University of Pittsburgh

SCHOOL OF Information Sciences

Gearing up for Data? Institutional drivers, challenges & opportunities

Professor Liz Lyon, School of Information Sciences, University of Pittsburgh

Senate Plenary Meeting, October 2014

RDM Agenda

1. Context and Drivers
2. Challenges
3. Opportunities and Benefits

Six reasons why you should care about managing your research data

theguardian

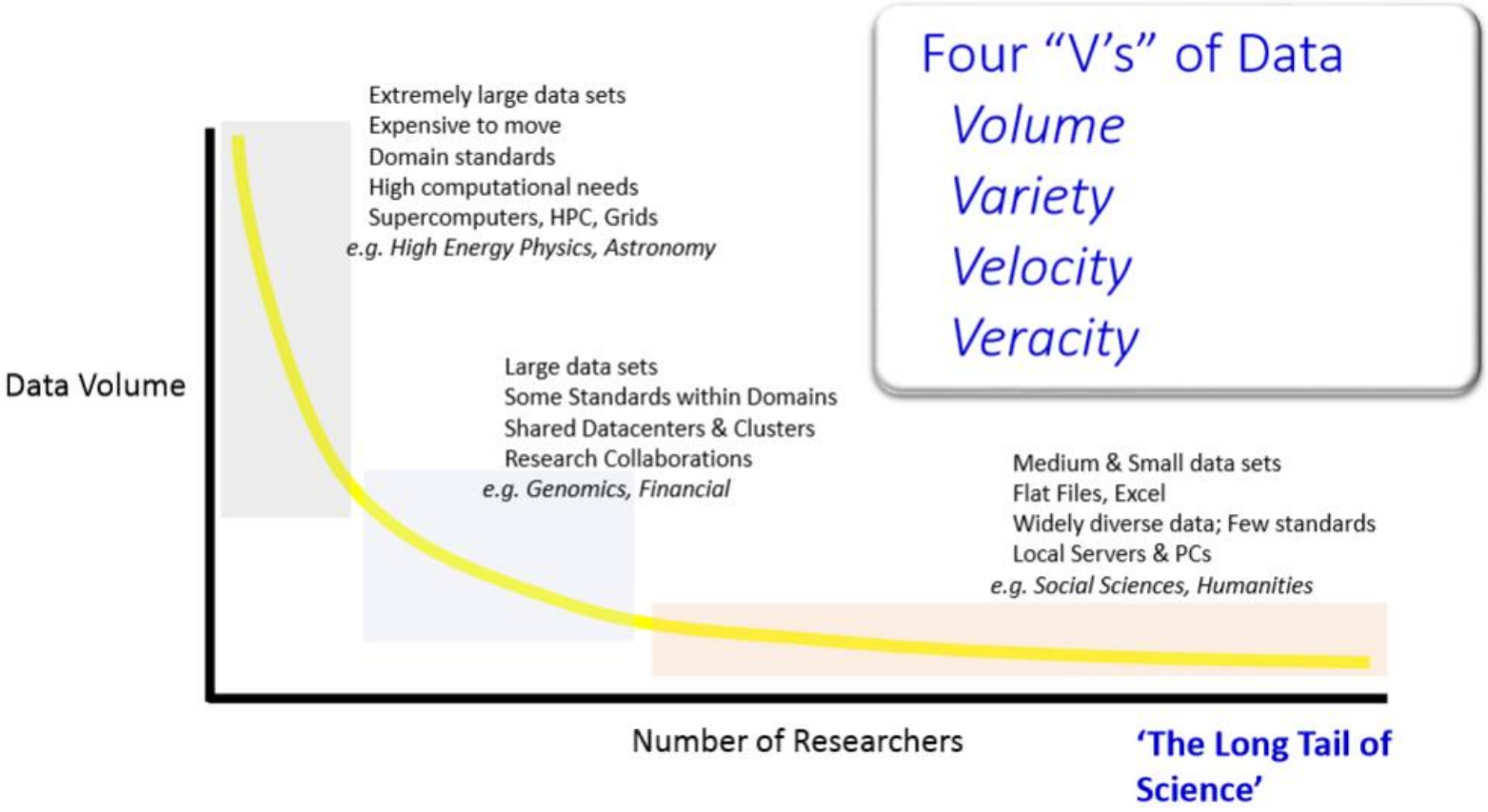
Nottingham university fire destroys new multimillion-pound chemistry building

Photo credits: theguardian.com

1.Risk: where is <u>your</u> data?

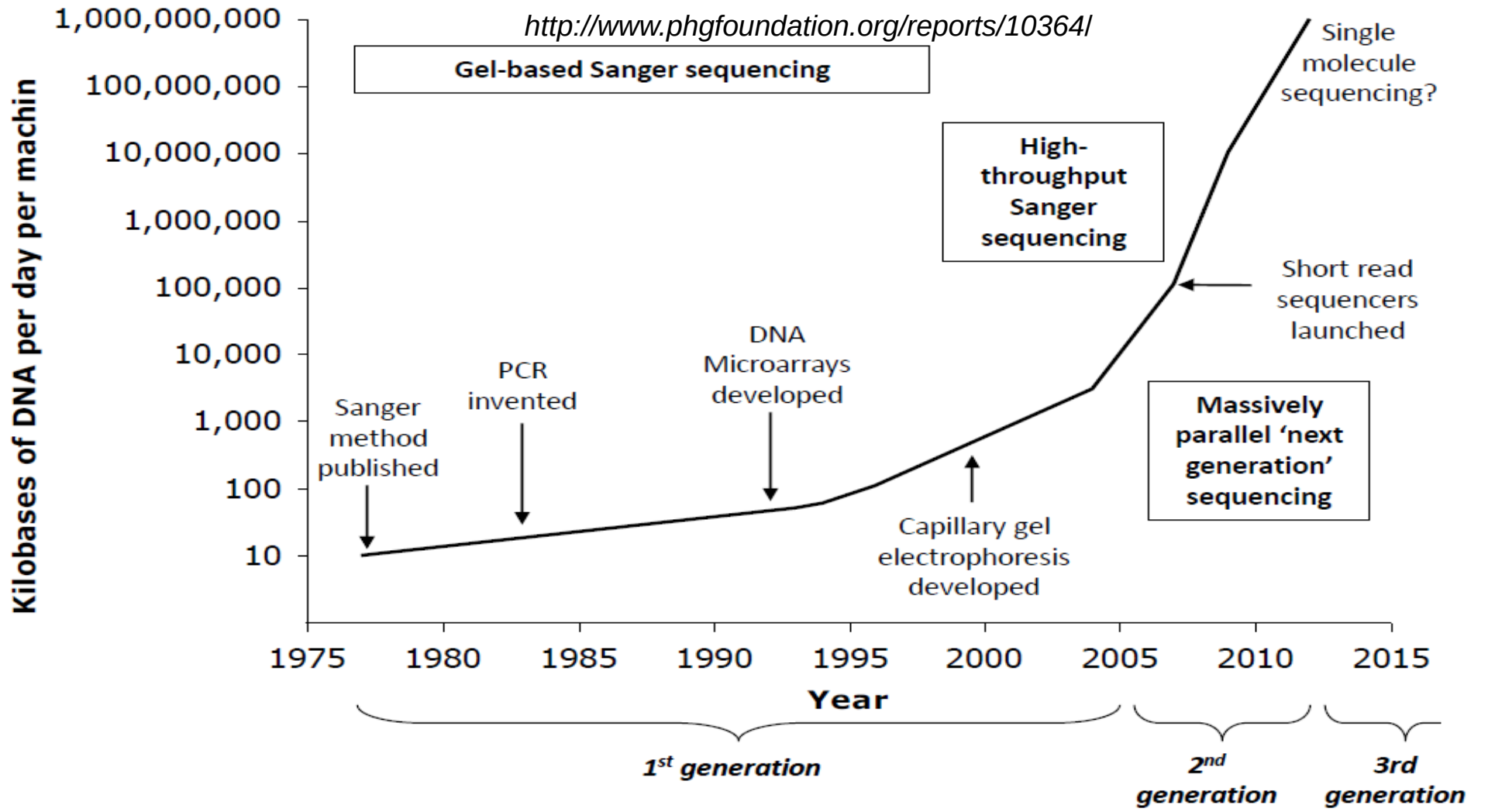
Photo credits: Harvey Rutt http://www.ecs.soton.ac.uk/regenesis/pictures/

