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

José Bayoán Santiago Calderón
Dylan G. Rassier

ASSA 2023 – Advances in Machine Learning on Online Job Postings
Friday, January 6, 2023 14:30 – 16:30 CST

The views express here are those of the authors and not necessarily those of the U.S. Department of Commerce or the Bureau of Economic Analysis.

Motivation

- Implications of data as an asset in productivity and predicted economic growth patterns (Farboodi and Veldkamp 2021; Jones and Tonetti 2020)
- Estimates of data (Goodridge, Haskel, and Edquist 2021)
- Treatment of data in the System of National Accounts (SNA) (Rassier, Kornfeld, and Strassner 2019)

  SNA08 10.113: The cost of preparing data in the appropriate format is included in the cost of the database but not the cost of acquiring or producing the data.

- How to measure own-account data assets in the business sector?

Sum-of-costs approach

Production costs include:

- Labor costs
- Capital costs
- Intermediate consumption

The strategy will consist of:

- Estimate time-use allocated by occupations (Blackburn 2021),
- Obtain a wage bill associated with the occupations and their time-use allocations to data-relevant activities,
- Apply a markup factor to the wage bill to incorporate full sum-of-costs, and
- Apply adjustment factors for capital formation and multiple counting

Full production costs (continued)

Production cost function

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

Time-use factor

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

Ratio of employees engaged in relevant activities

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

Similarity to closest landmark occupation

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

Full production costs (continued)

Effective time-use factor

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

Sum-of-costs function for production cost

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

Lastly, we apply industry-specific adjustments to obtain capital formation and mitigate multiple counting

Full production costs (continued)

- Employment and wage bill estimates from Occupational Employment and Wage Statistics (OEWS) program (U.S. Bureau of Labor Statistics 2021; Dey, S. Piccone Jr, and Stephen M. Miller 2019)
- Job ads data from Burning Glass Technologies (Burning Glass Technologies 2019)
- Model fitting using doc2vec for autocoder (Řehůřek and Sojka 2010; Le and Mikolov 2014)
- Markup and national accounts data from BEA published tables

Who does what?

"Anyone who actually writes software, please report to the 10th floor at 2 pm today." - Elon

Processing Information

Save Table: XLSX CSV

Compiling, coding, categorizing, calculating, tabulating, auditing, or verifying information or data.

Level examples:

- 85 Compile data for a complex scientific report
- 57 Calculate the adjustments for insurance claims
- 28 Calculate the costs for shipping packages

873 occupations shown Show Job Zones: All 1 2 3 4 5 Hide detailed work activities

Importance	Level	Job Zone	Code	Occupation
96	90	5	13-2099.01	Financial Quantitative Analysts
95	90	5	25-1067.00	Sociology Teachers, Postsecondary Bright Outlook • Compile specialized bibliographies or lists of materials.
94	93	4	15-2099.01	Bioinformatics Technicians
94	88	4	33-3021.06	Intelligence Analysts
93	92	5	19-1021.00	Biochemists and Biophysicists

Data – Potential Sources

O*NET OnLine

Lightcast

NLx National Labor Exchange

LinkedIn

Data – Considerations

- Job ads are from a perspective employer compared to resumes and workforce surveys which are from present or historical employee accounts;
- Coverage in terms of geography, temporal, included occupations, employer/industry, sample sizes, and detail;
- Identifying work activities vs skills-based.

Data – Sample

Item	Description
Dataset	Lightcast Job Ads (Formerly Burning Glass Technologies)
Geography	U.S. Based (includes territories)
Period	2010 – 2019
Sample Size	239M $\parallel$ For US States & DC $\wedge$ w/NAICS4 $\wedge$ O*NET $\approx$ 140M
Occupations	O*NETs 1k+ $\parallel$ OEWS 800+
Industries	NAICS4

Data preparation included sampling O*NET occupations by average composition based on OEWS data from 2015–2020. Each O*NET occupation had at least 1,500 job ads. Stratified sampling by occupation/industry and job posting order.

