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

José Bayoán Santiago Calderón Dylan G. Rassier

CRIW Conference on Technology, Productivity, and Economic Growth
March 17, 2022

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 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

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

Similarity to closest landmark occupation

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

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 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

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

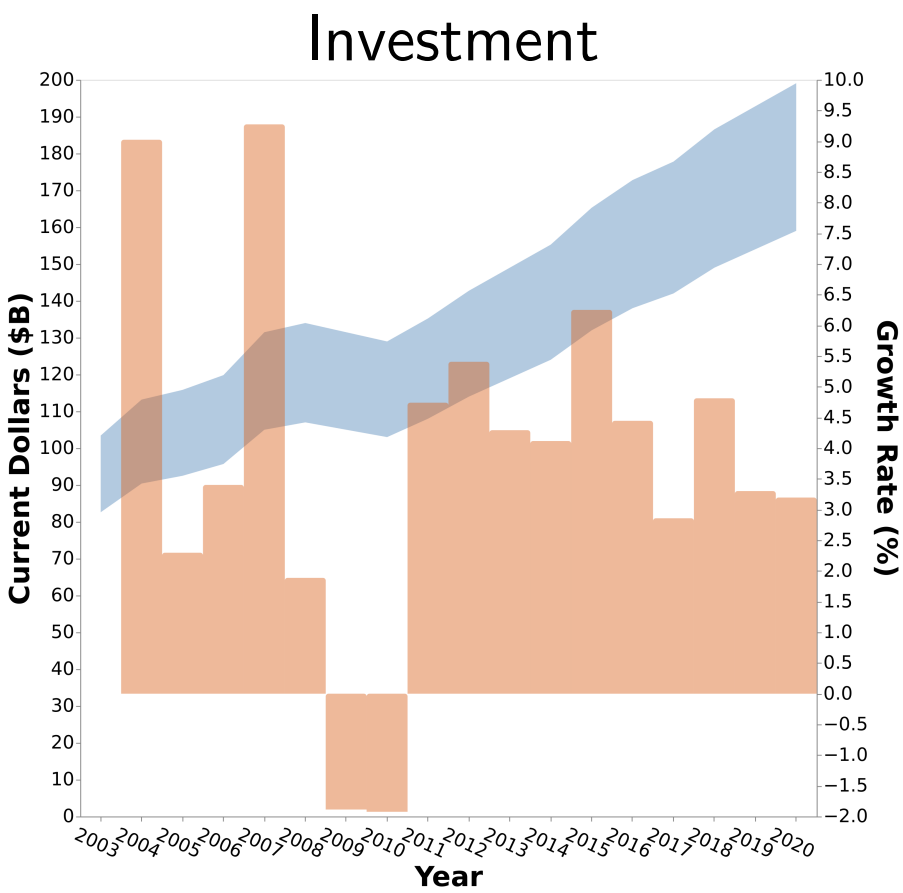
Weighted composite ratios for full sum-of-costs

