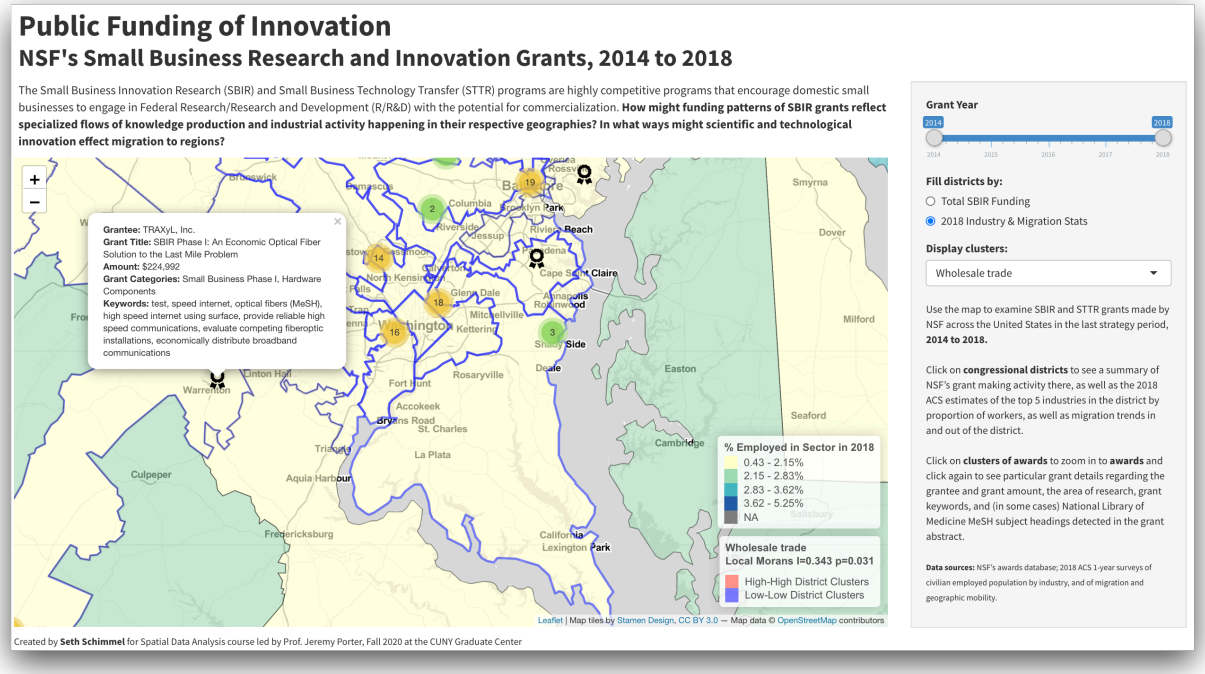
Creating a dynamic geo-social browser for exploring publicly funded research and development

Seth Schimmel
December 2020



This memo reflects my work to enrich, analyze and create and interactive visualization to display publicly available grants data from the United States **National Science Foundation (NSF)**. Inspired and informed by conversations with peers in grant-making and analyst roles at public and private funding institutions, the aim of these efforts is to facilitate discovery of information that can be used to describe research trends and develop a geo-social understanding of one funding initiative, the **Small Business Research and Innovation (SBIR)** and **Small Business Technology Transfer (STTR)** programs.

As this work remains ongoing and will be developed further for completion of my capstone project, at this time I cannot yet present any substantive analyses of these programs. Instead, before detailing the techniques and methodological choices that have gone into producing the web-app, I will discuss how recent scholarship—in fields like Bibliometrics and Scientometrics, Science and Technology Studies (STS), Science and Technology Policy Studies (STPS), and Economic Geography—has informed and motivated the work.

