Sharing Scientific Data to Advance Science

Claudia Bauzer Medeiros
Instituto de Computacao – UNICAMP
CCES – CEPID (eScience)

Sharing Scientific **Data** to Advance Science

Claudia Bauzer Medeiros

Instituto de Computacao – UNICAMP

CCES – CEPID (eScience)

Sharing

- WHAT -
- WHY -
- HOW –
- WHO –

Science and beyond

- “Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it’s going to get worse”
  - Office Science of Data Management challenge - DoE

Subject of the decade

@Fabio Porto

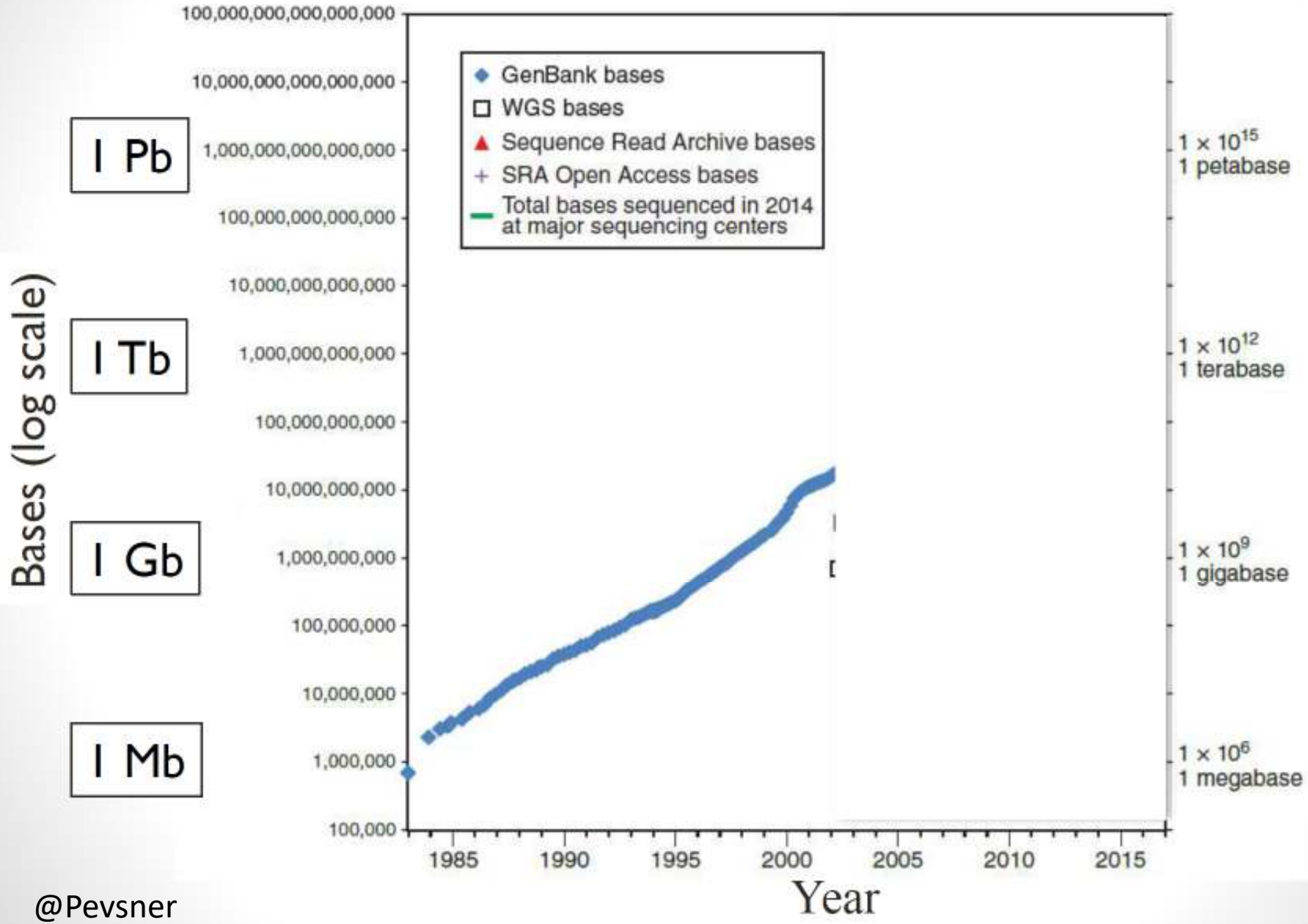
DCC because good research needs good data

The value of digital curation

Some £3 billion of public money is invested annually in research in the UK alone, yet the research data resulting from this considerable investment are seldom as visible as they might be.

