4-1-2020

GDPR and the Importance of Data to AI Startups

James Bessen

Stephen Michael Impink

Lydia Reichensperger

Robert Seamans

GDPR and the Importance of Data to AI Startups*

James Bessen, Technology & Policy Research Initiative, Boston University

Stephen Michael Impink, Stern School of Business, New York University

Lydia Reichensperger, Technology & Policy Research Initiative, Boston University

Robert Seamans, Stern School of Business, New York University

Abstract: What is the impact of the European Union's General Data Protection Regulation ("GDPR") and data regulation on AI startups? How important is data to AI product development? We study these questions using unique survey data of commercial AI startups. AI startups rely on data for their product development. Given the scale and scope of their business models, these startups are particularly susceptible to policy changes impacting data collection, storage and use. We find that training data and frequent model refreshes are particularly important for AI startups that rely on neural nets and ensemble learning algorithms. We also find that firms with customers in Europe are significantly more likely to create a new position to handle GDPR-related issues or to reallocate firm resources due to GDPR.

JEL codes: O33, J21, L10

Keywords: artificial intelligence, automation, technology, data, data regulation, data protection, EU data protection reform

*Thanks to Chen Meng for excellent research support. Thanks to Tim O'Reilly and O'Reilly Media, the Creative Destruction Lab, Philipp Hartmann, and Joachim Henkel for help in acquiring our sample. This work was funded by the Ewing Marion Kauffman Foundation. The contents of this publication are solely the responsibility of the authors.

As described in the AI Index 2018 Annual Report (Shomham et al. 2018), artificial intelligence ("AI") has advanced rapidly over the past decade. Many scholars believe that AI has the potential to boost human productivity and economic growth (Furman & Seamans 2019). Scholars also worry that these gains may come at a cost, potentially including labor displacement, income inequality and loss of privacy. AI algorithms rely on lots of data, often including data on individuals. In an effort to protect consumers' privacy, a number of regulators have passed or considered laws restricting use and sharing of data, including the European Union's General Data Protection Regulation ("GDPR") and California's Consumer Privacy Act ("CCPA"). Even though this increased regulation is intended to protect consumers' privacy, the legislation may negatively impact firms that need data to develop AI products. These burdens could be particularly costly for startups in Europe (Jia et al. 2019). In this paper we report results from a survey designed to assess which AI startups rely more heavily on data and how GDPR affects these startups.

AI relies on large quantities of data. These data are used to train and tune algorithms. Certain types of algorithms, such as neural network and ensemble learning algorithms, support more complex tasks and require more training data[1]. Also, certain technologies are more difficult to develop. For example, a startup that wants to train a chatbot's underlying natural language comprehension capabilities would benefit from using neural networks and relatively large amounts of training data, as compared with other less sophisticated technologies or algorithms. It is apparent through numerous contests, including those leading up to the prestigious Loebner Prize[2] for most "human-like" AI chatbots, that data is a key ingredient for success[3]. The need for data does not diminish with firm size; larger and smaller firms targeting the creation of similar AI products require similar data resources. However, larger firms may be able to access data more easily from supplier and customer relationships as they benefit from a breadth of supplier relationships and a

[1] https://www.eff.org/ai/metrics
[2] https://aisb.org.uk/new_site/?page_id=2
[3] https://aichat.com/2019/06/27/data-is-the-key-to-develop-a-truly-conversational-chatbot/

more developed customer ecosystem. Additionally, larger firms could benefit from complementary business models which provide data as an externality of normal business operations. So, to continue the example, firms that have user-based platforms as part of another business line may be able to reuse that customer chat data to develop their chatbot.

In addition to greater access to data through relationships, larger firms also have access to additional capital to hire computer scientists and engineers (Athey & Luca 2019). It's even difficult for high-growth potential startups to raise capital (Nanda 2016). Though startups benefit from cloud computing and other variable cost IT resources (Jin et al. 2018), larger firms could have an over-abundance of these IT resources.. Slack resources, such as excess cloud computing capabilities, could be used to run valuable experiments (Thomke 2003, Varian 2014) or to develop infrastructures that capture and store large amounts of customer data. More so, high technology firms may be able to avoid inertial tenancies to use outdated or aging IT resources

There is an apparent tradeoff between access to training data and consumer privacy. GDPR and other types of data regulation make it harder for firms to collect, store and analyze certain types of data, especially personally identifiable or employment data. Also, these regulations may impact the willingness of other firms to enter into data sharing collaborations.

In a prior paper, "The Business of AI Startups" (Bessen et al. 2019) we provide the results of a first-round survey of AI startups and discuss how these startups' AI products impact labor. In this paper, we provide the results of a second-round survey of AI startups, which includes an additional 7 questions on data importance and the impact of the European Union's General Data Protection Regulation ("GDPR"). We designed our survey to address two questions. First, we address the impact of the GDPR and data regulation on AI startups. Second, we examine the importance of data to AI product development. We find that training data is important for AI startups that rely on neural nets and ensemble learning algorithms. We also find that firms with customers in Europe are significantly more likely to create a new position to handle GDPR-related

issues or to reallocate firm resources due to GDPR. This implies that the GDPR imposes costs, perhaps substantial costs, on startup AI firms.

Our paper contributes to the literature by providing some of the first evidence on the relationships between data access, means of access, firm strategy, and technology choices. These findings have direct significance for data privacy policy. This paper proceeds as follows. In the next section, we discuss related literature on data privacy. We then describe the survey data and respondent demographics. Lastly, we describe our new findings on the importance of data to AI startups and explore how data privacy regulation can affect these firms.

Related Literature on Data Privacy

Data privacy and protection continue to be a point of intense debate in research, policy and mass media. Several high visibility data breaches, from Equifax[4] and Facebook/Cambridge Analytica[5] in 2017 to the alleged hacking of Jeff Bezos'[6] mobile phone more recently, have received significant attention in the news. Consumers continue to pressure regulators and legislators to more effectively safeguard their data and privacy. However, this increased regulation creates tradeoffs between safeguarding personally identifiable information and data access for entrepreneurial activities.

The European Union ("EU") passed the GDPR in 2016, but the United States ("US") and other similarly advanced countries have yet to pass substantial regulatory policies. Given the interconnectedness of world economies, many firms are compliant with GDPR regardless of their headquarters' location because they have customers or operations in the EU. GDPR is being used as a rubric to frame similar legislation in other countries. More recently, CCPA[7] which focuses on

[4] https://www.ftc.gov/enforcement/cases-proceedings/refunds/equifax-data-breach-settlement
[5] https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html
[6] https://www.nytimes.com/2020/01/22/technology/jeff-bezos-hack-iphone.html
[7] https://www.nytimes.com/2020/01/03/us/ccpa-california-privacy-law.html

the rights to access your personal data from technology providers, to know what personal data is being collected and stored by employers, to delete one's data, and to prevent the sale of one's data. Even though other states lack similarly exhaustive policies, firms that conduct interstate business or sell their products online often adhere to the most stringent state's guidelines.

Prior research argues that GDPR increases the costs of collecting and using customer data. Revenues from online sales for EU firms impacted by GDPR enforcement in 2018, dropped by 10% (Goldberg et al. 2019). Additionally, GDPR has asymmetrically impacted smaller firms in some industries. Enforcement of GDPR led to a reduction in the number of smaller web technology vendors used, leading to increased concentration of more established, larger firms in the web technology industry (Johnson & Shriver 2020). Rates of venture capital funding of startups in the EU also declined during this time period in comparison with the United States (Jin et al. 2019). After GDPR was legislated, many websites even outside of the EU were less likely to share personal data with web technology providers. At the same time, Google increased market concentration while smaller firms lost significant share, raising concerns over possible negative externalities to competition (Batikas et al. 2020).

