Lies, Damn Lies and Data Science

Recently, I find myself being asked by friends and acquaintances from outside the industry "What is Data Science?" and "What does a Data Scientist do?" I struggle to find a concise yet understandable answer for them which doesn't lead to more questions. How is Data Science different from Data Analytics? Surely all scientists use data, they don't actually research data itself? Isn't that just Statistics? At first glance, it seems like the term "Data Scientist" is a bit of a tautology; But if we accept science as a systematically organised body of knowledge on a particular subject, it makes more sense.

It has been decades since Jeff Wu first uttered "Data Science" in 1985. Originally an alternative term for statistics, it has evolved into a nebulous chimera of Statistics, Business Analysis, Mathematics, Operational Research, Programming and Visualization. Data Science was initially very much business driven, applying the scientific method and technologies to the data generated in a business environment. The purpose of the role was to add business value through improved decision-making resulting from the analysis of copious volumes of new data generated by modern business systems. Basically, to perform alchemy and turn the lead of raw business data into golden insights and competitive advantage.

Statistics is a well-established and essential branch of mathematics providing the theoretical rigour and tools to investigate real-world questions, address uncertainty and understand data. It has been the cornerstone of the scientific method for hundreds of years, providing the foundations and means to communicate empirical scientific research. Early Statisticians had a major impact on our understanding of the world, such as Florence Nightingale's insight that 16,000 out of 18,000 Crimean War mortalities were caused by preventable diseases due to bad hygiene, not trauma from battle wounds. Or Dr John Snow's data-driven discovery that cholera was transmitted via poor water quality which formed the early basis for our understanding of pathogens.

At its heart, Statistics transforms raw data into knowledge and understanding - fundamental to real-world insights and improved decision-making. Evidently Data Science and Statistics have much in common, they share the same objectives and methods; indeed, Data Science draws upon many statistical techniques, however there are some fundamental differences.

Statistics has traditionally focused on learning and understanding data through descriptive and inferential techniques; concerned with samples, populations, hypotheses, descriptions, inferences, etc... all of which can be implemented with a pencil, paper, some quality thinking, and a great deal of time. Historically, statistical techniques could not leverage any computational power and hence it was only possible to collect and manage smaller data sets such as experimental data or surveys.

Smaller data sets require very precise analysis and understanding of uncertainty as there is less data to elicit the signal from the noise. Quantifying uncertainty and accurately determining the relationship between independent and dependent variables is a critical element of Statistics.

Data Science on the other hand, was born of the need to handle the abundant supply of data generated by these new business systems. Exponential data growth and real-time, or close to real-time, data sources created a capability gap in terms of skills and technologies. In addition to the enormous volumes of rapid and accessible data, much of it is unstructured, with no underlying data model or schema and therefore lacking the nicely formatted rows and columns which conventional tools and techniques can manage. Technology, frameworks and infrastructure to handle large volumes of unstructured data pioneered by Google, Yahoo and Facebook were made available to all. It is now possible to consume and computationally analyse entire populations of unstructured data rather than sampling smaller data sets and inferring. But such technologies require specialist skills and knowledge of data and processing.

Furthermore, a multitude of sophisticated statistical algorithms became readily accessible through open-source libraries. It is possible to run complex machine learning from your client, but again, such technologies require specialist skills and knowledge of Statistics to fully harness and understand. Anyone familiar with basic programming can avail themselves of such technological advancements, but operating and comprehending these incredibly powerful algorithms is highly specialised. But not only are many advanced analysis methods available as open-source, with the advent of cloud everyone had vast computational resources at their fingertips to access, acquire, store, process and analyse data at scale. From this confluence of data, technology and the capability gap, Data Science emerged.

Technology and the exponential growth of data have accelerated Data Science far beyond Computational Statistics: the application of greater processing power, automation and iteration. It has opened up new horizons unconstrained by human perception, into the realms of pattern recognition and prediction through inductive reasoning. Through repeated processing of data and refinement through feedback, a machine can “learn” of its own accord and produce an accurate model. In this way, machine learning makes predictions through bottom-up processing, where incoming data is transformed step by step, to find patterns in data and develop a repeatable, high-level model. Given clear objectives, success criteria, and a clean corpus of data, a machine learning model can be extremely accurate in forecasting.

These machine learning algorithms can identify patterns in enormous data sets far beyond human interpretation, but they are highly sensitive and easily fooled by subtleties which humans would easily recognise. The human brain has an innate ability to find patterns, honed through countless generations of evolution and outperforms machines at inferring patterns from just a few simple examples. Human brains are poor at logic, calculations and recalling information, but they excel at pattern recognition, exceeding machine learning performance for many tasks. Humans find patterns and structure everywhere: Jesus appears in burnt toast, canals on Mars, dragons in clouds, or faces in shadows.

In summary, Statisticians analyse small data sets using simple mathematical models and rigorously ensure that the data is consistent with the assumptions of the model. They then focus on optimising the model to best fit the data, usually with the aim of making inferences about population parameters and the magnitude of their uncertainty. Data Scientists analyse and process data at very large scale to develop and refine a predictive model. Multiple machine learning models are tuned

and run against with each other with the most accurate model selected. Performance and results are the critical outcomes, and there is usually little or no consideration of uncertainty, which may not even be possible. Simplistically put, Data Scientists harness machine learning to build predictive models, Statisticians understand and infer from data.

Data Science will need to combine the creativity of people with technology and incorporate skills beyond STEM, such as soft skills and the humanities – the branches of knowledge concerned with human beings and human values and ethics. Communication and leadership skills will be critical to bridge business and technology and drive change. Next generation computing paradigms such as fog, serverless, cognitive and quantum computing environments promise immense new capability, but will invariably present their own Pandora's Black Box such as trust, adoption, infrastructure and security. Traditional challenges will inevitably continue to be translating business problems into computational structures.

What is clear is that no Data Scientist unicorn will be able to fulfil all of these burgeoning requirements. Success will require building collaborating teams with subject matter expertise and complementary specialisations such as Business Analysts, Data Engineers, Machine Learning Engineers, Deep Learning experts, Data Architects, Cloud Architects, UI experts and Statisticians.

At Fujitsu Data & AI, in addition to our core values of collaboration, honesty and integrity we prize problem solving, critical thinking and experience, not labels and roles. Fujitsu Data & AI's diverse team of specialists hail from diverse backgrounds. If you would like to understand what Data Science might do for your business, please contact a Fujitsu Data & AI specialist now.

It seems appropriate that the last word on Data Science should be provided by a machine: Generative Pre-trained Transformer 3 (GPT-3), an autoregressive language model developed by OpenAI, which can emulate human responses to prompts.

> **What is the difference between data science and statistics?**
> **Data science is a field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured. Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data.**

I may have found a concise yet understandable solution for my friends and acquaintances which doesn't lead to more questions.

Contact
Fujitsu Data & AI
+61 3 9924 3000