April 19, 2022

José Bayoán Santiago Calderón
1000 New Jersey Ave SE Apt 816
Washington, DC 20003-3368
Jose.Santiago-Calderon@bea.gov

The Editors
Review of Economics and Statistics
Harvard Kennedy School
79 JFK St # 142
Cambridge, MA 02138-5801
restat@hks.harvard.edu

Dear Selection Committee,

I desire to be considered for the NBER Summer Institute fellowship invitation and support to attend the second week (July 18-22). I serve as a research economist at the U.S. Bureau of Economic Analysis, mainly studying the digital economy and intellectual property products. I received my Ph.D. in economics from Claremont Graduate University in 2019 and held postdoctoral research associate appointments at the University of Virginia for two years. During my tenure at UVA's Biocomplexity Institute, I worked with federal, state, and local government partners on projects related to science policy (National Center for Science and Engineering Statistics), computational modeling (Defense Advanced Research Projects Agency), and program evaluation (Arlington County Police Department). My dissertation focused on econometrics and computational economics, which I have leveraged throughout my professional career in industry, academia, and public service.

My current research agenda includes projects for valuing the U.S. data economy and measuring the scope and impact of open-source software. Conference on Research in Income and Wealth (CRIW) events have served as outlets for both projects, most recently, my presentation for valuing data during this spring conference. I also presented my work on open-source software at the last ASSA conference. Some sessions I am interested in attending are CRIW, Innovation, IT and Digitization, and Science of Science Funding.

Valuing intangibles assets, especially those without market transactions such as own-account or non-rival cross-sectors produced, poses challenges that usually require an excellent understanding of the production process and related labor market to apply sums-of-costs approaches and assess productivity implications. Examples include using non-traditional data such as mining software repositories or applying data-driven techniques to estimate time-use factors for relevant activities using online job advertisements. I am also interested in measuring indicators and developing pertinent statistics for science policy. Intangible assets are an ever-growing component of the production process that powers modern economies. Having relevant statistics is vital to public economics (e.g., investment, policy, regulation).

The NBER SI presents a rare opportunity where a significant part of the community gathers in a remarkable conducive forum. This exposure, particularly at this stage in my career, would be highly beneficial. As an added benefit, the in-person component would allow me to connect with many with whom I have only interacted virtually.

Cordially,
José Bayoán Santiago Calderón

José Bayoán Santiago Calderón

Jose.Santiago-Calderon@bea.gov Nosferican jbsc.netlify.app
0000-0002-8406-6175

Education

2019	PhD Economics	Claremont Graduate University
2015	MA Economics	Claremont Graduate University
2014	BA Economics	Southwestern University

Appointments and Employment

Research Economist

Analysis & Research Group
National Economic Accounts Directorate — 2021 – Present
U.S. Bureau of Economic Analysis

Working on topics related to the digital economy and measuring intangible assets (e.g., software, data).

Statistics Consultant

Pumas-AI — 2018 – Present

Work on the development of components for the pharmaceutical modeling and simulation platform and in the consulting services branch for statistical analysis.

Postdoctoral Research Associate

Social & Decision Analytics Division
Biocomplexity Institute & Initiative — 2019 – 2021
University of Virginia

Worked on multiple projects with government partners including:

- National Center for Science and Engineering Statistics (NCSES)
  - Measuring the Scope and Impact of Open Source Software
  - Skilled Technical Workforce
- Defense Advanced Research Projects Agency (DARPA)
  - Computational Simulation of Online Social Behavior (SocialSim) – Summary
- Arlington County Police Department (ACPD)
  - Evaluation of the Arlington Restaurant Initiative

Assisted the infrastructure team on helping the team best use UVA computing resources (e.g., high-performance computing) and best practices (e.g., version control). Served as project lead and instructor for the Data Science for the Public Good Young Scholars Program (DSPG).

Data Scientist

Residential Energy and Water Intelligence (Res-Intel) 2016 – 2018

Conducted a program evaluation of the California Advanced Home Program (CAHP) commissioned by Southern California Edison and developed the module for residential benchmarking.

Dissertation

José Bayoán Santiago Calderón. 2019. "On Cluster Robust Models." PhD diss., Claremont Graduate University. DOI: 10.5642/cguetd/132

Selected Peer-Reviewed Publications

Behnaz Moradi-Jamei, Brandon Lee Kramer, José Bayoán Santiago Calderón, and Gizem Korkmaz. 2021. "Community Formation and Detection on GitHub Collaboration Networks." In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* 244–251. ASONAM '21. Virtual Event, Netherlands: Association for Computing Machinery. ISBN: 9781450391283. DOI: 10.1145/3487351.3488278

Mengchen Sam Yong, Lavínia Paganini, Huilian Sophie Qiu, and José Bayoán Santiago Calderón. 2021. "The Diversity-Innovation Paradox in Open-Source Software." In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. DOI: 10.1109/MSR52588.2021.00089

José Bayoán Santiago Calderón. 2020. "Econometrics.jl." *Proceedings of the JuliaCon Conferences* 1 (1): 38. DOI: 10.21105/jcon.00038

Selected Presentations

José Bayoán Santiago Calderón and Dylan G. Rassier. 2022a. *Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings.* 37th International Association for Research on Income and Wealth (IARIW) General Conference. Forthcoming. Luxembourg, August 23, 2022

José Bayoán Santiago Calderón and Dylan G. Rassier. 2022b. *Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings.* National Bureau of Economic Research (NBER) Conference on Research in Income and Wealth (CRIW) on Technology, Productivity, and Economic Growth (manuscript, slides, and presentation). Washington, DC, March 17, 2022

José Bayoán Santiago Calderón, Gizem Korkmaz, Brandon Lee Kramer, and Carol Ann Robbins. 2022. *Measuring the Cost of Open Source Software Innovation on GitHub.* Allied Social Science Associations (ASSA) General Conference (manuscript). Remote, January 8, 2022

Valuing the U.S. Data Economy Using Machine Learning and Online Job Postings

José Bayoán Santiago Calderón and Dylan G. Rassier

U.S. Bureau of Economic Analysis*

April 19, 2022

Abstract

With the recent proliferation of data collection and uses in the digital economy, the understanding and statistical treatment of data stocks and flows is of interest among compilers and users of national economic accounts. In this paper, we measure the value of own-account data stocks and flows for the U.S. business sector by summing the production costs of data-related activities implicit in occupations. Our method augments the traditional sum-of-costs methodology for measuring other own-account intellectual property products in national economic accounts by proxying occupation-level time-use factors using a machine learning model and the text of online job advertisements (Blackburn 2021). In our experimental estimates, we find that annual current-dollar investment in own-account data assets for the U.S. business sector grew from $72 billion in 2003 to $131 billion in 2020, with an average annual growth rate of 3.6 percent. Cumulative current-dollar investment for the period 2003–2020 was $1.8 trillion. In addition to annual current-dollar investment, we present historical-cost net stocks, real growth rates, and effects on value-added by industrial sector.

Keywords: Data Assets, System of National Accounts, Digital Economy, Intangible Capital, Measurement

JEL Codes: E22, O3, O51

*Author e-mail addresses: Jose.Santiago-Calderon@bea.gov and Dylan.Rassier@bea.gov. The views expressed in this paper are those of the authors and do not necessarily represent the U.S. Bureau of Economic Analysis (BEA) or the U.S. Department of Commerce. We thank Christopher Blackburn, former research economist at BEA, for developing the machine learning approach we use in the paper (Blackburn 2021). We also thank Marshall Reinsdorf, Jon Samuels, Brian Sliker, Brenda Bugge, and John Haltiwanger and other participants at the NBER-CRIW Pre-conference on Technology, Productivity, and Economic Growth for early comments.

1 Introduction

With the recent proliferation of data collection and uses in the digital economy, the understanding and statistical treatment of data stocks and flows is of interest among compilers and users of national economic accounts. During the last revision of the *System of National Accounts* (*SNA*) (United Nations 2010), which is the international standard for national economic accounts, the treatment of data stocks and flows was a topic of discourse among statistical agencies and international organizations, who ultimately settled on the treatment of *databases* as a subcategory of software in capital formation (Ahmad 2004; Ahmad 2005). Current global efforts for the next revision of the *SNA*, expected to be published in 2025, have a renewed focus on the valuation and recording of the information content of databases – i.e., the embedded *data* – in response to the presumed rapid increase in data stocks and flows over the last decade and longer.

The value of data is implied in the profits and market values of some firms. In 2020, two of the largest global data firms – Alphabet Inc. (Google) and Meta Platforms Inc. (Facebook) – had a combined net income before tax of $81.3 billion, which amounted to 3.7 percent of U.S. corporate profits before tax. In November 2021, the combined market capitalization of the two firms was $2.9 trillion, which amounted to 7.5 percent of the market capitalization of all S&P 500 firms.[1] While the value of data may be implied in these measures, data stocks and flows are not visible in national economic accounts under the current *SNA* treatment. Moreover, understanding and measuring the value of data presents challenges to economic statisticians.

In this paper, we measure the value of own-account data stocks and flows for the U.S. business sector by summing the production costs of data-related activities implicit in occupations. Production costs include labor costs, capital costs, and intermediate consumption. To estimate production costs, we apply a markup to an estimate of the wage bill for data-related

[1]Net income before tax comes from each firm's 10-K filings with the U.S. Securities and Exchange Commission for year-end December 31, 2020. Corporate profits come from BEA's National Income and Product Accounts table 6.17D. Market capitalizations come from YCHARTS (GOOG, FB) as of November 19, 2021.

activities, which is consistent with the Bureau of Economic Analysis (BEA) methodology for own-account software. Our method augments the traditional sum-of-costs methodology for measuring other own-account intellectual property products (IPPs) in national economic accounts by proxying occupation-level time-use factors using a machine learning model and the text of online job advertisements (Blackburn 2021).

The occupation-level time-use factors can be decomposed into two components: 1) the fraction of jobs in an occupation engaged in qualifying activities based on data-relevant skills revealed in the job advertisement and 2) the average share of time allocated to the data-relevant activities. Using online job advertisements from Burning Glass Technologies (BGT), skills in the BGT taxonomy that are relevant to data-related activities are identified, including data entry, storage, analysis, and management. The fraction of jobs in an occupation engaged in qualifying activities is given by the fraction of BGT job advertisements that contain at least one of the data-relevant skills. The average time allocation is based on the distance of an occupation to known data-intensive occupations that serve as "landmark" occupations (e.g., data entry keyers or statisticians). The landmark occupations are selected based on the top 20 occupations by highest rates of job openings listing data-relevant skills. A doc2vec model is then trained on the job advertisement text for each occupation to obtain a numerical representation of what the occupation-level job postings convey. Using the numerical representation, occupation-level pair-wise distances are obtained to measure how "close" or similar an occupation is to the landmark occupations. The product of the similarity to a landmark occupation and the ratio of job openings with identified data-relevant skills serves as the proxy for the occupation-level time-use factor. We then apply the time-use factors to the product of average annual wages and annual number of employees by occupation at the 3-digit North American Industry Classification System (NAICS) level available from the U.S. Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics (OEWS) to calculate the wage bill (i.e., wages and salaries, excluding employee benefits) for data-related activities.

The main challenges for measurement of own-account data stocks and flows that we address in the paper are similar to challenges imposed by other own-account IPPs that are already included in capital formation in the *SNA* and the U.S. National Income and Product Accounts (NIPAs). First, the scope of capital formation is not yet well-defined for own-account data. Second, own-account data and other own-account IPPs are at risk of multiple counting from sources such as overlap among categories of IPPs and non-rival use of data. A third challenge is what proportions of the sum-of-costs should be accounted for by labor costs, capital costs, and intermediate consumption in light of the role that capital services play in the collection, storage, analysis, and management of data. Finally, own-account data are not transacted in active markets, which means there are no observed transactions that are useful for measuring prices and depreciation. We discuss each of these challenges and our approach to addressing them in section 4 of the paper.

Our experimental results indicate that annual current-dollar investment in own-account data assets for the U.S. business sector grew from $72 billion in 2003 to $131 billion in 2020, which yields an average annual growth rate of 3.6 percent. Cumulative current-dollar investment for the period 2003–2020 was $1.8 trillion. Annual current-dollar investment in own-account data for the period averaged 0.8 percent as a share of business sector value-added, 3.8 percent as a share of investment in private fixed assets, and 13.6 percent as a share of investment in IPPs. Likewise, the historical-cost net stock of data assets grew from $174 billion in 2003 to $313 billion in 2020, which yields an average annual growth rate of 3.4 percent. The average annual growth rate in real data investment over the period was 3.8 percent, which yields an average annual increase in real business sector value-added growth of only 1 basis point. In contrast, growth in real data investment is lower than growth in real investment in IPPs and software, which yields a decline in average annual growth of real investment in IPPs and software of 23 basis points and 116 basis points, respectively. For NAICS sectors, the largest dollar investments were made in Finance and Insurance (NAICS 52) and Manufacturing (NAICS 31-33). The largest increase in average real value-added

growth by NAICS sector shows up for Management of Companies (NAICS 55).

The next section summarizes the *SNA* background on the treatment of own-account data stocks and flows and related literature. Section 3 provides details on our measurement, including source data and methodologies. Section 4 identifies challenges associated with measurement of own-account IPPs and how we address those challenges for own-account data. Section 5 reports our core experimental results and some additional experimental results of interest. Section 6 concludes.

2 National Accounts Background and Related Literature

2.1 *System of National Accounts*

The 1993 version of the *SNA* includes only a brief paragraph on the inclusion of "large databases that the enterprise expects to use in production over a period of time of more than one year" as part of the computer software category of capital formation (United Nations 1993, paragraph 10.93). There is no mention of embedded information content (i.e., data) in *SNA 1993*. Leading up to the 2008 version of the *SNA*, an *SNA* group of statistical agencies and international organizations considered the inclusion of embedded data in capital formation. To guide the discussions, Ahmad (2004) outlined two components of databases – supporting software and data stored in the database – and summarized practical challenges that countries encounter while trying to implement the vague *SNA 1993* recommendation. In light of the challenges, Ahmad (2005) described two definitions for databases considered by the *SNA* group. One definition included the value of the information content to be stored in databases as long as the information had a useful life of more than one year, and one definition did not include the value of the information content. The group recommended that the latter definition is preferable because the former definition would "open the door

to the capitalization of knowledge" (Ahmad 2005, p. 2). Based on the summary outlined in Ahmad (2005), the group primarily considered databases maintained by statistical agencies.

