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

José Bayoán Santiago Calderón
Dylan G. Rassier



CRIW Pre-Conference on Technology, Productivity, and Economic Growth
Virtual Meeting
September 30, 2021

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.

National Accounts Omit Data Assets

- Databases are within scope of the *SNA* boundaries
  - Subcategory of software
  - Exclude value of data in own-account databases
  - Include value of data in market purchases of databases

- Next revision of the *SNA* is expected to include own-account data in capital formation
  - Subcategory of software or databases
  - Sum-of-cost = labor cost + capital cost + intermediate inputs

- Addressing measurement issues for own-account IPPs
  - Some may be unique to data assets, but most are not

Measurement Issues

- Scope of capital formation (OECD 2021)
  - What costs to include?
- Double counting
  - Overlap with other IPPs and purchased data assets?
- Prices and depreciation
  - Are they different than software and databases?
- Capital cost component in sum-of-costs
  - Does capital play a much bigger role than labor?

BEA Measurement of Data as an Asset

- Our method captures resource costs
  - Consistent with other U.S. own-account IPPs
  - Data-related activities
    - Collecting, recording, organizing, storing, processing, distributing

- Components of resource costs
  - Labor cost
    - Identify occupations and time use using a machine learning model and job postings from Burning Glass Technologies
    - Industry-level estimates
  - Capital cost and intermediate inputs
    - Use data from BEA industry accounts
    - Assume resource use for capital formation in data is the same as other products for a given industry

Tackling Labor Costs Estimation for Data Activities

Time-use Labor Costs Estimation

$$\hat{E}_\tau = \sum_{\omega \in \Omega} \tau_\omega W_\omega H_\omega$$

What occupations work with data?

Inclusion based on tasks performed

Ad-hoc rather than data-driven

How often do they engage with data?

Time-use factors rarely observed

50% estimate commonly assumed

We use machine learning techniques to estimate Ω and τ_ω from online job postings

Labor costs estimation using online job text

Online job postings from Burning Glass to estimate

$$p_\omega = \frac{l_\omega}{L_\omega} \equiv \text{Fraction of workers in } \omega \text{ engaged in data-related tasks}$$

$$\sum_{i=1}^{L_\omega} \mathbb{I}(y_{i,\omega} = 1) \equiv \text{Data-related skills from Burning Glass}$$

