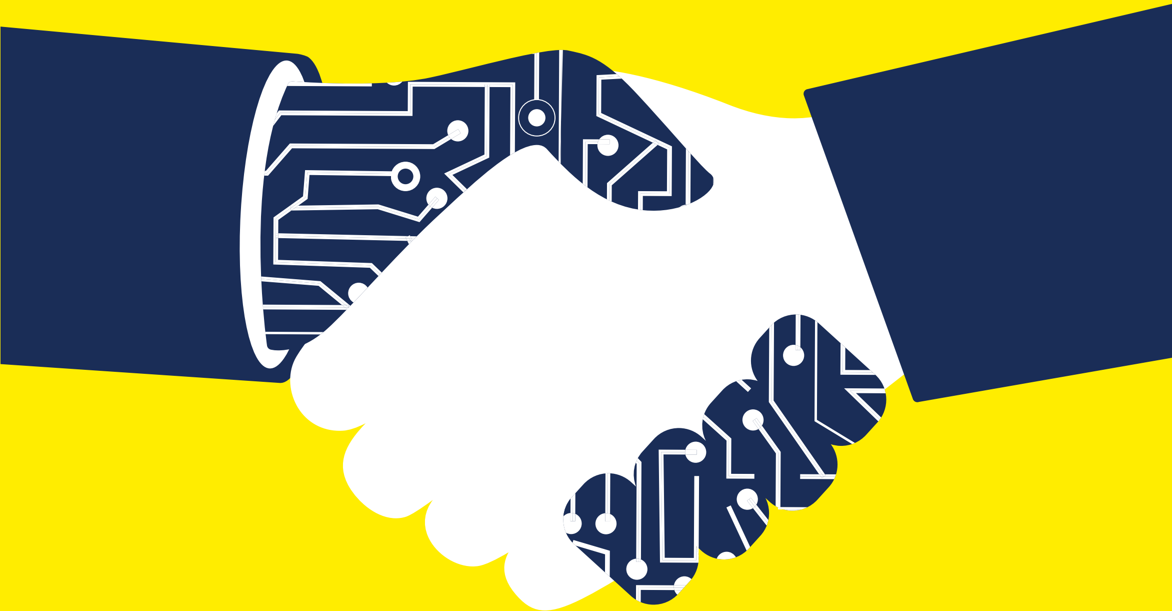

The Decision Maker's Handbook to Data Science

A Guide for Non-Technical Executives, Managers, and Founders

Second Edition

Stylianos Kampakis

Apress®

THE DECISION MAKER'S HANDBOOK TO DATA SCIENCE

A GUIDE FOR NON-TECHNICAL EXECUTIVES, MANAGERS, AND FOUNDERS

SECOND EDITION

Stylianos Kampakis

Apress®

The Decision Maker's Handbook to Data Science: A Guide for Non-Technical Executives, Managers, and Founders

Stylianos Kampakis
London, UK

ISBN-13 (pbk): 978-1-4842-5493-6 ISBN-13 (electronic): 978-1-4842-5494-3
https://doi.org/10.1007/978-1-4842-5494-3

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shiva Ramachandran
Development Editor: Rita Fernando
Coordinating Editor: Rita Fernando

Cover designed by eStudioCalamar

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit http://www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-5493-6. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

Contents

About the Author

Dr. Stylianos (Stelios) Kampakis is a data scientist who is living and working in London, UK. He holds a PhD in Computer Science from the University College London as well as an MSc in Informatics from the University of Edinburgh. He also holds degrees in Statistics, Cognitive Psychology, Economics, and Intelligent Systems. He is a member of the Royal Statistical Society and an honorary research fellow in the UCL Centre for Blockchain Technologies.[1] He has many years of academic and industrial experience in all fields of data science like statistical modeling, machine learning, classic AI, optimization, and more.

Throughout his career, Stylianos has been involved in a wide range of projects: from using deep learning to analyze data from mobile sensors and radar devices, to recommender systems, to natural language processing for social media data, to predicting sports outcomes. He has also done work in the areas of econometrics, Bayesian modeling, forecasting, and research design. He also has many years of experience in consulting for startups and scale-ups, having successfully worked with companies of all stages, some of which have raised millions of dollars in funding. He is still providing services in data science and blockchain as a partner in Electi Consulting.

