

# 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

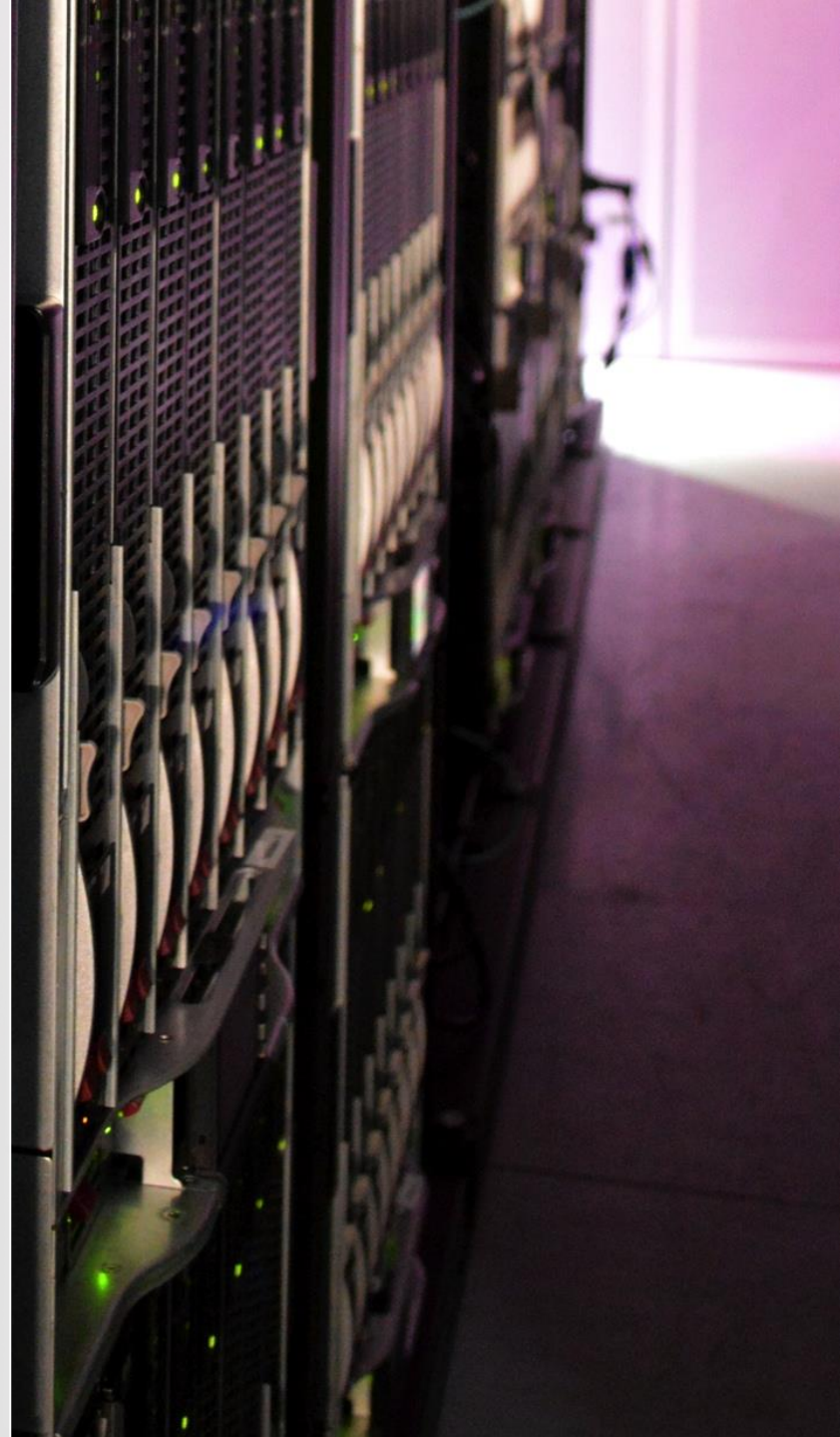
Associate member states:  
Argentina, Australia





