ESFRI/e-IRG Collaboration

Yannis Ioannidis

MaDgIK Lab

University of Athens & ATHENA Research Center

MADGIK

National and Kapodistrian UNIVERSITY OF ATHENS
DEPT. OF INFORMATICS & TELECOMMUNICATIONS

Ερευνητικό Κέντρο Αθηνά
Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών, της Γνώσης

ESFRI: European Strategy Forum on Research Infrastructures

- Set up by the EU Council of Research Ministers in 2002
- Brings together representatives of Ministers of the 28 Member States, 9 Associated States, and of the EC
- Supports a coherent and strategy-led approach to policy making on Research Infrastructures
- Mandate to develop a Roadmap (2006) and its updates (2008, 2010)
- Evaluation and new Roadmap in the plans (... 2016)

ESFRI – The Roadmap

- New pan-European RIs or major up-grades to existing Ris
- Needs of European research communities in the next 10-20 years
- All fields of Sciences and Technologies, regardless of possible location
- Published: 2006, updated: 2008 & 2010
  - 48 projects
  - Financial investment: ~20 b€, long term commitment for operation: ~2 b€/year

The ESFRI Process (1)

RI proposal by Member State(s)

ESFRI

TWG1 | TWG2 | TWG3

- Pan-European
- Science case
- Technical case
- Business case
- Overall maturity

Recommended projects and analysis of landscape

The ESFRI Process (2)

EC
Funding of
Preparatory Phase

- Finalise concept and design
- RI Organisation
- Legal framework
- Business plan
- Finance model

ESFRI
Incubator
role

Member States
Funding of Implementation
Phase and operation

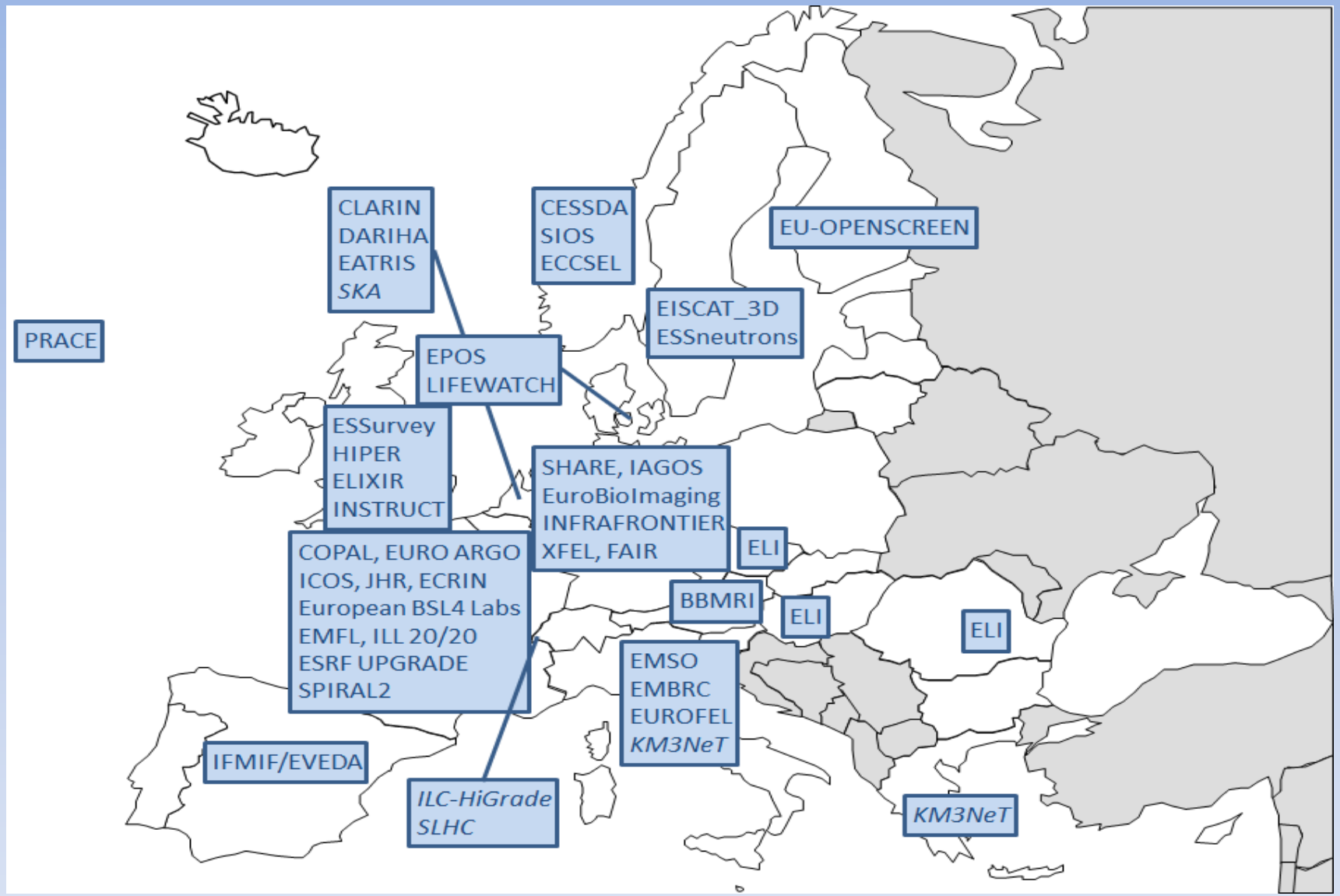
Support
from EC

ESFRI Roadmap 2010

Social Sc. & Hum. (5)	Life Sciences (13)		Environmental Sciences (9)		Energy (7)	Material and Analytical Facilities (6)	Physics and Astronomy (10)		e-Infra-structures (1)
SHARE	BBMRI	ELIXIR	ICOS	EURO-ARGO	ECCSEL	EUROFEL	ELI	TIARA	PRACE
European Social Survey	ECRIN	INFRA FRONTIER	LIFEWATCH	IAGOS	Windscanner	EMFL	SPIRAL2	CTA	
CESSDA	INSTRUCT	EATRIS	EMSO	EPOS	EU-SOLARIS	European XFEL	E-ELT	SKA	
CLARIN	EU-OPENSCREEN	EMBRC	SIAEOS	EISCAT_3D	JHR	ESRF Upgrade	KM3NeT	FAIR	
DARIAH	Euro BioImaging	ERINHA BSL4 Lab		COPAL	IFMIF	NEUTRON ESS	SLHC-PP	ILC-HIGRADE	
	ISBE	MIRRI			HiPER	ILL20/20 Upgrade			
	ANAEE				MYRRHA				

Distributed research infrastructures

Single sited research infrastructures


CLARIN
DARIHA
EATRIS
SKA
CESSDA
SIOS
ECCSEL
EU-OPENSCREEN
PRACE
EISCAT_3D
ESSneutrons
EPOS
LIFEWATCH
ESSurvey
HIPER
ELIXIR
INSTRUCT
SHARE, IAGOS
EuroBioImaging
INFRAFRONTIER
XFEL, FAIR
ELI
COPAL, EURO ARGO
ICOS, JHR, ECRIN
European BSL4 Labs
EMFL, ILL 20/20
ESRF UPGRADE
SPIRAL2
BBMRI
ELI
ELI
EMSO
EMBRC
EUROFEL
KM3NeT
IFMIF/EVEDA
ILC-HiGrade
SLHC
KM3NeT

Many ARE e-Infras

- CLARIN: Language resources … enabling eHumanities
- DARIAH: Digital RI for the Arts & Humanities
- LIFEWATCH: An e-science and tech infrastructure for biodiversity and ecosystem research
- ELIXIR: Platform for biological data
- …
- PRACE: HPC

All need e-Infras

- EURO-ARGO: Ocean floor - not only instruments ... but associated data streams and data centers
- IFMIF: Generation of a materials irradiation database ... for fusion reactors
- BBMRI: Bio-banks, bio-molecular resource centers ... and bio-computational tools
- ...

