Clouds in biosciences
Part II – grid usage in life sciences

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

A journey through CPU-intensive life sciences…

- Part I
  - Who am I?
  - Introduction to CPU-intensive life sciences
- Part II: Grid usage in life sciences
- Part III: Clouds in life sciences
- Part IV: Entering a new world

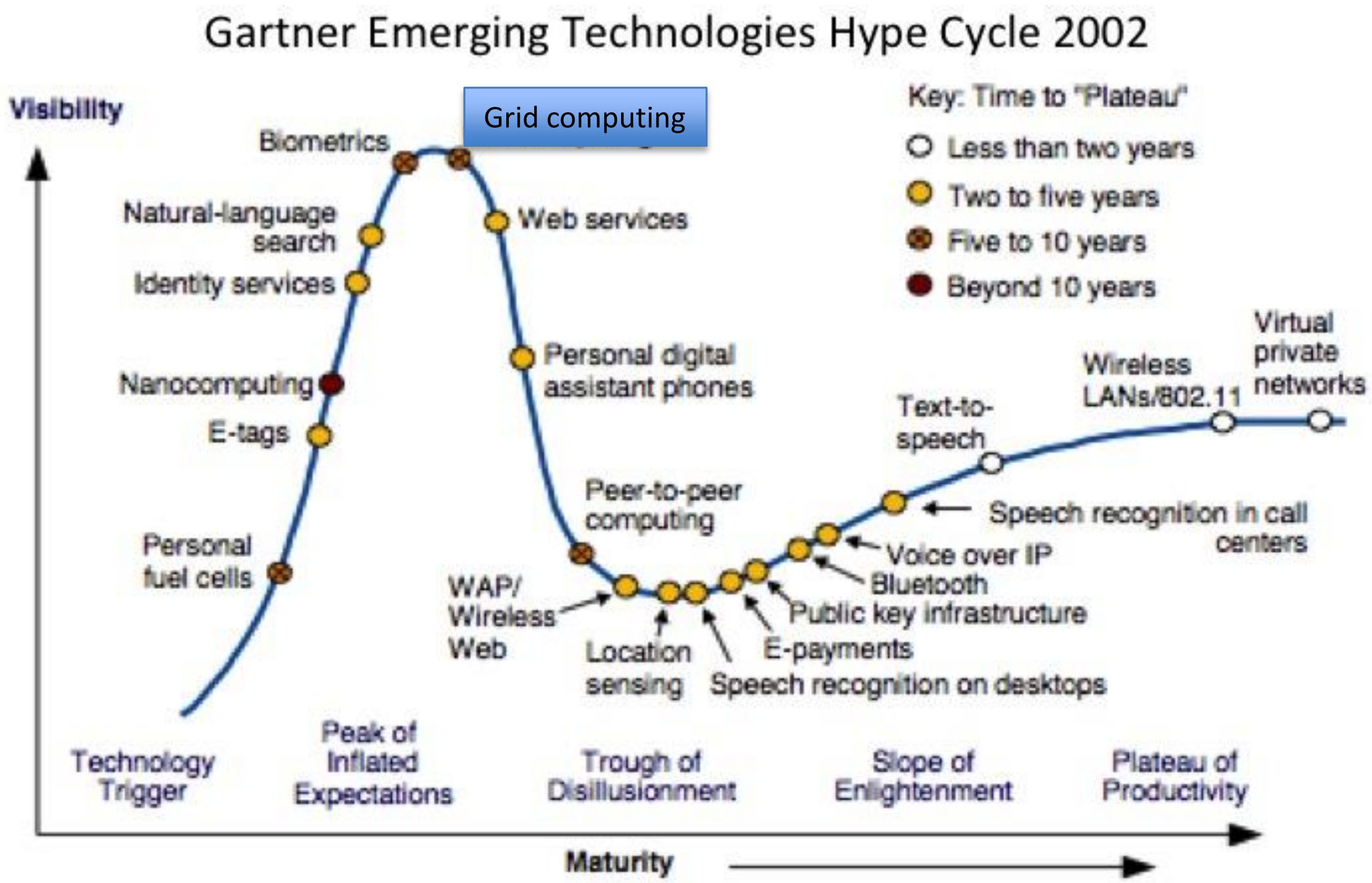
Session II: grid usage in biosciences

- Historical perspective: the different stages
- Examples at the different stages
  - First successes in life sciences
    - WISDOM (drug discovery)
  - Usage of grid on the plateau of maturity
    - WeNMR (structural biology)
    - VIP (medical imaging – neurosciences)

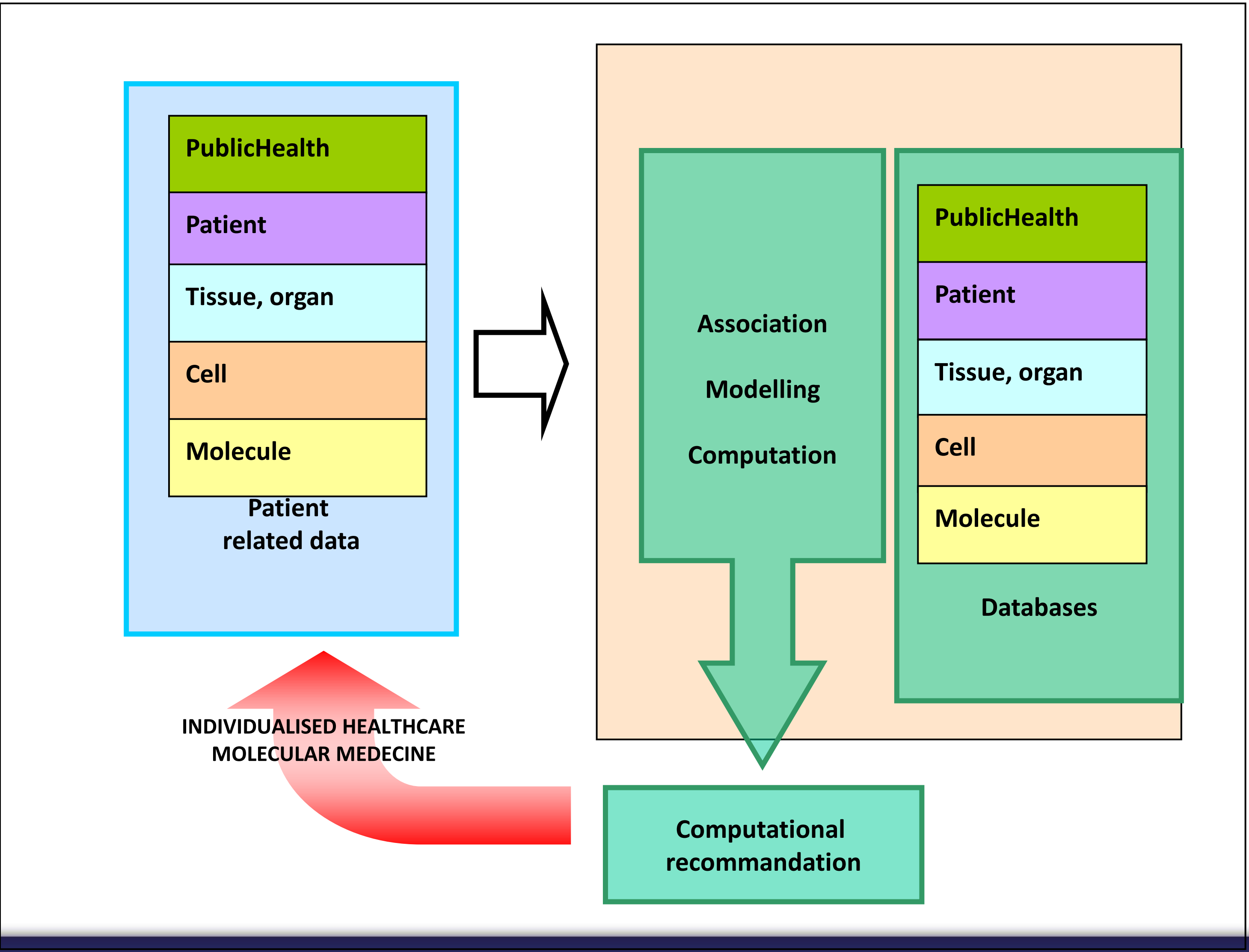
Historical perspective

- Three stages for life sciences
  - – Pioneering time : 2000-2005
  - – First successes : 2005-2010
  - – Plateau of maturity: 2010 - 2014

