A machine learning analysis of the non-academic employment opportunities for PhD candidates in Australia

Abstract:

Can Australia's PhD graduates be better utilised in the non-academic workforce? There has been a historic structural decline in the ability of PhD graduates to find work within academia for the last couple of decades (Forsyth 2014). Around 60% of PhD graduates in Australia now find jobs outside the academy, and the number is growing year on year (McGagh et al. 2016). The PhD is a degree designed in the early 20th century to credential the academic workforce. How to make it fit contemporary needs, when many if not most graduates do not work in academia, is a question that must be addressed by higher education managers and policymakers. Progress has been slow, partly because of the lack of reliable data-driven evidence to inform this work. This paper puts forward a novel hybrid quantitative/qualitative approach to the problem of analysing PhD employability. We report on a project using machine learning (ML) and natural language processing to perform a 'big data' analysis on the text content of non-academic job advertisements. This paper discusses the use of ML in this context and its future utility for researchers. Using these methods, we performed an analysis of the extent of demand for PhD student skills and capabilities in the Australian employment market. We show how these new methods allow us to handle large, complex datasets, which are often left unexplored because of human labour costs. This analysis could be reproduced outside of the Australian context, given an equivalent dataset. We give an outline of our approach and discuss some of the advantages and limitations. This paper will be of interest for those involved in re-shaping PhD programs and anyone interested in exploring new machine learning methods to inform education policy work.

The PhD – a problematic degree?

From the policymaker’s point of view, the need to smooth the transition of people between academia and industry is urgent. PhD students represent one of the most potent vehicles for enhanced collaboration and knowledge transfer between academia and industry, yet many of them remain in the academy, despite rampant under-employment due to the contingent nature of the workforce. Australia stands out amongst most developed countries for the disinterest that non-academic employers display towards PhD graduates; Australia lags significantly behind similar countries on the World Economic Forum competitiveness index (Schwab & Sala-i-Martin 2016).

The PhD was initially designed to train the next generation of academics, but this career outcome is looking less likely for today's graduates. There have been claims there is an over-supply of graduates for academic positions over the last decade at least (Coates & Goedegebuure 2010, Edwards 2010, Group of Eight 2013). The latest Australian data, showcased in the Australian Council of Learned Academies (ACOLA) report (McGagh et al. 2016), suggests that 60% of Australia's PhD graduates will not end up in academia, a finding consistent with other advanced economies. For example, a recent survey by the Vitae organisation (2013) in the UK showed that although the overall unemployment rate for PhD graduates was low (around 2%), only 38% of PhD graduates are now employed in academia after graduation.

The growing awareness of a range of possible career destinations for PhD graduates has led to widespread questioning of the ‘traditional' model of the PhD curriculum, with its emphasis on producing a written dissertation. In reality, however, people have been questioning the PhD's fitness for purpose for an extraordinarily long time. Edgar Dale wrote the earliest paper we could find; "The training of Ph.D.s" appeared in the Journal of Higher Education in April 1930 - some twenty-five years before the first PhD was awarded in Australia (Dale 1930). Disturbingly, Dale's 88-year-old paper focusses on complaints about PhD candidates not being fit to teach; a debate which is still live today in academic circles (for an in-depth analysis of this debate, see Probert, 2014). While Dale's critique focused on the lack of teacher training during the PhD and its relevance for academia, we can assume that graduates who go on to careers outside academia will need a broad range of graduate capabilities to compete for the non-academic jobs on offer. However, determining what these capabilities should be is no easy task. Platow (2016) notes that supervisors and candidates had different ideas about what these graduate capabilities mean and that, often, a formal framework for their use in teaching and learning activities is lacking. Refashioning the PhD around new skills and capabilities development requires an understanding of what all employers need, not just academic requirements.

If we are to develop PhD programs to better support graduates who transition out of academia on completion of the PhD, we will need better data to inform decisions. Academic managers typically take a 'roundtable meeting' approach to employer consultation, but this approach is methodologically limited as it relies on personal networks and retrospective self-report. It is hard to poll sufficient numbers of employers to ensure the prospective market for graduates has been adequately scrutinised. We suspect, based on findings reported later in this paper, that employers in Australia have a poor understanding of what PhD graduates could offer to their business. It is undoubtedly challenging to get around problems with retrospective self-reporting, where employers may only remember outstanding, or deeply problematic, encounters with PhD graduates.

