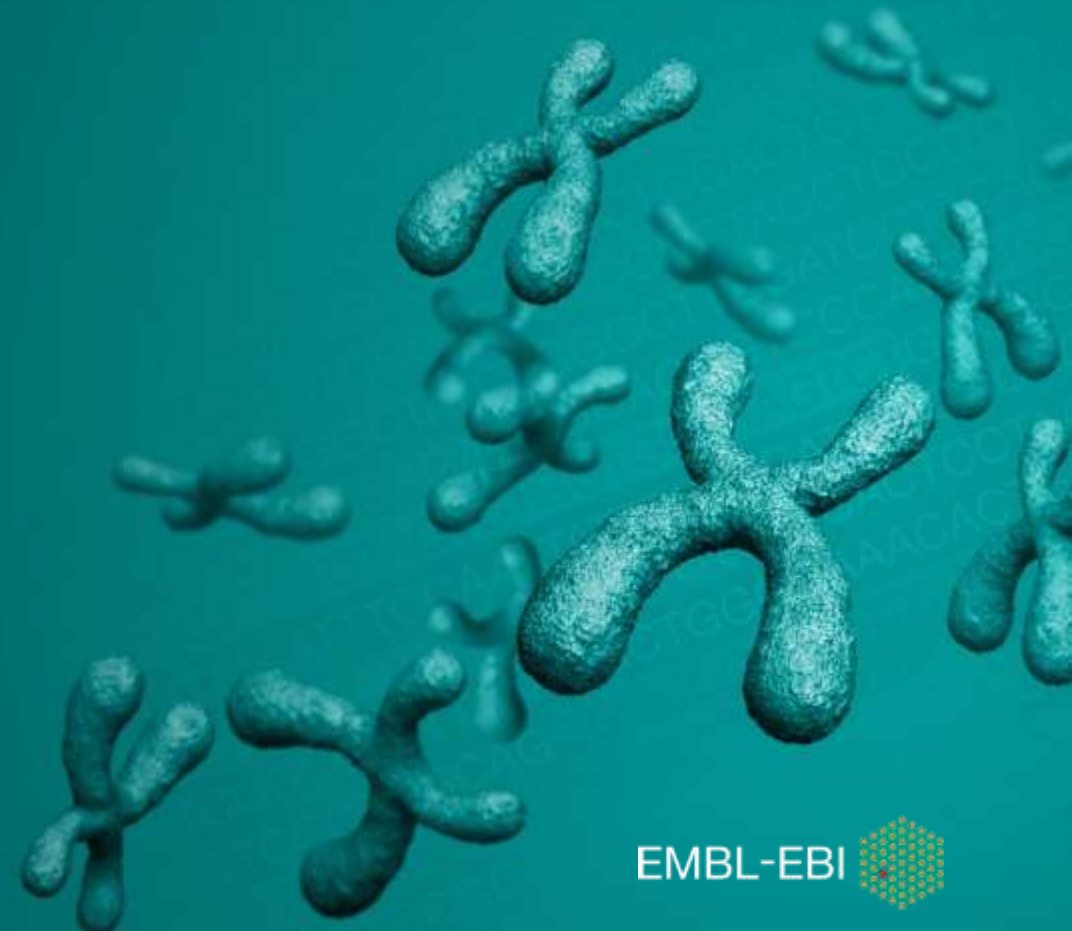


Putting Data Science in the Centre: Observations from EMBL, a multisite, international life science research organisation

Rolf Apweiler
Director, EMBL-EBI
www.ebi.ac.uk



What is EMBL-EBI?

- Europe's home for biological data services, research and training
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- Home of the ELIXIR Technical hub



EMBL Member States

Member states (27)

Austria 1974	Belgium 1990
Denmark 1974	Portugal 1998
France 1974	Ireland 2003
Germany 1974	Iceland 2005
Israel 1974	Croatia 2006
Italy 1974	Luxembourg 2007
Netherlands 1974	Czech Republic 2014
Sweden 1974	Malta 2016
Switzerland 1974	Hungary 2017
United Kingdom 1974	Slovakia 2018
Finland 1984	Montenegro 2018
Greece 1984	Poland 2019
Norway 1985	Lithuania 2019
Spain 1986	

