Big Data in BioMedical Sciences

Steven Newhouse,

Head of Technical Services, EMBL-EBI

Big Data for BioMedical Sciences

- EMBL-EBI: What we do and why?
- Challenges & Opportunities
- Infrastructure Requirements
- European Context
- The Future

EMBL-EBI

What we do and why?

The European Molecular Biology Laboratory

Heidelberg

Basic research

Administration

EMBO

Hamburg

Structural biology

Hinxton, Cambridge

Bioinformatics

Grenoble

Structural biology

Monterotondo, Rome

Mouse biology

EMBL staff:

1700 people

>60 nationalities

EMBL member states

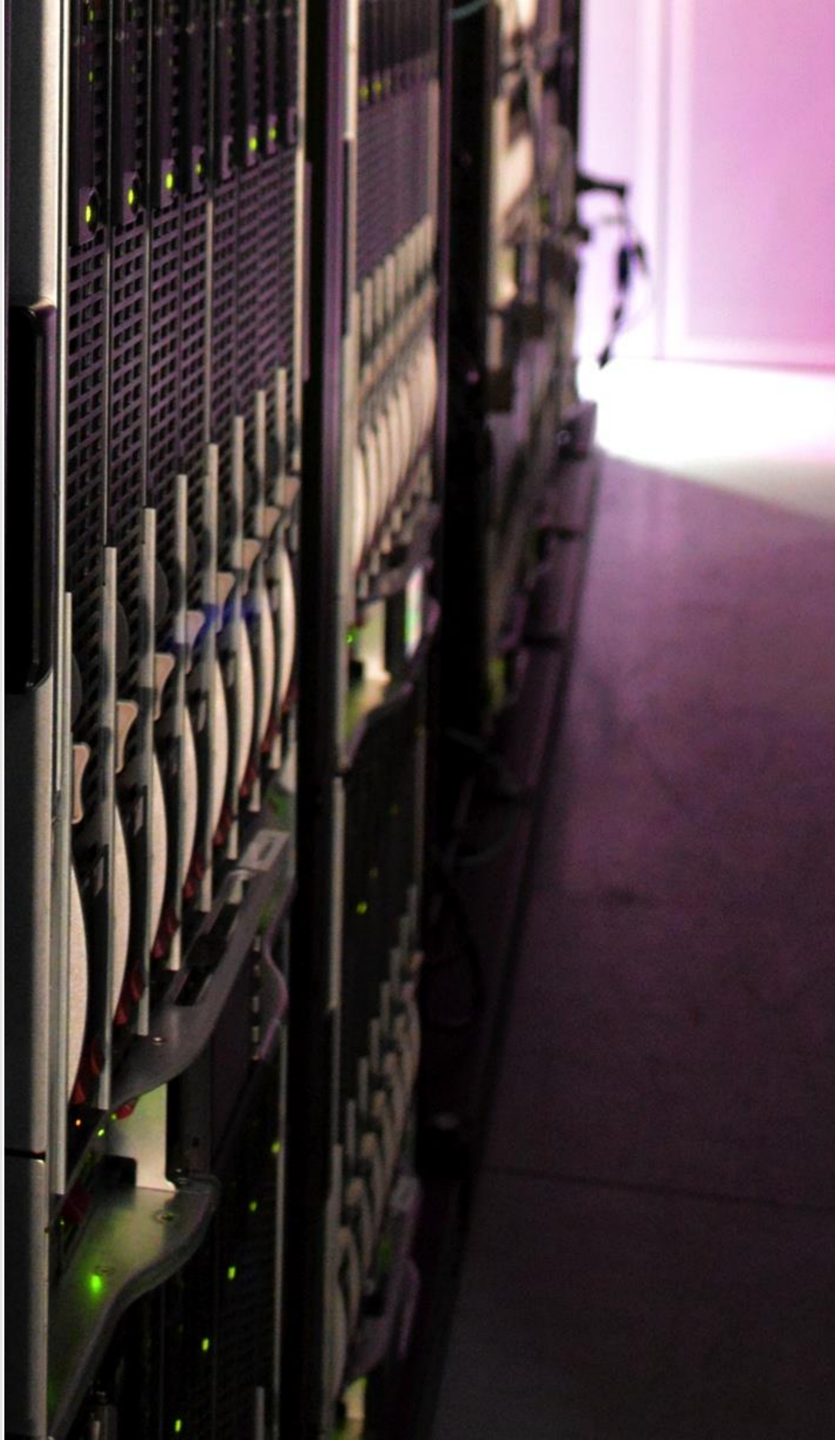
Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom

Associate member states: Argentina, Australia

ICELAND
FINLAND
NORWAY
SWEDEN
DENMARK
IRELAND
UNITED KINGDOM
NETHERLANDS
BELGIUM
GERMANY
LUXEMBOURG
AUSTRIA
FRANCE
SWITZERLAND
CROATIA
ITALY
PORTUGAL
SPAIN
GREECE
Associate Member State
AUSTRALIA
ISRAEL

EMBL-EBI MISSION

To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress



European Bioinformatics Institute (EBI)

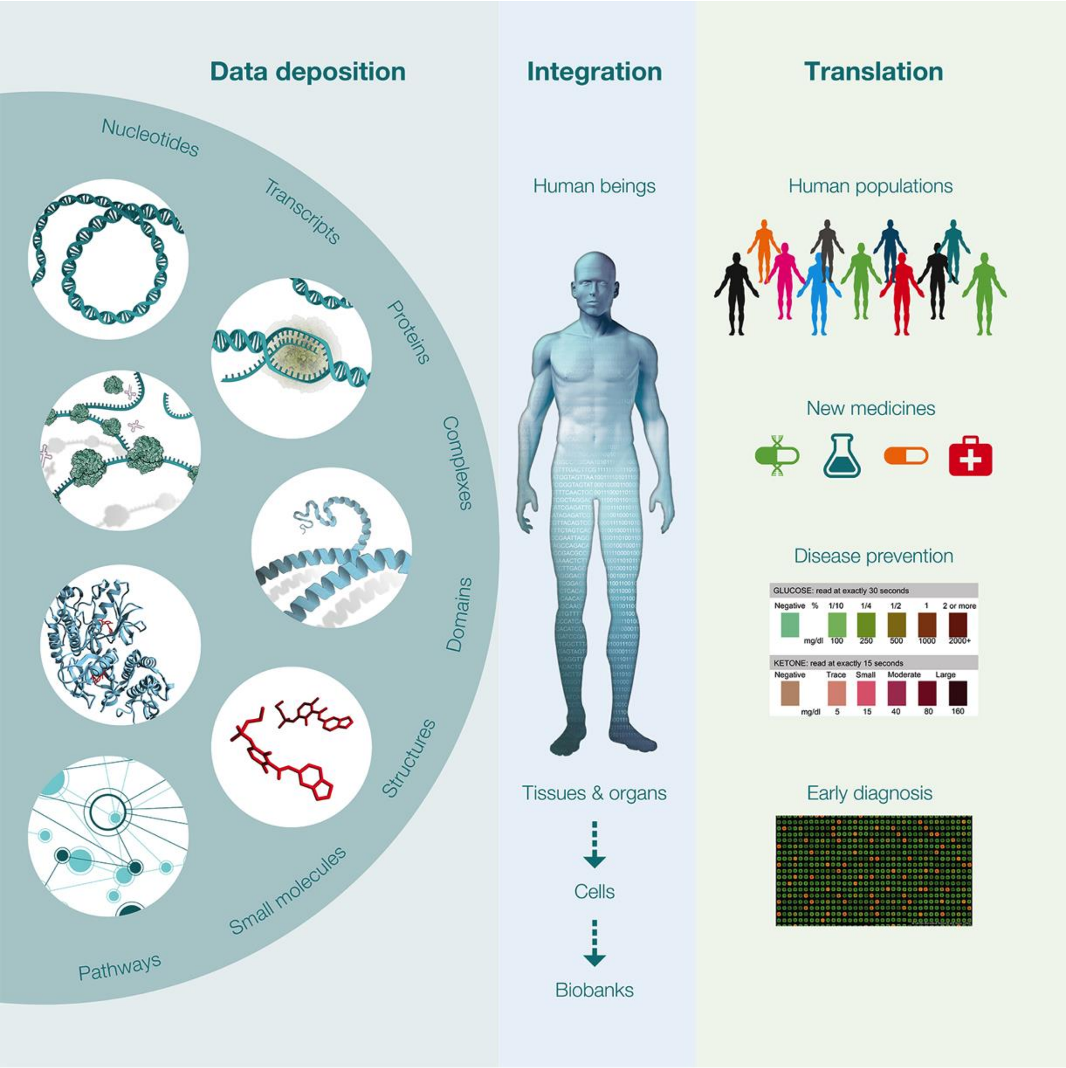
- International, non-profit research institute
- Europe’s hub for biological data services and research
- 500 members of staff from 53 nations
- Funded primarily by member states and research bodies (EC, USA, UK, Wellcome Trust)

What is bioinformatics?

- The science of storing, retrieving and analysing large amounts of biological information
- An interdisciplinary science involving:
  - biologists
  - biochemists
  - computer scientists
  - mathematicians.

Challenges & Opportunities

From molecules to medicine

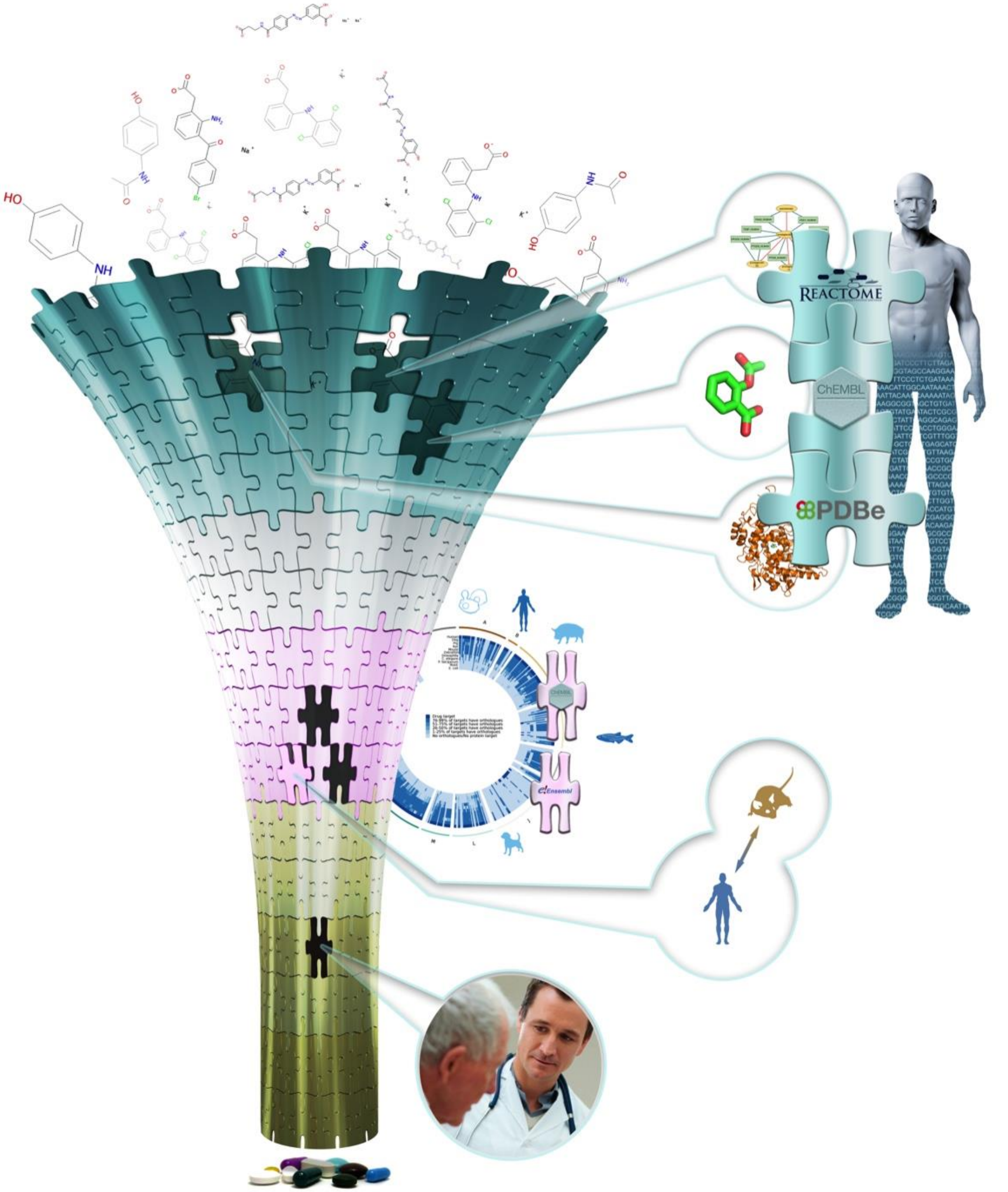


Biology is changing:

- Data explosion
- New types of data
- Emphasis on systems
- Applied biology:
  - molecular medicine
  - agriculture
  - food
  - environmental sciences.

