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

An Emerging Data Economy

The collection, analysis, and distribution of data is a hallmark of the modern economy

In terms of spending, how large is the data economy?

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

Determining Relevant Occupations

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Is a job working with data? Check the job posting!

Job title: Indicates specialization, e.g., Data Engineer, Data Scientist

Pre-Sales Data Engineer
TimeXtender 4 reviews - Bellevue, WA 98009
Apply Now

Love Your Work: BE PART OF OUR BUSINESS GROWTH IN NORTH AMERICA

TimeXtender is a privately held company with dual headquarters in Aarhus, Denmark and Bellevue, Washington, US. We are a global market leader in Data Management and were named a Microsoft Data Estate Modernization Partner of the Year finalist in 2019.

At TimeXtender, we do everything BECAUSE TIME MATTERS. We value diversity and bringing together talented people together who love what they do. We seek enthusiastic candidates who take responsibility and are eager to develop themselves in an international environment.

We are currently looking for a Pre-Sales Data Engineer to work closely with the US commercial team to enable partners and their customers to adopt, master & champion our proprietary data management solution, the Discovery Hub® platform. In this role you'll bring the technical know-how into the partner recruitment process and provide ongoing support of US-based partners & customers, enabling them to deliver analytics solutions with the Discovery Hub platform.

Your Key Qualifications:

- 1-3 years data engineering (data modelling/data warehouse) experience.
- Experience with SQL Server/ SSIS / Azure SQL preferred.
- Strong understanding of ERP/CRM applications & related business processes.
- Proven background with excellent performance working under pressure.
- Highly self-motivated with a desire to provide solutions.

Your Responsibilities:

- Effectively demonstrate our software platform's capabilities and business value to prospective partners & customers.
- Support the commercial & marketing teams by developing and delivering product demos and presentations in remote, on-site, and conference settings.
- Develop technical support & training content such as procedures, user guides, videos, and online training content.
- Provide ongoing technical account management to a portfolio of partners, ensuring successful delivery of projects & continued identification of new opportunities.
- Enable product sales & assist/teach partners through the delivery of small customer projects and "Proof-of-Concepts."

What You Get:

- The ability to grow and mold your role to support your strengths & interests, as well as your personal & professional goals
- Opportunity to work in a high-growth setting with an international team
- A uniquely mindful, challenging & innovative company culture
- A strong network of support from your co-workers (think "work-family")
- Flexibility to balance work & life with a high priority on holistic wellbeing
- 7-minute daily recharge breaks, mindful minutes before each meeting, and optional Friday morning meditation sessions
- Friday morning team breakfasts & periodic team lunches

Job duties: Tasks performed, e.g., Data Analysis, Modeling

Job experience: Necessary skills, e.g., SQL, Machine learning

A (Naïve) Classification Rule

Job Postings

Human Reader

Classifications

Yes

Yes

No

Data-Related Title?

No → **Data-Related Text?**

- No → No
- Yes → Yes

Yes → Yes

Machine Reader

Is the machine reader too naïve?

Statement 1: The data entry clerk inputs data into a database.

Data-related title: Yes Data-related text: Yes Data Classification: Yes

Statement 2: The sales representative enters customer data into a computer.

Data-related title: No Data-related text: Yes Data Classification: Yes

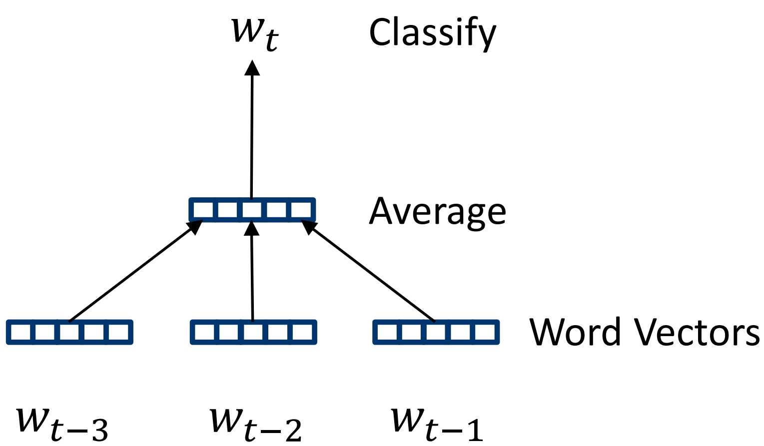
Statement 3: Applicant's data will not be shared with third-parties.