Associate member states

Australia 2008

Prospect member states

Estonia
Latvia



Six sites with almost 1800 people and >90 nationalities



EMBL-EBI

Bioinformatics

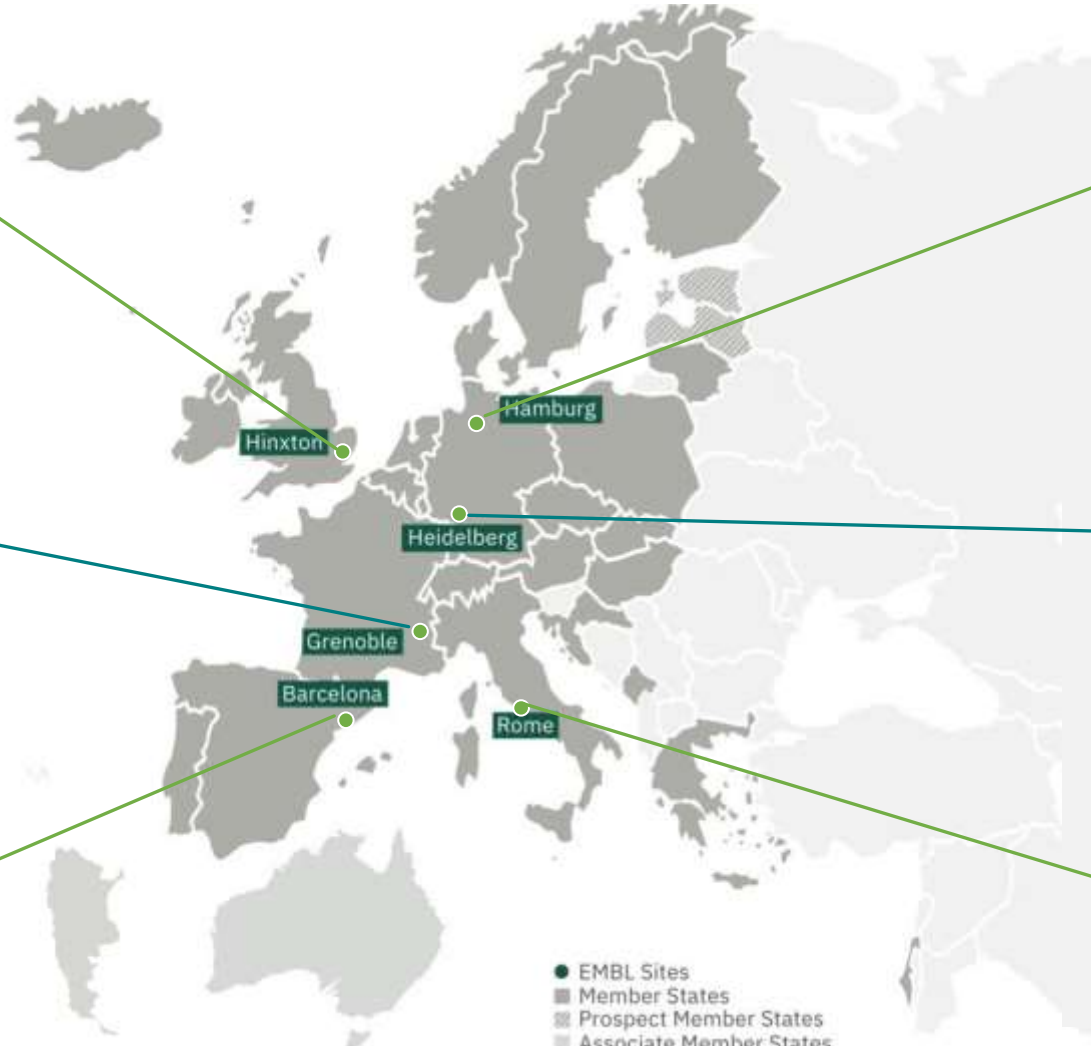


Grenoble

Structural
biology



Barcelona
Tissue biology
and disease
modelling



Hamburg

Structural
biology



Heidelberg

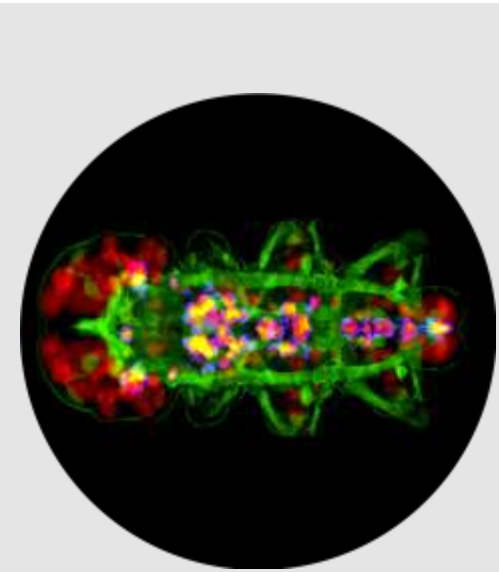
Life sciences



Rome
Epigenetics
and
neurobiology



EMBL's missions



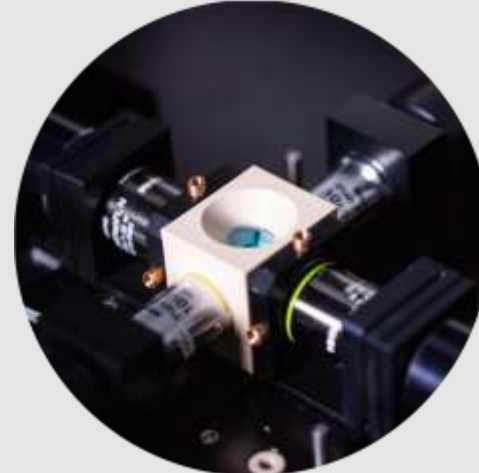
**Excellent
Research**



**Scientific
Services**



**Advanced
Training**

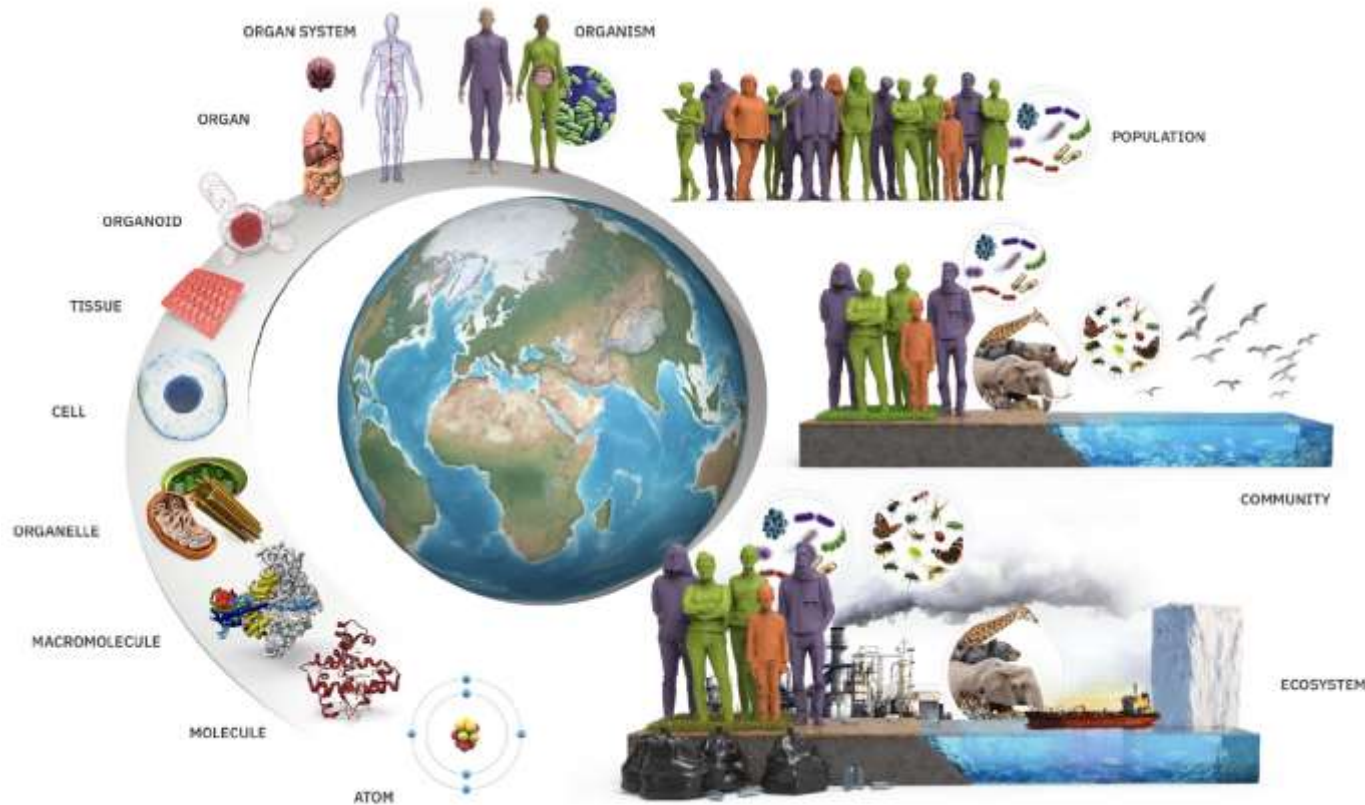


**Innovation and
translation**



**Integrating
life sciences**

EMBL's bold scientific vision: Molecules to Ecosystems



The next step is to understand:

Life in context

Molecular mechanisms

Phenotype: **G**enotype x **E**nvironment

Organisms in communities:

Symbiosis, Parasitism, Commensalism, Mutualism, Infection, Predation

Molecules and organisms in ecosystems

Response and adaptation
to changing environments

New Transversal Themes



Planetary Biology

Understanding at the molecular, cellular, organismal, and population levels how microbes, algae, plants, and animals interact with and respond to environmental change



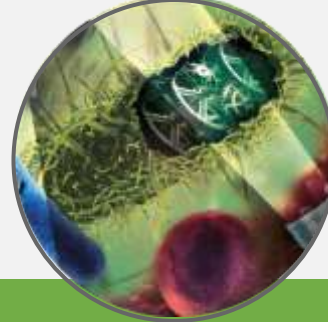
Human Ecosystems

Understanding how the environment impacts humans as individuals and within populations
How genotype and the environment influence human phenotypes and disease



