MODULE 1 UNIT 1

Defining data science

UNIVERSITY OF CAPE TOWN

IN COLLABORATION WITH getsmarter™

Table of contents

Tel: +27 21 447 7565 | **Fax:** +27 21 447 8344
Website: www.getsmarter.com | **Email:** info@getsmarter.com

Learning outcomes:

LO1: Recognise the fundamentals of data science.

LO2: Discuss the value of data science in a particular domain.

1. Introduction

In the current digital age, data is an important aspect of many businesses and other organisations. However, harnessing that data, and especially the vast amount of data that has recently become available, is not a simple task. It requires the specialised skills and tools contained within the field of data science.

Data science is a multidisciplinary field that requires its practitioners to master a number of scientific and domain-specific skill sets. The scientific core of any good data scientist consists of mathematics, statistics, and computer science, and their domain-specific expertise can span fields ranging from astronomy and physics to finance and biostatistics.

This set of notes will explore the core of data science, examine some examples of where data science is applied in industry, and outline popular statistical tools and software used in practice.

2. Data and processing power

Without data, the field of data science could not exist. Fortunately, there is no shortage of data in the current age. Consumers leave a trail of data behind them as they navigate through the modern world. Tech giants then collect and use this data to tweak adverts and product recommendations.

The collection of large swaths of data has been facilitated by the exponential decrease in the cost of data storage, as shown in Figure 1. In the late 1970s, when personal computers were becoming popular, the cost of storing a gigabyte of data was about $300,000 (McCallum, 2002:147). Think about how the price of data storage has changed in your everyday life. Nowadays, large capacity flash drives are given away as trinkets in conference goody bags.

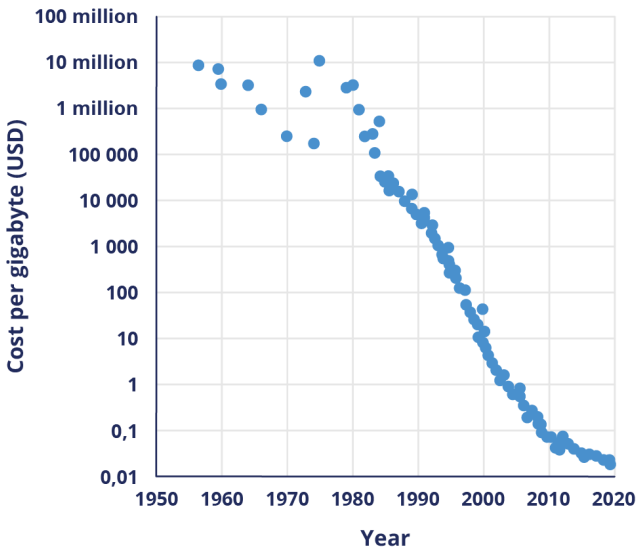

Figure 1: The decreasing cost of data storage. (Adapted from: McCallum, 2002:147)

While the cost of data storage has decreased, the speed at which computers can perform calculations has increased substantially, as shown in Figure 2. In particular, the use of special processing architectures such as graphics processing units (GPUs) and tensor processing units (TPUs), interlinked in huge computer clusters, have made it possible to process large amounts of data and conduct large-scale data analysis with astonishing efficiency.

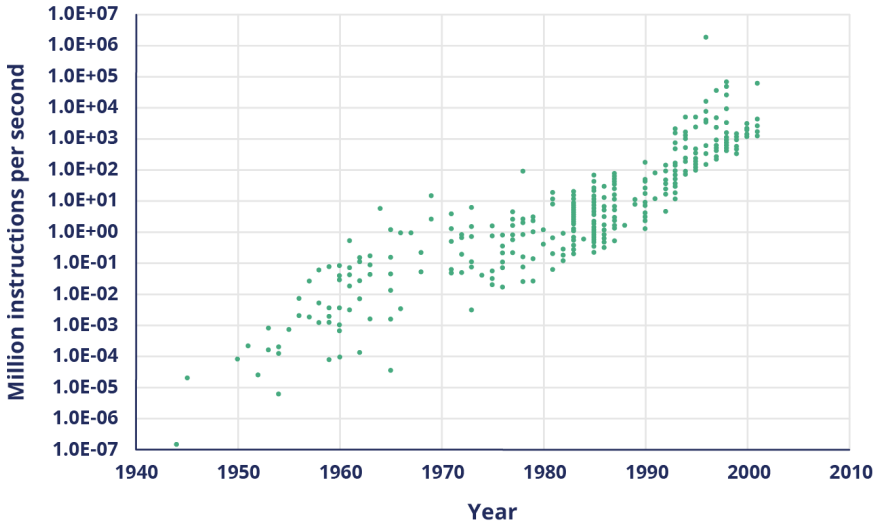

Figure 2: The increasing computational power from 1944 to 2002. (Adapted from: McCallum, 2002:144)