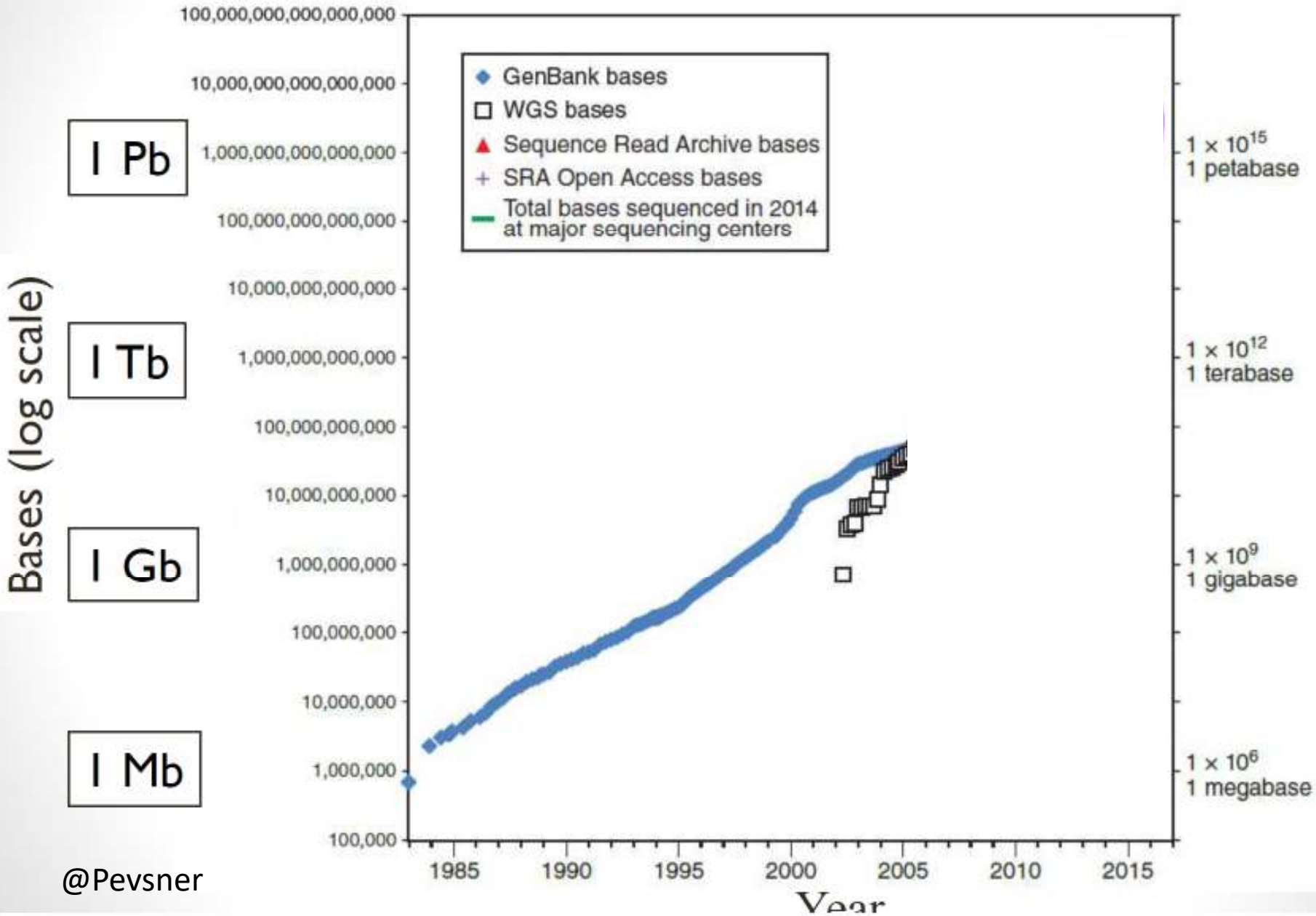
By learning how to preserve and share digital materials so others can effectively reuse them, you will maximise the impact of your research – and inspire confidence among the research councils and funding bodies that invest in your work.

