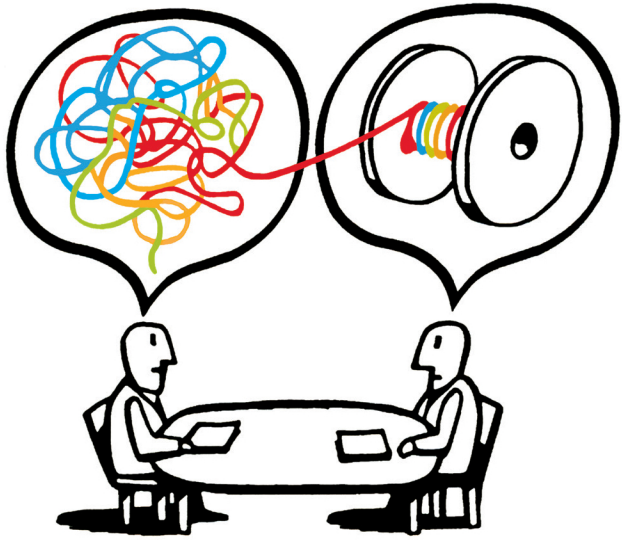
O'REILLY®

Thinking with Data

How to Turn Information into Insights



Max Shron

O'REILLY®

"*Thinking with Data* does a wonderful job of reminding data scientists to look past technical issues and to focus on making an impact on the broad business objectives of their employers and clients. It's a useful supplement to a data science curriculum that is largely focused on the technical machinery of statistics and computer science."

—John Myles White, Scientist at Facebook, author of *Machine Learning for Hackers* and *Bandit Algorithms for Website Optimization*

"This is a great piece of work. It will be required reading for my team."

—Nick Kolegraff, Director of Data Science, Rackspace

"*Thinking with Data* gets to the essence of the process, and guides data scientists in answering that most important question—what's the problem we're really trying to solve?"

—Hilary Mason, Data Scientist in Residence, Accel Partners, and cofounder of the DataGotham Conference

Thinking with Data

Many analysts are too concerned with tools and techniques for cleansing, modeling, and visualizing datasets and not concerned enough with asking the right questions. In this practical guide, data strategy consultant Max Shron shows you how to put the *why* before the *how*, through an often-overlooked set of analytical skills.

Thinking with Data helps you learn techniques for turning data into knowledge you can use. You'll learn a framework for defining your project, including the data you want to collect, and how you intend to approach, organize, and analyze the results. You'll also learn patterns of reasoning that will help you unveil the real problem that needs to be solved.

Max Shron runs a data strategy consultancy in New York, working with many organizations to help them get the most out of their data. His analyses of transit, public health, and housing markets have been featured in *The New York Times*, *Chicago Tribune*, *Huffington Post*, and other media outlets.

US $19.99 CAN $20.99
ISBN: 978-1-449-36293-5

Twitter: @oreillymedia
facebook.com/oreilly
oreilly.com

Data Science

Praise for *Thinking with Data*

"*Thinking with Data* gets to the essence of the process, and guides data scientists in answering that most important question—what's the problem we're really trying to solve?"

— *Hilary Mason*
Data Scientist in Residence at Accel Partners; co-founder of the DataGotham Conference

"*Thinking with Data* does a wonderful job of reminding data scientists to look past technical issues and to focus on making an impact on the broad business objectives of their employers and clients. It's a useful supplement to a data science curriculum that is largely focused on the technical machinery of statistics and computer science."

— *John Myles White*
Scientist at Facebook; author of *Machine Learning for Hackers* and *Bandit Algorithms for Website Optimization*

"This is a great piece of work. It will be required reading for my team."

— *Nick Kolegraff*
Director of Data Science at Rackspace

"Shron's *Thinking with Data* is a nice mix of academic traditions, from design to philosophy, that rescues data from mathematics and the regime of pure calculation. … These are lessons that should be included in any data science course!"

— *Mark Hansen*
Director of David and Helen Gurley Brown Institute for Media Innovation; Graduate School of Journalism at Columbia University

Thinking with Data

Max Shron

O'REILLY® Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

THINKING WITH DATA

by Max Shron

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Mike Loukides and Ann Spencer
Production Editor: Kristen Brown
Copyeditor: O'Reilly Production Services
Proofreader: Kim Cofer

Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

February 2014: First Edition

Revision History for the First Edition:

2014-01-16: First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449362935* for release details.

ISBN: 978-1-449-36293-5
[LSI]

Contents

Preface

Working with data is about producing knowledge. Whether that knowledge is consumed by a person or acted on by a machine, our goal as professionals working with data is to use observations to learn about how the world works. We want to turn information into insights, and asking the right questions ensures that we're creating insights about the right things. The purpose of this book is to help us understand that these are our goals and that we are not alone in this pursuit.

I work as a data strategy consultant. I help people figure out what problems they are trying to solve, how to solve them, and what to do with them once the problems are "solved." This book grew out of the recognition that the problem of asking good questions and knowing how to put the answers together is not a new one. This problem—the problem of turning observations into knowledge—is one that has been worked on again and again and again by experts in a variety of disciplines. We have much to learn from them.

People use data to make knowledge to accomplish a wide variety of things. There is no one goal of all data work, just as there is no one job description that encapsulates it. Consider this incomplete list of things that can be made better with data:

- Answering a factual question
- Telling a story
- Exploring a relationship
- Discovering a pattern
- Making a case for a decision
- Automating a process
- Judging an experiment

Doing each of these well in a data-driven way draws on different strengths and skills. The most obvious are what you might call the "hard skills" of working with data: data cleaning, mathematical modeling, visualization, model or graph interpretation, and so on.[1]

What is missing from most conversations is how important the "soft skills" are for making data useful. Determining what problem one is actually trying to solve, organizing results into something useful, translating vague problems or questions into precisely answerable ones, trying to figure out what may have been left out of an analysis, combining multiple lines or arguments into one useful result…the list could go on. These are the skills that separate the data scientist who can take direction from the data scientist who can give it, as much as knowledge of the latest tools or newest algorithms.

Some of this is clearly experience—experience working within an organization, experience solving problems, experience presenting the results. But these are also skills that have been taught before, by many other disciplines. We are not alone in needing them. Just as data scientists did not invent statistics or computer science, we do not need to invent techniques for how to ask good questions or organize complex results. We can draw inspiration from other fields and adapt them to the problems we face. The fields of design, argument studies, critical thinking, national intelligence, problem-solving heuristics, education theory, program evaluation, various parts of the humanities—each of them have insights that data science can learn from.

Data science is already a field of bricolage. Swaths of engineering, statistics, machine learning, and graphic communication are already fundamental parts of the data science canon. They are necessary, but they are not sufficient. If we look further afield and incorporate ideas from the "softer" intellectual disciplines, we can make data science successful and help it be more than just this decade's fad.

A focus on *why* rather than *how* already pervades the work of the best data professionals. The broader principles outlined here may not be new to them, though the specifics likely will be.

1. See *Taxonomy of Data Science* by Hilary Mason and Chris Wiggins (*http://www.dataists.com/2010/09/a-taxonomy-of-data-science/*) and *From Data Mining to Knowledge Discovery in Databases* by Usama Fayyad et al. (AI Magazine, Fall 1996).

This book consists of six chapters. Chapter 1 covers a framework for scoping data projects. Chapter 2 discusses how to pin down the details of an idea, receive feedback, and begin prototyping. Chapter 3 covers the tools of arguments, making it easier to ask good questions, build projects in stages, and communicate results. Chapter 4 covers data-specific patterns of reasoning, to make it easier to figure out what to focus on and how to build out more useful arguments. Chapter 5 takes a big family of argument patterns (causal reasoning) and gives it a longer treatment. Chapter 6 provides some more long examples, tying together the material in the previous chapters. Finally, there is a list of further reading in Appendix A, to give you places to go from here.

Conventions Used in This Book

The following typographical convention is used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/thinking-with-data*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

I would be remiss to not mention some of the fantastic people who have helped make this book possible. Juan-Pablo Velez has been invaluable in refining my ideas. Jon Bruner, Matt Wallaert, Mike Dewar, Brian Eoff, Jake Porway, Sam Rayachoti, Willow Brugh, Chris Wiggins, Claudia Perlich, and John Matthews provided me with key insights that hopefully I have incorporated well.