ELIXIR

Safeguarding the results of life science research in Europe

ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:

medicine

environment

bioindustries

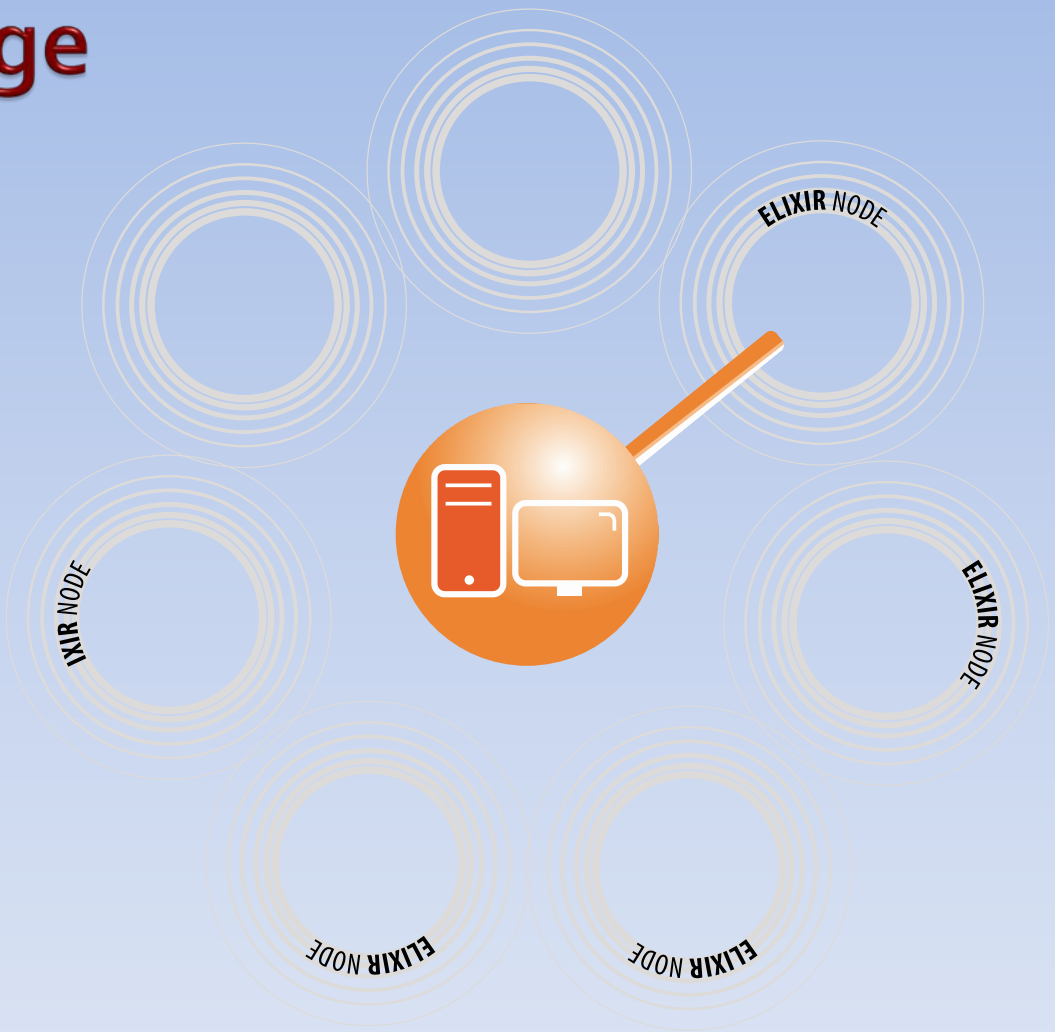
society

A distributed infrastructure to scale with the challenge

ELIXIR data infrastructure for Europe's life science research sector

ELIXIR Nodes build local bioinformatics capacity throughout Europe

ELIXIR Nodes build on national strengths and priorities



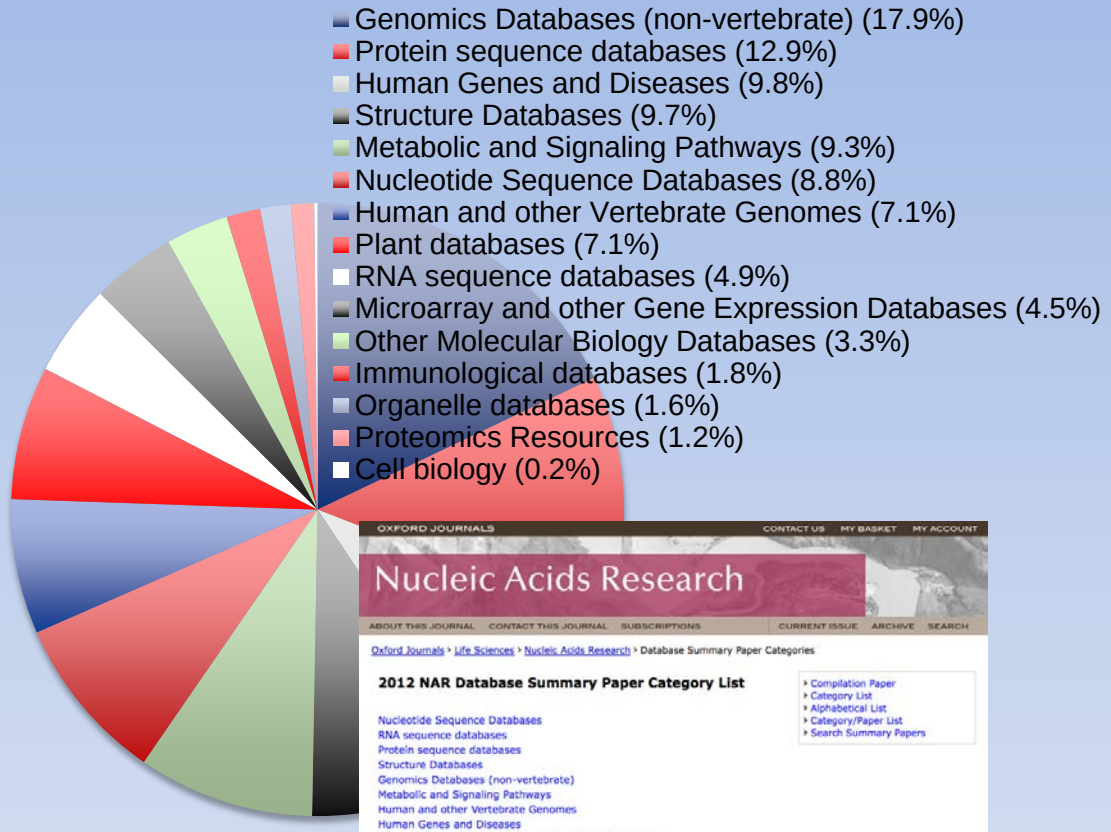
What is bioinformatics?

