Australian National University

TRACKING TRENDS IN INDUSTRY DEMAND FOR AUSTRALIA'S ADVANCED RESEARCH WORKFORCE

This report was created in collaboration between:

Australian National Centre for the Public Awareness of Science,
ANU College of Science

Research School of Computer Science,
ANU College of Engineering & Computer Science

Office of the Dean, Higher Degree by Research

Theory and Applications in Multimodal Pattern Analysis (TAMPA)
within Machine Learning, Data61, CSIRO

Prepared by:

Associate Professor Inger Mewburn
Adjunct Professor Hanna Suominen
Dr Will Grant

August 2017
Canberra, ACT, Australia

CONTENTS

Executive summary and recommendations 2

Project team 3

Key partners and acknowledgements 4

Aims and background 4

Addressing the challenge with a ‘big data’ approach 6

Results 8

Preliminary analysis of the state of Australian industry demand for research skills 10

Applications of The Machine and future opportunities 18

EXECUTIVE SUMMARY AND RECOMMENDATIONS

This project successfully used Machine Learning to analyse job ads in order to better understand Australian industry demand for highly skilled researchers. Though further research and development work is required, The Machine developed in this project can be used to perform a longitudinal examination of Australian industry response to the innovation agenda.

Key deliverables included:

> A Machine Learning (ML) and Natural Language Processing (NLP) based Machine able to 'read' a large set of job ads and assess the research skills intensity using a three-point ranking scale: *High*, *Medium*, and *Low*.

> An evaluation of The Machine using an *average swapped pairs percentage* (ASP%) - a measure of ranking error that takes values from zero for the best performance to 100 for the worst performance. Our testing showed a range from 21.22 for a system using the ad text alone, to a promising 9.42 for one that was enriched with skills highlights.

> An online demonstration visualisation system, called *PostAc*™ (App. No. 1872576) (jobs.t3as.org). *PostAc*™ (App. No. 1872576) ranked the job ads on knowledge intensity, and characterised the demand for PhD skills in Australia in terms of geographic location, industry sector, job title, working hours, continuity, and wage.

> The *Research Skills Annotation Schema*: a new way of describing generic research skills that complements existing frameworks, such as the *Researcher Development Framework*[1] and the *Research Skills Development for Curriculum Design and Assessment*[2].

The key findings in this report include:

> There is a large 'hidden job market' for PhD graduates in the Australian workforce. Only 20.7 per cent of non-academic job ads (2,770 of 13,379 unique job titles) in our dataset asked for a PhD qualification, yet as many as 43 per cent (210 of 483) of the unique job ads that were analysed[3] required a high level of research skills and capabilities, indicative of a PhD.

> Building on this, The Machine read 29,693 job advertisements, predicting 15,440 ads (52%), 10,689 ads (36%), and 3,564 ads (11%) as having a High, Medium, and Low Knowledge Intensity Bandwidth, respectively.

> The visual representations in *PostAc*™ (App. No. 1872576) generated by The Machine revealed some interesting patterns regarding demand for research skills, particularly in industries traditionally assumed to have low demand for PhD graduates, such as manufacturing, transport, logistics, marketing and communication. In addition, other industry sectors were shown as potentially ready to embrace more graduates with research skills, echoing the thinking of the innovation agenda.

Based on the information presented in this report we propose that:

1. The Machine could be refined and used to track changes in industry demand for Australia's higher degree research qualified workforce over a five year period in order to see if this approach is useful as a benchmarking process.
2. Further refinement of auto-coding function could be undertaken to improve The Machine's accuracy and additional engineering of *PostAc*™ (App. No. 1872576) to make it more accessible and useful to end users, including PhD students, graduates, employers and research educators.
3. The fostering of further collaboration between qualitative researchers and ML experts in the application of ML and NLP approaches to science, research and innovation (SRI) relevant datasets could lead to the development of improved tools and datasets, and subsequent improvements in SRI policy development.

[1] www.vitae.ac.uk/researchers-professional-development/about-the-vitae-researcher-development-framework

[2] www.adelaide.edu.au/rsd/framework

[3] This result was derived as part of the hand coding carried out during this project by two content experts

PROJECT TEAM

This interdisciplinary project was based on a strategic partnership across governmental, industrial, and academic sectors. It took place in the Australian National University (ANU) and Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO).

The project team in ANU and Data61 was multi-disciplinary, consisting of qualitative analysts and quantitative scientists, supported by an engineer. The team had expertise in the fields of computer science, research education, linguistics, and social science.

The Australian National University

Associate Professor Inger Mewburn

www.linkedin.com/in/ingermewburn

Inger is the Director of Research Training in ANU. Her team is located within the division of research and provides professional development opportunities for research students across the whole university. Inger has specialised in the study of research education and research student support since 2006. Inger is a senior fellow of the Higher Education Academic (UK) and an adjunct fellow of the School of Culture, History and Language in ANU.

Dr Will J Grant

www.linkedin.com/in/grantwill

Will is Senior Lecturer and Graduate Studies Convener at the Australian National Centre for the Public Awareness of Science ANU. His research and writing has focused on the intersection of society, politics and science, and how the relationships between these are changing with new technologies.

Mrs Stephanie Kizimchuk

anu-au.academia.edu/StephanieKizimchuk

Stephanie is currently a PhD candidate in cultural history and memory studies in ANU, a research officer within the ANU Research Training team, and an education focused academic based in the Teaching & Learning Directorate at the University of Canberra (UC). Stephanie is professionally recognised as an Associate Fellow of the Higher Education Academy.

Data61

Adjunct Professor Hanna Suominen

www.linkedin.com/in/hanna-suominen-b476507

Hanna works as the Team Leader of Natural Language Processing (NLP) and Senior Research Scientist in Machine Learning (ML) in Data61. She is interested in developing and evaluating methods and applications of ML and NLP. Her over 100 publications have been published in A* journals and awarded once for a 10%-elite PhD, twice for a best paper, and four times for the top ML methods for clinical NLP. Hanna's work has led to real-life products for hospitals and sports, and scored competitive grants together with business-plan and teaching-excellence awards. She has been a Co-Chair/Task-Leader of the ShARe/CLEF eHealth Evaluation Labs since 2012 and Assistant/Guest Editor for Artif Intell Med, J Pattern Recogn Res, and Int J Med Inform. Hanna is an Adj/Prof in the University of Turku (Finland), Department of Computer Science, Adj/Assoc/Prof in the University of Canberra, Faculty of Education, Science, Technology, and Maths, and Adj/Ass/Prof in the ANU, College of Engineering and Computer Science.

