Gaining maximum value out of the rising tide of data

Keith Russell

24 October 2017

ands

AUSTRALIAN NATIONAL DATA SERVICE

RDS www.rds.edu.au

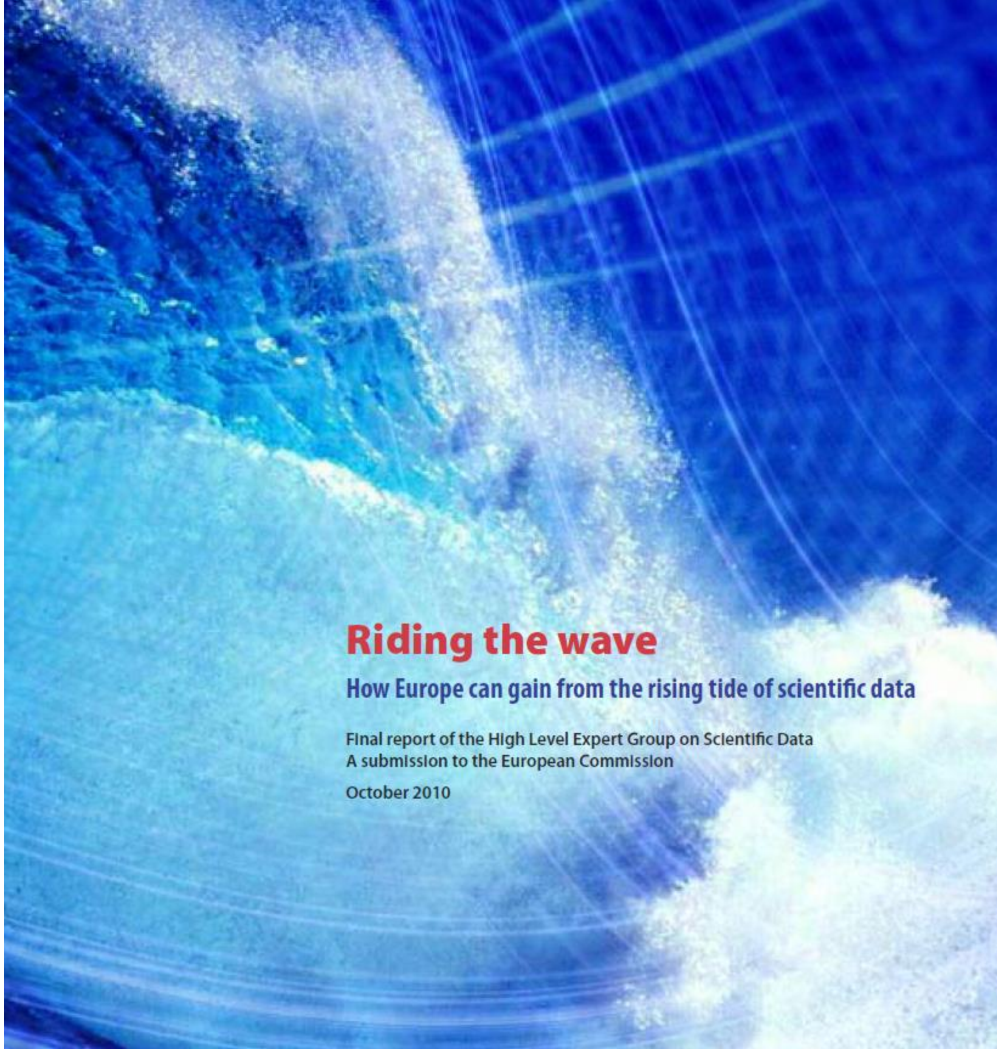
ands AUSTRALIAN NATIONAL DATA SERVICE

nectar

Rising tide of data

‘A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge.’

https://ec.europa.eu/eurostat/cros/content/riding-wave_en

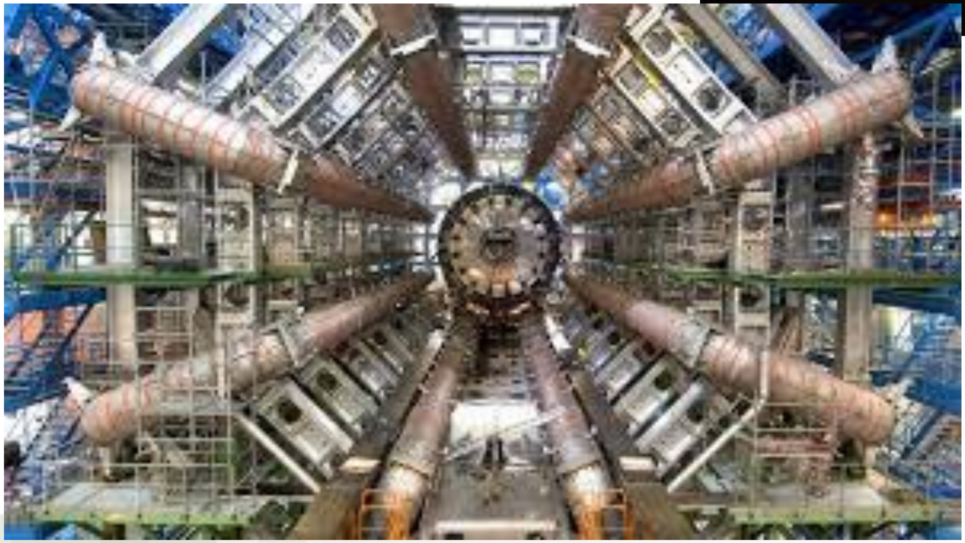


Riding the wave

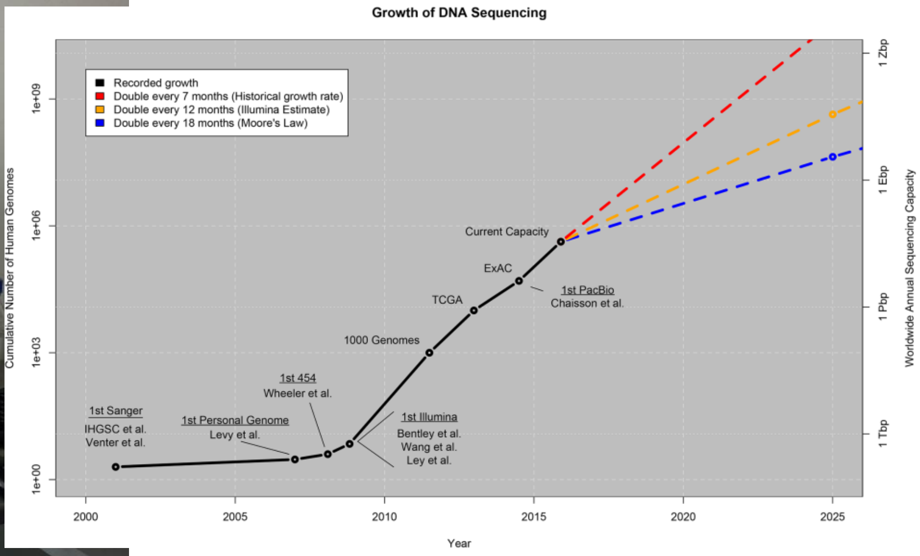
How Europe can gain from the rising tide of scientific data

