Clouds in biomedical sciences
Part III – clouds in biosciences

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

Session III: clouds in life sciences

- Generalities
- Deployment of life science applications on public clouds
- "De novo" deployment deployment of scientific applications on academic clouds
- Pilot jobs platform help hiding technical difficulties
  - examples

Summary of grid adoption in life sciences

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

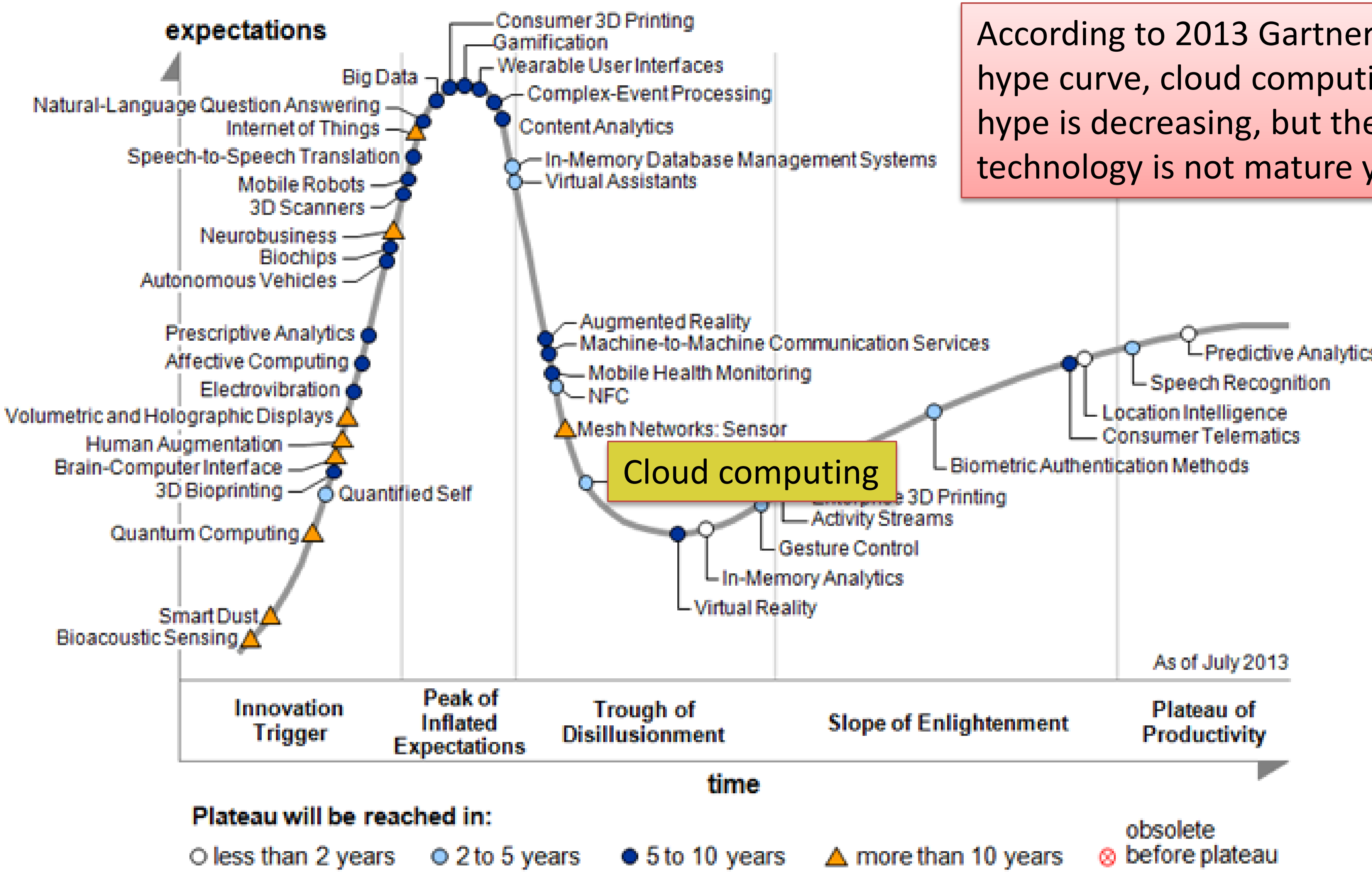
Cloud computing provides new opportunities (flexibility, reduced operational cost)

The promises of cloud computing

- Public clouds
  - No cost to operate IT infrastructure: only pay what you use
  - Computing capacity on demand
    - Unbound resources
  - Flexibility to upload favorite Operating System
- Academic (private) clouds
  - Reduced cost to operate IT infrastructure (compared to grid)
  - Flexibility to upload favorite Operating System

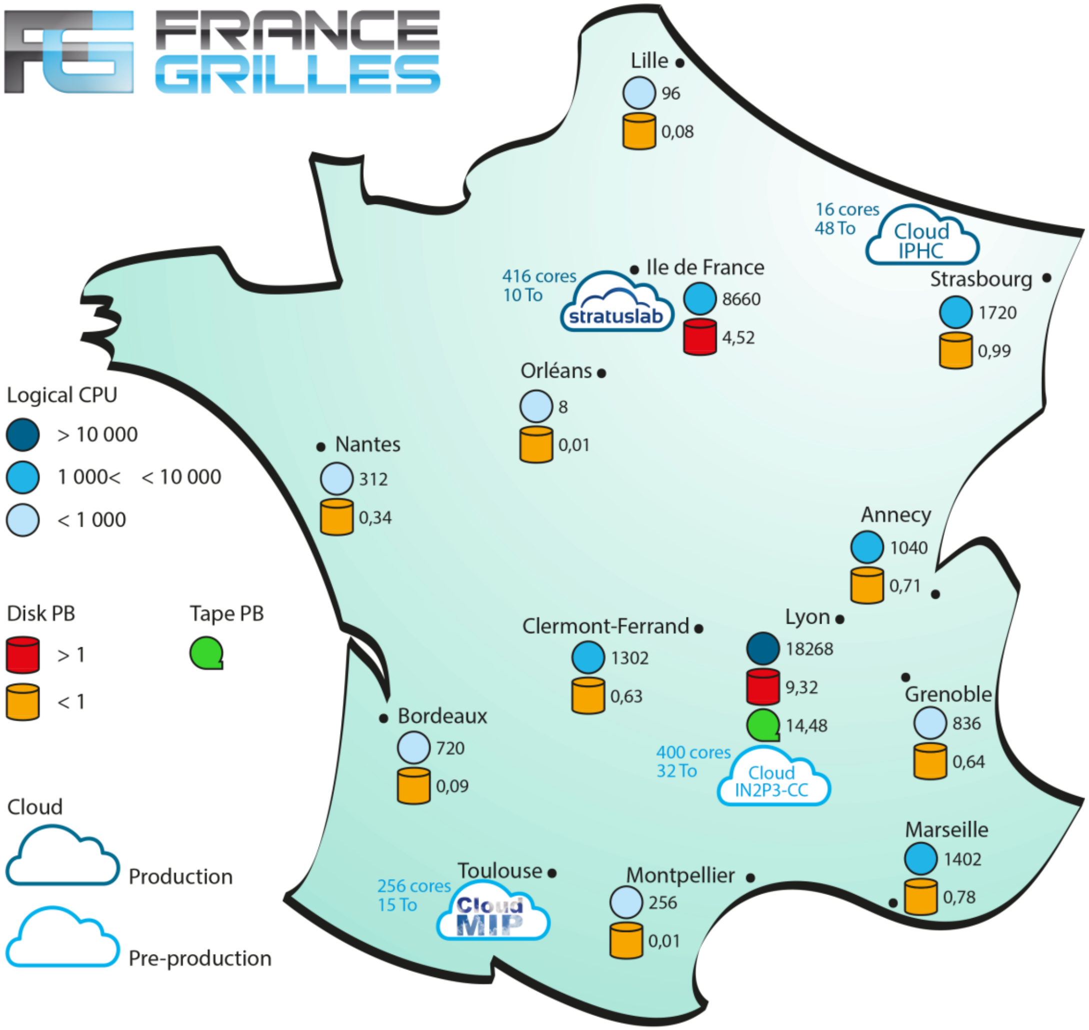
Where are we today?

According to 2013 Gartner hype curve, cloud computing hype is decreasing, but the technology is not mature yet



The situation in France

- French state put all cloud money in industry
- Federation of academic clouds started in 2012
  - OpenNebula
  - StratusLab
  - OpenStack



Adoption of clouds in the life sciences community in 2014 is very hard to assess

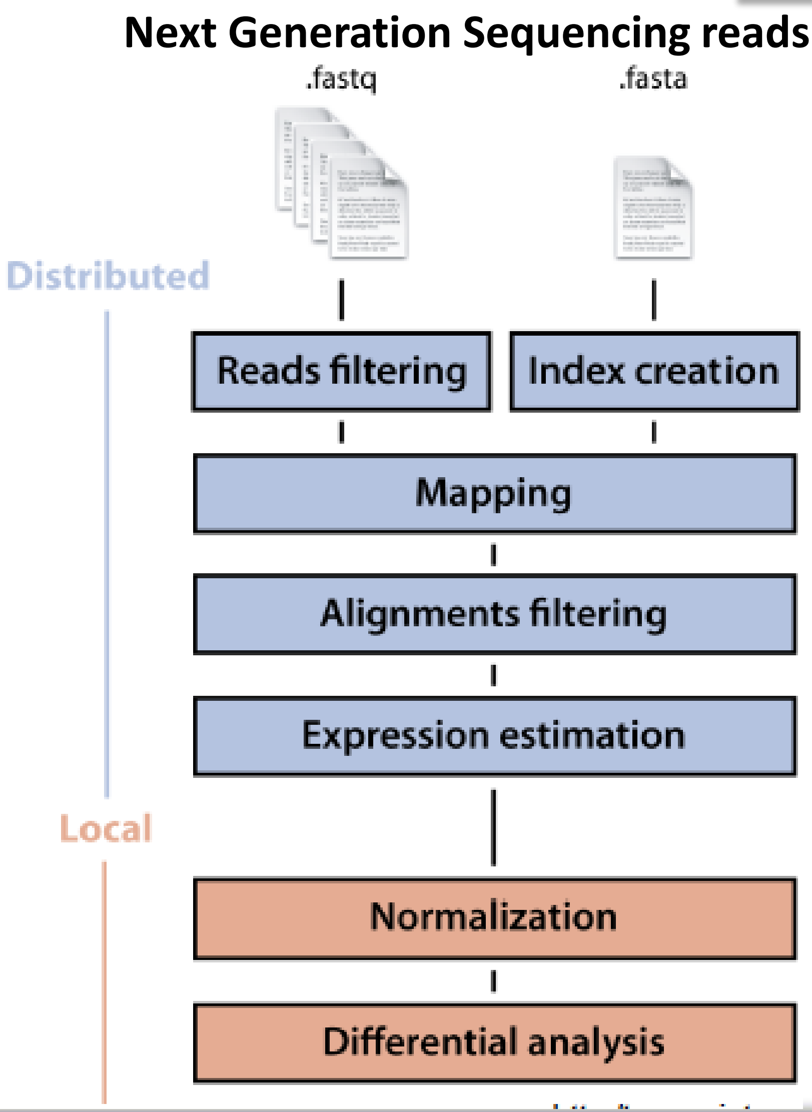
- Everything is now renamed cloud computing
  - Cluster computing
  - Grid computing
- Three scenarii:
  - Deployment of scientific applications on public clouds (Amazon)
  - De novo deployment of scientific applications on academic clouds
  - Migration to academic clouds of grid applications deployed using pilot agent platforms

