‘A data scientist’s perspective on roles and responsibilities in data curation’

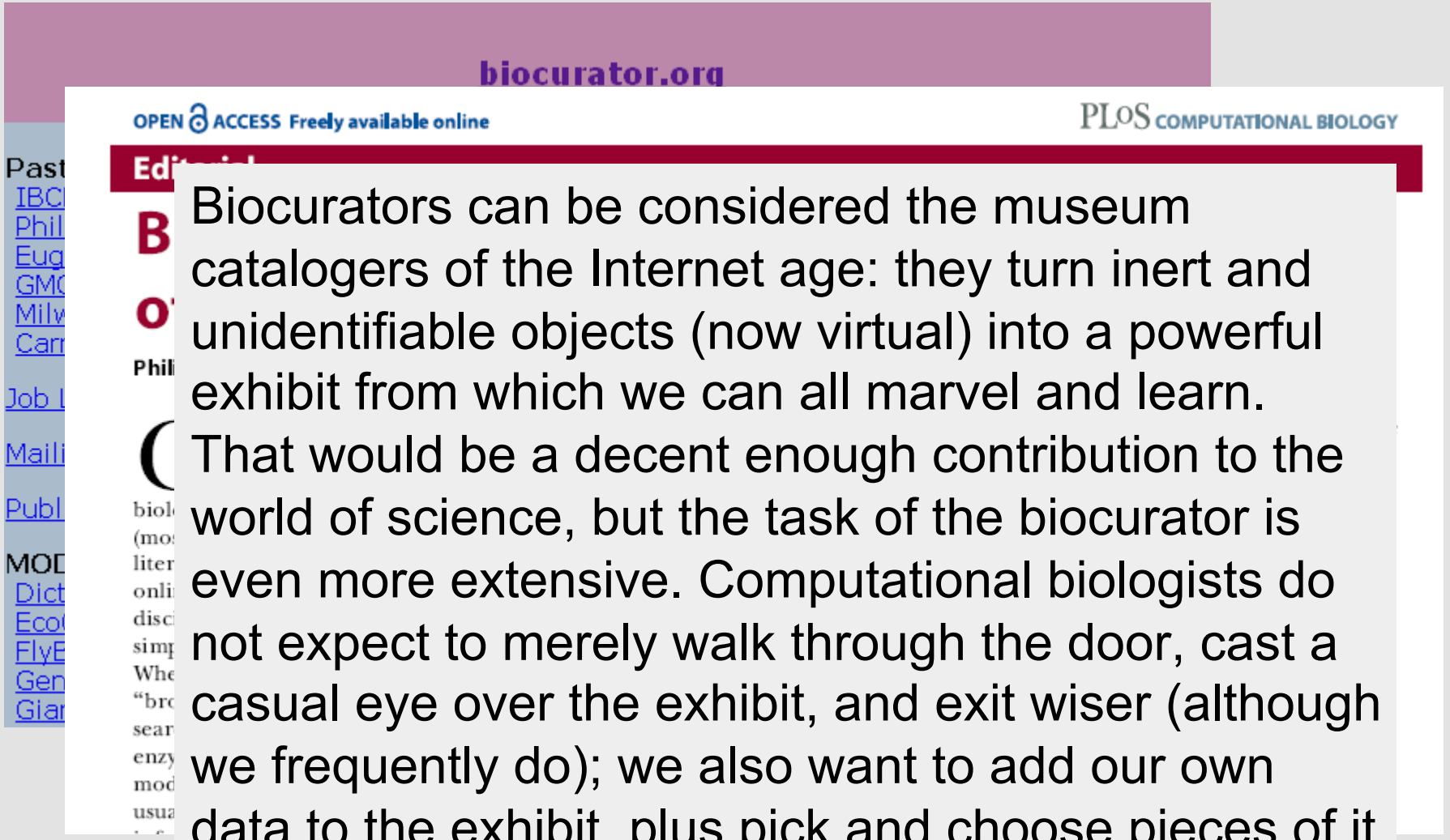
Helen Parkinson, PhD
Production Coordinator, ArrayExpress Database
European Bioinformatics Institute