# 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
- 570 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

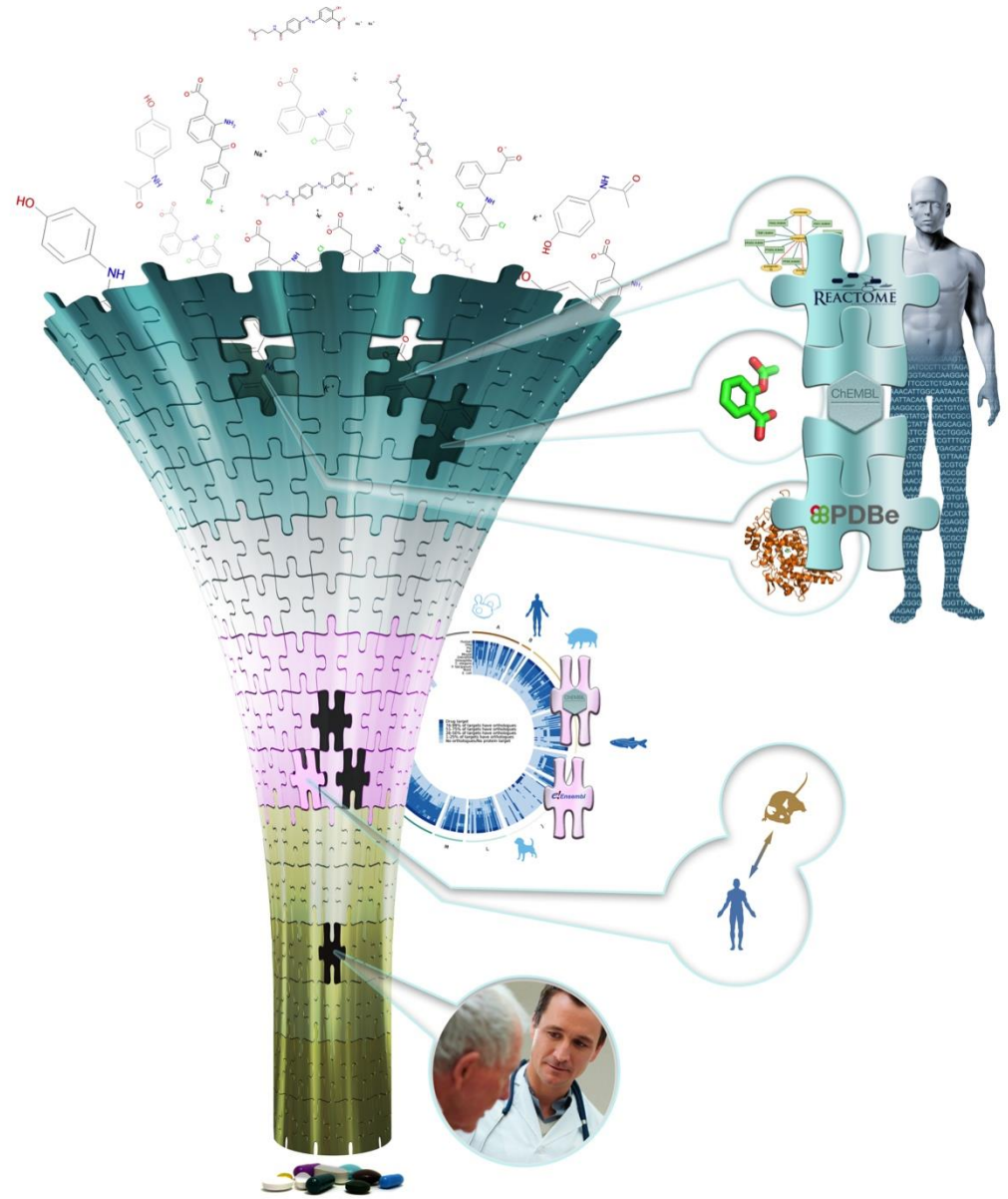




# Challenges & Opportunities

# 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.

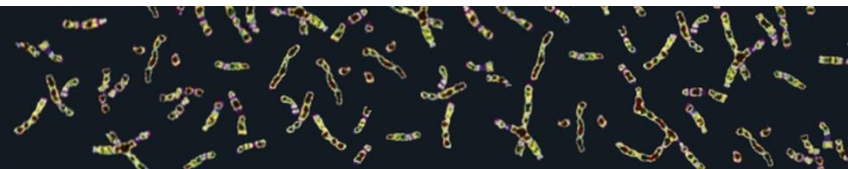


# Interpreting human variation

- 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.

**1000 Genomes**

A Deep Catalog of Human Genetic Variation



# 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

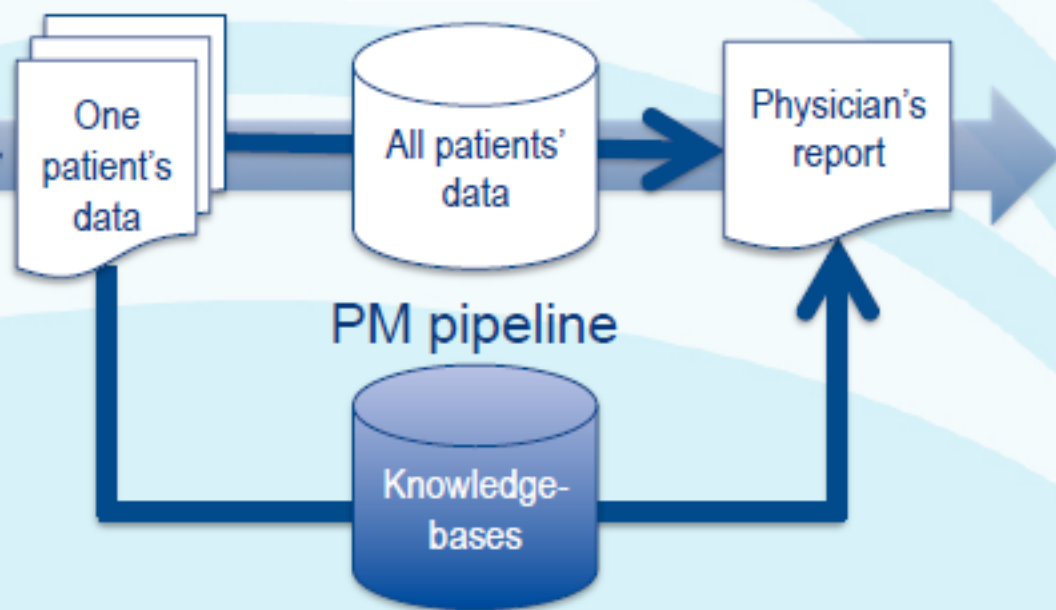


© Time, Inc.

