Table of Contents:

What’s your data culture?

Executing a Company Data Strategy Starts with Building a Data-driven Culture.

Creating a culture that embraces data-driven decision-making requires understanding your individual contributors, building your technical infrastructure, and recognizing the impact of data and organizational bias.

Adrian Lievano. December, 2019.

Executive Summary:

Data is everywhere -- in every industry, country, organization, and user of digital applications, data and the way we store, process, analyze, and share its insights with others can be used for great benefit. Executing a data strategy starts by building a stronger data culture. To build a stronger data culture, understand the differences between individual contributors in a data team, build your technical infrastructure to support collaboration across company departments, and always consider the impact of data and organizational bias when pursuing data projects.

Motivation:

Data is everywhere -- in every industry, country, organization, and user of digital applications, data and the way we store, process, analyze, and share its insights with others can be used for great benefit. Leaders across companies and prospective job seekers interested in information are on fertile grounds: the cost of data storage is exponentially decreasing, the amount and velocity of data is increasing, and the algorithms that open the valve on this spigot of value are more accessible with modern programming frameworks [1]. To capture this value, however, companies face considerable challenges such as hiring and retaining talent, using an organization’s structured and unstructured datasets, and much more [2]. The best way to tackle these problems is to have a data strategy: a strategy for organizing, governing, analyzing, and deploying an organization’s information assets [3].

A data strategy has multiple parts: addressing compliance and security, creating new products and services, or developing organizational analytics capabilities to name a few. A crucial element in creating an effective data strategy, however, involves setting your data culture; it influences the competitive advantage when your bring talent, tools, and decision making together [4]. There are multiple surveys of c-suite executives from various Fortune 500 companies, each adding a unique understanding of the makings of a strong data culture. In this report, however, we add to the conversation by providing insight into building technical teams and how your data infrastructure defines your data culture. As a result, I aim to empower executives with insights to advance their business goals.

Background:

Companies that prioritize data-driven decision-making create competitive advantages in their industries: Lyft, Didi Chuxing, Facebook, Google, Apple, among others, are examples of the most valuable businesses that leverage data and analytics to create new products, improve on existing products or services, or attract the best talent. Despite the economic opportunities present in data across industries, progress towards creating data-driven cultures is stagnant: of 64 surveyed c-level technology executives at some of the largest corporations, 72% report that they do not have a data culture, 69% are not data-driven, 53% are not treating data as an asset, and 52% do not believe they are competing on their data assets and analytic capabilities [10]. In attempts to address these issues, a staggering 93% of respondents identify people and process issues as the main obstacle. In another study from the McKinsey Global Institute report, 42%, 45%, and 36% of executives across industries listed ensuring senior management involvement, designing an appropriate organizational structure to support analytics activities, and designing an effective data architecture and technology infrastructure, respectively, as their top 3 most significant challenges.

Methodology:

The annual industry-wide Kaggle Data Science & Machine Learning survey contains 16,000, 23,859, and 19,717 responses in 2017, 2018, and 2019, respectively. A Kaggle data science and Jupyter notebook is used to analyze the survey fields. This report focuses on self-reported software engineers, data engineers, data scientists, analysts, and product manager respondents. I selected this audience because these are the more common contributors in a data team. All code, visualizations, and supporting resources can be found in the reference section.

Section I - Understanding Technical Contributors:

A shortage of the analytical and managerial talent to leverage data is an obstacle companies can begin to face in the short term. The United States alone faces a shortage of nearly 200,000 people with deep analytical skills and 1.5 million managers and analysts to analyze data and make decisions on their findings. These skilled workers require multiple years of mathematical training and programming experience, as well as the ability to ask targeted business questions and use data to support their conclusions.

“There is a continuing shortage of analytics talent.”

For companies to benefit from their data, a great first step starts with understanding the nuances between the people in a data team so that they can begin to build a strong analytics organization. In the modern data science team, some roles include machine learning engineers, data engineers, data scientists, product managers, analysts, and software engineers -- some teams look different depending on the size of the company and the datasets that they work with [6, 7, 8]. By understanding the differences in skills, education, and responsibilities, companies can source talent from multiple channels and avoid common mistakes that may cause these technical contributors to leave -- some examples might include poor job specificity, working in isolation, or unrealistic expectations [9]. The roles overlap, and vary in programming, mathematical, and communication skills, but each use data to accomplish business goals.

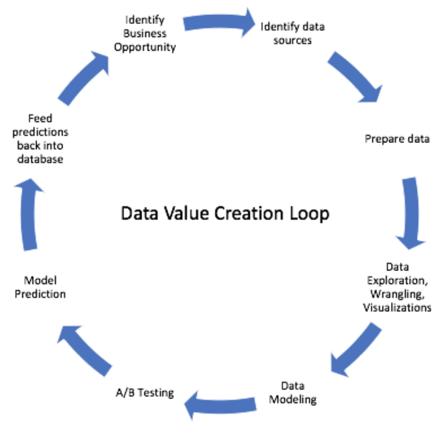


Exhibit 1: The data-value creation loop.

A business goal using data can be achieved, for example, by following a data value creation loop: a sequence of well-defined steps that involve generating revenue from data. Awareness of the data-value creation loop and its potential impact on a company's revenue is well understood: 92% of the c-level respondents reported an accelerating rate of investment into "artificial intelligence" and 55% of them report investments in Big Data and AI exceeding $50MM and growing [10]. There is a misunderstanding, however, because increasing investment dollars into AI without the foundation in the data value creation loop in place can have serious consequences -- it's putting the cart before the horse.

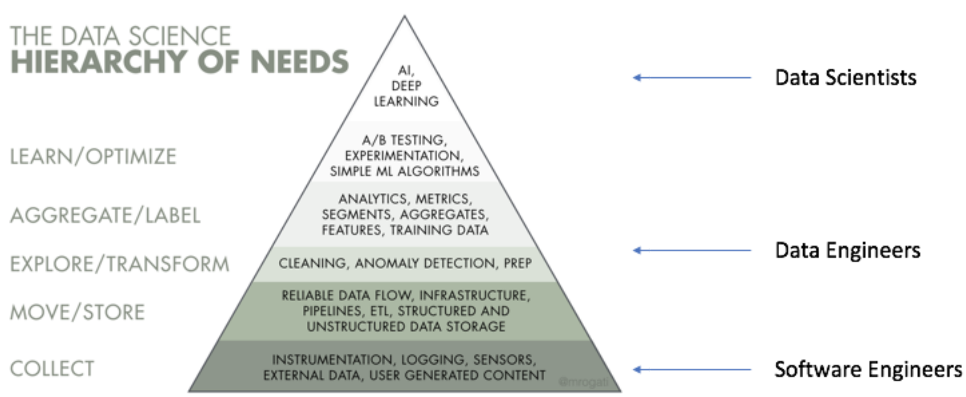


Exhibit 2: A data hierarchy of needs. Credit: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

At the bottom of the pyramid, software engineers interface with sensors located in devices (mobile devices, industrial machinery, etc.) to collect data; they build web and mobile user-applications. These user-interfaces collect data on user behavior. Data engineers interface with unstructured data in a variety of formats and program algorithms to extract, transform, and load the data into structured, accessible formats. It is at this point where more value can be captured — for example, analysts or data scientists can gather sample statistics, clean the data, or build visualizations to inform strategic initiatives. In the explore and transform part of the pyramid, dashboards can be presented to cross-functional teams and provide actionable insight based on company data. At the learn and optimize level, data scientists and machine learning engineers either design experiments or deploy prediction models . At the top -- the level where most of the corporate investment dollars go towards -- artificial and deep learning technologies are applied.

Data Scientists:

Mix the roles of a statistician, business advisor, and software engineer and out comes a data scientist: a unique position at companies that navigate unstructured and structured datasets to produce novel insights that drive business goals forward. 25% of data scientists (n = 4085) report spending the majority of their time analyzing and understanding their datasets. Close in second, 22.3% report spending their time building prototypes to explore their datasets. In fact, developing machine learning models, refining algorithms or preparing training sets are less of a priority based on this aggregate data. The problem is that a large majority of data science job postings are misleading, even so far that applicants are recommended to apply regardless of the requirements and tools they require candidates show projects that show problems in a similar domain [11].

