Designing and Building Data Science Solutions

A Practical Guide

Jonathan Leslie Neri Van Otten

2020-07-23

Contents

Welcome

Data science is a dynamic field that, when successful, can have a game-changing impact for a business. Yet designing and delivering a successful data science project can be difficult. Indeed, in a recent Gartner study, 85 percent of all data science projects fall short of expectations. The main cause of failure was found to be that replacing legacy systems and associated processes is not easy. Other found issues revolved around not having the right data for the project or not having the right talent. This grim statistic underscores the fact that, for a business, getting started on a data science journey can be risky and daunting.

With this guide, we wish to de-risk the process and help practitioners succeed by providing a step-through guide on how to best tackle the entire process from designing to delivering a successful project. This guide can also be useful for managers, executives and project managers, they will gain insight into how the entire data science project process works.

When building groundbreaking systems or conducting cutting-edge research, one can never be sure where the work will take you. As a data scientist, you can never take all of the uncertainty and risk away. But you CAN approach your projects in a sensible way that can give your projects the best chances of success.

This guide covers our approach. It includes two different frameworks: one for evaluating project success and the other for navigating a data science project lifecycle. We use this system in our own work and have found it to be invaluable

in helping us to avoid common pitfalls and to give our projects the best chances for impactful outcomes. We hope it is useful to you as well.

About us

Your authors have each worn many hats: data scientists, project managers, mentors and consultants. We have designed and executed hundreds of projects using data science, machine learning and artificial intelligence (AI), most of which we would consider "successful" (you can read much, much more on the topic of "success" in Chapter **??**).

Like many data scientists, we have both come to the field via routes in other disciplines. Jon obtained his PhD in Biology from the University of London, studying blood vessel formation at the Cancer Research UK London Research Institute. After 22 years in biomedical research, he turned to data science and founded a freelance consultancy business. In 2017 he joined Pivigo, where he is now the Director of Data Science. He is passionate about promoting open-source software and routinely volunteers as a mentor in the R-programming and data science communities.

NERI'S BIO GOES HERE

We met via the S2DS programme, where they worked together as project mentors.

We have also made countless mistakes along the way and have done our best to learn from them. Designing projects, ensuring stakeholder buy-in and delivering successful, relevant outcomes is a difficult task, and it doesn't come naturally to anyone. We hope that you can benefit from the mistakes we have made and the learnings we have acquired in our own careers.

Acknowledgements

We would like to thank a number of people who have contributed to this work. Without you, this would never have happened. We thank Andras Szabo and Kim Nilsson, who spent many hours carefully reading our drafts and providing invaluable feedback. We would also like to thank Ole Moeller Nilsson, Maryam Qurashi, Mandeep Soor, Jason Muller, Neil Forest and Deepak Mahtani for their many suggestions and useful comments. We are especially grateful to our families for their patience and support.

This book is written in RMarkdown with bookdown.

Chapter 1

Introduction

The past decade has seen an explosion of technological innovations, perhaps none of which more seismic for both businesses and individuals than the field of data science. Indeed, the ability to apply advanced analytics to business challenges can be exciting, fruitful and fun. With recent advances in computational capabilities and cost-effective data collection and storage solutions, applying data science to business challenges is now within the reach of most business owners.

Getting started with data science can be daunting, and for the non-specialist, exactly how to begin a data science journey can be unclear. Just like software engineering projects, data science projects require specific design strategies. Our personal experience designing data science projects in a wide range of industries and sectors has given us an understanding of how to make this journey successful and how to work with stakeholders in order to identify the most impactful business questions and formulate scientific approaches to answer them.

In this book, we share our learnings about sensible approaches to designing data science projects. We offer a framework that we have found to be useful in ensuring successful project outcomes and walk the reader through the process

of using this framework for their own data science endeavours. Our goal is to provide a resource that data science practitioners can use in their own work and give business leaders an insight into the steps that go into building a data science project from scratch.

1.1 The challenges when embarking on a data science project?

For data scientists who are new to the field, perhaps the single most challenging aspect of the role is not technical but rather conceptual: learning how to design a successful data science project. Often this means breaking down a complex business case into concrete objectives and specific questions. In our roles as mentors and project managers, we too-often see data scientists attack an analysis without first identifying the underlying questions. These can be specific questions about the scientific approach – "What is the hypothesis of this experiment?" or "What statistical question am I asking my model to answer? – to more general ones, such as, "Why are we doing this?", "What is an ideal outcome?", "What would constitute a failure?" or "What is the business problem we are trying to solve?" As we tell our mentees, this approach is like building a house without a blueprint: you might hammer together some bits of wood in a useful way, but without understanding the objectives of the project as a whole, success is essentially impossible.