In the academic domain, he is one of the foremost experts in the area of sports analytics, having done his PhD in the use of machine learning for predicting football injuries. He has also published papers in the areas of neural networks, computational neuroscience, and cognitive science. Finally, he is also involved in blockchain research and more specifically in the areas of tokenomics, supply chains, and securitization of assets.

Stylianos is also very active in the area of data science education. He is the founder of The Tesseract Academy,[2] a company whose mission is to help decision makers understand deep technical topics such as machine learning and blockchain. He is also teaching "Social Media Analytics" and "Quantitative

[1] http://blockchain.cs.ucl.ac.uk/
[2] http://tesseract.academy

Methods and Statistics with R" in the Cyprus International Institute of Management[3] and runs his own data science school in London called Datalyst.[4]

Finally, he often writes about data science, machine learning, blockchain, and other topics at his personal blog: The Data Scientist.[5]

In his spare time, Stylianos enjoys (among other things) composing music, traveling, playing sports (especially basketball and training in martial arts), and meditating.

[3] www.ciim.ac.cy/
[4] www.dataly.st/
[5] http://thedatascientist.com/

Introduction

What is *data science*? What is *artificial intelligence*? What is the difference between *artificial intelligence and machine learning*? What is the best algorithm to use for X? How many people should I hire for my data science team? Do I need a recommender system? Is a *deep neural network* a good idea for this use case?

Having dedicated my career to understanding data, and modeling uncertainty, these questions (and many similar ones) have popped up very often in conversations I am having with CEOs, startup founders, and product managers. There are always three common elements:

1. The people involved have a non-technical background.
2. Their business collects data, or is in a position to collect data.
3. They want to use data science but they don't know where to start.

Data science (and all the fields it encompasses such as AI and machine learning) can transform our world on every level: business, political, and individual. In more than one way, this is already happening. Online retailers know what you are going to like, through the use of recommendation engines. Your photographs get automatically tagged through the use of computer vision. Autonomous vehicles can drive us around with no driver in the seat.

However, this is still only a fraction of the things that are possible with data science. The benefits of this powerful technology will never be reaped, unless the entrepreneurs and the decision makers fully understand how to use it.

Data science is unique in the space of technology in two ways. First in contrast to software development, it is intangible. You can't see a flashy front end, but only the results of a model. Secondly, it is *science*, which means that, in contrast to engineering, it is difficult to set out a perfectly laid plan in advance. Uncertainty is an integral part of data science, and this can make estimations and decisions more difficult. These two factors make the understanding of data science more challenging.

The cloud of buzzwords currently dominating technology does not really help at all with that. I have seen buzzwords such as “big data,” “analytics,” “prediction,” “forecasting,” and many more used without any real understanding of the context surrounding them. This is partly due to aggressive sales tactics that end up confusing rather than illuminating. The result is that many entrepreneurs end up feeling more insecure, since they can’t understand what they need and how much is a fair price to pay. This is why I realized that it is upon us, the data scientists, to take up the role of educators.

This book is meant to be the ultimate short handbook for a decision maker who wants to use data science but is not sure where to start. All case studies outlined are always described with the decision maker in mind. The problem in business is not how to choose the right model from a scientific viewpoint but how to deliver *value*. The data scientist has to make decisions based on trade-offs such as the cost of development (which can include time and hiring), the interplay with business decisions, and the cost of data. The book explains how the decision maker can better understand these dilemmas and help the data scientist make the most beneficial choices for the business.

I hope that after reading this book, the world of data science will no longer be a dangerous landscape dominated by buzzwords and incomprehensible algorithms, but rather a place of wonder, a place where the future lies. I do hope that you will enjoy reading it as much as I enjoyed writing it.

CHAPTER

1

Demystifying Data Science and All the Other Buzzwords

In the business world, data has become a big thing. You hear all sorts of buzzwords being thrown around left, right, and center. Things like *big data*, *artificial intelligence*, *machine learning*, *data mining*, *deep learning*, and so on. This can get confusing, leaving you paralyzed as to what is the best technology to use and under what circumstances. In this chapter, we are going to demystify all these buzzwords, by taking a short stroll through the history of data science. You will understand how the history of data analysis gave birth to different schools of thought and disciplines, which have now all come together under the umbrella of the term **data science**.

S. Kampakis, *The Decision Maker's Handbook to Data Science*,
https://doi.org/10.1007/978-1-4842-5494-3_1

What Is Data Science?