EMBL-EBI

Talk content

- The Biocurator
- The EBI
- The work we do
- Case study on gene expression data
- Skills
- Training
- Future

The rise of the biocurator



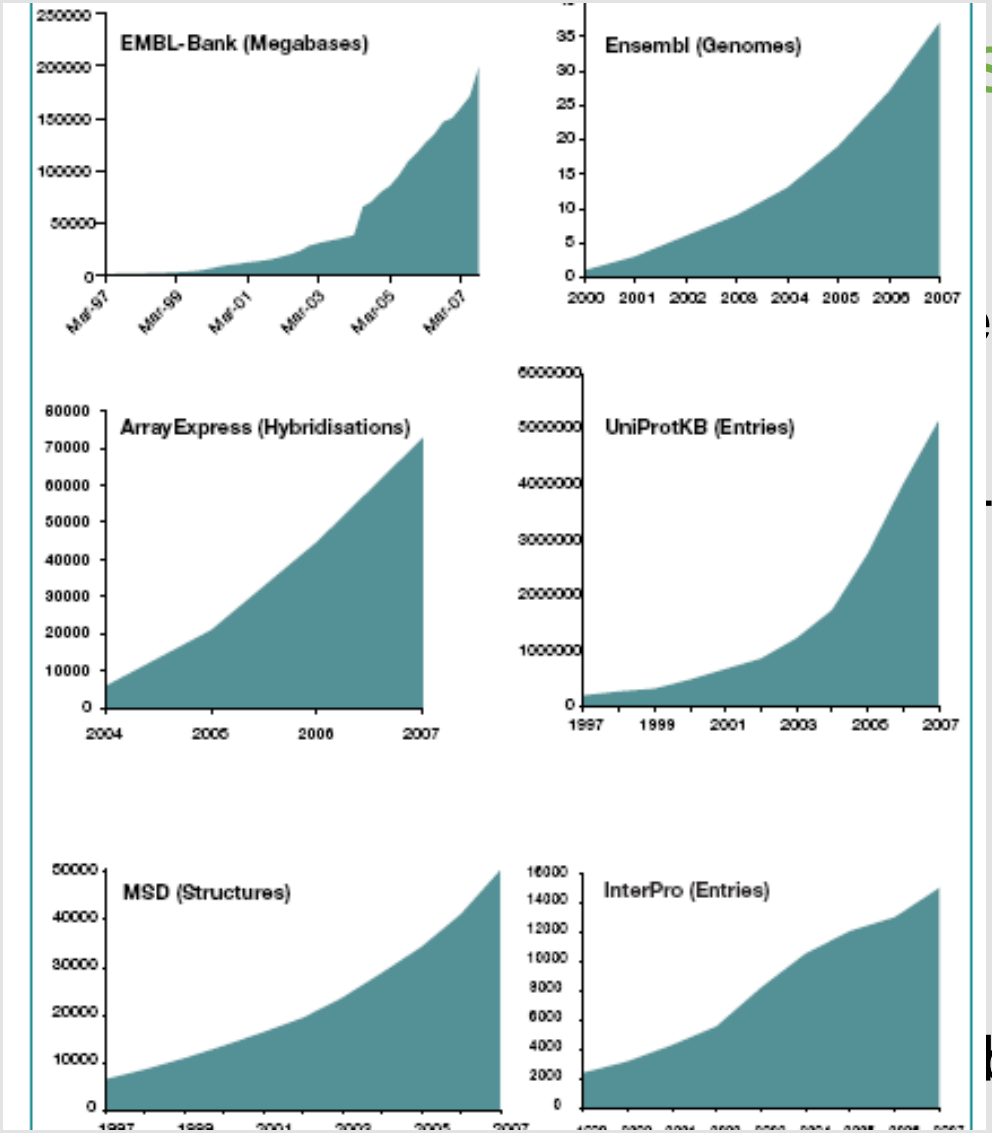
Biocurators can be considered the museum catalogers of the Internet age: they turn inert and unidentifiable objects (now virtual) into a powerful exhibit from which we can all marvel and learn. That would be a decent enough contribution to the world of science, but the task of the biocurator is even more extensive. Computational biologists do not expect to merely walk through the door, cast a casual eye over the exhibit, and exit wiser (although we frequently do); we also want to add our own data to the exhibit, plus pick and choose pieces of it to take home and create new exhibits of our own

