Bridging Decision Problems

Volume I: Framing the Problem

Warren B. Powell

Bridging Decision Problems

Volume I: Framing the Problem

Warren B. Powell

January 2026

Bridging Decision Problems Volume I: Framing the Problem

First Printing, 2026

Published by:
Warren B. Powell
Princeton, NJ 08540
https://tinyurl.com/BridgingDecisionProblems/

ISBN: 000-0-00-000000-0
Library of Congress Control Number: 2026999999

Cover design by: Tiago Romeu - Freire
Interior layout by: Warren B Powell

Printed in the United States of America

Contents

Preface

Problem solving generally starts with unstructured situations that emerge because of a desire to improve performance in some way. Problems do not always need mathematical models, but we are going to use the structure of mathematical models to guide how to think about problems. The degree to which problems need simple models on computers (think of spreadsheets), or more sophisticated models, varies from one problem to the next. Ultimately, we are going to use a very general modeling framework to guide how we think about problems.

For most of my career, I approached problems as shown in Figure 1. I would start with a physical problem, translate it to a mathematical model, and then implement it on the computer. The problem with this approach is that I would start with a particular mathematical modeling framework, and then collect the data needed to fill in my preconceived framework.

I have come to learn that this approach breaks down in the context of the richer problems that arise in practice. The limitations of my initial approach would reveal itself as it became clear that the model was not solving the real problem. This would start an iterative process that I called "From the lab, to the field, and back."

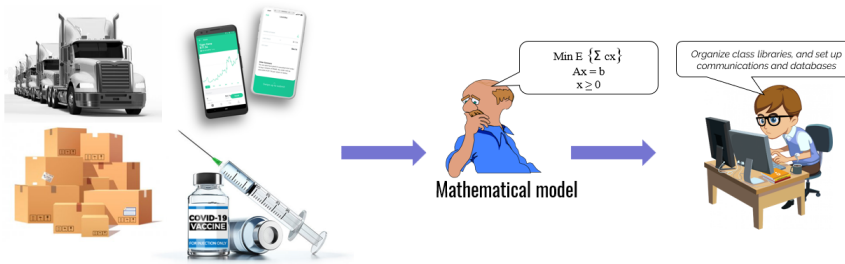


Figure 1: The bridge between the real world and the computer is a mathematical model.

As I tackled an ever-evolving range of problems, I managed to develop a more flexible modeling framework that captures a much richer set of problems that I now call sequential decision problems. This shifted the emphasis from modeling a particular problem to understanding the problem before the modeling process was started.

The process of understanding complex problems such as those that arise throughout business, but also health (especially public health), energy and some financial applications, has long been recognized by consultants who approach the problems under the umbrella of terms like "decision analysis" where they recognize a step they call "framing the problem." "Framing" is a familiar concept in the modeling community, but it has typically been used in a general way to describe a problem. In this book, we give "framing" a very specific definition that is independent of any tool.

Professional optimization specialists, on the other hand, approach the process as I used to, looking to fill in the blanks for a well-defined modeling framework that would fulfill the requirements of optimization software for solving any of a number of (typically deterministic) optimization problems. The problem is that the modeling process is performed within the limitations of the optimization software. The most prominent limitation has been the handling of uncertainty which is pervasive in real applications.

This book is motivated by the development of a very general modeling strategy we call the "universal modeling framework" which was designed to represent any sequential decision problem. The framework covers a virtually unlimited range of problems (static problems are just one special case), justifying the use of the adjective "universal." The defining philosophy of this framework is given by:

Model first, then solve.

This means we need to model the problem before deciding how to "solve" it (by this I mean, how to make the decisions that allow us to achieve an improved solution).

Sequential decision problems have been approached in the research literature by over a dozen distinct communities, using eight different notational systems, a number of modeling frameworks and a range of tools motivated by applications from different problem settings. Every one of these communities assumes that we are solving a problem with a well-defined objective function, pre-defined decisions, and a clear model of the uncertainties (which might mean ignoring them altogether). Even when we know this, different problem settings can have very different characteristics such as the dimensionality of the decisions (binary, discrete actions, continuous or integer vectors), the properties of the objective function (linear, convex, nonconvex), and the behavior of the problem that determine how decisions need to be made (simple rules or models that plan into the future).

This volume makes none of these assumptions. Most important, we do not even start with a mathematical model. Instead, we develop a process

that we refer to as "framing the problem" which consists of asking questions where the answers form the foundation of the mathematical model in the universal modeling framework. We design a framing process that, as with other efforts at framing, starts in English. But rather than the very general language of consultants wading through complex business problems, our framing process is guided by the universal modeling framework which provides a direct path to a computer model.

Our universal modeling framework has three critical features:

1) We start with the premise that any analysis activity is trying to improve one or more quantifiable metrics.

2) We next observe that any improvement requires making better decisions. These may come in any form, and while they may be static (we just make a decision once), decisions are typically made repeatedly over time, which makes it a sequential decision problem. We can handle *any* sequential decision problem.

3) The framework does not make any preconceived assumptions about how decisions will be made. Decisions are made with methods called "policies." We describe four classes of policies which, combined with hybrids of two or more classes, include *any* method for making decisions, including whatever method is being used in practice.

The modeling framework is so general that it allows us to approach any problem that involves making decisions. Instead of solving the math models created in the various communities working on stochastic optimization, we need a more general vocabulary that can capture the richness of all sequential decision problems. Ultimately, we are going to create a framework where we can formulate optimization problems that have never been formulated before.

To accomplish this, we are replacing figure 1, which steps from problem to model to software, to the one depicted in figure 2, where we start with a general statement of the original problem described in the language of domain experts. We then transition to the framing step, which involves posing a series of questions in English, but which produces answers that translate directly to the universal modeling framework, without any preconceived notion of how we are going to solve the problem.

There are three stages in the process of designing and implementing models, but this volume will focus on the first one that starts by looking for answers to three questions:

1) What are the performance metrics?

2) What types of decisions are being made (and who makes them)?

3) What are the sources of uncertainty that affect the performance of the system?

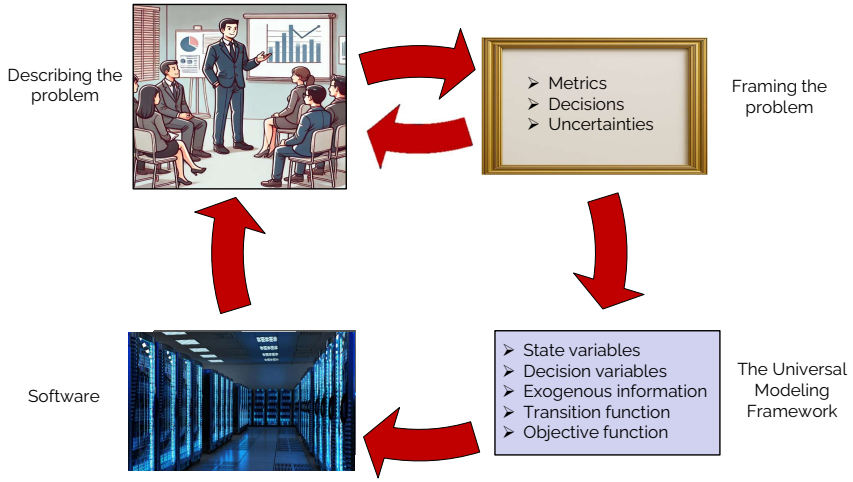


Figure 2: The modeling cycle: From problem, to framing, to model, to software, back to problem.

By insisting that this entire discussion be made in English, we avoid the subtle assumptions that insert themselves when using a mathematical model. However, in Volume II, we will show how to convert the answers to these questions into mathematics. We are going to use the structure of the universal modeling framework to structure the process of understanding the problem. The generality of the framework insures that we are not filtering out important dimensions of a problem, beyond the requirement that there are quantifiable performance metrics and decisions that affect some or all of these metrics.

Acknowledgments

This volume has grown out of a lifetime of research developing models and algorithms for a wide range of complex problems. This work could not have been done without the help of over 60 graduate students, post-docs and my valuable research staff. This volume has also been informed by many of the over 200 senior theses that I have supervised at Princeton University. This research laid the foundation for the universal modeling framework for modeling any sequential decision problem which is described in my 2022 volume (Powell 2022).

The idea of starting with the step of answering three questions (about metrics, decisions and uncertainties) has its origins in a talk I gave at a workshop for Professor Steven Platt at the University of Loyola in 2022, where I presented my universal modeling framework to a group of (very nonanalytical) business executives.

It was not until 2024 where, working with the Department of Supply Chain Management at Rutgers University, I realized that this department (also with nonanalytical students) could provide the incredibly valuable role of framing the problem which is a critical bridge between real applications and my modeling framework. I am grateful to Professor Lian Qi, chair of the Rutgers SCM department, for giving me the opportunity to interact with his students and faculty.

The last step was meeting Adam DeJans at Toyota (through LinkedIn) who demonstrated a real command of my universal modeling framework. From his position at the world's largest automotive manufacturer, Adam recognized its usefulness in the diverse and complex problems faced by Toyota. He invited me to give a talk at Toyota, and it was the preparation for this talk, to a broad audience including top-level executives, that filled in some of the remaining pieces that are reflected in this book.

Warren B. Powell
Princeton, New Jersey
January, 2026

Chapter 1

The Fundamentals of Framing

Humanity is comprised of a variety of processes, each of which encompasses a range of activities which can be evaluated in terms of one or more performance metrics. It seems to be a fundamental characteristic that people always want to do better. Athletes want to be faster or stronger; companies want to be more profitable; health professionals want to save more lives; the power grid wants to provide electricity at lower cost.

This book will be defined by the following statement:

If you want to run a better { *Supply chain*, *Energy system*, *Health system*, *Business process*, *Transportation system*, ..., *Anything* } you have to make better decisions.

We work from the premise that we are always working to improve things, and we can only do so by manipulating elements we control, otherwise known as *decisions*.

Figure 1.1 lists a range of human activities, each followed by a short list of metrics that might be used to evaluate performance (the list of metrics can be quite long). These applications hint at the universe of problems where we "want to do better" but the challenge has been creating a step-by-step path that leads to improved performance.

All real-world problem settings need to start with an unstructured, "plain English" description. By contrast, any mathematical model assumes that the problem has already been structured into a form that can be understood by a computer. What is missing is the inputs of people who actually

- Energy systems – Reduce cost, minimize outages
- Public health – Minimize deaths, maximize productivity
- Business applications – Maximize profits, minimize costs
- Supply chain management – Minimize costs, maximize revenue/productivity
- Manufacturing – Minimize cost, maximize yield, minimize defects
- Economics – Minimize inflation, maximize growth and employment
- Finance – Maximize returns, minimize risk
- Transportation systems (public) – Maximize coverage, minimize cost
- Freight transportation – Minimize cost, maximize service, meet driver needs
- Engineering – Maximize strength, minimize cost, maximize performance
- Drug discovery – Minimize deaths, negative health outcomes, cost
- Sports – Maximize wins, minimize player payrolls, maximize attendance
- Entertainment – Maximize views, minimize costs

Figure 1.1: This is a sample of the many problem settings with examples of objectives that capture performance.

understand the problem, creating the gap depicted by the unfinished bridge that is the front cover of the book.

Standard modeling practice today typically involves someone who is familiar with a "decision technology," which might be integer or nonlinear programming, or it might be machine learning, or Monte Carlo simulation (today we could throw in large language models which is technically a form of machine learning). When a company reaches out to an expert (whether from industry or academia), they will have an immediate tendency to view the problem from the perspective of their own expertise.

The technical expert will then ask the questions that fit their skill set. The integer programmer will ask about decision variables and a cost function; the machine learner will focus on unknown quantities that have to be estimated or forecasted; the simulation expert may identify design decisions that have to be evaluated using simulation.

This behavior is a form of bias that we call *expertise filtering*: learning about the problem in a way that reflects their expertise. It happens in virtually every project, since the domain expert will not have the expertise to identify the most appropriate technical expert. The technical experts always assume that their expertise is relevant, and look at problems through the lens of their training. This is not an issue of deception; it is simply human nature.

We take the position that all "problems" are motivated from a desire

to improve a process in some way. To improve a process requires making changes which are the result of decisions, and we would like to make better decisions. This perspective seems to turn every problem into an optimization problem since we always want to make the best decisions. This does not mean that we are going to use optimization tools. We do not even presume that we are going to do any formal analysis at all, but we will always keep the door for this open.

We are going to use a much more holistic process toward improving a process. We start by replacing the initial step familiar in the optimization community called "modeling" (translating real problems to mathematical models) with a step we call "framing the problem" which precedes modeling. "Framing" is a well-worn term in business problem solving, but we are going to give it a much more precise meaning.

Our version of framing will be a process that requires training people to ask the right questions which are easier to understand by domain experts (business people, health professionals, scientists, engineers), and which fill in specific elements of a mathematical model *if one might be required to solve the problem.* Framing should not be done by a technical expert precisely because of the risk of expertise filtering. However, our approach will result in answering questions that would be needed in the use of any analytical tool. We believe that our process of framing will, for many applications, introduce clarity that may help solve the problem even without a computer.

1.1 What is a "problem"?

Before we solve a problem, what do we even mean by a "problem"? While there are many varieties of problems, from the perspective of making decisions, we are going to identify two styles:

- Decision-focused problems – These are problems where the decisions we are making are clear:
  - Routing trucks
  - Ordering inventories
  - Pricing a product
  - Choosing a medical treatment
  - Choosing a battery storage technology

  - Locating a facility
  - Where to advertise a product, service or candidate
  - Choosing which state to visit (running for office)

- Metric-focused problems – These typically arise in more complex situations where we know what we want to achieve, although we may not initially know what decisions can be made to improve the metrics. Some examples of metric-focused settings are:
  - Lowering costs, raising revenues or improving profit margins.
  - Reducing inventories
  - Improving the yield of a manufacturing process
  - Reducing infections
  - Maximizing financial returns
  - Reducing risk
  - Improving utilization of people, equipment, and facilities
  - Maximizing vote counts (running for office)

Metric-focused problems are generally more complex, since goals are easier to state than the decisions required to achieve a goal. Often we may not even know what decisions might be used to help improve the metrics. Indeed, identifying the decisions that have the biggest impact on the performance metrics is an important step in framing a problem.

At the same time, identifying the right metrics can also be an important step in framing a problem. In fact, in settings with multiple decision-makers (as would happen in any organization), an important type of decision by a manager can be choosing the metrics to evaluate people and business units lower in the organizational hierarchy.

1.2 Settings for decision problems

Decision problems can arise in a number of ways:

- We need to make decisions to solve a particular problem at hand that just needs to be solved once.

- We have a well-defined set of decisions, and simply want to do better. In most cases, the decisions are being made by people, and there may be the hope that computers could do better.
- We want to improve our performance over time, especially when we are not meeting expected targets. For these problems, we may not even know in advance what decisions affect performance.
- We have a well-defined set of decisions being made by people, and we would like to automate the process to remove the manual component, possibly as a form of cost reduction (not having to pay for the people), or to gain more control over a process.
- We are simulating decisions for the purpose of planning the system in the future. This could support strategic planning applications, or understanding the impact of decisions made now on the future.

Any of these represent a perfectly sound motivation for identifying a problem to be solved, or an opportunity for improvement. A major challenge is making sure that you are paying attention to the right metrics, and then identifying all the ways you can influence the metrics. Anything that you control falls in the category of a decision.

1.3 The three stages of decision automation

An article in the USA Today during the COVID pandemic des-cribed the problem of distributing vaccines as "mind-boggingly complex." The reason for this statement is that people do not know how to think about complex problems. It is sometimes helpful to remember that medieval cathedrals were designed and built by people with no formal education; the problem with vaccine distribution is not the complexity - it is knowing how to think about it.

Figure 1.2: A medieval cathedral.

What has been missing is a structured way for thinking about how to make decisions over time. Our process involves breaking down the process of automating decisions into three stages given by:

Stage I Framing - Here we identify the core elements of a decision problem which begins by first answering the three questions:

1) What are the performance metrics?
2) What types of decisions are being made, and who makes them? We make decisions with a method we call the *policy*.
3) What are the sources of uncertainties that affect performance?

Stage II Modeling - The next step is to fill in the details of the universal modeling process. This starts by answering the following questions:

4) How do we make decisions? This done using a function we call the "policy." These will be designed from four classes of policies (given in Volume III).
5) What information is needed? This makes up the elements of our *state variable* (alternatively called the "state of knowledge") which consists of the information required to:
   i) Make a decision (which depends on the policy).
   ii) Compute any performance metrics.
   iii) Any other information needed to compute (a) and (b) in the future.

   Information can be divided between:
   i) What we know perfectly about quantities of physical and financial resources.
   ii) Parameters and functions used for various purposes.
   iii) What we have to estimate and represent in the form of beliefs.
6) How does the state variable evolve over time?

Stage III Implementation - This ranges from acquiring the information needed to implement and evaluate the decisions. This includes:

7) How do we acquire the information that is needed? There is information that is immediately available, information that has to be acquired from other sources, and information that has to be estimated (or forecasted).
8) How do we implement the decisions that we make using the policy?

9) How do we evaluate how well the decisions perform in the field?

We describe the three stages in the sections that follow.

1.3.1 Stage I: Framing the problem

We refer to the initial stage of the automation process as "framing the problem" which consists of answering the following questions:

1) What are the performance metrics?

2) What types of decisions are being made (and who makes them)?

3) What are the sources of uncertainties that affect the implementation of the decisions and performance of the system?

These three questions are not enough to solve a problem, but they are the starting point for any process that involves making and implementing decisions.

It helps to illustrate these questions for the setting of one of the most challenging games ever invented: playing chess (see figure 1.3). Answering our three framing questions is given by:

1) Performance metric – Winning the game.

2) Decisions – Allowable moves.

3) Uncertainties – Opponent's moves.

Figure 1.3: Playing chess may be hard, but it is very simple to model.

Of course, being trivial to model does not make playing chess easy, but chess was long used as a benchmark for demonstrating the power of algorithmic strategies such as "reinforcement learning."

Solving the problem arises in Step 4 (Stage II), and while this is quite hard, the remaining steps are also trivial.

Now consider some of the problems that we will identify in chapter 2:

- How do we reduce deaths due to fentanyl?
- How do we design a supply chain to minimize costs that is robust to different sources of uncertainty?

- How do we manage a fleet of trucks to maximize profits while providing on-time service?
- What is the best strategy for reducing CO2 emissions?
- How should a large manufacturer store and invest its money to maximize returns, while managing risk and meeting short and medium-term cash requirements?

Answering our three framing questions for these problems is a nontrivial exercise. For this reason, we have three chapters dedicated to the process of answering each question:

- Chapter 3 – Performance metrics
- Chapter 4 – Decisions
- Chapter 5 - Uncertainties

These issues are illustrated using the applications in chapter 2. We provide a brief peek into these three core elements by describing different types of metrics, decisions, and uncertainties in the subsections that follow.

1.3.1.1 Types of metrics

Metrics come in an endless variety and depend completely on the context.

- Business – Businesses are characterized by long lists of financial metrics, productivity metrics, performance metrics, labor metrics, and metrics capturing how the market is being served.
- Health – Illness/death, cures, side effects, mobility, strength, cost.
- Energy – Cost, quantity of energy provided, outages, demand curtailment.
- Manufacturing - Yield, product performance, speed, cost.
- Drug discovery - Performance, patent protection, market potential, side effects, health risks.
- Sports - Points scored, consistency, fan popularity, injuries, consistency.
- Freight transportation - Revenue, cost, service, labor requirements, exposure to market volatility.

Choosing the right metrics is a challenge of its own, whether you are using them to guide the behavior of a computer model, or to guide the behavior of people.

Separate from what a metric is measuring is how it is being used to guide the performance of the system. Metrics can be used in three different ways:

- Objectives - These are metrics we want to maximize or minimize.
- Targets - We may want the metric to get as close as possible to a target number, such as the temperature in a building or a patient's blood pressure.
- Limits - We may want a metric to stay under or over some limit. For example, we may want to keep a patient's blood sugar under a particular value; stockouts should stay under a certain level; financial portfolios need to keep volatility under a specified value.

1.3.1.2 Types of decisions

An initial list of different types of decisions is given by:

- Binary – These arise when we are choosing between two webpage designs (known as A/B testing), determining when to sell an asset (at each point in time we can hold or sell).
- Discrete choices – It helps to divide this category into three classes:
  - A small set of discrete choices - We might need to choose the best drug, the best supplier for a component, or the best location for a facility.
  - A set of discretized values of a continuous parameter – Examples are prices, dosages of a drug, or temperatures for baking a semiconductor wafer.
  - In some cases the number of discrete choices may be quite large, such as choosing which of 30,000 different molecules that might be used for a drug, or the choice of locations for different facilities spread among 100 different possible locations.
- Continuous choices – Prices, concentrations, dimensions, temperatures, ... These can be scalars (that is, a single parameter), or

vectors, where we might be optimizing across multiple (potentially many) continuous parameters.

- Vectors of discrete choices – We may have a set of M drivers that we are assigning to N loads, where we have to decide whether to assign driver m to load n.

A second dimension of decisions involves the timing of when a decision made now is implemented in the future. For example:

- A dispatcher assigns a driver to a load to be moved right now.
- A doctor may assign a blood sugar medication that requires several hours to take effect.
- A grid operator will plan today which steam generators should be running tomorrow.
- A supply chain manager places an order that may take several weeks or months to arrive.
- An airline may order new aircraft that may take two years to deliver.
- An investment with a private equity firm may tie up that capital for 8 to 10 years.

A third dimension of decisions involves identifying who makes a decision.

- The management of vaccine distribution involves decisions starting with federal and stage agencies, extending through hospitals, physicians and nurses that administer the vaccine.
- The manufacture of automotive engines involves the participation of a sequence of manufacturers who provide the materials, and make the various components of the engine, ultimately be moved to market through dealers that control the orders of cars.
- Clinical drug trials involve decisions by scientists, regulators, companies providing the funding, hospitals and clinics that administer the drug, and the patient.
- A trucking company performs dispatching using a team of dispatchers and load managers, which might be replaced with a single computer model that can coordinate these decisions throughout the company.

1.3.1.3 Forms of uncertainties

Arguably the most subtle aspect of making decisions involves understanding the uncertainties that invariably arise when implementing decisions in the field. Not surprisingly, the forms of uncertainties that arise is heavily dependent on the context. Some examples include:

- Financial trading – Here we are primarily interested in changes in asset prices, but traders are also interested in demand for assets, and changes in other metrics that might suggest where markets are headed such as changes in unemployment, interest rates, retail sales. Markets often move with expectations, which are notoriously hard to measure.

- Supply chain management – Here we have to deal with uncertainties in the market demand for a product, the strategies of competitors, the performance of suppliers, and the behavior of workers (especially when unionized). In addition there are outside influences such as weather, earthquakes and the spread of diseases.

- Public health – The spread of disease is a function of the source of the disease (this may be a single infection, or from many animals who were infected, from which a human strain evolved), the prevalence of the disease, the rate of transmission, how it affects patients, the development of drugs, the distribution of the drugs, and the response of the public in the acceptance of the drugs.

Each of these examples involves multiple sources of uncertainty. These feature different forms of uncertainty, such as:

- Fine-grained noise such as daily random demands.

- Changes in prices and weather typically exhibit spikes and bursts.

- There may be unexpected shifts to new plateaus reflecting shifts in technology, consumer behavior, or competitor decisions.

- Single, rare events such as an earthquake, or invention of a major new technology.

- Contingencies for events that might happen, but which have never actually happened

The consideration of uncertainty has to be considered in terms of how it affects the performance metrics. When we are making decisions in the presence of uncertainty, we have to make choices, such as how to make a decision, that work well on average given that we do not know what is going to happen in the future.

However, some forms of uncertainty introduce a new dimension called risk that captures factors that are not present in the performance metrics if uncertainty did not exist. Risk is a very popular topic in fields such as finance, supply chain management and health. There are many books that talk about risk, along with very sophisticated papers that model risk, without ever providing a formal definition of risk. We will provide this definition in chapter 3.

1.3.2 Stage II: Modeling

Our initial three questions (performance metrics, decisions, uncertainties) lay the foundation for what we are going to call our *universal modeling framework* (or UMF). The UMF can be used to model *any* decision problem, especially when we used the extended version to handle multiagent problems. For now, we place emphasis on capturing the evolution of decisions and information over time. In Volume II, we will describe the UMF using full mathematical notation (which is not as bad as it sounds), but for now, we are going to sketch it in plain English.

The Universal Modeling Framework consists of five elements:

1. State variables capture all the information we need to make decisions and compute our performance metrics. An understanding of the elements of a state variable informs the process of what information is needed to make decisions.

2. The decision variables represent what decisions we might make (building on the types of decisions we described when framing the problem). Note that we assume that we make decisions with some "policy" *which is to be designed later.*

3. The exogenous information is any new information that arrives after we make a decision, and before we make our next decision.

4. The transition function which describes how the state variable changes given what decision we have made, and given the exogenous information that arrived after we made a decision.

5. The objective function describes how to evaluate the performance of the system using the method we have chosen for making decisions.

Identifying the state variables requires choosing the method (called the policy) for making decisions, so this has to be done first. However, evaluating and tuning policies requires the entire universal modeling framework if we are going to use a simulator. Ultimately, the design of policies (which plays a major role in determining what information we need in the state variable), and the evaluation of the policies is an iterative process.

The universal modeling framework is covered in much more detail in Volume II where we introduce very basic mathematical notation.

If the universal modeling framework sounds obvious, it is. It is little more than a framework describing the evolution of what we know (the state variable) by decisions (which we control) and the exogenous information (which we do not control). What is perhaps stunning is that this is not standard in the research literature, although there are pockets where it can be found.

Arguably the most difficult step is designing the policy for making decision. We separate the process of evaluating a policy from designing the policy which differentiates our approach from that used in virtually every book on stochastic optimization. Fortunately, we have a strategy for overcoming this complexity.

1.3.3 Stage III: Implementation

Easily the most widely overlooked dimension in the design and solution of optimization models is the process of implementing them. The academic literature utterly overlooks that what counts is not how well we solve a problem in the computer, but rather the impact of decisions when they are implemented.

The key dimensions of implementations covers three areas:

1) Acquiring the data needed to fill out the state variable, which means the information we need to make decisions and compute the performance metrics (more details are provided in Volume II).

2) Implementing the decisions, which might mean getting people to follow instructions, or communicating the instructions electronically.

3) Evaluating performance. For complex systems, understanding how well

the system is performing, which presumably is affected by the decisions being made, can be quite difficult.

For complex problems in industry, implementation can be an exceptionally challenging process. Even if this is viewed as outside the scope of the modeling process, it helps for modelers to think about these steps. It may be that some decisions are simply never going to be made by a computer. For example, allocating resources in a public health setting involves negotiating between state, regional and local organizations which each have their own hidden priorities.

1.4 Three types of information

We first recognize that any quantity that can be represented on a computer is a form of information. We can identify three types of information from the perspective of how it evolves over time:

1. The information that we know at the time we make a decision which constitutes the state of our system (more precisely the state of knowledge). This is the information needed to make decisions and/or to compute the performance metrics, now or possibly in the future.

2. New information that we control. We define decisions formally in chapter 4 and describe six different types of decisions which helps in the process of identifying decisions.

3. New information which arrives from outside of our system beyond our control, although it may be influenced by our current state and/or the decisions we make. We call this exogenous information, and it can come from a number of sources:

- Natural phenomena such as weather and earthquakes.
- Markets, such as the demand for a product, stock prices and interest rates.
- Population dynamics such as the spread of diseases.
- The behavior of other companies or organizations.
- Decisions made by people (more generally, agents) from outside of our system, such as:

  - The production decisions of suppliers of inputs to a manufacturing plant.
  - Other divisions within a company (such as pricing and marketing, if we are in manufacturing or inventory planning).
  - The actions of a patient (if you are the physician).
  - Advertisements placed by the competing candidate in a presidential election.

Exogenous information is described in greater depth in chapter 5.

1.5 Decision making as a process

There is a vast literature that focuses on creating "optimization problems" that consist of:

- A decision (or set of decisions).
- An objective to be minimized or maximized.
- Constraints, which determine the set of allowable decisions.

Real decision making is a process, and understanding this process is critical to designing methods for making better decisions. We start by identifying the following elements

1) Sequential decision problems, which describe the process of making a particular set of decisions over time by a single agent.
2) The "information chain" which describes the process of creating the information needed to make a decision.
3) The steps involved in implementing decisions.
4) The process of evaluating performance.

Beyond the scope of this monograph is the challenge of coordinating across multiple decision-makers.

