Research Note:
Chief Data Scientist Survey

Rajeev Chand
Partner, Research
rajeev@wing.vc

Jake Flomenberg
Partner
jake@wing.vc

Olivia Rodberg
Research Associate
olivia@wing.vc

Wing Venture Capital
480 Lytton Avenue
Palo Alto, CA 94301

Data science is central to the success of all enterprises—from startups to corporations and across all industries. The field, dubbed the "sexiest job of the 21st century" by HBR in 2012, is undergoing dramatic growth and change. Innovation is occurring at a rapid pace, as evidenced by this summer's GPT-3 paper and demos. Further, the role of the data scientist is increasingly complicated with pre- and post-modeling workflows, severe talent shortages, and organization-wide stakeholders.

What are the biggest challenges facing data scientists and machine learning models? What is the importance of GPT-3 to data science, ML, and AI? How is the industry addressing data bias, and to what extent are the approaches working? What is the impact of automation to the need for data scientists? What are trends in data science budgets and projects? And, how will the role of the data scientist evolve in companies?

To address these and other questions, Wing conducted an exclusive survey of the senior-most data scientists at global corporations and venture-backed startups from October 20 to October 28, 2020. The respondents included 88 chiefs, vice presidents, and heads of data science, ML, and AI, each of whom led his/her company's data science organization.

In addition, Wing hosted the Wing Data Science Summit on October 27, 2020. The speakers included luminary AI researchers Peter Norvig (Google) and Sandy Pentland (MIT) and veteran corporate executive Bill Groves (Walmart) in a closed-door, Chatham House Rule setting. The participants included 320 senior-most data scientists, of which 140 were from public companies with $10B+ market caps and 55 were from private companies with $1B+ valuations.

In this research note, we present the results of our survey, summarize anonymized insights from our summit, and share our takeaways on what's next in data science. We conclude with an investor's perspective on commercial opportunities in data science.

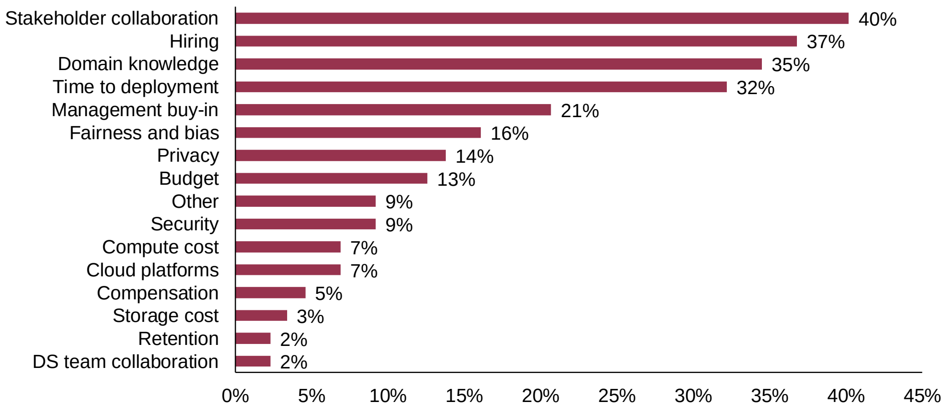
We would like to thank Will Uppington at Truera, Michelle Ufford at Noteable, and Shoaib Shaikh at L3Harris for their contributions to the Wing Chief Data Scientist Survey and Wing Data Science Summit.

The graphs from this research note are available at the following link.

Overall Challenges

Survey respondents indicated that stakeholder collaboration, hiring, and domain knowledge are their biggest challenges as data science leaders.

What are your biggest challenges as the data science organization leader?



We heard similar comments at our summit and in research meetings:

"I am leading a cultural transformation within our company."

- Stakeholder collaboration. A tech DS leader asked, "what are the best practices on the interactions between data science teams and other stakeholders?" Another tech leader said, "getting ML to work in a company is more of a cultural shift than a technology shift." A cable leader added, "I am leading a cultural transformation within our company."