- The science of storing, retrieving and analysing large amounts of biological information
- An interdisciplinary science involving biologists, biochemists, computer scientists and mathematicians
- **At the heart of modern biology**

New building for ELIXIR Hub now open

Data resources in life science

- Diverse
- Many
- Disperse



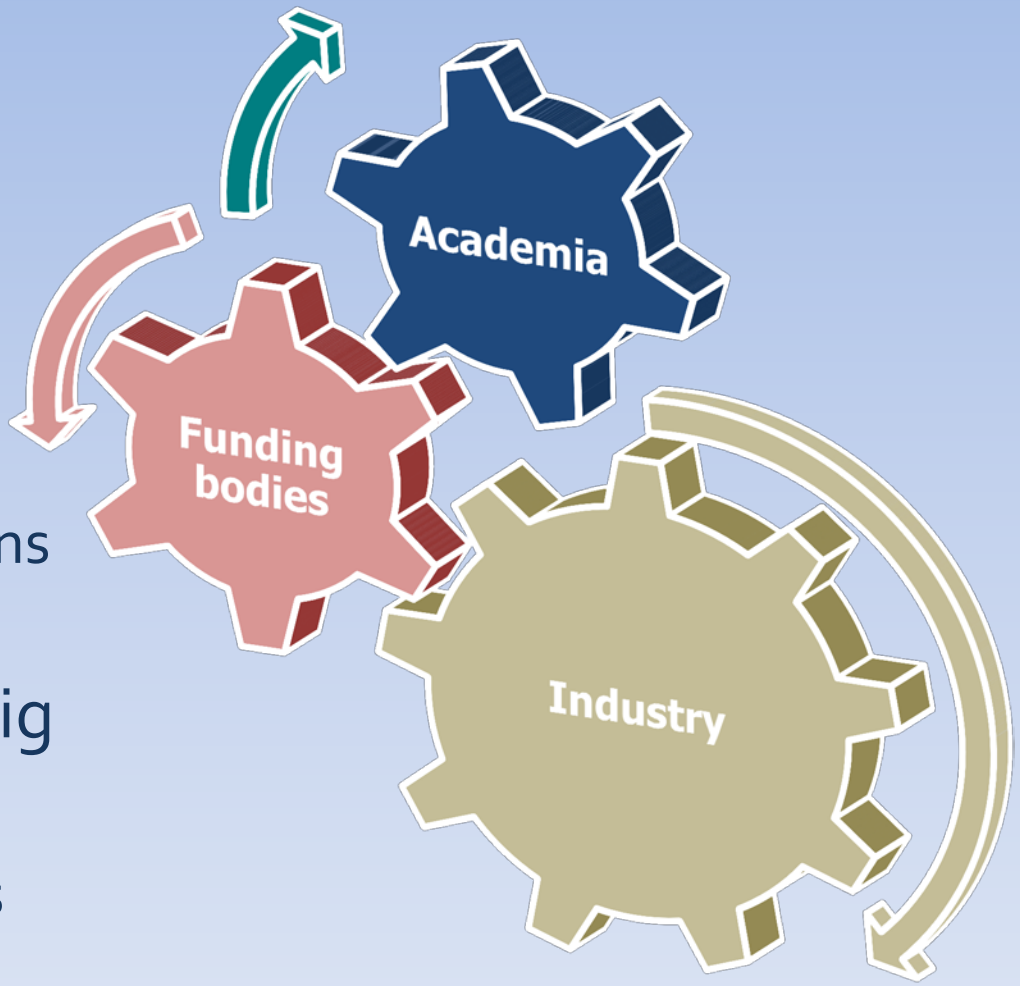
~1800

molecular biology data resources

Nucleic Acids Research annual Database Issue
and the NAR online Molecular Biology Database Collection in 2012.
MY Galperin, GR Cochrane – Nucleic Acids Research, 2011

Industry and Innovation

- Challenge to access and understand services
  - EU wide engagement programme - Build on strong local interactions through Nodes!
- Small companies – Big Data
  - ELIXIR Cloud for SMEs



A Life-science infrastructure of interoperable and integrated data-services

Will the improvement we see in an LPS-challenged mouse translate to a measurable clinical benefit?

elixir

Data interoperability, vocabulary and ontology services

user data

Europe PMC

EGA

Expression Atlas

HPA

Gene Ontology

UniProt

MAD IK

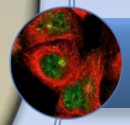
Future challenges for life-science data services

Scale and Sustain funding

Distributed infrastructure with >1M users

Managing and interoperate big and heterogeneous data

Capacity Compute. Capability Storage



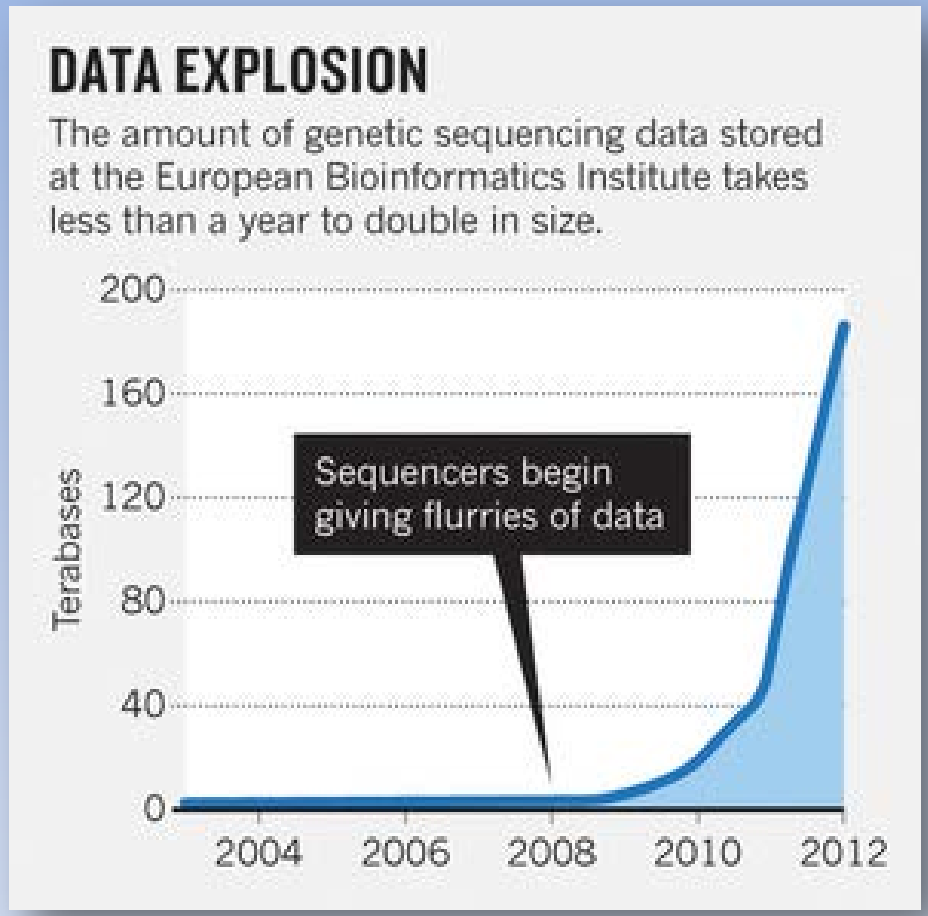
Integrating clinical and translational data

Privacy and ethical concerns

Algorithms to data – clouds, research environments...

