# Succeeding with AI

## How to make AI work for your business

Essential Excerpts

Veljko Krunic


+

# *Succeeding with AI*

## HOW TO MAKE AI WORK FOR YOUR BUSINESS

### Essential Excerpts

VELJKO KRUNIC

MANNING
SHELTER ISLAND

For online information and ordering of this and other Manning books, please visit www.manning.com. The publisher offers discounts on this book when ordered in quantity. For more information, please contact

Special Sales Department
Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964
Email: orders@manning.com



Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964

| | |
|---|---|
| Acquisitions editor: | Mike Stephens |
| Development editor: | Marina Michaels and Jennifer Stout |
| Technical development editor: | Al Krinker |
| Review editor: | Ivan Martinović |
| Production editor: | Anthony Calcara |
| Copy editor: | Carl Quesnel |
| ESL editor: | Frances Buran |
| Proofreader: | Keri Hales |
| Typesetter and cover designer: | Marija Tudor |



ISBN 9781633437524
Printed in the United States of America

# *brief contents*



Full book available at this link.

# contents

Full book available at this link.

# Drive mission-critical results and improve customer experiences with Azure AI today.

Azure AI offers AI solutions for businesses that are designed for performance and safety. Deploy artificial intelligence in the cloud, hybrid or in containers with flexible pricing options. Customise or develop machine learning models and try out products in demo Studios. Utilise Azure AI to create scalable and accessible experiences.

Start planning your strategy with Azure AI

Microsoft Azure

# *about the author*

VELJKO KRUNIC is an independent consultant and trainer specialising in data science, big data and helping his clients get actionable business results from AI.

He holds a PhD in computer science from the University of Colorado at Boulder and an additional MS in engineering management from the same institution. His MS degree in engineering management focused on applied statistics, strategic planning and the use of advanced statistical methods to improve organisational efficiency. He is also a Six Sigma Master Black Belt.

Veljko has consulted with or taught courses for five of the Fortune 10 companies (as listed in September 2019), many of the Fortune 500 companies, and a number of smaller companies, in the areas of enterprise computing, data science, AI and big data. Before consulting independently, he worked in the professional services organisations (PSOs) of Hortonworks, the SpringSource division of VMware and the JBoss division of Red Hat. In those positions, he was the main technical consultant on highly visible projects for the top clients of those PSOs.

# *foreword*

Hello, my name is Jessica Hawk, Corporate Vice President of Data and AI at Microsoft Azure. Thank you for downloading this special edition of 'Succeeding with AI: How to make AI work for your business.' In this book, AI consultant and visionary Veljko Krunic shares his tested process for planning and running cost-effective, reliable AI projects that produce real business results. Based on his experience working with dozens of start-ups, established businesses and Fortune 500 giants, this practical guide reveals secrets for maximising the return on data-scientist and developer hours and implementing effectiveness metrics to keep projects on track and resistant to calcification.

As a technical or business decision maker, you play a crucial role in adopting and implementing AI in your organisation. Azure AI, the world-class platform for enterprises from Microsoft, is being utilised in various ways to improve experiences for employees, users and customers. Our capabilities are backed by Microsoft research and tested at scale in our apps, including speech transcription and captioning in Teams, content and design production in PowerPoint, biometric detection and identity verification in Windows Hello, personalised recommendations in Xbox, content reading and writing experiences in Edge and M365, text-to-speech and speech-to-text in Office, real-time language translation in Skype and language detection and translation in Outlook. We are also excited about our new Azure OpenAI Service featuring enterprise ChatGPT.

Azure AI offers a simpler way to access responsibly built AI capabilities without requiring extensive data science and development skills. With Azure AI, you can deploy artificial intelligence in the cloud, hybrid or in containers, and access a range of pre-built and customisable services to meet your business needs. In the

last few years, we have been inspired by the way our customers achieve success with AI for a variety of use cases. A few highlights:

- Using Azure OpenAI Service, CarMax has streamlined the creation of text summaries for its car research pages, quickly providing customers with meaningful content that also boosts the pages' search engine rankings. The company's initial goal was customer review summaries for 5,000 car pages. With the existing manual process, it would have taken about 11 years of content generation. With OpenAI Service, CarMax hit the goal in just a few months.
- Fujitsu improved the performance and accuracy of its cloud scanning solution by incorporating Azure Form Recogniser, which boosted character recognition rates for handwritten text up to 99.9%
- Ecolab used Azure Bot Service, Language Understanding (LUIS), Azure Cognitive Search and QnA Maker to build an intelligent virtual agent for digital support, resulting in a 12% drop in call volume from the field
- KPMG used Azure Cognitive Services to reduce the time to identify compliance risks in contact centre calls from 14 weeks to two hours
- Volkswagen Group used Azure Translator and Cognitive Services for Language to translate an average of 325 documents per day, from one-page files to one-million-character books, without the need for manual review
- Twitter used Cognitive Services and the Speech-to-Text service to caption live conversations for accessibility on Spaces and reach broader audiences

In addition to providing top-quality AI solutions, Azure AI is committed to trustworthiness, reliability and accessibility. We understand the importance of data privacy and security, and we work hard to ensure that our products and services meet the highest standards in these areas.

After reading Krunic's book, we invite you to explore Azure AI and see how it can help you achieve mission-critical results and improve customer experiences. Whether you're looking to plan your AI and ML strategy with a specialist, demo what Azure AI can do for your business or train your team on AI, we have the resources you need to get started. Begin your journey with Azure AI at *aka.ms/azureai*.

– Jessica Hawk, Corporate Vice President, Data and AI

# 2 How to use AI in your business

***This Chapter covers***

- What project leaders must know about AI
- Finding which business problems benefit from the use of AI
- Matching AI capabilities with the business problems you're solving
- Finding the gap between the skills the data science team has and the ones your AI project needs

You can spend years learning about AI, but because of the fast evolution of this field, even fully proficient data scientists need to spend a significant portion of their time on continuous and ongoing learning. The market of AI books and papers is dominated by technical information about AI. With all that wealth of knowledge, it's difficult to distinguish between what you need to know to manage AI and the knowledge necessary to have if you're an engineer building an AI system.

This Chapter talks about aspects of AI and ML that are necessary to understand to lead an AI project. It also teaches you how to find business problems that benefit from the application of AI. It provides examples of how to make AI insights actionable by linking AI capabilities with the business actions you already know you can take.

I've chosen the examples given in this and subsequent Chapters from different business domains. It's possible that some of the examples will come from a business domain unfamiliar to you. This is a good opportunity to practise one of the primary skills in successfully applying AI: adapting AI capabilities to business situations that you encounter for the first time.

## 2.1 What do you need to know about AI?

AI projects are very complex, combining business, computer science, mathematics, statistics and machine learning. This section explains why technical knowledge about AI isn't the primary knowledge needed for managing AI projects. If you're an AI project leader who doesn't have an analytical background, it's understandable if you feel that you need to grasp all of those concepts to be able to make the best decisions.'

The situation could be even worse: not only are the data scientists talking about concepts with which you're unfamiliar, but those concepts might look like something *you're supposed to know*, but can't fully recall. The jargon they use is often rooted in (or related to) statistical terminology. You might have taken a statistics class or two during your MBA programme, and you might not have paid particular attention to all the topics covered. Don't worry. The most important decisions for project success don't require or even necessarily benefit from an extensive knowledge of statistics or the details of AI algorithms.

What you do need to know to manage AI projects is the same as with any other project: how to define metrics and processes that allow you to properly comprehend and monitor the direction and success of the project. Once you understand that, managing AI projects is similar to running those projects you've overseen before.

### Managing an AI project is another application of management science

To use an analogy from a well-understood domain, if you were managing a factory, you wouldn't think that you'd need to become as good a worker as your foreman to run it. For that matter, it's a safe bet that quite a few executives who successfully manage factories aren't remotely handy.

The same principles apply for IT projects. Do you really need to know your database as well as your database administrator (DBA)? Do you feel you need to become a DBA to manage a database project?[a] You manage database projects by separating the

[a] In case someone wants to argue that AI isn't a factory, no, it isn't, but then neither is a database project. We've learned how to manage database projects without executives needing to become DBAs. Management as a profession is based on some universal principles for running organisations and projects, and that body of knowledge applies to AI too.

***(continued)***

business and architectural aspects of such projects from the skills needed to maintain RDBMS systems.

Just like a factory manager benefits from knowing how a factory works, having technical knowledge about AI doesn't hurt the project leader. However, the factory manager can't focus on the details of their foreman's job as a substitute for *knowing how to manage a factory and actively managing it*. Likewise, an AI project leader must focus on *management considerations*.

Still, there's often a feeling that managing AI projects requires significant focus on the details of how AI works internally, while comparable focus on details isn't necessary when managing factories or database projects. To the extent that this is true, it isn't so much because AI is different from other fields, but because AI is simply a *much younger field*.[b] In the case of factories, we've had enough time to develop management theory to understand that management knowledge isn't the same as domain-level knowledge about manufacturing. Having more time has allowed us to build methods and systems that allow us to compartmentalise skillsets: those needed to run a factory versus those needed to build a product. The goal of this book is to help you to do the same with AI.

[b] Yes, finding good data scientists is difficult, and, currently, they're rare. Today, some of them might object to being compared to a foreman. But do you think that a shift foreman for a railway was a common skillset when early railways were built? Or that the DBA skillset was common when databases were introduced? That's what I mean when I say AI is a young profession.

Most AI concepts that are relevant to making executive decisions could be explained to businesspeople in business terms. Ideally, your data scientists should be able to do that. If they can't, you should supplement your project team with people who have expertise in both AI and business to help with communications.

**NOTE** If you're feeling that you need to better comprehend analytics to make business decisions, what you have isn't a knowledge deficiency, but rather a communications problem.

What you need to know to manage an AI project is how to relate AI concepts to business. Namely, you need to be able to answer the following questions:

- What can AI do, and how can I use that in my business?
- What type of AI project should I start with first?
- How will I measure how successful AI is in helping my business?
- How should I manage an AI project?
- What resources are scarce, and how should I best assign them?

The rest of this book shows you how to organise a data science project in such a way that you can apply the management skills you already possess with minimal modifications to run AI projects.

## 2.2 How is AI used?

You make money when you perform an *appropriate business action*. That leads us to where AI plays into any system its using – AI *directs you in which action to take*. This section explains how AI does that.

While AI, ML and data science are perceived as new, the role that they play in making businesses successful isn't new. There are quite a few professions that have historically used some form of data analysis to make money. Some examples of those professions include actuaries and quantitative analysts. Experts in the application of statistical methods to process engineering and quality improvement science also have a long history of using data analytics to improve business results. AI doesn't change the way analytics and business relate, it just changes the methods for doing the analysis and the capabilities (and cost) of the analysis. There are significant parallels between how AI fits in with business today (and in the future) and traditional uses of data analysis in business.

To understand how to identify an opportunity to apply AI to a business problem, you first need to understand how a successful application of AI to a business problem has to look at a high level. In any problem in which data is used to inform further actions, there's a common pattern describing that process. We collect data, analyse it and then react to it. This is simply an age-old control loop, and it's important to understand the elements of this loop. AI just adds new capabilities in the analytical part of that loop. Figure 2.1 shows how these elements interact.

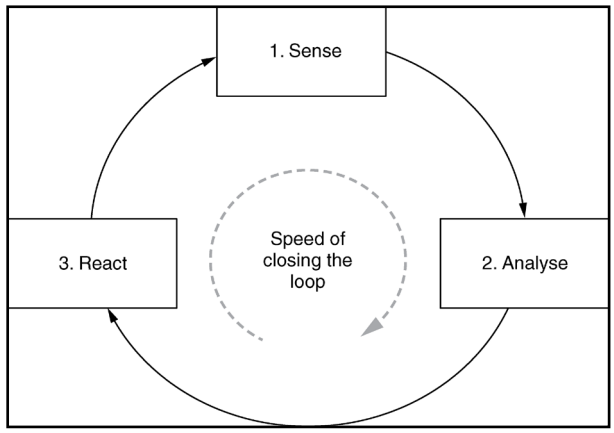


**Figure 2.1 The Sense/Analyse/React loop. Any successful analytical project must have all three elements of this loop.**

Elements of figure 2.1 are as follows:

- *Sense* – The sensor part of the loop is where you get the data that the analysis looks for. For the majority of enterprise systems in the pre-big data age, data was in disparate databases. For big data systems, it's common to store data in a data lake.

- *Analyse* – That's the box in which you now apply AI to your dataset. Before AI, we used simpler algorithms (for example, a PID controller [34]) or human intervention (for example, a manual loan approval in the context of a bank). Although the introduction of AI to help with analysis is perceived as a recent development, it isn't – AI research started in 1956 [35]. We've been using computerised systems to perform analysis for decades. What's new is that today, with modern AI techniques, computerised analysis is much more powerful!
- *React* – Reactor/effector is the part responsible for the action in the real world. That reaction might be performed by a human or by a machine. Examples of manual reaction include many decision-support scenarios in which a management decision is made based on the results of the analysis. Examples of automatic reaction include robotics systems, smart thermometers [36,37] and automated vehicles [38].

The speed of closing the loop is the time between the moment some event occurs and the time when a reaction is performed. How much the speed of closing the loop matters depends on the domain. In a high-frequency trading system, there may be a strong requirement for completing a loop with the utmost speed. In other situations (for example, if you're performing data analysis in the context of archaeological research), timing requirements may be much more relaxed. Sometimes the ability to guarantee that you'll meet time-critical deadlines matters too; an autonomous vehicle may require that your AI analytics never take longer than a specified time.

> **NOTE** The speed of closing the loop also depends on how often data is ingested into the system. Sometimes it's acceptable that data is periodically ingested into your system. In other cases, data must be analysed in real time as it's arriving into the system. (This is called *streaming analytics*.)

An important consideration in the application of a Sense/Analyse/React loop is the question of *who or what is reacting*. It could be the system itself in some automated fashion. (That's what a self-driving car [38] does.) Or, based on the results of the analysis, it could be a human. The latter case is much more common today within an enterprise use of data science.

### The Sense/Analyse/React loop is widely applicable

The Sense/Analyse/React loop is applicable across many scales. It could be applied on the level of a single device (as in the case of smart thermometers like Nest [36] and ecobee [37]), a business process, multiple departments, the whole enterprise, a smart city or an even larger geographical area. I believe that the Sense/Analyse/React pattern loop will, in the future, be applied to the level of whole societies, in systems such as disaster relief and the tracking and prevention of epidemics.

The Sense/Analyse/React pattern isn't limited to the domain of big data and data science. That pattern applies to the domain of development and organisational processes too. You might be aware of various forms of the control loops that

management sciences define and use. Examples of those loops are concepts like PDCA [39,40], OODA [41,42] and CRISP-DM [43], which have commonalities with and are further elaborations of this pattern. The Sense/Analyse/React pattern even applies to biology (for example, how octopuses and other animals behave [44]). In some domains, people might call the React part of the loop the Effector instead [45].

Automation of any business process is just an application of the Sense/Analyse/React loop. Using AI allows the application of that loop to some problem domain in which an automated reaction wasn't previously possible.

### Automated data analysis is a recent development?

Even uses of fully automated and rapid Sense/Analyse/React loops using complicated and computerised analysis are nothing new. Capital markets, especially combined with algorithmic trading, implement this pattern on a large scale. With the further advancement of the Internet of Things [46] and robotics, these large-scale, fully automated, closed control loops will become much more prevalent within the physical world.

## 2.3 What's new with AI?

The advance of AI broadened the applicability of the Sense/Analyse/React loop, because AI brought to the table new analytical capabilities. This section explains those new capabilities.

What's new with AI and big data is that automated analysis has become cheaper, faster, better and (using big data systems) capable of operating on much larger datasets. Analysis that used to require human involvement is now possible to do with computers in areas like image and speech recognition. Thanks to this new AI-powered capability, whole Sense/Analyse/React loops became viable in these contexts, when it wasn't economical to apply them before.

### Examples of AI making automation viable

The following are some examples where an introduction of AI makes it possible to automate tasks that previously required a human to perform them:

- *Automated translations from one language to another* – Language translation is nothing new and is something that humans have done since the beginning of time. What's new is that AI has reached a level at which automated translations are now viable and, as such, a translation web service becomes practical.[a]

[a] Note that the actual control loop in a real translation system typically requires that at least two control loops are present. One loop translates from one language to another, while another loop makes money for this service. That second loop may also work by collecting information about what translations you need, analysing them and performing some action that makes money for the provider of the translation service.

***(continued)***

- *Autonomous cars* – We've had some form of the motor car for the last 250ish years, always requiring a human operator.[b] What's new with AI is that we may be on the verge of being able to construct a car that doesn't need a human driver.
- *Ability to diagnose eye diseases* – We've all read letters at a distance for ophthalmologists and opticians and stared into the bright light on command. What's new is the ability of AI to detect diabetic retinopathy from simple retinal images [49].
- *Ability to read comments posted on the web* – If you read enough material in the comments section of a website, you can surmise whether people are enthusiastic or sceptical with regards to some topic. Now AI can do it too. AI can read a much larger number of comments faster and cheaper than a human ever could, and then tell you whether the audience is predominantly enthusiastic or sceptical. We call this capability *sentiment analysis*.[c]
- *Product recommendations* – Each one of us has friends that recommend books, films and products we might like. When AI does that (for example, on the Amazon website), it's called a *recommendation engine*.

Historically, when the size of datasets was small, humans were able to perform the same analysis that was done by AI. In some cases, what AI does is worse than what humans can do looking at the same dataset. But AI is more economical in the long run, and it can operate on datasets that are too big for humans to look at.

---

[b] The first self-propelled vehicle with wheels was invented in 1769 [47], with the first petrol-powered cars appearing in 1870 [48].

[c] At the time of this writing, AI isn't nearly as good a reader of web content as a human is, and it's struggling with cynicism and subtle messages in text – it often misses even the basic thesis of the message. However, for the purpose of answering the question, "Has sentiment about the product improved in the last three months?" AI is good enough and can provide an answer much more cheaply than you or I can.

What's *not* new or different with AI is that analysis still can't make money all by itself. Note that in none of the use cases given in the previous examples did I talk about how to make money. While some of those use cases are clearly straightforward to monetise (for example, an autonomous vehicle that drives better than humans), in others it may not be clear how to monetise AI.

### AI can't help you with a poor business case

Sometimes you won't be able to make a profit *regardless of how good your AI-powered analysis is*. Suppose that an ill-advised manufacturer of traffic signs decides to do sentiment analysis of the public opinion of those signs. The manufacturer would likely lose money on this analysis. It's not clear that drivers' feelings have a significant practical influence on the selection of traffic sign suppliers (or for that matter, that sentiment about the sign would be determined by the choice of supplier as opposed to where the sign is placed).

When you perform an analysis, you incur the cost of that analysis. Profit may happen when you react based on the results of analysis. If there's no business action you can take after getting the results of analysis, such an analysis is always a loss.

## 2.4 Making money with AI

If AI allows for improved analytics with the Sense/Analyse/React loop, how do you make money with AI? By finding a situation in which AI allows you to apply Sense/Analyse/React loops so that one of the business actions available to you could be automated using that loop. This section shows you how. Figure 2.2 presents the general process of making money with AI.

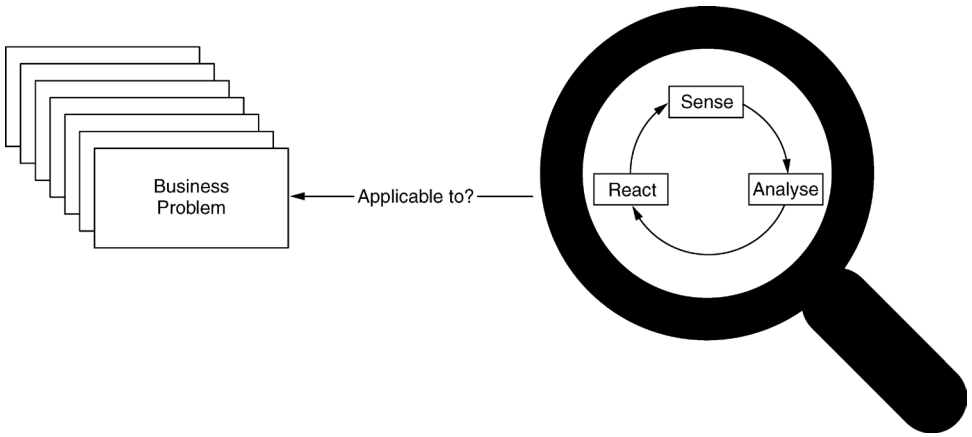


**Figure 2.2 Making money with AI is based on finding a business problem in which you can apply the Sense/Analyse/React loop to one of the actions you can take.**

You can apply this control loop in a new context because of the capabilities of AI. But to successfully apply the Sense/Analyse/React loop, you must make sure that all components of the loop are technically possible:

- On the Sense side, you must have the ability to collect the data that AI-supported analysis will need. Chapter 3 addresses how to ensure you've collected the appropriate data for your chosen AI method.
- On the Analysis side, you must make sure that you stay within the boundaries of what's possible with the available AI technology.
- On the React side, you must link the results of the analysis with one of the actions that you can actually implement in your business. You'll make a list of the possible business actions that you can take, and ask, "Is there an AI analysis that I can perform that will better inform this business action?"

Once you know that the Sense/Analyse/React loop is applicable to your business problem, you know that you have the ability to solve that business problem using AI. Let's start with an example.

### 2.4.1 AI applied to medical diagnosis

Suppose you're part of a software development team in a large hospital. Your team's goal is to apply AI to clinical and diagnostic procedures in the hospital. This section shows you how to find a use case in which AI can help.

To keep this example small and manageable, I'll concentrate on a single diagnostic workflow: a patient getting an eye exam. An image of a patient's retina is taken to check if there are any diseases. Assume that this procedure consists of the steps shown in figure 2.3.[1]

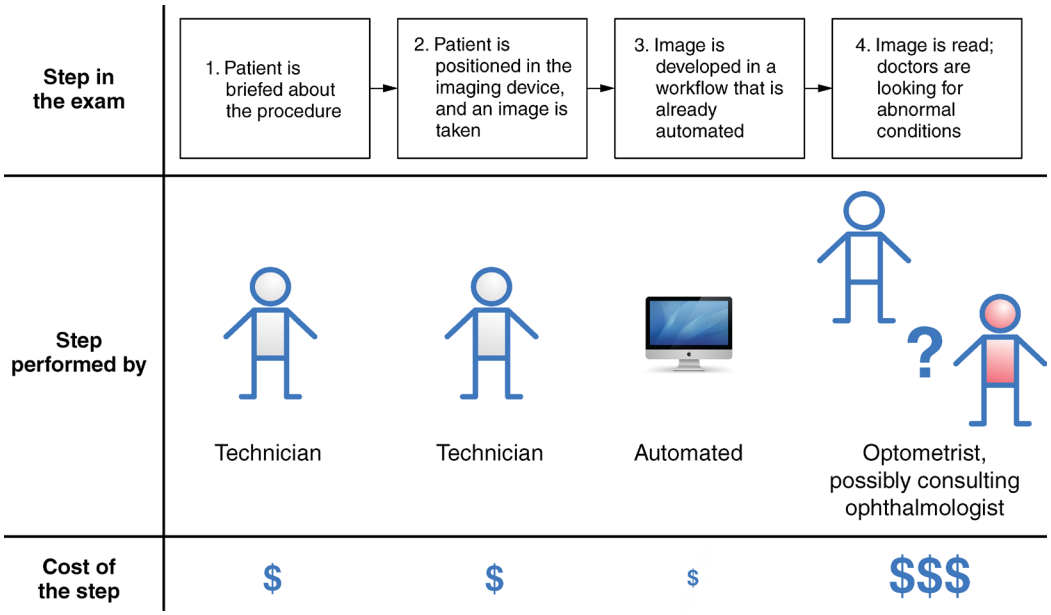


**Figure 2.3 A workflow of a routine optometry exam. We'll apply AI to automate part of this workflow.**

The workflow shown in figure 2.3 consists of the following steps:

1. Patient is briefed about the procedure. This step could be performed by a technician with minimal involvement of the optician.
2. Patient is positioned in the imaging device, and an image is taken. This step is also performed by the technician.

[1] An actual optometry/ophthalmology exam is more complicated and is simplified here for the sake of illustration.

3 Image is developed in a workflow that's already automated.
4 Image is read by the optician, looking for abnormal conditions. If necessary, additional doctors would be consulted.

Now you'd find a place where AI can help. In this workflow, you have three steps to which you could apply AI: two interactions with the patient and the final reading of the eye image.

### Never start by thinking which analysis to do!

To illustrate why you don't start by asking, "What analysis can I perform?" let's construct a scenario in which you start with an analysis based on knowing that AI can do something for you.

You might be aware of voice assistants like Apple Siri [50] and know that their voice recognition is getting better. What if you combine a voice assistant/voice recognition with the chatbot so the patient can be briefed by a machine? You're lucky to have a good data science team that's happy to work with this cool technology. This looks like a good application of AI, doesn't it? Let's build a quick prototype!

Unfortunately, any time you spend on such a prototype will be wasted. Replacing a technician has limited value – the cost of the time a technician spends on briefing the patient is relatively small. More importantly, you're serving a diverse patient population, including multiple languages, disabilities, ages and comfort with interacting with the machine. Human technicians are good at dealing with this population; today's AI isn't, if for no other reason than some segments of the population aren't used to talking with a voice assistant.

The idea you had was based on an interesting technology. The use case is inherently interesting and straight from sci-fi – many sci-fi stories feature patients talking with an AI doctor. The problem is that you've tried to apply AI to a situation in which a business action isn't profitable from the start, due to factors beyond your control.

This is a common trap. Everyone who works with AI in a business setting claims that they have a good business case, but often the business case was an afterthought, and the team's initial excitement about the project was caused by the opportunity to work with interesting AI technology.

In the worst case scenario, it might be impossible to monetise the project from the start, even if its technical portion was successful. Good AI implementation can't bail out a poor business case.

Let's use a systematic approach to better apply AI to this optometric exam. You start by enumerating the domain actions that you can perform and then see if you can apply the Sense/Analyse/React loop to those actions.

**TIP** Start by asking, "What are my viable domain actions?" As the number of domain actions that you can take is limited, you need to consider only a small number of use cases.

In this workflow, you have the two interactions of a technician with the patient, and you have the ophthalmologist/optician reading eye images to check for the presence of eye diseases. The interactions with the patient consist of an initial briefing and then positioning the patient in the imaging device so that a good image can be taken. You saw a moment ago why you can't automate the briefing. What about positioning the patient? That requires robotics expertise, and your executives are adamant that you're a software development company and not a robotics company. In your case, no action in the domain of direct patient interaction is viable.

What about interpreting the images? It turns out that interpreting images for certain eye diseases is complicated and that, in some cases, opticians may miss important conditions. Professional interpretation is also costly and something that your hospital would save money on if you could make an alternate system that's helpful when diagnosing eye diseases. This use case is worth further investigation.

Further research from your data science team shows that there has been significant progress in the application of computer vision to medical diagnosis. You find that Google's team created an AI capable of diagnosing cases of moderate to severe diabetic retinopathy [49]. You have enough data from past optometry exams that you can train AI on that data. To make sure the *Sense*/Analyse/React loop is applicable in this use case, you need to cover only the Sense part. That proves to be easy; you already have an image of the retina of the patient, and you can send that image to your AI system.

### 2.4.2 General principles for monetising AI

The previous example showed you how to find an opportunity for using AI in one business scenario. This section shows you what general principles you can extract from this example. Figure 2.4 shows those principles for applying AI.

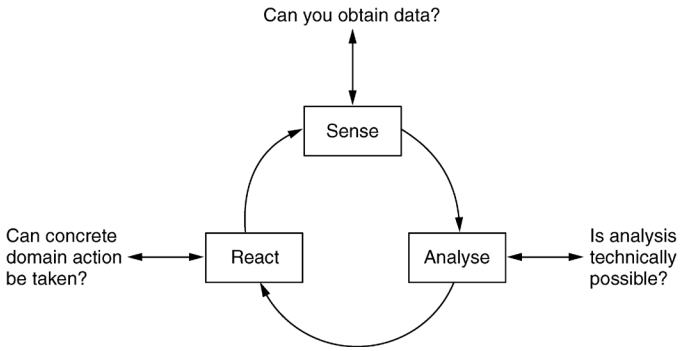


**Figure 2.4 General principles for applying AI to a business problem. The basic idea is to make sure that you'll be able to implement all parts of the Sense/Analyse/React loop.**

The approach shown in figure 2.4 covers each part of the Sense/Analyse/React loop:

- *Sense* – Can you collect the data you need? How much does it cost to collect that data?
- *Analyse* – Can AI do that analysis under ideal circumstances, or has anyone ever succeeded in doing something similar with AI? Is it well known that AI has such a capability? Does your team have expertise in applying those AI methods? How difficult is it to apply them?
- *Analyse* – Can a domain action that would be of value and possible for AI. What is the economic value of that action? This information lets you judge if automating that action with AI is economically viable.

Chapters 3 and 4 will discuss how to use business metrics to cover the economic aspects and the application of the Sense/Analyse/React loop. For the time being, lets concentrate on how to cover the React and Analyse parts of the loop. You need to answer the following two questions:

1. Is there a systematic way to think about your business that helps to find domain actions that can benefit from AI?
2. What are the high-level capabilities of AI?

Once you know the answer to these two questions, you can perform an analysis like the one shown in section 2.4.1 to find viable use cases for AI.

### Making money with AI isn't based on AI being smarter than humans

Examples in this Chapter (and Chapter 1) show why, to achieve success with AI, linking AI with business is much more important than specific algorithms and technology. AI isn't bringing superhuman intelligence to the table; it possesses human-like capabilities in limited domains, such as image recognition. It can also apply those capabilities economically and operate with larger datasets than any human can. But you still need to figure out how AI's capabilities translate into improving your business.

AI can sometimes find insights that escape human intelligence because of its ability to process large datasets. However, when operating in complicated domains, AI still lags behind humans. By itself, AI can't figure out how to make money.

Peter Drucker believed that it's more important to do the right thing than to do things right.[a] AI's job is to help you do better analyses, and it could help you do things right, but only you can ensure that AI is applied to the right problem.

[a] From the article 'Managing for Business Effectiveness' [4]: "It is fundamentally the confusion between effectiveness and efficiency that stands between doing the right things and doing things right. There is surely nothing quite so useless as doing with great efficiency what should not be done at all."

## 2.5 Finding domain actions

Now that you understand that the application of AI is simply a matter of applying the Sense/Analyse/React loop to some domain action, the next question is how you can systematically find domain actions that you can take. This section shows you how to find them.

There's a limited set of high-level roles that AI can play in your business. Figure 2.5 shows those roles.

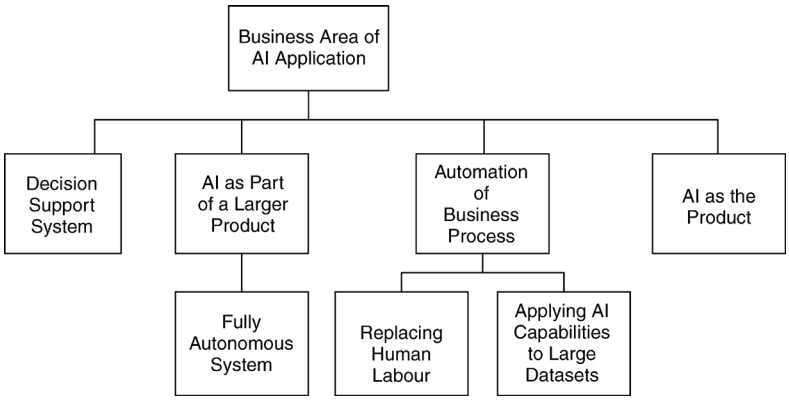


**Figure 2.5 AI taxonomy based on the high-level role it plays in business. You could use this taxonomy to guide you in eliciting available business actions you can help with AI.**

You can use AI as a part of the following:

- *Decision support system* – AI helps an employee or manager of your organisation to make better decisions. Uses of such systems range from helping management make decisions affecting the whole organisation to helping line employees in their day-to-day tasks.
- *Larger product* – AI could be just part of a larger product. Such a product has capabilities that AI may enable, but that aren't purely AI capabilities. An example here would be house cleaning robots (like Roomba [51]) or smart thermostats (like ecobee [37] and Nest [36]). In the case of a fully autonomous system, AI guides the system's operation and makes its decisions without needing human involvement.
- *Automation of the business process* – AI automates some steps in the business process. Sometimes this is done to replace human labour; other times, it's done to process datasets that are so large that humans can't possibly handle them.
- *AI as the product* – You can package AI tools as a product and sell them to other organisations. An example is an AI product capable of recognising images on traffic signs that will be sold to manufacturers of autonomous vehicles.

The rest of this section provides discussion of each of these bullet points.