Pioneering time: manipulating concepts and deploying test applications

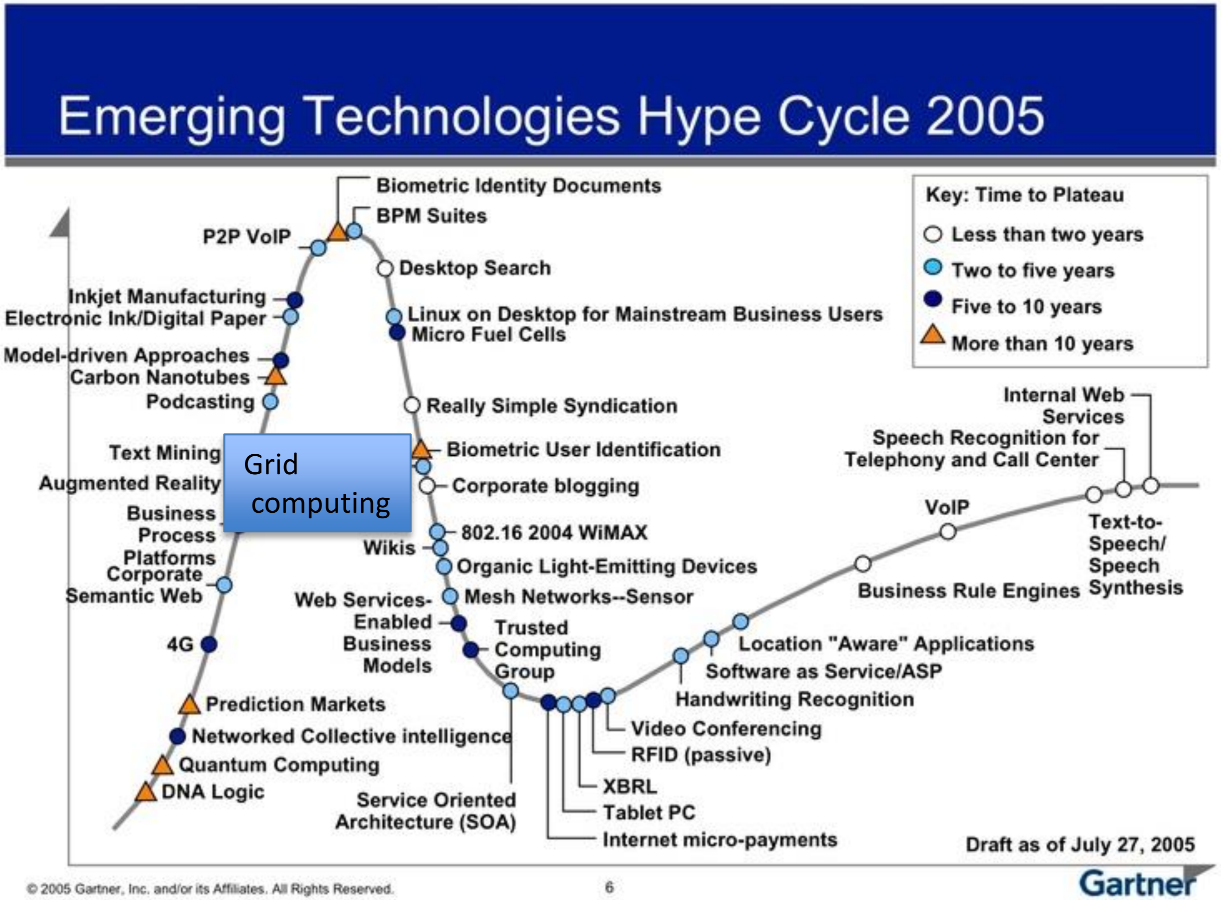


The challenges of tomorrow… in September 2002



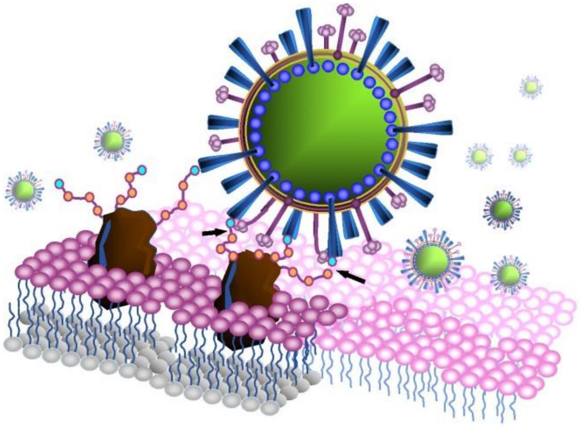
S. Norager
Y. Paindaveine
DG- INFSO

First successes (2005-2010)



WISDOM In silico Drug Discovery

- Goal: find new drugs for neglected and emerging diseases
  - Neglected diseases lack R&D
  - Emerging diseases require very rapid response time
- Need for an optimized environment
  - To achieve production in a limited time
  - To optimize performances
- Method: grid-enabled virtual docking
  - Cheaper than in vitro tests
  - Faster than *in vitro* tests



WISDOM, a highly successful drug discovery initiative on grids

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Wisdom-I Malaria Plasmepsin	DataChallenge Avian Flu Neuraminidase	Wisdom-II Malaria 4 targets	DataChallenge Diabetes Alpha-amylase, maltase	SARS 3C proteases					

GRIDS

EEGE, Auvergrid, TwGrid, EELA, EuChina, OSG EuMedGrid

EUROPEAN PROJECTS

Embrace
EGEE
BioInfoGrid

INSTITUTES

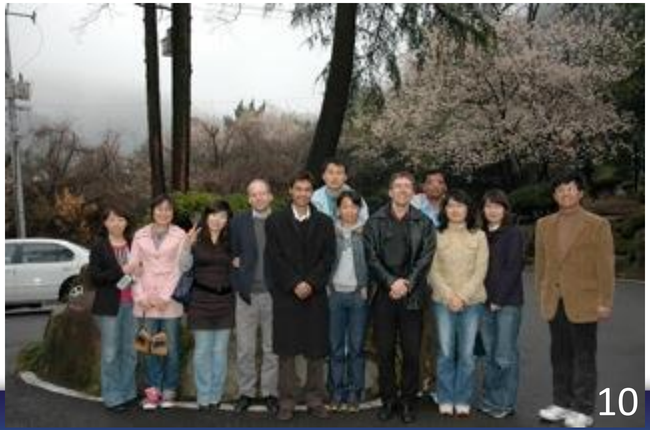
SCAI, CNU, Academica Sinica of Taiwan, ITB, Unimo Univ,, LPC, CMBA ,CERN-Arda, Healthgrid, KISTI

- New scientific applications
- New infrastructures and tools (Cloud, Supercomputers)
- Performance optimization

