

Paper SD100

How to be a Data Scientist Using SAS®

Charles Kincaid, Experis Business Analytics, Kalamazoo, MI

ABSTRACT

The role of the Data Scientist is the viral job description of the decade. And like LOLcats, there are many types of Data Scientists. What is this new role? Who is hiring them? What do they do? What skills are required to do their job? What does this mean for the SAS programmer and the statistician? Are they obsolete? And finally, if I am a SAS user, how can I become a Data Scientist? Come learn about this “job of the future” and what you can do to be part of it.

INTRODUCTION

When I was in 7th grade I told our newspaper/yearbook that I wanted to be a scientist when I grew up. Since then my career took me through priest, spy, president, nuclear engineering, nuclear physics, computer science, and finally statistics. Then I look at what I did as a statistician and think that I was not far off in 7th grade. I made important contributions to scientific research. Now there is this title out there called *data scientist* and it brings up questions. What does the science of data (dataology?) entail? Am I not as much of a scientist as I thought? Over the years, I have been following the discussion, often fan-boy-ish, about data science and started to consider what I do, what the people that work for me do, what companies are looking for in light of this hot new topic.

In this paper, we will analyze the data about data scientists and draw conclusions about this role. We will first define what a data scientist is by looking briefly at the history of the term, the way people define it and three of the more important components of the job. Then we will see what the market for a data scientist looks like by understanding what companies want them, what they are looking for in the data scientist and what they do in their role. Finally, we will see how you as a SAS programmer or statistician compare to a data scientist and what you can do to jump on this bandwagon.

The first two topics – what a data scientist is and what the market is like – will not address SAS itself. The author’s perspective is that the reader’s challenge in understanding how to be a data scientist using SAS is uncertainty about what a data scientist is and what companies are looking for. Therefore, this paper will focus first on those aspects and then discuss the steps that can be taken for the SAS user to reposition themselves in the marketplace.

WHAT IS A DATA SCIENTIST?

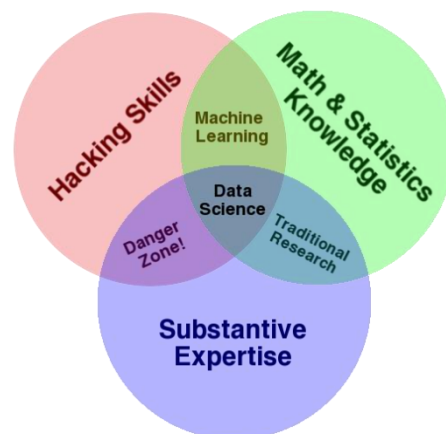
A BRIEF HISTORY

What is Data Science and where does the idea come from? The term is actually somewhat ambiguous with definitions depending on the context. Many people want data scientists, but sometimes they do not really know what it is they want. The term came into use about the turn of the century, but the ideas and sources have been around before that. As we will see, some particular changes in the last decade have made data scientists possible and exciting.

A poster on quora.com says it “is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.” A bit more rigorously, Mark Pitts from the *All Analytics Academy* recently defines the data scientist as

A person with advanced expertise and experience in at least these three disciplines: Information Technology, Business, and Advanced Analytics, and who applies this multidisciplinary skillset to solve business problems and create new approaches as needed. (Pitts, 2013)

Drew Conway is a leading expert in the application of computational methods to social and behavioral problems at large-scale. Similar to Pitts, Conway defines Data Science as the intersection of Hacking Skills, Math & Statistics Knowledge and Substantive Expertise (i.e. business knowledge). (Conway, 2010) We will cover each of these areas when we discuss the job market.



Going back a little earlier to 2008, Jeff Hammerbacher (then at Facebook) and D.J. Patil (then at LinkedIn) coined the term data scientist to describe the kind of work they had been doing (Davenport & Patil, 2012), but that wasn't the original source of the idea.

The phrase "Data Science" had been around before that and the concepts even earlier. Bill Cleveland, for example published "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" in 2001. (Cleveland, 2001) Even before that, 1998, Gerry Hahn and Roger Hoerl, while at GE R&D, published "Key Challenges for Statisticians in Business and Industry," in which they discuss the transitions happening in the field of Statistics. Hahn and Hoerl describe the following challenges:

- statistics done by non-Statisticians,
- the growth of Statistics into new areas such as healthcare and financial applications,
- greater expectations by management for statisticians to "be responsive and vital to today's business needs and to be able to prove their contributions quantitatively,"
- the requirements of analyses to be timely as well as appropriate,
- the need to work with immense databases, and
- adapting to new forms of communication. (Hahn & Hoerl, 1998)