The use of "big data" raises numerous critical questions for regulation and policies focused on safeguarding personal data (Boyd & Crawford 2012). Until GDPR1 and CCPA, there was little guidance for firms on managing data privacy. The topic of data privacy and protection is intertwined with that of AI due to the necessity of data in product development. Often, personally identifiable information is intermingled with other firm data prompting many legal scholars to discuss data ownership and exclusion rights. Privacy concerns arise when firms may analyze data that include information about specific individuals. Also, this more personal data could be used in a way that leads to biased decision-making (Cowgill & Tucker 2019). Computer scientists and programmers may not have the training needed to create models that are less biased; data that is less suited to the task could further exacerbate issues of bias.

There is some concern that GDPR and other data regulations, specifically in Europe, adversely affect entrepreneurial ventures (Jia et al. 2018). Even though some smaller firms with less than $1M in revenue are exempt from GDPR, this regulation has become the *de facto* standard. Ultimately these startups are targeting swift revenue growth and investors want to know that they can quickly be compliant to regulatory policies. Also, systems need to be initially designed correctly to enable adherence. For example, the ability to delete customer data requires that you are able to search all your data sources by a single customer's name or identifier.

Survey Data

We study these issues using unique data from two online surveys. The first survey was administered from May to September 2018 and the second survey was administered from October 2019 to January 2020, through a questionnaire. The first questionnaire contained 22 questions. The second questionnaire contained 29 questions, including additional questions on the impact of GDPR, regulation and data usage. Both surveys were pretested with half a dozen academics and practitioners associated with startups. Potential respondents were founders, CEOs, CTOs or other similar executives and were contacted via email and alternate methods of communication such as professional-social networks (LinkedIn, Twitter).

Respondents from our survey come from several sampling frames, however the largest frame of our sample comes from Crunchbase. We used Crunchbase to select firms that are tagged with "artificial intelligence" as a description keyword, have positive funding, are still operating, and have not yet experienced an IPO. The Crunchbase sample grew from 1,246 firms in May 2018 to more than 3,000 firms in September 2019. We also received contact lists of AI firms from the Creative Destruction Lab, a startup incubator based in Toronto, and Philipp Hartmann and Joachim Henkel (Hartmann & Henkel 2018). Also, for both rounds of the survey, O'Reilly Media ran a notice of the survey in its AI newsletter, providing a link to the online questionnaire.

The first-round of the survey includes 157 responses. The second-round of the survey includes 199 responses. The total number of responding AI startups amounted to 31 firms that responded to both rounds of the survey. No responses were received from China in the second-round of the survey. Across the two surveys, we reached out to an audience of 2,975 firms. We estimate that about 5% of these startups are not addressable in our study as they are located in China or are no longer in business (emails returned to sender) leading to a 12% response rate overall. We dropped two observations in which the respondent indicated that their firm was not involved with AI.

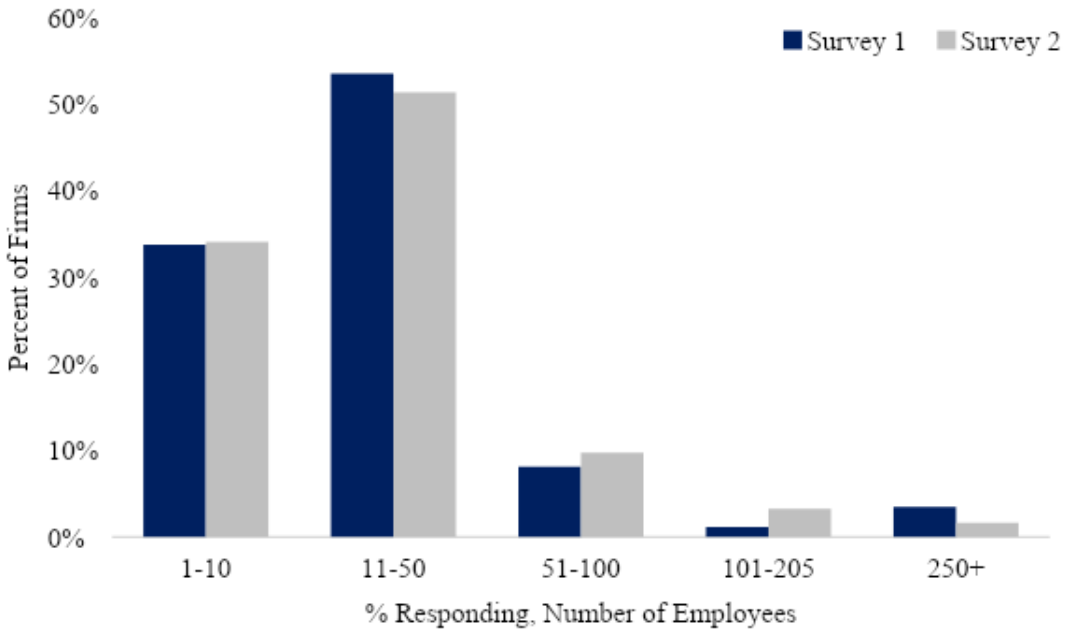
Comparison between respondents and survey rounds. We use t-tests to determine if the subsample means for several variables including number of employees (five size classes), age, number of investors, number of rounds of investment, and total amount of investment are significantly different for respondents and non-respondents. There are no differences significant at the 5% level. We also test to determine if measures and firm demographics are significantly different in the first and second round of the surveys. Of the measures tested, firm age is found to be significantly different at the 5% level. We do find, however, that a few of the firms in the first round of the survey are outliers in terms of their high number of employees (>250 employees).

We also run t-tests to determine if the sample of firms that responded to both surveys is significantly different from the sample of firms that responded to only one survey at both t=1 and t=2. There are significant differences in firm age and usage of proprietary data for firms responding to both surveys versus only responding at t=2, however there are no other significant differences in size, geographic location, customer location, customer size, use of algorithms, measures of product's ability to automate tasks, reduce labor costs or eliminate professions or measures of funding. Firms that completed both surveys are on average 6.1 years old, whereas firms that completed only one survey are on average 5 years old at t=2. All T-test results are reported in the Appendix.

Respondent Demographics

Size. In both rounds of the survey, median size is similar with the majority of firms having less than 50 employees. Around 35% of firms have less than ten employees (see Figure 1, below). Across both surveys, firms that are shipping a product are on average older, larger, at a later funding stage and have experienced more funding rounds. Firms based in the US are generally larger and older than firms based internationally. Most of the firms surveyed (187 firms) are at the seed investment funding stage.