Deployment of life science applications on public clouds

- Only a few research groups are using public clouds in France
  - Academic Research funding model is hardly compatible with credit card payment for computing capacity
- Feedback is not very positive
  - Public clouds perceived as expensive compared to academic clusters/grids

Eoulsan experience on AWS (Amazon)

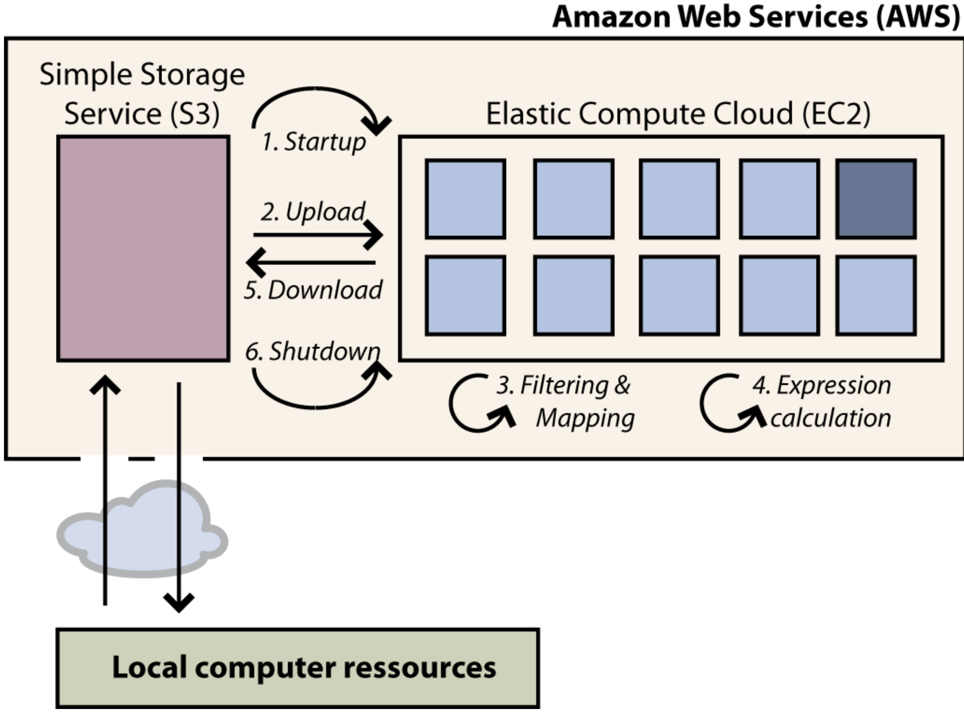
- Eoulsan is an analysis workflow of RNA-sequences
- Three steps:
  - Data upload (upload step)
  - Read mapping and filtering (filtermap step)
  - Transcript abundance estimation (expression step)
- Distributed calculations to speed up analysis
  - Parallelisation using Hadoop



Jourdren (2012) Bioinformatics Credit: Stéphane Le Crom

Eoulsan conclusion: Grid performances better than Amazon Web Service

- Comparison of Eoulsan running times (in minutes) between grid and Amazon cloud (AWS) for each analysis step
  - Human data
  - 888 Million reads corresponding to 88Gb data
- Conclusion: migration to EGI of the pipeline analysis



	Upload	filtermap	expression	Total
Standalone	154	1,146	4	1,304
Grid	53	388	2.5	467
AWS	80	810	64	1,120

Some considerations on public cloud storage prices

- Google Drive offer (⇔ external hard disk): 1$ per TeraOctet per month [1]
- Storage offers on commercial clouds: ≈ 300K$/PO/yr
  - Amazon S3[2] and Google[3] almost equivalent: ≈ 30$ per TeraOctet per month
  - Additional cost: billing of requests and data transfers
    - Amazon S3: 0,1 $ per GOctet of data transfered from S3 to internet (100K$/PO)
    - Google: ≈ 0,2 $ per GOctet of data transfered from S3 to internet (200K$/PO)

[1]: valid for 300 Toctets and above

[2]: http://aws.amazon.com/fr/s3/pricing/

[3]: https://cloud.google.com/products/cloud-storage/#pricing

- Ecclesiastes 1:9* The thing that hath been, it *is that* which shall be; and that which is done *is* that which shall be done: and *there is* no new *thing* under the sun.

Example: the e-Biothon initiative

- Telethon: every year, fund raising by french media for French Muscular Distrophy Association (AFM)
- From Telethon to Decrypthon
  - Computing infrastructure (IBM)
  - Research projects (CNRS)
  - Human resources (AFM)
- From Decrypthon to E-Biothon

E-Biothon: infrastructure

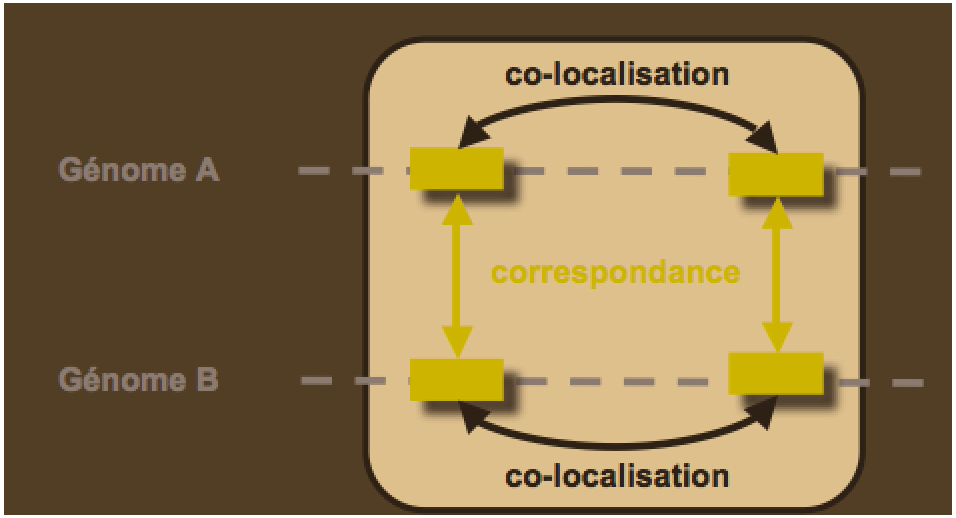
- 2 Blue Gene/P IBM racks with 200 TO storage
  - 2x1024 4-core nodes
  - up to 28 TFlops peak performance
- **SysFera-DS web access to computing resources**
- 2 modes:
  - Standard (MPI)
  - **HTC (1024 independent tasks in parallel)**

E-Biothon vision is to offer a service to the user communities in life sciences

- 2013-2014: first 3 projects
  - Jean-François Gibrat et al, (MIGALE platform, INRA Jouy-en-Josas)
  - Olivier Gascuel, Stéphane Guindon et Vincent Lefort (CNRS Montpellier)
  - Yec'han Laizet, Philippe Chaumeil, Jean-Marc Frigerio, Stéphanie Mariette, Sophie Gerber, Alain Franc (INRA BioGeCo – Bordeaux)
- > 2014: open call for projects (IFB)

Studying the synteny over a wide range of microbial genomes

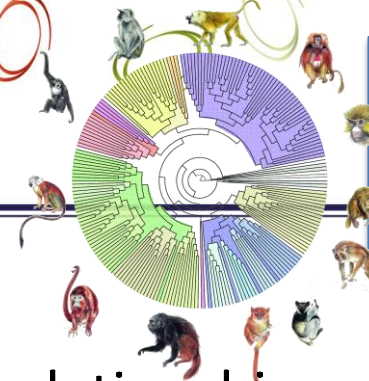
- Definition: similar blocks of genes in the same relative positions in the genome



- Interest: Study of synteny can show how the genome is cut and pasted in the course of evolution
- MIGALE team at INRA designed a pipeline analysis to compute synteny between 2 genomes and store it in a database
- **E-Biothon impact: change in scale - capacity to compute synteny between 2000 complete bacterial genomes (7 millions comparisons**)

Phylogeny.fr
Robust Phylogenetic Analysis For The Non-Specialist

PhyML



IDGC
Institut des Grilles et du cloud du CNRS

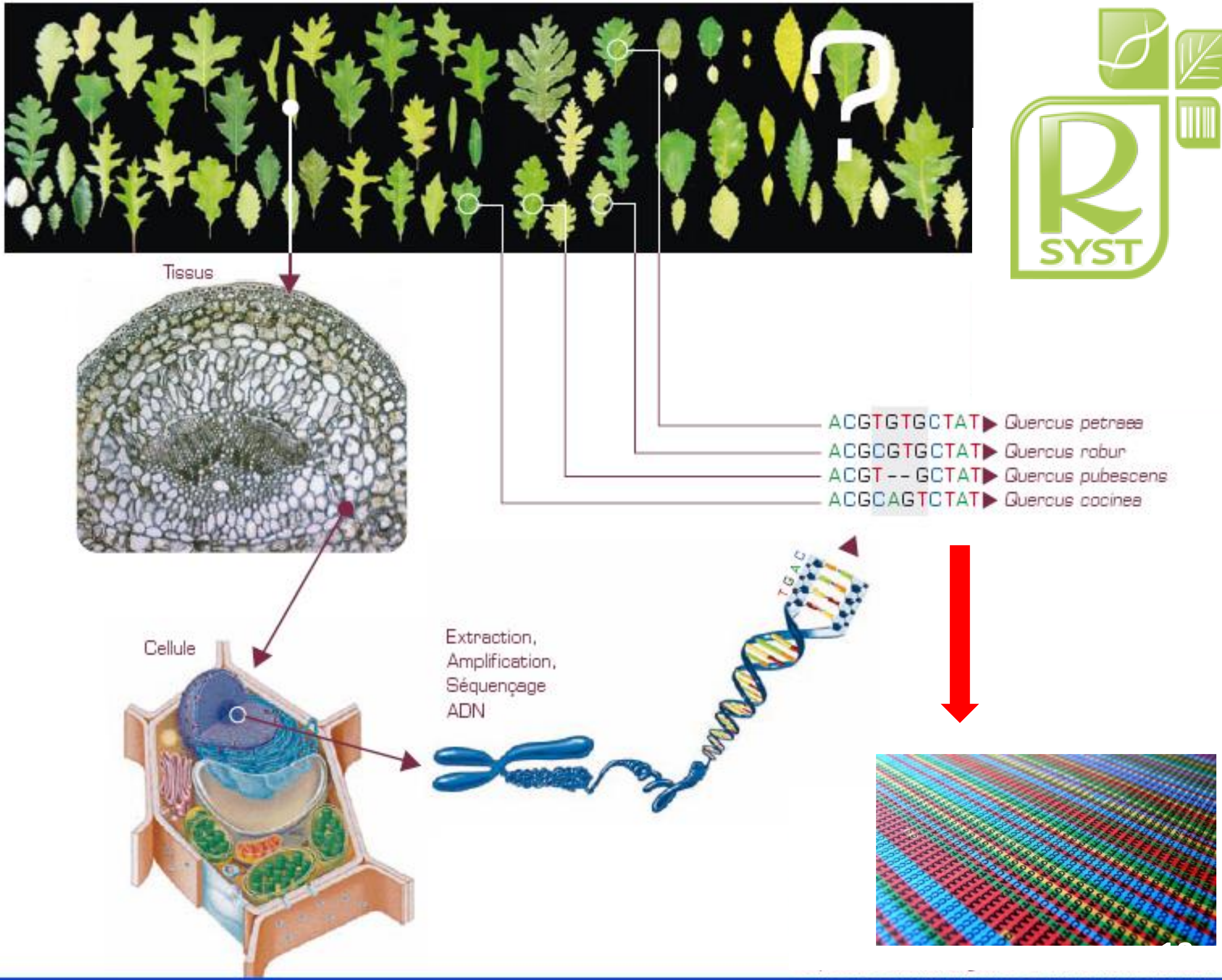
Philogenetics is the study of evolutionary relationships among groups of organisms

