The Changing Economics of Knowledge Production

Simona Abis and Laura Veldkamp
Columbia University, NBER and CEPR*

May 26, 2020

Abstract

Big data technologies change the way in which data and human labor combine to create knowledge. Is this a modest technological advance or a transformation of our basic economic processes? Using hiring and wage data from the financial sector, we estimate firms' data stocks and the shape of their knowledge production functions. Knowing how much production functions have changed informs us about the likely long-run changes in output, in factor shares, and in the distribution of income, due to the new, big data technologies. Using data from the investment management industry, our results suggest that the labor share of income in knowledge work may fall from 44% to 27% and we quantify the corresponding increase in the value of data.

Machine learning, artificial intelligence (AI), or big data all refer to new technologies that reduce the role of human judgment in producing usable knowledge. Is this an incremental improvement in existing statistical techniques or a transformative innovation? This nature of this technological shift is similar to industrialization: Industrialization changed the capital-labor ratio, allowing humans to be more efficient at goods production. Machine learning is changing the data-labor ratio, allowing humans to be more efficient at knowledge production. Economists model industrialization as a change in production technology: a move from a technology with starkly diminishing returns to capital, to one with less diminishing returns. One measure of the importance of the industrial revolution is the magnitude of the change in the production parameter that governs diminishing returns.

[*] Preliminary work. Numerical estimates will surely change. Comments welcome. Email: lv2405@columbia.edu or sa3518@columbia.edu. Columbia Business School, 3022 Broadway, New York, NY 10025. Thanks to Sahil Arora, Sagar Agarwal, Samrat Halder, Anshuman Ramachandran, Ran Liu and Saad Naqvi who substantially advanced the project with their research assistance. We acknowledge the generous financial support and helpful feedback from IBM and the Columbia University Data Science Institute, as well as seminar participants at Columbia, the IMF and Princeton. Keywords: FinTech, machine learning, artificial intelligence, big data, labor share.

Using labor market data from the financial sector, we estimate two production functions – one for classical data analysis and one for machine learning. The decline in diminishing returns shows up as an exponent on on data in the production function that is closer to one: We estimate that the data exponent rose from 0.560 to 0.734. The magnitude of the change in the diminishing returns parameter informs us about the importance of the innovation. As the economy transitions from producing knowledge using old statistical techniques to producing with new, machine learning technologies, this change in diminishing returns governs changes in input use, income shares and productivity. In other words, our results inform us about how the demand for labor and data will change, how to value each in the new economy, and how the distribution of income is likely to shift, absent policy intervention.

Estimating old and new knowledge production functions is challenging, because for most firms, we do not know how much data they have, nor how much knowledge they create, nor do they announce which technology or what mix of technologies they employ. What we can observe is hiring, skill requirements and wages. A simple model of a two-layer production economy teaches us how to infer the rest. The two layers of production are as follows: Raw data is turned into usable, processed data (sometimes called information) by data managers; processed data and data analyst labor combine to produce knowledge. Thus, we use hiring of data managers to estimate the size of the firm's data stock, the skills mix of analysts to estimate the mix of data technologies at work, and we bypass the need to measure knowledge by using wage data to construct income shares, which inform us about the returns, and the rate of diminishing returns, to each factor.

To estimate production functions, it is imperative that we precisely categorize job postings and match postings by employer. Unlike other work that measures machine-learning-related employment(e.g., Acemoglu and Restrepo (2018)), our work demands a finer partition of jobs. We need to distinguish between workers that prepare data to be machine-analyzed, workers that primarily use machine learning, and workers that use similar statistical skills, that are frequently co-listed with machine learning, but are of a previous vintage. We also need to know whether data managers are being hired by the same firm that is also hiring machine-learning analysts.

Because different industries have different job vocabularies, we can categorize jobs more accurately by focusing on one industry: finance, more specifically we focus on investment management. Since investment management is primarily a knowledge industry, with no physical output, it is a useful setting in which to tease apart these various types of knowledge jobs.[1] We use Burning Glass hiring data, including the textual descriptions of each job, to

[1] According to Webb (2019) and Brynjolfsson et al. (2018b), finance is also the industry with the greatest potential for artificial intelligence labor substitution.

isolate financial analysis jobs that do and do not predominantly use machine learning, as well as data management jobs, for each company that hires financial analysts. We adjust the number of job postings by a probability of job filling. That product is our measure of a company's desired addition to their labor force. This series of worker additions, along with job separations by job category, enables us to build up a measure of each firm's labor stock.

The next challenge is to estimate the amount of data each firm has. We consider data management work to be a form of costly investment in a depreciating data asset. Therefore, we use the job postings for data managers, the job filling and separation rates for such jobs, and an estimate of the initial data stock to construct data inflows (investments), per firm, each year. To estimate the 2010 initial stock of data of each financial firm, we estimate which stock best rationalizes the firm's subsequent hiring choices. Specifically, we choose an initial stock of data that minimizes the distance between each firm's actual hiring and the optimal amount of hiring in each category, dictated by the firm's first order conditions. Combining this initial stock, with a data depreciation rate and a data inflows series gives us an estimate of the size of the data stock that every financial firm has in its data warehouse.

Armed with data stocks, labor forces in each category, and wages, we estimate the data and labor income shares. These income shares correspond to the exponents in a Cobb-Douglas production function. We estimate a constant-returns Cobb-Douglas specification because we are exploring the analogy that AI is like industrialization. Therefore, we model knowledge production in a parallel way to industrialization, to facilitate comparison, while recognizing the non-rival nature of data. By comparing the estimated exponent for classical data analysis and machine-learning data analysis, we can assess the magnitude of the technological change. Of course, this knowledge is then combined with capital to generate excess financial returns. But it turns out that we do not need to model or measure this downstream value-creation to make inference about knowledge production. Just like we can determine the production function for milk, without knowing what factors are needed to turn it into ice cream, we can estimate the production of knowledge, without asking how that knowledge is turned into excess financial returns.

Our data reveals a shift underway in the employment of knowledge workers in the investment management sector. We see a steady increase in the fraction of the workforce skilled in new big data technologies. The number of old technology jobs in the sector has not fallen; it simply represents a smaller share of employment. While AI job postings were a tiny fraction of all analysis jobs through 2015, by the end of 2018, about $1/7^{th}$ of all financial analysts in investment management firms had big data or AI-related skills.

Related Literature A handful of recent working papers also use labor market data to investigate how machine learning and artificial intelligence are affecting labor demand. They primarily use a difference-in-difference approach. Acemoglu and Restrepo (2018), Babina et al. (2020) and Deming and Noray (2018) identify industries and/or regions that are more exposed to machine learning-related technology. Then, controlling for other labor-related variables, they report how many jobs have been lost or gained, relative to unexposed regions or industries. Others offer useful inputs in this exercise by reporting the number of AI jobs postings or patents by industry and occupation (Cockburn et al. (2018) and Alekseeva et al. (2020) paper). Agrawal et al. (2017) and Agrawal et al. (2018b) argue that machine learning is likely to be a general purpose technology, because of the breadth of industries in which it is being adopted.

Our paper contributes a structural, production function approach. Estimating how much the production function has changed allows us a more holistic understanding of the nature of the transformation. A structural model allows us to forecast, to make inferences about income redistribution, and to understand the social welfare effects, beyond job counts. The number of jobs gained or lost due to machine learning to date is an important question; it informs our work, but it is just one piece of our overall puzzle.

Others examine the productivity gains or potential discrimination costs that follow the adoption of AI techniques in providing credit (Fuster et al. (2018)), in equity analysis (Grennan and Michaely (2018)), or in deep learning more generally (Brynjolfsson et al. (2017) and Brynjolfsson et al. (2018a)). Our emphasis, on how inputs combine to create knowledge, is complementary to such studies that examine the outputs and effects of machine learning.

Berg et al. (2018) take a similar structural approach, with a more theoretical focus, on a somewhat different topic. They explore models with different elasticities of substitution between robots and manual workers. Our focus is on knowledge production, rather than manual task automation. The scope for computers to replace human thought and judgment may be quite different from their ability to replicate repetitive physical movements. However, our quantitative approach using hiring data could be applied to study robotics as well.

Models of the role of data in the process of economic growth (Jones and Tonetti (2018), Agrawal et al. (2018a), Aghion et al. (2017) and Farboodi and Veldkamp (2019)) share our model-based approach but equate data and knowledge. In these theories, firms accumulate a stock of useable knowledge that enhances productivity or facilitates prediction. In contrast, this study unpacks how raw data is transformed into that valuable output-enhancing knowledge.

Finally, our approach is related to work using Q-theory to impute the value of intangible assets (Crouzet and Eberly, 2020). Our approaches are different: Q theory backs out a production function exponent from asset prices and book values, while our approach builds up a production function from labor inputs. Our objectives are also different: Q-theory is decomposing the sources of value in a firm. We are interested in how much two technologies, often both used within the same firm, differ.

1 A Model for Measurement

The objective in writing down this model is not to provide insight into new economic mechanisms, nor it is to provide the most realistic, detailed description of financial knowledge production. Rather, the goal is to write down a simple framework that maps objects we observe into those that we want to measure. It needs to relate hiring to labor as well as quantities and prices of labor to data stocks and knowledge production. There are three types of workers: AI (artificial intelligence) analysts, old technology (OT) analysts, and data managers. We use AI as a shorthand to denote a diverse array of big data technologies. The data managers create structured data sets, which, along with labor, are the inputs into knowledge production. Among data managers we also include workers who select, purchase and integrate externally produced data sets into the firm's databases. We define as data (D) only information that is readily available for analysis. This production process is illustrated in Figure 1.

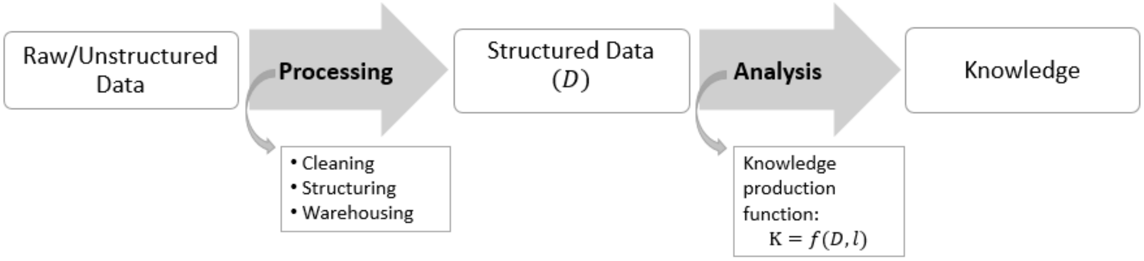


Figure 1: Production process for knowledge

The new technology knowledge production function is:

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

where D_{it} is structured data, L_{it} is labor input for data analysts with machine-learning skills, and K_{it}^{AI} is the knowledge generated using the new technology. The old technology knowledge production function is:

$$K_{it}^{OT} = A_t^{OT} D_{it}^{\gamma} l_{it}^{1-\gamma}, \tag{2}$$

where l_{it} is labor input for data analysts with traditional analysis skills, K_{it}^{OT} is the knowledge generated using the old technology. A_t^{AI} and A_t^{OT} are time-varying productivity parameters.

