Building Data Science Teams

The Skills, Tools, and Perspectives Behind Great Data Science Groups

DJ Patil

O'REILLY®

O'REILLY®
Strata
Making Data Work

SAN JOSE

LONDON

NEW YORK

SINGAPORE

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Building Data Science Teams

DJ Patil

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Building Data Science Teams
by DJ Patil

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Editor: Mike Loukides

Printing History:

ISBN: 978-1-449-31623-5
1316117207

Table of Contents

Building Data Science Teams

Starting in 2008, Jeff Hammerbacher (@hackingdata) and I sat down to share our experiences building the data and analytics groups at Facebook and LinkedIn. In many ways, that meeting was the start of data science as a distinct professional specialization (see "What Makes a Data Scientist?" on page 11 for the story on how we came up with the title "Data Scientist"). Since then, data science has taken on a life of its own. The hugely positive response to "What Is Data Science?," a great introduction to the meaning of data science in today's world, showed that we were at the start of a movement. There are now regular meetups, well-established startups, and even college curricula focusing on data science. As McKinsey's big data research report and LinkedIn's data indicates indicates (see Figure 1), data science talent is in high demand.

This increase in the demand for data scientists has been driven by the success of the major Internet companies. Google, Facebook, LinkedIn, and Amazon have all made their marks by using data creatively: not just warehousing data, but turning it into something of value. Whether that value is a search result, a targeted advertisement, or a list of possible acquaintances, data science is producing products that people want and value. And it's not just Internet companies: Walmart doesn't produce "data products" as such, but they're well known for using data to optimize every aspect of their retail operations.

Given how important data science has grown, it's important to think about what data scientists add to an organization, how they fit in, and how to hire and build effective data science teams.

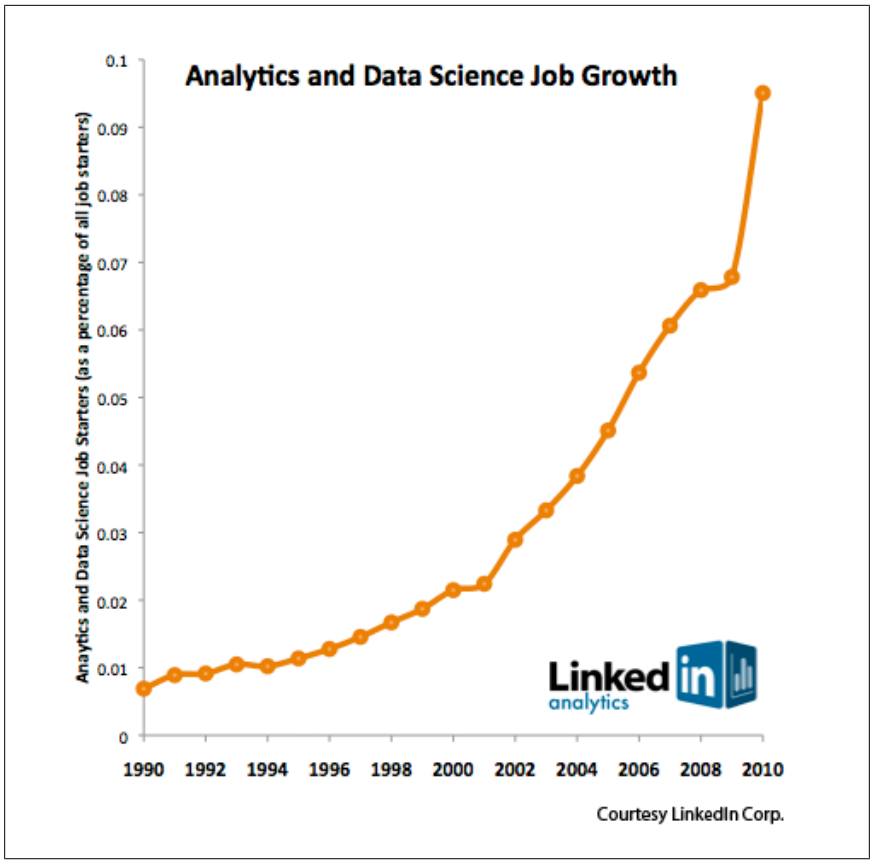

Figure 1. The rise in demand for data science talent

Being Data Driven

Everyone wants to build a data-driven organization. It's a popular phrase and there are plenty of books, journals, and technical blogs on the topic. But what does it really mean to be "data driven"? My definition is:

A data-driven organization acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape.