1.5.1 Sequential decision problems

We are now ready to write out, in English, a sequential decision problem, which we can state using:

State, decision, information; state, decision, information; . . . , state, decision, information.

Each triplet {state, decision, (exogenous) information} represents the information associated with a particular time period:

1. "State" is the information we know at the beginning of the time period.

2. "Decision" is our endogenously controllable information.

3. "(Exogenous) information" is the information that arrives after we make a decision, and before we make the next decision.

After we make a decision (sometimes after we observe the exogenous information) we stop and calculate performance metrics.

Of course, not all decision problems are sequential decision problems, although the vast majority decisions are made repeatedly over time. However, we can identify several categories of sequential decision problems from the perspective of the sequencing of decisions and information:

1. Make decision, stop.

2. Make decision, see exogenous information, stop.

3. Make decision, see information, make one more decision, stop.

4. Make decision, see information, make decision, see information, . . . , repeat T times, stop.

5. Make decision, see information, repeated infinitely.

Some comments:

- Category 1 describes, static, deterministic decision problems that have dominated what is known as the optimization community since the 1950s. The simplest version of these problems might involve finding the best of a set of choices, such as purchasing an item from the least cost supplier, as long as we assume that the item will perform exactly as we expect.

  More complex problem instances might involve finding the least cost allocation of supplies from multiple sources to serve different needs, or assigning different people or machines to perform different tasks, introducing the complexity of working in multiple dimensions. The complexity of these problems has led to the astonishing oversight

that the vast majority of applications are actually sequential decision problems, a property that has been completely ignored in the literature on this topic.

- Category 2 describes problems known as stochastic search, which represents one of the most widely studied class of problems. Examples of stochastic search problems include:
  - Pick a set of manufacturing facilities and warehouses, and then run a simulation to evaluate its performance.
  - Pick a treatment regimen for a patient, and then see how it unfolds.
  - Set an investment strategy for a stock portfolio, and then observe how well it works.

  All these can be described by "make choice" and then "see how well the choice works." If this is done in a simulator or a laboratory, we may be able to run these experiments over and over. In this case we have a fully sequential search process that falls in category 4.

- Category 3 describes a more general version of category 2 where we might make an initial decision, such as sending product to a set of warehouses. Next the demands for the product at retail outlets is revealed. Finally, we have the decision of shipping from warehouses to the retail outlets. This problem has been widely studied under the umbrella of stochastic programming.

- Category 4 is the most common form of sequential decision problem since it captures the repeated nature of making decisions followed by learning new information, but we stop after a specific number of time periods, typically for the practical reason that we are running a simulation that has to have a predefined stopping point.

- Category 5 is a popular topic in communities such as dynamic programming (specifically Markov decision processes) and stochastic control. The objective is usually the infinite discounted sum of costs or rewards. This literature typically assumes that the information arriving in each time period comes from the same distribution (this is known as a stationary distribution) and is useful for deriving a variety of theoretical results.

1.5.2 From optimizing decisions to policies

When we are solving a static, deterministic optimization problem, virtually every author represents the decision as a variable (typically a vector) "x" where we have to design an algorithm to find the best "x." By contrast, when we have a sequential decision problem, there is a fundamental lack of an understanding of what it is we are optimizing over. In a nutshell, with deterministic problems we are looking for the best decision x, while for sequential decision problems we are searching for the best function (that is, the policy) which represents a method for making decisions.

The idea of finding the best function for making decisions seems foreign in the optimization literature. By contrast, this is exactly what is done in machine learning, where the challenge is to find a function (often called a statistical model) that does the best job of fitting the data. Moving forward, we will refer to the functions for making decisions as policies, a topic that we address in greater detail in Volume II.

1.5.3 The information chain

Making a decision enjoys certain parallels with making physical products. To make a car (for example), it is necessary to make various parts, which often take multiple steps. Then, after we make the car we have to distribute it to the customer.

Decisions are "made" from information, which itself may need to be created (collected or estimated) through a series of steps. Deciding how many, and what types, of cars to make may require a forecast that is compiled from historical data, as well as economic forecasts and estimates from a sales force. This data then has to go through a set of methods that create the forecasts.

The flow of information is depicted in figure 1.4. "Information" might be observed inventories, a forecast created from history, the result of a decision to collect information through a market survey, or the result of a production planning process. The processing nodes are mathematical functions – sets of equations that act on the inputs to produce an output. The functions may do anything, from adding up numbers to producing forecasts to making decisions by solving an optimization problem.

Just as with physical processes, information processes typically consist of manual steps (such as entering inventories) combined with steps that

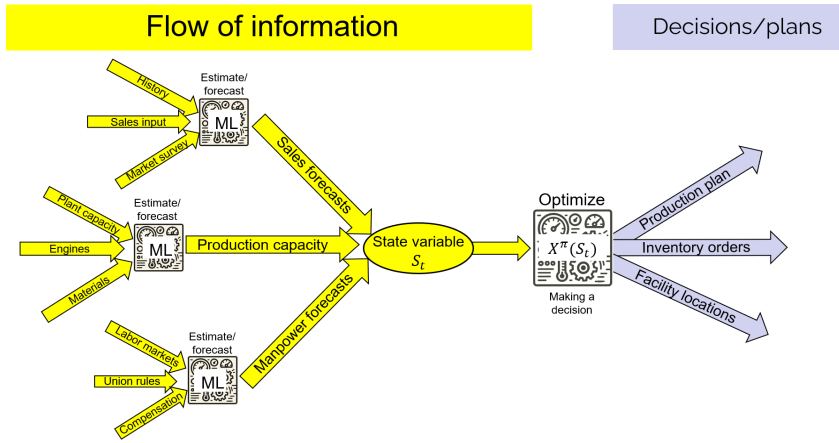


Figure 1.4: An illustration of information flowing from initial source, through stages of processing and estimation, up to the point where it is used to make decisions (another form of information).

are done on the computer (and hence are automated), such as running a forecast.

It is easy to think that, given the extensive use of computers, information processes should be almost completely automated. However, there are still a lot of white-collar workers, and they are not loading trucks or working on an assembly line.

1.6 Artificial intelligence

Ultimately the goal of thinking about a complex problem in a formal way is to use the power of the computer to improve the process. Most people will immediately suggest using "artificial intelligence" (often referred to as "AI"). The problem is that "AI" is a term that has been used since the 1950s, and has evolved steadily over the years, typically being applied to the latest invention coming out of the field of computer science.

We divide the major forms of AI into seven levels, depicted in figure 1.5. After we describe these seven levels, we will organize them into four fundamentally different classes of intelligence.

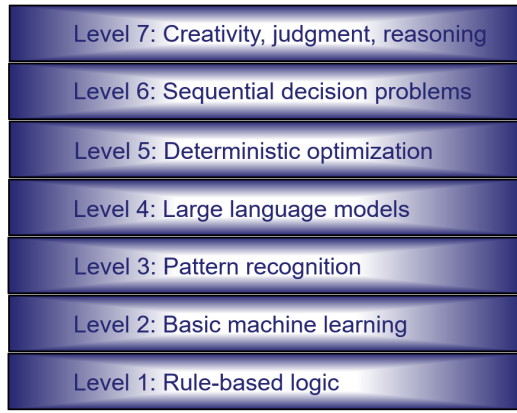


Figure 1.5: The 7 levels of artificial intelligence.

1.6.1 The seven levels of AI

Level 1: Rule-based logic - This first evolved in the 1960s and 70s, and emerged in the 1980s (as computers first became much more widely available) as "expert systems." These consist of rules specified by humans of the form "If {condition} then {action}." For example, the condition might be "eating red meat" and the action could be "drink red wine." Or the condition could be the attributes of a patient (symptoms, gender, age, weight, smoker?, blood pressure, ...,) and the action could be a medical treatment. This form of AI went through what has become known as the "hype cycle" where people fantasized how computers were going to take over the world.

The problem with rule-based systems is that as the number of elements making up a condition grew, the number of possible condition/action pairs increased exponentially (a behavior known as the "curse of dimensionality"). By the 1990s, this early form of AI was widely viewed as a failure, but in fact rule-based systems remain widely used even today. The only failure is that they did not live up to the initial hype. Rule-based systems are widely used today.

Level 2 – Statistics/machine learning - Under development since the early 1900s, statistics (known as machine learning in computer science) is the science of using data to estimate models. We might use observations of different prices of a hotel room to estimate the demand, or historical

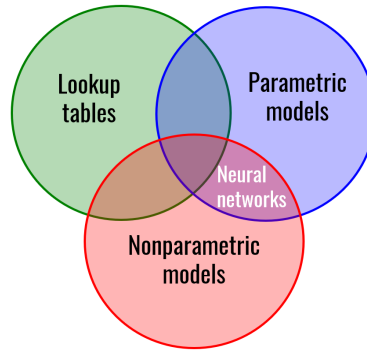


Figure 1.6: Every type of function for machine learning falls in these overlapping circles, including 1) lookup table functions, 2) parametric functions, and 3) nonparametric functions (locally parametric).

demands to forecast the future. This field grew explosively in the 1980s and 1990s (as computers became widely available).

Machine learning models come in a variety of styles, but these can be organized into three broad classes as illustrated in figure 1.6:

Lookup tables – These are of the form "If {input} then {output}," similar to rule-based systems.

Parametric models – These are analytical functions of inputs that produce one or more outputs using a mathematical function that depends on a set of unknown parameters. If the function is linear in these parameters, then this would be a linear model. More general models use functions that are nonlinear in the parameters. Figure 1.7 illustrates linear and nonlinear models.

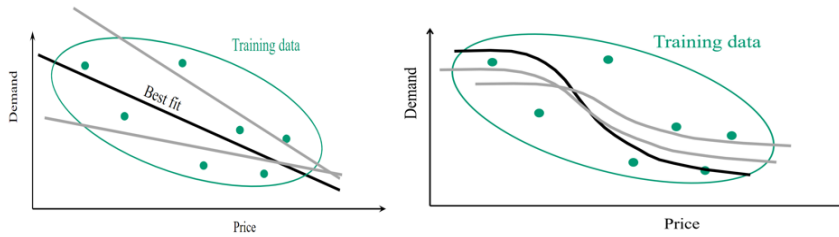


Figure 1.7: Illustrations of linear and nonlinear parametric functions used in machine learning.

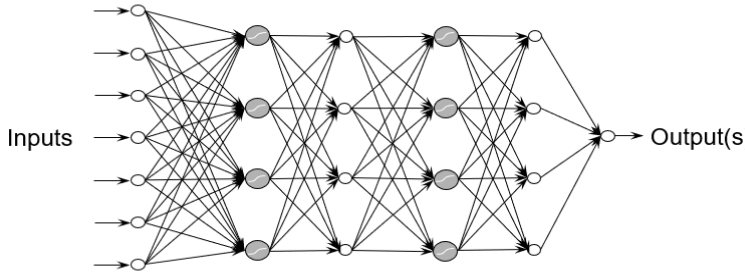


Figure 1.8: Illustration of a (very small) neural network. Each link carries a parameter that has to be tuned so that the output comes as close as possible to the label associated with the inputs in a training dataset.

An important class of parametric models which first emerged in the 1970s is neural networks (see figure 1.8). Neural networks have an input layer, where any set of inputs enter the network through the input nodes. These values are then transformed through the intermediate layers before producing one or more outputs. Each link in the network has a parameter associated with it, where the early neural networks often had thousands to a million parameters.

It is best to think of neural networks as a very high-dimensional nonlinear function that can be used to fit a virtually unlimited set of relationships, but at a cost of requiring large training datasets. Also, their flexibility limits their ability to be used in the presence of noise.

Nonparametric models – These are most easily envisioned as models that are local approximations of a function. For example, we may have estimates of a function at a set of points, and we then use linear extrapolations of these points to provide estimates of points for which we do not have an estimate.

Level 3 – Pattern recognition - The next level of AI emerged from the research community in 2010 addressing the problem of pattern recognition. Pattern recognition is just another form of machine learning that we saw in level 2, which involves using neural networks. However, these neural networks are much larger than the ones used in the 1990s. Instead of many thousands to a million parameters, these neural networks might have 10 to 100 million parameters. These were referred to as “deep neural networks.”

The inputs would be the pixels in an image (or the signals from a voice pattern), which is a much higher dimensional input. The hard part was

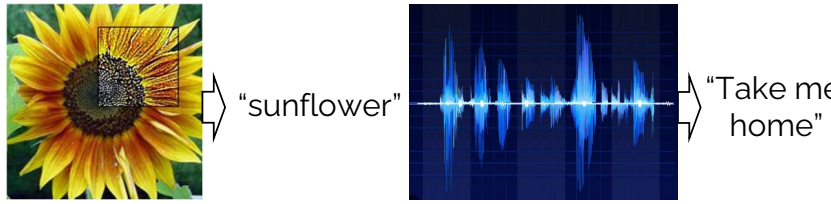


Figure 1.9: Illustration of the ability of a neural network to recognize the picture of a sunflower, or the voice pattern saying "Take me home."

creating a training dataset large enough to perform the parameter tuning. The training dataset had to consist of millions of images (which was easy to find on the internet) with the associated "labels" which identified the image such as "sunflower" or "Take me home" in figure 1.9. The hard part was getting the labels, which had to be generated by people.

The breakthrough in training came when a Princeton computer science professor, Fei-Fei Li, realized that a software environment created by Amazon called the "Mechanical Turk" made it possible to reach out to people around the world willing to work for very low wages to create these labels. In other words, the breakthrough was not so much the underlying analytics (neural networks were developed back in the 1970s) but rather access to enough data at low cost.

Level 4 – Large language models - Level 4 is simply another step up from image recognition, where instead of estimating (or "predicting") the identity of an image, the neural network was taking as input a sequence of words (anywhere from a few words to hundreds or thousands of words) to predict the next word (the models act on word fragments known as "tokens"). It does this by creating a probability distribution of the words that might come next given a sequence of words which could be an initial prompt provided by a user.

Figure 1.10 illustrates this, starting with the prompt "The best way to improve the robustness of a supply chain is to..." The neural network then produces a probability distribution of the word that might come next, based on the training dataset. The large language model (or LLM) then samples from this distribution, in proportion to the distribution. If it chooses the word "design" then the sequence "The best way to improve the robustness of a supply chain is to design..." is input to the neural network, which then

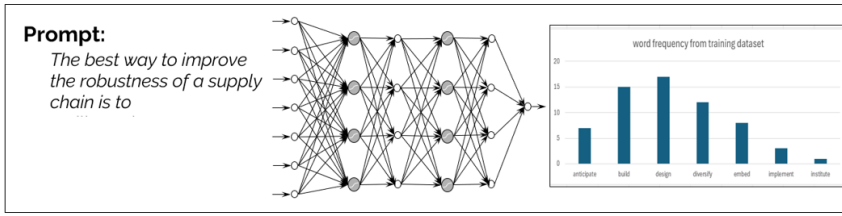


Figure 1.10: Large Language Models such as ChatGPT use a training dataset to construct a distribution of words that might follow a sequence of words, initiated with a prompt, but building on the sequence created by the LLM.

produces another distribution of words. The process of sampling a word, adding it to the previous sequence of words to produce a new sequence is repeated over and over. This is why the process is called "generative AI."

The neural networks used to generate the distribution of "next words" that follow a previous sequence are truly colossal. While a deep neural network for pattern recognition (Level 3) might have 10 to 100 million parameters (the black arcs in the figure), the neural networks used for LLMs might range between 10 billion and 1 trillion parameters.

It needs to be clear from this description that LLMs are not inherently intelligent; they are just mimicking word patterns from a training dataset. They sound intelligent because they are mimicking word patterns that come from an intelligent source (assuming a human wrote the words).

Level 5 – Deterministic optimization – This covers a substantial library of tools for solving hard decision problems. These problems are known as linear programs, integer programs, and nonlinear programs, and they all have the characteristic that a "decision" is a vector, which means it is a set of different decisions (a very large set). Some examples are:

- We may want to decide how much product to send from a set of 10 distribution centers to 200 warehouses, creating a 2,000-dimensional vector that has to be decided.
  item Airlines have to schedule their aircraft and the crews (both pilots and stewarding crews) over extended periods (typically quarterly) to maximize utilization while observing rules for maintaining aircraft along with rules for using people.

- A financial manager may be constantly juggling the allocation of capital among 10,000 different investments, giving us 10,000 buy or

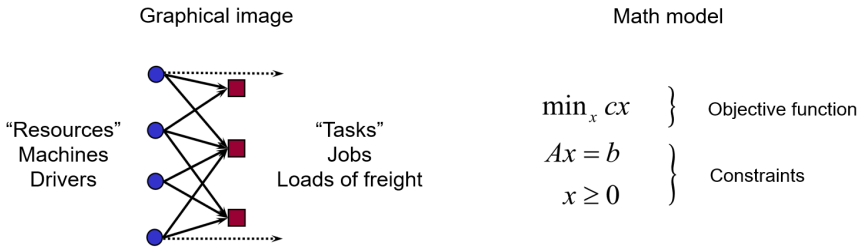


Figure 1.11: Graphical illustration of a type of decision problem (assigning resources to tasks), and its mathematical representation as a linear program.

sell decisions that are made daily.

Often these problems can be depicted pictorially, as we have done on the left in Figure 1.11, but there is a standard way of writing them mathematically, often starting with the notation on the right. While this mathematical notation will not be generally familiar, universities produce thousands of students each year who are trained to model problems in this format. Then, there are many computer packages, some available commercially while others are available for free, that can solve even large scale problems efficiently.

Level 6 – Sequential decision problems – The vast majority of decisions are made repeatedly over time, whether it is every few seconds, minutes or hours, daily, weekly, quarterly or yearly. Even our deterministic optimization problems in level 5 above are usually solved repeatedly over time, but most decisions are much simpler. Some examples of sequential decision problems are:

- Deciding when to sell an asset, and the expected value of holding the asset in the face of dynamically changing prices.
- Replenishing inventory, possibly with very long lead times, to satisfy uncertain demands in a dynamic marketplace.
- Determining the parameters for an automated financial trading policy.
- Choosing the right concentrations of materials, the right temperature for blending and the time to exposure the mixture to each temperature to produce a material with the highest strength.

- Choosing the best drug to treat a patient with specific characteristics.
- Deciding how much energy to store from a combination of wind farms, solar farms and the power grid to meet future loads (demands) at least cost.
- Truckload trucking requires determining which drivers to move which loads of freight.
- Choosing how much to invest in each of thousands of stocks and other investments.

Sequential decision problems arise throughout human processes. A decision may be binary (hold or sell), discrete (which drug), and discrete and continuous vectors. Choosing among the best of a discrete (or discretized) set of choices is easily the most common sequential decision problems, but many involve complex operational problems that arise in business logistics, energy systems, and distribution problems in health.

Level 7 – Creativity, reasoning, and judgment – Level 7 represents the highest level of intelligence. For example, while many of the problems in levels 5 and 6 can be quite complex, they always involve well-structured problems with clearly defined decisions and objectives. Level 7 is where we can pose complex problems such as reducing CO2 emissions, minimizing disease, and creating new products.

It is our firm belief that while many authors will talk about the future of "AI" in terms of replacing people, the reality is that computers will not be able to extend past well-defined problems. One activity that we believe is beyond the capability of computer intelligence (including the hyped skills of large language models) is framing complex decision problems. For this reason, we refer to level 7 as science fiction, something that is fun to talk about but will never actually happen.

1.6.2 Three classes of computer intelligence

The first six levels of artificial intelligence represent different forms of intelligence that can be implemented on a computer, while the seventh, in our opinion, remains the sole domain of humans. The first six levels can be divided into three distinct classes:

Class 1 – Human-specified behaviors – This class includes Level 1 in the seven levels of AI, and it can be used for two different purposes:

- Pattern recognition – A rule may specify that if a patient has a specific set of conditions, then it means they have a specific disease.
- Decisions – Similarly, a patient with a specific set of conditions should take a particular medication (which is a form of decision).

Rule-based logic does not distinguish between whether the rule is making a statement about the state of the world, or an action that should be taken. The conditions behind the rule and its outcome (whether it is a statement of the state or a decision) must be manually specified.

An important characteristic of Class 1 artificial intelligence is what it does not use:

- It does not use a training dataset.
- It does not require a model of the underlying decision problem.

Rules have to be directly specified by people, although it is possible for rules to be specified in a dataset. For example, we might have a dataset listing medical protocols, where for each patient condition a treatment is specified. However, imagine we have a dataset compiled of actual physician decisions which may conflict: different physicians may order conflicting treatments despite having patients with identical conditions. If we use this dataset to learn treatments, that would be an example of machine learning.

Class 2 – Machine learning – This class includes Levels 2, 3, and 4. Machine learning refers to the use of mathematical functions that consist of inputs and a set of tunable parameters that can be adjusted so that the function best matches a set of responses, also called labels (among many other names). Machine learning requires a user-specified function, along with a training dataset that consists of inputs and responses (labels).

While rule-based logic is limited in terms of the complexity of the inputs, machine learning can handle very complex inputs using models that have large numbers of parameters. Linear models can have dozens to hundreds of thousands of variables. Neural networks have been trained for large language applications with over a trillion variables. Of course, larger models require large datasets which has proved to be the major barrier limiting the use of neural networks for the highly complex task of language processing.

Class 3 – Optimization – This class includes levels 5 and 6 which address the problem of choosing the best decision from a set of choices, which may be discrete set or a high-dimensional vector space. Level 5 is

limited to static (deterministic) problems where all the data is known, and we seek the best decision (which is often a vector). Level 6 tackles the complex problem of choosing the best decisions over time which spans an absolutely vast range of problems.

The optimization class does not use a training dataset. Instead, it is necessary to specify a performance metric (often called an objective function) along with a set of equations that describe what decisions are allowable. For sequential decision problems, we also need equations that tell us how information evolves over time.

1.6.3 Summary

The methods in class 2, machine learning, aim at training mathematical functions to behave like a training dataset. If the training dataset consists of images such as breast X-rays along with human-generated "labels" of whether the breast displays evidence of cancer, the trained model will never be able to perform any better than the skills of the radiologists who provided the labels. For this reason, it is possible to say that Class 2 methods (machine learning) teach computers to behave like humans (more precisely, behave like the training dataset).

By contrast, the methods in Class 3 (optimization) are designed to produce decisions that outperform humans. The price of this higher-level performance is that we have to provide what is known as a model of the problem. In particular, these methods require a mathematical model, consisting of:

- A well defined set of decisions.
- A clear performance metric that makes it possible to assess whether one decision is better than another.
- The physics of the problem which describe:
  - What decisions can be made at a point in time.
  - How the system evolves over time.
  - How new information is arriving to the system.

This book addresses class 3, since this covers the methods that address making decisions. In particular, we are going to focus on sequential decision problems, since these are the most pervasive – virtually everyone makes

decisions, and we make them over time, making them sequential decisions. Static (deterministic) problems are just a special case of sequential decision problems, and the solution of sequential decision problems will draw heavily on the tools developed for static, deterministic problems.

Sequential decision problems represent an incredibly rich problem class. Invariably these tools depend on the methods in the first five levels of artificial intelligence. As with level 5 tools (deterministic optimization), we need a model of the underlying problem. However, since sequential decision problems are much richer than the static problems in level 5, the models need to be much richer and more complex, but this is an area where classical mathematical modeling has fallen short.

1.7 Traditional modeling frameworks

It helps to divide the modeling frameworks for making decisions into two broad categories:

- Static, deterministic models that assume that all information is known, where we strive to pick the decisions that work the best.
- Sequential decision models that capture the flow of decisions and information. Since we explicitly model information that arrives after we make a decision, this means that the decisions have to be made before information (presumably relevant to the performance of the decision) has arrived.

In this volume, all sequential decision problems explicitly capture the flow of information, which means that we are making decisions at each point in time before we know the information that may arrive in the future. For this reason, sequential decision problems are fundamentally *stochastic* (the fancy term for saying the future information is random).

1.7.1 Static, deterministic models

The literature for modeling static, deterministic problems is quite mature, with a substantial base of software built around variations of an optimization model that can be written:

$$\min_{x,y} C(x,y), \tag{1.1}$$

subject to:

$$g(x,y) = 0, \tag{1.2}$$
$$x \geq 0, \tag{1.3}$$
$$y \in (0,1). \tag{1.4}$$

We have allowed for the presence of both continuous variables x (that might take on a value such as 0.56) and discrete variables y that must be 0 or 1.

What happens when modeling deterministic optimization problems (level 5) is that we take the mathematical model given by equations (1.1) - (1.4), and we then go to the physical problem and fill in the elements of the model, which requires identifying the decision variables, the objective function, and the constraints. Imagine having a hammer and looking for nails. The tool is useful, but the process requires fitting the problem into the modeling framework.

Deterministic optimization has long emphasized the challenge of designing tools to find the optimal decisions given a model, with secondary attention given to creating the model itself. Note that the modeling framework does not provide any mechanism for capturing the evolution of decisions and information, or anything related to how decisions are organized.

1.7.2 Sequential decision models

Traditionally, the literature for sequential decision problems has tried to follow the same approach, but it has failed completely. In contrast with the well-defined modeling framework for deterministic optimization represented by equations (1.1) - (1.4), the optimization literature has not adopted a standard modeling framework for sequential decision problems. As of this writing there are over a dozen different communities using eight different notational systems, with fundamentally different styles for expressing what problem is being solved, or what we are solving for. For example, some communities will write out an objective function as is done in deterministic optimization, others write out a policy, and others will write out an optimality condition.

Our approach depends on the universal modeling framework sketched in section 1.3.2 which can be used to model *any* sequential decision problem. This modeling framework is described in detail in (Powell 2022) [chapter 9]. This book lays out the model before describing policies for mak-

ing decisions, which is done in chapter 11 (chapter 10 focuses on modeling uncertainty).

Volume II of this series will also cover the dimensions of the universal modeling framework in far more detail than we can in this volume, using modest levels of notation. However, the universal modeling framework cannot be used without answering the three questions addressed in this volume.

1.7.3 The most common decision problem

Decision problems, and in particular sequential decision problems, are an exceptionally rich problem class. However, often overlooked in the literature on optimization is that the vast majority of decision problems are described by figure 1.12, where we might have two choices (take an action or not), or a small set of choices (which drug to use, where to purchase a part), or a large number of choices (which product to advertise, which molecule to use when creating a new drug).

A distinguishing characteristic of this problem class is that while we may have an estimate about the performance of each choice, we are typically uncertain about the performance which will emerge after we make a choice. We may just have one chance at making the best choice, but often we make this decision repeatedly, and can learn from past experiences. There are many variations of this problem:

- Whether we are running offline experiments in a lab or simulator, or

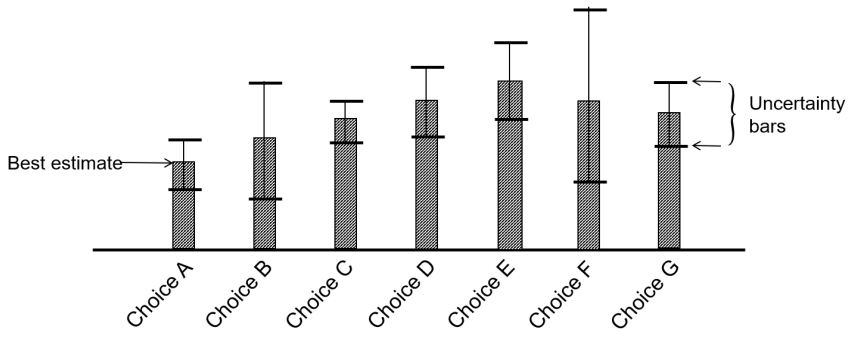


Figure 1.12: A decision problem with discrete choices, and uncertainty about the value of each choice.

if we have to learn while doing.

- The number of times we repeat the choice.
- What we learn from one choice may affect our beliefs about other choices, reflecting an underlying belief model.
- The structure of the belief model that captures any underlying structural relationships.
- The presence of physical resources that are being consumed or managed, such as setting up a machine to perform an experiment, consuming supplies, or requiring skilled personnel.
- The time and expense required to make and implement a choice.

The most common approach that people use when choosing from among a discrete set of choices is to simply pick the one that appears to be the best. This ignores the ability to learn from the choice to make a better decision in the future. The value of learning now on future decisions will depend heavily on how many times we will be facing the same set of choices. It may also be ignoring risks that may be associated with making a choice that may perform very poorly.

There are many settings where the decisions are quite important, and we have to live with the decision for a while. Examples might be deciding to develop a particular drug, or choosing a supplier that we have to live with for at least a year. For these problems, it is particularly important to spend some time developing the best set of beliefs about the possible performance of each choice.