The recommendation that was ultimately written into *SNA 2008* includes databases and computer software as separate categories of intellectual property products in capital formation (United Nations 2010, paragraphs 10.109–10.114). If a database is developed for own use, *SNA 2008* recommends a sum-of-costs approach to value the database. The sum-of-costs includes the cost of preparing data in a format that conforms to the database but excludes the cost of acquiring or producing the data. In addition, the sum-of-costs excludes the value of the database management system (DBMS), which is included instead with computer software. If a database is developed for sale or for license, the value should be determined by the market price, which includes the value of the information content. Thus, *SNA 2008* recommends an inconsistent treatment for data in capital formation depending whether a database is developed for own use or for sale or license. In addition, the value of data acquired or produced for inclusion in databases is not to be treated as intermediate consumption in the sum-of-costs approach for own-account databases, which is inconsistent with the inclusion of intermediate consumption in the usual sum-of-costs measurement. The overall conclusion drawn by the *SNA* group was that if data is an asset, it is a non-produced asset whose value should be limited in national accounts to measures of purchased goodwill. Thus, there should be no value of data reflected in production measures.

2.2 U.S. National Accounts

The U.S. national accounts are consistent with the *SNA* recommendations on intellectual property products, including computer software (U.S. Bureau of Economic Analysis 2020). Similar to general practice in other countries, the U.S. accounts do not include a separation between software and databases (i.e., the software that houses data) in published capital stock and flow measures. The value of any data included in purchased software is included in measures of investment and capital stock. The value of any data in own-account software

is excluded from measures of investment and capital stock.

BEA estimates three categories of software: 1) prepackaged, 2) custom, and 3) own-account. Benchmark estimates of prepackaged and custom software are determined using a commodity flow method based on receipts reported in the U.S. Economic Census for Software Publishers (NAICS 5112), Data Processing, Hosting and Related Services (NAICS 5182), and Computer Systems Design and Related Services (NAICS 5415). For non-benchmark years, estimates are based on receipts reported in the Census Bureau's Service Annual Survey. Benchmark estimates of own-account software are determined using a sum-of-costs methodology based on wage data in the BLS OEWS for four occupations - Computer Programmers, Computer System Analysts, Software Developers (applications), Software Developers (system software) - and based on the Economic Census. For non-benchmark years, estimates are primarily based on the OEWS data. In addition to labor costs and intermediate consumption, own-account software includes a cost for capital services (Chute et al. 2018).

2.3 Related Literature

Two strands of literature provide additional context for this paper. The first is literature on the value of data, which not only suggests that data has value but also suggests that the value of data may, at least in part, be a result of a production process rather than non-produced, which is the current perspective in national economic accounts. The second is emerging literature from statistical agencies, international organizations, and other sources that is beginning to recognize the likelihood that some data are produced assets and to generate renewed focus for recording and valuing data stocks and flows.

2.3.1 Value of Data

Varian (2018) describes a data pyramid that is a variation of the data-information-knowledge-wisdom (DIKW) hierarchy introduced by Ackoff (1989) and subsequently used in information science and economics (Rowley 2007; Boisot and Canals 2004). The data pyramid is used to

illustrate the relationships among data that is stored as bits, information that is stored in documents, and knowledge that is stored in the human mind (Mokyr 2013). Related to the data pyramid is the data value chain presented in OECD (2013) and subsequently expanded in Moro Visconti et al. (2017). The data value chain illustrates a production process for data from an unstructured form that has very little value to a structured form that can be leveraged in a business model or other usage (Bakhshi et al. 2014). The stages of the chain include collection, storage, processing, distribution, and usage. The Moro Visconti et al. (2017) version of the data value chain focuses on business users at the last stage of the chain with the monetization of data via a business model, which may be data-dependent or data-neutral. Data-dependent business models, such as online platforms, rely heavily on data for sources of revenue and profits (Li et al. 2019; Nguyen and Paczos 2020). Data-neutral firms do not depend on data for revenue but can still realize benefits from data that help improve existing products or offer new products. The OECD (2013) version of the data value chain includes business users and also includes government, non-profit, and household users.

Hughes-Cromwick and Coronado (2019) outline the value of U.S. government data to business decision-making. Likewise, the value of household data is evident in literature on the economics of personal privacy, which has re-emerged as an area of interest as summarized in Acquisti et al. (2016).

Farboodi and Veldkamp (2021) construct a growth model of the data economy in which data is an information asset that contributes to growth by reducing uncertainty and helping firms choose better production techniques via forecasts. The model demonstrates short-run increasing returns due to a feedback loop within the firms. However, long-run diminishing returns result in the absence of traditional technological progress in the Solow (1956) growth model because better forecasts are not a tool that can logically sustain long-run growth. In contrast to Farboodi and Veldkamp (2021), Jones and Tonetti (2020) model data contributing directly to productivity and generating long-run growth.

2.3.2 Recording and Valuing Data Stocks and Flows

Under the second strand of literature, statistical agencies and international organizations are renewing efforts to introduce guidelines for recording and valuing data as a produced asset in national economic accounts. Ahmad and van de Ven (2018) and Reinsdorf and Ribarsky (2020) provide background on the historical treatment of data stocks and flows and offer preliminary thoughts on moving forward with changes to the *SNA*. Rassier et al. (2019) also summarize considerations for treatment of data stocks and flows and present cursory estimates of data-related flows based on official statistical sources for the U.S. economy. Statistics Canada (2019b) carefully defines and categorizes data, databases, and data science and presents experimental estimates for each category using a sum-of-costs measurement methodology for the Canadian economy. The Australian Bureau of Statistics and the Netherlands Central Bureau of Statistics have compiled measures of data stocks and flows based on the Canadian methodology. Each of these previous efforts have resulted in preliminary guidance that is currently being considered by the *SNA* community (United Nations Inter-Secretariat Working Group on National Accounts 2022).

Goodridge et al. (2021) are the first to provide a harmonized set of cross-country estimates for data assets in European Union countries and then estimate the contribution of data capital deepening to growth in productivity. They find that about 43 percent of employment engaged in capital formation of software and data is unaccounted for in measured own-account software and databases, and the missing piece of capital formation is growing faster than the measured piece. Goodridge et al. (2021) also provide a summary of previous economic literature on information and knowledge.

3 Data and Methods

The *SNA* recommendation for valuing IPPs that lack an observable market transaction is a cost-based approach including labor costs, capital costs, and intermediate consumption.

For market producers, capital costs include a net return to fixed assets (or "normal profit") in addition to consumption of fixed capital. Our strategy builds on BEA's sum-of-costs methodology for own-account software. Estimates of aggregate production costs for data-related output have the general form

$$C_{i,t} = \alpha \sum \tau_{\omega} W_{\omega,i,t} H_{\omega,i,t} \tag{1}$$

where for each occupation ω, industry i, and year t, we calculate the wage bill by multiplying the annual number of employees ($H_{\omega,i,t}$) by the average annual wage ($W_{\omega,i,t}$) and an occupation-specific time-use factor (τ_{ω}) that reflects the time-effort that the occupation allocates to data-related activities. The parameter α is a markup that reflects employee benefits (not included in the wage bill), capital costs, and intermediate consumption, which yields the full production costs. The rest of this section discusses our estimation of each of the elements in equation 1.

3.1 Employment and Wage Data

The U.S. Bureau of Labor Statistics (BLS) has 12 survey programs that provide information on pay and benefits. We prioritize having wage data at an occupational level specific enough to capture the activities of interest.[2] The BLS Occupational Employment and Wage Statistics (OEWS) program produces employment and wage estimates annually for around 800 occupations and is well-suited for our purpose. In addition, the historical data temporal coverage dates back to the early 2000s, allowing us to generate a longer time series.

Occupational data collected by the U.S. federal statistical system is generally collected, calculated, and disseminated based on the Standard Occupational Classification (SOC) system. Employees are assigned to an occupation based on the work they perform and not on their education or training. The OEWS data are an SOC-based occupational system that

[2] `https://beta.bls.gov/comparison-matrix`

uses SOC codes to assign occupations at levels for which the data are published. Some SOC codes are aggregated into a single OEWS occupational code for reporting purposes. The OEWS system allows the data to be linked to other systems such as the U.S. Department of Labor (DOL) Occupational Information Network (O*NET) (Hopson 2021).

The OEWS program reports data using the North American Industry Classification System (NAICS). We use estimates at the NAICS 3-digit subsector levels that are representative of privately-owned business establishments, which excludes government and nonprofit institutions serving households (NPISH).[3]

To generate data-related employment and wage estimates for 2003–2020, we use model-based estimates (Dey et al. 2019) for 2015–2019 and official estimates for the period 2003–2014 and for 2020 (U.S. Bureau of Labor Statistics 2021). We use estimates of the average annual wage and annual number of employees for privately-owned establishments at the NAICS 3-digit level.

The wage series use several versions of the underlying occupational and industry systems across time. In order to obtain a consistent time series, we use the OES 2019 hybrid structure crosswalk and rely on the OES 2010–2011 classification crosswalk to account for the temporary codes in those two years. For example, if multiple occupations are aggregated in the latest taxonomy (OES 2019), the estimated average annual wage is an average of available wage data weighted by the employment estimates for the corresponding occupations. In cases when an occupation is attributed to multiple 2019 codes, which may occur due to special hybrid-code estimates, those are equally distributed among the corresponding 2019 codes. The resulting series provide us the employment and wage numbers by year and industry.

[3]The sectors and subsectors identified as NPISH include Educational Services (NAICS 61), Health Care and Social Assistance (NAICS 62), Arts, Entertainment and Recreation (NAICS 71), and Religious, Grantmaking, Civic, Professional, and Similar Organizations (NAICS 813).

3.2 Time-Use Factors

Time-use factors are important to the sum-of-costs methodology as they provide measures of time-effort allocated to the activities of interest such as data-related activities. Time-use factors have been used to examine quality of service, job satisfaction, and other outcomes in various domains such as education, health, and security. For example, the OECD Teaching and Learning International Survey (TALIS) collects data on the time-use allocation of teachers among categories: (1) Administrative and leadership tasks and meetings, (2) Curriculum and teaching-related tasks and meetings, (3) Student interactions, (4) Parent or guardian interactions, and (5) Interactions with local and regional community, business and industry. In the area of health, studies have collected information on the share of time physicians spend on direct care with patients or administrative work (Woolhandler and Himmelstein 2014). While most data on time-use allocations is collected in surveys, other methods include direct observation of the work performed such as rides with police officers (Parks et al. 1999).

For data as an asset, we estimate time-use factors using a methodology developed in Blackburn (2021) for data-relevant skills, which are reflected in activities including entry, storage, analysis, and management of data. Examples of specific activities include data cleaning, data wrangling, data manipulation, and data science.

Following Blackburn (2021), the time-use factor τ can be decomposed as follows

$$\tau_\omega = \frac{l_\omega}{L_\omega} s^*_\omega = \rho_\omega s^*_\omega \tag{2}$$

where the time-use factor for occupation ω is the product of the fraction of employees that engage in activities of interest (ρ_ω) and an estimate of how much time the occupation allocates to the activities (s^*_ω). Without time-use factors, the sum-of-costs methodology relies on identifying specific occupations assumed to best embody the activities of interest.

3.2.1 Online Job Advertisements

The method uses online job advertisement data from Burning Glass Technologies (BGT) to estimate time-use factors for a broad range of occupations.[4] The data not only contain the job advertisement text but also enhancements, including deduplication, identified skills, degree requirements, location, and information on the employer (e.g., industry). BGT also uses their own-developed occupational auto-coders to identify the occupation of the job advertisement.[5]

Lancaster et al. (2021) look at the data quality, suitability, and representativeness of BGT job advertisement data for research purposes. The study confirms several findings in the literature such as the over-representation of certain occupation groups in the BGT job advertisement data. One aspect of the BGT data that distinguishes them from other alternatives is the focus on skills rather than occupations as the main unit of analysis for understanding the job market (Burning Glass Technologies 2019). The BGT skills taxonomy includes over 17,000 skills as well as various skill properties such as whether a skill refers to data or software (or both).

The strategy uses the BGT job advertisement data to estimate equation 2. The first component refers to the fraction of employees in an occupation who engage in data-related activities. The concept is operationalized with the following definition:

$$\hat{\rho_{\omega}} = \frac{\sum_{j=1}^{L_{\omega}} \mathbb{1}\left(\hat{y_j}\right)}{L_{\omega}} \tag{3}$$

where $\mathbb{1}\left(\hat{y_j}\right)$ denotes $\exists \;\; s : s \in S$ such that s is a subset of skills identified for job advertisement y_j, and S is the set of skills identified as data relevant. In other words, a ratio is computed as the fraction of job advertisements with at least one of the data-relevant skills out of all job advertisements (L_{ω}) for each occupation. Blackburn (2021) manually identified 203

[4]Online job advertisements have been used in applications related to labor, education, and credential research. Examples of online job advertisement providers include Burning Glass Technologies, Indeed, and the National Labor Exchange (NLx) Research Hub. These providers collect job postings in "real-time" from various websites.

[5]The NLx Research Hub uses the O*NET SOC Code AutoCoder™ (`https://www.onetsocautocoder.com`).

Table 1: BGT skills identified as data relevant

3D Seismic Data	Data Dictionary System
Accenture Data Governance Framework	Data Documentation
Advanced Data Entry	Data Encryption
Assessment Data	Data Engineering
Big Data	Data Entry
Big Data Analytics	Data Entry Prioritization
Billing Data Analysis	Data Evaluation
Biological Database Search	Data Exploitation
Business Intelligence Data Modeling	DFHSM
Cascading Big Data Applications	Data Flow Diagrams (DFDs)
Climate Data Analysis	Data Governance
Clinical Data Abstracting	Data Integration
Clinical Data Analysis	Data Integrity
Clinical Data Exchange	Data Lakes / Reservoirs
CDISC	Data Loss Prevention
Clinical Data Management	Data Management
Clinical Data Review	Data Management Platform (DMP)
Clinical Data Understanding	Data Manipulation
Clinical Database Development	Data Mapping
Clinical Research Data Accuracy and Integrity	Data Migration
Cloud Security Data Protection And Privacy	Data Mining
Columnar Databases	Data Mining Industry Knowledge
Conceptual Data Models	Data Modeling
Customer Data Integration	Data Modeling Star / Snowflake Schema
Customer Service Database	Data Multiplex System (DMS)
Data Acquisition	Data Munging
Data Acquisition Systems	Data Operations
Data Analysis	Data Platform as a Service
Data and Safety Monitoring Board	Data Pre-Processing
Data Architecture	Data Privacy
Data Archiving	Data Protection Industry Knowledge
Data Buffers	Data Protection Planning
Data Capture	Data Protection Strategy
Data Center Hardware	Data Quality
Data Cleaning	Data Quality Assessment
Data Collection	Data Reports
Data Communications	Data Science
Data compression	Data Security
Data Conversion	Data Security Classification

Note: Top 80 skills by frequency out of 203 data relevant skills identified manually (Blackburn 2021).