We use a Cobb-Douglas production function for knowledge because it offers a clear mapping between incomes shares and the production function parameters and it facilitates our comparison between new data technologies and the changes induced by industrialization. Notice that our specification does embody the non-rival nature of data. Both technologies make use of the same data set, at the same time. In addition, it may well be that each form of knowledge production has increasing returns. If it did, we would need to make a host of other controversial assumptions about what the firm chooses as its optimal scale, how profits are shared, and why so many firms are competing in a sector that ought to give rise to a natural monopoly. In other words, the intense competition for the provision of investment management services suggests that returns are not greatly increasing in scale. But that is a set of questions best left for another project with a different aim.

This structure implies that the nature of the data inputs is the same for both types of analysis. This simplifies measurement, but the obvious counterfactual would be: Machine learning can make use of a broader array of data types than traditional analysis. One way to interpret this is that it is the source of greater decreasing returns to data from the old technology. Suppose that data is ordered, from easily usable to difficult to use. Once the easiest data is incorporated, the next additional piece of data for traditional analysis has very low marginal value. For machine learning, that next piece of data has higher marginal value. Thus, the difference in the usability of data could be the primary reason for the difference in returns to data.

Data management and Data Stocks. Data inputs for analysis are not raw data. They need to be structured, cleaned and machine-readable. This requires labor. Suppose that structured data, sometimes referred to as "information," is produced according to $A^{DM}\lambda_{it}^{1-\phi}$, where λ_{it} is labor input for data managers and A^{DM} is the productivity of data manager (DM) labor. Labor with diminishing marginal returns can turn raw or purchased data into an integrated, searchable data source that the firm can use. New processed data is added to the existing stock of processed data. But data also depreciates at rate δ. Overall, processed data follows the dynamics below:

$$D_{i(t+1)} = (1-\delta)D_{it} + A^{DM}\lambda_{it}^{1-\phi} = D_{i0}(1-\delta)^t + \sum_{s=0}^{t}(1-\delta)^{t-s}A^{DM}\lambda_{is}^{1-\phi}. \tag{3}$$

If we estimate the rate of diminishing returns to data management labor λ_{it}, initial data D_{i0} and the depreciation rate δ, we can recover D_{it} from data management labor λ_{it}.

Equilibrium We are interested in a competitive market equilibrium where all firms choose the three types of labor to maximize firm value. We can express this problem recursively, with the firm's data stock as the state variable. In this equilibrium, each firm i solves the following optimization problem:

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

$$\text{where } \; D_{i(t+1)} = (1-\delta) D_{it} + A^{DM} \lambda_{it}^{1-\phi}, \tag{5}$$

and $v(D_{it})$ is the present discounted value of firm i's data stock at time t. Note that we have implicitly normalized the price of knowledge to 1. This is not restrictive because knowledge does not have any natural units. In a way, we are saying that one unit of knowledge is however much knowledge is worth \$1. Seen differently, our A parameters measure a combination of productivity and price. We cannot disentangle the two and do not need to for our purposes.

Optimal firm hiring and wages. The first order condition with respect to new technology (AI) analyst labor L_{it} is

$$(1-\alpha) K_{it}^{AI} - w_{L,t} L_{it} = 0, \tag{6}$$

which says that total payments to new technology analysis labor $w_{L,t} L_{it}$ are a fraction $(1-\alpha)$ of the value of knolwedge output from AI analysis, K_{it}^{AI}. The first order condition with respect to old tech analyst labor l_{it} is

$$(1-\gamma) K_{it}^{OT} - w_{l,t} l_{it} = 0. \tag{7}$$

This says that the total payments to old technology analysis labor $w_{l,t} l_{it}$ are a fraction $(1-\gamma)$ of the value of total output K_{it}^{OT}. Taking the ratio of the two first order conditions implies that

$$\frac{(1-\alpha) K_{it}^{AI}}{(1-\gamma) K_{it}^{OT}} = \frac{w_{L,t} L_{i,t}}{w_{l,t} l_{i,t}} \tag{8}$$

This ratio varies by time t and it measures how much knowledge production technology has changed. The first order condition with respect to data management labor λ_{it} is

$$\frac{1}{r} v'(D_{i(t+1)}) (1-\phi) A^{DM} \lambda_{it}^{-\phi} = w_{\lambda,t}. \tag{9}$$

If the marginal value of data today and tomorrow are similar, we can solve for $v'(D)$ and replace $A^{DM}\lambda^{1-\phi}$ by the change in the data stock, to get[2]

$$\frac{(\alpha K_{it}^{AI}+\gamma K_{it}^{OT})(1-\phi)}{r-(1-\delta)}\frac{D_{i(t+1)}-(1-\delta)D_{it}}{D_{it}}-w_{\lambda,t}\lambda_{it}=0. \tag{10}$$

Intuitively, total payments to data management $w_{\lambda,t}\lambda_{it}$ are a portion of $(\alpha K_{it}^{AI}+\gamma K_{it}^{OT})(1-\phi)$, pdv (Gordon growth), or total output times the percentage increase in the data stock.

Using these expressions for optimal labor choices, we can derive an expression for the optimal stock of data for a firm. This is an expression we will use to impute the initial data stock of each firm. We start with (10) and substitute in $A^{DM}\lambda_{it}^{1-\phi}$, in place of $D_{i(t+1)}-(1-\delta)D_{it}$. Next, we need to replace K_{it}^{AI} and K_{it}^{OT} which are the unobserved knowledge produced with each technology. To do this, we use the first order conditions for AI and OT labor, (6) and (7), to substitute wage per worker expressions; $K_{it}^{AI}=w_{L,t}L_{i,t}/(1-\alpha)$ and $K_{it}^{OT}=w_{l,t}l_{i,t}/(1-\gamma)$. This yields an expression that relates firm i stock of data to production function exponents and observable hiring and wages:

$$D_{it}-\frac{\left(\frac{\alpha}{1-\alpha}w_{L,t}L_{i,t}+\frac{\gamma}{(1-\gamma)}w_{l,t}l_{i,t}\right)(1-\phi)}{r-(1-\delta)}\frac{A^{DM}\lambda_{it}^{-\phi}}{w_{\lambda,t}}=0. \tag{11}$$

2 Data and Estimation

Why look at the investment management industry? Our model is about knowledge production generally, in any industry. But as we turn to estimating this model, we use asset management industry labor and data estimates. One reason we do this is that the investment management industry is primarily a knowledge industry, where information is processed to form forecasts about asset returns and profitable portfolios. But the main reason is that finance is an early adopter of AI and big data technology. If we want to study the nascent adoption of this new technology, it is helpful to look in corners of the economy where adoption is most substantial. In independent studies with different methodologies, Felten et al. (2018) and Brynjolfsson et al. (2018c) both came to the conclusion that the finance/insurance industry was the one with the greatest potential for labor substitution with AI. Acemoglu et al. (2019) document that finance has the third most number of AI job postings, behind information and business services.

Finally, the financial industry is a useful laboratory because finance jobs are typically filled. JOLTS data tell us that finance is an industry with one of the highest vacancy conversion rates into new employment, presumably because the finance sector pays more than others. Thus,

[2] See appendix for step-by-step derivation.

when they want a worker with a specific set of skills, they can buy them. Since our work relies on job postings, it is helpful if many of these postings are, in fact, filled.

Of course, one could argue that we could include the investment management industry, as well as all other industries, to broaden our sample and sharpen our estimates. The problem with this approach is that distinguishing which workers combine data and labor to produce knowledge is tricky. Determining which workers use which technology is even more delicate. Different industries use different vocabularies to describe this type of work. The type of work that the investment management industry calls an analyst, the retail industry might call an online marketing expert. Both are using data and labor to make predictions that will enhance their company's profit. But because the language used to describe jobs differs, one needs a separate dictionary/model to identify relevant jobs in each context. Therefore, restricting our analysis to the asset management sector allows us to obtain a cleaner sample of job postings and improve the accuracy of our estimates.

Labor demand Our data is the job postings data set collected by Burning Glass, from January 2010 through December 2018. These postings are scraped from more than $40,000$ sources (e.g. job boards, employer sites, newspapers, public agencies, etc.), with a careful focus on avoiding job duplication. Acemoglu et al. (2019) show that Burning Glass data covers 60-80% of all U.S. job vacancies. The finance and technology industries have especially good coverage. It includes jobs posted in non-digital forms as well. Importantly, for a large portion of job postings, the data reports employer names, as well as the sector, job title, skill requirements, and sometimes the offered salary range. In addition to the structured data fields, we also make use of the full text of the job posting, as written by employers.

The total number of job postings for the employers in our sample is $507,971$, we categorize $143,809$ of them as searching for old-tech financial analysts, AI financial analysts, or data managers. The unique number of employers goes from 620 in January 2015 to 797 in December 2018. The total number of unique employers is 928.

In order to construct this data set of interest, we develop various data filters that (1) subset the Burning Glass data to candidate jobs in the financial industry, (2) identify which of those jobs require investment management skills, (3) assign all jobs to unique employers and (4) keep only job postings from employers that significantly hire in investment management. Finally, among all job postings for the employers of interest, we identify those searching for AI/old-tech financial analysts or data managers. After keeping only such observations for employers that in a given month have a non-zero stock of at least one of the three labor types, the total number of employer-month observations is: 33,610. This is the number of observations used for our estimation.

In our initial filter (1), we use the jobs' NAICS, O*NET and proprietary Burning Glass codes to restrict the Burning Glass data set to candidate jobs in the financial industry. More specifically, we first drop all job postings that do not belong to one of the following 2-digit NAICS codes: 'Professional, Scientific, and Technical Services', 'Finance and Insurance', 'Information' and 'Management of Companies and Enterprises'. We also keep all jobs for which the NAICS code is not available. Next we compile lists of O*NET codes and Burning Glass proprietary codes (BGT Occupation Group, BGT Career Area) of job categories that should clearly not be contained in our sample[3]. After eliminating all jobs belonging to those categories, we are left with a first sample of candidate fincance jobs.

