THE PUBLIC INNOVATIONS EXPLORER: A GEO-SPATIAL & LINKED-DATA VISUALIZATION PLATFORM FOR PUBLICLY FUNDED INNOVATION RESEARCH IN THE UNITED STATES

by

SETH SCHIMMEL

A master's capstone project submitted to the Graduate Faculty in Data Analysis and Visualization in partial fulfillment of the requirements for the degree of Master of Science, The City University of New York

2021

The Public Innovations Explorer: A Geo-Spatial & Linked-Data Visualization Platform For Publicly Funded Innovation Research In The United States

by

Seth Schimmel

This manuscript has been read and accepted for the Graduate Faculty in Data Analysis and Visualization in satisfaction of the capstone project requirement for the degree of Master of Science.

Date	Matthew Gold Capstone Project Advisor
Date	Matthew Gold Program Director

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

The Public Innovations Explorer: A Geo-Spatial & Linked-Data Visualization Platform For Publicly Funded Innovation Research In The United States

by

Seth Schimmel

Advisor: Matthew Gold, PhD.

The Public Innovations Explorer is a web-based tool created using Node.js, D3.js and Leaflet.js that can be used for investigating awards made by Federal agencies and departments participating in the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) grant-making programs between 2008 and 2018. By geocoding the publicly available grants data from SBIR.gov, the Public Innovations Explorer allows users to identify companies performing publicly-funded innovative research in each congressional district and obtain dynamic district-level summaries of funding activity by agency and year. Applying spatial clustering techniques on districts' employment levels across major economic sectors provides users with a way of examining patterns in the underlying economic activities of districts alongside Federally-funded innovation research activities taking place in a district. Finally, mathematical and dictionary-based text-mining techniques are used to derive district-level keyword details and provide users with access to some basic keyword stats for each district. Among other sources, the Explorer utilizes vocabulary sources from the European Commission, the United Nations and Leibniz Information Centre for Economics and builds on the National Institute of Health Office of Portfolio Analysis's NLPre Pipeline available on Github to index keywords extracted from the text of grant records. The project seeks to contribute to work in research fields like scientometrics, economic geography, and in the nonprofit and philanthropy sector by developing and documenting data processing techniques and a user-interface fit for exploring geographic and thematic trends across grant datasets.

Keywords: Small Business Innovation Research; grants data; innovation research; dictionary-based keyword extraction; linked data; spatial clustering; geospatial analysis; scientometrics; economic geography

ACKNOWLEDGMENTS

This project would not be possible without the mentorship and guidance of my professors, peers and colleagues. I am grateful for the enthusiasm Dr. Hussein Ghaly brought to our computational linguistics labs, where I learned NLP and Python. I began applying this learning to publicly available grant records while learning more about data mining from Professor Ashwin Satyanarayana. Professor Jeremy Porter's spatial analysis class provided an opportunity to engage with work at the intersection of economic geography and scientometrics, and I am grateful for the conversations I had with Dr. Porter and his insights on the research literature on industrial change and more generally his advice on choosing metrics and balancing presentation and analysis. I also want to acknowledge the immense learning opportunity I had while a part of Professor Ellie Frymire's data visualization studio. Her enthusiastic support, technical pointers and perspective on what a calendar might reasonably look like when learning Javascript on the fly gave me just the right balance of realism and ambition as I made final decisions as to the components I wanted to utilize for this project.

I also want to acknowledge the rich discussions I had with Dr. Howard Everson when I was just beginning to think about the project. His familiarity with the competitive landscape of private and public research funding and his experienced perspective as an analyst and evaluator helped me to hone in on defining a three-month project to complete, and his networking advice proved invaluable. I am thankful for the peers at research and funding institutions – Koby Oppenheim, Jason Rhody, Travis Hoppe – who so graciously offered up their time to jump on Zoom calls and help me think through this project in its infancy and who ultimately offered key insights that helped me imagine what an analyst, grant-maker or policy-maker looking at the data might like to see.

I am also grateful to a set of mentors that helped to make my years-long engagement at The Graduate Center possible while I was working full-time. I'm thankful for Marianna Azar, my manager while working at the GC, who encouraged me to get started ASAP on beginning my graduate studies and pointed me to the MALS program. I'm thankful also for Myka Carroll, who has continued to support my professional and academic development, pointed me to all kinds of workshops, conferences and trainings, and without whose thought-partnership I likely would've never taken on such a project. I'm grateful for the flexibility I've been afforded by my managers and colleagues while completing my studies. In the same way, I am grateful to my capstone advisor Dr. Matthew Gold, who shepherded me through the MALS program into the new MS program and whose honesty, calm and trust during my capstone work helped me to feel grounded and secure in the work I proposed and completed. I am grateful for the conversations we had throughout my years at The Graduate Center, as I thought about my path forward and reflected on what I was learning, and, much like our Zoom calls regarding this project, from which I always left feeling more balanced, secure in my pacing, and excited about taking the next steps.

Last, I am grateful to friends and family who supported me throughout the journey of obtaining this degree. I am grateful for the thought partners and friendships where conversation sparks inspiration, and grateful to have witnessed the unique trajectories these friends have taken over these last few years. And, last but most certainly not least, I am grateful for Müge, who listened as I rambled through ideas, gave me key bits of design advice, and whose love and friendship helped me to navigate these last few years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DIGITAL MANIFEST

I. Capstone Whitepaper (PDF)

II. WARC Files of Public Innovations Explorer Website

Archived version of The Public Innovations Explorer

(https://sethsch.github.io/innovations-explorer/app/index.html)

III. Code and other processed data files

Zip file containing the contents of the Github repository at time of deposit

(https://github.com/sethsch/innovations-explorer)

A NOTE ON TECHNICAL SPECIFICATIONS

The Public Innovations Explorer is a website currently hosted on Github Pages here: https://sethsch.github.io/innovations-explorer/app/index.html

The website utilizes Node.JS, D3.js, Leaflet.js, SlickGrid and the Bootstrap 5.0 UI Kit. The site depends on data resources accessible in the application data directory of the application folder at the Github repository here: https://github.com/sethsch/innovations-explorer

CHAPTER 1: BACKGROUND

Scholars in fields like economic geography and scientometrics regularly make use of publications and patents data to quantify aspects of scientific practice such as collaboration networks and the flow of new ideas and technologies. In each field, the ability to answer entire research questions can crucially depend on the relative cleanliness of either the qualitative subject headings that publishers apply to scholarly publications, or of the industrial classification codes that the Federal government applies to patent records. With consistent metadata schemes in place, researchers can then study the recombination of classification tags applied to records as a proxy for the recombinant social activity of knowledge exchange.

While many studies make use of these important *outputs* of knowledge production, far fewer investigate the *inputs* to knowledge production in the form of funding flows. Grant records are one such input that capture funding flows in a discrete way. As a type of data artifact, moreover, these records are comparable to publications or patents—including information like titles, abstracts, attribution and authorship details—but have far less consistent classification schemes. Taxonomic efforts like Candid's Philanthropy Classification System apply a subject area, populations and grant-making strategy taxonomy to awards data Candid solicits and collects through a hybrid editorial-algorithmic approach. The Philanthropy Classification System (https://taxonomy.candid.org/) is available to be applied at a cost as a natural language processing driven classifier through an API, and Candid's grant-level data lives behind a paywall and is largely used by those in the philanthropy and nonprofits sectors rather than researchers in scientometrics. This project began as an effort to apply techniques in natural language processing and data mining to explore qualitative patterns across various publicly available Federal grants datasets, initially with the National Science Foundation's grants data and ultimately the Small Business Administration's awards database for the SBIR/STTR grant making programs from all 11 participating Federal agencies and departments. In the end, the effort to curate richer qualitative data for the grant records became just one piece of a broader effort to facilitate data exploration and information discovery within a geospatial discovery platform.

The aim of the Public Innovations Explorer (also referred to throughout as “the Explorer”; available at: https://sethsch.github.io/innovations-explorer/app/index.html) is to provide a platform for various kinds of analysts as well as the interested public to discover information and perform exploratory data analysis concerning innovation research undertaken by recipients of Federal grants through the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs between 2008 and 2018. The Explorer presents this information geographically at the level of the Congressional Districts for the 116th United States Congress (2018), and simultaneously allows users to identify “hot-spots” and “cold-spots” across the country where particular labor and industry sectors are more or less prevalent.

I had two main sources of inspiration for the project. Firstly, I was inspired by my professional experience working in a private foundation, and my conversations with colleagues and peers at similar funding institutions. The choice of using congressional districts to display the data, for instance, was directly inspired by a grant-maker who suggested that this information could be useful in a policy and advocacy context whenever representatives want to know more about research taking place in their district by companies receiving Federal SBIR/STTR grant funding. Secondly, I was inspired by current research at the intersection of fields like economic geography, science and technology studies and science and technology policy studies and research methods and approaches in computational social sciences and scientometrics. I spent time engaging with such work while attending conferences such as the International Conference for Computational

Social Science 2019 and Network Science 2020, and in my coursework at the Graduate Center learning methods in geospatial statistical analysis. Specifically, this project was inspired by those works in the fields just mentioned where “knowledge spillovers” and the interplay between place, region and knowledge production are being studied. To that end, I wanted to come up with a way to enable users to explore a map using contextually relevant external data so that they could ask important strategic questions about the funding landscape in particular places and regions which are differentially impacted by certain issues or research agendas.

This white paper provides an account of the data exploration, processing, analysis, design and visualization decisions that generated the Explorer, and my assessment of the achievements and shortcomings of the work.

CHAPTER 2: PROCESS

EXPLORATION AND DESIGN

The data exploration and design work that led to the Explorer was the result of work with the National Science Foundation's (NSF) awards data as part of coursework in data mining and geospatial analysis classes. In short, NSF's awards includes a qualitative coding scheme that is *somewhat* intelligible to members of the interested public, but which serves administrative ends more so than information discovery ones. The codebook available for tying these "Program Element Codes" and "Program Reference Codes" to specific portfolios or initiatives leaves a lot to be desired, and is not wholly informative to any member of the interested public.[1] As part of my data mining coursework, I observed that in the database of over 450,000 awards the number of codes grew linearly over time and approached over 1,500 unique codes between 1974 and 2018 (see Figures 1A-C in Appendix C). Navigating this large number of codes would either require institutional knowledge, or application of techniques to narrow in the codes or extract key pieces of qualitative information from award abstracts to facilitate discovery and analysis.

As part of the data mining coursework, I applied topic modeling techniques and assessed the possibility of using topic models to explore changes over time. The rationale for this approach was that there were too many closely-related labels to meaningfully use a classification approach, and the mixture model afforded by topic modeling seemed more attractive than clustering the awards into discrete topics. After consultations with a professor familiar with the research funding

[1] See the codebook of the awards database at: https://www.nsf.gov/awardsearch/lookup?type=pec&letter=N*; National Science Foundation. (2020). Awards Database.

landscape, as well as a data scientist familiar with applying clustering techniques on the National Institute of Health's grants data, I decided to select a subset of awards to apply different techniques to and selected the SBIR/STTR awards made by NSF. For the 2,000+ SBIR/STTR awards made by NSF between 2008 and 2018, I decided to take a keyword extraction approach rather than a topic modeling approach and to apply a geospatial frame for information discovery. The result was a dashboard made in R with Leaflet and Shiny, a screenshot of which is displayed in Figure 2 in Appendix C (and which is still hosted online at: https://sethsch.shinyapps.io/sbir_clusters/). The dashboard integrated spatial cluster analysis at the congressional district level for employment and migration data, two variables I encountered while reviewing research in innovation and science policy.

In the later stages of developing the Shiny dashboard, I came across the full SBIR/STTR awards database hosted on SBIR.gov and decided that for the capstone project I would expand the scope to cover all of the research funded by the 11 participating Federal agencies and departments. Because platforms like StatsAmerica's Innovation Index (http://www.statsamerica.org/ii2/overview.aspx) offers users a variety of measures and indicators related to science, technology and innovation policy in a largely numerical format (see Figure 3 in Appendix C) and the SBIR website only offered aggregate analysis on the state level, I wanted to design a method of combining the two sources of information in a way that facilitated discovery and exploratory data analysis.[2] The aim would be to ensure users could evaluate what

[2] The Innovation Index is a product of StatsAmerica. StatsAmerica is a service of the Indiana Business Research Center (IBRC) at Indiana University's Kelley School of Business, and supported by the Department of Commerce's Economic Development Administration.
.

projects have been funded by the various Federal agencies in specific regions, and use contextual quantitative data to come up with boundaries of interest regarding place and region (e.g. districts with higher than average employment in manufacturing sectors, districts with lower than average employment in transportation sectors, etc.). By being able to ask such questions while viewing funding data, the hope is that a user's or analyst's imagination might be activated. What role do certain places or regions play in contemporary knowledge production? Across various places, what knowledge already exists through labor that researchers or innovators might seek to integrate with? These are obviously broad and even speculative questions, but questions for which having concrete pieces of information about past funding trends can be useful.

