Contents

Preface

I am not a data scientist. These might be strange words to open a book about data science, so please allow me to explain. This book is the result of a twenty-five-year career in civil engineering, building and managing structures in Europe, Africa, Asia and Australia. Most of my tasks involve managing and analysing large amounts of data. Cost estimates, volume calculations, modelling river flows, structural calculations, Monte Carlo simulations and many other types of number crunching are integral to my work as an engineer.

My journey towards what we now call data science started at university. When studying engineering in the Netherlands, I wrote computer code in the Pascal and BASIC languages. I loved spending time in the computer lab and write software to solve technical problems. The dean advised me to switch from civil engineering to computer science, but I enjoyed writing software to solve engineering problems, not for the sake of it, so I did not heed the advice.

In my first job as a civil engineer, my company introduced the now-defunct *Lotus 123* spreadsheet. When first using this package, I thought it was the best thing since sliced bread. Graphical output and managing data were complex tasks when writing code in those days. The allure of the spreadsheet was the ability to combine inputs with computer code and show the results in text or graphs, all in one convenient file.

Over the next two decades, I have written hundreds of spreadsheets to solve a myriad of engineering problems. I even developed a 'jungle' of interconnected spreadsheets to manage the logistics of a large river engineering project in Bangladesh. The complexity of this task took me to beyond limits of what spreadsheets can achieve.

Throughout my career, I had many nightmarish experiences trying to reverse engineer spreadsheets to figure out how they work, even ones I wrote myself. The combination of data, code and output that I loved at the start of my career was now a source of frustration.

My love affair with the venerable spreadsheet ended when writing my doctoral dissertation. Excel was incapable of helping me with complex statistics such as structural equation modelling or network analysis. A colleague suggested looking into this new thing called 'Data Science'. I decided to learn how to write code in R, a specialised computer language for statistical analysis. The R language is like a Swiss army chainsaw for engineers, with capabilities that far exceed anything a spreadsheet can do.

I now manage the data science function for a water utility in regional Australia. Through my experience with practical data analysis and expertise in management, I have developed a strategic approach to data science. I have published and presented my views on strategic data science at conferences in Australia and New Zealand. This book is an expanded version of an article I wrote for the journal of the Australian Water Association, which became one of the most downloaded papers. *Lifting the Big Data Veil* describes a back-to-basics approach on how to maximise the value we can extract from data assets.[1] This book dives deeper into the principles of data science first presented in this paper.

I write this book from the perspective of an engineer and a social scientist, but the same principles are valid for any field of human endeavour. All professionals and scientists rely on data to make decisions. Data science provides a systematic approach to making better business decisions and discovering new patterns in society or nature.

My approach in this book goes back to the basics of what it means to create value from data. This book is not a treatise about machine

[1] Prevos, P. (2017). Lifting the 'Big Data' Veil. Creating Value through Applied Data Science. *Water E-Journal*, 2(1), 1–5. DOI 10.21139/wej.2017.008.

learning, mathematics or developing software, but a practical guide to strategically and systematically using data to create a better world. This book is pragmatic because it doesn't dwell on the future promises of machine learning, artificial intelligence or quantum computing. The framework in this book is inspired by my current and desired practice as an engineer and social scientist, with a data science responsibility and best-practice in management.

The information in this book is by no means the final say on the topic of strategic data science. This book is opinionated in that the approach it describes is only one way of looking at the subject. This book is not an academic dissertation that evaluates various methods to the science and craft of analysing data. My experience and knowledge profoundly influence my views on the topic as a civil engineer and training in the humanities. My objective for this book is to encourage the reader to think beyond the algorithms and shiny data tools and seek a deeper understanding of what it is you want to achieve.

My primary objective is to introduce the reader to a wide range of considerations to improve the way they manage data and to provoke curiosity to research the topic in more depth. My objective is to inspire data professionals and managers to worry less about the technical details and think strategically about extracting value from data.

The raw text of this book and R code for the data visualisations is available on GitHub. I encourage anyone who discovers mistakes or likes to enhance the information in these pages to contact me. The LeanPub publishing system provides flexible opportunities to publish new versions when updated information is available. Anyone purchasing this book through LeanPub can register to be informed of future editions.

Preface to second edition

The second edition of this book contains many grammar fixes. Thanks to David Smith and Catherine Cousins for spotting these mistakes.

