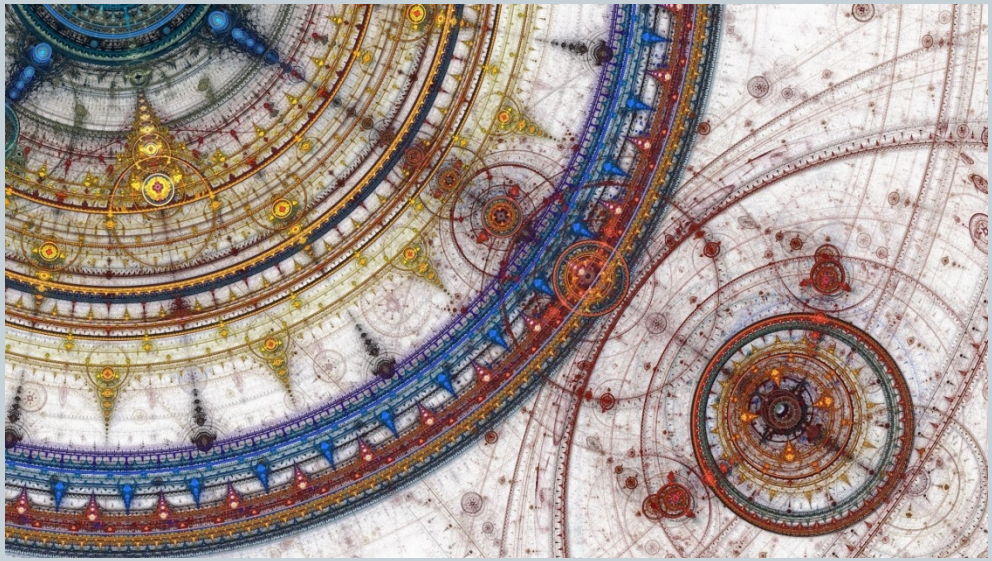
Going beyond the Grid to enable life science data analysis

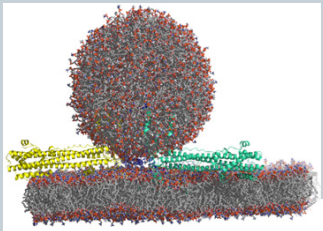
A SHORT OVERVIEW

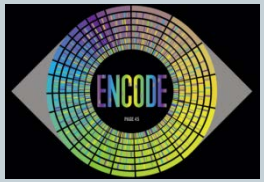


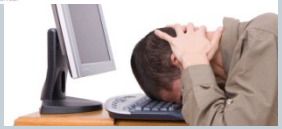
Shifting Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$J_i = \frac{dn_i/dt}{S_p} = -D_i\left(\frac{dc_i(x)}{dx} - c_i(x)\frac{z_i F}{RT}\frac{d\psi(x)}{dx}\right)$$







Jim Gray on eScience, The Forth Paradigm, Microsoft Research, 2009

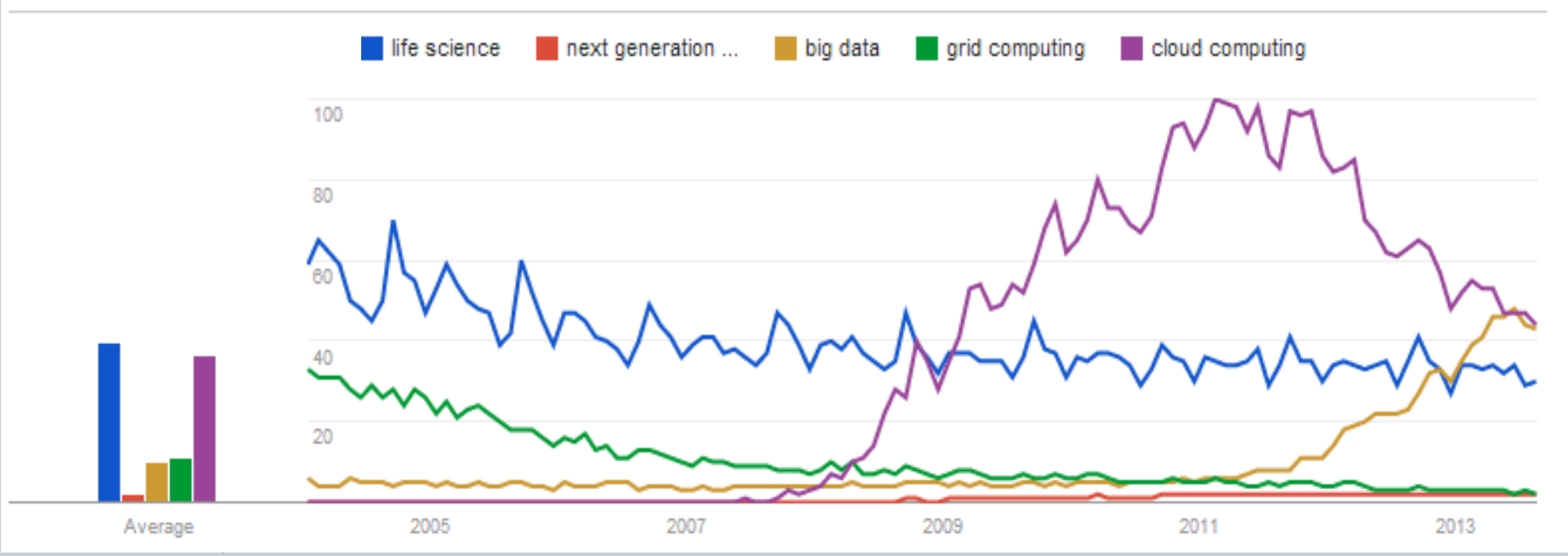
Big Data Biology

- The term “Big Data” is not only for size:
  - Speed
  - Volume
  - Computational and analytical capacity to manage data and derive insight

- The “**Forth Paradigm**” is at hand in Life Sciences
  - the analysis of massive data sets

“It’s the data, stupid”

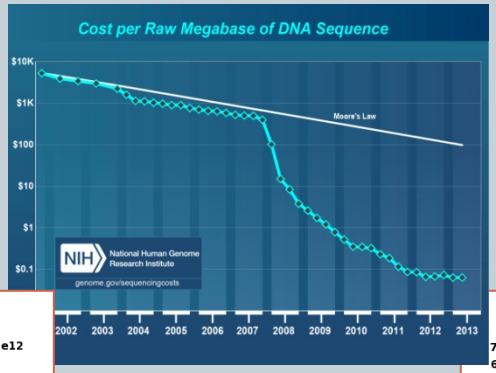
- It’s a new scientific methodology based on the power of

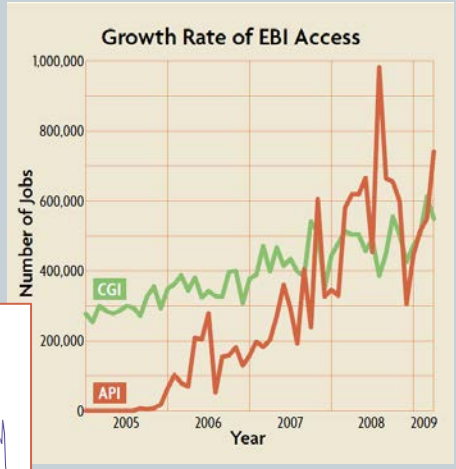


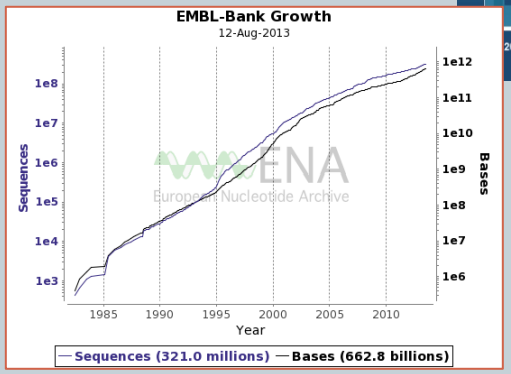
- At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order, but of dimensionally agnostic statistics.

