The EMBL-Bioinformatics and Data-Intensive Informatics

Graham Cameron

EMBL-EBI

EMBL-EBI

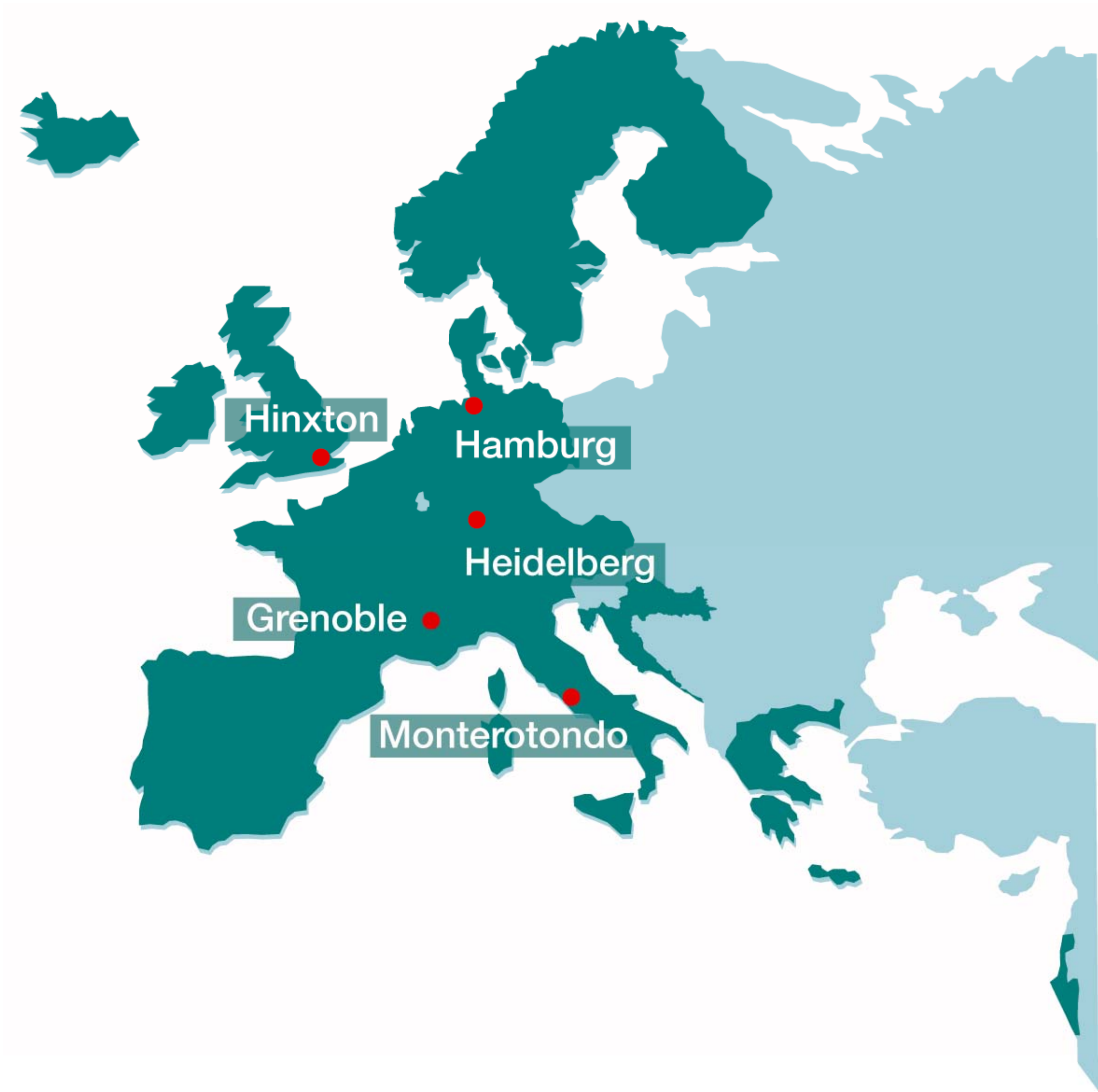
What is the EMBL-EBI?

- Non-profit organization
- Part of the European Molecular Biology Laboratory
- Based on the Wellcome Trust Genome Campus near Cambridge, UK

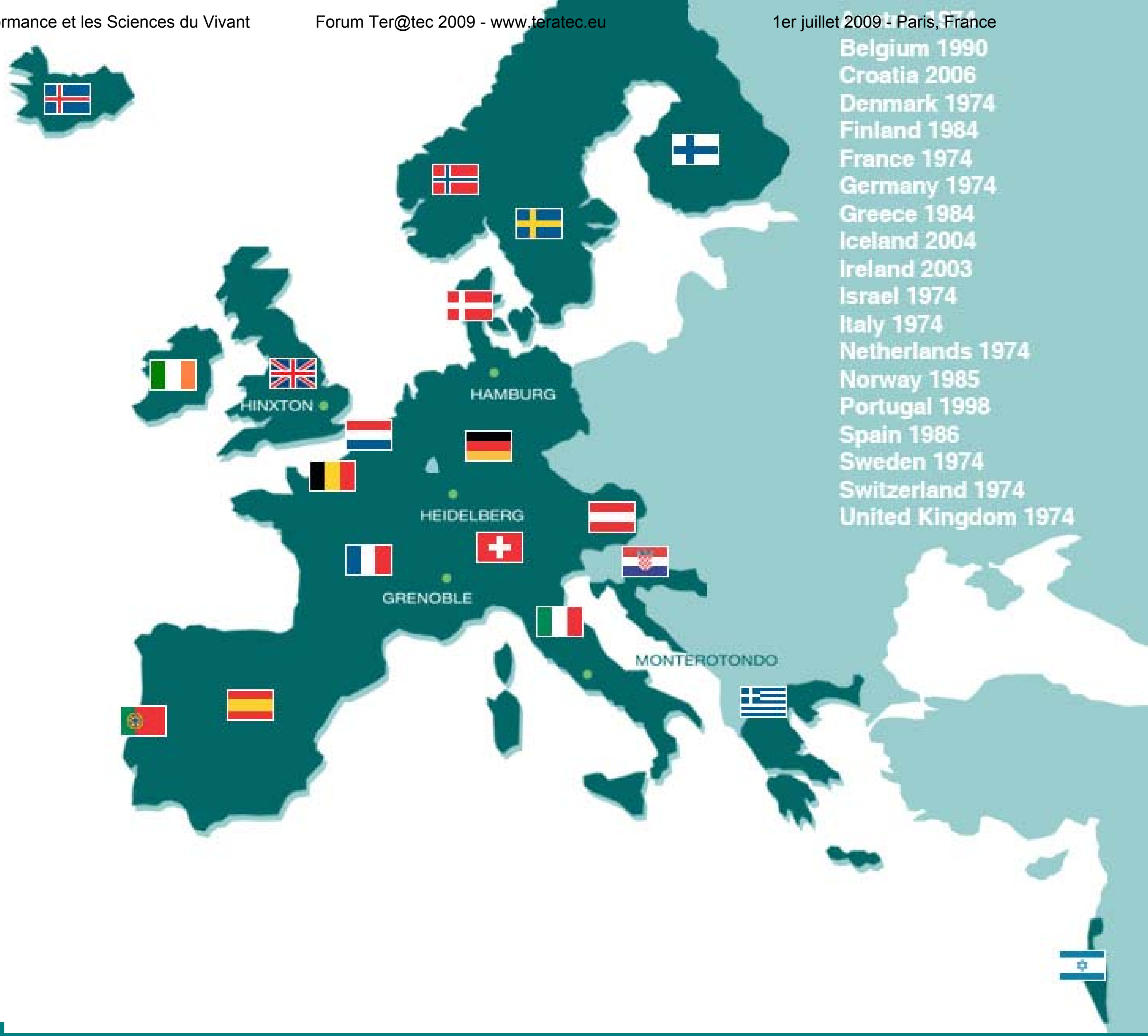
EMBL-EBI

Part of EMBL

The EBI is part of the European Molecular Biology Laboratory (EMBL), a basic research institute funded by public research funds from 19 member states.



EMBL Member States



Austria	1974
Belgium	1990
Croatia	2006
Denmark	1974
Finland	1984
France	1974
Germany	1974
Greece	1984
Iceland	2004
Ireland	2003
Israel	1974
Italy	1974
Netherlands	1974
Norway	1985
Portugal	1998
Spain	1986
Sweden	1974
Switzerland	1974
United Kingdom	1974

European Bioinformatics Institute (EBI)

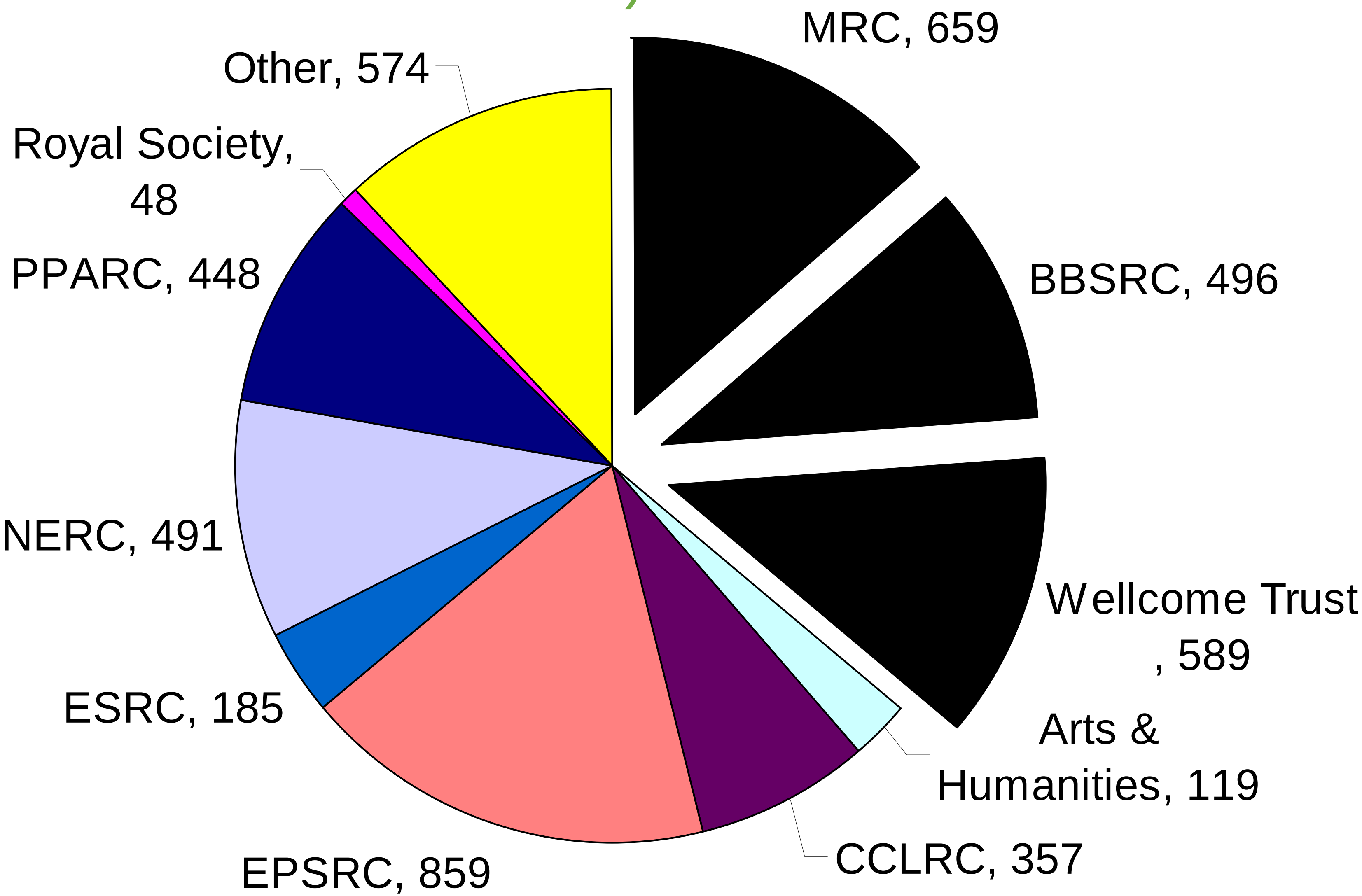
- Research
- Service
- Training
- Industry support

European Bioinformatics Institute (EBI)

- Research
- Service
- Training
- Industry support

Bioinformatics

€1.7 billion of UK Research funding is life science (total= € 4.8 billion 2007/8)



The central dogma

- Genomes contain genes
- Genes produce transcripts
- Transcripts translate to protein sequences
- Protein sequences form complex 3D structures.