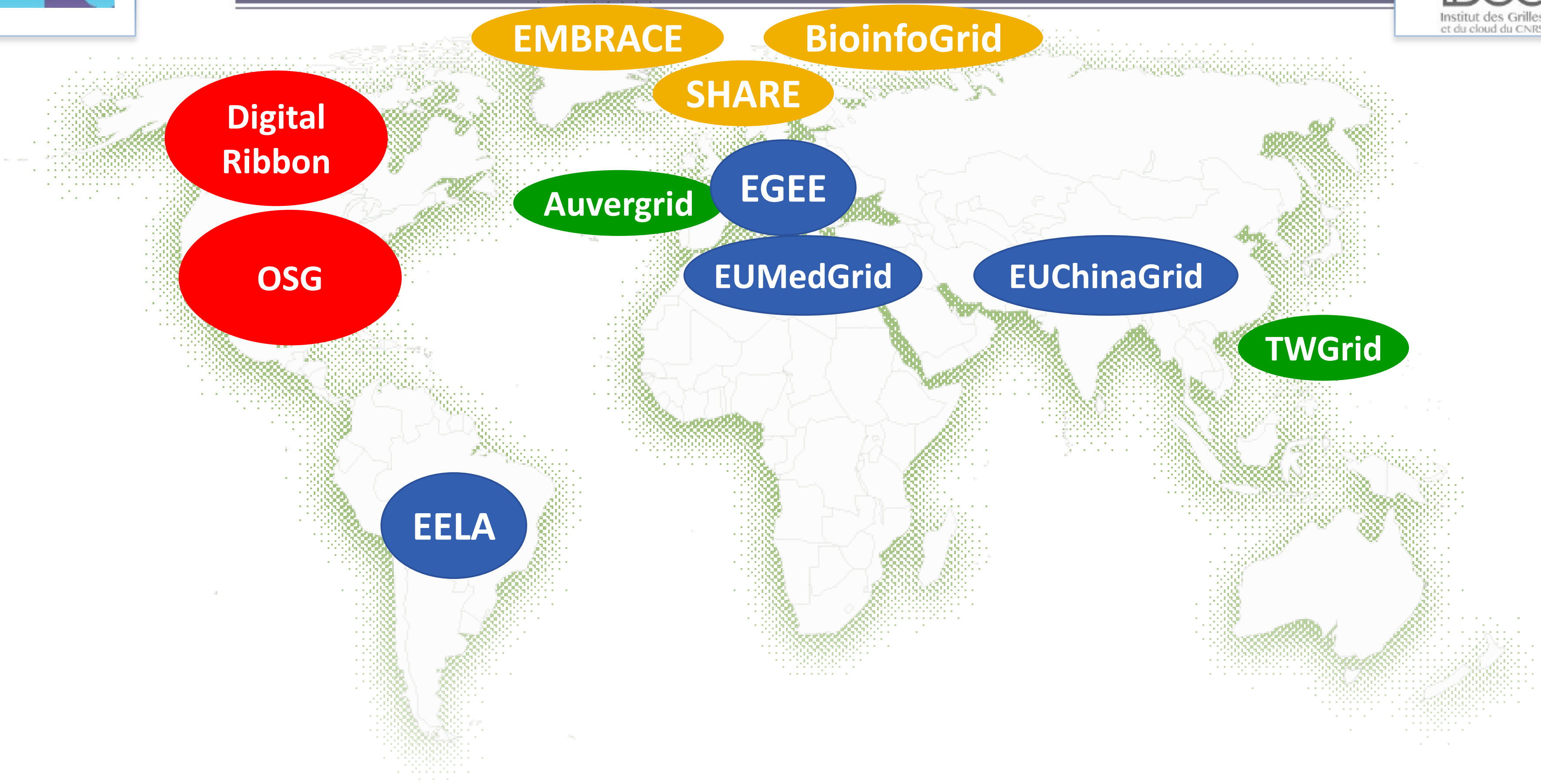
More than 15 papers in peer-reviewed scientific journals
5 patents on potential drugs against diabetes, malaria and SARS

What made WISDOM successful?

- The support of all grid infrastructures
  - As much CPU as needed: one century of CPU time as early as 2005
- The WISDOM Production Environment (Jean Salzemann)
  - First generation platform to push jobs on the grid
- The interest of Doman Kim and his team at Chonnam National University for testing *in vitro* the compounds selected *in silico*



Grid infrastructures and projects contributing to WISDOM



: EC funded grid infrastructure
: EC funded grid project
: Regional/national grid infrastructure
: US grid project

An unprecedented deployment on grid infrastructures

RESULTS ALREADY ACHIEVED IN 2009	
Number of docked compounds	> 150 million
Duration of the experience	2 months
Throughput of the experience	80,000/hour
Estimated duration on 1 PC	>400 years
Maximum number of computers	> 3000
Number of countries giving computers	27
Volume of data produced	1.6 TB

WISDOM received invaluable support from BioSolveIT, who has provided more than 3,000 free licenses for their commercial docking program FlexX.

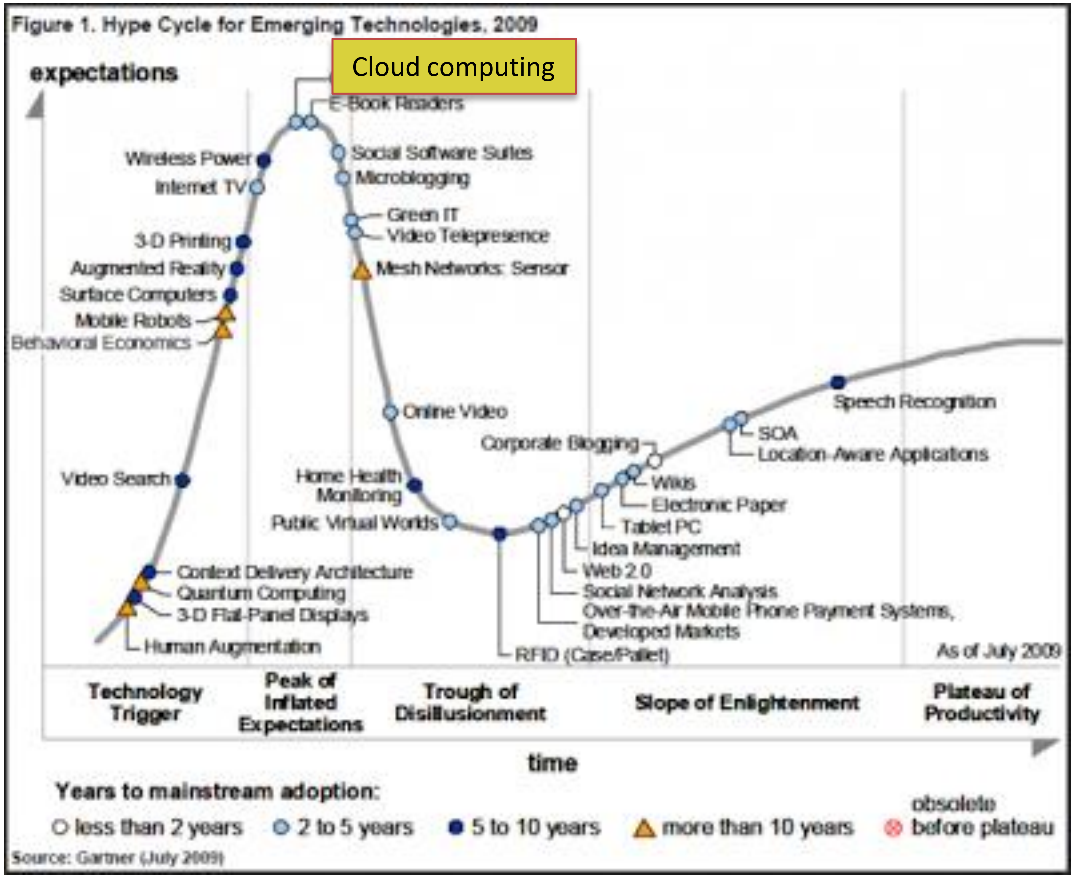
WISDOM: achievements and limitations

Grid added value	Grid limitations
Very large scale deployment : > 1 millenium of computation over 5 years	Security issues
	Grid fault tolerance (>30% failure rate)

What worked	What failed
In silico discovery of new active compounds against malaria, diabetes and SARS	Successful deployment of a virtual screening service
International deployment	Adoption by pharma

Grid infrastructures are excellent environments for in silico drug discovery but pharmaceutical laboratories are too concerned by IP issues to ever use them

Grid usage on the plateau of maturity (2010 -)