Mr Travis Simon

www.linkedin.com/in/travissimon

Travis works as a Senior Software Engineer in Data61. He is a software developer with 20 years of experience across a wide range of industries. Some highlights of his work includes building the olympics.com.au website for Channel 7 and the AOC, creating the software systems used for the T3 Telstra share offer, and providing significant contributions to the overhaul of Graincorp's logistics software. His current work involves digitising legislation for computer processing.

KEY PARTNERS AND ACKNOWLEDGEMENTS

This project was made possible via funding provided by the Australian Government Department of Industry, Innovation and Science (DIIS), and raw data provided by SEEK Limited (hereafter called SEEK). For both forms of support we are extremely grateful.

This ANU-Data61-SEEK-DIIS project was partly carried out in the National Information and Communications Technology (ICT) Australia (NICTA) and its successor organisation Data61. NICTA was supported by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centres of Excellence Program.

ANU provided in-kind contributions from Associate Professor Inger Mewburn from the Division of Research and Dr Will Grant from the Centre for the Public Awareness of Science to develop the research skills annotation schema and explore the application of the tools developed by Data61.

We express our gratitude to the following people: Dr Rachael Pitt, Graduate Education Manager at the Graduate Research School of the Griffith University, for assistance with data annotation. Dr Margaret Kiley (adjunct faculty, ANU), Professor Allyson Holbrook (University of Newcastle), Nigel Palmer (University of Melbourne) as the members of our expert panel to identify the initial categories for our annotation schema development and data annotation; and Adj/Prof Leif Hanlen, Technology Director at Data61, for supporting the project preparation in 2015.

Expert peer review and feedback on this report was provided by Dr Margaret Kiley, ANU; Professor Allyson Holbrook, University of Newcastle, and Mr Nigel Palmer, University of Melbourne. We are grateful for their contributions and insights.

AIMS AND BACKGROUND

This project aimed to produce data and methods to address the following two key challenges for Australia: 1) helping universities prepare graduates for workplaces outside academia, and 2) helping industry to recognise the value of the research skills developed by graduates of PhD and MPhil programs.

The PhD was originally designed to train the next generation of academics, but it is now more common for PhD graduates to leave the higher education sector than to go on to an academic career.[4] This is a positive development; highly skilled researchers working in a wider variety of industry sectors is important to our future economic prosperity. PhD students represent one of the most potent vehicles for enhanced collaboration and knowledge transfer between academia and industry. Yet Australia stands out amongst most developed countries for the disinterest non-academic employers' display towards PhD graduates, significantly behind similar countries on the World Economic Forum competitiveness index.[5] While the benefits of introducing more higher degree research graduates into industry are tangible for both the individual and the economy, there remains significant challenges before these benefits can be harnessed.

From the supply side, Australian universities must do more to prepare PhD graduates for a wider range of workplaces. Re-envisioning the PhD will involve changes to the form and content of PhD curriculum, as well as additional co-curricular opportunities. Unfortunately, there is not enough reliable data to inform this work. As a consequence, PhD programs still tend to privilege skills required for an academic career over those required by industry. From the demand side, there seems to be a lack of awareness of what skills and capabilities people develop during the PhD and perhaps a lack of trust in the qualification as producing 'work ready' employees. As noted by former Chief Scientist, Professor Ian Chubb, industry "seems to believe that the content of a degree is more important than the process of learning that underpins the content. You did physics? We don't need physics." [6]

Government can help by setting policy and creating incentives, but the data to inform this work is incomplete and not actionable. While data from the Australian Bureau of Statistics (ABS) and the Graduate Destination Survey (GDS) can show us the uptake of graduates in the business community, these are both lagging indicators. We need future-focused information that tells us about emerging industry needs in order to act by targeting our policy and education efforts appropriately.

In this project, we addressed this need for complete, actionable data with a 'big data' approach, by applying Machine Learning (ML) and Natural Language Processing (NLP), sometimes referred to as 'Text (Data) Mining', to analyse 29,693 authentic job ads.

[4] ACOLA review / GDS data

[5] *Building Australia's comparative advantages*, www.bca.com.au/publications/building-australias-comparative-advantages, *Productivity, Industry Engagement and the PhD Workforce*, www.chiefscientist.gov.au/2013/02/productivity-industry-engagement-and-the-PhD-workforce

[6] *Productivity, Industry Engagement and the PhD Workforce*, www.chiefscientist.gov.au/2013/02/productivity-industry-engagement-and-the-PhD-workforce

ADDRESSING THE CHALLENGE WITH A 'BIG DATA' APPROACH

Previous research on PhD graduate employability has largely concentrated on surveying and interviewing employers and academics about the fitness of graduates for the workplace.

Up to this point, research has provided useful qualitative and quantitative data on employer and graduate perceptions, motivations and career destinations. However, these conventional research methods have been unable to measure, evaluate and track employer demand for PhD graduates.

This project has sought to address this gap by using a 'big data' approach, enabled by intelligent ICT: specifically adopting ML and NLP techniques to 'read' a large set of job ads and assess employer needs for graduates with research skills. ML can be broadly defined as a suite of ICTs that address the challenge of making sense of the ever increasing volume of data generated in modern information dense society. Traditionally ML refers to symbolic computational approaches arising from the artificial intelligence community, or statistical pattern recognition. However, today it is seen as including allied data-intensive areas such as computational NLP that supports producing and using free-form, natural, human language in spoken and written forms (for example, in the context of speech recognition or web-search engine tasks). ML-based NLP researchers construct software artefacts, and as such need to be cognisant of aspects of software engineering and protocol design, as well as human-computer interaction design. ML-based NLP requires multi-disciplinary team based approaches which includes content specialists to supply information helpful to the design process.

In this project, our team employed ML and NLP to reveal the nature and extent of the 'hidden job market' for PhD graduates in a big dataset consisting of 29,693 authentic job ads.