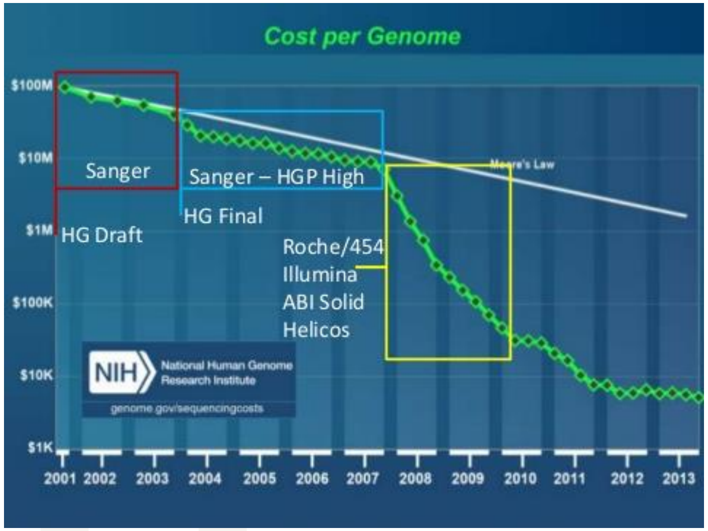
Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission

October 2010

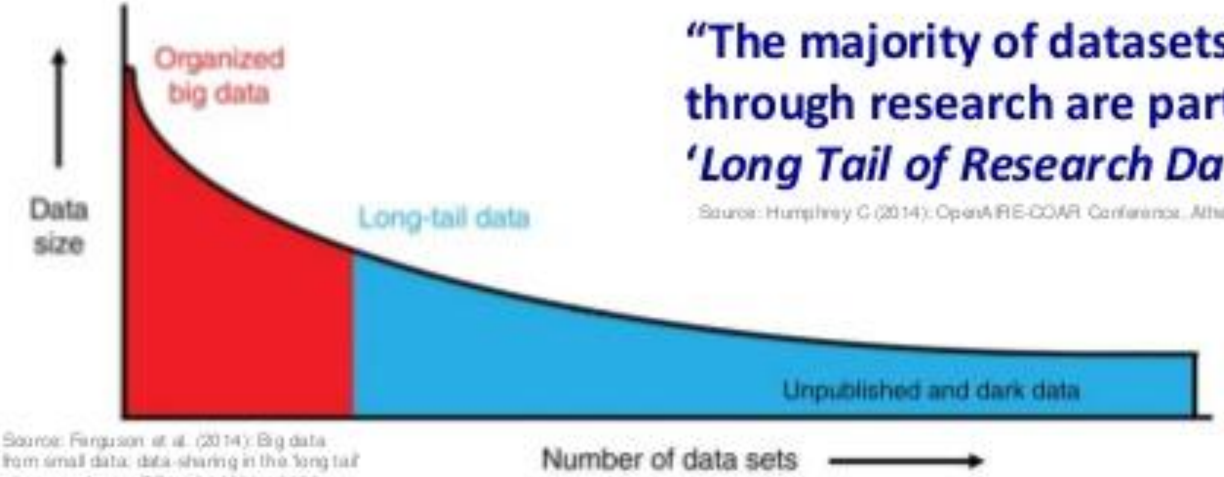


MiSeq
illumina


Growth of DNA Sequencing
⬛ Recorded growth
🟥 Double every 7 months (Historical growth rate)
🟧 Double every 12 months (Illumina Estimate)
🟦 Double every 18 months (Moore's Law)
Cumulative Number of Human Genomes
1e+00
1e+03
1e+06
1e+09
Worldwide Annual Sequencing Capacity
1 Tbp
1 Pbp
1 Ebp
1 Zbp
Year
2000
2005
2010
2015
2020
2025
1st Sanger
IHGSC et al.
Venter et al.
1st Personal Genome
Levy et al.
1st 454
Wheeler et al.
1st Illumina
Bentley et al.
Wang et al.
Ley et al.
1000 Genomes
TCGA
ExAC
1st PacBio
Chaisson et al.
Current Capacity


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Sanger
Sanger – HGP High
HG Draft
HG Final
Moore's Law
Roche/454
Illumina
ABI Solid
Helicos
NIH National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

RESEARCH DATA - “Long Tail“



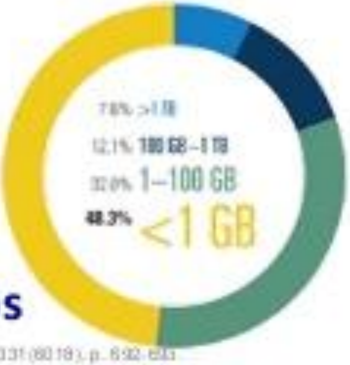
Source: Ferguson et al. (2014): Big data from small data: data-sharing in the 'long tail' of neuroscience. DOI: 10.1038/nn.3838

“The majority of datasets produced through research are part of the ‘*Long Tail of Research Data*’”