In 2017, we generated more data than we did over the previous 5000 years of our history.[1] That's a lot of data. And it's not surprising. Every device we own generates data, and all our interactions with said devices generate even more data.

So, you take a picture with your smartphone? You've generated data. You read the news on your tablet? You're generating data. You listen to a podcast on your laptop? You're generating data. You go on Facebook to update your status? You've generated data.

You get the point. There's a good chance that the only thing we do that doesn't generate data is breathe, but even that is debatable considering all the wearable devices that are available and which can track everything from heart rate to calories burned.[2]

What happens to all this data? It must have some use or things wouldn't be set up so that we generate it in the first place.

At the moment, some of it does and some of it doesn't. In fact, some studies have discovered that less than 0.5% of data is being analyzed and turned into actionable insights.[3] But more on that later.

The important thing is that all data *can* be used. We just need to figure out better ways of doing so.

And this is where data science comes in. There are many definitions of data science. One such definition is

> *Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.*[4]

Okay, so those are a lot of big and fancy words, which can get confusing. Let's boil it down to the simplest definition:

> *Data science is about using data to do useful stuff.*

[1] David Sønstebø, "IOTA Data Marketplace," *IOTA*, November 28, 2017, https://blog.iota.org/iota-data-marketplace-cb6be463ac7f

[2] www.wired.co.uk/article/hospital-prescribing-tech

[3] www.technologyreview.com/s/514346/the-data-made-me-do-it/

[4] Chikio Hayashi, "What is Data Science? Fundamental Concepts and a Heuristic Example," in *Data Science, Classification, and Related Methods*, eds. Chikio Hayashi, Keiji Yajima, Hans-Hermann Bock, Noboru Ohsumi, Yutaka Tanaka, and Yasumasa Baba (Tokyo, Japan: Springer-Verlag, 1998), 40-51.

Short and to the point. That's exactly what data science is all about. The methods we use are important, of course, but the essence of this discipline is that it allows us to take data and transform it so we can use it to do useful things.

For example, let's say someone visits a doctor because they are short of breath. They are experiencing heartburn and they have chest pains. The doctor will run the basic tests, including measuring blood pressure, but nothing seems out of the ordinary.

Since the patient is overweight, the doctor immediately assumes the symptoms are caused by the patient's size and recommends a healthier diet and exercise.

Three months later, the same patient is brought into the emergency room and ends up dying on the table because of a heart defect.

This might sound like an episode on your favorite medical show, that is, a work of fiction, but it happens much more often than you might think. In fact, in the United States, 5% of patients are misdiagnosed, while misdiagnosis cost the United Kingdom over £197 million in the 2014/2015 fiscal year.[5]

However, this situation can be avoided thanks to data science. The analysis of similar cases reveals that the symptoms our patient exhibited aren't just caused by obesity but also by some cardiovascular conditions.

Having access to systems that can analyze data and compare it with new data inputs could have helped the doctor identify the problem sooner without relying solely on their own personal experience and knowledge.

So, data science can be used to save lives, among many other applications, which is pretty useful.

Data Science Is Multidisciplinary

Data science involves multiple disciplines, which is why finding someone with the necessary skills to be a data scientist can be difficult.

Thus, data science involves everything from statistics and pattern recognition to business analysis and communication. It requires creative thinking as much as it requires analytical thinking.[6]

[5] Lena Sun, "Most Americans Will Get a Wrong or Late Diagnosis At Least Once In Their Lives," *Washington Post*, September 22, 2015, www.washingtonpost.com/news/to-your-health/wp/2015/09/22/most-americans-who-go-to-the-doctor-will-get-a-wrong-or-late-diagnosis-at-least-once-in-their-lives-study-says/; "The Top Misdiagnosed Conditions In NHS Hospitals In 2014/15", *Graysons*, www.graysons.co.uk/advice/the-top-misdiagnosed-conditions-in-nhs-hospitals/

[6] Take a look at this infographic by Brendan Tierney: https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/

So, data science involves discovering which data is useful as well as effective ways of managing it. It also requires determining how the data should be processed and what types of insights can be garnered from the massive amounts of data available.

Data science requires knowledge of programming and computing, but also visualization so that the insights can be presented in a way that everyone can understand.

Furthermore, business acumen is also a necessity because while data science can be applied to any field of business, it is critical to know what types of answers the business needs and how to present said insights so leadership can understand them.