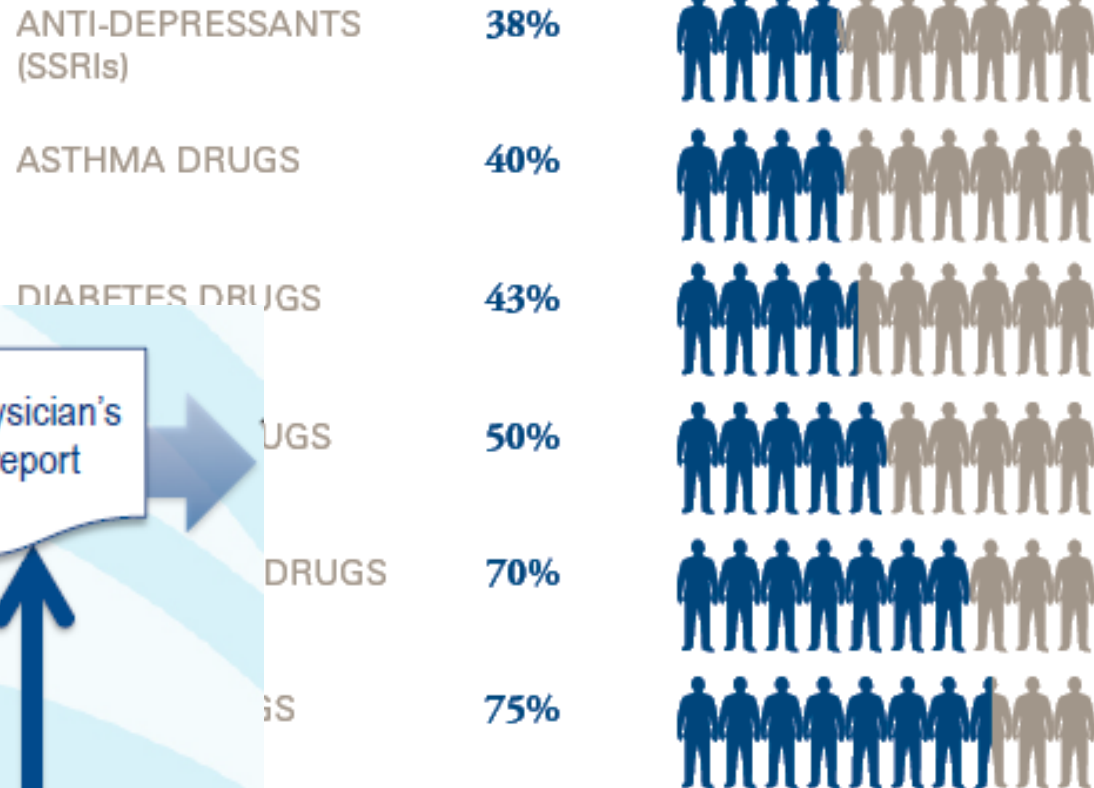


# Personalised Medicine

- 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



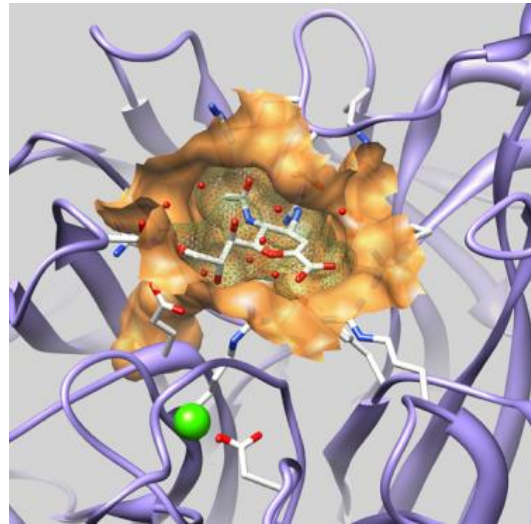
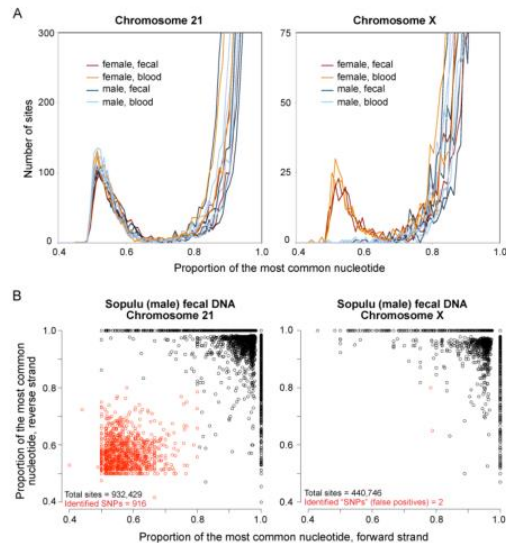
Source: The Case for Personalized Medicine, 3<sup>rd</sup> edition

# From Data to Knowledge in the Life Sciences

Molecular data

Information

Knowledge



Mutation X  
disrupts  
enzyme  
function, which  
causes Y  
disease

# Answering these Challenges

- 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
- Data Exploitation
  - Enable access beyond specialised researcher
  - Improve usability across broad user base

