PREDONx 2018 : Atelier sur la Préservation des Données Scientifiques

BigData Platform for Enhancing Life Imaging

Christophe CERIN, Leila ABIDI **-** Université Paris13
Salima Benbernou - Université Paris Descartes
Mehdi Bentounsi - Université Paris Descartes
Mourad Ouziri - Université Paris Descartes
Soror Sahri - Université Paris Descartes
Philippe Garteiser - INSERM
Mustapha Lebbah - Université Paris 13
Hanene Azzag - Université Paris 13
Michel Smadja - SisNCom

Montpellier, March 19, 2017

Outline: from Big Science to Big Population

- Motivations related to large scale platforms
- Background:
  - Criteria for observing platforms
  - High Performance Computing (HPC)
  - Apache suite for Big Data
  - Some projects
- Contribution:
  - CIRRUS overview: USPC platform
  - Overview of the Atlas IDV
  - Architecture of the Atlas IDV
  - Crowdsourcing
  - Linked Open Data based Semantic Enrichment
- Conclusion and perspective

Motivations related to large scale platforms

In experimental Sciences we need tips for using them appropriately / best practices

Definition of a large scale system

- System with unprecedented amounts of hardware (>1M cores, >100 PB storage, >100Gbits/s interconnect);
- Ultra large scale system: hardware + lines of source code + numbers of users + volumes of data;
- New problems:
  - built across multiple organizations;
  - often with conflicting purposes and needs;
  - constructed from heterogeneous parts with complex dependencies and emergent properties;
  - continuously evolving;
  - software, hardware and human failures will be the norm, not the exception.

Some speeches are ready for you

- Don’t believe them;
- Variety in the eco-system is the norm.

Since in our lives (Cloud Computing)

Software as a Service

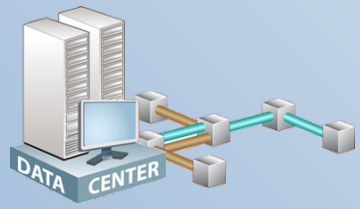
SaaS

Plateform as a Service

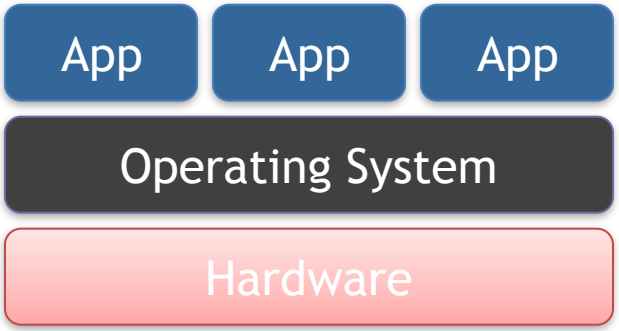
PaaS

Infrastructure as a Service

IaaS

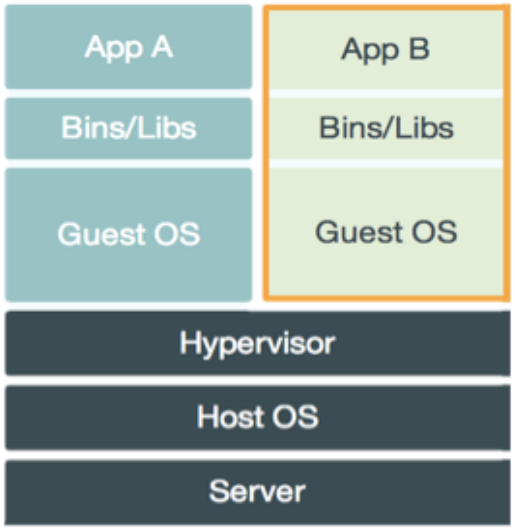


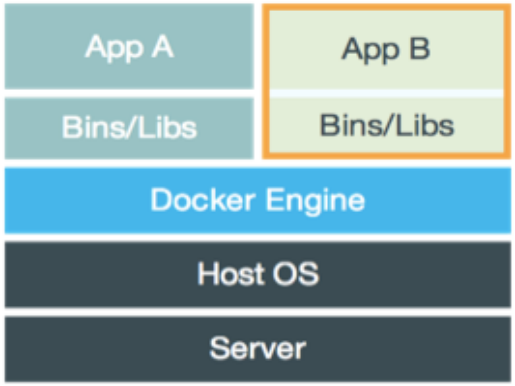
VM versus Container



Conventional Stack

- **Abstraction on resources in the past:** VM technologies to hide physical properties, interactions between system layer, application layer and the end-users

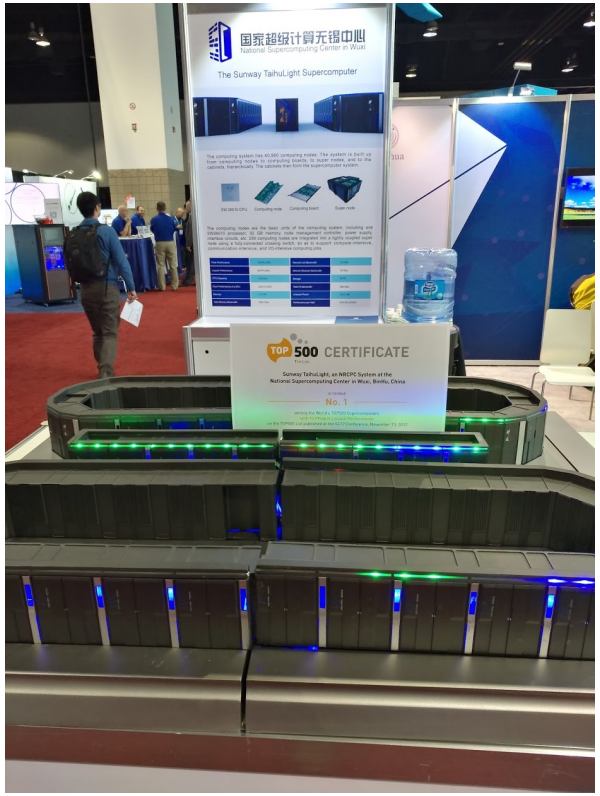




In Sciences (not in business)

- Experimental approach: modeling,…, experiments,…
- What is an experimental plan for large scale systems and reproducible research?
- Answers depend on the system used:
  - Cluster (dedicated; managed)
  - Cloud (almost non supervised… or by you)

Background


国家超级计算无锡中心
National Supercomputing Center in Wuxi
The Sunway TaihuLight Supercomputer
TOP 500 CERTIFICATE
No. 1


MERIDIEN
NVIDIA
DENVER
WELCOMES
SC17

Performance criteria for observing large scale system

- Cluster →Performance →Flops - Top500
- Cluster →Power →Flops per watt - Green500
- Cloud →Price →for computing, for storing…
- Cloud: the economic model may be strange (Amazon Spot Instances)
- Scientific issues: how to compare cluster and cloud? (be fair)

Performance (Top500)

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272
4	Japan Agency for Marine-Earth Science and Technology Japan	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz ExaScaler	19,860,000	19,135.8	28,192.0	1,350
5	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
6	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890

Performance (Green500)

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	259	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	842.0	50	17.009
2	307	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	788.2	47	16.759
3	276	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	149	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	1,350	14.173

Productivity

- **Productivity** traditionally refers to the ratio between the quantity of software produced and the cost spent for it.
- **Factors:** *Programming language used*, Program size, The experience of programmers and design personnel, The novelty of requirements, The complexity of the program and its data, The use of structured programming methods, Program class or the distribution method, Program type of the application area, Tools and environmental conditions, Enhancing existing programs or systems, Maintaining existing programs or systems, Reusing existing modules and standard designs, Program generators, Fourth-generation languages, Geographic separation of development locations, Defect potentials and removal methods, (Existing) Documentation, Prototyping before main development begins, Project teams and organization structures, Morale and compensation of staff

Programming model

- Parallel (http://press3.mcs.anl.gov/atpesc/files/2015/03/ESCTP-snir_215.pdf):
  - Shared memory (OpenMP):
    - CPU+vector
    - CPU+GPU
    - CPU+accelerators (Google Tensor processing unit…)
  - Message passing (MPI)
- Sequential with the call to « parallel data structure » (Scala)
- Map-reduce: Hadoop, Spark (In addition to Map and Reduce operations, it supports SQL queries, streaming data, machine learning and graph data processing.)

