C1.1 ARC Training Centre in Data Analytics for Resources and Environments (DARE)

C1.2 Aims and Background

The aim of the ITTC for Data Analytics in Resources and Environment (DARE) is to develop and deliver the skills and tools for Australia's resources industries: to be expert users of data and models; to quantify, explain and understand uncertainty; and to make the best possible evidence-based decisions in exploiting and stewarding the nation's natural resources and environment.

The resources industries – minerals, energy, and water in particular – are experiencing an unparalleled growth in the quantity and variety of data that is being made available to make informed decisions. However, natural resource environments are hugely complex: highly coupled and integrated, with data that is often uncertain and ambiguous, and where decisions are usually a balance of difficult compromises. Superficial use of conventional "Big Data" methods are wholly inadequate and often dangerously misleading as they can create a false sense of security. With a deep understanding of modern statistics and machine learning methods, and approaches to capturing uncertainty and ambiguity in data and models, more balanced trade-offs of cumulative risks can be defined [1]. These skills and tools – delivered by DARE – are the keys to understanding, modelling and managing complex integrative natural resource systems and ultimately delivering on sustainable evidence-based economic and societal outcomes for the resources industry.

DARE will deliver foundational skills spanning three of the six industrial transformation priorities – Mining Equipment and Technology (METS), Food and Agribusiness (FIAL), and Oil, Gas and Energy Resources (NERA). All of these priority areas are facing the common central issue of how data is exploited to build predictive and integrated resource models which can be used to make economic and sustainable decisions in the face of high levels of uncertainty. Modern data science methods, created through the convergence of mathematical, statistical, and computing disciplines, will underpin a profound change in all of these industry priority areas, creating huge opportunities for improved efficiency and competitiveness, and also for better stewardship and sustainability of our natural resources.

The natural environment is characterised by many and varied stakeholders, all with different priorities and values. Resources are finite, and stakeholders compete for access with vastly different outcomes. For example, the economic return per gigalitre of water used in mining is some fifty times that in agriculture and yet we need both for a sustainable economy. DARE uses data as a common language to build an exciting, truly multi-disciplinary partnership between industry, government, leading educators and researchers in data science, and environmental sciences. Together DARE's academic, industry and governments partners will drive transformational advances in the development and application of data science, building scale in training and human capacity, and will deliver translational outcomes for Australia's resources and environment industries.

World-wide there is a growing recognition that the new methodologies and techniques being developed within the discipline of data science are key strategic technologies with the potential to revolutionise the natural resources sector. Globally there is an emerging shift from "big data" to informed data science with a number of world leading institutions being established. In the UK the Alan Turing Institute (ATI) has been set up as the national data science institute, containing 13 university members. In the US, the University of Washington, the University of Michigan and Columbia University, have recognised this new opportunity and have established data science institutes focusing, in part, on natural environments.

In Australia there is a special opportunity, given the strength and reliance of the Australian economy on the resource sector. DARE has engaged broadly with the three Federal Government Growth Centres in resources (METS, FIAL, NERA), and a key range of industry partners in areas covering minerals, water and agriculture, and with a range of government agencies whose role is to manage, sustain and make best use of these resources. DARE will build on the need of all of these partners and will deliver the necessary skills and capabilities in modern data science to this community. DARE will achieve this by

- Developing Australia's first international cohort-based doctoral and postdoctoral training programme in data science methodologies to model and quantify uncertainty in the context of the cumulative impact of decisions in the natural resources.
- Developing a strategy for equality, diversity, and inclusion in order to attract and support the most able students from the widest talent pool possible, and to remove barriers in the 'pipeline' to anyone that wants to pursue a career in data science in the natural resources sector.
- Training students in a broad range of technical skills covering: computational analytics, Bayesian methods, modelling and inference, and machine learning, at the interface between natural sciences, computing, mathematics, and statistics so that they can address a diverse set of multi-disciplinary challenges in an increasingly complex, data connected, digital world.

- Training students in transferable and professional skills and endowing them with attributes that will make them highly-employable, and easily-deployable across multiple divisions of the resources sector. This will be achieved by embedding Higher Degree by Research (HDR) students within our diverse range of industrial and government collaborators, fostering opportunities for HDR candidates and postdoctoral fellows to pursue industrial training.
- Providing advice on industrial needs in terms of student training and research challenges, co-create training provision, sponsor and co-supervise student projects, and provide opportunities for internships and secondments.

C1.3 PROPOSED PROJECT QUALITY AND INNOVATION

DARE will create a unique world-leading Higher Research Degree training (HDR) hub addressing the urgent need the resource sector has for skills and deep expertise in data science. DARE will produce graduates that are able to innovate, operate and lead at the interface between natural resources and the data sciences: mathematics, statistics, complexity and computing. DARE will produce graduates that think probabilistically – that can quantify and understand data uncertainty, that can use data to build and maintain models of natural resource systems, and that can use probabilistic models to deliver economically and sustainably effective decision making. DARE students will be exposed to the multidisciplinary nature of the natural resources sector and will be mentored by academic, industry and government partners in trans-disciplinary application of deep data science skills to tackle emerging and contemporary Australian and global challenges facing the allocation of our natural resources.