Much of Science is now Data-Intensive



Slide : Tony Hey iConference, Berlin 2014

2. Scale: an explosion of data



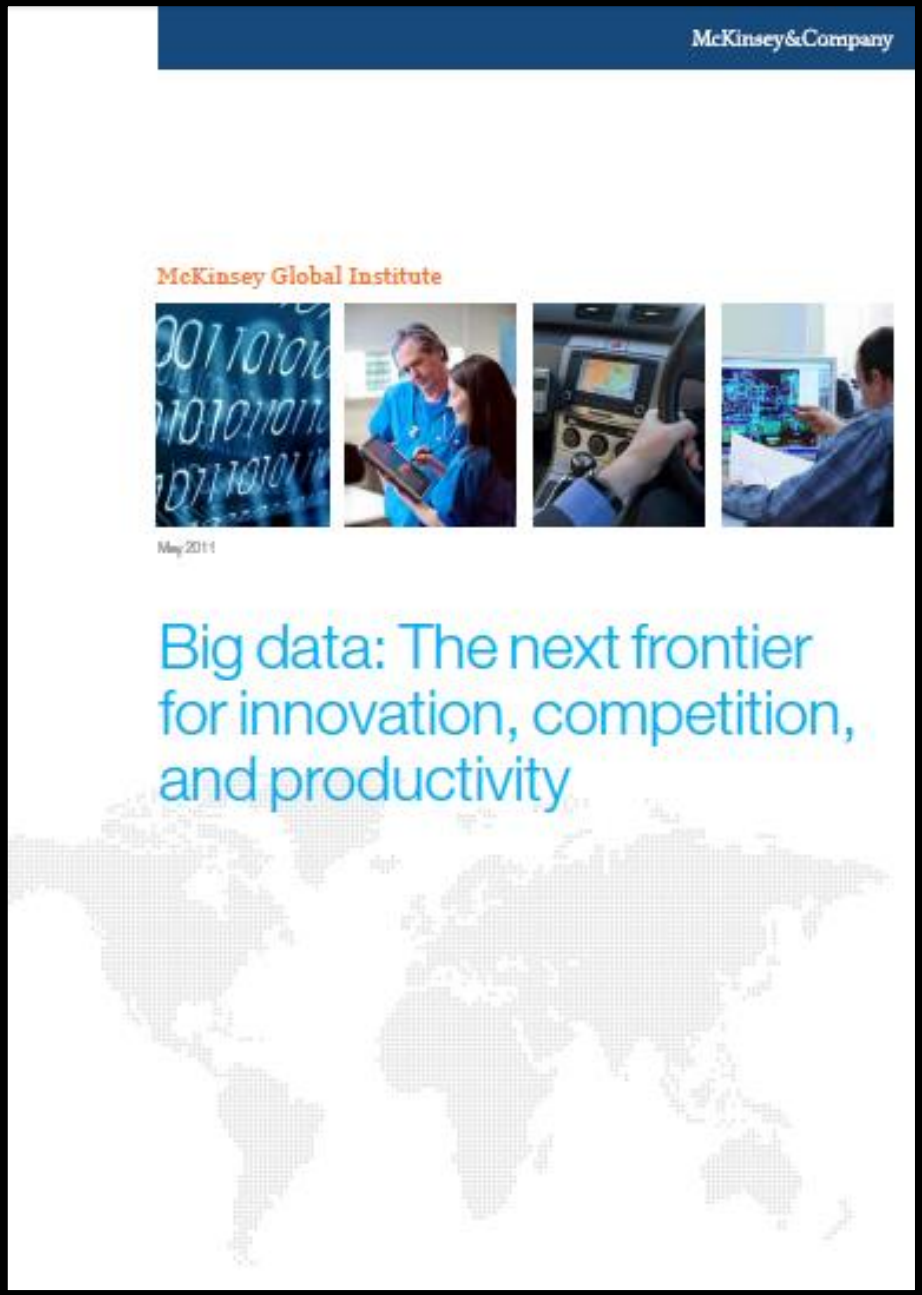
"A single sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project"