There are many ways to assess whether an organization is data driven. Some like to talk about how much data they generate. Others like to talk about the sophistication of data they use, or the process of internalizing data. I prefer to start by highlighting organizations that use data effectively.

Ecommerce companies have a long history of using data to benefit their organizations. Any good salesman instinctively knows how to suggest further purchases to a customer. With "People who viewed this item also viewed ...," Amazon moved this technique online. This simple implementation of collaborative filtering is one of their most used features; it is a powerful mechanism for serendipity outside of traditional search. This feature has become so popular that there are now variants such as "People who viewed this item bought ... ." If a customer isn't quite satisfied with the product he's looking at, suggest something similar that might be more to his taste. The value to a master retailer is obvious: close the deal if at all possible, and instead of a single purchase, get customers to make two or more purchases by suggesting things they're likely to want. Amazon revolutionized electronic commerce by bringing these techniques online.

Data products are at the heart of social networks. After all, what is a social network if not a huge dataset of users with connections to each other, forming a graph? Perhaps the most important product for a social network is something to help users connect with others. Any new user needs to find friends, acquaintances, or contacts. It's not a good user experience to force users to search for their friends, which is often a surprisingly difficult task. At LinkedIn, we invented People You May Know (PYMK) to solve this problem. It's easy for software to predict that if James knows Mary, and Mary knows John Smith, then James may know John Smith. (Well, conceptually easy. Finding connections in graphs gets tough quickly as the endpoints get farther apart. But solving that problem is what data scientists are for.) But imagine searching for John Smith by name on a network with hundreds of millions of users!

Although PYMK was novel at the time, it has become a critical part of every social network's offering. Facebook not only supports its own version of PYMK, they monitor the time it takes for users to acquire friends. Using sophisticated tracking and analysis technologies, they have identified the time and number of connections it takes to get a user to long-term engagement. If you connect with a few friends, or add friends slowly, you won't stick around for long. By studying the activity levels that lead to commitment, they have designed the site to decrease the time it takes for new users to connect with the critical number of friends.

Netflix does something similar in their online movie business. When you sign up, they strongly encourage you to add to the queue of movies you intend to watch. Their data team has discovered that once you add more than than a certain number of movies, the probability you will be a long-term customer is significantly higher. With this data, Netflix can construct, test, and monitor product flows to maximize the number of new users who exceed the magic number and become long-term customers. They've built a highly optimized

registration/trial service that leverages this information to engage the user quickly and efficiently.

Netflix, LinkedIn, and Facebook aren't alone in using customer data to encourage long-term engagement — Zynga isn't just about games. Zynga constantly monitors who their users are and what they are doing, generating an incredible amount of data in the process. By analyzing how people interact with a game over time, they have identified tipping points that lead to a successful game. They know how the probability that users will become long-term changes based on the number of interactions they have with others, the number of buildings they build in the first *n* days, the number of mobsters they kill in the first *m* hours, etc. They have figured out the keys to the engagement challenge and have built their product to encourage users to reach those goals. Through continued testing and monitoring, they refined their understanding of these key metrics.

Google and Amazon pioneered the use of A/B testing to optimize the layout of a web page. For much of the web's history, web designers worked by intuition and instinct. There's nothing wrong with that, but if you make a change to a page, you owe it to yourself to ensure that the change is effective. Do you sell more product? How long does it take for users to find the result they're looking for? How many users give up and go to another site? These questions can only be answered by experimenting, collecting the data, and doing the analysis, all of which are second nature to a data-driven company.

Yahoo has made many important contributions to data science. After observing Google's use of MapReduce to analyze huge datasets, they realized that they needed similar tools for their own business. The result was Hadoop, now one of the most important tools in any data scientist's repertoire. Hadoop has since been commercialized by Cloudera, Hortonworks (a Yahoo spin-off), MapR, and several other companies. Yahoo didn't stop with Hadoop; they have observed the importance of streaming data, an application that Hadoop doesn't handle well, and are working on an open source tool called S4 (still in the early stages) to handle streams effectively.

Payment services, such as PayPal, Visa, American Express, and Square, live and die by their abilities to stay one step ahead of the bad guys. To do so, they use sophisticated fraud detection systems to look for abnormal patterns in incoming data. These systems must be able to react in milliseconds, and their models need to be updated in real time as additional data becomes available. It amounts to looking for a needle in a haystack while the workers keep piling on more hay. We'll go into more details about fraud and security later in this article.