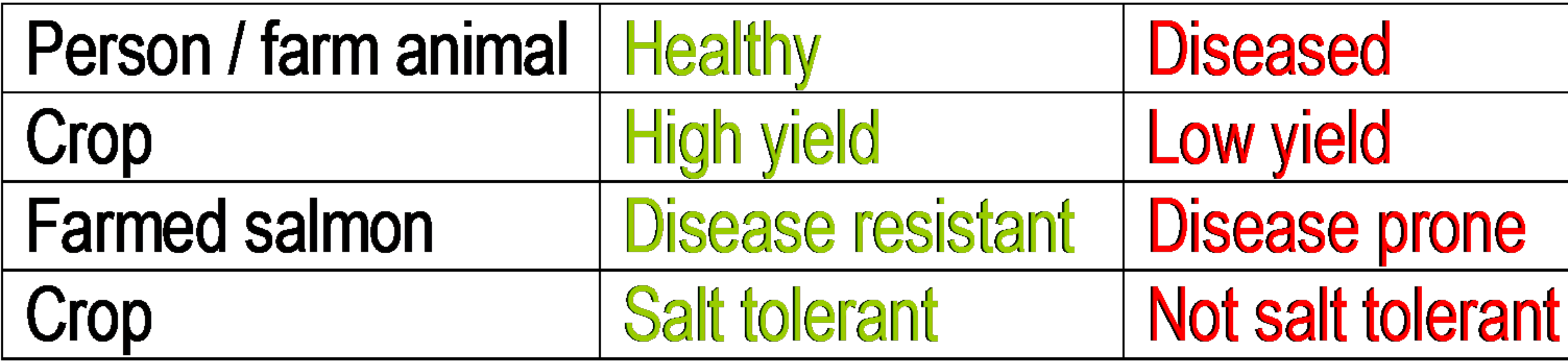
The data

- Protein Structure Database (PDB) created in 1971 for 3D structures of bio-macromolecules
- Early 80s DNA sequencing databases established
- Protein sequence databases
- Gene expression
- Proteomics
- Genetic variation
- Interactions and pathways
- Models
- Drugs
- Metabolites

Benefits

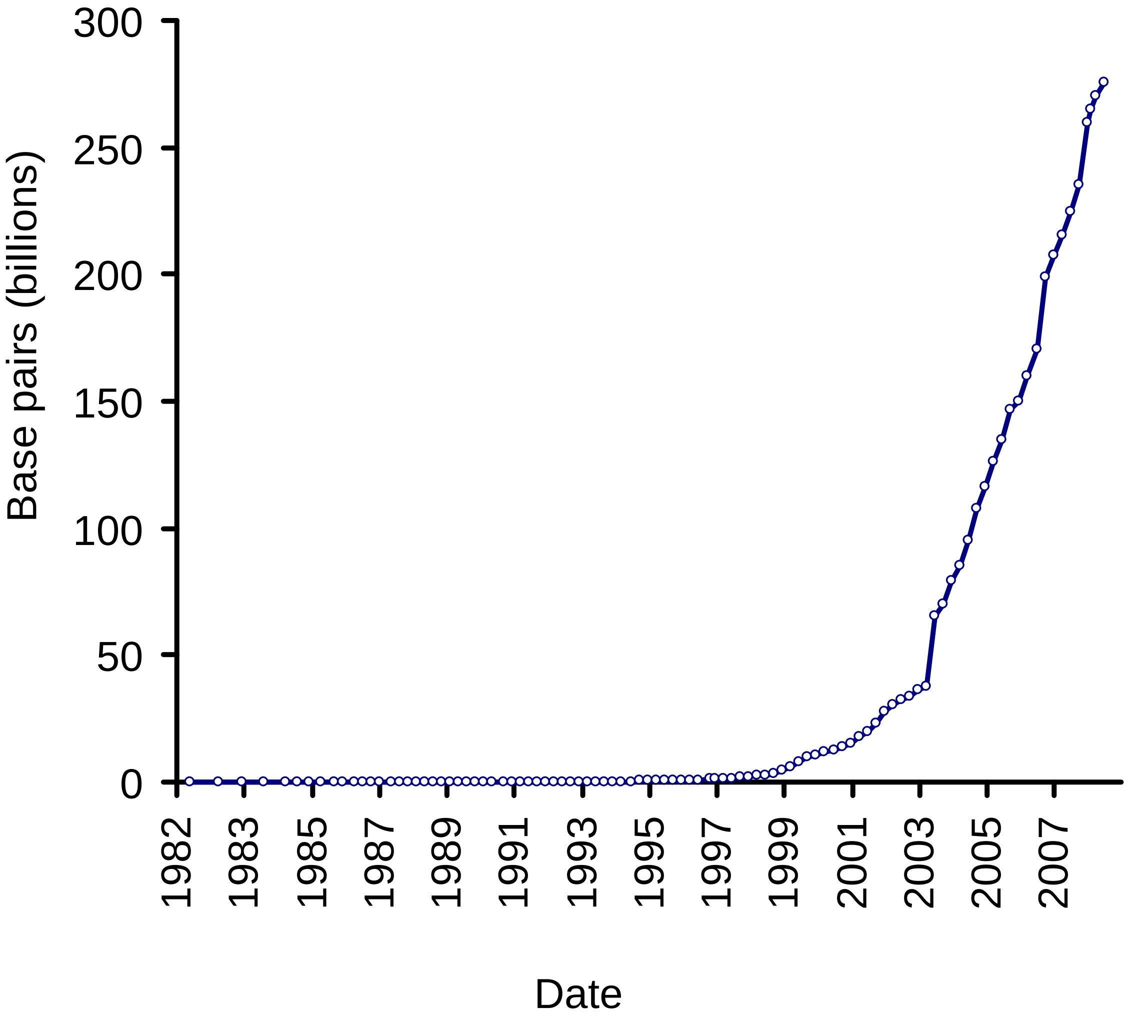
- Health and medicine
- Personal care
- Agriculture
- Food science
- Brewing and fermentation
- Forestry
- Fishery
- Environment

Use



Person / farm animal	Healthy	Diseased
Crop	High yield	Low yield
Farmed salmon	Disease resistant	Disease prone
Crop	Salt tolerant	Not salt tolerant

DNA Sequence database growth



Usage

- About three million web hits a day at the EBI
- A few hundred thousand users
- A new data acquisition every 2 seconds

Genomes are getting easy(ish)

Human genome

- 3 000 000 000 base pairs
- Draft released in 2000
- Cost about $3 billion
- Today’s sequencing centres can sequence that much in half a day!
- Massively parallel laboratory methods

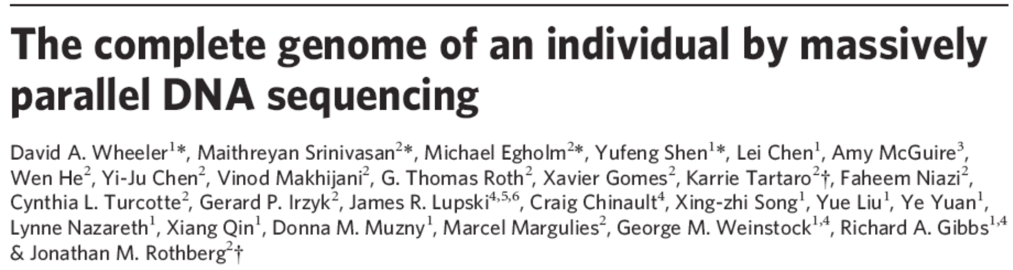
- Oops – informatics is now the bottleneck

2007: The Personal Genome Era Beings

Jim Watson
(Photo credit: Caltech

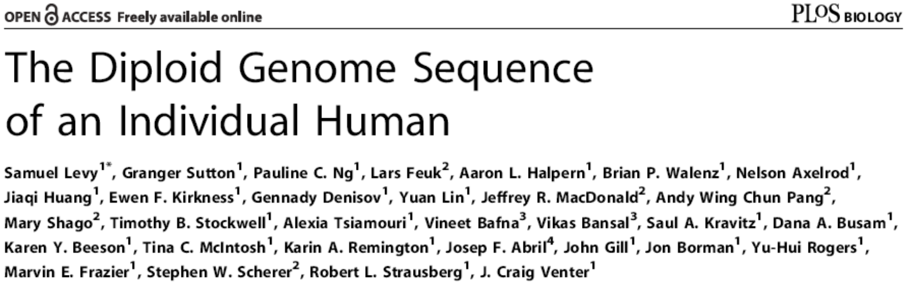
nature | 2008| doi:10.1038/nature06884

LETTERS



The complete genome of an individual by massively parallel DNA sequencing

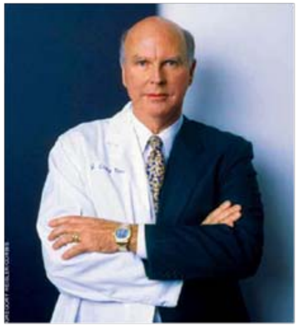
David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†



OPEN ACCESS Freely available online PLoS BIOLOGY

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1]*, Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]



Craig Venter
(Photo: BusinessWeek)

Genotyping (your very own genome)

- Codeine is metabolised into morphine by the cytochrome P450 2D6
- Often used as a painkiller after childbirth
- Most people have one working copy of the 2D6 gene
- A small number of people have two or even three working copies
- Mothers with multiple copies of 2D6 convert codeine to morphine so efficiently that their babies have been known to die of morphine poisoning through breast milk
- If you know the genotype you can prescribe the right drug

- Personalised medicine and theranostics

1000 Genomes Project

- Create a deep catalogue of human variation to provide a better baseline to underpin human genetics
- There is lots of undiscovered variation
- Say 100 times as much data as we had at the start of the project
- Expect approaching half a petabyte of data from this one project
- (this is after about a 100 fold reduction in what comes off the machines)

Data Transfer Infrastructure

- FTP does not work well for terabytes of data
- Send a hard drive
- Point to point leased lines
- Advanced technology solutions which don’t do all sorts of nannying accuracy checks

Aspera™

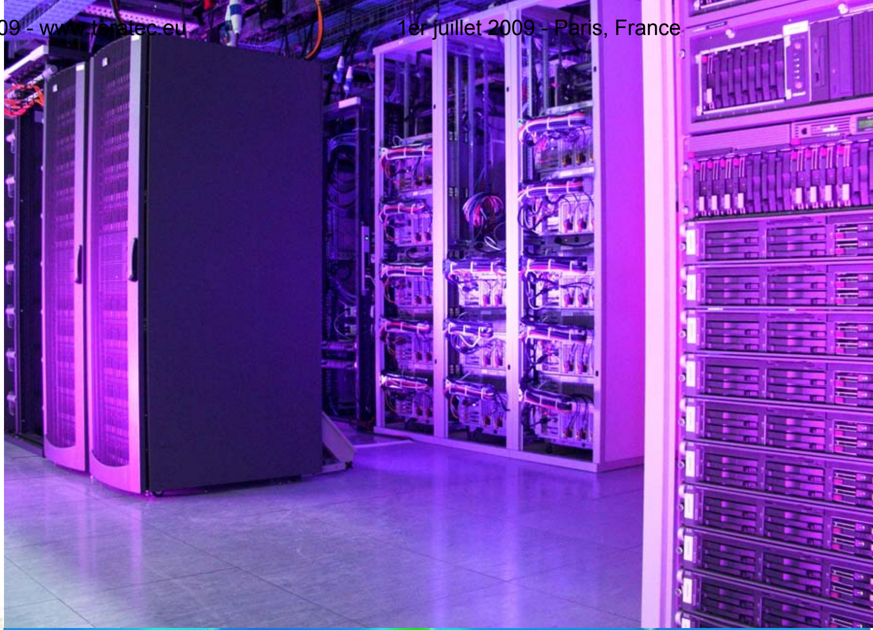
Supercomputing Data Centre

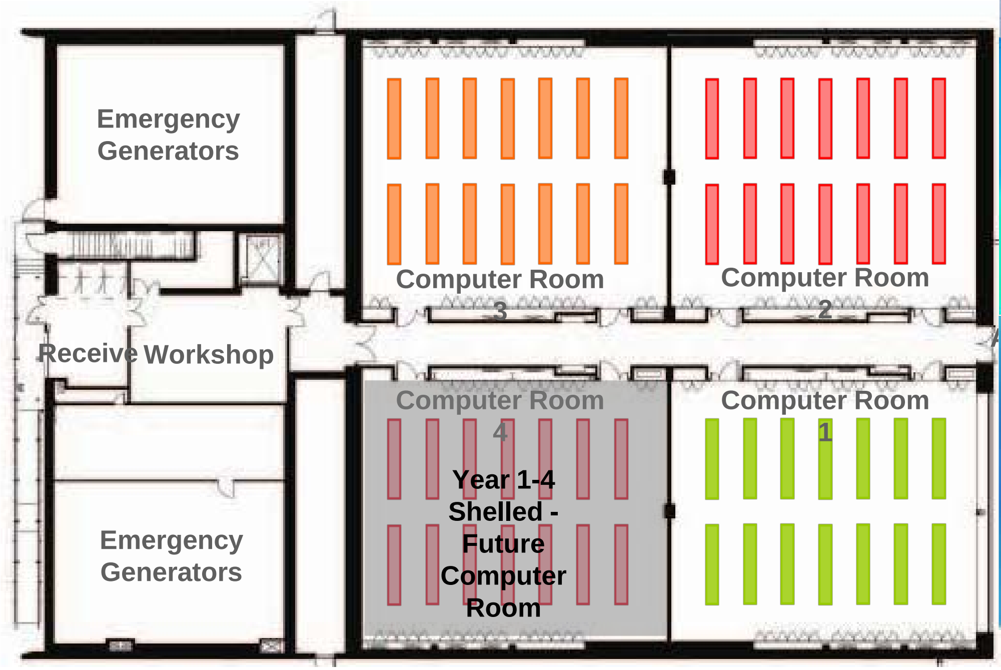
£45 million of computers, 1,000 square metres of computing equipment rooms