Infection Biology

Integrating multidisciplinary experimental and computational approaches to understand how pathogens and their hosts interact



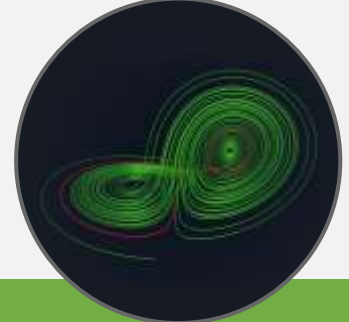
Microbial Ecosystems

Understanding the functional diversity of microbial species and strains, and the interactions and properties of microbial communities within their ecosystems



Data Sciences

Data science centres at all EMBL sites will facilitate research, provide support and training, advance novel data science methods, set technical standards, and offer public data resources

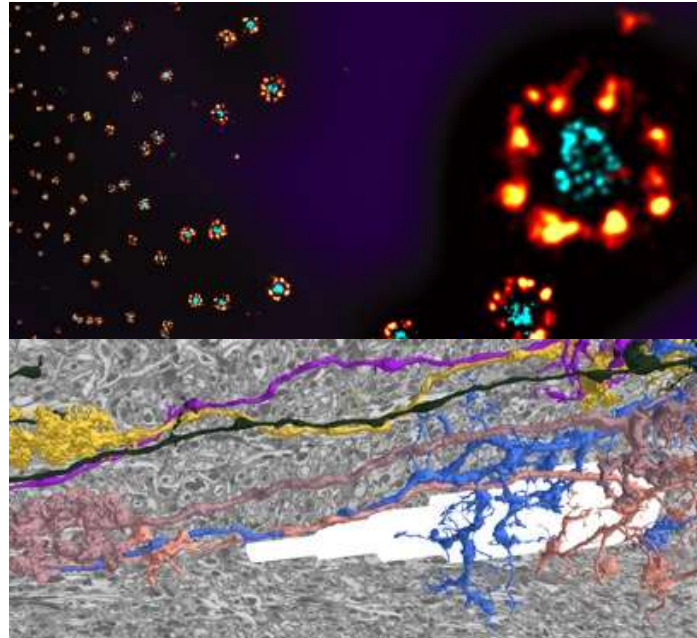
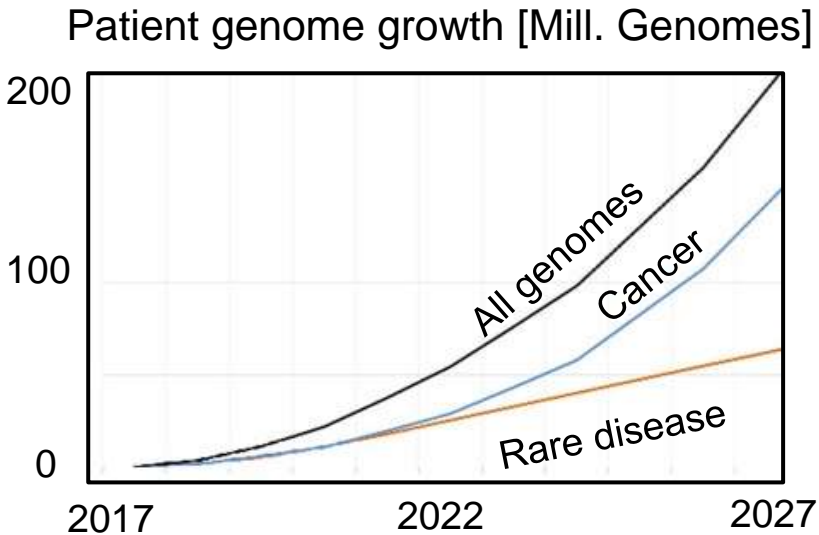


Theory

Explaining biological phenomena using mathematical formalism and models, turn data into understanding, and generate testable predictions

Data Sciences in Biology

Data doubling every 18 months in diverse biological fields



Global Alliance
for Genomics & Health

Many image
data types



Multiple data types
generated by the EMBL
Programme 2022-2026

Exponential Biological Data Growth

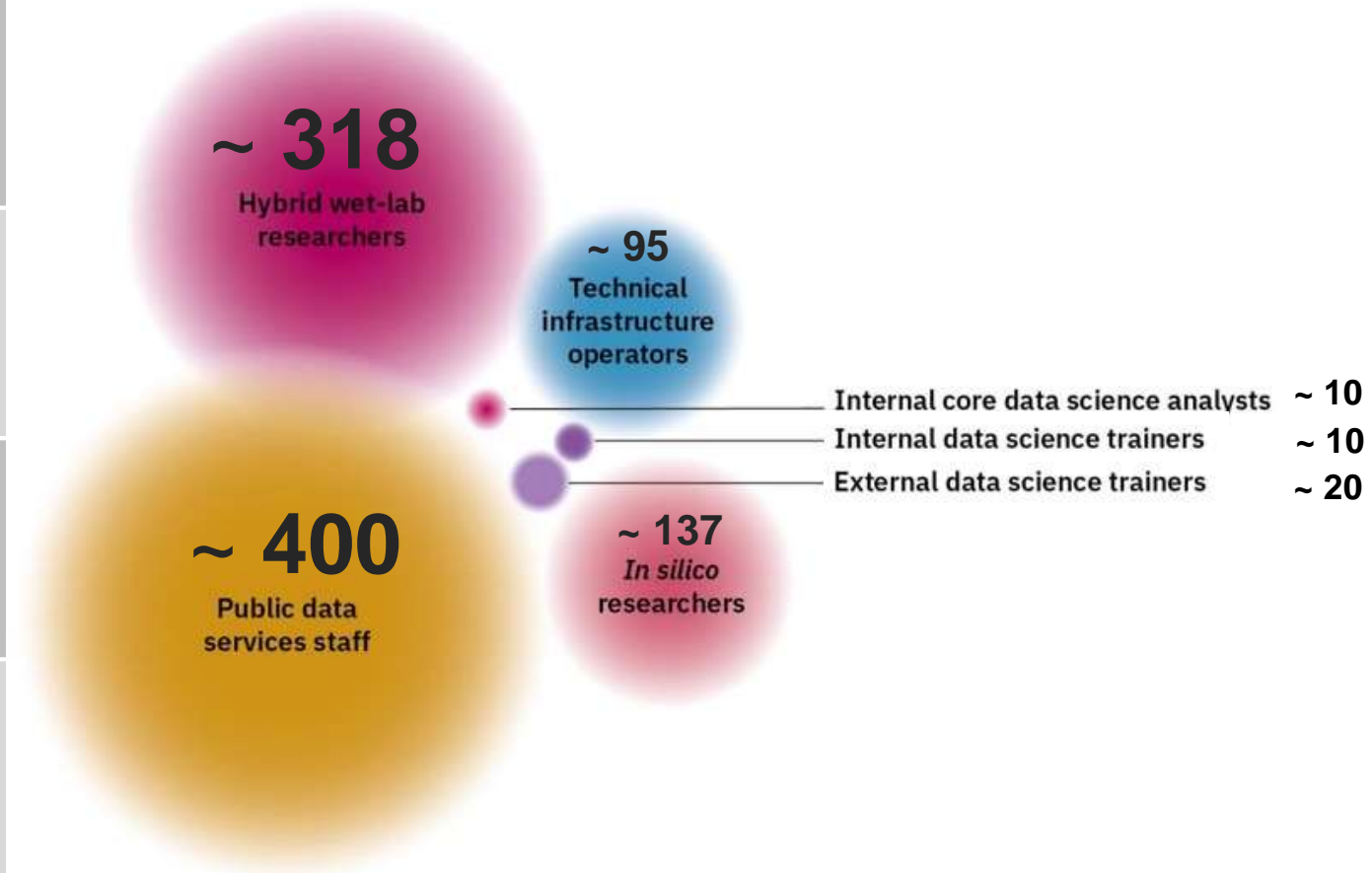
Researchers keen to maximise the utility of their data

