Complex Computing and Data Infrastructure Challenges: the Biology Case

e-IRG Workshop, Copenhagen, Jun 11, 2012

Kristoffer Rapacki

CENTERFOR BIOLOGICAL SEQUENCE ANALYSIS CBS

The 3 main points:

Paradigm shift in biology

Need for new infrastructure

Current effort to provide it

Paradigm shift in biology

From gene-by-gene analysis

to

all genes (Homo sapiens ~ 25,000)

to

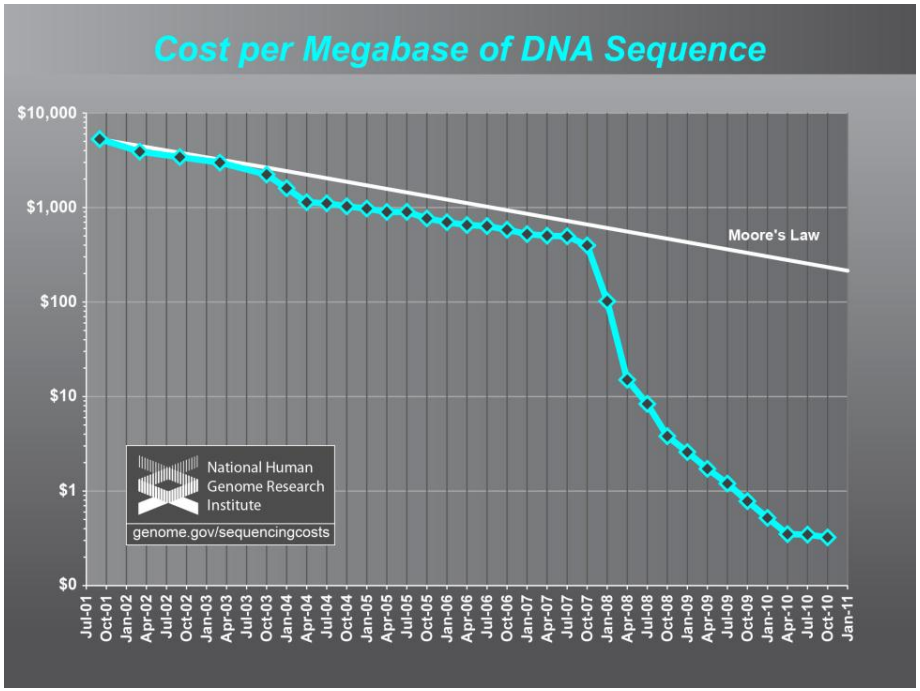
to entire genome (> 500,000 functional loci)

to

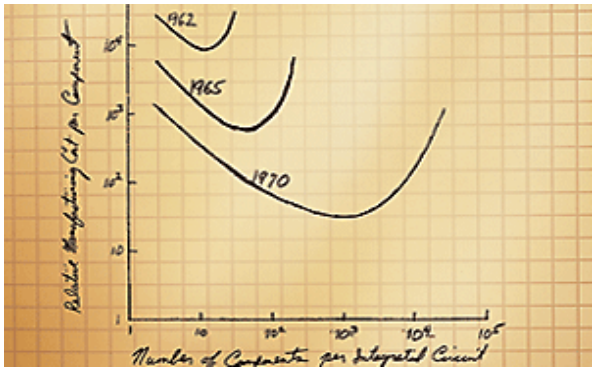
entire population (millions of genomes)

The planet is being sequenced

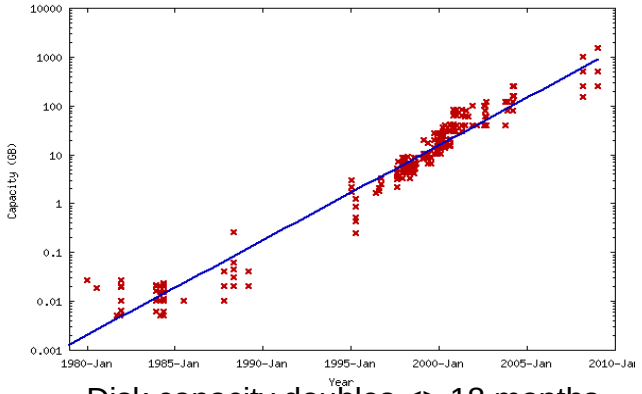
- UK10K
- Personal Genome Project 16K -> 100K
- Nurses health
- FarGen 100 -> 50K
- Dutch Genome Project
- Danish Genome Project
- ...
- The Cancer Genome Project
- ...
- The human microbiome project
- ...



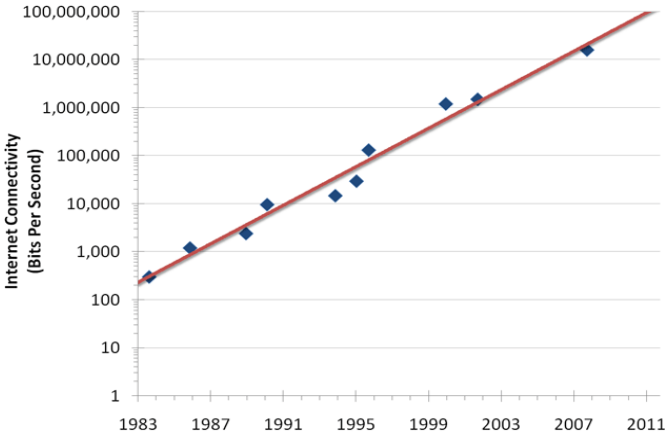
Data growth exceeds growth in IT capability



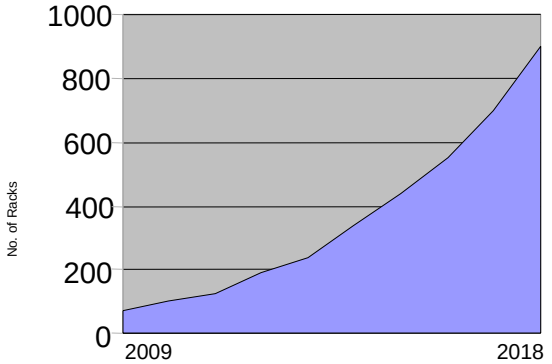
CPU Power doubles <> 24 months (Moores Law)



Disk capacity doubles <> 18 months

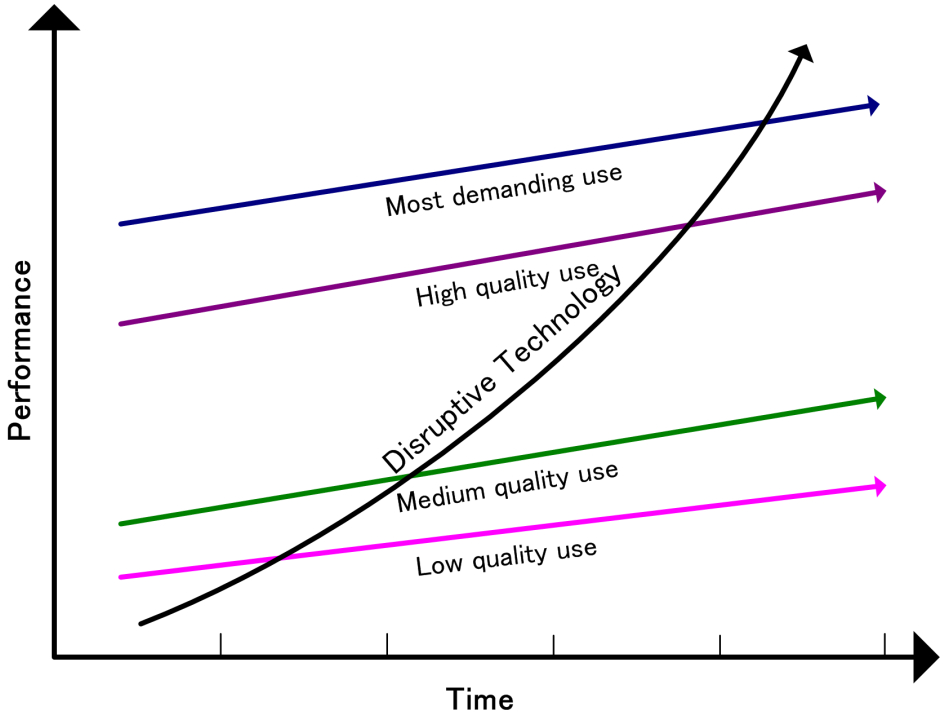


Network bandwidth doubles <> 20 months



Racks in EBI machine room double <> 12 months

Disruptive technologies



"A technology becomes disruptive when the rate at which it improves exceeds the rate at which users can adapt to the new performance."

