May 2016
No.2016-010-e

Journal of Integrated Creative Studies

The Science of Data Science

Jevin D. West[1]
[1]Information School, University of Washington
E-mail: jevinw@uw.edu

Abstract. This report summarizes two talks that I gave at the Advanced Future Studies at Kyoto University in February of 2016. One talk was for the Global Partnership on Science Education through Engagement. In this talk I focused on an emerging educational trend in the United States—the rise of Data Science at both the undergraduate and graduate level—and the effect it is having on research and industry. In the second talk, I spoke at the International Symposium on Advanced Future Studies symposium. In this talk, I provided an overview of an emerging research trend—the emergence of a new discipline called the Science of Science. In this new field, science is done at the level of millions of publications over many generations and disciplines using new tools from machine learning, computer vision, and network science. Both Data Science and the Science of Science require perspectives from multiple disciplines, which fit well with the general theme of both meetings in Kyoto.

Keywords: Data Science, Multidisciplinarity, Science of Science, Scientometrics, Citation Networks

1. Multidisciplinary Perspectives

Some of the world's most pressing problems in the environment, health, and economics will require conversations across disciplinary boundaries. Facilitating these multi-perspective discussions is one of the goals of the Advanced Future Studies program at Kyoto University. Its founder, Professor Masatoshi Murase, has been speaking for many years about the need for new forms of creativity that address big problems.[1] The conferences, including the one I attended this year, were Frenkel-esque[2] (Frenkel, 2013) in that attendees included mathematicians, physicists, biologists, psychologists, and even musicians.

In this report, I will summarize the two talks[3] I gave this year at the International Symposium on Advanced Future Studies symposium and the Global Partnership on Science Education through Engagement (GSEE) symposium and conclude with a couple suggestions on building multidisciplinary communities[4]. For the first symposium, I talked about the emergence of *Data Science* in research and education. In the second talk, I focused on the *Science of Science* -- another emerging discipline that studies the scientific enterprise itself and mines for hypothesis, not from single papers, but from millions of papers. Both emerging disciplines are a bit like the Advanced Future Studies; they are both new and require multiple areas of expertise.

2. Science of Science

In my lab (DataLab), we study the Science of Science – an emerging discipline that puts the scientific enterprise under the microscope. We examine science, not at the level of an individual paper or author, but at the level of millions of papers and hundreds of thousands of authors. Within this data are the scientists, journals in which they publish, the citations to other papers, and the funding agencies that support their work. We can ask questions about the origin of ideas, mechanisms that spur innovation, and incentives that promote the kind of science for solving big problems. We can also mine this data[5] for new hypotheses and patterns that only emerge when examining millions of papers and figures in aggregate rather than as individual objects.

I mentioned several example questions in this area. For example, how do we identify, measure, and visualize the origin of ideas and disciplines? How are metrics (impact factor, h-index) affecting science positively and/or negatively? What funding policies lead to better science? Can we build a better gateway to the scholarly literature than what is afforded by Google Scholar and other standard text-base search engines? These questions deal with what I consider "knowledge science" or "knowledge engineering". I see a synergy in working in both areas. What we learn from the sociology of science can tell us a lot about how to build better recommendation algorithms for navigating the literature influenced by that sociology.

[1] I attended two previous conferences in Kyoto: "What is Evolution?" and "What is creativity?".

[2] Edward Frenkel is a University of Berkeley mathematician who is known for writing "Love and Math" \cite{frenkel2013love}, a book that explores mathematics with references to art and music.

[3] The slides for two talks can be found at these two links:
http://www.jevinwest.org/presentations/data_gold_rush_jevin_west.pdf and
http://www.jevinwest.org/presentations/mapping_science_kyoto_jevin_west.pdf.

[4] There are distinctions between multidisciplinarity, interdisciplinarity and transdisciplinarity (Klein 1990). Here I will mostly refer to multidisciplinary, which involves researchers from multiple fields that apply their respective methodologies and ideas for solving a particular question. This is different than transdisciplinary, which tries to transcend the respective disciplinary, and interdisciplinary, which tries to integrate ideas from one discipline into another.