Jay Garlapati, Shauna Gordon-McKeon, Michael Stone, Brian Eoff, Dave Goodsmith, and David Flatow provided me with very helpful feedback on drafts. Ann Spencer was a fantastic editor. It was wonderful to know that there was always someone in my corner. Thank you also to Solomon Roberts, Gabe Gaster, emily barger, Miklos Abert, Laci Babai, and Gordon Kindlmann, who were each crucial at setting me on the path that gave me math. Thank you also to Christian Rudder, who taught me so much—not least of which, the value of instinct. As always, all the errors and mistakes are mine alone. Thanks as well to all of you who were helpful whose names I neglected to put down.

At last I understand why every author in every book on my shelf thanks their family. My wonderful partner, Sarah, has been patient, kind, and helpful at every stage of this process, and my loving parents and sister have been a source of comfort

and strength as I made this book a reality. My father especially has been a great source of ideas to me. He set me off on this path as a kid when he patiently explained to me the idea of "metacognition," or thinking about thinking. It would be hard to be grateful enough.

Scoping: Why Before How

Most people start working with data from exactly the wrong end. They begin with a data set, then apply their favorite tools and techniques to it. The result is narrow questions and shallow arguments. Starting with data, without first doing a lot of thinking, without having any structure, is a short road to simple questions and unsurprising results. We don't want unsurprising—we want knowledge.

As professionals working with data, our domain of expertise has to be the *full problem*, not merely the columns to combine, transformations to apply, and models to fit. Picking the right techniques has to be secondary to asking the right questions. We have to be proficient in both to make a difference.

To walk the path of creating things of lasting value, we have to understand elements as diverse as the needs of the people we're working with, the shape that the work will take, the structure of the arguments we make, and the process of what happens after we "finish." To make that possible, we need to give ourselves space to think. When we have space to think, we can attend to the problem of *why* and *so what* before we get tripped up in *how*. Otherwise, we are likely to spend our time doing the wrong things.

This can be surprisingly challenging. The secret is to have structure that you can think through, rather than working in a vacuum. Structure keeps us from doing the first things to cross our minds. Structure gives us room to think through all the aspects of a problem.

People have been creating structures to make thinking about problems easier for thousands of years. We don't need to invent these things from scratch. We can adapt ideas from other disciplines as diverse as philosophy, design, English composition, and the social sciences to make professional data work as valuable as possible. Other parts of the tree of knowledge have much to teach us.

Let us start at the beginning. Our first place to find structure is in creating the scope for a data problem. A scope is the outline of a story about why we are working on a problem (and about how we expect that story to end).

In professional settings, the work we do is part of a larger goal, and so there are other people who will be affected by the project or are working on it directly as part of a team. A good scope both gives us a firm grasp on the outlines of the problem we are facing and a way to communicate with the other people involved.

A task worth scoping could be slated to take anywhere from a few hours with one person to months or years with a large team. Even the briefest of projects benefit from some time spent thinking up front.

There are four parts to a project scope. The four parts are the *context* of the project; the *needs* that the project is trying to meet; the *vision* of what success might look like; and finally what the *outcome* will be, in terms of how the organization will adopt the results and how its effects will be measured down the line. When a problem is well-scoped, we will be able to easily converse about or write out our thoughts on each. Those thoughts will mature as we progress in a project, but they have to start somewhere. Any scope will evolve over time; no battle plan survives contact with opposing forces.

A mnemonic for these four areas is CoNVO: *co*ntext, *n*eed, *v*ision, *o*utcome. We should be able to hold a conversation with an intelligent stranger about the project, and afterward he should understand (at a high level), why and how we accomplished what we accomplished. Hence, CoNVO.

All stories have a structure, and a project scope is no different. Like any story, our scope will have exposition (the context), some conflict (the need), a resolution (the vision), and hopefully a happily-ever-after (the outcome). Practicing telling stories is excellent practice for scoping data problems.

We will examine each part of the scoping process in detail before looking at a fully worked-out example. In subsequent chapters, we will explore other aspects of getting a good data project going, and then we will look carefully at the structures for thinking that make asking good questions much easier.

Writing down and refining our CoNVO is crucial to getting it straight. Clear writing is a sign of clear thinking. After we have done the thinking that we need to do, it is worthwhile to concisely write down each of these parts for a new problem. At least say them out loud to someone else. Having to clarify our thoughts down to a few sentences per part is extremely helpful. Once we have them clear (or at least know what is still unclear), we can go out and acquire data, clarify our understanding, start the technical work, clarify our understanding, gradually converge on

something smart and useful, and...clarify our understanding. Data science is an iterative process.

Context (Co)

Every project has a context, the defining frame that is apart from the particular problems we are interested in solving. Who are the people with an interest in the results of this project? What are they generally trying to achieve? What work, generally, is the project going to be furthering?

Here are some examples of contexts, very loosely based on real organizations, distilled down into a few sentences:

- This nonprofit organization reunites families that have been separated by conflict. It collects information from refugees in host countries. It visits refugee camps and works with informal networks in host countries further from conflicts. It has built a tool for helping refugees find each other. The decision makers on the project are the CEO and CTO.
- This department in a large company handles marketing for a shoe manufacturer with a large online presence. The department's goal is to convince new customers to try its shoes and to convince existing customers to return again. The final decision maker is the VP of Marketing.
- This news organization produces stories and editorials for a wide audience. It makes money through advertising and through premium subscriptions to its content. The main decision maker for this project is the head of online business.
- This advocacy organization specializes in ferreting out and publicizing corruption in politics. It is a small operation, with several staff members who serve multiple roles. They are working with a software development team to improve their technology for tracking evidence of corrupt politicians.

Contexts emerge from understanding who we are working with and why they are doing what they are doing. We learn the context from talking to people, and continuing to talk to them until we understand what their long-term goals are. The context sets the overall tone for the project, and guides the choices we make about what to pursue. It provides the background that makes the rest of the decisions make sense. The work we do should further the mission espoused in the context. At least if it does not, we should be aware of that.

New contexts emerge with new partners, employers, or supervisors, or as an organization's mission shifts over time. A freelancer often has to understand a new context with every project. It is important to be able to clearly articulate the long-term goals of the people we are looking to aid, even when embedded within an organization.

Sometimes the context for a project is simply our own curiosity and hunger for understanding. In moderation (or as art), there's no problem with that. Yet if we treat every situation only as a chance to satisfy our own interests, we will soon find that we have passed up opportunities to provide value to others.

The context provides a project with larger goals and helps to keep us on track. Contexts include larger relevant details, like deadlines, that will help us to prioritize our work.

Needs (N)

Everyone faces challenges. Things that, were they to be fixed or understood, would advance the goals they want to reach. What are the specific needs that could be fixed by intelligently using data? These needs should be presented in terms that are meaningful to the organization. If our method will be to build a model, the need is not to build a model. The need is to solve the problem that having the model will solve.

Correctly identifying needs is tough. The opening stages of a data project are a design process; we can draw on techniques developed by designers to make it easier. Like a graphic designer or architect, a data professional is often presented with a vague brief to generate a certain spreadsheet or build a tool to accomplish some task. Something has been discussed, perhaps a definite problem has even been articulated—but even if we are handed a definite problem, we are remiss to believe that our work in defining it ends there. Like all design processes, we need to keep an open mind. The needs we identify at the outset and the needs we ultimately try to meet are often not the same.

If working with data begins as a design process, what are we designing? We are designing the steps to create knowledge. A need that can be met with data is fundamentally about knowledge, fundamentally about understanding some part of how the world works. Data fills a hole that can only be filled with better intelligence. When we correctly explain a need, we are clearly laying out what it is that could be improved by better knowledge. What will this spreadsheet teach us? What will the tool let us know? What will we be able to do after making this graph that we could not do before?

> When we correctly explain a need, we are clearly laying out what it is that could be improved by better knowledge.

Data science is the application of math and computers to solve problems that stem from a lack of knowledge, constrained by the small number of people with any interest in the answers. In the sciences writ large, questions of what matters within the field are set in conferences, by long social processes, and through slow maturation. In a professional setting, we have no such help. We have to determine for ourselves which questions are the important ones to answer.

It is instructive to compare data science needs to needs from other related disciplines. When success is judged not by knowledge but by uptime or performance, the task is software engineering. When the task is judged by minimizing classification error or regret, without regard to how the results inform a larger discussion, the task is applied machine learning. When results are judged by the risk of legal action or issues of compliance, the task is one of risk management. These are each valuable and worthwhile tasks, and they require similar steps of scoping to get right, but they are not problems of data science.