DARE has engaged with the relevant Industry Growth Centre leads to build a cohort of industry and government that is representative of the full breadth of challenges in the natural resources environment. DARE graduates will inherit this trans-disciplinary DNA – they will be co-supervised by both resources experts and by leading data scientists, and embedded in industry and government. DARE graduates will bring critical new skills and drive new data-driven approaches to innovation across the minerals, energy and food sectors.

C1.3.2 Conceptual Framework DARE has identified three resource exemplars in **Minerals, Water** and **Biodiversity**, and three data science capability areas **in Exploring Data, Building Models** and **Making Decisions** (Figure 1). The resource exemplars focus on key challenges for the ITPs where data science will have major impacts on competitiveness (Table 1). In each exemplar, DARE has built a partnership with key industry and government players and has supported this with leading academic expertise in the exemplar area. The three data science capability areas bring together outstanding capabilities in advanced statistics, machine learning and decision sciences with a long track record of innovation in teaching and training across all levels.

Top students will be recruited at the intersection of data science and domain science. DARE will use courses developed for core practice skills as the basis for a data science PhD training program. Additionally, students will be guided through key advanced courses in their chosen domain area. Uniquely, all graduates affiliated with DARE will thus receive top training and skills in both data science and domain science – regardless of ultimate discipline specialisation – and this will embed the principles of trans-disciplinary data science in future generations of researchers.

Resource Exemplars. Studying the three resource exemplars jointly is crucial. It highlights that an action taken to maximise some payoff function in one exemplar has consequences for the payoff function in another, thereby explicitly encoding the cumulative impact of an action. Additionally, although different in surface structure, the three exemplars share a need for probabilistic thinking and proper uncertainty quantification for optimal decision making.

Minerals Exemplar: Mineral discovery underpins large parts of the Australian economy where mining contributes over $60bn a year to national GDP. However, the Australian Identified Mineral Resources 2013 report [2] shows that there have been very few large-scale mineral discoveries in Australia in the past two decades and most of Australia's current mineral production and exports are sourced from deposits discovered during an exploration high more than two decades ago. The grand data science challenge is to exploit the vast amount of minerals exploration data to discover what mineralogy exists beneath the 80% of Australia where favourable geology lies below regolith or other barren cover. DARE will build the new skills and capabilities necessary to meet this challenge: in managing the vast amount of available geophysical information, in developing methods that can fuse this data in to uncertainty-quantified geological models, and in developing evidence-based approaches to characterisation of mineralisation.

Figure 1 Conceptual Framework

The minerals exemplar will be led by CI Lindsay, an ECR and DECRA holder, and an expert in 3D geological modelling. He will be supported by CI Cripps, S. Centre Director and expert in advanced Bayesian and MCMC methods applied to geology, and who has established a strong relationship with key mining industry partner Newcrest. Additional CIs include Ramos and Scalzo, with expertise in non-parametric and sample-based methods applied to minerology and mining, and CI Jessell, with expertise in geological modelling. Training, mentoring and capability development in the mineral's exemplar will be undertaken with key DARE partners including Newcrest, McKinsey, Natural Resources Commission, and Geoscience Australia.

Water Exemplar: Water is a fundamental resource, vital to the genesis and sustainability of communities, ecosystems and industry. Understanding the drivers behind water supply and usage and quantifying the joint uncertainty around supply, and demand is critical to many areas of the Australian economy.

The foremost challenge in water resource management today is to make uncertainty quantified predictions in a changing environment for applications such as ecosystem management, flood warning, and water allocation. This is a difficult task when one considers the uncertainty associated with modelling daily rainfall, hydrologic observations, and future climate.

The water exemplar will be led by CI Marshall, a leading water domain expert and DARE Deputy Director. She will be supported by CI Cripps E., a data scientist with a strong track record for developing statistical models for rainfall. Additional CIs contributing to this exemplar include domain specialists Johnston and Vervoort, and data scientists Cripps, S. Tanaka and Chandra. DARE will draw on deep expertise in water, hydrology and risk from key partners including IAG, McKinsey, Water NSW, Natural Resources Commission, Geoscience Australia, NSW OEH and DAC.

Biodiversity Exemplar: Biodiversity, the diversity of livings things on Earth, underpins and influences almost every product and service we value today and is essential to provide future generations of Australians with a healthy, sustainable economy and environment. Decisions for the management of biodiversity, the ability to prioritise actions and policy, must be informed by complex models based on relatively sparse measurement data where uncertainty quantification is key to decision making.

The biodiversity exemplar will be led by CI Wardle, a leading ecologist with extensive experience in uncertainty quantification. She will be supported by CI Tao, an Australian Laureate Fellow and world leader in machine learning and data mining. Additional CIs contributing to this exemplar include domain specialists Webster, with expertise in reef formation modelling and Greenville, with expertise in technology for ecology and environmental sciences, and data scientists Cripps E., Scalzo, and Azizi. DARE will draw on expertise from key partners including IAG, Natural Resources Commission, Water NSW, WABSI ,OEH, and DAC.