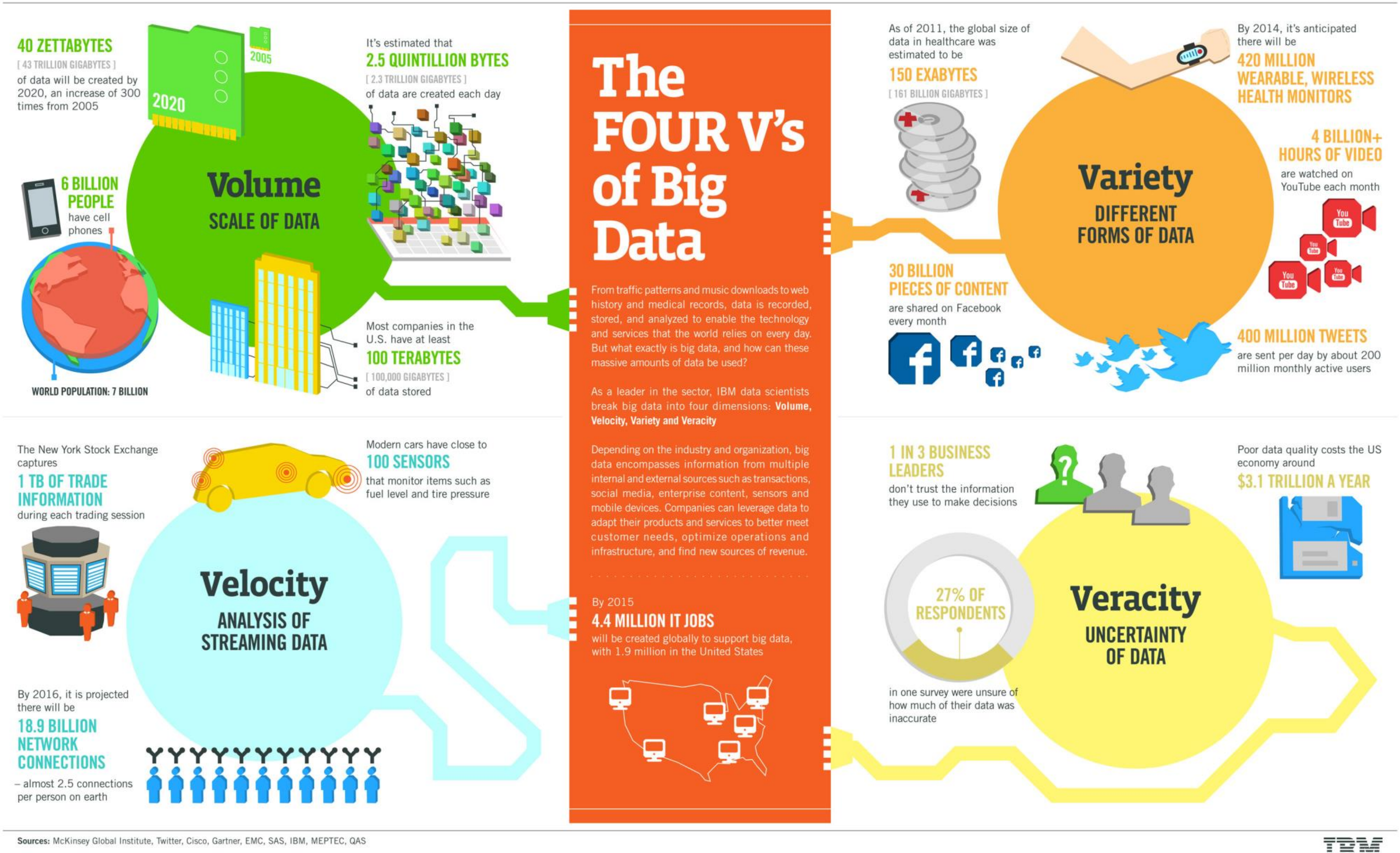
Source: Humphrey C (2014): OpenAIRE-COAR Conference, Athens

Science Survey 2011:

- **48 %** of respondents were working with datasets that were **<1GB in size**
- **50 % stored data exclusively! in labs**



Source: Science (2011): 331 (6018), p. 692-693
DOI: 10.1126/science.331.6018.692



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

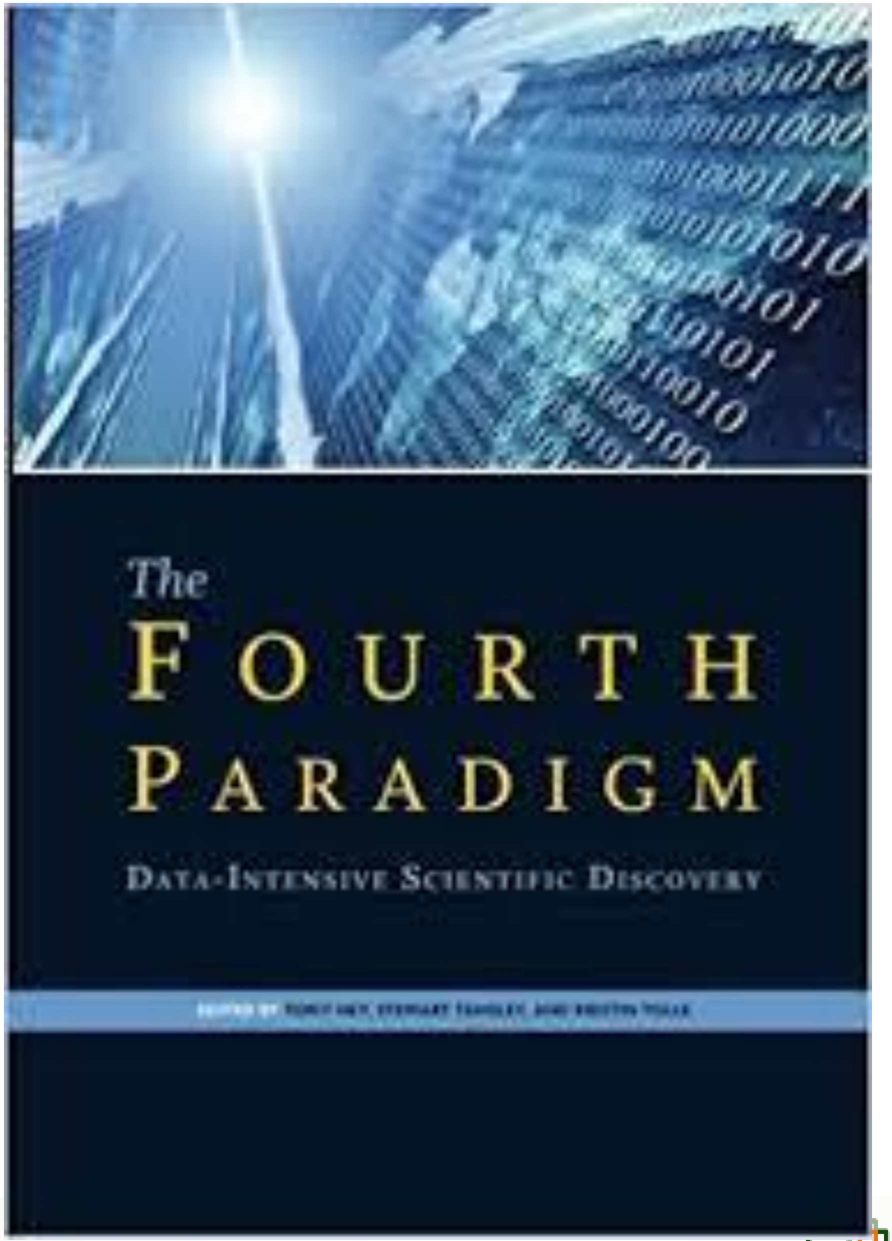
www.ibmbigdatahub.com/infographic/four-vs-big-data

ands
AUSTRALIAN NATIONAL DATA SERVICE

Fourth Paradigm

‘Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.’