PhyML is a software that estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences

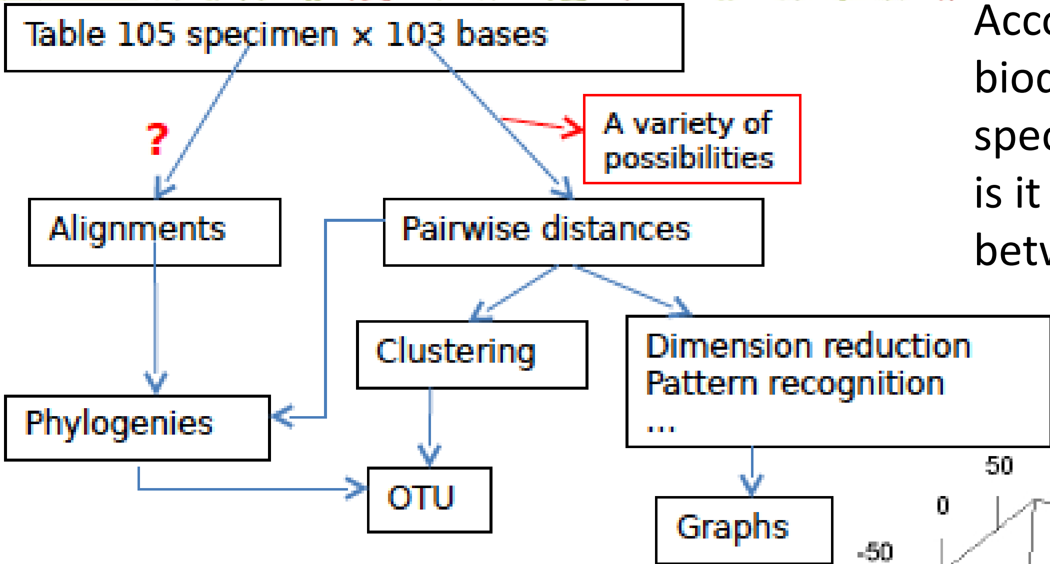
PhyML original publication in 2007 is the most cited in environment and ecology (> 6000 citations).

e-Biothon impact: change in scale in the resources made available to PhyML users

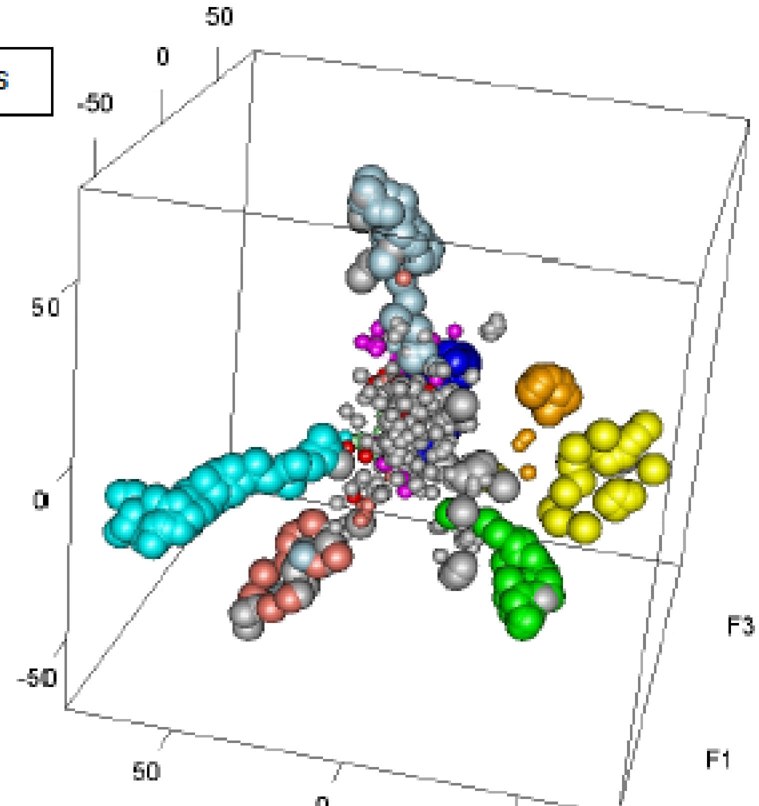
Characterizing biodiversity



09:15:44



According to botanic theory, biodiversity is organized in species, genders, families, orders: is it confirmed in the distance between sequences?



blue -> Mimosoideae

lightblue -> Lecythidaceae

cyan -> Chrysobalanaceae

green -> Annonaceae

lightgreen -> Caesalpinioideae

yellow -> Myrtaceae

orange -> Elaeocarpaceae

magenta -> Apocynaceae

salmon -> Burseraceae

red -> Malvaceae

~ 1000 individus

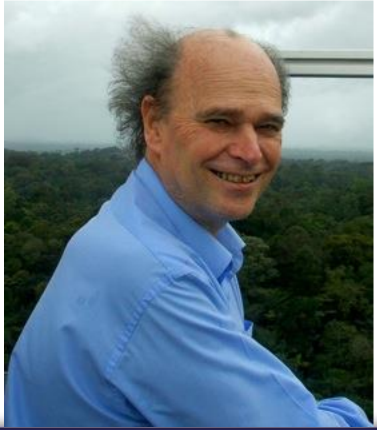
Study of biodiversity in Guyane

16000 different tree species in amazonian forest (≈ 300 in Europe)

More biodiversity in 10000 m^2 of forest in French Guyana than in Europe

E-Biothon added value

- Change in scale (from local Mesocenter in Bordeaux)
- Millions of reads
- Exact distance computation without heuristics (alignement scores)



Credit: Alain Franc et Yec'hran Laizet

Which global strategy for molecular biology ?

- Grid middleware and computing resources do not optimally fit the core needs of molecular biology
  - Genome assembly from Next Generation Sequencing raw data requires both RAM and large disk storage
  - Bioinformatics analysis requires much more flexibility than current grid infrastructures

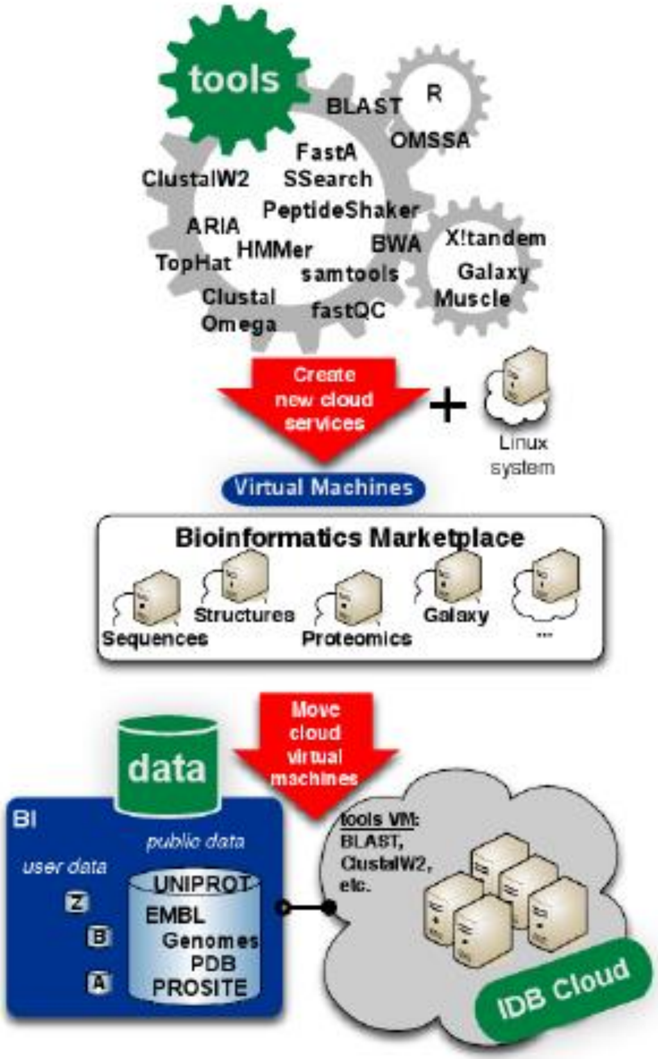
The french strategy for molecular biology

- France Genomique: an infrastructure to strengthen french capacities for High Throughput genomics
  - Central resource: HPC computing and storage resources @ TGCC (CEA)
- Institut Français de Bioinformatique: an infrastructure for the management and analysis of biological data
  - Central resource: academic cloud @ IDRIS
  - French node of ELIXIR, the European Research Infrastructure for Molecular Biology

France Genomique @ TGCC

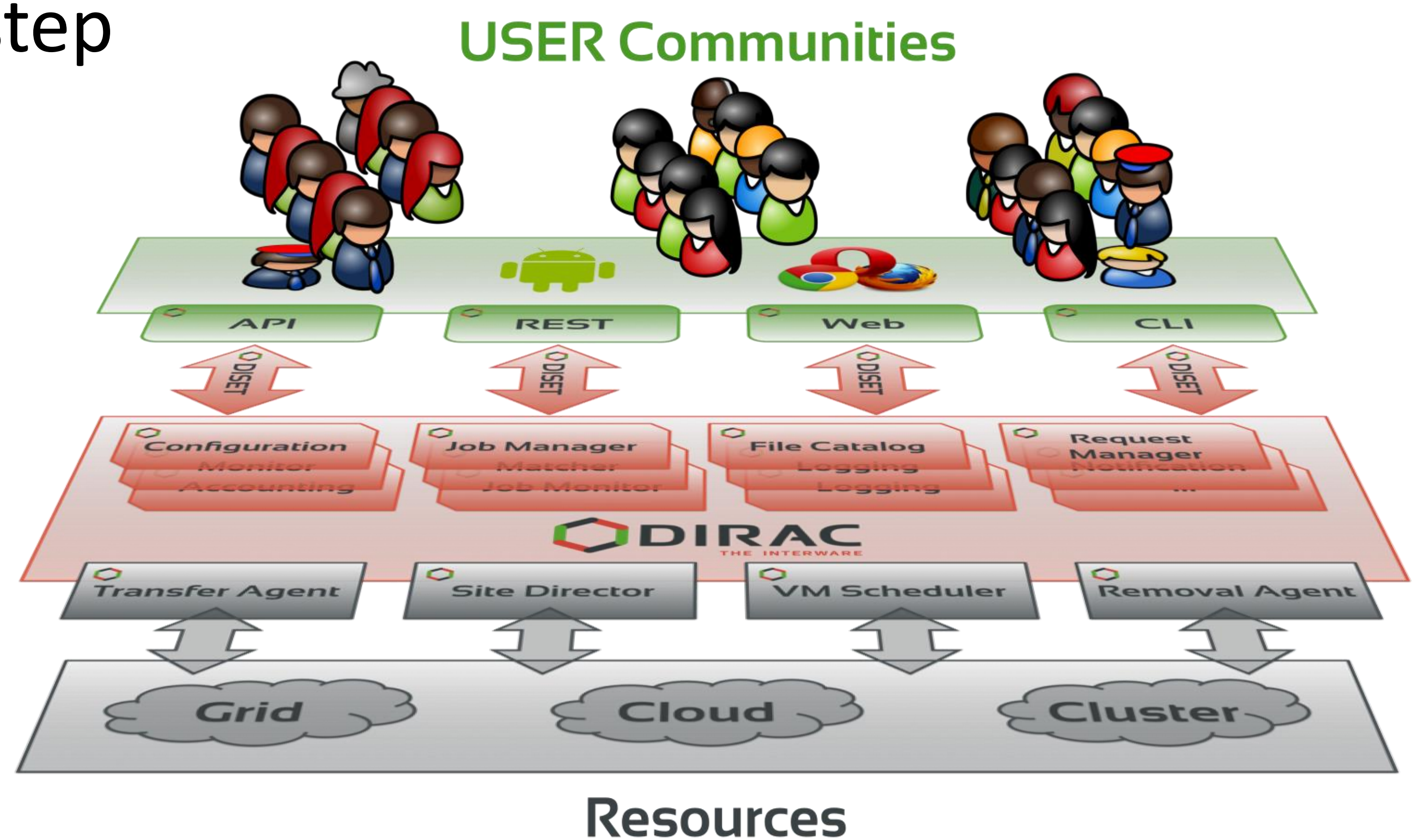
- Computing resources
  - 180 bi processors nodes (Intel Sandy Bridge E5-2680, 2.7 GHz, 8 cores) with 128 Go memory per node, equivalent to 2.880 cores (Bull)
  - 2 very large memory systems Bullx S6410 systems with 2 To memory
- Storage resources: 5 Po including 2 Po on disk
  - Hierarchical storage system Lustre + IBM HPSS