Data Science Capability Areas The three data science capability areas address common data science challenges, the key skills and capabilities identified by industry and government partners in the natural resources sector (Table 2). Fundamentally, these are concerned with probabilistic understanding of uncertainty – in data exploration, in model building and decision making. These are the core skills and training to be delivered by DARE. The simple conceptual sequence of the three themes represents a logically consistent and principled way of structuring a complex business problem. The structure forces the decision maker to think (i) What data do I have? and, how reliable is it? What do I already know, or believe I know, about the issue? (ii) How can I combine all the information in a sensible manner to predict uncertain outcomes and attach probabilities to those outcomes and update these predictions and probabilities as more information comes to hand? and (iii) given these outcomes and probabilities how do I make a decision considering different constraints and objectives. These are skills that the entire resource industry workforce needs.

1. Exploring data. This capability area aims to impart an appreciation among industry for the complexities and uncertainties of real data, and to develop student training and skills for quality control and extraction of features from data anterior to the main statistical modelling workflow. We have identified 3 concepts that need to be understood and addressed in this capability.

Concept 1(a) Data Visualisation and Communication: Visualisation and communication are fundamentals of data science best-practice, threading through the entire life cycle of analytics problem solving. A clear initial understanding of the problem to be solved, along with a sound assessment of the volume, kind, and quality of data available to solve that problem, begins with clear communication with stakeholders and with preliminary visualisations of the data. Interactive exploration of a model and its predictions can be used to refine the model iteratively. Finally, once a model is built to address the problem, querying the model for solutions and communicating their economic value to scientific colleagues, industry partners, policy makers, and the public relies on visualisation.

Concept 1(b) Data Reliability: Variation and uncertainty in data arises from many sources: sampling error, missing and incomplete data, treating interpolated data as if they were *ab initio* observations, data integration and data pre-processing. Common practice in the exemplar areas is that the uncertainty attached to data is at best underestimated and at worst completely ignored. Models that use these data naively produce biased and overconfident predictions, resulting in lost revenue for industry and poor planning for government.

DARE will develop new deep learning techniques (the training and deployment of neural networks with many hidden layers) for more accurate classification [3]. While traditional methods usually rely on features drawn from domain knowledge, deep neural networks are able to directly learn discriminative features or compressed representations of rich high-dimensional data. The uncertainty surrounding this classification will then be modelled by embedding deep neural networks within a probabilistic framework and estimated using variational techniques.

Concept 1(c) Data Granularity: The massive increase in data availability, attributable to the relative ease of data capture with remote sensing technology, and the inexpensive data storage, has not provided the expected insights into geological, ecological and hydrological processes that might have been anticipated. This is in part attributable to a lack of knowledge and training in developing models which can incorporate data with varying levels of granularity. The Bayesian paradigm represents a vehicle for doing this. Different temporal scales can be represented in a Bayesian hierarchical model where the different levels correspond to different scales. Furthermore, temporal scales and spatial scales are likely to be dependent. This dependency can be accommodated by having parameters at each level in the temporal hierarchy having spatially varying priors.[4] There is an opportunity to develop new techniques for incorporating data of different granularity and assess how this adds to our understanding the availability of Australia's natural resources

2 Building Models: The second objective is to develop skills in building probabilistic models to fuse all available information, and to **quantify uncertainty over possible models** which explain the data. Four key sub-themes relate to models of natural resource systems: the development of models making predictions across space and time (*spatiotemporal* models); the elicitation and encoding of prior knowledge in terms of probability distributions; the formulation of models that can make robust predictions when data are missing or aggregated; the capacity to include known deterministic behaviour of model subsystems either by embedding in likelihood functions or as priors in Bayesian modelling.

Concept 2(a) Models for Extremes: Natural Perils occur at the intersection of a number of unlikely conditions, and modelling this is difficult because, by definition of an extreme, data is sparse. This concept explores the use of time-varying copula models for multivariate extreme distributions for natural peril prediction [7]. In addition, this concept will explore how decisions based on these models change as we alter the definition of an extreme event. The definition of an extreme event for the farming or mining community will differ from the statistical definition of an extreme. The farming and mining communities think of an extreme event as one which will result in a change of management practice or a significant change in production. This in turn will be a function of several variables such as location, vegetation, water-table level and land-use. DARE will establish definitions for these critical levels at various locations and how these critical levels these change across time. DARE will extend the work of [4] and assess the usefulness of long-range climate indicators to forecast the probability of an extreme event, and the extent to which any increase is due to human activity.