Interestingly, the machine learning program I used for checking the text did not identify many of these mistakes in the first version. This experience strengthens one of the points in this book, which is that artificial intelligence is not a replacement for natural intelligence.

Acknowledgements

My career has been dynamic and varied but analysing data has been the one constant in my career. My toolkit to convert data to actionable intelligence has evolved significantly over the past twenty-five years. My former and current colleagues helped me to accumulate new approaches and skills, which led me to write this book.

In my first job as an engineer, Wim van Vliet in Johannesburg was a significant influence on my development as an engineer. Through him, I learned how to convert my theoretical knowledge into a systematic approach to solving real problems.

René Zekveld was my manager at Boskalis when I worked on marine engineering projects in Hong Kong and Bangladesh. René taught me tricks of the trade on how to interpret data to achieve sound business outcomes. He emphasised the importance of the relationship between reality and data. Our long discussions about data showed me that actionable intelligence and insight are essential outcomes.

After my international career, I moved back to the Netherlands and spent time with Rijkswaterstaat, the government agency responsible for keeping the country from flooding. Working for this organisation exposed me to probabilistic approaches in cost estimation and project management. We regularly worked with mathematicians, whose names are lost in time, to help us grasp the complexity of our analysis. Working for this organisation advanced my skills in using mathematics to create value from data.

At the start of the new millennium, I started my current job at Coliban Water, a water utility in regional Australia. Brad Dole,

a former IT manager, has been influential in how I now work with data. As an IT manager, Brad provided me with the freedom to maximise the capabilities of the company hardware. Brad also has a razor-sharp insight into data and how to convert business problems into code. Under Brad's guidance, we developed several in-house software solutions, which took my software development skills beyond a mere hobby.

Jenny Fogarty is my current colleague and the data architect at Coliban Water. We have worked on developing software and reporting mechanisms for many years. Through her expertise, I learned everything I know about managing data. It was on her advice five years ago that I started researching the topic of data science. Not only has her advice helped me to complete my dissertation about customer centricity, but it has also transformed the way I view my profession.

No vision of data science can ever be implemented without people doing the work. Gary Schurr has over the past few years been great at finding innovative ways to report information. His critical helps to straighten out some of my more impulsive ideas.

I started my formal journey into data science by doing a range of courses on the Coursera website. The data science specialisation by John Hopkins University taught me the principles of how to use the R language for statistical computing. The inspiring lectures by Roger Peng, Jeff Leek and Brian Caffo helped me to think more systematically in how I deal with data.

My manager David Sheehan often encourages me to develop new ideas and 'conquer the world'. This book is just another small step in that direction. It would not have existed without the freedom David provides me to shape the data science function at Coliban Water and the broader water industry in Australia.

Lastly, I thank all my colleagues at Coliban Water and the broader Australian water industry who indulge me in my geekiness and

enthusiasm to find data solutions to improve the way we service our customers. This book is the direct result of the positive feedback from the people that have attended my conference presentations and read my publications and articles on my Lucid Manager (lucidmanager.org) website.

1. What is Data Science?

The activity of analysing data is as old as human culture. The earliest known form of writing is not an epic poem or religious text, but data. The Ishango bone is an engraved fibula of a baboon which was carved in central Africa 20,000 years ago. Some scholars hypothesised that the carvings represent an early number system as it lists several prime numbers, while others believe it to be a calendar. Some researchers dismiss these ideas and believe the markings merely improve grip when using the bone as a club. Whatever their purpose, the groupings of the markings are distinctly mathematical (Figure 1).[1]

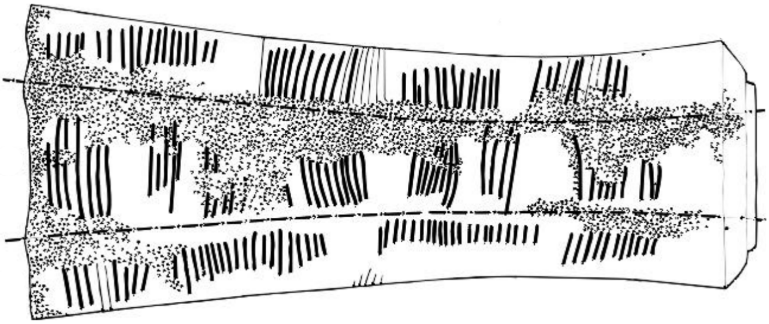

Figure 1: Markings on the Ishango Bone

Ancient cultures around the world collected data by observing nature and the stars to predict when they needed to move camp, start sowing crops, hunt seasonal animals or to obtain whatever other knowledge required for survival. These proto-scientific methods

