The Changing Economics of Knowledge Production

Simona Abis and Laura Veldkamp
Columbia University

August 2022
SHoF Annual Conference: AI & Machine Learning in Finance

Motivation

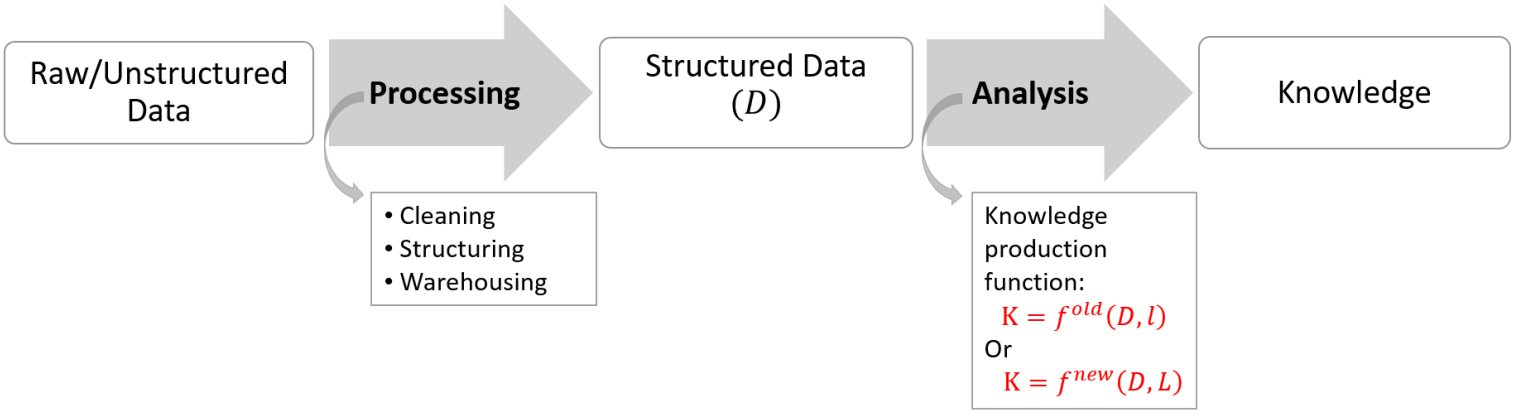
- Claim:
  "Big data technologies are the industrialization of knowledge production."
  - Is this claim accurate? Let's measure it like industrialization and see.

- Key feature of industrialization:
  It changed the relative intensity of labor and capital (data).
  - Is AI doing the same?
  - How much is AI changing the labor intensity of knowledge production?
  - This matters for employment / labor income share / firm size and competition...

- Challenge:
  We are in the early stages of adoption.

- Objective:
  Quantify the impact of big data technologies on an early adopting industry.
  - Investment Management is a good lab:
    - It's an early adopter.
    - It's a knowledge industry.

This Paper



- A Model for Measurement:
  We model knowledge production based on the above representation.
  - The objective is to quantify the parameters that regulate knowledge production for the old and new technologies.

- Measurement:
  - Measure how many Data Management and Analysis workers (old and new tech) each firm hired (job postings and BLS data).
  - Identify how much workers of each type are paid (crowdsourced salary data).
  - Structurally estimate the parameters regulating data processing and the two knowledge production functions.

Preview of Results

- Main Result:
  AI has significantly raised the productivity of analyzing larger data sets.
  - Labor share fell from 18% to 13%.
  - This change is substantial, if compared to similar estimates for the industrial revolution.

- How Does Labor Share Fall?
  A fall in the labor share could mean fewer workers, or could mean more data. Which was it?
  - Labor stock has been increasing steadily.
  - Data value rose 39% between 2015 and 2018.
  - Firms are becoming more productive at using data.

Related Literature

- Structural estimation to value intangible assets:
  E.g., Eisfeldt and Papanikolaou (2013); Peters and Taylor (2017); Belo, Gala, Salomao, and Vitorino (2021).
  - We contribute a hiring-based approach to overcome measurement issues in valuing data.

- Information in financial markets:
  E.g., Edmans, Goldstein, and Jiang (2015), Goldstein and Yang (2019), Dugast and Foucault (2020), Davila and Parlatore (2020).
  - They equate data/signals and knowledge. We identify them separately and measure the changing relationship between data and knowledge.

- Impact of AI on the labor market:
  E.g., Acemoglu and Restrepo (2018); Deming and Noray (2018); Cockburn, Henderson, and Stern (2018); Babina, Fedyk, He, and Hodson (2020); Alekseeva, Azar, Gine, Samila, and Taska (2020).
  - They mostly use a DiD approach to measure labor changes. We use labor changes as an input to our structural framework.