Data scientists cover the widest spectrum of undergraduate majors, but a majority of them studied mathematics, physics, an engineering field apart from computer science, or some degree of finance or economics. They also have the highest concentration of Master's and Doctorate degrees by nearly twice as compared to data engineers and software engineers. In addition, their perception of MOOCs relative to traditional brick and mortar education tends to be worse when compared to other contributors in a data science team: nearly 10% of data science respondents rated MOOCs as much worse than traditional education pathways. Despite having a larger percentage of respondents with negative ratings, nearly 35% rated MOOCs as slightly better or much better, and an overwhelming percentage (>70%) of data scientists are enrolled or completed a MOOC data science course. This supports the claim that as a role, data scientists thrive in environments where they can "build things" in addition to giving advice, and where they are given "room to experiment and explore possibilities" to tackle business problems. A key lesson from these points: data scientists come from varied backgrounds, but most of them are focused on using statistics, and high-level software engineering toolkits to rapidly pull together data to draw insights on important questions.

Data Engineers:

Data engineers build the infrastructure to move, edit, and deliver information to people in different departments. These individual contributors are more software engineers than statisticians, and are more likely to spend less time communicating to business stakeholders as opposed to engineering managers. 47% (n = 624) of data engineers responded that building or running the data infrastructure that their business uses for storing, analyzing, and operationalizing data is the most important part of their role. When it comes to spending time understanding machine learning models, a downstream application of the data they prepare, nearly 24% of data engineers view ML models as "black boxes and that there are other contributors in a team" that can explain model outputs (i.e., data scientists). This supports the idea that in a data team, data engineers are less focused on machine learning and more on enabling data scientists to do their advanced analytics techniques. When companies begin to assemble these teams, or support them, it is important to consider that a weaker culture will rely on an individual contributor to understand the full stack of data or require skills that are not typically expected in the industry for this given role.

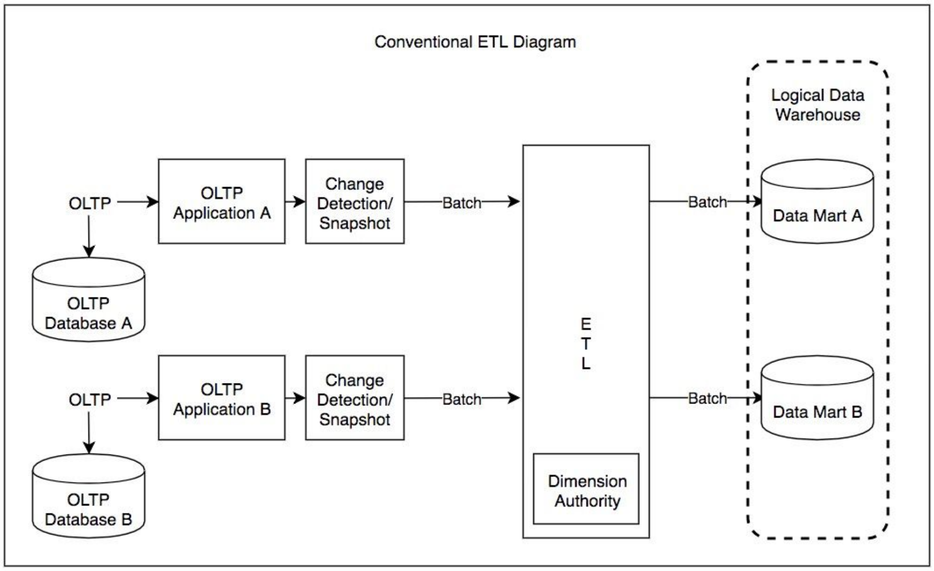


Exhibit 3: An industry example of a data pipeline.

Data engineers complete a narrower list of undergraduate degrees: 52% report having a computer science degree, 18.6% report some sort of engineering degree that is not computer-focused (chemical, bioengineering, mechanical, etc), and 8.7% report a math degree. In addition, nearly 88% of data engineers hold a Bachelor's or Master's degree -- only 7% hold doctoral degrees as compared to 19% of data scientists. Data roles that require a more research-oriented approach (i.e., designing experiments, setting appropriate sample sizes, or preparing study briefs to communicate to different stakeholders) is typically less aligned with the role of a data engineer. It is, however, promising to see that data engineers are spending as much time as data scientists continuing their education using MOOCs: 42.8% report spending the most time on Coursera. The difference however, lies in their perception of the quality of these courses relative to traditional brick & mortar education. 42.5% data engineers report MOOCs are "slightly better" as compared to 18% from data scientists. 0% report it as "much worse" as compared to 9% of data scientists. It's known that traditional education tends to focus more on mathematics and less exposure to a particular programming language [12]. This perception of MOOCs being slightly better for data engineers could provide a window into understanding the level of statistical rigor the roles require.

Software Engineers:

Software engineers are arguably the most understood role in a data team. From serving on an ad hoc basis for multiple data projects to shipping machine learning models, or building the user-interfaces that enable the collection of the data, software engineers are the hackers that bind the rest of the team together to produce usable software. They're different from data engineers and data scientists in a number of ways:

for example, 33.1% of software engineers (n = 2705) report spending less than 1 year writing code to analyze data. Another 29% report spending only 1-2 years. When compared to data scientists, for example, 31% report spending 3-5 years, 20% report 5 to 10 years, and 10.1% report less than 1 year. The difference is stark because the roles and expectations are different. Of a list of options to describe the important part of their role at work that include (i) analyze and understand their data, (ii) build or run a machine learning model, (iii) build the data infrastructure, (iv) build prototypes to explore models, (v) or do research that advances the state of the art in machine learning, 32.6% of software engineers report “none of these activities” are an important part of their role, compared to 1.4% of data scientists and 3.4% of data engineers. It is important to note, however, that the number of software engineers that report “none” of these activities being important in Kaggle’s 2019 Data Science Survey dropped to 10%, supporting the idea that although they focus less on building models, they still have to analyze data and understand the impact of what they ship.

It’s unsurprising to see that 25% of software engineers report basic statistical software like Microsoft Excel as a primary tool at work to analyze data. The expectations are different, which is an important thing to understand in a data team to build a strong data culture. Software engineers represent the largest population with Bachelor’s degrees (39.4%) and the lowest amount of master’s degrees at (43.7%) when compared to the other individual contributors in a data team in 2019. Of all the undergraduate degrees, 67% report studying computer science and 13% report non-computer focused engineering as their undergraduate degree; they are the least diverse when it comes to undergraduate degree. When compared to other individual contributors, 67% of data scientists and 57.9% of data engineers report python as their most used language.

The wide repertoire of programming languages with less formal years of schooling provides insight into the role of software engineers in data teams. Rather than split the focus of software engineers to do basic modeling techniques, it is better to have them learn the foundations so that they can communicate with data scientists and data engineers. Afterward, focus them around building the prototypes and tools needed to support data collection and shipping machine learning models.

Analysts:

Analysts, sometimes further specified as data or business analysts, are differentiated from the other roles in their need to be data storytellers: they are contributors on a team that are less attuned to advanced statistics or machine learning, but can quickly write a sequence of SQL queries to parse through data in a company with hypothesis or question in mind. According to the chief decision scientist at Google, the best analysts “surf vast datasets” to identify “useful gems” and have a “mastery of visual presentation” and storytelling [13]. These types of analysis are then validated by data scientists or statisticians, where insights are adjusted for risk and then presented to decision makers [14]. Similar to data scientists and data engineers but unlike software engineers, 68% analysts (n ~ 2000) report that most of time is spent analyzing and understanding data to influence product or business decisions. 50% of them report SQL being a language they use on a regular basis. In addition, besides python being the overwhelming recommended language to learn first, analysts suggest SQL more than any other technical contributor. SQL enable fast queries and allows analysts to gather data to support questions they seek to answer.

Python, however, is used more frequently by data engineers and data scientists because it allows advanced modeling techniques that SQL does not.

It is also interesting to note that 23% of data analysts (n = 1598) and 21% business analysts (n = 778) report “not knowing” which specialized hardware they use on a regular basis as compared to 8.6% for data scientists, 13% for data engineers, and 9.6% for software engineers. Though there are percentages of analysts that report knowing that they use CPUs, analysts represent the lowest percentage of the technical contributors that understand this difference and the lowest percentage of contributors that use GPUs. It’s important to understand this nuance: analysts use CPUs and GPUs, but relative to their more machine learning driven peers, they use them less or know which platform they work on less often. Analysts are not as technical: though they need to quickly parse through data and use similar tools, their main goal is to validate an initial hypothesis that aligns with a business objective.