Table 2: Landmark occupations

O*NET SOC 2010	Description	Time-use factor
43-9021.00	Data Entry Keyers	0.94
15-1111.00	Computer and Information Research Scientists	0.77
15-1141.00	Database Administrators	0.75
15-1199.06	Database Architects	0.72
19-1029.01	Bioinformatics Scientists	0.68
19-4061.00	Social Science Research Assistants	0.67
15-2041.00	Statisticians	0.66
15-1199.07	Data Warehousing Specialists	0.63
15-2041.01	Biostatisticians	0.63
15-1199.08	Business Intelligence Analysts	0.61
53-7073.00	Wellhead Pumpers	0.60
19-3022.00	Survey Researchers	0.59
43-9111.01	Bioinformatics Technicians	0.58
43-9111.00	Statistical Assistants	0.54
29-2092.00	Hearing Aid Specialists	0.54
15-2041.02	Clinical Data Managers	0.54
43-3021.01	Statement Clerks	0.50
51-8099.02	Methane/Landfill Gas Generation System Tech.	0.47
15-1199.05	Geographic Information Systems Technicians	0.44
33-3021.06	Intelligence Analysts	0.43

Note: For landmark occupations, the similarity to the nearest landmark is one, and thus the time-use factor $\hat{\tau}_\omega$ is the same as $\hat{\rho}_\omega$.

BGT skills that are data relevant, which excludes any skill deemed a software skill as a way to minimize potential overlap with capital formation in software. The top 80 data-relevant skills by frequency are presented in table 1.

A two-step approach is used to estimate (s^*_ω). The first step identifies the occupations with the highest rate of employees engaged in data-related activities (ρ_ω) based on the BGT skills. The top 20 occupations are denoted "landmark" occupations and assigned full time-effort. Table 2 shows the occupations determined to be "landmark" occupations. Many of the occupations are obvious landmark candidates such as *data entry keyers* and various research and analyst occupations. Some of the occupations are less obvious as landmarks. For example, *wellhead pumpers* may not be an occupation immediately associated with data, but their job activities rely on monitoring and assessing data in order to act on that information

promptly.

For non-landmark occupations, s^*_ω is estimated as the cosine similarity to the closest landmark occupation. Mathematically,

$$\hat{s^*_\omega} = \max_{w \in \mathbb{M}} \left\{ \frac{\mathbf{A}_\omega \cdot \mathbf{A}_w}{\|\mathbf{A}_\omega\| \|\mathbf{A}_w\|} \right\} \tag{4}$$

where $\mathbb{M}$ is the set of landmark occupations, A_ω (i.e., omega) is a numerical representation of occupation ω, and A_w (i.e., double-u) is the mathematical representation of occupation w (i.e., one of the landmark occupations). The effective time-use factor then becomes

$$\hat{\tau_\omega} = \hat{\rho_\omega} \hat{s^*_\omega} = \frac{\sum_{j=1}^{L_\omega} \mathbb{1}\left(\hat{y_j}\right)}{L_\omega} \max_{w \in \mathbb{M}} \left\{ \frac{\hat{\mathbf{A}}_\omega \cdot \hat{\mathbf{A}}_w}{\|\hat{\mathbf{A}}_\omega\| \|\hat{\mathbf{A}}_w\|} \right\}. \tag{5}$$

The estimate for the full production costs of data-related activities for a given industry i at time t is then as follows:

$$\hat{C}_{i,t} = \alpha \sum_{\omega \in \Omega} \left[\frac{\sum_{j=1}^{L_\omega} \mathbb{1}\left(\hat{y_j}\right)}{L_\omega} \left(\max_{w \in \mathbb{M}} \left\{ \frac{\hat{\mathbf{A}_\omega} \cdot \hat{\mathbf{A}}_w}{\|\hat{\mathbf{A}}_\omega\| \|\hat{\mathbf{A}}_w\|} \right\} \right) \hat{W}_{\omega,i,t} \hat{H}_{\omega,i,t} \right]. \tag{6}$$

3.2.2 Sample Design and Description

The BGT data enable observation of over 239 million job advertisements in the United States for 2010–2019. The data include over a thousand O*NET SOC 2010 occupations. The data are used to train an auto-coder model for O*NET SOC 2010 using job text to obtain the numerical representation for each occupation such that we can compute pair-wise similarity between occupations.

The sample design uses the occupation distribution across NAICS 3-digit subsectors based on the annual number of employees in privately-owned establishment from OEWS for 2015–2019. If observations available at the occupation-industry level are below the target (1,500 × percentage of occupations in a NAICS 3-digit subsector), all observations are included. Otherwise, we sample from the valid observations a targeted sample of 1,500 obser-

vations equally distributed based on the sequential order of the job posting dates. We only include occupations for which at least 100 job ads are available, meaning there is complete information for the job text, job posting date (ISO 8601), O*NET SOC 2010, and NAICS subsector. The sample design aims to be representative of the business sector in terms of the occupation-industry interactions and composition as well as temporal changes, while also maintaining desirable properties for modeling, such as avoiding extreme unbalanced classes or making inferences based on limited data. Our goal is different from traditional auto-coders, which may emphasize some metric like accuracy and may prefer to focus on having better performance for frequent occupations over having acceptable performance for less frequent occupations.

Considering the observations with complete information (O*NET SOC 2010 and NAICS 3-digit codes are not always available), we sample occupations for which at least 100 observations are available and limit the observations included per occupation to around 1,500. We also exclude military occupations (SOC major group 55) and those that are exclusive to the public sector like legislators and postal service employees. The sample includes 959 O*NET SOC 2010 occupational codes that have between 101 and 1,548 observations.

The final sample comprises 1.12 million observations for which we recover the job advertisement text as well as the skills for each of the job advertisements. We split the dataset by applying a clustered sampling design based on occupation-industry subsector with 90 percent of the sample being used as the training set and 10 percent as the test set.

3.2.3 Machine Learning Model

The modeling approach uses the doc2vec (Le and Mikolov 2014) implementation in gensim v4.0.1 (Řehůřek and Sojka 2010) for Python (van Rossum and Drake 2009) v3.8.5. The modeling technique uses paragraph vectors to capture the semantics of documents as an alternative to "bag-of-words" or "bag-of-n-grams" representations. In our case, each job advertisement is represented not just by applying a tokenization to the text but by also

including indicators for each job advertisement. After training the model, each occupation is represented by a multidimensional numerical abstraction. Once the model has been fitted, we extract the features that are a 1,000-dimensional numerical representation of each O*NET SOC 2010 code. The pair-wise cosine distance is then computed to obtain the occupation similarity matrix. The similarity matrix allows us to find the most similar "landmark" occupation.

3.2.4 Validation

Model Validation. Occupation auto-coders are automated models that can predict occupation codes based on inputs such as job advertisements and job titles. One common autocoder regarded as a gold standard is the O*NET-SOC AutoCoder™ developed by R.M. Wilson Consulting, Inc. for the U.S. Department of Labor. There are various ways to evaluate the performance of a multi-class classifier. For example, the classifier accuracy claims for the O*NET-SOC AutoCoder™ are stated based on the inputs whether it operates on a job advertisement and job title or on just a job title. For job titles and job advertisement text, the model accuracy for predicting O*NET SOC 2019 codes is 85 percent based on their internal test results. Since the model is using just the unstructured job advertisement text, we expect the performance to be lower than the gold standard using additional information. Another consideration is that the labeled data are also based on the BGT auto-coder for O*NET SOC 2010 codes which in turn contains errors from the BGT auto-coder. Given the licensing terms and industry practices, properly assessing the performance of these tools is difficult. For example, a measure of accuracy can be significantly driven by unbalanced classes. Performing well for common occupations in a sample will result in better accuracy even if the performance for many occupations is not as good if those occupations are less common. For the purpose of obtaining a good numerical representation for each occupation, it is more valuable to have good performance for many occupations rather than great performance for a select few. One metric we compute is the unweighted average F1 score on our

test set (out-of-sample) of around 120,000 observations for which the model predictions yield an estimate of 0.5 on a 959-dimensional classification problem.

Validation of Landmark Occupations. An alternative strategy for obtaining representative occupations and similarity between occupations is to use the O*NET program. The O*NET tasks file provides estimates of how frequently certain occupations perform job work activities. For example, for the occupation *data entry keyers* (43–9021.00) in the O*NET v26 data, there is a task relevant to data-related activities: "Compile, sort, and verify the accuracy of data before it is entered". For this occupational task, the O*NET data provide a breakdown of how frequently the task is performed: *Yearly or less* (1.49), *More than yearly* (0), *More than monthly* (2.67), *More than weekly* (41.69), *Daily* (18.6), *Several times daily* (18.6), *Hourly or more* (35.54). Each task can then be assigned an indicator of whether it is considered data-relevant or not and estimate a time-use factor for each occupation. Some of the heuristics we employ include assigning the frequency for each task per occupation to the category with the highest value (daily, several times daily, or hourly). At each of the different frequency categories, we compute the fraction of tasks deemed data-relevant. We then collapse the ratios assigning relative values of 10 percent, 20 percent, and 70 percent to daily, several times daily, and hourly, respectively. This back-of-the-envelope heuristic is used as an additional check on our rankings. To assess whether the landmark occupations are reasonable, we compute the Spearman's rank correlation coefficient and obtain a value around 0.22. The correlation coefficient reveals a negligible correlation, but the machine learning model seems to provide more reasonable rankings than the O*NET-based model.

3.2.5 Mapping Occupational Codes

Using equation 5, we obtain O*NET SOC 2010 level estimates for the time-use factors. However, employment and wage data are collected and prepared with OEWS Employment Projections (EP) 2019 codes. We use the crosswalk files to map the O*NET SOC 2010 codes to the SOC 2010 codes and then to OEWS 2019 codes (see subsection 3.1 for additional

information on the process and data files). If multiple occupations are mapped to a single OEWS-EP 2019 code, the time-use factor for that occupation is the mean value for all mapped occupations. We then obtain time-use factor estimates for 736 OEWS 2019 occupation codes. These estimates tend to be quite low with a mean lower than 10 percent and a 95 percent quantile of 8 percent. The larger values are almost exclusive to the landmark occupations shown in table 2.

3.3 Markup and Additional Adjustments

Markup. We estimate the full production costs by multiplying the data-related wage bill by a markup (i.e., the α in equations 1 and 6) that is designed to reflect employee benefits, capital costs, and intermediate consumption, none of which are included in the wage bill. Observing publicly available information, firms engaged in data aggregation, data sales, and other data-related market activities appear generally to be classified to industries in Information (NAICS 51), Professional, Scientific and Technical (PST) Services (NAICS 54), or Administrative & Support Services (NAICS 56).[6] Based on these observations, we use NAICS 51, NAICS 54, and NAICS 56 as the representative sectors for the markup in the full production cost estimates. In particular, we calculate α as a composite ratio of compensation, intermediate consumption (exluding materials), consumption of fixed capital (CFC), and net operating surplus to wages and salaries using data summed for select industries from BEA's annual industry accounts. In other words, the composite ratio is weighted by the size of the chosen industries.

We consider calculating the composite ratio using BEA industry accounts data at the NAICS 2-digit sector level - i.e., all industries in NAICS 51, 54, and 56 - or at a more detailed subsector level of industries. The sector level yields a ratio that reflects a broader

[6]For example, Axciom, Facebook, Localeze, and Burning Glass Technologies appear to be classified to NAICS 51 industries. In addition, Alphabet, Foursquare, Infogroup, Nielsen, Automatic Data Processing Inc. (ADP), The NPD Group, and International Data Corporation (IDC) all appear to be classified to NAICS 54 industries. Exceptions are the credit bureaus, which are classified to Administrative & Support Services (NAICS 56) industries.

representation of industries and may be driven by the production functions of products that are not data-related. The latter feature assumes the production function of data assets resembles the production functions of primary products for the representative industries. The subsector level allows us to choose industries that are more explicitly engaged in the production of data-related products and, thus, yields a ratio that more closely reflects the production functions of data-related products. The industries we choose from BEA's industry accounts include Data Processing, Internet Publishing and Other Information Services (NAICS 518), Computer Systems Design and Related Services (NAICS 5415), and Administrative & Support Services (NAICS 561).

At the sector level, the annual ratios range from 2.91 to 3.11 for 2003–2020, with a simple average of 3.03. At the subsector level, the annual ratios range from 2.22 to 2.61, with a simple average of 2.42. For our core experimental results, we use the subsector average of 2.42 as a time-invariant lower bound markup for each year 2003–2020. In addition, we present results using the sector average of 3.03 as a time-invariant upper bound markup for each year.

Additional Adjustments. We do not have an estimate of own-account data output that is used up in current production versus own-account data output that is used in capital formation. Thus, we reduce the full production cost estimate by 50 percent to capture capital formation.

We also adjust for potential overlap between data assets and R&D assets by reducing the capital-formation-adjusted production cost estimates by 50 percent for R&D intensive industries (Galindo-Rueda and Verger (2016)), including industries in Chemical Manufacturing (NAICS 325), Computer and Electronic Product Manufacturing (NAICS 334), Transportation Equipment Manufacturing (NAICS 336), Publishing (NAICS 511), and PST Services (NAICS 541). Likewise, we adjust for overlap between own-account data and purchased data by further reducing the adjusted production cost estimate for PST Services (NAICS 541) by 50 percent.

