# Machine learning could improve innovation policy

To the Editor — Understanding factors that affect the rate and direction of innovation has been a central aim of research in the study of science and innovation for more than half a century. However, substantially more progress has been achieved in understanding the factors that drive the rate of innovation than its direction — the set of research topics scientists or research institutions tackle at a given point in time (diversity) and over time (trajectory).

Economists have long understood that markets provide insufficient incentives for innovation because of the difficulty of fully appropriating the returns to innovation investments, particularly when new innovations build on older ones[1]. Similar features inhibit investment in the diversity of innovations[2,3], thus potentially impacting their trajectory. For example, investment in environmentally friendly technologies may have been inhibited by initial successes using fossil-fuel technologies and hence scarce investments in alternative technologies[3]. These insights show that more empirical research is needed to uncover what factors influence the direction of innovation.

Such research is challenging. For example, estimating changes in the direction of innovation requires the boundaries of research trajectories to be defined. However, this creates a paradox, as boundaries of research trajectories are part of the core unknown to be estimated. Recent developments in machine learning (ML) have the potential to address this limitation by helping researchers infer the structure of the knowledge space by quantifying the various research topics and the distances between them. A remaining challenge is adapting ML algorithms to the study of causal relationships, because research on the direction of innovation is interested in identifying the latent categorization of research topics in order to then identify factors that causally change this structure. This is a non-trivial difference from the core prediction purpose of ML algorithms.

We call on researchers to intensify their efforts in adapting ML algorithms to the study of the direction of innovation. Nascent efforts can be grouped in two categories: (1) off-the-shelf ML-based categorization schema developed by bibliographical data services (for example, the National Library of Medicine's PubMed Related Articles algorithm or the Microsoft Academic Graph's categorization schema[4]) and (2) customized algorithms that provide access to more granular data on similarity between corpuses of text and that can be applied to bibliographical datasets of choice. Efforts that fall under the latter category are scarce. In ref. [5], using a modified hierarchical Dirichlet process, we developed an algorithm that constructs measures of research diversity (the breadth of one's portfolio of research topics at time $t$) and research trajectory (the distance in knowledge space between one's portfolio of research topics at times $t - 1$ and $t$). We applied this to 14 years of academic publications and conference proceedings in computer science, electrical engineering and electronics to reveal that automation of certain research tasks leads to an increase in diversity of research topics and a shift in research trajectories, an outcome desirable for economic growth[5]. However, our algorithm can be used to develop measures of diversity and trajectory at various levels of analysis such as individual, organization or geographic region, and for any dataset of academic publications or patents.

While our algorithm addresses some of the limitations of off-the-shelf ML-based categorization schemas, more remain. For example, it is limited to a syntactic analysis of abstracts. Extensions and future work should consider a semantic analysis or one that takes into account the full body of text. Such steps could address the challenge that abstracts are subject to strategic behaviour that could obscure shifts in the direction of innovation — for example, as authors tend to highlight terms that are popular at that particular time. In addition, research on the direction of innovation would greatly benefit from techniques that generate hierarchies of topics and capture changes in such hierarchies over time. Last, all these techniques would need to be developed in a way that permits causal analysis, not just prediction.

We hope that bringing awareness of these potentially large benefits to a broader audience will incentivize interdisciplinary collaboration. Combining the technical knowledge of ML specialists with the domain expertise of innovation scholars could accelerate the development of empirical techniques for informing policymakers about factors that influence innovation and hence our living standards. ❐

Jeffrey L. Furman[1,2] and
Florenta Teodoridis[3*]
*[1]Boston University, Boston, MA, USA. [2]National Bureau of Economic Research, Cambridge, MA, USA. [3]University of Southern California, Los Angeles, CA, USA.*
**e-mail: teodorid@marshall.usc.edu*

Published online: 14 February 2020
https://doi.org/10.1038/s42256-020-0155-8

## References

1. Arrow, K. in *The Rate and Direction of Inventive Activity: Economic and Social Factors* 609–626 (Princeton Univ. Press, 1962).
2. Aghion, P., Dewatripont, M. & Stein, J. C. *Rand J. Econ.* **39**, 617–635 (2008).
3. Acemoglu, D. *Diversity and Technological Progress: NBER Working Paper No. 16984* (NBER, 2011).
4. Sinha, A. et al. in *Proc. 24th Int. Conf. World Wide Web* 243–246 (ACM, 2015).
5. Furman, J. L. & Teodoridis, F. *Organ. Sci.* https://doi.org/10.1287/orsc.2019.1308 (2020).

## Competing interests

The authors declare that there are no competing interests.