What Is Data Science, and What Does a Data Scientist Do?

Introduction

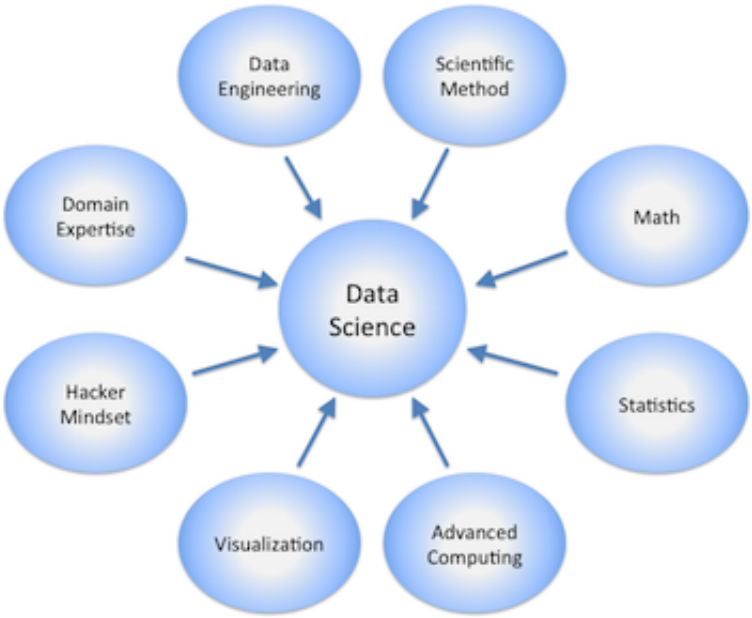
What profession did Harvard call the Sexiest Job of the 21st Century? That's right... the data scientist.

Ah yes, the ever mysterious data scientist. So what exactly is the data scientist's secret sauce, and what does this "sexy" person actually do at work every day?

This article is intended to help define the data scientist role, including typical skills, qualifications, education, experience, and responsibilities. This definition is somewhat loose since there really isn't a standardized definition of the data scientist role, and given that the ideal experience and skill set is relatively rare to find in one individual.

This definition can be further confused by the fact that there are other roles sometimes thought of as the same, but are often quite different. Some of these include data analyst, data engineer, and so on. More on that later.

Here is a diagram showing some of the common disciplines that a data scientist may draw upon. A data scientist's level of experience and knowledge in each, often varies along a scale ranging from beginner, to proficient, and to expert, in the ideal case.



By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

While these, and other disciplines and areas of expertise (not shown here), are all characteristics of the data scientist role, I like to think of a data scientist's foundation as being based on four pillars. Other more specific areas of expertise can be derived from these pillars.

Let's discuss them now.

The Pillars of Data Science Expertise

While data scientists often come from many different educational and work experience backgrounds, most should be strong in, or in an ideal case be experts in four fundamental areas. In no particular order of priority or importance, these are:

- Business domain
- Statistics and probability
- Computer science and software programming

- Written and verbal communication

There are other skills and expertise that are highly desirable as well, but these are the primary four in my opinion. These will be referred to as the data scientist pillars for the rest of this article.

In reality, people are often strong in one or two of these pillars, but usually not equally strong in all four. If you do happen to meet a data scientist that is truly an expert in all, then you've essentially found yourself a unicorn.

