Te Ara Paerangi Future Pathways Submission March 2022

Executive summary

We, the undersigned, feel that Data Science is not prioritised in our current research model, and the current consultation process has not sufficiently addressed the importance of Data Science for fast-tracking, and delivering impact from, Aotearoa/New Zealand's research. Data underpins all research carried out in Aotearoa/New Zealand and supporting those that work with data is imperative. The current model limits the potential of researchers to bring cutting-edge solutions to the problems faced by Aotearoa/New Zealand. The result of Aotearoa/New Zealand's current research funding model is that Data Science has been regarded as supporting science and research, rather than a field in its own right. We feel that a national strategy is needed to grow Data Science into a world-class field by supporting infrastructure, capability and best practices in security, diversity inclusion and process design. We propose establishing a hybrid physical/virtual National Centre of Excellence entity to support growing Data Science into a key sector in New Zealand. We firmly believe that Aotearoa/New Zealand's economy will only be able to fully benefit from Data Science research if the research, applications and infrastructure aspects are all supported with an end-to-end focus.

We contend that a data science COE would enable Data Science to thrive in Aotearoa/New Zealand, with a bustling, connected, collaborating workforce that have genuine impacts for national benefit.

Defining "Data Science"

Data science combines multiple fields, including statistics, scientific methods, artificial intelligence, and data analysis, to extract value from data (definition from Oracle). In this submission we extend this definition to also encompass methodologies used to translate research outcomes into tangible products and integration into business operations and the physical IT infrastructure required to support both the research and engineering scopes.

Background

We are a group of researchers and IT professionals working in the Data Science space in CRIs and NeSI. As members of CRIs and NeSI we are keenly aware that Aotearoa/New Zealand has some critical issues that need data solutions. We came together as a group to identify ways to better collaborate and improve best practices across the CRIs and eResearch system. Through these discussions we have identified that we are all independently addressing common issues, and there is a huge opportunity loss by not working together to better address Aotearoa/New Zealand's Data Science needs. We believe that a national strategy for Data Science alongside a National Data Research Centre is needed to focus on collaboration and world-class practices.

Data underpins everything and need research missions in their own right

Digital science is at the core of all science that is done in Aotearoa/New Zealand. Data and digital solutions should not be seen as 'supporting' science, but needs to be a key mission in their own right. Research in Data Science techniques such as AI, machine learning and deep learning (and their application in the New Zealand context) is cutting edge. These techniques unlock the enormous potential in the exponential growth of scientific data, and open doors to new insights, innovation

and impact. It is critical therefore that data science is an efficient, collaborative discipline with sufficient resources to grow Aotearoa/New Zealand's expertise and capacity to meet this opportunity. Data Science missions provide focus on the key contribution and provide a means for the visibility of Data Science as a career.

Challenges

We see much evidence already about the explosion of data streams that are becoming increasingly common in science today. This can range from small scale data acquisition everywhere to large global science technologies. Just as traditional paradigms are being influenced by this data and technology, so too is the growth of dedicated data-led science methods, tools, and infrastructure. Within the Aotearoa/New Zealand context there remains a lack of coordinated effort toward underpinning this critical paradigm shift for the benefit of all society. A fragmented and bottom-up approach whilst enthusiastically supported by active researchers, lacks the critical mass to scale and advance the benefits of data-led methods. To elevate our ambitions from institute and sector to national and international impact we need to consider the funding mechanism, signals, and drivers to achieve this step change. We have not seen in Te Ara Paerangi sufficient or clear acknowledgement of attention towards these aspects. The infrastructure, workforce, capability, and even organisational components don't expressly highlight the aspiration of Aotearoa/New Zealand to achieve a change in vision to a truly digital and data-led research capability. We believe that lack of central focus reduces the impact that Aotearoa/New Zealand researchers would otherwise achieve through innovation, scale, and agility.