The research reported in this paper builds on an exploratory study by Pitt and Mewburn (2016) which reported on the text analysis of academic job advertisements to see what skills and capabilities academic employers wanted from PhD graduates (Pitt & Mewburn 2016). To replicate this earlier study we needed to locate a dataset of non-academic job ads. It was in the process of solving this problem that the current project was born and the research started to take a different direction: to track overall employer demand for PhD graduates. Our previous work concentrated on one sector: academia. Jobs were easily sourced from university notice boards, but sourcing an appropriate set of non-academic job ad texts was a more complex task. The job opportunities and destinations outside of academia are multiple. Our ambition of shaping PhD programs via employer demand needs to start with a better understanding of the possible employment destinations. The 'Tracking Trends in industry demands for Australia’s advanced research workforce' project (2017) was a collaboration between the Australian National University (ANU) and the computer science research centre 'Data61', with assistance from the Australian Department of Industry and Australia's leading jobs marketplace, Seek.com.au. This project set out to explore the demand for PhD graduates in the Australian workforce. We believe this paper presents one of the earliest, if not the first, worked examples of the application of Machine Learning (ML) techniques to the problems of higher education curriculum development. ML has the potential to dissolve the divide between quantitative and qualitative data analysis by enabling qualitative techniques to be applied on an unprecedented scale. This study will, therefore, be of interest to those who would like to know more about applying these new methods to old problems in education.

Method

We obtained a dataset of job ads from the largest online employment marketplace in Australia, 'Seek.com.au'. The dataset consisted of 29,693 authentic job ads from 29 different industry categories. Each of these texts is a 'wish list' of sorts, reflecting employer aspirations for new hires and thus a rich starting point for an analysis of what employers are looking for when they hire a researcher. While a surface reading of the texts implied that many companies were looking for people with advanced research skills. Surprisingly, most did not mention the PhD as a qualification. The dataset was too big to sort by hand, so we explored other ways to identify the ‘PhD shaped’ (research skills intensive) jobs and isolate them for analysis. This dataset does not represent all job opportunities in Australia in 2015 as many jobs are never formally advertised. Our analysis of the dataset suggested that was skewed towards professional 'white collar' jobs while Australia has a large retail sector. However, since Seek.com.au is the most prominent site of its type in Australia, we considered this set to be appropriate for the use we were planning to put it to and an excellent sample of potentially available jobs for PhD graduates outside of academia.

Our ambition was to develop an algorithm that could ‘read’ this big dataset of job ads, sort the ads by research skills intensity and use the results to assess employer needs for graduates with research skills. We adopted a ‘big data' approach, enabled by computer cognition, adopting ML and Natural Language Processing (NLP) techniques. ML is a bracket term for a suite of information computer technologies (ICTs) that address the challenge of making sense of the ever-increasing volume of data generated in our information-dense society. Traditionally ML refers to symbolic computational approaches arising from the artificial intelligence community, or statistical pattern recognition. However, today it is understood to include allied data-intensive areas such as computational NLP that support 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). Far from being a 'press a button / get a solution' option, ML-based NLP research requires humans with a complex set of skills and capabilities. 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, including content specialists to supply information helpful to the computer scientists who perform experiments that enable the production of algorithms. The ANU/Data61 team consisted of two experts in research education who designed a coding schema for the job ads, a specialist panel to oversee the coding work, a computer scientist with ML-NLP expertise to perform experiments and a software engineer to construct a platform to visualise our results.

A discussion of how the algorithm is constructed is not possible here because the software is now commercial in confidence and the complex technical specifications would not, in any case, be of interest to this audience. Instead, we will summarise and explain our approach with a view to assisting others who may want to take a similar approach to higher education research in the future.

The main steps in the development of our coding schema and algorithm development were as follows:

1) We convened an expert workshop to help develop the initial ontology, which we have called the "Research Skills Annotation Schema" (RSAS). In ML, an ontology is a list of statements that are used to code text and produce a training set for ML. The RSAS describes the 'ideal graduate' from a PhD program. While we cannot share our exact schema because it would allow it to be backwardly engineered, our list included statements about general research capabilities and attitudes to work. From an initial eleven statements recommended by our panel, we refined the ontology to nine. We left out statements about writing as our expert panel could not agree on definitions that would apply to all PhD graduates. Academic writing is highly discipline-specific and writing in one domain may not be legible in another. Some disciplines, like mathematics, put minimal emphasis on writing and tend to produce short theses, while others, such as history, set high expectations. Our ML experiments showed that only three of our statements were critical to the decision making process.