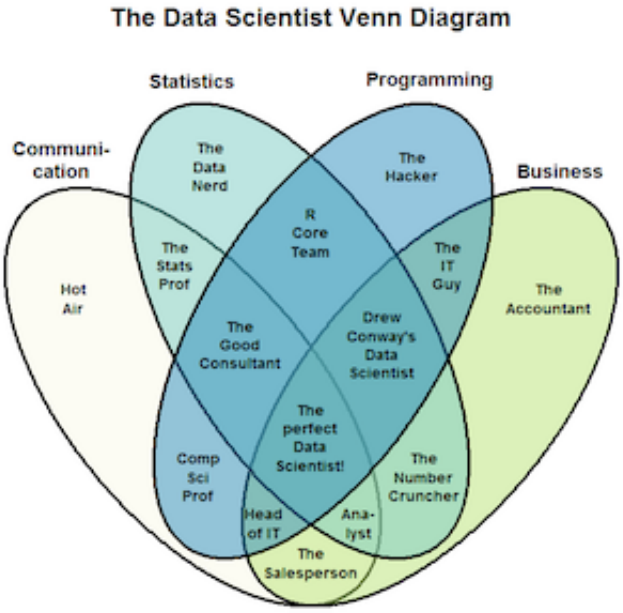
Based on these pillars, a data scientist is a person who should be able to leverage existing data sources, and create new ones as needed in order to extract meaningful information and actionable insights. These insights can be used to drive business decisions and changes intended to achieve business goals.

This is done through business domain expertise, effective communication and results interpretation, and utilization of any and all relevant statistical techniques, programming languages, software packages and libraries, data infrastructure, and so on.

Data Science Venn Diagrams

One can find many different versions of the data scientist Venn diagram to help visualize these pillars (or variations) and their relationships with one another. David Taylor wrote an excellent article on these Venn diagrams entitled, Battle of the Data Science Venn Diagrams. I highly recommend reading it.

Here is one of my favorite data scientist Venn diagrams created by Stephan Kolassa. You'll notice that the primary ellipses in the diagram are very similar to the pillars given above.



This diagram, and others like it, attempt to assign labels and/or characterize the person or field that lies at the intersection of each of the primary competencies shown, which I'm calling pillars here.

As this diagram shows, Stephan Kolassa labels 'The Perfect Data Scientist' as the individual who is equally strong in business, programming, statistics, and communication. I agree completely.

Data Science Goals and Deliverables

In order to understand the importance of these pillars, one must first understand the typical goals and deliverables associated with data science initiatives, and also the data science process itself. Let's first discuss some common data science goals and deliverables.

Here is a short list of common data science deliverables:

- Prediction (predict a value based on inputs)
- Classification (e.g., spam or not spam)
- Recommendations (e.g., Amazon and Netflix recommendations)

- Pattern detection and grouping (e.g., classification without known classes)
- Anomaly detection (e.g., fraud detection)
- Recognition (image, text, audio, video, facial, …)
- Actionable insights (via dashboards, reports, visualizations, …)
- Automated processes and decision-making (e.g., credit card approval)
- Scoring and ranking (e.g., FICO score)
- Segmentation (e.g., demographic-based marketing)
- Optimization (e.g., risk management)
- Forecasts (e.g., sales and revenue)

Each of these is intended to address a specific goal and/or solve a specific problem. The real question is which goal, and whose goal is it?

For example, a data scientist may think that her goal is to create a high performing prediction engine. The business that plans to utilize the prediction engine, on the other hand, may have the goal of increasing revenue, which can be achieved by using this prediction engine.

While this may appear to not be an issue at first glance, in reality the situation described is why the first pillar (business domain expertise) is so critical. Often members of upper management have business-centric educational backgrounds, such as an MBA.

While many executives are exceptionally smart individuals, they may not be well versed on all the tools, techniques, and algorithms available to a data scientist (e.g., statistical analysis, machine learning, artificial

intelligence, and so on). Given this, they may not be able to tell a data scientist what they would like as a final deliverable, or suggest the data sources, features (variables), and path to get there.

Even if an executive is able to determine that a specific recommendation engine would help increase revenue, they may not realize that there are probably many other ways that the company's data can be used to increase revenue as well.

It can therefore not be emphasized enough that the ideal data scientist has a fairly comprehensive understanding about how businesses work in general, and how a company's data can be used to achieve top-level business goals.