Current impediments limit our potential

- Each CRI duplicates efforts in Data Science research and infrastructure, as a data science solution can often be leveraged across a range of diverse sectors
- Funding encourages competition and siloed efforts which does not leverage existing knowledge
- Project funding does not allow for maintenance and continued development of dynamic Data Science solutions
- Contractual funding models create barriers to production of quality research output past minimum viable product and does not incorporate co-design thinking to solve relevant and important pain points for end-users
- Resource funding does not enable a healthy Data Science workforce, or clear career path, as there is little or no:
    - clear picture of what a career in Data Science looks like for the prospective workforce;
    - training for new members of the workforce or upskilling the existing workforce;
    - attraction or retention of talent when competing with significant salaries in private sector and other industries; and
    - appropriate success metrics and rewards to retain staff.

Collections are there to be used

The databases and collections that maintain our national and local scale data should not just be about holding information. They should be focussed on the future value and insights gained from their application towards our biggest challenges. Support of databases and collections needs to be forward-thinking about this need. Enabling creative ideas and data-led research that extracts maximum knowledge from these national datasets would open new ways of seeing and understanding these valuable underpinning resources. Models, research software, and the capability based around collections and national datasets are also as much a national resource as is the underpinning information they contain. These components need to be supported, curated, and made accessible in conjunction with the changing paradigm of data-led discovery. Future funding models and/or support for data collections, their codes, and tools will need to consider the needs (current and future) of a Data Science community. For this reason, there needs to be a strong linkage between national data and a national focus on supporting and enhancing Data Science capability. This is a very different funding model than simply funding to maintain a national dataset; this is a funding model which recognises and rewards innovation.

Our vision

Our overarching vision is for Data Science to be a major tool for good in Aotearoa/New Zealand:

- Data Science is embedded into the fabric of Aotearoa/New Zealand, and is harnessed to solve wicked problems, increase productivity, protect and enhance our natural resources and improve our social wellbeing.
- Data Science expertise, tools, and infrastructure are scaled-up and shared openly across the entire RSI, and more widely via international collaborations, to accelerate capability and multiply the effects of all Data Science research and development.
- Aotearoa/New Zealand's research, science and innovation system experiences a step change in impact by using Data Science to extract maximum value and knowledge from research efforts and artefacts, including our national databases and collections.
- A National Centre of Excellence that federates investment into operational Data Science capability spread out across organisations in the RSI ecosystem.
- A nationally governed and co-designed supporting physical IT infrastructure (including High-Performance Computing and research data platforms) and the capability to maintain it, and build skills and capabilities. This should operate in strong partnership with the data science community, and through co-design approaches that invest in fit for purpose infrastructure and data science platforms.

Data Science and Artificial Intelligence are often used interchangeably. Data Science is broader than Artificial Intelligence, but they have similar goals. The following specific goals are adapted from the NZ Artificial Intelligence Researchers Association whitepaper "Aotearoa New Zealand Artificial Intelligence: A Strategic Approach":

- Leading edge Data Science research and capability is harnessed quickly and effectively to help solve Aotearoa/New Zealand's most pressing problems, such as climate change, child poverty, equity, disaster response including natural disasters and pandemics.
- Productivity growth is enhanced by enabling private sector adoption of leading data science technologies.
- Government leads the adoption of Data Science in administration, healthcare, infrastructure and regulation in ways that give effect to Te Tiriti o Waitangi.
- Data Science is used to promote socially inclusive growth and encourage a Data Science community that is inclusive of diverse backgrounds and perspectives.
- Data Science use is ethical and culturally appropriate.
- Data Science is a well understood and desirable career pathway.

This vision is supported by:

- A focussed Data Science centre for research, development, and investment that operates across all stakeholders including CRIs, universities, iwi, industry, local and central government, and the community.
- An effective national data infrastructure with open data partnerships and datasets, while enabling and supporting Māori data sovereignty obligations, as well as commitments to create test environments and regulatory sandboxes.
- A compelling research ecosystem with the capacity to attract, retain, and train domestic and international Data Science talent.

