Australian Academy of Science

Advancing data-intensive research in Australia

Australian Academy of Science

October 2021

Australian Academy of Science

Advancing data-intensive research in Australia

Australian Academy of Science

October 2021

Acknowledgements

The Australian Academy of Science gratefully acknowledges funding provided by the Australian Research Council under the Linkage Learned Academies Special Projects (LASP) scheme to support the conduct of this project. The views expressed in this publication are those of the authors and are not necessarily those of the Australian Government or the Australian Research Council.

Production of the report was supported by the Australian Academy of Science's Ian Ross Bequest. Professor Ian Ross AO FAA (1926-2006) was responsible, as Deputy Vice-Chancellor, for the ANU acquiring in 1987 the first research supercomputer in Australia thereby stimulating modern data-intensive research.

The report was authored by:

- Emeritus Professor Michael Barber AO FAA FTSE
- Professor Jane Elith FAA
- Dr Danny Kingsley
- Dr Ayesha Tulloch

Brief backgrounds on the authors are given in the Appendix.

The project was overseen by an Expert Working Group who liaised closely with the Academy's National Committee for Data in Science. The members of these committees are thanked for their commitment to the project. Project management, research and editorial services were provided by Dr Al Usher, Mr Brent Davies, Dr Jana Phan, Dr Daniel Bouzo, Mr Chris Anderson, Dr Chris Hatherley, Dr Hayley Teasdale and Ms Robyn Diamond from the Secretariat of the Australian Academy of Science and are gratefully acknowledged.

The report was reviewed by:

- Professor Philip E. Bourne, Stephenson Dean of the School of Data Science, University of Virginia;
- Professor Doctor Deiter Kranzlmüller, Chairman of the Board of Directors, Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities;
- Professor Robyn Owens FAA, Emeritus Professor, University of Western Australia;
- Dr John Simpson, Senior Director Transition and Strategic Initiatives, Compute Canada

While the reviewers provided constructive comments and suggestions, they were not asked to explicitly endorse the content, nor did they see the final draft before release. Consistent with the Academy's policy, the final report was endorsed by the Executive Committee of the Council of the Academy.

ISBN 978-0-85847-845-9 (digital PDF version)
ISBN 978-0-85847-846-6 (print version)

How to cite this report:
Australian Academy of Science (2021). *Advancing data-intensive research in Australia*.

Australian Academy of Science
GPO Box 783
Canberra ACT 2601

Tel +61 (0)2 6201 9400
Email aas@science.org.au
www.science.org.au

Process and consultation

The Australian Academy of Science convened an Expert Working Group (EWG) comprising Fellows of Australia's learned academies and other leading Australian scientists. The EWG was co-chaired by Emeritus Professor Michael Barber AO FAA FTSE and Professor Jane Elith FAA, and had three face-to-face meetings. The EWG was supported by policy analysts from the Academy and liaised closely with the Academy's National Committee for Data in Science and its Chair, Dr Lesley Wyborn.

Expert Working Group

Professor Michael Barber AO FAA FTSE (Co-chair)	Professor John Mattick AO FAA FTSE FAHMS
Professor Jane Elith FAA (Co-chair)	Professor Kerrie Mengersen FAA
Dr Sue Barrell FTSE	Professor Toby Walsh FAA
Professor Jane Hunter	Professor Bob Williamson FAA

National Committee for Data in Science

Dr Lesley Wyborn (Chair)	Dr Danny Kingsley
Professor Ginny Barbour	Dr Steven McEachern
Dr Adrian Burton	Professor Andy Pitman AO FAA
Dr Simon Cox	Dr Francois Petitjean
Professor Darren Croton	Professor Shazia Shadiq FTSE (ex-officio, Chair National Committee for Information and Communication Sciences)
Professor Louisa Jorm	

Brief backgrounds on the members of these committees are given in the Appendix.

Over 2018, 2019 and early 2020, the project undertook desk-based research, attended a number of data science conferences (e.g. the Data61's 'D61+ LIVE' 2018 conference) and consulted widely. Over sixty formal and informal interviews were held with Australian government and non-government agencies and experts from academia and business. One co-chair (MNB) used several overseas trips to engage with leading data-intensive research organisations in the UK, USA and Singapore. More formal presentations on the project were given at various fora including the 2020 C3DIS Conference. Details of these consultations are on the project's legacy website.

Between September 2018 and February 2019, the Australian research community was invited to respond to '10 questions for big data in Australian research'. This survey elicited eighty-two responses across twenty-seven disciplines and sub-disciplines and gives a useful perspective on the state of data-intensive research and associated issues circa late 2018. The responses are archived on the project's legacy website.

A number of more detailed submissions or background papers were commissioned from a number of Australian experts. These papers are also available on the project's legacy website.

In March 2019, the Australian Academy of Science convened a Data in Health and Medical Research roundtable in collaboration with the Australian Academy of Health and Medical Sciences to inform the report. The roundtable was attended by leaders in Australian health and medical research representing more than forty organisations and resulted in a formal communique by the two Academies.[1]

In April 2019 a videoconference was held with early- and mid-career researchers who had attended the Academy of Science's 2018 Research Data Science Winter School to gauge whether responses from the consultation were representative of all levels of the academic research community.

As required by the ARC funding agreement, the project liaised closely with the related ARC LASSP project, The Use of Big Data for Public Policy, being concurrently conducted by the Academy for Social Sciences in Australia (ASSA) including presenting formally on the project at ASSA's Workshop on Big Data and Public Policy held at UNSW in March 2020.[2]

Acknowledgement of Country

The Australian Academy of Science acknowledges and pays respects to the Ngunnawal people, the Traditional Owners of the lands on which the Academy office is located. The Academy also acknowledges and pays respects to the Traditional Owners and the Elders past, present and emerging of all the lands on which the Academy operates and its Fellows live and work. They hold the memories, traditions, cultures and hopes of Aboriginal and Torres Strait Islander peoples of Australia.

Contents

Acronyms and abbreviations

A	**AARNet**	Australia's Academic and Research Network
	ABS	Australian Bureau of Statistics
	ACCESS	Australian Community Climate and Earth System Simulator
	ACM	Association of Computing Machinery
	AHMAC	Australian Health Ministers Advisory Council
	AI	Artificial intelligence
	AIHW	Australian Institute of Health and Welfare
	ALA	Atlas of Living Australia
	ANDS	Australian National Data Service
	ARC	Australian Research Council
	ARC LIEF	ARC Linkage Infrastructure, Equipment and Facilities
	ARDC	Australian Research Data Commons
	AURIN	Australian Urban Research Infrastructure Network
B	**BRAEMBL**	Bioinformatics Resource Australia-EMBL
C	**CARE**	Collective benefit, Authority to control, Responsibility and Ethics
	CODATA	Committee on Data for Science and Technology
	COPDESS	Coalition for Publishing Data in the Earth and Space Sciences
	CTS	CoreTrustSeal
D	**D2DCRC**	Australia's Data to Decisions CRC
	DCC	Digital Curation Centre, UK
	DDERP	Digital Data and eResearch Platform
	DMP	Data Management Plan
	DOI	Digital Object Identifier
	DORA	Declaration of Research Assessment
	DPT	Data Science Development Planning Tool
	DSA	Data Seal of Approval
	DSCF	Data science competency framework
E	**EFT**	equivalent full time
	EI	Engagement and Impact
	EOSC	European Open Science Cloud
	ERA	Excellence in Research for Australia
	ESRC	Economic and Social Research Council
F	**FAIR**	Findable, Accessible, Interoperable and Reusable
	FoR	Fields of Research, ANZSRC
G	**GA4GH**	Global Alliance for Genomics and Health
	Gbps	gigabit per second
	GDPR	General Data Protection Regulation
	GEO	Group of Earth Observations

H	**HASS**	humanities, arts and social sciences
	HPC	high-performance computing
I	**ICAC**	Independent Commission Against Corruption (NSW)
	ICT	information and communication technology
	IMOS	Integrated Marine Observing System
	IP	intellectual property
	ISC	International Science Council
M	**MADIP**	Multi-Agency Data Integration Project
	MIT	Massachusetts Institute of Technology
	ML	machine learning
N	**NASA**	National Aeronautics and Space Administration, USA
	NCDiS	National Committee for Data in Science, Australian Academy of Science
	NCI	National Computational Infrastructure
	NCRIS	National Collaborative Research Infrastructure Strategy
	NDRIO	New Digital Research Infrastructure Organization, Canada
	NEPS	National Environmental Prediction System
	NeSI	New Zealand eScience Infrastructure
	NeurIPS	Neural Information Processing Systems
	NHIS	National Health Information Strategy
	NHMRC	National Health and Medical Research Council
	NSF	National Science Foundation, USA
O	**OECD**	Organisation for Economic Co-operation and Development
P	**PHRN**	Population Health Research Network
	PI	Principal Investigator
R	**RDA**	Research Data Australia (data discovery service of ARDC)
	RDM	research data management
	RDS	Research Data Services
S	**SEO**	Socio-Economic Objectives, ANZSRC
	STM	International Association of Scientific, Technical and Medical publishers
T	**TEQSA**	Tertiary Education Quality and Standards Agency
	TERN	Terrestrial Ecosystem Research Network
	the Code	the Australian Code for the Responsible Conduct of Research
	ToA	Type of Activity, ANZSRC
	TRUST	Transparency, Responsibility, User focus, Sustainability and Technology
U	**UKRI**	UK Research & Innovation
	UKRN	UK Reproducibility Network
	UQ	University of Queensland
W	**WDS**	World Data System
	WHO	World Health Organization
	WiDS	Women in Data Science

1. Introduction: context, objectives, scope and recommendations

Data has always been the bedrock of scientific research.[3] Traditionally seen as an input to a research project, data is now recognised as a significant output of a research project increasingly on par with publication.[4] Technology-driven advances in data collection, together with advances in computation, communications and storage, are dramatically increasing the volume and the nature of data available for research. Combined with significant advances in data analytics, including the emergence of the new discipline of data science, these developments present new opportunities and some fundamental challenges.[5] It is plausible that the ensuing 'data revolution' will disrupt research and the research enterprise as much as any other sector of the economy and society.[6]

Australia and Australian research have not escaped these profound changes. They have spurred significant investment in enabling eResearch infrastructure described by the 2016 National Research Infrastructure Roadmap as 'critical infrastructure for modern research'.[7] Access to this infrastructure together with new data sources are opening up new opportunities for research that ask questions not answerable before. These research opportunities are extending from the science, technology, engineering and maths (STEM) disciplines through to the humanities, arts and social sciences (HASS) disciplines. Indeed, some of the most exciting research with greatest impact is arising from significant collaborations *across* fields. These new developments are demanding more sophisticated data skills in researchers in many disciplines. While the research opportunities afforded are profound, their realisation is accompanied by a range of challenges. Some are technical, others are cultural, and some go to the fundamental tenets of scientific research which, unless addressed, could significantly impact on trust in science.

What is eResearch infrastructure?

The 2016 National Research Infrastructure Roadmap describes eResearch infrastructure as a "cross-cutting capability that serves research collaboration, modelling, data, and data analysis needs. It comprises advanced networks, identity, access and authentication services, high performance and cloud computing resources, management of and access to research data; ... **and the integration of all those elements** to create digital environments researchers use every day".

1.1 Objectives of the study

In response to these significant developments, opportunities and challenges, the Australian Research Council, through the Linkage Learned Academies Special Project (LASP) scheme, funded the Australian Academy of Science for a project titled 'Big data in Australian research: Issues, challenges and opportunities'. The project's objectives were to:

- explore how data and associated data technologies are changing research
- investigate the state of data science in Australia and consider how data science can most effectively collaborate with other disciplines that are increasingly using data and data analytics
- document the current state of Australia's research data ecosystem to ensure Australian eResearch infrastructure is globally competitive and able to support high-quality research

- determine a strategic framework and roadmap for data-intensive research in Australia, including recommendations for immediate steps that can improve the research data ecosystem.

Early in the project it became apparent that it was advantageous to consider data beyond the strict definitions of big data. Researchers are already exploring the potential of integrated and diverse data that may comprise small or big data but remain challenging to process or analyse. One of the early workshops on big data in research convened by the US National Academy of Sciences in 2012 noted: "Big data has created a greater awareness of all aspects of data, including the life cycle of data and the importance of metadata. It has also generated new research energy around data."[8] That remains true today. Realising the full potential of data-intensive research requires a research system with a solid foundation of best practice concerning all aspects of research data. This study took a broad definition of data but focused on research that invariably involves significant computation to generate, manage or analyse data. We term such research 'data-intensive', with the title of the project and this report reflecting that focus.

The report should be of interest to researchers, their employing institutions particularly universities, the funding councils and agencies, the publicly funded research agencies (PFRAs), professional societies and the learned Academies, policy makers and all organisations charged with maintaining and building Australia's eResearch infrastructure. Wherever possible we have tried to suggest who we think should take carriage of our more specific recommendations. Nevertheless, advancing data-intensive research in Australia requires a coordinated response by the entire Australian research system. Our hope is that this report facilitates that response.

What is big data?

Conventionally, big data refers to data characterised by the 'three Vs'—volume, velocity and variety. Such data is so large or so complex that it is difficult to analyse using traditional statistical or data analysis methods. Scientifically, this includes new data being generated by advances in ecological monitoring and earth observational technologies as well as the traditional areas of genomics, astronomy, climate science, and health data.

1.2 Structure of the report

The report is structured as follows.

Section 2: Data and data analytics are changing research describes the key drivers of the data revolution in research and illustrates how these are changing research to create both exciting opportunities and significant challenges.

Section 3: Access to data is the foundation: FAIR the governing principle explores the current status of Australia's research data policies and practices and argues that significant reform is necessary to match global best practice.

Section 4: Data science: A new enabling discipline explores the emergence of data science as a new discipline and assesses the current state of Australian data science and how it should be developed.

Section 5: An integrated eResearch infrastructure is critical reviews the current state of Australia's eResearch infrastructure, describes global trends impacting the development of eResearch infrastructure and makes recommendations that should be considered in the next phase of investment in Australia's eResearch infrastructure.

Section 6: Skills and skilled people are vital but culture matters draws on global best practice to make recommendations on how to lift the data skills of researchers and describes the impact of culture.

Section 7: Data-intensive research spawns significant challenges describes how advances in data science and data-intensive research are increasing the importance of issues such as replicability and transparency in research. These advances are challenging some of the fundamental tenets of scientific research and potentially affecting public trust in science.

1.3 List of recommendations

Recommendation	Responsible entities
RESEARCH DATA	
Recommendation 3.1 Adopt and implement FAIR Universal adoption and implementation of the Findable, Accessible, Interoperable and Reusable (FAIR) principles and the Collective Benefit, Authority to control, Responsibility and Ethics (CARE) principles for Indigenous data governance and integrate them into a national research data strategy.	**Australian Research Sector** ARC, NHMRC, DESE, PFRAs Office of National Data Commissioner Universities (DVC-Rs); Learned Academies; professional societies; researchers, publishers
Recommendation 3.2 Reform research data policies The Chief Scientist be mandated to lead a national reform of research data policies, involving the Australian Research Data Commons (ARDC) and Australia's learned academies, and working with the National Committee for Data in Science as an 'expert advisory committee' and conduit to CODATA and the International Science Council.	**Chief Scientist** ARDC Learned Academies National Committee for Data in Science Office of National Data Commissioner
Recommendation 3.3 Data and grants FAIR and CARE be mandated for all research data resulting from government funding schemes and ensure that the costs of managing research data are recognised in funding policies for universities, agencies, grants and funding councils. This should be similarly mandated and funded by state and territory governments.	**ARC, NHMRC** Other Australian Government and State and territory governments funding schemes
Recommendation 3.4 Cultural change Drive cultural change to ensure uptake of FAIR principles and new research data policies. The learned Academies should lead discipline-specific consultation to provide advice on both the appropriate length of retention of data, curation of data and decisions on whether to discard data based on the nature of the research area. Specific consideration should be given to the costs associated with the collection, retention, and preservation of data.	**Universities/DVC(R)s, PRFAs** Researchers Professional societies National Committees
DATA SCIENCE	
Recommendation 4.1 Revise employment statistics Revises Australia's official employment statistics to recognise data scientists and other data professionals explicitly and ensure data scientists working in multidisciplinary areas are accurately represented.	**ABS**
Recommendation 4.2 Data science as a discipline Data science should be recognised as a discipline and a field of research in its own right. Australian universities should recognise data science as an academic discipline and review their organisational structure for data science to enable a cohesive data science discipline whilst encouraging collaboration within the institution and externally.	Universities Learned Academies
Recommendation 4.3 Data science centres Strengthen coordination between Australian data science centres with direct infrastructure support for the Australian Data Science Network.	**DESE (NCRIS)** ARDC
Recommendation 4.4 Data Science Society The data science research community establishes a Data Science Society with an inclusive membership.	**Data science research community** Australian Academy of Science STA
Recommendation 4.5 Data science field of research code Recognises data science at 4-level in Information and Computational Sciences. Under this code at 6-level should be aggregated all areas of research relevant to the data life cycle, including machine learning, deep learning, statistical data science, data management and data curation aimed at the development of generic knowledge and data science tools.	ABS Australian Research Council
Recommendation 4.6 Map Australia's data science capability Seek funding to map Australia's data science capability analogous to the recent report on Australia's climate science capability.	Australian Academy of Science Academy of Technology and Engineering

Recommendation	Responsible entities
Recommendation 4.7 Strengthen Australian research in foundational data science Strengthen Australian research in foundational data science through an Australian program similar to the NSF's Transdisciplinary Research In Principles Of Data Science (TRIPODS)[9] program and by considering a dedicated call for proposals to establish an ARC Centre in Data Science. At least half the research activity should be in foundational data science research and its translation to areas of research with significant data challenges beyond the capacity of existing data science.	Australian Government Australian Research Council
Recommendation 4.8 Gender imbalance in data science Implement the Women in STEM Decadal Plan to actively address the gender imbalance in data science by enhancing current initiatives and giving greater prominence to applications other than business.	Australian Government Learned Academies Research funders Universities
RESEARCH INFRASTRUCTURE	
Recommendation 5.1 Refresh national research roadmap The 2021 refresh of the National Research Infrastructure Roadmap takes a holistic view of eResearch infrastructure and prepares an integrated investment plan covering all elements of eResearch infrastructure. This plan should take account of global developments in data-intensive research but must also set realistic priorities.	**NRI Expert Working Group** (DESE)
Recommendation 5.2 Co-ordination of eResearch infrastructure The 2021 refresh of the National Research Infrastructure Roadmap considers recommending that oversight and coordination of the next phase of investment in Australia's eResearch infrastructure is vested in a National eResearch Infrastructure Council—adapted from the European EOSC and Canada's NDRIO. The Council should: a. involve all areas of the research sector and all agencies involved in the deployment of significant eResearch infrastructure as well as representatives from commercial players b. have as its primary purpose to monitor and update where necessary the comprehensive strategy for investment in the next phase of Australia's eResearch infrastructure including setting key priorities c. undertake the detailed planning needed for next generational models and applications, explicitly considering how to integrate data, compute software and skills d. take leadership of the discussion of how Australia can gain the benefits that will emerge with next-generation exascale computing e. potentially, have a more direct role in allocating funding for eResearch infrastructure.	**NRI Expert Working Group** (DESE) Chief Scientist
Recommendation 5.3 Discipline decadal planning National Committees of Science should: a. have strong engagement with relevant platforms or commons established within eResearch infrastructure b. when developing decadal plans, explicitly address data issues, explore new opportunities afforded by data-intensive research and consider the need for retention of data c. conduct more frequent reviews of the state of data and associated infrastructure in their discipline areas.	**Australian Academy of Science**

Recommendation	Responsible entities
DATA AND DIGITAL SKILLS	
Recommendation 6.1 Staff research data skills and employment Research organisations (particularly universities): a. review staff development programs to ensure all researchers have access to courses aimed at enhancing basic data skills as well as discipline-focused courses to enhance more advanced skills b. resolve the current ambiguity in the employment of data scientists noting recommendation 4.2 regarding the recognition of data science as a discipline c. ensure all PhD students have access to courses to develop generic data skills before they commence their research projects, with these courses including advice on how students should collaborate or consult with a data scientist d. champion a culture that supports data science skills and creates positive feedback for data-intensive approaches to research and collaboration.	**Universities** PRFAs
Recommendation 6.2 Employ data professionals The 2021 refresh of the National Research Infrastructure Roadmap recognises that all research infrastructure facilities now produce digital data and need resources to employ data professionals to manage and maintain this data and, where appropriate, to pre-process it to facilitate research access.	**NRI Expert Working Group** (DESE) NCRIS facilities
Recommendation 6.3 Data maturity of scientific disciplines National Committees for Science be given a specific mandate to: a. assess the data maturity of their relevant disciplines and work closely with the ARDC to develop general and discipline-specific approaches to improve data maturity b. challenge disciplinary cultures that inhibit the development of data literacy and data-intensive research c. promote exchange of good practices across the research community through discussions on digital skills, provision of training and reward structures.	**Australian Academy of Science** ARDC
Recommendation 6.4 Recognition in research outputs When disseminating research outputs, researchers acknowledge and give credit where it is due: to data generators, curators and stewards.	**Researchers** Universities (DVC-Rs) Scholarly publishers Research Funders
CHALLENGES AND OPPORTUNITIES	
Recommendation 7.1 Research integrity Strengthens the governance of research integrity and develops a national policy statement on ensuring research integrity for Australia. Such a statement should specifically address the issues raised by data-intensive research	**Australian Government** Academy of Science ARC, NHMRC

2. Data and data analytics are changing research: exciting opportunities but with significant challenges

The ever-increasing volume of data and the sophisticated techniques being developed to draw information and value from it, are changing and challenging how research is conducted. This change is evident across all research disciplines and is enabling researchers to answer questions not otherwise answerable.

This chapter introduces the critical drivers in data-intensive research and provides six case studies that illustrate the opportunities this change presents. It sets the scene for the succeeding chapters.

2.1 A time of massive change, even disruption—we are at a tipping point

In 2009, Microsoft published *The Fourth Paradigm: Data-Intensive Scientific Discovery*, a collection of essays that explored the likely impact on science of information technology and particularly the increasing flows of data.[10] The title referred to the potential emergence of a fourth paradigm—data-intensive scientific discovery—to stand alongside the more conventional scientific paradigms of experimental, theoretical and computational investigation.

The Fourth Paradigm essays reproduced a talk given two years earlier by Jim Gray (then a Technical Fellow at Microsoft Research and a Turing Award winner), in which he summarised the current state of data tools and presciently concluded:

"Almost everything about science is changing because of the impact of information technology. Experimental, theoretical and computational science are all being affected by the data deluge, and a fourth, 'data intensive' science, paradigm is emerging. The goal is to have a world in which all of the scientific literature is online, all of the science data is online, and they interoperate with each other."[10]

A decade or so on, this vision is becoming a reality and perhaps in ways that might have even surprised Jim Gray (who disappeared sailing off San Francisco in 2007). Certainly, 'data-led discovery' by which data are analysed—by computer-based algorithms—to yield relationships, models, and insight has emerged as a distinctly new approach to conducting research.

The traditional paradigms are also being transformed as more and more disciplines—both scientific and non-scientific—become, in various ways, 'digitised' and thus produce

or use increasingly large or complex streams of data that can only be explored and analysed with modern data analytic methods. This is allowing research to ask new and deeper questions, obtain better answers and create greater impact. At the same time, these developments are posing significant challenges that need to be addressed to fully realise the undoubted potential of data-intensive research.

2.2 Key drivers for change

The data revolution in research is being driven by the same technological trends that are driving the data revolution in the broader economy and society. Scientific instrumentation—large and small—now automatically produces output that is digital. Consumer and industrial electronic devices from smart phones and cameras to sensors and the wider 'internet of things' are being used as scientific instruments or at least to yield data for research. Conversely, developments in society that are built on technology, such as social media, are opening up new data sources for research. In addition to technical developments, reductions in cost of data generation, such as DNA sequencing, is an important driver. For example, the cost of sequencing the human genome has decreased from US$100 million in 2001 to less than US$1,000 today.[11]

New generations of scientific instruments are producing extremely large data streams. This is particularly true in astronomy. In 2019, the Event Horizon Telescope produced the first image of a black hole by combining observations involving 5 petabytes (PB) of data (see Section 2.3.1).[12]

Even greater data streams are envisaged with new instruments in planning, such as the Giant Magellan Telescope and particularly the Square Kilometre Array (SKA).[13] Data streams of these magnitudes will not only stress computational and storage capacities, they will also require new algorithms and a deep collaboration that brings together astronomy, supercomputing and data science.[14]

Awaiting call up. These Argo ocean observing robots or floats ready for deployment around Australia. PHOTO: BRUCE MILLER, CSIRO

At the other end of the scale, ubiquitous (and often cheap) ICT-enabled sensors are opening up new areas of research or significantly transforming that data available for research. The ARGO float program, consisting of over 4000 underwater robotic floats, has allowed direct measurement of ocean data in areas of ocean (such as the Southern Ocean) that are not frequented by shipping which has been the traditional source of such data.[15] Again, even greater volumes are on the horizon. The Defence Advanced Research Projects Agency (DARPA) in the US is beginning to deploy an 'Ocean of Things' involving thousands of small, low-cost floats distributed across the ocean.[16] Significantly, the development includes new computational and analytical tools to effectively handle and analyse the resulting data streams.

Remote cameras, smart sensors and increasingly drones are enabling wildlife to be observed remotely, and difficult or hostile environments to be explored. Such developments have important implications for environmental research.[17, 18]

"Remote cameras, smart sensors and drones are enabling wildlife to be observed remotely, and difficult or hostile environments to be explored."

The reduction in the cost of DNA sequencing is a major factor in the explosion of available genomic data. Combined with increases in compute power and sophisticated algorithms, this is transforming biology and medicine. The response to the COVID pandemic, from underpinning epidemiological modelling to guide public heath responses to the development of a vaccine in record-breaking time, would not have been possible without these developments in data-intensive science and the associated data technologies.[19]

Data streams for research are not only getting larger, but they are also becoming increasingly diverse. Smart and wearable devices are now commonplace. Such devices can yield data that potentially allows new ways to track clinical outcomes in clinical research and to enable new research in behavioural science and fields like nutrition.[20] However, data from wearables is relatively unstructured compared to that from more conventional instruments.

Even more diverse data arises from social media or techniques such as 'web scraping' that extracts data from websites. The analysis of such data sets often requires complex and sophisticated techniques from machine learning (ML) and artificial intelligence (AI). In addition to analytical challenges, the collection of such data can pose legal and ethical issues.[21] It can also be combined with other data sources, leading to new insights—such as leveraging social media posts in combination with news alerts and airline travel data to predict and map the locations of infectious disease outbreaks (see case study in Section 2.3.2).

The emergence of large digital archives, such as photographs, videos and MRIs, are opening up new areas of research although efficient and effective tools to search and extract pertinent information are still in their infancy.[22] See the Australian Acoustic Observatory for an interesting collection of sounds from zoology.[23]

From a theoretical perspective, these new data sources are encouraging the development of increasingly complex and sophisticated models that generate large data streams themselves or increase the demand on data sources. Weather and climate models are perhaps the best-known examples. More detailed weather forecasts (at a finer spatial scale) require significant increases in data and in computational power. Importantly, such models demand interoperability between data sets (see Section 2.3.4). However, the ability to build more complex data-intensive models is occurring across science and in many non-scientific fields. An example is the sophisticated models and

forecasts of elections that the US website FiveThirtyEight builds, powered by sophisticated data and data analytics.[24]

Finally, artificial intelligence and machine learning are both major consumers and sources of large-scale data. One recent estimate suggests that computing power needed to train AI applications is rising seven times faster than in earlier years.[25] This reflects the size of the data sets needed to effectively train a sophisticated AI system. Such developments have significant implications for the infrastructure required to support data-intensive research (see Section 5).

2.3 Great opportunities: six case studies

2.3.1 THE FIRST IMAGE OF A BLACK HOLE

In April 2019, the Event Horizon Telescope (EHT) released the first image of a black hole, a feat of massive data collection and analysis.[26] The black hole is in the centre of the elliptical galaxy Messier 87, which is the biggest galaxy in the centre of the Virgo cluster and 53 million light years away.

The EHT consists of eight telescopes at six locations across the world, from the South Pole to Spain. This created an aperture the size of Earth, required because the distance to M87 means the black hole is very small in the sky.[27] The image is the result of data collected using a technique called Very Long Baseline Interferometry. This works by synchronising the array of telescopes to focus on the same object.

