Evaluating Institutions for Innovation: Automated Content Analysis of Patents and Matching

Gabriel Chan

This paper applies a latent Dirichlet allocation (LDA) topic model to the corpora of U.S. patent abstracts to estimate the effect that public innovating institutions (the U.S. National Labs) have on subsequent innovation as compared to innovations from private sector institutions. Evaluating an institution's innovation effort is made difficult because the research scope of institutions often have partial but not complete overlap with each other, implying that innovation arising from one institution can be compared to only a very carefully selected subset of other innovations. Utilizing a matching algorithm on the modeled topic structure of patent abstracts, this paper identifies an appropriate subset of patents filed by the private sector that can be compared to patents filed by the National Labs. Then, subsequent citation rates between public and private sector patents are compared, holding the differences in the technological scope of the patents constant. For policymakers considering privatizing public R&D effort, this is a relevant metric for estimating the counterfactual outcome that would result if the same R&D that was conducted in a National Lab was instead conducted by the private sector. This is one of the first papers that combine natural language processing methods with matching methods in an applied context in the social sciences.