- Hiring. A tech DS leader stated, "the canonical issue especially in Fortune 1000 corporations is: who are people we can attract, who are people we can retrain internally, and how do we set them up for success."

- Domain knowledge. A CPG DS leader said, "when I first joined, I spent time learning the 101s of the business. I took classes for domain specialists, as I needed to understand how the entire business works. That is not really documented anywhere."

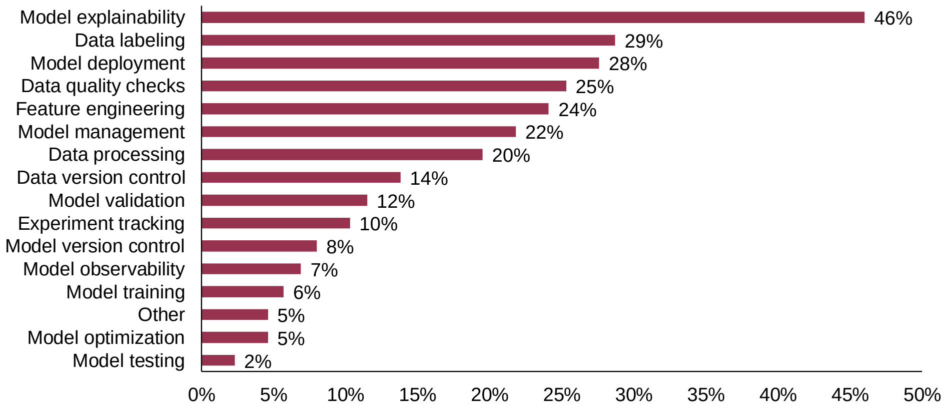
- Legacy corporations. As a beverage DS leader asked, "how do you plan and develop data science on a mature platform?" An enterprise networking leader commented, "when you're designing a new product, it's easier to build with labels in mind. It's much harder to retrofit an existing product that is not ML from the ground up."

- Privacy. A fintech DS leader commented, "how do we see data science changing in a privacy-first and somewhat anti-data mining world?" Another DS leader in marketing technology asked, "how do we ensure privacy without stifling data science?"

Model Challenges

When asked about models specifically, respondents listed explainability as the biggest challenge by a significant margin, followed by data labeling and model deployment.

What are your biggest challenges with models currently?



Our summit discussions included numerous comments on explainability, data, and deployment:

"You're doing this magic. I don't trust it."

- Explainability. A Fortune 10 DS leader said that business users tell his team, "You're doing this magic. I don't trust it yet. That's why I like to use the rules that we have used forever, even if your models show they are performing at only 30%-40%."

- Data labeling. A DS leader commented, "getting clean data in general and labels specifically is still very hard in traditional industries." Another consumer electronics leader stated, "data annotating and labeling is a big bottleneck. Being able to remove the human is very important." An information services DS leader emphasized, "automated data sourcing and categorization would have the largest real-world impact."

"For deployment, we are much more art than science right now."

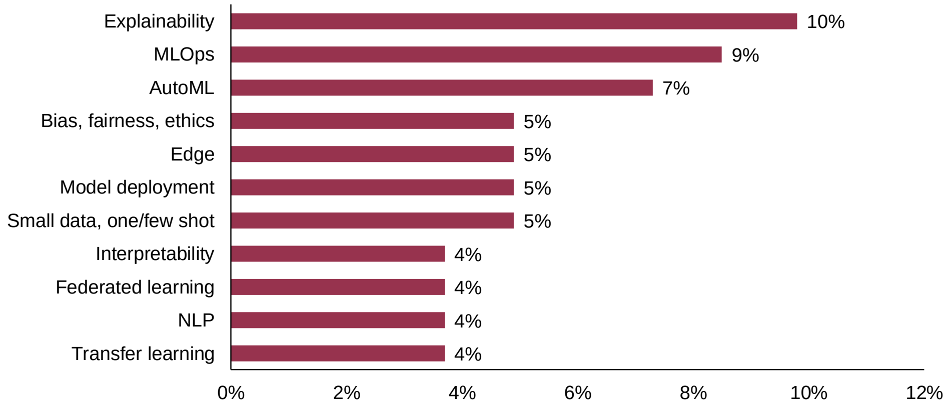
- Model deployment. A retail DS leader stated, "the challenge for us is to get ML models deployed to operational systems. How can we move quickly on deployment? We are much more art than we are science right now." Another leader added, "getting ML to be operational is still hard."

