Surfing the Tsunami

Biology Data and EBI's Infrastructure

Andy Jenkinson, Manuela Menchi

EMBL-EBI

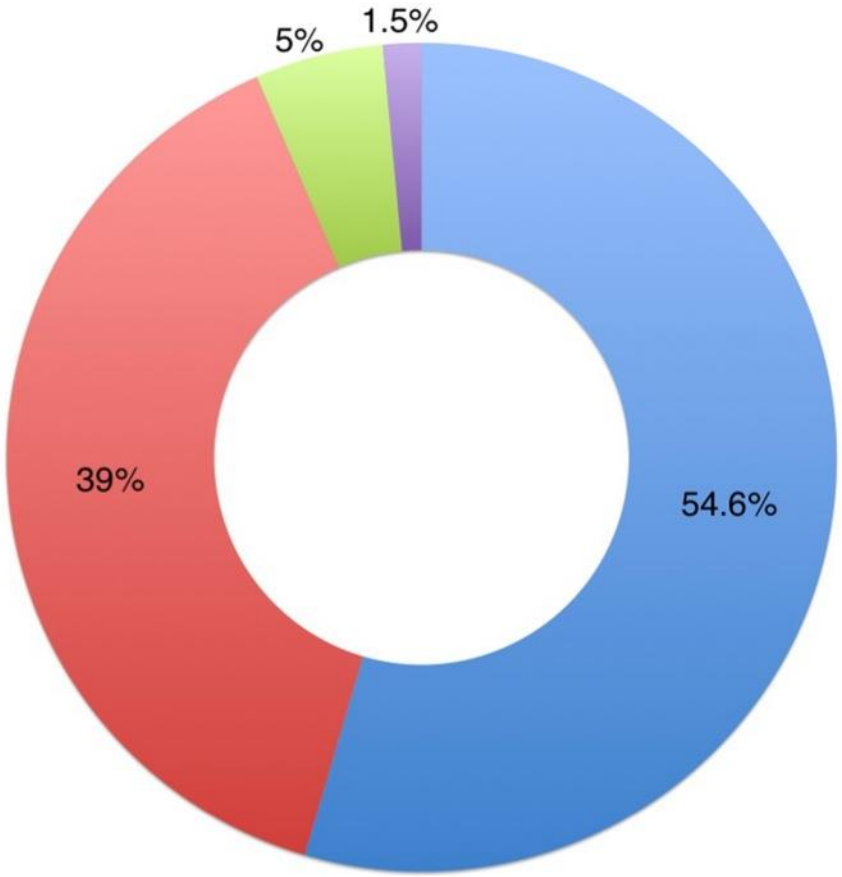
Overview

- Data in Molecular Biology
  - Bioinformatics and the EBI
  - What are the challenges for bio data?
  - Sequencing: a disruptive technology
- IT Architecture perspective

Introduction

- Bioinformatics
  - The application of IT to biology
  - Interdisciplinary science
  - Pervasive in modern molecular biology
- The European Bioinformatics Institute
  - Part of the European Molecular Biology Laboratory
  - International, non-profit research institute
  - Europe’s hub for biological data services and research

Who are our users?



- Bench laboratory researchers
- Bioinformaticians & computer scientists
- Roles related to research (trainers, teachers, outreach)
- Non-scientific roles (admin, funders)

Data resources at EMBL-EBI

Genomes & variation
- Ensembl
- Ensembl Genomes
- Genome-phenome archive
- Metagenomics

Nucleotide sequences
- European Nucleotide Archive (ENA)

