1. Introduction & Motivation

Fundamentals of Data Management

Dr Adam Carter
Project Manager, EPCC
adam@epcc.ed.ac.uk
+44 131 650 6009

Fundamentals of data management

- This course is about the management of digital research data
  - Data *analysis* comes next Semester

- After completing this course, you should:
  - Understand the digital data lifecycle, from creation to archive
  - Understand the different ways of storing and organising data
  - Understand the importance of metadata
  - Understand how all this happens in the digital science of today
  - …and how it's only going to get worse

Our challenge: the data-sharing panda

http://www.youtube.com/watch?v=N2zK3sAtr-4

Our challenge: the data-sharing panda

http://www.youtube.com/watch?v=N2zK3sAtr-4

Data management…

- It's about doing better science!
  - Or perhaps, doing science, better
- Today's science is complex
  - Methods are increasingly complex (often captured in software)
  - Experimental or simulated data are increasingly large and complex
- Scientific progress is built on openness
  - Publication of thesis – method – results
- But not just on a USB!
  - Openness ⇒ Intelligent Openness
    - Royal Society June 2012, Science as an Open Enterprise: http://royalsociety.org/policy/projects/science-public-enterprise/report/

The data don't fit

- A single paper is just too small to contain everything needed for effective
  - Scrutiny
  - Reproducibility
  - Validation, and
  - Re-use
- Standing on the shoulders of giants is getting harder and harder
- The paper increasingly needs to be augmented
  - With publication of the data
  - (Ideally) with publication of the workflows and software

Well managed, publically accessible data is important: why?

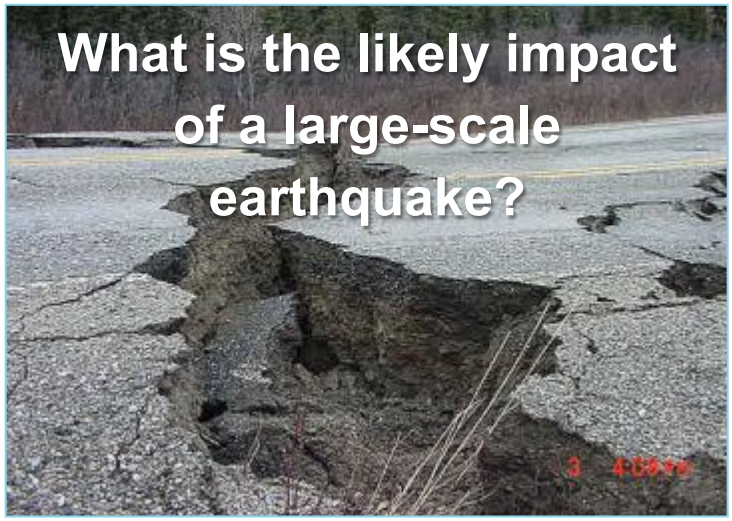
- Here are a few reasons (from the UK Data Archive):
  - Increases the impact and visibility of research
  - Promotes innovation and potential new data uses
  - Leads to new collaborations between data users and creators
  - Maximizes transparency and accountability
  - Enables scrutiny of research findings
  - Encourages improvement and validation of research methods
  - Reduces cost of duplicating data collection
  - Provides important resources for education and training

slide from DataONE

Reproducible science at risk

- John Ioannidis 2005: “Why Most Published Research Findings Are False”
  - doi:10.1371/journal.pmed.0020124
- 2011 “Stapel” scandal at Tilburg Uni., Netherlands
  - “Fabricated datasets, forged scientific studies, fake research assistants”
- 2012 “Penkowa” scandal at Uni. Copenhagen, Denmark
  - “misleading and false information”, “constructed data”
- International Union of Crystallography study (2013)
  - IUCr promote increasing use of automatic validation for structure data
  - Found, retrospectively, *over 100 fraudulent structures* published in *Acta Crys. E* from 2007-2009
    - http://www.iucr.org/__data/assets/pdf_file/0020/80273/MH_Publication-of-Small-Molecules.pdf

