Measuring Innovation and Innovation Activities Using Non-Survey Data Sources

Sallie Keller, Gizem Korkmaz, Bianica Pires, Stephanie Shipp
Social and Decision Analytics Laboratory (SDAL), Biocomplexity Institute of Virginia Tech
Gary Anderson, Karen Hamrick, Carol Robbins
National Center for Science and Engineering Statistics (NCSES), National Science Foundation

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference

Abstract

The National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES) is exploring opportunities to develop new ways to measure innovation, STEM (Science Technology, Engineering, Mathematics) pathways, and outputs in the economy not traditionally measured in official statistics, such as open source software. NCSES would like to understand the opportunities in natural data flows to supplement or enhance its current efforts in providing science and engineering indicators. Working collaboratively with the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech, NCSES is exploring the identification and collection of data that naturally exists for other reasons and repurposing these data to measure innovation and related concepts.

Introduction

We are exploring whether we can identify and use data that naturally exists for other reasons and repurpose these data to measure innovation and innovation activities. With these objectives in mind, the initial findings from three projects are briefly described in this paper.

1. Examining the feasibility of measuring the characteristics of Open Source Software (OSS) as a first step to measuring its scope, impact, and value to the economy.
2. Measuring product innovation using administrative and opportunity data, where data discovery is instrumental in assessing whether companies tell their innovation story publicly and if so how, such as through company websites, portals, news media, blogs, and databases that aggregate company information.
3. Identifying STEM pathways by finding and assessing the quality of administrative and opportunity data sources and combining these with survey data across multiple agencies to measure the pathways that individuals take to obtain a STEM position.[1]

The research described in this paper aims at benefiting NCSES in continuing its development of data collections and acquisitions to provide better statistical products to its data users. We are developing data collection methods beyond traditional survey collections, exploring administrative data and opportunity data (e.g., data scraped from the internet) with spatial or temporal data collection designs, such that metric creation from these data can be repeated.

We use a Data Science Framework to guide this work (Keller, Lancaster, and Shipp, 2017). This approach involves a disciplined process of identifying data sources, preparing them for use, and assessing the value of these sources for their intended use (Keller et al., 2017). As shown in Figure 1, the process starts with the research problem and then proceeds to iteratively undertake a well-defined collection of steps. The purpose of the Data Science Framework is to leverage the data revolution by creating repeatable and measurable processes for the use and repurpose of existing data sources to support new research questions.

[1] The Bureau of Labor Statistics, Standard Occupational Classification Policy Committee (SOCPC) uses the 2010 SOC to define STEM (Science, Technology, Engineering, Mathematics) occupations as those in STEM and STEM-related domains. STEM domains are the life and physical science, engineering, mathematics, information technology, and social science. STEM-related domains are architecture and health. The term STEM is used throughout the paper to include occupations in both STEM and STEM-related domains.

The steps in the Data Science Framework include the following.

- *Data Discovery* to identify data that naturally exist, including administrative, previously designed data collections (e.g., surveys), opportunity, and procedural data. These data sources were originally collected for other reasons and will require repurposing to measure concepts of interest to the research at hand. Once discovered, the data sources are *inventoried* and then *screened* to determine which are useful to *acquire* for the intended research questions.
- *Data Ingestion and Governance* invokes the data quality assessment of the data sources through *data profiling* to evaluate the representativeness, timeliness, accuracy, consistency, completeness, reliability, and relevance of the data. The ingestion process needs to capture all known metadata and provenance to help guide the profiling processes and inform the *data preparation* steps. The data may come from many difference sources (e.g., organizations and agencies). The governance (e.g., access and privacy) that surrounds the data needs to be captured and adhered to during the *data linkage steps*. *Data exploration* combines and explores the data to gain an understanding in the spatial and temporal biases and coverage relative to the intent of the research questions.
- *Fitness-for-Use Assessment* and the *Statistical Modeling and Analyses* are tightly coupled. Given a particular analysis, fitness-for-use of the associated data is a characterization of the information content in the data that can support the particular analysis. This is a function of the statistical model(s) used, the data quality needs of the model(s), and the data coverage needs of the model(s). Of note, statistical modeling and analyses is considered broadly in this framework and includes evaluation.