The design of the data preparation pipeline and the dashboard went hand-in-hand, and my goal was to prepare the data and have a fully functioning web site hosted on Github pages up and running in three months. Early and later mockups of the platform design are provided as Figure 4 and Figure 5 in Appendix C. Crucially, I ensured that all the data processing steps would yield the right data artifacts and keys needed to ensure the platform interactions would work correctly. The following sections detail the data processing steps and the criteria used when making key processing and design decisions.

AWARDS DATA

Award-level data for the SBIR and STTR programs is available through the SBIR.gov awards database (https://www.sbir.gov/sbirsearch/award/all). While the database provides users with the ability to search keywords, topic codes, and companies, and filter by key grant criteria like award year, program phase, federal agency and US state, the geographic information is limited to the state level, and users cannot browse the data according to localities and regions. Additionally, while users can browse proposal solicitations and the solicitation information pages link out to the full call for proposals, neither the linked topic area codes are not clearly tied to the solicitations page nor are the linked grants. I investigated the possibility of better linking this important qualitative data by examining the data structure of the awards, solicitations and topic records.

First, after having a bit of trouble getting the correct parameters set to make API calls to the awards database to download the data, I downloaded the awards data from the database manually. I then downloaded the related solicitations and topics datasets. The data dictionary, including all relevant fields can be found on the SBIR website (https://www.sbir.gov/data-resources) and a saved version is included in the Github repository.

To protect privacy of the persons mentioned in the award records, and because this information was not relevant to my task, I deleted data fields with personal identifiers from the awards data. I then joined all downloaded JSON files into a master JSON and CSV file to work with later on, using my script ***merge_SBIRjson.py***. The resulting dataset here captured the 65,749 grants awarded between 2008 and 2018 from the participating Federal departments and agencies. A table of the key features of the funding by each agency is included here as Table 1A and 1B in Appendix D.

Before attempting to join the solicitations and topic information onto the awards data, which would have provided an additional wealth of qualitative information, I examined the availability of the data. 81.4% of awards included a Solicitation Code and 94.1% of all awards included a Topic Code. In theory, these codes should allow for linking to the solicitations and topics information. However, while attempting to join the information using the script ***join_AwardsTopicSolic.py*** I realized that the identifiers for solicitations and topics were not consistently unique. While URLs in the SBIRTopicLink field were mostly unique where they existed, they resolved to landing pages that also included Topics that had non-unique Topic Codes. For Topics, I attempted to make a composite identifier, taking into account the grant program, agency, topic code and year. For Solicitations, I attempted to resolve formatting and punctuation differences that appeared between the awards, topics and solicitations datasets and caused mismatch. However, even after these efforts only 52.1% of awards were linked to solicitations data and 32.2% of awards were linked to topics data. Due to the incompleteness of the linking effort, I decided not to use solicitations and topics as a source of qualitative information about the grant.

If the Topics and Solicitations datasets reliably included consistently unique resource identifiers (URIs), additional analysis of the similarity between grant abstracts, solicitations and topics would be possible. One could also imagine examining the consistency or novelty of solicitations and topics over time as a proxy for examining thematic grant making priorities across agencies, or examining the geographic spread of cleanly defined topics. Given that the coding schemes were not so neat, I decided to use various methods of keyword extraction to enrich the data. I used two approaches to keyword extraction: 1) lookups from a variety of external authoritative controlled vocabularies and taxonomies; and 2) mathematical approaches to keyword extraction, TFIDF and textrank algorithms.

To prepare the awards data for keyword extraction, I first performed pre-processing on the award title and abstract using a suite of tools authored by data scientists at the National Institutes of

Health's Office of Portfolio Analysis. I found the NLPre package while searching for Github repositories with tools relevant to processing publications or patent data, and later consulted with one of the package authors over Zoom to discuss the utilities included. The library includes a number of utilities often used when working with scientific publications, including: acronym identification and replacement; parenthetical phrase identification; parenthetical phrase extraction; part of speech tokenization; Unicode to ASCII character conversions; reconnection of hyphenated words; depcapitalization of document and section titles; citation separation; URL replacement; replacement of mathematic characters with linguistic tokens (e.g. > →'greater than'; % → 'percent'); token replacement with phrases from user-defined dictionaries. I utilized the suite of preprocessing functions to yield processed text for each award and examined the ***replace_from_dictionary*** function included in the package to determine how best to use custom dictionaries.

The ***replace_from_dictionary*** function yields documents that have been scanned for keywords across a user-defined dictionary; by default the NLPre pipeline includes a CSV file of 192,358 terms the National Library of Medicine's Medical Subjects Heading list. The function expects a CSV with a "term" column and a "replacement" column, allowing for multiple spellings or labels corresponding to a single concept to be replaced for a single preferred label. The dictionary file includes a term column and a replacement column for over 190,000 terms. For the replacement, a given term label may either be swapped for a version where whitespaces are swapped for underscores, or the label may be swapped with either a label corresponding to a broader taxonomic category to which the entity belongs, or a preferred label for entities with multiple ways of being named. In a linked-data context, co-reference resolution is typically managed by the application of URIs to a given entity which can have a single preferred label and many alternate labels. Relations in the ontology usually map terms to broader or narrower concepts, with the multiple kinds of labeling schemes aiding machine understanding of the semantic relationships between the referred to entities and latent concepts. In conversation with one of the package's authors, I confirmed that the packaged MeSH dictionary uses simply uses all 2+ word terms and utilized some bit of semantic "rounding-up" for different label sets. For instance, in the example above "white blood cell count" is one of 16 labels that are all replaced by the more generic "Leukocyte_Count". Given a document as input, a user can define a prefix to be applied to all terms in the document that are found in the dictionary. Any term that is identified is replaced (if applicable) and appended the tag. In the Figure below, note that "white blood cell count" is replaced with "*MeSH*_Leukocyte_Count".

Based on the ***replace_from_dictionary*** function, I created an additional function that would apply the replacement over a series of dictionaries, checking for tagged ngrams and adding them to an index of documents and identified terms from the dictionary. The resulting function ***update_term_index*** is included in the repository of scripts. The function takes a tagged document and a dictionary object in and identifies all tagged tokens and adds them to a dictionary entry for the document, and subsequently de-tags the document so that it can be processed in another iteration of the replacement function. The output for each document in the index is a list of tokens for each vocabulary specified by the user; I wrote an additional function to retroactively count the term frequencies (***count_vocabIndex.py***). A sample of both functions' outputs are shown in Figures 6A and 6B in Appendix C.

Note that each different vocabulary has its own prefix and becomes a sub-entry in the index. This is useful in cases where we want to use several related ontologies or vocabularies to identify terms. In the future, though, to ensure that the replacement schemes for one dictionary don't obfuscate attempts by later dictionaries, I should either modify the replacement function or the indexer function to capture swapped terms and revert the text before it is called again with another vocabulary file. Given the self-imposed time constraints to have processed data ready to use for the dashboard, I neglected to work on optimizing this pipeline for efficiency and reusability. I address this in more detail in the final chapter.

VOCABULARIES, TAXONOMIES AND LINKED DATA ONTOLOGIES

The sources in Table 1 of Appendix D reflect a variety of statistical, economic, bibliographic and policy area classifications and specialized vocabulary that I reviewed to assess usability for the project. Terminology from each source can be extracted and indexed within a corpus, in order to enhance qualitative understanding. Dictionary based topic modeling using sources like these can potentially provide qualitative analysts with a more ready form of evidence to interpret. To enhance dictionary-based extraction in the future, though, a common set of pre-processing techniques (including lemmatization, stemming, etc.) might be applied to both the dictionary and the source text.

Selected vocabularies and ontologies were either downloaded in RDF/XML, TTL, XLSX or CSV formats in total or after completing a SPARQL query for the alternate and preferred label in the scheme. For a few resources, I used the SPARQL endpoint with the following queries to extract English labels. Both SPARQL queries that I used are included in the zipped repository files. For sources which did not have an immediately accessible SPARQL endpoint, I used the open source ontology management software Protege to open files in RDF/XML, NT, TTL formats, identify individual entities and extract relationships into CSV format. Given my novice SPARQL skills, I found it easier to manipulate the CSV using Excel and Python.

For the selected vocabularies, I ensured that all alternate labels appeared as a term to be replaced by a processed preferred label (with punctuation swapped for underscores). Additionally, all preferred labels themselves are included with a processed replacement label. For a few resources, including the STW Thesaurus of Economics and the UN Food and Agriculture Organization AGROVOC thesaurus, alternate and preferred label relations also encoded some level of taxonomic relationship. Besides allowing these replacement schemes to persist (so that the preferred label replaced all instances of alternate labels), I made no additional effort at this time to do any semantic "rounding-up" within taxonomies. In order to make more informed semantic text mining decisions, I'd like to gain more facility navigating and extracting elements from RDF.

Successfully running the edited pre-processing pipeline took me to the limits of my Python programming knowledge. Being aware of the utility and efficiency of the batched job queuing employed within the NIH word2vec pipeline, yet unable to fully grasp how to edit the package to successfully implement my own indexing function given time constraints, I resorted to taking a piecemeal approach. First, I ran the half of the processing pipeline using the command line data import and parse steps from the word2vec pipeline. This included steps to identify acronyms and abbreviations in the text, handle unicode decoding issues, handling hyphenated words, removing title caps, replacing acronyms with their expanded forms, and expanding parenthetical comments into standalone sentences. Next, using this partially processed output, I ran the replace and update dictionary steps within my iPython console from Spyder. I did this for seven of my vocabularies, excluding the Medical Subject Headings (MeSH) dictionary built in to NLPre (containing ~192K terms), which was much larger than my other dictionaries (ranging from 420 terms for the EIGE Gender Equality Thesaurus to ~42,000 terms for the United Nations AGROVOC thesaurus). In the future, I would re-write this script to do batch processing like the word2vec pipeline does; it took upwards of 36 hours to process the 65,000 records and create an index of all found vocabulary terms. This is quite obviously not optimally performant for a production-grade pipeline, but on the bright side it at least gave me time to step away from the computer. When I subsequently ran the MeSH dictionary using the in-built functions in the word2vec pipeline, it took only about 3 minutes to process the records. I then imported the tagged documents into the ***update_term_index.py*** script and updated my main document index with the results from the MeSH tagging. Last, on the tagged output, I ran the final pipeline steps -- including token replacement to take care of punctuation and special characters, and a part-of-speech tokenizer to remove words irrelevant to the subject matter of the award.

GEOPROCESSING AND SPATIAL CLUSTERING

I used the service Geocod.io to perform geocoding using the Address, City, State, ZIP code fields for all ~29K unique recipients in the awards dataset. The service added 2018 census information including Census Tracts, Blocks, Metropolitan Area Divisions, and congressional district information for the 117th Congress (2021). In approximately 14 minutes, I could download the file at a cost of approximately $40. Even though I was going to have to join additional congressional and county level identifiers in later, I thought it was worth the cost to get started. There were 75 of 29,000 recipients (0.25%) that required manual validation, which I did using Google Maps.

With the geocoded recipients output, I then used qGIS to join the additional congressional district and county level identifiers. Adding the county and district GEOIDs make the geographic queries possible on the dashboard tool. Once I had a fully geo-coded recipients file, I then joined the additional geographic information back into the awards file using the script ***join_geocodedRecips.py***.

The awards data includes repeat recipients that sometimes have several DUNS Numbers or multiple addresses. The qGIS output files included identifiers for the 113th and 116th Congresses, as well as county FIPS identifiers. I joined the output back into the awards data by using the recipient DUNS#, name, and address; only 160 of 65,749 (< 0.25%) required manual validation. Using the various iterations of geocoded recipients data, I was able to manually correct these 160 awards by triangulating across various outputs. I only had to manually validate 1 additional recipient to complete this step.

