

Data submissions and archiving raw data in life sciences
a Pilot with Proteomics Data

Rafael C Jimenez

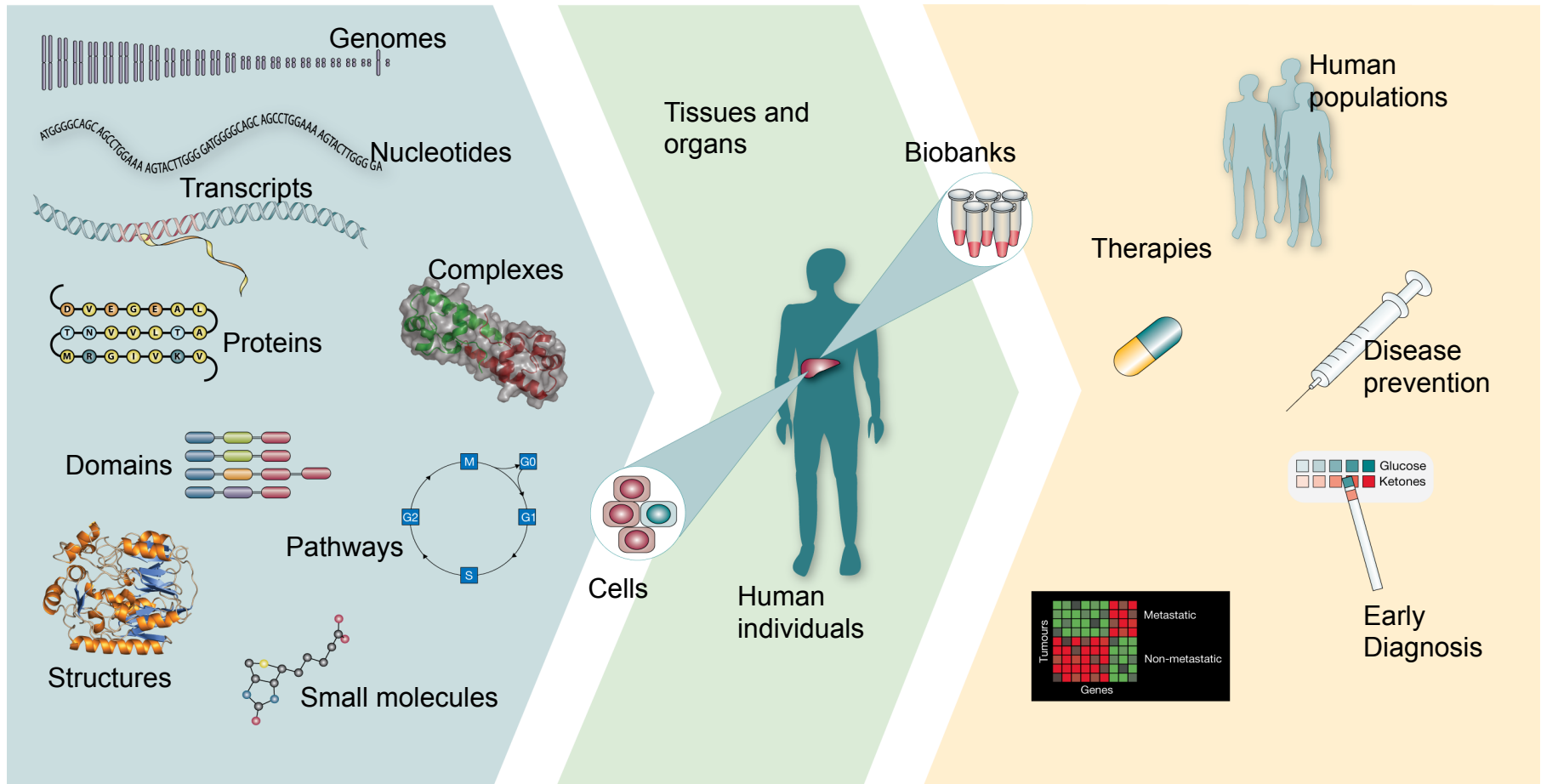
ELIXIR CTO



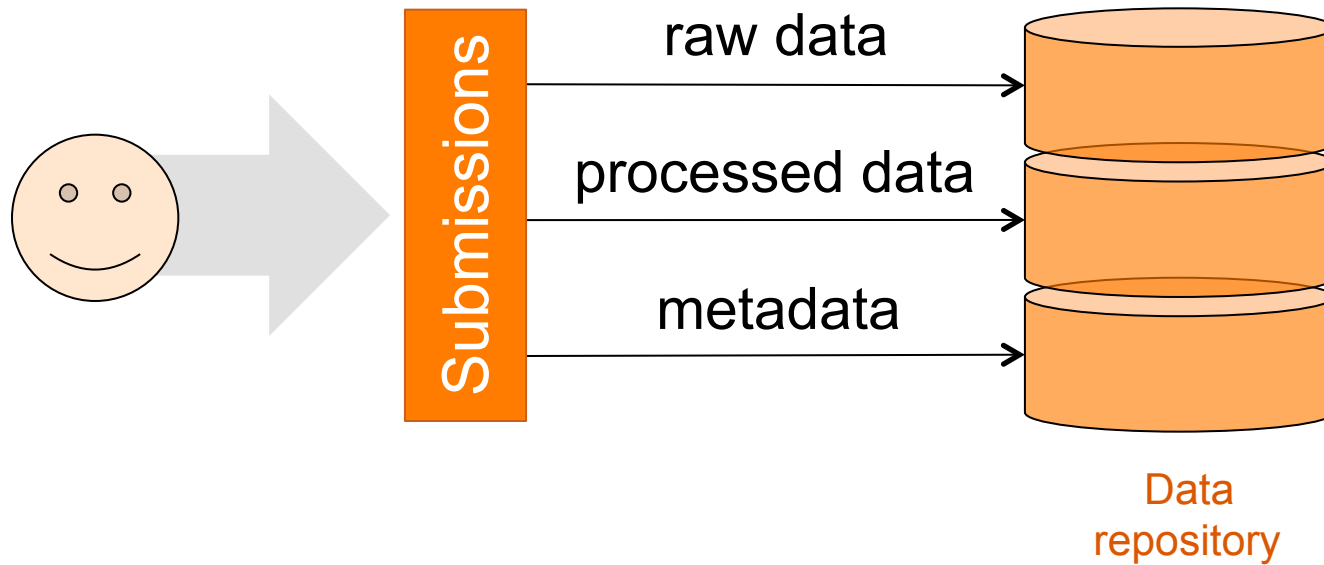
EUDAT User Forum, 23 April 2014

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

Data in biomedical sciences

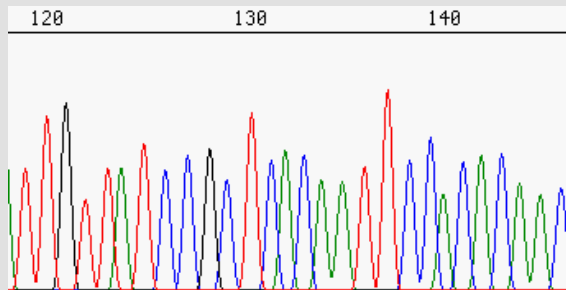


Data submissions



Data types examples

Raw data

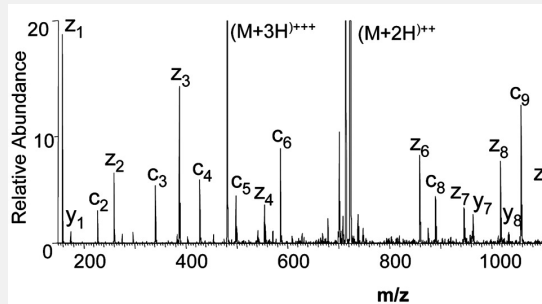


Process data

TTGTTATCCG...

Metadata

DNA
Human
Liver
Mitochondria
W. Smith
...



LPISASHSSK...