Modeling

- Target was 2010 O*NET SOC code from BGT autocoder.
- Model trained was a doc2vec trained on $\approx$ 1M observations training sample
- We compute the pairwise cosine similarity for each occupation using a 1000-dimensional feature representation

BGT skills identified as data relevant

Data Entry	Data Validation	Data Conversion
Data Analysis	Assessment Data	Data Privacy
Data Collection	Data Manipulation	Data Integrity
Data Management	Data Acquisition Systems	Master Data Management (MDM)
Database Management	Data Security	Data Documentation
Relational Databases	Big Data Analytics	Data Warehouse Processing
Database Administration	Data Capture	Clinical Data Interchange Standards Consortium(CDISC)
Data Warehousing	Data Governance	Data Trending
Data Quality	Data Communications	GPS Data
Data Mining	Geographic Information System (GIS) Data	Data Evaluation
Data Acquisition	Clinical Data Management	Data Cleaning
Material Safety Data Sheets (MSDS)	Database Schemas	Database Architecture
Data Science	Data Mapping	Enterprise Data Management
Big Data	Data Reports	Database Tuning
Database Design	Managing Student Data	Database Marketing
Data Modeling	Data Migration	Data Engineering
Data Transformation	Data Verification	Database Programming
Data Architecture	Clinical Data Review	Data Loss Prevention
Data Structures	Quantitative Data Analysis	Data Warehouse Development
Data Integration	Clinical Data Analysis	Data Archiving

Note: Top 60 skills by frequency out of 203 data relevant skills non-software manually identified in (Blackburn 2021).

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

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 $\hat{\rho}_\omega$.

Mark-up Factor

Table: Weighted composite ratio for full sum-of-costs

	Ratio	Share (%)
Compensation	1.15	46
Intermediate consumption	0.81	32
Consumption of fixed capital	0.29	11
Net operating surplus	0.27	11
Markup	2.52	

Note: All data are from BEA's annual industry accounts. Intermediate consumption excludes materials. The table reports the simple average for 2002-2021 of each annual measure summed for NAICS 518-519 and NAICS 5415 divided by annual wages and salaries summed for the same industries.

Adjustments

Adjusting for R&D

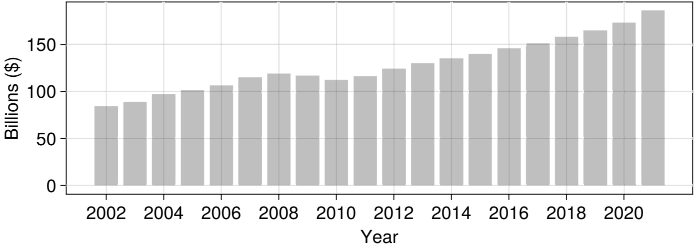
$$\hat{\tau_\omega}' = \hat{\tau_\omega}\left(1 - \hat{\rho_\omega}'\right) \tag{7}$$

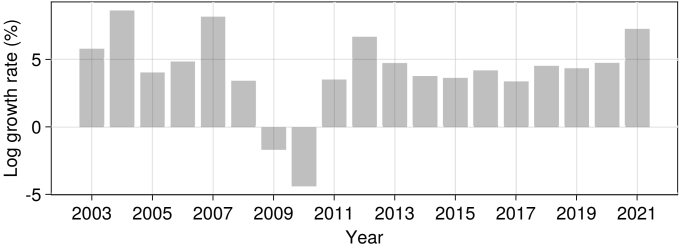
The effective time-use for occupations include accounting for a time-use factor (based on ratio of employees engaging in R&D).