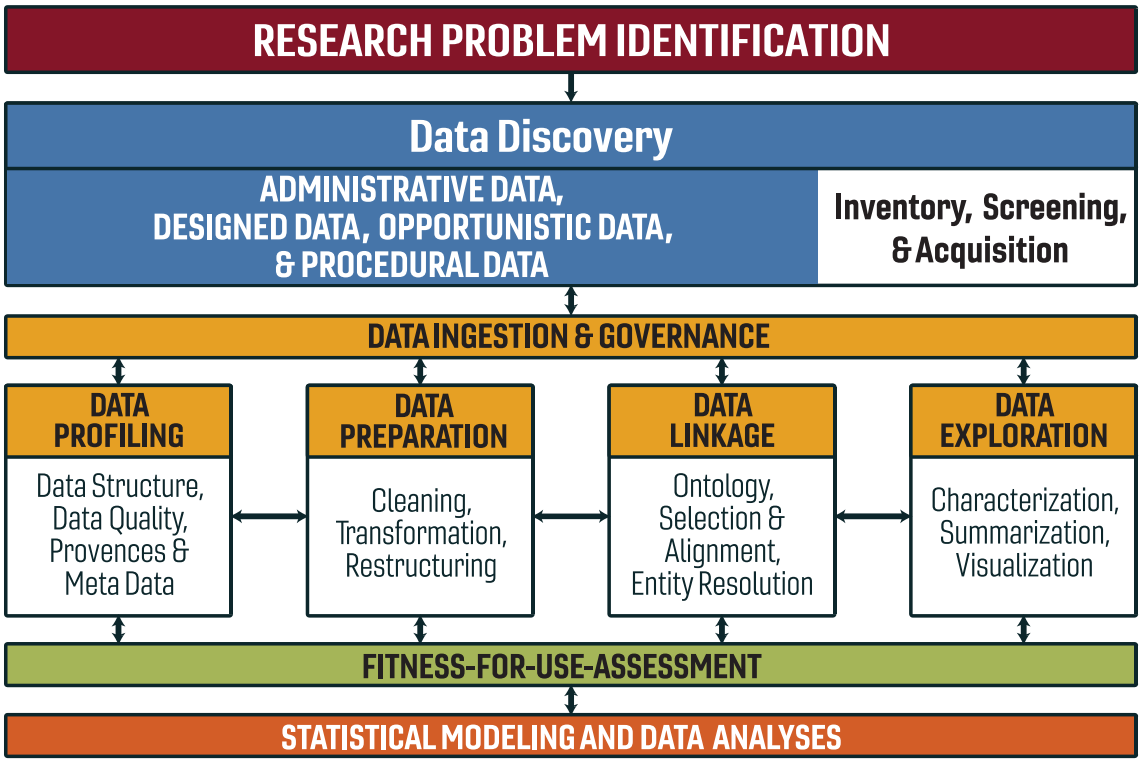


Figure 1. Data Science Framework. The process starts with the research question and continues through the following steps: data discovery, inventory, screening, and acquisition; data quality assessment (data profiling); data preparation and linkage; data exploration; assessment of the fitness-for-use; and statistical modeling and analyses.

In our collaboration with NCSES, we are addressing two overarching research questions

- Can we measure innovation and innovation activities using non-survey data sources?

- Can such measurement complement current survey approaches in a defensible and repeatable manner?

The Data Discovery phase of our Data Science Framework has been the primary focus of the early stages of the research described here.

Background – Defining Innovation

Innovation, and activities that contribute to innovation, have always been difficult to measure comprehensively. Innovation is believed to be the primary source of growth in the economy, sometimes measured directly but more often reflected indirectly or as a residual in our national statistics. The National Center for Science and Engineering Statistics is the NSF's statistical agency. As a supplement to survey data collection, NCSES is interested in assessing the feasibility of using non-survey measures to measure innovation and activities that lead to innovation.

Most definitions of innovation start with the ideas of Joseph Schumpeter in his *Theory of Economic Development* (1934). He emphasized the role of market forces in producing change through new products and processes, new markets, the discovery of new inputs, and changes in the organization of firms and markets.

Today, the Oslo Manual provides guidelines for internationally comparable statistics on innovation and is considered the primary source used to define innovation indicators (OECD/Eurostat 2005). The Oslo Manual defines innovation as a new product, process, or method of production that has been implemented. These guidelines have generally been applied to market activity in the business sector, and thus build from Schumpeter's defining work. The principal way that many national federal systems collect data about innovation is through the Community Innovation Surveys.[2] In the US, the NSF-NCSES Business Research & Development and Innovation Survey (BRDIS) survey has collected statistics on business innovation since 2008 (Kindlon and Jankowski 2010).

