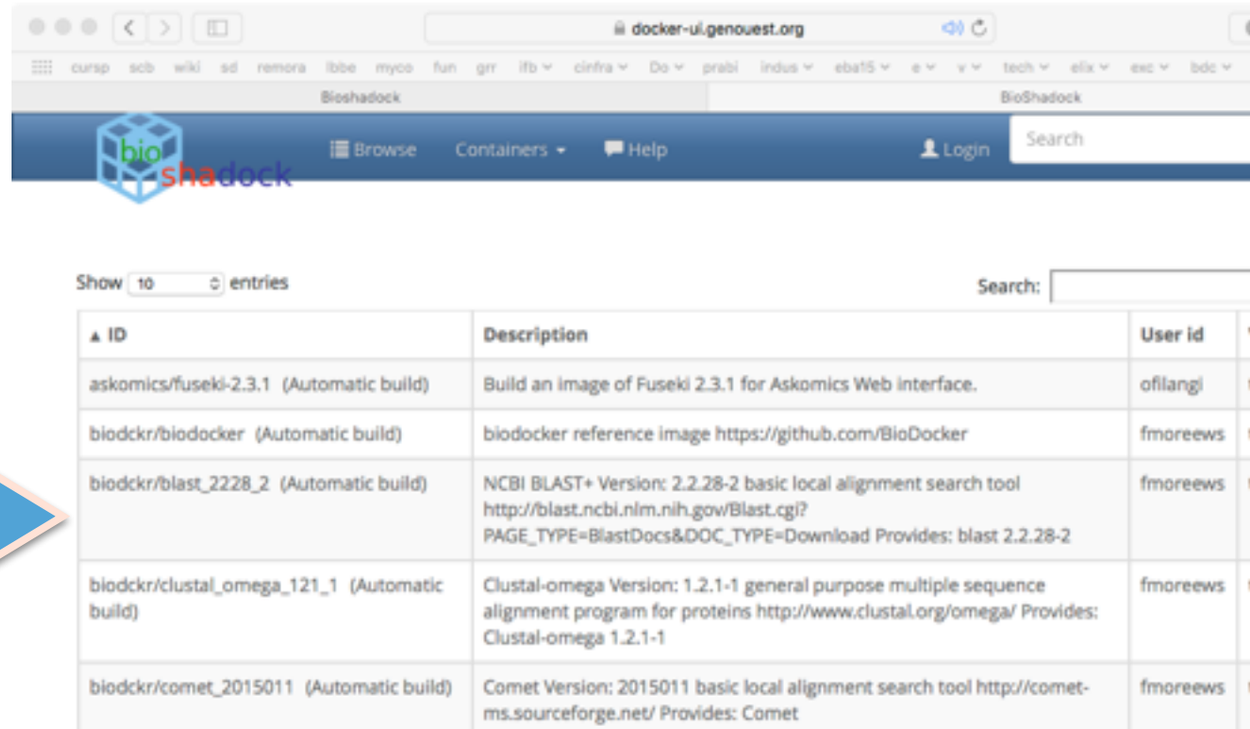


BioShaDock

IFB's docker hub

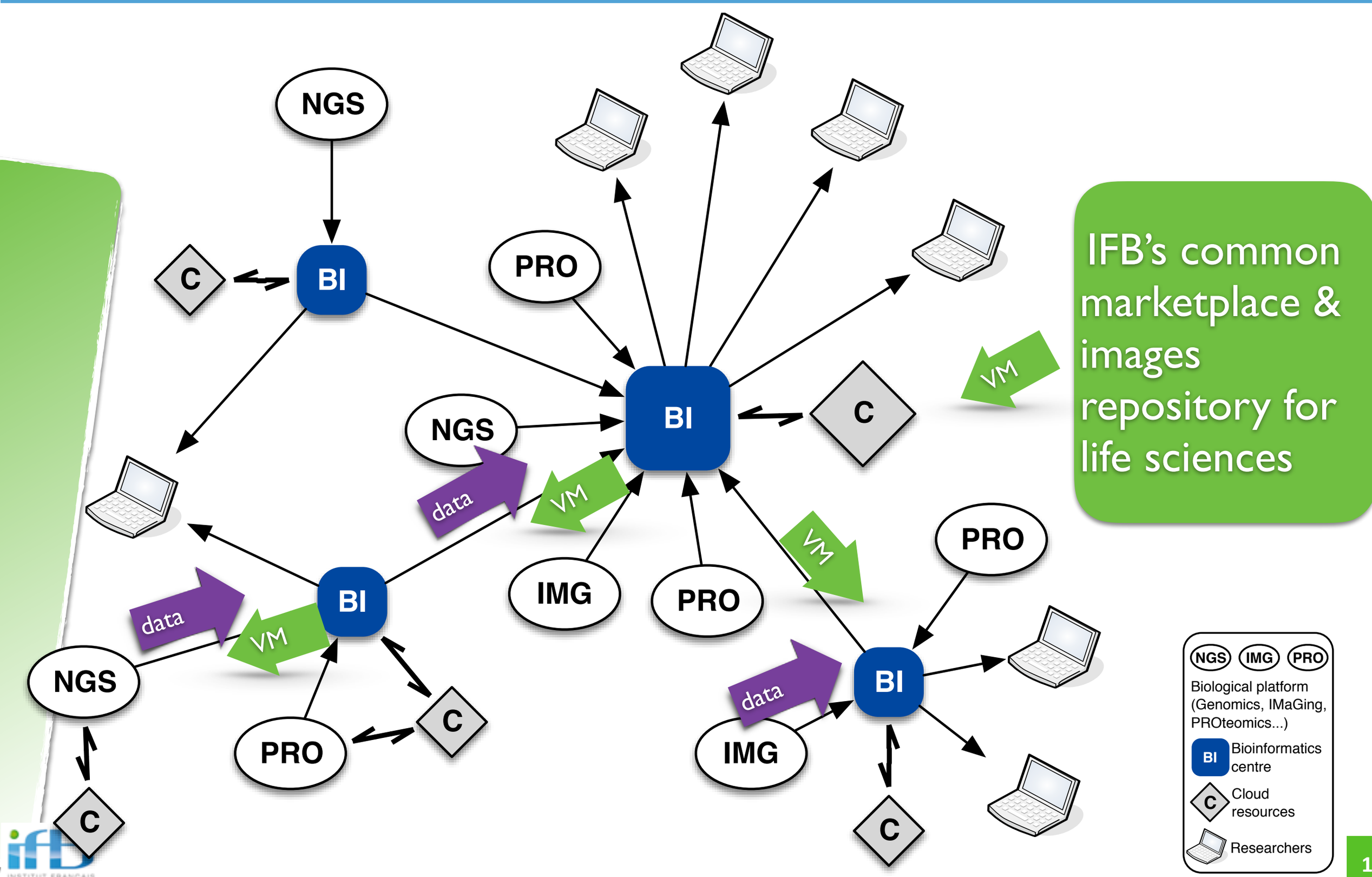
(hosted at GenOuest)





INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Move virtual images rather than data



IFB's common marketplace & images repository for life sciences

Use IFB's Bioinformatics Cloud

Public Data sources

BioMAJ

EMBL

UNIPROT

PROSITE

PDB

Genomes

Reference Datasets

common share

VMs VMs

Data

Cloud Credentials

Data

Personalized interfaces

e. martin

chb

j. doe

you

cg

...

User data

virtual disks

Author.

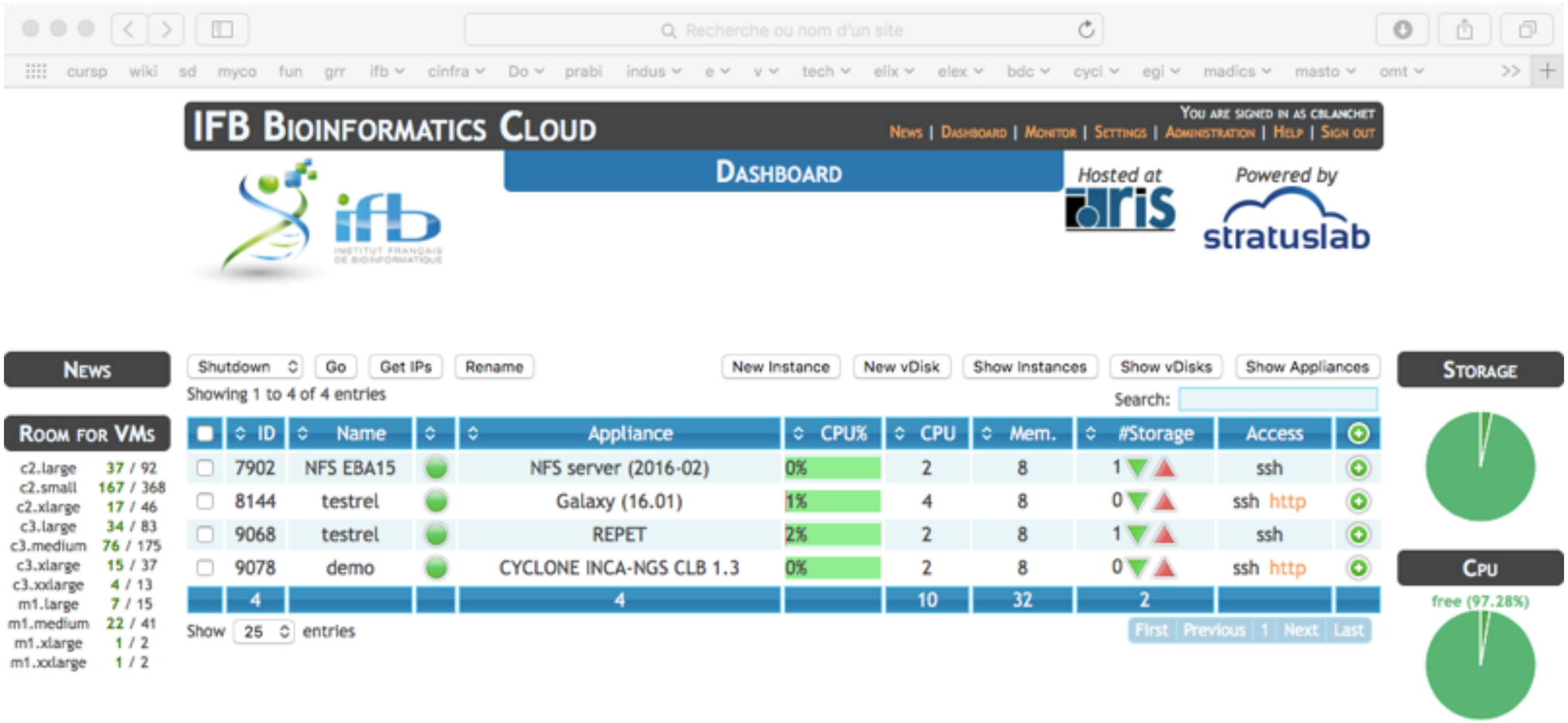
VMs

VMs

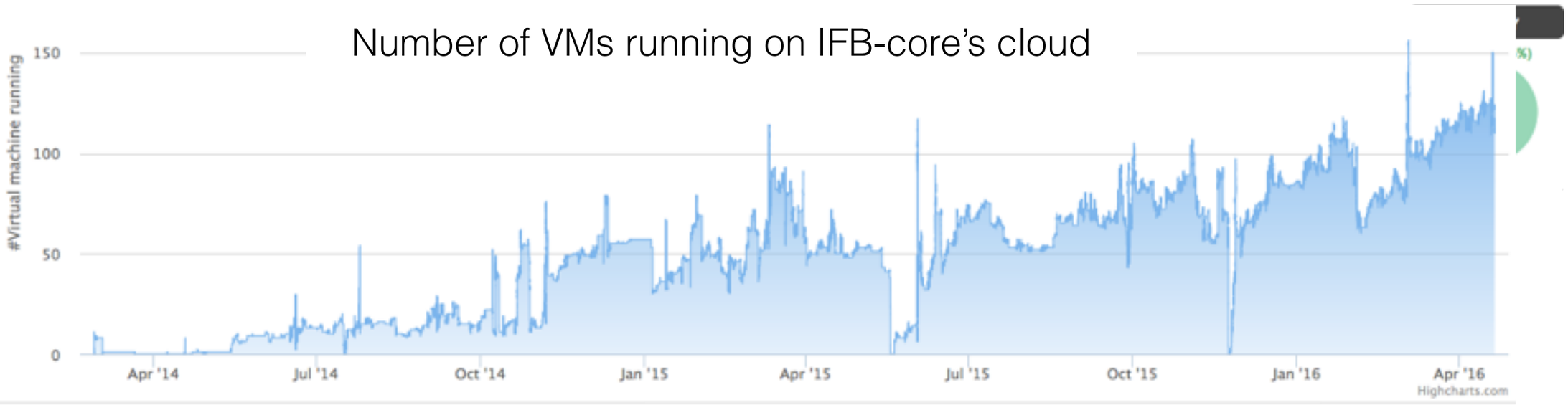
VMs

Cloud for Bioinformatics

IFB's Cloud Usage



Number of VMs running on IFB-core's cloud



App Biocompute

Standard bioinformatics node

With pre-installed standard bioinformatics tools

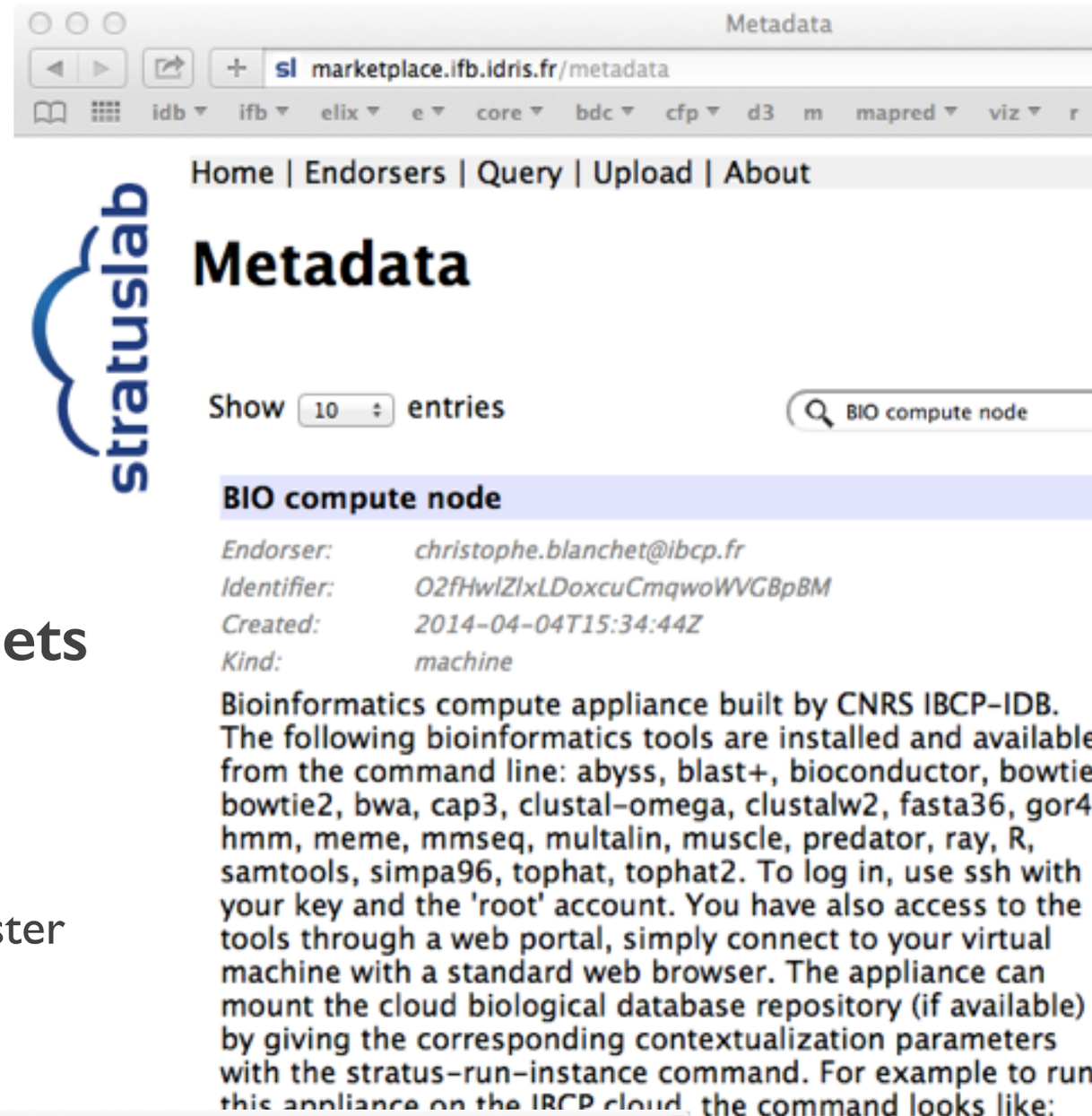
- BLAST, FastA, SSearch,HMM,...
- ClustalW2, Clustal-Omega, Muscle,..
- Bowtie(2), BWA, samtools, ...
- MEME, R, etc.

