

The Kinds of Data Scientist

by Yael Garten

NOVEMBER 06, 2018



MONTY RAKUSEN/GETTY IMAGES

In 2012, HBR dubbed data scientist “the sexiest job of the 21st century”. It is also, arguably, the vaguest. To hire the right people for the right roles, it’s important to distinguish between different types of data scientist. There are plenty of different distinctions that one can draw, of course, and any attempt to group data scientists into different buckets is by necessity an oversimplification. Nonetheless, I find it helpful to distinguish between the deliverables they create. One type of data scientist creates output for humans to consume, in the form of product and strategy recommendations. They are decision scientists. The other creates output for machines to consume like models, training data, and algorithms. They are modeling scientists.

1. **Data science for humans:** the consumers of the output are decision makers like executives, product managers, designers, or clinicians. They want to draw conclusions from data in order

to make decisions such as which content to license, which sales lead to follow, which medicine is less likely to cause an allergic reaction, which webpage design will lead to more engagement or more purchases, which marketing email will yield higher revenue, or which specific part of a product user experience is suboptimal and needs attention. These data scientists design, define, and implement metrics, run and interpret experiments, create dashboards, draw causal inferences, and generate recommendations from modeling and measurement.

2. **Data science for machines:** here the consumers of the output are computers which consume data in the form of training data, models, and algorithms. Examples of the work products of these data scientists are: recommendation systems which recommend what shirt a customer might like or what medicine a physician should consider prescribing based on a designed optimization function, such as optimizing for customer clicks or for minimizing readmission rates to the hospital. Depending on the engineering background of these data scientists, these work products are either deployed directly to the production system, or if they are prototypes they are handed off to software engineers to help implement, optimize and scale them.

The elusive full stack data scientists do exist, though they are hard to find. In most organizations, it makes sense for data scientists to specialize into one type or another. But data scientist are curious creatures who thrive from being able to creatively dabble; there are benefits to giving them flexibility to work on projects that touch both “types” - both for them and for the organization. (The sidebar offers more detail on how the two types of data scientists differ not only in their skills and the work they do, but in whom they partner with and their measures of success.)

Decision Scientist vs. Modeling Scientist

Who consumes the output?

Decision scientist: Humans.

Modeling scientist: Machines.

What is the output?

Decision scientist: Dashboards, presentations, memos, new metrics, predictive models to inform decision-making, opportunity analysis to determine what to invest in or prioritize, reports on the results of experiments including recommendations.

A more detailed look at data roles

In larger and more sophisticated data operations, more fine-grained roles are necessary. Here are five key areas that contribute to data science operations. In small organizations, one person will do several of these things. In slightly bigger teams, each of these may be a role staffed by one or more individuals. In larger operations, each may be a team unto itself. These roles cover the creation,

Modeling scientist: Models, training data, algorithms.

What are the measures of success?

Decision scientist: Improved decision-making in the organization.

Modeling scientist: Direct improvements in the product or business from the code developed and shipped.

What are some examples?

Decision scientist: Which content to license, which sales lead to follow, which medicine is less likely to cause an allergic reaction, which webpage design will lead to more engagement or more purchases, which marketing email will yield higher revenue, which specific part of a product user experience is suboptimal and needs attention.

Modeling scientist: recommendation systems that recommend what shirt a customer might like or what medicine should be prescribed based on a designed optimization function such as optimizing for customer clicks, or for minimizing return rates to the clinic.

What skills are required?

Decision scientist: Statistics, experimentation, analytical thinking, communication and collaborations skills to work with both technical and non-technical partners, knowledge of both scripting and query languages (e.g. Python, R, SQL), and ideally also formal computer science background.

Modeling scientist: Computer science, machine learning, production-grade coding skills, strong communication to work with both technical and non-technical partners

Who are their main partners on the job?

maintenance, and use of data, and are in addition to the data scientists described above (decision scientists and modeling scientists).

- **Data infrastructure:** data ingestion, availability, operations, access, and running environments to support workflows of data scientists. e.g. running Kafka and a Hadoop cluster
- **Data engineering:** determination of data schemas needed to support measurement and modeling needs, and data cleansing, aggregation, [ETL](#), dataset management
- **Data quality and data governance:** tools, processes, guidelines to ensure data is correct, gated and monitored, documented, standardized. This includes tools for data lineage and data security.
- **Data analytics engineering:** enabling data scientists focused on analytics to scale via analytics applications for internal use, e.g. analytics software libraries, productizing workflows, and analytic microservices.
- **Data product manager:** creating products for internal customers to use within their workflow, to enable incorporation of measurement created by data scientists. Examples include: a portal to read out results of A/B tests, a failure analysis tool, or a dashboard that enables self serve data and root cause diagnosing of changes to metrics or model performance.

Who to hire

So which kind of data scientist should you be recruiting? To answer that question, first decide what stage you are in with your data operation, and second ask how vital data is to your

Decision scientist: Decision makers (executives, business leaders, product managers), data engineers, software engineers responsible for the applications generating data.