Data are driving solutions to complex science and societal challenges



Earthquake simulations integrate
- Fault geometry data
- + Subsurface data
- + Structure data
- + Population data
- + Physical infrastructure data

Disease spread models couple
- Medical data
- + Population data
- + Envt. data

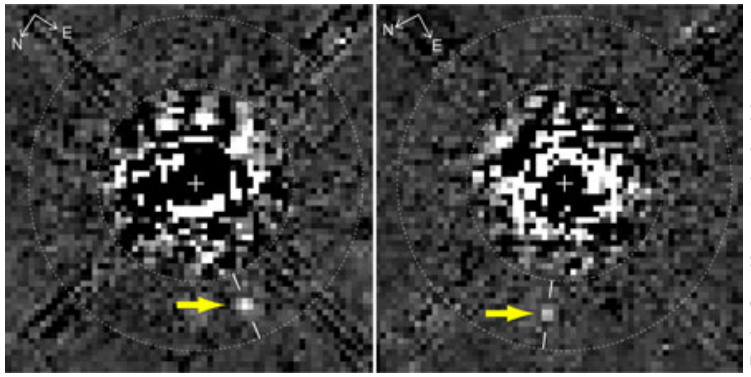
Productivity analysis leverages interoperability of
- Germplasm data
- + Genetic and phenotypic data
- + Statistical data
- + Bibliographic data

slide from John Wood, RDA

New discoveries

A new image processing technique reveals something not before seen in this Hubble Space Telescope image taken 11 years ago: A faint planet (arrows), the outermost of three discovered with ground-based telescopes last year around the young star HR 8799.D. Lafrenière et al., Astrophysical Journal Letters

"Planet hidden in Hubble archives" *Science News* (Feb. 27, 2009)