The data deluge

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade



Source: *Nature News & Comment*, June 2013

Future scale of genomic data

- 1 Human Genome is 10^9 bases
- Human genome for all UK citizens - 65m x 10^9 = $6x10^{16}$
- SNPs for all UK citizens - 65m x 10^6 = $6x10^{13}$
- 1 petabyte = 10^{15}
- EBI currently has ~16 petabytes of storage

MAD IK

Sequencing every child born in the EU?

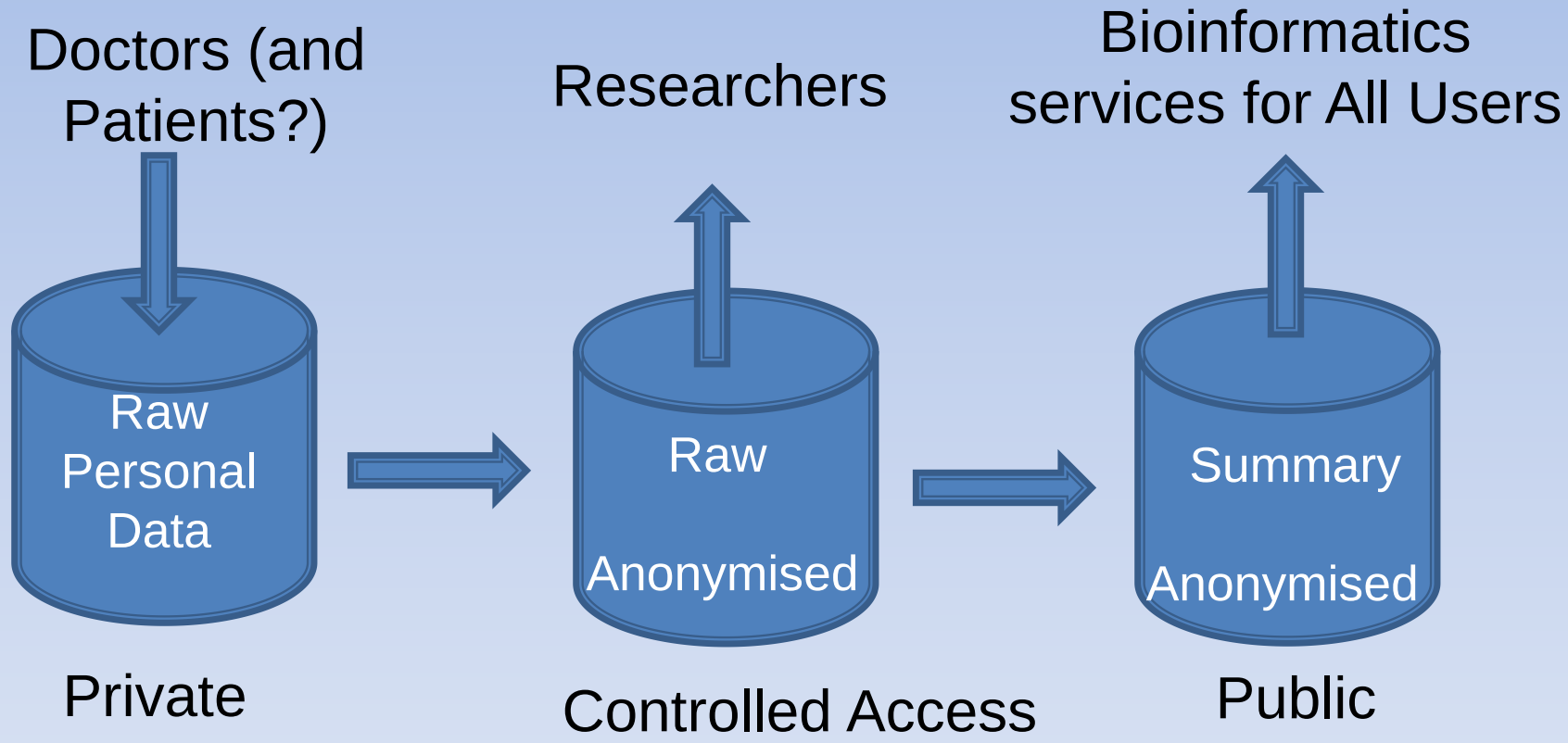
5 million births per year

9 petabases of DNA every week

3 petabytes of raw data every day

Storing only variants: much more feasible

Clinical Data
Raw, Anonymised and Summary Data



The scientific reason for ELIXIR

- Data is an essential commodity for life-science research
- Ten years ago, finding the connection between a gene and a characteristic (e.g., drought tolerance, risk of heart disease) could take years; now it takes minutes

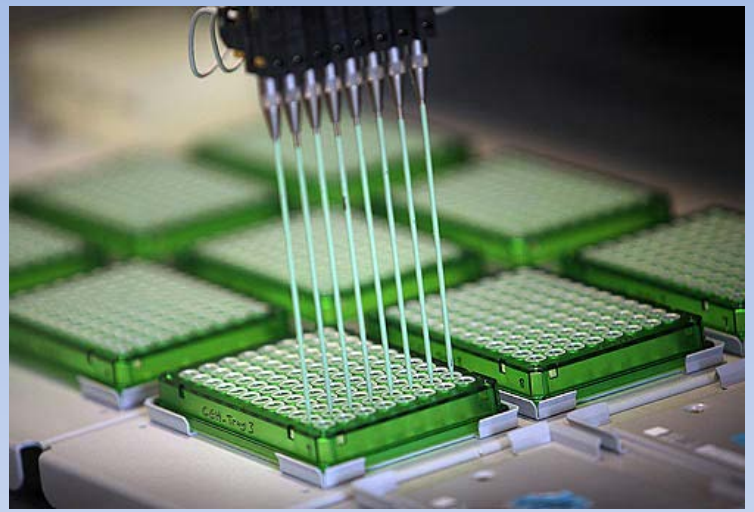


Image courtesy of Genome Research Ltd.