With our second filter (2) we identify investment management jobs in our sample of candidate finance jobs. For each job we consider the list of required skills as identified by Burning Glass. These are standardized skills extracted from the full text of the job postings. If the skill is mentioned at least once in the job posting then Burning Glass includes it in the list of skills required by the job. We first construct a list of all standardized skills required by any of the jobs in our sample. From that list we select all investment management related skills (a full list of the shortlisted skills can be found in Appendix A). If a job requires one or more of these skills, we categorized it as belonging to the 'Investment Management' category.

Out third step (3) is to assign all jobs to unique employer identifiers, which we develop through fuzzy matching of the provided employer names. We exclude jobs for which the employer is a recruiting company. Combining steps (2) and (3), in step (4) we keep all jobs for employers that posted at least one job requiring investment management skills between Jan 2010 and Dec 2018.

For all jobs in this sample, we then use the full text of the selected job postings in order to identify analysis jobs and data management jobs. We define 'data management' jobs as those requiring skills related to the cleaning, purchasing, structuring, storage and retrieval of data. What define as “analysis jobs" those jobs that combine structured data with skilled labor. We call these analysts because they analyze data in different ways. They are not necessarily what the financial industry calls analysts. Within the analysis jobs we further distinguish between those that mostly require old (Old Technology - OT) or new (Artificial Intelligence - AI) skills.

This classification is obtained by developing a dictionary of words and short phrases that indicate 'data management' or 'data analysis', and then counting the relative frequency of these words or expressions in each pre-processed job text.[4] Among the 'data analysis' keywords we

[3] Examples of excluded 6-digit O*NET codes that were still present in the sample: 'Bookkeeping', 'Accounting, and Auditing Clerks', 'Customer Service Representatives', 'Cashiers', 'Retail Salespersons' ...

[4] We pre-process the text of each job posting by first removing symbols, numbers and stop-words (e.g. is, the, and, etc.) and then stemming each word to its root using the Porter stemmer algorithm (thus, e.g. 'mathematic', 'mathematics', ... = 'mathemat').

further identify those clearly indicative of the old and new technologies and we assign jobs to 'Old Tech - OT' or 'Artificial Intelligence - AI' depending on the relative frequency of words of the two types present in the posting. The full dictionaries used are available in Appendix B.

While this last step is similar in nature to the decompositions by Acemoglu and Restrepo (2018) or Babina et al. (2020), working with one type of job in a single industry allows us to partition the data more precisely. The approach of these authors is to define a dictionary of big data related words in all industries. They then identify job postings that contain those words in the standardized skills list provided by Burning Glass. Those are categorized as AI jobs; everything else is non-AI. This approach does not work for our exercise: Burning Glass' skills list is not detailed enough to distinguish between different types of data analysis in finance. Misclassification that might wash out in a job counting exercise is more serious for us. We need to match data and labor stocks firm-by-firm. This is why we analyze the full text of the job posting. Analyzing the full text, rather than using the Burning Glass skills list, greatly improves our classification by allowing us to account for the frequency of mentions of each type of skill.[5]

Finally, we further restrict the sample to employers that posted at least 5 'Old Technology' or 'Machine Learning' jobs throughout the entire sample and employers for which at least 25% of all identified 'data analysis' jobs also belong to the investment management subset. Other types of analysis jobs include procurement, operations, marketing and sales analysts. This final filter is needed in order to identify employers for which investment management is a large fraction of their business. The reason we do this is because data may be collected and used for many purposes. We want to measure data collection that will primarily be used in combination with the labor we measure. Of course, we also do robustness checks on less restricted samples.

There is lots of entry in our data set. 58% of firms are in our data set in 2015. The remaining 42% appear for the first time in 2016-2018. That does not mean these 42% are all new firms. Instead, many of them are existing firms that enter our data set when they hire data workers for the first time.

Figure 2 illustrates the frequency of all keywords in the job postings categorized as belonging to each type. Note that even if all 'data analysis' and 'data management' keywords are included in all three word clouds, the keywords specific to the assigned category have a significantly higher relevance. The word overlap illustrates why counting word frequency is important. At the same time, the significat differences between the word clouds validates our approach. If a clear distinction between the three types of job postings did not exist we would observe that the most frequently mentioned words in each category would be less distinct.

[5] For instance a job that mentions 'Machine Learning' 10 times withing the job text and then also states "Masters in Statistics also accepted", in our approach would be clearly classified in the 'AI' category. Looking at the skills lists, instead, the categorization of the job would be ambiguous as it would appear to require both old and new technology skills in the same proportion: 'Statistics' and 'Machine Learning'.

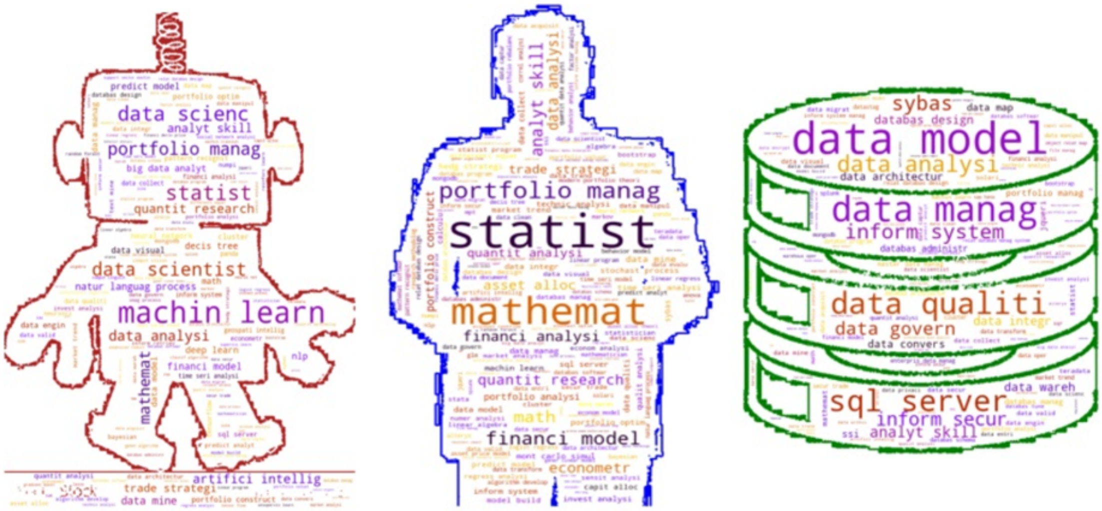


Figure 2: Keywords in the full text of the categorized machine learning, old technology and data management jobs. Larger fonts indicate a higher word frequency. Burning glass job postings, 2010-2018.

Wages Many, but not all jobs in Burning Glass list a salary range. We do not believe the listed salaries are representative of all jobs in this area. They are a starting point. We assume that they are biased for all types of jobs, in proportion to the listed salary. We typically use the median of the salary range listed as the salary for that job. We have robustness checks using the maximum and minimum of the salary range instead.

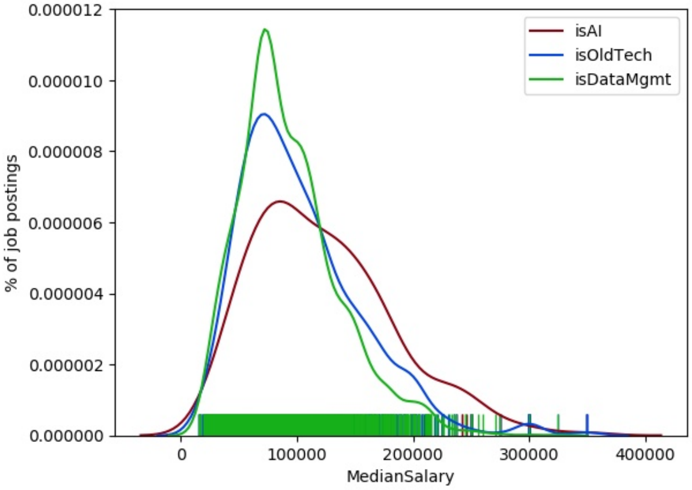


Figure 3: Distribution of wages for data managers, old technology analysts and machine learning analysts. Burning glass job postings, 2010-2018.

Figure 3 shows a distribution of wages (medians if the job lists a salary range) for data managers, old technology analysts and machine learning analysts. The key insight is that

AI jobs clearly pay more than traditional analyst jobs. This suggests that AI workers make more productive use of their data. This difference in wages is a key input that determines the difference in production function estimates.

Cumulating hiring to get labor. The data series we need in order to estimate production is the labor force working in a given month, for both knowledge and data processing workers. We do not observe the stock of labor. Therefore, we use the following procedure to estimate labor from observed job postings by firm. The number of observed job postings for the three categories of interest is displayed in Figure 4, together with the number of employers hiring in each category.

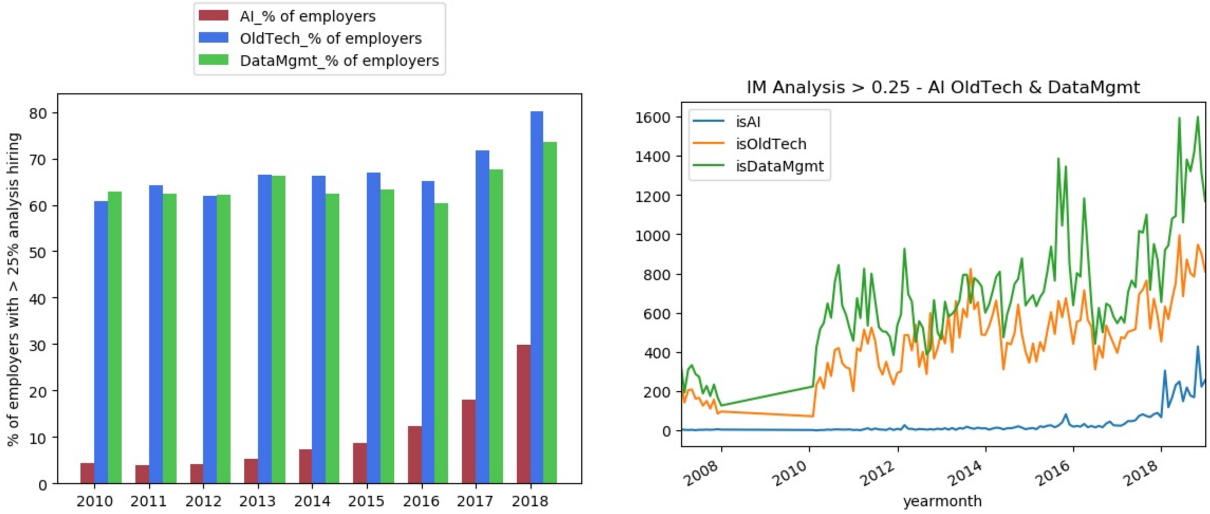


Figure 4: Job postings: Panel 1 show the fraction of employers hiring in each category. Panel 2 shows the numbers of job postings in each skill category. This sample includes only firms for which 25% or more of their analysis job postings required investment management skills.

