A Data Science Approach to Studying the Small Business Innovation Research (SBIR)/Small Business Technology Transfer (STTR) Program: A New Tool for Policy Researchers

Grant A. Allard
PhD Student
Clemson University

APPAM Fall 2018 Research Conference
Washington, DC
Abstract

The goal of this conference paper is to share preliminary findings addressing the research question, “What is the effect of agency budget changes on the implementation of the Small Business Innovation Research (SBIR) / Small Business Technology Transfer (STTR) program?” This question is significant because scholars have argued that one of the chief limitations of the SBIR/STTR program’s effectiveness is budgetary limitations. I investigate how budget changes affect the implementation of the program in terms of the number of awards, number of proposals, award size, and geographic distribution of the awards. One of the challenges in studying the SBIR/STTR program is data availability. The raw data available through sbir.gov is fragmented, unstructured, and requires extensive cleaning and structuring to be useful. Through this study, I make two contributions to the field of science and technology policy, and specifically the study of the SBIR/STTR program. First, I introduce a new data set that combines budget and agency data in order to promote the study of this program. Second, I provide an exploratory analysis of SBIR/STTR program budget variance and its effects on program implementation. My results do not provide statistically significant insights and further work is needed to continue to explore this relationship.
Introduction

The goal of this conference paper is to share preliminary findings addressing the research question, “What is the effect of agency budget changes on the implementation of the Small Business Innovation Research (SBIR) / Small Business Technology Transfer (STTR) program?” The SBIR/STTR program provides federally-funded grants for small businesses engaged in scientific research and innovation. This question is significant because scholars have argued that one of the chief limitations of the SBIR/STTR program’s effectiveness is budgetary limitations. I investigate how budget changes affect the implementation of the program in terms of the number of awards, number of proposals, award size, and geographic distribution of the awards. One of the challenges in studying the SBIR/STTR program is data availability. The raw data available through sbir.gov is fragmented, unstructured, and requires extensive cleaning and structuring to be useful. Being a new scholar in this field and knowing that several other studies have delved into the SBIR/STTR program (Lanahan, 2017; Tingle, 2016; National Academies, 2015; Link and Ruhm, 2009, among others), I turned to trying to find data through scholarly data repositories such as Harvard’s Dataverse and the University of Michigan’s Inter-University Consortium for Political and Social Research (ICPSR). Unfortunately, few scholars have shared access to their data sets to make their research replicable or reproducible through these means. A major part of my project became the production of a data set that integrates data at the level of individual awards, with agency-level budget data, and accurate geographic data. Through this study, I make two contributions to the field of science and technology policy, and specifically the study of the SBIR/STTR program. First, I introduce a new data set that combines budget and agency data in order to promote the study of this program. Second, I provide an exploratory analysis of SBIR/STTR program budget variance and its effects on program implementation.
In providing a new publicly-accessible data set to the community, I devote a lot of attention in this conference paper to the data collection, wrangling, and analysis methods. My goal is to practice the principles of open science, which make data collection and results reproducible and replicable. Part of replicability is about making data reusable for other scholars. Tenopir and colleagues (2015) argue that data reuse allows scholars to become more productive. Piwowar and Vision (2014) find that data reuse promotes usage and writing about specific problems in the discipline of genetics. As a policy scientist, my goal is to provide accurate and precise methodological accounts, in order to meet the high scientific standards in our community and to promote the study of the SBIR/STTR program.

The SBIR/STTR program is important to study because it is cited as one of the most effective R&D subsidy programs in the world. Tingle (2016) argues that one of the factors that limits the program from being even more successful is that agencies do not allocate more funding to support the program.