[5] It is just recently that the papers have been available in machine readable forms at this scale.

My colleagues and I have developed a set of algorithms that help us to map domains of science at the scale of millions of articles. This includes ranking algorithms on citation networks (West, 2010), community detection algorithms (Rosvall, 2008) and recommendation algorithms (West, 2016). These tools[6] can be used both for knowledge science and engineering tasks.

Figure 1 is a case study in how some of these tools[7] are used. The figure shows, what appears to be, the emergence of neuroeconomics – a discipline at the cross section economics and neurology – using large scale citation data. Economists are interested in the empirical data that can be collected from brain imaging experiments and neuroscientists are interested in the models and analyses that economics use to understand human incentives and behaviour (Camerer, 2005). Some have speculated the beginnings of stand-alone field called neuroeconomics (Fischman, 2012). The citation pattern differences between 1997 and 2010 provide evidence for the beginnings of this field (Figure 1). The maps represent hundreds of journals and hundreds of thousands of citations in neuroscience and economics. In 1997, there were zero citations between the two fields among the hundreds of thousands of possible citations. In 2010, there were more than 250 citations being shared between the fields. The radial diagram shown in my talk provides a different view[8]. The inner circle shows journals within different disciplines and the lines between the blocks in the inner circle represent the in and out-citations between the fields. There exists a thin line connecting the *Finance Journal* and *Neurology*. These data and tools provides examples of what can be tracked and measured when looking just at bibliographic data.

[7] The maps were built using MapEquation (Rosvall, 2008) and InfoMap codebase.
[8] One can explore the data interactively by going to the wellformed project at eigenfactor.org.

