Alison Yao (yy2564)

Dr. Andrew West and Dr. Ilana Goldberg

NODEP-UA 9982.SY1 Experiential Learning Seminar

21 April 2021

**A Comparison of Data Scientist Roles:  
Job Responsibilities and Business Impact in Enterprises vs. Startups**

1. **Introduction**

In the age of data explosion, more and more companies are relying on data for decision making and acquiring Data Scientists to facilitate business growth. Data Scientists, therefore, have been become one of the newer job inventions that almost every industry is yearning for. *Harvard Business Review* even described Data Scientists as the “sexiest job of the 21st century” [HBR].

Given the low supply and the increasing demand for Data Science talents, many universities and bootcamps started to provide Data Science programs for college students and young professionals. I, as one of them, also dream of becoming a Data Scientist in the future. However, as I learn more about Data Scientist as a profession and gain more experience working in the Data Science field, my glorified expectations have been challenged by reality. For example, instead of building machine learning models that predict risks and growth, I could be spending hours on data cleaning and data organization. Therefore, questions arise, such as, what exactly do Data Scientists do differently in enterprises and startups? More specifically, what are the differences in their job obligations and how significant is their business impact?

The project aims to answer 3 main questions through the means of literature review, digital ethnography and interview. The first question will present the duties of a Data Scientist in different contexts; the second question will dig into the reasons behind those differences; the third question make some recommendations for Data Scientist in terms of better preparing for these differences.

* How do the job responsibilities and business impact differ for Data Scientists in enterprises versus startups?
* What causes these differences to exist across enterprises and startups?
* Given these differences, how can Data Scientists better prepare themselves for success?

This paper aims to demystify the Data Science fiction in the media and gain a clearer perspective on my future career choices. The research outcome is designed to benefit a broader audience as well. Data Science students, for one, are typically wondering the same questions as I am. As they break into the Data Science field, they are also facing the choice of working for startups or enterprises. It is imperative that we understand what we are dedicating our time and efforts to because many junior Data Scientists end up leaving their jobs when their expectations of the job are not met. Moreover, Data Scientists sometimes become really frustrated when facing the mismatch between their expectation and the reality. This project will potentially help them better understand why discrepancy exists and therefore providing a new perspective on their career.

1. **Literature Review**

Before any analysis, we need to first define Data Science and Data Scientist. Although the term Data Science first existed back in 1960, it was used as a synonym of Computer Science. Later in 1996, the term officially appeared in public writing and gradually evolved to become the definition today. Data Science refers to the interdisciplinary field that utilizes mathematical methods, statistics, Computer Science and quantitative analysis to extract insights and knowledge from large volume of structured and unstructured data. In most scenario and in the business context especially, the purpose of Data Science is data-driven decision making through accurate data analysis. Data Scientists, therefore, are the people who identity real problems based on available data and access, manage and analyze data to create added value. [1] Data Scientists are generally required to be good with Math, Statistics, programming and a business domain of choice. They have been around in startups and enterprises years before the term Data Scientist was even coined in 2008, but the title itself has only started to drawn attention in the past decade. [HBR] Data Engineers are different from Data Scientists in that Data Engineer focuses more on building data infrastructure and architecture for data generation; Data Analysts are different because they do not deal with complicated model building process like Data Scientists do.

When researching the differences between Data Scientist roles in enterprises vs. startups, I narrowed down to two aspects: job responsibilities and business impact. The reason why business impact is such an important aspect for Data Science is because the final goal of Data Science is “improving decision making” and exploiting “the interaction between data science and business strategy” [qtd. 2]. It is imperative to keep the business goal in mind in further analysis.

In the following analysis, I will employ the concept of structure and agency. Structure and agency can help us understand human behavior and human identity. According to Dutta, “structure refers to the recurrent patterned arrangements that influence or limit the choices and opportunities available to people” and “agency is an inborn urge in every person to act independently and to make choices of their own free will” (qtd. Liwane, Rossouw) [4]. In this context, enterprises symbolize structure and startups symbolize agency.

1. **Methodology**

This research project is framed as an exploratory project aiming to answer research questions via data collection from interpretive designs such as interview (case research) and digital ethnography. Qualitative research methods are conducted because of the insufficiency of prior theories and a lack of structured quantitative data. [6] However, I included a small portion of quantitative research both as an example of a Data Science project to illustrate key findings and meaningful analysis to reinforce key ideas.

I aim to conduct an in-depth interview, which is “typically more open-ended and more flexible than constructed or semi-constructed interviews while being more focused on than natural conversations” [3]. Due to COVID-19, my internship is remote, so I conducted an online interview of my supervisor Andrii Galan, the one and only data scientist in the company. The interview lasted for 47 minutes, including the introduction and the ending. It was video-taped and audio-recorded. I have also contacted Senior Data Scientist Gang Su from Netflix through LinkedIn asking for an interview. He has read my message but never replied.