Outline

A Model for Measurement

- Knowledge is produced using either the old technology or big data tech (AI). Same data can be used for both. Technologies have different rates of diminishing returns and use differently-skilled labor:

$$K_{it}^{AI} = A_t^{AI} a_i^{AI} D_{it}^{\alpha} L_{it}^{1-\alpha}, \tag{1}$$

$$K_{it}^{OT} = A_t^{OT} a_i^{OT} D_{it}^{\gamma} I_{it}^{1-\gamma}. \tag{2}$$

  A large $(\alpha - \gamma)$ = big revolution

- Data inputs are not raw data. They need to be structured, cleaned and machine-readable. This requires labor (λ) with diminishing marginal returns.
- New structured data is added to the existing stock of structured data. But data also depreciates at rate δ:

$$D_{i,t+1} = (1-\delta) D_{it} + \lambda_{it}^{1-\phi} \tag{3}$$

Maximization

- Firms maximize value function:

$$v(D_{it}) = \max_{\lambda_{it}, L_{it}, I_{it}} A_t^{AI} a_i^{AI} D_{it}^{\alpha} L_{it}^{1-\alpha} + A_t^{OT} a_i^{OT} D_{it}^{\gamma} I_{it}^{1-\gamma} - w_{L,t} L_{it} - w_{I,t} I_{it} - w_{\lambda,t} \lambda_{it} + \frac{1}{r} v(D_{i(t+1)}) \quad (4)$$

  where (3) holds.

- First Order Conditions:
  - L_{it}: $(1-\boldsymbol{\alpha}) K_{it}^{AI} - w_{L,t} L_{it} = 0$.
  - I_{it}: $(1-\boldsymbol{\gamma}) K_{it}^{OT} - w_{I,t} I_{it} = 0$.
  - λ_{it}: $\frac{(\alpha K_{it}^{AI} + \gamma K_{it}^{OT})(1-\boldsymbol{\phi})}{r-(1-\delta)} \frac{D_{i(t+1)} - (1-\delta) D_{it}}{D_{it}} - w_{\lambda,t} \lambda_{it} = 0$.

- These first order conditions allow us to identify $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$.

State Variable Evolution

We have two challenges:

1. We don't observe firms' data stock (D_{it}) but we can infer it:

$$D_{it} = \frac{\left(\alpha \frac{w_{L,t} L_{i,t}}{(1-\alpha)} + \gamma \frac{w_{I,t} I_{i,t}}{(1-\gamma)}\right)(1-\phi)}{r-(1-\delta)} \frac{\lambda_{it}^{-\phi}}{w_{\lambda,t}}$$

$$= (1-\delta)^t D_{i0} + \sum_{s=1}^{t} (1-\delta)^{t-s} \lambda_{s-1}^{1-\phi}$$

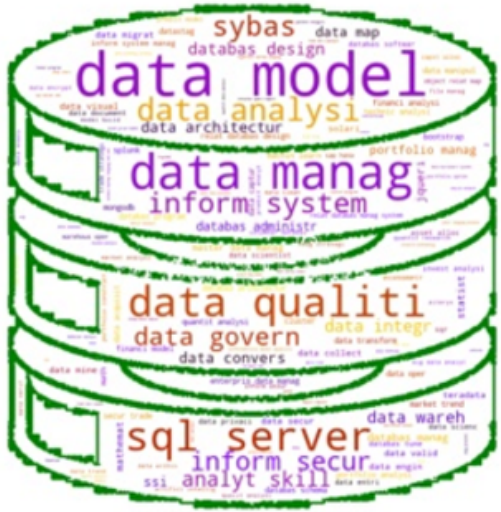
   - lhs: Data must be optimal given wages paid to workers.
   - rhs: Data accumulates in proportion to data management hires.
   - We express the D_{i0} of each firm as a function of an average $\bar{D}_0$, proportional to each firm's cumulated hiring between 2000-2014 (burn-in period).

2. We compute the four productivity parameters from the OT and AI FOCs:
   - A_t^{OT} and A_t^{AI} are computed using cross-sectional averages for each month.
   - a_i^{OT} and a_i^{AI} are computed using time-series averages for each firm.

Outline

Labor Demand: Job Postings Categorization

- Job postings sample:
  - Burning Glass Technologies (BGT), 2010 – 2018.
  - Filtered to jobs of predominantly investment management firms.
- Identify Data Management, AI or OldTech jobs:
  - Develop dictionaries of words and short phrases indicating data management or analysis (AI or OldTech) skills.
  - Identify jobs beloging to the three categories based on the relative frequency of skills in the full job text.



