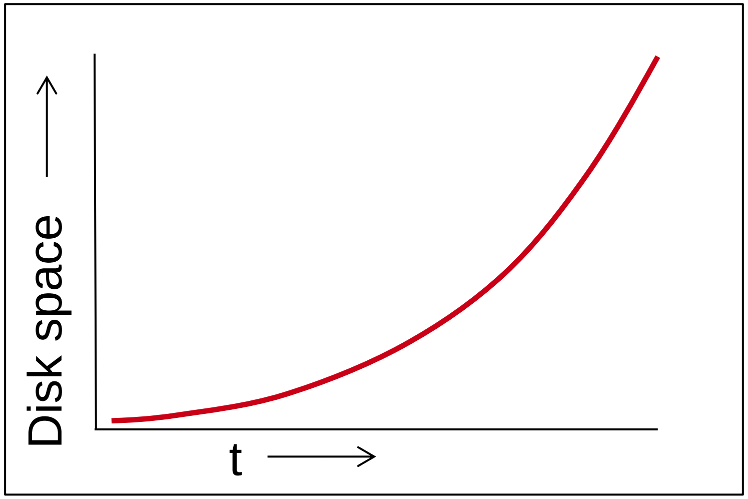
ELIXIR: a sustainable infrastructure for biological information in Europe

Workshop on the future of Big Data Management
The Blackett Laboratory, Imperial College, London
June 27 & 28, 2013
Andrew Lyall PhD, ELIXIR Project Manager

EMBL-EBI

European Bioinformatics Institute

- Outstation of the European Molecular Biology Laboratory
- International organisation created by treaty (cf CERN, ESA)
- 20 year history of service provision and scientific excellence
- EMBL-EBI has 500+ Staff, €50 Million Budget, at least a million users, 20 petabytes of data, 10,000 cpus
- Data are doubling in less than a year
- Bandwidth between disk and memory is at least as big an issue as obtaining sufficient CPU-cycles



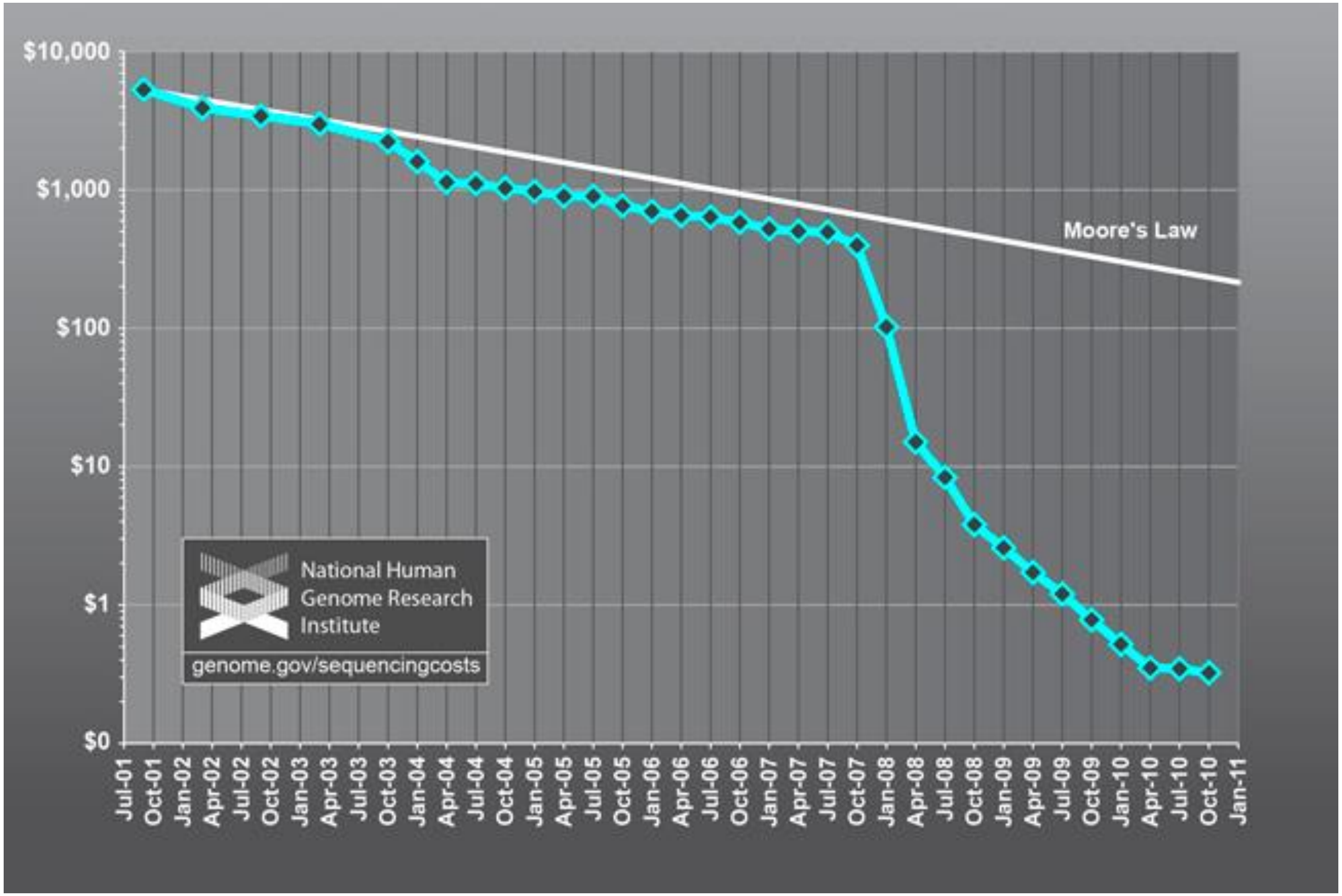
EMBL-EBI Mission Statement

- To provide **freely available data** and bioinformatics **services** to all facets of the scientific community in ways that promote scientific progress
- To contribute to the advancement of biology through basic investigator-driven **research** in bioinformatics
- To provide advanced bioinformatics **training** to scientists at all levels, from PhD students to independent investigators
- To help disseminate cutting-edge technologies to **industry**