Growth of DNA sequence in repositories



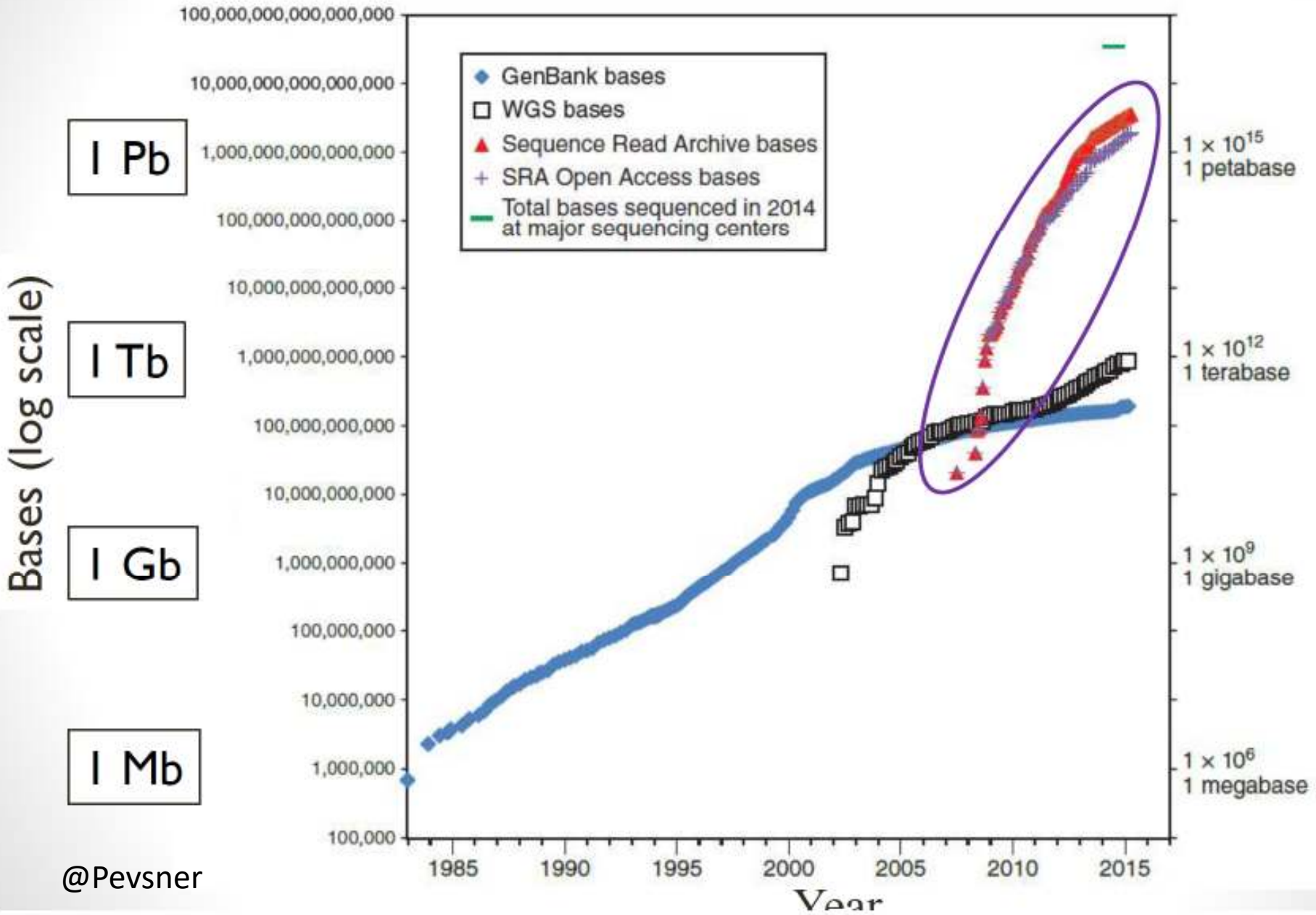
@Pevsner

Growth of DNA sequence in repositories



@Pevsner

Growth of DNA sequence in repositories

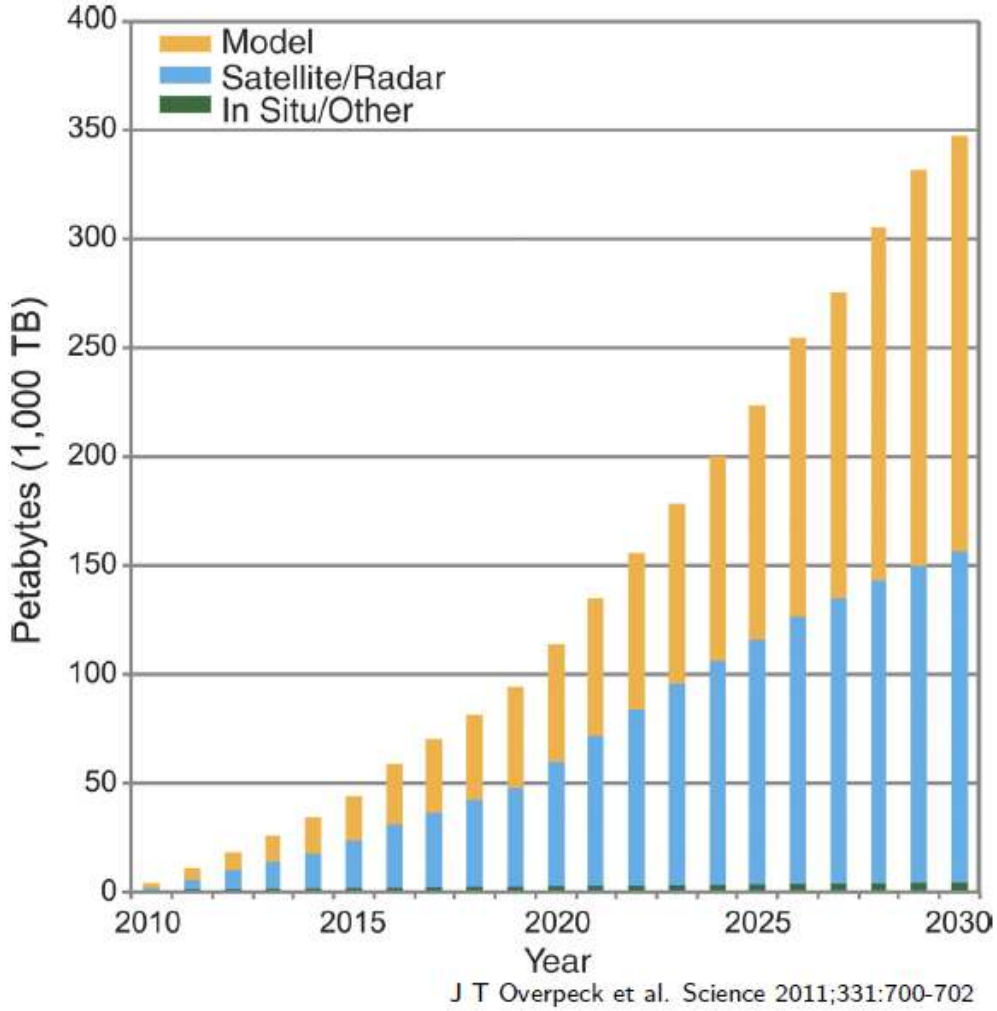


@Pevsner

More Data

Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