### 2.5.1 AI as part of the decision support system

One of the most common scenarios for the use of data science in the enterprise today is the one in which AI is used as a decision support system. This section shows you how to use AI as a part of such a system to find domain actions.

AI as a part of the decision support system is the easiest scenario for elicitation of domain actions. In any decision support system, you're already focused on the options you need to decide on. When using AI as a part of the decision support system, you should consider the user (or management team) whose decisions you're supporting. Then you enumerate a spectrum of the decisions they can make. Finally, you ask yourself this question: "What information is needed to choose between these possible options?" The project is then organised around providing that information.

#### AI helps the management team

Suppose you're supporting a large manufacturing operation. The operation has multiple large suppliers that ship thousands of components to it every day. A big cost concern for your organisation is that if a certain percentage of the supplier's components are faulty, the organisation will spend a lot of time in the manufacturing process on troubleshooting problems. Such troubleshooting is costly. Even worse, the quality of the manufacturer's own end product could suffer.

Although individual suppliers are a big part of your business, this sector is dominated by a few large suppliers, and your ability to force suppliers to improve the quality of their product is limited. How can AI help the management team for your manufacturing operation?

Start by enumerating the options regarding what the management team can implement and which ones are viable. Because your organisation has little leverage on an individual supplier, the only viable business action your organisation can take is to change suppliers.

What questions do you need to answer to change the supplier? Here we concentrate on one possible answer:

> *Ideally, you want to be proactive and switch suppliers before their quality deteriorates – by then, our manufacturing operation has already incurred costs. It's difficult to decide to terminate the relationship with the supplier, because you don't know what the cost of staying with the supplier will be. Ideally, you want to act proactively, based on where the trend of the supplier's quality is heading. You don't want to cut a supplier whose trend in quality is improving. Nor do you want to wait to switch suppliers if the trend is diving.*

Based on management response, you now know that if you could use AI to analyse the historical trend of quality and to predict a trend in future quality, you'd have a system that would be useful to management. This is an example of using AI as a part of the decision support system.

This example also illustrates why customising AI's use to your own business case is better than applying AI solutions that worked for someone else. If you were a much

***(continued)***

larger customer to those suppliers, your management team might be able to negotiate the terms of the relationship, as opposed to just switching suppliers. Examples of adjusting such relationships might include escalating problems to the supplier's management or asking for monetary compensation for defective parts.

While those might be viable actions for customers of your suppliers that are much bigger than your organisation is, they aren't viable actions for your organisation. Generic AI solutions tailored to much bigger organisations might focus on actions you can't take.

One final question in this scenario: Supposing that you're a large organisation with many departments, on which level of granularity should you request that business actions be supported by the decision support system? You should consider the options that are directly within the scope of responsibility and execution of the team that's performing the analysis.

**WARNING** It's critical that you choose the right level of organisation to look for possible actions you can take. If you finish with a list that enumerates 20 or more different options that you believe you can take, the granularity level on which you performed analysis was wrong.

When discussing the use of AI as a part of a decision support system, the danger lies in diving in too deep. If you're applying AI as part of a decision support system for the team of senior managers, then you should analyse the actions that senior managers take, not the actions that each individual worker working in their organisation can take. Don't analyse the actions an intern can take on their first day.

**TIP** The decision maker doesn't have to be a high-level manager. Imagine an AI that recommends to your sales force which customers to approach and displays a dashboard with the further information about each customer. However, that AI leaves the final choice to the individual sales professional. Such an AI is a decision support system.

### 2.5.2 AI as a part of a larger product

Another common situation occurs when AI capabilities are part of a larger product. In this situation, a key characteristic is that the end customer isn't buying the AI itself; they pay for some capability that the larger product wouldn't have without using AI. This section shows you how to use AI in the context of a larger product.

AI as a part of the product itself is already extremely important. Examples include products that range from smart speakers (Amazon Alexa [52], Google Home [53,54] and Apple HomePod [55]) to autonomous vehicles [38]. Although you can think of AI as a way to differentiate products, you're generally better off thinking about it as an *enabler* of your value proposition to the customer.

**TIP** Few people would buy a product specifically because it uses AI. The key question is, "What value is the product providing for your customers?"

There might have been a time when saying "We use AI" was a viable marketing/fundraising technique, but that time is over. Over time, AI will play the same role in autonomous products as an engine plays in a car today: you can't go anywhere without it. However, most car buyers don't care about a particular engine, but rather its ability to move the car from point A to point B.

### AI as a part of the product

An example of an AI product that also loops in humans is a home security company that uses an AI-powered device as a part of the security system. What are the relevant actions that such a system can take? For one, it can sound an alarm if it believes there's an intruder.

For various cost and liability reasons, management will probably require that the final action of sounding the alarm or calling the police must always be initiated by a trained, live operator in a monitoring centre. Management could also decide how many operators will be assigned to monitor the properties. This business would be much more profitable if a single person could monitor multiple secured properties.

AI could be leveraged in such a system to help the operator seated in the monitoring centre. If AI can recognise faces, it can also sound an alert when humans are in the house who are not part of the family that lives in the home. The AI can then notify the operator so that a check can be made and, if necessary, the operator can raise an alarm.

When AI functions as part of a larger product, that product operates somewhere within the physical world. Because the customer is paying for some capability the product has, not for the fact that the product is using AI, start with how the product functions. Which potential actions could the system carry out? Once you know the set of possible actions, the next question is, "When should the system take each one of those actions?"

**NOTE** When AI is part of a larger product, the product itself could be fully autonomous, or it could be a hybrid product that performs some functions automatically, while depending on humans for other tasks.

### AI in the fully autonomous product

An example of a fully autonomous system would be a vacuum cleaning robot like the Roomba [51]. In this case, the vacuum needs to clean the whole room. The relevant domain action is, "Where should I go, and what areas should I avoid?"

AI can be used to provide navigation capabilities for the device in its environment. Note that such an AI could range from a sophisticated navigation system to relatively

***(continued)***

simple operations. A robotic vacuum can use AI to learn the layout of your rooms and recognise changes in that layout. You can also trade sophisticated mappings of the room for a bigger battery, allowing obstacle avoidance using a time-intensive trial-and-error approach.

That bigger battery is another example of the whole system being more important than the choice of AI algorithms. A few years back, it was simpler (and cheaper) to add a larger battery to increase run time than to spend a lot of time and money on significantly improved AI navigation.

In the context of a fully autonomous product, you also need to consider not only what actions the product can take, but that some actions and outcomes are neither desirable nor permissible. You don't want to watch an expensive robotic vacuum such as the Roomba crashing down the stairs.

**How would capabilities of your product evolve?**

It's important when using AI as a part of a larger product to consider not only the capabilities you're planning to add in the initial product, but also the whole roadmap of product capabilities that you plan to add later.

Often, your product is a physical system shipped to the customer. For example, in the case of the AI-powered autonomous vehicle [38], you'd ship the vehicle itself. Once the vehicle is delivered, an additional capability could be added to it as a software upgrade. But you're stuck with the sensors and effectors (engine, brakes, steering mechanisms, horn, signal lights, headlights and so forth) that are shipped with the car. Once you distribute physical systems to your customers/users, it's often impossible (or expensive) to add the capacity to perform new actions that you didn't envision at design time. Whatever autonomous cars we have in the future, it's a safe bet that some of their capabilities will be fixed at the time the car is manufactured and will be difficult to change later.

### 2.5.3 Using AI to automate part of the business process

One of the uses of AI that's getting increasing attention in both industry and the popular press is the use of AI to perform actions that previously required humans. This section shows you how to apply AI to optimise existing business processes.

**AI automating part of the workflow**

Suppose you have a facility that's using CCTV cameras and security guards to monitor it. Looking at the screens is part of the workflow of the security guards. AI could be used to make this part of the security guard's workflow more efficient by monitoring the video streams and highlighting unusual situations.

When you're looking at using AI to automate part of the business process, start by sketching out that process and then ask, "Can any of these steps be made more efficient or eliminated using AI?" This is using AI to perform a one-to-one task replacement: the task that used to be done by humans is now done by AI.

As the capabilities of AI and humans differ, a one-to-one replacement of tasks performed by people with performing them using AI is complicated and expensive. In most workflows, a few tasks are essential, and they prove to be difficult to automate, even if the most time-consuming function of the job can be automated!

In practice, it's usually necessary not only to apply AI to steps in an existing process, but also to re-engineer business processes. Re-engineering should separate out operations that are easy to automate with current technology into a separate step of the workflow. Then you assign AI to only those parts of the process that are easy for AI, but time-consuming or error-laden for people.

**Creating new jobs with AI**

The use of AI for automation is a controversial topic. If AI replaces a human in performing some action, and that action is the primary purpose of that person's job, that job can now be in jeopardy.

There are significant costs that should be considered when eliminating jobs. Foremost are the costs to the people whose jobs disappear. There are also costs to your company, not just monetary, but also in the goodwill of both the public and your remaining employees. It's important to keep that human perspective in mind when you talk about automation of your processes, and to understand that this scenario is often a zero-sum game.

If you're limiting yourself when thinking about AI only to scenarios in which you're replacing jobs with AI, then you're actually missing an opportunity. AI can allow you to create new businesses that weren't possible or economical before. This scenario generates new jobs – not only jobs building and supporting that AI system, but also all other jobs that come with such a business.

Take, for example, using AI to monitor the behaviour of pets when owners are at work. At the moment, no one is doing this, because such monitoring isn't economically viable as a service if it has to be done by humans. An AI that's capable of monitoring the behaviour of pets and entertaining them requires people in the loop to handle some rare situations that AI can't (for example, situations in which the pet appears to have a medical issue). Such an AI creates jobs for the people monitoring those pets. These jobs weren't economically viable at all when 100% of the work had to be done by humans. Such jobs become viable once an AI handles most of the monitoring and humans handle the exceptions.

### 2.5.4 AI as the product

Sometimes you have an AI solution, or an infrastructure solution supporting AI, that you believe to be applicable to many business contexts and many different customers. When that happens, such an AI solution is valuable in and of itself and could be

packaged and sold as a standalone product. This section talks about some special considerations that apply when you're intending to sell an AI solution as a complete product.

You have a complete product when you have customers who are willing to pay for the AI capability that you can develop. There's a long history of companies offering various analytical products (such as SAS [56] or IBM's SPSS [57]), and AI-based products could be considered a continuation of this tradition, wherein complicated analytical capability is packaged in a format that customers can use.

> **TIP** You're selling a product. The key question is whether you can find customers that are willing to purchase this product. With regards to the sales cycle, the fact that the product itself is based on AI is secondary to all other sales considerations.

But there's a specific consideration that you must address when you base your product on AI. You must correctly assess the capabilities of your organisation and your team regarding their knowledge of AI. There's a vast difference between developing an unprecedented AI solution and applying known AI capabilities in a new and specific context.

Developing new AI solutions and capabilities is a different ballgame that requires significant prior expertise in the field. When you're selling AI as a product, you must assess not only the ability to deliver an initial version of the product, but also your ability to out-innovate the competition.

> **WARNING** Unless you have a team of experts in AI research working for you, stick to applying an existing AI capability to a new context. Avoid AI products that require you to develop new AI capabilities that no one else has demonstrated yet, because they're unpredictable, difficult and risky to develop.

On the other hand, if you understand a general AI capability, then there is much less risk in applying that capability to a product in some new field. For example, it's known that AI is getting very good at recognising the context of an image – that's a general capability. If you can apply that capability to a specific area, you might have a viable product. One example would be software that's able to recognise defects on a factory line. This could be invaluable, provided you know to whom you would sell it.

### Is my AI product widely applicable?

Some AI products are general frameworks that are (clearly) widely applicable, but others are specific to one category of problems.

If your AI solves one category of problems, it can stand alone as a product if you can find multiple examples where the use of your AI solution makes new business actions viable. If, instead, you find only a single example of AI producing a new business action, then you're better off thinking about what you have as an example of AI being part of a larger product.

When trying to figure out what new business actions AI makes viable, you'll be applying the techniques in sections 2.5.1, 2.5.2 and 2.5.3. However, instead of applying them to your own business, you'll apply them to your potential customer's business.

## 2.6 Overview of AI capabilities

Section 2.5 showed you how to find a business question on which you can act if you can pair it with the appropriate AI capabilities. This section presents the taxonomy of AI capabilities that helps you answer the question, "Is there a broad area of AI capabilities that could address my business problem?" Figure 2.6 presents such a taxonomy of AI methods.

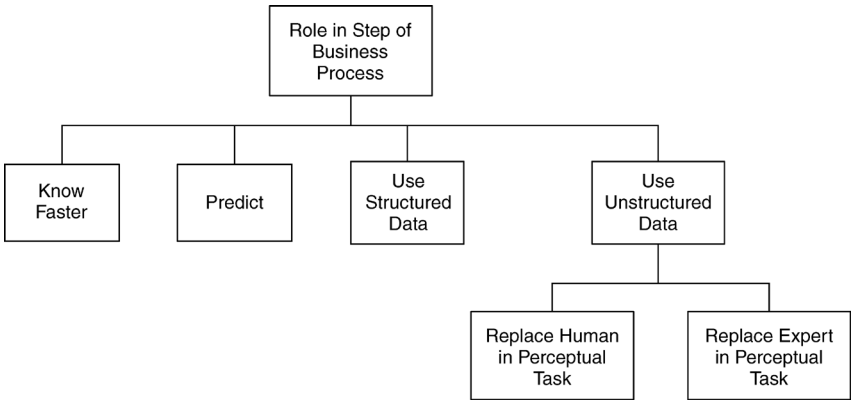


**Figure 2.6 Taxonomy based on AI capabilities. This framework groups broad areas of AI capabilities so that you can quickly check if any of them are applicable to the business problem you're addressing.**

This taxonomy is a modification of the taxonomy originally presented in Bill Schmarzo's books [58,59] with the 'Use of unstructured data' category expanded to highlight the use of AI in perceptual tasks. The main goal of this taxonomy is to guide a discussion between the AI expert and the business expert. Categories in this taxonomy follow:

- *Know results faster.* Here, AI helps you discover a result more quickly, and that has a business value in many scenarios [58,59]. Suppose that you run a car manufacturing plant and you're assembling cars from parts that are made in one area of your factory. If you know that some car part is defective as soon as it's made, you can discard it at once and never install it in the car. This is much better than learning that the part was defective after you've already installed it in the car and shipped that car to the customer.

- *Predict some event that occurs in the future, based on current trends.* You saw this technique used in section 2.5.1 when predicting the future quality of the supplier based on historical trends.
- *Use structured data.* Sometimes you can find the answer you're looking for in one of the relational databases that you already have, especially if you have a large volume of data [58,59]. There are also AI methods that work well with data that's already in tabular format.[2]
- *Use unstructured data.* AI methods can also help you process and comprehend a large quantity of unstructured data, such as text, images, video and audio [58,59]. In this case, you can use AI methods to recognise the context of the image, video or audio recording.
- *Replace humans in perceptual tasks.* This subcategory of unstructured data use is based on the fact that, in recent years, AI has matched and even exceeded human abilities on many simple recognition tasks, such as image recognition [62,63]. You can think about this category of AI as having the ability to perform simple perceptual tasks that humans easily and instinctively perform. An example of such a task is recognising objects in a photographic image.
- *Replace experts in perceptual tasks.* This subcategory of AI capability also comprehends unstructured data, but here AI performs perceptual tasks that otherwise require a high-level human expert. Such an expert uses skills that, after years of training, have become instinctual. An example of this would be using AI to interpret medical imaging. In recent years, AI has demonstrated an ability to interpret medical images on a level that in some cases rivals human experts [64,65].

Now you see how we've found AI solutions applicable to the business problems presented in section 2.5. In all those examples, you start by finding an actionable business problem and then the domain actions that could be taken. You ask the question, "Can we apply any of the six categories of AI capabilities shown in figure 2.6 to this business problem?"

### Can you enumerate all the individual AI methods out there?

There's no way to describe all the capabilities that AI has in any single book, including this one. AI is a rapidly developing area, and AI capabilities transform daily with the development of novel methods and applications. If you're interested in the details of individual AI methods, you need an experienced data scientist or consultant to guide you through the details of the latest capabilities of AI.

The taxonomy presented in this section isn't a substitute for AI expertise, but it's a systematic way to frame a discussion between a business expert and an AI expert.

---

[2] An example of such a method is *gradient boosting*. If you're interested in the technical details of this method, see the discussions on Wikipedia [60] and the Kaggle website [61].

It provides common terminology and concepts in a way that's easy to comprehend for the business users. If you're an expert in AI, you can use the taxonomy presented in this Chapter as a quick checklist for a class of methods and algorithms that should be checked for applicability to business questions.

## 2.7 Introducing unicorns

This Chapter has shown you how to determine which business problems can benefit from AI techniques, but does your particular development team have the knowledge necessary to implement the solution you just proposed? This section helps you answer that question.

The skills we're using on AI projects are still new (and rare), and there's still some amount of confusion in the industry about the skillsets that data scientists and data engineers should possess. Because of the rarity of those skills, there's a joke that such experts are *unicorns*. In this section, I'll start by describing the skills that are often *attributed* to unicorns. Then I'll explain why most real-world teams will never have all those skills. Finally, I'll show you how to make sure your team possesses all the skills that the specific AI project you're running requires.

### 2.7.1 Data science unicorns

Data science could be considered an umbrella term that covers many skills. A survey performed in 2013 lists 22 different areas that are part of data science [66]. Examples of those areas include topics like statistics, operational research, Bayesian statistics, programming and many others. It gets worse! Today, there are new areas that would certainly be considered important (for example, deep learning).

> **NOTE** Clearly, a data science unicorn should be a world-class expert in each one of those areas, right? No, these are individually very complex areas. Many distinguished professors in leading universities spend the totality of their time and effort to be an expert in just one of those areas. Most likely, no one in the world has expertise (defined as being on a level comparable to the skills of the aforementioned professors) in all those disciplines. Even if such a unicorn exists, which AI projects would have the budget for one?

Why are so many different skillsets part of data science? Because different practical problems benefit from different skills. No single ML method beats out all other methods across all possible datasets.[3] Each of these methods emerged because when the AI community tackled real, practical problems, some of them worked better than others. After many years, we use a combination of many methods from different disciplines.

[3] This is also known as the *No Free Lunch Theorem* [67].

**How to grow unicorns**

Did the leading data scientists start by learning all the methods that they know today? To be an accomplished data scientist, must you first build a skillset that emulates the skillset of a famous data scientist? No. Often, the skills of two accomplished data scientists don't match. Even among accomplished data scientists, it's virtually guaranteed that one will have expertise in at least one area with which the other is unfamiliar.

The skillset of accomplished data scientists is often acquired by working through problems that benefited from specific types of AI methods. They had to learn those particular AI methods *because they were necessary to solve a concrete problem in the domain in which they worked.* Each new project brings them new skills, sometimes in new areas that weren't previously part of their core domain of expertise. For example, in 2011, few people in the world, in business or academia, were working on what today is known as deep learning.

If you want to become a unicorn, work on problems worth solving. You'll acquire a strong skillset along the way.

As a manager, you should look for two things when hiring data scientists for your team. You should look for a candidate who has skills in the core domain that your initial AI project is likely to use, but you also need them to have a demonstrated ability to learn new skills. Chances are good that, along the way, your data scientist will need to learn many new methods. When hiring senior data science team members, don't just look for a strong background in one set of AI methods. Senior data scientists should have a history of solving concrete problems using a diverse set of methods.

**TIP** Data science is a team sport. To completely cover all of the knowledge that's part of data science, you need a whole team, so you must assemble a team with complementary skillsets.

How should you assemble your initial data science team? Your team needs both enough business expertise to understand your business problem and enough proficiency in AI methods to perform an initial analysis and determine if AI can address your problem. Keep in mind that on the way to delivering a full AI solution, the team will have to learn some new skills.

### 2.7.2 What about data engineers?

When we discuss AI, we often talk about operating with some datasets so large that they don't fit on a single machine and require a big data framework to manage. While data scientists are proficient with the use of big data frameworks, they're rarely experts in the details of those frameworks. As a result, you'll need specialists who are primarily focused on the use of the big data frameworks themselves. We call them *data engineers*.

Just like data science, big data is a large area. Let's take the example of just one popular product in the big data space – the Apache Hadoop framework [15]. A few

years back, the distribution of one of the leading Hadoop suppliers consisted of 23 separate components, each one of which was large enough that a separate book could be (and often has been) written about it [68].

The body of knowledge that falls under the umbrella of data engineering is much larger than any single big data framework. Data engineers often need to be able to operate in both an on-premise and cloud environment. Cloud services like Amazon AWS [11], Microsoft Azure [13] and Google Cloud Platform [12] have different platforms with significant differences between them. That means that in addition to the specialist skills in big data frameworks, data engineers you hire may also need to have a skillset in the cloud platform of your choice.

Clearly, the same limits that apply to data scientists also apply to the data engineers: they're also humans and can't know everything. Data engineers are characteristically experts in a few of the components of the leading big data frameworks.

### 2.7.3 So where are the unicorns?

I hate to break this to you, but it's highly unlikely that you'll find any single human that has a strong expertise in each one of the methods, products and technologies that are part of data science and data engineering. At best, you could hope to find a couple of senior people who have strong experience in individual data science and data engineering topics and who have enough familiarity with other related subjects to talk with specialists in areas in which they themselves aren't experts.

Although universities have started offering programs and degrees in data science and data engineering topics in recent years, it's not likely that this problem is solvable through better education. The field of knowledge is simply too large, so you should have a realistic expectation of what these institutions can teach their students.

> **WARNING** As a project leader, you must differentiate between the skillsets that your team possesses and skillsets that you need for your project. You should identify and close skill gaps. Don't assume that your senior data scientists and architects know everything in their fields, and don't impose expectations on them that they should. Such expectations only make people less likely to acknowledge skill gaps.

Project leaders must know where the knowledge gaps are in the team. Piloting an AI project that requires skills your team doesn't already possess means that you need to close these knowledge gaps. You do that by applying gap analysis [69]. An example of a gap analysis between the skillset a team presently has and the skillsets that are needed is shown in figure 2.7.

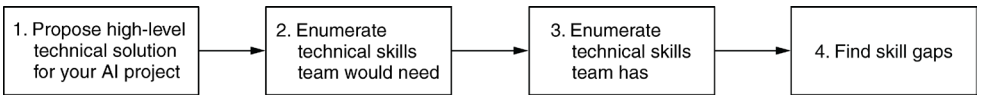


**Figure 2.7 Gap analysis between skills the team has and the ones needed. This analysis allows you to create a plan for how to address missing skills.**

You perform gap analysis by applying the following steps:

1. You first work with your technical team to sketch a high-level technical solution. Match the time spent on this technical solution with the likelihood that you'll implement it. If you're just considering the project, then the solution should be a high-level one. If you're planning to initiate the project soon, then your initial technical solution needs significant detail.
2. Based on the solution, take an inventory of the technical skills you expect the project will need to address your identified use cases. This summary of skills should be made by people who both are familiar with the business problem you're trying to solve and have enough technical expertise to quickly identify a high-level technical approach to solving it.
3. Which skills do your team members already have? Ask your team about their skillsets and avoid assumptions in this area – AI and data engineering are highly technical fields, and it's easy to have unfounded assumptions.
4. Find any gaps between the needed and current skillsets. These gaps are useful in estimating the project's difficulty level for your team. Keep this list. If you decide to proceed with a project that addresses this business question, you'll need to make a plan for how to close the gap. (You close knowledge gaps by training your team, hiring new team members or hiring consultants.)

Understand that gap analysis is always performed based on the current situation. If you're just thinking about an AI project as a possibility, you should perform this gap analysis on a coarse level with just an outline of a technical solution. For projects that are in progress, you need a much more detailed solution as a starting point. That means that throughout the project life cycle, you'll typically perform gap analysis multiple times.

### Beware how you ask

When you ask about gaps in technical skills, you're asking your technical staff to admit to areas in which they personally don't have expertise. If poorly handled, they might rightly consider it a landmine. Put some thought into this before you ask; it's your job as a leader to make sure that you create an atmosphere in which it's easy for team members to admit that they don't possess some technical skills.

One preferred technique to create such an atmosphere is based on building trust among team members so that they can talk about issues like this. Other techniques you might find useful are asking for the skillset in private, creating an anonymous survey or asking a trusted intermediary to approach the subject with your team members.

## 2.8 Exercises

The goal of this book is to help you develop practical skills you can use when running your project. To help you with that, the exercises in this section ask you to apply skills learned in this Chapter to new business scenarios.

### 2.8.1 Short answer questions

Please provide brief answers to the following questions:

**Question 1:** Think about a failed project in your enterprise. Would that project have failed in the exact same way if it also had a component based on AI?

**Question 2:** Do you personally have enough knowledge of data science and data engineering to understand the gap between the technical skills that your team has and the skills that they need for this project?

**Question 3:** Do you have a good enough relationship with your team members that they're comfortable admitting the limitations of their skillset to you?

### 2.8.2 Scenario-based questions

Answer the following questions based on the scenarios described:

**Question 1:** One of the important skills in applying a Sense/Analyse/React loop is to identify who will execute on the React part of the pattern. For the following scenarios, answer this question: Who or what will carry out the action and fulfil the React part of the Sense/Analyse/React loop?

- **Scenario 1:** You're making an automated car, and the AI that you're using will allow fully autonomous driving under all conditions (so-called Level 5 autonomy [38], in which there are no available controls for the driver).
- **Scenario 2:** You're writing a recommendation engine in which products are suggested to the customer.
- **Scenario 3:** You're writing an AI program to regulate a smart thermostat that controls the temperature in your home.

**Question 2:** Use AI to create a new job. Find an example of an AI capability that would let you offer a new service that your organisation doesn't yet provide. (For the job to count as a solution to this exercise, it must be a job that's so unrelated to the software development team that's building the AI, that the person hired for the job is unlikely to ever meet that team.)

**Question 3:** Suppose you're using an AI algorithm in the context of a medical facility – let's say a radiology department of a large hospital. You're lucky to have on the team the best AI expert in the field of image classification, who has you covered on the AI side. While you're confident that expert will be able to develop an AI algorithm to classify medical images as either normal or abnormal, that expert has never worked in a healthcare setting before. What other considerations do you need to address to develop a working AI product applicable to healthcare?

**Question 4:** Apply the previous example from a hospital setting to a classification problem in your industry. What are the new considerations that exist in your industry as compared to the healthcare industry?

**Question 5:** Provide an example of an AI that has replaced a human role, but doesn't provide as good of an experience as a human would.

**Question 6:** You're a manufacturer of security cameras, and you've developed an AI algorithm that can detect a person in a picture. Regarding the taxonomy of its role in your business, how would you classify this use of AI?

**Question 7:** You're an insurance company, and you've developed an AI program that, based on static images from an accident site, could recognise which parts of the car are damaged in a wreck. Can this replace an insurance adjuster?

## Summary

- Managing AI projects doesn't require expertise in the details of AI algorithms. Instead, you need to know how to explain the benefits of an AI project in business terms. What business problem is being solved? What business benefit does AI provide? How is that benefit measured?
- You can discover business actions you can take and those that may benefit from AI using a systematic process. Apply the taxonomy described in figure 2.5 to your organisation.
- AI capabilities are based on being able to know sooner, predict, process structured and unstructured data and perform perceptual tasks
- AI can help your business by performing analysis that informs some concrete business action. AI opportunities arise when you can apply a Sense/Analyse/React loop, with the Analyse part based on AI capability and the React part based on concrete business actions you can take.
- No individual is an expert on all topics of AI, data science and data engineering Project leaders must identify and close any relevant gaps in the knowledge and capabilities a team has.

# 3 Choosing your first AI project

***This Chapter covers***

- Selecting AI projects that are matched to your organisation's AI capabilities
- Prioritising your AI projects and choosing which AI project to run first
- Formulating a research question that's related to a business problem
- Pitfalls to avoid when selecting AI projects, and best practices of such projects

To develop a sustainable analytical organisation, you shouldn't start with an AI project that involves complex technical challenges. Instead, you should choose your initial project so it provides clear and actionable results quickly. Your whole process should be organised to optimise *time to success*.

This Chapter shows you how to select your first AI project. It also teaches you how to check if the research question that your AI project uses correctly reflects the business concerns it's supposed to address. Finally, it presents a list of common pitfalls that young AI teams might fall into.

## 3.1 Choosing the right projects for a young AI team

I assume that your long-range goal is to build an AI team that will help the success of your parent organisation by delivering a series of successful AI-related projects. To achieve that, you need to understand the journey that a successful AI team will take. This section explains that journey.

> **TIP** If you're after a one-off AI project, you might be better off buying an off the shelf solution or contracting with an outside partner to do it for you.

One of the most crucial decisions that you as a leader need to make is how you want to prioritise the order of the initial AI projects that your team must undertake. Before you're ready to make that decision, you need to understand its impact. To understand how AI teams succeed or fail, you first need to understand what success and failure look like.

### 3.1.1 The look of success

Leo Tolstoy wrote in *Anna Karenina* [70]:

> *All happy families are alike; each unhappy family is unhappy in its own way.*

Likewise, all successful AI teams are alike: your AI team is growing in expertise (and possibly headcount) and is solving more and more complicated problems. Unsuccessful AI projects result from an assortment of errors (many of which are described in section 3.4), and they may take your whole AI team down with them. This section explains why you should start with projects that are quick to deliver, but still provide significant value for your business.

If you're initiating AI efforts in your organisation (or even if you're part of an established analytical organisation), you're subject to three forces:

1. *Plentiful opportunities* – You're operating with technologies (AI and big data) that weren't present in business and industry historically, and you're the first to apply them to the many datasets your organisation has.
2. *Limited time and resources* – You have limited resources to devote to analysis. Chances are, you don't have enough qualified people to run AI projects you're thinking about.
3. *Success makes you stronger* – If you make money for your business, your analytical resources will increase over time. Management invests in teams with a good track record of providing value. Additional data scientists will want to join projects with a history of success. Solve some easy problems first, and you'll get the resources you need to address larger problems.

How do you succeed in an environment like this? Does it make sense to first concentrate on large wins regardless of how difficult they are (for example, projects that provide significant monetary value)? Clearly not. What's difficult today will be

easier tomorrow, so start with a project that has significant monetary value and can be delivered quickly.

> **TIP** The key is a fast turnaround on initial projects so that you can learn quickly. You're the first one in your company that has applied AI to your data and business problems, and there are plenty of opportunities to succeed with AI. Frankly, if you can't find an easy win in such a setting, then AI simply can't help your business.

Let's use another analogy here. The position your team is in (with respect to the opportunities) would be similar to a hunter that finds a rich hunting ground. If opportunities to make money were animals, and you were a prehistoric hunter, you'd be operating in a target-rich environment (see figure 3.1).

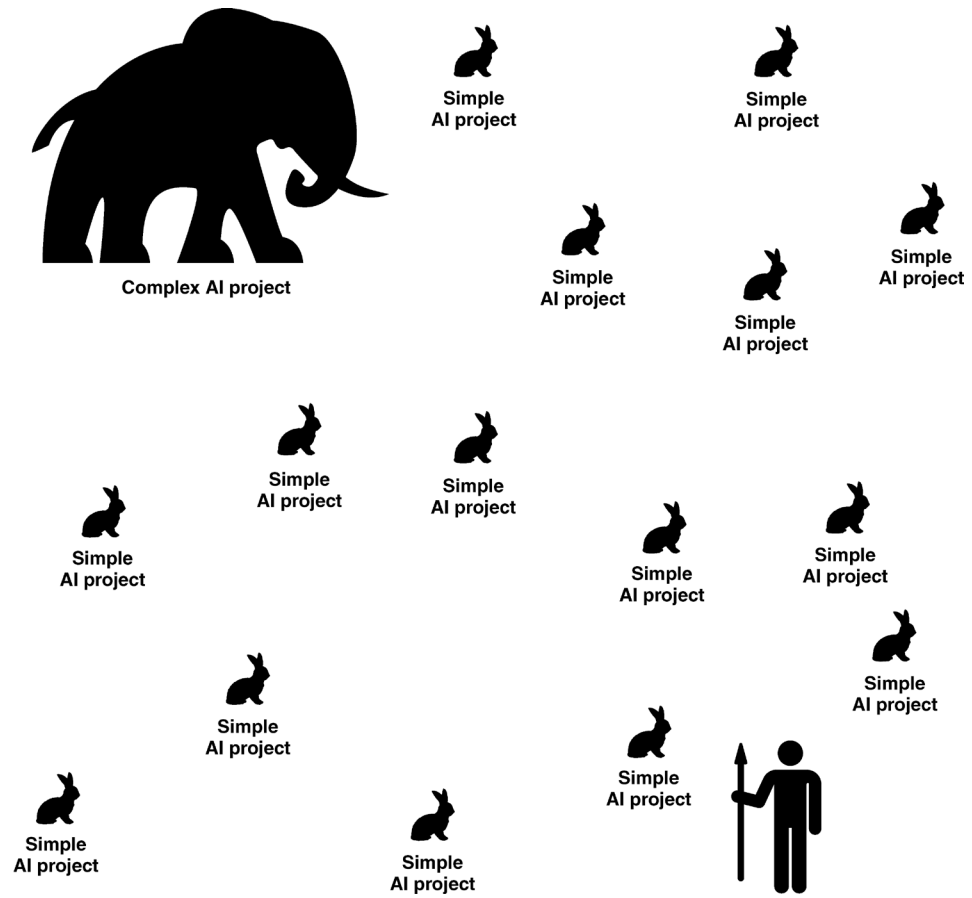


**Figure 3.1 You're in a rich hunting ground – plenty of rabbits and a big mammoth are in sight. Which animal should you try to catch first?**

Building a successful analytics organisation is similar to surviving and prospering as a hunter. Now, ask yourself, "If I were in that position, would I aim for the biggest animal in the field first?"