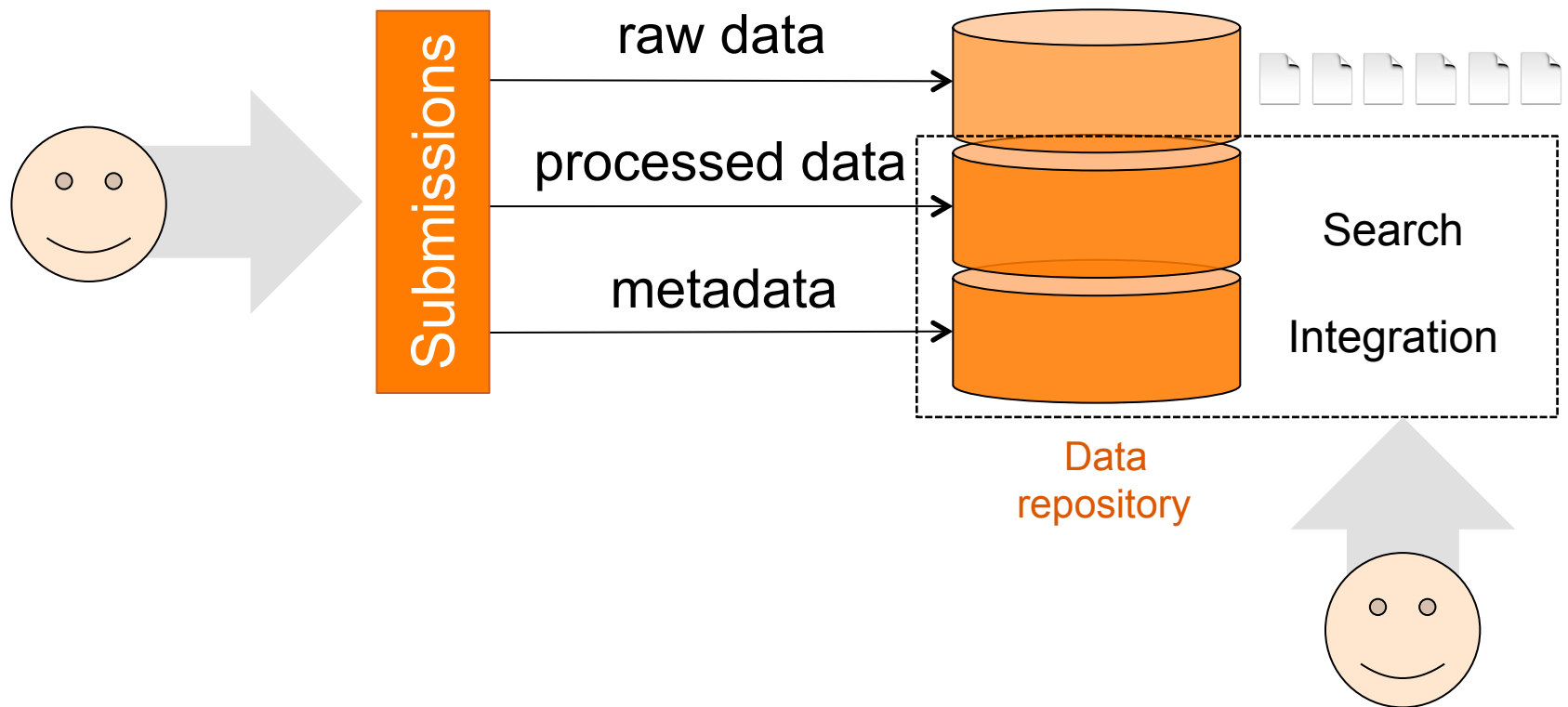
Peptide
Mouse
Heart
Nucleus
J. Heinz
...

...

...

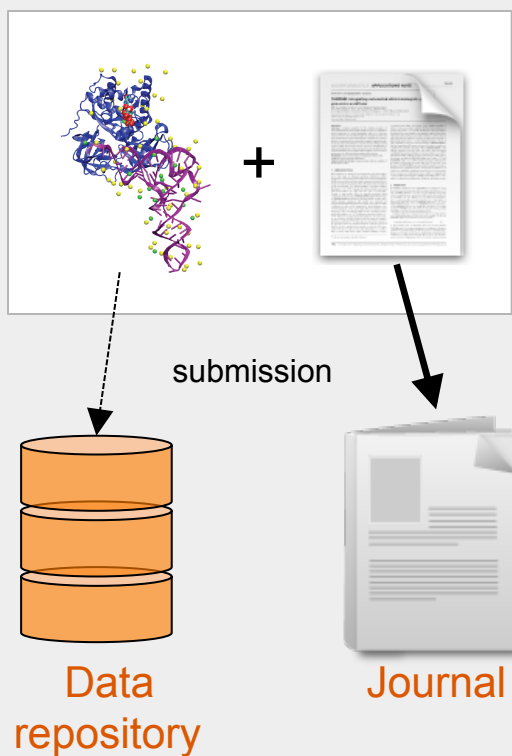
...