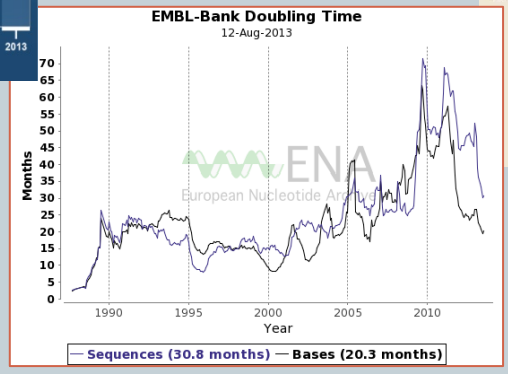
Big Data Biology

- Moving from traditional small-scale, focused experiments to more hypothesis-neutral studies
- Small biology labs can become
  - Big data generators
  - Big data users



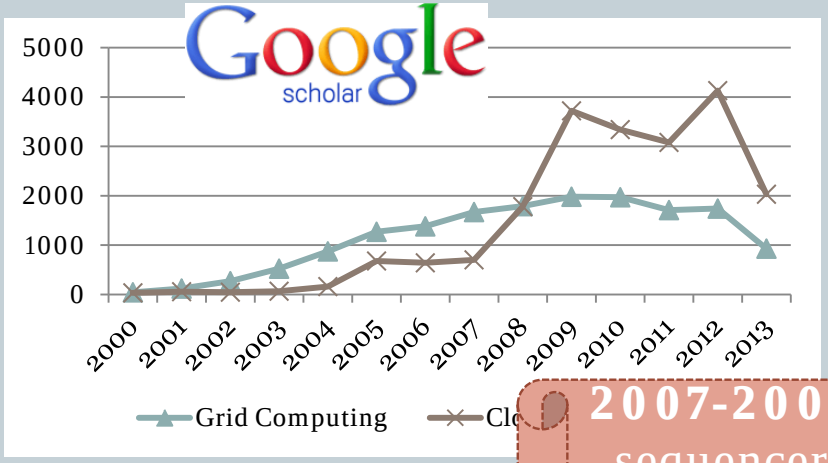


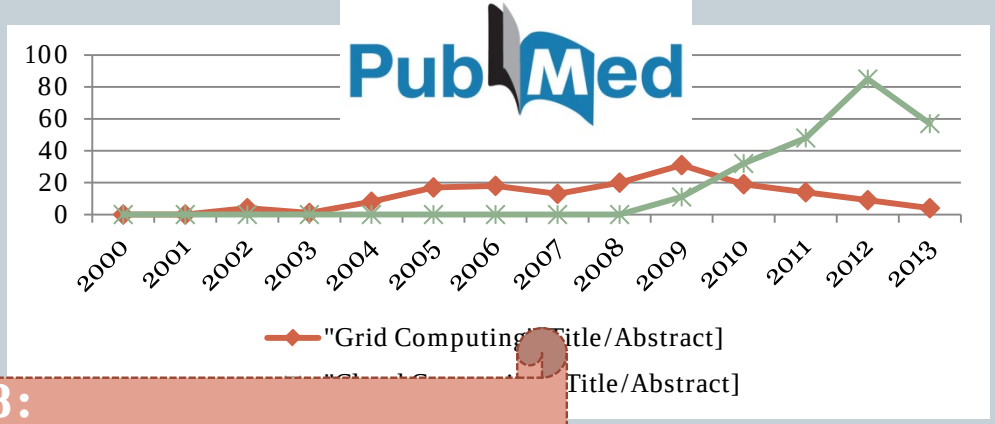




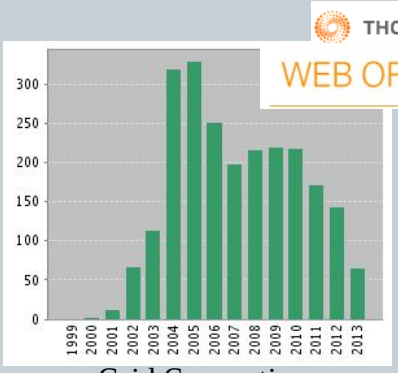
The story so far…

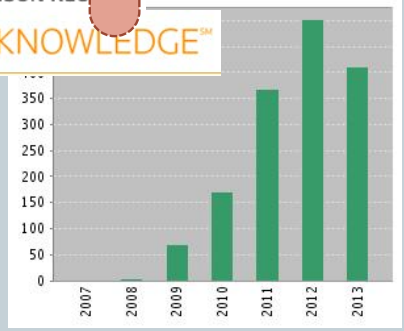
"We can know more than we can tell" Michael Polanyi (1891-1976)

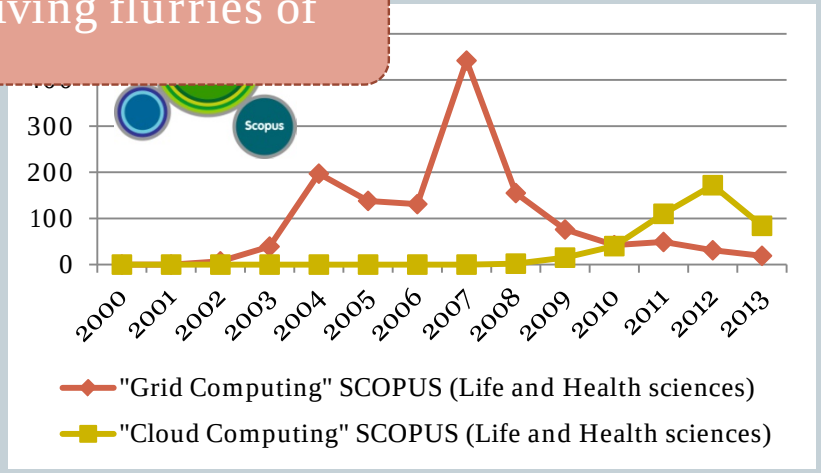




2007-2008:
sequencers begin giving flurries of data





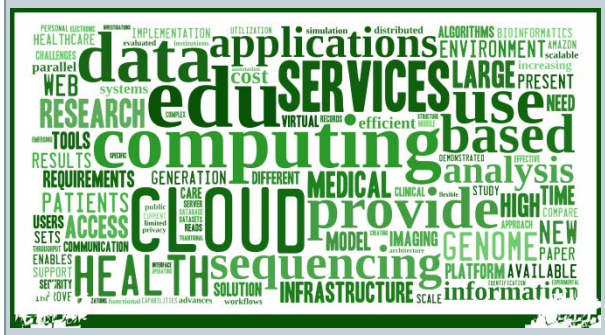


Words of the story...

- 391 abstracts from PubMed
- 4,770 unique terms

Word Cloud for "*Grid*" abstracts

Word Cloud for "*Cloud*" abstracts



Common terms

- comput
- data
- system
- provid
- technolog
- applic
- resour
- analysi

Grid terms

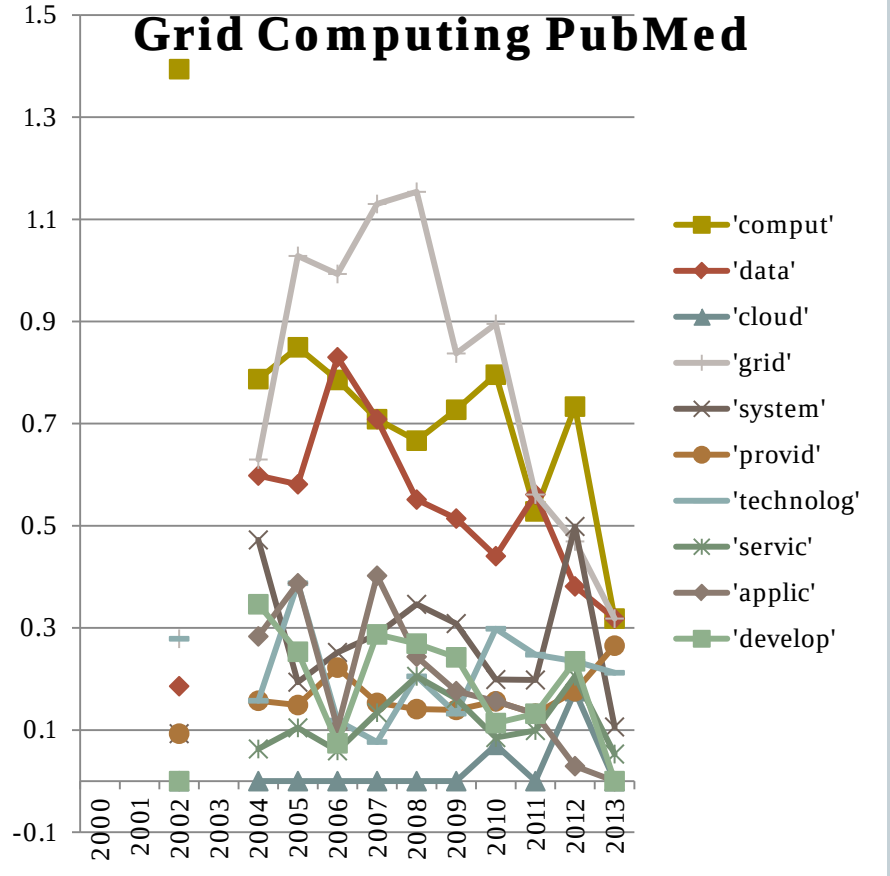
- grid
- model
- distribut
- bioinformat
- molecular