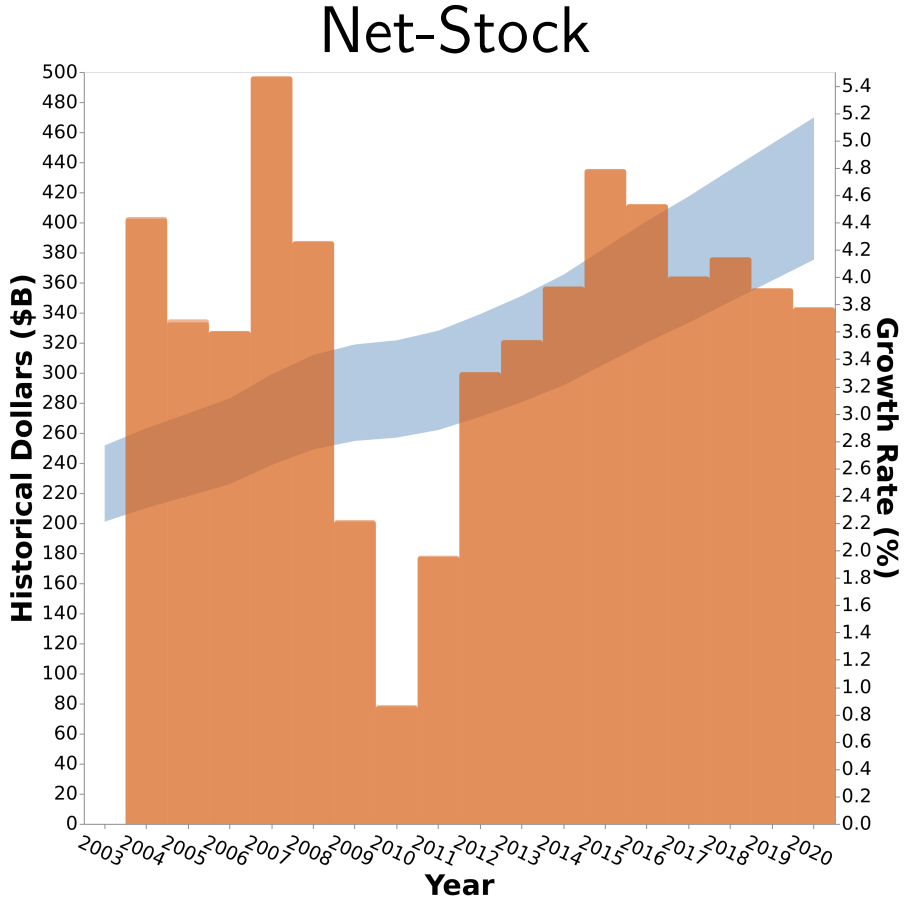
	Subsector (lower bound)		Sector (upper bound)	
	Ratio	Share	Ratio	Share
Compensation	1.17	48%	1.17	38%
Intermediate consumption	0.78	32%	1.07	35%
Consumption of fixed capital	0.18	8%	0.29	10%
Net operating surplus	0.29	12%	0.50	17%
Markup	2.42		3.03	