**Background**

The SBIR/STTR program is administered by the Small Business Administration (SBA) to “support scientific excellence and technological innovation through the investment of federal research funds in critical American priorities to build a strong national economy” (SBA, 2018, p.6). There are two different parts of the program, although both award grants through a competitive application process to small businesses to conduct research and development (R&D) activities. One part is the Small Business Innovation Research (SBIR) program, which focuses on the development of small businesses to conduct applied research and development, with the goal of creating a new technology that can then be commercialized. As of the time of my writing, eleven executive departments and independent executive branch agencies participate in this
program. The SBIR program is a three-phase program that allows a business to establish the “technical merit, feasibility, and commercial potential” of a new technology (SBA, 2018, n.p.). The other part is the Small Business Technology Transfer Research (STTR) program, which focuses on the development and transfer of new technologies for commercial purposes. Five executive departments and independent executive branch agencies participate in the STTR program. One of the most unique features of this program is that agencies are mandated by congress to set aside a percentage of their extramural research budgets for the SBIR/STTR program.

Existing research on the SBIR/STTR program has primarily evaluated the effectiveness of the SBIR/STTR program at producing its stated policy outcomes, rather than examining the role of politics in its implementation. These studies have focused on the effectiveness of the SBIR/STTR program from the perspective of firms. Specifically, these studies have answered questions about the SBIR/STTR program’s effect on firms’ revenues, human capital, and longevity (Tingle, 2016; Link and Scott, 2010; Link and Ruhm, 2009; Cooper, 2003; Audretsch et al., 2002). Studies have also looked for indicators of and determinants for identifying why some recipient firms perform better than others (Siegel and Wessner, 2012). The SBIR/STTR literature examined why the SBIR/STTR program increases firm performance (Howell, 2017). One of the main conclusions that emerges from analyzing the SBIR/STTR program from the perspective of firms is that it is an effective program for promoting entrepreneurship.

A second vein of existing scholarship has been to investigate the SBIR/STTR program from the perspective of policy implementation. Lanahan (2015) and Lanahan (2016) identified tensions between state and federal agencies, caused by volatility in the federal budget, that have pushed some states to develop their own SBIR/STTR supplemental programs. Feldman and
Lanahan (2015) found that federal agency budget volatility had a positive, statistically-significant effect for states to develop their own programs as a substitute for federal funding. The Lanahan-Feldman approach to studying the SBIR/STTR program showed that budgetary behavior matters and may be a driver of behavior.

Another approach to studying implementation is to understand the role of the bureaucracy in the SBIR/STTR program. Joshi, Inouye, and Robinson (2017) investigated how the demographic diversity of granting agencies impacts the demographic diversity of SBIR/STTR award recipient firms. They find that diversity in federal agencies is significantly correlated with increased diversity in Phase II SBIR/STTR awards, especially those given to women. This study highlighted that differences between agencies in their demography and behavior led to differences in how the SBIR/STTR program was implemented.

Research Design and Methodology

I took a data science approach to answering my research question, meaning that my focus is as much on the process of how I did what I did as it is on my results. Data science uses data, statistics and inference (Dhar, 2013) and attempts to both explore and exploit patterns in the data (Kelleher and Tierney, 2018). It is a natural approach for a policy scientist. First, as my research focus changed from data analysis to an exploratory study, I had to integrate workflows between data collection and data analysis. The data science approach focuses on creating replicable insights, including allowing other scholars to replicate the data collection phase of a study. Second, the data science approach allows me to conduct data exploration. Data exploration is still a rigorous and hypothesis-driven activity because it seeks to answer my overall research question about the effect of budget changes on SBIR/STTR program implementation, and it exploits the presence of other variables and relationships to identify new questions of interest.
that may be worth answering. My workflow is to collect my data, visualize it, conduct an initial test of relationships, design a more thorough statistical analysis, and then report the relationships. In this paper, I present findings from my data collection, visualization, and initial tests. I provide my methodology in exhaustive detail, in the interest of making my research replicable.

There are several tools that I use throughout the next three steps that are worth noting. First, I used the statistical modeling software $R$ for my data management and analysis. $R$ is useful because it provides rigorous audit trails of my work and is flexible enough to handle many different types of data wrangling and analysis tasks. I used a tidy approach to data management throughout my project, which was made possible through several packages such as “tidyrr” (Wickham, 2018), “leaflet” (Cheng et al., 2018), and “jsonlite” (Ooms, Lang, and Hilaiel, 2017).