2) We performed four iterations of hand annotations by six coders to refine an annotation schema. Two coders were PhD supervisors with extensive experience of working with early career researchers, one coder was an expert in research education, and the other three were PhD students. The majority of the work involved applying the RSAS to the job ad texts and then making a judgement about the advertisement as a whole, and categorising it as either high knowledge intensity, medium knowledge intensity, or low knowledge intensity (with high knowledge intensity signifying 'PhD shaped' jobs). The human coders needed to reach a ‘gold standard' (GS) intercoder agreement of at least 60% to produce useful input for our computer scientist. Only the two experienced PhD supervisors were able to reach the required level of inter-coder agreement.

3) The two coders who reached the required standard of inter-coder agreement in the sorting task continued alone. Each coder marked a dataset and then checked each other's work. This process generated 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).

4) Our computer scientist constructed the ML-based NLP algorithms towards learning to automate the data annotation process. From this we developed a sorting algorithm, which we applied to our full set of job advertisements, resulting in a ranking of all jobs on a scale of one to ten, with one being no discernable need for research skills and ten being highest research skills intensity jobs.

5) To make a useful and accessible tool for analysis of Australian employer demand for research skills, we engineered a pilot online demonstration visualisation system (POVS), which ranked the ads in our data set, and displayed the demand for PhD skills in Australia as

a series of histograms. These visual representations show the number of jobs in the dataset by geographic location, industry sector, working hours, continuity, and wage.

As this kind of research method is novel, a brief description of these problems will be helpful to others seeking to do similar work in the future. The following insights emerged during the course of this project:

1) The process of generating a useful and robust ontology is critical to the success of this kind of machine-assisted work. Human expertise and knowledge must be carefully validated and then translated. We employed a community approach to ensure the relevance and usefulness of our ontology. Convening an appropriate team who can provide advice and feedback was vital. We were able to draw on the knowledge and experience of some of the key experts in research education. Our initial panel consisted of Professor Alysson Holbrook, Dr Margaret Kiley and Nigel Palmer who helped us develop initial categories for the RSAS and feedback on our subsequent refinements and the final report. Dr Rachael Pitt provided valued advice throughout the project, in particular with the coding task. We would like to thank them for independent feedback that ensured the project team could have more confidence in our method and findings.

2) The process of coding is labour intensive and entirely different to standard procedures in qualitative research. The annotation agreement is required for a machine to be able to learn from the human coding. ML does not have the human capability to abstract from the provided solutions and generalise to unseen words. For example, if a given concept is sometimes codified to one class and other times to another class, the machine needs many, many examples to lean the classification rule or signal between these decisions or to rule out simple mistakes or differences of opinion or shuttle differences in interpretation.
The critical problem in our case was the human ability to generate the degree of accuracy required regarding the number of words highlighted in the text spans. In 'normal' qualitative coding work, inter-coder agreement is about the location of meaning in the text, not the specific words which are suggestive of that meaning. Despite four of the six coders having extensive experience with qualitative research, they were not accustomed to having the results of their work measured to 2 decimal places. The coders had to meet frequently to review results and argue about the meanings in and application of the schema. This resulted in at least three iterations of the schema before we found one that two coders could use to produce 60% intercoder agreement. We relied heavily on our computer scientist's experience in this process, which required us to invent tools and note-taking procedures to help the process along. Anyone contemplating this work will need to manage the team and interpersonal dynamics. The process of reaching inter-coder agreement, if done correctly, relies on people who can listen and critically engage with each other about ideas without succumbing to inter-personal tension.

3) Relating to point two, above, researchers should compare the labour costs of analysing a dataset without machine assistance before deciding if the ML approach is worth adopting as the human labour costs of training machines are substantial. Through many hours of repetition, our human coders did become more accurate and predictable as they performed the task (perhaps we could say they became more like human machines?). Accuracy and reliability was greatly increased through using a more user-friendly text mark-up tool (in this

case, *Dedoose*) rather than open source tools, which have much more rudimentary functions. Interestingly, in this project at least, the affordances of human intelligence and ML are surprisingly similar. Tasks that were hard for people (i.e., highlighting text) were hard for the machine too and the easier tasks (i.e., ranking) were easier for both people and machine. What differed was the speed the task could be completed. We estimate our coding work took around 1600 hours of human effort, but, in our case, using ML was the right decision. Each advertisement took between one and five minutes to analyse, with a median time of 3 and a half minutes; if one person attempted to analyse all the job advertisements the task would take at least 1750 hours. Although the time to complete this specific task is around the same for ML vs human only analysis, there are two benefits to using ML: accuracy and repeatability. We found most people can only apply the necessary focus and accuracy to coding job ads for about an hour at a time. Once the training set is produced the machine can replicate the task in seconds. A trained algorithm can repeat the analysis on different datasets, increasing our analytical reliability and the utility of the results. In other words, this kind of ML is labour intensive in the setup phase, but the effort is rewarded in the execution phase.