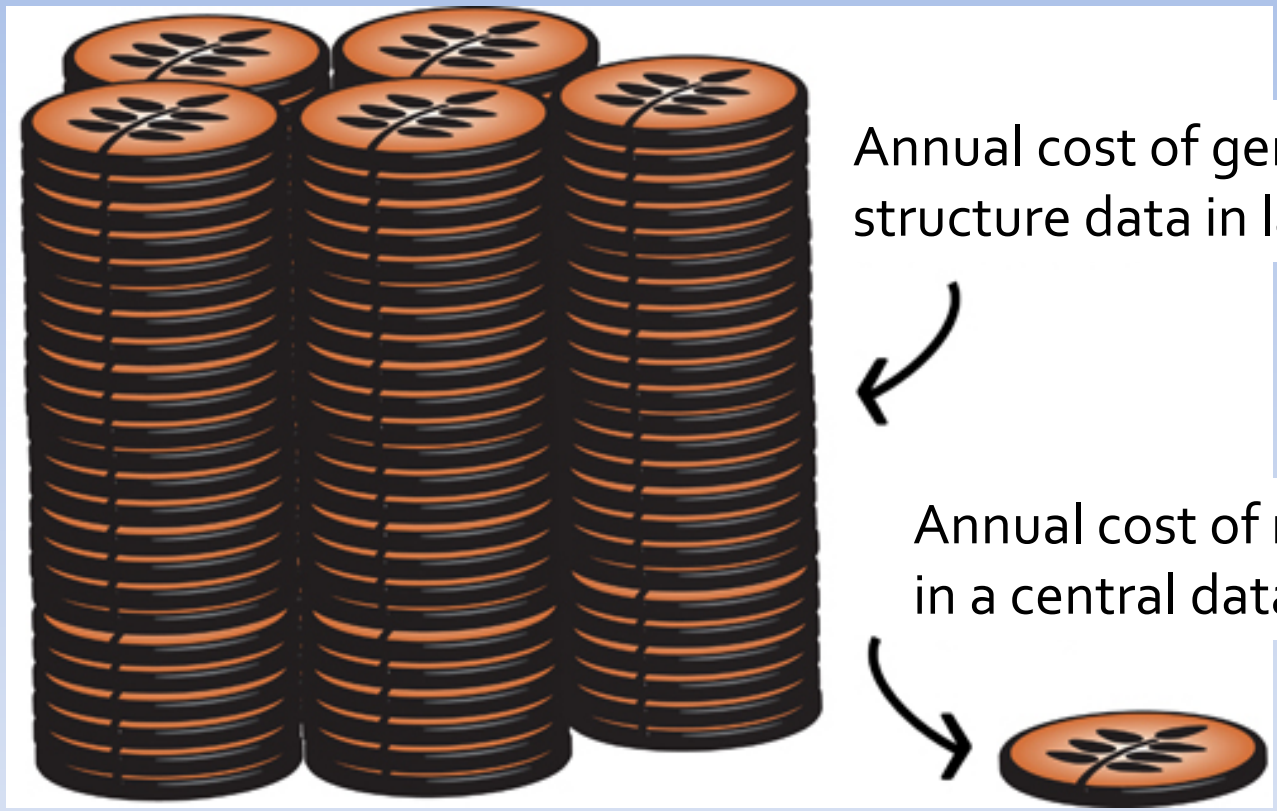
- Data analysis is now the bottleneck in life-science research
- ELIXIR is our only realistic hope of easing that bottleneck

Benefits of ELIXIR

ELIXIR will contribute to European innovation by:

- Optimising access to and exploitation of life-science data
- Ensuring longevity of the data, thereby protecting investments already made in research
- Enhancing the quality of European research by supporting national efforts to increase the competence and number of bioinformatics users through training
- Strengthening the global position and influence of Europe in life-science research in both in academia and industry

Europe has already paid for the science

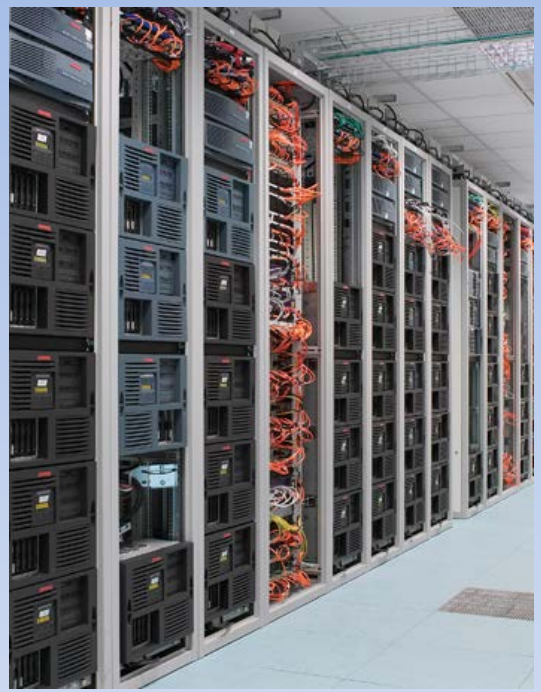


Annual cost of generating new protein structure data in labs around the world

Annual cost of maintaining the data in a central database

ELIXIR's data infrastructure will…

- Enable full data **integration** by making the best use of Europe's collective, expanding capacity
- Establish universal **principles** for **optimisation** of existing data capacity to meet rising demand - for example, assessing which data should be stored and made available to users
- Present a transparent **single interface** to a distributed infrastructure

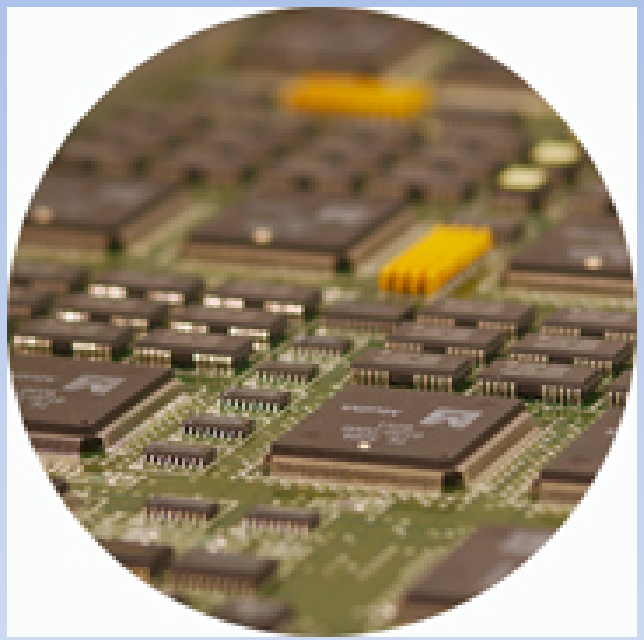


ELIXIR Data Centre

- UK funding has been used to purchase:
  - 5-year lease for Tier-3+ space in commercial high-security data centre in London
  - Compute & networking to establish EBI external services in this secure environment
- Benefits of this option:
  - Provides secure service and 24/7 delivery
  - Electricity requirement can be met
  - Scalable
  - Allows just-in-time purchasing
  - Good networking and connectivity

ELIXIR's compute infrastructure will…

- Optimise existing compute architecture to enable storage of **vast quantities** of homogeneous data
- Life scientists produce an enormous quantity of **heterogeneous** data which needs to be integrated
- An **order of magnitude increase** in compute infrastructure capacity is needed
- Without it, Europe's data infrastructure will **not** be adequate to respond to the data deluge