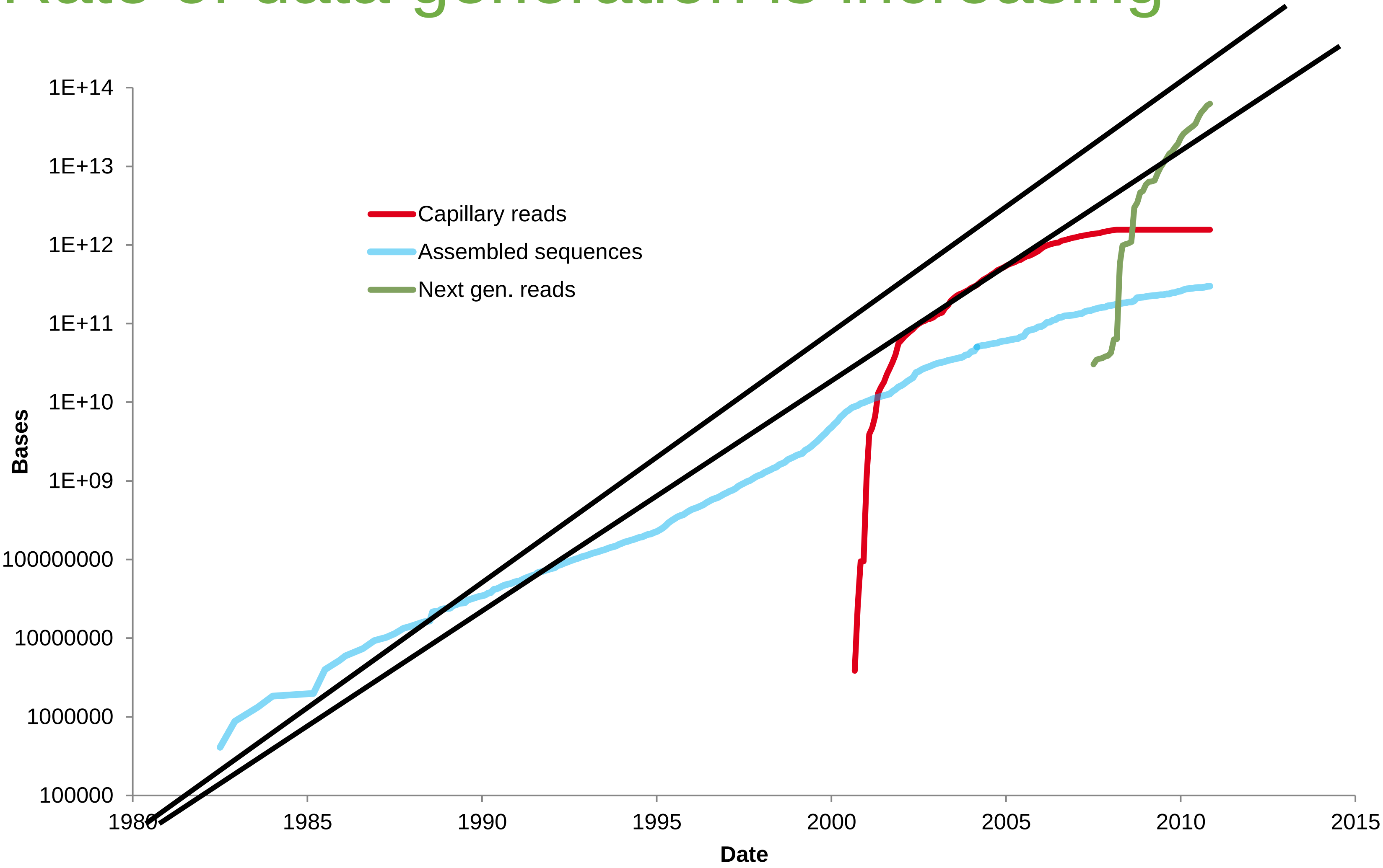
Human Genome project

- 1989 – 2000 – sequencing the human genome
    - Just 1 “individual” – actually a mosaic of about 24 individuals but as if it was one
    - Old school technologies
    - A bit epic
- Now
    - Same data volume generated in ~3mins in a current large scale centre
    - It’s all about the *analysis*