You can **find the Shiny web-app here.**

Background

*"Yet **specialization becomes increasingly necessary for progress**, and the effort to bridge between disciplines is correspondingly superficial. Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose." (3; my emphasis)*

*"A record, if it is to be useful to science, must be **continuously extended**, it must be stored, and above all it must be consulted." (4; my emphasis)*

Vannevar Bush, As We May Think, 1945

Head of the wartime Office of Scientific Research and Development and influential in the creation of the National Science Foundation, Vannevar Bush's observations about the lifecycle of research publications have since inspired a host of technical innovations in publishing and scholarly communication. Even in 2020, anyone who regularly follows scholarly or technical research knows the sometimes dizzying experience of tracking, collecting and making sense of the litany of resources available. Luckily, the work of curators, taxonomists, data scientists and metadata specialists on databases like EBSCO, ProQuest, JSTOR and others continues to open pathways for exploring the deluge. Moreover, as linked-data proliferates, so too have efforts to *continuously extend* scientific records grown more successful and effective. Services like the Web of Science, Altmetric, Scopus, Semantic Scholar and the Microsoft Academic Graph and others have made significant strides in reliably tracking the citation and co-authorship networks that exist between publications, researchers, research centers and funders.

Beside the practical impacts these service providers have had in facilitating discovery of research, efforts to deal with the masses of digital research publications and accompanying linked metadata have also opened up new possibilities for researchers in several disciplines. Researchers in **bibliometrics** develop statistical measures to evaluate the impact of publications, generating co-authorship and citation networks that make it possible to make content recommendations based on textual and network level information and to digitally trace genealogies of new ideas (Ganguly & Pudi, 2017; Satish et al., 2020; Vieira & Gomes, 2009). As a sub-field of bibliometrics, **scientometrics** seeks to apply these quantitative techniques to publication outputs in order to describe phenomena such as: the historical or current dynamics of national and international scientific collaboration (Gui et al., 2019; J. Li et al., 2020; Newman, 2001; Stek & van Geenhuizen, 2016); the emergence of innovations in particular fields like materials science, topics like big data, or even particular technologies like flash memory or solid-state drives (Gläser & Laudel, 2015, 2015; Y. Li et al., 2018; Ranaei et al., 2019); to explain how publication and citation activity differs based on gender, and relates to career outcomes for scientists (King et al., 2017; Nosek et al., 2010; Parker et al., 2013). Scientometrics has contributed techniques that are useful both for enhancing research monitoring and distribution platforms, and also for conducting sociological analyses of various dimensions of scientific

practice. Increasingly, though, there is also overlap between practitioners of scientometrics and others in fields like policy studies, economic geography and those conducing evaluations of funding programs.

Robust datasets, the expansion of linked data, and frequent collaboration with researchers in computational social sciences, natural language processing, network science, economics and geography enables researchers to develop actionable insights not only for research program managers and evaluators hoping to increase their team's impact, but also to policy makers more broadly. For instance, at the 7th Satellite on Quantifying Success at this year's Network Science Conference, researchers from Northwestern University's Center for Science of Science and Innovation presented work drawing linkages between funding of scientific research and publications' likelihood of being cited in government policy documents, by media outlets and in patents. Furnishing their analysis was 200 million articles from the Microsoft Academic Graph, and Altmetric's store of an additional 350K articles cited by over 40K government policy documents, 550K news media pieces citing scientific research, 1.3M United States patent records and an additional 5 million research projects sponsored by 400+ funding agencies worldwide. Their study of COVID-19 citation networks showed the carryover effects that policy document citation had on future citation, suggesting the influence policymakers can have in affecting the dynamics of scientific communication and citation (Yin et al., 2020).

While the linked data ecosystem has become more robust in recent years, the use of patent records as a proxy for measuring innovative activity is not new (Jaffe et al., 1993; Verbeek et al., 2004). Recent scholarship in economic geography and science and technology policy studies uses patent data to describe comparative advantages and disadvantages between different geographic regions and to offer policy makers and analysts one way of understanding of how scientific and technical *specialization* might effect policy planning and outcomes (Apa et al., 2018; Balland et al., 2019; Boschma et al., 2014; Castaldi & Los, 2017; Perruchas et al., 2020; Surana et al., 2020; Wouden & Rigby, 2020).

