White Paper

The Quest for Unicorns

Kenny Darrell
Lead Data Scientist

November 13, 2015

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

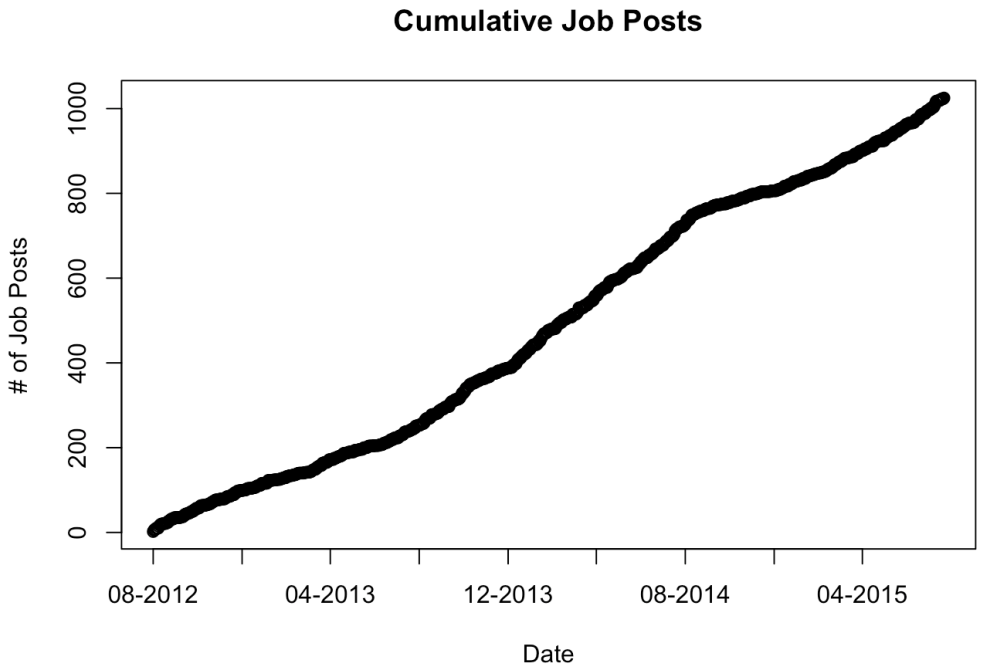
Will there be enough data scientists in the future? The question sounds like a subplot for a science fiction film, but it has received much attention over the past few years due to the forecast of a substantial shortfall in the industry. A McKinsey study has projected that “by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills.” This deficit is making it increasingly challenging to hire data analysts; because of their rarity, they are now beginning to be described as unicorns. A recent article even magnifies the issue by calling them purple unicorns, and another argues that we are looking for a platypus because we do not know what we want.

Predictions are difficult to make, especially about the future, but it is helpful to quantify the problem. As data scientists and companies seeking to hire data scientists, we should break down this anecdotal deficit notion to see what it means. Data Science is not a new concept; while the term is new, people have been performing these tasks for many years. So, we must ask ourselves if the deficit is related to the number of people available, or if the problem lies in how we define a Data Scientist.

Data Science seems to mean one with the full range of skills that could be needed in the industry, making it challenging for a company to find the right talent and nearly impossible for one person to have every skill required. Additionally, as new technologies emerge every year, the definition broadens; a perfect fit today may not have every skill required of a new hire a year from now. What’s the solution? Finding 200,000 people who can use Hadoop is a challenge, as it is still a somewhat new technology; however, finding 200,000 analytically-driven people is a different matter entirely. It could be that the deficit will shrink substantially if we simply temper our expectations and allow people the time and training they need to acquire new, specific skill sets.

The challenge of defining roles in a developing industry isn’t new. It has already appeared in the software industry, but eventually different labels were converged upon to define different roles. Front-end staff work on the parts of software that people see (javascript/html/css) and back-end staff work with servers and performance (java/scala), while full-stack employees have some knowledge in a lot of areas. Data Science is beginning to settle on different roles as well, as we can see in an excellent article by a member of the Twitter team. The article splits the industry into Type A (statistics-focused) and Type B (engineering-focused) Data Scientists. There has also been the growing but dwarfed effort to tag this Type B role as a “Data Engineer.” Another excellent article, Analyzing the Analyzers, attempted to define new categories by looking at what Data Scientists really do, resulting in roles that real humans can fill. One of its goals was to change what hiring managers were asking for and how they crafted job posts. Did their methods work? Are companies asking for too much, maybe even everything? Is the deficit a result of expecting skills few had had time to develop? These questions are essential to answer to define realistic roles within the industry. Despite these articles, however, many companies have not yet adopted new roles or updated their job responsibilities.