Drug discovery

- From discovering a target to a drug reaching the market: 12 years
- Bioinformatics shortens time to target discovery.
- EMBL-EBI services support all stages of drug discovery.

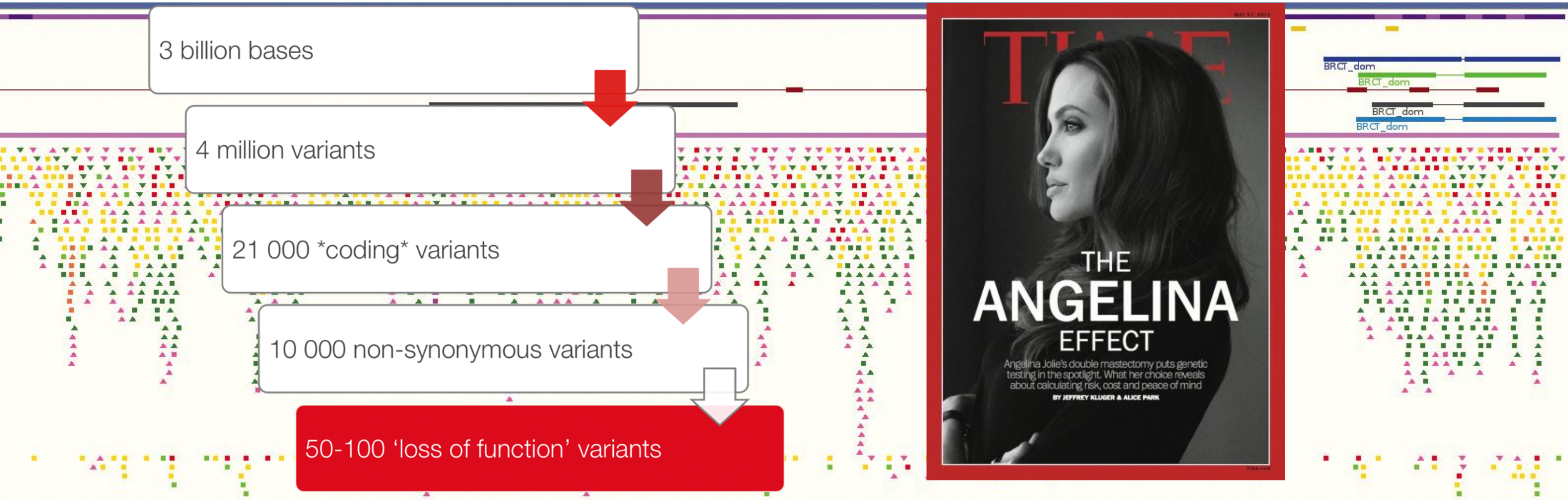


Interpreting human variation

- How and why do we differ from one another?
- What causes susceptibility to disease?
- We explore individual genomes by comparing them with reference genomes.
- Combining that information with other types of data provides valuable insights into human variation.
- This is largely a data-driven process.

Making choices

- There are 3 billion base pairs in the human genome.
- Figuring out which regions are involved in disease – and what they do – is a major challenge.



3 billion bases

4 million variants

21 000 *coding* variants

10 000 non-synonymous variants

50-100 ‘loss of function’ variants

BRCT_dom
BRCT_dom
BRCT_dom
BRCT_dom
TIME
THE ANGELINA EFFECT
Angelina Jolie's double mastectomy puts genetic testing in the spotlight. What her choice reveals about calculating risk, cost and peace of mind
BY JEFFREY KLUGER & ALICE PARK

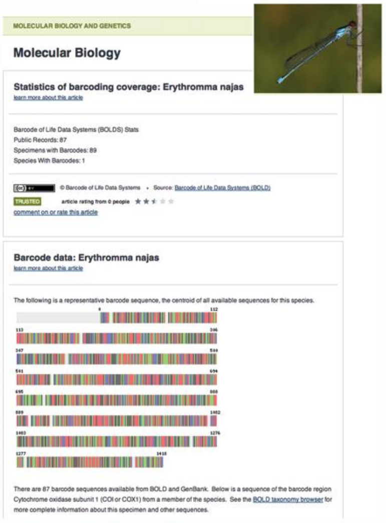
© Time, Inc.

Mapping biodiversity

Species in the wild
- what is it?

Extract DNA,
Generate barcode

Compare barcodes



MOLECULAR BIOLOGY AND GENETICS

Molecular Biology

Statistics of barcoding coverage: Erythromma najas
learn more about this article

Barcode of Life Data Systems (BOLDS) Stats
Public Records: 87
Specimens with Barcodes: 89
Species With Barcodes: 1

© Barcode of Life Data Systems • Source: Barcode of Life Data Systems (BOLD)
TRUSTED article rating from 0 people ★★☆☆☆
comment on or rate this article

Barcode data: Erythromma najas
learn more about this article

The following is a representative barcode sequence, the centroid of all available sequences for this species.

There are 87 barcode sequences available from BOLD and GenBank. Below is a sequence of the barcode region Cytochrome oxidase subunit 1 (COI or COX1) from a member of the species. See the BOLD taxonomy browser for more complete information about this specimen and other sequences.

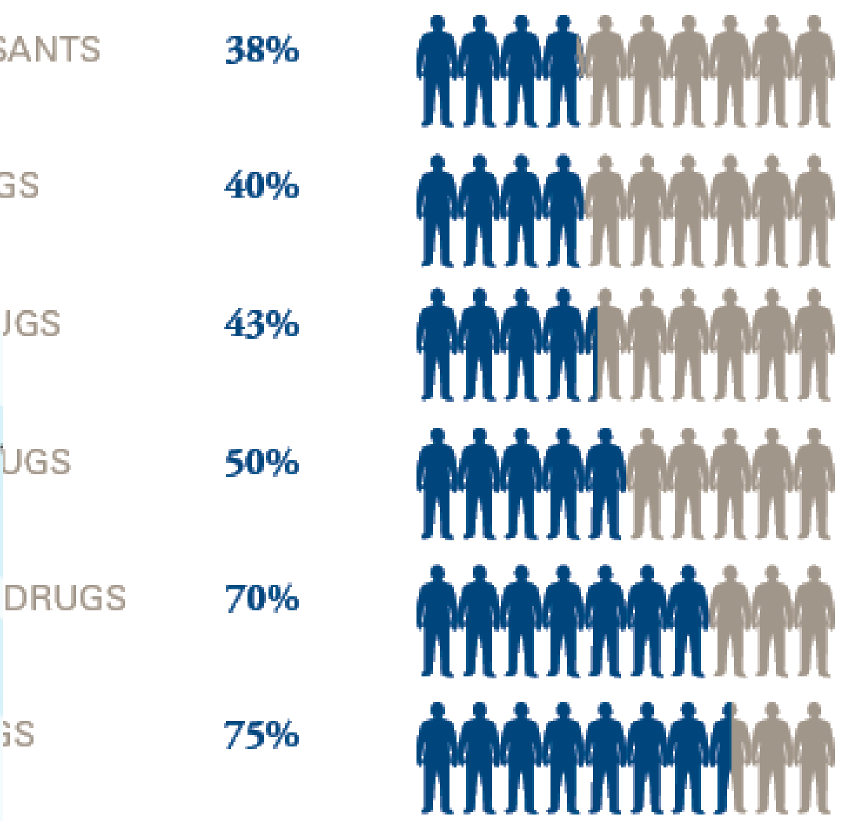
Species identified

EMBL-EBI provides the informatics and data needed to understand all aspects of an environment on the molecular level.

Personalised Medicine

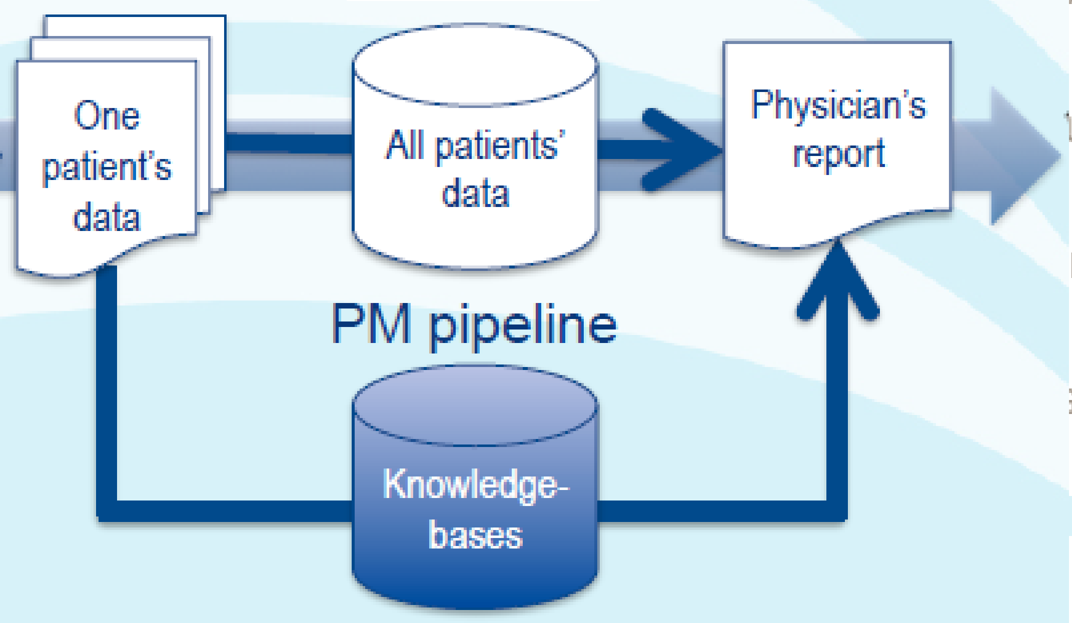
- Reduced sequencing costs enable new techniques
  - Past: 13 yrs & £2Bn
  - Now: 2 days & £1000