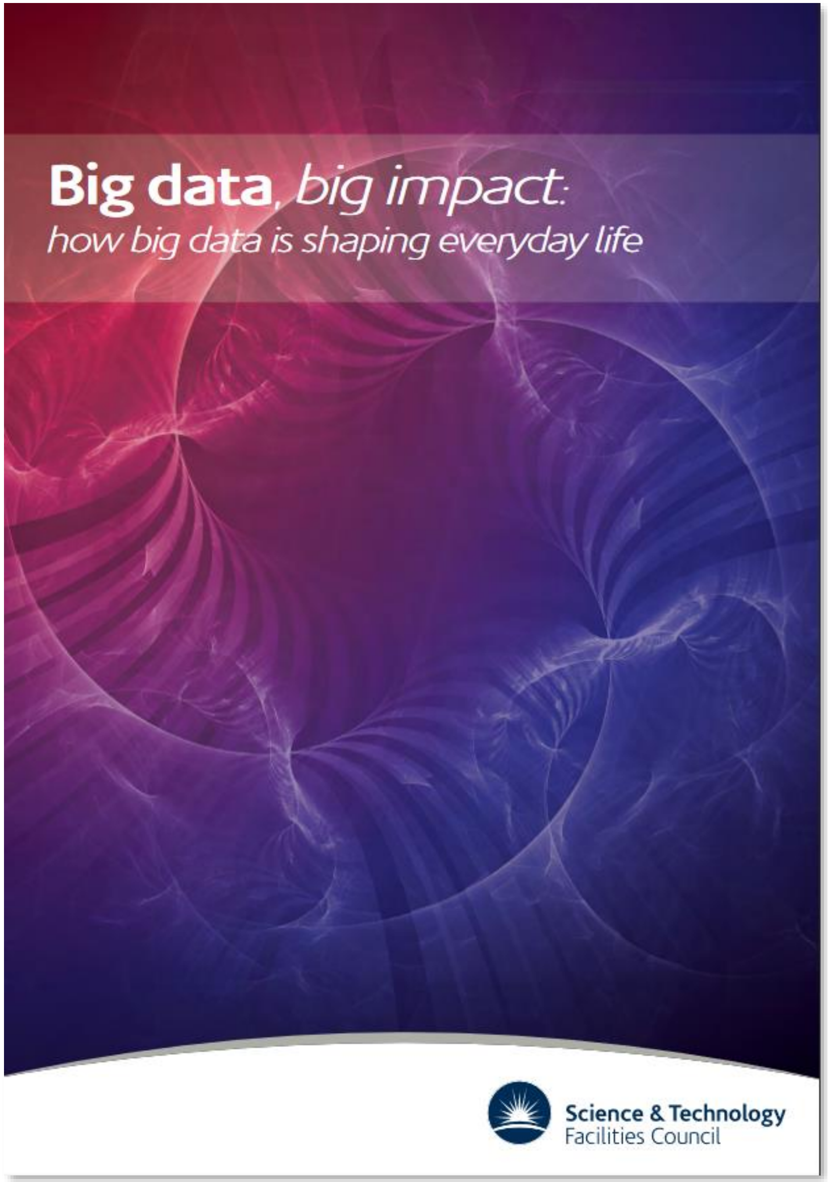
https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Big Data examples

‘Big data is helping us to learn more about the Universe we live in, and to answer some fundamental questions. Reaping all of the benefits that big data offers us means constant innovation in computing and communications.’

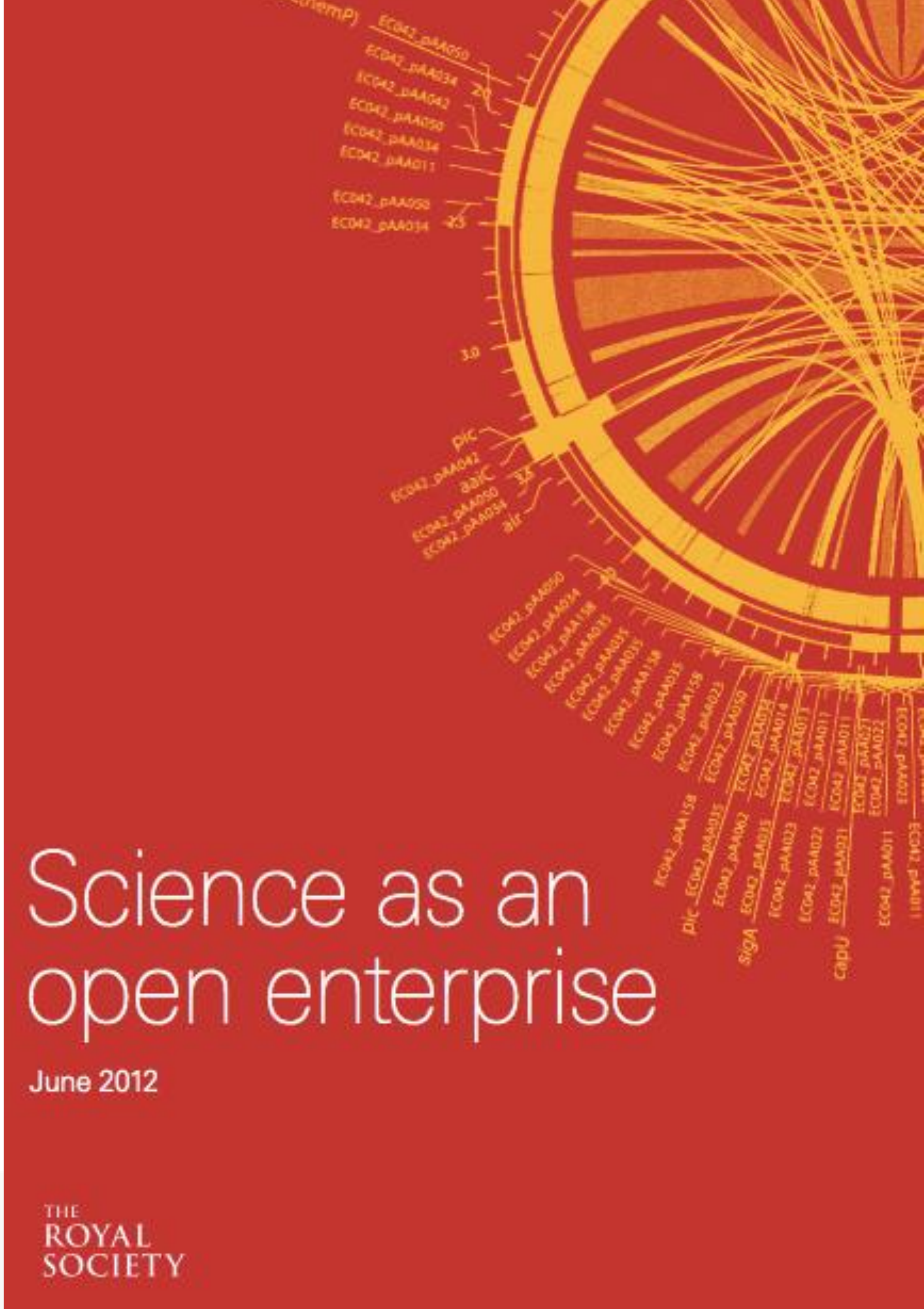
http://www.stfc.ac.uk/files/impact-publications/big-data-big-impact/