Job postings are not the same as net hiring. There are two key differences: the probability that a vacancy is filled and the probability that an employed worker separates from their job. We adjust for both of these using data on vacancy fill rates and job separation rates from the Bureau of Labor Statistics (BLS).

Each month, the BLS reports the job posting, job filling and separation rate for each occupation. The three occupation brackets present in the final sample are: 'Finance and Insurance', 'Professional, Scientific and Technical Services' and 'Information'. Since we want to map our job postings into expected hires, we multiply each job posting number by the fraction of job postings that results in a new hire (h).

Of course, machine learning jobs are not an occupation. We need a way to map our technology-based job classification into the BLS occupation classification. Fortunately, each Burning Glass job posting has a listed occupation. Of course, different postings have different classifications, even within machine learning, old technology or data management jobs. Thus, we measure the proportion of jobs in each of our samples that belongs to each occupation. Each month we compute a vector of occupation weights for machine learning jobs, one for old tech jobs and one for data management jobs that is the fraction of jobs in each category that belongs to each occupation. We multiply this weight vector by each of the fill and separation rates that month, to get the imputed fill and separation rates for machine-learning financial analysis jobs (h_t^{AI} and s_t^{AI}), the imputed fill and separation rates for old technology financial analysis jobs (h_t^{OT} and s_t^{OT}) and those for data management jobs (h_t^{DM} and s_t^{DM}). See Appendix C for more detail on how BLS data is mapped into our job categories and how h and s are derived from BLS reported rates.

For $type = [AI, OT, DM]$, if s_t^{type} are separation rates by type-month, and h_t^{type} are the fraction of posted vacancies filled by type-month and j_t^{type} are Burning Glass job postings rates by type-month, we cumulate labor flows into stocks as follows:

$$L_{it} = (1 - s_t^{AI})L_{i(t-1)} + j_{it}^{AI} h_t^{AI}, \tag{12}$$

$$l_{it} = (1 - s_t^{OT})l_{i(t-1)} + j_{it}^{OT} h_t^{OT}, \tag{13}$$

$$\lambda_{it} = (1 - s_t^{DM})\lambda_{i(t-1)} + j_{it}^{DM} h_t^{DM}. \tag{14}$$

To use this cumulative approach, we need to know the initial number of workers of each type (L_{i0}, l_{i0} and λ_{i0}). That information is unfortunately not available, but we know that the initial number of workers becomes less relevant the further we are from initialization. For this reason we start the initialization from zero for all job types and we use the first 5 years of data $[2010 - 2014]$ as a burn-in period. We then use the last 4 years $[2015 - 2018]$ for the structural estimation of the model's parameters.

Table 1: **Labor Summary Statistics**.

	Data Management λ_{it}	AI analysts L_{it}^{AI}	Traditional analysts L_{it}^{OT}
mean	22.25	0.84	14.95
stdev	97.26	8.45	56.78
minimum	0	0	0
median	4.27	0	2.89
maximum	1986	594	945
Observations	33,610	33,610	33,610

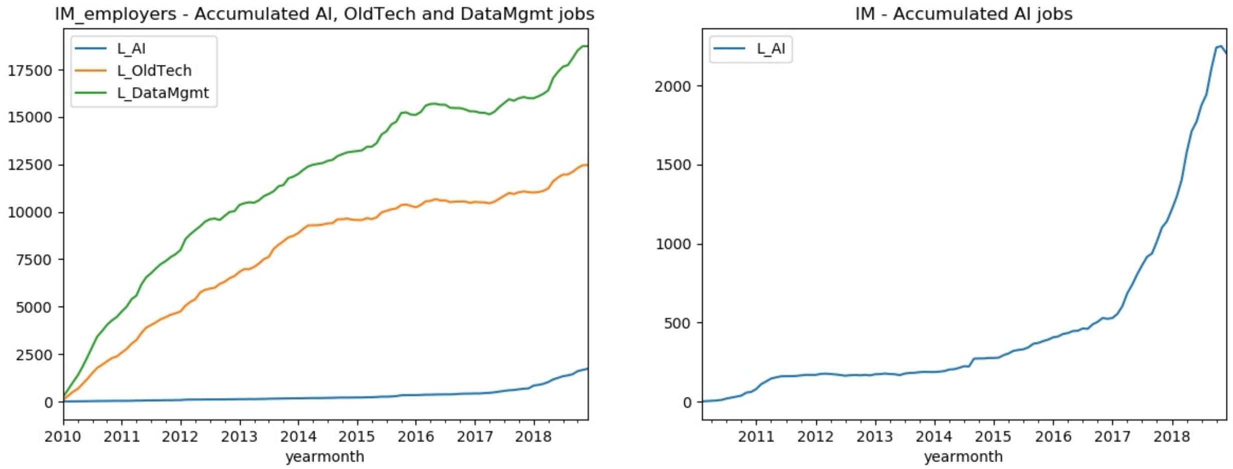


Figure 5: Labor stocks. Panel 1 shows the labor stock in each category, measured as a number of job postings, cumulated, adjusting for filling and separation rates as in (12). This is estimated on 33, 610 relevant job posting observations.Panel 2 shows the same plot for AI jobs jobs only.

Figure 5 shows the imputed labor stocks for each job type. After an initial increasing phase, notice that by 2015, the effect of the initial distortion due to the accumulation period has disappeared. AI workers in finance are still a small fraction of the overall labor supply, suggesting that the transition to a new model of knowledge production is just in its beginnings. Table 1 reports the summary statistics for the stock of each type of labor. What is salient in all three categories is the large disperion. This is helpful because the cross-firm heterogeneity will allow us to estimate the technology parameters.

Cumulating data management to get structured data stocks We measure each firm's stock of data in each period by adding the data management inputs to the depreciated stock of yesterday's data:

$$D_{it} = (1-\delta)^t D_{i0} + \sum_{s=0}^{t} (1-\delta)^{t-s} \lambda_{is}^{1-\phi}. \tag{15}$$

We fix the depreciation rate of data at $\delta = 0.0.25$, which is a 2.5% depreciation rate per month. We also report results for 1% and 10% deprecation. This represents some high-frequency data, whose value lasts for fractions of a second, as well as longer term data used to value companies. In future iterations, we will experiment with other values for depreciation.

To use this approach, we need information about firms' initial data stocks. We estimate this initial stock, by finding the initial stock that makes all subsequent data levels closest to the firm's optimal level. Specifically, the initial data stock of each firm is the D_{i0} that best fits the sequence of the firm's data optimality condition (11).

If we estimate this recursive system of data stocks, production parameters and data inputs for every firm in our sample, the problem quickly becomes unmanageable. At the same time, we do not want to lose the interesting cross-firm heterogeneity. Therefore, instead of estimating D_{it} for each firm in our sample, we compute it for the average firm and use a rule to map the average into a firm's initial data. We use the initial data stocks to estimate the production function parameters. Then, given the parameters, we can recover the best-fit initial data and cumulate up a data stock for each firm easily.

Specifically we express the D_{i0} of each firm as a function of a unique average data stock by setting each firm's initial data proportional to the average data stock and to their cumulative hiring in data management from 2010-2015. In other words, we take the estimated data management labor stock in 2015, $\lambda_{2015,i}$, multiply it by the data management productivity parameter and raise it to the production function exponent to turn it into an amount of data produced: $A^{DM}\lambda_{2015,i}^{1-\phi}$, and then choose a constant ι so that the average initial data stock is the estimated average stock: $(1/N)\sum_i \iota A^{DM}\lambda_{2015,i}^{1-\phi} = \bar{D}_0$.

Then we can express equation 15 as follows:

$$D_{it} = (1-\delta)^t \iota A^{DM}\lambda_{2015,i}^{1-\phi} + \sum_{s=0}^{t}(1-\delta)^{t-s}\lambda_{is}^{1-\phi}. \tag{16}$$

where ι is a function of $\bar{D}_0$. For each firm we then cumulate up the data management flows to construct a stock of data.

$$D_{it} = (1-\delta)^t \iota A^{DM}\lambda_{2015,i}^{1-\phi} + \sum_{s=0}^{t}(1-\delta)^{t-s}A^{DM}\lambda_{is}^{1-\phi} \tag{17}$$

The initial data stock that best explains the sequence of data management hiring is the $\bar{D}_0$ that minimizes the sum of squared errors or the right hand side of (11), for each firm i.

Estimating production functions The key variables of interest are the two production function exponents, α and γ from (1) and (2). There are four variables we need to estimate: α, γ, the exponent ϕ on data management in the structured data production function (3), and finally, we need the initial average data stock $\bar{D}_0$. For three of our moment conditions, we use the first order conditions for each of the three types of labor (6), (7) and (10), for the fourth, we use the optimal data stock condition, (11).

When we estimate the machine learning labor first order condition, we use only firms that employ some machine learning workers and some data management workers. Requiring that the firm currently employs a type of worker does not imply they hired someone that month. Rather, it means that some worker was hired at some time in the past. If we do not exclude these firms, our production exponent estimate would be heavily influenced by the many observations with zero labor and abundant data, or vice-versa. Similarly, when we estimate the traditional labor first order condition, we use only observations from firms that have, at some point, hired a data manager and a traditional analyst.

We also need to solve for the productivity parameters A_t^{AI}, A_t^{OT} and A^{DM}. Given a set of guessed parameters (α, γ, ϕ and $\bar{D}_0$), we solve for A^{DM} from equation 11 computed on cross-sectional and time-series averages. We solve for A_t^{AI}, A_t^{OT} using the first order conditions 6 and 7 computed on cross-sectional averages. In other words, the A parameters reconcile the average magnitudes of knowledge with average wages, while the production exponents are identified off of the cross-firm heterogeneity.

We then substitute the computed productivity parameters into the four conditions and compute a vector of residual using the full time-series and cross-sectional variation. The residual vector contains $(33,610 \times 4)$ observations.

Finally we use non-linear least squares to iterate over different combinations of α, γ, ϕ and $\bar{D}_0$. The algorithm converges when it finds the combination of parameters that yields the smallest sum of squared errors.