Job ads are a rich and robust source of data about employer needs and expectations, but without ML and NLP, there is no simple, accurate, and objective way to measure non-academic employer demand for PhD graduates. A job ad is a pitch to a potential applicant which outlines the type of work that is required and the type of person that the employer would like to hire. However, if employers do not use "PhD" as a keyword in a job ad, a simple keyword search will not retrieve all ads aimed at PhD graduates. As a consequence, there is a large 'hidden job market' for PhD graduates in the Australian workforce. Only 20.7 per cent of non-academic job ads (2,770 of 13,379 unique job titles) in our dataset asked for a PhD qualification, yet as many as 43 per cent (210 of 483) of the unique job ads that were analysed required a high level of research skills and capabilities, indicative of a PhD (see 'preliminary analysis' starting on page 14). Whilst this dataset does not encompass the full extent of the Australian employment landscape (ads of this dataset appeared to skew towards managerial and professional jobs) it does reflect a dramatic gulf in Australian perceptions of the PhD.

This project sought to design a ML-based NLP algorithm that could learn what a 'PhD shaped job' looks like; highlight these within a large, complex dataset supplied by SEEK and enable interactive search and visualisation of this information as a web demonstration system. In addition, we aimed to teach The Machine to analyse the job ad and tell us what skills and capabilities were most important to employers. This report summarises the activities undertaken and the results obtained in this ambitious project so far.

RESULTS

In this section, we will first provide a summary of the project deliverables and use The Machine to produce a preliminary analysis of the hidden job market for PhD graduates.

Summary of deliverables

The Machine

The primary outcome of this project is a prototype ML-based NLP service on the web for mapping industry demands for PhDs, using a sample of 29,700 job advertisements. This ML-based NLP algorithm is hence hereafter referred to as 'The Machine' and the interface is called 'PostAc™ (App. No. 1872576)'.

We used an expert-annotated Gold Standard (GS), using the Research Skills Annotation Schema, on approximately 500 ads for ML. We evaluated the applicability of Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) to automate these tasks on a hold-out test set of 105 ads with the average swapped pairs percentage (ASP%) in ranking from 21.22 for a system using the ad text alone to a promising 9.42 for one enriched with the skills highlights. The ranking considered the following three-point scale to assess the Knowledge Intensity Bandwidth: High, Medium, and Low. The ASP% is a measure of ranking error that takes values from zero for the best performance to 100 for the worst performance.

The Machine can assess job ad text, decide what knowledge intensity is represented in the ad, and sort the data in a variety of ways. There is ample scope to refine and extend the capabilities and utility of the tools and approaches we have developed in this project.

To summarise, the main steps in the development of The Machine were as follows: 1) convening an expert workshop to help develop the initial ontology, 2) four iterations of hand annotations by expert coders to refine an annotation schema, 3) extraction of a final expert-annotated set of 483 unique ads which was declared as the GS, that is, the ground truth in ML and its evaluation, and 4) experimenting the ML-based NLP algorithms towards learning to automate the data annotation process.

PostAc™ (App. No. 1872576) visualisation

The Machine is coupled with an interactive web interface (i.e. PostAc™ (App. No. 1872576)) that displays the outcomes of The Machine's analysis visually according to a number of scales. PostAc™ (App. No. 1872576) displays The Machine's analysis of a range of job ads from 2015. Later in this report we have used PostAc™ (App. No. 1872576) to produce a snapshot of Australian industry demand for research skills. This preliminary analysis provides a baseline for a longitudinal comparison of the demand for research skills in the Australian workforce.

Research Skills Annotation Schema

As part of the development of The Machine, we created a Research Skills Annotation Schema. This schema was used to hand code the job ad data and train The Machine. Our new annotation schema complements existing ontologies and frameworks that categorise and explain research skill sets, such as the Researcher Development Framework.

Other outcomes

During this research the team also produced the following three additional outcomes:

1. A new method for assessing graduate skills, which can be adopted for other higher education cohorts, either automatically or by hand,
2. A methodology for further development of The Machine, with possible commercial applications, and
3. Several research outcomes that have been or are to be published in the fields of ML, NLP, and research education.

Development of The Machine

The Machine was developed using expert coders from ANU who applied an unique Research Skills Annotation Schema to 500 job samples. Using this schema The Machine learned to auto-coded job ad text, including the skills highlighting on a High-ranked ad of the GS. The Machine auto-coded each ad using what it learned from the expert coders' application of the research skills annotation schema.

We used this trained ML model to predict the ranking for these and the remaining (over 29,500) advertisements. Of the entire set of 29,693 ads, this system predicted 15,440 ads (52%), 10,689 ads (36%), and 3,564 ads (11%) as having a High, Medium, and Low Knowledge Intensity Bandwidth, respectively. The lower threshold values of ≥ 0.8612 and ≥ -0.1820 separated the High rank from the Medium rank and Medium rank from the Low rank, respectively.

PostAc™ (App. No. 1872576): a visual interactive search, exploration, and analysis tool

To make The Machine a useful and accessible tool for analysis, we engineered an online demonstration visualisation system, called PostAc™ (App. No. 1872576). PostAc™ (App. No. 1872576) ranked the ads in the dataset, quantified the most sought-after PhD skills, and characterised the demand for PhD skills in Australia in terms of geographic location, industry sector, job title, working hours, continuity, and wage. At present PostAc™ (App. No. 1872576) does not report The Machine coded Research Skills Annotation Schema information, meaning that specific research skills contained in the job ad texts cannot be displayed to the user. This could, however, be engineered in future versions of the tool, and would provide significant value to a range of end users.

The PostAc™ (App. No. 1872576) demonstration system is available for viewing at http://jobs.t3as.org/. We have included a walk-through of the system below to aid exploration.

The PostAc™ (App. No. 1872576) offers the following five functionalities, placed as the top panel of the website: All, Categories, Locations, Salaries, Job Types, and Configuration. The All functionality presents an analysis of all job ads as percentile distributions by their:

iv. Knowledge Intensity

v. Job Category (e.g., ICT or Manufacturing, Transport, and Logistics as in Examples 1 and 2) and this intensity score

vi. Location (e.g., ICT as in Example 1) and this intensity score

vii. Salary (e.g., 100,000-120,000 AUD per annum) and this intensity score, and

viii. Job Type (e.g., Full time) and this intensity score.

Hovering a mouse over the distribution allows further insight into the data and clicking the data labels allows the user to include / exclude data (Figure 1).

The Categories, Locations, Salaries, and Job Types functionalities allow limiting these distributions to a given job category, location, salary, and job type (Figures 2, 3 and 4).