The answer is probably no, if for no other reason than that you're at the end of a long and distinguished line of ancestors that have succeeded in passing their genes to the next generation. They had the common sense and the skills to survive. You don't initially have the hunting skills that bring you success if you go after the mammoth first! If you're like me, the author, chances are that by the time that mammoth turned around, give up the attempt, run far away and convert to vegetarianism for the rest of your life. But I'd like to believe that even I can be successful snaring a rabbit. All successful hunters should be able to catch rabbits (see figure 3.2), right?

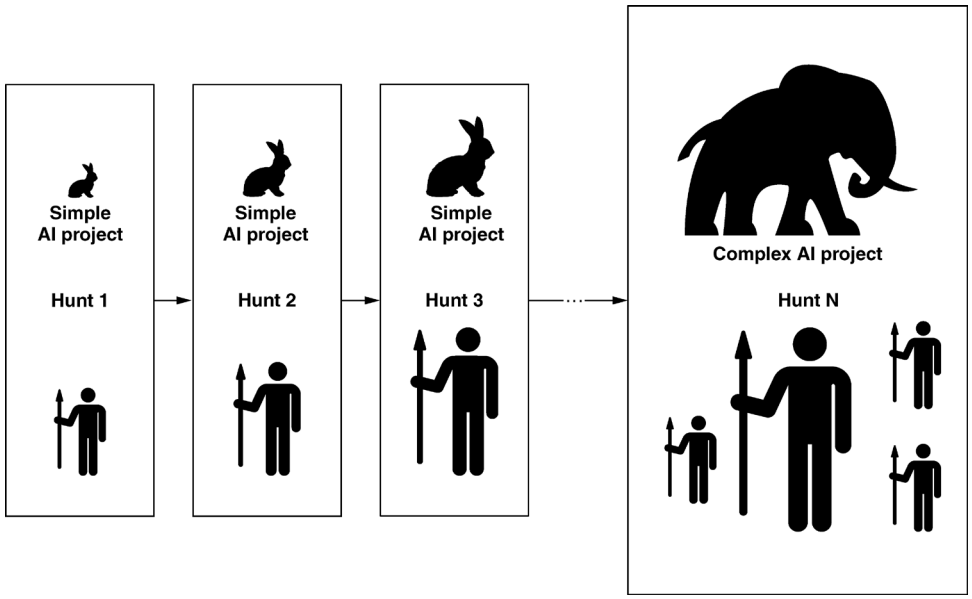


**Figure 3.2 Start with easy projects. Success with those projects enhances your skills and reputation within the rest of the company, allowing you to attempt more difficult hunts later.**

You don't want to start with a technically difficult project that requires a long time to deliver, even if that project is perceived to have a high business value. If your team is the hunter, the easier projects are the rabbits of the project world.

**TIP** Once you're known as a good hunter, the rest of the tribe is going to be more willing to help you with hunting mammoths. As your AI team learns and builds a solid reputation with the executive team, you'll get more resources. That's the time to take on difficult AI projects.

### Start simple and build from there: An example

A large and established engineering company building heavy machinery was interested in using AI. The initial project was relatively simple, and the AI technologies used were something that you'd find in almost every machine learning (ML) introductory textbook. Analysis consisted of the basic clustering of problems encountered with the equipment and a basic trend prediction. But the volume of data about equipment failure was large, so that type of analysis wasn't something that could be performed manually across the complete line of equipment manufactured by this company. The details of the business case were specific to that operation, but a monetisation case was simple and straightforward. The higher level management agreed that it was a good idea to start that project and agreed to take business actions based on what AI advised. The business action to be taken (based on analysis) was changing the allocation of resources for future equipment maintenance.

A small team finished the technical side of the project quickly. It was simple to persuade managers to immediately adapt the results. A strong business case had helped the AI team leader to navigate the many organisational constraints and bureaucratic obstacles, which are a fact of life in any large organisation. (And the AI team learned how to avoid many of those obstacles the next time around.)

The solution provided high visibility to the AI team and an excellent relationship with the executive team. The AI team now had access to all the resources it needed and was able to hire new people and take on more difficult projects. Regardless, the next project selected was again relatively simple technically and, again, had large business consequences. This is a virtuous cycle: the AI team becomes more highly respected and has access to even more resources. Today, as you can imagine, the team is much larger and works on some of the most complicated AI projects in its industry.

### 3.1.2 *The look of failure*

What must you avoid to prevent a massive failure? Yes, there are many ways in which individual small AI projects can fail, but there's only one general way in which a whole AI team fails. Total team failure occurs when the AI team places all their bets on a single AI project that subsequently fails.

Let's extend our previous analogy of the hunter. How does the hunter fail? At the end of the day, a successful hunter has something to eat, and an unsuccessful one doesn't. Why do unsuccessful hunters starve in a target-rich environment?

In a target-rich environment, you don't starve because your hunt for the biggest animal failed. You starve because you spent too much time chasing the biggest animal, overlooking smaller ones that would have been an easy dinner. The time to hunt big animals is when you've honed your skills on the smaller ones and have a full belly.

**NOTE** When you're just starting your AI efforts, you'll have a hard time assembling a large enough team that's capable of tackling your biggest, toughest opportunity. The dangerous approach, 'Let's go for the big opportunities first', too often becomes a risky bet that can destroy your team.

Choosing to start with the technically challenging projects, even if they have a higher perceived value, is dangerous. If your project is complicated, your analytics team might not have enough resources to run any other significant projects. All of your eggs are in one basket.

### Medical diagnostics is a difficult problem

Suppose you're part of a hospital team that's interested in AI, and you try for the biggest *moonshot* – using AI to help in the oncology department. You form a large project team to build a decision support system for the oncology department. But cancer is a complicated illness, and in hospitals, within clinical guidelines, doctors have the last word in how things get done.

So now you're working on a complicated project requiring millions of dollars of investment, and you're trying to address an overly broad problem. You also didn't build trust incrementally with your end user (the oncologists), so they're sceptical of the results of your AI system. Even worse, they're right to be sceptical! Early system prototypes working on a different problem provided poor results. You're stuck in a vicious cycle, with no alternative, but to double down on the project into which you've already invested millions of dollars.

Your team might have done much better if they'd concentrated on a simpler problem and built good relations with the doctors first. To start with, cancer isn't a single disease, but a large group of illnesses. Significant advances have recently occurred in AI's use in medical imaging [64], allowing for a good diagnosis of heart arrhythmia, for example. Why not build a good relationship with some cardiologists and take on more difficult projects later? It certainly won't hurt your chances of success if the head of cardiology is willing to recommend your expertise to their colleagues in other departments.

Figure 3.3 shows what can happen if you take on a project that's too hard for the initial skills of your team.

Sometimes, if you start with a complicated project, you might get lucky. Perhaps you'd be able to keep management's trust for long enough to be able to deliver a successful project, and success can help when tackling larger efforts later. But is betting

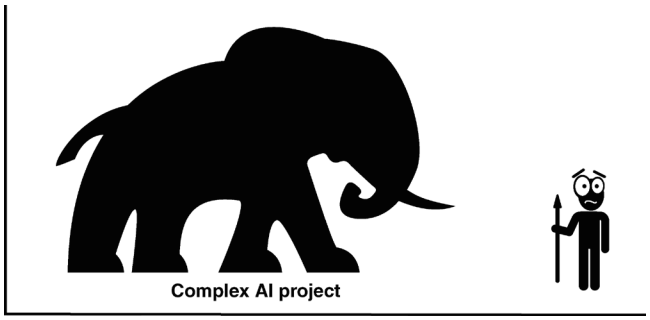


**Figure 3.3 You've cornered the mammoth on your first hunt. What are you going to do now?**

on that kind of support wise, or is it a huge risk in which your most precious resource (your team's time) is spread thin on one large attempt? Furthermore, remember that this type of success will also set high expectations for future accomplishments. Even if your first project succeeds, now you have to find another large and risky project. How long will it be before your luck runs out?

When running a data science team, the real dangers lie in taking on a complex project, persisting too long on the wrong track, monopolising your scarce resources and having nothing to show for your efforts. All the while, you're incurring significant costs. Also, you're putting management in the position that they have to continue supporting an expensive project that doesn't deliver any result quickly. What if they decide that pulling the plug is the rational thing to do?

> **WARNING** When you're building an AI project that will be used as a decision support system, your business organisation needs time *to learn how to implement* the results of your AI-powered analytics. If your management team gets nervous before the technical part of the project is completed, your project is dead on arrival.

## 3.2 Prioritising AI projects

How do you choose the right first AI project? Simple: it must be, from a business standpoint, viable, valuable and simple. That means each project must be able to deliver a result that's actionable for your business. It must have a significant business value, and you must be able to estimate how difficult it is to deliver. This section shows you how to create a list of projects that meet that criteria. Figure 3.4 describes the process that you should use to create a list of projects.

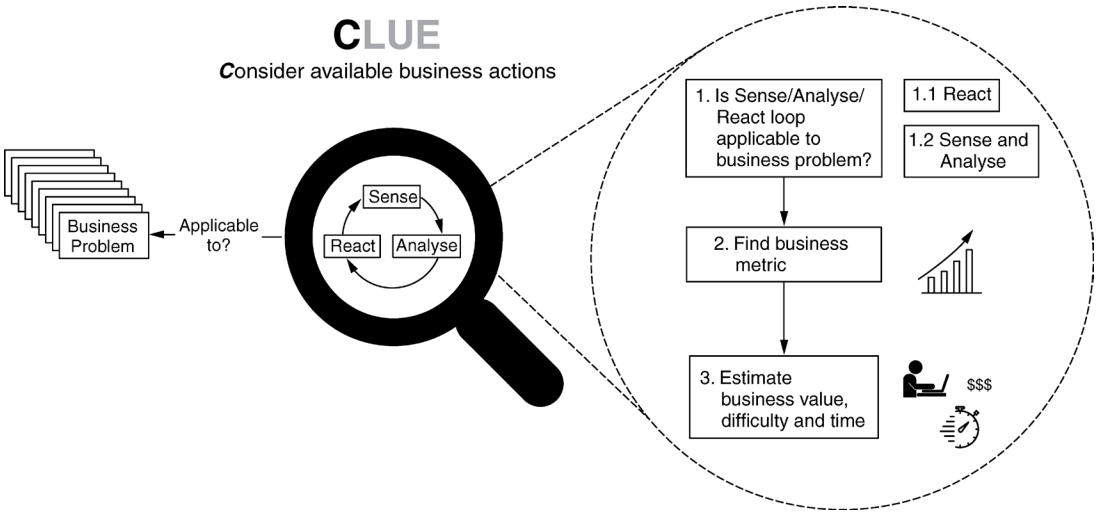


**Figure 3.4 The *C* part of the CLUE allows you to create a list of viable AI projects and estimate their complexity.**

Figure 3.4 shows these elements:

1. Start by looking at all the business areas your team is responsible for.
2. In which of those areas is it possible to apply AI and make sure you cover all elements in the Sense/Analyse/React loop? (A good reference for this step is figure 2.4 in Chapter 2.)
   - Start with React, by finding the available actions that can be taken. (See section 3.2.1 for details.)
   - Then make sure that you can cover the Sense and Analyse side of the loop. (See section 3.2.2 for details.)
3. Determine which business metric you'll use to measure how much your AI project is helping you to achieve the business goal. (See section 3.2.3 for details.)
4. Estimate the business value of the given AI project.
5. Estimate the difficulty of implementing this business case and how long it will take to implement (section 3.2.4).

I'll assume you're able to estimate the business value (step 4) of the AI project on your own, as I don't know your organisation or your business as well as you do. The rest of this section shows you how to implement the other steps in this workflow.

### 3.2.1 React: Finding business questions for AI to answer

After reading Chapter 2, you're already aware of AI taxonomy based on the role AI plays in business. Using that taxonomy is a good way to elicit business actions that can be used in the React loop. This section shows you how to apply that taxonomy.

You find the appropriate React part of the Sense/Analyse/React loop by using a process of elimination. First, look at all the areas of your business. Then determine which of those areas will benefit from AI by applying an AI taxonomy at a high level (as described in figure 2.5 in Chapter 2). Figure 3.5 shows you how to use that taxonomy to help facilitate discussions to discover the domain actions that cover the React part of the loop.

Figure 3.5 simply applies the existing business-related taxonomy of AI and asks a question designed to find a business problem that needs to be answered by AI. I'll show you an example of how you can use it.

Imagine you're working with a retailer that's nominally independent, but part of a bigger franchise. Any process changes require the franchise owner's approval, and store management isn't willing to ask for that until you've shown them that AI can improve the bottom line. Consequently, the retailer's management isn't willing to change or automate their business processes yet, but it can change the *product mix* (how many products of which type are in the store).

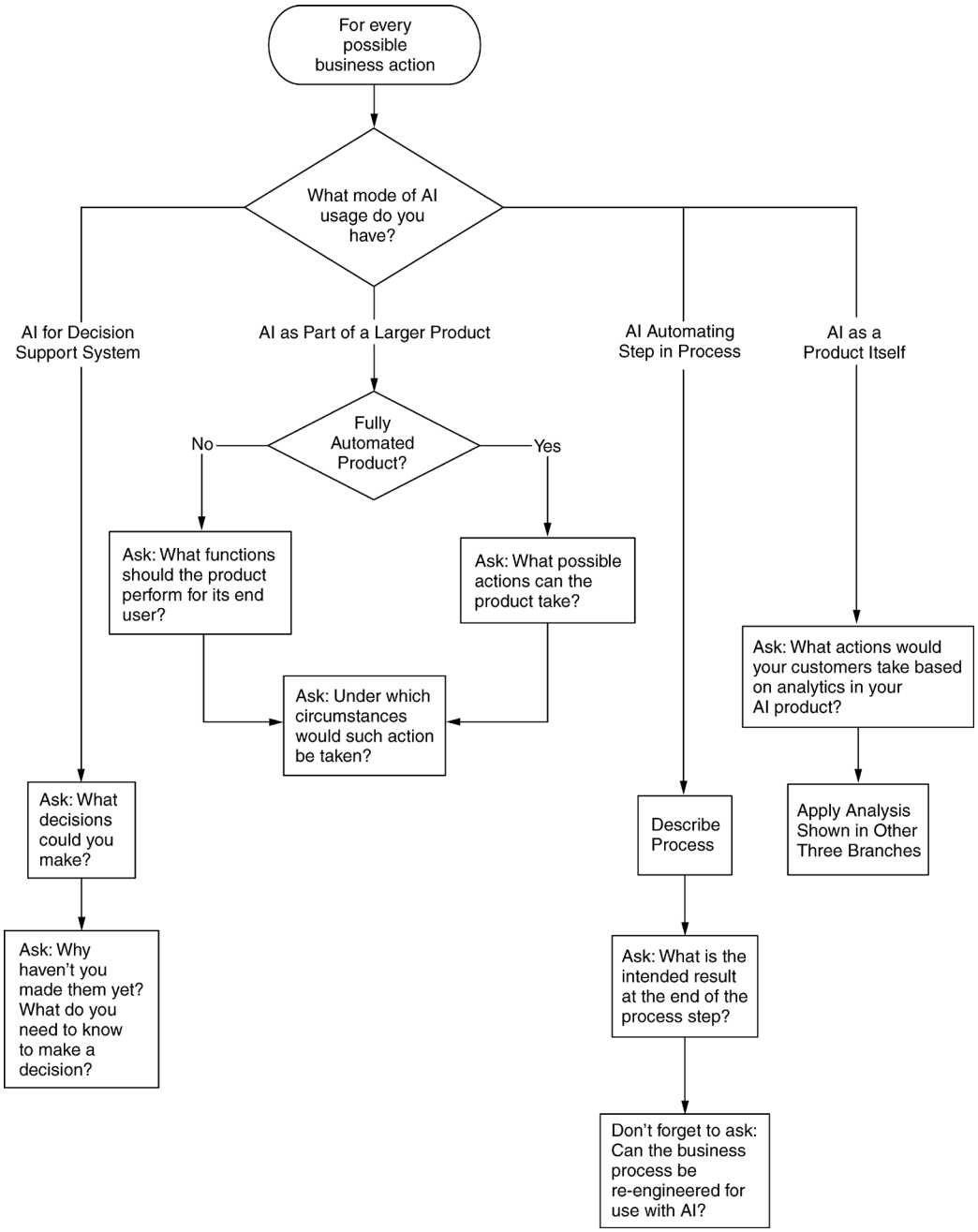


**Figure 3.5 The React part of the Sense/Analyse/React loop: finding business problems that AI can react to. Once you've identified the role AI plays in your business, ask the questions provided here.**

**TIP** Management's scepticism toward an AI-based solution is an issue that often needs to be addressed in the real world. Technologists often take AI capability for granted, but businesspeople can be sceptical. You need to earn their trust before larger and more effective AI projects can be adopted. Put yourself in the shoes of the retailer's management. How would you feel about going to the franchise owner and opening up with, "Can we change the processes that impact all the stores in the franchise? I want to try something called AI in my store, but I don't know how it will work out."

In this example, the only use of AI you can make is helping the store management itself. This is clearly an example of creating a decision support system. Look at figure 3.5 in the branch AI for Decision Support System. Questions applicable to this branch are which decisions can management make and why haven't they made them yet? That's where you learn that all your management team can do right now is change the product mix – you can put different products on different shelves. This is the React part of the Sense/Analyse/React loop.

The business question you're answering is, "What is the most profitable product mix in my store, based on historical sales?" Now you can move to the Sense/Analyse part of the loop.

### Don't stop as soon as you find the first business problem

Before we proceed, let's see if there are other options for using AI in this retail store. Look again at figure 3.5. Is there any way to use the AI as Part of a Larger Product branch? Well, there's already a video surveillance system in the store. Can you use that video surveillance system with AI? If management wants to optimise the store's product mix, what function of AI combined with video surveillance can you perform for management? This is how you find additional actionable business questions:

- Did a customer look at the product and walk away? Is that product more expensive than the competition's?
- Did a customer look for a product that's out of stock? (They approached the area where that product is displayed, saw that you're out of the product and walked away from the store.)

If you can use AI to answer these questions, you may have found a viable use case that can help you with the product mix optimisation. When using figure 3.5, remember that each area of the business can generate multiple use cases.

But before you spend too much time on using AI for video stream analysis, be aware that this isn't an easy problem to solve and it will take time to implement such an AI project.

Before you start the prototype, you decide to ask the retailer's management how they feel about the use case you've found. Good that you asked! Management tells you that they're worried about legal and public relation aspects of this use of AI in their store. They aren't interested in using video analysis. You've just saved the company the cost of implementing an AI solution that wouldn't be used even if it was successful.

### 3.2.2 Sense/Analyse: AI methods and data

Once you've established the React part of the loop, you must decide which AI algorithms you'll use and make sure that you have sufficient data to use them. This section talks about the relationships between AI methods and data.

This is one area where, if you aren't proficient in AI yourself, you need an expert with a strong command of a broad range of AI and ML algorithms to help you. The field of AI is simply too large and changing too fast for any single book to teach you enough to replace that expert. Still, the taxonomies presented in figure 2.6 in Chapter 2 can be useful to frame the discussion and remind you of high-level AI capabilities that you can apply to the business questions you want to address.

Once you have an idea of the type of AI methods you'll use, you need to be aware of the relationship between those methods and data, as shown in figure 3.6.

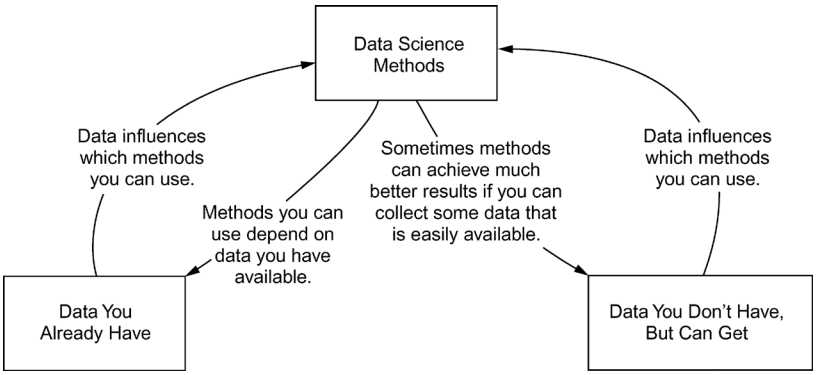


**Figure 3.6 Data science methods and data are interconnected and influence each other. Never discuss a method without asking where you can get the data needed to train it.**

You must always consider the data and AI methods that you're planning to use together. You can divide data into two groups: data that your team has and data that your team can collect.

> **TIP** Data you can collect isn't only data that you have somewhere in your organisation, but that isn't immediately available to your team. It can also be data you can acquire from sources external to your organisation or data you can purchase from business partners. Getting access to such data often requires negotiation and signed contracts.

### Considerations when collecting data

Data collection has many pitfalls, and you must carefully govern it. Ask at least the following questions:

- What does your chosen ML algorithm need to train on this data? What data format does it require? What volume of data does this algorithm need to be trained? What quality?
- What sources provide this data?
- Who owns the dataset?
- What is the cost of acquiring this dataset? How long would it take to acquire it? Is it necessary to negotiate (or even sign legal contracts) to get access to that data?
- How closely does this dataset conform to the data format that you'd obtain in a production system? Do you need to pre-process the training data before it can be used, and does the data need to be labelled?
- How big of a data infrastructure will you need to store the dataset?
- How can you collect new data after the initial dataset is constructed?
- Is it possible that your organisation has some data, but that your team isn't able to access it? It often happens that you're not able to access some of the data that your organisation already has. Data can be confidential for reasons of ethics, regulations or company privacy policy.
- What are the legal and ethical considerations (copyright, privacy policy, expectations and so forth)? You should always consider ethics, organisational policy and regulations (with GDPR [71] and HIPAA [72,73] being some examples) when collecting data.

Once you go through this checklist, you'll notice that, in some cases, you're able to easily collect the data that your ML algorithms need. In other cases, you may find that the data is unavailable or too expensive to collect.

In the retail example given in the previous section, your data scientist assures you that you can use one ML method to predict sales trends and another to optimise product mix. The names of these algorithms and methods don't mean much to you; some terms like ARIMA, LSTM and operational research might pop up. Those methods require data about past sales. Such data is available internally in your organisation, and your project can access it immediately. You've now identified a use case for which all elements of the Sense/Analyse/React loop are covered – this is a viable use case for an AI project.

### AI that recognises your food

Another example of the relationship between data and the algorithm used can be seen when using AI to do image recognition of food. The context can be a food processing plant or a smart, internet-connected oven [74] that has a camera inside it and uses AI to automatically recognise the food you put into the oven.

To make such an oven, you need to use an AI algorithm that can recognise an image (at the time of this writing, usually some form of convolutional neural network), and you need data to train such an AI method. This data consists of pictures of various kinds of food.

When you're beginning a project, you won't have many pictures of food in-house. But there may be some external sources from which you can collect data. Such sources are websites that feature pictures of food, or even pictures of the foods that users of your oven are cooking. There are additional considerations when collecting pictures of food from these sources that are typical to data you don't have, but can potentially collect. You need to make sure that copyright and privacy laws are respected.

Another interesting aspect of collecting data is that some data you collect is subtly different from the ideal data you'll want for training your AI. The position and type of the camera in the oven make the pictures of food it takes look a little different from pictures of food on a plate (which is what you'll typically find on the web). Also, ovens are greasy places, and grease on the glass may impact the image of the food in the oven. No picture of food on the web is shot through a greasy lens!

What if some of your business questions could have benefited from some AI technique that isn't known to the best data scientist you were able to find? If the best AI experts you can find aren't even remotely familiar with that particular AI technique, drop it from consideration, as it's unlikely that you can assemble a team that's strong enough to deliver using it.

**Big data or small data?**

Big data has a large mindshare, and most AI conversation occurs in the context of big datasets. Big data is certainly necessary: if you're going to store hundreds of pictures per person, at high resolution, taken by millions of people, you certainly need a large storage capacity.

But big data is just one type of data your AI algorithms can use, so you shouldn't think solely of big versus small data. Think instead about *all the data necessary to make decisions*. Sometimes you don't need (or can't get) big datasets. For example, quarterly results happen once per quarter. A drug study won't be able to recruit millions of patients. Car accidents are common, but (fortunately) aren't measured in trillions per year.

Some datasets may be small, but they still hold information about important outcomes. They also may be expensive to collect. Consider a reinsurance market like Lloyd's of London. When claims are measured in the hundreds of millions, the dataset is (hopefully) small, but important and expensive to acquire.

### 3.2.3 Measuring AI project success with business metrics

By making sure that all parts of the Sense/Analyse/React loop are covered for a specific use case, you've verified that this case is technically possible and actionable.

But how can you know how it affects the bottom line? This section shows you why you should use business domain metrics to measure the outcome of your AI project.

AI is metric driven, but you must use AI to satisfy a business goal. That goal should be represented by a business metric. That business metric should, in turn, indicate how valuable an answer to your question is when used to improve your business. The measured metric doesn't have to be a single, exact number like 'Profit improved 10%'. It can also be estimated, such as 'Profit improved between 8% and 12%'.

> **WARNING** The business metric must be defined for every single AI project. AI methods are, by their nature, quantitative methods, in the sense that they operate only with hard data. Don't attempt to use AI if you're unable to define a business metric first that you can use to measure the project's result.

Being able to choose an appropriate business metric is a business skill that's anything but trivial. But the good news is that if your organisation is using metrics correctly, you should already have a business metric defined for you – the same one that already measures business results in the area to which you're trying to apply AI. That metric should also measure how much AI improves your business.

### Examples of possible business metrics

As a word of caution, the correct business metrics are always *specific* to your organisation, not something you pick up from someone else just because it worked for them. The following metrics aren't exceptions to that rule – they're *examples* of what can work for some organisations.

- When choosing between different suppliers, one possible metric could be the *total cost of using their parts in your product.* Note that this is different from the *price* of their parts, as it includes other related costs you'll incur due to the use of those parts. This includes the cost of support or repair of your product when those parts break.
- If you're a book publisher debating how many new books to print, one good metric could be the *expected profit from the sales of physical books.* This is different from profit per book sold, as it includes factors such as the cumulative cost of the printed books, the cumulative cost of storing the books in the bookshop or some other storage facility, the schedule and price you expect to sell the books at and the cost of capital.
- Suppose you're a retailer optimising your product mix, as in the example given in section 3.2.1. One correct business metric for that retailer was *how much did the change in net income relate to all items in the product mix?* The net income is impacted not only by the sales volume of individual items in the mix, but also by the costs of changing that mix, storage, transportation and many other expenses specific to each individual retailer.

A good business metric should be customised to your organisation's needs and the concrete business outcome that you're measuring.

This book can't provide all the best practices for building good organisational metrics; that would be a book in itself (Luftig and Ouellette's book [1] and Ries's book [28] discuss the topic of business metrics). But I'd like to point out that a good business metric should be specific to your organisation, quantifiable, measurable, relevant to the desired result and free of unintended consequences.

> **TIP** Sometimes your organisation is using a business metric, but you suspect that what it's measuring is wrong and isn't helpful for running your business. If you're in such a situation, fix the metric *before* you start an AI project, and use that fixed metric to measure the end results of your AI project. Otherwise, you'll optimise AI to produce the wrong business results.

Once you've selected a business metric that you can use to measure the business contribution of your AI project, you can define the threshold. This *threshold* represents the minimum value that AI directed action must accomplish for your project to be worthwhile. As an example, if the business metric you choose to use is 'profit increase', your threshold might be that your profit must increase by at least USD 2M/year for the AI project to be worthwhile.

### Threshold for the retailer example

Thresholds are always organisation-specific, as they depend on your organisation's cost and profit structure. You need to obtain them from the business team. The following is an example of you getting targets for the retailer example given in section 3.2.1.

**You:** If the AI project provides an increase in net income of 1%, would you be willing to change the product mix?

**Retailer's manager:** While I'm willing to change the product mix, signing on new suppliers is costly. I need to account for the costs of signing the supplier: not only monetary costs, but also management attention and the time required to sign them. Our metric should account for how many new suppliers I need to sign on. Specifically, I need my *net income to increase by 0.3%* to justify signing on a single new supplier. So I can't say across the board that 1% is enough to change the product mix – it's enough if I need to sign up to three new suppliers, but not if I need to sign 20 new suppliers.

In this example, *net income* is the metric, and 0.3% for each new supplier is the threshold.

Now that you have the metrics that you intend to use with your project, you need to confirm that it's possible to measure the results of your AI project using those metrics. Present the business metrics to your AI expert, and request that they confirm that their team will be able to report the result of the AI project using those metrics. Your AI expert needs to establish a link between that business metric and one of the

technical evaluation metrics (as in RMSE, for example) that they intend to use in the AI project.

**TIP** A business metric is appropriate for an AI project when it correctly measures the business result you want to achieve, and when technical experts know how to report their technical progress using that business metric.

#### What if you can't find an applicable business metric?

The inability to easily recognise business metrics by which an AI project should be measured raises a big red flag. If you can't quantify the business result you're hoping to achieve, you have to ask yourself and your colleagues, "Should we start an AI project in the first place?"

Without the ability to define your business metric, you also can't define any value threshold for your AI project. Without the threshold, you don't know if the results of your project are substantial enough to use in your business. Without a business metric and threshold, you also have no way to estimate the business value of your AI project, and you won't know if it's cost-efficient. In all cases, in the absence of a good business metric, management of the AI project will degenerate into a series of decisions made by gut feeling.

The inability to select a business metric for your AI project may be a sign of a poorly constructed business metric, creating a situation in which you may provide value to the business, but can't measure it. It can also indicate that you're not able to provide a business value at all. Or it may indicate that the AI project is so disconnected from the core business that the business has no idea of what to do with the results. Such projects are risky at best.

Sometimes there is a clear business value, but management may perceive it as intangible and something that can't be measured. Examples of intangibles might be employee morale and brand value. This happens when management isn't aware that intangibles could be measured by using a range instead of a single number. See Hubbard's book [75] for many examples of the best practices for using ranges to measure so-called intangible quantities in business.

### 3.2.4 Estimating AI project difficulty

Now you've confirmed that the AI project you're considering is technically possible, and you have a way to measure its business impact and its business value. To determine if the AI project is viable, you need to know its cost, the difficulty of its implementation and how long it will take. This section details considerations in estimating those quantities.

To estimate difficulty, you need to sketch an outline of the technical solution that will be used in the AI project. You need to have representatives from your data science and data engineering teams, plus your software architect, work together on this outline. Your goal is to provide a high-level outline of the solution, to compare a selection of possible AI projects.

Once you have this outline, use it to estimate the difficulty, cost and length of time to deliver the project. These are, again, rough estimates intended for comparing different AI project options.

### Considerations for estimating AI project difficulty

When estimating AI project difficulty, be aware of the following considerations:

- Account for the time required to collect the data you need.
- Do you have the infrastructure necessary for your data size? Do you even need a big data framework?
- If you're using large datasets, don't forget to account for the time necessary to process the data and train AI algorithms.
- Does your team have all the skills necessary to cover this use case? What are the gaps in their skillsets, if any? (As advised in Chapter 2, a team leader should be aware of knowledge gaps in the team.)
- Is it certain that the project is even technically possible? Do you understand the proposed AI methods enough to be positive that your team can build it, or do you just know that area of AI enough to assume that the project is possible?

Once you can account for the specifics of the AI project, you can use any estimation methodology that you're familiar with in your organisation to estimate other software projects.

**TIP** Remember that people aren't particularly good in estimations [75], are worse if the estimate is based on just a sketch of the solution and are worse still if they must estimate in technical areas they know little about. Your estimate will be very rough by necessity. It's only intended to compare different AI project options, and you should make no strong commitments to management based on this estimate.

At this point, you have all the information you need to create a list of viable and actionable AI projects. You know how to determine whether the proposed project is actionable and technically possible. You know how to measure the business value of the AI project, and you can make a rough estimate of the cost, difficulty and duration. Now it's time to select the first AI project to run. The next section guides you through the selection and preparation for running your first AI project.

