Valuing the Data Economy using Machine Learning and Online Job Postings

Christopher J. Blackburn

NEA Research Group

March 16th, 2021

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

The views expressed in this presentation are those of the author and not necessarily those of BEA or the Department of Commerce

Tackling Labor Costs Estimation for Data Activities

Time-use Labor Costs Estimation

$$L_\tau = \sum_{\omega\in\Omega} \tau_\omega W_\omega H_\omega$$

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

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Online job postings from Burning Glass to estimate

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

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

Proxy time-use using distance to “landmark” occupations

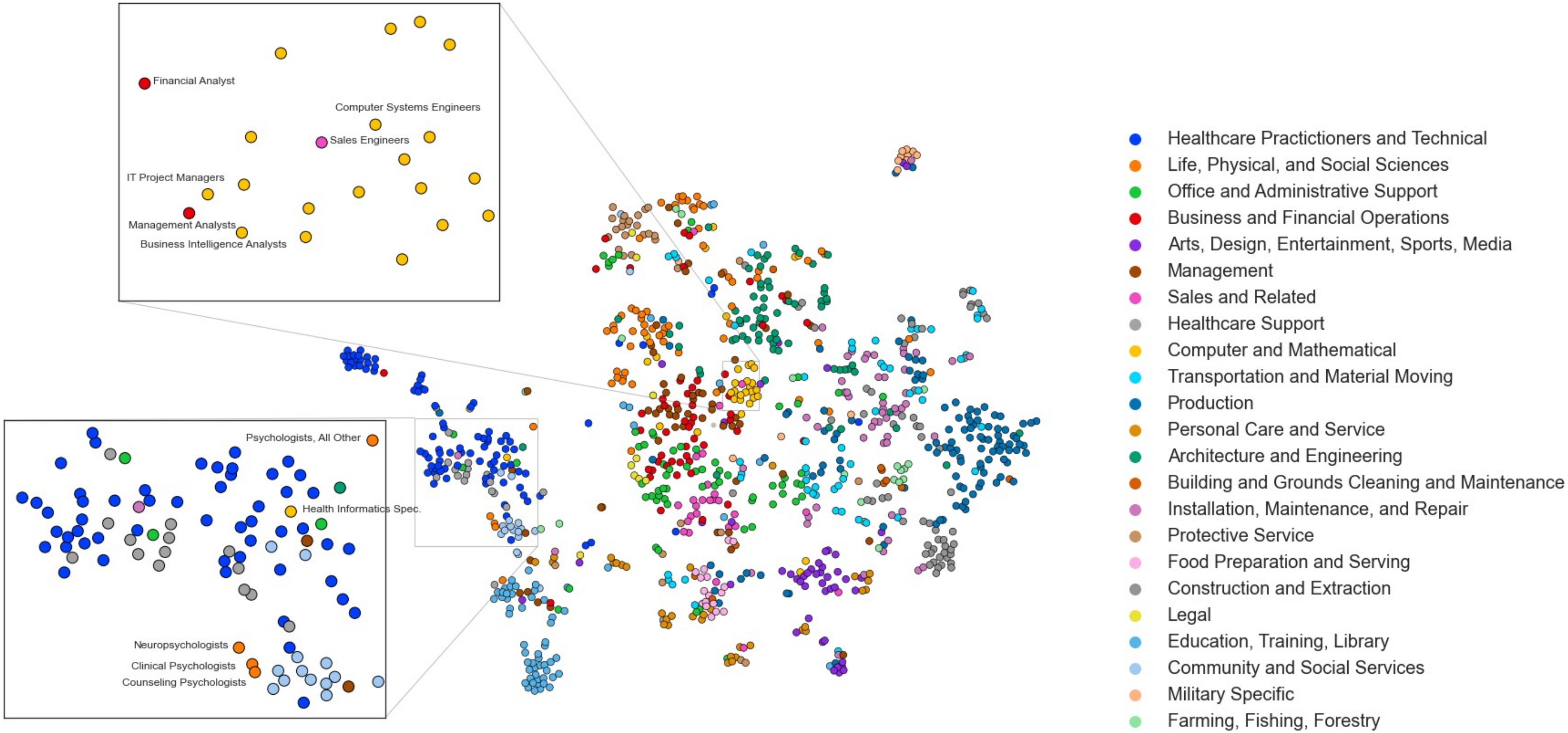
$$\tau_\omega = \frac{h_\omega / l_\omega}{H_\omega / L_\omega} p_\omega \approx \min\left(d_{\omega,1}, d_{\omega,2}, \dots, d_{\omega,L}\right) p_\omega$$