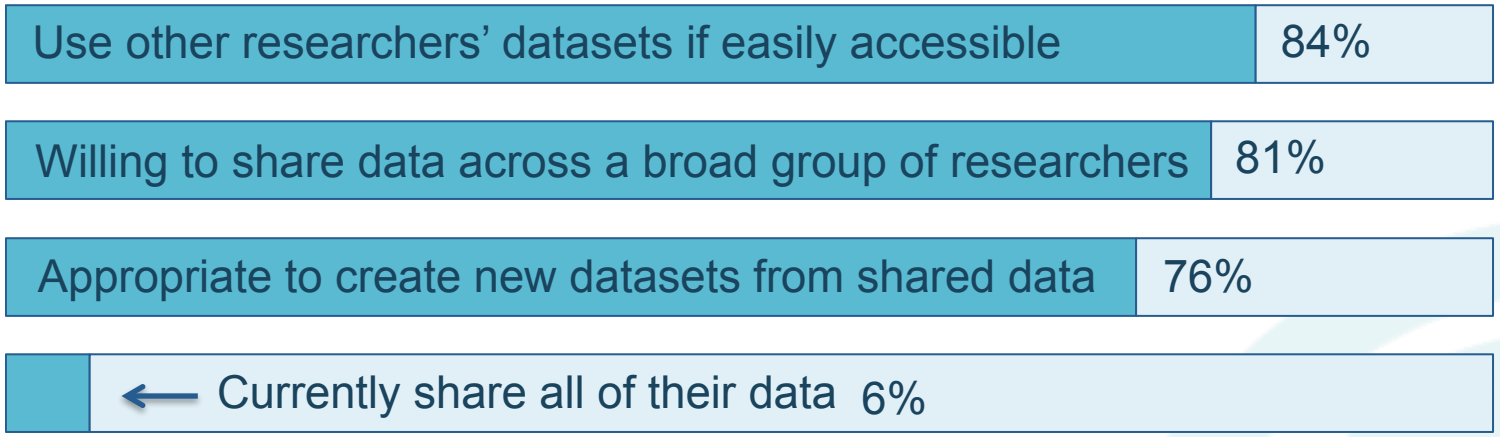


D. Lafrenière et al., ApJ Letters

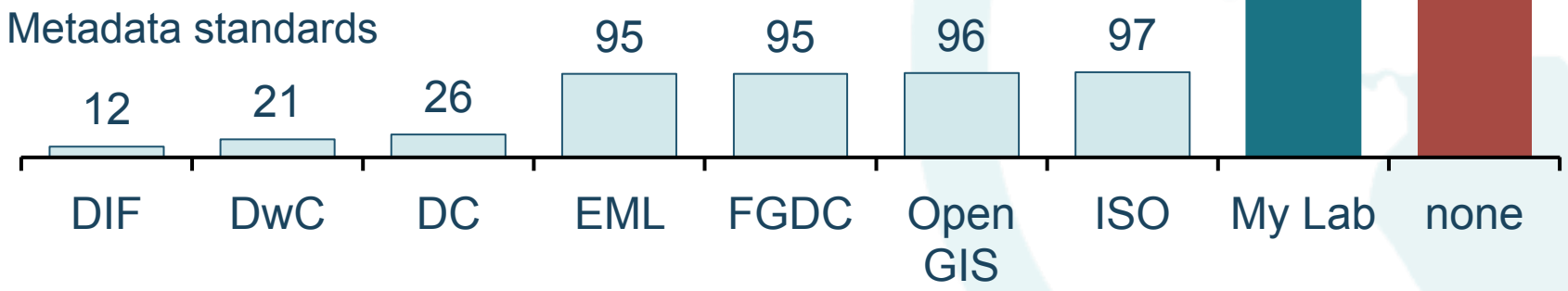
"The first thing it tells you is <u>how valuable maintaining long-term archives can be</u>. Here is a major discovery that's been lurking in the data for about 10 years!" comments Matt Mountain, director of the Space Telescope Science Institute in Baltimore, which operates Hubble.

"The second thing its tells you is having a well calibrated archive is necessary but not sufficient to make breakthroughs — it also takes a very innovative group of people to develop very smart extraction routines that can get rid of all the artifacts to reveal the planet hidden under all that telescope and detector structure."