Historically, measurement challenges have limited the ability of national statistics to provide robust indicators of innovation activities, including the intangible capital, such as the creation of knowledge products of the organization capital and the tacit knowledge of its workers (Robbins 2016). There has been a flurry of research over the last few decades in both the business and economics literature about the importance of measuring intangibles in the economy, (e.g., Haskel & Westlake 2018). The US national accounts now measure computer software, R&D, and artistic and literary originals as capital inputs. Related to the accurate measurement of innovation and innovation-related activities is the collection of data about the science and engineering workforce and the condition and progress of STEM education in the United States.[3]

Project 1 - Measuring open source software innovation

Open- source software (OSS) is computer software with its source code that is made available to anyone with a license. The license gives the user the rights to examine, modify, and distribute the software to others for any purpose.[4] Its reuse makes it a kind of intangible asset that is both an input to innovation and an innovation activity.

Software overall is included as business investment in national statistics, though outside the business sector this investment is incompletely measured. The open source component of software investment is not separately measured, and this includes the OSS created with public spending by the government and academic sectors. We are developing a framework to measure characteristics of this open source software as a first step to measuring its scope, impact, and value in the economy.

Our first step is to discover data sources to assess whether the scope and impact of OSS could be measured using publicly available data. For this early research, we identified six data sources and screened them to determine if they were useful to address our questions: for example, (1) can the stock and flow of OSS be measured? And (2) can

[2] http://ec.europa.eu/eurostat/web/microdata/community-innovation-survey
[3] NCSES' central role is the collection, interpretation, analysis, and dissemination of objective data about the entire science and engineering enterprise (R&D, innovation, STEM education and workforce, and competitiveness). (https://www.nsf.gov/statistics/about-ncses.cfm).
[4] Open Source Initiative (1998) (https://opensource.org/osd).

we examine the sectors, collaborators, and users of OSS? Based on these criteria, we selected three OSS repositories for this research. To access the data from the three repositories, we scraped data from the internet through the use of APIs.[5] The repositories are large:

- Black Duck/Open Hub has over 675,000 OSS projects
- SourceForge has 450,000 OSS projects,
- Depsy, funded by NSF, has almost 10,000 R packages.

After we discovered and acquired these three data sources, we profiled the data to assess its quality. We then cleaned and conducted initial exploratory analysis to understand the data. By doing this, we are able assess what is feasible in terms of measuring open source software production and usage.

Selected descriptive results for the three sources of data are presented to help us understand what is feasible in terms of measurement of OSS creation and usage

For BlackDuck/OpenHub, the distribution of projects by total lines of code, total commits,[6] and contributors are presented in Figure 2. These data are compared for a random set of selected projects, and those that are labeled as "relevant" by OpenHub. Figure 3 shows the number of the top 10 languages from the relevant projects groups. Additional results are available in our publications and presentations (Keller 2017; Korkmaz, et al. 2018; Robbins, et al. 2018; Robbins 2018).

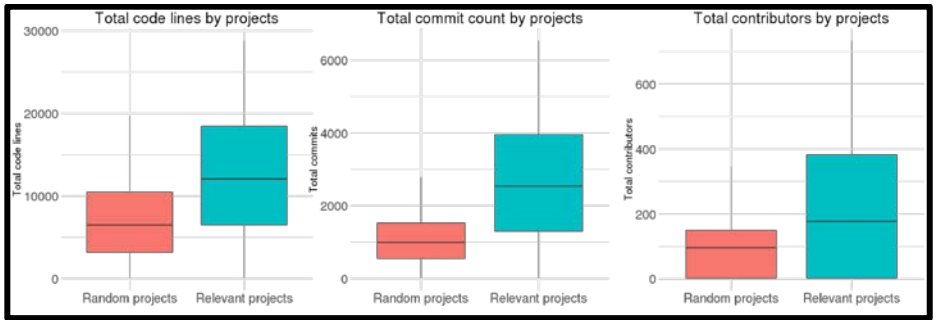
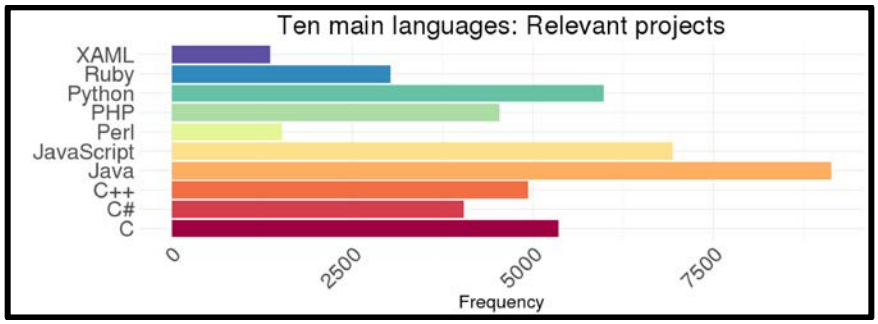


Figure 2. Development activity of open source software projects on Black/Duck OpenHub repository. For each group by project, the figure presents the total code lines, total commit counts (see footnote 6 for definition), and total contributions. These characteristics can be used to measure the value of the open source software when combined with other data, such as the average wage of software developers.