PERCENTAGE OF THE PATIENT POPULATION FOR WHICH A PARTICULAR DRUG IS INEFFECTIVE, ON AVERAGE



ANTI-DEPRESSANTS (SSRIs)	38%
ASTHMA DRUGS	40%
DIABETES DRUGS	43%
UGS	50%
DRUGS	70%
S	75%

Source: The Case for Personalized Medicine, 3rd edition



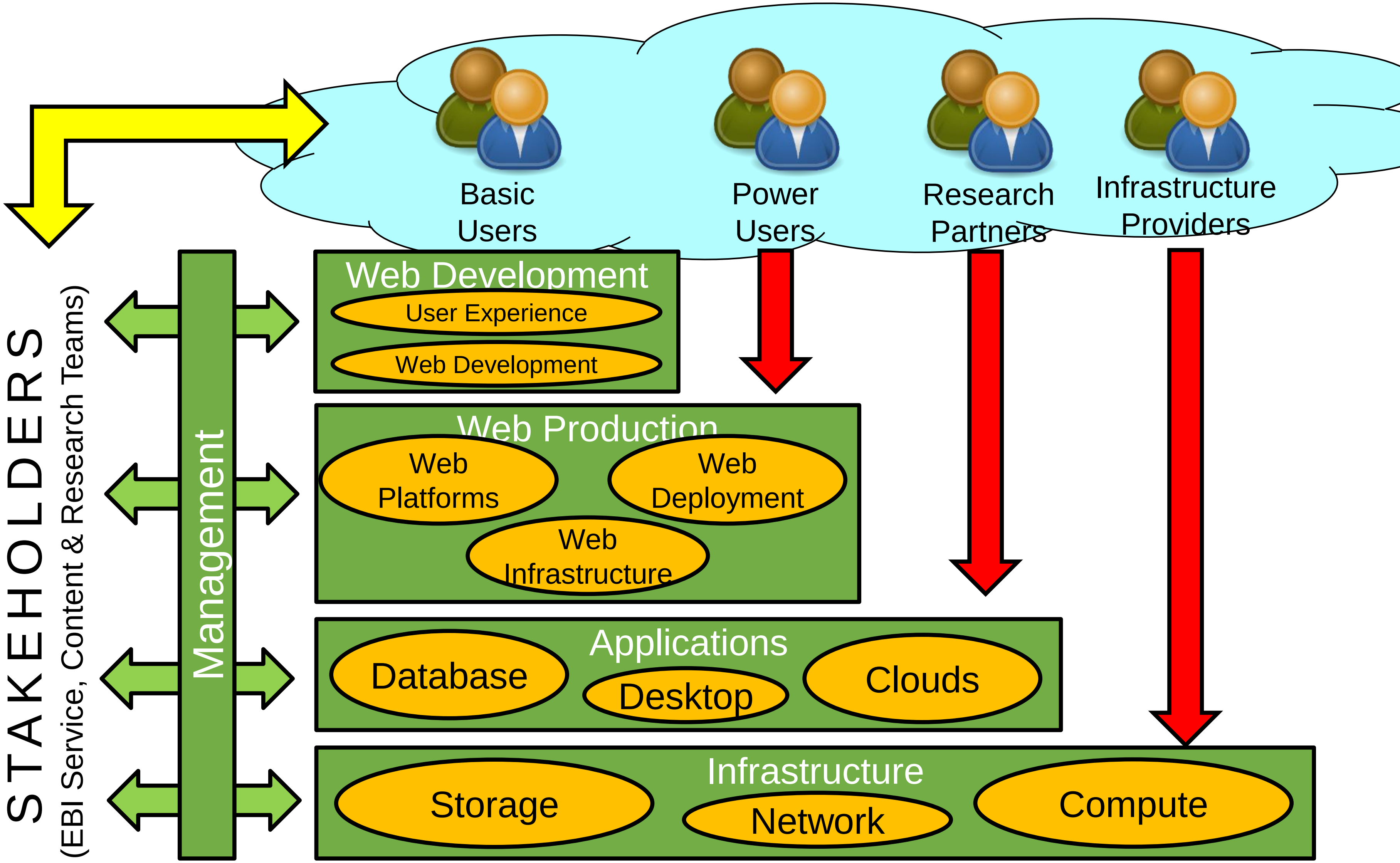
Requirements Summary

- Data Infrastructure
  - Distributed trans-national sources (→1000s+ sequencers)
  - Individually and collectively producing LOTS of data
  - Incredibly sensitive meta-data (i.e. your medical records)
- Data Consumption
  - Annotating data leads to information and knowledge
  - Improve usability and responsiveness to broad user base
    - Move from specialised researcher to other domain expert

Short Break

Pizza as a Service

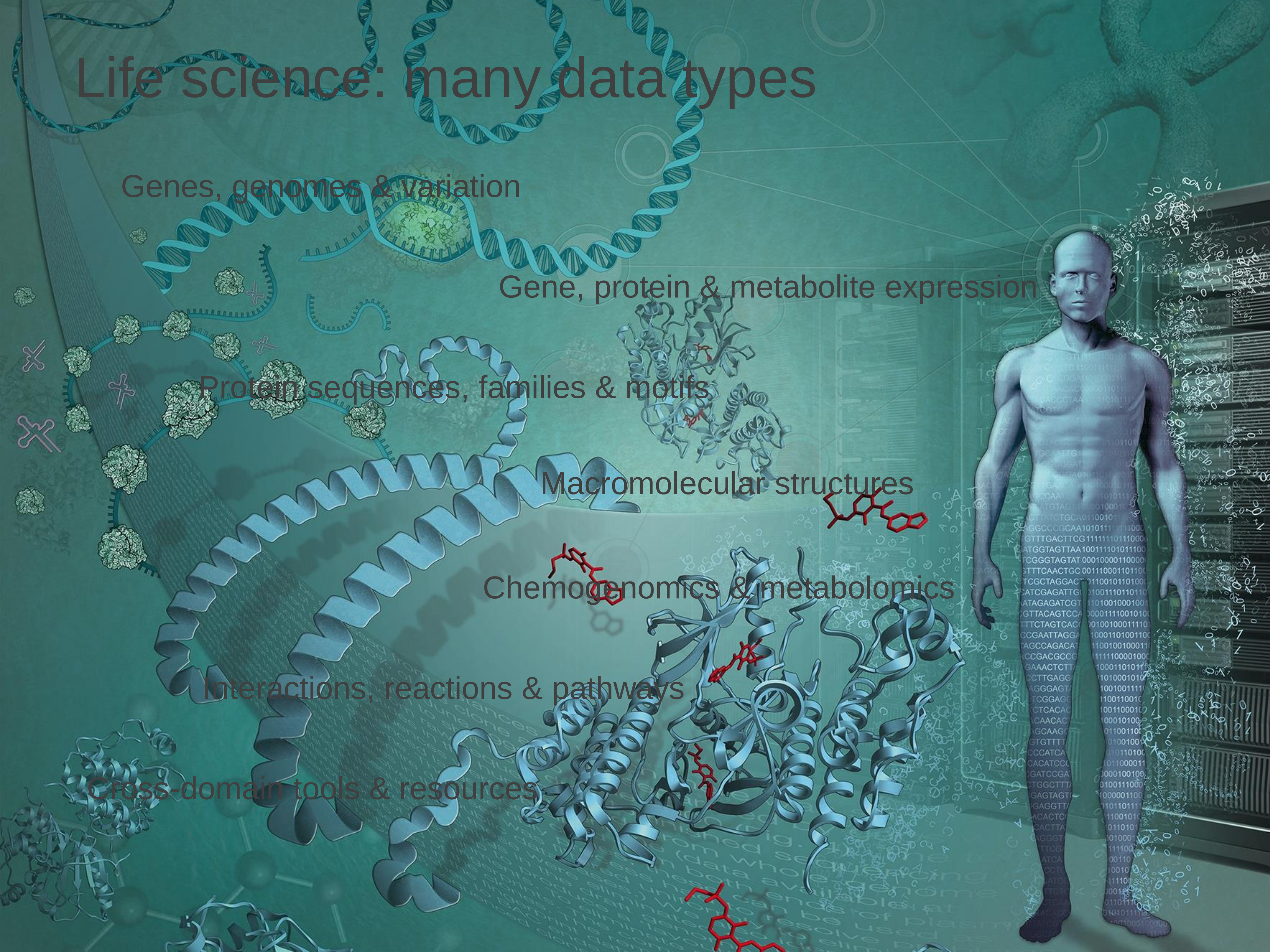
🟦 You Manage 🟩 Vendor Manages


Basic Users
Power Users
Research Partners
Infrastructure Providers
STAKEHOLDERS
(EBI Service, Content & Research Teams)
Management
Web Development
User Experience
Web Development
Web Production
Web Platforms
Web Deployment
Web Infrastructure
Applications
Database
Desktop
Clouds
Infrastructure
Storage
Network
Compute