Consider some descriptions of some fairly common needs, all ones that I have seen in practice. Each of these is much condensed from how they began their life:

- The managers want to expand operations to a new location. Which one is likely to be most profitable?
- Our customers leave our website too quickly, often after only reading one article. We don't understand who they are, where they are from, or when they leave, and we have no framework for experimenting with new ideas to retain them.
- We want to decide between two competing vendors. Which is better for us?
- Is this email campaign effective at raising revenue?
- We want to place our ads in a smart way. What should we be optimizing? What is the best choice, given those criteria?

And here are some famous ones from within the data world:

- We want to sell more goods to pregnant women. How do we identify them from their shopping habits?

- We want to reduce the amount of illegal grease dumping in the sewers. Where might we look to find the perpetrators?

Needs will rarely start out as clear as these. It is incumbent upon us to ask questions, listen, and brainstorm until we can articulate them clearly and they can be articulated clearly back to us. Again, writing is a big help here. By writing down what we think the need is, we will usually see flaws in our own reasoning. We are generally better at criticizing than we are at making things, but when we criticize our own work, it helps us create things that make more sense.

Like designers, the process of discovering needs largely proceeds by listening to people, trying to condense what we understand, and bringing our ideas back to people again. Some partners and decision makers will be able to articulate what their needs are. More likely they will be able to tell us stories about what they care about, what they are working on, and where they are getting stuck. They will give us places to start. Sometimes those we talk with are too close to their task to see what is possible. We need to listen to what they are saying, and it is our job to go beyond listening and actively ask questions until we can clearly articulate what needs to be understood, why, and by whom.

Often the information we need to understand in order to refine a need is a detailed understanding of how some process happens. It could be anything from how a widget gets manufactured to how a student decides to drop out of school to how a CEO decides when to end a contract. Walking through that process one step at a time is a great tactic for figuring out how to refine a need. Drawing diagrams and making lists make this investigation clearer. When we can break things down into smaller parts, it becomes easier to figure out where the most pressing problems are. It can turn out that the thing we were originally worried about was actually a red herring or impossible to measure, or that three problems we were concerned about actually boiled down to one.

When possible, a well-framed need relates directly back to some particular action that depends on having good intelligence. A good need informs an action rather than simply informing. Rather than saying, “The manager wants to know where users drop out on the way to buying something,” consider saying, “The manager wants more users to finish their purchases. How do we encourage that?” Answering the first question is a component of doing the second, but the action-oriented formulation opens up more possibilities, such as testing new designs and performing user experience interviews to gather more data.

If it is not helpful to phrase something in terms of an action, it should at least be related to some larger strategic question. For example, understanding how users of a product are migrating from desktop to mobile versions of a website is useful for informing the product strategy, even if there is no obvious action to take afterward. Needs should always be specified in words that are important to the organization, even if they're only questions.

Until we can clearly articulate the needs we are trying to meet, and until we understand how meeting those specific needs will help the organization achieve its larger goals, we don't know why we're doing what we're hoping to do. Without that part of a scope, our data work is mostly going to be fluff and only occasionally worthwhile.

Continuing from the longer examples, here are some needs that those organizations might have:

- The nonprofit that reunited families does not have a good way to measure its success. It is prohibitively expensive to follow up with every individual to see if they have contacted their families. By knowing when individuals are doing well or poorly, the nonprofit will be able to judge the effectiveness of changes to its strategy.
- The marketing department at the shoe company does not have a smart way of selecting cities to advertise to. Right now it is selecting its targets based on intuition, but it thinks there is a better way. With a better way of selecting cities, the department expects sales will go up.
- The media organization does not know the right way to define an engaged reader. The standard web metric of unique daily users doesn't really capture what it means to be a reader of an online newspaper. When it comes to optimizing revenue, growth, and promoting subscriptions, 30 different people visiting on 30 different days means something very different from 1 person visiting for 30 days in a row. What is the right way to measure engagement that respects these goals?
- The anti-corruption advocacy group does not have a good way to automatically collect and collate media mentions of politicians. With an automated system for collecting media attention, it will spend less time and money keeping up with the news and more time writing it.

Note that the need is *never* something like, "the decision makers are lacking in a dashboard," or predictive model, or ranking, or what have you. These are potential solutions, not needs. Nobody except a car driver needs a dashboard. The need is not for the dashboard or model, but for something that actually matters in words that decision makers can usefully think about.

This is a point that bears repeating. A data science need is a problem that can be solved with knowledge, not a lack of a particular tool. Tools are used to accomplish things; by themselves, they have no value except as academic exercises. So if someone comes to you and says that her company needs a dashboard, you need to dig deeper. Usually what the company needs is to understand how they are performing so they can make tactical adjustments. A dashboard may be one way of accomplishing that, but so is a weekly email or an alert system, both of which are more likely to be incorporated into someone's workflow.

Similarly, if someone comes to you and tells you that his business needs a predictive model, you need to dig deeper. What is this for? Is it to change something that he doesn't like? To make accurate predictions to get ahead of a trend? To automate a process? Or does the business need to generalize to a new case that's unlike any seen in order to inform a decision? These are all different needs, requiring different approaches. A predictive model is only a small part of that.

Vision (V)

Before we can start to acquire data, perform transformations, test ideas, and so on, we need some vision of where we are going and what it might look like to achieve our goal.

The vision is a glimpse of what it will look like to meet the need with data. It could consist of a mockup describing the intended results, or a sketch of the argument that we're going to make, or some particular questions that narrowly focus our aims.

Someone who is handed a data set and has not first thought about the context and needs of the organization will usually start and end with a narrow vision. It is rarely a good idea to start with data and go looking for things to do. That leads to stumbling on good ideas, mostly by accident.

Having a good vision is the part of scoping that is most dependent on experience. The ideas we will be able to come up with will mostly be variations on things that we have seen before. It is tremendously useful to acquire a good mental library of examples by reading widely and experimenting with new ideas. We can expand our library by talking to people about the problems they've solved, reading books

on data science or reading classics (like Edward Tufte and Richard Feynman), following blogs, attending conferences and meetups, and experimenting with new ideas all the time.

There is no shortcut to gaining experience, but there is a fast way to learn from your mistakes, and that is to try to make as many of them as you can. Especially if you are just getting started, creating things in quantity is more important than creating things of quality. There is a saying in the world of Go (the east Asian board game): lose your first fifty games of Go as quickly as possible.

The two main tactics we have available to us for refining our vision are mockups and argument sketches.

A mockup is a low-detail idealization of what the final result of all the work might look like. Mockups can take the form of a few sentences reporting the outcome of an analysis, a simplified graph that illustrates a relationship between variables, or a user interface sketch that captures how people might use a tool. A mockup primes our imagination and starts the wheels turning about what we need to assemble to meet the need. Mockups, in one form or another, are the single most useful tool for creating focused, useful data work (see Figure 1-1).

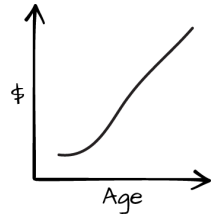


Figure 1-1. A visual mockup

Mockups can also come in the form of sentences:

Sentence Mockups

The probability that a female employee asks for a flexible schedule is roughly the same as the probability that a male employee asks for a flexible schedule.

There are 10,000 users who shopped with service X. Of those 10,000, 2,000 also shopped with service Y. The ones who shopped with service Y skew older, but they also buy more.

Keep in mind that a mockup is not the actual answer we expect to arrive at. Instead, a mockup is an example of the kind of result we would expect, an illustration of the form that results might take. Whether we are designing a tool or pulling data together, concrete knowledge of what we are aiming at is incredibly valuable.

Without a mockup, it's easy to get lost in abstraction, or to be unsure what we are actually aiming toward. We risk missing our goals completely while the ground slowly shifts beneath our feet. Mockups also make it much easier to focus in on what is important, because mockups are shareable. We can pass our few sentences, idealized graphs, or user interface sketches off to other people to solicit their opinion in a way that diving straight into source code and spreadsheets can never do.

A mockup shows what we should expect to take away from a project. In contrast, an argument sketch tells us roughly what we need to do to be convincing at all. It is a loose outline of the statements that will make our work relevant and correct. While they are both collections of sentences, mockups and argument sketches serve very different purposes. Mockups give a flavor of the finished product, while argument sketches give us a sense of the logic behind the solution.

For example, if we want to know whether women and men are equally interested in flexible time arrangements, there are a few parts to making a convincing case. First, we need to have a good definition of who the women and men are that we are talking about. Second, we need to decide if we are interested in subjective measurement (like a survey), if we are interested in objective measurement (like the number of applications for a given job), or if we want to run an experiment. We could post the same job description but only show postings with flexible time to half of the people who visit a job site. There are certain reasons to find each of these compelling, ranging from the theory of survey design to mathematical rules for the design of experiments.