[1] Pletser, V. (2012). Does the Ishango Bone Indicate Knowledge of the Base 12? An Interpretation of a Prehistoric Discovery, the First Mathematical Tool of Humankind. Eprint ArXiv:1204.1019.

were the first attempts at science as these early researchers collected data to explain the world in logical terms. These primitive forms of science helped these people to understand their world and control their destiny, which is precisely what contemporary science seeks to achieve.

Mathematics was an integral part of ancient civilisations. Sumeria, Egypt, Rome and other advanced ancient civilisations used mathematics to manage their society and build their elaborate cities. The origins of civilisation as we now know it lies in Mesopotamia, current-day Iraq. Archaeologists have excavated thousands of clay tablets that record their day to day activities such as land sales, delivery of goods and other commercial transactions. Around that same time in Pharaonic Egypt, the first census took place, recording demographic data about its inhabitants.[2] These examples show that collecting data and using it to control and improve our world is an ancient human activity.

This time was also a period of the first significant mathematical discoveries and inventions. Mathematics was, however, more than a language to model the world. To the great Greek mathematician Pythagoras, numbers possessed meaning beyond their ability to describe quantity. In these early days of intellectual exploration, divination was the most popular method to predict the future. Astrologers mapped the skies or studied the entrails of a bird to find a relationship between these patterns and their world. In these divination systems, mathematics was practised as a tool to manage society through engineering and bookkeeping, not as a tool to describe the world.

The scientific revolution of the seventeenth century replaced divination with a mathematical approach to understanding the world. Since the work of René Descartes, mathematics is a method to

[2] Kelleher, J.D., & Tierney, B. (2018). *Data science.* Cambridge, Massachusetts: The MIT Press.

describe the world and to predict its future.[3] This revolution in how we perceive the world mathematically enabled the industrial revolution. Early technology enhanced our physical capabilities with machines, while modern technology improves our minds with computers. Machines make us stronger and faster, and their development revolutionised society during the first industrial revolution. Computers enhance aspects of our mental abilities, and we are in the middle of a second industrial revolution which is not fuelled with oil and coal but with data.

The idea that data can be used to understand the world is thus almost as old as humanity itself and has gradually evolved into what we now call data science. We can use some basic data science to review the development of this term over time.

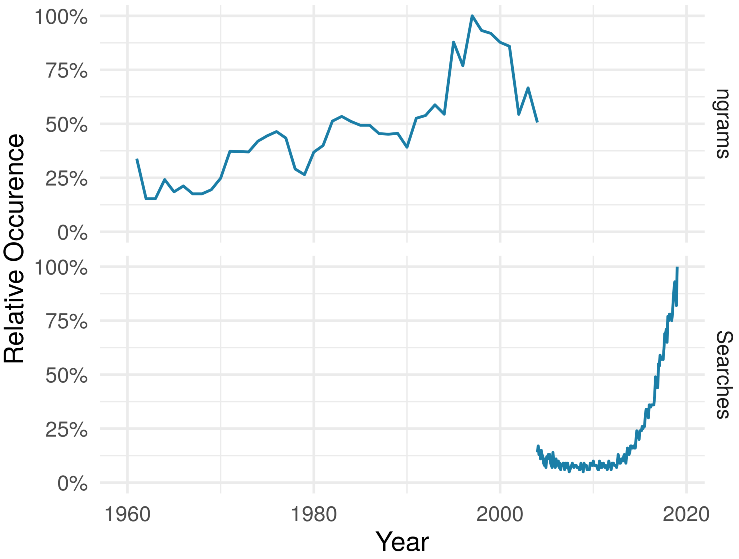

Figure 2: Frequency of the bi-gram 'data science' in literature and Google searches in the percentage of highest occurrence.

[3] Davis, P.J., & Hersh, R. (1990). *Descartes' Dream. The World According to Mathematics*. London: Penguin.