An area of interest for me when it came to exploratory data analysis was to provide relevant contextual geographic data as a way to navigate the map and think about the knowledge production patterns. While working in the philanthropy sector, this has become an area of interest to me. I n discussions with others working in the philanthropy sector, I asked when and where external data sources are used to analyze geographic funding choices. In some instances, I heard that geographic data was analyzed on an ad-hoc basis to provide insight into the demographic makeup of places and regions, as well as the workforce makeup of regions (e.g. where artists are located). These discussions also brought me in contact with the National Endowment for the Arts' Arts Data Profile Series (https://www.arts.gov/impact/research/arts-data-profile-series) and the Indiana Business Research Center's StatsAmerica Regional Innovation Indices and profilers (http://www.statsamerica.org/Default.aspx). These two curated sets of data and indicators related to arts economies and artistic production, and regional innovation and R&D activity, respectively, inspired my interest in looking for external contextual data to provide a backdrop for exploration.

Ultimately, the choice to specifically include workforce composition data specifically was influenced by work on R&D knowledge spillovers. This work examines how the proximity of innovative firms to each other effects the spatial distribution of innovative activity and associated economic effects (Anselin et al., 2000; Bonaccorsi & Daraio, 2005; Jaffe et al., 1993; Wallsten, 2001). However, while many researcher in the literature explain the relevance of spatial proximity by pointing to the tacit, often "non-codified" social dimension of knowledge exchange, I encountered others who rather suggested that a more complex interplay between labor markets, institutional collaborations and legal arrangements, and other political-economic factors are as much a cause of spatial clustering of innovative activities as are more rudimentary social and communicative factors (Breschi and Francesco, 2001). Seeing that quite a few researchers have remarked on the localization of specific industries and meanwhile also disaggregate findings of innovation spillovers or localizations across specific types of research domains. (Anselin et al., 2000; Audretsch & Feldman, 1996; Boschma et al., 2014), I thought it would be interesting to identify spatial clusters by workforce and provide users with this as a backdrop to guide browsing and investigation into the research activities being funded in different parts of the country. While the Innovation Index from StatsAmerica, for instance, included a battery of variables and indicators in their platform, I ultimately chose to select one for this application--the workforce composition of each county and congressional district by industrial sector.

I obtained the workforce data from the American Community Survey's 1-year and 5-year estimates and proceeded to explore the data in qGIS and R. As there clearly were non-random spatial distributions of higher density by industry (e.g. high proportions of manufacturing labor in the Midwest; high proportions of labor in Finance and Real Estate, and Science, Management and Technology sectors in particular large metropolitan areas), I proceeded to identify the clusters using geo-spatial statistical techniques. Using my R scripts ***LISA_stats_acsIndusry.R*** I performed clustering using the Univariate Local Moran's I test to produce a local indicator of spatial association (LISA) statistic for each unit of space, a clustering strength indicator (Moran's I) and a significance value of the clustering (p-value). I explored the data at both the county and congressional district levels.

The aforementioned script produced LISA statistics and the congressional district level and I wrote a variant ***LISA_stats_acsIndustry_COUNTY.R*** to obtain the stats at a county level. For each spatial unit I obtained a variable that I named according to a generic scheme `"LSAcl + _ + [industry]"`, so that an area's cluster group value could be obtained for any given industry as the user would change industries on the choropleth. As shown in the data here, I found there to be significant clustering across all industries, but there were stronger clustering effects (higher Moran's I values) at the congressional district level than at the county level. This made sense intuitively, since, on the one hand, congressional districts could integrate large swaths of sparsely populated counties, and on the other hand, districts could be relatively small and densely populated and next to many other small densely populated districts where spillover effects are likely more significant as compared to sparsely populated areas.

I ultimately chose to use the congressional district level as the unit of analysis, because it lends itself to asking questions about how R&D policy agendas may or may not be taken up by representatives. Exploring the labor composition of a district alongside its SBIR/STTR funding might provide policy analysts with a sense of the interests representatives have in certain future-oriented policy areas. Additionally, exploring the workforce data at the congressional district level is interesting because of the more or less constant population of districts; it is interesting to compare the differences in labor forces over varying population densities and varying sized districts. This choice was also influenced by a conversation with a peer at a funding institution, who highlighted the potential utility in advocacy and policy contexts of being able to quickly get summary stats regarding funding activity at the district level. The Public Innovations Explorer was ultimately designed to be able to deliver on this premise.

Along the way, I also produced % change measures for both units by comparing the 5-year ACS estimates for 2013 and 2018. The rationale for doing so was to identify whether there was any statistically significant spatial association with regard to changes in workforce composition. This is a relevant question for anyone interested in studying regional industrial change, economic policy or innovation policy. In this context, the Moran's I value provides evidence of whether the percentage of the population employed in a given sector tended to change more or less depending on whether the percentage employed in that sector in a neighboring spatial unit also underwent some level of change. On a practical level, if this clustering were indeed spatially significant, we could use it to identify hot-spots (strong increases) and cold-spots (strong decreases) of workforce composition change and ask whether certain underlying socio-economic factors or local policy changes in the corresponding places had some effect on the regional labor force changes. However, I ultimately found that there were not significant spatial effects in % changes, and I left that analysis and line of questioning behind.

UI DESIGN & VISUALIZATION

During the data processing phase I was also tweaking the user-interface design by exploring reusable D3.js charts and libraries, creating mockups in Adobe Illustrator and XD, and looking through freely available UI kits. I decided to build the project from Node.js mostly out of a desire to continue learning Javascript coding, and also because it afforded the most customizability as compared with using a more out-of-the-box product like Tableau or PowerBI.

I settled on using a parallel coordinates chart as my secondary UI element besides the map for a few reasons. A parallel coordinates chart's main draw is that it can afford a user with the ability to see relationships between many variables at once. The visual logic of the chart is that each line in the chart represents an object or observation (in my case, a congressional district), and where the line passes through an axis represents the value for pertaining to the district for the given attribute (in my case, a proportion of its labor force working in the given sector, or the funding received by a given agency). As compared to a scatter or line plot, where you only have two dimensions in play, with the parallel coordinates chart you have as many dimensions as you might want, with the only limit being visual clutter. As compared to a (faceted) violin plot, which can help visualize a distribution of observations and the variance, with the interactive parallel coordinates chart you can better understand the inter-relatedness of observations across attributes and understand the attributes as inter-related continuous "levels" rather than conceiving of them as potentially separate discrete observations.

I encountered two variations of a parallel coordinates chart package for D3, and was happy with the features already available to help reduce visual clutter. Those features were path bundling (meaning the paths would be less spread out around the axes so a user sees how paths cluster within certain dimensions), axis reordering, and color switching. By reordering the axes, a user could evaluate relationships between certain attributes seem and focus in on them. Extending this functionality, I added a UI element that a user could click to totally remove the axis, allowing them to focus in further on just a few axes. Additionally, the ready-to-use functionality that allows a user to click on an axis and switch the color palette of the chart means that a user can focus on how districts that are high or low in one dimension are observed across other dimensions simply by following groups of lines with a certain color range. Using a z-score computation to base the color palette for my labor data and a quantized scale for my funding data meant that my color scale went from blue ("strongly below average" or "none to low"), to yellow ("average" or "medium"), to red ("strongly above average" or "highest"). Since the color palette can only apply to one variable at a time, for the selected axis a user will always see the red values at the top of the scale and blue values at the bottom. Those lines then continue on to other axes in the same color, allowing a user to follow the lines and see whether inverse relationships or any kind of clustering relations seem to exist.

I found a few examples of successful implementations of the parallel coordinates chart where a user could interact with the chart to filer and select data within other components of a display (see, for instance, the examples section here: http://syntagmatic.github.io/parallel-coordinates/). In one project exploring malaria rates (https://realimpactanalytics.github.io/d4g-hackathon-malaria-viz/), the chart was being used to filter counties in Kenya. Applying brushes to select portions of the axes would highlight certain counties on a map of Kenya and would also filter a corresponding table. I sought to replicate this functionality and forked the code and utilized the chart and table interaction in my own code.

I explored using D3’s geojson library (https://github.com/d3/d3-geo) to dynamically display sub-regions of the United States on a user selection, but I was unhappy with the re-scaling and I also wanted to avoid displaying districts outside of a holistic geographic context since some districts are wholly unrecognizable shapes. Spending more time experimenting might have resolved these issues. However, since the underlying map offers lots of useful contextual information that makes for a more geographically grounded viewing experience, I thought using Leaflet (which I had used previously for the R/Shiny dashboard referenced above) was the way to go and simply had to reorient myself to the Javascript syntax.

With the components researched and tinkered-with to ensure viability, I turned to exploring UI kits for the first time. I settled on Bootstrap because it had all of the components I could have wanted, and I played around copying and pasting the elements to make a design in Adobe XD. I then began scaffolding the interface within HTML to ensure all of my elements would have a place. Once that was ready, I began chugging away at my Javascript code to load the components and set up the full range of interactions. I learned a great deal of HTML and CSS along the way, as this is the largest web-development undertaking I have completed to date. Additionally, while I had some experience compartmentalizing my Javascript code into cleaner modules, I didn’t go into this project with a strong application development background and ended up having one large Javascript file housing all of my components.

Regarding the database design, I initially wanted to allow a user to dynamically filter or locate recipients based on keyword information as would be displayed in the Topics menu, but I realized that doing so would require a setup difficult to reconcile with using Github Pages for hosting. Github pages allows free hosting of static webpages, with the underlying repository serving as a database. I explored other hosting solutions in consultation with the Graduate Center's digital librarians, but given the time constraints and my success in hosting the page on Github up to that point decided it was not worth the trouble if I wanted to finish the project on schedule. As a solution, I created individual files for each district that included aggregations of keyword frequency by agency and year. This allows a user to filter dynamically by agency and year and see the corresponding total frequency for keywords, but does not create a linkage allowing them to identify the pertinent grants where those terms were mentioned. To reduce loading time, I ensured that these files would only be loaded from the Github repository when a user selected a district. In theory, because I had included a list of grant identifiers in these files, I could have created a way for the pertinent grants to be queried. However, in my attempts to create 65,000+ individual JSON files with extracted keyword information, I encountered a serious bottleneck that suggested it would be futile to attempt to load these files into a Github repository. In the future, I'd like to learn more about database structure for text indexes, and better understand use cases and trade-offs associated with using SQL databases versus non-SQL databases like MongoDb or graph databases. At one point, I had also hoped to be able to deliver on featuring keyness analysis when displaying topics. Keyness analysis refers to a set of computational linguistic techniques to identify the key terms most unique to a given corpus as compared to another, and in this context could have applied in comparing the awards for the selected district to the rest of the districts as a whole (Gabrielatos, 2018). I have used this in narrower analytic contexts, but it is a non-trivial computational operation to make available on-the-fly while filtering the corresponding corpora among multiple dimensions and would have required more intensive consideration of how to design the right database structure. In lieu of offering this feature, I integrated a feature that allows a user to click to get the underlying data generating the Topics section for the selected district, so that they can proceed to perform their own analyses.

CHAPTER 3: OUTCOMES

The Explorer has been tested on a variety of computers and browsers, and has always loaded in a reasonable amount of time (~2.5 seconds). A final user-interface diagram for the deployed version of the Explorer can be found as Figure 7 in Appendix C. Additionally, the introductory text presenting the Explorer takes a user time to read and in that same time the initial dataset—which, by design, does *not* include all Federal agencies by default—loads. I have done my best to address and limit lingering UI/UX issues that a more seasoned web developer would surely be able to address—most significantly to me, ensuring that the components resize with the window. For the most part the components do resize appropriately, and where the parallel coordinates chart glitches the reload button included in the Profiler Menu redraws the chart appropriately for however the window is currently sized.

There are a number of features built to provide a user with additional context and guide them through using the tool. Included with the main dashboard is a User Tips popup that includes GIF images of features on the parallel coordinates chart, information about using the map, and the funding details sections. Additionally, there is a more detailed User Guide that sits as a standalone page, navigable by clicking the About dropdown menu. The User Guide proceeds feature by feature and includes details on how a user might utilize the components to answer particular questions about innovation research funding. Last, I have included a standalone page with Background information about the project and its sources of inspiration, and another with information on the Data Sources.