Adjusting for own-account software: we exclude occupations used for estimating own-account software

Adjusting for purchased data We apply a 50% discount to NAICS 518 (Data Processing, Hosting, and Related Services)

Current-dollar annual investment in data assets





Current-dollar investment in data assets by NAICS sector

NAICS	Description	($B)
11	Agriculture, Forestry, Fishing and Hunting	4
21	Mining, Quarrying, and Oil and Gas Extraction	29
22	Utilities	28
23	Construction	95
31-33	Manufacturing	353
42	Wholesale Trade	183
44-45	Retail Trade	141
48-49	Transportation and Warehousing	81
51	Information	159
52	Finance and Insurance	338
53	Real Estate and Rental and Leasing	51
54	Professional, Scientific, and Technical Services	646
55	Management of Companies and Enterprises	179
56	Administrative & Support and Waste Management & Remediation Services	210
72	Accommodation and Food Services	36
81	Other Services (except Public Administration)	30
	Total	2,563

Current-dollar investment in data assets NPISH

NAICS	Description	($B)
61	Educational Services	149
62	Health Care and Social Assistance	329
71	Arts, Entertainment, and Recreation	23
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations	51
	Total	552

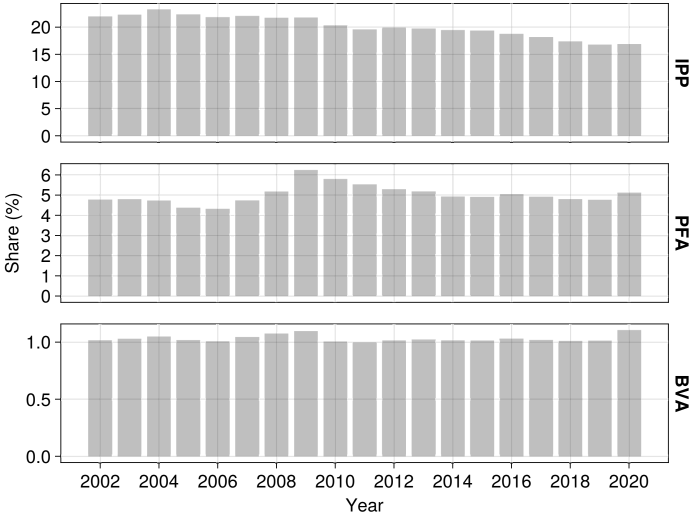
Note: Current-dollar estimates summed for 2002–2021.

Occupational Shares of Investment in Data

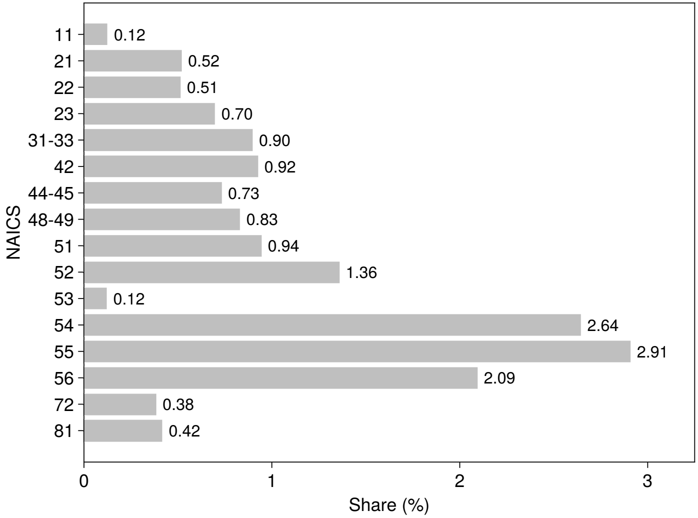
OEWS 2021	Description	Share (%)
43-9061	Office Clerks, General	5.68
13-1111	Management Analysts	5.27
11-1021	General and Operations Managers	4.48
43-9021	Data Entry Keyers	4.26
11-3021	Computer and Information Systems Managers	4.15
43-3031	Bookkeeping, Accounting, and Auditing Clerks	3.28
43-4051	Customer Service Representatives	3.21
43-6014	Secs and Admin Assistants, Except Legal, Medical, and Executive	2.85
13-1161	Market Research Analysts and Marketing Specialists	2.68
15-1242	Database Administrators	2.58
15-1243	Database Architects	2.38
15-1244	Network and Computer Systems Administrators	2.18
11-3031	Financial Managers	2.17
13-2011	Accountants and Auditors	2.03
43-1011	First-Line Supervisors of Office and Admin Support Workers	1.73
15-1299	Computer Occupations, All Other	1.41
11-2021	Marketing Managers	1.12
15-1241	Computer Network Architects	1.12
11-9041	Architectural and Engineering Managers	1.05
	Total	53.63