Astronomy: Square Kilometre Array

Telescopes in Australia & South Africa

By 2025: Exabytes data / year, 1 PB / day

https://www.skatelescope.org/

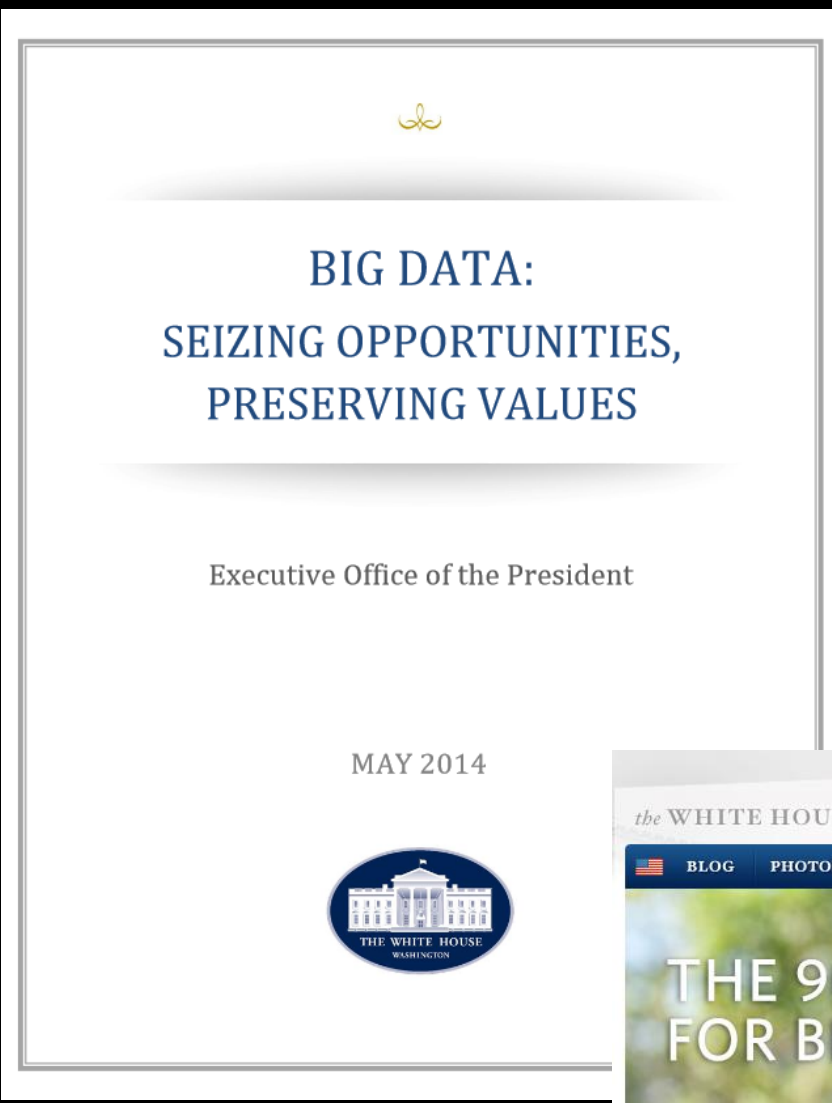


McKinsey Global Institute

Big Data Report 2011

Implications of “Big Data” and data science for organisations in all sectors

http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation



May 2014
Big Data and Privacy

http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

NITRD

The Networking and Information Technology Research and Development (NITRD) Program

Search

HOME | NITRD PROGRAM | NITRD GROUPS | MEMBER AGENCIES | NITRD EVENTS | NITRD INVESTMENTS | PCAST | NCO | LAWS | PUBLICATIONS

Request for Input (RFI) National Big Data R&D Initiative

A Notice by the National Coordinating Office, Networking and Information Technology Research and Development Big Data Senior Steering Group on October 2, 2014

More ▸

1 2

Request for Input (RFI)-National Big Data R&D Initiative

A Notice by the National Coordinating Office, Networking and Information Technology Research and Development Big Data Senior Steering Group on October 2, 2014

Notice in Federal Register (https://www.federalregister.gov)

ACTION: Request for Input (RFI).

The National Big Data R&D Initative

Vision and Actions to be Taken

The following represents a preliminary, draft vision statement and possible actions to be taken as formulated by federal agencies participating in the Networking and Information Technology R&D Program Big Data Senior Steering Group (BDSSG) with input from external stakeholders. This framework outlines a vision and how agencies might move forward achieving this vision, both separately and collaboratively. Within these two thrusts, BDSSG agencies have identified some areas of high interest and potential impact.

VISION STATEMENT: **We envision a Big Data innovation ecosystem in which the ability to analyze, extract information from, and make decisions and discoveries based upon large, diverse, and real-time data sets enables new capabilities for federal agencies and the nation at large; accelerates the process of scientific discovery and innovation; leads to new fields of research and new areas of inquiry that would otherwise be impossible; educates the next generation of 21st century scientists and engineers; and promotes new economic growth. To this end, the NITRD agencies should consider how to most effectively:**

- Create next generation capabilities by leveraging emerging Big Data foundations, technologies, processes, and policies
- In addition to supporting the R&D necessary to create knowledge from data, emphasize support of R&D to understand trustworthiness of data and resulting knowledge, and to make better decisions and breakthrough discoveries and take confident action based on them
- Build and expand access to the Big Data resources and cyberinfrastructure-- both domain specific and shared-- that are needed for agencies to best achieve their mission goals and for the country to innovate and benefit
- Improve the national landscape for Big Data education and training to fulfill increasing demand for both analytical talent and capacity for the broader workforce

The agencies will consider how to create new and enhance existing connections in the current national Big Data innovation ecosystem by, for example:

- Fostering the creation of new partnerships that cross sectors and domains
- Creating new gateways that enable the interconnection and interplay of Big Data ideas and capabilities across agency missions
- Ensuring the long term sustainability, access, and development of high value data sets and data resources

Deadline: 14 November 2014

PSC

The Data Exacell (DXC)

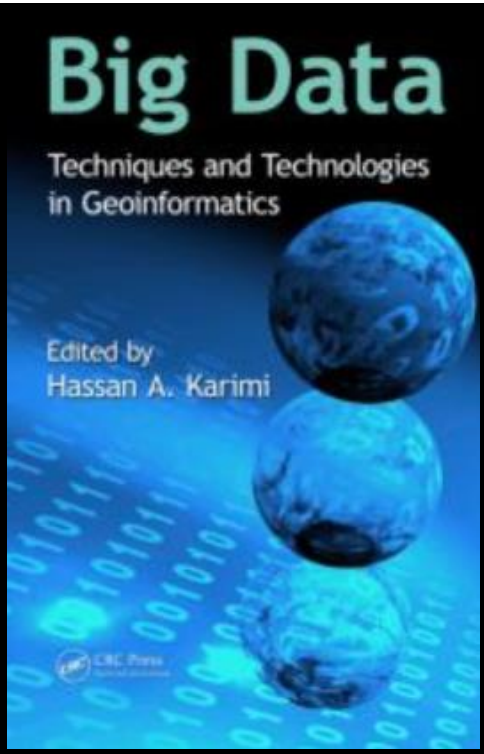
Blacklight: very large shared memory (up to 2×16 TB) for Java, R, Python, MATLAB, C/C++, Fortran, and other Linux-based analytics

Shared storage:
- high bandwidth
- low latency
- high reliability
- high capacity

Sherlock: accelerated graph analytics using RDF and SPARQL or C/C++; for graphs of up to 5 billion edges

DXC

PROJECT TYCHO
DATA FOR HEALTH

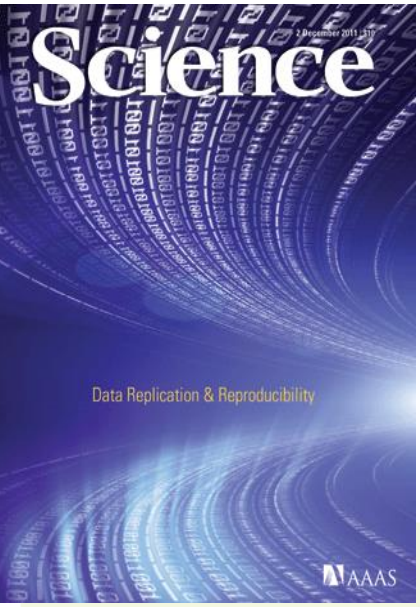


University of Pittsburgh

University of Pittsburgh
Schools of the Health Sciences Media Relations

Pitt Gets $11 Million from NIH to Lead Center of Excellence in National Big Data Research Consortium

3. Research quality: data gold standard



Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Share data & code

Reproducibility Spectrum

Publication only	*Publication +*			Full replication
	Code	Code and data	Linked and executable code and data	