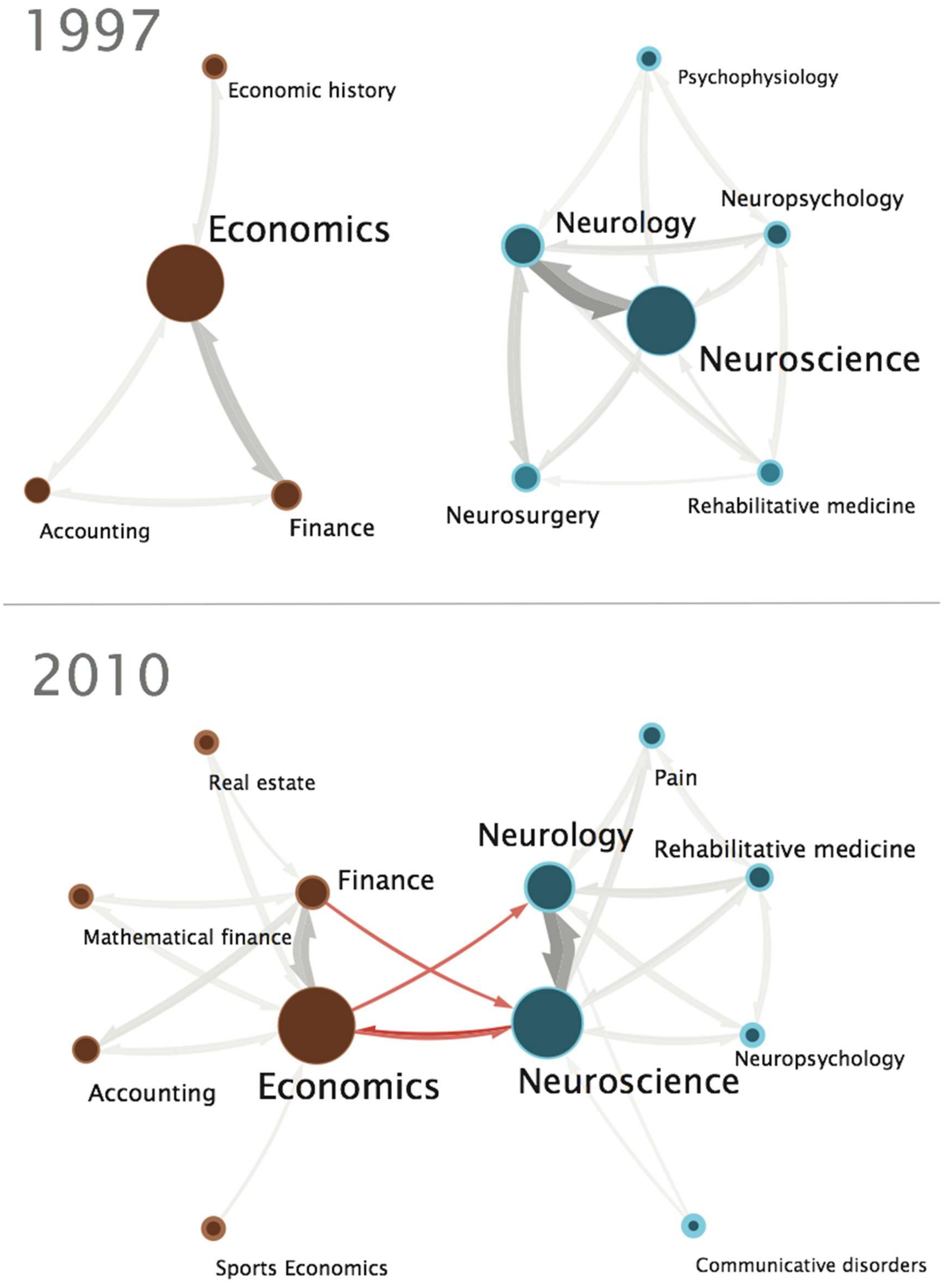

Figure 1. Emergence of neuroeconomics. The figure above looks at the cross-talk between the fields of economics and neuroscience. The journal-level citation data comes from Thomson Reuters' Journal Citation Reports. The top panel from 1997 shows no citations between neuroscience and economics.

The bottom panel in 2010, there exist 195 citations from economic journals to neuroscience journals and 74 citations from neuroscience journals to economic journals. This map was noted in a Chronicle of Higher Education article titled "The Marketplace in Your Brain." (Fischman, 2012). The figure was co-created with Carl Bergstrom at the Eigenfactor.org Project.

The origin of ideas has many roots, but their development depends on sociological factors. We are looking at some of these factors including incentives and reward structures within promotion and funding. We are also looking at the role of gender in science. In a recent paper noted in my talk, my colleagues and I examine gender differences in authorship over the last several hundred years (West, 2013). We find an increase in overall authorship of females, but do not see an equal increase at the last author position (i.e., the principal investigator in many fields). We used the JSTOR corpus as our data source. It consists over 8 million full text scholarly articles from the last several hundred years in the social sciences, ecology and evolution, statistics, molecular biology, and many others. We also found that the increases in authorship were not consistent across all domains[9].

In addition to bibliographic data, we can also utilize the full text, equations, and figures within scientific papers. Up until recently, there have been barriers to this kind of data. Some of this has been technological reasons (e.g., optical character recognition limits at scale), but the primary reason have been publisher paywalls. Publishers have not allowed access to full text articles in bulk. This has changed a bit in the last decade with open access journals, the NIH Mandate (English, 2008) and national repositories such as PubMed Central, and archivists like JSTOR. We have been utilizing various aspects of this full text. This includes the gender project mentioned above, but also mapping jargon differences across different domains and mining figures in the biomedical literature[10] in my talk. This is a project where we mine the figures for patterns in the biomedical sciences. We find that high impact papers tend to have higher density of diagrams and schematics. We have also build a figure-centric search engine for better exploring visual information in the literature (Lee, 2016). Using standard information theory (e.g., cross entropy), we develop a set of methods for measuring differences in language across the landscape of science (Vilhena, 2014). We find that evolutionary biology and ecology tend to be more balkanized where the social sciences use similar language across the different domains. Figure 2 is an example jargon map that I showed in my presentation.