The EBI Mission

- To provide bioinformatics **facilities and services** for the Scientific Community
- To become a flagship laboratory for basic investigator-driven **research** in bioinformatics
- To provide advanced bioinformatics **training** to individual scientists at all levels, from PhD students to independent investigators
- To help disseminate cutting edge technologies to **industry**
- **To ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress**

Dramatic Changes in Biology over last 5 years

- Data Explosion & New Types of Data
- High-Throughput Biology
- Systems Biology
- Much larger community – often naïve users
- Growth of Applied Biology – molecular medicine, agriculture, food, environmental sciences
- Diversity of use cases – more analysis than archiving

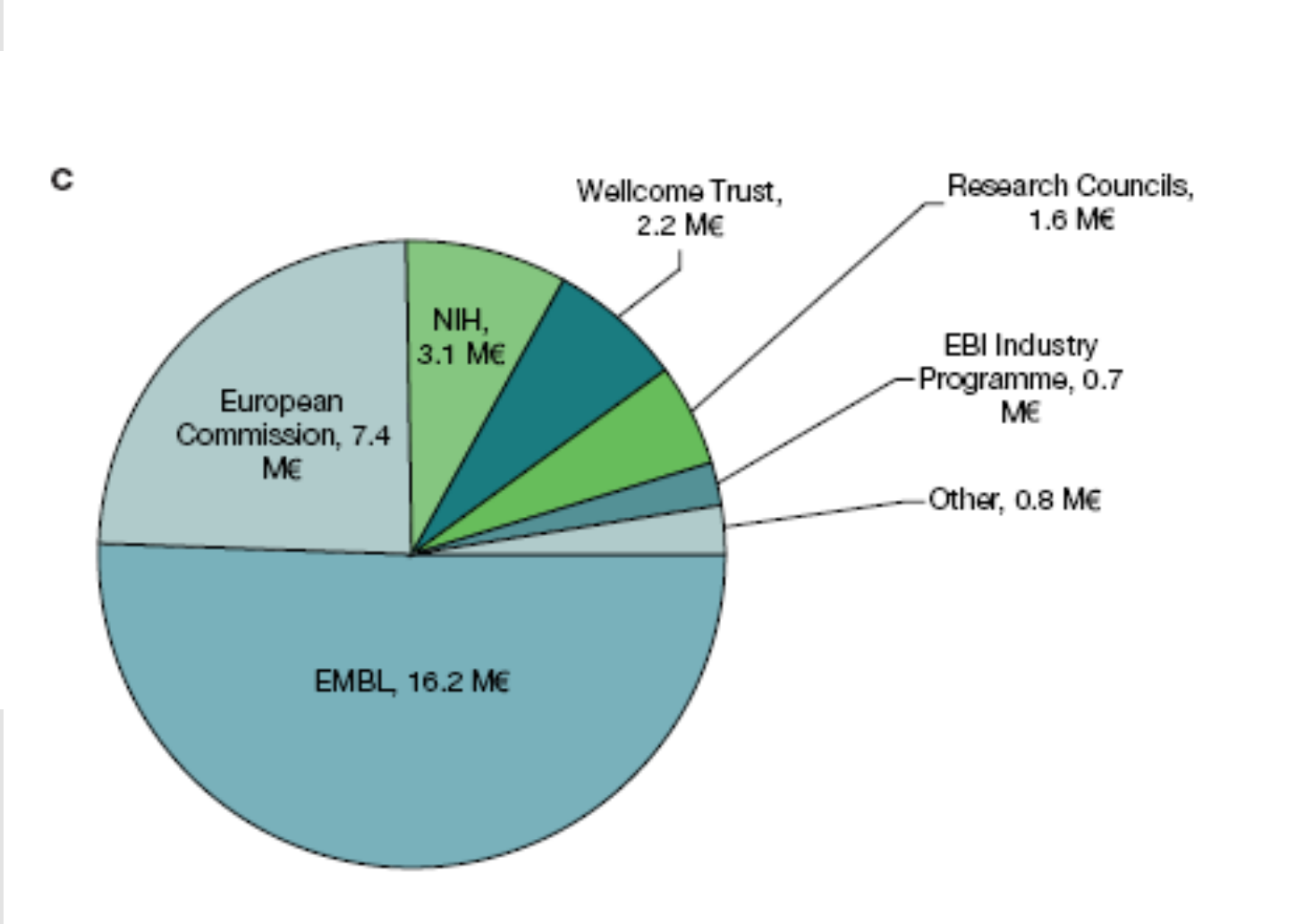


nces

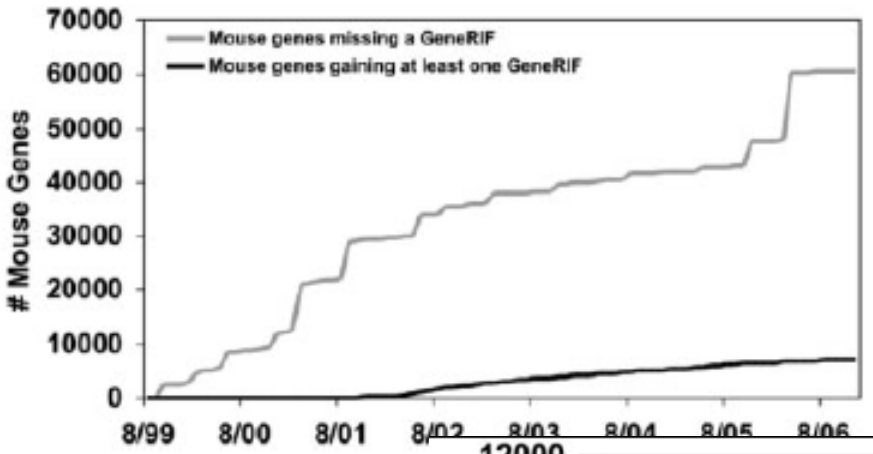
PIR) - proteins

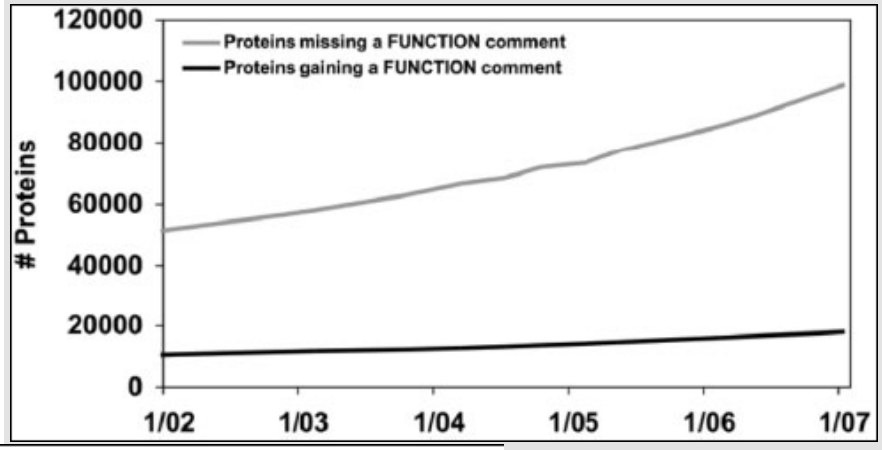
base – protein structure

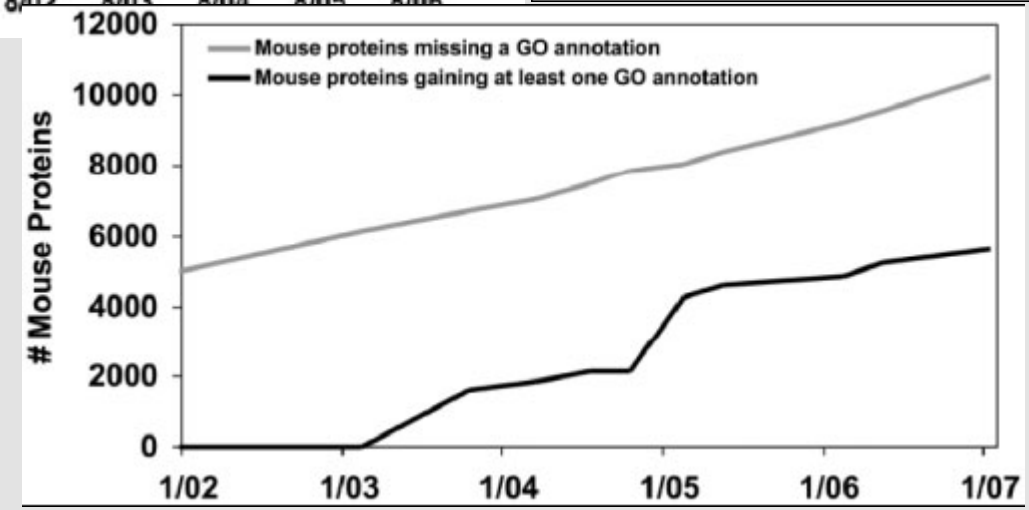
Funding


C
Wellcome Trust, 2.2 M€
Research Councils, 1.6 M€
NIH, 3.1 M€
EBI Industry Programme, 0.7 M€
European Commission, 7.4 M€
Other, 0.8 M€
EMBL, 16.2 M€

Mind the (annotation) gap







Baumgartner et al, BionformaticsVol. 23 ISMB/ECCB 2007, BIOINFORMATICS doi: 10.1093/bioinformatics/btm229