Google and other search engines constantly monitor search relevance metrics to identify areas where people are trying to game the system or where tuning is required to provide a better user experience. The challenge of moving and processing data on Google's scale is immense, perhaps larger than any other company today. To support this challenge, they have had to invent novel technical solutions that range from hardware (e.g., custom computers) to software (e.g., MapReduce) to algorithms (PageRank), much of which has now percolated into open source software projects.

I've found that the strongest data-driven organizations all live by the motto "if you can't measure it, you can't fix it" (a motto I learned from one of the best operations people I've worked with). This mindset gives you a fantastic ability to deliver value to your company by:

- Instrumenting and collecting as much data as you can. Whether you're doing business intelligence or building products, if you don't collect the data, you can't use it.
- Measuring in a proactive and timely way. Are your products, and strategies succeeding? If you don't measure the results, how do you know?
- Getting many people to look at data. Any problems that may be present will become obvious more quickly — "with enough eyes all bugs are shallow."
- Fostering increased curiosity about why the data has changed or is not changing. In a data-driven organization, everyone is thinking about the data.

It's easy to pretend that you're data driven. But if you get into the mindset to collect and measure everything you can, and think about what the data you've collected means, you'll be ahead of most of the organizations that claim to be data driven. And while I have a lot to say about professional data scientists later in this post, keep in mind that data isn't just for the professionals. Everyone should be looking at the data.

The Roles of a Data Scientist

In every organization I've worked with or advised, I've always found that data scientists have an influence out of proportion to their numbers. The many roles that data scientists can play fall into the following domains.

Decision sciences and business intelligence

Data has long played a role in advising and assisting operational and strategic thinking. One critical aspect of decision-making support is defining, monitor-

ing, and reporting on key metrics. While that may sound easy, there is a real art to defining metrics that help a business better understand its "levers and control knobs." Poorly-chosen metrics can lead to blind spots. Furthermore, metrics must always be used in context with each other. For example, when looking at percentages, it is still important to see the raw numbers. It is also essential that metrics evolve as the sophistication of the business increases. As an analogy, imagine a meteorologist who can only measure temperature. This person's forecast is always going to be of lower quality than the meteorologist who knows how to measure air pressure. And the meteorologist who knows how to use humidity will do even better, and so on.

Once metrics and reporting are established, the dissemination of data is essential. There's a wide array of tools for publishing data, ranging from simple spreadsheets and web forms, to more sophisticated business intelligence products. As tools get more sophisticated, they typically add the ability to annotate and manipulate (e.g., pivot with other data elements) to provide additional insights.

More sophisticated data-driven organizations thrive on the "democratization" of data. Data isn't just the property of an analytics group or senior management. Everyone should have access to as much data as legally possible. Facebook has been a pioneer in this area. They allow anyone to query the company's massive Hadoop-based data store using a language called Hive. This way, nearly anyone can create a personal dashboard by running scripts at regular intervals. Zynga has built something similar, using a completely different set of technologies. They have two copies of their data warehouses. One copy is used for operations where there are strict service-level agreements (SLA) in place to ensure reports and key metrics are always accessible. The other data store can be accessed by many people within the company, with the understanding that performance may not be always optimal. A more traditional model is used by eBay, which uses technologies like Teradata to create cubes of data for each team. These cubes act like self-contained datasets and data stores that the teams can interact with.

As organizations have become increasingly adept with reporting and analysis, there has been increased demand for strategic decision-making using data. We have been calling this new area "decision sciences." These teams delve into existing data sources and meld them with external data sources to understand the competitive landscape, prioritize strategy and tactics, and provide clarity about hypotheses that may arise during strategic planning. A decision sciences team might take on a problem, like which country to expand into next, or it might investigate whether a particular market is saturated. This analysis might, for example, require mixing census data with internal data and then building

predictive models that can be tested against existing data or data that needs to be acquired.

One word of caution: people new to data science frequently look for a "silver bullet," some magic number around which they can build their entire system. If you find it, fantastic, but few are so lucky. The best organizations look for levers that they can lean on to maximize utility, and then move on to find additional levers that increase the value of their business.

Product and marketing analytics

Product analytics represents a relatively new use of data. Teams create applications that interact directly with customers, such as:

- Products that provide highly personalized content (e.g., the ordering/ranking of information in a news feed).
- Products that help drive the company's value proposition (e.g., "People You May Know" and other applications that suggest friends or other types of connections).
- Products that facilitate the introduction into other products (e.g., "Groups You May Like," which funnels you into LinkedIn's Groups product area).
- Products that prevent dead ends (e.g., collaborative filters that suggest further purchases, such as Amazon's "People who viewed this item also viewed ...").