[9] One can explore the data interactively with the gender browser.

[10] I mentioned the Vizometrics project in my talk.

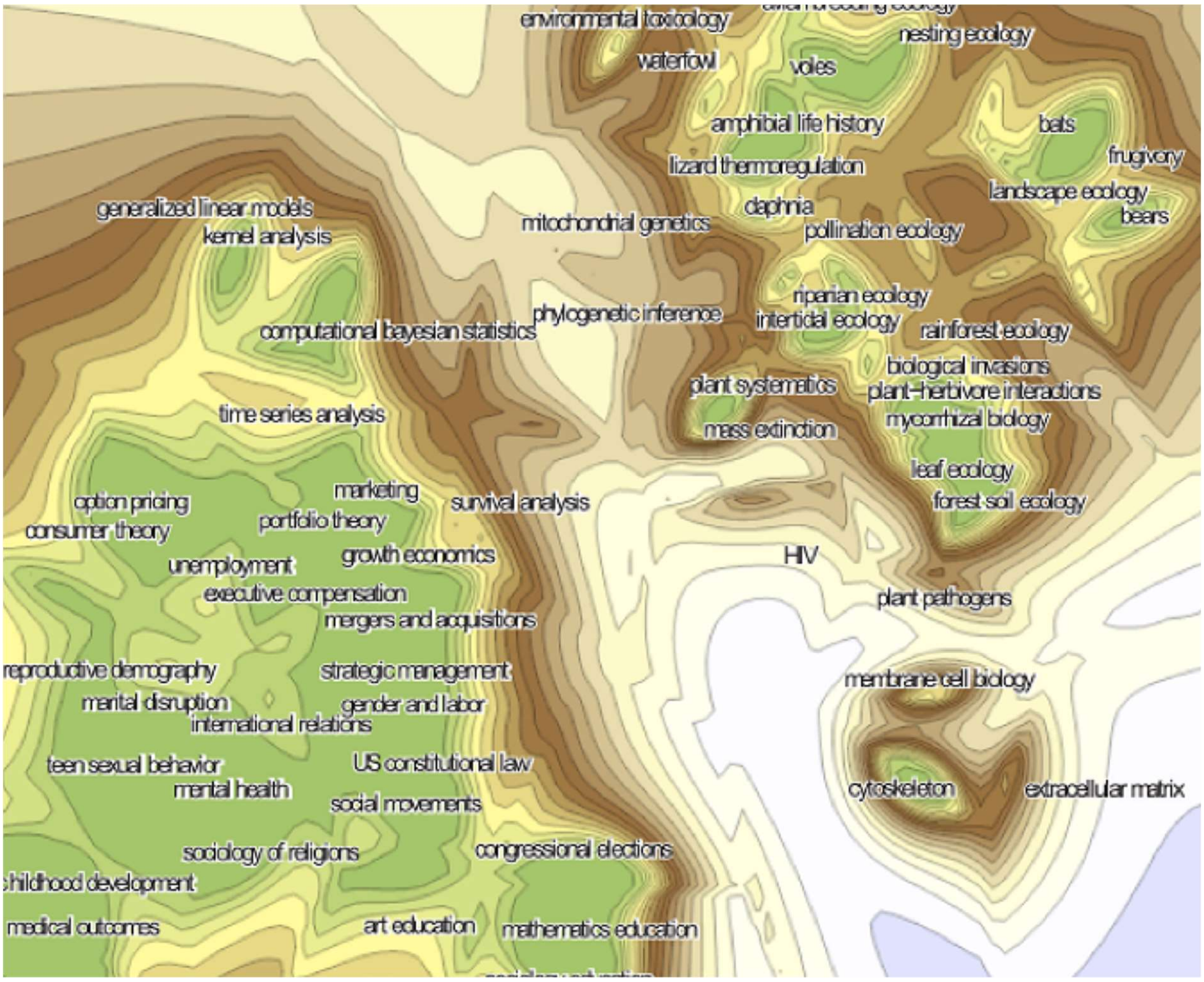

Figure 2. Jargon map of science. The map looks at the differences in jargon among fields in the social and biological sciences. The full text data comes from the JSTOR corpus. The distance between fields is determined by InfoMap (Rosvall, 2008) and the heights (isoclines) in the topographical maps are determined by the jargon differences between neighbouring fields. More details of the methods can be found in the paper by Vilhen et al (2014). The figure noted here was created with Daril Vilhena and Carl Bergstrom at the Eigenfactor.org Project. This was an early, inverted version of the map before the published maps at Sociological Science (Vilhena, 2014).

Lessons learned from studying the sociological factors governing science can be applied to knowledge engineering questions. For example, can we build a better interface for navigating the scholarly literature? To address this, we are building recommendation engines and new interactive visualizations based on the social connections between authors and papers. One of the goals of our methods is to locate classic papers for graduate students or scholars moving into new disciplines (West, 2016). The recommendation project is called Babel. Here scholars can search for related papers, but more importantly publishers can connect to the service and provide recommendations to their big audiences without having to construct a complicated search engine infrastructure. The overall goal is to help scholars find papers they never knew they always needed and to help publishers deliver these new kinds of tools from data science to their users.

The establishment of the field is also indicated by new sections of high profile journals. For example, PLoS Biology now has a new section called "Meta-Research" that encourages research in the Science of Science (Kousta, 2016). This is driven by the need for improving reproducibility in science and the need to better define statistical significance. One of the most downloaded articles in all PLoS journals is a paper published by Ioondinis, titled "Why are all research false" (Ioannidis, 2005). This speaks to the need to these issues of reproducibility and the utility of studying science at the level of the enterprise itself.

3. Science of Science