Note: Shares of investment in data are included for occupations with at least 1 percent share.

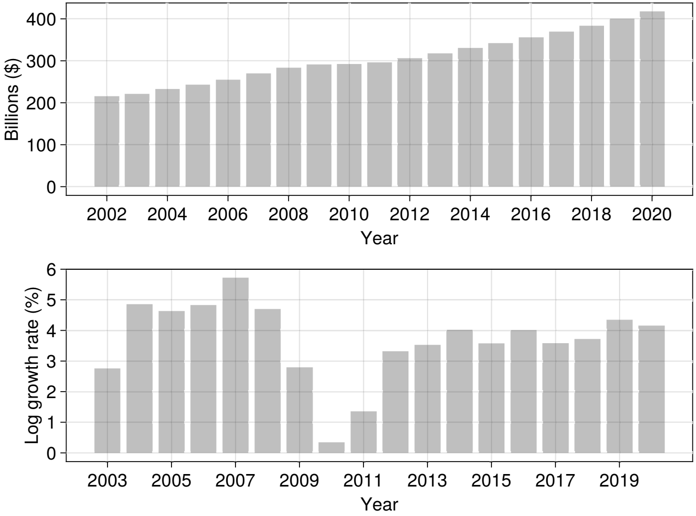
Investment in data assets as a share of NIPA aggregates



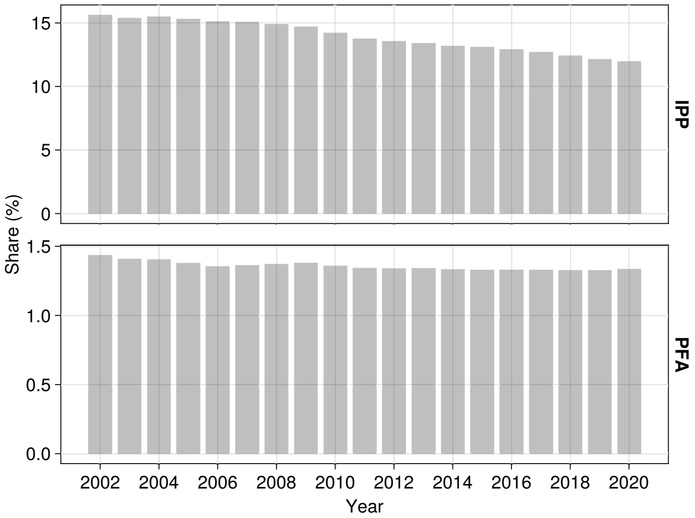
Investment in data assets as a share of value-added by NAICS



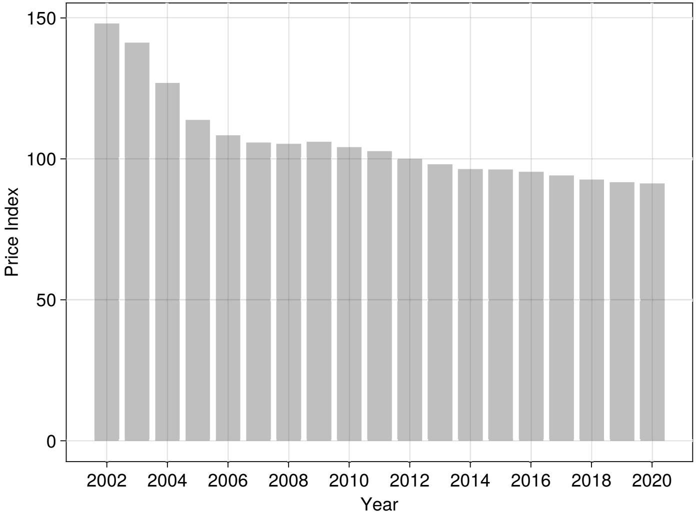
Historical-cost annual net stocks of data assets