Scientists do want to share data



…but don't know how to!



slide from Bill Michener, DataONE

Science is increasingly data driven

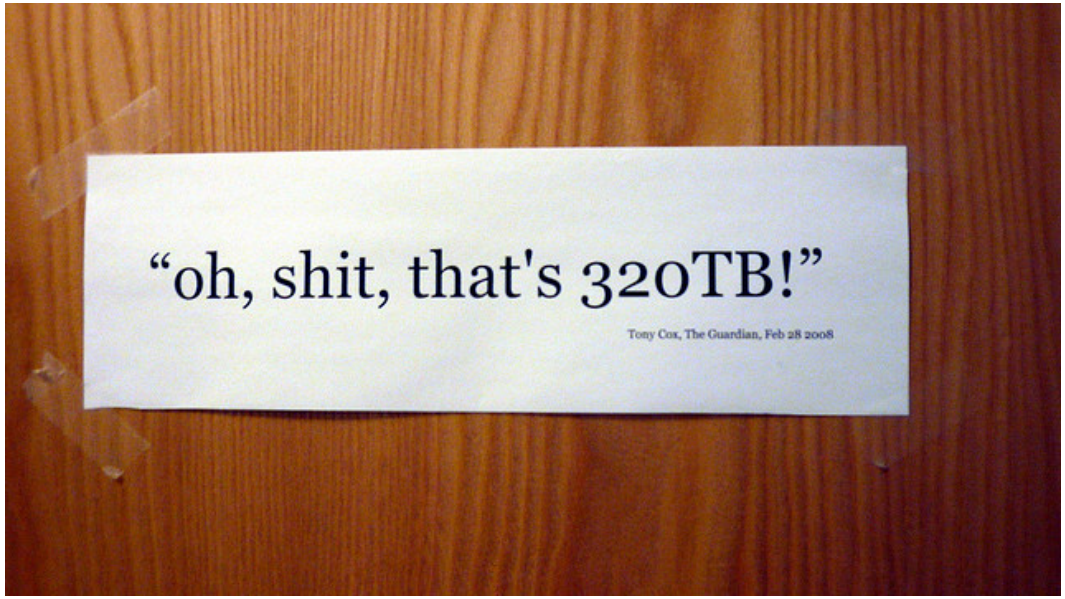
- And this means Big Data!
- And Big Data means…
  - *Volume*, which means
    - Large files (or, more generally, “digital objects”) in terms of bytes
    - Large *numbers of* files or digital objects
  - *Velocity*, which means
    - Rapidly changing datasets
    - Think tweets, data streams from scientific instruments, etc.
  - *Variety*, which means
    - Complexity, in the data structures
    - Semi-structured data (eg. XML, Web)
    - Unstructured data (email, PDF, images, video, audio)

