Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples

Sallie Keller[a,1], Gizem Korkmaz[a], Carol Robbins[b], and Stephanie Shipp[a]

[a]Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Arlington, VA 22203; and [b]The National Center for Science and Engineering Statistics, National Science Foundation, Alexandria, VA 22314

Edited by Katy Börner, Indiana University, Bloomington, IN, and accepted by Editorial Board Member Pablo G. Debenedetti August 25, 2018 (received for review February 16, 2018)

Measuring the value of intangibles is not easy, because they are critical but usually invisible components of the innovation process. Today, access to nonsurvey data sources, such as administrative data and repositories captured on web pages, opens opportunities to create intangibles based on new sources of information and capture intangible innovations in new ways. Intangibles include ownership of innovative property and human resources that make a company unique but are currently unmeasured. For example, intangibles represent the value of a company's databases and software, the tacit knowledge of their workers, and the investments in research and development (R&D) and design. Through two case studies, the challenges and processes to both create and measure intangibles are presented using a data science framework that outlines processes to discover, acquire, profile, clean, link, explore the fitness-for-use, and statistically analyze the data. The first case study shows that creating organizational innovation is possible by linking administrative data across business processes in a Fortune 500 company. The motivation for this research is to develop company processes capable of synchronizing their supply chain end to end while capturing dynamics that can alter the inventory, profits, and service balance. The second example shows the feasibility of measurement of innovation related to the characteristics of open source software through data scraped from software repositories that provide this information. The ultimate goal is to develop accurate and repeatable measures to estimate the value of nonbusiness sector open source software to the economy. This early work shows the feasibility of these approaches.

intangibles | measurement | open source software | data science | nonsurvey data

Many inputs to innovation in the form of research and development (R&D) expenditures are well-measured; however, broader flows of inputs for innovation within government, industry, academic, and household sectors are less well-captured (1). Innovation is typically captured and measured using surveys, patent analysis, case studies, and peer reviews, and most available statistics are focused on the business sector. Definitions of innovation focus on the creation of new and improved products, processes, and marketing and organizational business mechanisms. Innovation is measured through its incidence (survey measurement), activities (primarily science, technology, engineering, and mathematics education and workforce), outputs (products and processes), and outcomes (economic growth and societal benefits) (2). Because of the link to economic growth, policymakers and researchers are interested in understanding and supporting activities that lead to innovation.

Traditional approaches to measuring innovation leave many types of innovation and inputs to innovation uncaptured, because they are produced within firms, are not commercialized, and often represent intangible assets that are hard to put a price on, such as knowledge, core competencies, and business processes, including organizational innovation (3). Today, there are many examples of innovative outputs that are not sold in the market, such as open source software (OSS) and free online education (4). Furthermore, activities in the household sector, including inventions and social innovation (e.g., food delivery to poor rural children during the summer), are not included in summary data on innovation, because they are outside of the scope of business activity (5). There are many nonsurvey data sources created and used in the business and nonbusiness sectors that may provide signals that can lead to new measures of innovation.

Key Terminology

Innovation. Innovation is the implementation of a new or significantly improved product (good or service) or process, a new marketing method, or a new organization method in business practices, workplace organization, or external relations. This definition is based on the Oslo Manual, which provides guidance for internationally comparable measurement and is produced by the Organization for Economic Cooperation and Development (OECD) and Eurostat (6). The ideas of Schumpeter (7) are fundamental to this understanding of innovation, emphasizing 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.

Intangible Capital. Intangible capital is a nonphysical factor expected to generate future benefits to the entities that control their use (6).

Intellectual Property Products. Intellectual property product is a classification category in national economic accounts in which some intangible capital is measured. These products include computer software and databases, R&D, and entertainment and artistic originals (5).

Through a process of data discovery, acquisition, statistical data integration, and visualization, we use nonsurvey data sources, such as administrative records that capture business transactions and websites that capture repositories, to identify

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Modeling and Visualizing Science and Technology Developments," held December 4–5, 2017, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/modeling_and_visualizing.

Author contributions: S.K., G.K., C.R., and S.S. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. K.B. is a guest editor invited by the Editorial Board.

Published under the PNAS license.

[1]To whom correspondence should be addressed. Email: sak9tr@virginia.edu.

Published online December 10, 2018.

intangible inputs to innovation. Our focus is to assess accessibility and quality of the data to develop innovation measures. The ultimate goal is to evaluate the feasibility of creating scalable and repeatable metrics of innovation in the economy using these nonsurvey data sources to supplement and enhance survey data collected by agencies, such as the National Center for Science and Engineering Statistics within the National Science Foundation (NSF).

Measuring Intangible Capital

Many nations collect firm-based survey data through statistical agencies, requiring extensive research and testing to tune survey language to both the needs of statistical users and the data that firms are able to provide. In the United States, data about company innovation are collected through the NSF Business Research & Development and Innovation Survey (8). The community innovation survey is conducted every 2 y by national statistical offices throughout the European Union and in Norway and Iceland (9). The surveys are harmonized and designed to provide information about innovation by sector and region. The data are used to create the European Innovation Scoreboard and study innovation. Other countries also conduct community innovation surveys, such as Japan, Australia, and Canada (10).

These innovation surveys ask companies if they have produced innovative processes or products or developed new organizational processes or business models. Many observers argue that an unknown and potentially large part of innovation activity is currently unmeasured or undermeasured. Indicators focused on business sector activity are likely to miss the innovation that is incremental or that takes place in universities and government laboratories and by individuals, obscuring a potentially large set of contributions and linkages.

Martin (11) has referred to this as "dark innovation," the amount of innovation activity that is outside the scope of current measurement. A broader approach to defining innovation would encompass not only business activity but also, the innovation that takes place in households, universities, and governments. In these cases, innovation occurs when the product is used, rather than sold in the market (12), and is referred to as free innovation (4) or household production (13).