Cost of data generation decreasing exponentially



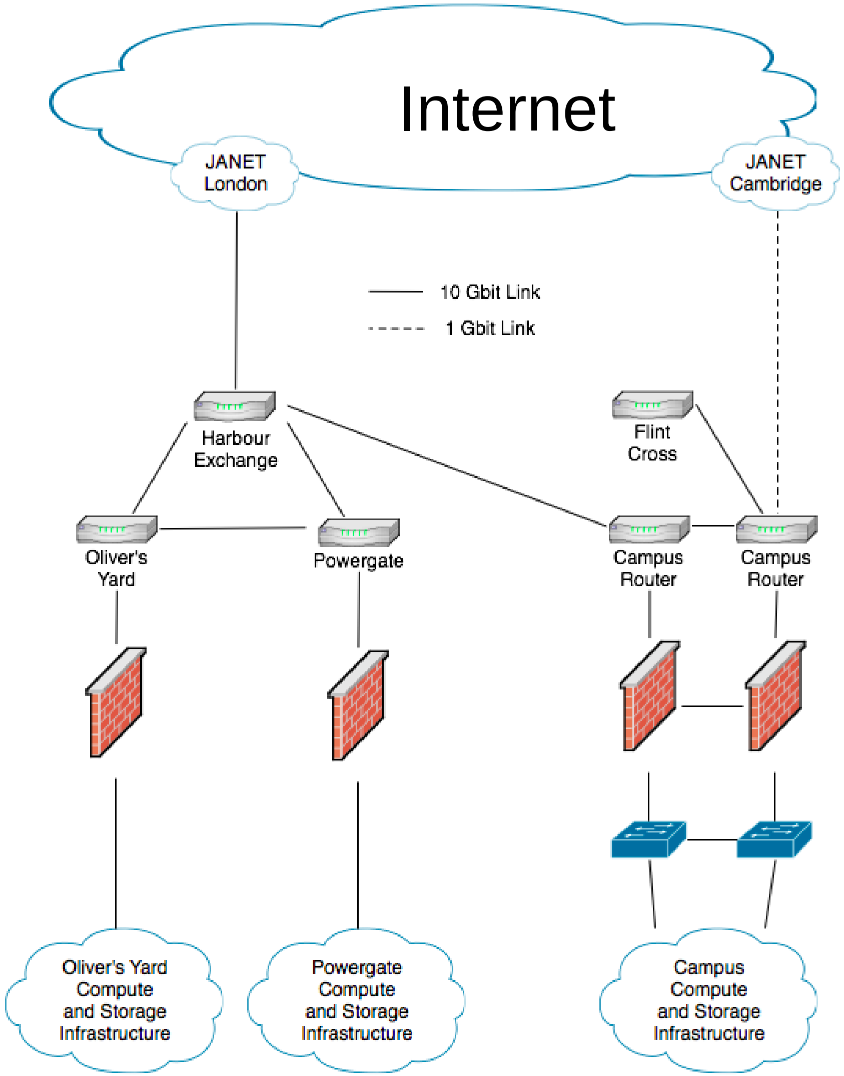
Rate of data generation is increasing

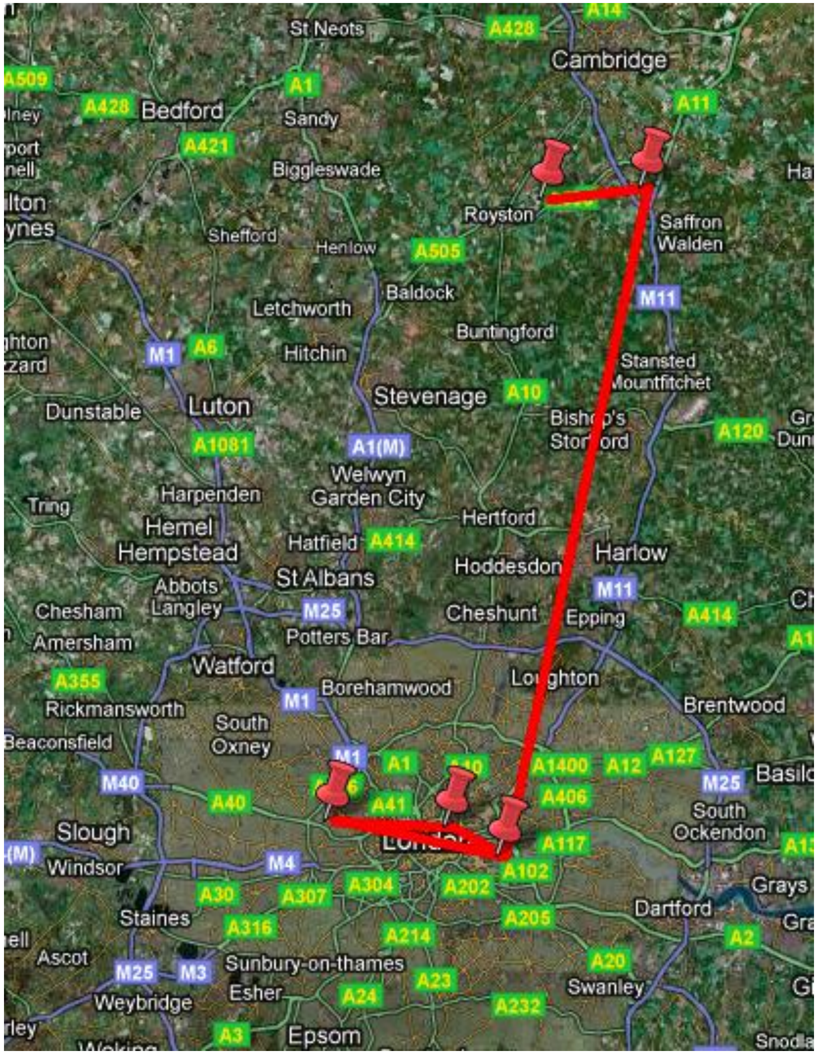


EBI’s technical infrastructure

- 20 PB of “raw” disk
  - Big archives on two systems, no tape backup (analysis is recovery would be very hard; disaster recovery by institutional replication in US)
- ~20,000 cores in 2 major farms
- A Vmware Cloud (“Embassy Cloud”) allowing remote users to directly mount large datasets (in pilot mode)
- 4 machine rooms; 2 in London, 2 in Cambridge
- Janet uplink at 10 Gb/sec

Machine room architecture


Internet
JANET London
JANET Cambridge
10 Gbit Link
1 Gbit Link
Harbour Exchange
Flint Cross
Oliver's Yard
Powergate
Campus Router
Campus Router
Oliver's Yard Compute and Storage Infrastructure
Powergate Compute and Storage Infrastructure
Campus Compute and Storage Infrastructure



Biology already has a data infrastructure

- For the human genome
  - (…and the mouse, and the rat, and… x 150 now, 1000 in the future!) - Ensembl
- For the function of genes and proteins
  - For all genes, in text and computational – UniProt and GO
- For all 3D structures
  - To understand how proteins work – PDBe
- For where things are expressed
  - The differences and functionality of cells - Atlas

And, it is growing fast …

- We have to scale across all of (interesting) life
  - There are a lot of species out there!
- We have to handle new areas, in particular medicine
  - A set of European haplotypes for good imputation
  - A set of actionable variants in germline and cancers
- We have to improve our chemical understanding
  - Of biological chemicals
  - Of chemicals which interfere with Biology

Core data collections at EMBL-EBI

Genomes & Genes

1. **Ensembl:** Joint project with Sanger Institute - high-quality annotation of vertebrate genomes
2. **Ensembl Genomes:** Environment for genome data from other taxons
3. **1000 Genomes:** Catalogue of human variation from major World populations
4. **EGA*:** European Genotype Archive* – genotype, phenotype and sequences from individual subjects and controls
5. **ENA:** European Nucleotide Archive – all DNA & RNA, nextgen reads and traces

Transcription

6. **ArrayExpress:** Archive of transcriptomics and other functional genomics data
7. **Expression Atlas:** Differentially expressed genes in tissues, cells, disease states & treatments

Protein

8. **UniProt:** Archive of protein sequences and functional annotation
9. **InterPro:** Integrated resource for protein families, motifs and domains
10. **PRIDE:** Public data repository for proteomics data
11. **PDBe:** Protein and other macromolecular structure and function

Small molecules

12. **ChEBI:** Chemical entities of biological interest
13. **ChEMBL:** Bioactive compounds, drugs and drug-like molecules, properties and activities

Processes

14. **IntAct:** Public repository for molecular interaction data
15. **Reactome:** Biochemical pathways and reactions in human biology
16. **Biomodels:** Mathematical models of cellular processes