Table 3: Effective factors applied to the wage bill

NAICS	Markup	Capital formation	R&D	Purchased data	Effective factor
325	2.42	0.50	0.50	N/A	0.605
334	2.42	0.50	0.50	N/A	0.605
336	2.42	0.50	0.50	N/A	0.605
511	2.42	0.50	0.50	N/A	0.605
541	2.42	0.50	0.50	0.50	0.3025
All other	2.42	0.50	N/A	N/A	1.21

Note: The table summarizes by NAICS subsector the effective factors that are applied to the wage bill to calculate the full production costs. *All other* excludes NPISH and general government: Educational Services (NAICS 61), Health Care and Social Assistance (NAICS 62), Arts, Entertainment and Recreation (NAICS 71), Religious, Grantmaking, Civic, Professional, and Similar Organizations (NAICS 813), and Public Administration (NAICS 92).

Finally, we completely exclude from the data wage bill the wages of the four occupations that BEA uses to estimate own-account software–Computer Programmers, Computer Systems Analysts, Software Developers (applications), and Software Developers (system software). Since we also exclude software skills in our estimates of time-use factors, no additional adjustment is required for potential overlap between data assets and software assets.

We choose 50 percent as a placeholder for each of the additional adjustments to reflect our acknowledgment of the adjustments until future empirical evidence comes available. Table 3 summarizes by NAICS subsector the effective factors that are applied to the wage bill to calculate investment in data assets for our core experimental results. Section 4 has more discussion on the markup and additional adjustments.

4 Addressing Measurement Challenges

The main challenges for measurement of own-account data stocks and flows that we address in the paper are similar to challenges imposed by other own-account IPPs that are already included in capital formation in the *SNA* and the U.S. NIPAs. We identify here four challenges and how we address the challenges for own-account data.

4.1 Scope of Costs in Capital Formation

The scope of costs to include in capital formation is important for own-account data and other own-account IPPs.[7] The stages of the data value chain (OECD 2013; Moro Visconti et al. 2017) yield insights into activities that underlie the production process for data – including collection, storage, processing, distribution, and usage – some of which are reflective of costs that embody capital formation. The scope that has been recently introduced by the OECD (2021) is whether capital formation should include both recording and processing costs and the costs of procuring access to data or only include recording and processing costs. Procurement costs may include either explicit purchases or the value of "free" digital products exchanged for access.[8] OECD (2021) considers recording and processing costs to be production activities that could result in capital formation. However, they call into question whether procurement costs should be considered production activities or simply costs to improve the quantity or quality of access for a more valuable data asset, and thus, outside the production boundary.

We address this challenge by identifying skills in the BGT taxonomy that are relevant to data-related activities, including data entry, storage, analysis, and management. These skills are embodied by the occupations that underlie our estimates of the data wage bill. In addition, we reduce the full production cost estimate for data by 50 percent to account for data-related output that is used up in the current period and never becomes capital formation.[9]

[7]For example, OECD (2010) identifies eight stages of activities in the production of own-account software: feasibility analysis, functional analysis, detailed analysis, programming, tests, documentation, training, and maintenance. Of the eight stages, only costs for functional analysis, detailed analysis, programming, tests, and documentation are within scope for capital formation.

[8]The latter procurement costs are identified in Farboodi and Veldkamp (2021) as a measurement challenge, and their model moves beyond price-based valuation by assigning a value to goods and data that have a zero transaction price.

[9]Another approach would be to use a depreciation rate for data that reflects a more rapid decline in value of investment flows as a result of obsolescence. However, such an approach would disregard the accounting difference between intermediate consumption and consumption of fixed capital.

4.2 Multiple Counting

Own-account data and other own-account IPPs are at risk of multiple counting from several sources.

4.2.1 Overlap Among Categories of IPPs

One source of multiple counting is overlap among categories of IPPs that are separately measured, such as data, software, and research and development (R&D).

Software. We address overlap between data and software by estimating the time-use factors for data, which are intentionally based on unique data-relevant skills that are distinguishable from software-relevant skills. In addition, we completely exclude from the data wage bill the wages of the four occupations that BEA uses to estimate own-account software–Computer Programmers, Computer Systems Analysts, Software Developers (applications), and Software Developers (system software).[10] For comparison with data-relevant skills presented in table 1, table 4 presents the top 80 BGT skills identified as software relevant. Likewise, table 5 shows the top time-use factors for occupations engaged in software-related activities to compare with landmark occupations for data reported in table 2.

While tables 1 and 4 demonstrate no overlap for BGT skills by design, there is overlap of seven occupations with high time-use factors in tables 2 and 5. The occupations that show up in both tables include Business Intelligence Analysts, Computer and Information Research Scientists, Database Administrators, Database Architects, Data Warehousing Specialists, Geographic Information Systems Technicians, and Statisticians. In addition, the data and software time-use factors sum to more than 100 percent for the first five occupations, which implies the occupations and their time-use factors may be used to measure either data- or software-related activity but not both. However, there is no overlap between landmark

[10]The time-use factors for these four occupations are relatively modest - each less than 20 percent - and we think the occupations may reasonably embody some data-related activities that accompany own-account software development but are not included in BEA's measures of own-account software. However, we choose to completely exclude the four occupations for a more conservative estimate.

occupations for data and the four occupations BEA currently uses to estimate own-account software.

Overlap is expected when activities are intertwined. One example is software and R&D. Based on the 2018 Business Enterprise Research and Development Survey (BERD) table 19, $165.6 billion (37.5 percent) of domestic R&D was software products and embedded software R&D.[11] In the NIPAs, this overlap is included with R&D investment and excluded from software investment. There may also be job activities at the intersection of data, software, and R&D, such as software programming to prepare a data set for research. Observing independently-estimated time-use factors that together exceed 100 percent is not problematic in itself for our estimates because we know that own-account software estimation is limited to four specific occupations with relatively low time-use factors. However, the outcome does suggest an incentive to jointly estimate own-account data and own-account software to ensure consistency and prevent double counting if the scope of own-account software is expanded to include more occupations in the future.

Research & Development. In contrast to software, the BGT skills dataset does not include R&D-relevant skills to distinguish them from data-relevant skills, which may yield some overlap between data assets and R&D assets. For example, Goodridge et al. (2021) and Statistics Canada (2019a) both argue that data science activities meet the *SNA* and OECD (2015) definition of R&D. However, Goodridge et al. (2021) suggest that data science is not included in measured R&D in practice by some countries because R&D is generally measured based on surveys designed for known performers of traditional scientific R&D activities. Likewise, Statistics Canada (2019a) suggests that the survey they use to measure

[11] According to guidance in the BERD survey, R&D activity in software includes 1) software development or improvement activities that expand scientific or technological knowledge and 2) construction of new theories and algorithms in the field of computer science. Likewise, the BERD survey requires that R&D activity in software exclude 1) software development that does not depend on scientific or technological advance, such as supporting or adapting existing systems, adding functionality to existing application programs, and routine debugging of existing systems and software, 2) creation of new software based on known methods and applications, 3) conversion or translation of existing software and software languages, and 4) adaptation of a product to a specific client, unless knowledge that significantly improved the base program was added in that process.

Table 4: BGT skills identified as software relevant

.NET	Microsoft Sql Server Integration Services (SSIS)
Active Server Pages (ASP)	Microsoft Visio
Agile Development	Middleware
AJAX	MySQL
Amazon Web Services (AWS)	Object-Oriented Analysis and Design (OOAD)
AngularJS	Oracle
Apache Hadoop	Oracle PL/SQL
Apache Tomcat	PERL Scripting Language
Apache Webserver	Platform as a Service (PaaS)
Application Design	Python
ASP.NET	Relational DataBase Management System (RDBMS)
Atlassian JIRA	Ruby
C++	Salesforce
COBOL	SAP
Computer Engineering	SAS
Crystal Reports	Scrum
Debugging	Shell Scripting
Eclipse	Software Architecture
Enterprise Resource Planning (ERP)	Software as a Service (SaaS)
Extensible Markup Language (XML)	Software Development
Extensible Stylesheet Language XSL	Software Engineering
Firmware	Software Testing
Git	SQL
HTML5	SQL Server
Hypertext Preprocessor (PHP)	SQL Server Reporting Services (SSRS)
IBM WEBSPHERE	Systems Analysis
Informatica	Systems Development Life Cycle (SDLC)
Java	Teradata DBA
Java Server Pages (JSP)	Transact-SQL
JavaScript	Unified Modeling Language (UML)
JavaScript Object Notation (JSON)	UNIX
jQuery	UNIX Shell
JUnit	User Acceptance Testing (UAT)
Linux	User Interface (UI) Design
Microsoft Access	Visual Basic
Microsoft Azure	Visual Studio
Microsoft C#	VMware
Microsoft Operating Systems	Waterfall
Microsoft Project	WebLogic
Microsoft SQL	Windows Server

Note: Top 80 software skills by frequency identified from landmark occupations.

Table 5: Top time use factors for software

O*NET SOC 2010	Description	Time use factor
15-1133.00	**Software Developers, Systems Software**	**0.95**
15-1132.00	**Software Developers, Applications**	**0.94**
15-1131.00	**Computer Programmers**	**0.85**
15-1121.00	**Computer Systems Analysts**	**0.82**
15-1134.00	Web Developers	0.65
15-1199.02	Computer Systems Engineers/Architects	0.60
15-1199.08	Business Intelligence Analysts	0.55
15-1199.06	Database Architects	0.55
15-1141.00	Database Administrators	0.52
15-1199.01	Software Quality Assurance Engineers and Testers	0.52
15-1199.09	Information Technology Project Managers	0.50
15-1143.00	Computer Network Architects	0.49
13-1111.00	Management Analysts	0.48
15-1199.07	Data Warehousing Specialists	0.46
15-1111.00	Computer and Information Research Scientists	0.45
15-1142.00	Network and Computer Systems Administrators	0.41
15-1199.00	Computer Occupations, All Other	0.40
15-1122.00	Information Security Analysts	0.40
11-3021.00	Computer and Information Systems Managers	0.39
17-2061.00	Computer Hardware Engineers	0.38
15-1199.04	Geospatial Information Scientists and Techs	0.38
15-1199.03	Web Administrators	0.37
15-1152.00	Computer Network Support Specialists	0.37
15-1151.00	Computer User Support Specialists	0.36
15-2041.00	Statisticians	0.34
13-2099.01	Financial Quantitative Analysts	0.32
15-1199.11	Video Game Designers	0.32
15-1143.01	Telecommunications Engineering Specialists	0.32
17-3029.07	Mechanical Engineering Technologists	0.30
17-1011.00	Architects, Except Landscape and Naval	0.30
13-2051.00	Financial Analysts	0.28
15-1199.10	Search Marketing Strategists	0.28
13-1081.02	Logistics Analysts	0.28
17-2199.08	Robotics Engineers	0.27
27-1021.00	Commercial and Industrial Designers	0.27
27-3042.00	Technical Writers	0.27
17-3031.02	Mapping Technicians	0.27
15-1199.05	Geographic Information Systems Technicians	0.27

Note: Bold occupations denote landmark occupations that are by design the same occupations BEA currently uses to estimate own-account software.

R&D needs to be examined for updates because the survey was developed years ago and is biased toward the selection of firms engaged in more traditional forms of R&D activities, such as pharmaceutical firms, and biased away from a growing number of firms in diverse industries engaged in data science activities.

For R&D measures in the U.S. national accounts, the survey used to collect information on performance of business R&D is the BERD survey.[12] The 2019 BERD survey explicitly includes software development or improvement activities that expand scientific or technological knowledge and construction of new theories and algorithms in the field of computer science. In addition, the 2019 BERD survey asks respondents to report the percentage of domestic R&D expenditures paid for and performed by them that was for artificial intelligence (AI), which includes speech recognition, machine vision, machine learning, text analytics, and natural language generation and processing. Thus, data science may not be absent in practice from the BERD survey and may be included in BEA's measures of R&D based on the BERD survey.

To adjust for potential overlap between data and measured R&D in our estimates, we reduce the capital-formation-adjusted production cost estimate by 50 percent for R&D intensive industries (Galindo-Rueda and Verger 2016), including Chemical Manufacturing (NAICS 325), Computer and Electronic Product Manufacturing (NAICS 334), Transportation Equipment Manufacturing (NAICS 336), Publishing Industries (NAICS 511), and Professional, Scientific, and Technical Services (NAICS 541).[13] Consistent with the suggestions of Goodridge

[12]The 2019 BERD survey defines research and development as follows: "*Research* is defined as experimental or theoretical work undertaken primarily to acquire new knowledge or understanding of phenomena and observable facts. Research may be either "basic", where the goal is primarily to acquire new knowledge or understanding of a given topic without a specific commercial application in mind, or "applied", where the goal is to solve a specific problem or meet a specific commercial objective. *Development* is defined as the systematic use of research and practical experience to produce new or improved goods, services, or processes. In simple terms, the intended output of research is ideas and the intended output of development is products."

[13]As an alternative, we also consider completely excluding from the data wage bill the wages of four SOC 2010 occupations that appear to be R&D occupations based on "research" in the descriptions: Computer and Information Research Scientists (SOC 15–1111), Operations Research Analysts (SOC 15–2031), Survey Researchers (SOC 19–3022), and Social Science Research Assistance (SOC 19–4061). However, we think the occupations may reasonably embody some data-related activities that accompany R&D, and we are not able to determine if excluding these four occupations based on "research" is as comprehensive as adjusting the R&D intensive industries.

et al. (2021) and Statistics Canada (2019b), the resulting dollar adjustments for these industries are modest.[14]

4.2.2 Non-Rival Use of Data

Another source of multiple counting is overlap between own-account data and purchased data, which is also a source of multiple counting for software and R&D. We address this challenge by further reducing the adjusted production cost estimate by 50 percent for PST Services (NAICS 541), which we determine to be the NAICS code used by most data aggregators and brokers and is the NAICS code with the highest data wage bill (prior to adjustment) under our methodology.[15]

Finally, a source of multiple counting that has been cited as a concern specifically for data as an asset is that the same piece of information can be used in multiple databases. However, a single piece of information is transformed when it is added to a record with other pieces of information. Likewise, we are not valuing pieces of information but rather the activities associated with making the information available in a record, which is unique to every firm for own-account measures.

4.3 Measuring Capital Costs

A challenge that may be more relevant (but not unique) to data stocks and flows is how to measure the capital cost component of the sum-of-costs. There is a question of what proportions of the sum-of-costs should be accounted for by labor costs, capital costs, and intermediate consumption, given the role of capital services in the collection, storage, analysis, and management of data.