~990 data scientists at EMBL

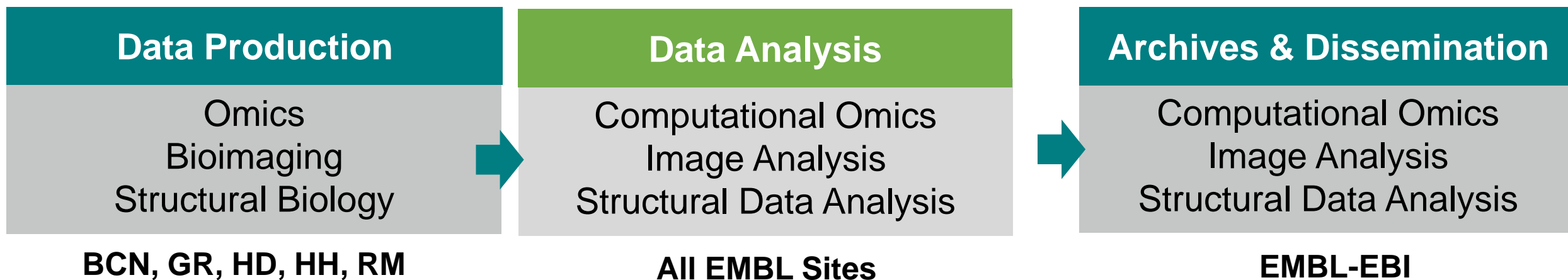
Interdisciplinary mindsets

Teamwork

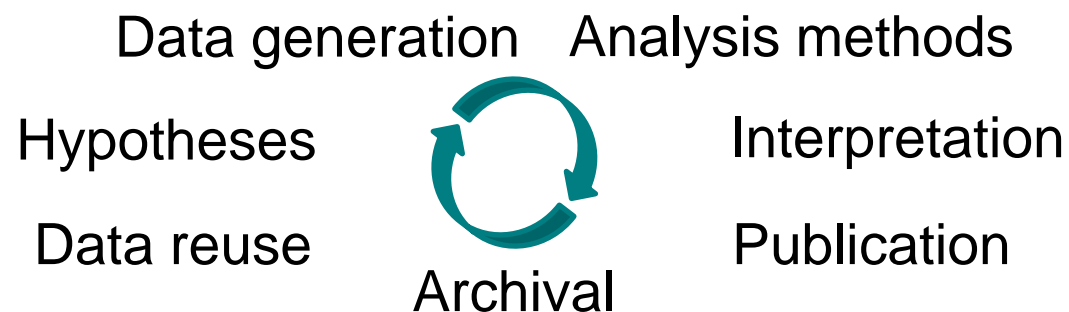
Hybrid “wet-dry” scientists
on the rise in different subfields



EMBL Embraces the Whole “Biodata Life Cycle”



10+ PB of data/year (Exabyte-level data/year by 2026)



**Data coordination across sites
both internally & externally**

**Data-driven research & services
at EMBL**

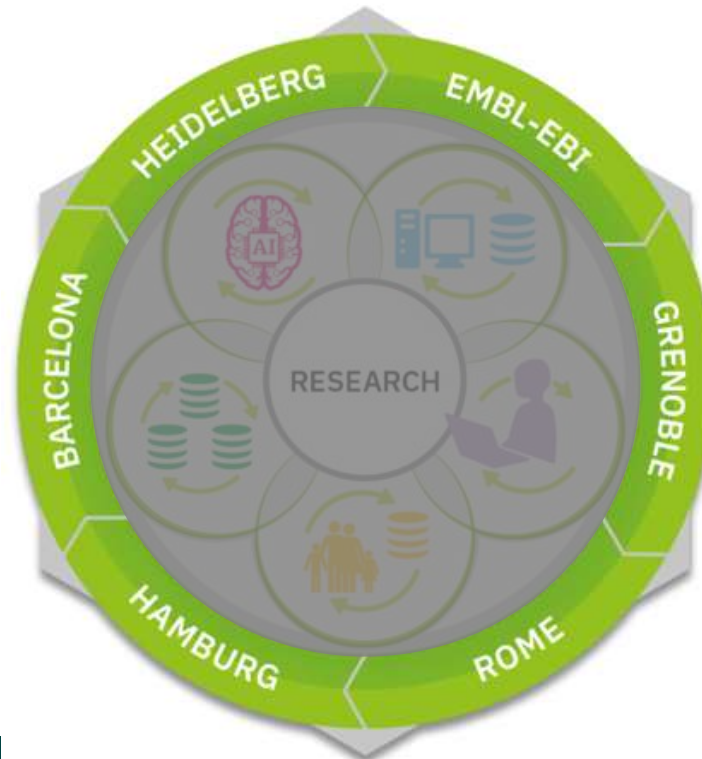
The Opportunity: EMBL's Unified Approach to Data Science

Use research algorithms and methods to extract knowledge from structured and unstructured biodata. Pursue research yielding insights from emerging properties of the big data.



The Opportunity: EMBL's Unified Approach to Data Science

Use research algorithms and methods to extract knowledge from structured and unstructured biodata. Pursue research yielding insights from emerging properties of the big data.