Belief models are typically complicated by correlations. Choosing which drug to develop may require comparing different types of cancer drugs that serve a similar market. We may have to choose among a number of suppliers clustered by country which share the same risks of increased tariffs, disease outbreaks, and currency changes.

1.7.4 Static versus sequential decision problems

In the academic literature, there is a strong sense of competition between the community doing deterministic optimization, and the fragmented communities doing optimization under uncertainty. Most deterministic optimization models are deterministic approximations of stochastic problems,

and a byproduct is that people using deterministic optimization can be quite defensive when faced with the ways that uncertainty affects their problem.

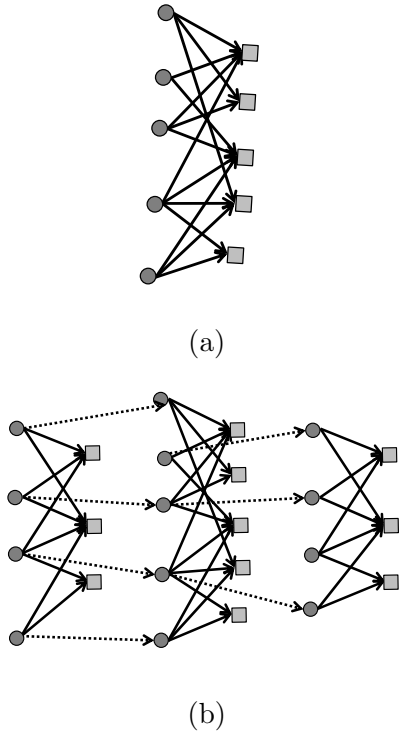
We urge readers to keep the following in mind:

- A static, deterministic optimization model is just a special case of a sequential decision problem.
- We will show (in Volume II) that deterministic optimization tools are widely used in the solution of general sequential decision problems.
- By far the most common decision problem that arises in practical applications is the one depicted in figure 1.12 where we have to choose the best of a set of choices. Even when we capture the uncertainty in our beliefs about the choices, the different methods for solving this problem still reduce to solving sequences of deterministic optimization problems.
- The problem is not that a deterministic approximation is used; the error is in how the decisions are evaluated. The performance of the decisions have to be evaluated over time as new information arrives.

The most common mistake made in the use of deterministic optimization models is to overlook when the problem has to be solved repeatedly over time. An example of this arises in a problem class called an "assignment problem" where we are assigning "resources" (people, trucks, machines) to "tasks" (job assignments, loads to be moved, jobs to be completed). These are never solved just once; as time moves forward, the resources make progress finishing the tasks, new tasks are called in, and the machines may undergo failures changing how long they need to finish a task.

Figure 1.13(a) depicts the problem as a static, deterministic problem. When this problem was first solved by George Dantzig in the 1950s, it was considered a great breakthrough (which it was). However, even 70 years later, top professionals ignore that the problem is never solved just once; it has to be solved repeatedly over time, and it is virtually always the case that the solution at one point in time affects the problems that need to be solved in the (uncertain) future.

The real problem is depicted in figure 1.13(b), where we illustrate the problem being solved sequentially over time. We note that it is impossible to determine if a deterministic optimization problem needs to be solved sequentially just by looking at the mathematics of the model; it requires understanding the problem in English.



(a)

(b)

Figure 1.13: (a) Static assignment problem; (b) dynamic assignment problem.

1.8 Stages of modeling

We begin by recognizing three different ways of viewing a problem:

- The real world – This is where decisions are implemented, and where we collect information describing the true performance of our system.
- The base model – This is typically in the form of a simulator that is designed to mimic the real world as closely as possible. Simulators (sometimes called "digital twins") are powerful, but they can be very expensive to develop, which is the reason why we often need to design methods for making decisions without the benefit of a simulator to test the performance of our policy.
- A lookahead model – Lookahead models are used only for making decisions where we have to approximate the impact of a decision made now on the future. Lookahead models are widely used in some

form (Google maps uses an approximate lookahead model to plan a path to the destination), but they are not used universally.

In section 1.3.1 in our discussion of framing the problem, we described three stages that were involved in understanding the different dimensions of a decision problem. In this section, we are going to focus specifically on developing a computer model, should one be needed.

1. **Framing the problem:** – Any attempt to model a decision problem requires the elements identified in our framing process:

   a. A plain English narrative – It is important to always start a description in the words of a domain expert with no training in even the process of modeling.

   b. Answer the three framing questions:

      i. What are the performance metrics? If you cannot articulate quantifiable performance metrics, you may have one of those complex, unstructured problems that is not amenable to a formal analysis process.

      ii. What types of decisions are being made (and possibly who makes them)? At this point, does simply listing potential decisions make it apparent what choice you should make?

      iii. What are the types of uncertainties that may affect the performance of the system? This can be a complex question that will take time to articulate and then analyze to understand the effect of these uncertainties on the performance of different decisions.

   At this point you may feel that the choice you should make is obvious. If not, continue to the next step.

2. **The universal modeling framework** – Understanding the different elements of the universal modeling framework (described in section 1.3.2) can provide a more complete understanding of your problem. Specifically:

   a. You will need to pull together the information you need to make a decision and compute your performance metrics (also known as the state variables). Since this depends on how you are going

to make decisions (the policy), you typically will not be able to identify all the elements of the state variables right away.

Pay attention to information that you would like to have, but cannot observe directly (at least not with any accuracy). These may represent opportunities to use statistical estimation/machine learning.

b. Understand what decisions you are allowed to make. Later you will address the problem of making decisions (designing the policy).

c. List the types of information that will arrive after you make a decision.

d. You will need to think about how your information in the state variable changes over time. Of course, this evolves as our understanding of what information we need. This is the transition function.

e. Finally, you will need to understand how you are going to evaluate the performance of your system. This constitutes your objective function.

3. Modeling uncertainty – This is often the most subtle dimension of modeling a sequential decision problem, often because uncertain quantities may not be immediately obvious. Although there are many potential sources of uncertainty, there are only two ways that it enters the model:

a. Uncertainty in quantities and parameters within the state variable, which carries the information needed to make a decision and/or calculate performance metrics.

b. Uncertainty in the information that may arrive after a decision is made, but before the next decision is made (we have been calling this the exogenous information process).

There are different ways to capture uncertainty:

a. Use observations of uncertain quantities (prices, demands, travel times) from history, and use these samples to calibrate and tune our model.

b. Create a mathematical model of the uncertainty, and then generate samples from the mathematical model.

We deal with uncertainty in much greater depth in chapter 5, where we will identify a number of different sources of uncertainty as a guide to naming the uncertainties that apply to your specific application.

4. **Designing policies** – Here we address the very rich challenge of designing methods for making decisions. Chapter 4 describes four classes of policies, which capture fundamentally different methods for making decisions, but we defer to Volume III a complete discussion of the process of designing policies.

   Note that while we introduce the idea of a state variable in the universal modeling framework, the state variable is partly defined by the information needed by the policy.

5. **Computer implementation** – Once we have designed policies, we have to decide how we are going to test them. The choices are:

   a. Testing in the field – In this case all we have to do is to implement the policy on a computer, which also means pulling together the data needed to make a decision.

   b. Computer simulation – How we test policies on the computer depends on the complexity of the system. Typically we are choosing between:

      i. Spreadsheet implementation – Most problems are relatively simple allowing us to test ideas in a spreadsheet. Spreadsheets can even be the basis of a production system.

      ii. General purpose programming environments – If the problem is too complex for a spreadsheet, we will need to turn to any of a wide range of programming environments. This requires the skills of expert programmers.

6. **Evaluating and calibrating models and policies** – At this point we have to decide if we can develop a simulator to evaluate the policy, or if we need to implement the policy so it can be used in the field:

   a. Developing a computer-based simulator – At this point we have everything we need to simulate the performance of the policy. A computer simulator is simply a software implementation of

the elements in the universal modeling framework. We can then run this simulator either on historical data, or data generated from a mathematical model.

b. Field testing – It is often the case that we may not have the time or resources to develop a simulator. Instead, we implement the policy and then monitor how well it works in practice.

Creating a computer-based simulator offers significant advantages, but introduces the difficult dimension of model calibration. By contrast, implementing a policy directly in the field means we are testing it in an environment which requires no calibration. The problem with a field implementation is that searching over different classes of policies, and in particular tuning any parameters, can be painfully slow. Tuning policies in the field has received very little attention in the research literature.

1.9 Types of analytics

If we are solving deterministic optimization problems, we can draw on a substantial family of solvers, from commercial packages such as Gurobi or FICO Xpress to any of a wide range of packages that can be downloaded for free. For example, Google offers their "OR Toolbox" at no cost, even for commercial users.

There is very little in the way of commercial tools for optimizing decisions over time as would be necessary in a sequential decision problem. However, we will typically be drawing on several toolboxes as we build custom systems for specific problems. These include:

- Deterministic optimization – Just because we are trying to make decisions over time, under uncertainty, does not mean that the tools of deterministic optimization are no longer used. In fact, most (but not all) sequential decision problems involve solving sequences of optimization problems that are solved using deterministic solvers.
- Simulation – Typically this refers to Monte Carlo simulation which is a body of tools and techniques for estimating functions of random variables. Monte Carlo tools are particularly well suited for high-dimensional, complex problems, making them one of the most powerful tools for modeling the evolution of information.

- Statistical estimation/machine learning – Stat/ML tools, which include linear or nonlinear regression, tree regression, locally parametric models, and neural networks in a variety of sizes. Stat/ML can be described as a set of tools to estimate something we do not know, using information we do know.

These tools are typically described as coming from different communities, of which only one (deterministic optimization) is viewed as solving decision problems. And yet, it is important to understand the role of each for the purpose of making decisions.

Simulation models, for example, are almost always built to help with understanding the behavior of some process, which might be anything from a manufacturing plant or the spread of disease in a population. In both of these examples we are looking to see how the design of the system (the plant layout, where vaccine inventories are held) or the control of the system (how jobs are routed, placing vaccine replenishment orders). We might also simulate the path of hurricanes, and while we cannot change their paths, this information can be used to help guide evacuations which would also have to be simulated.

In short, it helps to think of simulation models as objective functions, or forecasts of future events to be used to make decisions.

So what about statistical estimation/machine learning? While these fields use optimization to fit a model, the goal is simply to estimate a quantity or parameter. But why are we creating these estimates?

We might be estimating the nature of a tumor, or how many people approve of the performance of a president, or the probability a circuit will work. Or we might be estimating events in the future, such as how many people might purchase a product, or the generation of energy in a wind farm. In all these cases, we are creating an estimate or forecast to help make a decision now.

Each of these tools may also provide services with intrinsic value beyond helping to make better decisions. This is most easily seen with the large language models that are rapidly evolving as of this writing. For example LLMs can help with a growing set of tasks such as doing research, processing requests, and creating images, but not making decisions.

1.10 Closing notes

This chapter has laid the foundation for thinking about the complex array of problems known as sequential decision problems. We start from the following premise:

"If you want to run a better {anything} you have to make better decisions."

The vast majority of settings that involve making decisions fall in the broad category of sequential decision problems, where we make decisions repeatedly over time as new information is arriving (note that a special case of a sequential decision problem is a problem where we just make one decision).

The goal of the volume is building bridges from any problem where there is interest in performing better (presumably by making better decisions), and computer software that can help with those decisions. Computers require mathematical models that capture the problem, and these models need to understand the problem, expressed in English, but using terms that capture the problem in a way that can be translated into the language of models.

A critical feature of this process is the use of a general modeling strategy that we call the universal modeling framework. It is our claim, based on decades of working on a very wide class of problems from many settings, that this modeling framework can capture the features of any problem where computers may be useful. This is an important caveat, since there are problems with poorly defined, or nonexistent, metrics: Whether to marry someone? What field to choose as a major in college? What restaurant to choose when hosting a visitor?

The universal modeling framework formalizes choices such as how to make a decision (called a policy), which then helps answer what information is needed. The UMF also provides a basis for comparing policies which do not need a forecast (buy-low, sell-high in finance, order-up-to inventory policies) against those that do, and to evaluate the value of more accurate forecasts.

Using computers to make decisions opens the door to the use of "artificial intelligence" which is a term that is widely used in the public press without being properly defined. We cover the seven levels of artificial intelligence which clearly differentiate between tools based on machine learning such as large language models (such as ChatGPT), and tools for making de-

cisions such as deterministic optimization (level 5) and sequential decision problems (level 6).

1.11 Exercises

Review questions

1.1 — What are the three stages of decision automation?

1.2 — What are the three questions posed in Stage 1 of framing the problem? Illustrate these in a problem setting of your choosing.

1.3 — What are the five elements of the universal modeling framework?

1.4 — What is the definition of a decision? Give three examples in different settings that you encounter in your personal activities.

1.5 — Briefly describe the seven levels of artificial intelligence, divided into the four different classes as organized in the chapter.

1.6 — What are the three classes of statistical models?

1.7 — What is the difference between a base model and a lookahead model? Use the context of making a long trip by car using Google maps to illustrate both of these.

1.8 — What are the six stages of modeling? **Modeling questions**

1.9 — Give an example of a sequential decision problem that you encounter in your daily activities and do the following:

a) Identify at least one performance metric that you would like to improve.

b) Provide at least one decision that affects the performance metric.

c) Describe any uncertainties that might interfere with how the decision performs when it is implemented.

1.10 — Give three examples of problems involving physical resources and identify the decisions that arise in each setting.

1.11 — You are going to play 15 games of tic-tac-toe where the goal is to force the opposing player to get three in a row (at which point you win).

Neither of you have ever played this game before, and you want to capture how the other player learns your strategy of playing. Remember that tic-tac-toe usually ends in a tie, so it will be necessary to get your opponent to believe you are going to make a mistake. Answer the questions below in English (no math allowed).

a) Design a performance metric that captures the results of the 15 games.

b) What decisions do you have to make?

c) What are the uncertainties?

d) Describe the information you would have after playing several games that you would want to have to design a policy?

Chapter 2

Applications

The first step in improving any product, process or service is to provide a basic description, and then identify possible performance metrics, types of decisions, and sources of uncertainty. We are going to illustrate these first steps using a variety of application settings. We are then going to draw on these applications throughout the rest of the book to illustrate different modeling devices.

The beauty of sequential decision problems is that they arise throughout activities that involve people. Figure 2.1 is a snapshot of some of the problem settings that describe the activities of the author, and which served as the motivational foundation for the work in this book. Each picture represents a host of decision problems. This is in sharp contrast with problem classes such as linear, integer, and nonlinear programming, which are important and powerful tools, but which solve only a very narrow subset of decision problems.

We note that there is a natural bias to focus on the management of physical resources since this is what we see. To be sure, managing physical resources poses many opportunities for making better decisions, but there are other decisions that directly address collecting information, along with managing the often significant flows of money required to support these operations. In this chapter we are going to review the following problem settings:

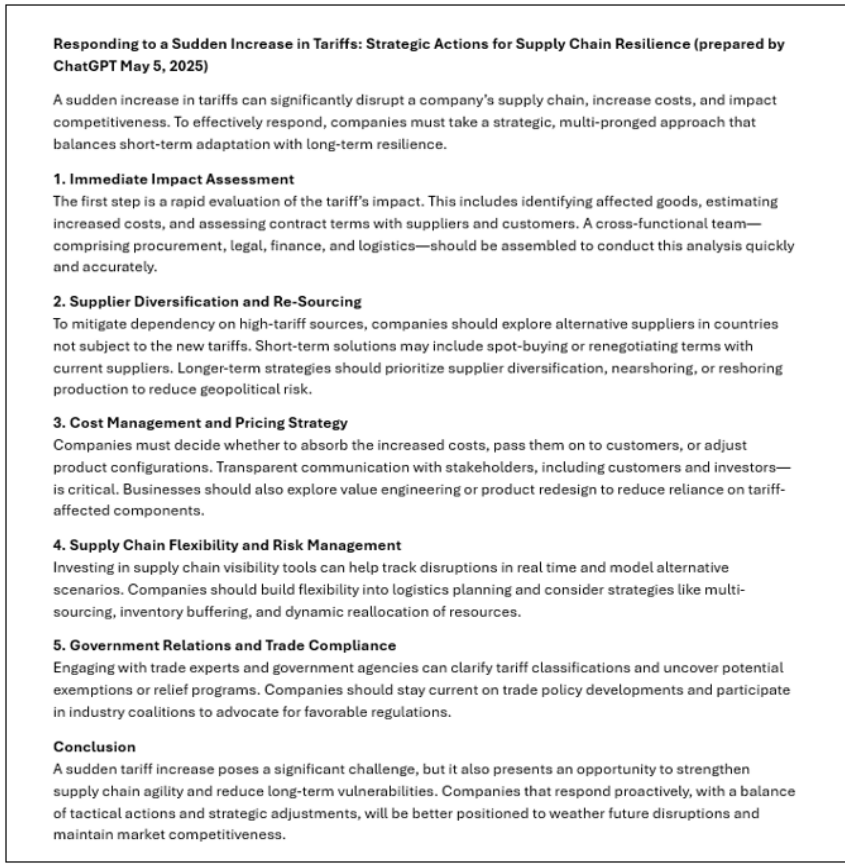
- Inventory planning
- Demand management
- Electric power management

Figure 2.1: An illustration of the many settings for making decisions.

- Hotel revenue management
- Health applications
- Presidential elections
- Truckload fleet management
- Mutual fund cash management
- Supply chain finance
- Intelligent trial-and-error (many settings)

Many of these can be described as meta-problem domains, since they contain subareas which by themselves represent major fields of human activity. Our goal is to create a diverse set of applications, partly to illustrate the range of problems that fall under the umbrella of sequential decision problems, and partly to provide a diverse set of decision settings that will motivate the modeling framework that we will present in the remainder of the book.

At this stage we are not ready to describe complete models – that will come in Volume II. For each application, we are going to offer answers to the three questions in the framing process, recognizing that this is just a sample to get the reader thinking about the process.



Responding to a Sudden Increase in Tariffs: Strategic Actions for Supply Chain Resilience (prepared by ChatGPT May 5, 2025)

A sudden increase in tariffs can significantly disrupt a company's supply chain, increase costs, and impact competitiveness. To effectively respond, companies must take a strategic, multi-pronged approach that balances short-term adaptation with long-term resilience.

1. Immediate Impact Assessment
The first step is a rapid evaluation of the tariff's impact. This includes identifying affected goods, estimating increased costs, and assessing contract terms with suppliers and customers. A cross-functional team—comprising procurement, legal, finance, and logistics—should be assembled to conduct this analysis quickly and accurately.

2. Supplier Diversification and Re-Sourcing
To mitigate dependency on high-tariff sources, companies should explore alternative suppliers in countries not subject to the new tariffs. Short-term solutions may include spot-buying or renegotiating terms with current suppliers. Longer-term strategies should prioritize supplier diversification, nearshoring, or reshoring production to reduce geopolitical risk.

3. Cost Management and Pricing Strategy
Companies must decide whether to absorb the increased costs, pass them on to customers, or adjust product configurations. Transparent communication with stakeholders, including customers and investors—is critical. Businesses should also explore value engineering or product redesign to reduce reliance on tariff-affected components.

4. Supply Chain Flexibility and Risk Management
Investing in supply chain visibility tools can help track disruptions in real time and model alternative scenarios. Companies should build flexibility into logistics planning and consider strategies like multi-sourcing, inventory buffering, and dynamic reallocation of resources.

5. Government Relations and Trade Compliance
Engaging with trade experts and government agencies can clarify tariff classifications and uncover potential exemptions or relief programs. Companies should stay current on trade policy developments and participate in industry coalitions to advocate for favorable regulations.

Conclusion
A sudden tariff increase poses a significant challenge, but it also presents an opportunity to strengthen supply chain agility and reduce long-term vulnerabilities. Companies that respond proactively, with a balance of tactical actions and strategic adjustments, will be better positioned to weather future disruptions and maintain market competitiveness.

Figure 2.2: ChatGPT's version of how a supply chain should respond to tariff increases.

2.1 Getting started – framing the problem

It is very common to discuss complex problems using general terminology. Figure 2.2 (prepared by ChatGPT) answers the question:

> *Prepare a 1-page discussion of how a company should respond to a sudden increase in tariffs that will disrupt their supply chain.*

The problem with these general discussions is that they never provide a clear pathway to improving the process. Everything in the discussion is probably true, but it lacks specific actions that can be taken to solve a problem. The response provided by ChatGPT (in 2025) reflects the type of

generic chatter that are widely found in business books, which could never be the basis of a formal model.

This chapter illustrates the first stage of framing the problem, which consists of four elements:

The narrative: This is a short discussion that describes a problem in the style that might be used by someone within the problem domain.

Performance metrics: We provide a list of performance metrics which will need to be prioritized (this is covered in Chapter 3).

Decisions: Next we provide a list of decisions that impact one or more of the metrics. Decisions are covered in Chapter 4

Uncertainties: Finally, we describe the forms of uncertainty that can the effect of a decision can be distorted when implemented, or as the process progresses forward in time. Uncertainties are covered in Chapter 5.

At this point we are not going to attempt to describe how we might solve the problem (that is, make the decisions). For this, we need other material that will be developed later. The goal at this stage is to use a variety of problem settings to illustrate the process of identifying metrics, decisions and uncertainties in a general way. Identifying these three elements is the key to solving any decision problem, so we have to develop the habit of them first.

2.2 Capturing interactions

While the identification of metrics, decisions and uncertainties is a valuable starting point, it is also important to understand how they interact.

- The effect of decisions on metrics – Each decision should have some impact on at least one metric, and every metric should be affected by at least one decision.
- Uncertainty in the performance metrics given the decisions – We might want the shortest path, but the travel time depends on congestion; we might want to choose a drug that will reduce an infection, but a patient may not respond to a particular medication; an investor cannot predict the exact return when purchasing a stock.

Decisions\Metrics	Sales revenue	Product costs	Holding costs	Stockouts
When/how much to order	H	H	M	M
Purchase currency hedge?	N	L		N
Discounting	M	N	L	M
Market product on social media	H	L	L	M

Figure 2.3: Interaction matrix for decisions and metrics for an inventory problem with small lead times.

- The uncertainty may constrain what decisions we can make – A trucking company has to move loads, but these are called in at random; a hotel may hold back rooms for business travelers who may or may not reserve rooms; a utility may count on energy from wind, but the amount of energy that can be generated is uncertain.
- The uncertainty changes the dynamics of how the system evolves over time – The disease in a population may spread in an uncertain way; the economy may evolve in an uncertain way that affects the value of the dollar; uncertain competitor behavior can decrease sales.

2.2.1 Impact of decisions on metrics

A useful exercise is to create a spreadsheet where different decisions are listed on the left and metrics are listed across the top. Then, using pure judgment, enter one of the following in each cell to capture what you think describes the impact of each decision on each metric:

- H – Decision has a high impact on the metric.
- M – Decision has a medium impact on the metric.
- L – Decision has a low impact on the metric.
- N – Decision has no impact on the metric.

Figure 2.3 illustrates how this might look for a small inventory problem. The spreadsheet can be downloaded from:

https://tinyurl.com/FramingInteractionMatrix/

Start by listing the metrics from left to right in order of importance. We are then going to use the matrix to identify the most important decisions, and the metrics that are most impacted by the decisions you have listed.

The interaction matrix can be used in two ways:

Uncertainty\Metrics	Sales revenue	Unit costs	Holding costs	Stockouts
Sales (units sold)	H	L	M	M
Lead times	L	L	M	H
Forecasting errors	L	N	M	M
Inventory shrinkage	M	N		N

Figure 2.4: Interaction matrix for uncertainties and metrics given a decision, for an inventory problem with small lead times.

1) List all decisions, and then assess the impact of each decision on each metric. From this, identify the decisions that seem to have the greatest impact on the most important metrics.

2) For complex problems, listing all decisions may be impractical. Instead, use the set of metrics to help identify the decisions that are most relevant to the problem. Then return to (1) to help prioritize the most important decisions.

The exercise of filling in tables such as this can help guide the process of understanding the role that decisions play in improving performance, but before moving forward with the expensive and complex step of collecting data and building a computer model.

2.2.2 Impact of uncertainty given the decision

Imagine that we have made a decision (which means it is fixed). We need to understand the forms of uncertainty that affect the metrics produced by the decision. For our simple (short lead time) inventory problem, we might obtain the matrix given in figure 2.4.

Modeling uncertainty simply means understanding information that may arise in the future which we do not know yet. This simple observation is often overlooked in discussions of uncertainty, which can become buried in sophisticated mathematics ("stochastic modeling") and the quantification of risk (which few understand).

As with decisions, we can start by listing every source of uncertainty that we can think of, and then use the interaction matrix to prioritize the ones that are most important. Alternatively, we can use our set of metrics to help guide the identification of important sources of uncertainty.

2.2.3 Impact of uncertainty on decisions

The most common setting where uncertainty impacts what decisions you are allowed to make arise in the context of resource allocation problems where we are managing some resource (people, machines, product supplies, drugs) to serve tasks (jobs, patients, customers). In Figure 2.5 we are illustrating the assignment of trucks (with drivers) to move loads of freight. The main source of uncertainty is the flow of loads being called in by shippers to be moved, but this could be any task. We might assign a driver to a load that is not attractive in terms of profitability, but which ties up the driver for several days, preventing him from being used on a better load that might be called in later in the day.

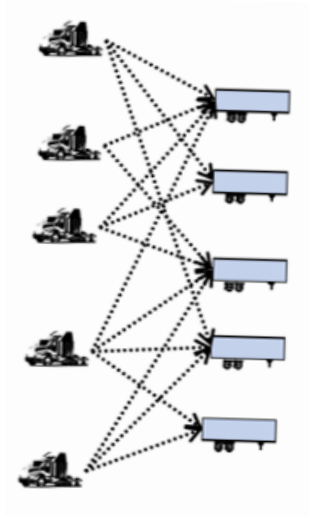


Figure 2.5: A simple assignment problem.

The flow of customer requests is a major source of uncertainty that arises in:

- Supply chain management (demand for products).
- Health (patients needing treatment).
- Hotels (requests for rooms to rent).
- Energy (the demand for electricicty or gas for heating).
- Finance (deposits and withdrawals of cash).

An important characteristic of the flow of demands that need to be served is how these demands become known to the system. Some variations include:

- No advance warning, immediate service (sales of any retail product).
- No advance warning, backlogging possible (online purchases).
- Advance request, immediate commitment required (booking of hotel rooms).

- Advance commitment with cancellation terms (expensive purchases, such as aircraft).

There can also be uncertainty in the availability of resources used to satisfy customers:

- Doctors, nurses may be ill.
- Machinery may fail.
- A truck driver can decline an assignment to move a load.
- An investment may decline in value.

2.2.4 Uncertainty in system dynamics

Whatever we know at one point in time may change as we step forward in time, and if it changes, we are typically unsure about how it is changing. Some examples of uncertainty in the evolution of the system include:

- Changes in costs, prices and other performance metrics.
- Changes in the status of people, equipment and facilities over time. People may quit or get sick, equipment may break down, a facility may be damaged in a storm.
- Changes in market attitudes, the presence of disease in a population, how people might vote for a candidate.

It is important to recognize that uncertainty about how the system evolves over time can be divided into two categories:

- Uncertainty in the *function* describing the evolution of the system. Some refer to this as "model uncertainty." If we are managing the distribution of vaccines, we may use different models for how the disease propagates through the system. When we are planning evacuations for a hurricane, we can choose from different models of how the storm will progress.
- Uncertainty in the *parameters* that determine the behavior of the function.

2.2.5 Uncertainty in forecasts

There are many problems (but not all) where making a decision now requires projecting what might happen in the future. Of course, the future is almost always uncertain, but it is our choice whether to use a "best estimate" of what might happen in the future, or to explicitly model this uncertainty to help us make a decision now. We return to this issue in Chapter 4 when we discuss ways of making decisions.

2.2.6 Comments

Uncertainty is easily the most subtle issue when understanding a decision problem. Often people have an intuitive sense that a type of uncertainty is important; this section helps to refine how a form of uncertainty actually impacts a decision problem.

2.3 Inventory planning

2.3.1 Narrative

One of the most widely studied problems in operations research (as well as stochastic optimization) is the inventory problem which is typically posed as determining when to place an order to replenish, and how large the order should be. The classical textbook description of an inventory replenishment problem is depicted in figure 2.6, which shows the increase in inventories when new product arrives, followed by the depletion as product is consumed. A stockout, where inventory drops to zero, is depicted.