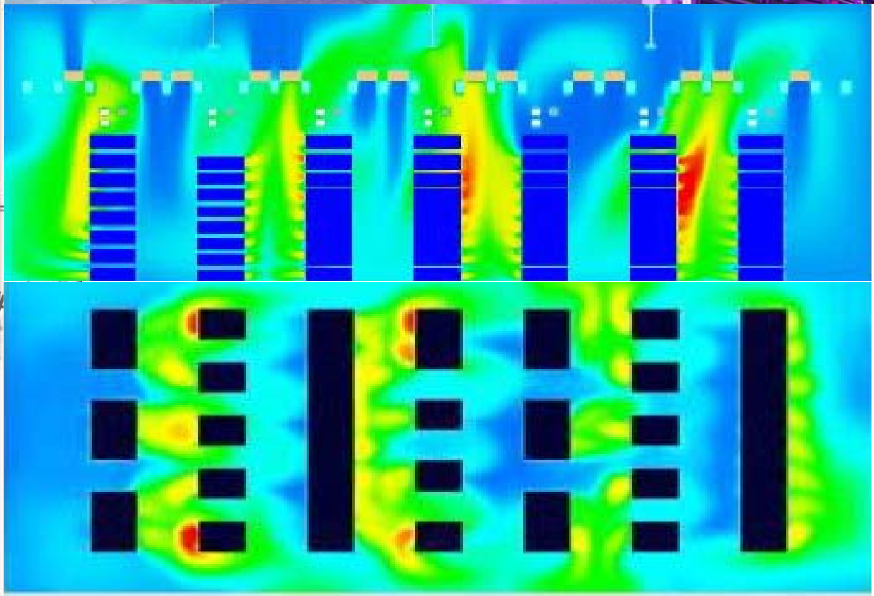
Computing power: 2,000 watts per square metre

3.4 megawatts for the total facility

Not the way we would do it today







Strategy/trends

- 7000 cores (14000 total on campus)
- About 5 petabytes of data
- Multiple disks for speed not storage

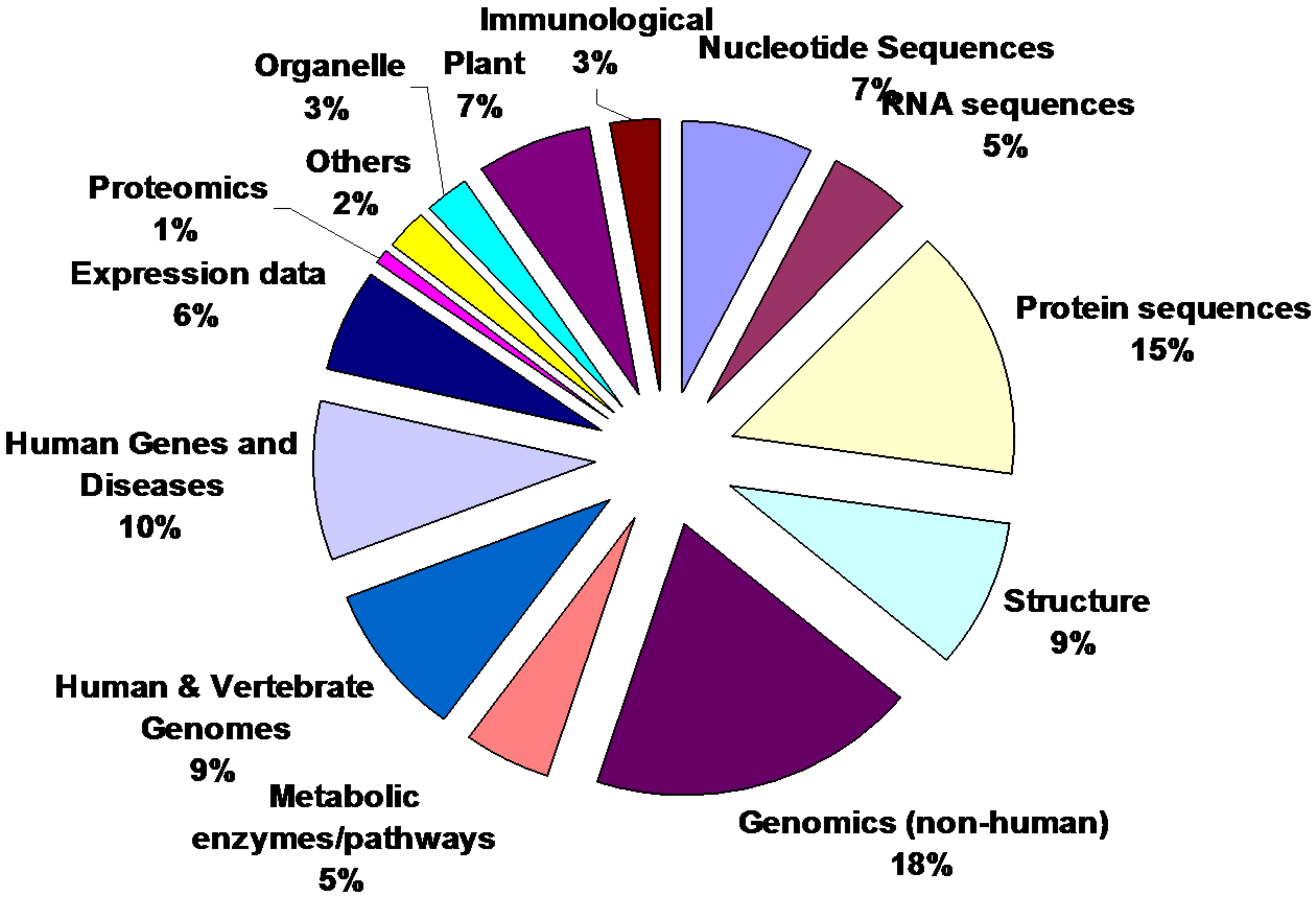
- Massive shared compute farms
- Replicated data storage
- Outsourcing compute (more expensive than our own cloud)

- All human genomes are pretty similar
- Large regions of even one genome are similar to each other
- Even genomes of different species have lots of similarity
- Storing that information, exploring it, and presenting results of searches can be made hugely efficient by utilising data structures which directly represent all the relationships between identical subsequences of genomes (store them only once)

- Prototype de Bruijn graph methods currently use 200 gigabytes of physical memory
- We really need about 5 terabytes of physical memory!

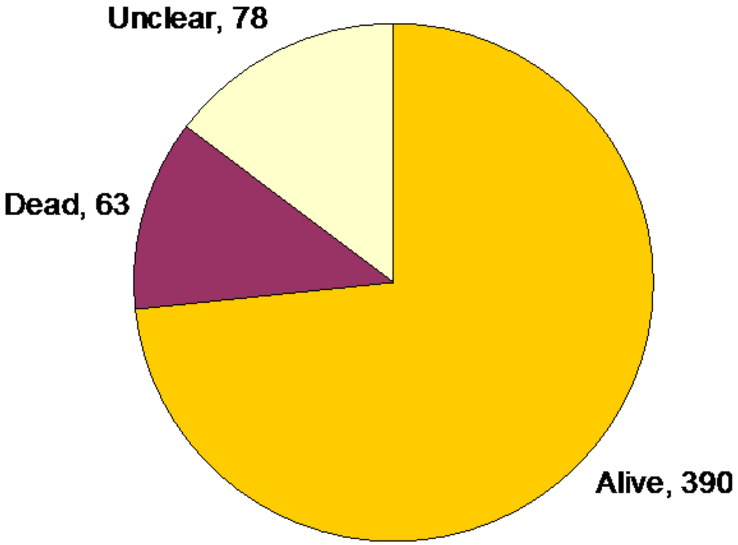
Information Infrastructure

1000 databases (Galperin 2008)

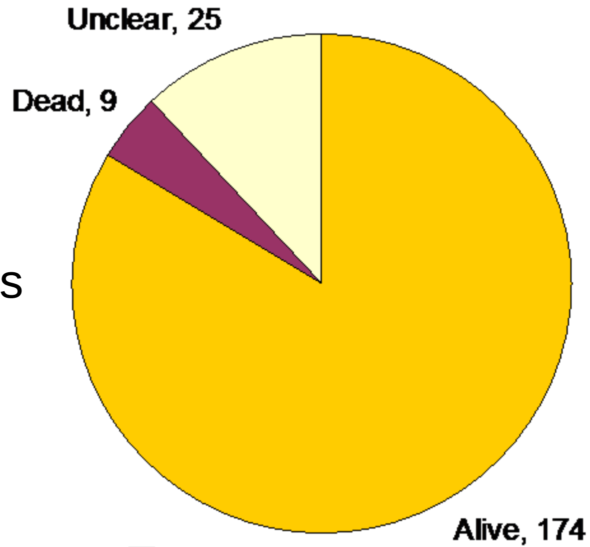


531 Databases surveyed

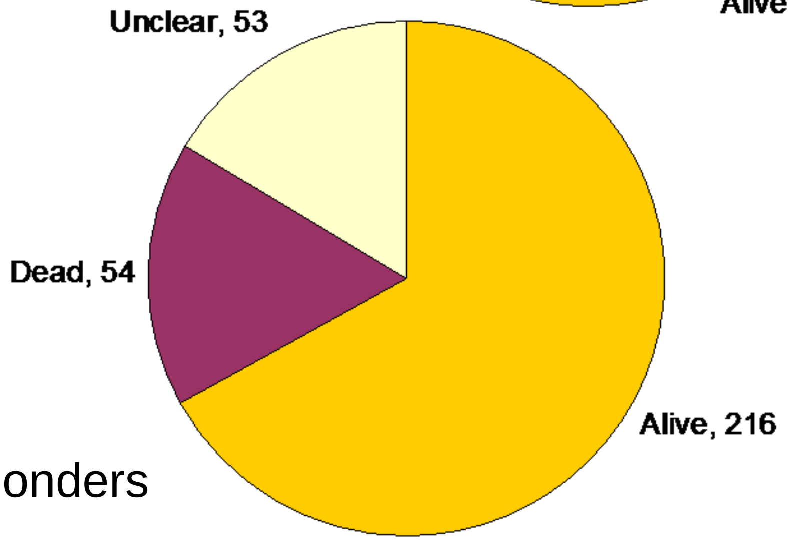
208 Responded, 323 did not



Responders



Dead = no update since 2005



Non-responders

A few big databases

Size	NDB
0 to 0.5 gigabytes	47
0.5 to 1 gigabytes	32
1 to 2 gigabytes	21
2 to 4 gigabytes	13
4 to 6 gigabytes	6
6 to 8 gigabytes	5
8 to 10 gigabytes	7
10 to 20 gigabytes	8
20 to 50 gigabytes	17
50 to 100 gigabytes	14
100 to 200 gigabytes	6
200 to 500 gigabytes	8
500 to 1000 gigabytes	6
1000 to 2000 gigabytes	2
2000 to 3000 gigabytes	2
3000 to 4000 gigabytes	0
4000 to 5000 gigabytes	5