Effective factors applied to the wage bill

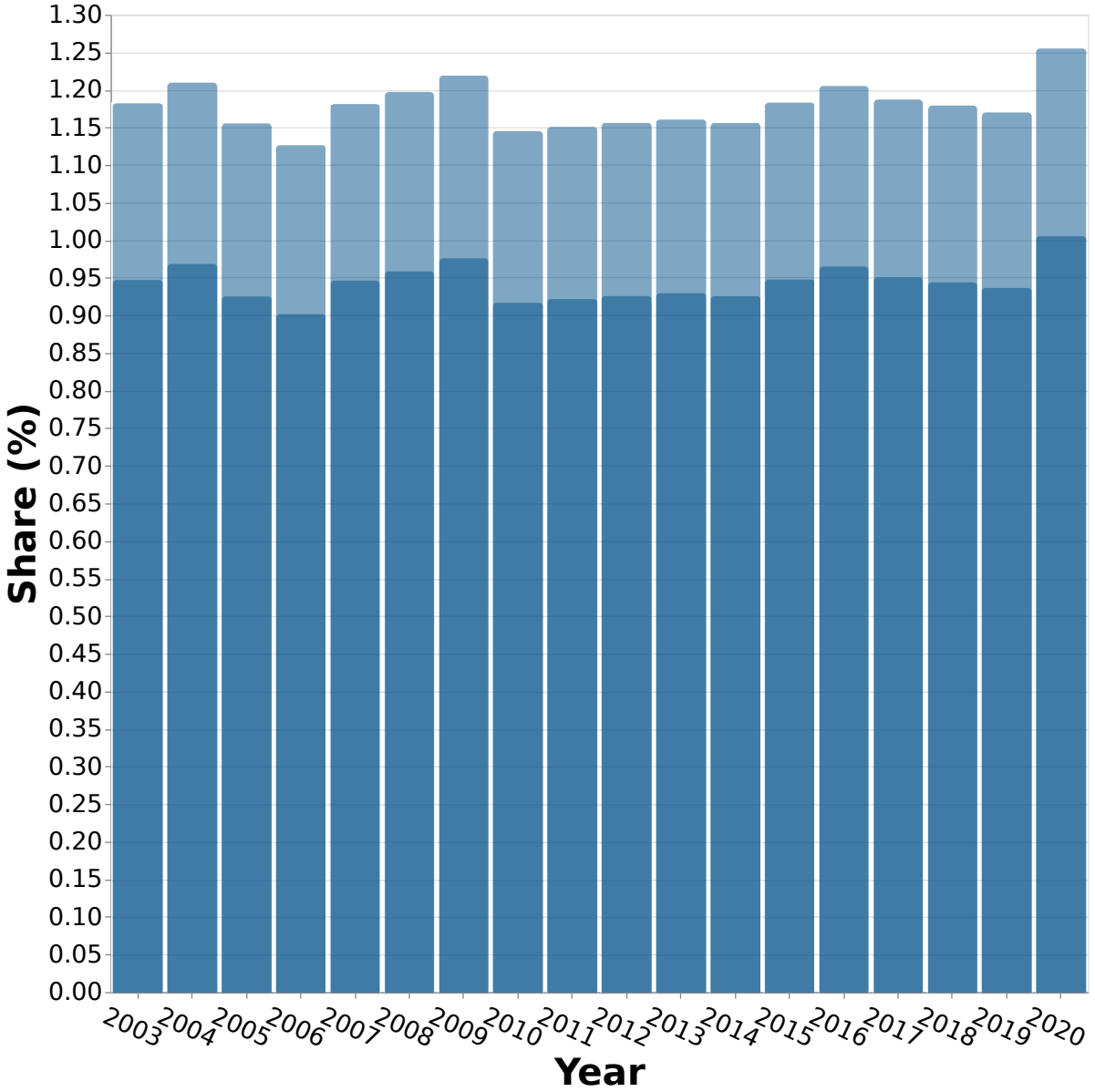
NAICS	Markup	Capital formation	R&D	Purchased data	Eff. factor
325	2.42	0.50	0.5	N/A	0.605
334	2.42	0.50	0.5	N/A	0.605
336	2.42	0.50	0.5	N/A	0.605
511	2.42	0.50	0.5	N/A	0.605
541	2.42	0.50	0.5	0.5	0.3025
All other	2.42	0.50	N/A	N/A	1.21