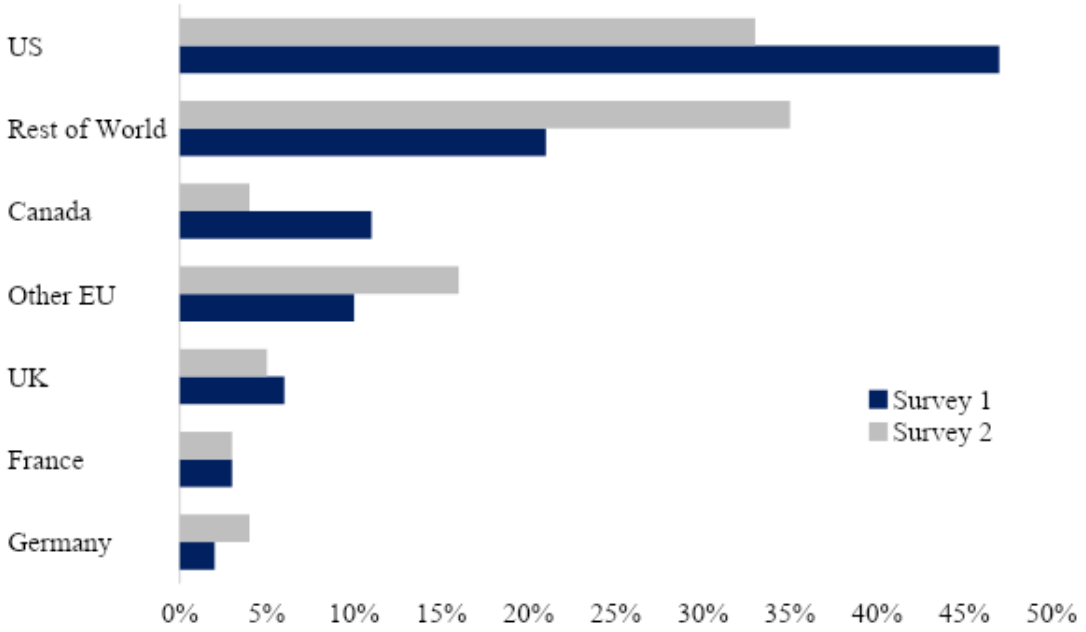
Figure 1. Employees per firm, respondents in Survey 1 and Survey 2



Headquarters Location. We use a Pearson's chi-squared test to determine that response rate and aggregate geographical groupings of Europe and the US are independent, so these subsets can be reported separately or aggregated. Response rates for the US, where the majority of companies surveyed is around 8%. Rest of World and parts of Europe were more highly represented, partly due to an increase in Crunchbase's reporting of international AI firms. As noted in Figure 2 below, the proportion of respondents from the US drops by more than 10%, with the bulk of the drop being

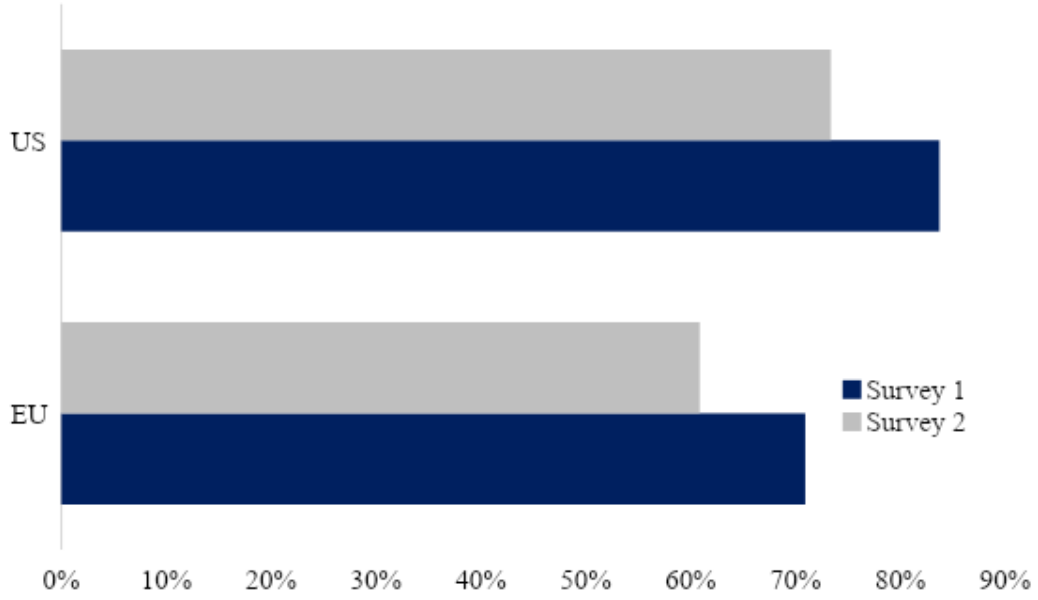
picked up by the Rest of World grouping.

Figure 2. Geographic Distribution of Startup Firms



Customer Location. In survey two, partly due to the increased number of respondents from Rest of World, the percent of firms with customers in Europe and the US declines. However, these differences are not significant. For the total survey population, around 80% of respondents have customers in the US and 65% have customers in Europe. Thus, the majority of firms in our sample will be impacted by some aspects of GDPR. Firms that operate in the US will likely be impacted by aspects of CCPA which are similar to GDPR.

Figure 3. Geographic Distribution of Startup Firms Customers



Findings

Data Protections. We asked questions on data protection in both rounds of the surveys but focused our analysis on the second round of the survey. To access additional training data, around 50% of startups retain secondary reuse rights to their customer's data. Most firms that have secondary reuse rights to customer data report adhering to a data retention policy. Customer's data could in many cases be personally identifiable and must be adequately safeguarded. Firms use a variety of technical means to protect and control data access, including de-identification, encryption, passwords, access logs, and application program interfaces (see Figures 4 and 5, below).

Survey respondents were asked to select all types of data protections used. Across all types of data protection, startups with customers in the EU report using data protection at about the same rate as startup firms without customers in the EU. Furthermore, using OLS regression controlling for HQ location and firm age, we find no significant relationship between having customers in Europe and the use of particular types of data protection (Table 1). Also, the comparison is reported

in Figure 4, below. There are some differences in data protection across firm size, reported in Figure 5. However, using OLS regression controlling for HQ location, we do not find a significant relationship between startup size and data protection (Table 2).

Table 1. Data Protection by Customer Location

DV:	Legal Contract	Deidentify	Encryption	Password	Logged Access	App
Customers Europe	-0.06	0.05	0.00	0.10	0.05	-0.02
	(0.08)	(0.10)	(0.09)	(0.06)	(0.07)	(0.02)
HQ Location Control	Yes	Yes	Yes	Yes	Yes	Yes
Firm Age Control	Yes	Yes	Yes	Yes	Yes	Yes
N	137	137	137	137	137	137

** p < 0.01; * p < 0.05.

Table 2. Data Protection by Firm Size (Employees)

DV:	Legal Contract	Deidentify	Encryption	Password	Logged Access	App
Firm Size (<10)	0.06	-0.03	-0.05	0.03	-0.09	0.07
	(0.08)	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)
HQ Location Control	Yes	Yes	Yes	Yes	Yes	Yes
N	149	149	149	149	149	149