Concept 2(b) Models for Space and Time: DARE will develop capabilities in nonparametric models for phenomena that evolve through space and time. For example, in the mineral's exemplar, a key goal is to find similar, but not necessarily adjacent, geological regions, which are defined by properties such as density and porosity, and where the boundaries of these regions are possibly sharp. One approach is to develop priors over random partitions, so that the posterior distributions of these partitions can admit, possibly non-linear, sharp or slowly varying boundaries. Other data is based on repeated measures over time, the goal will be to develop flexible trajectory models for multivariate longitudinal data.

Concept 2(c) Model for Priors and Missing Data: Expert knowledge is an essential component of any model for the solution of complex problems with sparse data. The experience of the specialist in understanding the natural processes helps to reduce ambiguity in the solution and account for information not directly observable. However, experts frequently have very different interpretations for the same set of existing measurements. DARE will use novel Bayesian expert crowd sourcing and prior elicitation methodologies to capture the uncertainty surrounding expert opinion. Missing observations, or incomplete multivariate observations in longitudinal data, are common in natural resource datasets. In a Bayesian framework, the missing observations become part of the unknown set of model parameters, and the uncertainty associated with them is accounted for by integrating over all those possible values of those missing observations with respect to their posterior distribution. DARE will use prior elicitation techniques to develop new models of missingness when the data are not missing at random.

Concept 2(d) Model Fusing and Verification: Model Fusing and Verification addresses two fundamental issues; how to fuse deterministic and probabilistic models and how to estimate uncertainties associated with models; whether that model is probabilistic, deterministic or a hybrid of the two. We call this model verification.

Almost all exemplars have models for physical systems; the formation of reservoirs, the evolution of landscapes, the growth of coral reefs, the formation of mineral deposits, and the effects of various activities on groundwater. These models are among the most important and specific types of expert knowledge available about natural resources. Bayesian inference provides a framework in which the simulated output from these models can be compared with actual data to (i) obtain probability distributions over the initial conditions, (ii) to attach probabilities to the mathematical models thereby verifying the models themselves. Model verification is absolutely critical to establish trust between different stakeholders in natural resources, and to (iii) create hybrid, custom-built statistical and mathematical models to answer questions of deep importance for our industry and government partners.

This area presents some of the most challenging problems in data science. The posterior distributions are extremely difficult to sample from using existing Markov chain Monte Carlo (MCMC) methods [6] and are not well approximated by variational inference techniques. Additionally, the complexity of the physical models makes techniques such as MCMC computationally infeasible.

3 Making Decisions A decision to permit exploration and development may bring economic prosperity but adverse environmental or social consequences. Equally the decision to prevent exploration may have detrimental economic and developmental consequences. DARE's third objective will develop and train PhD students and industry decision makers to create a framework for evaluating competing criteria in a formal manner, for quantifying outcomes and attach probability distributions to these outcomes. Most existing multi-criteria decision making (MCDM) methods focus on decisions when the multiple outcomes are assumed certain. The uncertainty surrounding the outcomes as well as the uncertainty surrounding the relationship between the outcomes is mainly ignored.

Concept 3 (a) Adaptive Decisions: What is the next piece of information that should be acquired to reduce uncertainty while maximizing some objective function? And how do I measure that uncertainty? A mining company might seek to maximize profit; a government agency in charge of environmental protection might seek to minimize disruption to biodiversity; a water utility might seek to minimize service disruption, subject to water quality constraints; an insurance company might seek to minimize risk.

Then, given this new piece of information, the strategy of a mining company or of a water authority may change in response to that information. This type of adaptive decision making is based on the notion of Bayesian updating and generalizes across all exemplars, allowing effective knowledge translation between otherwise siloed communities and transforming the entire process by which natural resources are managed across Australia. DARE will explore the use of how different acquisition function, and different measure of uncertainty affect this decision-making process.

Concept 3 (b) Competing Decisions: A decision to permit exploration and development may bring economic prosperity but adverse environmental or social consequences. Equally the decision to prevent exploration may have detrimental economic and developmental consequences. DARE will train PhD students and industry decision makers to create a framework for evaluating competing criteria in a formal manner, for quantifying outcomes and attach probability distributions to these outcomes. Most existing multi criteria decision making (MCDM) methods focus on decisions when the multiple outcomes are assumed certain. The uncertainty surrounding the outcomes as well as the uncertainty surrounding the relationship between the outcomes is mainly ignored. DARE has carefully chosen industry partners and government departments to bring together those companies making the investment decisions and those responsible for policy development to begin a conversation surrounding these issues.

Concept 3 (c) Weighting Outcomes of Decisions: How much is a national park worth? What is the dollar value we place on a body of water? What is the value to the Australian economy of a body of ore? These are highly subjective quantities. DARE will use the prior elicitation techniques described in concept 2(b) to obtain distributions, with associated uncertainties to weight outcomes. This theme brings together the notion of a cumulative impact of an action and will serve as a capstone piece of all PhD projects.