4) Getting access to appropriate data is difficult and, when it comes to text-based analysis, no dataset will be perfect. We know that many jobs are never advertised and filled via networks and recruiters, so no analysis of job ads will be a complete picture of the labour market. It is worth noting that Seek.com.au is the largest online employment marketplace in the country, so the number of ads was probably the most we could source from any one supplier. The main limitation of our pilot visualisation system is the composition of the set, which differs significantly from Australian Bureau of Statistics (ABS) data both in composition and in the way data is 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". By contrast, the Seek.com.au dataset had "education and training" with the highest number of job ads, followed by "healthcare and medical" and "information and communications technology". The Seek.com.au dataset has a distinct skew towards professional and managerial jobs; the so-called ‘knowledge economy' rather than low skilled, manual labour work. While the limitations of this dataset need to be acknowledged, it is sufficient for this particular project. PhD graduates are likely to seek professional and managerial jobs on graduation and research suggests employers are likely to use formal means to hire these sought after, highly skilled employees: internet advertising was ranked equal first with personal contact as methods for businesses to recruit researchers (Allen Group, 2010). In the next section, we show how our pilot online demonstration visualisation system can be used to produce a snapshot of Australian industry demand for research skills, as represented in the Seek.com.au dataset.

Results

Using this ML approach we were able to generate many more results than might be possible using standard qualitative methods. One of the challenges of working with the results of big data is how to represent and navigate them in a way that enables insights to be generated. For reasons of space, and because this paper aims to discuss this ML method and its utility, we will only be able to include highlights in this paper. We have included screenshots of our pilot visualisation system to illustrate our results. These graphs are a working example of how to portray ML-NLP results to non-technical audiences. The results we feature here are to demonstrate the utility of the approach and insights of immediate value for policy makers and planners, rather than a complete report of the state of Australian industry demand for researchers in 2015, as this would require a more detailed discussion of workforce dynamics.

Of the entire set of 29,693 ads supplied by Seek.com.au, 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 ≥ -12 0.1820 separated the High rank from the Medium rank and Medium rank from the Low rank, respectively.

In the interest of simplicity, our pilot visualisation system displays the machine ranking output on a scale of one to ten, with one being lowest research skills intensity and ten being the highest. It then remained to choose where the ‘PhD threshold' might lie: the cut-off above which a job ad could be called ‘PhD shaped' with confidence. In our analysis, we treated x = 5 as the cut off where the jobs that would be most appropriate for PhD graduates. This decision resulted in approximately 25% of the Seek.com.au dataset falling within our 'PhD shaped' definition. Further work would need to be done to see if x = 5 is a reasonable assumption, given the nature of the data set, which skews towards managerial and high paid jobs.

Table 1, below, shows the number of non-academic ‘PhD shaped' jobs sorted by the machine into employment categories used by Seek.com.au:

We used our pilot demonstration system to produce histograms of the job ad rankings. In the diagram below, the x-axis represents increasing research skills intensity; 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 higher levels of research skills. The diagram below, shows a representation of all the jobs in the Seek.com.au data set as a histogram. The coloured section under the curve is a graphical representation of how many jobs we deemed 'PhD shaped':