For our core experimental results, we apply a conservative markup of 2.42 to the wage bill as explained in section 3.3. In addition, we apply an alternative markup of 3.03 to

[14]The resulting dollar adjustment for 2020 is $29.4 billion or approximately 5.75 percent of business R&D investment in the U.S. national accounts.

[15]The reduction for PST Services (NAICS 541) to adjust for non-rival use of data is in addition to the reduction of 50 percent to adjust for R&D overlap.

Table 6: Weighted composite ratios for full sum-of-costs

	Subsector (lower bound)		Sector (upper bound)	
	Ratio	Share (%)	Ratio	Share (%)
Compensation	1.17	48	1.17	38
Intermediate consumption	0.78	32	1.07	35
Consumption of fixed capital	0.18	8	0.29	10
Net operating surplus	0.29	12	0.50	17
Markup	2.42		3.03	

Note: All data are from BEA's annual industry accounts. Intermediate consumption excludes materials. The lower bound column reports the simple average for 2003-2020 of each annual measure summed for NAICS 518, 5415, and 561 divided by annual wages and salaries summed for the same subsectors. The upper bound column reports the simple average for 2003–2020 of each annual measure summed for NAICS 51, 54, and 56 divided by annual wages and salaries summed for the same sectors.

the wage bill, which we treat as an upper bound result. Since the markups are weighted composite ratios of compensation, intermediate consumption (excluding materials), CFC, and net operating surplus, we can decompose the markups into the proportionate share of each component in the sum-of-costs. The decomposition is presented in table 6.

Aside from the overall numeric difference between the two markups - i.e., 2.42 and 3.03 - table 6 demonstrates two important differences between the two markups that result from the NAICS levels chosen for each. First, the ratios for labor costs (i.e., compensation) are the same between the two markups (i.e., 1.17). Second, the ratios for capital costs (i.e., net operating surplus plus CFC) and intermediate consumption are both higher for the upper bound markup, which yields higher proportionate shares of capital costs and intermediate consumption and a lower proportionate share of labor costs for the upper bound markup. A gap of 10 percentage points exists in the proportionate shares of labor costs between the lower- and upper-bound markups, and the gap is offset by 7 percentage points attributable to capital costs and 3 percentage points attributable to intermediate consumption.

4.4 Prices and Depreciation

Own-account data are not transacted in active markets, which means there are no observed transactions that are useful for measuring prices and depreciation. While measuring prices and depreciation for own-account data is beyond the scope of this paper, we utilize international guidelines and U.S. practice for deflating and depreciating own-account software and databases as a starting point for deflating and depreciating own-account data, which reflects the inclusion of own-account data as a subcategory of software and is consistent with preliminary guidance emerging from the *SNA* community's current work on revising the *SNA*.

Prices. In the absence of deflators for own-account software, OECD (2010) recommends that deflators for custom software be used as a proxy until own-account software deflators are developed. In addition, OECD (2010) suggests three options for own-account database deflators. One option is to use a price index of a related activity for which there is a price index of reasonable quality. The other two options are an input cost index, one with zero productivity growth and another with an adjustment based on productivity growth of a similar industry.

BEA does not compile a price index for own-account databases because own-account databases are not measured separately from own-account software in the U.S. NIPAs. For own-account software, BEA currently compiles a price index from a weighted average of the prepackaged software price and a BEA input cost index (U.S. Bureau of Economic Analysis 2020). The input cost index is compiled from BLS data on wage rates for Computer Programmers and Computer Systems Analysts and the intermediate consumption associated with the production of software. The input cost index reflects an explicit adjustment for changes in productivity that is based primarily on a BLS multifactor productivity index.[16] To address the lack of prices for own-account data, we propose using BEA's current price

[16]When BEA first introduced capital measures of own-account software into the NIPAs in 1999, an input cost index was compiled from a weighted average of compensation rates for Computer Programmers and Computer System Analysts and the intermediate consumption associated with their work (Moulton and Sullivan 1999). No productivity adjustment was made.

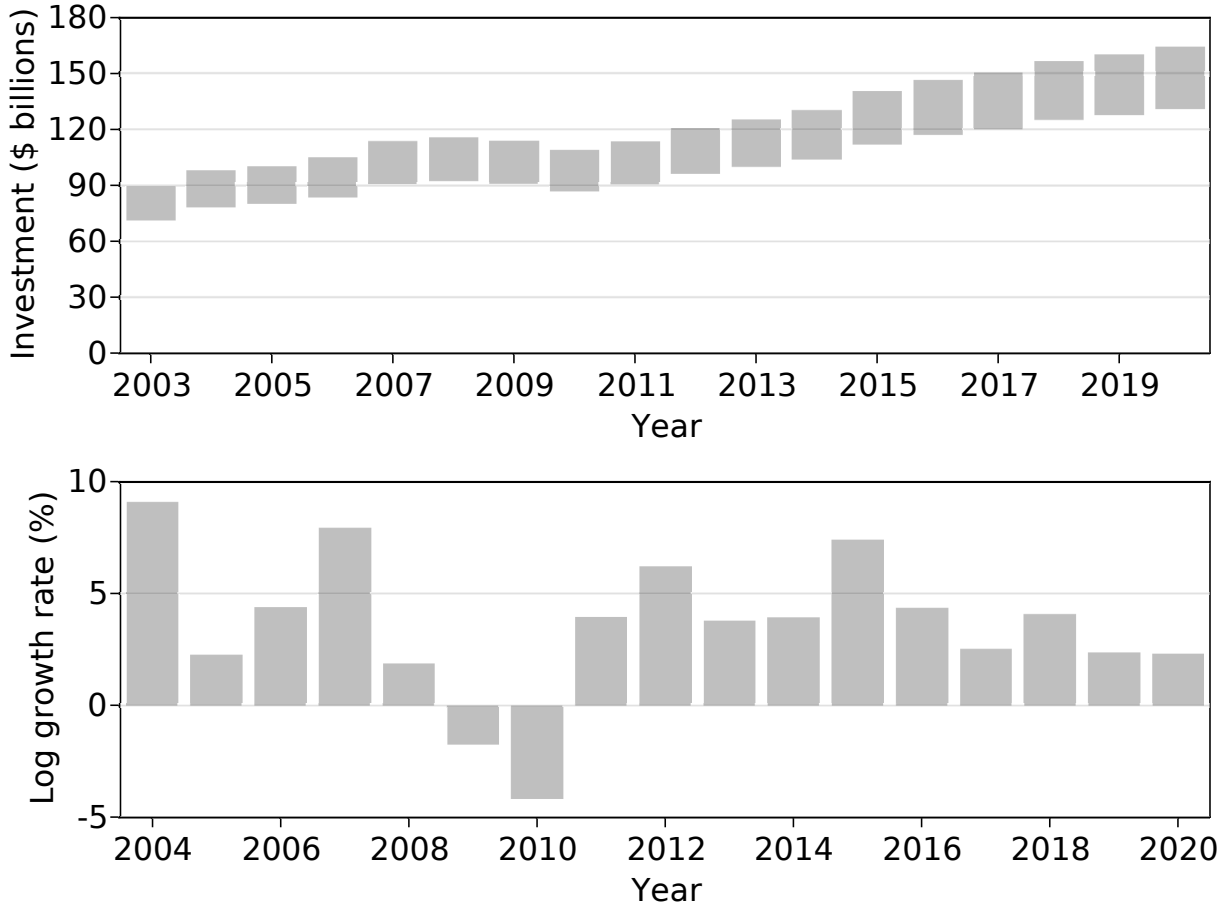
index for own-account software as a proxy to deflate current-dollars. In the future, an input cost index with a productivity adjustment could be developed specifically for data.

Depreciation. OECD (2009) recommends several options for determining depreciation parameters for an asset class, which generally requires information on prices or service lives. One option uses information on the service life and makes an additional assumption about the functional form of the depreciation pattern. A functional form that is commonly used for pragmatic reasons is the geometric model of depreciation, which yields a pattern of constant percentage decline in an asset's value. In the absence of econometric estimates of the geometric depreciation rate, the geometric depreciation rate can be estimated using a simple declining balance method as $\delta = R \ / \ S$, where R is an estimated declining balance rate and S is an average service life (Hulten and Wykoff 1996). The declining balance rate may be either estimated econometrically or is sometimes assumed to be 2. For software and databases, OECD (2010) recommends the average service life be obtained by surveying software users, surveying software suppliers, or hiring software consultants.

BEA does not compile a measure of depreciation for own-account databases but does compile a measure of depreciation for own-account software for the U.S. Fixed Assets Accounts (FAAs). To estimate depreciation for own-account software, BEA uses a 5-year service life and a declining balance rate of 1.65 to determine the geometric depreciation rate. The service life is based on estimates of the relationship between computer expenditures and software expenditures, anecdotal evidence about how long software is used before replacement (including an informal survey of business uses of software), and tax-law-based service lives of software.[17] The declining balance rate is borrowed from the Hulten-Wykoff methodology that is used for most equipment in the FAAs (Hulten and Wykoff 1981; Hulten, McCallum, et al. 1981; Wykoff and Hulten 1979; Fraumeni 1997). To address the lack of depreciation rate for own-account data, we use BEA's current geometric depreciation rate of 0.33 for own-account

[17]A summary of BEA's depreciation estimates is available at: `https://apps.bea.gov/national/pdf/BEA_depreciation_rates.pdf`.

Figure 1: Current-dollar annual investment in data assets



Note: The lower bound for investment in the upper panel uses a markup of 2.42, and the upper bound uses a markup of 3.03. Growth rates in the lower panel reflect the lower bound markup of 2.42.

software as a proxy.

5 Results

5.1 National Aggregates

5.1.1 Experimental Current-Dollar and Historical-Cost Estimates

Experimental estimates for aggregate business sector current-dollar investment in data assets for 2003–2020 are presented in figure 1. The upper panel in the figure reflects current-dollar levels and the lower panel reflects current-dollar log growth rates. The lower bound in the

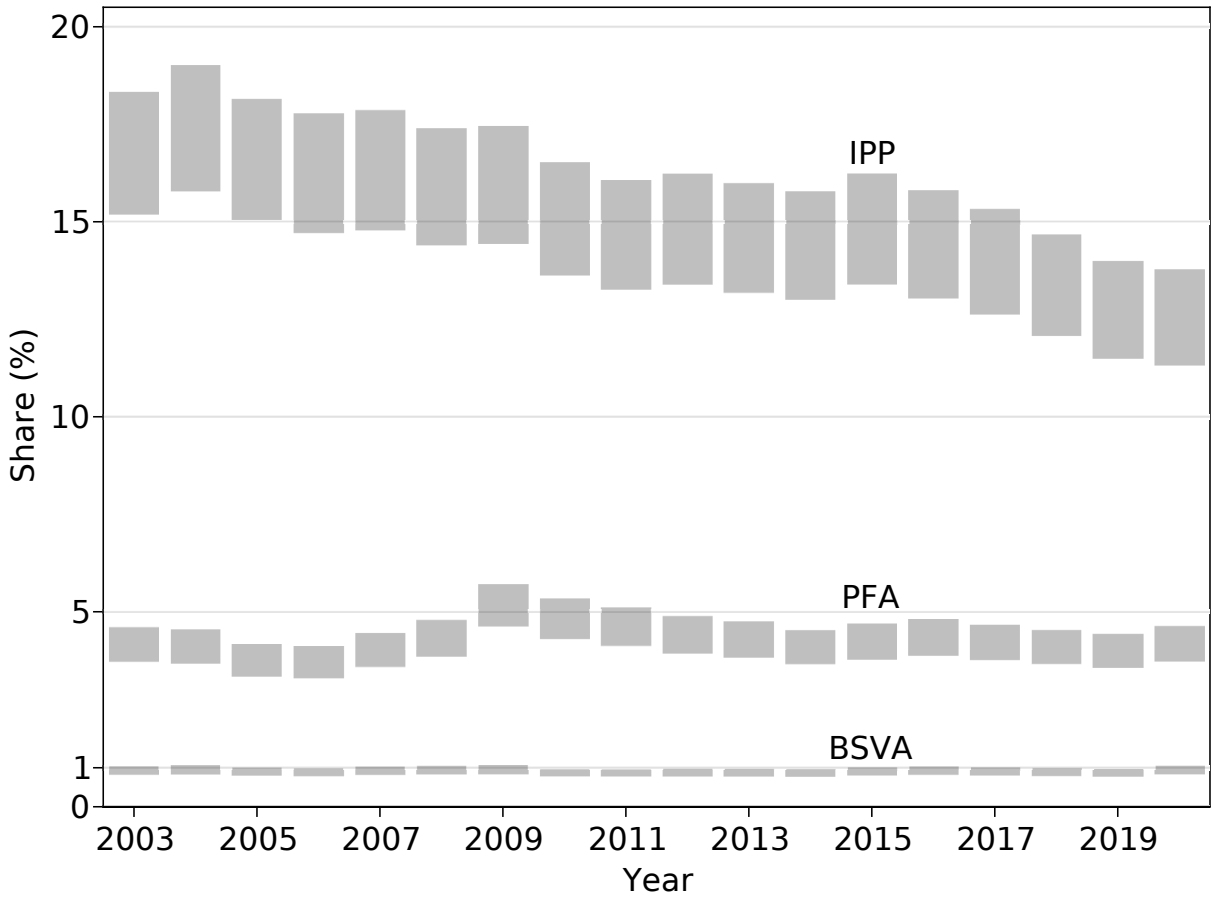
upper panel uses a markup of 2.42, and the upper bound uses a markup of 3.03. With the lower markup, estimated current-dollar investment in 2003 is $72 billion and in 2020 is $131 billion. Growth of current-dollar investment is as high as 9.1 percent for 2004 and as low as −4.2 percent for 2010. The average annual growth rate in current-dollar investment for 2004–2020 is 3.6 percent. Prior to the recession of 2007–2009, the average annual growth rate for 2004–2007 was 5.9 percent, which declined to 2.8 percent for 2008–2020. Average annual growth for the ten-year period 2011–2020 was 4.1 percent.