1. Bayesian Optimization for Sequential Drilling Decisions
2. Bayesian inference to determine optimal sampling strategies for subterranean fauna.
3. Three-dimensional structure of formation boundaries for fracture detection
4. Modelling of Extremes for Better Farm Practices
5. Machine Learning approximations of geophysical forward models for computational tractability.
6. Covariate dependent mixture models for approximation to ecohydrological models.
7. Ecohydrology of flood water and groundwater dependent terrestrial vegetation systems.
8. Scaling hydrological models from continent to catchment.
9. Copula models for Multivariate Extreme distributions to Predict Natural Perils.
10. Heterogenous Dynamic Bayesian Networks for detecting change points is ecosystems
11. Integrated geophysical and geological interpretation and 3D model construction in the
12. Quantifying uncertainty in hydrologic systems under change.
13. Quantifying and communicating uncertainty in flood damage estimation
14. Uncertainty quantified deep neural nets for vegetation and land use maps for Pilbara
15. One Tree Reef; The past and future impacts of sea level and climate change on reef evolution.
16. Models for Earth surface evolution informing what lies beneath.
17. Understanding water losses related to mine-impacted landscapes.

Table 1 Potential PhD Projects

C1.3.3 Training Approach

The uniqueness of DARE stems from its focus on cohort-based PhD experience in Australia, training of highly-qualified graduates in a novel synthesis of mathematics, statistical sciences (computational statistics, data science, and machine learning), geological and environmental science, combined with an appreciation for ethics and responsible research and innovation. Students will engage industry and government organisations to gain a deep appreciation of the issues facing the natural resources sector. These organisations include Newcrest Mining (Newcrest), Insurance Australia General (IAG), McKinsey and Co (McK), National Resources Commission (NCR), the NSW Office of Environment and Heritage (OEH), The NSW Data Analytics Centre (DAC), The Western Australia Biodiversity Scientific Institute (WABSI), Geoscience Australia (GA), and Water NSW. DARE will train the decision-makers and thought-leaders of the future for these organisations thereby generating a positive impact for the Australian economy, society, and the environment and be at the forefront of world's best practise in data science in the resources sector.

The DARE programme will begin with course work in the first 6 months of the programme, coinciding with the launch of Sydney University's new coursework requirement for PhD students. Upon successful completion of the coursework students will then choose a topic. A project must be grounded in both a data science area as well as a domain area. Our partner investigators have identified a number of projects.

All these projects will develop new data science methodologies in the context of an industrial or environmental issue. The programme is co-designed with our industry and government partners and anticipated to take 15 students over the life of the centre.

Training Course Modules

Data science is still an emerging field and it is anticipated that students will come from a variety of backgrounds, with varying exposures to mathematics, statistics, computing and earth sciences. To accommodate this the centre will commence with a three-week boot-camp, designed to even out exposures and provide cohort-based activities for students, industry and government partners and academics. The boot camp will cover basic probability theory an statistical inference as well as courses in R, Python and other data-centric computing languages, and training in the use of High-Performance-Computing. Over an initial 4 months students will take a series of 3 compulsory courses as follows

- **Core Module 1:** This module will involve advanced statistical inference and will include linear mixed models, random effect models, marginal GLM/GEE, model selection and inference, comparison of Bayesian and frequentist methods.

- **Core Module 2:** This module will train students in fundamentals of statistical machine learning. It will involve the following topics: data visualisation and presentation, generative modelling framework, predictive modelling framework, R, Python, C++, Fortran, RMarkdown, Jupyter notebooks, complex computational pipelines, data representation, supervised and unsupervised techniques.
- **Core Module 3:** This module will train students in scientific computing, and data analytics. The main topics here will be nonparametric methods, convex optimisation, matrix decomposition, orthogonal representations, sparse regression, bootstrap, state-space models, Kalman filters, Markov Chain Monte Carlo methods, Multilevel Monte Carlo methods (for introduction to uncertainty quantification), and variational methods.
- **Core Model 4:** This module will train students in a Research Project, which will allow students to put the concepts, knowledge, and experience acquired through Core Modules 1-3 into practice. Students will work in groups of two or three, guided by an academic mentor. Assessment will be via submission of a journal-style paper a short newspaper article/press/media release, and wrap-up sessions in which students explain their projects to the whole cohort, and receive feedback, thus ensuring that all students derive some benefit from all the projects. Datasets will be provided by DARE partners to ensure the relevance of the work.

Students will be required to pass all modules and the Research Project in order to progress in the programme. Student progress will be reviewed formally at the end of every year by a team of DARE academics, including the 'primary' supervisor. The formal requirements for the rest of the research component will follow the host institution's practice. In each of Years 2 and 3, they will take 2x20 hour elective modules, one from each of the clusters below such as Bayesian Data Analysis, Graphical Models, Statistics for Extreme Events, Non-Parametric Smoothing, Advanced Monte Carlo techniques. Figure 2 shows a possible governance structure for DARE.

<u>HDR Subject Areas</u>

All HDR projects will contain at least one element from each of the three core capability areas and with one or more of the resource exemplars, and will focus on specific ITP Knowledge Priority (KP) focus areas from Table 1.
Typically, it is expected that HDR students will spend equal time at their host academic institution and at the industry or government department hosting the project. Table 3 below provides an overview of subject areas. These will be refined in consultation with industry and government partners as DARE develops over time.

