THE DATA ANALYTICS

HANDBOOK

DATA ANALYSTS + DATA SCIENTISTS

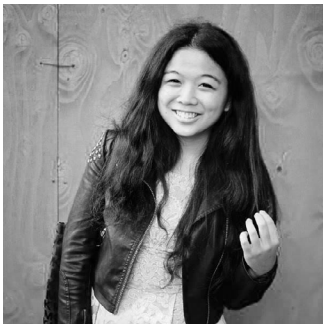
ABOUT THE AUTHORS

BRIAN LIOU Content

Brian graduated from Cal with simultaneous degrees in Business Administration at the Haas School of Business and Statistics with an emphasis in Computer Science. He previously worked in investment banking before he transitioned into Data Analytics at MightyHive, an advertising technology company backed by Andreessen Horowitz.

TRISTAN TAO Content

Tristan holds dual degrees in Computer Science and Statistics from UC Berkeley (‘14). He first began working as a quantitative technical data analyst at Starmine (Thomson Reuters). From there he worked as a software engineer at Splunk. He has experience working with various Machine Learning models, NLP, Hadoop/Hive, Storm, R, Python and Java.



ELIZABETH LIN Design

Elizabeth is in her third year studying Computer Science at Cal with a focus in design. She is President of Innovative Design, a student group of visual designers and photographers, and Layout Director for BARE Magazine, Cal’s fashion and arts publication. She has interned at LinkedIn, TechValidate, and UC Berkeley’s EECS Department. She will be a Product Design Intern at Khan Academy this summer.

What exactly do the sexy "Data Scientists" do? We start with this simple question. What other professions are there in Big Data? What tools do they use to accomplish their tasks? How can I enter the industry if I don't have a Ph.D. in Statistics? The genesis of "The Data Analytics Handbook" stems from our own internal frustrations with these questions; as recent graduates of UC Berkeley (Go Bears!) in statistics, we saw a burgeoning industry craving curious minds but disconnected from its potential employees because of a lack of understanding of what the Data Analytics industry is and of what it requires. And so, we set out to discover the answers for ourselves by reaching out to industry leaders, academics, and professionals.

This handbook is the first of three parts and will focus on the experiences of current data analysts and data scientists. What we discovered in our research is that while each interviewee's response was highly informative, their knowledge was fundamentally incomplete; the truth is that the Big Data industry is still so nascent that there is no singular definition of a data scientist or data analyst. In fact, the two terms are often used interchangeably.

While we continue to ponder the big questions surrounding Big Data, we believe that our research has shone light in an area unexplored, making our interviews that much more valuable to you. The firsthand experiences of those in the trenches, those leading the troops, and those crazy Ph.Ds, have not been shared until now. We hope you enjoy reading the interviews as much as we enjoyed conducting them!

From Tristan, Brian, & Elizabeth

TOP 5 TAKEAWAYS

FROM THE DATA ANALYSTS & DATA SCIENTISTS

1. Communication skills are underrated

If you can't present your analysis into digestible concepts for your CEO to understand, your analysis is only useful to yourself.

2. The biggest challenge for a data analyst isn't modeling, it's cleaning and collecting

Data analysts spend most of their time collecting and cleaning the data required for analysis. Answering questions like "where do you collect the data?", "how do you collect the data?", and "how should you clean the data?", require much more time than the actual analysis itself.

3. A Data Scientist is better at statistics than a software engineer and better at software engineering than a statistician

The greatest difference between a data scientist and a data analyst is the understanding of computer science and conducting analysis with data at scale. That being said, data scientists only need a basic competency in statistics and computer science. Not all data scientists are Ph.D.'s, and newly developed tools are empowering more and more people to be able to do data science.

4. The data industry is still nascent, if you want to work with a variety of stakeholders in a more freeform role, the time to do so is now.

Data scientists and data analysts all say they interact with a many parts of the company from engineering to business intelligence to product managers. The roles of data scientists and data analysts are largely undefined and vary by your own skill set and the company's needs.

5. Both roles require a curiosity about working with data, a quality more important than your technical abilities.

The ability to discover trends and patterns previously unseen is what truly makes you valuable. Having a curiosity enables you to ask the creative questions necessary for transcendent analysis. As practice, when given a dataset, ask yourself what questions do you have about the data and how would you answer them?

TABLE OF CONTENTS

ABRAHAM CABANGBANG
DATA SCIENTIST AT LINKEDIN

ABRAHAM graduated from Stanford University with a Bachelors and Masters in Chemical Engineering. He is currently part of a decision sciences/insights team focusing on data quality and cross product analysis at LinkedIn

LINKEDIN connects the world's professionals to make them more productive and successful

What is your background and how did you end up at LinkedIn doing Data Science?

I graduated with a B.S. and M.S. in Chemical Engineering in 2011. After that I started working for a digital advertising platform as a data analyst and about year ago started at LinkedIn.

So how did you develop the skills to become a data scientist at LinkedIn if you studied chemical engineering?

I learned some statistics through my degree's coursework, but most of it was from working as a data analyst before I began at LinkedIn. I picked up SQL pretty quickly at my previous job and used a fair amount of R. At LinkedIn I have learned Pig and Hadoop.

How would you define the role of a data analyst and how would you define the role of a data scientist?

It's definitely a gray area. At my previous company I did both analyst and scientist jobs and as an analyst we were more customer facing; the tasks we did were directly related to the tangible business needs—what the customers wanted/requested. It was very directed. The scientist role is a little more free form. The first thing I did as a data scientist is work on building out internal dashboards, basically surfacing information that we were tracking on the back end, but weren't being used by the data analysts for any reasons; for example, we might have lacked the infrastructure to display it, or the data was just not very well processed. It really wasn't anything tailored out from a customer need, but came from what I noticed the analyst team needed in order to do their job.

Going into your role at LinkedIn, what are aspects of the job you and enjoy and what are aspects that you find frustrating?

One thing I enjoy is that the team at LinkedIn is so big that I have to work with a lot of different teams to get things done. My work involves working with PMs (product managers), the BI team, data services, and other data scientists so it's a very open role where you do a lot of different things. As for something I don't like, there isn't a major thing that comes to mind.

What are some of the characteristics that differentiate data scientists from each other to be more effective?

I don't think a data scientist at LinkedIn doesn't mean one thing, we all have a baseline set of skills, and apart from that it's pretty project specific. Some data scientists specialize in machine learning, others specialize in visualizations, but they are all effective data scientists.

How important do you think having a background in Statistics is?

It's important, but it really depends on the project you're working on. You can be a great data scientist without being an expert statistician. The skills you bring to the table kind of define the projects you do at LinkedIn.

What are the tools you use in your job?

I use SQL, Pig, and Hive regularly. I use R for visualizations or predictive modeling. I have used Micro Strategy and Tableau as well.

Can you describe a typical day and a typical project you might work on?

Since I work on a team that focuses on reporting and data quality, if there is a new product out there we might want to incorporate that into one of our major dashboards so that will involve working first with product mangers to work out what is important to the product, engineers to make sure the relevant data is being tracked, and then working with our data services team to do ETLs (extract, transform, and load) and visualizations.

Could you breakdown the percentages of how much time you spend coding, meetings, etc.?

A third to a half of my time coding, a quarter to a third meeting with the various parties, and the rest of the time split between data visualizations and other tasks.

What advice can you give to aspiring young professionals trying to enter the field?

I would start thinking about what questions you would ask about data and also what data would you want after using certain products. That would be a good way to start. A lot of the technical stuff can be learned. What is important is knowing how to ask the right questions and knowing how to go about answering them.

JOSH WILLS

SENIOR DIRECTOR OF DATA SCIENCE AT CLOUDERA

JOSH holds a degree in Mathematics from Duke University as well as an M.S.E. in Operations Research from UT Austin. Before coming to Cloudera, he worked as a software engineer as well as a statistician at companies including Google, Zilliant, and IBM. He is also the founder of Apache Crunch.

CLOUDERA is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop™

Can you describe your background?

I studied math at Duke University. During most of college, I thought that I would become a math professor of some sort. I took a couple of probability courses, but didn't take any statistics until my last semester at Duke, but I ended up falling in love with it. I minored in philosophy, and statistics felt like quantitative epistemology; given some data, what conclusions can we draw? What can we say that we know?

On the computer science side, I spent a couple of summers at Carnegie Mellon building brain simulations in MATLAB. After college, I went to IBM in Austin, Texas and worked on low-level system processing in C++. I got pretty bored, so I started working on a master's in operations research at UT Austin, and then I went on to work at a couple of different startups- Zilliant, OneSpot, and Indeed- sometimes as a data analyst and sometimes as a software engineer. During my brief stint at Indeed, I became really interested in ad auctions and started reading all of this auction theory stuff, but I felt like I didn't really know what I was doing, and I wanted to learn how auctions worked in the real world, which led me to Google.

I was hired at Google as a statistician, but as soon as I got in the door, all I did was write code and run experiments on the core auction module. After nine months of that, I made a lateral move over to Google's software engineering ladder, and ended up working on things like logging systems, recommendation engines, and Google's multivariate experiment framework. After that, I headed to Cloudera to become their first Director of Data Science, the first time I had "data science" as part of my job title. In summary, I moved between data analyst and software engineering roles during my entire career. I'd say that I was a data scientist the entire time, there just wasn't a term for it yet.

You indicate that data science does not have a structured way of learning, for example you did not have a PhD in statistics. How did you learn the formal skills required to create defensible statistical analyses?

Everything boils down to understanding the fundamentals. In addition to the core of statistics, reasoning probabilistically and understanding conditional probability are really important. I think the same thing is true for computer science; understanding data structures and computational complexity is fundamental. Once you understand them it makes moving in a new direction

or exploring a new problem relatively easy. It is a little bit trivial to say it, but if you have a good understanding of the fundamentals, then that's all you really need.

What is your definition of a data scientist?

"Someone better at statistics than any software engineer, and someone better at software engineering than any statistician." The vast majority of statisticians can write code, or rather they think they can write code in SAS, or R, or Python. However, it is usually really bad code that is only intended to be used by the person who wrote it. I think that what makes you a good software engineer is the ability to write code for other people. To be good at coding you must understand how other people are going to use it. Statisticians don't always have that inherent skill.

On the other hand, software engineers conceptually understand models such as linear regression, but they don't really understand the underlying assumptions of statistical modeling. They think that they are doing data analysis and finding insights, but what they're finding is nonsense, or at least not what they think they're finding. It's easy to make this mistake and think you're doing analysis by just applying statistical models to data without understanding the assumptions.

The funny thing is, people who don't know anything about statistics or data analysis typically do the right thing- they look at the data. Novices will put data into a spreadsheet, look at it, and make some basic plots and charts. People who learn about t-tests and regression get a little sophisticated, and start applying models without looking at the data. They think that they can just apply this technique and they don't have to look at the data anymore. You do that a few times, you get burned by some bad data, and you learn to go back and start looking at the data again.

How would you define the role of a data analyst?

When I was a data analyst, I spent most of my time working in Excel and R. The volume and structure of the data was not too extreme or unusual, so whether the algorithm I was running was linear or O(N^2) did not matter. I think the transition into a data scientist happens when the data cleansing process or the volume of the data becomes so extreme that you need to worry about the computational complexity of your algorithms in order to get an answer in a

reasonable time frame. So the line between data analyst and data scientist is when an understanding of computer science makes you much, much better at your job.

Per your answer, I argue that there are companies such as BigML that are creating a layer of abstraction, so that analysts don't need to understand the software engineering portion of the job. Do you think the emergence of such tools will move data analysts into data scientists?

Great question. I think, that as servers get bigger and we gain access to cloud systems like BigML, the size of problems that data analysts can tackle will continue to increase. That being said, the richness, the volume, and the complexity of the data that we are collecting is also increasing at a healthy rate. So, I don't really worry about the future of data scientists or data analysts; there is plenty of data and plenty of problems for everybody to work on.

You mentioned the shortage of data analysts and data scientists. Where do you think the next generation of these people will come from, given that academia is not fully preparing graduates?

It's a great question, and I wish I had a simple answer for you. If we go back to our definition of data scientists, we can talk about the two key skill shortages. Obviously, we already have a shortage of software engineers. But beyond that, we also have a shortage of people who are statistically literate. This is true across the board, from managers to analysts to software engineers. What makes the data scientist shortage so acute is that it exists at the intersection of both of these shortages.

I'm inclined to believe that there is a contingent of software engineers like myself, who studied math in college, and who will take on more and more data science positions. I can't solve the problem of creating more software engineers. However, I would like to work on solving the problem by making people more statistically literate. I've toyed on and off about putting together a short course for software engineers on how to think like a statistician.

With the emergence of Big Data technologies, it seems like everyday someone else is coming out with a new tool. How important would you say it is to be up to date with all of these tools in order to work within the industry?

Again, I would bring it back around to the fundamentals. The thing to understand about "Big Data stuff," is that it's driven by economics. Specifically, it is the fact that disk has gotten insanely cheap. The cost vs. performance of the disk is falling faster than Moore's Law. All computer systems have a mix of resources, primarily network bandwidth, CPU, memory, and disk. For a long time, every one of these resources was scarce, and so we designed systems to use each of them as carefully as possible. Now we are in a mode where disk is really cheap, CPUs are great, memory is getting cheaper all the time, and network bandwidth is the primary bottleneck for almost everything interesting that we want to do. When you talk about new Big Data technologies, it's important to point out that the ideas behind things like MapReduce and Spark are not new. It is just that the economics of storage have shifted so that these architectures make economic sense in a way that they really didn't before.

I think that over time, we will teach MapReduce in the same way that we teach bubble sort. It's a very simple way to think about distributed computing that is elementary and accessible to everyone. If you can think in MapReduce, it prepares you to be able to think in Spark. Spark is a faster, more optimized model, but it's still fundamentally similar to MapReduce. Ultimately, we really want people to develop the MapReduce-style of thinking about computations.

The only way that we won't need MapReduce-style thinking is if we make some kind of advancement in networking technology that would enable huge clusters of machines to act as a single machine. When and if that happens, it will blow away all of this stuff. If having a machine in LA and communicating with another machine in New York was virtually instantaneous, then there is no need for a distinction between a data analyst and a data scientist. As long as that gap exists, we will need people with the ability to think in computer science terms when approaching analytical problems.

BEN BREGMAN

PRODUCT ANALYST AT FACEBOOK

BEN graduated in 2011 from Stanford with degrees in math and physics. After graduation, Ben spent time traveling and working at a trading firm before joining Facebook in 2013. He currently works as a Product Analyst at Facebook.

FACEBOOK'S mission is to give people the power to share and make the world more open and connected.

What is your role as a data analyst at Facebook?

I use data to help drive product development. At Facebook, product teams are generally composed of some mix of product managers, designers, engineers, and data analysts. The analyst's role is to bring a data-driven perspective to the conversation. This might include understanding who is using our product, what value users are deriving from our product, and where we should go next. We then work directly with the PMs, designers, and engineers to make sure the team is prioritizing and moving as effectively as possible to hit those goals.

What are the tools you use most at Facebook?

One philosophy at Facebook is that if you need something, then build it! To that end, members of the analytics team have created some awesome tools to facilitate analysis of our data, including real-time monitoring tools and fast analysis tools for easy deep dives into data. We use Python to write scheduled analysis scripts, HiveQL to query the data itself, and R for statistical analysis. Finally, we have a few in-house visualization tools that we use to effectively aggregate and disseminate data information through the org.

How would you differentiate your position from a data scientist?

From what I've seen, the mapping of title to job function differs company by company and team by team. In general, I perform 3 functions: 1) setting up and debugging logging infrastructure in our core code bases, 2) setting up downstream processes for analysis of the data, and 3) using the results to build a story and sync up with the rest of our team on the state and future of our product. The first two functions probably line up most closely with the typical image of a "data scientist," while the third function speaks more directly to the product analyst's role as a key member of the product development team.

Do you see applications for your job in industries other than technology? Which ones?

The process of using large data sets to understand the value that people are deriving from our product lends itself most naturally to the technology industry. At the end of the day, this kind of information is useful to any company that wants to get feedback from its users about where to grow and develop. For example, that is why Facebook has developed tools like Page Insights, which allows business owners to understand who is visiting and interacting with their page on Facebook.

Could you describe a typical day and a typical project you might work on?

My typical day will vary depending on where we are in the lifecycle of a product release. If we are actively rolling out a new feature, I will be monitoring and digging into metrics to understand where we are under/outperforming. If we are developing a new feature, I will be working with engineers to ensure that our logging is up to par and communicates as expected with any backend services involved in the feature. If we are brainstorming future direction for a product, I will be pulling data and performing analyses that help inform the conversation. It's awesome being involved in the product lifecycle from beginning to end, and great to see when users are really enjoying and benefiting from a new feature.

How were you interviewed for the position? Programming questions? Statistics questions?

I went through about 7 or 8 interviews end-to-end for this position. Technical aspects of the interviews included basic coding (data structure-type questions, any language), technical data analysis (how to extract and handle information from large data sets, included SQL), and general analytics (what questions to ask to learn about a product, some quick math during examples). The interviews also contained a strong focus on cultural fit, both for Facebook and for the Analytics org at Facebook.

LEON RUDYAK

BUSINESS INSIGHT ANALYST AT YELP

LEON is part of Yelp's Business Insight team involved with driving strategic improvements in product, revenue operations and sales. He graduated from UC San Diego with a degree in Economics.

YELP operates an "online urban guide" and business review site.

Can you talk about your background; how did you end up working at Yelp and how did you end up specifically in data analytics?

I graduated as an Economics major in 2010, which was one of the worst years to graduate given the economy. I started working at an executive compensation research company where I went through a couple of different roles. My work involved using SQL and sorting through large sets of financial data, which I really enjoyed. However, I really wanted to apply my skills to the consumer internet space and Yelp was a great fit because I loved the product and they were looking for analysts comfortable with SQL and working with large data sets.

How long have you been with Yelp?

1 year and 3 months.

Can you talk about the aspects of the job you enjoy and aspects that are frustrating?

I enjoy the fact that my team gets insight into all aspects of Yelp's business. At larger companies with large analytics teams, individual analysts usually only get exposed to specific parts of the business and don't really get to see the big picture. Yelp is a public company, but our BI team is only 5 people so we get to work on projects in product, sales, marketing and etc. A frustrating part of my job would be monthly reporting, which usually involves manual and repetitive work. We have to get through the reporting quickly to get to work on the cooler projects.

Can you talk about a one specific project that is representative of your overall job role and also something you enjoyed at Yelp?

One of my favorite projects was analyzing how positive feedback on the site (likes or comments) influences user contribution on Yelp via writing reviews or posting photos. We did some testing with the community managers (people who interact with elite Yelp users) to see how their involvement affected user contributions. The analysis was very impactful and resulted in changes to both community operations and product.

What kind of tools do you utilize to accomplish a project like the one you mentioned? Are you still heavily reliant on SQL? Do you

have other tools that you've been using?

A big part of my job is tying together data from the different sources. Our main sources of information are the SQL database, Salesforce and some internal company tools and documents. In the project I mentioned earlier, I used SQL to get the raw data of user activity that I then synthesized in Excel to create charts and graphs.

How much Statistics do you use in your job? Do you think more statistical knowledge will help you in your work?

It is not necessary for what I do, but it definitely helps. We've done a couple of projects in R that involved linear regressions but it's not something we do frequently. Mainly a background in statistics gives you a better feel for the data and helps you determine the validity of the results.

Can you talk about the workflow at Yelp, and how your role fits in?

My team reports to VP of Strategy and Operations. He presents us with the main problem we're trying to solve for or some part of the business we are trying to explore more deeply. Then my manager provides the outline for what we need to do to solve the problem and the kind of outputs we need to show. My role is to pull and clean the data, create the analysis with expected outputs and transform the analysis into a digestible PowerPoint. Then we report back to our VP, who gives us feedback. We iterate if we need to.

What are some of the softer skills you need as a data analyst?

Communication is important. I know the term "effective communication" gets thrown around a lot, but I do think it's very important for a data analyst. You need to be able to clearly communicate the methodology and results of your analysis to someone who isn't involved with the data. When communicating over emails, you have to be very specific in the question you're asking. Not communicating properly can easily start a wave of confusion.

Do you have any advice for undergrads/young professionals entering the field? What advice would you give to someone looking to get a similar job as you?

I think new grads focus too much on what title they need to get instead of what experience will be the most valuable. In my opinion you should spend the early years of your career maximizing your learning opportunities and

trying to surround yourself with people who are smarter than you. I would also advise someone looking for an analyst position to get a "hard skill" (SQL, Excel modeling, Python, R). Everyone applying for analyst jobs has college degrees, but having a technical skill allows you to differentiate yourself tremendously.

How are you specifically interviewed at Yelp?

I had two phone interviews with the recruiter and my manager to gauge my interest in the position and see how much I knew about the company. I then had two onsite interviews with multiple people from my team, finance and engineering that involved Excel, SQL and critical thinking business problems. I also had a take home Excel project.

PETER HARRINGTON

CHIEF DATA SCIENTIST AT HG DATA

PETER is the author of Machine Learning in Action, a best selling book on the most important machine learning algorithms. He holds both a Bachelors of Science and a Masters degree in electrical engineering.

HG DATA indexes, extracts, and renders unstructured data to improve B2B technology sales and marketing programs.

How would you define the role of a data analyst? How would you define the role of a data scientist?

The short answer is that a data analyst doesn't know how to code. A data analyst doesn't know how to code but instead is expected to be proficient in industry tools such as Excel or if you say work in Finance a Bloomberg terminal. A data scientist definitely has a much higher understanding of computer science and is expected to develop tools on their own or put to use some non-standard tools for the products needs or the company's needs.

In terms of skills what skills do you think a data analyst needs, do you think a skill like R will become the new baseline skill instead of Excel?

Yes I think so and I think it's more because the tools are becoming easier to use than people are becoming smarter. We had a conversation the other day about this; the typical programmer ten years ago was a very nerdy person. A programmer today isn't so stereotypically nerdy anymore. Now is this because programming has gotten cooler? Or because the tools have gotten easier and we think its because the tools have gotten easier. The greater accessibility of these tools will inevitably allow a larger audience to participate.

In response to that, do you think the bar to become a data scientist is getting lower so that professionals without advanced degrees in Statistics or Computer Science can become one?

Companies right now ask for Ph.D.'s because up to now these techniques for applying machine learning and data mining in industry haven't been well defined, but I think a lot of the techniques are becoming more standard and accessible for public use.

How does your company use data analytics as a comparative advantage?

HG Data does a lot of text analysis. We do a lot of modeling that is similar to how insurance providers use data analytics, an example for us would be: if this company does XYZ what is the likelihood that they will switch CRM (Customer Resource Management) providers. We index and aggregate that data so that enterprise sales professionals can be more efficient in their work.

How would you describe the data science work that you do?

We do some predictive analytics but we have found that a lot of the customers want to do the prediction on their own. We focus on natural language processing work. The predictive models we make have got to be super transparent and in most cases our clients have their own information they want to use. The biggest barrier we have found is when we create predictive models the client wants to understand under the hood what's going on and without the expertise we have, they can't understand it and thereby don't trust it.

Could you describe a typical day or project in your role?

A typical project would be if we have found a new source of data but it's not in the form that can be stored in our database. So we work to transform the data into the form we need. A student may think, "Oh, well you just have to reformat it." But it's not that easy because there are nondeterministic things that need to be done and needs to be done with high accuracy. Since we are a startup I actually probably spend 60% of my time coding, 5% I'm looking at results, and 35% I am researching new ways to fill in the gaps in my analysis.

How would you describe the value you and other data scientists bring to companies?

The value is making data driven decisions. It varies by industry how accurate or data based these decisions are going to be but in most cases it determines whether your company is profitable or un-profitable. I think what we do at HG Data will be the future for how all companies will operate and we already see sign of companies moving slowly towards that.

JOHN YEUNG

DATA ANALYST AT FLURRY

JOHN is a Data Analyst at Flurry. He previously worked at Trial Pay and was a Surgical Research Associate at Stanford University

FLURRY is a market-leading analytics software for smartphone and tablet apps consumer behavior.

Can you talk about who you are, and how you ended up in the field of data analytics?

I studied Philosophy and Parasitology as an undergraduate. I more stumbled upon the field of data science. I was admitted to medical school, but I didn't know if I wanted to commit to the schooling required to become a doctor. I eventually went into a business development ad operations role for another company. That's where issues with Big Data came up, data that you couldn't open up in Excel. I couldn't do my job without being able to aggregate and crunch these numbers together quickly and it would take Excel hours. That's also when SQL became the modern day Excel tool for analysts. From there I hopped onto Flurry during the mobile boom and I got a lot of exposure building predictive models in the advertising space.

So did you self-teach the programming and statistical knowledge you use now?

Yes, I had to learn R and SQL and you sort of self teach with online resources followed by professional training. I'd always recommend getting professional training to really solidify your foundation.

Have you considered doing a Masters program?

Yes, definitely. But right now the Mobile Analytics/Ad space is moving so fast that I don't want to take a hiatus from work just yet.

Can you talk about the aspect of the job you enjoy, and the aspect of the job that are frustrating?

From a high level, I enjoy the problem-solving aspect of this field, and this actually goes into the frustrating part as well. Big Data and the technologies resulting from it are relatively recent. For the business side of things, the marketing people and acquisition managers or consultants, for them to extract meaningful data and insights means you have to have infrastructure in place to get this reporting. What Flurry does is provide analytics for developers so they can understand their users' usage behavior better. Developers will include the Flurry SDK in their app, which in turn allows them to slice and dice the data of their users through our dashboard. Now extrapolate that across 450,000 apps, across 1.3 billion devices that we track on a monthly level. Processing that and deriving insights from that is what I do. The data is in petabytes. So Excel is useless here. To get answers to a simple question such as how many people

actually went and made a purchase in your app might take hours or days to answer if you don't have the proper infrastructure. To me the result of this has redefined the data analyst role and the impact they can have in a company.

What about the job of a data analyst most interests you? Is it the technical challenge or the non-technical challenges?

Personally, I think that the business insight aspect of the job is something I lean towards. It is more directly tied to revenue. Some advice I give to undergraduates is that it's always good to be as closely tied to revenue as possible. Engineers will always be able to build and code better than you, but as analysts, being able to interpret the data and make recommendations on strategy is where you can contribute the most. Today, it's difficult to really understand the business side of things without a good technical background and understanding of metrics and data.

Do you use any other tools other than R/SQL?

That depends on whom you're talking to in the company. We're one of the biggest data companies; the amount of tools we can use is actually limited. The data is too big; therefore it's difficult to plug in BI tools like Tableau. I still use Excel, because some things work well in Excel. I will use Python/Linux to splice raw data, remove os and clean. Then I load the data into SQL. From there, I do another layer of massaging and move to Excel or I do my calculations in SQL. It ends with making a power point slide deck so you can convey your ideas and discoveries. Some people are really good at data pulling/data mining, but can't communicate the esoteric and crazy analysis to the CEO or the sales team. This is what separates the really good analysts from the rest.

Can you describe a typical day while working on a typical project?

Some of the interesting projects I've worked on include some of the largest gaming companies; we do one-off consulting projects for them. Generally a gaming company will have a portfolio of games, and they're always looking to expand the user base or find there the industry is heading. So a lot of times they will turn to Flurry to get a sense of where the market is heading towards. One example is when there are different companies where they own different games, but the genre is rather concentrated. Now, if they want to acquire more users, they have to decide what investment will get them the best ROI. So, if a

company is specializing in strategy games, they would try to figure out where overseas, is a good place to expand and acquire new users. We can look at users in those countries and see that people in those countries are over indexing in a specific game type.

Where do you see other applications of the data analyst role outside of technology and gaming?

In mobile, there is a huge industry in shopping; that industry hasn't been fully understood. Mobile shopping hasn't successfully taken off by using analytics. There is also a huge opportunity in Biotechnology. The human genome project for example. Also you can look at manufacturing companies such as jet engine producers, the amount of data that a turbine produces per day is enormous and if you could improve engine efficiency for example that would be a huge impact.

If you were to interview a data analyst, what questions/skills would you ask and look for?

When I'm interviewing other candidates, I try to figure out what they know. I.E. on a scale of 1-10, how well do you know SQL? And this is a loaded question actually, because based on their response and given background, you can assess their competency, analytical ability, and how experienced they are. Many candidates will say they're an 8, but I would say it is very rare to come across someone that is a real solid "8" since there are extremely limited environments out there that would give you the ability to become an 8. I also think it is preferable for someone to have a strong technical background, there is no need to know how to build apps or know how to code, but you'll need some relevancy by being proficient in SQL, know python, or you have a good statistics background. But, if you're a junior or 1-2 years out of school, there might not have been enough time for you to have that full spectrum to be fully developed in the data side and the business side. I'd probably pick one and focus on that side. Try to sell yourself on your ability and hunger to learn to increase your knowledge. You can be the greatest SQL guru, but if you are a poor communicator, and have no ability to extrapolate your work into business insight, you will hit a ceiling quicker than you expect.

SANTIAGO CORTEZ

SENIOR DATA ANALYST AT HG DATA

SANTIAGO is a senior data analyst at HG Data. He enjoys playing soccer in his free time.

HG DATA indexes, extracts, and renders unstructured data to fundamentally improve B2B technology sales and marketing programs.

Can you talk about your background and how you ended up working at HG Data?

I was going to school for Computer Science but to be honest I didn't want to be in an office all day long. I was in Santa Barbara when I saw an ad for a job on Craigslist. I thought I had the skills needed for it so I applied. The company originally was a database for non-profit donor information. We indexed donor information into a searchable format. A non-profit would come to our company and seek out where the best donors are. Trying to use analytics to automate this process of collecting donor information is where it all started. I believe that our competition at the time was manually collecting pamphlet data and typing it into the computer, and our use of analytics gave us a step on the competition.

Now that you're in the industry, why are you passionate about it? Why you believe in Information Revolution?

I believe it in because I've seen first hand how paying attention to patterns can have huge impact for a company. We identify and use patterns of strings that can help make our algorithms get smarter. The powerful thing is you can have a small team, yet if you correctly utilize the data, it enables you to do so much more than you ever could with a large team without the analytics. This is one aspect of the job that is different from the other ones I've seen. For example, let's say we are looking for a particular Microsoft product. Using analytics we were able to extract multiple valuable data points from a document, rather than just the single piece of data we originally searched for. We were able to find that certain strings of text contained keywords that had a tendency towards good hits, and others to bad. This is an example of a small clue in the data that changed our overall process. Ultimately we were able to increase the value of our product.

It seems like what you're describing is an "Aha" moment with data. How frequently does that happen?

It definitely is great. Depending on the process we're working on, these moments inevitably come up. You have a general map, and you want to get to a point. But along the path you find different ways you can add values to the end goal. It is a developing process. I don't think we have ever had a method/process that didn't evolve; most things become completely different than what they were on day one. It is an ever-changing process. That is another cool part

of the job. Since we're a smaller company, we can change quickly. New ideas are sprouting everywhere.

Can you then talk about what kind of skills/tools you use day-to-day?

The skills I use the most is in-house proprietary software that we have. We don't use too much outside technology. Besides maybe Microsoft Word or PDFs, most of tools are custom made by the engineering team. One of my main jobs is to refine some of the natural language processing related queries and overall quality control.

When you notice those things to change, do you prescribe things for the engineering team to edit the tools?

Yes, we have meetings and we talk about what we noticed. We come up with tools/ideas and we present them to the engineers. If we agree that there is enough values in it and we have the resources, it gets implemented.

What do you use to figure out these patterns? Excel and pivot tables?

Some of it is note taking. When you start seeing things, you write it down. If you see strings appear with frequency you write it down. It is about sifting through the data and either taking a mental note or writing it down. You try different searches and see what yields the best results. The real tool is the way we setup our Quality Control System, how we have people look at the data and our communication system to make judgment calls. We have an in-house handbook that we created that guides these judgment-calls because even if you see the trend, sometimes you're not sure if you're the only one seeing it. Having that open line of communication helps.

What is your workflow? Does your team figure out the keyword and the engineering team implements the tools that gather intelligence based on your findings?

It sort of starts with an order for a technology. It starts with our research librarian. He determines the best keyword for information retrieval. Then it goes through QC. We then determine if the result is viable/valuable. Depending on the results it gets fed back into the QC for another iteration until we have the clean data we want.

What advice would you give to undergrads/young professionals looking to enter the data analytics field?

It is a huge/growing field. With my case, it was a leap of faith. I would say that it is a great field to get into, if you're the type of person who wants to control the chessboard. If you like to see how different pieces move, you play around with different strategies and yet you love looking at the larger picture and use data as efficiently as possible. Especially with a smart startup (of course it varies from person to person, but my experience has been great), the excitement payoff is worth it for me. The opportunity of the wild-west frontier aspect of it is very rewarding.

The Data Analytics Handbook:
CEOs + Managers edition

COMING SOON

with interviews from...

Mike Olson	Rohan Deuskar	Greg Lamp
Cloudera	*Stylitics*	*Yhat*

---

and many more!

CHECK OUT OUR BLOG AT

statsguys.wordpress.com