ELIXIR's tools will…

- Harmonise the staggering number of **analytical software tools**
- Support typical analyses using multiple tools **linked together** into data processing 'pipelines'
- Be developed based on the following principles:
  - discoverability
  - ease of use
  - benchmarking
  - Interoperability

Maintaining open access

- Open access to life science data is essential for advances in many areas of research
- Open access to bioinformatics resources provides a **valuable path to discovery**, one that in many other areas of research is limited by commercial confidentiality
- Charging for that data, or seeking to **restrict access** through exercising Intellectual Property (IP) rights, would impede progress
- ELIXIR will guarantee that open access to biological data is maintained
- Speaking with a **single voice** will strengthen Europe's influence in such global discussions

OpenAIRE

Deposit Publications & data

Visualize - Manage Enhanced Publications

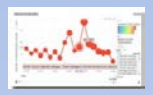
Curate & collaborate

Research impact Citations, usage statistics +++

Search & Browse

Linked Content Statistics +++

Get support (NOADs)



CERN zenodo

Link | Classify

Text Mine

De-duplicate | Cite

APIs

OpenAIRE HUB

Metadata ***Usage data***
on publications

6,500,000 OA publications
345+ validated repositories

Guidelines for Publications

Publication repositories
Institutional & Thematic
Open Access Journals

Guidelines for Funding Info

CRIS Systems

EC funding

National funding

ResearchID (ORCID,

OpenDOAR

...

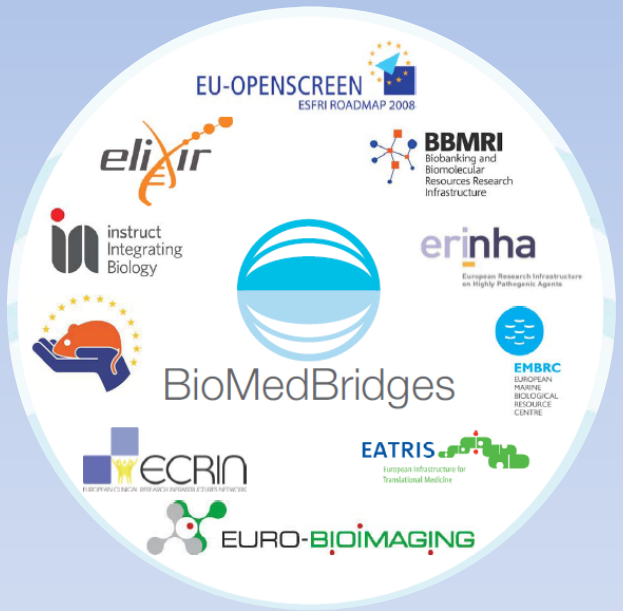
Metadata
on data

Guidelines for Data Providers

Data repositories
Data Journals

ELIXIR bridges life sciences and ICT

Millions of Life Sciences users



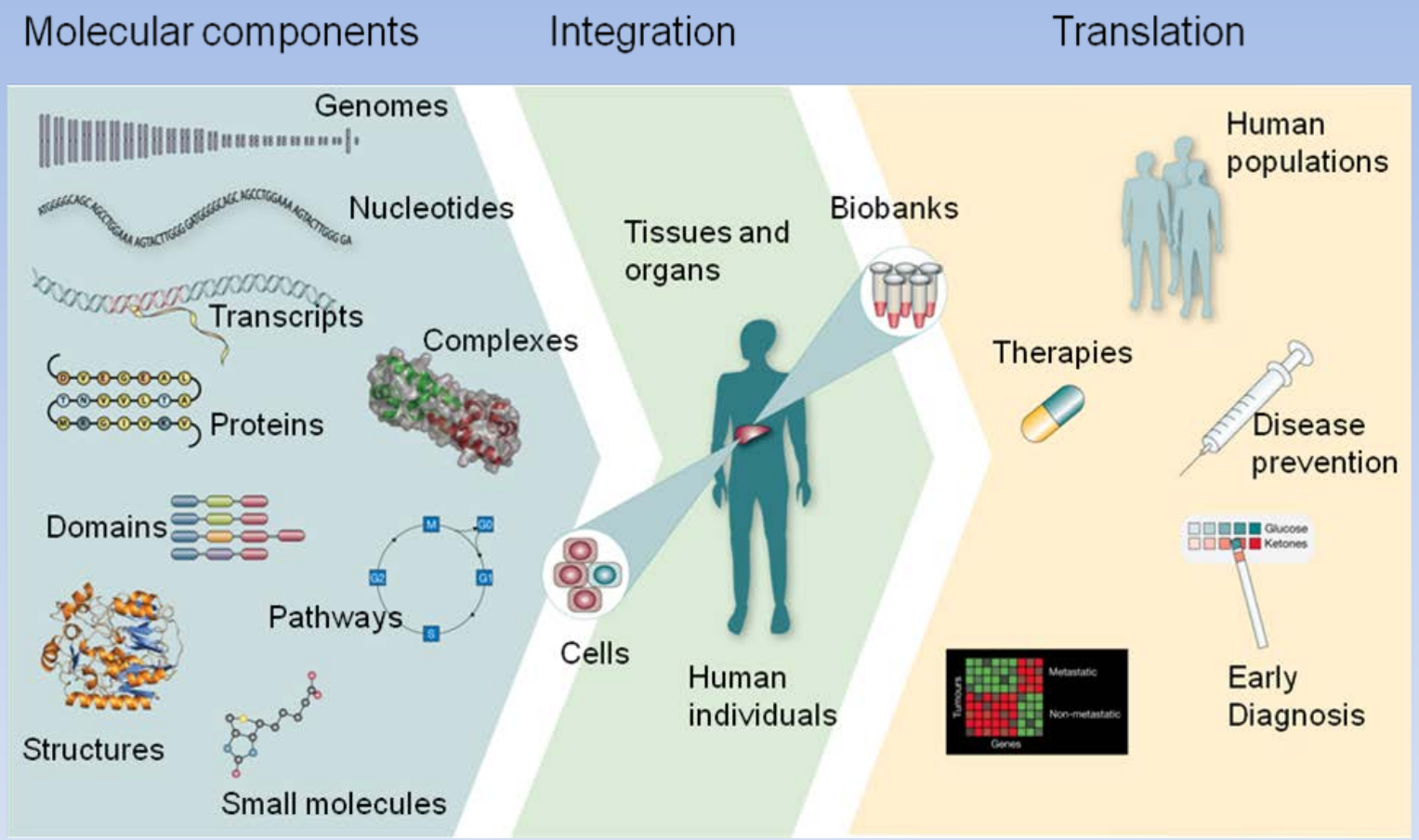
ICT PROVIDERS

GEANT - Network
EGI - GRID
PRACE – Supercomputing
EUDAT – Data Storage
OpenAIRE – OA: Pubs & Data