Not reproducible ⟵ ⟶ Gold standard

Fig. 1. The spectrum of reproducibility.

http://www.sciencemag.org/content/334/6060/1226.full.html

“Duke University scandal”

- Anil Potti paper in Nature Medicine 2006
- Independent audit of the research by Baggerly & Coombes (bio-statisticians)
- Duke IRB Inquiry & Report
- Lessons learned include (Ince 2011):
  - *Sloppiness in data curation & software storage*
  - *Duke reviewers did not verify the provenance of the data*
  - *Duke data was not released*
  - *Duke report was not published*

Reproducibility Initiative

Initiative to validate 50 landmark cancer studies

Founding partners:

PLoS, Mendeley, figshare, Science Exchange

$1.3M grant from Laura & John Arnold Foundation

4. Reputation: data access, FOI

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Society › Smoking

Tobacco firm demands university's research on children and smoking

Stirling University fighting attempt by Philip Morris to gain access to research under freedom of information laws

Severin Carrell, Scotland correspondent
guardian.co.uk, Thursday 1 September 2011 15.02 BST
Article history

Philip Morris International, which makes Marlboro cigarettes, has asked for Stirling University's research on teenagers and smoking. Photograph: Paul Sakuma/AP

BBC NEWS

Queen's University has been told to hand over research data

University told to hand over tree ring data - April 15, 2010

G8 Endorses OA

Open Data Charter

Policy Paper

18 June 2013

"Open Data by Default"

Cabinet Office

Policy paper

G8 Open Data Charter and Technical Annex

Published 18 June 2013

Contents

1. Principle 1: Open Data by Default
2. Principle 2: Quality and Quantity
3. Principle 3: Usable by All
4. Principle 4: Releasing Data for Improved Governance
5. Principle 5: Releasing Data for Innovation
6. Technical annex

"One of the things we're doing to fuel more private sector innovation and discovery is to make vast amounts of America's data open and easy to access for the first time in history. "

USA: Obama Administration Executive Order May 2013

PROJECT OPEN DATA

Open Data Policy — Managing Information as an Asset

5. Partnerships

The New York Times

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

Alzheimer's Disease Neuroimaging Initiative: a unique (open) $60M partnership between NIH, FDA, universities and drug companies.

It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

6. Research awards $$$: Funder compliance

Mandate for Data Management Plans

Data Sharing policy

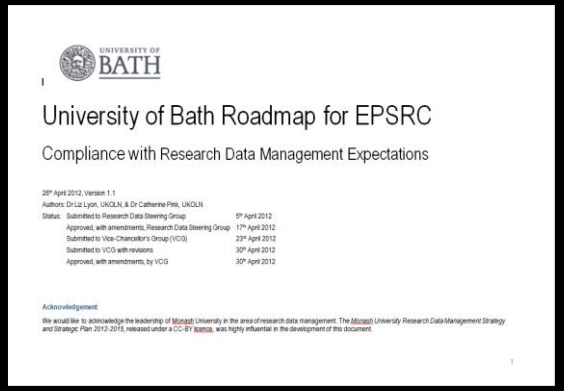
NIH National Institutes of Health
Turning Discovery Into Health

Policy Framework on Research Data

EPSRC

Engineering and Physical Sciences Research Council

- UK Research funder perspective EPSRC
- Unique emphasis on institutional responsibility for RDM (rather than the PI)
- Letter sent to all UK V-Cs in 2011
- Required an RDM Roadmap by May 2012
- And to address expectations…



UNIVERSITY OF BATH

University of Bath Roadmap for EPSRC

Compliance with Research Data Management Expectations

26th April 2012, Version 1.1

Authors: Dr Liz Lyon, UKOLN, & Dr Catherine Pinti, UKOLN

Status: Submitted to Research Data Steering Group 5th April 2012

Approved, with amendments, Research Data Steering Group 17th April 2012

Submitted to Vice-Chancellor's Group (VCG) 23rd April 2012

Submitted to VCG with revisions 30th April 2012

Approved, with amendments, by VCG 30th April 2012

Acknowledgement

We would like to acknowledge the leadership of Monash University in the area of research data management. The *Monash University Research Data Management Strategy and Strategic Plan 2012-2015*, released under a CC BY licence, was highly influential in the development of this document.

Expectations summary

EPSRC
Engineering and Physical Sciences Research Council

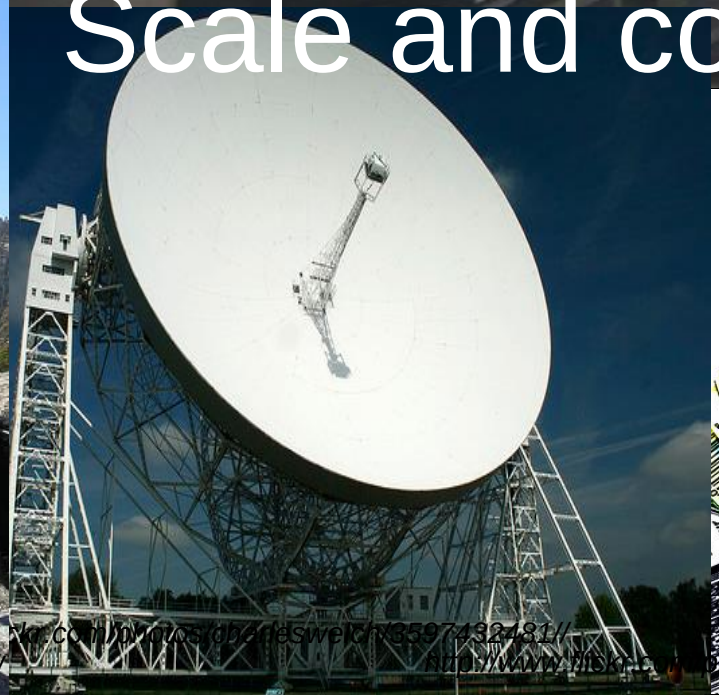
- Awareness of regulatory environment
- Data access statement
- Data policies and processes
- Data storage
- Structured metadata descriptions
- DOIs for data
- Data securely preserved for a minimum of 10 years

Challenges

http://www.google.co.uk/imgres?q=illumina+bgi&hl=en&client=firefox-a&hs=Jl2&rls=org.mozilla:en-GB:official&biw=1366&bih