Hardware landscape

- Even more complicated: Cray has added "ARM Option" (i.e. CPU blade option, using the ThunderX2) to their XC50 supercomputers
- We need consolidations:
  - in hardware
  - in software
  - in algorithms
  - in applications
  - in experimental methods
  - hpc facilities
  - training

Apache suite for Big Data

1- Bases de Données
- 1.1 Apache Hbase
- 1.2 Apache Cassandra
- 1.3 CouchDB
- 1.4 MongoDB

2 Accès aux données/ requêtage
- 2.1 Pig
- 2.2 Hive
- 2.3 Livy

3 Data Intelligence
- 3.1 Apache Drill
- 3.2 Apache Mahout
- 3.3 H2O

4 Data Serialisation
- 4.1 Apache Thrift
- 4.2 Apache Avro

5 Data integration
- 5.1 Apache Sqoop
- 5.2 Apache Flume
- 5.3 Apache Chuckwa

6 Requetage
- 6.1 Presto
- 6.2 Impala
- 6.3 Dremel

7 Sécurité des données
- 7.1 Apache Metron
- 7.2 Sqrrl

8.2 MapReduce
8.3 Spark
9.1 Elasticsearch
10.1 Apache Hadoop
10.6 Apache Mesos
10.10 Apache Flink
10.12 Apache Zookeeper
10.15 Apache Storm
10.20 Apache Kafka

Example of a project mixing HPC and Big Data considerations

Where theory comes before data:

- physics
- chemistry
- Materials
- ...

Where data comes before theory:

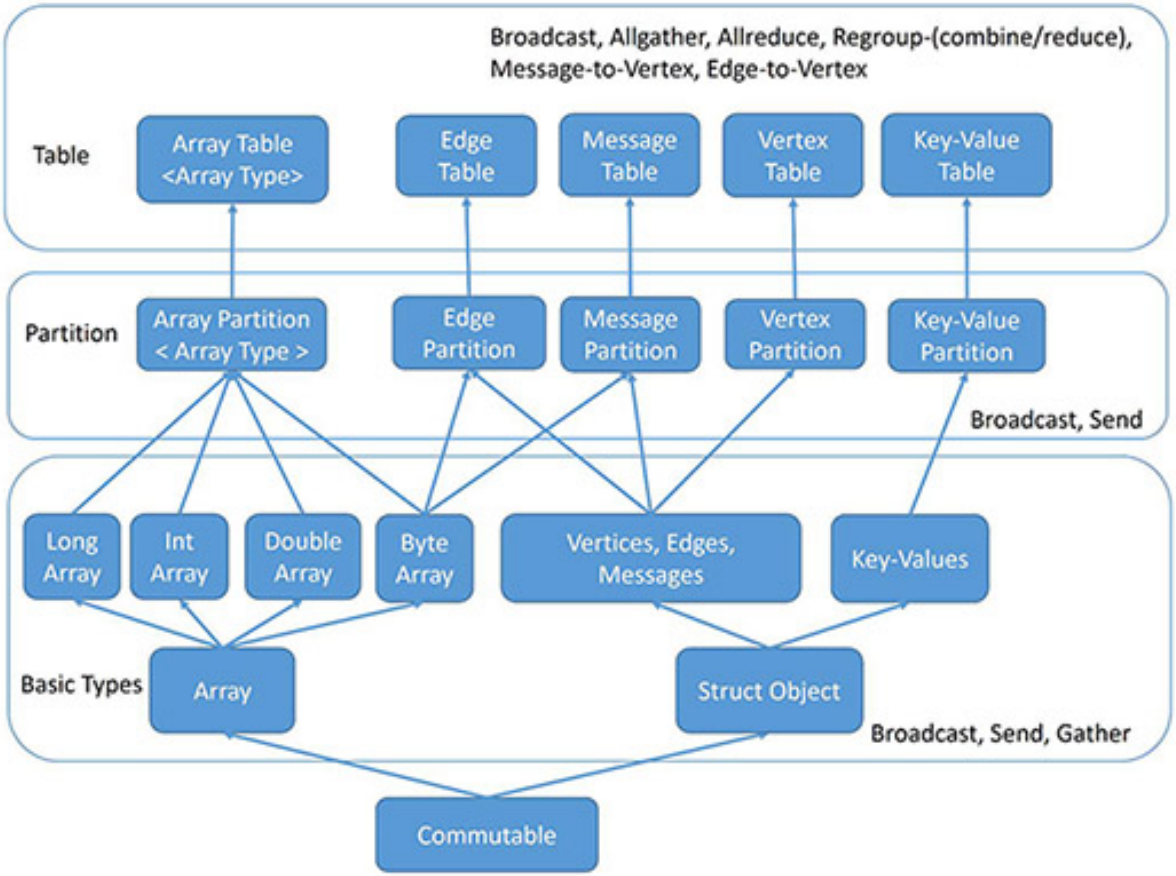
- Medicine
- business insights
- autonomy (smart car, building, city)

Harp: Hadoop and Collective Communication for Iterative Computation

- Project http://salsaproj.indiana.edu/harp/
- Motivation:
  - Communication patterns are not abstracted and defined in Hadoop, Pregel/Giraph, Spark
  - In contrary, MPI has very fine grain based (collective) communication primitives (based on arrays and buffers)
- Harp provides data abstractions and communication abstractions on top of them. It can be plugged into Hadoop runtime to enable efficient in-memory communication to avoid HDFS read/write.