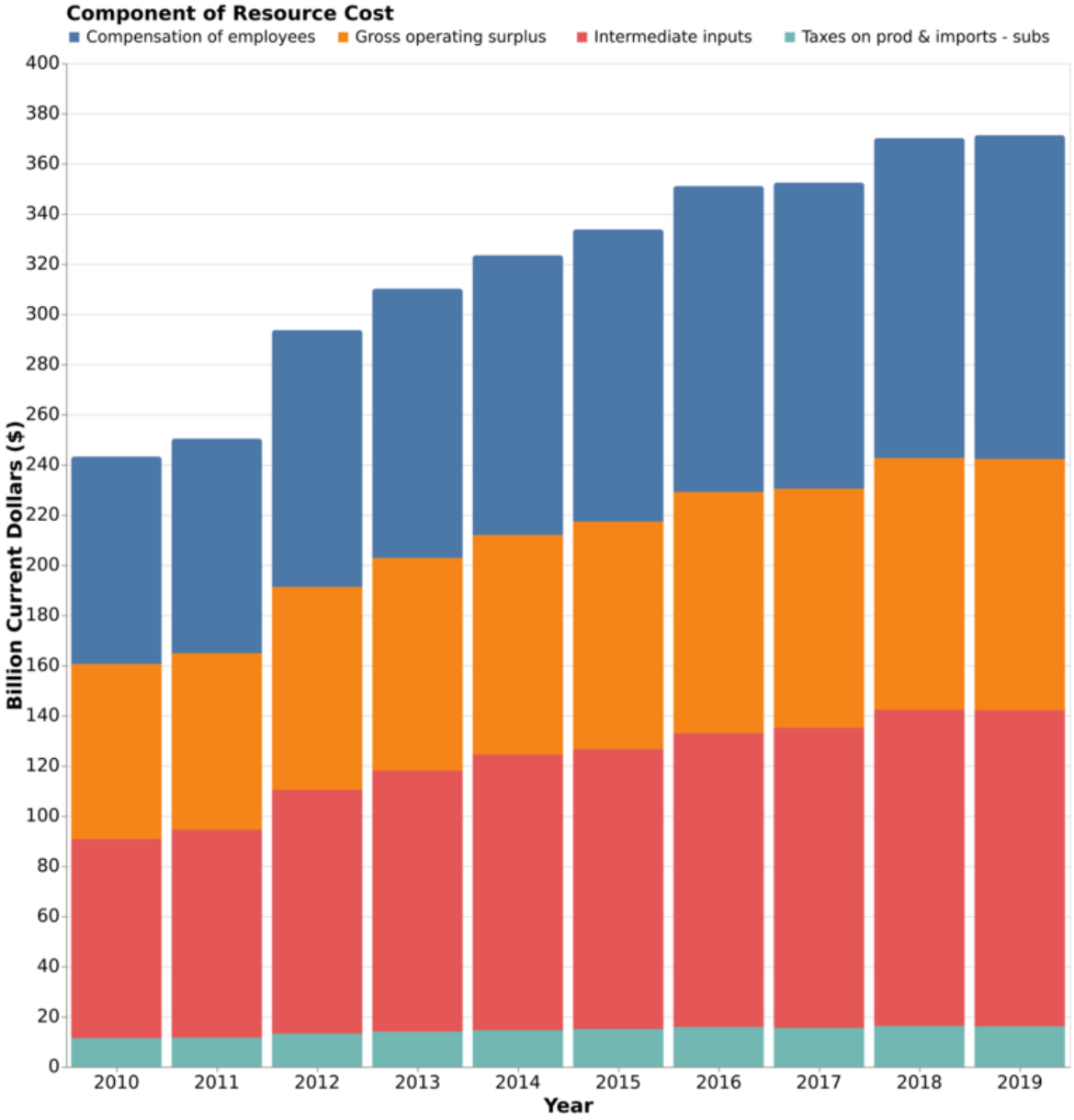
Proxy time-use using distance to "landmark" occupations

$$\tau_\omega = \frac{h_\omega / l_\omega}{H_\omega / L_\omega} p_\omega \approx \min\left(d_{\omega,1}, d_{\omega,2}, \ldots, d_{\omega,L}\right) p_\omega$$

