



THE DATA ANALYTICS

HANDBOOK

CEO'S + MANAGERS

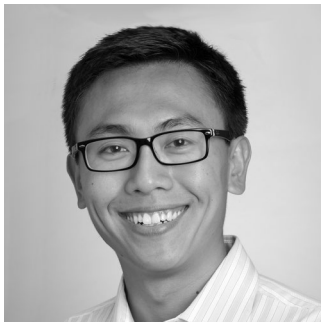
ABOUT THE AUTHORS



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 CEOS AND MANAGERS

1. The Information Revolution is real.

Data analytics will be applied everywhere but it will not surface itself to the end user like most good technologies. Data oriented decisions is a reality for all decision making in the future.

2. Data analytics is becoming increasingly accessible.

Technology is lowering the barriers to becoming both a data scientist and a data analyst. Tools being developed at BigML, Y-hat, Cloudera, and many more empower data analysts to work as data scientists and business analysts to work as data analysts.

3. Do your own projects to break into the industry.

The truth is, even in a quantitative major you are not taught what you need to know to work in data analytics. There is a learning gap between academia and industry that is best filled by doing projects. Find some sports statistics and do your own analysis. Learn R so that you can complete this analysis, not just to learn R itself. Also try Kaggle.

4. Statistics > Programming.


The development of tools and popularity of programmers has caused black box statistical analysis usage. Understanding selection bias vs. sampling bias and the underlying assumptions to which statistical functions are built on will make your opinions matter and your work invaluable.

5. The most important skill is being able to ask the right questions.

The power of data analytics is in taking open response questions and framing them to be multiple choice. Therefore if you have the ability to filter a million questions into options A through D, you are a data scientist for hire.



TABLE OF CONTENTS

- 05** DEREK STEER
MODE ANALYTICS
 - 10** DEAN ABBOTT
SMARTER REMARKETER, INC.
 - 14** TOM WHEELER
CLOUDERA
 - 18** MIKE OLSON
CLOUDERA
 - 22** ROHAN DEUSKAR
STYLITICS
 - 26** MARY ELLEN GORDEN
FLURRY
 - 30** GREG LAMP
YHAT
 - 34** ALI SYED
PERSONTYLE
 - 38** DAVE GERSTER
BIGML
- 

DEREK STEER

CEO + CO-FOUNDER AT MODE ANALYTICS

DEREK is the CEO and co-founder of Mode. Previously, he worked on the Analytics teams at Yammer and Facebook, both with a focus on monetization.

MODE ANALYTICS is building tools that eliminate the costs of sharing and discovery to make analysts and data scientists more productive.

Can we go into your background first? You have a unique perspective in that you have both the experience of a data analyst at Yammer and the experience as CEO of Mode Analytics.

I graduated from Occidental College where I majored in economics. My brain is just wired for economics – I think of everything in terms of opportunity cost and tradeoffs. I wanted to take a job where I could apply economic thinking, so I ended up working for an economic consulting firm doing analysis to support expert testimony in antitrust cases and mergers. I worked there for a few years, then made my way to Facebook. In consulting, managers tell you the beginning and the end of the problem you’re solving, and it’s up to you to figure out the middle part. At Facebook, I spent a lot more time thinking about the beginning (how to frame a problem) and the end (what the output/recommendation should look like). I wanted to be somewhere smaller, though, so I joined Yammer as one of the early members of its Analytics team.

At Yammer, I worked almost exclusively on sales, marketing, and customer-success related projects – a natural fit after working on monetization at Facebook. Yammer was similar in that they sought to “democratize data.” A part of my job was to enable my internal customers to make better decisions using data without necessarily having to ask me for recommendations every time. I did this by creating lightweight tools – either web apps or reusable SQL queries – that my customers could run on their own. When a particular request required ad-hoc analysis, I spent most of my time on the beginning and ends of the problem, and almost no time executing. That is to say, I thought of clever ways to solve problems by counting and summing things rather than building complex models. Model building takes a lot more time and doesn’t necessarily lead to better business decisions. My former boss, Peter Fishman, deserves a lot of credit for enforcing this type of behavior on the team, a behavior that, I believe, was critical in making the Analytics team at Yammer successful.

I left Yammer after two and a half years or so along with a couple of teammates there to start Mode.

What are the most important skills you think a data analyst needs?

Creative thinking and the ability to solve problems. I hate that answer though because it’s so vague. It’s true, but it’s vague. I will try to give a concrete answer.

I think it makes sense to explain this in the context of an interview question, because I'm specifically trying to measure problem solving ability.

The way I assess creative problem solving in an interview is to ask about a specific project the analysts completed at a previous job and request that he walk me through the tradeoffs he faced and how he determined the correct path. When I do this, I'm simply listening for sound logic. If anything sounds odd, I'll probe deeper. If I'm not an expert in the type of business problem the candidate is solving, he should be able to easily answer all my questions about how to perform the best analysis.

I also use some hypothetical questions. I like to ask people about pricing because it's very nuanced. "What should we charge for Yammer?" It requires some fundamental knowledge of Yammer's business model, so it's also a quick check to see if you've done your research about how Yammer works. Assuming you have, it's still a multi-faceted problem. Good candidates are occasionally able to talk through the complex work they might do to find the right price, but what I really love is when people break down problems into the simplest elements. The most important thing about setting price is understanding what you are trying to optimize: You're optimizing revenue, which is price times quantity. It's a simple equation but even just saying that gives you an understanding of the levers you have to pull and what they're going to affect. If you can nail that part, many other parts of the problem become clearer.

What are aspects of the job as a data analyst that you enjoyed and what are aspects that you didn't enjoy?

What I have enjoyed the most is working with great people. I would do any job as long as I liked the people I was working with.

Within the world of data analysis, I might be a little weird in that I don't particularly enjoy learning new technical skills. It's something I do because I have to, but what I really enjoy is the feeling of solving a new problem that leads to some sort of action. If other people are applying my work and seeing returns from it, that makes me happy.

How pervasive do you see the demand for data analytics becoming? Do you see the demand going past the metropolitan hubs of SF/NY?

I have a friend who lives in Arizona and works for CVS where they have hundreds of people who do data analysis. It's already there. I'll give you

another example. SAS has 70,000 customer sites in the US. It's not just a NY/SF thing. The question is whether people are doing it well. My suspicion is that the data analysis that people are doing across the US is actually relatively similar in nature. Even the data looks very similar from company to company within a given vertical. There's a huge opportunity to bring together all of these companies, as well as individuals who are doing analysis in a non-professional setting. That's part of what we're trying to do at Mode.

What are some of the softer skills required for a data analyst?

To some degree, analysis or data science is all about trust. You can deliver the best insights but if your customers don't trust you, it's useless. Analysis that does not turn into a decision is basically a waste of time. Establishing that trust is a tricky thing to do and requires a fair degree of interpersonal skill. It's partially about your customers believing that you're smart and capable enough to actually deliver accurate insights. Once you've got that down, it's about communicating ideas clearly. I love linear regression. It's fantastic because it's very easy to explain to a non-technical person what a coefficient means. You can identify which variables make a difference, and whether they are positively or negatively correlated with the outcome you want to affect. You don't even need to show your customer the regression output – you can just explain it in plain English. More complex models make this a bit tougher. And, of course, counting and summation are the easiest things to explain.

How granular do you see the applications of data analytics getting?

I think it's going to be applied everywhere but I don't think it's going to surface itself to most people as data analysis. The best applications of data science in products are things like Jawbone's UP, which is processing a ton of data behind the scenes to say "here's what you should do at this moment to be healthier." The recommendations are very powerful and you know they are based on data, but it's not something you think about as you interact with the product. Maybe even better is a sort of system where data gets you 90% of the way and you make the ultimate decision. The power of data science is how one can turn open-ended problems into multiple-choice questions. That's what data does really well.

What advice would you give a non-technical person trying to gain more technical skill?

It's pretty easy to learn new technical skills. Or, at least, it can be. I don't really believe in learning or applying skills simply for the sake of doing hard stuff. I actually think it's dangerous, because that's the type of thinking that can lead you down rabbit holes when you're doing something important that's time-sensitive. I think that the best way to learn skills so that you can apply them practically in the future is to start with a project, then learn all the skills necessary to complete it as you go. If you can't think of a project that incorporates the skill you're trying to learn, you should probably ask yourself why you're even bothering to learn it in the first place.

DEAN ABBOTT

CO-FOUNDER AT SMARTER REMARKETER, INC.

DEAN holds a B.S. in computational mathematics from Rensselaer Polytechnic Institute and an M.A.M. in Applied Mathematics from University of Virginia. In his early career, he worked as a research scientist at companies including Lockheed Martin and Elder Research, Inc. He also founded Abbott Analytics. He is now the co-founder and chief data scientist at Smarter Remarketer, Inc. He also recently published a book: “Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst” [Wiley].

SMARTER REMARKETER collects comprehensive behavioral data. This includes traditional “web analytics” style data, but also a compliment of specialized data required for behavioral marketing best practices. Through Smarter Remarketer’s Metrics feature, the company shows retailers exactly how people are behaving, why they are abandoning and how much revenue can be recaptured.

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

Data Science is a relatively new way to describe analytics. Data analysts are more business focused, likely to be less academic focused, have a bachelors degree, and are trying to solve business questions. Data Scientists are more dominated by advanced degrees holders, more likely to work with big data tools (such as NoSQL, Hadoop, MongoDB, and Cassandra), utilize R/Python, and also more likely to be serious programmers. The data science position may be a superset of most analyst roles, whereas data analysis may be a subset of data science.

Would you say that data scientists and data analysts work in disparate groups?

I think so. The data scientists I work with are more research oriented. They are more cutting-edge and are working to create new ideas and new algorithms. They also try to solve problems where the traditional databases don't work well. Data analysts are more likely to work within a business group. In certain situations where there is a specific data science group, people are more likely to be labeling themselves data scientists.

Do you see the data analyst job becoming more standardized?

Data analysts / business analysts (other times known as predictive modelers or data miners), are more likely to work within business oriented groups. A lot of these people are not coming out of college wanting to become data analysts; generally they have a business related degree. But as they work, they realize that they like math and data. Over time these people cluster towards data analytical jobs. Just recently, one of my clients hired people from the outside to train in-house employees to become the "data analyst".

How deep of statistical understanding / programming skills do you need to become a data analyst?

More math is always better, especially if you're doing cutting edge work. Nonetheless algorithms such as Decision Trees, Neural Networks, and Support Vector Machines will work right out the box even if you don't know how they work mathematically. I have a book coming out called "Applied Predictive Analytics", which teaches people predictive analytics while using minimal

mathematics (some algebra and basic stats terminology). It is important to know when/how to turn the knobs in statistical models. Ultimately, math isn't the most important metric for an analyst's quality. But rather love for data and an innate curiosity make for a great analyst; I call it the "freakonomics mindset". Also, knowing the details/strength/weakness of the algorithm is always helpful. This is especially important when the model doesn't work: why is the decision tree not working? Can the data be skewed? You learn to ask the right questions.

How do you develop the skills required to be an effective data analyst? Your previous work involved your client bringing in an outside expert (you!), but how can an undergrad/recent-grad develop the same skills?

Two ways. First, take a few classes to get some fundamental understanding of data. Not necessarily databases, but rather about data and what data means. Basic stats and basic courses in machine learning can help (know thy data!) In university classes you learn theoretical details and implementations etc. But the details are not required and are no longer used once you're out working in the industry; university educations don't prepare you for business applications. Some schools with application-based courses (UCSD, UCI) have after-hours programs for business people, which are helpful. Second, start building models. Work on projects. It helps to work with someone who has done it before. Data preparation is harder to teach because there are so many ways for you to do it incorrectly. It is hard to teach in a way where you cover all "incorrect" approaches. It is a bootstrapping method; try co-op, internships, working in a company etc.

How do I make a switch into data analytics after I've been working in the industry?

Take online courses. Read some books ("Data Mining Techniques" – Linoff, or my "Applied Predictive Analytics" book). There are books that will teach you applications, good principles, and how to approach the problems. Once you understand the process better, get on Kaggle! "Handbook of Statistical Analysis and Data Mining applications" – Elder, Miner is also a good read.

Which is more important for a data analyst: intellectual curiosity/intuition V.S. deep statistical knowledge?

The curiosity. The understanding of statistics can be gained, but the curiosity is more innate; if you're not naturally into working with data, you can be a fantastic analyst, but you become a specific type of analyst. These analysts may not care enough to build a better model. They lack creativity and the determination to go the extra mile. They'll need handholding and will struggle to resolve new problems.

With curiosity about the data and an interest in why the data looks the way it looks, you can overcome a lack of mathematics training. I once worked with an analyst who had only high school math but became a great modeler. In the beginning, he didn't even know algebra well and certainly didn't know any calculus or linear algebra. But he had good intuition and determination, and ultimately became an excellent analyst and modeler.

TOM WHEELER

SENIOR CURRICULUM DEVELOPER AT CLOUDERA

TOM spent most of his career as a software engineer working on the design and implementation of software for major financial, healthcare, and aerospace firms. He started using Apache Hadoop in 2008 after he recognized that it solved a common problem he'd observed at several companies; namely, how to store and process ever-increasing amounts of diverse data at reasonable cost. In 2011, he joined the curriculum development team at Cloudera so that he could share what he'd learned with others.

CLOUDERA is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

What are the most valuable skills an aspiring data analyst should develop?

The most important skills are understanding which factors help a business succeed and determining the key metrics by which those can be evaluated. These skills ultimately provide the foundation needed for constant experimentation and iterative improvements that will help uncover important patterns and help a business thrive.

Our understanding is that Data Analysts require a certain level of technical skill and experience; do you not feel that is the case?

While Data Analysts do require some amount of technical skill, it's different than what a programmer might need. Instead of developing an e-commerce application in Java as a programmer would, the data analyst might write queries in a language like SQL to analyze the profit margins or sales trends for that Web site. Even though these skillsets are a bit different from one another, both require specialized training.

Let me also contrast this with the role of a Data Scientist. Given a set of information like the order data for an e-commerce site, the Data Analyst knows how to answer important questions like, "how many products did we sell this year compared to last year" or "which product categories have the highest profit margin?" The Data Scientist, on the other hand, knows how to gather new data as well as query it. They approach the problem scientifically, designing experiments, forming hypotheses, and building systems to collect the data needed to validate the results. In other words, the Data Scientist knows how to ask questions as well as answer them, and has the software engineering skills needed to build new tools when needed. That's a rare combination of skills, but one that's increasingly in demand as businesses become more data-driven.

Can you expand on what you mean by a Data Analyst being able to ask the right questions?

Data analysts need to understand the constraints of a business and what makes it successful. For example, if you're working for an airline, then you'd want to know which things affect costs, which things affect revenue, and which things help you to attract and retain customers. Fuel is a major cost for an airline,

so you might do some analysis to project the future cost and buy in quantity when the price is low. In terms of increasing revenue, consider that customers are typically more sensitive to price when buying tickets for personal travel than for business, since the fare comes out of their own pockets. You might therefore attract new customers by offering low-cost fares to popular vacation destinations like Florida or Hawaii and offering higher-cost fares with perks that appeal to business travelers going to destinations like New York City.

Do you think there is a less experienced version of a data scientist that an undergraduate degreed person could achieve?

Certainly. Not all data scientists have a Ph.D. They tend to have a mix of skills in statistics, software development, research, business, and so on. That's a really rare combination and there are currently very few PhD programs that give you exposure to all of them.

Just like there are lots of amazing programmers who don't have a PhD in computer science, so too are there amazing data scientists who started working after getting a Masters or Bachelors degree in one of those areas. If they have an inquisitive personality and a lot of self-motivation, they tend to can quickly gain any other skills they need through real-world experience.

How would a young professional best prepare himself to be a data analyst?

Aside from general business experience, it's important to learn how to use a variety of tools. People who only know one tool really well tend to see it as the solution to every problem, but people with a broader range of experience are versatile and adapt quickly.

Although data analysts often use graphical business intelligence tools such as Tableau or Microstrategy, it's worth the time to also learn how to write your own queries directly in SQL. This lets you go beyond what a specific tool might support and allows you to directly query any modern database or even a "Big Data" system like Cloudera Impala.

Learning how to work with Big Data is increasingly important, and fortunately, many universities are starting to offer courses on it. Cloudera has an academic partnership program with dozens of leading universities that provides professors with material from Cloudera's professional training courses. As part of this program, students also receive a discount on certification exams,

and getting certified is a good way for people to demonstrate their expertise and stand out in a competitive job market.

Finally, self-education is a great option for learning new skills. Udacity recently partnered with Cloudera to offer an “Intro to Hadoop and MapReduce” course, for example, and they’ve got some others that are really relevant to people just getting started in analytics, such as Intro to Statistics, Intro to Data Science, and a range of machine learning courses.

What do you think of R as a skill for data analysts?

R is definitely a valuable skill to have. Although R was designed specifically for statistical analysis, it still is a programming language. Every Data Scientist definitely ought to know R, but not every Data Analyst will use it. I’d say that SQL is the more valuable skill for Data Analysts in general, but once they’re comfortable with SQL, it’s worth the time to learn R too.

How clearly do you see the role of the data analyst as being defined in today’s job market? Do you see it becoming a more standardized position?

Data Analyst is one of those job titles that mean slightly different things for each industry or organization. It’s kind of a general term like programmer; for example, some programmers design new programs while others do maintenance on existing ones. Some are doing low-level things like writing device drivers and others are doing high-level things like developing Web applications.

Although I don’t really see the “Data Analyst” role becoming well-defined across all industries, I do see demand for that role increasing across all industries. To paraphrase something that Google’s Chief Economist Hal Varian said in an interview for the Freakonomics blog in 2008, the key to success is being able to offer value around some commodity that’s becoming ubiquitous and cheap. Organizations are producing exponentially more data than they did even ten years ago and this trend will continue far into the future. The people who have the skills needed to extract value from all of this data are going to be in very high demand.

How would you describe the value that data analysts bring to a company? How about data scientists?

Data is newly available in volumes and in variety previously impossible to capture. That explosion in growth is a big opportunity – companies that figure out how to process and analyze it will rack up real competitive advantages over those that don't. Every organization needs to become data driven.

Fortunately the tools available have evolved quickly right alongside the data. It's now easy and inexpensive to store, process and analyze that data. The skills required to do so are still reasonably thin on the ground, and the profession of "data scientist" has emerged to fill that gap. It's a rich area for innovative smart young folks to get into.

In the long term, the data science expertise that's between the ears of people needs to be rendered into software, so that we can scale out. It's hard to make huge numbers of really smart new workers; it's easier to capture smarts in applications and tools and make them available to traditional businesspeople who can use them to explore their markets, evaluate their competitors and engage with their customers better.

So I believe the best data scientists will also be competent software developers.

Cloudera sells its big data infrastructure to banks, hospitals, etc. but who takes care of the analysis? How much analytic consulting does Cloudera have to provide?

About two thirds of our revenue today comes from software subscriptions, and we see that fraction growing over time. Our annual renewals model means that most of our installed base renews every year, and the work we do to find new customers pays us off for years to come.

Of the remaining third, the money is split about equally between training and consulting. We view both of those as ways to drive our subscription numbers up – we make customers smarter with training, and we help them build analytics with consulting, and they'll consume more of our software.

We have a rich ecosystem of systems integration and consulting partners including big names like TCS, Accenture and Cap Gemini, and smaller firms like Think Big and Mu Sigma. They do a lot of work with our customers. One financial analyst estimates that for every dollar spent on software licenses in big data, fourteen dollars are spent on hardware and integration services.

Overall Cloudera seems to answer the question of big data capture and storage. What do you see as the answer for the question of analysis? Put another way there is a clear talent shortage in data analytic skill, how do see this gap being filled?

There is a lot of innovation in that space right now. Tons of startups are attacking different vertical-specific analytic opportunities. Others are building horizontal capabilities – machine learning, for example, is useful across most of our installed base.

As I said before, those techniques need to get rendered into software and made consumable by traditional enterprises. Business people are plenty smart enough to use the tools and interpret the results, they just need help from data scientists who build the apps that do machine learning by way of random forests and present the results in useful ways.

When relational databases were brand new, no one knew SQL, there were no apps and business people couldn't use them very easily. That changed with the huge economic opportunity that the database market created for app and tool developers. Same thing is happening today in big data. The landscape is way, way better than it was when we started six years ago!

Currently it seems like big data technology is really only necessary or possible with the largest technology companies or corporations but do you see big data applications (using Hadoop/MapReduce) for small businesses eventually? If so, how?

I think it's already there. Most of the public cloud adoption of Apache Hadoop is by small and medium-sized enterprises.

I also think that people will consume machine learning as a service, bundled into business apps they require. Think about how you use Amazon and Google. There's tons of analytics and personalization behind the scenes there. You're getting Hadooped all the time. You may not run that software yourself but you are using it extensively by way of the online services you consume. More cloud apps will mean more big data analytics, made available in useful ways to regular people and small companies.

What are the limitations of Big Data? What are the pessimists of the information revolution arguing?

Clearly privacy is a big deal, and we have work to do there. Technology determines what can be collected; we must formulate a policy and ethics that tell us what is okay to do with it, and we need a governance framework that can impose meaningful penalties on misuse. Plenty of opportunity but the need here is pretty urgent.

The problem is worse because of inconsistent laws among countries. The EU has a pretty comprehensive data policy, with each country having adopted a consistent framework at the national level. The US has essentially no Federal-level data policy. We spell out policy in some key verticals – HIPAA, FCRA – but leave it to states, and worse to companies, to craft and enforce policies generally.

That patchwork makes it very hard to do business. The internet doesn't really know about geopolitical borders. When EU data privacy and safe harbor provisions impose rules on data that crosses borders, it creates enormous friction for businesses and NGOs. Friction equals cost.

What advice do you have for undergraduates or young professionals aspiring to enter the industry?

My advice to young people pursuing any course of study is the same: Find problems that really interest you, really blow your hair back, and find a way to work on them. Chase interesting. Don't chase money. Money will come if you love what you do so much that you become a world-class expert and plow huge effort and time into your work.

ROHAN DEUSKAR

CEO + CO-FOUNDER AT STYLITICS

ROHAN is the CEO & Co-Founder of Stylitics. His goal is to help consumers and retailers leverage the data in consumer closets to power smarter decision-making, better recommendations, and in general, bring the fashion industry into the digital age. Rohan was selected as a 2014 New York City Fashion Fellow and has spoken on the intersection of fashion and data at events like NY Fashion Week and DataBeat. He has an MBA from the Wharton School.

STYLITICS is a fashion analytics company that leverages a proprietary stream of consumer closet data to help retailers improve CRM, recommendations, and targeting. Stylitics has built the world's largest consumer clothing panel using its popular digital closet platform.

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

As we see it, a data analyst is somebody who can sit down with a set of data, in some simple form like CSV, and is able to do two things. Firstly, understand the patterns and trends in the data. Secondly, be able to manipulate the data through various tools in order to understand the fundamentals of the dataset. That could be knowing about data that is missing, knowing additional data that is needed, or some other kind of insight about the characteristics of the data we're working with. Basic tools like pivot tables are a requirement, but a good analyst is somebody who can come back to me and say "I'm worried because I've noticed that average price is too low for this product category. It tells me the data is weak here. Or maybe we're counting blank entries as zeroes." The best data analysts act like detectives.

Given trends in technology are you requiring higher technical skills for your data analysts with languages such as R or SQL?

Knowing SQL is an advantage because it brings you one level closer to the database, allowing you to craft your questions in a way you are unable to when using pivot tables. Given that, I would rather find somebody who doesn't know SQL but has a really good eye for finding patterns within the data, than find someone who is technically proficient but doesn't have the right attitude or mindset towards it...that innate curiosity.

How would you define the role of a Data Scientist?

Data Scientist is a very broad term. Our company tends to approach it from a business intelligence perspective. At this point, we work with new data streams rather than trying to do very complex analysis on existing data streams. A data scientist shares the curiosity and intuition I mentioned for analysts, but is much more technically proficient and has a very solid grounding in statistics. They should be able to say, "Yes on the surface this looks like a trend but for this reason it's not." I would also look for someone who is comfortable with the latest machine learning methodologies and has applied those methods to real world data problems.

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

Our approach has always been different because we have a unique data set: post-purchase behavior in fashion, provided by consumers themselves. This is hugely valuable for retailers and has never been available before at scale. Our approach to data is to make it accessible and relevant for our end customer. We think of things from the customers' point of view. What is the day-to-day of the fashion business? What is the data we have that relates to their business needs? Can we answer their questions better or faster than they can using the data that we have? We want to give them data they can pull up on demand in the 5 minutes before they go into a meeting.

We also try to be very disciplined on data cleanliness. There have been many chances to get more consumer-generated data with less cleanliness. For example, on our mobile apps we could let people simply write in the product name in a free-form text field versus making them select from dropdowns. It would make adding items easier and generate more data. But then we'd have much less ability to easily sort and categorize the data and provide useful insights. Other companies have opted to go with letting users tag photos in an unstructured way...as a result, our data set is marketable whereas theirs is not.

Can you elaborate more on how analytics surfaces in your product?

Like the name says its style + analytics, so the analytics piece is baked into everything we're doing. For example in our consumer platform, as you add your outfits to your calendar it's calculating the cost-per-wear for each item of clothing and the number of times you've worn it. We also help you find brands that you may like from your friends through a similarity score. My personal opinion so far from everything we've seen is that you don't need very heavy-duty data science. There's a simpler way to leverage your data to deliver a lot of value to users or customers. It's just about understanding what data point is most useful to users of your product at that moment.

So how is what you're doing special in your industry?

We specifically cater to the fashion industry that historically has been primarily an intuition driven business. Our approach is to bring competitive and consumer data to complement human expertise and intuition. It takes some education because retailers tend to be of two kinds: 1) The vast majority of retailers and brands in fashion who make their decisions mostly based on personal experience and opinion. 2) A newer set of retailers who track dozens

of metrics religiously and base their decisions only on data. We're helping companies to use a combination of data and human judgment, which is where I think the real success happens.

What kind of interview questions would you ask to hire a data analyst?

We are actually in the process of hiring a couple analysts right now. First of all I want to know, have they worked with large data sets before and to what extent? The next thing I would do is show them a data set, for example all the details for the clothing purchases made by everybody in your university, and ask them what are some of the insights you could get from that? I want to know how they structure the questions about data themselves, without prompting from me. e.g. "I know you asked only for the impact of weather on category sales, but I decided to also check if shades of color sales were affected by weather and found something interesting..."

We will also give them a raw data set to take home and have them share five interesting things they see in the data. They would also be asked to present their findings in a couple of PowerPoint slides because part of the data analyst role according to me is being able to convey your findings to people who haven't spent the time you have on the data.

What are your thoughts the development of technology outpacing the demand for data analysts?

Yes, in a sense our technology removes some of the need for data collection and analysis by humans. That's part of a macro trend. But with more companies focusing on personalization, targeting, CRM, and other data-driven approaches, the overall need for data-savvy people will grow as well.

The existence of data mining and analysis technologies shouldn't dissuade students. For example, spell check exists, but you still need to know how to spell words and use them in the right context...

It's important for everyone to know how to think about data, how to frame the right questions, and how to work with large datasets. All of that will serve you well regardless of the specific tools that are being used.

MARY ELLEN GORDON

DIRECTOR OF INSIGHTS AND ANALYTICS AT FLURRY

MARY ELLEN has spent her career in marketing-related research and analytics. She previously worked at Flurry and owned a market research company that conducted commercial, government, and academic research for clients in the U.S., U.K. and New Zealand. She has also worked in data-oriented jobs in technology, pharmaceutical, and consumer product companies.

FLURRY is optimizing mobile experiences for people everywhere. Flurry's market-leading analytics software is in over 400,000 smartphone and tablet apps on over 1.2 billion devices worldwide, giving the company deep understanding of mobile consumer behavior.

What led you to work in data analytics and why are you passionate about it?

For me it goes back to when I was an MBA student and took a marketing research course. At the time conjoint analysis, a form of tradeoff analysis, was gaining popularity. We gauged how much customers will pay for different features and how much they value one product attribute against the other. In doing so I saw the value of analysis when it came to making good marketing decisions. I realized I was more interested in the research side of marketing rather than the tactical side so I ended up doing my PhD and getting a lot more training in sophisticated analytical methods. Since then I have worked on the analytical side of marketing.

What aspects of your job at Flurry do you enjoy, and what aspects do you not enjoy?

Flurry has big data, and I enjoy seeing what I can do with that much data in terms of really honing in on differences that you might not be able to see with self-reported survey data. On the downside, the data management required when you have that much data is a bit annoying. I come from more of a survey research background where the data is much more clean and more structured. With our data there is a lot more work that goes into pulling the data you want off of our system. Sometimes that's even impossible or requires an engineer. Sometimes the file that you generate is too big to open in a single computer. So before you do actual analysis, you have to do a lot of work to get it ready and into the format you need before you get to that "aha" moment.

That "aha" moment you're describing, how often does it happen?

Some discoveries are more interesting than the others. They all kind of add to a bigger puzzle. The smaller "aha" moments give you insights on different parts of the picture and then when you link discoveries together to get even more information you get the really cool insights that come from seeing the fuller picture.

Why should a recent grad work in this field, as opposed to a more traditional job?

I'm actually not sure they have to choose and I would almost encourage them not to choose. I believe what will be increasingly valuable in the future is

not what is valuable now, but rather a combination of quantitative skills plus domain-specific skills (such as marketing). There probably will be more jobs for people with that combination of skills than for people who are in pure data science. Your data comes from a specific context; the context gives you clues as to what your data might be trying to tell you. The quantitative tools are there to help you better understand the context.

How do you think the role of a data analyst has changed, since it has become popular, and how do you think it will evolve?

I think there will more emphasis on understanding statistics. Analysts will have to be able to tell you the best technique to extract answers, given a specific data set. People will have to know about the different assumptions in regression; what different distributions looks like and when to use different clustering algorithms. That kind of statistical knowledge will help you when you're trying something and it isn't working. It helps you diagnose the problem. Many data science people now have a more solid understanding of programming than statistics. Many people know R and Python and know how to fit a clustering algorithm to a data set, but lack much knowledge of the underlying statistics that they are applying so are almost using them as a black box.

How do you think your company uses data analytics as a comparative advantage to other companies?

One of our products is analytics, so in that sense we help provide a comparative advantage to other companies by helping them understand their users. Beyond that we are also taking the data we generate through our analytics product and using it to power advertising. The better our analytics are, the better we are going to be at showing ads that are relevant and interesting to you and therefore have a good return for advertisers.

How soon do you think skills related to Big Data (i.e. Hadoop/Hive/Pig) will be a necessary skill for data analysts?

It would awesome to have those skills now. I don't know MapReduce and Hadoop and it would be really useful if I did because I would be much more self-sufficient in terms of being able to go and get the data I need rather than being dependent on someone else to get it for me. Over time, I suspect that tools like Pig/Hive will get developed (and ones that are even more user friendly), that will let you query even very large non-relational databases without knowing the

programming concepts behind them. The amount of data that we have and the uses for it are such that we can't be dependent on the small number of people who have high-level programming skills. Also, if you've invested enough time in your career to have that level of software expertise you may not have subject matter and statistical expertise. What it gets down to at the moment is having a team of people who have these different skillsets. Over time, the tools will catch up and you won't have to know how to program to get the data.

What do you look for in a data analyst, and what specific skills or experience do you favor? How do you interview them?

They would need to know in which context to apply what kind of tools. Do they know the context for factor analysis, or when regression is appropriate? The next step, now that you know which tools to apply when, is how do you deal with challenges such as your data is incomplete, or where you have results and they haven't come out in a way that you expected. What is your process then? Do you have a process for checking for bad data? Finally I check for the ability to translate results from data to a solid explanation of "what does that mean"?

What is the range of technical ability do you look for?

The skills will depend on the specific job. It is unrealistic to expect everybody to have the domain expertise and the programming expertise and the statistical expertise. So what is the job, what are the specific skills that are required for them to perform, and what do they have already? What kind of person can we bring on to speed up the process we already have? For example right now we sometimes depend on engineers to get the data we need, so someone who knows MapReduce and Hadoop could speed up our workflow. If we had that person then we would probably look for someone with a strong statistical background to work with the data that is being pulled appropriately. Ultimately, I look for different skills depending on the situation and don't expect the whole package in a single person.

GREG LAMP

CO-FOUNDER + CTO AT YHAT

GREG graduated from University of Virginia with a B.S. in Systems Engineering as well as a B.A. in financial mathematics. He worked as a senior data analyst at comScore and also as a product manger at On Deck Capital. He is now the co-founder and CTO of Yhat.

YHAT is a cloud solution for deploying predictive analytics at your company. Data scientists create pipelines to ingest and clean data and train and deploy models. Models are instantly accessible via REST, low latency streaming API, and batch mode.

Based on you experience, how do data analysts differ from data scientists?

I generally lump them in the same category. This is because a lot of companies are saying that they are hiring “data scientists”, but they’re actually looking for entry-level data analysts. It is a sexier name; people want to become data scientists. I would classify a data analyst as someone who can use business intelligence to generate reports and metrics. Data scientists are capable of doing BI/reporting, but they can use statistic and machine learning techniques to better understand their data and business problems. I’ve also found that while a typical data analyst is comfortable gathering data from a pre-packaged source such as a data warehouse, data scientists are much more adept at collecting their own data via web scraping, surveys, etc.

Who do you foresee using your product: analysts, or data scientists? How will it improve the current process?

*What empowers users to deploy production-ready statistical models to the cloud

Right now we’re catering to the data scientists. However we’re looking to expand and offer a tool that is more accessible to data analysts and business people. Once you go out in the industry, often times you’ll see data scientists’ works getting left on the shelf. There are engineering barriers where less technical people can’t utilize advanced analytical work. So we’re building a tool that will enable a company to utilize the prototypes that the data scientists built, and turn them into a deployed data product that can then be used by everyone in the company.

I think you’re implying that there is a technical gap between the new tools and the current generation of analysts. To fill this gap, will the tools become easier to use or will analysts have to become more technical?

I think it’ll go both ways. Having a technical analyst is very helpful. Nonetheless tools have also gotten better, particularly with python and R (to a lesser extent). Tools are becoming mainstream and easier to use. Ultimately, the idea of becoming “data scientists” is something data analysts are going to aspire to, thus pushing them to increase their technical abilities. It is getting easier for analysts to make the jump into technical analysts and eventually data scientists.

There is also a wide spectrum of tools being built: from a excel plugin-ins to high-end machine learning platforms (namely GoodData, MapR, Cloudera etc). I strongly believe that there is no such thing as “data science in a box”. You can’t just take data and stick it into an algorithm that will predict everything (think Google prediction API). Therefore the humans in the process add tremendous value; these individuals provide domain expertise and make judgment calls. No matter where the industry goes, I think that both side of the equation (tools and analysts) are important.

So do you think that this space will always require human interaction?

Yes, I think certain activities will get easier and become automated, but we’re not going to get to a point where “plug in data science software” happens.

From a recent graduates point of view, how do you develop the skills to work as a Data Analyst?

It depends on the industry and how technical you want to be. Learning R/Python will never be a bad thing. The other aspect is picking up a project and doing it. That is the best way to pick up new skills. Without a defined goal, it is challenging to put in the extra effort to complete the project. For example, here is what I did when I was learning R: I picked a project, which was to automate our company’s quality control process. Up until then, they were manually made in excel. Ultimately pain and suffering lead to learning.

A lot of people have the idea that “domain specific knowledge” is more important than the statistics itself. Do you agree?

This actually comes up a lot when talking to colleagues: “is the data more important or the algorithm more important”. I personally think that there is no replacement for domain specific knowledge. I previously worked for OnDeck, where we used technology to improve the way small business loans were made. I learned that even if you have a bunch of data about the individual, it didn’t necessarily improve the lending process. What really matters is asking the right questions, and finding the right data to answer them.

Any last advice to our readers?

The good news is that breaking into the field is easy; we have a new hire starting in a month. He is a recent grad from San Francisco State University. He

simply e-mailed me looking for a project to work on and started volunteering his own time. That is a great way to get exposure; startups are going to be open to people looking for work experience. Reach out and don't be afraid. When we were starting Yhat, we emailed 20 companies seeking advice. Almost everyone was willing to meet with us. We got 10-15 meetings on the calendar during the first few days. People are generally very helpful.

ALI SYED

FOUNDER AND CEO AT PERSONTYLE

ALI is a data science enthusiast and transformation leader with more than a decade of experience assisting organizations in using insights and intelligent predications as a value amplifier. He loves drinking good coffee, reading great books and meeting interesting people.

PERSONTYLE was born from the idea of creating a platform for the people, by the people, to share passion, knowledge, theory and practices of scientifically analyzing data. Persontyle's fundamental belief is that helping people learn and apply Data Science enables freedom, and yields important development benefits for individuals and society. Persontyle offers a comprehensive portfolio of Data Science learning opportunities and services for organizations (profit, non-profit, and government) to deliver breakthrough impact and meaningful value using data.

As someone who is offering coursework in data science, what do you think is the most important skill that an analyst has to acquire?

Fundamentally, what sets a creative data scientist apart from the rest is an obsession to challenge the status quo. Great data scientists never compromise with mediocrity. Not just data scientists, but all of us should be able to use data to ask relevant questions and challenge the way things have always been done. If we don't break the rules then the rules will break us. There are too many mediocre and boring things in life to deal with and data science shouldn't be one of them.

Data Science starts with creative imagination. Our ability to imagine and think beyond the obvious is one of our extraordinary powers as humans. It is why we are different from other beings on the planet. We are creatures of action; we build and create things. We don't live in the world directly; we live our lives through our imagination, which is full of ideas, concepts, theories and ideologies.

One can learn tools and algorithms to deal with this data deluge and get hired based on skill-sets reflected on a resume, but you will only be able to add real and meaningful value based on your ability to improvise, adapt, and create. Last thing we need is a breed of data scientists who are too skills-centric and have not dwelled much on the idea of being the biggest skeptic of data. Before we end up with an army of tech-savvy data scientists who are nothing more than clever calculators, we need to ensure that data science remains a science and does not degrade into some gibberish practice that is more of a collection of techniques rather than a discipline. To conclude, if data science doesn't require thinking, challenging, questioning, and rebelling, then it's just a useless tech to protect status quo and mediocrity.

Why did you decide to approach educating students in data science via in-person classes rather than an eLearning course or MOOC?

I was expecting this question. In the last 6 months I've been asked this question so many times that at one point I was seriously thinking about putting it on the back of my business card. I have a short answer and a long answer. Short answer is more positive and long answer includes detailed criticism.

We encourage people to attend MOOCs and use any possible channel (i.e. do whatever it takes) to get as much information as you can. However I do think that the classroom style learning will always remain the best format for educating minds. MOOCs are not there to challenge the existing educational channels (universities and the like), but rather to complement them. Just like how eBooks will never replace books, MOOCs will never replace formal educations. And the day this happens, we may bid adieu to future and humanity.

There are more than 2 billion potential learners around the world today, and more than 70% of them are unable to afford college degrees. So we have to try all the possible means and channels to make education affordable and accessible to those people. MOOCs and eLearning are one of the ways to address this challenge. Also MOOCs are exciting people into trying new disciplines that they might have not previously had access to.

What is your opinion on using higher education to gain the skills necessary for Data Science?

Schools' sole purpose is not to train us for industry. The thought that the sole purpose of education is to prepare us for industry is incorrect, limited and very capitalistic. Going to school is an absolute must; we won't ever encourage a young professional not to go to school. Having said that, the challenge for all of us is making sure that the education we deliver is not commoditized.

Since your company is based in London, have you found that the market for data scientists to be different in UK than in Silicon Valley?

Data science is an emerging discipline and has the potential to become a profound branch of knowledge. In the last year or two, we witnessed universities create departments and offer programs related to data science. This is not just happening in the US or Europe; in fact it's happening everywhere. In Asia, Turkey, Middle East, just to name the few. Data, as a fabric of this digital age, underpins all that we do. Using data as a strategic resource is essential for all, regardless of where you are geographically. So it's definitely not just restricted to the Silicon Valley.

How do you think the role of a data scientist will change in the next 5 years?

I can't say anything about future with certainty. I can only speculate. I think

the skills will become much more standardized and pervasive.

Do you have any last advice for young professionals?

Be awesome, be yourself, read complex/complicated books, strive to learn the true meanings of life, break the rules, challenge the status quo, work on meaningful problems, fix the world, and create your own future. Remember to not waste a single moment of your life working on meaningless pursuits and listening to the BS of your bosses.

DAVE GERSTER

VP OF DATA SCIENCE AT BIGML

DAVE is the VP of Data Science at BigML. He attended Harvard and also has an MBA from Haas School of Business. He has taken various roles at companies including Yahoo, Cloudera, Groupon and more.

BIGML offers a highly scalable, cloud based machine learning service that is easy to use, seamless to integrate and instantly actionable.

What problem is BigML trying to solve?

We're trying to bring data science to a much larger audience. There's been so much hype about the mythical "data scientist" who has a quantitative PhD, deep background in statistics and machine learning, mastery of both relational databases and Hadoop, the ability to program in Java and Python, and so forth. These qualifications are extremely hard to find, so unless you're Google, Facebook or LinkedIn, hiring a team with this background is simply not feasible. A better solution is just to give data analysts better tools, so they can do the deep analysis that was previously the domain of these scarce data scientists.

At BigML, we started with the area in data science called predictive modeling. We did this for two reasons. First, there are a lot of business problems that naturally take the form of trying to line up a collection of inputs with some known output. Second, training a predictive model is a well-defined workflow that really lends itself to automation. You always start with a training set, use an algorithm to train a model, evaluate that model using holdout sets, and then maybe upgrade your model into an ensemble to get a better accuracy. Finally once you're happy with your model, you use it to make predictions. Again, all of this falls into a well-defined and automatable workflow.

I think both data analysts and data scientists will like what we've created. With BigML, data analysts can now train predictive models, which was previously the domain of data science. We even let you export your models to Tableau or Excel VBA, so you can make predictions using your favorite analytical tools.

Likewise, data scientists are liberated from the drudgery of model training and validation. This frees them up to worry about bigger problem. If you're a data scientist wrestling with R on your laptop, BigML will greatly simplify your life: we give you an automated workflow, train your models on beefy server class machines in the cloud, and make it easy to export your models as code. We also have a powerful API that lets you integrate BigML with your existing workflow, so if you're already doing everything in Python, that's no problem.

What are the challenges for BigML when creating these tools?

The biggest overall challenge is design: we're trying to take this complicated modeling process and make it simple for the end user. One specific challenge is making sure users don't overfit their models, so we've paid a lot of attention to making model evaluation really easy. With BigML, you can evaluate a model

using a holdout set in just a few mouse clicks. Personally, I'd like to reduce that to zero mouse clicks—just automatically run an evaluation in the background every time someone trains a model.

Are there problems that better tools can't solve?

I think we'll get to the point where BigML can train the best possible model with a single mouse click. The only constraint is how much computing power we provide on the backend, so I can easily envision something much more powerful and more automated than what we're doing now. For example, you can extract a few more points of accuracy just by having the right number of nodes in your decision tree, or the right number of trees in your ensemble. Why not tune these parameters automatically? Again, it's just a question of compute resources, which are constantly getting cheaper.

Would you say that BigML reduces the amount of statistical knowledge required to do advanced analyses?

I think there will always be demand for people who understand the theory behind statistics and machine learning. But in the future, these highly qualified people won't spend their time on tasks that are automatable. If it's automatable, then why do it manually?

Would you say that data analysts are one of the key audiences for BigML?

Yes, definitely. A data analyst has a good understanding of the business and wants to test out their hypotheses, but they don't know the details of how a decision tree algorithm works, nor should they be expected to. An analyst does need to understand the concepts behind predictive modeling, and I think the simplicity of BigML helps with this understanding. Ultimately, our goal is to liberate the analytical brainpower of these folks who already know the business, the data, and the right questions to ask.

There is a McKinsey report that predicts a 1.5M shortfall in "data-savvy analysts and managers". What do you think is contributing to that shortfall?

Part of the challenge is that it's hard to learn practical data skills in an academic environment. For example, the skills required to wrangle data into usable form aren't something you learn in academia, even if you're pursuing a

quantitative degree. You'll undoubtedly learn a lot of valuable theory, but you probably won't learn how to write a complicated SQL query.

Looking ahead, how do you see BigML's business developing?

Mountains of data are just sitting untapped right now, at companies big and small. Large companies have the most money at stake, so I see them as early adopters of BigML. As advanced data analysis continues to become more mainstream, I think smaller companies will contribute an increasing share of our business. These are companies who might be using Excel or Tableau, so BigML really opens up a whole new world of analysis for them.

*The Data Analytics Handbook:
Researchers + Academics Edition*

COMING SOON

with interviews from...

Peter Norvig

*Director of Research
at Google*

Hod Lipson

*#7 Most Powerful Data
Scientist by Forbes*

Hal Varian

*Chief Economist
at Google*

and many more!

LEARN INSIGHT IN DATA

<http://www.teamleada.com/>