Having more data and more efficient computers has led to the birth of data science. While the statistical methods used in data science are not novel, the influx of data and increased computational power has enabled innovative breakthroughs in several domains. In the 1960s, NASA needed a whole department to do calculations that modern computers can now do in a few seconds (NASA, 2019b). Nowadays, data and powerful computers allow NASA to pursue projects such as detecting water on planets 1,300 light years away (NASA, 2019a).

3. What is data science?

The term "data scientist" as a job title was coined in 2008 by DJ Patil and Jeff Hammerbacher (Patil, 2011). The data scientist job title is new, and how it fits into the old science of statistics has not been well defined. However, it is clear that, as a data scientist, you will need to know how to operate in the realm of classical statistics (O'Neil & Schutt, 2014).

Further reading:

Cathy O'Neil and Rachel Schutt explore what data science is in the first chapter of their book, *Doing Data Science*.

3.1 The purpose of data science

The purpose of the majority of data science business applications is ultimately to increase revenue or reduce expenses. Data science can increase revenue by identifying new or underutilised markets. Points of growth or inefficiencies in expenses can also be identified. For example, data science methods can give insight into which products will be the most profitable, or identify areas where expenses can be reduced in operations, advertising, and product offerings. Data science can further help to optimise logistic routes or manufacturing plant operation modes, determine the most effective advertising campaigns, and identify products and services that have become redundant.

In order to gain these insights and thus achieve the purpose of data science, a data scientist needs a diverse skill set.

3.2 The key skills of a data scientist

Data science is best described as the intersection of statistics and applied mathematics, computer science, and domain knowledge. One of the first people to clearly articulate the key skills required for data science was Drew Conway (2013), who used a Venn diagram to illustrate his definition of data science. An adaptation of Conway's diagram is shown in Figure 3.

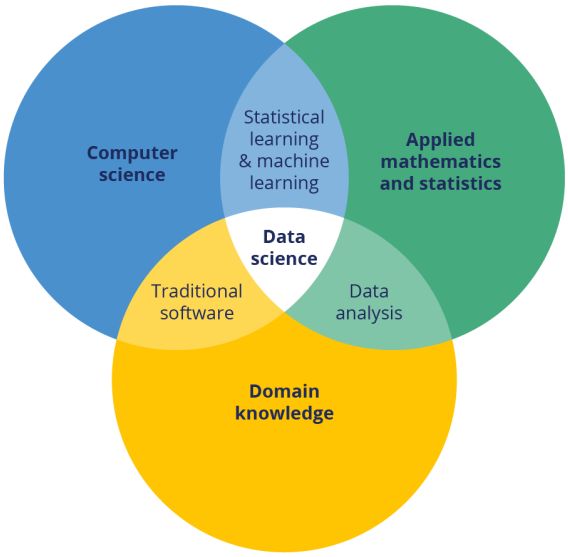

Figure 3: The data science Venn diagram. (Adapted from: Conway, 2013)

The three key skills are applied mathematics and statistics, computer science, and domain knowledge. The combination of all these skills makes a data scientist.

3.2.1 Applied mathematics and statistics

From the fields of statistics and applied mathematics, data science draws from probability theory, applied probability (statistical modelling), linear algebra, calculus, and numerical analysis. It is not necessary to be an expert in any of these fields or to be able to recite concepts from memory, but you should understand how to use the necessary tools. If a model is not working, you should be able to diagnose where it might be going wrong.

3.2.2 Computer science

The need for programming skills stems from the fact that most data is electronically stored, and that the large volumes of data would be difficult to process manually. While you don't need a degree in computer science to perform coding tasks in data science, you should be able to think algorithmically, code in an object-orientated or functional programming language, and understand how to use programming language manuals (commonly known as documentation) (Conway, 2013).