[5] Application Programming Interface (API) is a set of subroutine definitions, protocols, and tools for building application software. In general terms, it is a set of clearly defined methods of communication between various software components.

[6] To contribute source code on most large projects, one must make modifications and then "commit" those changes to a central repository

Figure 3. Top languages used in the open source software projects on Black Duck/OpenHub open source software repository. Java and Java Script are the most prevalent languages used.

We use SourceForge data to identify different categories/uses of open source software projects. Figure 4 illustrates the number of projects in the top categories and the average downloads per project in each category.

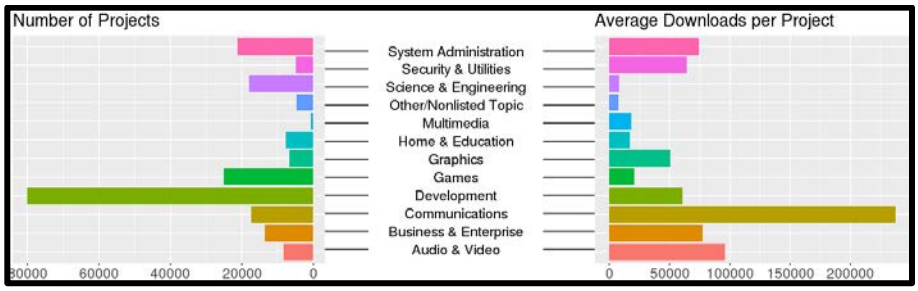


Figure 4. The number of projects on SourceForge open source software repository by category and the average downloads per project in each category. The group with the largest number of projects is development. Communications projects have the largest number of downloads.

We use the information from the Depsy repository to learn about the R packages that are required for the development of other packages (i.e., dependencies) to generate the network of R packages where a directed edge from j to i indicates that the package j requires i to be installed to function. We obtain a network with 7,389 nodes and 20,235 directed edges. Figure 5 uses a subgraph (starting at MASS[7] as the root node) with 149 nodes for illustrative purposes. Models developed to measure the value and cost of OSS projects need to take into account these dependencies.

There are many ways that one might measure the value of OSS outside of the business sector. Our initial focus will be to estimate the costs to produce OSS based on the number of hours, lines of code, and other characteristics of OSS. This approach is similar to how other intangible capital is measured in the national accounts. We will then expand our models to measure diffusion by including network measures such as centrality or number of uses of one software package by another.

[7] MASS is the R-package `Modern Applied Statistics with S'