# Short Break



# Pizza as a Service

You Manage    Vendor Manages

# Infrastructure Requirements

Role of the Technical Services Cluster

# Life science: many data types

Genes, genomes & variation

Gene, protein & metabolite expression

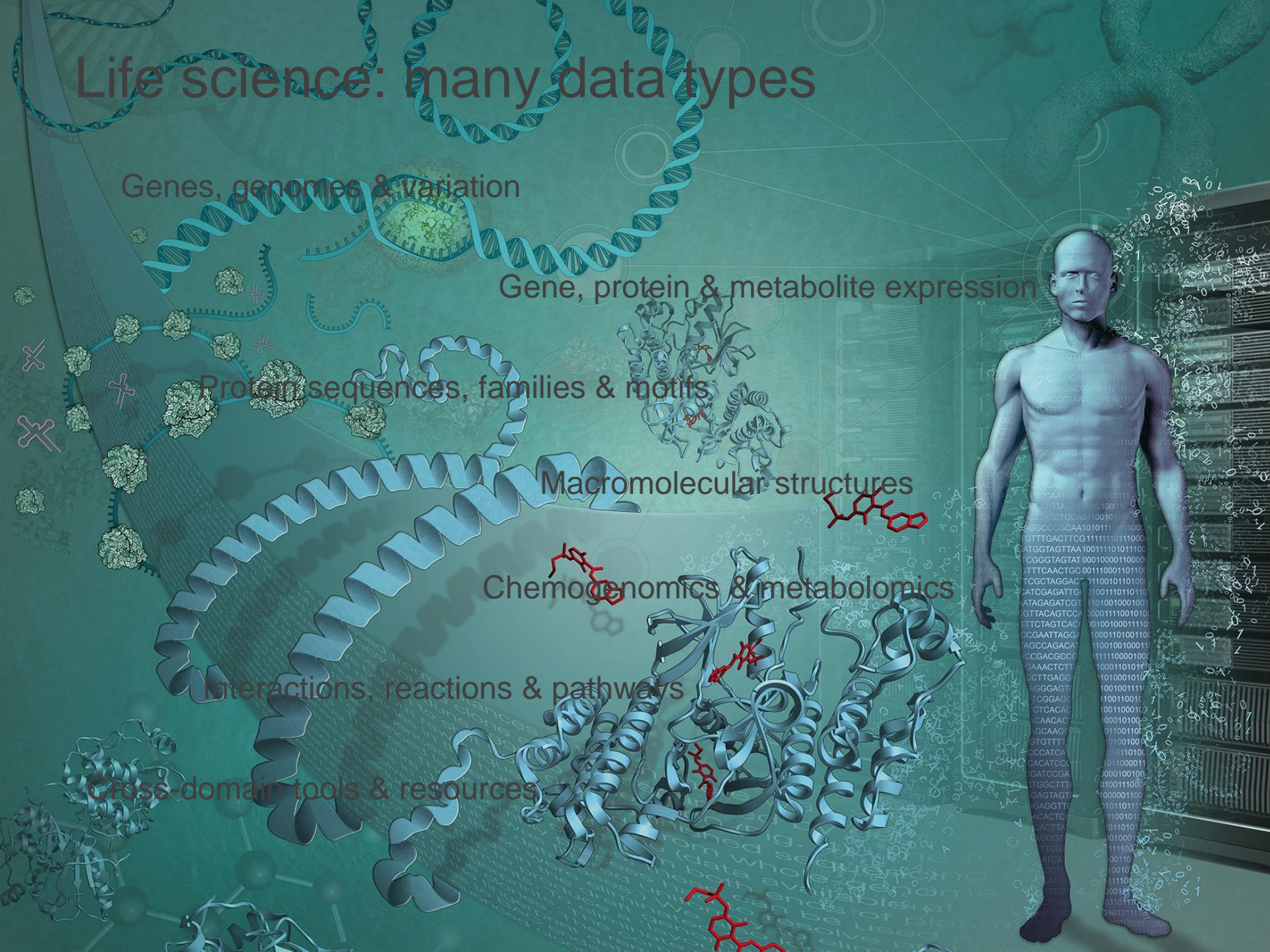
Protein sequences, families & motifs

Macromolecular structures

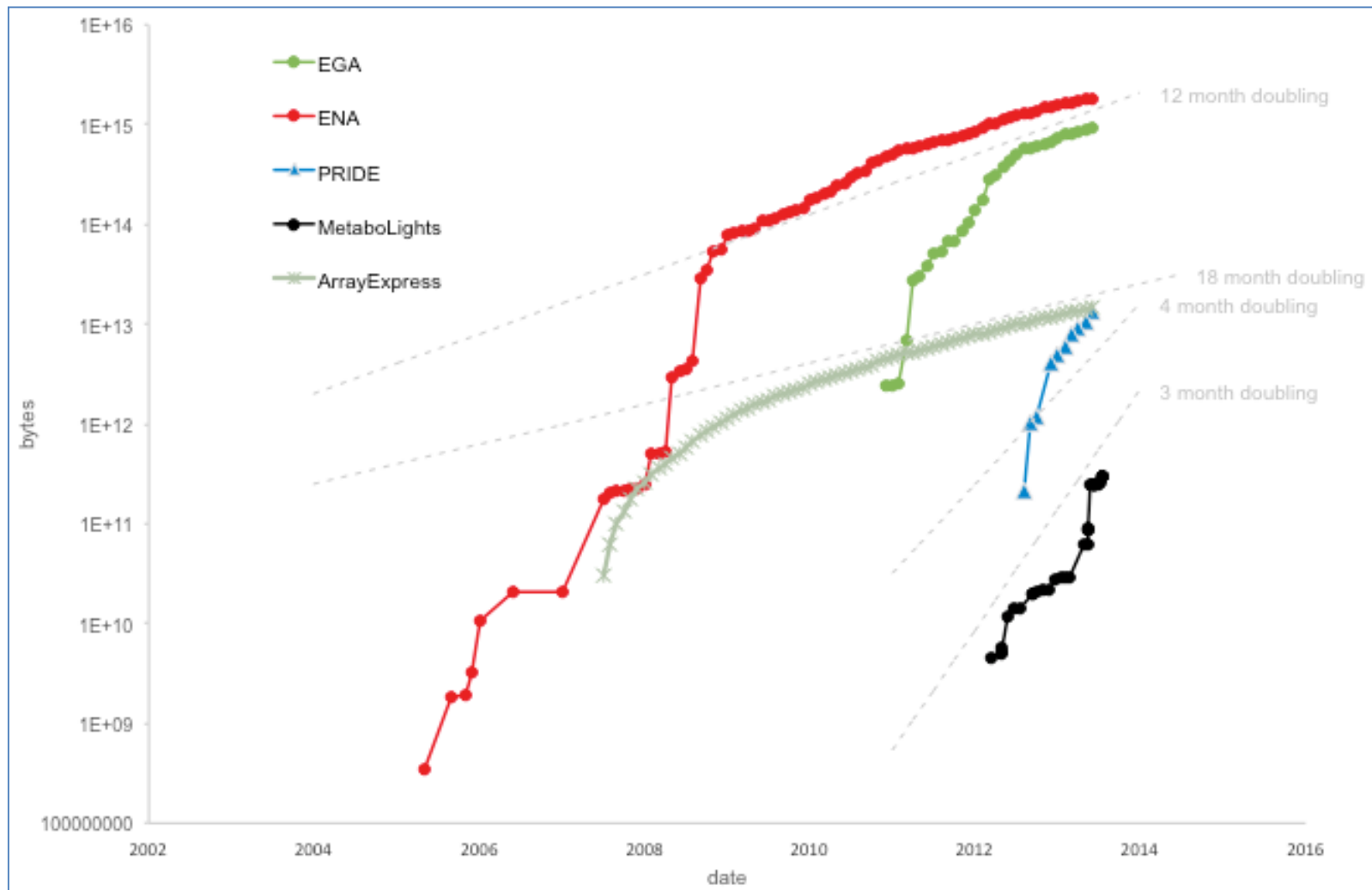