Infrastructure Requirements

Role of the Technical Services Cluster

Life science: many data types



Genes, genomes & variation

Gene, protein & metabolite expression

Protein sequences, families & motifs

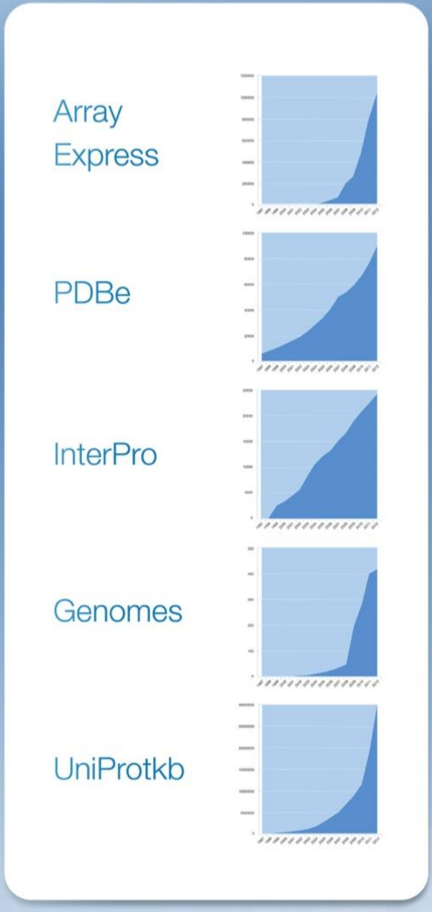
Macromolecular structures

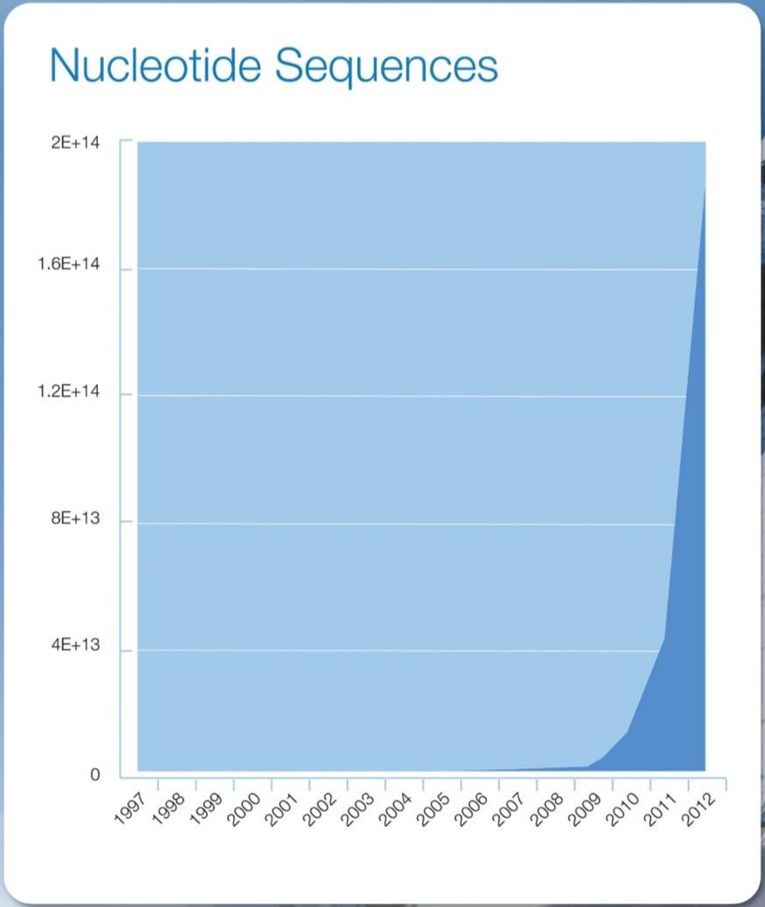
Chemogenomics & metabolomics

Interactions, reactions & pathways

Cross-domain tools & resources

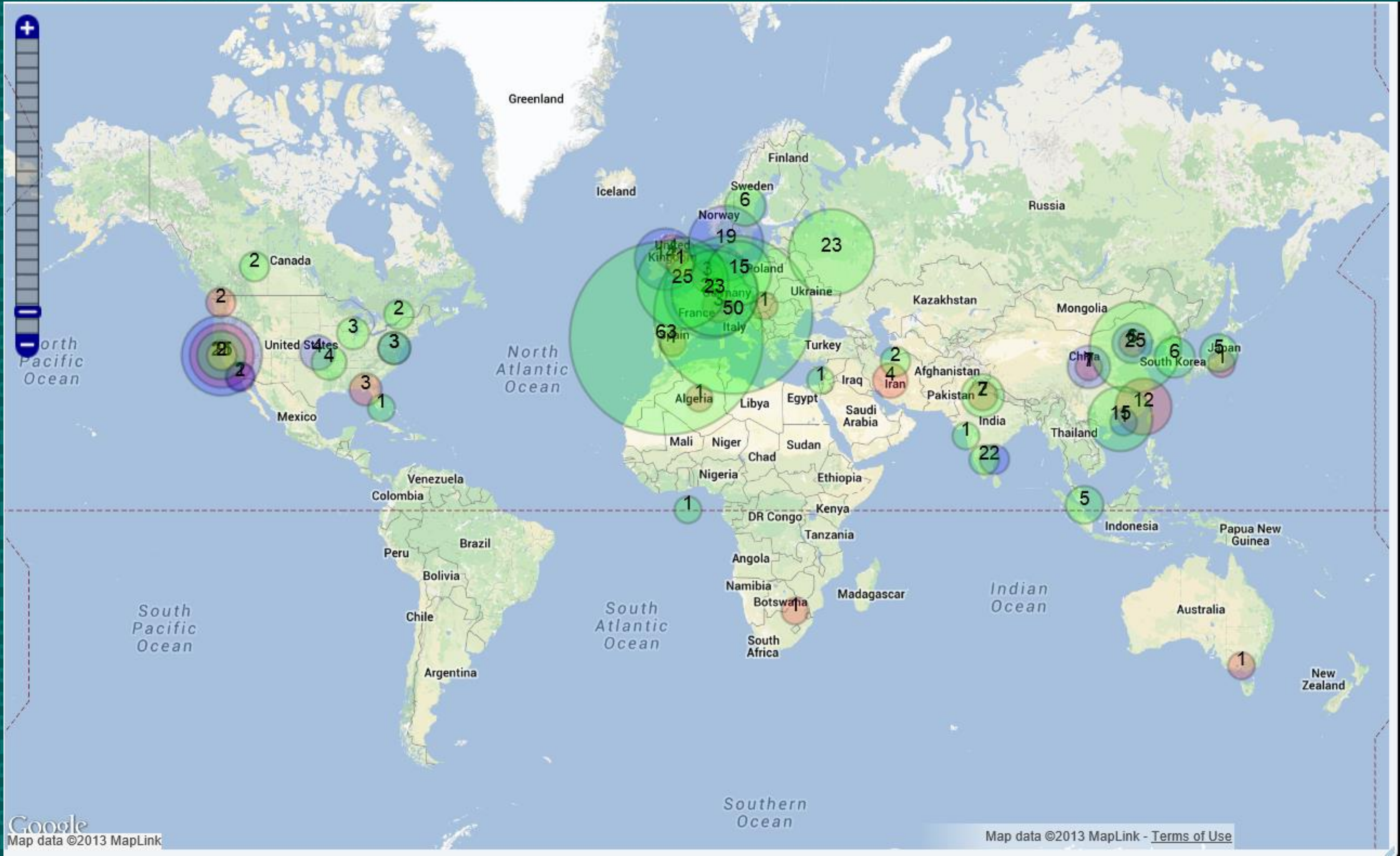
Bigger and bigger data

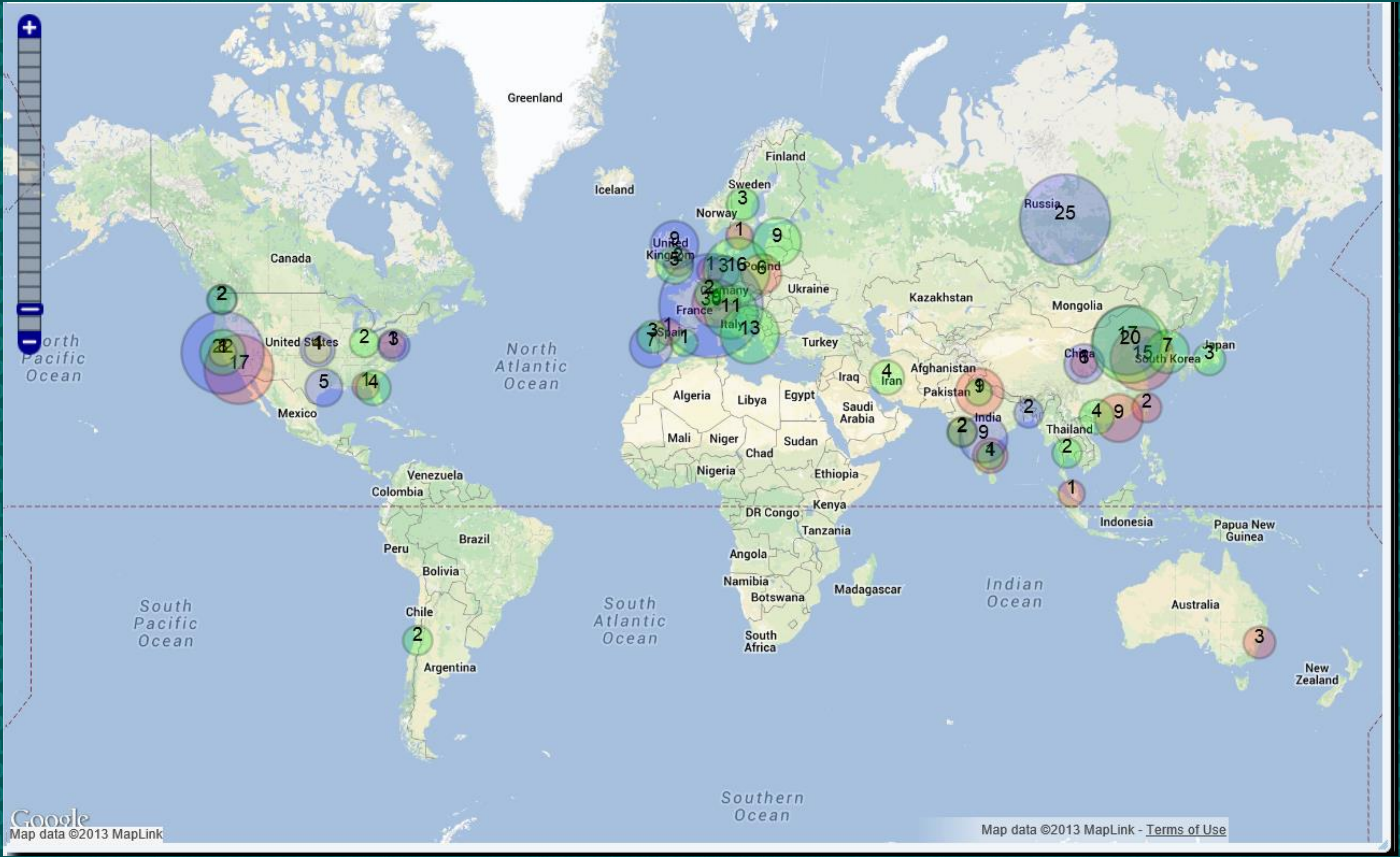


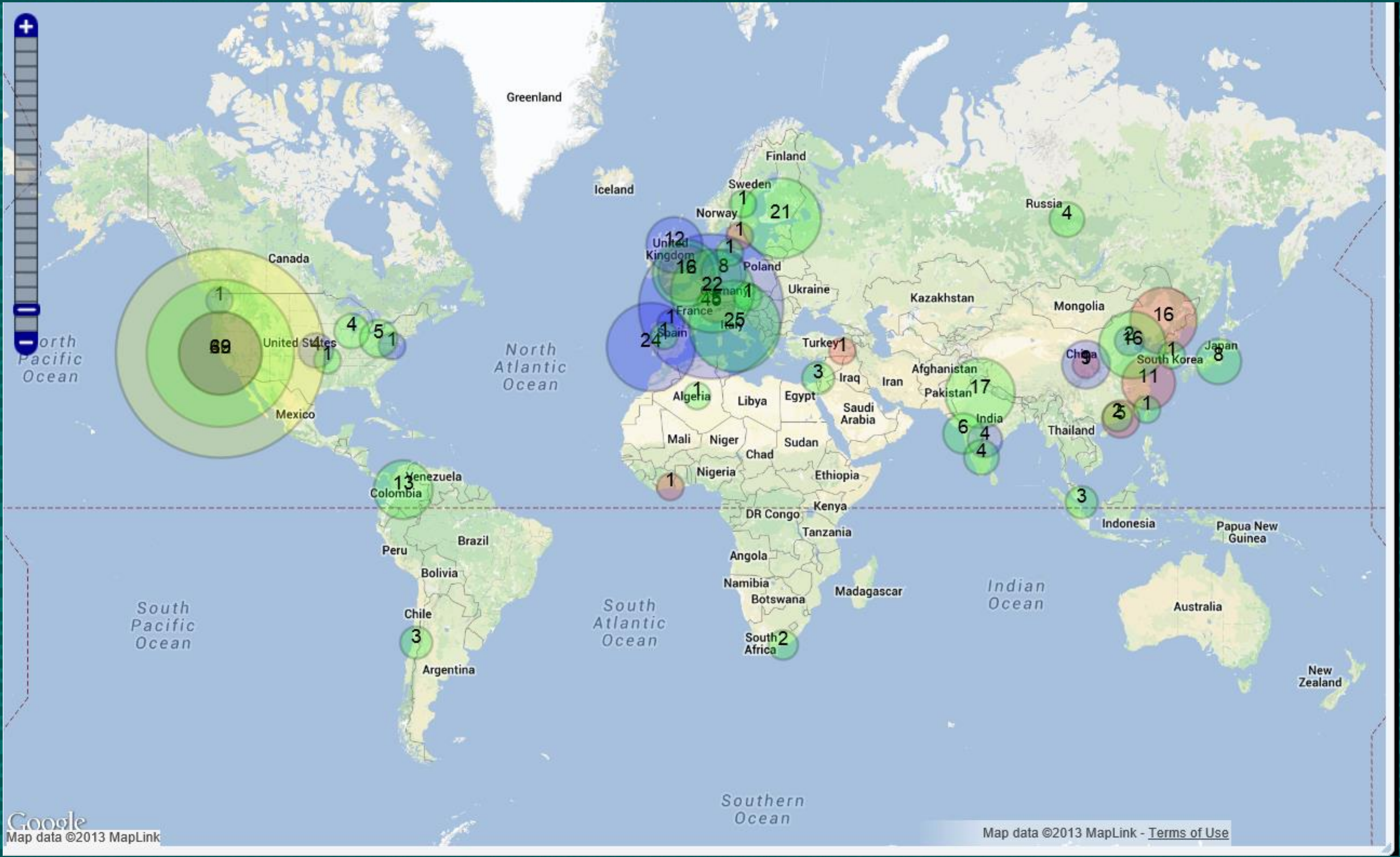


EBI Provides Services and Data Resources

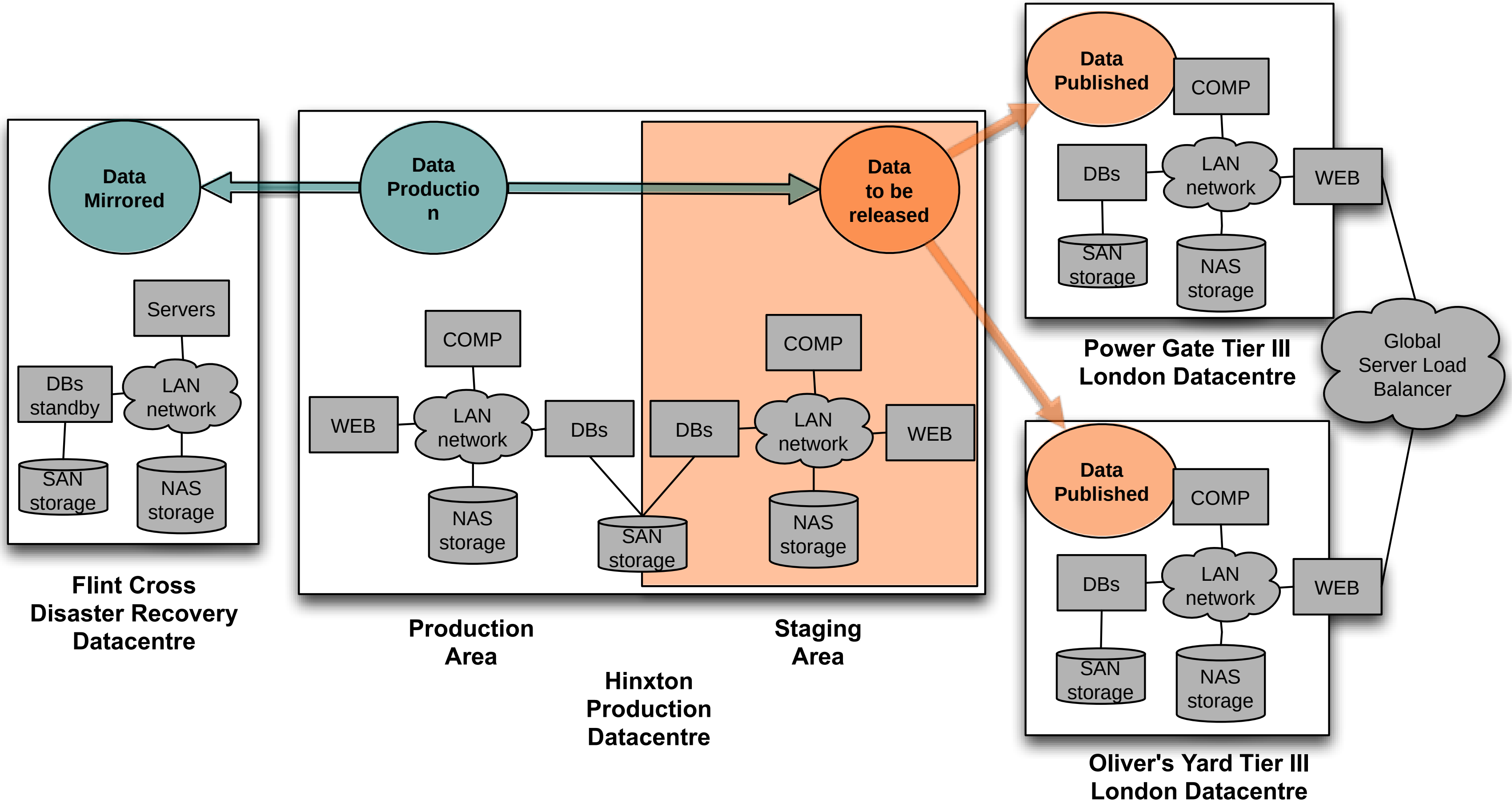
- Data Resources
  - Public and Managed Access
  - Individual sequence or bulk download
- Services
  - Web & programmatic access for common tools
  - Run ‘jobs’ on EMBL-EBI hardware
- Volume and variety of genomic data expanding
  - EMBL-EBI data doubling every year - replication is challenging
  - Infrastructure currently 50,000 CPUs & 46PB


Greenland
Iceland
Finland
Sweden
6
Norway
19
23
United Kingdom
1
25
15
Poland
23
Germany
Ukraine
France
50
1
Italy
63
Spain
1
Algeria
Libya
Egypt
Turkey
1
Iraq
2
4
Iran
Afghanistan
Pakistan
7
India
1
22
Saudi Arabia
Kazakhstan
Russia
Mongolia
China
7
25
6
South Korea
5
Japan
1
12
15
Thailand
5
Indonesia
Papua New Guinea
Australia
1
New Zealand
2
Canada
2
25
2
United States
4
4
3
2
3
3
1
North Pacific Ocean
Mexico
North Atlantic Ocean
Venezuela
Colombia
Brazil
Peru
Bolivia
Chile
Argentina
South Pacific Ocean
South Atlantic Ocean
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
1
DR Congo
Kenya
Tanzania
Angola
Namibia
Botswana
1
South Africa
Madagascar