Second, I used the programming language Python for web scraping and managing a mySQL database. I will be citing the usage of particular tools throughout the paper to promote the replicability of my work and to acknowledge the creators of these tools. $R$ and Python are maintained by open source communities, and it is standard practice to cite the creators of particular tools when they are referenced.

**Data Collection**

The phase of data collection posed significant challenges because there were several data sources and they yielded mostly messy data. One of the main contributions of this project to the field of science and technology policy is the creation of a new data set that links and standardizes several existing and unlinked data sets. Linking data sets matters because it is a time-consuming step for answering questions that may require data from two or more data sets. An example of this problem is that we need data at both an agency level-of-analysis and an individual award
level-of-analysis if we are to understand the geography of SBIR/STTR awards, given budget changes. Budget changes are in one data set and the geographic coordinates are in another. While they are related, they are not linked. While scholars have written and published extensively about this program, there are few data sets accessible through scholarly data repositories, including for peer-reviewed articles in which both replicability and reproducibility are valued. The lack of available and clean data meant that I had to collect my own, using the SBIR/STTR program website.

My first step was to identify all of the data that I would need to conduct the desired analyses. First, I needed to analyze the impact of compliance or non-compliance with the congressional set-aside requirement, on the implementation of the program in terms of the number of awards and average award size. Second, I needed to understand the impact of congressional set-aside requirements—and politics more broadly—on the geographic distribution of awards. Third, I needed to understand variations in the implementation of the program between agencies, specifically executive departments such as the Department of Agriculture and independent executive branch agencies such as the National Aeronautics and Space Administration. Table 1 presents a summary of the data that I needed to collect. I needed to collect a lot of data.

My next step was to create a data collection and storage strategy. This step is important and forgotten in most research because it is easy to overlook how the data’s structure and storage can bias analyses. In my data collection strategy, I identified what data I could collect from each source. My sources included laws passed by Congress, existing scholarship, databases accessible through the sbir.gov website, and the sbir.gov website itself. I am thankful that the Small Business Administration has invested resources in making these data accessible. There are two
main sources that could be used for collecting the information I needed. First, \textit{sbir.gov} offers an application programming interface (API) that allows for users to access large amounts of data with relative efficiency. One area in which the API could be improved is to provide a better guide to using its query language, which is not intuitive. Second, \textit{sbir.gov} provides the Annual Reports Dashboard from which users can download data in a tabular form. There are several challenges for users who interact with the Graphical User Interface (GUI) of the dashboard. First, there is no variable within the downloaded data specifying the agency with which the data is associated, which requires users to be very disciplined about their download and naming conventions to avoid errors. There is no audit trail if this step is done manually. Second, when a user selects several agencies, their data sums rather than appends. It is impossible to download distinct data for multiple agencies except by downloading data one agency at a time.

Because one of my goals was to contribute a cleaned, ready-to-analyze data set, I needed to develop a data storage plan. Storing large amounts of data carries with it unique considerations. For example, Microsoft Excel will only allow approximately 65,000 URLs to be contained in each worksheet of a workbook. My data storage plan had three phases. During my data collection, I stored data locally on my personal computer in a comma-separated values (CSV) worksheet and I backed this data up to an external hard drive regularly. This step allowed me to flatten the data into rows and columns that made it very easy to read into statistical programming software such as R and Python. In my exploratory analysis phase, I stored data on my local computer and regularly backed it up to a Git Hub repository. GitHub is a useful tool for data science because it gives users the ability to store data and code while providing a detailed audit trail of what changes were made since the last update (i.e., “push commit” is the technical programming term). In the production phase, I stored the data, including the finished data set
and metadata containing the procedures I used to clean it, in a Harvard Dataverse repository. I placed my data and code in Dataverse because it makes my work available to other scholars, while also protecting the work as my intellectual property, which means that users must attribute use of the data in the same way that they would attribute a journal article, conference paper, or correspondence.