Chemogenomics & metabolomics

Interactions, reactions & pathways

Cross-domain tools & resources

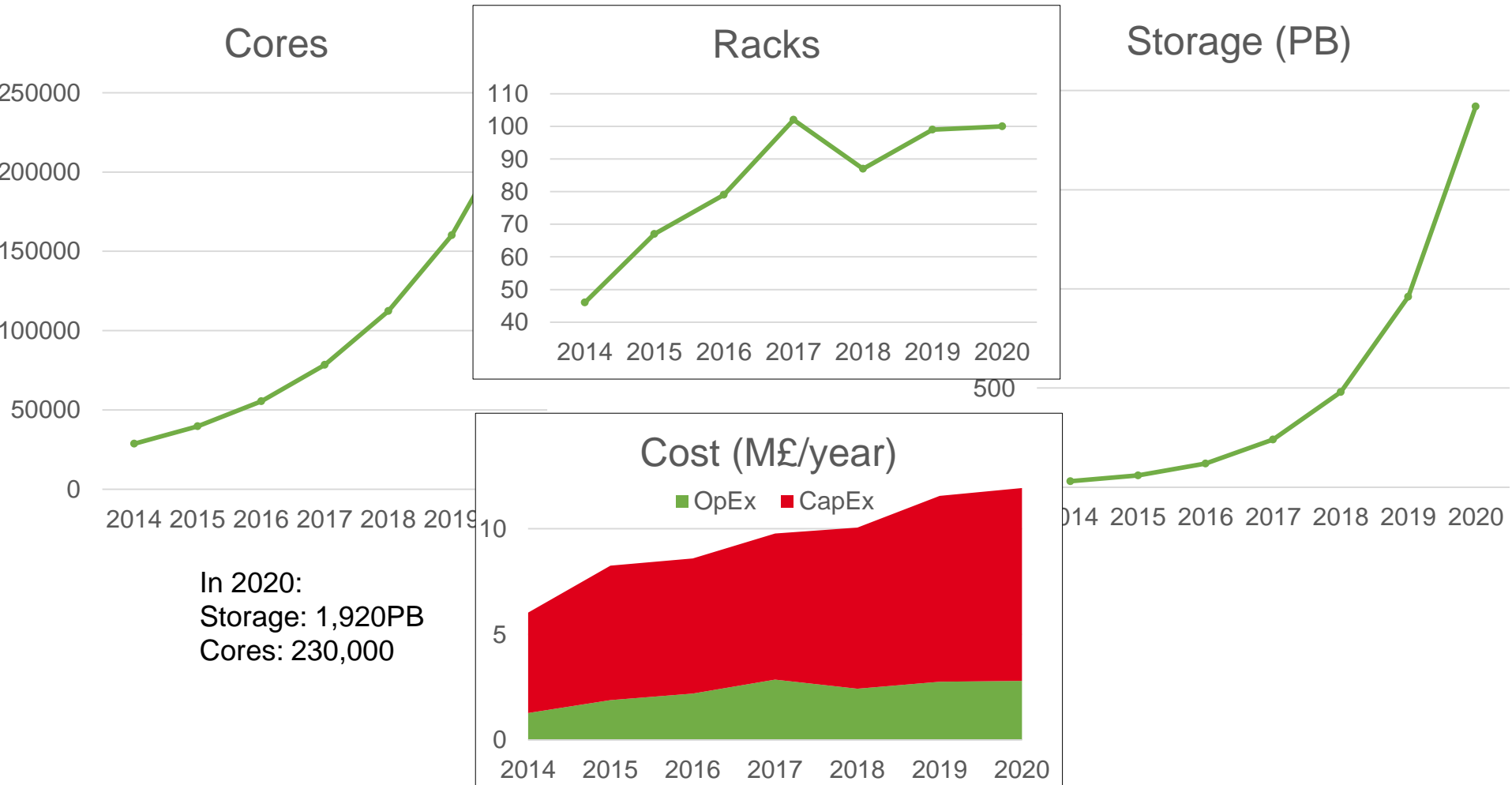


# Data Growth: A Community Challenge

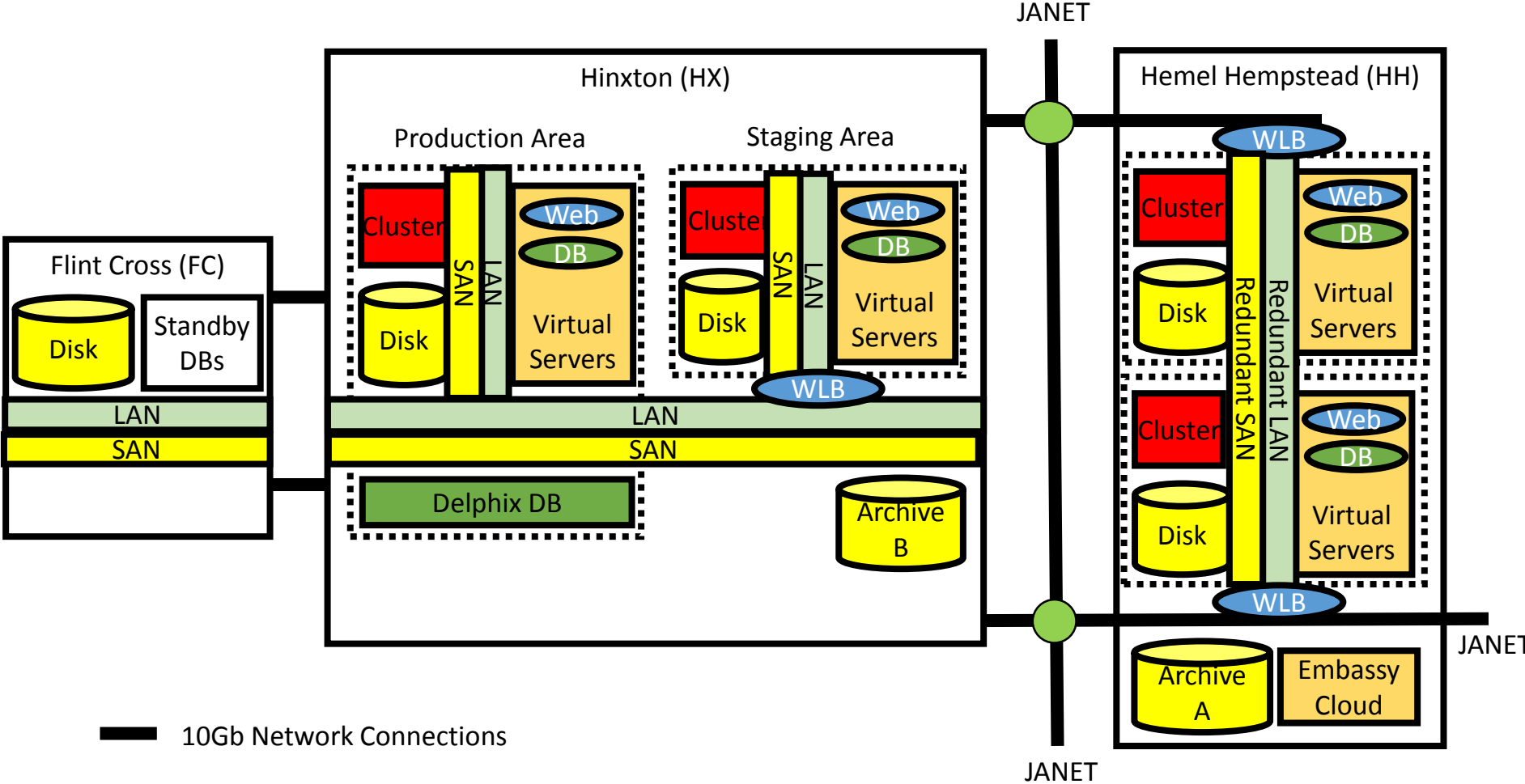




# Challenges of Big Data Infrastructure 2015+



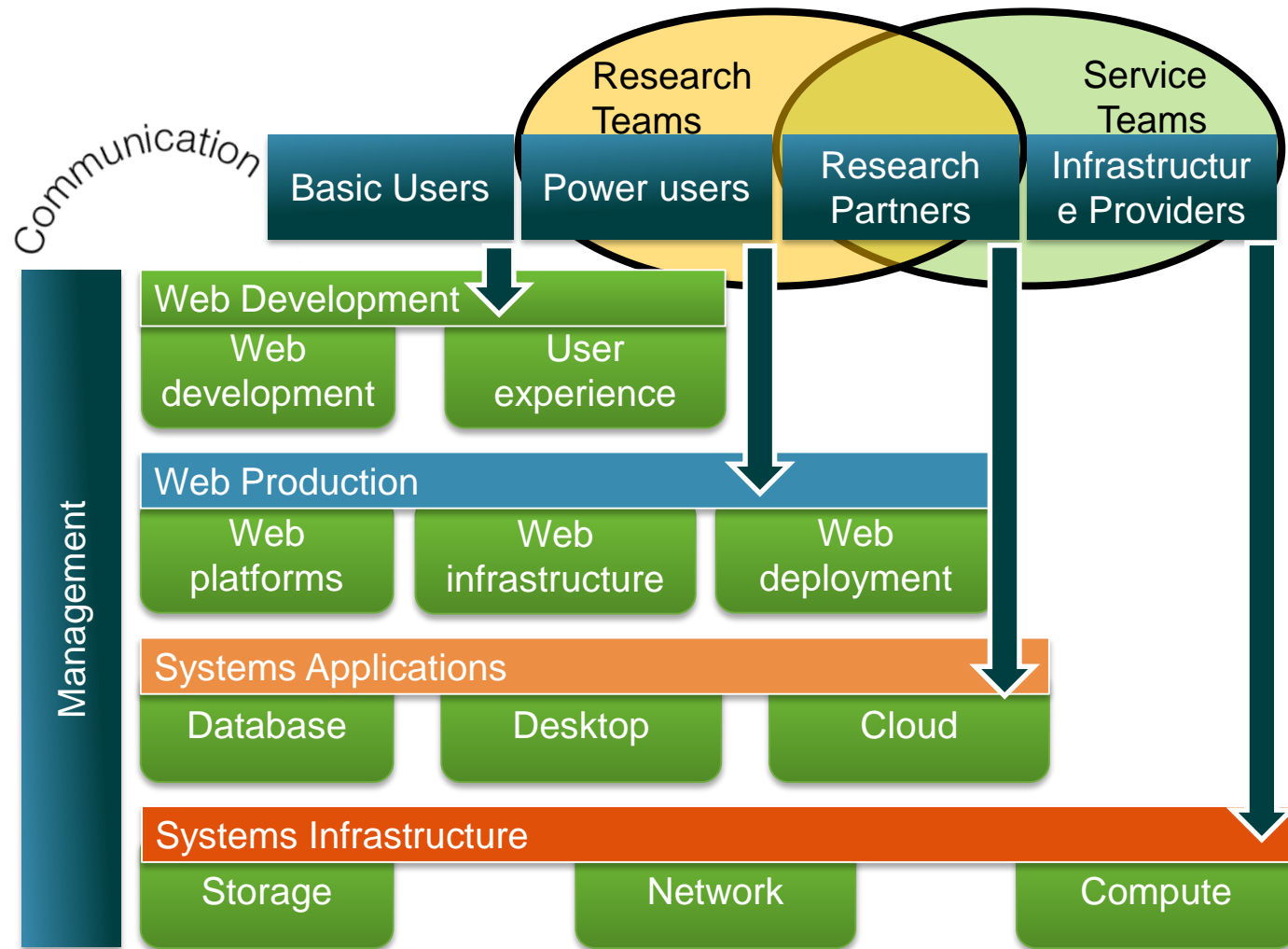
EMBL-EBI Data Centre Infrastructure