Connected to public reference datasets

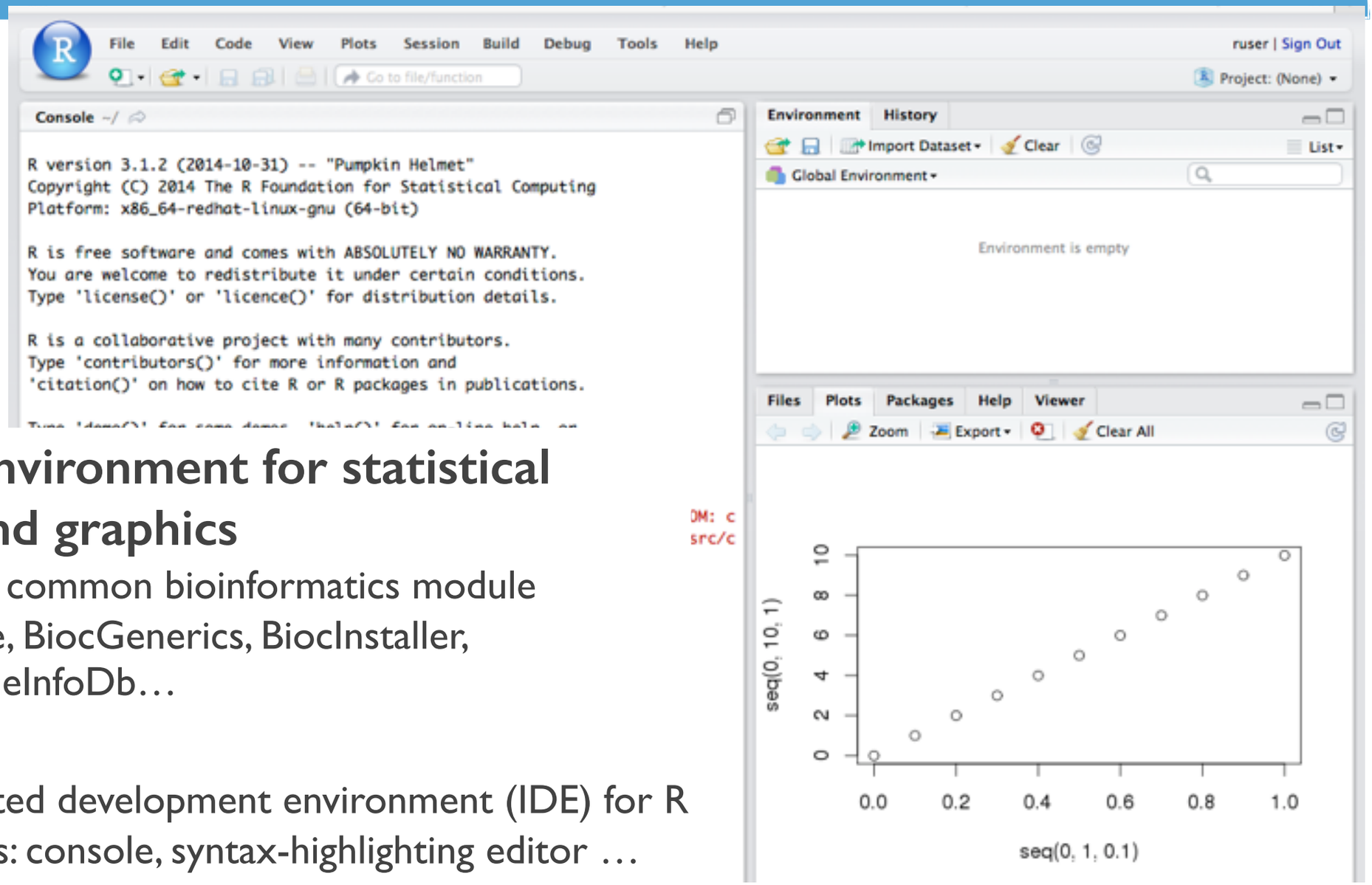
- Uniprot, EMBL, genomes, PDB, etc.
- Automatically shared with the VMs

Cluster mode

- turn several instances in a single virtual cluster
- shared file system
- batch scheduling



App R Statistical Computing



R software environment for statistical computing and graphics

- include common bioinformatics module
- Biobase, BiocGenerics, BiocInstaller, GenomeInfoDb…

RStudio IDE

- integrated development environment (IDE) for R
- features: console, syntax-highlighting editor …

Shiny web framework

- powerful web framework for building web applications using R.
- without requiring HTML, CSS, or JavaScript knowledge.

Contact: Stéphane Delmotte (IFB PRABI-LBBE)