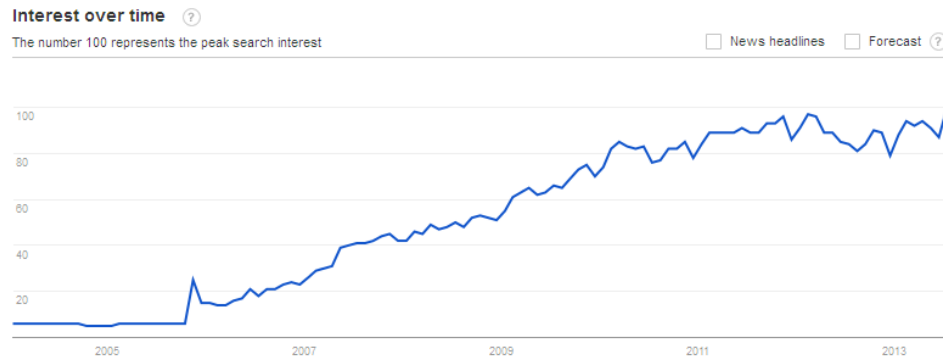
As you can see, these challenges are pre-cursors to the idea behind the data scientist. One might ask then, whether the data scientist is different from the statistician described by Hahn and Hoerl? In many ways, it is not. The reason that the data scientist has branched off is that most statisticians have stayed limited in their approach because they have traditional roles in traditional companies doing traditional statistics. It is the Facebooks and Amazons and Googles that have made the role so "sexy," because those companies themselves are sexy. As dotcoms, they also have a clearer understanding of the value of their data. They have so much of it and it is staring them in the face. They do not have their normal business to distract them. It IS their business.

Three dimensions have developed to a point that is making data scientists possible and exciting. We see all three in these companies just mentioned: Analytics, Big Data and High Performance Computing.

ANALYTICS

Statistics as a tool has been used by people since the 18th Century. It has long been recognized in the field that when practicing statistics, just knowing the theory or being able to crunch the numbers is not enough. Knowing the discipline being analyzed and being able to present the results in a way that is meaningful and relevant to the client are important parts of statistical consulting.

The use of the term analytics began much more recently and like data science it can be ambiguous. As the Google Trends chart shows, it started to be a word of interest back in 2006, which is right before Thomas Davenport's book, *Competing on Analytics* came out.



Analytics is a word that describes a collection of skills or tasks to be accomplished. Evan Stubbs differentiates it from *business analytics*. According to him, “Analytics can be considered any data-driven process that provides insight.” However, insight by itself is not enough. Analytics have to be relevant to the business and provide measurable value in order to be business analytics. In that sense, his definition of business analytics is not far from data science. In fact, the best data scientist is the one who not only derives insights, but also creates real value for the business. (Stubbs, 2013)

In terms of the statistical techniques, its meaning depends on the sophistication of the company. Analytics can mean anything from basic summarization and simple searches to predictive models and optimization.

BIG DATA

Big Data is a very hot topic, as well. In fact the two – Big Data and data scientists – almost go hand in hand, as it is Big Data that has done the most for heralding in the data scientist role. All that data is sitting around and people do not know what to do with it. They see the data scientist as the one who will come in and take advantage of it.

Big Data also has many meanings. It is not just having large amounts of data (Volume), though that is the first place the term was used. The dotcoms such as Google, Facebook, and eBay collect terabytes of data weekly. The term has also come to mean any source of data that requires greater computing power. That includes data that gets refreshed daily, hourly or even every minute (Velocity) or data that does not have one consistent structure (Variety).

But these are not the only characteristics of the data sources that a data scientist would use. The data scientist is expected to get creative in finding sources of data. They may pull data from the web or from multiple departments within the organization. They are also expected to be innovative in how they see “old” data so that it can be leveraged to new advantage for the company.

HIGH PERFORMANCE COMPUTING

Finally, the infrastructure that brings these pieces – Data and Analysis – together is computing power. As each increase in computing capabilities occurred, the kinds of questions that could be answered and the methods used to answer the questions grew. We went from pencil and paper to manual calculating machines; calculating machines to mainframes; mainframes to desktop computers; and desktop back to grids of servers with ever faster processors and more memory and more space.

This increase in computing power has helped the problem and exacerbated it. We can store and analyze more and more data, but these same computers have also been the primary cause for generating so much data.

DATA SCIENTIST

WHO IS HIRING?

There has been great excitement about analytics, big data and data science within organizations. Many companies across the spectrum of industries are looking for data scientists. There are even companies who do not really know what Big Data and data scientists do, but they know they want it.