The Innovator's Dilemma. Clayton M. Christensen. Harvard Press. 1997

Novel types of infrastructure needed!

Disruptive technologies in biology

- Next-generation DNA sequencing
  - Data will be 1,000 <> 1,000,000 times cheaper to produce
  - Data production rates will be 1,000 <> 1,000,000 more by the end of the ESFRI period.
- Protein sequencing by Mass Spectrometry may also be disruptive
- There will probably be others
  - Macromolecular structure determination by Electron Microscopy
  - Imaging of various kinds
  - etc

Exponential growth in data…

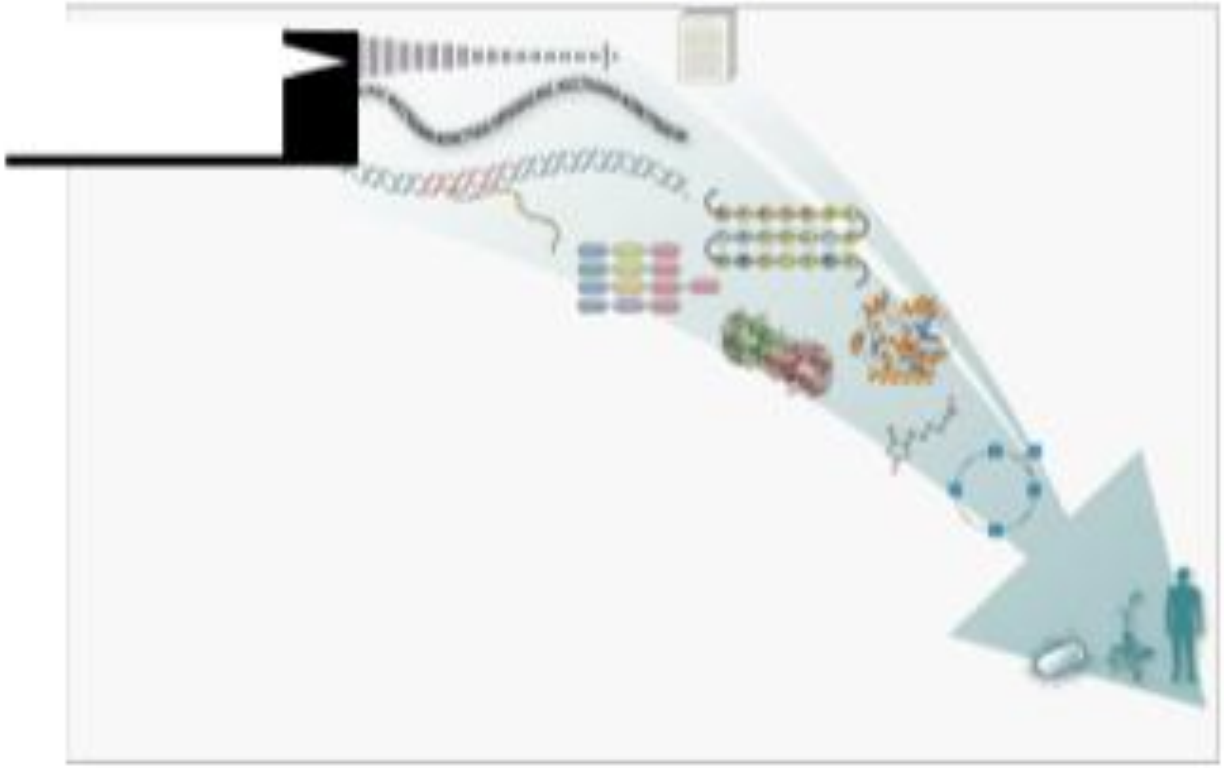
Cannot equate to exponential growth in funding, so

1. Link budgets for data generation and data processing
   - Only produce as much data as you can deal with

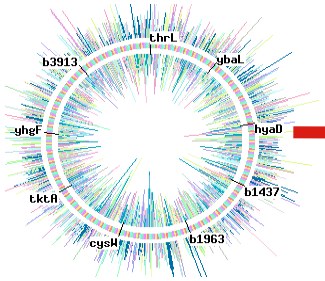
2. Take steps to control staff growth
   - Automation of annotation and curation
   - Implement distributed annotation (DAS)
   - Use web services and distributed resources
   - Support for metadata deposition

3. Take steps to control IT resource requirements.
   - Develop policies for which data are to be kept (& which not)
   - Develop data compression techniques

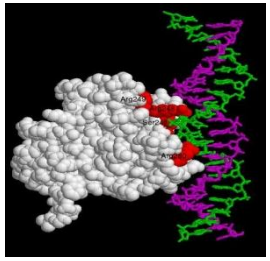
Enormous data diversity



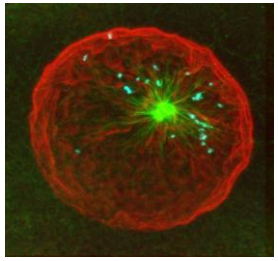
Modern biology requires data integration



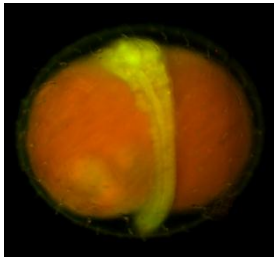
Genome



Protein



Cell



Embryo

Fruitfly

Mouse

Development, Ageing, Disease

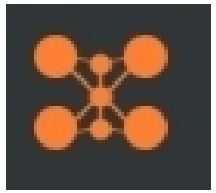
The 3 main points:

Paradigm shift in biology

Need for new infrastructure

Current effort to provide it

What is Elixir?