Ontologies

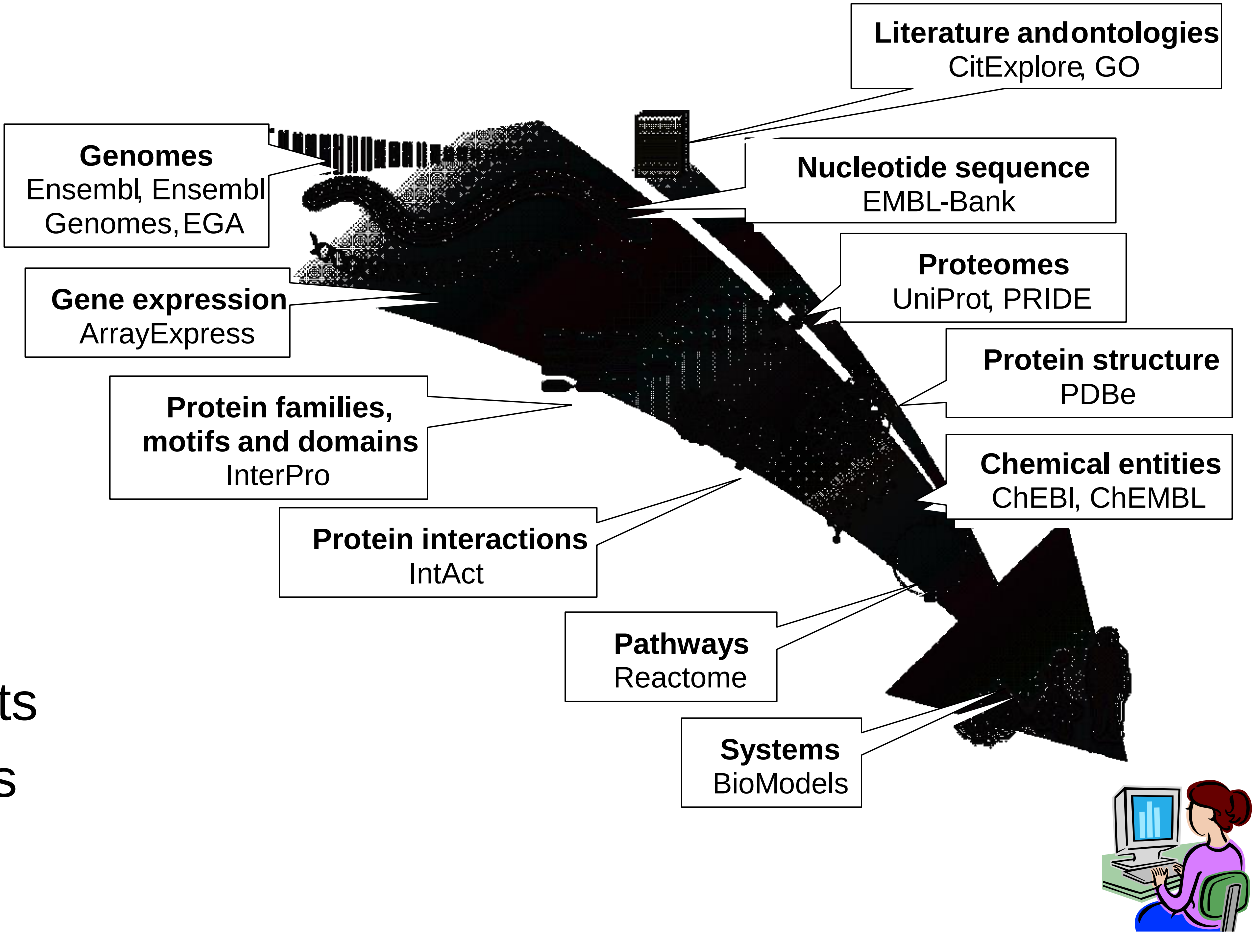
17. **GO:** Gene Ontology, consistent descriptions of gene products

Scientific literature

18. **CiteXplor:** Bibliographic query system

Universal, comprehensive, integrated

- Life sciences
- Medicine
- Agriculture
- Pharmaceuticals
- Biotechnology
- Environment
- Bio-fuels
- Cosmaceuticals
- Neutraceuticals
- Consumer products
- Personal genomes
- Etc…



Literature and ontologies
CitExplore, GO

Genomes
Ensembl, Ensembl Genomes, EGA

Nucleotide sequence
EMBL-Bank

Proteomes
UniProt, PRIDE

Gene expression
ArrayExpress

Protein structure
PDBe

Protein families, motifs and domains
InterPro

Chemical entities
ChEBI, ChEMBL

Protein interactions
IntAct

Pathways
Reactome

Systems
BioModels

Biology is a big data science

- Not perhaps as big as high energy physics, but “just” one order of magnitude less (~35 PB plays ~300PB)
- Hetreogenity/diversity of data far larger
  - Lots and lots of details
  - Always have “dirty” data even in best case scenarios
- Big Data => scaleable algorithms
  - CS ‘stringology’ – eg, Burrows-Wheeler transform
- Biological inference => statistics
  - “just” need to collapse a 3e9 dimensional problem into something more tractable
- We are often I/O not CPU bound

Big Data Example: Personalised medicine

- Personalised medicine will require sequencing of the genomes of large numbers of patients and volunteers
- It will be necessary to compare at least some of these genomes with the reference data collections
- Most hospitals and clinical research institutes will not wish to maintain up-to-date copies of the reference data collections
- It will be therefore be necessary to send these genomes to the institutes that hold the reference data collections
- It seems likely that this will be achieved using secure VMs and secure clouds holding the reference data collections
- EMBL-EBI is engaging with stakeholders to evaluate opportunities in this area.