** p < 0.01; * p < 0.05.

Figure 4. Data Protection by Customer Location

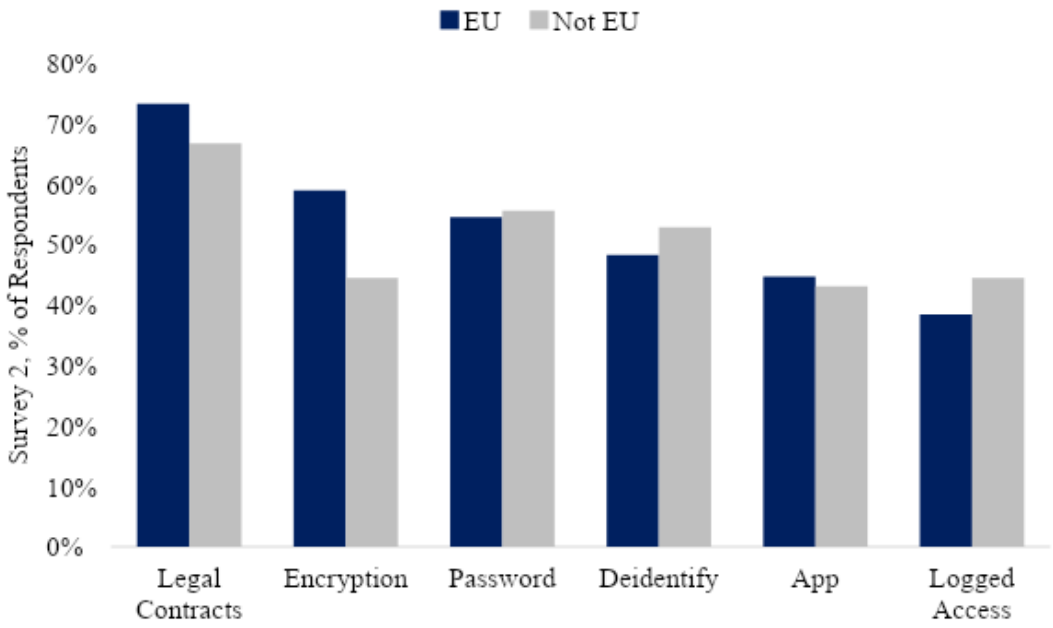
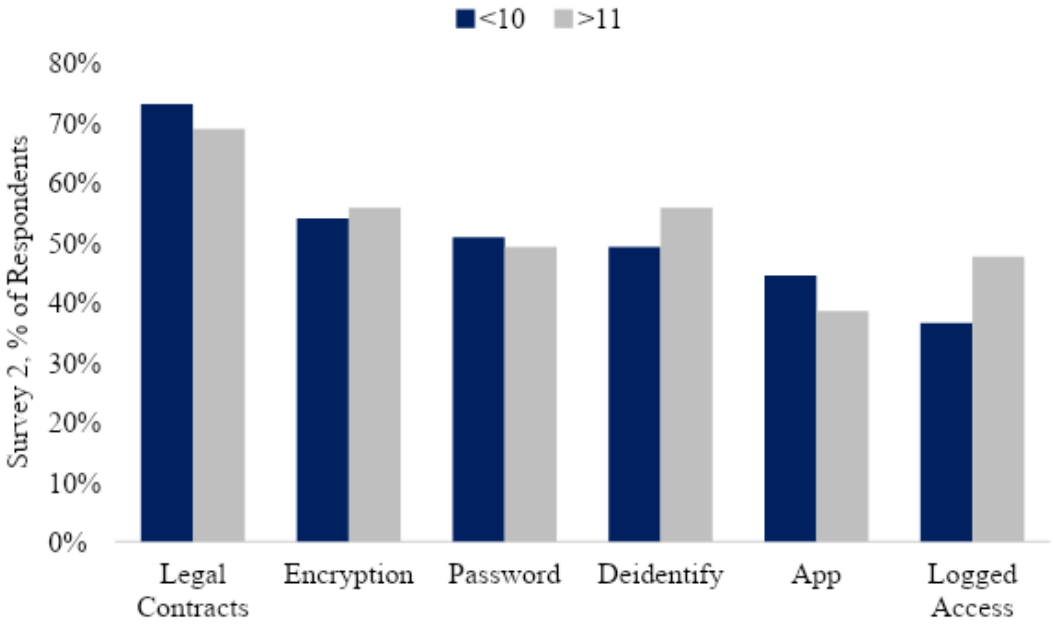
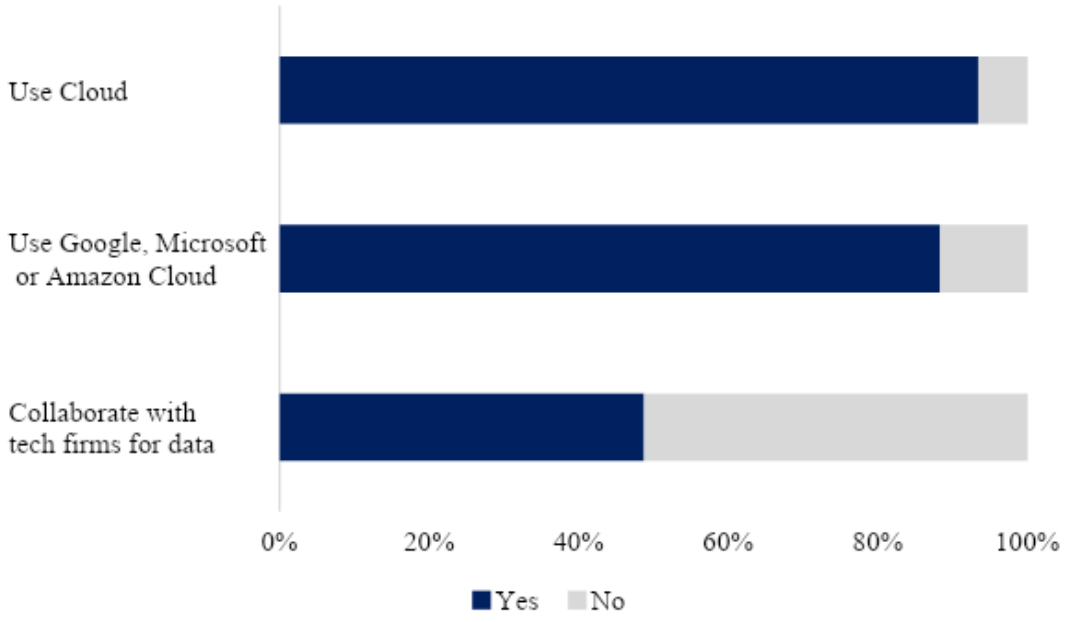


Figure 5. Data Protection by Firm Size (Employees)



Collaboration with Large High-technology Firms. Relationships with other firms and with customers are necessary to accessing the data needed to develop, train and refine AI products. Large high-technology firms have the data (usually captured from platforms with a large existing user base) and the engineering expertise needed to assist startups in developing AI products. Almost 90% of firms use Microsoft, Google or Amazon cloud services. Nearly 50% of respondent firms actively collaborate with large high-technology firms in order to access data. Firms with headquarters based internationally are more likely to collaborate for data than firms based in the US. Possibly the competitive or institutional landscape of the US is less conducive to these types of collaborations.