In the business sector, indicators for related but different activities, such as R&D performance and patenting, are often used as proxies for innovation. Both have limitations, as the innovation survey (used since 2008 in the United States) has focused on R&D performers, whereas patenting is more prevalent for some technologies than others. According to data collected using the Oslo Manual definition (6), one in six US firms (17%) introduced a new or significantly improved product or process between 2013 and 2015 (14). While these data do not provide cost savings from process improvements, firm revenue from the sale of products that were new to their markets was $1.2 trillion (8).

Innovation leads to the creation of economically useful knowledge in the form of intangible assets that can be an output of a productive process as well as an input into the creation of new output. These include creative works, scientific works, discoveries, inventions, and computer software as well as systems created within businesses. In addition to being outputs, these intangible assets also have the capacity to contribute to production of goods or services or are intended to generate future benefits to the entities that control their use (15). Intangibles are more likely to create spillovers and synergies than tangible capital (16).

Not all intangible investment becomes a successful innovation; however, the creation of these intangibles is a key activity in the process that brings new discoveries into use (16). When the resources used to create intangibles are measured, a cost-based measure of intangible investment can be estimated. As currently measured in US economic accounts, the magnitude of intangible investment is comparable with that of tangible investment in machinery, business equipment, and other capital equipment.

While US private sector investment in equipment (excluding that used in homes) was $1.1 trillion in 2017, investment in three intangible assets (intellectual property products) was almost $800 billion in 2017: R&D accounted for $335 billion, artistic originals accounted for $85 billion, and computer software accounted for about $375 billion according to Bureau of Economic Analysis (BEA) data. Of this software investment, more than one-third is created internally for the firm's own use, one-third is on custom software, and less than one-third is on prepackaged software (17).

Economists and others argue that there are important other types of intangible capital, including firm investments in human capital embedded in people (18). Examples of intangible capital investments are formal investments to create designs, develop and protect brands, train human capital in firms, and change organizational processes (19, 20).

Business accounting treats intangible investment somewhat differently from gross domestic product (GDP) accounts based on different accounting objectives (21). National accounts guidelines, designed to provide an aggregate picture of economic activity from the perspective of both buyers and sellers, recommend treating expenditures on intangibles, such as R&D, as investment, because they contribute to future production and income generation (22). In contrast, generally accepted accounting principles call for immediate expensing of R&D expenditures, because future benefits are uncertain (21).

Science of All Data

Hard-to-measure things are an increasing share of economic output. The data science framework, described in this section, gives us a way to think about measurement using existing data to capture undermeasured innovation. The objective of the data science framework is to leverage the data revolution by creating repeatable and measurable processes for the use and repurposing of existing data sources to support research questions. We have adopted the term "all data revolution," since our research focuses on data of all sizes, not just big data (23). This is an important distinction.

We categorize data into four categories and provide examples in the context of this paper (24, 25). Categorizing data by types can accelerate the data discovery phase (described below).

Designed data involve statistically designed and intentional observational data collections, such as from surveys, experiments, and registers. The NSF's Business Research and Innovation Survey is one example of a statistically designed survey to measure innovation (8), and the R&D Satellite Account (26) to the GDP is an example of an intentional data compilation that provides statistics on R&D investment.

Administrative data are collected for the administration of an organization or program by entities, such as government agencies, as they provide services, companies to track orders, and universities to record registered students.

Opportunity data are derived from internet-based information, such as websites and social media. For example, software repositories (e.g., OpenHub) provide not only the code but also, documentation and information about the creation and the use of the software, such as contributors and downloads.

Procedural data focus on processes and policies, such as a change in health care coverage or a data repository policy that outlines procedures and metadata required to store data.

Through our research, including research on intangible measurement, we have developed a data science framework (Fig. 1) to create methods to identify and integrate multiple data sources to address specific research questions (27).

Data discovery identifies data that naturally exist, including administrative, previously designed data collections, opportunity,

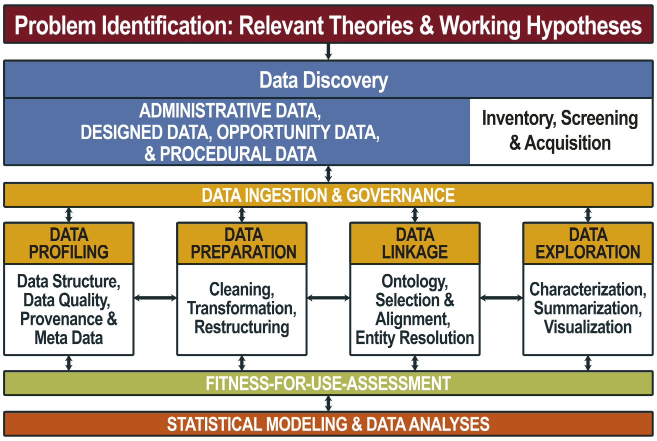
Problem Identification: Relevant Theories & Working Hypotheses
Data Discovery
ADMINISTRATIVE DATA, DESIGNED DATA, OPPORTUNITY DATA, & PROCEDURAL DATA
Inventory, Screening & Acquisition
DATA INGESTION & GOVERNANCE
DATA PROFILING
Data Structure, Data Quality, Provenance & Meta Data
DATA PREPARATION
Cleaning, Transformation, Restructuring
DATA LINKAGE
Ontology, Selection & Alignment, Entity Resolution
DATA EXPLORATION
Characterization, Summarization, Visualization
FITNESS-FOR-USE-ASSESSMENT
STATISTICAL MODELING & DATA ANALYSES

Fig. 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 data analyses. Adapted with permission from refs. 24 and 27.

and procedural data. These data sources are likely to have been collected for other reasons other than the problem at hand and require repurposing to measure concepts of interest. After 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 involve the 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. 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 specific research questions.