Facebook might be the typical company we think of when we think of hiring for data scientist. The dotcom companies have the entrepreneurship, the passion for adventure and the “just do it” attitude. “Here’s a bunch of data. Go figure out something with it.” This Christopher Columbus type mentality is often tied to data scientists, and they had the data to do this. If your data fits into a spreadsheet, then there is probably not much mystery to it. Facebook, Google, and LinkedIn were the Big Data hotshots. They developed and popularized the computing capabilities and software to handle the analysis of large data sets. They certainly hire data scientists, even before they knew whom it was they were hiring. DJ Patil from LinkedIn said

[W]e realized that as our organizations grew we both had to figure out what to call the people on our teams. “Business analyst” seemed too limiting. “Data analyst” was a contender, but we felt that title might limit what people could do.... [T]he focus of our teams was to work on data applications that would have an immediate and massive impact on the business. The term that seemed to fit best was data scientist: those who use both data and science to create something new. (Patil, 2011)

However, they really were not the first and they are not the only companies with Big Data. Telecom, for example, as well as retail and financial companies also have big data and some, like AT&T have been dealing with Big Data for decades. They are also hiring.

We know that the government has a Big Data Research and Development Initiative. Their purpose for the initiative is “to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning.” They have the same focus on extracting “knowledge and insights from large and complex collections of digital data.” (Kalil, 2012) Many of their data scientists come through consulting companies like Booz Allen Hamilton.

NASA has also had Big Data for many years. Their major categories of data are Satellite Remote Sensing Data, Assimilated Datasets (Validation Data), Model Output, and Climate Projections. (Pryor, 2012)

Startups are looking for data scientist-types, but it is one of the 5 most difficult positions to fill according to Mashable. (Hartwig, 2013)

Other companies are looking for data scientists, too, not always with Big Data, but wanting to hire data scientists to help them take advantage of the data they do have.

In a recent survey of one job posting website the employers looking for data scientists range from traditional dotcoms – Amazon and Facebook – to new startups – Boxfish – to creators of video games – Electronic Arts – and even to traditional employers – American Express, Walt Disney World, and The New York Times Company- who are trying to be analytically competitive. (Unless otherwise stated quotes come from the survey of these job postings.)

As can be seen, companies have to be highly competitive in order to attract the best and the brightest. They try to sell their company as the employer of choice for data scientists with phrases such as the following...

The opportunity is to participate as a member of a small team with a simple objective. Given one of the largest data sets in the world Innovate!

-- AT&T

This is a high impact role in a talented and close knit team... The ideal candidate should have one hand on the white-board writing equations and one hand on the keyboard writing code.

-- Amazon

Do you have a passion for creating data-driven solutions to the world's most difficult problems?

-- CIA

It means more opportunities to unleash your creative genius, be inspired by those around you and ignite your path in any direction you choose.

-- Electronic Arts

The ideal candidate is an unabashed data geek. You enjoy searching the Internet for datasets that you can explore and mashup to tell interesting stories.

-- MaxPoint

The LifeLock Product team sets strategy, owns their numbers, and builds innovative products that consumers and professionals don't just use, but love. They thrive on data and on the opportunity to push the boundaries of web and mobile products...

-- LifeLock

You may be a great fit for our team if you are excited about working on high impact, real world problems using huge (sometimes slightly messy) data sets, including billions of transactions, to unlock valuable insights and power new products.

-- The New York Times Company

If Nate Silver regularly seeks you out for forecasting advice, please apply.

-- Chegg

It is clear that these companies are looking for more than just a standard number cruncher or programmer. They realize that great opportunities are possible for the company that leverages their data for competitive advantage.

This excitement has created an apparent lack of talent as described by the McKinsey Global Institute. They find that "In the United States alone ... the demand for people with the deep analytical skills in big data (including machine learning and advanced statistical analysis) could outstrip current projections of supply by 50 to 60 percent." (Brown, Chui, & Manyika, 2011) Many companies are looking for data scientists even if they do not really understand what they do or can actually take advantage of their skills. That growth has been phenomenal according to Wanted Analytics, who tracks job market trends. They indicate that since April 2009, data analysis positions have gone up 246% and there are now over 31,000 online openings.

If we focus on Big Data skills, then there are about 8,000 jobs openings with 15 candidates for each job ad. Nationwide hiring, though, scores 75 out of 99, with 99 being the most difficult, indicating that it is very difficult to recruit for Big Data positions. (Rowe, 2013)



There are many candidates for these positions, but as we can see, it is also difficult to find good talent. Some candidates are only excited by these positions and want to be part of the buzz around them. Many will have some of the skills, but lack one dimension or another, which brings us to the next question.