With significant business domain expertise, a data scientist should be able to regularly discover and propose new data initiatives to help the business achieve its goals and maximize their KPIs.

The Data Science Process

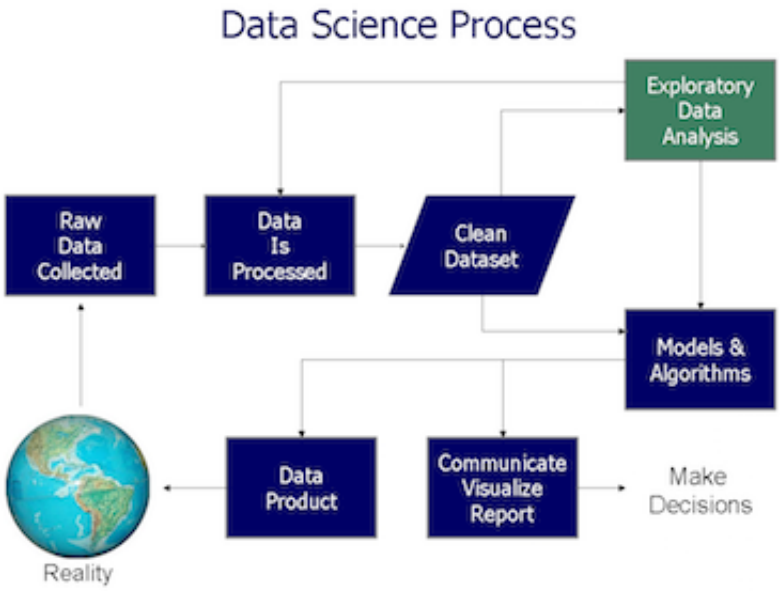
The data science process can be a bit variable depending on the project goals and approach taken, but generally mimics the following.

The data science process involves these phases, more or less:

- Data acquisition, collection, and storage
- Discovery and goal identification (ask the right questions)
- Access, ingest, and integrate data
- Processing and cleaning data (munging/wrangling)
- Initial data investigation and exploratory data analysis (EDA)
- Choosing one or more potential models and algorithms

- Apply data science methods and techniques (e.g., machine learning, statistical modeling, artificial intelligence, ...)
- Measuring and improving results (validation and tuning)
- Delivering, communicating, and/or presenting final results
- Business decisions and/or changes are made based on the results
- Repeat the process to solve a new problem

Here is a diagram representing a simpler version of this process.



Farcaster at English Wikipedia [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

That's the process in a nutshell. So how do these pillars come into play here?

Data Scientist Pillars, Skills, and Education In-Depth

We've already discussed the business domain and communication pillars, which represent business acumen and top notch communication skills. This is very important for the discovery and goal phase. It's also very helpful in that data scientists typically have to present and communicate results to key stakeholders, including executives.

So strong soft skills, particularly communication (written and verbal) and public speaking ability are key. In the phase where results are communicated and delivered, the magic is in the data scientist's ability to deliver the results in an understandable, compelling, and insightful way, while using appropriate language and jargon level for her audience. In addition, results should always be related back to the business goals that spawned the project in the first place.

For all of the other phases listed, data scientists must draw upon strong computer programming skills, as well as knowledge about statistics, probabilities, and mathematics in order to understand the data, choose the correct solution approach, implement the solution, and improve on it as well.

One important thing to discuss are off-the-shelf data science platforms and APIs. One may be tempted to think that these can be used relatively easily and thus not require significant expertise in certain fields, and therefore not require a strong, well-rounded data scientist.

It's true that many of these off-the-shelf products can be used relatively easily, and one can probably obtain pretty decent results depending on the problem being solved, but there are many aspects of data science where experience and chops are critically important.