The category is filtered in by using the menu placed just below the top panel before presenting the distributions. Again, hovering a mouse over the distribution allows further insight into the data and clicking the data labels allows including/excluding data.

The Configuration functionality enables the user to include/ exclude jobs inside/outside academia in order to, for example, analyse only non-academic job ads. For purposes of comparative studies (e.g., longitudinal or sector by sector), this functionality can also be used to lock the scale on the y-axis. Setting the Domain allows the website user to specify the Knowledge Intensity Bandwidth scores (i.e., from the minimal/lower-limit to the maximal/upper-limit value) to be analysed and setting the Range controls the scale of the x-axis of the distributions.

A further development of PostAc™ (App. No. 1872576) could include visualising the location information as an Australian map. This could be built on the existing Australian National Map software, already operated by the Australian Government, but developed at NICTA/Data61 and the NICTA/Data61 Speech to Clinical Text Demonstration (figures 5 and 6).

PRELIMINARY ANALYSIS OF THE STATE OF AUSTRALIAN INDUSTRY DEMAND FOR RESEARCH SKILLS

Using PostAc™ (App. No. 1872576) we generated visualisations based on the 2015 job ad data supplied by SEEK. In this section we have generated a series of histograms which demonstrate the kind of analysis that PostAc™ (App. No. 1872576) is capable of generating.

A key limitation of PostAc™ (App. No. 1872576) at present is the nature of the job ad dataset, which differed significantly from the Australian Bureau of Statistics' (ABS) data both in composition and in the way data was categorised. The ABS derives industry participation from tax office data and their latest release (2014-15) shows that "retail trade" is the largest employer of Australian workers, followed by "healthcare and social assistance" and "construction".

In contrast, our dataset had "education and training" with the highest number of job ads, followed by "healthcare and medical" and "information and communications technology" (see method section for further details). The dataset had a distinct skew towards professional and managerial jobs; the so called 'knowledge economy' rather than low skilled, manual labour work.

While the limitations of the dataset need to be acknowledged, the dataset can still be considered appropriate for this particular project. PhD graduates are likely to seek professional and managerial jobs on graduation and employers are likely to use formal means to hire these sought after, highly skilled employees. Previous Department of Industry research suggests that internet advertising was equal first with personal contact as methods for businesses to recruit researchers. Hence, the dataset was deemed a good enough representation of the non-academic job ads that are potentially visible to PhD graduates.

In the representations shown in this section, the x-axis represents increasing research skills intensity and the y-axis represents the number of jobs at each level. The further along the x-axis that a job falls, the more likely it is to require high levels of research skills. For the purposes of the discussion here, we will treat x=5 as the cut off where the jobs that would be most appropriate for PhD graduates: approximately 25% of the dataset.

Further work would need to be done to see if x=5 is a reasonable assumption, given the nature of the dataset, which was skewed towards managerial and high paid jobs. Table 1 shows the number of non-academic 'PhD shaped' jobs sorted by The Machine into categories employed by SEEK to store their data.

Table 1: number of jobs predicted by The Machine from the SEEK dataset

SEEK's industry category	Number of non-academic jobs requiring high levels of research skills (x=5 and above) in 2015	Highest x value for a non-academic job in that set	% of jobs in set that are x=5 or above (i.e. number of 'PhD shaped jobs available)
Accounting	32	x=7	8%
Administration and office support	36	x=6	3%
Advertising, arts and media	19	x=7	9%
Banking and Finance	144	x=10	34%
Call centre and customer service	3	x=5	1%
CEO and management	94	x=8	52%
Community Services and development	30	x=8	3%
Construction	5	x=6	1%
Consulting and strategy	129	x=10	36%
Design and Architecture	74	x=7	28%
Education and training	811	x=8	21%
Engineering	96	x=7	16%
Farming, animals and conservation	14	x=6	7%
Government and defence	249	x=8	18%
Healthcare and medical	1033	x=10	18%
Hospitality and tourism	2	x=6	1%
Human resources and recruitment	38	x=7	6%
Information and communication technology	622	x=10	21%
Insurance and superannuation	99	x=6	10%
Legal	47	x=7	24%
Manufacturing, transport and logistics	116	x=8	14%
Marketing and Communications	851	x=10	40%
Mining, resources and industry	13	x=6	2%
Real Estate and Property	1	x=5	0.10%
Retail and consumer products	0	x=4	0%
Sales	2	x=5	0.05%
Science and technology	559	x=9	30%
Sport and recreation	19	x=6	44%
Trades and services	9	x=6	2%

The table of results generated by The Machine might be compared to an estimate of the available PhD candidates from the graduating cohort of PhD students from the same year (n= 5334). Table 2, below, shows domestic research doctoral completions for 2015 grouped by broad field of education (Department of Education and Training, 2016). These are broadly grouped into what might be regarded as 'HASS' (Humanities, Arts and Social Sciences) and 'STEM' (Science, Technology, Engineering and Mathematics) respectively.

Table 2: Domestic doctoral completions in 2015 by broad field of education

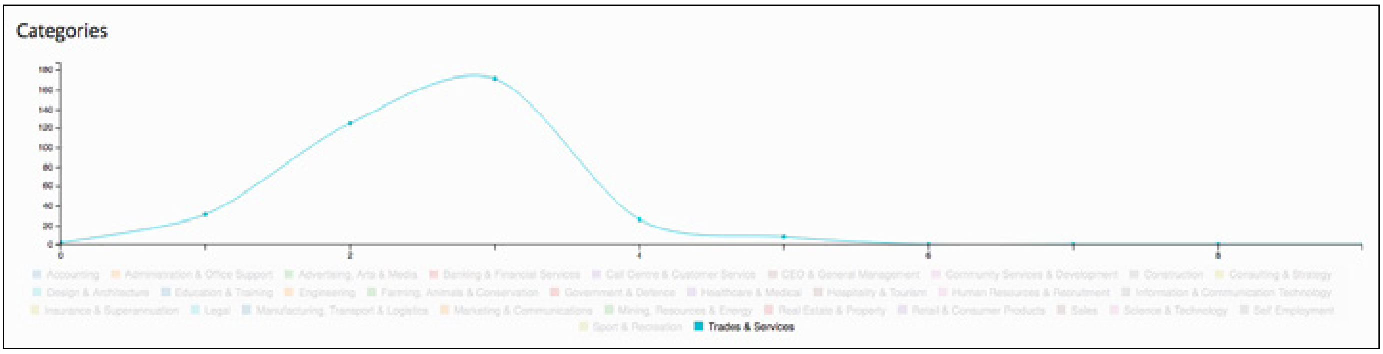
Domestic doctoral completions 2015	Doctorate by Research	HASS	STEM
Agriculture, Environmental and Related Studies	217		217
Architecture and Building	72		72
Creative Arts	287	287	
Education	303	303	
Engineering and Related Technologies	602		602
Health	944		944
Information Technology	146		146
Management and Commerce	301	301	
Natural and Physical Sciences	1,186		1,186
Society and Culture	1,276	1,276	
All 2015 domestic research doctoral completions	5,334	2167	3167
		41%	59%