<u>International Engagement</u>

DARE has a close collaboration with the prestigious ATI in London. Prof Mark Girolami, Director of the Data Centric Engineering group at the ATI is a PI in DARE. Students with satisfactory progress will be able to spend six months at Alan Turing Institute as part of their training. DARE will aim to expand this international program to other universities engaged in related training area.

C1.4 FEASIBILITY AND COMMITMENT. Delivering DARE's ambitious and broad remit is challenging but achievable via three key innovations. The first innovation is in the selection of projects. DARE recognizes that data analysis and uncertainty quantification is a common theme facing decision makers in the natural resource sector, and yet different elements of the sector have resource specific challenges. The projects have been chosen to exploit both the commonality and the specificity of challenges faced by decision makers in the resource sector. Second, the mix of the CIs and PIs and partner organisations leverages strong existing collaborations to deliver early success while creating new networks for future sustainability. Third the design of the PhD and industry training programme encourages peer learning and ensures a general upskilling of research capability in data science.

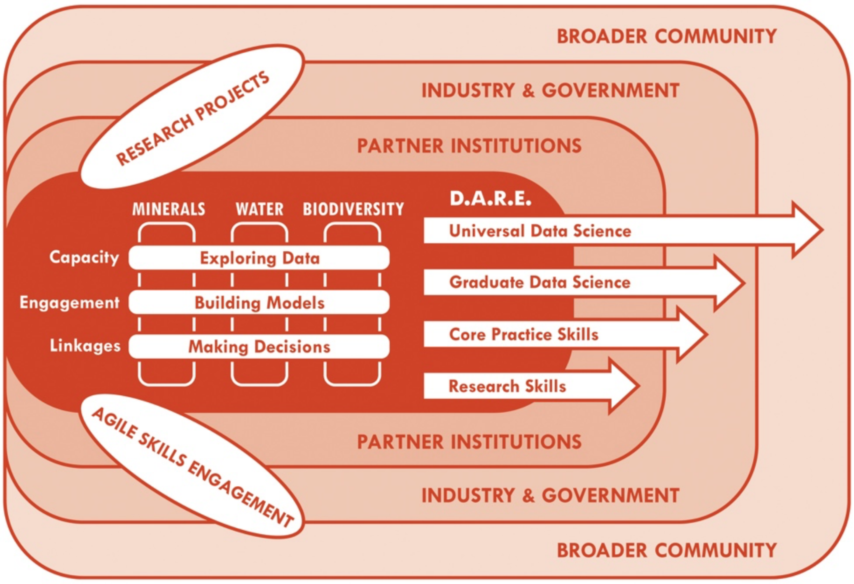
<u>C1.4.1 Design of Programme</u> *Choice of Projects* The scope of projects is wide; projects range from determining the structure of the crust and lithosphere in mineral exploration to modelling the usefulness of climate indicators to predict water storage; from understanding the cumulative impact of a decision to mine on a location's groundwater and biodiversity to understanding the drivers behind landscape and basin evolution; from eliciting and quantifying prior beliefs for evidenced based decision making to the prediction of floods across time and space. Although these may seem disparate when observed from a specific domain, they have striking similarities when observed from a data science perspective. They all require skills development in DARE's three training capabilities, discussed in section **C1.3.2** And it is these similarities in data science which provide the structure of DARE and mitigate against the risk associated with tackling problems across such varied domains. To deliver this suite of projects requires data scientists who understand the commonality of the data science issues as well as domain specific understanding of the physical process. DARE has such a team.

<u>C1.4.3 Partners</u>

Academic.

Industry.

Government.



C1.5 BENEFIT. The benefit of the DARE ITTC is to deliver new insights driven by data science, insights into the deep geology of Australia and ways to explore it, insights into the fundamental mechanisms which govern Australia's resources, insights into the management and allocation of Australia's water supply, insights into the value of biodiversity. Insights into the cumulative impact of an action on Australia's future. These insights are only obtainable by creating a workforce that understands and can quantify uncertainty and is trained to think probabilistically.

The benefit to the mining industry is enormous. DARE will develop the techniques to discover ore bodies that otherwise would have remained hidden. Our partner Newcrest has set a target to have five tier one gold ore bodies by 2025. Ambitious targets like this require substantial improvements in exploration techniques, such as those presented in this proposal. The economic flow of just one additional ore body is in the billions of dollars with subsequent benefits for mining services, employment, government (royalties) and all sectors impacted as a result of increased mining activities. DARE will also substantially reduce exploration costs by sequentially optimizing the locations for expensive data acquisition.