The data was gathered over four days, creating 5 petabytes which was greater than the level the internet is able to process, so the data was transported physically to a central processing location. The image was supplemented by data collected from multiple other sources including observations from NASA space telescope missions. This created models of the jet and disk around the black hole that could be compared with the EHT observations.[28]

2.3.2 PREDICTING THE PANDEMIC

A Canadian health monitoring platform called BlueDot correctly predicted the coronavirus outbreak first detected in Wuhan, sending out warnings before the World Health Organisation notified the general public in January 2020.[29, 30] The platform uses natural-language processing and machine learning techniques to analyse information coming from multiple sources in multiple languages. By analysing global airline ticketing data, BlueDot can predict where and when infected people are potentially going, with the associated spread of the disease. A similar approach was used in 2015 to identify high-risk international pathways for the dispersion of Zika virus.[31]

This program was developed by a team with a wide skill set including, in addition to data scientists, epidemiologists, engineers, ecologists, geographers and veterinarians. They spent a year teaching the computer to detect 150 deadly pathogens. Picking up a report in a Chinese business paper of the occurrence in Wuhan of some cases of a disease not in this data base, the algorithm issued an alert. Within a matter of seconds of scientists realising this could be an outbreak, the algorithm was able to analyse the flights of the more than 800,000 travellers leaving Wuhan on 31 December, identify the destination cities and the number of people who travelled there.

2.3.3 IDENTIFYING ENDANGERED ANIMALS AFFECTED BY BUSHFIRES

In late 2019 and early 2020, Australia was engulfed by the worst bushfires on record. In addition to the loss of human life, homes and crops, more than a billion birds, mammals and reptiles were estimated to have been killed. But even as the fires were still burning there was a need to quickly understand where the greatest losses of habitat were for Australian threatened species.

New generations of scientific instruments, such as the Square Kilometre Array, are expected to produce extremely large data streams. Image shows antennas of CSIRO's ASKAP telescope at the Murchison Radio-astronomy observatory in Western Australia.
CREDIT: CSIRO SCIENCE IMAGE

Burned or fire-affected areas are characterised by deposits of charcoal and ash, removal of vegetation and the alteration of vegetation structure. NASA's Terra satellite, using its Moderate Resolution Imaging Spectroradiometer (MODIS) instrument, takes images of vegetation that has been affected by bushfire. MODIS can overlay reflective bands of colour to highlight areas burned and distinguish them from areas of regular vegetation. These create images that clearly demonstrate bushfire-affected areas.[32]

The Centre for Biodiversity and Conservation Science at the University of Queensland used this publicly available NASA satellite imagery of the burnt areas and intersected it against the approximate distributions of all the threatened animals and plants listed under the *Environment Protection and Biodiversity Conservation Act*. They found that more than 6 million hectares of habitat had been lost, endangering 250 already threatened species. Birds, mammals, reptiles, and fish were all affected but the largest group of threatened species were plants. This finding contributed to the government's bushfire recovery planning.[33]

2.3.4 INTEROPERABILITY IN SUPPORT OF EARTH SYSTEM MODELLING

'Data' in the meteorological context refers mainly to that gathered from systematic observational networks, spanning in-situ and remote sensing instruments on surface-based, airborne and satellite-based platforms. Measurements of diverse parameters are made with varying spatial and temporal frequency and sensitivity. Variability is also introduced depending on the instrument and environment, intended application, and reporting standards, according to designated data formats and available communications methodologies and protocols. The data are often then shared globally according to varying intergovernmental agreements. By incorporating data into numerical modelling of the Earth system, researchers address global, national and local questions from climate predictions to aviation route forecasting to flood or fire preparedness. Data interoperability is essential because data, models and questions span global climate scales down to high-resolution scales.

Early numerical meteorological models were developed at individual (and competing) institutions, but improved data communications and computing power has since facilitated greater collaboration including the sharing of data. Access to the extensive suite of global meteorological data sets is essential for developing and evolving advanced climate and weather models. The utility of the exchanged data and its incorporation into models is assisted by interoperability protocols covering all aspects of model development: data and metadata formats, data communication, software environments, architectures, algorithm formulation and coding.

Importantly, interoperability does not mean that models need to be identical; there is value in diversity. However, the number of components and their complexity means that shared development is required, as each system is only as good as its weakest part. Shared development ensures interoperability by enabling models at different scales or based on other data to 'talk' to one another. Shared development places extra demands on interoperability, since comparing results between partnerships to advance overall development requires the exchange of input data and package code interfaces, outputs, diagnostics and associated software. The international numerical meteorological modelling community has developed strategies to ensure the necessary protocols are established. All 192 member countries of the World Meteorological Organization (WMO) work with and are supported by the Global Data Processing and Forecasting System, which will soon offer cloud-based data and computational servers to every nation for meteorological modelling support and sharing.

2.3.5 TRANSCRIPTION OF ENDANGERED LANGUAGES

At the ARC Centre of Excellence for Dynamics of Language (CoEDL), researchers are employing AI to assist with automated transcription of endangered language recordings held in archives that contain 50,000 hours of recordings in over 130 different languages. Manual transcription is extremely laborious for linguists trying to document archives of recordings. To fast-track this process, CoEDL researchers collaborate with Google and the open-sourced AI platform, TensorFlow, to develop ML speech-recognition algorithms that process audio recordings automatically. Algorithms are improved by linguists and community elders to build manually annotated training models that accelerate word detection in untranscribed recordings. Typically, nine hours of transcribed recordings are required for model training, and it takes two days to build a model, but the outcome is rapid keyword detection of over hundreds of hours of untranscribed recordings.

CoEDL researchers aim to build AI language recognition software into their robot, Opie, designed to teach children to understand Australian Indigenous languages. CREDIT: ARC CENTRE OF EXCELLENCE FOR DYNAMICS OF LANGUAGE (COEDL)

The greatest challenges are poor data quality and multi-lingual recordings, dialects and accents, small training corpuses and privacy or access restrictions. Indigenous communities can be reluctant for their recordings to be stored or processed on the cloud due to security, privacy and confidentiality concerns. Despite these challenges, to date, these advances have enabled linguists to train models and obtain usable first-pass transcriptions of data for 16 endangered languages across the Asia-Pacific region.[34] The long-term ambition of the project is to build the AI language recognition software into CoEDL's robot, Opie, designed to teach children to understand Australian Indigenous languages.

2.3.6 BRINGING TOGETHER DISPARATE DATA ON THREATENED SPECIES

Through the Australian Government's National Environmental Science Program, a Threatened Species Index (TSX) for Australian birds was produced in 2018. For the first time in Australia, this index can provide reliable and rigorous measures of trends across

Australia's threatened species. To build the TSX, years of monitoring data on Australia's threatened birds from researchers in academia and state government environment agencies to private ornithologists had to be collated. More than 110 individual data sharing agreements were developed with data custodians. Signing of data agreements involved legal teams from multiple institutions and took two years from the project's inception, due to the number of custodians that needed to be approached, and to perceived or real data sensitivities.[35]

The TSX project was not only constrained by access, but also by data availability and standardisation. Many datasets were simply not up to date. A lag of up to 5 years in digitisation of surveys by custodians meant that trends were possible for only 66 species (24% of all of Australia's threatened bird species) despite monitoring data being available for 76% of Australia's birds.[36] The project was further delayed by the need to aggregate disparate data supplied in many formats (and often with limited or no metadata) into a single data base, as no standards currently exist for archiving population trend data.[37]

Despite the challenges, the collated dataset contains hundreds of thousands of surveys dating back to 1985 and is Australia's first national aggregated database of trend monitoring for threatened species. Collaboration with data scientists from the Terrestrial Ecosystem Research Network (TERN) enabled the team to develop a transparent, repeatable data pipeline that pulls disparate data for individual sites and individual species into a scientific workflow for analysis. The TSX team developed guidance for data collectors and custodians on how to archive trend data to maximise its re-use in national and global efforts to track biodiversity change over time.

2.4 Challenges

The six case studies in the previous section illustrate not only the breadth of opportunities presented by data intensive research but also some of the challenges faced by researchers conducting such research. These challenges include:

- access to data, particularly government and public data, with valuable data locked in institutional silos
- a lack of uniform policies in critical areas such as privacy and data ownership
- integration of large and diverse data sets across domains
- storage and curation of data—the 'unfunded' part of the eResearch infrastructure
- the need to optimise the interactions between data, algorithms/software and compute
- skills and skill development with implications for research training and development
- impediments arising from culture and practice—individual, disciplinary and institutional
- issues concerning research integrity, particularly reproducibility and replicability.

How Australia addresses these challenges will significantly influence Australia's ability to extract full benefit from data-intensive research. The remainder of this report analyses these challenges in greater detail and advances several recommendations to overcome them.

3. Access to data is the foundation, FAIR the governing principle

Data-intensive research generates and requires access to data—both data **from** research and data **for** research. Some of the most valuable insights generated from data-intensive research come from linking and analysing complex and heterogeneous data sets—large and small.

This section explores the principles and policies that facilitate the management and sharing of research data and outlines international efforts to encourage open science and the sharing of research data. Compared to other jurisdictions Australia currently has multiple and inconsistent policies referring to the management and sharing of research data, with no coherent national research data policy to govern access and management of data.

Major reform of Australia's data research policies is warranted. This reform should be based upon universal acceptance of the FAIR (Findable, Accessible, Interoperable and Reusable) Principles augmented by CARE (Collective benefit, Authority to control, Responsibility and Ethics). A major priority should be to improve access to Government data for research.

3.1 What are the main issues

Some of today's most innovative and exciting data-intensive research involves linking very diverse data. Meeting the challenges associated with data-intensive research requires, as a foundation, a comprehensive research data policy and modern data management practices.

Unfortunately, despite the increasing importance of data in research and public policy, **Australia has no coherent national research data policy**. Instead, there are currently multiple and disparate policies from institutions, government funders, professional organisations and others referring to the management and sharing of research data. There is inconsistency, and in most cases, policies are not enforceable or enforced. No single body in Australia has oversight of the research data agenda, either relating to policies on managing research data or coordinating the supporting infrastructure.

Globally, how research disciplines deal with data, data management and data practices is a live conversation. As data has become more central to research, so has the debate about how to manage data. Often this conversation has been binary—should data be completely open or closed? Or is there a different way between the binary choices of open or closed?

3.1.1 'OPEN SCIENCE' AND 'OPEN ACCESS'

Internationally and domestically, the research community has shown increasing support for open science, also referred to as 'open scholarship'. Open science and open access are not the same.[38] Open science is a much broader term that captures the conduct and dissemination of research. Open science captures all aspects of the research life cycle: open access publications, open data (including code and data pipelines) and other research and training materials, and the avoidance of restrictive intellectual property.[39]

In Australia, open science has attracted some support. In the October 2018 House of Representatives Committee's Standing Committee on Employment, Education and Training Inquiry into Funding Australia's Research, recommendation (12) stated that "... the Australian Government develop a more strategic approach to Australia's open scholarship environment".[40] Similarly, the 2017 Australian Government Productivity Commission's inquiry report on Data Availability and Use make several recommendations regarding the release of publicly funded datasets.[41]

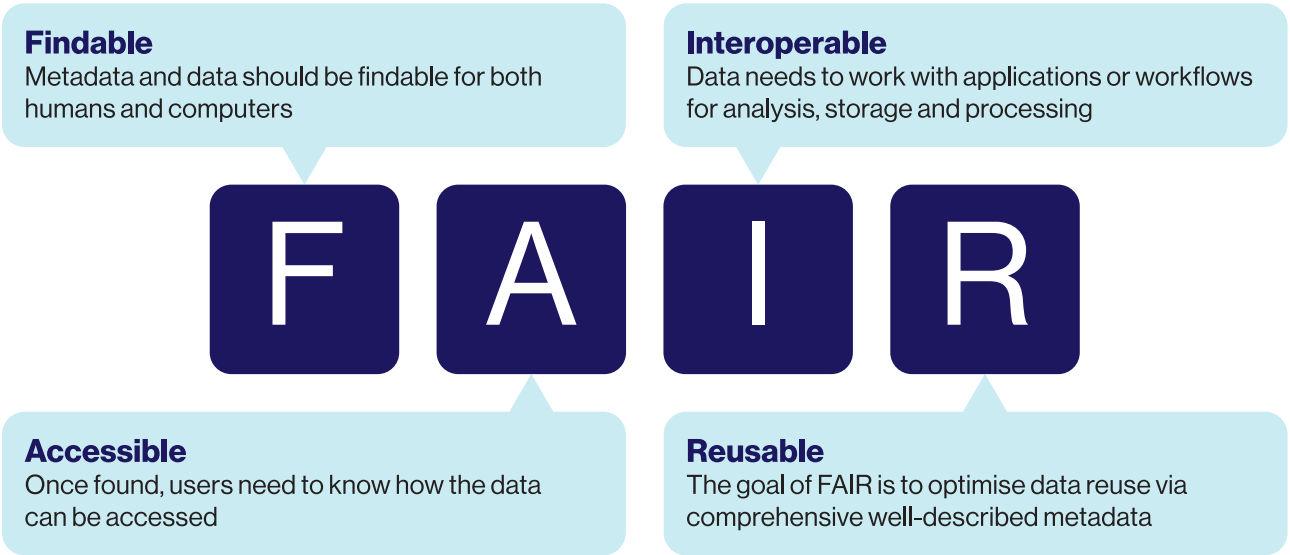
In 2015, Australia endorsed the Open Government Partnership[42], which has a goal of making public sector data open by default. The Australian Government's 2015 Public Data Policy Statement states that "Australian Government entities will make non-sensitive data open by default to contribute to greater innovation and productivity improvements across all sectors of the Australian economy".[43] Following this, the Department of the Prime Minister and Cabinet conducted consultations and developed an *Open Government National Action Plan 2018-20*,[44] which built on the first plan released in 2016. Through the *Open Government National Action Plan*, the Australian Government "has committed to implementing a simpler, more efficient data sharing and release framework...".[45]

3.1.2 DISTINGUISHING BETWEEN FAIR AND OPEN DATA

One of the pillars of open science is open data. There are many instances where open is not an appropriate mechanism for researchers or other owners of data. Areas where regulated access is more appropriate are access to Indigenous data and access to sensitive public data, in particular health data, as well as defence data.

While open data is a laudable objective, the FAIR Guiding Principles (data that is Findable, Accessible, Interoperable and Reusable) offer a more feasible option for research data.[46] In the case of access to Indigenous data, integrating FAIR with CARE Principles — Collective benefit, Authority to control, Responsibility and Ethics — would allow regulated access while respecting Indigenous people's rights and interests.[47]

The FAIR Guiding Principles for scientific data



Taken together, these principles allow for the regulation of access, the fundamental principle being that **there should be a pathway to access, and access should not be unnecessarily or unjustifiably withheld**. Each principle can be broken into measurable requirements.[48] Importantly, data provenance is key to the FAIR Guiding Principles and is necessary to ensure data quality.

Internationally, the Beijing Declaration on Research Data articulates a set of principles to achieve FAIR research data.[49] The Declaration came from a meeting of the International Science Council's Committee on Data (CODATA) held in Beijing in September 2019.[50] Australia is represented on CODATA by the Chair of the Australian Academy of Sciences National Committee for Data in Science (NCDiS).[51] By proxy, Australia is a 'signatory' to the Beijing Declaration. The OECD Council, which Australia is a member of, has also recently updated its recommendation to access research data from public funding.[52]

In the Australian context, a policy statement on FAIR access to Australia's research outputs (i.e. not just data) was released in January 2017, developed by a working group convened by the Council of Australian University Librarians and the Australasian Open Access Strategy Group (now Open Access Australasia) under the auspices of the Universities Australia Deputy Vice-Chancellors (Research).[53]

Despite several Australian organisations endorsing the policy statement and FAIR principles,[54] it is far from universally accepted as an Australian position on research data management of stewardship. The most significant adoption of the principles is within the National Health and Medical Research Council's (NHMRC) Open Access Policy,[55] which "support[s] the overall intent of the FAIR Access to Australia's Research Statement". Last updated in 2018, the policy is undergoing revision.

However, many Australian institutions, including universities, have given only tepid endorsement to the FAIR principles. Guidelines[56] and frameworks[57] to support and assess FAIR practices have been developed, however, **it is time that all players in the Australian research community and those agencies holding Government or public data commit to the FAIR principles** in a coordinated manner to ensure consistency.

Accepting FAIR as the goal allows for feasibility, purpose, cost (financial and environmental), and the data life cycle to be explicitly considered. This includes considerations from the perspective of data-intensive research and how 'accessibility' includes machine accessibility.

Recommendation 3.1 The Australian science and research system universally adopts and implements the Findable, Accessible, Interoperable and Reusable (FAIR) principles and the Collective Benefit, Authority to control, Responsibility and Ethics (CARE) principles for Indigenous data governance and integrate them into a national research data strategy.

3.1.3 ACCESS TO GOVERNMENT-HELD DATA

A key message from the consultations conducted for this report, particularly with the health and medical research community, is Australia's need to make government (public) data available for research. Making government data open access requires that it be available in non-proprietary standards and formats with a clear definition of licensing and access constraints to promote findability, accessibility, interoperability and reusability (FAIR).[58]

The data held by the Australian Government and state and territory governments forms an incredible resource that can, through research, inform and influence public policy and health outcomes.[59] There should be an accessible pathway for this data, for **researchers to request access to data in a format with minimal manipulation or conversion, as well as high-resolution data products to facilitate their broad use in scientific research**. One possible structure to categorise different levels of data processing is the

USA's NASA Earth Science Data Systems data processing levels.[60] These levels range from Level 0, which include unprocessed data at full resolution, through Level 4, which include outputs of lower data processing levels as models.

Before data can be moved into an archive, it needs to be curated in a modern, FAIR and machine-readable standard, particularly where government data is maintained for longitudinal studies such as environmental data. This is good research practice, providing equity of access to data and ensuring research quality and value for research funding, as well as of compliance with government policies and strategies such as the *Archives Act (1983)*.[61]

Within Australia, two bodies oversee the curation and archive process for public data. The National Archives of Australia (NAA) is responsible for Australian Government records and government agency information management policy.[62] For example, the NAA Digital Continuity 2020 Policy "seeks to support efficiency, innovation, interoperability, information re-use and accountability".[63] The NAA has also recently published *Building trust in the public record: managing information and data for government and community*,[64] a policy to support how Australian Government information is managed. The second body, the Office of the National Data Commissioner (ONDC), has responsibility for government agency data exchange and sharing policy.[65] The ONDC's Sharing Data Safely document describes principles and provides a framework for government agencies to share public data safely.[66]

Without access to health data, for example, it is difficult to see how initiatives like the Genomics Health Futures Mission can succeed.[67] In addition to research access to health data, there is a need to invest in managing crucial health data assets that can support and enhance the research capacity that exists in Australia. Prior to the COVID-19 pandemic the Australian Health Ministers Advisory Council (AHMAC)[68] funded the National Health Information Strategy (NHIS)[69], and the Health Studies National Data Asset program was funded by the ARDC. These were significant steps forward.[70]

There is an opportunity for Australia to harness the power of data to drive improvements in health outcomes through research by **creating an environment in which the safe and secure use of patient data is balanced with the rights and interests of individuals**.[59] Advances in health and medical technologies and research methods that enable linked data sets will allow Australia to realise innovation and improve preventative health medicine. Linked data sets for research, appropriately managed and de-identified, are essential. Such data assets need to allow for disaggregation by demographics, location and health issue. Where existing data sets do not have relevant 'flags', efforts should be made to improve them.

"There is an opportunity for Australia to harness the power of data to drive improvements in health outcomes through research by creating an environment in which the safe and secure use of patient data is balanced with the rights and interests of individuals."

However, there are significant and justifiable concerns about the management of and access to public data, particularly health data, due to its potentially sensitive nature and the lack of 'unbreakable' de-identification technologies.[71] Investment is required to increase the capability of the Australian Institute of Health and Welfare (AIHW) to manage public health data in collaboration with the national research agenda(s).[69] Agencies such as the AIHW should work with governments at all levels and other specialist agencies such as the Australian Bureau of Statistics (ABS) to coordinate and evaluate appropriate secure data-sharing technologies.

The increasing use of artificial intelligence (AI) to aggregate and process data opens specific challenges for the use of public and health data.[72] Surmounting these challenges is essential for both the credibility of AI applications to health, and minimising risks from these new technologies.

There has been some progress in Australia on access to government data, including for research. The May 2017 Productivity Commission's report on Data Availability and Use[41] and the November 2019 Data Sharing and Release Legislative Reforms Discussion Paper sought to advance access to data for research.[73] A National Data Commissioner was appointed in 2018 to "provide oversight and regulation of Australia's new national data sharing and release framework, including monitoring and reporting on the operation of the framework and enforcing the accompanying legislation".[74] The Data Commissioner is supported on "ethical data use, community expectations, technical best practice, and industry and international developments" by a National Data Advisory Council.[75] Its membership includes the Australian Chief Scientist and several health researchers, suggesting that research access to government data is also on the agenda.

As of 7 June 2021, the proposed Data Availability and Transparency Bill is currently before the Australian Parliament.[76] Various states have recently reviewed their legislation concerning data access and privacy, and several jurisdictions have established data analytics teams such as the Data Analytics Centre in New South Wales.[77]

In the absence of an overarching legislative framework to allow researchers access to Australian Government data, access is available through the AIHW, and more recently, the Multi-Agency Data Integration Project (MADIP) led by the ABS. The AIHW administers numerous significant national health data collections that include large amounts of state and territory data, releases significant data resources, operates a data linkage team and has a crucial function making data available.[78, 79] The ABS's MADIP allows researchers access to unidentified data from a range of Australian Government departments, including health, education, the Australian Taxation Office, and data from the Census.

While national leadership, through the Office of the National Data Commissioner and the National Data Council, is welcome, Australia is a federation and requires a collaborative effort to maximise the value drawn from our health data resources. This must include buy-in from the Australian Government and state and territory governments and associated stakeholders, including the research community. The challenge is to **ensure that the various initiatives about improving access to Government data for research are effectively coordinated to enable data sharing and avoid unnecessary duplication**.

Recommendation 3.2 The Chief Scientist be mandated to lead a national reform of research data policies, involving the Australian Research Data Commons (ARDC) and Australia's learned academies, and working with the National Committee for Data in Science as an 'expert advisory committee' and conduit to CODATA and the International Science Council.

3.1.4 THE IMPORTANCE OF INTEROPERABILITY AND GLOBAL STANDARDS

Interoperability is embedded in the FAIR principles and refers to the ability of disparate computational systems to connect, exchange and interpret information. Interoperability stands in contrast to data remaining in silos, whether institutional or discipline based. For data to be interoperable, researchers must have the ability to access, exchange, interpret and assimilate it from all providers within a specific modelling framework.

A lack of interoperability complicates the linkage of Australia's research data. Different systems collect data in different formats and file types and there is inconsistency between data, tools and methodological innovations. The development of agreed national

standards has lagged behind the data boom, with data management policies for archives being informed by the sustaining partners' priorities rather than the public good. Data collected in one place by one institution may not be able to be read by another system elsewhere.

Data interoperability is essential for effective research collaboration that enables the successful operation and application of many models and processes dependent on big data. Examples include numerical weather prediction systems and global climate models. The European Open Science Cloud Research Pilot Program identified data interoperability as a specific task for stakeholders to ensure scientific data's ongoing availability.[80] Lack of interoperability means wasted research time processing and standardising data from different sources.

For research data spanning different domains, institutions, platforms, space, and time scales to be aggregated and interoperable, there need to be well-defined data standards, enabling agreements and policies. Such tools must be developed collaboratively to ensure success. Without data standards, it will be impossible for multiple organisations or countries to collaborate in creating global models critical for current and future research.

"Strategies to ensure data interoperability are urgently needed to support Australia's data linkage within many disciplines, with the goal of linkage across disciplines and even with international data ecosystems."

Strategies to ensure data interoperability are urgently needed to support Australia's data linkage within many disciplines, with the goal of linkage across disciplines and even with international data ecosystems. Examples of the value of interoperable data in a global context are demonstrated in earth system modelling, weather forecasting and climate modelling (see case study in 2.3.4).

In these examples, data is sourced from multiple systematic observational networks with varying spatial and temporal frequency and sensitivity. Data interoperability is essential because data, models and questions span global climate scales down to high-resolution local scales. To meet this specific need, the Australian and New Zealand Location Information Metadata Working Group has formed to improve interoperability and data exchange, especially between research and government sector data, and bridge differences between commonly used standards.[81] The group has developed a good practice guide for the spatial community to improve understanding of locational metadata and communicate its benefits.[82]

3.1.5 FUNDER RESPONSIBILITIES

Policies for Australian funding bodies to encourage FAIR research data management and stewardship are currently weak and inconsistent. Responsibility for research data resides within the 2018 Australian Code for the Responsible Conduct of Research (the Code).[83] The Code includes the principle of transparency (Principle 3), which is to "share and communicate research methodology, data and findings openly, responsibly and accurately".

Several major science funding agencies in Australia have data management policies based on recommendations rather than monitoring and compliance. For instance, in an explicit example of weak policies for data management and stewardship, the ARC states in its data management requirement that it "does not require that full, detailed data management plans be submitted for assessment, but from 2020 will require that such plans are in place prior to the commencement of the project. Currently, the ARC does not mandate open access to data".[84] The NHMRC Open Access Policy is also weak, pushing

responsibility onto researchers, and it is primarily focused on publication and not research data.[55] This policy is currently under review.

There is a reference to FAIR in the National Collaborative Research Infrastructure Strategy (NCRIS) Guidelines,[85] which states that "data generated, created, captured or stored by NCRIS funded projects will be made available to the wider research community based on the FAIR principles, appropriately implemented for individual research communities". However, again, with no monitoring and no sanctions for non-compliance, all these policies are recommendations at best, according to well-understood principles of successful policies.[86]

There are clear solutions for this situation. Both the ARC and the NHMRC could require fully costed research data management plans (DMPs) and software management plans for all grant applications, which should be independently reviewed by data professionals, not just instrument infrastructure experts. DMPs must be machine actionable and accessible. Consideration should be given to the issue of capturing ancillary data, machine settings and compiler options. This is a significant challenge because of the need to capture more than the data and the software. In addition to this, funding organisations must recognise and support the costs of ensuring data are FAIR.[87]

Funding policies must be aligned and strengthened from recommendations to requirements for sharing research under FAIR principles, with systems in place for monitoring compliance and sanctions on future funding for non-compliance.

3.1.6 INSTITUTIONAL RESPONSIBILITIES

Under the Australian Code for the Responsible Conduct of Research (the Code), institutions are responsible for developing policies to implement the Code. According to the Management of Data and Information in Research guidelines, "research institutions have a responsibility to develop and implement policies and provide facilities and processes for the safe and secure storage and management of research data and primary materials ...".[88] The Code notes that institutions and researchers should, "where possible and appropriate, allow access and reference to [research data] by interested parties". However, the interpretation of 'access' can be broad in this context, and these responsibilities are far from a requirement to make the research data FAIR.

Across institutions, the policy response to this responsibility is highly variable. Many institutions have no policies, and those that do are often focused on published research outputs rather than the source data.[89]

There is a need for Australian universities to address their policies relating to research data, ensuring that they address the question of research data and publication, with an equal emphasis on both FAIR and CARE.

Australian universities and FAIR data

There is some commitment of intent from the universities to making Australian research data FAIR. In early 2020, the Group of Eight signed[i] the Sorbonne Declaration on research data rights.[ii] The International Alliance of Research Library Associations,[iii] which includes the Council of Australian University Librarians,[iv] has also supported the Sorbonne Declaration.[v] This declaration commits to "encouraging our universities and their researchers to share data as much as possible" and "advocating that these principles be integrated into institutional research data policies".

Of the universities that constitute the Group of Eight, only the Australian National University policy states that research data is made openly accessible by default via its open access repository. The University of Sydney states that researchers should make completed research data sets openly available for re-use by other researchers via an appropriate repository, and UNSW encourages researchers to make research data descriptions, small datasets and software and computer code openly available via the institutional repository. The University of Western Australia's Research Integrity Policy states that researchers will "make available for discussion with other Researchers all Research data related to publications in accordance with FAIR Principles" and researchers are able to publish their research data in the UWA repository. Monash University's policy and procedure states that data must be made accessible for use and reuse and should be stored in a repository. The University of Adelaide's Research Data and Primary Materials Policy states that "The University is committed to enabling open access to its scholarly output. Where the University is the owner of research data, it will make the data-sets available under the Creative Commons CC-BY licence v4.0."

The other Group of Eight universities policies on open access to research data are less specific. The University of Queensland's Research Data Management policy states that "Research Data and Primary Materials should be made available by Researchers for use by other researchers and interested persons for further research, after reasonable periods following the completion of the research." The University of Melbourne requires deposit of metadata in an accredited repository and supports sharing of research data.

Beyond policy, the Publication and Dissemination of Research: a guide supporting the Australian Code for the Responsible Conduct of Research suggests that "institutions should support researchers to ensure their research outputs, which must include data and not just publications, are openly accessible in an institutional or another online repository, or on a publisher's website".[90] Although not explicitly noted, material made openly available in an institutional repository needs to be accessible to researchers outside of the hosting institution.

In Australia, the joint NHMRC, ARC and Universities Australia guide on *Management of Data and Information in Research* recommends that "institutional policies should address the storage, retention and disposal of all research data, whether held in an institutional repository or externally".[88] However, the guide notes that the specific type and purpose of the research will affect the minimum period for the retention of research data, suggesting variations from 12 months to 15 years, to retaining data permanently.