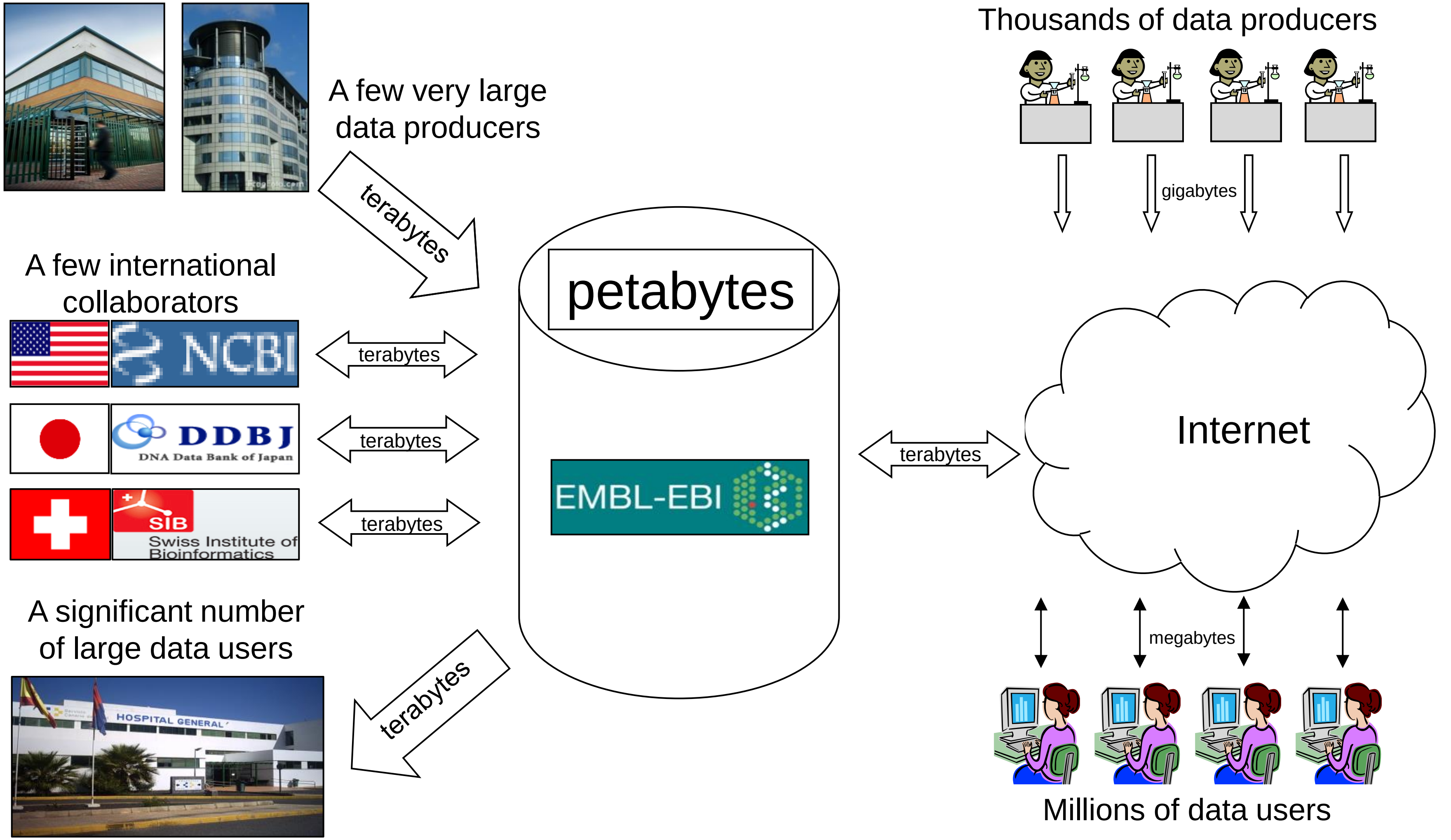
Collaborator “Embassy” Clouds

- Pharmaceutical companies put significant effort into creating secure “EBI-like” services on their own infrastructure
- Many other users with high computational requirement do not wish to recreate our infrastructure on their own site
- A secure cloud environment providing “Cloud-Embassies” at EMBL-EBI would obviate this
- Embassy owners would have complete control over their virtual infrastructure
- Embassy owners could bring their own data and software to compute against EMBL-EBIs data and services
- Such services would be managed with legally acceptable collaboration agreements.

The Helix-Nebula Science-Cloud

- Three members of EIROforum (CERN, EMBL & ESA)
- Thirteen European IT providers (more are joining)
- A pan-European partnership of academia and industry to create cloud solutions and foster innovation in science
- Stimulate the creation of a cloud computing market in Europe (cf USA)
- Two year pilot phase after which it will be made more widely available to commercial and public domain
- EMBL will use it for the analysis of large genomes

Many different modes of data generation & utilisation



A few very large data producers

terabytes

A few international collaborators

terabytes

terabytes

terabytes

A significant number of large data users

terabytes

petabytes

EMBL-EBI

Thousands of data producers

gigabytes

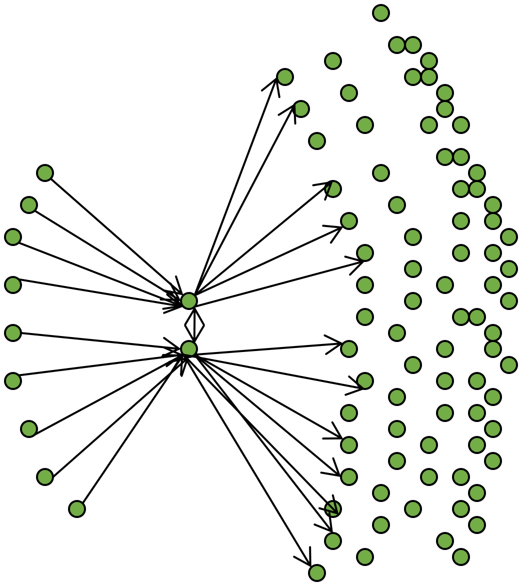
terabytes

Internet

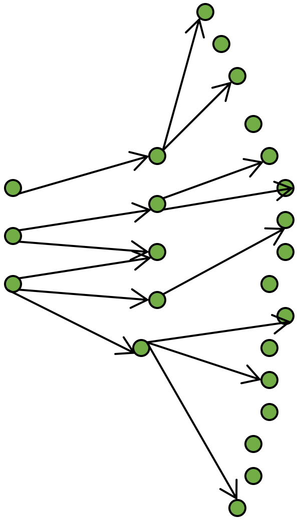
megabytes

Millions of data users

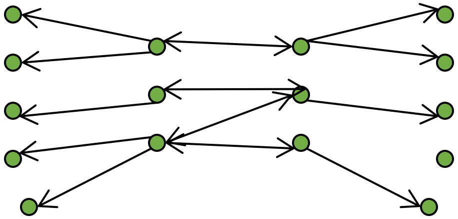
Data distribution patterns are discipline dependent