Big data, big impact:
how big data is shaping everyday life
Science & Technology Facilities Council

Changing attitudes to data

‘Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories - and of the experimental and observational data on which they are based - permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge. Science’s powerful capacity for self-correction comes from this openness to scrutiny and challenge.’

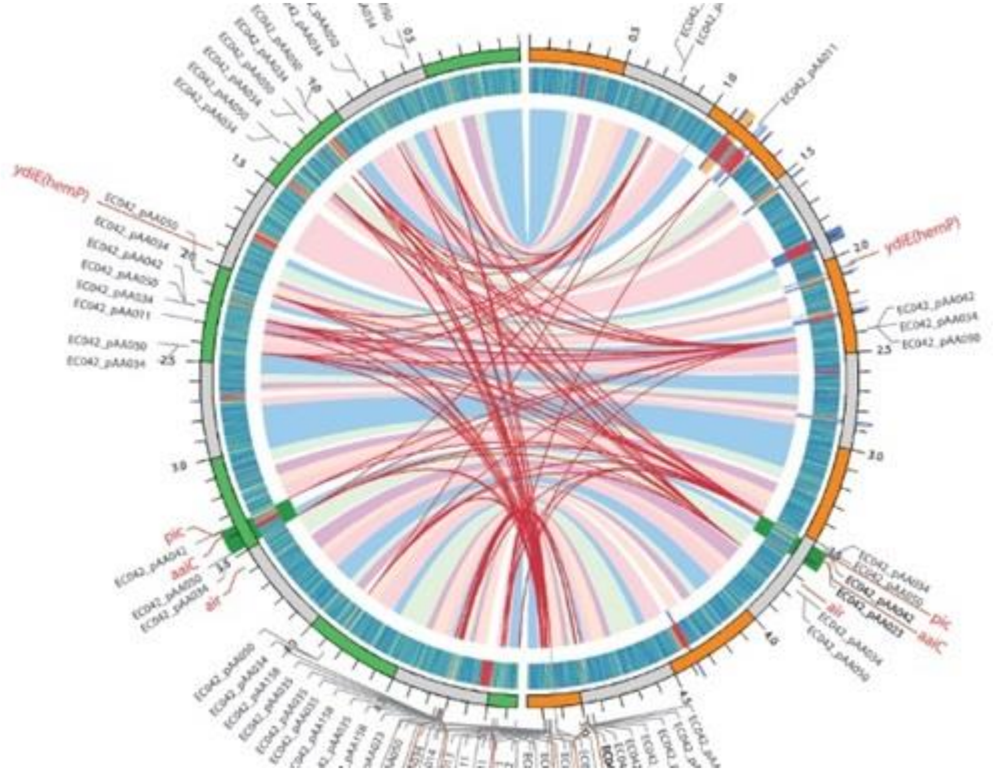
https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/



Science as an open enterprise

June 2012

THE
ROYAL
SOCIETY

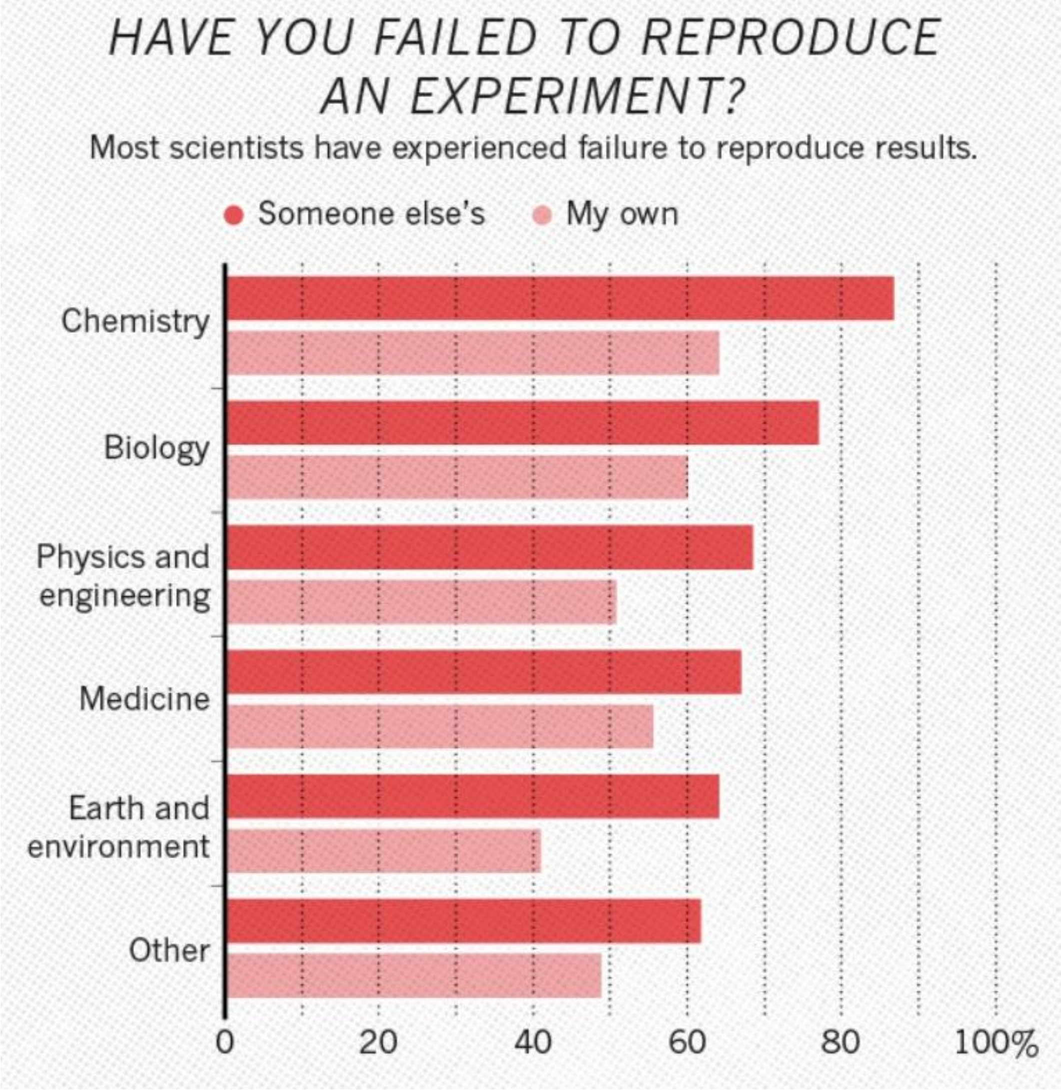


The Spanish Cucumber E. Coli. This genome was analysed within weeks of its outbreak because of a global and open effort; data about the strain’s genome sequence were released freely over the internet as soon as they were produced.