Project design is difficult: it can be loosely structured and often has no single correct answer. This can often be disorienting. How-best to approach this task has been considered and discussed often, and many learnings have been derived from the methodologies found in fields such as product design or design engineering. For an excellent discussion of this topic, we recommend eppisodes 63 - 70 of the podcast Not So Standard Deviations, in which hosts Roger Peng and Hilary Parker discuss the Nigel Cross book Design Thinking and draw

parallels to applications in data science.

We often think of data science as a process by which we frame a business problem as a scientific question and apply scientific methodologies to answer that question and derive insights. But how do we identify the business question in the first place? As a data scientist, a good first step is to ask different stakeholders. Yet in our experience, this can often be an unsatisfying approach: more often than not, our clients/managers will not have a clearly-defined business problem or a concrete objective for the project. We believe our framework can help drive this conversation, setting the stage for well-planned project design and giving projects the best chances for success.

1.2 A note on data science, machine learning and artificial intelligence

"Data science", "machine learning" and "artificial intelligence" are terms that can be used in imprecise ways and can have overlapping meanings. Many will be familiar with the phenomenon of a job advertisement that is billed as a data scientist role but in reality is more of a data analyst or an IT specialist. AI is perhaps the most liberally-used of the terms, sure to increase one's chances of writing a successful grant application or tender for a piece of work. It is safe to say that all of these terms can be susceptible to over-hype, and the choice of which to use often appears as a matter of marketing.

But there are differences between them and it's important to understand what those differences are. One of our favourite discussions about how these terms relate to one another comes from David Robinson:

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

To expand on this slightly, the goal of data science work is to gain insights and understanding into data. While the numerical patterns revealed may be clear and objective, the way these findings are interpreted requires a human in the loop. One way it differs from machine learning is that data science doesn't necessarily involve modelling. We would also argue that it doesn't necessarily involve coding or programming: there are plenty of data scientists that combine domain expertise with statistical inference using spreadsheets to acquire valuable insight.

Machine learning extends data analytics into the realm of predictive modelling. This can be done in a highly automated way, although no machine learning system should be allowed to impact decision-making without a human in the loop. Machine learning models can range in complexity and interpretability. For example, linear regression is at the straight-forward end of the spectrum, so much so that many do not consider it machine learning at all. Deep learning models reside at the opposite end of the spectrum, with inner workings that are so opaque that it is essentially impossible to understand how the model makes its predictions. The difference between data science and machine learning is clearly not black and white but rather a spectrum or two domains that tend to overlap.

xkcd.com Machine Learning

Artificial Intelligence is the most common term that is most widely understood but is possibly also the hardest to distinguish between. Everyone will have their own view on this but to distinguish it from the other two terms we will define artificial intelligence as an actional application. This can be anything from robotics used in industries, chatbots providing customer service to game-playing algorithms using reinforcement learning.

For the purpose of the book, we use the term data science throughout, just because we feel that this most accurately represents the type of projects we tackle but the reader can substitute the term with machine learning or artificial

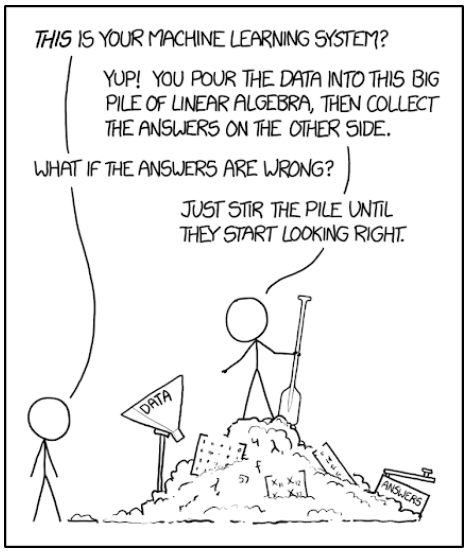

Figure 1.1: (ref:xkcd)

intelligence to suit their own needs.

1.3 Our framework

Our framework has two related components: **project execution** and **project evaluation**. While the temptation may be to view these in series, we have found that these components are in fact intertwined; each has a use on its own, but the framework yields the best results when they are used together. In the following chapters we outline these components, counterintuitively starting with evaluation and then moving on to execution. Why do we start with evaluation? We have found that thinking about what you want the final successful project to look like can help in planning how to get there. In short, we recommend starting with a zoomed-out view of the project as a whole and the context into which fits before considering the details of how you might get there. We can think of the analogy of taking a road-trip across the country: a sensible approach is to first think about what you want to get out of the journey and what are the

major milestones you want to achieve before determining exactly which roads you will take to get there. This has echoes in Test Driven Development (TDD), a common approach to software development in which one first defines the criteria that the final code should adhere to and creates the corresponding tests before writing the actual code.