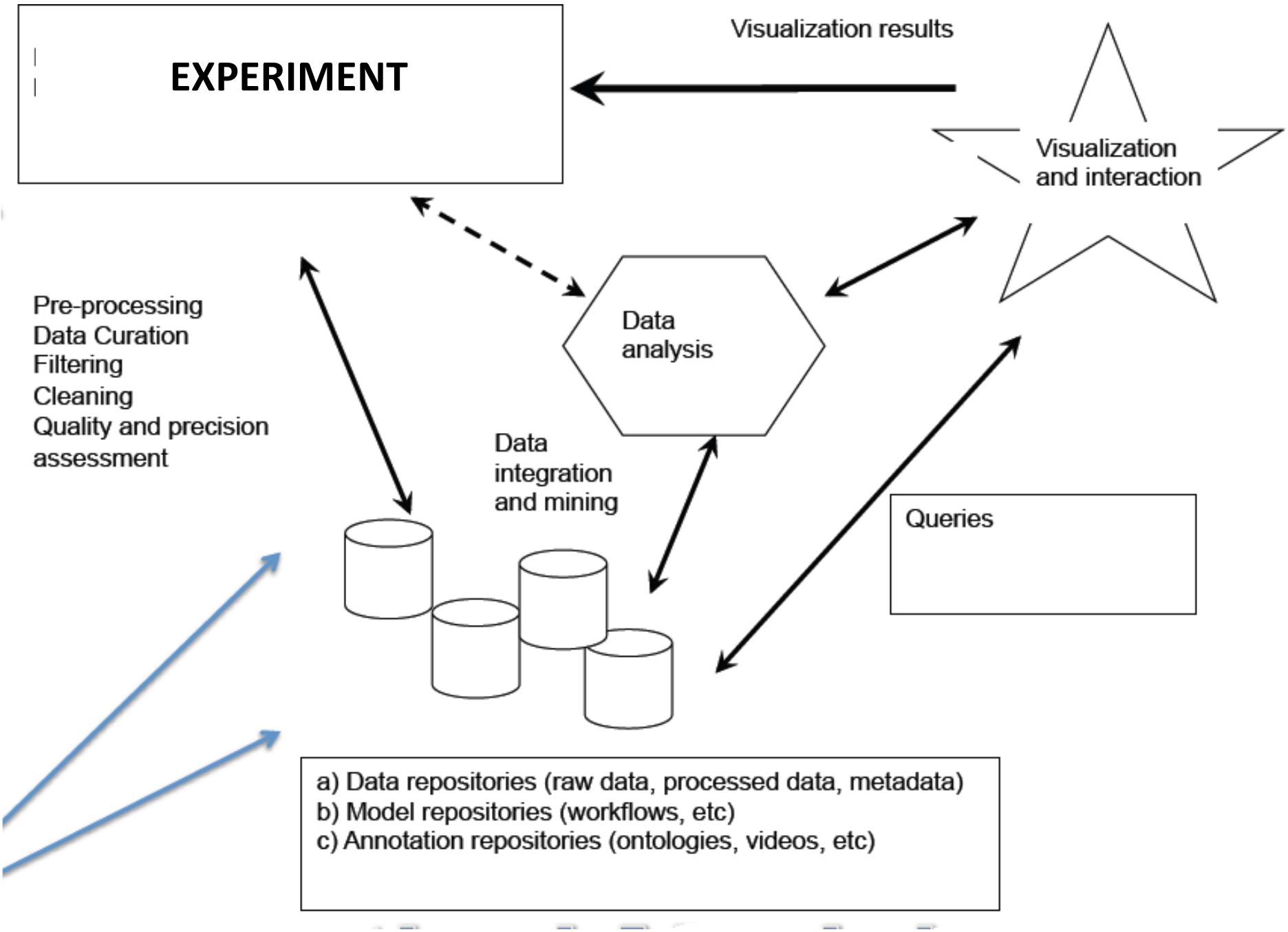
(BNL: Even if you are?)



J T Overpeck et al. Science 2011;331:700-702

Sharing

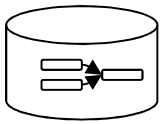
- WHAT - DATA
- WHY -
- HOW –
- WHO –

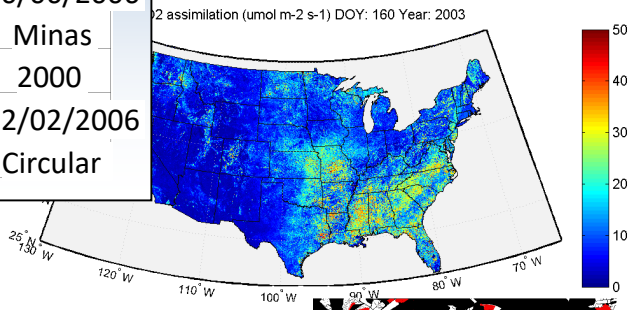

EXPERIMENT
Visualization results
Visualization and interaction
Data analysis
Pre-processing
Data Curation
Filtering
Cleaning
Quality and precision assessment
Data integration and mining
Queries
a) Data repositories (raw data, processed data, metadata)
b) Model repositories (workflows, etc)
c) Annotation repositories (ontologies, videos, etc)

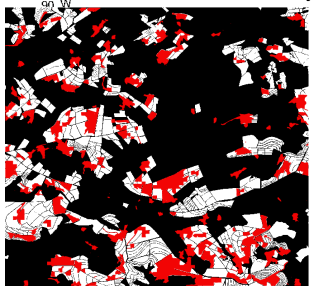
WHAT (is) Data?

