Australian Government
Department of Education

NDRI Investment Plan Consultation Survey Summary

Compute and AI

Q9 - What are the current gaps and weaknesses in Australia's NDRI technological capabilities?	• Lack of research software engineer (RSE) workforce and long-term funding for humanities and social sciences to enhance sustainability. • Cumbersome process for accessing high performance compute (HPC) resources. • No Tier 0 facilities for Australian researchers to access. • Lack of collaboration with the US and EU. • Significant weakness in access to large scale private compute for AI research in healthcare. o There's a limited understanding of AI technology as applied to healthcare research. • Lack of AI-focussed expertise and experience that intersects with the National Research Infrastructure (NRI) workforce. • Skills gap in digital literacy and data management among researchers. • Insufficient access to open-access peak HPC resources comparable to peer countries. • Inconsistent data management practices and standards across institutions. • Inadequacy in existing cybersecurity and data privacy protection measures. • In current funding and merit schemes for computing infrastructure, there are vast gaps between research institutions. • Lack of an entity that can provide a leadership role in defining a national strategy for computing grand challenges. • No incentive mechanism to facilitate the sharing of HPC capabilities. • Lack of visibility and awareness about the importance, scale, sophistication and long-term sustaining capabilities being provided by institutional research data repositories. • Non-availability of easily accessible and affordable storage and compute for smaller institutions who don't have their own infrastructure and human resources to provide support. • Multiple HPC providers across the Australian research sector, with each provider employing different workflows and access processes. • A fragmented user experience for researchers with different identities, access management approaches, and workflows in use at each HPC. • Current NCRIS-funded supercomputers lack specialized systems required for secure handling of sensitive health data. • Lack of sufficient GPU based computing to enable training of AI based weather, climate and ocean models. • Insufficient tools/processes to transform "not-research ready" data into "research ready" datasets suitable for robust analysis. Lack of cross-team/institution collaboration for easier data access. • Absence of a national repository for research data archiving and rewards for properly constructing reusable datasets. • Lack of integration across Tier-1 and Tier-2 computing facilities. • The national computational resources that are available to researchers are split between supercomputers (NCI & Pawsey) and generic cloud resources (Nectar).

Q10 - What should Australia's supercomputing/high-performance compute landscape look like in 5 years from now?	• Cost-effective investment would be allocating more of the available computing time to open merit-based processes to improve return of investment. • Aligning with international standards, enabling equal access to computing resources for all researchers regardless of institutions. • Availability of more user-friendly options. • Federated AI and compute capability, with HPC nodes embedded within all research settings. • Availability of an onboarding and training pipeline for staff to learn about and get access to compute. • Investment in GPU support for CPU-based software and workflows. • Continued investment in energy-efficient technologies to ensure that HPC systems are sustainable and environmentally friendly. • Improved access to HPC resources for researchers across all disciplines, including humanities and social sciences. • Development of integrated platforms that allow seamless data sharing and collaboration across different research institutions and disciplines. • Robust cybersecurity measures to protect sensitive data and ensure the integrity of computational research. • Ensuring that HPC investments respect Aboriginal and Torres Strait Islander people data sovereignty principles, providing their communities with control over their data and its use in research. • An aligned/common governance model for Australia HPC as a whole. • Common entry method for seamless transition for users between Tier-1 facilities. • Increased availability of Tier-2 HPC resources accessible to a wider range of researchers, complementing Tier-1 national systems like NCI and Pawsey. • Augmented Tier-1 HPC capabilities through international partnerships and compute resource sharing agreements with the US and EU coupled with exabyte-scale data systems and collaborative software environments operating as an integrated "laboratory" for sharing tools, data, and methods. • Specialised large-scale clusters optimised for specific domains/architectures developing alongside flagship systems. • Significant growth in access to GPU nodes, essential for processing large datasets and running complex simulations, will be vital to support the diverse needs of the research community. • A robust and scalable HPC framework that includes broad Tier-2 capacity and strategically developed Tier- 1 compute resources to accommodate growing demand, including to support exascale computing.

Q11 - To what extent should NCRIS investments extend to Tier-2? What would make these "collaborative" and "national"?	• To support and encourage a federated compute and AI capability. This can include flexible models of technical, scientific, and health governance in response to local capability. • Project-based support for access to Tier-2 capabilities, especially for projects that would not make full use of Tier-1 facilities. o Enabled by a model akin to a voucher access scheme, which is responsive to demand, without allocating scarce long-term operational funding. • Tier-2 facilities should continue to be funded, and the number of such facilities should be increased. o These new facilities should be geographically distributed and could be optimised for discipline specific cases (e.g. GPU clusters for AI/ML processing). These play a critical role in training higher degree researchers (including students) and early career researchers in HPC and data. • To provide more researchers, especially those in smaller institutions and diverse disciplines, with access to HPC resources. o This democratizes access to advanced computational tools, fostering innovation across the board. • To support regional and smaller institutions, ensuring that research capabilities are not concentrated in major cities alone. o This promotes equitable development and utilization of research infrastructure across the country. • NDRI investment into Tier-2 computing needs to be accompanied with institutional accountabilities. • NCRIS investments can be made “collaborative and “national” through: o Integration under a single strategy and oversight structure for Australian HPC. o Clearly defined purpose of NCRIS investment. o Focussed investment to specialise rather than duplicate efforts. • Directing NCRIS investment into hardware will facilitate long-term planning for Tier-2 and support for personnel. It will drive sharing of expertise across facilities to create a national network. • A merit-based national access model similar to NCI would be required, rather than allocation tied to institutional investment. • A focused training program is essential to equip users with the skills needed to effectively leverage Tier-2 facilities. o This program should also facilitate interactions with Tier-1 facilities, ensuring that researchers and industry professionals can fully utilise these advanced computing resources.