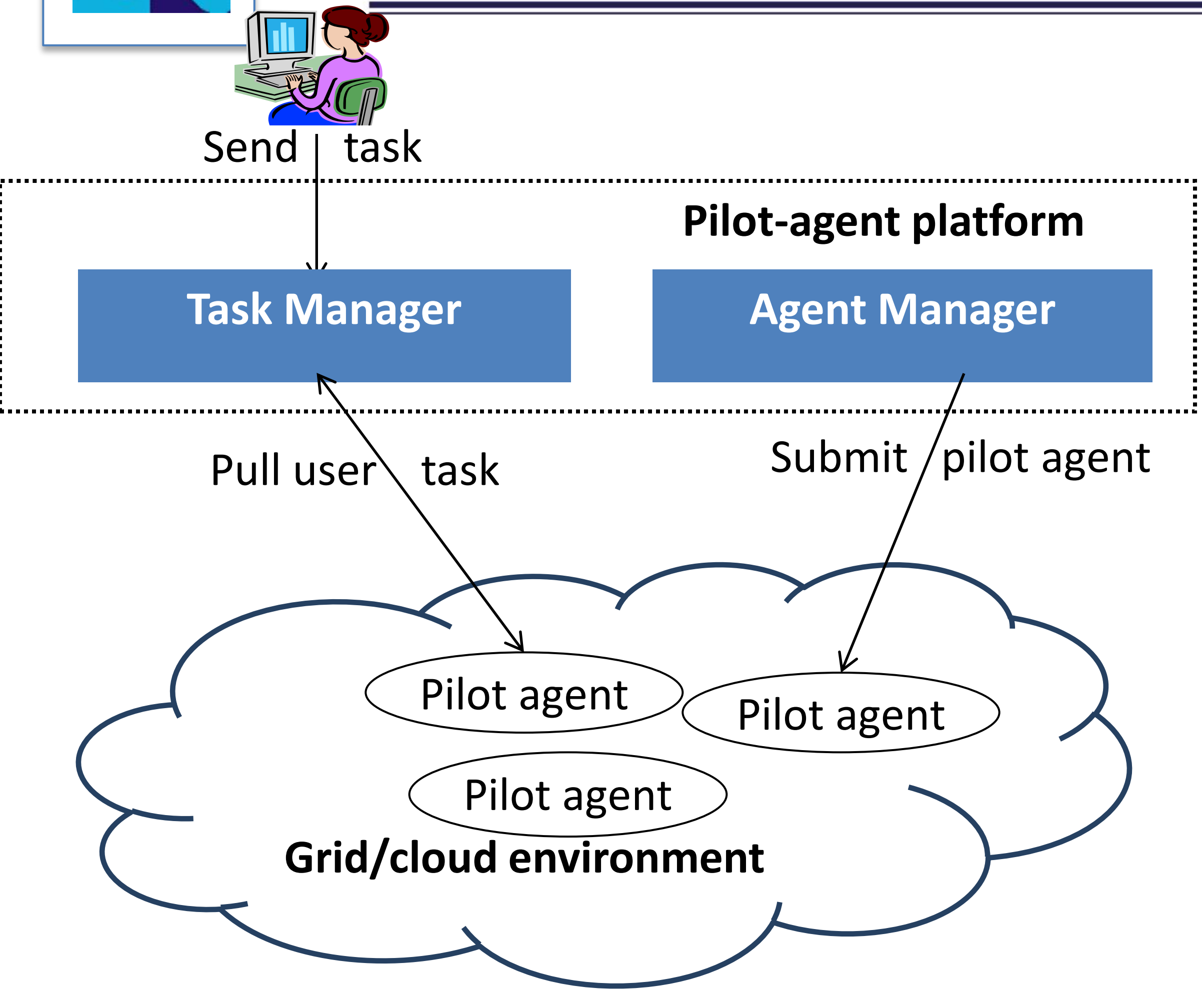


Grids had already disappeared from Gardner hype cycle for emerging technologies in 2009

What did change around 2010 (from a user point of view)?

Positive	Negative
Grid infrastructure became production quality for LHC data analysis	Pressure on resources considerably increased
Emergence of platforms hiding grid limitations - in terms of failure rate - in terms of information systems	
Emergence of web portals hiding grid complexity - no need for a certificate - “transparent” grid usage	Security ?

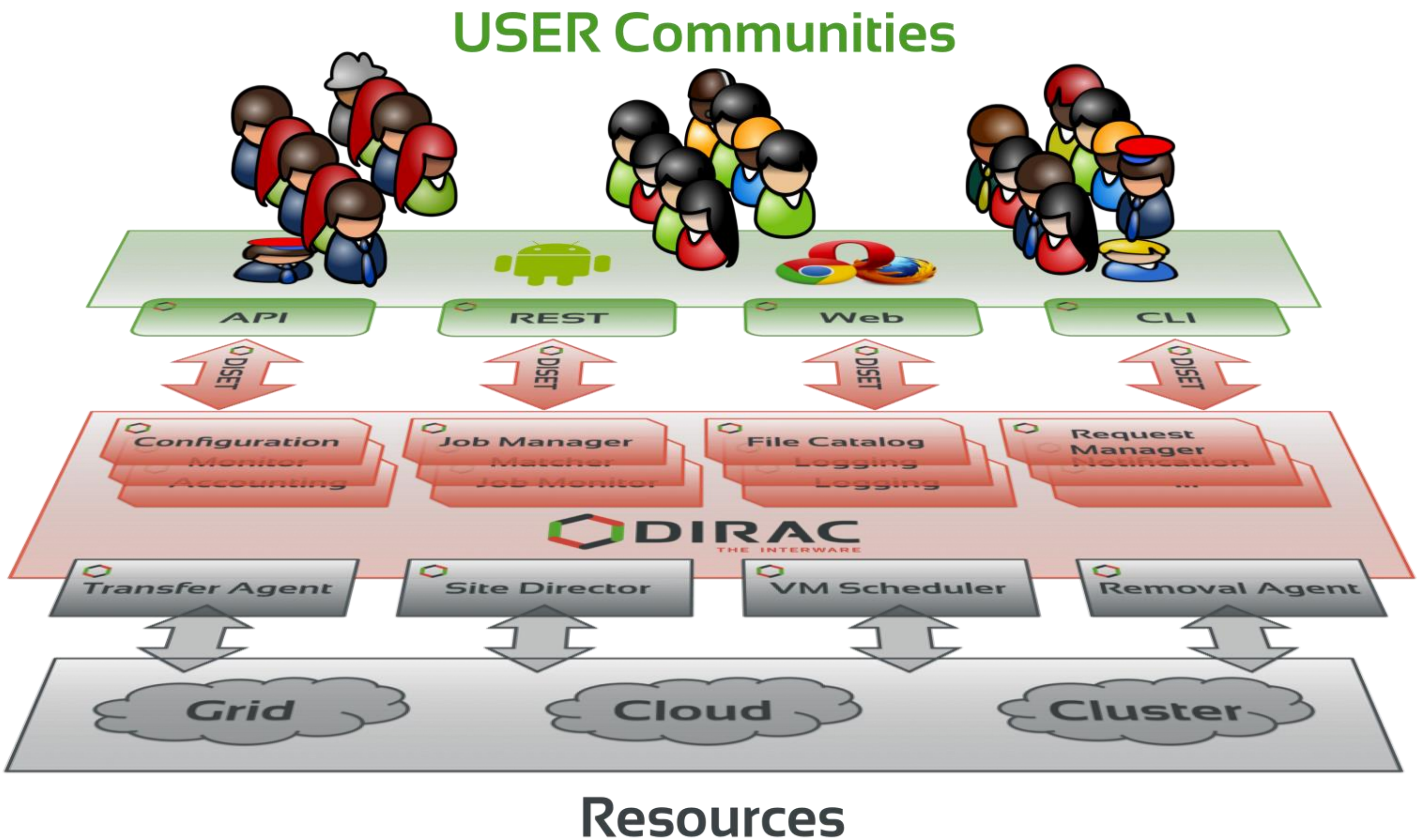
The winning strategy for grid users: pilot agent platforms



- Users submit their docking tasks to a central pool
- Pilot jobs are submitted to the grid and pull user tasks from the central pool
- Tasks in central pool are pulled according to a scheduling policy

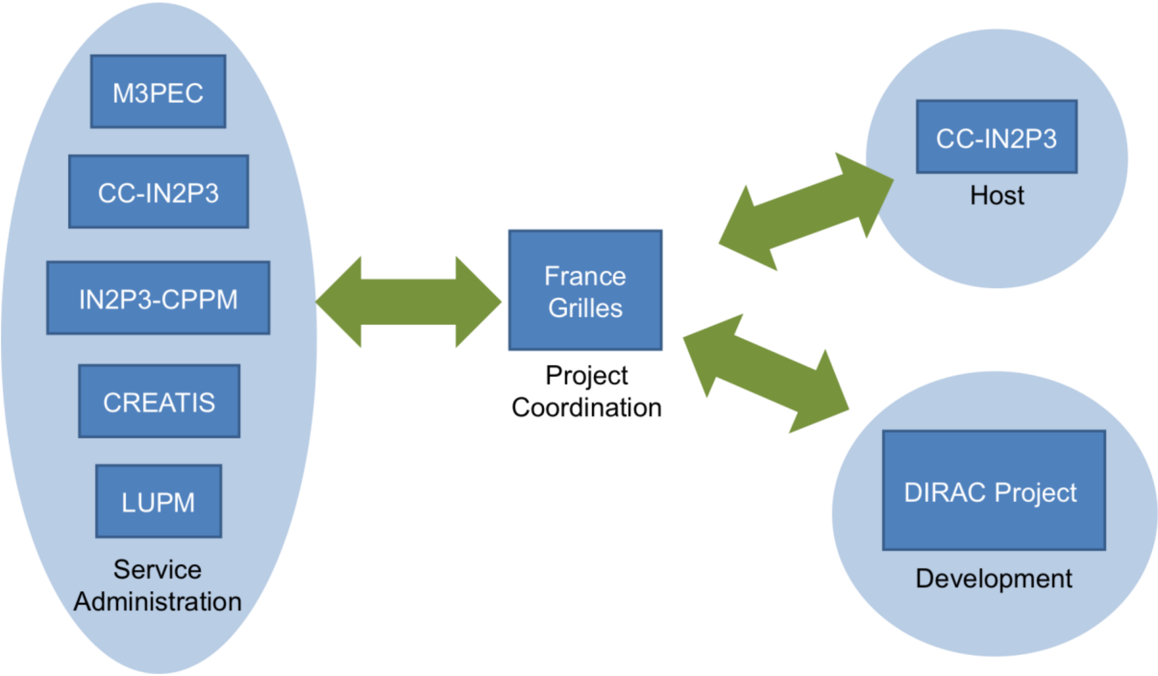
DIRAC

- A pilot agent platform developed for LHCb, now widely adopted



The France Grilles DIRAC service

- Hosted by the CC/IN2P3
- Distributed administrator team
  - 5 participating universities
- 18 VOs, ~100 registered users
- In production since May 2012
  - > 7 millions jobs



