Literature Review

Bhargav Srinivasa Desikan

May 2019

1 Introduction

In what ways are modern skills and knowledge created and forged for the 21st century data economy? Is the primary crucible the academic arena (skill discovery), the industrial domain (skill invention), or the educational space of teaching (skill translation)? This complex interplay of industry, academic research and teaching in the context of what trends emerge in the skill marketplace has not been quantitatively studied, and certainly not with the context of massive textual databases.

We have seen that academic reflection and strategic or accidental investigation yield insights and the most successful (e.g., topic modeling Blei and Lafferty [2009], Western blot analysis Mahmood and Yang [2012]) harden into theorized protocols and skills deemed valuable in the marketplace.

Alternatively, industry experimentation and refinement yield patterns that generate value and demand greater theoretical insight (e.g., neural word embeddings Mikolov et al. [2013], plant genetic transformation). We also see that the act of translating and transmitting skills through educational instruction helps to encapsulate and crystallize them.

Either can influence the other and this is reflected in the modern history of invention.

In this literature review I will start by discussing some of the existing literature linking these fields together before moving on to the literature regarding the machine learning and textual analysis tools we will be using.

2 Literature on Knowledge and Skill Formation

Early literature on the study of relationship between academia and industry (and innovation in them) often studies particular discoveries, or spoke of knowledge creation or scientific revolution as a whole, and less about the organisational or institutional relationships. Roberts [1988] speaks of how invention has been previously approached in the social sciences or study of science, and offers his own view on how invention or innovation can be measured or talked about. The interplay between market tensions, the customer, and scientific discovery is poked at, and also discussed the importance of management of these resources for the production of more 'useful' science.

Etzkowitz and Leydesdorff [2000] studies what they call a "triple helix", with how university-industry-government influence the dynamics of innovation. It is a qualitative discussion of the interplay between these three forces and they posit that they are not related in traditional relationship models but instead form a complex helix with the three pulling and pushing on each other.

Carlsson et al. [2009] provides a historical review on the role of universities, particularly engineering schools and land-grant universities, industrial laboratories, and corporate research and development (RD) laboratories prior to World War II, and after. It offers a glimpse into how periods of economic growth can lead to innovations in science. Lamoreaux and Sokoloff [1999] is another landmark paper discussing the history of science and innovation in the United States of America, and discusses how the patent system and the market of technology influenced how innovation happens. Lamoreaux and Sokoloff [2005] also discussed how entreprenurial innovation fits into this.

Patents start to become an important way to measure innovation - you can count them, you can see who collaborates with whom, and the content of the patents. Previous research on linking invention done in cities to city size and growth (Bettencourt et al. [2007a], Bettencourt et al. [2007b]) does precisely this, and are both very important examples of how we can start to quantitatively measure innovation.

More recent publications from Sociology include Evans [2010], where he has also discussed how academia and industry interact with a case study (of the Arabidopsis thaliana plant), and other work by Foster et al. [2015] shows us a large scale data analysis of papers from MEDLINE and discusses

possible policy interventions to increase innovation. Wu et al. [2017] discusses innovation in the context of the size of teams, in both academia and industry, as well as on online code collaboration platforms.

Building on older qualitative research and gathering hints from more recent quantitative research will help me to formulate a way to be able to measure and talk about innovation in the context of academia and industry.

3 Literature on Methods

I will be using a wide variety of data sources and methods to be able to ask these questions. The Web of Science Reuters [2012] dataset includes metadata of research articles published from the year 1900, a syllabus data set which includes details of syllabi of courses being taught at universities worldwide, and details of job postings across all major American cities. I will be receiving all three of these datasets through the Knowledge Lab at University of Chicago. The details and metadata of these datasets are available on request.

I plan to use machine learning techniques applied to text to understand how different universities, cities and companies produce knowledge, how this relates to what is currently being taught at universities, and how both of these may be related to the demand for skilled workers by companies. In particular, I will use Topic Modelling (Blei and Lafferty [2009]), and Word2Vec Mikolov et al. [2013]) to represent the words as vectors to compare between the various cities and organisations.

These methods are often popularly used in textual analysis of documents. Evans and Aceves [2016] is a good review of the kind of textual analysis tools one can employ in the social sciences.

Other standard mathematical tools from linear algebra and machine learning will be used via the wonderful NumPy (Van Der Walt et al. [2011]) and scikit-learn (Pedregosa et al. [2011]) toolkits.

References

Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007a.

Luis MA Bettencourt, Jose Lobo, and Deborah Strumsky. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research policy*, 36(1):107–120, 2007b.

David M Blei and John D Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

Bo Carlsson, Zoltan J Acs, David B Audretsch, and Pontus Braunerhjelm. Knowledge creation, entrepreneurship, and economic growth: a historical review. *Industrial and Corporate Change*, 18(6):1193–1229, 2009.

Henry Etzkowitz and Loet Leydesdorff. The dynamics of innovation: from national systems and "mode 2" to a triple helix of university–industry–government relations. *Research policy*, 29(2):109–123, 2000.

James A Evans. Industry induces academic science to know less about more. *American Journal of Sociology*, 116(2):389–452, 2010.

James A Evans and Pedro Aceves. Machine translation: mining text for social theory. *Annual Review of Sociology*, 42:21–50, 2016.

Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80 (5):875–908, 2015.

Naomi R Lamoreaux and Kenneth L Sokoloff. Inventors, firms, and the market for technology in the late nineteenth and early twentieth centuries. In *Learning by doing in markets, firms, and countries*, pages 19–60. University of Chicago Press, 1999.

Naomi R Lamoreaux and Kenneth L Sokoloff. The decline of the independent inventor: a schumpterian story? Technical report, National Bureau of Economic Research, 2005.

Tahrin Mahmood and Ping-Chang Yang. Western blot: technique, theory, and trouble shooting. *North American journal of medical sciences*, 4(9): 429, 2012.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Thomson Reuters. Web of science. 2012.

Edward B Roberts. What we've learned: Managing invention and innovation. *Research-Technology Management*, 31(1):11–29, 1988.

Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.

Lingfei Wu, Dashun Wang, and James A Evans. Large teams have developed science and technology; small teams have disrupted it. 2017.