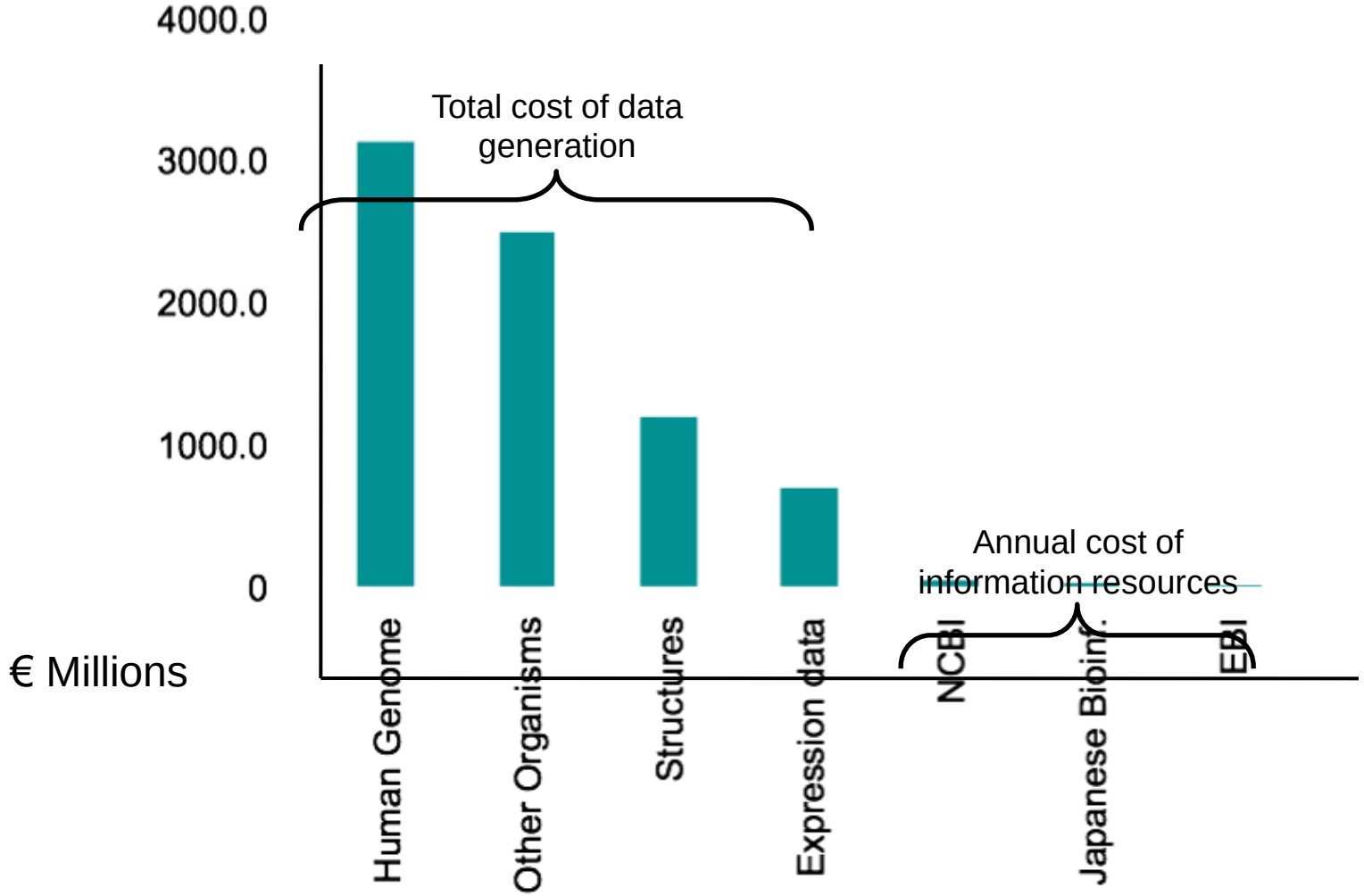


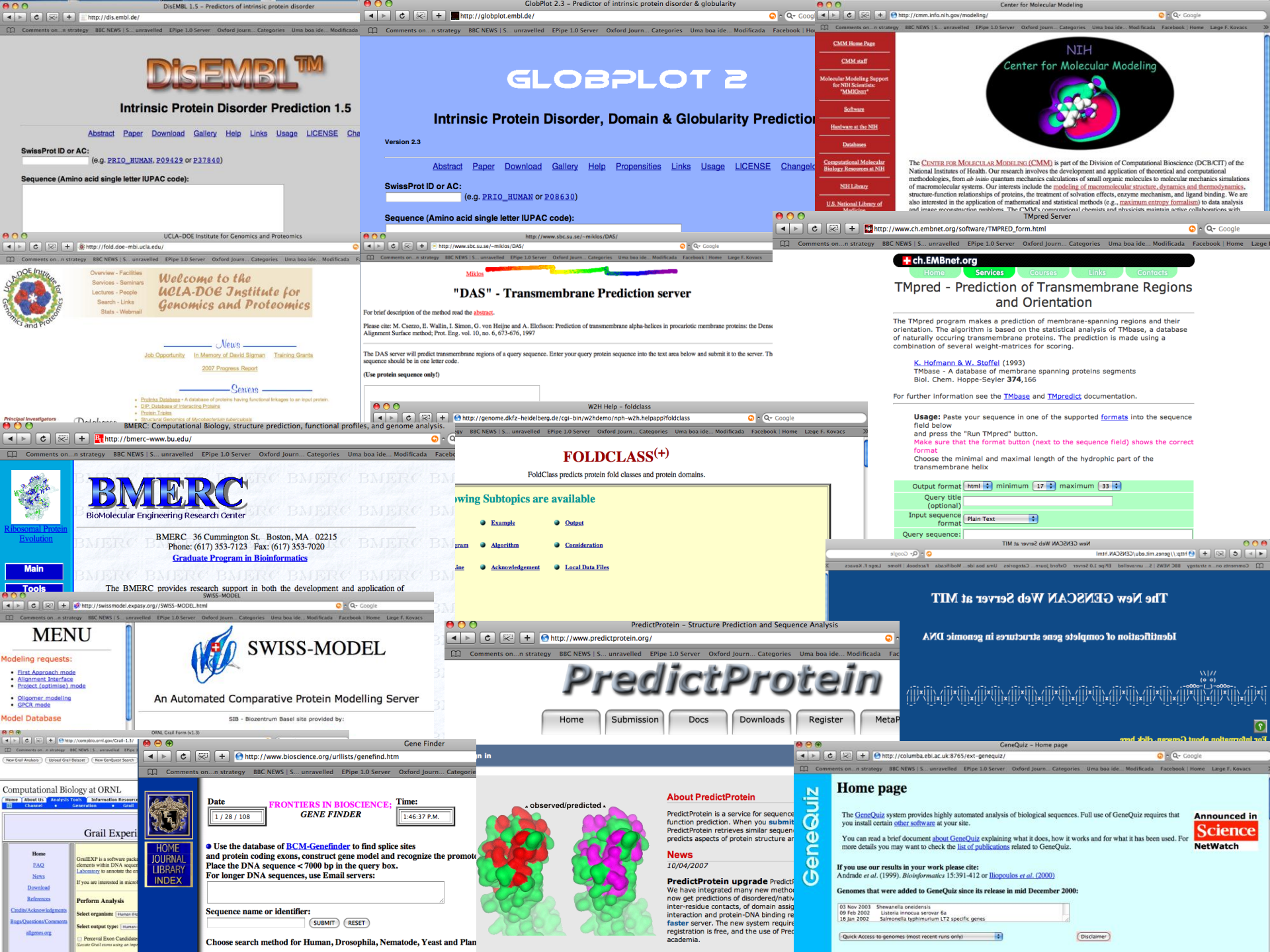
- 1 of 35 initial phase ESFRI projects
- Started as an EU FP7 Preparatory Phase Project
- ELIXIR, in its interim form led by Prof. Søren Brunak, Denmark, and coordinated by Prof. Janet Thornton, Director EMBL-EBI, UK
- To construct a plan for the operation of a sustainable infrastructure for biological information in Europe
- 32 member consortium engaging many of Europe's main bioinformatics funding agencies and research institutes
- Deliverables are memoranda of understanding to fund the implementation phase which could cost €500 million
- More than 10 countries have already made committments, including the UK, Sweden and Denmark
- International Consortium Agreement being made for ELIXIR
- Interested parties should register as stake-holders via the ELIXIR Website: www.elixir-europe.org

ELIXIR Objectives

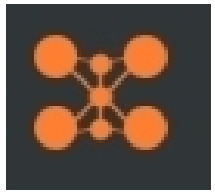
- to secure funding commitments from government agencies, charities, industry and intergovernmental organisations throughout Europe, to strengthen and sustain a world-class infrastructure for the management and integration of information in the life sciences.

Good value for money




DisEMBL™
Intrinsic Protein Disorder Prediction 1.5
GLOBPLOT 2
Intrinsic Protein Disorder, Domain & Globularity Prediction
NIH
Center for Molecular Modeling
"DAS" - Transmembrane Prediction server
TMpred - Prediction of Transmembrane Regions and Orientation
BMERC
BioMolecular Engineering Research Center
FOLDCLASS(+)
SWISS-MODEL
An Automated Comparative Protein Modelling Server
PredictProtein
The New GENSCAN Web Server at MIT
Home Page
Home
Services
Courses
Links
Contacts
Main
Tools
Home
Submission
Docs
Downloads
Register

Elixir Tool integration marketing bait



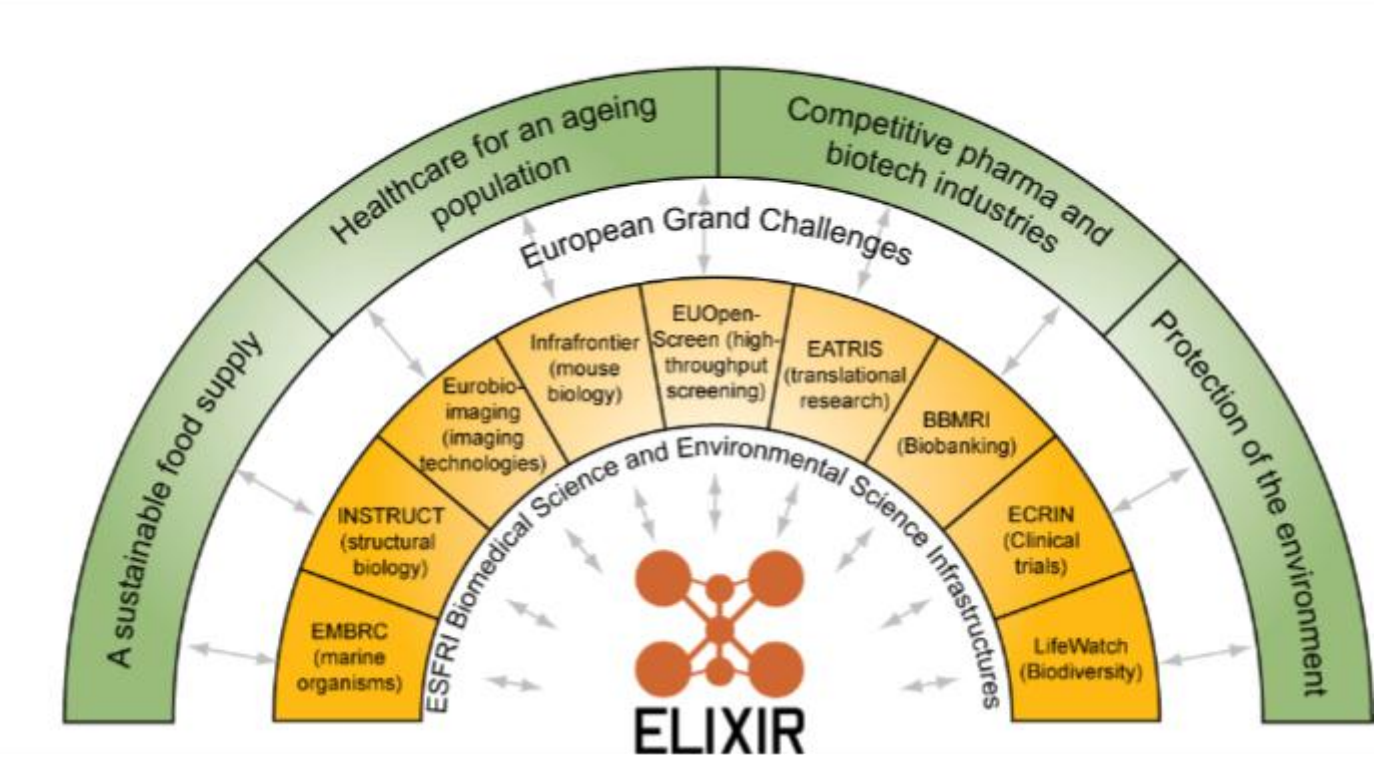
”Parsed on arrival”