Fitness-for-use assessment, statistical modeling, and data 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, the statistical modeling and analyses step is considered broadly in this framework and includes evaluation.

The components in the data science framework are based on scientific principles from the engineering and computational fields (e.g., statistical process control, Total Data Quality Management, visualization tools, and simulation experiments) (28). The processes represented in our framework are critical for creating defensible and repeatable measures.

In this paper, we highlight the value and use of the data science framework through the presentation of two case studies that focus on intangible capital as an input to innovation. These studies, described below, show that it is challenging yet possible to create intangible innovations and to uncover existing intangible assets using nonsurvey sources of data.

Case Studies

Organizational Process Innovation in a Fortune 500 Company to Synchronize the Supply Chain Using Administrative Data. Developing organizational processes in a company is an investment that produces assets that are expected to be used repeatedly. Sometimes, these developments are routine and based on traditional models and data sources, such as simulation models for supply-chain synchronization based on material flows. In contrast, this case study is an example of using and linking administrative data (e.g., customer orders) and procedural data (e.g., holiday schedules) across existing data silos in a capital-intensive Fortune 500 company to synchronize the supply chain. The data sources used are traditionally used in business analytics applications and not for supply-chain synchronization. This process innovation uses a combination of Bayesian Hierarchal Modeling and discrete event simulation to simulate the supply-chain process from orders through shipments. The activities to undertake this research require tangible (e.g., data) and intangible (e.g., knowledge) inputs that lead to tangible outputs (cost reductions due to more efficient processes, on-time delivery rates) and intangible outputs (satisfied customers who receive their shipments accurately and on time). The expectation is that the benefits of these intangible outputs will lead to higher total shareholder returns paid through dividends to their shareholders. Procter & Gamble (P&G) uses similar language in defining how changes affect their bottom line through improvements in productivity and costs (29).

Although this research has already been published (30), here we are highlighting why the research produced intangible innovations in the creation of new knowledge and processes. In this case study, P&G, a manufacturer of consumer nondurables goods, sought to optimize their supply-chain processes through business transaction data (i.e., administrative data). The motivation for this work was to understand the interactions of inventory, profit, and service. Working collaboratively, we developed a set of integrated models to create a data informatics synchronization of the supply chain. The data include thousands of orders for products that are produced at and/or shipped from one of their P&G facilities to customers who are primarily retailers.

Traditional approaches to synchronize the supply chain are largely based on historical data. In the model presented here, we acquired both current and historical data and used these data to inform the model structure and as inputs to the models. This acquisition and use of the data in this case study show that implementing the data science framework is not a linear but rather, an iterative process. During data discovery and inventory phases, we held many discussions with subject matter experts (those responsible for operations at each step of the supply-chain process) in the company and by "walking" the supply chain through site visits to the suppliers, the factory floor, distributors, and retailers. Data profiling, cleaning, linking, and exploring the data involved documenting and identifying gaps, leading to the creation of a conceptual data model illustrated in Fig. 2.

The conceptual data model was a critical component of the research, providing a deep understanding of the supply-chain process through the mapping of the data flows. Using 1 y of administrative data, the data model brought together numerous data sources on customer orders, production, raw materials, inventory, shipments, and deliveries. The data tables are large (some of which contain over a million records), complex, and frequently, poorly documented. Our criteria for using the data were whether metadata that clearly articulated the relationships between the tables could be created and used to produce clean, accurate, and reliable tables for use in the model building.

Data preparation entailed standardizing the data to common units and exploring the data to inform the model development, leading to the creation of four interconnected simulators: (*i*) the orders simulator, (*ii*) the production simulator, (*iii*) the production planner simulator, and (*iv*) the shipment simulator. These simulators capture the supply chain, including estimating actual

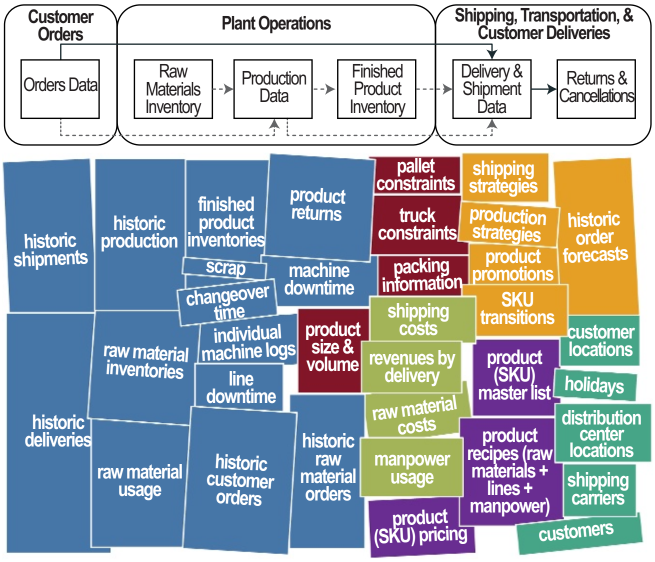

Fig. 2. Two different but complementary views of the administrative and procedural data flows identified to support the supply-chain synchronization. (*Upper*) A notional diagram of administrative data flows. Solid lines indicate where direct linking was found, and dashed lines indicate where modeled links were created. (*Lower*) A depiction of the many data sources associated with activities across the supply chain that were used directly in the models or used to inform the modeling. The data are color coded to align with the different features of the supply-chain process. Starting in the top right-hand corner, the orange boxes represent strategies and forecasts, the teal boxes show procedural data, the purple boxes are product-related data, the chartreuse boxes are costs and revenues, the maroon boxes are constraints (primarily shipping constraints), and the blue boxes are a variety of other measures from the supply-chain system. SKU, stock-keeping unit.

and simulated profits to validate the model. Much of the dynamics may be the result of the fact that humans are involved in all of the activities represented by the data. Although we have not yet tested this hypothesis, we believe that these dynamics are implicitly folded into the analysis as a result of using the business transaction data.