Table 2. Reported sizes of databases

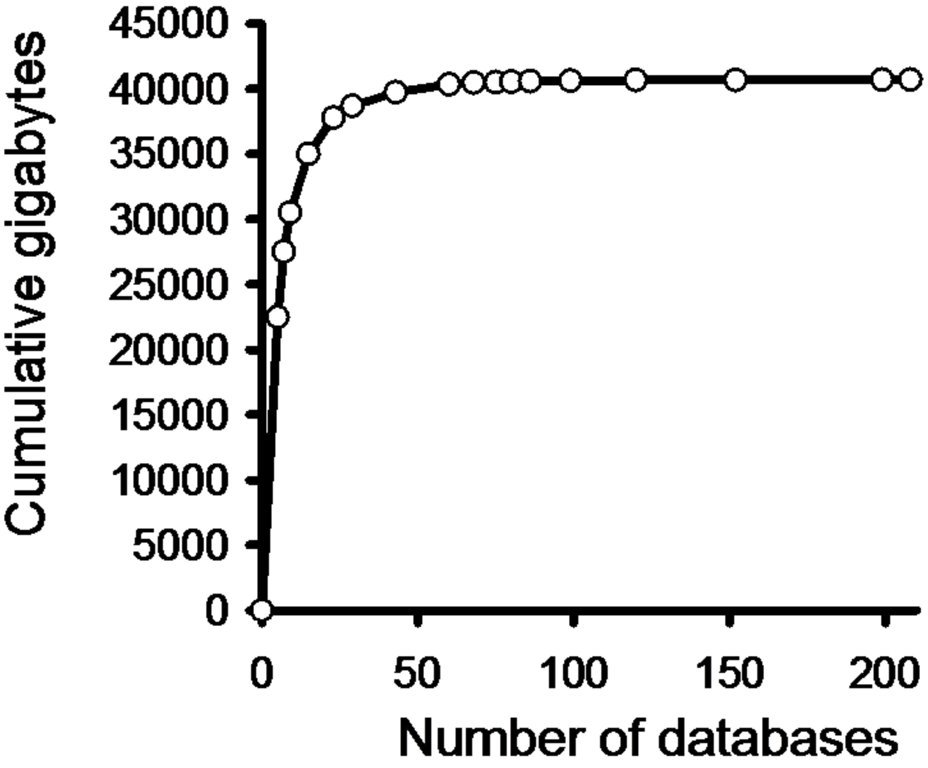


Figure 6. Cumulative gigabytes by database

Usage - Europe

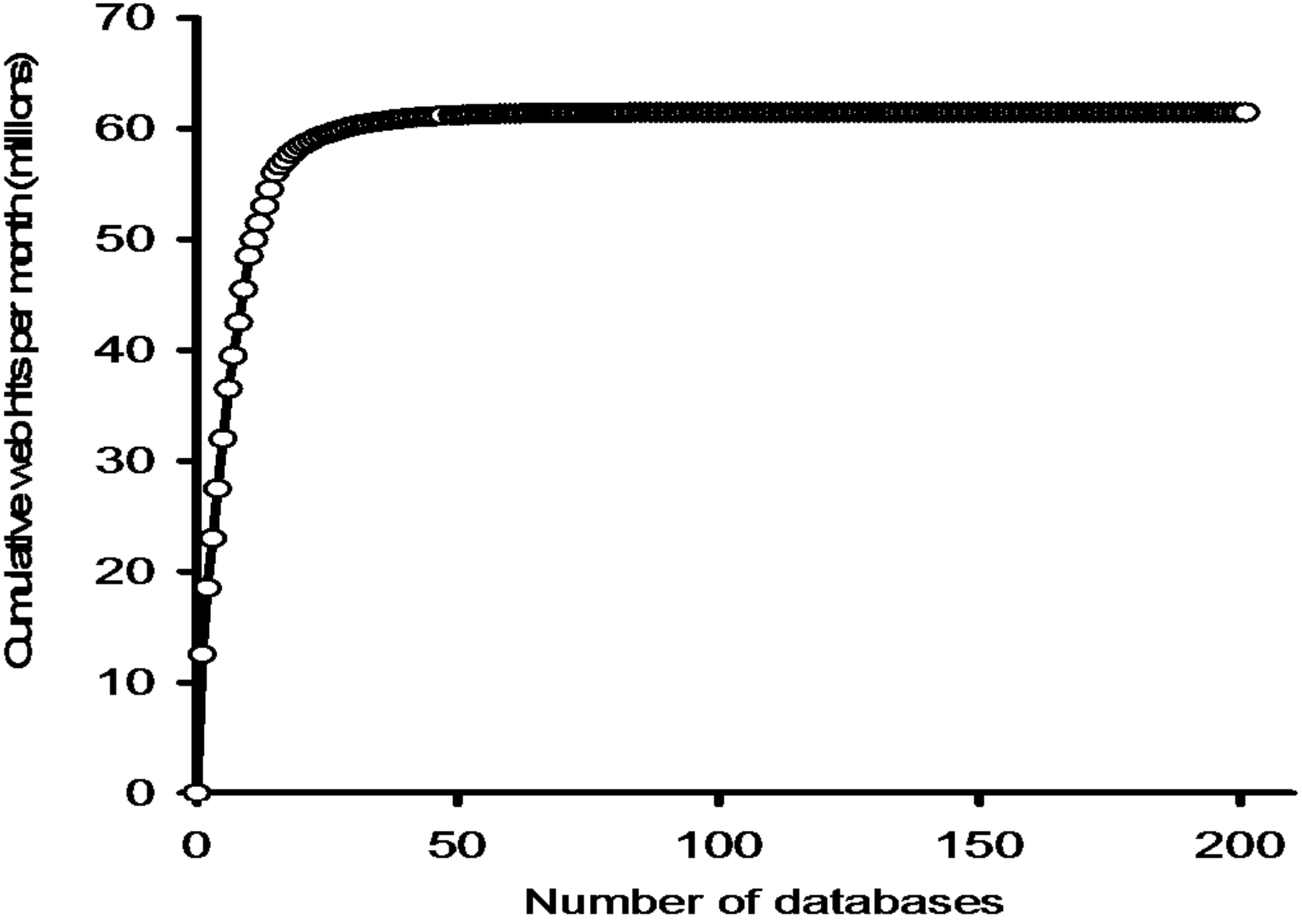


Figure 8. Cumulative web hits by database

Lots of databases to come

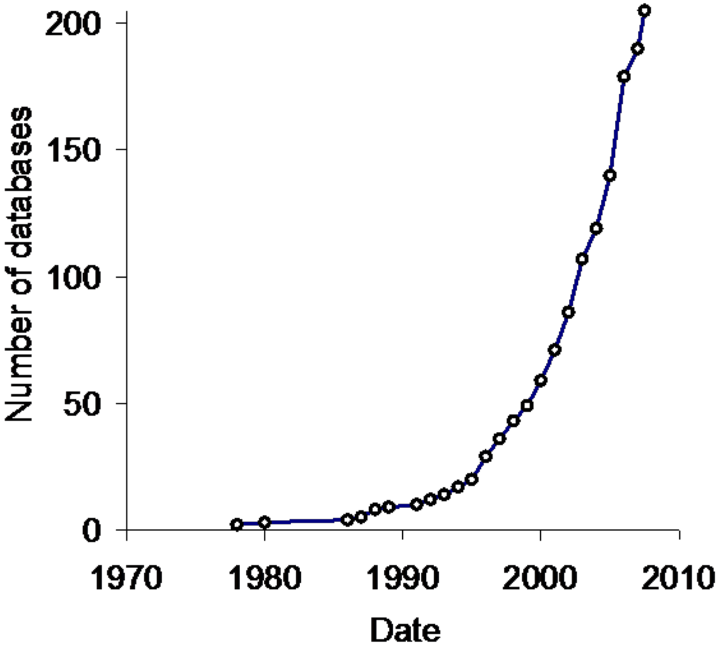


Figure 15. Growth in the number of databases

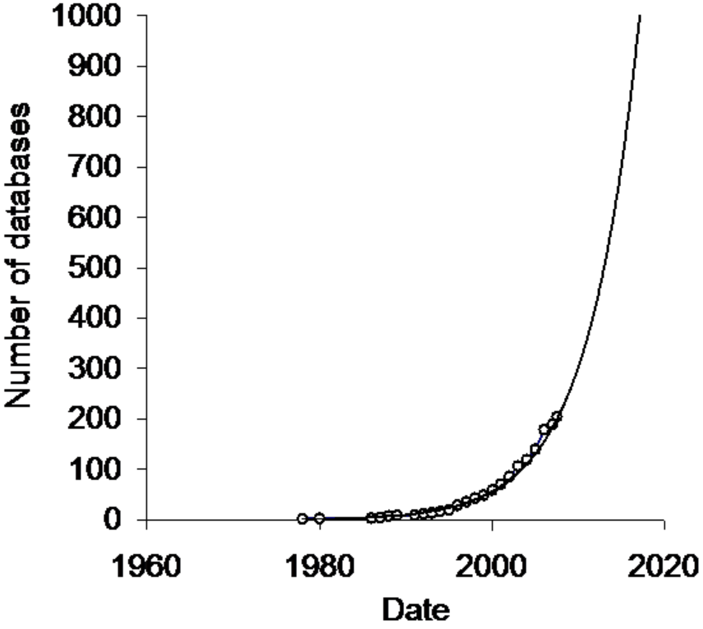


Figure 16. Trend in the number of databases

Big databases

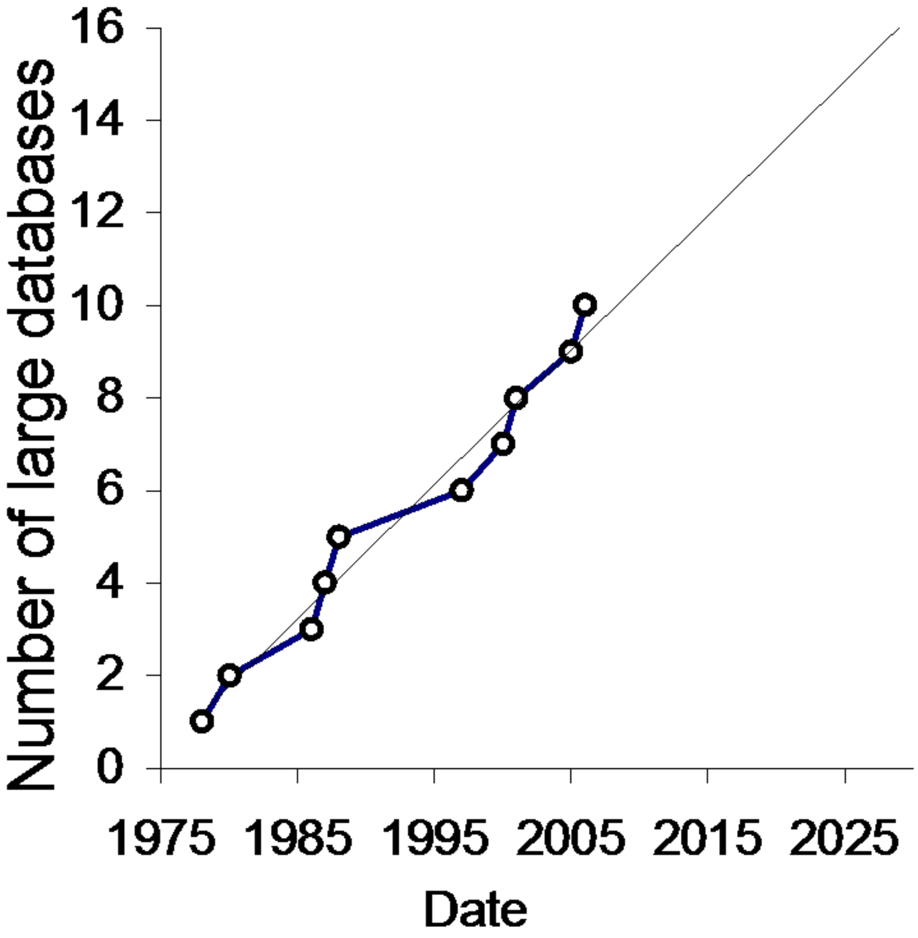
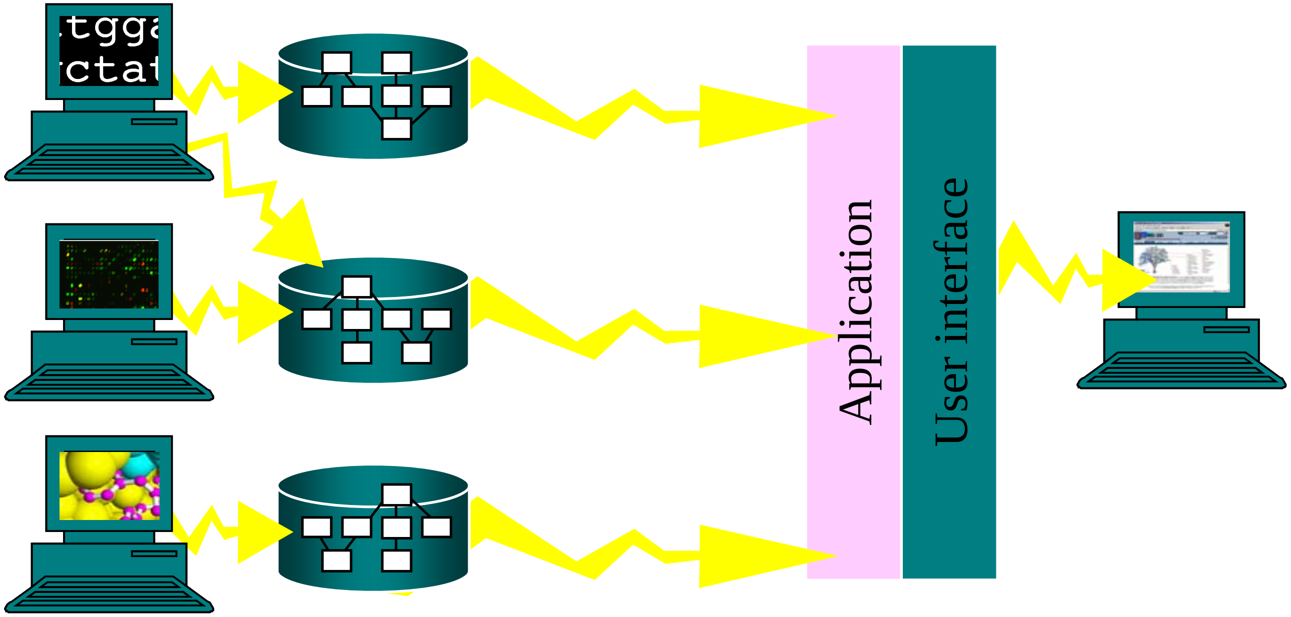