My next step was to collect my data. Figure 1 presents a table of all the information I collected, the data sets I wrangled, and the final outputs. First, I created a data set that contains information regarding the congressional set-aside requirement of each agency for each year. I sourced the data in this data set by using Tingle (2016) and [enter laws]. I created this data set using Microsoft Excel because it is a useful tool for data entry, and then read it into R. Then I created a data frame, which I sued for joining this data into other data sets.

![Figure 1. Data Workflow from Collection to Output](image-url)
Second, I collected data that gives insights about each agency’s SBIR/STTR program budget performance. Because of the limitations of using the GUI, I developed a slightly more effective way to collect this data. I noticed in the URL that the dashboard was requesting data from a database using a SQL query. I wrote a script that downloaded data for each agency and read it into R. By using R for this step, I was able to create a clear audit trail for which data belonged to which agency. For each individual agency data set, I added a column with its three-letter abbreviation, which is the same abbreviation I use for each agency throughout the rest of my work. Next, I bound the rows together to create a long-form data set. I pulled this data set from the SBIR Program’s Annual Reports Dashboard. I have data for a total of eleven agencies starting in 1990 and running through 2015, with the exception of 1993 and 1999. Each observation consists of data for a single agency in a single year. I call this unit of analysis an agency-year. For each agency-year, I have information for its extramural research budget in thousands of dollars, the amount of money obligated to the SBIR/STTR program in thousands of dollars, the overall budget for the SBIR/STTR program in thousands of dollars, the deficit in thousands of dollars between the amount that is obligated and the program budget, the percent of the extramural research budget that the program budget represents, the number of proposals, and the number of awards. This data set consists of 252 agency-year observations. However, the data is not complete. There are three problems that I had to address to make this data useful.

The first problem that I faced was that all agency-year observations for the years 1993 and 1999 are missing. Neither the SBIR Dashboard, nor supporting documentation such as the Annual Reports themselves, explain why this is the case. It does not seem coincidental that these are the only years within the range for which there are no SBIR Annual Reports. The best alternative for completing these cases is to find the missing data and insert it. I reviewed the
Annual Reports for several years, including 1994 and 2000, which are the first reports after the missing observations, and did not find data for the variables in which I am interested. Therefore, simply filling in the missing values was not an option. There were several alternatives that I could choose to pursue.

First, I could continue to do the analysis with only the data that I had. The strength of this approach is that it limits the risk of introducing bias, although I would not know how my data was biased by the missing observations. The weakness of this approach is that it would limit the total number of year-over-year changes that I could analyze because each of the missing years also eliminates the year immediately following it from my data set. I needed the prior year’s agency-year observations to calculate the year-over-year change. In effect, this strategy eliminates approximately 13.3% (i.e., 36/270 observations) of my data from the analysis. While the standard for acceptable data loss is largely arbitrary, 5% or less is considered the standard number in most data science projects (Tierney, 2017). Alternatively, I could impute the missing values for the variables extramural budget and number of proposals. There are several methods that I could use for this process (Moritz and Bartz-Beielstein, 2017). Most processes such as Multiple Imputation (Rubin, 1987), Expectation-Maximization (Dempster et al., 1977), Nearest Neighbor (Vacek and Ashikaga, 1980) use correlations between variables to estimate values over time. While these methods are plausible alternatives, I opted for a simpler method: univariate time-series imputation. In a univariate time-series imputation, I treated the variable with the missing value as a function of time. I had the expectation based on the budgetary politics literature that budgets move incrementally over time (J.E. Anderson, 2015; Lee et al., 2013; S. Anderson and Harbridge, 2010). It is reasonable to assume that I could impute the missing values as a function of time. Another assumption that I made was that the time between observations
was of equal duration. Then I used the “imputeTS” package in R to impute the missing observations for each agency. I chose the Kalman Smoothing method for two reasons: first, Mortiz and Bartz-Beielstein (2017) show it to be one of the most effective algorithms for time-series data; and second, because it gave me results that—on visual inspection of the time series—fit the general trend of each variable, while other algorithms gave me results that were clearly unlikely. In future work, I hope to improve the accuracy of these imputations. Once the imputations were complete, I calculated the remaining variables—deficit in dollars and extramural percentage-- using arithmetic.