Data-related title: No Data-related text: Yes Data Classification: Yes

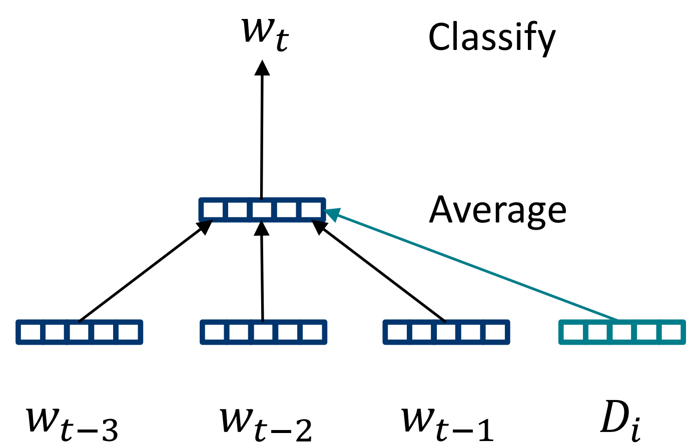
"Statements 1 and 2 are semantically identical, and Statement 3 is not relevant for the classification."

Semantic Similarity and Document Embeddings

Word2Vec (Mikolov et al. 2013)



Doc2Vec (Le and Mikolov 2014)



[0.2, 0.7, …, 0.8, 1.2, 3.1]

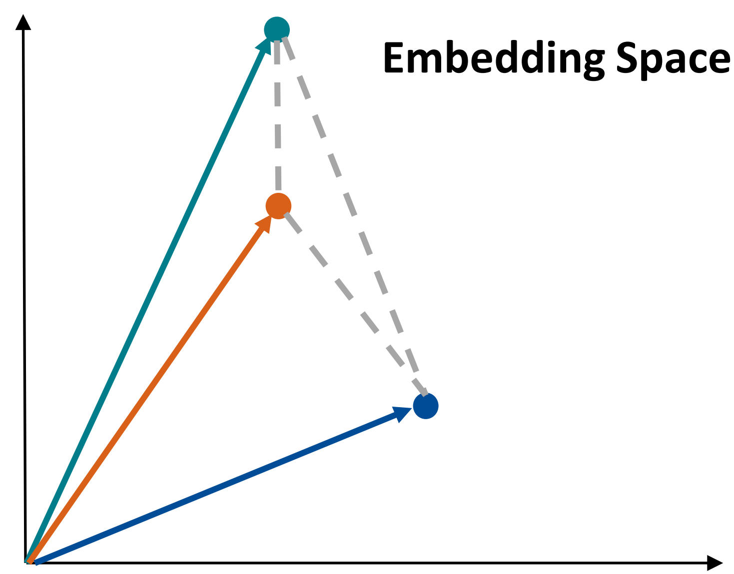
Statement 1: The data entry clerk inputs data into a database.

[0.8, 1.2, …, 0.1, 2.6, 0.7]

Statement 2: The sales representative enters customer data into a computer.

[1.8, 0.2, …, 3.5, 1.1, 2.4]

Statement 3: Applicant's data will not be shared with third-parties.



Labor costs estimation using online job text

Online job postings from Burning Glass to estimate

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

$$\sum_{i=1}^{L_\omega} \mathbb{I}\left(y_{i,\omega} = 1\right) \equiv \text{Output of naïve dictionary-based classifier}$$