Indian Ocean
Southern Ocean
Google
Map data ©2013 MapLink
Map data ©2013 MapLink - Terms of Use

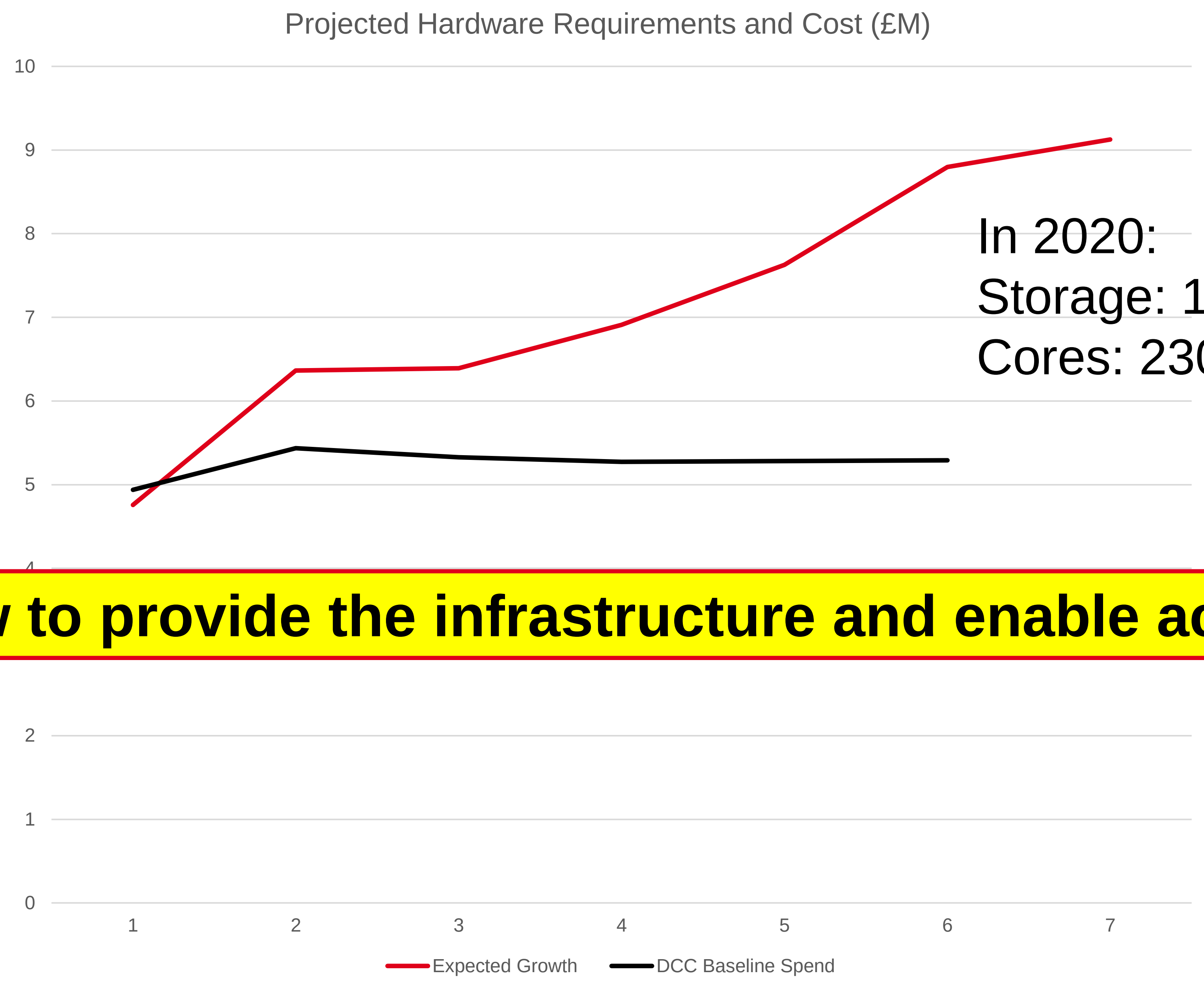

Greenland
Iceland
Finland
Sweden
3
Norway
1
9
United Kingdom
9
2
3
1
316
6
Poland
2
Germany
30
11
France
Italy
13
3
7
Spain
1
1
Ukraine
Kazakhstan
Russia
25
Mongolia
17
20
15
7
South Korea
Japan
3
China
8
Turkey
Iraq
4
Iran
Afghanistan
Pakistan
9
India
2
9
4
2
Thailand
4
9
2
2
1
Indonesia
Papua New Guinea
Canada
2
United States
32
17
4
2
3
5
14
Mexico
North Pacific Ocean
North Atlantic Ocean
Algeria
Libya
Egypt
Saudi Arabia
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Tanzania
Angola
Namibia
Botswana
South Africa
Madagascar
Venezuela
Colombia
Peru
Brazil
Bolivia
Chile
2
Argentina
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Australia
3
New Zealand
Southern Ocean
Google
Map data ©2013 MapLink
Map data ©2013 MapLink - Terms of Use


Greenland
Iceland
Finland
Sweden
Norway
Russia
United Kingdom
Poland
Ukraine
Germany
France
Spain
Italy
Kazakhstan
Mongolia
Canada
United States
Mexico
North Pacific Ocean
North Atlantic Ocean
Turkey
Iraq
Iran
Afghanistan
Pakistan
China
South Korea
Japan
India
Thailand
Algeria
Libya
Egypt
Saudi Arabia
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Tanzania
Venezuela
Colombia
Brazil
Peru
Bolivia
Chile
Argentina
Angola
Namibia
Botswana
Madagascar
South Africa
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Indonesia
Papua New Guinea
Australia
New Zealand
Southern Ocean
69
1
4
4
5
1
1
13
3
21
1
1
12
1
16
8
22
1
46
1
25
1
24
1
3
1
4
17
6
4
4
16
16
2
9
1
8
11
2
1
3
2
1
Google
Map data ©2013 MapLink
Map data ©2013 MapLink - Terms of Use