3 Results

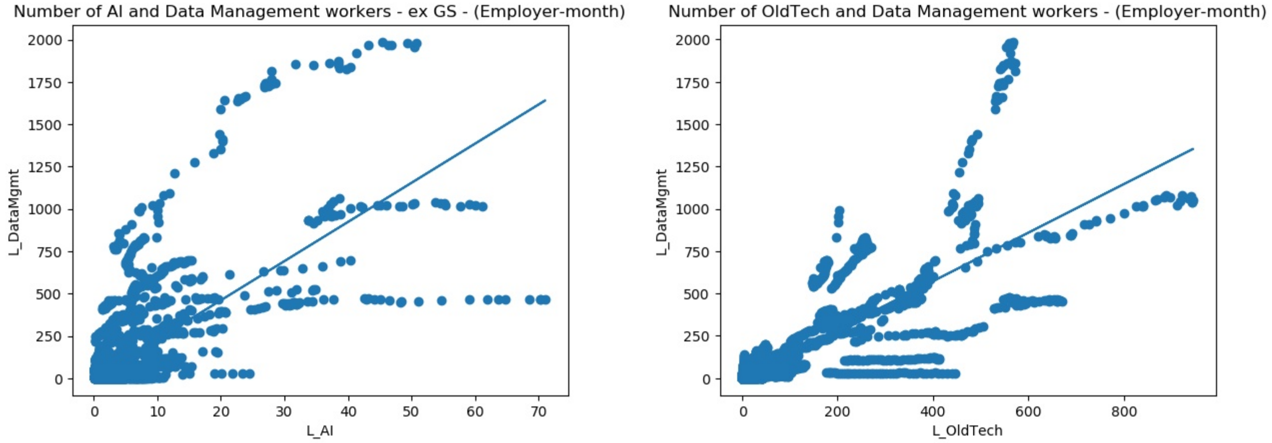


Figure 6: Firms with more structured data hire more AI analysts (left panel) and more traditional analysts (right panel). The left panel excludes Goldman Sachs simply because their hiring is an order of magnitude larger than others. Excluding them makes the rest of the data set more visible. Source: Burning Glass, 2015-2018.

The first question we ask of our data is whether a very basic premise is satisfied: Do firms that hire more data management workers, and thus presumably have larger structured data sets, also hire more analysis workers? Figure 6 shows that the answer is clearly yes. That tells us that at least, our data makes sense.

Table 2: **Main Results: Production Function Exponents on Data**. The estimates are for the exponents on data in the knowledge production functions in (1) and (2) and the production of structured data in (3). Standard errors in parentheses.

		$\delta = 1\%$	$\delta = 2.5\%$	$\delta = 10\%$
Data Management	ϕ	0.172	0.190	0.144
		(0.0025)	(0.0019)	(0.0022)
AI Analysis	α	0.806	0.734	0.613
		(0.0013)	(0.0026)	(0.0038)
Old Technology Analysis	γ	0.458	0.560	0.567
		(0.0024)	(0.0017)	(0.0006)

The other parameter we estimate is the average initial data stock, which is $(0.015, 0.00021, 0.000555)$ for $\delta = (0.01, 0.025, 0.1)$. From here on, we present results for the medium depreciation case of $\delta = 2.5\%$ and report results for the other two cases in the appendix.

Our main question is: What are the production function exponents from each technology? Table 2 reports these main results. The exponents α and γ represent the diminishing returns to data in the old and new technologies. The fact that $\alpha > \gamma$ means that the rate of diminishing returns to data is less with the new AI technology. In other words, new data technology has significantly raised the productivity of analyzing larger data sets. That is not surprising. The fact that the exponent rose by 31% of its previous value suggests that the improvement is not trivial.

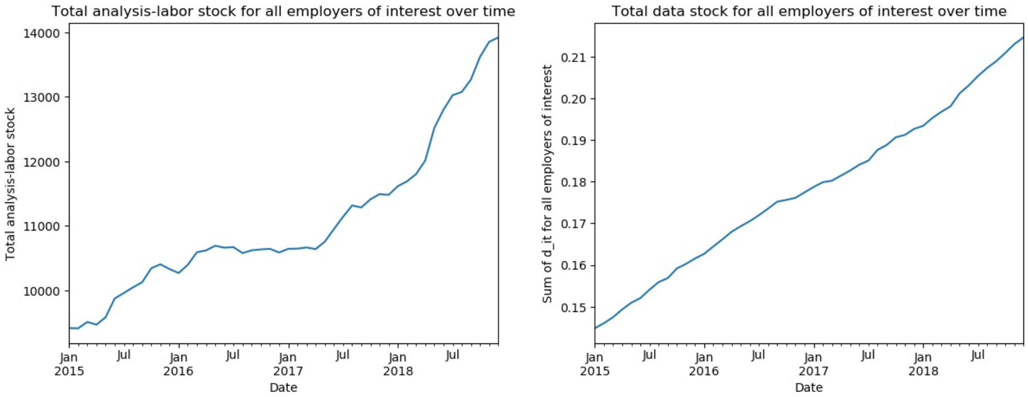
How can we gauge the size of this change in knowledge production. Since this paper is pursuing an analogy between knowledge production with big data technologies and the change in physical production in the industrial revolution, a historical comparison seems most relevant. Klein and Kosobud (1961) estimate that between 1900 and 1920, the labor share of income fell from 0.909 to 0.787. Since the labor share of income corresponds to one minus the exponent on capital in the production function, this estimate suggests that the capital exponent in the production function rose by 0.122. Our rise of 0.174 is even higher than the industrial revolution value. That simple comparisons suggests that the magnitude of the technological change in the big data revolution is at least comparable to that of the industrial revolution. Even when assuming a very high depreciation rate of data (10% monthly) we still obtain a sizable decrease of 0.046, which represents a third and a half of the industrial revolution value.

The labor first order conditions (6) and (7) tell us that these exponents also govern the distribution of income to factor owners. Our results imply that owners of data have gained enormously from this technological change. While they used to be paid 56% of the value of the knowledge output, they can now extract 73.4% of that value. In addition, since more knowledge is being produced, this is 73.4% of a larger number. This finding is consistent with the overall economic trend of a decrease in the labor share of income (Karabarbounis and Neiman, 2017).

Of course, owners of data stock had to pay data managers to build these data sets, just like owners of capital had to pay for the investment in their capital stocks. But once they own these data stocks, they get the income associated with their factor.

3.1 Data Stocks and Labor Stocks

One of the main concerns people have with new data technologies like AI is that they might be labor replacing. Our results do not support that concern.


Total analysis-labor stock for all employers of interest over time

Total data stock for all employers of interest over time

Figure 7: Data Stocks and Labor Stocks. The left panel displays the aggregate stock of analysis labor, including both old and new technology skills. The right panel is the sum of all data stocks, estimated for each firm in our sample. Data depreciation is 2.5% per month ($\delta = 0.025$).Source: Burning Glass and authors' estimates, 2015-2018.

Figure 7 illustrates the aggregate stock of analysis labor and the aggregate stock of data. Both grow rapidly, with data slightly outpacing the growth in analyst labor. Analysis labor here includes both old tech and AI-skilled analysts.

What is most striking about these estimates is that data is not replacing labor. To the contrary, this technological progress in data processing is accompanied by a hiring boom of workers to work with the increasing stock of information. The growing labor force is not an artifact of our parameter estimates. Growing labor is also a feature of the results with 1% or 10% data depreciation. It is also not dependent on model assumotions. The growing labor result comes from simply counting up the new hires and adjusting for BLS-reported departures.

It is true that some of this increase comes from there being more firms in our sample. But the growth of firms working with financial data is hardly a sign of low labor demand. To the contrary, this technology seems to be increasing not decreasing labor demand. Although it is true that AI jobs grew at a faster rate (from about 0 to 2000), they account for only about half of the increase. The other half comes from more hiring of old technology analysts who are also made more productive by the abundance of structured data.

Data in the Cross-Section of Firms Figure 8 illustrates the evolution of the data stock of firms in each percentile of the cross-firm distribution. One thing the results make clear is that the distribution of data is quite skewed. A few firms have enormous troves of data and many have very little.

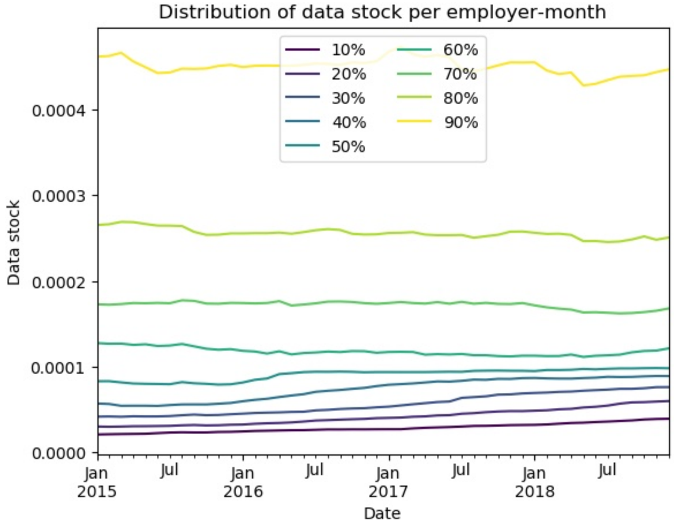


Figure 8: Estimated Stock of Data, Across Firms ($\delta = 2.5\%$), 2015-2018.

Notice in Figure 8 what is not happening. It is not increasing. What is going on here is two-fold. First, the sample of firms is growing over time. Many firms are starting to hire data workers, and thus entering our sample. As a result, the top decile of firms has a lot more firms in it at the end and the firm at the 90th percentile is much lower in the rankings. Second, much of the data accumulation is happening at the top of the distribution.[6] The top 1% of firms is not reflected. The 90th percentile is not an average of the top 10% of firms. It is the stock of the single firm at the 90th percentile. The take-away is that, while the aggregate stock of data is growing rapidly, for many firms, their stock of data is quite stable.

[6] Our data allows us to put names on the firms with these enormous data stockpiles. We hope to be able to report those in the future.

3.2 Estimating the Value of Data

One of the big questions in economics and finance today is how to value firms' data stocks. Four of the five largest firms in the U.S. economy, by market capitalization, have valuations that are well beyond the value that their physical assets might plausibly justify. These firms have future expected revenues based on their accumulated stocks of data. Our structural estimation offers a straightforward way to compute this value.

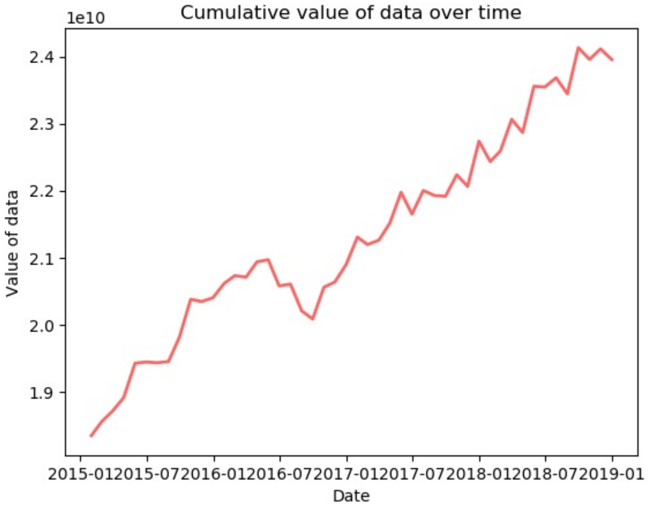


Figure 9: Estimated Value of the Aggregate Stock of Data, in billions of current U.S. dollars, 2015-2018.