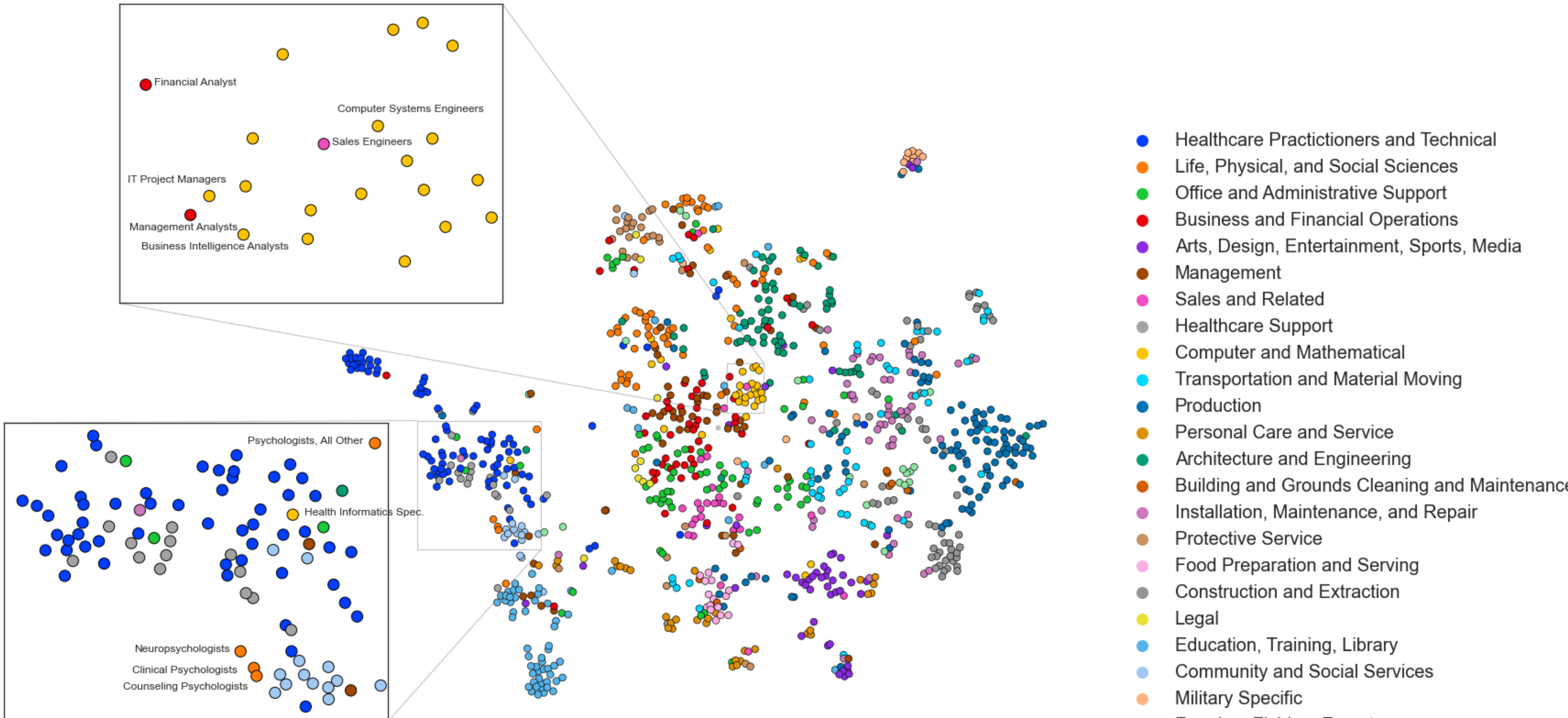
Proxy time-use using distance to “landmark” occupations

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

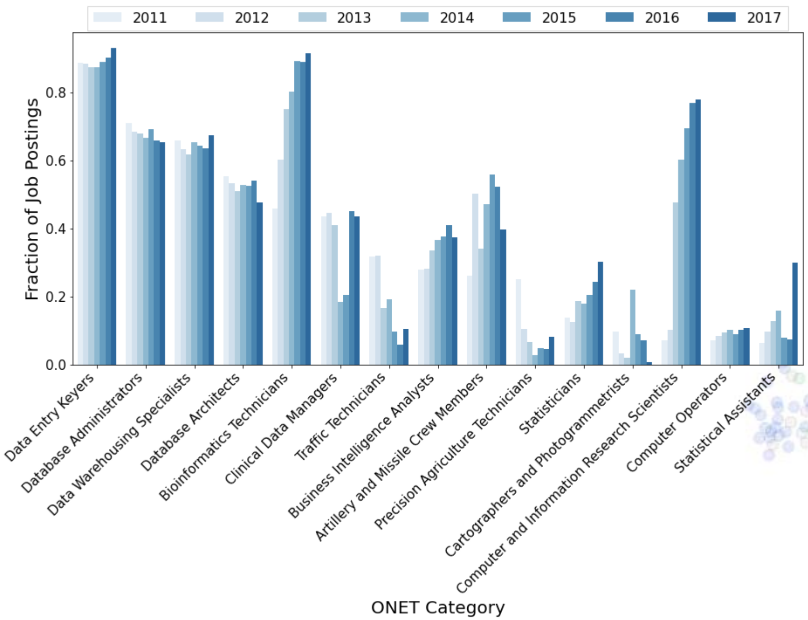
Construct labor costs estimates for data activities