Literature & ontology
- Europe PubMed Central
- Gene Ontology

Expression
- Array Express
- Expression Atlas
- PRIDE
- R-Workbench

Molecular structures
- Protein Data Bank in Europe
- PDBsum
- ProFunc

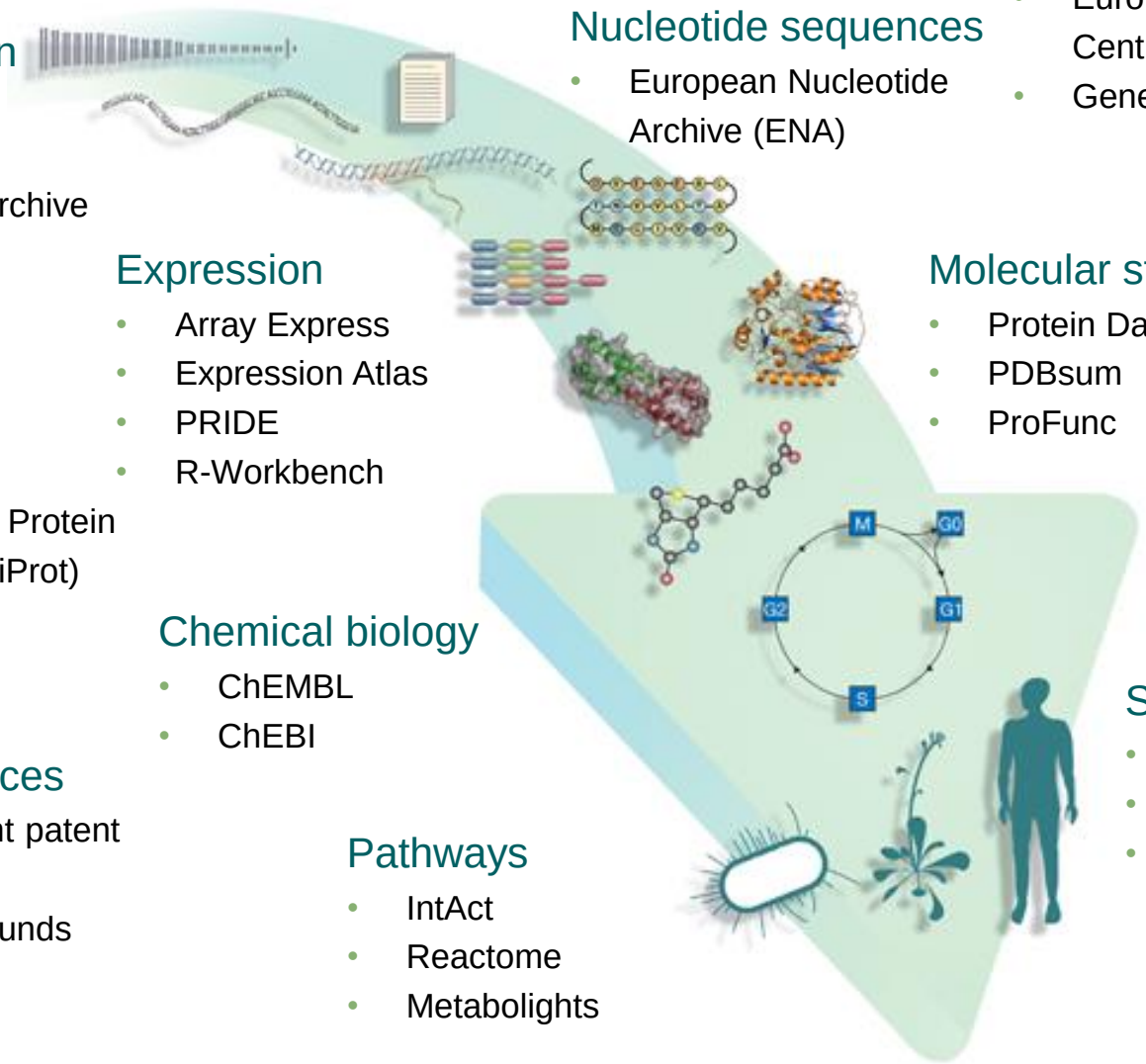
Proteins
- The Universal Protein Resource (UniProt)
- InterPro

Chemical biology
- ChEMBL
- ChEBI

Patent sequences
- Non-redundant patent sequence dbs
- Patent compounds

Pathways
- IntAct
- Reactome
- Metabolights

Systems
- BioModels
- Enzyme Portal
- BioSamples



Molecular Biology Data

- Huge variety of methods, each with its own output
- Protein structure
  - Atomic coordinates
  - Volume maps
- Small molecule structures
- DNA, RNA sequences
- Hybridisation (which proteins interact with each other?)
- Microarrays (which genes are expressed in this cell?)
- Mass spectrometry (which proteins are in this cell?)
- Images

Global sequence archive

- International Nucleotide Sequence Database Collaboration (INSDC)
  - NCBI (US), DDBJ (Japan), EMBL-EBI (Europe)

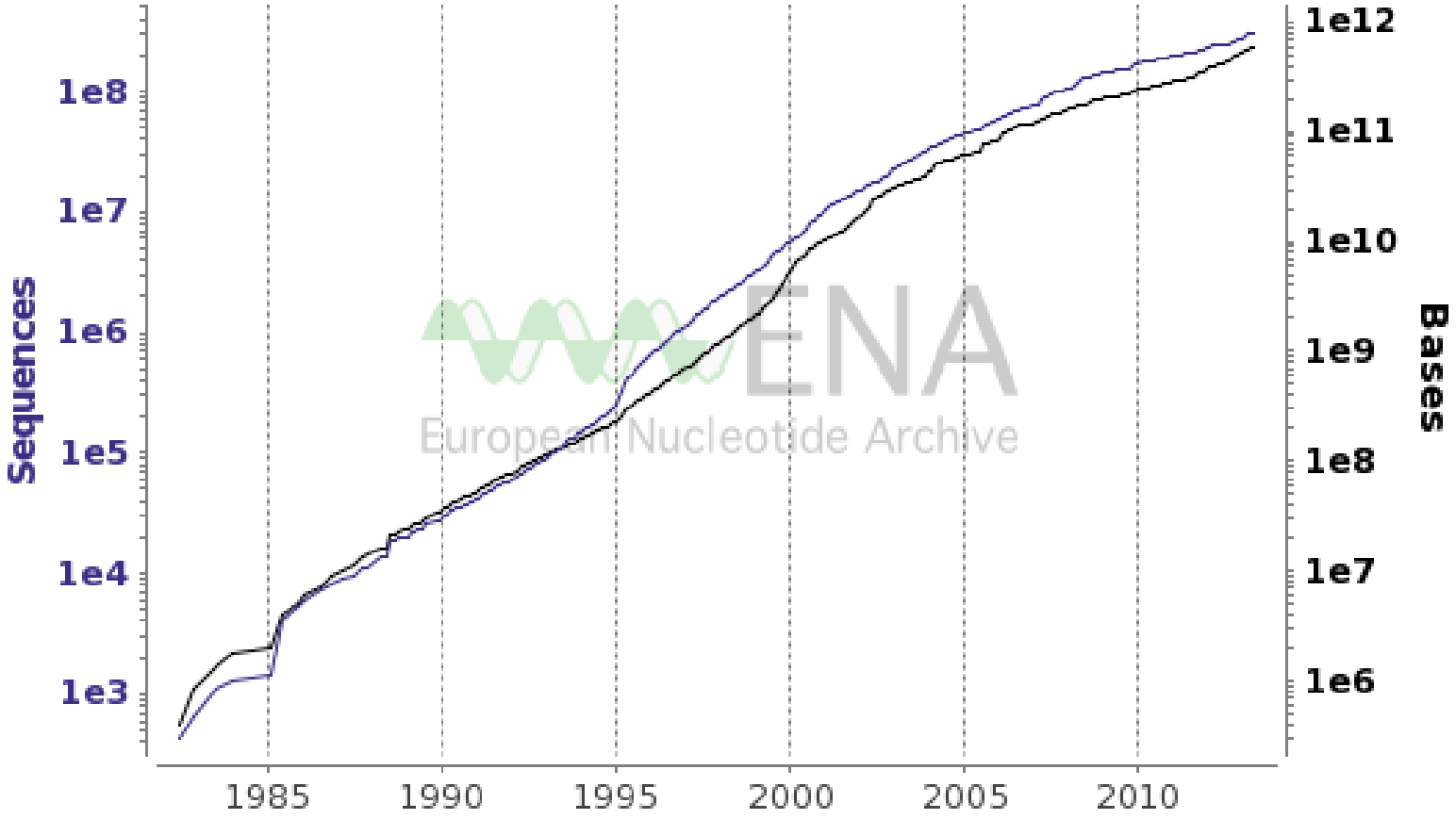
- 1982: EMBL Data Library - 500,000 base pairs
- 1990: Human Genome Project starts
- 1995: EMBL Nucleotide Sequence Database relocated to Hinxton, Human Genome Project
- 2007 or so: second-generation sequencing