Construct labor costs estimates for data activities

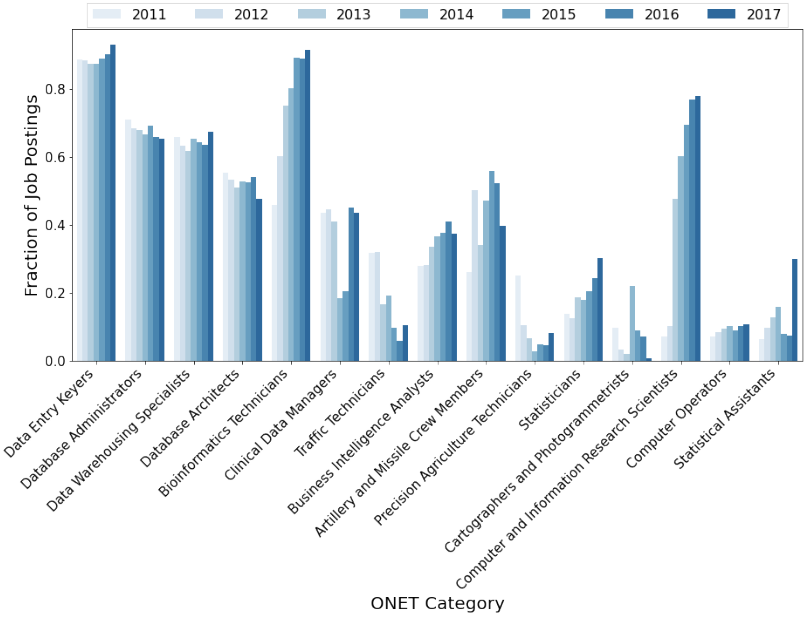
$$E_\tau \approx \sum_{\omega \in \Omega} (1 - d^*_\omega) p_\omega W_\omega H_\omega$$

Landmark Occupation Vector Space

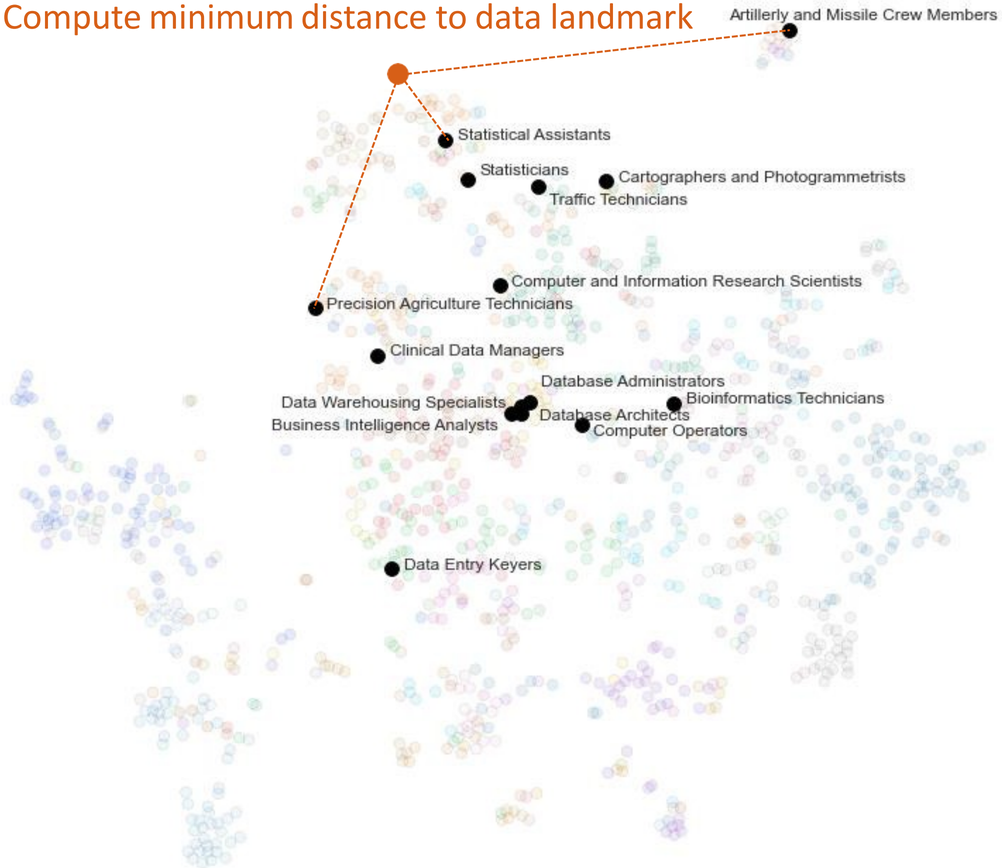
bea
Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE



Distance to Landmark “Data” Occupations



Compute minimum distance to data landmark



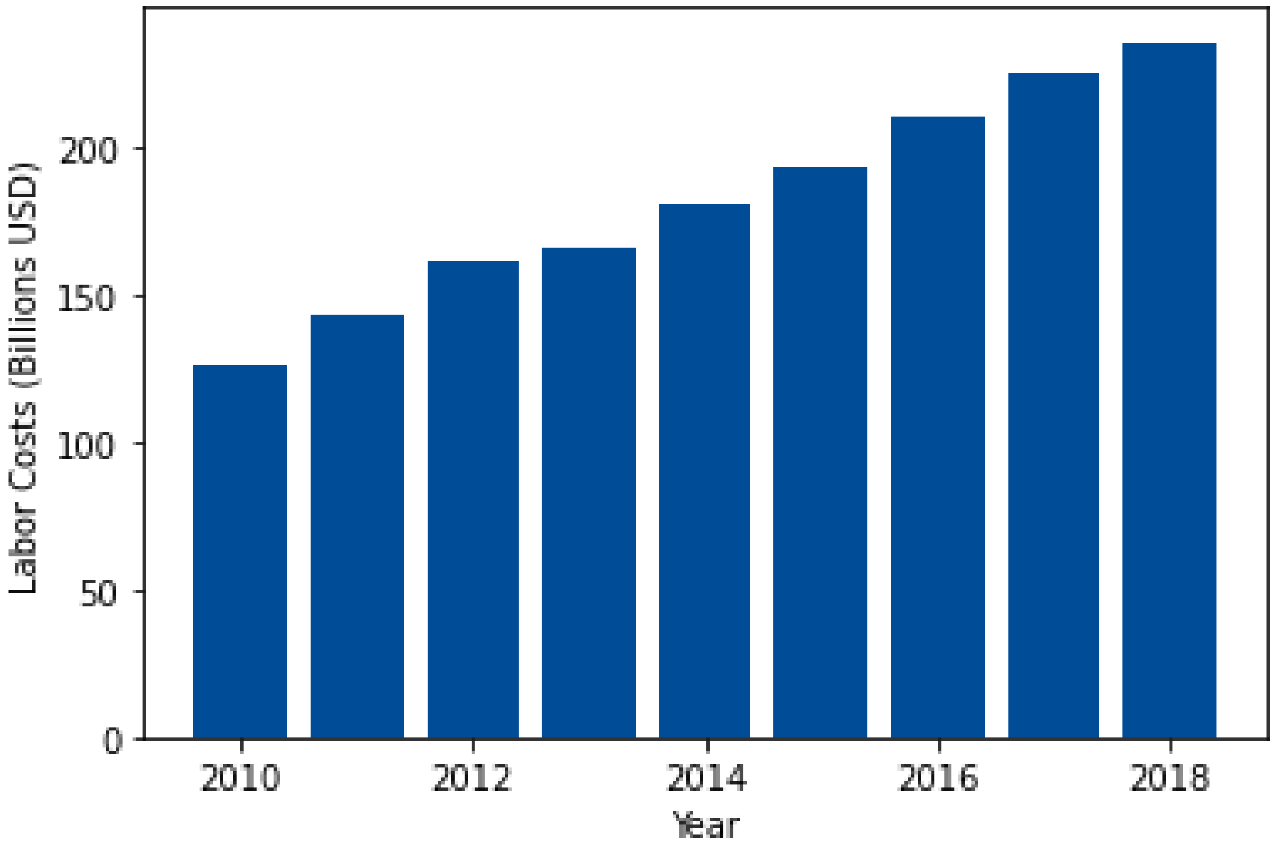
Distance function

$$d_{i,d} = 1 - \cos\left(\theta_{i,d}\right) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

$$E_\tau \approx \sum_{\omega \in \Omega} \cos(\theta^*_\omega)\, p_\omega W_\omega H_\omega$$