Data submissions

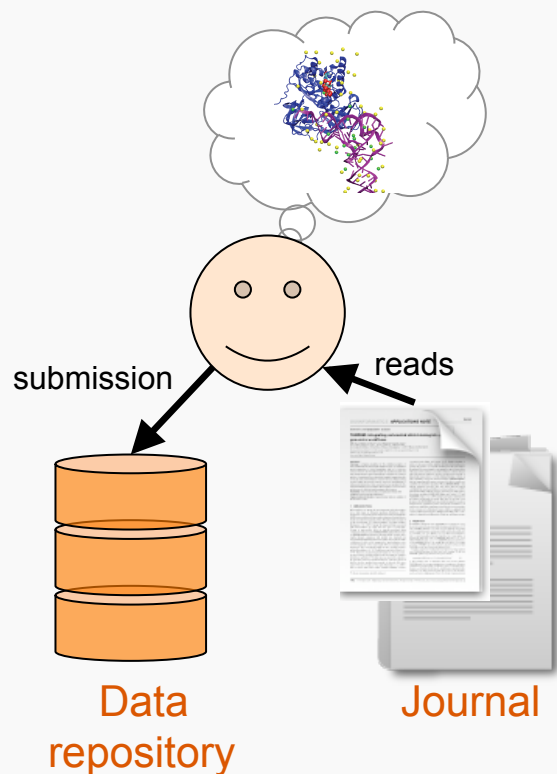


Public data submissions

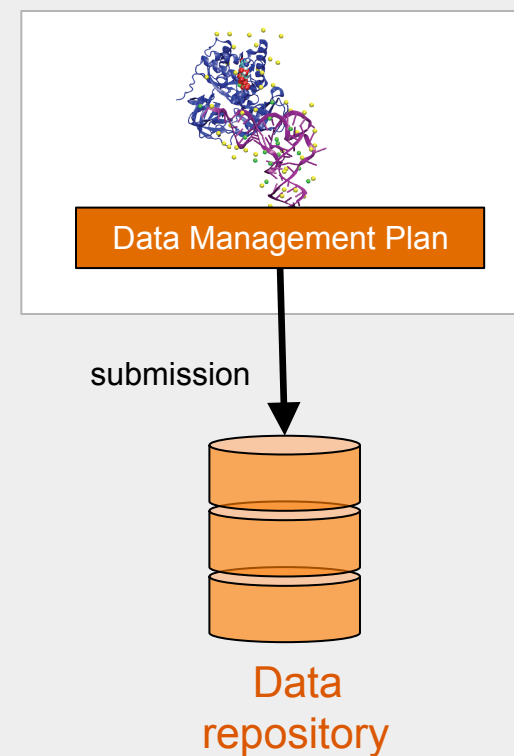
Journal request



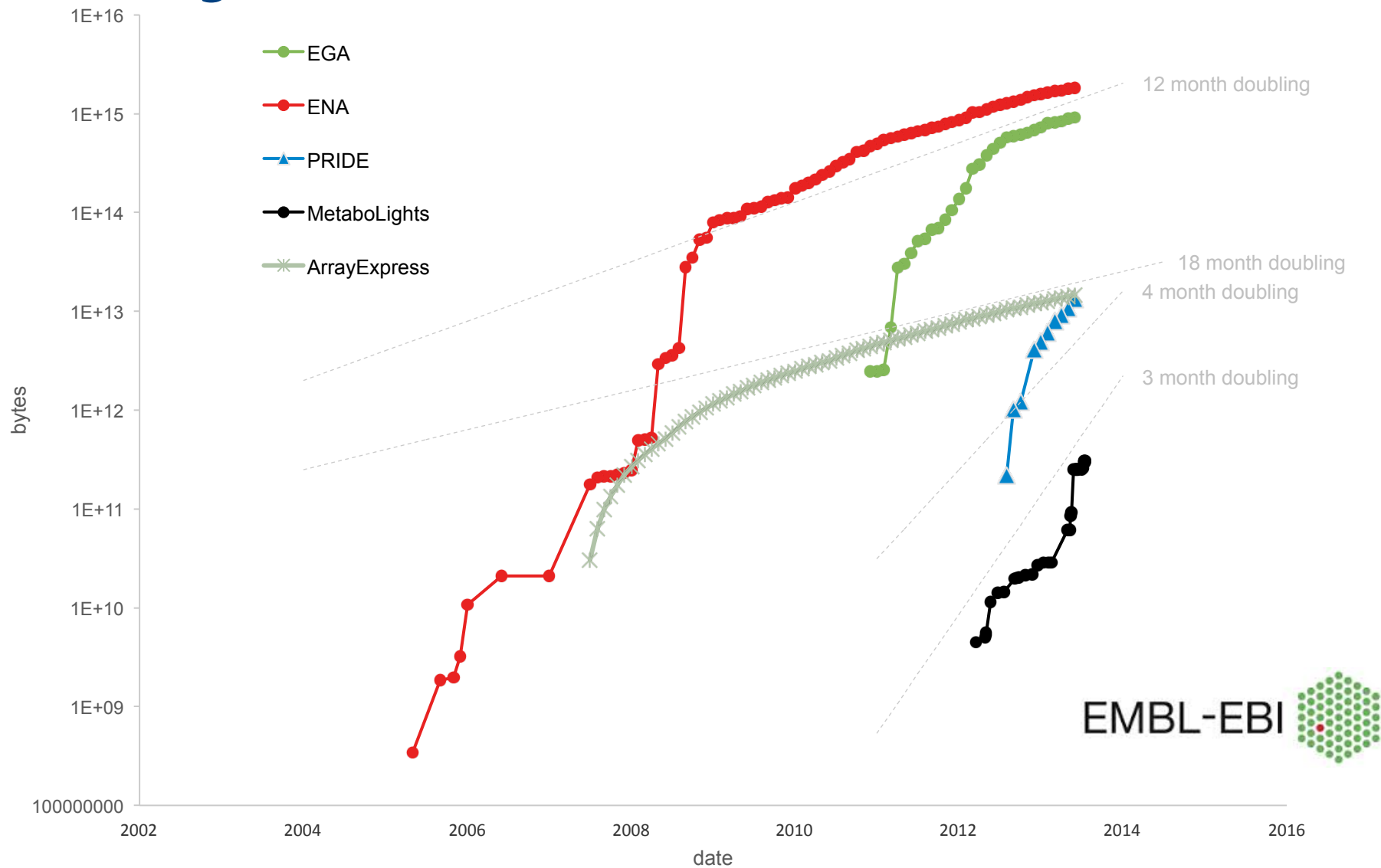
Curator



Data management



Growing data

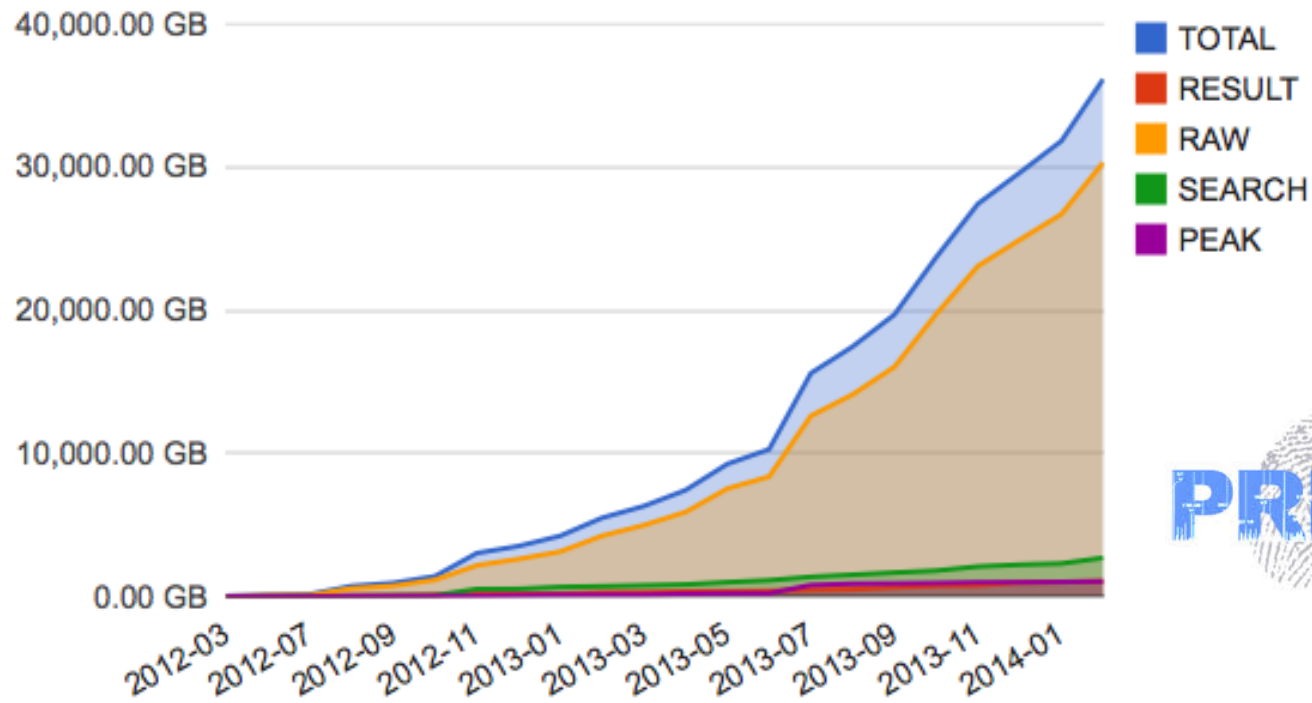


EMBL-EBI



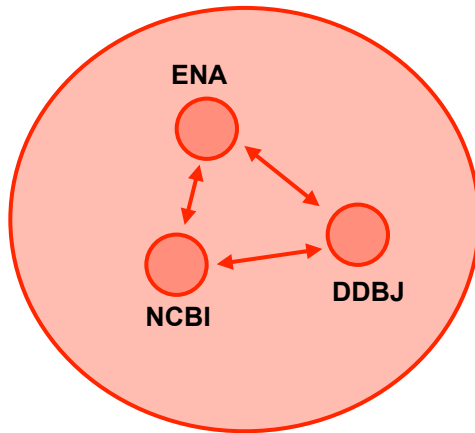
Proteomics data in PRIDE

~85% raw data



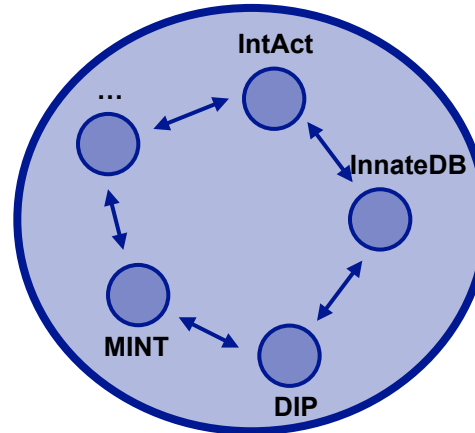
Collaboration among life sciences data providers

Replication



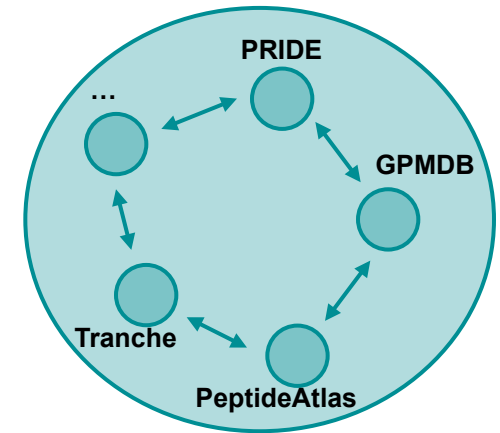
Nucleotide sequences
INSDC

Federation



Molecular interactions
IMEx

Specialization

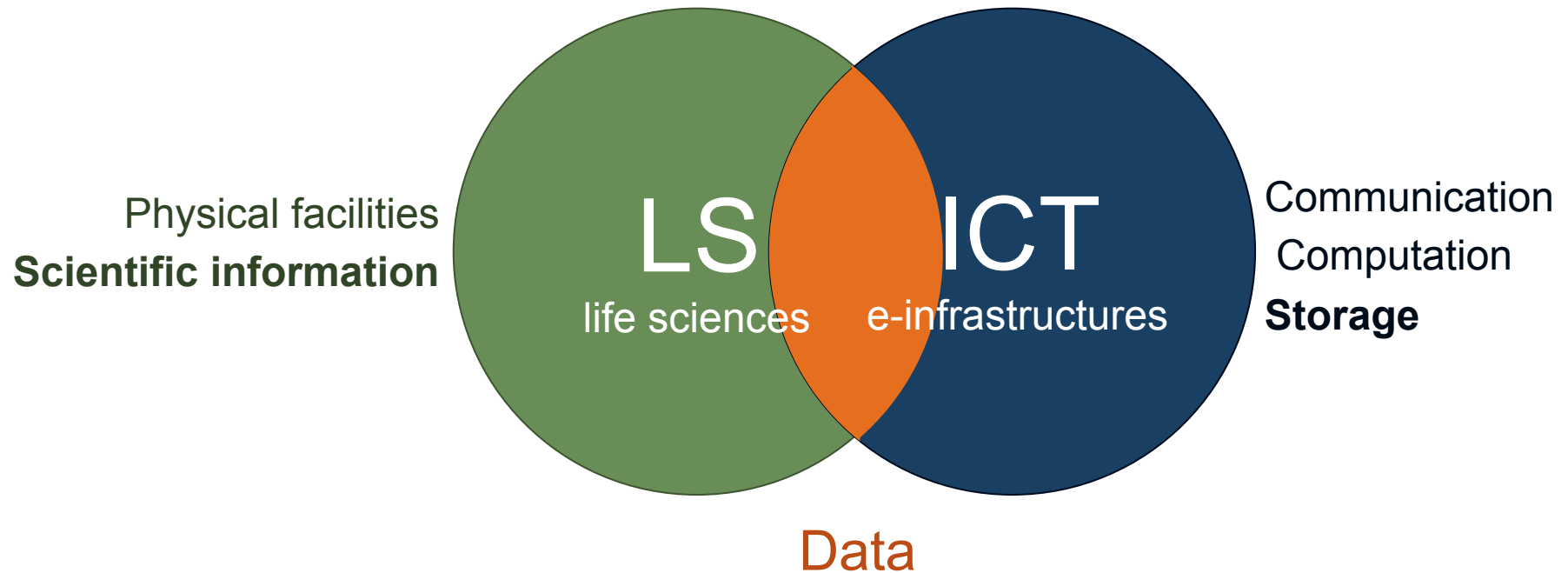


Protein identifications
ProteomeXchange

Managing life sciences data is a bigger task than any individual organisation can take on alone