ELIXIR Work Packages.

Elixir is organised into 14 work packages which have committees of (mainly) European experts associated with them. It is organising two surveys, one of users and one of data-providers, and five technical-feasibility studies. The Elixir Steering Committee is associated with WP1 and has oversight of the whole project. WP3 has four committees; for Bioinformatics Communities, for Data Providers, for Industry and for Interactions with the rest of the World (International). There will be regular Stakeholder meetings intended to encourage the widest possible participation.

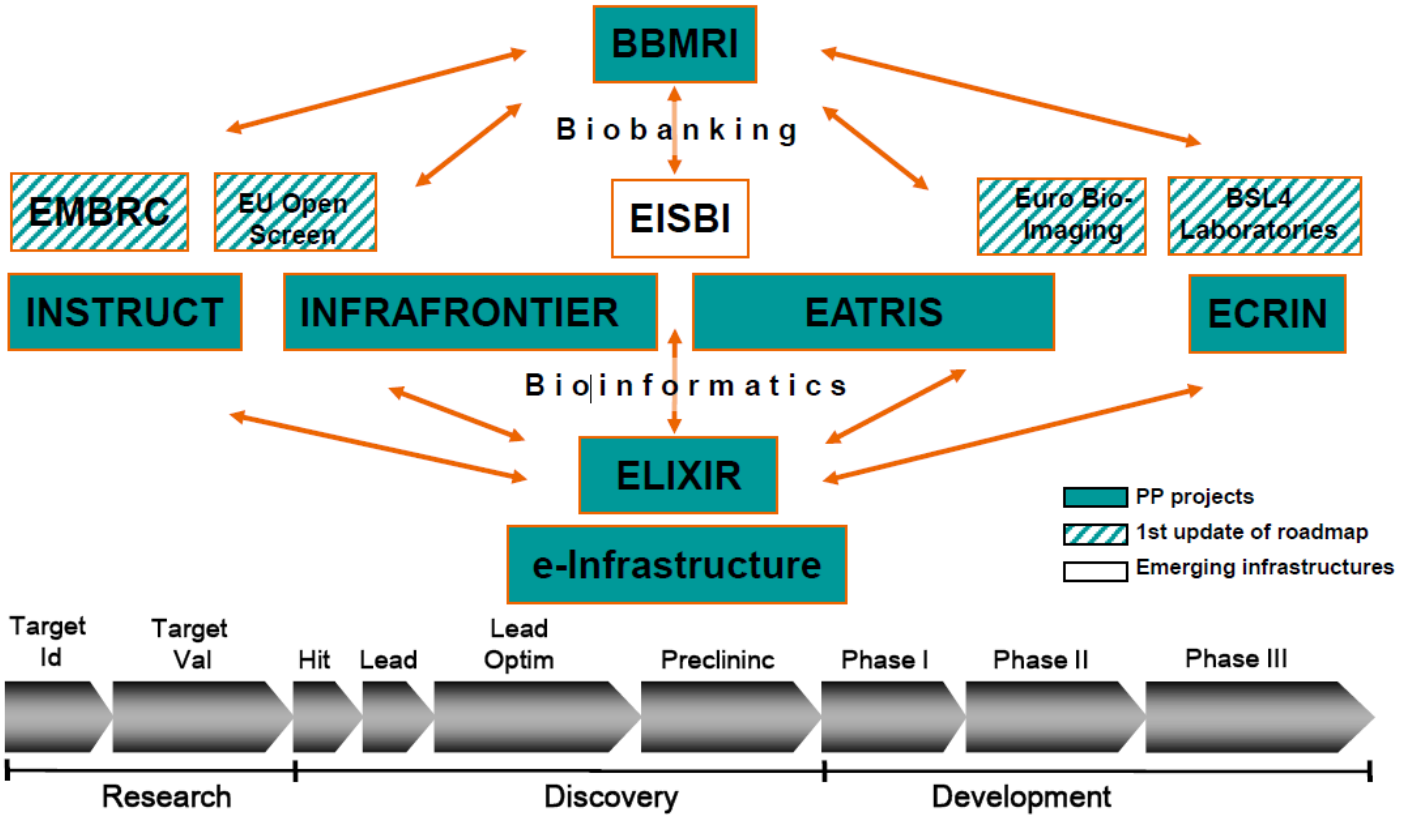
1. Project management
2. Data providers
3. User communities
4. Organisation and Legal
5. Funding
6. Physical infrastructure
7. Data interoperability
8. Literature
9. Healthcare
10. Chemistry & Environment
11. Training
12. Tools integration
13. Feasibility studies
14. Reporting and negotiation