Digital Ethnography utilizes the Internet as both a “field site” and a “research tool”, meaning that the Internet serves both as a platform of “participant observation” and “a means of data collection” [3]. I collected data from YouTube, Medium and Quora and analyze Data Scientists who are active online through videos, blog posts and online discussion. YouTube videos include Ken Jee, Ian BlumenFeld, Jeevan Mokkala, Shae Wang and Katie Bauer while the blogs are written by Randy Au and Eryk Lewinson.

I collected data from 1 online interview conducted by myself, 5 YouTube videos (4 of which are interviews conducted by Springboard interviewers), 2 blog posts and Quora. Here are the work experience of the interviewees and bloggers:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Andrii Galan | Ken Jee | Ian Blumenfeld | Jeevan Mokkala | Shae Wang | Katie Bauer | Randy Au | Eryk Lewinson |
| Startup /  Smaller Companies | √ | √ | √ |  |  |  | √ | √ |
| Enterprise /  Larger Companies | √ | √ |  | √ | √ | √ |  |  |

My sample includes 8 Data Scientist in total, 2 of which has worked in both startups and the rest splits equal between startup only and enterprise only. The definition of startups and enterprises can be a little blurry here because I use them interchangeably with small companies and large companies. Further reasons will be discussed below. These Data Scientists come from a wide range of education background. Some went through a formal education in Data Science while others come from Physics, Computer Science and even Social Science and Business. All interviewees appear comfortable during their interviews and their responses are informative and well-constructed.

1. **Qualitative Analysis**

After going through all the data multiple times, I find that the major differences of Data Scientists roles in different companies concerns job responsibilities and business impact. First of all, Data Scientists in smaller startups are typically generalists who deal with much more than just Data Science, while Data Scientists in bigger enterprises are quintessentially specialists who may only focus on one part of Data Science. Ken Jee, a Data Scientist (Manager) who has worked for both startups and fortune 100 organizations, summarize his interpretation of the differences as follows:

“(00:35) In smaller startups, usually data scientists are responsible for a lot more breadth in terms of activities. So, you're maybe not going to be doing data science all the time. You could be helping with even marketing strategy or designing the infrastructure of the databases. There's a lot of things you can do a lot of ways that you can pitch in. […] (01:17) In large organizations, big corporations, the role of the data scientist is generally a lot more defined you're probably going to work in one fairly specific area with fairly specific constraints.” [Ken]

Eryk Lewinson, a Data Scientist formerly at a startup, summarizes the same idea as the “width vs. depth trade-off”, which is reflected through many other interviews and blogs. Randy Au, a former Data Scientist with 12 years of startup experience, emphasizes his work of building databases and data infrastructure before doing Data Science. Data Scientist in startups can mean a mixture of Data Science and Data Engineering with a splash of everything else. On the other hand, Jeevan Makkala from Facebook worked on the “Auction & Delivery” team and focused solely on ads optimization algorithm, a portion of model building. Shae Wang from Uber worked with increasing the bike and scooter conversion rate, again a small portion of model building. Senior level Data Scientist like Katie Bauer from Reddit and Ian Blumenfeld, focus more on Data Science project management. Therefore, Data Scientists at enterprises typically funnel their energy to one specific part of Data Science, while Data Scientists at startups spread out their efforts on things other than Data Science.

Second, Data Scientists in startups tend to have a bigger business impact than those in enterprises. Data Science has the potential to have a big business impact if done correctly. Ken Jee mentions that he prefers working at startups so much more because he was “able to really make an impact” [Ken]. One reason behind that is probably because Data Scientists exists the role of a mere consultant and becomes a developer from time to time. According to Ian, being able to interact with software products by writing production code yields high-level impact [Ian]. When Data Scientists are more involved with more aspects of a product, they are more likely to make a bigger impact. On the other hand, Brian D'Alessandro, Director of Data Science at Zocdoc, says on Quora that “your direct marginal impact to the org will be much smaller” if Data Scientists work at enterprises because they are likely to “deal with more process and bureaucracy” [Quora]. It is unfortunate but very often that a Data Scientist’s model can be brutally disregarded when it has to be approved by higher-level management. Months or even years of hard work would go to waste, let alone business impact. As Eryk Lewinson puts it, what is rewarding for Data Scientist is to “see that their work is actually used (as opposed to some report buried in an avalanche of emails) and provide clear value-added to the business”. Linking to the social science concept of structure and agency, the rules and regulations in enterprises seem to hinder and limit what Data Scientists’ contribution to a certain degree. But Data Scientists in startups can practice their agency for freely without constraint. Therefore, compared with a corporate structure of enterprises, startups may cultivate a better environment where Data Scientists can be exposed to all parts of businesses and actually implement their creations.