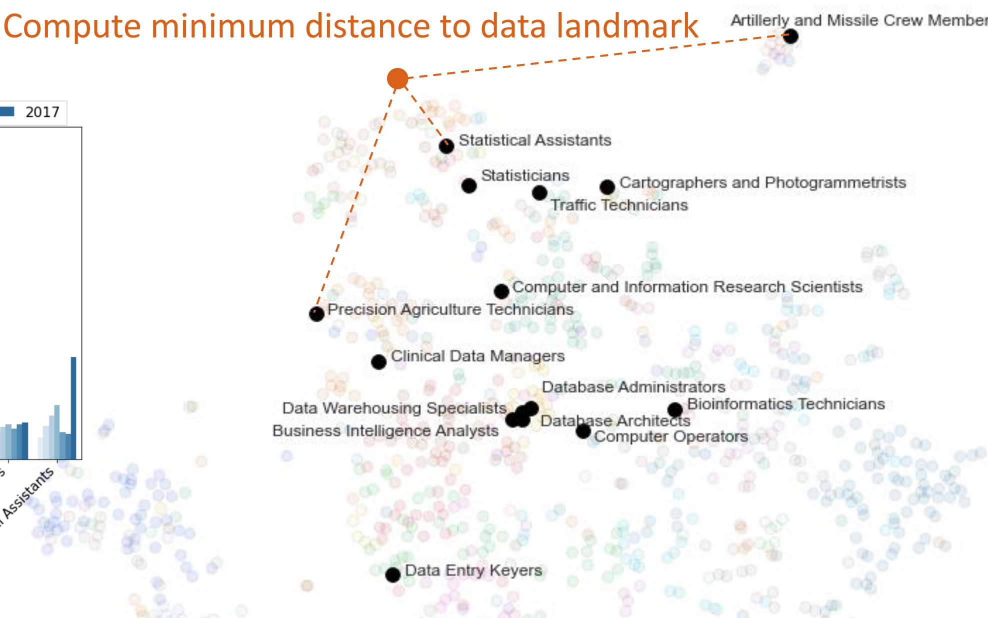
$$E_\tau \approx \sum_{\omega \in \Omega} (1 - d^*_\omega) p_\omega W_\omega H_\omega$$

Landmark Occupation Vector Space (LOVeS)