Thinking concretely about the argument made by a project is a valuable tool for orienting ourselves. Chapter 3 goes into greater depth about what the parts of an argument are and how they relate to working with data. Arguments occur both in a project and around the project, informing both their content and their rationale.

Pairing written mockups and written argument sketches is a concise way to get our understanding across, though sometimes one is more appropriate than the other. Continuing again with the longer examples:

Example 1

- Vision: The nonprofit that is trying to measure its successes will get an email of key performance indicators on a regular basis. The email will consist of graphs and automatically generated text.
- Mockup: After making a change to our marketing, we hit an enrollment goal this week that we've never hit before, but it isn't being reflected in the success measures.
- Argument sketch: The nonprofit is doing well (or poorly) because it has high (or low) values for key performance indicators. After seeing the key performance indicators, the reader will have a good sense of the state of the nonprofit's activities and will be able to adjust accordingly.

Example 2

Here are several ideas for the marketing department looking to target new cities, depending on the details of the context:

Idea 1

- Vision: The marketing department that wants to improve its targeting will get a report that ranks cities by their predicted value to the company.
- Mockup: Austin, Texas, would provide a 20% return on investment per month. New York City would provide an 11% return on investment per month.
- Argument sketch: The department should focus on city X, because it is most likely to bring in high value. The definition of high value that we're planning to use is substantiated for the following reasons....

Idea 2

- Vision: The marketing department will get some software that implements a targeting model, which chooses a city to place advertisements in. Advertisements will be targeted automatically based on the model, through existing advertising interfaces.
- Mockup: 48,524 advertisements were placed today in 14 cities. 70% of them were in emerging markets.

- Argument sketch: Advertisements should be placed proportional to their future value. The department should feel confident that this automatic selector will be accurate without being watched.

Idea 3

- Vision: The marketing department will get a spreadsheet that can be dropped into the existing workflow. It will fill in some characteristics of a city and the spreadsheet will indicate what the estimated value would be.
- Mockup: By inputting gender and age skew and performance results for 20 cities, an estimated return on investment is placed next to each potential new market. Austin, Texas, is a good place to target based on age and gender skew, performance in similar cities, and its total market size.
- Argument sketch: The department should focus on city X, because it is most likely to bring in high value. The definition of high value that we're planning to use is substantiated for the following reasons....

Example 3

- Vision: The media organization trying to define user engagement will get a report outlining why a particular user engagement metric is the ideal one, with supporting examples; models that connect that metric to revenue, growth, and subscriptions; and a comparison against other metrics.
- Mockup: Users who score highly on engagement metric A are more likely to be readers at one, three, and six months than users who score highly on engagement metrics B or C. Engagement metric A is also more correlated with lifetime value than the other metrics.
- Argument sketch: The media organization should use this particular engagement metric going forward because it is predictive of other valuable outcomes.

Example 4

- Vision: The developers working on the corruption project will get a piece of software that takes in feeds of media sources and rates the chances that a particular politician is being talked about. The staff will set a list of names

and affiliations to watch for. The results will be fed into a database, which will feed a dashboard and email alert system.

- Mockup: A typical alert is that politician X, who was identified based on campaign contributions as a target to watch, has suddenly showed up on 10 news talk shows.
- Argument sketch: We have correctly kept tabs on politicians of interest, and so the people running the anti-corruption project can trust this service to do the work of following names for them.

In mocking up the outcome and laying out the argument, we are able to understand what success could look like. The final result may differ radically from what we set out to do. Regardless, having a rough understanding at the outset of a project is important. It is also okay to have several potential threads at this point and be open to trying each, such as with the marketing department example. They may end up complementing each other.

The most useful part of making mockups or fragments of arguments is that they let us work backward to fill in what we actually need to do. If we're looking to send an email of key performance indicators, we'd better come up with some to put into the email. If we're writing a report outlining why one engagement metric is the best and tying it to a user valuation model, we need to come up with an engagement metric and find or develop a user valuation model. The pieces start to fall into place.

At the end of everything, the finished work will often be fairly simple. Because of all of the work done in thinking about context and need, generating questions, and thinking about outcomes, our work will be the right kind of simple. Simple results are the most likely to get used.

Because of all of the work done in thinking about context and need, generating questions, and thinking about outcomes, our work will be the right kind of simple.

They will not always be simple, of course. Having room to flesh out complicated ideas is part of the point of thinking so much at the outset. When our work is complicated, we will benefit even more from having thought through some of the parts first.

When we're having trouble articulating a vision, it is helpful to start getting something down on paper or out loud to prime our brains. Drawing pretend graphs, talking through examples, making flow diagrams on whiteboards, and so on, are all good ways to get the juices flowing.

Outcome (O)

We need to understand how the work will actually make it back to the rest of the organization and what will happen once it is there. How will it be used? How will it be integrated into the organization? Who will own its integration? Who will use it? In the end, how will its success be measured?

If we don't understand the intended use of what we produce, it is easy to get lost in the weeds and end up making something that nobody will want or use. What's the purpose of all this work if it does nobody any good?

The outcome is distinct from the vision; the vision is focused on what form the work will take at the end, while the outcome is focused on what will happen when we are "done." Here are the outcomes for each of the examples we've been looking at so far:

- The metrics email for the nonprofit needs to be set up, verified, and tweaked. Sysadmins at the nonprofit need to be briefed on how to keep the email system running. The CTO and CEO need to be trained on how to read the metrics emails, which will consist of a document written to explain it.
- The marketing team needs to be trained in using the model (or software) in order to have it guide their decisions, and the success of the model needs to be gauged in its effect on sales. If the result ends up being a report instead, it will be delivered to the VP of Marketing, who will decide based on the recommendations of the report which cities will be targeted and relay the instructions to his staff. To make sure everything is clear, there will be a follow-up meeting two weeks and then two months after the delivery.
- The report going to the media organization about engagement metrics will go to the head of online business. If she signs off on its findings, the selected user engagement metric will be incorporated by the business analysts into the performance measures across the entire organization. Funding for existing and future initiatives will be based in part on how they affect the new engagement metric. A follow-up study will be conducted in six months to verify that the new metric is successfully predicting revenue.

- The media mention finder needs to be integrated with the existing mention database. The staff needs to be trained to use the dashboard. The IT person needs to be informed of the existence of the tool and taught how to maintain it. Periodic updates to the system will be needed in order to keep it correctly parsing new sources, as bugs are uncovered. The developers who are doing the integration will be in charge of that. Three months after the delivery, we will follow up to check on how well the system is working.

Figuring out what the right outcomes are boils down to three things. First, who will have to handle this next? Someone else is likely to have to interpret or implement or act on our work. Who are they, what are their requirements, and what do we need to do differently from our initial ideas to address their concerns?

Second, who or what will handle keeping this work relevant, if anyone? Do we need to turn our work into a piece of software that runs repeatedly? Will we have to return in a few months? More often than not, analyses get re-run, even if they are architected to be run once.

Third, what do we hope will change after we have finished the work? Note again that "having a model" is not a suitable change; what *in terms that matter to the partners* will have changed? How will we verify that this has happened?

Thinking through the outcome before embarking on a project, along with knowing the context, identifying the right needs, and honing our vision, improves the chance that we will do something that actually gets used.

Seeing the Big Picture

Tying everything together, we can see that each of these parts forms a coherent narrative about what we might accomplish by working with data to solve this problem.

First, let's see what it would look like to sketch out a problem without much structured thinking:

> *We will create a logistic regression of web log data using SAS to find patterns in reader behavior. We will predict the probability that someone comes back after visiting the site once.*

Compare this to a well-thought-out scope:

This media organization produces news for a wide audience. It makes money through advertising and premium subscriptions to its content. The person who asked for some advice is the head of online business.

This organization does not know the right way to define an engaged reader. The standard web metric of unique daily users doesn't really capture what it means to be a reader of an online newspaper. When it comes to optimizing revenue, growth, and promoting subscriptions, 30 different people visiting on 30 different days means something very different from 1 person visiting for 30 days in a row. What is the right way to measure engagement that respects these goals?

When this project is finished, the head of online business will get a report outlining why a particular user engagement metric is the ideal one, with supporting examples; models that connect that metric to revenue, growth, and subscriptions; and a comparison against other metrics.

If she signs off on its findings, the selected user engagement metric will be incorporated into the performance measures across the entire organization. Institutional support and funding for existing and future initiatives will be based in part on how they affect the new engagement metric. A follow-up study will be conducted in six months to verify that the new metric is successfully predicting revenue, growth, and subscription rates.