Modeling scientist: backend engineers, product managers (to determine what to optimize for), other modeling-scientist colleagues who share techniques, decision scientists on what features to consider and datasets to use.

product. If you're a small organization just starting off and hiring your first data scientist, try to hire someone who can span as many of these roles as possible – the elusive full stack data scientist. If you're larger or farther along in your data operation, the answer will depend more on how essential data is to your product. If your product is going to depend on machine learning from inception, you'll need machine learning expertise in your first hire, or your first leader. If, by contrast, you're looking to identify product opportunities or to improve general

decision-making throughout the organization, you'll need someone more trained in decision science, descriptive and predictive analytics, and statistics, and someone who can translate how to use data across the leadership team and to non-technical partners.

Finally, if you don't have internal data in a format that is consumable or reasonable, you will need a data scientist with a strong enough engineering or computer science background that they can work with engineers to guide what data must be captured and how, before they can start their work.

How to organize

Much has already been written about how data science functions should be organized. Perhaps the most important point is that if data science is a strategic differentiator for the organization, the head of the data science unit should ideally report into the CEO. If this is not possible, they should at least report into someone who understands data strategy and is willing to invest to give it what it needs. Data science has its own skillset, workflow, tooling, integration processes, culture; if it is critical to the organization it is best to not bury it under a part of the organization with a different culture.

The other big question is whether and how to embed data science into the different business lines. There are three basic models: centralized in one data science team, distributed throughout the business lines, or a hybrid between the two where you have a centralized team reporting into one head, but physically co-locate and embed teams of data scientists into business units long

term. Unless your data operation includes several hundreds of employees, it's pretty clear at this point that the hybrid model is most effective. (If you reach this scale, a fully distributed model can make sense, but very few companies work this way.)

In the hybrid model, the centralization in reporting structure enables data scientists to have career progression and growth in a ladder specialized for data scientists, to grow with and be assessed against their peers, and to facilitate and ensure that best practices will be shared across them since they are not each in their own silos. (Establishing this peer group is key; data scientists are curious creatures that want to grow and learn from each other.) Due to the reporting structure, it also enables the leader to more easily promote internal mobility across business groups; this cross-pollination across the company is usually a large benefit.

At the same time, embedding within business groups enables data scientists to establish themselves as domain experts in their business group, and develop a rapport with business partners as an essential long-term part of the team. This partnership will provide the data scientists with rich business context, enabling them to have maximal impact by truly understanding and guiding what business priorities should be addressed using data, and how.

What data scientists need to succeed

Although different kinds of data scientists may have different specialties or duties, there are a few things they all need to succeed. They need business partners who can help them integrate into the core business line and product line. They need data partners – such as software application engineers and data infrastructure engineers – who help ensure the necessary foundational data instrumentation and data feeds are correct, complete, and accessible. And they need leaders willing to invest in the foundations necessary for their work, including data quality, data management, data visualization and access platforms, and a culture of expecting data to be part of the process of business and product development. Key to this is allotting appropriate (and often underestimated) time within the development process for data and measurement. Far too often, product and software teams think of data and measurement as something they can quickly “add on” at the end.

A final piece of advice for those hiring data scientists: Look for people who are in love with solving problems, not with specific solutions or methods, and for people who are incredibly collaborative. No matter what kind of data scientist you are hiring, to be successful they need to

be able to work alongside a vast variety of other job functions – from engineers to product managers to marketers to executive teams. Finally, look for people who have high integrity. As a society, we have a social responsibility to use data for good, and with respect. Data scientists hold the responsibility for data stewardship inside and outside the organization in which they work.

Yael Garten is the Director of Siri Data Science and Engineering at Apple.

This article is about ANALYTICS

 FOLLOW THIS TOPIC


Related Topics: TECHNOLOGY

Comments

Leave a Comment

POST

0 COMMENTS

 **JOIN THE CONVERSATION**

POSTING GUIDELINES

We hope the conversations that take place on HBR.org will be energetic, constructive, and thought-provoking. To comment, readers must sign in or register. And to ensure the quality of the discussion, our moderating team will review all comments and may edit them for clarity, length, and relevance. Comments that are overly promotional, mean-spirited, or off-topic may be deleted per the moderators' judgment. All postings become the property of Harvard Business Publishing.

Harvard Business Review Notice of Use Restrictions, May 2009

Harvard Business Review and Harvard Business Publishing Newsletter content on EBSCOhost is licensed for the private individual use of authorized EBSCOhost users. It is not intended for use as assigned course material in academic institutions nor as corporate learning or training materials in businesses. Academic licensees may not use this content in electronic reserves, electronic course packs, persistent linking from syllabi or by any other means of incorporating the content into course resources. Business licensees may not host this content on learning management systems or use persistent linking or other means to incorporate the content into learning management systems. Harvard Business Publishing will be pleased to grant permission to make this content available through such means. For rates and permission, contact permissions@harvardbusiness.org.