Some of these include having the ability to:

- Customize the approach and solution to the specific problem at hand in order to maximize results, including the ability to write new

algorithms and/or significantly modify the existing ones, as needed

- Access and query many different databases and data sources (RDBMS, NoSQL, NewSQL), as well as integrate the data into an analytics-driven data source (e.g., OLAP, warehouse, data lake, ...)
- Find and choose the optimal data sources and data features (variables), including creating new ones as needed (feature engineering)
- Understand all statistical, programming, and library/package options available, and select the best
- Ensure data has high integrity (good data), quality (the right data), and is in optimal form and condition to guarantee accurate, reliable, and statistically significant results
- Avoid the issues associated with garbage in equals garbage out
- Select and implement the best tooling, algorithms, frameworks, languages, and technologies to maximize results and scale as needed
- Choose the correct performance metrics and apply the appropriate techniques in order to maximize performance
- Discover ways to leverage the data to achieve business goals without guidance and/or deliverables being dictated from the top down, i.e., the data scientist as the idea person
- Work cross-functionally, effectively, and in collaboration with all company departments and groups
- Distinguish good from bad results, and thus mitigate the potential risks and financial losses that can come from erroneous conclusions

and subsequent decisions

- Understand product (or service) customers and/or users, and create ideas and solutions with them in mind

Education-wise, there is no single path to becoming a data scientist. Many universities have created data science and analytics-specific programs, mostly at the master's degree level. Some universities and other organizations also offer certification programs as well.

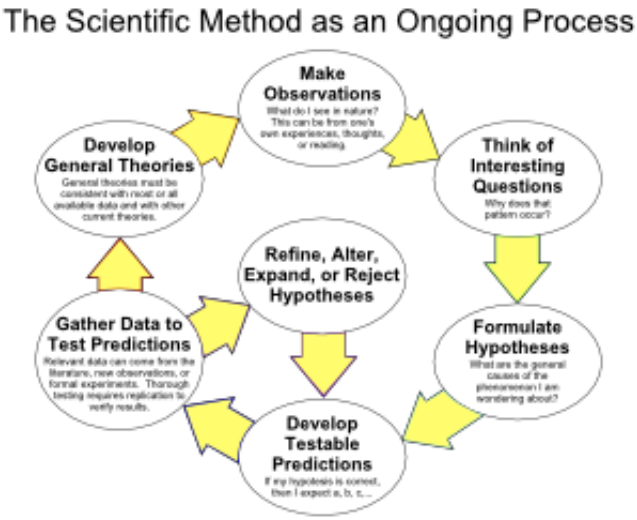
In addition to traditional degree and certification programs, there are bootcamps being offered that range from a few days or months to complete, online self-guided learning and MOOC courses focused on data science and related fields, and self-driven hands-on learning.

No matter what path is taken to learn, data scientist's should have advanced quantitative knowledge and highly technical skills, primarily in statistics, mathematics, and computer science.

The "Science" in Data Science

The term science is usually synonymous with the scientific method, and some of you may have noticed that the process outlined above is very similar to the process characterized by the expression, scientific method.

Here is an image that visualizes the scientific method as an ongoing process.



By ArchonMagnus (Own work) [CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Generally speaking, both traditional scientists and data scientists ask questions and/or define a problem, collect and leverage data to come up with answers or solutions, test the solution to see if the problem is solved, and iterate as needed to improve on, or finalize the solution.

Data Scientists vs. Data Analysts vs. Data Engineers

As mentioned, often the data scientist role is confused with other similar roles. The two main ones are data analysts and data engineers, both quite different from each other, and from data science as well.

Let's explore both of these roles in more detail.

Data Analyst

Data analysts share many of the same skills and responsibilities as a data scientist, and sometimes have a similar educational background as well. Some of these shared skills include the ability to:

- Access and query (e.g., SQL) different data sources

- Process and clean data
- Summarize data
- Understand and use some statistics and mathematical techniques
- Prepare data visualizations and reports

Some of the key differences however, are that data analysts typically are not computer programmers, nor responsible for statistical modeling, machine learning, and many of the other steps outlined in the data science process above.