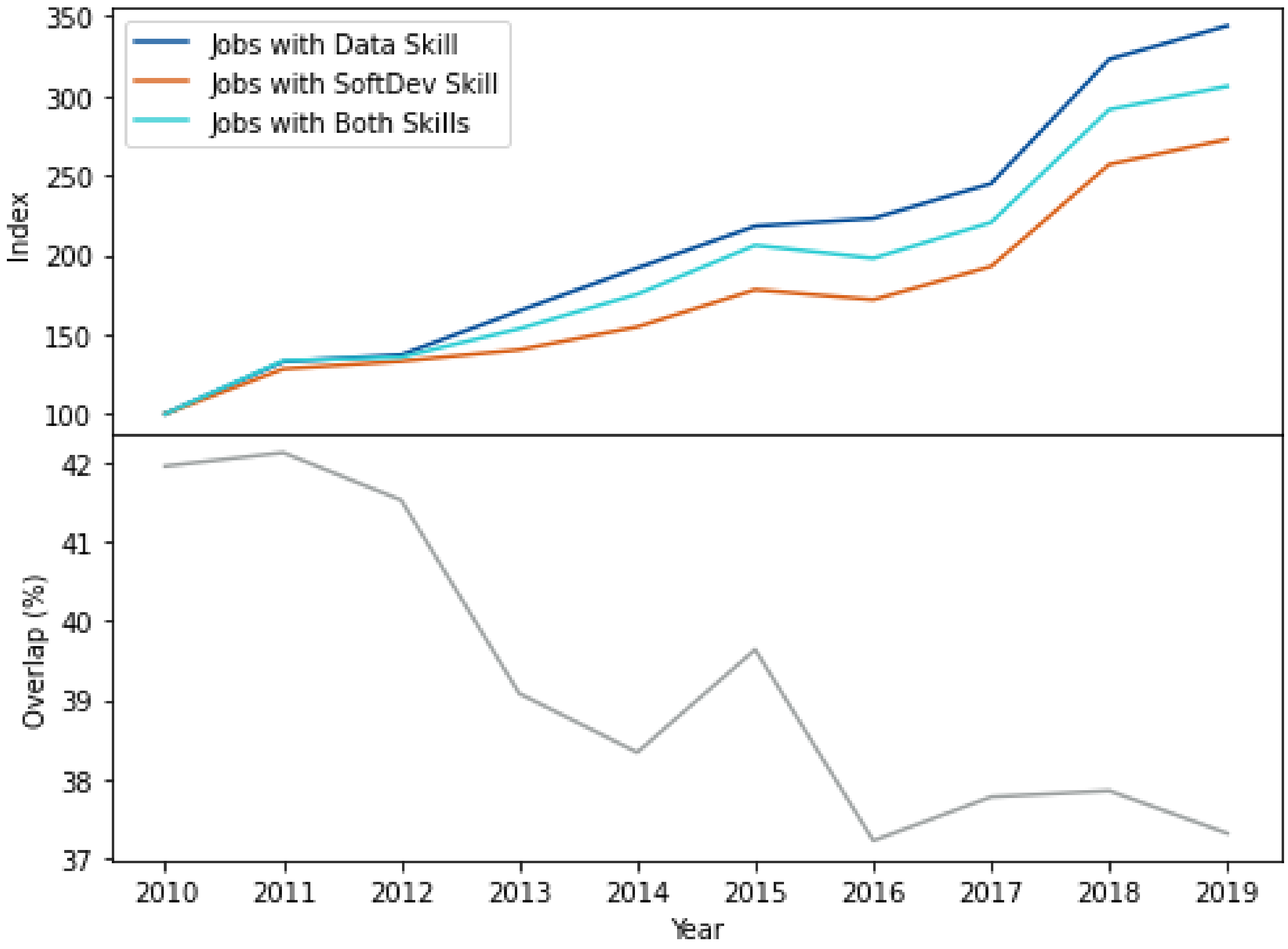
Labor Costs Estimates for Data-Related Activities

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



Assessing Overlap with Software Development

38% of data-interfacing jobs have software development skill



Conclusions and Future Work

Combine ML with online job postings to estimate labor costs of data activities

…Annual spending ranges depending on the technique
…Similarity adjusted spending estimates come in around $200 billion annually

Future work aims to address overlap between data, R&D, and software investment

…National accounts may already capture spending on data, but how much?
…Preliminary estimates suggest around 38% of SoftDev jobs overlap

Combining estimates using similar NLP techniques could yield more reliable estimate

…Many document embedding/similarity approaches exist, e.g. LDA, WMD
…Ensemble approaches usually yield more reliable estimators

Data is ubiquitous, but not nearly as exciting as popular anecdotes suggest

…Think data collected from oil changes, customer call records
…Data is everywhere, but will it show up in the productivity statistics?