How big is Big?

- In the 10 years to 2008, the largest current astronomical catalogue, the Sloan Digital Sky Survey, produced 25 TB of data from telescopes – 2.5 TB/year average
- In 2013 Facebook users uploaded three billion photos monthly for a total of 3,600 terabytes annually – 3.6 PB/year
- In 2013 YouTube users uploaded 48 hours of footage every minute – c. 17.3 GB/min or about 9.1 PB/year
- In 2012 LHC collision data was being produced at approximately 25 PB/year

Big is pretty big

- In 2008 the Sanger Institute's gene sequencers produced raw data at 45 GB/s
  - So over a 2 hour run…



…said Tony Cox, head of Sequencing Informatics, Sanger Institute quoted in the Guardian newspaper

Image by

Cary Doctorow, Wellcome-Sanger, http://www.flickr.com/photos/doctorow/2698389465 /

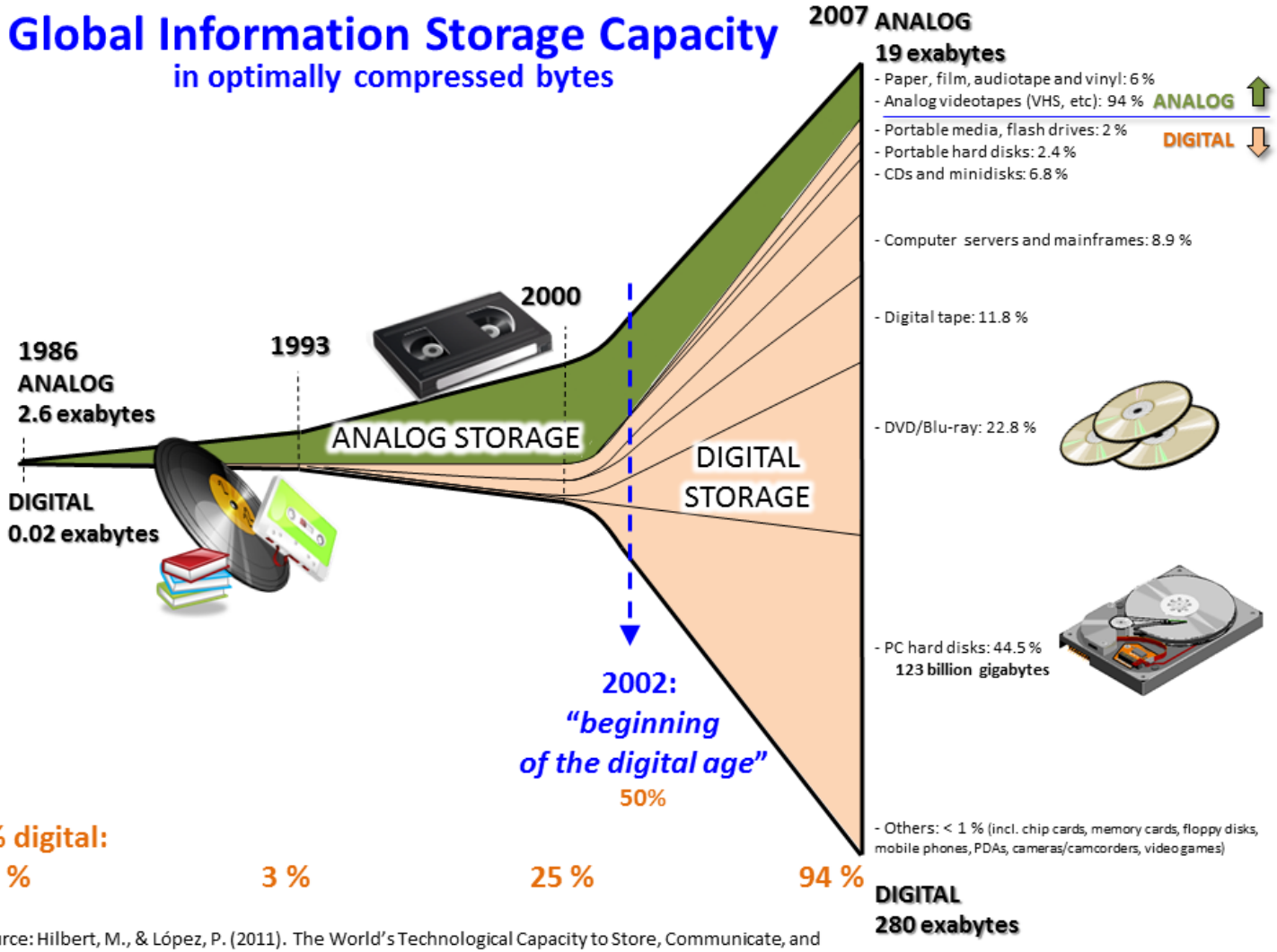
Big keeps getting bigger

- By 2014, it is anticipated that the Large Synoptic Survey Telescope will produce 20 TB each night, say 5 PB/year
- By the year 2019, The Square Kilometre Array radio telescope is planned to produce 50 TB/day of processed data – say 10 PB/year – from a *raw data rate* of 7,000 TB/s
- EISCAT_3D is a three-dimensional imaging radar to be located in the northernmost parts of Europe; from 2020 it will generate data of c. 100 PB/year

Computational science & modelling

- Not quite so bad in terms of file volume
  - Raw file sizes are limited by system memory
  - Even so, 1 simulation can produce 100GB files
- But ensemble runs can generate 100s of files in parallel
- German climate modelling centre DKRZ store c. 10 PB/year
- Velocity tends to be low
- Variety can be cripplingly high
  - Arbitrary file formats, random complexity!

How much can we store?



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Data management fundamentals