Labor Demand: Measurement innovation

- Investment management focus:
  - Match BGT to investment mangement firms from 13F filings and Preqin.
  - Novel algorithm tailored for employer names matching, utilizes both BGT employer field as well as Employer mentions in full job posting text.
  - Focusing on one industry allows to more precisely account for context-specific language and labor market.
- Jobs categorization:
  - Given the specific context the dictionary-based approach allows for a more fine-grained categorization (required for our estimation).
  - Utilizing the text of each job posting allows accounting for relevance of skills mentioned (e.g., "Machine Learning knowledge is a plus" not an AI job).
  - Labor & computationally intensive.

An Example: Two Sigma - OldTech

Two Sigma – Aug 2010 – Quantitative Analyst:

We are looking for world-class quantitative modelers to join our highly motivated team. Quant candidates will have exceptional quantitative skills as well as programming skills, and will write production quality, high reliability, highly-tuned numerical code. Candidates should have: a bachelor's degree in mathematics and/or computer science from a top university; an advanced degree in hard science, computer science, or the equivalent (a field where strong math and statistics skills are necessary); 2 or more years of professional programming experience in Java and C, preferably in the financial sector; strong numerical programming skills; strong knowledge of computational numerical algorithms, linear algebra and statistical methods; and experience working with large data sets. (...)

- Keywords:
  - AI: None
  - OldTech: mathemat (x1), math (x1), statist (x2), algebra (x1)
  - DataMgmt: None

An Example: Two Sigma - AI

Two Sigma – Mar 2018 – Quantitative Researcher In Machine Learning:

Two Sigma combines massive amounts of data, world-class computing power, and statistical expertise to develop sophisticated trading models. We believe that the scientific method is the best way to approach investing. We are looking for talented researchers who can apply and develop machine learning algorithms for financial datasets. Researchers work on: Developing trading strategies using statistical and machine learning algorithms. Advancing existing initiatives and opening opportunities to pursue new research topics. Designing solutions for challenges in analyzing real world, large datasets. Minimum qualifications: PhD in quantitative disciplines. Expertise in statistics and machine learning. Intermediate programming skills in Java, C++, or Python. (...)

- Keywords:
  - AI: data machin learn (x4)
  - OldTech: statist (x3)
  - DataMgmt: None

An Example: Two Sigma - DataMgmt

Two Sigma – Dec 2013 – SQL Data Analyst:

*(...) Technology drives our business it's our main competitive advantage and as a result, software engineers play a pivotal role. They tackle the hardest problems through analysis, experimentation, design, and elegant implementation. Software engineers at Two Sigma build what the organization needs to explore data's possibilities and act on our findings to mine the past and attempt to predict the future. We create the tools at scale to enable vast data analysis; the technology we build enables us to engage in conversation with the data, and search for knowledge and insight. (...) You will be responsible for the following: * Capturing and processing massive amounts of data for thousands of different tradable instruments, including stocks, bonds, futures, contracts, commodities, and more; (...)*

- Keywords:
  - AI: None
  - OldTech: None
  - DataMgmt: explor data possibl, enabl vast data analysi, data specialist, data team

Cumulating Postings to Labor Stocks

s_t^{type}: separation rates by type-month (from BLS, match NAICS codes by type)
h_t^{type}: fraction of posted vacancies filled by type-month (BLS, same)
j_t^{type}: Burning Glass job postings rates by type-month

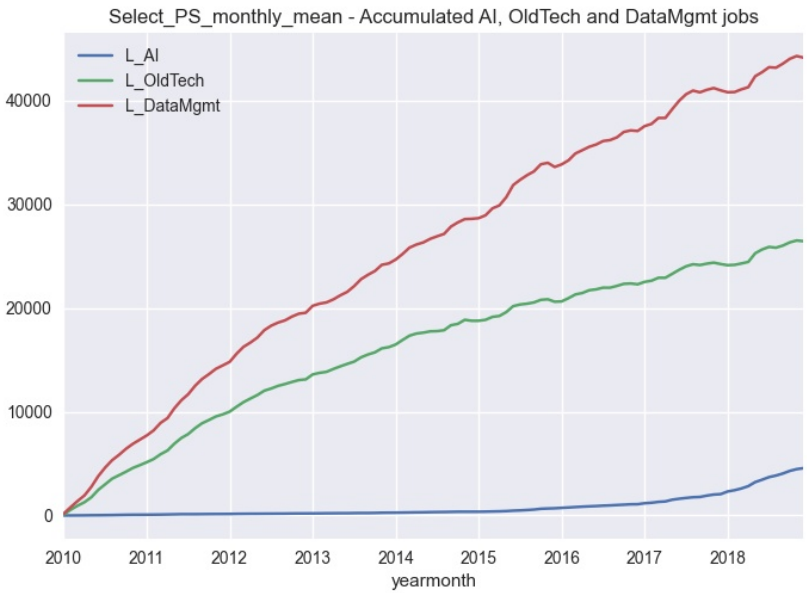
$$L_{it} = (1 - s_t^{AI}) L_{i(t-1)} + j_{it}^{AI} h_t^{AI}, \tag{5}$$