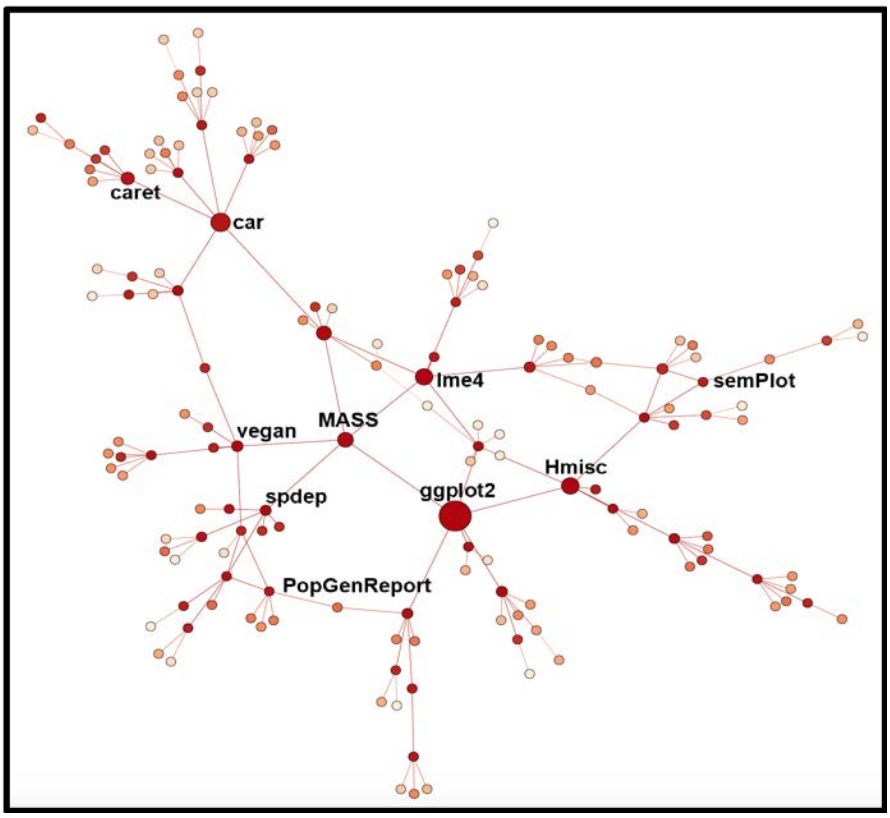


Figure 5. A subgraph of the dependency network of R packages on Depsy open source software repository. A link between two packages indicates that one package requires the other to be installed to function. In this figure, the color represents the impact of the package given by Depsy (calculated as the average of citations, downloads, and the software reuse), and the size of the node corresponds to the number of downloads.

Project 2 - Measuring company innovation with a focus on product innovation

The current survey approaches measure innovation overall and some innovative activities, such as patenting and funded research. For example, the NSF BRDIS survey asks companies to report on new or significantly improved goods, services, methods of production, logistics, or support activities. They ask if these new or significantly improved products, services, or methods are new to one of the company's markets or only new to the company (Kindlon and Jankowski 2010).

The advent of data science and access to non-survey data, including administrative and opportunity data, create the potential to complement survey-based measures of innovation with a broader and perhaps richer set of innovation measures. To leverage these opportunities, we are assessing the feasibility of whether we can count and measure product innovation through non-survey sources, such as product announcements, press releases, websites and other information obtained through web scraping or queries of selected companies.

To undertake our research, we examined the academic literature and business reports and reviewed innovation awards and used this to guide our data discovery process. The Data Discovery phase of our Data Science Framework is a critical step in this iterative process. Given the research question, the process involves identifying as many data sources as possible, screening using a series of questions about the applicability and usability of the data sources, and then selecting those data sources that meet the screening criteria. To assess whether it is possible to identify product innovation, we screened each data source using the following questions:

- Can searches be conducted by industry sector?
- Are NAICs codes identified for each sector and each company?
- Are annual company reports available?
- Do news stories and reports report on product launches (new products) or is the focus primarily financial picture of company/companies?

As part of data discovery, we conducted in-depth reviews of three different kinds of companies to learn about what information is publicly available. The three companies chosen are known for creating innovative products. The companies are Procter & Gamble (P&G) , Capital One, and Luna Innovations.

- P&G is a Fortune 50 consumer goods company whose motto is that "Innovation starts with the customer." They have developed innovative products such as the Gillette Flexball and Tide Pods.
- Capital One is a large bank whose tagline is "to design new experiences that empower customers to better manage their finances." One example of a recent innovation is ENO, a new text message-based assistant that can help customers with a variety of banking and credit card activities.
- Luna Innovations is a small research company that produces fiber optics and has won many grants and awards for innovative research.

Through these in-depth reviews, we evaluated whether we could identify new products introduced to the market that are announced on their platforms, (referred to as internal data sources, e.g., websites, social media accounts), and use these particular launches to validate external data sources that mention product launches by these companies and by others (e.g., news websites, databases).

We then examined external sources to evaluate their feasibility to develop repeatable and reliable measures of product innovation. Table 1 presents some examples of innovations from each of these companies, the internal platform of the announcement and the external data source identified.

We next undertook research to assess what we could find by examining data sources for the pharmaceuticals industry. We pulled and tabulated data from the Food and Drug Administration (FDA) database for pharmaceutical approvals. During 2013-2015, 976 unique companies had a total of 13,074 submissions approved. The top 25 companies with the most approvals are presented in Figure 6. Teva and Mylan were the only two companies that had over 100 approvals. We also pulled and tabulated data about new products from Pharmacy Times and Pharmacy Today, two industry journals. We are currently comparing and reconciling the findings.