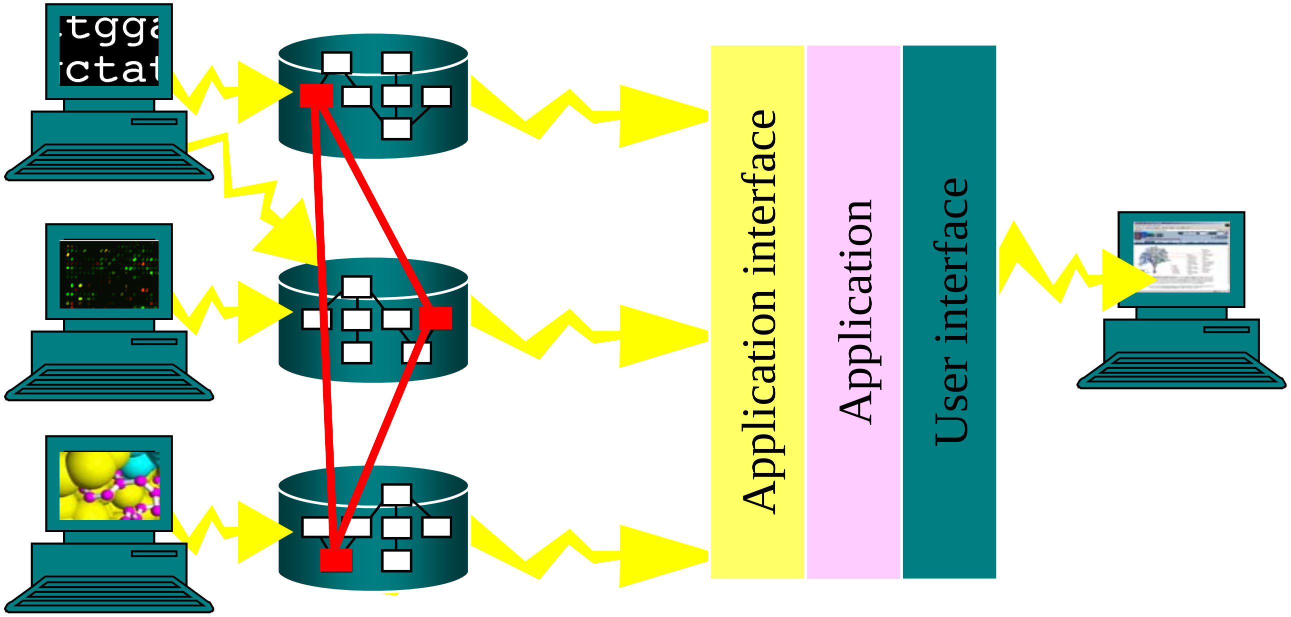
Figure 17. Trend in the number of "large" databases