Current-dollar investment in data assets as a share of NIPA current-dollar business sector value-added (BSVA) is reported in figure 2. With the lower markup, the share of data investment for the period averages 0.8 percent of business sector value-added. With the higher markup, the share of data investment averages 1.0 percent of business sector value-added.

We also report in figure 2 current-dollar investment in data assets as a share of NIPA current-dollar investment in IPPs and as a share of NIPA current-dollar investment in private fixed assets. Lower bound data investment as a share of investment in IPPs declines over the period from 15.2 percent in 2003 to 11.3 percent in 2020 - the share averages 13.6 percent for the period. Lower bound data investment as a share of investment in private fixed assets is 3.7 percent in 2003 and 2020 with some variation over the period - the share averages 3.8 percent for the period.

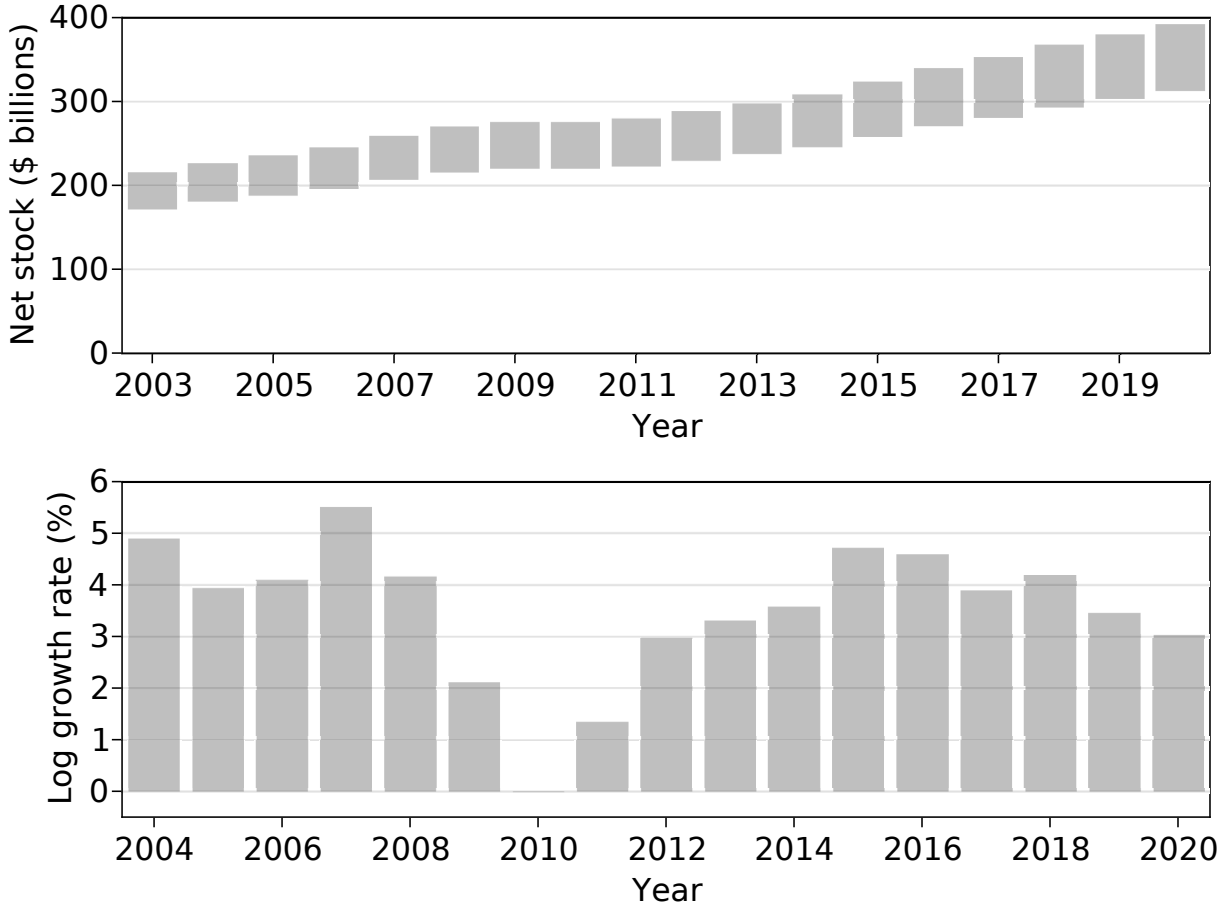
Experimental estimates for aggregate business sector historical-cost net stocks of data assets for 2003–2020 are presented in figure 3. The historical-cost net stocks are calculated using a perpetual inventory method (PIM) with geometric depreciation. The PIM is a sum of annual depreciated historical-cost investment in data assets. Annual historical-cost investment in data assets prior to 2003 is backcast using the annual growth in current-dollar investment in own-account software as an indicator. Depreciation of annual historical-cost investment is calculated using BEA's depreciation rate for own-account software of 0.33, assuming new investment in data assets is placed in service at midyear. The calculation for

Figure 2: Investment in data assets as a share of NIPA aggregates



Note: The lower bound for each share uses a markup of 2.42, and the upper bound uses a markup of 3.03. The numerator for each share is the current-dollar investment in data assets. The denominator for each share is the current-dollar NIPA aggregate plus the current-dollar investment in data assets. Business sector value-added (BSVA) is from line 2 of "National Income and Product Accounts: Table 1.3.5. Gross Value Added by Sector". Investment in intellectual property products (IPP) is from line 1 of "Fixed Assets Accounts: Table 3.7I. Investment in Private Intellectual Property Products by Industry", adjusted to exclude NPISH by subtracting lines 66, 67, and 72. Investment in private fixed assets (PFA) is from line 1 of "Fixed Assets Accounts: Table 3.7ESI. Investment in Private Fixed Assets by Industry", adjusted to exclude NPISH by subtracting line 66, 67, and 72.

Figure 3: Historical-cost annual net stocks of data assets



Note: The lower bound for net stock in the upper panel uses a markup of 2.42, and the upper bound uses a markup of 3.03. Growth rates in the lower panel reflect the lower bound markup of 2.42.

annual depreciated historical-cost investment, i.e., net stock N, in year t for investment I placed in service in year h can be summarized as follows

$$N_{t,h} = I_h \left(1 - \frac{\delta}{2}\right) (1 - \delta)^{t-h} . \tag{7}$$

The upper panel in figure 3 reflects historical-cost levels and the lower panel reflects historical-cost log growth rates. The lower bound in the upper panel uses a markup of 2.42, and the upper bound uses a markup of 3.03. With the lower markup, estimated historical-cost net stock in 2003 is $174 billion and in 2020 is $313 billion. Growth of historical-cost net stocks is as high as 5.4 percent for 2007 and as low as 0 percent for 2010. The average

annual growth rate in historical-cost net stocks for 2004–2020 is 3.4 percent.

We report the historical-cost net stocks of data assets as a share of the FAA historical-cost net stocks of IPPs and as a share of the FAA historical-cost net stocks of private fixed assets in figure 4. Lower bound data stocks as a share of IPP stocks decline over the period from 11.3 percent in 2003 to 8.6 percent in 2020 - the share averages 10.0 percent for the period. Lower bound data stocks as a share of private fixed assets stocks is 1.2 percent in 2003 and 1.0 percent in 2020 with some variation over the period - the share averages 1.1 percent for the period.

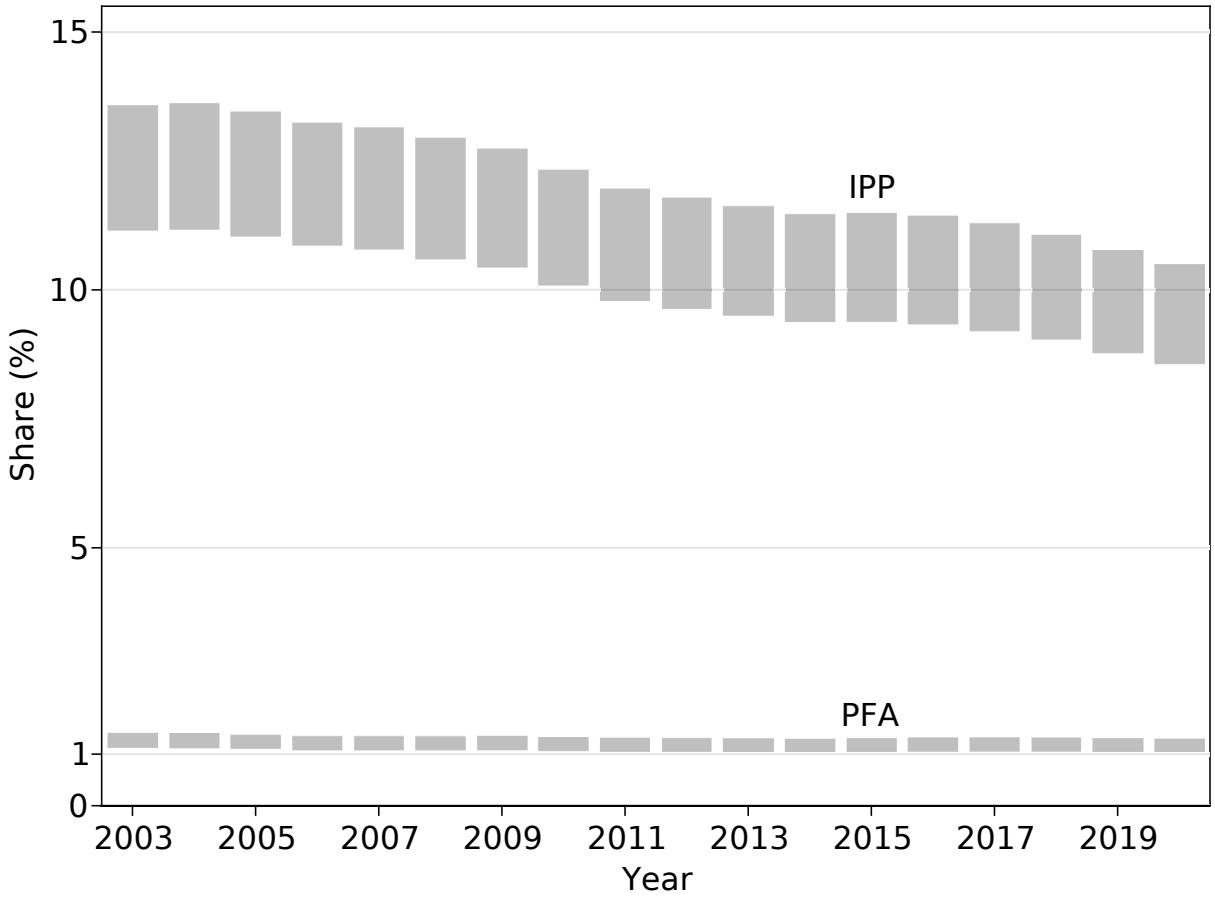
5.1.2 Real Growth

We compare the log growth in real U.S. aggregate measures with and without our experimental estimates of business sector investment in data assets for 2004–2020. Real aggregate measures include business sector value-added, private sector investment in IPPs, and private sector investment in software. We use the price index for own-account software as a proxy for own-account data and recalculate measured aggregate prices using Törnqvist expenditure shares. We report average and cumulative annual growth in table 7.

With the lower markup, the largest annual growth rate in real data investment is 10.8 percent for 2004 and the smallest is -2.8 percent for 2010. The average annual growth in real data investment for 2004–2020 is 3.8 percent and the cumulative annual growth for the period is 64.0 percent. When data investment is added to business sector value-added, the average annual change in real value-added over the period is an increase of 1 basis point, and the cumulative annual change is an increase of 9 basis points. Results are similar with the higher markup.

In contrast to business sector value-added, growth in real investment in IPPs and in software are both higher than real data investment. When data investment is added, the average and cumulative growth in real measures decline. With the lower markup, the decline in average annual growth in real investment in IPPs for 2004–2020 is 23 basis points (i.e.,

Figure 4: Net stocks of data assets as a share of FAA aggregates



Note: The lower bound for each share uses a markup of 2.42, and the upper bound uses a markup of 3.03. The numerator for each share is the historical-cost net stock of data assets. The denominator for each share is the historical-cost FAA aggregate plus the historical-cost net stock of data assets. Net stocks of intellectual property products (IPP) are from line 1 of "Fixed Assets Accounts: Table 3.3I. Historical-Cost Net Stock of Private Intellectual Property Products by Industry", adjusted to exclude NPISH by subtracting lines 66, 67, and 72. Net stocks of private fixed assets (PFA) are from line 1 of "Fixed Assets Accounts: Table 3.3ESI. Historical-Cost Net Stock of Private Fixed Assets by Industry", adjusted to exclude NPISH by subtracting lines 66, 67, and 72.

Table 7: Growth in real measures with and without investment in data assets 2004–2020 (%)

	Average			Cumulative		
	With data	W/o data	Δ	With data	W/o data	Δ
Markup of 2.42						
Data	3.76			63.96		
Value-added	1.74	1.73	0.01	27.83	27.74	0.09
IPPs	4.83	5.06	−0.23	77.28	80.90	−3.63
Software	6.42	7.58	−1.16	102.78	121.30	−18.53
Markup of 3.03						
Data	3.76			63.96		
Value-added	1.74	1.73	0.01	27.88	27.74	0.14
IPPs	4.78	5.06	−0.27	76.51	80.90	−4.40
Software	6.23	7.58	−1.36	99.61	121.30	−21.70

Note: The table reports average and cumulative log growth rates in real data investment along with changes in growth for business sector real value-added, private sector real investment in IPPs, and private sector real investment in software with and without data investment for 2004–2020. Price indexes for own-account software are used as a proxy for own-account data and aggregate price indexes are recalculated using Törnqvist expenditure shares.

0.23 percentage point) and in cumulative average annual growth is 363 basis points (i.e., 3.6 percentage points). The decline in average annual growth in real investment in software for the period is 116 basis points (i.e., 1.2 percentage points) and in cumulative average annual growth is 1,853 basis points (i.e., 18.5 percentage points). The declines are steeper with the higher markup. The changes in growth of business sector real value-added and private sector real investment in IPPs and software all reflect the modest growth in measured own-account data investment for the period.[18]

5.1.3 Shares of Investment by Occupation

In order to assess overlap between our estimates of data investment and investment in other IPPs - i.e., R&D and software - we calculate shares of investment in data by occupation. Table

[18]Goodridge et al. (2021) report labor productivity with and without data investment for 13 European Union countries and find a modest average increase in labor productivity of only 5 basis points with data investment for 2011–2016. The largest average annual increase by country is 9 basis points attributable to Germany, and largest average annual decrease by country is 9 basis points attributable to Estonia.