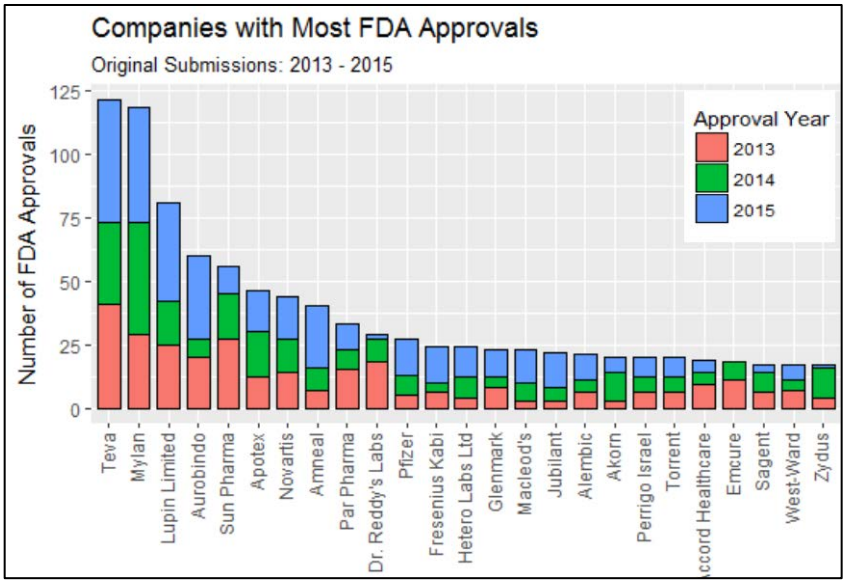


Figure 6. Companies with the most FDA approvals.

Table 1. Innovation Examples from In-depth Internal Company Sources and External Sources for Three Innovative Companies, Procter & Gamble, Capital One, and Luna Innovations.

Company	Product	Type	Company Platform	Source Type	External Data Source	Source Type	Potential
P&G	Tide PODS with Downy	Detergent	Latest Innovations	Innovation Webpage	Wall Street Journal	News Website	yes
			Link to the announcement	News Release			
	Febreze with OdorClear Technology	Home Care	Latest Innovations	Innovation Webpage	Business Wire	News Website	yes
					Savings with Denise	Blog	no
					WKYC-TV	Online TV	no
	Vicks Sinex	Health Care	Latest Innovations	Innovation Webpage	Business Wire	Awards	yes
	Luvs Diapers with NightLock Plus	Diapers	Latest Innovations	Innovation Webpage	Wall street Journal	News Website	yes
Capital One	Auto Navigator	Car Technology	Cars	Website	Owler	Database	yes
					vvdailypress	Blog	no
	Digital Home Loan	Mortgage			PR Newswire	News Website	yes
					Owler	Database /reports	yes
	Eno	Virtual Assistant	Apps/Eno	Website	Reuters	News website	yes
Luna Innovations	Odisi	Fiber Optic Sensing Platform	Website	News Release			
			Website	Blog Post			

We learned that companies announce their innovations on their websites, and although this may be a challenging way to capture information, it might provide a way to validate data collected from other sources. Importantly, we found that the use of large databases that aggregate news stories, press releases, Securities and Exchange Commission (SEC) filings, and other publicly available information may be another promising approach.

The next steps are to continue to explore the use of these resources for two industries – pharmaceutical manufacturing and software. For pharmaceuticals, we will continue to use and reconcile data from the Food and Drug Administration (FDA) approvals database, trade journals, and SEC filings. For the software industry, we are primarily using databases that aggregate multiple sources of data.

Project 3 - Identifying and measuring education and training pathways to STEM occupations

Science, technology, engineering, and math (STEM) higher education and research is vital for a competitive workforce and innovation (NSB, 2015; NSB, 2016). Today, the STEM workforce includes individuals with post-secondary degrees in STEM areas, as well as individuals that acquire the necessary STEM knowledge, skills, and abilities for a particular job that requires this training. Promoting a STEM capable workforce requires knowledge of the diverse set of pathways available to individuals desiring to enter the STEM workforce (NSB 2015). Similar to the Bureau of Labor Statistics Standard Occupational Classification system[8] we are defining STEM occupations as occupations in STEM and STEM-related domains. STEM domains are the life and physical science, engineering, mathematics, information technology, and social science. STEM-related domains are architecture and health. Simply put, a STEM occupation is a job or profession that requires some form of STEM training.

[8] https://www.bls.gov/soc/

Traditional and alternative pathways to STEM occupations are shown and compared in Figure 7. A traditional pathway to a STEM occupation includes high school to baccalaureate to post-baccalaureate degrees. However, pathways to a STEM occupation can be much more complicated. There are many avenues to a STEM occupation that may or may not involve a baccalaureate degree. For example, some high school graduates obtain a two-year community college degree and are employed in STEM occupations, while others obtain industry certifications as their entrée to STEM occupations. The role of massive open online courses (MOOCs), such as Coursera, and other online certification is still a largely unexplored area.