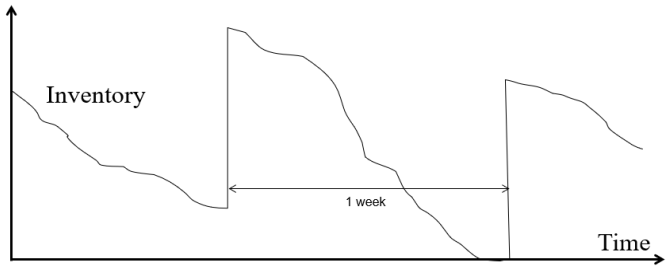


Figure 2.6: Illustration of a classical inventory problem with short lead times.

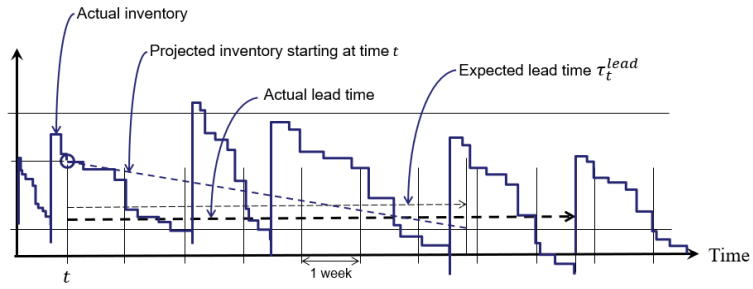


Figure 2.7: Illustration of an inventory problem with long lead times.

A more realistic version of an inventory problem is illustrated in figure 2.7, which depicts an inventory problem that might arise in a setting where the product is coming from a distant location (such as from China to the eastern U.S.). We might have to wait 6-8 weeks, but weather delays can extend this even more. Long-distance shipping typically involves movement by ocean container ships for port-to-port moves, rail (common within the U.S.) and then truck.

Planning inventories has to be coordinated with strategies for managing demand, which can be influenced through pricing, discounts, promotions, and marketing. Inventory management has to deal with a number of sources of uncertainty, ranging from the usual day-to-day variability in demand, to market shifts due to competitor behavior, new technologies, and both losing suppliers as well as the emergence of new sources of supplies. In addition, there can be significant variations in transportation times due to weather, mechanical failures, and labor actions at ports. Excessive delays may be managed by using fast modes such as air freight as an alternative to container shipping, and truckload trucking as an alternative to rail.

2.3.2 Metrics

We separate metrics between "base metrics" which are captured through routine reporting, and "risk metrics" that specifically account for significant events (typically negative) which, in the judgment of management, are not properly captured in the base metrics.

- Base metrics

  - Inventory holding costs, which cover a range of items including cost of the capital tied up in inventory, warehousing costs (heating/AC, manpower handling, overhead of the warehouse and equipment), insurance costs, cost from spoilage, theft, obsolescence.
  - Shipping costs, including packaging, transportation and insurance.
  - Revenue from meeting demand, which needs to reflect any discounting.
  - Forecasting accuracy metrics.
  - Customer service metrics, such as delayed or lost demand that is not met from lack of inventory, and product returns (e.g. due to quality issues).
  - Cost of promotions, coupons, marketing and advertising.
  - Utilization of facilities (are they full?), people and equipment.
  - Labor issues, including labor productivity, need for overtime, hiring costs and layoffs.

- Risk metrics:
  - Currency risks when the product is purchased from another country in a different currency.
  - Significant stockouts that force customers to competing products.
  - Theft, cyberattacks.
  - Significant disruptions (interruptions at a supplier, damage to facilities) that prevent delivery to customers or employment.

2.3.3 Decisions

It helps to organize decisions based on whether we are solving a single inventory problem, or addressing network level issues. Textbook inventory models typically focus on operational decisions such as when to place an order and how much. However, the perspective changes when we have long lead times, where a decision now impacts the system months into the future.

The list of decisions that are relevant to inventory planning is quite long. In section 2.3.5 we are going to use a tool we call “interaction matrices” to identify the most important decisions.

	Physical	Financial	Informational
Whether to observe/verify inventory			x
Who from the set of available suppliers to place the order (if there are multiple suppliers)			x
When to place a replenishment order.	x		
How much to order.	x		
How to package it (ocean container, half-container, pallets, boxes).	x		
How to finance order (cash transfer, bank loan, ...)		x	
The choice of transportation modalities for products from abroad to intermediate storage facilities	x		
The choice of transportation modalities for domestic distribution to customers	x		

(a) - Operational

Whether to purchase currency hedges for products from abroad.		x	
Discounts/promotions (to reduce inventory)	x	x	
Product pricing.	x	x	
Marketing/displays (shelf space, end-cap display, advertising (various forms)			x
Running market tests for features, design, ...			x
Design and implement marketing campaign			x
Choice of supplier (for each material or component), including whether to have multiple suppliers. This determines the possible suppliers.			x
Equipment maintenance (increases scheduled downtime, decreases unscheduled down time)	x		

(b) - Tactical

	Physical	Financial	Informational
Contracts with inventory visibility platforms (where is my shipment)?			x
Choice of demand forecasting methodology (statistical methods, involvement of different people across the organization).			x
Product design (which determines the required materials and components)	x		
Market identification (who are we selling to)			x
How much connectivity (information sharing) to seek with manufacturing supply chain partners			x

(c) - Strategic

Figure 2.8: Different types of decisions for inventory problems for each time scale.

2.3.3.1 Single inventory problem:

These decisions are made on different time scales: operational (hourly, daily, weekly), tactical (monthly), and strategic (quarterly, yearly).

- Operational - These are decisions that might be made in real-time, but are typically made either daily or weekly (figure 2.8(a)).
- Tactical – These decisions are made monthly (figure 2.8(b)).
- Strategic – These decisions are made quarterly or yearly (figure 2.8(c)).

The decisions that have to be considered for each time scale depends on how far into the future a decision now impacts the system. For example, weekly decisions for ordering inventory (which would be operational) will be approached differently if the lead times are one week or eight months.

	Physical	Financial	Informational
Where to locate buffer stocks and how to rebalancing them	x		
Closing existing facilities	x		
Where to purchase/lease/build/expand manufacturing facilities	x		
Which manufacturing facilities to close/sell, terminate leases	x		
Where to purchase/lease/build/expand warehouses and distribution centers	x		
Whether to introduce materials handling automation in DCs and warehouses	x		
Investment in information technologies for information sharing and coordination			x
Arrange significant line of credit or other source of backup financing		x	

Figure 2.9: Supply chain design decisions.

2.3.3.2 Supply chain design

There are decisions related to the design of supply chain networks that cut across many (tens to thousands) of individual inventory decisions. These are decisions that are typically made on longer time scales. Figure 2.9 provides some examples of network-level decisions for designing the supply chain.

2.3.4 Uncertainties

Uncertainties also occur on different time scales. We include a special category for major disruptions that may occur, but not on a regular basis.

- Hourly to daily variations (figure 2.10(a)).
- Weekly variations (figure 2.10(b)).
- Monthly to yearly variations (figure 2.10(c)).
- Major distributions(figure 2.10(d)).

Identifying the different sources of uncertainty is a particularly rich area for complex problems such as supply chains. Not only are there a wide range of uncertainties, they come in different styles such as fine-grained volatility, regime shifting, spikes, bursts and rate events. We discuss these behaviors in more detail in chapter 5.

2.3.5 Interactions

A powerful exercise that helps with developing an understanding of the different elements of decision problems is to subjectively assess the strength of different types of interactions, an idea we first introduced in section 2.2. We start with describing the interactions between decisions and metrics

	Physical	Financial	Informational
Day-to-day variations in customer demands	x		
Errors in measuring inventories			x
Inventory "shrinkage" (theft, loss, spoilage, breakage, ...)	x		
Yield from shipment (how many items/how much material met specifications)	x		
Transportation delays due to weather, equipment failures	x		
Forecasting errors			x
Cost of raw commodities		x	
Cost of inputs from suppliers		x	
Power (electricity, fuels) outages	x		
Communication errors, human execution errors			x
Day-to-day variations in company stock price		x	
Financial fraud in individual transactions		x	
Day-to-day availability of available-to-allocate capacity	x		

(a)

	Physical	Financial	Informational
Shifts in the mean demand due to technology shifts, competitor behavior, market shifts	x		
Changes in the selling price of a product (affects demand and profit flows)	x	x	
Changes in commodity prices		x	
How the market responds to pricing changes	x		
Delays due to strikes at ports, railyards, international crossing points	x		
Shifts in behavior of large customers	x		
Shifts in attitudes on Wall St (e.g. from "growth" to "stable" to "recession")			x

(b)

	Physical	Financial	Informational
Emergence of new information technologies (AWS, AI, visibility platforms)			x
Emergence of new manufacturing/material handling technologies (e.g. robotics)	x		
Emergence of new competitors	x		
Shifts in population patterns (e.g. growth of immigration)	x		
Shifts in demand patterns (increase in demand for high-end products)	x		
Treaties governing trade	x	x	
Changes in labor availability	x		

(c)

	Physical	Financial	Informational
Emergence of new information technologies (AWS, AI, visibility platforms)			x
Emergence of new manufacturing/material handling technologies (e.g. robotics)	x		
Emergence of new competitors	x		
Shifts in population patterns (e.g. growth of immigration)	x		
Shifts in demand patterns (increase in demand for high-end products)	x		
Treaties governing trade	x	x	
Changes in labor availability	x		

(d)

Figure 2.10: Different types of uncertainties for inventory problems in operational (a), tactical (b) and strategic (c) time scales, plus major disruptions (d) that do not occur on a regular basis.

Decisions\Metrics	Sales revenue	Product costs	Holding costs	Stockouts	Inventory turns	Operating margin	Sales growth
When/how much to order	H	H	M	M	M	M	L
Purchase currency hedge?	N	L	N	N	N	M	N
Discounting	M	N	L	M	L	M	M
Market product on social media	H	L	L	M	M	L	M
Choice of supplier	L	M	L	L	L	M	L
Pricing	H	N	L	L	L	M	M
Currency hedges?	N	L	N	N	N	L	N
Inventory sensors	L	L	L	M	L	L	N
Use visibility platforms to track inbound product?	N	L	L	M	L	L	N
Product design	M	H	L	M	M	M	H

Figure 2.11: Interaction matrix for decisions and metrics for an inventory problem with long lead times.

for an inventory problem with long lead times, shown in figure 2.11. We emphasize that filling out this matrix is completely subjective, since it helps us identify the most important decisions, as well as the metrics that we have the greatest chance of improving.

What we are doing is replacing what is often a completely invisible step of choosing what decisions to focus on, with a process that makes this choice explicit, even if it is made subjectively.

The interaction matrix for uncertainties and metrics given a decision might look like that given in figure 2.12. Here, we make a point of holding a decision fixed to avoid blending the effect that uncertainty has on which decision we make.

Uncertainty\Metrics	Sales revenue	Unit costs	Holding costs	Stockouts	Inventory turns	Operating margin	Sales growth
Sales (units sold)	H	L	M	M	H	M	H
Lead times	L	L	M	H	M	L	M
Forecasting errors	L	N	M	M	M	H	L
Inventory shrinkage	M	N	L	N	M	L	L
Changes in commodity prices	N	H	N	N	N	M	L
Market response to price	M	N	N	N	N	M	L
Work stoppages	M	L	N	M	L	L	N
Competitor pricing behavior	M	N	L	L	L	M	M

Figure 2.12: Interaction matrix for uncertainties and metrics given a decision for an inventory problem with small lead times.

2.4 Demand management – selling furniture

2.4.1 Narrative

The flip side of managing the flow of goods through the different steps of manufacturing and distribution is the challenge of managing demand. The largest producers of furniture are China (by far), the United States (mostly for domestic consumption), Germany (mostly for Europe), Italy (high end

furniture) and Poland (for lower cost furniture). Furniture sellers have to work with long lead times, highly seasonal demand and customization, as well as a competitive marketplace. While they will use all the usual tools to manage the flow of physical product, it is important to use various strategies for managing demand to help balance supply with the marketplace.

Some of the demand-side issues that furniture sellers must deal with include:

- Highly variable demand, due in part to variations in people moving into new homes.
- Evolving customer preferences as customers respond to design trends and changing styles, along with new products and materials.
- Price sensitivity, which reflects both the state of the economy and the competition.
- Market response to advertising and visibility in social media.
- Strategies for search engine optimization.
- Partnering with home décor influencers who can showcase products.
- The ability to offer discounts and promotions to reduce excess inventories.

2.4.2 Metrics

Metrics always depend on the perspective of who is being evaluated, but some that we would expect in this setting might be:

- Sales (in units, and total revenue).
- Net revenue – Sales, minus cost of goods and advertising.
- Stockouts, delayed order fulfillment.
- Web traffic – There are a number of metrics used to evaluate e-commerce portals such as visits, click-through rates, bounce rates, time on site, and conversions.
- Number of likes, shares, comments, as well as brand mentions and sentiment analysis in online discussions.
- Engagement – Use of virtual reality previews.

2.4.3 Decisions

We are envisioning that we are the outlet manager for a furniture store:

- Which furniture items to stock.
- Pricing.
- Promotions and discounts – e.g. discount for a furniture setting.
- What marketing channels to use – social media, TV, print mailings, in-store marketing.
- Marketing budgets for each channel.
- A/B testing of webpage designs.
- Conducting market surveys – offering specific packages at a subset of stores.
- Staffing decisions (how many, what skills).

2.4.4 Uncertainties

Some examples of uncertainties that might arise when selling furniture include:

- Deviations between actual demand and forecast for furniture at different levels of aggregation.
- Supply lead times.
- Product quality issues.
- Market response to price, discounts and promotions.
- Customer willingness to substitute products of higher or lower quality.
- Variations in consumer preferences.
- Variations in web traffic and conversion rates.

2.5 Electric Power Grid management

Energy systems is an umbrella term spanning the vast network that supplies the energy that supports modern society. We are going to focus our attention on the flow of electricity, but this includes power generation that can come from different sources, primarily gas (but still some coal and oil), nuclear and a growing presence of energy from wind, solar and hydroelectric facilities.

2.5.1 Narrative

The backbone of any electrical system is the power grid which consists of high-capacity transmission lines that move power over long distances at high voltages, from 69kv (that is, 69,000 volts) up to 345kv, with ultra-high voltage lines as high as 765kv. Power is then sent to businesses and residences using local distribution networks with voltages between 4kv and 14kv.

Power comes from a "fleet" of power generators that may include nuclear, coal, steam generators and gas turbines, along with hydroelectric power (there is a strong imprint of naval vocabulary because of the presence of nuclear power). These generators are differentiated by the speed with which they can be turned on ("dispatched") or off, and how readily they can be run faster or slower. The other important characteristics are the fixed cost and operating costs. For example, nuclear power is high fixed cost, low operating cost, and they have to be run continuously except for maintenance periods. Gas turbines have much lower fixed costs but higher operating costs, and they can be turned on in under an hour. Steam generators, on the other hand, need 8-12 hours to heat up, and as a result they are typically planned a day in advance.

The growing use of energy from wind and solar has introduced a degree of uncontrollable variability that power grids have not been exposed to before. The way this variability can be handled is with storage which comes in different forms, but the most visible is grid-level battery storage. Australia and Florida are two regions that have invested heavily in battery storage, but this is starting to become a common investment accompanying the development of large solar fields and wind farms.

Storage, however, comes in other flavors, including:

- Pumped-hydro storage, where water is pumped uphill, and then used on demand to generate electricity by flowing downhill.
- Battery-to-grid storage, where the batteries in cars and residences are used as a form of battery storage.
- Thermal storage, where energy is stored by heating up a liquid in a large tank.
- Demand management (or demand response) – We can "store" the need for electricity by deferring activities such as running washing machines and driers or cooling down rooms (such as libraries) to use the cool air later.

Energy is a particularly rich problem domain in terms of managing different forms of uncertainty, using different technologies for generating power which require dramatically different time frames in terms of advance notification (literally from 2 seconds for varying the output of a gas turbine to a year for changes in maintenance schedules for nuclear power plants).

As this book is being written, the power grid has come under pressure to meet the growing demands from the use of "AI" tools, which require massive computing centers to handle the demands for calculating neural networks with tens of billions of parameters using the types of specialized chips from companies like Nvidia. There is also growth in the use of air conditioning to handle increasing temperatures, along with the computing demands of cryptocurrencies.

2.5.2 Metrics

Among the rich set of metrics for power generation would include:

- The cost of electricity – This is easily the most important metric used to evaluate energy systems, although this is for societies that can assume 24-hour availability of electricity. It is important to distinguish between the fixed cost of an investment (nuclear power plants are very different from gas turbines and solar panels) and the operating costs.
- Demand coverage/outages – There are some regions of the world that have access to electricity for only a portion of each day.

- Meeting temperature targets – People like to live in environments where the temperature stays in a narrow range. A building manager may face penalties for the periods when the temperature in an apartment falls outside of a specified range. Some food and drugs need to be refrigerated to certain temperatures, with penalties when these are violated.
- Impact on the environment, ranging from net CO2 emissions, heating water, consuming land, impact on local flora and fauna (the list is quite long).
- Reliability – The frequency and severity of outages.

2.5.3 Decisions

Decisions in the power sector span time frames from seconds (to smooth out voltage variations) to years, for long term agreements for purchasing power:

- Adjustment of power generators for frequency regulation, which occurs at 2-second intervals.
- Power purchase decisions (typically at 5-minute intervals) which may involve buying power from the grid or selling power back to the grid.
- Decisions to purchase or sell power given current grid prices.
- Purchase and storage of gas, oil and coal (in some cases, hydrogen).
- Installation of grid sensors to understand the state of transmission lines.
- Power purchase agreements, which are contracts to buy or sell power over multi-year periods.
- The location, type and capacity of energy generator, from gas turbines and nuclear power plants to wind farms and solar fields
- The location, type and capacity of energy storage.
- Grid transmission capacity, which controls how much power can be transmitted at a point in time.

2.5.4 Uncertainties

Energy systems offer an exceptionally rich set of uncertainties that affect both infrastructure investments to the daily operation of the energy system.

- Weather, especially temperature and humidity, affects demand in a region.
- Wind direction and speed for wind turbines.
- Cloud cover that can change solar intensity.
- Generator failures due to weather events, mechanical failures, and sabotage.
- Grid prices, which can vary by both 5-minute intervals (the frequency of updates to grid prices) and 2-second intervals (for power regulation).
- Human activities such as a football game or concert.
- Regulatory changes that can affect tax incentives to penalties and outright restrictions (e.g. on offshore wind, or building new pipelines).
- The cost of equipment (solar panels, wind turbines, batteries, gas turbines, and nuclear power plants) evolves continuously over time.
- The emergence of new technologies, such as small nuclear power plants and new battery technologies.

The emphasis of renewables has raised the visibility of uncertainties. Figure 2.13 shows solar output on an hourly basis, over an entire year, which communicates both seasonal variations, familiar daily cycles, and the effects of cloud cover. Of particular importance is the predictability of the different forms of uncertainty. We know when the sun will set decades in the future, but cloud cover is particularly difficult even on very short time horizons.

2.6 Hotel revenue management

2.6.1 Narrative

Hotels face the need to manage reservations for rooms for up to a year in the future, although most bookings arrive in the last few months, and in

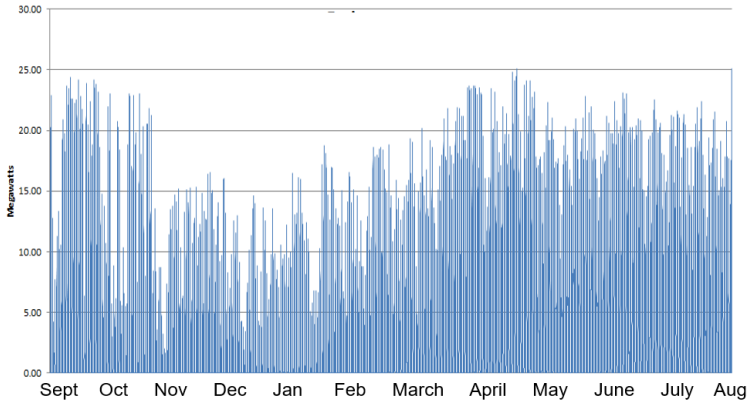


Figure 2.13: Hourly solar energy generation over an entire year.

some cases, the last few weeks. As time passes, hotels can increase rates as the hotel fills up. Normally the hotel will start by offering lower rates, but these rates have to reflect the possibility that the hotel may fill up, which means possibly turning away people traveling on business with a much higher willingness to pay.

There is more to managing hotels than just the price charged for a room. Hotels can offer a variety of services, from free breakfast, access to gyms and pools, and tickets to local services such as ski slopes or travel tours.

An important advertising channel is on social media outlets such as Google and Facebook. These outlets run sophisticated auctions where advertisers have to bid dynamically for the right to post links to their webpage for a period of time.

2.6.2 Metrics

Some of the metrics for hotel revenue management include:

- Total revenue each booking day, minus the costs of services offered.
- Amount spent on internet ad searches (Google, Facebook, ...).
- Room utilization.
- Declined customers.
- Unused rooms.

2.6.3 Decisions

The decisions that might be made by a hotel manager typically include:

- How much to charge for a room τ days into the future.
- Which e-commerce outlets to advertise on.
- How much to bid to have their ads posted on each e-commerce outlet.
- What services to offer at different rates.
- How to design the webpage.

2.6.4 Uncertainties

Hotel managers have to face several sources of uncertainty:

- Total bookings each day for a particular stay-date.
- The acceptance rate for a room given the price and service offerings.
- How often a bid to advertise is accepted given the size of the bid (or the bidding policy).
- The success rate for customers that see a webpage design.

2.7 Health applications

Health is a massive topic that literally touches every human being. We have a strong incentive to make decisions that maintain or improve our health, while keeping within budgets. The topics below are just a tiny snapshot of the rich set of decision problems that arise in this setting.

2.7.1 Managing Type 2 diabetes

2.7.1.1 Narrative

Approximately 10 percent of the global population has Type 2 diabetes, which reflects an inability to control blood sugar (glucose) levels in the blood. Type 2 diabetes arises when the pancreas does not produce enough insulin, or when the body becomes resistant to insulin. A failure to control the resulting elevated levels of blood sugar can be a host of health conditions, including heart and kidney failure, damage to blood vessels in the

eyes which can lead to glaucoma and blindness, foot problems from poor circulation (sometimes requiring amputation), and increased incidence of dementia.

Short term spikes in blood sugar (known as hyperglycemia), which may occur shortly after eating certain types of food, can produce blurry vision, headaches, fatigue, and difficulty concentrating. Drops in blood sugar (hypoglycemia) can produce dizziness, rapid heart rate, fainting, seizures, and even a coma.

Diabetes, then, is a disease that has to be managed both over the long term, as well as the short term. Elevated blood sugar over long periods of time can produce permanent organ damage, while short-term variations can create medical conditions requiring immediate treatment ((Hsih 2010), (Brown 2010)).

2.7.1.2 Metrics

As with most medical conditions, a number of the metrics capture the state of the patient, but there are others.

- Blood sugar, measured in mg/dL (typical range is 70-130), or mmol/L (typical range 3.9-7.2), which is an instantaneous measurement often taken after meals.
- Fasting blood sugar – This is the blood sugar level after 8 hours of fasting.
- Time in range (TIR) – This is used with continuous blood sugar monitors and measures the time that the blood sugar stays within an acceptable range.
- Time above range (TAR) and time below range (TBR) – Percent of time the blood glucose level is above or below the acceptable range.
- Hemoglobin A1c (HbA1c) – This test reflects a rolling 2-3 month average, where desirable values are under 6.5 to 7 percent.
- Glucose variability – The standard deviation of the blood sugar.
- The cost of treatment (physician visits and medications).
- Frequency of need to visit a physician.
- Medical consequences from diabetes, spanning foot pain (neuropathy), vision loss, amputation and death.

2.7.1.3 Decisions

We deviate from our normal style of just listing decisions, and list medical decisions made by the physician separately from the decisions made by the patient.

Medical decisions (made by the physician)

- Choice of drugs, dosage levels and timing. Metformin is the standard drug of choice which is used by 50 to 80 percent of patients who are taking medication. However, many patients cannot tolerate it, and have to turn to a range of other medications including insulin, sulfonylureas, meglitinides, DPP-4 inhibitors, and so on.
- Prescribing technology-assisted treatments, such as
  - Continuous glucose monitoring devices.
  - Insulin pumps, which provide precise insulin delivery.
  - Automated insulin delivery systems.
- Surgical interventions, such as bariatric surgery and pancreatic islet transplantation.

Patient decisions

- Seeing a physician.
- Following physician instructions.
- Submitting to testing, investing in home testing equipment.
- Administering drugs.
- Diet choices – This of course represents a wide range of decisions affecting the type of food and quantity.
- Exercise choices – What type, how frequently, how intensely.

2.7.1.4 Uncertainties

- Side effects of a medication.
- How well does a patient respond to a medication (change in blood glucose level).
- How well a patient adheres to a diet and exercise program.

- Patient ability (and willingness) to follow treatment instructions.
- Long-term progression of the disease as a patient ages.
- Availability of new medications.

2.7.2 Public health – Managing naloxone kits

2.7.2.1 Narrative

While drug use and overdoses have been a problem for decades, there was a dramatic spike in overdose deaths due to synthetic opioids starting around 2013, quickly outpacing deaths from all other drugs by a wide margin. Much of this increase was due to the introduction of Oxycontin by Purdue Pharmaceuticals in 1996. Oxycontin contained oxycodone, which was less addictive than other painkillers.

Oxycodone had a long-lasting formulation that did not provide the quick "hit" that drug users were looking for. However, the public found that the drug could be crushed and misused, a practice that exploded in use after 2013. Below we summarize the metrics, decisions and uncertainties from the perspective of a public health officer working for the state or municipal government ((Pi 2019)).

2.7.2.2 Metrics

- Number of opioid overdoses where:
  - No one present had naloxone (person survived or died).
  - Naloxone was present, but naloxone was not administered (person survived or died).
  - Naloxone was administered (person survived or died).
- Number of overdoses where EMS responded.
- Number of overdoses where person had to be transported to the hospital.
- Cost of the naloxone kits.
- Cost to healthcare system.
  - Overdose is handled outside of the hospital (e.g. by EMS).

  - Overdose requires transporting person to the hospital (very expensive).

- Enforcement costs.

2.7.2.3 Decisions

The decisions below are from the perspective of the state government:

- How many naloxone kits should be allocated to different types of organizations:
  - Harm reduction agencies, treatment providers.
  - Direct services organizations.
  - First responders (EMS, law enforcement, fire).
  - Other community-based organizations that interact with people who use drugs (faith-based orgs, housing providers, food pantries, etc).
  - Pharmacies, hospitals.
  - Jails and prisons.
- How many kits should be allocated to needle exchange programs by region:
  - Overdose hot spots.
  - Rural vs urban.
  - Different counties/regions.
  - At risk populations, such as Tribal lands.
- Who to train on how to recognize and reverse an overdose? Who to train how to use the kits?
- How to advertise the availability of naloxone kits?
- How to fund naloxone distribution strategy?
- Who to submit proposals to obtain funding?

2.7.2.4 Uncertainties

- Usage rates and patterns by people/patients. This is affected by:
  - Awareness – people may not know that naloxone is available.
  - Trust – people may not be comfortable disclosing that they need it.
  - How people respond to opioid use and treatment.
  - Availability of drugs in the market.
  - Transportation barriers – people may not be able to get to a distribution point.
- Contaminants in the supply that have an unknown impact on naloxone and overdose reversals.
- Budget allocated for preventive measures such as naloxone kits and the staff capacity to distribute them.

2.7.3 Running clinical trials for drug testing

2.7.3.1 Narrative

As of 2024, there were almost 500,000 clinical trials testing various drugs and treatments for effectiveness. There are three phases of a clinical trial:

Phase I: Safety and dosage testing ($5 - $10 million) – A small group of healthy people are used to test for toxicity at different dosage levels and identify possible side effects. Researchers may also compare different methods of administering a drug, such as pills, patches or injections.

Phase II: Evaluation of efficacy and side effects ($20 - $100 million) - The treatment is applied to a larger group of patients who have the disease or condition that is the target of the treatment. Guided by what is learned in Phase I, this phase provides an initial indication of the effectiveness of the treatment. Side effects are observed, and the results will be compared to existing treatments.

Phase III: Large-scale testing ($100+ million) – Using pools of hundreds, often thousands, of patients drawn from different regions, the

treatment is compared to competing therapies to evaluate effectiveness and further observe for adverse reactions. Additional data is gathered for regulatory review.

Clinical trials are not only very expensive, they also take a lot of time. During this evaluation, the 20-year clock on patents is ticking, creating an incentive to draw a (hopefully positive) conclusion to go to market.