- Development of an academic cloud dedicated to the management and analysis of molecular biology data
  - 10.000 cores
  - 1PO storage
- Cloud stack: Stratuslab (OpenNebula)
  - Successful prototyping at IBCP
- Testing started early 2014



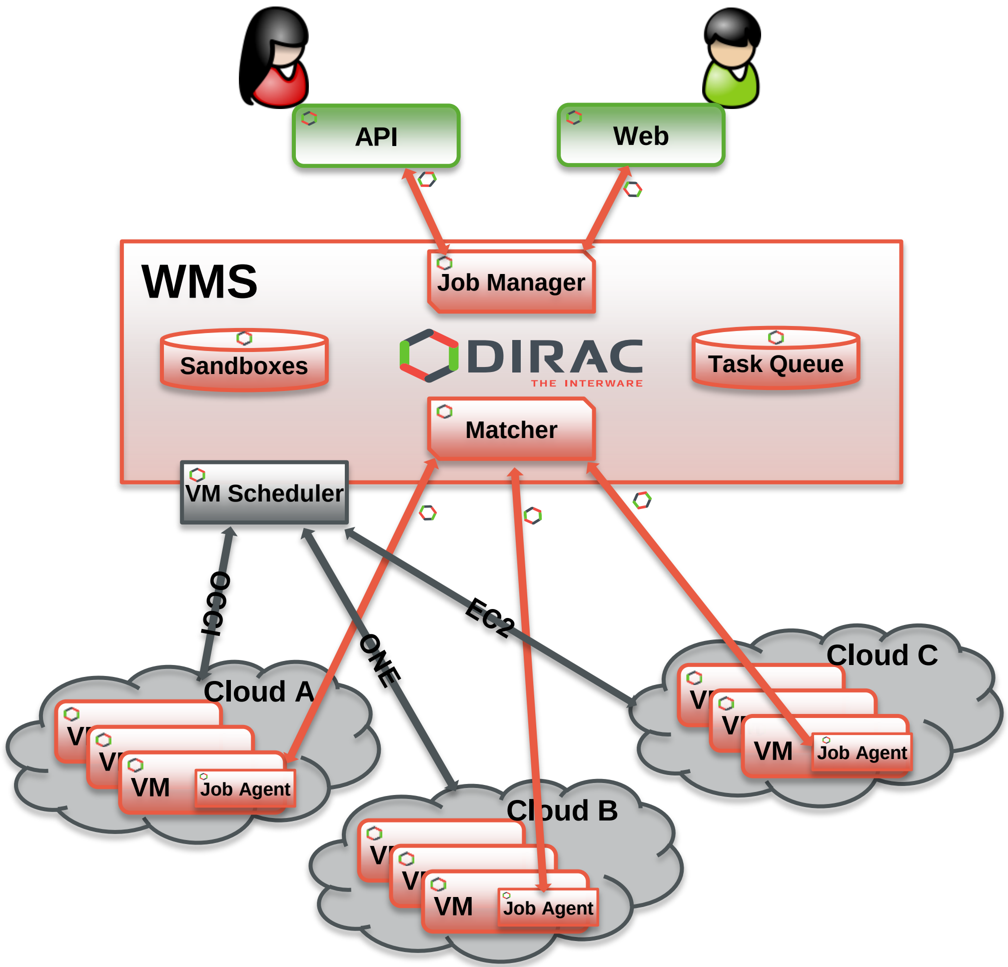
Migration of scientific gateways from grids to clouds