Genomics



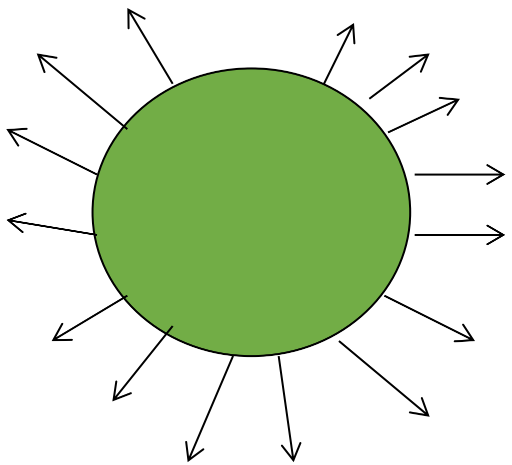
High Energy
Physics



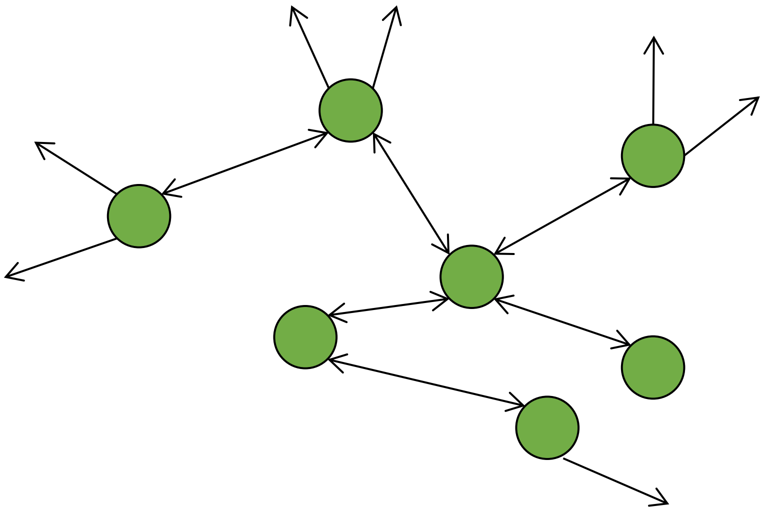
Astronomy

How to coordinate biological data provision?

Fully Centralised



Fully Distributed



Pros:

- Easier to benefit from re-use
- Easier to standardise

Cons:

- Impossible to centralise all necessary expertise
- Risk of bottlenecks
- Lack of diversity

Pros:

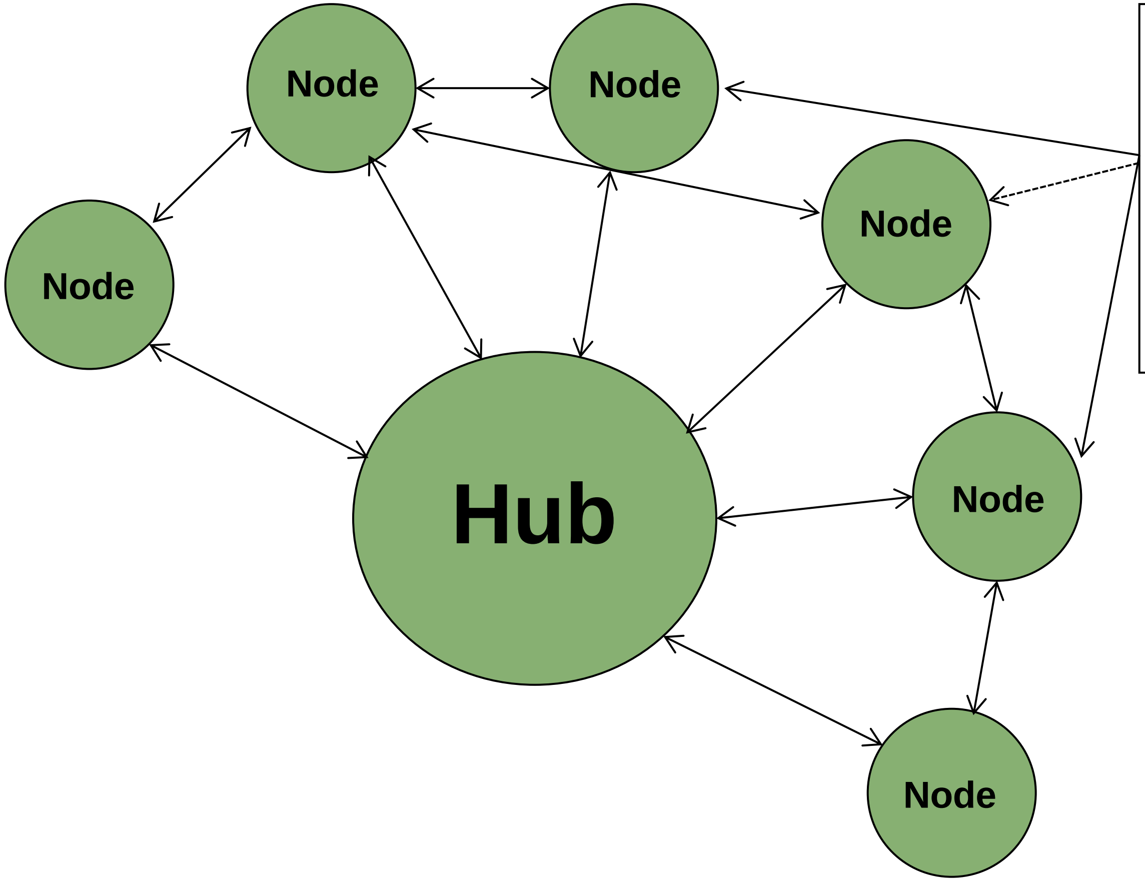
- More responsive
- Better support for Human languages

Cons:

- Communication & coordination overheads
- Harder to support end users
- Harder to provide multi-decade stability