53% of business analysts and 51% of data analysts have master’s degrees; 32.6% and 33.8% of these two groups, respectively, have Bachelor’s degrees. Though a smaller percentage of them also have PhDs (3.50%, 6.5%), they also represent the group with the lowest percentage of doctorate degrees. They are also the group with the highest percentages of having no formal education (1.94%) when compared to data scientists (0.53%). Analysts also represent the largest group with undergraduate majors in a business discipline (27% for business analysts and 14% for data analysts). This does not mean that they do not study computer science or a mathematics degree; it is an observation that they study other less engineering-focused majors more often than data scientists, data engineers, or software engineers.

This data supports the claim that overemphasizing machine learning and statistics will cause companies to lose analysts. A strong data culture depends on having clear business goals and a clear delineation of responsibilities for each person on the team. By focusing analysts to sift through data to identify reasons to fund a new or current project instead of building machine learning models, companies can focus their data scientists on the deep-technical work and their analysts on connecting the dots between the data and the business objective.

Product/Project Managers:

Product manager roles in a traditional sense are known to vary across products, company size, industry, and more. Without having direct authority over an engineering team, product managers must deliver new product features on a regular cadence while balancing the needs of diverse teams like engineering, design, management, marketing, and sales. In the context of a data team, however, product managers are also expected to have domain expertise in data science, data modeling, infrastructure, statistics, and machine learning [15]. Their role differs from an analyst because they need to deliver products or leverage data assets to accomplish business objectives. Product managers create the plan, a timeline, and is usually the decision-maker along the way. They are also expected to be able to write their own SQL and interpret results presented by their data scientists or analysts. Similar to analysts and data scientists, 57% of product/project managers (n = 723) said that “analyzing and understanding” their data is an important part of their role. 37% of product/project managers say that “building prototypes to explore applying machine learning to new areas is an important” part of their role. 35.7% of product/project managers, similar to analysts, also say that basis statistical software is their primary tool to analyze data. When asked about programming languages or specific machine learning toolkits that contributors used, product/project managers were not far behind: 46.4% report Scikit-learn, a popular library across all roles for machine learning, as the library they used the most. In addition to the expectation of having high emotional intelligence so that they can conduct customer interviews, run design sprints, prioritize features, allocate resources, etc., product managers in a data team must understand the common tools, libraries, and different use cases of machine learning so that they can better target business opportunities, define success metrics, and develop pricing and revenue models, for example.

An overwhelming percentage of product/project managers have master's degrees (55.4%) and 9.6% hold doctorate degrees. As undergraduates, 32.7% of product/project managers majored in computer science, 25.9% in an engineering discipline that is non-computer focused, and 9.6% in business (finance, economics, etc.). When compared to data engineers, data scientists, or analysts, product/project managers represent the lowest percentage of respondents that major in mathematics or statistics (6.8%). The data suggests that although product manager roles are typically more customer or management-facing, they are still expected to be highly technical contributors and to collaborate with data scientists, engineers, or analysts. Another observation shows that product/project managers rated MOOCs as 'much better' than traditional brick & mortar education more than any other group (35%) -- 0% of product/project managers rated them as much worse. the reasons could be many, but it does suggest that MOOCs are great options for product managers to learn these key data science and analytical skills.

Data scientists, data engineers, software engineers, analysts, and product managers are some of the contributors in an effective analytics team. A strong data culture understands their differences because these roles require different support policies and habits in place. It's silly to hold software engineers to the same performance metrics as analytics, or to expect that the needs of a data scientist are equivalent to those of a product manager. Strong data culture is built to accommodate for people of varying skills, responsibilities, years of experience, and more. With this knowledge, managers can incentive the right behavior, or design new programs to support their skills development. The challenge today includes differentiating these roles enough so that we can take such action. With the information above, for example, some programs might also include broadening the hiring requirements for certain roles while narrowing it for others, or incentivizing online learning through stipends so that these teams can stay up-to-date and relevant in the fast-changing pace of this industry. By doing so, companies establish a healthy data culture that can execute on a data strategy.

Section II - Data Infrastructure & Bias:

Purpose:

The different levels of successfully integrating technology infrastructure are described to vary between a state of exploration (i.e., collecting data) and using machine-learning models to automate the bulk of decision-making [16]. In between these two ends, organizations develop and standardize their data assets, build data dashboards to enable dynamic decision making, or expand the decision-making process to include data from a global network of users. As described in a series of blog posts, the "minimum viable product" of data infrastructure consists of pipelines that extract, transform, and load data to multiple stakeholders, a data warehouse that is designed to be queried, and any additional business intelligence tools to help derive insights to inform decision-making [17]. Data culture is tied to infrastructure because it either empowers employees across teams or it continues to only support specialized workers like data scientists, data engineers, or analysts.

A data-driven culture starts with collective data empowerment: a culture where (i) less-technical employees can work with and benefit from data, (ii) anyone can contribute knowledge to an analysis, and (iii) more people can discover, adapt, or reuse prior work. Without the infrastructure or tools to support creating this collective empowerment, common issues include: (i) siloed, unused data, (ii) delays, (iii) lost knowledge from prior analysis, (iv) needless repetition, (v) poor data literacy, and more. The inefficiencies compound, and tools are emerging to support collaboration within data teams. To execute on any potential data strategies, companies need to make an active decision and integrate with tools that enhance the ability to collaborate with internal and distributed data teams.

Bias & Ethics: Awareness of it is Key to a Healthy Data Culture

Corporations seeking to leverage big data and machine learning to capture the value in their industries and build defensible business-moats need to consider data bias. A strong data culture considers the impact of data and machine learning bias.There are plenty of resources that discuss the different types of bias that exist in data and machine learning models [18]. To summarize, bias creeps into data in a few main ways:

1. Reporting Bias: a result of having data skewed to represent a group you're analyzing
2. Automation Bias: when you favor a machine's prediction over a non-automated system.
3. Selection Bias: when a data set does not represent their real-world distribution.
4. Group Attribution Bias: the tendency to generalize what is true for individuals to a larger group which they belong.
5. Implicit bias: the result of making an assumption based on your personal experiences that do not generalize to other groups.

A company that neglects to train or to incentive its data teams to consider bias in their daily work promotes a weak data culture. Of software engineers, data engineers, data scientists, business and data analysts, and product managers, greater than 60% of survey respondents (n > 10,000) believe the ability to explain model outputs or predictions as "very important". In addition, in all respondents, more than 54% believe perceive fairness and bias are "very important" topics in machine learning. Less than 5% of respondents believe it to "not important at all". There is, however, an inconsistency: less than 15% of respondents say that a data project's success is tied to evaluating unfair bias. Instead, "revenue or business goals" or "model accuracy" are reported more than 50% of the time to evaluate the success of a project. There is a misalignment between individual contributors that build with data and metrics set by managers that neglect understanding the limitations and edge cases of their models.

As managers build their data cultures, it is important to remember that data is not conclusive; bias can creep in and companies that build employee confidence will invest the time and resources to understand the limitations of their data sets and their trained models so that edge cases that could cause project failures are understood. This thoughtful, considerate approach will build confidence across cross-functional teams and give credibility to an analytics department.

Becoming a Data-Driven Organization:

Most of the insights in this report come from a wide variety of past and recent publications. Most of the data presented comes from nearly 40,000 surveys of respondents from Kaggle's annual data science and machine learning survey in 2018 and 2019. It is, however, important not to weigh these statements as final. This data represents the frequency of these observations. There is no data that supports how a given team composition -- and how you select skills or responsibilities for a given role -- or how the sophistication of a company's data infrastructure contributes to the overall success of a business -- do we track revenue, market durability? This is, however, an exploration of the current state of the labor market and current business practices, so it might serve as the first step companies take to build a stronger data culture to advance their data strategies. Most companies are lacking in this regard, so any action taken to learn about data teams, their contributors, and the different levels of data infrastructure quality can go far in leveraging more data assets and accomplishing greater business goals. Data is meant to support qualitative decision-making because it is never complete. It is clear, however, that after acknowledging these risks and assumptions, every organization can find tremendous value in strengthening their data cultures. If we heed these insights, maybe we can all become data-driven and solve more of our greatest problems.

Acknowledgements:

Thank you to my friends and mentors for reading drafts of this post and for providing feedback. I am grateful for their support. I'd also like to thank Kaggle for preparing such detailed data sets from their annual data science and machine learning survey. It is such hard work to clean, prepare, and present data for a community of over 1 million data enthusiasts to analyze.