The process of running trials poses a large-scale logistical problem to administer the trials and requires substantial financing which also means considerable financial risk. The entire process has to be conducted in the presence of considerable uncertainty about the performance of a drug or treatment on a large scale.

Clinical trials may fail at any of the three levels because of:

- Lack of efficacy – The drug does not work as hoped.
- Safety concerns – There may be significant side effects.
- Regulatory hurdles – The drug may encounter regulatory problems.
- Commercial or strategic reasons – A company may not pursue a drug because of financial projections, financial risk, or competitive issues.

Typical success rates are:

- Transition from Phase I to Phase II: 60 percent.
- Transition from Phase II to Phase III: 30 percent.
- Transition from Phase III to approval: 50-60 percent.

The overall success rate through the entire process is around 10 percent.

2.7.3.2 Metrics

There are a variety of metrics that go into the evaluation of a drug:

- Successful transitions from each of the three phases to the next stage.
- The cost of each phase.
- The cost of gaining regulatory approval at each stage.
- The effectiveness of the drug or treatment.
- The presence of side effects.
- Manufacturing cost of the drug.

- Cost of distributing the drug (it may require refrigeration).
- Cost of administering the drug. (By mouth? Injection?)
- Marketing costs.

2.7.3.3 Decisions

We describe decisions from the perspective of the company that owns the drug with an interest in bringing it to market:

- At each phase, each week there is a decision to continue testing, stop and terminate the review (the drug fails), or stop and transition to the next stage (success).
- How many patients to interview and invite to become a part of the trial.
- Choice of hospitals to use as clinical testing locations.
- Decision to open testing locations (e.g. in a strip mall).
- Pricing of the drug.
- Marketing strategies: To physician? Direct to the market?.

2.7.3.4 Uncertainties

Decisions have to be made while keeping the following uncertainties in mind:

- The rate at which eligible people can be identified.
- The response of people to the treatment.
- The decisions of regulatory committees.
- Anticipated acceptance of the drug by physicians.
- Decisions made by competitors that can affect the sales of the drug.

2.8 Running a presidential election

2.8.1 Narrative

Anyone who has watched the series "West Wing" (or carefully follows presidential elections) has seen the challenge of managing a presidential campaign. Invariably it is a complex operational problem that requires managing candidates and staff, often either collecting information (such as running polls) or disseminating information (making speeches), and always in a budget-constrained environment.

2.8.2 Metrics

Some of the most important metrics include:

- Whether the candidate wins the election or not.
- The number of votes from the electoral college.
- Polls in each state (especially the swing states).
- The amount of money on hand each week.
- Donations each week.
- Donations in response to social media posts.
- Weekly expenditures.

2.8.3 Decisions

The campaign manager has to make a number of decisions, including:

- Where to give speeches each day.
- What themes to emphasize.
- Choice of vice-presidential candidate.
- What advertising channels to use (television, social media, billboards), and spend rates.
- Expenditures on printed promotional material (signs, mailings, pamphlets).
- How many people to hire at different levels, by region.

- Where to set up field offices.
- When and where to run polls, what questions to ask.

2.8.4 Uncertainties

Presidential elections have to be managed in the presence of a number of uncertainties:

- How many votes the candidate receives.
- Change in favorability ratings over time, and after major events (e.g. national convention).
- Change in favorability ratings after highlighting different themes.
- Donations overall, and in response to specific calls for donations (e.g. through text messages).
- Anticipated biases in the polls.
- Events in the news that impact public perception (favorably or unfavorably) of the candidates policies.
- Attack ads by opponents.
- Large donations to favorable or competitive super-PACs.
- Adverse health events impacting the candidate.

2.9 Truckload fleet management

2.9.1 Narrative

In the U.S., freight primarily moves in the form known as full truckload trucking, where a shipper fills what is typically a 53-foot trailer that will pull up to 46,000 pounds (depending on the type of freight) from one location to another. They operate similarly to taxis – the truck driver (with a tractor) will move empty to pick up a load of freight at one location and then drive it to another where the trailer is either unloaded or dropped off to be unloaded later. A driver might move one or two loads in a single day, but most loads take anywhere from 1 to 5 days.

Once a driver drops a load, the challenge is to minimize the number of miles the driver has to move empty to pick up another load. Three issues really complicate running a truckload carrier:

- The movement of freight is not balanced. There are regions of the country that produce more freight than is consumed (this is particularly true of the Midwestern U.S.) and regions that are primarily consuming regions (typically the coasts and major cities). As a result, the market is willing to pay much more to move freight from producing regions to consuming regions, while loads out of consuming regions may not even pay enough to run the truck (but it is better than moving empty).

- Truck drivers need to observe strict rules on how many hours they can drive each day and each week. In addition, they need to return home, either daily, or weekly, or, for long-haul drivers, once or twice a month.

- The booking of freight is highly dynamic. Most loads are booked one to three days into the future. A trucking company may have to hold drivers to meet the needs of a top shipper who only calls in loads a day in advance.

There are over 2 million drivers working in the truckload industry. Most trucking companies operate with fewer than five drivers, while others have 10,000 drivers or more.

2.9.2 Metrics

The most commonly reported performance metrics include:

- Operating profit per week or per mile.

- Revenue per driver per week or per mile.

- Empty miles as a percent of total miles.

- Miles per driver per week.

- Fraction of time that drivers get home on time.

- Percent of time loads are picked up and delivered on time.

- Driver turnover (number of drivers quitting per week).

2.9.3 Decisions

Decisions from the perspective of the manager in charge of dispatch and load planning might include:

- Which load a driver should be assigned to.
- Whether to accept a load that is offered to be picked up in the future.
- Whether a load should be handled by the carrier's own drivers or a brokerage division (which finds owner-operators who can move the load).
- What price should the trucking company offer to move freight for a shipper in a particular traffic lane (origin-destination pair) in the upcoming year? This is part of an annual "bidding process" that determines the preferred carrier for each shipper for each lane.
- How many drivers to hire who live in a particular location (called a driver domicile).
- How many tractors and trailers should the fleet be operating.

2.9.4 Uncertainties

Some of the uncertainties faced in truckload trucking include:

- How many loads will be offered each day, in each lane, by the primary shippers that the carrier is serving?
- How many loads will be available to be moved, and at what price, on public "load boards" that any carrier can choose from?
- Will a driver accept an assignment to a particular load?
- Are freight volumes trending up or down?
- What are current spot prices?
- Will a load on an external load board actually be available if the carrier elects to move it?

Professor,

As discussed briefly during the break in Saturday's class I am trying to apply the newsvendor inventory service level formula to optimize the level of cash that is held by my mutual fund in order to meet redemption requests from investors.

Problem

Mutual funds hold a certain percentage of their assets in cash in order to meet redemptions from investors. The exact amount is determined more as a guesstimate rather than systematically. I believe the problem is a variation of the Newsvendor problem discussed in class:

Cost of shortfall: If not enough cash is held, the fund must sell some of its holdings and will experience transaction costs. These costs are deterministic and can be assumed to amount to 0.2% of the transaction volume. We can assume that there are no financing costs if there is a cash shortfall. Cost of excess cash [Ce]: Holding to much cash leads to an opportunity cost of not participating in the market. Daily returns on the fund's portfolio are stochastic, and therefore I am not sure whether the Newsvendor formula we saw in class can be applied. The 'cost' of excess cash may even be a gain on some days when the portfolio is down.

Complications

I think reducing the problem to a Newsvendor problem is a good first approximation, but I see several complications:

(i) There will be some correlation between the error term in the daily return on the portfolio and the probability function of getting redemptions.

(ii) Reducing the problem to a single-period is a simplification for which I don't have a good feeling whether it is significant or not.

(iii) The demand function is likely to be skewed by a few large redemptions. Although there is a large number of atomistic retail investors who would redeem small amounts each, a few large institutional investors might redeem large amounts at a time.

(iv) The zero financing cost assumption does not apply for all redemptions, and in particular may not apply to large redemptions by institutions. Their redemptions proceeds need to be wired the following business day, while corresponding sales of portfolio securities take three business days to settle. However, small redemptions by retail investors are paid by check which take several days to mail and clear, by which time any securities sales will have settled.

I would appreciate if you could point me to some literature that treats Ce as a stochastic variable. Maybe there is already a published solution to my problem (I'm not aware of any)?

Thank you for your help.

Regards,

Part-time MBA student

President & Portfolio Manager

Figure 2.14: Email from a mutual fund manager and former MBA student seeking advice on how to manage the cash balance.

2.10 Mutual fund cash management

2.10.1 Narrative

A mutual fund manager who had taken an operations planning course for his MBA was introduced to a classic problem known as the "newsvendor problem." Newsvendor problems arise when you have to decide on a quantity of a resource (for example, newspapers) to allocate to serve a demand that is not known when you make your decision. If you allocate too much, you will have resources left over, where we assume they cannot be held for the future (just as today's newspapers are of no value tomorrow). If we allocate too few, then we will have unsatisfied demand.

After finishing his MBA (at a top business school), the mutual fund manager faced the problem of deciding how much cash to keep on hand to handle requests for redemptions. The problem is summarized in the email shown in figure 2.14, but the core elements are as follows:

- The mutual fund has to maintain enough cash to meet requests for redemptions. If there is not enough cash on hand when a redemption request comes in, they will have to liquidate stocks, incurring transaction costs, and possibly being forced to sell at a lower price. If they hold too much cash, then they are missing out on the potential growth of investments in the market.
- There are two types of customers: retail and institutional investors. Redemptions for retail investors may take several days to clear, while the larger redemption requests from institutional investors have to be settled the same day.
- Deposits and redemption requests are correlated with market performance. Growth in the market can attract new deposits, while drops can trigger sudden requests for redemptions.

2.10.2 Metrics

The metrics involved in this exercise include:

- Overall return on the portfolio each day, net of operating costs (transactions costs, redemption expenses).
- Amount of cash being held.
- Sales required to cover redemption requests.

2.10.3 Decisions

The decisions faced by the mutual fund manager are:

- How much cash to hold.
- Which assets to sell to raise cash.
- Which assets to purchase when there is too much cash on hand.

2.10.4 Uncertainties

The decisions have to be made in the face of the following uncertainties:

- Deposits by retail or institutional investors.
- Redemption requests by retail or institutional investors.

- Changes in market indices.
- Changes in interest rates.

2.11 Supply chain finance

2.11.1 Narrative

Every supply chain transaction involving the purchase or sale of commodities, components and final products implies a flow of money, creating a complex network of flows between buyers and sellers (at all levels of the supply chain), along with third-party financial partners who may supply financing and insurance.

The steps in a financial transaction typically include:

- Supplier sends goods and invoices to the buyer.
- Buyer approves the invoice in their ERP system.
- Once approved, the supplier may choose to be paid early by the financier (which might be a bank).
- Financier pays the supplier (typically at a discount).
- Buyer pays the financier at invoice maturity (perhaps 60 or 90 days later).

There are a variety of financial transactions that may occur, such as:

- Invoice approval – Buyer confirms invoice is valid and due for payment.
- Receivable assignment – Supplier assigns the invoice to the financier.
- Early payment – Financier pays supplier before due date.
- Maturity payment – Buyer pays financier on agreed due date.
- Transaction fees/discounts – Financier earns a fee from the transaction.

There are a number of sources of uncertainty in supply chain management that have an impact on finances. Companies can protect themselves using different forms of insurance. Some examples are:

- Inventory insurance - Protects goods held in warehouses or in-transit (including 3rd-party logistics centers) against theft, damage, or loss.
- Currency hedges to protect against changes in the relative value of different currencies when importing from other countries.
- Trade credit insurance - Protects suppliers or lenders against the risk of buyer non-payment due to insolvency, protracted default, or political events.
- Marine cargo insurance - Covers physical loss or damage to goods in transit—via land, sea, or air—during international or domestic shipment.
- Protects against losses due to political instability, such as expropriation, currency inconvertibility, import/export restrictions, war or civil unrest.
- Performance bond insurance - Guarantees that a supplier or contractor will meet contractual obligations. Insures buyers against supplier failure.
- Credit default swaps - Used by financial institutions to hedge against counterparty credit risk.

2.11.2 Metrics

There is quite a long list of financial metrics used by larger companies. A sample of those that are directly related to the financial management of a supply chain include:

- EBITDA – Earnings before interest, taxes, depreciation and amortization. This is a high-level metric that captures cost of goods sold (COGS), revenues, and all costs incurred to manage the flow of cash and capital.
- Return on equity (ROE) and earnings per share (EPS).
- Free cash flow.
- Working capital and cash reserves.
- Debt to equity ratio.
- Interest expense.

2.11.3 Decisions

A sample of decisions made by a chief financial officer include:

- Choice of financing strategies for different transactions.
- Choice of forms of insurance (see list above).
- How much cash to maintain, and in which accounts.
- Dividend payments.
- Capital allocation.
- Debt vs. equity financing.

2.11.4 Uncertainties

Again, a small sample of different forms of uncertainty arising in supply chain finance include:

- Payment defaults by buyers and sellers.
- Currency variations.
- Changes in tariffs and trade restrictions.
- Recession risk, shifts in overall sales (up or down).
- Interest rate volatility.
- Credit market volatility.

2.12 Intelligent trial and error

2.12.1 Narrative

There is a massive problem class in decision-making that can be best described as "intelligent trial and error." These arise when there is a set of discrete choices, and where the performance of each choice is uncertain. Examples of problem settings where this arises include:

- Materials science
  - What chemicals to blend to make a new material.
  - What temperature to run a process.

  - What steps in the manufacturing process.

- Health
  - What drug to try to treat a condition.
  - Whether to run a test (imaging, blood test).
  - Where to locate a clinic for distributing naloxone kits.

- E-commerce
  - Which of two web page designs to use.
  - Which product to advertise on a webpage to maximize revenue.
  - What price to charge for a product (out of a set of possible prices).

- Manufacturing
  - Optimizing a semiconductor manufacturing process (temperatures, time in a chemical bath, chemical concentrations, diameter of the silicon wafer).

- Finance
  - Finding the best settings for the parameters of a trading policy.
  - Which supplier to use, given the risk of default.
  - How much reserve capital to maintain.

- Supply chain management
  - Which supplier to use for a product given uncertainty about product quality.
  - Setting the reorder points for stock replenishment.
  - What advertising channels to use.

- Choice of people
  - Baseball – Who should bat fourth, or play catcher.
  - Basketball – Who should play each position.
  - Portfolio managers – Who gets the best results managing a portfolio.

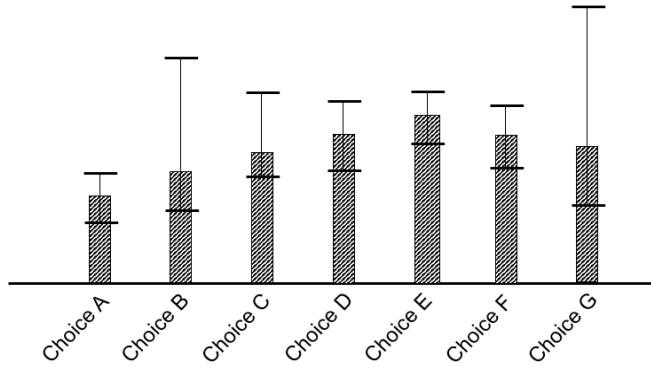


Figure 2.15: A set of discrete choices.

Each of these contexts involves choosing from among a set of choices. We want to choose the one that works the best, but we are not sure how well each will perform. The situation is depicted in figure 2.15. There may be two choices, dozens, hundreds, and many thousands.

This basic problem comes in a variety of flavors:

- Belief model
  - Independent beliefs – This is where our belief about one choice is unrelated to the beliefs about other choices. In real applications, this is relatively rare.
  - Correlated beliefs – Choices may share features, such as drugs from the same family, a style of shirt with different colors, or the "closeness" of two choices, especially if they represent a discretized price, or concentration, or geographical location.
  - Parametric models – We may construct our beliefs using a parametric model, such as a linear model relating different prices to estimated demands.
- Cost of running a test
  - Inexpensive experiments – Observing how many times an ad is clicked on a webpage is a very inexpensive way to conduct an experiment. Time frames may span microseconds to seconds to minutes.

  - Expensive experiments – A laboratory experiment may take a day to a week (or more). Complex computer simulations may take hours to a week or more.

- Noise level
  - Low noise experiments produce accurate estimates from a single trial.
  - High noise experiments produce very noisy outcomes, requiring multiple tests with the same or similar choices.

- Offline vs. online learning
  - Offline learning describes experiments done in a lab or computer simulation, where we can tolerate poor performance from an experiment.
  - Online learning describes learning done in the field, where we have to live with the outcome of an experiment (such as testing a price of a product, or how a drug works on a patient).

- Presence of physical or financial resources – Basic learning problems are linked from one experiment to another purely on the basis of what we learn. However, it is possible that problems are linked by a physical (or financial) resource:
  - There may be a fixed budget for running experiments. Each experiment consumes some portion of the budget.
  - Physical experiments may require inventories of ingredients that have to be available.
  - An experiment may require a machine that is set up to perform a specific task, which means it is easier to do other experiments that need the same setupl

- Sequential or parallel experiments
  - Sequential experiments:
    * A patient can be used to test one drug at a time to determine which works best on that patient.
    * A manufacturer may be able to test one process at a time to determine which produces the highest yield.

  - Parallel experiments:
    * A retailer can run multiple promotional campaigns (e.g. in-store advertising) at different stores to learn which works best.
    * A scientist can test dozens or hundreds of different compounds on a single plate to see how they react to a particular type of cancer cell.

- Instantaneous vs. lagged learning
  - Instantaneous learning - We make a choice (e.g. to run an experiment) and learn the results immediately.
  - Lagged learning - There is a time lag between when we run an experiment vs. learn the outcome. Lags can be minutes in high-speed settings, up to a year or more, as would occur when a bank offers a loan, and has to wait years to learn if the loan recipient misses payments or defaults.

Any of these settings can still be described by our trio of metrics, decisions and uncertainties.

2.12.2 Metrics

Any "experiment" is assumed to return an observation of a performance, whether it is the number of ad-clicks, or the response of a patient to a drug, or the yield of a process for manufacturing semiconductors. Of course, there may be more than one metric to describe performance, which we may wish to optimize some combination. However, we should distinguish two important dimensions of performance:

- The cost of trying each choice.
- Average performance over some horizon.
- The variability around the average, which captures the reliability of a process.
- The likelihood of "poor" outcomes.
- Other performance metrics, such as side effects of a drug, or the potential of significant losses in market share.

2.12.3 Decisions

This is simple – it is the set of choices. These might be:

- Binary – Such as
  - Whether to take an action (sell a company, launch a new product, send a drug to clinical trials) or not.
  - Whether to hold or sell an asset.
  - Which of two webpage designs to use (often called A/B testing)
  - Whether to give a patient a drug, or not.
- Discrete set – This could be a set of suppliers, a choice of different drug treatments, different marketing channels to advertise a product, or any of a set of thousands of molecular compounds to be tested in drug development.
- A discretized set of values of a continuous parameter, such as the price of a product, the concentration of a chemical, the temperature for baking a semiconductor.

There are problems where the set of choices is not obvious. For example, we may be looking for a supplier who can make a specialized component out of a new material which requires working at high temperatures. Or we need a very special chemical to make a new vaccine, or an extremely pure form of a gas that is needed in the process of making the latest semiconductor chips. Finding suppliers, or materials, or chemicals, to fit a need can be extremely challenging.

Then there are going to be problems where we know our performance metric, but do not know how to improve it. A cement manufacturer may need to cut costs to be competitive, but does not have a clear strategy for how to achieve this. A physician wants to treat a condition in a patient but does not know what treatment to pursue.

2.12.4 Uncertainties

Uncertainties for discrete choice (trial-and-error) problems can come in two forms:

- The performance of a choice, which typically differs from how we thought it would perform when decided to use the choice. We may

have a point estimate of the metric(s) for each choice, or some form of distribution. The actual performance is typically different than the point estimate, and if we are given a distribution of possible outcomes, the actual outcome may not necessarily be drawn from an assumed distribution.

- Whether the choice is available – Some examples are:
  - The choice may be a supplier, who is unable to bid on a contract.
  - The choice may be a person to fill a job, but they may not be willing to take a job.
  - We may want to use a type of material, but supply chain issues may restrict its availability.

2.13 Exercises

Review questions

When an exercise asks for an interaction matrix, you can use the template for the "Framing Interaction Matrix" that can be downloaded from https://tinyurl.com/FramingInteractionMatrix/

2.1 — For the inventory problem, pick a product you are familiar (for example food, clothing, household items, drugs, or hardware) and answer the following:

a) Identify metrics, decisions and uncertainties that seem relevant to your problem, using the lists of each dimension from the inventory section as a guide.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.2 — For the demand management problem:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced by a store manager at a retail furniture outlet.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.3 — For the power grid problem:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced when performing daily planning of power generators.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.4 — For the hotel revenue management problem:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced when managing bookings for rooms over a two-month planning horizon.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.5 — For the problem of managing type 2 diabetes:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced by a physician making decisions about a patient with type 2 diabetes.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.6 — For the naloxone kit management problem:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced by a state government planning the allocation of naloxone kits to different counties using funding from the federal government.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.7 — For the problem of running a presidential election:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced by the campaign manager for a candidate running for president.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.8 — For the problem of managing a truckload fleet:

a) Choose a set of metrics, decisions and uncertainties that you think would be faced when planning the problem of accepting which loads to move (typically performed up to seven days in the future).

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.9 — Consider the mutual fund cash balance problem:

a) The email from the mutual manager suggests a way of deciding how much money to hold in cash. Write out that formula.

b) Use the Framing Interaction Matrix to create interaction matrices to capture your best estimate of the impact of each type of decision on each performance metric.

c) Repeat (b) to capture your best estimate of the impact of each type of uncertainty on each performance metric.

2.10 — Name an example of a "trial-and-error" problem that you encounter in your own experience, where you have to make the same choice repeatedly.

a) Describe the context of the trial-and-error problem, and what triggers the need to make the decision again.

b) Describe the metrics (one or more if necessary), the set of choices, and all forms of uncertainty that arise in the process of making decisions.

c) Suggest how you would go about making a choice.

Chapter 3

Performance metrics

There is an old adage in management:

> "You cannot manage what you cannot measure."

Any time we wish to improve the performance of a process or system, it is important that we have a clearly defined performance metric, recognizing that there are often multiple metrics. Before we start down the road of discussing the complex issues associated with metrics, we have to first acknowledge that there is no shortage of problems that do not have well-defined metrics, such as who to marry, what job to take when graduating from college, or choosing what to paint or which sculpture to purchase. If you have trouble identifying at least one clearly quantifiable metric, it is likely that your problem belongs in the complex domain of human decision problems which are not going to benefit from analytical thinking.

But if you can identify at least one clear, quantifiable metric, keep reading.

3.1 Categories of metrics

There is a vast range of metrics, so it helps to try identifying major categories of metrics. Some of the more popular categories are:

1) Financial metrics - These include any metric measured in a currency. They might be measured:
   - Total quantity - Cash on hand, loan guarantees, investments.

  - Per unit time (day, month, quarter, year), typically representing cost, revenue or profit.
  - Per unit of a resource - Dollars per person, machine, facility, or per share.

2) Productivity metrics - These are non-dollar denominated metrics that may also be measured per unit time (patients seen, units produced, miles traveled) and per unit of a resource (per person, per machine, per facility).

3) Effectiveness metrics - Strength of a material, performance of a drug, yield from a manufacturing process, mean time between failures of a machine.

4) Service performance - How well we are serving markets or external agents, such as demand covered, performance ratings by customers, placement of students.

5) External performance ratings - The ranking of a school, reliability of products made by a company, sales ranking, rating of hospitals.

6) Behavior metrics - Deviation between actual decisions from predetermined directions or guidelines.

7) Estimation metrics - How well do we estimate or predict quantities (future demand, rainfall, amount in inventory) or parameters (patient diagnostics, cost of production). These assume we have some way of comparing a prior estimate to an observation of actual performance.

There are two ways to evaluate each category of metric:

- Average performance - These are totals or averages over time, capturing what would actually be experienced.
- Risk metrics - These measure events that are not properly represented by an average.

Average performance and risk metrics are discussed further in section 3.5.

It helps to provide specific examples. Below is a list of metrics from different categories.

- Financial metrics
  - Profitability metrics

        * Net income.
        * Gross profit margin.
        * Operating profit margin.
        * Return on assets.
        * Return on equity.
        * EBITDA – earnings before interest, taxes, depreciation and amortization.
    - Liquidity metrics
        * Current ratio (current assets/current liabilities).
        * Quick ratio (current assets-inventory/current liabilities).
    - Efficiency metrics
        * Asset turnover ratio (revenue/total assets).
        * Inventory turnover (cost of goods sold/average inventory).
    - Solvency metrics
        * Debt to equity ratio (total liabilities/shareholders equity).
        * Interest coverage ratio (EBIT/interest expense).
    - Valuation metrics
        * Earnings per share (EPS) – net income/average outstanding shares.
        * Price-to-earnings ratio (P/E) – Market price per share/ earnings per share.

- Productivity metrics
    - Fraction of time the asset is being used.
    - Number of jobs/tasks completed per week.
    - Number of jobs/tasks completed on time or late.
    - Mean time between failures (MTBF).
    - Mean time to repair.

- Effectiveness metrics
    - Machines come in a vast array of styles, from cars to air conditioners to blenders. In all cases, there is an assessment of whether it "works," although complex machinery such as a car

can fail in a variety of ways, from not starting to a blown tire to the defroster not working. A car may work, but the gas mileage may be lower than expected.

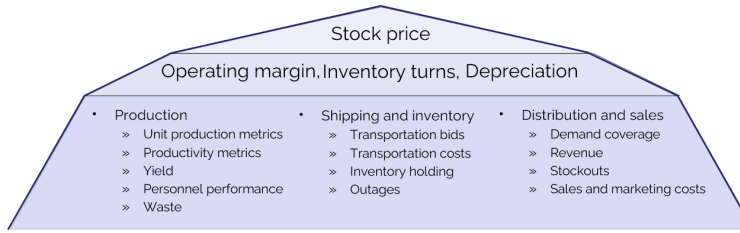
  - A plastic may be required to be heated to a given temperature without melting.
  - A laptop may need to perform at a certain speed.
  - Performance of a medication (e.g. for weight reduction).
  - Strength of a material.

- External performance ratings
  - Product sales (unit sales or revenue), rate of growth.
  - Cost of customer acquisition.
  - Number of positive reviews.
  - Rate of product returns.
  - Rate of customer churn (customers declining to renew the contract).

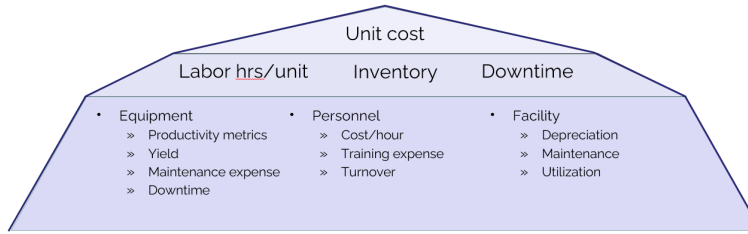
- Labor performance
  - Number of pieces attached/inspected per hour (manufacturing).
  - Monthly sales within someone's region or product line (sales).
  - Whether a project is finished on time and on budget (management).
  - Number of calls handled/customer rating (call centers).
  - Employee retention/turnover.
  - Rate of positive employee assessments in annual HR surveys.
  - Salary required to attract and retain people.

3.2 The metric pyramids

It is common, especially in business, to compile lists of metrics, where these lists can be quite long. It is very important to prioritize the metrics, which can be done fairly easily by organizing them into pyramids, as shown in figure 3.1. Figure 3.1(a) illustrates a potential set of metrics for someone



(a)



(b)

Figure 3.1: (a) A pyramid of metrics that might be used at the executive level of a publicly traded company. (b) A pyramid of metrics that might be used in a manufacturing plant.

working at the highest levels of a company (often called the "C-suite") where the most important goal is to maximize the quarterly stock price. These metrics, however, do not provide much guidance for someone working in a manufacturing plant where the most important metric might be cost, followed closely by production and quality. Figure 3.1(b) illustrates how a different pyramid can be created for someone who might work in manufacturing, where there is more emphasis on cost.

The organization of metrics into a pyramid is largely subjective, but there should be a single metric at the top which is felt to be the most important metric. The top metric should be one that is either maximized or minimized, but the same is not necessarily true of all of the other metrics, an issue we address next.