We need the best of both models…

Robustly connected nodes coordinated by a strong hub

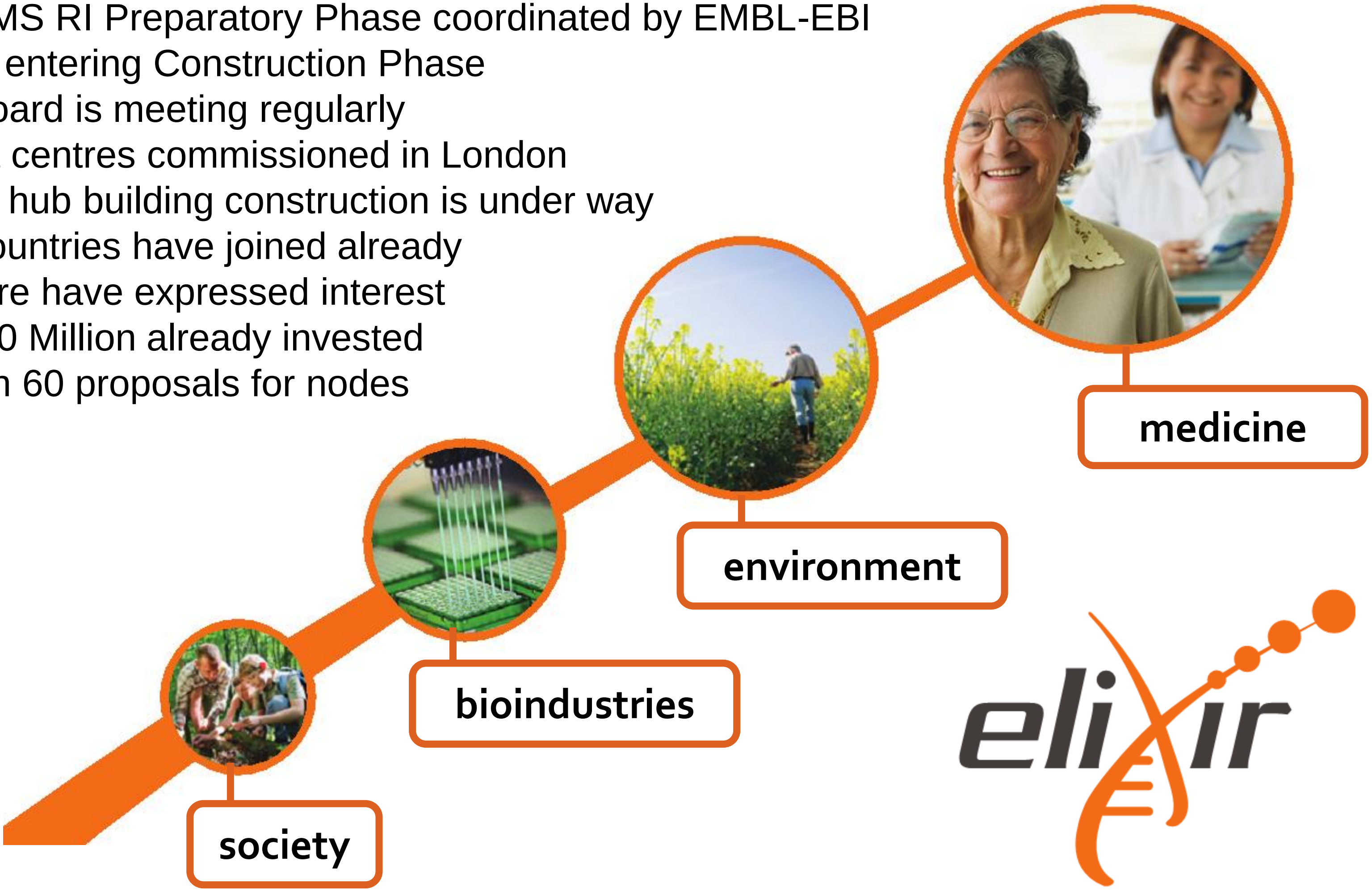


Diversity of Nodes:

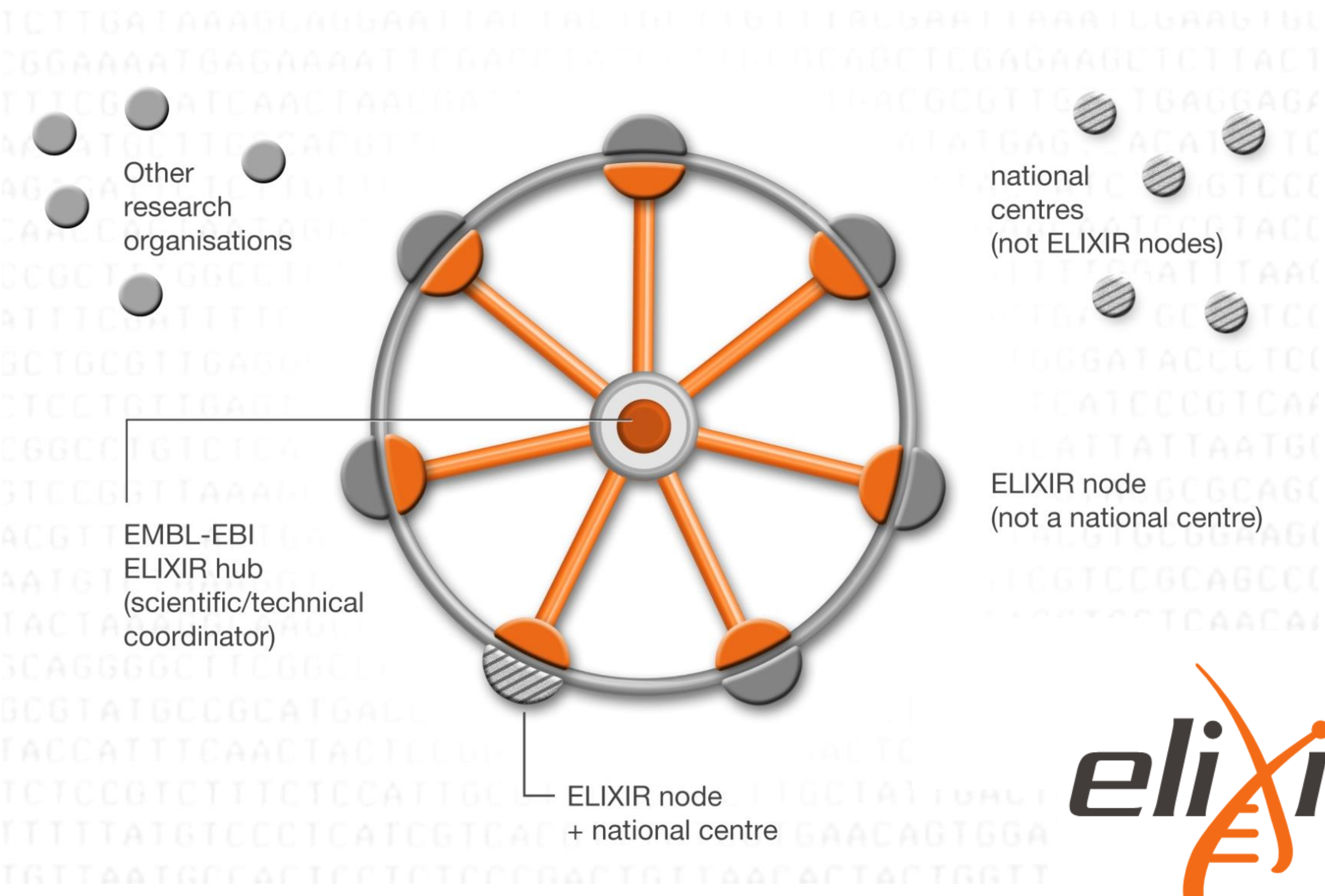
- Domain Specific
- Regional
- National
- Training
- Etc…

