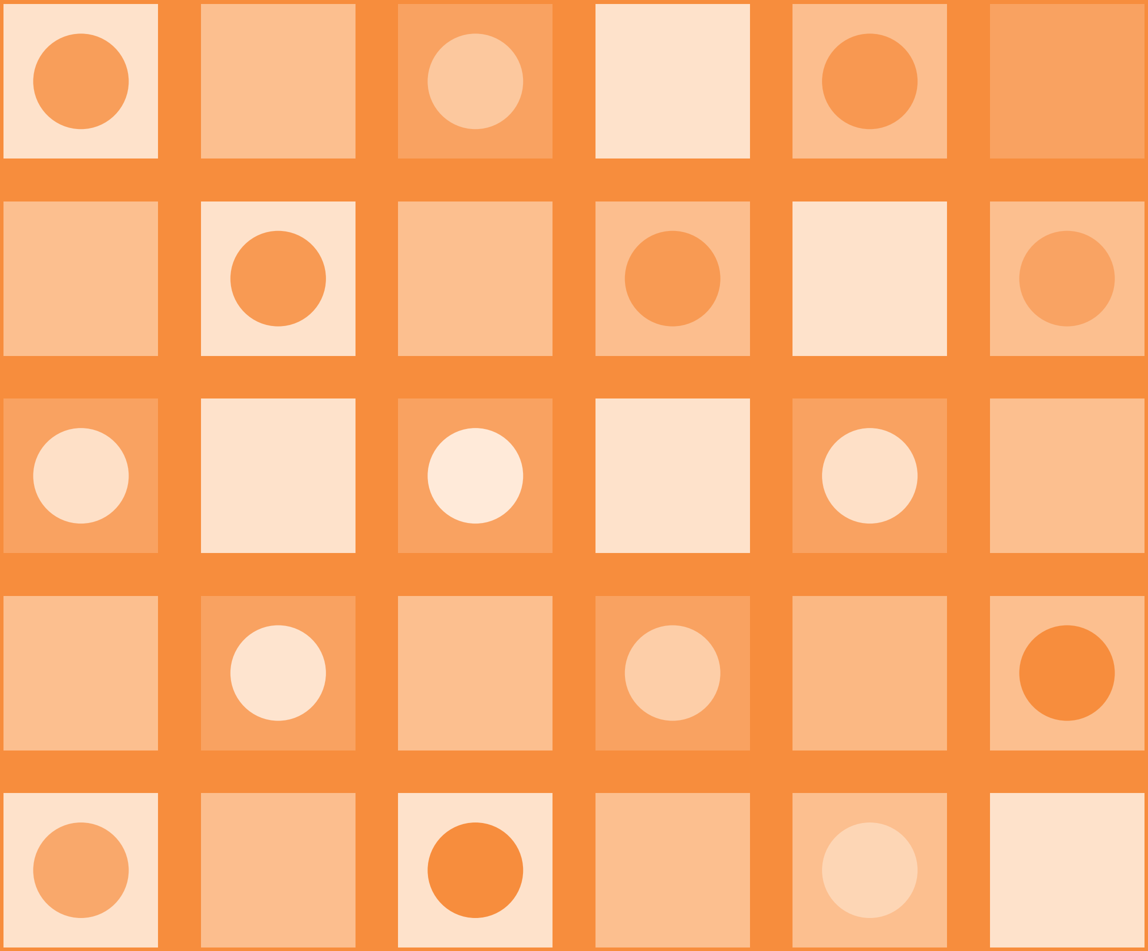


THE DATA ANALYTICS

HANDBOOK

RESEARCHERS + ACADEMICS

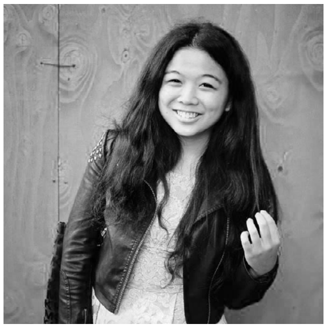
ABOUT THE AUTHORS

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.

When we began writing the Data Analytics Handbook, we set out to understand the role of the data scientist in industry by compiling the knowledge and expertise of over 30 professionals ranging from the data analyst at Yelp to the chief economist at Google. We hope that our work elucidates this burgeoning profession to you, our readers. We hope you now see, as we do, how ambiguous and misleading the phrase "big data" is.

If there is a single takeaway we could distill from our handbook, it would be this – to introduce a new term that we feel should be even more pervasive than the rest: Data Literacy. Data literacy is important because more than words you are interpreting numbers every day. Just as skillful rhetoric can change your beliefs so can skillful analysis. Can you read the morning New York Times without stumbling over a statistic or interpret a visualization? In your work, did you make a decision where you wished you had more data to bolster your confidence or even worse, did you make a decision with data you weren't certain of? How regressive is our education system with its emphasis on calculus rather than statistics given that we use means 10 times more often than integrals?

Data literacy encompasses two abilities: a statistical understanding and the experience of application. It requires an understanding of the basic rules of probability and sampling that is utilized in every experiment. It is also imperative that everyone should have experience manipulating data, whether that's done in Excel or now, more commonly through programming.

Data literacy is a fundamental skill for all professionals, but at its basest level empowering individuals with data literacy is a consumer right. We believe it's so important that we have created an e-Learning application that aims to develop the statistical and programming skills necessary to become data literate professionals. Learn insight in data at www.teamleada.com.

Who is the data literate professional?

You have an understanding of the nuances of statistics. You are familiar with munging with data. You are curious and hypothesis-driven. You are focused on how data can improve or result in action. You are product-focused.

From Tristan, Brian, & Elizabeth

TOP 5 TAKEAWAYS

FROM RESEARCHERS AND ACADEMICS

1. There are wrong questions to ask about data.

Included informally with Type I and Type II error in hypothesis testing should be Type III error. Type III error is asking the wrong questions about data or attempting to discern answers from data that actually isn't available.

2. Data Science is a strategic initiative.

The huge demand for data scientists is a result of companies early investment in Big Data and wanting to get returns from those investments. As more companies invest in Big Data it will result in the strategic recruitment of more data scientists and data science departments.

3. Data professionals must be humble.

Not only are humble people better to work with, but a data literate professional must be humble to its data. He/she must be willing to accept when hypotheses are disproven and be skeptical of results. He/she must recognize that data is the main channel in which users communicate with a company now.

4. Analytics is a basis for competition.

The effective use of data is going to form the basis for competition for every industry in every organization.

5. For data science, learn how to learn.

We are still in the early stages of data science so the tools will constantly evolve, therefore education is a continuing process and should not be tied to any specific tool. As more tools get commercialized, the build, buy, or outsource decision firms must make is impossible to predict so to be competitive become adept at learning new tools.

TABLE OF CONTENTS

05 MICHAEL CHUI
MCKINSEY GLOBAL INSTITUTE

08 PRASANNA TAMBE
NYU STERN SCHOOL OF BUSINESS

11 HAL VARIAN
GOOGLE

14 JIMMY RETZLAFF
UC BERKELEY'S SCHOOL OF INFORMATION

18 DAVID SMITH
REVOLUTION ANALYTICS

23 GREGORY PIATETSKY
KDNUGGETS

26 TIM PIATENKO
COMR.SE

31 TOM DAVENPORT
BABSON COLLEGE

MICHAEL CHUI

SENIOR FELLOW AT MCKINSEY GLOBAL INSTITUTE

MICHAEL is a partner at the McKinsey Global Institute (MGI), McKinsey's business and economics research arm. He leads research on the impact of information technologies and innovation on business, the economy, and society. Michael has led McKinsey research in such areas as Big Data, Web 2.0 and collaboration technologies, and the Internet of Things.

Why did you choose to make your expertise on Big Data?

My role at McKinsey now is to lead the research on the impact of long term technology trends. We do believe that the effective use of data, big data or open data, is going to be incredibly important for the economy. We really think it's going to be a basis for competition going forth in pretty much every sector and for every organization in every company. As a result we thought it would be important to understand this phenomenon and we think it will be huge going forward.

Your report talks about a large short fall of talent and how deep the demand is going to be in the sector. Do you think it is possible that the concept of information revolution is overblown?

No. Broadly speaking, we do believe that effective use of data is going to form the basis of competition for every company in every sector. But we also don't think that the effective use of data is the only thing that matters. Big Data does not substitute for leadership. It does not substitute for creativity. It can help, but it will not substitute. It will not substitute the creation of relationships between people. I also think that some people have questions to whether or not the increase in need for talent will be as urgent or intense as some of us make it out to be. The tools will get better, and then we won't need as much skill. My argument is that there will always be a need to utilize tools effectively. You need to understand how to create experiments for example.

You cite the short fall of demand in talent specifically in the context of Big Data. Do you think the shortfall applies to general data analytic talent as well?

Yes. I think people can fight about the definition of Big Data. I tend to be more liberal with my definition when I include the larger business intelligence technologies. I think some of the more modern avant-garde technologies, whether Hadoop or Spark, or various type of unstructured data analysis is going to be important. But at the same time, I think the more general point about a more effective use of data in all of its forms is going to be important.

Are there any major update since you wrote the report?

We did publish a report in the last few months on Open Data. The basic insight there is that if data is going to be increasingly important, what happens when data becomes more liquid, more shared between the organizations etc. We all know about open data when the government releases the data, but there is a wider phenomenon when data is transferred between companies. There is also the concept of "My Data", when an organization allows an individual or other organization to hold access to that data specifically.

Then how pervasive do you think data literacy will be? Is it your belief that society as a whole will have to become more data literate?

I think that is an excellent statement and I think it is an absolutely true. I joke sometimes because this isn't something I truly believe, but I joke that we should be teaching less calculus. When was the last time someone besides engineers, used integrals outside of the classroom? On the other hand when was the last time, you as an individual citizen, as a shopper, student, or businessperson had to make an important decision based on imprecise dirty data? That is exactly the type of skills you learn in statistics class. We should be teaching people statistics to everyone at a young age. People are really bad at conditional probability. People can't tell the difference between causation and correlation. It's terrible. These are basic skills for us to be better citizens and better employees. Yes I am making a gross generalization, but I really do think that we need to understand data better in all of us.

What kind of general advice would you give an undergrad about to enter the job market?

Nothing beats persistence. Here is a quote by Calvin Coolidge I really like:

> *"Nothing in this world can take the place of persistence. Talent will not: nothing is more common than unsuccessful men with talent. Genius will not; unrewarded genius is almost a proverb. Education will not: the world is full of educated derelicts. Persistence and determination alone are omnipotent."*

It has nothing to do with statistics or machine learning, but it is hard getting a job and it is hard doing well at your job. You will always face expected and unexpected difficulties and there is nothing that beats keep going at it again and again.

PRASANNA TAMBE

ASSISTANT PROFESSOR OF INFORMATION, OPERATIONS, AND MANAGEMENT SCIENCES AT NYU STERN SCHOOL OF BUSINESS

PROFESSOR TAMBE'S research focuses on the economics of IT labor and on how the diffusion of technical skills affects workers, firms, and regions. In recent papers, Professor Tambe has studied the effects of offshoring on the demand for technical skills and how the flow of technical skills within the US economy affects productivity and growth through the diffusion of new IT innovations.

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

There is significant overlap between what we have traditionally called data analysts and what we're now calling data scientists. One change is the scale of the data. The tools being used are different because of the scale. Tools like machine learning and data visualization are becoming more important, almost necessary. Firms also need to understand the context of the data; data scientists are sometimes called storytellers and to do that you need to know what questions to ask.

In your paper, you focus on the shortfall of big data skill specifically; do you think the aspiring data analyst/data scientist should focus on that?

From industry I keep hearing that firms are no longer hiring for skills, they are hiring for the ability to learn new skills. That's different from telling someone to focus on Hadoop. It's useful to get familiar with the current generation of skills because that's what's being used right now but people have to be aware that it's a continuing education process. Especially at this stage of the technology, the tools are going to continue to evolve and the role of the data scientist is going to change quite a bit because these tools are being commercialized, and this changes the build, buy, or outsource decision that a business makes, and consequently the roles of data scientists.

Do you think tools will outgrow analysts, so that analysts are no longer required?

It is going to change the definition of analyst, at end of the day firms are going to integrate data into the decision making process at a fundamental level which means they'll require people who understand analytics. That doesn't mean they need a technical person, but at a minimum it must be a statistical person with analytic intuition and an understanding of the firm and its priorities. So you may need fewer people to build Hadoop clusters, but someone will need the statistical intuition and that may continue to be a scarce commodity. It is also the case that how each company uses data is very idiosyncratic to the firm. Wal-Mart's analysis is different from what a petroleum company's analysis would be so being able to combine that domain expertise with analytic intuition is going to continue to be valuable.

How pervasive do you think this need for statistical intuition and analytics will be?

I think it’s going to be potentially very pervasive and a big shift in the labor market beyond data analysts and data scientists. It could take a while, diffusion processes can take decades. We are already hearing a little bit in quantitative industries such as finance that they don’t want to hire people at any level who don’t know how to work directly with data. You can see this diffusing into other sectors of the economy as statistical intuition becomes a core skill.

So in terms of a team structure, how do you envision data analysts fitting within a company?

A good model for this is sports teams--take the movie Moneyball. Sports teams share a lot with HR groups and many now have data analysts that crunch the numbers and provide recommendations. I think having at least some employees who are comfortable with statistics and analytics and can combine that with expertise related to the firm and industry is going to be a common model.

Can you speak more about the lag in diffusion of the big data skill you reference in your research?

What is occurring is very typical of technology cycles where you have a lot of investment occurring right now in tools, skills, and techniques. A lot of firms are thinking about how to build tools and these tools are going to slowly diffuse out into the “using” sectors of economy, and these firms then have to figure out how to incorporate these tools into their workflow, whether it be retail, manufacturing, or healthcare. When that diffusion begins is when you see the adaptation to specific industry contexts and the innovation that I think you are referencing. We are still really in the first phase. Even so there has been evidence of significant big data driven innovation with retailers such as Wal-Mart.

HAL VARIAN
CHIEF ECONOMIST AT GOOGLE

HAL is Chief Economist at Google. He is also an Emeritus Professor at the School of Information (iSchool), the Haas School of Business, and the Department of Economics at the University of California, Berkeley. From 1995-2002, he served as the founding dean of the Information School.

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

The terminology isn't really standardized. I think that the necessary skills are the ability to manage data using various sorts of databases, analyzing data using various statistical and machine learning tools, document what you did so others can build on it, and communicate what you did effectively. This involves skills in coding, software engineering, statistics, and visualization.

How granular do you see applications of data analytics getting?

In doing data analysis, it's sometimes hard to know when to stop. But ultimately the data is being analyzed because it can contribute to a decision, so at that point you have to stop the analysis and make the decision.

What are the limits of data analytics? At what level of granularity can data analytics not help or just becomes more complicated than it's worth?

In statistics we talk about type I and type II error. But there's also type III error --- asking the wrong question. Sometimes people waste a lot of time trying to answer questions that just can't be answered with the data available.

What industry do you envision data analytics having the largest impact on in the next five years?

I think that we will see the entertainment industry become much more analytic. They've got a lot of data, but they haven't really been using it effectively.

Where do you envision Big Data/Data Science heading in 5 years?

We're going to see a lot more "self optimizing" or "learning systems" that run experiments and improve their performance without any human intervention.

How would you change your advice you undergraduates and young professionals from what you said in 2008 to Freakonomics?

I still think that statistics is the sexy job of the next decade. However, I would urge statistics students to supplement their learning with courses in machine learning, databases, and visualization.

As the market consolidates, tools will change, along with the general analysts skillset. How do you foresee the interaction between future tools and next generation of analysts?

I think that database tools will incorporate more elements of statistics, and statistical tools will incorporate more elements from databases. I think more attention will be paid to experimentation, since that's really the gold standard for causality, and causal inference is critical for decision-making.

You were quoted saying that statisticians are the sexiest jobs of the 21st century, how fundamental do you think an understanding of statistics will be for the average worker in society?

I think that statistics should be a required course in high school. It's not just important for jobs, it's also important for citizens. People need to learn "defensive statistics" and other forms of critical thinking in order to make intelligent decisions both on their jobs and in their life.

JIMMY RETZLAFF

PROFESSOR AT UC BERKELEY'S SCHOOL OF INFORMATION

JIMMY is the Engineering Manager for the Ads Team at Yelp. He was previously an engineer working on the first 3 generations of Kindle at Amazon's Lab126 and has worked at several small startups including his own that developed software for visualizing investment sales activity geographically. He is driven by a desire to create software that improves the professional and private lives of its users.

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

I will answer this in the context of Yelp where I work and we have both now. An analyst is looking at data and manually trying to pull out information from it. Our data scientists are trying to build systems that can automatically pull information from data. Traditionally we've had software engineers who do data science in their jobs; we're now just forming the role of data scientist that is a product management role where they use data science to help inform their product decisions and help prototype things that will eventually be used in production.

An example of this is if there is some desirable metric such as how long do people spend on the site, we want to find certain features that are correlated with that and emphasize those features to boost those metrics. We also ask questions such as can we build it into product features such as if you look for a review of a restaurant we will prompt you to review it from the Yelp.com homepage. Both how we decide which restaurant to show you, because you may have looked at several places, and user interface elements have a huge impact on whether you actually write a review.

I'm trying to understand how the use of data analytics today is termed as "new" because data analysis has existed for decades now, even in business?

I think bringing in the analysis of data deeper into the functions of the business world hasn't been done enough yet. It's definitely started and certainly exists in the Bay Area but has this occurred in Detroit?

So you see data analytics expanding beyond the metropolitan hubs and becoming more available?

I don't think that farmers toiling in the fields will be doing analysis, but his tools might be. There are so many things that we can optimize that has never been optimized before.

Why do you think recent grads should look to work in this field instead of other ones?

We're actually now in a transitional period, where we're learning to use new tools. I think if you ask this question again in 10 years, it won't be a choice. I think it'll be a part of how we think about any significant decision. It'll be pervasive and it will just be an aspect of everything. We are not there right now, so depending on where you are you might end up becoming the advocate for it. By actively going towards it now, you're beating the crowd. If you're going somewhere counter to this, you're going to have a tough time.

What do you look for when you're hiring a data analyst? Are there any specific skills that you look for?

I'm hiring software developers who work with data. They have a feel for statistics and probability and a certain background in math. I'm not always worried about their statistical understanding to be deeply academic. I'm more interested in if they can learn to apply new things quickly. Humility is another big part; I want to work with humble people, but there is also humility in being willing to listen to data. We'll have ideas about what makes for a good advertisement and what makes for a good feature but the data might contradict us. You need to be able to accept that you might have been wrong. At the same time, you need to have a skeptical eye about if the data tells you your wrong is that also true. Data is the main way that users can talk to us now.

Does that mean you ask probability questions or technical data science question? What would the interview process look like?

For me, I actually teach them something, specifically I teach them MapReduce and then have them apply it. It is best if they don't know it coming it. Then I gauge how fast they learn it, how well they learn it, and if they start abstracting and begin thinking about how to apply it. I'll throw in a question like, how do you calculate median in this context. Are you going to be creative? You don't have to have an answer on the spot, but you should have a few ideas.

So do you think a technical level such as R will be required for business people?

Yes, I think more and more it is becoming so. If they don't want to become dependent on a technical person being there to assist them, which will make them at risk of being replaced by that technical person, they should learn R. When push comes to shove, whoever has both technical and business sides will end up with a job. We also have a product management side, where we want people with a business degree, but also a programming background. For example, my current main product manager has a CS degree from Georgetown, but also an MBA from Harvard. People with this combination are hard to find and having this interdisciplinary background makes you inherently valuable.

Can you speak a bit about interests in your class, as compared to where the industry is going?

Again, I think the data mining skills are going to be more and more pervasive. As Excel gets replaced by more advanced statistical methods, people will have to learn new skill sets. This is evidenced by my class, which is formed by I-School students, CS undergrads and MBA candidates. I also have a person doing a post-doc in Chemical Processing. The diverse student body outside of the traditional I-school students is evidence of the pervasiveness of the topic.

DAVID SMITH

CHIEF COMMUNITY OFFICER AT REVOLUTION ANALYTICS

DAVID began as a statistician at the University of Adelaide in Australia. His work in data science first began with the S programming language where he wrote an EMACS interface and co-published (with Dr. Bill Venables) an "Introduction to the S Language" which was widely used at the time. He later began working at Revolution Analytics as the Chief Community Officer.

It seems like you were an early adopter of statistical programming languages. When did you start believing that statistical programming was going to become mainstream?

I think the question is more appropriately: when did statistics become mainstream? For a long time statistics was a back office type of job. You worked as a consultant inside a big pharmaceutical company or financial services company, but it wasn't anything at the front and center. The Big Data movement changed all of that. Companies started to realize that the data they had was valuable, and they should collect it and they should save it and they should analyze it. That's when the data science movement kicked off. I call myself a data scientist now even though I was trained as a statistician. Data science to me means the applied side of the statistics and using computers to do that. That's also when R started taking off on the commercial side after becoming so popular on the academic sector.

How pervasive do you see R being used by young professionals?

R itself has a somewhat steep learning curve so I don't necessarily see business analysts using it. You have to learn a lot of statistics, a programming language, and good data visualization skills to use it effectively. I will say that a lot of companies are hiring a lot more skilled data scientists because its strategic: it's the way they get value out of the investment of collecting Big Data. Secondly businesses are also realizing that what data scientists do is also a really important part of running other parts of the business beyond just research. You've got to put these things into practice. In a financial organization for example it's the algorithms and the methods that the data scientists developed that get used by the traders, the money managers, and analysts. Those professionals are probably using Excel because that is what they are familiar with — they aren't programmers — but the data scientists are embedding their algorithms directly into these spreadsheets. A lot of the time that's driven by R behind the scenes and the business professionals don't realize it.

Where would a young professional who has experience in R fit into that job structure of a company?

I think they'll actually be the data scientists. I think if someone learns R in school, they should branch out beyond the statistical side of things. They should learn about databases, Hadoop, data platforms, and integration into data systems to be a bit more rounded. But I see R users working in a research group or a quantitative analysis group or a data scientist group in a lot of companies today. These groups provide companies advanced statistical analysis through either reports, which is the more traditional route, or more often these days by integrating their work into applications embedded in business intelligence dashboards or websites.

Why would you suggest bolstering your skill in R, as opposed to some other language such as Python/Pandas?

First of all, R is a domain specific language. It is for data scientists. If what you're doing is working with data, building statistical models, and visualizing data R was explicitly designed to do those things. There are also personal considerations. For me, R lets me think with code in such a way that is not possible with any other language. R is really a lingua franca of data science in the same way that English is the lingua franca for business. For example, if someone publishes a research paper, that person is likely to also publish the R code. So when you work in R its not only a language which allows you to think intuitively and collaborate easily, you also have this breadth of the open source community that has created these packages and scripts and tutorials that are ready for you to use.

I saw on the Revolution Analytics website that the R user base was 2 million which was smaller than I would have guessed. What do you think are the impediments of R becoming more mainstream?

I don't think there are impediments, because R is mainstream today! As an open source project anyone can download R from hundreds of website which makes it difficult to track its pervasiveness, but I think 2 million is easily an underestimate. If you look at the most recent Coursera course that used R, I think they had 300,000 registered students. That's just a single course! I think 2 million is a lower bound. In an O'Reilly survey last year, R is now the most used language in data science, other than SQL. It's used by 70% of data miners. R is growing faster than any other data science language. A Dice.com survey also revealed that professionals with R skill had the highest average salaries.

Where do you see the data science field going in 5 years?

I see the demand for the data scientist growing and growing. You have to keep in mind that Revolution Analytics works with companies that have already figured out how to collect Big Data and are looking to tap into its power. Many mainstream companies are now saying, "we need data scientists and a data science department". Then there are a lot of other companies that are behind in that trend and haven't yet figured out their Big Data strategy and haven't gotten to the level that they need data science. Regardless every company is figuring out that if you're not collecting and analyzing data, you're at a competitive disadvantage. Currently there is more demand than supply for data scientists, but more and more universities are coming up with accredited data science programs to try and fill that shortage. I think the supply will catch up to the demand over the next 5 years.

In 5 years, which industry do you think the data science will have the most impact on?

This is a hard question — it's difficult to single out any one industry because Big Data and the need for data science is everywhere. I can say the field is changing rapidly. Traditionally, we've seen statisticians and data scientists in financial services, pharmaceuticals, and to a certain extent manufacturing. But today, there are so many industries where data science is important. Digital media is the one that has seen the most growth over the last few years. Advertising, gaming, and other new-media industries are also getting really deep in data science. As for what is coming up, the media industry, a lot more for manufacturing, and analytics services in general. The services companies will provide other companies the much-needed analytical insights and do so on their behalf.

Can you speak more about the resources that young professionals can use to become data scientist? What do you think is the most affective?

For me, as an employer the skills I look for are certainly statistics, data science, databases, computer systems and so forth — but the thing that I really look for is practical experience. I look for somebody that has actually demonstrated the ability to apply the methods to the data and have gotten the results. Even if it is someone who worked on weekends on data projects or at a non-profit like DataKind, as long as they can show that they know how to apply it, I value that. I have more confidence in them being able to join and start working. Any amount of practical experience you can do. That’s something I look for.

So you’re talking about things such as Kaggle competitions?

Kaggle is certainly one of them. But, though I like Kaggle, I think the competition they run is a bit sterilized; the data you get is prepared, nice and organized. In the real world, it is a lot messier than that.

GREGORY PIATETSKY SHAPIRO

FOUNDER OF KDNUGGETS

GREGORY is the President of KDnuggets, which provides consulting in the areas of business analytics, data mining, data science, and knowledge discovery. Previously, he led data mining and consulting groups at GTE Laboratories, Knowledge Stream Partners, and Xchange. He has extensive experience developing CRM, customer attrition, cross-sell, segmentation and other models for some of the leading banks, insurance companies, and telecommunication companies.

What led you to work in the field of Data Science?

I started in research in artificial intelligence and I was working in a company that was developing databases. I began to think about how I could apply my interest in AI and Machine Learning to databases and that led me to organize the first workshop on knowledge discovery and databases in 1989 (I came up with the term "Knowledge Discovery in Data" which is still popular for research conferences). The workshop eventually became a conference and now we have 20 years of KDD conferences. KDD -1995 was the first conference on topics of Data Mining and Knowledge Discovery.

Why has Data Science & Data Analytics become so popular?

I think people have an intrinsic interest in trying to understand the world. So, we can say data analysis is the "second oldest profession" in the world. What is different now is that there is more and more data and we are developing the tools to analyze all types of data.

The names have changed - from Data Mining to Knowledge Discovery to Predictive Analytics to Big Data to Data Science, but the essence has remained the same - understanding data, finding interesting and actionable information in data. Now there are several hundred conferences on these topics all over the world - see www.kdnuggets.com/meetings/ page.

People who used to call themselves data miners now call themselves data scientists, and there are many more new people who want to be data scientists - which was called "The sexiest profession of 21st century". I think this term is here to stay.

Can you speak more about data analysts in particular?

The name data analyst is used differently and there are several levels of complexity. One level is a person who assembles the data and organizes it into the database. They might more appropriately be called Data Engineers. Another level of data analyst is someone who analyzes it and runs reports - perhaps this person can be called the Business Analyst. Data scientists are people who can build predictive models, use machine learning methods, and try to make sense of the data by asking the right questions. Being a data scientist requires three types of skills: coding, business, and analytical/math/statistics. An ideal Data scientist is this mythical creature who has all three of these skills.

How has the data analyst job changed since its popularity and

how do you expect it to change in the future?

Data analysts now have access to more powerful tools that has created this democratization of analytics. Companies such as BigML, Skytree, Alpine, SiSense, Tableau, and even big companies like Microsoft are producing tools that are easy to use, directed at business users or business analysts rather than PhD scientists. There is also increasing automation in industry. The current meme of "Software eating the world" also applies to data jobs - software tends to eat many of those jobs especially for tasks that can be automated. So there is downward pressure and upward pull on those jobs.

Why do you feel like a recent graduate should work in the field of data science?

I don't think everyone should work in data science. (See also Is Data Scientist the right career path for you? Candid advice)

For a prospective applicant interviewing into a data science team, what should they look for in a data team?

Great question because very often data science skills are hard to find in one person so it's done in teams. You want to make sure the team covers the different skills that are required so I would look for a team that has experience in doing large data installations, engineering the data, presentation skills, implementation and deployment, and security and privacy. My opinion also is that I think advancement in science always creates specialization and what will most likely occur is that the most effective data science results will come from a team of them.

Any final advice for undergraduates?

Learn R and Python and relevant data science packages, and then try competitions. Kaggle is the leading data science competition platform and they offer competitions at all levels from beginners to experts. It's a very effective way to learn and gain experience.

TIM PIATENKO
CTO OF COMR.SE

TIM holds a BS in Mathematics from Cornell as well as a PhD in Physics from California Institute of Technology. Tim made the transition into the industry after finishing his PhD and doing research at Stanford Linear Accelerator Center. He worked at companies including like.com, eBay, Zynga, and Badgeville. Tim currently serves as the CTO at comr.se.

Can you talk a little about your background in Research?

I started in Elementary Particle Physics at Caltech, moving to the Stanford Linear Accelerator Center (SLAC) to work on my thesis. The project I was working on — BaBar — was in the later part of its life cycle. This means the main particle searches had been done, and the new batch of students had to look for more rare events, thus increasing the need for more sophisticated analysis tools. Therefore I ended up building various machine-learning models.

Can you talk about how you transitioned to Analytics/ Statistics?

My first job, which I actually transitioned into gradually as I was wrapping up my thesis, was at a Comparison Shopping Engine called Like.com. I learned to work with SQL databases and started learning my way around Python and R — the tools that would be the basis for my toolbox going forward. While the techniques I learned in Particle Physics translated pretty well, the modern tools were very refreshing! I quickly realized that I liked working with people data more than particle data. On top of that, you don't need to be nearly as precise as you do in hard science. Also I didn't have to worry about peer reviews!

After that, I went to eBay, where I worked in Site Analytics, trying to understand how people interacted with the search pages. We ran a ton of A/B tests and wrote a lot of SQL code. However, the problems all seemed a bit too confined. The linear path through a traditional website only get you so far. So my next gig at Zynga was quite an eye opener from the data perspective. The nonlinear nature of a game produces variety and complexity that rivals those of elementary particle decays. That led me to Badgeville, where it was absolutely essential to understand the elements of human psychology. I learned how data collection and storage should be tailored to the design of the product you are trying to build, or else you can end up with the wrong data to analyze.

Though I've worked in different places, the concepts I learned back at SLAC are still the foundation for everything I do. What has changed over time is my appreciation for domain expertise, without which the real world business problem cannot be properly tackled. Analysis should be tailored to the problem you are trying to solve. My motto has always been that the data you've collected can only tell you what you did, not what you could have done. To know what would work better, you need to try other things; only the domain expertise (and sometimes an open mind) can guide you to try the right ones.

I saw that you instructed a few internal trainings within your organization; I know a lot of skills are learned on the job. How else can one acquire the necessary skills?

Having a background in statistics is important, of course. I was lucky in that I had time during grad school to learn at my own pace. I also had smart people around me to learn from. If you took stats classes in college, dig up your textbook or go on Wikipedia to refresh your memory. If you never did, consider taking one online. I've never tried any of the more advanced online machine-learning/data-mining courses, but I imagine there are some really good practical ones as well.

But most importantly, I think you need to lean more about your field. When you do get a job in analytics, don't just concentrate on learning R, Python, running regression models, SQL queries… learn the business problems. Absorb as much as you can. Change jobs (or roles in your company) to get exposed to new settings with new approaches. It's the ability to continuously learn and adapt that matters the most - not any of your particular skills. But having a foundation helps a lot.

Based on your experience, how would you define the role of a data analyst? What about a data scientist? Is it possible to start out as one and make transition?

An analyst is someone who is given predefined problems to solve with a relatively well-defined set of tools. An analyst fits into an organization and performs a set of tasks. An analyst refines his skills over time. An analyst should excel at what he does, but what he does may be pretty narrow. A scientist is someone who finds the problem to solve/methods to solve it. A scientist doesn't necessarily fit into an existing org structure, but augments it. A scientist continuously picks up new skills and could make big transitions from one project to the next. A scientist is good at a lot of things (i.e. diverse skillset), but may not be the very best at something specific. I suppose a transition can be made, though I think it also depends on the candidate's characteristics, the mindset, and the academic background. A combination of those factors can make the transition pretty difficult.

Since you've been working in the field for so long, how do you think the field has changed over the last few years, and what do you think will happen over the next few?

Like everywhere else in the industry, lines are becoming more blurred all the time. Data Science is a pretty nebulous term these days; five years ago it wasn't even really a term in use. I think as we get through the craziness of this Big Data era and people begin successfully deriving true values from the data, we'll end up with better-defined job functions and titles. Right now, it's the Wild West era for Data; Five years ago everyone seemed to know how to analyze Web data. Now, no one does! In five years, I hope we'll know once again.

Given that the whole field is segmented and the technology is quickly changing, how should one adapt? How can one keep up with the new tools that are rolling our every day?

You can't keep up, so don't try too hard! Build a solid foundation, and you'll learn the next big thing when you need to. If you know R, SQL, some basic programming, and maybe be familiar with map-reduce framework, you'll be fine for a while. If you understand the technology behind the scene, it's even better. Just assume there will be a new kid on the block tomorrow, and everyone will be asking you when you'll switch over. Give it a shot when you get a chance, but don't go chasing your tail. SQL is still here. R will be around for a long time. Statistics books from 100 years ago still speak the truth, and Tufte's book from 1983 is still the best read for conveying information.

Some say Data Science requires 3 skills: programming, statistics, and domain expertise. Which one would you say is most important?

If you don't know basic statistics, you can't be a data analyst or scientist. So that's a given. Programming can be learned, though preferably in some formal way, and the earlier the better. Domain expertise is really what provides the value in the end. Otherwise, you're just a replaceable human calculator. By itself, it's useless. There's nothing you can do to add value, even if you know how. But without it, you aren't a unique asset. Trainings and skills let you move forward. Domain expertise lets you excel.

SPECIAL TOPICS ON GAMIFICATION

Can you tell me a bit about gamification?

Gamification is the idea of bringing elements from traditional and social games into other products and services, as well as workflows. It's a combination of sociology, psychology, design, and data analysis/optimization. You start with some basic drivers, such as competition/collaboration/achievement/social validation, and see which ones apply to your current problem. For example, sales teams like being competitive, so you might give them a leaderboard. On the other hand, a food critic may be posting restaurant reviews to gather the largest loyal following possible, combining personal achievement (of being the most prolific content generator) with the social component (of being validated by the readers) might work well. Classic examples of gamification are Foursquare, Yelp, and Stack Overflow.

How does it work?

Think Casinos. You want the bulk of your users to follow a well-thought-out pattern that is beneficial to you, while being fun for them. It's Behavior Management that we talked about at Badgeville that has been in practice for quite some time. First figure out what scenario you are in, so you get the design right and the flow is appropriate. Then implement the data acquisition system to match, so you're collecting the most relevant pieces of data. Make sure the data is processed and analyzed in a timely matter, so the findings can be applied before the experience becomes stale. Keep things as simple as you can for as long as you can. This isn't high-frequency trading, so you don't need to solve crazy differential equations. But you're dealing with human beings data, which require special attention.

Why is it better than the traditional approach?

It views your users as people. They're not treated as just random generators of stochastic signals (though they are that too, of course!) — you need to understand what drives them and how you can drive them better. I've also argued in the past that if done correctly (in terms of data collection and segmentation that match what you're trying to accomplish) elements of gamification can introduce self-ordering into your system that makes data much more digestible. This requires a significant amount of work upfront, but pays off in the long term when your users become self-labeling. All you have to do then is to follow a manageable number of distributions over time and see how their shapes change.

TOM DAVENPORT

PROFESSOR IN MANAGEMENT AND INFORMATION TECHNOLOGY AT BABSON COLLEGE

TOM has written some of the most influential and popular books and articles on analytics and big data in business. He is the President's Distinguished Professor in Management and Information Technology at Babson College, co-founder of the International Institute for Analytics, a Fellow of the MIT Center for Digital Business, and a Senior Advisor to Deloitte Analytics. Tom earned a Ph.D. from Harvard University in social science. Tom has been named one of 10 "Masters of the New Economy" by *CIO Magazine*. He recently published Big *Data at Work: Dispelling the Myths, Uncovering the Opportunities.*

How would you differentiate the use of data analytics in the past (in the form of business intelligence) with data science?

I recently published a book, Big Data at Work, where I actually talk about how at one point I didn't think that there was any difference between "Big Data" and data analytics. As I did more research though I found there to be significant differences. I would say that the analysis process is not that different; the statistics are still the same. People in the "Big Data" world are more interested in visual analytics than what is historically the case in traditional analytics, which I found to be interesting.

What is different is all of the structuring activities that you have to do with "Big Data" to put it to a form in which it can be analyzed. "Big Data" comes in text, video, server logs, and all other types of formats, so if you want to do any analysis you have to spend a lot of time and effort putting it into structured form. These new data sources also allow you to augment your analysis and your models in ways that were not previously possible.

Another difference I would say is the broad use of open source tools for structuring data like Hadoop, Pig, and Hive among data scientists. They are very likely, particularly in startup companies, to use these tools, and that tends to mean they use open-source statistical tools too—R in particular. Some of them also like open-source data management and machine learning tools.

How would you define the role of a data scientist and contrast that role from what a business analytics person might do?

People who are data scientists typically do analytical work as well. The complicating factor when defining roles is that people who are traditionally quantitative analysts are now calling themselves data scientists. Real data science I would say has a very high component of data management and data structuring related activities. When I interviewed a number of data scientists for an HBR article Data Scientist: The Sexiest Job of the 21st Century, I wrote that you could almost describe these guys as data plumbers rather than data scientists because they would talk about how they would really have to get their hands dirty with the data which included not very highly analytical activities. Typically most statisticians have not been very interested in those types of activities, they would rather work with more structured data to begin with.

There is a McKinsey Big Data report that cites a 1.5M shortfall in "data-savvy managers" in the US. Where is this shortage

coming from?

I think the demand for data talent has come from two sources. First using analytics in business was already on a roll before “Big Data” came along so companies were already attempting to bolster their analytics teams. Secondly this data science movement has only added to the demand for analytic talent and the reality is that everything is being lumped together in these generic terms. There is a website called icrunchdata.com that aggregates the online listings for “Big Data” jobs and it is an impressive number—510,000 jobs at last count. I asked the author of this website a couple weeks ago what categories does he include in this index and he said “data scientists, data analysts, business intelligence, SAS, and a lot more.” What that suggests is that this site, and other sites, are using “Big Data” or “Data Science” as the broad category for what used to be called business analytics, and before that business intelligence.

What is your opinion on the masters programs in analytics that are being initiated?

There was a study last year that counted over 100 programs in the U.S. alone that were either master degrees or concentrations in BI, analytics, or data science. It’s really only the startup community that is demanding a data scientist rather than someone who knows about analytics and data management in general. So I would jokingly say that if you are in San Francisco, you want a masters in data science and if you are elsewhere, a masters in analytics. The earliest programs like the NC State one, which I think is one of the best, are slowly adding more data science and big data concepts and skills. In general I would say to attain these top jobs you are going to require some sort of masters degree. The contents of these programs have begun to slowly specialize, for example at the University of Tennessee they have a concentration in supply chain analytics and then there are those that are more machine learning oriented. Unfortunately I think it’s impossible in any one of these programs to include all of the content you need to know about.

What is your opinion on the need for increasing statistical literacy for society today?

People have been saying that for a while and I certainly think its true. There is both the need for the 1.5M data-savvy managers and analysts but also a need for consumers to make intelligent decision-making. An example of this is that a consumer of medical information needs to know the difference between correlation and causation. This need has been going on for a while and unfortunately statistics courses traditionally have been poorly taught in most schools. Xiao-Li Meng a professor of statistics at Harvard was able to expand the interest in studying statistics by changing the curriculum to be more applied and fun, and I think we need a lot more academics that can open up the world of statistics to a broader audience. I think most academics don't really know enough about how statistics are applied in the workplace to be able to create high quality content for teaching purposes.

Any last advice to young professionals about developing their careers?

I think its really good idea to focus on being more interdisciplinary in your education. What employers want is not how a statistic is derived but how can you use it to solve business problems. Unfortunately if you don't get that in statistics classes you have to create that by double majoring or taking classes in multiple departments. The combination of business and statistics, or statistics and medicine is really valuable. Secondly in the report I previously mentioned, 300 employers were asked what capabilities were they looking for in terms of hires, and the one most demanded was the ability to communicate statistics and analytical results effectively. Therefore you should look for opportunities to try to present your ideas and your results in a form that non-statistically literate people can understand and act on them.

Special Thank you to All of our Interviewees

Peter Harrington
HG Data

Santiago Cortes
HG Data

Ben Bregman
Facebook

John Yeung
Flurry

Josh Wills
Cloudera

Leon Rudyak
Yelp

Abraham Cabangbang
LinkedIn

Greg Lamp
Yhat

Rohan Deuskar
Stylitics

Mary Ellen Gorden
Flurry

Derek Steer
Mode Analytics

Mike Olson
Cloudera

Dean Abbott
Abbott Analytics

Tom Wheeler
Cloudera

Dave Gerster
BigML

Ali Syed
Persontyle

Prasanna Tambe
NYU

Gregory Piatetsky
KDNuggets

Michael Chui
McKinsey

Jimmy Retzlaff
Cal

Hal Varian
Google

Tim Piatenko
comr.se

David Smith
Revolution Analytics

Tom Davenport
HBS

LEARN INSIGHT IN DATA
http://www.teamleada.com/