Data-driven research

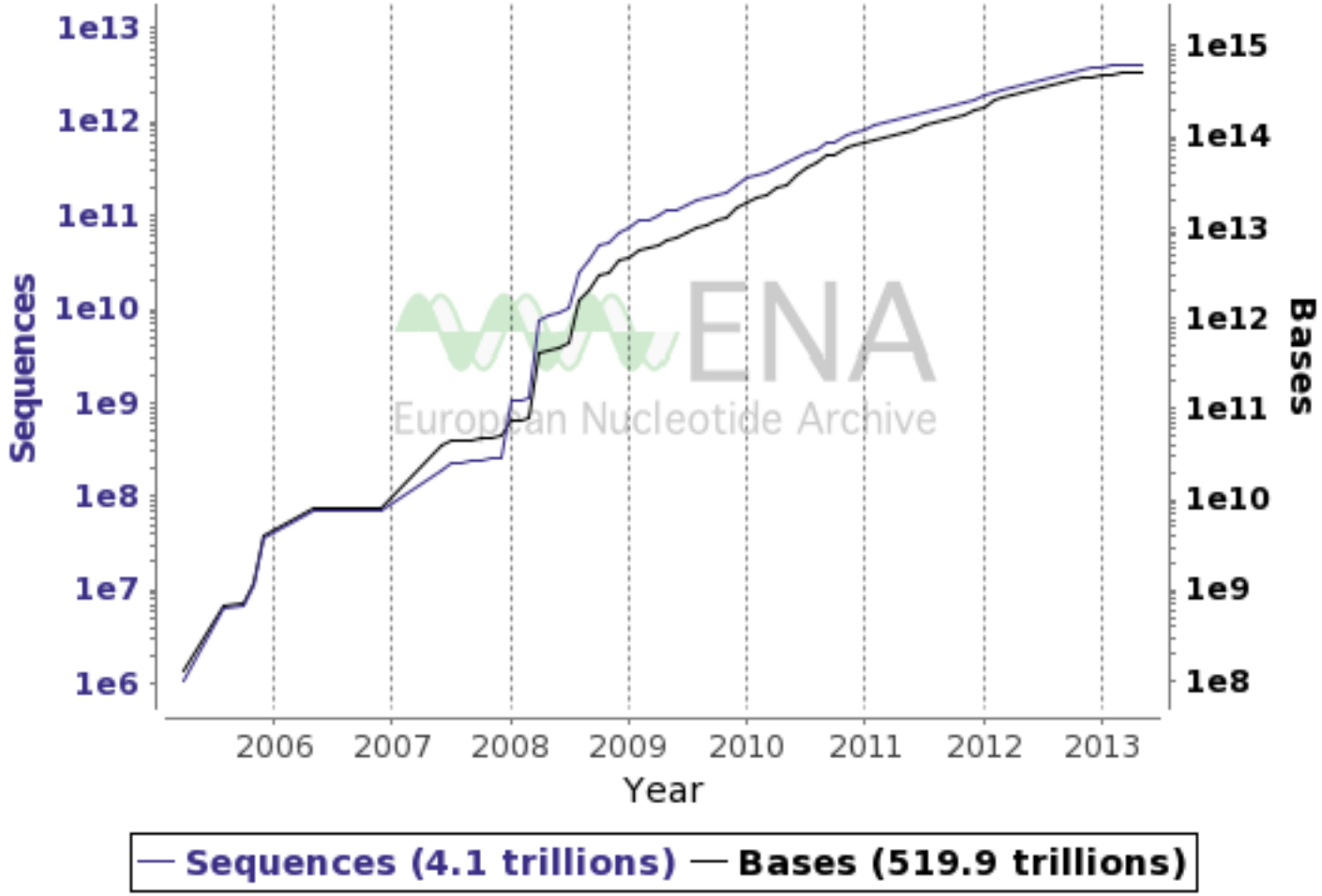
- Faster: human genome 13 years -> 1 day
- Cheaper: $3 billion -> $1500

- Genome centres can sequence more
- Any bio lab can sequence
- Sequencing is now a generic and pervasive method
- Do something to your sample, sequence it
  - DNA methylation (bind with a blocking agent, sequence)
  - Gene expression (sequence the RNA)
- Data driven research