My first talk during my Kyoto visit was at the Global Partnership on Science Education through Engagement. In this talk, I focused on the rise of Data Science[11] in both research and education. Over the last several years, Data Science programs have sprouted on campuses and companies across the globe (West, 2016b). My university is no exception. At the University of Washington (UW), we have deployed several new programs in Data Science at the undergraduate, masters and PhD programs. We also house the eScience Institute to meet the data science needs on our campus. This has been buoyed by the Moore-Sloan $40 million grant to support data science for science at UW, University of California at Berkeley and New York University. The coordination efforts in education have required department heads from statistics, computer science, information sciences, HCDE, Sociology, Biology, Oceanography, and Astronomy to come together in one room for a common purpose—a bit like the Kyoto meetings. One of the goals for the Moore-Sloan program is to create a model that other universities can look to.

This data science trend is not unique to the University of Washington. We are seeing massive proliferation across the United States and the world. This is mostly in response to the increased demand for quantitative talent within industry and in reference to reports like the McKinsey report on big data (Manyika, 2011). In the US, there has been a surge of new programs in Data Science (West, 2016b). In 2007, there were about 4 programs with the name "Analytics" but this has changed to Big Data and now Data Science. In 2015, there are scores of programs at universities across the country. One would be hard pressed to find an area of science that has grown as fast in such little time.

Data Science by definition is interdisciplinary and requires not just method disciplines (statistics, computer science, mathematics, information sciences); it also requires the domain sciences. We encourage in our program project-based learning, data ethics training, and domain expertise. This data science movement is a perfect example of how researchers from different disciplines can come together to solve problems that require more than just one's knowledge of their own discipline. In our programs at the eScience Institute at UW, we are emphasizing the need to train students in multiple disciplines. We want our students to be pi-shaped; want them to have a 'foot' of training and expertise in multiple domains. It doesn't so much matter what they are. We just want them to be able to compare, contrast and connect ideas from multiple disciplines. If one only has training in one discipline, then the student does not have the ability to see the assumptions and limitations of methods in their field as well. Plus, the world is changing so fast that we want students to speak multiple languages in science.

4. Disciplinary Risk

[11] The name of this area has changed many times throughout the last five years. Big Data held for some time, but we are seeing a convergence on "Data Science"' at universities and within industry. This is the term we will use in this paper.