Other ESFRI Research Infrastructures

ELIXIR coordinates BioMedBridges

From molecules to medicine



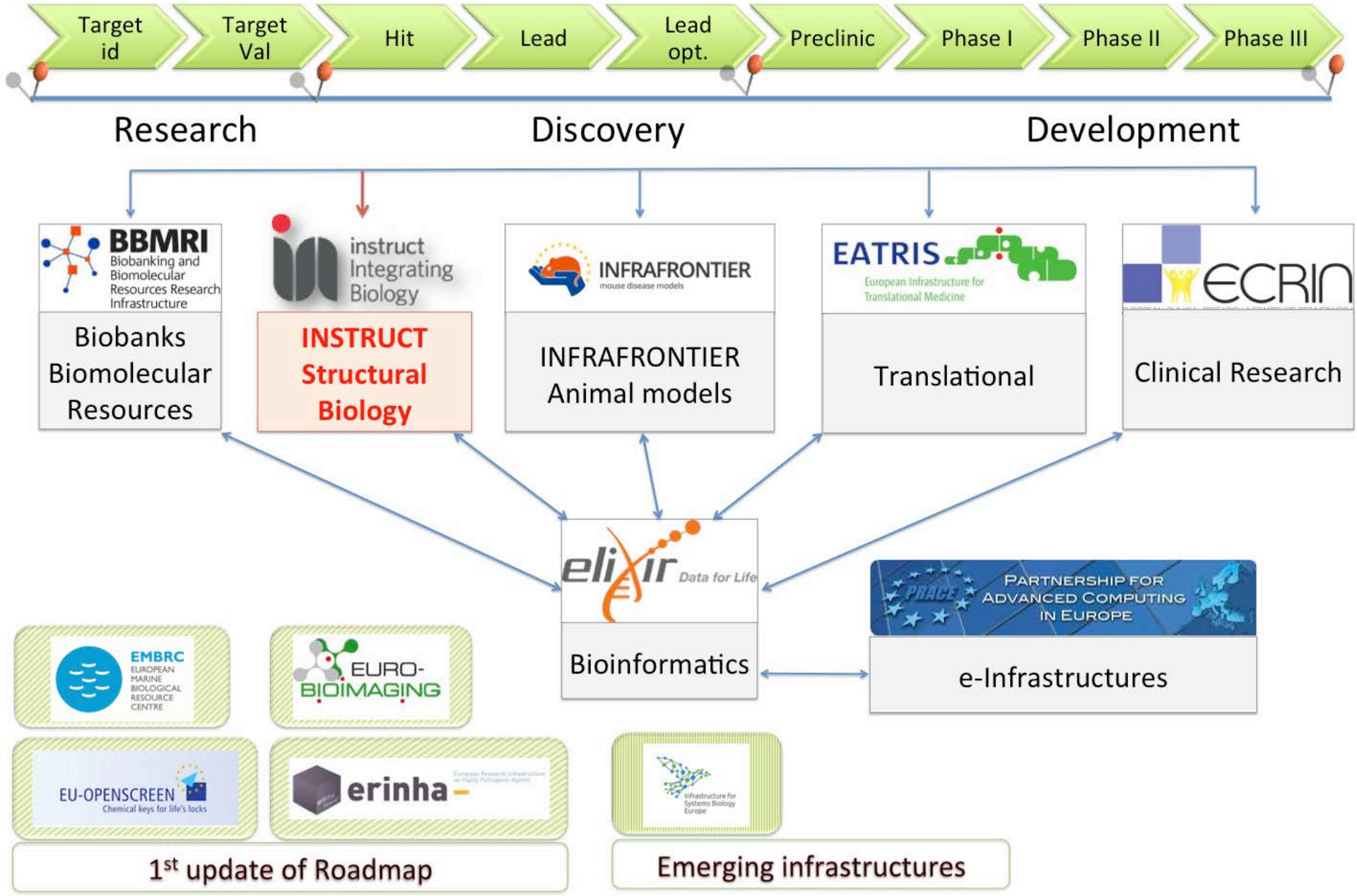
The BMS Research Infrastructures *constitute a well coordinated series of facility providing technologies and expertise relevant in the biological, medical, translational and clinical domains*."