NCSES currently measures traditional STEM pathways through surveys, such as the National Survey of College Graduates and the Survey of Earned Doctorates (NSF 2017). Their focus is on those who receive a bachelors, masters, or PhD degrees in science and engineering fields. However, there is strong interest in understanding *all* STEM pathways, including pathways from high school, community college, apprenticeships, online courses, and other non-traditional entries into STEM occupations. NCSES would also like to measure the knowledge, skills, and abilities gained through varying STEM pathways and how they match to employer requirements.

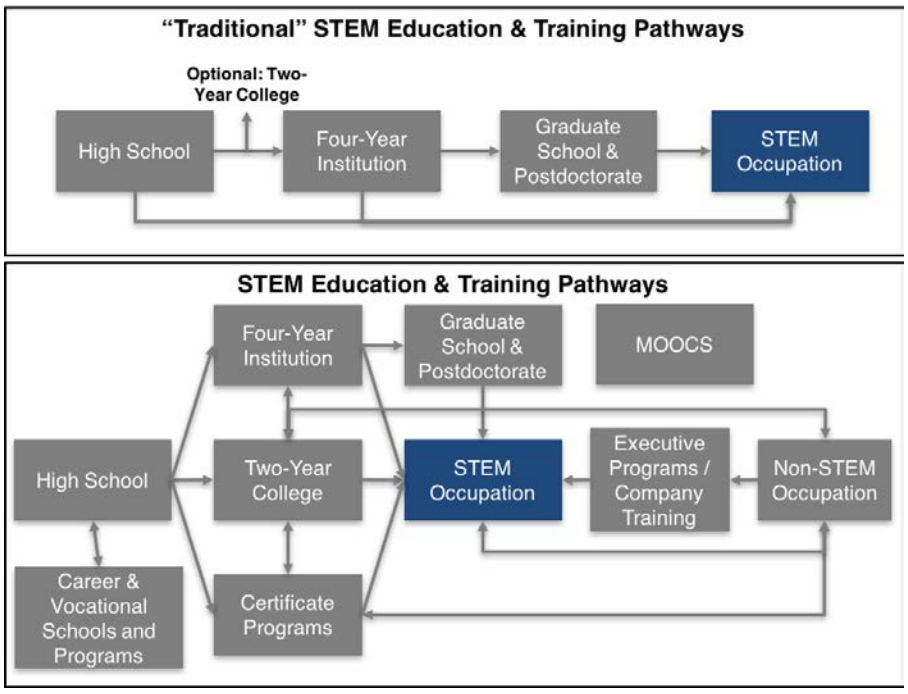


Figure 7. A mapping of traditional and alternative STEM pathways.
The figure provides a mapping of the traditional STEM education and training pipeline (high school, undergraduate university, graduate school, STEM occupation) in comparison to the diffuse set of pathways an individual can take towards the STEM workforce, e.g., high school to community college to STEM occupation.

In this project, we are assessing the feasibility of identifying data to measure *all* STEM pathways. Through our Data Discovery process, we are discovering and examining the quality of existing local, state, and federal data sources (both administrative and survey) as well as opportunity data, such as job postings, that could be repurposed in new ways to measure post-secondary education and training pathways.

One challenge is that there is no single standard for what occupations are defined as STEM. For example, NSF defines *Science & Engineering* occupations as any college graduate employed in an S&E or S&E-related occupation, regardless of field of study. The Bureau of Labor Statistics, Standard Occupational Classification Policy Committee (SOCPC) uses the 2010 SOC to define *STEM occupations* as those in STEM and STEM-related domains. Rothwell (2013) categorizes occupations into *High-STEM* and *Super-STEM* using STEM knowledge scores from the Department of Labor Occupational Information Network (O*NET). *Skilled Technical* occupations

are defined as STEM occupations that require a high level of knowledge in a technical domain but do not require a bachelor's degree (Rothwell 2015).

One avenue to consider is the knowledge, skills, and abilities (KSAs) necessary for a particular occupation from two perspectives (1) those that employees acquire in their paths to a STEM occupation and (2) those required by employers hiring candidates for STEM jobs. We are just beginning to conduct this research by examining resumes and job postings.

We explored over 70 data sources across federal, state, and local government as well as industry. To document what we found, we created a data inventory that describes each data source, for example, the type of data, timeframe, unit of analysis, and accessibility. (For related research, see Ziemer et al. 2017.)

As part of the Data Discovery phase, we explored what the STEM jobs postings landscape looks like geographically depending on the occupation classification system we use. For example, the Data at Work initiative at the University of Chicago provides publicly accessible aggregated counts of jobs advertisements by occupation by state.[9] We linked each occupation to the three classification systems (BLS, NSF, Rothwell).