The Opportunity: EMBL's Unified Approach to Data Science

Areas EMBL will grow as part of the Unified Data Science Theme

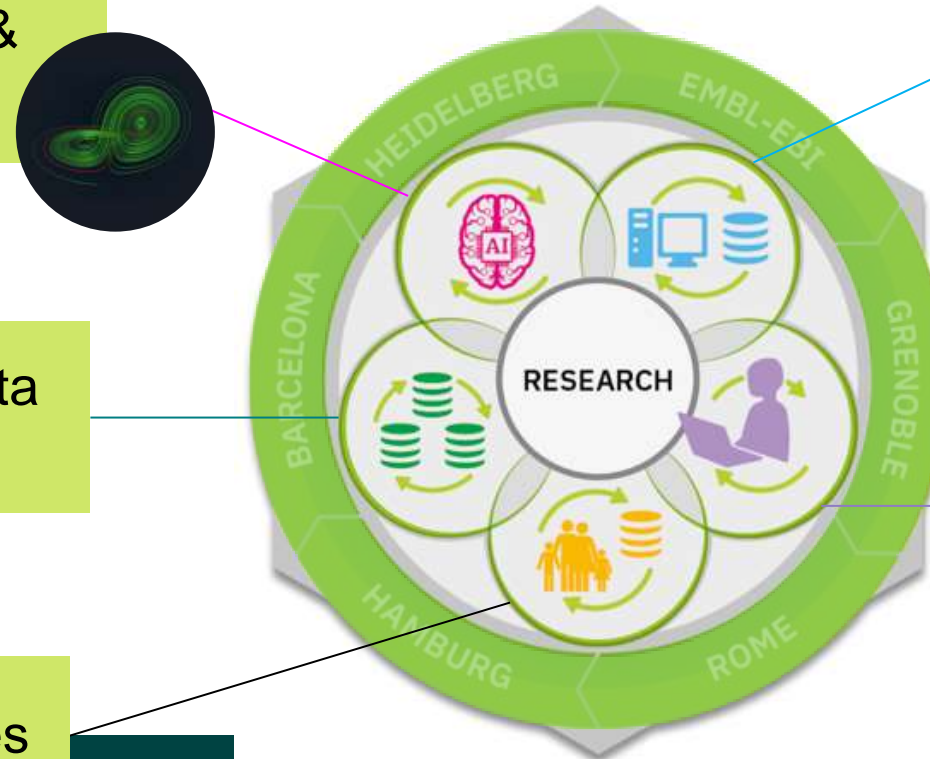
I. Data Science research & methods

II. Integrated research data management

IV. Expand EMBL's Services

III. IT infrastructure solutions

V. People: build the next generation of data scientists



I. Data Science Research and Methods



Data science research needs in the new EMBL programme:

Unsupervised, semi-supervised, and self-supervised learning to tackle lack of training data

Reference dataset generation, and labelling (data curation)

Data integration: e.g. environmental, omics, and imaging data (factor analysis, dimension reduction, interpretable machine learning)

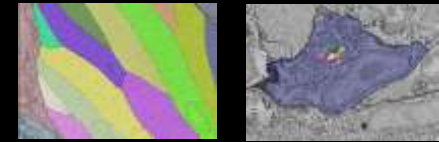
Visualisation, multi-scale image data browsing, streaming

**Electron
Microscopy:**

**11,500
cells**



**11,500
nuclei**



**Spatial gene
expression
resolved**

**Leading edge technology:
whole organism segmentation of *Platynereis***

I. Data Science Research and Methods



Data science research needs in the new EMBL programme:

Unsupervised, semi-supervised, and self-supervised learning to tackle lack of training data

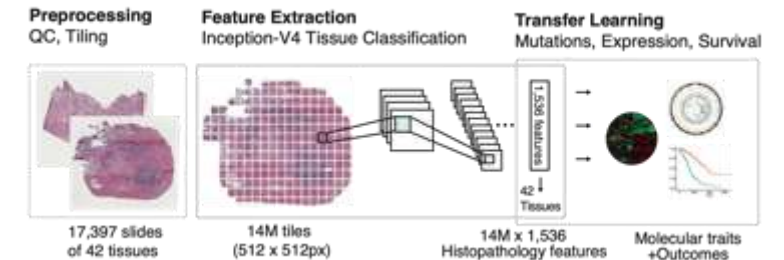
Reference dataset generation, and labelling (data curation)

Data integration: e.g. environmental, omics, and imaging data (factor analysis, dimension reduction, interpretable machine learning)

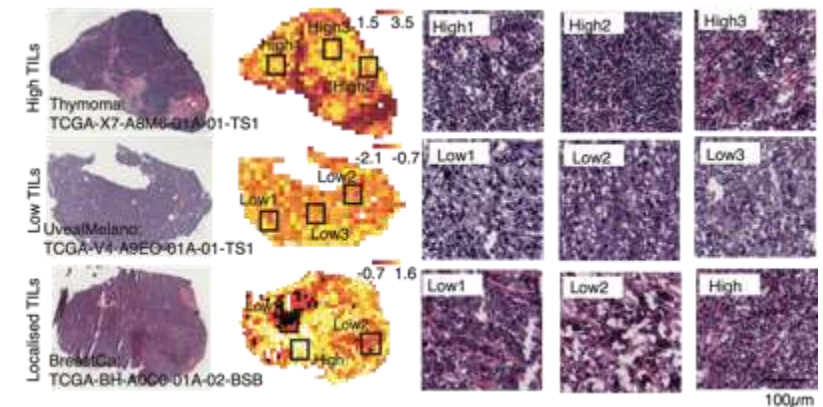
Visualization, multi-scale image data browsing, streaming

Derive causal and mechanistic insights (causal models, e.g. to dissect the molecular basis of GxE)

Histopathological patterns of mutations...