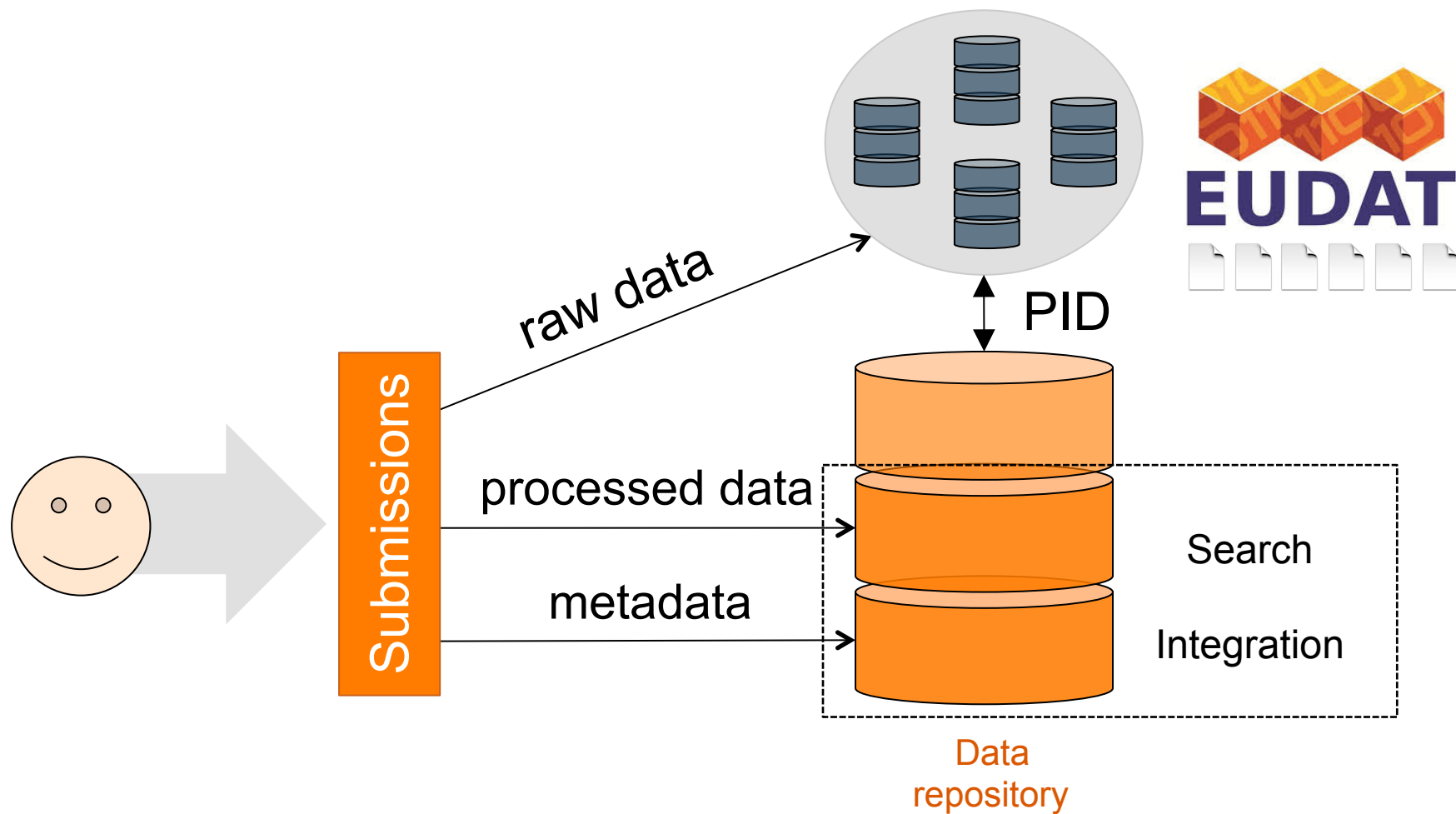
Research infrastructures

Facilitate research



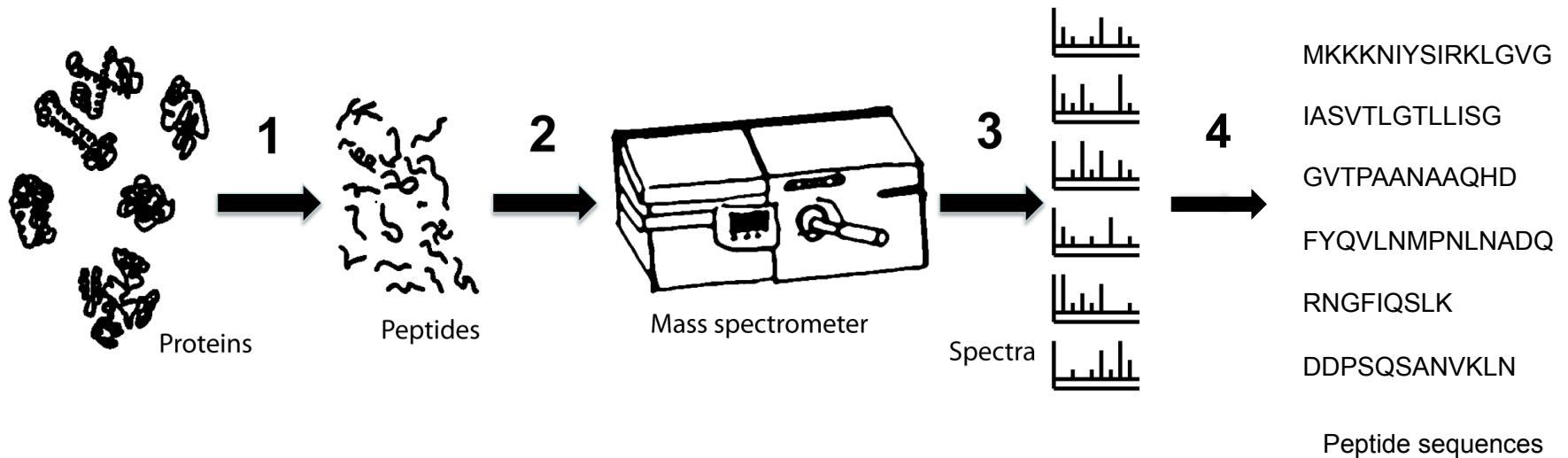
Collaboration proposal

Delegation of raw data



a Pilot with Proteomics Data

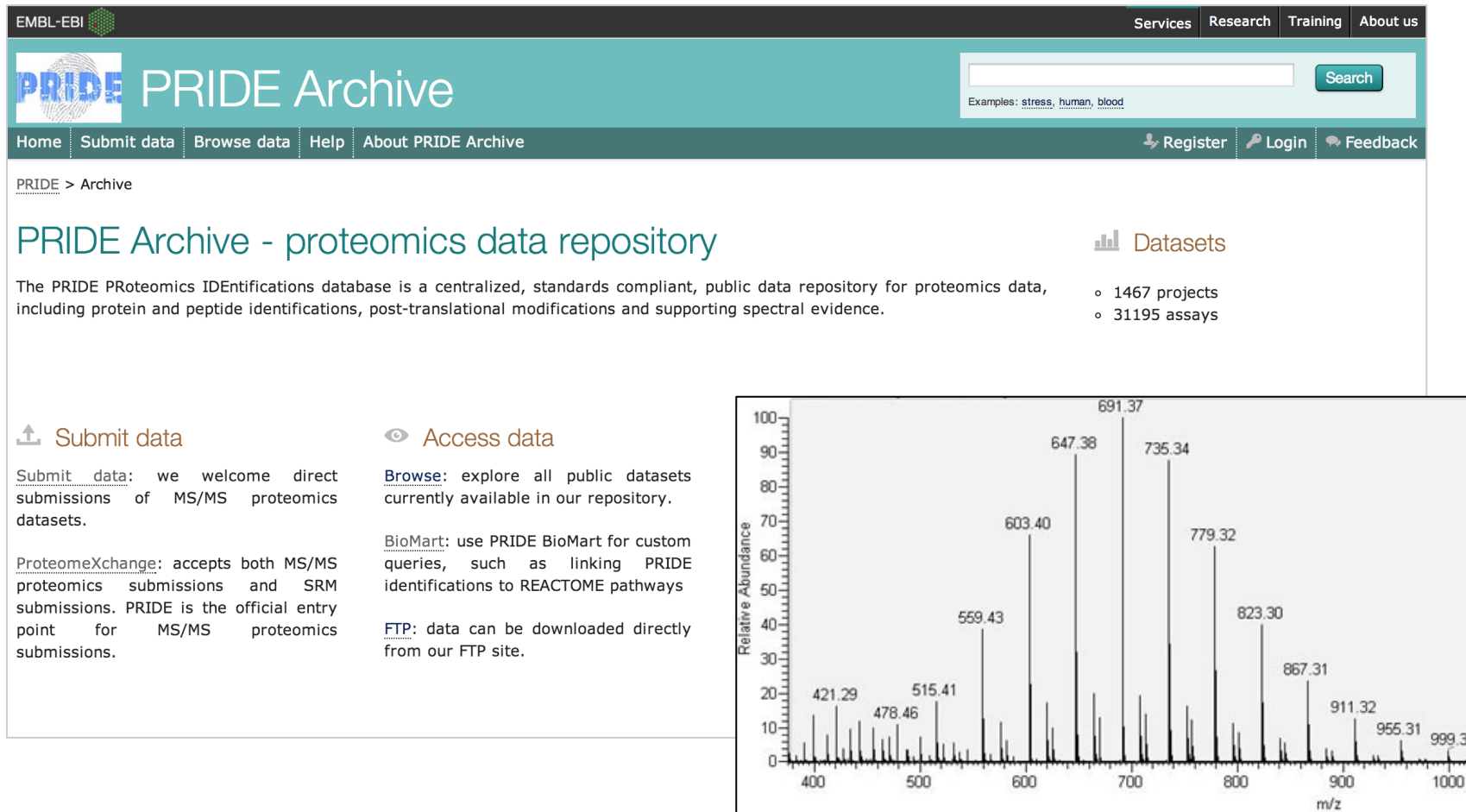
Overview of shotgun proteomics data production