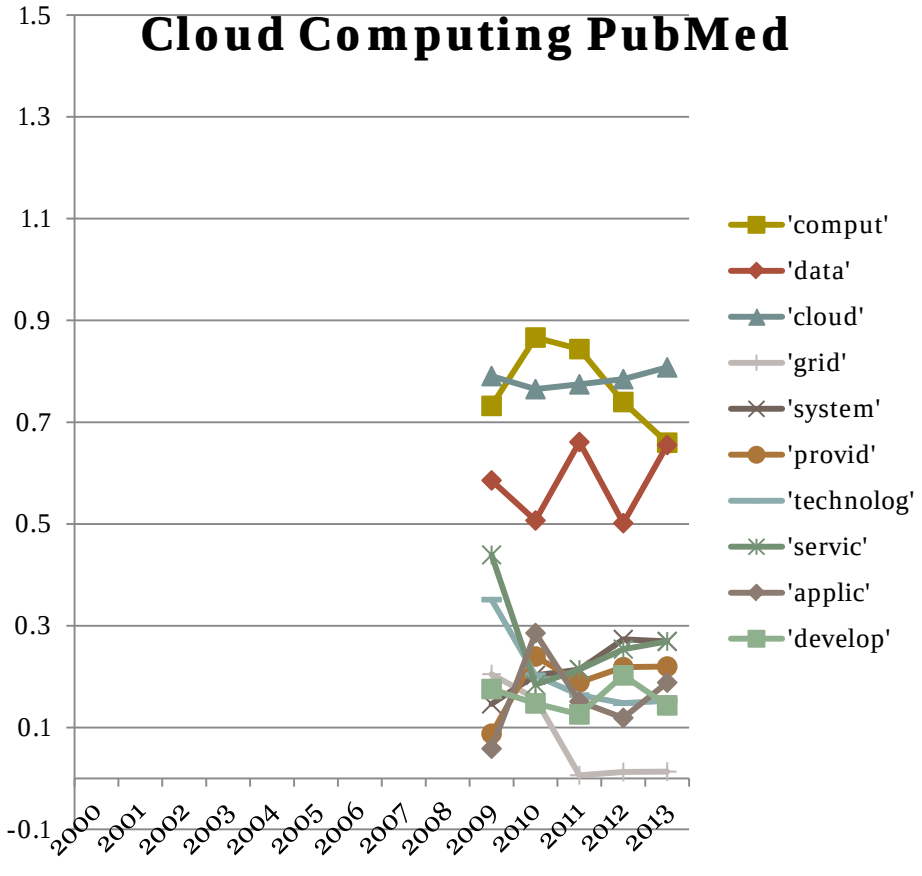
Cloud terms

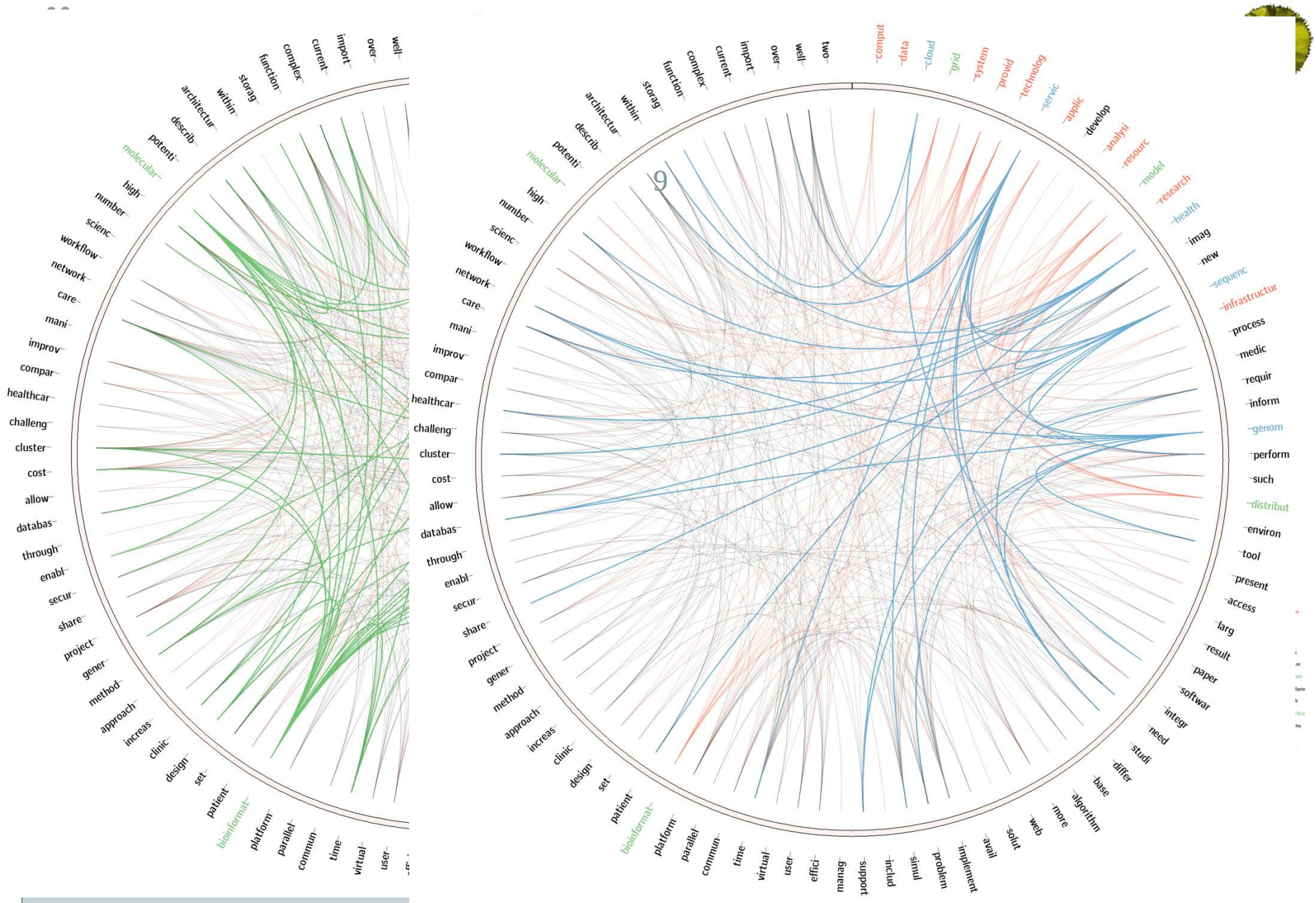
- cloud
- servic
- sequenc
- health
- genom

Word associations

- Plotting word frequencies across the decade:

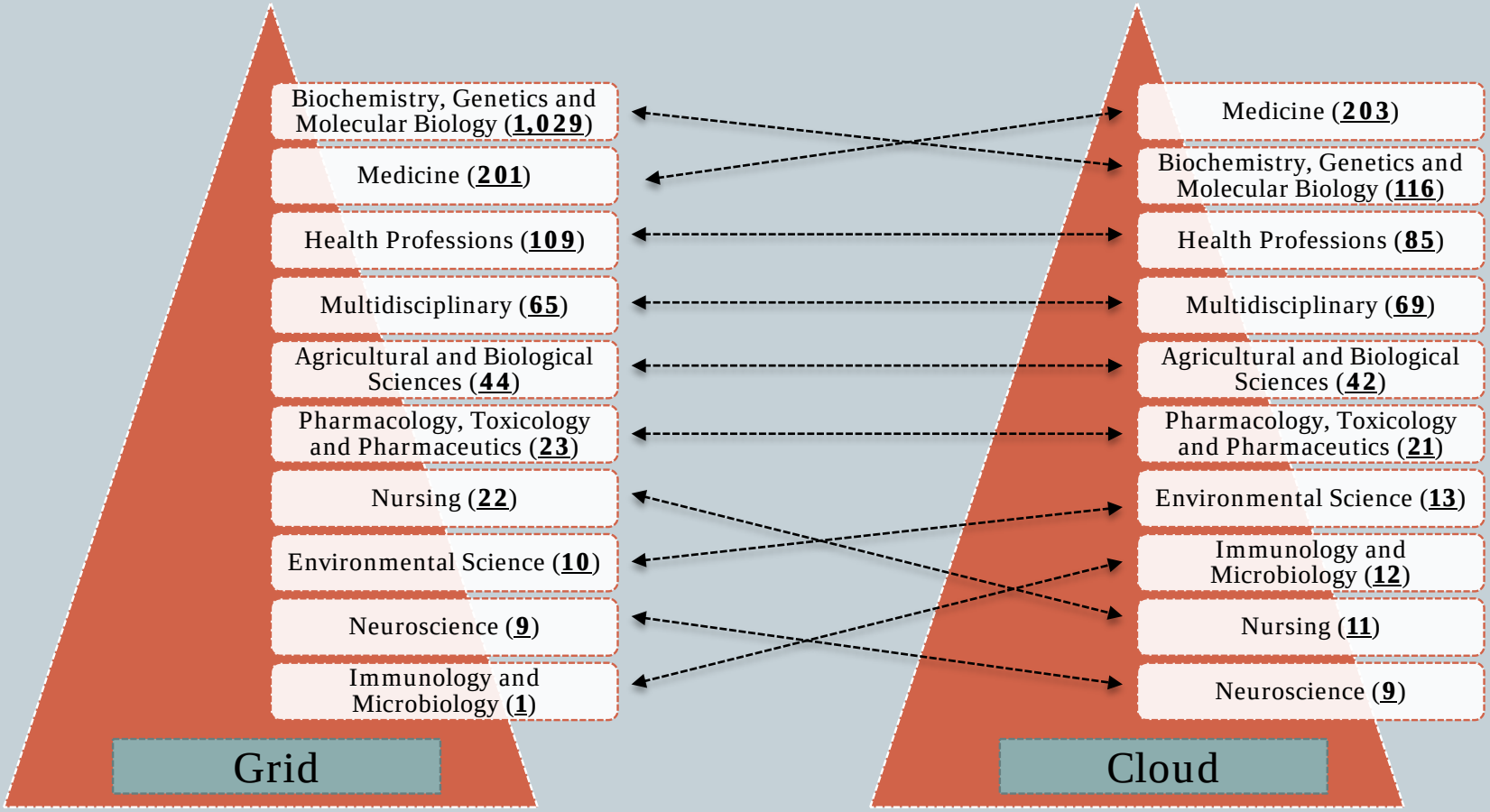





well
over
import
current
complex
function
storag
within
architectur
describ
potenti
molecular
high
number
scienc
workflow
network
care
mani
improv
compar
healthcar
challeng
cluster
cost
allow
databas
through
enabl
secur
share
project
gener
method
approach
increas
clinic
design
set
patient
bioinformat
platform
parallel
commun
time
virtual
user
effici
manag
support
includ
simul
problem
implement
avail
solut
web
more
algorithm
base
differ
studi
need
integr
softwar
paper
result
larg
access
present
tool
environ
distribut
such
perform
genom
inform
requir
medic
process
infrastructur
sequenc
new
imag
health
research
model
resourc
analysi
develop
applic
servic
technolog
provid
system
grid
cloud
data
comput
two
9

Any field in particular?

- Research areas from SCOPUS

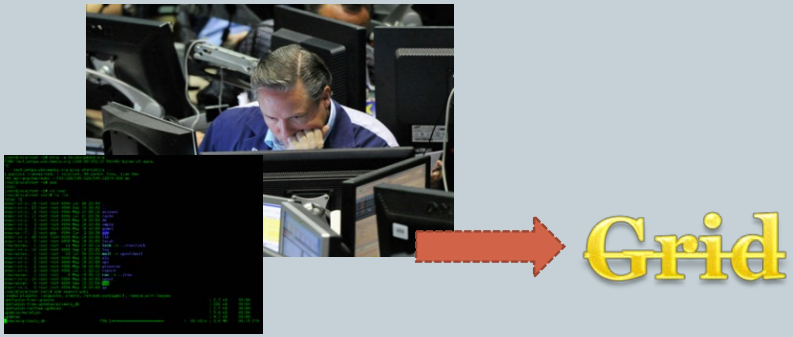


Making the bridge…

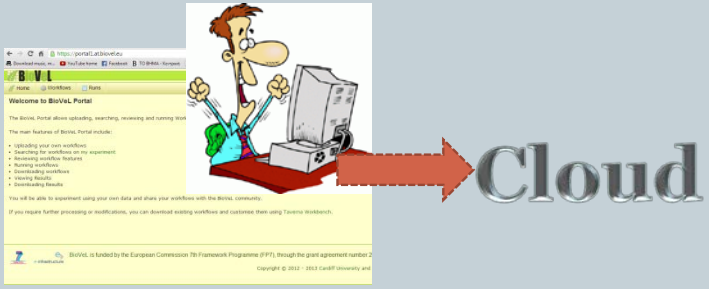
"Ba: Knowledge creation requires a time and place in which people share knowledge and work together as a community."

Kitaro Nishida

x "Grid computing" in 2004:



✓ "Cloud computing" in 2014:



Interlude

EGI? What is EGI?

Grid vs. Power Grid

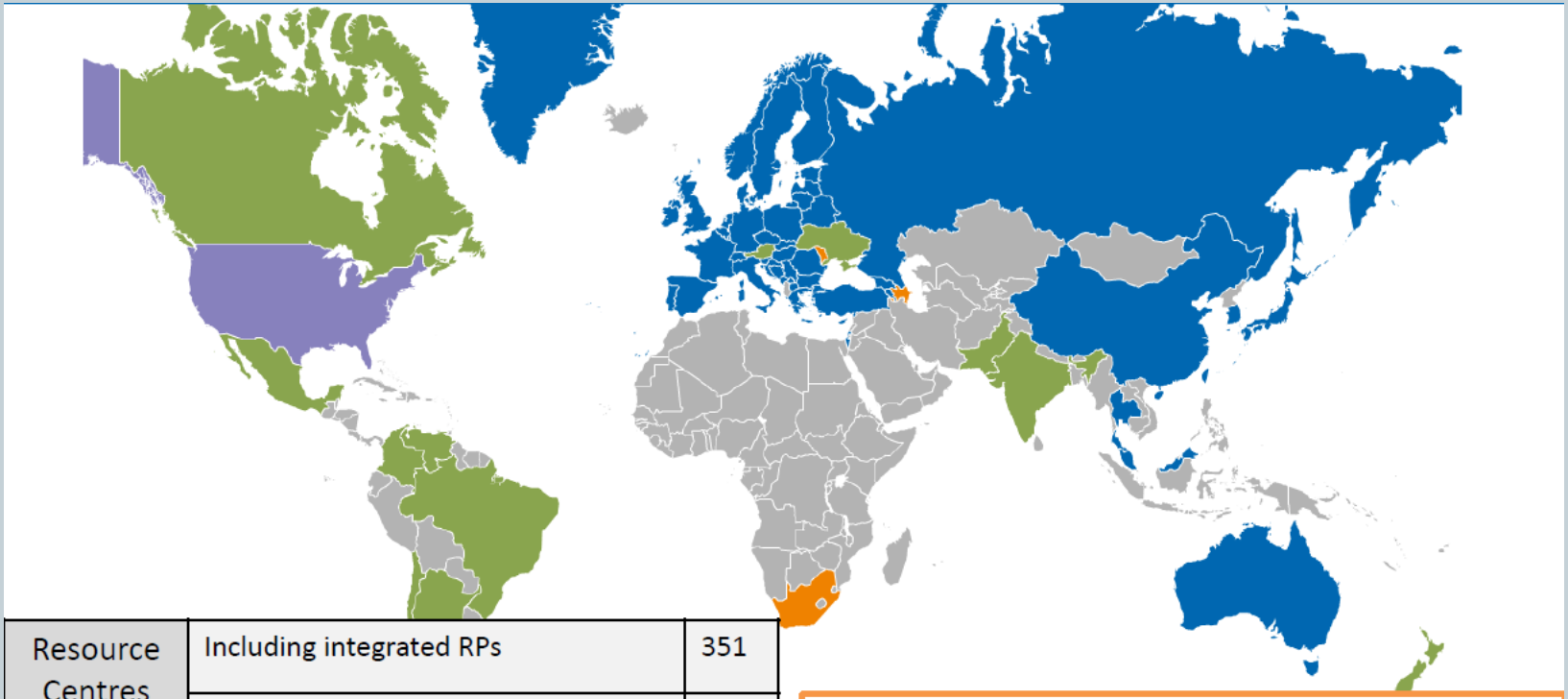
Power Grid	Computational Grid
You **never worry** about where the electricity you are using comes from. You simply know that when you plug your toaster in to the wall socket, it will get the electrical power you need to do the job.	You would **never worry** about where the computer power you are using comes from. You simply know that when you plug your computer in to the Internet, it will get the computer power you need to do the job.
The **infrastructure** that makes this possible is called "the power grid". It links together power plants of many different kinds with your home.	The **infrastructure** that makes this possible is called "the Grid". It links together computing resources.
The power grid is **pervasive**: electricity is available essentially everywhere and you can simply access it through a standard wall socket.	The Grid is be **pervasive**: remote computing resources would be accessible from different platforms, including laptops, PDAs and mobile phones, and you will simply access the Grid through your web browser.
The power grid is a **utility**: you ask for electricity, and you get it. You also pay for what you get.	The Grid is a **utility**: you ask for computer power or storage capacity and you get it. You also pay for what you get.

EGI, EGI.eu, EGI-InSPIRE

- EGI – an open collaboration
  - To support the digital European Research Area through a pan-European research infrastructure based on services federed from the NGIs
- EGI.eu – a Dutch foundation owned by the NGIs
  - To coordinate the work of EGI (operations, technology, user support, policy, community & communications, administration)
  - Sustainable small coordinating organisation
- EGI-Engage – an H2020 project
  - EGI-Engage aims to accelerate the implementation of the Open Science Commons by expanding the capabilities of a European backbone of federated services for compute, storage, data, communication, knowledge and expertise, complementing community-specific capabilities.

- European
  - Over 35 countries
- Grid
  - Secure sharing
- Infrastructure
  - Computers
  - Data
  - Instruments
  - .... and beyond!!