The second problem was several missing observations for both the deficit and program budget variables. The good news was that I could calculate the values for the missing data from data that was already in the data set. For the cases that are missing values for the deficit variable but no other variables, I could compute the budget deficit by subtracting the obligated variable from the program budget variable. For the remaining observations with missing values across two variables, I could multiply these values for the observations that are missing values for the program budget variable. Then I subtracted the obligated variable from the program budget variable to calculate the deficit variable for these observations. Another decision I made was to remove the Department of Defense data from this data set since the Department of Defense is not present in my other datasets.

My third data set gives me information about each award that has been made since 1983. I pulled this data using the API provided by the Small Business Administration on the SBIR website. In this data set, each observation consists of an individual award. Each award can be uniquely identified through its URL. There are no duplicates of URLs in this dataset. The awards data that is available from the API gave me information about each individual award such as its
As of October 21, 2018, there are 89,000 observations in this data set. In reviewing the data, awards granted by the Department of Defense are not available through the API. Therefore, this dataset consists of data from the Department of Homeland Security (DHS), Department of Commerce (DOC), Department of Energy (DOE), Department of Transportation (DOT), Department of Education (ED), the Environmental Protection Agency (EPA), Department of Health and Human Services (HHS), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and Department of Agriculture (USDA). Another notable exception to this data is that it does not contain information that is important to my analysis such as the firm’s location, the amount of the award, or the DUNS number. The SBA provides a firm-level API for pulling this type of data. Unfortunately, the data that was pulled from the firm-level API was unusable, because there are no unique identifiers by which to tie each firm-level observation to an award. An example is that if you call a firm such as Radiation Monitoring Devices, Inc., which appears across many agency-year observations, then you get back four different entries, some of which could contain slightly different information about location and ownership status. There is no way to know which entry belongs to which award. This led me to construct my own data in my third dataset.

Third, I constructed a data set by programming a web scraper to collect information that is not available through the API from the website for each individual award. A web scraper is an automated computer program that interacts with a website to identify and capture information. I programmed my web scraper using Python. I chose Python because of its flexibility, and used
the approach pioneered by Rao (2017). I used the list of URL links in the awards database as my frame for the scraper. Each website that I was scraping had the same format, which made coding my scraper and cleaning the data much more manageable. The scraper’s workflow was to find a website on the list, find and capture specific information, convert it from HTML to readable text, create an observation within a dataframe, and then send the observation to a MySQL server. The variables that I had the web scraper collect were the amount of the ward, the firm’s DUNS number, firm’s address, contact information, and information about the firm’s status as a HUBZone firm, woman-owned firm, or socially and economically disadvantaged-owned firm.

In deploying my web scraper, one challenge was that the SBIR website requires all web scrapers to put at least 30 seconds of rest in between scrapes. With over 88,500 results to scrape, this would take one computer roughly 737 hours and 30 minutes (30.7 days), assuming that it works continuously with no interruptions. An example of the type of interruption that could make this process take longer is a power outage that knocks out the Wi-Fi network on which the computer relies. Originally, I had deployed my web scraper on a local computer that I could run continuously for roughly 6.9 days per week. I found this to be unsustainable because of interruptions caused by electricity, Wi-Fi, the computer’s own operating system, or my own clumsiness. I found that a more effective way of completing the task was to deploy the web scraper to a computing cloud. I segmented the total list of links that I intended to scrape. Then I used Jupyter Notebooks to deploy my Python code for each web scraper to Amazon’s Elastic Computing Cloud. I did not have previous experience in using Amazon’s cloud computing resources and this step took considerably longer than expected. I used Amazon Elastic Computing (EC) virtual machines that provide very little memory to keep my costs manageable.