Now that we have seen what Data Scientists say about the differences in their job responsibilities and business impact, the next question is to dig deeper and ask why. I believe there are two reasons. The first reason is the differences in team sizes due to company resources. Data Scientists are compelled to be generalists when the team size is small. For startups with few people, it is common for Data Scientists to wear multiple hats at the same time. Big enterprises, in contrast, can afford more people and a much bigger team. Jeevan from Facebook points out that he is “constantly working with other Data Scientists on the team to get their perspectives” and that it is fun to collaborate with others [Jeevan]. Ian, former Chief Data Scientist at Clover Health, a startup, says that he used to do everything by himself, but as his team grows, he gives others more opportunities to step up [Ian]. Andrii, in contrast, used to work in big Data Science teams where communication and presentations are more often, is now working alone and caring most of the responsibility for the company.

The second reason is the different stages of a Data Science lifecycle that a company is currently at cause the business goal to be different. In order to perform data analysis and data modeling, Data Science needs clean datasets. But to have clean datasets, companies need to have functioning data infrastructure that captures and documents data. There is a temporal procedure that cannot be tempered with in the different stages of a Data Science lifecycle. Randy Au calls it “a common trap” when fresh-graduated Data Scientists expect to use fancy models in their work but only to find out that there not enough data to work with.

Diagram

Description automatically generated

Randy Au notes that one cannot “expect each layer to be done before moving on to the next”. The lifecycle of a data-driven company starts from the bottom of the pyramid which guarantees a working product and data pipelines so that important data are automatically documented in a database. The business goal is simply capturing data. Then, data visualization and analysis reports are created for different departments in a company regarding performance update and decision making. Now, the business goal is accessing and some easy decision making. Once these reports are automated and more manpower is released, we can finally arrive on the top to the pyramid and use all the sexy models to do Data Science. The ultimate business goal is to derive data-driven insights that are too intricate for human eyes to see. Please note that any company can go back to a former stage and start over when deemed necessary.

Most of the big company Data Scientists never mentioned Data Engineering in their interview. Shae from Uber mentioned that IOT censors are already installed on bikes and scooters, so her responsibility is to use data to piece together a user story. In Jeevan’s interview, he mentioned a challenge for him is to determine “what data can be used”, which implies Facebook has too much data instead of too little. These tech giants are already very mature in terms of Data Science lifecycle. Most of the startup Data Scientists in my data like Randy Au and Eryk Lewinson, on the other hand, emphasize data engineering when explaining their responsibility outside of Data Science.

The two differences of team size and lifecycle stage happen to coincide in the comparison of startups versus enterprises, because typically, startups have smaller teams and is still in an early stage of building data infrastructure. As you can see from the graph below, most cases in this project follows the trend, but the cases of Andrii Galan and Katie Bauer are outliers.

Diagram

Description automatically generated

Therefore, although the differences in job responsibilities of Data Scientists in startups versus enterprises exist, whether or not the company is a startup or an enterprise does not 100% dictate what a Data Scientist does. My supervisor at Paypro Global, Andrii, has been working as a Data Scientist for 3 years. Paypro Global was founded in 2006 but the size of the company stays between 50-250 and Andrii is the only Data Scientist working full-time at the company. Established over a decade ago and thriving in ecommerce, Paypro has been collecting and organizing data for years. Therefore, Andrii is working on generating reports, automating reports and model buildings all at once. By the time I started my internship there, I have been able to use data from maintained SQL database and focus on Data Analysis and Data Science. The case of Katie Bauer is the exact opposite. She worked at Reddit from 2017 to 2020 and by 2017, Reddit had already been a well-known company established 10 years ago. But according to Katie, who was a founding Data Scientist at Reddit, their hive database was old, inconvenient and error-prone, so they “migrated all of [their] data into the Google Cloud System” and “started building [their] actual analytics infrastructure”. This Data Engineering-centric part of her work experience was renovating an earlier stage of the lifecycle today for better Data Analysis and Data Science tomorrow. Therefore, when digging into the reasons why Data Scientist roles differ in startups and enterprises, I find that the startups or enterprises is merely a superficial reason. The deeper causes are the differences in team sizes and stages of data lifecycle.

Building on the theory above, any Data Scientist can encounter all stages of a lifecycle in different teams. Despite the differences in job responsibilities and companies mentioned above, I have noticed much more similarities regarding the attitude, the motivation and the qualities among the Data Scientists I observed. The first interesting finding is that Data Scientists dislike dirty data, but they see them more as a potential opportunity rather than a total nightmare. Shae Wang speaks for everyone that “one thing that all Data Scientists hate the most is messy data that they have to clean up and somehow debug” [Shae]. Dirty data requires a lot of time and laborious efforts on data cleaning, which is the less sexy part of the job because it does not have any business impact pre-analyzed. But our optimistic Data Scientists see dirty data in a brighter light. Shae thinks of dirty data as “an opportunity to improve the platform and infra”. Andrii thinks that even though model building is fascinating, switching to other things for a while is refreshing and conducive to new ideas. These are two very good reasons to defend something that Data Scientists may not enjoy as much. The second finding, which is reassuring to me, is that the job as a Data Scientist is overall really enjoyable. 6 out of the 8 Data Scientists in my research have switched companies but they remain Data Scientists or they are promoted to senior-level Data Scientists. When they are talking about their job in the interview, they speak with pride and satisfaction. The conversation becomes especially engaging when they start to explain how to use Data Science in domains that they are really interested in. Jeevan’s face lights up when talking about basketball and sport analytics; Katie and her interviewer laugh when Katie talks about using Data Science to prove that people care more about cats than dogs. Through my research, you can see that Data Scientists, no matter where they work, are a bunch of patient, optimistic, adaptive and creative problem solvers, business-centric techies and lifelong learners who find challenges interesting and exciting. Those are everything I aspire to be.

1. **Discussion & Findings**

In retrospect, it was quite intuitive in the beginning to look for the differences of Data Scientist roles in the enterprises vs. startups. On the surface, there are indeed a lot of differences, especially in job responsibilities and business impact. However, as I investigate further into the reasons behind those differences, it became clear that whether a company is an enterprise or a startup is not the real cause. Rather, the size of Data Science team and the stage of Data Science lifecycle a company is at are the reasons behind those differences. Since any Data Scientist is likely to encounter different team sizes and lifecycle stages, it is imperative that they think positively about the less appealing part of their job and practice their nature as the problem solvers that they are.

1. **Conclusion**

I have fulfilled the goal that I started out with to my satisfaction, with some pleasant surprises on the way. I have answered my research questions both from a lower factual level and a higher causal level. I have a much clearer idea of what I can expect in a Data Science job and how I may impact business growth. More importantly, I am also constantly assessing my qualities and personalities to see if I might be an excellent Data Scientist in the future. And I think the answer is quite positive. Moreover, I now have a more defined plan regarding how to help my startup grow during the coming summer. Thanks to Firebase, we already have some basic analytics data pipelines setup, but our data infrastructure is far from complete. Any second before the infrastructure is finished is a second we let key customer data go to waste. Therefore, during the summer, as the only Data Scientist in our startup, I plan to start from the early stage of the Data Science lifecycle and build our automated data-capturing infrastructure. In the following year or two, I will slowly bring our company into a more mature stage of the Data Science lifecycle.

One pleasant surprise and also enlightenment when doing this research project is a whole new perspective on Data Science and Data Scientists. I never realized how tool-oriented my college education was. A Data Science undergraduate student, all 12 courses I need to graduate teach me how to use tools. The first ever major course is devoted entirely to teach Python and my Machine Learning class literally moved from one algorithm to another every week. With the shiny tools in hand, like many new Data Scientist, I am eager to use these tools and change the world, while the reality might be there is not even any data to start with. This is actually one of the motivations for this project. So not surprisingly, I panicked when I found out there was no data out there to quantitatively analyze my research questions and to use my little tools on. Therefore, in a way, I have no choice but to resort to qualitative methods and to convince myself to embrace subjectivity, which is unfamiliar and untrustworthy from a data perspective. But I got so much useful information from those interviews, blogs and online discussion. I saw through the texts and videos and looked directly at those Data Scientists in the eyes. And that would never have happened if I just looked at cold numbers and charts. When taking a completely different way to answering my research question, I realized that any tools and methods are secondary compared to the problem that you are really interested in. Using the fanciest tool to solve a wrong problem is worthless. And the senior Data Scientists in the interviews, Katie and Ian, echoed that thought. Instead of asking how I can accomplish it, they ask what I should accomplish. They spend more time thinking about what the problems are than actually solving them. Therefore, the biggest takeaway is that Data Scientists are still scientists and scientists are researchers and problem solvers. We are people who solve problems by using a large amount of data. A good Data Scientist should never be defined by the tools he or she uses.

Works Cited

[1]

[2]

[3]

[4] agency and strusture

[6] Bhattarcherjee's Social Science Research\_ Principles Methods and Practices