- Data processing. Several DS leaders commented that "adding a data product manager was the #1 most impactful thing we did in the past year."

Near-Term Technology Priorities

We asked respondents in an open-text field for the most important areas for technology development over the next 1-2 years. Explainability, MLOps, and AutoML were the top three responses.

What technology development is the most important for DS/ML over the next 1-2 years?



Participants in our summit shared similar perspectives on near-term priorities:

- Explainability. Numerous DS leaders discussed explainability. One tech leader asked, “how do we debug multilayered neural networks and give precise answers for AI decisions” and “can we use explainable AI to meet GDPR requirements?”

- MLOps. We found a clear, significant shift to ML operationalization. A tech DS leader commented, “the majority of the effort is going to shift to monitoring and maintaining, rather than just the earlier stages of data labeling and processing.” An enterprise software leader stated, “you need to monitor and update daily. You would think it’s 10% of the effort, but it takes 40% in reality.” Other DS leaders mentioned the need for “Simple, Easy, Transparent monitoring” and “handoffs in data science workflows.”

“You would think monitoring and updating is 10% of the effort, but it takes 40% in reality.”

- AutoML. A startup tech DS leader asked, “to what extent is AutoML impacting your day-to-day work?” Other tech leaders commented on the need for “AutoML at scale” and “tools to automate models from data.”

- Federated learning. An airline DS leader stated, “federated learning could dramatically accelerate progress in ML-based healthcare, transportation, etc., because of the access to much more data.” Another leader commented, “you have to be in federated learning. Otherwise, you will be toast.” A tech DS leader asked, “federated learning is a hot area, but can self-supervised learning be used to label data at the edge?”

In addition, participants highlighted several emerging technology areas, which did not rank in the top survey results:

- Multi-modal. A defense DS leader commented that “multi-modal training provides models additional context such as interaction effects, penalties, etc., and leads to enhanced inference.” Numerous other leaders also discussed multi-modal labeling.

“Causality has been around the corner for some time, but people are starting to be able to work with it.”

- Causal inference. A tech DS leader commented, “causality is something that has been around the corner for some time, but people are starting to be able to work with it.” Another tech leader said, “causal inference is fascinating. We built a casual model to understand product usage drivers in order to improve our engagement strategies. The model worked in certain cases, and not in certain cases.”

- Neural-symbolic. A healthcare DS leader stated, “I’m most fascinated in the combination of neural networks and symbolic reasoning techniques to create human-aware, interpretable AI systems to augment daily decision-making.”

Finally, numerous DS leaders discussed lessons learned for the pandemic and preparedness for the current surge. A startup DS leader asked, “how are organizations dealing with the impact of COVID on their data and AI?” A telecom DS leader asked, “what could the data science community learn from the pandemic to adapt to future mass behavioral changes sooner?” Another leader asked, “what are we anticipating for the current new surge, and are models prepared?”

