

At Airbnb, Data Science Belongs Everywhere

Insights from Five Years of Hypergrowth



AirbnbEng [Follow](#)

Jul 7, 2015 · 11 min read

By Riley Newman



. . .

Five years ago, I joined Airbnb as its first data scientist.

At that time, the few people who'd even heard of the company were still figuring out how to pronounce its name, and the roughly 7 person team (depending on whether you counted that guy on the couch, the intern, and the barista at our favorite coffee shop) was still operating out of the founders' apartment in SOMA. Put simply, it was pretty early stage.

Bringing me on was a forward-looking move on the part of our founders. This was just prior to the big data craze and the conventional wisdom that data can be a defining competitive advantage. Back then, it was a lot more common to build a data team later in a company's lifecycle. But they were eager to learn and evolve as fast as possible, and I was attracted to the company's culture and mission. So even though we were a very small-data shop at the time, I decided to get involved.

There's a romanticism in Silicon Valley about the early days of a startup: you move fast, make foundational decisions, and any good idea could become the next big thing. From my perspective, that was all true.

Back then we knew so little about the business that any insight was groundbreaking; data infrastructure was fast, stable, and real-time (I was querying our production mysql database); the company was so small that everyone was in the loop about every decision; and the data team (me) was aligned around a singular set of metrics and methodologies.

But five years and 43,000% growth later, things have gotten a bit more complicated. I'm happy to say that we're also more sophisticated in the way we leverage data, and there's now a lot more of it. The trick has been to manage scale in a way that brings together the magic of those early days with the growing needs of the present — a challenge that I know we aren't alone in facing.

So I thought it might be worth pairing our posts on specific problems we're solving with an overview of the higher-level issues data teams encounter as companies grow, and how we at Airbnb have responded. This will mostly center around how to connect data science with other business functions, but I'll break it into three concepts — how we characterize data science, how it's involved in decision-making, and how we've scaled it to reach all sides of Airbnb. I won't say that our solutions are perfect, but we do work every day to retain the excitement, culture, and impact of the early days.

Data Isn't Numbers, It's People

The foundation upon which a data science team rests is the culture and perception of data elsewhere in the organization, so defining how we think about data has been a prerequisite to ingrain data science in business functions.

In the past, data was often referenced in cold, numeric terms. It was construed purely as a measurement tool, which paints data scientists as Spock-like characters expected to have statistics memorized and available upon request. Interactions with us would therefore tend to come in the form of a request for a fact: how many listings do we have in Paris? What are the top 10 destinations in Italy?

While answering questions and measuring things is certainly part of the job, at Airbnb we characterize data in a more human light: **it's the voice of our customers**. A datum is a record of an action or event, which in most cases reflects a decision made by a person. If you can recreate the sequence of events leading up to that decision, you can learn from it; it's an indirect way of the person telling you what they like and don't like — this property is more attractive than that one, I find these features useful but those.. not so much.

This sort of feedback can be a goldmine for decisions about community growth, product development, and resource prioritization. But only if you can decipher it. Thus, data science is an act of interpretation — we translate the customer's 'voice' into a language more suitable for decision-making.

This idea resonates at Airbnb because listening to guests and hosts is core to our culture. Since the early days, our team has met with community members to understand how to make our product better suit their needs. We still do this, but the scale of the community is now beyond the point where it's feasible to connect with everyone everywhere.

So, data has become an ally. We use statistics to understand individual experiences and aggregate those experiences to identify trends across the community; those trends inform decisions about where to drive the business.

Over time, our colleagues on other teams have come to understand that the data team isn't a bunch of Vulcans, but rather that we represent the very human voices of our customers. This has paved the way for changes to the structure of data science at Airbnb.

Proactive Partnership v. Reactive Stats-Gathering

A good data scientist is therefore able to get in the mind of people who use our product and understand their needs. But if they're alone in a forest with no one to act on the insight they uncovered, what difference does it make?

Our distinction between good and great is impact — using insights to influence decisions and ensuring that the decisions had the intended effect. While this may seem obvious, it doesn't happen naturally — when data scientists are pressed for time, they have a tendency to toss the results of an analysis 'over the wall' and then move on to the next problem. This isn't because they don't want to see the project through, but with so much energy invested into understanding the data, ensuring statistical methods are rigorous, and making sure results are interpreted correctly, the communication of their work can feel like a trivial afterthought.

But when decision-makers don't understand the ramifications of an insight, they don't act on it. When they don't act on it, the value of the insight is lost.

The solution, we think, is connecting data scientists as tightly as possible with decision-makers. In some cases, this happens naturally; for example when we develop data products (more on this in a future post). But there's also a strong belief in cross-functional collaboration at Airbnb, which brings up questions about how to structure the team within the broader organization.

A lot has been written about the pros and cons of centralized and embedded data science teams, so I won't focus on that. But suffice to say we've landed on a hybrid of the two.