... and transcriptomic changes

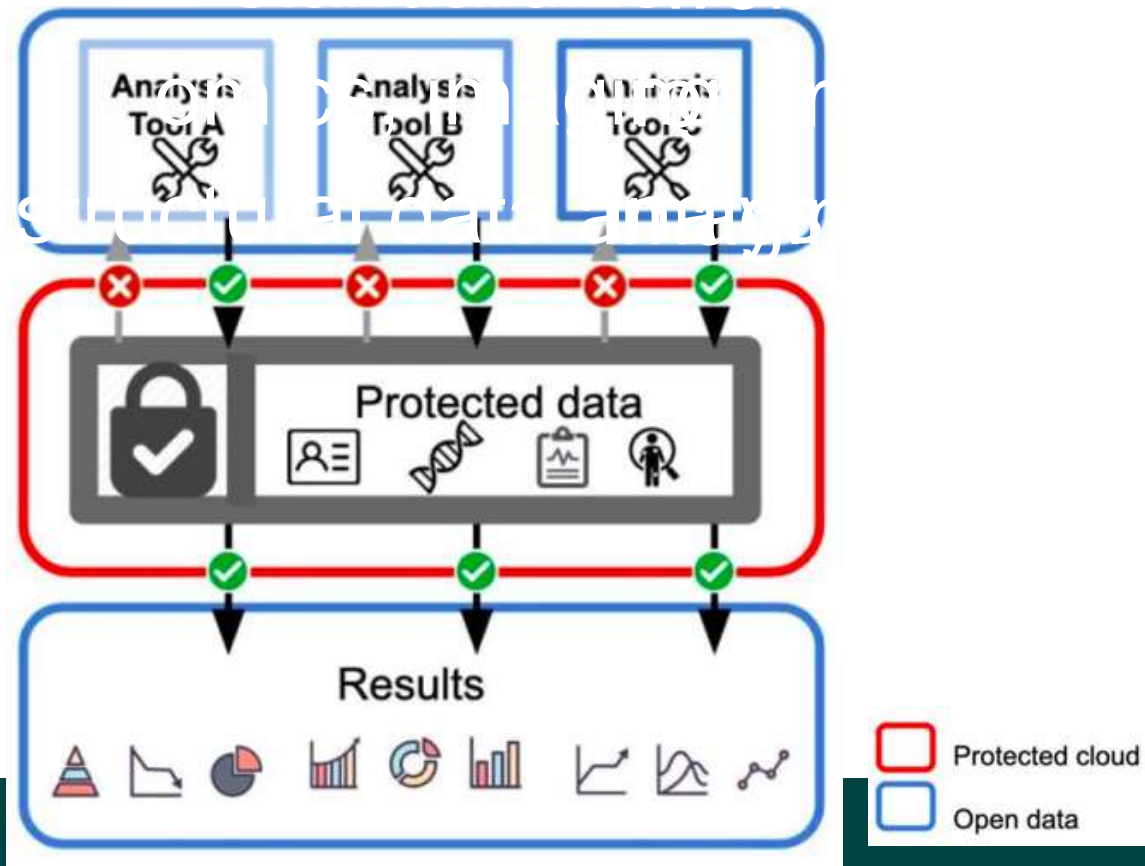


Deep transfer learning:
integrate histopathology, mutations, transcriptomes

I. Data Science Research and Methods: Democratising Resource Access



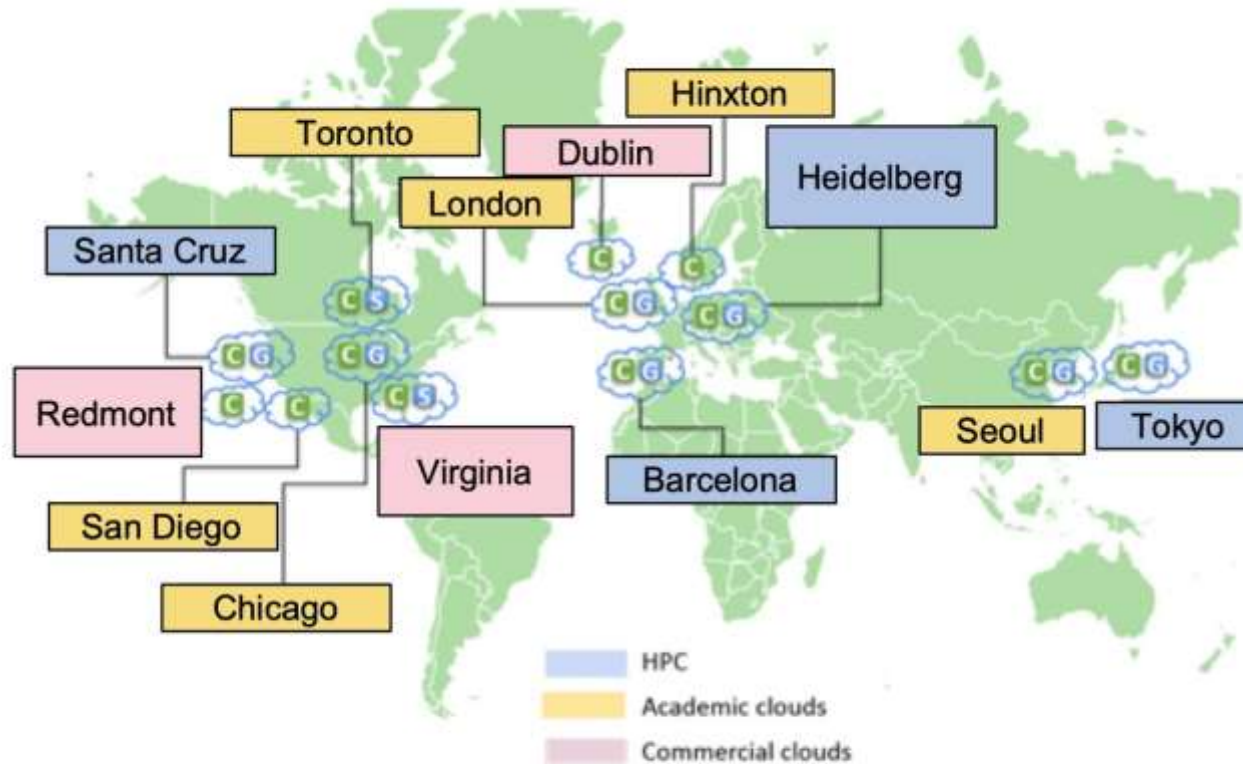
Containerized workflows:



Shared workflows for recurring analyses,
for various EMBL use cases

Method dissemination,
reusability & reproducibility

I. Methods in Data Science: Democratising Cancer Biology via Containerised Workflows



38 tumor types
2,600 cancer genomes
1300 participants

Rodriguez-Martin *et al. Nat Genet* 2020
Cortes-Ciriano *et al. Nat Genet* 2020
Yuan *et al. Nat Genet* 2020
Akdemir *et al. Nat Genet* 2020
Zapatka *et al. Nat Genet* 2020
Yakneen *et al. Nat Biotechnol* 2020
Li *et al. Nature* 2020
Gerstung *et al. Nature* 2020
Phillips *et al. Nature* 2020
Rheinbay *et al. Nature* 2020
Alexandrov *et al. Nature* 2020
Calabrese *et al. Nature* 2020
PCAWG Consortium, *Nature* 2020

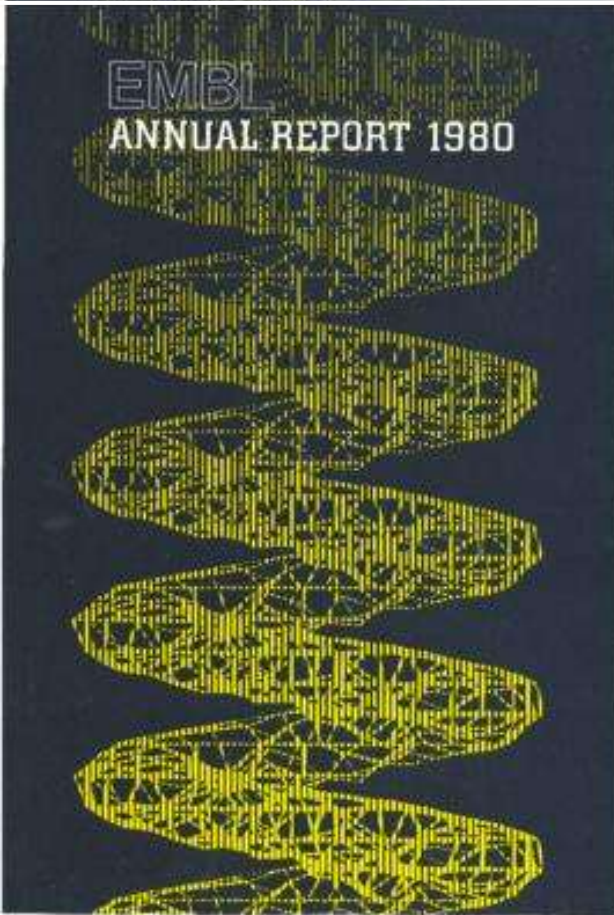
Tumours have 4-5 driver mutations.
Intergenic cancer drivers in 25% of patients.

Study foundation: shared “containerised” workflows

II. Integrated Research Data Management: Brief Look into EMBL History - Open Data & Standards




User manual: EMBL nucleotide sequence database, first release:



This manual and the database it accompanies may be copied and redistributed freely, without advance permission, provided that this statement is reproduced with each copy.

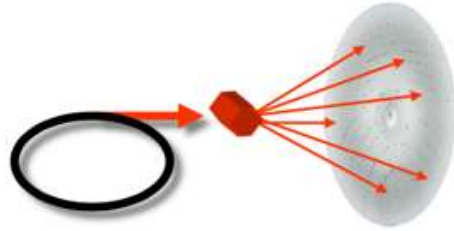
FAIR Principles: Findable. Accessible. Interoperable. Reusable.