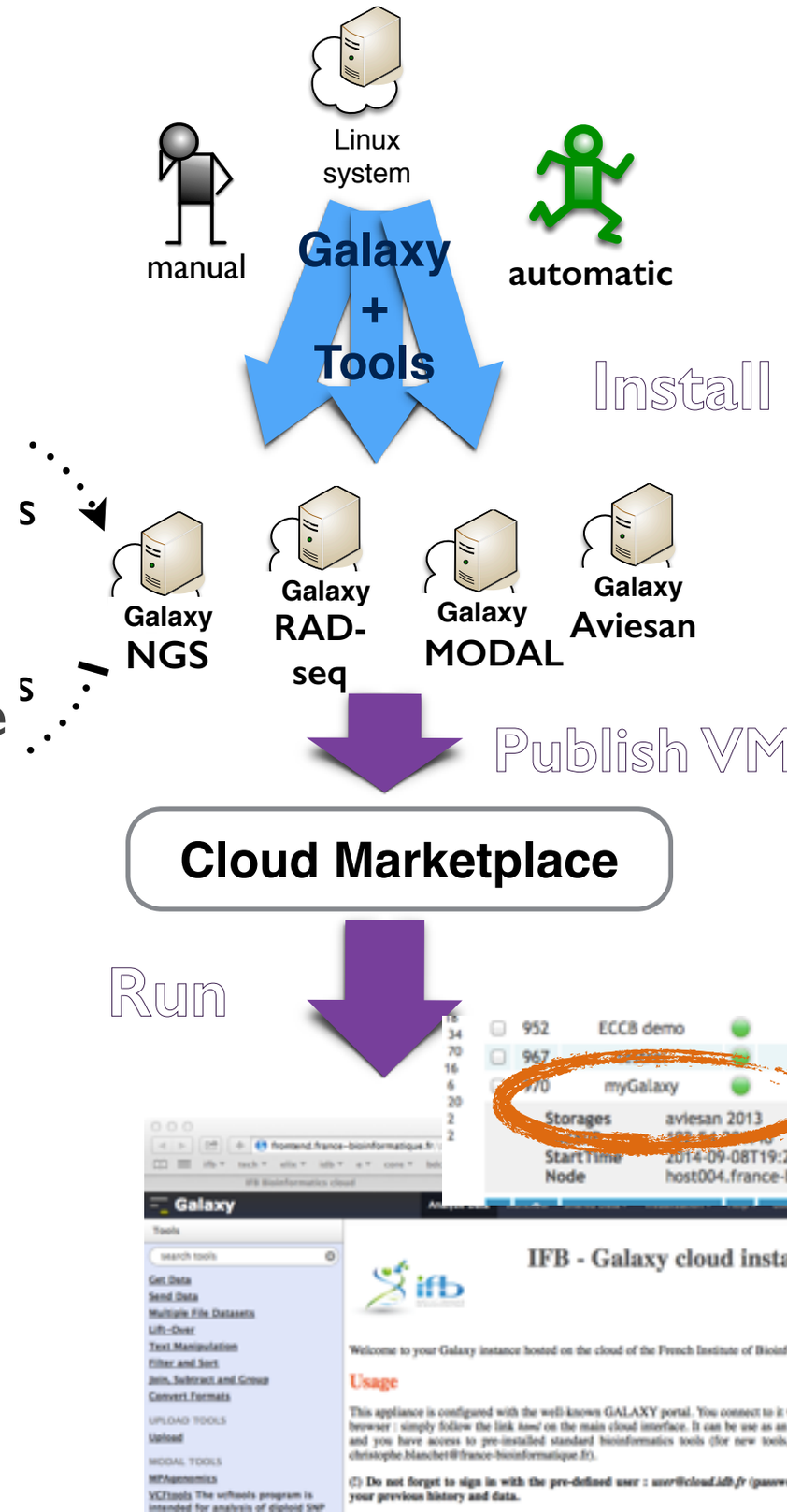
AppS Cloud Galaxy Portal

Galaxy portal is widely used in the community

- analyse NGS data (mainly but not only)
- connected to community knowledge: data and indexes, tools, workflows

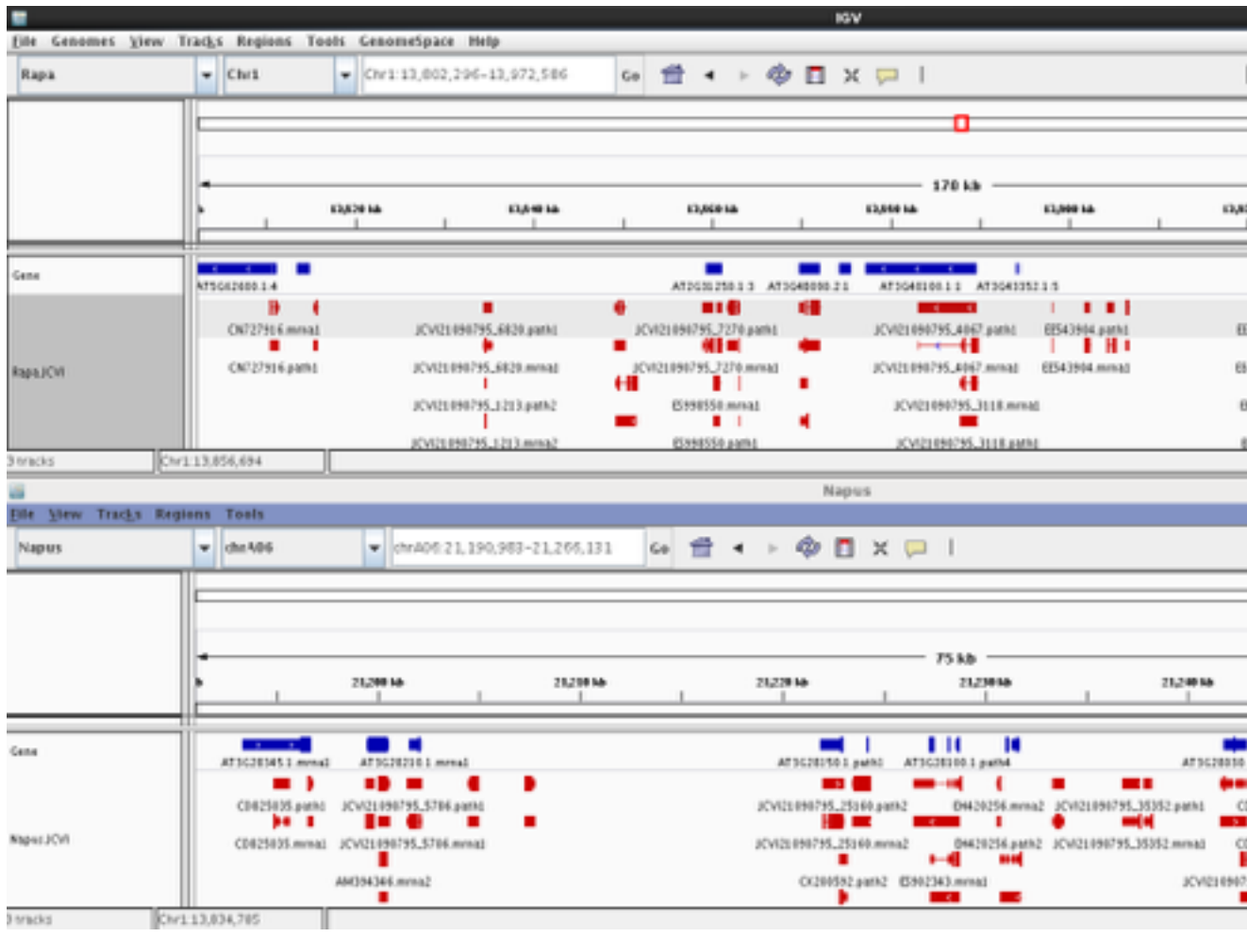
Cloud advantages :