# 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 & 55+PB

# Technical services



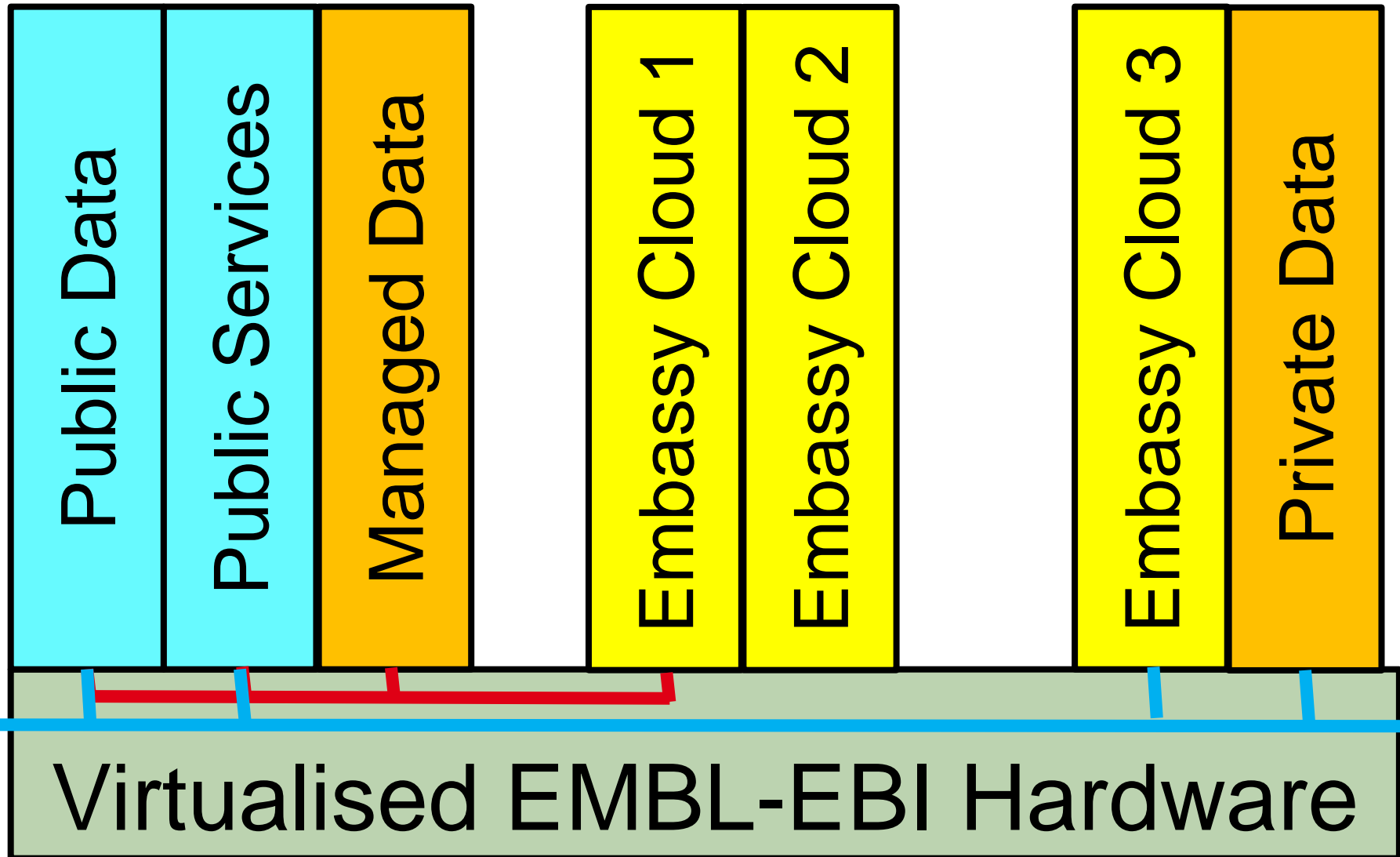


# The Challenge Facing Services @ EMBL-EBI

- 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 Concept

PanCancer



# 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.



# 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, 1.0 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.