Noble WS, MacCoss MJ (2012) Computational and Statistical Analysis of Protein Mass Spectrometry Data. PLoS Comput Biol 8(1): e1002296. doi:10.1371/journal.pcbi.1002296

<http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002296>

PRIDE (PRoteomics IDEntifications) database



EMBL-EBI

Services Research Training About us

PRIDE Archive

Examples: stress, human, blood

Home Submit data Browse data Help About PRIDE Archive

Register Login Feedback

PRIDE > Archive

PRIDE Archive - proteomics data repository

The PRIDE PRoteomics IDEntifications database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.

Datasets

- 1467 projects
- 31195 assays

Submit data

Submit data: we welcome direct submissions of MS/MS proteomics datasets.

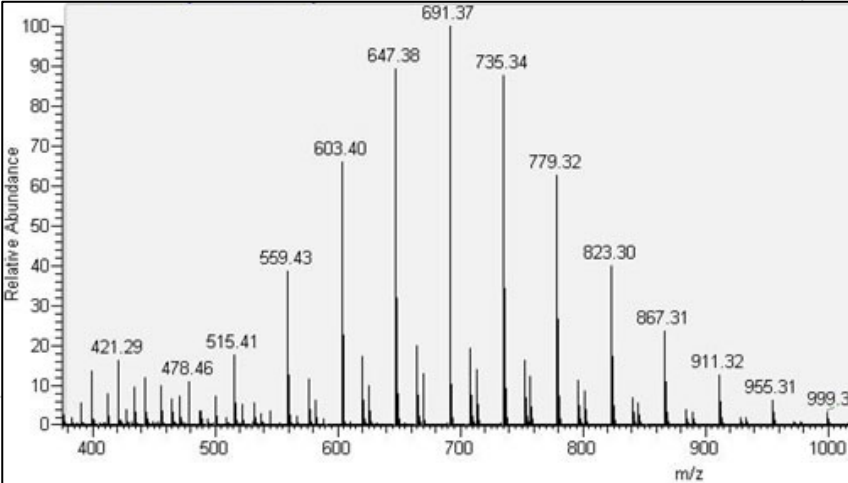
ProteomeXchange: accepts both MS/MS proteomics submissions and SRM submissions. PRIDE is the official entry point for MS/MS proteomics submissions.

Access data

Browse: explore all public datasets currently available in our repository.

BioMart: use PRIDE BioMart for custom queries, such as linking PRIDE identifications to REACTOME pathways

FTP: data can be downloaded directly from our FTP site.



Mass spectrometry

821 submitted datasets by April 1st 2014



Origin:

152 USA
108 Germany
67 United Kingdom
53 Switzerland
48 Netherlands
42 China
42 Canada
41 France
36 Spain
33 Belgium
25 Australia
23 Sweden
17 Japan
16 Denmark
13 Norway
12 Finland
12 India
12 Taiwan
10 Italy
9 Republic of Korea
8 Austria
8 Ireland
8 Brazil
7 Singapore
5 Israel
5 Russia ...

Type:

273 PRIDE complete
501 PRIDE partial
47 PeptideAtlas/PASSEL complete

Submissions/year:

2012: 102
2013: 527
2014: 192

Access:

38.3% PRIDE public
5.3% PASSEL public
56% PRIDE private
0.4% PASSEL private

Top Species studied by at least 8 datasets:

381 *Homo sapiens*
100 *Mus musculus*
31 *Arabidopsis thaliana*
26 *Saccharomyces cerevisiae*
16 *Escherichia coli*
14 *Rattus norvegicus*
12 *Mycobacterium tuberculosis*
11 *Drosophila melanogaster*

~ 215 species in total

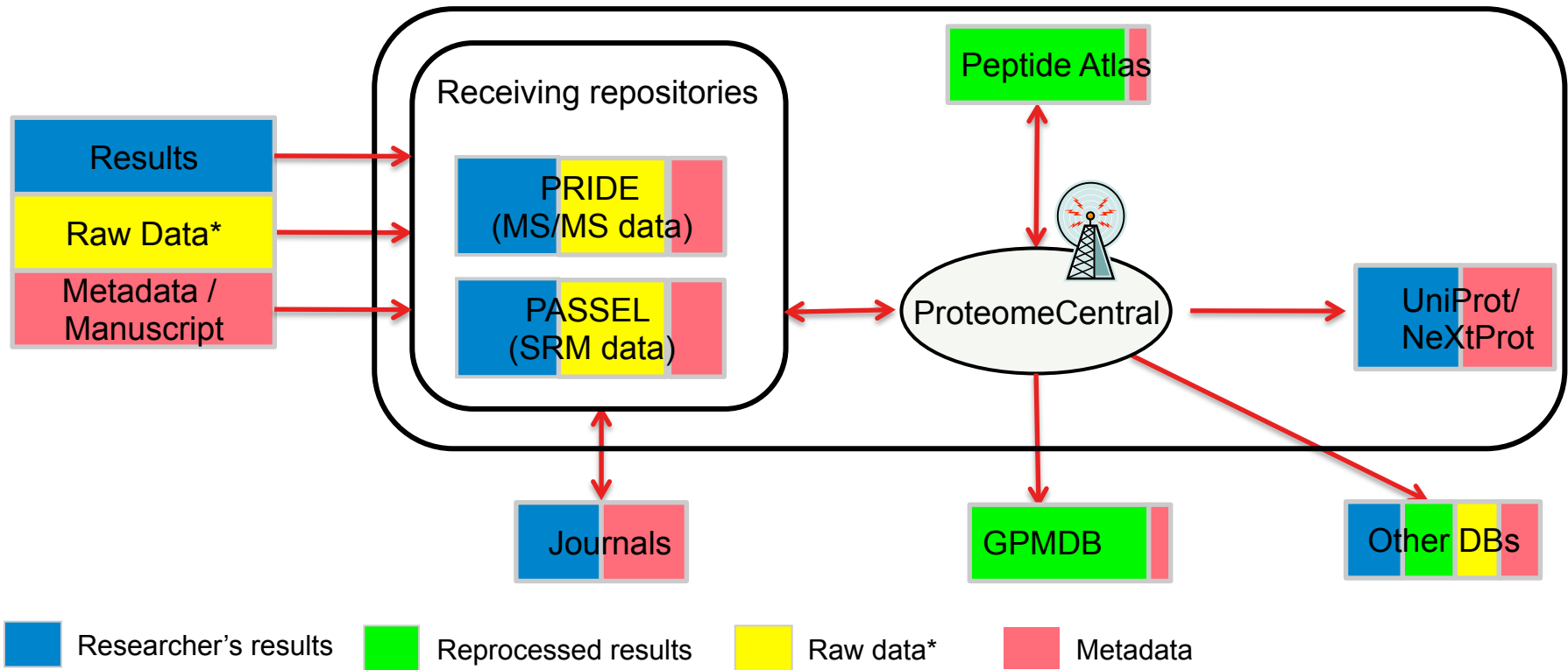
Data volume:

Total: >40 TB
Number of all files: >120,000
PXD000320-324: ~ 5 TB
PXD000065: ~ 1.4TB