- User is **administrator of his/her own Galaxy** instance: he/she can install data and tools
- Preserve **workflows and results in cloud storage**
- Help the integration of monthly updates and new tools
- Different appliances can be available at the same time:
  - ⋆ a basic one with common tools for NGS
  - ⋆ specific ones for a domain or a set of tools
    e.g. Galaxy-MODAL, Galaxy-RADseq, **EBA-ChIP-Seq**
  - ⋆ or for training: create a special appliance with dedicated datasets, tools or workflows
    e.g. AVIESAN school 2015



App Multi-genomes browser

Integrative Genomics Viewer



Based on IGV

Ready to deploy in the cloud close to the datasets

Remote virtual desktop

- transfer only graphical visualization
- based on NX protocol

Contact: Marie-Laure Franchinard (IFB MIGALE)

Funded by the French BIODATACLOUD project.

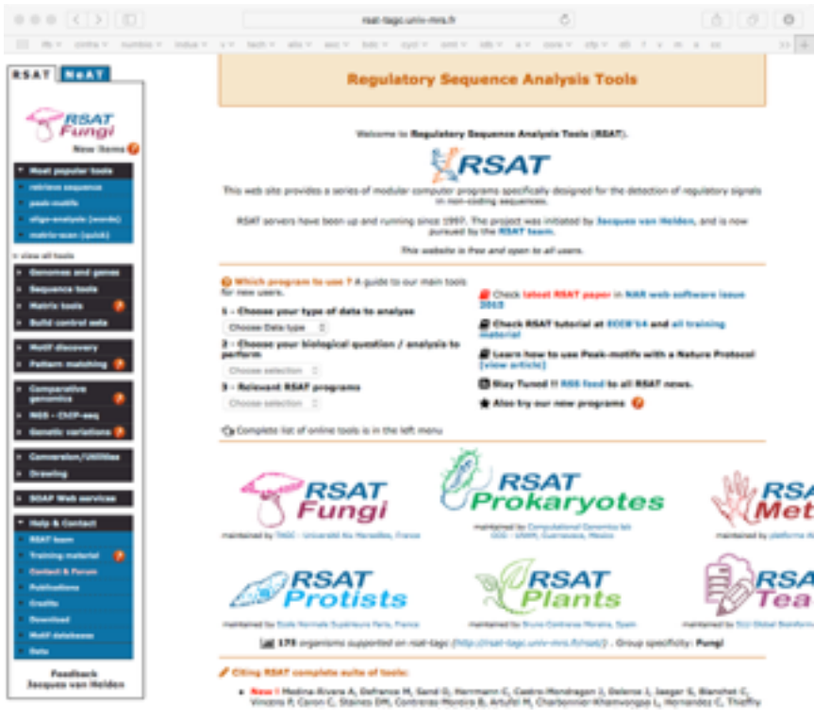
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