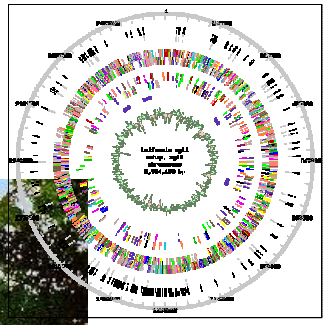
Table of Product Characteristics

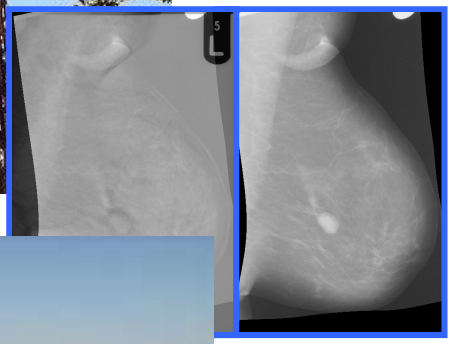
id	Property name	Value
MilkProd	productsrep	MilkA
MilkProd	quantity	10000
MilkProd	validity date	10/06/2006
CheeseProd	productsr	Minas
CheeseProd	e quantity	2000
CheeseProd	validity date	12/02/2006
CheeseProd	shape	Circular

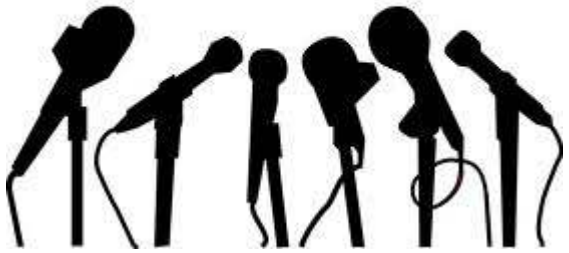


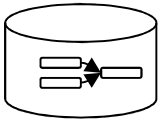






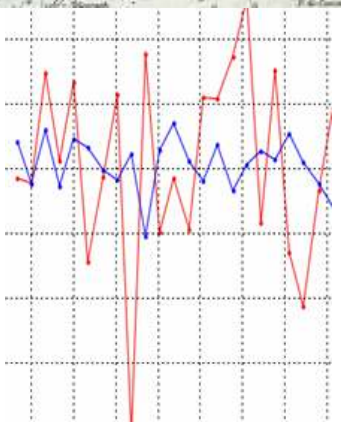






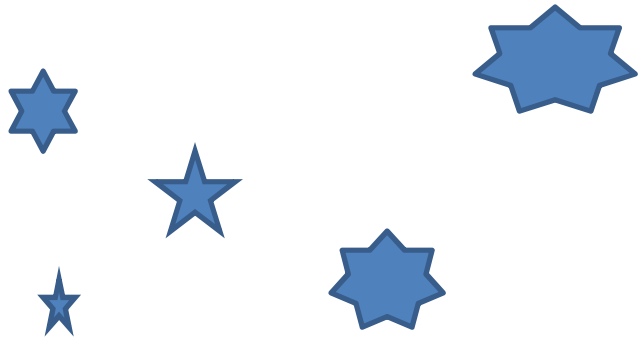
Data?