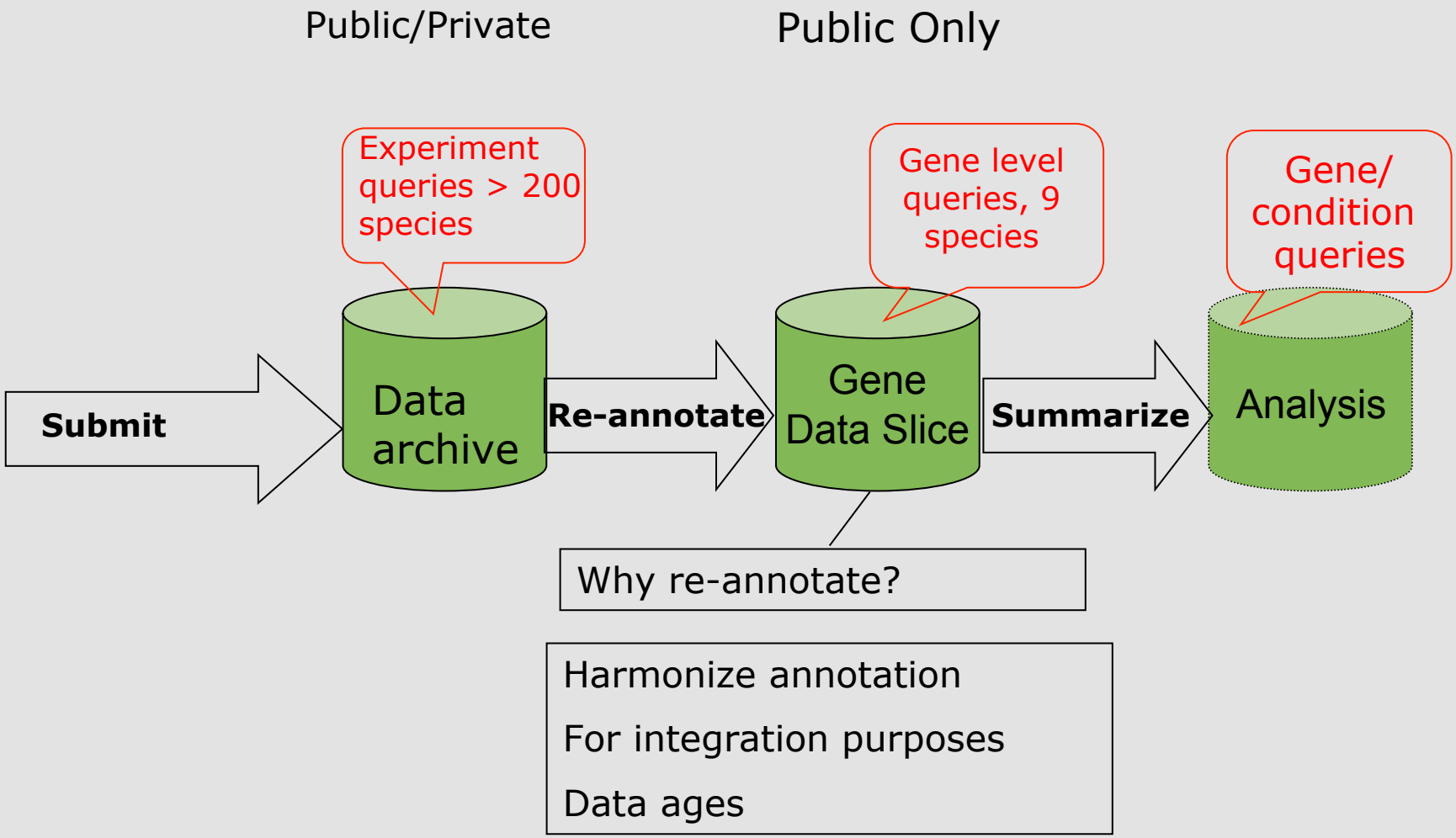
www.ebi.ac.uk/arrayexpress

EMBL-EBI

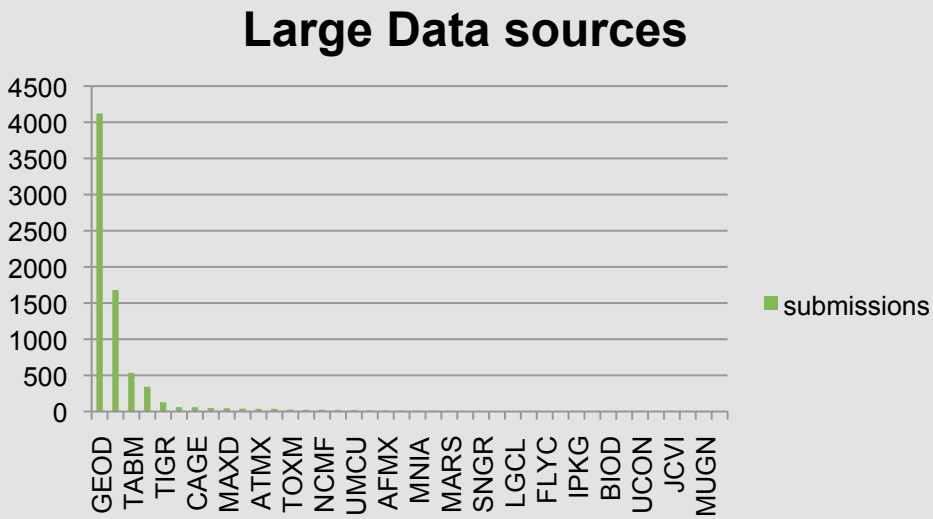
Process Case Study

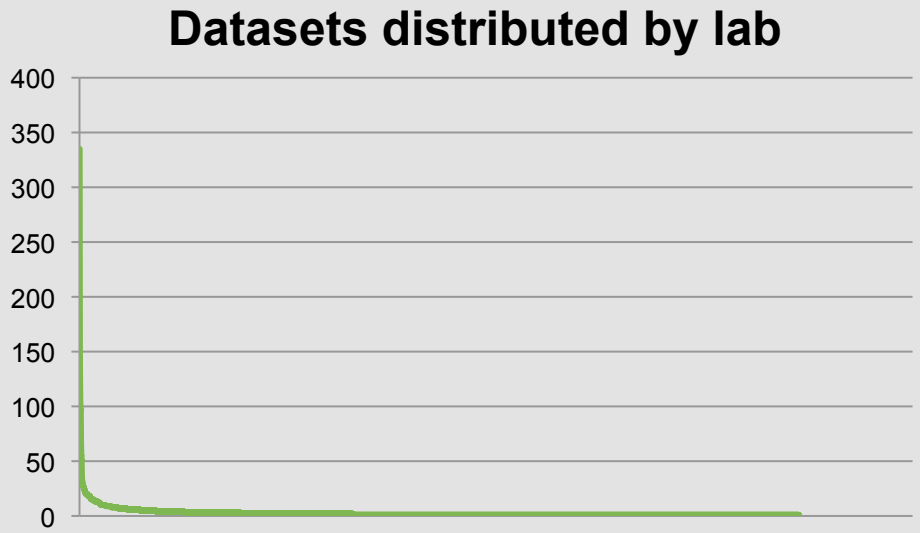
- ArrayExpress - transcriptomic
  - Acquisition – small labs, databases
  - Curation – qc, error detection
  - Annotation – vs. the genomes
  - Analysis – quality
  - Integration – vs internal and external resources
  - Presentation – multiple views on the data
- Uniprot - protein
  - Acquisition – sequence data from genome projects/users
  - Curation
  - Annotation
  - Analysis – detection of protein families
  - Integration – vs internal and external resources
  - Presentation – multiple views on the data

ArrayExpress: Overview



Long tail on the data





www.ebi.ac.uk/arrayexpress

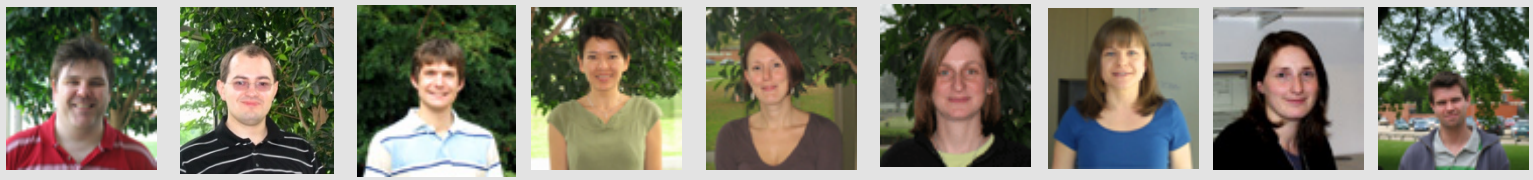