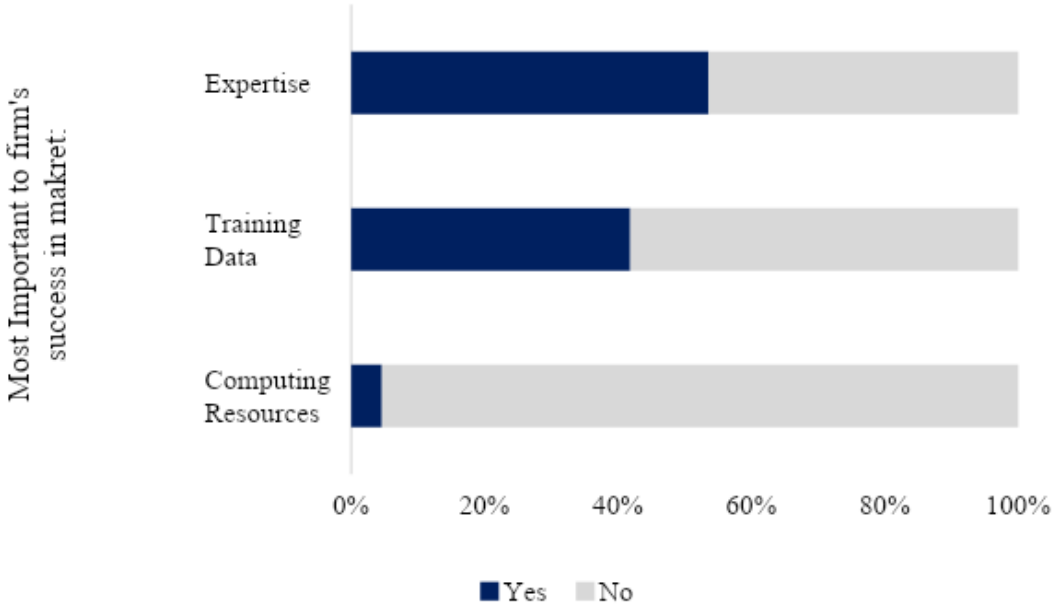
Figure 6. Collaboration with High Technology Firms



Importance of Data. In response to feedback from the first round of the survey, we asked respondents about how important data is to their firm. The survey question specifically asks, "what is most important to your firm's success in your market: training data, data science expertise or computing resources." More than half (54%) of firms responded that expertise is most important, which is unsurprising given the value of highly skilled data scientists and engineers who have the

capabilities needed to build AI products. Very few customers (4%) agree that computing resources are most important to their firm's success. Possibly this is due to the use of variable-cost cloud-based capabilities which positively impact entrepreneurial ventures (Jin & McElheran 2019). Lastly, about 42% of firms agree that training data is most important to their firm's success.

Figure 7. Importance to Firm's Success



We asked respondents if they desire to refresh their model more often. If customers, hypothetically, have unlimited access to data, 60% respond that they would choose to refresh their model more often. More generally, across both surveys, firms that use neural networks and ensemble learning are more likely to refresh their models more often. Lastly, 68% of firms report that data ownership is a major advantage in their market. These findings point to a continued need for training data, particularly when using more advanced algorithm technologies.

Figure 8. Refresh More Often, Neural Network and Ensemble Learning

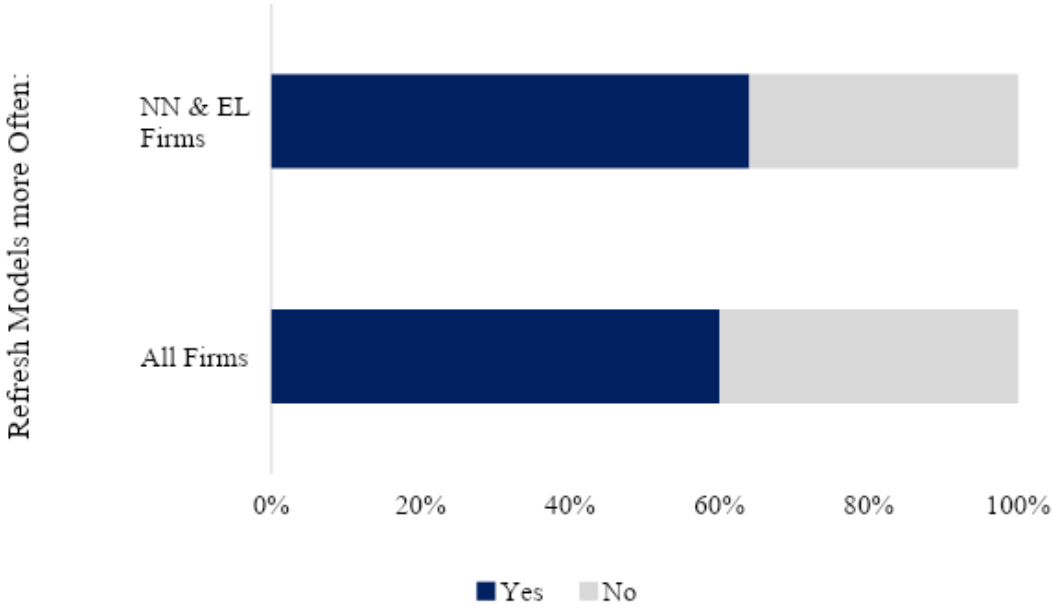
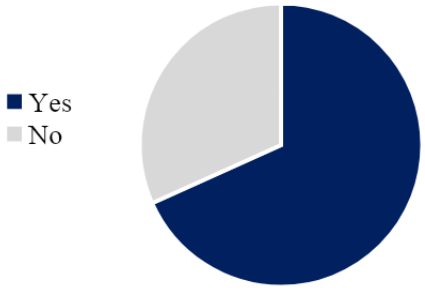


Figure 9. Advantage from Data



Neural network and ensemble learning algorithms have been frequently associated with developing technologies that can complete a task as well as a human. Other less sophisticated algorithms have so far been unable to attain human equivalency. We find that firms using neural networks or ensemble learning algorithms are significantly more likely to respond that data is more important than labor or expertise for success in their market. To provide additional rigor to this result, we use OLS regression to confirm that the use of neural network or ensemble learning

algorithms are significantly related to if the firm reports that "training data is most important". Also, we model if the similar use of these more sophisticated algorithms is significantly related to if firms, given unlimited data, would like to "refresh their models more often". Controlling for age, size and customer location, we find that the use of neural networks or ensemble learning is significantly related to these measures (see Table 3).

Table 3: Importance of Data

DV:	Data Adv	Training Data Imp	Top Cloud Provider	Refresh More	Top Tech Co Collaborate
Use of NN or EL (Dummy)	0.02	0.39 **	0.00	0.29 *	0.01
	(0.15)	(0.12)	(0.09)	(0.14)	(0.15)
Firm Age Control	Yes	Yes	Yes	Yes	Yes
Firm Size Control	Yes	Yes	Yes	Yes	Yes
Cust. Location Control	Yes	Yes	Yes	Yes	Yes
N	84	84	85	83	85

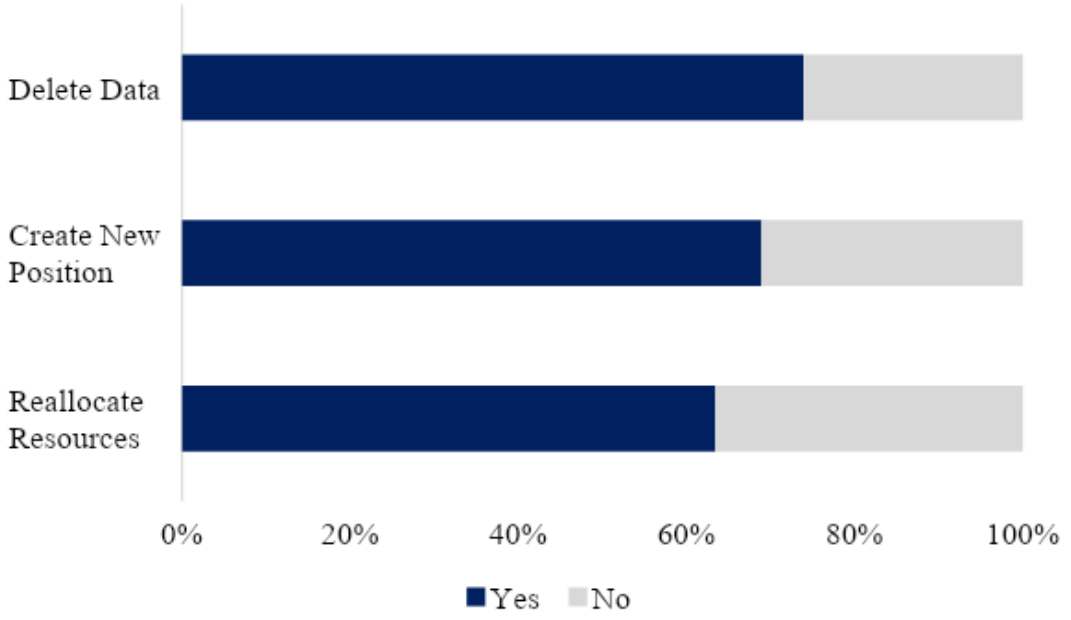
** $p < 0.01$; * $p < 0.05$.