- Pilot agent platforms hide the technological step

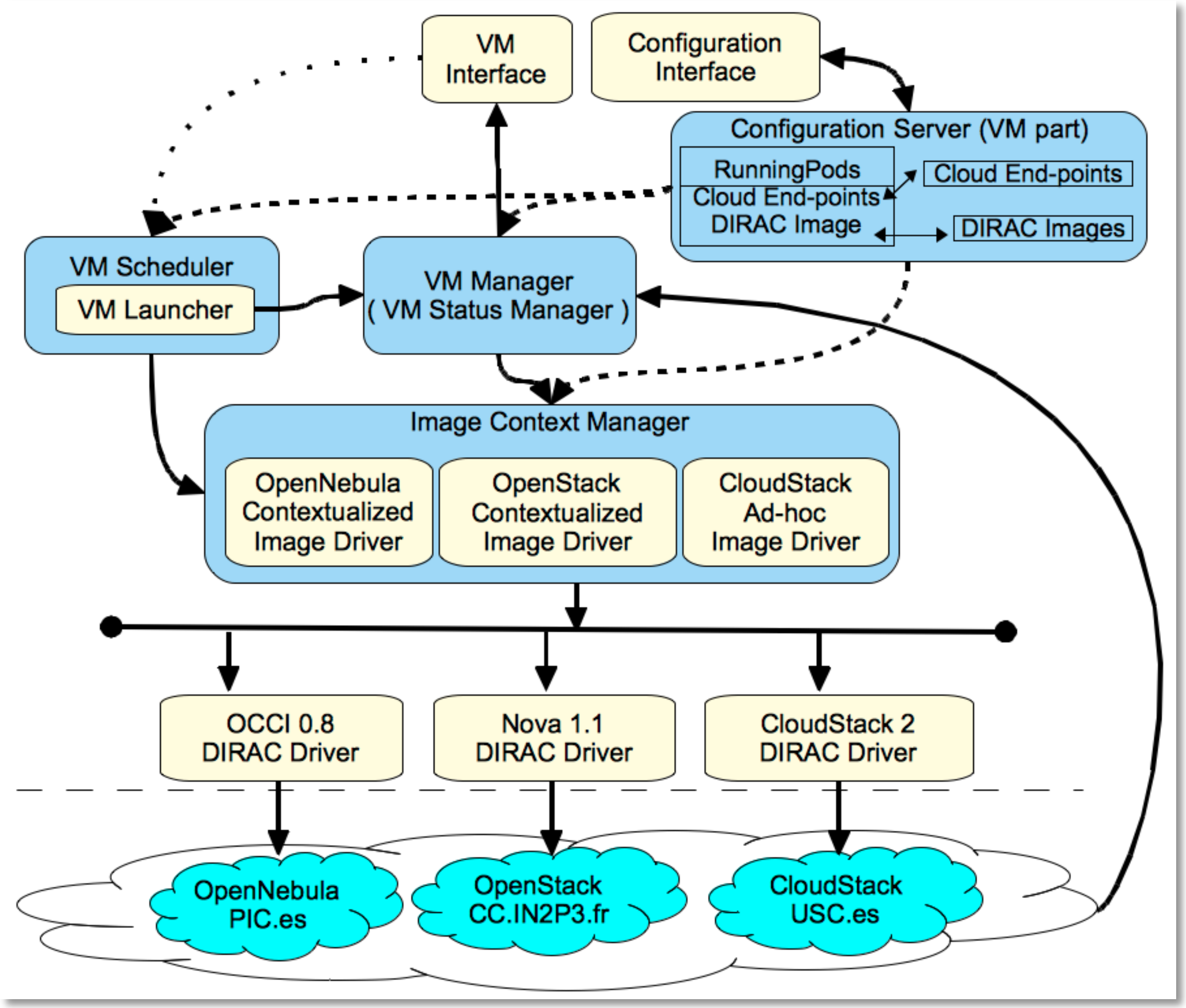


Credit: Vanessa Hamar

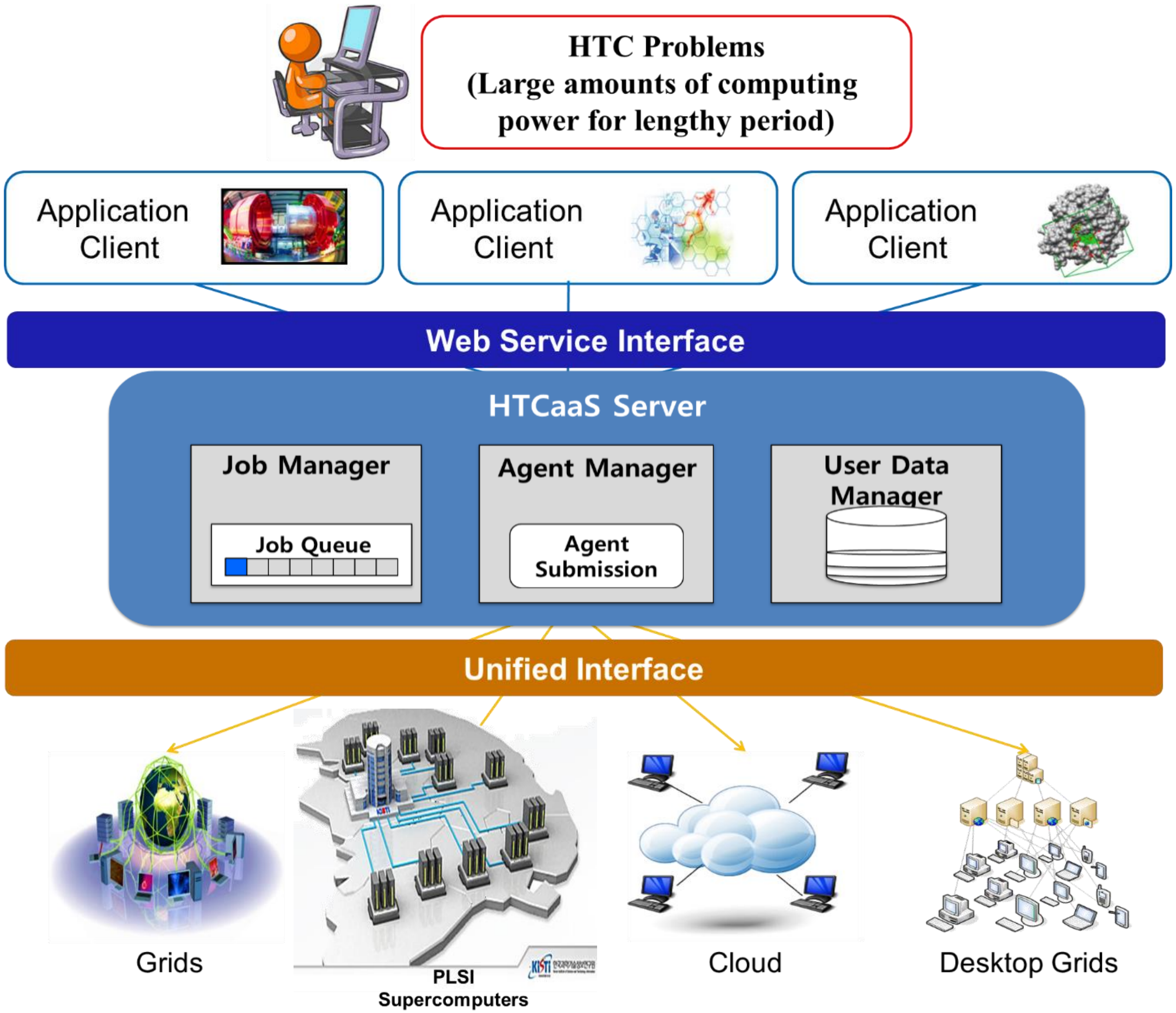
DIRAC & Clouds



Federated Cloud Test



WISDOM follow-up: High Throughput Computing as a Service



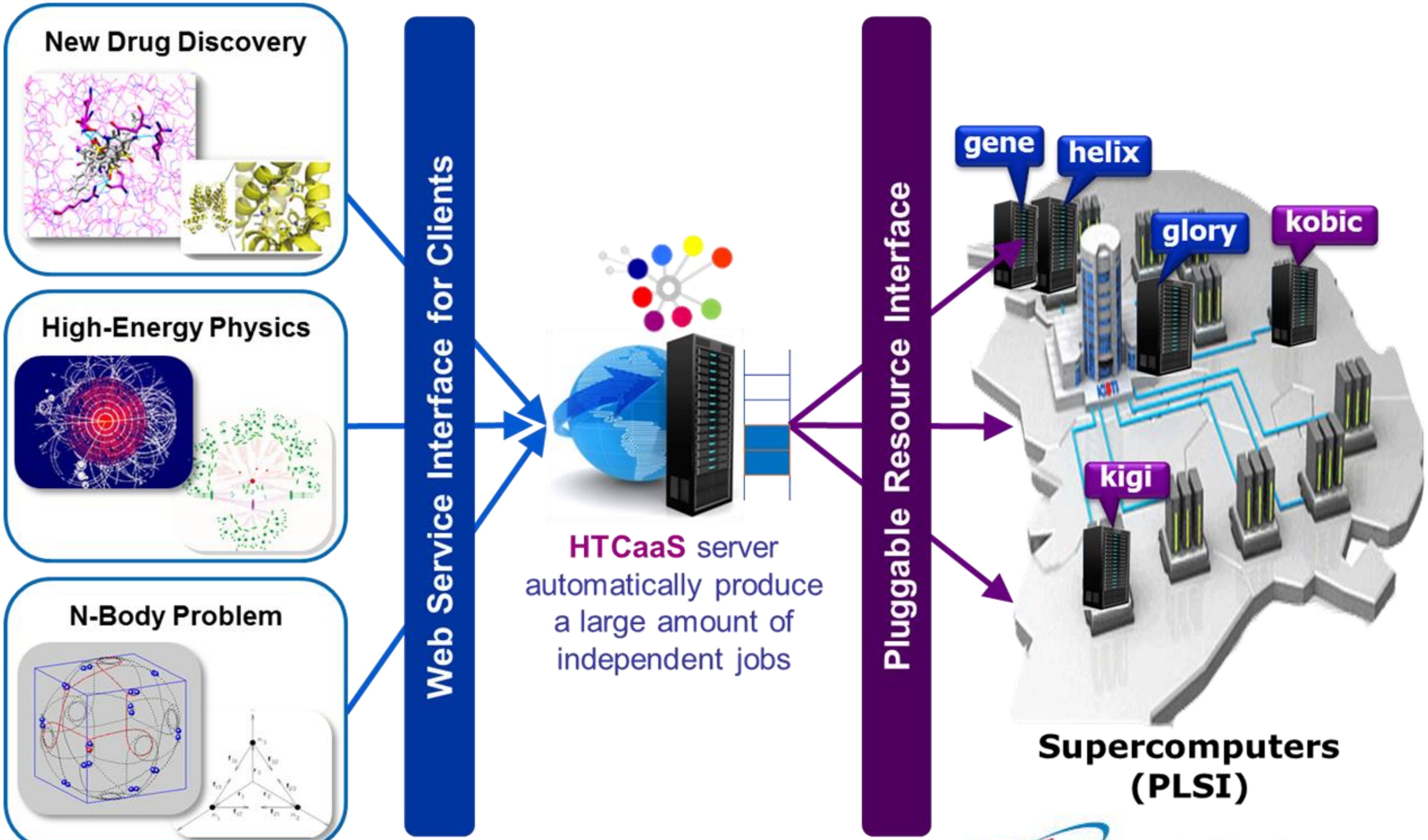
Credit: Soonwook Hwang

Korea Institute of Science and Technology Information

PLSI: Partnership & Leadership for the nationwide Supercomputing Infrastructure

- Consortium of 14 HPC Computing Centers in Korea
- ~100 TF computing capacity by combining 17 computing resources at 9 partner sites over a dedicated high-performance network

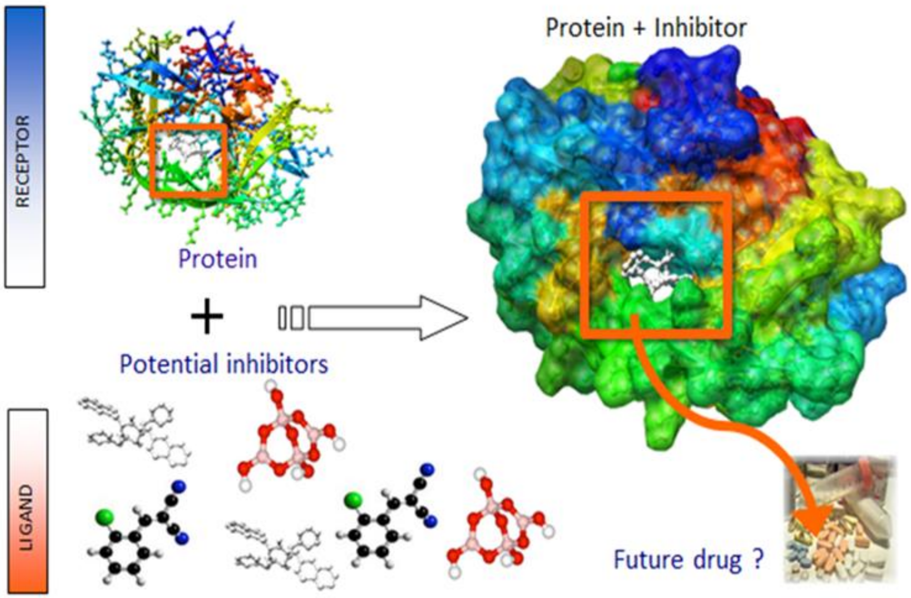
Pilot job-based High Throughput Computing(HTC) Environment running on top of PLSI



Protein-Ligand Docking using HTCaaS

Virtual Screening using Molecular Docking

- Autodock3/4, a suite of automated docking tools
  - perform the docking of ligands to a set of target proteins to discover new drugs for several serious diseases such as SARS or Malaria



No	Target Protein	PDB code	Ligand	Number of ligand	Meta Job ID	Protein preparation: Gene Cloning	Protein preparation: Protein Expression
1	Neuraminidase N1	3TI3	Chembridge	11455	125	O	O
				39533	123		
				47027	126		
				66141	127		
				68880	128		
				75099	129		
			Natural compounds	2720	124		
2	3C-like protease SARS	2ZU5	Natural compounds	2720	140	O	O
3	Human intestinal maltase	2QMJ	Natural compounds	2665	8	O	O
			Carbohydrate	14473	29		
			Marine Compounds	6154	25		
4	Malaria	3BPF	Natural compounds	2720	130	O	O
		1YVB	Carbohydrate	14473	27		
			Marine compounds	6154	24		
			Natural compounds	2665	6		

Key messages

- Grid computing has allowed building a truly multidisciplinary distributed IT infrastructure
- Cloud computing allows extending the grid functionalities
  - Life sciences will benefit even more
  - Public cloud prices and performances are not so appealing
  - Still a long way to the plateau of maturity for academic clouds
  - Pilot agent platforms allow a smooth transition from grids to clouds for users
    - Use of HPC resources through pilot agent platforms for High Throughput Computing

Clouds in biomedical sciences Part IV – entering a new world

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

Session IV: the future

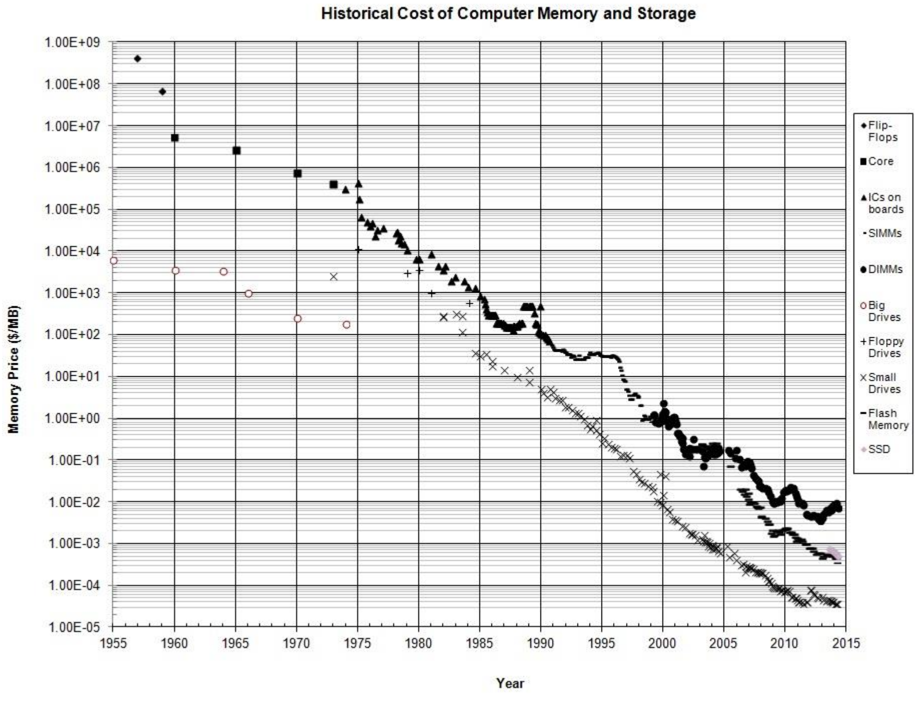
- Welcome to a new world
- Learn from history to prepare future: an introduction to Big Data
- What I do of my spare time…

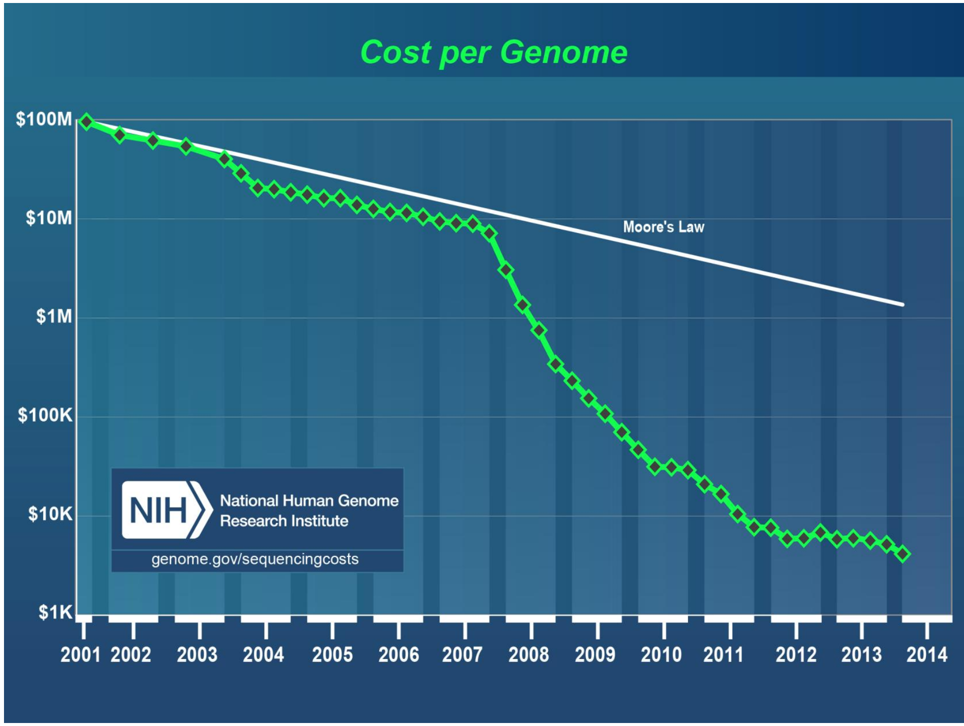
A new world beyond the standard model

- For more than 30 years, validation of the standard model
  - Electroweak physics at LEP
  - Top quark discovery at TEVATRON
  - Higgs Boson discovery at LHC
- New exploratory phase beyond the standard model
  - Where is the new physics?

