Data Standardization and Entry

Min Ren

PhD candidate
Management and Strategy
Kellogg School of Management
Northwestern University
m-ren@kellogg.northwestern.edu

Cumulative innovation is thought as a driving force of economic growth (Romer 1990; Jones 1995). In order to build upon previous knowledge, however, scientists need to devote time and effort to understand the details and contributions. I define the costs involved in acquiring such prior knowledge as access costs. If there is a reduction in access costs, scientists are more likely to enter certain research fields.

A particular form of access costs comes along with the challenge of big data. The increasing use of digitized information has affected scientific research, especially life sciences research. We have generated enormous data resources in terms of genomic information since the Human Genome Project, and we think these enormous data offers lots of potential for innovation. However, a barrier for using these data is that many of these data are poorly documented. It is difficult for other researchers to understand the data and build new work based on the data. Therefore standardized data reporting can be potentially useful in this scenario. Standardized data reporting makes it easier to understand the data, and the increase in data transparency may encourage more scientists to enter related research fields. I propose to study the impact of data reporting standards on the entry into research fields.

Understanding how data standardization affects entry into is crucial to government funding of scientific research. We spent a lot of money in generating the genomic data, both from the public sectors and private sectors. It is important to make sure these data are easy to use, and to understand the effect of standardization on entry into scientific fields. Currently, most of the public finding goes to new research projects that possibly generate more new data, instead of standardizing existing data. Therefore understanding the trade-off between funding for new projects and funding for data standardization is very important to policy, and my proposal on data standardization can help us think about this question.

The paper builds on the strand of recent papers that documents the relationship between access costs and cumulative knowledge production. First, Furman and Stern (2011) emphasize the impact of institutions on knowledge accumulation. They examine a

biological resource center that reduces access costs by certifying, preserving, and providing access to standardized biological materials. They find an amplification in cumulative knowledge production. However, they do not separate the effect of standardization from the multiple functions of the institution. Second, Murray et al. (2009), Galasso and Schankerman (2013) and Williams (2013) discuss the effect of intellectual property rights on subsequent innovation, and they stress the crucial role of openness in knowledge accumulation. Data reporting standards could play a similar role in making data more open to third-party investigators. Finally, Agrawal and Goldfarb (2008) study the effect of a decrease in collaboration costs resulting from the adoption of Bitnet on university research collaboration. Similar to Bitnet, data reporting standards may be another way to reduce communication costs between researchers.

I plan to use the introduction of Minimum Information About a Microarray (MIAME) as the identification following Ren (2013). MIAME is a data reporting standard developed by the Functional Genomics Data (FGED) Society in 2001. It specifies the minimum information required to describe a microarray experiment. The end goal of MIAME is to ensure that every researcher can interpret the experimental results in an unambiguous way. Starting at the end of 2002, some journals such as Nucleic Acids Research endorsed MIAME and required authors publishing with them to comply with the MIAME protocol. Over time, more journals adopted MIAME and the variation in the timing of adoption is mostly due to journal editors' preferences. This is the source of variation that I exploit for identification.

I plan to study whether the patterns of entry into fields are different between articles with and without standardized articles. I will use the PubMed Related Citations Algorithm [PMRA] to find the related articles to articles as in Azoulay, Furman, Krieger and Murray (2013).

References:

Agrawal, Ajay, and Avi Goldfarb. "Restructuring Research: Communication Costs and the Democratization of University Innovation." American Economic Review, 2008: 98(4): 1578-1590.

Azoulay, Pierre, Jeffery Furman, Joshua Krieger, Fiona Murray, “Retractions”, NBER Working paper, w18499.

Furman, Jeffrey L., and Scott Stern. "Climbing Atop the Shoulder of Giants: The Impact of Institutions on Cumulative Research." American Economic Review, 2011: 1933-1963.

Galasso Alberto. and Mark Schankerman. "Patents and Cumulative Innovation: Causal Evidence from the Courts." CEPR discussion paper 9458, 2013.

Jones, Benjamin F. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" Review of Economic Studies, 2009: 283-317.

Jones, Charles I. "R&D-Based Models of Economic Growth." Journal of Political Economy, 1995: 759-784.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, et al. "Of Mice and Academics: Examining the Effect of Openness on Innovation." 2009 March, NBER Working Paper 14819.

Ren, Min, "When Big Data Meets Life Sciences: Data Reporting Standards and Innovation", Kellogg School of Management working paper.

Romer, Paul M. "Endogenous Technological Change." Journal of Political Economy, 1990: S71-102.

Williams, Heidi L. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." Journal of Political Economy, 2013 121(1): 1-27.