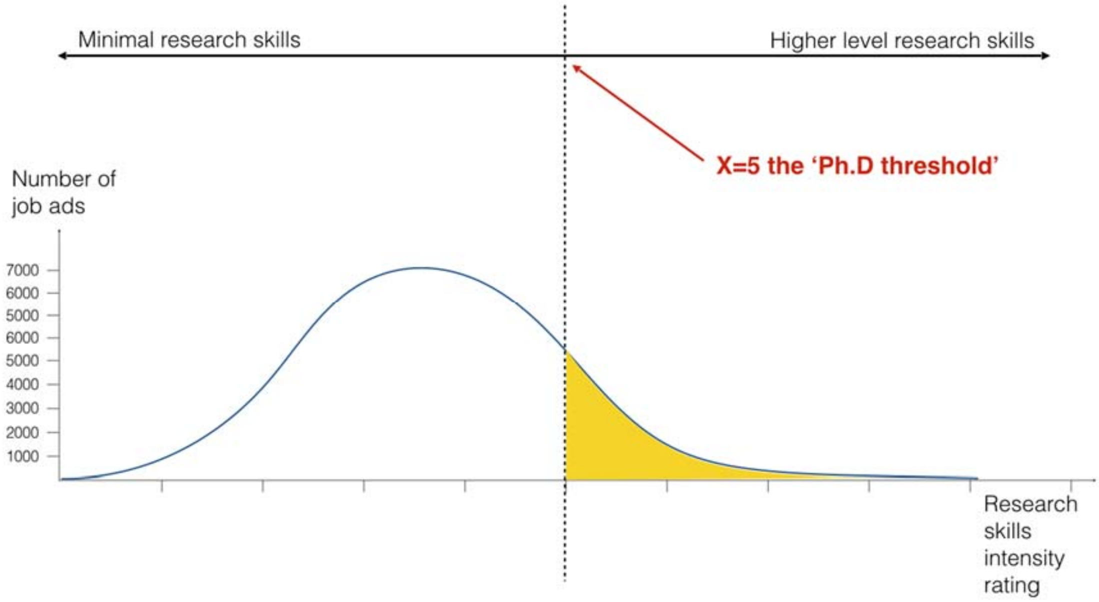

Figure one: distribution of 'PhD shaped jobs' from 2015 seek.com.au dataset

As expected, some industries showed distinct patterns of higher and lower demand for research skills. For example, "trades and services" (Figure 2, below) has an abrupt drop off at x = 4 and hits zero at x = 6, which shows, as one might expect, there are very few jobs in trades and services requiring high-level research skills:

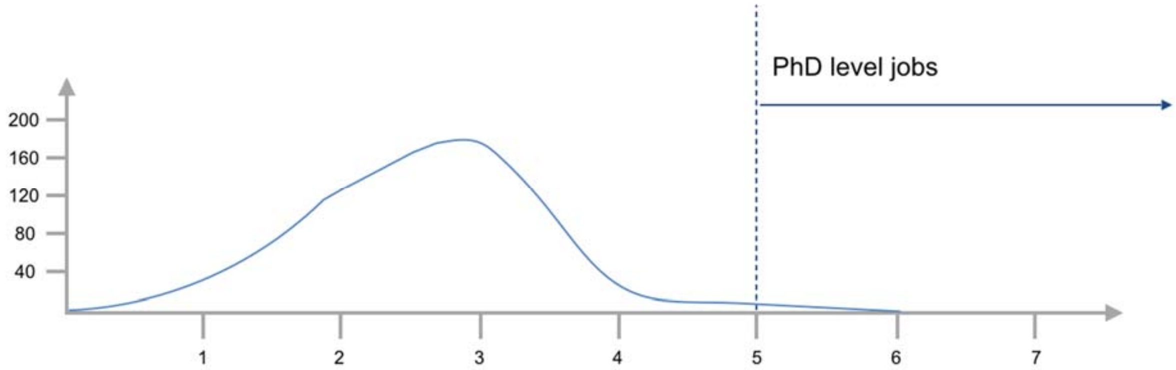

Figure two: distribution of 'PhD shaped jobs' in trades and services from 2015 seek.com.au dataset

Similar industries generated broadly similar histogram curve shapes, increasing confidence in the accuracy of our ML-NLP tools. For example, while 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. Similarly, knowledge intensive sectors such as "education and training" and "science and technology" show striking similarity in curve shapes

Discussion

If it is true that most jobs are never formally advertised online, our research suggests universities are currently under-supplying the Australian workforce with PhD candidates as our analysis showed that the number of ‘PhD shaped’ jobs was roughly equivalent to the graduating cohort who had the right to work in Australia that year. Assuming 40% of these people would take a position in academia, there is likely to be a shortfall of research skilled people, year on year. This high level of demand is counter to a dominant narrative, common in many pessimistic news reports, that there is a ‘glut' of overqualified PhD graduates with nowhere to go (see for example "The disposable academic" in The Economist 2010). This mismatch might be going unnoticed because employers have low levels of awareness (or perhaps appreciation) for the PhD as a qualification. Our research shows that 80% of job ads looking for employees with high levels of research skills did not mention the PhD as a qualification. While this is good news for universities, a note of caution should be sounded. We might not be producing enough, but are we producing the right kind of graduates? To answer this question we need to start looking more carefully at the ads themselves and the distribution of jobs on the research intensity spectrum.