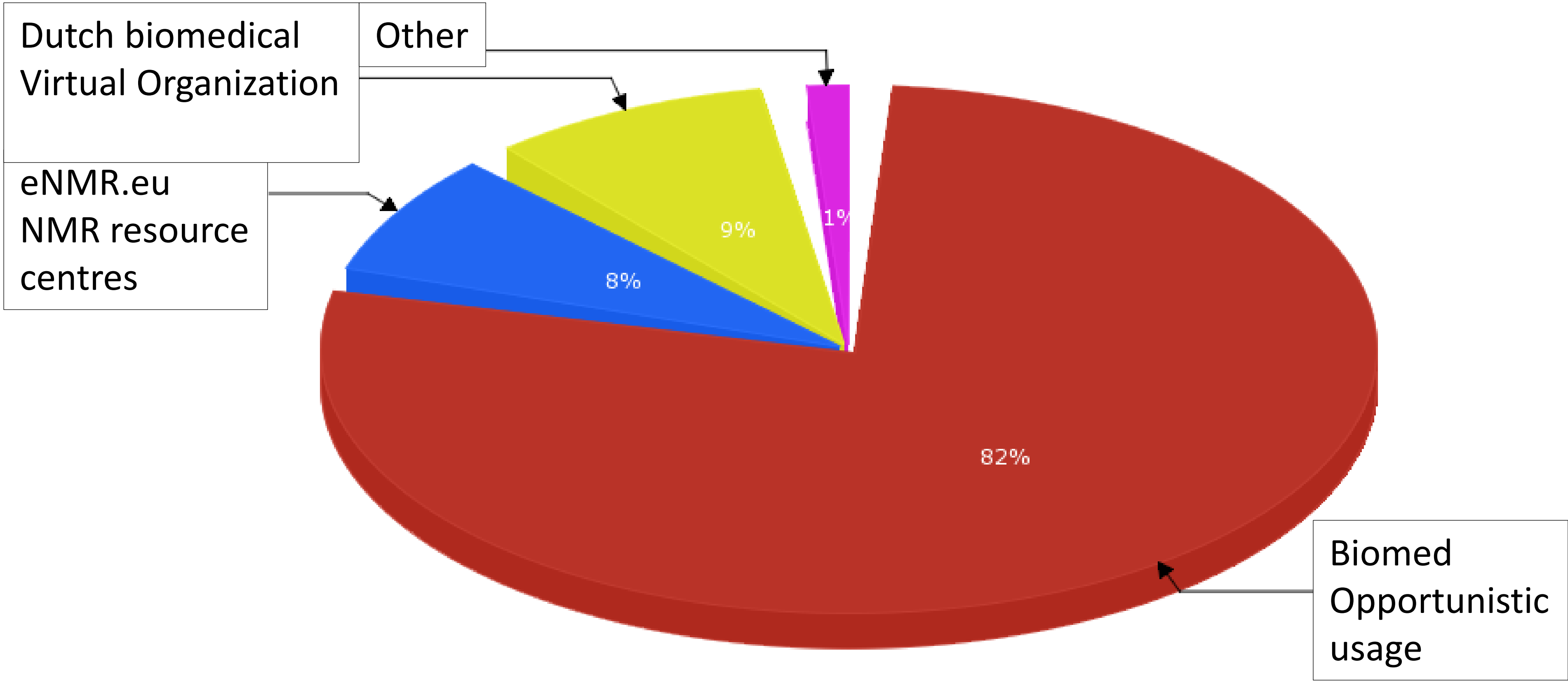
How is the grid used today?

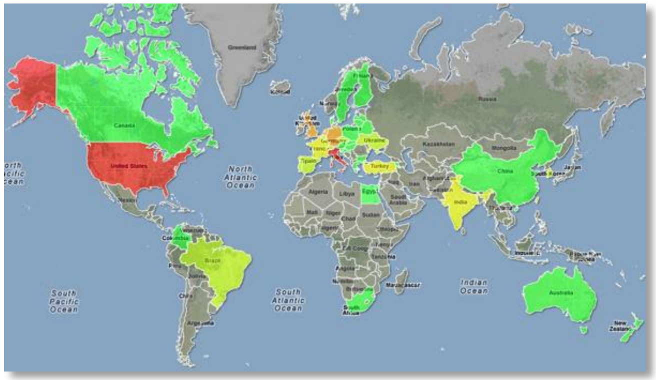
- Access to resources
  - Dedicated Virtual Organizations providing their own resources
    - We-NMR for structural biology
    - N4U for neurosciences
  - catch-all Virtual Organizations for all life sciences with opportunistic usage
    - International: Biomed Virtual Organization
- User friendly user interfaces
  - Science gateways with hundreds of users
  - Pilot agent platforms integrated into the gateways

Virtual Organization = dynamic set of individuals or institutions defined around a set of resource-sharing rules and conditions

Opportunistic usage is still dominant

Distribution of the normalized CPU-time per Virtual Organization in the life sciences



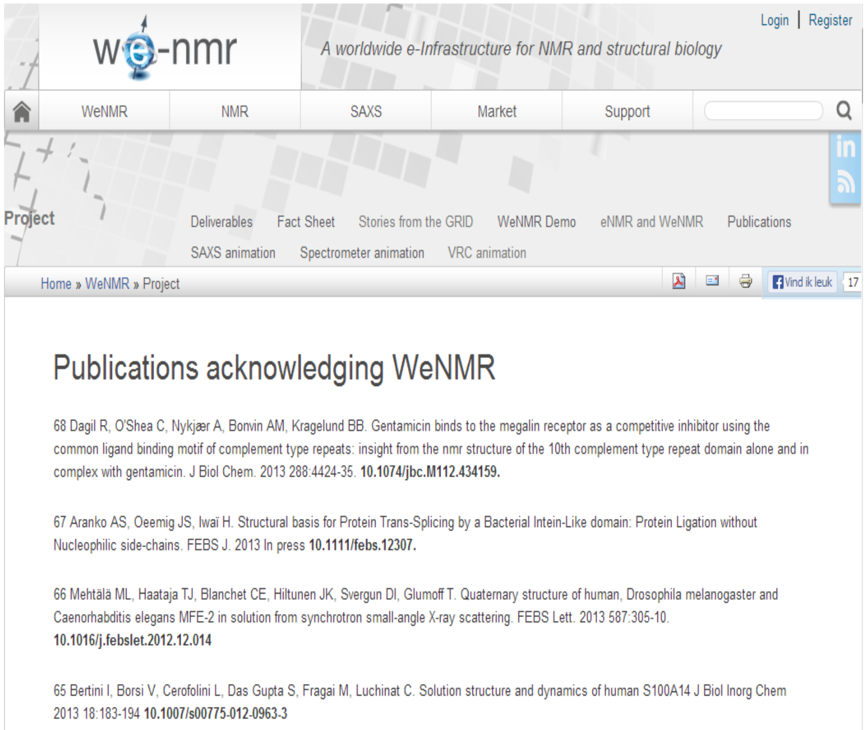


WeNMR VRC (Sept. 2013)

- Largest VO in the life sciences
- > 575 registered users (35% outside EU)
- ~ 90 000 CPU cores via EGI resources
- > 4.7M CPU hours over the last 12 months
- > 1.8 million jobs over the last 12 months
- User-friendly access to Grid via web portals

www.wenmr.eu

68 publications since 2011 acknowledging WeNMR (or eNMR)


Login | Register

we-nmr

A worldwide e-Infrastructure for NMR and structural biology

WeNMR | NMR | SAXS | Market | Support

Project

Deliverables | Fact Sheet | Stories from the GRID | WeNMR Demo | eNMR and WeNMR | Publications