---

1 A good practice in web scraping is to check the /robots.txt file and adjust your scraper accordingly or program your scraper to comply with it always.
(i.e., greater memory = more cost) and I stored my data in a MySQL Database using Amazon’s Relational Database Services (RDS). I kept logs for all my scrapers and manually tracked their performance to aid with the reproducibility of future research.

Once the data was collected, my biggest challenge was to convert unstructured web data (albeit in text readable form) into data that makes sense to a human being. An example of unstructured data is that the addresses were in slightly different formats because some were in a format such as Company Name, Building Number, Street, City, State, Zip Code while others were in Street, Company Name, City, State, Zip Code format. Then, I faced the challenge of making a lot of raw data, that was primarily in the form of text, into usefully-structured data that could be used in quantitative and spatial analyses. To restructure this data, I used the “stringr” package (Wickham, 2018) to employ natural language processing to identify patterns in text such as zip codes, addresses, and amounts of money. After cleaning this data, I was able to create new variables for the zip code, award amount, woman-owned business, and socially and economically disadvantaged- business.

I had one more data collection phase, which was to add geospatial information into my dataset. Using geospatial data can be a bit tricky because there are different types of variables depending on whether I wanted to map the SBIR/STTR awards as points, areas, or lines. For example, a zip code does not represent a line of addresses on a map (rather than an area), which makes it a relatively inaccurate frame for conducting spatial statistics. My web crawling data gave me zip codes, so I decided to use them as a key for linking them with my SBIR/STTR data. I used the R “zipcode” package (Breen, 2012) to add geospatial data such as city, state, and the latitude and longitude of each zip code, into the data set.
At this point, I had a lot of data but it was divided into separate containers, and needed to be consolidated into useful data sets. I merged the Awards Data with the Web Scraper Data by using the unique URL link as a key. I also mapped variables that were of interest at an agency-year level, but were currently presented at an individual-award level, such as information about whether the firm is HUBzone-owned, woman-owned, or socially and economically disadvantaged-owned, from the award data set to my budget data set.

**Data Description and Exploration**

My final data set consists of data regarding agency budgets, awards, and geography. I provide summary statistics and a description of each aspect of the data set.

For the budgets data, I have collected 246 observations of 21 variables from ten agencies, for the years 1990-2015. There were originally 51 observations with missing values, 18 of which were missing observations in their entirety. Through imputation and calculation, I was able to address issue of missing data. I then calculated several new variables to indicate year-over-year changes in the program budget as a percent of the extramural research budget, the congressional set-aside percent, the number of awards, and number of proposals. This data set needs to be curated once per year because it will change annually upon the publication of a new report.

For the awards data, I have collected 85,530 observations of 27 variables. One of the challenges of working with this data set is that it is constantly changing, as new awards are made and added to the population. Since I started this phase of the project in early October, over 1,000 new awards have been added to the dataset, so it has grown from 88,674 to 89,286 potential awards. Based on my original pull of data, the finished data set consists of complete data for 96.9% of the awards that were pulled.
I will be posting these data sets on Harvard’s Dataverse before APPAM’s 2018 Fall Research Conference along with the necessary code for collecting the data and analyzing the data sets.

**Exploratory Analytical Models**

I conducted two exploratory analyses of my data. First, I analyzed the agencies and their behaviors over the course of time, because I want to understand if there are discernible trends in budgetary changes and changes in program implementation. The goal of an exploratory data analysis is to identify preliminary trends and discuss whether they worth investigating further. The dependent variables that I investigated are the number of awards and the average award amount. I chose these variables because they are policy outputs. I will control and condition my analyses for compliance with the congressional set-aside requirement, firm ownership, and political partisanship.