ELIXIR: A sustainable infrastructure for biological information in Europe…

- ESFRI BMS RI Preparatory Phase coordinated by EMBL-EBI
- Currently entering Construction Phase
- Interim board is meeting regularly
- New data centres commissioned in London
- Technical hub building construction is under way
- Fifteen countries have joined already
- Many more have expressed interest
- Over €100 Million already invested
- More than 60 proposals for nodes



ELIXIR: Distributed nodes with a hub



Other research organisations

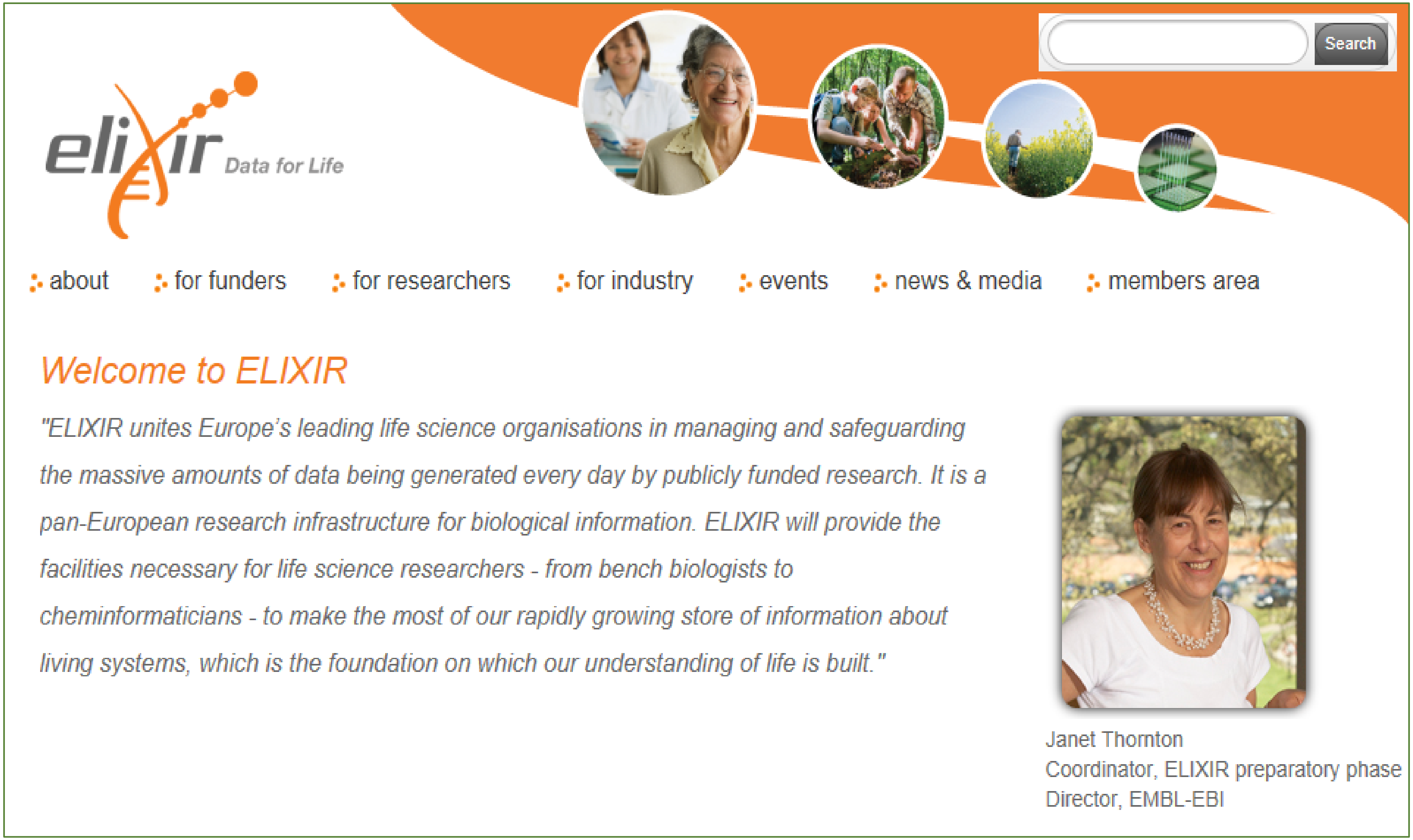
national centres (not ELIXIR nodes)

EMBL-EBI ELIXIR hub (scientific/technical coordinator)

ELIXIR node (not a national centre)

ELIXIR node + national centre

http://www.elixir-europe.org/



Welcome to ELIXIR

"ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information. ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built."

Janet Thornton
Coordinator, ELIXIR preparatory phase
Director, EMBL-EBI

Conclusions

- Data management is becoming a significance challenge in biology: size, complexity, ELSI…
- High-throughput data-generators and users will be situated all round Europe
- The environment will be very heterogeneous with complex data and many different modalities of use
- Organising I/O from disk to memory is as big a challenge as obtaining sufficient CPU-cycles

nature International weekly journal of science

Biology: The big challenges of big data

Vivien Marx

Nature, Volume: 498, Pages: 255–260, Published:***13 June 2013***

Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists.

(& thank you for your attention)