  Products that are stand alone (e.g., news relevancy products like Google News, LinkedIn Today, etc.).

Given the rapidly decreasing cost of computation, it is easier than ever to use common algorithms and numerical techniques to test the effectiveness of these products.

Similar to product analytics, marketing analytics uses data to explain and showcase a service or product's value proposition. A great example of marketing analytics is OKCupid's blog, which uses internal and external data sources to discuss larger trends. For example, one well-known post correlates the number of sexual partners with smartphone brands. Do iPhone users have more fun? OKCupid knows. Another post studied what kinds of profile pictures are attractive, based on the number of new contacts they generated. In addition to a devoted following, these blog posts are regularly picked up by traditional media, and shared virally through social media channels. The result is a powerful marketing tactic that drives both new users and returning users. Other companies that have used data to drive blogging as a marketing strategy include Mint, LinkedIn, Facebook, and Uber.

Email has long been the basis for online communication with current and potential customers. Using analytics as a part of an email targeting strategy is not new, but powerful analytical technologies can help to create email marketing programs that provide rich content. For example, LinkedIn periodically sends customers updates about changes to their networks: new jobs, significant posts, new connections. This would be spam if it were just a LinkedIn advertisement. But it isn't — it's relevant information about people you already know. Similarly, Facebook uses email to encourage you to come back to the site if you have been inactive. Those emails highlight the activity of your most relevant friends. Since it is hard to delete an email that tells you what your friends are up to, it's extremely effective.

Fraud, abuse, risk and security

Online criminals don't want to be found. They try to hide in the data. There are several key components in the constantly evolving war between attackers and defenders: data collection, detection, mitigation, and forensics. The skills of data scientists are well suited to all of these components.

Any strategy for preventing and detecting fraud and abuse starts with data collection. Data collection is always a challenge, and it is tough to decide how much instrumentation is sufficient. Attackers are always looking to exploit the limitations of your data, but constraints such as cost and storage capacity mean that it's usually impossible to collect all the data you'd like. The ability to recognize which data needs to be collected is essential. There's an inevitable "if only" moment during an attack: "if only we had collected x and y, we'd be able to see what is going on."

Another aspect of incident response is the time required to process data. If an attack is evolving minute by minute, but your processing layer takes hours to analyze the data, you won't be able to respond effectively. Many organizations are finding that they need data scientists, along with sophisticated tooling, to process and analyze data quickly enough to act on it.

Once the attack is understood, the next phase is mitigation. Mitigation usually requires closing an exploit or developing a model that segments bad users from good users. Success in this area requires the ability to take existing data and transform it into new variables that can be acted upon. This is a subtle but critical point. As an example, consider IP addresses. Any logging infrastructure almost certainly collects the IP addresses that connect to your site. Addresses by themselves are of limited use. However, an IP address can be transformed into variables such as:

- The number of bad actors seen from this address during some period of time.

- The country from which the address originated, and other geographic information.
- Whether the address is typical for this time of day.

From this data, we now have derived variables that can be built into a model for an actionable result. Domain experts who are data scientists understand how to make variables out of the data. And from those variables, you can build detectors to find the bad guys.

Finally, forensics builds a case against the attackers and helps you learn about the true nature of the attack and how to prevent (or limit) such attacks in the future. Forensics can be a time-consuming process where the data scientists sift through all of the data to piece together a puzzle. Once the puzzle has been put together, new tooling, processes, and monitoring can be put in place.

Data services and operations

One of the foundational components of any data organization is data services and operations. This team is responsible for the databases, data stores, data structures (e.g., data schemas), and the data warehouse. They are also responsible for the monitoring and upkeep of these systems. The other functional areas cannot exist without a top-notch data services and operations group; you could even say that the other areas live on top of this area. In some organizations, these teams exist independently of traditional operations teams. In my opinion, as these systems increase in sophistication, they need even greater coordination with operations groups. The systems and services this functional area provides need to be deployed in traditional data centers or in the cloud, and they need to be monitored for stability; staff also must be on hand to respond when systems go down. Established operations groups have expertise in these areas, and it makes sense to take advantage of such skills.

As an organization builds out its reporting requirements, the data services and operations team should become responsible for the reporting layer. While team members may not focus on defining metrics, they are critical in ensuring that the reports are delivered in a timely manner. Therefore, collaboration between data services and decision sciences is absolutely essential. For example, while a metric may be easy to define on paper, implementing it as part of a regular report may be unrealistic: the database queries required to implement the metric may be too complex to run as frequently as needed.