My first step was to describe my data, especially several variables that will play a prominent role in my analysis. I am exploring if a positive relationship does exist between extramural research budgets and the number of awards. This relationship is the one that is theorized by scholars who argued that the greatest limitation of the SBIR/STTR program is limited funding. For my initial regression, I used a simple OLS linear regression to establish if a relationship exists between the percent of extramural research budgets (i.e., its program budget) dedicated to the SBIR/STTR program and the number of awards and average award size. I used the percent of extramural research budget as my independent variable rather than the program budget as a dollar amount, because it controls for differences in the scales of agency extramural research budgets. I did not find a statistically significant relationship between the agency’s SBIR/STTR program budget and the number of awards made (See Figure #). This is a
moderately surprising result because at face value one would expect that changes in program budgets would also change the number of awards. This finding brings up several possibilities that must be further investigated. First, the absence of this relationship could be caused by noise in the data that is accounted for by fixed variations between time, agency, program, or phase. I can control for these factors in a more detailed model. Second, this could be caused by agency behavior. Looking at the scatterplots in Figures 1 and 2, it is clear that most of our results cluster around 2.43%, which is the mean of the program budget expressed as a percentage of extramural research budget. It may not be a coincidence that the mean congressional set-aside percentage for the data set is 2.40% (marked in red on the panels below). Statistically speaking, there is no difference between these mean values. The congressional set-aside requirement could be shaping agency behaviors in terms of the quantity of awards they make. Third, there could also be no actual relationship between the program budget and the number of awards.

Next, I did the same analysis for the percent of an agency’s extramural research budget allocated for the SBIR/STTR program, and the size of the awards and the average size (in thousands of dollars) of the awards. There is a statistically-significant relationship (p<.001) between an agency’s extramural research budget allocated for the SBIR/STTR program and the size of the awards and the average size (in thousands of dollars) of awards. This result is not surprising since one would expect that agencies with more resources would be able to—on average—provide larger awards.
Next, I chose a fixed-effects regression model because I expected there to be changes in my dependent variable because of time and agency. The fixed-effects regression model treats the data as a panel and controls for time in-variate effects for each group and time (Woolridge, 2016). I used the same dependent variables as before and used two different models. My first model only used the same independent variables but controlled for time and agency. My second model controlled for agency compliance with the congressional set-aside requirement because there is reason to believe that the congressional set-aside requirement could influence agency behavior. In both models, I removed data for the Department of Health and Human Services because this agency is well above the mean and is skewing the data. I recommend that Health and Human Services be analyzed in its own context. I do not find a statistically significant
relationship between my independent and dependent variables in either model. As this is an exploratory data analysis, I do not have enough evidence to make a conclusion. The three possibilities I outlined after my OLS regression are still possible and I do not have enough evidence to outright reject or support them. More work is needed to continue exploring this topic.

**Next Steps**

There are many avenues for further research. One avenue is to analyze the geographic distribution of SBIR awards (see figure #). One reason for why the implementation of the SBIR/STTR program may change could be the pressure for the politically-appointed leadership of agencies to support the electoral goals of the president. I chose this analytical frame because SBIR/STTR awards are government subsidies that are allocated by executive branch agencies, which face the challenge of making decisions on merit rather than on political partisanship. I have reason to believe that executive agencies may act on partisanship, based on prior research by Kriner and Reeves (2016, 2015), who showed that presidents—through executive agencies—tend to favor spending the federal government’s money in some geographical locations more than they do in others. Kriner and Reeves (2016), in their study of the distribution of all federal grant funding, found that presidents tend to spend more in geographical areas that supported them in elections throughout their presidency, and in geographical areas that are thought to be important swing districts in the year before an election. While science policy has typically been seen as a distributional endeavor, few scholars have examined the effects of partisanship and particularism within science policy.
This preliminary spatial analysis shows SBIR/STTR awards made in 2012 in yellow and maps them to the electoral vote data in the 2012 presidential election. Red counties were won by Mitt Romney and blue were won by Barack Obama.


Lanahan, Lauren (University Of Oregon); Feldman, Maryann P. (University Of North Carolina At Chapel Hill). (2017). *SBIR project level data.tab* [Data set]. Harvard Dataverse.
https://doi.org/10.7910/DVN/UC2ENJ/Y3A0YK


https://doi.org/10.3386/w5753

https://doi.org/10.1093/scipol/scs052

https://doi.org/10.1080/10438590802208166


https://arxiv.org/abs/1403.2805