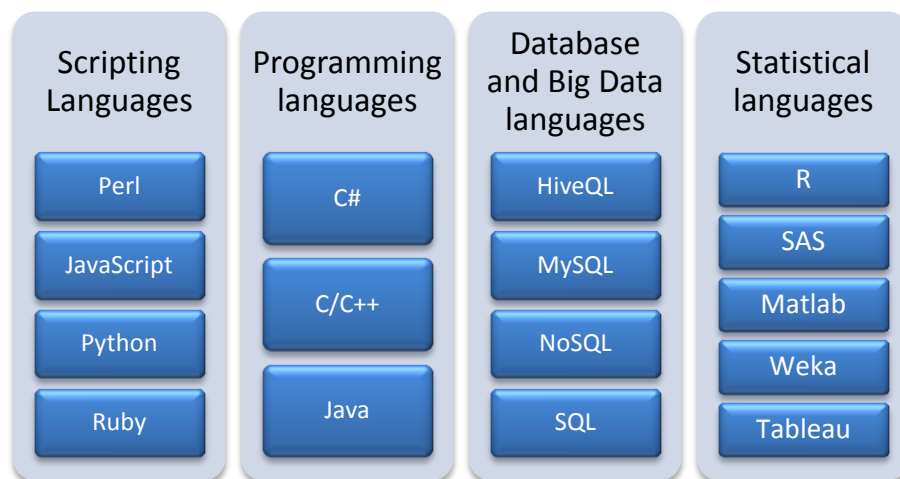
WHAT SKILLS ARE EMPLOYERS LOOKING FOR?

As described above, data scientists are seen as a combination of computer programming, analytics and business skills. Many of the applicants out there have expertise in one or two of these areas, but not all. Each company emphasizes the aspect of these skills that is more important to them. Some positions require strong programming and others stronger statistics. In the job site survey, the requirements were

categorized into the general areas listed below. The top five skills focused on the most were Analytics-related and Soft skills (about 22% each), Programming and Big Data (both about 15%) and Communication (as its own category ~8%).

Programming

There are multiple places that programming, called hacking by Conway, is necessary in the lifecycle of a data scientist's work. These include getting the data from sources that include RDBMS, Hadoop, or the internet, cleaning the data, performing the statistical analysis and then presenting the results. Companies are looking for expertise with the following to perform these tasks.



Not all of these are requested by every company. Some will allow the data scientist the flexibility to use the languages they prefer. Many of them are freely available and, after all, the data scientist must be kept happy. Which database language is required depends on the environment in which the work is to be performed. SQL can be translated from one SQL platform to another, but NoSQL works differently. In addition, most companies realize that when a candidate knows one scripting language, for example, then they can more easily learn others.

Some companies wanted strong software developers, for example candidates with a “deep understanding of the orders of algorithms and their scaling behaviors.” These companies have the understanding that realizing full value from insights requires operationalizing the analytical results so that they are robust and repeatable.

Other companies were looking for only mid-level programmers, saying, “You do not have to be able to write highly scalable production code (you will be working with engineers who can do that) but you do need to be able to build prototypes and be able to understand the existing code base.”

We might make mention of why Conway uses the term “hacking” to describe these skills. We often think of hacking as related to computer security, but it was also used to describe “people who use computers for fun” (Steele Jr., Woods, Finkel, Crispin, Stallman, & Goodfellow). Richard Stallman takes it even further with a description that could almost describe a data scientist. (Hacker (programmer subculture))

What they had in common was mainly love of excellence and programming. They wanted to make their programs that they used be as good as they could. They also wanted to make them do neat things. They wanted to be able to do something in a more exciting way than anyone believed possible and show "Look how wonderful this is. I bet you didn't believe this could be done."

Analytics

Analytics includes a wide variety of techniques that have been developed in multiple disciplines. Sometimes companies are not specific in the types of techniques they are looking for, whether because

they do not know what is needed or because they do not want to limit the candidates is not known, but can sometimes be guessed. For example, Facebook states that they want someone with “[e]xtensive experience solving analytical problems using quantitative approaches.” Their emphasis is more on Big Data techniques.

The common analytical techniques that are requested are statistical modeling, data mining, machine learning, and natural language processing. Statistical modeling could include time series forecasting, regression, and hypothesis testing based on experimental design. Data mining “is an interdisciplinary field at the intersection of artificial intelligence, machine learning, statistics, and database system.” (Chakrabarti, et al., 2011) Its purpose is to extract knowledge from large databases and can include Bayesian methods, neural networks, association rule mining and segmentation. The field has extended its reach to include many of the technical aspects of Data Science.

Machine learning is a branch of artificial intelligence. As we have seen so far, many of these techniques overlap one another. For example, this branch includes decision trees, association rules, neural networks and clustering techniques, and is thought to be part of data mining.

Natural language processing is used to turn unstructured data into structured data. It is often based on the statistical machine learning techniques already mentioned.

The emphasis around analytics varies from posting to posting, likely indicating the level of analytical maturity of the company. For example, Walt Disney World is a leader in analytics, hosting an analytics summit each year. Their job posting has more advanced requirements than others in the area of analytics. For example, they look for someone with the “[a]bility to develop advanced analytical data mining, forecasting, and optimization models.” It is rare for companies to value the full capabilities of optimization models. They also look for the “[a]bility to design and conduct studies for comparison and benefit assessment of various advanced analytical models.” Many companies will perform simple experiments such as A/B Testing, but advanced experimental design shows a greater degree of analytical sophistication.