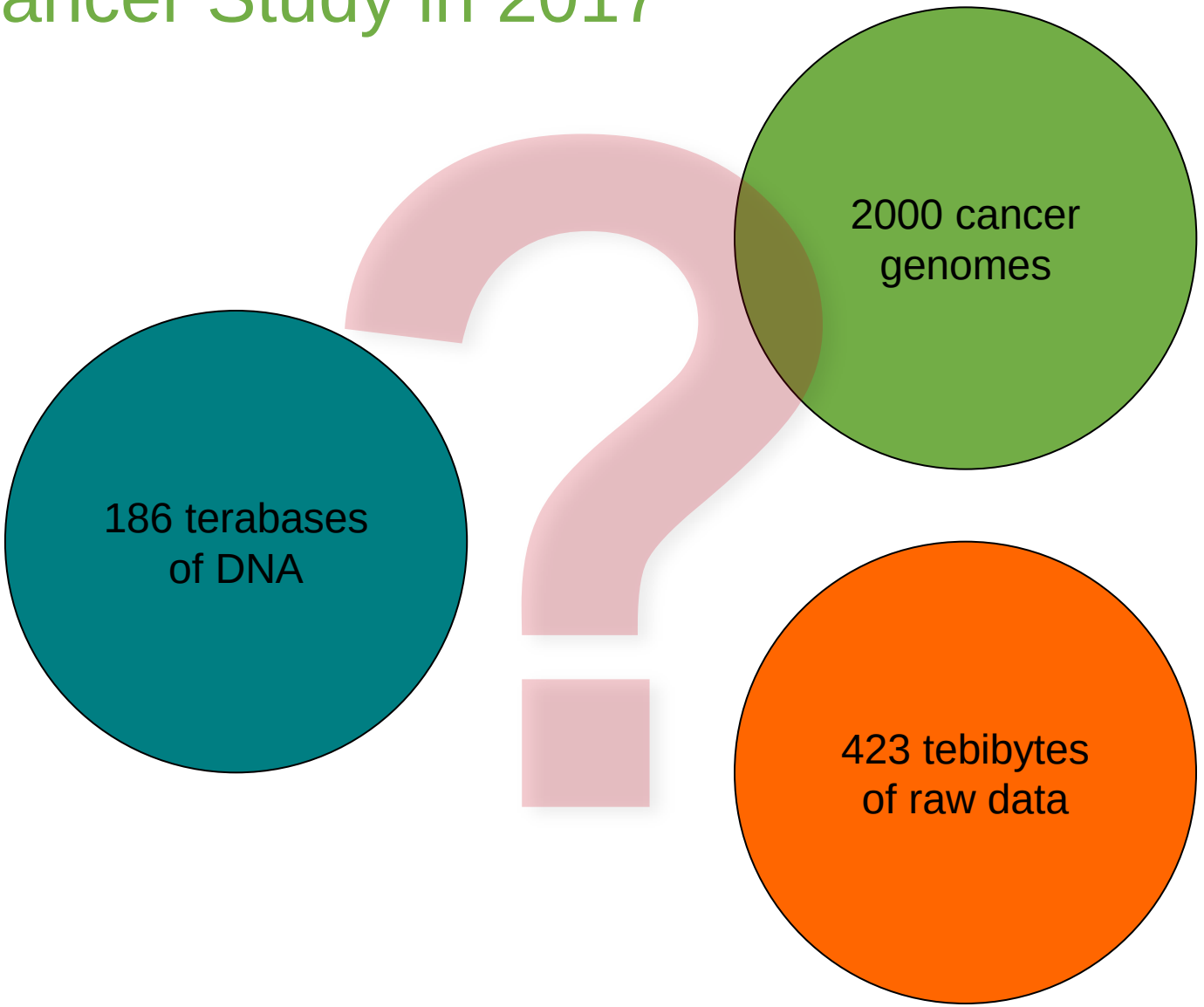
EMBL-Bank (assembled sequences)



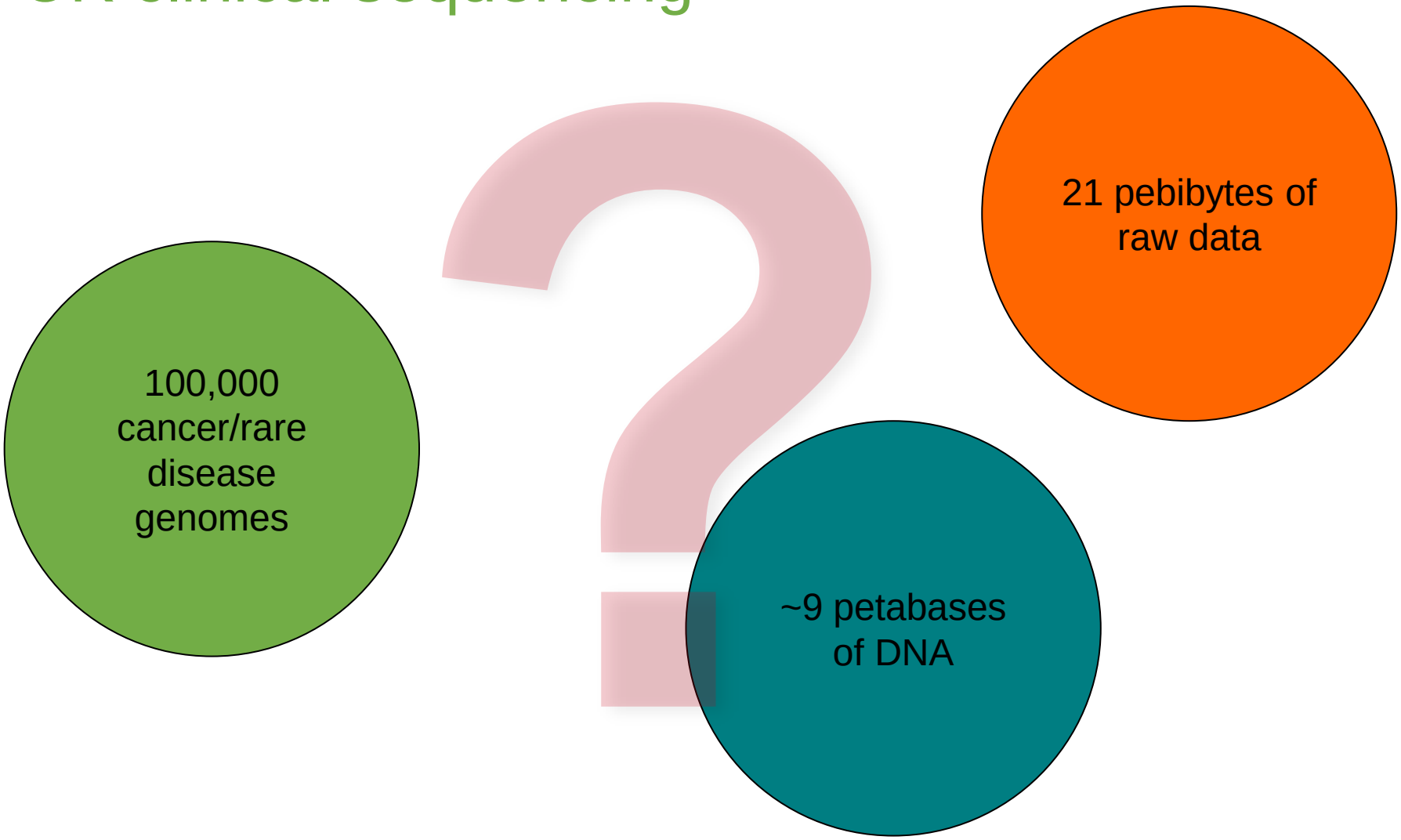
Sequence Read Archive (2nd gen raw data)