After calibrating the production planner simulator to reproduce recent historical results, strategies were explored to optimize profit and service characteristics. The model was parameterized so that the relative weights of demand, profitability, and rarity (products rarely ordered) could be adjusted, affecting how the production planner assigns production runs to lines given the customer orders (Fig. 3). We carried out a series of simulations to seek an optimal setting for the production planner. The results show a reduction in inventory and increases in profits and on-time delivery (service), thus increasing total shareholder return. Exact results cannot be shown due to the proprietary nature of the data.

This study shows that administrative data can be used to create a business process innovation through the development of models that are faithful to the supply-chain process and take into account the human interactions inherent in the data as well as the mechanical processes. This use of separate but linked statistical and discrete event models allows different sources of data to inform a manager about the supply-chain system and to predict its behavior under new conditions.

This approach allows flexibility to control and optimize the supply-chain process while providing a window into the entire supply-chain process. It captures the variations in the models and forecasts. Ultimately, these measures affect cash flow and total shareholder returns of the company. A next phase is to explore how it could be applied to other capital-intensive industries, like P&G, and in other types of industrial settings.

Measuring OSS, Innovation Activity, and Intangible Capital. This case study explores the possibility of measuring the value of the creation and use of an important intangible, OSS, through nonsurvey data sources. The Open Source Initiative (https://opensource.org/) defines OSS as a computer software with its source code available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone for any purpose.

Corrado et al. (20) describe intangible investment as expenditures intended to increase future output and consumption, and it is not dependent on whether the output is shared. However, sharing is an essential feature of OSS, and these motivations lead to software tools and activities with the fundamental characteristics of intangibles as capital investment.

Existing US statistical data on computer software investment account for business and government investment and do not separately identify OSS. Within the United States, investment in software by government and the private sector was nearly $400 billion in 2016 (31, 32). Investment in this internal use software is estimated based on compensation of employees and input costs (33).

While empirical data on the full extent of OSS are absent, evidence suggests that the scope and impact of public investments are substantial. For example, Apache is estimated to hold the largest market share of domains (35%) and active websites (41%) as of June 2018 (34). Apache was originally developed and used at the National Center for Supercomputing Applications at the University of Illinois, and it included federally supported research. The software was not developed in the business sector, but now, it is a significant contributor to business sector activities. Greenstein and Nagle (35) estimated the capital stock of Apache software (having a replacement cost in terms of the nearest available substitute) to be between $512 million and $12.8 billion in 2013. To the extent that the magnitude of this investment is significant, Greenstein and Nagle (35) note two measurement concerns: first, the omission of an investment good with a value but no price, and second, the attribution problem. When productive inputs are not

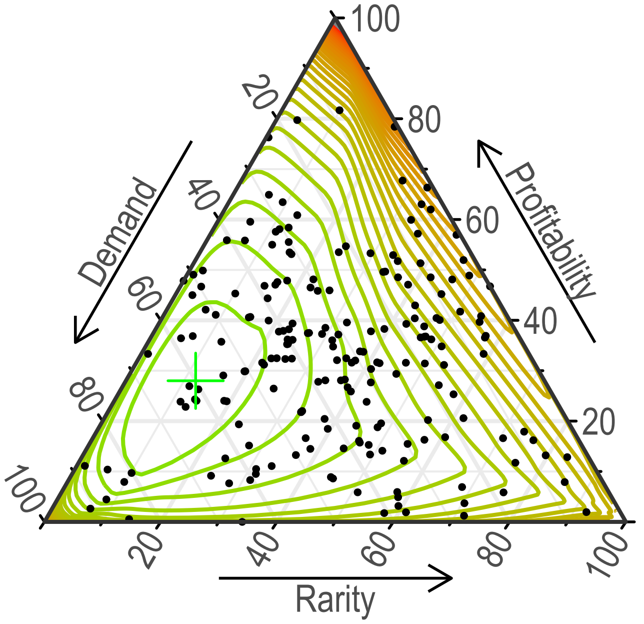

Fig. 3. Simulated profit for all on-time deliveries over the course of the test month as parameters of the production planner are varied. A response surface was fit to the simulation output, suggesting optimal planner settings.

