Seth Schimmel
January 2021

Project 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

For my capstone project, I will conduct analysis and create exploratory interactive visualizations to display publicly available grants data reflecting the activities of various federal agencies who participate in the Small Business and Innovation Research (SBIR) and Small Business Technology Transfer (STTR) grant making programs. 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 these funding initiatives. The intended audience for this work is twofold: firstly, grant makers and policy analysts looking for quantitative and qualitative insight into the body of over 66,000 grants made between 2008 and 2018; second, policymakers looking to get a quick brief on innovation funding in their congressional districts. The goal is to build a visualization tool that can also offer fact-sheet print-outs for each district, which can be useful for policy analysts and policymakers who want to know what research has recently been funded in their district.

This work is practically inspired by my professional work as a data science analyst at a private foundation, and theoretically inspired by several fields of research and library and information science. Firstly, I am taking after researchers in **bibliometrics** who develop statistical measures to evaluate the impact of publications, sometimes 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). More specifically, 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); 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); 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). 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. Extending these areas of work, my hope is to apply to grants data a "tech-mining" approach that seeks to identify trending innovations, technologies, and research topics using methods in text mining, natural language processing and linked-data management. To do so, I will make use of various openly available specialized vocabularies, taxonomies and ontologies related to the sciences, environment, social and economic policies from sources such as the United Nations, the European Union, and various policy centers and institutes.[1]

Increasingly, there is significant overlap between practitioners of scientometrics, fields like policy studies and economic geography and those conducing evaluations of funding programs. While I may develop descriptive statistics that help compare and contrast aspects of different funding initiatives over time and space, my theoretical approach here will not be "monitoring and evaluation" or "learning and assessment." While these approaches are useful, important and popular within public and private foundations, my aim is analytics development rather than arriving at evaluative conclusions. To facilitate a richer analytical description of the grants, though, I will use a geospatial theoretical frame as I analyze and curate the dataset. By using American Community Survey data on labor sectors by congressional district, I will develop a geospatial cluster analysis of waxing and waning labor sectors that will be used to foreground questions about the geography of industrial and economic change. As part of my accompanying analysis I will examine the qualitative and/or quantitative differences in the funding activity that happened within different geospatial clusters, sectoral "hot spots" or "cold spots" where certain economic activities and corresponding labor distributions tend to be highest or lowest, or increasing or decreasing most significantly.

This approach is influenced by scholarship in **economic geography** and **science and technology policy studies,** which 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). Using patent records as a proxy for measuring innovative activity is not new, and various researchers have sought to derive insight about innovative activity from patents data (Jaffe et al., 1993; Verbeek et al., 2004). Moreover, researchers have also used patents data to study the localization of specific industries and utilize the Standard Industrial Classification codes applied to patent records to further disaggregate findings of innovation spillovers or localizations across *specific* types of activities. (Anselin et al., 2000; Audretsch & Feldman, 1996; Boschma et al., 2014).

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

[1] A sample of the sources are included in the Data Sources section at the end of this document.

efforts that *explicitly analyze funding data*. Since the taxonomic descriptors available for the grants on SBIR.gov is less robust than patents data, I will use existing linked-data vocabularies and ontologies (included in the data sources bibliography) to do targeted text mining and will evaluate the effectiveness of this methodology in the whitepaper. This work also takes inspiration from scholars in economic geography and spatial analysis by visualizing the data geographically to enable exploration by proxy of some relevant economic indicators.

Resources and tools

I have completed parts of this work during my coursework. I began text mining grants data using the National Science Foundation's open grants data in Fall of 2019 in a data mining course. In the following semester, I developed interactive visualizations in D3 that showcased differences in keyword trends across that dataset. Most recently, I used Shiny and R to complete a geospatial analysis and exploratory visualization that can identify labor sector clusters by congressional district and additionally puts grants on a map of the United States. I have developed this project meanwhile through conversation with peers at public and private foundations in data scientist, data analyst, and program officer roles, as well as through observations made while attending conferences such as the Network Science 2020 Conference and Empirical Methods in Natural Language Processing 2020 Conference. Last, I will modify an existing text-mining pipeline available on Github and developed by the NIH Office of Portfolio Analysis for doing text analytics on grant and publication data. To host the final project, my tentative plan is to use Github pages.