A good story about a project and a good scope of a project are hard to tell apart.

It is clear that at the outset, we do not actually know what the right metric will be or even what tools we will use. Focusing on the math or the software at the expense of the context, need, vision, and outcome means wasted time and energy.

What Next?

With a basic understanding of the four areas of a project scope (context, needs, vision, and outcome), we turn our attention to filling in the details of the project. By thinking deeply before digging into the data, we maximize our chances of doing something useful as opposed to simply the first things that come to mind.

Working with data is a process that you lose yourself in. There is a natural tension between going into exploration as quickly as possible and spending more time thinking and planning up front. When balanced properly, they are mutually beneficial. However, diving in quickly and getting lost in the data exerts a natural siren song on those of us who work with data professionally. It takes effort and patience to put time into thinking up front, but it is effort that is duly rewarded.

Before we start down rabbit holes that may or may not take us somewhere useful, and after we have a rough project scope, we need to take some more steps to clarify the details of the problem we are working on. That process is the focus of this chapter. This includes important discussions with decision makers and implementers, figuring out how to define key terms, considering what arguments we might make, posing open questions to ourselves, and deciding in what order to pursue different ideas.

There is no particular order to these steps. A project might be so simple that every area is obvious and we don't need to engage with anybody else or do any more thinking before we dive into the data work. *This is rare*. More than likely, there will be things that need clarification in our own heads (and in the minds of others) to avoid wasted effort.

It's possible to know everything you need to know for a small, personal project before you even begin. Larger projects, which are more likely to cause something important to change, always have messier beginnings. Information is incomplete, expectations are miscalibrated, and definitions are too loose to be useful. In the same way that the nitty-gritty of data science presumes messier data than is given for problems in a statistics course, the problem definition for large, applied problems is always messier than the toy problems we think up ourselves.

As we move on to the rest of the project, it's critical to remember to take careful notes along the way. There are minor intellectual and technical decisions made throughout a project that will be crucial in writing the final documentation. Having a final, written version of the work we do means a much greater chance to reproduce our work again months or years down the line. It also means we are more likely to catch our own errors as we put our ideas down into words.

Refining the Vision

The vision we expressed in our first pass at a scope is often sufficient to get started, but not complete enough to guide our actions.

We refine our vision by improving our intuition about the problem. We improve our intuition by talking to people, trying out ideas, gathering questions, and running simple experiments. We want to spend time up front maximizing our understanding. It pays to make our early work investigative rather than definitive.

Pointed questions explore the limits of our current knowledge, and focusing on question generation is a good use of time. Good questions also offer up new ways to frame a problem. At the end of the day, it is usually how we frame the problem, not the tools and techniques that we use to answer it, that determine how valuable our work is.

Some of these questions will be preliminary and serve to illustrate the breadth of the problem, such as knowing whether there are ten thousand or ten million purchases per month to study. Others will form the core of the work we are looking to undertake, such as how exactly those purchases are related over time for the same customer.

One technique for coming up with questions is to take a description of a need or of a process that generated our data and to ask every question that we can think of—this is called *kitchen sink interrogation*. In a kitchen sink interrogation, we are generating questions, not looking for answers. We want to get a sense of the lay of the land. A few minutes up front can save days or weeks down the line.

If our customers leave our website too quickly, why do they leave? What does it mean to leave? At what points do they leave? What separates the ones who leave from the ones who stay? Are there things that we have done before that have changed customer behavior? Are there quick things we can try now? How do we know what their behavior is? How reliable is that source? What work has already been done on this problem?

If we're trying to understand user engagement, what metrics are already being used? Where do they break down? What are they good at predicting? What are some

alternative metrics that we haven't looked at yet? How will we validate what a good metric is? By collecting questions with a kitchen sink interrogation, we start to get a sense for what is known and what is unknown.

Another technique, *working backward*, starts from the mockups or argument sketches and imagines each step that has to be achieved between the vision and where we are right now. In the process of working backward, we kick up a number of questions that will help to orient us. When we're lucky, we will figure out that a certain task is not feasible long before we've committed resources to it.

The same techniques discussed in Chapter 1 do not go away once we have a basic sense of the vision. Mockups and argument sketches are continuously useful. Having a clear vision of what our goal looks like—whether it's in the form of a sentence describing what we would learn or a hand-drawn sketch of a graph—is incredibly instructive in its production and a wonderful guiding light when we are deep in the trenches. Having a clear idea of what numbers we expect to come out of a process before we start it also means that we will catch errors right away.

We can also borrow tactics that we used to refine needs. Walking through a scenario or roleplaying from the perspective of the final consumer is a good way to catch early problems in our understanding of what we are aiming at. If we are producing a map or a spreadsheet or an interactive tool, there is always going to be someone on the other side. Thinking about what their experience will be like helps keep us focused.

Once results start to come in, in whatever form makes sense for the work we are doing, it pays to continually refer back to this early process to see if we are still on track. Do these numbers make sense? Is the scenario we envisioned possible?

Techniques for refining the vision

Interviews
: Talk to experts in the subject matter, especially people who work on a task all the time and have built up strong intuition. Their intuition may or may not match the data, but having their perspective is invaluable at building your intuition.

Rapid investigation
: Get order of magnitude estimates, related quantities, easy graphs, and so on, to build intuition for the topic.

Kitchen sink interrogation
: Ask every question that comes to mind relating to a need or a data collection process. Just the act of asking questions will open up new ideas. Before it was polluted as a concept, this was the original meaning of the term brainstorming.

Working backward
: Start from the finished idea and figure out what is needed immediately prior in order to achieve the outcome. Then see what is prior to that, and prior to that, and so on, until you arrive at data or knowledge you already have.

More mockups
: Drawing further and more well-defined idealizations of the outcome not only helps to figure out what the actual needs are, but also more about what the final result might look like.

Roleplaying
: Pretend you are the final consumer or user of a project, and think out loud about the process of interacting with the finished work.

Deep Dive: Real Estate and Public Transit

An extended example will be instructive. Suppose that a firm in New York that controls many rental properties is interested in improving its profitability on apartment buildings it buys. It considers itself a data-driven company, and likes to understand the processes that drive rental prices. It has an idea that public transit access is a key factor in rental prices, but is not sure of the relationship or what to do with it.

We have a context (a data-driven New York residential real estate company) and a vague need (it wants to somehow use public transit data to improve its understanding of rental prices). After some deep conversation, and walking through scenarios of what it might do if it understood how transit access affects rental prices, it turns out the company actually has several specific needs.

First and simplest, it wants to confirm its hunch that rental prices are heavily dependent on public transit access in New York. Even just confirming that there is a relationship is enough to convince the company that more work in this area is warranted. Second, it wants to know if some apartments may be under- or overpriced relative to their worth. If the apartments are mispriced, it will help the

company set prices more effectively, and improve profitability. And third, the company would love to be able to predict where real estate prices are heading.

Note that the latter two needs did not mention public transit data explicitly. It may turn out in the process of working with this data that public transit data isn't useful, but all the other data we dig up actually is! Will the real estate company be disappointed? Certainly not. Public transit data will be the focus of our work, but the goal isn't so much to use public transit data as it is to improve the profitability of the company. If we stick too literally to the original mandate, we may miss opportunities. We may even come up with other goals or opportunities in the course of our analyses.

Before we go too far, what is our intuition telling us? Knowing the subject matter, or talking to subject matter experts, is key here. Reading apartment advertisements would be a good way to build up an understanding of what is plausible. Apartment prices probably are higher close to transit lines; certainly listings on real estate websites list access to trains as an amenity. Putting ourselves in the shoes of someone getting to work, we can realize that the effect likely drops off rapidly, because people don't like to walk more than 10 or 15 minutes if they can help it. The effects are probably different in different neighborhoods and along different transit lines, because different destinations are more interesting or valuable than others.

Moving on to the vision, we can try out a few ideas for what a result would look like. If our final product contained a graph, what would it be a graph of? Roughly speaking, it would be a graph of "price" against "nearness to transit," with price falling as we got farther away from transit. In reality it would be a scatterplot, but drawing a line graph is probably more informative at this stage. Actually sketching a mockup of a basic graph, with labels, is a useful exercise (Figure 2-1).