While quantitative measures of scientific research outputs alone may not be sufficient to when evaluating funding decisions, several recent works have derived actionable findings from their analysis of grant proposals and funding programs. For instance, Myers developed an econometric model for the costs of switching research topics given the probability of increased payout by winning an award from one of the National Institutes of Health's targeted funding opportunities. To build the model, Myers used text analytics to develop similarity measures between researcher's prior published work and the text of the call for proposals (Myers, 2020). Hoppe et. al (2019) recently used word2vec and other topic clustering techniques to evaluate how differences in topics proposed by African American / Black researchers versus White researchers, and to identify how topical differences emerged at particular stages of the application process (Hoppe et al., 2019). In each of these cases, the findings can be applied in a monitoring and evaluation context to design more attractive and equitable funding opportunities and programs.

My work intends to draw on and engage with scholars, analysts and funders for whom the work above is of interest. While much has been done analyzing patent data, publications and citations as proxies for particular scientific outcomes or to develop regional comparisons, much less work has been done to analyze publicly available grants data (Kardes et al., 2014; Talley et al., 2011; Zhang et al., 2016) and to assess relationships between geographical specialization and funding flows (Chausse Vázquez de Parga, 2018). This project adds to those handful of efforts that explicitly analyze funding data, using techniques from text mining and linked data to enrich data. This work also takes inspiration from scholars in economic geography and spatial analysis by visualizing the data geographically alongside potentially related contextual indicators.

In my view, a data scientist can contribute to these issues by putting into code the diversity of perspectives and questions being posed to potentially related datasets. By doing so, the data scientist curates like a digital librarian, prepares data for their own analysis but also for analysis by other future analysts, and designs visualizations for other practitioners interested in these topics. I hope to do a little bit of each by extending this work on a few different fronts, detailed in the methodology section below.

Methods

Getting the data

NSF's historic to present **awards data** is openly available at the grant level as XML files, and includes details such as the NSF directorate, division and program officer responsible for making the award, award title, abstract, principal investigator, recipient organization, award instrument (standard grant, continuing grant, fellowship, fixed price award, etc.), and several administrative codes that add thematic or program-level details. I modified existing python scripts found on Github authored by University of Pennsylvania doctoral student and science of science researcher Titipat Achakulvisut to parse the XML files into a single dataframe and ensure all relevant fields were present.

Other data sources include:

- Census.gov shapefiles for county and 116th Congress congressional district
- ProPublica Congress API to obtain senator and representative information for each district
- 2018 American Community Survey 1-year estimate data at the congressional district level for:
    - Geographical mobility in the past year
    - Industry for the civilian employed population aged 16 and older

Cleaning and processing the awards text data

Directorate and division names were cleaned for parity, taking liberties except where potential program name changes signaled historical transitions. While "program reference codes" and "program element codes" offer some additional details about thematic or

strategic aspects of the grants, they are messy and administratively opaque. Codes were cleaned for parity using the alphanumeric identifiers affixed to them. If the taxonomic logic behind these codes was more transparent, thematic analysis could likely happen with less emphasis on text mining methods.

To identify grants belonging to the SBIR/STTR program, I queried text from the program codes, grant title and abstract fields. This program was chosen for its relevance to broader issues in science, technology and innovation policy studies. It is also an interesting subset of the entire collection, as it is one of the few NSF programs where the grantees are not typically colleges or universities. The period of 2014 to 2018 was selected because it reflects the last full strategy period, and it may be worthwhile in the future to ground the analysis by comparing localization and topical trends by strategy period.