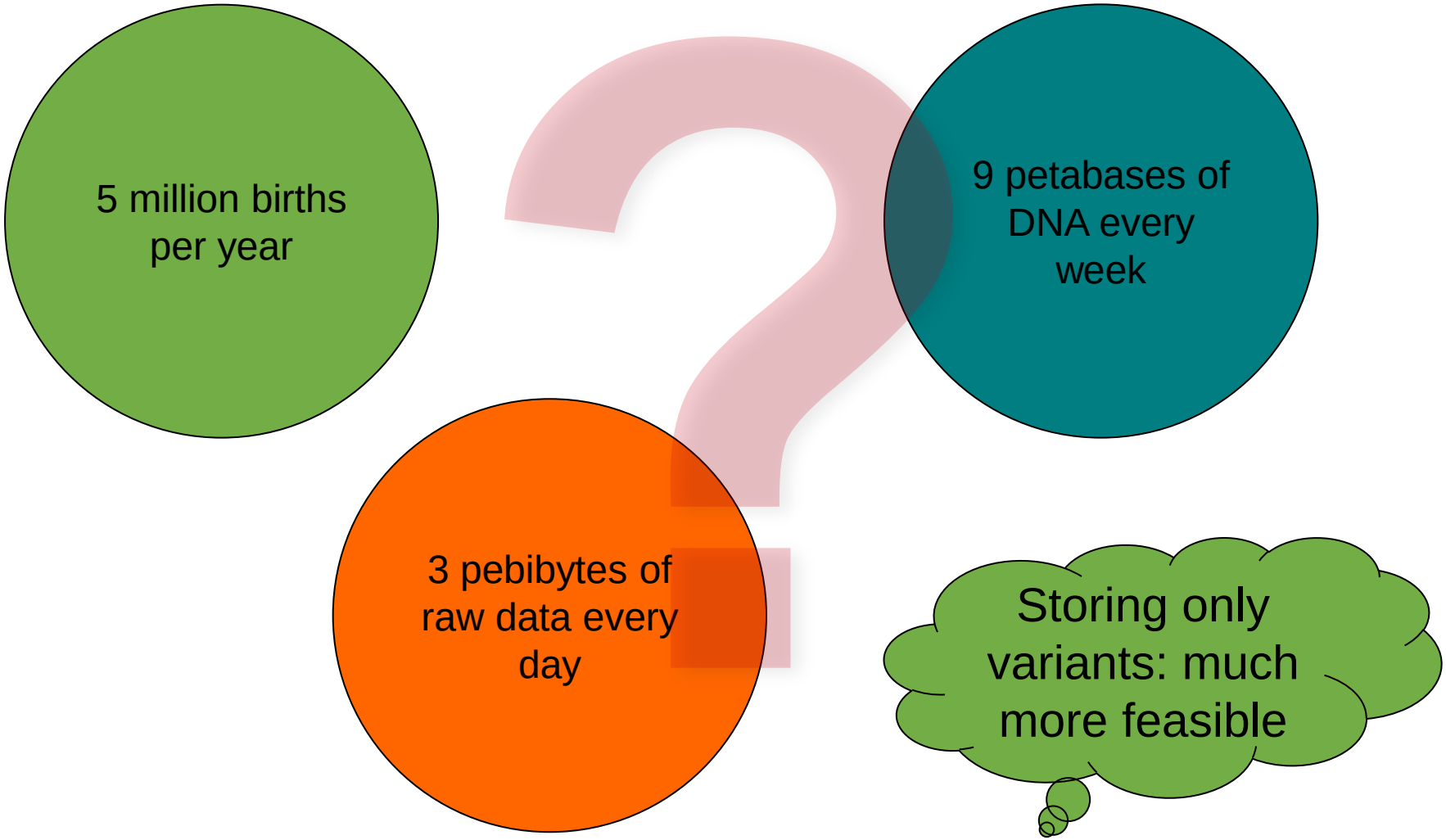
A Cancer Study in 2017



UK clinical sequencing



Sequencing every child born in the EU?