BIODATACLOUD

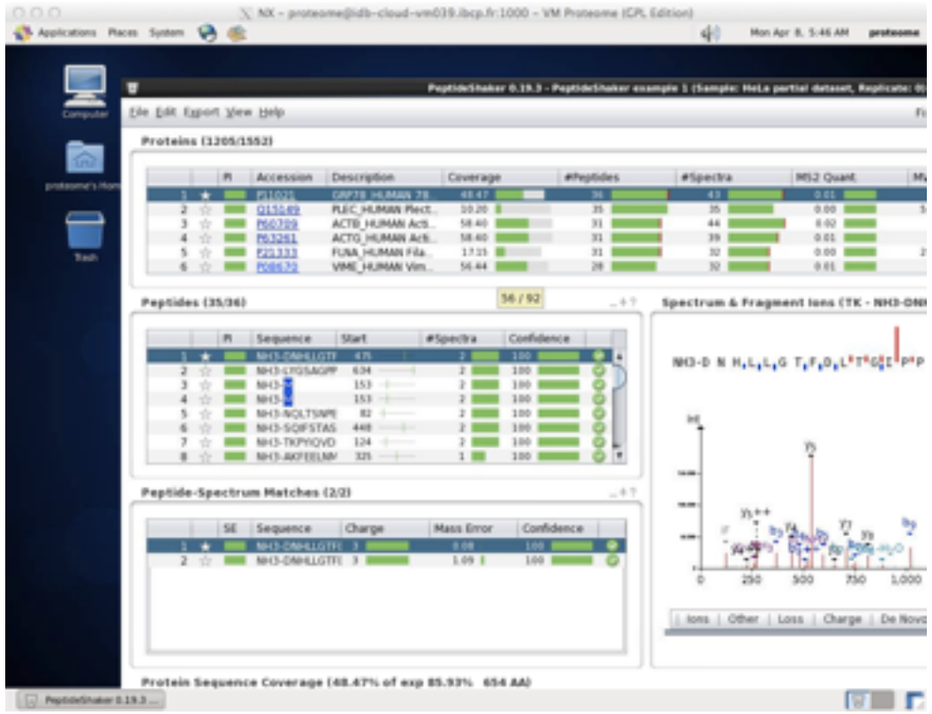
bpifrance

INVESTISSEMENTS D'AVENIR

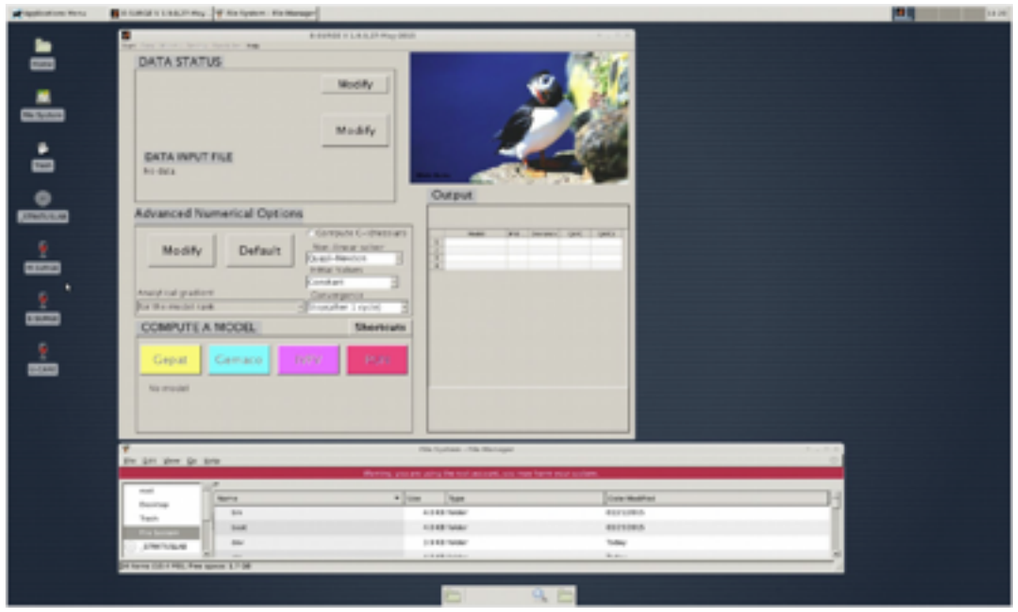
And other apps …



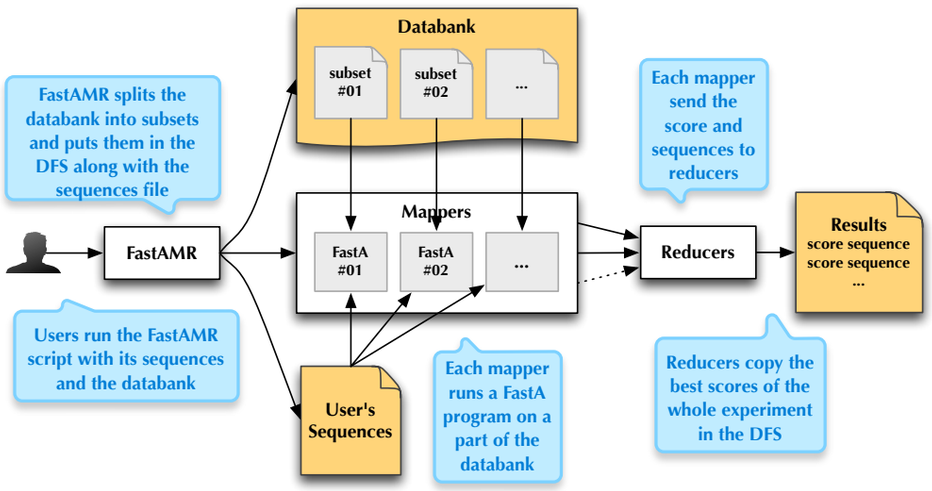
RSAT



Proteomics



Ecology of populations



Hadoop

Cloud training sessions 2014-15

Title	Teachers	Dates	Part.	Level
Tutoriels Cloud pour la Biologie *Gif-sur-Yvette*	IFB-core, GenOuest	juin 2014	23	débutant
Tutoriels Cloud pour la Biologie Rennes	GenOuest, IFB-core	nov. 2014	20	débutant
GRISBi-27 *Gif-sur-Yvette*	IFB-core	mars 2015	27	admsys, développeur
★ Ecole Cumulo Numbio *Aussois*	IFB-core, LRI, Grid5000, FranceGrilles, KerData	juin 2015	53	mixte
Journées work packages bioinformatique IFB *Gif-sur-Yvette*	IFB-core	juin 2015	15	développeur bioinformatique
Journée IBC *Montpellier*	IFB-core	nov. 2015	20	débutant

IBI-1 - IFB cloud basics

Sessions
- 3rd February 2016
- 16th April 2016

Public : tout public

Fréquence : bimestrielle

Niveau requis

- débutant sur le cloud,
- utilisateur de services bioinformatiques

Objectifs

- Savoir utiliser le cloud IFB pour des analyses de données biologiques : exécuter ses propres machines virtuelles et transférer ses données entre son poste de travail et le cloud et récupérer les résultats.

Contenu:

- Présentation du cloud IFB ;
- Présentation du tableau de bord ;
- Déploiement des machines virtuelles ;
- Gestion des données avec des disques virtuels, leur gestion ;
- Les différents types de connexion aux VMs (SSH, Web et bureau à distance).
- Pratique : utilisation de l'appliance Galaxy et d'un bureau virtuel

Documents:

- Diapos : Présentation du cloud IFB
- Fascicule : Utilisation du cloud IFB

IBI-2 - IFB cloud advanced

Session
- 3rd May 2016

Public : Utilisateur du cloud

Fréquence : trimestrielle

Niveau requis

- Avoir suivi IBI-1 (ou équivalent) et maitriser la ligne de commande

Objectifs

- Savoir déployer une application complexe pour l'analyse intensive de données biologiques de grande taille.
- Savoir adapter les machines virtuelles disponibles pour répondre à des besoins plus complexes.