3.3 Objectives, targets and limits

We next need to specify what we are trying to achieve with each metric. There are three ways we can approach using metrics to evaluate perfor-

mance:

- Maximize/minimize – We often want to maximize or minimize a metric where bigger (or smaller) is always better. We may want the strongest material, or highest energy density, or the lowest cost.
- Targets – Here we are trying to hit a particular value, which might be a patient's body temperature, or the voltage in an electrical transmission line. Companies might want to hit projected targets for revenue or profitability to help control volatility.
- Upper/lower limit – A pre-diabetic patient may wish to keep their A1c level (a measure of blood sugar) below 6.0. A furniture retailer may wish to sell their current inventory (but no more). A truckload carrier would like to let each driver move 2000 miles per week, but no more, since the carrier would never be able to sustain a larger number, and the driver can be disappointed when high-mileage weeks are not repeated.

While there can be different ways to measure performance, ultimately a computer has to be able to look at a set of decisions and choose which one is best.

3.4 Handling multiple objectives

Often there are multiple objectives to be maximized or minimized. While there is an extensive literature on multi-objective optimization, ultimately it will become necessary to combine these metrics into a single utility function that requires assigning weights to each metric. The metric at the top of the pyramid (which tends to be one that needs to be maximized or minimized) typically serves as the base, while other metrics are weighted relative to the top metric.

When multiple metrics have to be combined into a single utility function, it raises the issue of how to weight them. We recommend that the weight on the metric at the top of the pyramid be set equal to 1.0, which means that the weights of other metrics (which are not necessarily in the same units) have to be scaled relative to the top metric. Initially these weights can be set subjectively, but eventually this will lead to a set of decisions that produce a level of performance in each of the dimensions that is being maximized or minimized. If a domain expert is not happy with the

performance in some dimension, the usual path is to adjust the weight and then reassess after seeing a new set of decisions.

3.5 Average performance vs. risk

If we run 20 simulations to evaluate some process for making decisions, we are evaluating our method based on average performance. We can do this if we have access to a simulator, but an alternative is just to watch how it works in the field for a period of time. In this case, we are following a single sample of observations and using the actual performance to evaluate our method. We could say that watching the actual performance is like taking an average of just one observation.

The actual performance in the field is what we experience. For companies, this is captured in their profit and loss statements, as well as any other reports summarizing their other performance metrics such as the various financial KPIs, along with statistics on inventories, and the utilization of facilities and equipment. In a health setting we might be looking at the average rate of new infections or deaths from overdoses. A hotel will look at room utilization and revenue. Truckload fleets will collect statistics on revenue per driver and empty miles.

Now, consider the problem of a sudden disruption in the normal operations of the company. It could be an earthquake or tsunami that destroys a major manufacturing plant, or the emergence of a disease such as COVID disrupting consumption patterns. A tariff war could break out, severely disrupting global trade.

We have already recognized the presence of different sources of uncertainty, as we did throughout Chapter 2, so why are we drawing attention to these new sources of uncertainty? Isn't it the case that if one of these major events happens, that its effect will be captured as we accumulate our performance metrics over time?

The simple answer is: no. Imagine that there is a major disruption in a supply chain so that we have to go through a period of time where we cannot serve our market. Most important is that we may lose customers to competitors as they may not be willing to wait until the problem is fixed. In addition, we may have to furlough significant numbers of employees because we do not have the parts needed to run the factories. This is a hardship for the employees, leading to dissatisfaction, and the best employees may find

Figure 3.2: A sample of books on supply chain risk and resilience.

better jobs.

These issues are not captured by the usual accounting processes that track corporate performance. It is this reason that there is an array of books addressing what is widely referred to as "risk" (or "resilience" which refers to the ability of companies to bounce back from major events) as shown in figure 3.2. These books are typically qualitative descriptions of different types of risk, often (but not always) without a formal process for handling risk.

Risk is a term that is generally used whenever decisions have to be made in the presence of uncertainty. Some examples of risk that might arise in the context of applications in chapter 2 might be:

- Bud Light once introduced a marketing campaign aimed at the LGBTQ community. Their sales dropped by 25 percent as many of their conservative customers reacted badly, which represented a severe disruption to their entire production process.

- The scheduling of power generators considers the potential that a generator (such as a nuclear power plant) might fail, but they would not be able to handle two failures of this magnitude, which would result in rolling blackouts.

- An incorrect diagnosis for a patient might result in the patient's death.

- A failure to provide sufficient cash reserves for a manufacturer could result in bankruptcy given a major drop in the economy, as happened in 2008 with the auto industry.

These events are simply not properly accounted for by accumulating the usual performance statistics. For this reason, it is necessary to account for these events, which may have very low probability, separately from average or actual performance.

The management of the power grid provides a nice illustration. The companies that manage their grids are required to schedule enough power to handle the event that their largest generator (which would be a nuclear power plant) fails, which would produce blackouts. There is no attempt to quantify the economic impact of a blackout. Instead, they simply put it in a separate category, and require that they can handle a major outage.

The literature on risk can be roughly divided into two categories:

- General domain-specific discussions, such as the books in figure 3.2 for supply chain applications, which typically provide lists of events that people in the field would agree constitute "risk."
- The mathematical research literature, largely focused on finance, which uses well-defined risk metrics such as the probability that the financial return falls below some target (typically represented as "VaR" or "CVaR"), or simply the standard deviation of a performance metric (such as the financial return).

The first group, which consists of the types of books shown in figure 3.2, use plain English to describe events that most would generally agree represent examples of risk that should be avoided. The second group consists of books and papers that are often highly theoretical, but which limit their characterizations of risk to extreme values of well-defined probability distributions.

Surprisingly, neither literature provides what could be described as a formal definition of risk that applies broadly to the wide range of settings in which risk seems to be an issue. We offer such a definition here, but we begin by providing a formal name for our original objective, which is based on a simulation of our process, either in a simulator or in the field:

The base objective – This is how we would evaluate our decision-making process over time in the field through the normal accumulation of performance metrics (including, but not limited to, profit and loss statements).

For companies this is typically (but not always) in monetary units, but it could be deaths in a public health setting, loaded miles per driver for a trucking company, and votes in a presidential election.

Now we are ready to define risk:

Risk – Risk consists of two dimensions:

- Risk events – These are events which in the subjective judgment of domain experts (managers, physicians, politicians) are not properly captured by the base objective. Risk events are not quantitative measures – they are characterizations of events in English.
- Risk metrics – This is where we turn a risk event into one or more metrics that quantify the impact of an event on the current or long term performance of a system. Risk metrics are not necessarily in the same units as the base performance objective. Risk metrics can be added to the objective using a scaling factor, or handled as limits.

In most applications, risk is captured as its own metric, which might be power outages exceeding some limit, a drop in the supplies of parts that would require a production shutdown, or events that lead to serious health outcomes. Most of the time the goal is to keep risk below some user-specified limit (we assume that we are always trying to reduce our risk metric(s)).

Interestingly, the mathematical literature on risk usually combines the base objective and the risk metric into a single utility function using a tunable risk parameter. While this may possibly be a valid approach to combining two metrics, we are not ready to make this assumption.

We note that while the base objective is always an average or sampled estimate of an average, a risk metric is often computed not as an average (or expectation), but rather as an event that may happen, possibly with a completely unknown probability.

3.6 At a point in time vs. over time

There is a vast literature on problems that make decisions over time. Inventory management is clearly a problem that has to be solved over time, balancing inventory holding costs with the possibility of stockouts as new orders become known. But what if we have a problem of assigning drivers

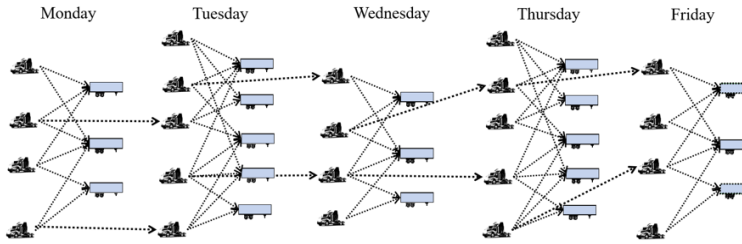


Figure 3.3: Here we illustrate the reality that the assignment problem has to be solved repeatedly. Further, the decisions made at one point in time has an impact on future problems.

to loads, or balancing an investment portfolio, or deciding where to build warehouses? In the 1950s solving any of these problems at a single point in time represented a major challenge.

Today, we have software packages that can solve even large instances of these problems very quickly, but this still leaves us with the challenge of making decisions that work well over time. For example, assigning a driver to a load that goes to Montana, which is very isolated, can create problems when the driver finishes the load and needs to find another load. We have to think about whether we even want to accept the request to move this load, and if we do, which driver do we assign to it? Solving sequences of assignment problems is depicted in figure 3.3. Our stock portfolios have to work well even as asset prices vary over time, and warehouse locations have to anticipate future demand patterns.

At the time that this book is being written, we understand that inventory problems have to be optimized over time. However, the community that specializes in optimization models for complex problems, such as the driver assignment problem, portfolio management problem, and warehouse location problem, each have to reflect the effect of new information, and the impact of decisions now on the future. We have powerful software packages to solve these problems at a point in time, but nothing to optimize performance over time.

When we have problems that have to be solved repeatedly we have to capture:

- The impact of a decision now on later decisions.
- The arrival of new information that we did not know in advance.

The arrival of new information introduces a significant complication when making decisions over time (and virtually all problems that are solved over time have to do so in the presence of new information). For example, in our assignment problem the new information might be the arrival of new loads to be moved. The new loads to be moved on Tuesday would not be known when we are making decisions on Monday. As a result, if we are assigning drivers on Monday, the loads that might be called in on Tuesday are uncertain. We might also say that the loads to be called in the future are "random" (or "stochastic," a term favored by the mathematical modeling community).

For the problem of optimizing the assignment of drivers to loads, we have to capture the arrival of new loads, and how the driver assignments on previous days affects the status of drivers today. To evaluate performance, we would run a simulation that would consist of the steps:

- Start by solving the problem on Monday with the loads that are known at that time.
- Step forward to Tuesday, and then observe the loads that happen to be called in.
- Optimize the assignment of drivers to loads on Tuesday using what is known.
- Step forward to Wednesday and repeat the process.
- Repeat until we reach the end of our simulation period.

Now imagine that we repeat this entire process again starting over on Monday, but as we step forward, we sample different sets of loads being called in. This means that we could simulate making decisions over time, and we would get completely different results.

There are different ways of making decisions to assign drivers to loads that might account for the impact of decisions now on the future. Imagine we have three methods. We could evaluate each method by running repeated simulations and then taking an average. For example, we might perform 20 simulations of each method for making decisions and use this average to evaluate the methods.

Problems that involve making decisions over time are routine; in fact, it might be the case that they represent the vast majority of decision problems. Every single application in Chapter 2 is a sequential decision problem. For example:

- Inventory planning – We have to repeatedly order inventory which arrives after some (typically random) lead time, during which we still have to satisfy orders that arrive. The rule we use to place orders (typically known as an "inventory policy") has to be evaluated over time. If lead times are, say, three months, we would need to simulate the policy for several years, and do this repeatedly.

- Pricing and advertising decisions have to be made over time, as we observe how the market responds to these incentives. A decision at one point in time yields information (such as market response) that can be used to inform future advertising choices. At the same time, these decisions deplete the advertising budget.

- Decisions for generating and storing power have to be made over time as we observe variations in the weather, generator failures, and how the public responds to changes in the weather. Decisions about which generators to turn on or off changes the physical state of the system in the future.

- Medical treatments involve decisions about running tests and experimenting with different treatments to see how the patient responds.

- The allocation of naloxone pens to handle opioid overdoses have to be made over time as we observe how health officials and drug users adapt to the availability of this resource.

- Presidential campaigns have to make decisions about advertising and scheduling candidate visits while looking at polls to see how voters are responding.

- Mutual fund managers have to adjust how much cash they keep on hand while observing changes in the market, and the pattern of deposits and withdrawals that their customers are making.

It is somewhat surprising that while the literature on solving static decision problems is incredibly mature, the academic research community that works on these problems has not adopted a standard framework for modeling and solving sequential problems.

We are going to return to these issues in Volume II when we start using a little notation. Without notation, the discussion reduces to a lot of hand-waving.

3.7 Psychological performance metrics

Virtually the entire optimization literature assumes that there is a well-defined performance metric, called the objective function, that can be used to evaluate decisions. The objective function may not be known exactly, but we assume that the uncertainty can be quantified or at least sampled. By contrast, most of the literature on the psychology of decision making focuses on how people evaluate complex alternatives, which is probably explained by the fact that these are the most interesting and challenging decision problems.

In this section we will start by identifying some complex metrics, followed by a sampling of theories of how people handle these complex metrics. We close with a brief discussion of how brains "optimize."

3.7.1 Complex metrics

Some examples of complex decision problems from the applications in chapter 2 include:

- What is the best supplier for a complex component for a jet engine that requires special expertise in materials?
- What is the best way to market a consumer product to maximize sales?
- What is the best medical treatment to handle stage 3 lung cancer?
- What is the best allocation of resources in a presidential election (marketing, trips to give speeches).
- What is the best strategy to market a complex dispatch system to truckload motor carriers?

Each of these can be presented as an example of an "intelligent trial and error" problem (section 2.12) where there is a set of discrete choices. What makes these choices hard is a) they are important and b) we have considerable uncertainty about how well each will perform.

We can divide problems with complex alternatives into three classes:

a) We know the metrics we want to use, but do not know their values.

b) There are multiple metrics, but we do not know (precisely) their relative importance.

c) We are not even able to articulate some or all of the metrics to evaluate each choice.

(a) has attracted considerable attention from the optimization literature, but it remains a very common context that people face frequently, where they fail to use the best methods for handling uncertainty. (b) is another common topic and typically involves posing different alternatives to a decision-maker who is then asked to make a choice. (c) is a common problem in the psychology literature since there are problems (such as those listed above) where someone may have a gut feel for what choice they want to make, without being able to articulate why it is best.

3.7.2 Some theories for metric formation

Examples of different theories for evaluating alternatives include:

Prospect theory - Key principles of prospect theory (from (Kahneman & Tversky 1979)) include:

- Loss aversion – People feel the pain of losses more intensely than the pleasure of equivalent gains. For example, losing $100 feels worse than the joy of gaining $100.
- Reference dependence – Decisions are made relative to a reference point rather than absolute outcomes. Gains and losses are perceived in relation to this reference.
- Risk aversion in gains, risk seeking in losses – When faced with potential gains, people tend to prefer certain outcomes over risky ones. However, when dealing with losses, they often take greater risks to avoid a definite loss.
- Diminishing sensitivity – The impact of changes in wealth decreases as amounts grow. The difference between losing $100 and $200 feels more significant than between losing $1,000 and $1,100.
- Probability weighting – People overestimate the likelihood of rare events (e.g., winning the lottery) and underestimate the probability of common events.

Mental accounting theory - This refers to the way people mentally organize, categorize, and evaluate financial decisions by creating sepa-

rate "accounts" in their minds, rather than treating money as fully interchangeable/fungible (Thaler 1985). Some key concepts include:

- Categorization: People assign money to different mental budgets (e.g., rent, groceries, entertainment) and often make spending decisions based on the category rather than overall financial position.
- Framing: The same amount of money can be valued differently depending on how it was acquired — for example, a $100 windfall may be spent more freely than $100 earned from work. This explains why some people might splurge with a tax refund while being strict about everyday expenses.
- Sunk cost fallacy: People often continue with a losing endeavor (like attending a bad concert they paid for) because they have mentally "spent" the money, even though it is unrecoverable.

Attribution theory - Attribution theory explains how people interpret the causes of their own and others' behavior, especially in achievement contexts like success or failure. For example, consumers assign reasons to their purchasing decisions, influencing brand perception and loyalty (Weiner 1986). Three dimensions of causal attributes include:

- Locus – Is the cause internal (e.g., ability, effort) or external (e.g., luck, task difficulty)?
- Stability – Is the cause stable (consistent over time) or unstable (variable)?
- Controllability – Can the person control the cause (like effort), or is it uncontrollable (like innate ability or luck)?

These attributions influence emotions and future motivation. For example:

- Attributing success to internal, controllable factors (like effort) boosts motivation and pride.
- Attributing failure to internal, uncontrollable factors (like lack of ability) can lead to shame and discouragement. Weiner's theory is widely used in education, sports, and organizational settings to understand how beliefs about causes shape behavior and performance.

3.7.3 How the brain learns to optimize

There has been a tremendous tendency to attribute intelligence to the neural networks that are used to learn word patterns. While recognizing patterns is in fact an important form of intelligence, it is distinctly different from the process of making decisions, which humans are quite capable of doing. Making decisions requires the ability to maximize rewards that are specific to achieving some goal that might be related to eating, being comfortable, avoiding pain, winning a contest, or solving a problem.

It turns out that the brain has very specific functions that help with optimizing an objective that has nothing to do with simply matching a pattern. This is done with parts of the brain known as reward receptors, which are specialized proteins that respond to neurotransmitters, such as dopamine, to help the brain experience pleasure, motivation and positive or negative reinforcement. Neurotransmitters are like molecular locks that are activated when the right chemical key binds to them, which then triggers neural activities that guide behavior.

Types of reward receptors are:

- Dopamine receptors – These are the most important elements of the reward system, which come in different forms:
  - D1 receptors, which promote reinforcement learning and help maintain long-term motivation.
  - D2 receptors, which are involved in social bonding and mood regulation.
  - D3 receptors, which play a role in motivation and goal-directed behavior.
- Opioid receptors – These respond to endorphins and other natural opioids, contributing to feelings of euphoria and pain relief.
- Serotonin receptors – These influence mood and emotional reward and are often used as a target by antidepressants.
- Glutamate receptors – These help encode reward-related learning and memory.

Obviously any discussion of these incredibly complex mechanisms is far beyond the scope of this book. The point we are making is that the brain has specific mechanisms for maximizing rewards, which is how it can choose

the best decision. This is a process that is distinct from the brain's powerful mechanisms for identifying patterns. By contrast, the neural networks used by large language models are trained to identify patterns or text; these use a single objective function, which captures the similarity between a function (the neural network) and the training dataset.

3.8 Setting performance goals for others

An important dimension of metrics is setting goals for the purpose of evaluating the performance of people or groups. This inherently implies a multiagent setting, where one decision-maker has the authority to set performance goals for another unit in an organization. In a multiagent setting, a goal of one agent may be the decision made by another (presumably higher level) agent.

The setting of goals is a particularly rich area in the context of organizations with multiple decision-making units, typically organized in a hierarchical fashion. We will return to this topic in a future volume.

3.9 Exercises

Review questions

3.1 — Name five examples of performance metrics.

3.2 — What is meant by objectives, targets and limits? Give an example of metrics that would fall into each of these three categories. You may use metrics from any of the examples in chapter 2 (they do not all have to come from the same example).

3.3 — What is a risk event? Give examples of risk events if you are:

a) The operator for the power grid for a region.

b) A physician working with a patient to manage their diabetes.

c) The chief financial officer for a supply chain.

3.4 — Name two theories for how people evaluate choices.

3.5 — Name four receptors that the brain uses to reward specific behaviors.

3.6 — For each of the examples of risk events in exercise 3.3, design a risk metric for that event.

Modeling questions

For each question below, design a metric pyramid using the metrics provided for the designated application, given the decision maker specified.

3.7 — You are a supply chain manager in charge of picking suppliers for the components of an air conditioner. Suppliers may be anywhere in the world.

3.8 — You have to restock inventories of different types of furniture for a furniture retail outlet.

3.9 — You have to plan investments in new power generating capacity (these orders are placed up to five years in the future).

3.10 — You are the revenue manager for a hotel.

3.11 — You are the physician choosing the treatment for a diabetic patient.

3.12 — You are the state official who has to allocate naloxone kits across counties in your state.

3.13 — You are a drug company who has to choose which drugs to put into Phase II trials.

3.14 — You are the campaign manager for a presidential election.

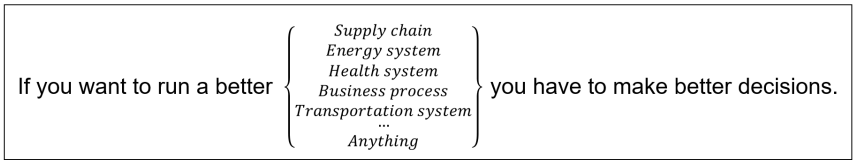
3.15 — You are the vice-president of operations for a truckload carrier who has to manage dispatchers (who assign drivers to loads) and load managers (who handle which loads to move).

3.16 — You are the mutual fund manager who has to determine how much cash to keep on hand.

Chapter 4

Decisions

This entire book is based on the statement:



It goes without saying that before you can address the problem of identifying the best decisions, you have to know what decisions you are making.

Remember from Chapter 1 that there are two types of "problems": metric-focused and decision-focused. Examples of each are:

- Metric-focused problems:
  - Supply chain management – Minimize inventories, maximize operating margin.
  - Electric power grid – Minimize energy generation costs.
  - Public health – Minimize deaths.
  - Managing a truckload fleet – Maximize net operating revenue per driver per week.
  - Managing a hotel - Maximize operating profit.
  - Running a presidential campaign - Winning the election.
- Decision-focused problems:

- Supply chain management – How much to order, which supplier to use.
- Demand management – How to price a product, what marketing channels to use.
- Electric power grid – Which generators to schedule for operation, which gas turbines to use.
- Managing diabetes – Which medication to use to control blood sugar, what dosage.
- Mutual fund cash management – How much cash to keep on hand to handle redemptions, what stocks to invest in.

If we start with a metric, our challenge is to identify the decisions that will help us improve the metric. If we start with decisions, then the problem is to design the metric. However, even when we think we know the decisions, we have to be sure that we have not missed any.

There is an extensive mathematical literature on the topic of optimizing decisions, but even these books lack a standard definition of what a decision is. Instead, the books will introduce notation such as decision vector "x," or control "u," or action "a," after which they will give examples and hope that the reader "gets it." While this works for simple problems, it creates a barrier between the mathematical model and real applications.

For complex applications such as managing supply chains or solving public health problems, identifying decisions is much more challenging than identifying metrics. This is not meant to trivialize the identification of metrics, but the concept of metrics is well understood by both domain experts and modelers alike. When asked what decisions are involved, business executives, medical professionals, engineers and scientists often return blank stares. While the word "decision" is familiar to everyone, it does not appear to be a term that they use in daily problem-solving, while everyone understands "metrics" in one form or another.

4.1 Decisions and the English language

It seems as if a good starting point for a chapter on "decisions" would be to offer a definition. It helps to note that standard definitions such as those by Webster will include the usual variety of meanings for a word as it is used in the English language. For example, winning a baseball game is referred to

as a "decision." In this book, we only use "decision" to refer to situations where we have a set of choices, and we have to make the best choice which, of course, implies the identification of performance metrics.

We start by noting that decisions are always a form of information. It helps to put all information into three broad classes:

1) Information that we already know at a point in time. We refer to this information as the state of our system (more precisely, the state of knowledge).

2) Information that we control that changes the state.

3) New information that arrives that we do not control (although we may influence it).

Information in classes (2) and (3) produce an updated state variable (class 1). We are now ready to define a decision:

Definition (formal): A **decision** is an endogenously controllable information class.

So, decisions (which are contained in "decision variables") represent information that we create by naming one out of a set of choices.

Our formal definition requires a lot of overhead for what should be a very simple concept, so we offer a second definition:

Definition (informal): A **decision** is something we control.

This definition avoids "information class" by using "something," but it gets the point across.

Both our definitions raise the question of who is making the decision, which is inseparable from the concept of a decision. Classical mathematical models avoid this question, but it is central to the modeling of most real systems.

Given the importance of decisions in human activities, it should not be surprising that there are a number of terms in English that capture the concept of a choice. Table 4.1 lists a number of words that imply making a choice in a general setting. Under the column "collecting information" are terms that arise when deciding what experiment to run, who to listen to, what to observe (and so on). The column labeled "Acting on resources" list a variety of terms that arise in the context of managing resources (such

General terms	Collecting information	Identification decisions	Acting on resources
Action	Experiment (which?)	Identify	Promote (who, how much)
Choice	Listen (to what?)	Classify	Acquire (which, how much)
Control	Observe	Finding	Sell (to whom, how much)
Decision	Test (which one)	Conclude	Reward (how much)
Design	View/scan	Label	Criticize (who, how)
Intervention (medical)			Move (to where)
Option			Trade (which, to whom)
Move (where)			Treatment (which one)
Response (which one)			Accept/decline
Task			Recommend
Trade (finance)			

Table 4.1: The English language offers a variety of words that all mean the freedom to choose.

as people). For example "promote" implies the decision of whether or not to promote someone (and to what level).

This table is not intended to be a comprehensive list of words that imply a choice, but it hints that "decisions" come in many ways in the English language.

4.2 Identifying decisions

Understanding all the different words that imply (or require) making a choice is important when identifying the decisions that are available to be made. It is important to recognize that decisions do not come with bright labels attached to them. Campbell's Soup Co. recognized the challenge of making consumers aware that they were making decisions in a famous series of commercials in the 1970s labeled "*I could have had a V8!*". Their marketing department realized that people would often pick up a can of soda without realizing that they could have chosen to have a V8 instead. The commercials helped to make consumers aware that drinking a soda was a decision.

People in almost any problem setting fall into the habit of solving problems a certain way, without realizing that they have choices. One might say that this is how we get through the day, since evaluating choices to identify the best one takes time. The challenge we face is to first be aware of when we are making a decision, and then identifying the decisions that have the

Figure 4.1: A challenge is to work in any of a variety of problem settings (such as those on the right) and identify the decisions that are being made.

biggest impact on performance.

The behavior of passively making decisions is absolutely pervasive, but this creates an opportunity. Imagine that you are in any problem setting (such as those illustrated to the left in figure 4.1). Now assume that you want to improve performance, whether it is profitability, productivity, improved health outcomes, better drugs, or improving agriculture. Then remember our basic line:

> *If you want to run a better anything you have to make better decisions.*

To make a better decision, you have to recognize when you are making a decision. A good exercise is to create your "decisions book" and then make mental notes as you recognize when a decision is being made (which is to say, there was a choice, and different choices could be made).

4.3 Types of decisions

Our approach to framing requires being able to identify all decisions, not just decisions that can be handled by a particular methodology. To guide this process, we list below six types of decisions which, to our knowledge, cover every form of information that we control.

1) Physical and financial decisions – These decisions arise in the management of physical and financial resources, spanning people, equipment, facilities, products, commodities, water, energy, in addition to

cash, investments, loans, ... Decisions include buying, selling, moving, and modifying resources. This class is the domain of operations research, engineering control, and finance, and draws heavily on tools such as linear, integer and nonlinear programming.

2) Discrete actions – This is a general term designed to cover activities that may involve complex projects such as launching a new product, submitting a drug to clinical trials, or purchasing a company. Sometimes called "projects," these may involve a series of changes to performance metrics, resources, finances, and system dynamics. Special cases can be simpler problems, such as picking a price or who to hire for a leadership position. These problems are popular in the decision analysis literature, and typically involve relatively small sets of actions that are difficult to evaluate.

3) Information acquisition/observation decisions – These include decisions to acquire or observe information by running experiments in the lab, field, or with computer simulations. It helps to distinguish two settings in which we may acquire information:

   – Offline learning - These are activities that are conducted in a test environment. Offline information acquisition can include research efforts, internet searches, or hiring domain experts.
   – Online learning - This covers decisions to run and observe processes in the field using a "learning while doing" approach, which involves observing a process as it evolves, such as how a market responds to advertising or pricing, or how a patient responds to a treatment.

   Both styles of information acquisition imply making decisions specifically to acquire information. Information acquisition has been studied under names such as design of experiments (static or sequential), stochastic search, active (or optimal) learning, multiarmed bandits, and Bayesian optimization.

6) Information communication/sharing decisions – These come in two forms:

   a) Messaging – This reflects what we say in text, video and/or audio. A modern example of messaging includes prompt optimization.

b) Channels and timing – This reflects the choice of channel (text/emails, publication (print or online), social media, or advertising channels) along with the timing and frequency.

5) Choosing functions – Often overlooked as a decision, functions may be methods to make decisions (policies), the formulation of optimization models, the choice of performance metrics, methods for forecasting or estimation, or transition functions (such as how disease spreads). This category covers the choice of function, which means its structure.

6) Setting parameters – Functions are typically characterized by one or more parameters (typically continuous, but not always) that can be tuned to improve predictive accuracy (when fitting statistical models) or optimized to improve performance (when tuning a policy for making decisions). Parameters may be associated with a function; they can be the weight on a performance metric, or they could be a target (or limit) for a performance metric.