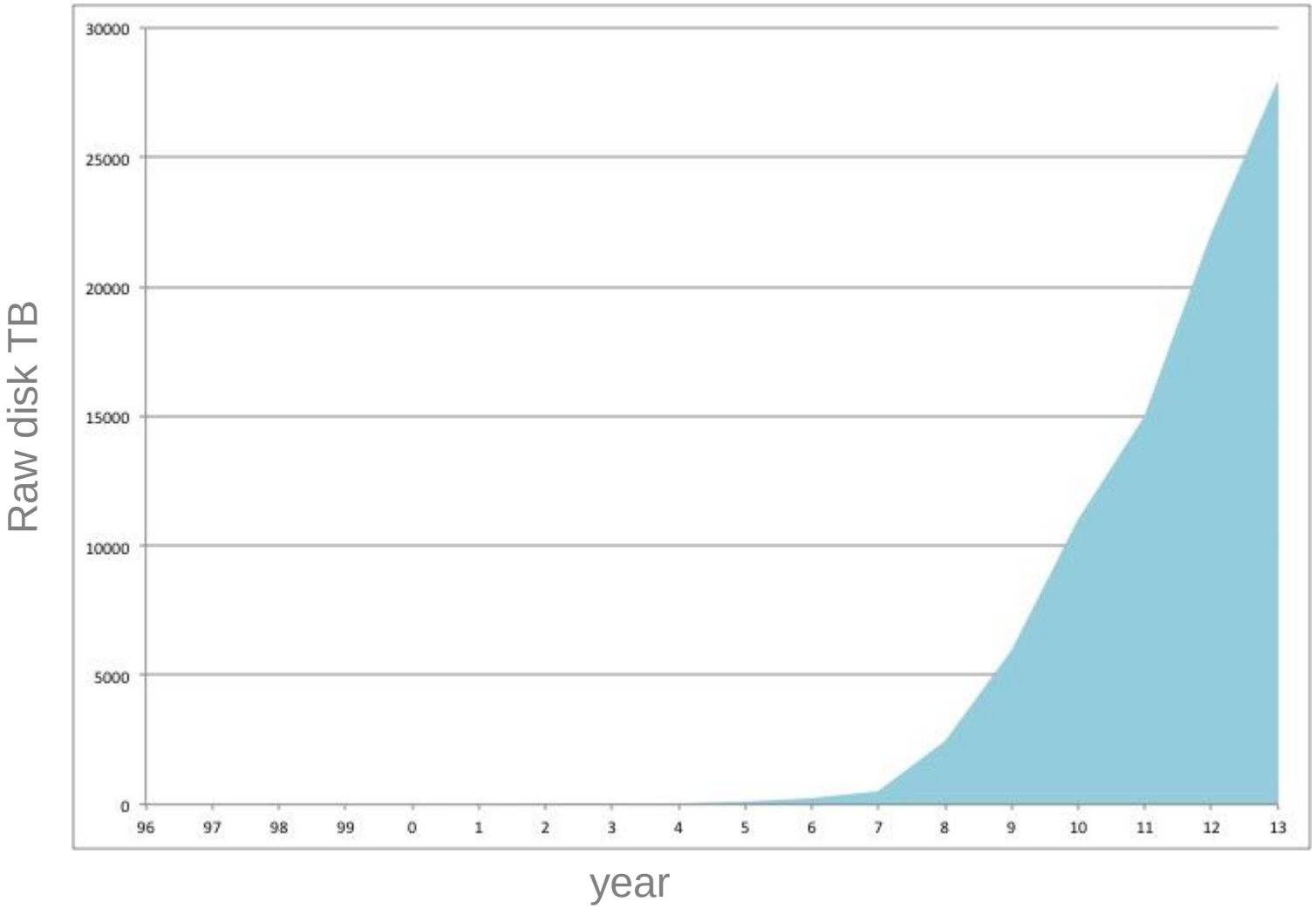
IT Architecture Perspective

This part of the contribution comes from the IT Systems team of EMBL-EBI and provides an overview of Data IT Architecture and relevant challenges at EMBL-EBI.