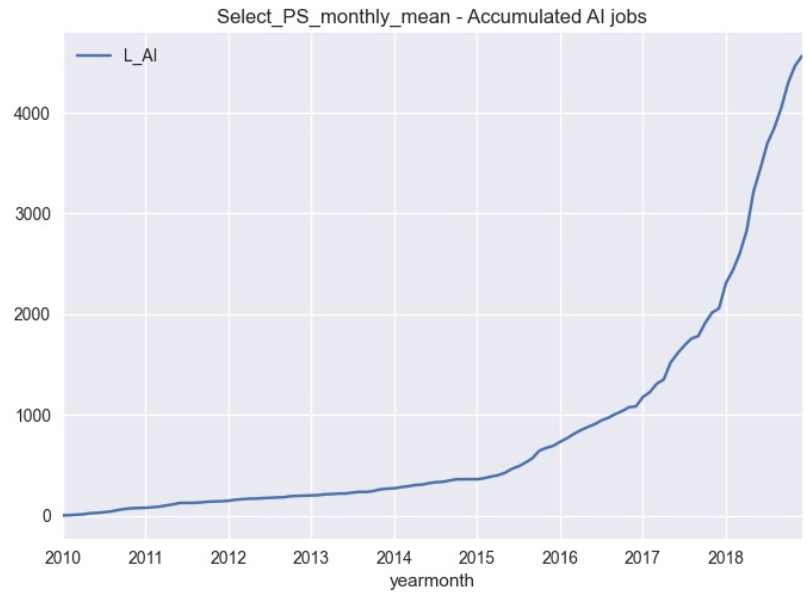
$$l_{it} = (1 - s_t^{OT}) l_{i(t-1)} + j_{it}^{OT} h_t^{OT}, \tag{6}$$

$$\lambda_{it} = (1 - s_t^{DM}) \lambda_{i(t-1)} + j_{it}^{DM} h_t^{DM}. \tag{7}$$

- Remaining question: What is the initial stock of labor? 2 possibilities:
  - Baseline: Start all initial employment at zero
  - Robustness: Assume that the sector in 2007 is in steady state. Then hiring is equal to the expected number of separations: $h_{i0} = s_t L_{i0}$ and $H_{i0} = S_t \lambda_{i0}$. Use initial hiring to impute initial stocks.
- In both cases, we use 2010-14 as burn-in and start estimation in 2015.

Labor Stocks: Growth in AI employment





- AI employment has been growing at a faster rate since 2015. We use 2015-2018 as our estimation period.
- Between 2015 and 2018 AI employment rose 13 times (from 350 to 4,537 jobs).

Steady-state initialization

Wages: PayScale

- Crowd-sourced salary data from PayScale salary surveys.
  - Detailed salary information at the industry, job title and skills-mix level.
  - 925,480 survey responses in the industries of interest.
  - Subset to match O*NET, job titles and employers in job postings.
  - Final sample: 11,041 surveys categorized as AI (2,585), OldTech (2,817) and DataMgmt (5,639), in 2015-2018.