4.4 Flavors of decision variables

Decisions are going to come in different styles, but decision variables can typically be put in one (or more) of the following categories:

- Binary – Here we have just two choices, which might be:
  - To perform an action or not.
  - To hold or sell an asset.
  - A/B testing for webpage design, where we need to choose between a current design and a new or modified design.
  - Whether to continue testing a drug or treatment in a clinical trial, or terminate the trial.
- Discrete set of choices or actions – This is easily the most common form of decision problem, and it arises when we have a set of discrete choices or actions, such as:
  - Choosing a supplier for a part.
  - Choosing a drug or medical treatment.
  - Choosing a marketing channel.

  - Choosing a location for a facility.

- Continuous scalar – Examples are:
  - Setting the price of a product.
  - Choosing the dosage of a drug.
  - Deciding how much to spend on advertising in a market for a presidential campaign.
  - Choosing how much cash to keep on hand for a mutual fund.

- Discrete vectors – There are many problems that involve the management of discrete resources such as people, machines, and jobs. When we have a single set of discrete choices such as where to purchase a product, it is easy to enumerate all the choices. But when we have to decide how to schedule, say, 100 machines to handle hundreds of jobs, then we need specialized algorithms.

- Continuous vectors – There are problems with a small number of continuous decisions, such as controlling a car, aircraft or rocket. Then there are problems with large numbers of continuous parameters, such as allocating funds across many asset classes, or allocating large numbers of naloxone kits to a hundred counties in a state. There are powerful search algorithms for solving these problems.

4.5 How decisions impact the system

It does not make sense to talk about "decisions" as an abstract concept. We first recognize that a decision changes the system in some way, but how?

There are three ways that a decision can impact a system:

- Physical resources – This where we buy, sell or modify in any way anything physical, which could be people, equipment, facilities, food, water, or energy.

- Financial – This can be cash, investments, and loans; insurance contracts, and currency hedges; and prices.

- Informational – This is a category that might include a decision to run an experiment in a lab, computer simulation, or a field test which

is used to update estimates or beliefs; it might involve setting performance targets, designing metrics, or specifying the terms of a sales contract.

There is some overlap in the categories, such as the distinction between currency hedges and the terms of a sales contract. What is important is the breadth of ways that we can affect how a system evolves over time.

As we progress in our modeling framework, we are going to need to understand the following about any decision:

- How does the decision affect our performance metrics now?
- What effect will a decision now have on the state of the system before making the next decision?
- Will the decision impact new information that arrives after the decision is made?

In Volume II we describe these points using mathematical notation.

4.6 Timing of decisions

One of the most important but challenging attributes of decisions involves time, specifically:

- How frequently decisions are made – We can divide decisions into two broad classes:
  - Design decisions, which are made just once (initially) over the planning horizon. In practice, even design decisions evolve over time, but it is common to have decisions that are made just once within what is considered a reasonable planning horizon.
  - Control decisions – These are decisions that are made repeatedly over time, but there are complex systems where a variety of decisions are being made at different time intervals. For example, grid operators plan the scheduling of steam generators once each day; gas turbines are planned hourly; adjustments to the speed of certain generators are made every 5 minutes; and signals to adjust the voltage levels are sent every 2 seconds.
- Lag times – When a decision is made, there is often a lag before it impacts the system. For example:

  - Ordering inventory may require weeks or months to arrive.
  - Administering a drug might take minutes, hours or days before it affects a patient.
  - Grid operators plan the schedules for running steam plants the day before, while decisions to turn gas turbines on require 30 minutes notice.
  - Pricing changes might not be seen in sales for days or weeks, and may affect markets for months.

- Advance planning of lagged decisions – In addition to the dimensions of when a decision is made and when it impacts the system, we have to think about the timing when we are planning into the future. For example:

  - A manufacturer may face lead times of eight months when ordering from Asia, but can get much faster turnarounds for smaller quantities (at higher cost) when there are shortages. When thinking about how much inventory to keep on hand, the manufacturer would have to keep much larger inventories without the option of ordering from the high cost but closer supplier. However when this option is available, the manufacturer can consider the option of using the closer supplier in the event of a surge in demand.
  - Airlines often need to plan aircraft purchases up to 10 years into the future, but can negotiate faster deliveries at a higher cost. This allows the airline to consider this as an option if passenger volumes rise faster than planned. Or they can cancel contracts at a cost depending on how long they wait to exercise this option.

4.7 Who makes decisions

There are many settings where there is more than one decision-maker (or agent). Examples of multi-agent settings include:

- Two equal decision-makers (often called players) as might happen in negotiations between a manufacturer and a supplier or a customer, or in interactions between a physician and a patient.

- Two decision-makers where one has a controlling position. For example, a "field agent" may ask for resources from a "central agent" who has control over how much of the request to satisfy.
- Several agents, as might arise when a few companies are competing against each other (examples arise in industries that sell cars or industrial chemicals), or when there are multiple organizational units at the same level in a company.
- Multiple agents, as arises in a supply chain with different manufacturers providing components to make a part such as an engine or entire car.
- A single agent learning about an unknown environment, which is how we might model any problem involving uncertainty. The unknown environment could be the weather, the presence of disease in a population, or a market purchasing a product.

We return to multiagent problems later in the series, where we show how to extend the notation (presented in Volume II) to handle multiple decision makers. For now, we are going to focus on a single decision-maker, who may be one of two or more decision-makers.

We have several reasons for avoiding the explicit identification of decision-makers at this stage:

- The organization of decisions can vary, even within the same industry (such as trucking or supply chain management) or problem domain (such as public health).
- If your goal is to develop a computer model, you may be looking to change how decisions are organized. A modeler may wish to treat a set of decisions as if they are being made by a single agent either as a simplification, or because if may produce better results.
- The goal of listing different types of decisions is not to tackle all of them in a single modeling project. Rather, it is necessary to identify the goals of a model and then choose the decisions that are relevant to the goals of the project.
- We recommend that the reader approach these projects from the perspective of a single decision maker, which does not necessarily have to align with how decisions are actually made within an organization.

This will be supported by the initial presentation of the universal modeling framework in Volume II.

For now, when faced with a multiagent setting we recommend treating each agent separately to identify their own metrics and decisions. Uncertainties often affect the broader environment, although each agent may have uncertainties that are relevant to their own decisions and performance metrics.

4.8 Making decisions with computers

Computers have a very straightforward way of making decisions. It starts by knowing the types of decisions and the set of possible (feasible) decisions. It then uses a prespecified method to "make" the decision, which means a particular choice from the set of feasible (or allowable) decisions.

We start by introducing how we refer to these decision-making methods:

Definition: A **policy** is a method for choosing an allowable decision using the information that is available at the time the decision is made.

There are two broad strategies for designing policies, each of which can be divided into two classes, creating four classes of policies that include *any* method for making decisions. These are:

Policy search - This strategy creates functions that have to be tuned to work well over time. They make decisions without directly planning into the future. These can be divided into two classes:

1) - Policy function approximations, or PFAs.

2) - Cost function approximations, or CFAs.

Lookahead policies - This strategy tries to make the best decision now by optimizing over the performance of the decision now, plus an approximate of the impact of the decision now on the future. These can also be divided into two classes:

3) - Policies based on value function approximations, or VFAs.

4) - Direct lookahead approximations, or DLAs.

Each of these policies is described below.

4.8.1 Policy function approximations (PFAs)

Policy function approximations (PFAs) represent any analytical function which, given inputs from what we know, produces as an output the action we should take. Some examples are:

- Inventory ordering policies often place an order when the inventory falls below a level "s" at which point they place an order to bring the inventory up to "S." "s" and "S" are parameters that have to be tuned.

- A physician may prescribe insulin injections when a patient's A1c (which reflects a 3-5 month rolling average of blood sugar) goes above 6.5, and stops when it falls below 6.0. Again, these numbers need to be varied to find the values that work best.

PFAs can be any analytical function such as a linear or nonlinear function. What a PFA cannot include, which will be found in each of the three remaining classes of policies, is an imbedded optimization problem. PFAs can be simple rules, but they can also be very high dimensional nonlinear functions such as a neural network.

4.8.2 Cost function approximations (CFAs)

There are many problems where the best approach for making decisions is to use a deterministic approximation at one point in time which has been modified using various parameters which, when properly tuned, produce decisions that work well over time. This is an approach that is widely used in practice, although often without recognizing a) the ability to introduce parameters to help improve the decisions and/or b) a failure to recognize that the parameters can be tuned to produce better results.

The simplest example of this approach is illustrated in Figure 4.2, where we need to choose which product to advertise on social media (we could substitute any problem with discrete choices listed in section 2.12). We have a point estimate for the value of each product based on past experience, which we have learned can involve a lot of noise, resulting in some poor estimates. We can also use past experience to estimate a standard deviation which is a measure of the spread of uncertainty. Typically we are 95 percent sure that the truth is within plus or minus 2 standard deviations.

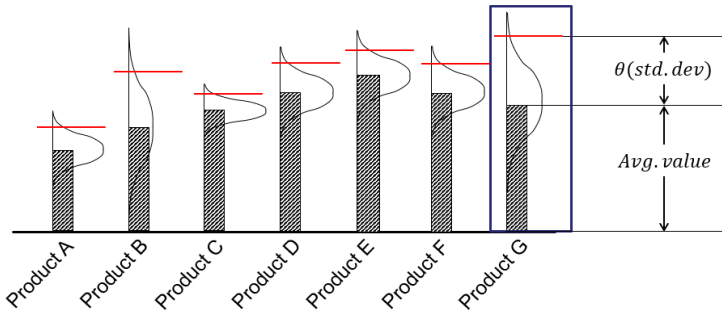


Figure 4.2: When faced with a discrete set of choices, the uncertainty in the belief about how each performs may be described by its average value, and the standard deviation which captures the spread of the belief.

What we are going to do is to create an "index" for each product x given by:

$$Index_x = Avg.value_x + \theta(std.dev_x)$$

We are then going to choose to advertise the product x that has the highest value of "$Index_x$". We find product x by solving the following optimization problem (which is deterministic):

$$\max_x \{Avg.value_x + \theta(std.dev_x)\}$$

Solving this optimization problem is fairly simple – we just have to sort the values $Avg.value_x + \theta(std.dev_x)$ and find the product x that has the highest value (and here you thought deterministic optimization had to be hard!).

The challenge, then, is choosing the tunable parameter θ. If we use $\theta = 0$, then that means we are just using our current estimate. The problem is that if our estimate "$Avg.value_x$" is low because of a run of bad luck, we might never try advertising product x again. If we use $\theta = 2$, then we are using a very optimistic estimate of the value of product x which will encourage trying products where there is a high level of uncertainty (which is not necessarily a bad strategy).

The idea of using a parameterized deterministic approximation is exceptionally powerful. Airlines use it when they optimize their schedules, where they have to use an estimate of the weather delays for each flight. If they use the median, then half the time the delay will be more than what

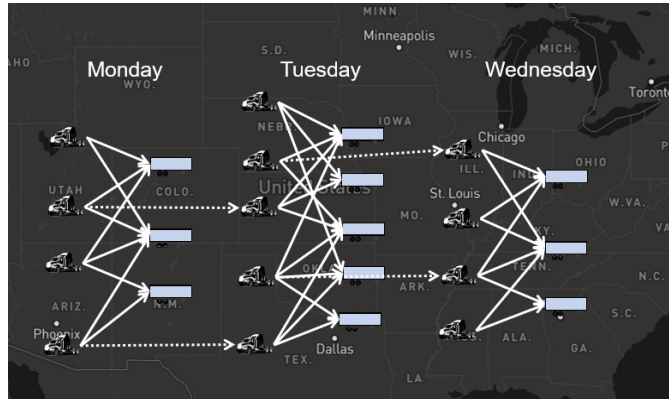


Figure 4.3: Illustration of the problem of assigning trucks over a three day period.

is anticipated by the schedule, which will then produce a large number of late arrivals of aircraft, delaying subsequent flights. However, if we use the 90th percentile, then we may be introducing too much slack in the schedule, resulting in poor utilization of aircraft.

It is easiest to think of tuning a set of parameters θ in a simulator, but it is often the case (such as the airline scheduling problem) where the problem is much too complicated. For this reason, it may be necessary to do online learning, which means testing different values in the field and observing the actual performance.

4.8.3 Value function approximations (VFAs)

Imagine that we are dispatching a fleet of trucks where we have to assign drivers to move the loads of freight from a pickup location to a delivery location. Figure 4.3 illustrates how this problem has to be solved repeatedly over time. What we decide to do on Monday will change the locations of drivers on Tuesday and Wednesday. Each day, shippers call in new sets of loads that are not known in advance, so the carrier has to make assignment decisions on Monday without knowing what is going to happen on Tuesday or Wednesday.

Optimizing over a multiple day horizon in the presence of the uncertainties is an incredibly complex task. Instead, we can approximate the value of drivers in the future, as shown in Figure 4.4. This can be done by running simulations into the future, and then calculating the value of

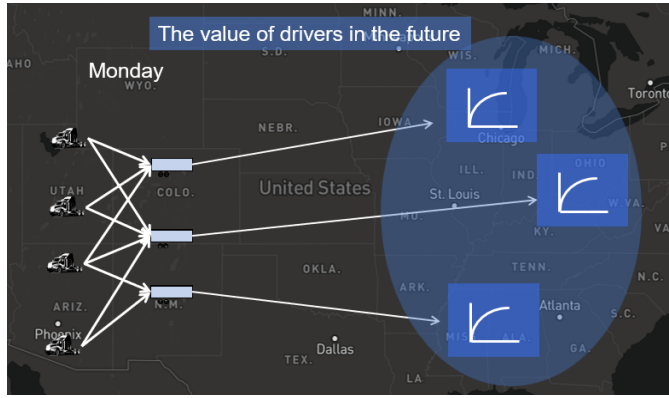


Figure 4.4: Assigning drivers to loads using estimates of the value of drivers in the future.

drivers in different locations. When we include these values (called "value function approximations"), the problem we now have to solve on Monday is no more complicated than if we completely ignored the impact of sending drivers into different locations.

Approximating the value of landing in a particular state is a strategy that is very popular in the research literature, but its success is highly dependent on the structure of a particular problem, and it tends to work well for a small number of specially structured problems.

4.8.4 Direct lookahead approximations (DLAs)

There are many problems where we simply have to plan into the future to make a decision now. One of the most familiar examples of a direct lookahead policy is when we use Google maps to plan a path to the destination.

DLA policies can be divided into two subclasses:

a) Deterministic lookaheads – This is where we use point estimates of any uncertain quantities such as traffic delays.

b) Stochastic lookaheads – Here we want to explicitly model the uncertainty that we face, such as the potential traffic delays that may arise as we are driving to our destination. It helps to further divide this class into two types:

   b.i) Problems with discrete actions - These are problems that are

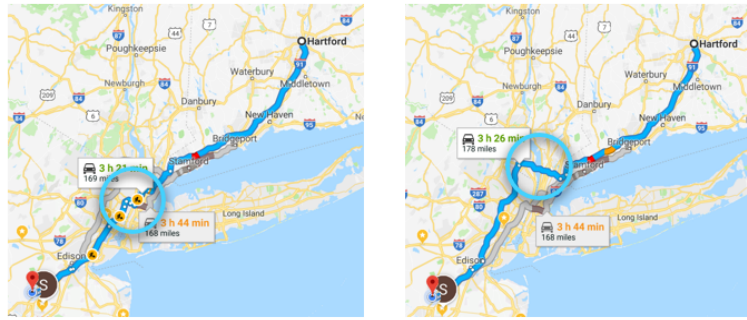


Figure 4.5: Path planned by Google maps based on estimated travel times (left); alternative path based on perceived risk of driving through New York city (right).

typically solved with decision trees.

b.ii) Problems where decisions are vectors - Here we need to use the tools of math programming to search over a multidimensional space.

Note that we do not need to subdivide deterministic lookaheads since, even if the decision in each time period is a scalar, the entire lookahead model requires optimizing over the vector of decisions spanning the time periods over the planning horizon.

Figure 4.5(left) shows an example of Google maps planning a path from Hartford, Connecticut (upper right) to Princeton, New Jersey (lower left), departing at 4pm in the afternoon. Note that the path goes right through New York city, which would occur right around 5pm when traffic is expected to be heaviest. Google maps uses a point estimate, and still feels that this is the shortest path.

Of course, any knowledgeable traveler would understand that there is tremendous uncertainty surrounding the actual travel times through New York at 5pm. Figure 4.5(right) shows an alternate route that Google provides, giving a traveler the opportunity to choose between a path that is expected to be shorter, but with a risk of being much longer, versus a slightly longer path that is expected to be close to the time that Google estimates.

The first path, then, is an example of a deterministic lookahead, but the user can introduce the uncertainty as he assesses the recommendation. By choosing the second path, we are solving, in an admittedly ad hoc way,

a stochastic lookahead.

When we are planning into an uncertain future, there is a wide range of strategies for modeling this process to help make a decision now. One strategy is to use a point forecast (that is, a deterministic lookahead) but introduce tunable parameters that can make the solution more robust.

4.8.5 Hybrid policies:

In addition to the four classes of policies, we can create a variety of hybrids that combine two, three or even all four of the classes. Some examples using a supply chain setting are:

- CFAs with PFAs - Choosing least cost supplier, but with rules to exclude high-risk companies.
- Lookahead (DLA) with VFA - Optimize seasonal production plan, with functions capturing value of ending inventories
- Parameterized deterministic direct lookaheads (DLA/CFA) - Plan seasonal production plan using θ-percentile (say, 80th percentile) demand forecasts.
- VFA policy using PFA - Distribution planning using VFAs to value inventory at each warehouse, but using rules (PFAs) to force deliveries to specific locations.
- VFA with CFA - Start with a VFA-based policy with a linear model, and then tune the parameters of the linear VFA to get the best results using a simulator.

While these policies may sound complicated, it is possible to describe specific settings where human decision making is using each of them. For example, the most complex policy uses a stochastic lookahead, which we illustrated above using the navigation problem with Google maps, where a longer path was chosen to avoid the risk of congestion in New York City.

4.8.6 Which policies are most widely used?

Discussing policies can sound complicated and confusing. It is important to remember that:

a) Everyone makes decisions. We all face situations on a day-to-day basis, whether to get through the day or decisions that come up in our jobs.

b) When we make decisions, our brain is using some method which belongs to one of the four classes (and possibly a hybrid).

Start by dividing the fourth class, DLAs, into two types: deterministic lookaheads and stochastic lookaheads. We are then going to divide the last type, stochastic lookaheads, into two subtypes: problems where decisions are one of a set of discrete choices, and problems where decisions are vectors, such as allocations of assets among investments, or assigning machines to tasks.

This gives us six types of policies that we divide into four categories:

Category 1 - This category includes three types of policies:

- Policy function approximations (PFAs), which include all the simple rules such as "when it is cold, wear a coat" or "buy a product when it goes on sale." PFAs can be rule-based "if in this state, take this action" or it can be an analytical function, a topic we return to in Volume III.
- Cost function approximations (CFAs), which include any method where we have to solve a deterministic optimization problem (typically an approximation of the real problem that involves uncertainty) such as our discrete choice problem in figure 4.2.
- Deterministic direct lookahead approximations (Det-DLAs), where we plan into the future as is done by Google maps, using point estimates of any uncertain quantity.

CFAs and Det-DLAs both involve solving deterministic optimization problems; the only difference is that CFAs do not plan into the future, while DLAs do.

Category 2 - Stochastic lookahead policies where decisions are discrete choices. Here we explicitly model the uncertainty in the evaluation of each choice. These are widely studied using the device of decision trees.

Category 3 - Policies based on value function approximations, where a decision now considers the immediate cost or reward plus an estimate of the future value from transitioning to some state. This is an advanced and computationally difficult class of policies which are needed for a small set of specialized problems.

Category 4 - Stochastic lookahead policies where decisions are vectors. This is a very complex class of problems that requires complex strategies, since a stochastic lookahead is just another stochastic optimization problem, with simplifications introduced to reduce the computational complexity.

The policies in category 1 are used by everyone, regardless of formal training. These policies are the simplest, but this requires the introduction of parameters that have to be tuned, which can be hard.

Human brains have evolved the natural ability to use all four classes of policies, at least in the context of discrete choices. We even know how to switch between the classes without our realizing it. If we are playing chess (and we have some experience with the game), we probably are doing the first few moves from memory (expert players are able to execute quite a few moves from memory). This is a pure PFA. However, at some point we start to think about what our opponent might do, which involves a direct lookahead policy, typically combined with VFAs which can capture the value of losing major pieces.

4.9 Exercises

Review questions

4.1 — Give the formal definition of a decision, the informal definition, and three examples of decisions.

4.2 — Explain what is meant by an "endogenously controllable information class." What are the other two classes of information that are described in the same context?

4.3 — Give examples of decisions that fall in each of the following categories:

a) Binary.

b) A set of discrete choices with at least five choices.

c) There are at least 10,000 different decisions that have to be made at a time.

4.4 — Name five examples of continuous decisions.

4.5 — Give three examples of each of the three types of decisions:

a) Decisions impact physical resources.

b) Decisions impact financial resources.

c) Decisions impact the collection or distribution of information.

4.6 — Name three settings where decisions have to be made at different time scales. Describe the setting and the timing of decisions.

4.7 — Summarize in your own words the four classes of policies, and given an example of each for some problem setting.

Modeling questions

4.8 — Give an example of a decision that falls in each category:

a) A decision that has to be made every minute (or more quickly).

b) A decision that has to be made daily.

c) A decision that has to be made yearly.

4.9 — Identify the decisions implied in each setting, and the decision-maker who makes each decision.

a) An individual has to take medication prescribed by their doctor, who is following protocols worked out by the drug developers.

b) The power grid has to tell a utility which steam plants to turn on, and when. The scheduling decisions are made by a computer model run the day before.

c) A mutual fund manager has to decide how much cash to keep on hand to respond to deposits and redemption requests by individual investors (small amounts) and retail investors (large amounts).

4.10 — Give three examples of policy function approximations. Describe the setting and how the PFA would work.

4.11 — Give an example of a cost function approximation for a discrete choice problem.

4.12 — You are using Google maps to find a path so you arrive at work by 8:30am. You also have to decide how much time to allow so you arrive on time. Describe the decisions being made, and which type of policy is being used to make each one.

4.13 — Describe as many classes of policies you might use if you were going to design a computer program to play chess. Describe how each class of policy you have identified would be applied.

Chapter 5

Uncertainties

Sequential decision problems invariably have to deal with uncertainty, which is typically the most challenging dimension of making decisions over time. There are three ways that uncertainty affects the performance of our system:

1) The decision we make at a point in time is not implemented correctly.

2) The performance of the system given our decision is not the same as what we estimated when we made the decision.

3) The impact of a decision now on the future is not estimated correctly because of changes as we step forward in time.

Although we list only three ways that uncertainty affects performance, uncertainty arises in many forms, which is why some forms of uncertainty are often overlooked in the modeling process. In fact, most uses of optimization tools ignore all forms of uncertainty, typically reflecting the dramatic increase in uncertainty introduced by explicitly modeling any form of uncertainty.

A goal of this chapter is to highlight the different ways that uncertainty can arise. This does not mean that models need to incorporate all forms of uncertainty. However, the decision to ignore a form of uncertainty should be an explicit choice, and not just because a modeler overlooked it.

5.1 The 12 classes of uncertainty

One way to approach the identification of sources of uncertainty is to work backward from a mathematical model. Below are 12 classes of uncertainty

created from the perspective of how uncertainty can enter a model. Complex problems, such as managing a supply chain, an energy system or solving a public health problem, will involve all 12 classes, while simple problems such as playing chess may involve just one.

There is some overlap in the classes, so do not worry if there is some ambiguity in terms of where to list a source of uncertainty. What is important is to identify as many different forms of uncertainty as possible.

1. **Observational errors** – These represent errors in quantities and parameters that we have to observe from the environment. Some examples might be:
    - The current inventory of a product as represented in the computer, which may not match what is actually on hand.
    - Medical X-rays of a patient to detect cancer.
    - The fraction of voters who prefer a particular candidate for political office.
2. **Exogenous uncertainty** – This is information that will arrive to our system after making a decision, such as:
    - The demand for a product being sold in the market.
    - The change in price of a stock.
    - The time required to drive from one city to the next.
    - The amount of cash that may be deposited or withdrawn tomorrow.
    - How a patient responds to a type of medication.
3. **Prognostic uncertainty** – This is errors in forecasts of demands, prices, travel times (any quantity that we might be forecasting).
4. **Inferential uncertainty** – This captures uncertainty in our estimates of the state of the world right now. This could include:
    - How the market might respond to a change in price. We might think there is a 10 percent drop in demand for a 5 percent increase in price, but the true value might be that demand will drop by 12 percent.

- We think that a cancer patient is stage 2, but it might be stage 3. We might detect breast cancer, but overlook that it has spread to other organs.
- A presidential campaign may think that $10 million in ad spending in a major market might produce a 2 percent increase in favorability of a candidate, but the reality may be higher or lower.

5. **Experimental uncertainty** – This describes the variation from running repeated experiments either in a lab, a simulator, or the field:
   - A manufacturer runs experiments of a process for manufacturing silicon wafers. The test may be repeated 10 times, producing a spread of yields between 70 and 90 percent.
   - A business is evaluating a new marketing campaign by running it in five different test markets. There will be variations across the markets, and over time.
   - A computer simulator is used to test the performance of an inventory ordering policy. Each pass of the simulator will produce different results.
6. **Model uncertainty** – This is an umbrella that can cover multiple sources of uncertainty, but one of the most important is uncertainty in the model of how a process evolves over time. Examples might be:
   - How the climate responds to changes in policies for controlling carbon.
   - How a patient responds to injections of insulin.
   - How a disease spreads in a population in response to changes in policies regarding the distribution of vaccines.
7. **Transitional uncertainty** – This is noise in how a system responds to a control. The simplest example would be controlling the path of a rocket or aircraft, which is buffeted by wind. We typically assume that the evolution of the system is known and deterministic, but is affected by an exogenous process (such as wind).
8. **Implementation uncertainty** – There can be a difference between what we decide to do, and the decision that is actually implemented in the field. For example:

- The physician orders a particular medication, but the patient does not take it, or takes the wrong dose.
- A scientist wants to test a particular combination of materials, but the intern orders an incorrect item (errors like this can produce major breakthroughs!).
- The power grid orders that a generator be turned on at 1pm, but the local operator does not turn on the generator until 2pm.

9. **Communication errors** – Instructions to the field can be simply miscommunicated. The person receiving the instruction may think they are doing what is requested, but they just did not hear or understand an instruction.

10. **Algorithmic instability** – There are some settings where running an algorithm repeatedly can return different solutions:

    - Complex problems often require the use of sophisticated algorithms that introduce an element of variability, which often arises when an algorithm uses parallel processing. The speed of parallel processors may affect who finishes first, which can affect the overall path of the algorithm.
    - Algorithms for solving stochastic optimization problems often depend on Monte Carlo sampling which will produce different results each time the algorithm is run (this is seen when running large language models).

11. **Goal uncertainty** – Companies that require groups of people to make decisions (dispatching trucks, trading financial assets, bidding on energy contracts) can exhibit variations because different people emphasize different performance metrics.

12. **Environmental uncertainty** – Here, "environment" might reflect climate, or a political environment (which might impact policies or tariffs), or new management at a company (which results in a change in priorities).

5.2 Examples from selected applications

It helps to see examples of each of the 12 classes for some of the applications we introduced in Chapter 2. For each application, we describe one or more examples of the uncertainties for each class, noting that simpler applications will not have uncertainties for all 12 classes. It is important to remember that the real goal here is to recognize as many sources of uncertainty as possible. How these uncertainties are reflected in the process of making decisions will come in future volumes.

5.2.1 Cash management for a mutual fund

A mutual fund has to determine how much cash to keep on hand to meet redemption requests, and as deposits are made, by both individual and institutional investors.

	Classes of uncertainty	Mutual fund cash balance
1	Observational uncertainty	
2	Exogenous uncertainty	Deposits, redemptions, market indices
3	Prognostic uncertainty	Forecasts of deposits, redemptions, market indices, interest rates
4	Inferential uncertainty	Estimating how redemptions change with market performance
5	Experimental variability	Testing different policies for holding cash
6	Model uncertainty	
7	Transitional uncertainty	Updating how much cash is on hand
8	Implementation uncertainty	
9	Communication errors	
10	Algorithmic instability	
11	Goal uncertainty	Balancing maximizing investment returns, minimizing stock sales for redemptions
12	Environmental uncertainty	Changes in interest rates

5.2.2 Finding the best diabetes treatment

Diabetes patients have to manage their blood sugar using a combination of medications (perhaps using an insulin pump) and diet.