There are some excellent examples of systematic developments for the management of research data within Australian universities. The UQ Research Manager system provides the University of Queensland (UQ) research community with a collaborative, safe and secure large-scale storage facility to practise good stewardship of research data.[91] It facilitates collaboration across the entire UQ research community, external research institutions and industry partners. The European Commission report, Turning FAIR into Reality, cites UQ's Research Manager as an example of, and approach to, good research data management practices.[92] Many other Australian institutions use ReDBox built by the Queensland Cyber Infrastructure Foundation.[93] ReDBox is an open-source research data management (RDM) platform that helps researchers and institutions plan, create and publish their research data assets.

The 2019 Review of the Higher Education Provider Category Standards discussed the requirements of an institution to be able to use the title 'university' and conditions for the research component of the university's work.[94] The review noted that the current lack of definitions for research quality and quantity, and recommended a threshold benchmark of quality and quantity of research be included in the Higher Education Provider Category Standards. Despite **the dependence of research quality and reproducibility on the good management of research data**, the review made no mention of research data. It would seem appropriate that at least minimal standards about the management of research data be included in threshold standards.

Standardising global health data

The World Health Organization (WHO) is progressing policy that standardises how it receives data from global institutions through the WHO Forum on Data Standardization and Interoperability.[vi] The Global Alliance for Genomics and Health (GA4GH) is working on how to standardise approaches to allow researchers to share data through platforms and workflows that recognise different sets of rules in various communities.[vii] GA4GH Connect is a five-year strategic plan aiming to accelerate the uptake of standards and frameworks for genomic data sharing within the research and healthcare communities to enable the responsible sharing of clinical-grade genomic data by 2022.[viii]

3.2 What is global 'best practice' telling us?

3.2.1 INTERNATIONAL APPROACHES TO RESEARCH DATA MANAGEMENT

Other countries are increasingly eclipsing Australia concerning a national approach to RDM and sharing. The Finnish 2020 Declaration for Open Science and Research,[95] Ireland's 2019 National Framework on the Transition to an Open Research Environment,[96] the UK's 2016 Concordat on Open Research Data,[97] Europe's 2013 guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020,[98] Sweden's 2015 Proposal for National Guidelines for Open Access to Scientific Information[99] and the 2013 G8 Open Data Charter serve as good examples.[100] The European Commission has focused on open science (in understanding the word 'science' to incorporate all scholarship) as part of the Horizon 2020 funding round.[101]

National support for data management is crucial to success. The UK Data Service[102] is funded by the Economic and Social Research Council (ESRC) to meet the data needs of researchers, students and teachers from all sectors. Services include guidance and training to develop skills in data use and the development of best practice data preservation and sharing standards.[103] The equivalent in Australia is the Australian Research Data Commons (ARDC).[104]

In the UK, the Digital Curation Centre (DCC)[105] was launched in 2004 arising from the JISC Continuing Access and Digital Preservation Strategy,[106] which argued for establishing a national centre for solving digital curation challenges that individual institutions or disciplines could not tackle. It is now a world-leading centre of expertise in digital information curation.

The US Data Coordination Network is a group of 12 leading US Academic institutions and non-profit data repositories aiming to make academic research data FAIR and available to the public.[107] Funded by the Sloan Foundation, the institutional data curators collaborate with researchers to make data FAIR. The network also pools its data curation experts to validate and curate datasets collectively.

3.2.2 INTERNATIONAL FUNDER APPROACHES

The UK Research and Innovation's Common Principles on Data Policy demonstrate that consistency between funder policies provides a clear message and more straightforward implementation of policies across institutions.[108]

There is evidence that even 'light touch' compliance checking results in significant behavioural change. In the UK, the Engineering and Physical Sciences Research Council[109] announced in February 2015 that it would be undertaking "'dip-stick' checking of availability of data underpinning published research",[110] which had been a requirement of their funding since 2011. It has resulted in the **establishment of research data services in institutions across the nation**. The existing NHMRC requirement that research metadata is lodged in a repository and listed in the final report would potentially provide a mechanism for this type of spot-checking to be done in Australia.

The US National Science Foundation (NSF)[111] Dissemination and Sharing of Research Results policy requires the preparation of data management plans describing how the proposal will conform to NSF policy on the dissemination and sharing of research results.[112] The policy notes that "investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants". That said, the NSF recognises that "putting data in a form that others can use may require work that goes above and beyond a stated research activity" and notes that "PIs (principal investigators) may budget ... for the work needed to prepare research data for distribution". Grantees are expected to report on their implementation of the dissemination and sharing of research results. **This policy puts the management and sharing of research data on an equal footing to the research publications that arise from a research project. The ARC and NHMRC should adopt similar expectations and practices**.

3.2.3 MEETING INTERNATIONAL REPOSITORY STANDARDS

There is international support for developing standards through international scientific unions and the International Science Council's (ISC) CODATA.[50] The ISC *Science as a Global Public Good Action Plan 2019-2021* envisages how a federated global commons might develop and operate to the benefit of science.[113] It includes decadal programs for data-driven interdisciplinarity and discusses global data with plans to convene a group of technical experts and representatives of the principal data-holding sectors.[114] The group would explore the potential for global data commons, make recommendations on the

optimal principle of data governance for adoption by the widest numbers of data holders and propose a program for further action.

As a country that produces a disproportionate research output ratio globally, Australia's infrastructure must meet world standards.[115] There is increasing consensus that repositories should meet standards—one example is the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS)[116] commitment statement[117] and the recent call for TRUST principles for digital repositories (Transparency, Responsibility, User focus, Sustainability and Technology).[118] One international standard for repositories that hold research data is the CoreTrustSeal (CTS) certification,[119] which identifies a repository as meeting the multiple CTS Trustworthy Data Repositories Requirements.[120]

Those repositories in the science and social science space with CTS certification can apply for automatic qualification for membership of the World Data System (WDS),[121] an interdisciplinary body of the International Council for Science with a vision for "universal and equitable access to scientific data and information by verifying members against international standards".[122]

There needs to be a plan to implement the Beijing Declaration on Research Data's intention, which includes the principle that "publicly funded research data should be interoperable, and preferable without further manipulation or conversion, to facilitate their broad reuse in scientific research". Such a plan has particular significance for publicly funded research agencies such as Geoscience Australia, the Bureau of Meteorology and CSIRO. An essential step in this direction would be for all publicly funded agencies or facilities that produce observation or other instrumental data to develop an access policy and procedure that give access to unprocessed data at full resolution for research purposes. Such an approach would not prevent the agencies from developing data products for commercial purposes. The licence for research access should recognise the originating data provider's interest in any subsequent commercialisation of research outputs. Such a policy has been adopted by NASA.[123]

Given this increased activity, Australian funders should consider associating funding for repositories with CTS certification with its associated trust assurance. Significant research data assets funded by these funding organisations should be housed in CTS certified repositories as a priority. Also, NCRIS facilities and ARC-Linkage Infrastructure, Equipment and Facilities (LIEF) funded equipment should consider adopting NASA guidelines.

Disciplinary approaches to requiring supporting data be made available

There are some exemplary examples of domains taking the lead in requiring research data that support papers submitted to journals to be available from a trusted repository and linked via identifiers. The Earth and Environmental Science community has made publishing conditional upon data's concurrent availability.[ix] The International Union of Pure and Applied Chemistry endorsed the FAIR Principles in 2019,[x] and the Ecological Society is investigating data sharing.[xi] There has been some work on domain-specific information on implementing FAIR in neuroscience[xii] and some early activity within materials sciences,[xiii] biology and life sciences.[xiv]

More broadly, the publisher association STM (International Association of Scientific, Technical and Medical publishers) designated 2020 as its Research Data Year.[xv] It is working with publishers and other partners to boost the effective sharing of research data by increasing journals with data policies and articles with Data Availability Statements. Another initiative is the Scholarly Link Exchange (Scholix), which is a framework designed to link datasets with research articles.

3.2.4 RETENTION AND DISPOSAL OF DATA—AN INTERNATIONAL PERSPECTIVE

An issue that needs addressing is the increased costs associated with the generation, retention and preservation of research data. The US National Academies of Sciences Committee of Forecasting Costs for Preserving, Archiving, and Promoting Access to Biomedical Data released a report on this issue, making valuable observations about the breadth of considerations in this area.[124] The report frames the question around "the cost of retaining versus replacing data; the value of retained data; the costs of data curation and storage; and potential costs borne by future data users". The report notes that paying attention to and estimating these costs initially allows better investment in scientific knowledge production.

With the introduction of the General Data Protection Regulation (GDPR) in Europe in 2018, new rules apply to individuals' personal information.[125] The European Commission recommends that personal data be kept for "the shortest time possible". However, the situation is different when it comes to research data.[126]

In the UK, there is some consensus on how long research data should be retained. The general research data retention requirements from the UK Research Councils are that data underpinning findings in publications should be accessible for at least 10 years after publication.[108] Various funding councils in the UK require data retention for between 10 and 25 years, depending on the nature of the subject.

A 2019 JISC study,[127] looking at what research data should be kept, found two primary use cases and drivers for what should be kept: research integrity and reproducibility (availability of the data supporting the findings in research); and the potential for reuse (availability of data for sharing with other users). A checklist for appraising research data created by the Digital Curation Centre offers a process to make these decisions.[128]

3.3 What should Australia do, and who should do it?

Australia needs national reform of its research data policies and processes to make them globally competitive. **All players in the research data ecosystem should adopt the FAIR guidelines**, and policies need to be strong, so there are consequences for non-compliance.

To be fully effective, policy and associated practices and protocols need to address data generated by the research community itself and access to other data for research purposes, particularly access to data generated by government and public institutions. An example of a comprehensive data governance landscape analysis is that undertaken in 2017-18 by the British Academy and Royal Society.[5]

Given international developments, designing a comprehensive research policy framework for Australia should not be difficult, although its implementation will challenge many embedded institutional and disciplinary cultures and practices. A survey of researchers in Australia noted that "the responsibility for strengthening transparency and openness must be undertaken not only by scientists and researchers, but also research funding and delivery agencies, and even those beyond the research and innovation sector".[129] Without such system-wide reform, the full potential of data-intensive research will not be achieved. Just as importantly, a robust national research data policy and associated procedures will help mitigate the risks associated with the new data technologies, thereby strengthening trust in science and research more broadly.

Recommendation 3.3 The Australian Government mandates FAIR and CARE for all research data resulting from government funding schemes and ensure that the costs of managing research data are recognised in funding policies for universities, agencies, grants and funding councils. This should be similarly mandated and funded by state and territory governments.

Recommendation 3.4 Discipline and research leaders such as librarians and senior university staff drive cultural change to ensure uptake of FAIR principles and new research data policies. The learned Academies should lead discipline-specific consultation to provide advice on both the appropriate length of retention of data, curation of data and decisions on whether to discard data based on the nature of the research area. Specific consideration should be given to the costs associated with the collection, retention, and preservation of data.

4. Data science — a new enabling discipline

While 'data scientist' is a well-established profession (although not differentiated in Australia's official employment statistics) and 'data science' a recognised field of study, debate continues on whether data science is a discipline and field of research in its own right.

This section discusses the origins of data science (in statistics and computer science) and explores global trends in data science. The section puts forward arguments that it is time to more formally recognise data science as a discipline and in Australia's research data collections. The invisibility of data science in Australia's research data collections means that it is difficult to assess the current state of Australia's data science capability. A detailed mapping of Australia's data science capability would be beneficial for the development of research policy and funding schemes.

Compared to global developments, Australia's investment in data science research is focussed primarily on the application and translation of data science. While this is essential, data science and data-intensive research would benefit from focussed investment in 'fundamental data science'. The section makes some suggestions for possible research areas that could be considered.

4.1 The rise of the data scientist

In 2008, DJ Patil and Jeff Hammerbacher (then at LinkedIn and Facebook respectively) coined the term 'data scientist.'[130] Four years later, the term was popularised in a Harvard Business Review article as "the sexiest job of the 21st century."[131] Since then and particularly over the last five years, there has been an explosion in interest in 'data science'.

Data science has been described as everything from a fad and hype to hope and the answer to every problem. Nevertheless, the influence of data scientists has become pervasive. Using new and powerful tools, data scientists now impact almost every aspect of life, including business, health science and social justice.[6] Academia and research — scientific and non-scientific — have not escaped this data revolution driven by data science and the increasing volumes of data generated by the ubiquitous deployment of information and communication technology (ICT).

Data analysis, particularly sophisticated statistical analysis, has long been a feature of research. The difference today is the scale and range of new tools and methodologies made available by data science. These tools draw on a blend of disciplines, including (but not limited to) mathematics, statistical science, information science and computer sciences. They allow the analysis of both big data and diverse data. The tools are usually automated and often with little or limited transparency. Thus, while data science opens up exciting new research opportunities, including a new paradigm of data-led discovery, the tools also challenge core aspects of research practice (see Chapter 7).

In the context of data-intensive research, this report focuses on data science in research and the academic community. This focus is not to downplay the significance of the broader ramifications and impact of data science. However, a 2019 report by the Royal Society emphasised that a strong and externally engaged *academic* data science capability was vital for the United Kingdom's future economic development.[132] That is no less true for Australia. This report's findings and recommendations could guide the development of Australian data science (and data science research) so that the academic data science community can fulfil its critical role in the future economic, social and environmental development of Australia.

4.2 What is data science?

An agreed consensus on a definition of data science does not exist. Suggestions include "the processes that deal with the extraction of meaning or knowledge from data,"[133] "the study of data...to gain insights and knowledge from any type of data—both structured and unstructured"[134] or, simply, "the study of extracting value from data".[135]

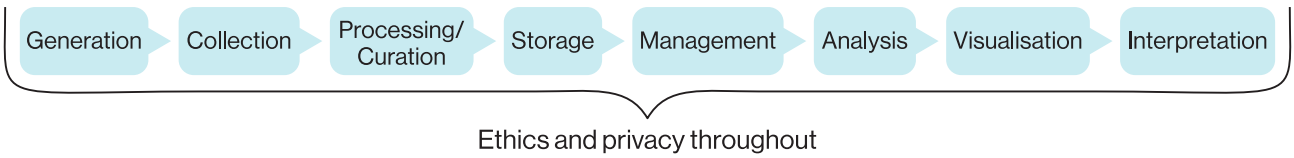
CODATA's *Data Science Journal*—the first refereed journal devoted to data science—regards data science as "the 'science of data:' the evidence-based study of the socio-technical developments and transformations that affect science policy; the conduct and methods of research; and the data systems, standards, and infrastructure that are integral to research."[136]

Such definitions are helpful, but so broad as to capture conventional statistics and well-established disciplines such as information science in not particularly helpful ways. As a result, such encompassing definitions run the risk of obscuring significant issues raised by the development and deployment of new data technologies and methods produced by data science.

Rather than trying for a precise definition of data science, it is more useful to identify the key characteristics of data science. There would seem to be reasonable consensus on the synthesis by Blei and Smyth: "Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention. It emphasizes the value and necessity of approximation and simplification. It values effective communication of the results of a data analysis and of the understanding about the world that we glean from it. It prioritizes an understanding of the optimization algorithms and transparently managing the inevitable trade-off between accuracy and speed. It promotes domain-specific analyses, where data scientists and domain experts work together to balance appropriate assumptions with computationally efficient methods."[137]

Importantly, data science involves the full data life cycle from the generation and/or collection of data through its curation and management to analysis and interpretation and ultimately communication including visualisation.[138, 139] Importantly, throughout this process, attention—and indeed increasing focus—needs to be given to ethical issues and privacy. While the data life cycle is a useful construct, strict linearity to the pipeline should not be assumed. Instead, in practice, the stages interact in more complex ways (see the critique by Williamson[140]).

The Data Life Cycle (after Wing)



4.2.1 DATA SCIENTIST AS A PROFESSION

Notwithstanding the absence of a cogent definition of data science, the profession of data scientist has become well established and well remunerated globally, including in Australia. While there has been some codification of skills,[141] competencies and attributes, perhaps the simplest definition is from the World Economic Forum: "Data scientists are the talent that provide the ability to extract, refine and deploy this new source of value (data) in the global economy," provided 'the global economy' is replaced with 'society', thereby recognising the ubiquitous impact of data science.[142]

Data scientist is perhaps the epitome of how the digital revolution is changing the nature of work. A 2019 report by the Royal Society found that demand for workers with data skills had increased by 230% over the past five years while demand for all workers increased by 36%.[132] Similarly, the 2018 *The Future of Jobs Report* by the World Economic Forum predicted that the top two emerging jobs by 2022 would be data analysts/scientists and AI/ML specialists.[143] In Australia, a 2018 report by Deloitte Access Economics estimated that the data science workforce would grow by 2.4% annually between 2016–17 and 2021–22, compared to a 1.5% annual growth anticipated for the total Australian labour force over the same period.[144] Universities have responded to this external demand through a range of courses in various faculties. While the COVID-19 pandemic has affected these predictions, in many ways the response to the pandemic has reinforced the need for data skills in the workforce and the importance of data scientists and other data professional.[145]

In Australia, data scientists find employment in a diverse range of opportunities and are well remunerated. A search for 'data scientist' on seek.com.au on 3 March 2021 returned 867 positions, an increase of over 100% since a similar search in March 2019.[146] Of the positions advertised on 3 March 2021, 633 were full time with a median annual salary of approximately $120,000 and 5% offering over $200,000 pa. Seek classified 48% of the full-time positions in ICT (predominantly software engineering, developers/programmers or database managers) and 15% in Science and Technology. However, the latter is an underestimate since Seek lists positions in CSIRO and other government research agencies under 'Government and Defence' and university positions under 'Education and Training'. Nevertheless, the breadth of demand for data scientists is evident.

The crisis in higher education caused by the pandemic is suppressing university recruitment, with only four research fellowships advertised in the 3 March sample. However, of the 407 Australian 'data scientist' jobs advertised in March 2019, only 3% were offered within the university sector and were mostly fixed-term appointments with an average salary of $91,000. The demand from the private and government sectors and an ensuing mismatch in salaries (along with an ambiguity of the position of many data scientists within the traditional academic framework) has serious implications for the health of data science in universities.[147]

Although data scientist is recognised on employment web sites such as Seek (and differentiated from other 'data professions' such as data engineer, data analyst), official employment statistics have yet to adjust. A 2020 report on the STEM workforce released by the Office of the Chief Scientist does not even mention 'data scientist', presumably because the underlying employment data does not.[148] It would be very desirable for data scientist (and other data professionals) to be separated from other mathematical/statistical/computational jobs in Australia's official employment statistics.

Recommendation 4.1 The Australian Bureau of Statistics (ABS) should revise Australia's official employment statistics to recognise data scientists and other data professionals explicitly and ensure data scientists working in multidisciplinary areas are accurately represented.

4.2.2 DATA SCIENCE AS A SCIENTIFIC DISCIPLINE

While data scientist is a recognised profession and data science a recognised field of study, there is still debate about whether 'data science' is a new academic or scientific discipline or a field of research in its own right.[149, 150, 151] Indeed, Irizarry and Meng have argued that given the breath of data science and its characteristics, it should be regarded not as a single discipline but rather as "an umbrella term to describe the entire complex and multistep processes used to extract value from data" and this makes it unique.[152, 153] While data science certainly has wide application, this argument could equally be mounted by applied mathematics, statistical science or computer science, all of which are rightly regarded as disciplines but with broad applicability.

Extensive academic literature exists on the nature of disciplines, their evolution and the emergence of new ones.[154, 155, 156] This research has relevance to an objective assessment of the status of data science both currently and how it might develop.

Most established disciplines are characterised by:

- a particular object of research (although this may be shared with other disciplines)
- a body of accumulated specialist knowledge, referring to their object of research, which is specific to them and not generally shared with another discipline
- theories and concepts that can organise the accumulated specialist knowledge effectively
- specific terminologies or a specific technical language adjusted to their research object
- specific research methods according to their particular research requirements
- an integrated set of subjects taught at universities
- organisational recognition both in universities and through professional associations connected to the discipline.

Many of these apply to data science and suggest that it is, at least, on a pathway towards discipline status. Indeed, the evolution of data science is following that of other disciplines. According to Areekkuzhiyil, for example, disciplines evolve, and new ones emerge in response to a number of factors that include:[155]

- the merger of two or more branches of knowledge that then develops its distinct characteristics to form a new discipline
- the convergence of several disciplines into a field of activity that is enriched by the two-way flow of ideas from the constituent disciplines
- recognition of social or professional activities that involve areas of application for several disciplines and is recognised as an independent field of study
- development of significant inventions that transcend existing disciplines, such as nanotechnology.

Many of these features are already met by data science or are emerging as it develops.

Other hallmarks of a discipline exist. There are data science journals, data science conferences and a growing number of data science associations.[136] Many of the associations are professional associations, such as the US Data Science Association,[157] or more narrowly focused and linked to AI or ML, such as the Data Science and AI Association of Australia rather than a more conventional scientific society.[158] Nevertheless, they demonstrate the existence of a community of practice typical of a discipline.

The existence of a community of practice with a body of shared and common expertise is an important characteristic of most disciplines. As argued in the context of psychology but with broader applicability: "The concept of a scientific discipline is an important and enduring one. It implies that there is a body of knowledge to master and skills to be acquired before one can proclaim disciplinary expertise."[159]

Perhaps the missing ingredient is a sense of a theoretical framework or some 'big questions' that help define a scientific discipline. These are emerging and, in some ways, driven by the breadth of the applications of data science. Indeed, being diverse in application does not imply that a common knowledge basis cannot exist. The existence of a common foundation of generic concepts can enhance application by encouraging the transfer of techniques from one discipline area to another.

4.2.3 ORIGINS OF DATA SCIENCE

The origin of (statistical) data science is often traced to a paper by John Tukey in 1962 in which he envisaged a different form of data analysis than conventional statistics;[160] see also Cleveland, who is attributed with the first use of 'data science'.[161] Donho and Cao have described the evolution of data science from its origins in statistics and data analysis.[162, 163]

The contending narrative traces the origins back to work[164] on neural nets in the 1940s and particularly Turing's work[165] in the 1950s. It emphasises the emergence of ML as a critical, even defining, feature of modern data science. The claim is that data science has more in common with other ICT disciplines than the mathematical/statistical sciences.

While these two sides of data science have the common objective of extracting information from data, their philosophy and approach are somewhat different. These differences have important implications for data-intensive research and, indeed the development of data science. The differences arise from both the nature of that data and the methodologies employed. Put most simply, "statistics draws population inferences from a sample, and machine learning finds generalisable predictive patterns".[166] Machine learning aims to extract information by an automated algorithmic approach.

On the other hand, the statistical approach is built on the fundamental concepts of inferences and probability. They share a common feature in attempting to allow the data to be analysed without or minimal *a priori* assumptions. On the statistical side, significant developments have drawn on nonparametric statistics and particularly Bayesian statistics.

"... much of the most innovative data-intensive research comes from a deep collaboration between data scientists and domain experts, which will be encouraged if data science has its own identity."

In reality, both approaches are important, and their combination is a core strength of data science. The most powerful tools are increasingly drawing on both 'parents'. Notably, much of the most innovative data-intensive research comes from a deep collaboration between data scientists and domain experts, which will be encouraged if data science has its own identity.

4.2.4 WHAT DOES THE REST OF THE WORLD TELL US?

International developments suggest a greater recognition of data science as a scientific discipline and as a field of research in its own right than is the current situation in Australia. In 2017, following a five-year program of investment in supporting infrastructure, methodologies and capacity building, the US National Science Foundation recognised data science as a discipline in its own right.[167] Subsequently, and guided by a forward-looking report,[168] NSF has made significant investments through its Harnessing the Data Revolution program in data science research that bridges the contributing disciplines.[169] The US National Institute for Health has an Office of Data Science[170] that in 2018 released a Strategic Plan for Data Science noting that "data science holds significant potential for accelerating the pace of biomedical research."[171]

In the UK, the Royal Society—in its influential report on data skills—implicitly recognised the existence of data science by its vision for the UK to be "a leading data science research nation".[132] The Alan Turing Institute,[172] the UK's national institute for data science and artificial intelligence, has a major program to build "data science at scale".[173]

Sometimes data science is grouped together with artificial intelligence (AI) and machine learning (ML). The Alan Turing Institute rejects this nomenclature and implies that it is limiting to combine data science and AI. Data science also involves more than ML and has a much wider impact than on AI alone. The recent investment[174] of £425 million by the UK's Engineering and Physical Sciences Research Council (EPSRC) to advance artificial intelligence includes investment in data science.[174]

More conventional academic structures are beginning to emerge. The University of Virginia has established a formal School of Data Science,[175] claimed to be the first in the US. The new school aims to be interdisciplinary in focus and build an academic core of expertise in data science across the data life cycle.[176, 177]

In 2018, the University of California at Berkeley announced the creation of a new division of data science[178]—now entitled the Division of Computing, Data Science and Society[179]—headed by an associate provost. Incorporated into the new division are the School of Information, the Berkeley Institute for Data Science and the departments of statistics, electrical engineering and computer science.[180] As with the University of Virginia school, the new division intends to strengthen both the core expertise in data science and enable its wide application in all of UC Berkeley's other divisions and programs. Both are models that Australian universities should consider.

4.3 The state of data science research in Australia

Since data science is not recognised as a research field, it is hard to assess Australian research in data science on conventional research metrics. The 2019 decadal plan for ICT, *Preparing for Australia's Digital Future*, attempted to evaluate the quantity and quality of Australian research in ICT.[181] It concluded that "there is insufficient information to provide definitive answers at a fine-scale about the quality and quantity of Australian research in ICT". Nevertheless, the report identified several ICT research areas of relevance to data science in which there were significant strengths (both in quantity and quality). These included data mining, big data, data analytics and visualisation, and artificial intelligence and machine learning.

A potentially more concerning conclusion on the state of Australian research in machine learning comes from an analysis of patent filings. A 2019 report by IP Australia for the Australian Computer Society analysed patent filings relating to machine learning.[182] It concluded that Australia ranked 10th globally as a patent destination with 295 patent families filed between 2012 to 2017. China was the leading destination with 73% (26,758) of global filings. On the other hand, only 59 patent families were filed by Australian organisations. China again was first with 25,319, followed by the USA (6590), South Korea (1630) and Japan (1503). The top Australian originator in the IP Australia data set was CSIRO with five families, while health care (with 11 filings) was the leading application area. Unfortunately, this comparison of patent filings with patent origination is in line with broader analyses that Australia tends to be an adopter of technology rather than a significant developer.[183]

While some of the subfields of statistics relevant to data science are recognised, notably computational statistics, it is still difficult to assess Australian statistical data science using

conventional metrics. Australian statisticians, notably the late Professor Peter Hall[184] AO FAA FRS and his collaborators, have contributed[185] and continue to contribute to the development of fields such as nonparametric statistics and high-dimensional statistics, which are the foundations of the statistical analysis of big data. Similarly, Australia has a number of researchers who have contributed to the revitalisation of Bayesian statistics, a critical component of statistical data science.

Rectifying this inability to assess the quality of Australian activities in data science would be highly desirable, given the pervasive role that data science plays in data-intensive research and the broader economy and society. How this might be achieved is outlined in Section 4.3.

4.3.1 DATA SCIENCE RESEARCH AND THE FRONTIERS OF DATA SCIENCE

A common issue that arises in many discussions or debates on the status of data science concerns whether there are research questions in data science that are not the province of any of the disciplines that contribute to data science—as with applications, drawing on multiple disciplinary expertise does not exclude a field being identified as a research field in its own right. Indeed, many now well-identified fields of research have evolved in precisely this way. See, for example, Areekkuzhiyil.[155]

Several lists of the challenges of data science have been proposed. In particular, Wing has put forward a list of the top 10 challenges in data science with the intention "to start the community discussing what a broad research agenda for data science might look like". Her list and similar discussions include some deep questions, the resolution of which would have profound implications for data science and data-intensive research. These include the following.

Automation of the front end of the data life cycle. While unlikely to be perceived as exciting as the discovery stages, automation of the initial steps of the cycle involved in preparing and curating data for analysis would have an immense return. It is not enough that the data required for data science applications or data-intensive research is available. Data needs to be cleaned, curated and managed. Ideally, the curation and management should comply with the FAIR principles, including machine accessibility.[46] Expediting this process by developing tools that researchers could use themselves or in data repositories would be very useful. Some tools have been developed, and standards are beginning to emerge. Further research is needed; however, there is a review of developments in the automation of curation, including some Australian work sponsored by the Data to Decisions CRC.[186, 187] Australian developments need to be carefully coordinated with international developments, and Australia must be active in the development of standards.

Methods to enhance privacy protection. While no existing de-identification method or algorithm is foolproof,[71] data scientists and others are beginning to explore methods that enhance privacy protection and increase security through encryption or other methods. As noted by Wing, a novel example of differential privacy was introduced for the US 2020 census. Noise was deliberately added to a query result, thereby maintaining, at least to a greater extent, the privacy of individuals participating in the census.