1.4 Organisation of this book

This book is divided into two parts. In Part 1 we discuss our approach to designing and executing data science projects. Chapter **??** covers the four levels of project evaluation in more detail and provides examples of how to assess each. In Chapter **??** we move on to the phases of project delivery. Two of these phases are then explored in more detail: Project Definition (Chapter **??**) and Project Execution (Chapter **??**).

Part 2 focuses on offering more concrete, practical advice. In Chapter **??** we provide an example of a project proposal that we might use when working with a client to define a project. Chapter **??** is primarily aimed at independent contractors/freelancers and covers how to build up a client base and find work. Chapter **??** lists some useful resources for where you can find help in the wider data science community.

Part I

Chapter 2

What success looks like: the four levels of project evaluation

When approaching a data science project, we often look to the end: knowing what a successful outcome would look like is a good way to determine the direction of a project and the steps required to get there. The idea is that if you meet those criteria – if you deliver something that resembles this desired outcome – you have succeeded. However, while faithful production of the agreed-upon deliverables is, indeed, important, this is often only part of a much larger picture.

In early conversations with a stakeholder, you may have agreed upon a certain set of outcomes and deliverables for the project based on the circumstances and understanding at the time. However, circumstances can change. Similarly, simply because you have achieved a predetermined goal for a project does not necessarily mean that this goal is the best outcome for the business. Part of our jobs as data scientists, especially if tasked with designing a project, is to

understand what the business needs and help stakeholders clearly see what is the most beneficial. Thus, simply focusing on deliverables falls short when assessing how successful a project has been. Similarly, we often find that while we may not find exactly what we expect, we almost always come up with something valuable.

Eskander Howsawi and colleagues have proposed a framework for evaluating project success that has four levels, termed context, business, product and project process (**ref; Figure 1). In our approach to data science project design, we have found striking parallels to the Howsawi framework. (For a more detailed explanation of these levels and this framework, we recommend reading the paper.)

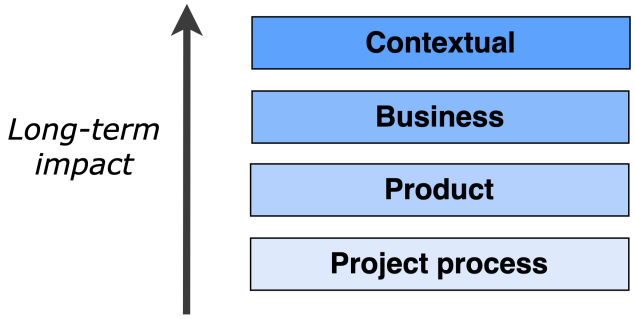

Figure 2.1: **The four levels of project evaluation.** We evaluate project outcomes at four levels: project process (relating to the execution of the project), product (relating to the delivered final product), business (related to what value the project brings to the business) and contextual (related to circumstances surrounding the project). The higher levels (contextual, business) are more abstract, but also have more relevance to the business value of the work. We recommend you consider all of these when defining, and evaluating, a project. (Adapted from Eskander Howsawi and colleagues.)

In the sections below we discuss each of these levels in detail. If you are involved with defining and scoping a project, we strongly recommend you think carefully about how your proposed work satisfies each of these levels. As a consultant, this will be an important part of your role and an important step in developing a project that has real value to its stakeholders. Even if your specific role does

not involve project design, we recommend you take the time to consider these levels at both the beginning and the end of a project: if nothing else, it's a good thought exercise that will be important if you are ever involved in project design.

2.1 Contextual level

The first and highest level of the project evaluation framework is termed contextual and is the most abstract. This relates to the circumstances surrounding a project and the externalities that affect it. While taking this into consideration may seem beyond the remit of the data scientist's role, this level is arguably the most impactful: if a project delivers business value, but the circumstances upon which that value is based change, the realised value may change dramatically. Consider, for example, how Brexit may affect UK businesses: project outcomes that may have been valuable in 2016 may not hold up after the UK leaves the European Union. Similarly, the COVID-19 pandemic has had seismic effects on many businesses, changing the ways in which they operate and the landscape of business opportunities before them. Projects aimed at pre-COVID business cases may no longer have contextual relevance. Understanding and adjusting for these changing circumstances helps ensure that your work is topical, relevant and impactful.