Classes of uncertainty	Managing blood sugar
Observational uncertainty	Measuring A1c levels
Exogenous uncertainty	What a patient eats
Prognostic uncertainty	Anticipating changes in blood sugar levels after a meal
Inferential uncertainty	Estimating how a patient's blood sugar responds to medication
Experimental variability	Changes in blood sugar for different types of medication
Model uncertainty	Modeling how a patient responds to a type of medication
Transitional uncertainty	
Implementation uncertainty	Whether a patient follows their physician's instructions
Communication errors	Whether a patient misunderstands the physician's instructions
Algorithmic instability	
Goal uncertainty	Balancing blood sugar reduction vs. digestion issues
Environmental uncertainty	

5.2.3 Supply chain management

Supply chains require managing inventories that have to be coordinated across the system.

	Classes of uncertainty	Supply chain management
1	Observational uncertainty	Measuring inventory
2	Exogenous uncertainty	Market demand, weather, transit times,
3	Prognostic uncertainty	Forecasting demands, production, resignations
4	Inferential uncertainty	Market response to price, machine failure rates
5	Experimental variability	Simulation errors, testing new materials, test marketing
6	Model uncertainty	How information spreads in the marketplace, how employees respond to incentives
7	Transitional uncertainty	Updating inventories
8	Implementation uncertainty	Failure to follow instructions
9	Communication errors	Incorrect instructions to suppliers
10	Algorithmic instability	Variations in optimal solution from production schedules
11	Goal uncertainty	Differences in priorities toward production cost vs. covering demand
12	Environmental uncertainty	Changes in tariffs, currency exchange rates, interest rates

5.2.4 Allocating naloxone kits

State agencies have to allocate naloxone kits to meet the needs of local clinics and medical professionals who are treating patients.

Classes of uncertainty	Management of naloxone kits
Observational uncertainty	The number of naloxone kits in inventory
Exogenous uncertainty	The number of events requiring uses of naloxone kits
Prognostic uncertainty	Estimates of changes in patterns of drug use
Inferential uncertainty	Estimates of how the availability of kits affects their use
Experimental variability	
Model uncertainty	Understanding how drug use patterns change over time
Transitional uncertainty	Changes in naloxone kit inventories from week to week
Implementation uncertainty	Whether kits are used properly; whether instructions to allocate are followed
Communication errors	Whether field representatives follow instructions in handing out kits
Algorithmic instability	
Goal uncertainty	Prioritizing who to supply with naloxone kits
Environmental uncertainty	Availability of funding for naloxone kits

5.2.5 Managing a fleet of trucks

Truckload trucking companies have to determine which loads to move, with what driver.

Classes of uncertainty	Managing a fleet of trucks
Observational uncertainty	
Exogenous uncertainty	New loads from shippers; refused assignments by drivers; traffic delays
Prognostic uncertainty	Forecasts of loads in the future
Inferential uncertainty	How the market will respond to changes in spot prices
Experimental variability	Running simulations of changes in driver allocations
Model uncertainty	
Transitional uncertainty	Changes in number of available loads; updates to driver availability
Implementation uncertainty	Whether a dispatcher follows the instruction of the model
Communication errors	Whether dispatchers follow the instructions of their managers
Algorithmic instability	Changes in the solution from updates of estimates of driver values
Goal uncertainty	Balancing empty miles against shipper commitments against getting drivers home
Environmental uncertainty	Changes in hours-of-service rules by the Dept. of Transportation

5.2.6 Planning an electric power grid

The power grid has to work with utilities to determine which generators should be turned on to meet the anticipated demands placed on the grid.

Classes of uncertainty	Managing the electric power grid
Observational uncertainty	Estimating temperature, weather, customer attitudes
Exogenous uncertainty	Changes in weather, generator failures
Prognostic uncertainty	Forecasts in temperature, wind, cloud cover
Inferential uncertainty	Estimating how power demand changes as grid prices change
Experimental variability	Variability in the response to changes in model parameters
Model uncertainty	Errors in evolution of wind speeds over geographical region
Transitional uncertainty	Difference between expected wind power and actual
Implementation uncertainty	Differences between instructions to utilities and what they do
Communication errors	Errors in understanding of instructions communicated to utilities
Algorithmic instability	Variations in the performance of the integer programming algorithm
Goal uncertainty	Balancing the use of nuclear vs. coal vs. renewables
Environmental uncertainty	Changes in policies for reimbursement of excess solar generation

5.3 How uncertainty affects performance

While we have identified 12 classes of uncertainty, there are only three ways that uncertainty affects the behavior of a model:

1. How decisions are made.

2. The performance metrics from the decisions chosen in the model.

3. The evolution of the system in the model after a decision is made, and before the next decisions have to be made.

Then there are the ways that uncertainty affects performance in the field:

4. The decisions that are implemented in the field.

5. The actual performance metrics for the decisions that are implemented in the field.

6. The evolution of the system in the field.

There are many ways uncertainty affects performance, from random costs to how a patient responds to a drug to the price of an investment. For now, we are just going to focus on identifying how uncertainty affects performance.

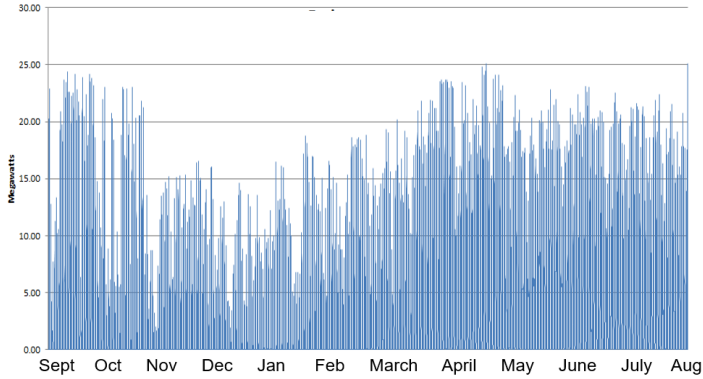


Figure 5.1: Hourly energy output from solar over an entire year, demonstrating both with-day and seasonal variability.

5.4 Different forms of uncertainty

The first step in understanding uncertainty requires listing the different sources of uncertainty, as we have done above. The next step, then, is describing the different forms that the uncertainty arises. Below is a sampling of these:

- Fine-grained variability – This might arise at time scales of seconds (even fractions of a second), minutes, hours, or daily. Examples of fine-grained variability are:
  - High-frequency trading in finance - These decisions are made several times per second.
  - Frequency regulation for the power grid - These are signals sent every two seconds to generators to make adjustments so that the power voltage stays within a narrow range.
  - Hourly sales of different restaurant food choices which may require some preparation in advance of service.
  - The hourly variations in wind speeds, shown in figure 5.1. This figure would also capture hourly to daily variations in cloud cover, all in the context of predictable seasonal variations.
  - Daily sales of a retail product.
  - Daily to weekly variations in hospital admissions with the flu.

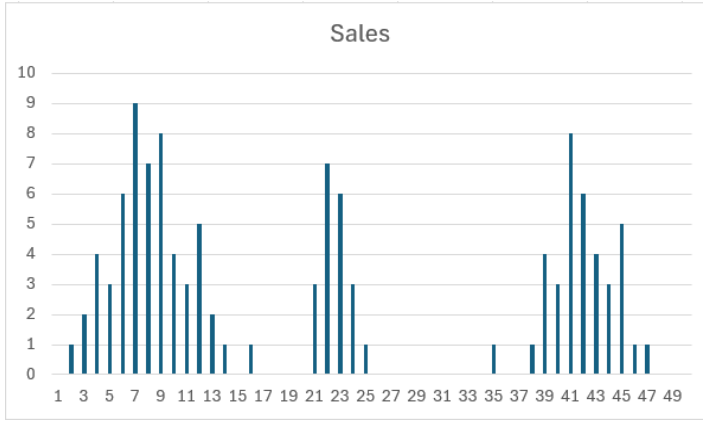


Figure 5.2: Illustration of bursts of activity.

- Shifts – The fine-grained variability of a process typically represents variations around a mean, but there are times when the mean will shift from time to time. Examples are:
  - The random demands for a retail product may shift as a result of a change in pricing for either the product, or a competitive product.
  - The rate of hospital admissions for an infectious disease will shift as the disease moves through a population near the hospital.
  - The demands for redemptions from a mutual fund, which vary by the minute, will shift when the broader stock market responds to a changing economy.
  - The number of people making bids on houses (say, for a given realtor), will shift to different levels as interest rates change.

- Bursts, intermittent demands – These describe patterns where there is little or no activity, but then undergoes a burst until it dies down again (see figure 5.2). Examples of bursts include:
  - Spread of diseases such as measles – When a disease enters a region, there will be a period of increased infections as the disease moves through the most vulnerable part of the population.

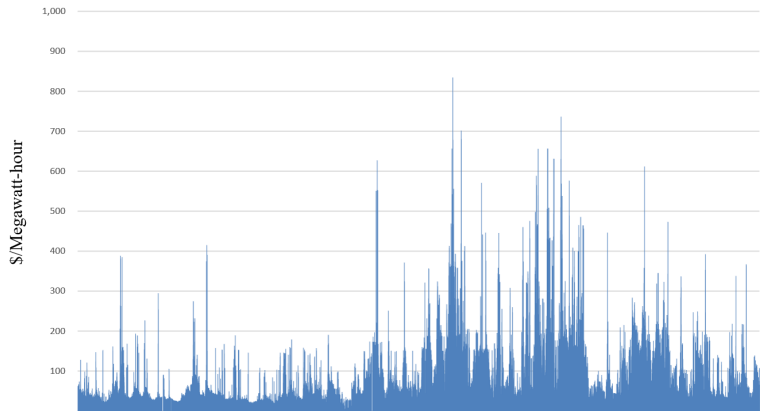


Figure 5.3: Real-time electricity prices, updated every five minutes, in February, illustrating extreme volatility.

  - A product may not be selling, until someone happens to buy it and then spreads the word when they have a good experience. This will spread through their network until it is saturated.

- Spikes – A process may reflect two driving sources. One produces modest outcomes from a well-defined distribution. The second represents infrequent outcomes that are much larger than the first distribution. For example:

  - The price of electricity on the grid is updated every 5 minutes. Figure 5.3 shows real-time grid prices for the month of February. It shows a steady sequence of random changes, with occasional spikes that are much larger than the typical variations.
  - A storm creates a rush of purchases of milk, eggs and toilet paper.
  - A storm system moving past an airport can result in a number of flight cancellations, which in turn can create a large number of last-minute requests for hotel rooms.

- Spatial events (weather, diseases, regulatory) – There are numerous examples of random processes that are regional in nature. Some examples are:

  - Weather – Storms can create a range of random events in a region that has been struck by bad weather, or where bad weather is forecast.
  - Diseases – Since disease propagation often requires physical contact, outbreaks typically follow a regional pattern.
  - Regulations – Changes in regulations typically follow political boundaries, which could be for a country, or a state, district or province within a country.

- Systemic events – These are events that can affect an entire company (spanning international boundaries), an entire country, or even have a global impact, such as:
  - Cyberattacks, which can impact information flows for an entire company.
  - Public perception – Public events can produce rapid positive or negative perceptions of a company. For example, a beer company undertook a campaign to promote the LGBTQ community, which produced a sudden backlash by their conservative customers which impact sales across the company.

- Rare events – Rare events can arise from a number of sources such as earthquakes, disease outbreaks, or terrorist attacks. These tend to be recognized events that occur quite rarely, but which can have a major impact on an organization when they do happen.

- Contingencies – This category refers to events that might happen, but for which there is no history. For example, grid operators will plan for a failure of nuclear power plants. While this may not have ever happened within a country, the grid operator may still want to prepare for the event if it does happen.

5.5 Seasonality

A different form of variability is captured under the general term "seasonality" which comes in various forms:

- Daily cycles – Also known as diurnal cycles, these are all ultimately traced to solar cycles, but these can induce strong daily patterns in

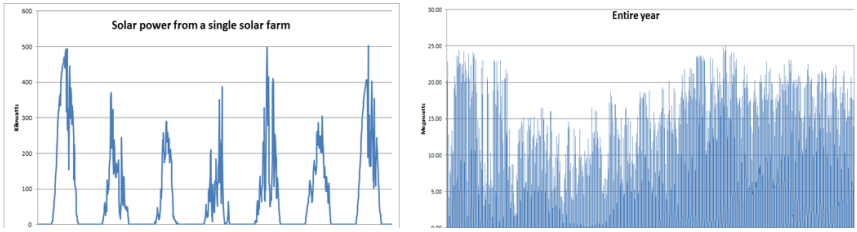


Figure 5.4: Daily solar energy over a week (left), and annual solar energy (right).

human activities. Daily cycles are typically discretized into hours, but finer discretizations (5 minutes, 1 minute) can arise.

- Day of week – This reflects the daily patterns in human behavior set around the different days of the week.
- Hour of week – Hourly patterns may easily depend on the day of week as well as hour of day to capture effects such as Monday morning, Friday afternoon, and daily patterns on weekdays versus weekends.
- Week of month – Manufacturing often has a push to maximize production by month, creating an incentive to push product out before the end of the month. This creates a surge toward the end of the month, followed by a lull.
- Month of year – This captures the familiar seasonal patterns of winter, spring, summer and fall.
- Week of year – Seasonal changes can occur within a month, encouraging the use of week-of-year as a seasonal time increment.

Figure 5.4(left) shows solar energy production over the course of a week, illustrating both the familiar and highly predictable pattern created by the sun, which is interfered by the highly stochastic presence of cloud cover. Figure 5.4(right) shows hourly solar energy over the entire year, where we can clearly see the reduction in solar energy during the winter season.

5.6 Creating beliefs

If we are modeling uncertainty on the computer, we have to find a way to represent it. Below are several popular strategies.

Price	Cum prob
$ 44.40	0.1
$ 51.00	0.2
$ 55.00	0.3
$ 66.20	0.4
$ 76.00	0.5
$ 86.33	0.6
$ 92.36	0.7
$ 98.00	0.8
$100.00	0.9
$112.00	1

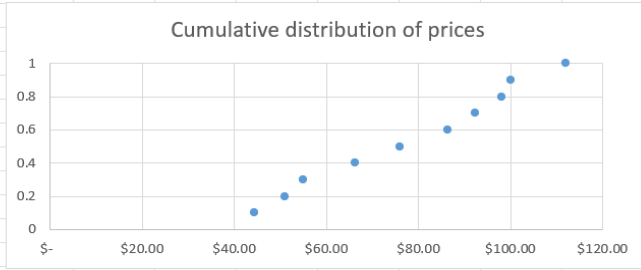


Figure 5.5: Computing an empirical CDF from a set of observations.

- Historical data may be used to fit a known probability distribution – There is an entire family of probability distributions we can use to fit to historical data, the best known being the normal distribution. We return to this rich topic later.

- Use historical data to create a sampled belief model – Imagine we have travel times ranging from 50 to 80 minutes for a trip, depending on the traffic. We can use any of several probability distributions to represent this uncertainty, or we may simply use a sample of past observations, such as:

  (52, 63, 78, 59, 71, 68)

- Use historical data to create a quantile distribution from which samples can be drawn – Assume we have a sample of 10 observations of electricity prices, given in figure 5.5 (left). After sorting the prices from smallest to largest, we then show the cumulative probability given in figure 5.5 (right). So, we would say that 60 percent of the observations are $86.33 or lower. These are then plotted in the cumulative distribution on the right.

  It is also possible to manually create a cumulative distribution using judgment.

- Use manually created outcomes to represent events that might happen – When do not have data, we can simply make up possible outcomes. For example, we may be shipping product from Taiwan, which normally takes four weeks. However, we can envision various forms of delays, from hurricanes to backups at the Suez Canal, labor problems at ports or even terrorist attacks. We might feel that we have

to allow for the possibility that the shipment might take as long as nine weeks, and then plan for this contingency.

5.7 The problem of correlations

The previous section is a brief snapshot of ways of representing the uncertainty in an estimate (see chapters 3 and 10 in (Powell 2022)). However, once we go down the road of recognizing uncertainty, we have to face the far more complex issue of correlations.

It helps to have some examples of information processes in mind to illustrate different forms of correlation. Assume we might be considering any of the following streams of data:

1) Customers purchasing a retail product across many sales locations.

2) The lead time from placing an order and receiving it.

3) The energy generated from a wind farm

4) The rate of new infections from the latest strain of flu.

5) The number of truckload movements tendered by a customer to different locations.

These are just a small handful of the types of information streams we will have to deal with. Below we use these examples to talk about three different types of correlations:

- Correlations over time.
- Correlations over geography.
- Correlations over attributes.

5.7.1 Correlations over time

All sequential decision problems involve the element of time, which may be at virtually any time scale, from seconds, minutes, hours, and days to weeks, months and even years.

Correlation over time can arise in each of our five problem settings as follows:

1) An incoming snow storm can create a surge in demand for snowblowers; negative publicity can create a period of reduced demand.

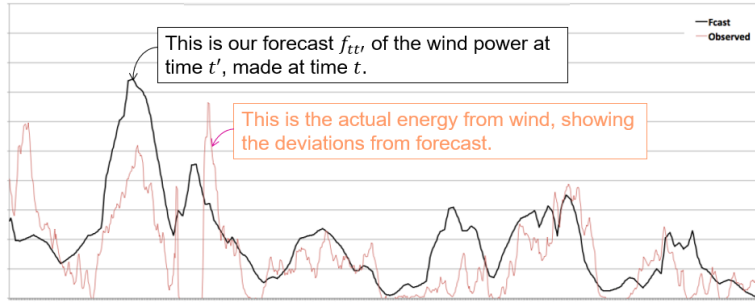


Figure 5.6: Actual vs. forecast, showing crossing times.

2) A port strike can create backlogs that increase unloading times for months.

3) Rain storms can create periods of increased wind generation that may last for days.

4) As a virus enters a region, it will create a period of elevated infections that can last from weeks to months.

5) If a plant is shut down for maintenance, there may be a drop in loads out of a location for a week.

Figure 5.6 illustrates how the energy generated from wind may exceed, or fall below, the forecast over a period of time as weather systems move through a region. It is important that we replicate not just the error between actual and predicted, but also the amount of time we stay above or below the forecast, a quantity known as the "crossing time."

It is fairly common for random signals to be viewed as variations from a base mean, which is usually treated as a constant which has to be estimated. In reality, the "base mean" may also be varying, but on a different time scale. For example, customers walking into a retail store represent random outcomes on a fine time scale, since the behavior of each customer is independent. But they may be responding to market signals (advertising, word-of-mouth) that is also changing, but more slowly.

Arguably the biggest challenge with correlation over time is that it can occur at multiple time scales, at the same time. Independent events (such as how many people come into a store each hour requesting cough medicine) are quite easy to model. The variations that happen on longer time scales

are harder because they create what appear to be correlations across time at smaller time scales.

5.7.2 Correlations across geography

Customer purchase decisions, disease outbreaks, and weather are all examples of random processes that vary geographically. Sometimes political boundaries may limit the correlations, but most of the time it is simply distance that governs the strength of the correlation.

Spatially distributed processes typically occur in very high dimensions (there are a lot of spatial locations!). What simplifies geographical correlations is that it is typically fairly easy to capture. Geography may be a pure function of distance, but it can also reflect geographical boundaries as well as population movement patterns. Fortunately, there are powerful mathematical tools help identify and capture these correlations.

Correlation across geography can arise in each of our five problem settings as follows:

1) The surge in demand for snowblowers will also be regional since it is responding to snowstorms (which are regional).

2) A port delay can produce reduced supplies in the region served by the port, with higher correlations for points closer to the port.

3) Rainstorms are also regional, and will create surges in energy from wind farms in the areas affected by the storm. Similarly, hot spells (which are also regional) will produce periods of low wind.

4) The spread of flu will be regional since it passes between people who are close to each other.

5) Freight is generated either by changes in a manufacturing plant (which is located at one point) or changes in demand, which may be driven by regional forces.

5.7.3 Correlations across attributes

Most of our examples involve activities that are characterized by a set of attributes:

1) Demand for clothing will have correlations between garments with similar style but different colors.

2) Products that share common inputs (such as materials for clothing, chips for cars, rare earths for motors) may exhibit similar lead time delays when there is a shortage of the input.

3) (No apparent use of correlation across attributes for wind energy.)

4) New infections may be correlated across people who share features such as age or medical conditions.

5) The flow of truckload movements can be correlated when they are moving common commodities or products.

It is often the case that when we expand all the attributes, we find ourselves with so many combinations that the number of observations for a particular combination of attributes may be quite small, and possibly zero. These problems lend themselves to the use of hierarchical estimation methods, where we create different time series by neglecting one or more attributes, and then using weighted combinations.

5.8 Exercises

Review questions

5.1 — Name the 12 classes of uncertainty, giving one example of each from any application.

5.2 — Name seven forms of uncertainty that can describe random processes, and describe a context that might produce each one.

5.3 — What are the ways that uncertainty can impact the performance of a system, and give an example of each.

5.4 — Name four forms of seasonality.

5.5 — Create a cumulative distribution of wind speeds from the following observations:

(17, 8, 2, 12, 9, 28, 10, 8, 35, 12, 15)

Modeling questions

For each of the questions below, try to find as many forms of uncertainty within each of the classes for the following settings, following the tables given in section 5.2.

5.6 — The inventory planning problem in section 2.3.

5.7 — The furniture demand management problem in section 2.4.

5.8 — Planning clinical trials in section 2.7.3

5.9 — Running a presidential election in section 2.8.

5.10 — Supply chain finance in section 2.11.

5.11 — Choose a problem setting of your own, ideally one with some complexity, and identify as many types of uncertainty using the 12 classes as a guide.

Chapter 6

Closing notes

This volume has focused on answering three questions to help with framing pretty much any problem involving a decision. The questions are:

- What are the performance metrics? These should be presented using the following guidelines:
  - They should be organized into pyramids as an informal way of capturing relative performance.
  - They should be identified as objectives (to be minimized or maximized), targets, or limits.
  - Finally, they should be separated between base metrics and risk metrics.
- What types of decisions are being made (and who makes them)? These should be identified by:
  - Are the decisions acting on physical resources, financial resources, or information?
  - Are the decisions discrete or continuous, scalar or vectors?
  - At what frequency are decisions made, and when are they implemented?
- What are the sources of uncertainty that affect performance? These should be characterized on the basis of:
  - Which of the 12 classes of uncertainty are relevant to the problem?

- How does each source of uncertainty impact how decisions are made, and how do they impact the performance metrics?
- How does each source of uncertainty behave? This should capture how they behave over time, and any correlations that need to be represented.

Of course these questions sound interesting and relevant, but this is all we cover in Volume I. We gave a brief tour of the four classes of policies for making decisions, but we will not return to this topic until Volume III. Before we can address making decisions, we have to fill in other details such as identifying the information needed to make a decision, and how the system evolves over time. This is covered in Volume II, which also sets the foundation for evaluating the policies that we will present in Volume III.

6.1 Decisions, decisions

We make so many decisions that we are often working our way through problems without even recognizing that we have choices. Arguably the first decision we have to make is what sort of analysis to do in order to make a decision. Four important categories of decision settings include:

1) Decisions where there is potential for simply doing a better job:
   - A trucking company has to decide which loads to book to maximize revenue while meeting the needs of their drivers. This may involve deciding which shippers they should offer bids for to try to win their freight, which also requires specifying what prices to charge, and whether to make changes in their fleet.
   - A hedge fund wants to automate what had been a manual process for picking investments on a day-to-day basis. Anticipated benefits are better performance with fewer, reducing administrative costs.
   - A manufacturer wants to do a better job managing inventories for their supply chain.
2) High volume decisions which require automation:
   - A major retailer has to manage inventories for 50,000 items, requiring daily inventory reviews and replenishment decisions.

- A hotel has to update pricing on over 10,000 different room/service offerings on their website.
- The power grid has to plan the schedules for hundreds of power generators on a rolling basis.
- A mutual fund manager has to decide which of 10,000 stocks to invest in.
- An online bank might have to evaluate thousands of loan applications each day.

These are decisions that are made in which high volume, where manual decision making is cumbersome, and may require training a large number of people.

3) High value, high risk decisions - These are decisions that require analysis because the choices are high value, with high uncertainty, which means there is considerable risk:

   - Should a company purchase another company? There is a lot of money involved, and uncertainty in the performance of the new markets they are acquiring, and how well corporate cultures will blend.
   - Should a drug company push a drug into clinical trials? Total costs can be over 100 million dollars, and there is on average only a 10 percent probability the drug will ultimately be successful.
   - A patient is suffering from a serious disease, but the only treatment puts the patient's life at risk.

   These are the kinds of problems that are typically the subject of careful analyses using decision trees, sometimes with the help of external consultants.

4) Decisions made without any analysis - These are often decisions affecting complex activities where formal analysis is not likely to be of value, and people have strong intuition about what choices to make:

   - A startup needs to increase sales. After a meeting of the executive team, they decide to increase the marketing budget, add two sales people, and include a promotional package to allow people to try the software at very low cost.

- A public health expert is trying to address a surge of drug overdoses. She decides to undertake an information campaign, provides additional funding to harm reduction groups, and talks to local police and health officials.
- As a clothing manufacturer in the U.S., you source most of your fabric from Bangladesh, which is being threatened with a dramatic increase in tariffs. If they go through, you will not be able to operate profitably. What do you do?
- The campaign manager for a presidential campaign has to decide where to schedule the speeches for a candidate over the next two weeks.

In each of these, the decision-maker is moving forward on gut instinct, without even making a list of the alternative choices that may be required. While an argument can be made that the decision is drawing on past experience, there is typically uncertainty and some thought should go into thinking about strategies given different outcomes.

We would make the argument that all decisions benefit from simply understanding the metrics for evaluating performance (including risk), what types of decisions can be made, and the uncertainties that may affect performance. Whether these are then subjected to more formal analysis will be a judgment call by the decision maker, which is the first decision that has to be made for a project.

The goal of this volume is to avoid falling in the trap of framing a problem based on the familiarity of the person (or team) doing the framing with specific tools, whether they are decision trees or large integer programs. Framing needs to be completely independent of any toolbox.

6.2 Next steps

Framing a problem in terms of metrics, decisions, and uncertainties is a critical first step, one that may contribute additional clarity to help understand a problem, even if there is no subsequent use of quantitative analysis. However, there will be problems that either call for more careful analysis, or there is a clear need for automation (such as the examples above).

When there is an interest in moving to the computer for making decisions, we have to anticipate the following steps:

Step 1) Identifying the metrics, decisions, and uncertainties that we want to include in our model to address the ultimate goal(s) of the project. At this point a choice has to be made: use the improved understanding to make a decision, or move forward with further analysis.

Step 2) Mathematical modeling the chosen problem using the universal modeling framework, including the modeling of uncertainty. This is covered in Volume II.

Step 3) Designing the policies to determine the decisions identified in Step 2, and tuning them using the model developed in Step 3. This step will help identify the information that is needed. This is covered in Volume III.

Step 4) Designing the processes for collecting the information needed to make decisions (compute the policy) and evaluate performance.

Step 5) Implementing decisions in the field. This requires communicating instructions and designing the processes to implement decisions. This is where we observe and manage compliance.

Step 6) Evaluating performance of the process.

It is possible to simulate all these steps in the computer which can help serve as a test environment. Simulators (sometimes known as "digital twins") can be helpful for evaluating and comparing policies, but they can be difficult to build and validate. As a field implementation there are significant steps for creating data collection processes, as well as systems for implementing and compliance monitoring.

References

Brown, A. (2010), Beat the Curve: Designing Adaptive Blood Glucose Management Strategies for Non-Pump Patients with Type 1 Diabetes, Senior thesis, Princeton University.

Hsih, K. (2010), Optimal Dosing Applied to Glycemic Control for Type 2 Diabetes, Senior thesis, Princeton University.

Kahneman, D. & Tversky, A. (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica* **47(2)**, 363–391.

Pi, S. (2019), An optimal learning model for state-level optimization of naloxone kits with non-convex response rates, Senior thesis, Princeton University.

Powell, W. B. (2022), *Reinforcement Learning and Stochastic Optimization: A unified framework for sequential decisions*, John Wiley and Sons.

Thaler, R. H. (1985), 'Mental Accounting and Consumer Choice', *Marketing Science* **4(3)**, 199–214.

Weiner, B. (1986), *An attributional theory of motivation and emotion*, Harper and Row.