Data Ethics and Management Plan

Data gathered from the American Community Survey contains no personal identifiers, and will be used with proper attributions. Linked-data resources, vocabularies and taxonomic resources collected do not contain any personal identifiers, and will be used with proper attributions. Data collected from SBIR.gov does contain personal identifiers corresponding to the Principal Investigators associated with a given grant record. As this information is not necessary to the work, to minimize exposure of personal information I will remove this data from my dataset as early as possible before continuing the processing, analysis and visualization steps.

Regarding data management, I plan to keep a Github repository for the project where I will document ideations, design sketches, all processing scripts, data and visualization scripts. Additionally, as I work I will develop documentation through markdown files that reflect the evaluations driving my processing and data manipulation choices.

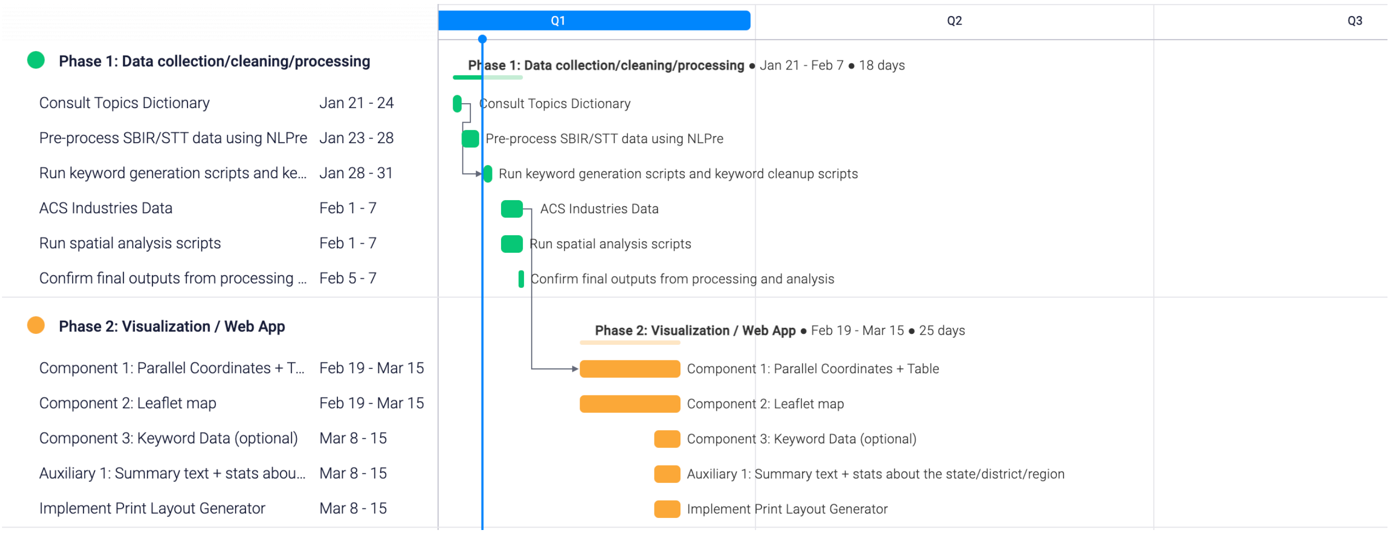
Archiving and Digital Deposit Considerations