Distance to Landmark “Data” Occupations

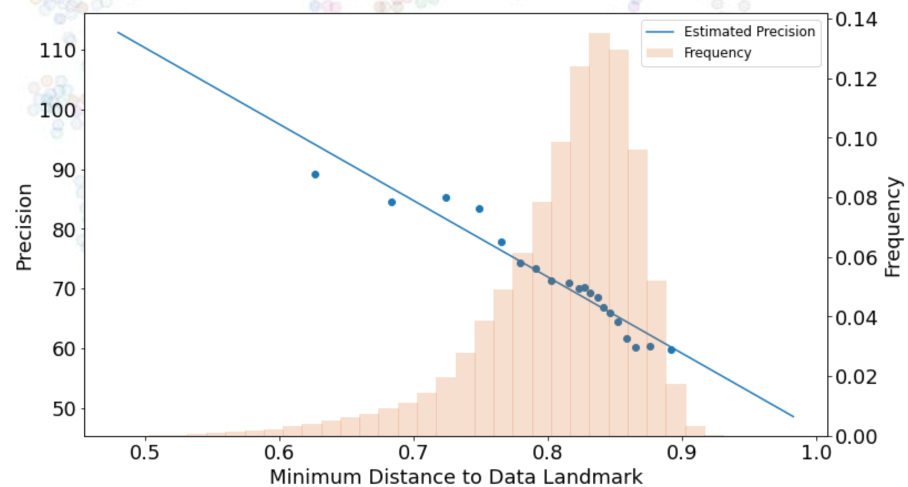




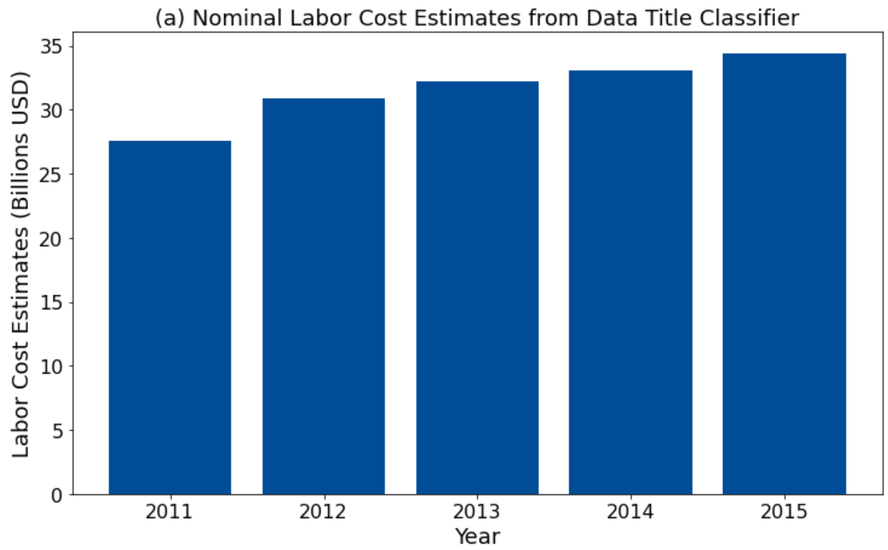
Distance function

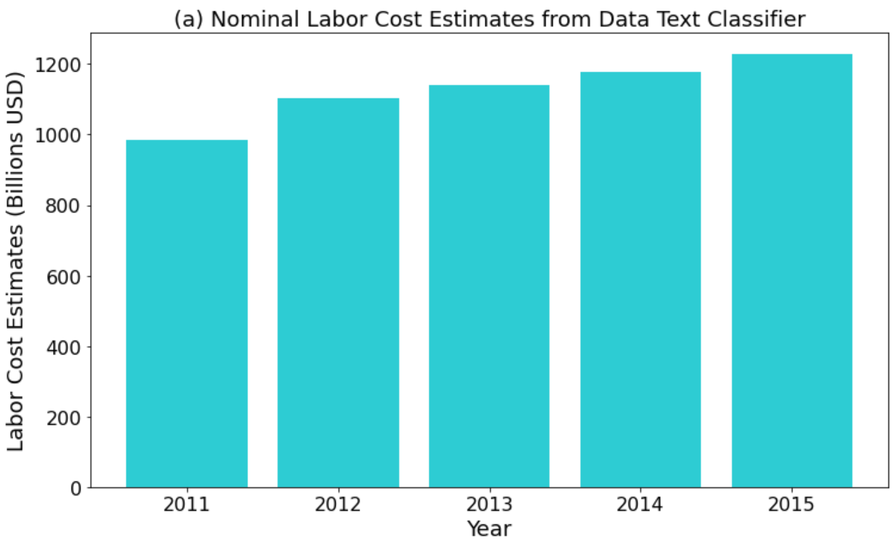
$$d_{i,d} = 1 - \cos(\theta_{i,d}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

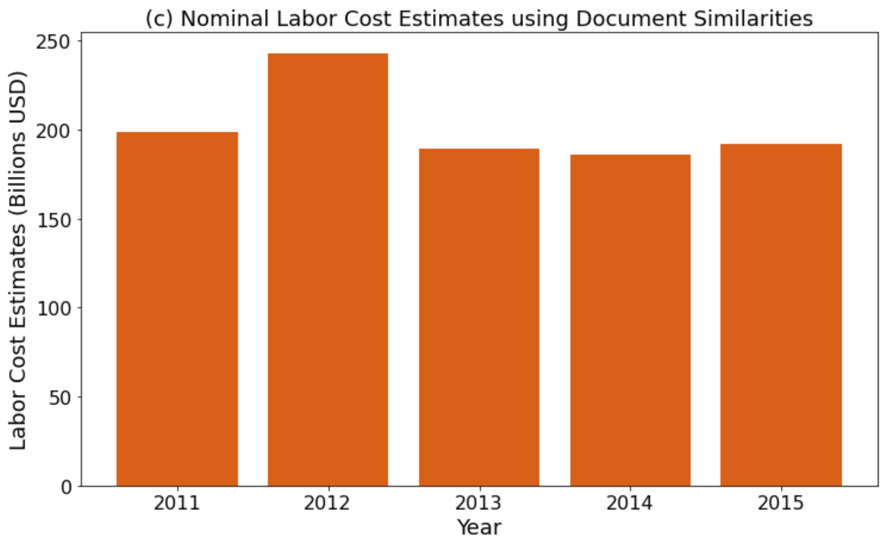
$$E_\tau \approx \sum_{\omega \in \Omega} \cos(\theta^*_\omega)\, p_\omega W_\omega H_\omega$$



Labor Costs Estimates for Data-Related Activities







Data Title Mean Estimate: $31 Billion

Data Text Mean Estimate: $1.1 Trillion

Doc2Vec Mean Estimate: $200 Billion

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?

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?

Some Important Caveats

Our method assumes tasks within job postings reflect underlying composition

…Emerging tasks and responsibilities will be overrepresented

…Can potentially overestimate p_ω relative to true composition

Validity of our estimate relies on representativeness of job posting data

…Some occupations over/underrepresented $\left(\frac{L_\omega^T}{L} = \alpha_\omega \frac{L_\omega^B}{L}\right)$

…Ratios might help but bias could still exist $\left(\frac{L_{d,\omega}^T}{L_\omega^T} = \frac{\beta_{d,\omega}}{\alpha_\omega} \frac{L_{d,\omega}^B}{L_\omega}\right)$