Resource Infrastructure Providers

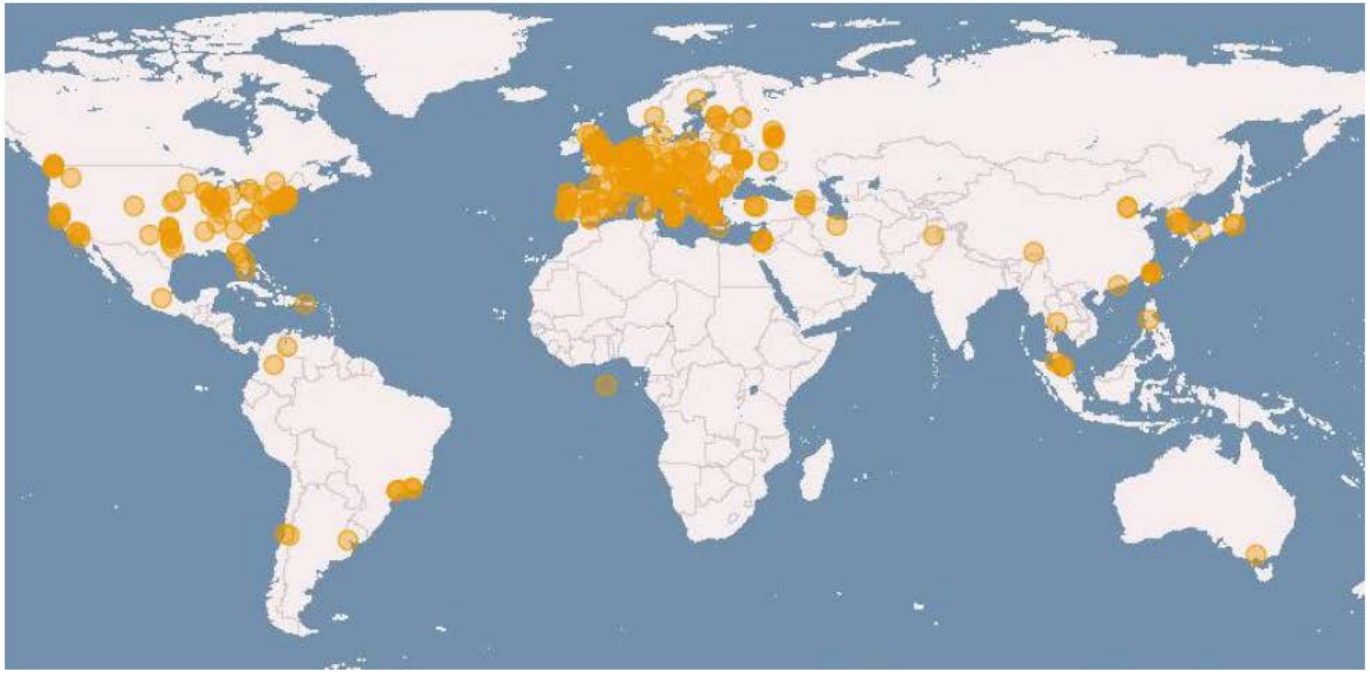


Resource Centres	Including integrated RPs	351
	Supporting MPI	87
Countries	EGI-InSPIRE & EGI Council members	43
	Including integrated RPs	59

Installed Capacity

Logical CPUs	Value
EGI-InSPIRE and Council Participants	306,000
Including integrated and peer RPs	429,000

Storage	Value
Disk (PB)	155 PB
Tape (PB)	150 PB

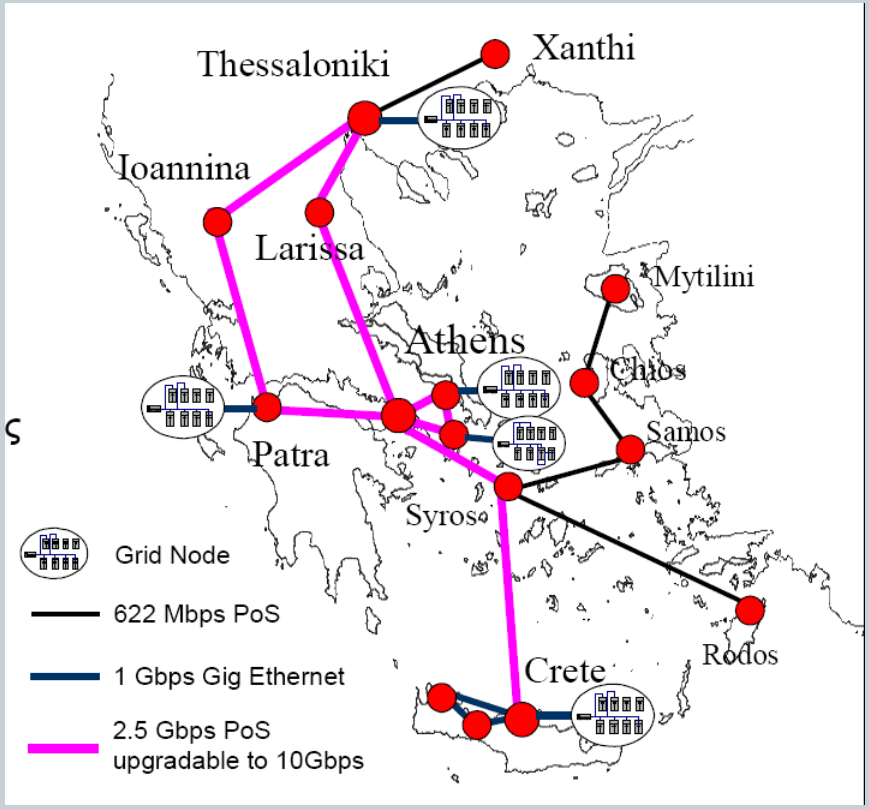


CPU Usage

Usage metrics Nov 2012		Value
CPU wall clock time	Million hour/day	50.6
Jobs	Average Job/day (Million)	1.8
Distribution of usage (main disciplines)	High-Energy Physics	88.23%
	Astronomy and Astrophysics	2.00 %
	Life Sciences	1.11%
	Remaining disciplines	8.40%

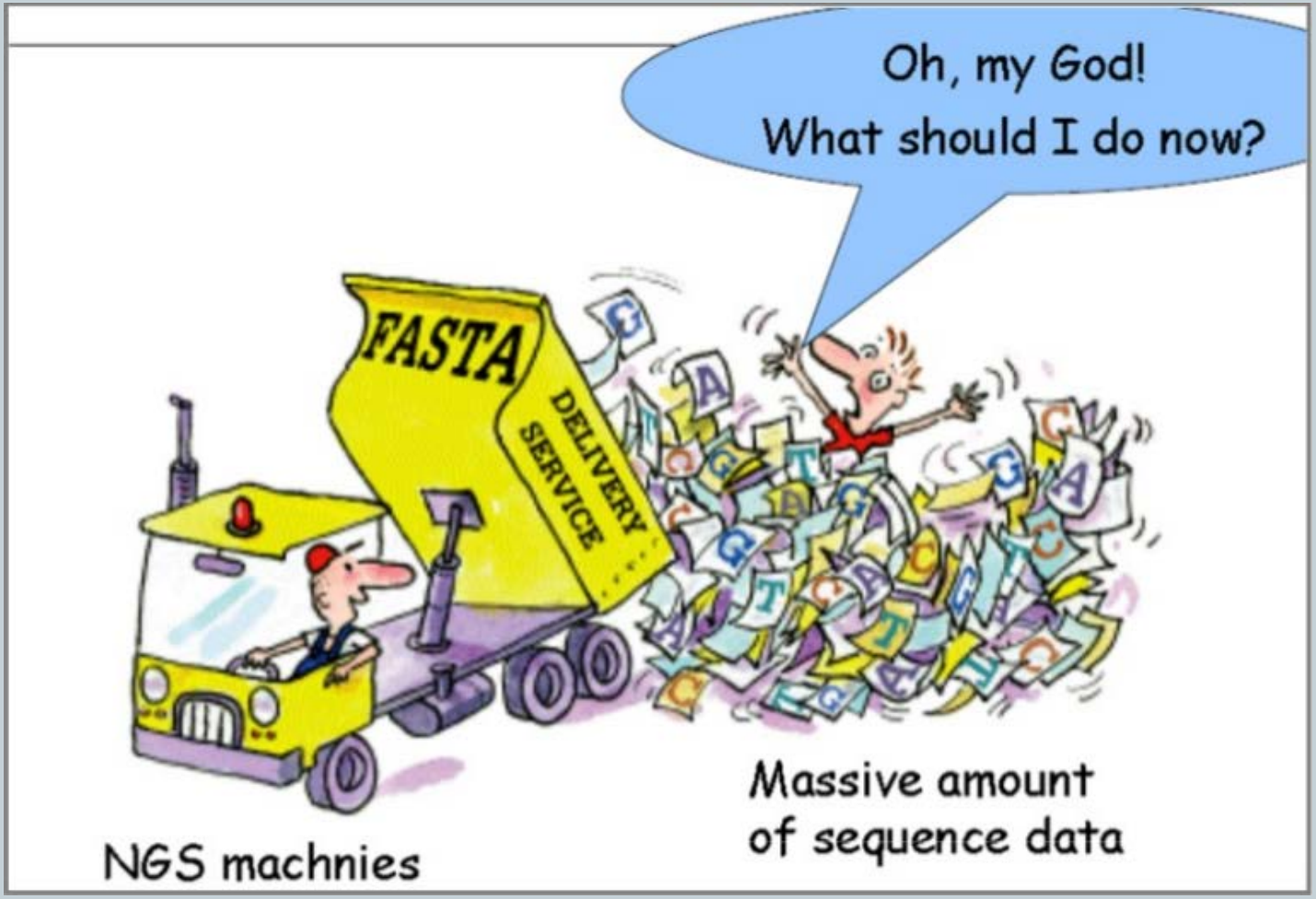
Hellasgrid

- 6 clusters of computational and storage resources
  - Athens
    - HG-01-GRNET
    - HG-02-IASA
    - HG-06-EKT
  - Thessaloniki
    - HG-03-AUTH
  - Patras
    - HG-04-CTI-CEID
  - Heraklion
    - HG-05-FORTH
- >750 Cores
- ~80 TB Storage (SE)
- ~5 TB Scratch (UIs)