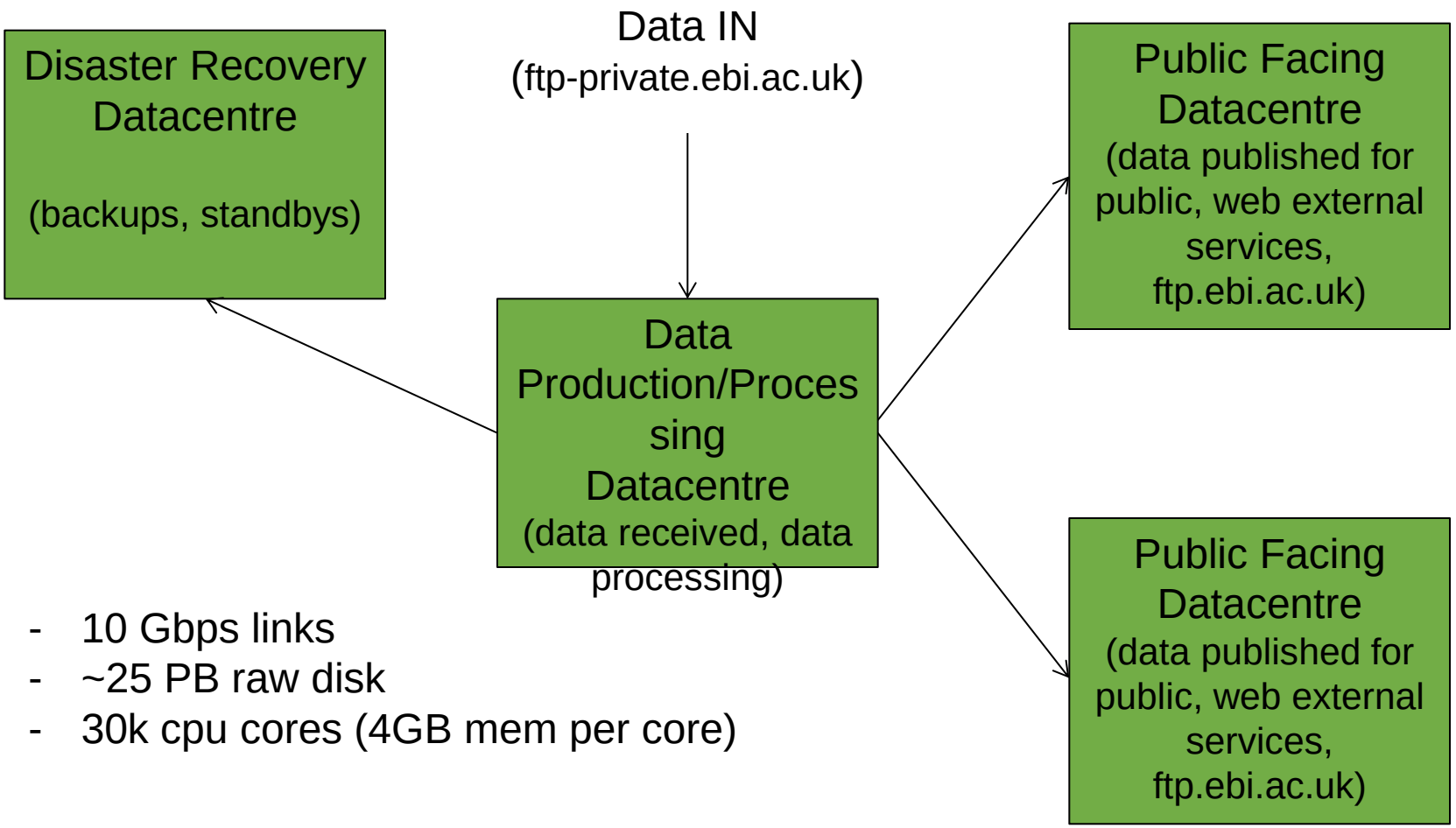
The EBI Systems Group

- EBI Systems Team is dedicated to plan, procure, implement, maintain IT infrastructure and services:
- core and desktop support: total 25 people/20 core
  - **Core hardware infrastructure** datacentre**,** disaster recovery site, computing servers, storage, network (LAN/WAN)
  - **Core services infrastructure** network (ftp, aspera, rsync, email), LSF, database administration, backups
  - **Procurement of central infrastructure** (equipment/datacentres)
  - **Core user support** of EMBL-EBI service groups in their daily activities. The group works closely with all project groups maintaining and planning their specific

How has EBI disk grown

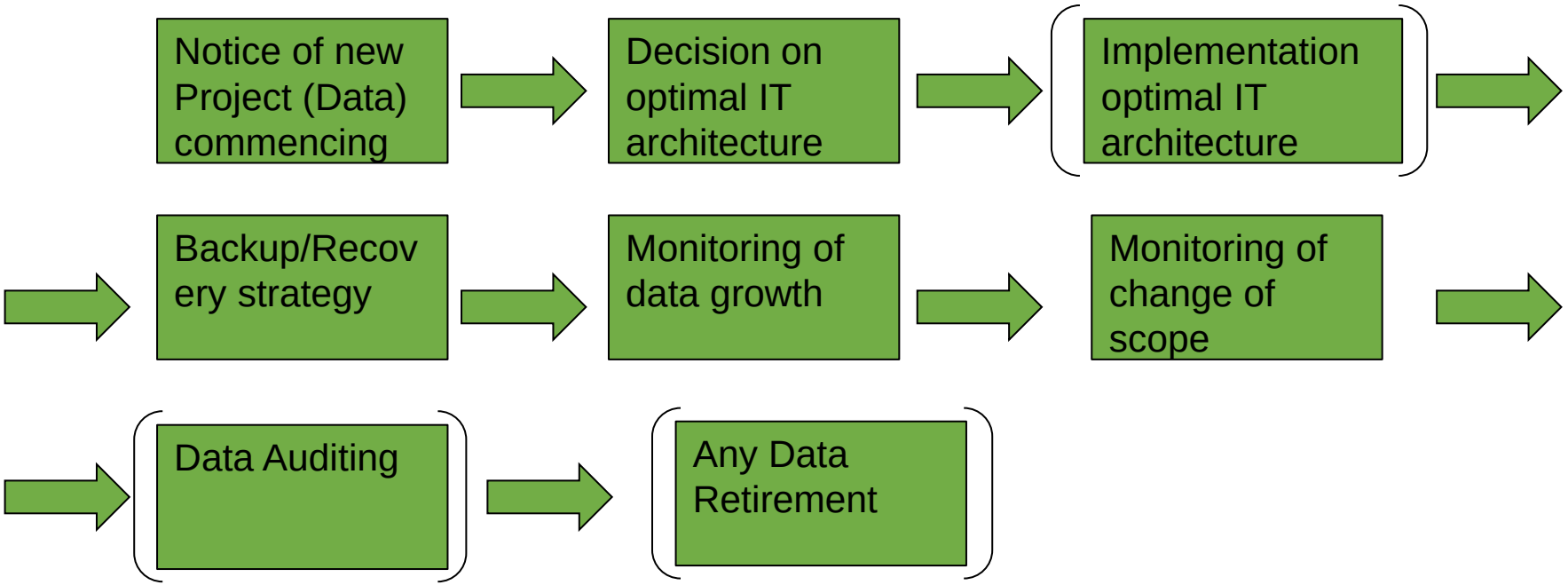


Relevant EMBL-EBI facilities



- 10 Gbps links
- ~25 PB raw disk
- 30k cpu cores (4GB mem per core)