SAXS animation | Spectrometer animation | VRC animation

Home » WeNMR » Project

Vind ik leuk 17

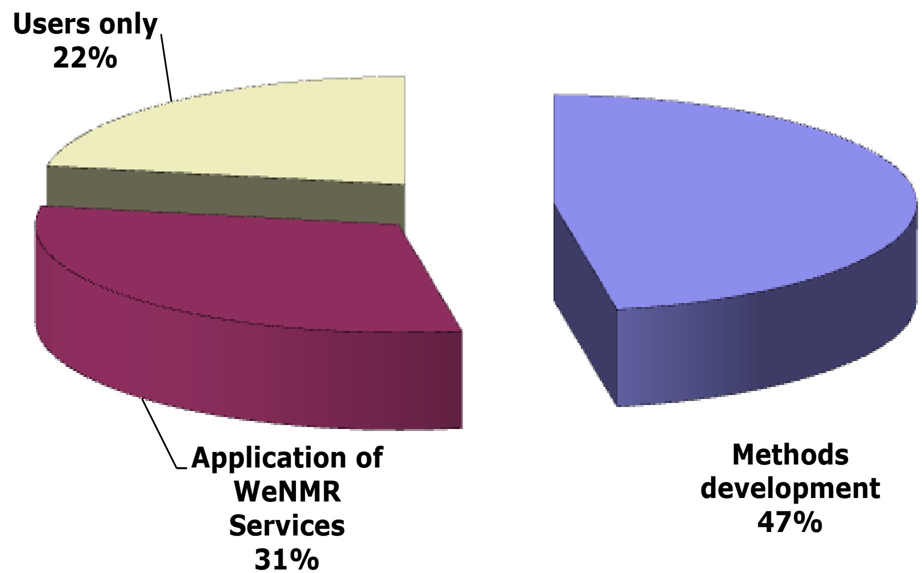
Publications acknowledging WeNMR

68 Dagil R, O'Shea C, Nykjær A, Bonvin AM, Kragelund BB. Gentamicin binds to the megalin receptor as a competitive inhibitor using the common ligand binding motif of complement type repeats: insight from the nmr structure of the 10th complement type repeat domain alone and in complex with gentamicin. J Biol Chem. 2013 288:4424-35. **10.1074/jbc.M112.434159.**

67 Aranko AS, Oeemig JS, Iwaï H. Structural basis for Protein Trans-Splicing by a Bacterial Intein-Like domain: Protein Ligation without Nucleophilic side-chains. FEBS J. 2013 In press **10.1111/febs.12307.**

66 Mehtälä ML, Haataja TJ, Blanchet CE, Hiltunen JK, Svergun DI, Glumoff T. Quaternary structure of human, Drosophila melanogaster and Caenorhabditis elegans MFE-2 in solution from synchrotron small-angle X-ray scattering. FEBS Lett. 2013 587:305-10. **10.1016/j.febslet.2012.12.014**

65 Bertini I, Borsi V, Cerofolini L, Das Gupta S, Fragai M, Luchinat C. Solution structure and dynamics of human S100A14 J Biol Inorg Chem 2013 18:183-194 **10.1007/s00775-012-0963-3**

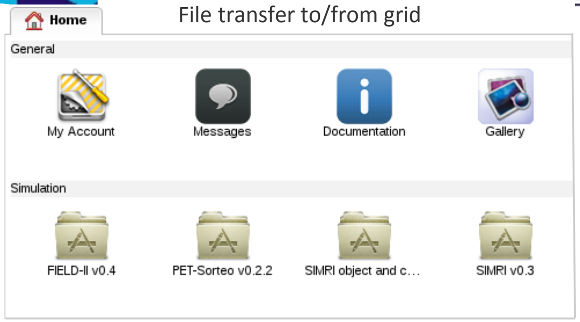


Application of WeNMR services
= collaborations between WeNMR staff and users

Virtual Imaging Platform

http://www.creatis.insa-lyon.fr/vip

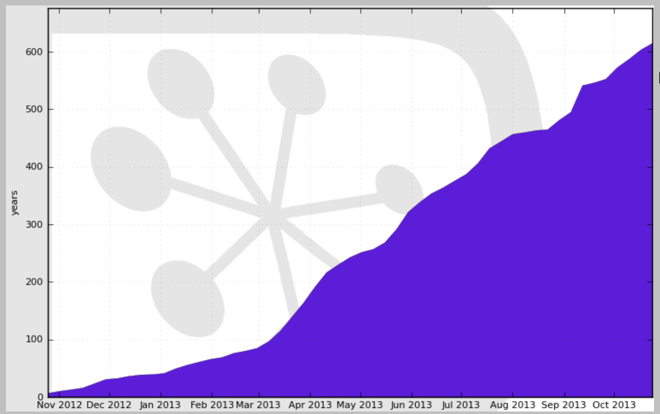
Application as a service
File transfer to/from grid



Web portal

Infrastructure

Supported by EGI Infrastructure
Uses biomed VO (most used EGI VO for life sciences in 2013)
VIP accounts for ~**25% of biomed's activity**
VIP consumes ~**50 CPU years** every month

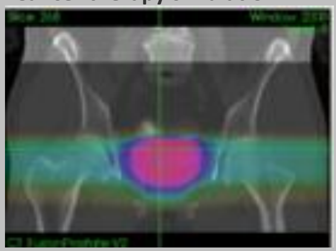


France-Grilles

DIRAC

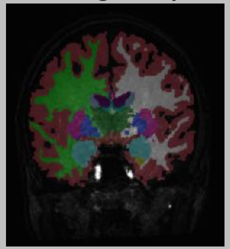
Scientific applications

Cancer therapy simulation



Prostate radiotherapy plan simulated with GATE(L. Grevillot and D. Sarrut)

Neuro-image analysis

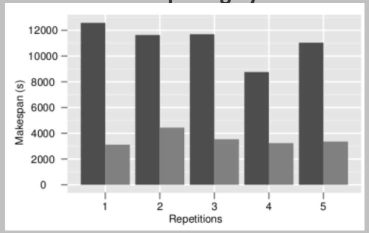


Brain tissue segmentation with Freesurfer

Image simulation

Echocardiography simulated with FIELD-II (O. Bernard *et al*)

Modeling and optimization of distributed computing systems

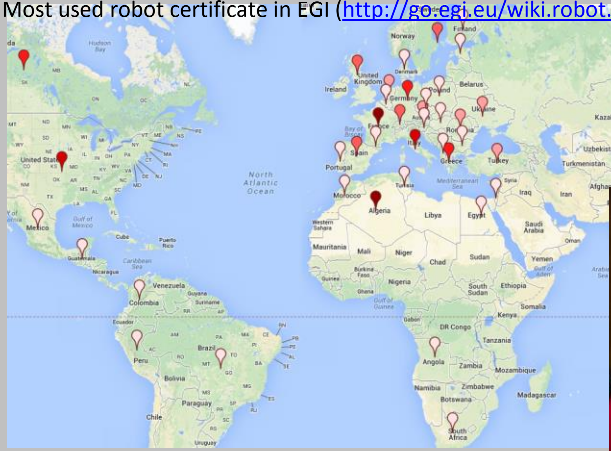


Acceleration yielded by non-clairvoyant task replication (R. Ferreira da Silva *et al*)