References:

[1] Manyika, James, et al. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." *McKinsey Global Institute*, 2011, www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx.

[2] Henke, Nicolaus, et al. "The Age of Analytics: Competing in a Data-Driven World." McKinsey Global Institute, 2016, www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.ashx.

[3] Dallemule, Leandro, et al. "What's Your Data Strategy." Harvard Business Review. May, 2017. https://hbr.org/2017/05/whats-your-data-strategy

[4] Ransbotham, Sam, et al. "The Talent Dividend." MIT Sloan Management Review, 2015, https://sloanreview.mit.edu/projects/analytics-talent-dividend/

[6] Hu Samson. "Building the Analytics Team at Wish." Medium. January, 2018. https://medium.com/wish-engineering/scaling-analytics-at-wish-619eacb97d16

[7] Ng, Andrew. "AI Transformation Playbook. How to Lead Your Company to Success in the AI Era." Landing.ai. December, 2018. https://landing.ai/ai-transformation-playbook/

[8] Thorpe, Ryan. "How to Structure a High Performance Analytics Team." Medium. Towards Data Science. February, 2018. https://towardsdatascience.com/how-to-structure-a-high-performance-analytics-team-f564c92a1aaa

[9] Brooks-Bartlett, Jonny. "Here's why so many data scientists are leaving their jobs." Medium. Towards Data Science. March, 2018. https://towardsdatascience.com/why-so-many-data-scientists-are-leaving-their-jobs-a1f0329d7ea4

[10] Bean, Randy, et al. "Companies Are Failing in Their Efforts to Become Data-Driven." Harvard Business Review. Feb. 2019, https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven

[11] Harris, Jeremie. "The Problem with Data Science Job Postings." Medium. Towards Data Science. March, 2019. https://towardsdatascience.com/the-problem-with-data-science-job-postings-8a3542f38724

[12] Bartham, Ammon. “Bootcamps vs. College.” Triplebyte Blog. May, 2016.
https://triplebyte.com/blog/bootcamps-vs-college

[13] Kozyrkov, Cassie. “What Great Analysts Do -- and Why Every Organization Needs Them.” Harvard Business Review. December, 2018.
https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them

[14] Diaz, Alejandro, et al. “Why Data Culture Matters.” McKinsey Quarterly. September, 2018.
https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/why-data-culture-matters

[15] Causey, Trey. “Rise of the Data Product Manager.” Medium. Towards Data Science. July, 2017. https://medium.com/@treycausey/rise-of-the-data-product-manager-2fb9961b21d1

[16] Palantir. “Leveling up your company: A Lexicon for Digital Transformation Success. Medium. November 22, 2019.
https://medium.com/palantir/levels-9be772098942

[17] Bolard, Christopher. “Data Engineer vs. Data Scientist.” Medium. Towards Data Science. December, 2018. https://towardsdatascience.com/data-engineer-vs-data-scientist-bc8dab5ac124

[18] Google Developers. “Fairness: Types of Bias” Google Developer Blog.
https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

[19] Crowdflower, Inc. “2016 Data Science Report.” Crowdflower, 2016,
https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Supporting Resources:

[20] Davenport, Thomas, et al. “Data Scientist: The Sexiest Job of the 21st Century.” Harvard Business Review. October, 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century?referral=03759&cm_vc=rr_item_page.bottom

[21] Manyika, James, et al. “What Do We Do About the Biases in AI.” Harvard Business Review. October, 2019. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[22] Dowsett, Chris. “It’s Time to Talk About Organizational Bias in Data Use.” Medium. Towards Data Science. April, 2018. https://towardsdatascience.com/lets-talk-about-organizational-bias-in-data-use-92ba83bb2c59

[23] Krishnamurthy, Prabhakar. “Understanding Data Bias.” Medium. Towards Data Science. September, 2019. https://towardsdatascience.com/survey-d4f168791e57

[24] Srivastava, Sanjay. “Competing in a Digital First World.” CIO. July, 2018. https://www.cio.com/article/3294216/recognizing-and-solving-for-ai-bias.html

[25] Rodrigues, Marlon. “What is the Role of an AI Software Engineer.” Medium. Towards Data Science. Aprl, 2018. https://towardsdatascience.com/what-is-the-role-of-an-ai-software-engineer-in-a-data-science-team-eec987203ceb

[26] Austin, Julia. “What it Takes to Become a Great Product Manager.” Harvard Business Review. December, 2017. https://hbr.org/2017/12/what-it-takes-to-become-a-great-product-manager

[27] Berinato, Scott. “Data Science and the Art of Persuasion.” Harvard Business Review. Feb, 2019. https://hbr.org/2019/01/data-science-and-the-art-of-persuasion

[28] Davenport, Tom, et al. “Data Not Leading to Insights? Culture Might be to Blame.” Wall Street Journal. Sep, 2019. https://deloitte.wsj.com/cmo/2019/09/29/data-not-leading-to-insights-culture-may-be-to-blame/

[29] Rogati, Monica. “The AI Hierarchy of Needs.” Hackernoon. June, 2017. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Code Used for Analyzing Data and Creating Visualizations:

[30] Lievano, Adrian. "Adrianlievano/kaggle_data_science_2018_survey." *GitHub*, 2019, github.com/adrianlievano/kaggle_data_science_2018_survey.

I encourage forking off this GitHub repository to continue analyzing the kaggle datasets. I wrote a CONTRIBUTING.MD file that targets some additional areas which should be further dissected (i.e., growth rates over the years for different parameters of interest). I also added notes about points in the code that can be improved. I also add a list of questions that future annual Kaggle surveys should include if we are to better understand the nuances between technical contributor roles and how data infrastructure plays a role in data culture.

Public Data & Support Code:

[31] Crawford, Chris, et al. “2018 Kaggle ML & DS Survey.” *Kaggle*, 3 Nov. 2018, www.kaggle.com/kaggle/kaggle-survey-2018.

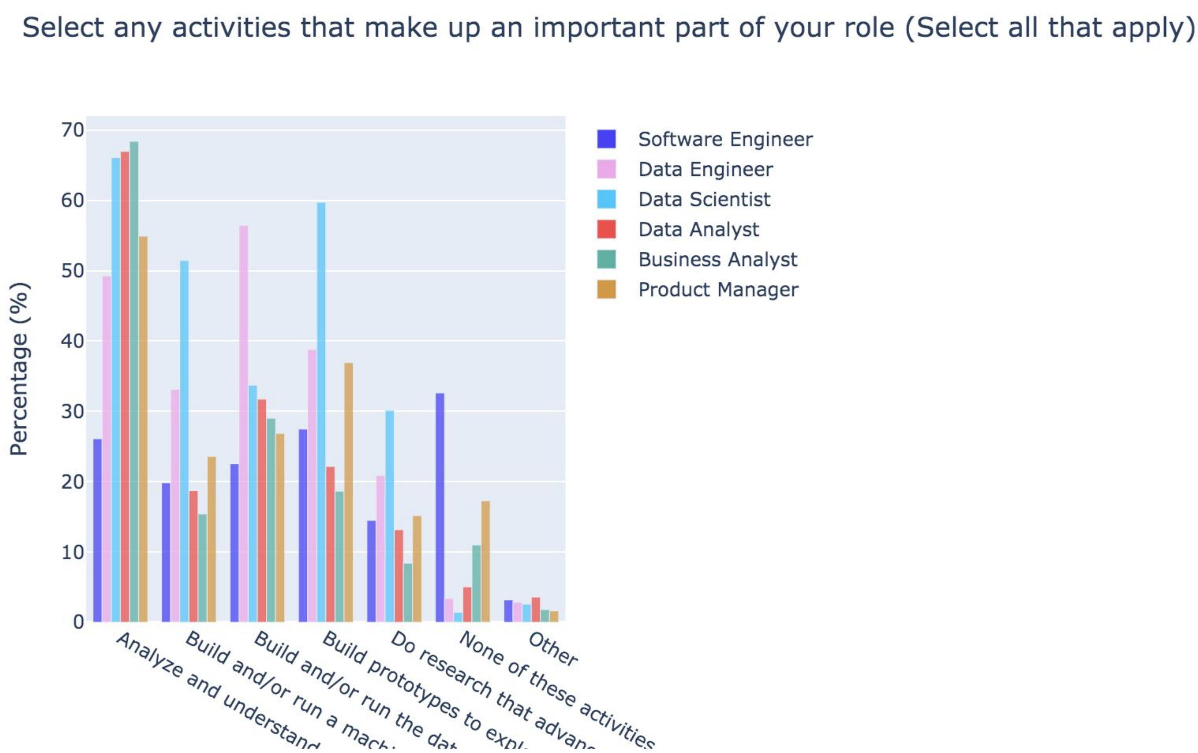
[32] Team, Kaggle. “2019 Kaggle ML & DS Survey.” *2019 Kaggle ML & DS Survey*, 2019, www.kaggle.com/c/kaggle-survey-2019/.