In some cases the broad field of education recorded for graduating candidates appears to align with SEEK's industry categories(Table 1). At a very general level these may be used to provide an overall view of research graduates relative to positions requiring research skills. For example, we might observe there were 944 domestic candidates graduating with a research doctorate in health-related disciplines in 2015, with what appear to be 1033 non-academic "healthcare and medical" jobs requiring high levels of research skills for them to apply for.

While it is tempting to 'map' each graduate field of education industry sectors and vice versa, attempting to map each of the fields at a higher level of detail would lead to misleading results. For example, industry categories do not always neatly align with what are very broad fields of education (such as Society and Culture, and Natural and Physical Sciences). It would also be misleading to assume that graduate pathways between field of education and field of employment neatly align. A PhD in a health-related discipline may lead to a position with government just as a PhD in IT may lead to employment in the healthcare and medical area. In the absence of longitudinal unit-level graduate destinations data we can do little more than compare the number of candidates graduating from each of the broad discipline areas with the broad industry categories where jobs requiring high levels of research skills appear to be available (as outlined in table 1 and 2 above).

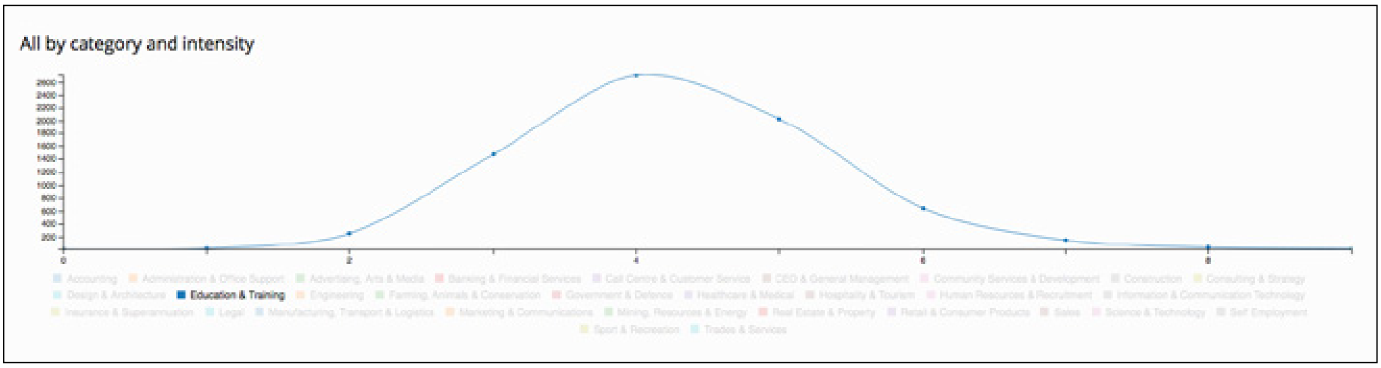
As expected, some industries showed distinct patterns of higher and lower demand for research skills. For example, "trades and services" (Figure 7, below) has an abrupt drop off at x=4 and hits zero at x=6, which shows, as one might expect, there are no "trades and services" jobs outside of academia requiring high levels of research skills.

Figure 7: histogram of research skills intensity for “trades and services” from the SEEK set



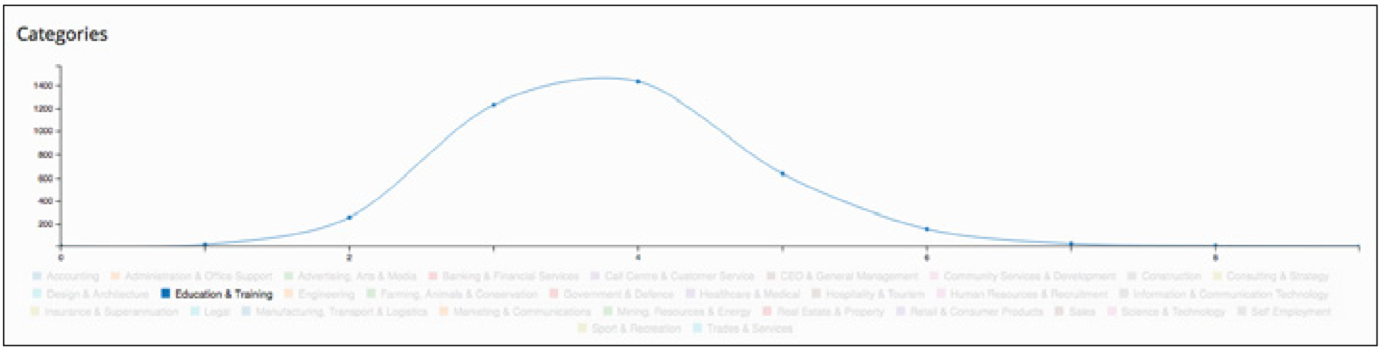
Likewise all the jobs categorised as “education and training” show a distinct bulge at x=5 and x=6, consistent with the hypothesis that there is a high need for research skills in this knowledge intensive industry, shown in Figure 8 below:

Figure 8: histogram of research skills intensity for “education and training” for all of the SEEK dataset