eScience and interoperability

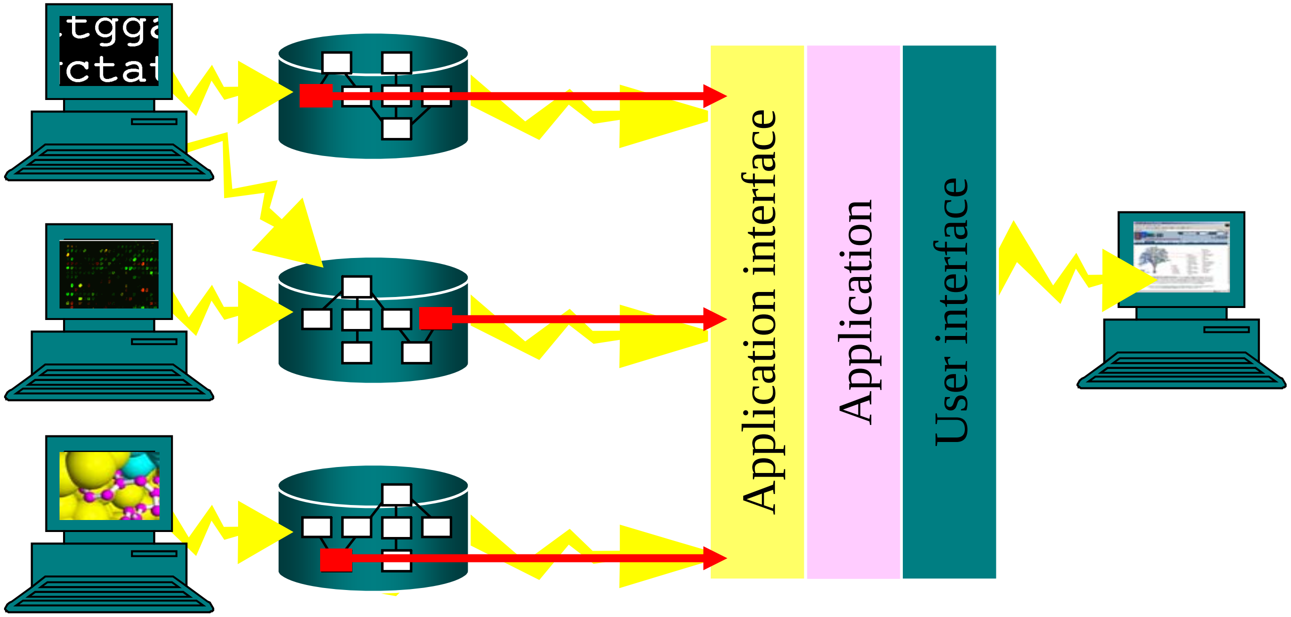
Databases



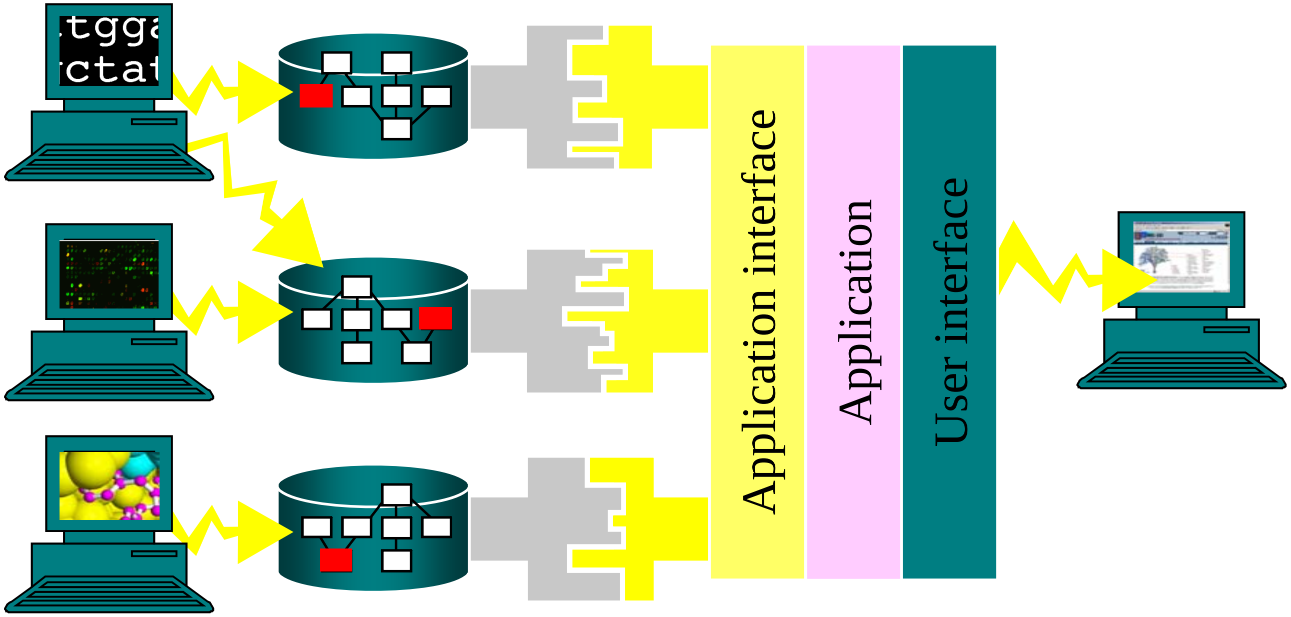
Interoperability



Communicate objects and their identities

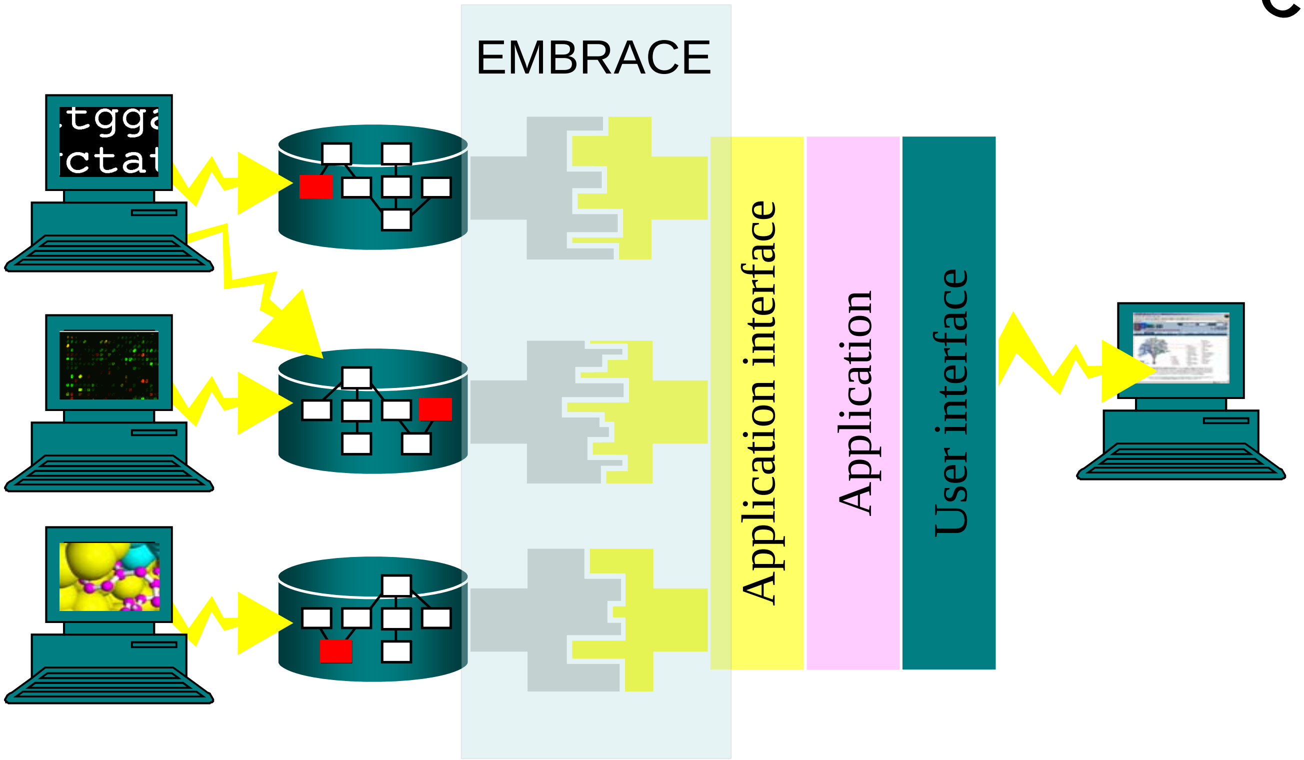


Using standard protocols

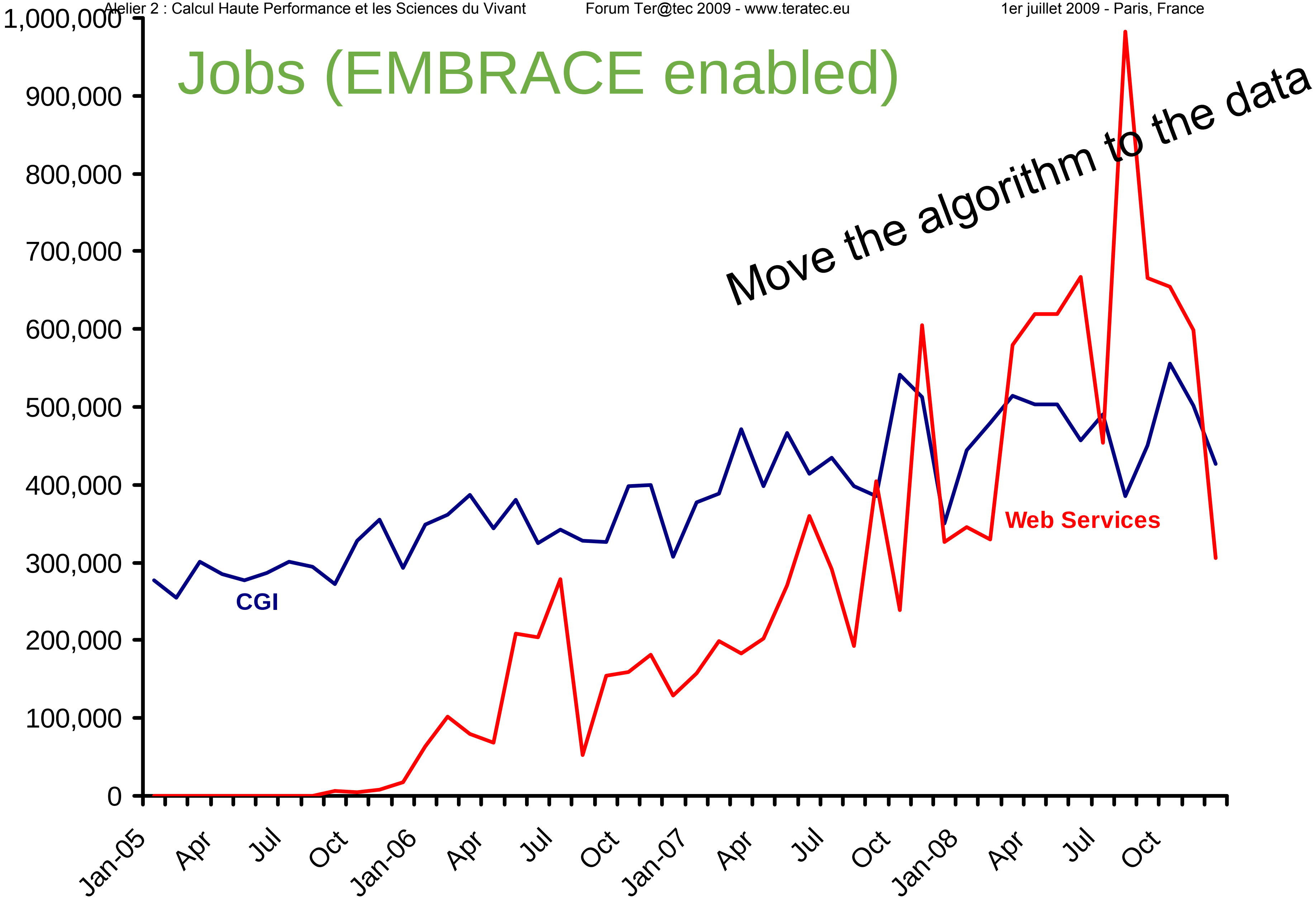


Using standard protocols

Web Services



Jobs (EMBRACE enabled)

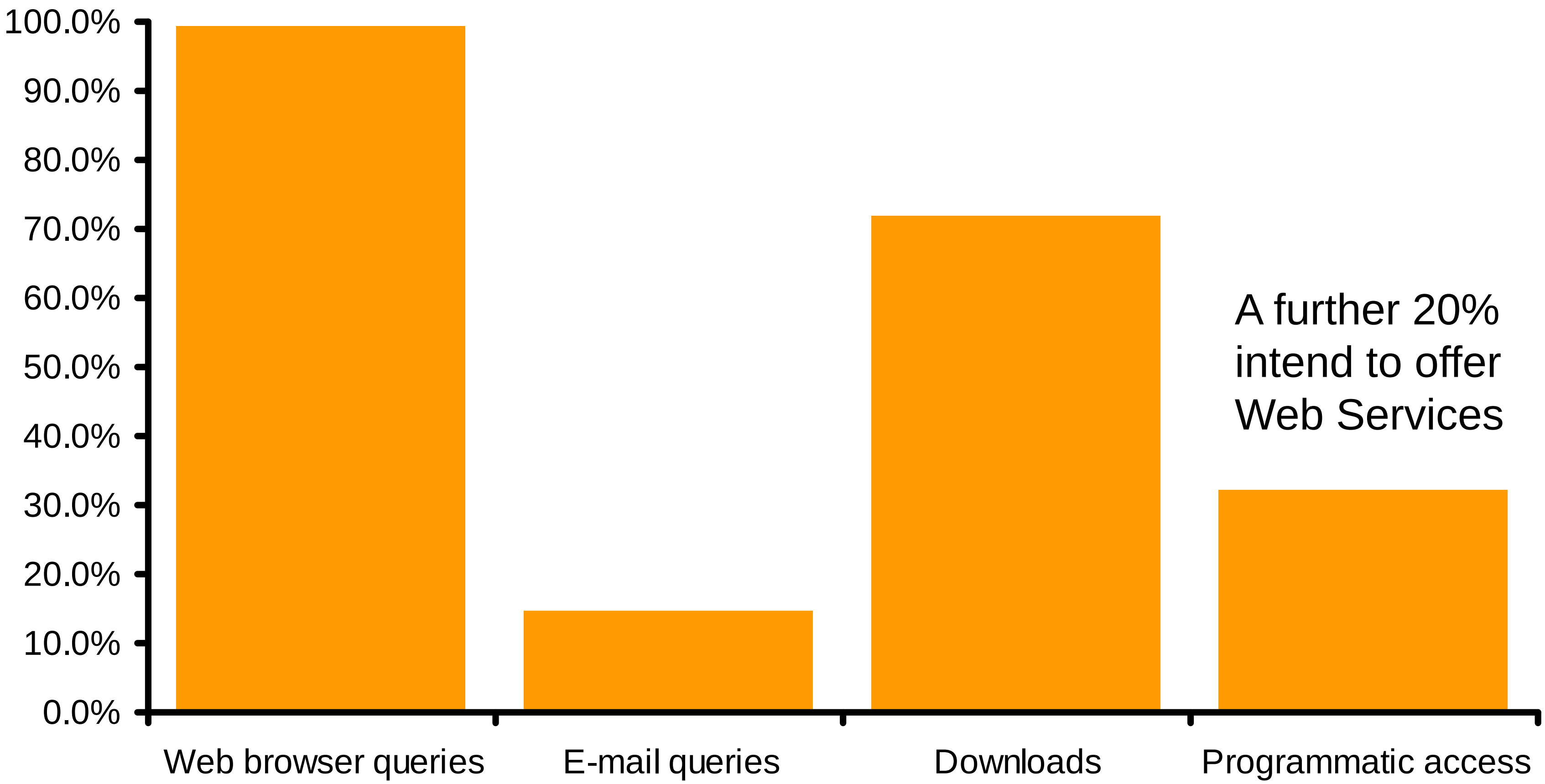


Move the algorithm to the data

Usage (2009 so far)

- 3740119 jobs at EBI
- 60773027 internal jobs at EBI
- Unique users:
  - 2008: 6004
  - 2009: 5865

Modalities (databases surveyed)

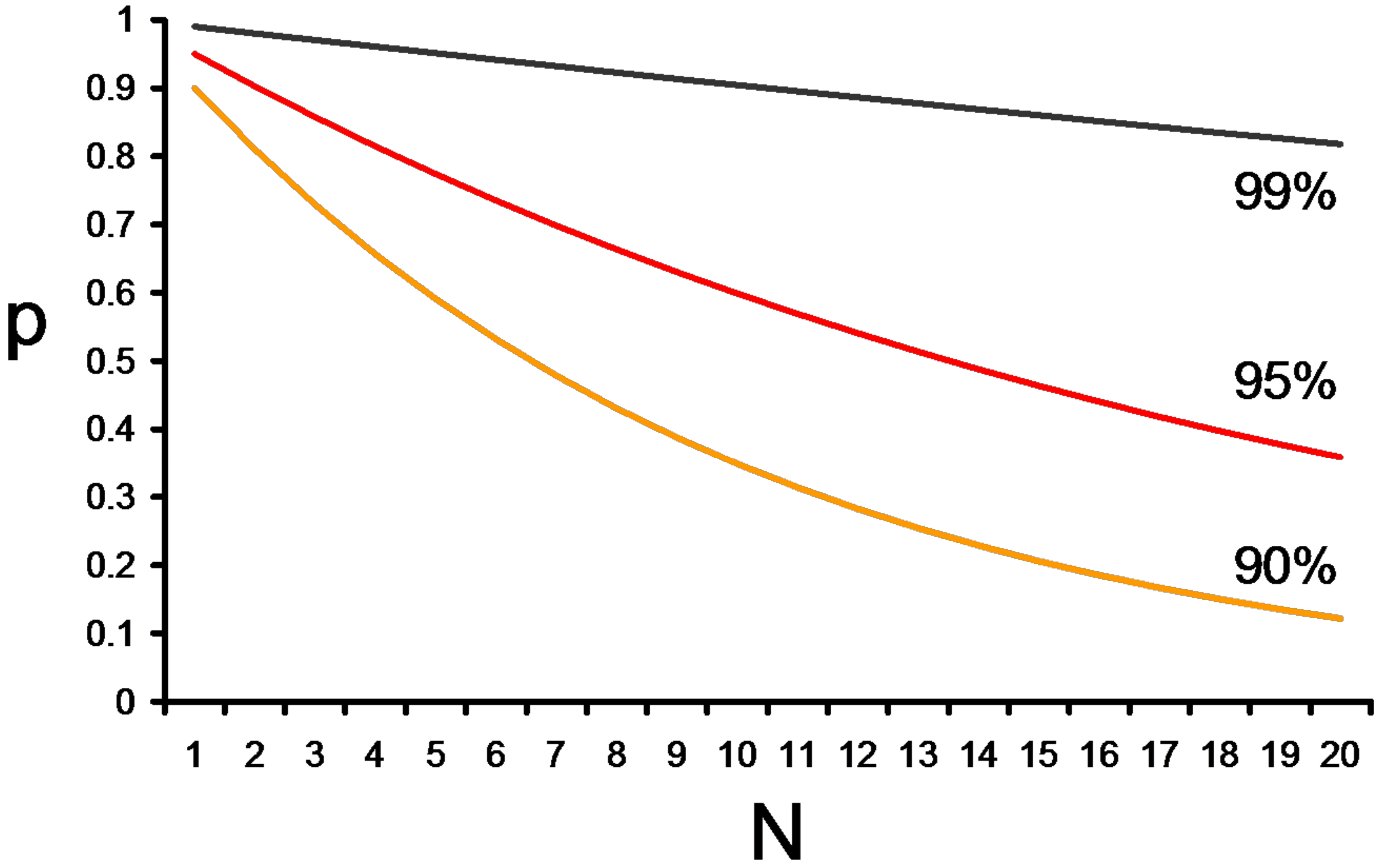


A further 20% intend to offer Web Services

EMBRACE REGISTRY

- 782 services
  - Some of them serve many resources, eg 60 or 70 databases
- Only about 50% of them are from the EMBRACE partners
- www.embraceregistry.net