[33] Team, Kaggle. “The State of ML and Data Science 2017.” *Kaggle*, 2017, www.kaggle.com/surveys/2017.

[34] Pandey, Parul. “Geek Girls Rising : Myth or Reality!” *Kaggle*, Kaggle, 18 Nov. 2019, www.kaggle.com/parulpandey/geek-girls-rising-myth-or-reality/data?utm_medium=email&utm_source=intercom&utm_campaign=kaggle-survey-2019.

[35] Amin. “Student Community in Kaggle.” *Kaggle*, Kaggle, 26 Nov. 2019, www.kaggle.com/amiiiney/student-community-in-kaggle/comments.

Appendix:



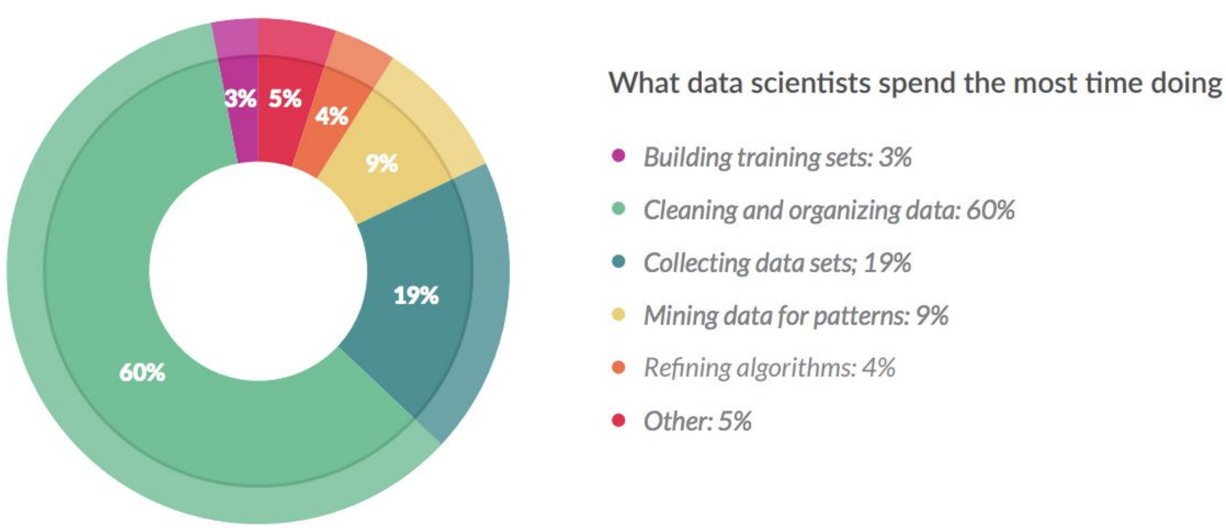


Exhibit 1: Right: Top: percentage breakdown of the typical responsibilities of a data scientist from Kaggle survey. Bottom: Data scientist programming time break down according to a 2016 CloudFlower survey [19].

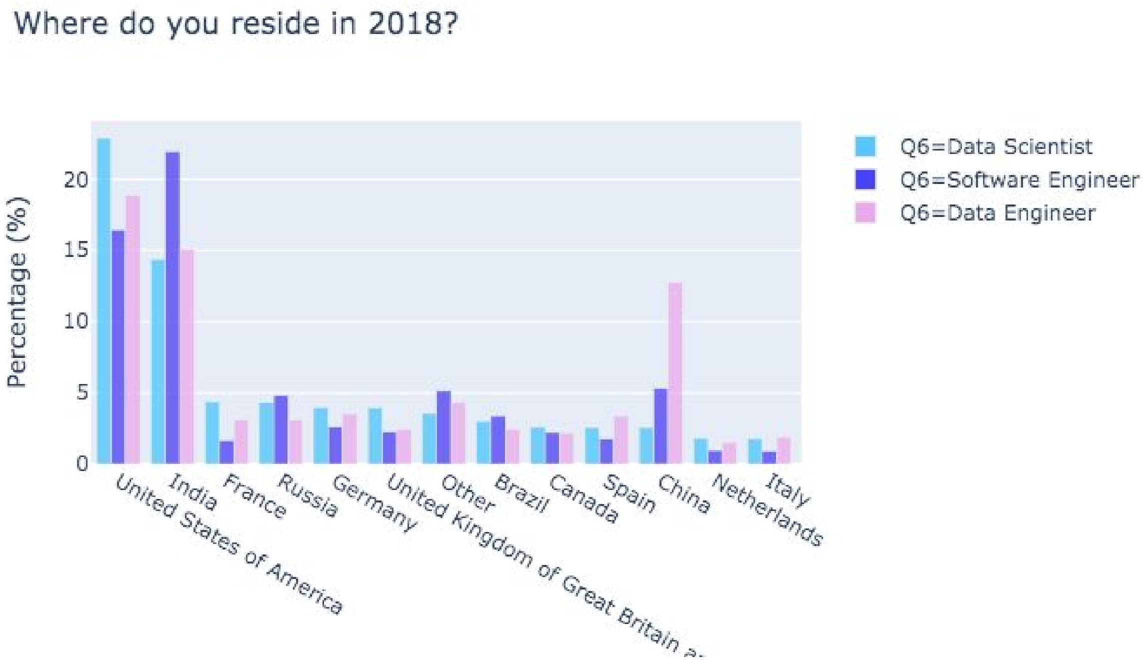


Exhibit 2: Geographic concentration of self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

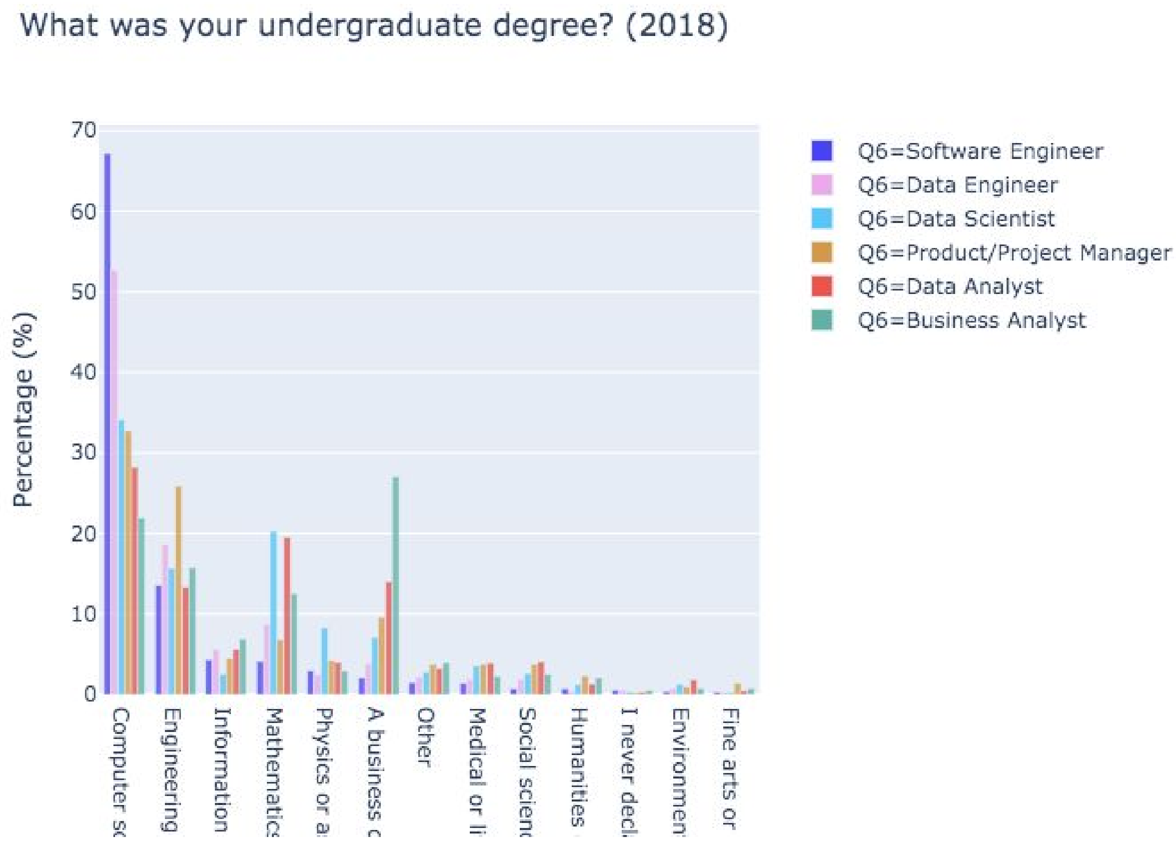


Exhibit 3: Undergraduate degree breakdown of self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

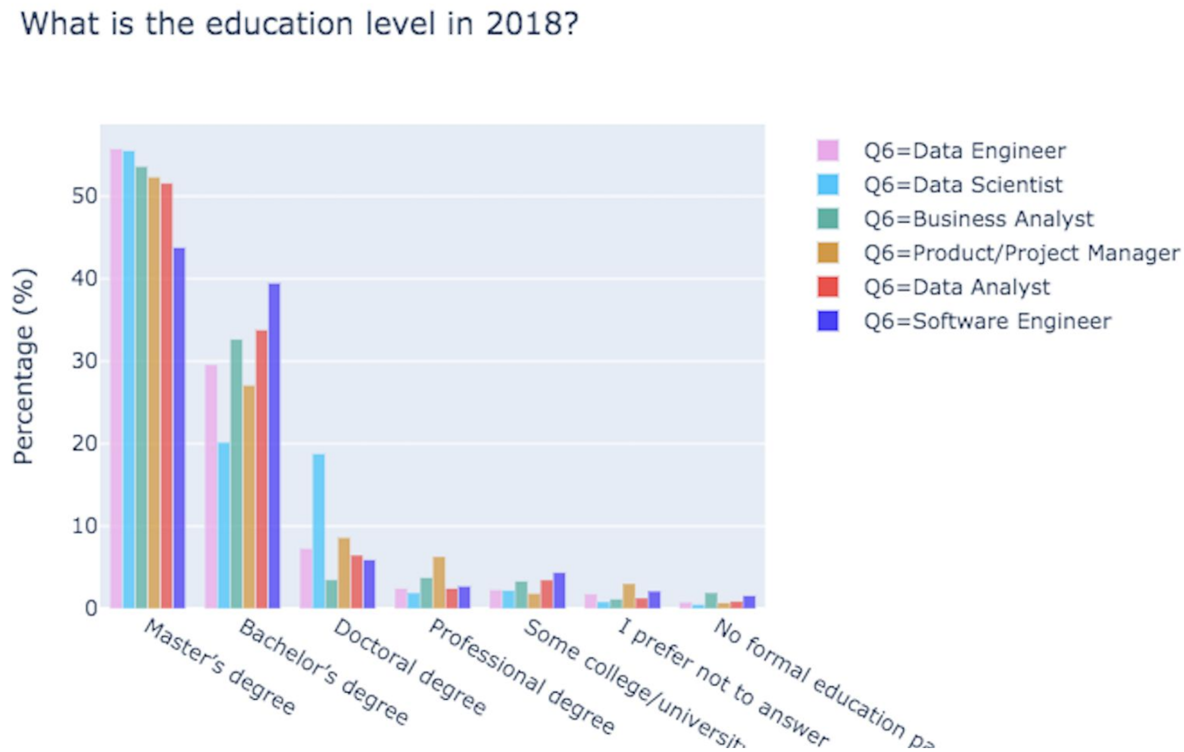


Exhibit 4: Level of education for self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

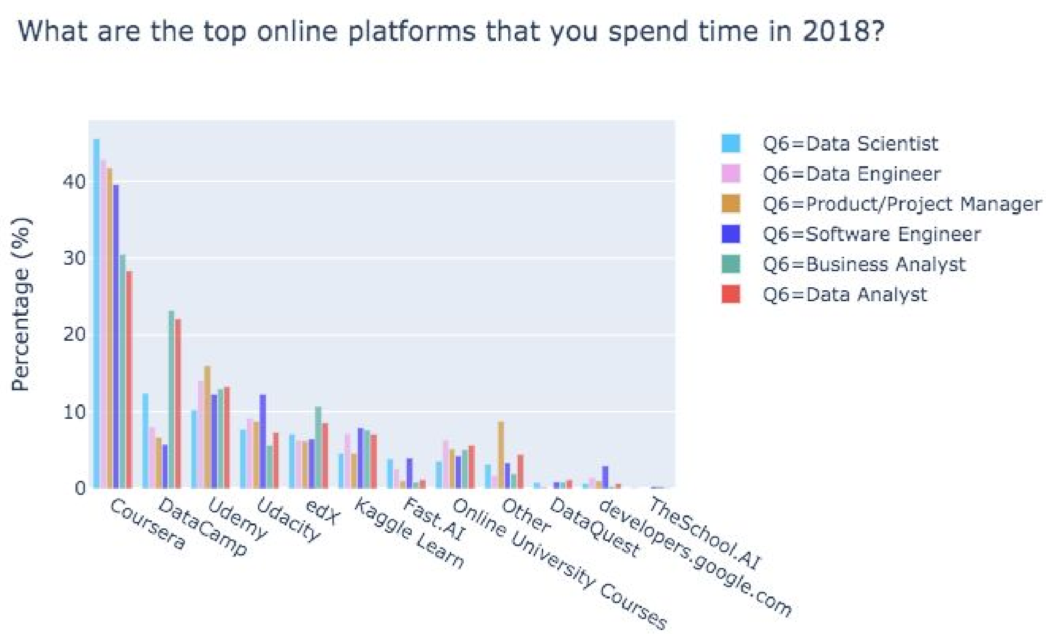


Exhibit 5: Most popular MOOCs based on total time spent self-reported by data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

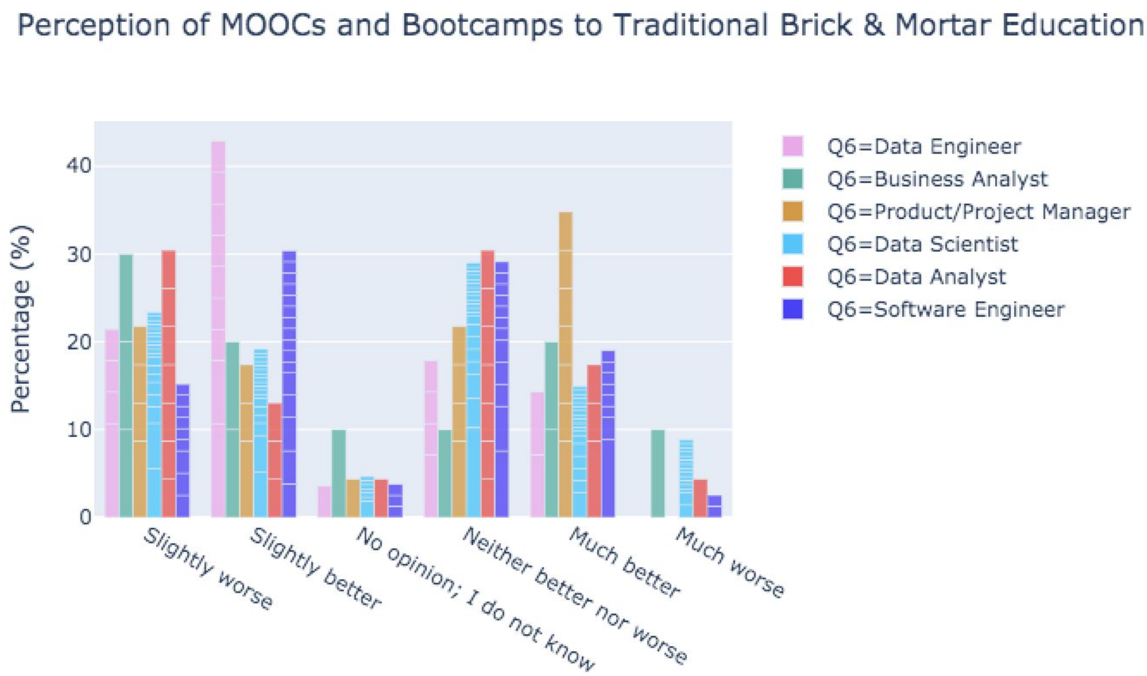


Exhibit 6: The Perception of the Top-5 most popular Massive Open Online Courses (MOOCs) compared to traditional brick & mortar institutions. 'Much Better' indicates that a MOOC is 'Much Better' than a traditional education. This ranking includes responses from data scientists, data engineers, and software engineers separated by annual self-reported salary. Blank squares indicate no data was available.

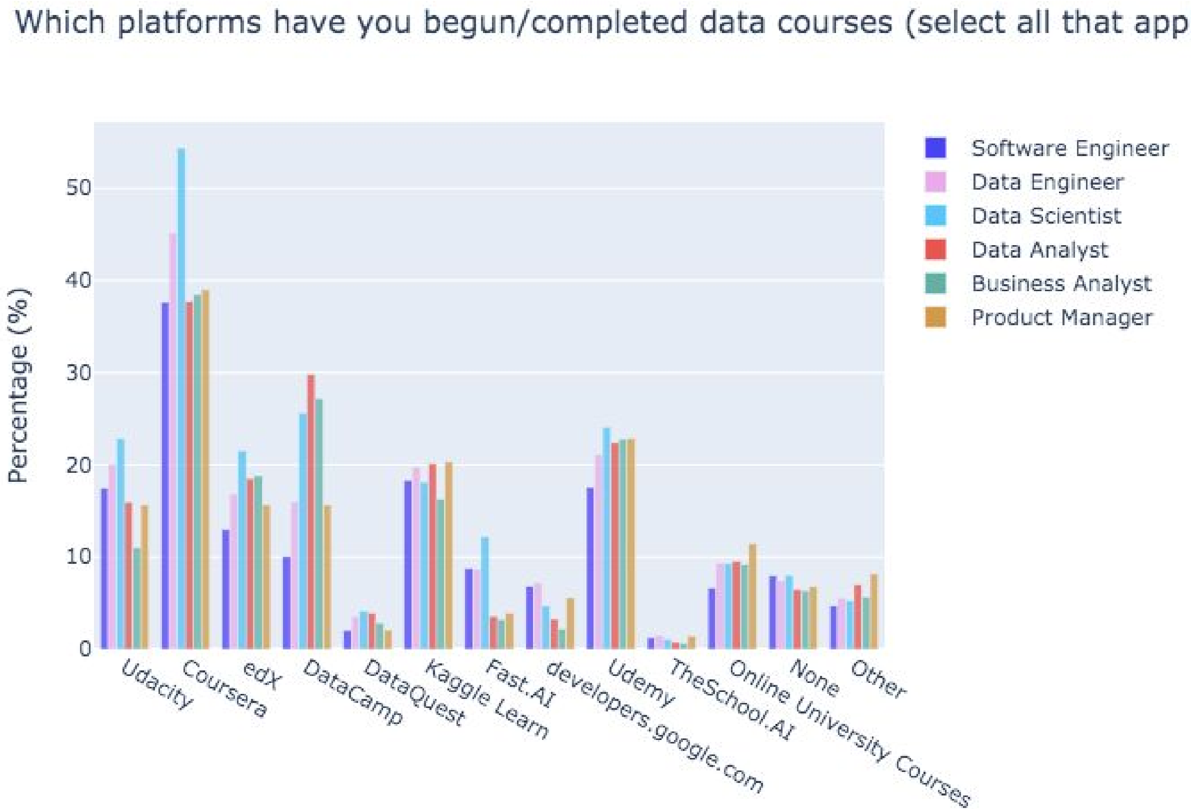


Exhibit 7: Most popular Massive Open Online Courses (MOOCs) that were completed per role from the 2018 annual Kaggle Data Science Survey. Respondents selected all that applied. Respondents can select all that apply.

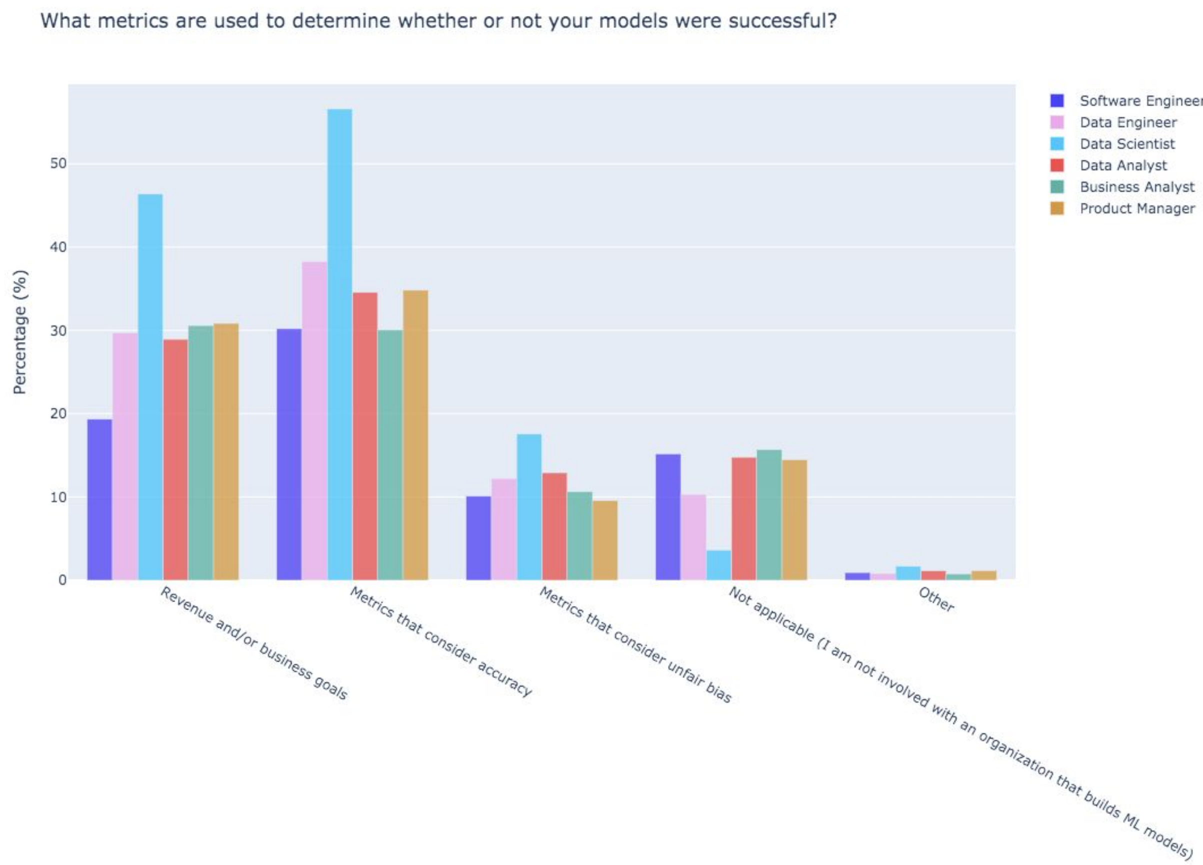


Exhibit 8: A percentage breakdown of respondents for the metrics they use to evaluate the success of machine learning models from the 2019 Kaggle Data Science Survey.