More:
http://www.hellasgrid.gr/about/

… and back to the story


Oh, my God!
What should I do now?
FASTA
DELIVERY SERVICE
NGS machnies
Massive amount
of sequence data

NGS pushes bioinformatics needs up

- Need for large amount of CPU power
  - Informatics groups must manage compute clusters
  - Challenges in parallelizing existing software or redesign of algorithms to work in a parallel environment
  - Another level of software complexity and challenges to interoperability
- VERY large text files (~10 million lines long)
  - Can't do "business as usual" with familiar tools such as Perl/Python
  - Impossible memory usage and execution time
  - Impossible to browse for problems
- Need sequence Quality filtering

Data Management Issues

- Raw data are large. How long should be kept?
- Processed data are manageable for most people
  - 20 million reads (50bp) ~ 1 Gbyte
- More of an issue for a facility: HiSeq recommends 32 CPU cores, each with 4GB RAM

- Certain studies much more data intensive than others
  - Whole genome sequencing
    - A 30X coverage genome pair (tumor/normal) ~ 500 Gbyte
    - 50 genome pairs ~ 25 TB

So what?

- In NGS we have to process really big amounts of data, which is not trivial in computing terms.
- Big NGS projects require supercomputing infrastructure
- Or put another way: it's not the case that anyone can study everything.
  - small facilities must carefully choose their projects to be scaled with their computing capabilities.

Intermediate Solution #1: Cloud Computing

- Pros:
  - Flexibility
  - You pay what you use
  - Don't need to maintain a data center

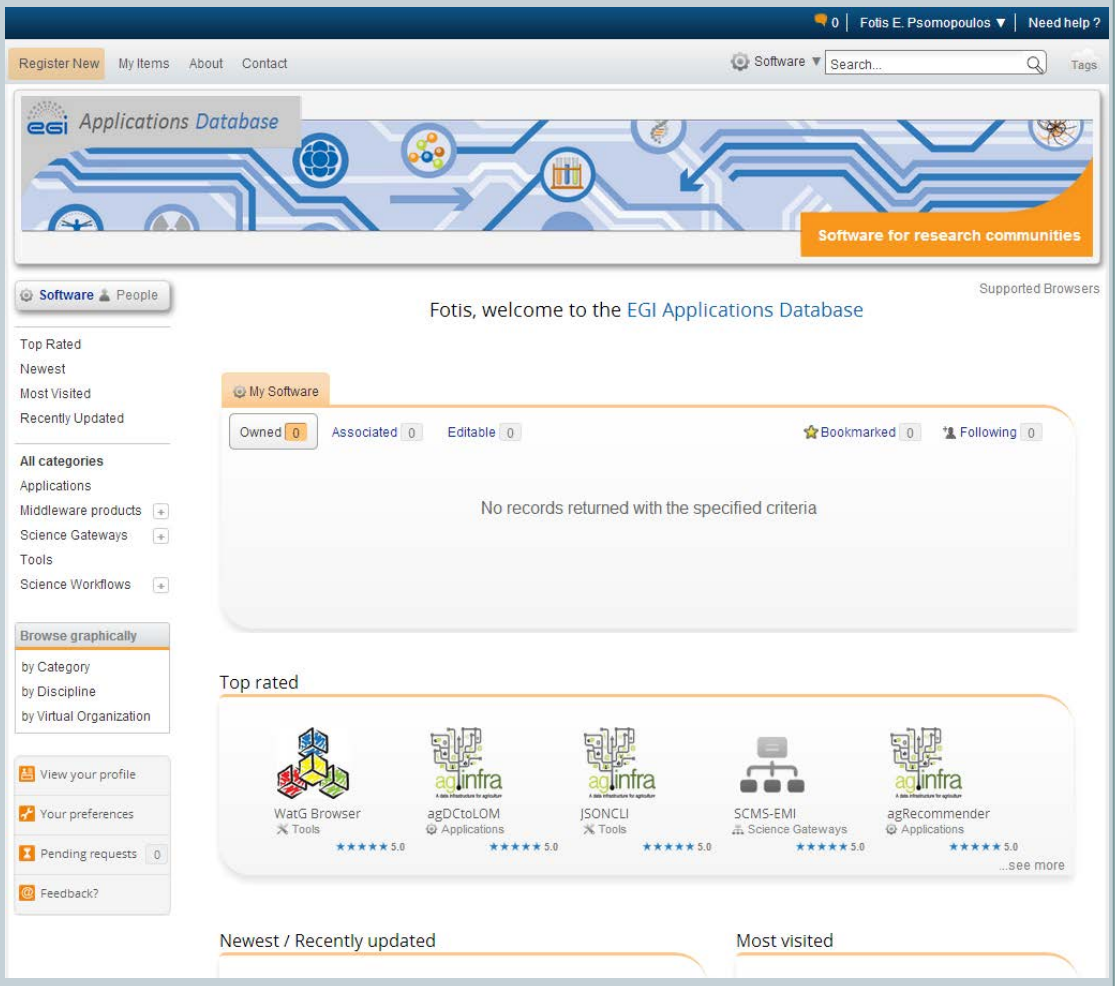
- Cons:
  - Transfer big datasets over internet is slow
  - You pay for consumed bandwidth. That is a problem with big datasets
  - Lower performance, specially in disk read/write
  - Privacy/security concerns
  - More expensive or big and long term projects

Intermediate Solution #2: Grid Computing

- Pros
  - Cheaper
  - More resources available
- Cons
  - Heterogeneous environment
  - Slow connectivity
  - Much time required to find good resources in the grid

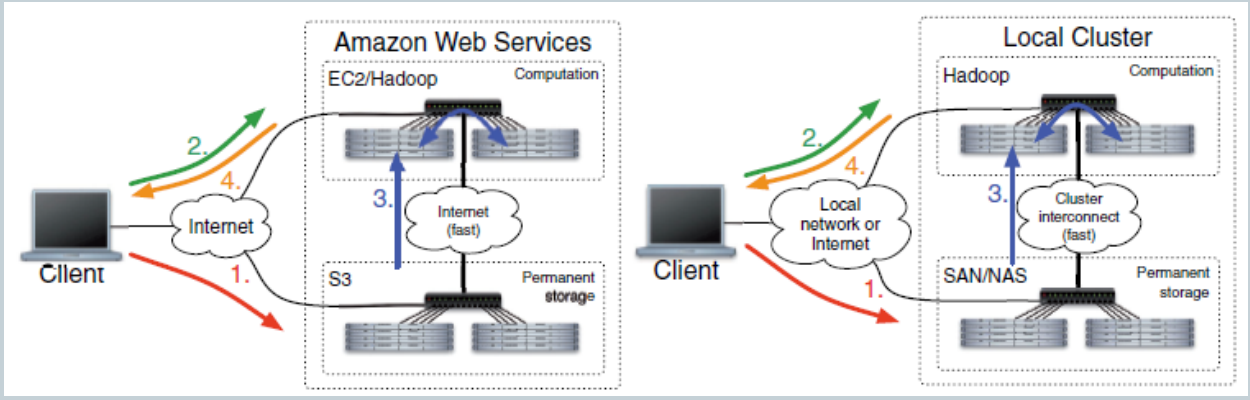
AppDB: Ready-to-use Apps in EGI

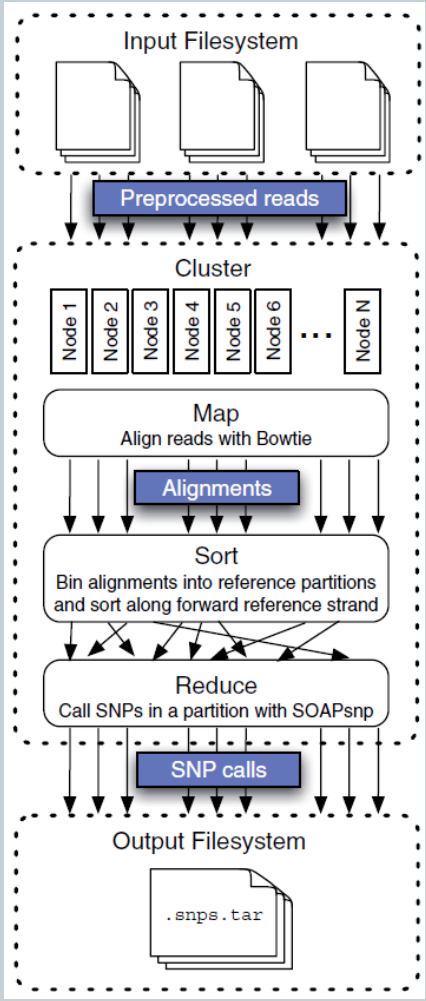
- The EGI Applications Database (AppDB) is a central service that stores and provides to the public, information about:
  - software solutions for scientists and developers to use,
  - the programmers and the scientists who developed them, and
  - the publications derived from the registered solutions