Research on privacy-preserving analytics is an area where CSIRO's Data 61 and collaborators are active with some recognised significant research achievements. For example, "the world's most efficient blockchain protocol that is both secure against quantum computers and protects the privacy of its users and their transactions". This research contributes significantly to data policy, such as through the development of a data de-identification guide.[188]

Federated analytics is another approach that,[189] while not a panacea,[71] offers enhanced privacy protection. Under a federated analytical approach, data is retained locally and only the outcomes from the analysis are transferred to a central aggregator or collection

point. An example is OpenSAFELY, an analytics platform developed to analyse health records in the UK's National Health System during the coronavirus emergency.[190, 191] Notably, the software is open source and is available for scientific review.

More computationally efficient algorithms to handle the steadily increasing volumes and variety of data. Of particular importance to many research areas would be analysing what Wing terms "precious data". Precious data includes data that is expensive to generate, for example, by advanced radio telescopes and data streams containing rare events with a low signal to noise ratio. As Wing noted, focused research in combining multiple sources of data would have an extraordinary impact. New methods to handle noisy and incomplete data would also be of significance to data-intensive research and more broadly. More generally, integration between simulation and data analytics offers many opportunities.[192]

As crucial as these examples (and others on Wing's list or similar lists) are individually, more significant are three common features they exhibit:

- they involve two or more disciplines from those that underpin data science in a way that is more than the 'sum of the parts'
- their impact is more significant on data science than on any of the contributing disciplines
- the outputs are generic with potential application across a wide range of domains.

Taken together, these characteristics define a very respectable, challenging and highly valuable 'research agenda' for data science as a field of research in its own right.

"... perhaps the most significant sign of data science maturing as a research field in its own right is the emergence of active research on the foundations of data science itself and the tools that define the discipline and its applications."

However, perhaps the most significant sign of data science maturing as a research field in its own right is the emergence of active research on the *foundations* of data science itself and the tools that define the discipline and its applications. This subfield of data science is asking some fundamental questions of data science itself, such as:

- Why does machine learning and deep learning work (and perhaps more importantly, under what circumstances will they not work)?
- How can data science move from a focus on identifying correlations to the identification of causality? From the perspective of data-intensive research, this would be particularly significant since science remains driven by understanding the causality between events and observations, not just their correlation.

Machine learning and now deep learning is proving to be incredibly successful both in data-intensive research and in wider applications of data science. However, research on machine learning and deep learning remains largely empirical.[193] With a *theoretical* understanding of machine learning and deep learning, only now is there beginning to be a serious focus of research on why deep learning algorithms work better than one might initially expect.[194] Such theoretical research offers the potential for more efficient and robust algorithms. While new computationally efficient algorithms will not alone relax the computational limits emerging in deep learning, they would help considerably.[195]

Currently, machine learning's strength lies in its ability to detect patterns and correlations, particularly in large data sets. While often sufficient in business applications and not insignificant in research, integrating causal reasoning into machine learning (see, for example Pearl, Luo et. al.), would be valuable in a wide range of fields.[196, 197] Since most real-world observations are due to multiple factors, new tools to explore multiple causal inferences would be particularly useful.[198]

Indeed, suppose one wanted to define data science by a 'great challenge' (such as understanding the universe defines astronomy). In that case, a candidate might understand and automate Judea Pearl's 'science of causal inference' as described in his book, *The Book of Why*.[199] While fundamental to the development of AI, such developments have far wider ramifications on how data is processed, analysed and interpreted. Pearl argues a critical element of implementing the science of causal inference is the ability to generate and analyse counterfactuals, that is, to analyse what may have happened if some event had not occurred or had occurred differently.[200] While essential to the development of 'strong AI', such a way of analysing and interpreting data changes the nature of scientific prediction, particularly in contentious areas such as climate change or other wicked problems.

Two final comments on recognising data science as a field of research are important.

Not all research necessary for data-intensive research or data science applications should be captured within a recognised field of data science. Some crucial areas of research are better recognised as genuinely interdisciplinary between data science (as a discipline) and one or more other disciplines. Important examples are:

- research in data ethics involves data science, law and ethics
- the design of computational systems to support advanced data analytics requires close collaboration between data science, computer engineering and software engineering.

A greater focus within data science on foundational research does not (or should not) devalue the importance of the research that data scientists conduct with domain specialists. This collaboration remains vital. Generic data science tools and methods often need to be adapted, even significantly, in the light of the nature of data being analysed or even generated. Moreover, the necessary adaptation or modification may be needed along the whole of the data life cycle, with that integrated view being a key contribution that data scientists bring to such collaborations. Conversely, such collaborations and insights they generate are essential to developing the research agenda for foundational data science research.

These issues concerning the balance between 'basic' and 'applied' data science research have important implications for assessing data science research, including the assessment of individual data scientists. However, they do not detract from a conclusion that it is time to recognise data science as a discipline and field of research in its own right. Blei and Smythe described data science as a child of computer science and statistics.[137] Continuing this metaphor, it is time to let data science grow up retaining its parents' genes but now taking on its own identity. How this should be carried out in Australia and the issues that will need to be resolved are discussed in the next section.

4.4 What should Australia do, and who should do it?

For a field that relies on data, it is ironic that there is so little data that allows a detailed assessment of Australian activity in data science, the scale of investment in data science research, the balance of that research between foundational research and applied or translational research, the quality of that research, and its international competitiveness. Indeed, data science is essentially invisible in Australian research data. Correcting this situation is crucial, not only for data science to prosper in its own right but also to ensure that the policies and programs supporting Australian research, as it becomes increasingly data-dependent, are appropriate and remain relevant.

In this section, we discuss five critical actions that need to be taken urgently. These are to:

- recognise data science as a discipline and field of research in its own right
- recognise data science formally in Australia's research data collections and research assessment
- map Australia's data science capability

- establish a program to support foundational research in data science
- address gender imbalance in data science.

Specific action is needed by:

- universities and other research agencies
- funding agencies
- the data science community itself
- the research assessment system.

4.4.1 RECOGNISE DATA SCIENCE AS A DISCIPLINE AND FIELD OF RESEARCH

Without recognition and effective implementation, it will be harder to implement other critical actions to develop data science in Australia. This recognition needs to come from all elements of the Australian research system, but particularly universities. The recognition also needs to be reflected in changes in the policies and programs that underpin the system.

"How universities respond to the emergence of data science as a discipline and field of research will be critical to the long-term health and development of data science in Australia"

How universities respond to the emergence of data science as a discipline and field of research will be critical to the long-term health and development of data science in Australia. This is a complex issue with several critical threads—ranging from data scientists' recognition and employment conditions to how a university organises data science internally and presents a coherent external view of its data science capability. Organisational structure and culture matter since, as McKinsey has demonstrated: "organizational culture can accelerate the application of analytics [and] amplify its power."[201]

Business has explored a wide range of organisational structures to enhance data science's business impact.[202, 203] The inherent challenge is to balance the creation of a critical mass of data scientists (giving data science visibility) and their decentralisation through business units to engage more effectively with those units. Some of these models may be relevant to universities where the preferred model in Australian research is a 'centre' or 'institute' that cuts across the more formal organisational structure—faculties, schools, departments—of the university.

The advantage of this structure lies in recognising the increasing pervasiveness of data science, and that data science should not be 'captured' by any of the conventional disciplines. On the other hand, by lying outside the formal governance structures, resourcing can become problematic. As a result, many centres rely heavily on external funding, exacerbating the difficulties in retaining data scientists and often resulting in the perception, if not the reality, that their primary purpose is the translation of data science research. Another unfortunate consequence can be the fragmentation of the university's data science capability, making access from outside more difficult.

The new models at the University of Virginia (UVA) and UC Berkeley (described in Section 4.1.4) are designed to bring data science within the conventional organisational structure while respecting its pervasiveness. While the UVA and UC Berkeley developments are worth close consideration, each Australian university needs to develop its appropriate organisational arrangement (including the funding) of data science. Whatever that structure, it must be designed, as at UVA, to be porous, thereby facilitating internal and external collaboration while building a critical mass of data scientists and supporting foundational data science.

Recommendation 4.2 Data science be recognised as a discipline and a field of research in its own right. In particular, Australian Universities should recognise data science as an academic discipline and review their organisational structure for data science so as to enable a cohesive data science discipline whilst encouraging collaboration within the institution and externally.

Recommendation 4.3 ARDC and NCRIS strengthen coordination between Australian data science centres with direct infrastructure support for the Australian Data Science Network.

Recommendation 4.4 The data science research community (possibly under the Academy or STA auspices) establish a Data Science Society with an inclusive membership.

4.4.2 RECOGNISE DATA SCIENCE IN AUSTRALIA'S RESEARCH DATA COLLECTIONS AND ASSESSMENT SCHEMES

The Australian and New Zealand Standard Research Classification is a set of three data collections:[204] Type of Activity (ToA); Fields of Research (FoR), and Socio-Economic Objectives (SEO). The purpose is to collect data on research and development activities, including investment, in Australia and enable the comparison of that data between Australia and New Zealand and globally, through the Organisation for Economic Co-operation and Development (OECD).

The disaggregation of Australian research activity by the FoR codes has become the foundation on which fields of research are compared and analysed.[205] Assessment schemes such as Excellence in Research for Australia (ERA)[206] and Engagement and Impact (EI)[207] are also based on the FoR codes. The structure is hierarchical with 23 top-level codes below which there is increasing refinement at 4- and 6-digit levels.

While the FoR code structure was substantially modified in 2020,[208] the current structure does not allow an accessible or integrated assessment of data science research. As it stands, 'data science' is explicitly mentioned at:

- 4605 Data Management and Data Science; 460501 Data Engineering and Data Science (below 46 Information and Computing Science;) and
- 490509 Statistical Data Science (below 49 Mathematical Sciences; 4905 Statistics).

and implicitly in a number of other 4-level codes (e.g. Computational Biosciences (3102), Geoinformatics (3704), Econometrics (3802)).

In contrast, machine learning, a critical component of data science and comparable in importance to statistical data science, has its own 4-digit code (4611) with considerable refinement at the 6-digit level. Other research areas central to the data science life cycle are distributed in Information and Computing Sciences or elsewhere. For example, digital curation and preservation, a critical component of the front-end of the life cycle, appears at a 6-digit level (461001) in Library and Information Studies (4610).

A comprehensive review of the codes in the immediate future to better reflect data science as a field of research in its own right is unlikely to be feasible given the extensive consultation required to produce even the modifications made. A more formal integration or at least reporting of codes relevant to data science would be invaluable as both an assessment of data science research and as an important input to major policy decisions that will need to be made in the next few years, such as the recently announced refresh of the national research infrastructure roadmap (see Section 5.3.2).

A possible interim solution would be to recognise data science at 4-level in Information and Computational Sciences. Under this code, at 6-level, should be aggregated all areas

of research relevant to the data life cycle, including machine learning, deep learning, statistical data science, data management and data curation that is aimed at the development of *generic* knowledge and data science tools.

The emphasis on generic understanding also helps differentiate data science research (classified in this proposed 4-code) from more domain-specific work that would continue to be recorded in a range of existing fields, such as bioinformatics or geoinformatics. In this sense, the proposed classification is similar to the situation pertaining to applied mathematics.

We concede that this will be perceived as contentious by some areas, such as statistics, with the explicit inclusion of statistical data science under Information and Computational Sciences, and computer science with the 'demotion' of machine learning. Indeed, research in both fields fits naturally and appropriately in core statistics and computer science. However, in both cases, from the perspective of data science, they have a common purpose: to extract information from data predominantly by computational means.

Notably, the proposed structure is similar to the structure recommended by a review of the Sectional Committees of the Australian Academy of Science—groups of Academy Fellows that represent diverse specialised fields.[209] After considerable debate, this review recommended data science be placed under Sectional Committee 6: Information and Communication Sciences rather than Sectional Committee 1: Mathematics. However, candidates with a mathematical focus on data science should continue to be considered by the latter Sectional Committee.

Importantly, the *impact* of data science research must continue to include the assessment of its impact on a domain area. Indeed, in many cases, the impact of data science should be assessed fully within that domain. A critical feature of research classified as 'data science research' should be its likelihood of impacting multiple domains. Again, that is nothing new. It is already practised in fields such as applied mathematics, statistics and computer science.

Recommendation 4.5 The ABS, as an interim solution to a full review of the Australian Field of Research (FOR) codes, recognise data science at 4-level in Information and Computational Sciences. Under this code at 6-level should be aggregated all areas of research relevant to the data life cycle, including machine learning, deep learning, statistical data science, data management and data curation aimed at the development of generic knowledge and data science tools.

4.4.3 MAP AUSTRALIA'S DATA SCIENCE CAPABILITY

A mapping of Australia's data science capability, analogous to the recent report on Australia's climate science capability, should be conducted.[210] Doing so would establish a baseline from which to guide data science development in Australia. Terms of reference would need to be developed and refined in consultation with the data science community and key agencies, but should include:

- estimating the number of data scientists employed in publicly funded research organisations and universities in Australia, whether they are working in sub-units that are primarily focused on data science or are embedded in other (domain-based) research units
- assessing the distribution of data science research over the spectrum: foundational, strategic basic research, applied research and the translation of existing data science; and with regard to the sub-components of the data life cycle.

Importantly, the mapping should adopt a broad definition of data science (and data scientist), including bioinformaticians, geoinformaticians and other scientists practising data science. Inclusion of international experts on the oversight or expert working group would be highly desirable.

Recommendation 4.6 The Academy (possibly in collaboration with ATSE) seeks funding to map Australia's data science capability analogous to the recent report on Australia's climate science capability.

4.3.4 ESTABLISH A PROGRAM IN FOUNDATIONAL RESEARCH IN DATA SCIENCE

While the lack of relevant research data prevents a definitive data-based conclusion, our consultations suggest that Australia is lagging in other jurisdictions in investment in research on the foundations of data science. The importance of establishing a vibrant program in foundational research in data science was emphasised in a 2016 report from an advisory committee established by the NSF.[168] The report highlighted that foundational research in data science enables "tremendous opportunities to adapt data to provide value beyond simple questions". Adopting this report's recommendations has led the NSF to make significant investments in data science research that bridges the contributing disciplines.

To fully exploit data science's potential for both research and its broader applications, Australia needs to do the same. Despite the passage of five years, the NSF report's conclusions remain valid for Australia and particularly for the ARC and the NHMRC (given the importance of data science to advancing fundamental biomedical fields, particularly genomics). While the ARC and the NHMRC are the relevant agencies to pursue and fund a significant research agenda in foundational data science, the ARDC has a potential role through its Platforms initiative[211] running from 2019 to 2023. Although the objective of the program is somewhat more functional, with the goal of increasing "the number of researchers with access to platforms", establishing a program in foundational research in data science aligns with its aspirations to establish "transformative platforms that will enable radical changes in the way research is conducted, or dramatically increase the speed in which research is done".[212]

Recommendation 4.7 The ARC (supported by the NHMRC) strengthen Australian research in foundational data science through an Australian program similar to the NSF's *Transdisciplinary Research In Principles Of Data Science* (TRIPODS) program and by considering a dedicated call for proposals to establish an ARC Centre in Data Science. At least half the research activity should be in foundational data science research and its translation to areas of research with significant data challenges beyond the capacity of existing data science.

4.4.5 ADDRESS GENDER EQUITY IN DATA SCIENCE: AN URGENT ISSUE

The benefits of achieving diverse research teams have been discussed extensively.[213] In brief, gender diversity increases a group's problem-solving capabilities,[214] innovation[215] and overall economic prosperity.[216] Despite these known benefits, women remain underrepresented in STEM, and data science is no different,[217]–with estimates suggesting women comprise 15% to 26% of data science professionals.[218, 219] The susceptibility of ML and AI algorithms to bias (including gender and racial), such as inappropriately chosen training sets, is a particularly pertinent and urgent issue for data science.

In addition to addressing the broader challenges[220] associated with achieving gender equity in STEM, implementation of the *Women in STEM Decadal Plan*[213] should give explicit and urgent attention to women in data science guided by the policy work[221] of the Alan Turing Institute to address the imbalance of women in data science. Initiatives such as Women in Data Science[222] (WiDS) are important and should be encouraged as they give profile to women working in data science. In this regard, Australian data science is well-positioned since women hold several senior positions including as Fellows of the Australian Academy of Science. WiDS or similar events can also help alleviate perceptions[218] that data science is a 'nerdy' computer-based discipline, too abstract for people to consider

Dr Kate Robertson of the Geological Survey of South Australia setting up a magnetotelluric instrument in the Flinders Ranges during the AusLAMP South Australia Program. CREDIT: MILLICENT CROWE / AUSCOPE

as a career pathway or too focused on business applications. One possible solution to overcome this final perception is to profile applications of data science to social issues such as Data for Social Good[223] and Big Data for Sustainable Development.[224]

Recommendation 4.8 The implementation of the Decadal Plan for Women in STEM needs to actively address the gender imbalance in data science by enhancing current initiatives and giving greater prominence to applications other than business.

While improving gender equity in data science is the final recommendation of this report, it might be the most critical. Indeed, gender bias in data science may be more pernicious than just the imbalance in the population of data scientists—as crucial as that is.

There is considerable evidence that much population data collected is heavily biased against women, to the point where women are often invisible.[225] This has significant implications for machine learning and deep learning models. Since such models predict what they have been trained to predict, their predictions are only as good as the data they are trained with. Failure to recognise bias in training used in machine learning or deep learning can yield misleading and even harmful outcomes. While the ultimate solution must involve redesign of public data collections, the data science community can play an important role by developing and adopting codes of practice such as the Position Statement on Population Data Science,[226] debating and appreciating the values that should guide data scientists and ensuring wherever possible the underlying data is available for scrutiny.[227]

5. An integrated eResearch infrastructure is critical

Data-intensive research is underpinned and enabled by eResearch infrastructure consisting of physical elements (compute facilities, data stores, networks), software and standards and advanced tools.

This section describes the current structure of Australia's eResearch infrastructure and global trends—both driven by data-intensive research and by technological developments—that will need to be considered in the next phase of investment. While Australia has invested significantly in its eResearch infrastructure, a more strategic and integrated approach will be required in the next phase of investment.

The section makes several recommendations for consideration in the current refresh of Australia's National Research Infrastructure Roadmap.

5.1 Research infrastructure has become data intensive

The Australian Government's 2016 National Research Infrastructure Roadmap emphasised the importance of Australia's eResearch infrastructure* for modern research. The roadmap described eResearch infrastructure as a "cross-cutting capability that serves research collaboration, modelling, data, and data analysis needs. It comprises advanced networks, identity, access and authentication services, high performance and cloud computing resources, management of and access to research data; … **and the integration of all those elements** to create digital environments researchers use every day".[7] In short, it provides a roadmap closely aligned to Gray's vision of 2007. While increasingly an enabler of all research, eResearch infrastructure is vital to data-intensive research. Data-intensive research often pushes the limits of existing eResearch infrastructure, stimulates and justifies new investments, and spurs innovation with significant spillover to broader use.

Significant investments in Australia's eResearch infrastructure have been made over the past decade or so, but establishing a precise figure is difficult. The Australian Government has been the primary source, mainly through the National Collaborative Research Infrastructure Strategy (NCRIS).[85] Since 2017–18, NCRIS has directly invested $219 million to support core compute and data services (see Section 5.1.1), with an additional $8.9 million announced in 2020 to develop the HASS and Indigenous eResearch platforms[228]. In addition, many if not all of the other NCRIS investments have significant elements of eResearch infrastructure within them, and all generate research data. Investments have also been made and are continuing to be made by institutions and increasingly by state and territory governments. These investments have created a

* Note the term eResearch infrastructure is synonymous with cyber infrastructure (used more often in the US).

strong foundation for data-intensive research in Australia. The challenge ahead is building and adapting that infrastructure to exploit data-intensive research across a widening range of research fields. It is not clear that the approach that has served Australian research well to date will continue to do so in the decade ahead.

According to Innovation, Science and Economic Development Canada: "To maintain Canada's science and research excellence and make sure we can benefit from [data], we must coordinate our national computing power and connectivity with the best software and storage services for data."[229] This is also true of Australia.

5.2 Australia's current eResearch infrastructure

As emphasised above, eResearch infrastructure involves integrating various elements: hardware and software; compute and data storage facilities; and policies and practice. However, to assess the current status of Australia's eResearch infrastructure and how it should develop, it is helpful to consider the various elements separately.

5.2.1. THE DIGITAL DATA AND ERESEARCH PLATFORMS (DDERP)

Colloquially but appropriately referred to as 'the four cornerstones', DDeRP consists of the following.

Australia's Academic and Research Network (AARNet)

Australia's Academic and Research Network (AARNet) links all Australian universities and CSIRO sites with a minimum connectivity of 10Gbps (gigabit per second) and between major capital cities up to 100Gbps. In addition, AARNet holds right-to-use licenses on sub-sea cables linking Australia to North America and has invested directly in the Indigo consortium that laid a cable linking Perth to Singapore.[230] This connectivity—both nationally and internationally—plays a critical role in data-intensive research by transferring data and enabling remote access to computational resources and other services. On the other hand, inequities remain in access, particularly from rural and remote regions, thereby significantly limiting research, research collaboration, and applications.[231]

Australian Access Federation (AAF)

The Australian Access Federation (AAF) provides access, authentication and identity verification services, enabling trusted electronic communication between education and research institutions nationally and internationally.[232]

National Computational Infrastructure (NCI) and Pawsey Supercomputing Centre

Australia's two Tier 1 high-performance computing (HPC) facilities are the National Computational Infrastructure (NCI) at the ANU in Canberra and the Pawsey Supercomputing Centre in Perth.

Since 2018, both NCI and Pawsey have each received $70 million through NCRIS to replace their supercomputers. NCI's new system, called Gadi, went live in January 2020 and in June 2020 was rated 24th on the global TOP500 List of research supercomputers.[233] Pawsey is in the process of a similar capital refresh with the first phase of its new supercomputer—Setonix—due for operation in late 2021.[234] When fully installed in 2022, Setonix is predicted to have a peak speed that will position it well into the top 20. As a result, Australia's HPC facilities currently compare very favourably with other nations.

Australian Research Data Commons (ARDC)

The Australian Research Data Commons (ARDC) was created in 2018 from a merger of (the previously NCRIS-funded) Australian National Data Service (ANDS), Research Data Services (RDS) and Nectar (National eResearch Collaboration Tools and Resources).

ARDC has a broad remit in research data and the aspiration "to facilitate the involvement of the sector as a whole in the development and implementation of a national research data commons".[235] ARDC operates several national information services, including a national catalogue of global identifiers and national scientific terminology registers. ARDC responds to proposals from the community through its funding calls, which have a particular focus on projects that align to the ARDC's strategic themes of data and services, platforms and software, storage and compute and people and policy.[236, 237, 238, 239] ARDC also has an increasing and important role in ensuring that Australia's research data is globally linked.[240]

While the roles of AARNet (network), NCI and Pawsey (high-performance computing) and the ARDC (data) are ostensibly differentiated, in practice, they significantly overlap.

- AARNet operates CloudStor offering 1TB of free storage to individual researchers at AARNet-connected institutions[241]
- Both NCI and Pawsey host significant data stores. NCI hosts and curates over 20 petabytes of data, including substantial national geoscience and environmental data, and has developed a suite of sophisticated tools to enable researchers to access and use this data. Similarly, Pawsey currently houses more than 40 petabytes of data storage resources, including data from the West Australian based radio telescopes, and has recently been designated Australia's space data facility.
- ARDC has continued to operate and, in 2019 refreshed, the Nectar Research Cloud, which provides cloud computing infrastructure and associated services.[242]

This overlap is understandable and to be expected. Despite the bandwidth offered by AARNet, the steady increase in data volumes maintains a tension between the cost of data transmission to compute facilities and the cost of storing the data at a facility. Similarly, cloud computing and HPC are not distinct approaches to the compute requirements of data-intensive research. Indeed, both NCI and Pawsey operate clouds in addition to (and as part of) their more 'conventional' HPC. The challenges are to ensure, to the greatest possible extent, that unnecessary duplication is avoided, and information is provided to the research community to ensure that the most appropriate and cost-effective choices are made. Thus, the announcement at the 2021 eResearch NZ conference of the formalisation of the DDeRP Executive Group is welcome.[243] This should hopefully mitigate unnecessary duplication and improve the coordination of current investments, actions that should strengthen the case for the future investment that will be needed.

Notwithstanding their critical role as the 'cornerstones' of Australia's eResearch infrastructure, the financial sustainability of the DDeRP is mixed and will need to be addressed in the next phase of investment (see Section 5.3.2). AARNet and AAF are reasonably secure, with AARNet having a solid balance sheet.[244, 245] Importantly for data-intensive research, AARNet's business model is not a typical telco model based on volume-based charging but a subscription model with subscriptions tied to a broad index of research activity. Consequently, the cost of data transmission on AARNet is essentially independent of the volume being transmitted. This is critical for data-intensive research. For example, research using the Square Kilometre Array will depend critically on AARNet's network to transfer extremely large amounts of data and enable international research collaboration.[246]

While the core supercomputers of NCI and Pawsey have been secured for at least the immediate future, both facilities face significant challenges meeting the capital requirements for associated infrastructure such as storage and operational costs. The current arrangements by which such costs are met by an annual grant from NCRIS (primarily to support merit-based access by individual researchers), multiple research partners and, in Pawsey's case the WA State Government, are not sustainable. More

National Computational Infrastructure (NCI) Gadi supercomputer.
CREDIT: NATIONAL COMPUTATIONAL INFRASTRUCTURE

significantly, the question of whether Australia can continue to afford two Tier 1 facilities will continue to be relevant.

Despite its critical mission, ARDC appears to be the least secure of the DDeRP platforms. While now a company limited by guarantee with primarily universities as members, its balance sheet appears to consist almost entirely of Australian Government funding made available through the current round of NCRIS funding.[247] Unfortunately, this vulnerability extends to many of the ARDC's initiatives that, for the most part, are based on leveraging other investment; see more specific comments on ARDC's platforms at the end of Section 5.2.3.

5.2.2. DATA REPOSITORIES AND AGGREGATORS

The fragmentation and complexity of Australia's eResearch infrastructure are most apparent when observing the data 'sub-infrastructure'. Our consultations suggest there is a confusing landscape of data repositories, aggregators and curators—a 'zoo of acronyms' in one commentator's words—with complex funding arrangements. As recommended in Section 3, data-intensive research would be significantly facilitated if all bodies in this sub-infrastructure committed to the FAIR principles and developed a plan to implement them, including protocols to enable access for research purposes.

Many of Australia's national data archives are underpinned by infrastructure that is directly or indirectly funded by the Australian Government. These data archives include

repositories of important national data held and managed by Geoscience Australia, the Bureau of Meteorology, the Australian Bureau of Statistics (ABS), Australian Institute of Health and Welfare (AIHW) and the publicly funded research agencies (CSIRO, ANSTO and the Australian Institute of Marine Science).[79, 248, 249, 250, 251, 252, 253] State and territory governments or state-based agencies hold significant data of importance to research, particularly in health, education, justice and the environment.

Critical data repositories, aggregators, curators and increasingly data analytical hubs for Australian data-intensive research include several NCRIS facilities, notably the Integrated Marine Observing System (IMOS) for ocean sciences, Terrestrial Ecosystem Research Network (TERN) for environmental sciences, Atlas of Living Australia (ALA) for biodiversity, the Australian Urban Research Infrastructure Network (AURIN) for urban planning and social science, BioPlatforms Australia for biological science and Population Health Research Network (PHRN) for population health.[254, 255, 256, 257, 258, 259] Other NCRIS facilities, such as AuScope for earth sciences and Astronomy Australia, play essential roles by overseeing investment strategies that include data repositories or associated facilities.[260, 261] It is arguable that all NCRIS facilities are now, in the digital age, producers of data and need to adopt the FAIR research data guidelines and have explicit data management plans as part of their operations.

Funding to support data acquisition, associated data infrastructure and data analysis for genomics and health research are flowing through the Genomics Health Futures Mission and other Medical Research Future Fund initiatives.[67, 262] However, these initiatives appear to have little strategic engagement with other aspects of the eResearch infrastructure. As discussed in Section 3, these initiatives will be impeded unless the critical issues of research access to government and health data can be resolved.

Whatever the source of funding, funding agreements underpinning data repositories and aggregators should explicitly recognise the importance of data curation, stewardship and storage and, as recommended in Section 3, the adoption of the FAIR principles.[87, 263]

5.2.3. INFRASTRUCTURE TO FIND DATA FROM RESEARCH AND FOR RESEARCH

The challenge of navigating the ever-growing plethora of data sources *from* research or data of potential interest *for* research is leading to the emergence of technologically driven search tools and associated infrastructure.