Data engineering and infrastructure

It's hard to understate the sophistication of the tools needed to instrument, track, move, and process data at scale. The development and implementation

of these technologies is the responsibility of the data engineering and infrastructure team. The technologies have evolved tremendously over the past decade, with an incredible amount of collaboration taking place through open source projects. Here are just a few samples:

- Kafka, Flume, and Scribe are tools for streaming data collection. While the models differ, the general idea is that these programs collect data from many sources; aggregate the data; and feed it to a database, a system like Hadoop, or other clients.
- Hadoop is currently the most widely used framework for processing data. Hadoop is an open source implementation of the MapReduce programming model that Google popularized in 2004. It is inherently batch-oriented; several newer technologies are aimed at processing streaming data, such as S4 and Storm.
- Azkaban and Oozie are job schedulers. They manage and coordinate complex data flows.
- Pig and Hive are languages for querying large non-relational datastores. Hive is very similar to SQL. Pig is a data-oriented scripting language.
- Voldemort, Cassandra, and HBase are data stores that have been designed for good performance on very large datasets.

Equally important is the ability to build monitoring and deployment technologies for these systems.

In addition to building the infrastructure, data engineering and infrastructure takes ideas developed by the product and marketing analytics group and implements them so they can operate in production at scale. For example, a recommendation engine for videos may be prototyped using SQL, Pig, or Hive. If testing shows that the recommendation engine is of value, it will need to be deployed so that it supports SLAs specifying appropriate availability and latencies. Migrating the product from prototype into production may require re-implementing it so it can deliver performance at scale. If SQL and a relational database prove to be too slow, you may need to move to HBase, queried by Hive or Pig. Once the application has been deployed, it must be monitored to ensure that it continues meeting its requirements. It must also be monitored to ensure that it is producing relevant results. Doing so requires more sophisticated software development.

Organizational and reporting alignment

Should an organization be structured according to the functional areas I've discussed, or via some other mechanism? There is no easy answer. Key things to consider include the people involved, the size and scale of the organization,

and the organizational dynamics of the company (e.g., whether the company is product, marketing, or engineering driven).

In the early stages, people must wear multiple hats. For example, in a startup, you can't afford separate groups for analytics, security, operations, and infrastructure: one or two people may have to do everything. But as an organization grows, people naturally become more specialized. In addition, it's a good idea to remove any single points of failure. Some organizations use a "center-of-excellence model," where there is a centralized data team. Others use a hub-and-spoke model, where there is one central team and members are embedded within sponsoring teams (for example, the sales team may sponsor people in analytics to support their business needs). Some organizations are fully decentralized, and each team hires to fill its own requirements.

As vague as that answer is, here are the three lessons I've learned:

1. If the team is small, its members should sit close to each other. There are many nuances to working with data, and high-speed interaction between team members resolves painful, trivial issues.
2. Train people to fish — it only increases your organization's ability to be data driven. As previously discussed, organizations like Facebook and Zynga have democratized data effectively. As a result, these companies have more people conducting more analysis and looking at key metrics. This kind of access was nearly unheard of as little as five years ago. There is a down side: the increased demands on the infrastructure and need for training. The infrastructure challenge is largely a technical problem, and one of the easiest ways to manage training is to set up "office hours" and schedule data classes.
3. All of the functional areas must stay in regular contact and communication. As the field of data science grows, technology and process innovations will also continue to grow. To keep up to date it is essential for all of these teams to share their experiences. Even if they are not part of the same reporting structure, there is a common bond of data that ties everyone together.

What Makes a Data Scientist?

When Jeff Hammerbacher and I talked about our data science teams, we realized that as our organizations grew, we both had to figure out what to call the people on our teams. "Business analyst" seemed too limiting. "Data analyst" was a contender, but we felt that title might limit what people could do. After all, many of the people on our teams had deep engineering expertise. "Research scientist" was a reasonable job title used by companies like Sun,

HP, Xerox, Yahoo, and IBM. However, we felt that most research scientists worked on projects that were futuristic and abstract, and the work was done in labs that were isolated from the product development teams. It might take years for lab research to affect key products, if it ever did. Instead, the focus of our teams was to work on data applications that would have an immediate and massive impact on the business. The term that seemed to fit best was data scientist: those who use both data and science to create something new.

(Note: Although the term "data science" has a long history — usually referring to business intelligence — "data scientist" appears to be new. Jeff and I have been asking if anyone else has used this term before we coined it, but we've yet to find anyone who has.)