A new world without Moore's law

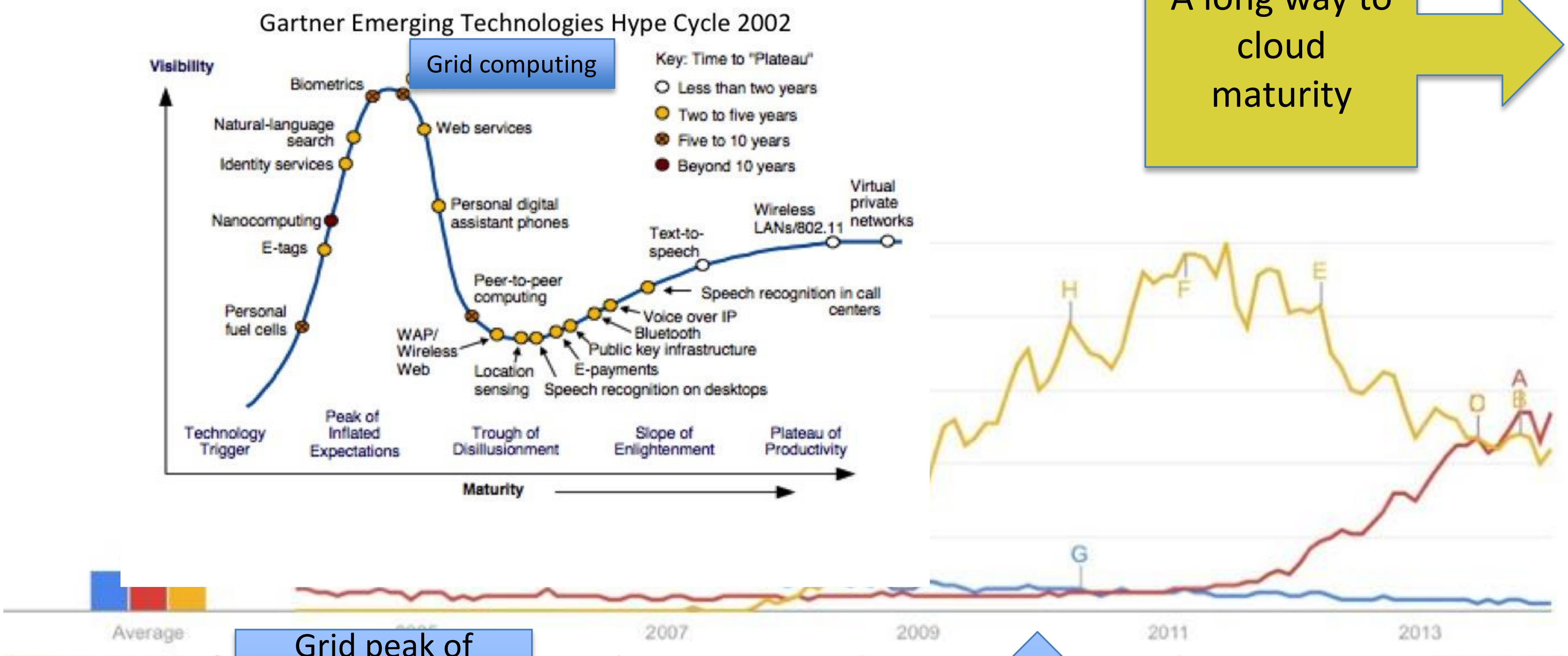
- Moore's law does not apply any more to storage capacities… nor to sequencing data production





It takes many years from hype to production quality

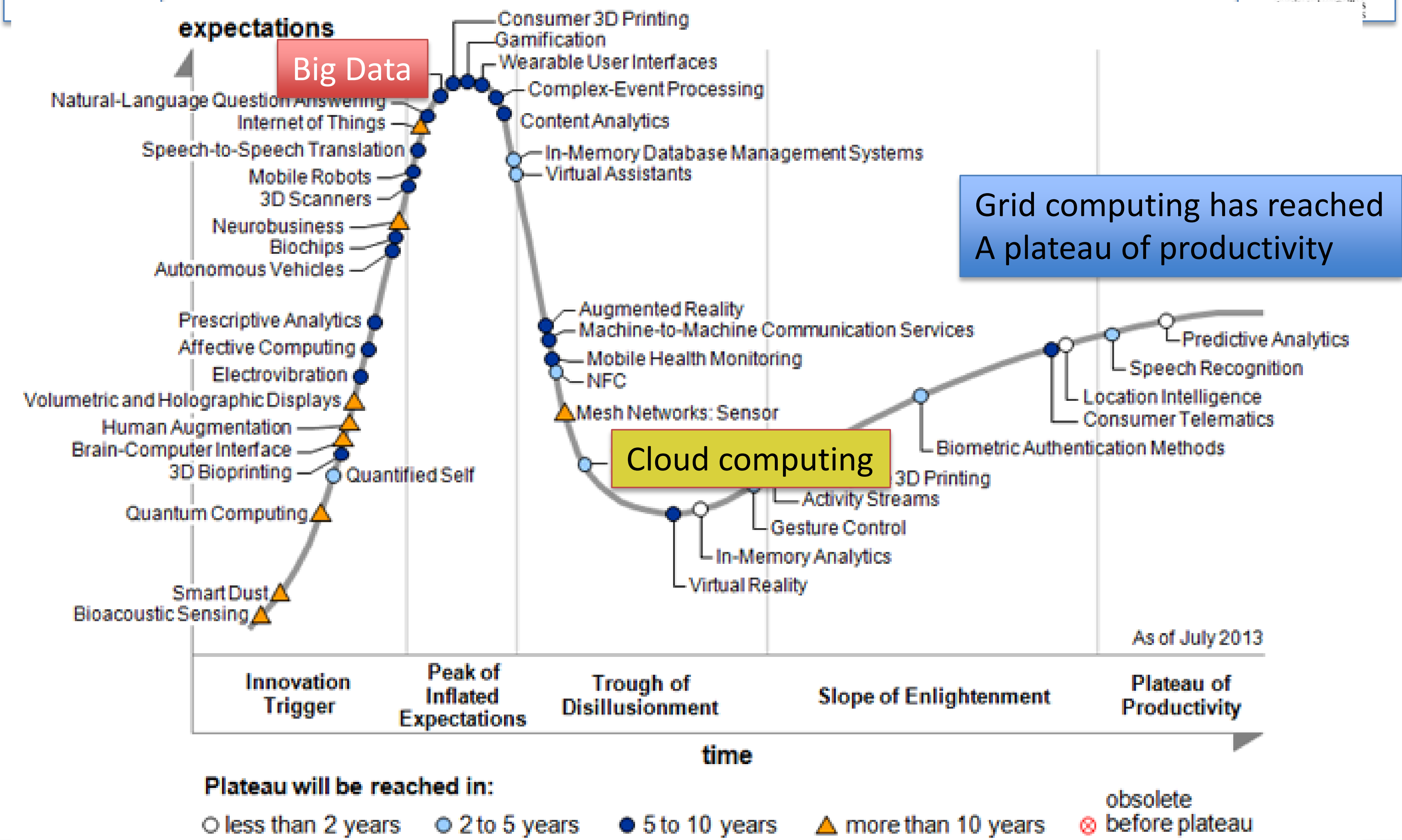
A long way to cloud maturity



Grid peak of expectations back in 2002

Grid maturity

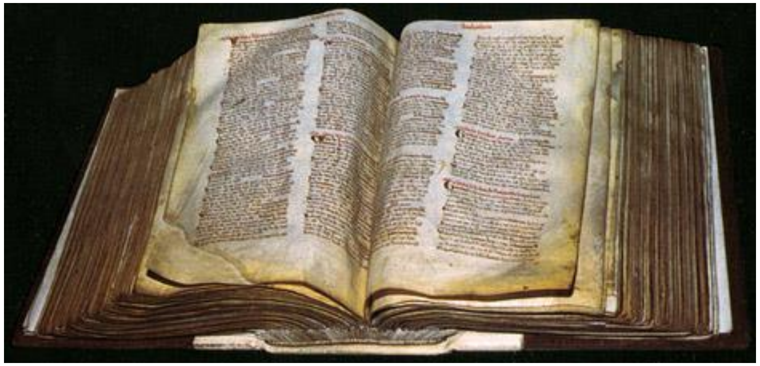
Gardner hype curve for 2013



Learning from history to build the future

- The greatest achievement of grids is not the capacity it has built
  - Obsolescence in three years for the hardware
  - Obsolescence of the grid middleware
- The greatest achievement are the human networks it has created
  - Fantastic human adventure

Learning from history: the Domesday Book (1087)



- Manuscript record of the great survey, completed in 1086 on orders of William the Conqueror

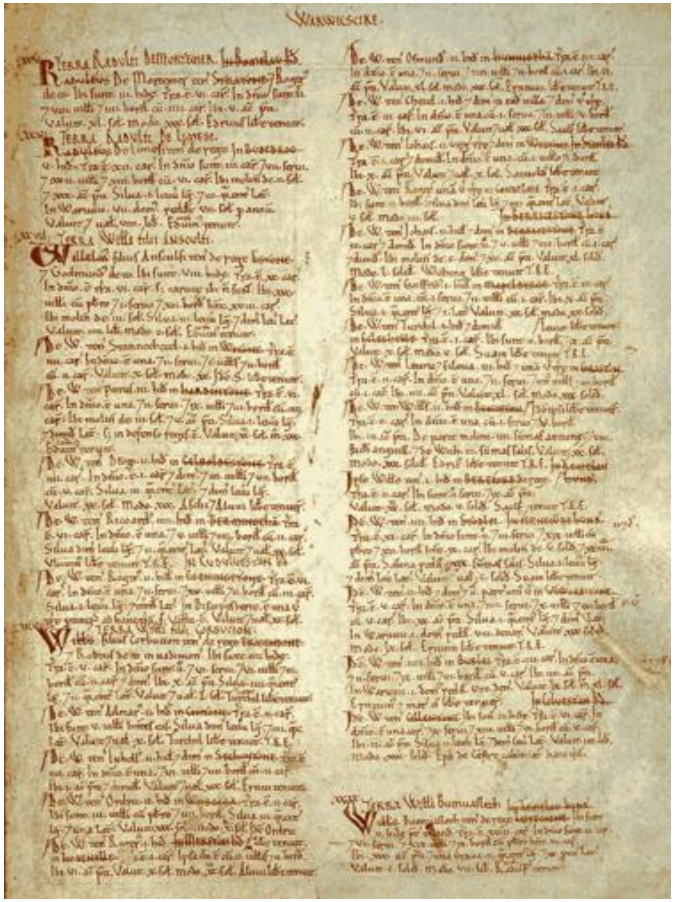
«While spending the Christmas time of 1085 in Gloucester, William had deep speech with his counsellors and sent men all over England to each shire to find out what or how much each landholder had in land and livestock, and what it was worth» *Anglo-Saxon chronicle*