Once we have estimated production parameters and data stocks, we can put them back into our value function, and approximate the value of each firm's stock of data in each month. This value is in nominal dollar units, since those are the units of the wages we use. Figure 9 plots this aggregate value. This is our estimate of the value function in (4) for the aggregate stock of data. These results are presented with an important caveat: The wage data we have is sparse. Therefore, it is incredibly volatile. Since the value of data depends very much on the wages of the workers who work with it, results might change once we repeat the estimation using better wages data, which we are in the process of acquiring.

The units of Figure 9 are tens of billions of U.S. dollars. Over the time period, 2015-2018, we see a rise in the value of this data stock from about \$ 18 billion to about \$ 24 billion, a 33% increase in value.

Where does this increase in value come from? The first source is simply the accumulation of data. The right panel of Figure 7 reveals that the aggregate stock of data rose just over 50%. More than half of the increase in the value of data comes from this rise in the size of the structured data stock. A second contributor to the increase in the value of data is the increase in financial analysts that work with data. The more workers there are, the higher is the marginal value of data and the more valuable the stock of data is. The left panel of Figure 7 reveals that the financial analyst labor force grew enormously, almost as much as the data stock did.

Finally, firms are becoming more productive at using data. More productivity also contributes to the rise in the value of data. Figure 10 reports our estimates of the analysis productivity parameters, A^{AI} and A^{OT}, for each month. While productivity with the old technology show no trend over time, the productivity of working with the new (AI) data technologies displays a clear jump in 2017. This productivity jump is additional evidence of the transformative power of new big data technologies.

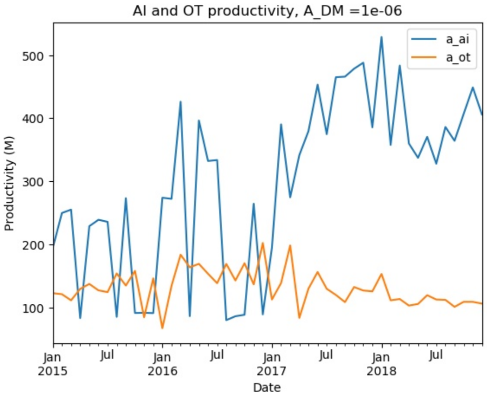


Figure 10: Productivity of Financial Data Analysis, reported for old tech and AI technologies, 2015-2018.

4 Conclusion

Modern discourse describes new big data technologies as the next industrial revolution, or more specifically, as the industrialization of knowledge production. What does that mean? Industrialization was the adoption of new production technologies that involved less human input and less diminishing returns to capital. In other words, the key feature of industrialization is that factor shares changed. Thus if big data technologies are the industrialization of knowledge production, they should offer less diminishing returns to data.

We explored this hypothesis by modeling the production of knowledge, in the same why economists model industrial production. Instead of mixing capital and labor with a Cobb-Douglas production function to produce goods, we described how labor and data can be mixed with a Cobb-Douglas production function to produce knowledge. Then, just as 20th-century economists estimated the exponents of the industrial production function using labor income shares, we similarly measure the exponents of the knowledge production function using wages and labor flows in a particular type of knowledge production, financial analysis. We find a substantial change in the production function, of magnitude larger than the change due to industrialization. Thus, describing this change as a new industrialization seems to be a fair comparison.

Adoption of AI and big data technologies, as well as the accumulation of stocks of data vary widely by firm. The firms with more data are more prone to hire more big-data or AI workers. This supports the idea that this is a technology that is changing the factor mix of production. This finding has important implications for the future of the income distribution: It changes the future labor share of income. In a model that did not have constant returns to scale, such a change would alter the optimal size of a firm: Firms with less diminishing returns to data may well take on a larger optimal size. It also tells us that knowledge will be significantly more abundant going forward.

Two extensions of the model would be useful next steps. One would be to relax the assumption of constant returns to scale in knowledge production. It is possible that doubling data and doubling data workers more than increases the production of knowledge. It is also possible that there is a form of knowledge crowd-out, where it gets harder and harder to produce new knowledge (Bernard and Jones, 1996). We use constant returns because it facilitates a comparison with industrialization, which typically used such production functions. Constant returns also yields a clear mapping from labor shares to production function exponents. In the absence of constant returns, there is considerable dispute about the best way to determine market wages or factor shares. Getting caught up in that debate would distract from the simple main message of this paper.

Another extension would be to consider market power. Owners of data extract rents because data is not perfectly substitutable. Knowledge producing firms also produce differentiated products that allow them to profit. Market power does interact with equilibrium wages. Correcting for it would complicate the mathematics of the model, but could also sharpen the production function estimates.

Of course, this estimation was for workers doing one type of work in one sector. In other sectors, big data might be more or less of a change to output. It may also be too early to tell since machine learning is not widely adopted in most other sectors. Much work in this area remains to be done to understand the magnitude and consequences of the technological changes in data processing that we are currently experiencing.

References

Acemoglu, Daron and Pascual Restrepo, "Artificial Intelligence, Automation and Work," Working Paper 24196, National Bureau of Economic Research January 2018.

_ , David Autor, and Jonathon Hazell, "AI and Jobs: Evidence from Online Vacancies," Working Paper, Massachusetts Institute of Technology October 2019.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," 2017. Stanford GSB Working Paper.

Agrawal, Ajay, John McHale, and Alexander Oettl, "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth," in "The Economics of Artificial Intelligence: An Agenda," National Bureau of Economic Research, Inc, 2018.

_ , Joshua Gans, and Avi Goldfarb, *What to expect from artificial intelligence*, MIT Sloan Management Review, 2017.

_ , _ , and _ , "The economics of artificial intelligence," *McKinsey quarterly*, 2018.

Alekseeva, Liudmila, José Azar, Mireia Gine, Sampsa Samila, and Bledi Taska, "The Demand for AI Skills in the Labor Market," 2020.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "How Does Artificial Intelligence Affect Jobs? Evidence from US Firms and Labor Markets," Technical Report, Working Paper 2020.

Berg, Andrew, Edward F Buffie, and Luis-Felipe Zanna, "Should we fear the robot revolution?(The correct answer is yes)," *Journal of Monetary Economics*, 2018, *97*, 117–148.

Bernard, Andrew B and Charles I Jones, "Comparing apples to oranges: productivity convergence and measurement across industries and countries," *The American Economic Review*, 1996, pp. 1216–1238.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics," Technical Report, National Bureau of Economic Research 2017.

_ , _ , and _ , "The productivity J-curve: How intangibles complement general purpose technologies," Technical Report, National Bureau of Economic Research 2018.

_ , Tom Mitchell, and Daniel Rock, "What can machines learn, and what does it mean for occupations and the economy?," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 43–47.

_ , _ , and _ , "What can machines learn, and what does it mean for occupations and the economy?," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 43–47.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern, "The impact of artificial intelligence on innovation," Technical Report, National bureau of economic research 2018.

Crouzet, Nicolas and Janice Eberly, "Rents and Intangible Capital: A Q+ Framework," 2020. Northwestern University Working Paper.

Deming, David J and Kadeem L Noray, "Stem careers and the changing skill requirements of work," Technical Report, National Bureau of Economic Research 2018.

Farboodi, Maryam and Laura Veldkamp, "A Growth Model of the Data Economy," 2019. Working Paper, MIT.

Felten, Edward W, Manav Raj, and Robert Seamans, "Linking Advances in Artificial Intelligence to Skills, Occupations, and Industries," in "AEA Papers and Proceedings" 2018.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? the effects of machine learning on credit markets," *The Effects of Machine Learning on Credit Markets (November 6, 2018)*, 2018.

Grennan, Jillian and Roni Michaely, "Fintechs and the market for financial analysis," *Michael J. Brennan Irish Finance Working Paper Series Research Paper*, 2018, (18-11), 19–10.

Jones, Chad and Chris Tonetti, "Nonrivalry and the Economics of Data," 2018. Stanford GSB Working Paper.

Karabarbounis, Loukas and Brent Neiman, "Trends in factor shares: Facts and implications," *NBER Reporter*, 2017, (4), 19–22.

Klein, Lawrence R and Richard F Kosobud, "Some econometrics of growth: Great ratios of economics," *The Quarterly Journal of Economics*, 1961, *75* (2), 173–198.

Webb, Michael, "The Impact of Artificial Intelligence on the Labor Market," *Available at SSRN 3482150*, 2019.

A Appendix: Identifying Investment Management Jobs

We identify Investment management jobs as those that require at least one skill belonging to the following Burning Glass skill clusters: 'Asset Management Industry Knowledge', 'Electronic Trading Systems', 'Investment Management', 'Financial Trading', 'Financial Trading Industry Knowledge', 'Investment Services Industry Knowledge', 'Financial Advisement'.

This list of skill clusters was compiled by tabulating all skill clusters required by any of the jobs in our sample and selecting those most related to investment management.

Since sometimes skills clusters are missing, we compile a list of all skills ever present in the list of relevant skill clusters and also classify as 'investment management' those jobs that require at least one of those underlying skills.

We finally check the full list of skills required by the selected jobs and exclude those jobs which require the following skills, as we believe these jobs are not likely to be actual investment management jobs: 'Marketing Strategy', 'General Marketing', 'Urban Planning', 'Technical Support', 'Telemarketing', 'Business-to-Business (B2B) Sales', 'Marketing Automation', 'Litigation', 'Retail Sales', 'Billing and Invoicing', 'General Administrative and Clerical Tasks', 'Journalism', 'Claims Processing', 'Merchandising', 'Carpentry', 'Animation and Game Design', 'Basic Customer Service', 'Cash Register Operation', 'Real Estate and Rental', 'Marketing Software', 'Online Marketing', 'Accounts Payable and Receivable', 'Packaging and Labeling', 'Inventory Management', 'Advanced Customer Service', 'Payroll', 'Underwriting', 'Marketing Management', 'Supply Chain Planning'.

B Categorizing Jobs

Jobs are first categorized into 'data management' (DM) and 'data analysis' by looking at the relative frequency of the 'data management' vs. 'data analysis' keywords listed below in the full text of the underlying job postings. Jobs identified as 'data analysis' are further categorized (where possible) as AI or old technology (OT), by looking at the relative frequency of the AI and OT keywords listed below - these are subsets of the 'data analysis' keywords.