CHAPTER 4: REFLECTIONS & FUTURE WORK

Creating the Public Innovations Explorer afforded me the opportunity to learn and develop my skills on multiple fronts. Through this capstone project I have learned more about the relationship between information management and data modeling that exists through metadata curation and maintenance and have a deepened appreciation for how upstream modeling decisions not only affect information discovery but also have impacts on issues like public accountability and transparency and the possibilities of scientific research. Related to the issue of data modeling and information management are the nuances of database structure and design, and the way that leveraging contemporary data science techniques requires planning for a series of transformations in data structure. While researchers and analysts know this and appreciate it at some level, typically when one undertakes an analysis they seek to mitigate any issues or messiness and move on to doing whatever it takes to answer the questions they set out to answer. From a data or information management point of view, or from an analytics development point of view, it is important to anticipate the range of possible inquiries and plan ways to keep open as many pathways to questions and answers as is possible. Reflecting on conversations and brainstorming sessions I had with professors and colleagues before sketching out the project, I realize how this tension between *doing analysis* and *developing analytics* animated many of the conversations. It was important for me to realize how visualization in service of particulars explanations or concrete, finite analyses is a distinct practice from visualization in service of information discovery and exploration.

As I previously mentioned, this project took me to the limits of my current knowledge when it comes to application development and database design. Both in appreciating the necessity of carefully designing a data model and database structure that could potentially allow for on-the-fly keyness measures when it comes to topics and keywords, and also in loading a large dataset directly in the browser to populate map popups for each recipients, I had to deal with the limitations of using a static website hosting service and not having thought out a true backend database. To make the user-interface sufficiently performant when allowing users to click recipient icons and see their respective award details, for instance, rather than populate all recipient popups in Leaflet with information at the initial load I designed a function that grabbed the applicable awards from the award dataset and populated the pop-up only as needed. In this instance, and also in the instance described above regarding the development of my vocabulary indexing utility, this project helped me to develop a more keen sense of dealing with computational limits and has spurred a desire to learn more about efficient and performant scaling of prototype-level code. Practically speaking, especially in an organizational context, it might also be the case that using pre-built tools like the Semantic Web Company's PoolParty (https://www.poolparty.biz/) or OntoText's Metaphacts (https://metaphacts.com/) would be the most efficient manner of performing vocabulary, taxonomy or ontology-driven text mining and that developing code to do this, while perhaps more cost-effective, might not be sustainable.

When it comes to the user interface and the imagined audience of analysts and interested members of the research and funding communities, I was always afraid of having attempted to do too much. I acknowledge the extent to which my concern with the economic geography and scientometrics literature is a fairly specialized interest, and that that interest may not have been translated as well as it could be to end users. For instance, when creating the map, I used my LISA statistic clustering groups to draw red borders around districts belonging to so-called "hot-spots" and blue borders for districts in "cold-spots." However, I ultimately chose not to draw the user's attention to specifically to these borders. My rationale was based on a few considerations: 1) the spatial clustering is not *of funding data* but rather of this background contextual data, and it might confuse users (as some early feedback from a peer indicated); 2) I didn't want to draw so much attention to the method in case it would distract users from the overall utility of the tool or confuse them; and 3) the clustering data is not a conclusion in itself that leads to a quick, easily digestible conclusion, but rather offers one additional way to browse the map. In line with the 3rd reason here, this is also why I chose not to proceed to explore a battery of variables for presentation but rather to go with one measure only. In future work, I'd like to continue exploring the use of geo-spatial statistics to produce statistically-sound features that can guide geo-spatial data exploration. Especially with smaller units of space, I think that these methods can help to identify non-intuitive, non-categorical groupings of space. Moreover, I appreciate how the parallel coordinates plot allows a user to explore many inter-related dimensions at once on a map and hope that that functionality alone will be an interesting takeaway for others. In making a tool specifically geared toward analysts, having statistically tested clusters accompany choropleths can speed up exploration and hypothesis formation. I hope that the design of the map and parallel coordinates interaction in the Public Innovations Explorer might be seen as a reusable and reproducible one, useful for other inquiries across other domains of geospatial data analysis.

In taking up the suggestion to present funding information at the congressional district level, I also encountered another point of tension with regard to data analysis and information design. Accompanying this suggestion was the idea that the browser could generate fact-sheet PDFs ready to be printed out and used in a policy or advocacy context by policymakers interested in making arguments related either to research funding and practice in their district or the specific scientific and technological contributions being made by companies located in their district. I explored this idea with one of my data visualization professors, and learned about the technical requirements likely associated with this idea. Realizing the idea would have likely required use of unfamiliar Javascript frameworks, and I was advised that setup and debugging, even for someone familiar with the frameworks, would likely have taken an additional four to six weeks. Because that was not realistic within the scheduled timeframe to complete the project, it was excluded from consideration. Instead, I integrated a feature in the Topics display that allows a user to access the aggregated vocabulary statistics file for the selected district from the underlying Github repository and use it as they would like. Meanwhile, while not quite as seamless, I imagine that the information displayed in the Explorer for a district in the District Summary section and in the other tabs could still be grabbed manually by any analyst interested in putting together a quick fact sheet. I would like to learn more about automated reporting and print-layout generations in the future.

In conclusion, I am satisfied with the level of interoperability, information discovery and pipeline development I was able to pull off individually in more or less in three months. The Public Innovations Explorer aims to facilitate information discovery and exploration of funding flows, and aid analysis regarding the research agendas receiving funding so that one can imagine what else might be funded in the future. Taken as a whole, I hope that this the project can also facilitate conversation and exploration within the research and funding communities and serve as a prototype for future analytics development that integrates text mining with external indicators and data on funding flows.

APPENDIX A: LIST OF VARIABLES

cd116_5yearACS2018_LISAclust.js – geojson data

STATEFP	State FIPS code, string.
AFFGEOID	Full GEOID for district, string.
GEOID	Last 4-digits of AFFGEOID, integer.
CDSESSN	Congressional District Session (“116”), string.
DISTR_NAME	Full-text district name, string.
IND_PROFILE	String with district’s top-5 sectors by proportion of total workforce, and corresponding national rank in that sector.
LSAcl_.pct_[*SECTOR*]	For each industrial sector, with sector name in [*SECTOR*], the LISA clustering statistic: “1” denotes hot-spot, “2” denotes cold-spot, “5” denotes neutral.
GEOMETRY	List of projected coordinate strings in EPSG3857 for polygon geometry.

[*AFFGEOID*].json – aggregated vocabulary statistics for each district, by AFFGEOID

AGENCY	Funding agency, string.
YEAR	Funding year, integer.
WD_CT	Total word count of included grants, integer.
AW_CT	Award count of total included grants, integer.
IDS	List of award identifier strings (concatenated award tracking number and contract number).
[*VOCAB_METHOD*]	For each vocabulary or extraction method, contains a dictionary of key, value pairs where the key is a keyword and the value is its raw total frequency.

acs2018_industry_congdist.csv – ACS labor statistics for each district

GEOID	District GEOID, string.
YEAR	Funding year, integer.
DISTR_NAME	Full-text district name, string.
IND_PROFILE	String with district's top-5 sectors by proportion of total workforce, and corresponding national rank in that sector.
TOTAL_WORKER_POP	Total population of in workforce in district, integer.
PCT_[*SECTOR*]	For each labor sector, the percentage of the workforce employed in that sector, float.
IND_PROFILE	String with district's top-5 sectors by proportion of total workforce, and corresponding national rank in that sector.

sbir_2008to2018_geoRefed.csv – geo-referenced award-level data

_REF	Reference ID outputted from NLPre pipeline, integer from 0 to 65,749.
COMPANY	Company name, string.
AWARD_TITLE	Award title, string.
AGENCY	Agency name, string.
BRANCH	Agency branch name, string.
PHASE	Funding phase, string.
PROGRAM	"SBIR" or "STTR", string.
AWARD_YEAR	Award year, integer.
AWARD_AMOUNT	Award amount, integer.
DUNS	Company DUNS identifier, string.
HUBZONE_OWNED	If company is located in an SBA Hubzone, "Y," else "N", string.

SOCIALLY_AND_ECONOMICALLY_DISADVANTAGED	If company is at least 51% owned by socially or economically disadvantaged groups, “Y,” else “N”, string.
WOMAN_OWNED	If company is at least 51% owned by women, “Y,” else “N”, string.
NUMBER_EMPLOYEES	Number of employees at company, integer.
ADDRESS1	Address line 1, string.
ADDRESS2	Address line 2, string.
CITY	City name, string.
STATE	State name, string.
ZIP	Company Zip code, string
RESEARCH_KEYWORDS	List of keywords provided by agency, string.
LATITUDE	Company latitude, float.
LONGITUDE	Company longitude, float.
GEOID_CD113	GEOID of congressional district company is located in for 113th Congress (2013), string.
AFFGEOID_CD116	GEOID of congressional district company is located in for 116th Congress (2018), string.

cd116_sbirRecipients_epsg4326.csv – list of unique award recipients and locations

STATEFP	State FIPS code, integer.
CD116FP	Two digit FIPS Code for congressional district in 116th Congress, integer.

GEOID	Four digit GEOID for congressional district in 116th Congress, integer.
AFFGEOID	Full GEOID for congressional district in 116th Congress, string.
DISTR_NAME	District full name, string.
COMPANY	Company name, string.
ADDRESS1	Address line 1, string.
ADDRESS2	Address line 2, string.
CITY	City name, string.
STATE	State name, string.
ZIP_1	Company Zip code, integer
COUNTY	Name of county the company is located in, string.

cd116_agency_year_fund_aggs.csv – list of aggregated funding stats by agency and year for each congressional district

NAME	District Name, string.
YEAR	Funding year, integer.
DOD	Total funding from Department of Defense, integer.
ED	Total funding from Department of Education, integer.
HHS	Total funding from Department of Health and Human Services, integer.
DOT	Total funding from Department of Transportation, integer.
DOE	Total funding from Department of Energy, integer.
NASA	Total funding from NASA, integer.

NSF	Total funding from National Science Foundation, integer.
USDA	Total funding from Department of Agriculture, integer.
EPA	Total funding from Environmental Protection Agency, integer.
DHS	Total funding from Department of Homeland Security, integer.
DOC	Total funding from Department of Commerce, integer.
GEOID	District GEOID, string.

APPENDIX B: GLOSSARY OF FUNCTIONS

Data Manipulation and Pre-Processing

agg_awardDistKeyword.py: Aggregates award-level keyword stats for each district by Federal agency and year.

analyze_vocabCoverage.py: Compares vocabulary coverage between dictionaries and mathematical algorithms across awards.

break_keywordIndexJson.py: Creates individual JSON files at the congressional district and award level for vocabulary data.

convert_json2csv.py: Converts JSON of awards data to a CSV.

count_vocabIndex.py: For every award and its corresponding vocabulary index, generates term frequencies for all extracted keywords.

get_mathyKeywords.py: Performs keyword extractions on award abstracts and titles using TF-IDF and Textrank algorithms.

get_prefLabel_altLabel_En.sparql: SPARQL query to get preferred and alternate labels in English from select SPARQL endpoints for various linked-data taxonomies and ontologies.

get_SBIR_distAgencyYear_aggs.py: Takes the geo-coded awards data and creates aggregated funding stats by congressional district, year and funding agency.

get_SBIRstats.py: Generates summary statistics regarding funding by award-program, organization type and other criteria for each funding agency in the awards database.

join_AwardsTopicSolic.py: Script to explore the viability of joining award level data to available Topics and Solicitations data from SBIR.gov.

join_geocodedRecips.py: Joins award data back up to geo-coded recipient data.

join_vocabIndices.py: Joins the vocabulary indices for each award back up to the award data.

LISA_stats_acsIndustry_COUNTY.R: Performs Univariate Local Moran's I test and joins LISA statistics for a series of variables to a geographic shape file; this variation of the script was used while working with county level data.

LISA_stats_acsIndustry.R: Performs Univariate Local Moran's I test and joins LISA statistics for a series of variables to a geographic shape file; this variation of the script was used while working with congressional district level data.

merge_SBIRjson.py: Merges JSON downloads from SBIR.gov into a single JSON file.

prep_vocabFiles.py: A variety of cleaning and processing steps used for transforming exports from Protégé ontology editor into CSV files ready for use with the **replace_from_dictionary** and **update_term_index** scripts.
replace_from_dictionary(prefix, csv_dictionary_path)(document): Replaces terms found in dictionary CSV with replacement from dictionary and appends prefix (Note: this function is from NIH's word2vec/NLPre pipeline).

update_term_index.py: Provides a class to use with the NLPre pipeline that updates a user-defined dictionary with a vocabulary index for a specified document; the output is an updated dictionary with document IDs as a key and a sub-dictionary containing the vocabulary name and a list of keywords found.

updateDictionary_NLPre.py: Runs the **replace_from_dictionary** and **update_term_index** functions on the awards file to produce JSON indexes for each award in the awards dataset.