A national Data Strategy is required

Data Science has been shown to be a force multiplier for great science, and for this to push Aotearoa/New Zealand forward we need to enable data science nationally. Realising our potential from Data Science requires a national data strategy, partnered with Māori from the beginning. Data is a taonga and therefore Māori have tino rangatiratanga over data. Thus decisions about data and data infrastructure must be made with Māori. There are many interconnected areas of work carried out nationally that need to be addressed in the proposed strategy to enable a way forward. Inclusion of best practice in Māori Data Sovereignty, Data Security and Privacy (e.g., ISO27018), data access and integration, interoperability, data ownership and the rights to use and reuse data and models. Solving this at the national level would amalgamate knowledge and enable its continuous and sustainable development.

Data is integral to all research projects across the science and research sector and Data Science tools have become requirements to gain the most insight from it. Our CRI Data Scientists meet this need, developing innovative solutions to a diverse range of applications including public health, climate change, disaster resilience, our natural environment and primary production across public health, safety, the environment and primary production sectors. Whilst advanced off-the-shelf software and solutions are increasingly being made available from large organisations such as Google and Microsoft, it requires substantial effort and innovation to apply them to the Aotearoa/New Zealand context, as well as specialist skills to consume those services in a cost-effective and secure manner.

A coordinated National Centre for Data Science

We advocate for the establishment of a Centre of Excellence-type (CoE) organisation to bring focus on the applications of Data Science research to deliver impact to New Zealand industries and to support the needs of New Zealand Data Science researchers.

Our vision is to bring together capability in Data Science to reduce redundancy, share knowledge, and bring together developers of Data Science products to allow for co-ordinated efforts across multiple applications. This would enable:

- a cohesive community and network of researchers, connecting the pipeline of talent into impact
- Data Science would be a career path in its own right
- The sector would be skilled enough in basic Data Science understanding to leverage a Data Science approach for the benefit of NZ
- The network of researchers would enable a shared view of what groups are working on, reduced redundancy
- A front door for a Data Science conversation

This is not a proposal to bring all forms of Data Science under one roof, but to highlight the need for a point of focus, a champion at the national scale for all researchers, and an elevator of a national conversation about the fourth-paradigm: Data-intensive Scientific Discovery.

We strongly support national infrastructure for Data Science, but any infrastructure needs to be positioned alongside researchers and have a strong focus on capability development and co-developed with Māori. The work NeSI has done in enabling trainers for software carpentry is an excellent example of integration of capability with infrastructure. In addition, suitable career pathways need to be defined and supported for those operating in the digital technology space, who do not currently have a clear way forward in traditional researcher roles. Global movements such as the Society of Research Software Engineering attempt to fill this gap by explicitly acknowledging and recognising the vital role of software in research and people in research software. Recognition is growing that a researcher needs to have knowledge in Data Science and research software engineering practices as well as traditional research skills. Any capability development needs to cover all of these skill sets.

We aspire to these skills being applied with an entrepreneurial confidence and culture. We're seeking a common approach within the sector of how to embrace and manage high failure rates in translating deep technologies into impact. Focusing on the culture and skills to enable this, a source of inspiration is an established programme in the United States based on lean start-up - a

programme initiated in 2011 by the National Science Foundation called i-Corps which now has 99 institutions involved nationwide, and operates in a hub model as an innovation network. This programme has been replicated into other agencies, including i-Corps at NIH (National Institutes of Health), and i-Corps at the Department of Defense. This proven approach offers significant untapped potential to our research and innovation system, across a broad range of technologies.

Alongside capability development and co-design of any infrastructure with researchers and Māori, there need to be incentives built into the system to support research teams and institutions to consume that infrastructure at scale. Many of the existing capabilities of teams and of infrastructures are currently highly fragmented, with investment following along local lines. This could be balanced by incentives to coordinate investments into shared capabilities and assets where it makes sense. An example of this comes from the Canadian Digital Alliance, where their Contributed Systems policy has a range of incentives to encourage national coordination of underlying enabling infrastructure.