Table 8: Occupational Shares of Investment in Data 2003-2020

OEWS 2019	Description	Share(%)
43-9061	Office Clerks	6.3
11-1021	General and Operations Managers	4.7
43-9021	Data Entry Keyers	4.6
15-1245	Database Administrators and Architects	4.5
43-4051	Customer Service Representatives	3.9
13-1111	Management Analysts	3.6
43-3031	Bookkeeping, Accounting, and Auditing Clerks	3.5
11-3021	Computer and Information Systems Managers	3.5
43-6014	Secretaries and Administrative Assistants	3.0
15-1299	Computer Occupations, All Other	2.6
43-3021	Billing and Posting Clerks	2.4
13-2098	Financial and Investment Analysts	2.2
13-1161	Market Research Analysts and Marketing Specialists	2.2
15-1244	Network and Computer Systems Administrators	2.0
43-1011	First-Line Supervisors of Office and Administrative Workers	1.9
11-3031	Financial Managers	1.7
13-2011	Accountants and Auditors	1.7
11-2021	Marketing Managers	1.0
53-7062	Laborers and Freight, Stock, and Materials Movers	1.0
15-1232	Computer User Support Specialists	1.0
15-1241	Computer Network Architects	1.0
11-1011	Chief Executives	0.9
13-1020	Buyers and Purchasing Agents	0.9
13-1071	Human Resource Specialists	0.9
41-4012	Sales Representatives, Wholesale and Manufacturing	0.9
13-1198	Project Mgmt and Business Ops Specialists, All Other	0.9
11-3010	Administrative Services and Facilities Managers	0.8
11-2022	Sales Managers	0.8
43-5071	Shipping, Receiving, and Inventory Clerks	0.8
11-9041	Architectural and Engineering Managers	0.8
11-9198	Personal Service Managers, All Other	0.8
15-1221	Computer and Information Research Scientists	0.8
17-2112	Industrial Engineers	0.7
51-9061	Inspectors, Testers, Sorters, Samplers, and Weighers	0.7
41-3031	Securities, Commodities, and Financial Services Sales Agents	0.6
43-6011	Executive Secretaries	0.6
11-3051	Industrial Production Managers	0.6
15-2031	Operations Research Analysts	0.6
15-1212	Information Security Analysts	0.6
	Total	71.6

Note: Shares of investment in data are included for occupations with more than 0.5 percent share.

8 reports shares of investment in data for occupations with more than 0.5 percent share for 2003–2020. We choose the cutoff to fit the table onto a single page. The 39 occupations in table 8 account for 71.6 percent of our investment estimate for the period. There are 651 additional occupations included in our estimate but excluded from table 8.

We note two observations in table 8. First, even if an occupation makes the list of landmark occupations in table 2 does not imply the occupation will account for a large share of the investment estimate. Two landmark occupations excluded from table 8 are *wellhead pumpers* and *hearing aid specialists*, which may be useful references to calculate a numeric similarity for non-landmark occupations but do not play a big role in data investment. In contrast, landmark occupations like *data entry keyers* and *database administrators and architects* account for a large share of the data investment estimate. Second, R&D occupations and software occupations are mostly absent from table 8, which reflect in part the adjustments to exclude R&D intensive industries and software occupations. Occupations with low shares such as *computer and information research scientists*, *computer occupations (all other)*, and *computer network architects* may contribute to capital formation in data as well as capital formation in R&D and software, but occupations with the highest shares are unlikely to contribute to capital formation in R&D and software.

5.2 NAICS Sector Aggregates

5.2.1 Experimental Current-Dollar Estimates

Experimental estimates for current-dollar investment in data assets by NAICS sector for 2003–2020 are presented in table 9. Table 10 reports the average annual growth in current-dollar investment in data assets by NAICS sector for 2004–2020. The current-dollar estimates as a share of current-dollar value-added by NAICS sector for 2003–2020 are shown in figure 5. Since our methodology starts with an estimate of the occupation-based wage bill, results by NAICS sector reflect industries that employ occupations engaged in data-related activities.

For the period 2003–2020, table 9 shows the total current-dollar investment in data was

Table 9: Current-dollar investment in data assets by NAICS sector 2003–2020

NAICS	Description	($B)
11	Agriculture, Forestry, Fishing and Hunting	4
21	Mining, Quarrying, and Oil and Gas Extraction	26
22	Utilities	25
23	Construction	80
31-33	Manufacturing	246
42	Wholesale Trade	169
44-45	Retail Trade	127
48-49	Transportation and Warehousing	74
51	Information	139
52	Finance and Insurance	301
53	Real Estate and Rental and Leasing	46
54	Professional, Scientific, and Technical Services	149
55	Management of Companies and Enterprises	162
56	Administrative & Support and Waste Management & Remediation Services	192
72	Accommodation and Food Services	31
81	Other Services (except Public Administration)	27
	Total	1,798

Note: Current-dollar estimates summed for 2003–2020 by NAICS sector. Estimates for NPISH and general government are excluded: Educational Services (NAICS 61), Health Care and Social Assistance (NAICS 62), Arts, Entertainment and Recreation (NAICS 71), Religious, Grantmaking, Civic, Professional, and Similar Organizations (NAICS 813), and Public Administration (NAICS 92).

$1.8 trillion. The largest dollar investments were made in Finance and Insurance ($301 billion), Manufacturing ($246 billion), and Administrative Services ($192 billion). The smallest dollar investments were made in Agriculture, Utilities, and Mining.

Table 10 demonstrates that average annual growth in current-dollar investment in data assets was as low as 1.9 percent for Retail Trade and as high as 6.3 percent for Management of Companies. Average annual growth rates by NAICS sector exceeded the aggregate average annual growth rate of 3.6 percent for Mining (4.6 percent), Construction (4.1 percent), Transportation and Warehousing (4.0 percent), Real Estate (4.2 percent), PST Services (5.9 percent), and Management of Companies (6.3 percent).

In figure 5, the three largest shares of NAICS sector value-added show up for Management of Companies (NAICS 55), Administrative Services (NAICS 56), and Finance and Insurance

Table 10: Average annual growth in current-dollar investment in data assets by NAICS sector 2004–2020

NAICS	Description	(%)
11	Agriculture, Forestry, Fishing and Hunting	3.6
21	Mining, Quarrying, and Oil and Gas Extraction	4.6
22	Utilities	2.9
23	Construction	4.1
31-33	Manufacturing	2.0
42	Wholesale Trade	3.0
44-45	Retail Trade	1.9
48-49	Transportation and Warehousing	4.0
51	Information	3.3
52	Finance and Insurance	3.6
53	Real Estate and Rental and Leasing	4.2
54	Professional, Scientific, and Technical Services	5.9
55	Management of Companies and Enterprises	6.3
56	Administrative & Support and Waste Management & Remediation Services	3.4
72	Accommodation and Food Services	2.7
81	Other Services (except Public Administration)	3.1

Note: Annual log growth rates for current-dollar investment averaged for 2004–2020.

(NAICS 52). The smallest shares of NAICS sector value-added show up for Agriculture (NAICS 11), Real Estate (NAICS 53), Accommodation and Food Services (NAICS 72), and Other Services (NAICS 81).

5.2.2 Real Growth

We compare the log growth in real value-added for NAICS sectors with and without our experimental estimates of business sector investment in data assets for 2004–2020. We use the price index for own-account software as a proxy for own-account data and recalculate measured NAICS sector prices using Törnqvist expenditure shares. We report average and cumulative annual growth in real value-added growth by NAICS sector in table 11.

Consistent with the result for business sector real value-added in table 7, the change in average annual growth is positive for each NAICS sector except Information (NAICS 51). When data investment is added, the largest increases in average real value-added growth show up for Management of Companies (NAICS 55, 11 basis points), Construction (NAICS

Table 11: Growth in real value-added with and without investment in data assets by NAICS sector 2004–2020 (%)

	Average			Cumulative		
NAICS	With data	W/o data	Δ	With data	W/o data	Δ
11	2.26	2.25	0.00	38.35	38.33	0.02
21	3.36	3.35	0.00	57.05	57.02	0.03
22	1.85	1.84	0.01	31.44	31.32	0.11
23	−0.84	−0.88	0.03	−14.34	−14.93	0.58
31-33	1.39	1.39	0.01	23.71	23.62	0.09
42	1.30	1.28	0.02	22.11	21.79	0.32
44-45	0.92	0.91	0.01	15.61	15.46	0.15
48-49	1.29	1.26	0.03	21.89	21.46	0.43
51	5.51	5.53	−0.02	93.69	94.01	−0.32
52	1.74	1.71	0.03	29.6	29.15	0.44
53	1.91	1.90	0.00	32.44	32.38	0.05
54	2.99	2.97	0.02	50.85	50.49	0.36
55	2.43	2.33	0.11	41.36	39.53	1.83
56	2.42	2.39	0.03	41.07	40.64	0.43
72	−0.69	−0.7	0.01	−11.69	−11.91	0.21
81	−1.07	−1.09	0.02	−18.19	−18.49	0.30

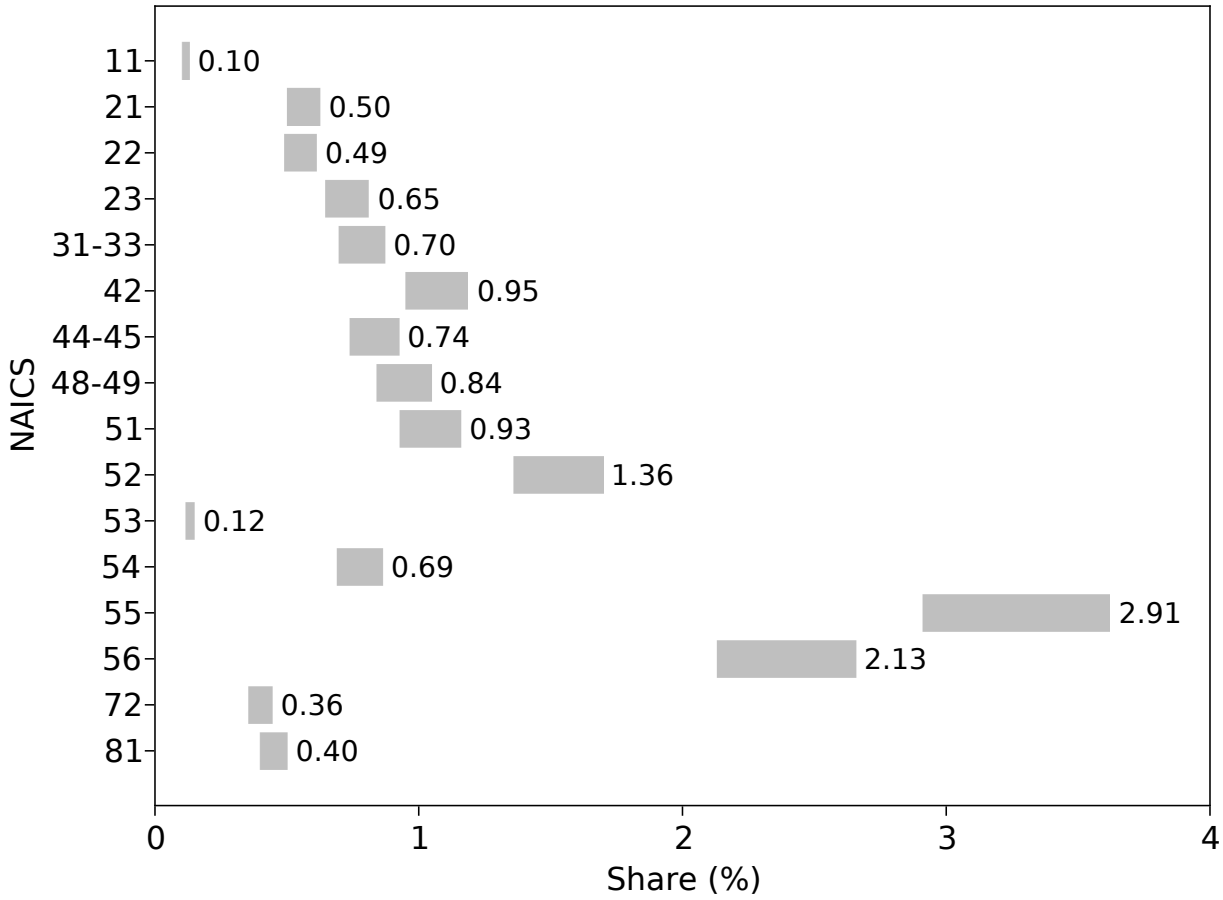
Note: The table reports average and cumulative log growth rates in real value-added by NAICS sector with and without data investment for 2004–2020. Price indexes for own-account software are used as a proxy for own-account data and NAICS price indexes are recalculated using Törnqvist expenditure shares.

23, 3 basis points), Transportation and Warehousing (NAICS 48-49, 3 basis points), Finance and Insurance (NAICS 52, 3 basis points) and Administrative Services (NAICS 56, 3 basis points). Increases in cumulative real value-added growth are also largest for those NAICS sectors. The smallest increases in average and cumulative real value-added growth show up for Agriculture (NAICS 11), Mining (NAICS 21), Information (NAICS 51), and Real Estate (NAICS 53).

5.2.3 Non-profit Institutions Serving Households

While core experimental results for the paper are limited to the business sector, table 12 reports supplemental experimental results for current-dollar investment in data assets for NAICS sectors that represent the NPISH sector.

Figure 5: Investment in data assets as a share of value-added by NAICS sector 2003–2020



Note: The lower bound for each share uses a markup of 2.42, and the upper bound uses a markup of 3.03. Enumerated shares reflect the lower bound share. The numerator for each share is the current-dollar investment in data assets summed for 2003–2020. The denominator for each share is the current-dollar investment in data assets plus NAICS sector value-added summed for 2003–2020 from the U.S. Bureau of Economic Analysis, "Interactive Access to Industry Economic Accounts Data: Table 1 Value Added by Industry".

Table 12: Current-dollar investment in data assets for NPISH 2003–2020

NAICS	Description	($B)
61	Educational Services	150
62	Health Care and Social Assistance	318
71	Arts, Entertainment, and Recreation	21
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations	45
	Total	534

Note: Current-dollar estimates summed for 2003–2020.