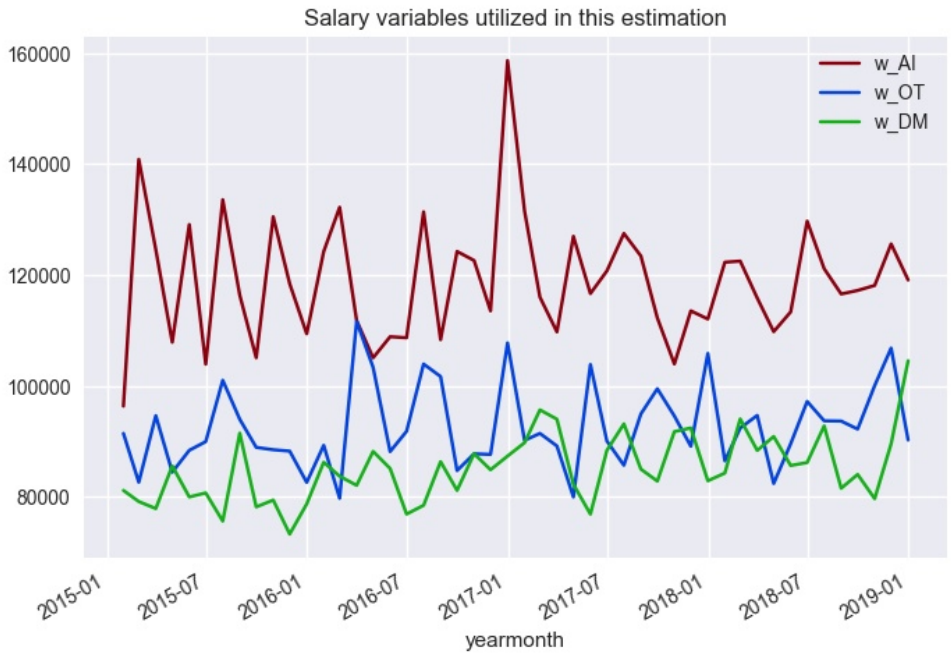


Figure: Average total compensation (salary + bonus) for AI, OldTech and DataMgmt workers. PayScale, 2015-2018.

Structural Estimation

- From the model's solution we have $4 \times 33,392$ equations to be set to zero.
- Procedure:
  - Non-linear least squares to iterate over different combinations of the diminishing returns parameters (α, γ and ϕ) and the average initial data stock ($\bar{D}_{i0}$).
  - In each iteration, we back out the production parameters from:
    - Cross-sectional averages each month of the AI and OT FOCs (A_t^{AI} and A_t^{OT}).
    - Time-series averages for each firm of the AI and OT FOCs (a_i^{AI} and a_i^{OT}).
  - We also use a grid search to check global convergence.

Outline

Main Results: Greater Productivity of Data

		$\delta = 1\%$	$\delta = 3\%$	$\delta = 10\%$
AI Analysis	α	0.916	0.867	0.690
		(0.0010)	(0.0015)	(0.0027)
Old Technology Analysis	γ	0.690	0.816	0.640
		(0.0033)	(0.0020)	(0.0032)
Data Management	ϕ	0.280	0.453	0.010
		(0.0063)	(0.0058)	(0.0085)
Change in Labor Share	$\gamma - \alpha$	-23%	-5.1%	-4.9%

Table: α and γ are the diminishing returns to data in the new and old technologies.

- $\alpha > \gamma$
  - AI has significantly raised the productivity of analyzing larger data sets.
  - Labor share fell from 18% to 13% (for $\delta = 3\%$).
    Same size change for $\delta = 10\%$.
- Technological change is substantial.
  - Industrial revolution: capital exponent estimated to have risen of $0.05 - 0.20$. We estimate an increase of 0.05 in the data exponent.

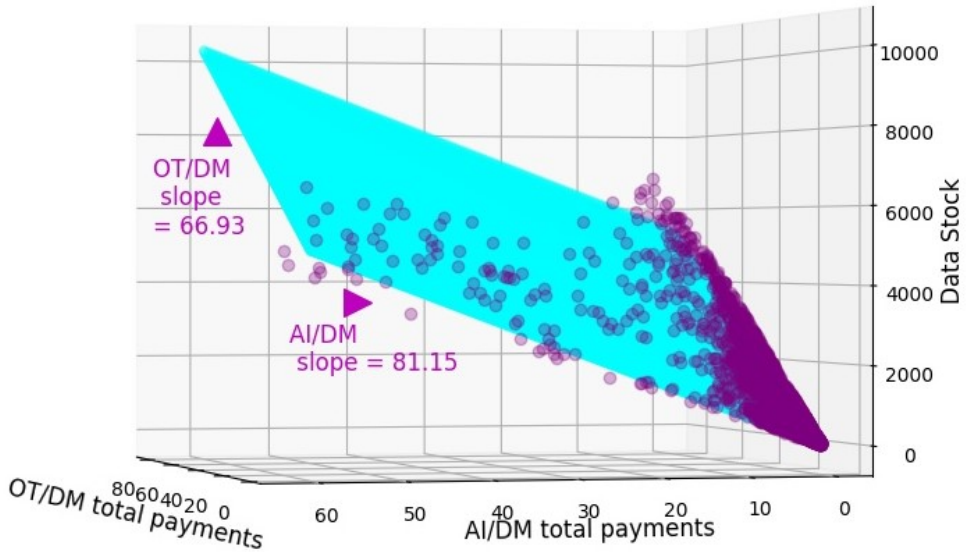
Robustness

What Data Features Matter?