- Absolute authority to define property rights since Middle Age

for as the sentence of that strict and terrible last account cannot be evaded by any skilful subterfuge, so when this book is appealed to ... its sentence cannot be quashed or set aside with impunity. That is why we have called the book 'the Book of Judgement' ... because its decisions, like those of the Last Judgement, are unalterable. *Richard Fitzneal, Dialogus de Scaccario, 1179*

Big data issues (I/II)

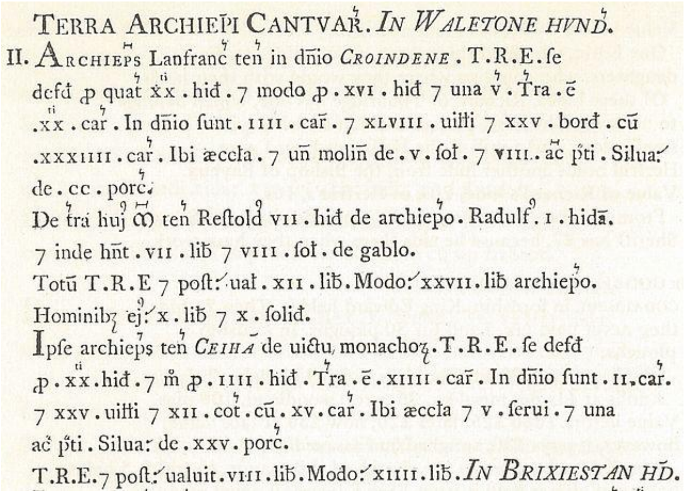
- Data collection
  - Every shire visited by a group of royal officers (1085-1086)
  - The unit of inquiry was the Hundred (a subdivision of the county)
- Data veracity
  - return for each Hundred was sworn to by twelve local jurors, half of them English and half of them Normans.
- Data analysis
  - names of the new holders of lands and assessments on which their tax was to be paid
  - national valuation list, estimating the annual value of all the land in the country



Crédit: Wikipedia

Big Data issues (II/II)

- Data presentation
  - Properties listed by fiefs
  - Properties listed by owner categories
    - king's holdings
    - holdings of churchmen and religious houses
    - Aristocrats
    - Lay men
- Data preservation
  - Preservation in the Royal Treasury in Westminster till 19th century
  - Stored at UK National Archives in Kew
  - 1986: digital version
  - 2002: access problem to digital version



TERRA ARCHIEPI CANTVAR. *IN WALETONE HVND.*

II. ARCHIEPS Lanfranc ten in dñio *CROINDENE* . T.R.E. ſe defđ p quat xx . hiđ . 7 modo p . xvi . hiđ 7 una v . Tra . ē .xx . car . In dñio ſunt . IIII . car . 7 XLVIII . uitti 7 XXV . borđ . cũ .XXXIIII . car . Ibi æccla . 7 un molin de . v . ſot . 7 VIII . ac pti . Silua: de . cc . porc.

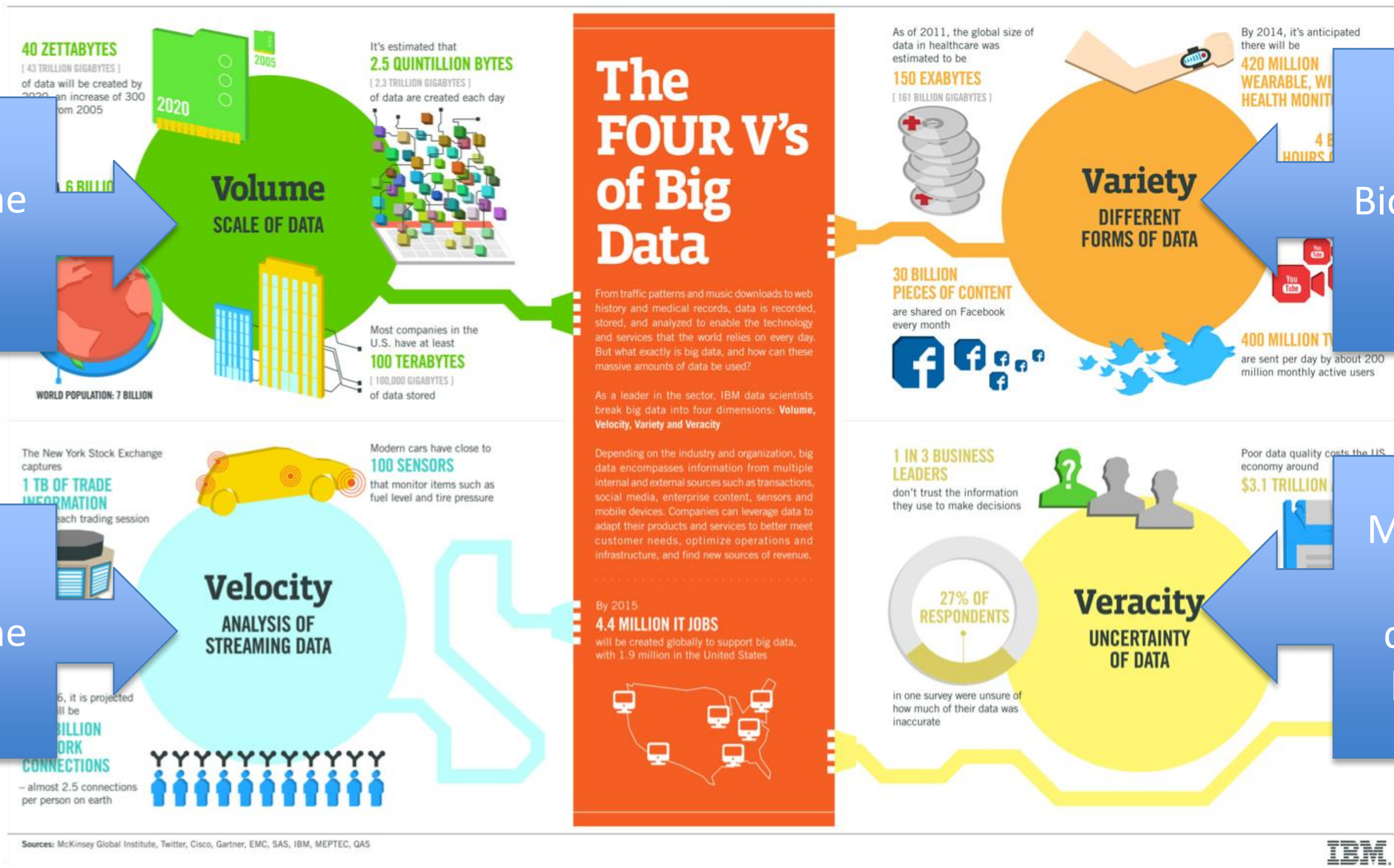
De tra huj m ten Reſtold VII . hiđ de archiepo . Radulf . I . hidā. 7 inde hnt . VII . lib 7 VIII . ſot . de gablo.

Totũ T.R.E 7 poſt: uat . XII . lib . Modo: XXVII . lib archiepo. Hominib ej: x . lib 7 x . ſoliđ.

Ipſe archieps ten *CEIHA* de uictu monachoꝝ . T.R.E. ſe defđ p . xx . hiđ . 7 m p . IIII . hiđ . Tra . ē . XIIII . car . In dñio ſunt . II . car. 7 XXV . uitti 7 XII . cot . cũ . XV . car . Ibi æccla 7 v . ſerui . 7 una ac pti . Silua: de . XXV . porc.

T.R.E. 7 poſt: ualuit . VIII . lib . Modo: XIIII . lib . *IN BRIXIESTAN HD.*

Big Data 4 Vs



LHC experiments

LHC experiments

Biodiversity

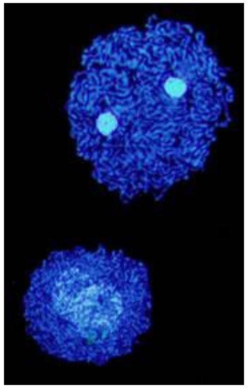
Molecular biology data are bugged

Data volumes: the example of metagenomics

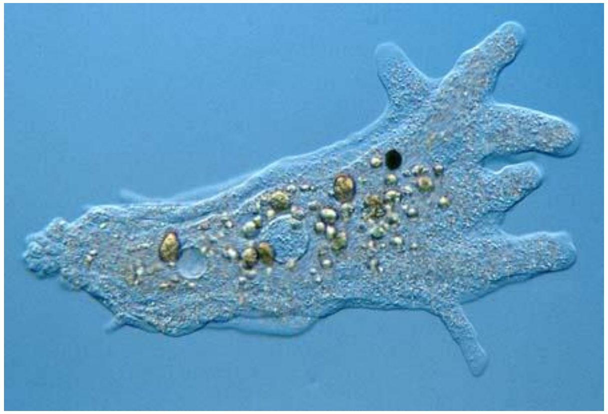
Metagenomics is the study of genetic material recovered directly from environmental samples.

Evolution of sequencing techniques

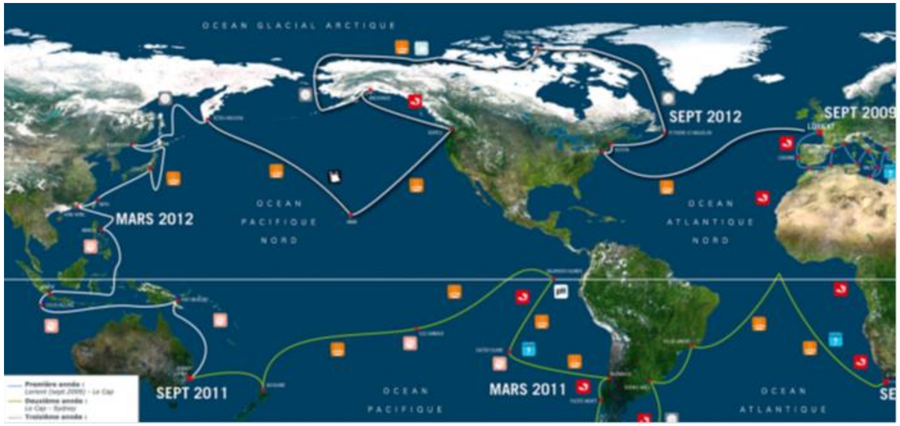
Sanger technology	500 base pairs (bp)
454 technology	10^5 400-600 bp reads
Illumina Technology	10^6 100 bp reads
TARA project	10^7 100-400 bp reads



Smallest non viral genome: *Carsonella ruddii* (0,16Mbp)