Lesser emphasis could also indicate that the data scientist is on a team that includes members with higher analytical skills. They may be working with statisticians or econometricians and need to know what they are doing, but do not have to come up with the new, advanced techniques themselves.

Business Knowledge and Soft Skills

It is important for a data scientist to understand that they have to provide value to the business. They cannot ignore the bottom line and still be successful. In fact, companies hire for people who specifically know the importance of adding that kind of value and want to do so. They have to have the “[a]bility to understand business problems from a high level” and to “drive key business results.” They not only have to think well, but also need “an understanding of business research and analysis.” As we saw above, these positions are considered “high impact” that “drive new and innovative ideas to our business units.”

To meet this requirement the candidate does not need the specific subject matter expertise as much as they need the attitude and desire behind the requirement. Companies realize that their candidates may not come from the same type of company or the same type of industry. Though that is often desired, many companies want to do things that no one has done before. A smart hiring company welcomes cross-pollination when they hire someone who has been successful in similar jobs outside of the industry.

One of the more common words seen in the job postings is “passion.” Think of the dotcoms and startup companies. Since they are likely to be high stress and low pay, a person’s passion keeps them going. As above, companies often want their analytics groups to be like startups (maybe not the low pay part, though) with creativity and innovation. A person with a passion for data and analytics is going to make strong contributions to their company.

Communication is another of the business / soft skills that companies are looking for. In fact, almost all companies are looking for it as part of their open positions. It is the #1 most commonly required skill in online job ads according to Wanted Analytics. (Lombardi, 2013) The data scientist works on a “high-involvement, collaborative team” that includes the business, IT, Hadoop developers, business analysts, software engineers, product managers, and subject matter experts in their industry. The data scientist must be able to work well in the team and communicate advanced analytical concepts in ways that others

on the team can understand. They must also be part of presentations of their insights and data products to the business and upper management. Often these positions provide for attendance and presentations at external events, such as conferences and forums, as well.

It seems obvious that companies would require problem solving from their data scientists and so they do. It is the #4 most commonly required skill in online job ads, so companies realize that everyone's job is about solving problems in one way or another. In this case, it is the definition of the data scientist to solve business problems.

Degrees

Because the data scientist is expected to bring advanced analytics to solve the business problems, companies specify a PhD as the minimum education almost 2 to 1 over Bachelor or Master degree. The field is often flexible, though, and must be. Strong candidates do not only come out of Statistics, but also Computer Science, Physics, Engineering and Economics. These disciplines are quantitative in nature and have overlap in the analytical and computational methods they employ. However, some companies do not need a PhD and want someone with either a Bachelor or Master degree. Some will even accept "equivalent level of experience" rather than a degree.

There are starting to be degree programs specifically for analytics or data scientists, such as the M.S. in Analytics from the Institute for Advance Analytics at North Carolina State University. Coursera offers a course in Data Science from the University of Washington. (By the way, Coursera is looking for an Analytics Engineer.) Other universities, such as Oklahoma State University and Berkeley, also have analytics and data science programs or courses.

WHAT RESPONSIBILITIES DOES THE DATA SCIENTIST HAVE?

We have already discussed some of the responsibilities of the data scientist: innovate, work collaboratively on a team. The primary responsibility of the data scientist is seen in the definition above, *"applies this multidisciplinary skillset to solve business problems and create new approaches."* The other responsibilities are in support of this primary one, which is not as simple as it might sound. The programming languages are a required skill set because they are responsible to "collect, explore and identify the right data to be analyzed from internal and external sources." The sources could be

- internal RDBMS platform, usually more than one
- Hadoop, Teradata, EMC or another Big Data platform
- unstructured data from a call center or from Twitter
- websites that need to be scraped
- real-time data feeds

The collection process also requires cleaning (very few data sets are not messy), transforming, and coding as any ETL process would. Even before that, this task often requires finding new internal and external data sources, which may require the data scientist to think creatively about what could be considered data and to think of innovative ways to use data they already have.

Once the data is in hand, or in Hadoop, the data scientist has the responsibility to mine the data, understand "complex data attributes and constraints," and "[c]reate informative visualizations that intuitively display large amounts of data and/or complex relationships." None of this is insight just for the sake of insight. These tasks must be guided by and directed towards what is valuable to the business as they turn the "data into new product offerings and improved monetization." A data scientist must recognize that what they do has to drive "business impact through actionable analytic insight."