For the period 2003–2020, table 12 shows the total current-dollar investment in data by NPISH was $534 billion. The largest dollar investments were made in Health Care and Social Assistance ($318 billion), which is bigger than Finance and Insurance in the business sector. The second largest dollar investments by NPISH were made in Educational Services ($150 billion).

6 Conclusions

In this paper, we measure the value of own-account data stocks and flows for the U.S. business sector by summing the production costs of data-related activities implicit in occupations. In our experimental estimates, we find that annual current-dollar investment in own-account data assets for the U.S. business sector grew from $72 billion in 2003 to $131 billion in 2020, which yields an average annual growth of 3.6 percent. Cumulative current-dollar investment for the period 2003–2020 was $1.8 trillion. Overall, our results indicate that business sector investment in own-account data grew only a bit faster than other business sector economic activity and slower than business sector investment in other IPPs.

The method we use in the paper augments the traditional sum-of-costs methodology for measuring other own-account intellectual property products in national economic accounts by proxying occupation-level time-use factors using a machine learning model and the text of online job advertisements (Blackburn 2021). The method appears to be a feasible method for identifying occupations engaged in data-related activities and for estimating the time-effort

that occupations allocate to data-related activities. The time-use factors we develop for occupations engaged in own-account data appear to have some overlap with the time-use factors we develop for occupations engaged in own-account software, which suggests an incentive to jointly estimate own-account data and own-account software to ensure consistency and prevent double counting if the scope of own-account software is expanded to include more occupations in the future.

In the future, we plan to expand the scope of estimation to include the NPISH and government sectors. In addition, this paper focuses on estimating current-dollar values and uses the price index and depreciation rate for BEA's measures of own-account software as proxies to calculate real growth rates and net stocks, respectively. Thus, an important area of development in the future will be a price index and depreciation rate specific to own-account data. Likewise, an important area for further development will be estimates of purchased data assets and more precise adjustments for capital formation and overlap with R&D and software investment.

References

Ackoff, Russell L (1989). "From data to wisdom". In: *Journal of applied systems analysis* 16.1, pp. 3–9.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman (June 2016). "The Economics of Privacy". In: *Journal of Economic Literature* 54.2, pp. 442–92. DOI: 10.1257/jel.54.2.442.

Ahmad, Nadim (Dec. 8, 2004). "The Measurement of Databases in the National Accounts". In: Second Meeting of the Advisory Expert Group on National Accounts. SNA/M2.04/04. New York, USA: United Nations Statistics Division. URL: https://unstats.un.org/unsd/nationalaccount/aeg/m2-04.asp.

Ahmad, Nadim (July 19, 2005). "Follow-Up to the Measurement of Databases in the National Accounts". In: Third Meeting of the Advisory Expert Group on National Accounts. Vol. SNA/M1.05/19.1. UNESCAP, Bangkok. URL: https://unstats.un.org/unsd/nationalaccount/aeg/m1-05.asp.

Ahmad, Nadim and Peter van de Ven (Nov. 9, 2018). "Recording and measuring data in the System of National Accounts". In: Meeting of the Informal Advisory Group on measuring GDP in a digitalised economy. Vol. SDD/CSSP/WPNA(2018)5. Working Party on National Accounts. OECD Statistics, Data Directorate Committee on Statistics, and Statistical Policy. URL: https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/CSSP/WPNA(2018)5&docLanguage=En.

Bakhshi, Hasan, Albert Bravo-Biosca, and Juan Mateos-Garcia (Mar. 2014). *Inside the Datavores: How data and online analytics affect business performance*. URL: https://www.nesta.org.uk/report/briefing-inside-the-datavores/.

Blackburn, Christopher J. (Mar. 17, 2021). "Valuing the Data Economy Using Machine Learning and Online Job Postings". In: The Sixth World KLEMS Conference 2021. Vol. Digital Economy. Virtual. URL: https://scholar.harvard.edu/files/jorgenson/files/valuing_data_klems.pdf.

Boisot, Max and Agustí Canals (Jan. 1, 2004). "Data, information and knowledge: have we got it right?" In: *Journal of Evolutionary Economics* 14.1, pp. 43–67. ISSN: 0936-9937, 1432-1386. DOI: 10.1007/s00191-003-0181-9.

Burning Glass Technologies (2019). *Mapping the Genome of Jobs: The Burning Glass Skills Taxonomy*. URL: https://www.burning-glass.com/research-project/skills-taxonomy.

Chute, Jason W., Stephanie H. McCulla, and Shelly Smith (Apr. 2018). "Preview of the 2018 Comprehensive Update of the National Income and Product Accounts". In: *Survey of Current Business* 98.4.

Dey, Matthew, David S. Piccone Jr, and Stephen Stephen M. Miller (Aug. 27, 2019). "Model-based estimates for the Occupational Employment Statistics program". In: *Monthly Labor Review*. ISSN: 19374658. DOI: 10.21916/mlr.2019.19.

Farboodi, Maryam and Laura Veldkamp (Feb. 2021). *A Growth Model of the Data Economy*. Working Paper 28427. National Bureau of Economic Research. DOI: 10.3386/w28427.

Fraumeni, Barbara M. (July 1997). "The Measurement of Depreciation in the U.S. National Income and Product Accounts". In: *Survey of Current Business* 77.7. URL: https://apps.bea.gov/scb/pdf/national/niparel/1997/0797fr.pdf.

Galindo-Rueda, Fernando and Fabien Verger (July 16, 2016). *OECD Taxonomy of Economic Activities Based on R&D Intensity*. OECD Science, Technology and Industry Working Papers 2016/04. DOI: 10.1787/5jlv73sqqp8r-en.

Goodridge, Peter, Jonathan Haskel, and Harald Edquist (Sept. 28, 2021). "We See Data Everywhere Except in the Productivity Statistics". In: *Review of Income and Wealth*. ISSN: 0034-6586, 1475-4991. DOI: 10.1111/roiw.12542.

Hopson, Amy (Aug. 19, 2021). "Mapping Employment Projections and O*NET data: a methodological overview". In: *Monthly Labor Review*. ISSN: 19374658. DOI: 10.21916/mlr.2021.18.

Hughes-Cromwick, Ellen and Julia Coronado (Feb. 2019). "The Value of US Government Data to US Business Decisions". In: *Journal of Economic Perspectives* 33.1, pp. 131–46. DOI: 10.1257/jep.33.1.131.

Hulten, Charles R., Janice McCallum, and Urban Institute, eds. (1981). *Depreciation, inflation, and the taxation of income from capital*. Washington, D.C: Urban Institute Press. 319 pp. ISBN: 978-0-87766-311-9.

Hulten, Charles R. and Frank C. Wykoff (Apr. 1981). "The estimation of economic depreciation using vintage asset prices". In: *Journal of Econometrics* 15.3, pp. 367–396. ISSN: 03044076. DOI: 10.1016/0304-4076(81)90101-9.

Hulten, Charles R. and Frank C. Wykoff (Jan. 1996). "ISSUES IN THE MEASUREMENT OF ECONOMIC DEPRECIATION INTRODUCTORY REMARKS". In: *Economic Inquiry* 34.1, pp. 10–23. ISSN: 00952583, 14657295. DOI: 10.1111/j.1465-7295.1996.tb01361.x.

Jones, Charles I. and Christopher Tonetti (Sept. 2020). "Nonrivalry and the Economics of Data". In: *American Economic Review* 110.9, pp. 2819–58. DOI: 10.1257/aer.20191330.

Lancaster, Vicki Ann, Devika Mahoney-Nair, and Nathaniel Ratcliff (Mar. 10, 2021). *Review of Burning Glass Job-ad Data*. TR 2021-013. Proceedings of the Biocomplexity Institute. DOI: 10.18130/V3-KADT-5D77.

Le, Quoc and Tomas Mikolov (June 22, 2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.

Li, Wendy C.Y., Nirei Makoto, and Yamana Kazufumi (Feb. 19, 2019). *Value of Data: There's No Such Thing as a Free Lunch in the Digital Economy*. Working Paper. Bureau of Economic Analysis. URL: https://www.bea.gov/research/papers/2018/value-data-theres-no-such-thing-free-lunch-digital-economy.

Mokyr, Joel (Sept. 4, 2013). "The Knowledge Society: Theoretical and Historical Underpinnings". In: Ad Hoc Group of Experts Meeting on Knowledge Systems for Development. New York, USA: United Nations Department of Economic and Social Affairs. URL: https://web.archive.org/web/20051216205602/http://www.unpan.org:80/dpepa-kmb-ksranda.asp.

Moro Visconti, Roberto, Alberto Larocca, and Michele Marconi (2017). "Big Data-Driven Value Chains and Digital Platforms: From Value Co-Creation to Monetization". In: *SSRN Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.2903799.

Moulton, Brent R. and David F. Sullivan (Sept. 1999). "A Preview of the 1999 Comprehensive Revision of the National Income and Product Accounts: New and Redesigned Tables". In: *Survey of Current Business* 79.9. URL: https://apps.bea.gov/scb/pdf/NATIONAL/NIPA/1999/0999niw.pdf.

Nguyen, David and Marta Paczos (2020). *Measuring the economic value of data and cross-border data flows*. Series: Digital Economy Papers Vol. 297. OECD. DOI: 10.1787/6345995e-en.

OECD (2009). *Measuring capital: OECD manual 2009*. 2. ed. Paris. ISBN: 978-92-64-02563-9. URL: https://unstats.un.org/unsd/nationalaccount/docs/OECD-Capital-e.pdf.

— (2010). *Handbook on deriving capital measures of intellectual property products*. Paris. ISBN: 978-92-64-07290-9. URL: https://unstats.un.org/unsd/nationalaccount/docs/OECD-IPP.pdf.

— (Apr. 2, 2013). *Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value*. OECD Digital Economy Papers 220. DOI: 10.1787/5k486qtxldmq-en.

— (Oct. 8, 2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. The Measurement of Scientific, Technological and Innovation Activities. OECD. ISBN: 978-92-64-23901-2. DOI: 10.1787/9789264239012-en.

— (Apr. 7, 2021). "An update on recording and measuring data in the system of national accounts". In: 15th Meeting of the Advisory Expert Group on National Accounts. Vol. SNA/M1.21/7.4. Virtual. URL: https://unstats.un.org/unsd/nationalaccount/aeg/2021/M15.asp.

Parks, Roger B. et al. (Sept. 1, 1999). "How officers spend their time with the community". In: *Justice Quarterly* 16.3, pp. 483–518. ISSN: 0741-8825, 1745-9109. DOI: 10.1080/07418829900094241.

Rassier, Dylan G., Robert J. Kornfeld, and Erich H. Strassner (May 10, 2019). "Treatment of Data in National Accounts". In: BEA Advisory Committee. Vol. Measuring Data in the National Accounts. BEA's headquarters in Suitland, Maryland. URL: https://www.bea.gov/system/files/2019-05/Paper-on-Treatment-of-Data-BEA-ACM.pdf.

Řehůřek, Radim and Petr Sojka (May 22, 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50.

Reinsdorf, Marshall and Jennifer Ribarsky (Jan. 3, 2020). "Measuring the Digital Economy in Macroeconomic Statistics: The Role of Data". In: ASSA 2020 Annual Meeting. Vol. Big Data: Value, National Accounts, and Public Policy. San Diego, CA. URL: https://www.aeaweb.org/conference/2020/preliminary/2179.

Rowley, Jennifer (Apr. 2007). "The wisdom hierarchy: representations of the DIKW hierarchy". In: *Journal of Information Science* 33.2, pp. 163–180. ISSN: 0165-5515, 1741-6485. DOI: 10.1177/0165551506070706.

Solow, Robert M. (Feb. 1956). "A Contribution to the Theory of Economic Growth". In: *The Quarterly Journal of Economics* 70.1, p. 65. ISSN: 00335533. DOI: 10.2307/1884513.

Statistics Canada (June 24, 2019a). *Measuring investment in data, databases and data science: Conceptual framework*. 13-605-X201900100009. URL: https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00008-eng.htm.

— (July 10, 2019b). *The value of data in Canada: Experimental estimates*. 13-605-X201900100009. URL: https://www150.statcan.gc.ca/n1/pub/13-605-x/2019001/article/00009-eng.htm.

U.S. Bureau of Economic Analysis (2020). "Chapter 6: Private Fixed Investment". In: *NIPA Handbook: Concepts and Methods of the U.S. National Income and Product Accounts*. URL: https://www.bea.gov/resources/methodologies/nipa-handbook.

U.S. Bureau of Labor Statistics (2021). *Occupational Employment Statistics: National industry-specific and by ownership.* URL: https://www.bls.gov/oes/tables.htm.

United Nations (1993). *System of national accounts 1993*. Rev. ed. Brussels/Luxembourg : Washington, D.C. : Paris : New York : Washington, D.C. 711 pp. ISBN: 978-92-1-161352-0. URL: https://unstats.un.org/unsd/nationalaccount/sna1993.asp.

— (July 19, 2010). *System of National Accounts 2008*. United Nations. ISBN: 9789210544603. DOI: 10.18356/4fa11624-en.

United Nations Inter-Secretariat Working Group on National Accounts (2022). *Recording of Data in the National Accounts*. URL: https://unstats.un.org/unsd/nationalaccount/RAdocs/DZ6_GN_Recording_of_Data_in_NA.pdf.

van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Varian, Hal (Jan. 2018). "Artificial Intelligence, Economics, and Industrial Organization". In: *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, pp. 399–419. URL: http://www.nber.org/chapters/c14017.

Woolhandler, Steffie and David U. Himmelstein (Oct. 2014). "Administrative Work Consumes One-Sixth of U.S. Physicians' Working Hours and Lowers their Career Satisfaction". In: *International Journal of Health Services* 44.4, pp. 635–642. ISSN: 0020-7314, 1541-4469. DOI: 10.2190/HS.44.4.a.

Wykoff, Frank C. and Charles R. Hulten (July 26, 1979). *Tax and Economic Depreciation of Machinery and Equipment: A Theoretical and Empirical Appraisal, Phase II Report*. 28. Washington, D.C: The U.S. Treasury Department Office of Tax Analysis. URL: https://home.treasury.gov/system/files/131/WP-28.pdf.