Harp

- The data abstraction has 3 categories horizontally and 3 levels vertically



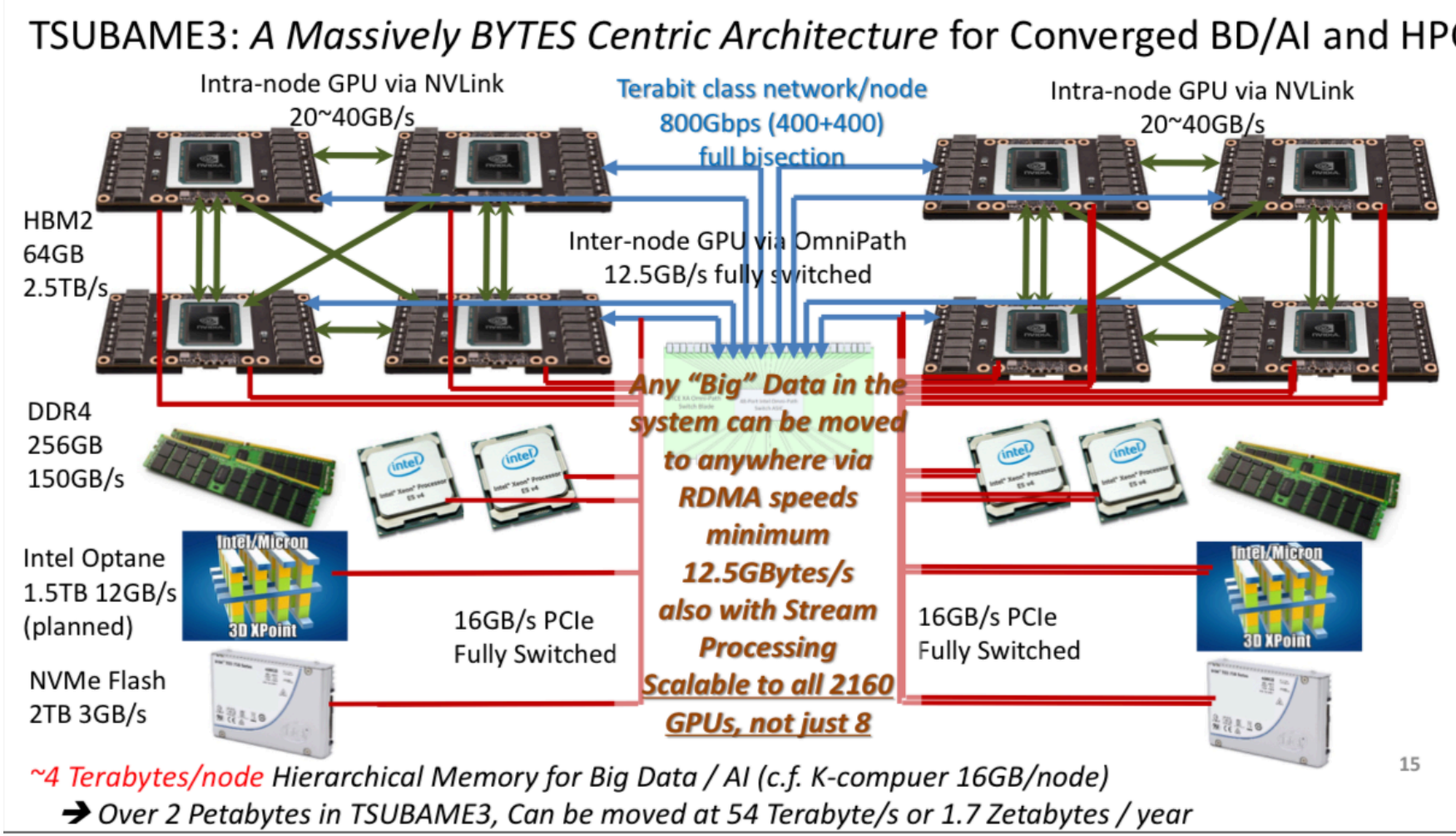
Harp

- Harp works as a plugin in Hadoop. The goal is to make Hadoop cluster can schedule original MapReduce jobs and Map-Collective jobs at the same time.
- Collective communication is defined as movement of partitions within tables.
- Collective communication requires synchronization. Hadoop scheduler is modified to schedule tasks in BSP style.
- Fault tolerance with checkpointing.

BYTES-oriented: what does it means?

- Satoshi Matsuoka (Tsukuba, Japan): On the convergence of HPC and Big Data (video on Youtube)
- Being "BYTES-oriented" in HPC leads to an open big data/ AI ecosystem and further advances into the post-moore era. BigData 2017 speech;
- http://web.cse.ohio-state.edu/~lu.932/hpbdc2017/slides/hpbdc17-satoshi.pdf
- Inference, training, generation, learning… as HPC tasks:
  - Need BOTH bandwidth and capacity (BYTES) in an HPC-BD-AI machine
  - we lack the cutting edge infrastructure dedicated to AI and Big Data

BYTES-oriented: what does it means?



BYTES-oriented: what does it means?

- Okay. Imagine we know how to build scalable architecture for data centric computing (both in terms of capacity and bandwidth)
- Big issue: *what are the benchmarks to use?*
  - no one can say!
  - nobody wants to know (?)

RosettaHub
(a focus on Big Population)

- For the data-scientist of nowadays;
- Hub to facilite access to Amazon cloud resources, to share data, to collaborate between many people through notebooks;
- RosettaHUB: https://www.rosettahub.com
- Available through Amazon Educate program (https://s3.amazonaws.com/awseducate-list/AWS_Educate_Institutions.pdf)

RosettaHUB, from the cloud to the open science cloud a universal open platform for data science

Computational Components

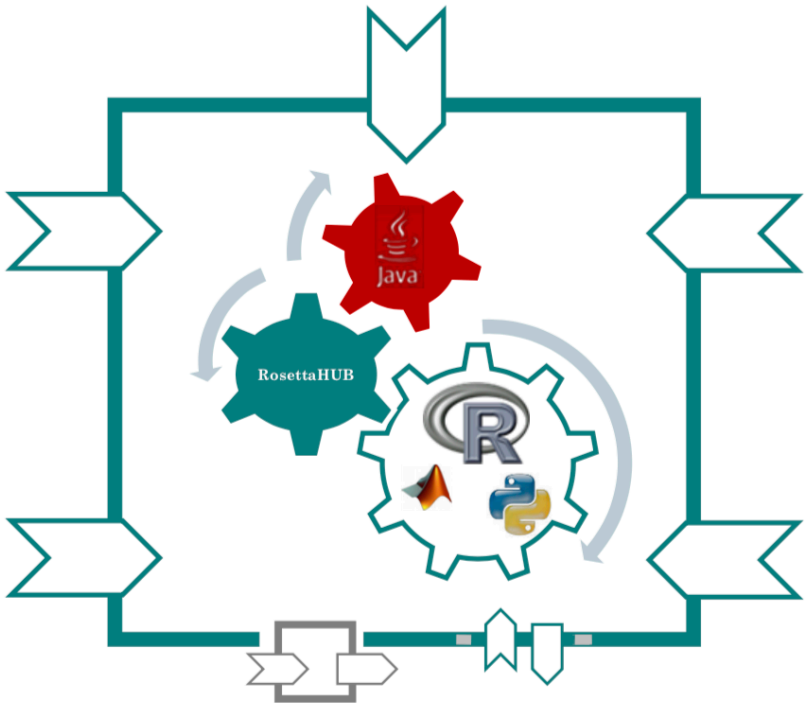
R packages, Wrapped C,C++,Fortran code, Python modules, Matlab Toolkits…