- We can re-write the data first order condition as

$$D_{it} = \beta_1 \frac{w_{l,t} l_{i,t}}{\lambda_{it}^{\phi} w_{\lambda,t}} + \beta_2 \frac{w_{L,t} L_{i,t}}{\lambda_{it}^{\phi} w_{\lambda,t}}; \tag{8}$$

$$\textit{where} \quad \beta_1 = \frac{(1-\phi)}{r-(1-\delta)} \frac{\gamma}{1-\gamma} \quad \textit{and} \quad \beta_2 = \frac{(1-\phi)}{r-(1-\delta)} \frac{\alpha}{1-\alpha}. \tag{9}$$



- Greater sensitivity of data stock to AI workers payments is the one feature of the data that mostly explains our main findings.

Results: Not a Labor Replacing Technology

A fall in the labor share could mean fewer workers, or could mean more data. Which was it?

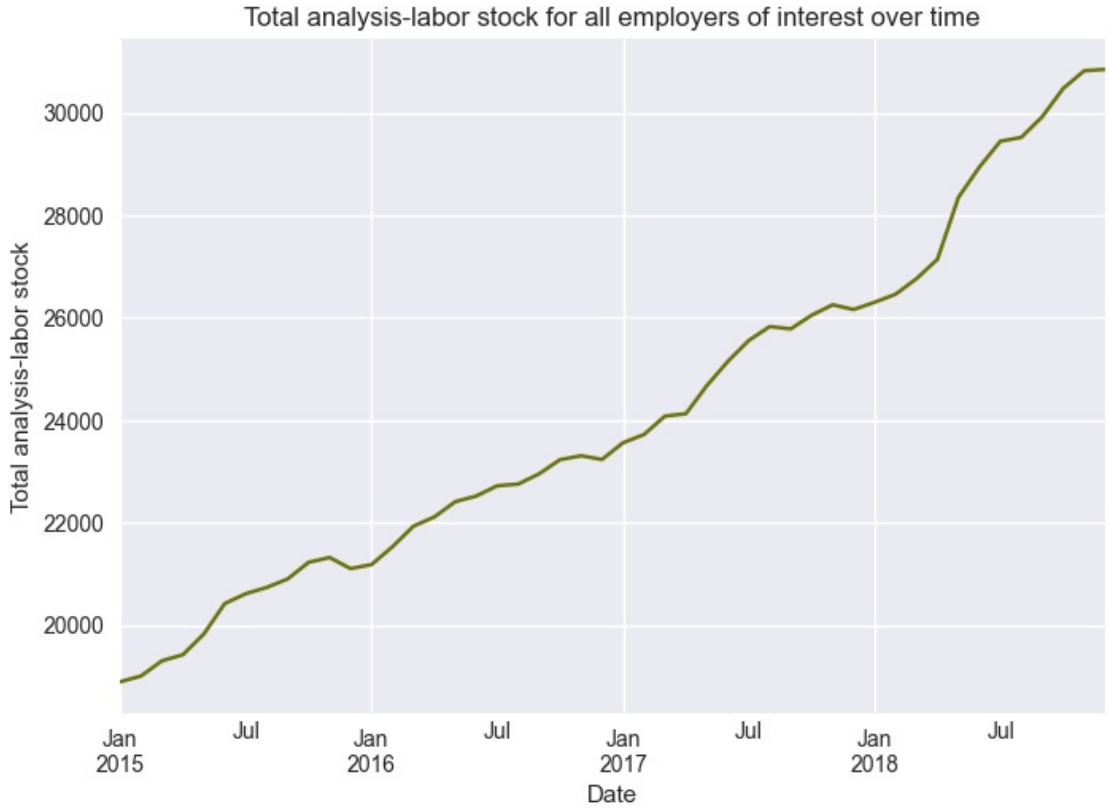


Figure: The aggregate stock of analysis labor (AI and OldTech).

- Labor stock has been increasing steadily, split about evenly between AI and OldTech analysts

Results: Our Methodology Can Value Data

How has this shift impacted the value of data?

- Substitute estimated parameters and data stocks into our value function.
- Data value rose 39% in 4 years.