ELIXIR Support of the European Grand Challenges

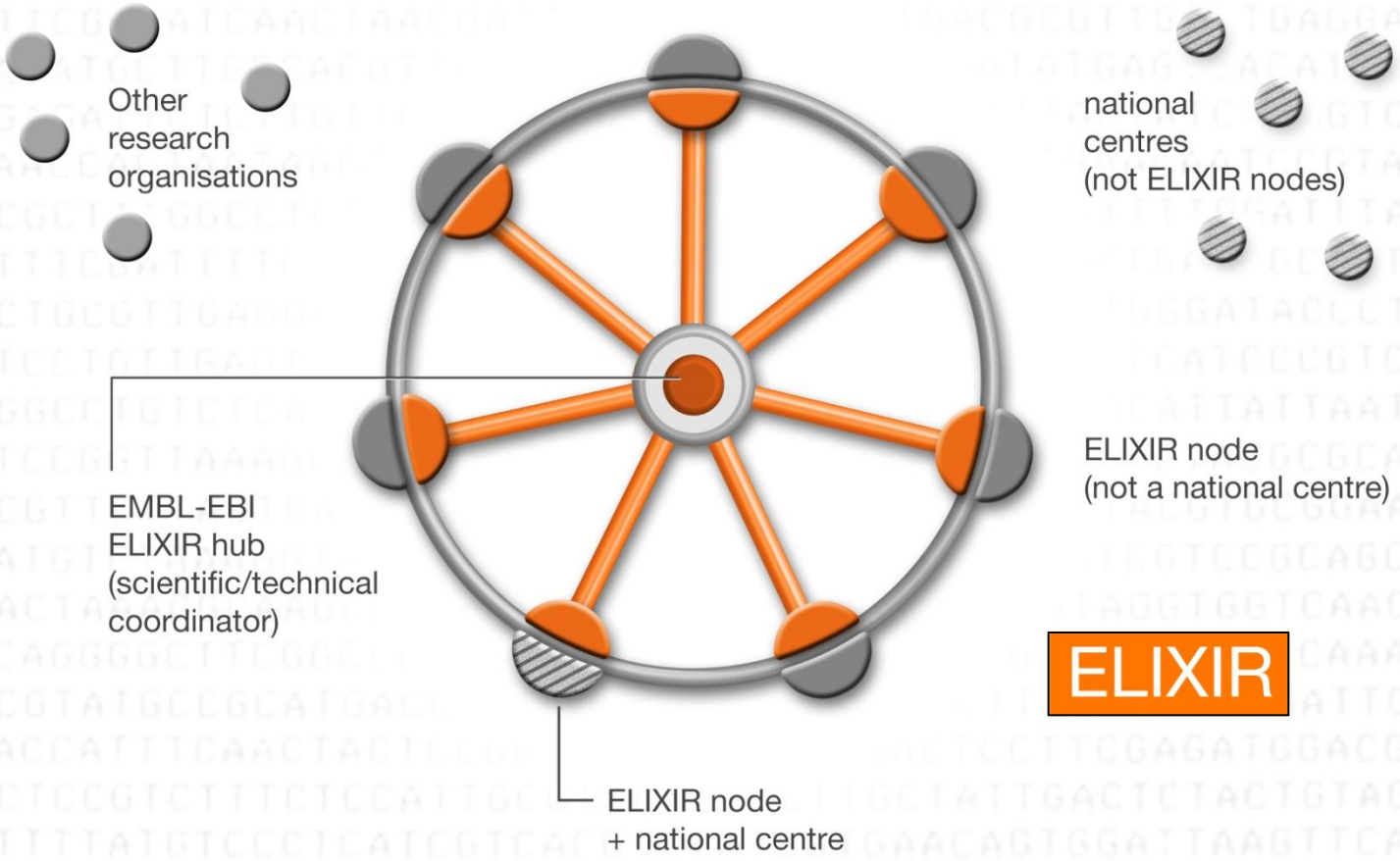


ELIXIR supports the European Grand Challenges by providing Infrastructure for the other ESFRI Biology Projects.

Synergies between BMS RI

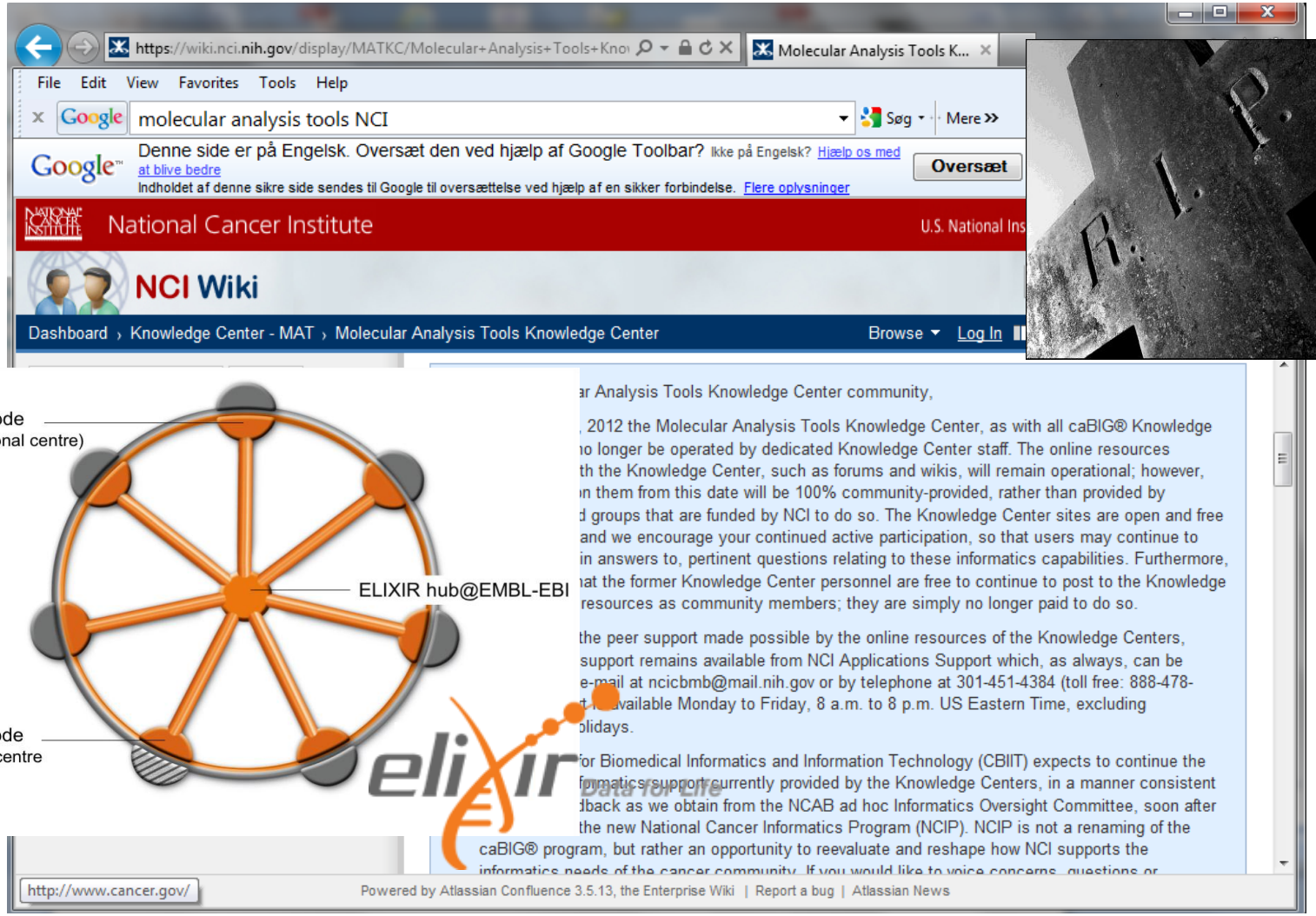


ELIXIR Scientific & Technical Structure

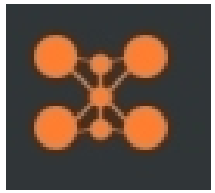


Bottom-up versus top-down

- caBIG infrastructure terminated



Paradigm shift in biology



From gene-by-gene analysis

to

all genes (Homo sapiens ~ 25,000)

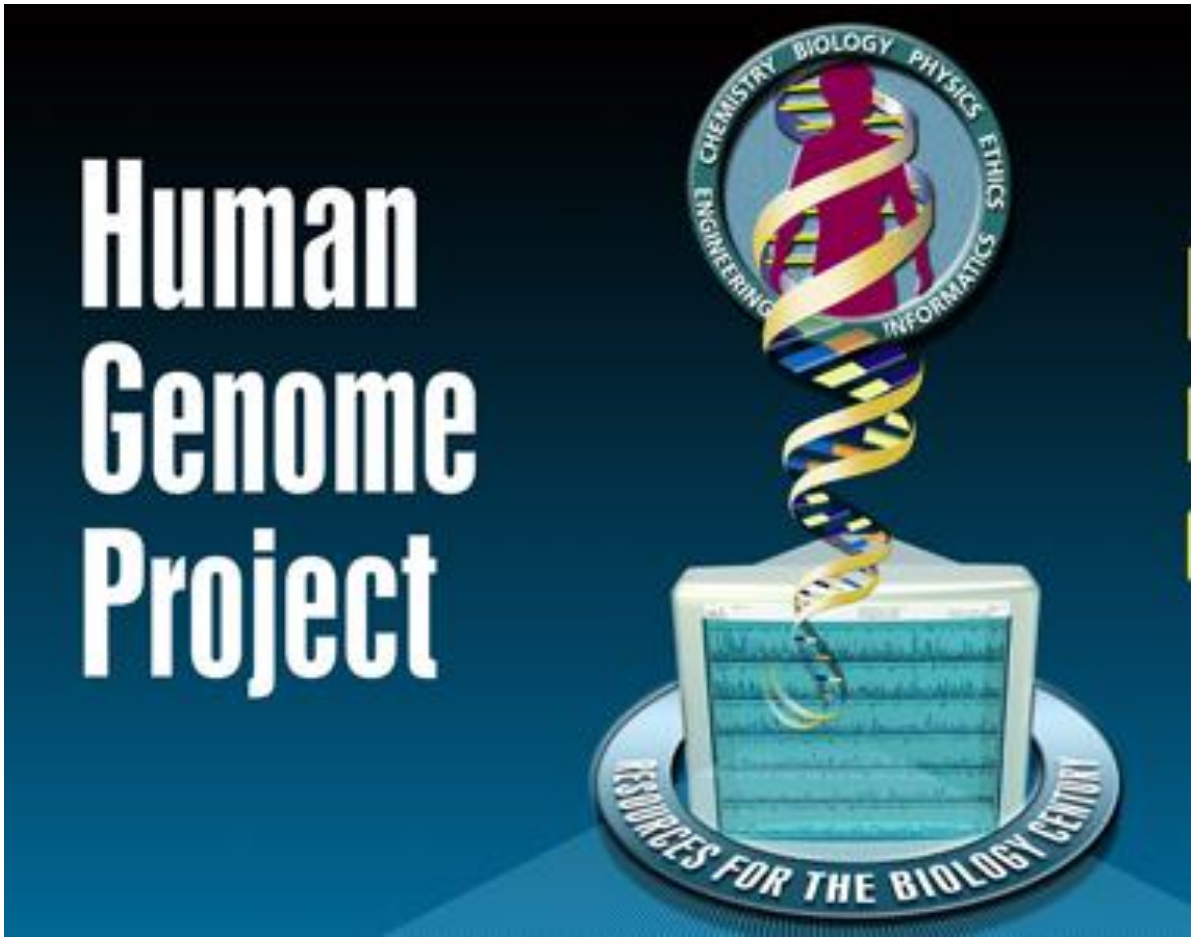
to

to entire genome (> 500,000 functional loci)