Open source or commercial

Computational Resources

Clusters, grids, private or public clouds

Free or pay-per-use



Computational GUIs

HTML5 and Desktop Workbench

Built-in views /Plugins /Collaborative views

Open source or commercial

Computational Storage

Local, NFS, FTP, Amazon S3, EBS, HDFS

Computational Scripts

R / Python / Matlab / Groovy

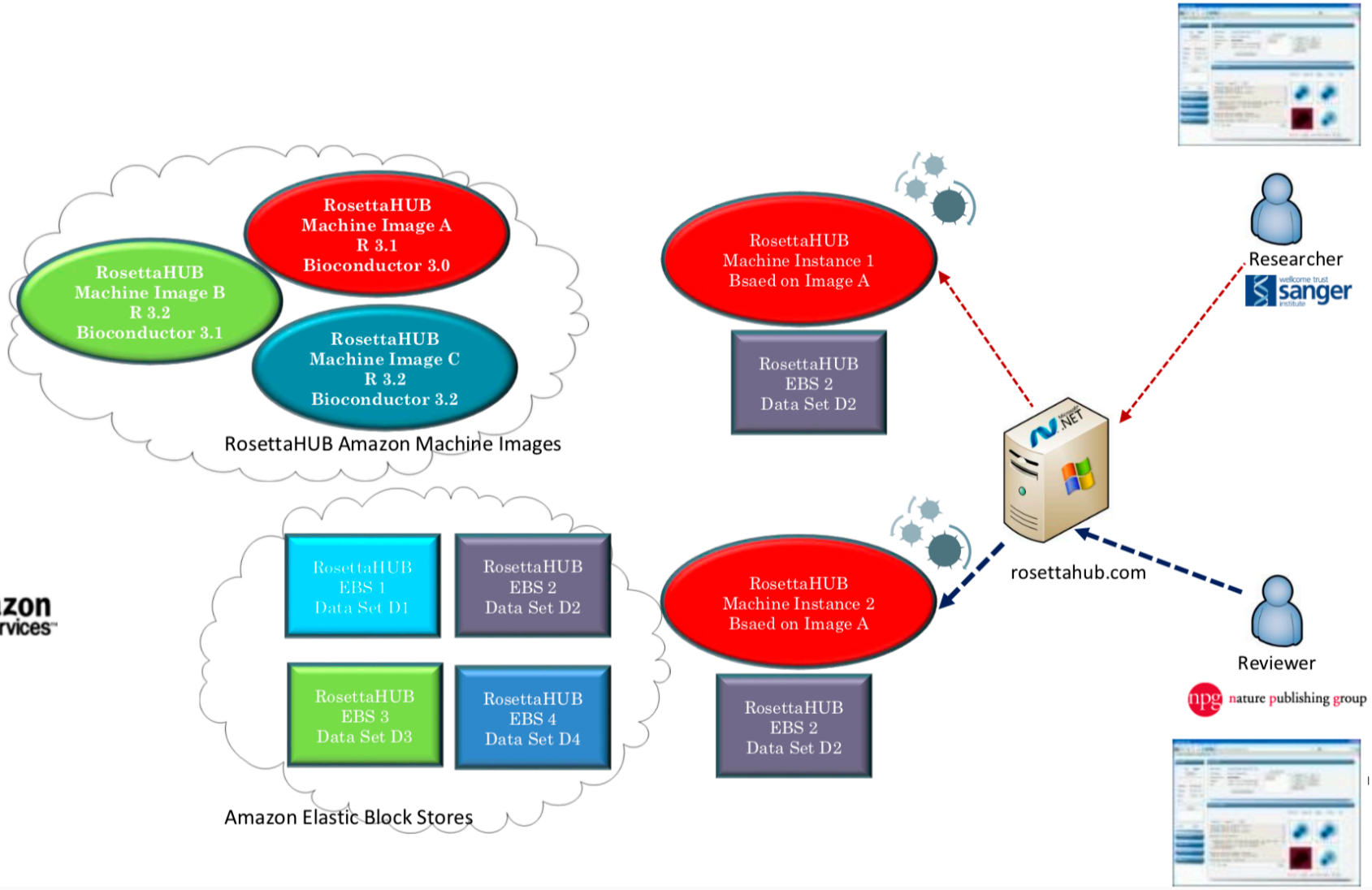
Generated Computational Web Services

Stateful or stateless, mapping of R objects/functions

Computational APIs

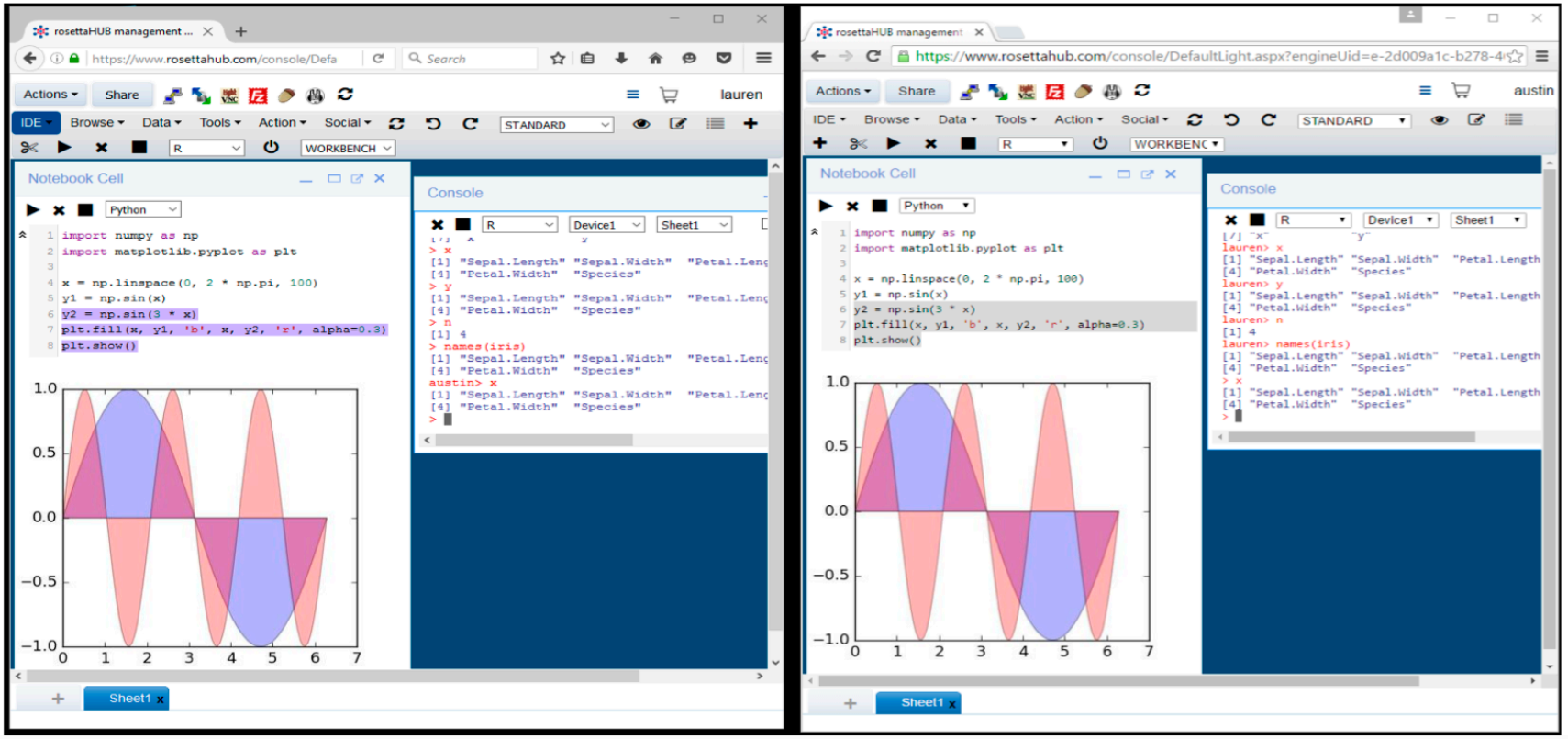
Java / SOAP / REST, Stateless and stateful