Net stocks of data assets as a share of FAA aggregates



Own-account data price index



Growth in real measures with and without investment in data assets 2003–2020 (%)

	Average			Cumulative		
	With data	W/o data	Δ	With data	W/o data	Δ
Data	7.47			134.42		
Value-added	1.99	1.95	0.04	35.89	35.15	0.74
IPPs	5.28	4.97	0.31	95.08	89.48	5.60
Software	7.45	7.71	−0.26	134.07	138.72	−4.65

Growth in real value-added with and without investment in data assets by NAICS sector 2003–2020 (%)

	Average			Cumulative		
NAICS	With data	W/o data	Δ	With data	W/o data	Δ
11	2.57	2.57	0.00	46.28	46.24	0.04
21	2.52	2.50	0.02	45.32	44.95	0.37
22	1.66	1.64	0.02	29.93	29.55	0.38
23	−0.68	−0.73	0.05	−12.22	−13.2	0.98
31-33	1.65	1.61	0.04	29.69	29.06	0.63
42	1.54	1.50	0.05	27.81	26.98	0.83
44-45	1.17	1.14	0.03	21.03	20.52	0.51
48-49	1.44	1.39	0.05	25.92	25.01	0.91
51	5.41	5.40	0.02	97.47	97.16	0.31
52	1.61	1.54	0.07	28.92	27.63	1.29
53	1.91	1.91	0.01	34.45	34.32	0.13
54	3.05	2.89	0.17	54.98	51.95	3.03
55	2.57	2.38	0.19	46.35	42.89	3.46
56	2.73	2.65	0.08	49.11	47.65	1.46
72	−0.51	−0.54	0.03	−9.2	−9.68	0.48
81	−1.15	−1.19	0.03	−20.74	−21.33	0.59

Note: The table reports average and cumulative log growth rates in real value-added by NAICS sector with and without data investment for 2003–2020. NAICS price indexes are recalculated using Törnqvist expenditure shares.

Conclusion

- We find that annual current-dollar investment in own-account data assets for the U.S. business sector grew from $84 billion in 2002 to $186 billion in 2021, which yields an average annual growth of 4.2 percent.
- Our results indicate that business sector investment in own-account data grew moderately faster than other business sector economic activity and slower than business sector investment in software.
- Identified a seemingly feasible method for identifying occupations engaged in data-related activities and for estimating the time-effort that occupations allocate to data-related activities.

Future work

- Harmonized estimates for own-account data and own-account software.
- Estimates of depreciation rates for own-account data.
- Definition boundaries between data and other related potential asset classes (e.g., A.I. / trained models)

Acknowledgments

- We would like to acknowledge Christopher Blackburn, former research economist at BEA, for developing the machine learning approach we use in the paper.
- We also the participants at the NBER-CRIW Preconference on Technology, Productivity, and Economic Growth as well as those of the 37th IARIW General Conference.

Happy to take questions!

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Works cited

- 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.
- Burning Glass Technologies (2019). *Mapping the Genome of Jobs: The Burning Glass Skills Taxonomy*. URL: https://www.burning-glass.com/research-project/skills-taxonomy.
- 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.
- 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.
- 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.
- 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.
- 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*. Valletta, Malta: ELRA, pp. 45–50. URL: http://is.muni.cz/publication/884893/en.
- U.S. Bureau of Labor Statistics (2021). *Occupational Employment Statistics: National industry-specific and by ownership*. URL: https://www.bls.gov/oes/tables.htm.