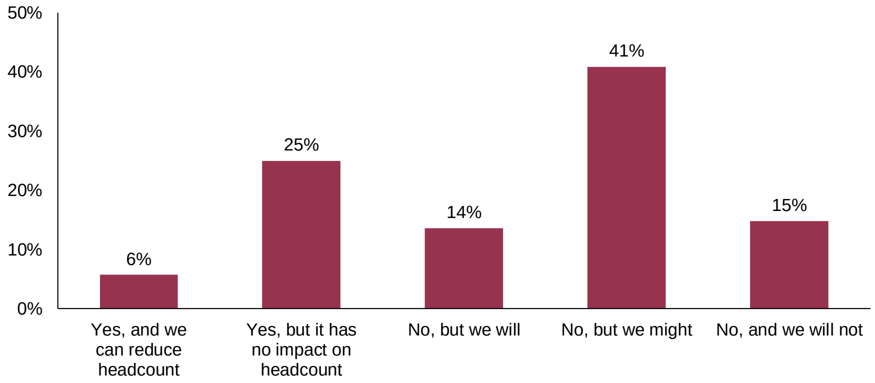
AutoML

31% are currently using AutoML, of which 81% have had no impact on headcount.

AutoML is a key theme for data science, as highlighted in the prior survey response. We asked respondents specific questions on whether they are using AutoML platforms currently and if so, what has been the impact to headcount.

31% of respondents indicated that they are currently using AutoML, of which 81% said that AutoML has had no impact on headcount. 55% indicated that they are planning or considering to use AutoML, and only 15% indicated that they will not use AutoML.

Is your company currently using AutoML platforms?



There were two key debates on AutoML at our summit and in our recent meetings:

- The first was, “will there be a profession called ‘Data Scientist’ in 5 years,” as asked by a tech DS leader. A retail DS leader also asked, “what is the role of a data scientist in a world powered by AutoML?” In our conversations, the vast majority of DS leaders believed that the demand for data scientists will only grow over time, as the volume and complexity growth of data and data science more than offsets the efficiency improvement from automation.

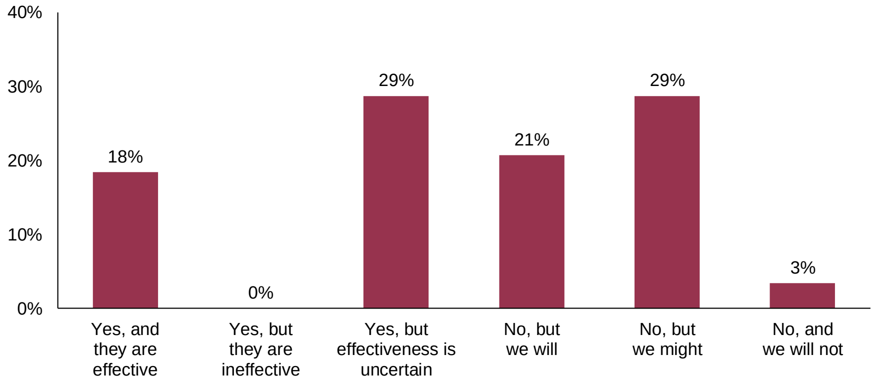
- The second and more interesting debate was the applicability and efficiency of AutoML. A tech DS leader argued, “if you are good with accuracy up to 80%, then you can use AutoML. If you need 85% or higher, then you need data scientists.” Another tech leader stated, “AutoML is not cheap. It is expensive in terms of carbon footprint.” Numerous other DS leaders had very different perspectives, saying “AutoML is great” and “we are using AutoML in my team.”

Data Bias

47% are currently implementing anti-bias solutions, of which 62% are unclear as to their effectiveness.

47% of respondents indicated that they are currently implementing anti-bias solutions, of which 62% said that the solutions’ effectiveness is unclear. 50% indicated that they are planning or considering anti-bias solutions, and only 3% indicated that they will not implement anti-bias solutions.