Traceable and reproducible data science

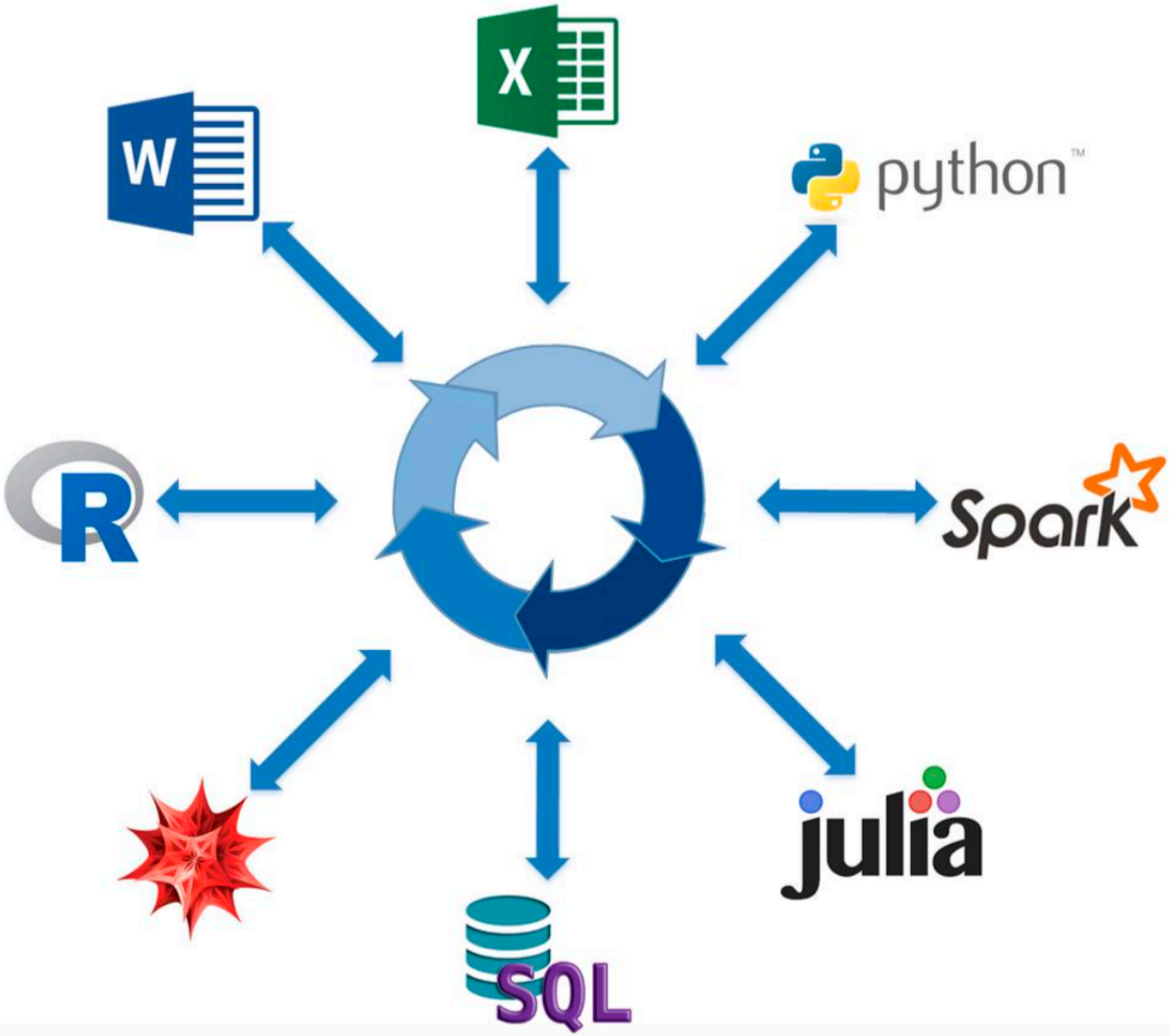


Google Docs-like real-time collaboration

- Collaborative spreadsheets
- Collaborative scientific graphics canvas
- Collaborative dashboards
- Collaborative widgets



A multi-language framework / Hybrid Engines



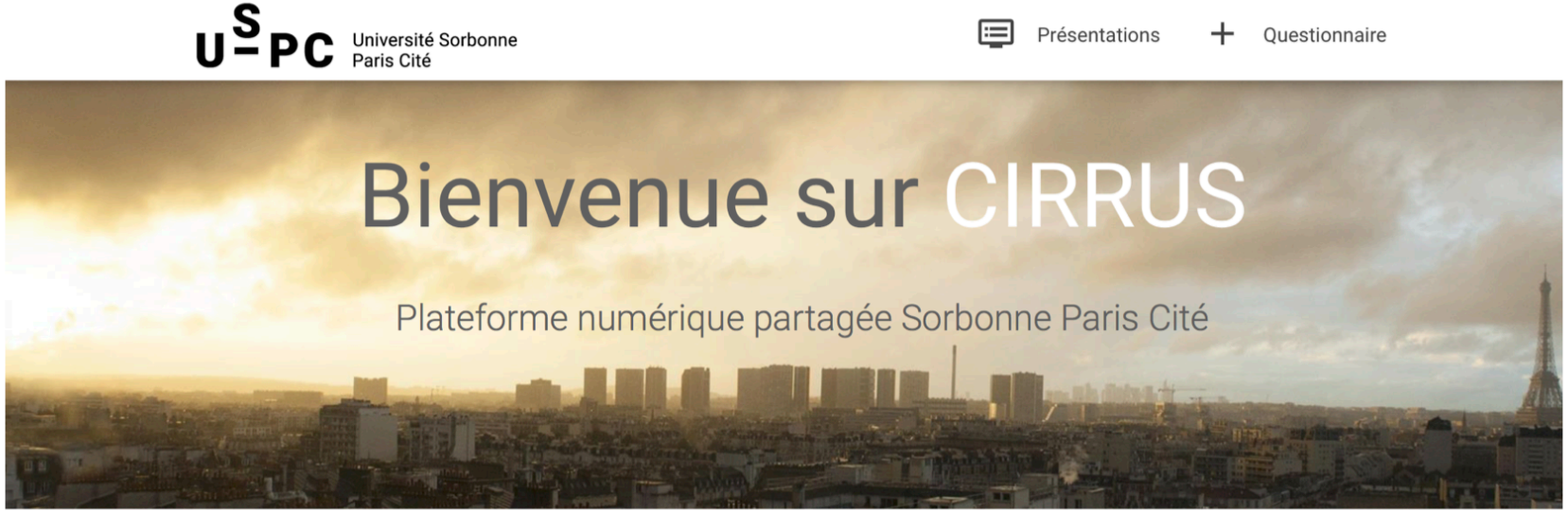
SPC and the deployment of virtual machines

IDV experience

(Imageries du vivant)

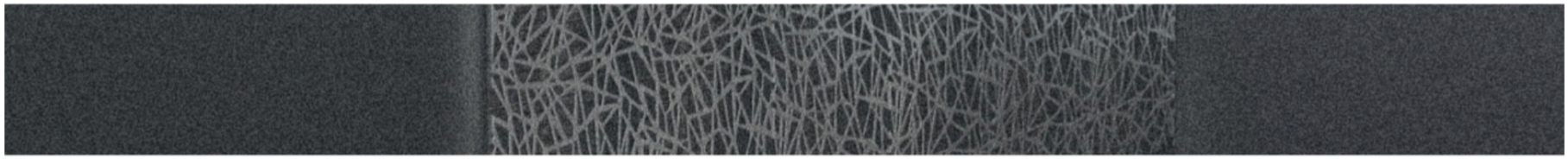
http://cirrus.uspc.fr

LA PLATEFORME SPC



Les multiples appels à projets lancés par SPC rendent possible de fortes synergies entre les équipes de recherche. Pour autant, les équipements et compétences numériques sont encore largement dispersées dans les laboratoires. Mutualiser une partie du support numérique à la recherche améliorera les services offerts aux chercheurs et réduira sensiblement leurs coûts liés au numérique. Par ailleurs, une infrastructure numérique partagée facilitera les collaborations ainsi suscitées entre les établissements membres et avec les EPST.

LES PLATEFORMES ETABLISSEMENTS



MAGI

USPC s'appuie sur le centre interdisciplinaire de calcul de Paris 13 pour offrir des moyens de calcul intensif et distribué à la communauté scientifique. Cet outil

S-CAPAD

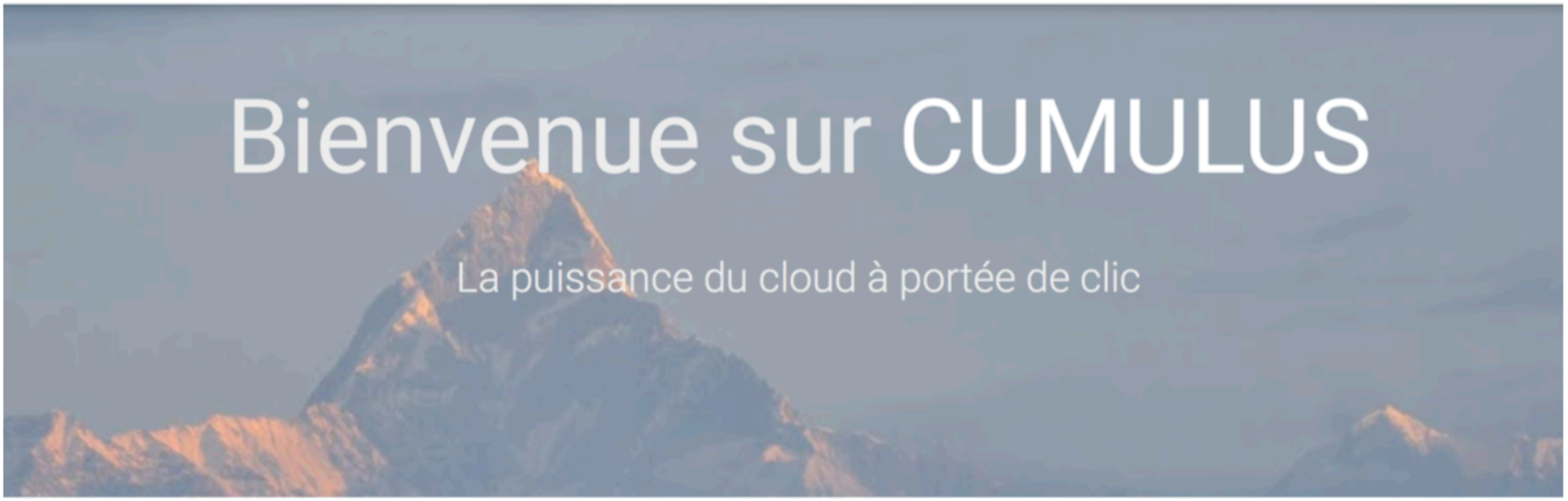
USPC s'appuie sur le Service de Calcul Parallèle et de Traitement de Données en sciences de la Terre de l'Institut de Physique du Globe de Paris (IPGP). Il fournit les

Cumulus

USPC s'appuie sur la plateforme de l'université Paris Descartes et la développe pour mettre en œuvre un cloud privé à destination de ses chercheurs. Elle offre aujourd'hui

UNIVERSITÉ PARIS DESCARTES

Se connecter



Pourquoi?

Cumulus est une plateforme de virtualisation dédiée aux chercheurs. Son objectif : vous offrir des serveurs puissants, sécurisés,

Comment?

En pratique, tout personnel de recherche associé à Sorbonne Paris Cité peut créer un compte[1], puis choisir le type d'environnement dont

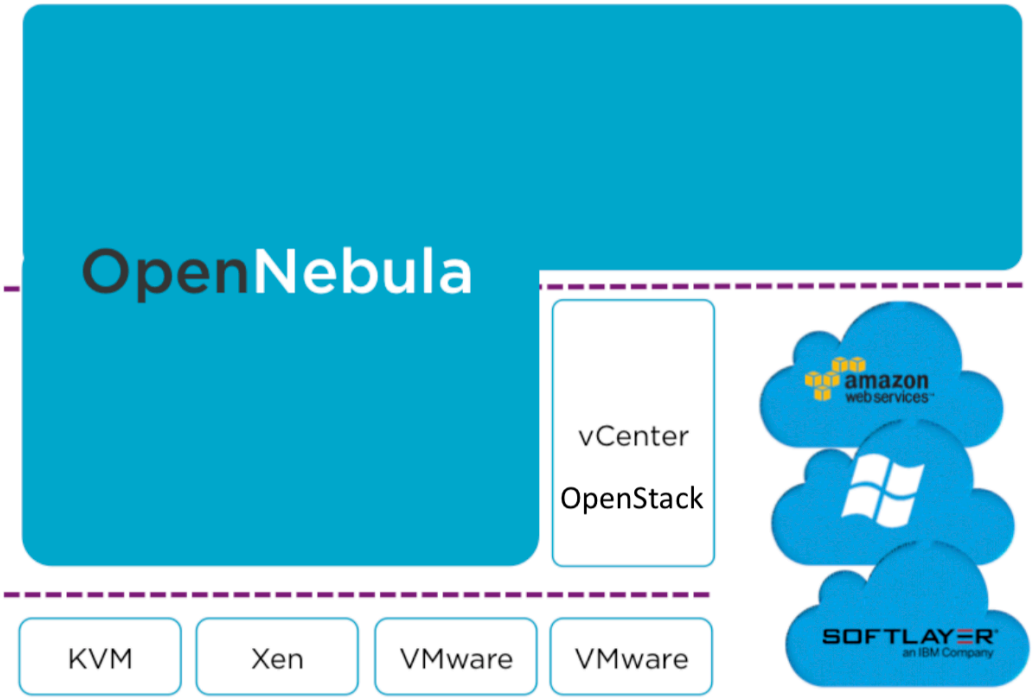
Qui peut m'aider?

Tout personnel de recherche associé à Sorbonne Paris Cité peut créer un compte[1], puis choisir le type d'environnement dont il a

OpenNebula Technology

Cloud Management
- Multi-tenancy
- Simple cloud GUI and interfaces
- Service elasticity/provisioning
- Federation

Virtual Infra Management
- Capacity management
- Virtual appliance management
- Resource optimization
- HA and business continuity



Examples (Jan 2018)

- 600 researchers use CIRRUS - More than 250 ongoing projects in all areas, among them IDV projects:
  - Plateforme d'imagerie ePad
    http://pf-01.lab.parisdescartes.fr:1243/dcm4chee-web3/
  - Plateforme PLM dans le cadre du projet DRIVE-SPC
    http://pf-01.lab.parisdescartes.fr:2345/tc/webclient
  - Plateforme Sis4web de SisNCom
    http://pf-01.lab.parisdescartes.fr:1323/sis4web/login.php
  - Benchsys
    http://pf-01.lab.parisdescartes.fr:9001/
  - Owncloud
    http://pf-01.lab.parisdescartes.fr:1253/owncloud/index.php/login
- …

BigData Platform for Enhancing Life Imaging

IDV

Imageries du vivant

General objectives

- Develop computational models and methods for imaging and quantitative image analysis, and
- *V*alidate the added diagnostic and therapeutic value of new imaging methods and biomarkers.
- Observations: Imaging is characterized by a large diversity in the types of data (multiple acquisition devices, formats, modalities, scale and finality)

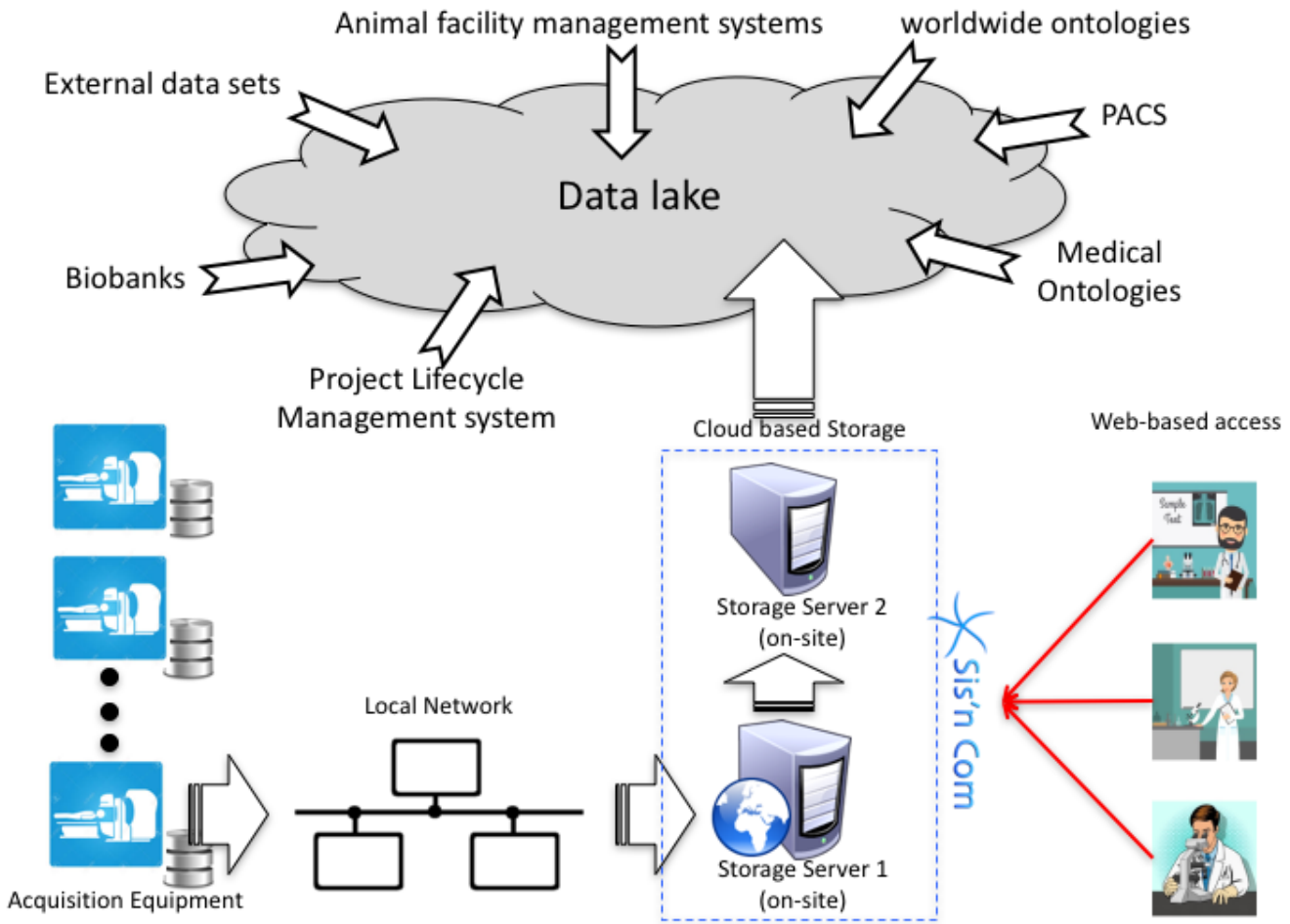
The Atlas IDV initiative

- 20 research groups, 10 sites (Saint Pères, Biomedical center, HEGP, Necker Hosp, Bichât-Beaujon, Cochin…)
- The integrated imaging operational system promotes intra and inter-sites collaborative research work.
- Based on the CIRRUS platform

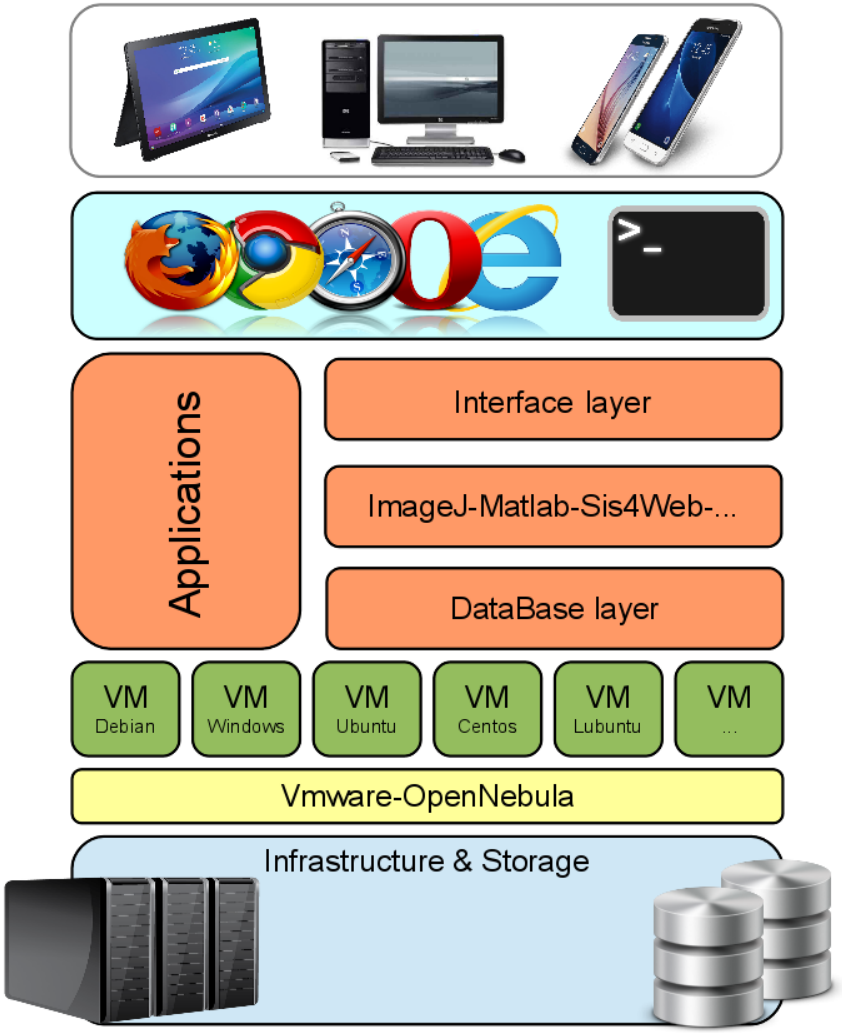
The Atlas IDV initiative

- To improve the image sets quality and indexing, native metadata are enhanced using 3 main strategies:
  - A robust annotation scheme (using web-based standardized annotation tools)
  - An enrichment of images sets via crowdsourcing
  - A semantic enrichment using linked open data and medical ontologies
- Privacy-enhancing technologies: preserve patient privacy by anonymizing images and metadata + control the identification risk for the entire lifecycle of the Atlas IDV