In Agriculture, the Key Industrial Transformation priority of food innovation rests on the reliable access to water. The impact of water on our economy is huge – in agriculture alone the annual value of Australian production is more than \$56 billion, and more than \$15 Billion of this is in irrigated production. At the same time, the 2011 Queensland floods were estimated to cost Australia \$9Billion in exports.
It is difficult to put a value on a balanced ecosystem. But common sense suggests that the cost of not maintaining biodiversity is enormous. [4] states "*total global ecosystem services in 2011 is \$125 trillion/yr …and \$145 trillion/yr … both in 2007 \$US. From this we estimated the loss of eco-services from 1997 to 2011 due to land use change at \$4.3–20.2 trillion/yr, depending on which unit values are used…*"

The foremost challenge in water resource management today is to quantify the joint uncertainty surrounding Australia's water supply, from water sources including rainfall, groundwater and aquifers, and from water usage including the impact of agriculture, mining and other activities. DARE will develop the training, skills and capabilities to apply probabilistic data science approaches to water with the aim of radically improving decision making in our use of this key resource – water allocation, ecosystem management and flood warning.

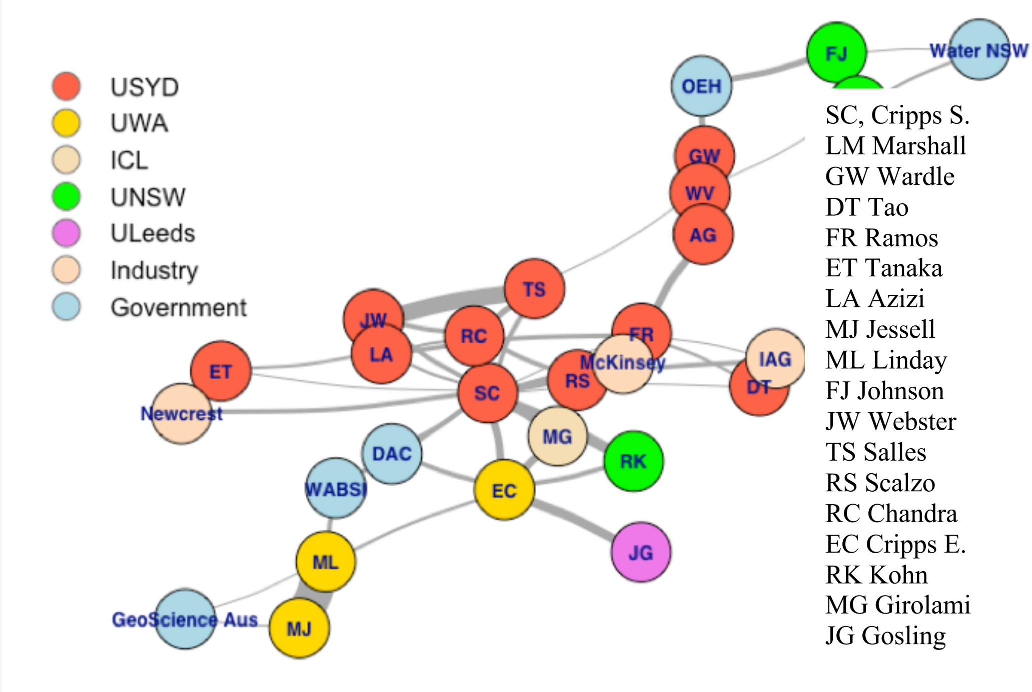
C1.5.1 Growth Centre Benefits The industry growth centres are direct beneficiaries of the DARE ITTC. METS and NERA sector members will benefit through projects delivered in partnership with Geoscience Australia (GA) noting their Strategic Priority to Build Australia's Resource Wealth. In the Food and Agribusiness sector DARE will deliver skills impacting on many of FIAL's strategic priorities, especially: Development of complex data science models for agriculture optimisation (rainfall predictability, soil models, biomass and others), delivery of data science models to support changing climatic conditions and dynamic agricultural systems that need to adapt to changing consumer demands. These are further developed in Table 3 below.
C1.5.2 Market Opportunity and Intended Transformation for Australian Industry. Table 3 below identifies the key skills and graduate capabilities that will be delivered by DARE in the against stated Knowledge Priority areas of the three Industrial Transformation Priorities (ITPs). These are linked to the DARE training and skills concepts – and how these therefor build competitiveness for the sector.
C1.5.3 Benefits to industries and Communities. The beneficiaries of Australia's natural environment also include the community and industries beyond mining and agriculture. For example, Australia's natural environment is an integral part of Australia's global image underpinning our world leading value as a tourism leader. Tourism is a \$54.7b industry for Australia underpinned by pour abundant natural resources and environmental riches. The DARE ITTC will provide for the first time a collective modelling capability that explores biodiversity and natural resources values which can enable more efficient and sustainable methods of exploration, mining and farming. This world first capability will demonstrate that Australia is using innovative and novel methods to understand and quantify the impacts of all users and beneficiaries of our natural resources.

The very structure of the DARE ITTC offers immense collaboration even with sectors that do not typically engage. By using data science as the universal language and examining the natural resources environment in its entirety, we bring to together stakeholders who are more often adversarial. This engagement is cross sectors (e.g. mining, agriculture, natural resource management), between industry and government and nationwide.