Core Fields of Data Science

Data science has three core fields, namely, artificial intelligence, machine learning, and statistics.

Artificial intelligence is all about replicating human brain function in a machine. The primary functions that AI should perform are logical reasoning, self-correction, and learning. While it has a wide range of applications, it is also a highly complicated technology because to make machines smart, a lot of data and computing power is required.

Machine learning refers to a computer's ability to learn and improve beyond the scope of its programming. Thus, it relies on creating algorithms that are capable of learning from the data they are given. They are also designed to garner insights and then make forecasts regarding data they haven't previously analyzed.

There are three approaches to machine learning, namely, supervised, unsupervised, and reinforcement learning, plus some subfields (such as semi-supervised learning). Here, we will be talking only about supervised and unsupervised learning, since this is what is mainly used in business.

Let's say you want to sort all your photographs based on content. In supervised learning, you would provide the computer with labeled examples. So, you'd give it a picture of a dog and label it animal. Then you'd feed it a picture of a person and label it human. The machine will then sort all the remaining pictures.

In unsupervised learning, you'd just give the machine all the photos and let it figure out the different characteristics and organize your photos.

In reinforcement learning, the machine learns based on errors and rewards. Thus, the machine analyzes its actions and their results. A good example of reinforcement learning is a rat in a maze that needs to navigate its way to a piece of cheese. The learning process that helps the rat achieve this can be implemented in a machine through reinforcement learning. This is one of the most esoteric types of machine learning.

Statistics is an essential tool in the arsenal of any data scientist because it helps develop and study methods to collect, analyze, interpret, and present data. The numerous methodologies it uses enable data scientists to

- Design experiments and interpret results to improve product decision-making
- Build signal-predicting models
- Transform data into insights
- Understand engagement, conversions, retention, leads, and more
- Make intelligent estimations
- Use data to tell the story

Let's take a closer look at all three.

Artificial Intelligence: A Little History

In 1954, the field of AI research came into being at a workshop at Dartmouth College. There, the attendees discussed topics that would influence the field for years to come.

As we've already explained, the goal of artificial intelligence is to create a "thinking machine," that is, one that emulates human brain function. To do this, of course, one needs to understand the human mind, which is why AI is closely related to the field of cognitive science.

Cognitive science involves studying the human mind and its processes, including intelligence and behavior. Memory, language, perception, attention, emotion, and reasoning are all studied, and to gain greater understanding of these faculties, scientists borrow from other fields, including

- Linguistics
- Psychology
- Philosophy
- Neuroscience
- Anthropology
- Artificial intelligence

Artificial intelligence played a key part in the development of cognitive science, and there was a large interplay between cognitive psychology and AI. The understanding of human cognition helped us improve our understanding of how to transfer this inside machines. Vice versa, the computational theory of the mind[7] was one of the dominant paradigms in cognitive science. According to this theory, the mind works like a computer, with processes and limited memory. While this is now considered outdated, it drove research for decades.

The AI Dream

It all started with a grand vision. Marvin Minsky, the head of the artificial laboratory at MIT who was considered the father of AI, stated that "artificial intelligence is the science of making machines do things that would require intelligence if done by men."[8] In other words, the dream was to create an intelligent machine.

So, how would one go about doing it? First, we start with intuition, that is, the human's distinct ability for logical reasoning.

Logical reasoning is, essentially, the capacity to reason based on various premises to reach a conclusion that has logical certainty. For example:

> *If all men are mortal, and Socrates is a man, then Socrates is mortal.*

Or:

> *If it's raining outside and I don't have an umbrella, I will get wet if I go out.*

To translate this logical deduction ability to a machine, a rule-based or symbolic approach is used. This involves humans constructing a system of rules with which the computer is programmed. Using these rules, reasoning algorithms are capable of deriving logical conclusions.

A good example is MYCIN,[9] which was an early successful working system based on reasoning algorithms. It was used to diagnose infections and determine the type of bacteria that was causing the problem. It was never used in a clinical setting but is an excellent example of an expert system and a predecessor to machine learning.

[7] https://plato.stanford.edu/entries/computational-mind/

[8] Blay Whitby, *Reflections on Artificial Intelligence* (Exeter, UK: Intellect Books, 1996).