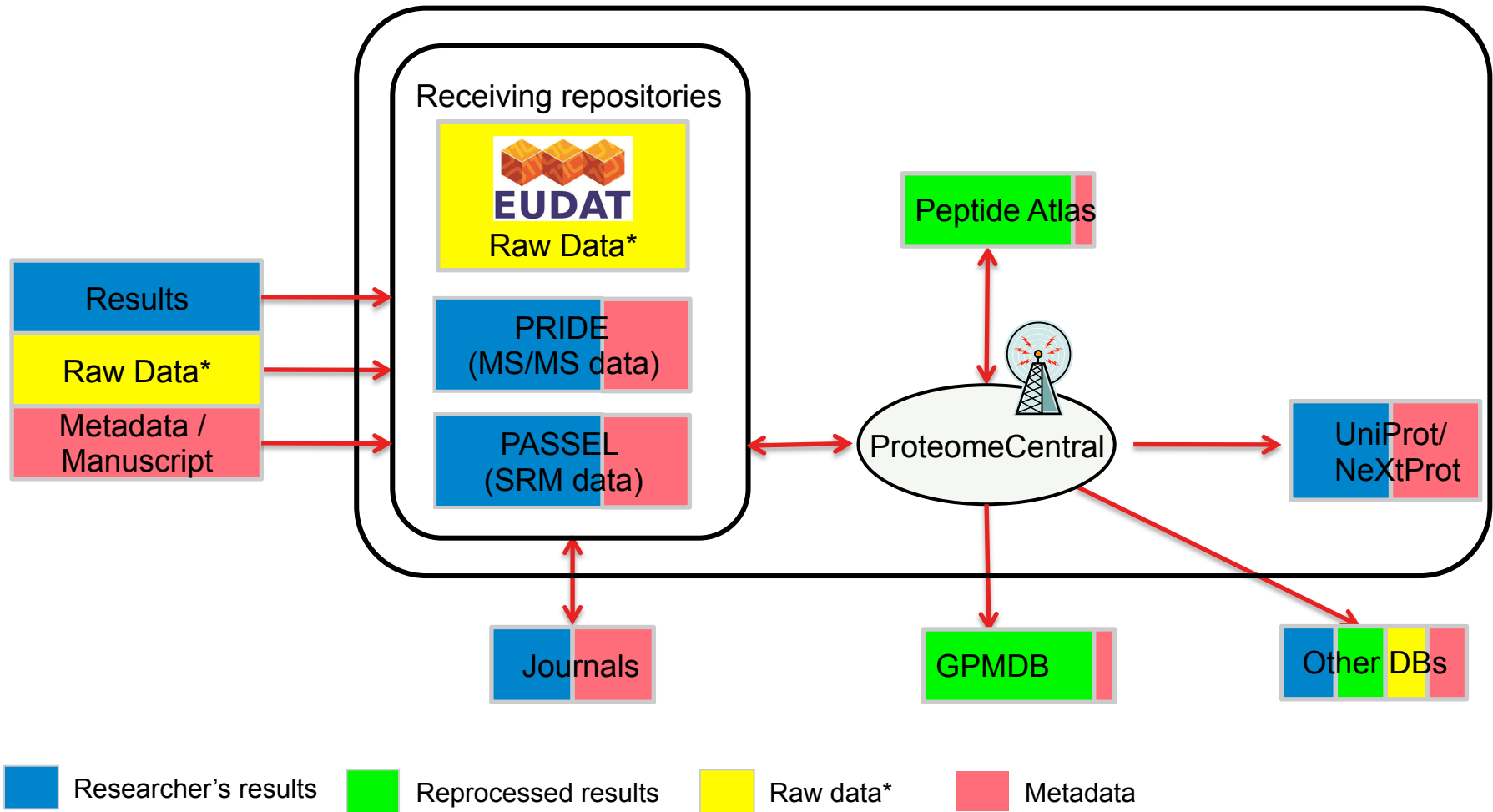
ProteomeXchange

- **Framework** to enable **standard data submission and dissemination pipelines** between the main existing proteomics resources.



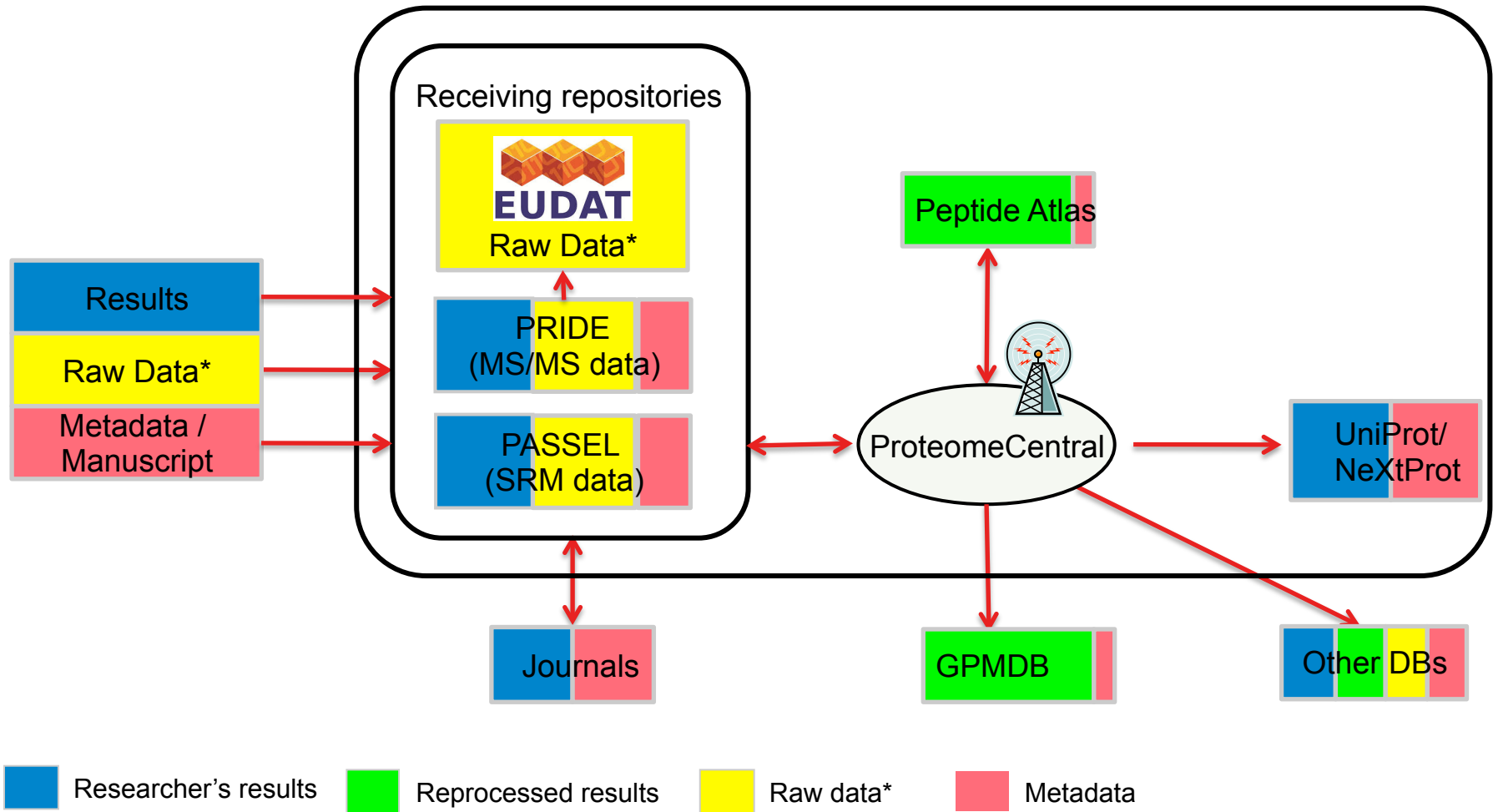
ProteomeXchange

Idea ...



ProteomeXchange

Pilot ...



Important points to explore

- EUDAT programmatic access for data submission
- Synchronization for data updates
- Keep unpublished datasets private until publication
- Keep up with the growth (40TB, doubling every 4 months)
- Use HPC facilities through EUDAT to process raw data
- Example for other repositories in life sciences

Acknowledgements

PRIDE team

Juan Antonio Vizcaino
Henning Hermjakob
Attila Csordas
Rui Wang
Florian Reisinger
Tobias Ternent
Noemi del Toro