Overview EMBL-EBI IT infrastructure

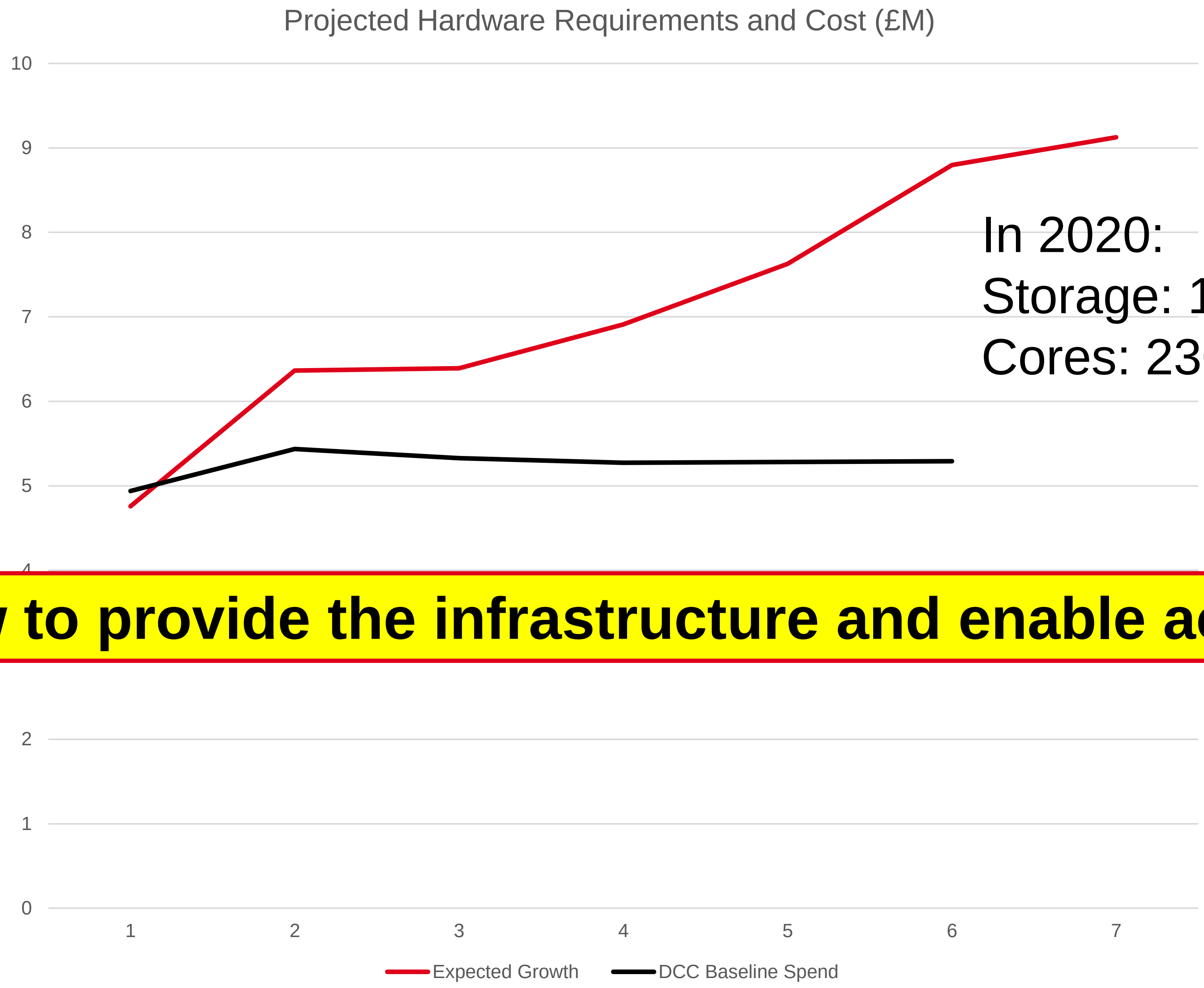


Data centre virtualised throughout with VMWare



In 2020:
Storage: 1,920PB
Cores: 230,000

How to provide the infrastructure and enable access?


Projected Hardware Requirements and Cost (£M)

In 2020:
Storage: 1,920PB
Cores: 230,000

How to provide the infrastructure and enable access?

EMBL-EBI

The Challenge Facing Services @ EMBL-EBI

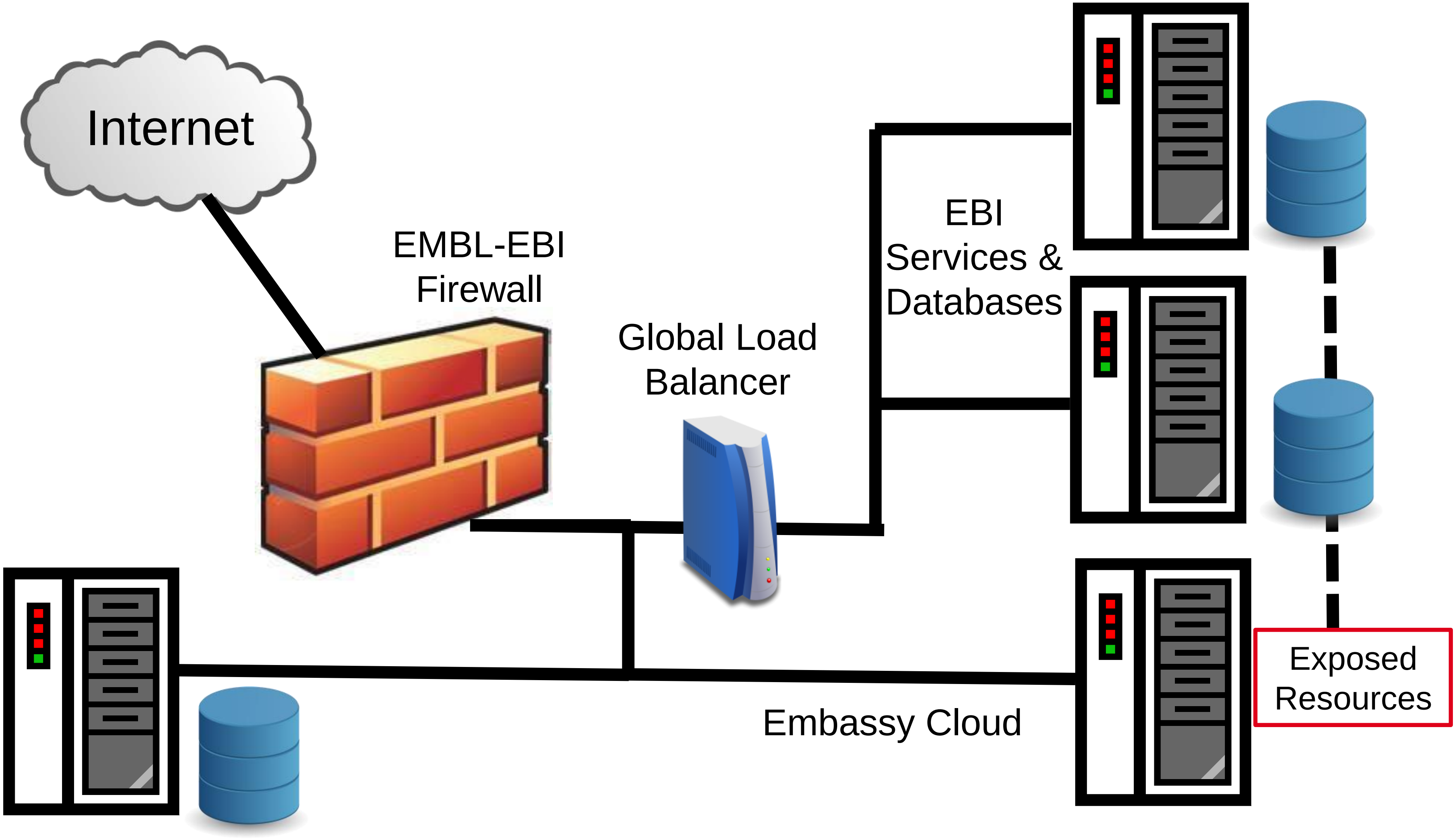
- Need to support complex analysis scenarios
  - Access to both public and managed access data sets
  - Bespoke workflows and tools across a variety of domains
  - Issues with disk to memory bandwidth
  - Web and programmatic access to services (3M unique users)
- Hard for users to replicate data sets for local analysis
  - Use the ‘cloud’ to bring local analysis to EMBL-EBI data

Embassy Cloud

- Our partners work directly beside EMBL-EBI data.
  - High bandwidth
  - Low latency
  - Robust, secure environment.
- Not in competition with commercial cloud services.
- Other cloud initiatives:
  - ELIXIR-facing cloud support
  - HELIX Nebula.

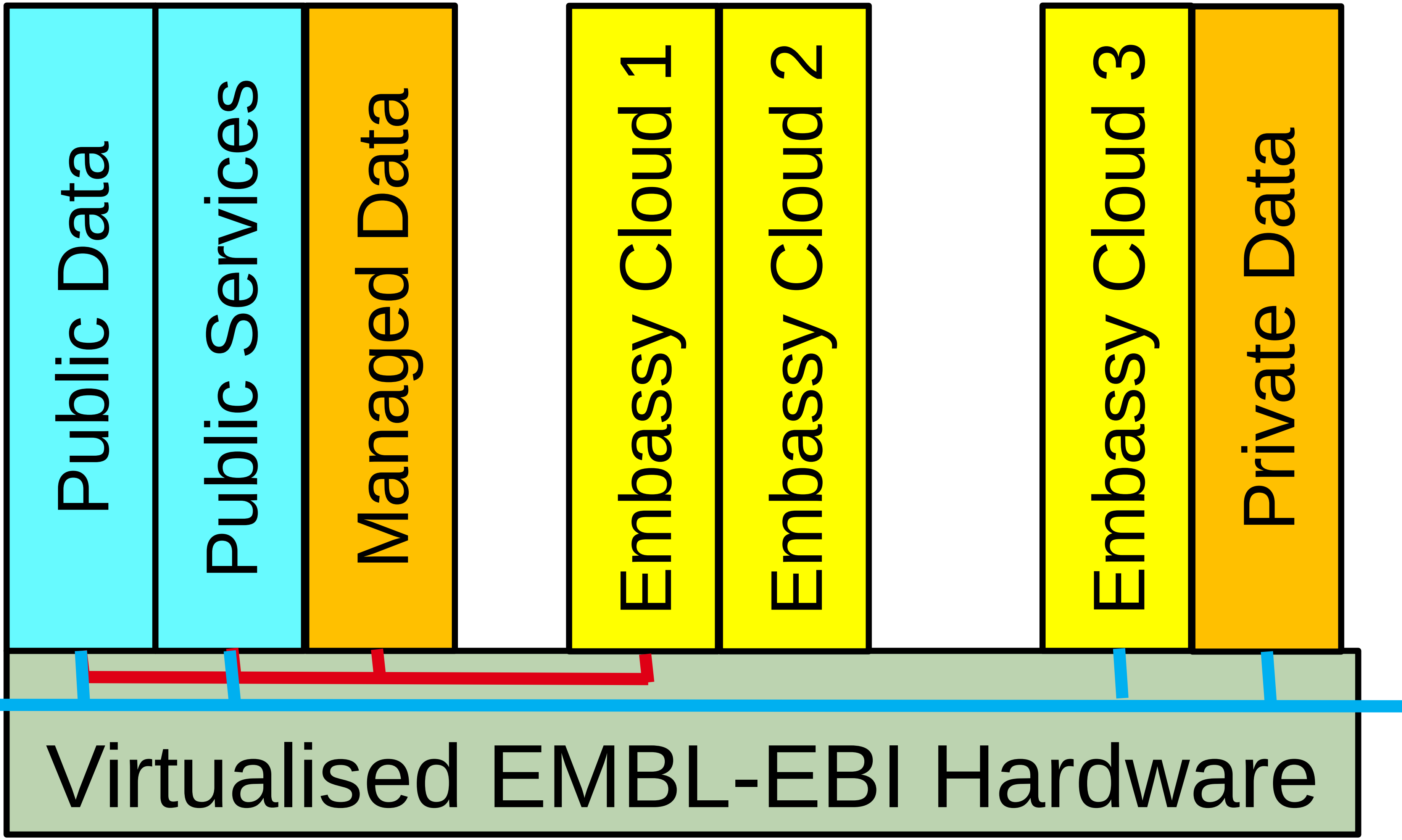
Genes & genomes · Expression · Proteins
Structures
Pathways
Literature · Ontology & taxonomy · Cross
EMBL-EBI
Embassy Cloud