 © 2001 Nature Publishing Group <http://genetics.nature.com> *commentary*

Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma¹, Pascal Hingamp², John Quackenbush³, Gavin Sherlock⁴, Paul Spellman⁵, Chris Stoeckert⁶, John Aach⁷, Wilhelm Ansorge⁸, Catherine A. Ball⁴, Helen C. Causton⁹, Terry Gaasterland¹⁰, Patrick Glenisson¹¹, Frank C.P. Holstege¹², Irene F. Kim⁴, Victor Markowitz¹³, John C. Matese⁴, Helen Parkinson¹, Alan Robinson¹, Ugis Sarkans¹, Steffen Schulze-Kremer¹⁴, Jason Stewart¹⁵, Ronald Taylor¹⁶, Jaak Vilo¹ & Martin Vingron¹⁷

II. Integrated Research Data Management: Brief Look into EMBL History - Open Data & Standards

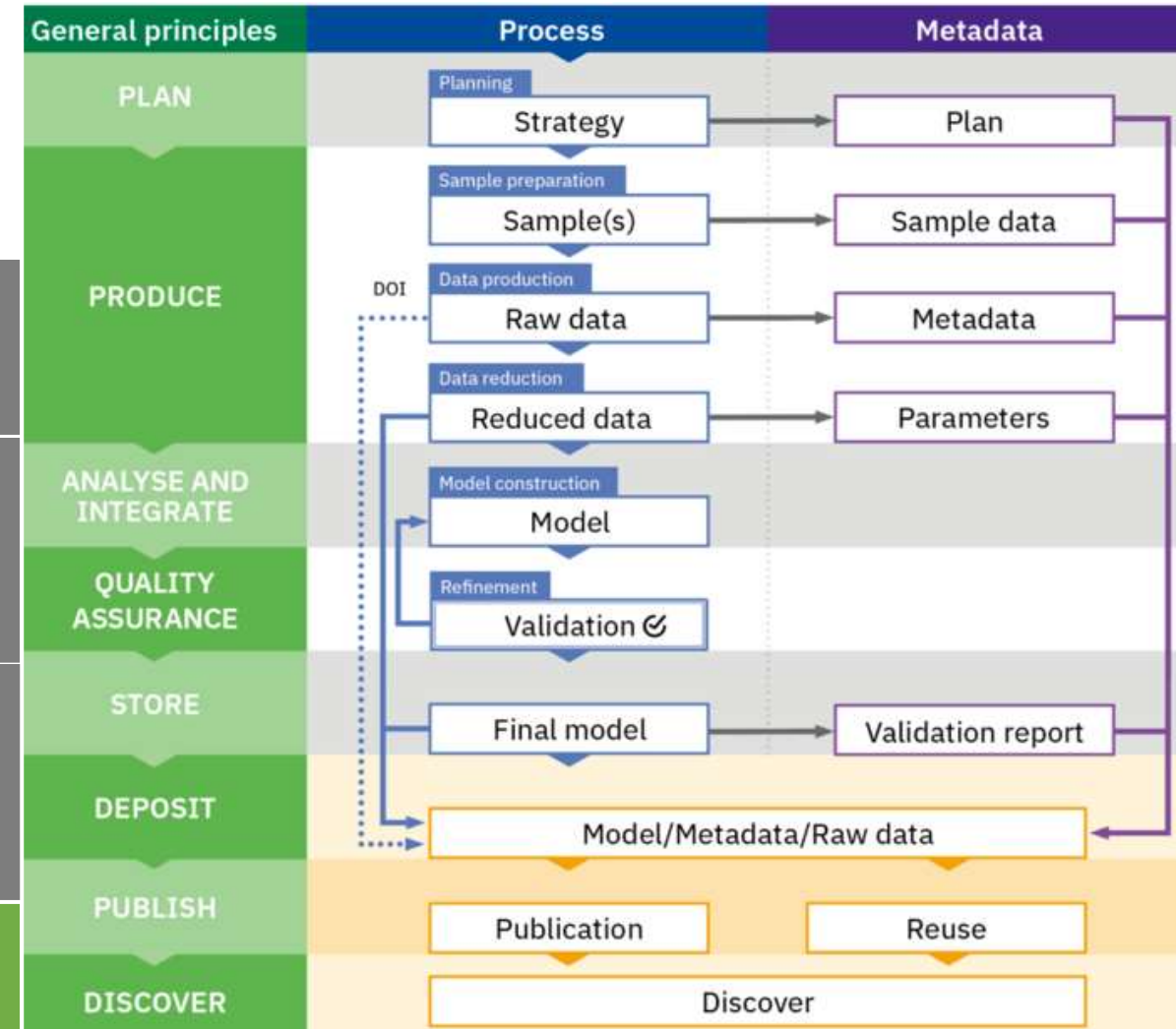


Controlled parameters and metadata (reproducibility and reuse): from reduced data to structural models

Data Science pilot study: record data from protein production and crystallisation conditions to raw X-ray data – *i.e.* track all data

Roll out to other EMBL use cases, including the “Bioimaging Revolution” (where data is historically less shared than in omics)

Promote Open Science



III. IT Infrastructure Enabling Data Science at EMBL



Large-scale distributed data analysis

(Pan-Cancer, Human Cell Atlas)

EMBL-EBI



Heidelberg



“EMBL Science Cloud”

(3D Cloud, de.NBI/Elixir-DE associated cloud)



Embassy Cloud

EMBL Science Cloud

When sensible, move data to Clouds

Provision GPUs for AI tools

Explore advanced data storage (object store)

III. European Open Science



"We are creating the European Open Science Cloud. ...with a pool of information leading to a web of research insight."