Reliability



Paying for it all

Paying for it all (public funding)

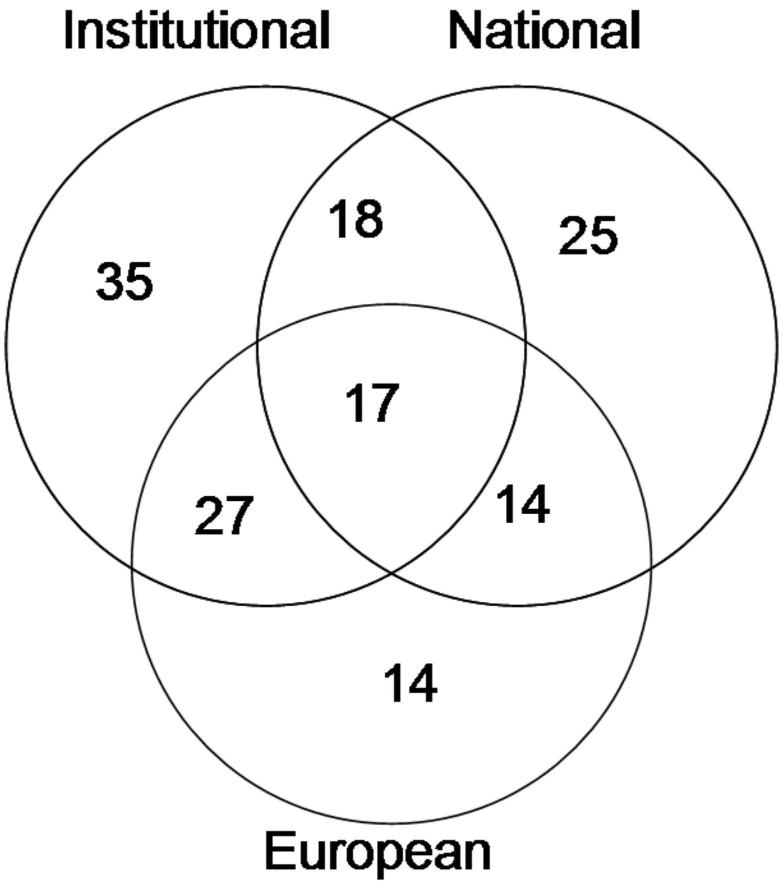


Figure 10. Sources of public funding

Commercial funding

	Has no commercial income	Has commercial Income	Total
Academic but charges commercial users	21	10	31
Free to all	171	6	177
Total	192	16	208

Table 3. breakdown of commercial income.

Costs to date (Europe)

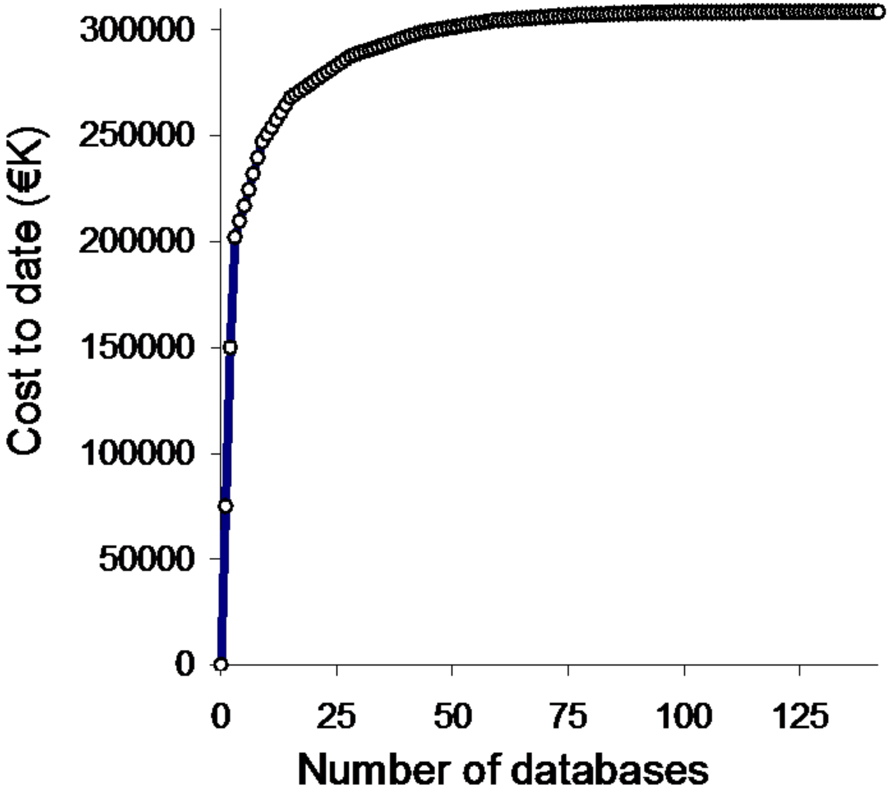


Figure 11. Cumulative cost to date of databases

Annual cost

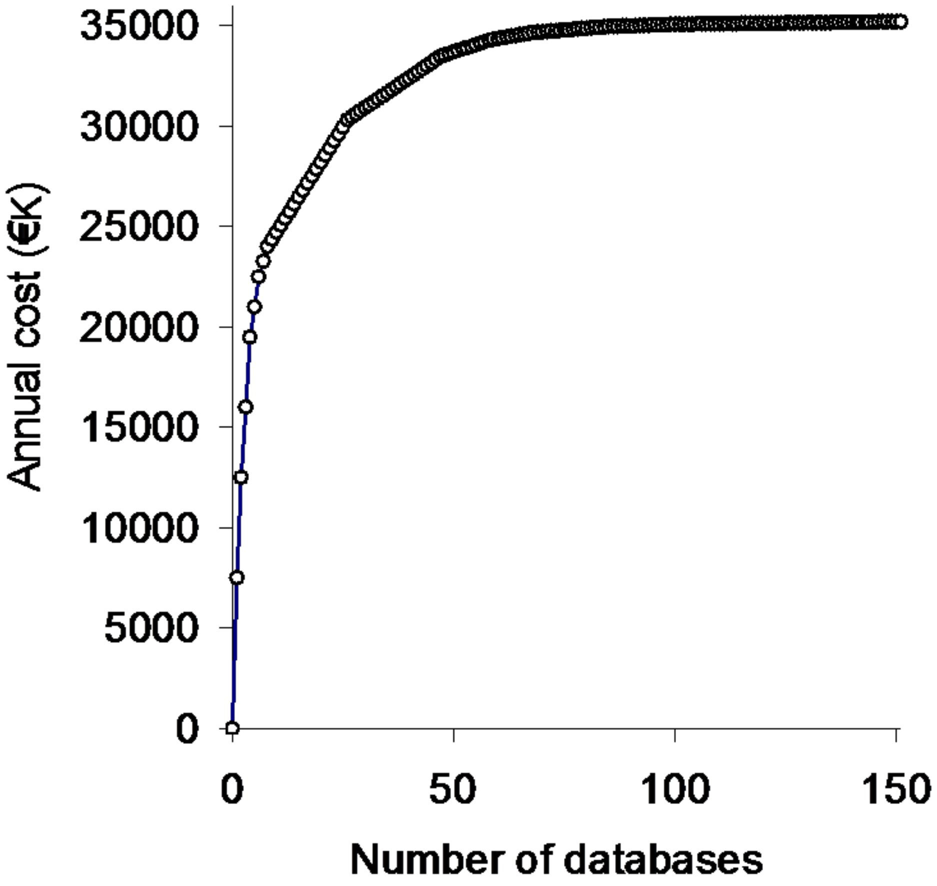
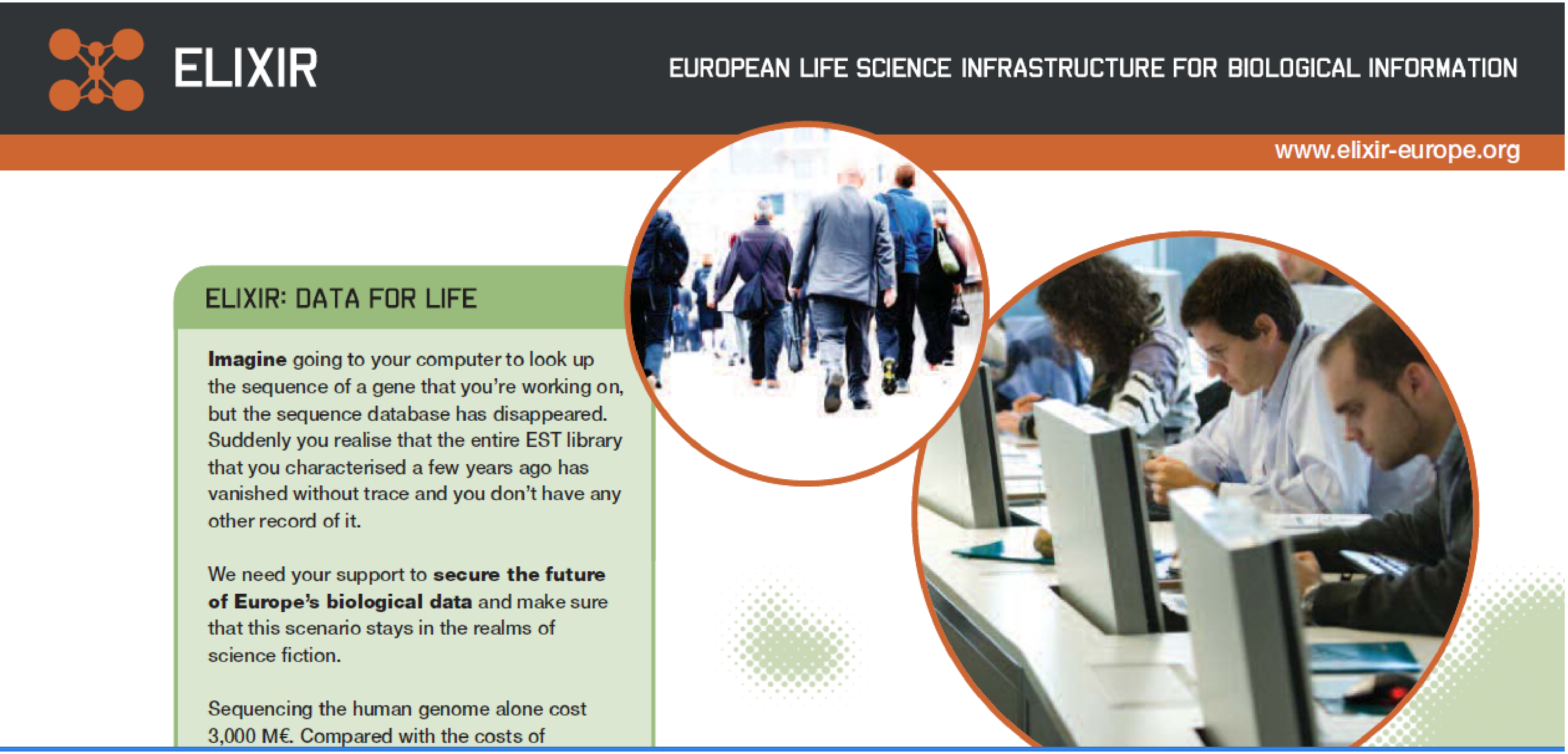