to

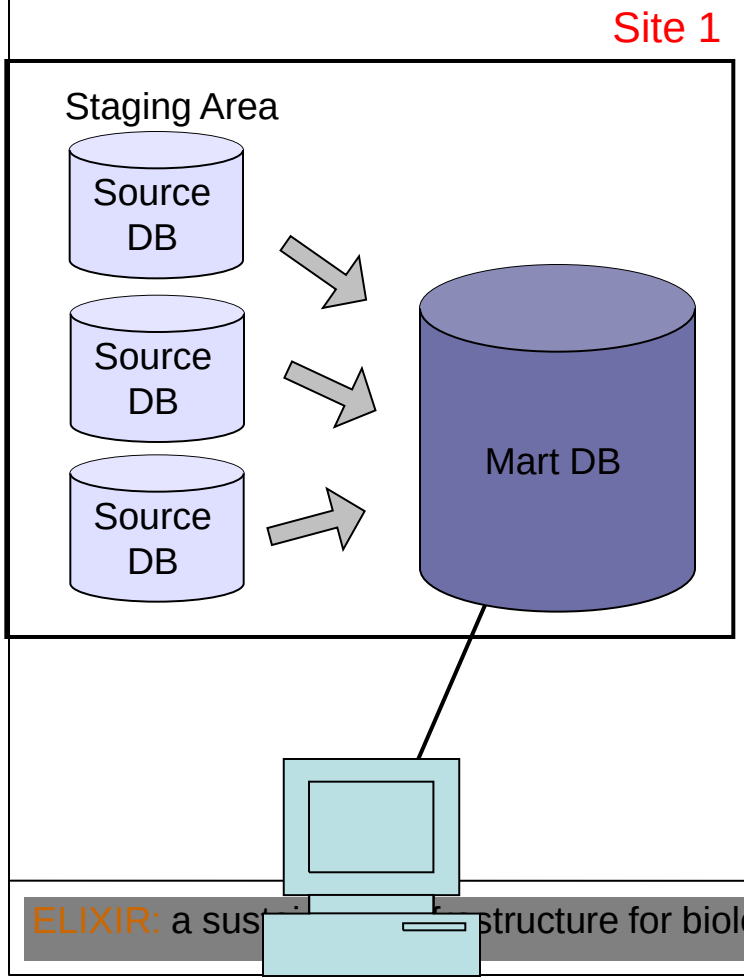
entire population (millions of genomes)

Data security! Person-sensitive!

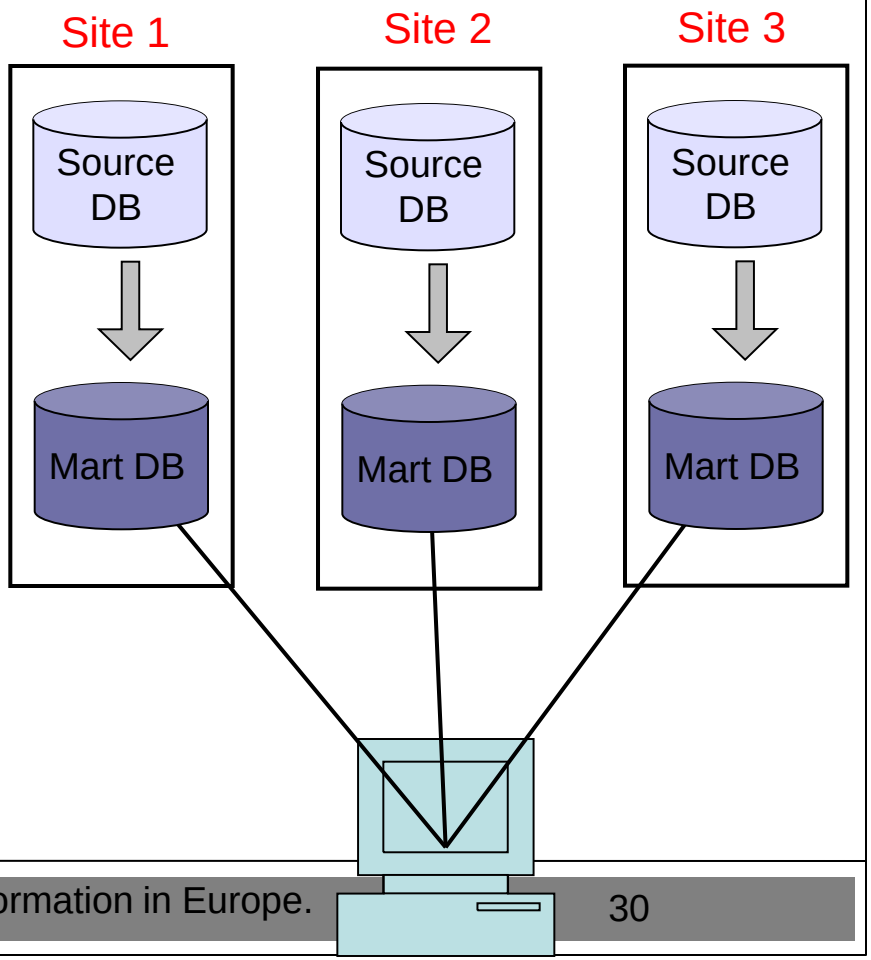


Federate or Centralize?