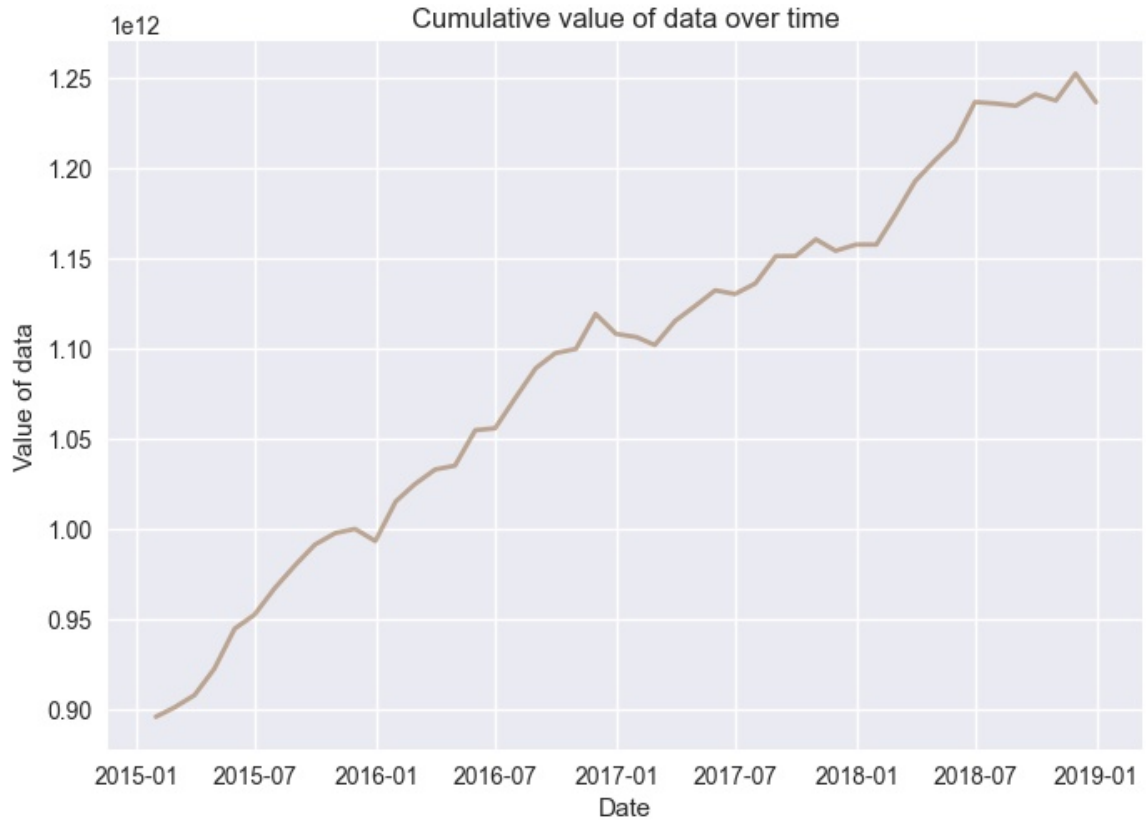


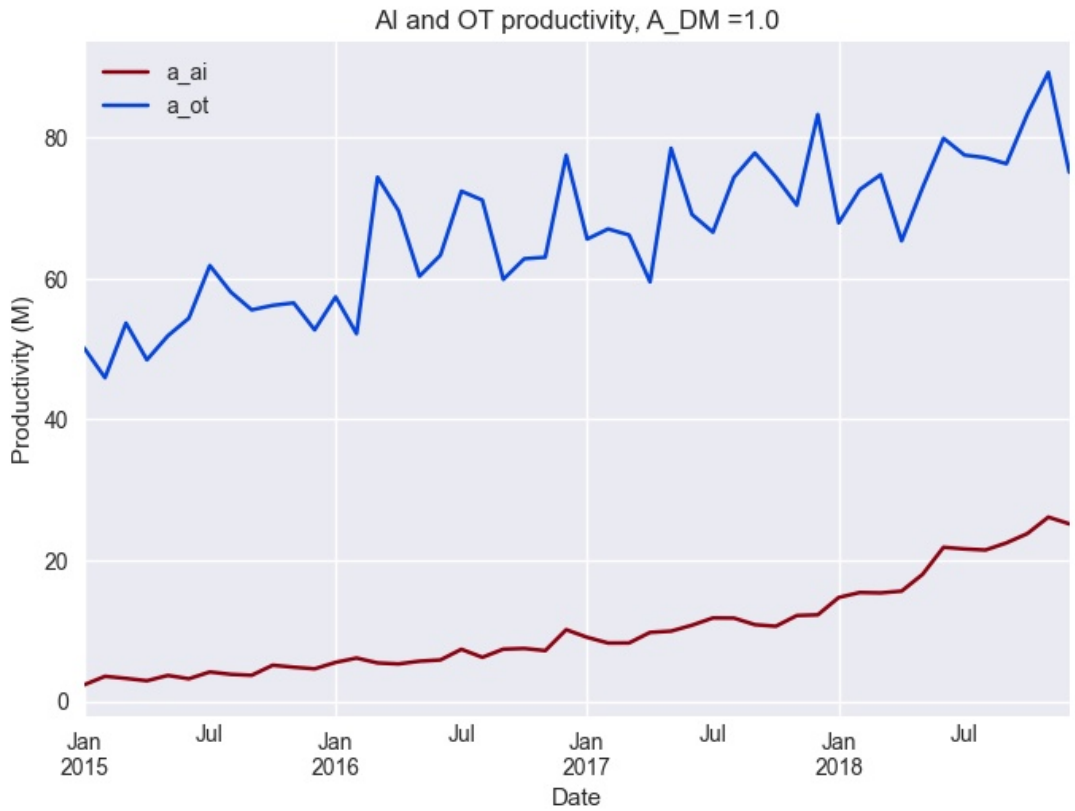
Figure: Estimated Value of the Aggregate Stock of Data, in current U.S. dollars, 2015-2018.