Ursula von der Leyen
World Economic Forum - Davos
22 January 2020



Initial life-science use case: Pan-Cancer



EOSC-Life
Implementation project



COVID-19 Data Portal

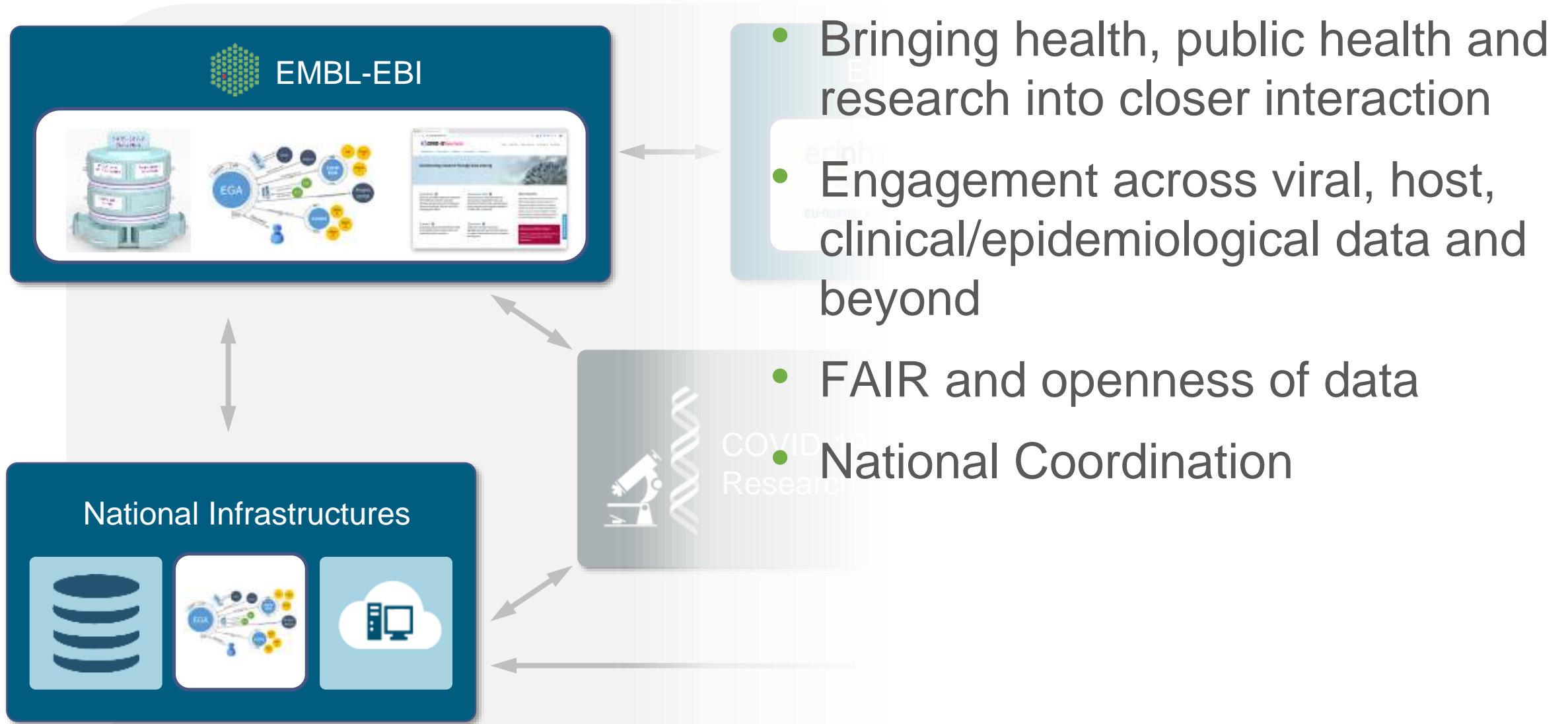
European COVID-19 Data Platform

- Open and rapid access to data, tools and workflows
- Built upon open standards
- Global data coverage and global access
- Enabling diverse research to fight COVID-19

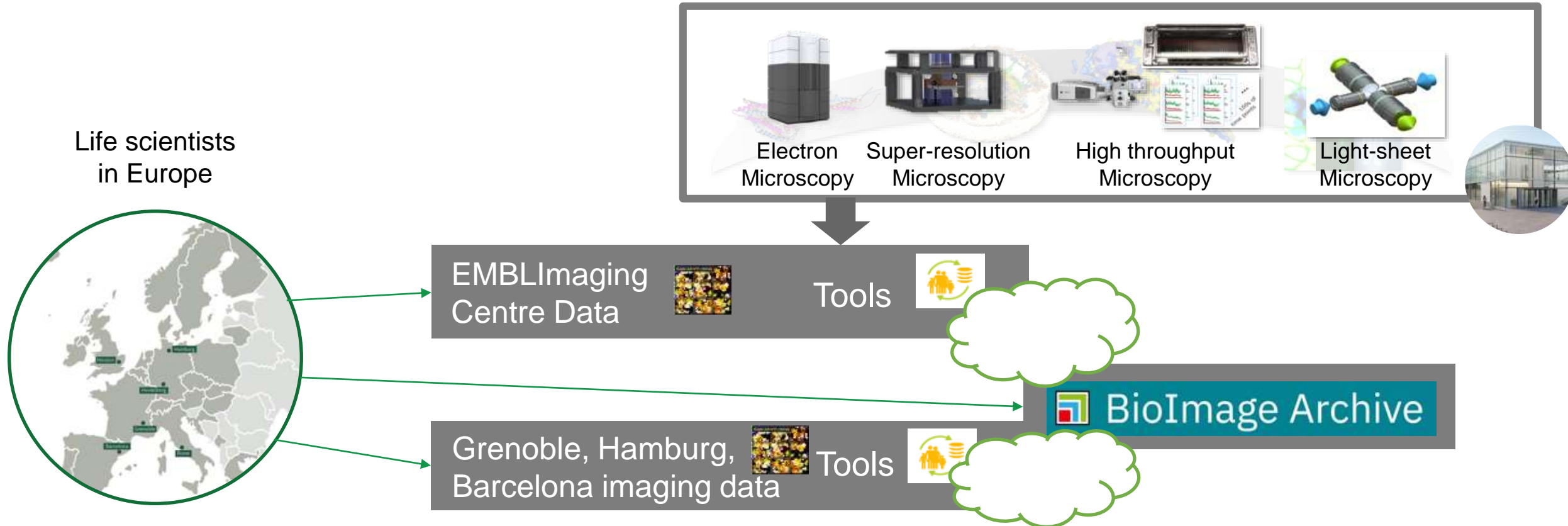


<https://www.covid19dataportal.org/>

Connecting at the national level



IV. Expand EMBL Services: Enabling the Bioimaging Revolution



Empower Service Users

IV. Expand EMBL Services: Enabling the Bioimaging Revolution

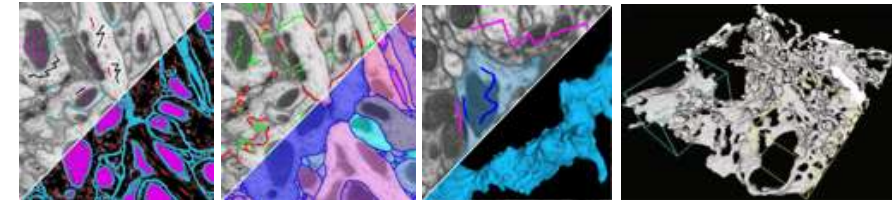
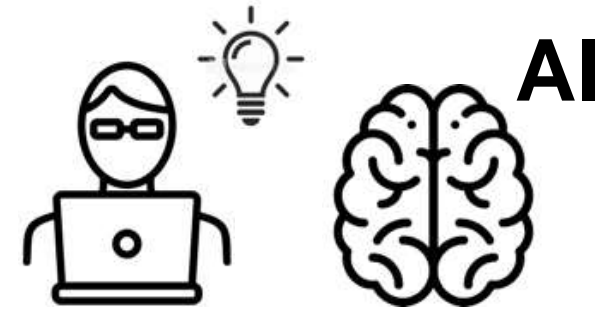


Foster image data analysis,
using advanced machine learning techniques

Provide User Support

Methods to promote multi-modal
data integration

Connect software with experimental platforms



IV. Expand EMBL services: Sustainable Research Software



Aim to identify EMBL software of strategic value
needing to be maintained as a service

Downloaded
286,000x
DESeq2
for gene expression
analysis

Downloaded
60,000x
ilastik
for image
segmentation

Used
10,000x
webPRANK
a phylogeny aware
multiple-sequence aligner

Sustained tools to receive common “EMBL brand”

V. People: Build the Next Generation of Data Scientists



New Data Science personnel required:
pan-EMBL support team(s), data stewards, data deposition
brokerage & representation

Attract best talent in strategically important areas

Internal & External Training (EICAT):
Develop more hybrid scientists

Opportunity for co-funding via new ARISE accelerator
programme for Research Infrastructure Scientists



e l l i s
h e a l t h

pan-European nonprofit organization for promoting AI

EMBL's Unified Approach to Data Science: European impact

Cutting edge data science research

Facilitate collaborations with local partners,
and more generally with member states

Expand EMBL's services

Framework to develop European data science careers

Continue as a catalyst for European Open Science



Stay in touch

www.ebi.ac.uk

Twitter: @emblem
Facebook: EMBLEBI
LinkedIn: /company/ebi
YouTube: EMBLMedia