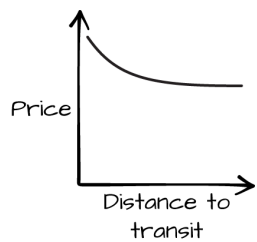


Figure 2-1. Mockup graph

We can recognize from this that we will need some way to define price and proximity. This presentation is probably too simple, because we know that the

relationship likely depends on other factors, like neighborhood. We would need a series of graphs, at least. This could be part of a solution to the first two needs, verifying that there is a strong relationship between public transit and the housing market, and trying to predict whether apartments are under- or overpriced.

Digging into our experience, we know that graphs are just one way to express a relationship. Two others are models and maps. How might we capture the relevant relationships with a statistical model?

A statistical model would be a way to relate some notion of transit access to some notion of apartment price, controlling for other factors. We can clarify our idea with a mockup. The mockup here would be a sentence interpreting the hypothetical output. Results from a model might have conclusions like, "In New York City, apartment prices fall by 5% for every block away from the A train, compared to similar apartments." Because we thought about graphs already, we know that one of the things we will need to control for in a model is neighborhood and train line. A good model might let us use much more data. For example, investigating the government data archives on this topic reveals that turnstile data is freely available in some cities.

A model has the potential to meet all three of our needs, albeit with more effort. Model verification would let us know if the relationship is plausible, outlier detection would allow us to find mispriced apartments, and running the model on fake data would allow us to predict the future (to some extent). Each of these may require different models or may not be plausible, given the data that is available. A model might also support other kinds of goals—for example, if we wanted to figure out which train line had the largest effect on prices.

If our vision is a transit map, it would be a heat map of apartment prices, along with clearly marked transit lines and probably neighborhood boundaries. There would need to be enough detail to make the city's layout recognizable. Depending on the resolution of the map, this could potentially meet the first two needs (making a case for a connection and finding outliers) as well, through visual inspection. A map is easier to inspect, but harder to calibrate or interpret.

Each has its strengths and weaknesses. A scatterplot is going to be easy to make once we have some data, but potentially misleading. The statistical model will collapse down a lot of variation in the data in order to arrive at a general, interpretable conclusion, potentially missing interesting patterns. The map is going to be limited in its ability to account for variables that aren't spatial, and we may have a harder time interpreting the results. Each would lend itself to a variety of arguments.

What we finally end up with will probably be more complicated than the basic things we outline here. There may actually be a combination of two or all three of these, or some output we haven't considered yet; maybe a website that the firm can use to access the model predictions with a few knobs to specify apartment details, or a spreadsheet that encodes a simple model for inclusion in other projects. Any graph, model, or map we make for this project will depend on additional bits of analysis to back up their conclusions.

Another way to explain this process is to say that we begin with strong assumptions and slowly relax them until we find something we can actually achieve. A graph of price against proximity, or a user interface with two buttons, is almost certainly too simple to be put into practice. To make such a project work requires much stronger assumptions that we can make in practice. That shouldn't stop us from trying to express our ideas in this kind of clean way. Sometimes the details, even when they take up most of our time, are only epicycles on top of a larger point that we will be worse off if we forget.

Don't forget the utility of a few concrete examples in spurring the imagination. Before building a map, we should try plugging a few intersections into real estate websites to get a feel for how the aspects of homes might vary with distance and price. The same goes for reading classifieds. There may be entire aspects of apartments that were not obvious at first glance, like proximity to highly regarded schools, that will be mentioned in the apartment description and could have a huge effect on price. Always seek to immerse yourself in some particular examples, even if it just means reading the first 10 or 20 lines of a table in depth before building a model.

Always seek to immerse yourself in some particular examples, even if it just means reading the first ten or twenty lines of a table in depth before building a model.

Deep Dive Continued: Working Forward

Having imagined the end of our work, it is helpful to think about what kind of data is appropriate for defining the variables. Having spread our wings, it is time to get a little realistic and start working forward from what we have.

What will we use for apartment prices? It is common in the real estate industry to use price per square foot, to normalize against differences in apartment size. Finding historical price-per-square-foot data across an entire city may be as simple

as purchasing a database, or it could be a much more involved process of connecting public and private data together.

And what is transit access? Note that, despite the easy way we were able to draw that initial graph, it is not clear at first blush how to even define the term transit access! A little kitchen sink interrogation is useful here.

First, what is transit? The initial conversation was sparked from subway lines. Do buses count? Bus access will be much harder to show on a map than train access, but buses are a necessity in areas that are less well connected to trains. Knowing where the company actually operates might be useful here. How long do people actually walk? Where do people in each neighborhood actually go? Is that information available? Are there people we could talk to about getting end-to-end transit data, maybe from existing surveys? Could employment records be useful?

"Transit access" itself could be about walking distance to train or bus lines, or it could be about average travel time from a point to important landmarks, like the Empire State Building or Wall Street in New York City. Which one we pick will make a big difference!

In refining the vision we can also recognize that this is a causal question of sorts (how much does being near a subway station increase prices *compared to an identical apartment that was farther away*?), and therefore calls for a causal argument pattern. Chapters 4 and 5 cover argument patterns in detail, but for our purposes we can recognize that we will, at a minimum, need to acquire additional information to help distinguish the effect of proximity to transit from, say, higher prices on more luxurious apartments. More luxurious apartments may have been built closer to the subway to take advantage of the better location, and so on.

Further refining the vision, we know that apartment prices will be a continuous variable, neighborhood will probably be an important confounder, and each transit line will probably contribute a different amount. We will need locations of apartments and transit stops, information on subways accessed by each stop, and, if we build a model, a reasonable distance or travel time function to tie things together. If we want to understand how these things change over time, we will need not only a snapshot, but also a historical record. The breadth of making a full model starts to become clear in a way it might not have been at the start.

At this stage we may become aware of the limitations we are likely to face. It will probably be hard to encode an "apartment quality" measure. A proxy metric, like some sense of how recently or frequently an apartment was refurbished, requires additional data like city records. Our results may be hard to interpret without a great deal of work, but it may be good enough for our needs. And if we want to

understand historical relationships between transit connectivity and apartment prices, we have to figure out how far back to go and how to handle the additional complexities inherent in working with time data.

Thinking hard about the outcome can clear this up. What will be different after we are done? Might the easiest need be sufficient for now? A more purely observational study would be fine. Or might there be enough buy-in to get this work widely used within the firm? And is the time component really that valuable? Each of these goals is different, the arguments that are needed are different, and they will call for different levels of investment of time and energy. If we don't think about how the work will be used after we finish, we may end up working on something pointless.

Who will maintain this work after we finish? Keeping a map up-to-date is probably easier than a model with a dozen separate data sources. Are all the sources we are interested in available programmatically, or would we have to take weeks of time to get them again next year?

How will we know if we have done a good job? How do we cross-check our results? For example, we could look at how quickly or slowly each apartment was rented, as a way of verifying that we predicted over- or underpricing correctly. Naturally, this is complicated by the speed with which the rental market moves in a big city, but it is worth a thought nevertheless.

Deep Dive Continued: Scaffolding

Having elaborated our vision and what the pieces are that we plan to work with, the next step is to consider our project's scaffolding. How can we go about our tasks so that at each step we can evaluate what is taking shape and see if we need to change direction? We want to avoid looking back in horror at having wasted our time on something useless.

Especially at the beginning, we want to find things to do that will be fast and informative. The simple truth is that we don't know in advance what will be the right things to pursue; and if we knew that already, we would have little need for our work. Before we do anything slow, or only informative at the margins, we want to focus on building intuition—and eventually that means poking around with data.

If we have already collected some data, simple tabulations, visualizations, and reorganized raw data are the best way to quickly build intuition. Just combining and slicing various relevant data sets can be very informative, as long as we do not get stuck on this as our main task.

Models that can be easily fit and interpreted (like a linear or logistic model), or models that have great predictive performance without much work (like random forests), serve as excellent places to start a predictive task. Using a scatterplot of latitude and longitude points as a first approximation map is a great way to start a geospatial project. And so on.

It is important, though, to not get too deep into these exploratory steps and forget about the larger picture. Setting time limits (in hours or, at most, days) for these exploratory projects is a helpful way to avoid wasting time. To avoid losing the big picture, it also helps to write down the intended steps at the beginning. An explicitly written-down *scaffolding plan* can be a huge help to avoid getting sucked deeply into work that is ultimately of little value. A scaffolding plan lays out what our next few goals are, and what we expect to shift once we achieve them.

It also helps when we understand the argument or arguments we are looking to make. Understanding the outline of our argument will lead us to discover which pieces of analysis are most central. Chapter 3 discusses the details of arguments, including transformation, evidence, justifications, and arranging claims. These let us solve potentially complicated needs with data. With a sketch of the argument in place, it is easier to figure out the most central thing we need to work on. The easiest way to perform this sketching is to write out our ideas as paragraphs and imagine how we will fill in the details.