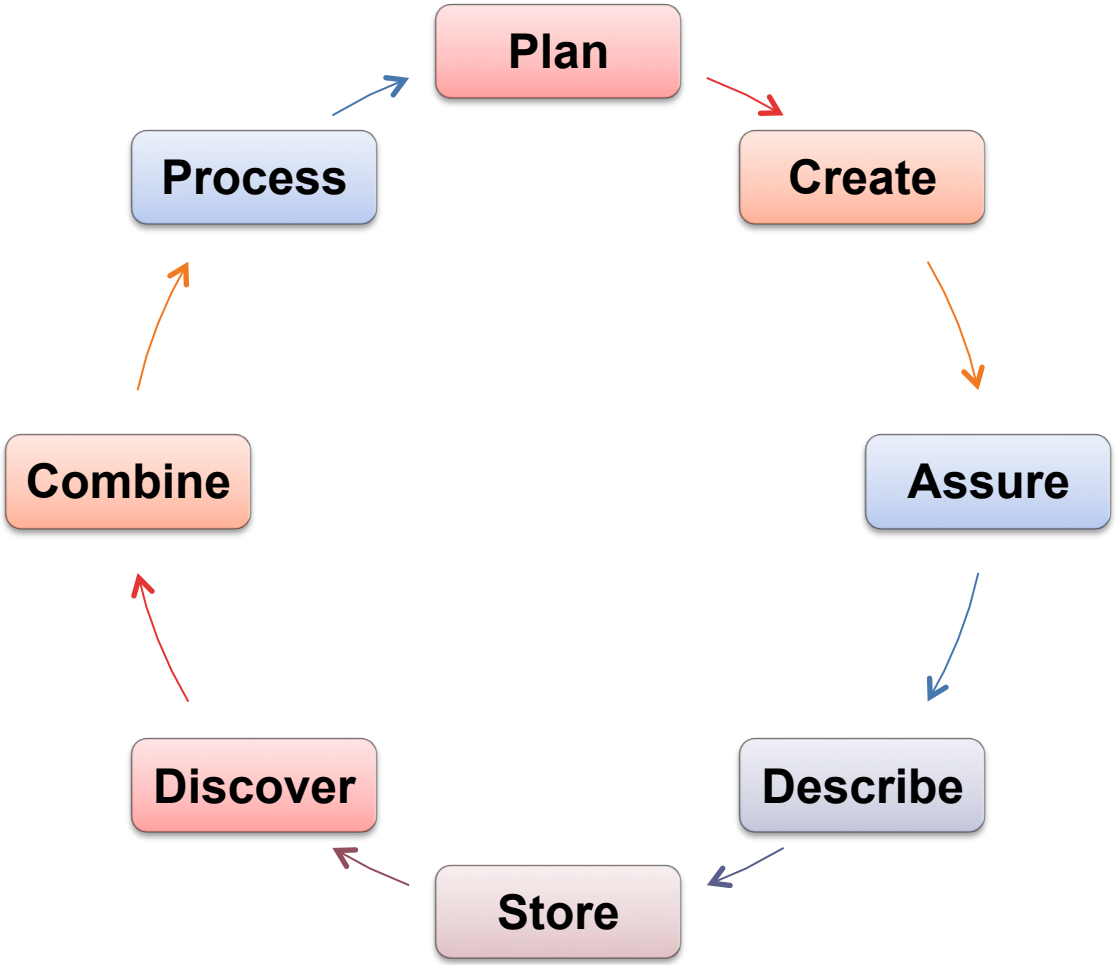
- Capturing and storing science data is one thing
- Organising and managing them is another
- This course focuses on organising and managing
  - Because the power to do that better is in your hands!

Data management fundamentals

- Structured approaches to research data
- Files & file formats
  - what is a file? Common research data file formats
- Relational databases, SQL
  - files are not the only storage
- Web formats, XML & JSON
  - formats for the web and wide-area distributed data
- NoSQL databases, key-value stores
  - from big data to scientific databases
- Resource description format & semantic data
  - linked data models

Data management fundamentals

- Data lifecycle and elements of data management planning



Data management fundamentals

- Data management & movement in HPC
  - iRODS, gridFTP
- Metadata, standard formats
  - Why good metadata are important
- Persistent identifier systems
  - Finding data after 50 years
- Publication and citation of research data
  - Maintaining the scientific record in a digital world
- Archiving and preservation
  - Looking after data long-term

Data management fundamentals

- Legal aspects of research data
  - Ownership, copyright, licensing, certification
- Data management hardware
  - Distributed infrastructure, hierarchical storage, Amdahl systems
- Distributed AA
  - A brief overview of authentication and authorisation
- Future challenges
  - Scale, economics, how long term is long-term?
- Guest lectures
  - University Data Library
  - Astronomy

Course team

- Lecturers
  - Dr Rob Baxter
  - Dr Adam Carter
  - Dr Charaka Palansuriya
  - Mr Albert Heyrovsky
- Labs and tutorials, also starring
  - Dr Amy Krause
  - Mr Kostas Kavoussanakis






Please stop talking now…