The next task is applying the analytics to the data. We have already discussed the variety of analytical techniques that could be performed. Any and all of these are possible to use as long as it is appropriate for the setting and generates value to the business. Specific applications of analytics include

- "personal recommendations, machine learned search rankings, and email targeting algorithms"

- “analyzing A/B test results, developing metrics and systems to monitor the ongoing quality of the models”
- “segmenting donors into interesting subgroups and modeling individual-level behavior such as giving, event participation, and volunteering”
- “providing robust analytics to develop new digital partnerships and enhance our ecommerce capabilities”
- “define new metrics and help drive feature teams to use the right metrics”
- “analyzing user and social behavior and determining patterns in the data”
- “perform sensitivity analysis on a variety of options, and recommend to senior leadership the best ways to move forward”

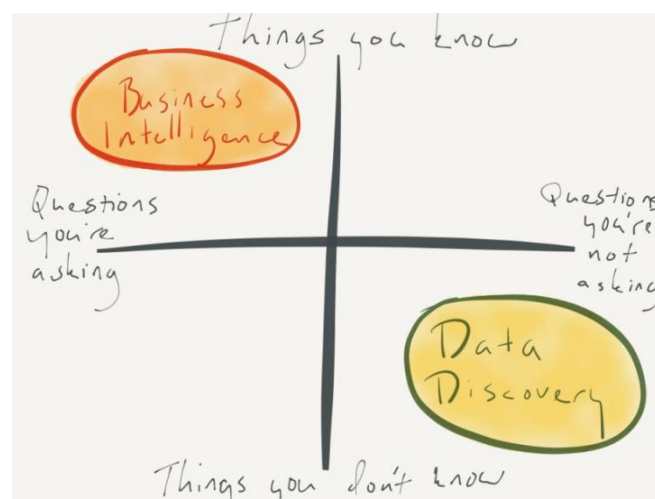
The “productionization” of analytics is an important part of the data scientist’s responsibilities, as well. They may do this on their own or work with software engineers or BI specialists to do so. They are responsible for maintaining the algorithms and updating them when they become out of date. Often the data scientist is still responsible for finding new data sources for or improving the analytics behind the algorithms in production. Where they spend most of their times depends on whether they favor more data or better algorithms. (Tawakol, 2012)

Finally, the data scientist is often responsible for presenting their results in many forums, primarily to the business units or production managers. These presentations may require training or communicating advanced computational, data and analytical concepts to a lay audience. They often also include “representing the institution in public forums, professional conferences, and publications.” Related to this is their continued education at these forums and conferences, as well as keeping up with publications and the literature. The data scientist is expected to be on the cutting edge of analytics so that their company can maintain the competitive advantage.

WHAT DOES THE DATA SCIENTIST PRODUCE?

From the previous discussion, it may be clear what the data scientist produces, so we will discuss it only briefly here. From a strategic or high-level view, the data scientist produces solutions to business problems. How do we get more customers? How do we generate greater response to our marketing? How do we provide our users with greater satisfaction? How do we make our games more enticing to players? To solve these problems the data scientist will create algorithms or models that can be put into production.

However, not all the “problems” that data scientists solve are known ahead of time. Sometimes they have to discover what they do not know in the data, which is where the creativity and innovation again come to play. As Jason Kolb shows in the diagram below, their field is the intersection of things the company does not know and questions they are not asking. (Kolb)



HOW TO BE A DATA SCIENTIST

Now that we understand what a data scientist is and what they do, we come to the main question of the paper: How can the reader become a data scientist using SAS? It is hoped that the answer to the question is fairly obvious. Many data scientists are already using SAS, though they also like to use other public domain software, as do smaller, startup companies. It is also true that many SAS users can already be considered as data scientists, though they may not have thought of themselves in that way. In this section, we will first review the possible gaps between SAS users and “official” data scientists and offer some suggestions for closing that gap. Finally, we will review the SAS products to determine which ones can be used by the data scientist. Hint: the answer is all of them.

CLOSING THE GAP

A key aspect of the data scientist position that is different from most SAS users is that the data scientist is involved in the entire business lifecycle “from ideation to research to development to productionization.”

The typical SAS user is involved in only one stage, whether it is ETL or reporting. Becoming a data scientist, then, may mean stretching yourself into areas that are not familiar or may not be comfortable. To do this, you can look for opportunities in your current position to take on tasks outside of your niche. Make sure that these tasks are high impact, though, and provide value to the business.

Another aspect is the creativity and innovation around data. Most SAS users have their data sent to them and there is a well-defined output to be created. Look for new ways to use the data and for new sources of data that can help solve the problem better.

Much of the data scientist’s data comes from the internet or Web 2.0 sources, such as social media. Start learning more about these data sources and how to make them usable. SAS can consume almost all of them, but you have to know how it is done.