EMBL-EBI

Curation or Annotation?

- Correction of errors, typos, impossible technical conditions - curation
- Added value – added annotation, maps to ontology terms
- Semantic integration – mapping between ontology terms
- Cross database integration – links into Ensembl/Uniprot
- Specialized presentation - Atlas, images, summaries
- Updates
- Application of quality metrics
  - Social problems
  - Internal only

9 People and their skills

ants biochemist bioconductor bioinformaticians biologist clinical computer-
scientist curator flies geneticist genome gui-development
html human information java knowledge-engineers md
medic molecular mouse msc mysql ontologies oracle
perl phd post-doc rails ruby
sequence software-engineer workflows xml



Training for data scientists

- *'Ask a computer scientist for a pint of milk, and he will start with setting up a dairy farm'*
- Bio -> Comp science OR Comp science -> Bio
- Training models that work hard for the community
  - train the trainer
  - Intensive workshops
  - support bioinformaticians
  - E-learning
  - Training for teachers
  - Mainly analysis and resource training
  - Cost recovery model for roadshows

Challenges for the future

- Retaining skilled personnel
- Ensuring training is adequate for researchers esp. clinicians
- Promoting data sharing esp. in clinical communities
- Ethical issues and data protection
- Semantic data integration – sample dimension
- Dealing with the long tail cost effectively
- Embracing new technologies
- Dealing with ethical issues
- Staying relevant

Acknowledgements