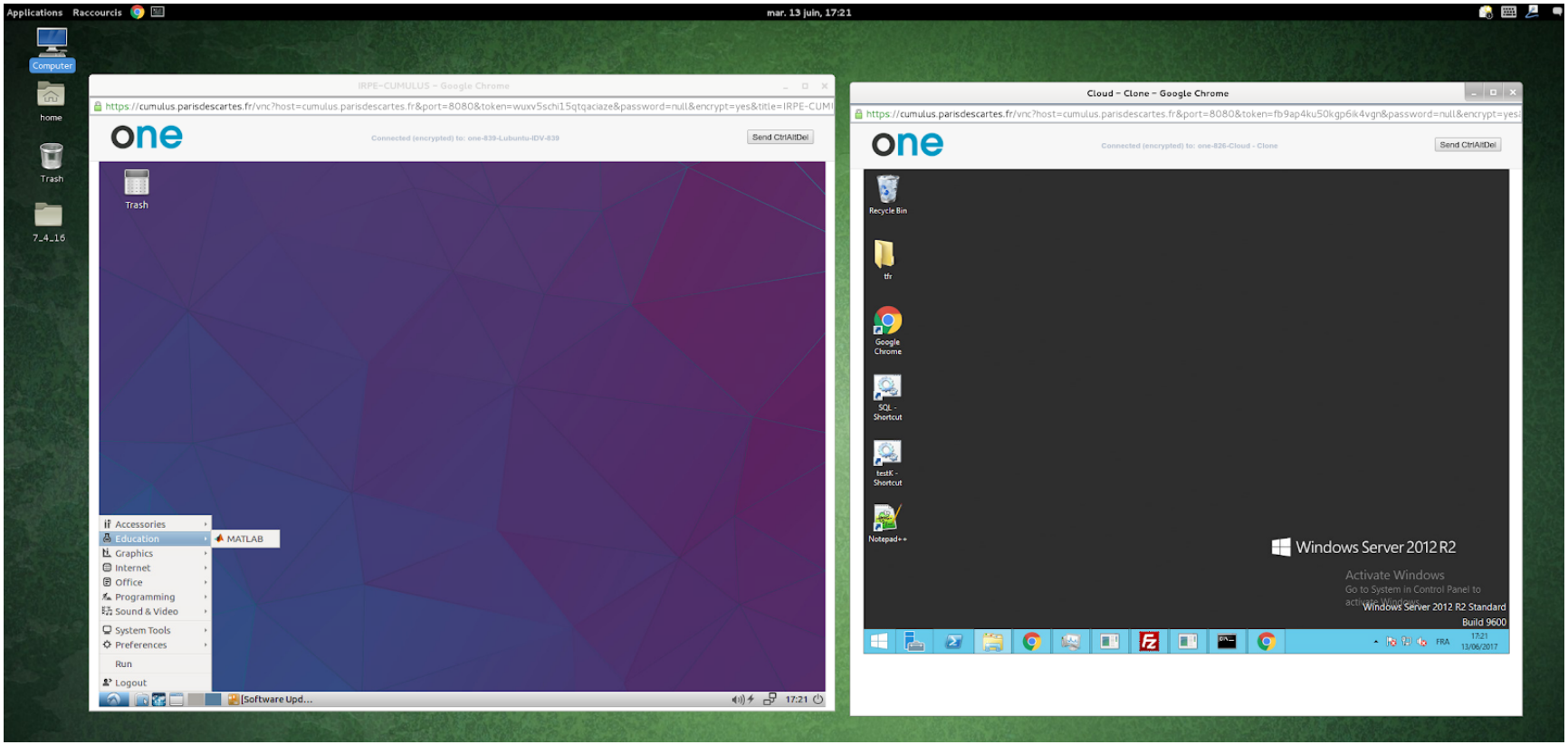
The Atlas IDV architecture



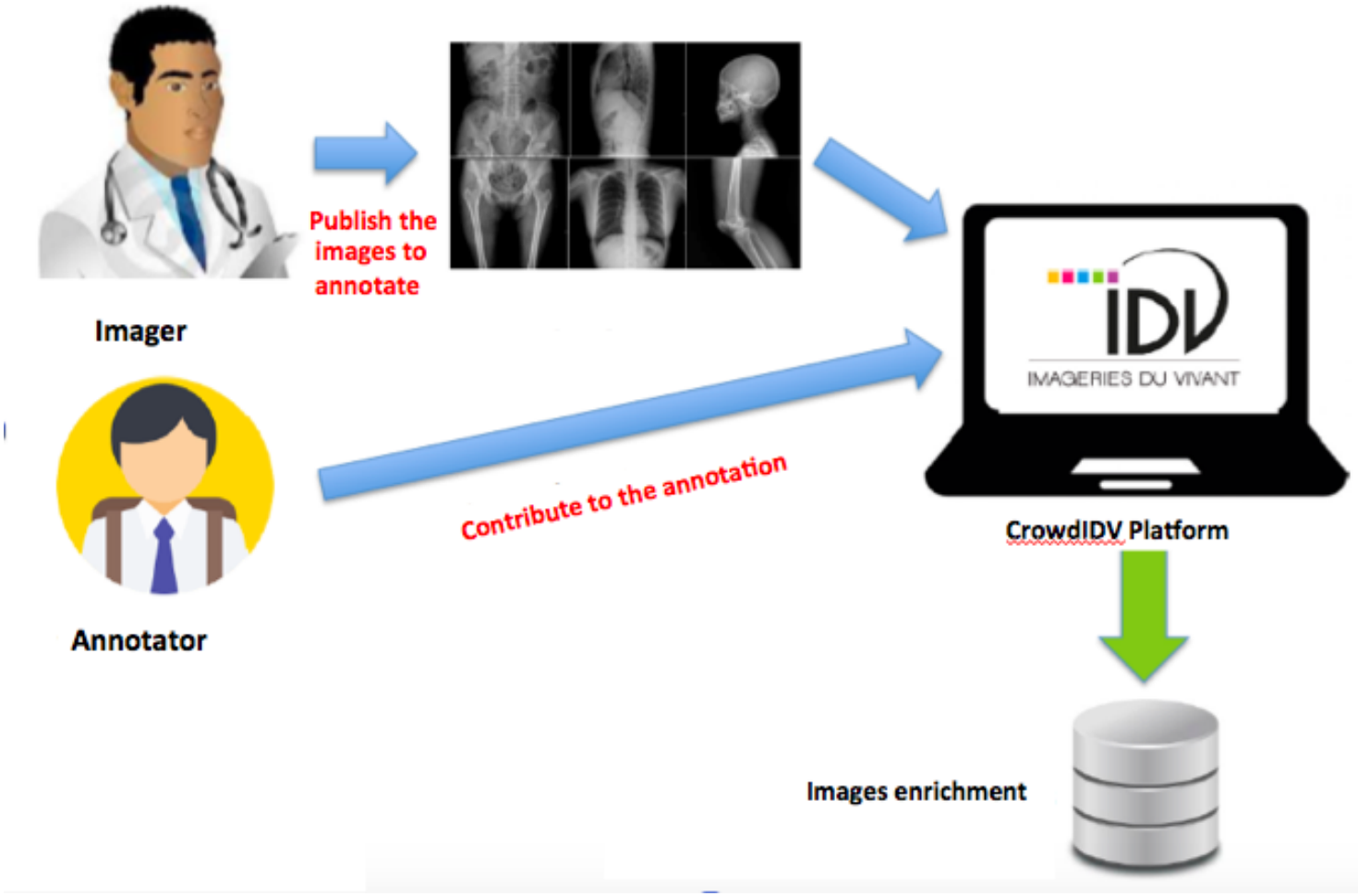
The Atlas IDV: building the dedicated cloud infrastructure



The Atlas IDV: building the dedicated cloud infrastructure



Crowdsourcing



Crowdsourcing

- Relevant annotations:
  - each imager submits its task for only its students
  - Apriori like algorithm (mine frequent itemsets)

Linked Open Data based Semantic Enrichment

- Brute images… but also data related to multi-modal and multi-scale life imaging including text, pre-clinical data and images of cells etc.
- Different banks of images are stored in their acquisition place, and such spatial fragmentation leads to under exploitation
- For clinicians, the aim is to enrich their clinical medical data and images with pre-clinical data and images, and vise versa for biologist.

Linked Open Data based Semantic Enrichment

Clinical images

Clinical image annotation

Clinical images datasource:ds1
--Incity(adr03,paris)
-Living(Bob, adr03)
-Has-Id-echo(IE4,Bob)
-Patient(Bob)
-Type-IRM(lung,adr03)
-Has-Disease(IE4, **Cancer**)
-Disease(Cancer)

Incomplete result= Bob has **cancer**

Linking image annotations

air quality open datasource:ds3
-Location(adr03, dep75)
-Type-pollen(P10)
-Propagate (P10,adr03)

Entity resolution

Need information about patient x?

Pre-clinical images

Cellular image annotation

Linking image annotations

Pre-clinical image datasource: ds2
-IsCell (**C20**)
-Patient(P15)
-IsPartOf(C20, Human)
-Belongs-to(C20,P15)
-Type-Cell(C20, human)
-Infection(C20,Hazel)
AllergyP(Hazel)
-Medicine (**M25**,C20)

Complete result with entity resolution= {**cancer**, **Paris** , **hazel** ,**M25** ,**C20**,}

Linked Open Data based Semantic Enrichment

- Life imaging data sources are linked at two levels:
  - *At the data level*, when querying data sources, the data sources linking process aims to identify the same real-life entities (such as patients, a human organ) and connect them using the owl relationship *sameAs*. This process is named *Entity Resolution* (ER).
  - *At the semantic level*. It is achieved using appropriate inference mechanisms to deal with big data and at the same time cleaning the big data produced when processing data fusion

Conclusion and Future work

Many, many opportunities

Conclusion and perspective

- Computer hardware will continue to evolve; Trend: accelerators instead of GPUs;
- Platforms will continue to evolve to facilitate the management of data, the computation on data;
- Need for convergence/resonance between HPC and Big-data eco-systems;
- SPC: needs for training, using, transforming the scientific approaches… with research engineers for accompany the change;
- SPC: needs for projects… and to share between communities.

Conclusion and perspective

- Google, Amazon, Facebook, Apple
- Microsoft, IBM
- and…

Alibaba Launches European Supercomputing Cloud Service, Quantum Computing Platform

Michael Feldman | March 2, 2018 07:23 CET

Chinese-base tech giant Alibaba is challenging American cloud providers in Europe with an HPC service designed for users running a variety of compute-intensive and data-intensive workloads. The company also unveiled a new cloud-based quantum computing platform.

Read more

Thanks to Olivier Waldek, Roland Chervet