Researchers engage in a 3D visualisation of tectonic plate subduction simulation at Monash University. CREDIT: AUSCOPE / PROF. LOUIS MORESI AND OWEN KALUZA

In Australia, ARDC supports Research Data Australia (RDA), enabling a search of a range of Australian data sources by multiple facets, including discipline, geographic region, time period, licence and select keywords.[264] As of June 2021, RDA covers 178,998 datasets from 100 contributors. The ABS has launched several Data Integration Projects that integrate various government datasets and are available to appropriately qualified researchers through ABS's DataLab.[250, 265]

In 2018, Google AI launched Google Dataset to improve data discoverability and enable easy access to data from anywhere in the world.[266, 267] Google Dataset is underpinned by guidelines based on a community-maintained standard, Schema.org, that allow data providers to describe their data so that Google and other search engines can understand its content.[268] ARDC has implemented the standard at RDA, so all Australian research organisations contributing their records to RDA are syndicated directly to Google Dataset Search. For example, many Australian data repositories, such as TERN, have adopted the Schema.org standard and are working with Google to link Australian data into global tools such as Google Earth.[269, 270]

Compared to other forms of search, database search is an active area of research with various sophisticated tools beginning to emerge; see for example this recent survey.[271] Such developments, particularly as they become mainstream, have significant implications for data-intensive research and research practices more broadly. Significantly, this research has raised issues for the producers and curators of datasets to ensure that their practices align with the FAIR Guiding Principles (see Section 3). Critically, the emphasis has been placed on the need for up-skilling of researchers and particularly the training of research students to know how to use these emerging tools and the limitations that exist.

Neither an RDA or a Google Dataset search explicitly certifies the quality or veracity of the data accessed. Indeed, this is difficult outside the context of very specialised discipline-specific repositories. However, RDA does provide guidance for a minimum set of standards.[272] These standards generally address the quality of data description and expression, but not the quality (accuracy) of the data content. While the FAIR principles note the importance of data provenance, researchers ultimately need to convince themselves of the veracity and reliability of any data they use, whether generated by them or sourced from a database or data integrator. They then need to communicate the data limitations in any ensuing publication.

ARDC has launched two important parallel programs to support the aim of producing and managing findable and accessible research data:

- the National Data Assets program aims to develop "a portfolio of national-scale data assets that support leading-edge research".[273]
- the ARDC Platforms program aims "to connect and provide access to a range of resources to researchers and industry".[237] ARDC defines a platform as "a set of services, often with associated integration and/or orchestration functions and connections to specific data resources, that is intended to enable researchers to carry out some of their research activities".

Underpinning ARDC Platforms is the concept of creating, within the comprehensive 'research data commons', discipline-focused domains or commons that integrate the various elements of eResearch infrastructure and allow for legitimate differences between disciplines and the relative maturity of a discipline. In doing so, they can address cultural impediments and drive researcher upskilling.

The most advanced and comprehensive example of a discipline-focused data commons is the Australian BioCommons, which was initiated through the investment of $20 million of NCRIS funding by BioPlatforms Australia.[258, 274] The Australian BioCommons aims to deliver digital support for Australian research on the molecular basis of life across environmental, agricultural and biomedical sciences.

Not all platform investments by ARDC are on the scale of the Australian BioCommons. These potentially important initiatives run the risk they will add another layer of complexity to the Australian research data ecosystem with little long-term financial sustainability and viability given the limited commitments made by ARDC. The platforms' effectiveness might also be enhanced by stronger engagement with other planning and priority-setting mechanisms within the Australian research sector. Suggestions and recommendations to this effect are discussed in section 5.3.2.

Internationally, efforts to ensure data quality have emerged. In 2017, the World Data System of the International Science Council (WDS) and the Data Seal of Approval (DSA) launched a new certification organisation, CoreTrustSeal.[120, 122] As of July 2021, according to the website, only five Australian data repositories had achieved certification. The repositories with accreditation are the Australian Antarctic Data Centre, Space Weather Services, CSIRO Data Access Portal, Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Australian Data Archive.[251, 275, 276, 277, 278] To promote certification, ARDC has established a community of practice to support trusted data repositories; the group provides peer support for repositories applying for CoreTrustSeal certification.[279]

5.3 What is global 'best practice' telling us?

5.3.1. NATIONAL INITIATIVES: A COMPARISON

Since the articulation in 2017 by the US National Science Foundation of a bold vision for research data infrastructure that the USA would need for the 21st century, other nations have developed similar visions and plans.[280] A 2019 report by New Zealand eScience Infrastructure (NeSI) assessed eight national eResearch ecosystems: Australia, Canada, Netherlands, New Zealand, Singapore, Sweden, Switzerland and the United Kingdom.[281, 282] While all involved the same elements, the report demonstrated a complex landscape, reflecting each country's history and organisational traits. Broadly, the NeSI analysis identified the following three possible approaches to the development and delivery of eResearch infrastructure and associated services.

Multiple organisations. This was the most common approach, into which Australia fell, and was characterised by:

- weak coordination and complementarity in services
- the fact that researchers, themselves, need to navigate, identify and integrate the services they require
- competition for funding exacerbating complexity and fragmentation
- unclear roles of each organisation and unbalanced investment.

A single national private provider coordinating all eResearch infrastructure efforts and operating with an explicit government mandate—as exemplified by Finland—which:

- ensures that the full range of eResearch services are available to all researchers
- minimises competition and overlap
- ensures that demanding service requirements are met.[283]

Federated structures that were emerging and seemed promising as a bridge between multiple, relatively independent providers and a single national provider. Such organisations:

- could avoid competition and overlap
- leverage complementary investments, but
- struggled with issues regarding governance, coordination, accountability and funding, which can, in turn, limit the services offered and effectiveness.

Since the NeSI report was published, developments in Europe and Canada have moved towards more formal federated structures, suggesting this could be an option for Australia. The European Open Science Cloud (EOSC) aims to provide the European scientific community with "seamless access to data and interoperable services that address the whole research data cycle, from discovery and mining to storage, management analysis and re-use across borders and scientific disciplines".[284] EOSC is tightly connected with Europe's Horizon 2020 R&D projects including Europe's exascale initiatives.[285] Similarly, Canada has recently reorganised its eResearch providers into a new organisation, the New Digital Research Infrastructure Organization (NDRIO), which will fund coordinated activities in advanced research computing, data management and research software.[286]

Applying either the European or Canadian model to Australia would require some modification to account for the direct role of the Australian Government in university funding compared to Canada, where university funding is the responsibility of the provinces. In addition, the publicly funded research agencies, particularly CSIRO, play a more significant role in Australia than Canada. Whether Australia adapts an existing model to suit the Australian context or develops one, coordinating resources and services will be key to advancing data-intensive research and critical to the next phase of investment.

5.3.2 RESEARCH AND TECHNOLOGY TRENDS STRESSING ERESEARCH INFRASTRUCTURE

Whatever the organisational structure of Australia's eResearch infrastructure, it needs to prepare for several significant trends that are already becoming apparent in data-intensive research globally. The following trends are already beginning to stress all aspects of eResearch infrastructure and will continue to do so.

Data variety and volumes

Continual increases in the variety of data and data volumes increasing faster than compute or network capacities, thereby maintaining a tension between the cost and ease of transmission of data to compute facilities and the cost of storing the data at a facility or processing via cloud computing. In astronomy, for example, it is anticipated that the Square Kilometre Array (radio telescope) will require a dedicated supercomputer to simply run the data analytics involved in the initial screening of the data streams. This trend is not restricted to the sciences. For example, ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) in the Netherlands links large social databases to a supercomputer increasingly used by social scientists across a range of disciplines and projects.[287]

Data repositories

The convergence of HPC, cloud computing and data analytics, the integration of simulation and data analytics and increasing demand on computational facilities by advanced AI tools requires access to significant data repositories.[25, 192, 288, 289] Even in business, this is calling into question the cost-effectiveness of cloud solutions and seeing a return to hybrid solutions involving the cloud and in-house HPC facilities.

Evolution of supercomputers

Trends in the design and evolution of supercomputers themselves from a homogenous 'compute box' to a heterogeneous mix of processors, with this trend expected to accelerate over the next generations of machines.[290] For example, the expected increase in compute speed of Pawsey's new supercomputer, Setonix, over NCI's Gadi is attributed to an increase in the ratio of graphical processing units (GPUs) to conventional computational processing units (CPUs) with the GPUs of a new generation.[291] However, accessing this brute compute power for applications is not trivial and involves significant re-writing of codes.[292]

5.4 What should Australia do, and who should do it?

To date, investment in Australia's eResearch infrastructure has not been highly coordinated; however, in some ways this has been a strength. A 2015 assessment of the NCRIS eResearch capability noted that the investments (until that date) had been "catalytic for much research activity in Australia. Innovative and in many ways world-leading, ... [in particular] its collaborative intent has been realised in implementation and practice to a degree unmatched [in some respects] in other developed countries".[293]

While this approach has arguably served Australia well to date, it is unlikely to do so in the future, particularly in a period of extremely restricted government funding. Our key finding (see Section 3 concerning research data policy) is that it is time for a more strategically coordinated approach to investment in Australia's eResearch infrastructure. The aims should be to:

- minimise the risk of unproductive duplication, fragmentation and competition for limited funds
- determine the disciplines requiring peak data environments and decide if Australia wants a footprint in this area and, if so, resource it
- balance support for the broad base of Australian research (as more disciplines become more significant users of data-intensive research methodologies) with continued investment in leading-edge technological developments to ensure Australian research remains globally competitive, particularly in research areas critical to Australia's future
- set priorities but allow appropriate diversity in disciplinary requirements and encourage innovation.

The 2016 National Research Infrastructure Roadmap included a recommendation to: "establish a National Research Infrastructure Advisory Group to provide independent advice to Government on future planning and investment for a whole of government response to national research infrastructure".[7] While such a comprehensive mandate may have value, eResearch infrastructure would benefit greatly from the establishment of such a body.

5.4.1. GHOSTS OF REPORTS PAST

The lack of effective coordination in Australia's eResearch infrastructure is not a new observation. Over the past decade, several reports, roadmaps and assessments of eResearch infrastructure have contained calls to establish mechanisms to ensure a coherent overview of the infrastructure and the associated investment.

The 2011 Strategic Roadmap for Australian Research Infrastructure articulated the need to identify "the priority research areas for national, collaborative research infrastructure over the next five to ten years (capability areas)."[294] Stemming from this roadmap, in 2013, the NCRIS Research Data Infrastructure Committee published The Australian Research Data Infrastructure Strategy – The Data Revolution: Seizing the Opportunity, which contained a key recommendation (number 6) to "Establish a national research data infrastructure advisory committee to review, coordinate and provide coherence to the implementation of research data infrastructure investments".[295]

There are also significant reports commissioned by the Australian Government that have never been released. These include Rhys Francis's 2016 report on the next generation eResearch framework; a 2018 report on the design of, and transition to, what became the ARDC; and a 2018 report on HPC governance. From conversations with the authors or others involved with these reports and the limited information in the public domain (such as a March 2018 presentation on the transition plan or discussion papers), it is clear these were well-researched reports with implications worthy of wider consideration.[296] In particular, we understand that the need for the strategic coordination of investments was a recurring theme. It is disappointing that these reports have not been released.

Unfortunately, the practice continues. For example, while the Australian Community Climate and Earth-System Simulator (ACCESS) has been funded, the underlying scoping study has not been released.[297] Similarly, the report of the recent scoping study for a National Environmental Prediction System, which will have significant implications for data-intensive research in the environmental sciences, has not been released.[298] Even if the government, for whatever reasons, decides not to pursue some or all specific recommendations, the release of the underpinning research and analysis would still be of considerable value and consistent with the principle that publicly funded research data should be available.

5.4.2. THE NEED FOR STRATEGIC PLANNING: A PRECURSOR TO THE NEXT PHASE OF INVESTMENT

To address the requirement for better coordination of investment in Australia's eResearch infrastructure, we recommend three complementary activities:

- Develop an **integrated strategic plan**, incorporating all eResearch infrastructure components, to underpin the next phase of investment from 2022 –27.
- Consider **establishing a new body** to set strategy and oversee the next phase of investment for all eResearch infrastructure—tentatively entitled the National eResearch Infrastructure Council—adapted from EOSC and Canada's NDRIO.
- Strengthen **engagement** between eResearch infrastructure planning (particularly the development of discipline-focused domains) and other forms of discipline planning through the existing discipline-based National Committees of the Australian Academy of Science and associated decadal plans.[299]

Significantly, these activities should strengthen the case for the substantial future investment that will be needed.

Given the ubiquitous nature of data-intensive research and the need to ensure that the leading edge of Australia's eResearch infrastructure remains globally competitive, the current refresh of the 2016 National Research Infrastructure Roadmap must take a holistic view of eResearch infrastructure and not silo it into its components.[300] The 2021 Roadmap Expert Working Group needs to be bold and not beholden to current arrangements. It should be prepared to assess the effectiveness and sustainability of various existing initiatives and programs or, at least, establish a process by which such assessments can be made before further investment.

One area that the Expert Working Group should explore is what role the private sector, both international (such as Amazon Web Services and Google) and particularly Australian providers of high-performance computing and data services, could play. In order to advance data-intensive research in Australia, this engagement should not just be as suppliers of data services or vendors of computing equipment. With the correct policy settings, more strategic engagement between Australian commercial HPC providers and Australian eResearch could advance both.

In the longer term, the Expert Working Group should consider recommending establishing a National eResearch Infrastructure Council adapted from international examples (such as the European EOSC and Canada's NDRIO) to oversee and coordinate the next phase of investment. To be effective, the Council's scope should comprise all research areas and all agencies involved in the deployment of eResearch infrastructure, importantly, including agencies from the health sector, given its growing role.

While the primary purpose of the Council would be to articulate and monitor a comprehensive strategy for the investment in the next phase of Australia's eResearch infrastructure development, including setting key priorities, the Council could have a more direct funding role as NDRIO does in Canada. With such a brief, its establishment would take some boldness and confidence from the government that a significant sum of

federal funding can be invested with less direct departmental involvement and a willingness by many existing agencies to give up some degree of their independence. However, the value to Australia's research is expected to be considerable.

Such a Council could also work with domain experts to guide and review the planning of next generation eResearch infrastructure. Ahead of development or deployment, these projects need to explicitly consider how to integrate data, compute, software and skills. The recent exercise to scope ACCESS is a case in point. As already noted, that report has not been released despite its considerable interest in an important area of data-intensive research. Commissioning by a body like the new Council would hopefully give sufficient distance from departmental sensitivities concerning 'cabinet-in-confidence' or budget negotiations to allow the release of such reports as a matter of course.

Finally, the Council should take leadership of a discussion of how Australia can gain the benefits that will emerge with next-generation exascale computing. While it will be challenging for Australia to invest in its own exascale facility, the cost of not doing so may be even higher, given the likely advances in areas of critical importance to Australia's future such as climate change, manufacturing, agriculture and astronomy. Strong arguments that are not inhibited by the self-interest of existing agencies or current structures will be needed.

What is exascale computing?

'Exascale' is the capability to perform a billion billion (a quintillion, 10^{18}) operations per second.[xvi] Exascale computers will be able to quickly analyse massive amounts of data, faster than existing petascale supercomputers (which perform a quadrillion, 10^{15}, operations per second). The increased memory, storage and compute power of exascale systems will enable modern scientific advancement in a wide range of fields. The United States, Europe, Japan and China are currently pursuing exascale computer programs.[xvii]

While the deployment of operational exascale computing and data facilities are a few years away, other nations are already launching ambitious projects in anticipation. An example of particular relevance to Australia is ChEESE, a Centre of Excellence in the domain of Solid Earth (SE), which is targeting the preparation of 10 flagship European codes for the upcoming European exascale supercomputer anticipated in 2022.[301] Similarly, the European Union's Horizon 2020 research and innovation programme is investing in EVOLVE.[302] This project, comprising 19 partners from 11 European countries, is designed to process massive data that require demanding computation across commercial markets, including automotive services, agricultural production and maritime surveillance. While Australia cannot match this level of European investment, it cannot be left behind in such developments.

In strengthening the relationship between the planning and provision of eResearch infrastructure and broader disciplinary planning, there is an important role for the Academy's National Committees and their development of decadal plans.[303] Most recent plans have touched, to various degrees, on the impact that data is having on developing the relevant discipline and the research opportunities that are opening up. For example, the Decadal Plan for Australian Astronomy 2016-2025 mid-term review includes the recommendation to "establish a long-term sustainable, distributed and interoperable set of HPC and data centre arrangements that span the requirements of... funding for commensurate training and education in data science and code development".[304]

Given the pace of innovation in data and data technologies, it would be helpful for the National Committees to review this aspect of decadal plans in a much shorter time frame. This disciplinary perspective could assist in a more strategic assessment of potential domains. In addition, once a domain is firmly established, it would be helpful to have cross membership between any scientific advisory committees and any relevant National Committees, such as the scientific advisory committee of the Australian BioCommons and the National Committees in the life sciences.

Recommendation 5.1 The 2021 refresh of the National Research Infrastructure Roadmap takes a holistic view of eResearch infrastructure and prepares an integrated investment plan covering all elements of eResearch infrastructure. This plan should take account of global developments in data-intensive research but must also set realistic priorities.

Recommendation 5.2 The 2021 refresh of the National Research Infrastructure Roadmap considers recommending that oversight and coordination of the next phase of investment in Australia's eResearch infrastructure is vested in a National eResearch Infrastructure Council—adapted from the European EOSC and Canada's NDRIO. The Council should:

a. involve all areas of the research sector and all agencies involved in the deployment of significant eResearch infrastructure as well as representatives from commercial players
b. have as its primary purpose to monitor and update where necessary the comprehensive strategy for investment in the next phase of Australia's eResearch infrastructure including setting key priorities
c. undertake the detailed planning needed for next generational models and applications, explicitly considering how to integrate data, compute software and skills
d. take leadership of the discussion of how Australia can gain the benefits that will emerge with next-generation exascale computing
e. potentially, have a more direct role in allocating funding for eResearch infrastructure.

Recommendation 5.3 The Australian Academy of Science ensures that all National Committees:

a. have strong engagement with relevant platforms or commons established within eResearch infrastructure
b. when developing decadal plans, explicitly address data issues, explore new opportunities afforded by data-intensive research and consider the need for retention of data
c. conduct more frequent reviews of the state of data and associated infrastructure in their discipline areas.

6. Skills and skilled people are vital, but culture matters

Research today requires digital skills to supplement discipline-specific skills, necessitating a lift in the basic data skills and literacy of all researchers.

This chapter discusses the challenges of ensuring the required skills for successful data-intensive research, both skills within the research community and skills of the research support community. Training and educational options, from foundational through to PhD training of researchers, need to be enhanced and offered widely to provide skilled and data-literate people in Australia to enable modern research.

Vital to this success is a research culture, that recognises, supports and credits data science skills.

6.1 Data skills have become ubiquitous

The digital revolution brought about by computation power and information technology is changing the nature of work. A 2019 report by the Royal Society, Dynamics of data science skills,[132] found that demand for workers with data skills had increased by 230% over the past five years, while demand for all workers had only increased by 36%. Similarly, the 2018 *The Future of Jobs Report* by the World Economic Forum predicted that the top two emerging jobs by 2022 would be data analysts or scientists and AI or ML specialists.[305] In Australia, a 2018 report by Deloitte Access Economics estimated that the data science workforce would grow by 2.4% annually between 2016–17 and 2021–22,[144] compared to a 1.5% annual growth anticipated for the total Australian labour workforce over the same period. Australian universities have responded to this external demand through a range of courses in various faculties.[306] While the COVID19 pandemic has affected these predictions, in many ways the response to the pandemic has reinforced the need for data skills in the workforce and the importance of data scientists and other data professionals.[145]

6.2 Data literacy for all researchers

Research, particularly scientific research, is no longer a solitary pursuit, with research now conducted by a team of people bringing a range of skills to a project. Data-intensive research requires additional technical skills, including data stewardship or custodian work, research software engineers, and ethical or legal skills. The specific skills required are continually evolving as the discipline continues to emerge.

In 2020, the OECD published a report, *Building digital workforce capacity and skills for data-intensive science*,[133] which provides a comprehensive analysis of the digital workforce's global requirements concerning data-intensive science. This section refers to some of the relevant points made in the OECD report and puts an Australian focus on them.

6.2.1. THE NEED FOR A DIGITALLY SKILLED RESEARCH WORKFORCE IN AUSTRALIA

There is a pressing need for digital skills development within the Australian research sector. Australia is third in the world after Japan and the USA for indicating "having access to the right set of skills" as the most important challenge to researchers.[307] Inadequate access to skilled people will constrain research productivity. An analysis in Australia of research IT support, commissioned by the ARDC in 2019, calculated the number of equivalent full-time (EFT) staff providing digital support to researchers.[308] It reported there was one EFT per 90 researchers "for management or stewardship of research", one EFT per 60-70 researchers "for collection and analysis of research data", one EFT per 250 researchers for "providing training [or] advice on research data", one EFT per 100 researchers "involved in various software engineering roles" and one EFT per 200 researchers "providing training and advice on software engineering". Irrespective of the exact support required, this is well below one European estimate that suggests the need for digitally skilled research support professionals, as a proportion of the research workforce, has been estimated as high as one per 20 researchers, in order for research data to meet the FAIR principles.[87]

A marked concern about lack of expertise and access to expertise in bioinformatics were the key conclusions of a 2013 survey of over 200 Australian life scientists by Bioinformatics Resource Australia-EMBL (BRAEMBL).[309] The respondents indicated **training and community building as the most sought after services**. The report itself is no longer available online. Still, the OEDC report (page 16) quotes the survey, "more than 60% of researchers surveyed said that their greatest need was additional training, compared to a meagre 5% who need access to additional compute power...".[133]

Demonstrating the applicability of data-intensive skills to the entire research sector, data and technical literacy were identified as the most important areas for skills development for the future humanities workforce in a 2019 consultation by the Australian Academy of the Humanities, Future Humanities Workforce.[310] The skills and capability gaps in training for humanities graduates and researchers identified in the consultation included quantitative and statistical research skills, competency with textual analysis, visualisation tools, data processing and machine learning. In addition, "more advanced use of technology can also augment data preservation, collection, management and storage practices".

6.2.2. A COMPETENCE FRAMEWORK FOR A DIGITALLY SKILLED RESEARCH WORKFORCE

The competencies for data-intensive research reach further than researchers. It is not just researchers who need digital skills; there is a range of research support staff required to work with data. Data-intensive research requires researchers with skills in computer science, statistics and data stewardship to acquire and clean or transform unstructured and continuously streaming data, regardless of its format, size or source.[311] In addition, strong analytic and data management skills are also needed to work with large, complex data sets throughout the research pipeline, from data curation and visualisation to modelling and interpretation.[312] Oral and written communication skills are necessary to engage with diverse audiences about real-world problems, work in teams, and participate in effective problem solving for technical and ethical dilemmas encountered in data science. This section focuses on the role of data scientists in a digitally skilled research workforce.

A 2016 industry report noted that data scientists spend a majority (60%) of their time cleaning and organising data (or data wrangling),[313] with the remainder of their time spent collecting data sets (19%), mining data for patterns (9%), refining algorithms (4%), building training sets (3%) and 'other' (5%). The diverse competencies required for data-intensive research further highlight that it is not individuals we need to train but rather to equip an entire research workforce with digital skills.

Considering the broader scope of skills required for data-intensive research, the categories of 'data science' and 'data stewardship' can be used as a high-level taxonomy. As discussed in Section 4 of this report, there is little consensus on the definition of a 'data scientist', but the OECD report distinguished between data science and data stewardship in terms of the manipulation of data versus the management and preservation of data.[133]

There have been multiple attempts to define the skills and competencies in data science. For example, in 2017, Australia's Data to Decisions CRC (D2DCRC) produced Australia's first Data Science Competency Framework (DSCF) which defines a series of competencies focused on describing the abilities, knowledge and experience required by data scientists.[141] This framework is the basis for the CRC's Data Science Development Planning Tool (DPT), an online self-assessment system that enables people to identify gaps in their competencies, ranking competencies from least mature to most: practitioner, senior, lead and director. The DSCF and DPT were made available as an online trial in 2018, and updated in 2019, however with the closure of the CRC in June 2019 this remains a static document.

Acknowledging the variable competencies in data science will assist in planning and developing data science capabilities. This is particularly important for policy development, training, developing professional communities and clear career paths.

The European Commission's 2013 Framework for Developing and Understanding Digital Competence in Europe describes a framework for digital work.[314] However, there is no mention of how data generated in research should be managed, or the analysis or curation of data for research purposes. The OECD report notes this framework is missing competencies required in a publicly funded research context in the areas of information and digital literacy, digital content creation, safety, and communication and collaboration.[133]

6.2.3. RESEARCH CULTURE IS VITAL TO DEVELOPING DATA SCIENCE

All stakeholders within the research sector have a responsibility to create an ecosystem capable and supportive of data-intensive research. The OECD report (page 25) argues for the need to define skills and roles to assist policy development and the "establishment and recognition of professional communities and associated career paths".[133]

"All stakeholders within the research sector have a responsibility to create an ecosystem capable and supportive of data-intensive research"

The ambiguity of being a data scientist—professional or academic staff?

Culture within institutions influences this discussion. Within universities and their funding agencies, there is ambiguity in the role and status of data science and data scientists. This ambiguity manifests as tension between appointments as 'technical staff' (or sometimes 'professional' or 'general') instead of 'academic staff'. Data science positions require both the focused technical drivers associated with professional roles and the creative aspect of academic roles.

The OECD report recommends that research organisations **develop and adopt frameworks to employ digital research support professionals in stable positions with opportunities for career advancement and the development of metrics that recognise all members of research teams' contributions**, amongst others.[133] Aligned with this recommendation, an informal virtual event was held in 2020 to set up a dialogue about the 'third space' in which data scientists work, as quoted by them: "a space between professional and academic spheres in which lateral interactions, involving teams and networks, occur in parallel with formal institutional structures and processes, giving rise to

new forms of management and leadership".[315] The event was co-organised by Melbourne Data Analytics Platform at the University of Melbourne, and Sydney Informatics Hub at the University of Sydney. There were over 80 attendees, including international interest.

The ambiguity regarding data scientists' roles and status is further reflected in the lack of structure around reward and recognition systems. The inability to categorise these 'hybrid' roles that sit between academia and the professional staff have meant that, to date, research organisations have struggled to recognise these roles across institutions. Furthermore, current research reward and recognition systems rely on 'journal impact factor'.[316] Journals dedicated to data science research are relatively new and hence have low impact factors. As such, the reward system encourages researchers working in the 'third space' to publish their research results in journals that do not necessarily support efforts to publish the technical aspects of data science techniques—hindering the open science endeavour.

The Declaration of Research Assessment (DORA) attempts to improve how research is assessed.[317] Only a few Australian institutions have signed DORA.[318] Future discussions about altering reward structures in academia need to consider hybrid technical and professional roles within universities notwithstanding our recommendation that data science should be recognised as scientific discipline. Given the demand for data scientists from the business sector without attractive conditions of employment for (and appropriate recognition of) data scientists in research organisations Australian data-intensive research is likely to be impeded.

6.3 Existing training options for knowledge and skills in data science

Developing a digitally skilled research workforce requires support for the entire spectrum of education: from foundational studies to extending existing researchers' skills.

6.3.1. FOUNDATIONAL AND SECONDARY EDUCATION

In a 2016 report, Australian students ranked last in technology skills and interest in technical jobs across a global sample of countries.[319] Since that report, substantial change has been initiated at an elementary and secondary level, with a new digital technologies curriculum.[320] This is an important first step that will feed into tertiary education with the aim of building a world-leading capability in data science and analytics in Australia.

One challenge in rolling out the new curriculum for elementary and secondary education is that much of the teaching workforce is unprepared to teach new data management capabilities, problem-solving using algorithms, information and communication technology (ICT), and ethical understanding of data and digital technologies.[321] Numerous support programs have been funded through the Australian Government's National Innovation and Science Agenda,[322] such as the Digital Technologies Challenges—free classroom-ready resources for years 3 to 8 produced by the Australian Computing Academy,[323] industry partnerships through CSIRO STEM Professional in Schools,[324] and online professional development for teachers through the University of Adelaide's CSER Digital Technologies Education,[325]—massively open online courses (MOOCs) in digital technologies.

However, inequities exist as not all schools have been able to embrace these opportunities. To ensure that young people have the foundational knowledge required for them to progress to data science in tertiary education and contribute to Australia's industry and research activities, all teachers need to have the time, resources and support to engage with the new curriculum in a meaningful way.

The Digital Technologies curriculum is an important first step that will feed into tertiary education with the aim of building a world-leading capability in data science and analytics. CREDIT: FLICKR "MACHINE LEARNING STUDIO" (CC BY-NC-ND 2.0) BY ARS ELECTRONICA

6.3.2. TERTIARY EDUCATION

Most Australian universities now offer undergraduate and postgraduate programs in data science (or computer science or IT with a data science focus). A desktop survey of offerings in 2020 shows a breadth of approaches to undergraduate data science education, ranging from bachelor degrees completely dedicated to data science, to data science offered broadly as a major in one of many bachelor degrees or as a double degree. With these majors or double degrees, the emphasis is most commonly on pairing a data science degree with a bachelor such as business, engineering or science. The term 'data science' is also, at times, used as a badge for presenting all computing options to prospective students.