During my visit, I spoke about two emerging fields of study—the Science of Science and Data Science. Both require expertise and skills from multiple fields. The need for transdisciplinarity in research and education has become a common theme, which is why we need institutes like the Advanced Future Studies program at Kyoto University. We need institutes willing to take this disciplinary risk to bring researchers together on common themes and big questions (e.g., What is creativity in Science?). There will be communication challenges across disciplines, but big problems like climate change will require researchers to step outside their discipline, listen to other ideas, and borrow methods from other domains.

The disciplinary challenges are real. Creating an institution that facilitates productive discussions between a high energy physic and game theory economist or an educational psychologist and an evolutionary biologist will require changes in how researchers communicate ideas and findings. A norm of communication for interdisciplinarity meetings needs to be established. Details at meetings like the Kyoto meeting are only as important as their contribution to the bigger ideas. It is important to convey the larger concepts and to actively draw connections to the disciplines represented in the audience. This may require a set of 'rules' at interdisciplinary meetings. The rules could be as simple as 'big picture' ideas only. The Advanced Future Studies program at Kyoto University could play a role in changing this culture of communication. For example, an economist may use Regression Discontinuity Design (RDD) to identify causal elements but this may also be useful for policy makers who want to elucidate the impact of various funding programs. Understanding the basic goals of this method are what are important, not all the caveats, limitations, and Greek symbols in describing the method.

In sum, there is a strong need for researchers to come together from different disciplines. It works especially well for me because I study how scientists communicate across different fields and how education institutions experiment with their infrastructures. The world's big problems are too complex for one discipline to solve. The Advanced Future Studies at Kyoto University is trying to do this. Putting together symposia with researchers from fields as diverse as condensed matter physics, educational psychology, and the science of science is difficult. It will take time to figure out how to make this work effectively, but it is good to know that there are programs and universities willing to take this risk. Solving hard problems requires breaking down disciplinary boundaries and taking risks. I hope to see the Advanced Future Studies succeed and provide a model for other institutions.

5. References

Camerer, C., Loewenstein, G., and Prelec, D., Neuroeconomics: How neuro- science can inform economics. *Journal of economic Literature*, pages 9–64, 2005.

English, R., The NIH mandate: An open access landmark. *College & Research Libraries*, 2008.

Fischman, J., The marketplace in your brain. *Chronicle of Higher Education*, 2012.

Frenkel, E.*, Love and math: The heart of hidden reality.* Basic Books, 2013.

Klein, J.H.*, Interdisciplinarity: History, theory, and practice*. Wayne State University Press, 1990.

Kousta, S., Ferguson, C., and Ganley, E., Meta-research: Broadening the scope of plos biology. *PLoS biology*, 14(1), 2016.

Ioannidis, J., Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

Lee, P., West, J.D., and Howe, B., Viziometrix: A platform for analyzing the visual information in big scholarly data. In *Proceedings of the 25th International Conference on World Wide Web. ACM*, 2016.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., Big data: The next frontier for innovation, competition, and productivity. 2011.

Rosvall, M. and Bergstrom, C.T., Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118– 1123, 2008.

Vilhena, D., Foster, J., Rosvall, M., West, J.D., Evans, J. and Bergstrom, C., Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1(June):221–238, 2014.

West, J.D., Bergstrom, T.C., and Bergstrom, C.T., The eigenfactor metrics: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3):236–244, 2010.

West, J.D., Jacquet, J. King, M.M., Correll, S.J. and Bergstrom, C.T., The role of gender in scholarly authorship. *PloS One*, 8(7) :e66212, 2013.

West, J.D., Wesley-Smith, I. and Bergstrom, C.T., A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, 2016.

West, J.D., and Portenoy, J., The data gold rush in higher education. In C.R. Sugimoto, H. Ekbia, and M. Mattioli, editors, *Big Data is Not a Monolith*, chapter 6. MIT Press, 2016b.