Table 1. Data discovery, inventory, and acquisition of software repositories

Data Discovery	Data Inventory		Data Acquisition
Software Repository	**Data Description**	**Potential Uses**	**Data Collection Methods**
SourceForge.net is a community resource hub to help open projects grow. The site facilitates users to develop, download, review, and publish open source software. • ∼450,000 projects • 3.7 million users Run by Slashdot Media.	Descriptive metrics, e.g., • number of downloads • date of last update • number of reviews • average rating • sector • popularity • release date Users can post comments or questions on individual project forums.	• Analyze popularity, downloads, ratings, and software sectors • Conduct sentiment analysis using the forum text to explore user opinions about specific projects	• API (Application Programming Interface) to collect the project names • Web-scraping to collect detailed information on each project.
OpenHub.net is an online community and a public directory that provides statistics on: • 676,523 Free/Libre and OSS (FLOSS) projects • 698 organizations and 2,850 portfolio projects • 291,782 people; users and contributors Run by Black Duck Software and describes itself as "not a forge — it does not host projects and code."	Descriptive metrics, e.g., • number of lines of code added to and removed • commits by contributors and organizations • rates and times of updates • size and the activity level of the software developer team • user information, project usage Users can ask questions and provide feedback.	• Analyze project development and popularity over time • Create developer/contributor networks • Analyze sentiments about projects and sectors using forum text	• API to collect information on projects: – A subset of the most popular projects on OpenHub (8.2% of all projects) – A randomly selected subset (6.8% of all projects)
Depsy.org is a website that compiles R and Python packages to quantify coding impact in the scientific community. Initially funded by NSF and run by Impactstory.	Descriptive metrics on packages: • contributors • number of commits • number of downloads • number of citations • reuse/dependency • tags Metrics on developers: • number of packages contributing to R & Python • number of citations	Analyze *impact score* calculated as the average of these three: • downloads from software repositories • software reuse via reverse dependencies • and literature reuse (citations) Generate network of developers and packages, identify influential nodes, and OSS communities	• API is provided on Depsy.org. – Collected all of the R packages hosted on CRAN using Depsy.org. – Information about 9,810 R packages are scraped from Depsy.org (last update in Sept. 2015).
GitHub.com is a website hosting both private repositories and free accounts, commonly used to develop OSS projects and to store computer code. Provides access control and collaboration features such as bug tracking, feature requests, task management, and wikis. Run by GitHub, Inc.	Descriptive metrics, e.g., • users and github pages • followers/following • tags, stars • repositories • organizations • activities; commits, updates, push/pull requests	• Analyze project development and contributions over time • Explore popularity (impact) of projects using starred projects • Study developer influence using follower/following information • Analyze sentiments about projects using messages	• API is provided, but the rate limit is low. • Data collection is in progress.
OSalt.com (Open Source ALTernative) lists OSS that can be used in lieu of a proprietary software. Created by Anders Ingeman Rasmussen, an "entrepreneur" at Airflake.	• Proprietary software packages and substituting OSS • User comments on the relative utility of each one	Estimate approximate value of the OSS using the price of the proprietary software.	• Webcrawler and webscraping
Stack OverFlow is a forum-based tool used to spread knowledge about programming languages. It is primarily used to help solve coding problems. Run by Stack Exchange and financial assistance provided by corporate donors.	• Tags for different languages, methods, and coding related issues. • Compilation of user-contributed questions and answers about OSS projects and coding languages.	• Identify which terms and languages are most commonly associated with specific OSS. • Generate network of relevant projects using tags. • Determine popularity/use based on the activity • Conduct sentiment analysis for specific methods or programming languages to determine impact of packages.	• Webcrawler/Scraping/Text Mining. • There are challenges in scraping tags.

This table presents a description of six software repositories found during the data discovery and inventory phase of this research. The data provided in each repository and potential uses of the data are presented. The first three, Depsy.org, SourceForge.net, and OpenHub.net, were chosen for the initial study based on the variables of interest. The data were collected in July 2017.

measured, their impact can be attributed incorrectly to measured inputs (35).

The investments made in the private sector and in government in developing OSS are conceptually accounted for in BEA's estimates through the compensation of the software programmers and their related costs. What is missing is the value of the created OSS itself. For example, the Linux Foundation development report notes that, since 2005, over 15,000 developers from over 1,500 companies have contributed to the Linux kernel, with more than 200 companies and 1,500 programmers participating on a particular version (36). The part of OSS that is created and developed within universities and federal laboratories and by individuals is less well-understood and not measured.

Current and comprehensive survey data do not exist for the contributions of OSS. However, as it is disseminated online, a wealth of information is available to be scraped (i.e., opportunity data that include both metadata and information embedded in repositories and in the code and headers of the software programs themselves). Our contribution is to show how these data can be used to develop measures of OSS, shedding light on the impact of OSS innovation, which is currently not well-understood. The characteristics of OSS from software repositories are collected using publicly available data with the goal to test whether accurate and repeatable models can be built to estimate the value of nonbusiness OSS to the economy.

Data discovery, inventory, and acquisition. The data discovery phase for measuring OSS was guided by the following dimensions that are used for measuring creation and use of software (16).

Stock measures: How much OSS is in use?
Flow measures: How much is created each year?
Categories: What types can be identified?
Sectors and collaborators: Who creates it?
Users: Who benefits from its development?

The data discovery involved looking for data sources that might have information to capture these dimensions. Table 1 presents the results of our data discovery, inventory, and acquisition steps. Six software repositories (SourceForge, OpenHub, Depsy, GitHub, OSalt, and Stack OverFlow) are described with respect to their potential use to inform our research. The acquisition component of Table 1 summarizes the complexity of acquiring the discovered data sources. SourceForge, OpenHub, and Depsy were initially chosen for our study based on the variables of interest.

Table 2. Categories, numbers of projects developed, and subcategories of OSS projects on SourceForge

Category	Projects	Top subcategories
Development	80,134	www/http (26%), software development (8%)
Games	25,117	Games/entertainment (23%), role playing (13%)
System administration	21,218	Networking (19%), storage (15%)
Science and engineering	18,007	Bioinformatics (14%), artificial intelligence (12%)
Communications	17,302	Chat (32%), email (21%)
Business and enterprise	13,536	Enterprise (27%), financial (24%)
Audio and video	8,254	Sound/audio (68%), video (27%)
Home and education	7,639	Education (39%), computer-aided instruction (15%)
Graphics	6,797	3D rendering (19%), 3D modeling (12%)
Security and utilities	4,975	Security (59%), cryptography (35%)
Other/nonlisted Topic	4,811	No subcategories given
Multimedia	708	No subcategories given

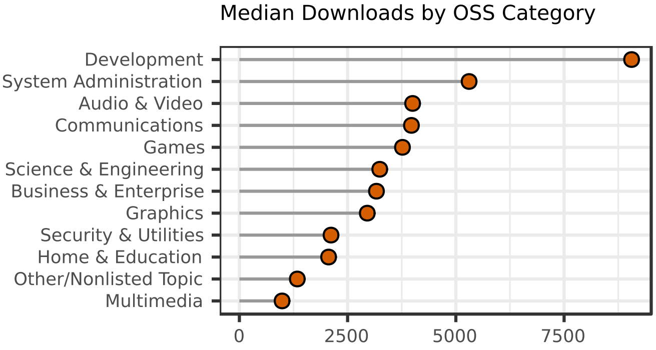

Fig. 4. The median number of downloads of projects on SourceForge by category.