All keywords lists are obtained by first tabulating all Burning Glass skills present in the selected sample and identifying skills that best map to the types of jobs described by the model. We then also inspected the text of selected job postings requiring most of the selected skills in order to refine the keywords and phrases to best reflect the format in which they are most frequently present in the text.

Before computing relative frequencies both the keywords lists and the underlying text are pre-processed and stemmed to their root using the Porter stemmer.

Data Management keywords : 'Apache Hive', 'Information Retrieval', 'Data Management Platform (DMP)', 'Data Collection', 'Data Warehousing', 'SQL Server', 'Data Visualization', 'Database Management', 'Data Governance', 'Data Transformation', 'Extensible Markup Language (XML)', 'Data Validation', 'Data Architecture', 'Data Mapping', 'Oracle PL/SQL', 'Database Design', 'Data Integration', 'Teradata', 'Database Administration', 'BigTable', 'Data Security', 'Database Software', 'Data Integrity', 'File Management', 'Splunk', 'Relational DataBase Management System', 'Teradata DBA', 'Data Migration', 'Information Assurance', 'Enterprise Data Management', 'SSIS', 'Sybase', 'jQuery', 'Data Conversion', 'Data Acquisition', 'Master Data Management', 'Data Capture', 'Data Verification', 'MongoDB', 'Data Warehouse Processing', 'SAP HANA', 'Data Loss Prevention', 'Data Engineering', 'Database Schemas', 'Database Architecture', 'Data Documentation', 'Data Operations', 'Oracle Big Data', 'Domo', 'Data Manipulation', 'Data Management Platform', 'DMP', 'HyperText Markup Language', 'Data Access Object (DAO)', 'Structured Query Reporter', 'SQR', 'Data Dictionary System', 'Data Entry', 'Data Quality', 'Data Collection', 'Information Systems', 'Information Security', 'Change data capture', 'Data Management', 'Data Governance', 'Data Encryption', 'Data Cleaning', 'Semi-Structured Data', 'Data Evaluation', 'Data Privacy', 'Dimensional and Relational Modeling', 'Data Loss Prevention', 'Data Operations', 'Relational Database Design', 'Database Programming', 'Information Systems Management', 'Database Tuning', 'Object Relational Mapping', 'Columnar Databases', 'Datastage', 'Data Taxonomy', 'Informatica Data Quality', 'Data Munging', 'Data Archiving', 'Warehouse Operations', 'Solaris', 'Data Modeling', 'data feed management', 'data discovery', 'exporting large datasets', 'exporting datasets', 'database performance', 'disigning relational databases', 'implementing relational databases', 'designing and implementing relational databases', 'database development', 'data production process', 'normalize large datasets', 'normalize datasets', 'create database', 'Develop database', 'data onboarding', 'Data Sourcing', 'data purchase', 'data inventory', 'cloud Security', 'negotiating data', 'data attorney', 'data and technology attorney', 'reliability engineering', 'reliability engineer', 'data specialist', 'enable vast data analysis', 'enable data analysis', 'Data team', 'capturing data', 'processing data', 'Supporting data', error free data sets', 'error free datasets', 'live streams of data', 'data accumulation', 'Kernel level development', 'large scale systems', 'Hadoop', 'distributed computing', ' multi database web applications', 'connect software packages to internal and external data', 'explore data possibilities', 'architect complex systems', 'build scalable infrastructure for data analysis', 'build infrastructure for data analysis', 'solutions for at scale data exploration', 'solutions for data exploration', 'information technology security', 'security engineer', 'security architect', 'architect solutions to allow modelers to process query and visualize higher dimensional data'

Analysis keywords

- General Analysis: 'Regression Algorithms', 'Regression Analysis', 'Quantitative Analysis', 'Clustering', 'Time Series Analysis', 'Economic Analysis', 'Model Building', 'Quantitative Research', 'pandas', 'numpy', 'Hedging Strategy', 'Quantitative Data Analysis', 'Investment Analysis', 'Economic Models', 'Predictive Analytics', 'Market Trend', 'Portfolio Optimization', 'Portfolio Rebalancing', 'Financial Derivatives Pricing', 'Active Alpha Generation', 'Financial Data Interpretation', 'Alteryx', 'Predictive Models', 'Exploratory Analysis', 'Sensitivity Analysis', 'News Analysis', 'Asset Allocation', 'Research Methodology', 'Mathematical Software', 'Portfolio Construction', 'Portfolio Analysis', 'Portfolio Analyst', 'Market Analysis', 'Data Techniques', 'Capital Allocation', 'Financial Modeling', 'Algorithm Development', 'Securities Trading', 'Trading Strategy', 'Statistical Programming', 'Data Mining', 'Social Network Analysis', 'Dimensionality Reduction', 'Principal Components Analysis (PCA)', 'Statistical Software', 'Portfolio Management', 'Numerical Analysis', 'Time Series Models', "Asset Allocation Theory", 'Analytical Skills', 'Financial Analysis', 'Financial Modeling', 'Modern Portfolio Theory', 'MPT', 'Portfolio Valuation', 'strategic portfolio decisions'

- Old Technology: 'Linear Regression', 'Logistic Regression', 'Statistic', 'STATA', 'Emacs', 'Technical Analysis', 'Qualitative Analysis', 'Qualitative Portfolio Management', 'Data Trending', 'Stochastic Optimization', 'Multivariate Testing', 'Bootstrapping', 'Time Series Models', 'Factor Analysis', 'Durations analysis', 'Markov', 'HMM', 'Econometrics', 'Stochastic Processes', 'Calculus', 'Statsmodels', 'Linear Algebra', 'Mathematics', 'Maths', 'Monte Carlo Simulation', 'Generalized Linear Model', 'GLM', 'Linear Programming', 'Bayesian', 'Analysis Of Variance', 'ANOVA', 'Behavioral Modeling', 'Black-Scholes', 'Behavior Analysis', 'Discounted Cashflow', 'Numerical Analysis', 'Correlation Analysis', 'E-Views', 'Differential Equations', 'Algebra', 'Value at Risk', 'Asset Pricing Models', 'Statistician', 'Mathematician', 'Econometrician'

- AI: 'Artificial Intelligence', 'Machine Learning', 'Natural Language Processing', 'NLP', 'Speech Recognition', 'Gradient boosting', 'DBSCAN', 'Nearest Neighbor', 'Supervised Learning', 'Unsupervised Learning', 'Deep Learning', 'Automatic Speech Recognition', 'Torch', 'scikit-learn', 'Conditional Random Field', 'TensorFlow', 'Tensor Flow', 'Platfora', 'Neural Network', 'CNN', 'RNN', 'Neural nets', 'Decision Trees', 'Random Forest', 'Support Vector Machine', 'SVM', 'Reinforcement Learning', 'Torch', 'Lasso', 'Stochastic Gradient Descent', 'SGD', 'Ridge Regression', 'Elastic-Net', 'Text Mining', 'Classification Algorithms', 'Image Processing', 'Natural Language Toolkit', 'NLTK', 'Pattern Recognition', 'Computer Vision', 'Long Short-Term Memory', 'LSTM', 'K-Means', 'Geospatial Intelligence', 'Big Data Analytics', 'Latent Dirichlet Allocation', 'LDA', 'Backpropagation', 'Machine Translation', 'Caffe Deep Learning Framework', 'Word2Vec', 'Genetic Algorithm', 'Evolutionary Algorithm', 'Data Science', 'Sentiment Analysis / Opinion Mining', 'Maximum Entropy Classifier', 'Neuroscience', 'Computational Linguistics', 'Semi-Supervised Learning', 'Data Scientist'

C Constructing the Labor Inflows Data

Job openings, filling and separation data Our data comes from

`https://www.bls.gov/news.release/jolts.tn.htm`

Job Openings Rate: Job openings information is collected for the last business day of the reference month. A job opening requires that: 1) a specific position exists and there is work available for that position, 2) work could start within 30 days whether or not the employer found a suitable candidate, and 3) the employer is actively recruiting from outside the establishment to fill the position. The job openings rate is computed by dividing the number of job openings by the sum of employment and job openings and multiplying that quotient by 100.

Hiring Rate: The hires level is the total number of additions to the payroll occurring at any time during the reference month, including both new and rehired employees, full-time and part-time, permanent, short-term and seasonal employees, employees recalled to the location after a layoff lasting more than 7 days, on-call or intermittent employees who returned to work after having been formally separated, and transfers from other locations. The hires rate is computed by dividing the number of hires by employment and multiplying that quotient by 100.

Separations Rate: The separations level is the total number of employment terminations S occurring at any time during the reference month, and is reported by type of separation - quits, layoffs and discharges, and other separations. The separations rate is computed by dividing the number of separations by employment and multiplying that quotient by 100: $s = S/E \cdot 100$.

Deriving the probability of filling an opening. If n_O is the total number of posted job openings, n_E is total employment and n_H is the number of new hires in this sub-occupation and month, then the BLS hiring rate is defined to be $r_h = n_H/n_E$, while the job opening rate is $r_o = n_O/(n_E + n_O)$. What we need to adjust the openings data from our model, is the fraction of openings that result in hires, $h = n_H/n_O$.

To solve for h, note that rearranging the definition of the opening rate yields $r_o = (1 - r_o)n_O/n_E$. Dividing r_h by this expression yields $r_h/r_o = (n_H/n_E)/((1-r_o)n_O/n_E) = (n_H/n_O) \cdot 1/(1-r_o)$. Therefore, we can express the n_H/n_O rate we want as $h = r_h(1-r_o)/r_o$.

Time to Fill a Job Vacancy In our calculations, we have implicitly equated a job posting with a one-month job vacancy. We do that because most of our job postings remain up and unfilled for approximately one month. Below, we report the distribution of the average time that job postings remain open in our data set. This data is for jobs that have the same occupations and regions as our sample for the years 2015, 2016 and 2017. The average time to fill is available for 86% of all the occupation (SOC) - region (MSA) combinations in our sample. Below is the distribution of the average time a Burning Glass job posting stayed online for all the SOC-MSA combinations in our sample for 2015-2017.

Table 3: **Time to Fill Posted Vacancies**.

mean	35.6857
std	7.1003
min	14.0000
1%	21.0000
5%	24.0000
10%	27.0000
15%	28.0000
20%	30.0000
25%	31.0000
30%	32.0000
35%	33.0000
40%	34.0000
45%	35.0000
50%	35.0000
55%	36.0000
60%	37.0000
65%	38.0000
70%	39.0000
75%	40.0000
80%	41.0000
85%	43.0000
90%	44.4000
95%	48.0000
99%	54.0000
max	75.0000

If we weight each of these fill times by the number of jobs present in our sample for each the SOC-MSA combinations, we get an average fill times of 38.12 days.