Another key aspect that companies are looking for in the data scientist is the passion. We saw this in almost all of the job descriptions. SAS users who have been doing the same job for a number of years may find it hard to exhibit that passion and understandably so. Somehow you have to find or create the parts of your job that you are passionate about. Remember the phrases we saw earlier:

- “passion for creating data-driven solutions to the world's most difficult problems”
- “unleash your creative genius, be inspired by those around you”
- “unabashed data geek. You enjoy searching the Internet for datasets that you can explore and mashup to tell interesting stories.”
- “thrive on data and on the opportunity to push the boundaries of web and mobile products”

Find ways to exhibit these characteristics in your job and it will surely help your career, even if you do not become a data scientist.

If you are comfortable presenting your work to others or, more importantly, they are comfortable hearing you present them, then that is great. As we saw, communication is very important in every job. An excellent venue to practice is the local and regional SAS user group meetings. If you think you do not have anything to speak about, then re-read the parts of this paper about passion.

Finally, the data scientist has to think analytically and do so well. If you have an advanced degree in a quantitative field, then you have a good start. How much have you used your statistics or computational algorithms? How much do you know anymore? Look for places to learn more and opportunities within the company to apply them. The SAS conferences always have a statistics track. The American Statistical Association and the Institute for Operations Research and the Management Sciences provide analytical training and hold analytical conferences. O'Reilly's Strata Conference is billed as the “essential training and information source for data science and big data.” (O'Reilly) We also saw previously that Coursera has data science and other analytic classes. There are many other places to learn more analytics. Make sure that the source is a credible one, though. Statistics that are misapplied can cause damage to a company and your career.

SAS PRODUCTS TO USE

Creativity, thinking and passion are only part of what is needed. The data scientist has to have the right tools to do the job. One of the exciting things about SAS is that it has all the tools you need. Both SAS Foundation and Visual Data Discovery (EG and JMP) can be used to read the sources of data we have been discussing, and then clean and explore the data.

However, the data scientist will want to take advantage of some of the more powerful SAS tools: Visual Analytics, SAS/Access to Hadoop, and their High-Performance Analytics made up of Grid Computing, In-Database Processing and In-Memory Analytics.

For the advanced analytics, such as building models and doing optimization, the data scientist has the best in class tools available to them: SAS Enterprise Miner, SAS/OR, SAS Forecast Studio, and the suite of SAS Text Analytics (SAS Enterprise Content Categorization, SAS Ontology Management, SAS Sentiment Analysis and SAS Text Miner).

The following table will help the reader understand the connections between the data scientist responsibilities and the SAS suite. Note that we do not address the SAS solutions, which are intended for the enterprise and do not fit with the data scientist mentality. Some tools, such as DI Studio, may not seem to fit with this mentality either, but there are strong enough reasons for using them that they are included. Note also that these responsibilities (obviously) will not include any of the business knowledge or soft skills. There are many papers from the SAS conference proceedings that cover the use of these tools, as well as the usual documentation.

	Responsibility	SAS tools
Acquiring data	Integrate data from multiple sources	SAS Foundation, DI Studio, SAS/Access to <insert RDBMS here!>
	Extract data from the web	FILENAME, PROC SOAP, PROC HTTP, LIBNAME
	Work with large data	SAS/Access to Hadoop, SAS Grid, SAS/Access to Teradata, SAS/Access to Greenplum
	Processing unstructured data	Text Analytics (Enterprise Content Categorization, Ontology Management, Sentiment Analysis, Text Miner)
	Clean data	Dataflux, SAS Foundation, Regular expressions
Exploring	Exploratory data analysis	JMP, Enterprise Guide
	Create mashups	Enterprise Guide, SAS Foundation
	Create data visualizations	ODS Graphics, JMP, Visual Analytics
Analytics	Forecasting	ETS, SAS Forecast Studio
	Neural networks, data mining, clustering, segmentation	Enterprise Miner, SAS/STAT
	Optimization	SAS/OR
	Natural language processing	Text Analytics
	Big analytics	SAS High-Performance Analytics, SAS/STAT
Presentation	Data products	Stored processes

Web services	SAS BI Web Services
Visualization	SAS EBI, VA

Not all SAS users have all of these tools available to them. Many companies do take advantage of them, though, such as the New York City-based insurance firm Chartis Inc. (Kelly, 2010)

CONCLUSION

Data Science is a hot, new career because of the growth in Analytics, Big Data and High Performance Computing. The data scientist role is specifically intended to take advantage of the large amounts of data available to companies in order to give them a competitive advantage in their industry. They bring to the company a combination of skills that includes analytics (statistical techniques and machine learning), business knowledge and soft skills (passion, creativity, innovation) and programming (SAS, SQL, perl, java). Sometimes companies see the data scientist as a wildcatter, someone who drills exploration oil wells in areas not known to be oil fields, or maybe a gold miner, digging for nuggets underground. Each term evokes images of exploration, daring and risk that generate high impact by “striking it rich.” Often companies have stored data without really understanding what they have. When a data scientist digs into the data and gets creative about using it, then they just might find insights that do strike it rich for the company. Sometimes the data scientist is seen as an inventor who creates innovative data products that again provide strong value to the company.