Analysis. This step starts with data profiling and assessing the quality of the data followed by cleaning and linking various datasets. This allows us to map the dimensions into variables that we can obtain and measure using the data sources. The mappings from variables to the dimensions could be listed as the following.

Stock measures: Active completed projects, downloads
Flow measures: New projects (annual and cumulative), lines of code, commits (submitted edits to the code), and man-hours
Categories: Type of software (purpose of the package)
Sectors and collaborators: Contributors' sectors (business, government, academic, nonprofit, individual, foreign)
Users: Downloads, citations, other developers (reverse dependencies)

Exploratory analysis. This initial exploration of the data helps us understand what is feasible in terms of measurement of OSS creation and use.

SourceForge data were used to explore the scope of OSS, as it includes information about the categories (types) and the downloads of the OSS projects. We collected information about 449,274 projects and 50,000 contributors over the span of 18 y (1999–2017). SourceForge uses categories and subcategories to characterize the OSS projects. This allows us to analyze the type (purpose) of the projects and measure their respective demands using downloads. Table 2 presents the total number of OSS projects and top subcategories for various categories on SourceForge. The number of projects developed (supply side) in the development category (80,134 projects) is significantly higher than those in the remaining categories. This category is followed by games, system administration, science and engineering, and communications. When the download volume (demand side) is analyzed, we observe a few projects with high downloads [such as Microsoft's TrueType core fonts (2,079 million downloads), eMule (685 million), and Apache (238 million)], and many projects with a low number. Fig. 4 illustrates the median download volumes by category. SourceForge as a data source contains valuable information for eventually modeling the value, scope, and impact of OSS.

OpenHub data are used to delve into the development of OSS projects, as it includes information about the lines of code, devoted time, the number of commits, the contributors, and the languages that they are written in. OpenHub includes development and activity information on 676,523 projects, 291,782 users and contributors, and their activity (e.g., commits). We have information about 698 organizations and their portfolio of 2,850 projects. We used their application programming interface (API) to collect information on two random sets of projects: (*i*) a subset

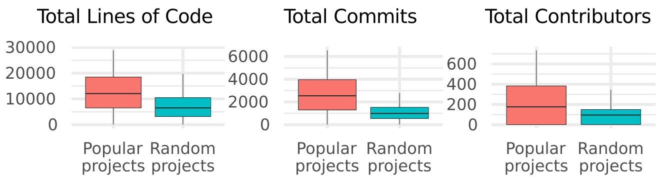

Fig. 5. The distribution of total lines of code, commits, and contributors for two random sets of projects on OpenHub: (*i*) a subset of the most popular projects based on number of users (55,217 projects) and (*ii*) a subset of all projects (46,005 projects).

of the most popular projects based on number of users (55,217 projects) and (*ii*) a subset of all projects (46,005 projects). For each of these projects, we collected the total lines of code, total number of commits, and total number of contributors. Fig. 5 illustrates the distributions of these variables for the two subsets of projects. OpenHub data show the feasibility of capturing significant information that could be used in models to assess the value of OSS. For example, these variables (e.g., total lines of code and number of commits) could be used to estimate the effort and the cost of production, as they capture the inputs to development (man-hours and the number of workers), as is done in constructive cost models for proprietary software (37).

Network analysis. As mentioned before, collaboration and sharing are essential features of OSS development.

Methods to measure the value and cost of these projects need to take into account the interactions (i.e., collaborations between contributors and the dependencies between packages). Network analysis allows us to incorporate the structural properties of these interactions in models that estimate the impact and cost of OSS.

Depsy creates a very rich source of data, including information about contributors, commits, downloads, citations that can be used to study the development activities, pattern of complex interactions among community members, and package dependencies to evaluate OSS projects (38).

We use Depsy to collect information on R packages. Some of the top downloaded and top cited packages are presented in Table 3. The average number of downloads is 45,775, and the median is 8,508; the average number of citations is five, and the median is zero. Citations are used to measure the impact of patents and publications and are frequently used to estimate the research outputs of universities and government institutions. However, it is not a common practice to cite the software packages that are used in published studies (39); hence, it is problematic to use citations to evaluate OSS projects (40, 41). We find that most R packages (74%) do not have any citations in Depsy, although they have a large number of downloads and reverse dependencies.

Depsy provides information on the packages that are required for the development of others (i.e., dependencies or reuses). We have generated the network of R packages where a directed edge $i \rightarrow j$ indicates that the package j requires i to be installed to function. We obtain a network with 7,547 nodes and 20,641 directed edges. The average degree (indegree and outdegree) of the dependency network is 2.74. This means that, on average, R packages depend on (and are used by) 2.74 other packages. Table 3 illustrates the top packages with the highest outdegree (number of reuses). MASS, one of the packages in the standard library of R used for statistical analysis, has the highest outdegree of 955. MASS is followed by ggplot2, a widely used package for visualization. Matrix is used for matrix operations, and plyr is a main package for data wrangling and exploration. The outdegree is an important factor to take into account when developing measures of impact, as the packages build on these software packages (they need to be compatible). Around 70% of the packages are not used by other packages. However, the maximum number of dependencies that a package has (i.e., indegree) is 27, with a median of 2. These should be taken into account when estimating the cost of the package, as these are necessary inputs for their development.