Annual investment and net-stock of data assets

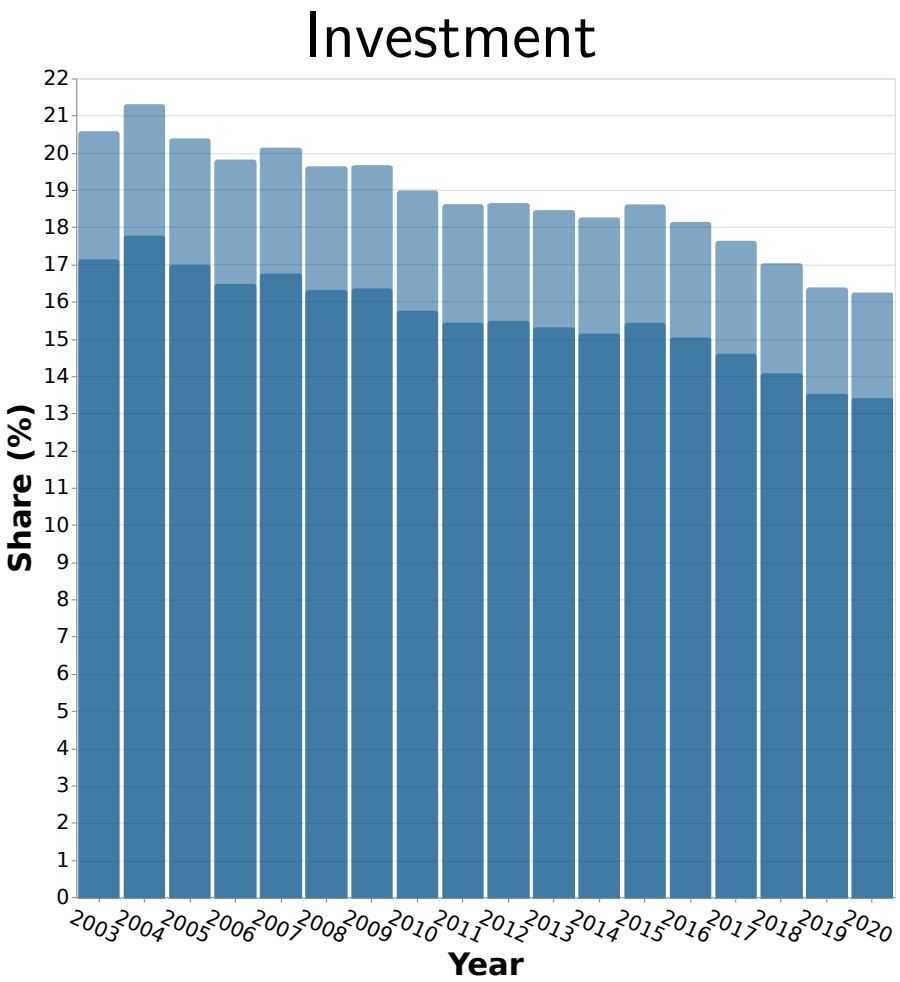


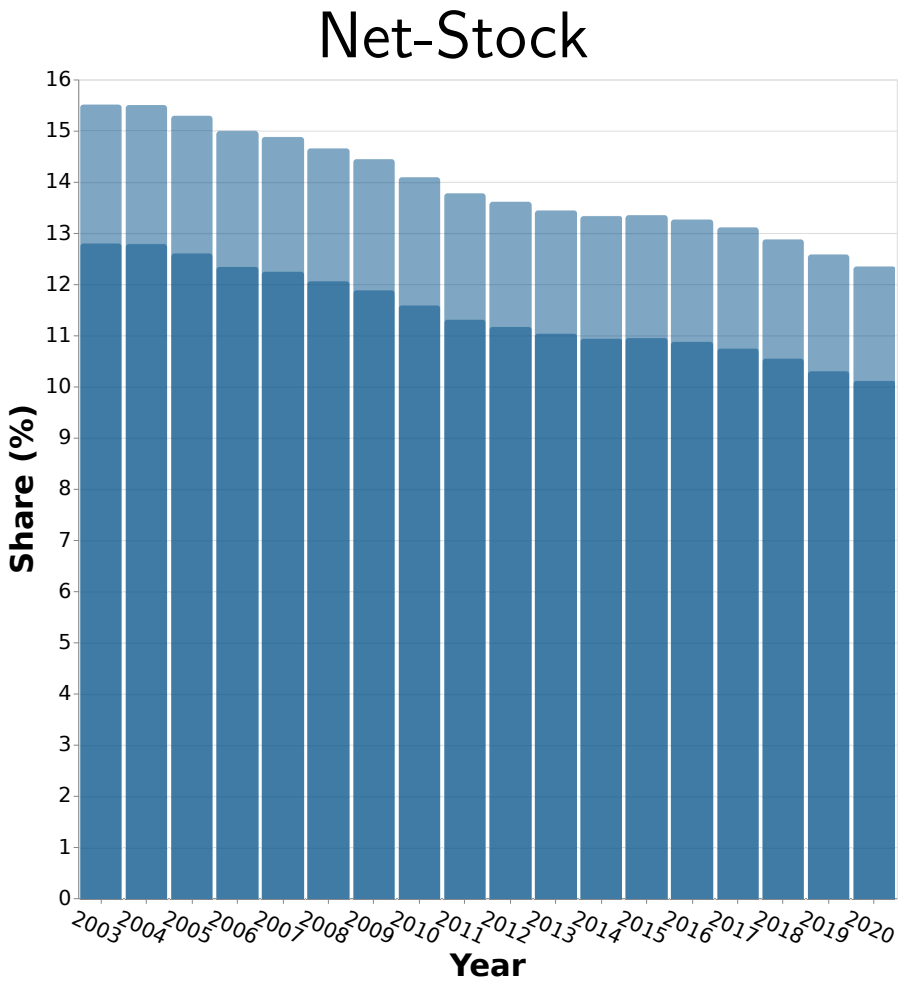


Business sector value-added

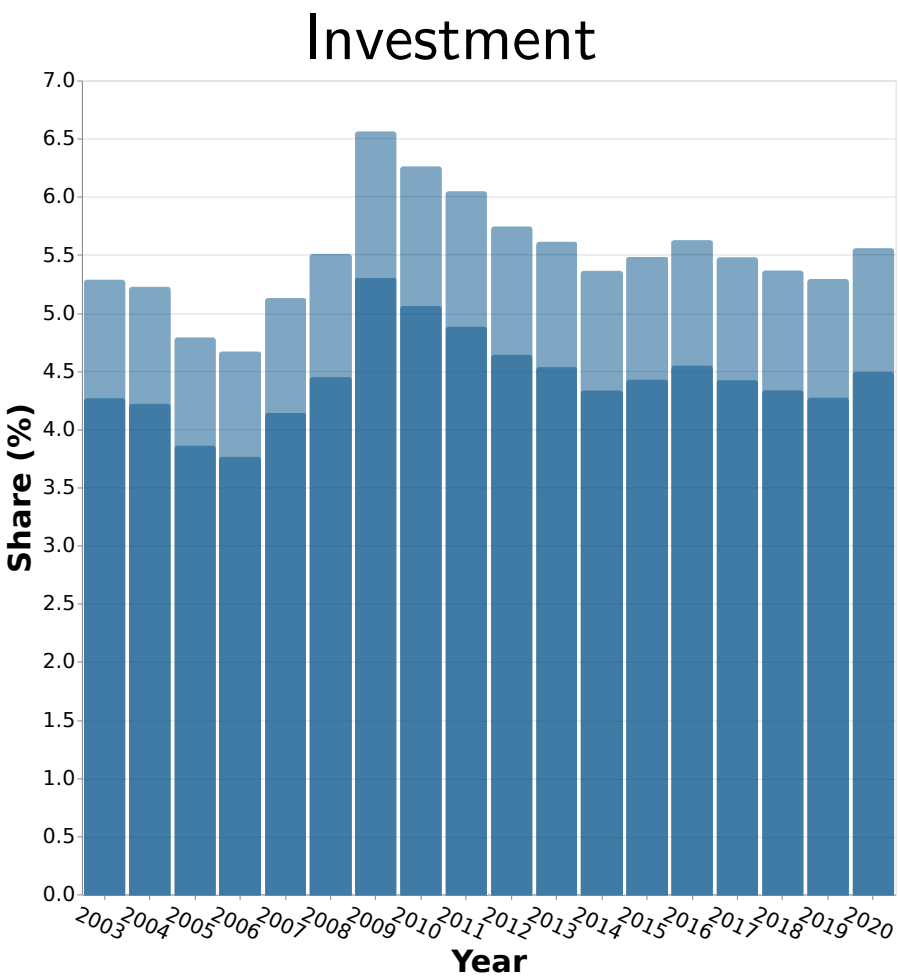


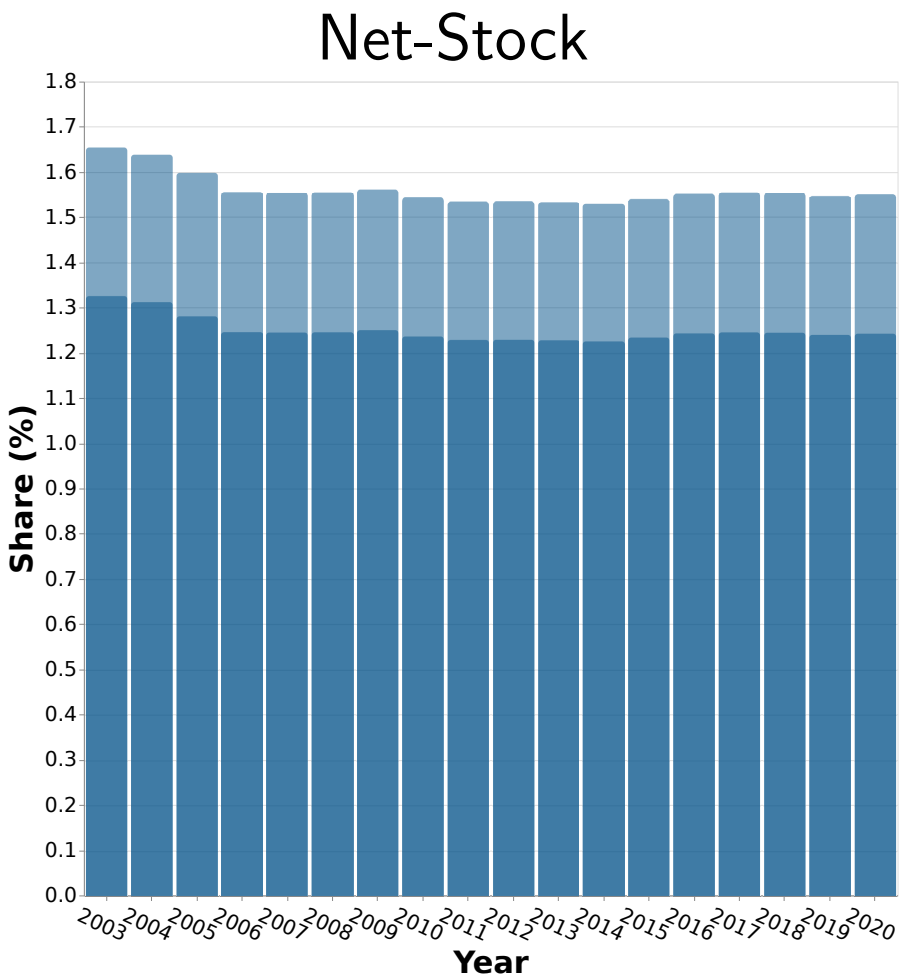
Impact on IPPs





Impact on private fixed assets





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
Data	4.07			69.13		
Value-added	1.74	1.73	0.01	27.92	27.74	0.18
IPPs	4.84	5.06	−0.21	77.50	80.90	−3.40
Software	6.38	7.58	−1.20	102.08	121.30	−19.23

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

NAICS	Description	($B)
52	Finance and Insurance	354
31-33	Manufacturing	283
54	Professional, Scientific, and Technical Services	214
56	Admin. & Support and Waste Management & Remediation Services	214
51	Information	207
55	Management of Companies and Enterprises	195
42	Wholesale Trade	194
44-45	Retail Trade	132
23	Construction	80
48-49	Transportation and Warehousing	76
53	Real Estate and Rental and Leasing	46
72	Accommodation and Food Services	31
21	Mining, Quarrying, and Oil and Gas Extraction	27
22	Utilities	27
81	Other Services (except Public Administration)	27
11	Agriculture, Forestry, Fishing and Hunting	3
	Total	2,110