Q12 - To what extent should future investments consider AI sympathetic architecture?	• Investment into architectures that can be dual use for both AI and non-AI applications to maximise flexibility. • Future investments (at least at the Tier-2 level) should be towards new computers with AI sympathetic hardware. ○ Matched by investment in AI specialists who can support researchers to take advantage of these powerful techniques. • Investments should focus on architectures such as: ○ Allocating resources to handle varying AI demands, ensuring efficient use of computational power. ○ Energy-efficient architectures, such as those using advanced cooling systems and renewable energy sources, will help meet sustainability goals and reduce operational costs. ○ Robust data management and storage solutions that ensure data is easily accessible, secure, and compliant with FAIR principles. • Patient-centred Care: AI can help design systems that prioritise patient needs and experiences. • NDRI investment in AI-sympathetic hardware needs critical assessment in prioritisation, taking into consideration other national funding streams such as the Australian Research Council's Linkage Infrastructure, Equipment and Facilities (LIEF) scheme and the Australian Government Department of Health and Aged Care's Medical Research Future Fund (MRFF). • Investment in the upskilling of users & growing the skilled software engineering expertise needed to realise the potential of exascale computing as well as expanding access to AI architectures. • Investment in international partnerships or partner with companies like Google to resource foundational large language model research.

Q13 - How can Australia sustainably invest in new NDRI technologies that align with the Australian Government's Net Zero greenhouse gas emission target by 2050?	• Investment in energy-efficient data centres and HPC systems. This includes using advanced cooling technologies, renewable energy sources, and energy-efficient hardware to reduce the carbon footprint of digital research infrastructure. • Ensuring that any design of new hard infrastructure takes full account of current and future sustainability requirements at the design phase. Post-build modification is extremely expensive. • Reducing the demand for compute resources by supporting efficient code writing. Efficient code uses less power. • Ensuring significant compute resources are powered via renewable resources (for example, through power purchase agreements). • Supporting the construction of efficient compute facilities. • Exploring the use of compute facilities as a demand management tool (for example, adjusting compute speed according to time of day and/or percentage renewable power). • Investment in research and development of sustainable technologies and practices. This includes developing new methods for reducing energy consumption in data processing and storage. • Adopting sustainable procurement practices for NDRI technologies. This involves: ○ Selecting vendors and products that prioritise sustainability, energy efficiency, and minimal environmental impact. • Implementing lifecycle management practices for NDRI technologies to ensure that equipment is used efficiently and recycled responsibly at the end of its life. This reduces electronic waste and promotes a circular economy. • Investment in skilled supercomputing specialists who can assist users in fine-tuning workflows and algorithms to efficiently utilise cutting edge supercomputing resources. • Investments in developing robust energy consumption monitoring and control systems at Tier-1 facilities. • Integrating carbon targets into infrastructure investment guidelines and NCRIS reporting processes.

Q14 - What are the priority NDRI investments in compute and AI that would enhance Australia's research efforts?	• Establishing quality data assets and making it accessible. • Developing capability and enabling outreach from existing centres of technical and intersectional expertise that enable federated AI and compute platforms. • Investment in: o GPU architectures and in the specialists (data, software, and platform) that can help to migrate the research community's codebase to take advantage of these architectures. o Training to ensure researchers are trained in best practice coding, and facilitate collaborations between research teams and data, software, platform specialists. o Science platforms that bring data and compute into a common hardware and software infrastructure, making deployment of AI driven analysis and curation as straightforward as possible. • Expanding and upgrading HPC infrastructure to support large-scale simulations, complex data analyses, and AI model training. • Developing and integrating cloud computing services that provide scalable and flexible resources for researchers. • Investment in robust data management and storage solutions to handle the vast amounts of data generated by AI and computational research. • Enhancing cybersecurity measures to protect sensitive research data and ensure the integrity of AI models. • Investment in NRI capability to establish foundational AI/ML models as operational research infrastructure, commencing with a use case around biodiversity data synthesis and ecosystem prediction. • Developing a sound data and compute infrastructure capability for biobanking for clinical data and environmental and resources data. • Investment needs to address the trust and identity requirements of a heterogenous computing environment and associated storage platforms (local or multi-site) across both Tier-1 and -2 facilities through a system-wide approach: o interoperable access o identity assurance o strong authentication o open and trusted access. • Leveraging AI for data cleaning, transformation, error-checking in analyses, and automating repetitive tasks like optimal character recognition, speech-to-text (across languages), and video recognition. • Establishing a national curated data repository with pathways from various disciplines, tools for proper record creation, and adopting systems. • Investing in enabling Australian researchers to actively participate and collaborate with international research groups working on pre-exascale and exascale infrastructures to facilitate modern data standards adoption.

	• A coordinated approach across NCRIS projects to set up an image data analysis hub that develops AI-based tools will uplift researchers and keep them abreast with best practices and applications of AI in imaging data. • Investing in domain-specific AI applications: ○ A balanced investment strategy should support both mature AI and early development initiatives, to allow research domains to evolve according to their needs.