Many Australian universities now offer a Master of Data Science and options for shorter courses of study leading to graduate certificates or diplomas. Most master degrees are two years, and a major distinction in offerings is in prerequisites. Some universities emphasise training or experience in mathematics and computer science as necessary prerequisites, whilst some soften this to include 'related disciplines' (usually mentioning physics, business, engineering or STEM broadly). In contrast, others only require a bachelor degree in any discipline.

The labelling of these degrees as data science, and the wide range of offerings, indicates that universities now view data science as a highly sought-after skill. Despite the increased availability of data science courses, inequities exist in how easily students can access data science training. Depending on how courses are structured, only students enrolled in a specific degree may access a data science course. This is particularly the case when a student wants data science skills without a full focus on data science.

6.4 Improving data science knowledge for researchers

LEARNING OPPORTUNITIES: WITHIN UNIVERSITIES

Beyond training a new generation of students in data science, a more complex issue is how existing researchers can be given appropriate data science support or training. Needs vary enormously, from those with no experience who want to begin on a path of upskilling, to those with research programs already using big data who need additional staff and expertise for new initiatives. PhD students and early- and mid-career researchers can be restricted in developing data science skills in research groups or at universities that do not value or enable upskilling, or make relevant courses available.

Some recent initiatives in universities have been attempting to bring elements together to support research data and data science. These include Melbourne's Petascale initiative,[326] the Sydney Informatics Hub,[327] the QUT Centre for Data Science[328] and the Australian Data Science Network.[329]

LEARNING OPPORTUNITIES: GOVERNMENT-FUNDED OR COORDINATED

There have been arguments overseas that the public sector's role is to provide 'framework conditions' that facilitate access to financing by reducing the cost of capital or access to skilled people by making sure there is an existing pool of expertise.[330] Similarly, the OECD report (page 9) recommends that national governments can take action by "recognising at the policy level the need for a digitally skilled workforce in research, and the importance of strategic planning that integrates the five key areas that are necessary to build and maintain this workforce: defining needs; provision of training; community building; career paths and rewards; and broader enablers".[133]

There is a serious undersupply of digital skills training in Australia. The OECD report (page 27) noted the findings from a 2019 ARDC Summit poll of 79 digital skills research training initiatives showed that demand exceeded supply for more than 75% of training initiatives, with 9% stating that demand was more than five times higher than supply.[133] However, in Australia, there is no coordinated national focus for upskilling existing researchers.

Some initiatives do exist, for example, the Data to Decisions CRC trained over 2000 scientists during its operation between 2014 and 2019, according to the legacy website.[141] Other examples include the provision of services by the ARDC, which has established a skilled workforce initiative.[331] This initiative is the primary response to the 2016 National Research Infrastructure Roadmap that specifically recommended the need to "recognise that a skilled workforce is critical to all national research infrastructure.[7] Ongoing commitment to training and career progression not only by the facilities and projects but also by the universities and research institutions that harness them is essential". Another example initiative is within the ARC Industrial Transformation Training Centre schemes.[332] While the current priorities for this scheme do not include data science, some centres recently funded are relevant—e.g. 2017 ARC Training Centre in Cognitive Computing for Medical Technologies,[333] 2018 ARC Training Centre for Transforming Maintenance through Data Science,[334] 2019 ARC Training Centre in Data Analytics for Resources and Environments,[335] 2020 ARC Training Centre for Information Resilience.[336]

LEARNING OPPORTUNITIES: A COMMUNITY OF PRACTICE AND VOLUNTEER INITIATIVES

In the absence of formal training options, volunteer services have emerged to fill the gaps. For example, a joint program between the Research Data Alliance and CODATA had delivered nine two-week schools on four continents by February 2020. The program used a curriculum for foundational Research Data Science skills for Early-Career Researchers and is entirely volunteer run.[337] Similarly, the annual FORCE11 Scholarly Communication Institute which runs multiple courses focused on data science skills attracts a world-wide audience and relies on volunteer instructors.[338]

In 2015, Cambridge University created a program of Data Champions, recruiting volunteers from the early-career researcher and principal investigator community to work locally among their peers to increase data skills and awareness of issues related to FAIR data.[339]

Grassroots movements make data science upskilling resources accessible and available but formal arrangements are needed. The predominantly volunteer-based workforce raises concerns regarding scalability and long-term sustainability.

A breadth of self-learning resources to upskill are also available within research groups in domains using, but not specialising in, data science. For instance, students and researchers in the Quantitative and Applied Ecology Group (QAEco) at the University of Melbourne:

- practise self-education: textbooks, free online coding books, manuals and courses
- help each other: the group runs a fortnightly coding and maths club, has Slack channels dedicated to help in coding, and regularly contributes to running workshops for others outside the group
- networks: data science has a large Twitter presence; local meet-up groups like RLadies[340]
- attend workshops (e.g. Software Carpentry, courses offered[341] by the university, workshops linked to conferences (ecologists usually offer one to several data skills workshops linked to their conferences, both nationally and internationally; data science conferences like useR! also offer accessible courses))[342]
- use university-supplied expertise through collaborations with data science specialists in other schools, or the Melbourne Data Analytics Platform[343]

While some of these options are available to all researchers Australia-wide, this degree of interest and available options are likely rare, so there is considerable inequity in the ability to learn on the job. Similarly, for a volunteer-based training workforce, it has been noted that these types of 'self-help' approaches to meeting data skill training needs are not scalable.

6.5 What is global 'best practice' telling us?

'Skills and training' are one of the capability factors identified in the Community Capability Model Framework—a tool developed by UKOLN, the University of Bath and Microsoft Research to assist research organisations, funders and researchers in growing the research community's capability in data-intensive research.[344] The framework notes that "... training is most effective when it is fully embedded as part of the early-education and continuing professional development of researchers".

"While Australia does not have a broad digital skills strategy or targeted funding for data skills, there has been activity in this area overseas."

While Australia does not have a broad digital skills strategy or targeted funding for data skills, there has been activity in this area overseas. Europe has taken a broad approach to digital skills, creating a pan-European Digital Skills and Jobs Coalition, under which 23 National Coalitions exist.[345] The coalitions "work on the ground and connect different actors—companies, government, training providers and NGOs—to improve citizens' digital skills and prepare them for the ongoing digital transformation of our economy and society".

Looking more specifically at data skills, an example of a practical approach to increase skills in data stewardship is the European Open Science Commission which has funded the development of a FAIR core competence centre,[346] which is described as "a shared

hub of expertise in implementing FAIR data stewardship principles, offering leadership, coordination and cataloguing services to connect relevant people, guidance, learning resources and curricula in different thematic areas".[347]

In addition, the European Commission has identified 'Open Science Skills' as part of its Digital Education Action Plan in response to the finding that three out of four researchers have no training in open access or open data management.[348] The plan runs between 2019 and 2021 and is funding training to undergraduates in open data; open access; and open and FAIR data management, analysis, use, reuse and publishing. Courses on open science for educators in higher education at all career levels are also being developed.

In another example, France's National Plan for Open Science from 2018 describes multiple actions the country should undertake to move to open science.[349] Actions include "develop open science skills, especially in postgraduate schools" and "create the conditions for, and promote, the adoption of an Open Data Policy for articles published by researchers".

There are two examples from the US where research organisations have prioritised institutional change to improve the management and use of data-intensive research:

1. The Moore Sloan Data Science Environments[350] seek to address the substantial systemic challenges in academia needed to be overcome to maximise the impact of data-driven research and create supportive environments for researchers using and developing data-intensive practices. The initiative supports cross-disciplinary academic data scientists at research institutions. There is a description of some of these projects at New York University; University of California, Berkeley; and the University of Washington.[351]
2. The University of California, Berkeley has appointed an Assistant Provost—Data with comprehensive oversight of all academic data and related issues. In 2018, the Division of Data Science and Information was established[178] and the 10-year Strategic Plan included investment in data science and its integration across the campus's signature initiatives.[352] The university's first Associate Provost for the Division of Data Science and Information and Dean of the School of Information took office in 2020.[353]

In addition to countries developing national initiatives, international collaborations to expand and strengthen the community of practice have also been established. These include collaborations between international data common initiatives such as the ARDC, European Open Science Cloud (EOSC),[284] African Open Science platform[354] and the National Institutes of Health Data and Research Center.[355]

Recommendation 6.1 Research organisations (particularly universities):

a. review staff development programs to ensure all researchers have access to courses aimed at enhancing basic data skills as well as discipline-focused courses to enhance more advanced skills
b. resolve the current ambiguity in the employment of data scientists noting recommendation 4.2 regarding the recognition of data science as a discipline
c. ensure all PhD students have access to courses to develop generic data skills before they commence their research projects, with these courses including advice on how students should collaborate or consult with a data scientist
d. champion a culture that supports data science skills and creates positive feedback for data-intensive approaches to research and collaboration.

Recommendation 6.2 The 2021 refresh of the National Research Infrastructure Roadmap recognises that all research infrastructure facilities now produce digital data and need resources to employ data professionals to manage and maintain this data and, where appropriate, to pre-process it to facilitate research access.

Recommendation 6.3 The Academy's National Committees for Science be given a specific mandate to:

a. assess the data maturity of their relevant disciplines and work closely with the ARDC to develop general and discipline-specific approaches to improve data maturity

b. challenge disciplinary cultures that inhibit the development of data literacy and data-intensive research

c. promote exchange of good practices across the research community through discussions on digital skills, provision of training and reward structures.

Recommendation 6.4 When disseminating research outputs, researchers acknowledge and give credit where it is due: to data generators, curators and stewards.

7. Data-intensive research spawns challenges for research integrity

Developments in data science and data-intensive research are drawing attention to research integrity and the processes and procedures that support and govern it.

How data is collected, managed and analysed have implications for critical issues such as reproducibility, replicability, transparency, representativeness, research ethics and mitigation of the risk of falsification or fabrication of data.

Addressing these issues is critical to promoting responsible practices in data science and ensuring research credibility and integrity in data-intensive research.

7.1 Data is challenging research culture

There is increasing global recognition that the research community needs to strengthen the processes and practices that support research integrity. Issues such as transparency, the integrity of and representativeness in data sets, reproducibility and replicability, bad research practices, and, unfortunately, research fraud are not new or restricted to data-intensive research. However, data-intensive research and developments in data science are exacerbating concern with these issues. The risks also increase as tools, such as machine learning, become 'commoditised' and are used by scientists with limited awareness of their potential pitfalls. Many of these challenges go to the heart of the scientific method and, if not addressed, have the potential to not only discredit data-intensive research but erode trust in science. This was demonstrated in late 2019 when political activity in Australia attempted to "discredit the integrity of scientists".

It is beyond the scope of this report to discuss research integrity in detail, except to note that other jurisdictions are more advanced than Australia, both in policies and practices. For example, the UK has a strong track record on research integrity issues. The UK Research Integrity Office was established in 2006 as an independent charity supporting the public, researchers and organisations to further good practice in academic, scientific and medical research.[356] Universities UK released the Concordat to Support Research Integrity in July 2012.[357] The following two sections describe some of the aspects of research integrity that either arises from the use (or misuse) of data or from the new tools made available by data science.

7.2 Data issues

Issues with data, including its falsification, are not new to research. A 2009 meta-analysis of surveys about the fabrication and falsification of data showed that with self-reporting studies, nearly 2% of scientists had fabricated, falsified or modified data or results at least once. This figure increased to 14% in surveys about the behaviour of colleagues.[358] Other

questionable research practices were identified in 34% and 72%, respectively. The author noted: "considering that these surveys ask sensitive questions and have other limitations, it appears likely that this is a conservative estimate of the true prevalence of scientific misconduct".

Early in 2020, the Editor in Chief of *Molecular Brain* noted that he had requested the raw data in 41 of the 180 manuscripts managed since 2017.[359] He concluded that "more than 97% of the 41 manuscripts did not present the raw data supporting their results when required by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases".

With the advent of COVID-19, the political situation concerning scientific research has intensified, as has attention to researchers' work. There have been concerns that the focus on a fast response is causing players in the research system (both researchers and publishers) to respond to 'imperfect incentives'.[360] For example, during the global scientific response to COVID-19, the retraction of a study from *The Lancet*,[361] caused the World Health Organization to temporarily suspend clinical trials[362] on hydroxychloroquine. The retraction resulted from questioning the underlying data, which was sourced from an analytics company.[363] This retraction led to subsequent investigations into multiple papers published using data from the same company.[364] This incident serves as an acute reminder of the need for data experts to be embedded through the research process: from research planning through to publishing. A review by the *New England Journal of Medicine* on a related retracted paper noted the limited experience the journal has with studies using an extensive database, noting "in the future, our review process of big data research will include reviewers with such specific expertise". This is, in fact, a broad challenge that many publishers face and is not a new problem—academic journals have never had the necessary infrastructure to publish the underpinning data-intensive research.

Universal adoption of the policies and process recommended in Section 3 would go a long way to making it more difficult to hide poor data practices, including falsification. Technical solutions—driven in part by data science—also could play a valuable role. Statistical tools can be used to detect data fabrication. Diverse methods have been developed, including a 2016 systematic approach using statistical techniques to assess randomisation outcomes to evaluate data integrity.[365] Other statistical methods to identify data fraud have been demonstrated in *Structure Validation in Chemical Crystallography*,[366] proposed in 2009, and another in 2013 to assess numerical data.[367] More recently, a paper in 2019 offers a standard for data fraud analysis and interpretation.[368]

Australia has not been free of claims of serious research misconduct involving, to some extent, data falsification or fabrication. Poor research practice has been identified at Australian institutions, including the following examples.

Allegations of scientific misconduct and fraud by Bruce Hall, Professor of Medicine at the University of New South Wales, were broadcast on the ABC by Dr Norman Swan in 2002.[369] Despite an investigation determining that Professor Hall was guilty of scientific misconduct, the UNSW Vice-Chancellor determined that Professor Hall only committed errors of misjudgement sufficiently serious in two instances to warrant censure.[370] This finding caused calls for the need to have external and independent inquiries in these cases.[371] The 2007 Australian Code for the Responsible Conduct of Research, a joint endeavour between the National Health and Medical Research Council (NHMRC), the Australian Research Council (ARC) and Universities Australia, was partly in response to this case.[372]

In 2016, two[373] University of Queensland researchers were convicted of fraud, with one charged[374] with using public and private research money for a 2011 publication on a new treatment for Parkinson's disease for which the clinical trial was not conducted.[375]

More recently, a UNSW researcher has had multiple papers retracted or withdrawn due to unresolved concerns over missing or manipulated data.[376] The six-year investigation by the university was the subject of some criticism, which calls for the case to be considered by the NSW Independent Commission Against Corruption (ICAC).[377]

Another internal investigation at the Swinburne University of Technology on the retraction of dozens of papers of one of their researchers over duplication of data resulted in the researcher losing their position and attracted calls for an external investigation.[378]

7.3 Reproducibility and Replicability

The use, management, curation and transparency of data lie at the heart of the cases above. Data is also central to the issues of reproducibility and replicability of research outputs—concepts that go to the heart of the scientific method.[379] Again, this issue is not restricted to data-intensive research, with the new tools developed by data science, particularly machine learning and deep learning, exacerbating the problem. The underlying training sets must be adequately described along with associated algorithms in such cases. This is far from common practice but hopefully beginning to change.

"Data is also central to the issues of reproducibility and replicability of research outputs—concepts that go to the heart of the scientific method."

The issue of reproducibility and the subsequent need to adopt more transparent research practices is being taken seriously in several countries. For example, in the USA, the National Academies of Science, Engineering and Medicine have established a committee to explore reproducibility and replication in scientific and engineering research.[380]

Many data-intensive disciplines are beginning to seriously address the issue of reproducibility in the following ways.

- The Association of Computational Machinery (ACM) has instituted formal processes for artefact review—defining an artefact to mean a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself, which includes software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyse results.[381]
- The top experimental machine learning conference, KDD, assesses submitted papers on reproducibility in addition to other research excellence metrics.[382] Authors are strongly encouraged to make their code and data publicly available whenever possible. The ECML and PKDD conference is the top European machine learning conference and provides 'reproducible research' badges to papers.[383]
- Since 2017, the ACM SIGMOD Reproducibility Awards have been given to the most reproducible papers presented at the SIGMOD conference the previous year.[384] The criteria include all results being verified, ease of reproducibility, flexibility and portability across platforms.
- Since 2018, pVLDF has introduced pVLDB Reproducibility awards with three goals: increasing the impact of database research papers, enabling easy dissemination of research results and enabling easy sharing of code and experimentation set-ups.[385]
- The machine learning community in the USA in 2014 established the annual event, Fairness, Accountability, and Transparency in Machine Learning,[386] in 2014, which provides the community with a venue to explore and address challenges regarding rigorous computational methods.
- The third Reproducibility Challenge for the Neural Information Processing Systems (NeurIPS) conference was held in 2019 to address the challenge in machine learning research to "ensure that published results are reliable and reproducible".[387]

7.4 Data-intensive research raises ethical issues

Related to the broad issue of research integrity is the issue of ethics. Section 3 discussed the need to adopt CARE as a guiding principle along with FAIR. From a practical and technical perspective, data-intensive research raises ethical issues that may not be adequately or appropriately dealt with by the current institutional ethics committees that oversee research involving humans.

For example, are current research consent practices and protocols sufficient to regulate research that accesses health data? A critical issue in the ethical approval of such research should be to efficiently and effectively balance the benefit of research access to health data and the potential risk of privacy violation. While the Five Safes framework[388] provides guidelines on managing access to sensitive data, it does not explicitly address the validity of the underlying data and the proposed research methodology. Should ethics committees have that responsibility? One suspects the *Lancet* hydroxychloroquine case might never have started if an ethics committee had asked that question (and answered it). However, as currently constituted, few ethics committees would have the necessary skills to make such a determination.

More broadly, should all research projects that involve a potential impact on humans, such as research on brain-computer interfaces or facial-recognition algorithms, require ethics clearances? A 2019 report from the Australian Academy of Health and Medical Sciences argued that artificial intelligence is starting to transform healthcare and there is a need to "act now to set our path through this new landscape".[389] An academic summit held in September noted the seriousness of the ethical issues associated with artificial intelligence, including bias, given the increased automation of previously decided activities by humans.[390]

Recommendation 7.1 The Australian Government should strengthen the governance of research integrity and develop a national policy statement on ensuring research integrity for Australia. Such a statement should specifically address the issues raised by data-intensive research.

Appendix: Background on key personnel

Authors

Emeritus Professor Michael Barber AO FAA FTSE Emeritus Professor Flinders University and UWA; Council Member & Treasurer, Australian Academy of Science (2017-21); Chair, National Computational Infrastructure (2015-20)

Dr Danny Kingsley Scholarly Communication Consultant and Visiting Fellow at the Australian National Centre for the Public Awareness of Science (Australian National University)

Professor Jane Elith FAA Member, US National Academy of Sciences; Professor of Quantitative Ecology, University of Melbourne

Dr Ayesha Tulloch ARC Discovery Early Career Research fellow, School of Life and Environmental Sciences, University of Sydney

Expert Working Group

Em Prof Michael Barber AO FAA FTSE (co-chair)

Prof Jane Elith FAA (co-chair), The University of Melbourne

Dr Sue Barrell FTSE former Chief Scientist, Bureau of Meteorology, Board Member, Australian Research Data Commons

Professor Jane Hunter Adjunct Professor, School of ITEE, the University of Queensland

Professor John Mattick AO FAA FTSE FAHMS HonFRCPA SHARP Professor of RNA Biology, University of New South Wales, CEO Genomics England (2018-19); Director, Garvan Institute (2012-18)

Professor Kerrie Mengersen FAA Professor of Statistics, Queensland University of Technology; Deputy Director, ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS); Convenor, Australian Data Science Network

Professor Toby Walsh FAA Scientia Professor of Artificial Intelligence, University of New South Wales and Data61

Professor Bob Williamson FAA formerly Professor of Computer Science, Australian National University; Now Professor of Foundational Machine Learning Eberhard Karls University Tübingen;

National Committee for Data in Science

Dr Lesley Wyborn (Chair), Honorary Professor, National Computational Infrastructure Facility and Research School of Earth Sciences, Australian National University

Professor Ginny Barbour Director Open Access Australasia, and Co-Lead Office for Scholarly Communication, Queensland University of Technology

Dr Adrian Burton Director, Data, Policy and Services, Australian Research Data Commons

Dr Simon Cox Research Scientist, Commonwealth Scientific and Industrial Organisation (CSIRO)

Professor Darren Croton Centre for Astrophysics and Supercomputing, Swinburne University of Technology

Professor Louisa Jorm Director, Centre for Big Data in Health, UNSW

Dr Danny Kingsley Visiting Fellow at the Australian National Centre for the Public Awareness of Science

Dr Steven McEachern Director, Australian Data Archive, ANU Centre for Social Research and Methods

Professor Andy Pitman AO FAA Director, ARC Centre of Excellence for Climate Extremes, UNSW

Dr Francois Petitjean Senior Research Fellow in Machine Learning, Monash University

Professor Shazia Shadiq (ex-officio, Chair National Committee for Information and Communication Sciences), Director, ARC Training Centre for Information Resilience, The University of Queensland

References

1 Australian Academy of Science & Australian Academy of Health and Medical Sciences. *Improving accessibility and linkage of data to achieve better health outcomes for all Australians*. (2019).

2 Academy of the Social Sciences in Australia. ARC funding for The Use of Big Data for Social Policy. https://socialsciences.org.au/wp-content/uploads/2017/11/ASSA-Press-Release-10Nov17-ARC-Funding-for-The-Use-of-Big-Data-for-Social-Policy.pdf (2017) [Accessed 12 July 2021].

3 Leonelli, S. Data - from objects to assets. *Nature* **574**, 317–320 (2019).

4 Cousijn, H. et al. A data citation roadmap for scientific publishers. *Sci. Data* **5**, 1–11 (2018).

5 British Academy and The Royal Society. *Data management and use: governance in the 21st century*. https://royalsociety.org/-/media/policy/Publications/2017/Data_management_and_use_governance_in_the_21st_century_2017_seminar_report.pdf (2017).

6 Mayer-Schönberger, V. & Cukier, K. *Big data: a revolution that will transform how we live, work and think*. (John Murray, 2013).

7 Australian Government Department of Education Skills and Employment. *2016 National Research Infrastructure Roadmap*. (2016).

8 Monino, J.-L. & Sedkaoui, S. The Big Data Revolution. in *Big Data, Open Data and Data Development* 1–21 (John Wiley & Sons, Inc., 2016). doi:10.1002/9781119285199.ch1.

9 National Science Foundation. Harnessing the data revolution: transdisciplinary resesearch in principles of data science phase I. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505347 [Accessed 12 July 2021].

10 Hey, T., Tansley, S. & Tolle, K. *The Fourth Paradigm: data-intensive scientific discovery*. *Microsoft Research* https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/ (2009).

11 Wetterstrand KA. DNA Sequencing Costs: Data. *Natl. Hum. Genome Res. Inst*. 28–31 (2016).

12 Event Horizon Telescope. Event Horizon Telescope. https://eventhorizontelescope.org/ (2021) [Accessed 10 June 2021].

13 Square Kilometere Array. The SKA Project. https://www.skatelescope.org/the-ska-project/ (2021) [Accessed 17 February 2021].

14 Scaife, A. M. M. Big telescope, big data: towards exascale with the Square Kilometre Array. *Philos. Trans. A. Math. Phys. Eng. Sci.* **378**, (2020).

15 ARGO. Welcome on ARGO.NET, the International Argo Program Homepage. http://www.argo.net/ (2021) [Accessed 10 June 2021].

16 Waterston, J. Ocean of Things. https://www.darpa.mil/program/ocean-of-things (2021) [Accessed 10 June 2021].

17 Elliott, K. C. et al. Drone use for environmental research. *IEEE Geosci. Remote Sens. Mag.* **7**, 106–111 (2019).

18 Australian Research Data Commons (ARDC). Australia's Scalable Drone Cloud – ARDC. https://ardc.edu.au/project/australias-scalable-drone-cloud/ (2021) [Accessed 10 June 2021].

19 Gertner, J. Genome Sequencing and Covid-19: How Scientists Are Tracking the Virus - The New York Times. *New York Times* (2021).

20 Dimitratos, S. M., German, J. B. & Schaefer, S. E. Wearable technology to quantify the nutritional intake of adults: Validation study. *JMIR mHealth uHealth* **8**, (2020).

21 Krotov, V. Legality and Ethics of Web Scraping. in *Twenty-fourth Americas Conference on Information Systems* (2018).

22 Litsa, T. Searching for Video, Images, Audio, Gifs, Podcasts, Memes & Radio: a directory of search engines, finders & generators - Search Engine Watch. *Search Engine Watch* https://www.searchenginewatch.com/2016/05/20/searching-for-video-images-audio-gifs-podcasts-memes-radio-a-directory-of-search-engines-finders-generators/ (2016) [Accessed 10 June 2021].

23 Australian Acoustic Observatory. A2O – Australian Acoustic Observatory (A2O) is a continental-scale acoustic sensor network, recording for a five-year period across multiple Australian ecosystems. *Australian Acoustic Observatory* https://acousticobservatory.org/ (2021) [Accessed 10 June 2021].

24 Silver, N. FiveThirtyEight | Nate Silver's FiveThirtyEight uses statistical analysis — hard numbers — to tell compelling stories about politics, sports, science, economics and culture. https://fivethirtyeight.com/ (2021) [Accessed 10 June 2021].

25 Hao, K. The computing power needed to train AI is now rising seven times faster than ever before. *MIT Technology Review* (2019).

26 Akiyama, K. et al. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. *Astrophys. J.* **875**, L1 (2019).

27 Drake, N. First-ever picture of a black hole unveiled. *National Geographic* (2019).

28 Lutz, O. How Scientists Captured the First Image of a Black Hole - Teachable Moments | NASA/JPL Edu. *Jet Propulsion Laboratory* https://www.jpl.nasa.gov/edu/news/2019/4/19/how-scientists-captured-the-first-image-of-a-black-hole/ (2019) [Accessed 10 June 2021].

29 Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* **395**, 689–697 (2020).

30 Niller, E. An AI Epidemiologist Sent the First Warnings of the Coronavirus | WIRED. *Wired* https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/ (2020) [Accessed 10 June 2021].

31 Bogoch, I. I. et al. Anticipating the international spread of Zika virus from Brazil. *The Lancet* vol. 387 335–336 (2016).

32 Jenner, L. Kangaroo Island Shows Burn Scars On One Third of the Land Mass. *NASA* http://www.nasa.gov/feature/goddard/2020/kangaroo-island-shows-burn-scars-on-one-third-of-the-land-mass (2020) [Accessed 10 June 2021].

33 Deacon, B. & Carbonell, R. Inside the race to protect 250 threatened species hit by bushfire - ABC News. *ABC News* https://www.abc.net.au/news/2020-01-19/inside-the-race-to-protect-threatened-species/11877990 (2020) [Accessed 10 June 2021].

34 Foley, B. et al. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages* 200–204 (2018). doi:10.21437/sltu.2018-42.

35 Tulloch, A. I. T. et al. A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nat. Ecol. Evol.* **2**, 1209–1217 (2018).

36 Legge, S. et al. *Monitoring Threatened Species and Ecological Communities.* (CSIRO Publishing, 2018).

37 Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological Data Should Not Be So Hard to Find and Reuse. *Trends Ecol. Evol.* **34**, 494–496 (2019).

38 Suber, P. What Is Open Access? *Open Access* https://en.unesco.org/open-access/what-open-access (2019) doi:10.7551/mitpress/9286.003.0003.

39 Gold, E. R. et al. An open toolkit for tracking open science partnership implementation and impact. *Gates Open Res.* **3**, (2019).

40 Parliament of Australia. *List of recommendations. Australian Government Funding Arrangements for Non-NHMRC Research* https://www.aph.gov.au/Parliamentary_Business/Committees/House/Employment_Education_and_Training/FundingResearch/Report/section?id=committees%2Freportrep%2F024212%2F26656 (2018) [Accessed 12 July 2021].

41 Australian Government Productivity Commission. Data availability and use. https://www.pc.gov.au/inquiries/completed/data-access#report (2017) [Accessed 12 July 2021].

42 Open Government Partnership. Australia. https://www.opengovpartnership.org/members/australia/ [Accessed 12 July 2021].

43 Australian Government. Australian Government public data policy statement. (2015).

44 Australian Government Department of the Prime Minister and Cabinet. Open Government Partnership Australia. https://ogpau.pmc.gov.au/ [Accessed 12 July 2021].

45 Australian Government Department of the Prime Minister and Cabinet. Open Government Partnership Australia - Commitment. https://ogpau.pmc.gov.au/national-action-plans/australias-second-open-government-national-action-plan-2018-20/improve-sharing (2020) [Accessed 12 July 2021].