Based on a conversation with Stephen Klein... HAPPENING MONDAY FEB 1

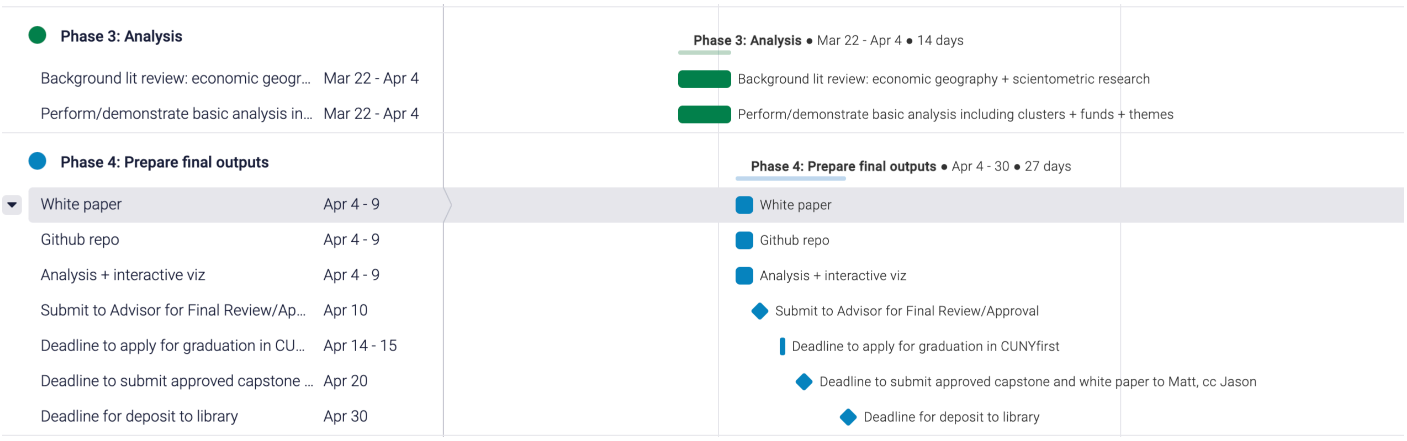
Seth Schimmel
January 2021

Workplan

I am using Monday.com to chart my progress for the anticipated completion date of mid-April. Shown here is the Gantt chart for the project – about 2/3 of phase 1 has already been completed, and the various visualization components are prepared in initial forms.



Phase 1: Data collection/cleaning/processing	
Consult Topics Dictionary	Jan 21 - 24
Pre-process SBIR/STT data using NLPre	Jan 23 - 28
Run keyword generation scripts and ke...	Jan 28 - 31
ACS Industries Data	Feb 1 - 7
Run spatial analysis scripts	Feb 1 - 7
Confirm final outputs from processing ...	Feb 5 - 7
Phase 2: Visualization / Web App	
Component 1: Parallel Coordinates + T...	Feb 19 - Mar 15
Component 2: Leaflet map	Feb 19 - Mar 15
Component 3: Keyword Data (optional)	Mar 8 - 15
Auxiliary 1: Summary text + stats abou...	Mar 8 - 15
Implement Print Layout Generator	Mar 8 - 15



Phase 3: Analysis	
Background lit review: economic geogr...	Mar 22 - Apr 4
Perform/demonstrate basic analysis in...	Mar 22 - Apr 4
Phase 4: Prepare final outputs	
White paper	Apr 4 - 9
Github repo	Apr 4 - 9
Analysis + interactive viz	Apr 4 - 9
Submit to Advisor for Final Review/Ap...	Apr 10
Deadline to apply for graduation in CU...	Apr 14 - 15
Deadline to submit approved capstone ...	Apr 20
Deadline for deposit to library	Apr 30

Phase 3: Analysis • Mar 22 - Apr 4 • 14 days
Background lit review: economic geography + scientometric research
Perform/demonstrate basic analysis including clusters + funds + themes
Phase 4: Prepare final outputs • Apr 4 - 30 • 27 days
White paper
Github repo
Analysis + interactive viz
Submit to Advisor for Final Review/Approval
Deadline to apply for graduation in CUNYfirst
Deadline to submit approved capstone and white paper to Matt, cc Jason
Deadline for deposit to library

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

European Environment Agency (EEA). (2021). GEMET - *General Multilingual Environmental Thesaurus.* Downloaded January 28, 2021. Retrieved from: https://www.eionet.europa.eu/gemet/en/about/

European Institute for Gender Equality (EIGE). (2021). *Gender Equality Glossary and Thesaurus*. Downloaded January 28, 202. Retrieved from: https://eige.europa.eu/thesaurus/about

Food and Agriculture Organization of the United Nations (FAO). (2021). *AGROVOC*. Downloaded January 28, 2021. Retrieved from http://www.fao.org/agrovoc/access

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

Small Business Innovation Research / Small Business Technology Transfer Programs (2021). Awards Database. Downloaded January 8, 2021. Retrieved from: https://www.sbir.gov/sbirsearch/award/all

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

Publications Office of the European Union. (2021). *European Science Vocabulary (EuroSciVoc).* Downloaded January 28, 2021. Retrieved from: https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/euroscivoc

Publications Office of the European Union. (2021). *EUROVOC.* Downloaded January 28, 2021. Retrieved from: https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

Renewable Energy and Energy Efficiency Partnership (REEP). (2021). *Clean Energy Linked Open Data*. Downloaded January 28, 2021. Retrieved from: http://poolparty.reegle.info/PoolParty/sparql/glossary

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