But how do you find data scientists? Whenever someone asks that question, I refer them back to a more fundamental question: what makes a good data scientist? Here is what I look for:

- *Technical expertise*: the best data scientists typically have deep expertise in some scientific discipline.
- *Curiosity*: a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- *Storytelling*: the ability to use data to tell a story and to be able to communicate it effectively.
- *Cleverness*: the ability to look at a problem in different, creative ways.

People often assume that data scientists need a background in computer science. In my experience, that hasn't been the case: my best data scientists have come from very different backgrounds. The inventor of LinkedIn's People You May Know was an experimental physicist. A computational chemist on my decision sciences team had solved a 100-year-old problem on energy states of water. An oceanographer made major impacts on the way we identify fraud. Perhaps most surprising was the neurosurgeon who turned out to be a wizard at identifying rich underlying trends in the data.

All the top data scientists share an innate sense of curiosity. Their curiosity is broad, and extends well beyond their day-to-day activities. They are interested in understanding many different areas of the company, business, industry, and technology. As a result, they are often able to bring disparate areas together in a novel way. For example, I've seen data scientists look at sales processes and realize that by using data in new ways they can make the sales team far more efficient. I've seen data scientists apply novel DNA sequencing techniques to find patterns of fraud.

What unifies all these people? They all have strong technical backgrounds. Most have advanced degrees (although I've worked with several outstanding

data scientists who haven't graduated from college). But the real unifying thread is that all have had to work with a tremendous amount of data before starting to work on the "real" problem. When I was a first-year graduate student, I was interested in weather forecasting. I had an idea about how to understand the complexity of weather, but needed lots of data. Most of the data was available online, but due to its size, the data was in special formats and spread out over many different systems. To make that data useful for my research, I created a system that took over every computer in the department from 1 AM to 8 AM. During that time, it acquired, cleaned, and processed that data. Once done, my final dataset could easily fit in a single computer's RAM. And that's the whole point. The heavy lifting was required before I could start my research. Good data scientists understand, in a deep way, that the heavy lifting of cleanup and preparation isn't something that gets in the way of solving the problem: it *is* the problem.

These are some examples of training that hone the skills a data scientist needs to be successful:

- Finding rich data sources.
- Working with large volumes of data despite hardware, software, and bandwidth constraints.
- Cleaning the data and making sure that data is consistent.
- Melding multiple datasets together.
- Visualizing that data.
- Building rich tooling that enables others to work with data effectively.

One of the challenges of identifying data scientists is that there aren't many of them (yet). There are a number of programs that are helping train people, but the demand outstrips the supply. And experiences like my own suggest that the best way to become a data scientist isn't to be trained as a data scientist, but to do serious, data-intensive work in some other discipline.

Hiring data scientists was such a challenge at every place I've worked that I've adopted two models for building and training new hires. First, hire people with diverse backgrounds who have histories of playing with data to create something novel. Second, take incredibly bright and creative people right out of college and put them through a very robust internship program.

Another way to find great data scientists is to run a competition, like Netflix did. The Netflix Prize was a contest organized to improve their ability to predict how much a customer would enjoy a movie. If you don't want to organize your own competition, you can look at people who have performed well in competitions run by others. Kaggle and Topcoder are great resources when looking

for this kind of talent. Kaggle has found its own top talent by hiring the best performers from its own competitions.

Hiring and talent

Many people focus on hiring great data scientists, but they leave out the need for continued intellectual and career growth. These key aspects of growth are what I call *talent growth*. In the three years that I led LinkedIn's analytics and data teams, we developed a philosophy around three principles for hiring and talent growth.

Would we be willing to do a startup with you?

This is the first question we ask ourselves as a team when we meet to evaluate a candidate. It sums up a number of key criteria:

- *Time*: If we're willing to do a startup with you, we're agreeing that we'd be willing to be locked in a small room with you for long periods of time. The ability to enjoy another person's company is critical to being able to invest in each other's growth.
- *Trust*: Can we trust you? Will we have to look over your shoulder to make sure you're doing an A+ job? That may go without saying, but the reverse is also important: will you trust me? If you don't trust me, we're both in trouble.
- *Communication*: Can we communicate with each other quickly and efficiently? If we're going to spend a tremendous amount of time together and if we need to trust each other, we'll need to communicate. Over time, we should be able to anticipate each other's needs in a way that allows us to be highly efficient.

Can you "knock the socks off" of the company in 90 days?

Once the first criteria has been met, it's critical to establish mechanisms to ensure that the candidate will succeed. We do this by setting expectations for the quality of the candidate's work, and by setting expectations for the velocity of his or her progress.