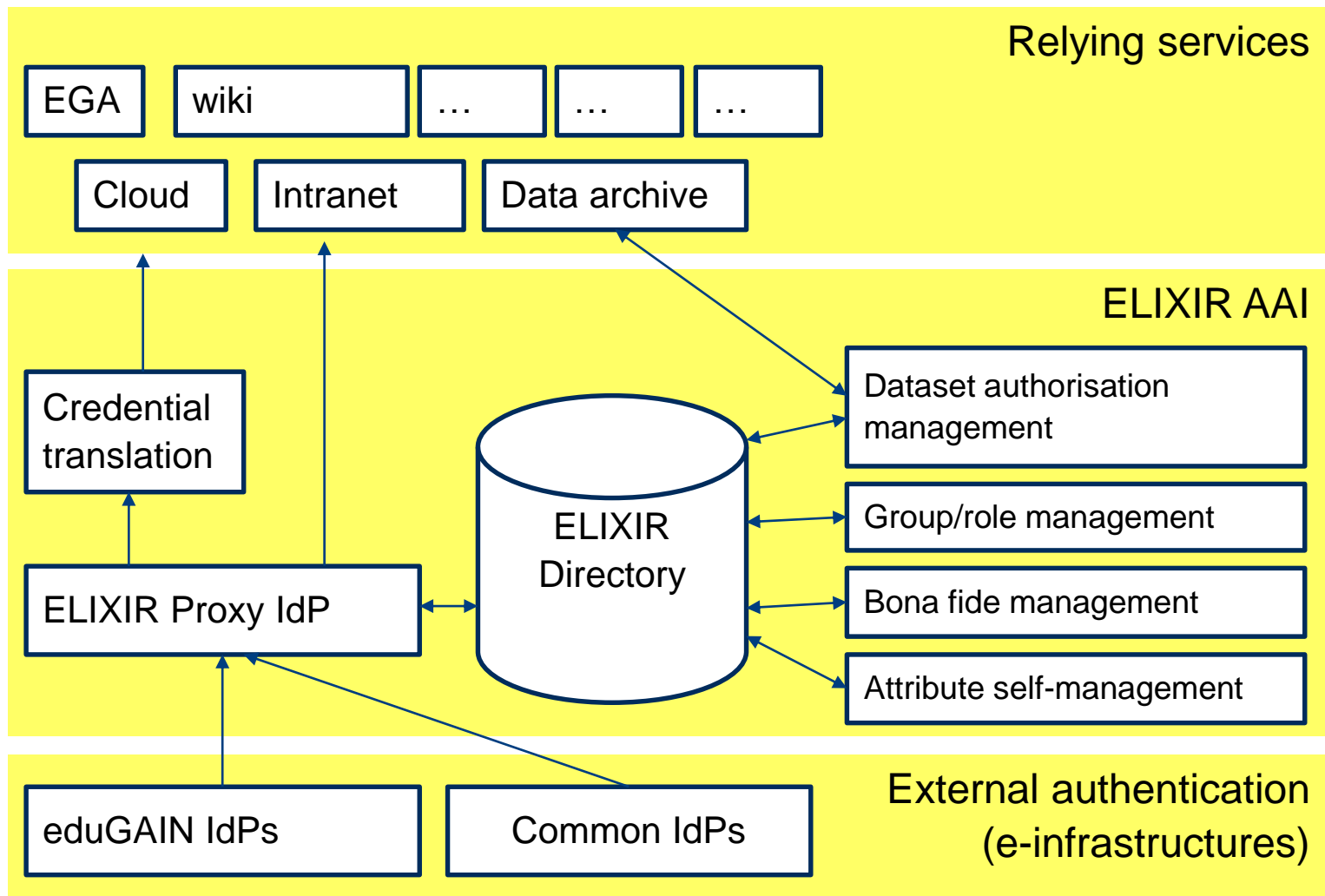


# Building capacity in Europe

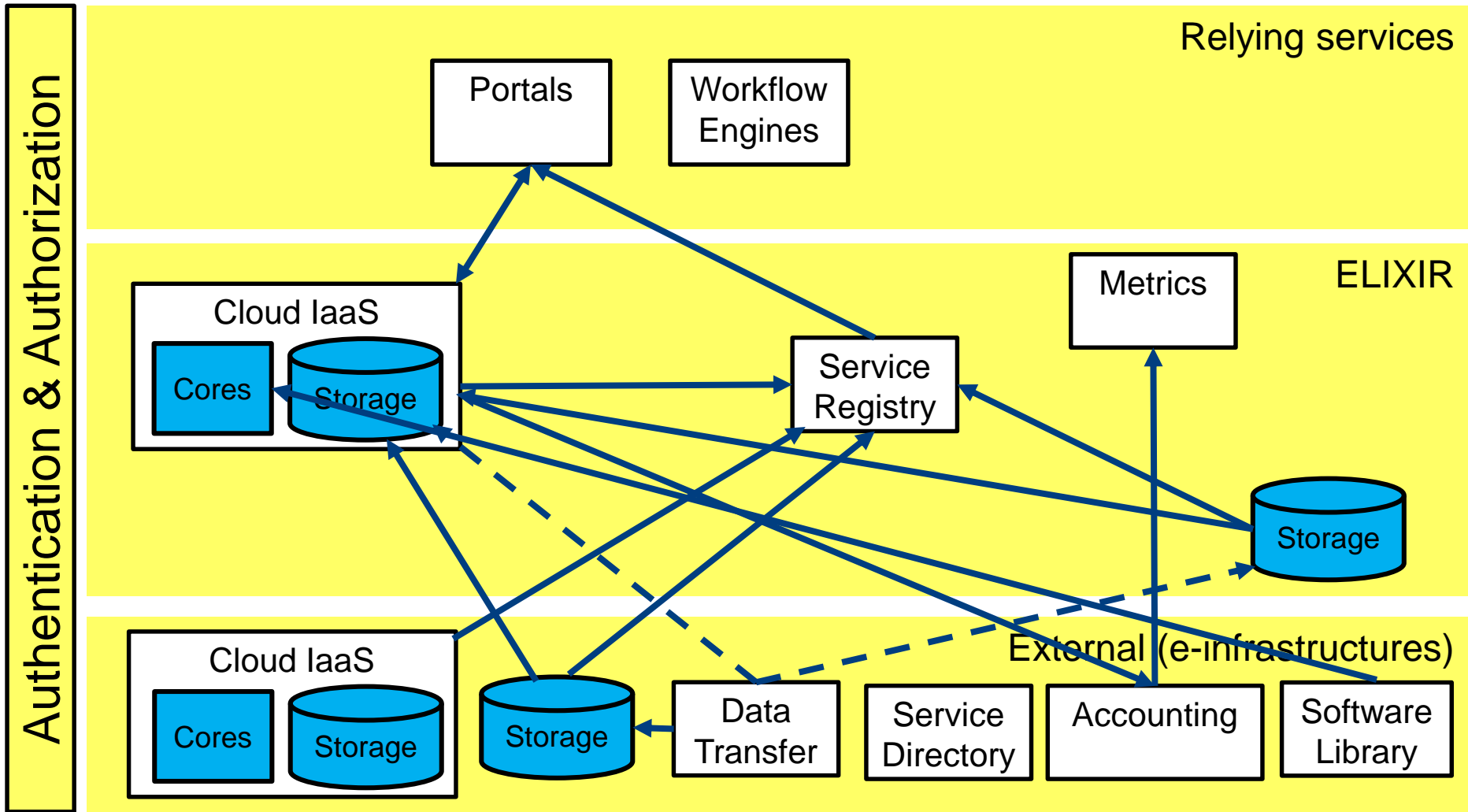
- ELIXIR: a sustainable infrastructure for biological information in Europe.
- Supporting life science research and its translation to
  - medicine
  - agriculture
  - the environment
  - the bioindustries
  - society.