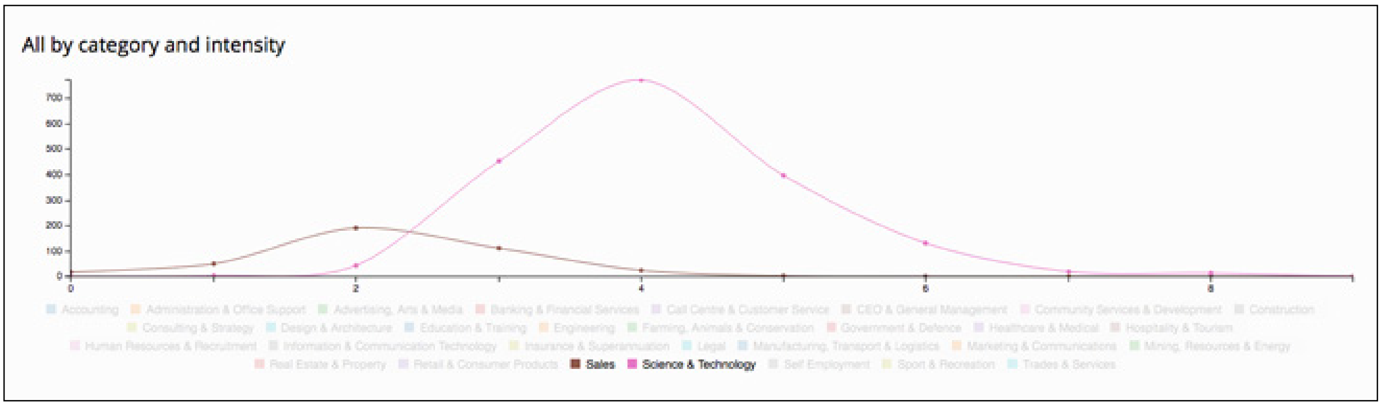
When the job set is filtered to remove most of the academic jobs (see note on ‘data cleansing’ in method below), the bulge of research intensive jobs moves back along the x-axis, indicating the need for research skills is lower in areas of “education and training” other than higher education, see Figure 9 below:

Figure 9: histogram of research skills intensity for “education and training” with academic job titles removed from the SEEK set



As expected, the machine shows us that different industries have different levels of demand for research skill intensity, which can be compared using the web interface. By turning on and off various industries, we can quickly compare the extent and nature of demand for research skills in different industry sectors. For example, Figure 10, below, shows a comparison of “sales” and “science and technology”; a comparison of the nature of demand for research skills between two industry sectors with quite different demands for research skills:

Figure 10: histogram of research skills intensity for “Sales” superimposed on “science and technology”



As we might expect, similar industries generate broadly similar histogram curve shapes, increasing confidence in the accuracy of the machine. For example, Figure 11, below, is a comparison of two low research intensity industries, “sales” and “call centre and customer service”. Although there are relatively more research skills intensive jobs in the “call centre and customer service” sector, both have a relative absence of high research intensive jobs above the cut off of x=5:

Figure 11: histogram of research skills intensity for “Sales” superimposed on “call centre and customer service”

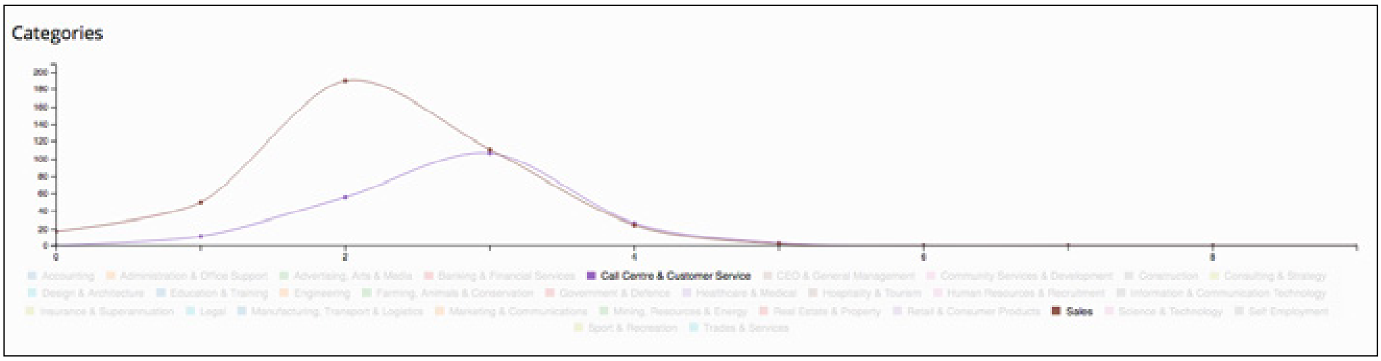
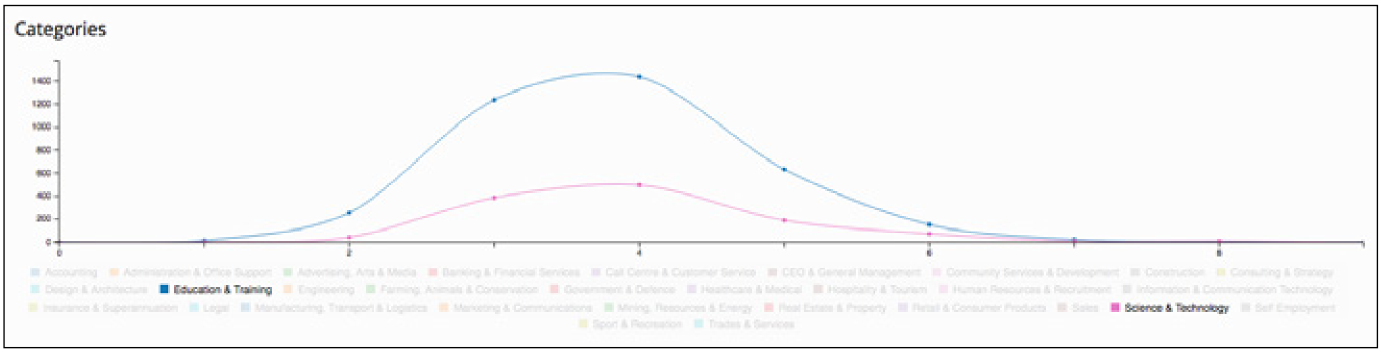


Figure 12, below, is a comparison of two high research skills intensity industry sectors, “education and training” and “science and technology”. The similarity in curve shape shows there is an almost equal relative demand for people with research skills, despite different numbers of jobs overall, as can be seen by the difference in the y-axis (education and training had the highest number of jobs in our dataset).

Figure 12: histogram of research skills intensity for “education and training” superimposed on “science and technology”



By superimposing all the different industry sectors the web interface user can see at a glance the overall composition of job ads in the dataset. Figure 13, below, shows “education and training” (blue) and “healthcare and medical” (purple) and “information and communications technology” (grey).