First, the "knock the socks off" part: by setting the goal high, we're asking whether you have the mettle to be part of an elite team. More importantly, it is a way of establishing a handshake for ensuring success. That's where the 90 days comes in. A new hire won't come up with something mind blowing if the team doesn't bring the new hire up to speed quickly. The team needs to orient new hires around existing systems and processes. Similarly, the new hire needs to make the effort to progress, quickly. Does this person ask questions when

they get stuck? There are no dumb questions, and toughing it out because you're too proud or insecure to ask is counterproductive. Can the new hire bring a new system up in a day, or does it take a week or more? It's important to understand that doing something mind-blowing in 90 days is a team goal, as much as an individual goal. It is essential to pair the new hire with a successful member of the team. Success is shared.

This criterion sets new hires up for long-term success. Once they've passed the first milestone, they've done something that others in the company can recognize, and they have the confidence that will lead to future achievements. I've seen everyone from interns all the way to seasoned executives meet this criterion. And many of my top people have had multiple successes in their first 90 days.

In four to six years, will you be doing something amazing?

What does it mean to do something amazing? You might be running the team or the company. You might be doing something in a completely different discipline. You may have started a new company that's changing the industry. It's difficult to talk concretely because we're talking about potential and long-term futures. But we all want success to breed success, and I believe we can recognize the people who will help us to become mutually successful.

I don't necessarily expect a new hire to do something amazing while he or she works for me. The four- to six-year horizon allows members of the team to build long-term road maps. Many organizations make the time commitment amorphous by talking about vague, never-ending career ladders. But professionals no longer commit themselves to a single company for the bulk of their careers. With each new generation of professionals, the number of organizations and even careers has increased. So rather than fight it, embrace the fact that people will leave, so long as they leave to do something amazing. What I'm interested in is the potential: if you have that potential, we all win and we all grow together, whether your biggest successes come with my team or somewhere else.

Finally, this criteria is mutual. A new hire won't do something amazing, now or in the future, if the organization he or she works for doesn't hold up its end of the bargain. The organization must provide a platform and opportunities for the individual to be successful. Throwing a new hire into the deep end and expecting success doesn't cut it. Similarly, the individual must make the company successful to elevate the platform that he or she will launch from.

Building the LinkedIn Data Science Team

I'm proud of what we've accomplished in building the LinkedIn data team. However, when we started, it didn't look anything like the organization that is there today. We started with 1.5 engineers (who would later go on to invent Voldemort, Kafka, and the real-time recommendation engine systems), no data services team (there wasn't even a data warehouse), and five analysts (who would later become the core of LinkedIn's data science group) who supported everyone from the CFO to the product managers.

When we started to build the team, the first thing I did was go to many different technical organizations (the likes of Yahoo, eBay, Google, Facebook, Sun, etc.) to get their thoughts and opinions. What I found really surprised me. The companies all had fantastic sets of employees who could be considered "data scientists." However, they were uniformly discouraged. They did first-rate work that they considered critical, but that had very little impact on the organization. They'd finish some analysis or come up with some ideas, and the product managers would say "that's nice, but it's not on our roadmap." As a result, the data scientists developing these ideas were frustrated, and their organizations had trouble capitalizing on what they were capable of doing.

Our solution was to make the data group a full product team responsible for designing, implementing, and maintaining products. As a product team, data scientists could experiment, build, and add value directly to the company. This resulted not only in further development of LinkedIn products like PYMK and Who's Viewed My Profile, but also in features like Skills, which tracks various skills and assembles a picture of what's needed to succeed in any given area, and Career Explorer, which helps you explore different career trajectories.

It's important that our data team wasn't comprised solely of mathematicians and other "data people." It's a fully integrated product group that includes people working in design, web development, engineering, product marketing, and operations. They all understand and work with data, and I consider them all data scientists. We intentionally kept the distinction between different roles in the group blurry. Often, an engineer can have the insight that makes it clear how the product's design should work, or vice-versa — a designer can have the insight that helps the engineers understand how to better use the data. Or it may take someone from marketing to understand what a customer really wants to accomplish.

