Data-Driven Search and Innovation

Matteo Tranchero, UC Berkeley-Haas School of Business, 4th year PhD student, graduation expected in May 2024, m.tranchero@berkeley.edu

How does data shape innovation? Most of the existing literature has looked at the impact of data as an amorphous input into production (Jones and Tonetti 2020; Farboodi and Veldkamp 2020), managerial decision-making (Brynjolfsson and McElheran 2016), entrepreneurship (Bessen et al. 2021), and science (Nagaraj et al. 2020). However, innovation is far from being homogeneous in nature, and data might be affecting the direction as well as the specific type of technological knowledge generated. A full account of the role of data in innovation requires a deeper understanding of the ways in which innovators use data to search for new ideas and how this in turn shapes the nature of their realized innovations.

In this paper, I argue that data can affect novelty generation not just as direct input, but also if it leads to alternative innovation search strategies with the potential to redirect attention toward novel questions. To elucidate this idea, it is useful to conceptualize innovation as a process of recombination of technological components (Fleming, 2001). Novel combinations are usually found either by marginally tinkering with existing ones or by using insights from scientific knowledge to guide experimental attempts. Against this backdrop, I suggest that data can enable a radically new search strategy for useful recombinations, which I define data-driven search. Being unconstrained by both past experimentation and theoretical maps, data-driven search might lead to diversify research and explore novel avenues that could greatly benefit overall knowledge accumulation. I theorize that data-driven search can lead to explore a wider portion of the technological landscape. This is important in light of the natural proclivity of firms and inventors toward incremental, exploitative search that might forfeit important innovation opportunities. Moreover, I argue that searching guided by data is particularly helpful in unknown areas of the technological landscape where technological components are harder to recombine.

Investigating these set of questions is difficult because the innovation search process is usually not observable to the researcher. To overcome this challenge, I study the role of data-driven search considering a specific type of scientific study in genomics, called genome-wide association studies (GWAS), that are employed to discover novel gene-disease associations to guide follow on research and drug development. The method involves scanning the genomes from many different people and looking for genetic variations that are highly correlated with the presence of a disease. As such, these studies do not target any one gene ex ante, but rather they are hypothesis-free and their findings completely determined by the data analyzed, approximating the ideal notion of a data-driven search. In my project, I use new data on the discovery of new gene-disease associations that are established using the GWAS approach. Using OLS regressions, I compare the characteristics of data-driven findings vis-a'-vis associations discovered using a traditional candidate-driven approach.

First, I show that GWAS-established gene-disease links have different characteristics relative to those uncovered by traditional search methods. Data-driven findings are much more likely to recombine previously overlooked genes which nonetheless hold great potential for pharmaceutical research. Second, I show that novel combinations uncovered with data-driven search are especially useful to innovate in less known areas of the technological landscape where components are harder to recombine. Finally, data-driven search leads to impactful findings that however consolidate existing research paradigms instead of opening completely novel research trajectories. Taken together, these findings suggest that data-driven search strategies can have an important role in the search for innovation within a specific research paradigm.