## 3.3 Your first project and first research question

As discussed in section 3.1, if your goal is to build an AI team that's a long-term asset to your business organisation, then initial projects should be simple and fast to deliver. The criteria for selecting your first AI project to run is therefore simple: choose the project that's fast to deliver and has significant business value. When you've chosen that project, you need to do the following:

- Define the research question that the project would answer (section 3.3.1).
- Organise the project so that if it fails, it fails fast (section 3.3.2).

The rest of this section shows you how to define the research question and explains what fail fast means.

### 3.3.1 Define the research question

You've chosen your first AI project. That project has a clear business question that needs to be answered, and that question is written in a form that a business decision maker will understand. Now that question needs to be translated into a format that AI can understand–the 'research question'. This section shows you how to ensure that your research question matches your business question.

Suppose you're a manufacturer, and your research question is, "Should I go with supplier A or supplier B based on the quality of their product?" You need AI to answer this question, but there's a problem: AI has no idea what the concept of *supplier* means.

#### AI DOESN'T UNDERSTAND BUSINESS CONCEPTS

People unfamiliar with AI capabilities are often under the impression that AI can find some novel business reaction that's escaping human capacity. With the current level of AI, that's rarely possible.

There's no AI algorithm that could look at a retailer and figure out how to improve profits. The reason is that AI doesn't know what the words *retailer, product* and *profit* mean. AI has no idea what a supplier even is, much less what makes one supplier better than another. Those are business concepts. Nor does AI understand that there could be ethical, public relations and legal considerations regarding the use of AI in analysing surveillance video in the store.

Business concepts may be understandable to humans, but the data regarding those concepts must be packaged in a format that AI/ML algorithms expect. That's your data science team's job. To do that, they must first formulate a research question. You can view the research question as a translation of the business question into a form that AI can understand.

#### THE CONTRACTUAL LANGUAGE OF THE TECHNICAL DOMAIN

AI methods operate in a technical domain. Language used in that domain is contractual in nature and is of the form, "If you present me an input in format X, I guarantee that I'll provide answer Y." Those contracts are often convoluted and require an expert in the field to really understand their precise meaning, as well as all the implications of what is said.

Let's demonstrate this contractual language in some concrete examples:

- AI that's based on classical statistical methods will use the language of hypothesis testing: "Is there a statistically significant difference between part samples from supplier A and supplier B at $p = 0.05$?"
- AI that's based on image recognition will express in the language of ML: "If this is a picture of your part, I can tell you with 95% confidence that it's much more similar to the class of parts you labelled as defective, compared to other classes."

AI can also tell you that, overall, it reaches 98% accuracy in correctly classifying parts between classes.

- AI used in the publishing industry, to predict how many physical books should be printed in a second batch if the first batch sold in three months, might be based on a time series model. Depending on the model used, one research question can be, "Predict with a 95% confidence interval the book sales in the next three months, based on the sales in the previous three months."

If you aren't a data scientist, your head is probably spinning right now. What do those sentences even mean? Sounds geekish? That's because it is! AI methods are defined in an abstract format that's intended to be understood by computers and data scientists, not by business users. To solve business problems, you need to translate them into *AI language*. That's what I mean by *formulation* of the problem when I say that ML is a combination of formulation, optimisation and evaluation. Figure 3.7 shows this process.

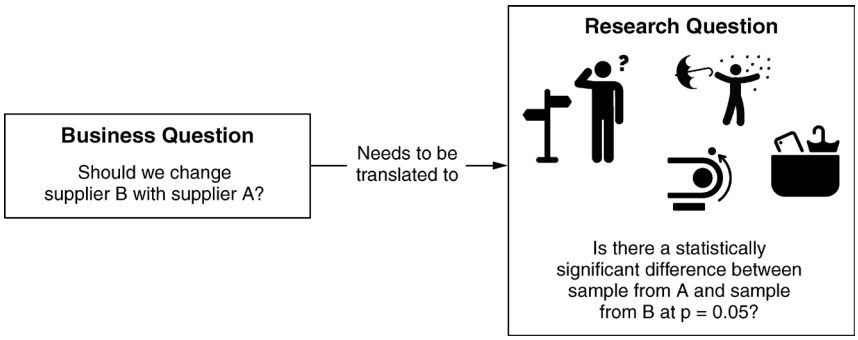


**Figure 3.7 The translation of a business question into a research question. AI doesn't understand business concepts. If you aren't familiar with statistics, a research question formulation might be difficult to understand.**

The job of your AI experts is to select the appropriate research question and translate it into a format the AI methods can answer without compromising its relationship to your business question.

> **WARNING** Projects often go awry when the business and data science teams don't communicate closely. Business questions can be incorrectly translated to poor research questions, causing you to get an answer to a question you didn't ask. The problem is compounded if you then use that answer to take the wrong business action.

It's important to understand that this translation is a highly complex activity requiring that your team share an understanding of both the business and AI domains. This translation isn't straightforward. It's almost impossible to devise a translation that

evaluates all possible business actions. It's the job of the business leader to guard against misalignment between the business question and the research question.

### Misaligned business and research questions

Here's an example of how business and research questions can get misaligned in a way that a business leader or a data scientist is unlikely to catch unless they talk directly.

**Business leader:** Give me an example of one possible answer that you'd provide when you finish the analysis.

**Data scientist:** We have enough statistical evidence to infer that there's only a 5% chance that we'd get the result we did if supplier A wasn't indeed better than supplier B.

**Business leader:** So, you're telling me that supplier A is better than supplier B? I expect that, three months from now, we'll have a big project that we need a good supplier for. I plan to replace supplier B with supplier A. We'll increase orders from supplier A 100 times. Is your analysis sufficient to support such a business decision?

**Data scientist:** Well, what we know is that on the sample we tested, A was likely to be better than B....

**Business leader** [thinking to himself]: *This is interesting. What does the technical jargon mean? This isn't a simple yes or no. Let's dig deeper...*

**Business leader:** Wait. How is that different than what I just said? Give me an example of a situation where you'd report that we have enough evidence to *interfere*.... and it would still be wrong to drop supplier B.

**Data scientist:** Well... when I say we have enough statistical evidence to... um... *infer,* we cover only the sample that was given to us last week. We can't say that would still be the case in three months. We also can't say that supplier A could deliver the same level of quality if the order you make is 100 times the size of the sample we tested.

**Business leader:** I see. What if it's important to me to consider trends of improvement in suppliers A and B? I need to know what's likely to happen in three months, if current trends continue. I'll place an order of 10,000 parts to be delivered by supplier A in three months. Does your analysis support these actions?

**Data scientist:** No, it doesn't. We need to perform a different type of analysis...

#### BEST PRACTICES FOR THE BUSINESS LEADER OF AN AI PROJECT

The 'Misaligned business and research questions' sidebar shows an example of a project that could have gone horribly wrong, as the business leader and data scientist are talking about totally different questions. But until that conversation happens, you won't notice that you're not talking about the same thing. The following provides best practices for business leaders who are starting an AI project:

- Never sign a document that proposes a highly technical research question you don't understand. Instead, call a meeting with the data scientists. Ideally,

someone who has a strong background in both business and AI should facilitate this meeting. If there's no such person in your organisation, you should enlist a consultant to help you.

- Ask your data scientist to provide you with a couple of the possible answers that the proposed research question can produce.
- Play out a scenario analysis (as in the previous example). Take the answer to a research question and describe what you think it says. Then, state the exact business decisions you'd make, explaining all your assumptions.
- Always repeat how you interpret the answer that the data scientist gives you. Use simple, non-technical terminology.
- Don't worry about looking stupid if you misunderstand something. Finding misunderstandings at this stage saves a ton of money later. You'd really look stupid if you waited until the wrong AI project was run.
- State clearly the business actions that you intend to take, based on their answers. Ask if those business actions are reasonable.
- If the data scientist responds to your planned business action with anything other than a simple yes or no, investigate in which circumstances it would be wrong for you to take such a business action, based solely on the results of the analysis.
- Don't be afraid to further explore your research question, together with your AI experts. Get educated about the details of what the research question means; by the same token, educate your data scientist on the details of your business.
- Understand that the mapping between a business question and a research question is never perfect. Research questions rarely correctly capture all aspects of the business problem. The question is, "Are those aspects important to you?"
- Beware of discussing business problems using highly technical terminology. Terminology that isn't shared between all meeting participants is an excellent vehicle to hide misalignment between technology and business. If a business doesn't have properly thorough discussions to ensure that research questions and business problems are aligned, they're in fact leaving it up to the technology team to fill in critical business details.
- Always make your business experts available to the technical team for consultation about clarification of important details. It isn't realistic to expect that the technical team will make optimal business decisions when the specification of the business problem is vague.

**WARNING** Always perform a scenario analysis of a research question *before* the AI team starts its work. The chance of a project succeeding if you ask the wrong research question is about zero.

A final point to remember when defining a research question is that correspondence between business questions is not "one research question per one business question." You might need multiple research questions to cover a single business question. It's also possible (although rare in practice) to have a single research question that can answer multiple business questions.

### 3.3.2 If you fail, fail fast

You're starting this first AI project with the assumption that it will be an easy project to deliver. However, your estimates were coarse; it's quite possible that the project, once you start it, will be more difficult to deliver than you initially calculated. This section explains how you should manage that initial AI project so that if the project is unexpectedly difficult, that becomes immediately obvious.

> **TIP** You should optimise your project delivery process for *speed to success*. When hunting in rich hunting grounds, the more you hunt, the more you catch. In the early days of your AI efforts, you shouldn't persist on projects that looked easy early on, but on closer examination were difficult. Instead, stop them early and use the remaining project's time to start an easier project instead.

Your project should start with a proof of concept that builds a quick prototype. That prototype serves four purposes:

1. It demonstrates to you that the engineering team has the technical expertise needed to deliver the project.
2. It gives you a concrete AI implementation, which is the Analyse in the Sense/Analyse/React loop you identified. Now you can test the React part of the loop (for example, you can test how difficult it would be to implement the required business action).
3. The prototype can be analysed to determine how your proposed system solution behaves when exposed to either more data or different ML algorithms. Chapters 6 and 7 of this book show you how this analysis is performed.
4. The prototype shows you how difficult it will be to implement your AI project. If the level of difficulty is much greater than you expected, that should be quickly obvious.

Until you have an experienced team that has passed through such processes many times, don't get stuck on projects that take a lot of time to implement. As a hunter, you won't starve because you failed to catch a single prey; you'll starve because you were trying to catch a single prey for way too long.

> **TIP** If your project is more difficult than expected, pause it and choose an easier one instead. *If you stumble upon the Gates of Hell, turn around and run!*

## 3.4 Pitfalls to avoid

When running an AI project, there are some common pitfalls that you should avoid. Some of the most important ones are as follows:

- Not communicating with the organisational actors that own the React part of the Sense/Analyse/React loop, or, even worse, not working with them at all until your AI project is well on its way
- Transplanting use cases (and metrics) from other projects or organisations
- Running fashionable AI projects that are likely to grab headlines

- Believing that you can buy a tool, any tool, that will give you a sustainable advantage
- Hoping that throwing random analysis at your data will produce results
- Selecting which project to run based on a 'gut feeling' instead of the results of analysis

This section discusses each one of these pitfalls in more detail.

### 3.4.1 Failing to build a relationship with the business team

When using AI as a decision support system, it's never enough to just deliver a good analysis; you need to execute well on the specific business actions recommended by an AI-powered analysis. That means that executive attention must hone in on the link between the analytical result and the business action. This section highlights why the AI team must build good relationships with the department of your organisation that will take business actions, based on your AI analysis.

> **TIP** Analytics is just like a speed gauge in a car. If the speedometer is telling you that you're going too fast, the driver must reduce the car's speed. Who's the equivalent of the driver in your project/organisation?

A non-specialist can misinterpret analytical results. A classic problem is that non- specialists won't understand the limits of the analysis and when the assumptions basic to the analysis are violated. An example of that problem was shown in section 3.3.1, when the research question and business question were misaligned.

I've personally witnessed several organisations hand off an analysis report to separate business teams. These business teams proceeded to take business actions without the input of data scientists. This is always a mistake.

> **TIP** Analytical expertise should always be represented in any group that's discussing how to react to analytical results. You need to ensure that the business team understands your analytical results and prescriptions fully and correctly. The analysis result must be valid in the light of the intended business actions.

If you're the leader of an analytical project, your job isn't done when you deliver the results. Your job is done when the analyst's prescription is successfully implemented. You must build good working relationships with departments that will implement the analysis. If the analysis has prescribed a particular business action, don't underestimate the need for you to help and follow through with its execution.

### 3.4.2 Using transplants

People are often tempted to copy what worked for the people and organisations that surround them. As a result, you see what I call *transplant projects*. Here, an enterprise decides to form an AI team and embarks on some AI project they've heard other organisations similar to theirs performed. This section explains why transplants are a bad idea.

Examples of transplant projects abound. Some examples are projects like 'let's have our own recommendation engine' or 'let's do sentiment analysis of customer feedback'. Sometimes these projects make sense in the context of the business, but all too often they're just vanilla use cases that you heard about from someone else and didn't analyse in the context of your own business.

> **NOTE** For some reason, people have more common sense when they're thinking about real transplants as opposed to business transplants. You'd never get a kidney transplant just because it worked well for your neighbour. Why should you behave differently in your business?

Instead of just blindly adapting a project that worked well for someone else, consider it to be just one of many possible AI projects. Use the analytical approach presented in this Chapter to determine which AI project you should start first.

### 3.4.3 *Trying moonshots without the rockets*

Many of the world's largest technology companies have made fortunes based on the use of data. In the core, companies such as Google, Microsoft and Baidu are heavily dependent on AI for their success. They have significant research capabilities and have a vested interest in ensuring that they won't miss the train of important AI advancements. This section explains why your organisation shouldn't blindly follow those companies.

#### Imagine that you're a CEO

Suppose you're running a company that's making USD 30 billion a year, and you're in a business that's associated with AI. Let's go a step further and assume that there's a 1% chance that someone in the next 10 years might invent something approaching a strong, human-level AI – so called Artificial General Intelligence (AGI) [76]. If the search for AGI fails, there may still be an autonomous vehicle [38] as the consolation prise. Finally, you know that your competitors are investing heavily into AI.

Will you invest substantial money into AI and hire accomplished researchers to help you advance the frontiers of AI knowledge? Or will you opt not to invest in AI, and accept the risks that:

- Your competitors develop AGI or autonomous vehicle technology. Your company may have been better positioned, but you failed to even try!
- Your error will be taught in every business school for many years to come.

While the logic from the sidebar 'Imagine that you're a CEO' applies to businesses such as Google, Baidu or Microsoft, there's an unfortunate tendency for many enterprises to emulate these companies without understanding the rationale behind their actions. Yes, the biggest players make significant money with their AI efforts. They

also invest a lot in AI research. Before you start emulating their AI research efforts, ask yourself, "Am I in the same business?"

If your company were to invent something important for strong AI/AGI [76], would you know how to monetise it? Suppose you're a large brick-and-mortar retailer. Could you take full advantage of that discovery? Probably not – the retailer's business is different from Google's.

Almost certainly, your company would benefit more from AI technology if you used it to solve your own concrete business problems. This means that instead of teams populated by the smartest researchers and processes oriented toward the acquisition of new AI knowledge, your organisation needs people who know how to make money in your business domain *with existing AI technologies*.

Don't emulate organisations richer than yours without first understanding how you'd exploit success. For most organisations, the road to success isn't found in advancing the frontiers of AI knowledge, but in knowing *how to tie AI results into their business*. You need a data science team focused on applications, not research. That doesn't mean that you shouldn't hire bright PhDs, but that the leadership of your AI teams must primarily be experts in applying AI to the task of making money.

### 3.4.4 It's about using advanced tools to look at the sea of data

Another common pitfall is the belief that you can buy an AI or big data tool that will make it trivial to look at your data, find insights and then monetise the insights found. Some organisations adopting AI might even take the attitude that the main focus of early AI efforts should be on finding the right tools. This section explains why this is a pitfall to avoid.

> **TIP** If monetisation is trivial, so is explaining how it happens. Ask suppliers detailed questions until you really understand the finest points of how to apply an out-of-the-box tool all the way from the point of purchase to the endpoint where your organisation makes profit as a result of use of that tool.

In most business verticals, it isn't trivial at all to monetise AI. And while many tools can help you get there, it's unlikely that these tools can solve monetisation problems for you. Even if there are tools that let you monetise by just installing and running them, what you're dealing with is a commoditised use case. Heck, someone already has a product that does it!

> **TIP** The early focus for your business should be on finding AI projects that provide a concrete business value. Tools are enablers of those projects.

A salesman might advise you to "Build a large data lake and unleash your data scientists on it; there has to be something in all that data." You might even have been given an example of the unexpected insights that only analytics on a big dataset can provide. However, those situations are rare and unpredictable. Don't count on the

tooth fairy. Don't start with the Analyse part of the Sense/Analyse/React loop. In our framework, always start with the React part.

> **WARNING** It's always possible that there might be something special lurking deep within your data. With the proper analysis, it might give you some unexpected business idea that you can implement and with which you can make a ton of money. While *possible*, this is certainly not *guaranteed*, isn't *predictable* and there's a big question of i*s it likely enough to justify it as a main strategy you should adapt?* Worse, the lucrative bluefin tuna you hope to catch might turn out instead to be a slimy monster of the deep. Instead of going on a fishing expedition, organise your early AI projects for predictable success.

### 3.4.5 Using your gut feeling instead of CLUE

Often a decision about running an AI project is made in a haphazard way, as little more than a technical idea that excites the team. Running an AI project primarily because you want experience with the underlying technology is the tech equivalent of buying a sports car. This section explains why following your gut may result in poor business results.

#### Video analysis of the behaviour of retail customers

Let's return to our retailer from section 3.2.1, for whom you're optimising a product mix. There were two proposed approaches: one based on predicting sales trends and another one based on video recognition of customers' behaviour.

If I put my data scientist hat on, I'd have to admit that video recognition of customer behaviour is a more technically interesting project for me. That project would excite a lot of the technology teams today. It uses cutting-edge AI video recognition abilities, whereas sales prediction may make do with older time series analytics methods.

Sometimes that technical allure is all it takes for a team to decide to build a prototype, and the data scientist in me certainly understands this urge. However, this is a classic 'gut' or 'Oh, shiny!' approach to project selection.

To see why this would be a mistake in this example, recall what happened in section 3.2.1. When you're talking with management, you may learn that your business isn't comfortable with the legal and public relations ramifications of doing video analysis of customers' behaviour. Your effort is unlikely to be adopted even if it's technically successful. This doesn't take long to learn when you bother to talk with business leaders about your proposal before building a prototype.

Also, even if you somehow persuade management to allow you to continue building your AI prototype, you failed to define a business metric for measuring success. Now you've created problems in managing your project. Suppose your project is in progress and has achieved some initial success. How would you know if it's good enough to release? How precisely does it need to recognise customer behaviour? Can it make recognition mistakes, and, if so, under which circumstances? Which mistakes are most damaging?

Be extremely sceptical about counting on intuition to select which AI project to run first. Section 3.2 showed you the steps necessary to correctly determine which is the best project to run. When selecting a first project, there are simply too many moving parts to consider for gut feeling to provide the right answer. You need to verify that the project is actionable, technically possible and business valuable. You need to know its cost, as well as the outline and difficulty of the proposed technical solution. It's highly unlikely that you can assess all those attributes of a proposed AI project by just thinking about the problem for a minute or two.

**TIP** Above all, be on the lookout for any situation in which, during a well-attended and important company meeting, everyone immediately exclaims, "This looks like a great idea!" Such a social situation isn't exactly conducive to encouraging people to perform the careful analysis needed to disprove the group consensus. In short, beware of groupthink.

### But we're using an MVP approach!

Some teams are Agile and/or use Lean Start-up [28] methodologies for developing their software projects. In a Lean Start-up methodology, the team is encouraged to dice projects into small chunks of work that can be presented to the customer for feedback. This chunk of work is called the *minimum viable product* (MVP). Part of the Lean Start-up methodology is that if you find that your MVP isn't what the customer wants, you can then try something else – the so-called pivot.

Some will argue that because you're building an MVP, you should quickly select some initial AI idea, show it to the customer and see what the customer says. *Don't do that!*

Using MVP has many advantages with real-world products, and CLUE can combine well with a Lean Start-up strategy. However, MVPs taken alone aren't solving the same problems that CLUE is addressing; for example:

- If you choose an MVP based on a gut feeling, you've started a project before knowing if the business is willing and able to adopt your analytical solutions.
- Although MVP can show you that you're on the wrong track faster, you're on the wrong track, right from the beginning.
- If your gut feeling was to think about the analysis you can do (versus starting with the React part of the Sense/Analyse/React loop), you're playing *analysis roulette*. You're hoping and praying that the analysis you do will yield a result that your business can somehow implement.

MVPs aren't valid excuses to promote a gut feeling approach to selecting and running an AI project. MVPs reduce the cost of finding out that you're on the wrong track, but that's just *reducing the price of failure*. The ability to pivot isn't an excuse to run projects haphazardly, hoping that with enough random AI ideas you'll stumble on something that just happened to be actionable.

The CLUE approach can be integrated, and is compatible, with MVPs and Lean Start-up. The C part of CLUE analysis is about selecting a first AI project, and such a project can be an MVP – an MVP that's based on analysis, not gut feeling.

The biggest cause of failure of AI projects today might be technical. But even among technically successful projects, there are far too many that aren't even used by the businesses that paid for them. Those AI projects shouldn't have been started at all and were usually started because a gut feeling about their value was wrong.

## 3.5 Exercises

The following questions each give a concrete business scenario and then ask follow-up questions about that scenario. Please answer the following questions:

**Question 1:** Suppose you're working in the publishing industry, and you're wondering if it's better to release printed, electronic and audiobooks at the same time or one after another. Also, if delivery is staged so that printed books are released first, how long should you wait before releasing the other formats? Within this setting, answer the following question: "What business metrics should you use?"

**Question 2:** If you're a business leader, define a business question and an appropriate metric to measure it. Think about some hypothetical scenarios not directly applicable to your organisation (for example, some scenarios related to philanthropy). Think about actions that you can take while running a non-profit organisation. Use the techniques introduced in Chapter 3 to select your first hypothetical business question, as well as the metrics you'd use to measure success.

**Question 3:** Once you've identified your business question from the previous exercise, take your senior AI expert to lunch and talk about the business problem. Ask them how they'd formulate a research question. Use the process described in Chapter 3 to check whether or not the answer supports the business action you intend to take. And, while you're having that lunch, talk about how you'd find a dataset to answer such a research question. Do you think you can acquire that dataset?

## Summary

- AI, when introduced to new businesses, falls on rich hunting grounds. Don't start by chasing difficult projects that tie up all your resources and destroy you if they fail. Start instead with simple projects that have big business value and are quick to complete.
- Use CLUE to select and organise AI projects. The *C* part of CLUE (figure 3.4) allows you to create a list of actionable AI projects that you can implement and helps you estimate their size and value.
- The business question that AI needs to answer must be translated into a technical format by defining a research question. If that translation is incorrect, it can wreck the business outcomes. Before starting a project, check your research question using scenario-based testing.

- Use business metrics to measure the progress of your AI project. Business metrics should be customised for your project and organisation. Don't start an AI project if you don't have the business metric for measuring its success.
- Organise your AI project so that if it fails, it fails fast.
- Start with the proof of concept. If the project happens to be more difficult than you thought, stop it and work on an easier project instead. The goal is to *optimise time to the next success.*
- There are common pitfalls to avoid when running AI projects. They include failure to build a relationship with relevant business leaders, transplanting use cases, adopting a 'moonshot' project, but missing the rockets, placing too much hope in random tools (or random analysis) and substituting a gut feeling for CLUE.

# 6 Analysing an ML pipeline

***This Chapter covers***

- Determining if you have the right ML pipeline before it ossifies
- Economising resources in your AI project
- Performing MinMax analysis on the ML pipeline
- Interpreting the results of a MinMax analysis

It is essential for you to ensure you don't have a wrong and inadequate ML pipeline that's incapable of fulfilling the business goals of your ML project. The most critical business question about the ML pipeline is, "How well does this pipeline do in business terms?" This Chapter shows you how to analyse an ML pipeline and get an answer to that question.

Once you know how well your ML pipeline does in a business sense, you can analyse it to determine if it can meet your business goals. The name of the analysis

you'll want to perform is MinMax.[1] It can do MinMax analysis early in the project, and it consists of two parts, each of which answers a different question:

- *The Min part of MinMax analysis* – If your life depended on releasing your AI project tomorrow, how well would the simplest implementation of your ML pipeline do? Can such an implementation meet your business goals?
- *The Max part of MinMax analysis* – What's the best possible result you can get with the current structure of your pipeline? Before you put effort into the best possible implementation of each stage in that pipeline, would that implementation meet your business goals?

As practical people, we need the answer to these questions in business terms, not in the form of technical metrics (such as 99.9543% accuracy).

In section 6.1, you'll learn why you should care about analysing your ML pipeline. Section 6.2 shows you how to economise resources devoted to your ML pipeline, and section 6.3 shows you how to use MinMax analysis to determine if you have the right pipeline that's capable of solving your business problem. Section 6.4 shows you how to interpret the results of a MinMax analysis, and section 6.5 shows you how to perform a MinMax analysis. Finally, section 6.6 presents FAQs on MinMax analysis.

## 6.1 Why you should care about analysing your ML pipeline

By now, you know that an AI system is more important than the sum of its parts and that an ML pipeline is one of the primary software artefacts that determines the behaviour of the system. An ML pipeline rapidly ossifies and allowing the wrong ML pipeline to ossify is a costly mistake. That's why managing the ML pipeline must be data driven. The analysis of an ML pipeline gives you that data.

All project management decisions are made under a time constraint (see figure 6.1). When a project is in progress, every day costs both money and opportunity. When accounting for the costs, you must account for all costs – the cost of changing something, and the cost of staying on the current course.

> **WARNING** Inaction is the decision to stay on the current course and can sometimes be as dangerous as taking the wrong action.

The form of ML pipeline you use in your project is one of the most critical project management and software architecture decisions you have to make. If you choose the wrong ML pipeline, you (and your wallet) will know. However, by that time, it'll be too late. Analysing an ML pipeline early in the project gives you confidence that you're using the right one. To be useful to a team in business and industry, such an analysis must do the following:

- Be easy to learn how to perform
- Be cheap to perform

[1] Just in case you're familiar with the Minimax algorithm from game theory [114], you shouldn't confuse MinMax analysis with the Minimax algorithm – they're totally different concepts.

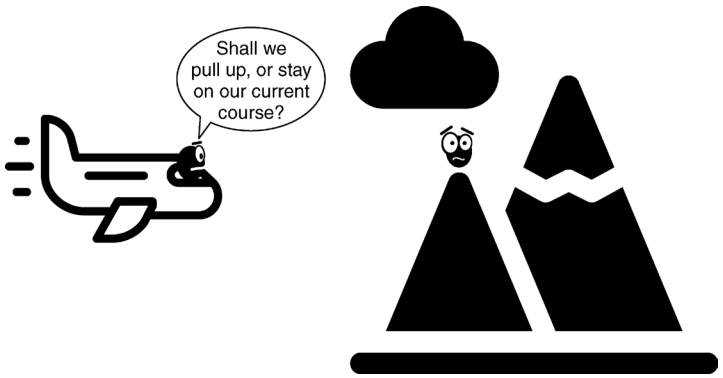


**Figure 6.1 Project management decisions are made under time constraints. Some projects are like the plane in this figure and must correct their course before they crash. Doing nothing can sometimes be as dangerous as taking the wrong action.**

- Provide results that are intuitive to interpret
- Provide a reasonable level of confidence from the information it returns

In this Chapter, I'll show you an analysis method that meets these four requirements. The next Chapter provides yet another. You, as a leader, need to learn what type of analysis to ask your team to perform and how to interpret the results of such an analysis.

**TIP** An analysis must be *cheap* to perform. It must balance the cost of analysis (cost of asking the questions) with the value of knowing the answer. This balance is what D. W. Hubbard [75] refers to as the *expected value of perfect information* [79].

Analysing an ML pipeline is something you should repeatedly do for each AI project and for each ML pipeline you consider for each project. Even if you're a non-technical reader, spend the time now to understand how analysis works. The understanding you acquire will help you make the best decisions, not only for your current AI project, but for future projects as well. In Chapter 8, I'll show you that the methods you'll learn have a far broader application than simply an analysis of the ML pipeline.

### Similarities with the stock market

There's another field in which you make decisions under time constraints and with limited and imperfect information. That field is investing.

Let's hear what Ray Dalio, who built the biggest hedge fund in the world, has to say about how to make decisions under uncertainty [29]:

> *"'He who lives by the crystal ball is destined to eat ground glass' is a saying I quoted a lot in those days. Between 1979 and 1982, I had eaten enough glass to realise that what was most important wasn't knowing the future – it*

***(continued)***

*was knowing how to react appropriately to the information available at each point in time."*

When building an AI capability for your organisation, you'll make many decisions. The goal is to tilt the balance of probability in your favour.

## 6.2 Economising resources: The E part of CLUE

How do you know that the ML pipeline you're using is the proper pipeline to use long-term? Should you continue using the current ML pipeline, or should you replace it with a different pipeline before it ossifies? Which part of the pipeline should you improve first? You should make such decisions based on the data, and this section gives you an overview of the tools that answer those questions. Those tools are:

- *MinMax analysis* – Answers the question, "Do I have the right ML pipeline to meet my business goals?"
- *Sensitivity analysis* – Answers the question, "How much would my business result change if I modified the implementation of a single stage of the ML pipeline?"

This Chapter concentrates on MinMax analysis; Chapter 7 presents sensitivity analysis. Together, those two analyses let you allocate your development resources toward the right ML pipeline and the right stage of that pipeline.

Concentrating on your ML pipeline is a logical next step of your project. So far, you've applied the *C*onsider (available business actions), *L*ink (research question and business problem) and *U*nderstand (the technical answer in a business context) parts of the CLUE process. By performing these parts of that process, you've ensured the following:

- Your AI project can viably affect the business. You know that, in your project, there's a way to apply the Sense/Analyse/React loop to your problem and that the React part is possible. Chapter 3 covered this material by describing the *C* part of CLUE.
- Your AI project defines the ML pipeline that you're planning to use.

What you still need to do is *economise* your scarce resources during the construction of the AI project. To *E*conomise (the *E* part of CLUE), you need to determine that you're using a reasonable ML pipeline for solving your business problem. You also need to decide what the best stages of the ML pipeline are to improve.

MinMax analysis answers the question, "Do I have the right pipeline?" It lets you know what the best possible business result is that you could hope for from your current ML pipeline. It lets you know that the ML pipeline you're making is the one that could support your business goals, and it lets you know that before the ML pipeline starts to ossify.

Unless you already have an ML pipeline that completely solves your business problem, chances are, you need to improve it. You have limited resources, and you need to assign them optimally to get the best overall project results. Sensitivity analysis [115-117] answers the question of which stage of the ML pipeline you should invest in next to get the maximum return on your investment. Figure 6.2 shows you how to integrate your project with MinMax and sensitivity analysis.

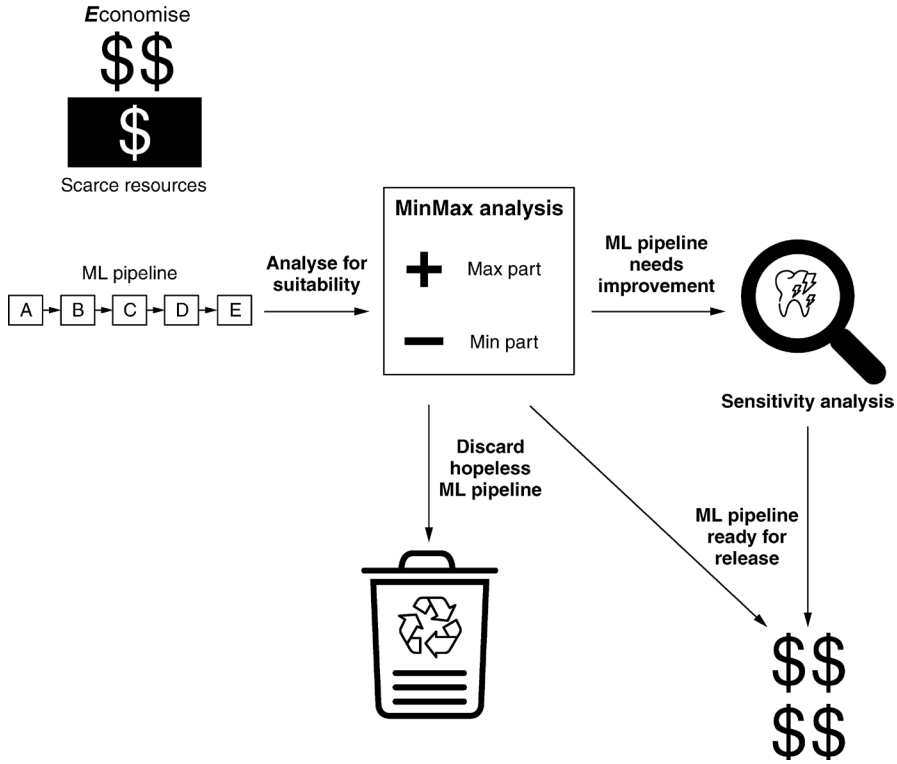


**Figure 6.2 The Economise part of the CLUE process. MinMax and sensitivity analyses let you apply your efforts to the right parts of a business-viable ML pipeline.**

Figure 6.2 first applies MinMax analysis to determine if your ML pipeline is already producing a viable business result. If not, does the pipeline need further improvement to do so? Or is it, by its nature, incapable of creating such an outcome?