The silos that have traditionally separated data people from engineering, from design, and from marketing, don't work when you're building data products. I would contend that it is questionable whether those silos work for any kind of product development. But with data, it never works to have a waterfall process in which one group defines the product, another builds visual mock-

ups, a data scientist preps the data, and finally a set of engineers builds it to some specification document. We're not building Microsoft Office, or some other product where there's 20-plus years of shared wisdom about how interfaces should work. Every data project is a new experiment, and design is a critical part of that experiment. It's similar for operations: data products present entirely different stresses on a network and storage infrastructure than traditional sites. They capture much more data: petabytes and even exabytes. They deliver results that mash up data from many sources, some internal, some not. You're unlikely to create a data product that is reliable and that performs reasonably well if the product team doesn't incorporate operations from the start. This isn't a simple matter of pushing the prototype from your laptop to a server farm.

Finally, quality assurance (QA) of data products requires a radically different approach. Building test datasets is nontrivial, and it is often impossible to test all of the use cases. As different data streams come together into a final product, all sorts of relevance and precision issues become apparent. To develop this kind of product effectively, the ability to adapt and iterate quickly throughout the product life cycle is essential. To ensure agility, we build small groups to work on specific products, projects, or analyses. When we can, I like to seat anyone with a dependency with another person in the same area.

A data science team isn't just people: it's tooling, processes, the interaction between the team and the rest of the company, and more. At LinkedIn, we couldn't have succeeded if it weren't for the tools we used. When you're working with petabytes of data, you need serious power tools to do the heavy lifting. Some, such as Kafka and Voldemort (now open source projects) were homegrown, not because we thought we should have our own technology, but because we didn't have a choice. Our products couldn't scale without them. In addition to these technologies, we use other open source technologies such as Hadoop and many vendor-supported solutions as well. Many of these are for data warehousing, and traditional business intelligence.

Tools are important because they allow you to automate. Automation frees up time, and makes it possible to do the creative work that leads to great products. Something as simple as reducing the turnaround time on a complex query from "get the result in the morning" to "get the result after a cup of coffee" represents a huge increase in productivity. If queries run overnight, you can only afford to ask questions when you already think you know the answer. If queries run in minutes, you can experiment and be creative.

Interaction between the data science teams and the rest of corporate culture is another key factor. It's easy for a data team (any team, really) to be bombarded by questions and requests. But not all requests are equally important. How do you make sure there's time to think about the big questions and the

big problems? How do you balance incoming requests (most of which are tagged "as soon as possible") with long-term goals and projects? It's important to have a culture of prioritization: everyone in the group needs to be able to ask about the priority of incoming requests. Everything can't be urgent.

The result of building a data team is, paradoxically, that you see data products being built in all parts of the company. When the company sees what can be created with data, when it sees the power of being data enabled, you'll see data products appearing everywhere. That's how you know when you've won.

Reinvention

Companies are always looking to reinvent themselves. There's never been a better time: from economic pressures that demand greater efficiency, to new kinds of products that weren't conceivable a few years ago, the opportunities presented by data are tremendous.

But it's a mistake to treat data science teams like any old product group. (It is probably a mistake to treat any old product group like any old product group, but that's another issue.) To build teams that create great data products, you have to find people with the skills and the curiosity to ask the big questions. You have build cross-disciplinary groups with people who are comfortable creating together, who trust each other, and who are willing to help each other be amazing. It's not easy, but if it were easy, it wouldn't be as much fun.

About the Author

Dr. DJ Patil is a Data Science in Residence at Greylock Partners. He has held a variety of roles in academia, industry, and government. These include the Chief Scientist, Chief Security Officer and Head of Analytics and Data Teams at the LinkedIn Corporation. Additionally he has held a number of roles at Skype, PayPal, and eBay. As a member of the faculty at the University of Maryland, his research focused on nonlinear dynamics and chaos theory applied to numerical weather prediction. As an AAAS Science & Technology Policy Fellow for the Department of Defense, Dr. Patil directed new efforts to leverage social network analysis and the melding of computational and social sciences to anticipate emerging threats to the US. He has also co-chaired a major review of US efforts to prevent bioweapons proliferation in Central Asia and co-founded the Iraqi Virtual Science Library (IVSL). More details can be found on his LinkedIn profile and he can be followed on Twitter (@dpatil).

Be at the forefront of the data revolution.

February 28 – March 1, 2012
Santa Clara, CA

Strata offers the nuts-and-bolts of building a data-driven business.

- See the latest tools and technologies you need to make data work
- Find new ways to leverage data across industries and disciplines
- Take big data from science project to real business application
- Understand the career opportunities for data professionals

Strata Conference is for developers, data scientists, data analysts, and other data professionals.

Registration is now open at strataconf.com/strata2012

20% Exclusive Discount for Strata Alumni
Use code STR12ALM

O'REILLY®

O'REILLY®
Strata
Making Data Work