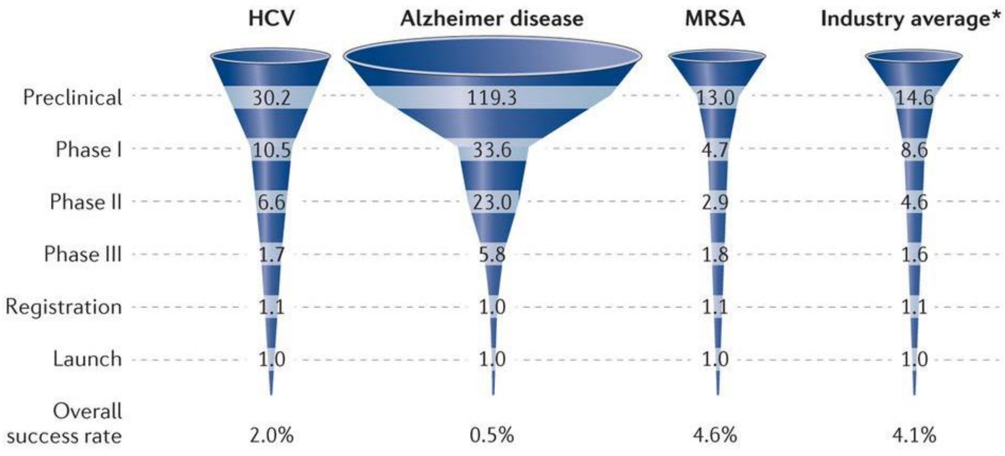
Reproducibility crisis

Science appears to have an issue with reproducibility. A survey by Nature revealed that 52% of researchers believed there was a “significant reproducibility crisis” and 38% said there was a “slight crisis”.

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

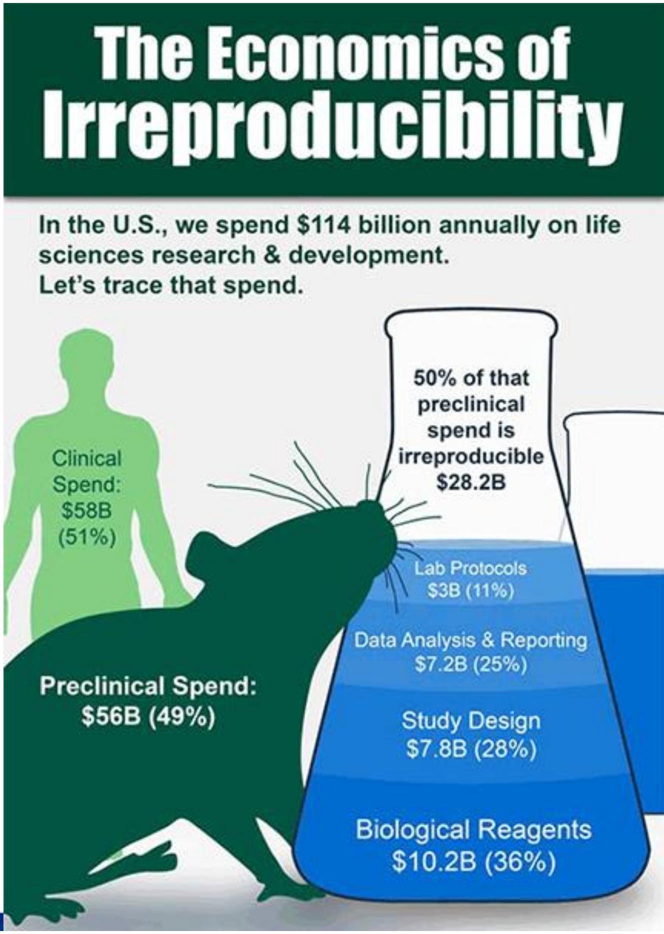


Productivity & Irreproducibility



Nature Reviews | Drug Discovery

Paul et al. (Nature Rev. Drug Discov. 9, 203~~2~~14; 2010
Calcoen D, Elias L, Yu X. (Nature Rev. Drug Discov. 14. 161-2; 2015



The Economics of Irreproducibility

In the U.S., we spend $114 billion annually on life sciences research & development.
Let's trace that spend.

Clinical Spend: $58B (51%)
50% of that preclinical spend is irreproducible $28.2B
Lab Protocols $3B (11%)
Data Analysis & Reporting $7.2B (25%)
Preclinical Spend: $56B (49%)
Study Design $7.8B (28%)
Biological Reagents $10.2B (36%)

Research data as a valued output

- Funders are seeing research data a publishable output
- Journals are requesting data alongside the article
- They expect data to be managed (Code for responsible conduct of research)
- They expect it to be available for further research