GDPR. There is a large difference in the impact of GDPR on firms with and without customers located in Europe. This difference is exacerbated if the firm has its headquarters based in the US. Internationally based firms are more likely to respond that they have been impacted by GDPR. Furthermore, smaller firms with customers in Europe report that they have been more impacted by GDPR than larger firms (similarly with customers in Europe). Size is significant across all these measures, which lends to concerns that this increased regulation, despite the revenue exclusion, asymmetrically impacts smaller firms.

GDPR impacts how AI startups run and structure their business. In the survey, 69% of respondents' customers answered that they have created a new position to handle GDPR-related issues within their firm. Many firms (63%) report that they have had to reallocate resources due to the impact of GDPR. Regulation such as this limits the types of data that can be stored. Almost three quarters of responding firms have deleted data due to GDPR. Given the high value that firms

place on use and access to the training data, deleting data could seriously impact the ability for AI startups to innovate and dampen AI advancement.

Figure 10. Impact of GDPR



Using OLS regression, we find that firms with customers in Europe and, more specifically, larger firms with customers in Europe are significantly more likely to create new positions to handle GDPR. Larger firms and firms with customers in Europe are significantly more likely to reallocate resources. Larger firms, and more specifically larger firms within Europe, are significantly more likely to have data retention policies which help to manage customer and personally identifiable data (see Table 4).

Table 4: Impact of GDPR

DV:	Data Policy	GDPR New Position	GDPR Reallocate Resources	GDPR Delete Data
Customers in Europe	0.23	0.35 **	0.34 **	0.06
	(0.13)	(0.1)	(0.12)	(0.11)
Interaction Dummy: Large xCustomers EU	-0.47 *	0.22	0.25	-0.06
	(0.20)	(0.33)	(0.21)	(0.35)
Firm Large (>50)	0.53 **	0.21	-0.19 *	0.14
	(0.11)	(0.29)	(0.09)	(0.30)
Firm Age Control	Yes	Yes	Yes	Yes
HQ Country Control	Yes	Yes	Yes	Yes
N	81	84	84	83

** $p < 0.01$; * $p < 0.05$.

Conclusion

Data is important to the success of AI startups. They need access to training data to run sophisticated algorithms, that in some cases are at the cutting edge of technology, such as neural networks and ensemble learning. The majority of these firms indicate that their current training data are deficient in some way and that they would benefit from refreshing their models more often. Though GDPR drives changes that are important to safeguarding personally identifiable information, there are also costs associated with this increased regulation. These startups are reallocating their limited resources and creating new positions to deal with the implications of this regulation. Given that more than 65% of firms included in the survey have fewer than 50 employees, hiring and resource shuffling could be detrimental to longer term success. Also, firms could be required to delete or not collect certain data that could better train their algorithms, enabling the creation of more economically impactful products. There are tradeoffs between increased consumer protection and innovation that must also be considered.

Our research has implications for AI startups and policymakers. Decision makers at AI startups may want to consider the benefits of reaching a wide European customer base against the

costs of complying with GDPR regulation. Policy makers may want to weigh the potential benefits to consumers from enhanced privacy protections against the costs imposed on AI startups, which may ultimately result in fewer startups competing against established firms. As the world continues to globalize, more small firms are expanding internationally in order to compete for customers. If these small firms find compliance untenable, they may choose to focus growth efforts on other geographies, impacting innovation and access to economically valuable products in the EU. Further research is needed to understand how entrepreneurial innovation and data regulations can coexist, enabling firms to access data needed for their success.

References

Athey, S., & Luca, M. (2019). Economists (and Economics) in Tech Companies. Journal of Economic Perspectives, 33(1), 209-30.

Batikas, M, Bechtold, S, Kretschmer, T and Peukert, C. 2020. 'European Privacy Law and Global Markets for Data'. London, Centre for Economic Policy Research.

Bessen, J. E., Impink, S. M., Reichensperger, L., & Seamans, R. (2018). The Business of AI Startups. Boston Univ. School of Law, Law and Economics Research Paper, (18-28).

Comerford, R., & Sokol, D. D. (2016). Antitrust and regulating big data. George Mason Law Rev, 23, 1129-1161.

Cowgill, B., & Tucker, C. E. (2019). Economics, Fairness and Algorithmic Bias. Journal of Economic Perspectives, forthcoming.

Donnelly, R., Ruiz, F. R., Blei, D., & Athey, S. (2019). Counterfactual Inference for Consumer Choice Across Many Product Categories. Stanford University Working Paper.

Furman, J., & Seamans, R. (2019). AI and the Economy. Innovation Policy and the Economy, 19(1), 161-191.

Goldberg, Samuel and Johnson, Garrett and Shriver, Scott, Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes (July 17, 2019). Boston University Working Paper.

Hartmann, P., & Henkel, J. (2018). Really the New Oil? A Resource-based Perspective on Data-driven Innovation. Academy of Management Global Proceedings, (2018), 142.

Jia, J., Jin, G. Z., & Wagman, L. (2018). The short-run effects of GDPR on technology venture investment (No. w25248). National Bureau of Economic Research.

Jin, W., & McElheran, K. (2019). Economies Before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing. Rotman School of Management Working Paper, (3112901).

Johnson, Garrett and Shriver, Scott, Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR (January 21, 2020). Boston University Working Paper.

Lambrecht, A., & Tucker, C. E. (2015). Can Big Data protect a firm from competition?. SSRN Working Paper.Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., ... & Bauer, Z. (2018). The AI Index 2018 annual report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.

Nanda, R. (2016). Financing high-potential entrepreneurship. IZA World of Labor.05530.

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., ... & Bauer, Z. (2018). The AI Index 2018 annual report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.