Figure 13: all seek.com.au industry sectors superimposed

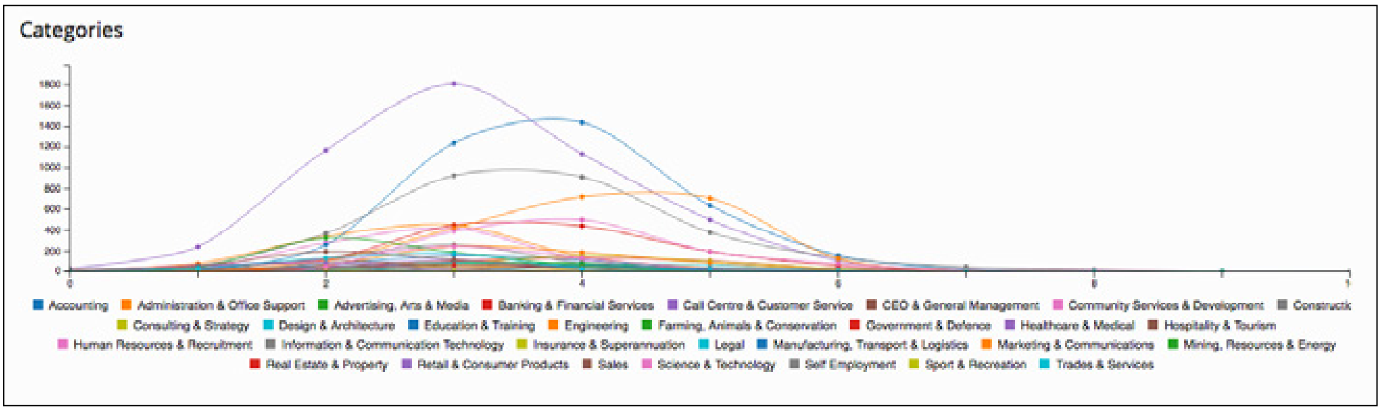
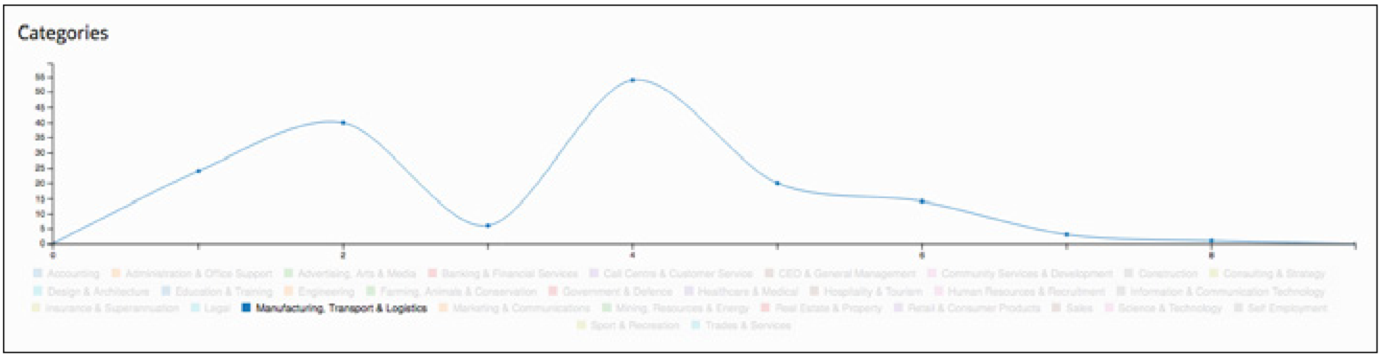


Figure 13 shows how our dataset departs significantly from the dataset from the Australian Bureau of Statistics (ABS), which derives industry participation from tax office data. ABS's latest data (2014-15) shows that "retail trade" is the largest employer of Australian workers, followed by "healthcare and social assistance" and "construction".

The representations generated by The Machine was able to reveal some interesting patterns regarding demand for research skills, particularly in industries traditionally assumed to have low demand for PhD graduates. For example, Figure 14 shows an intriguing histogram for "manufacturing, transport and logistics", revealing a diverse spread of both low and high knowledge intensity jobs in this industry sector.

Figure 14: histogram showing "manufacturing, transport and logistics"



A review of job titles in Figure 14 shows there is demand for people to fill a range of roles, from hands on work like fruit picking, to researchers and strategic thinkers tasked with managing the complexities of contemporary logistics chains and planning for the future. Job titles from different parts of histogram are summarised in Table 3, below:

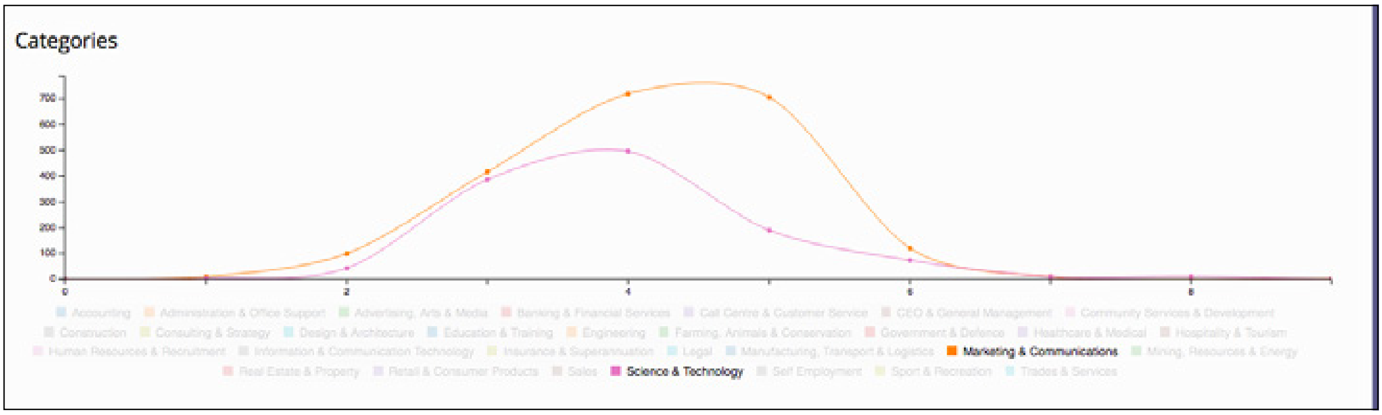
Table 3: selection of machine sorted job titles from "manufacturing, transport and logistics"