The tools used are usually different as well. Data analysts often use tools for analysis and business intelligence like Microsoft Excel (visualization, pivot tables, ...), Tableau, SAS, SAP, and Qlik.

Analysts sometimes perform data mining and modeling tasks, but tend to use visual platforms such as IBM SPSS Modeler, Rapid Miner, SAS, and KNIME. Data scientists, on the other hand, perform these same tasks usually with tools such as R and Python, combined with relevant libraries for the language(s) being used.

Lastly, data analysts tend to differ significantly in their interactions with top business managers and executives. Data analysts are often given questions and goals from the top down, perform the analysis, and then report their findings.

Data scientists however, tend to generate the questions themselves, driven by knowing which business goals are most important and how the data can be used to achieve certain goals. In addition, data scientists typically leverage programming with specialized software packages and employ much more advanced statistics, analytics, and modeling techniques.

Data Engineer

Data engineers are becoming more important in the age of big data, and can be thought of as a type of data architect. They are less concerned with statistics, analytics, and modeling as their data scientist/analyst counterparts, and are much more concerned with data architecture, computing and data storage infrastructure, data flow, and so on.

The data used by data scientists and big data applications often come from multiple sources, and must be extracted, moved, transformed, integrated, and stored (e.g., ETL/ELT) in a way that's optimized for analytics, business intelligence, and modeling.

Data engineers are therefore responsible for data architecture, and for setting up the required infrastructure. As such, they need to be competent programmers with skills very similar to someone in a DevOps role, and with strong data query writing skills as well.

Another key aspect of this role is database design (RDBMS, NoSQL, and NewSQL), data warehousing, and setting up a data lake. This means that they must be very familiar with many of the available database technologies and management systems, including those associated with big data (e.g., Hadoop and HBase).

Lastly, data engineers also typically address non-functional infrastructure requirements such as scalability, reliability, durability, availability, backups, and so on.

The Data Scientist's Toolbox

We'll finish with an overview of some of the typical tools in the data scientist's proverbial toolbox.

Since computer programming is a large component, data scientists must be proficient with programming languages such as R, Python, SQL,

Scala, Julia, Java, and so on. Usually it's not necessary to be an expert programmer in all of these, but R, Python, and SQL are definitely key, and others like Scala for big data are becoming more prominent as well.

For statistics, mathematics, algorithms, modeling, and data visualization, data scientists usually use pre-existing packages and libraries where possible. Some of the more popular ones include Scikit-learn, e1071, Pandas, Numpy, TensorFlow, Matplotlib, D3, Shiny, and ggplot2.

For reproducible research and reporting, data scientists commonly use notebooks and frameworks such as Jupyter, iPython, Knitr, and R markdown. These are very powerful in that the code and data can be delivered along with key results so that anyone can perform the same analysis, and build on it if desired.

More and more these days, data scientists should be able to utilize tools and technologies associated with big data as well. The most popular examples include Hadoop, Spark, Hive, Pig, Drill, Presto, Mahout, and so on.

Finally, data scientists should know how to access and query many of the top RDBMS, NoSQL, and NewSQL database management systems. Some of the most common are MySQL, PostgreSQL, Redshift, MongoDB, Redis, Hadoop, and HBase.

Summary

Harvard was right about data scientists. It's an extremely important and high-demand role that can have significant impact on a business' ability to achieve its goals, whether they are financial, operational, strategic, and so on.

Company's collect a ton of data, and much of the time it's neglected or underutilized. This data, through meaningful information extraction and

discovery of actionable insights, can be used to make critical business decisions and drive significant business change. It can also be used to optimize customer success and subsequent acquisition, retention, and growth.

As mentioned, data scientists can have a major positive impact on a business' success, and sometimes inadvertently cause financial loss, which is one of the many reasons why hiring a top notch data scientist is critical.

Hopefully this article has helped demystify the data scientist role and other related roles.

Cheers!