# AAI Architecture



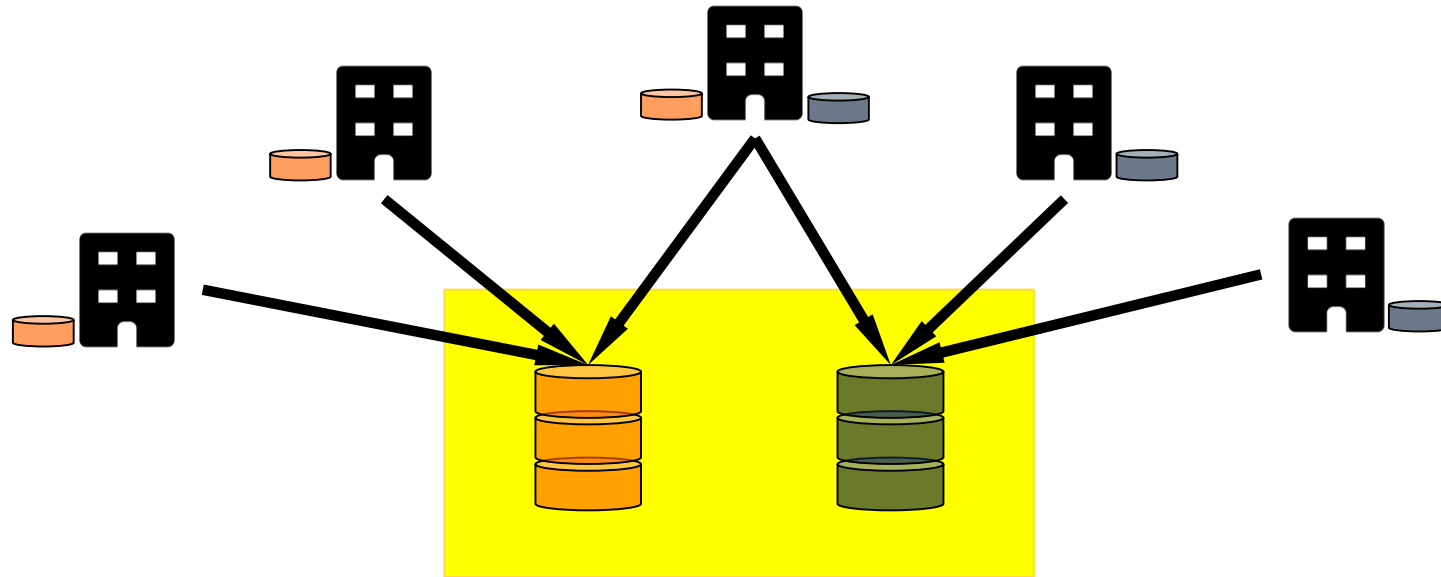
# Cloud & Storage Architecture



# Initial Prioritisation and Grouping Analysis

- AAI
  - 0: Federated ID, Other ID
  - 1: Elixir ID
  - 2: Credential Translation, Group/Attribute Mgmt, Endorsed Attributes
- Cloud Compute
  - 1: Cloud IaaS, HTC Cluster, Cloud Storage, Federated Cloud IaaS
  - 2: Infrastructure Service Registry & Directory, VM Library
  - 3: Operational Integration, Resource Accounting
- Data Transfer
  - 1: Network File Storage, File Transfer
  - 2: Data Set Replication, PID & Meta-data Registry

# Centralization & specialization



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



Data production



Data centralization

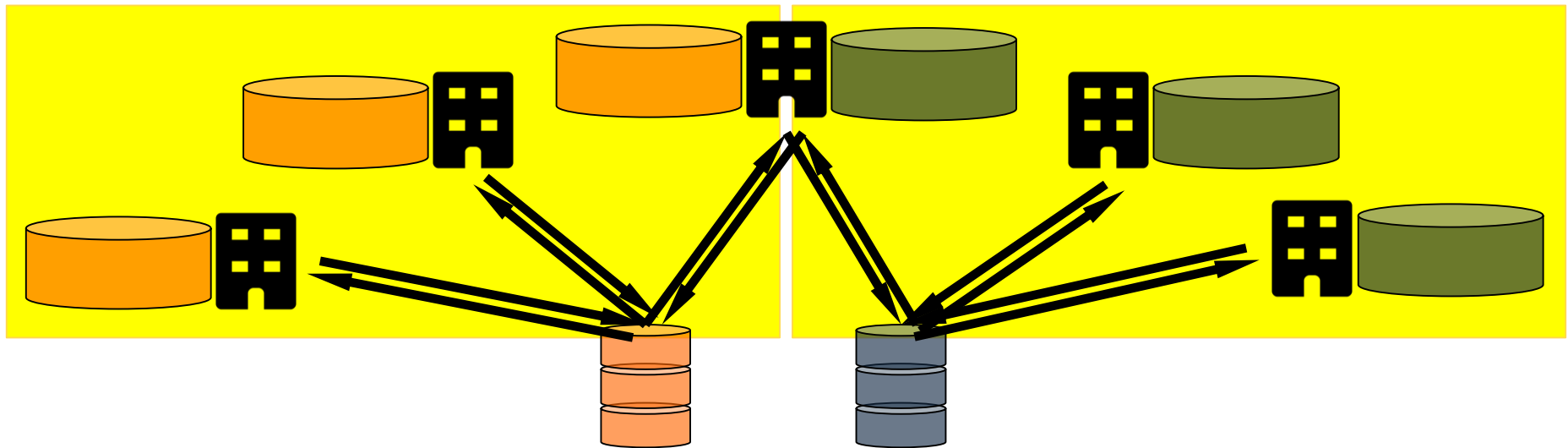


Data





# 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



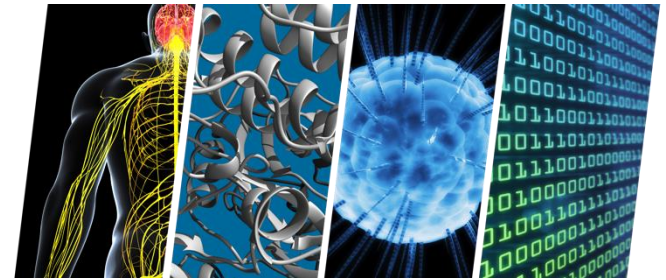
Data



# 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.



# Technical Services Cluster: IT as a Service

- ‘Professionalising’ our IT Services
  - Looking at lightweight FitSM portfolio
- 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

# External Infrastructure Activities

- EUDAT
  - Leverage CDI to strategically replicate large ‘popular’ data sets
- HelixNebula
  - Meet future needs through hybrid model
  - External hosted data centres with direct cloud access
- EGI: European Grid Infrastructure
  - Provides federated cloud & cluster resources
- CERN OpenLab
  - Identifying and building common IT services for science
- Cancer Research UK
  - Bringing islands of protected data together for cloud based analysis

# 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

Virtualised Infrastructure

Virtualised Infrastructure

Virtualised Infrastructure

Elixir Service

Storage

Compute

Storage

Compute

Data

Storage

Compute

Data

Data

Cloud Providers

EMBL-EBI Data Centre

Infrastructure Provider

Data Provider

Geant Network

# Thank you

[steven.newhouse@ebi.ac.uk](mailto:steven.newhouse@ebi.ac.uk)