Sherlock Holmes claimed, “It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.” To adhere to this rule of one of the first Data Scientists, we need some data before we continue theorizing. On second thought, before we label Sherlock Holmes a Data Scientist, we need to determine whether he was proficient in the ways of Hadoop and Spark, had a PhD Statistics with nine years of DBA experience, and could write efficient code object oriented and functional code for the Babbage Engine. But I digress. The data we will use to examine this problem includes about 1,000 job postings for Data Scientist-related positions over roughly three years.



The postings are pretty uniform over time to avoid over-emphasizing any period more than any other. Our first step is to ensure that these postings are related to our discussion. To examine this, we can look at the position titles for the postings.

Number	Position Title	Quantity
1	Data Scientist	476
2	Engineer	142
3	Analytics	119
4	Analyst	118
5	Learning	66
6	Machine	65
7	Research	60
8	Software	48
9	Big data	39
10	Business	32
11	Predictive	29
12	Statistician	27
13	Modeler	27

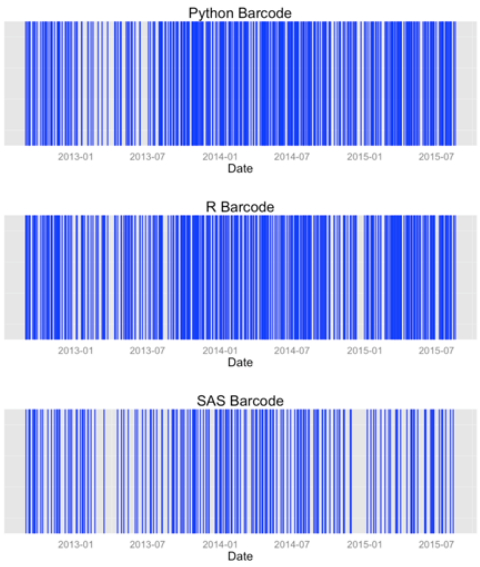
14	Developer	22
15	Quantitative	20
16	Architect	19
17	Intelligence	17
18	Mining	9
19	Database	8
20	Hadoop	8

All of these terms are related to the field of Data Science. We can see from the data that many companies are starting to ask for more precise roles. While all of the jobs are related to the field of data science, some may be asking for a software developer or system architect; however, the go-to term still seems to be Data Scientist. What type of experience are companies looking for?

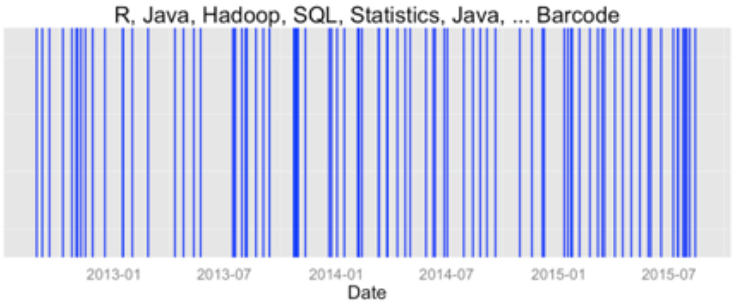
Number	Position Level	Quantity
1	Senior	102
2	Principal	14
3	Manager	65
4	Lead	47
5	Junior	14
6	Director	19

An important observation here is that significantly more people want senior staff than junior Data Scientists. This is part of the problem: many companies are not willing to train people or allow them to learn on the job.

These charts don't show us how requirements have shifted over time, so the barcode plots below show vertical lines for each day that a post requested that skill.