The exciting thing for the SAS user is that they already have at hand almost all the tools they need to be a data scientist. Even if they just have Enterprise Guide and JMP, they can still go far. As we have seen, the success of the data scientist is partly the technology, especially working with Big Data, but it is also a large part one’s attitude. Re-think the way you work and you can easily be a data scientist.

REFERENCES

- Brown, B., Chui, M., & Manyika, J. (2011, October). *Are you ready for the era of ‘big data’?* Retrieved from McKinsey & Company: https://www.mckinseyquarterly.com/Strategy/Innovation/Are_you_ready_for_the_era_of_big_data_2864
- Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., et al. (2011, October 11). *Data Mining Curriculum: A Proposal (Version 1.0)*. Retrieved from Sig KDD: <http://www.kdd.org/curriculum/CURMay06.pdf>
- Cleveland, W. S. (2001). *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*. Bell Labs.
- Conway, D. (2010, September 30). *The Data Science Venn Diagram*. Retrieved from DrewConway.com: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>
- Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, 5.
- Hacker (programmer subculture)*. (n.d.). Retrieved 2013, from Wikipedia: [en.wikipedia.org/wiki/Hacker_\(free_and_open_source_software\)](http://en.wikipedia.org/wiki/Hacker_(free_and_open_source_software))
- Hahn, G., & Hoerl, R. (1998). Key Challenges for Statisticians in Business and Industry. *Technometrics*, Vol 40, No 3, 195-200.
- Hartwig, E. (2013, February 02). *What to Know Before You Apply to a Startup*. Retrieved from Mashable: <http://mashable.com/2013/02/02/startup-before-you-apply/>
- Howe, B. (n.d.). *Introduction to Data Science*. Retrieved from Coursera: <https://www.coursera.org/course/datasci>
- Institute for Advance Analytics. (n.d.). Retrieved June 2013, from Institute for Advanced Analytics: <http://analytics.ncsu.edu/>
- Kalil, T. (2012, March 29). *Big Data is a Big Deal*. Retrieved from The White House: <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>
- Kelly, J. (2010, August 17). *Analytics skills in demand -- and analytics pros demanding top salaries*. Retrieved from TechTarget: <http://searchbusinessanalytics.techtarget.com/news/2240021017/Analytics-skills-in-demand-and-analytics-pros-demanding-top-salaries>
- Kolb, J. (n.d.). *The new reality for Business Intelligence and Big Data*. Retrieved from Applied Data Labs:

- <http://www.applieddatalabs.com/content/new-reality-business-intelligence-and-big-data>
- Lombardi, A. (2013, April 8). *Are You Recruiting for These In-Demand Skills?* Retrieved from Wanted Analytics: <http://www.wantedanalytics.com/insight/2013/04/08/are-you-recruiting-for-these-hot-skills/>
- O'Reilly. (n.d.). *O'Reilly Strata Conference*. Retrieved June 2013, from <http://strataconf.com/>
- Patil, D. (2011, September 16). *Building data science teams*. Retrieved from O'Reilly Radar: <http://radar.oreilly.com/2011/09/building-data-science-teams.html>
- Pitts, M. (2013). *Building the Big-Data Analytics Architecture*. Retrieved 2013, from All Analytics Academy: http://www.allanalytics.com/lecture.asp?doc_id=258001
- Pryor, L. H. (2012, July 18). *Summer School Big Data Talk 071812*. Retrieved from NASA Ames Research Center: https://c3.nasa.gov/nex/static/media/other/Summer_School_Big_Data_Talk_071812.pdf
- Rowe, A. (2013, July 6). *The Big Search for Big Data Candidates*. Retrieved from Wanted Analytics: <http://www.wantedanalytics.com/insight/2013/07/06/the-big-search-for-big-data-candidates/>
- Steele Jr., G. L., Woods, D. R., Finkel, R. A., Crispin, M. R., Stallman, R. M., & Goodfellow, G. S. (n.d.). *The Hacker's Dictionary*. Retrieved from jargon-file.org: <http://jargon-file.org/archive/jargon-1.5.0.dos.txt>
- Stubbs, E. (2013). *Delivering Business Analytics: Practical Guidelines for Best Practice*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Tawakol, O. (2012, September 7). *More Data Beats Better Algorithms — Or Does It?* Retrieved from All Things D: <http://allthingsd.com/20120907/more-data-beats-better-algorithms-or-does-it/>

ACKNOWLEDGMENTS

The author would like to thank his colleagues at Experis who helped collect the data for this paper: Jared Slain, Laura Oliver and Cindy Gardner.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chuck Kincaid
Experis Business Intelligence and Analytics
5220 Lovers Lane
Portage, MI 49002
(269) 553-5140
chuck.kincaid@experis.com
www.experis.us/analytics

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.