Figure 12. Cumulative annual cost of databases

European context



ELIXIR

EUROPEAN LIFE SCIENCE INFRASTRUCTURE FOR BIOLOGICAL INFORMATION

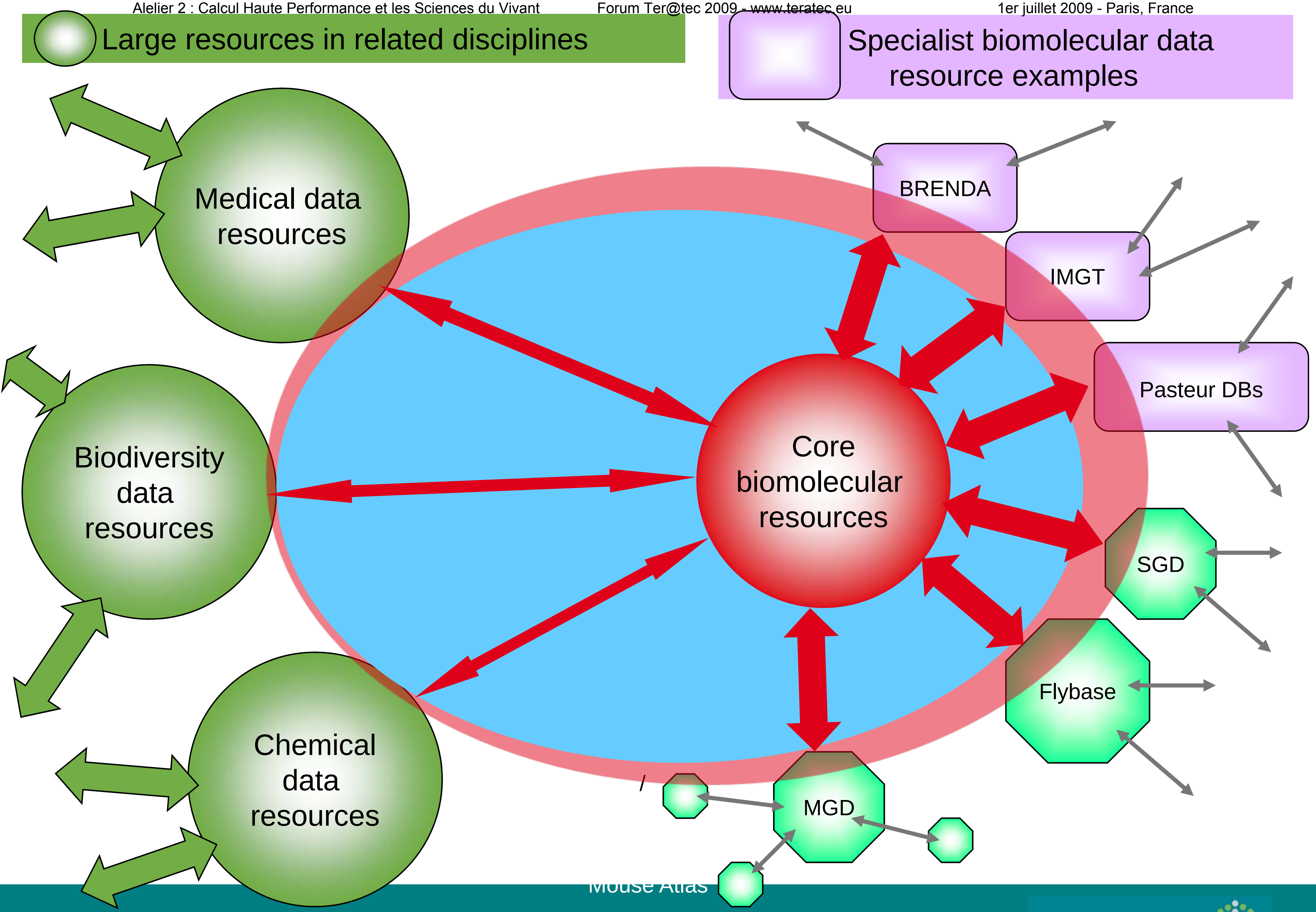
www.elixir-europe.org

ELIXIR: DATA FOR LIFE

Imagine going to your computer to look up the sequence of a gene that you're working on, but the sequence database has disappeared. Suddenly you realise that the entire EST library that you characterised a few years ago has vanished without trace and you don't have any other record of it.

We need your support to **secure the future of Europe's biological data** and make sure that this scenario stays in the realms of science fiction.

Sequencing the human genome alone cost 3,000 M€. Compared with the costs of



Large resources in related disciplines

Specialist biomolecular data resource examples

Medical data resources

Biodiversity data resources

Chemical data resources

Core biomolecular resources

BRENDA

IMGT

Pasteur DBs

SGD

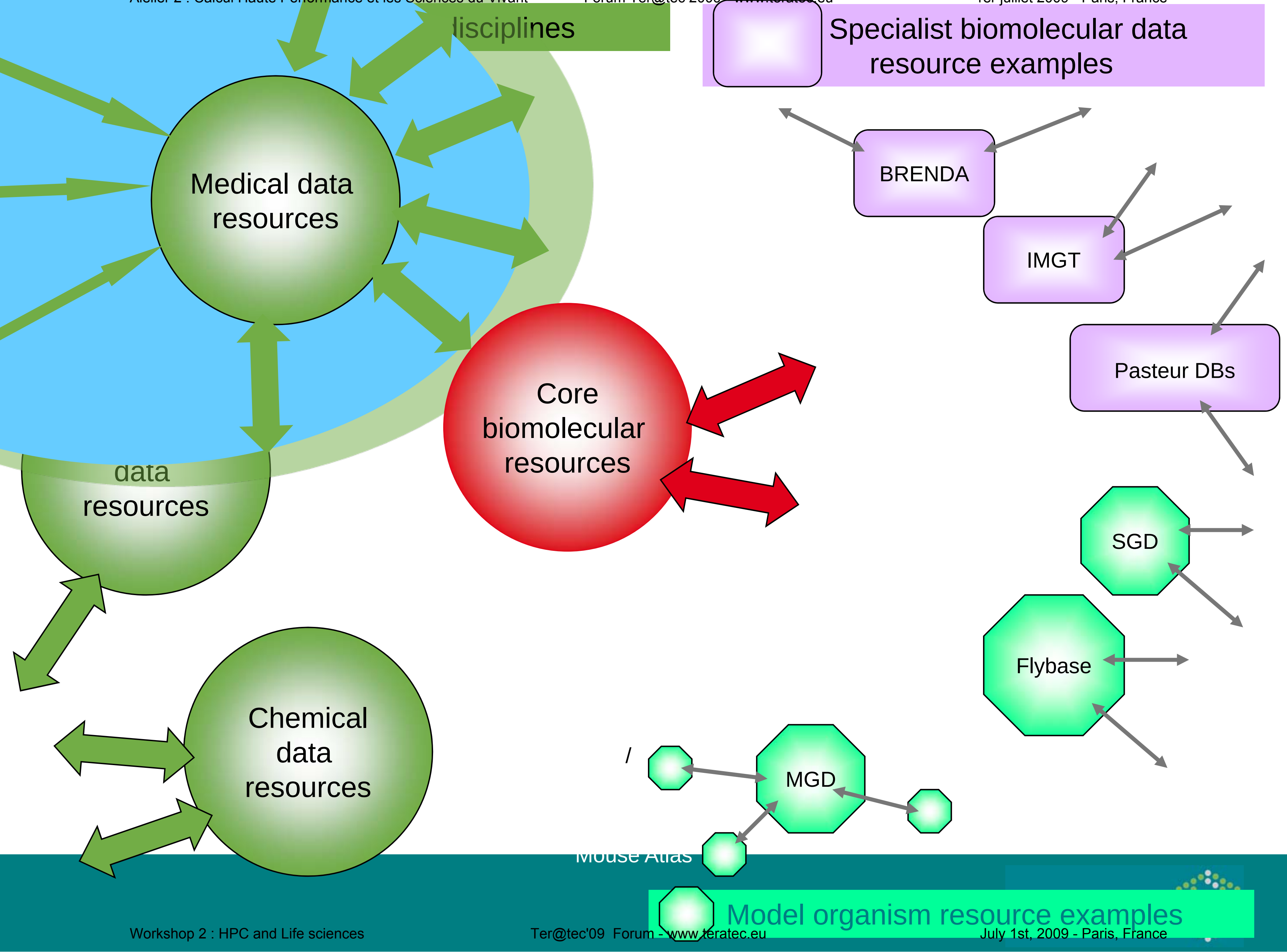
Flybase

MGD

/

Mouse Atlas

Model organism resource examples



lisciplines

Specialist biomolecular data resource examples

BRENDA

IMGT

Pasteur DBs

Medical data resources

data resources

Core biomolecular resources

Chemical data resources

SGD

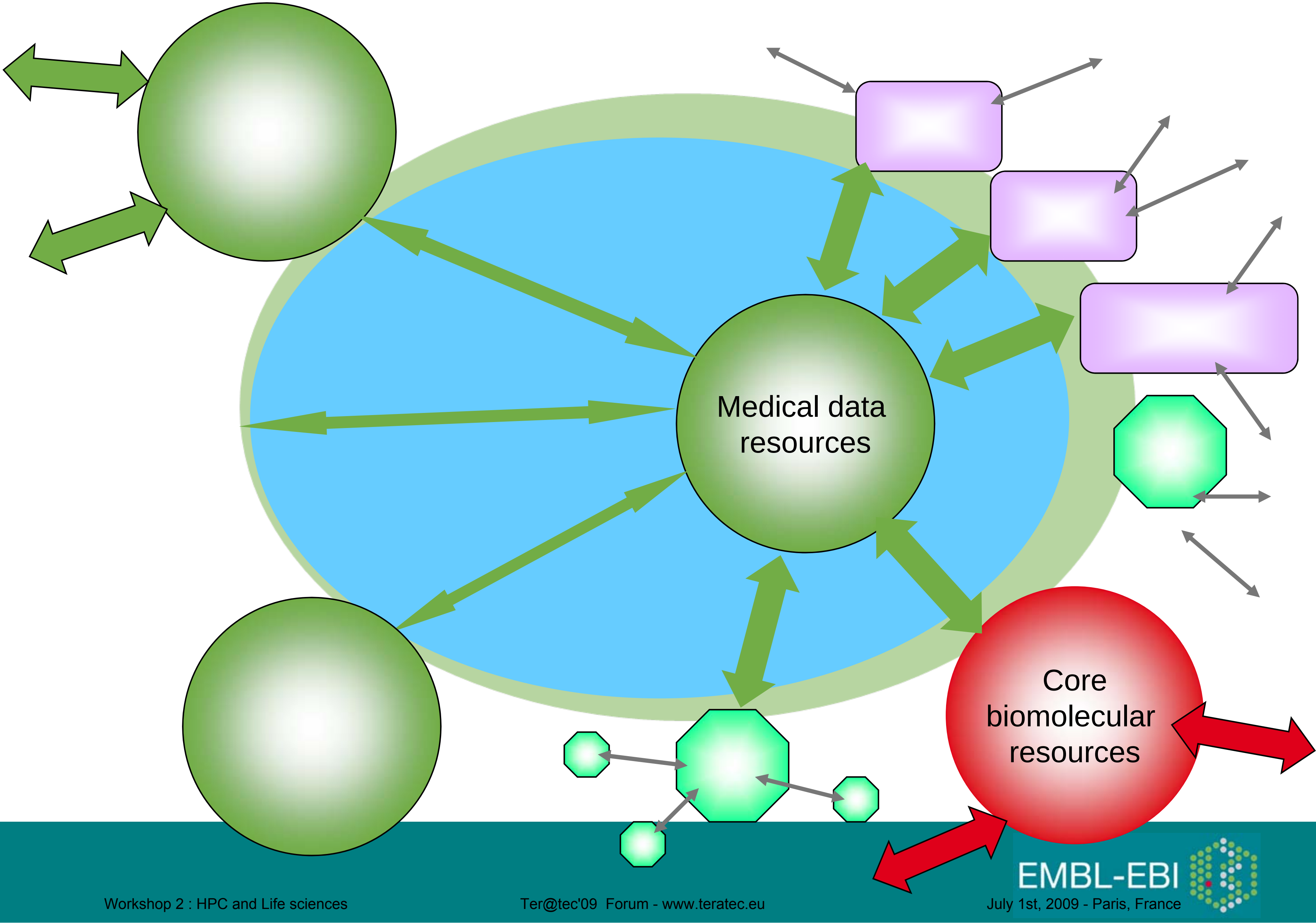
Flybase

/

MGD

Mouse Atlas

Model organism resource examples


Medical data resources
Core biomolecular resources

- EBI
- Chris Southan (Survey)
- Rodrigo Lopez-Serrano (Web services)
- Peter Rice (EMBRACE)

- The scientists

- EMBL
- European Union
- Wellcome Trust
- UK Research Councils
- National Institutes of Health (USA)