Visualization

main.js: Loads SBIR awards data, ACS data, and district-level aggregated vocabulary files to generate the Public Innovations Explorer while utilizing libraries D3.js, Leaflet, Leaflet Marker Clusters and SlickGrid. Major functions utilized by the file are described below *in order of appearance* in the script:

mergeVocabs(vocabdata): Takes a set of keyword and frequencies for a series of vocabularies and merges their frequencies; called when a user opts to show All sources of topics.
init(): Initializes map, information bar, awards and labor data and UI elements.

draw(): Draws parallel coordinates chart, sets the default color and wires up interactivity between UI elements, the table and chart.

changeYears(): Called when the user changes selects update after changing the selected year range, changes the state to reflect the year range.

changeVocab(): Called when the user changes the selected vocabulary source options, changing the tabs appropriately and calling the **showCdVocab** function to change vocab that is displayed.

getCdStats(district_id): Generates a funding summary for the selected district that can be used to display funding by agency, a table of recipients or KPIs for grant details; calls **getCdTextSumary** by passing the district_id and its set of awards and passes the summaries to **showCdGraph**.

updateHides(dimension): Called when a user clicks the icon to remove an axis from the parallel coordinates chart; adds that dimension to the list of hidden axes and re-draws the chart.

change_map_color(): Called when a user selects a new dimension on the parallel coordinates chart, to also change the color scale used for coloring the choropleth map.

change_color(dimension): Called when a user selects a dimension on the parallel coordinates chart to make that dimension the color scale.

choroQuantize(data): Called when the parallel coordinates chart is displaying funding data; generates a quantized scale given the district-aggregated funding data of the selected agency for use with the choropleth.

quantizeColor(data, dimension): Called when the parallel coordinates chart is displaying funding data; generates a quantized scale for use with the parallel coordinates chart.

zcolor(data, dimension): Returns a color function based on the data for the selected dimension by calling the **zscore** and **zcolorscale** functions.
zscore(col): Converts values to a zscore for the given column of the data.
dimensionStats(col): Returns the mean and standard deviation of the given column of the data to supply the information bar with statistics pertinent to the selected or hovered-over dimension.

stringFilter(item, args): Returns the index of all observations (districts) matching the string inputted into the states/districts search bar.

comparer(a,b): Called to sort values in the table.

gridUpdate(data): Called to update the table when the parallel coordinates chart is brushed or the search bar is being used.

preClear(searchString): Resets the map zoom and unmarks/unhighlights districts in the parallel coordinates chart when the search string is being removed and is less than the buffer length.

showQueryPaths(): Highlights the districts in the parallel coordinates chart within the set brushed or the input search string and zooms to the bounding region on the map.

toTitleCase(str): Converts string to title case; used for normalizing company names on popups.

initParcoords(): Called by the main **init** function, to initialize the parallel coordinates chart and the connected SlickGrid table.

initAwardsData(): Loads the award-level data and the geo-coded recipients data from respective CSV files, then calls **filterAwardsRecips.**

filterAwardsRecips(): Filters the awards and recipients based on the selected agencies and years, calls **addRecipsToMap, getCdStats** and **getCdVocab**.

switchParcoordsData(): Called when a user switches the parallel coordinates chart display from displaying district-level labor data to funding data, or vice versa; calls out to re-initialize and re-draw the parallel coordinates chart.

addRecipsToMap(d): Adds geo-referenced recipients data to the map, within marker clusters.

getMarkerAwards(id): Called when a user clicks a recipient icon on the map to generate popup info on demand, because doing so all at once is too computationally expensive. Passes the id of the marker and grabs corresponding awards so that a user can scroll through received awards on the popup.

initIndustData(): Loads the geojson that includes the ACS labor data and spatial clustering data.

style(feature): Generates style for choropleth polygons (districts).

getFillOpacity(feature): Gets fill opacity for choropleth districts, depending on selection state.

get_infobar_stats(dimension): Places stats for the selected or hovered-over dimension of the parallel coordinates chart inside the information bar.

highlightFeature(e): Coordinates highlight on mouseover of the choropleth map with the parallel coordinates hover.

resetHighlight(e): Resets highlight on mouseout for the chloropleth map and parallel coordinates chart.

zoomToFeature(e): Called on district click to zoom in the map.

onEachFeature(feature, layer): Applies the highlighting and zooming behavior to each map polygon.

brushMap(): Connects the parallel coordinates chart's brushing behavior to the map highlighting behavior.

ordinal_suffix_of(i): Generates ordinal suffixes for tidy information display in the district summary text.

getCdTextSummary(district, awards): Given the district and its set of awards, generates a textual summary of the funding activity in the selected district.

getCdVocab(district): Loads the corresponding aggregated vocabulary statistics file from the Github repo database any time a new district is selected, then calls **showCdVocab**.

showCdVocab(data): Creates UI pill-badges for each vocabulary item in the data; called to change the keywords displayed in the Topics panel any time a user selects the vocabulary source.

getFeaturesInView(map): Returns the map layers so that when a user clicks a recipient in the recipients table the map can open the popup from the recipient marker layer.

showCdGraph(fundSummary, countSummary): Given funding summaries with dollar amounts and award counts, generates the funding agency bar graph, recipients table or grant details KPIs within the Funding Details display.

getMarkers(map): Returns the markers currently on the map.

getDistricts(map): Returns the district polygon layers from the map.

APPENDIX C: FIGURES

Figures 1A, 1B, 1C: “Program Element Codes” are used by NSF to add qualitative tags to grant records denoting the area of work or portfolio the award belongs to. Figure 1A (top left) shows the number of tags applied to awards; Figure 1B (top right) shows the total number of unique tags existing in the database year by year; Figure 1C (bottom) shows how the number of tags in use per

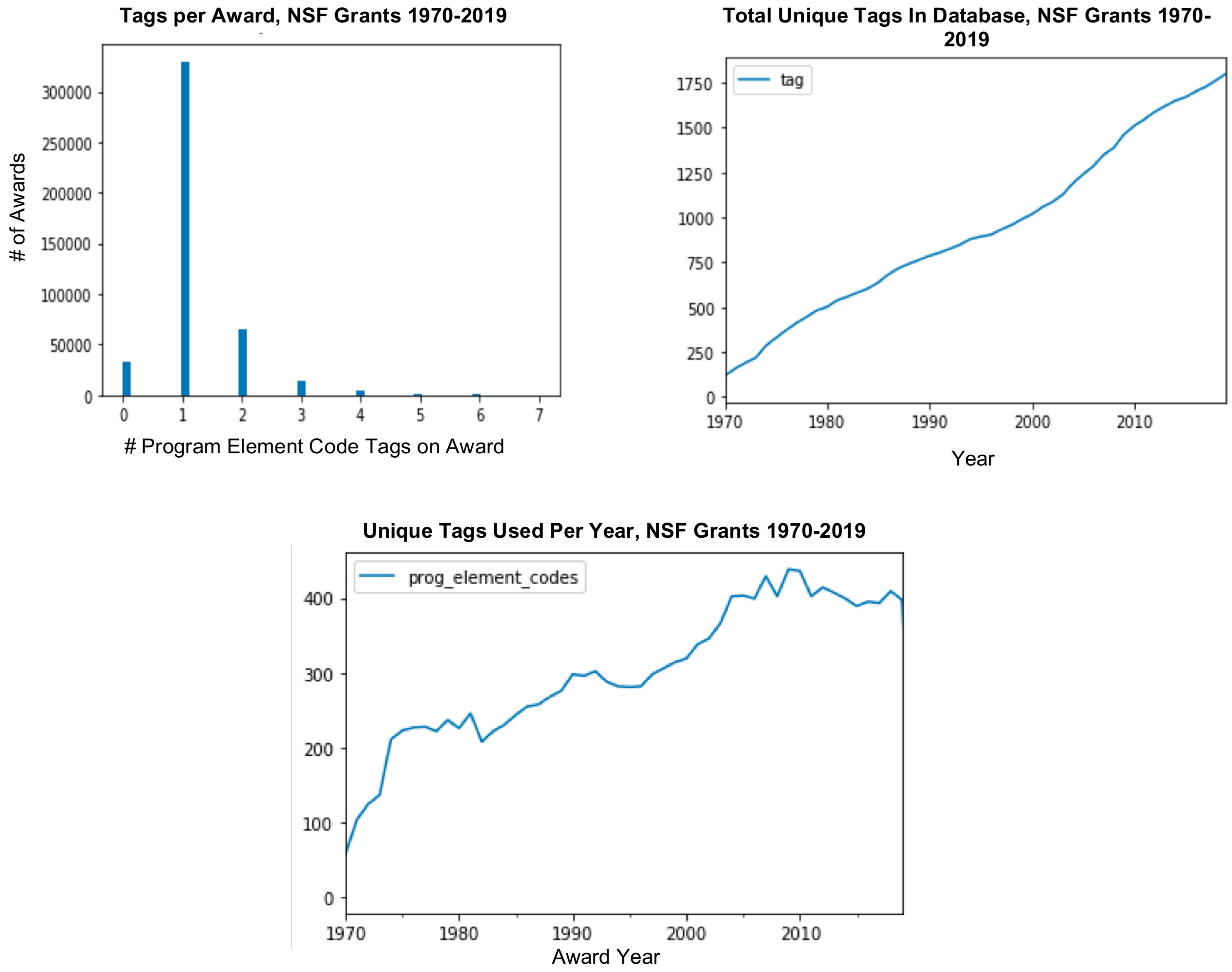
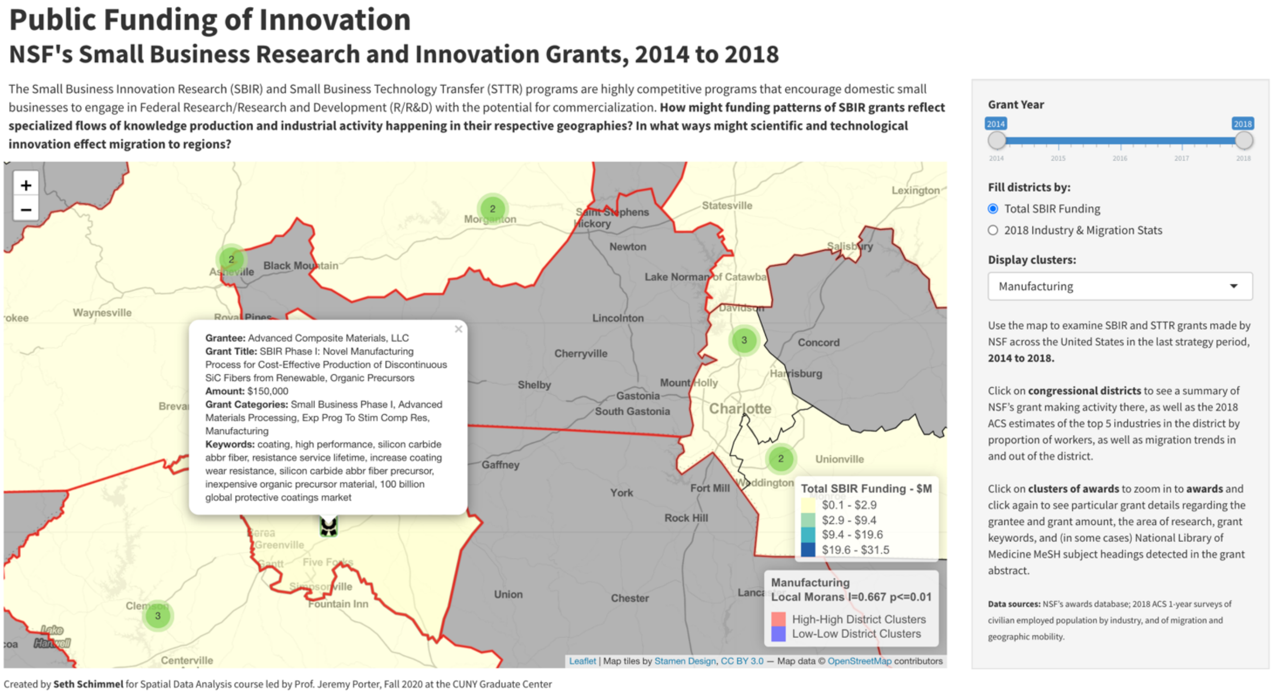


Figure 2: Screenshot from the Shiny dashboard prototype, with a popup showing award details and cluster display highlighting Manufacturing-heavy districts.