If the pipeline is hopeless, you discard it and try a different pipeline (or a different AI project). Once you know you're working with an adequate ML pipeline for solving

the business problem, you use sensitivity analysis to improve the pipeline repeatedly until you're satisfied with the business result.

All the earlier steps in figure 6.2 require the close cooperation of data science, data engineering and the business team. The team that works on the analysis of the pipeline should initially consist of representatives from all three areas. The goal is to reach a point where a shared understanding about the link between business and technical metrics emerges. The team that manages the ML pipeline should consist of management and engineering leaders (to include data scientists and data engineers). The goal of forming this team is to have an ongoing shared understanding of the technical characteristics of the data science pipeline and to guide deployment of resources based on those characteristics.

**TIP** This approach of analysing and improving the ML pipeline is by necessity iterative. You might perform early iterations of analysis while the proof of concept (POC) is still in progress. If you're using Agile [118,119] or iterative software development methodologies, you could consider analytical work to be part of the first iterations of the project.

## 6.3 MinMax analysis: Do you have the right ML pipeline?

A fundamental analysis you should always perform on any ML pipeline, MinMax analysis answers the question, "What's the best and what's the worst result that an ML pipeline of the given structure can achieve?" This section provides an overview of MinMax analysis.

**NOTE** I use the term *MinMax analysis*, but know that this type of analysis is sometimes also known as *Best Case/Worst Case analysis*. Section 6.6.4 elaborates on this terminology.

MinMax analysis shows in an early stage of the pipeline's life the expected range of results it can achieve. For the question, "Should I continue with this ML pipeline?", the analysis answers with either "yes", "no" or "maybe". Knowing that you have the right ML pipeline so you don't spend a lot of time and money on building the wrong ML pipeline is a beautiful thing.

**DEFINITION** MinMax analysis is a type of analysis that determines if your ML pipeline is already meeting your business goal, if it needs improvement to be able to meet the goal, or if it's incapable of meeting the goal.

MinMax looks at the structure of your ML pipeline and assesses its business viability. For the ML pipeline that's analysed, the Min part answers the question: "If I release the simplest implementation I can make, what would be the business result?" The Max part answers the question: "What's the business result I'd get with the best possible implementation?"

**TIP** Before you try anything, it's always an excellent idea to ask, "What's the best that can happen if I succeed, and, knowing this, is it worth it to even try?" The Max part of MinMax analysis answers that question.

To perform a MinMax analysis, you need an ML pipeline structure and a profit curve. With the profit curve, you have a business metric, which is a way to relate a technical metric to business. You also have the threshold (the minimum level of a business metric that you need to reach). Figure 6.3 shows an overview of MinMax analysis.

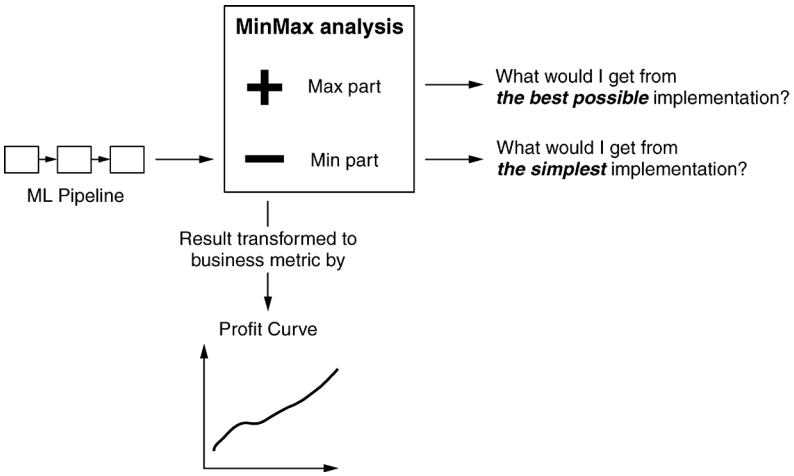


**Figure 6.3 A MinMax analysis consists of the Min part (–) and the Max part (+). It uses a profit curve to transform the technical metric into an expected business result. The result of this analysis determines the viability of your ML pipeline.**

At a high level, MinMax analysis consists of the following:

1. It measures how well your ML pipeline does on the problem you're trying to solve. You measure the ML pipeline using a technical metric.
2. It expresses the output of the ML pipeline in business terms. You use the profit curve to form an *U*nderstanding (*U* of the CLUE) of what the measured result is in business terms
3. It repeats the earlier two steps twice more: once for the Max part and once for the Min part. The Max part uses the best possible implementation at every stage of the ML pipeline. The Min part uses the most straightforward implementation at every stage.

Once you've performed the analysis, you'll know what the most straightforward ML pipeline can achieve in terms of the business. You'll also understand what the best implementation of your current ML pipeline can achieve. If the best isn't enough to make the business viable, you know it's time to discard the ML pipeline.

The following sections show you how to perform a MinMax analysis. But first, I'll present an example ML pipeline that we can use for this analysis and then show you how to interpret the results.

## 6.4 How to interpret MinMax analysis results

MinMax analysis answers the question, "Is it worth continuing with the development of the current ML pipeline?" Therefore, every team leader must know how to interpret the results of a MinMax analysis. Performing this analysis is of interest only to a subset of readers who want to perform the analysis themselves or to better understand the details. Consequently, I'll first show you how to interpret the MinMax analysis' results, and then, in a later section, you'll learn how to perform the analysis.

This section first gives you a concrete scenario of an ML pipeline that solves a real business problem. Then it asks you to make concrete decisions from this scenario: should you continue development of the ML pipeline or not? These examples highlight that the result of the analysis can point not only to the ML pipeline being adequate or inadequate for addressing your business problem, but also to the need to improve it before it can solve the business problem. Finally, once you've seen an example of interpreting the MinMax analysis, it provides a summary of rules for performing MinMax.

> **Understanding the details**
>
> Deciding if your technology is able to solve a business problem is inherently a multidisciplinary problem. You must understand the details of the business problem because decisions in the business can't be made without considering the financial impact and the business and domain rules. First you need a basic understanding of the technical solution. Then you need to decide if that solution is profitable.
>
> As a result, a realistic scenario for a MinMax analysis would be a little bit more complicated than other scenarios in this book. Taking the effort to comprehend the details of the analysis is worth it, though, as the easiest way to save a ton of money on an AI project is to abandon the wrong ML pipeline before it costs you dearly.

### 6.4.1 Scenario: The ML pipeline for a smart parking meter

Let's talk about a concrete business scenario with all of the details of a business problem that needs to be solved, and the financial implications of various decisions your system makes. I'll start by describing the business side of the problem and then show you the outline of the ML pipeline that solves the problem.

Your company makes smart parking meters. The parking meters have a camera, which can be used to take photos of licence plates. Your client is a city. It has many plans and possibilities for these parking meters, but for the time being, it issues automatic citations if someone overstays their visit.

The city is primarily interested in compliance with the parking regulations, not citation revenue. For our scenario, the city has agreed to pay for the initial installation of the meters and give you a yearly bonus for any meter with fewer than 50 overstays per year. The bonus in itself is large enough that your company is perfectly content with that bounty being the sole source of revenue related to smart parking meters for the city.

The economics of the smart parking meter is that you make a profit of USD 3 per citation, and you have to pay the city USD 20 for every incorrect citation that it issues. For the smart parking meter business to be practical for your company, the value threshold is USD 100/year (per meter) or getting the bonus from the city, which is the preferred choice. The best current estimate is that there are at least 300 overstays per year for each parking spot, with the maximum profit per meter at USD 900 if all citations are correctly issued.

**NOTE** As you learned in Chapter 1, you should be careful not to extrapolate past data into future data when the deployment of AI technology might change the reality of that speculation. Would the same number of overstays continue once you started issuing citations? Alternatively, would parking overstays disappear instantly?

To be on the safe side, you decide to assume that the number of overstays would collapse. Instead of using 300/year, you assume there would be at least 51 overstays/year. You're hedged by the city if there are 50 or fewer overstays, so 51 overstays/year is the worst-case scenario for you. Figure 6.4 shows the simple ML pipeline you'll use.

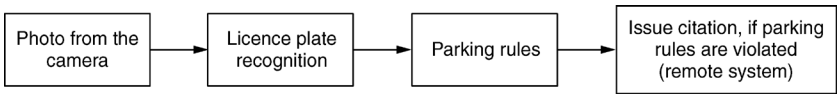


**Figure 6.4 A simple ML pipeline for the automated parking meter that takes a picture of the licence plate, checks if parking is legal and issues a citation if not. The examples in this Chapter use this pipeline for the analysis.**

The ML pipeline shown in figure 6.4 uses a camera and image recognition. However, because we're working with a physical device in the real world, there are a few complications:

- The quality of the camera image depends on the time of day and weather. (Reflections from the sun, light and rain are all factors.)
- Citations are issued using a wireless network, so the parking meter has a wireless modem. For this example, we'll assume that the ticketing stage of the pipeline always works perfectly.[2]

[2] In practice, it's possible that some citations would fail because of communication errors or problems with the citation system. To simplify this example, I'm disregarding those issues.

Parking rules are a bit more complicated than what I initially described. In fact, they're complicated enough that the ML pipeline shown in figure 6.4 needs a rule engine to support those parking rules

> **NOTE** In this Chapter, we're initially covering a basic case of MinMax analysis. In the basic case, it's safe to assume that your profit curve is such that a higher value of the technical metric would always result in a higher value of the business metric. Section 6.5.5 will show you what to do if this assumption isn't met.[3]

In some cases, you can make decisions based only on the Min or the Max part of analysis. In others, you'll need to examine both the Min and the Max part of the analysis before making a decision.

> **NOTE** Before you see the set of hypothetical results of the MinMax analysis, remember that for the parking meter to be profitable, the threshold that you need is at least USD 100 per year from each meter to make it practical. That threshold was based on assuming there would be 51 overstays per year, which is your worst case of overstays and that would not be few enough for the bonus from the city to kick in. (The city pays for no more than 50 violations.)

In the first scenario for this example, you know that with the most straightforward meter you can construct, and the simplest implementation of video recognition of the licence plate, 97% of citations will be correct. Unfortunately, 3% will be wrong. Based on your profit curve, your data scientists tell you that they've completed the Min part of the MinMax analysis and found that the expected profit per meter is USD 117.81 per year (for 51 illegal parking citations per year). Because your threshold was just USD 100 per year, you know that your ML pipeline is viable, and that the AI system would make a profit in an 'as-is' form of the pipeline. You don't need to worry about the Max part of MinMax, because the ML pipeline you have today is already good enough to support your business goals.

### To release or not to release?

If the Min part of your MinMax analysis shows that your current ML pipeline is already producing a value that exceeds your threshold, you might choose to release your product. Alternatively, for various business reasons, you can decide not to release it. This is now a business decision that might require more analysis.

As another consideration, parking meters have parts that can complicate business decisions; for example, once the meters are deployed, you can't easily replace the cameras. However, if you know that you already have the best camera you can get, and you can update the software in the meter remotely, you can treat the meter as a conventional software system. In this case, you should release at once and, if needed, build a better vision recognition system down the road.

[3] For data scientists in the audience – your profit curve is monotonic, but there's no requirement for the relation between the technical metric and the business metric to be linear.

Let's now look at an alternative scenario: someone has decided that instead of putting a camera in the parking meter itself (so that it's just a few inches from the licence plate), they'd reuse a security camera on the roof of the building near the car park to look at all the parked cars. With such a system, you don't even need to install new physical parking meter devices with cameras!

The city mandates that they like the idea and would use it, as it would save them from having to pay for installing new parking meters with cameras. All other parameters, such as the cost of citations and the USD 100 value threshold, stay the same. Your company decides that it would still want to do business with the city under those conditions.

Then you get the images from the existing security camera on the roof. It turns out that the camera takes fuzzy and distorted images only once every few seconds. It's a fisheye lens camera intended for security systems. There are also obstructions that affect the view of the car park, and you now need to map the image from the camera to determine which car is in which parking space.

You ask engineering what they think. Your team performs a MinMax analysis again. This time, they know they're facing a difficult technical challenge and wonder if even the best implementation of the current ML pipeline would be able to solve it.

The team first performs the Max part of the MinMax analysis. The analysis shows that with the images you have, your system will issue 89% correct citations and 11% incorrect ones. With 51 overstays per year, the profit per meter would be USD 23.97 per year. Your ML pipeline isn't going to work.

Panic time! Can engineering construct a different ML pipeline? They try, but none are business practical if you use images from the existing security camera. If you can't get better pictures, you don't have a viable AI project.

> **TIP** Cancelling a project in cases like this isn't a bad thing at all. This example shows you exactly why you should perform the MinMax analysis early in the project. When would you rather talk with your boss about cancelling the project: when you've spent 3% of the budget (time and cost) and can prove it can't work, or when you've spent 105% of the budget and realise it's going nowhere? If you're destined to fail, fail fast and start something more productive.

Your boss is reasonable and understands that the idea of using images from the security camera installed on the roof is a non-starter. The boss makes some phone calls and comes back to you with good news! You'll be able to use much better cameras placed around the car park instead. Multiple cameras will be positioned in such a way that you don't have to worry about obstructions. You ask the team to repeat the MinMax analysis with the same ML pipeline, but under the assumption that you'd be using better cameras.

This time, the Max part of MinMax shows that the best that's possible with such a system is 98% correct citations, with 2% incorrect. And at 51 overstays per year, the profit per meter would be USD 129.54 per year. (Remember, the Max part provides a result corresponding to the best implementation of the ML pipeline.) The Min

part, however, analysing the simplest pipeline you can construct, shows that you'd have only 91% correct citations, with a profit of USD 47.43 per year. The simplest implementation of the ML pipeline isn't going to work, but maybe you can improve it.

The project should continue, but there's a question of how difficult it will be to improve the ML pipeline. We'll discuss that next.

### 6.4.2 What if your ML pipeline needs improvement?

The results of the MinMax analysis are often conclusive about the business value of your ML pipeline. In the previous section, you saw an example where the Min analysis showed that the pipeline was already providing viable business results. You also saw an example in which the ML pipeline was inadequate to meet the business goal. However, sometimes results are inconclusive: your ML pipeline needs improvement to meet your business goal.

This section describes the last of these situations. Figure 6.5 shows an overview of applying MinMax analysis for the case in which an ML pipeline needs improvement.

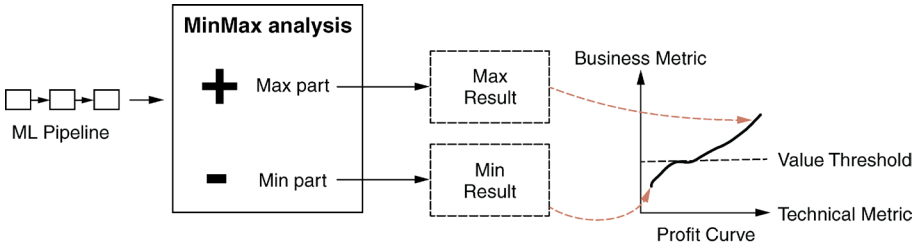


**Figure 6.5 Here, the Min analysis doesn't reach the value threshold, but the Max analysis exceeds it. It might be possible to improve this pipeline enough to make it business-viable.**

In figure 6.5, the result of the analysis shows that with the ML pipeline you're using, the Min implementation isn't good enough, but the Max analysis shows that the current structure of the ML pipeline can be improved to provide an acceptable business result. In this situation, you can say that the ML pipeline passes the Max analysis and fails the Min analysis.

An ML pipeline that makes you money with a minimum effort on your part is a good thing to have. An ML pipeline that doesn't make you money even when you use the best techniques possible in each of its stages is a good thing to abandon early. However, what happens when you're in between and you know that the Max is good enough, but your Min isn't? You need to improve your ML pipeline before you can release your AI product, so you'll need to perform a sensitivity analysis on it (as detailed in Chapter 7).

### 6.4.3 Rules for interpreting the results of MinMax analysis

In the previous section, you learned how to interpret the results of the MinMax analysis. If we were to summarise how we made those decisions, we could form a set of rules for interpreting the results of the MinMax analysis. This section summarises those rules. Let's first agree on some terminology:

- We'd say that the Min part of the MinMax analysis *passed* if the most minimal implementation of the ML pipeline you can make already has a business value that exceeds the value threshold on the profit curve. The minimal ML pipeline is business-viable.
- We'd say that the Max part of the MinMax analysis *passed* if the best possible implementation of the ML pipeline will have enough business value to exceed the value threshold on the profit curve. The best possible implementation of your current ML pipeline will be business-viable.

Different combinations of the results of the Min and Max analysis have different business meaning. Table 6.1 summarises the possible results of a MinMax analysis and their business impact.

**Table 6.1 Summary of the possible results of a MinMax analysis. Each of the results has direct implications for your business.**

| Min result/Max result | Max passed | Max failed |
|---|---|---|
| | The ML pipeline is business-viable. | This combination can't happen. |
| | The ML pipeline needs improvement to be business-viable. | The current ML pipeline isn't suitable for solving the business problem. |

## 6.5 How to perform an analysis of the ML pipeline

Now that you know how to *interpret* the results of the MinMax analysis, let's talk about how to perform the analysis. How did we get the numbers from the previous section that are the result of the MinMax analysis? By analysing an ML pipeline. This section shows you how to perform such an analysis.

> **NOTE** If you're a manager without an engineering background, you might want to skim through the rest of the description of MinMax analysis to get a basic understanding of what your team will do during it. If you're a manager with an engineering background, you should be able to understand (or even perform) this analysis.

The first step in analysing an ML pipeline is to make sure you have all the prerequisites. You need a logical diagram of your proposed ML pipeline. You also need the technical and business metrics you're using on the project and the profit curve.

You analyse an ML pipeline by running data through it and then using a profit curve to measure the result in business terms. The general process for the analysis (figure 6.6) is the same for both the Min and Max part of the MinMax analysis, and even applies to the sensitivity analysis of the ML pipeline.

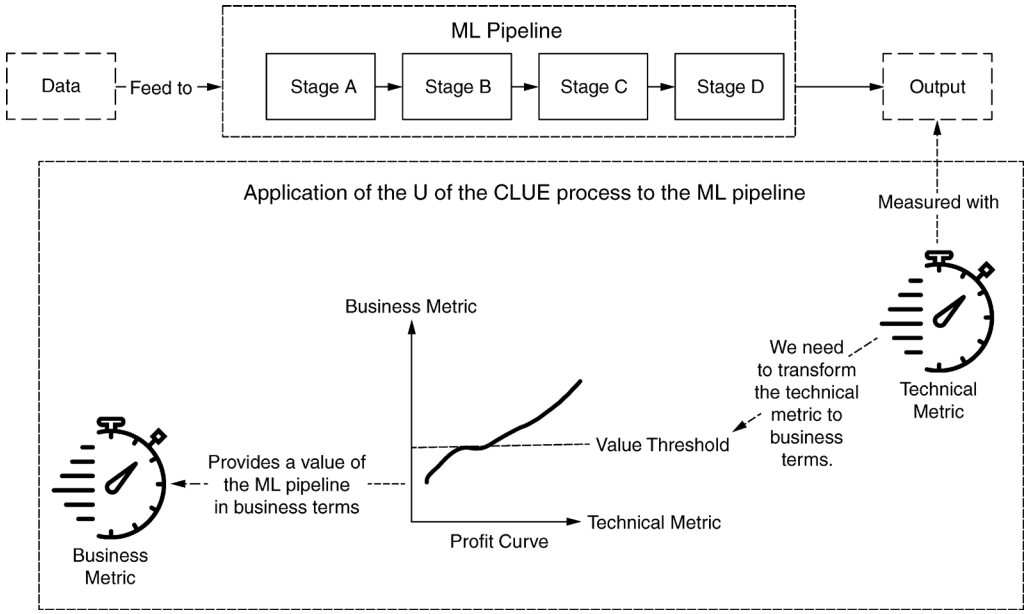


**Figure 6.6 Analysis of an ML pipeline. This analysis tells you what the current ML pipeline can achieve for the business. Both the Min and Max part of the MinMax analysis use the same process.**

The analysis shown in figure 6.6 measures the results of the ML pipeline in business terms. That's done by processing real data with your current ML pipeline, measuring the output using technical metrics and then transforming that value into the business metric. Another way to think about this analysis is that you're feeding the ML pipeline with real data and are then applying the *U*nderstand part of the CLUE process to measure the result.

The same process from figure 6.6 would apply to both the Min and Max parts of the MinMax analysis. What changes is not how you analyse the ML pipeline, but the implementation of the stages of the ML pipeline.

Now that you understand the overall process for analysing an ML pipeline, let's see how to perform the actual analysis. Section 6.5.1 shows you how to perform the Min part of MinMax analysis, and section 6.5.2 shows you how to perform the Max part. Section 6.5.3 discusses how to account for trend estimates and safety factors during the performance of the MinMax analysis. Section 6.5.4 introduces different types of

profit curves you may encounter in practice. Finally, section 6.5.5 shows you how to apply MinMax analysis to profit curves that have a complex shape.

### 6.5.1 Performing the Min part of MinMax analysis

To perform the Min part of the analysis, you should first construct the simplest implementation of your ML pipeline. The goal is to build an ML pipeline implementation that you can test with real data.

How much effort should you put into preparing the ML pipeline for the Min part of the MinMax analysis? The effort should be insignificant compared to the size of the whole project. The general rule is that it should take no more than a few people a couple of weeks of effort, regardless of the project size, costing not more than 5% of the total project budget. If you're a small company, it can mean the best that one person can do in a few days.

The liberal use of duct tape is the key. Use a pipeline implementation that you can construct fast. If commercial off-the-shelf (COTS) products can help you, use them. Ask yourself, can you use someone else's product at any stage of the pipeline to implement needed functionality? You're looking to get the system working with the minimum effort needed so you can see what business result is possible with such a system.

Once the analysis is complete, it's time to apply techniques that you learned in section 6.4 and interpret the results of the analysis. If you already have an ML pipeline that's giving you results that exceed the threshold value of the business metric, your pipeline is business-viable as-is.

> **NOTE** What happens if the amount of effort you need to spend to construct even the simplest ML pipeline is large? Without that simple implementation, you can't perform the Min part of the analysis. Therefore, the result of your Min analysis is zero – *the current* implementation of your ML pipeline has zero business value. Without making improvements, you have nothing.

### 6.5.2 Performing the Max part of MinMax analysis

If what you have today isn't good enough to release, you clearly need to improve your AI system before it's business-viable. It's time to perform the Max part of the MinMax analysis and determine if your ML pipeline is something you can improve enough to achieve your business goal. This section shows you how to perform this part of the analysis.

For the Max part of MinMax analysis, the idea is that for an ML pipeline with the given structure, the best possible result for the pipeline as a whole is the one in which every single one of its stages has the best possible implementation. Look at stage B of the ML pipeline shown in figure 6.6. Suppose that stage is given an image and handles visual recognition of digits on the image. For the Max analysis, you'd look at what the best result is that anyone else has achieved recognising digits and use that as a proxy for what the best possible result is that your stage B could accomplish if you were to put considerable effort into it.

### Best-so-far is an upper limit of reasonable expectation

For the sake of argument, let's assume that the best result anyone has got so far for some stage of the ML pipeline (let's call it stage B) is 99.22% accuracy. Suppose you construct an ML pipeline using that stage, and when you perform the Max part of the MinMax analysis, you find that your ML pipeline isn't business-viable. However, if you could improve the accuracy of stage B by just 0.5% (to an accuracy of 99.72%), you'd have a business-viable ML pipeline. Should you assume that your team would be able to achieve that 99.72% accuracy?

For most teams in the industry, it's risky to bet that a team can improve on the best published result achieved in the field of AI so far. That means that any ML pipeline that can't produce acceptable business results when every single stage is using the best possible implementation known is, by its structure, unsuitable for solving your business problem. Such an ML pipeline has failed the Max part of the MinMax analysis. If you need to exceed the best historical result to get even a minimally viable product, it's well past time to abandon such an ML pipeline.

The Max part of the analysis is especially crucial in the early stage of building a new business or product. At this stage, you're estimating the total costs of your project and your pipeline, and, chances are, you're overlooking some costs. If your ML pipeline fails the Max part of the MinMax analysis even when your total costs may be underestimated, that ML pipeline surely can't support a viable AI product.

The most important question during the Max analysis of the ML pipeline is, "What is a reasonable proxy for the best possible result in such a stage?" Let's look at some of the ways to find a good proxy.

The Max part of MinMax analysis has its root in competitive benchmarking [1,120]. Competitive benchmarking means that you're looking for a proxy, someone who has a similar problem and you measure how well they solved it. Here are some sources we'll investigate in more detail that you can use as a proxy:

- The best results achieved so far in business, industry or academia on a problem similar to one you're trying to solve
- A person performing the task you need your AI to perform

#### WHAT'S THE BEST RESULT FOR PROBLEMS LIKE YOURS IN ACADEMIA OR INDUSTRY?

What's the best published result in academia or business for a task like yours? What you're looking for is an organisation that had the same problem as the one you have and how well they solved it.

When looking at the industry, you're looking for not only the COTS products you can buy, but also what's best that other companies (even if they aren't your direct competitors) were able to achieve when tackling the problem you're facing. When you're looking at academic papers, you're looking at the best published result. In both cases, the key is that you're trying to find what's the best someone else has achieved when solving the problem *that looks as much as possible like yours.*

**TIP** If you're considering a COTS product for some stage in the pipeline, ask the supplier what the best result is that they've seen achieved with their product for a problem such as yours. A supplier that's confident that they're supplying a tangible business value for a business case like yours should recognise that question as an opportunity to distinguish themselves.

The idea of using academia or industry as a proxy is that if the world's best experts were unable to achieve more than 80% accuracy on some task, it's reasonable to assume that you won't do better than 80%.

As for which proxy to choose, you're typically better off using a proxy that's more like your problem – proven industry use is always a stronger proxy than an academic paper. On the other hand, an academic paper describing a situation that's precisely like yours might be a better proxy than industry use on a less-related problem.

### How similar is your situation?

You must work carefully to ensure that the problem you choose to use as a proxy is similar to your situation. The proxy must be closely related, and you must understand under which conditions the advertised result was achieved. Pay special attention to any simplifications that the authors of an academic paper might have introduced.

B. Hu et al. present an example in their paper on time series classification [121]. Accelerometer data can be used to recognise the motions and gestures of an actor. The question here is, how do you know when one motion starts and another one ends? The answer to that particular academic community, at one point in time, was that they'd use what's technically known as *pre-segmentation*.

That community was interested in classifying gestures under the condition that it was already known when the gesture started and stopped. That was known because the actor was asked to perform the gestures on cue. *To get more precise pre-segmentation, some actors were even using a metronome!*

The only problem is that recognising gestures if accelerometer data is pre-segmented with the help of the metronome is much easier than understanding accelerometer data coming from a smart watch (such as an Apple Watch). If you were using a former community's results as a proxy for what's possible with a smart watch accelerometer, you'd stop wondering why you were unable to get gesture recognition accuracy anywhere close to what that academic community was able to achieve.

What if during the Max analysis your team misses the absolute best result published? For example, it was in some obscure scientific paper. That doesn't matter; the finding of Max that your team made is still considered the Max for your organisation, and that Max isn't affected by the existence of some obscure paper. What you're looking for during Max analysis isn't the absolute Max known to humankind. You're looking for a practical or industrial Max – the one your actual team might be able to get when they try to implement the given stage of the pipeline. Your team can't implement algorithms from papers they don't know about.

#### HOW WELL WOULD A HUMAN DO SOLVING YOUR PROBLEM?

What if you have no example whatsoever that's like your problem? If no one has solved a problem like yours before, it's dangerous to assume that you can make an AI product that would do better than a human. Instead, see what a person can do when given a small subset of the data. Use that person as your proxy.

Take a human and show them the same data that the stage of the pipeline sees. Ask them to play the role of the pipeline stage, then measure what the human-powered stage achieves. That's your estimate of the maximum that an AI algorithm would be able to do.

> **NOTE** Although it might be possible to get a *better-than-human* effect on some tasks, at the time of this writing, such situations are infrequent, often newsworthy and, even when achieved, usually produced by teams that already consist of people who are among the best AI researchers in the world. It's much more common that the results from your AI algorithm would be worse than that which a person could achieve.

### 6.5.3 Estimates and safety factors in MinMax analysis

The nature of the Max part of the MinMax analysis is that you're estimating what can be achieved if every stage of your ML pipeline has the best possible implementation. This section addresses those questions that often present themselves when you're estimating what you can achieve. For example:

- When you're performing a MinMax analysis and your project is to be released in 18 months, should you account for trends in the improvement of hardware or AI algorithms?
- How should you use expert opinions during a MinMax analysis?
- Should you add any safety factors to your results?

#### USING TREND ESTIMATES IN A MINMAX ANALYSIS

Sometimes you might look at a trend in improvement and estimate what the continuation of such a trend would mean at the time you deploy your project. For example, the cost of data storage is declining, and if your project is shipping in two years, you can account for a decline in data storage costs and conclude that you'd be able to afford to store more data than you can today.

You can also apply trends to the improvement in AI algorithms. In some cases, there might be a clear trend that AI algorithms are getting better. For example, right now, AI is getting better in the recognition of images and video streams, and it's doing that rapidly. Estimating trends is especially tempting to do if the results of your Max analysis are failing with a small margin from the business viability threshold you're trying to reach. For example, you need accuracy of 96%, and currently the best you have is 95.5%, but you think it will be 96% in two years.

I'm not a fan of estimating trends in your early AI efforts; I tell most of my clients not to attempt to do so. To predict trends in some field successfully, you need more expertise than you're likely to possess at the time you're starting early efforts in that

field. I'm talking about not only technical knowledge, but also knowledge of how good your organisation is at assimilating new AI technology. Finally, for trends to be significant, the project typically needs to be long enough that it would be longer than your first AI project should take.

#### ESTIMATES BASED ON EXPERT OPINION

It's always a good idea to ask an expert for their estimate of the Max result that you can get in some stage of the pipeline. Such experts could be consultants or academics. They might also point out not only what the best possible result is in some stage of the pipeline, but how to achieve it.

Another advantage of an expert opinion is that they might know an area well enough to tell you not only what the current best result is, but what the trends are in the particular research area. Is this an area that's rapidly improving its capabilities (such as the image recognition community)? How much training is necessary for your team to achieve those results?

Be careful to contextualise expert advice. Did the expert answer the single question in isolation, and does the expert understand your particular situation well?

Experts are expensive and busy, and it's tempting to try to save money by asking them a straightforward question that your team constructed. The problem is that now your team has to handle contextualising both the question and the answer in an area where they have less expertise than the expert.

Answers you get in such situations may also give you a false sense of security: you're likely to believe that the answer is correct because it's coming from a recognised expert. However, the contextualisation part (done by lesser experts) could be an essential part of the problem. Asking the right questions is difficult, and a lot of what makes an expert an expert is that they know which questions to ask.

You encountered the problem of contextualising the question before in Chapter 3, when we were formulating research questions for solving business problems. Just like the executive and the data scientist needed to talk in Chapter 3, you'll need to speak with an expert to make sure that the question (asked correctly) reflects your needs.

I believe that saving money on expert advice is a 'penny wise, pound foolish' approach. Budget what it takes to get enough time with the expert to explain the problem you're facing so that expert can contextualise the answer. If you have the right expert, you can explain to them the specifics of your business situation in a few hours.

Finally, be careful to understand what the form of the answer is that you get from an expert. Is it fact, expert opinion on the situation, or just an estimate? You already know that I'm not a fan of trend estimates on initial AI projects, and my opinion doesn't change much just because the estimator is a technical expert in the field.

**WARNING** If you get an estimate, be aware that most people (and most experts) aren't particularly accurate estimators. It's worth reading D. W. Hubbard's book [75] for data showing that most people are poor estimators, a discussion on why humans are poor estimators and ways for people to become more accurate.

#### SHOULD YOU ADD SAFETY FACTORS DURING YOUR ANALYSIS?

Sometimes academic results can't be replicated fully in the industry setting. You might also be sceptical that your team can create an AI solution that will produce results on some task that are close to what a human working on the same task can do.