Table 3. Top downloaded and cited R packages in Depsy and the packages with the highest outdegree (i.e., number of reuses)

Top downloaded packages	No. of downloads	Top cited packages	No. of citations	Highest outdegree	Outdegree
Rcpp	6,683,565	vegan	4,275	MASS	955
ggplot2	6,255,500	lme4	4,023	ggplot2	737
stringr	5,366,703	nlme	2,916	Matrix	514
plyr	5,345,308	ggplot2	1,702	plyr	447

Network analysis also allows us to identify groups of contributors or projects that form communities based on structural features of the networks. These communities could involve individuals that develop similar projects or use similar tools/languages, that collaborate with common authors, that apply similar scientific methods (e.g., spatial analysis), or that belong to similar organizations. Evaluation of academic articles takes into account different publication conventions of different scientific fields, such as the number and order of authors. Similarly, the OSS communities could be characterized for a better evaluation of the projects and authors.

We identify communities in the dependency network of R packages using the modularity algorithm (42) implemented

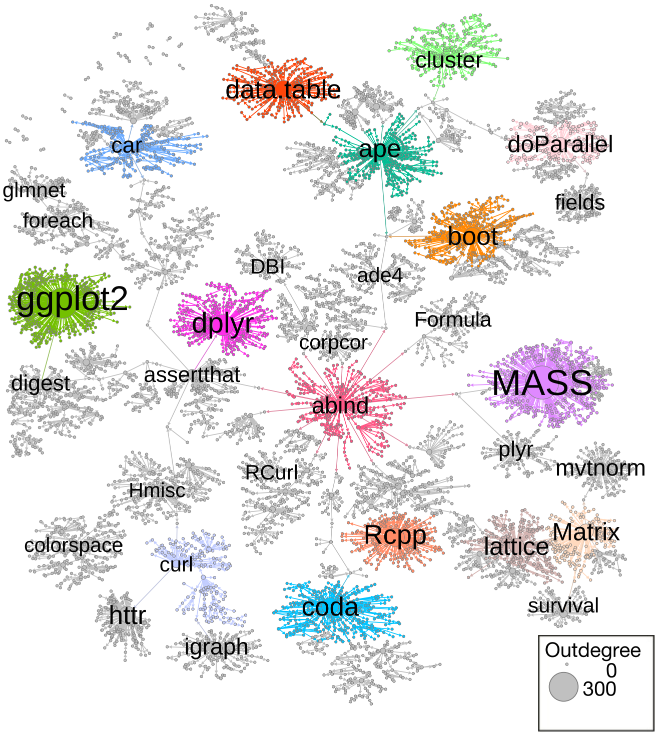

Fig. 6. Reduced dependency network of R packages in Depsy. The full network includes 7,547 nodes and 20,641 directed edges. MST-Pathfinder Network Scaling algorithm (44) was used for edge reduction. The reduced network has 7,491 edges. The size of the node is proportional to the outdegree (number of reuses) in the reduced network, and the different colors indicate communities identified using modularity (42). Communities that make up less than 2% of the nodes are in gray.

in Gephi (43). For illustrative purposes, we used the MST-Pathfinder Network Scaling algorithm (44) for edge reduction. Fig. 6 illustrates the reduced network (44) with 7,491 edges, where the size of the node is proportional to the outdegree (number of reuses) in the reduced network and the different colors indicate communities that make up at least 2% of the nodes (the remaining ones are gray). The communities correspond to different uses of R packages, such as statistical analysis (MASS), visualization (ggplot2), and data wrangling (dplyr). Methods to estimate the value of packages need to take into account the use/purpose and the centrality of the packages in the dependency network.

Next steps. Our next focus is to estimate the cost of OSS development based on the number of hours, lines of code, and other features of OSS. Wages of software developers or the price of private software will be used for the estimation of cost and value. 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. We may never be able to measure the value of all OSS, but even a baseline will show that its value is substantial.

Conclusion

Most intangible inputs are considered assets, because they are used repeatedly in production. Not valuing these intangibles misses changes in firms and the economy, and it leads to underestimation of productivity, misallocation of resources, and inaccurate financial statements.

Two case studies are presented that show promising approaches to create indicators based on intangible investments: (*i*) the creation of a new intangible (an organizational process innovation in a Fortune 500 manufacturing company) and (*ii*) the characterization of an unpriced intangible input to innovation (OSS). These case studies show that the value of intangibles can be measured, thus providing the details of business innovation that survey and other methods do not capture. Our methods provide an opportunity to extend existing measures beyond those already used by federal statistical agencies, shining additional light into the shadows of dark innovation.

ACKNOWLEDGMENTS. We thank the Data Science for the Public Good Program graduate students Claire Kelling, Sayali Phadke, and Daniel Chen for their contributions. We acknowledge our undergraduate researchers Alex Gagliano, John Higgins, Romcholo Macatula, Eirik Iversen, and Ben Swartz. This work was supported by US Department of Agriculture Grant 58-3AEU-7-0074.