Selection of job titles where X < 4	Selection of job titles where X > 5
Storeperson Customer service officer Qualified train drivers Supply officer Fruit packers Process operator Shuttle bus driver	Compliance and reporting analyst; Head of development, Nutritionals Sourcing officer (contract negotiations) National manager, Operations and maintenance business intelligence Operations and supply chain management Manager of strategic market analytics

Although, as previously noted, low intensity jobs may not be well represented in our dataset, the data from “manufacturing, transport and logistics” does appear to reflect an industry undergoing transition. We could hypothesize that this distribution of low and high research skills intensive jobs reflects how one sector might be responding to changes in technology, in particular the affordances it offers. Researchers are useful workers who can potentially find novel ways to make the transport industry more efficient and profitable, so as technology is introduced we could expect to see demand for these workers increase. It is in this area that performing a longitudinal comparison, year on year, may have value as it will enable emerging trends in demand for research skills tracked.

It should be noted that some of the high knowledge intensity job titles imply a speciality and depth of experience in a field. This kind of finding can inform new strategic initiatives, such targeting industry PhD scholarships by showing the value of adding research skill sets to a prior professional knowledge base. The so called ‘digital disruption’ trend is likely to be coupled with a higher demand for skilled knowledge workers, including those with research skills. Using The Machine we can start to explore whether the effects of digital disruption are starting to be felt more in some industries than others. For example, somewhat surprisingly, “marketing and communications” showed an even higher demand for research skills than “science and technology” in the non-academic job market, see Figure 15, below:

Figure 15: histogram showing ”marketing and communications” superimposed on ”science and technology”



A closer examination of job ad text in the “marketing and communications” set shows high demand for people with skills in statistical programming tools (such as ‘R’), qualitative and skills against workforce, giving them an evidence base to determine their own skills gaps and identify targeted opportunities for professional development. The need for research skills and the connection with creativity can be seen in a comparison of job titles, shown in Table 4 below:

Table 4: selection of machine sorted job titles from “Marketing and communications”

Selection of job titles where X < 4	Selection of job titles where X > 5
Events Administrator	Senior advanced analytics professional
Community Engagement Marketing Officer	Innovation consultant
Media & Communications Coordinator	Senior Quantitative Account Manager: research problem solving research
Marketing & Communications Coordinator	Market research and insights manager
Schools and Communities Engagement Manager	Qualitative account director - unique senior role in innovation and brand
	Market research and insights manager

APPLICATIONS OF THE MACHINE AND FUTURE OPPORTUNITIES

The Machine makes visible:

- the desire for high end research skills across different industry sectors,
- where these jobs are located, and
- what incomes can be expected by successful applicants.

It already enables us to do further analysis of the otherwise 'hidden' jobs suitable for PhD graduates, and, with extension, it would allow a range of other outcomes and applications to be realised.

In its current form, The Machine can be used to perform a detailed analysis of demand for research training. Supplied with up to the minute job data sources, The Machine could provide real time details on demand for higher degree knowledge workers usable in policy development. Using The Machine policy makers could:

- Perform fine-grained analyses of demand for skilled researchers in a given industry sector;
- Target further analysis of the specific researcher skill sets required by various industries;
- Measure the alignment between employer demand and targeted funding initiatives in the area of research training, enabling more forensic targeting of initiatives such as scholarships, internships, and industry incentive packages;
- Gain a better understanding of high research skill job opportunities by region, where such analyses could be useful for strategic planning of initiatives that target growth, specifically in regional and remote Australia;
- Identify the industry sectors that are most and least open to knowledge transfer with the academy.

Universities and policy makers need data to inform research degree curriculum design. Using The Machine with the current dataset, research education experts and statisticians could:

- Produce data to inform evidence based curriculum audits to explore how teaching and learning activities align with current industry demand;
- Identify emerging trends in demand for research skills to inform course design and planning around student load;
- Use the insights from an analysis of the workforce to help research students identify and develop in-demand skill sets and capabilities which make them employable in a range of industry sectors outside academia;
- Generate benchmarking data for researchers interested in labour force dynamics to enable cross sector and international comparisons.

Using a refined version of The Machine with additional data, researchers in educational and social sciences together with those in (digital) humanities could potentially address the following questions:

- Are we training the right kinds of graduates for industry? How, if at all, does industry demand map onto academic disciplines as represented in FoR codes and student load?
- How does the situation in Australia compare with that faced in other countries? Are employers starting to respond to global trends towards more highly trained graduates?
- How might this data be thought of as a measure of how Australian industry is responding to the innovation agenda?

Furthermore, The Machine, if shared, could help increase employer awareness of the value of graduates with high level research skills. The Machine could be used by representative councils and industry groups to educate industry stakeholders on the value of PhD graduates to their workforce. Highlighting case study employers or industry sectors could prove incredibly valuable in demonstrating value to others.

Beyond this, the data generated by this project and The Machine itself open up new lines of inquiry and research opportunities, specifically:

- Generating benchmark data for a longitudinal study of Australian industry demand for skilled researchers. Such data could be used to see the effects of policy initiatives to increase industry collaboration with the academy.
- Enabling longitudinal study of various industry sectors, specifically to explore how demand for research skills might change in response to factors such as digital disruption and structural changes in response to other economic drivers.
- Enabling targeted studies to inform policy interventions to drive innovation, particularly in regional areas and industries with currently low levels of demand for highly skilled workers.
- Adapting the methodology to make machines that will enable us to analyse the prospects for other cohorts in higher education, such as undergraduate and Masters degrees.

Further development of the machine into a web portal could help current PhD candidates by:

- Making the emerging research workforce trends more visible to those considering or undertaking a degree so they can make informed decisions about courses of study.
- Provide PhD students and educators examples of available jobs based on individual graduates' skills and interests.
- Enable PhD graduates audit and assess their own skills against industry demand.

CONTACT US

Australian National Centre for the Public Awareness of Science

ANU College of Science

Peter Baume Building #42A
Linnaeus Way
The Australian National University
Acton ACT 2601 Australia

T +61 2 6125 0498
E cpas@anu.edu.au
W cpas.anu.edu.au

CRICOS Provider #00120C

MO_RS173309