Contextual considerations often are related to business strategy, a good understanding of which is key when designing a project that is going to have an impact well into the future. A company's business strategy describes it's vision, culture and image. At its core is an understanding of the business's goals and how its leaders intend to achieve those goals. Naturally, business goals are often centred around performance: attracting customers, increasing profits and reducing costs are almost always major driving forces. But the strategy can also extend beyond that to include things such as organisational culture, brand and image and the company's place in the wider community. It is important

to note that the underlying goals driving strategy can vary across sectors. For example, the goals of a government agency will surely be very different from those of private-sector or not-for-profit organisations.

Understanding an organisation's strategy will help you to more clearly see how a data science project fits in. Often such projects are part of an overall move towards innovation, so it can be helpful to clarify what the business is hoping to achieve with that initiative. In our experience, we generally turn to a number of key questions that can help paint a picture of what the business strategy is:

- What are your business drivers and your strategic imperatives?
- What are the main pain points or challenges in delivering your strategy?
- Who are your competitors, and what do you think you need to do to stay or get ahead of them?
- Where do you think you may be missing opportunities?
- Do you have any particular business objectives in mind, such as increasing revenue, reducing costs or improving your products or services?
- Do you have any ongoing data science work already? What is its focus, and how does the current project fit into it?

These questions are ones that we, ourselves, often use to help understand the context of the project. They are gleaned from a number of existing frameworks that can help you to identify the business strategy. We have highlighted a few below.

2.1.1 SWOT analysis

SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis is a simple, yet powerful tool that is often used to develop a business strategy. While your job as a data scientist or a consultant is not to define your client's overall business strategy, the exercise of going through a SWOT analysis can help you to understand how the company views its position in the market. In short,

this analysis allows you to identify internal capabilities (strengths and weaknesses) and external factors (opportunities and threats) in order to understand the business's competitive advantage and the factors that are favourable or unfavourable to achieving its objectives. LivePlan has an excellent blog post that provides an overview of the process along with example and questions to help drive the conversation.

2.1.2 Porter's Five Forces

Porter's Five Forces framework has echoes in the SWOT analysis described above. Indeed, it's originator, Micheal E. Porter, developed it in reaction to SWOT analysis, which he felt fell short in analyzing competition of a business. It is generally used to assess an industry in terms of its potential for profitability. While this framework can be a powerful tool, in our experience it is less useful in identifying the contextual environment of how data science, or technical innovation in general, fits into a business's strategy. Nevertheless, we mention it here for the benefit of those readers who would like to learn more.

2.1.3 PESTEL Analysis

PESTEL analysis is used to understand the external forces that an organisation faces. The acronym stands for Political, Economic, Social, Technological, Environmental and Legal. The premise is that organisations that are more tuned-in to the changes in these forces will be better positioned to compete. It is often used in conjunction with a SWOT analysis and, when used well, allows an organisation to not only identify these relevant forces but also to assess the potential impact that they may have.

We have outlined some tools that can help give you a better understanding of the circumstances and business motivations driving an organisation's decision to embark on a data science project. You don't necessarily have to run through

the formality of a consultation session or a workshop with your client, but our very strong advice is to at least go through the exercise of considering some of the questions we have provided. It will give you a better understanding of how your work fits into the business as a whole and will demonstrate to your client that you are willing to take the time to fully understand the forces that drive their decisions. Usually your client is looking to you to guide them on their data science journey, and knowing the larger context of your work will go a long way to ensuring that the outcome is relevant and valuable.

2.2 Business level

The second level of the framework corresponds to the business. In short, this describes how much value the project brings to the business. Unlike concrete deliverables, which are generally tangible, business value can be hard to measure. Success on this level may not be realised immediately, but rather may only be understood well after the project has been completed.

How does one plan for business-level success when designing a project? This is a complicated question with no single right answer, however, this is often where the creative beauty of project design comes into play. To do this well, you will want to engage with stakeholders to identify opportunities for business value based on an understanding of the organisation's strategy, therefore the identification of business-level goals should be thought of as a collaborative effort between the data scientist and the business partner.

Naturally, projects must be feasible in order to generate business value. Thus when defining the business case in Phase 1 (Chapter 4), you will often find yourself moving between high-level discussions about what sorts of outcomes would be useful to the organisation and more concrete discussions about feasibility in terms of objectives, budget, data availability and appetite for risk. We discuss ways to drive this conversation during the stages of project design in Chapter 4.