In a second example, we used Open Data, Open Jobs data collected by Virginia Tech and George Mason University, in collaboration with the state of Virginia.[10] We mapped the percentage of job postings within a county that are for the STEM jobs using the three classification systems. Both explorations provided interesting insights about employer demand and allowed us to test the feasibility of using data from job postings but there are too many uncertainties and gaps in these data sources to pursue using them in the analysis. Instead, we are exploring for-profit data aggregators that collect data on job postings and resumes from a multiplicity of internet sites.

Our next steps are to choose one, harmonize the three, or create our own STEM occupation classification system and to extend this to include skilled technical occupations. We are also acquiring resume and job posting data to assess what knowledge, skills, and abilities are required for different kinds of jobs and whether they match the knowledge, skills, and abilities found on resumes.

Conclusions - Why measure innovation using non-traditional data sources?

NSF's National Center for Science and Engineering Statistics (NCSES) is interested in learning how non-survey sources of data can enhance or supplement current survey data. Innovation is key to economic growth and perhaps using non-survey sources of data may provide new insights. We are testing the feasibility of measuring sources of innovation in three case studies (1) non-business Open Source Software, (2) product innovation from databases and websites, and (3) STEM career pathways using new data sources such as resumes and job postings. Our conclusion is that these approaches are ambitious but feasible.

References

Haskel, J, Westlake, S. (2018). *Capitalism Without Capital: The Rise of the Intangible Economy* Princeton University Press.

Keller, S. (2017) "Modeling, Infrastructures, and Standards: Enabling New Opportunities to Observe and Measure." Arthur M. Sackler Colloquia of Sciences: Modeling and Visualizing Science and Technology Developments, Irvine, CA, December 4-5, 2017.

Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data-driven governance: Creating a new foundation for democracy. *Statistics and Public Policy*, 1-11.

[9] https://dssg.uchicago.edu/2017/08/03/introducing-the-datawork-research-hub/
[10] http://opendata.cs.vt.edu

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Reviews of Statistics and its Applications,* 4:85-108.

Kindlon, A and Jankowski, J. 2017. Rates of Innovation among U.S. Businesses Stay Steady: Data from the 2014 Business R&D and Innovation Survey, NSF 17-321. August. https://www.nsf.gov/statistics/2017/nsf17321/

Korkmaz, G, Kelling, C. Robbins, et al. (2018) "The Scope and Impact of Open Source Software as Intangible Capital: A Framework for Measurement with an Application Based on the Use of R Packages." National Bureau of Economic Research Conference on Research in Income and Wealth (CRIW) Pre-Conference: Big Data for 21st Century Economic Statistics, Cambridge, MA, July 18, 2018.

National Science Board. 2016. *Science and Engineering Indicators 2016*. Arlington, VA: National Science Foundation (NSB-2016-1).

National Science Board (NSB). (2015) *Revisiting the STEM Workforce*. February 4. NSB 201510.pdf

National Science Foundation (NSF). 2018. *2016 Doctorate Recipients from U.S. Universities*. March. https://www.nsf.gov/statistics/2018/nsf18304/static/report/nsf18304-report.pdf

OECD/Eurostat (2005) Oslo manual: Guidelines for collecting and interpreting innovation data. 3rd Edition, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264013100-en

Robbins CA (2016) Using new growth theory to sharpen the focus on people and places in innovation Schumpeter, J. (1934) *Theory of Economic Development, Cambridge University Press,* Cambridge, MA.

Robbins, C., Korkmaz, G., Kelling, C. et al. (2018) "Illuminating the Scope and Impact of Open Source Software: A Framework for Measurement with an Application Based on the R Program and Its Impacts." International Association for Research in Income and Wealth, Copenhagen, Denmark, August 20-25, 2018.

Robbins, C. (2018) "New Opportunities to Observe and Measure Innovation, Modeling, Infrastructure, and Standards." Interagency Council on Statistical Policy (ICSP) Big Data Day, Committee on National Statistics (CNSTAT), Washington, DC, May 10-11, 2018.

Rothwell, J. (2013). The Hidden STEM Economy. Metropolitan Policy Program at Brookings

Rothwell, J. T. (2015). Defining skilled technical work. Brookings Institute.

Ziemer, K. S., Pires, B., Lancaster, V., Keller, S., Orr, M., & Shipp, S. (2017). A New Lens on High School Dropout: Use of Correspondence Analysis and the Statewide Longitudinal Data System. *The American Statistician*.