Crossbow

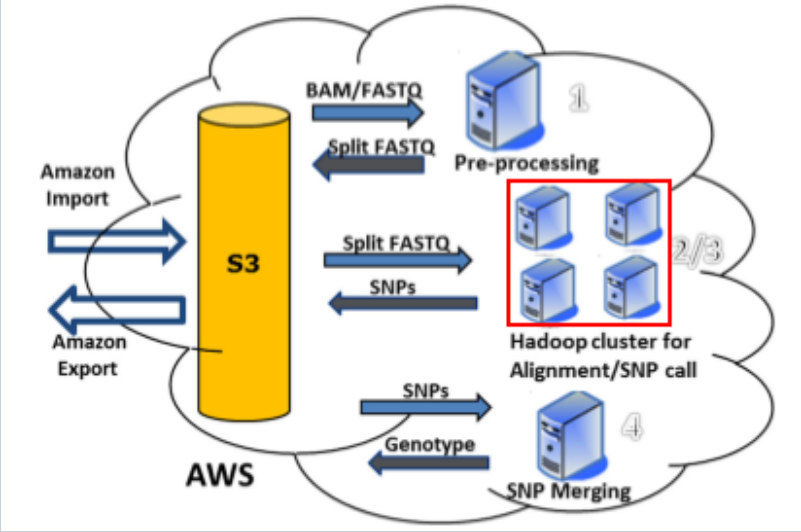
- Identifies SNPs from high-coverage, short-read resequencing data
- Combines the Aligner Bowtie and the SNP caller SOAPsnp
- Hadoop MapReduce approach
- Amazon EC2 / Local Cluster





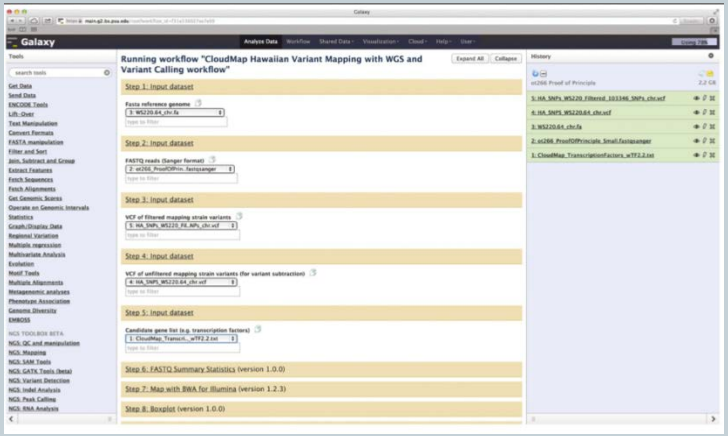
Rainbow

- Large scale Whole Genome Sequencing (WGS) analysis
- Supports FASTQ and BAM input
- Load balancing
- Active workflow monitoring
- Amazon EC2



CloudMap

- Greatly simplifies the analysis of mutant whole genome sequences
- Offers predefined workflows to pinpoint variations in animal genomes
- Available on the Galaxy web platform
- Amazon EC2 / Local Cluster

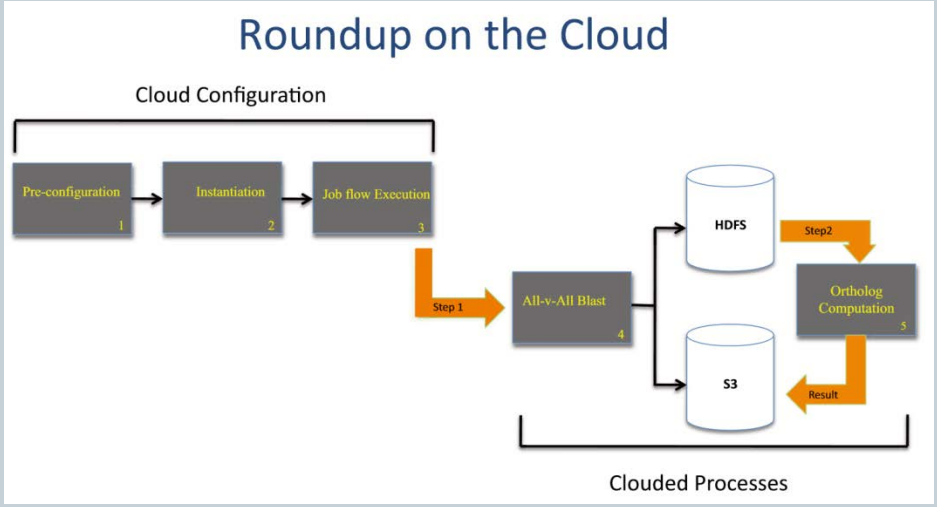


CloudBurst

- Parallel read-mapping algorithm optimized for mapping NGS data to the human and other reference genomes
- Modeled after the short read-mapping RMAP program
- Parallelization overcomes computational barriers and allows deeper analysis
- Hadoop MapReduce approach
- Almost linear increase in performance to the number of CPU cores available

RSD-Cloud

- Large comparative genomics analysis tool
- Redesigned the reciprocal smallest distance algorithm (RSD) to run on a cloud computing environment
- Fast and cost efficient solution
- Amazon EC2



Cloud BioLinux

- Publicly accessible VM
- Platform for developing bioinformatics infrastructures on the cloud
- Quick provision of on-demand infrastructures for HPC in bioinformatics
- Pre-configured tools and GUI
- Tested on Amazon EC2, Eucalyptus, Okeanos and Virtual box

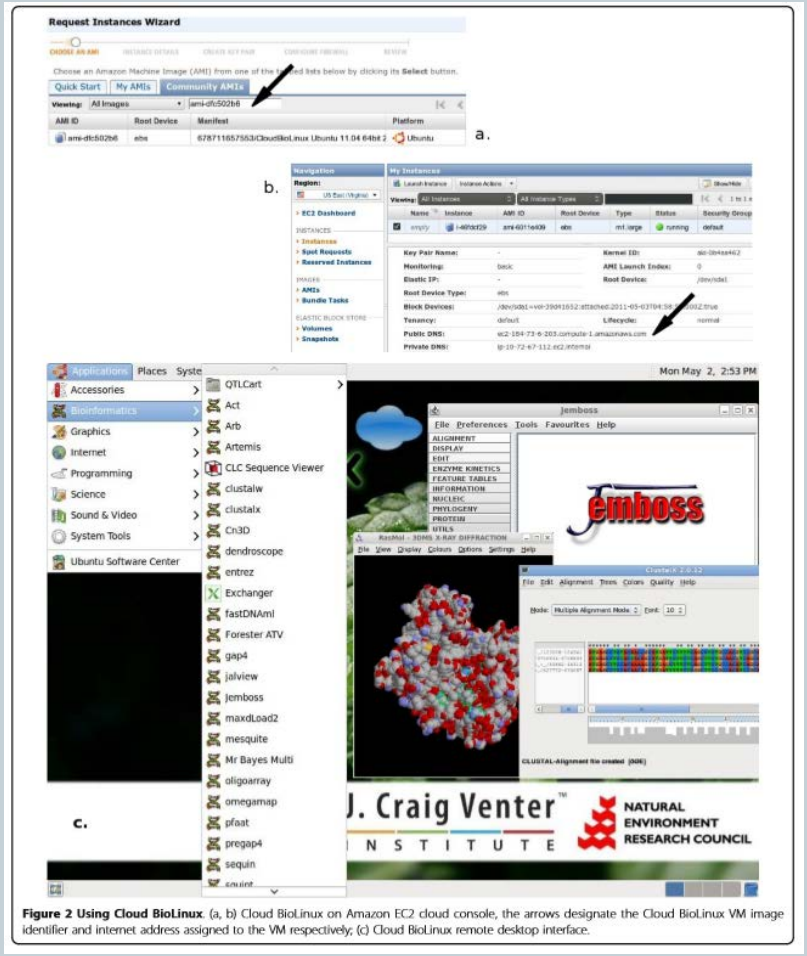
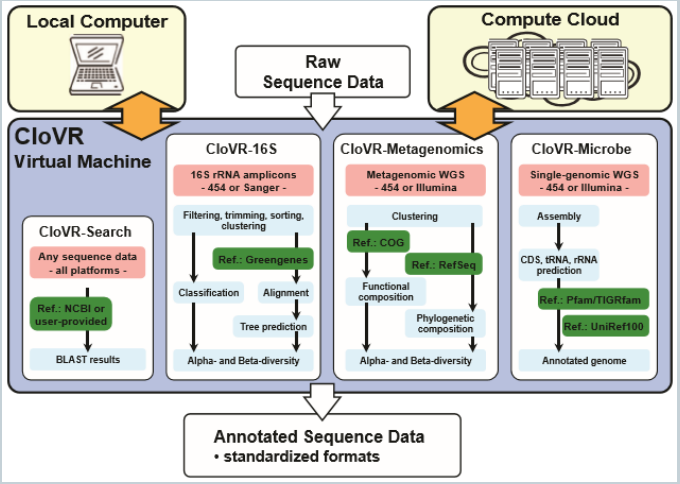


Figure 2 Using Cloud BioLinux. (a, b) Cloud BioLinux on Amazon EC2 cloud console, the arrows designate the Cloud BioLinux VM image identifier and internet address assigned to the VM respectively; (c) Cloud BioLinux remote desktop interface.