To clean and pre-process the text from grant title and abstracts, I relied on the NLPre pipeline made publicly available on Github by the NIH Office of Portfolio Analysis team. The pipeline does standard text cleaning, stopword removal and tokenization, but also expands and logs abbreviations and marks terms in the text that also appear in the National Library of Medicine's Medical Subject Headings (MeSH). While unsupervised text mining techniques are useful for their ability to reveal frequently used terms and phrases, relying on dictionary-based methods of extraction can be advantageous to the extent it enables inter-operability and cross-platform comparisons. In the future I would like to utilize other taxonomies, ontologies and RDFs to enhance term extraction and create linkages to additional descriptive information and related resources. Choice of sources may depend on the subset of grants to be studied, but could possibly include:

- GESIS Thesaurus for the Social Sciences
- American Economic Association's Journal of Economic Literature (JEL) subject codes
- The European Comission's EuroVoc and EuroSci vocabularies
- The UN's AgroVoc agricultural and food sciences vocabulary
- The UN's linked-data Sustainable Development Goals taxonomy
- The European Institute for Gender Equality's Gender Equality Thesaurus

Using the processed grant text, I used three different techniques for finding keywords and phrases: TF-IDF, the RAKE algorithm, and textrank. In this context, TF-IDF is useful because it takes into account corpus-level document frequency measures in order to uncover the frequent keywords that appear in a particular award and infrequently across the corpus. The RAKE algorithm is suitable for finding variable length phrases, and uses co-occurrence frequency and degree measures for finding optimal phrases. Textrank operates at the document level only, and, like RAKE, evaluates optimal keywords and phrases by ranking them within a network representation of the document. The outputs of all three methods were compared and merged so as to keep a limited number of unique phrases and/or the longest phrase derived from the most frequent unigrams.

Georeferencing and spatial cluster analysis

The awards were georeferenced using the Geocod.io web service, using the recipient address in order to obtain latitude and longitude estimates. Using the latitude and longitude estimates, the data was merged with various shapefiles using qGIS so that the awards data could also include county level and congressional district level identifiers and be joined up with the ProPublica congressional district details and ACS survey data.

Congressional districts were chosen as the unit of analysis based on conversation with grant makers who fund social sciences and humanities research, who expressed the potential utility of presenting the information to policy analysts and policy makers interested in understanding the publicly funded research activity in their district. This choice also has a benefit when conducting spatial cluster analysis that congressional districts are administrative boundaries whose populations are all comparable – each reflects an area of roughly 700,000 persons. This may help to mitigate the visual distortion in choropleths where area and population can be confounded.

For the ACS data, each industry and type of migration to a district (from the same county, from another county in the state, from out of state, from abroad) was transformed into a percentage of the total population before being mapped. Using a series of spatial data analysis packages in R, I generated a Queen's continuity matrix and performed a series of Univariate Local Moran's I tests to evaluate whether there were significant high or low valued clusters where particular migration patterns were similar in 2018 or where a similar distribution of worker specialization is observed.

The choice to include industry and migration statistics specifically was influenced by work on R&D knowledge spillovers that examine how proximity of innovative firms effects the overall spatial distribution of innovative activity and its associated economic effects (Anselin et al., 2000; Bonaccorsi & Daraio, 2005; Jaffe et al., 1993; Wallsten, 2001). While Jaffe and Wallsten emphasized, in part, how the tacit, social and non-codified dimensions of knowledge sharing may thought of as the reason why like-minded researchers would tend to cluster, others have avoided such theories in favor of more political-economic theories dealing with labor market localization or even research subcontracting arrangements between universities and other firms. Breschi and Francesco (2001), for instance, warned that spillovers of *knowledge* may not be unique simply because "knowledge" is supposedly a unique kind of product or service that depends on social transmission and collaboration, stating: "we suggest that spatial proximity of innovators, when found to be significant, may not depend upon any intrinsic feature of knowledge, such as its "tacitness" or "codification", but on a much more complex interplay between those characteristics, the labour market for scientists and technologists" (3). Following those whose work remarks on the localization of specific industries and meanwhile also aim to disaggregate findings of innovation spillovers or localizations across *specific* types of activities. (Anselin et al., 2000; Audretsch & Feldman, 1996; Boschma et al., 2014). Identifying spatial clusters where employment in specific industries tends to be highest is a

good first step for guiding inquiry into how labor market features affect the kinds of innovative activities that come out of particular regions.