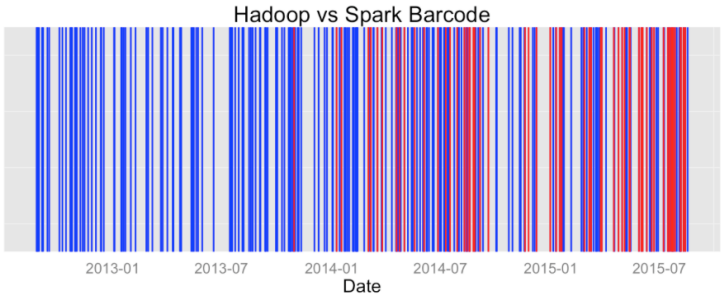


Python and R are very common job requirements, and SAS is often mentioned as well. The next question is: how many jobs ask for all three of them?
This chart looks very similar to the chart showing job descriptions only mentioning SAS. For many companies, if you want a Data Scientist with one of these skills, you ask for all of them. It would be unfair to say this is a bad thing; companies could be acknowledging that if you know one, you can quickly learn the others, rather than requiring knowledge in all three. To be certain, we would need to know if the post asks for 'Python, R, and SAS' or 'Python, R, or SAS.' Since this is a relatively minor difference, we will give them the benefit of the doubt. What is more concerning is what happens when we throw statistics, SQL, Hadoop, and Java into the mix, as you can see in the chart below:



Companies appearing in this chart aren't really looking for a specific candidate they need: they are looking for everything and the kitchen sink. They know that their data needs will evolve and become increasingly important, so they want to have people in place who have exposure to all the current languages of data science. The result of this trend is that companies who require experience modeling Hadoop may not even use Hadoop; they know that having these skills in their talent pool could help them in the future. It is a good strategy to look for someone who can do a bit of everything; however, when it comes to the way jobs are posted, what solicitations are calling for, what managers say they really want, the position requirements aren't reflecting the true needs of the client.

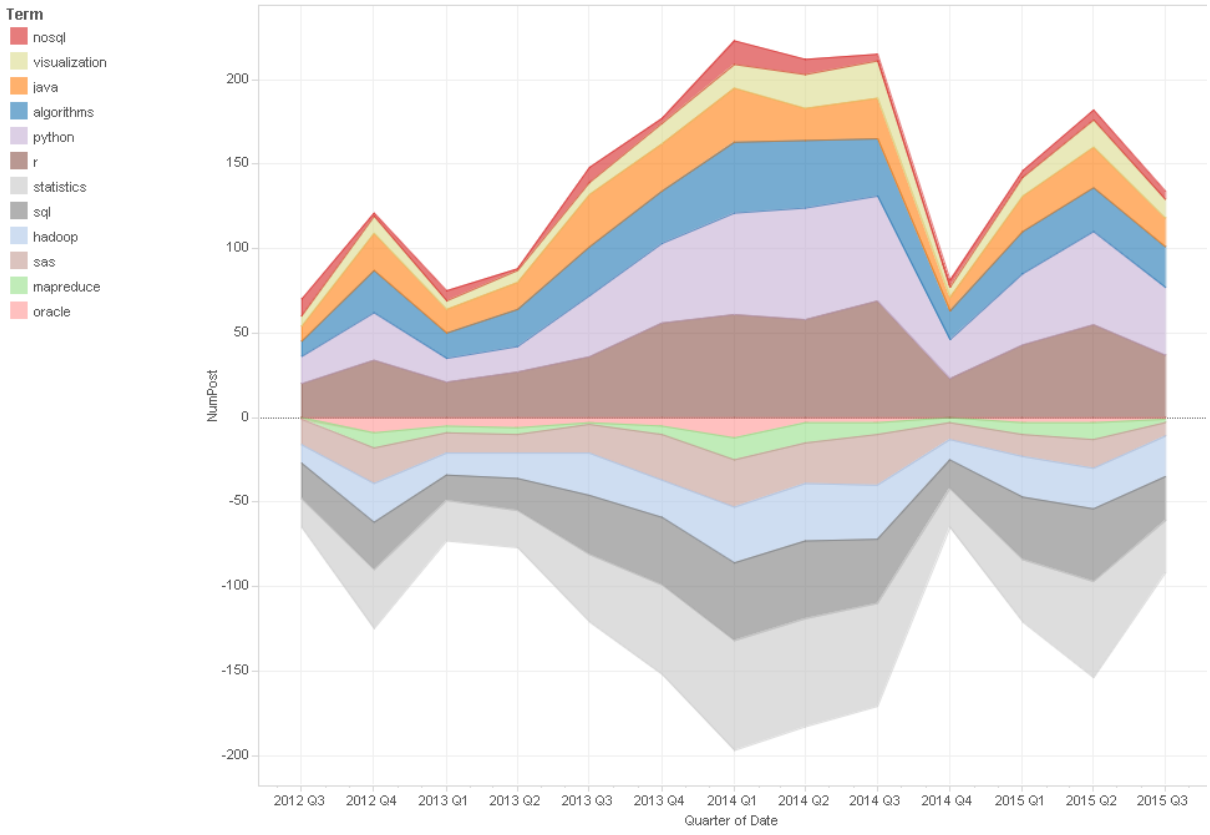
The data supports our claim that companies are not using well-defined roles in their job descriptions. The next question is, "How relevant are the skill sets currently required by hiring managers?"