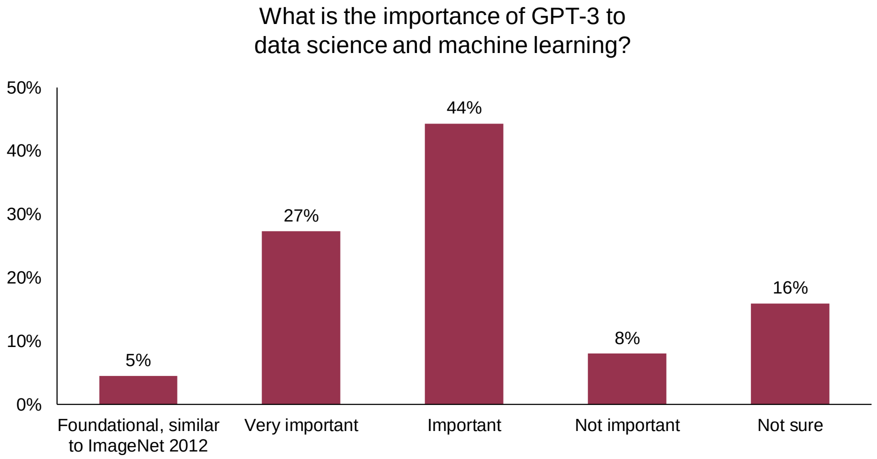
Are you currently implementing solutions to address data bias?



Several DS leaders at our summit were skeptical about the effectiveness of current solutions to address data bias. A security DS leader commented, “we are doing anti-bias. Our view is that it is better to know the problem rather than try to address it in the data or algorithms. Solving it creates more murky waters.” Another tech DS leader said, “building in fairness does not really solve the problem. Sometimes the fix is worse than the original.”

GPT-3

GPT-3 was the data science announcement of the year with a high-profile launch over the summer and an accompanied level of hype. We asked respondents how important is GPT-3 to data science and ML. 71% of respondents indicated that GPT-3 is very important or important. Only 8% of respondents indicated that GPT-3 is not important.



“What is the GPT-3 for inference reasoning?”

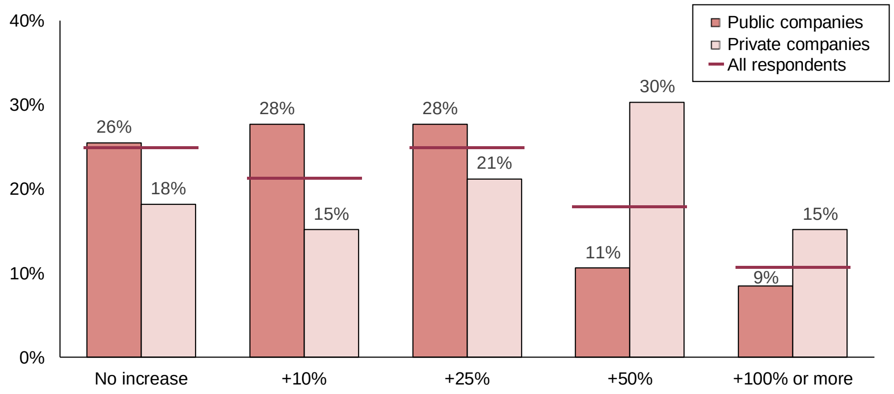
In our conversations, DS leaders continued to be impressed with GPT-3, albeit with a more moderated sense of the hype. Most importantly, leaders separated language processing models from understanding models, with GPT-3 addressing the former. A tech DS leader commented, “the model capacity is amazing. It can memorize everything, but I wouldn’t call it human understanding.” A financial services DS leader asked, “what does GPT-3 exactly ‘understand’?” A consumer electrics DS leader posed the question, “what is the GPT-3 for inference reasoning?”

Budgets and Objectives

48% of public companies indicated DS budget growth of 25% or greater for next year.

Enterprises are increasing investments in data science budgets at meaningful rates. 48% of respondents at public companies indicated that DS budgets will grow at 25% or greater next year. 45% of respondents at private companies indicated DS budget growth of 50% or greater next year.

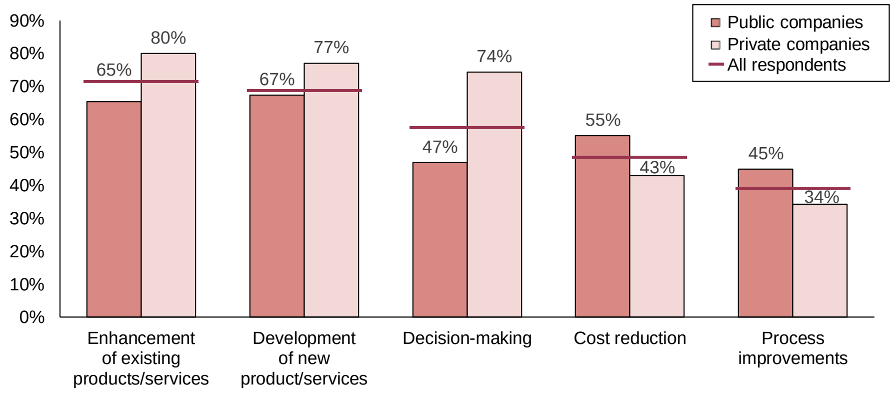
How will your DS/ML budget increase from 2020 to 2021?



74% of private companies use DS for decision-making, as compared to 47% of public companies.

In terms of DS objectives, private companies use DS for decision-making at a significantly higher rate than public companies. 74% of respondents at private companies listed decision-making as a DS objective, as compared to 47% of respondents at public companies.

What are the objectives of your DS/ML projects?



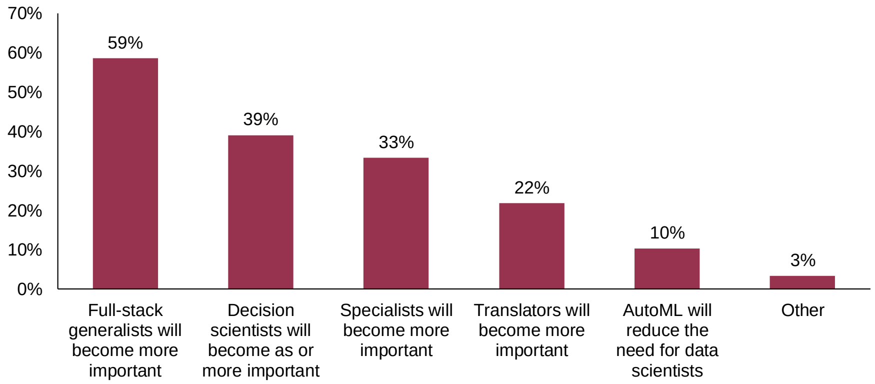
At our summit, it was clear that many companies are experiencing disillusionment with low or negative return on investment from data science projects. As revealed by a BCG report earlier this month, only 10% of companies have seen significant financial benefits.

Separately, we noted that many of the models in production today use traditional methods as opposed to deep learning methods. A CPG DS leader commented, “we are a traditional ML practice, established five years ago. We use linear regressions, decision trees, KNNs. Perhaps five years from now, when we have trust, we can go to the modern generation of ML models.” Another commented, “only 25% of models in production today are deep learning. Deep learning will ramp to 50% of more over the next two years.”

Role of the Data Scientist

As stated by a summit participant above, the role of the data scientist is a canonical issue in corporations. We asked respondents which shifts will define the future of the role. Respondents indicated that 1) full-stack generalists and 2) decision scientists will become more important over time.

How will the role of the data scientist evolve?



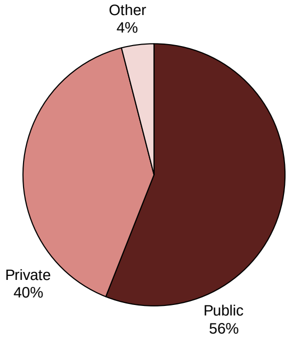
“Data scientists who refuse to build business sense and flex into adjacent functions will be increasingly irrelevant.”