Jose A. Dianes
Yasset Perez-Riverol
Qing-Wei Xu
Ilias Lavidas

PeptideAtlas Team

Eric Deutsch
Terry Farrah
Zhi Sun

Andrew R. Jones
Lennart Martens
Juan Pablo Albar
Martin Eisenacher
Gil Omenn

And many other PX partners and stakeholders



Preparing for the data deluge in life sciences

BioMedBridges workshop

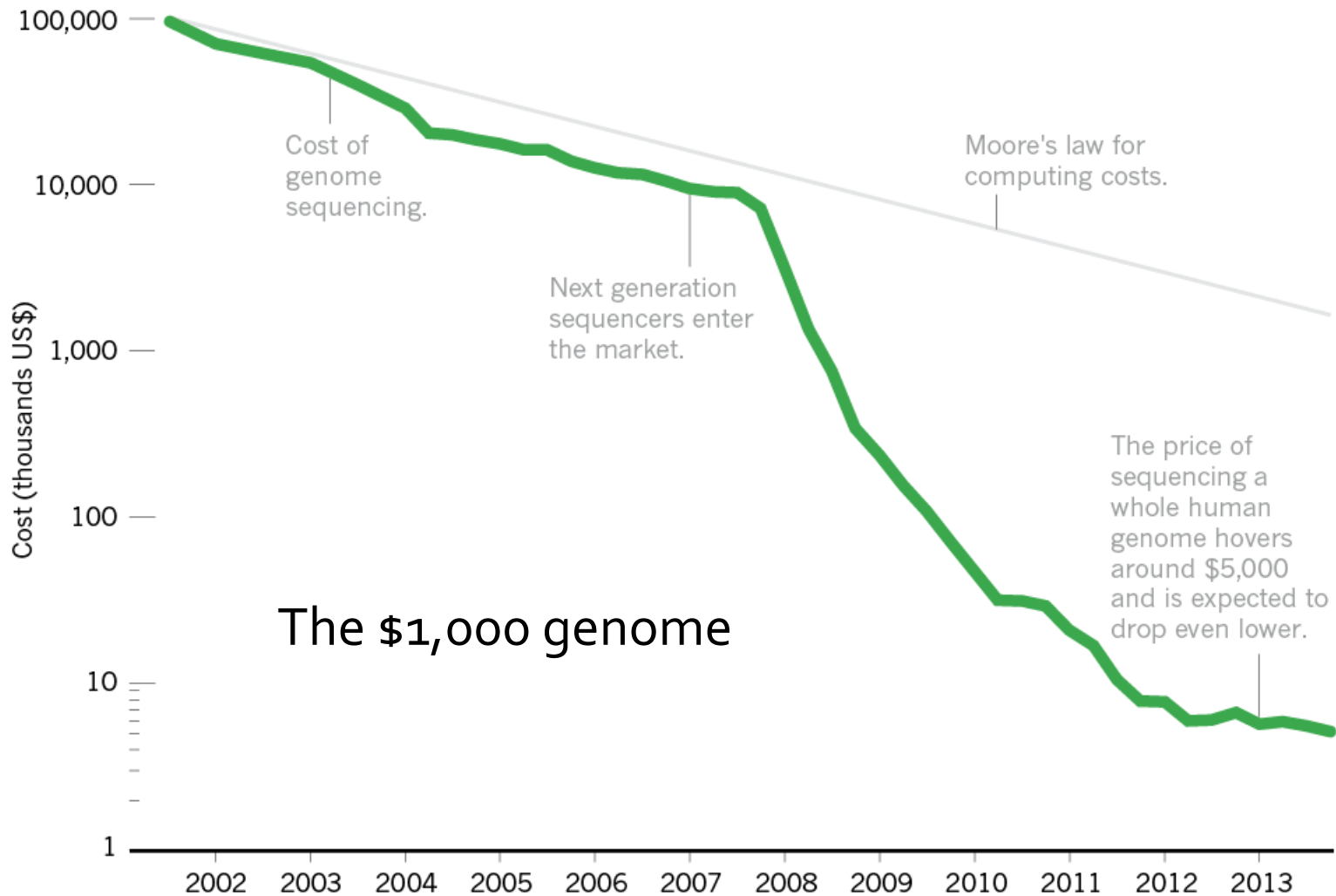
15-16 May 2014

Wellcome Trust Genome Campus

Hinxton, UK

Request attendance to Stephanie Suhr <ssuhr@ebi.ac.uk>

Production of data cheaper than storing data



Production of data faster than transferring data

		Network File Transfer		
24 hours		1 Gb	100 Mb	10 Mb
	DNA sequencing ~100 GB	~30 min	~5 hours	~2 days
	Mass spectrometry ~4 TB	~9 hour	~4 days	~5 weeks
	Microscopy ~4 TB	~9 hour	~4 days	~5 weeks

Preparing for the data deluge in life sciences

- Data too big to **store, exchange & compute**
 - Production of data **cheaper than storing** data
 - Production of data **faster than transferring** the data
 - Production of data **doubles faster than compute power**

BioMedBridges workshop

15-16 May 2014

Wellcome Trust Genome Campus

Hinxton, UK

Request attendance to Stephanie Suhr <ssuhr@ebi.ac.uk>



Thank you for your attention

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

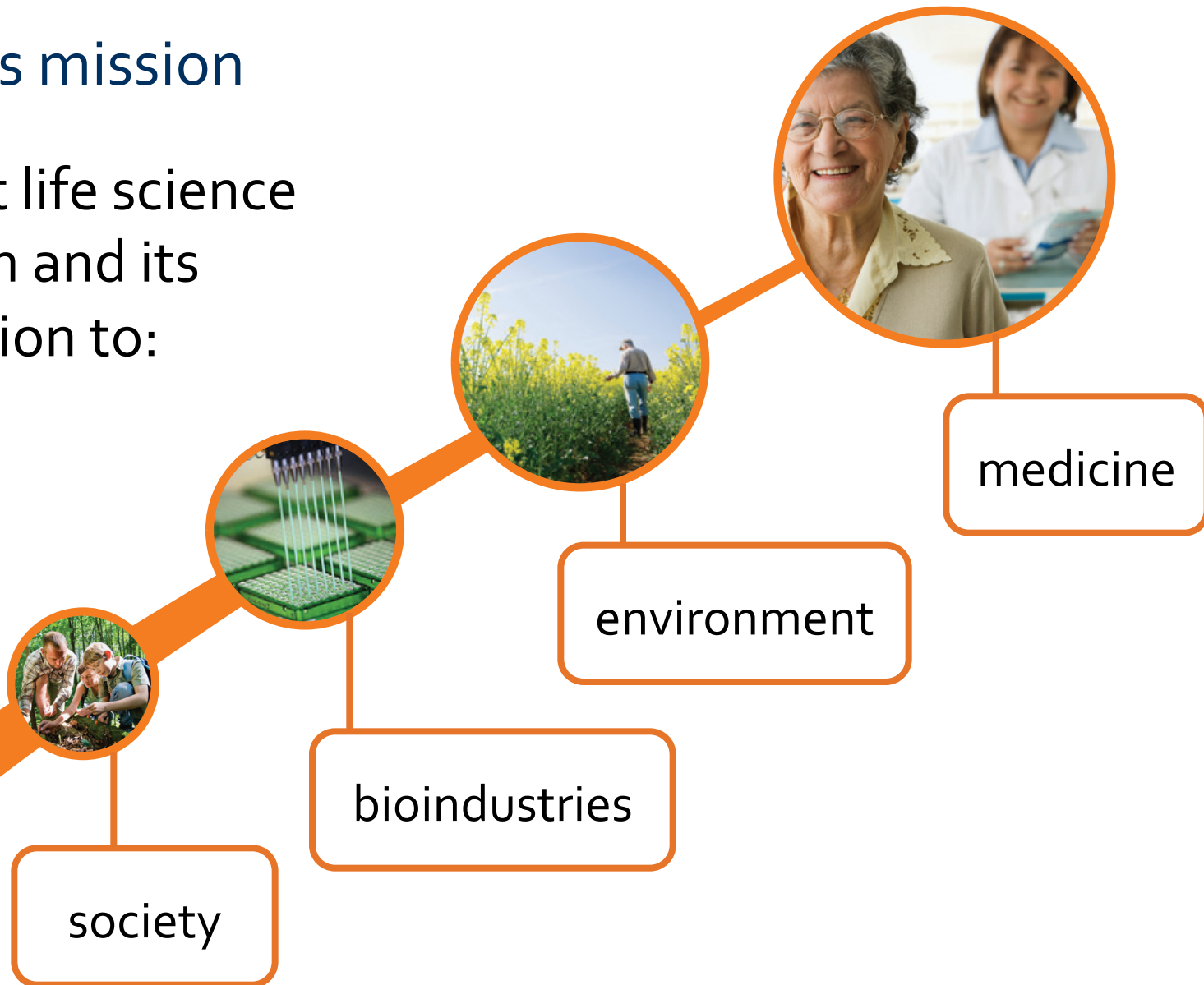


ELIXIR

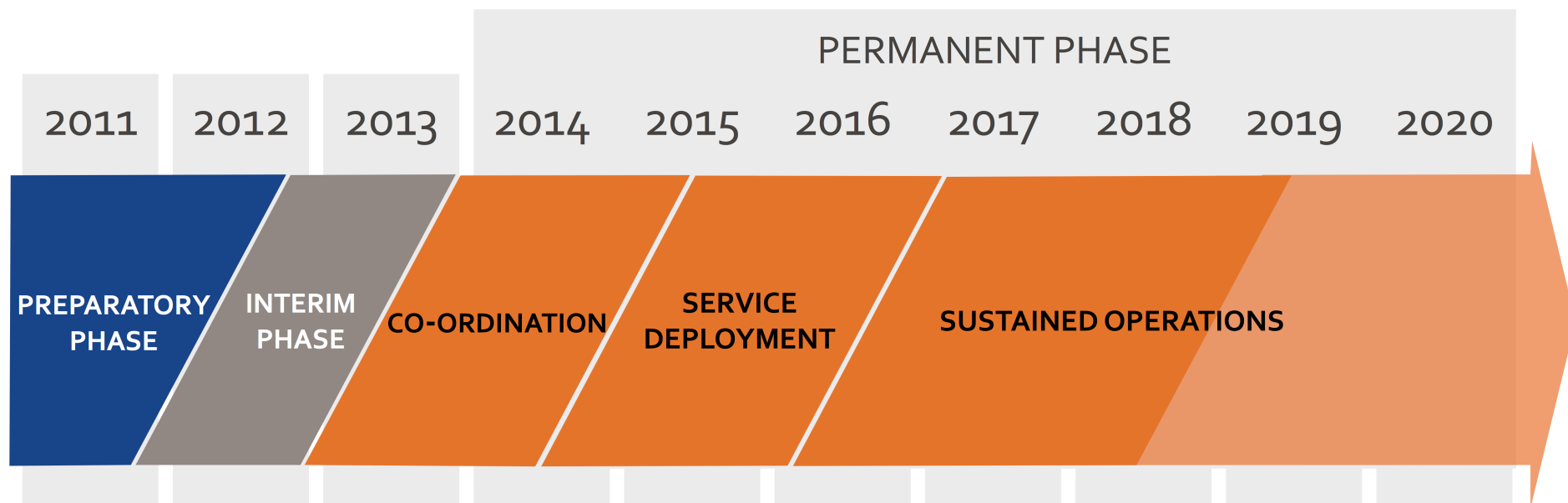
- European **life sciences** research infrastructure for **biological information** to **facilitate research**
- **Safeguard data** and build **sustainable data services**
- Participated by major bioinformatics service providers and supported by **17 EU member states**
- Creating a robust infrastructure for biological information is a **bigger task than any individual organisation** or nation can take on alone