The Spark vs. Hadoop plot is very telling – it demonstrates how quickly specific skill set requirements are evolving. Even if we find a purple unicorn today, six months from now we need a green unicorn, or maybe even a Pegasus or a dragon. This leads to a different problem: Data Science encompasses many tools and skills that do different things, and companies are seeking different collections of those skills based on both

current needs and projected future needs. However, people often have unique combinations of skills rather than fitting cleanly into these job descriptions. For example, if you need somebody to build a production recommendation engine, you may want two staff members: a data scientist to analyze the data, find the segments in users and products, and create some hybrid of content-based filtering and collaborative filtering; and a data engineer to collect and process new data and provide recommendations to users in a consumable way. However, building a production recommendation engine requires seven or eight unique skills, and it would be a poor assumption to think that we could divide this evenly among two people. Each person has a set of skills that he or she has picked up from different roles. Strategic candidates will have attempted to fit the unicorn persona and expand their learning to disparate areas; however, they may have some of the analytics skills needed and some of the engineering abilities. Finding a second person to exactly fit the remaining skills gaps may require another data scientist just to work on the optimization problem.

Here is a better view of how these skills have changed over time.



Every software engineer, no matter how good, was originally trained in how to write code. The same goes for statistics; it was taught. Technical skill sets can be taught; they don't need to be already mastered in every new hire. As new tools are required, we should acknowledge that we can train people to use them. We should also realize that we need a team of people to work on complex problems rather than one person who knows everything.

Sherlock Holmes claimed the mind has a finite capacity for information storage, and learning useless things reduces one's ability to learn useful things. When he learned that the earth revolved around the sun, he immediately set out to try to forget it since it was not relevant to his work. The skill sets companies seek are far from useless, but the principle remains: no one person can possibly do it all.

Much like your stock portfolio, you have to diversify. Today, the ideal Data Scientist is one who has experience with a little bit of everything in an IT space that is still evolving. They are a purple unicorn, born from a Pegasus mother and a Minotaur father. It's possible that Data Science will become increasingly specialized, much like software engineering, but for now, diversification is the name of the game.

If you hire a diverse team of excellent people, they can develop the skills that are valuable to your business. Otherwise, you may spend some time forcing them to forget some things. So, if technical skill sets shouldn't be the primary requirement, what should be? The answer is simple: as a company seeking Data Scientists, you should look for skills that cannot be taught. Qualities like curiosity, tenacity, skepticism, and empathy will get you farther as a team than expertise in Hadoop (especially if the candidate lacks those soft skills). If you're curious about the research surrounding the importance of soft skills in Data Science, the links below will allow you to dig deeper.
Curiosity 1, 2, 3
Skepticism 4, 5
Tenacity 6, 7
Empathy 8, 9

About the Author

Kenny Darrell is a Lead Data Scientist at Elder Research, the US's largest and oldest data science consultancy, where he leads projects primarily for federal government clients. He enjoys all aspects of data science; from problem definition and model construction to presenting the results in data products. He tries to keep a balance between hacking code and power points, and is a fan of learning new things and trying to do old things in new ways. Previously, Kenny was a Control Systems Engineer for the Air Force Research Laboratory and CDI Corp working on image recognition, rare event detection and sensor data fusion.

Mr. Darrell earned a BS in Aerospace Engineering and a MS in Quantitative Analysis from the University of Cincinnati, where his research focused on ensemble methods — combining data mining algorithms to increase performance. Outside of work he enjoys cooking, weight lifting and adding new items to his ever growing list of things he gets very excited about for a few weeks (flying, skydiving, boating, mountain biking …). He is grateful to his wife, an artist, for balancing his extremely rational approach to life.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING