# **What Makes a Data Scientist?**

*Author: Alex Labram*

You’ve heard it in the news. You’ve seen it on the internet. You’ve read it in The Actuary magazine. Data science is the hot topic of the last decade, and data scientist is the new rock-star job.

But what is it, who does it, and does it really merit all the hype?

## **How did we get here?**

To quote Wikipedia, data science is “a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data”.

The term “data science” technically originated in the 1960s as a synonym for computer science, but was first used with its current meaning in the mid-90s. Its increased popularity in the early 2000s was driven by the reduced cost of computing power and data storage, by the wider availability of statistical programming and data visualisation tools, and as a reaction to the strong distributional assumptions often made by classical statistics. (Not everything is Gaussian!)

By 2012, data science had attained buzzword status, and since then the number of data scientists has grown exponentially. Despite this, demand still consistently exceeds supply, driving high salaries for skilled practitioners.

## **What goes into data science?**

Data science is a multi-disciplinary field. In the same way that architects are required to understand both the dynamics of building residents and the material properties of concrete, a data scientist is expected to understand[1]:

* **Data visualisation** (aka “viz”). Where this discipline was traditionally split cleanly between highly-paid *artistes* on the one hand and crummy Excel chart producers on the other, over time it has become far more democratic. Anyone can now produce an infographic or a dashboard; a data scientist will know how and when to do so.
* **Mathematics** and **statistics**. Many data scientists have an academic background in this area. Those who don’t have these credentials must develop an excellent grasp of key statistics and distributions, alongside concepts such as Bayesian analysis and hypothesis testing, and general mathematical skills such as differential calculus and linear algebra.
* **Machine learning**. This discipline covers a range of techniques for non-parametric or semi-parametric modelling, not limited to the methods arising from classical statistics. It is often split into statistical learning (appropriate for medium-sized datasets) and deep learning (appropriate if you’re Google). This will be discussed in detail in later articles.
* **Computer science**. The right design choice can make the difference between a five-minute model run and a three-day impossibility. A data scientist will understand the philosophy behind the programming: data structures and algorithms; relational database use and management; parallelisation and distribution of large models; and performance analysis of code and processes.
* **Communications**. A message is not delivered until it is understood. Data scientists should be able to explain their methods and findings clearly and intuitively to a range of stakeholders including non-technical managers, regulators and end users.
* **Domain expertise**. Despite the best efforts of machine learning enthusiasts, it is impossible to effectively analyse data without having at least some clue what it means! Data science skills are comparatively transferrable between sectors, but practitioners still often focus on a given industry of which they have concrete knowledge.

The work of data scientists may be complemented by that of related practitioners:

* **Data engineers**. Where data scientists typically start out as statisticians who have learned to program, data engineers are usually programmers who have learned the data requirements of statisticians. They have exceptional computer science skills, with a focus on the construction and maintenance of data pipelines including in a Big Data context.
* **Machine learning engineers**. Once a data scientist has identified a high-performing model, it must be “productionised”: made usable to stakeholders as needed. The effort required to roll the model out, provide convenient access (e.g. setting up an API), ensure high throughput and monitor the model’s accuracy over time is distinct enough from bread-and-butter data science work that it is often treated as a separate field.
* **Regulatory & Ethics specialists**. In our post-GDPR world, and with an increasing number of scandals surrounding the use and abuse of data, extra effort may be required to ensure that no regulations are breached and no vulnerable groups are disadvantaged.

## **Actuaries vs data scientists**

There’s quite a lot of overlap in the above with the actuarial skillset. In particular, a qualified actuary should have excellent domain expertise, a broad knowledge of statistics, and (hopefully) the ability to fluently communicate their findings.

However, there are also significant differences. Just from the above list: actuaries traditionally have minimal exposure to data viz best practices, scant knowledge of computer science principles, and limited experience of machine learning algorithms and implementations.

Stepping back a bit, actuaries differ from data scientists in four major ways:

* Different **datasets**. Data scientists frequently work with larger, messier and more heterogeneous datasets than would be considered viable for an actuarial modelling process.
* Different **toolsets**. Where actuaries typically rely on a combination of Excel and specialist actuarial software (Prophet, RiskAgility, Mo.net), data scientists are expected to have strong knowledge of statistical programming languages (R, Python), data transformation languages (SQL), and data viz tools (Tableau, Power BI, D3).
* Different **skillsets**. Where actuaries often have deep knowledge of a few job-relevant areas of statistics – GLMs, for example – data scientists will generally have broad knowledge of statistical and machine learning techniques across a range of use cases. Data scientists will also have a higher level of general programming ability than is commonly seen in the actuarial world.
* Different **mindsets**. Actuaries have inherited the model-led approach of classical statistics: first assume a distribution; then fit it to the data; then review goodness-of-fit. Data scientists operate in a more data-led fashion, with an emphasis on exploratory data analysis and rapid iteration over different possible models, to avoid baking their presuppositions into their output.

## **But is it ~~art~~ science?**

This last point – the iteration over models – is really what justifies the use of “science” in the field’s name. Data scientists follow similar best practices in testing a model to those used by medical researchers, for example, in testing a new drug or treatment.

They have a keen awareness of risks to the statistical validity and business relevance of their work. These include, but are not limited to: data limitations, over-fitting, p-value fishing, failure to converge, inappropriate performance measures, model bias, observer effects, out-of-sample prediction, and auditability.

By abstracting the analysis of data away from the idiosyncratic concerns of particular business sectors and use cases, by incorporating cutting-edge tools and algorithms, and by framing their work so as to avoid statistical bear-traps, data scientists can develop and communicate game-changing insights.

For further information, please see the following resources:

* [R for Data Science](https://r4ds.had.co.nz/) (Wickham & Grolemund, 2017, O’Reilly Media)
* [Doing Data Science](https://www.oreilly.com/library/view/doing-data-science/9781449363871/) (O’Neil & Schutt, 2013, O’Reilly Media)
* [Nine Must-Have Skills to Become a Data Scientist](https://www.kdnuggets.com/2018/05/simplilearn-9-must-have-skills-data-scientist.html) (KDNuggets)
* [Data Scientist: The Sexiest Job of the 21st Century](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century) (Harvard Business Review)

[1] Source: Doing Data Science, chapter 1.