Lifecycle of Data Management in IT Systems



Lifecycle of Data Management in IT Systems

What is data? Data is “bytes” with “information” associated
Information typically asked/provided by any new service group project:

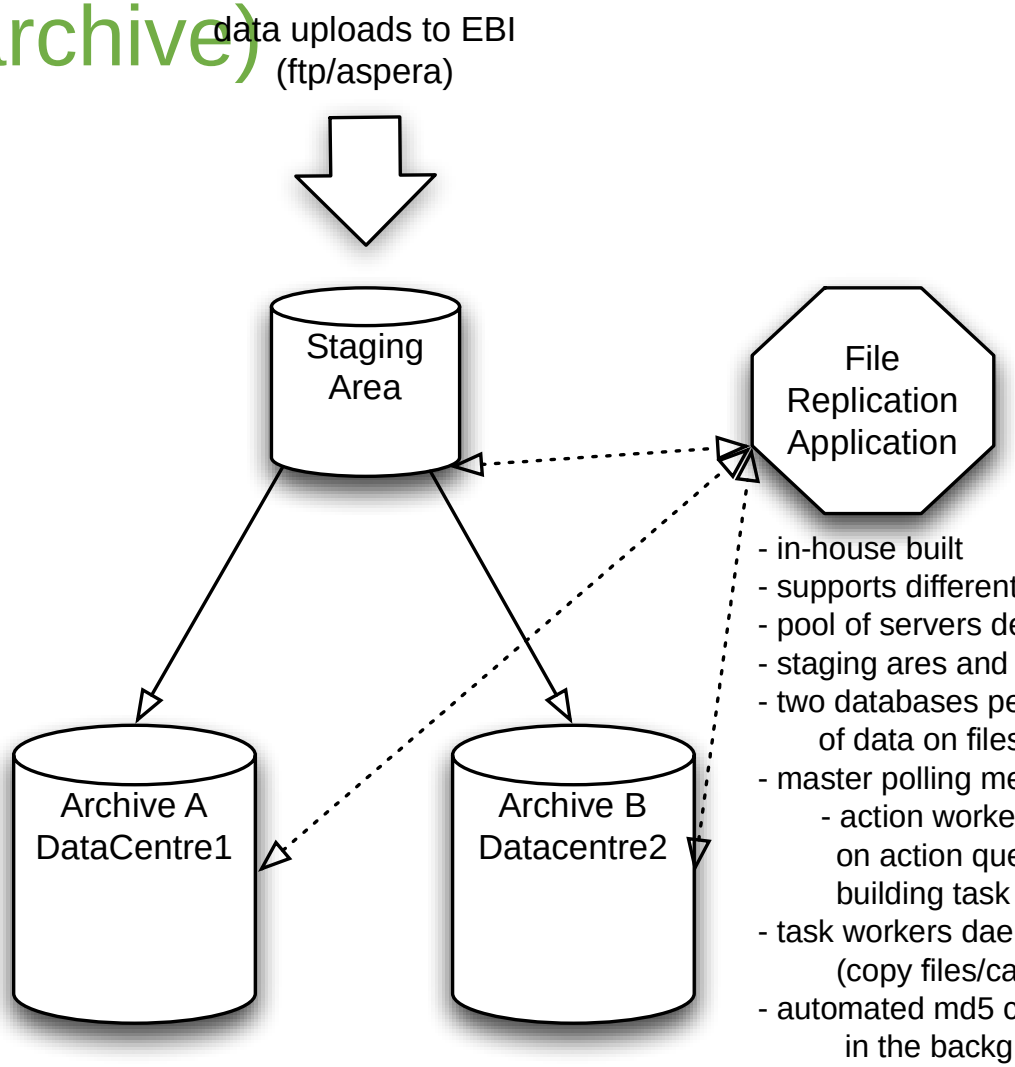
- Initial dataset size
- Data lifecycle /Scope of the data within the project (Archiving, backend of public facing application, HPC, Database, temporary, data transfer, etc)
- Backup or no backup
- Growth Forecast
- Ownership and access list

Optimal IT solution for data/project

Decision making process: focus on balance between simplicity of central IT data infrastructure as a whole and optimal solution per individual dataset/project. Continuous research of market and evaluation of solutions.

- Consolidation into a reduced number of storage solutions: traditional NAS, scale-out NAS, SAN, flash memory arrays, parallel storage
  - **Scalability**: size and IOPS: general data and concurrent access from HPC farms and **latency**: sensitive applications (databases)
- Storage Archiving:
  - **Scalability** for sizes order PBs: SRA archive ca 6PB (was ~1.6PB in 2011) and **Data integrity** (checksums and hardware strategy where possible)
- Data migrations/mirroring: disaster recovery, external facing facilities, standby sites, technology refresh, geographical distance
- Floor space and electricity, cooling

FIRE: File Replication system (inc SRA archive)



- in-house built
- supports different projects for archving
- pool of servers dedicated to FIRE operations
- staging ares and two distinct archives
- two databases per Project: metadata actions DB and metadata of data on filesystem DB
- master polling metadata actions DB for actions
  - action workers daemons performing metadata actions on action queue (archive/dearchive/migrate) and building task queue
- task workers daemons processing task queue (copy files/calculate checksums/delete files ..)
- automated md5 checks on archived files continuously running in the background