In some cases, the broad field of education recorded for graduating candidates appears to align with industry categories used by Seek.com.au. Comparing the number of jobs to the number of graduates in that field can provide an overall view of research graduates relative to positions requiring research skills. For example, around 900 domestic candidates graduated with a research doctorate in health-related discipline in 2015. Our algorithm calculated there were close to 1000 non-academic healthcare and medical jobs requiring high levels of research skills. If 40% of these graduates are likely to end up in academia, this leaves a shortfall of approximately 300 people. However, this analysis may be too crude to be useful. While it is tempting to ‘map' each graduate field of education industry sectors and vice versa, it would 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.

The representations generated by our pilot online demonstration visualization system reveals some interesting patterns regarding demand for research skills, particularly in industries traditionally assumed to have low demand for PhD graduates. For example, Figure 10 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, which is currently undergoing significant innovation trends in the form of ‘just in time' logistics chains and 3D printing:

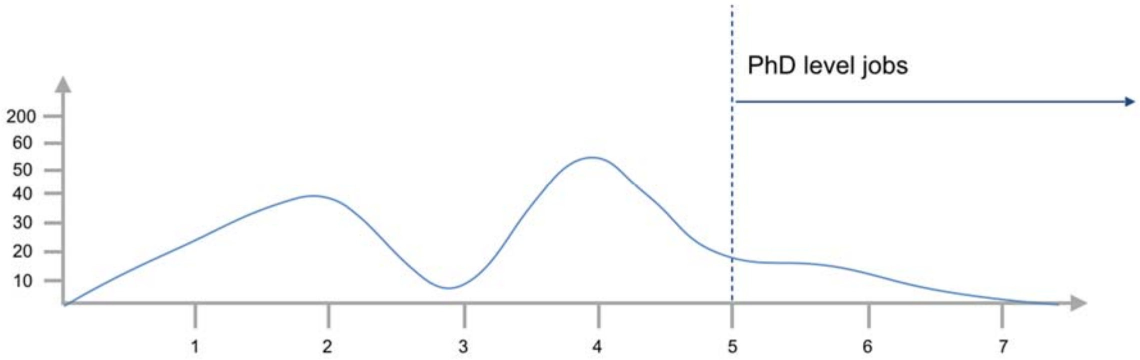

Figure 3: distribution of jobs in Seek.com.au dataset for "manufacturing, transport and logistics"

The 'hump' closest to the y-axis is truck drivers and other employees who do not require research skills to do their jobs. The hump in the middle is largely composed of office positions and the hump on the far right, past x = 5 are jobs requiring high levels of research skills. Although, as previously noted, low-intensity jobs may be under-represented in the seek.com.au data, the analysis of "manufacturing, transport and logistics" does appear to reflect an industry undergoing transition. Our diagram is suggestive of so-called 'digital disruption' trend being coupled with a higher demand for skilled knowledge workers, including those with research skills. A review of job titles in manufacturing, transport and logistics 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. The job titles we found at the high research skills end of the spectrum (eg: 'Compliance and reporting analyst', 'Head of development', 'Operations and maintenance business intelligence', 'Operations and supply chain management', 'Manager of strategic market analytics') are suggestive of high level operations jobs.

We could hypothesise that the distribution of low and high research skills intensive jobs reflect how one sector might be responding to changes in technology. Researchers are useful workers who can potentially find novel ways to make the transport industry more efficient and profitable, so as more technology is introduced, we could expect to see demand for these workers increase. ML and big data approaches are both valuable and efficient for longitudinal research. While it is time-consuming to design the algorithm the first time, once this work is done, the analysis can be performed at lower cost, year after year, to track emerging trends. 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" (orange) showed an even higher demand for research skills than "science and technology" (pink), see Figure 4, below:

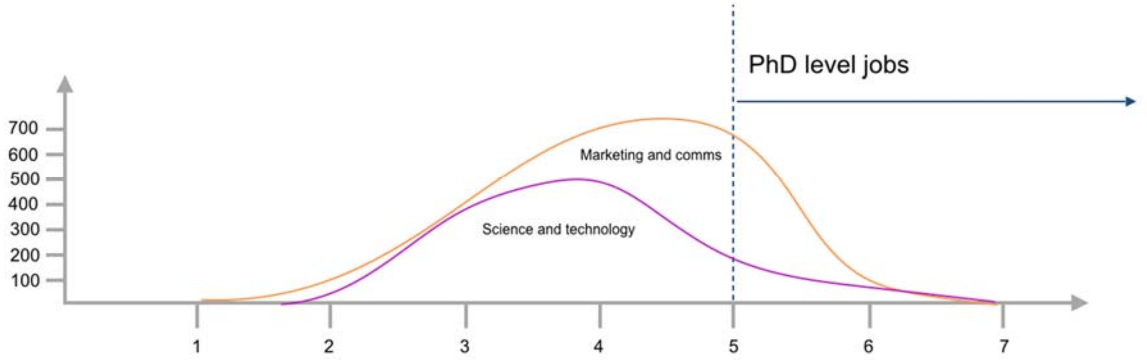

Figure 4: comparison of distribution of jobs in Seek.com.au dataset "marketing and communications" and "science and technology