In such a situation, you might want to introduce a safety factor, such as, "We assume that we can reach only 50% of the best published result." Of course, it's not clear what this safety factor should be – why have we chosen 50% as a safety factor and not 80%? A safety factor is a good idea when you understand the problem well enough that you can describe where the uncertainty is coming from and how much uncertainty there is. However, when you don't know how much uncertainty there is, a safety factor is just a guess.

> **WARNING** Be careful when you encounter safety factors that are round numbers (like 2, 3, 10 or 50%). They look too suspiciously round to be the result of a technical analysis of the problem, and those numbers might be just a guess.

### 6.5.4 Categories of profit curves

In the examples of profit curve analysis presented thus far, we've assumed that the profit curve is monotonic – that when technical metrics increase, the business metric (such as profit) also increases. This is the most common real-life situation, but this section will demonstrate what to do when the profit curve and technical metrics have more *complicated* relationships.

> **NOTE** This section is of interest to both the general reader and the mathematically inclined reader. To facilitate the broadest audience, I'll use everyday terminology to describe some of the concepts more simply. I ask for patience from my mathematically inclined readers – you already know the basic concepts I'm describing, as well as the corner cases, and can add the underlying mathematical rigour yourself.

Profit curves can have different shapes. We're particularly interested in the four categories in figure 6.7. Let's discuss the categories of profit curves shown in figure 6.7:

- In a linear profit curve, the relationship between the business metric and the technical metric is a straight line.
- In the monotonic profit curve, when the technical metric increases, so does the business metric. However, the function describing the relationship between the technical metric and the business metric is not a straight line – the profit curve can take many shapes. Every linear profit curve is a monotonic curve, but not vice versa.
- The next more complicated profit curve is the non-monotonic profit curve. This curve has segments in which the business metric increases when the technical metric does and other segments in which the business metric decreases when the technical metric increases. Graph (c) shows one type of non-monotonic curve you may encounter in practice.

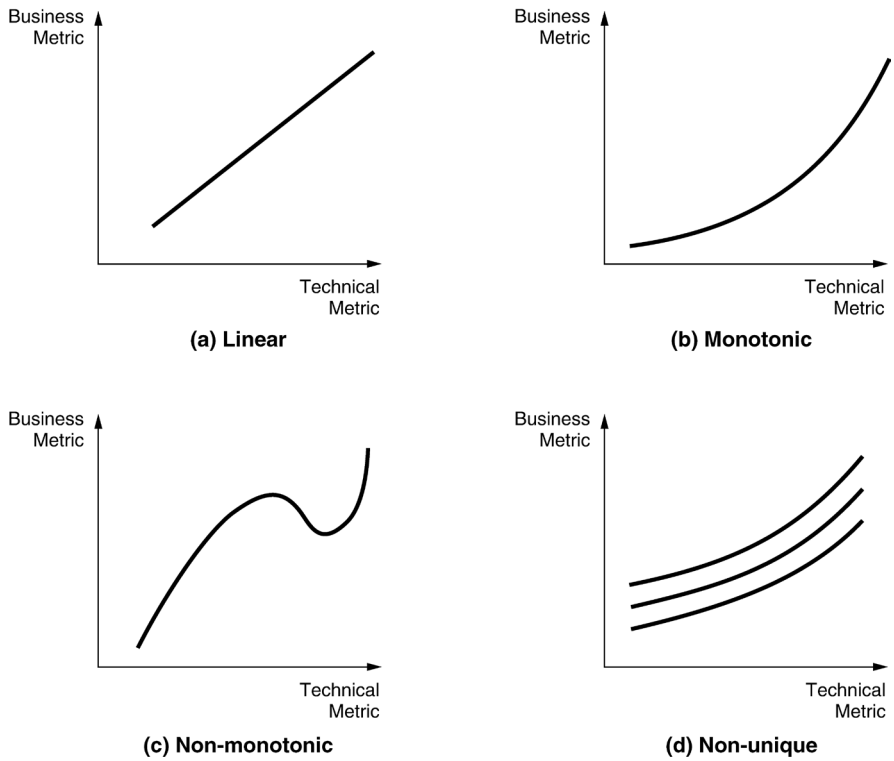


**Figure 6.7 Various types of profit curves. The linear and monotonic profit curves are simpler to optimise than the non-monotonic profit curve. Avoid non-unique profit curves.**

- The final example is a profit curve that's non-unique (ambiguous) – there's no unique relationship between a given technical metric and revenue. You'll encounter this type of profit curve when a technical metric fails to measure business considerations vital to you.

*You know the shape of the profit curve for your project; your team constructed it.*[4] Let's give some concrete examples of when you'll encounter the kinds of profit curves shown in figure 6.7:

- You'll often encounter a linear profit curve (graph (a)) when a simple and direct relationship exists between your technical metric and your business metric. For example, a linear profit curve would happen when you have a rule of the form, "Our yearly cost is given using the formula *USD 100K* RMSE*."

[4] While a profit curve may in some cases be experimentally derived, that's an advanced technical topic that isn't practical to cover in this book because of the intended target audience and space. Some of the technical topics relevant to experimental derivation of profit curves include design of experiments, response surface analysis and Bayesian optimisation.

- You may encounter a monotonic (but non-linear) profit curve when you're predicting some value – the better your prediction, the more valuable it is. Graph (b) shows an exponential relationship between a technical metric and a business metric – where a small increase in prediction accuracy will result in a massive increase in profits. You'll face that situation often in the financial markets.[5]
- Another example of a non-monotonic curve is in robotics, where the profit curve will look like the curve shown in graph (c). In robotics, there's a concept of *the uncanny valley* [122]. Consider two robots, the first with a doll-like face, and the second robot with facial features (and facial movement) similar to a human's, but not a perfect facsimile. We tend to think that the second robot with more sophisticated facial modelling would be more popular, but the opposite may be the case [122]. Some people have an instinctive, negative reaction to the second robot.
- Another example of a non-monotonic curve is of automated systems that require close human supervision. Think of a security system that uses AI to highlight suspicious activity. Still another is when workers are supervising automated machines and robots. Transportation systems (such as planes) can also be subject to the same phenomena [123].

  When (for legal or practical reasons) you have to have a human fully dedicated to supervising your system (and being bored for a significant amount of time), you may see shapes similar to graph (c). The combined system of AI and a human will perform *worse* when an imperfect AI makes occasional errors (but still makes them) than when the AI makes more frequent errors. The reason lies not in the performance of the AI system, but in the performance of the human supervising the operation – humans tend to get bored and inattentive and therefore may be slow to correct problems when they finally do occur. Or they may simply be out of practice in how to react to the errors of the AI system.
- Yet another example of a non-monotonic profit curve is in some situations where the business metric is profit. Sometimes, reaching a higher value in the technical metric may be expensive and require additional capital expenditure. Suppose that further improvement of your technical metric would require you to buy access to costly additional data sources. This creates a 'dip' in your profit curve where the further improvement of your technical metric is possible only after you buy costly additional data (and, therefore, reduce profit at that point).

As a general rule, monotonic profit curves are simpler to deal with, and if you have a choice, you should prefer them. If you have two *equally good* technical metric/business metric combinations to choose from, of which one is monotonic and the other is not, choose the monotonic combination.

---

[5] Note that the reason for an exponential relationship may be that large improvement may be difficult, or even be believed to be impossible. It's likely that because of competition, once you show that the improvement is possible, your competition will try to catch on and the shape of the curve may quickly change. Chapter 7 shows you how to operate with a profit curve that changes over time.

> **WARNING** *Never make business decisions based only on a technical metric that you can't relate to a business consideration you care about!* Otherwise, you're deciding to maximise the number (technical metric), not to optimise business outcomes.

In your early AI projects, or while your AI team is still gaining experience, you should prefer use cases that have simple profit curves. That's for both technical reasons (monotonic profit curves are simpler to analyse) and business reasons.

> **NOTE** A complex profit curve may be an indication of a weak business case for AI, in which monetisation isn't straightforward. But it can also be an indication of a business case that's so strong that it's worth your while to perform a detailed analysis of the profit curve. Or it can be merely a technical coincidence – as in "That's just what the relationship between these particular business and technical metrics is." I don't start with any preconceived notion, but when I see a non-monotonic profit curve, I always ask myself, "Why does the profit curve have this shape?"

Sometimes you don't have a choice – the only technical metric/business metric combination that works for you is non-monotonic. The techniques shown in the next section will help your data science team address that situation.

### 6.5.5 Dealing with complex profit curves

Now let's talk about the details needed to construct the more complex profit curves. This section describes the technical aspects of dealing with non-monotonic and non-unique profit curves.

> **NOTE** I assume in this section that the reader is already familiar with confusion matrices and F-scores in the context of natural language processing (NLP). You can find more information in Leon Derczynski's paper [124].

Let's first deal with how to recognise a non-unique profit curve. A non-unique profit curve happens when no unique mathematical relation exists between the business metric and the technical metric that you're using. An example from the legal field is *e-discovery,* in which AI is used to help to save on the cost of lawyers reviewing documents. We can use AI to check a large number of documents. If an AI system can *reliably* flag the text that's unrelated to the litigation, it can produce dramatic time savings for lawyers and cost savings for the litigant.

Suppose now that you're working for a law firm. Your business question is, "Can you estimate the maximum amount of money that AI can save me during the discovery phase of litigation?"[6] Your savings in e-discovery are proportional to the percentage of the documents that AI correctly classifies as "unrelated to the lawsuit" – the true negatives.

AI systems analysing documents are part of the broader field of NLP. One metric that the NLP community typically uses is the F-score [124]. Unfortunately for our business

[6] To keep this example simple and on-point, let's assume that this is the only question you care about and disregard the costs of AI making a legal mistake. Of course, if enough cost savings are coming from this first question, and the legal firm decides to explore this AI system, further business questions will be asked.

case, the F-score doesn't account for true negatives! It's quite possible for two different AI systems, with a widely varied number of true negatives, to have the same F-score! In our business case, that means that there's no unique relationship between the F-score and the amount of savings AI can provide. The same F-score may save 10% of lawyer time or 80%! You can't use this F-score to construct a 'cost-saving/F-score' profit curve. Although you can use the F-score to measure other characteristics of this system, it's certainly not a good technical metric for a profit curve in which the business metric is cost savings.

> **NOTE** If this is the case, why do people use F-score at all? Because the F-score makes sense in many areas of information retrieval, *but not in our particular business case.* F-score is often used in the context of NLP [124], so if you're debating which *technical metrics* to use, it's a reasonable starting point. The broader teaching point is that just because a certain technical metric is widely used, doesn't automatically make it a useful metric for your profit curve.

Now, let's explore some non-monotonic profit curves. You must perform MinMax analysis of the non-monotonic profit curve differently than the analysis of the monotonic curve. If your curve is non-monotonic, then the MinMax analysis approach you should undertake is shown in figure 6.8.

Remember that the term *Min* in the MinMax analysis refers to the minimum configuration of your ML pipeline (every stage has a simple, minimal implementation), not to the minimum value of the business metric. You always want to get the best value of the business metric you can get with your ML pipeline! Therefore, when performing

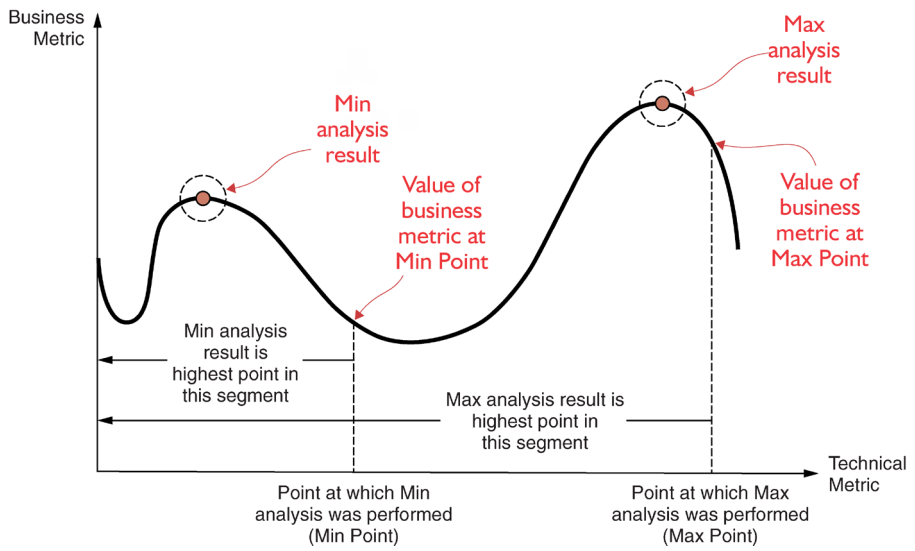


**Figure 6.8 MinMax analysis of a non-monotonic profit curve. The Min analysis result is the best value of the business metric on the segment [0, Min Point]. The Max analysis result is the highest value of the business metric on the segment [0, Max Point].**

MinMax analysis in figure 6.8, the result of your Min analysis is *the best value you've seen on the whole interval* [0, Min Point] *between the start of the curve and the point at which you've performed the Min analysis.* Similar logic applies when performing the Max part of the MinMax analysis on the whole interval [0, Max Point].

Once you get a handle on MinMax analysis overall, MinMax analysis on the non-monotonic profit curve isn't difficult to perform. However, it's more labour intensive than working with a monotonic profit curve! And the work doesn't stop with MinMax analysis – a project with a non-monotonic profit curve typically has special considerations. For example, if the profit curve has the shape in figure 6.7 part (c), because a human supervisor may become bored and inattentive, how are you going to address that boredom?

### When mental maths doesn't work

The more you climb the ladder of complexity in the profit curves, the more difficult it is to comprehend relationships. Simple non-linearity in a profit curve already makes mental maths impractical. It's much worse when the curve is non-monotonic!

When you fail to present a well-defined profit curve, you're forcing meeting participants on your team to perform mental maths to figure out what the technical metric means in business terms! I know I can't concentrate on the central topic of most meetings while simultaneously performing the mental maths required for non-linear, non-monotonic profit curves as a 'side activity'. I suspect that many people stuck in such a position will skip the mental maths altogether and settle for "Let's just improve the technical metric." Any non-linearity in the relationship between the technical metric and the business metric then gets ignored, or at best *approximated.*

That means that any AI project that fails to construct a profit curve, but happens to have a non-monotonic profit curve, is optimising for the wrong thing! That's especially unfortunate when the precise optimisation of the technical metric was a costly activity, which was then followed by an approximation!

I heard a joke about the series of successive approximation, back when I was in my first year of college. The joke was: "Engineering is about using a micrometer screw gauge to take a measurement, then marking where to cut with a piece of chalk, and finally using an axe to cut at the marked spot!"

## 6.6 FAQs about MinMax analysis

When you understand the basics of performing MinMax analysis, we need to address some practical questions about performing that analysis. This section presents answers to the following common questions:

- Should MinMax be the first analysis of the ML pipeline when a more complex analysis might provide more precise results?
- Should I perform the Min or the Max part of the MinMax analysis first?
- Is a MinMax analysis something that only big companies can afford to do?
- Why do you use the term MinMax analysis? Why don't you call it Best Case/Worst Case analysis?

### 6.6.1 Should MinMax be the first analysis of the ML pipeline?

Let's start with the question of whether MinMax should be the first analysis of the ML pipeline. If you're familiar with the fields of systems engineering and industrial process control, you might be familiar with the several types of analysis that could be suitable when analysing the ML pipeline. So why use MinMax analysis instead, when a more complex analysis might provide more precise results?

I'll discuss some alternative types of ML pipeline analysis in Chapter 7. Those types of analysis might be more potent than MinMax is, as described here. However, MinMax is simple to learn, is cheap to perform and provides a reasonable estimate of the suitability of your pipeline. It's a good trade-off between complexity of analysis and precision of results.

> **WARNING** There's no point in performing an analysis of an ML pipeline if that analysis is overly complex to learn and expensive to implement, regardless of how precise the results of such an analysis might be. Analysis that perfectly predicts the future is worth something only if its results will be available before the future has already arrived.

### 6.6.2 Which analysis should you perform first? Min or Max?

Let's talk about the order in which you should perform a MinMax analysis. Do you schedule the Min or Max part of MinMax analysis first? These parts are independent, so for the validity of the analysis, the order of the Min and Max parts of the analysis doesn't matter much. When I implement MinMax, I use the process shown in figure 6.9 to account for factors such as the difficulty of Min and Max parts and prior opinions about expected outcomes.

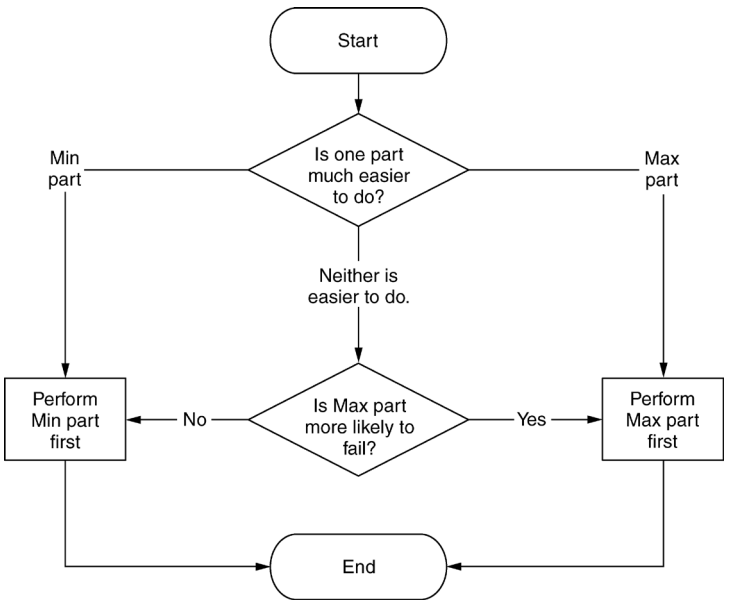


**Figure 6.9 The order in which you should perform a MinMax analysis on the ML pipeline. Perform the part of the analysis that can provide conclusive answers first.**

Each one of the Min and Max parts can be conclusive in itself regarding whether you're using the right pipeline. The conclusiveness of either part of MinMax analysis helps you to avoid the need to perform the other part. If the Max analysis is easy to do and it fails, you don't need to perform the Min analysis. Figure 6.9 shows scheduling the parts of a MinMax analysis so that you start with the more straightforward component of the analysis, which also provides conclusive results.

### 6.6.3 Should a small company or small team skip the MinMax analysis?

Another question is, "Shouldn't small companies minimise process overhead by working on AI algorithms and the ML pipeline itself and then see what happens when they deliver it?" Yes, it's better to skip analysis of the ML pipeline. if you can be certain you're working on the right ML pipeline! The problem is that in the absence of the MinMax (or equivalent) analysis, you *can't* be certain that you're working on the right ML pipeline. Because smaller companies and teams have less money and resources to recover from mistakes, *MinMax analysis is even more critical for small teams.*

When you're constructing an ML pipeline for an AI project, you're working in a new area that inherently has significant risks and for which best practices are still emerging. Very few people have done enough AI projects to be able to intuitively construct the right ML pipeline on the first try, and for many problems, the number of ML pipelines that wouldn't work overwhelms the number of ML pipelines that would work.

So how do you know that you're working on the right ML pipeline? You can take a risk, but if you're wrong, then you'll finish with a flawed ML pipeline into which you've put much effort and that is now ossifying.

> **TIP** Knowing if you have an adequate ML pipeline is even more critical in the context of the small company! A proper MinMax analysis is the most important technical step of an AI project, and choosing to skip it is taking a risk, not streamlining the process.

A similar argument applies to sensitivity analysis: How do you know which pipeline stage you should improve? It's vital to use sensitivity analysis in companies of all sizes. The next Chapter elaborates on this.

### 6.6.4 Why do you use the term MinMax analysis?

Multiple terms are used in business and industry to refer to MinMax analysis. It's also known as Best Case/Worst Case analysis. These terms are often heard in ML circles.[7] However, while working in the industry, I've found that the Best Case/Worst Case terminology is problematic because it's often unclear from which point of view (business or engineering) the best case is to be measured.

---

[7] As an example of use of Best Case/Worst Case in the context of ML algorithms, the best case result for an ML algorithm is discussed in [112].

Look at the ML pipeline we've analysed in this Chapter. What is the best case, from the business point of view? That the most straightforward technical implementation of ML pipeline would solve the business problem. That's the ML pipeline we used for the Min part of the MinMax pipeline. However, from the engineering point of view, you could argue that the Min result is the worst case. I've known many engineers and statisticians who argue precisely that. The ML pipeline you have today establishes the lowest limit (worst case) of how well the AI of your product would behave if you released your product today.

All of this discussion is of keen interest to some engineers, and they have good reasons for that: imprecise terminology is a root cause of a lot of the problems in their profession. While I empathise with the argument, from a pragmatic point of view, I prefer to use the term MinMax analysis.

## 6.7 Exercises

The following exercises help you get a better understanding of the concepts introduced in this Chapter. By its nature, interpretation of the results of a MinMax analysis straightforward. Analysing the ML pipeline is a *technical and business skill*, so it's time to form a team consisting of a business specialist and an engineer and do some of these exercises together. All of the exercises in this Chapter use the ML pipeline given in figure 6.10.

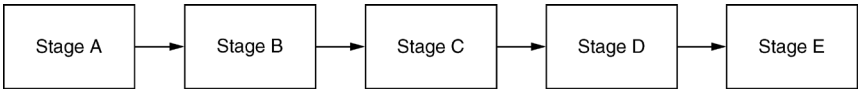


**Figure 6.10 An example of an ML pipeline. We'll use this pipeline for the exercises in this Chapter.**

You'll also need to refer to the information provided earlier in table 6.1 (which is repeated here for your convenience as table 6.2).

**Table 6.2 Summary of the possible results of MinMax analysis**

| Min result/Max result | Max passed | Max failed |
|---|---|---|
| | The ML pipeline is business-viable. | This combination can't happen. |
| | The ML pipeline needs improvement to be business-viable. | The current ML pipeline isn't suitable for solving the business problem. |

Answer the following questions.

**Question 1:** Note that in table 6.2, you don't have any guidance for the situation in which the Min part of the MinMax has passed, but the Max part of the MinMax failed. Explain why this is the case.

**Question 2:** For the ML pipeline in figure 6.10, assume that the value threshold at which the project becomes business-viable is USD 1 million. Determine whether the pipeline is worth pursuing if the results of the MinMax analysis are as follows:

- **Scenario 1:** The Min part is USD 2.3 million, and the Max part is USD 23 million.
- **Scenario 2:** The Min part is USD 500 K, and the Max part is USD 1 million.
- **Scenario 3:** The Min part is USD 500 K, and the Max part is USD 2 million.
- **Scenario 4:** The Min part is USD 1.1 million, and the Max part is USD 900 K.
- **Scenario 5:** The Min part is USD 500 K, and the Max part is USD 900 K.

**Question 3:** If you're a data scientist or technical manager, take a technical problem of your choice and construct an ML pipeline for it. Perform the Max part of the MinMax analysis for it.

**Question 4:** If you're a data scientist or technical manager, look at the examples given in section 6.4.1 and perform a MinMax analysis as described in that section. Determine where the dollar amount given in that section comes from. Hint: a profit curve was constructed from the confusion matrix of the classifier.

**Question 5:** How would you classify the use of AI in the context of saving litigation costs during the e-discovery process described in section 6.5.5? Use the taxonomy of AI uses introduced in section 2.5. It's shown in figure 2.5, duplicated here as figure 6.11, which summarises the taxonomies discussed in that section.

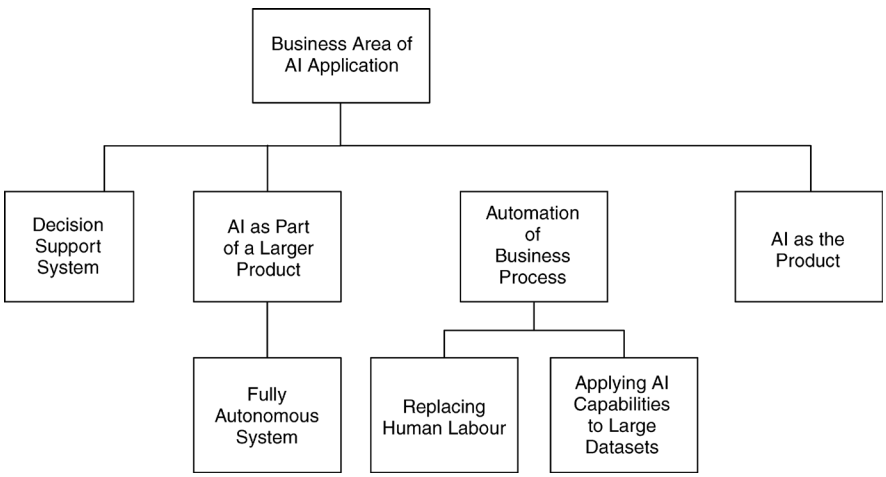


**Figure 6.11 AI taxonomy based on the high-level role it plays in business. You could use this taxonomy to guide you in eliciting available business actions you can help with AI. (This figure is a repeat of figure 2.5.)**

## Summary

- Project management is about making the best decisions based on the information you have now, and usually with a time constraint. To get an early indication of the business value of your ML pipeline, you should analyse it using MinMax analysis.
- To economically allocate resources in your AI project, you need to determine if you're using the right ML pipeline and then improve the right stages of that ML pipeline as needed. The former is done using a MinMax analysis, and the latter using sensitivity analysis. This is the *E*conomise part of the CLUE process.
- MinMax analysis allows you to determine if your ML pipeline already meets business goals, needs improvement to be able to meet those goals or is incapable of meeting those goals.
- MinMax analysis helps you implement the 'If you're going to fail, fail fast!' policy on AI projects.

# 7 Guiding an AI project to success

***This Chapter covers***

- Performing sensitivity analysis on the ML pipeline
- Assessing advanced sensitivity analysis methods
- Accounting for the effects of time in your pipeline
- Organising a project so that 'If you fail, you fail fast!'

This Chapter answers the questions: "What should I do when my ML pipeline needs improvement, and how do I know that I'm improving the right stage of the ML pipeline?" Such issues almost always arise for an AI product that's already on the market, and your goal is to continue to improve an AI product's user experience.

The same questions arise during the initial development of the AI project when your current ML pipeline needs to improve to meet business goals. Technically, this situation happens when the Min part of MinMax analysis is failing, and the Max part is passing. (Section 6.4.3 describes details of such a scenario.)

In this Chapter, I'll show you how to improve your ML pipeline. The key is to correctly decide the stage of an ML pipeline on which you should concentrate your

improvement efforts, which allows you to economise your resources. The ***E****conomise* part of the CLUE process addresses how to best direct your resources.

- In section 7.1, we look at how sensitivity analysis shows you which stage of your ML pipeline needs improvement.
- Section 7.2 completes our journey through the CLUE process.
- Section 7.3 discusses advanced methods for performing sensitivity analysis and when you should use them.
- Section 7.4 shows you how to manage the growth and maintenance of the ML pipeline past the release of your AI project.
- Section 7.5 addresses how you should balance a set of AI projects and your current project by reinforcing your winner projects and cutting off your losers.

## 7.1 Improving your ML pipeline with sensitivity analysis

When you know that you need to enhance the results of an ML pipeline, the question arises as to which part of the pipeline you should upgrade. Do you need cleaner data with fewer errors or a better AI algorithm? You have limited resources and can't just say, "Let's improve everything in the ML pipeline at the same time and see what happens." You need to choose a stage of the ML pipeline to improve. This section introduces the tool that would guide you through the quest for what is the best stage of the ML pipeline to improve next. That tool is called *sensitivity analysis*.

> **DEFINITION** *Sensitivity analysis* shows how refining a single stage of the ML pipeline improves the result of the pipeline as a whole. Sensitivity analysis guides you to the stage of the ML pipeline you should upgrade first.

At its core, sensitivity analysis answers the business question: "In which stage of my ML pipeline should I invest for the maximum pay-off?" Knowing the answer to that question allows you to economise resources put into pipeline improvement.

Let's look at an example. Suppose you have the pipeline in figure 7.1, and you want to improve the result (the output from stage E) of the pipeline.

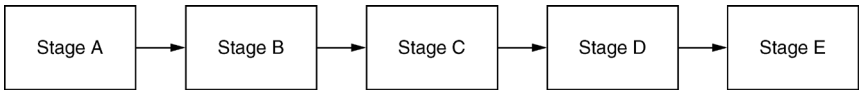


**Figure 7.1 An example of an ML pipeline. We use this pipeline as a base example for sensitivity analysis. (This repeats figure 6.10 for the reader's convenience.)**

In effect, you're asking, "What's the pay-off if I improve each stage of the pipeline?" For example, if stage A is 1% better, what's the improvement in the overall pipeline? You'd then ask the same question for stages B, C, D and E. Sensitivity analysis is the tool you use to answer these questions, and it allows you to replace intuition and gut feeling with the data-driven approach.

> **TIP** Sensitivity analysis gives you the business value of developing a stage in the ML pipeline. Sensitivity analysis of every stage allows you to rank the pipeline's stages by how much overall improvement in your business each stage will cause when it's improved.

Based on the results of sensitivity analysis, you can prioritise the stages of the ML pipeline and create a backlog of tasks needed to improve it. Once you have that backlog, you'll use the project management methodology you ordinarily use to manage the rest of the project. With sensitivity analysis, you transform the problem of improving the ML pipeline into management decisions similar to the ones that you make daily. This approach applies regardless of whether your environment is Agile or not. You're balancing the cost and time of upgrading a stage in the ML pipeline with the business benefits.

> **NOTE** I've written the next sections, 7.1.1 and 7.1.2, for readers with an engineering background, so some simple introductory calculus-level concepts are ahead. You can skip over the details of those sections if you're only looking to get an idea of how your team should perform sensitivity analysis.

In the following sections, I present two advanced methods for how to do sensitivity analysis:

- Section 7.1.1 covers local sensitivity analysis. Your team should use local sensitivity analysis when you expect that only incremental improvement in a stage of the ML pipeline is possible.
- Section 7.1.2 covers global sensitivity analysis. Your team should use global sensitivity analysis when you expect that a wide range of improvement in a stage of the ML pipeline is possible.
- Section 7.1.3 presents an example of interpreting the results of sensitivity analysis.

A manager making the decision about which stage of the pipeline to improve next is primarily interested in interpreting the results of sensitivity analysis (which is described in section 7.1.3); the type of analysis performed is of secondary interest. A technical expert performing a sensitivity analysis, however, must know how to choose the correct form of analysis – local or global.

### 7.1.1 Performing local sensitivity analysis

Local sensitivity analysis answers the question: "What would be the business result if I were to improve one stage of my current ML pipeline just a little?" This type of sensitivity analysis is appropriate when you expect that only small, incremental improvements in a specific stage are possible. An example is when you (or the broader community) already put so much effort into improving a specific stage of the pipeline that you believe that the days of substantial improvement in the results of that stage are behind you.

As an example, let's assume we want to improve the ML pipeline in figure 7.1 and want to find which stage of the pipeline, when improved, would result in the best overall performance of that pipeline. To do this, we would apply sensitivity analysis to every stage of the ML pipeline.

> **NOTE** Sensitivity analysis of a single stage is much faster to perform than it would be to actually improve (for all possible results) a particular stage of the ML pipeline. Therefore, we can use sensitivity analysis as a guide for which stage of the pipeline to improve next.

Let's look at analysing stage B from figure 7.1. Suppose that stage B classifies its input into a couple of categories. Let's assume that the technical metric for measuring stage B's ability to classify the input into the correct categories should be classification accuracy. Furthermore, suppose that you're currently achieving an x% classification accuracy in stage B. Figure 7.2 illustrates this method of sensitivity analysis.