The level of coding experience needed in data science was succinctly described by Josh Wills, the director of data engineering at Slack:

> Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
>
> (Anderson, 2015:62)

3.2.3 Domain knowledge

In order to understand the goal of a data science project, substantial domain knowledge is required. If you are given a set of data about life expectancy in South Africa, you may come up with several questions that could be answered by the data, but it may turn out that none of them are useful. A clear problem should be outlined. For example, an insurance company might want to update its life expectancy models to structure its products that provide a fixed income during retirement. Your model predicts that the average life expectancy in South Africa is 65 years old.

As the data scientist, you need to understand the context surrounding the data. Is the distribution of the insurance company's customers the same as the whole country? If there are more female customers, the relevant life expectancy is likely to be higher than the country-wide average. Conversely, if the majority of customers work in high-risk jobs, their life expectancy may be lower.

4. Challenges in data science

The main challenge in data science is that it is interdisciplinary; practitioners need to be adept at several fields. Given the high level of skill required to perform data science, the field has a tendency to be academically orientated, with a gap in knowledge of how to transfer these skills to a business context (Mukherjee, n.d.). Since it is difficult to find people with this diverse skill set, there is some debate about whether it is better to have specialists in the different fields who work together or to have generalists who may not be experts in any of the three fields (Colson, 2019).

In theory, there is no particular model class that should always outperform all other model classes on a specific learning problem. You can infer from the nature of the problem (non-linearity, computational efficiency, etc.) what model classes are likely to perform well on a statistical learning task, but even then, what constitutes the "best" model can only be established in a relative sense. That is, you can compare a set of models and assess their relative performance, but that does not guarantee finding the "best" model in any absolute sense. In fact, most of the time, models are not perfectly realistic representations of the underlying process, only educated approximations. For these reasons, it is important to remain agnostic with respect to the model class applied to a given problem and let the analysis guide you to a solution.

Another challenge in data science is the common lack of proper communication between data science teams and other business units; communication is one of the key skills not captured by Conway's Venn diagram (Mayo, 2016). The problem for statisticians is that they struggle to communicate their findings without using jargon. Executives, on the other hand, may oversimplify the results presented to them and extract the wrong message. Poor cooperation between business units and the data science team in asking valid and useful questions results in poor data insights (Berinato, 2019).

5. Applications of data science

In Video 1, Dr Etienne Pienaar explores examples of data science in various industries and business areas, such as finance, HR, marketing, sales, operations, and supply chains.

Video 1: Applications of data science. (Access this set of notes on the online campus to engage with this video.)

Were you aware of all the ways data science is being used in these business domains? The number of applications has increased significantly in the last two decades, mostly due to the advances in technology, as mentioned in Section 2, but also due to the software that was developed. The implementations discussed in the video are only possible because of the software tools available in the digital era.

6. Software

As mentioned, computer science is an important part of data science, since most data is stored electronically. There are many programming languages, but some are more suited to performing data science tasks than others. A summary of programming languages and their strengths and weaknesses in the context of data science is listed in Table 1.

Tel: +27 21 447 7565 | **Fax:** +27 21 447 8344
Website: www.getsmarter.com | **Email:** info@getsmarter.com

Table 1: Comparison of popular programming languages.

Programming language	Considerations
Python	• Free • Large community • General purpose language, potentially cumbersome for data-oriented operations • Low entry barrier and easy to learn • Large set of mature libraries • Low computational speed • Possible to integrate with production apps and services • Often used for prototyping before porting to other languages • Reliance on external libraries for statistical operations
R	• Free • Large community • Statistical computation-oriented • Low entry barrier and easy to learn • Large set of mature statistical and visualisation libraries • Low computational speed • Easy interface with C++, C, Fortran • Difficult to integrate with production apps and services • Some statistical learning techniques are ported to R from Python, and are thus delayed and not always well supported

Julia	• Free • Small community • Computation-oriented • High computational speed • Large set of libraries • Libraries are not mature and not all of them are well supported • Lack of documentation for some libraries • Interfaces with C, C++, R, Python • Difficult to integrate with production apps and services
Java, C, JavaScript	• Free • Large community • General purpose languages, not particularly suited for data science • High computational speed • Easy integration with production apps and services
MATLAB	• Expensive • Large community (mostly engineers) • Computation-oriented • Mature and well-maintained commercial libraries • Easy integration with engineering computations • Detailed and well-maintained documentation • Difficult to integrate with production apps and services (unless they are written in MATLAB)

Tel: +27 21 447 7565 | **Fax:** +27 21 447 8344
Website: www.getsmarter.com | **Email:** info@getsmarter.com

While there are many languages used for data science, Python has become one of the most popular languages, and is one of the most consistent competencies listed in data-related job adverts (Cass, 2019; Hayes, 2019).