Contenu:

- Déploiement d'une application bioinformatique complexe comprenant plusieurs machines virtuelles
- Installation de logiciels à partir d'archives (codes source ou binaires) ou à l'aide de scripts (approver)
- Utilisation de conteneurs docker pour l'installation de logiciels bioinformatiques
- Intégration des ressources de données (génome, annotation ...) grâce à l'appliance BiomaJ ou avec d'autres solutions
- Gestion des données avec des disques virtuels en NFS

Documents:

- Diapos : présentation des fonctionnalités avancées du cloud IFB
- Fascicule : cas pratique des fonctionnalités les plus fréquemment utilisées

IBI-3 - IFB cloud for developers

Sessions
- 16th mars 2016 (FG)
- *July 2016*

Public :
- Développeurs avec une pratique du cloud IFB

Fréquence : semestrielle

Niveau requis
- avoir suivi IBI-2 (ou équivalent) et connaitre le système d'exploitation Linux

Objectifs
- Savoir intégrer un logiciel ou un pipeline bioinformatique dans une machine virtuelle pour une diffusion et mise à disposition sur le cloud IFB.

Contenu:
- Présentation des bonnes pratiques de création d'appliance.
- Présentation des fonctionnalités avancées disponibles dans le cloud IFB : le montage automatique des collections de données publiques de référence, la contextualisation d'un portail web, la configuration des disques virtuels pour la conservation des paramètres d'un logiciel ou portail …
- Présentation des différents modèles d'intégration: archives (codes source ou binaires), scripts d'installation automatique (approver, puppet), conteneurs (docker).
- Création de conteneurs docker pour le déploiement de logiciels bioinformatiques
- Choix et configuration de l'interface pour les utilisateurs (CLI, portail web, bureau virtuel à distance)
- Rédaction d'une description pour le référencement dans le cloud IFB

Documents:
- Diapos : Présentation des bonnes pratiques de création d'appliance sur le cloud IFB
- Fascicule : Exemples de cas pratique

Master courses

Title	Teachers	Dates	Students	Special apps (*)
Master cours « cis-régulation » (M2) *Univ. Marseille*	J. van Helden, A. Griffon	oct. 2014	45	RSAT
Master Bioinformatique (M1) *Univ. Rouen*	S. Gallina	jan.-fév. 2015	10	Non *BIO ComputeNode Galaxy*
Polytech Biotech *Univ. Marseille*	D. Puthier	déc. 2015	38	TAGC Cours-Unix
Master Bioinformatique ADC (M1) *Univ. Lyon*	P. Veber	déc. 2015	18	Jupyter Notebook
Master Bionformatique AMI2B (M2) *Univ. Paris-Saclay*	D. Gautheret, F. Lemoine, C. Billerey	déc. 2015 -jan. 2016	20	COURS M2 Paris-Saclay 2015

** The special appliance was built by the teachers with the help of the IFB-core staff.*

Conclusion

Pilote infrastructure is running since 2014

- 2 clouds in IFB's datacenters, 2 others in collaboration with academia

36 bioinformatics appliances already available

- for different life science domains, several developments in progress

Scientific usage - 357 users

- 9500+ VMs created since 2014 on the IFB-core's cloud

Training (2014-16)

- Users & developers : 8 cloud sessions, 4 scientific trainings
- Students : 5 master courses (130 st.), 3 already planned for 2016

To whom it is available

- Members of IFB and French life science community
  - ⋆ academic and commercial
- Partners of academic/industry infrastructures and projects:
  - ⋆ national: BioDataCloud, ProFi, MetaboHub…
  - ⋆ European: ELIXIR-EXCELERATE, CYCLONE, EGI-ENGAGE…
- *Including trainers and trainees.*

Prospects

Extend the national cloud site

Help regional centers to deploy a cloud

Federate the national and regional clouds

Expand the catalogue of bioinformatics cloud services (with new VM/container images)

Make resources access simple with AAI and multi-cloud deployment

Provide users with a collaborative working space

Set up a continuous integration environment for the development and validation of bioinformatics images

Questions ?

http://www.france-bioinformatique.fr

Acknowledgments

- **IFB** members
  - IFB-core : **Awa, Bryan, Dominique, Jean-François, Jonathan, Frédéric, Mohamed, Patricia, Sandrine, Victoria** (*Alumni : Marie, Maxime, Quentin)*
  - Working group **IFB-GRISBI :** Olivier Collin and members
- **Appliances developers**
  Samuel Blanck (Inria Lille), **Jacques** van Helden (TAGC), **Stéphane** Delmotte (PRABI-LBBE), **Bruno** Spataro (PRABI-LBBE), **Marie-Laure** Franchinard (MIGALE), **Anis** Djari (BioinfoGenoToul), **Bertrand** Néron (Institut Pasteur), **Adrien** Josso (MicroScope), **Thomas** Lacroix (MIGALE), **Christian** Baudet (CLB), **Germain** Paimparay & **Baptiste** Brault (CFB), **Olivier** Inizan (URGI), **Jocelyn** Brayet (Institut Curie), **Guillaume** Brysbaert (Bilille), **Guy** Perrière (PRABI) …
- **CNRS IDRIS, StratusLab** developers

- IFB is funded by French programs **PIA INBS 2012**
- EU H2020 projects: **CYCLONE (644925), EXCELERATE (676559) and EGI-Engage (654142)**