[9] A good old reference for MYCIN by John McCarthy can be found here: www-formal.stanford.edu/jmc/someneed/someneed.html

The system was developed in the 1970s at Stanford University and had approximately 600 rules.[10] Users were required to provide answers to various questions and the program would then provide a list of potential bacteria that could be causing the problem, sorted from high to low probability. It would also provide its confidence in the probability of each diagnosis as well as how it came to the conclusion. Finally, it would provide the recommended course of treatment.

It had a 69% accuracy rate, and it was claimed that the program was more effective than junior doctors and on the same level as some experts.[11]

The program was created by interviewing a large number of experts who provided their expertise and experience. It used rules of the IF (condition) THEN (conclusion) form. For example, IF (sneezing and coughing or headache) THEN (flu).

One limitation the program had was computing power. It took approximately 30 minutes to go through the system, which was too much wasted time in a real-world clinical setting.

Another issue was also raised, namely, that of ethics and legal issues. Thus, the question arose of who would be held responsible of the program made the wrong diagnosis or recommended the wrong treatment.

Though it was never used, MYCIN still had a very important role in bringing us to where we are today as it was one of the early successes of AI, proving what is possible and strengthening.

Automated Planning

Planning is a vital component of rational behavior. Automated planning and scheduling is an area of artificial intelligence that involves the creation of a system that is capable of selecting and organizing actions to achieve a certain outcome.

An example is the Missionaries and Cannibals problem,[12] which is a classic AI puzzle. It is defined as follows:

> *On one bank of a river are three missionaries and three cannibals. They all wish to cross to the other side of the river. There is one boat available that can hold up to two people. However, if the cannibals ever outnumber the missionaries on either of the river's banks, the missionaries will get eaten.*

[10] Bruce G Buchanan and Edward H Shortliffe, *Rule-Based Expert Systems* (Reading, MA: Addison-Wesley, 1985).

[11] Victor L. Yu, "Antimicrobial Selection By A Computer," *JAMA* 242, no. 12 (1979): 1279, https://jamanetwork.com/journals/jama/article-abstract/366606

[12] You can play a version of the game here: www.novelgames.com/en/missionaries/

> *How can the boat be used to safely carry all the missionaries and cannibals across the river?*

It's a little gruesome and might also seem trivial, but a similar approach can be used for schedule planning.

Other similar problems in that are the towers of Hanoi[13] and the travelling salesman problem. The travelling salesman problem is a legendary benchmark in optimization, where the objective is to find a path for a salesman to go across all the cities in a country (or some other geographical regions). This path should never go through the same city twice, and at the same time it should as short as possible. It is easy to see how this relates to vehicle routing in real life. In Figure 1-1, you can see an example of the travelling salesman for all major cities in Germany.

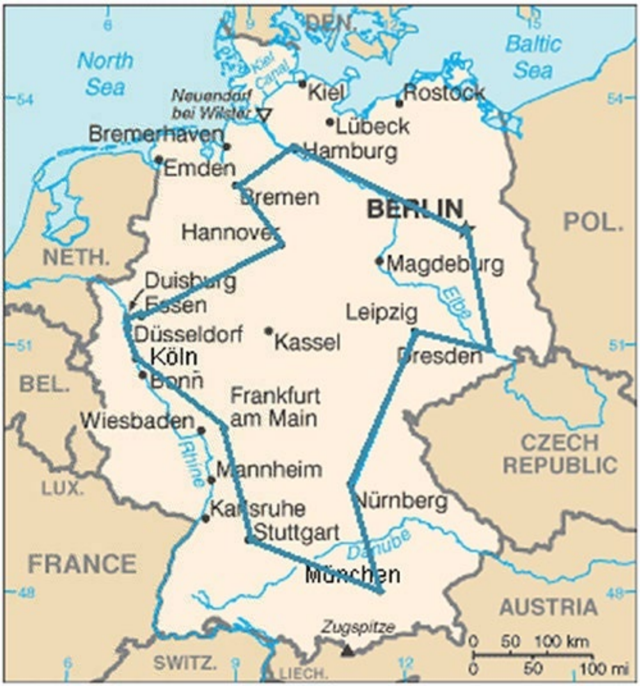

Figure 1-1. Example of the travelling salesman problem

Essentially, what the computer does is analyze each possibility, discarding the one that doesn't fulfill the parameters while presenting all the options that do.

[13] www.mathsisfun.com/games/towerofhanoi.html