Analytics and visualization choices + directions for future work

The visualization is built using Leaflet and Shiny in RStudio. By default, the map is shaded to reflect the SBIR program funding for the full 2014 to 2018 period across districts. The user can change the date range and the color scale and legend will dynamically change. Awards are clustered and become district as the user zooms in, at which point they can click to see award details. Clicking the district offers a summary of the area's total SBIR funding for the selected time period, as well as the other NSF funds received in that same period. The popup for the district also includes a summary of how the district compares to others across industries and for migration patterns. These summaries reflect the top five industries by ranked percentage. The relatively constant population between districts makes this measure a fair one for judging what industries are most prevalent in a given district.

Spatial clusters are visible on the map, and the user can select which clusters they'd like to be highlighted from the dropdown list. The user can also switch from seeing the districts shaded by SBIR funding to a view where districts are shaded by the contextual variable's values. Additional analysis is needed to fully explore whether topics different between clusters. However, I believe these non-intuitive clusters provide interesting units of spatial analyses where areas can be demarcated not be compass regions but by areas of common industrial activity. Questions to **analyze, explore** and **foreground** may include:

- In what industrial sectors do SBIR-recipient dense regions tend to specialize?
  - **Technique: Rank Statistics** -- Are SBIR-recipient dense regions **ranking** more highly in some industries versus other?
- Are the SBIR dense regions more or less similar to each other with regard to industrial profiles?
  - **Technique: Clustering** districts based on similarity scores computed on % population across the 16 sectors
  - **Visualization: tSNE scatter plot** (w districts as dots, filterable on region/state/district, shape as region, color as cluster); **(radial charts/parallel coordinates plot)** (w/ years as opaque layers, animate or user-select year layers, plot area as raw pct working pop, or as heat-mapped above/below average; radial helps by making industry a position on the circle, whereas it may be too many colors on a steam or bar graph)
- Are regions with "*more or less similar*" industrial profiles being funded to produce "*more similar"* research?

    - **Technique: Use above clustering results,** then compare directorate/division level funding breakdown and the intra-directorate **linguistic similarity scores** of grants using word2vec/doc2vec
    - **Technique:** Identify which linked data vocabulary terms are present, examine by industrial occupation cluster, see whether there are any patterns
- How do/can these small businesses take advantage of the environmental, social and policy landscape of their region?
    - **Technique:** Qualitative (policy analysis)
    - **Technique:** Use OpenStates API to find state-level legislation with subject areas related to either STEM+innovation policy OR the subject area(s) related to the keywords or program codes
- Are regions receiving more SBIR funding experiencing migration in or out? Is this the same or different for NSF funding across all programs?
    - **Technique:** Hypothesis testing, regression
- Do SBIR grants patterns coincide or differ from the funding patterns of other programs? Does SBIR follow the overall funding distribution? Do SBIR grantees spread out or cluster near areas where funding is already high, take advantage of knowledge spillovers?
    - **Technique:** Descriptive stats, breakdown of directorate/division funding data, aggregation of funding, number of unique grants, number of unique grantees spatially aggregated (perhaps at smaller unit level like county)
- Do regions experiencing more migration to/from each also produce "*more similar*" NSF funded science? In what districts was it more or less common for there to be funded PIs migrating between the two areas?
    - **Technique:** Use ACS migration data, compare intra-directorate similarity scores using doc2vec, regress similarity on migration
    - **Technique:** Identify PIs in grants data who are listed as affiliated with more than one institution over time period t0-t1, are regions with PIs in common producing more similar intra-directorate? Are regions with PI migration likely to have more similar industrial profiles?

Future work may include breaking down the total NSF funding by division or directorate, and aggregating the keyword data at the district level. The former task would be relatively straightforward and involve creating a panel with an additional chart, while the latter requires further processing of keywords data to allow for dynamic filtering at the date and district level. Producing district-level keyword trends data would be advantageous toward building additional functionality whereby a funder or policy maker could use the web-app to find their district, select the date range and quickly obtain a printout with funding details.