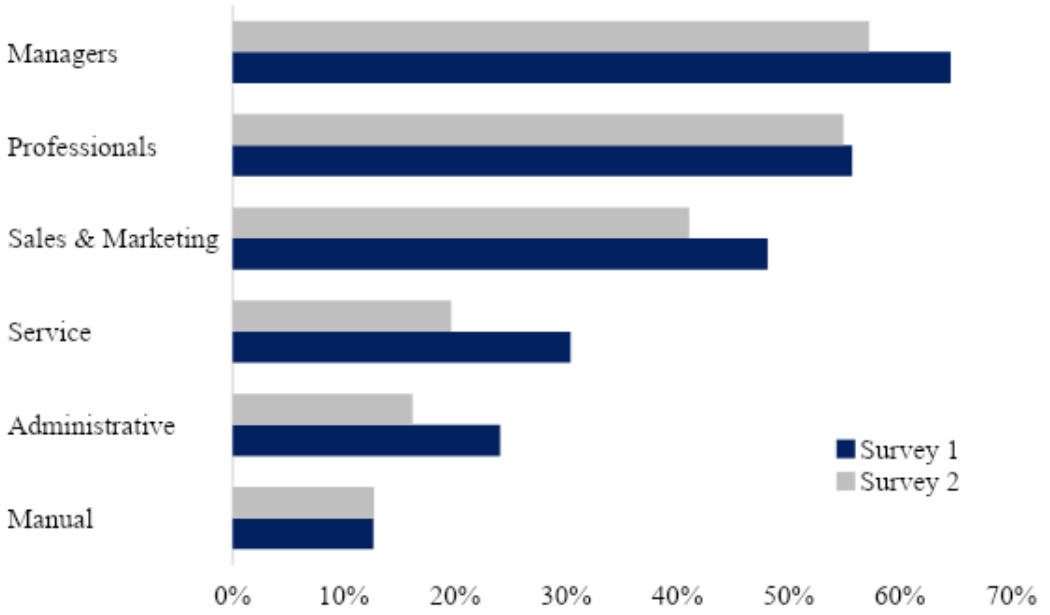
Thomke, S. H. (2003). Experimentation matters: unlocking the potential of new technologies for innovation. Harvard Business Press.

Varian, H. R. (2014). Beyond big data. Business Economics, 49(1), 27-31.

Appendix: Comparison across Surveys

Occupations. There are no significant differences in the percent of reported product users that are managers or professionals. We continue to see that the use of AI requires a certain level of base skills and is more likely to be used by higher skill workers. Individuals working as manual labors remain least likely to use AI products in their occupation (<15%).

Figure A. User Occupations



Labor Impact. Across both rounds of the survey, the three most frequently reported benefits provided by AI products are the capabilities to make predictions or decisions, to manage and understand data, and to create new and improved products and services. These AI-enabled activities enhance human capabilities and augments the abilities of the user. Other measures such as reduction of labor costs and the automation of routine tasks provides insight into the labor replacing aspects of AI.

Firms that are based internationally are more likely to respond that their product leads to gaining new capabilities, automating tasks and reducing labor in survey two than in survey one.

This trend does not hold for firms with headquarters in the US, who show a reduction across all measures. All subsets show an increase in products leading to the elimination of managers, yet the absolute impact is still relatively small at 11%. In both surveys more than 50% of respondents strongly agree that their products automate routine tasks and reduce labor costs; however, it remains that augmentation of skills and capabilities is a large component of the AI products created. The ability for AI to complete some aspect of human labor makes it valuable to customers. We continue to interpret this result as AI augmenting human labor.

Benefits to customers, above in Figure 5, map to the broader conversation on AI-products impacting labor but do not make a strong statement as to the exact nature of that impact. For more details, we asked respondents if their AI products replace or create jobs for certain occupations. Results point to similar levels of jobs creation and destruction in both surveys. These measures continue to differ in similar ways by location, industry and occupation. AI is more likely to create jobs for managers and professions. These higher skilled workers are more able to be trained and use this technology, especially if the technology is complex, sometimes requiring expert skills.

Survey two reports less substantial labor destruction for sales & marketing professional and administrative workers; however, the impact increases for general service workers. On average more than 15% of respondents strongly agree that their AI products create roles for professionals, managers, and in sales and marketing occupations. Clearly, AI is not only about destroying jobs. In some cases, jobs will be eliminated, especially in the three occupational groups that use AI relatively less. However, some respondents reply that jobs will be created in some occupations that replace jobs lost in other occupations within the same firms.