In the case of the apartment prices and public transit, finding or plotting a map of apartment prices next to a base layer of transit connections is probably the easiest thing to do first. By looking at the map, we can see whether such a relationship seems plausible, and start to gain intuition for the problem of making scatterplots or building a model.

Building exploratory scatterplots should precede the building of a model, if for no reason other than to check that the intuition gained from making the map makes sense. The relationships may be so obvious, or the confounders so unimportant, that the model is unnecessary. A lack of obvious relationships in pairwise scatterplots does not mean that a model of greater complexity would not be able to find signal, but if that's what we're up against, it is important to know it ahead of time. Similarly, building simple models before tackling more complex ones will save us time and energy.

Scaffolding is the art of prioritizing our aims and not going too far down that rabbit hole. How can we proceed in a way that is as instructive as possible at every step?

Verifying Understanding

In any scaffolding plan, it is important to build in explicit checks with the partners or decision makers to ensure that we understand their needs properly. It keeps us focused, and it builds trust. It is better to overcommunicate, especially with new partners, than it is to assume that we are on the same page for a week only to find we have built something pointless.

Find a convenient medium and explain the partners' needs back to them, asking if you have understood things properly. Mention generally what your thoughts are around your vision, such as the form that the results would take, without going into too much detail. The goal is conceptual agreement, not a detailed critique of the project, unless they are data-savvy and particularly interested. The details will evolve throughout the project anyway.

Explicitly ask them if they agree that you have understood what they are looking for, and if they have any more questions. You should feel confident that you are doing something that they will use. This doesn't need to be a formal meeting; it can often be accomplished with a quick conversation.

If one or more people will be needed to implement the final work, talk to them and make sure that their requirements are being represented. If someone else will be installing a script we create into production software, who will be in charge of cleaning it up and keeping it running? Talk to them and make sure that you understand what they need.

We will go through basically the same process on a regular basis for large projects, and at least when all the work is done for a project of any size, so pay attention to how these discussions go.

Partners and decision makers often have intuitive understandings of the processes they are looking to understand better, or at least have some idea of what concrete examples of what they're interested in will look like. Intuition like that is invaluable, and should never be overlooked. Spending lots of time talking to people who deal with a process is a smart way to get the intuition needed to build a data-based argument that can create real knowledge.

We know that we have grasped the problem well when we can explain our strategy to these partners in terms that matter to them (even if they have no knowledge of data science), and we receive back enthusiastic understanding.

Getting Our Hands Dirty

Once we have data, we may find that our initial ideas were wrong or that the need can be met even more easily than we thought. Regardless, thinking explicitly before we dive into them will make what we do far more productive.

We need to spend time data gathering: actually acquiring the data we will need. This step might be easy, or it may take a long time. We might have one database or API call, or we may need to strategize about how to store all the data we will require. We may need to contact people in positions of authority in order to acquire data. We may need to make trade-offs of accuracy against price or time to acquire new data.

Once some data is gathered, we can begin the transformations. We usually put raw data into a common format, then transform the data into graphs, models, tables, and summaries that can serve as evidence for a larger argument. These steps can take the longest amount of time. As part of the scaffolding, we should plan to start with easy transformations (like exploratory graphs or summary statistics) and then easy models, before moving on to more sophisticated or complicated models. Often, easy transformations will serve well enough to make a valid argument and additional work is not necessary.

Once the data is transformed and ready to serve as evidence, we can evaluate the strength of our arguments. By updating the few paragraphs we wrote out at the beginning with new information, we will know if what we have done fits into the larger goal or not. Is there another thing we forgot to justify or that we need to explore? Do we need to continue on to make a more complicated model? Are there alternative interpretations of what we have found that require us to find something else to help us decide between them? Does this argument even make any sense, given the data we have collected?

Say, for example, that we're outlining our argument to ourselves after collecting public transit and apartment price, and we realize that we're not sure if we have an unbiased sample of apartments. We have choices; do we want to acknowledge that bias and claim the conclusions to be more limited? Or do we perhaps want to make an argument as to why the data we have is actually representative of the city as a whole? We might find that most apartments are actually listed on Craigslist, or that, for the demographic that will be interested in using this information, nearly all apartments are listed on Craigslist. Strong arguments will rarely consist only of a single piece of evidence, though other strands of the arguments around them may be implied or gestured at.

After we have arrived at a useful point, we can arrange the results into a pleasing form, keeping in mind the needs of our audience. Will we only show the most damning graph, or does it make more sense to present a chorus of models and summaries? Should we present a series of claims each chained to another, or present a number of claims in parallel? If we're building a tool, which things should be optional and which fixed? If we're writing a narrative, which examples should we use to reinforce our point?

Finally, we need to present what we've done. The actual written copy, the final form of the graphs, the neat table, the interactive tool with a carefully designed UI —these are all part of having a good presentation. Presentation matters tremendously. At the very least, there are genre conventions in every field that are worth following in order to be taken seriously; tone is important in presentation. Highly polished, beautiful graphics may be considered fussy in an academic setting, but are necessary for an earnings report for a design company. A very abstract problem presentation is inappropriate for a board meeting, but useful for demonstrating a technique to colleagues. And so on.

As in any creative field, working with data is not a linear process where we proceed from a grand statement of the problem at hand and gradually fill in the pieces until we are satisfied. Sometimes we are lucky and problems flow like that, but more often (and more interestingly), there is an interplay of clarification and action that slowly brings us to a better place than where we started.

Arguments

Data consists of observations about the world—records in a database, notes in a logbook, images on a hard drive. There is nothing magical about them. These observations may prove useful or useless, accurate or inaccurate, helpful or unhelpful. At the outset, they are only observations. Observations alone are not enough to act on. When we connect observations to how the world works, we have the opportunity to make knowledge. Arguments are what make knowledge out of observations.

There are many kinds of knowledge. Sometimes we have an accurate, unimpeachable mental model of how something works. Other times we have an understanding that is just good enough. And other times still, the knowledge is not in a person at all, but in an algorithm quietly puzzling out how the world fits together. What concerns us in working with data is how to get as good a connection as possible between the observations we collect and the processes that shape our world.

Knowing how arguments work gives us special powers. If we understand how to make convincing arguments, we can put tools and techniques into their proper place as parts of a whole. Without a good understanding of arguments, we make them anyway (we cannot help ourselves, working with data), but they are more likely to be small and disconnected.

By being aware of how arguments hang together, we can better:

- Get across complicated ideas
- Build a project in stages
- Get inspiration from well-known patterns of argument
- Substitute techniques for one another
- Make our results more coherent
- Present our findings
- Convince ourselves and others that our tools do what we expect

Thinking explicitly about arguments is a powerful technique, with a long history in philosophy, law, the humanities, and academic debate. It is a more fleshed-out example of using a structure to help us think with data. Thinking about the argument we are making can come into play at any point in working with a problem —from gathering ideas at the very beginning, to ensuring that we are making sense before releasing something into the wild.

Audience and Prior Beliefs

Only in mathematics is it possible to demonstrate something beyond all doubt. When held to that standard, we find ourselves quickly overwhelmed.

Our ideal in crafting an argument is a skeptical but friendly audience, suitable to the context. A skeptical audience is questioning of our observations, not swayed by emotional appeals, but not so skeptical as to be dismissive. The ideal audience is curious; humble, but not stupid. It is an idealized version of ourselves at our best, intelligent and knowledgeable but not intimately familiar with the problem at hand.

With the skeptical ideal in mind, it becomes easier to make a general argument, but it is also easier to make an argument to a specific audience. After making an argument for an ideal audience, it is easy to remove some parts and emphasize others to meet the needs of one or more particular audiences. Simplifying or expanding on certain things for an audience is fine, but lying is not. Something that good data work inherits from the scientific method is that it is bad form to cheat by preying on gullibility or ignorance. It is bad form, and in the long run it will cause the ruin of a business (or maybe a civilization).

An argument moves from statements that the audience already believes to statements they do not yet believe. At the beginning, they already agree with some statements about the world. After they hear the argument, there are new statements they will agree to that they would not have agreed to before. This is the key insight as to how an argument works—moving from prior belief to new belief to establish knowledge in a defensible way.