DARE ITTC offers a broad level of research capability operating across all pillars of data science and with a number of different domain exemplars (e.g. water, vegetation, resources etc). This cross-sector model will provide opportunities for each researcher to identify common data science challenges and leverage the insights deliverable by focusing on different sectors. This unique model will enable **exponential improvements** in data science discovery, magnifying the value that each insight delivers as it can be applied for multiple stakeholders in the natural environment. For food and fibre production, agriculture and the objectives of FIAL while this may well deliver a direct financial benefit, other insights such as shifting climatic conditions and adaptive production models may offer something far more valuable. A climate risk adaptive model for all of Australian agriculture.

Every investment by the Australian Government must give regard to IP limitations and patent opportunities to fully leverage the commercial opportunities arising from new innovation. The DARE ITTC has a unique new element that optimises potential IP opportunities. The first is the operation of the research around data science. This enables research insights to be published with reference specifically to the data science elements. The **application** of these data science models to Partner Organisation data sets and specific domain exemplars offers a second tier of research insights and the potential to delineate from the more commercial applications. This aspect of the DARE ITTC is the **data application** element. Fundamentally this capability is often the least developed in research – the "what does this mean for business" question. DARE ITTC have identified this and incorporated the data application as a core capability that all researchers will be upskilled in, delivering PhD graduates with exemplary appreciation of commerciality and intellectual property considerations.

C1.6 INVESTIGATORS The DARE team was carefully selected to exploit existing networks while developing new ones. Figure 4 shows the prior connectedness of DARE.



C1.6.2 Leadership

DARE's leadership team comprises CIs Cripps S., Marshall, Ramos, Wardle, Tao, and Lindsay. Together this leadership team provides the skills, experience, and capacity to lead DARE to become an international leader in delivering data science capacity for the natural resources and environment sectors. The leadership team has been deliberately chosen to target emerging leaders:

Diversity of the leadership team has been a key selection criterion. The team is diverse in three dimensions; expertise, gender and seniority. The leadership team has a 50-50 split in expertise; there are three data scientists and three domain scientists. The data science team have different areas of expertise.

The Team and ROPE Evidence. Data Science Team:

Domain Science Team:

Partners

C1.6.4 The Team and Data Science Capability Programme

C1.7 COMMUNICATION OF RESULTS

Developing a communications strategy will be a key early deliverable for the centre manager. The usual avenues of journal publications, press releases, website, socials, and annual reports, will be targeted. In addition to this, DARE's centre manager, will develop a communications programme around understanding risk which will be the first of its kind in Australia and have the following components;

DARE's first objective will be to **communicate the notion of uncertainty and risk** to our partners in industry and government, to policy makers and to the general public. Effective communication between academia and the general public is crucial in all science, but it is particularly challenging in the context of decision making in the natural resources sector. This is because the decision often concerns the use of land which is jointly owned by all members of society, and different groups have different priorities and beliefs, and, most importantly, the decision is made in the presence of uncertainty. That is, a highly probable outcome may not be realised, and equally, a highly improbable outcome may eventuate. The mismatch between prediction and outcome can lead to a lack of public trust in the scientific community and the decision makers. Conveying the concept of uncertainty is difficult, but crucial if trust between all stakeholders in Australia's natural resources and a social licence to operate are to be established.

There is no single right way to convey the quantification of uncertainty, the context is important and multiple means of communication need to be targeted. CI Cripps, S. is organising a conference in November 2019 in Sydney on Communicating Uncertainty Quantification in Natural Resources. DARE will work with organizations such as the Winton Centre for Risk and Evidence Communication, at the University of Cambridge to attract international speakers in this area.

The **visualisations** developed in Project 1 will be used to convey results after the first year. DARE will develop interactive and narrative visualisations of data and models. Narrative approaches are increasingly used by journalists and other communicators to communicate complex nuanced ideas. Narrative visual tools need to facilitate scientists in getting an essential message across while enabling interested viewers to drill down for increased detail, understanding and explanation. Narrative and storytelling using algorithmic visualisation and visualisation animations are new directions in data visualisation which utilise techniques from film making and graphic design to emphasize meaning in visualisations. DARE will explore this in combination with decision support technologies such as argument mapping to record and explain decisions made during the analysis.

The **Executive masterclasses, and Annual Conference** are aimed at bringing together our industry and government partners and academic for workshops focusing on skills such as shaping policy for government, journalistic communication and visualisation, creativity and commercialisation, best practice in research software engineering, and leadership and management across the natural resource These are illustrative examples; the actual workshop material will be developed through consultation with the DARE leadership team and will draw on resources from academic and industry partners designed to be complementary to the training the research students will receive at their academic and industry and government partners.

A large proportion of our partners as government institutions so there will be limited, if any, restriction on the communication of results. For our industry partners, the methods by which results are communicated to and between the partners will depend on commercial sensitivity. For the majority of our industry partners approval mechanisms already exist with the academic institutions to assist with this and developing a mutually agreeable contract will be a key deliverable for DARE.

C1.8 MANAGEMENT OF DATA