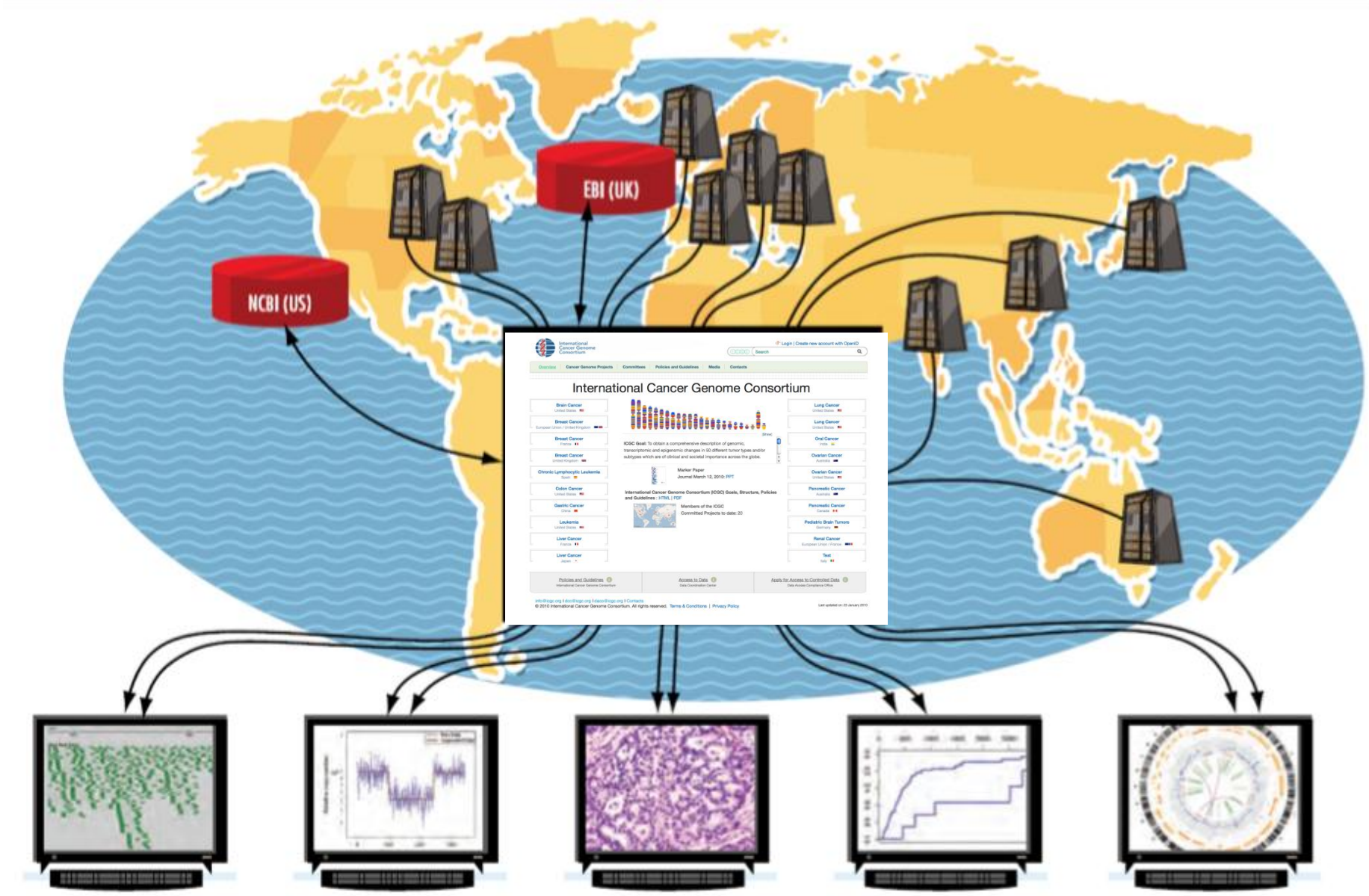
Centralized Model



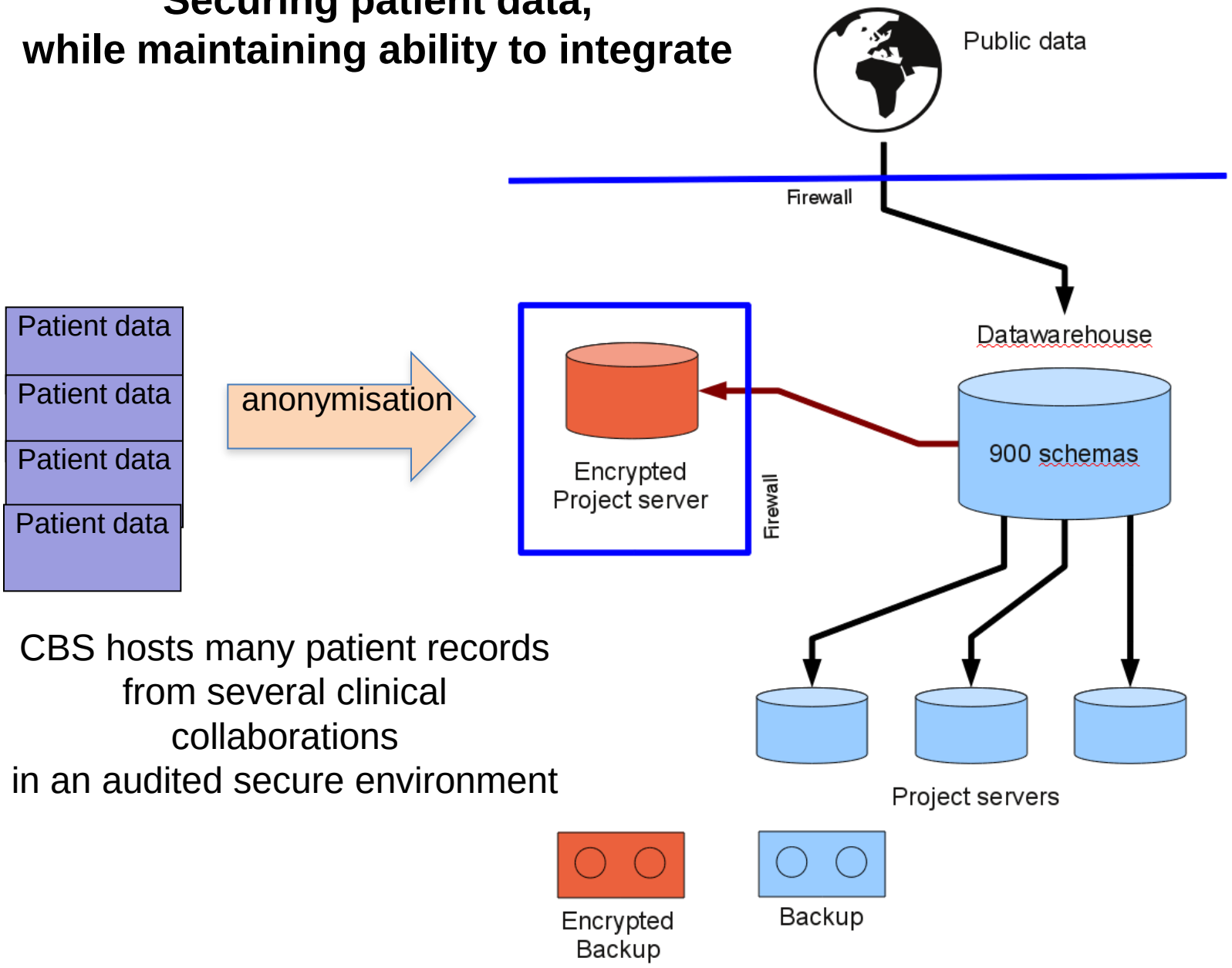
Federated Model



ICGC data is distributed, but is accessible through a common portal

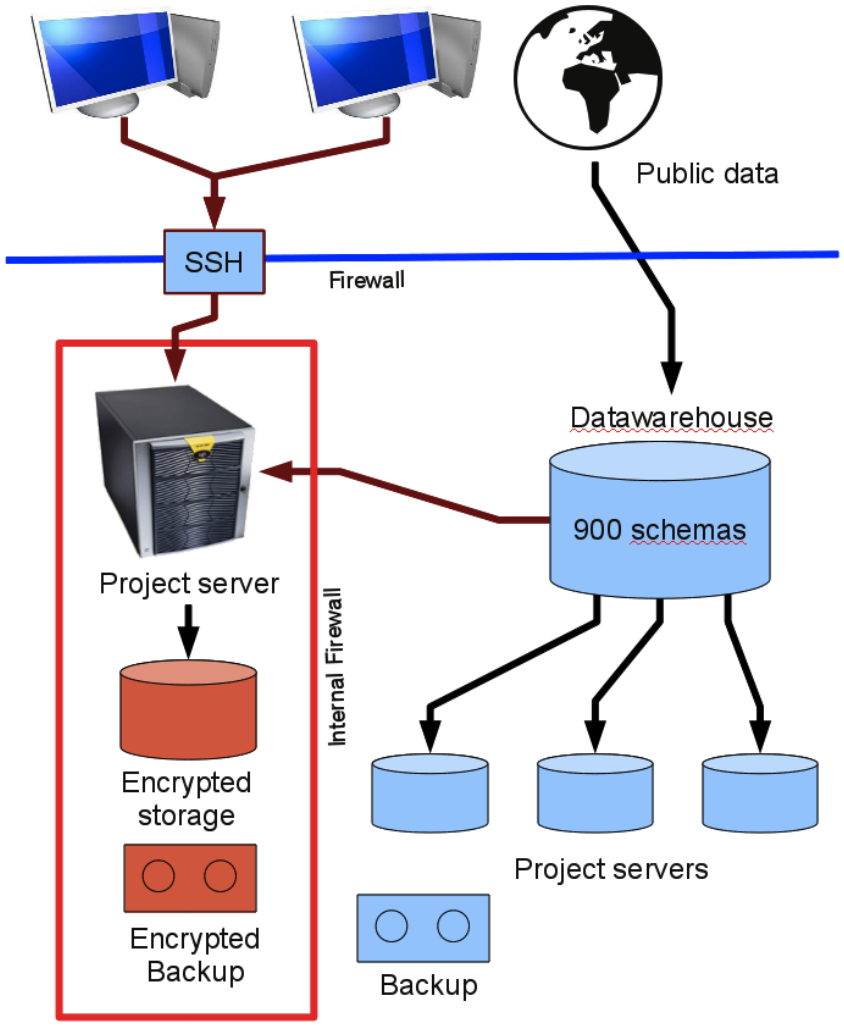


Securing patient data, while maintaining ability to integrate



CBS hosts many patient records
from several clinical
collaborations
in an audited secure environment