Public Funding of Innovation

NSF's Small Business Research and Innovation Grants, 2014 to 2018

The Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs are highly competitive programs that encourage domestic small businesses to engage in Federal Research/Research and Development (R/R&D) with the potential for commercialization. **How might funding patterns of SBIR grants reflect specialized flows of knowledge production and industrial activity happening in their respective geographies? In what ways might scientific and technological innovation effect migration to regions?**

Grantee: Advanced Composite Materials, LLC
Grant Title: SBIR Phase I: Novel Manufacturing Process for Cost-Effective Production of Discontinuous SiC Fibers from Renewable, Organic Precursors
Amount: $150,000
Grant Categories: Small Business Phase I, Advanced Materials Processing, Exp Prog To Stim Comp Res, Manufacturing
Keywords: coating, high performance, silicon carbide abbr fiber, resistance service lifetime, increase coating wear resistance, silicon carbide abbr fiber precursor, inexpensive organic precursor material, 100 billion global protective coatings market

Total SBIR Funding - $M

🟨	$0.1 - $2.9
🟩	$2.9 - $9.4
🟦	$9.4 - $19.6
🟦	$19.6 - $31.5

Manufacturing
Local Morans I=0.667 p<=0.01

🟥	High-High District Clusters
🟪	Low-Low District Clusters

Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Grant Year

2014 – 2018

Fill districts by:

● Total SBIR Funding
○ 2018 Industry & Migration Stats

Display clusters:

Manufacturing ▾

Use the map to examine SBIR and STTR grants made by NSF across the United States in the last strategy period, **2014 to 2018.**

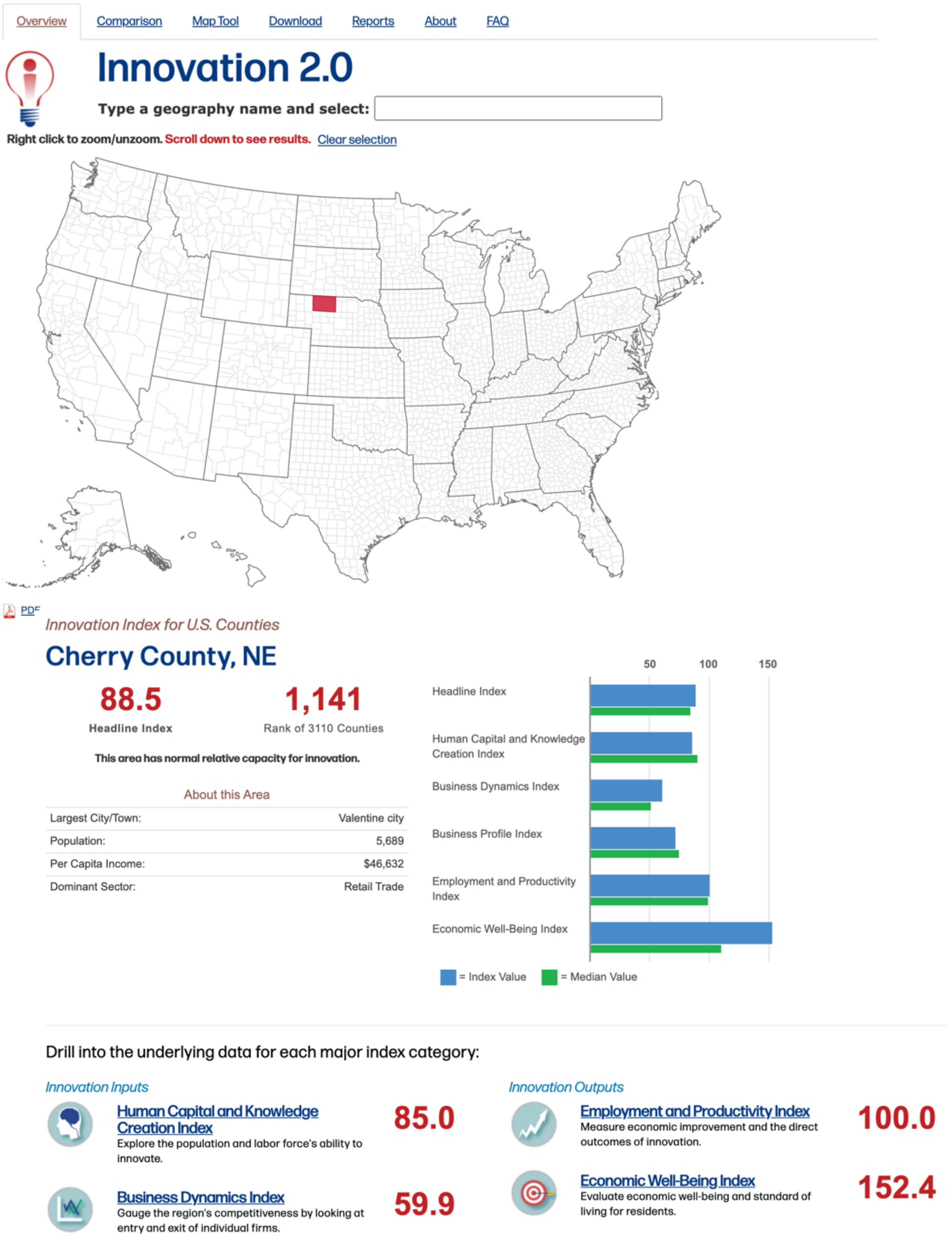
Click on **congressional districts** to see a summary of NSF's grant making activity there, as well as the 2018 ACS estimates of the top 5 industries in the district by proportion of workers, as well as migration trends in and out of the district.

Click on **clusters of awards** to zoom in to **awards** and click again to see particular grant details regarding the grantee and grant amount, the area of research, grant keywords, and (in some cases) National Library of Medicine MeSH subject headings detected in the grant abstract.

Data sources: NSF's awards database; 2018 ACS 1-year surveys of civilian employed population by industry, and of migration and geographic mobility.

Created by **Seth Schimmel** for Spatial Data Analysis course led by Prof. Jeremy Porter, Fall 2020 at the CUNY Graduate Center

Figure 3: Screenshot of StatsAmerica's Innovation Index 2.0 data portal. Users can click on a United States county to see innovation-related indices and data on the county.



Innovation 2.0

Type a geography name and select:

Right click to zoom/unzoom. Scroll down to see results. Clear selection

Innovation Index for U.S. Counties

Cherry County, NE

88.5 Headline Index

1,141 Rank of 3110 Counties

This area has normal relative capacity for innovation.

About this Area

Largest City/Town:	Valentine city
Population:	5,689
Per Capita Income:	$46,632
Dominant Sector:	Retail Trade

Drill into the underlying data for each major index category:

Innovation Inputs

Human Capital and Knowledge Creation Index — 85.0
Explore the population and labor force's ability to innovate.

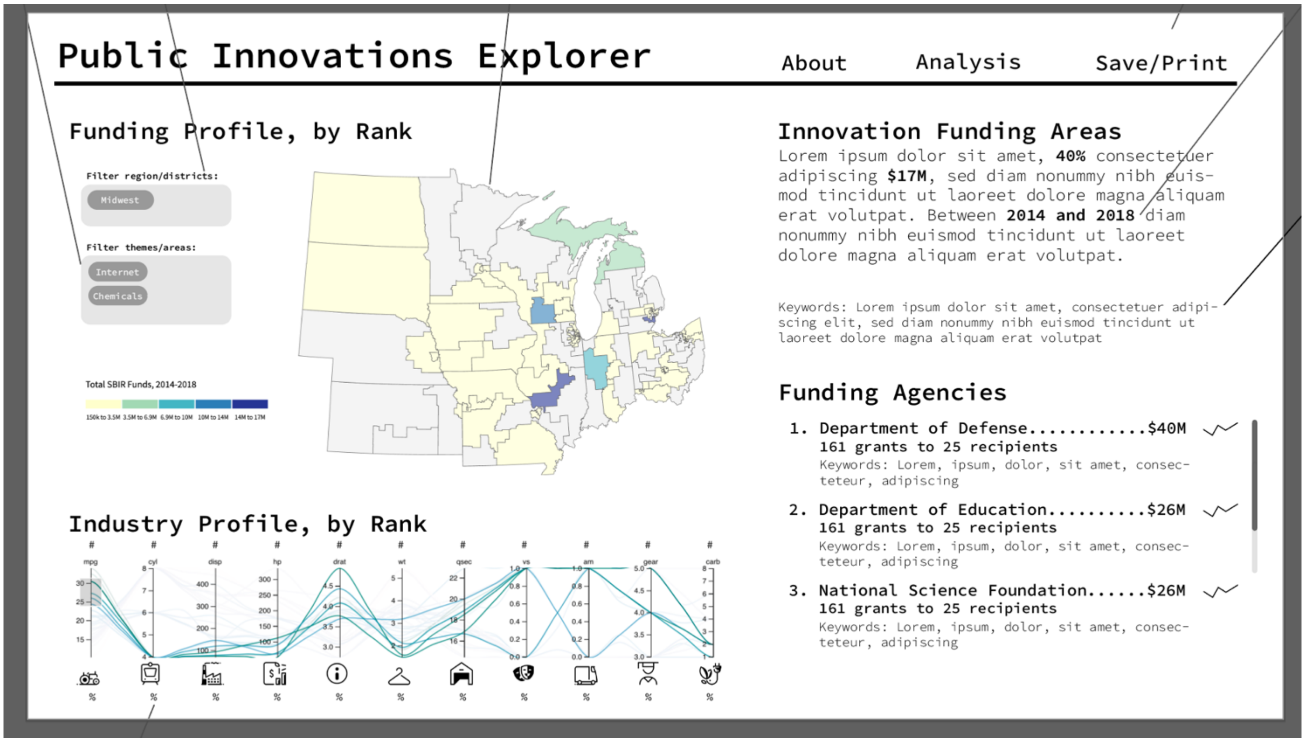
Business Dynamics Index — 59.9
Gauge the region's competitiveness by looking at entry and exit of individual firms.

Innovation Outputs

Employment and Productivity Index — 100.0
Measure economic improvement and the direct outcomes of innovation.

Economic Well-Being Index — 152.4
Evaluate economic well-being and standard of living for residents.

Figure 4: An early mockup of the layout for the Public Innovations Explorer I drafted in Adobe Illustrator. This mockup shows the ideas for sub-region filtering and thematic filtering.



Public Innovations Explorer

About Analysis Save/Print

Funding Profile, by Rank

Filter region/districts:

Midwest

Filter themes/areas:

Internet

Chemicals

Total SBIR Funds, 2014-2018

🟨	🟩	🟦	🟦	🟦
150k to 3.5M	3.5M to 6.9M	6.9M to 10M	10M to 14M	14M to 17M

Industry Profile, by Rank

Innovation Funding Areas

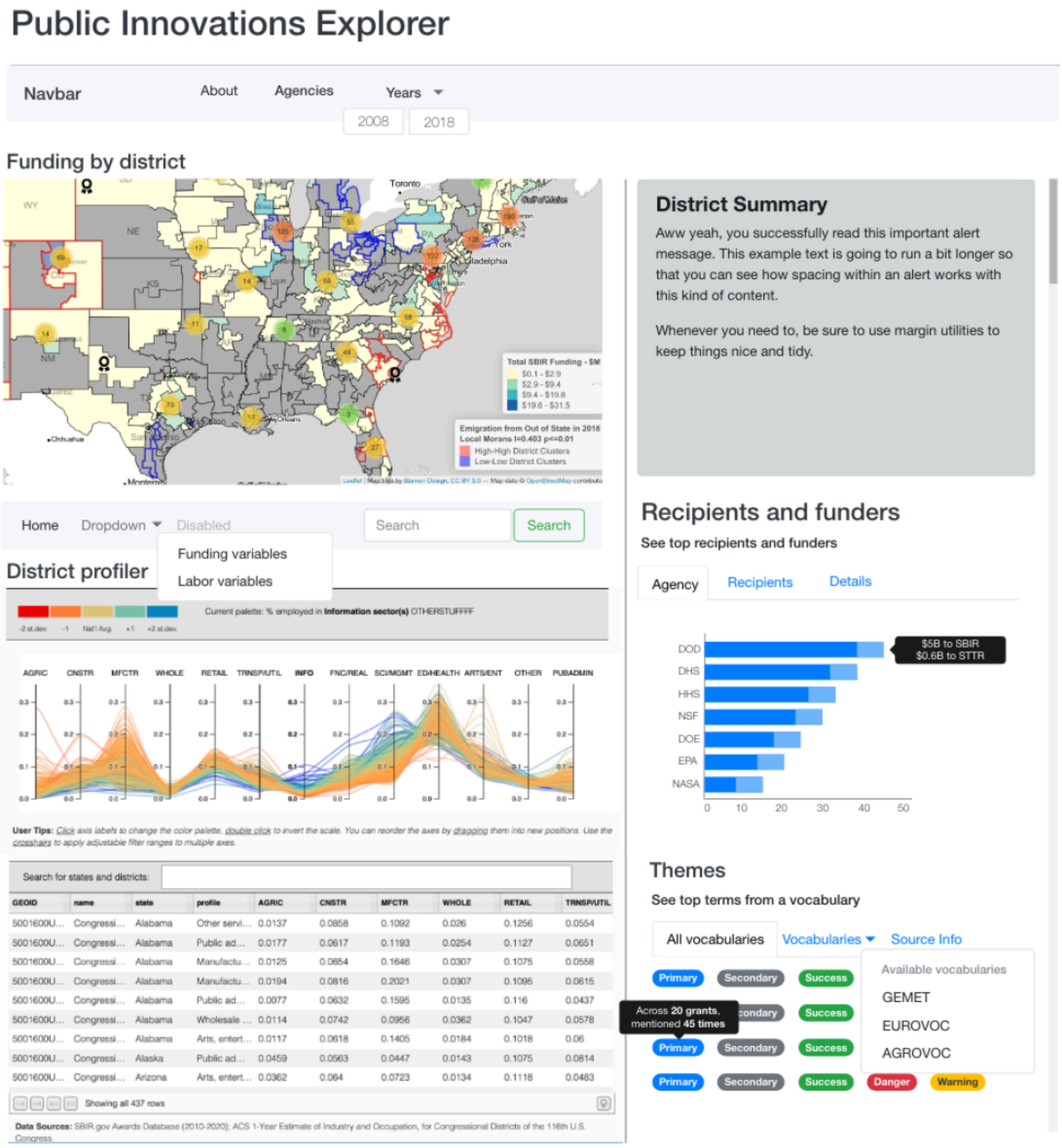
Lorem ipsum dolor sit amet, **40%** consectetuer adipiscing **$17M**, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Between **2014 and 2018** diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Keywords: Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat

Funding Agencies

1. Department of Defense............$40M
   161 grants to 25 recipients
   Keywords: Lorem, ipsum, dolor, sit amet, consecteteur, adipiscing
2. Department of Education..........$26M
   161 grants to 25 recipients
   Keywords: Lorem, ipsum, dolor, sit amet, consecteteur, adipiscing
3. National Science Foundation......$26M
   161 grants to 25 recipients
   Keywords: Lorem, ipsum, dolor, sit amet, consecteteur, adipiscing

Figure 5: A later mockup of the Public Innovations Explorer, utilizing screenshots of an already-tested parallel coordinates chart and the map from the Shiny dashboard.



Public Innovations Explorer

Funding by district

District Summary

Aww yeah, you successfully read this important alert message. This example text is going to run a bit longer so that you can see how spacing within an alert works with this kind of content.

Whenever you need to, be sure to use margin utilities to keep things nice and tidy.

District profiler

Current palette: % employed in **Information sector(s)** OTHERSTUFFFF

User Tips: *Click axis labels to change the color palette, double click to invert the scale. You can reorder the axes by dragging them into new positions. Use the crosshairs to apply adjustable filter ranges to multiple axes.*

Search for states and districts:

GEOID	name	state	profile	AGRIC	CNSTR	MFCTR	WHOLE	RETAIL	TRNSP/UTIL
5001600U...	Congressi...	Alabama	Other servi...	0.0137	0.0858	0.1092	0.026	0.1256	0.0554
5001600U...	Congressi...	Alabama	Public ad...	0.0177	0.0617	0.1193	0.0254	0.1127	0.0651
5001600U...	Congressi...	Alabama	Manufactu...	0.0125	0.0654	0.1646	0.0307	0.1075	0.0558
5001600U...	Congressi...	Alabama	Manufactu...	0.0194	0.0816	0.2021	0.0307	0.1095	0.0615
5001600U...	Congressi...	Alabama	Public ad...	0.0077	0.0632	0.1595	0.0135	0.116	0.0437
5001600U...	Congressi...	Alabama	Wholesale ...	0.0114	0.0742	0.0956	0.0362	0.1047	0.0578
5001600U...	Congressi...	Alabama	Arts, entert...	0.0117	0.0618	0.1405	0.0184	0.1018	0.06
5001600U...	Congressi...	Alaska	Public ad...	0.0459	0.0563	0.0447	0.0143	0.1075	0.0814
5001600U...	Congressi...	Arizona	Arts, entert...	0.0362	0.064	0.0723	0.0134	0.1118	0.0483

Showing all 437 rows

Data Sources: SBIR.gov Awards Database (2010-2020); ACS 1-Year Estimate of Industry and Occupation, for Congressional Districts of the 116th U.S. Congress.

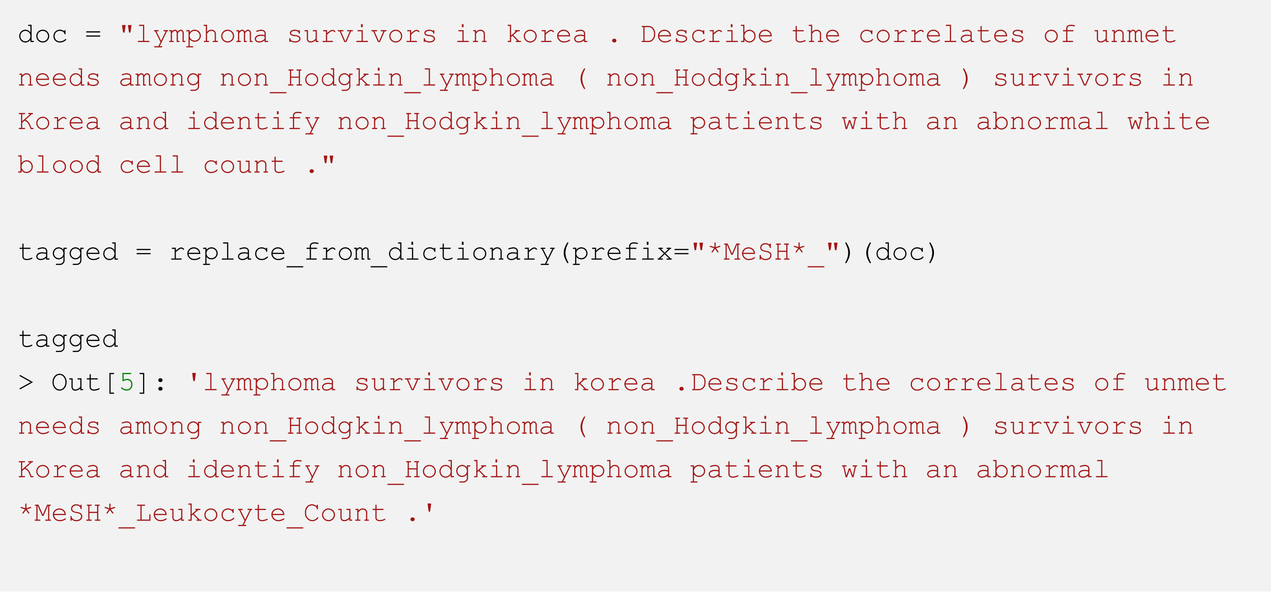
Recipients and funders

See top recipients and funders

Themes

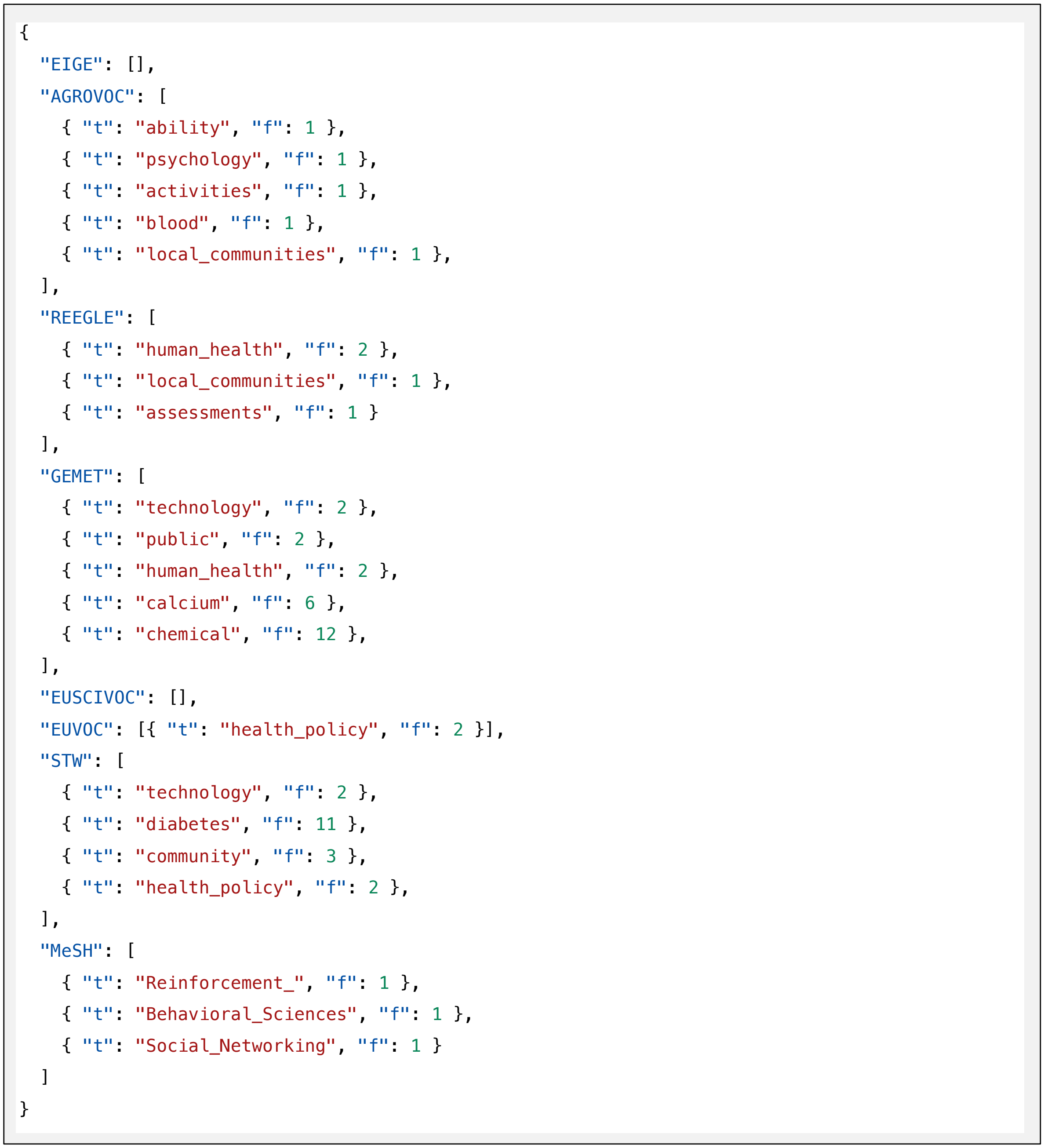
See top terms from a vocabulary

Figure 6A: Example showcasing the functionality of the NLPre replace_from_dictionary()