1. Hall BH, Jaffe AB (2018) Measuring science, technology, and innovation: A review. *Annals of Science and Technology Policy*, 10.1561/110.00000005.
2. Aizcorbe M, Moylan CE, Robbins CA (2009) BEA briefing: Toward better measurement of innovation and intangibles. *Survey of Current Business* (Bureau of Economic Analysis, Department of Commerce, Washington, DC).
3. Damanpour F (1991) Organizational innovation: A meta-analysis of effects of determinants and moderators. *Acad Manage J* 34:555–590.
4. von Hippel E (2016) *Free Innovation* (MIT Press, Cambridge, MA).
5. Organization for Economic Cooperation and Development (OECD) (2010) Handbook on deriving capital measures of intellectual property products. *OECD*. Available at www.oecd.org/sdd/na/44312350.pdf. Accessed September 12, 2018.
6. Organization for Economic Cooperation and Development (OECD)/Eurostat (2005) *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data* (OECD, Paris), 3rd ed.
7. Schumpeter JA (1934) *The Theory of Economic Development—An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle* (Harvard Univ Press, Cambridge, MA).
8. Kindlon A, Jankowski J (2017) *Rates of Innovation Among U.S. Businesses Stay Steady: Data from the 2014 Business R&D and Innovation Survey* (NSF, Arlington, VA), pp 17–321.
9. Eurostat (2017) Innovation statistics. Eurostat statistics explained. Available at ec.europa.eu/eurostat/statistics-explained/index.php?title=Innovation_statistics. Accessed September 12, 2018.
10. Stone A, Rose S, Lal B, Shipp S (2008) *Measuring Innovation and Intangibles: A Business Perspective* (Institute for Defense Analysis, Science and Technology Policy Institute, Washington, DC).
11. Martin BR (2016) Twenty challenges for innovation studies. *Sci Public Pol* 43:432–450.
12. Gault F (2018) Defining and measuring innovation in all sectors of the economy. *Res Policy* 47:617–622.
13. Bockstael NE, McConnell KE (1983) Welfare measurement in the household production framework. *Am Econ Rev* 73:806–814.
14. National Science Board (2018) *Science and Engineering Indicators 2018* (NSF, Alexandria, VA), NSB-2018-1.
15. Blair M, Wallman S (2001) *Unseen Wealth* (Brookings Institution, Washington, DC).
16. Haskel J, Westlake S (2017) *Capitalism Without Capital: The Rise of the Intangible Economy* (Princeton Univ Press, Princeton).
17. Bureau of Economic Analysis (BEA) (2017) NIPA national income and product accounts. Available at https://www.bea.gov/resources/methodologies/nipa-handbook. Accessed September 12, 2018.
18. Robbins CA (2016) Using new growth theory to sharpen the focus on people and places in innovation measurement. *Blue Sky Forum, Informing Science and Innovation Policies, Towards the Next Generation of Data and Indicators*. Available at www.oecd.org/sti/124%20-%20Focusing_on_People_and_Places_Robbins.pdf. Accessed September 19, 2018.
19. Corrado C, Hulten C, Sichel D (2005) *Measuring Capital and Technology: An Expanded Framework in Measuring Capital in the New Economy* (Univ of Chicago Press, Chicago), pp 11–46.
20. Corrado C, Hulten C, Sichel D (2009) Intangible capital and US economic growth. *Rev Income Wealth* 55:661–685.
21. Lev B (2000) *Intangibles: Management, Measurement, and Reporting* (Brookings Institution, Washington, DC).
22. Rassier D (2014) Treatment of research and development in economic accounts and in business accounts. Available at https://apps.bea.gov/scb/toc/0314cont.htm. Accessed September 19, 2018.
23. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: Traps in big data analysis. *Science* 343:1203–1205.
24. Keller S, Lancaster V, Shipp S (2017) Building capacity for data-driven governance: Creating a new foundation for democracy. *Stat Public Pol* 4:1–11.
25. Keller S, et al. (2018) Harnessing the power of data to support community-based research. *WIREs Comput Stat* 10:e1426.
26. Bureau of Economic Analysis (BEA) (1994) A satellite account for research and development. Available at https://apps.bea.gov/scb/pdf/national/niparel/1994/1194od.pdf. Accessed September 12, 2018.
27. Keller S, et al. (2016) Leveraging external data sources to enhance official statistics and products. Report prepared for US Census Bureau. Available at cdn.vbi.vt.edu/mc/SDAL/leveraging-external-data-sdal-2016.pdf. Accessed September 12, 2018.
28. 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. *Annu Rev Stat Its Appl* 4:85–108.
29. Procter & Gamble (2017) Annual report. Available at https://us.pg.com/annualreport2017/annual-report.html#/Productivity. Accessed.
30. Pires B, et al. (2017) A Bayesian simulation approach for supply chain synchronization. *Proceedings of the 2017 Winter Simulation Conference*, 10.1109/WSC.2017.8247898.
31. Bureau of Economic Analysis (BEA) (2017) Relation of private fixed investment in intellectual property products (by type) in the fixed assets accounts to the corresponding items in the national income and product accounts. Available at https://www.bea.gov/resources/methodologies/nipa-handbook. Accessed September 12, 2018.
32. Bureau of Economic Analysis (BEA) (2018) Fixed asset tables, 7.5.b. Investment in government fixed assets. Available at https://apps.bea.gov/iTable/index_FA.cfm. Accessed September 12, 2018.
33. Moylan C (2001) *Estimation of Software in the US National Accounts: New Developments in STN/NA (2001)* (OECD, Paris), Vol 25, pp 9–12.
34. Netcraft (2018) Web server survey. Available at https://news.netcraft.com/archives/category/web-server-survey/. Accessed February 1, 2018.
35. Greenstein S, Nagle F (2014) Digital dark matter and the economic contribution of Apache. *Res Pol* 43:623–631.
36. Linux Foundation (2016) Linux kernel development report. Available at https://www.linuxfoundation.org/events/linux-kernel-development-2016/. Accessed September 12, 2018.
37. Boehm BW (1981) *Software Engineering Economics* (Prentice-Hall, Englewood Cliffs, NJ), Vol 197.
38. Zhao R, Wei M (2017) Impact evaluation of open source software: An altmetrics perspective. *Scientometrics* 110:1017–1033.
39. Howison J, Bullard J (2016) Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *J Assoc Inf Sci Tech* 67:2137–2155.
40. Howison J, Deelman E, McLennan MJ, Ferreira da Silva R, Herbsleb JD (2015) Understanding the scientific software ecosystem and its impact: Current and future measures. *Res Eval* 24:454–470.
41. Singh Chawla D (2016) The unsung heroes of scientific software. *Nat News* 529:115–116.
42. Lambiotte R, Delvenne JC, Barahona M (2008) Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770.
43. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. Available at https://gephi.org/publications/gephi-bastian-feb09.pdf. Accessed September 12, 2018.
44. Schvaneveldt RW, Dearholt D, Durso F (1988) Graph theoretic foundations of Pathfinder networks. *Comp Math Applic* 15:337–345.