46 GO FAIR. FAIR Principles. https://www.go-fair.org/fair-principles/ [Accessed 12 July 2021].

47 Global Indigenous Data Alliance. CARE Principles for Indigenous Data Governance. https://www.gida-global.org/care [Accessed 12 July 2021].

48 FORCE11. The FAIR data principles. https://www.force11.org/group/fairgroup/fairprinciples [Accessed 12 July 2021].

49 CODATA. *The Beijing Declaration on Research Data*. https://zenodo.org/record/3552330#.XnRexJMzbOQ (2019) doi:10.5281/zenodo.3552330.

50 International Science Council Committee on Data (CODATA). https://codata.org/ [Accessed 12 July 2021].

51 Australian Academy of Science. National Committee for Data in Science. https://www.science.org.au/supporting-science/national-committees-science/national-committee-data-science.

52 OECD. Recommendation of the OECD Council concerning access to research data from public funding. http://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.htm [Accessed 12 July 2021].

53 FAIR Steering Group. Policy statement on F.A.I.R. access to Australia's research outputs. https://www.fair-access.net.au/fair-statement (2016) [Accessed 11 June 2020].

54 Supporting organisations. https://www.fair-access.net.au/background-information.

55 National Health and Medical Research Council. Open access policy. https://www.nhmrc.gov.au/about-us/resources/open-access-policy [Accessed 12 July 2021].

56 ARDC. FAIR data guidelines for project data outputs. https://ardc.edu.au/about_us/policies-and-guidelines/fair-data-guidelines-for-project-data-outputs/ [Accessed 12 July 2021].

57 Wilkinson, M. D. et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Sci. Data* **6**, 174 (2019).

58 Australian Academy of Science. Submission - Consultation on data sharing and release legislative reform. https://www.science.org.au/supporting-science/science-policy-and-analysis/submissions-government/aas-atse-submission-data-sharing-and-release (2019).

59 Australian Academy of Science. Academies call on health ministers to resolve health data issues. https://www.science.org.au/news-and-events/news-and-media-releases/academies-call-health-ministers-resolve-health-data-issues (2019).

60 NASA. Earth Science Data Processing Levels. http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/ (2016) [Accessed 12 July 2021].

61 Archives Act 1983 (Cth). https://www.legislation.gov.au/Details/C2019C00005.

62 National Archives of Australia. https://www.naa.gov.au/about-us/our-organisation [Accessed 12 July 2021].

63 National Archives of Australia. Digital Continuity 2020 Policy. https://www.naa.gov.au/information-management/information-management-policies/digital-continuity-2020-policy [Accessed 12 July 2021].

64 National Archives of Australia. Building trust in the public record: managing information and data for government and community. https://www.naa.gov.au/information-management/information-management-policies/building-trust-public-record-policy/building-trust-public-record-managing-information-and-data-government-and-community#intro [Accessed 12 July 2021].

65 Office of the National Data Commissioner. Our purpose. https://www.datacommissioner.gov.au/about/about-us [Accessed 12 July 2021].

66 Department of the Prime Minister and Cabinet Commonwealth of Australia. *Sharing Data Safely*. (2019).

67 Australian Government Department of Health. Genomics Health Futures Mission. *Department of Health* https://www.health.gov.au/initiatives-and-programs/genomics-health-futures-mission (2019) [Accessed 12 July 2021].

68 COAG Health Council. Health Chief Executives Forum. https://www.coaghealthcouncil.gov.au/Health-Chief-Executives-Forum/Introduction [Accessed 12 July 2021].

69 Australian Institute of Health and Welfare. National Health Information Strategy independent expert panel. https://www.aihw.gov.au/our-services/committees/strategic-committee-for-national-health-informatio [Accessed 12 July 2021].

70 ARDC. Health studies national data asset program. https://ardc.edu.au/collaborations/strategic-activities/national-data-assets/health-studies-national-data-asset-program/ [Accessed 12 July 2021].

71 The Royal Society. *Protecting privacy in practice: the current use, development and limits of privacy enhancing technologies in data analysis*. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf (2019).

72 ACOLA. *The effective and ethical development of artifcial intelligence: an opportunity to improve our wellbeing*. https://acola.org/wp-content/uploads/2019/07/hs4_artificial-intelligence-report.pdf (2019).

73 Office of the National Data Commissioner. Data Sharing and Release Legislative Reforms Discussion paper. https://www.datacommissioner.gov.au/resources/discussion-paper [Accessed 12 July 2021].

74 Australian Government Department of the Prime Minister and Cabinet. National Data Commissioner. https://pmc.gov.au/public-data/national-data-commissioner [Accessed 12 July 2021].

75 Australian Government Office of the National Data Commissioner. National Data Advisory Council. https://www.datacommissioner.gov.au/about/advisory-council [Accessed 12 July 2021].

76 Department of the Prime Minster and Cabinet. Data sharing and release reforms. https://www.pmc.gov.au/public-data/data-sharing-and-release-reforms [Accessed 12 July 2021].

77 NSW Government Data.NSW. NSW Data Analytics Centre. https://data.nsw.gov.au/nsw-data-analytics-centre [Accessed 12 July 2021].

78 Australian Bureau of Statistics. Multi-Agency Data Integration Project (MADIP). https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Multi-Agency Data Integration Project (MADIP) [Accessed 12 July 2021].

79 Australian Institute of Health and Welfare. Accessing government health & welfare data. https://www.aihw.gov.au/about-our-data/accessing-australian-government-data [Accessed 12 July 2021].

80 EOSC Pilot. EOSC data interoperability ensure availability of scientific data. https://eoscpilot.eu/news/eosc-data-interoperability-ensure-availability-scientific-data (2019) [Accessed 12 July 2021].

81 Intergovernmental Committee on Surveying and Mapping. Metadata working group. https://www.icsm.gov.au/what-we-do/metadata-working-group [Accessed 12 July 2021].

82 ANZUC Committee on Surveying & Mapping. *ICSM ISO19115-1 Metadata good practice guide*. https://www.icsm.gov.au/sites/default/files/5a-Good Practice document.pdf (2019).

83 National Health and Medical Research Council. Australian Code for the Responsible Conduct of Research. https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018 (2018) [Accessed 12 July 2021].

84 Australian Research Council. Research data management. https://www.arc.gov.au/policies-strategies/strategy/research-data-management [Accessed 12 July 2021].

85 Australian Government Department of Education Skills and Employment. National Collaborative Research Infrastructure Strategy (NCRIS). https://www.education.gov.au/national-collaborative-research-infrastructure-strategy-ncris [Accessed 12 July 2021].

86 Larivière, V. & Sugimoto, C. R. Do authors comply when funders enforce open access to research? *Nat. Comment* **562**, (2018).

87 Mons, B. Invest 5% of research funds in ensuring data are reusable. *Nature* **578**, (2020).

88 Australian Government National Health and Medical Research Council Australian Research Council and Universities Australia. *Management of data and information research: a guide supporting the Australian Code for the Responsible Conduct of Research*. https://www.nhmrc.gov.au/sites/default/files/documents/attachments/Management-of-Data-and-Information-in-Research.pdf (2019).

89 Open Access Australasia. Open access in Australia. https://aoasg.org.au/open-access-policies/ [Accessed 12 July 2021].

90 Australian Government National Health and Medical Research Council Australian Research Council and Universities Australia. Publication and dissemination of research: a guide supporting the Australian Code for the Responsible Conduct of Research. (2020).

91 The University of Queensland. Research data manager (RDM). https://research.uq.edu.au/rmbt/uqrdm [Accessed 12 July 2021].

92 European Commission. *Turning FAIR into reality: final report and action plan from the European Commission Expert Group on FAIR Data*. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf (2018).

93 ReDBox. About ReDBox. https://www.redboxresearchdata.com.au/about/ [Accessed 12 July 2021].

94 Australian Government Department of Education Skills and Employment. Final report - Review of the higher education provider category standards. https://www.dese.gov.au/higher-education-reviews-and-consultations/resources/final-report-review-higher-education-provider-category-standards (2019).

95 Federation of Finnish Learned Societies. Declaration for open science and research 2020-2025. https://avointiede.fi/en/policies/declaration-open-science-and-research-2020-2025 (2019) [Accessed 12 July 2021].

96 National Open Research Forum. National framework on the transition to an open research environment. *Digital Repository of Ireland* http://norf-ireland.net/wp-content/uploads/2019/07/NORF_Framework_10_July_2019-2.pdf (2019) [Accessed 23 December 2019].

97 *Concordat on open research data*. https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (2016).

98 European Commission. *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (2017).

99 Swedish Research Council. *Proposal for national guidelines for open access to scientific information*. https://www.vr.se/english/analysis/reports/our-reports/2015-03-02-proposal-for-national-guidelines-for-open-access-to-scientific-information.html (2015).

100 UK Government Cabinet Office. *G8 Open data charter.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf (2013).

101 European Commission. The EU's open science policy. https://ec.europa.eu/research/openscience/index.cfm (2018) [Accessed 12 July 2021].

102 UK Data Service. UK Data Service. https://ukdataservice.ac.uk/ (2019) [Accessed 13 July 2021].

103 UK Research & Innovation. Economic and Social Research Council. https://esrc.ukri.org/ [Accessed 25 March 2021].

104 ARDC. Australian Research Data Commons. https://ardc.edu.au/ [Accessed 16 March 2021].

105 Digital Curation Centre. DCC | Because good research needs good data. https://www.dcc.ac.uk/ (2021) [Accessed 9 June 2021].

106 Beagrie, N. The continuing access and digital preservation strategy for the UK Joint Information Systems Committee (JISC). *D-Lib Mag.* **10**, (2004).

107 Data Curation Network. We are the Data Curation Network. https://datacurationnetwork.org/ [Accessed 25 March 2021].

108 UK Research & Innovation. Your responsibilities if you get funding. https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/ [Accessed 25 March 2021].

109 UK Research & Innovation. Engineering and Physical Sciences Research Council. https://epsrc.ukri.org/ [Accessed 25 March 2021].

110 Ryan, B. Research Data Management: The EPSRC Policy Framework. https://prezi.com/kflylbtkcgvu/rdm-principles-and-expectations/ (2015) [Accessed 25 March 2021].

111 National Science Foundation. National Science Foundation. https://www.nsf.gov/ [Accessed 25 March 2021].

112 National Science Foundation. Dissemination and sharing of research results - NSF data management plan requirements. https://www.nsf.gov/bfa/dias/policy/dmp.jsp [Accessed 25 March 2021].

113 International Science Council. *Action Plan 2019 - 2021*. https://council.science/actionplan/ (2019) doi:10.24948/2019.09.

114 CODATA. Decadal programme: making data work for cross-domain grand challenges. (2019).

115 Australian Government Department of Industry Science Energy and Resources. *Australian Innovation System Monitor, 3.4.1 Share of world scientific publications*. https://publications.industry.gov.au/publications/australianinnovationsystemmonitor/science-and-research/research-output/index.html (2019).

116 Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). About COPDESS. http://www.copdess.org/home/about-copdess/ [Accessed 25 March 2021].

117 Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). Commitment statement in the earth, space, and environmental sciences. https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/ [Accessed 25 March 2021].

118 Lin, D. et al. The TRUST Principles for digital repositories. *Sci. Data* **7**, 144 (2020).

119 CoreTrustSeal. https://www.coretrustseal.org/ [Accessed 25 March 2021].

120 CoreTrustSeal. CoreTrustSeal trustworthy data repositories requirements. https://www.coretrustseal.org/why-certification/requirements/ [Accessed 25 March 2021].

121 World Data System. Community. https://www.worlddatasystem.org/community/membership [Accessed 25 March 2021].

122 World Data System. About. https://www.worlddatasystem.org/organization [Accessed 25 March 2021].

123 NASA Earthdata. Open data, services and software policies. https://earthdata.nasa.gov/collaborate/open-data-services-and-software [Accessed 25 March 2021].

124 The National Academies of Sciences Engineering Medicine. *Life-cycle decisions for biomedical data: the challenge of forecasting costs*. https://www.nationalacademies.org/our-work/forecasting-costs-for-preserving-archiving-and-promoting-access-to-biomedical-data (2020).

125 Privacy Europe. General data protection regulation (GDPR). https://gdpr-info.eu/ [Accessed 25 March 2021].

126 European Commission. For how long can data be kept and is it necessary to update it? https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/how-long-can-data-be-kept-and-it-necessary-update-it_en [Accessed 25 March 2021].

127 Beagrie, N. *What to keep: a Jisc research data study*. https://repository.jisc.ac.uk/7262/ (2019).

128 Digital Curation Centre. Five steps to decide what data to keep. https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep (2014) [Accessed 25 March 2021].

129 Lacey, J., Coates, R. & Herington, M. Open science for responsible innovation in Australia: understanding the expectations and priorities of scientists and researchers. *J. Responsible Innov.* **7**, 427–449 (2020).

130 Cao, S. What on earth is a data scientist? The Buzzword's inventor DJ Patil spills all. *Observer* (2019).

131 Davenport, T. H. & Patil, D. Data scientist: the sexiest job of the 21st century. *Harvard Business Review*. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (2012) [Accessed 16 February 2021].

132 The Royal Society. *Dynamics of data science skills: how can all sectors benefit from data science talent?* https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/ (2019).

133 OECD. *OECD Global Science Forum Building digital workforce capacity and skills for data-intensive science*. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/STP/GSF(2020)6/FINAL&docLanguage=En (2020) doi:10.1787/e08aa3bb-en.

134 Christensson, P. Data science definition. *Tech Terms* https://techterms.com/definition/data_science (2017) [Accessed 16 March 2021].

135 Wing, J. M. What is data science? *Columbia University Data Science Institute* https://www.datascience.columbia.edu/news/2018/what-is-data-science/ (2018) [Accessed 16 March 2021].

136 CODATA. Data Science Journal. https://datascience.codata.org/ [Accessed 16 March 2021].

137 Blei, D. M. & Smyth, P. Science and data science. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8689–8692 (2017).

138 Wing, J. M. The data life cycle. *Harvard Data Sci. Rev.* **1**, (2019).

139 Stodden, V. The data science life cycle: a disciplined approach to advancing data science as a science. *Commun. ACM* (2020) doi:10.1145/3360646.

140 Williamson, R. Process and purpose, not thing and technique: how to pose data science research challenges. *Harvard Data Sci. Rev.* (2020) doi:doi.org/10.1162/99608f92.6e525663.

141 Data to Decisions CRC. *Data Science Competency Framework*. *Data to Decisions* CRC https://www.d2dcrc.com.au/data-science-competency-framework (2017).

142 World Economic Forum. *Data science in the new economy: a new race for talent in the Fourth Industrial Revolution*. http://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf (2019).

143 World Economic Forum. *The future of jobs report 2018*. http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf (2018).

144 Deloitte Access Economics. *The future of work: occupational and education trends in data science in Australia*. https://www2.deloitte.com/au/en/pages/economics/articles/future-of-work-occupational-education-trends.html (2018).

145 Asplen-Taylor, S. & Pugh, R. Using data science to come out of Covid-19 stronger than the competition. *Information-Age* https://www.information-age.com/using-data-science-come-out-covid-19-stronger-than-competition-123490054/ (2020) [Accessed 26 March 2021].

146 SEEK. Seek: data scientist jobs. *Seek* https://www.seek.com.au/data-scientist-jobs (2021) [Accessed 11 June 2021].

147 Rodríguez-Sánchez, F., Marwick, B., Lazowska, E. & VanderPlas, J. Academia's failure to retain data scientists. *Science (New York, N.Y.)* vol. 355 357–358 (2017).

148 Australian Government Office of the Chief Scientist. *Australia's STEM workforce*. https://www.chiefscientist.gov.au/sites/default/files/2020-07/Australias STEM Workforce - Final.pdf (2020).

149 Wing, J. M. Ten research challenge areas in data science. *Columbia University Data Science Institute* https://datascience.columbia.edu/news/2019/ten-research-challenge-areas-in-data-science/ (2019) [Accessed 16 March 2021].

150 Ley, C. & Bordas, S. P. A. What makes data science different? A discussion involving Statistics2.0 and computational sciences. *Int. J. Data Sci. Anal.* **6**, 167–175 (2018).

151 Chamchoun, Y. Should data science be considered as its own discpline? *The Data Scientist* https://thedatascientist.com/data-science-considered-own-discipline/ (2019) [Accessed 16 March 2021].

152 Irizarry, R. A. The role of academia in data science education. *Harvard Data Sci. Rev.* (2020).

153 Meng, X.-L. Data science: an artificial ecosystem. *Harvard Data Sci. Rev.* (2019).

154 Stichweh, R. Differentiation of scientific discplines: causes and consequences. *Encycl. Life Support Syst.* **1**, (2006).

155 Areekkuzhiyil, S. Emergence of new disciplines. *Edutracks* **17**, 20–22 (2017).

156 Krishnan, A. *What are academic discplines? Some observations on the disciplinarity vs. interdiscplinarity debate*. https://core.ac.uk/download/pdf/229991542.pdf (2009).

157 Data Science Association. About the Data Science Association. *Data Science Association* https://www.datascienceassn.org/ [Accessed 16 March 2021].

158 Data Science and AI Association of Australia. https://dsai.org.au/ [Accessed 16 March 2021].

159 Breckler, S. The importance of disciplines. *Psychol. Sci. Agenda* (2005).

160 Tukey, J. W. The future of data analysis. *Ann. Math. Stat*. **33**, 1–67 (1962).

161 Cleveland, W. S. Data Science: an action plan for expanding the technical areas of the field of statistics. *Int. Stat. Rev*. **69**, 21–26 (2001).

162 Donoho, D. 50 years of data science. *J. Comput. Graph. Stat*. **26**, 745–766 (2017).

163 Cao, L. Data science: a comprehensive overview. *ACM Comput. Surv*. **50**, (2017).

164 McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol*. **52**, 97–99 (1990).

165 Turing, A. M. I — Computing machinery and intelligence. *Mind* **59**, 433–460 (1950).

166 Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **15**, 233–234 (2018).

167 National Science Foundation. Critical techniques, technologies and methodologies for advancing foundations and applications of big data sciences and engineering (BIGDATA). https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767 [Accessed 16 March 2021].

168 Berman, F. et al. *Realizing the potential of data science*. https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf (2016).

169 National Science Foundation. Harnessing the Data Revolution. https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp [Accessed 29 March 2021].

170 National Institutes of Health. National Institutes of Health Office of Data Science Strategy. https://datascience.nih.gov/ [Accessed 16 March 2021].

171 National Institute of Health. *NIH strategic plan for data science*. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf (2018).

172 The Alan Turing Institute. The Alan Turing Institute. https://www.turing.ac.uk/ (2021) [Accessed 16 March 2021].

173 The Alan Turing Institute. Data science at scale. https://www.turing.ac.uk/research/research-programmes/data-science-scale [Accessed 29 March 2021].

174 Clark, G., Hancock, M., Hall, W. & Pesenti, J. *Policy paper AI Sector Deal*. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal (2019).

175 University of Virginia. University of Virginia School of Data Science. https://datascience.virginia.edu/ (2021) [Accessed 16 March 2021].

176 University of Virginia School of Data Science. A school without walls. https://datascience.virginia.edu/about [Accessed 16 March 2021].

177 Data Science Institute. *School of Data Science - Phase II faculty senate submission*. https://api.dsi.virginia.edu/sites/default/files/attachments/2019-09/schoolofdatascience-190429155451.pdf (2019).

178 Manke, K. Berkeley inaugurates Division of Data Science and Information, connecting teaching and research from all corners of campus. *Berkeley News* https://news.berkeley.edu/2018/11/01/berkeley-inaugurates-division-of-data-science-and-information-connecting-teaching-and-research-from-all-corners-of-campus/ (2018) [Accessed 16 March 2021].

179 Manke, K. A fresh new name for UC Berkeley's data science division. *Berkeley News* https://news.berkeley.edu/2020/02/05/fresh-new-name-for-uc-berkeleys-data-science-division/ (2020) [Accessed 16 March 2021].

180 Berkeley Institute for Data Science. https://bids.berkeley.edu/ [Accessed 16 March 2021].

181 Australian Academy of Science and the Australian Academy of Technology & Engineering. *Preparing for Australia's digital future*. https://www.science.org.au/files/userfiles/support/reports-and-plans/2019/preparing-for-australias-digital-future.pdf (2019).

182 IP Australia. *Machine learning innovation: a patent analytics report*. https://www.ipaustralia.gov.au/sites/default/files/reports_publications/patent_analytics_report_on_machine_learning_innovation.pdf (2019).

183 Barlow, T. *The Australian miracle: an innovative nation revisited*. (Picador Australia, 2006).

184 Robinson, J. & Welsh, A. H. Peter Gavin Hall 1951-2016. *Hist. Rec. Aust. Sci*. **28**, 171–182 (2017).

185 Cheng, M.-Y. & Fan, J. Peter Hall's contributions to nonparametric function estimation and modeling. *Ann. Stat*. **44**, 1837–1853 (2016).

186 Beheshti, S.-M.-R., Tabebordbar, A., Benatallah, B. & Nouri, R. On automating basic data curation tasks. in *Proceedings of the 26th International Conference on World Wide Web Companion* 165–169 (International World Wide Web Conferences Steering Committee, 2017). doi:10.1145/3041021.3054726.

187 Freitas, A. & Curry, E. Big data curation. in *New Horizons for a Data-Driven Economy* (eds. Cavanillas, J., Curry, E. & Wahlster, W.) (Springer, 2016). doi:https://doi.org/10.1007/978-3-319-21569-3_6.

188 Pearce, R. OAIC, Data61 partner for data de-identification guide. *Computer World* https://www.computerworld.com/article/3471065/oaic-data61-partner-for-data-de-identification-guide.html (2017).

189 Ramage, D. Federated analytics: collaborative data science without data collection. *Google AI Blog* https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html (2020) [Accessed 14 February 2021].

190 OpenSAFELY. Home - Summary. https://opensafely.org/ [Accessed 29 March 2021].

191 The OpenSAFELY Collaborative et al. OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. *medRxiv* (2020) doi:https://doi.org/10.1101/2020.05.06.20092999.

192 The National Academies of Sciences Engineering Medicine. *Opportunities from the Integration of Simulation Science and Data Science: Proceedings of a Workshop*. https://www.nap.edu/catalog/25199/opportunities-from-the-integration-of-simulation-science-and-data-science (2018).

193 Flender, S. Why does deep learning work so well? *Towards Data Science* https://towardsdatascience.com/why-does-deep-learning-work-so-well-6550f3aa22c6 (2020) [Accessed 29 March 2021].

194 Sejnowski, T. J. The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci.* **117**, 30033 LP – 30038 (2020).

195 Thompson, N., Greenewald, K., Lee, K. & Manso, G. F. The computational limits of deep learning. *arXiv* (2020).

196 Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62**, 54–60 (2019).

197 Prosperi, M. et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* **2**, 369–375 (2020).

198 Wang, Y. & Blei, D. M. The Blessings of Multiple Causes. *arXiv* (2018).

199 Pearl, J. & Mackenzie, D. *The book of why: the new science of cause and effect*. (Basic Books, Inc., 2018).

200 Verma, S., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: a review. *arXiv* (2020).

201 Diaz, A., Rowshankish, K. & Saleh, T. *Why data culture matters*. (McKinsey Quarterly, 2018).

202 Noorzad, P. Models for integrating data science teams within companies: a comparative analysis. https://djpardis.medium.com/models-for-integrating-data-science-teams-within-organizations-7c5afa032ebd (2019) [Accessed 16 March 2021].

203 Grossman, R. & Siegel, K. P. Organizational models for big data and analytics. *Struct. Dimens. Organ. Behav. eJournal* (2014).

204 Australian Bureau of Statistics. Australian and New Zealand Standard Research Classification (ANZSRC), 2020. https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/1297.0 (2020) [Accessed 16 March 2021].

205 Autralian Research Council. NCGP Trends: areas of research. https://www.arc.gov.au/grants-and-funding/apply-funding/grants-dataset/trend-visualisation/ncgp-trends-areas-research [Accessed 16 March 2021].

206 Australian Research Council. Excellence in Research Australia. https://www.arc.gov.au/excellence-research-australia (2021) [Accessed 16 March 2021].

207 Australian Research Council. Engagement and Impact Assessment. https://www.arc.gov.au/engagement-and-impact-assessment (2021) [Accessed 16 March 2021].

208 Australian Research Council. ANZSRC Review. https://www.arc.gov.au/anzsrc-review (2020) [Accessed 16 March 2021].

209 Australian Academy of Science. Sectional Committee. https://www.science.org.au/fellowship/elections/sectional-committee (2020) [Accessed 16 March 2021].

210 Australian Academy of Science. *Australian climate science capability review*. https://www.science.org.au/files/userfiles/support/reports-and-plans/2017/climate-science-capability-review-2017.pdf (2017).

211 Investing in new platforms. *Australian Research Data Commons* https://ardc.edu.au/collaborations/strategic-activities/platforms/ [Accessed 23 May 2021].

212 Australian Research Data Commons. Request for proposal: Platforms program open call 2020. https://ardc.edu.au/wp-content/uploads/2020/06/RFP_Platforms_2020.pdf (2020) [Accessed 12 July 2021].

213 Australian Academy of Science. *Women in STEM Decadal Plan*. https://www.science.org.au/files/userfiles/support/reports-and-plans/2019/gender-diversity-stem/women-in-STEM-decadal-plan-final.pdf (2019).

214 Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *Science (80-.).* **330**, 686–688 (2010).

215 Sastre, J. F. The impact of R&D in teams' gender diversity on innovation outputs. *Int. J. Entrep. Small Bus.* **24**, (2015).

216 Maceira, H. M. Economic benefits of gender quality in the EU. *Gend. Equal.* **52**, 178–783 (2017).

217 UNESCO. *UNESCO science report: towards 2030*. (2015).

218 Sylvain, D. et al. What's keeping women out of data science? *Boston Consulting Group* https://www.bcg.com/en-au/publications/2020/what-keeps-women-out-data-science (2020) [Accessed 16 March 2021].

219 Better Buys. Why we need women in data science. https://www.betterbuys.com/bi/women-in-data-science/ [Accessed 16 March 2021].

220 Finkel, A. & Harvey-Smith, L. Chief Scientist: women in STEM are still far short of workplace equity. COVID-19 risks undoing even these modest gains. *The Conversation* https://theconversation.com/chief-scientist-women-in-stem-are-still-far-short-of-workplace-equity-covid-19-risks-undoing-even-these-modest-gains-143092 (2020) [Accessed 16 March 2021].

221 The Alan Turing Institute. Women in data science and AI. https://www.turing.ac.uk/research/research-projects/women-data-science-and-ai [Accessed 16 March 2021].

222 Women in Data Science. Women in Data Science (WiDS) Worldwide Initiative. https://www.widsconference.org/ [Accessed 16 March 2021].

223 Data Science For Social Good. Data science to create social impact. http://www.datascienceforsocialgood.org/ [Accessed 16 March 2021].

224 United Nations. Big data for sustainable development. https://www.un.org/en/global-issues/big-data-for-sustainable-development [Accessed 16 March 2021].

225 Criado-Perez, C. *Invisible women: data bias in a world designed for men*. (Abrams Press, 2021).

226 McGrail, K. M. et al. A position statement on population data science: the science of data about people. *Int. J. Popul. Data Sci.* **3**, 1–11 (2018).

227 Meng, X.-L. What Are the Values of Data, Data Science, or Data Scientists? *Harvard Data Science Review* https://hdsr.mitpress.mit.edu/pub/bj2dfcwg/release/2 (2021) doi:10.1162/99608f92.ee717cf7.

228 Australian Government Research Infrastructure Investment Plan. *Facilities for the future: underpinning Australia's research and innovation*. https://2021nriroadmap.dese.gov.au/resources/ (2018).

229 Government of Canada. Digital research infrastructure. https://www.ic.gc.ca/eic/site/136.nsf/eng/home (2019) [Accessed 16 March 2021].

230 AARNet. The Indigo Project. https://www.aarnet.edu.au/network-and-services/the-network/indigo-project (2021) [Accessed 7 June 2021].

231 Fleming, A., Jakku, E., Lim-Camacho, L., Taylor, B. & Thorburn, P. Is big data for big farming or for everyone? Perceptions in the Australian grains industry. *Agron. Sustain. Dev.* **38**, (2018).

232 Australian Access Federation. Home - Australian Access Federation. [*On-line*] https://aaf.edu.au/ (2021) [Accessed 7 June 2021].

233 TOP500.org. TOP500 List - June 2020. *TOP500* https://www.top500.org/lists/top500/list/2020/06/ (2020) [Accessed 11 June 2021].

234 Pawsey Supercomputing Centre. Keeping pace with global advances in supercomputing technology. https://pawsey.org.au/about-us/capital-refresh/ (2021) [Accessed 16 March 2021].