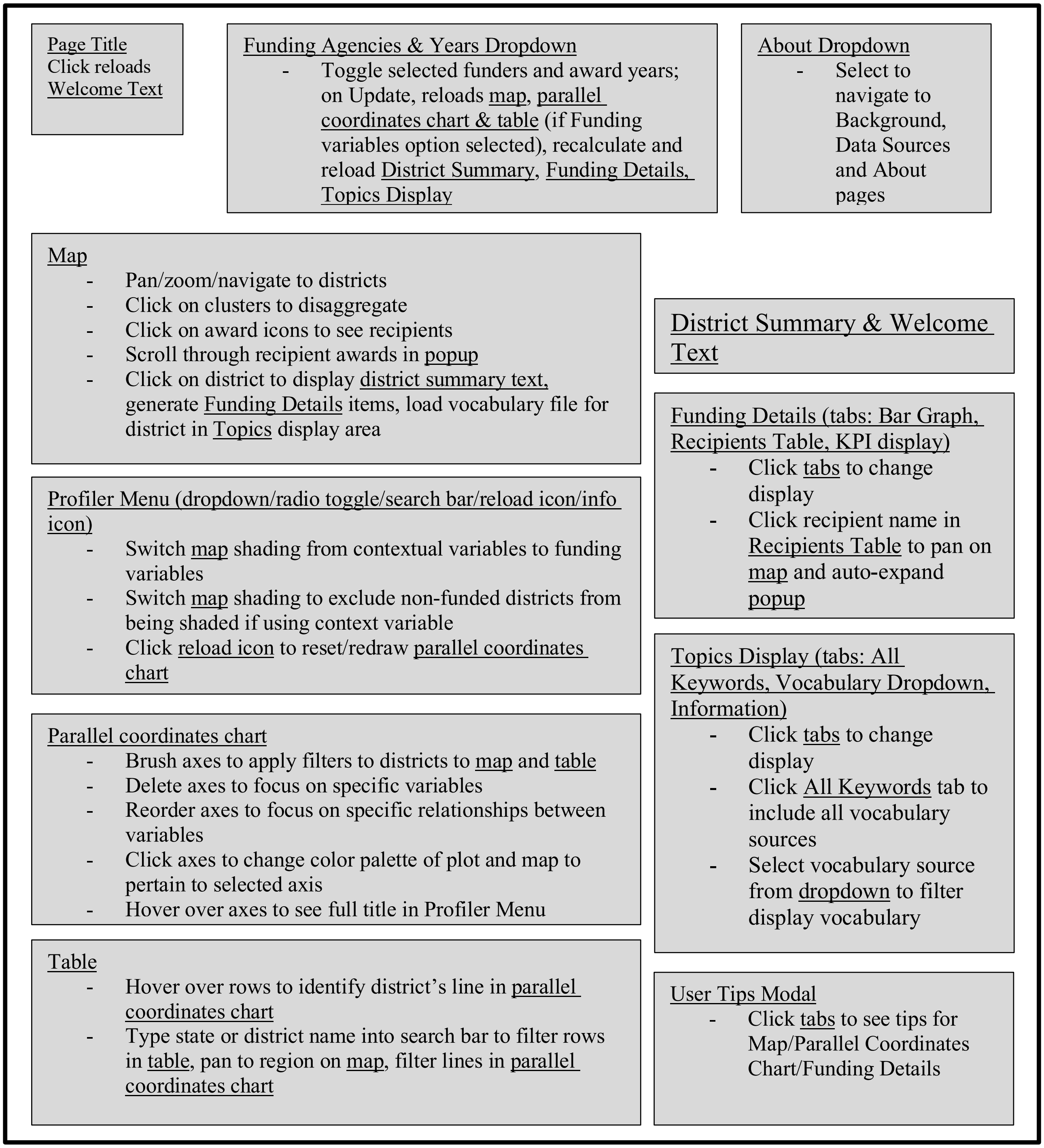
```
doc = "lymphoma survivors in korea . Describe the correlates of unmet
needs among non_Hodgkin_lymphoma ( non_Hodgkin_lymphoma ) survivors in
Korea and identify non_Hodgkin_lymphoma patients with an abnormal white
blood cell count ."

tagged = replace_from_dictionary(prefix="*MeSH*_")(doc)

tagged
> Out[5]: 'lymphoma survivors in korea .Describe the correlates of unmet
needs among non_Hodgkin_lymphoma ( non_Hodgkin_lymphoma ) survivors in
Korea and identify non_Hodgkin_lymphoma patients with an abnormal
*MeSH*_Leukocyte_Count .'
```

Figure 6B: Example showcasing a sample entry within the document-term index created as output of the function update_term_index function that I wrote to augment the NLPre pipeline. For each document, the index would contain a list of terms and their frequencies for each vocabulary source specified by the user.



```
{
  "EIGE": [],
  "AGROVOC": [
    { "t": "ability", "f": 1 },
    { "t": "psychology", "f": 1 },
    { "t": "activities", "f": 1 },
    { "t": "blood", "f": 1 },
    { "t": "local_communities", "f": 1 },
  ],
  "REEGLE": [
    { "t": "human_health", "f": 2 },
    { "t": "local_communities", "f": 1 },
    { "t": "assessments", "f": 1 }
  ],
  "GEMET": [
    { "t": "technology", "f": 2 },
    { "t": "public", "f": 2 },
    { "t": "human_health", "f": 2 },
    { "t": "calcium", "f": 6 },
    { "t": "chemical", "f": 12 },
  ],
  "EUSCIVOC": [],
  "EUVOC": [{ "t": "health_policy", "f": 2 }],
  "STW": [
    { "t": "technology", "f": 2 },
    { "t": "diabetes", "f": 11 },
    { "t": "community", "f": 3 },
    { "t": "health_policy", "f": 2 },
  ],
  "MeSH": [
    { "t": "Reinforcement_", "f": 1 },
    { "t": "Behavioral_Sciences", "f": 1 },
    { "t": "Social_Networking", "f": 1 }
  ]
}
```

Figure 7: User-interface diagram for deployed version of the Public Innovations Explorer.



Page Title
Click reloads
Welcome Text

Funding Agencies & Years Dropdown

- Toggle selected funders and award years; on Update, reloads map, parallel coordinates chart & table (if Funding variables option selected), recalculate and reload District Summary, Funding Details, Topics Display

About Dropdown

- Select to navigate to Background, Data Sources and About pages

Map

- Pan/zoom/navigate to districts
- Click on clusters to disaggregate
- Click on award icons to see recipients
- Scroll through recipient awards in popup
- Click on district to display district summary text, generate Funding Details items, load vocabulary file for district in Topics display area

District Summary & Welcome Text

Profiler Menu (dropdown/radio toggle/search bar/reload icon/info icon)

- Switch map shading from contextual variables to funding variables
- Switch map shading to exclude non-funded districts from being shaded if using context variable
- Click reload icon to reset/redraw parallel coordinates chart

Funding Details (tabs: Bar Graph, Recipients Table, KPI display)

- Click tabs to change display
- Click recipient name in Recipients Table to pan on map and auto-expand popup

Parallel coordinates chart

- Brush axes to apply filters to districts to map and table
- Delete axes to focus on specific variables
- Reorder axes to focus on specific relationships between variables
- Click axes to change color palette of plot and map to pertain to selected axis
- Hover over axes to see full title in Profiler Menu

Topics Display (tabs: All Keywords, Vocabulary Dropdown, Information)

- Click tabs to change display
- Click All Keywords tab to include all vocabulary sources
- Select vocabulary source from dropdown to filter display vocabulary

Table

- Hover over rows to identify district's line in parallel coordinates chart
- Type state or district name into search bar to filter rows in table, pan to region on map, filter lines in parallel coordinates chart

User Tips Modal

- Click tabs to see tips for Map/Parallel Coordinates Chart/Funding Details

APPENDIX D: TABLES

Table 1A: Award making details by agency and program, 2008 to 2018, SBIR Program.

Small Business Innovation Research (SBIR) Program												
	All Agencies	**NSF**	**DHS**	**DOD**	**ED**	**DOC**	**USDA**	**NASA**	**DOT**	**DHS**	**DOE**	**EPA**
Awards	**57,182**	**4,121**	**12,369**	**27,425**	**335**	**516**	**1,166**	**5,126**	**307**	**697**	**4,682**	**437**
Total Amount ($M)	**$24,983**	**$1,163**	**$7,891**	**$11,651**	**$102**	**$96**	**$237**	**$1,478**	**$93**	**$282**	**$1,933**	**$56**
% with Topic Codes	93.9	100.0	79.9	99.9	38.5	77.1	81.7	98.1	96.7	88.2	96.5	81.5
% with Research Keywords	43.5	0.0	0.4	69.5	60.0	0.4	38.7	88.6	0.3	60.5	0.4	28.1
% to Women Owned Businesses	12.9	14.5	12.4	14.6	18.5	9.1	9.9	10.7	22.5	10.9	6.4	7.6
% to Businesses Owned by Socioeconomically Disadvantaged Groups	59.8	9.8	4.1	6.1	0.9	5.2	2.7	8.3	19.9	7.3	4.3	5.5
% to Businesses in Hubzones	2.4	7.9	0.1	1.8	3.0	5.0	7.8	1.7	3.3	3.7	6.3	2.3

Table 1B: Award making details by agency and program, 2008 to 2018, STTR Program.

Small Business Technology Transfer (STTR) Program												
	All Agencies	**NSF**	**DHS**	**DOD**	**ED**	**DOC**	**USDA**	**NASA**	**DOT**	**DHS**	**DOE**	**EPA**
Awards	**8,567**	**700**	**2,090**	**4,435**	0	**0**	**0**	**678**	**0**	**1**	**663**	**0**
Total Amount ($M)	**$3,226**	**$189**	**$1,034**	**$1,543**				**$205**		**$1**	**$254**	
% with Topic Codes	95.7	96.0	84.3	99.9				98.7		100.0	99.8	
% with Research Keywords	43.8	0.0	0.0	71.5				85.1		0.0	0.8	
% to Women Owned Businesses	12.3	15.3	12.3	12.6				11.4		0.0	8.9	
% to Businesses Owned by Socioeconomically Disadvantaged Groups	7.2	8.7	2.1	8.8				11.1		0.0	7.4	
% to Businesses in Hubzones	2.5	8.9	0.1	2.1				1.8		0.0	6.9	

Table 2: List of controlled vocabulary, taxonomy or linked-data ontologies reviewed and considered for use in keyword-extraction pipeline; a handful of vocabularies were ultimately chosen based on my interest in comparing coverage between sources and based on the ease with which I could prepare the vocabulary data for use in the pipeline.

Source	**Brief Description**	**Used?**
EUROVOC Thesaurus of Activities related to the EU	Governmental, social, political, legal and economic classifications from the European Commission	**Y**
STW Thesaurus of Economics(http://zbw.eu/stw)	Areas of economics	**Y**
Food and Agriculture Organization of the United Nations - AGROVOC Thesaurus	Food systems and agricultural classifications from the UN	**Y**
EuroSciVoc - European Science Vocabulary	Science related classifications from the European Commission	**Y**
European Institute for Gender Equality (EIGE) Glossary & Thesaurus	Gender equality thesaurus from the European Commission	**Y**
European Environment Agency General Multilingual Environmental Thesaurus (GEMET)	Environmental issue classifications from the European Commission	**Y**
REEGLE Clean Energy Linked Data	Clean energy and environmental area thesaurus from REEP/REEGLE	**Y**
GESIS Thesaurus of Social Sciences	Areas of social sciences	N
American Economic Association JEL classifications	Areas of economics	N
European Commission Skills, Competencies, qualifications and Occupations	Skill, labor sector and occupational classifications from the European Commission	N
EU Statistical classification of products by activity, 2.1 (CPA 2.1)	Statistical classifications of products, from the European Commission	N
UN International classifications on economic statistics -- Central product classification (CPC) & International Standard Industrial Classification of All Economic Activities	Classifications related to economics, industrial areas and products, from the UN	N
UNBIS Thesaurus	Thesaurus of issues related to the work of the UN	N
Geological Survey of Austria Geological Thesaurus (Minerals, Mineral Resources, Lithology)	Geological classifications from the Geological Survey of Austria	N
U.S. DEPARTMENT OF AGRICULTURE Agricultural Thesaurus and Glossary	Food systems and agricultural classifications from the United States Dept. of Agriculture	N
North American Industry Classification System	Industry area and economic classifications from the United States Census	N

BIBLIOGRAPHY

Datasets, scripts and 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

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

Real Impact Analytics. (2015). *Malaria Data Exploration: Kenya counties in 2006 – 2013.* Retrieved from: https://realimpactanalytics.github.io/d4g-hackathon-malaria-viz/

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

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

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

Yun, X., Timelyportfolio, Jacobson J., Chopson, D. (2019). Parcoords-es: ES6 module of Syntagmatic's Parallel Coordiantes. Github Repository, https://github.com/BigFatDog/parcoords-es

Research and Scholarly Works

Anselin, L., Varga, A., & Acs, Z. (2000). Geographical Spillovers and University Research: A Spatial Econometric Perspective. *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

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

Breschi, S. (2001). Knowledge Spillovers and Local Innovation Systems: A Critical Survey. *Industrial and Corporate Change, 10*(4), 975–1005. https://doi.org/10.1093/icc/10.4.975

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

Gabrielatos, C. (2018). Keyness Analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review*. Routledge.

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

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

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*. 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). 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

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). 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, 122,* 215-247. 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*. 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

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

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