EMBL-EBI Embassy Cloud



Embassy Cloud Concept

PanCancer



Public Data

Public Services

Managed Data

Embassy Cloud 1

Embassy Cloud 2

Embassy Cloud 3

Private Data

Virtualised EMBL-EBI Hardware

Typical Uses

- Web Application Hosting
  - Limited need for resources & VMs
  - CTTV: Host intranet, databases, …
- Data Staging
  - Undertake submission from local machine (following data staging) rather from remote location
  - BRAEMBL: Remote submission unreliable due to file upload
- Data Analysis
  - Large scale management and analysis of data
  - PanCancer: 1,000 cores, 2.5 TB RAM, 0.5 PB HDD

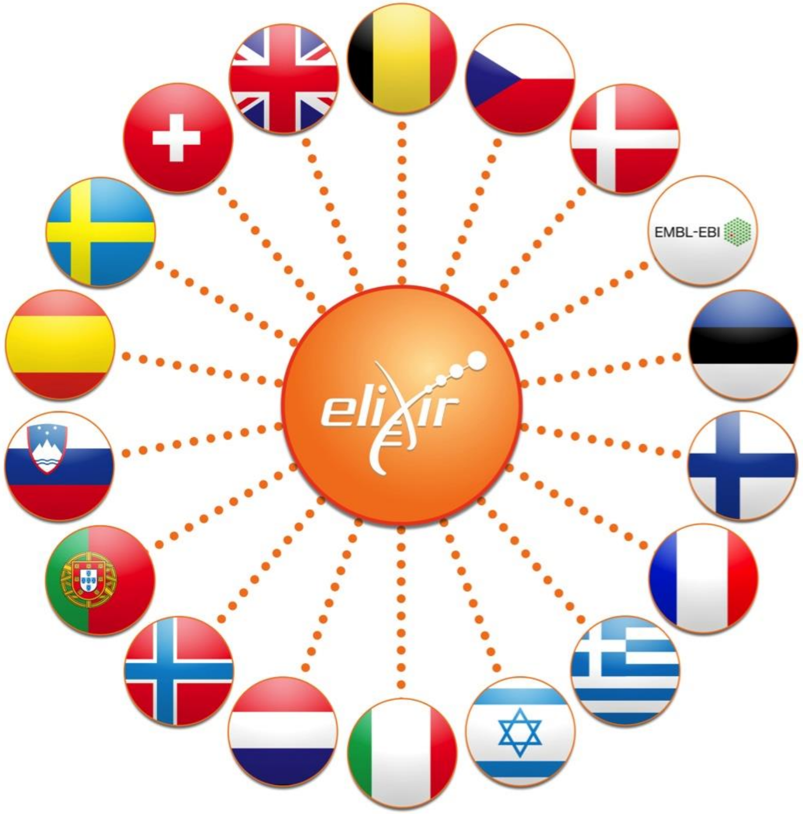
European Context

ELIXIR: a distributed data infrastructure

- EMBL-EBI is a major driver in ELIXIR, the pan-European research infrastructure for biological information.
- Central Hub at EMBL-EBI, with Nodes at centres of excellence throughout Europe.
- The goal of ELIXIR:
  - Build a sustainable European infrastructure for biological information
  - Support life science research and its translation to medicine, the environment, the bioindustries and society.

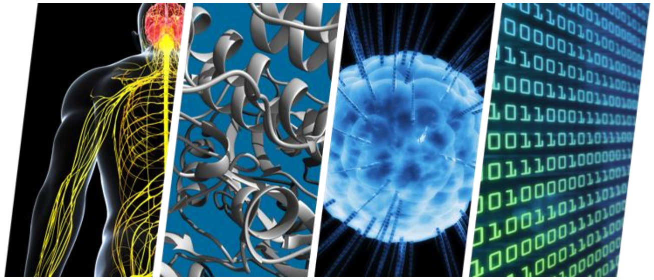
Building capacity in Europe

- ELIXIR: a sustainable infrastructure for biological information in Europe.
- Supporting life science research and its translation t(
  - medicine
  - agriculture
  - the environment
  - the bioindustries
  - society.



The Future

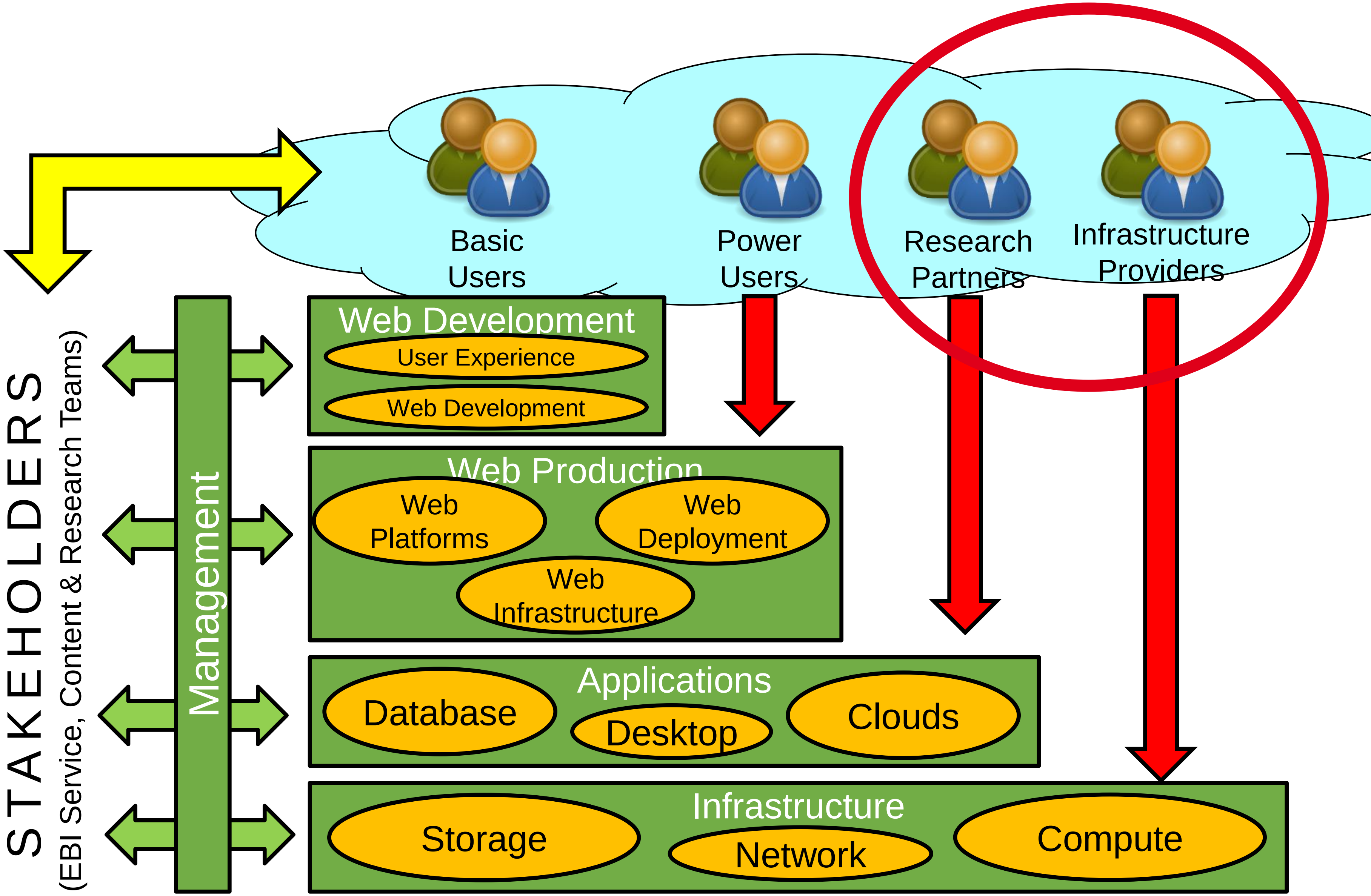
Centre for Therapeutic Target Validation



- Collaboration to pinpoint the processes in the human body that have a demonstrable effect on disease.
- Public-private initiative:
  - GSK: expertise in disease biology
  - EMBL-EBI: expertise in life science data integration and analysis
  - Wellcome Trust Sanger Institute: expertise in the role of genetics in disease.

CTTV

Centre for Therapeutic Target Validation



STAKEHOLDERS

(EBI Service, Content & Research Teams)

Management

Basic Users

Power Users

Research Partners

Infrastructure Providers

Web Development

User Experience

Web Development

Web Production

Web Platforms

Web Deployment

Web Infrastructure

Applications

Database

Desktop

Clouds

Infrastructure

Storage

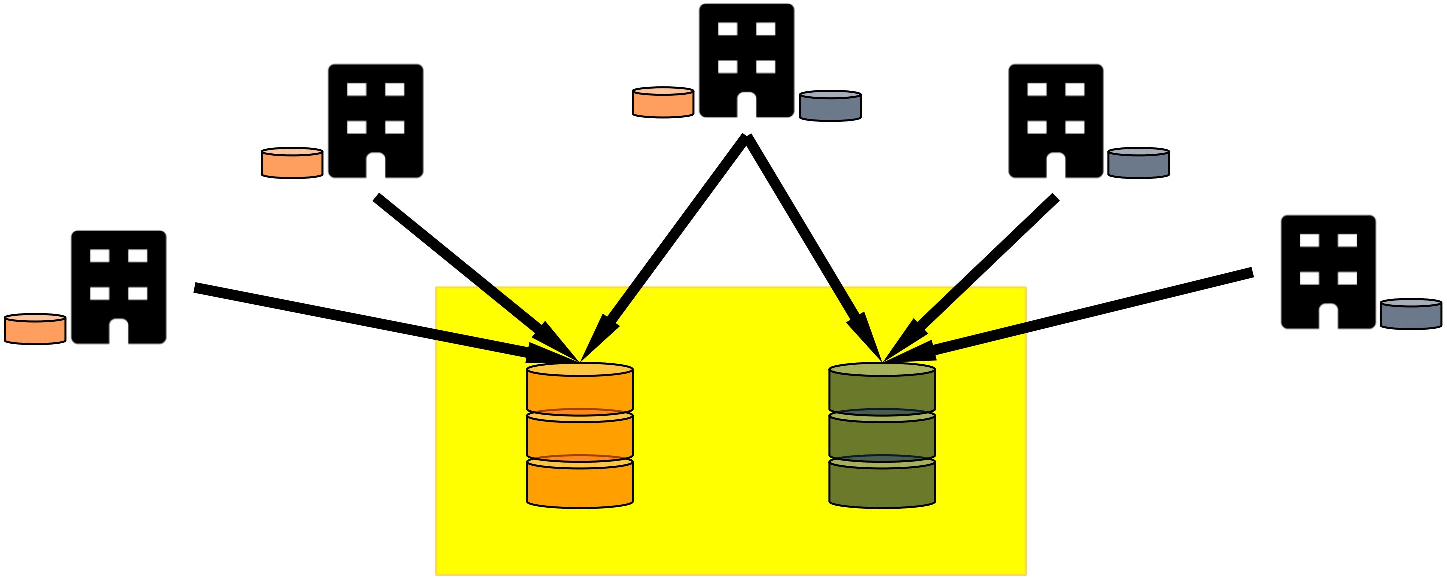
Network

Compute

Technical Services Cluster: IT as a Service

- Developing a Service Portfolio
  - Internal and External (Public & Private) Users
- Putting the Service Portfolio into a governance process
  - To manage and communicate change of the portfolio
- Contributing to the Elixir Research Infrastructure