Australian Government
Australian Research Council

NIH
National Institutes of Health

HORIZON 2020

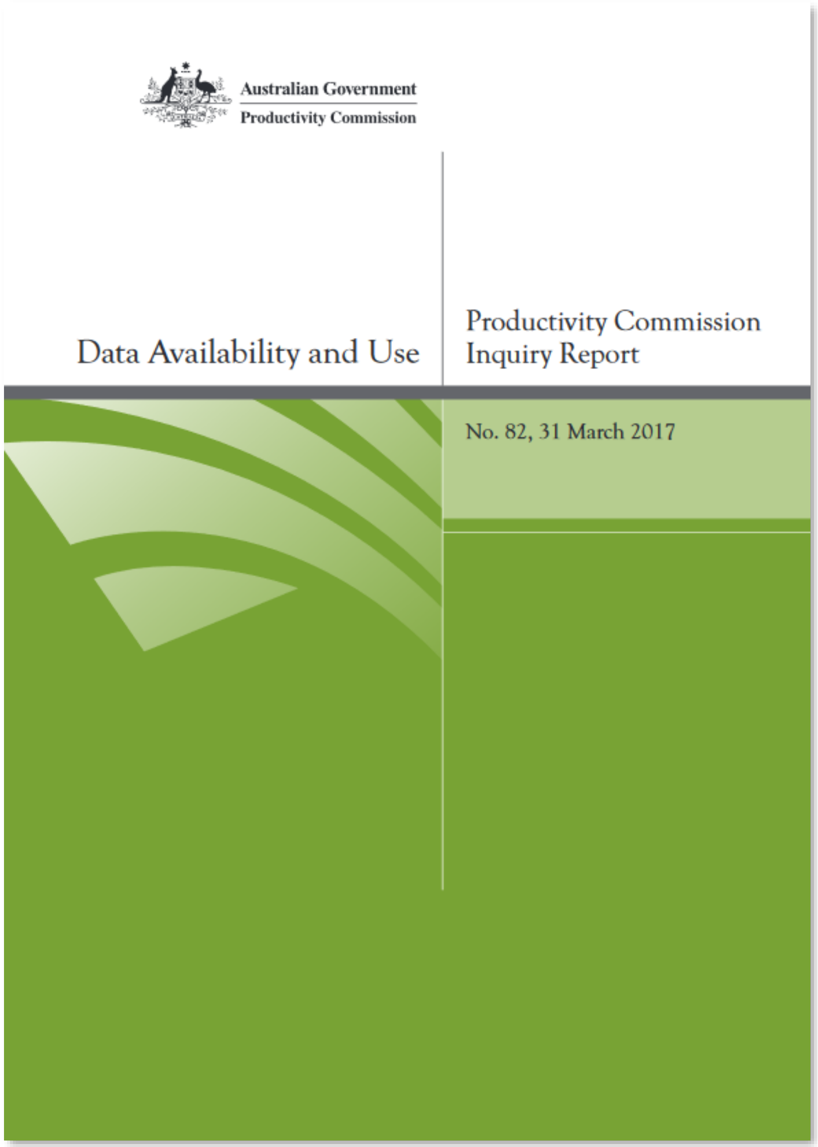
nature
International weekly journal of science

PLOS

Productivity commission report

Extraordinary growth in data generation and usability has enabled a kaleidoscope of new business models, products and insights. Data frameworks and protections developed prior to sweeping digitisation need reform. This is a global phenomenon and Australia, to its detriment, is not yet participating.
Improved data access and use can enable new products and services that transform everyday life, drive efficiency and safety, create productivity gains and allow better decision making.

https://www.pc.gov.au/inquiries/completed/data-access/report



Australian Government
Productivity Commission

Data Availability and Use

Productivity Commission
Inquiry Report

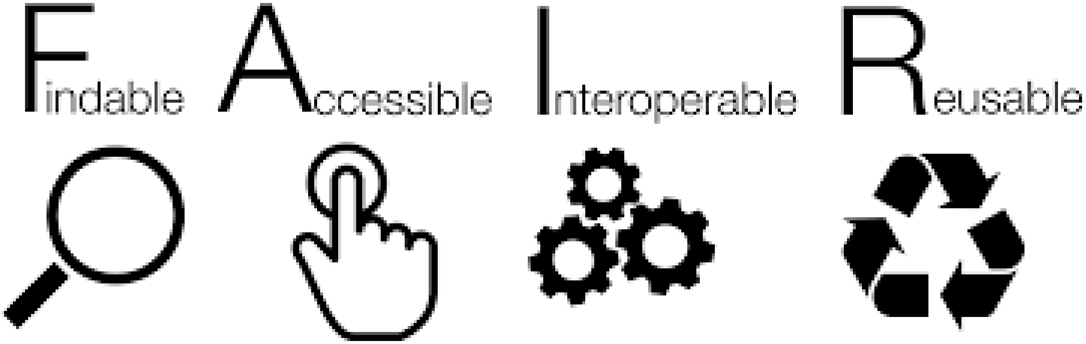
No. 82, 31 March 2017

Four transformations

- Building a data advantage
- Innovative approaches and tools
- Increase (inter)national collaboration
- Translating research outcomes

Requires FAIR data

What are the FAIR data principles?



https://www.force11.org/group/fairgroup/fairprinciples

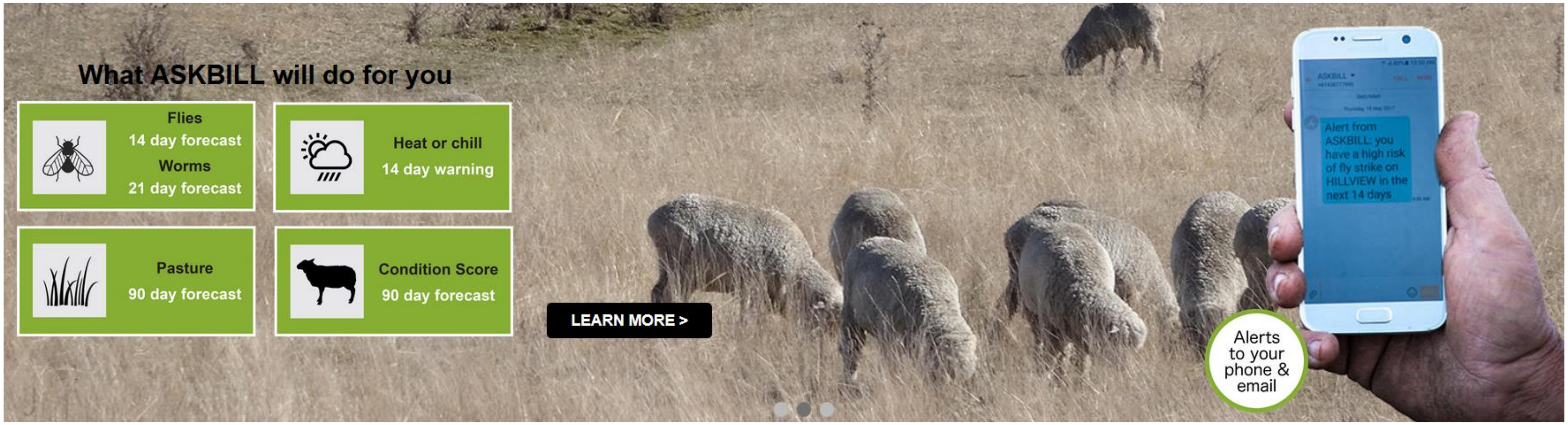
PetaJakarta project

The project was acknowledged by the US Government when their Federal Register cited SMART's PetaJakarta.org project as an example of best practice for using crowdsourced information in an emergency situation.