Datasets, Scripts, APIs

Achakulvisut, T., Ruangrong, T., Acuna, D. (2018). Grant database: downloader, preprocessor, parser and deduper for NIH and NSF grants. GitHub Repository, https://github.com/titipata/grant_database

Hoppe, T.A., Baker, H. (2019). Natural Language Preprocessing (NLPre), GitHub Repository, https://github.com/NIHOPA/NLPre

National Science Foundation. (2020). Awards Database. Retrieved from https://www.nsf.gov/awardsearch/download.jsp

ProPublica. (2020). ProPublica Congress API. https://projects.propublica.org/api-docs/congress-api/

U.S. Census Bureau. (2018). *Geographical mobility in the past year by age for current residence in the United States*. Retrieved from https://data.census.gov/cedsci/table?q=ACSDT1Y2019.B07001&tid=ACSDT1Y2019.B07001&hidePreview=true

U.S. Census Bureau. (2018). *Industry for the civilian employed population 16 years and over*. Retrieved from https://data.census.gov/cedsci/table?q=Industry%20for%20the%20civilian%20employed%20population%20aged%2016%20and%20older&g=0100000US.50016&tid=ACSST1Y2018.S2403&hidePreview=true

Bibliography

Anselin, L., Varga, A., & Acs, Z. (2000). Geographical Spillovers and University Research: A Spatial EconometricPerspective. *Growth and Change*, *31*(4), 501–515. https://doi.org/10.1111/0017-4815.00142

Apa, R., De Noni, I., Orsi, L., & Sedita, S. R. (2018). Knowledge space oddity: How to increase the intensity and relevance of the technological progress of European regions. *Research Policy*, *47*(9), 1700–1712. https://doi.org/10.1016/j.respol.2018.06.002

Audretsch, D. B., & Feldman, M. P. (1996). R&D Spillovers and the Geography of Innovation and Production. *The American Economic Review*, *86*(3), 630–640. JSTOR.

Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies*, *53*(9), 1252–1268. https://doi.org/10.1080/00343404.2018.1437900

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, *63*(1), 87–120. https://doi.org/10.1007/s11192-005-0205-3

Boschma, R., Balland, P.-A., & Kogler, D. F. (2014). Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, *24*(1), 223–250. https://doi.org/10.1093/icc/dtu012

Castaldi, C., & Los, B. (2017). Geographical patterns in US inventive activity 1977–1998: The "regional inversion" was underestimated. *Research Policy*, *46*(7), 1187–1197. https://doi.org/10.1016/j.respol.2017.04.005

Chausse Vázquez de Parga, I. (2018). *A geographical analysis of research trends applying text mining to conference data* [Escola Tècnica Superior d'Enginyeria Industrial de Barcelona]. https://upcommons.upc.edu/handle/2117/169067

Ganguly, S., & Pudi, V. (2017). Paper2vec: Combining Graph and Text Information for Scientific Paper Representation. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, & J. Tait (Eds.), *Advances in Information Retrieval* (pp. 383–395). Springer International Publishing. https://doi.org/10.1007/978-3-319-56608-5_30

Gläser, J., & Laudel, G. (2015). A bibliometric reconstruction of research trails for qualitative investigations of scientific innovations. *Historical Social Research*, *40*(3), 299–330. https://doi.org/10.12759/hsr.40.2015.3.299-330

Grandjean, M., Benz, P., & Rossier, T. (2017). *Elites académiques et (re)définition des frontières disciplinaires. Collaborations interdisciplinaires et structure du pouvoir académique*. 7e Congrès de l'Association Française de Sociologie. https://halshs.archives-ouvertes.fr/halshs-01525575

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum*, *105*, 1–12. https://doi.org/10.1016/j.geoforum.2019.06.017

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., Davis, A. F., Lauer, M. S., Valantine, H. A., Anderson, J. M., & Santangelo, G. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances*, *5*(10), eaaw7238. https://doi.org/10.1126/sciadv.aaw7238

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, *108*(3), 577–598. JSTOR. https://doi.org/10.2307/2118401