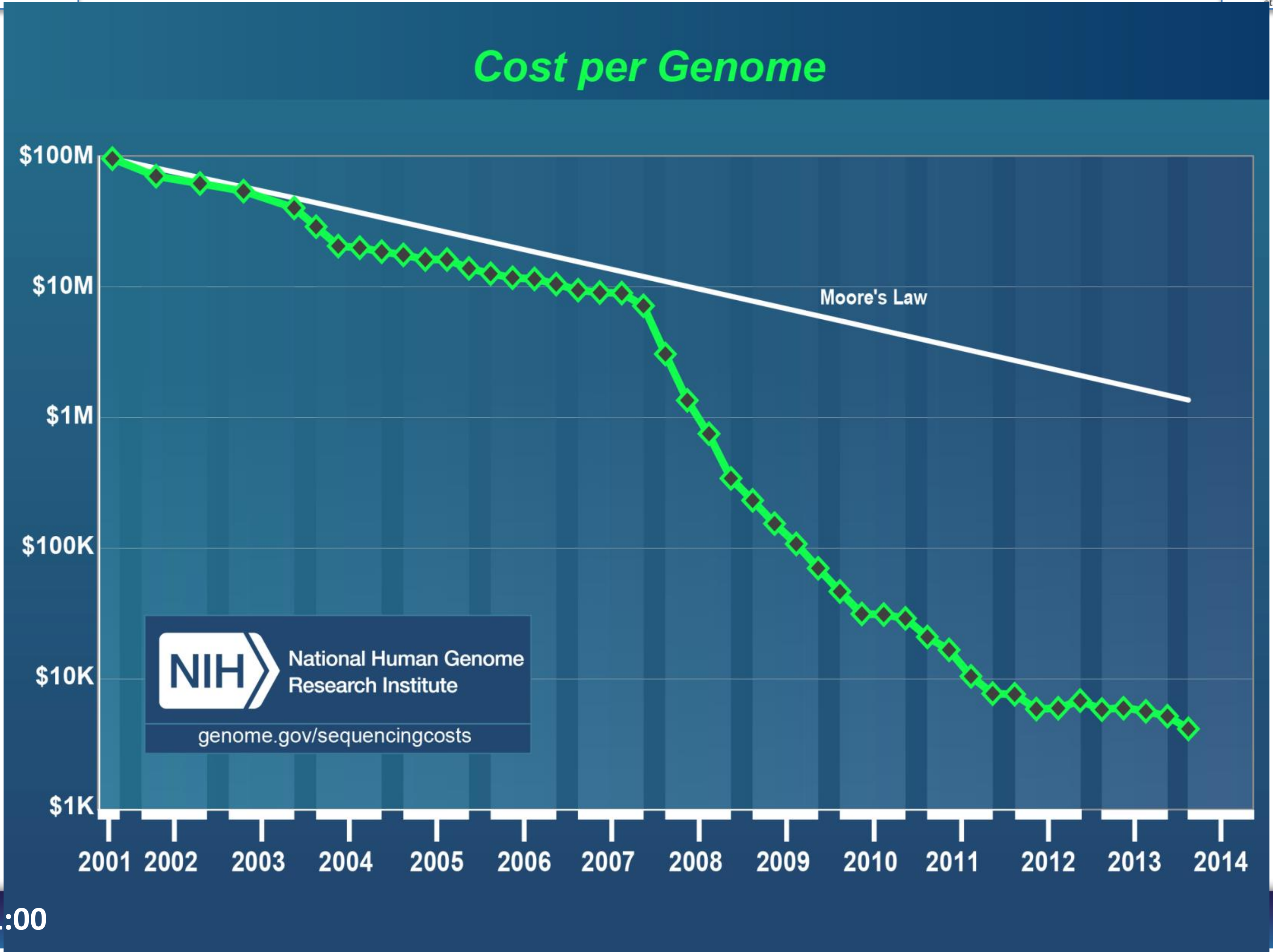


Largest genome: *Polychaos dubium* (670Gbp)

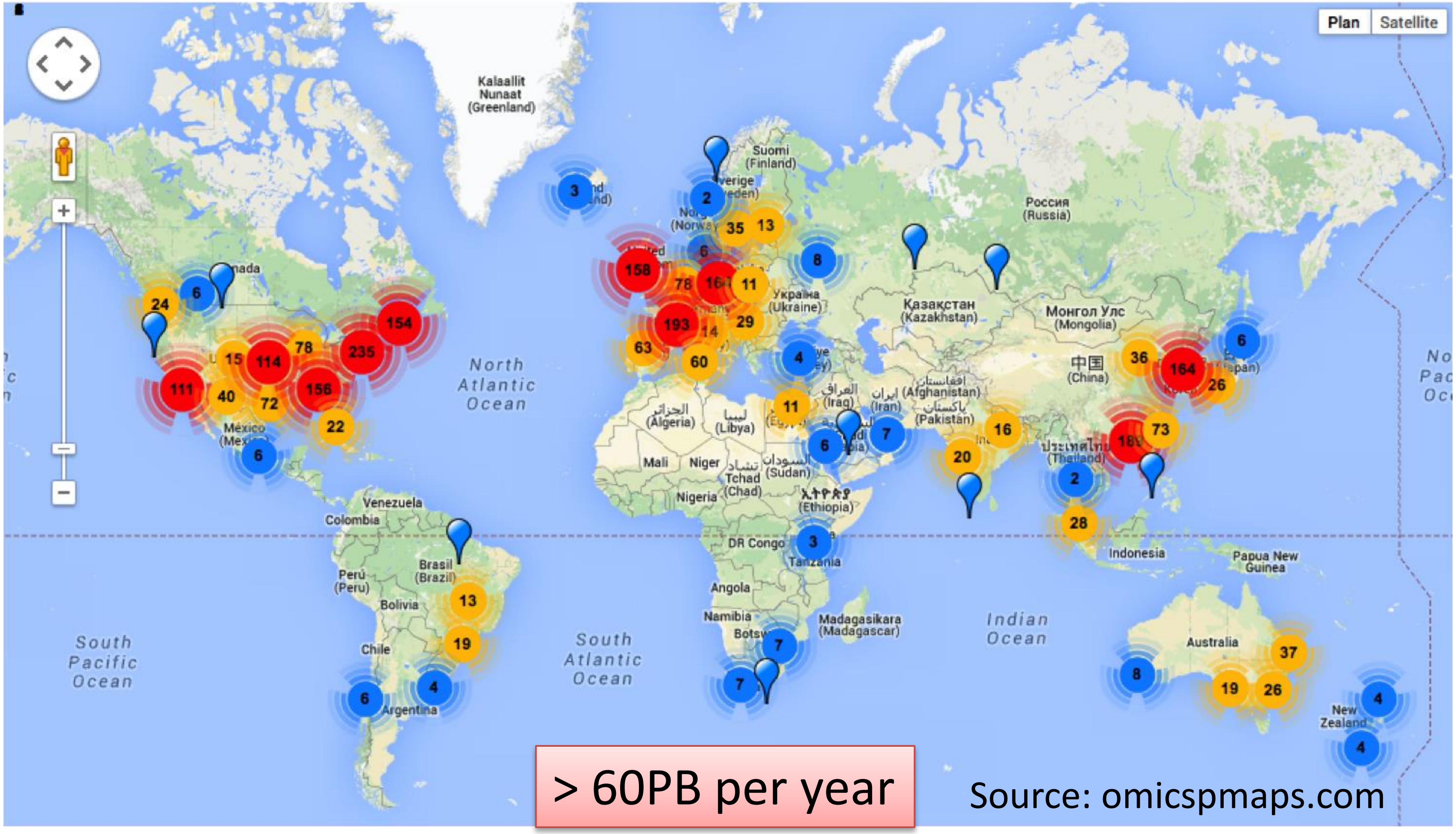


Tara @ http://oceans.taraexpeditions.org/

Cost per Genome is decreasing faster than Moore's law



Consequence: over 2500 Next Generation Sequencing machines in 900+ research centers in the world



> 60PB per year

Source: omicspmaps.com

Welcome to Auvergne, at the heart of France

1,35 Million inhabitants
26013 km^2

AUVERGNE
la région juste et grande

09:21:00

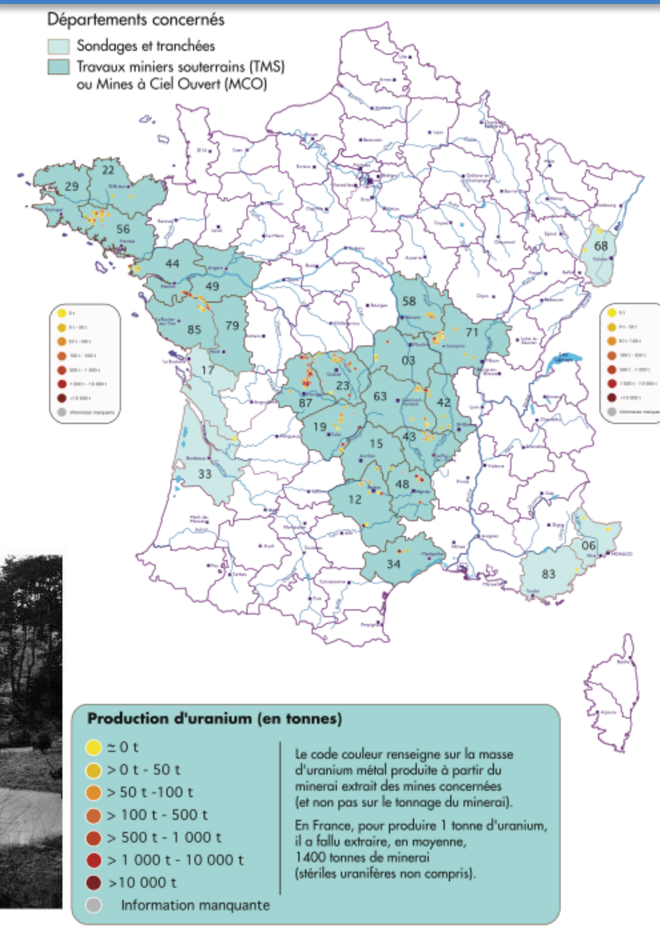
Auvergne at the heart of Uranium production in France

Map of uranium mines in metropolitan France

1949: first attempt to extract uranium ore in France in Lachaux (Auvergne)

In 50 years:

- 53 Million tons extracted in France till 2001
- 76000 tons of uranium ore produced in > 200 mines



ZATU, a Long Term Ecological Research dedicated to life under natural ionizing radiation

Natural radioactivity

Storage sites of uranium ore extraction residues

- Society in uranium rich territories
  - Social impact of uranium extraction
  - Preserving the long term memory
- Characterization, behavior and transfer of radionucleids
  - long term future of radionucleids in storage sites
- Impact of radiation on living systems
  - Multigenerational effects of chronic exposure to radiation

Impact of chronic exposure to low dose ionizing radiation on living organisms

- From the Chernobyl environment, a coherent picture of predictable radiation-induced effects for low-dose-rate exposures has not emerged
  - Contradictory experimental evidences from Chernobyl exclusion zone
- Need to collect more data from Chernobyl exclusion zone but also from other ecosystems under chronic low dose exposure
  - Radioactive water sources
- Point 0: what happens in “total” absence of radioactivity?

Photographs of abnormalities in barn swallows. (a) Normal phenotype. (b–d) Partially albinistic plumage. (e) and (f) Deformed beak. (g) Deformed air sacks. (h) and (i) Bent tail feathers.

Proasellus cavaticus

ZATU strategy

Multidisciplinary long term observation of selected sites in Auvergne, Massif Central and Massif Armoricain

- Radionucleid chemical speciation
- Industrial heritage
- Biodiversity survey

Characterization

Transfer

- Radionucleid migration
- Interaction of radiation with living organisms
- Territory administration and responsabilities

- Interactions and retroactions between matter and living systems
- Risk evaluation
- Prevention tools

Environmental impact

Significant production of scientific data (geography, ecology, biology, metagenomics, chemistry, physics, social sciences)

How to make all these data speak to each other is a huge challenge

Conclusion

- Grid computing has allowed building a truly multidisciplinary distributed IT infrastructure
  - Greatest achievement: human networks
- Cloud computing allows extending the grid functionalities
  - All sciences will benefit even more
  - Still a long way to the plateau of maturity
  - Scientific gateways and pilot agent platforms allow a smooth transition from grids to clouds
- Big Data is the next frontier
  - Volume will not be necessarily the most difficult challenge

Which data produced today will still be used in 900 years?