Centralization & specialization



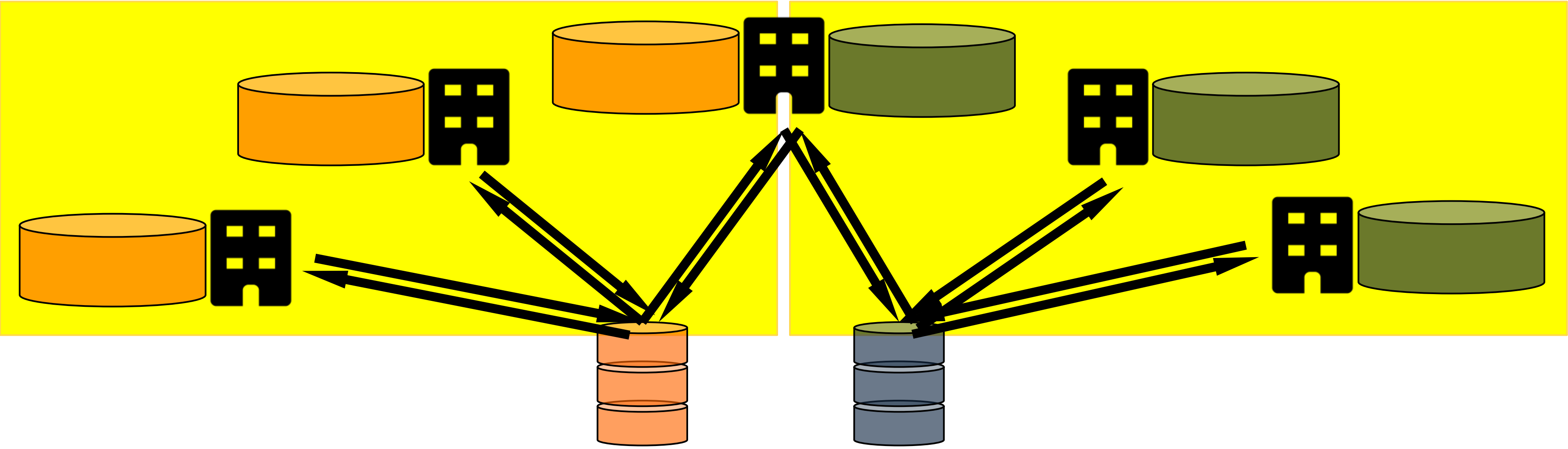
- Data is submitted to specialized centralized repositories.
- Current situation.

Data production

Data centralization

Data

Federation



- If data gets bigger, the **data might have to stay** where it is produced.
- We might have to provision data producers with **storage and computation**.
- Data might be pulled instead of pushed into centralized repositories.

Data production

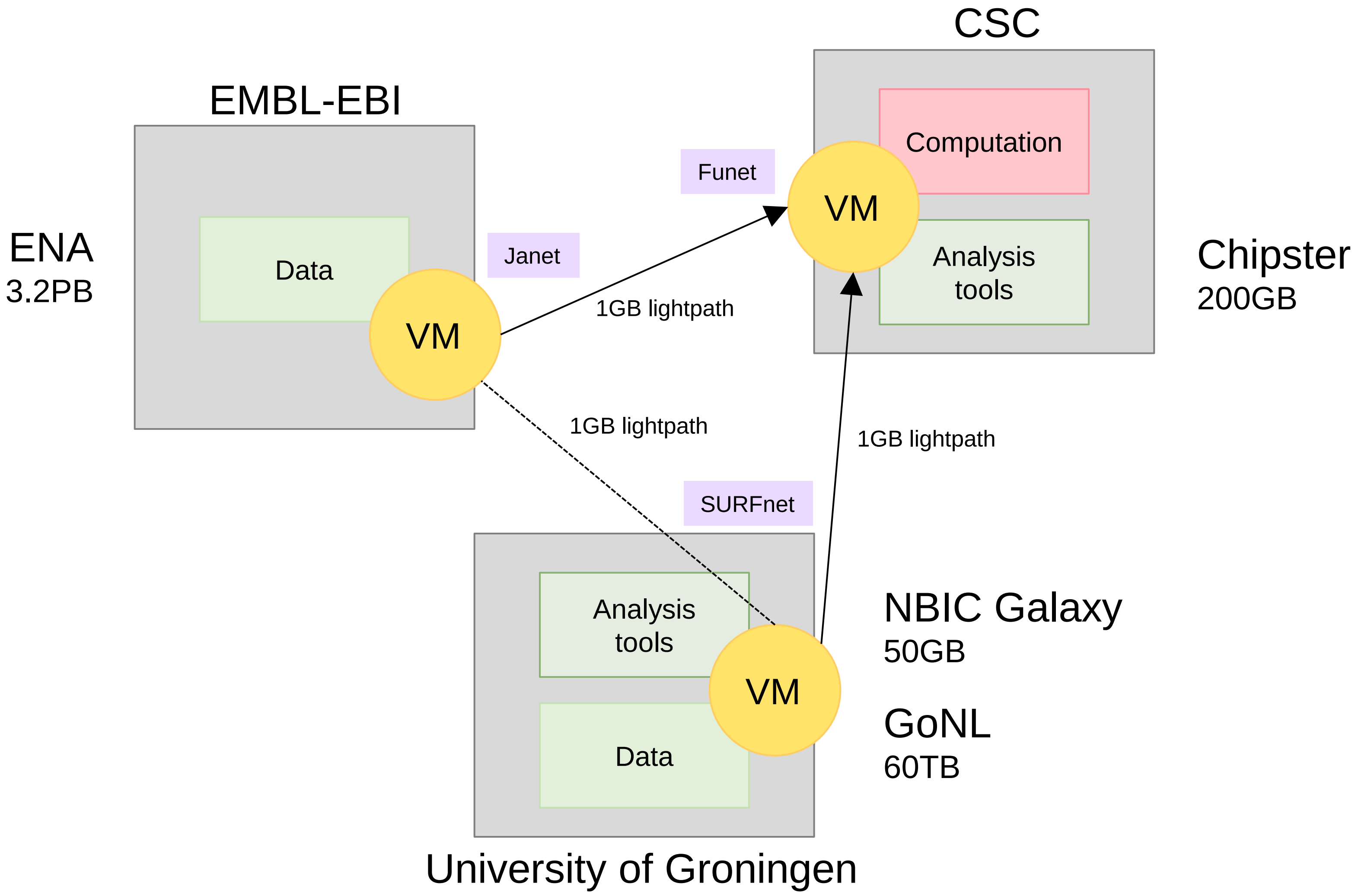
Data centralization

Data

Enlighten Your Research (GEANT)

- Explore cross-site VM operation using light-paths
  - Sites in NL, UK & FI
  - Provision networks on demand
- Use Case: Analysis needs significant resources and data
  - Moving beyond the scope of local clusters
- Goal: Distribute analysis and data over multiple clouds
- Activity since November 2013:
  - Liaising with sites and NRENs for bandwidth on demand
  - CSC & EMBL-EBI using existing light-path and different data movement protocols (e.g. GridFTP, Aspera)

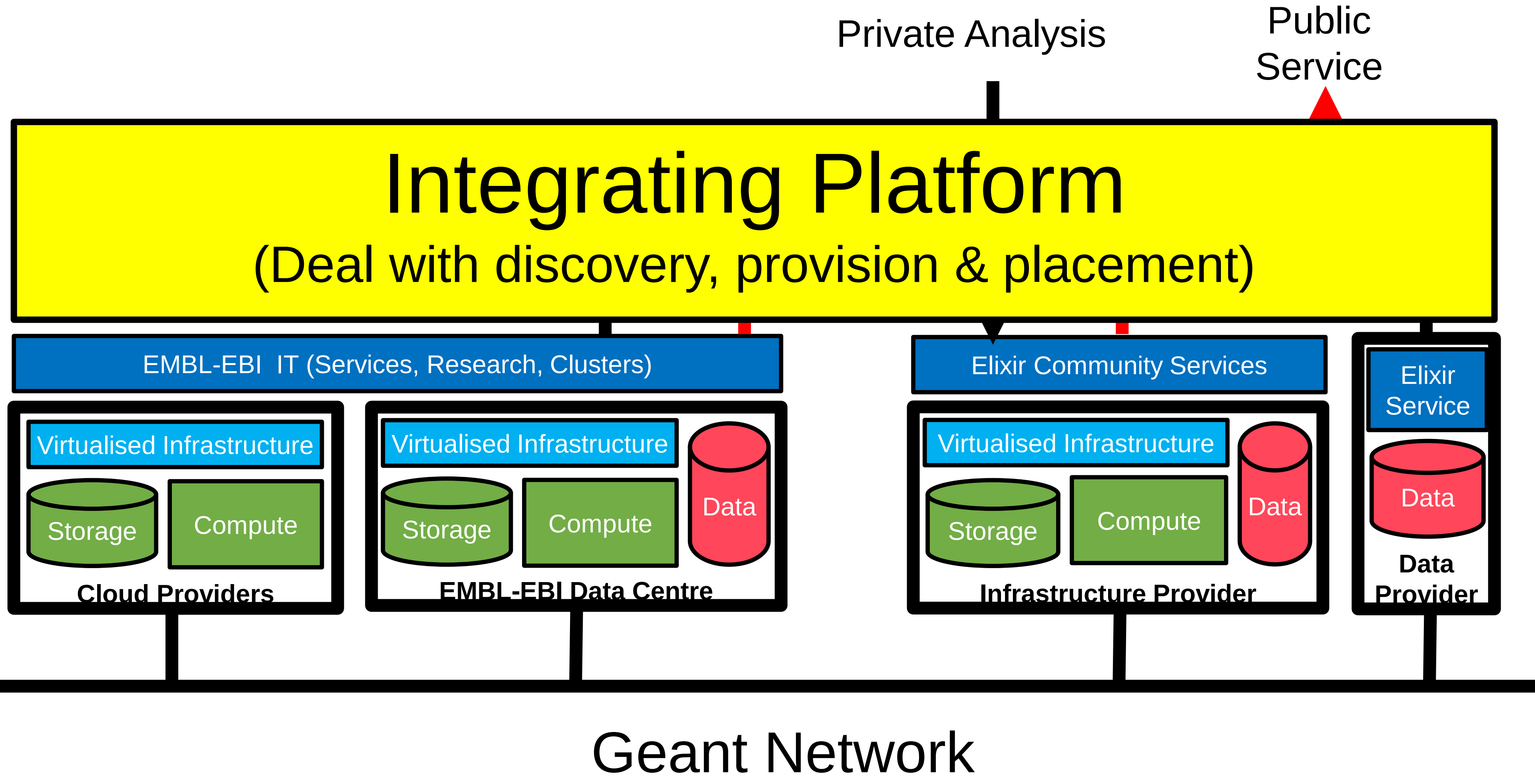
Cross-site VM Operation



Other Cloud Activity at EMBL-EBI

- Use Amazon to provide geographical distribution
  - Direct link to globally replicate databases
- HelixNebula
  - Integration of commercial cloud providers with big research
- Benefit of additional security assurances
  - For use by pharmaceutical companies
  - For on-demand personalised medicine
- Explore using IaaS to supplement/replace data centres
  - Put DC on cloud, scale out services (service + database), etc.

The Future



Private Analysis
Public Service
Integrating Platform
(Deal with discovery, provision & placement)
EMBL-EBI IT (Services, Research, Clusters)
Elixir Community Services
Elixir Service
Virtualised Infrastructure
Storage
Compute
Cloud Providers
Virtualised Infrastructure
Storage
Compute
Data
EMBL-EBI Data Centre
Virtualised Infrastructure
Storage
Compute
Data
Infrastructure Provider
Data
Data Provider
Geant Network

Thank you

Steven.Newhouse@ebi.ac.uk