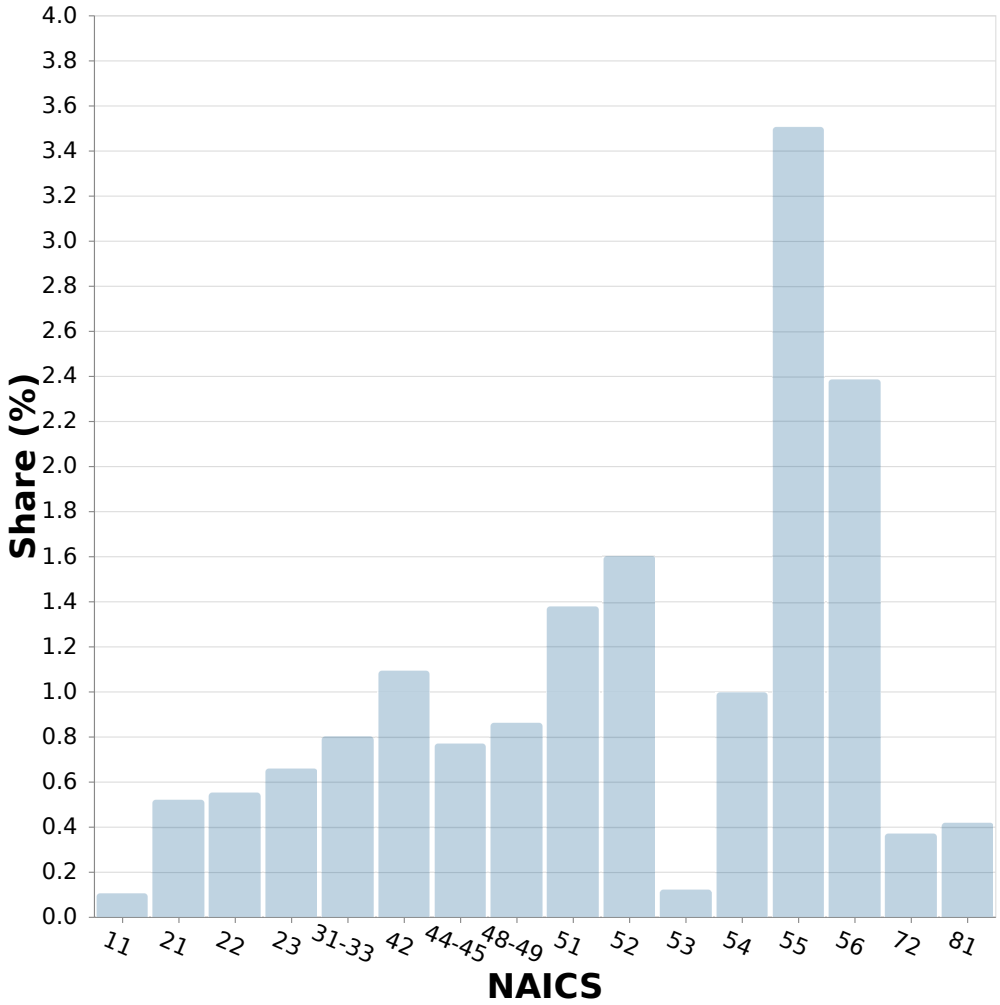
Industry current dollar average growth by NAICS sector 2004–2020

NAICS	Description	(%)
55	Management of Companies and Enterprises	6.4
21	Mining, Quarrying, and Oil and Gas Extraction	5.1
54	Professional, Scientific, and Technical Services	4.9
53	Real Estate and Rental and Leasing	4.3
48-49	Transportation and Warehousing	4.2
23	Construction	4.1
52	Finance and Insurance	4.1
51	Information	3.8
11	Agriculture, Forestry, Fishing and Hunting	3.6
56	Administrative & Support and Waste Management & Remediation Services	3.6
81	Other Services (except Public Administration)	3.1
42	Wholesale Trade	3.0
22	Utilities	2.8
72	Accommodation and Food Services	2.7
31-33	Manufacturing	2.3
44-45	Retail Trade	2.2

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

NAICS	Description	($B)
61	Educational Services	157
62	Health Care and Social Assistance	326
71	Arts, Entertainment, and Recreation	21
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations	47
	Total	551

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



Conclusion

- We find that annual current-dollar investment in own-account data assets for the U.S. business sector grew from $82.6 billion in 2003 to $159.5 billion in 2020, which yields an average annual growth of 3.9 percent.
- Our results indicate that business sector investment in own-account data grew marginally faster than other business sector economic activity and slower than business sector investment in other IPPs.
- 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

- Develop an input-cost own-account data price index
- Explore applying the methodology to other potential IPPs such as own-account software

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 thank the participants at the NBER-CRIW Preconference on Technology, Productivity, and Economic Growth for early comments.

Happy to take questions!



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.