Kardes, H., Sevincer, A., Gunes, M. H., & Yuksel, M. (2014). Complex Network Analysis of Research Funding: A Case Study of NSF Grants. In F. Can, T. Özyer, & F. Polat (Eds.), *State of the Art Applications of Social Network Analysis* (pp. 163–187). Springer International Publishing. https://doi.org/10.1007/978-3-319-05912-9_8

King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time. *Socius*, *3*, 2378023117738903. https://doi.org/10.1177/2378023117738903

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2020). Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *Journal of The Royal Society Interface*, *17*(165), 20200135. https://doi.org/10.1098/rsif.2020.0135

Li, Y., Li, H., Liu, N., & Liu, X. (2018). Important institutions of interinstitutional scientific collaboration networks in materials science. *Scientometrics*, *117*(1), 85–103. https://doi.org/10.1007/s11192-018-2837-0

Myers, K. (2020). The Elasticity of Science. *American Economic Journal: Applied Economics*, *12*(4), 103–134. https://doi.org/10.1257/app.20180518

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Nosek, B. A., Graham, J., Lindner, N. M., Kesebir, S., Hawkins, C. B., Hahn, C., Schmidt, K., Motyl, M., Joy-Gaba, J., Frazier, R., & Tenney, E. R. (2010). Cumulative and Career-Stage Citation Impact of Social-Personality Psychology Programs and Their Members. *Personality and Social Psychology Bulletin*, *36*(10), 1283–1300. https://doi.org/10.1177/0146167210378111

Parker, J. N., Allesina, S., & Lortie, C. J. (2013). Characterizing a scientific elite (B): Publication and citation patterns of the most highly cited scientists in environmental science and ecology. *Scientometrics*, *94*(2), 469–480. https://doi.org/10.1007/s11192-012-0859-6

Perruchas, F., Consoli, D., & Barbieri, N. (2020). Specialisation, diversification and the ladder of green technology development. *Research Policy*, *49*(3), 103922. https://doi.org/10.1016/j.respol.2020.103922

Ranaei, S., Suominen, A., Porter, A., & Carley, S. (2019). Evaluating technological emergence using text analytics: Two case technologies and three approaches. *Scientometrics*. https://doi.org/10.1007/s11192-019-03275-w

Satish, S., Yao, Z., Drozdov, A., & Veytsman, B. (2020). The impact of preprint servers in the formation of novel ideas. *BioRxiv*, 2020.10.08.330696. https://doi.org/10.1101/2020.10.08.330696

Stek, P. E., & van Geenhuizen, M. S. (2016). The influence of international research interaction on national innovation performance: A bibliometric approach. *Technological Forecasting and Social Change*, *110*, 61–70. https://doi.org/10.1016/j.techfore.2015.09.017

Surana, K., Doblinger, C., Anadon, L. D., & Hultman, N. (2020). Effects of technology complexity on the emergence and evolution of wind industry manufacturing locations along global value chains. *Nature Energy*, 1–11. https://doi.org/10.1038/s41560-020-00685-6

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., Leenders, A. G. M., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443–444. https://doi.org/10.1038/nmeth.1619

Verbeek, A., Debackere, K., & Luwel, M. (2004). Science cited in patents: A geographic "flow" analysis of bibliographic citation patterns in patents. *Scientometrics*, *58*(2), 241–263. https://doi.org/10.1023/a:1026232526034

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, *81*(2), 587–600. https://doi.org/10.1007/s11192-009-2178-0

Wallsten, S. J. (2001). An empirical test of geographic knowledge spillovers using geographic information systems and firm-level data. *Regional Science and Urban Economics*, *31*(5), 571–599. https://doi.org/10.1016/S0166-0462(00)00074-0

Wouden, F. van der, & Rigby, D. L. (2020). Inventor mobility and productivity: A long-run perspective. *Industry and Innovation*, *0*(0), 1–27. https://doi.org/10.1080/13662716.2020.1789451

Yin, Y., Dong, Y., Wang, K., Wang, D., & Jones, B. F. (2020, September 17). Quantifying uses of science beyond science. *The 7th Satellite on Quantifying Success*. NetSci 2020.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179–191. https://doi.org/10.1016/j.techfore.2016.01.015