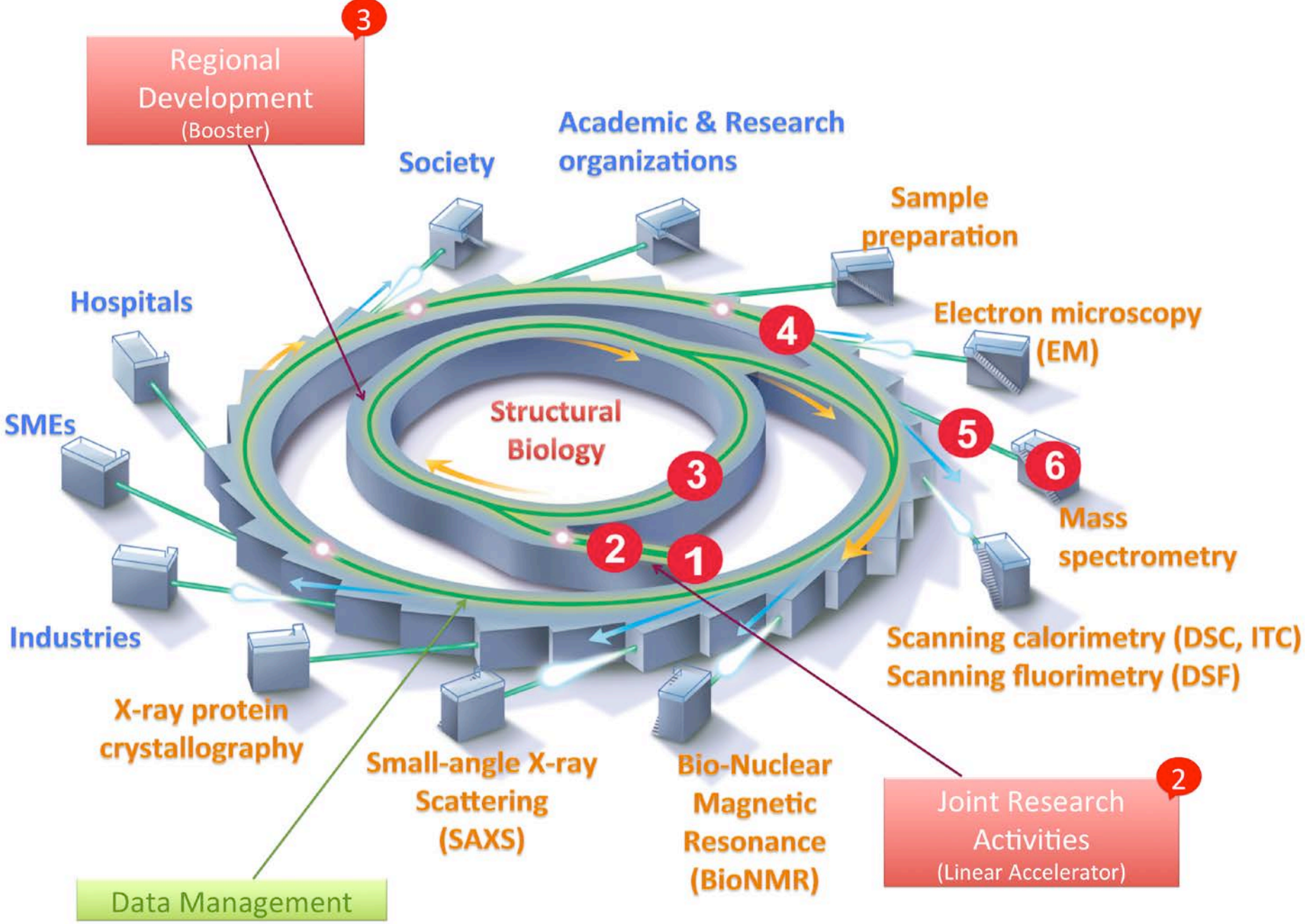
BioMedBridges

HOME AIMS NEWS WORK P

Building data bridges from biology to medicine in Europe

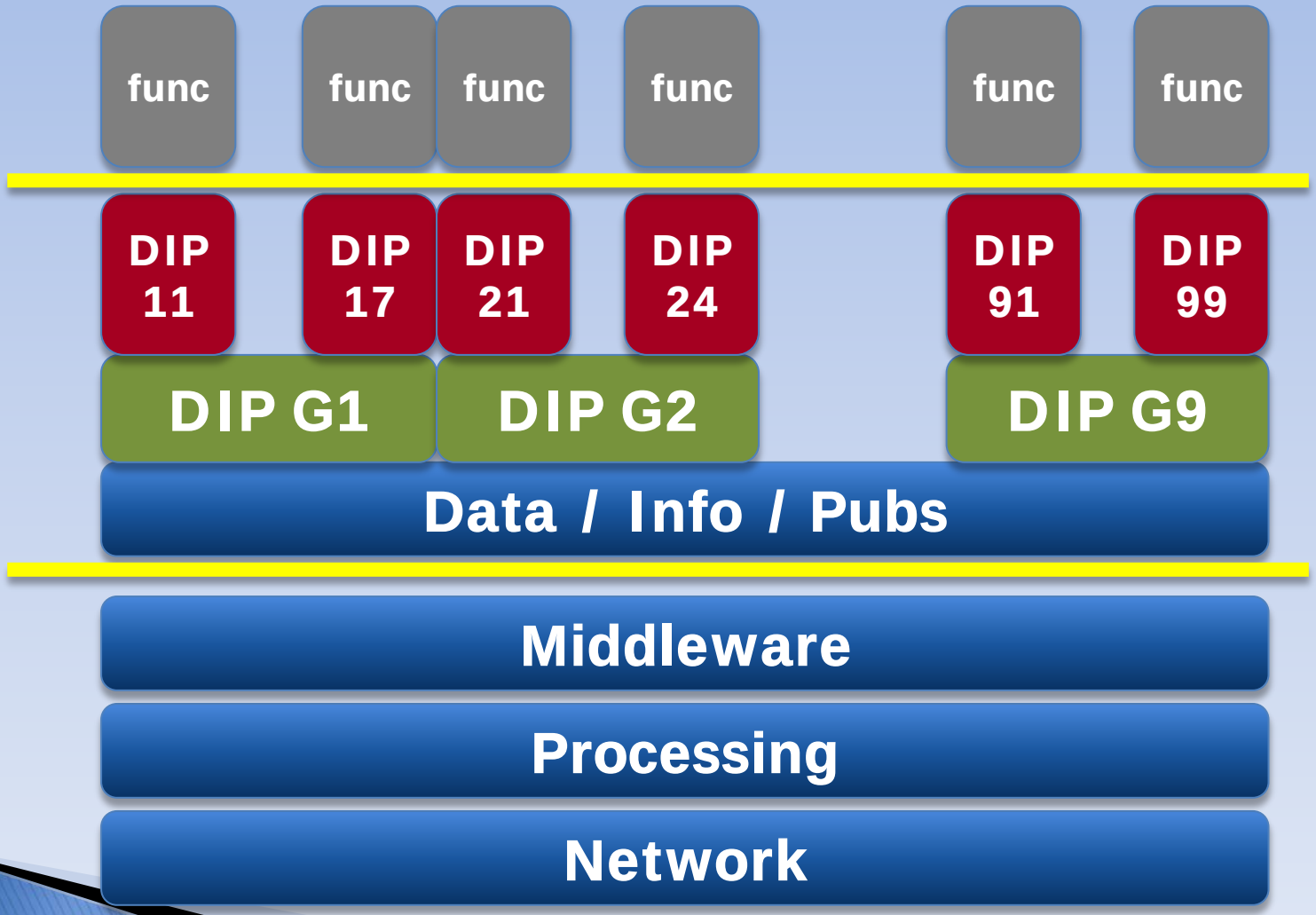
BioMedBridges is a joint effort of twelve biomedical sciences research infrastructures on the ESFRI roadmap. Together, the project partners develop the shared e-infrastructure—the technical bridges—to allow data integration in the biological, medical, translational and clinical domains and thus strengthen biomedical resources in Europe.


Target id
Target Val
Hit
Lead
Lead opt.
Preclinic
Phase I
Phase II
Phase III
Research
Discovery
Development
BBMRI
Biobanking and Biomolecular Resources Research Infrastructure
Biobanks Biomolecular Resources
instruct Integrating Biology
INSTRUCT Structural Biology
INFRAFRONTIER
mouse disease models
INFRAFRONTIER Animal models
EATRIS
European Infrastructure for Translational Medicine
Translational
ECRIN
Clinical Research
elixir Data for Life
Bioinformatics
PRACE
Partnership for Advanced Computing in Europe
e-Infrastructures
EMBRC
European Marine Biological Resource Centre
EURO-BIOIMAGING
EU-OPENSCREEN
Chemical keys for life's locks
erinha
1st update of Roadmap
Infrastructure for Systems Biology Europe
Emerging infrastructures



1. Electron gun 2. Linear accelerator 3. Booster ring
4. Storage ring 5. Beamline 6. End station

From General to Special



Common ESFRI Reqs for e-RIs

- Single sign-on: consistent access to resources
- Virtual organisations (collaboration)
- Persistent storage: long-term preservation of data and its access
- Data Management services
- Standards – web services
- Workflows – support of access to HPC/grid/network resources (compute+data) across Europe
- Training
- Global scope: beyond Europe

Research Paradigms

- Hypothesis ➔ Experiment/Observation ➔ Data collection

- Data collection ➔ Data analysis ➔ Hypothesis

Strategic Decisions Needed

- Technological
- Architectural
- Financial
- Political
- Legal
- Cultural / Social
- ...al

BIG DATA

- Volume (high)
- Velocity (high)
- Variety (great)
- Veracity (lack of)
- Value (hard to extract)

Architectural

- Where does data reside?
- Where does computation occur?

- Past/present researcher common behavior
  - discover datasets through metadata queries
  - download *all* data for local processing
  - keep analysis results locally
- Moving data to computation does not scale
- BIG DATA

Architectural

- Future researcher behavior
  - discover datasets through metadata and data queries
  - upload computation to RI for remote processing
  - keep analysis results at RI
  - keep computation at RI
- Computation and data e-infras together
- Cultural/social issues
- Political/legal issues

Thank you!