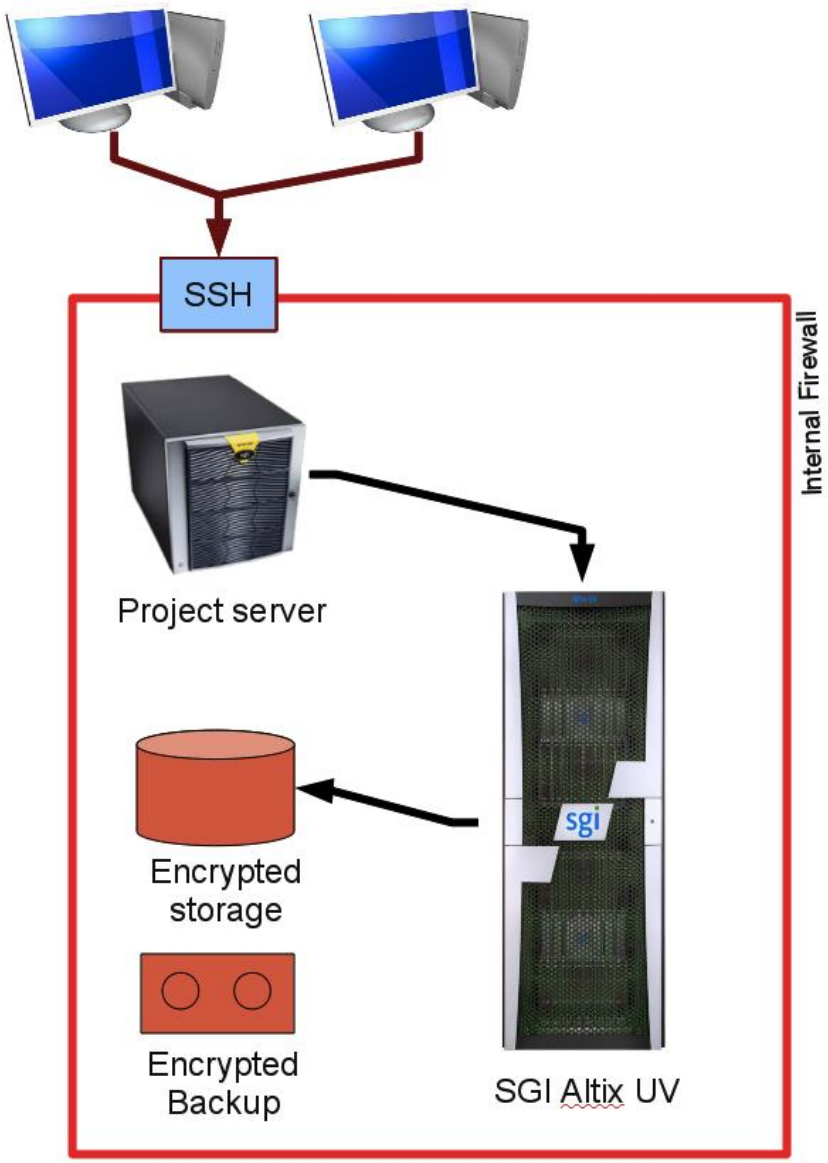
Secure Staging and Analysis



SSH/VPN data submissions to secure data server, or via secured couriers..

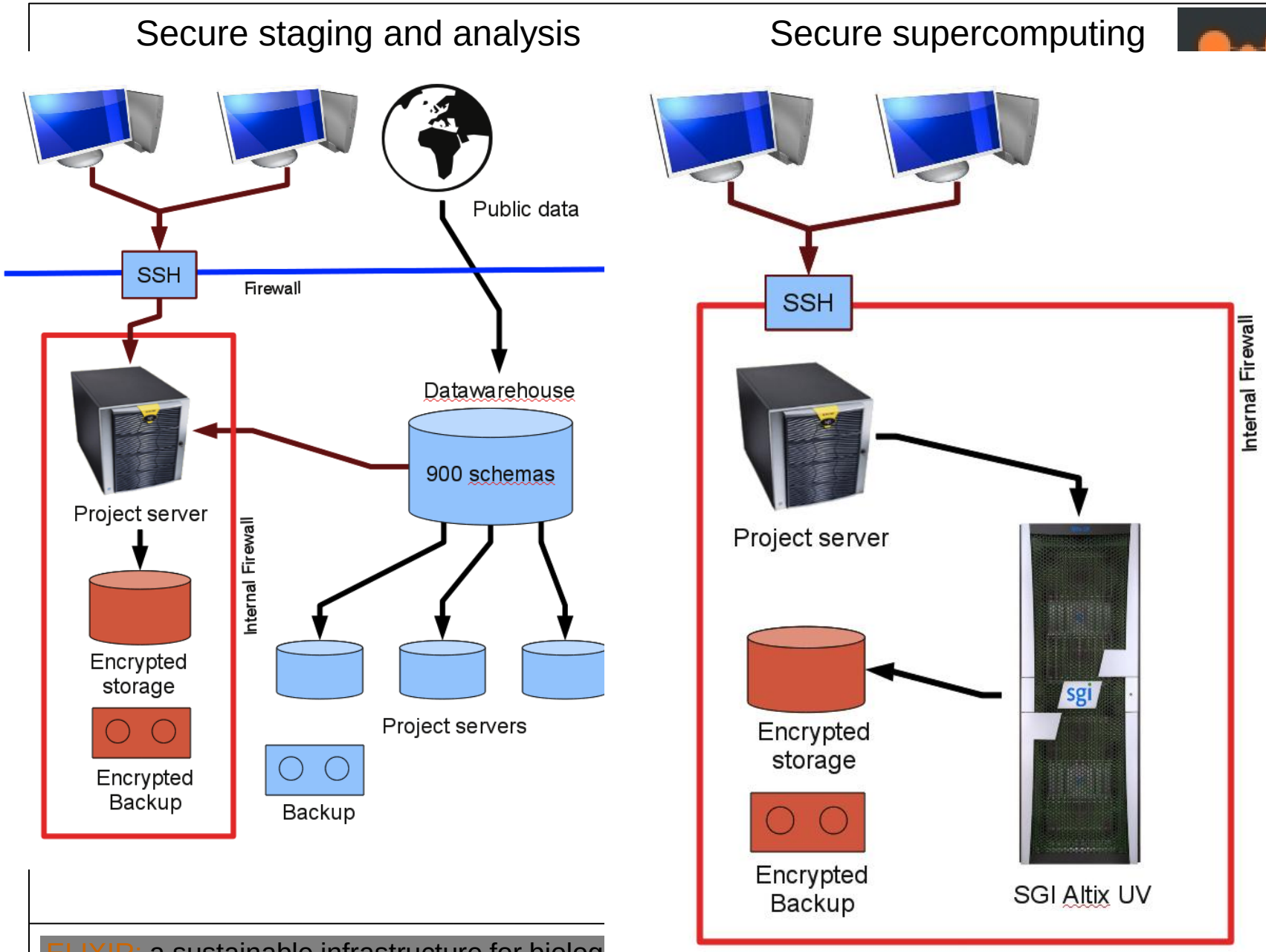
Staging server will have access to both encrypted and public data, and will be powerful enough for pilot runs and small production runs

Secure Supercomputing

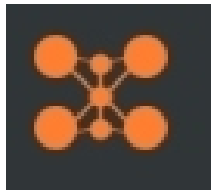


Secure connection to high shared memory (8TB) server when required, in a sandbox environment.

Large production runs can be carried out on the large memory server, which will have appropriate software availability.


Secure staging and analysis
Secure supercomputing
Public data
SSH
Firewall
Datawarehouse
900 schemas
Project server
Internal Firewall
Encrypted storage
Encrypted Backup
Project servers
Backup
SSH
Internal Firewall
Project server
Encrypted storage
Encrypted Backup
sgi
SGI Altix UV

Summary



- Europe is facing unprecedented (grand) challenges.
- The solutions are (mainly) biological
- There are emerging (disruptive) technologies that offer ways forward
- These are very demanding of IT (e-Infrastructure)
- Biology has not been here before
- It will be necessary to move quickly to arrange the necessary funding streams
- Action is needed at every level (scientific, technical, funding, political, ...).

- Open access to data and literature is fundamental
- Integration of data with literature is a major goal.

www.elixir-europe.org