235 Australian Research Data Commons (ARDC). Our Strategy. https://ardc.edu.au/our-strategy/ [Accessed 16 March 2021].

236 Australian Research Data Commons (ARDC). Data and services. https://ardc.edu.au/our-strategy/data-and-services/ [Accessed 16 March 2021].

237 Australian Research Data Commons (ARDC). Platforms and software. https://ardc.edu.au/our-strategy/software-and-platforms/ [Accessed 16 March 2021].

238 Australian Research Data Commons (ARDC). Storage and compute. https://ardc.edu.au/our-strategy/storage-and-compute/ [Accessed 16 March 2021].

239 Australian Research Data Commons (ARDC). People and policy. https://ardc.edu.au/our-strategy/people-and-policy/ (2020) [Accessed 16 March 2021].

240 Australian Research Data Commons. Bridging continents to share and re-use research data. https://ardc.edu.au/news/bridging-continents-to-share-and-re-use-research-data/ (2021) [Accessed 7 June 2021].

241 AARNET. CloudStor. https://www.aarnet.edu.au/network-and-services/cloud-services/cloudstor [Accessed 16 March 2021].

242 Australian National Data Service. The ANDS, Nectar and RDS partnership. https://www.ands.org.au/about-us/ands-nectar-rds [Accessed 16 March 2021].

243 Hicks, R., Marks, H., Stickells, S. & Hancock, C. Level UP – increasing collaboration across Australia's Digital Data and eResearch Platforms (DDeRP). in *eResearch NZ 2021* (2021).

244 AARNET. AARNet Pty Ltd Financial Report and Directors' Report 2020. https://www.aarnet.edu.au/images/uploads/resources/AARNet_Annual_Report_2020_Financials.pdf (2021) [Accessed 7 June 2021].

245 Australian Access Federation. Financial report 2020. https://aaf.edu.au/wp-content/uploads/AAF-Financial-Report-2020.pdf (2021) [Accessed 7 June 2021].

246 AARNET. Networking the Square Kilometre Array. https://www.aarnet.edu.au/case-studies/Networking-the-Square-Kilometre-Array (2020) [Accessed 7 June 2021].

247 Australian Charities and Not-for-profits Commission. AUSTRALIAN RESEARCH DATA COMMONS LIMITED - Financial and Documents. https://www.acnc.gov.au/charity/c6a37974698b21752689f58f9aaeef79#financials-documents (2021) [Accessed 7 June 2021].

248 Geoscience Australia. Data and publications search. https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/home [Accessed 16 March 2021].

249 Bureau of Meteorology. Climate and oceans data analysis. http://www.bom.gov.au/climate/data-services/ [Accessed 16 March 2021].

250 Australian Bureau of Statistics. Data integration project register, Australia. https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1900.0main+features5Australia (2019) [Accessed 16 March 2021].

251 CSIRO. Data access portal. https://data.csiro.au/collections [Accessed 16 March 2021].

252 Australian Institute of Marine Science. AIMS data. https://www.aims.gov.au/docs/data/data.html [Accessed 16 March 2021].

253 ANSTO. What we do. https://www.ansto.gov.au/about/what-we-do [Accessed 16 March 2021].

254 Integrated Marine Observing System. Integrated Marine Observing System. http://imos.org.au/ [Accessed 16 March 2021].

255 TERN. TERN ecosystem research infrastructure. https://www.tern.org.au/ [Accessed 16 March 2021].

256 Atlas of Living Australia. Open access to Australia's biodiversity data. https://www.ala.org.au/ [Accessed 16 March 2021].

257 AURIN. Australia's spatial intelligence network. https://aurin.org.au/ [Accessed 16 March 2021].

258 Bioplatforms Australia. Welcome to Bioplatforms Australia. https://bioplatforms.com/ [Accessed 16 March 2021].

259 PHRN. What is PHRN? https://www.phrn.org.au/ [Accessed 16 March 2021].

260 AuScope. AuScope. https://www.auscope.org.au/ [Accessed 16 March 2021].

261 National Research Infrastructure for Australia. Astronomy Australia Ltd. http://www.astronomyaustralia.org.au/ [Accessed 16 March 2021].

262 Australian Government Department of Health. Medical Research Future Fund. https://www.health.gov.au/initiatives-and-programs/medical-research-future-fund [Accessed 16 March 2021].

263 Teperek, M., & Dunning, A. (2018). Research data should be available long-term...but who is going to pay? *London School of Economics*. https://blogs.lse.ac.uk/impactofsocialsciences/2018/09/07/research-data-should-be-available- (2018) [Accessed 16 March 2021].

264 ARDC. About Research Data Australia. https://researchdata.edu.au/page/about [Accessed 16 March 2021].

265 Australian Bureau of Statistics. About the DataLab. https://www.abs.gov.au/ausstats/abs@.nsf/mf/1406.0.55.007 [Accessed 16 March 2021].

266 Google. Google AI. https://ai.google/ [Accessed 16 March 2021].

267 Google. Dataset search. https://datasetsearch.research.google.com/ [Accessed 16 March 2021].

268 Schema.org. Welcome to Schema.org. https://schema.org/ (2021) [Accessed 13 May 2021].

269 Google. Google Earth. https://www.google.com/earth/ [Accessed 17 June 2021].

270 Morris, B. Private communication to M. Barber. (2020).

271 Chapman, A. et al. Dataset search: a survey. *VLDB J.* **29**, 251–272 (2020).

272 Levett, K. & Brady, C. Metadata for impact: make RIF-CS work for you. *Australian Research Data Commons* https://documentation.ardc.edu.au/display/DOC/Metadata+for+Impact%3A+make+RIF-CS+work+for+you (2020) [Accessed 16 March 2021].

273 ARDC. National data assets. https://ardc.edu.au/collaborations/strategic-activities/national-data-assets/ [Accessed 16 March 2021].

274 Australian BioCommons. Australian BioCommons. https://www.biocommons.org.au/ [Accessed 16 March 2021].

275 Bureau of Meterology. Space Weather Service. http://www.sws.bom.gov.au/World_Data_Centre [Accessed 16 March 2021].

276 Australian Antarctic Data Centre. Australian Antarctic Data Centre: data management and spatial data services. https://data.aad.gov.au/ [Accessed 16 March 2021].

277 PARADISEC. Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). https://www.paradisec.org.au/ (2016) [Accessed 16 March 2021].

278 Australian Data Archive. The Australian Data Archive. https://ada.edu.au/ [Accessed 16 March 2021].

279 Atkins, N. et al. Trusted data community: working together to certify Australia's repositories. in *eResearch Australia Conference* (eResearch Australasia, 2019).

280 National Science Foundation. Cyberinfrastructure framework for 21st century science and engineering (CIF21). (2012).

281 Dietrich, M., Jones, N. & New Zealand eScience Infrastructure. Mapping eResearch ecosystems: the international situation is intensifying! (2019) doi:10.6084/m9.figshare.8066942.v1.

282 New Zealand eScience Infrastructure. NeSI. https://www.nesi.org.nz/ [Accessed 16 March 2021].

283 CSC IT Centre for Science. CSC: ICT solutions for brilliant minds. https://www.csc.fi/en/home (2020) [Accessed 16 March 2021].

284 European Open Science Cloud (EOSC). *European Commission* https://eosc-portal.eu/ (2020) [Accessed 1 July 2021].

285 European Commission. Horizon 2020 projects. https://ec.europa.eu/programmes/horizon2020/en/h2020-sections-projects [Accessed 16 March 2021].

286 NDRIO. NDRIO: supporting a world-class, collaborative and competitive Digital Research Infrastructure (DRI) community in Canada. https://www.engagedri.ca/ [Accessed 16 March 2021].

287 ODISSEI. ODISSEI – Open Data Infrastructure for Social Science and Economic Innovations. https://odissei-data.nl/en/ (2020) [Accessed 7 June 2021].

288 National Science and Technology Council. *The Convergence of High Performance Computing, Big Data, and Machine Learning: Summary of the Big Data and High End Computing Interagency Working Groups Joint Workshop*. https://www.nitrd.gov/pubs/Convergence-HPC-BD-ML-JointWSreport-2019.pdf (2018).

289 Asch, M. et al. Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *Int. J. High Perform. Comput. Appl.* **32**, 435–479 (2018).

290 Vetter, J. S. et al. *Extreme Heterogeneity 2018 - Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity*. http://www.osti.gov/servlets/purl/1473756/ (2018) doi:10.2172/1473756.

291 Caulfield, B. CPU vs GPU: What's the difference? *The Official NVIDIA Blog* https://www.intel.com.au/content/www/au/en/products/docs/processors/cpu-vs-gpu.html (2009) [Accessed 7 June 2021].

292 Osuna, C. & Sawyer, W. Status of the ICON Climate Model port to GPUs. *4th ENES Work. High Perform. Comput. Clim. Weather* (2016).

293 Australian Government. *Status report on the NCRIS eResearch capability*. (2015).

294 Australian Government Department of Innovation Science and Research. *2011 Strategic Roadmap for Australian Research Infrastructure*. https://apo.org.au/node/24415 (2011).

295 Research Data Infrastructure Committee. *The Australian research data infrastructure strategy.* https://apo.org.au/node/42792 (2013).

296 Cook, R. NRDC future design: HASS workshop. in *Humanities, Arts & Culture Data Summit* (2018).

297 Australian Government Bureau of Meteorology and CSIRO. ACCESS. https://www.cawcr.gov.au/research/access/ [Accessed 16 March 2021].

298 The University of Queensland. Research infrastructure investment plan scoping study: National Environmental Prediction System (NEPS). https://science.uq.edu.au/neps [Accessed 16 March 2021].

299 Australian Academy of Science. National Committees for Science. https://www.science.org.au/supporting-science/national-committees-science [Accessed 16 March 2021].

300 Australian Government Department of Education Skills and Employment. 2021 National Research Infrastructure Roadmap Consultation. https://2021nriroadmap.dese.gov.au/ (2021) [Accessed 7 June 2021].

301 Centre of Excellence for Exascale Supercomputing in the area of the Solid Earth. Objectives | ChEESE - Centre of Excellence for Exascale Supercomputing in the area of the Solid Earth. https://cheese-coe.eu/about/objectives (2019) [Accessed 7 June 2021].

302 EVOLVE. EVOLVE - Introduction to the Project. https://www.evolve-h2020.eu/about/introduction-to-evolve/ (2020) [Accessed 7 June 2021].

303 Australian Academy of Science. Decadal plans for science. *Australian Academy of Science* https://www.science.org.au/supporting-science/science-policy-and-analysis/decadal-plans-science [Accessed 16 March 2021].

304 Australian Academy of Science National Committee for Astronomy. *Decadal Plan for Australian astronomy 2016-2025 mid-term review*. https://www.science.org.au/files/userfiles/support/reports-and-plans/2020/astronomy-decadal-plan-mid-term-review-07-2020.pdf (2020).

305 World Economic Forum. *The future of jobs report*. https://www.weforum.org/reports/the-future-of-jobs-report-2018 (2018).

306 Skoolville. 27 universities with masters in data science in Australia - 2019. https://skoolville.com/blog/australia-universities-with-masters-in-data-science/ (2019) [Accessed 26 March 2021].

307 Bello, M. & Galindo-Rueda, F. *Charting the digital transformation of science*. https://www.oecd-ilibrary.org/science-and-technology/charting-the-digital-transformation-of-science_1b06c47c-en (2020) doi:10.1787/1b06c47c-en.

308 Buchhorn, M. *Surveying the scale of the research-IT support workforce - a survey and report comissioned by the Australian Research Data Commons (ARDC)*. https://ardc.edu.au/wp-content/uploads/2019/07/ARDC-National-Workforce-report-final-v3.pdf (2019).

309 Rosenthal, N. First PhD Course a success; bioinformatics survey results; and a new language for a new biology: EMBL Australia in August. *Science in Public* (2013).

310 Australian Academy of the Humanities. *Future humanities workforce: consultation summaries*. https://www.humanities.org.au/wp-content/uploads/2019/12/AAH-FHW-Consultation-Summaries.pdf (2019).

311 Burtch, L. *The Burtch Works study: salaries of data scientists & predictive analytics professionals*. https://www.burtchworks.com/wp-content/uploads/2019/06/Burtch-Works-Study_DS-PAP-2019.pdf (2019).

312 National Academies of Sciences Engineering and Medicine. *Envisioning the data science discipline - The undergraduate perspective: interim report*. https://www.nap.edu/catalog/24886/envisioning-the-data-science-discipline-the-undergraduate-perspective-interim-report (2018).

313 CrowdFlower. *Data science report*. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf (2016).

314 Ferrari, A., Punie, Y. & Brecko, B. *DIGCOMP: a framework for developing and understanding digital competence in Europe*. (2013) doi:10.2788/52966.

315 Melbourne Data Analytics Platform (MDAP). Defining the third space. *University of Melbourne* https://mdap.unimelb.edu.au/2020/09/17/defining-the-third-space/ (2020) [Accessed 11 December 2020].

316 Paulus, F. M., Rademacher, L., Schäfer, T. A. J., Müller-Pinzler, L. & Krach, S. Journal impact factor shapes scientists' reward signal in the prospect of publication. *PLoS One* **10**, e0142537 (2015).

317 Raff, J. W. The San Francisco Declaration on Research Assessment. *Biol. Open* **2**, 533–534 (2013).

318 Declaration on Research Assessment. Signers. https://sfdora.org/signers/ (2021) [Accessed 14 May 2021].

319 Infosys. *Amplifying human potential: Education and skills for the fourth industrial revolution*. https://imagesec.infosys.com/Web/Infosys/%7B8adf71d4-ce0c-48e1-8299-83e1dbd8c0c4%7D_Infosys-Amplifying-Human-Potential-new.pdf?elqTrackId=dd2dc4dfc43b432683b7b3b80355353a&elqaid=1049&elqat=2 (2016).

320 Australian Curriculum. Australian Curriculum: digital technologies. https://www.australiancurriculum.edu.au/f-10-curriculum/technologies/digital-technologies/ (2019) [Accessed 7 June 2021].

321 Hyndman, B. Ten reasons teachers can struggle to use technology in the classroom. *Sci. Educ. News* **67**, 1–6 (2018).

322 Commonwealth of Australia. The Agenda | National Innovation and Science Agenda. i–18 https://www.industry.gov.au/data-and-publications/national-innovation-and-science-agenda-report (2015) [Accessed 14 March 2020].

323 Australian Computing Academy. Australian digital technologies challenges: online and unplugged activities for your classroom. https://aca.edu.au/projects/dt-challenges/ (2019) [Accessed 7 November 2020].

324 CSIRO. STEM professionals in schools. https://www.csiro.au/en/education/programs/stem-professionals-in-schools (2021) [Accessed 13 April 2021].

325 University of Adelaide. CSER Digital Technologies Education. *University of Adelaide* https://csermoocs.adelaide.edu.au/ (2017) [Accessed 19 October 2020].

326 Open Working from 4TU.ResearchData & TU Delft Library. Petascale Campus Initiative. *University of Melbourne* https://openworking.wordpress.com/2020/03/10/day-1-from-petascale-campus-to-community-driven-training-a-myriad-of-innovative-data-initiatives-at-the-university-of-melbourne/ (2020) [Accessed 11 June 2021].

327 The University of Sydney. Sydney informatics hub. *University of Sydney* https://www.sydney.edu.au/research/facilities/sydney-informatics-hub.html (2021) [Accessed 11 June 2021].

328 QUT Centre for Data Science. QUT Centre for Data Science. *QUT* https://research.qut.edu.au/qutcds/ (2021) [Accessed 11 June 2021].

329 Australian Data Science Network. Australian Data Science Network. https://www.australiandatascience.net/ (2AD) [Accessed 16 March 2021].

330 Öster, P. Opening plenary, Wednesday 23 October 2019. in *RDA 14th Plenary* (2019).

331 ARDC. Skilled workforce. https://ardc.edu.au/collaborations/skilled-workforce/ (2019) [Accessed 2 December 2020].

332 Council, A. R. Industrial Transformation Training Centres. *Australian Research Council* https://www.arc.gov.au/grants/linkage-program/industrial-transformation-research-program/industrial-transformation-training-centres (2020) [Accessed 14 October 2020].

333 ARC Training Centre in Cognitive Computing for Medical Technologies. https://aimedtech.org.au/ (2017) [Accessed 11 June 2021].

334 Hodkiewicz, M., Small, M., Loxton, R., Griffin, M. & Klump, J. ARC Training Centre for Transforming Maintenance through Data Science. https://www.maintenance.org.au (2020) [Accessed 3 December 2020].

335 ARC Training Centre in Data Analytics for Resources & Environments (DARE). https://darecentre.org.au/ [Accessed 14 March 2020].

336 Australian Research Council. New ARC training centres to lead Australian industry research. https://www.arc.gov.au/news-publications/media/media-releases/new-arc-training-centres-lead-australian-industry-research (2020) [Accessed 14 July 2020].

337 Bangert, D. A curriculum for foundational research data science skills for early career researchers. *Research Data Alliance* https://rd-alliance.org/group/rdacodata-summer-schools-data-science-and-cloud-computing-developing-world-wg/outcomes-0 (2019) [Accessed 17 November 2020].

338 2020 online - course list. *FORCE 11* https://www.force11.org/fsci/2020/2020-online-course-list (2020) [Accessed 13 November 2020].

339 Higman, R., Teperek, M. & Kingsley, D. Creating a community of data champions. *Int. J. Digit. Curation* **12**, (2017).

340 R-Ladies Global. R-Ladies Global. https://rladies.org/ (2012) [Accessed 11 June 2021].

341 ResBaz. Research Bazaar, 2019. *ResBaz* https://resbaz.github.io/resbaz2019/ (2019) [Accessed 11 June 2021].

342 The R Foundation. R Conferences. *The R Project* https://www.r-project.org/conferences/ (2020) [Accessed 12 November 2020].

343 The University of Melbourne. Melbourne Data Analytics Platform. https://mdap.unimelb.edu.au/ (2020) [Accessed 19 October 2020].

344 Lyon, L., Ball, A. J., Day, M. & Duke, M. Developing a community capability model framework for data-intensive research. in *iPRES 2012: 9th International Conference on Preservation of Digital Objects* (2012).

345 European Commission. What are national coalitions? Who do they bring together and in order to do what? https://eufordigital.eu/e-card/what-are-national-coalitions-who-do-they-bring-together-and-in-order-to-do-what/ (2018) [Accessed 4 November 2020].

346 Newbold, E. et al. D6.2 Initial core competence centre structures. *Zenodo* https://zenodo.org/record/3732889#.YOunYugzYuU (2019) [Accessed 13 March 2021].

347 Newbold, E. et al. D6.1 Overview of needs for competence centres. *Zenodo* https://zenodo.org/record/3549791#.YOuucegzYuU (2019) [Accessed 4 March 2021].

348 European Commission. Digital Education Action Plan 2021-2027, Factsheet. *Dg Eac* (2020).

349 Vidal, F. *National Plan for Open Science*. https://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (2018).

350 Moore-Sloan Data Science Environments. *MSDSE* http://msdse.org/ (2019) [Accessed 3 April 2020].

351 Moore-Sloan Data Science Environments. *Creating institutional change in data science*. http://msdse.org/files/Creating_Institutional_Change.pdf (2018).

352 UC Berkeley. Strategic plan. https://strategicplan.berkeley.edu/ [Accessed 15 April 2021].

353 UC Berkeley taps Microsoft Research Fellow Jennifer Tour Chayes to lead new data science division. *Berkeley Computing, Data Science, and Society* https://data.berkeley.edu/news/uc-berkeley-taps-microsoft-research-fellow-jennifer-tour-chayes-lead-new-data-science-division (2019) [Accessed 11 June 2021].

354 Participants of African Open Science Platform Stakeholder Workshop September 2018. The Future of Science and Science of the Future: Vision and Strategy for the African Open Science Platform (v02). in *Zenodo* (2018). doi:10.5281/zenodo.2222418.

355 National Institutes of Health. Data and research center. https://allofus.nih.gov/funding-and-program-partners/data-and-research-center (2016) [Accessed 18 December 2020].

356 UK Research Integrity Office. Welcome to the website for the UK Research Integrity Office. https://ukrio.org/ (2020) [Accessed 8 August 2020].

357 Universities UK. *The concordat to support research integrity*. https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Pages/the-concordat-for-research-integrity.aspx (2019).

358 Fanelli, D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* **4**, e5738 (2009).

359 Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24 (2020).

360 Armstrong, S. Research on covid-19 is suffering "imperfect incentives at every stage". *BMJ* **369**, m2045 (2020).

361 Mehra, M. R., Desai, S. S., Ruschitzka, F. & Patel, A. N. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* (2021) doi:10.1016/S0140-6736(20)31180-6.

362 World Health Organization. Coronavirus disease (COVID-19): Hydroxychloroquine. https://www.who.int/news-room/q-a-detail/q-a-hydroxychloroquine-and-covid-19 (2020) [Accessed 30 March 2021].

363 The Lancet Editors. Expression of concern: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* **395**, (2020).

364 Davey, M. Surgisphere: mass audit of papers linked to firm behind hydroxychloroquine Lancet study scandal. *The Guardian* https://www.theguardian.com/world/2020/jun/10/surgisphere-sapan-desai-lancet-study-hydroxychloroquine-mass-audit-scientific-papers (2020) [Accessed 30 March 2021].

365 Bolland, M. J., Avenell, A., Gamble, G. D. & Grey, A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology* **87**, 2391–2402 (2016).

366 Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallographica Section D: Biological Crystallography* vol. 65 148–155 (2009).

367 Pitt, J. H. & Hill, H. Z. *Statistical detection of potentially fabricated numerical data: a case study*. https://arxiv.org/pdf/1311.5517.pdf (2013).

368 M Fleming, R., R Fleming, M. & K Chaudhuri, T. Establishing data validity: statistically determining if data is fabricated, falsified or plagiarized. *Acta Sci. Med. Sci.* **3**, 169–191 (2019).

369 Swan, N. Scientific & financial misconduct. *ABC The Science Show* https://www.abc.net.au/radionational/programs/scienceshow/scientific--financial-misconduct/3505366 (2002) [Accessed 30 March 2021].

370 Dayton, L. Hall found guilty of lesser misconduct. *Science (80-.).* **303**, 298 (2004).

371 van der Weyden, M. B. Managing allegations of scientific misconduct and fraud: lessons from that 'Hall affair'. *Med. J. Aust.* **180**, 149–151 (2004).

372 Australian Government National Health and Medical Research Council Australian Research Council and Universities Australia. The Australian Code for the Responsible Conduct of Research, 2007. https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2007 (2007) [Accessed 30 March 2021].

373 The State of Queensland Crime and Corruption Commission. *Australia's first criminal prosecution for research fraud*. https://services.anu.edu.au/files/guidance/Australias-first-criminal-prosecution-for-research-fraud-final.pdf (2017).

374 Nutt, A. A. Australian neuroscientist Bruce Murdoch nearly went to jail for making up data. *The Sydney Morning Herald* https://www.smh.com.au/technology/australian-neuroscientist-bruce-murdoch-nearly-went-to-jail-for-making-up-data-20160405-gnydf6.html (2016) [Accessed 30 March 2021].

375 Australian Broadcasting Corporation. 'Brazen' UQ research fraudster Caroline Barwood given suspended sentence. *ABC News* https://www.abc.net.au/news/2016-10-25/uq-dr-caroline-barwood-avoids-jail-on-fraud-charges/7963178 (2016) [Accessed 3 March 2021].

376 Oransky, I. An Australian university cleared a cancer researcher of misconduct. He's now retracted six papers. *Retraction Watch* https://retractionwatch.com/2019/01/14/an-australian-university-cleared-a-cancer-researcher-of-misconduct-hes-now-retracted-six-papers/ (2019) [Accessed 30 March 2021].

377 Worthington, E. & Taylor, K. UNSW skin cancer researcher Levon Khachigian hit with string of retractions. *ABC News* https://www.abc.net.au/news/2019-10-17/unsw-skin-cancer-levon-khachigian-allegations-and-retractions/11585768 (2019) [Accessed 30 March 2021].

378 Worthington, E. Swinburne University researcher has 30 papers retracted, loses job. *ABC News* https://www.abc.net.au/news/2019-10-26/swinburne-university-researcher-has-30-papers-retracted/11641136 (2019) [Accessed 30 March 2021].

379 Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

380 The National Academies of Sciences Engineering Medicine. *Reproducibility and replicability in science*. https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science (2019).

381 Association for Computing Machinery. Artifact review and badging - Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current (2020) [Accessed 30 March 2021].

382 Association for Computing Machinery. KDD 2020 Virtual Conference. https://www.kdd.org/kdd2020/ (2020) [Accessed 30 March 2021].

383 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2019. Accepted papers. https://ecmlpkdd2019.org/programme/accepted/ (2019) [Accessed 30 March 2021].

384 ACM Special Interest Group on Management of Data. ACM SIGMOD Reproducibility Award. http://sigmod.org/sigmod-awards/sigmod-most-reproducible-paper-award [Accessed 30 March 2021].

385 Proceedings of the VLDB Endowment. PVLDB reproducibility. https://vldb.org/pvldb/reproducibility/ [Accessed 30 March 2021].

386 FATML. Fairness, accountability, and transparency in machine learning. https://www.fatml.org/ [Accessed 30 March 2021].

387 Reproducibility Challenge. Reproducibility challenge @ NeurIPS 2019. https://reproducibility-challenge.github.io/neurips2019/ [Accessed 30 March 2021].

388 The Five Safes. The five safes: effective decision-making for data and risk management. http://www.fivesafes.org/ [Accessed 25 March 2021].

389 The Australian Academy of Health and Medical Sciences. *Artifcial Intelligence in Health: Exploring the Opportunities and Challenges*. https://aahms.org/policy/roundtable-report-on-ai-in-health/ (2020).

390 Ross, J. AI advocates 'not angry enough' about university funding. *Times Higher Education* https://www.timeshighereducation.com/news/ai-advocates-not-angry-enough-about-university-funding (2020) [Accessed 31 March 2021].

References for box stories

i Group of Eight Australia. Go8 joins Paris talks to push for greater global open access to research data. https://go8.edu.au/go8-joins-paris-talks-to-push-for-greater-global-open-access-to-research-data (2020) [Accessed 24 June 2020].

ii Research Data Rights Summit. Sorbonne declaration on research data rights. https://www.leru.org/files/Sorbonne-declaration.pdf (2020) [Accessed 25 March 2021].

iii International Alliance of Research Library Associations. https://iarla.org/ [Accessed 25 March 2021].

iv Council of Australian University Librarians. Welcome to CAUL. https://www.caul.edu.au/ [Accessed 25 March 2021].

v International Alliance of Research Library Associations. IARLA supports the Sorbonne Declaration on Research Data Rights. https://iarla.org/2020/04/iarla-supports-the-sorbonne-statement-on-research-data-rights/ [Accessed 25 March 2021].

vi World Health Organization. *WHO forum on health data standardization and interoperability*. https://www.who.int/ehealth/WHO_Forum_on_HDSI_Report.pdf?ua=1 (2012).

vii Global Alliance for Genomics & Health. Enabling genomic data sharing for the benefit of human health. https://www.ga4gh.org/about-us/ (2020) [Accessed 25 March 2021].

viii Global Alliance for Genomics & Health. *GA4GH Connection: a 5-year strategic plan*. https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf (2017).

ix Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). Enabling FAIR data project. http://www.copdess.org/enabling-fair-data-project/ (2018) [Accessed 25 March 2021].

x International Union of Pure and Applied Chemistry. IUPAC endorses the chemistry GO FAIR manifesto. https://iupac.org/iupac-endorses-the-chemistry-go-fair-manifesto/ (2019) [Accessed 25 March 2021].

xi Gentes, Z. Ecologists ask: should we be more transparent with data? *The Ecological Society of America* https://www.esa.org/blog/2018/10/26/ecologists-ask-should-we-be-more-be-more-transparent-with-data/ (2018) [Accessed 18 December 2020].

xii Martone, M. Data sharing in neuroscience - Challenges and opportunities for moving neuroscience towards open and FAIR. https://www.med.uio.no/imb/english/research/news-and-events/events/distinguished-seminar/2018/maryann-martone.html (2018) [Accessed 25 March 2021].

xiii Draxl, C. & Scheffler, M. Big-data-driven materials science and its FAIR data infrastructure. in *Handbook of Materials Modeling* (eds. Andreoni, W. & Yip, S.) (Springer, Cham, 2020). doi:10.1007/978-3-319-44677-6_104.

xiv Griffin, P. C. et al. Best practice data life cycle approaches for the life sciences. *F1000Research* **6**, 1618 (2017).

xv STM Research Data. Share-Link-Cite. https://www.stm-researchdata.org/ (2021) [Accessed 24 June 2021].

xvi Exascale Computing Project. Understanding Exascale. https://www.exascaleproject.org/what-is-exascale/ (2021) [Accessed 8 July 2021].

xvii Mann, A. Core Concept: Nascent exascale supercomputers offer promise, present challenges. *Proc. Natl. Acad. Sci.* **117**, 22623–22625 (2020).