A closer examination of job ad text in the 'marketing and communications' set shows high demand for people with quantitative data analysis and skills in statistical programming tools (such as 'R'). These new jobs are in addition to the kind of creative image and text creation skills normally associated with this industry. A scan of the difference in job titles between those below and above the x = 5 cut off is suggestive of this new 'layer' of data science in a traditionally 'human centric' industry. While less research-intensive jobs are suggestive of people focussed roles ('Community Engagement', 'Marketing Officer', 'Media & Communications Coordinator', 'Marketing & Communications Coordinator', 'Schools and Communities Engagement Manager'), jobs above the cut off were suggestive of 'data science' driven roles: 'Innovation consultant', 'Senior Quantitative Account Manager', 'Market research and insights manager', 'Qualitative account director', 'Market research and insights manager').

These snapshots of the evolving demand for research skills perhaps raise more questions than they answer. Are we training the right kind of PhD graduates? How, if at all, does industry demand map onto academic disciplines and student load? Given access to more data sets, our method could be used to ask broader questions, such as: 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? Our work in this area is ongoing.

Conclusion

The ML-NLP tools and the pilot online demonstration visualization system developed by this project is a significant advance on currently available forecasting tools for graduate employability. Our analysis makes visible the desire for research skills across different industry sectors, where these jobs are located, and what incomes can be expected by successful applicants. In its present form, the pilot online demonstration visualization system enables us to perform further analysis of the otherwise 'hidden' jobs suitable for PhD graduates. Once specific job ads have been identified as 'PhD shaped' they can be subjected to further forms of analysis, which can highlight the specific researcher skill sets required by various industries.

This type of ML analysis enables an exploration of 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. Since geographic data is included in the Seek.com.au data set, the visualisation system interface enables policymakers to get a better understanding of high research skill job opportunities by region. Such analyses could be used for strategic planning of initiatives that target growth, specifically in regional and remote Australia, a boon for universities located in the regions which wish to add value to their local economies. Using ML tools within search engines can assist in the employer awareness problem around the skills and capabilities of PhD graduates; if we can identify the industry sectors that are most and least open to knowledge transfer with the academy, we can better target our awareness initiatives.

If mobility between academia and industry could be encouraged, PhD graduates could be making a greater contribution to the global economy. Pittayachawan, Macauley and Evans (2016) found candidates were likely to have a ‘hidden' multi-disciplinary capacity to carry out research. This finding suggests that candidates even in traditional disciplines like mathematics or history, could potentially find employment in a range of different occupations that have little to do with their core subject matter knowledge and disciplinary knowledge. So-called 'transferable skills' training has been introduced in most universities to help candidates adjust and thrive in academic settings. There is much potential in to extend these programs to address employability outside academia too. MacAlpine and Mitra (2015) found that there were multiple sites of doctoral candidate learning and that candidates showed much agency in how they arranged their learning environments and experiences. With the appropriate information, candidates might choose to explore more potential employers. With knowledge of what these job ads are asking for, graduates could choose to take up more of the diverse learning opportunities while they undertake their PhD, rather than focussing exclusively on the dissertation. For example, historians might take advantage of classes in programming languages, and mathematicians might take up some business studies.

The recent release of the ACOLA review (2016) and the implementation of a working party to inform government puts a sense of urgency behind efforts to reform PhD curriculum in Australia. A lack of evidence hinders this work. Without data, universities rely on advice from industry partners. While employer views are valuable, this approach does not scale. It would be impossibly time consuming to survey and gather data from the huge range of employers that our graduates could end up working for - but this data does exist in the job ads, which reflect employer wishes and desires. Using new tools like this, we can perform curriculum audits to better align our teaching and learning activities with current industry demand and be more responsive to change. Armed with data, educators can identify emerging trends in demand for research skills to inform course design and planning around student load. Our machine analysis can help research students identify and develop in-demand skill sets and capabilities which make them employable in a range of industry sectors outside academia. Our research has demonstrated it is possible to generate benchmarking data for researchers interested in labour force dynamics to enable cross-sector and international comparisons of industry disruptions and the demand for research skills. We are currently exploring how to make this analysis available as a data service that the various stakeholders we have identified can use, on demand, to make strategic and informed decisions about the future of PhD study in Australia, and perhaps beyond.