Users

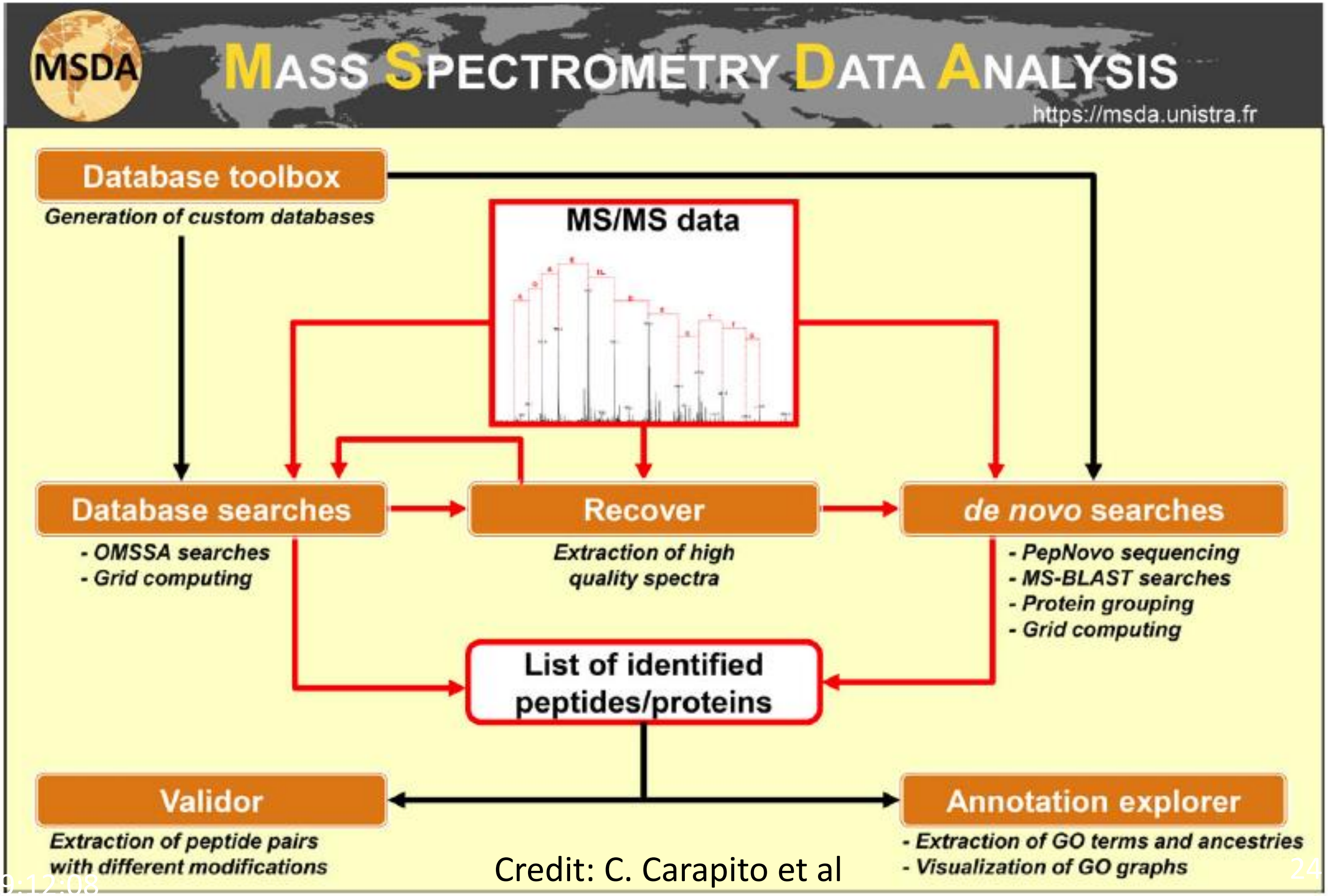
479 registered users in Nov 2013 (**175** in France)

Most used robot certificate in EGI (http://go.egi.eu/wiki.robot.users)



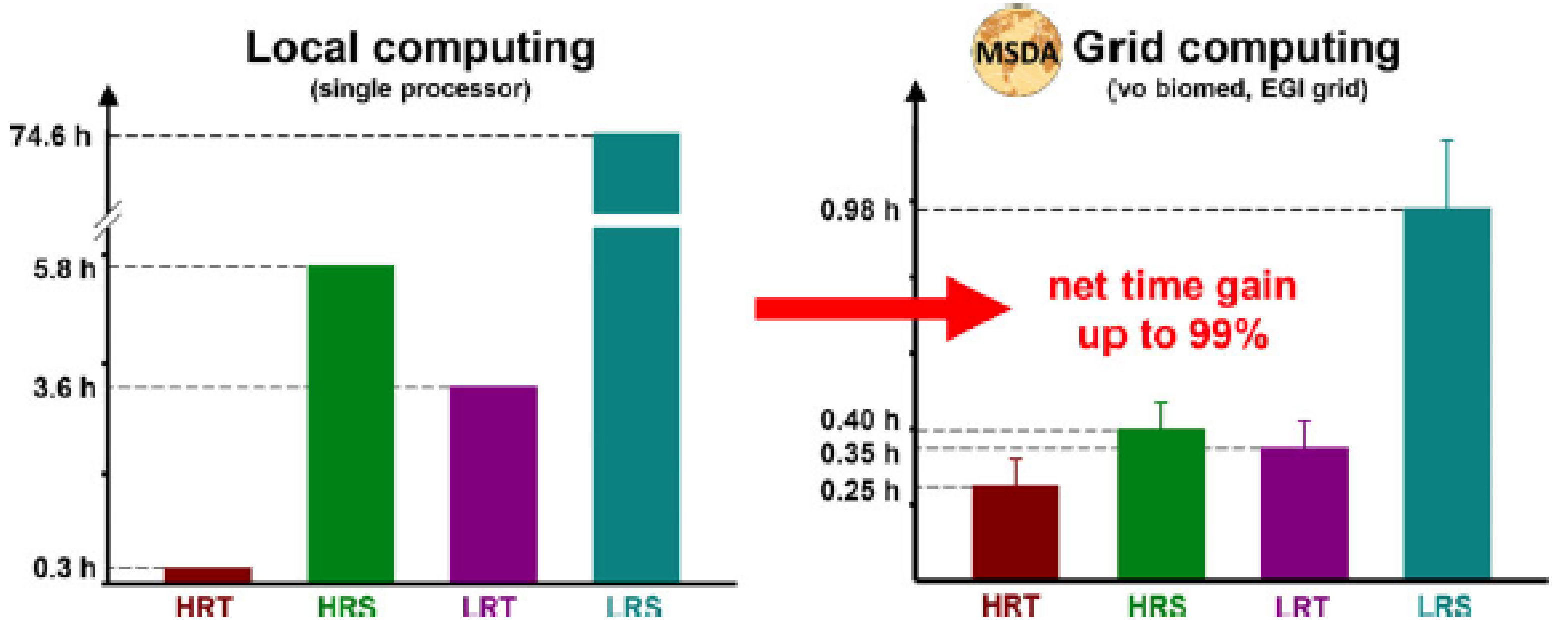
09:12:08

MSDA portal for Mass Spectrometry data analysis



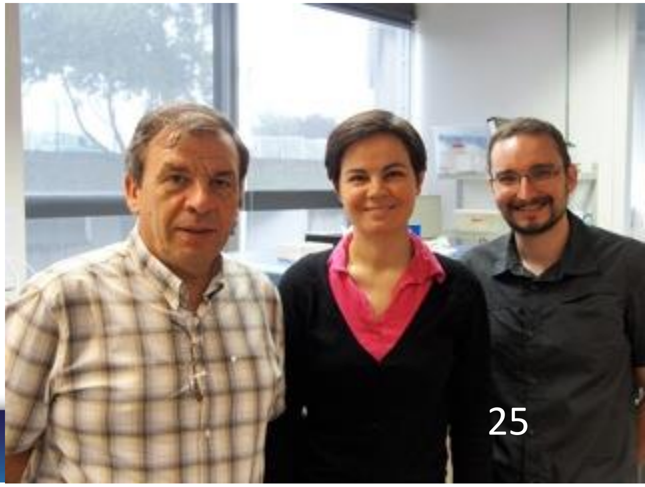
Credit: C. Carapito et al

Grid performances



Processing times of four typical shotgun proteomics datasets using a local laboratory computer versus grid computing on EGI