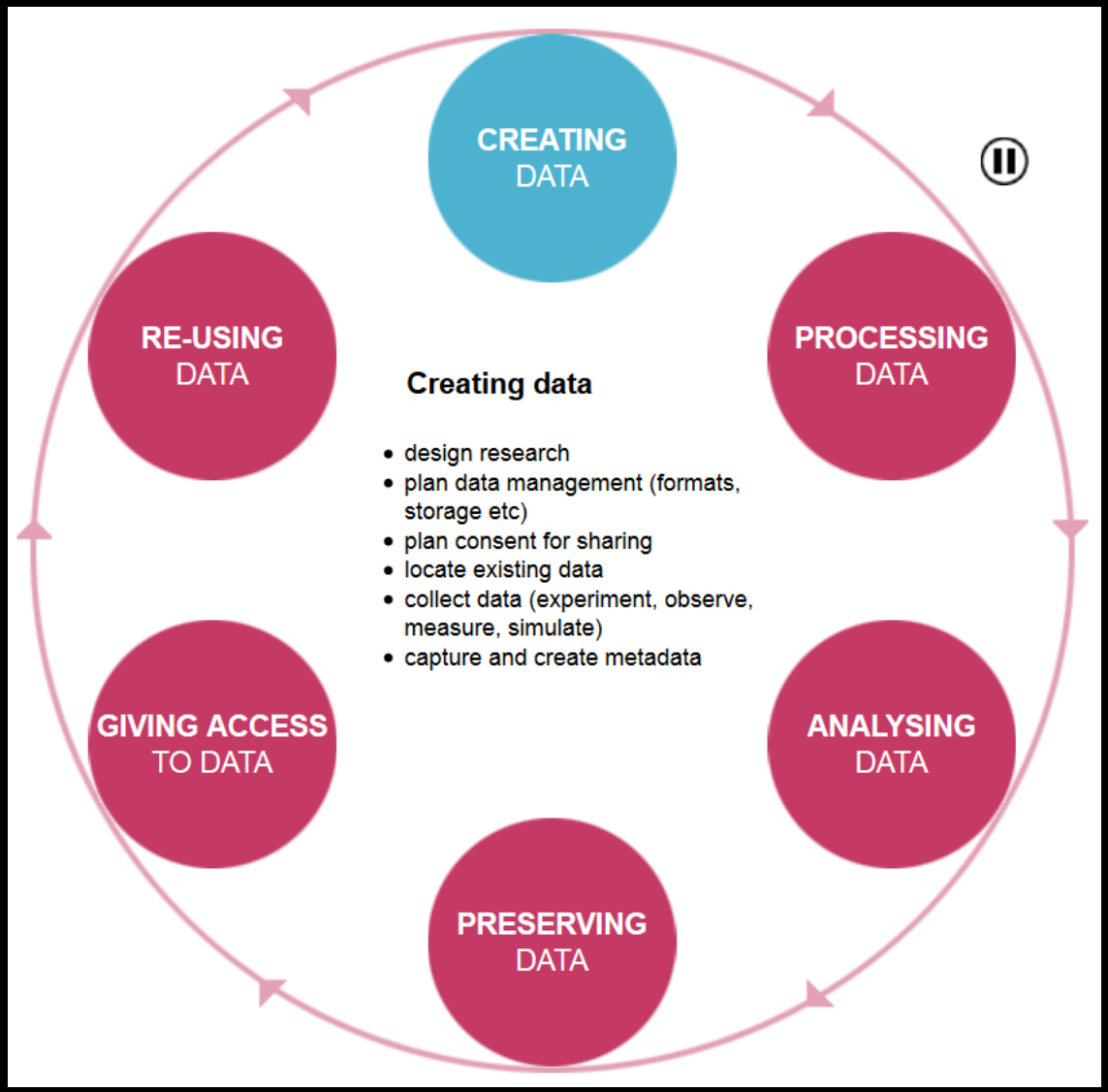
Disciplinary diversity: Scale and complexity

http://www.flickr.com/photos/usfsregion5/4546851916/



http://www.flickr.com/photos/rpgoldman/3597432481/

http://www.flickr.com/photos/wasp_barcode/4793484478/



Research Data Lifecycle

Disciplinary diversity:
data workflow, pipelines, procedures, methodologies

UK Data Archive
http://www.data-archive.ac.uk/create-manage/life-cycle

Large & complex organisation

- 16 Schools / Colleges
- 125 Departments / Programs
- ~200 Centers / Institutes
- 38 Administrative & Business Offices
- 2 Library Systems
- 4,450 f/t Faculty
- 813 p/t Faculty
- > 7000 Staff