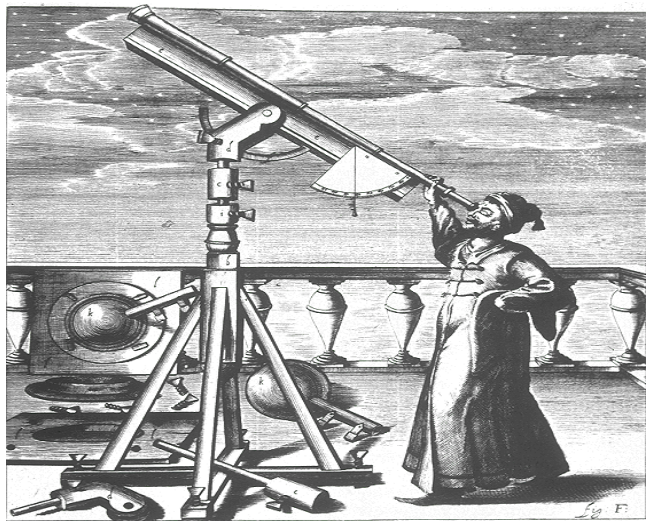
- Direct and indirect observations

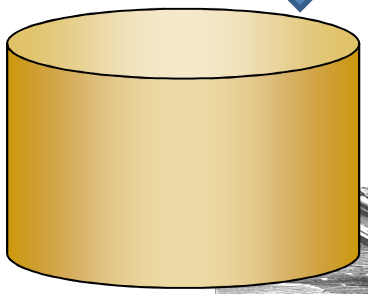


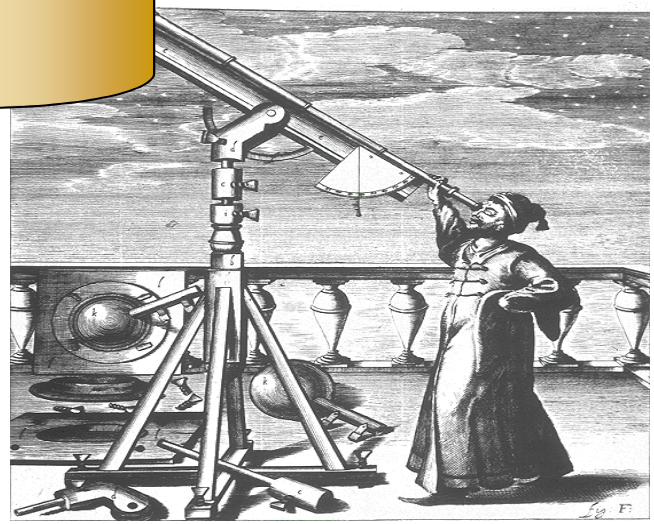
Data

Lights
Robotic Arm
Imaging Sonar
HD Stereo Cameras
Chemistry/Biology Lab
Fiber Optic Cable
cev