We found several, often-animated, debates on the role of the data scientist at the summit and in recent meetings:

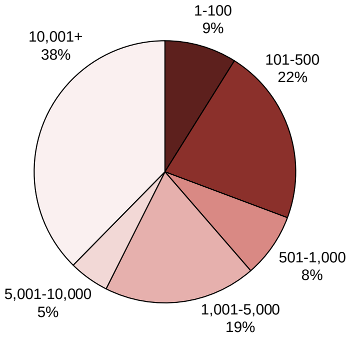
- Full-stack vs. specialist. A fintech DS leader commented, “data scientists who refuse to build business sense and flex into adjacent functions such as data engineering will be increasingly irrelevant.” Other DS leaders asked, “what is your data science dream team” and “will data science titles fragment into specialties?”
- Model. A tech leader commented, “we started in a CoE model, but we have since pushed the data scientists into the business units, still reporting to me. The centralized model works well for shared learnings, but the decentralized model works well if the business units are familiar with how to use machine learning.”
- Structure and culture. Another tech leader said, “data science is science, not engineering. Culturally, a lot of companies shove DS into engineering and apply engineering tools and culture to them.”

Methodology

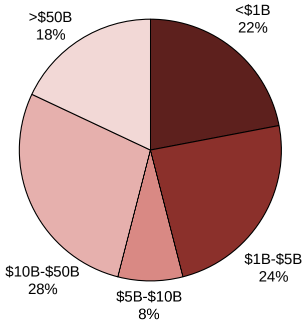
The 88 respondents spanned publicly-held corporations and venture-backed, privately-held companies. The respondent demographics by company type, number of employees, and company valuation are below:



Respondents by Company Type



Respondents by Number of Employees



Respondents by Company Valuation

Conclusion—An Investor's Perspective

Data science is a pillar of the modern enterprise. The field, which is relatively new in many corporations, is growing in importance and in budget. However, there are many challenges across technology, culture, resources, and other fronts. At Wing, we believe that data science is still more art than science and we see an opportunity for entrepreneurship to help address some of the pain points throughout the machine learning model lifecycle.

We see an opportunity for entrepreneurship to address pain points throughout the machine learning model lifecycle.

Model explainability was highlighted by participants as the biggest challenge with models by a wide margin. Amongst the issues raised was trust across many parties—stakeholders in the organization, regulators, and end users. Participants also raised the issue of data bias. While there is a strong desire to take action, there are concerns as to whether current strategies to deal with data bias are effective and fair. Explainability is an area in need of not only continued scientific research and socialization of best practices but also commercial solutions to help companies implement the state of the art.

Data labeling was the next most cited challenge with ML models. Interestingly, data science teams at all levels of maturity described data labeling and annotations as a significant bottleneck in achieving the desired velocity and results from data science initiatives. Hiring a data PM and planning in advance for label collection on greenfield projects were top cited strategies for a return on investment. With the recent progress in academia (e.g. auto-labeling, few-shot learning), it is clear that more of this process can be automated and that the number of things requiring labels can be reduced. We are excited to see how new solutions can be brought into industry to help companies.

Model deployment proved to be the third biggest challenge cited by participants. Not only is getting models deployed to operational systems a painful, manual process, but also model monitoring and updating is an area where industry is not investing enough. ML engineers seem to be building similar solutions across companies. This is an indication that robust new commercial platforms should be able to help.

Thank you to all participants in our Chief Data Scientist Survey and Wing Data Science Summit. We look forward to hearing from you on your thoughts on the aforementioned areas and other areas of interest.

Wing Venture Capital is an early stage venture capital firm focused on seed and series A investments in enterprise technology. As disclosure, a list of Wing's portfolio companies is available on Wing's website here. To subscribe to Wing research, please click here.

THIS PAGE LEFT INTENTIONALLY BLANK