To run the code for any of these languages, you need one of various applications suited to this task. One tool that is commonly used to not only run code, but also visualise plots, is Jupyter Notebook. Graphs and tables are generated in the same environment as the code, making it easy to explore data.

Explore further:

Jupyter Notebook has become so popular that big companies, such as Netflix, use them in their data science teams.

To simplify the task of coding, there are also a number of pre-written tools, called libraries, that can be used in your own program (depending on the programming language). The following are some of the common libraries used in data science:

- **NumPy:** This library creates arrays that allow for more efficient operations on the data (SciPy community, 2019).
- **Pandas:** This library is used to import and wrangle data into clean NumPy arrays.
- **Scikit-learn:** This library provides a variety of models that can be directly applied to new data.
- **Matplotlib:** This is a popular library used for visualising plots.

7. Conclusion

The era of data science has brought insight to many business problems, facilitated by the decreasing cost of data storage and increased computational power of current computers. Data scientists can bring great value to organisations, so they are in high demand. However, there is a scarcity of people who have the broad skill set required to perform data science optimally. Depending on the needs of an organisation, data science teams may be the way forward, where there are a number of people with knowledge in all three areas, but each person is stronger in a different field.

Continue to the enrichment activity in this unit to explore further examples of where data science is used, and to consider where data science is applied in your business.

8. Bibliography

Anderson, C. 2015. *Creating a data-driven organization: practical advice from the trenches*. Sebastopol, CA: O'Reilly Media, Inc.

Bansal, H. n.d. *Best languages for machine learning in 2020!* Available: https://becominghuman.ai/best-languages-for-machine-learning-in-2020-6034732dd242 [2019, November 11].

Berinato, S. 2019. Data science and the art of persuasion. *Harvard Business Review*. January–February.

Cass, S. 2019. *The top programming languages 2019*. Available: https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019 [2019, September 8].

Colson, E. 2019. Why data science teams need generalists, not specialists. *Harvard Business Review*.

Conway, D. 2013. *The data science Venn diagram*. Available: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram [2019, September 7].

Gonfalonieri, A. n.d. *Why is machine learning deployment hard?* Available: https://towardsdatascience.com/why-is-machine-learning-deployment-hard-443af67493cd [2019, November 11].

Hayes, B. 2019. *Programming languages most used and recommended by data scientists*. Available: https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists [2019, September 8].

Matloff, N. n.d. *R vs. Python for data science*. Available: https://github.com/matloff/R-vs.-Python-for-Data-Science [2019, September 8].

Mayo, M. 2016. *The (not so) new data scientist Venn diagram*. Available: https://www.kdnuggets.com/2016/09/new-data-science-venn-diagram.html [2019, September 7].

McCallum, J. C. 2002. Price-performance of computer technology. In *The Computer Engineering Handbook*. V. G. Oklobdzija, Ed. Florida, USA: CRC Press. 136-153.

Mukherjee, A. n.d. *Minimum viable domain knowledge in data science*. Available: https://towardsdatascience.com/minimum-viable-domain-knowledge-in-data-science-5be7bc99eca9 [2019, September 8].

NASA. 2019a. *Discovery alert! Two new planets – found by AI*. Available: https://exoplanets.nasa.gov/news/1565/discovery-alert-two-new-planets-found-by-ai [2019, September 17].

NASA. 2019b. *Human computers*. Available: https://crgis.ndc.nasa.gov/historic/Human_Computers [2019, September 17].

O'Neil, C. & Schutt, R. 2014. *Doing data science: straight talk from the frontline*. Sebastopol, CA: O'Reilly Media, Inc.

Patil, D. J. 2011. *Building data science teams*. Available: http://radar.oreilly.com/2011/09/building-data-science-teams.html [2019, September 17].

SciPy community. 2019. *What is NumPy?* Available: https://docs.scipy.org/doc/numpy/user/whatisnumpy.html [2019, September 8].