We began with the centralized model, tempted by its offering of opportunities to learn from each other and stay aligned on metrics, methodologies, and knowledge of past work. While this was all true, we're ultimately in the business of decision-making, and found we couldn't do this successfully when silo'd: partner teams didn't fully understand how to interact with us, and the data scientists on our team didn't have the full context

of what they were meant to solve or how to make it actionable. Over time we became viewed as a resource and, as a result, our work became reactive — responding to requests for statistics rather than being able to think proactively about future opportunities.

So we made the decision to move from a fully-centralized arrangement to a hybrid centralized/embedded structure: we still follow the centralized model, in that we have a singular data science team where our careers unfold, but we have broken this into sub-teams that partner more directly with engineers, designers, product managers, marketers, and others. Doing so has accelerated the adoption of data throughout the company, and has elevated data scientists from reactive stats-gatherers to proactive partners. And by not fully shifting toward an embedded model we're able to maintain a vantage point over every piece of the business, allowing us to form a neural core that can help all sides of the company learn from one another.

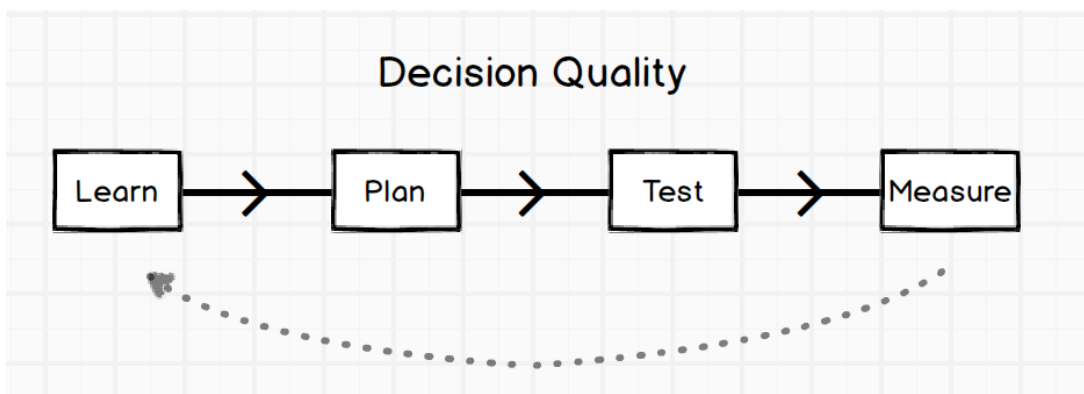
Customer-driven decisions

Structure is a big step toward empowering impactful data science, but it isn't the full story. Once situated within a team that can take action against an insight, the question becomes how and when to leverage the community's voice for business decisions.

Through our partnership with all sides of the company, we've encountered many perspectives on how to integrate data into a project. Some people are naturally curious and like to begin by understanding the context of the problem they're facing. Others view data as a reflection of the past and therefore a weaker guide for planning; but these folks tend to focus more on measuring the impact of their gut-driven decisions.

Both perspectives are fair. Being completely data-driven can lead to optimizing toward a local maximum; finding a global maximum requires shocking the system from time to time. But they reflect different points where data can be leveraged in a project's lifecycle.

Over time, we've identified four stages of the decision-making process that benefit from different elements of data science:



1. We begin by learning about the context of the problem, putting together a full synopsis of past research and efforts toward addressing the opportunity. This is more of an exploratory process aimed at sizing opportunities, and generating hypotheses that lead to actionable insights.
2. That synopsis translates to a plan, which encompasses prioritizing the lever we intend to utilize and forming a hypothesis for the effect of our efforts. Predictive analytics is more relevant in this stage, as we have to make a decision about what path to follow, which is based on where we expect to have the largest impact.
3. As the plan gets underway, we design a controlled experiment through which to roll the plan out. A/B testing is very common now, but our collaboration with all sides of the business opens up opportunities to use experimentation in a broader sense — operational market-based tests, as well as more traditional online environments.
4. Finally, we measure the results of the experiment, identifying the causal impact of our efforts. If successful, we launch to the whole community; if not, we cycle back to learning why it wasn't successful and repeat the process.

Sometimes a step is fairly straightforward, for example if the context of the problem is obvious — the fact that we should build a mobile app doesn't necessitate a heavy synopsis upfront. But the more disciplined we've become about following each step sequentially, the more impactful everyone at Airbnb has become. This makes sense because, ultimately, this process pushes us to solve problems relevant to the community in a way that addresses their needs.