References

Allen Group (2010). Employer Demand for Researchers in Australia. Report. Dept. of Innovation, Industry, Science and Research. https://trove.nla.gov.au/work/37558865?selectedversion=NBD45799384. Accessed March 2018.

Group of Eight Universities (2013) The Changing PhD: demand and supply. Report. https://go8.edu.au/sites/default/files/docs/the-changing-phd_final.pdf. Accessed March 2018.

Coates, H., & Goedegebuure, L. (2010). The real academic revolution. LH Martin Institute, Melbourne. Report. www.lhmartininstitute.edu.au/userfiles/files/research/the_real_academic_revolution.pdf Accessed March 2018.

Dale, E. (1930). The Training of Ph.D.'s. *The Journal of Higher Education*. 1(4), 198–202.

The Economist (2010) This disposable academic. Article. https://www.economist.com/node/17723223 Accessed March 2018.

Edwards, D. (2010). The future of the research workforce: estimating demand for PhDs in Australia. *Journal of Higher Education Policy and Managemen*t. 32(2), 199 – 210.

Forsyth, H. (2014). *A history of the modern Australian University*. Sydney: New South Press.

Vitae. (2013). What do researchers do? Early career progression of doctoral graduates. Vitae (UK). Report. https://www.vitae.ac.uk/vitae-publications/reports/what-do-researchers-do-early-career-progression-2013.pdf. Accessed March 2018.

London. McAlpine, L., & Mitra, M. (2015). Becoming a scientist: PhD workplaces and other sites of learning. *International Journal of Doctoral Studies*. 10, 111-128.

McGagh, J., Marsh, H., Western, M., Barber, M., Franzmann, M., Gallois, C., et al. (2016). Securing Australia's Future: Review of Australia’s Research Training System. Australian Council of Learned Academies (ACOLA). Report. https://acola.org.au/wp/PDF/SAF13/SAF13%20RTS%20report.pdf . Accessed March 2018

Mewburn, I., Grant, W. & Suominen, H. (2017) Tracking trends in industry demand for Australia’s advanced research workforce, Centre for the Public Awareness of Science, ANU: http://cpas.anu.edu.au/files/Mewburn%2C%20Suominen%20and%20Grant%202017%20Tracking%20Trends%20in%20Industry%20Demand%20for%20Australia%27s%20Advanced%20Research%20Workforce.pdf Accessed March 2018.

Pitt, R., & Mewburn, I. (2016). Academic superheroes? A critical analysis of academic job descriptions. *Journal of Higher Education Policy and Management*. 38, 1–14.

Pittayachawan, S., Macauley, P., & Evans, T. (2016). Revealing future research capacity from an analysis of a national database of discipline-coded Australian PhD thesis records. *Journal of Higher Education Policy and Management*. 38, 1–14.

Platow, Michael J. (2012). PhD experience and subsequent outcomes: a look at self-perceptions of acquired graduate attributes and supervisor support. *Studies in Higher Education*. 37(1),103-118.

Probert, B (2014) Becoming a university teacher: the role of the PhD. Office of Learning and Teaching, Australia: http://www.olt.gov.au/resource-becoming-university-teacher-role-phd-2014 Accessed march 2018.

Schwab, K., & Sala-i-Martin, X. (2016). World Economics Forum: The Global Competitiveness Report 2016–2017. Geneva. Report. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjsx-vtpP_ZAhVIwFQKHTVnDS4QFggpMAA&url=http%3A%2F%2Fwww3.weforum.org%2Fdocs%2FGCR2016-2017%2F05FullReport%2FTheGlobalCompetitivenessReport2016-2017_FINAL.pdf&usg=AOvVaw2TPNLCIPzZgRtaPA7DvXsp. Accessed March 2018

Acknowledgements

The project team would like to thank the Australian Department of Industry and Seek.com.au for their generous support of this project. The team would like to acknowledge the intellectual input and assistance of Professor Alysson Holbrook, Dr Margaret Kiley, Mr Nigel Palmer and Dr Rachael Pitt. We would also like to thank Dr Kailing Shen for initial comments on this article. Finally, special thanks to Dr Lindsay Hogan for her invaluable assistance with IP issues arising from this project.