Many stakeholders

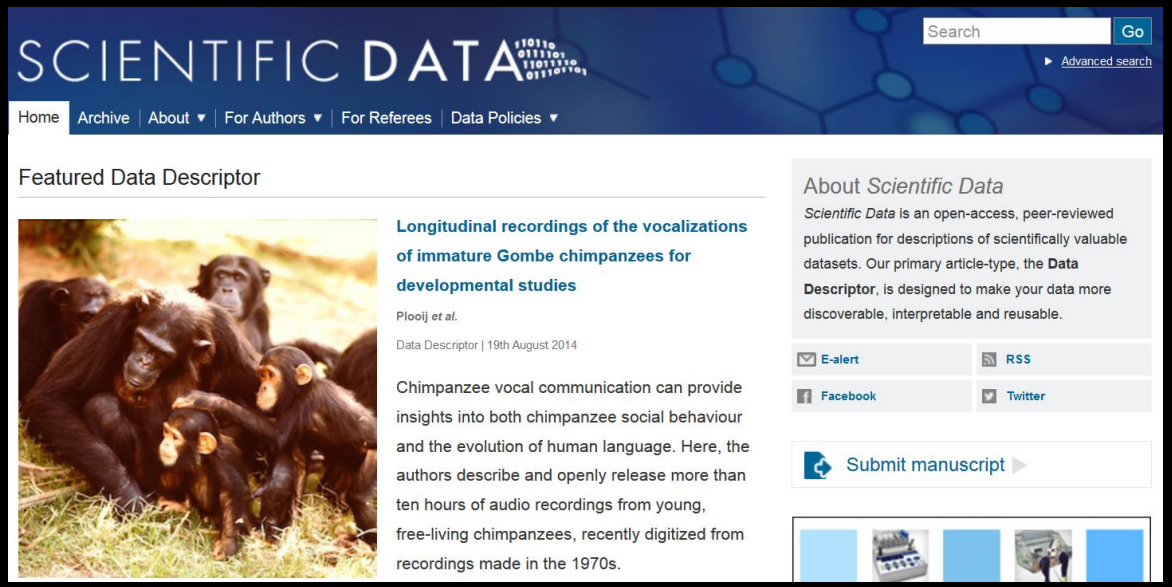
Risk of silos

Co-ordination is vital

Strategic approach

pitt.box.com

Technology Services



Launched in 2014

http://www.nature.com/sdata/

New data requirements

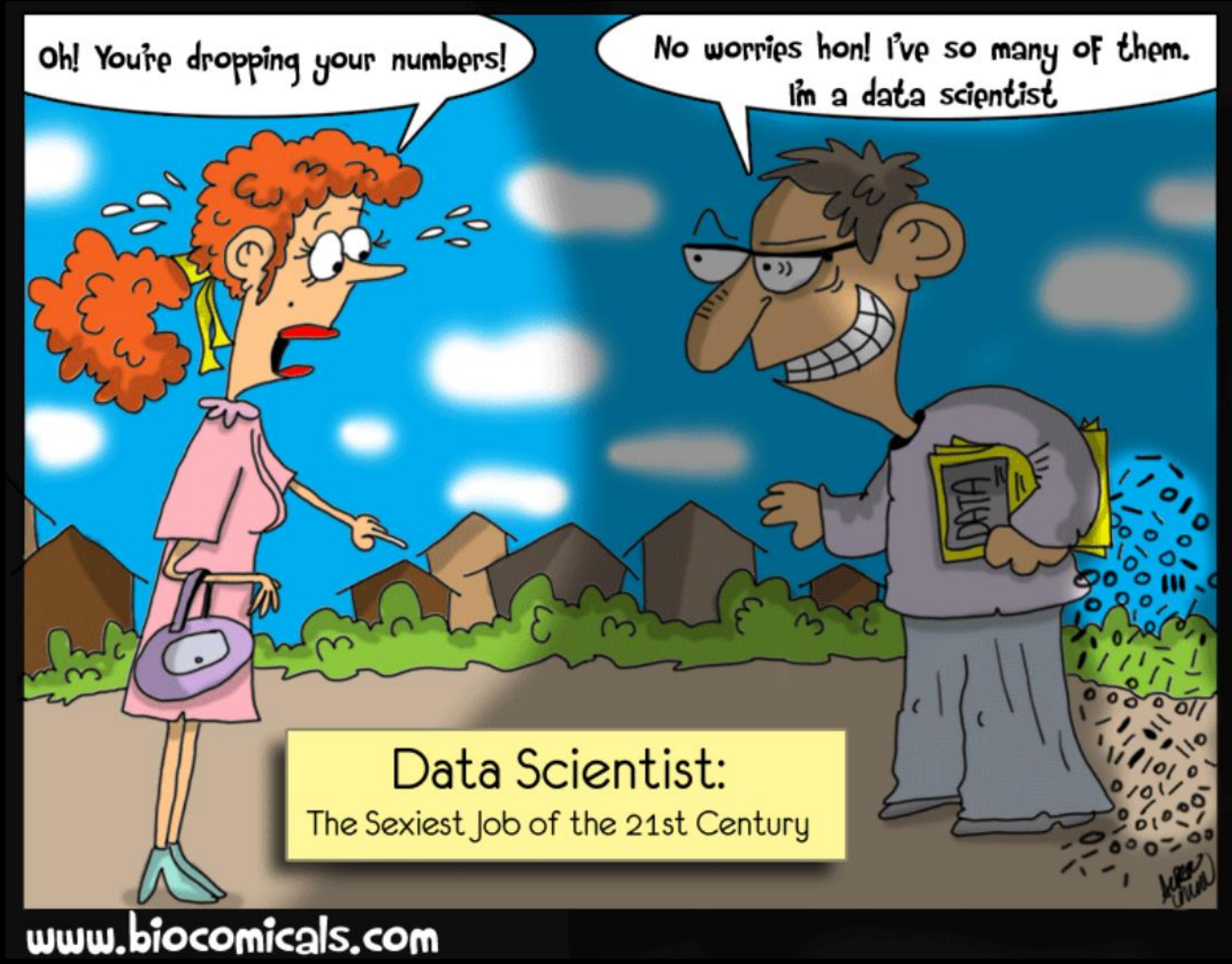
New services

New infrastructure

New tools

New publication platforms

New roles
New skills



…data librarian, research data services manager, data scientist, technical data co-ordinator, data curator, data analyst, data steward, chief data officer....

Developing data capability: talent gap?

HM Government

Seizing the data opportunity

A strategy for UK data capability

Data as a career

As well as ensuring that we equip school leavers and graduates with the key skills, we also need to ensure that data analytics is considered an exciting and rewarding career to pursue – amongst schoolchildren and graduates, but also parents and the wider media.

Career pathways and progression routes

In 2011, the Harvard Business Review published an article referring to the data scientist as the sexiest job of the 21st century[22], yet as described earlier in this chapter, this is an area which is currently experiencing skills shortages.

October 2013

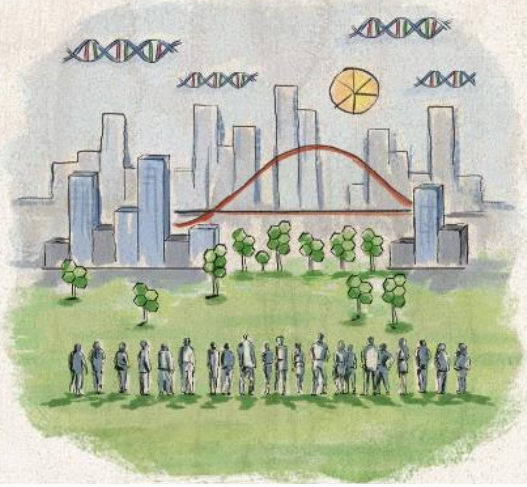
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf

Nesta...

creative skillset
Developing world class talent

MODEL WORKERS

How leading companies are recruiting and managing their data talent



5. Improve the supply of data talent with hybrid skill sets from education

Our research strongly supports the idea that UK universities are failing to produce graduates with the skills mix that data-driven businesses want. Perhaps this should not be a big surprise: there have long been concerns about how funding and organisational factors create hurdles to interdisciplinary teaching and research in UK universities.[37] While, say, the US major/minor system allows students to choose two fields of specialisation when they do a degree, UK students typically specialise in a single field. Anecdotally, one of our interviewees mentioned that PhD programmes in the US are in general broader and less specialised than in the UK too.

http://www.nesta.org.uk/sites/default/files/model_workers_web_2.pdf

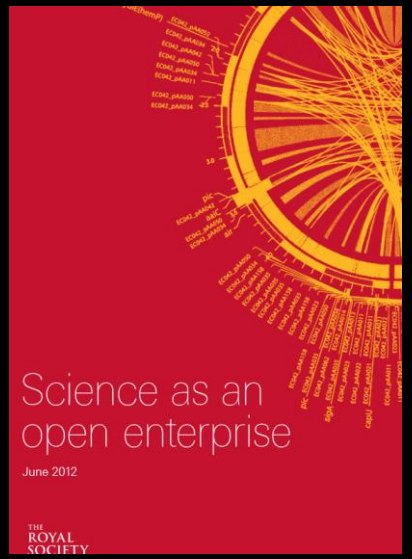
RDM Infrastructure investment $$?

- Long-term costs of data curation are still unclear
- Human & technical resources
- Capacity & capability provision
- Seek from grant funding?
- Internal pump-prime?
- Re-skill, re-align, re-structure?
- Business Case? Pitt RDM Roadmap?