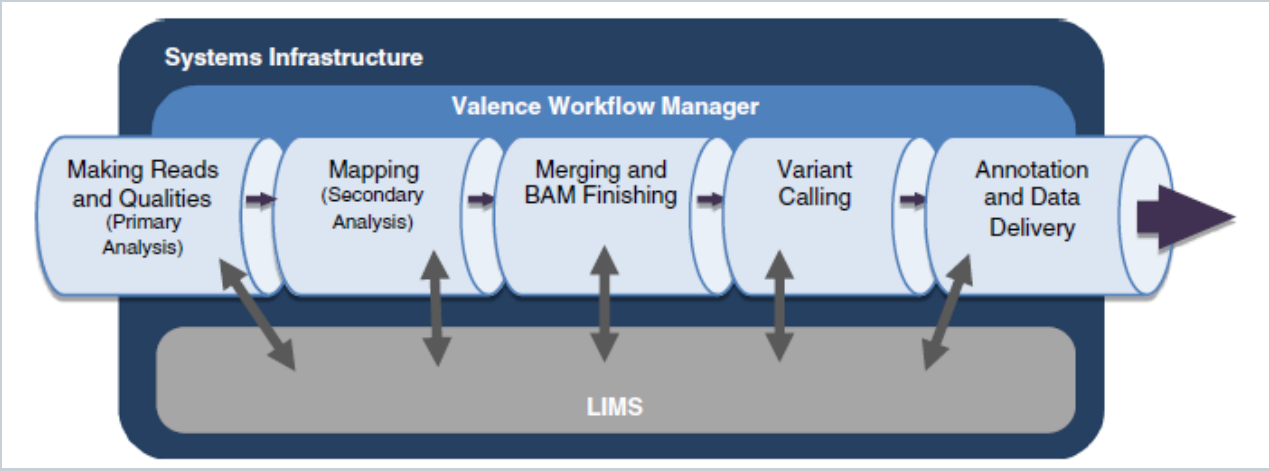
CloVR

- Portable VM
- Several automated analysis pipelines for microbial genomics provided, including 16S, whole genome and metagenome sequence analysis
- Run on a local PC but also supports use of remote cloud computing resources on multiple cloud computing platforms.



Mercury

- Integration of multiple sequence analysis tool in a single DNAnexus based platform
- Simplified workflow construction GUI
- Applet based workflows
- Amazon EC2 / Local Cluster



Taverna

- General purpose open source and domain-independent Workflow Management System
- Combines distributed web services and local tools into complex analysis pipelines.
- Execution takes place either locally or in a grid or cloud environment using the Taverna server
- Widely adopted in bioinformatics workflows, typically in the areas of high throughput omics analyses like proteomics, transcriptomics and evidence gathering methods involving text or data mining.

Galaxy

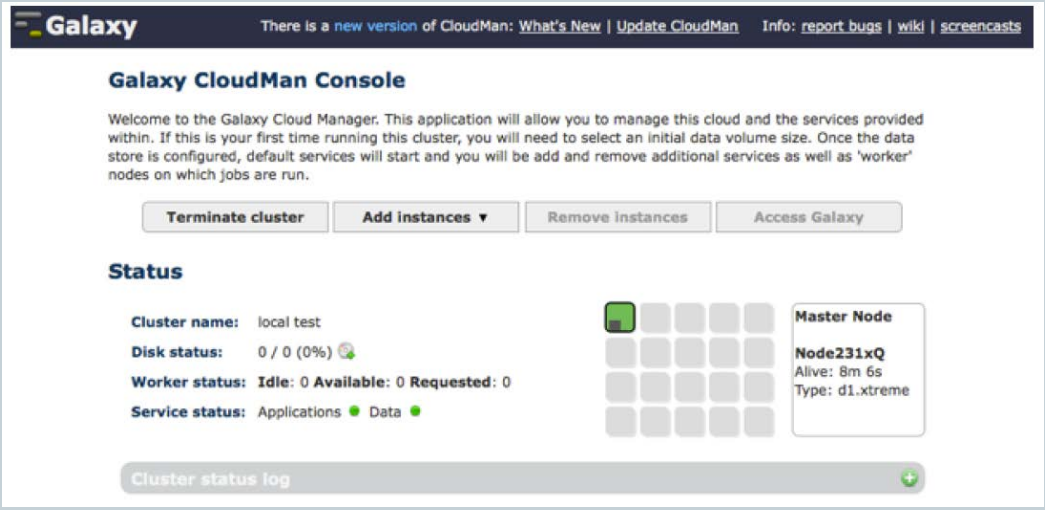
- Offers genome analysis resources for cloud computing platforms
  - Amazon EC2
  - Virtual Box
  - Eucalyptus
  - Okeanos
- Freely available and community maintained
  - software images and
  - data repositories
- Widely adopted in the bioinformatics community

Tavaxy

- Stand alone pattern based workflow system
- Integrates the use of Taverna and Galaxy workflows in a single environment
- Hierarchical workflows and workflow patterns approach
- Utilization of local and cloud HPC resources
- Currently centered on sequence analysis. Transcriptomics and proteomics analyses in development.

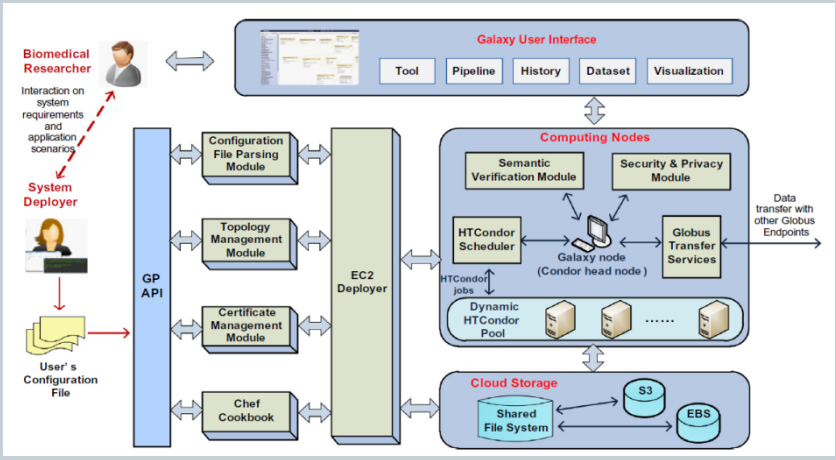
CloudMan

- Custom deployment of cloud resources
- Browser creation and control of an arbitrarily sized compute cluster on Amazon EC2
- Minimal informatics experience required for use
- Large number of tools available, packaged by the NERC Bio-Linux team.
- Based on Galaxy



Bioinformatics Cloud Platform

- Large scale NGS analyses platform
- Based on Galaxy
- Data management capabilities through Globus Transfer
- Automatic cloud deployment and on demand resource allocation
- Cloud provisioning and auto scaling through the HTCondor scheduler
- Supports external clusters



The take-home points...

- Life Sciences and Big Data are irrevocably linked
- A lot of Life Sciences infrastructure projects (ELIXIR, LifeWatch etc) are already looking towards Grid/Cloud solutions
- Although techniques are here to stay, there is a narrow window of opportunity for researchers to stay ahead of the curve
- If interested, do ask for more... ☺

Thank you for your patience!

PERSPECTIVE ARTICLE

Front. Genet., 23 June 2015 | http://dx.doi.org/10.3389/fgene.2015.00197

Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis

Afonso M. S. Duarte[1*†], Fotis E. Psomopoulos[2,3*†], Christophe Blanchet[4], Alexandre M. J. J. Bonvin[5], Manuel Corpas[6], Alain Franc[7], Rafael C. Jimenez[8], Jesus M. de Lucas[9], Tommi Nyrönen[10], Gergely Sipos[11] and Stephanie B. Suhr[12]