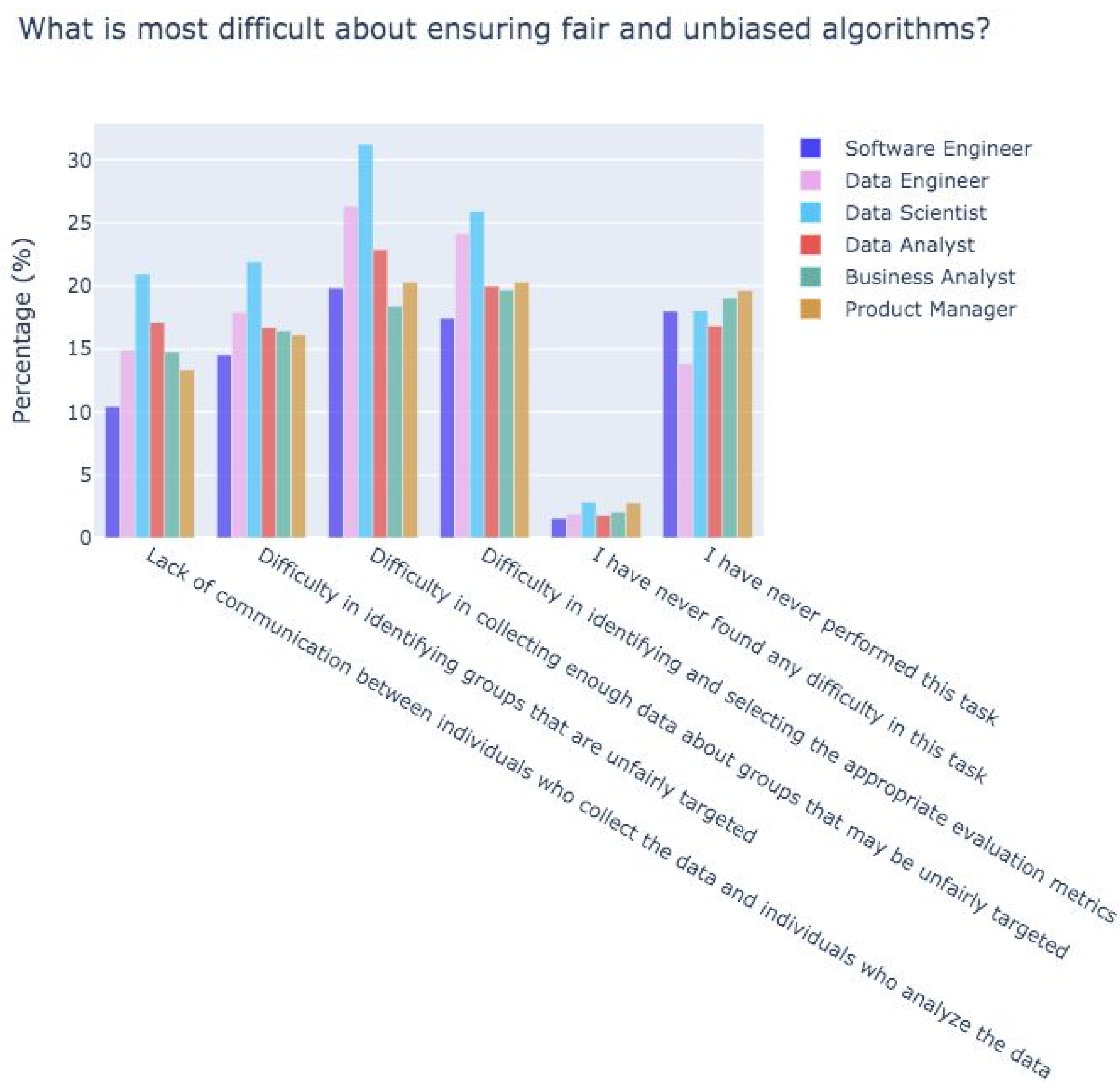


Exhibit 9: A percentage breakdown of respondents for different roles expressing the difficulty with ensuring fair and unbiased algorithms in data science and machine learning from the 2018 Kaggle Data Science Survey.

Exhibit 11: A percentage breakdown of the amount of time different roles spend exploring model insights/predictions for machine learning models from the 2018 Kaggle Data Science Survey.

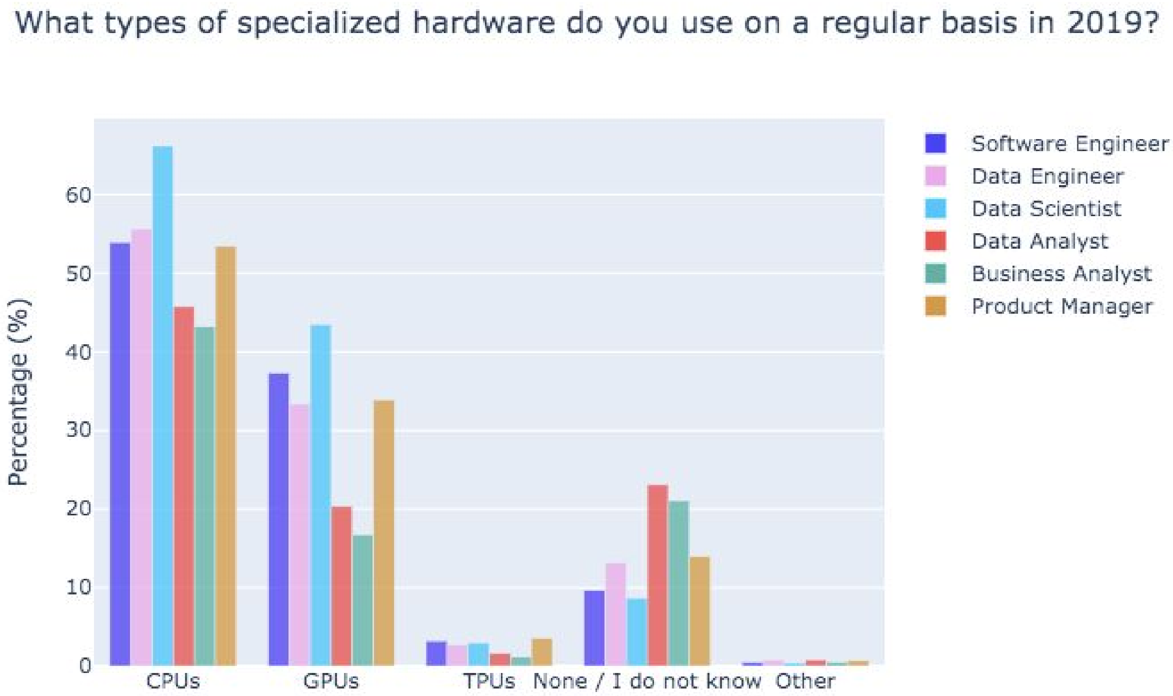


Exhibit 12: The usage of specialized hardware per role self-reported from the 2018 Kaggle Data Science & Machine Learning Survey.

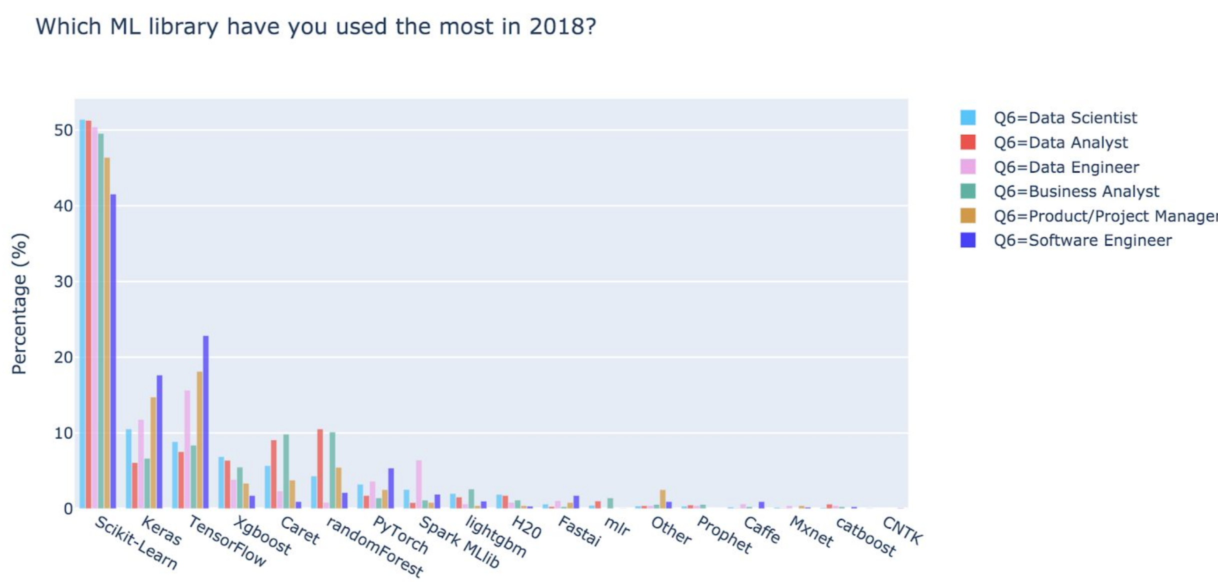


Exhibit 13: The most used machine learning libraries per role from the 2018 Kaggle Data Science & Machine Learning Survey.