ELIXIR's mission

Support life science
research and its
translation to:

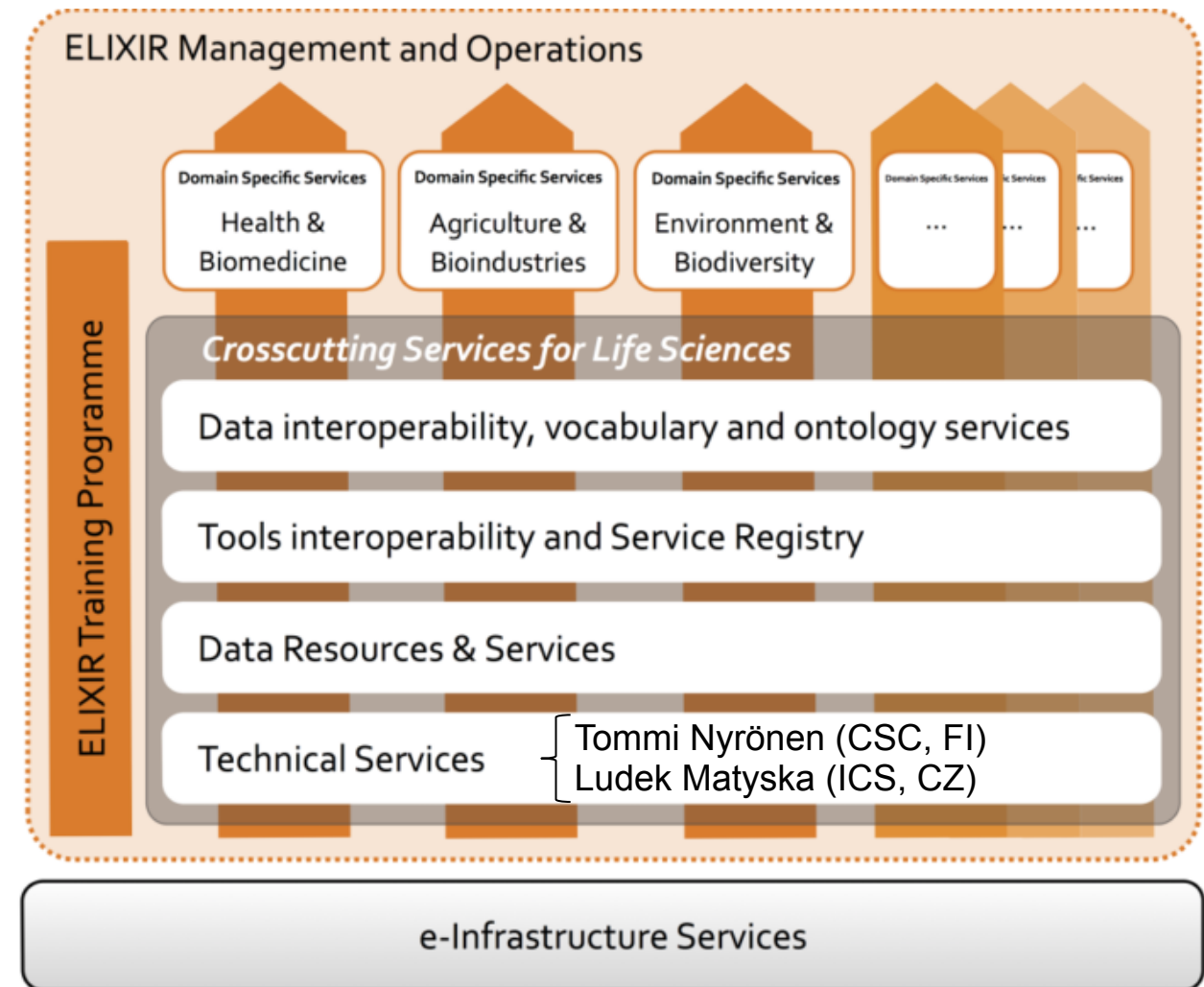


Acceleration towards sustained operations

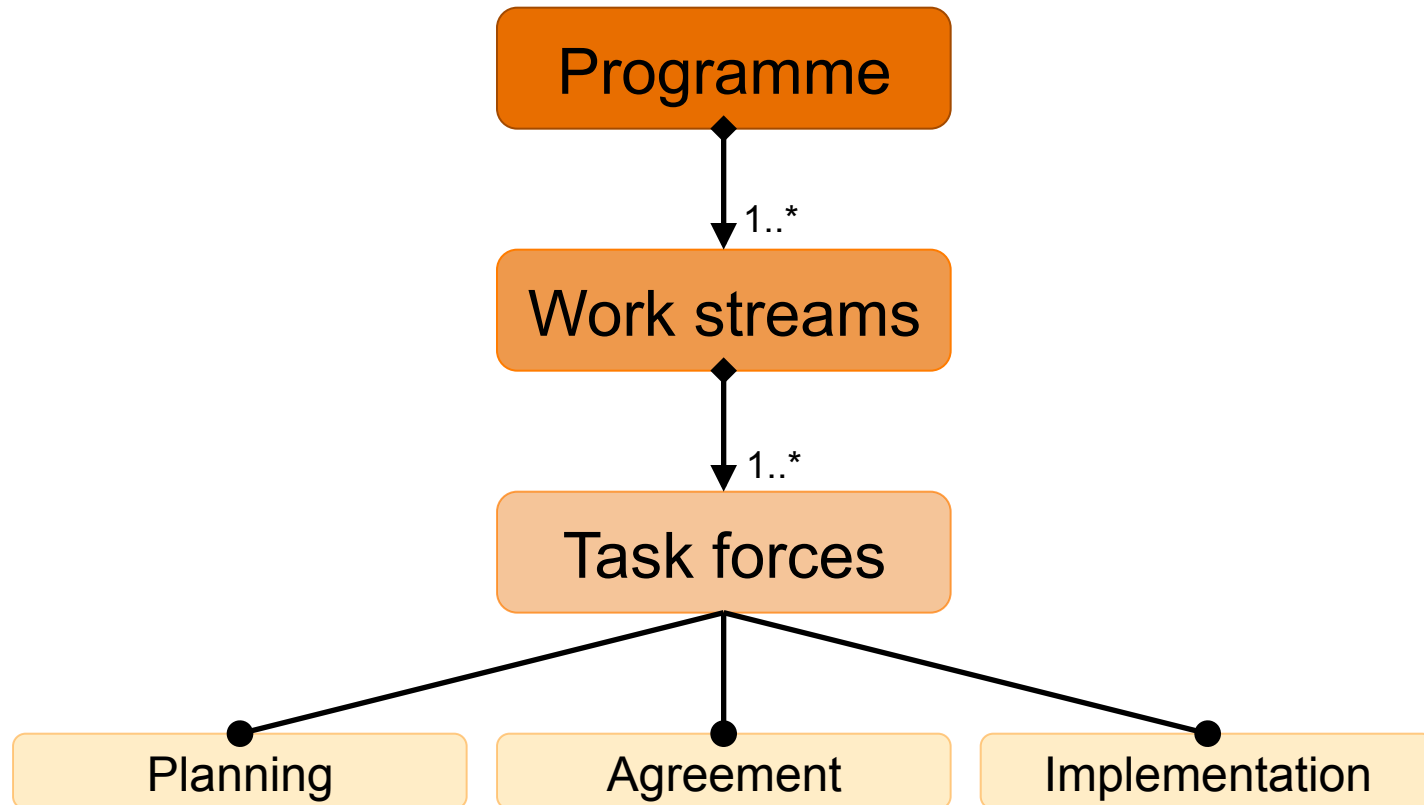


- Key objectives:
 - In 2014: Build on success of Interim Phase, formal establishment of Nodes
 - In 2015: Start delivering ELIXIR services
 - From 2016 onwards: sustained operations across ELIXIR

Work streams



Work streams and technical context



Prioritized task forces

- **Cloud**
- Service registry
- Communication
- Metrics, monitoring & quality control
- Website
- **Storage**
- **Authentication and authorisation**
- Training portal
- e-Learning and Training

Collaborations



ELIXIR Pilot Projects

- Five short-term Pilot Actions are underway to act as **test beds for integration of ELIXIR services**:
 1. ELIXIR Facing **Cloud Support and Virtual Machines** - with SIB
 2. ELIXIR Data IO to pilot the **continuous transfer of major archive resources** to a remote European location - with CSC, Finland
 3. Establishing EGA **Distributed authentication** - with CSC, Finland
 4. Establishing **EGA** as joint venture – with CRG, Spain
 5. **Improving links** between Human Proteome Atlas (HPA) and EMBL-EBI resources