https://msda.unistra.fr



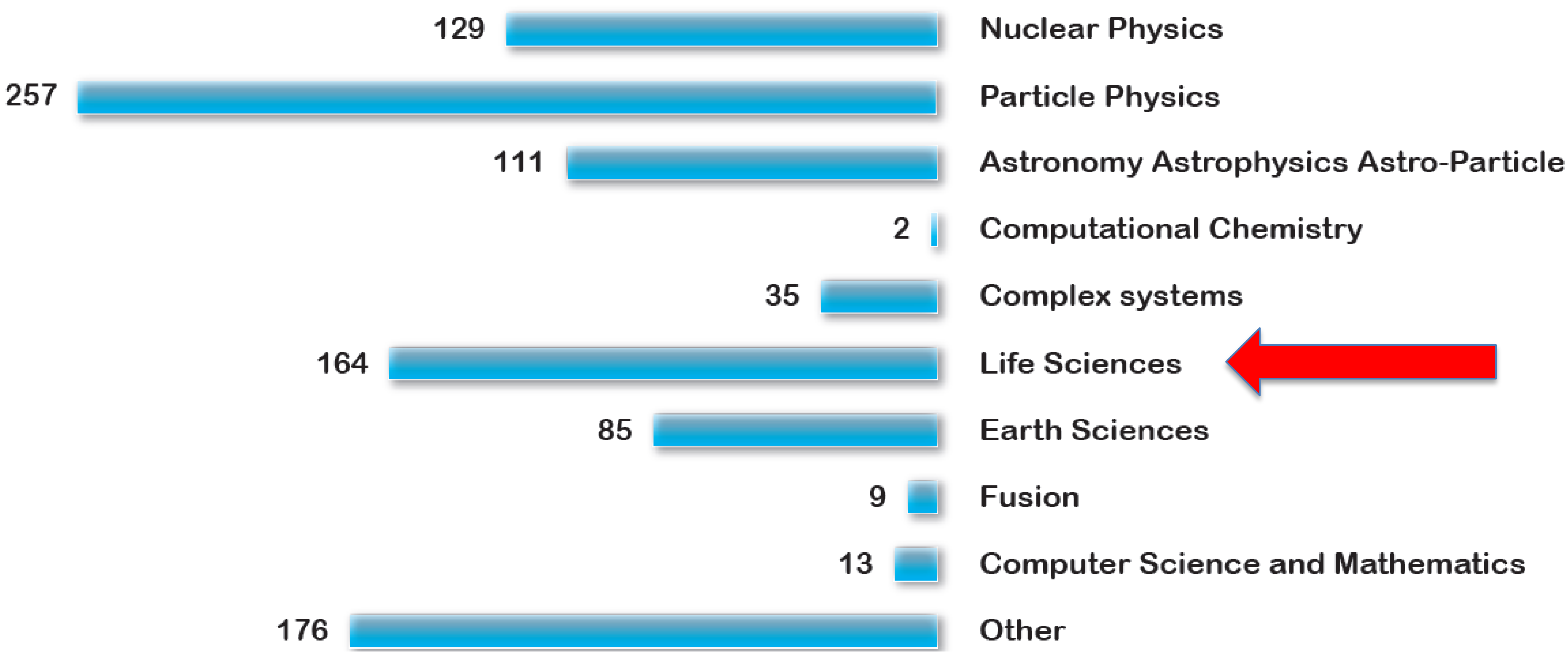
On the plateau of maturity: working on EGI takes from zero to three steps

- Get a certificate from a national Certificate Authority
  - Step not needed if you access the grid through a scientific gateway
- Learn how to use a platform (DIRAC)
  - Step not needed if you access the grid through a scientific gateway
- Access services like FG-Dirac or EGI-DIRAC
  - Open to the “long tail” of science
  - Not needed if you access the grid through a scientific gateway

Maturity: truly multidisciplinary

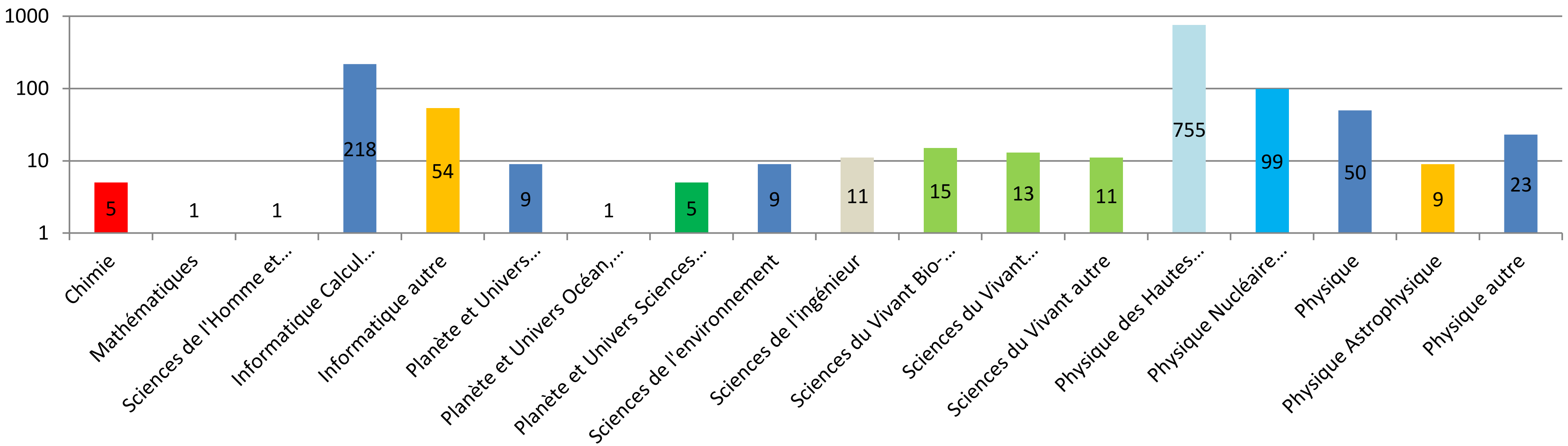
User communities

Owners of certificates delivered by the French Certificate authority in the last 12 months



Maturity: very significant scientific production

Over 1200 scientific publications co-signed by French researchers june 2010 – April 2014



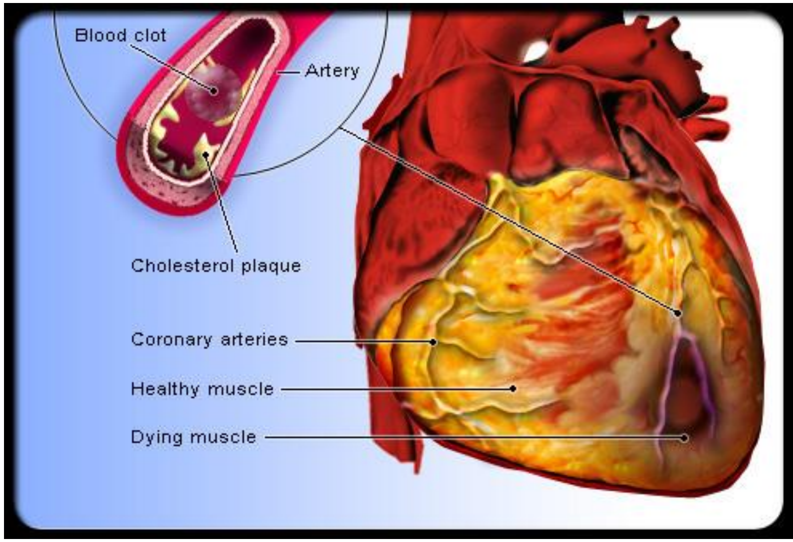
What about molecular biology?

- Early involvement
- Limited impact
  - Technical issues
  - Political issues
- Some success stories

	gLite
Banques internationales	~ oui
Espace personnel	~ oui
Espace commun	~ oui
Accès simple au stockage	non
Distribution des calculs	WMS
Intégration cluster l'existant	~ oui
Déploiement des logiciels	SWAREA
Workflow/pipeline	~ DAG
Gestion des identités et accès	vo.renabi.fr
Interface facile à utiliser	~ CLI
Interface publique: accès anonyme sur portail et web services	non

Genome Wide Haplotype analyses of human complex diseases with the EGEE grid

- Goal: study the impact of DNA mutations on human coronary diseases
- Very CPU demanding analysis to study the impact of correlated (double, triple) DNA mutations
- Deployment on EGEE Grid
  - 1926 CAD (Coronary Artery Diseases) patients & 2938 healthy controls
  - 378,000 SNPs (Single Nucleon Polymorphisms = local DNA mutations)
  - 8.1 millions of combinations tested in less than 45 days (instead of more than 10 years on a single Pentium 4)
- Results published in *Nature Genetics March 2009* (D. Tregouet et al)
  - Major role of mutations on chromosome 6 was confirmed



Summary

Scientific subdiscipline	Achievements	Limitations
Structural biology	100s of users through scientific gateways	Grid operational cost
Drug discovery	Large scale deployment of docking computations	IP issues have stopped adoption
Medical imaging (simulation)	100s of users through scientific gateways	Grid operational cost
Neurosciences	Emergence of grid-enabled scientific gateways	Protection of medical data – grid operational cost
Molecular biology - bioinformatics	Limited adoption	Grid middleware OS – Data management – grid operational cost - RAM

Cloud computing provides new opportunities (flexibility, reduced operational cost)

Conclusion of session II

- Grid computing has allowed building a truly multidisciplinary distributed IT infrastructure
  - Life sciences have benefitted and are benefitting from it
  - Human network across scientific disciplines
- Cloud computing allows extending the grid functionalities
  - Life sciences will benefit even more