Bringing Aotearoa/New Zealand Data Science practitioners into a Centre of Excellence would reduce duplication and multiply the impact of Data Science research, while retaining connectivity to the rest of the RSI system by being the "glue" that joins Data Science researchers together and facilitates the flow of ideas. This Centre of Excellence should be closely aligned to the national Data Science infrastructure, so that it supports the development and delivery of sophisticated tools for extracting knowledge from data in the New Zealand context.

There are models elsewhere that can be looked towards for pathways NZ might take. Some current exemplars include:

- **Australian Research Data Commons** - an inclusive national research data commons incorporating discipline specific research data platform capability building including for nationally significant databases. ARDC supports the research community and industry access to nationally significant, data intensive digital research infrastructure, platforms, skills and collections of high quality data.
- **Canadian Digital Research Alliance** - a broadly integrated national computational and data platform inclusive of skills and capability building
- **Alan Turing Institute** - a hybrid virtual / physical data science institute focused on skills and capability building
- **Data 61** - a research institution that seeks to solve Australia's data-driven challenges through the maintenance of strong industry relationships, using an industry focused approach to translate research data science into commercial outcomes

Institutional barriers should be minimised

We believe that there should be mechanisms to foster mobility of Data Scientists between universities, CRIs, industry, iwi corporations and community groups. This would enable better understanding of industry and community pull for research needs, and better delivery of impact. A

Centre of Excellence would support stronger collaborations and relationships through a thriving community of practice. It would also contribute to reducing complexity in existing remuneration mechanisms by explicitly recognising the hybrid “research plus IT” nature of the roles involved.

Any solution must be structured to be inclusive of any organisations and centres with a Data Science interest or undertaking Data Science work now (and in the future), allowing organisations to participate easily.

Base funding allows global impact

A level of base funding seems appropriate for developing and maintaining Data Science tools and software tailored to the Aotearoa/New Zealand context. There is already a vast trove of such research project artefacts across Aotearoa/New Zealand’s science communities, yet only a very few are supported to grow and mature into useful products under the current settings (R, Weka, Cylc, STRMix, Beast). Data Science research impact is strongly multiplied when developed tools and techniques can be generalised or modularised for reuse across other domains and data sources.

Research advancements in Data Science are moving at a lightning fast pace on the global research arena. A base-funding model would provide a means for researchers to excel in the global platform in data science research and deliver impacts to Aotearoa/New Zealand, by enabling:

- Aotearoa/New Zealand Data Science researchers to develop and test new ideas quickly
- Support the development, maintenance and enhancement of tools and techniques that can be generalised across other domains
- Support delivery of impact from research to product after the life of a funded research project or programme to allow data science missions to deliver on their promised impacts after the research phase
- The impacts of non-traditional outputs such as research software and code are recognised as a first class research output
- Support early career Data Science researchers in growing and establishing themselves in the competitive research field
- Support work where a solution is not clear at the outset and is more suitable to a co-designed and agile-development framework. This contrasts with the current model which is based on a possible marketable solution at the onset of the project.

Signatories:

Alan Tan, NIWA - National Institute of Water and Atmospheric Research

Brent Martin, MWLR - Manaaki Whenua Landcare Research

Claire Stewart, Scion

Elizaveta Graevskaya, Scion

Eric Burgueño, PFR - The New Zealand Institute for Plant & Food Research, Ltd.

Georgina Rae, NeSI - New Zealand eScience Infrastructure, the University of Auckland

Janet Stacey, ESR - The Institute of Environmental Sciences and Research Ltd.

Linley Jesson, PFR - The New Zealand Institute for Plant & Food Research, Ltd.

Nauman Maqbool, AgResearch

Nick Jones, NeSI - New Zealand eScience Infrastructure, the University of Auckland

Nick Spencer, MWLR - Manaaki Whenua Landcare Research

Peter Green, AgResearch

This submission represents our views as individual researchers, and may not necessarily reflect the view of our employer institutions.