2012 UK Royal Society Report

"Recommendation 2. Universities and research institutes should play a major role in supporting an open data culture by: .....developing a data strategy and their own capacity to curate their own knowledge resources and support the data needs of researchers……."

http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf



Open Science “Champions of Change”

(Culture)

Includes Paul Ginsparg, Stephen Friend, Atul Butte….
June 2013 at the White House

Office of Science and Technology Policy

Opportunities & Benefits

NSF: Products not publications

- NSF policy from 14 January 2013
- List your “research products” not your publications
- “*Acceptable products must be citable and accessible”*
- Datasets, software, electronic lab notebook, patents and publications

- Credit and reward for researchers?
- Build into professional career structures?

http://www.nature.com/nature/journal/v493/n7431/full/493159a.html

Philip Bourne (2005)
Skaggs Professor of Pharmacy, UCSD
Co-Director Protein Data Bank

The structure of human deoxyhemoglobin is one of the most downloaded structures in the PDB—in one year, it has been downloaded more times than the original paper has ever been cited thus far. Yet from the authors' perspective, the Nobel Prize does not come from constructing the PDB database entry, but from an eloquent description of the relationship between structure and function that was presented most completely in the literature. A tenure committee does not award tenure based on the number of deposits a faculty member has made to a biological database, but rather the number of papers they have published in leading journals.

http://www.ploscompbiol.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pcbi.0010034&representation=PDF

Citation advantage - Benefits to authors

Heather Piwowar (2007) & (2013)

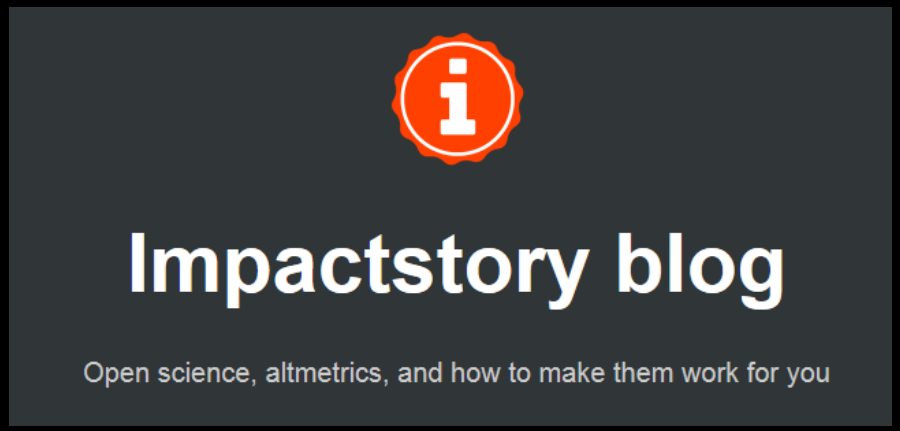
- Explored relationship between citation rate and whether data was publically available
- Study 1 used 85 cancer microarray clinical trials papers published between 1999 and 2003
- *“Cancer clinical trials which share their microarray data were cited about 70% more frequently than clinical trials which do not.”*
- Study 2 used larger sample than 1st study: 10,557 articles between 2001-2009
- *More conservative estimate of citation advantage 9%*

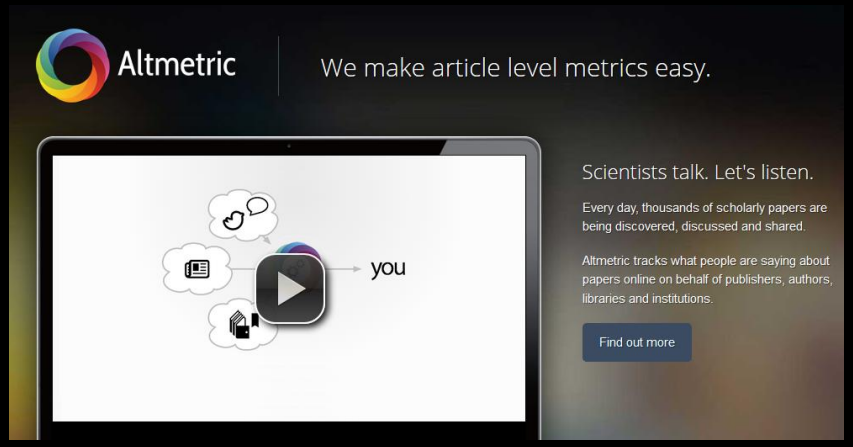
http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0000308&representation=PDF

https://peerj.com/articles/175.pdf

New Data Metrics & Tools

A role for research products like datasets in future tenure decisions?





Institutional profile & rankings

A role for research data in ranking metrics?

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

Timekeeper | Like 20k | Tweet 2,089

"I see a train wreck looming"

Daniel Kahneman

(referring to inability to replicate many priming experiments in psychology)

Better Research

Validated:

data peer review?

CENTER FOR OPEN SCIENCE

Home | About us

Improving the Credibility of Scientific Research: Social Psychology releases special issue of 15 Registered Reports attempting to replicate important results in social psychology

BBC Mobile News | Sport | Weather | iPlayer | TV

NEWS POLITICS

Home | World | UK | England | N. Ireland | Scotland | Wales | Business | Politics | Health | Entertainment & Arts

5 December 2011 Last updated at 21:22

1.3K Share

Everyone 'to be research patient', says David Cameron

New knowledge discovery

"Let me be clear, this does not threaten privacy, it doesn't mean anyone can look at your health records, but it does mean using anonymous data to make new medical breakthroughs.

PharmaTimes online

News | Magazine | JobSearch | Competitions | Events/Meetings | Reports

WORLD NEWS | FREE NEWS HEADLINES | FREE RSS FEED

Where am I? > Article

GSK outlines plan to share patient data online

UK NEWS | MAY 07, 2013

Tweet 0 | g+1 0 | Share 11

BEN ADAMS

GlaxoSmithKline is to establish a new online system allowing researchers access to patient level data from the firm's clinical trials.

This comes after the UK-based firm signed the AllTrials register in February, signalling its commitment to open up access to clinical trial data, and promising to make clinical study reports available.

But this new system goes beyond its AllTrials commitment, with th system, available at: https://clinicalstudydata.gsk. researchers the ability to rec anonymised patient level da behind the results of clinical

Related Links

GSK backs AllTrials campaign

AllTrials campaign raises the game on clinical-trial transparency

Pharma should not publish raw data, says former ABPI head

May 2013 GlaxoSmithKline sharing patient data from clinical trials

- De-identified patient data
- >200 studies from 2007-date
- Raw data + analysis-ready data provided to regulatory authorities

http://www.nejm.org/doi/full/10.1056/NEJMsr1302541

The NEW ENGLAND JOURNAL of MEDICINE

- Clinical study documents incl. protocols, analysis plan, methods, data specifications, variables…