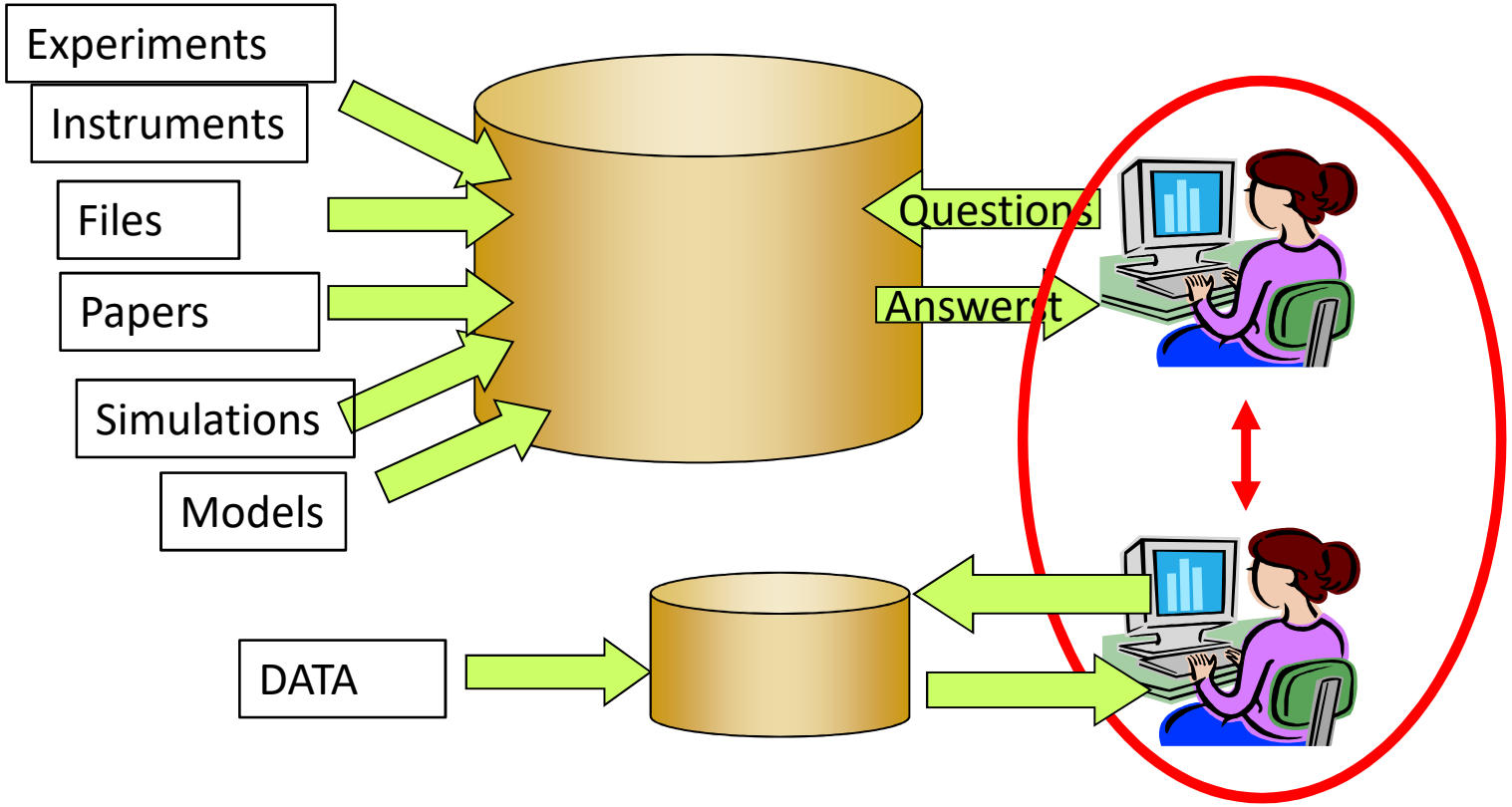






Standards
Formats
Algorithms
Architectures

Slide adapted from Jim Gray



Data-driven science

"Collaboratory"

Sharing

- WHAT - DATA
- WHY - ADVANCE SCIENCE
- HOW –
- WHO –

WHY - advance science

- Data sharing has become part of scientific methodology:

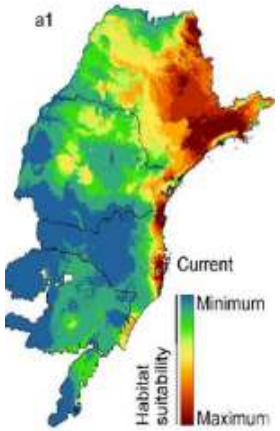
- Reproducibility / provenance
- Reuse (and save resources)
- Test new/alternative hypotheses
- Explore topics not envisioned by the initial investigators
- Create new data sets (combination of others)
- Detect fraud

Sharing advances science

Correlations across distinct experiments

– Sturge-Weber Syndrome and uveal melanoma (congenital neurocutaneous disorder; eye cancer)

- Mandacaia bee, climate change and coffee



Sharing

- WHAT - DATA
- WHY - ADVANCE SCIENCE
- HOW – CURATED REPOSITORIES
- WHO –

HOW – Curated Repositories

RCSB PDB
PROTEIN DATA BANK
An Information Portal to
124286 Biological
Macromolecular Structures

NCBI Resources

GenBank

BIPMed Brazilian Initiative on PRECISION MEDICINE

View data repositories

HOW – Data Journal

- Biological sciences:
  - nucleic acid sequence; protein sequence; molecular & supramolecular structure; neuroscience; omics; taxonomy & species diversity; mathematical & modelling resources; cytometry; organism-focused resources

- Health sciences
- Chemistry & chemical biology
- Earth and environmental sciences
- Physics, astrophysics & astronomy
- Social sciences
- Generalist repositories
- Institutional or project-specific repositories

HOW – Institutional Repositories

Stanford Digital Repository

HOW – Tools and Infrastructures

Software development

Secure file sharing

Share data and results

(Harvard) Explore research data

Citation manager

Sharing

- WHAT - DATA
- WHY - ADVANCE SCIENCE
- HOW – CURATED REPOSITORIES
- WHO – DATA LIBRARIAN

WHO - (Research) Data Librarians

- Organization
- Curation
- Support of Researchers

Bodleian Data Library

Bodleian Libraries

UNIVERSITY OF OXFORD

Homepage

Sharing

- WHAT - DATA
- WHY - ADVANCE SCIENCE
- HOW – CURATED REPOSITORIES
- WHO – DATA LIBRARIAN

Obrigada