AI is Raising the Value of Data: 3 Reasons

1. A larger data stock determines a higher cumulative value of data

2. More analysis workers make each data point more valuable

3. Firms are becoming more productive at using data:



Outline

Take-Aways: Let the Revolution Begin...

- We infer how much data each firm has from data management hiring.
- We infer how much diminishing returns there is by asking what exponent would make their hiring patterns closest to optimal.
  The magnitude looks like the industrial revolution – but for knowledge production.
- The change in diminishing returns matters for the value of data as an asset, for inequality and for firm size/competition.

Backup Slides

Estimates with Steady State Accumulation

Table: Structural estimates of model parameters initializing labor variables from steady-state in 2010. All models are estimated for the depreciation rate of data $\delta = [1\%, 3\%, 10\%]$.

Productivity	δ	α	γ	ϕ	$\bar{d}_0$	$(\alpha - \gamma)$
Steady-State Initialization of Old Tech and Data Management Labor Stock						
FirmTime	0.01	0.8943	0.6159	0.1197	2145	0.2784
		(0.0013)	(0.0035)	(0.0072)	(80)	
FirmTime	0.03	0.8299	0.7492	0.2579	626	0.0807
		(0.0016)	(0.0023)	(0.0066)	(19)	
FirmTime	0.1	0.7009	0.6389	0.01	625	0.062
		(0.0028)	(0.0032)	(0.006)	(48)	

Back

Alternative Estimation Methods

1. **None:** No productivity parameters were added to the Cobbb-Douglas production functions.

$$K_{it}^{AI} = D_{it}^{\alpha} L_{it}^{1-\alpha}; \quad K_{it}^{OT} = D_{it}^{\gamma} L_{it}^{1-\gamma} \tag{10}$$

2. **Time:** Time-varying productivity parameters were added to the Cobbb-Douglas production functions. They are computed as cross-sectional averages of the AI and old tech FOCs for all relevant firms each month (t).

$$K_{it}^{AI} = A_{t}^{AI} D_{it}^{\alpha} L_{it}^{1-\alpha}; \quad K_{it}^{OT} = A_{t}^{OT} D_{it}^{\gamma} L_{it}^{1-\gamma} \tag{11}$$

3. **Firm:** Firm-specific productivity parameters were added to the Cobbb-Douglas production functions. They are computed as time-series averages of the AI and old tech FOCs for each firm (i).

$$K_{it}^{AI} = a_{i}^{AI} D_{it}^{\alpha} L_{it}^{1-\alpha}; \quad K_{it}^{OT} = a_{i}^{OT} D_{it}^{\gamma} L_{it}^{1-\gamma} \tag{12}$$

Back

Estimates Using Alternative Methods

Table: Structural estimates of model parameters for different productivity parameters settings. All models are estimated for data depreciation $\delta = [1\%, 3\%, 10\%]$.

Productivity	δ	α	γ	ϕ	$\bar{d}_0$	$(\alpha - \gamma)$
Zero Initialization of Labor Stock						
None	0.01	0.9188 (0.0019)	0.6992 (0.0059)	0.2979 (0.0115)	737 (36)	0.2196
None	0.03	0.8777 (0.0034)	0.8321 (0.0046)	0.4992 (0.0132)	191 (8)	0.0457
None	0.1	0.6909 (0.007)	0.6403 (0.0081)	0.01 (0.032)	500 (103)	0.0506
Time	0.01	0.895 (0.0005)	0.631 (0.0018)	0.1525 (0.0012)	1417 (15)	0.264
Time	0.03	0.7993 (0.0008)	0.7073 (0.0008)	0.1469 (0.0009)	798 (5)	0.0919
Time	0.1	0.7165 (0.0012)	0.675 (0.0004)	0.1422 (0.0009)	223 (1)	0.0415
Firm	0.01	0.908 (0.0011)	0.6656 (0.0032)	0.2368 (0.0064)	929 (28)	0.2424
Firm	0.03	0.818 (0.0015)	0.7199 (0.002)	0.2212 (0.0061)	496 (15)	0.0981
Firm	0.1	0.6886 (0.0025)	0.6392 (0.0027)	0.01 (0.001)	482 (34)	0.0494