Figure B. Benefits to Customers, Respondents in Survey 1 & Survey 2

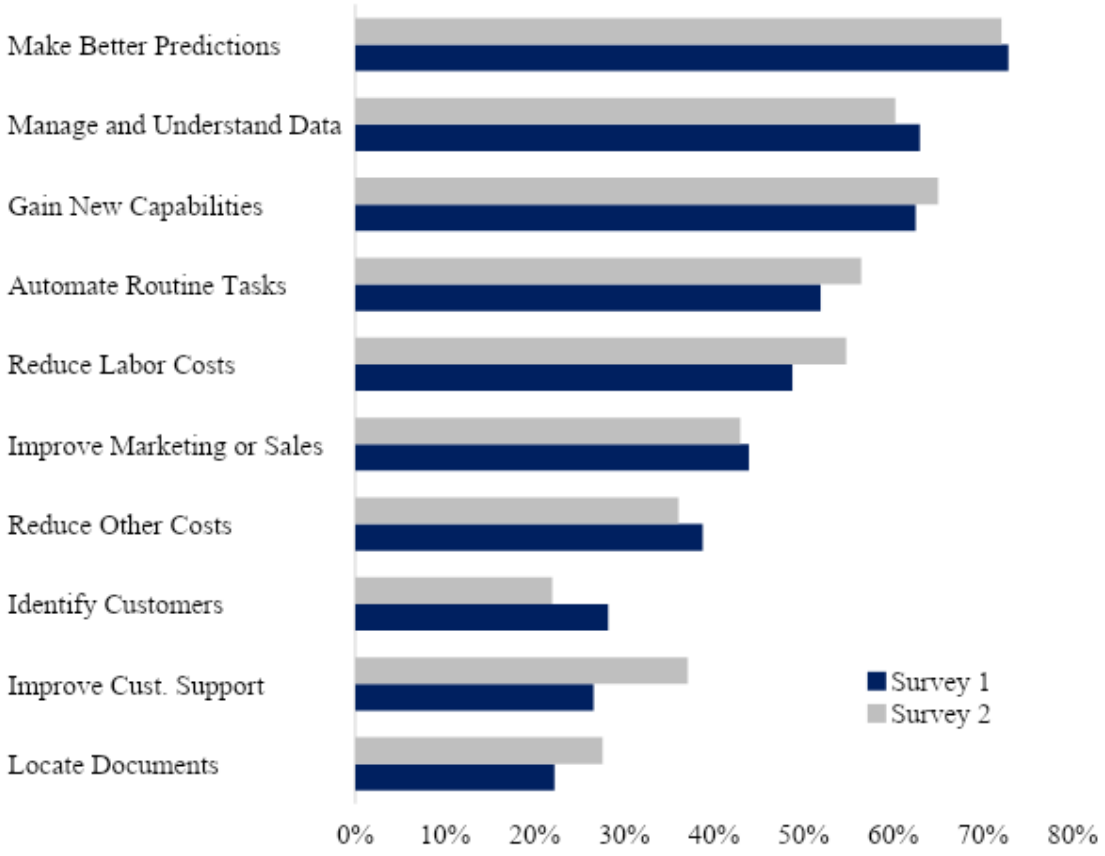


Figure C. Job Creation

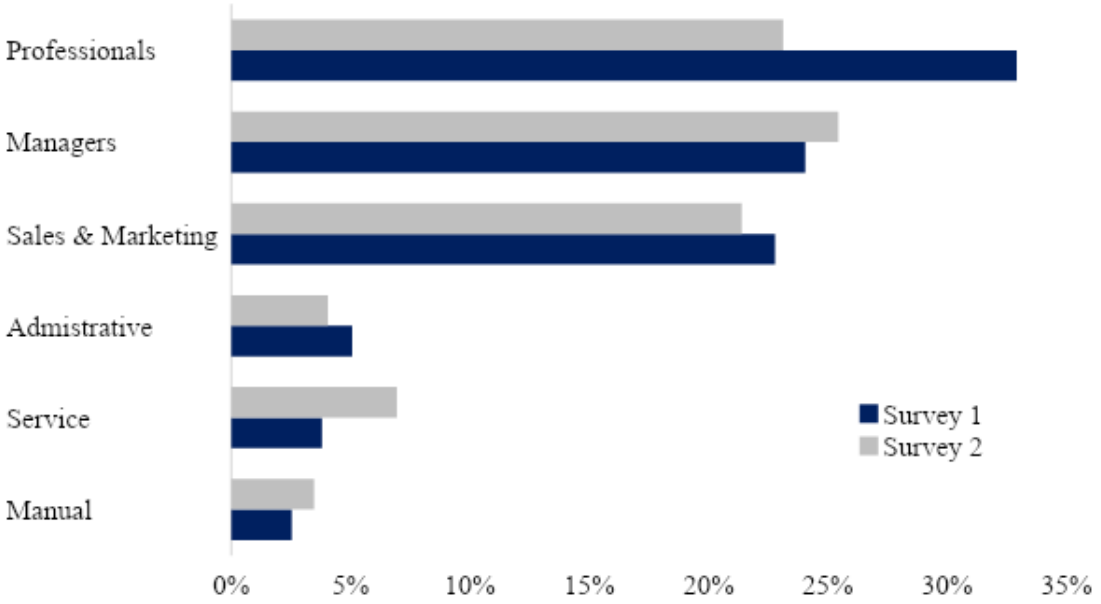


Figure D. Job Destruction

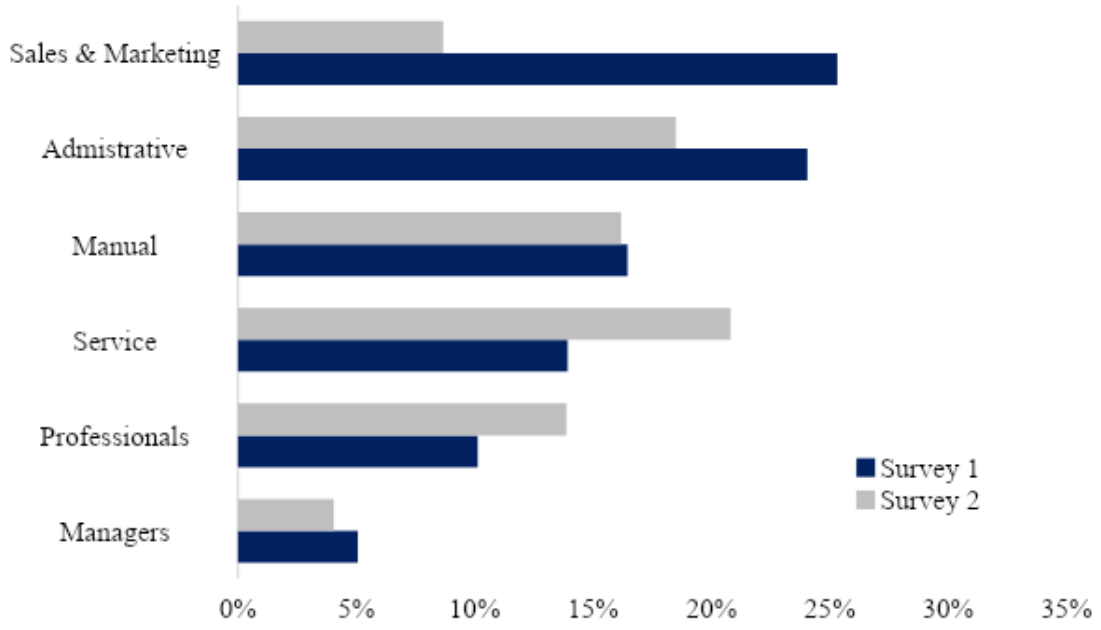


Table A: Measures

Measure	Source	Question	Construction	% of Firms /Mean
Only Proprietary Data	Survey	16	Product is developed with only proprietary firm and customer data	14%
Proprietary Customer Data	Survey	16	Product is developed with only proprietary customer data	8%
Proprietary Cust. Data	Survey	16	Product is developed with proprietary firm or customer data	74%
Internal Tech	Survey	11	Uses internally developed techologies (any) in product development	56%
More Complex Tech	Survey	11	Uses more complex techologies (natural language translation, sentiment/emotion analysis, virtual agents/chatbots) in product development	43%
Use of Neural Networks or Ensemble Learning	Survey	12	Use neural networks or ensemble learning algorithms in product development	85%
Reduces Labor Costs	Survey	15	Strongly agrees (5, on 1-5 scale) that their product reduces labor costs for customers	52%
Automates Routine Tasks	Survey	15	Strongly agrees (5, on 1-5 scale) that their product automates tasks for customers	55%
Eliminates Managers	Survey	23	product tend to emliminate manager positions	10%
Used by Managers	Survey	8	Managers use your product	61%
Data leads to Market Adv	Survey 2	18.2	Data provides a major advantage in my firm's market	68%
Training Data Most Important	Survey 2	5.8	Training data is the most important (compared with data science expertise and computing resouces) to success in my firm's market	42%
Wants to Refresh More	Survey 2	14.1	You would refresh your model more ofent if you have unlimited data	59%
Product Shipping	Survey	5	Product is currently in the market	70%
Change in Funding Round	Crunchbase		Shift in funding round (seed, early, late) from t=1 to t=2	9%
Funding Rate Increase	Crunchbase		Dummy varibale for if the funding rate increased after 2018 compared with 2017 and prior	7%
Log of Total Funding	Crunchbase		Log (Total Funding Rate in USD)	14.47

Table B: T -Tests

T-Test, P-value < .05?	Crunchbase vs. Respondent Sample	Survey Round 1 vs Survey Round 2	Responded to both rounds vs only one round (t=1)	Responded to both rounds vs only one round (t=2)
Demographics				
Firm Age (From CB)	Not Comparable	Yes	No	Yes
Customers Large (>250)	No	No	No	No
Customer US	No	No	No	No
Custoemr EU	No	No	No	No
HQ US	No	No	No	No
HQ EU	Yes	No	No	No
Large (>50 Employees)	No	No	No	No
Very Small (<10 Employees)	No	No	No	No
Measures				
Firm Data	No	No	No	No
Proprietary Customer Data	No	No	No	Yes
Prop Firm Data or Cust. Data	No	No	No	No
Public Data	No	No	No	No
Internal Tech (NA added in Round 2)	Not Comparable	Not Comparable	Yes	No
More Complex Tech (NA added in Round 2)	Not Comparable	Not Comparable	Yes	No
Internal More Complex Tech (NA added in Round 2)	Not Comparable	Not Comparable	Yes	No
Use of Neural Networks or Ensemble Learning	No	No	No	No
Reduces Labor Costs	No	No	No	No
Automates Routine Tasks	No	No	No	No
Eliminates Managers	No	No	No	No
Used by Managers	No	No	No	No
Data leads to Market Adv (S2 Only)	No	Not Comparable	Not Comparable	No
Training Data Most Important (S2 only)	No	Not Comparable	Not Comparable	No
Wants to Refresh More (S2 Only)	No	Not Comparable	Not Comparable	No
Change in Funding Round (From CB)	Not Comparable	No	No	No
Funding Rate Increase (From CB)	Not Comparable	Yes	No	No
Log of Total Funding (From CB)	Not Comparable	No	No	No