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS

SPECIAL REPORT

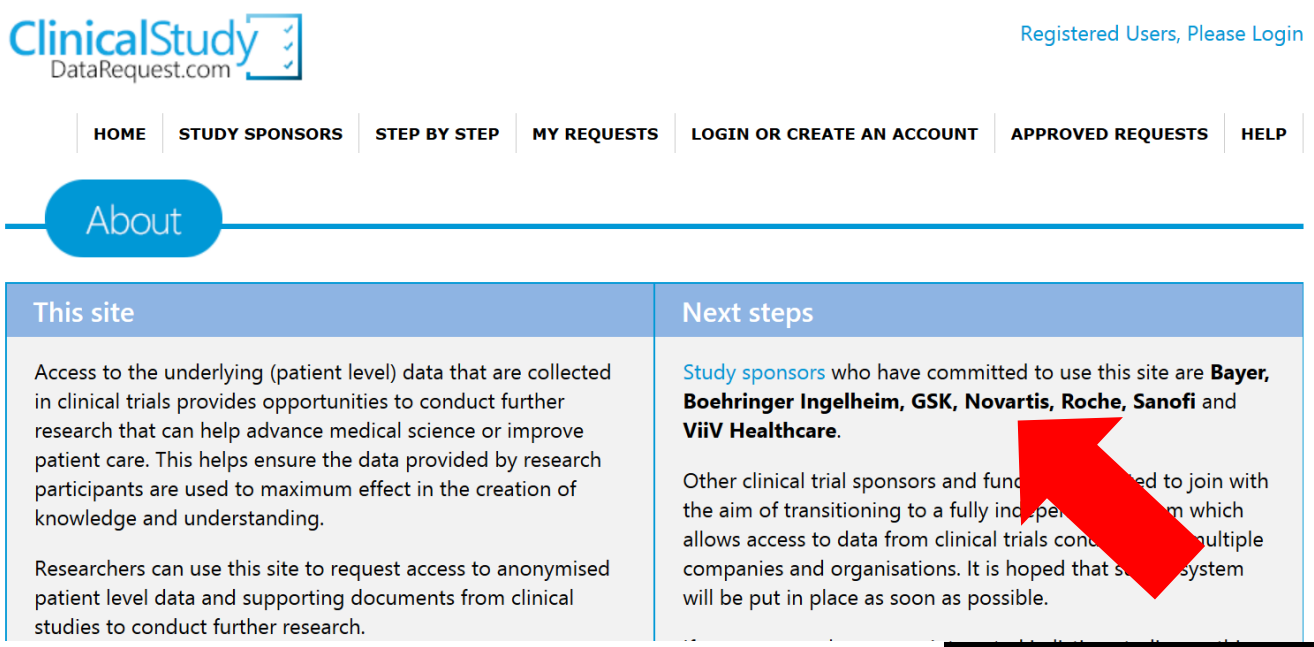
Access to Patient-Level Data from GlaxoSmithKline Clinical Trials

Perry Nisen, M.D., Ph.D., and Frank Rockhold, Ph.D.

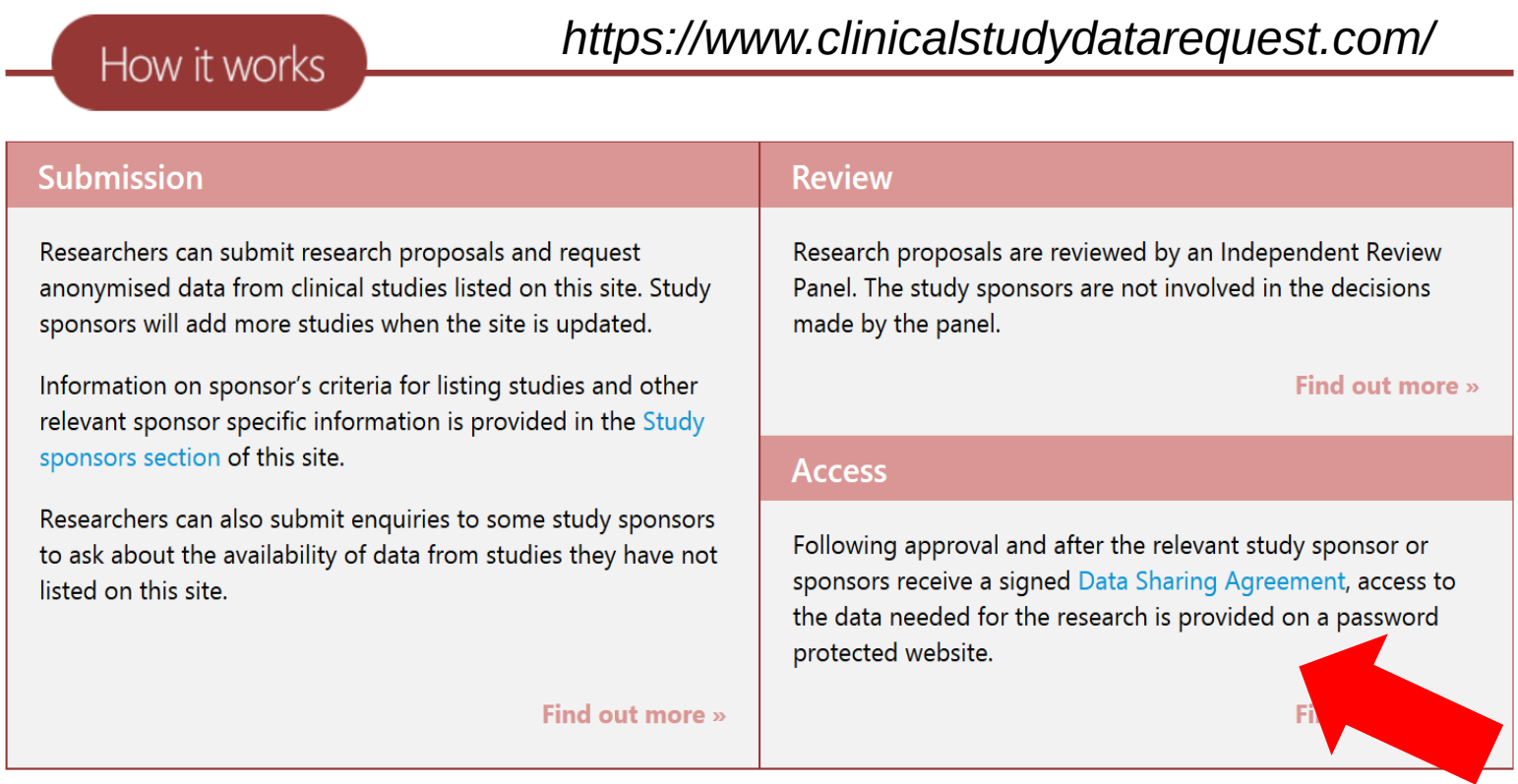
N Engl J Med 2013; 369:475-478 | August 1, 2013 | DOI: 10.1056/NEJMsr1302541

2014 Multi-sponsor data request site

- Independent review panel for requests
- Data Sharing Agreement



https://www.clinicalstudydatarequest.com/



Data access for 12 months

Privacy & confidentiality controls

Analysis software provided (R, SAS)

For the institution

- RDM infrastructure supports research and researchers (it could save you time)
- Economies of scale & efficiencies achieved through co-ordinated and shared services
- Legacy data is curated and preserved as part of the Scientific Record
- Data Management Plans inform annual strategic planning & investment decisions
- New-entrant researchers are “data-savvy” through RDM advocacy & training programs

Philip Bourne (2014)
NIH Associate Director for Data Science

"Big Data represents the emergence of the digital enterprise – the ability for an organization to take full advantage of its digital assets – which collectively can be described as large amounts of data and more."

What *Big Data* means to me (Editorial)
http://jamia.bmj.com/content/21/2/194.full.pdf+html

University of Pittsburgh

SCHOOL OF
Information
Sciences

Thank you….

Senate Plenary Meeting, October 2014

Professor Liz Lyon, School of Information Sciences, University of Pittsburgh