http://smart.uow.edu.au/projects/petajakarta-org/index.html

SheepCRC
Ramselect and AskBill

RamSelect Plus



What ASKBILL will do for you

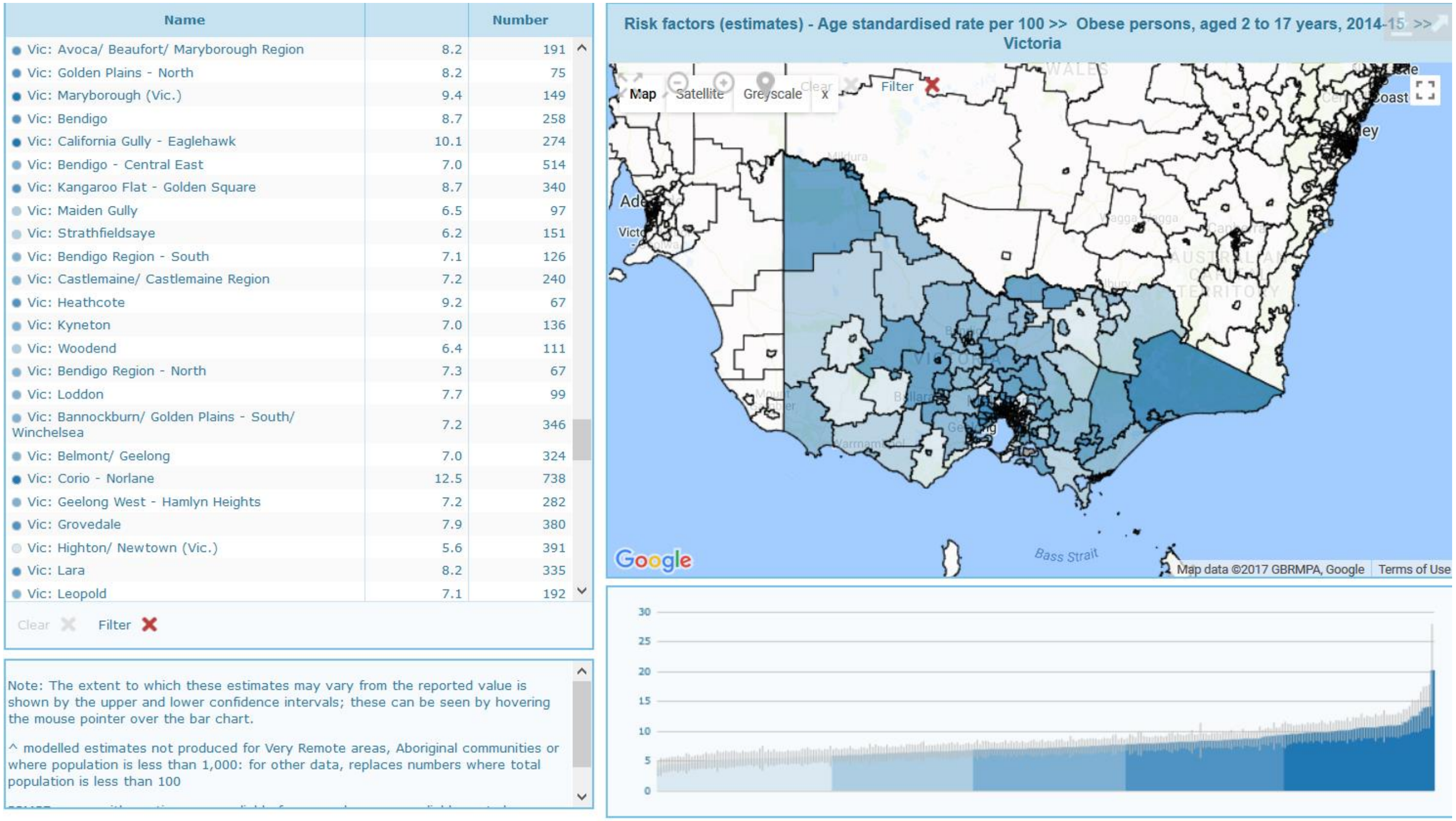
- Flies 14 day forecast, Worms 21 day forecast
- Heat or chill 14 day warning
- Pasture 90 day forecast
- Condition Score 90 day forecast

ASKBILL

Alert from ASKBILL: you have a high risk of fly strike on HILLVIEW in the next 14 days

http://www.farmonline.com.au/story/4667853/askbill-answers-producers-most-important-questions/

Health Tracker project



Name		Number
Vic: Avoca/ Beaufort/ Maryborough Region	8.2	191
Vic: Golden Plains - North	8.2	75
Vic: Maryborough (Vic.)	9.4	149
Vic: Bendigo	8.7	258
Vic: California Gully - Eaglehawk	10.1	274
Vic: Bendigo - Central East	7.0	514
Vic: Kangaroo Flat - Golden Square	8.7	340
Vic: Maiden Gully	6.5	97
Vic: Strathfieldsaye	6.2	151
Vic: Bendigo Region - South	7.1	126
Vic: Castlemaine/ Castlemaine Region	7.2	240
Vic: Heathcote	9.2	67
Vic: Kyneton	7.0	136
Vic: Woodend	6.4	111
Vic: Bendigo Region - North	7.3	67
Vic: Loddon	7.7	99
Vic: Bannockburn/ Golden Plains - South/ Winchelsea	7.2	346
Vic: Belmont/ Geelong	7.0	324
Vic: Corio - Norlane	12.5	738
Vic: Geelong West - Hamlyn Heights	7.2	282
Vic: Grovedale	7.9	380
Vic: Highton/ Newtown (Vic.)	5.6	391
Vic: Lara	8.2	335
Vic: Leopold	7.1	192

Note: The extent to which these estimates may vary from the reported value is shown by the upper and lower confidence intervals; these can be seen by hovering the mouse pointer over the bar chart.

^ modelled estimates not produced for Very Remote areas, Aboriginal communities or where population is less than 1,000: for other data, replaces numbers where total population is less than 100

Risk factors (estimates) - Age standardised rate per 100 >> Obese persons, aged 2 to 17 years, 2014-15 >> Victoria

http://www.theage.com.au/victoria/health-tracker-do-you-live-in-victorias-fittest-postcode-20170429-gvvd5v.html

Services and skills required

- Need for high reliability data
- Need for high reliability data services
- Need for high reliability data computation
- Need partnerships between researchers and skilled data technologists

Links on Privacy and Ethics

- ANDS guide on sensitive information and deidentification ands

AUSTRALIAN NATIONAL DATA SERVICE

ands.org.au

NCRIS

National Research Infrastructure for Australia

An Australian Government Initiative

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program. Monash University leads the partnership with the Australian National University and CSIRO.

Keith Russell

keith.russell@ands.org.au