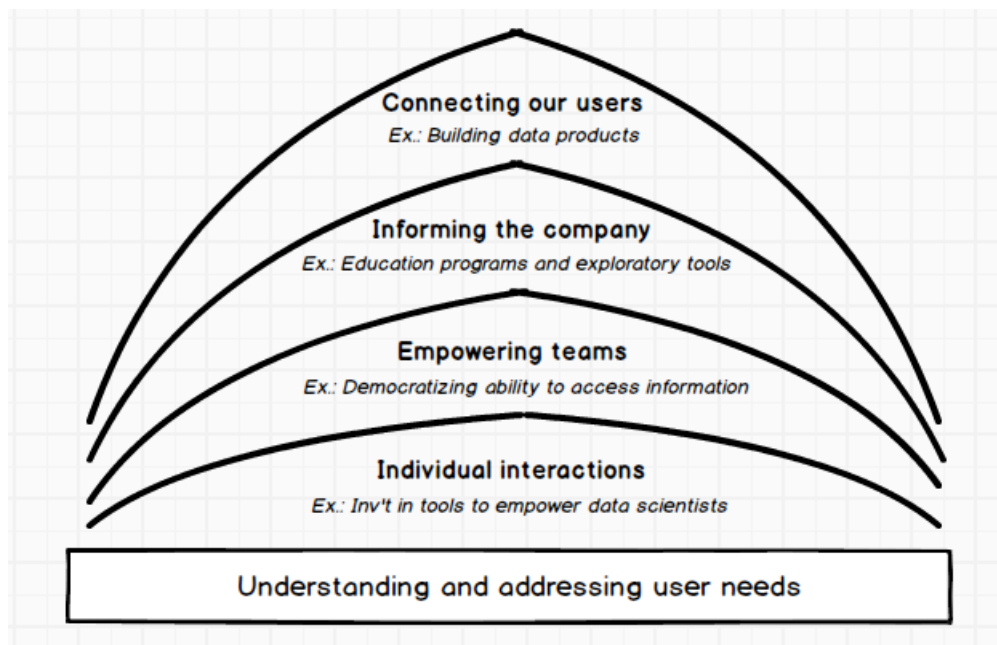
Democratizing Data Science

The above model is great when data scientists have sufficient bandwidth. But the reality of a hypergrowth startup is that the scale and speed at which decisions need to be made will inevitably outpace the growth of the data science team.

This became especially clear in 2011 when Airbnb exploded internationally. Early in the year, we were still a small company based entirely in SF, meaning our army of three data scientists could effectively partner with everyone.

Six months later, we opened over 10 international offices simultaneously, while also expanding our product, marketing, and customer support teams. Our ability to partner directly with every employee suddenly, and irrevocably, disappeared.

Just as it became impossible to meet every new member of the community, it was now also impossible to meet and work with every employee. We needed to find a way to democratize our work, broadening from individual interactions, to empowering teams, the company, and even our community.



Doing this successfully requires becoming more efficient and effective, mostly through investment in the technology surrounding data. Here are some examples of how we've approached each level of scale:

1. Individual interactions become more efficient as data scientists are empowered to move more quickly. Investing in data infrastructure is the biggest lever here — adopting faster and more reliable technologies for querying an ever-growing volume of data. Stabilizing ETL has also been valuable, for example through our development of Airflow.
2. Empowering teams is about removing the burden of reporting and basic data exploration from the shoulders of data scientists so they can focus on more impactful work. Dashboards are a common example of a solution. We've also developed a tool to help people author queries (Airpal) against a robust and intuitive data warehouse.
3. Beyond individual teams, where our work is more tactical, we think about the culture of data in the company as a whole. Educating people on how we think about Airbnb's ecosystem, as well as how to use tools like Airpal, removes barriers to entry and inspires curiosity about how everyone can better leverage data. Similar to empowering teams, this has helped liberate us from ad hoc requests for stats.
4. The broadest example of scaling data science is enabling guests and hosts to learn from each other directly. This mostly happens through data products, where machine learning models interpret signals from one set of community-members to help guide others. Location relevance was one example we wrote about, but as this work is becoming more commonplace in other areas of the company, we've developed tools for making it easier to launch and understand the models we develop.

Scaling a data science team to a company in hypergrowth isn't easy. But it is possible. Especially if everyone agrees that it's not just a nice part of the company, it's an essential part of the company.

Wrestling the train from the monkey

Five years in, we've learned a lot. We've improved how we leverage the data we collect; how we interact with decision-makers; and how we democratize this ability out to the company. But to what extent has all of this work been successful?

Measuring the impact of a data science team is ironically difficult, but one signal is that there's now a unanimous desire to consult data for decisions that need to be made by technical and non-technical people alike. Our team members are seen as partners in the decision-making process, not just reactive stats-gatherers.

Another is that our increasing ability to distill the causal impact of our work has helped us wrestle the train away from the monkey. This has been trickier than one might expect because Airbnb's ecosystem is complicated — a two-sided marketplace with network effects, strong seasonality, infrequent transactions, and long time horizons — but these challenges make the work more exciting. And as much as we've accomplished over the last few years, I think we're still just scratching the surface of our potential.

We're at a point where our infrastructure is stable, our tools are sophisticated, and our warehouse is clean and reliable. We're ready to take on exciting new problems. On the immediate horizon we look forward to shifting from batch to realtime processing; developing a more robust anomaly detection system; deepening our understanding of network effects; and increasing our sophistication around matching and personalization.

But these ideas are just the beginning. Data is the (aggregated) voice of our customers. And wherever we go next—wherever we belong next—will be driven by those voices.



Check out all of our open source projects over at airbnb.io and follow us on Twitter: [@AirbnbEng](https://twitter.com/AirbnbEng) + [@AirbnbData](https://twitter.com/AirbnbData)

. . .

This post originally appeared on [Venturebeat](https://venturebeat.com).

Then was reposted on nerds.airbnb.com on July 7, 2015.