D Model derivations

Firm i faces the following optimizing problem:

$$v(D_{it}) = \max_{\lambda_{it},L_{it},l_{it}} D_{it}^{\alpha} L_{it}^{1-\alpha} + D_{it}^{\gamma} l_{it}^{1-\gamma} - w_{L,t}L_{it} - w_{l,t}l_{it} - w_{\lambda,t}\lambda_{it} + \frac{1}{r}v(D_{i(t+1)}) \tag{18}$$

$$\text{where} \;\; D_{i(t+1)} = (1-\delta)D_{it} + \lambda_{it}^{1-\phi}. \tag{19}$$

Here the state variable is structured data D_{it}, and the control variables are data management labor λ_{it}, the machine learning analyst labor L_{it} and the old technology analysis labor l_{it}. Plugging (19) into (18), we have

$$v(D_{it}) = \max_{\lambda_{it},L_{it},l_{it}} D_{it}^{\alpha} L_{it}^{1-\alpha} + D_{it}^{\gamma} l_{it}^{1-\gamma} - w_{L,t}L_{it} - w_{l,t}l_{it} - w_{\lambda,t}\lambda_{it} + \frac{1}{r}v\left((1-\delta)D_{it} + \lambda_{it}^{1-\phi}\right) \tag{20}$$

Taking partial derivative with respect to L_{it}, we have

$$(1-\alpha)D_{it}^{\alpha}L_{it}^{-\alpha} - w_{L,t} = 0 \implies \frac{(1-\alpha)K_{it}^{AI}}{L_{it}} = w_{L,t}. \tag{21}$$

Taking partial derivative with respect to l_{it}, we have

$$(1-\gamma)D_{it}^{\gamma}l_{it}^{-\gamma} - w_{l,t} = 0 \implies \frac{(1-\alpha)K_{it}^{OT}}{L_{it}} = w_{l,t}. \tag{22}$$

Taking partial derivative with respect to λ_{it} and rearranging, we have

$$\frac{1}{r}v'(D_{i(t+1)})(1-\phi)\lambda_{it}^{-\phi} = w_{\lambda,t}. \tag{23}$$

We then total differentiate (20) to get

$$v'(D_{it}) = \frac{\alpha K_{it}^{AI}}{D_{it}} + \frac{\gamma K_{it}^{OT}}{D_{it}} + \frac{1}{r}v'(D_{i(t+1)})(1-\delta). \tag{24}$$

If we further assume that the marginal value of data today and tomorrow are similar, then

$$v'(D_{it}) = \frac{(\alpha K_{it}^{AI} + \gamma K_{it}^{OT})}{D_{it}} \frac{r}{r-(1-\delta)}. \tag{25}$$

Plugging it back to the first order condition (23) and combining it with the structured data dynamics (19), we arrive at

$$\frac{(\alpha K_{it}^{AI} + \gamma K_{it}^{OT})(1-\phi)}{r-(1-\delta)} \frac{D_{i(t+1)} - (1-\delta)D_{it}}{D_{it}} = w_\lambda \lambda_{it}. \tag{26}$$

E Robustness

Figures 11 and 12 illustrate the evolution of the data stock of firms in each percentile of the cross-firm distribution for 1% and 10% monthly rates of data depreciation.

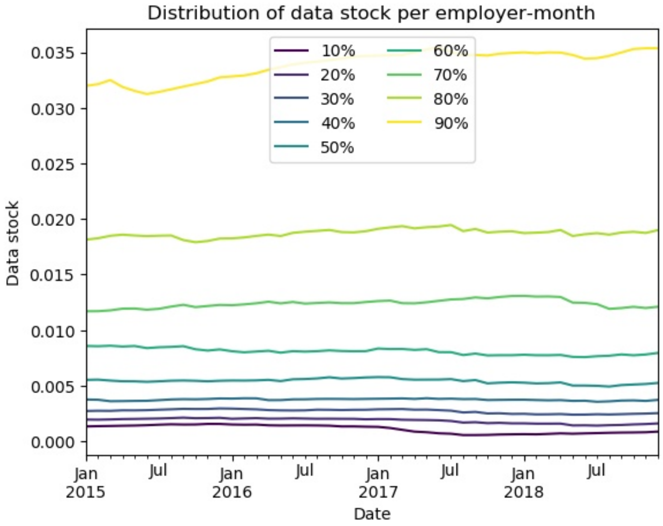


Figure 11: Estimated Stock of Data Processing Workers Per Firm with 1% data depreciation. Source: Burning Glass, 2015-2018.

Once we have estimated production parameters and data stocks, we can put them back into our value function, and approximate the value of each firm's stock of data in each month. This value is in nominal dollar units, since those are the units of the wages we use. Figures 13 and 14 plot this aggregate value for data depreciation of 1% and 10% per month. This is our estimate of the value function in (4) for the aggregate stock of data. The units of Figures 13 and 14 are billions and tens of billions of U.S. dollars respectively. Over the time period, 2015-2018, we see a rise in the value of this data stock.

Finally, firms are becoming more productive at using data. More productivity also contributes to the rise in the value of data. Figures 15 and 16 report our estimates of the analysis productiity parameters, A^{AI} and A^{OT}, for each month, for data depreciation rates of 1% and 10% per month. While productivity with the old technology show no trend over time, the productivity of working with the new (AI) data technologies displays a clear jump in 2017.

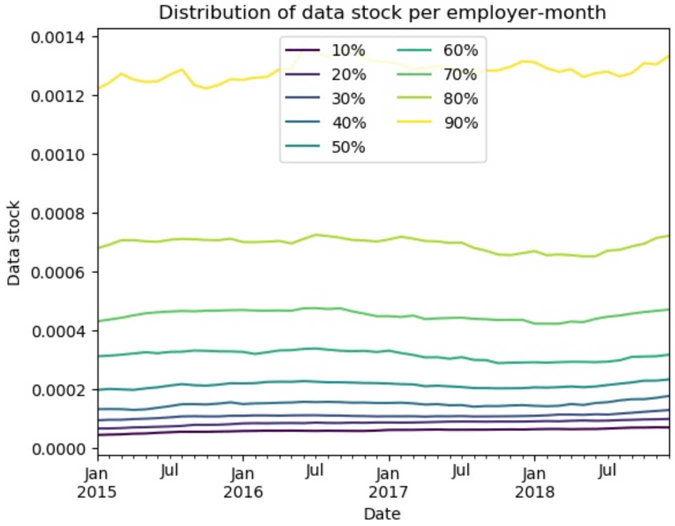


Figure 12: Estimated Stock of Data Processing Workers Per Firm with 10% data depreciation. Source: Burning Glass, 2015-2018.

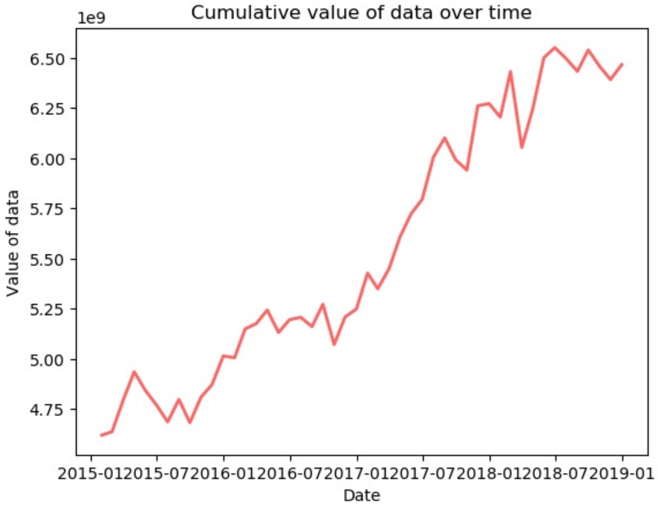


Figure 13: Estimated Value of the Aggregate Stock of Data with 1% data depreciation, in billions of current U.S. dollars, 2015-2018.

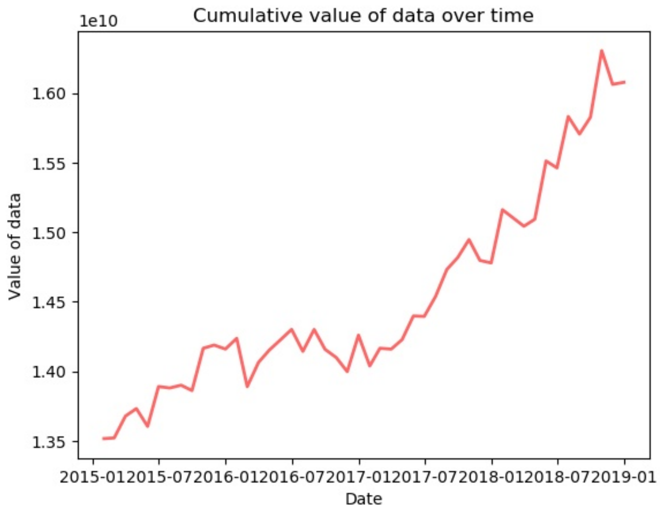


Figure 14: Estimated Value of the Aggregate Stock of Data with 10% data depreciation, in billions of current U.S. dollars, 2015-2018.

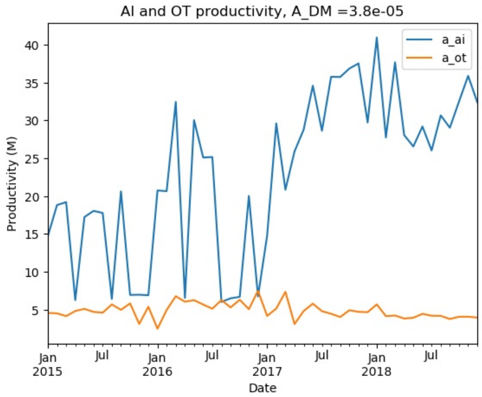


Figure 15: Productivity of Financial Data Analysis, reported for old tech and AI technologies with 1% data depreciation, 2015-2018.

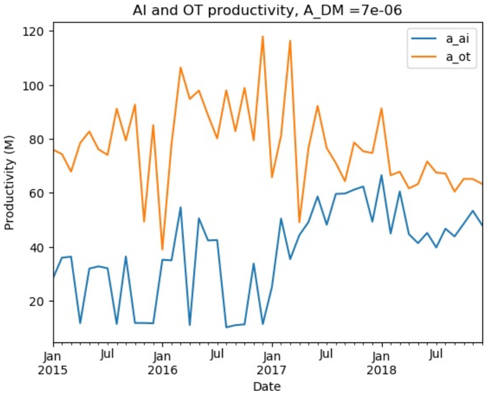


Figure 16: Productivity of Financial Data Analysis, reported for old tech and AI technologies with 10% data depreciation, 2015-2018.