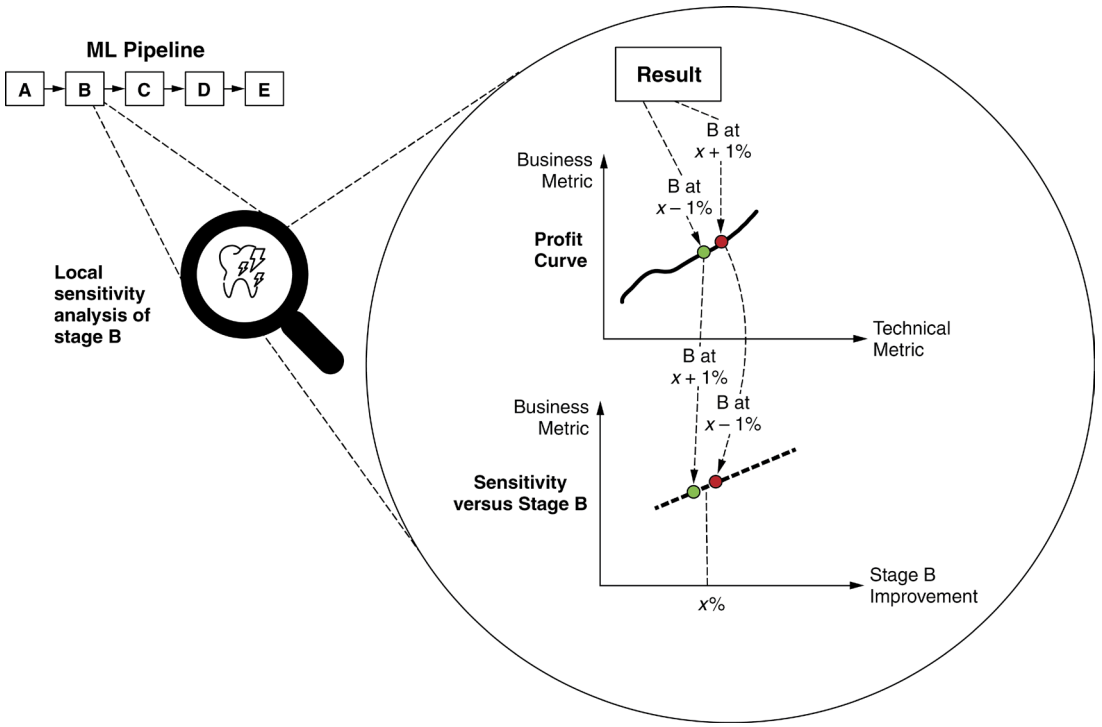


**Figure 7.2 Local sensitivity analysis. For small improvements in the response of stage B, this analysis assumes a linear response in the ML pipeline. If a gain of 0.5% in stage B results in a 1% improvement in the pipeline's result, then a 1% gain in stage B would result in a 2% improvement in the entire ML pipeline.**

What happens to the output of the last stage (stage E) of the pipeline as you change the classification accuracy of stage B in the vicinity of $x$%? What happens at $x-1$% and $x+1$%? You perform the local sensitivity analysis by simulating improvement in the results of stage B.

You perform this type of sensitivity analysis locally, near the point at which the accuracy of your classifier is $x$%. You get the result of the whole pipeline when you change the output of stage B to have an accuracy of $x-1$% and $x+1$% and then measure the effect of the change in stage B to the output of the ML pipeline. You use the profit curve to transform technical metrics into business metrics, and, finally, you plot the relationship between changes in stage B and the business results of the whole pipeline.

**WARNING** The profit curve and the *sensitivity versus stage* curve are two totally different curves. The former shows you how the business metric changes when the technical metric measuring the whole ML pipeline changes. The latter shows you how the business metric changes as a function of the improvement in the individual stage.

### How to get 1% worse (or better)

To get a 1% worse output, you take the classifier as-is and run the output of the classifier through a random number generator. The random number generator should be tuned to decrease the total accuracy of the classifier by 1%. To get a 1% increase in accuracy, you can use humans to review the results and correct them.

You can also use a commercial off-the-shelf (COTS) product or service that achieves better results than you're currently able to produce.[a] Such a COTS product or service doesn't have to be capable of playing a permanent role in the production version of your ML pipeline. For the purpose of the analysis, it doesn't matter if, for the reasons of cost, performance or the complexity of integrating the COTS product with the pipeline, you can't incorporate the COTS product in the production version of the ML pipeline. The COTS product just needs to be useful for experiments that can yield a one-time increase in accuracy of 1%.

[a] You should always survey the commercial landscape before investing significant resources in building in-house solutions. Therefore, you should already know what these COTS solutions are.

When performing local sensitivity analysis, you're assuming that in the vicinity of $x$%, the response of the whole ML pipeline is linear. If a 1% change in stage B causes a 1% change in the output of the ML pipeline, then a 2% change in stage B would cause a 2% change in the output. With this approach to sensitivity analysis, you understand the behaviour of the whole ML pipeline's output *in the vicinity of stage B's classification output being x%*.

This approach to sensitivity analysis is most appropriate when you're expecting that any improvement in the results of the sensitivity analysis would be incremental and that increasing performance by 1% might be non-trivial.

> **NOTE** A typical example of this situation is when the pipeline stage you're improving is an implementation of some AI algorithm. Your team may also have worked on improving that stage for a while. Here, the assumption is that even a small improvement in the metric might be a struggle.

In the example I've just used to illustrate local sensitivity analysis, there's nothing magical about a 1% metric improvement. You don't have to use a 1% improvement; you could also use 0.1%, 1.5%, 2% or some other percentage. What's essential is that you use a percentage increment that's a fraction of the total possible improvement in the stage. For example, if the potential gain is limited to 0.1%, you can use the increment of 0.05%.

### 7.1.2 Global sensitivity analysis

Often, there's no reason to believe that you'd be limited to small improvements to a particular stage of the ML pipeline. Maybe it's possible to improve a stage by 30% or 60%? Global sensitivity analysis helps you understand how your ML pipeline reacts to drastic improvements in a single stage.

One example of the source of a potential dramatic improvement in a stage of the pipeline is when you don't have any implementation of that stage yet, but you know that there are multiple possible implementations of the stage that would provide drastically different results. Another example is when you're improving the pipeline stage that's performing data cleansing. At least, theoretically, if you put enough effort into such a stage, you can completely clean the data.

> **NOTE** In practice, there are limits to how high a data quality you can afford and achieve. Still, you have significant control over how much you could improve the quality of your input data.

When significant improvements in the results of one stage are possible, local sensitivity analysis isn't appropriate, because you're looking at the ML pipeline's response over a wide range. Instead, you should test the whole range of values you have available with an interval between test points. Figure 7.3 shows this approach to sensitivity analysis.

Global sensitivity analysis is performed similarly to local sensitivity analysis, but instead of carrying out analysis at two points ($x-1\%$ and $x+1\%$), you analyse over a range of values that the stage can produce. With this approach, you can better accommodate non-linear relationships between changes in a single stage of the ML pipeline and the output of the ML pipeline as a whole. The disadvantage is that you need to put more work into the sensitivity analysis because you're analysing more points. So how many points (and in what interval) should you use?

As far as the interval is concerned, it comes from comparing the current level of performance to the maximum level of performance you think is possible (or 100%, if you have no reason to suspect you can't reach 100%). Regarding how many points you

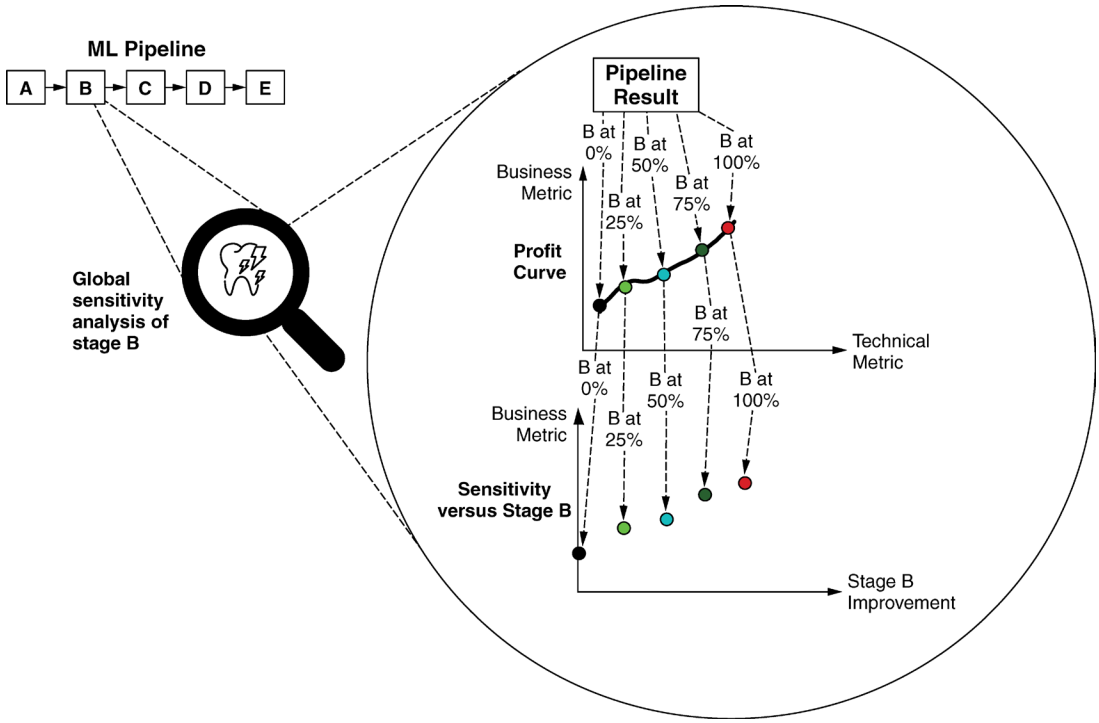


**Figure 7.3 Global sensitivity analysis uses a wide range of values. You perform global sensitivity analysis when there's no reason to believe that only small improvements are possible in some stages of the pipeline (in other words, you believe drastic improvements are possible).**

need to perform the analysis, that's determined by how much effort you need to carry out the analysis per test point. The recommendation is to start with at least three points and use as many points as you can accommodate within the available time frame.

Once you've completed sensitivity analysis for every stage of your pipeline, you have data on how changes in individual stages of the pipeline can affect your business value. This data gives you a massive advantage over people who improve the output of their ML pipelines based on 'experience and intuition'.

### 7.1.3 Example of using sensitivity analysis results

Suppose you have the ML pipeline from figure 7.1, and a MinMax analysis has shown that you need to improve it further to reach your business goal. You have an estimate from the development team about how long it would take to improve each stage of the pipeline (and how much each stage could be improved). Those improvements would take significant time: something on the order of occupying a majority of the team for weeks to improve each stage. You want your product released soon, so you want to implement only the improvements for which the pay-off is reasonable.

Let's suppose that the business metric is profit in dollars per unit, and you need an improvement of USD 3/unit to reach the value threshold. Also, assume you have the following results from the sensitivity analysis:

1. Stages A and B could be substantially improved. It was simple to simulate various improvements in them. However, after performing global sensitivity analysis, neither one of these stages significantly affects the results of the pipeline.
2. Stage C can be improved by only about 1-2%. It takes the engineer an hour of work to perform sensitivity analysis at a single point. The result of that analysis is that for every 1% improvement in stage C, based on local sensitivity analysis, the whole ML pipeline would improve by USD 10/unit/%.
3. Stage D also can be improved only slightly (1-2%), and it takes the whole team two days to perform analysis at a single point. Local sensitivity analysis at two points has shown that sensitivity for a 1% improvement in this stage is USD 0.05/unit/%. The difficulty of improving stage D is comparable to that of improving stage C.
4. Stage E can't be improved at all beyond its current level. (It's just providing notifications.)

In this example, I would choose to improve stage C: it has high sensitivity and appears to be equally easy to upgrade as stage D. Most importantly, I believe that a reasonable and achievable amount of improvement in stage C would bring me to my business goal.

## 7.2 We've completed CLUE

As you learned in section 6.2, the combination of MinMax analysis and sensitivity analysis is how you *Economise* your scarce resources, which is the final *E* of CLUE. MinMax analysis tells you that you're working on the right ML pipeline. Sensitivity analysis helps you work on the right stage of that pipeline.

> **TIP** If you're a data scientist, you may have noticed recursion here. You can think about MinMax and sensitivity analysis as if you're using data science to predict how your ML pipeline would behave.

CLUE is an integrated process, and each one of its stages depends on the previous stage. Figure 7.4 shows the dependencies in CLUE.

The *Economise* part of CLUE uses the *Understand* part (in the form of a profit curve as shown in figure 4.4). The *Understand* part requires business and technical metrics to be *Linked*. Finally, you worked on the right business problem, which is where the *Consider* part of CLUE comes in.

> **NOTE** The end goal of the CLUE process is a reduction in regret down the road. CLUE organises your AI project in a way that ties business and technology. CLUE prevents foreseeable waste, such as working on the wrong business problem or chasing technical solutions that can't deliver the anticipated business results.

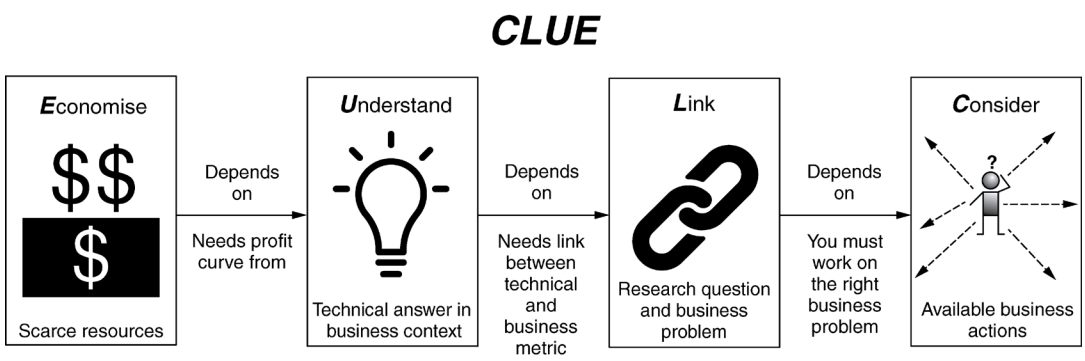


**Figure 7.4 Dependencies between the parts of CLUE – subsequent stages depend on the correct implementation of the previous stages. Following CLUE enables you to work on the right business problem, choose the proper ML pipeline to solve that problem and always work on improving the right stage of the pipeline.**

At every point during the project, CLUE helps you to focus on making informed decisions based on the information that's cheap to collect, but has a predictive power for the technical outcomes that are possible for your ML pipeline. It allows you to answer questions such as, "How likely is this ML pipeline to deliver acceptable business results?"

> **NOTE** CLUE helps you to make decisions based on the best information that's either available at present or easy to collect quickly.

Even if you elect to use a different process other than CLUE, such a process must address the same problems that CLUE does: you must work on the right business problem, understand the results in business terms and economise your resources based on the information you have (as opposed to gut feeling). If you fail to address any of these considerations, you're taking a chance with your project results.

### In the absence of data, personalities take over!

Not using data to manage the development of your ML pipeline means that you're handling its development based on gut feeling or on how much you trust personalities in your team who are advocating for particular actions.

Remember that it's typical for few team members to understand the whole ML pipeline, but most team members only have an extensive understanding of the part of the pipeline they're personally working on. It's human nature for team members to be more comfortable advising on how to improve the part of the system they understand instead of the part they don't understand.

Adjudicating technical proposals in the area where you're missing technical expertise is the worst kind of situation for the manager to sort out. While advocating for their

***(continued)***

opinions, team members will present qualities of leadership, integrity, maturity, persuasion power and expertise. Those qualities are genuine. *They're also entirely unrelated to what part of the ML pipeline might be the most productive stage to improve*. Deciding based on understanding people can put you on the wrong track here.

The gods of ML pipelines have a sense of humour. They will often assign the stage of the ML pipeline that's the most productive stage to improve to the least persuasive member of your team. Then those gods will happily let you live with the consequences.

The whole point of using a process like CLUE is to substitute the need for an intimate knowledge of AI algorithms and details of AI systems with a set of metrics that allows you to understand technical decisions in business terms. Remember the example of the factory manager from Chapter 2? The manager who wasn't as good a factory worker as the shift foreman, but still knew how to run the factory? That manager used data and management know-how to run the factory.

With the data CLUE provides, you can also use data and management skills to run your AI project. Applying CLUE offers a more scalable approach than asking managers to learn intricate details of AI and data engineering at the level that would be necessary if technical knowledge were the only tool you'd be using to adjudicate technical arguments.

### Managers are human too

Naysayers may tell you that no matter what, you need an intimate understanding of AI to lead an AI project. I beg to differ, especially if those AI skills come at the expense of leadership skills.

Besides, if such naysayers are right, we'd never have a widespread AI revolution in many areas of business and industry. Do you believe that there are enough people who have the ability, time, persistence and focus (and some of the other, less flattering qualities needed) to quickly get to a PhD level of understanding AI? Ah yes, those people are supposed to have also learned how to be good leaders. But AI is also improving rapidly, necessitating extensive and continuous technical education to stay current with it. After finding time to learn all of that, where would such people get the time to do the needed work?

We're asking for a combination of qualities that few people possess, and 'let's teach the details of AI to managers' doesn't scale. On top of that, we need such managers not only in top technology companies, but in every field that's supposed to be applying AI in the next few years.

I don't think it's realistic to expect that we'd be able to teach the details of AI to as many leaders as we'd need to deliver that AI revolution. We need to find a way to lead AI projects that doesn't require significant AI expertise, or we won't have many successful AI projects.

Maybe there are managers who have a fantastic intuitive feel and work on projects in new areas for which there has been little opportunity for experience (such as is the current case with most AI projects). I'd let you take a guess as to how numerous such managers are. However, I'd say that for people who don't have such an intuitive feel, managing projects based on processes such as CLUE makes them better stewards of management and software architectural responsibilities than managing on gut feeling alone.

## 7.3 Advanced methods for sensitivity analysis

The previously introduced methods for pipeline analysis are quick to do and could help you learn early on if your pipeline is the right one to use long-term. I recommend that when you're starting with CLUE and AI, you use the sensitivity analysis methods described in section 7.1. However, as you get more proficient with sensitivity analysis, you might be interested in more precise (but also much more complicated) ways to perform sensitivity analysis. This section discusses those methods.

### Complex topic ahead

This section describes advanced methods that require substantial process engineering experience to apply successfully. It's intended for advanced readers who already have a process engineering background or for teams that have already implemented CLUE and want to get better at it.

My goal is to teach you to recognise when you have a situation where these advanced methods might be needed, and give you just enough understanding of these methods so that you know what kind of help to ask for from the experts.

Sections 7.3.2 and 7.3.3 are the only parts of this book that I feel are relevant only for the large company with a significant investment in AI tools, technology and infrastructure. You would need an expert in the area of process engineering to apply many of the topics described in these sections, and there's no way to learn that expertise from a single Chapter or, for that matter, from a single book.

If you don't have the budget to afford such experts, I also include some references that will give you a head start on how to learn to do these things by yourself. To get the most from those references, you'll still need an engineering background (and much patience).

Once you've developed the initial ML pipeline and invested a lot of time and money into it, you may be willing to spend more time analysing ways in which you can improve that pipeline. This is where the more advanced analysis methods are useful. In specific scenarios, those methods could provide better analytical results, albeit at the price of an increase in the complexity of the performed analysis. A roadmap through the rest of this section follows:

- Section 7.3.1 shows you how to detect the presence of non-linearity.
- Section 7.3.2 talks about interactions in the ML pipeline.

- Section 7.3.3 introduces the concept of design of experiments, a technique that can discover and address interaction, but requires significant process engineering knowledge to apply successfully.
- Section 7.3.4 addresses common critiques that you might encounter when performing sensitivity analysis.
- Section 7.3.5 advises on the best practices for improving your ML pipeline by enhancing the quality of your data.
- Section 7.3.6 presents the applicability of some recent advancements in the field of sensitivity analysis for ML pipelines.

There are two significant sources of errors in sensitivity analysis: non-linearity [125, 126] and the interaction between pipeline stages. When you encounter non-linearity, the results of your local sensitivity analysis are subject to error. When you meet interactions, changing two stages of the pipeline can result in significantly different behaviour if you change them together rather than if you change one at a time. Let's look at each source briefly.

### 7.3.1 Is local sensitivity analysis appropriate for your ML pipeline?

Local sensitivity analysis, described in section 7.1.1, assumes that the response of the whole ML pipeline to the change in a single stage is linear: if a change of 0.5% in stage B produces a 1% improvement in the entire pipeline, then a 1% improvement in stage B would result in a 2% improvement in the ML pipeline. If the assumption of the linearity of the pipeline's response were violated, then your sensitivity analysis results would have an error in them. This section shows you how to detect when the assumption of linearity is broken. When that is the case, local sensitivity analysis isn't appropriate, and you should replace it with global sensitivity analysis.

Informally, a non-linear response means that the output on the profit curve could change faster (or slower) than would happen if the response were linear. Figure 7.5 shows a situation in which the system's profit curve increases in non-linear (superlinear or convex) fashion and, for a given percent increase in the performance of a single stage of the pipeline, you get more than that percent increase in the response on the profit curve.

Why does convexity of the response matter? Because, as you can see in figure 7.5 when you have a convex response, the farther away you move from the point at which you perform the analysis, the larger the error. In an extreme case, the error could be so significant as to invalidate the results of the analysis. You might even miss that the improvement in that stage of your ML pipeline would provide a considerable return. Such low-probability, high-impact events are commonly referred to in business circles as *black swans* [127].[1] So non-linearity matters when the significant pay-off is missed. Figure 7.6 shows a situation in which local sensitivity analysis misses the convexity.

[1] In Europe, it was believed for a long time that all swans were white, until someone travelled far enough to see a black swan. A single black swan, while rare, had a large impact on dispelling this theory.

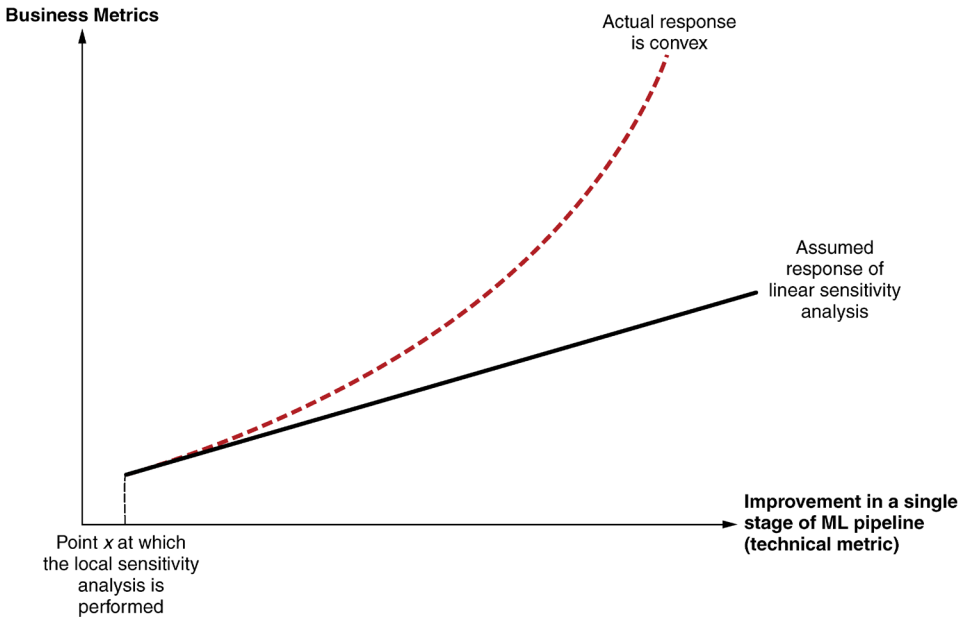


**Figure 7.5 Convexity in the ML pipeline's response. The further you move away from the point at which analysis was performed, the more significant the error in your analysis. Never extrapolate far from the point from where your local sensitivity analysis was performed.**

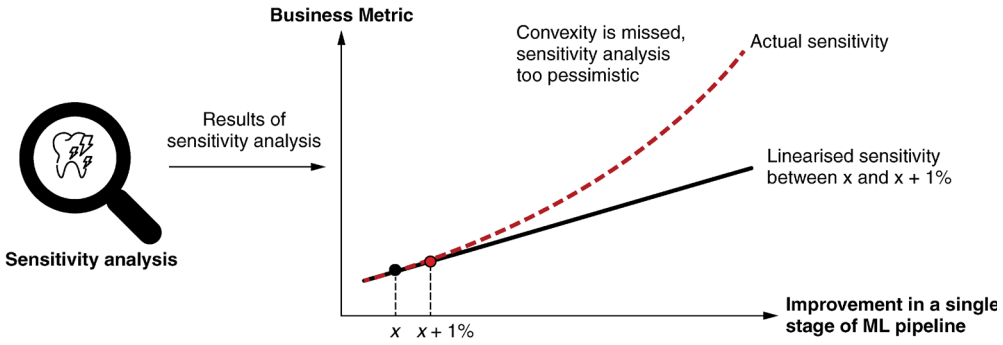


**Figure 7.6 Sensitivity analysis with convexity present. Localised sensitivity analysis performed at only two points, x and x + 1%, has missed the convexity because you can always draw a line between two points.**

There are heuristic techniques that you could use to show that you might be having a non-linear response and that are particularly appropriate for situations in which non-linearity is likely to highly skew the result. One heuristic consists of replacing local sensitivity analysis with performing global sensitivity analysis at three points and seeing

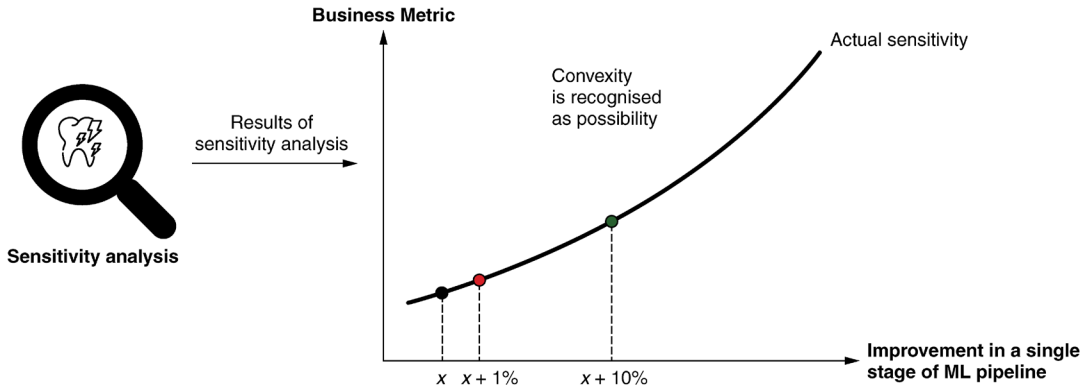


**Figure 7.7 Global sensitivity analysis in the presence of convexity. At a price of the increased complexity of the analysis, global sensitivity analysis could detect the presence of non-linearity of response.**

whether the response is linear or if there are indications of convexity/concavity. Details of this technique are explained by Taleb et al. [126]. Figure 7.7 shows the application of that heuristic. Coincidentally, any global sensitivity analysis could apply the same technique (described in the Taleb et al. paper [126]) to detect the non-linear response.

How much do the possible errors in the linear sensitivity analysis matter for your ML pipeline, and does the fact that linear sensitivity analysis could miss convexity mean that you should always perform global sensitivity analysis? I take a pragmatic view and remember that all project management decisions need to be made under time constraints. The key difference between local sensitivity analysis and global sensitivity analysis is that the former requires two points (such as $x$ and $x + 1\%$) to perform, while global sensitivity analysis (and the heuristic approach described in Taleb et al.'s paper [126]) requires at least three points to complete.

> **TIP** If I know that performing sensitivity analysis at one more data point is cheap and straightforward, I always perform global sensitivity analysis. If it's expensive (for example, when I have to use a human as a proxy to perform the analysis), I use local sensitivity analysis where appropriate and perform it at just two points.

If sensitivity analysis of the ML pipeline at additional data points would be expensive, the only situation in which I worry about convexity is when I don't have a clear winner after analysing the rest of the pipeline. However, in such a case, I schedule improvement of the stage that has convexity in front of the stages that have a similar sensitivity analysis result, but don't exhibit a convex response. That gives the (possibly) black swan a chance to work in my favour.

**NOTE** You shouldn't think about non-linearity as a danger to your project. Convexity means that it would be easier to achieve a business goal than if the sensitivity of the stage was linear. The impact of missing concavity during the sensitivity analysis is limited. Your project should be organised so that, 'If you fail, fail fast'. Therefore, if a stage improves slower than you expect it to, you find that out early and stop working on that stage.

### 7.3.2 How to address the interactions between ML pipeline stages

Sometimes, the result of changing two things at the same time is very different from changing them one at the time. That happens when there's an interaction between two variables. This section provides an example of interactions and advice on how to address them in the context of analysis of the ML pipeline.

One example of interaction is that when you buy a laptop, you care about both the weight of the laptop and the speed of the processor. The lighter laptop is always better, all other things being equal. The faster processor is always better, all other things being equal. However, if you put a speedy processor in a tiny (and light) laptop, the processor might start overheating, as there's not enough volume in the laptop for the processor to cool appropriately. Making a small, but powerful laptop is also expensive. As a result, a tiny laptop with a mighty processor might not be worth building (or buying).

#### What's the effect of the interactions?

In the presence of interactions, an analysis that changes only a single variable at a time (such as the output of a single stage of the ML pipeline) can be invalid. Furthermore, you'll encounter statisticians and data scientists who would point out that the presence of interactions affects the results of MinMax and sensitivity analyses.

That's true, but how is it relevant to the analysis your team could perform? You can perform interaction analysis only if your team knows how to perform interaction analysis. Even if interactions are present, if you don't know how to find them, you would have to accept the risk of interactions (and its effect on MinMax and sensitivity analysis).

My advice is that determining how much to care about interactions depends on the team you have. If your organisation has significant knowledge of process engineering and the ability to quickly analyse the behaviour of your ML pipeline at multiple points, then I would advise that you perform interaction analysis. In practice, that usually means a well-funded team in a large corporation working on a project in which a small change in the ML pipeline could provide a huge financial pay-off.

For teams just starting with AI and sensitivity analysis, my advice is to not worry about interactions initially and concentrate on what happens with the ML pipeline when you change only one factor at a time. Two Six Sigma resources from ASQ [21,22] give some starting points on the process and a profile of people who have good experience in designing experiments for detecting interactions, which brings us to the broader topic of the design of experiments ([24]).

### 7.3.3 Should I use design of experiments?

Design of experiments (DOE) [24] is a methodology that has been successfully used in the area of process engineering for decades to improve quality, cost and the efficiency of processes such as manufacturing physical objects. When you have a factory line costing millions of dollars to run, you want to know that it's running optimally. DOE is about conducting experiments, the results of which show you how to improve your factory line. This section introduces you to DOE and advises when DOE is applicable to an AI project.

Historically, software development didn't use DOE much. Some reasons for that are that DOE is a complex topic that's unfamiliar to software engineers. More importantly, instead of DOE, the cost of implementing a pseudo-experiment ('let's just try it and see what happens') was small in terms of software; for example, it was easy to change a configuration parameter to see how your database reacted.

> **NOTE** Pseudo-experiments have many weaknesses. For example, they could miss interactions. Also, they're sensitive to background processes; if a background process runs at an inconvenient time, it can impact the results of your pseudo-experiment. Unlike pseudo-experiments, a proper experiment conducted under the DOE methodology would provide correct answers even in the presence of interactions and background processes.

With recent AI projects, the cost of 'just trying something' has drastically increased. It's not unheard of that a large project in a large corporation might use hundreds of machines to train some complicated AI algorithms [128]. Running such a system costs a lot of money, and, as the price of hardware infrastructure starts rivalling the cost of running a factory, the methods used for running the factory may become relevant for software engineers too.

DOE allows for many advantages compared to pseudo-experiments. Using DOE enables you to better manage an expensive ML pipeline development.

So should you use DOE on your project? For most AI projects, I don't recommend DOE because it's complicated to perform right and requires experts with specialised training to implement it correctly. The cost of the experiments (and those experts) is too high to justify this approach on an average AI project today. Furthermore, if you've never thought about the management of your ML pipeline systematically, the methods already presented in this book would be a huge step forward.

> **NOTE** If you look at the history of how the field of process engineering developed, it also started with simple experiments and built DOE theory and knowledge later.

However, if you're running a massive AI operation, and your capital investment in the team, hardware and infrastructure rivals the cost of running a factory, and if the results of every decision you make regarding the management of your ML pipeline are of correspondingly high stakes, I'd reverse that recommendation. Instead, my advice

would be to work with an expert who knows both AI and process engineering. Such an expert can advise you on a case-by-case basis about what's right for your system. You're already spending so much money on running your system that you should design experiments in it properly.

### 7.3.4 One common objection you might encounter

This section is about a criticism that you're likely to face and why that criticism is both correct and mostly irrelevant for the practical use case of ML pipeline analysis. The criticism is that it's possible to construct an example in which sensitivity analysis would indicate results that, while useful, aren't the best possible result.

Let's look at an example. Suppose you have a situation in which you performed global sensitivity analysis at five points, and there's an unusual situation in which interactions have caused the sensitivity curve in figure 7.8.