Construct labor costs estimates for data activities

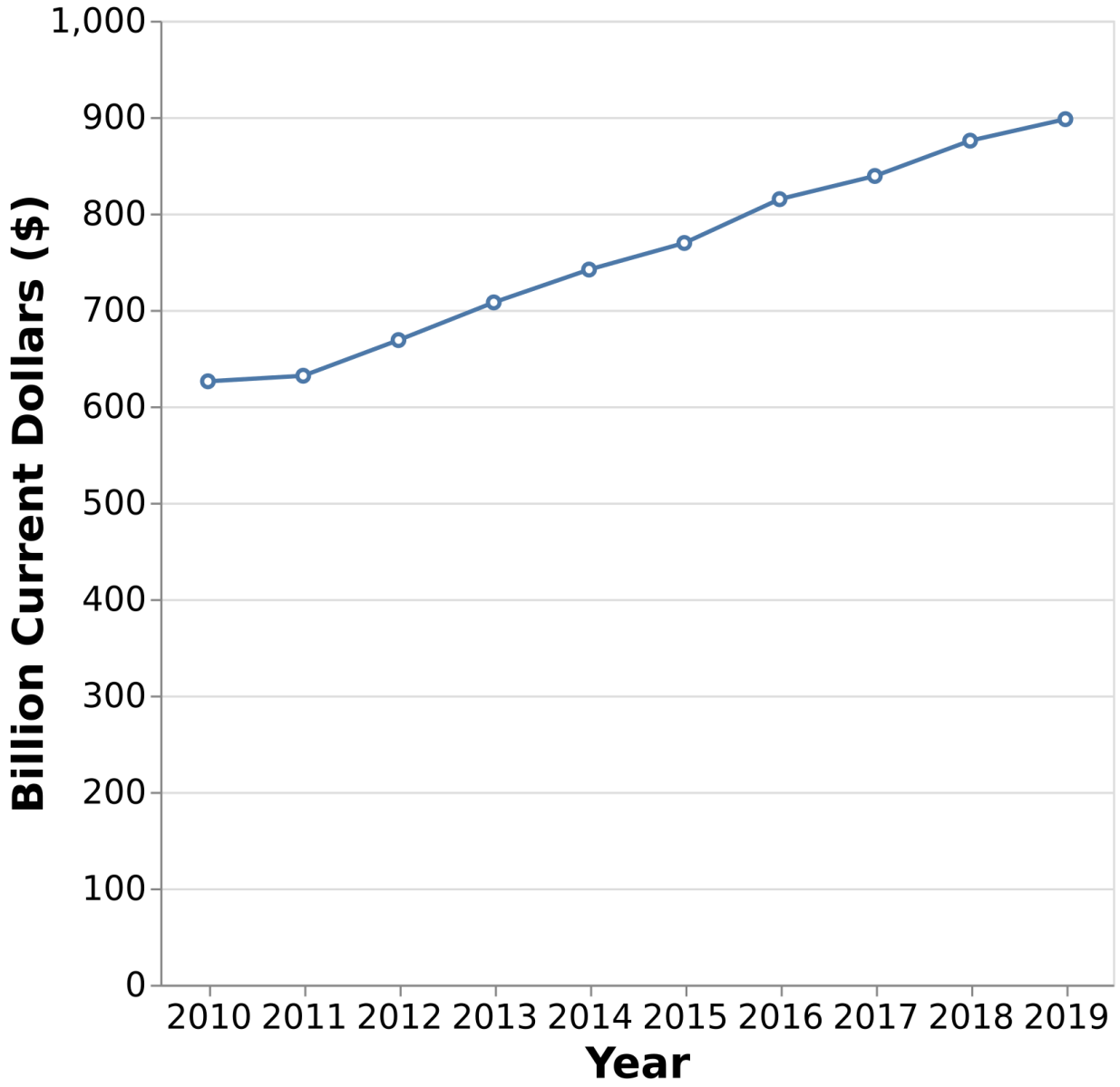
$$E_\tau \approx \sum_{\omega \in \Omega} (1 - d_\omega^*) p_\omega W_\omega H_\omega$$

Preliminary Estimates: Annual Flows



Average annual growth rate
(2010 – 2019): 4.5%
(2012 – 2019): 3.28%

Preliminary Estimates: Annual Stocks



Average annual growth rate
(2010 – 2019): 3.9%
(2012 – 2019): 4.1%

Outline of Paper - Completed

- Background for Measurement
- Time-Use Model
  - Assessment of Burning Glass Technologies data
    - Job text
    - Occupation code (SOC/OCC)
    - Skills
  - Model trained on nationally representative private sector
  - Time invariant estimates (based on 2010 – 2019 data)
  - Over 550 occupational groups
- Resource Cost Methodology
  - Blow-up factors by industry from effective wages and salaries to resource cost (intermediate inputs exclude materials)

Outline of Paper – In Progress

- Results for the Private Sector
  - Implications for production: aggregate and industry
    - Need to confirm overlap with R&D and software is resolved
    - Need to confirm overlap with purchased data is resolved: NAICS 5191, 5416, 5614 include data aggregators/brokers
      - Capitalize both data originals and data copies?
  - Implications for productivity: aggregate and industry
    - Recalculate labor productivity
    - Recalculate MFP using KLEMS framework
      - Growth in new output ≈ growth in new inputs?
- Robustness Checks and Validation
  - Nationally representative blow-up factors for capital costs
  - Validating Burning Glass SOC codes
  - Expanding employment and wages data for occupations pre-2012