Challenges

Answer: Agree

Comments: Specific comments are listed for each of the outcomes below.

Outcome 1. Underpinned by training frameworks for researchers and NRI workforce

Answer: Agree.

Comments:

1) True data science expertise involves the intersection of domain expertise (e.g. in physics, chemistry, mathematics, health, etc), computer science and mathematics. It is very common in data science training to see computer science presented as the only necessary part of the equation. A coherent national framework is the ideal place to develop and promote cohesive training that ties together domain-specific knowledge with the relevant maths and computing elements.

2) We have an urgent need to upskill existing workers in addition to training undergraduates and postgraduates. This is further complicated by the wide disparities in digital literacy that one encounters across scientific disciplines. We therefore encourage the government to think of incentives for creating expert training at all career levels, which could intersect with existing grant schemes or funding councils. e.g. awarding dedicated training centres in specific disciplines.

3) It is becoming clear that schemes such as ARC Discovery projects and ARC Centres of Excellence have shifted towards more applied areas in recent years. However, fundamental science is traditionally a power house of R&D in big data methods, high performance computing, statistical inference and machine learning due to the unique challenges of the huge datasets taken at the research frontier. This includes the exchange of methods, data and ideas with massive networks of international collaborators (e.g. the Large Hadron Collider experiments at CERN, large-scale astronomy experiments, the LIGO gravitational wave observatory). Many of our domestic data scientists graduate from fundamental science research programmes, making a strong fundamental science sector a key driver of NDRI. The government could consider formalising this relationship by either providing guidance to funding councils on how to recognise these contributions (e.g. through modified national interest test requirements or national research priorities), or by providing bespoke funding schemes that marry data science development with fundamental domain research.

2. Responsive to disruptive technologies and societal shifts

Answer: Agree.

1) Novel computing architectures (e.g. the rise of GPU-based systems over CPU-based systems) often require many person hours to port existing software code or rewrite it from the ground up. An investment in personnel to do this is therefore crucial if researchers or other stakeholders are to realise the potential of breakthroughs in hardware. There are some great examples already of national support for optimising software (e.g. the PaCER scheme from the Pawsey Centre for Extreme Scale Readiness). Embedding these, and expanding on them, in the nationally-coordinated strategic planning initiative should be a priority. This also relates to Outcome 6.

2) Little mention is made of the ethical dimension of data science research (outside of a reference to public acceptance of AI, and issues surrounding data collection). The development and deployment of AI systems raises many complex ethical problems, e.g. the displacement of workers, the challenges of information warfare and how to make decisions in a self-driving car. The section on

societal shifts should therefore include a statement about how the ethical dimension will be addressed.

3. Consistent in its standards for data collection, curation and access

Answer: Neutral.

Comments:

1) We find it unlikely that a “single, consistent framework across all research disciplines” will be workable in practise. For example, researchers on Large Hadron Collider experiments inherit all procedures and hardware specifications from CERN, and the same must be true for scientists on other large-scale international projects. It is not within the remit of Australian science to make these consistent. It might be possible to establish a single framework for e.g. NCRIS-funded projects, in which case there should be an explicit separation between data taken and stored within Australia vs data from overseas. Even this might prove insurmountable however – data storage methods vary massively between disciplines, as does the literacy of typical researchers in storing and processing raw data.

2) Designing systems for storing and parsing data long after it was taken is itself an interesting challenge. Data analysis methods and systems that work now will likely be obsolete in 10-20 years time, but it will still be important to access archived data. Any coherent framework for data curation must therefore have enough flexibility to be future-proof. It should also be continuously reviewed.

4. Integrated across levels of computing and data

Answer: Strongly agree.

Comments:

1) Greater integration of access to different tiers of computing capability and shared data is a laudable goal, and is probably achievable. It will also probably lead to efficiencies of scale if such integration can encompass procurement.

2) Integration of institutional facilities with national facilities is ambitious, but is likely to dramatically improve the technical support available to e.g. universities running HPC systems. These systems often suffer from low staffing levels for technical operation, and being able to access a national network of expertise would be highly beneficial.

3) Integrated access across levels of computing could be coupled to Outcome 1 (training). Ideally, users with less proficiency would cut their teeth on smaller systems, ensuring that larger systems are used more efficiently by trained users.

5. Cybersecure, particularly for national-scale data and computing

Answer: Strongly agree.

Comments:

1) It is beyond doubt that Australia’s national infrastructure is at risk of cyber attack. The proposed policies are sensible and proportionate.

6. Maximised by openly available research software tools

Answer: Strongly agree.

1) We strongly support open-science efforts. There is no doubt that open-source software with a wider user base benefits from multiple interactions that improve and further develop the code base.

2) There many examples in Australian physics of open-source codes that are world-leading, and enhance the visibility of Australian science – such initiatives have a key role to play in enhancing Australia's visibility on the world stage. Support for career paths in these areas would be very beneficial – pure software work is not always part of the standard academic track for example.

3) The NDRI strategy should prioritise open-source solutions where they are available. This includes the use of Linux-based operating systems, along with associated training for new users.

4) It seems an oversight not to mention intellectual property rights arising from software developed within the Australian research community, particularly given the ongoing desire to commercialise more Australian research. Perhaps this is a separate outcome that could be added to the strategy – how do we best support commercialisation of Australian software, consistent with the rest of the Outcomes in the document?

NDRI Strategy Overall Response

Answer: Agree

Comments: See the response to Outcome 6 regarding commercialisation – it would be good to consider how to meet the other Outcomes whilst also developing a competitive and rich ecosystem for generating commercial software outcomes. Adding an extra Outcome on this would be good.