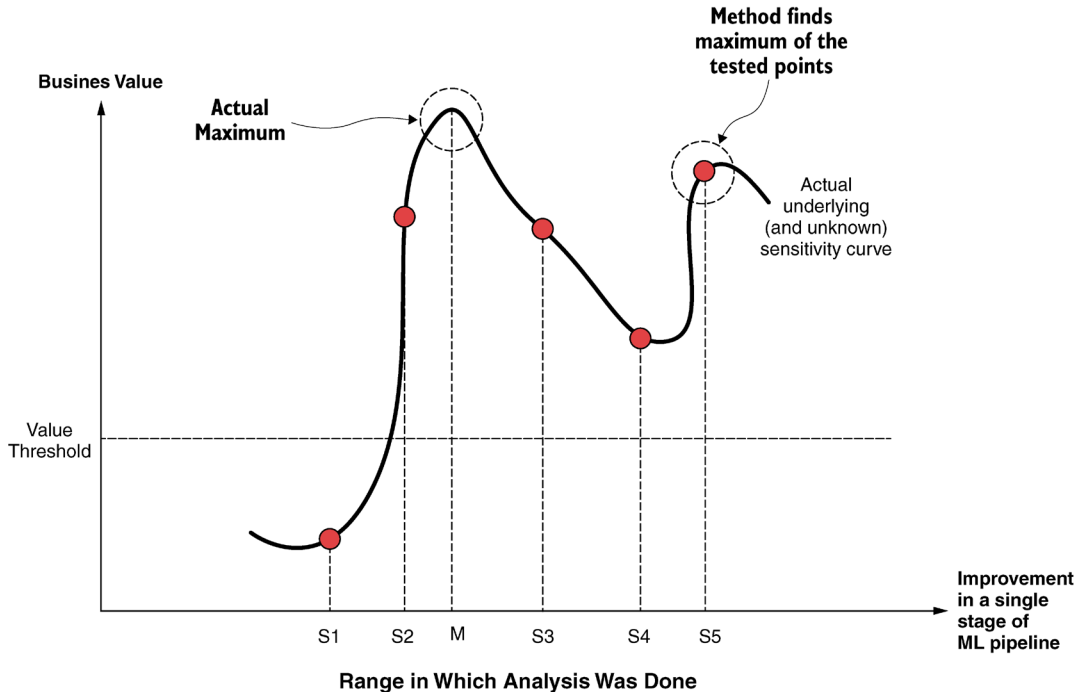


**Figure 7.8 Sensitivity analysis performed in locations S1-S5, but but missing the actual maximum at point M. It doesn't matter; you're above the value threshold, so you're still making money. No cheap analysis performed on only a few points can avoid this problem.**

In figure 7.8, you've performed the analysis at five points, S1-S5. The best result that you got is the maximum of the five points you looked at (S5). However, it's not the absolute maximum of the underlying function, which is the location *M*.

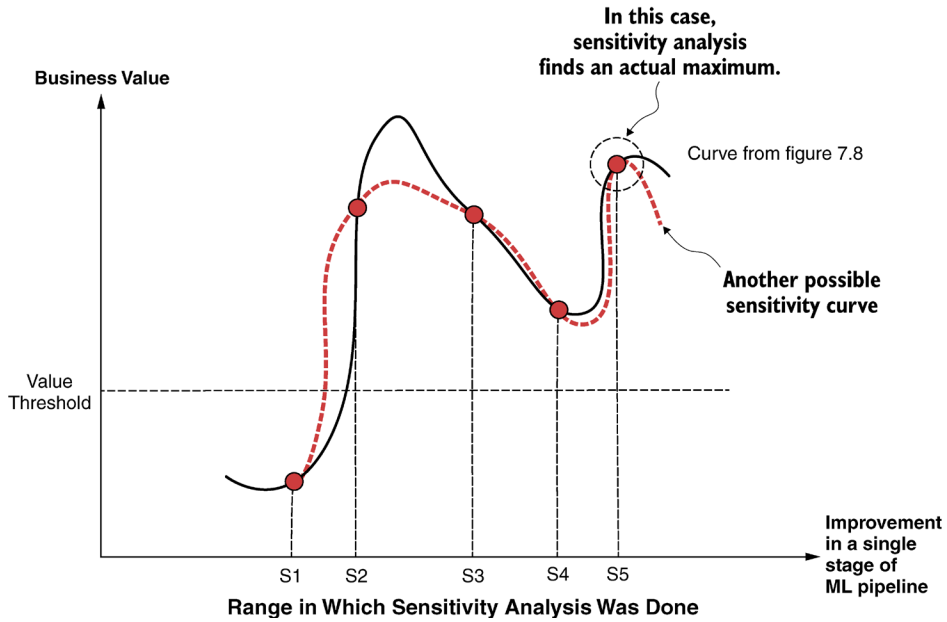


**Figure 7.9 If the underlying curve has a shape given in this figure (as opposed to figure 7.8), sensitivity analysis will find the actual maximum. You never know the shape of the underlying curve, so it doesn't matter what the maximum of that curve is. What matters is that you're above the value threshold.**

While the previous example looks convincing, think again. In practice, do you have access to the 'truth', represented as the actual underlying sensitivity curve in figure 7.8? No, you don't! That real sensitivity is unknown and unavailable to your project; all you have is the result of sensitivity analysis in points S1-S5. The underlying sensitivity curve stays hidden and might as well have been the curve in figure 7.9.

I could construct many counterexamples, but these have one thing in common: the criticism I mentioned assumes that you already know the whole shape of the curve in figure 7.8! In practice, you never have access to the underlying sensitivity curve. If you did, you wouldn't wonder about the stage in which to invest in the first place, you'd just read it from the sensitivity curve.

Which brings us to the difference between mathematical theory, which is what the original objection in this section was based on, and practical project management. The objection raised is a theoretical objection without any actual advice on what to do. While critics can point to some methods that could find a maximum value of the underlying curve, when you ask for details, you'd learn that such methods would require performing sensitivity analysis at many points. The number of points at which you analyse is a determinant of the cost of analysis, so such methods are often too expensive to be practical for guiding improvement decisions in an ML pipeline.

Practical people need solutions, and in this case, the answer is recognising that you're already familiar with the problem. It's the same problem as if I ask you, "So, what's the maximum amount of money you could have made in life?" Well, maybe if you were introduced to the right people, you might have founded a company more prosperous than Google is today, but you'd never know. You don't know the 'income sensitivity' curve of your life that answers that question.

> **NOTE** In life, you never know if you've made as much money as it was possible to make. The only thing you know is if you made enough money to live comfortably or not.

Like with life, with a project the question isn't "What's the maximum value of the curve?" Remember, you're on a rich hunting ground (section 3.1.1), and you want to make sure that the AI project you invested in is profitable. You also need to decide based on the best information available *at the time when you must make an investment decision*. If sensitivity analysis doesn't find the maximum value, but still allows for constructing a profitable pipeline, that's called success in the business.

> **NOTE** It's *rare* that you encounter an ML pipeline in which sensitivity analysis would miss so many profitable areas that you'd be stuck and unable to improve it.[2] It's even rarer that an ML pipeline in which that happens is easy to bring to profitability. Such a rare pipeline isn't a rabbit that's easy to hunt, and you're on the rich hunting ground. Try something else.

Sensitivity analysis involves maximising the information available to you at present and allows you to determine if the next action you take is profitable or not. That's all you need to run a successful AI project.

### What about unsupervised learning?

You may be asked if sensitivity analysis (as well as MinMax analysis) is only applicable to supervised learning. The answer is that it's relevant to any type of AI because you can always construct a profit curve for unsupervised learning too.

Suppose that your product is analysing data and creating clusters from this data. Afterwards, the clusters are presented to humans, who use them as one of the inputs to a decision that needs to be made under a considerable time constraint. An example of that would be a system that uses AI to cluster types of fault in a complex transportation system.

In such a situation, the more clusters you present (and the more difficult it is to see what's common in each cluster), the less value the system has for the user. Clearly, there's a relationship between the output of a system and the value to the user . It may happen that you'd have to perform an experiment with the help of actual users to determine the value to the user. The results of that experiment can be described in the form of a profit curve.

[2] You'd have to encounter such a situation not only in a single stage of the ML pipeline you're analysing, but in multiple stages at the same time.

### 7.3.5 How to analyse the stage that produces data

Some of the stages in your pipeline are likely to be operations with the data, and every AI algorithm you use would take data as an input. You can typically improve the quality of such input data. This section provides advice on how to analyse how data improvement affects your ML pipeline.

Your goal when conducting global sensitivity analysis or the Max portion of a MinMax analysis should be to get the best data you can. Here, the best is across all dimensions – larger dataset, better-targeted data, cleaner dataset.

> **NOTE** If you're building an AI-powered physical device, such as a camera, better might mean having a superior sensor in it. If you have a problem with a fuzzy picture, can you get a better camera? If you have a problem with obstructions, can you get more cameras?

It's often the case that better data could beat the better AI algorithm [129], so checking what happens when you have cleaner data is essential.

#### How to clean all that data?

An issue sometimes arises in the big data space: how do you measure what the impact of getting cleaner data would be if your data volume is substantial? It's not like you can ask humans to clean 1 PB manually! Moreover, while you may launch a project to clean 1 PB of data, by the time you get the answer, you've also spent a ton of time and money to do so, so the economic value of that information is low.

Fortunately, there's a simple solution: collapse two stages of the pipeline into one. Suppose that one stage of the pipeline ingests image data and another applies object recognition based on deep learning to those images. It's exceedingly difficult to answer the question, "What would be the result of applying this deep learning network architecture if I had perfectly clean image data?" So trying to improve the ingestion phase is difficult.

However, it's much simpler to answer the question, "What's the best result that we can achieve with image recognition if we look at the data and algorithm together?" You simply look at the best vision recognition results achieved so far on any dataset. By collapsing two stages of the ML pipeline into one (data ingestion and recognition), you've transformed a complex question into a simple one.

### 7.3.6 What types of sensitivity analysis apply to my project?

Sensitivity analysis is a complex topic and an area of active research in the computer science community (for example, see *Global Sensitivity Analysis: The Primer* [117]). You're likely to be interested in how some of the latest research can help you perform better sensitivity analysis. This section presents the criteria you should use to determine its applicability for your project.

The most important question that you should ask when presented with any research in sensitivity analysis is, "How much work would it be to apply this method?" Only the methods that are easy to implement (compared to the total size of your project) have a practical value for project management of the ML pipeline.

**TIP** If your analysis is so complex that performing it costs you as much as building the ML pipeline, you might as well just build the pipeline and see what happens.

The largest source of cost in ML pipeline analysis is the number of points at which the analysis needs to be performed. Consequently, global sensitivity analysis techniques that require analysing thousands of points are far less applicable (especially for an AI team with limited experience in process engineering) than the methods described in section 7.1.

### A handy trick

While it can be difficult to make data or results better, it's often straightforward to make them worse by introducing errors into them. The trick is that instead of making the result of the ML pipeline's stage 1% better and analysing at the points x and x + 1%, you make the result 1% worse and analyse at the points x – 1% and x%. If you use this trick, you're assuming that the behaviour of your pipeline is as similar when the output of a stage slightly improves as it would be if the output slightly declined.

Suppose that you're conducting global sensitivity analysis at points 33%, 66% and 100%. Once you complete analysis at 100%, you could purposely corrupt the output data of that stage to perform analysis at points 33% and 66%.

The same trick applies to local sensitivity analysis. If your pipeline is already producing results in a stage that are, for example, 95% accurate, don't conduct sensitivity analysis at points 95% and 96%. Instead, perform it at 94% and 95%. It's far easier to introduce an error in the outputs of your current stage than it is to improve it.

You could use this same trick to adapt sensitivity analysis methods that require evaluation at thousands of points to an analysis of an ML pipeline as you construct the results for a single point (the very best point), and then degrade those results to simulate other points.

However, be advised that the technique presented in this sidebar has subtleties and traps that are easy to fall into. Errors you introduce in the output aren't just random errors. They must have similar statistical properties as the errors that the actual implementation of a stage in the ML pipeline would have. You need experts to avoid this trap.

My advice is that you shouldn't attempt this technique until you have people on staff who have significant experience with sensitivity analysis, process engineering and analysis of statistical distributions.

## 7.4 How your AI project evolves through time

The techniques of MinMax and sensitivity analysis presented so far were focused on AI projects that would be delivered fast. All other things equal, you should prefer projects that can be delivered to your real customers quickly [28], and that's especially the case with your initial AI projects. However, once delivered, that AI project could be in the market for a long time. Furthermore, sometimes your AI project is breaking new ground, and it simply takes a long time to deliver it to the market. This section shows how you should modify methods presented so far when you're leading long-running AI projects.

Section 7.4.1 discusses how time affects your project. Section 7.4.2 shows you how to modify the Understand part of CLUE to account for the influence of time in long projects. Section 7.4.3 shows you how to diagram a change in the business value of the project through time.

### 7.4.1 Time affects your business results

In managing a project, we often focus on the sequence of steps we need to execute to succeed. We think about time in the form of the project deadline, and we think about deadlines as a responsibility of engineering. This section shows you a different way of thinking about time so that you can consider together the engineering and management decisions that affect your project.

In many projects, time becomes an afterthought that's present only as a deadline. The management and engineering teams negotiate the deadlines. Once settled, deadlines become the problem of project managers and engineering. That results in a divorce between the impact of time on the value of what's delivered and the technical management of the AI project. Engineering focuses on not missing deadlines. Management concentrates on making contingency plans if deadlines are missed, and it might occasionally surprise engineering with the request for a new feature. Instead of integrating technical decision-making with the business results, the relationship between teams becomes detached, if not outright politely adversarial.

A better way to address the time dimension is by including a time to complete the project directly in the metrics that you're managing and optimising. If the time needed to complete the project matters, you should be able to quantify at least a range of how much it matters. Once you quantify the influence of time on your project, you can incorporate time in the profit curve.

> **TIP** If you can't quantify precisely, use estimates. See D. W. Hubbard's book [75] for ways to quantify 'intangibles' in the business.

The value of the project (and the value threshold it should meet) changes over time. For example, AI capable of indexing and searching the internet had immense business value in 1998 (before Google). Today, the value of such AI is much smaller. Both the profit curve and the value threshold of your project, therefore, evolve over time.

Those variations become significant when you're talking about more extended periods, such as successful AI projects that could be on the market for many years.

If a project faces a range of delivery dates, you shouldn't use a single profit curve for the whole duration of the project. You should have multiple profit curves that reflect the changes in the business value at various times. Figure 7.10 shows a set of profit curves for a project running for two years.

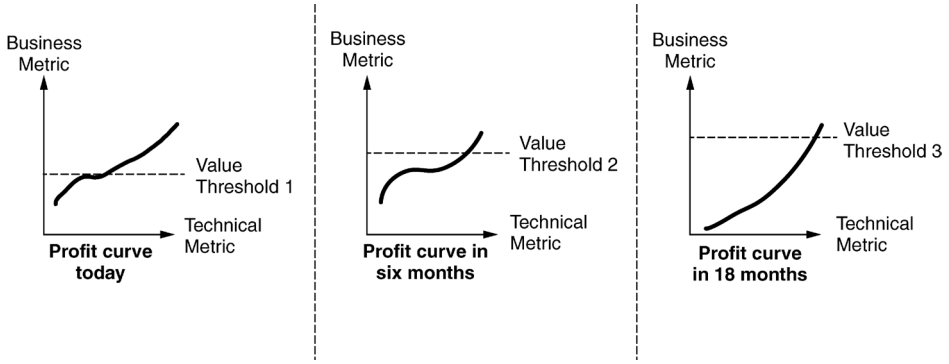


**Figure 7.10 A set of profit curves for a long-running AI project. Both the shape of the profit curve and especially the value threshold change with time.**

You should manage the ML pipeline with time in mind. Even an approach as simple as 'if done before June 1, accuracy *x* is worth USD *y*; if done later, it's worth USD z', when applied to the profit curve, would give you a quick way to address the time dimension.

> **WARNING** A deadline is at best an imperfect way to account for time. An Agile process by itself doesn't address such a dynamic of deadlines; it only enforces more common checkpoints between engineering and business teams. Even on Agile projects, you must purposely focus on the longer term implications of the ML pipelines you're choosing.

### 7.4.2 Improving the ML pipeline over time

At some point, you'll deal with much longer AI projects. Those projects could last for an extended period, and, as seen in section 7.4.1, the profit curve can change as time passes.

> **WARNING** On a longer project, the profit curve typically shifts over time to account for the opportunity cost of a delay . You must account for that shift when evaluating the business value of the project.

To account for that change, you estimate how long it would take to deliver an improvement in a stage of the delivery pipeline, and then you use the appropriate profit curve that reflects the value the improvement would have *at the time the AI project is released* (as opposed to today). Figure 7.11 describes this process.

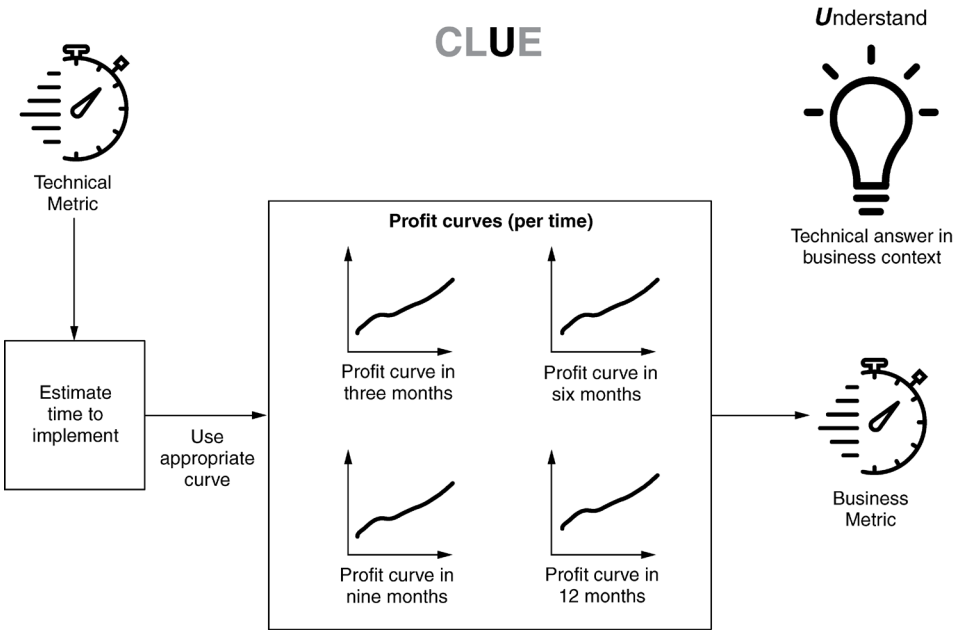


**Figure 7.11 Modification of the Understand part of CLUE to account for long delivery times. The project value changes with the time it takes to deliver it. You must use the profit curve corresponding to the time when you would release software to calculate the value of the improvement.**

Once you account for the influence time has on your profit curve, the scheduling of the improvement in your pipeline is simple. You schedule the stages in the ML pipeline to be improved in the order that allows you to reach the value threshold as soon as possible and to stay above the value threshold as the project is progressing.

> **NOTE** Just like the value of the improvement of a stage in the ML pipeline changes with time, so does the value threshold. For example, your value threshold could be USD 5/unit in the first six months, and then decline to USD 4/unit. Don't forget to account for the changes in the value threshold when building your schedule.

### 7.4.3 Timing diagrams: How business value changes over time

On longer projects, the business value of the project changes over time as well. You can represent the change of business value over time with the help of a *timing diagram*. This section gives an example of constructing such a diagram.

This example assumes that the value threshold is based on the value of your AI product to your end user and is expressed in the profit your end user makes per unit.

I would also assume that you're trying to capture a rapidly expanding market that's expected to have a high lifetime value for the company that establishes the standard solution. Therefore, the goal of your corporation is market presence. Your goal is to release a viable product as soon as possible, and to keep it continuously viable, not worrying about profit in the next 24 months. Corporate leadership expects that profit will come later, when your product is established as a standard.

Let's look at a scenario in which a change in the value thresholds over time is given in table 7.1. You can extract that information by applying the process shown in figure 7.11.

**Table 7.1 Values of stage improvements for your end user**

| Stage name | Improvement value today | Improvement value in six months | Improvement value in 12 months | The time needed to complete stage improvement |
|---|---|---|---|---|
| A | USD 7 | USD 4 | USD 3 | Two months |
| B | USD 30 | USD 27 | USD 21 | 11 months |
| C | USD 14 | USD 10 | USD 7 | 3 months |
| D | USD 10 | USD 8 | USD 6 | 6 months |

The value and time needed to improve a pipeline stage are given in table 7.2. You can construct the value threshold change by reading the value thresholds from the corresponding profit curves in figure 7.12.

**Table 7.2 The value threshold the unit has for your customer. You must exceed the value threshold for a customer to buy your product.**

| Value threshold now | Value threshold in six months | Value threshold in 12 months |
|---|---|---|
| USD 5/unit | USD 14/unit | USD 15/unit |

Figure 7.12 shows the timing diagram of the value that your ML pipeline is expected to have for the end user. Note that there are two reasons why the business value of the ML pipeline changes: improvement in the stage of the ML pipeline and the passage of time. Increases in the pipeline utility at 2, 5, 11 and 23 months are caused by completing improvements in stages A, C, D and B. Dips at 6 and 12 months are caused by the passage of time as the business value of the improvements in your pipeline stages decline.

A timing diagram allows you to determine what's expected for the value of your ML pipeline at every point in the future. That information helps you answer the question of in which order you should improve your pipeline stages if you have a long-running project.

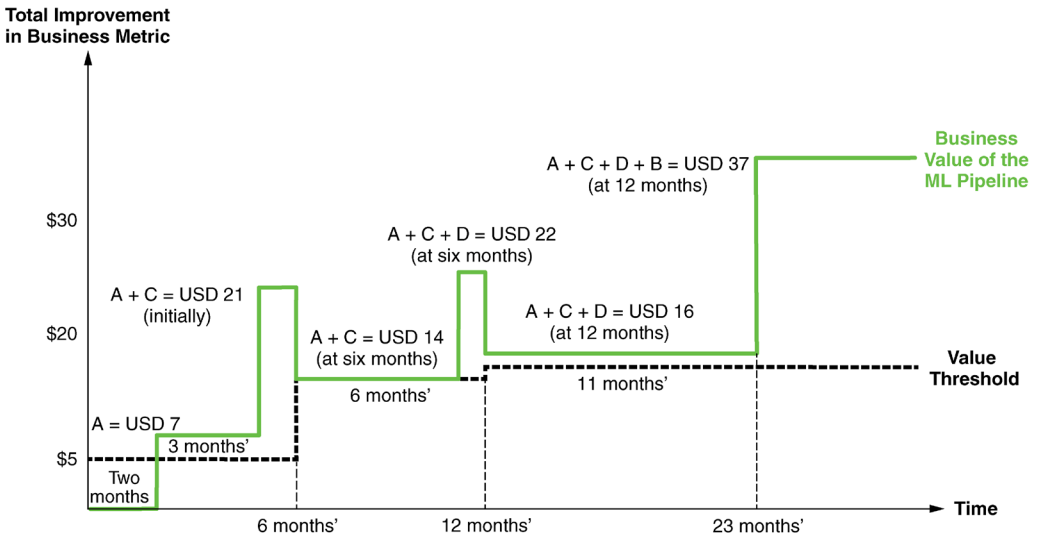


**Figure 7.12 Improving the ML pipeline accounting for the time parameters given in tables 7.1 and 7.2. The order of improvement of the ML pipeline stages should be first stage A, then stage C, then stage D and finally stage B. That order allows you to release a viable product after only two months.**

This technique is useful both in the early stages of the project, when you're choosing the best ML pipeline for your research question and when you need to manage the development of a new ML pipeline that would replace the current ML pipeline in your project. In the latter case, you could use a diagram like the one in figure 7.12 to tell you how to manage incremental improvements that you want to make in the old ML pipeline and estimate the point when you could expect that a new pipeline would be able to take over the job of the old pipeline.

## 7.5 Concluding your AI project

Managing projects requires many estimates, such as how long some functionality would take to implement, how complicated some AI algorithms would be to apply and how much business value implementation would provide. Sometimes, everything happens as estimated and planned. Other times, reality refuses to comply with our wishes, and we find that estimates don't work out. This section shows you what to do when the problem is much more complicated to solve than you initially thought it would be.

Today, if you're running your initial AI projects, your team is likely to be the first team in your company to use powerful AI techniques to address your business problems. This also means that if the project proves to be challenging to complete, your initial estimate of its complexity, although correct in the light of what you knew then, will need to be modified based on what you've learned now. Instead of continuing a project that's difficult to implement, you should pause it and try something simpler.

On early AI projects, you're on the rich hunting grounds, and you should think like a hunter (section 3.1.1) – don't spend time chasing mammoths, catch rabbits instead. If you're hunting for a while and figure out that an animal you're following is a mammoth that's quite good at camouflaging itself as a rabbit, you should abandon the chase and find a rabbit.

### If you're going to fail, fail fast

Your project management approach should be biased toward failing fast. You should make the trade-off of accepting the possibility of giving up too early and missing a potential solution if that means you'll avoid situations in which you're stuck for a long time on something that doesn't work in the end.

Remember that the primary way AI initiatives die is that they persist in problematic projects for far too long and have nothing to show for it at the end (section 3.1.1).

You should always *timebox* how long to allow your AI project to proceed before you pull the plug if you encounter difficulties. If research questions turn out to be much more difficult to implement than initially estimated, you shouldn't persevere in pursuing the answers. Instead, pause the current project and start working on more straightforward research questions instead.

With this approach, you're trading the possibility of putting on hold a project that potentially could result in a functional solution (but could also finish as a considerable time waster with nothing to show for it) to try more straightforward projects first. However, when using this technique, it's vital to understand what you found when you put the research question on hold. *You've decided to put that research question on hold; you haven't found that there's no business value in pursuing that research question further.*

Unfortunately, organisations often have a habit of classifying results of research projects as binary categories – 'yes/no', 'works/doesn't work'. To correctly use the timebox approach, you need to understand that the initial analysis of research questions has three possible results:

1. *Yes* – This approach is worth pursuing further. We should put many resources into it.
2. *No* – We've tried enough things to be confident that this is the wrong approach, and it's not expected to work. Don't put any additional resources into it.
3. *Maybe* – With the time we put into the *initial* investigation, we were unable to show that this approach works. However, we didn't investigate this long enough to know that it won't work even if we try harder. If we have more money and time later, we should revisit this problem.

**NOTE** It's crucial that your results be reported and tracked in the three-state logical form of yes/no/maybe. That's because you'd never reopen questions answered with 'no', but might resume questions answered with 'maybe' at a later date. Correctly making this distinction is the only way the 'be willing to abandon difficult projects early' approach works.

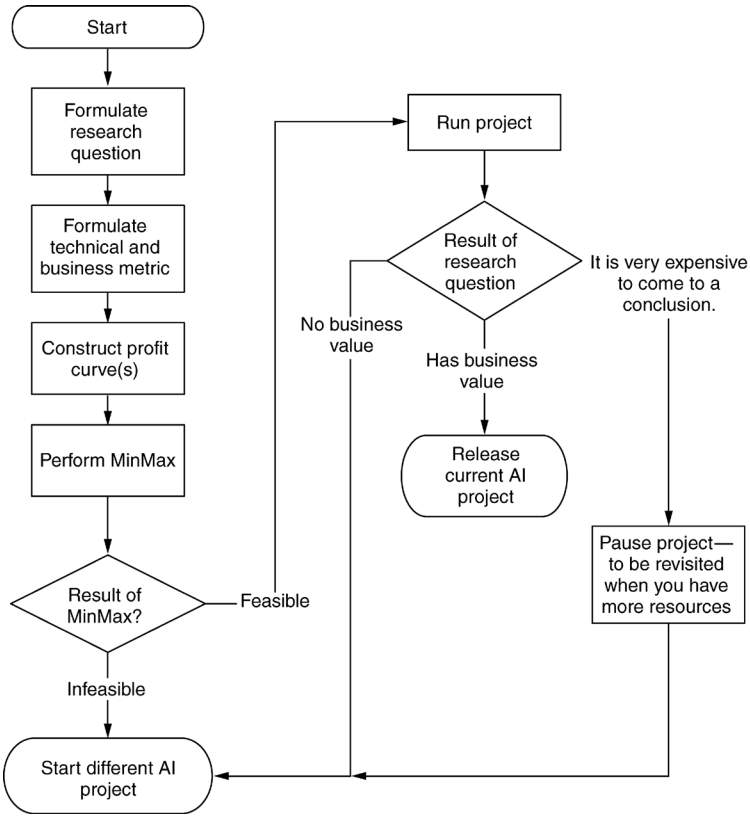


**Figure 7.13 Running an AI project with the ability to put challenging projects on hold. This approach allows you to quickly cut your losses on projects that prove to be more difficult than anticipated and, instead, allows you to try a more straightforward project.**

Down the road, when you have a successful solution that you're looking to improve later with more resources, you might decide that some of the maybes are worth a second look. Figure 7.13 summarises the process of running a project using this three-state, yes/no/maybe classification of results.

## 7.6 Exercises

The questions in these exercises refer to the ML pipeline in figure 7.14, which is a reproduction of figure 6.10 (and figure 7.1).

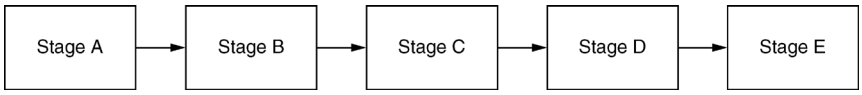


**Figure 7.14 An example of an ML pipeline. We will perform sensitivity analysis of this pipeline. (This is a repeat of figure 6.10 for the reader's convenience.)**

**Question 1:** This question gives you the results of the sensitivity analysis for the pipeline in figure 7.14. Assume that the business metric is profit and the value threshold is USD 2 million/year. The results of your MinMax analysis are the Min part being USD 1.9 million/year and the Max part being USD 3 million/year. You decide to perform a sensitivity analysis. Why is it necessary to perform the sensitivity analysis? You've worked on all the stages for a while, and you've reached a point where it's more and more challenging to improve any of the stages. Determine in which stage of the pipeline you should invest if the results of the sensitivity analysis are as follows:

- Stage A would require six months to improve by 1%. When you improve stage A, the overall improvement in the ML pipeline will be USD 10 K/%.
- Stage B would require two months to improve by 1%. When you improve stage B, the overall improvement in the ML pipeline will be USD 200 K/%.
- Stage C would require one year to improve by 1%. When you improve stage C, the overall improvement in the ML pipeline will be USD 800 K/%.
- The ML pipeline doesn't show any appreciable improvement in results when stages D and E are improved. When does such a situation occur in practice?

**Question 2:** This question gives you the results of the sensitivity analysis for the pipeline in figure 7.14. Assume that the business metric is profit and the value threshold is USD 2 million/year. The results of your MinMax analysis are the Min part being USD 1.9 million/year and the Max part being USD 3 million/year. You decide to perform a sensitivity analysis. You haven't constructed any prototype or tried to clean the data. Determine in which stage of the pipeline you should invest if the results of the sensitivity analysis are as follows:

- Stage A would require three months to improve by 2%. When you improve stage A, the overall improvement in the ML pipeline will be USD 200 K/%.
- Stage B would require two months to improve by 1%. When you improve stage B, the overall improvement in the ML pipeline will be USD 100 K/%.
- Stage C would require one year to improve by 1%. When you improve stage C, the overall improvement in the ML pipeline will be USD 800 K/%.
- The ML pipeline doesn't show any appreciable improvement in results when stages D and E are improved.

**Question 3:** Your AI project is investigating if, by installing an IoT sensor to monitor a vehicle's sound, you'd be able to determine what kinds of changes in tone would indicate a mechanical problem in the vehicle. You've deployed a sensor in 150 vehicles and waited for a month. Only a single vehicle had a mechanical problem. After the month long investigation, your data scientists tell you that from the data collected, they can't predict breakage of the vehicles, and that a single broken vehicle is an insufficient dataset. Does this mean you can't make an AI that can predict vehicle breakage?

**Question 4:** Suppose you have two ML pipelines. Your business metric is revenue. The value threshold is constant at USD 10 million/year. You have two parallel teams that could work on both ML pipelines. Pipeline 1 would deliver USD 20 million/year, andspipeline 2 would provide USD 30 million/year. The cost of the team to develop the pipeline is small compared to the lifetime profit expected from the AI project. Your organisation can implement pipeline 1 in four months and pipeline 2 in one year. Determine which of the two pipelines you should release, and when. Also, draw a timing diagram showing these two pipelines.

## Summary

- Sensitivity analysis answers the question, "In which stage of my ML pipeline should I invest?" There are two forms of sensitivity analysis: local sensitivity analysis and global sensitivity analysis.
- Local sensitivity analysis is applicable when you believe you can improve a stage of the ML pipeline only a little.
- You should perform global sensitivity analysis when you think that a stage in the pipeline could be significantly improved.
- CLUE is an integrated process that addresses important considerations of managing an AI project. Each part of the CLUE process depends on the previous sections of CLUE, so you must perform the ***C***, ***L***, ***U*** *and* ***E*** in order. To make informed decisions based on data, you need a process such as CLUE.
- On a long-running project, the business value of your solution changes with time, so you'd need to construct multiple profit curves to account for value at different times. You can use timing diagrams to visualise how the business value of your ML pipelines evolves over time.
- The answer to your research question isn't limited to yes/no. It could also be, "Unknown at this time with the resources we can devote to answering the question." Don't be afraid to put such a project on hold and revisit it at a later date.