No audience, neither our ideal nor a real one, is 100% skeptical, a blank slate that doesn't already believe something. Many things are already background knowledge, taken for granted. Consider an argument that a rocket is safe to launch. There are certain statements that any reasonable audience will take for granted. The laws of physics aren't going to change mid-flight. Neither will multiplication tables. Whether those laws and our understanding of metallurgy and aerodynamics will result in a safe launch requires an argument. The background knowledge of the equations of motion and laws of chemistry do not.

Most of the prior beliefs and facts that go into an argument are never explicitly spelled out. To explicitly spell out every facet of an argument is silly. Russell and Whitehead famously took 379 pages to *set up the preliminaries* for the proof that $1+1=2$[1]. If the audience lacks some knowledge, either because of ignorance or reasonable doubt, it is important to be able to provide that (while not going overboard).

Prior belief, knowledge, and facts extend to more than just scientific laws. Audiences have dense webs of understanding that pre-date any argument presented to them. This collection of understanding defines what is reasonable in any given situation.

For example, one not-so-obvious but common prior belief is that the data in an argument comes from the same source that the arguer says it does. This is typically taken for granted. On the other hand, there may have been corruptions in what was collected or stored, or an intruder may have tampered with the data. If these would be reasonable possibilities to our skeptical audience (say the analysis involves an experimental sensor, or the audience is full of spooks), then any argument will need to address the question of validity before continuing. There needs to be something that the audience is tentatively willing to agree to, or else there is no way forward.

The algebra or mathematical theory behind specific techniques constitute another kind of common prior knowledge. Most arguments can safely avoid discussing these. Some real audiences may need their hand held more than others, or will be at our throats for execution details. But for the most part, the details of techniques are safely thought of as background knowledge.

Another source of prior or background knowledge is commonly known facts. Chicago is a city in the United States of America, which is a nation-state in the Northern Hemisphere on Earth, a planet. When it is necessary to compare Chicago to the whole US, their explicit relationship is rarely brought up. Of course, what is commonly understood varies by the actual audience. If the audience is in India, the location of Chicago in America may be an open issue. At that point, the audience will believe the atlas, and we'll be back to something that is commonly accepted.

"Wisdom" is also taken for granted in many arguments. When I worked at an online dating site, it was commonly taken for granted in internal discussions that men are more likely to send messages to women than the other way around. It was something that we had verified before, and had no reason to think had changed drastically. In the course of making an argument, it was already commonly under-

1. Volume 1 of *Principia Mathematica* by Alfred North Whitehead and Bertrand Russell (Cambridge University Press, page 379). The proof was actually not completed until Volume 2.

stood by the audience and didn't need to be spelled out or verified, though it could have been.

Not all wisdom can be verified. In the online dating example, we assumed that most of the people who had filled out their profiles actually corresponded to genuine human beings. We assumed that, for the most part, their gender identity offline matched their identity online. It may be painful to take some ideas for granted, but it is a necessity. We should aspire to act reasonably given the data at hand, not require omnipotence to move forward. People rarely require omnipotence in practice, but it might surprise you how many people seem to think it is a prerequisite for an explanation of how arguments work.

Building an Argument

Visual schematics (lines, boxes, and other elements that correspond to parts of an argument) can be a useful way to set up an argument, but I have found that sentences and fragments of sentences are actually far more flexible and more likely to make their way into actual use.

Prior facts and wisdom will typically enter into an argument without being acknowledged, so there is not much to show here at this point. As we go through the parts of an argument, examples of how each idea is produced in written language will appear here, in these sidebars.

To make the ideas more transparent, I have also marked the concepts introduced in each section with tags like *(Claim)*. In a real argument, it's rare to explicitly call out claims and evidence and so on, but it is nevertheless instructive to try while developing an argument. One more thing to note: the following example is made up. It is plausible, but unresearched. Anybody found citing it as truth down the line will be publicly shamed in a future edition of this book.

Claims

Arguments are built around claims. Before hearing an argument, there are some statements the audience would not endorse. After all the analyzing, mapping, modeling, graphing, and final presentation of the results, we think they should agree to these statements. These are the claims. Put another way, a claim is a statement that could be reasonably doubted but that we believe we can make a case for. All arguments contain one or more claims. There is often, but not necessarily, one main claim, supported by one or more subordinate claims.

Suppose that we needed to improve the safety of a neighborhood that has been beset by muggings. We analyze the times and places where muggings happen, looking for patterns. Our main claim is that police officers should patrol more at these places and times. Our subordinate claims are that there is a problem with muggings; that the lack of police at certain places and times exacerbates the problem; and that the added cost of such deployments is reasonable given the danger. The first and last points may or may not require much of an argument, depending on the audience, whereas the second will require some kind of analysis for sure.

Note that the claim is in terms that the decision makers actually care about. In this case, they care about whether the lack of police in certain places and times exacerbates muggings, not what model we built and what techniques we used to assess that model's fit. Being able to say that our model has good generalization error will end up being an important part of making a tight argument, but it functions as support, not as a big idea.

In the details justifying our claim, we could swap out another technique for assessing the quality of our model to show that we had a reasonable grasp of the patterns of muggings. Different techniques[2] make different assumptions, but some might be chosen purely for practical reasons. The techniques are important, but putting them into an argument frame makes it obvious which parts are essential and which are accidental.

Let us turn our attention back to the task of predicting how public transit affects real estate prices over time. Here is where thinking about the full problem really shines. There is no single statistical tool that is sufficient to create confidence in either a causal relationship or the knowledge that a pattern observed should continue to hold in the future. There are collections of things we can do, held together by strong arguments (such as sensitivity analysis or a paired design) that can do the job and make a case for a causal relationship. None of them are purely statistical techniques; they all require a case be made for why they are appropriate beyond how well they fit the data.

2. In this case, for example, Bayes factors for cross-validation.

Claims

(Claim) A 5% reduction in average travel time to the rest of the city results in a 10% increase in apartment prices, with the reverse true as well, an effect which persists. *(Subclaim)* We know this is true because we have looked back at historical apartment prices and transit access and seen that the effect persists. *(Subclaim)* More importantly, when there have been big shocks in transit access, like the opening of a new train stop or the closing of a bus route, there have been effects of this magnitude on apartment prices visible within a year.

Evidence, Justification, and Rebuttals

A key part of any argument is evidence. Claims do not demonstrate themselves. Evidence is the introduction of facts into an argument.

Our evidence is rarely raw data. Raw data needs to be transformed into something else, something more compact, before it can be part of an argument: a graph, a model, a sentence, a map. It is rare that we can get an audience to understand something just from lists of facts. Transformations make data intelligible, allowing raw data to be incorporated into an argument.

A transformation puts an interpretation on data by highlighting things that we take to be essential. Counting all of the sales in a month is a transformation, as is plotting a graph or fitting a statistical model of page visits against age, or making a map of every taxi pickup in a city.

Returning to our transit example, if we just wanted to show that there is *some* relationship between transit access and apartment prices, a high-resolution map of apartment prices overlaid on a transit map would be reasonable evidence, as would a two-dimensional histogram or scatterplot of the right quantities. For the bolder claims (for example, that the relationship is predictable in some way and can be forecast into the future), we need more robust evidence, like sophisticated models. These, too, are ways of transforming data into evidence to serve as part of an argument.

Evidence and Transformations

A 5% reduction in average travel time to the rest of the city results in a 10% increase in apartment prices, with the reverse true as well, an effect which persists. We know this is true because we have looked back at historical apartment prices and transit access and seen that the effect persists. *(Transformation)* This graph, based on *(Evidence)* 20 years of raw price data from the City, demonstrates how strong the relationship is.

More importantly, when there have been big shocks in transit access, like the opening of a new train stop, or the closing of a bus route, there have been effects on apartment prices visible within a year. *(Transformation, Evidence)* Average prices of apartments for each of the following five blocks, with lines indicating the addition or closure of a new train route within two blocks of that street, demonstrate the rapid change. On average, closing a train stop results in a *(Transformation, Evidence)* 10% decline in apartment prices for apartments within two blocks away over the next year.

If I claim that the moon is made of cheese and submit pictures of the moon as evidence, I have supplied two of the necessary ingredients, a claim and evidence, but am missing a third. We need some justification of *why* this evidence should compel the audience to believe our claim. We need a reason, some logical connection, to tie the evidence to the claim. The reason that connects the evidence to the claim is called the *justification*, or sometimes the *warrant*.

The simplest justification of all would be that the claim is self-evident from the evidence. If we claim that Tuesdays have the highest average sales in the past year, and our evidence is that we have tabulated sales for each day over the past year and found that Tuesday is the highest, our claim is self-evident. The simplest factual arguments are self-evident; these are basically degenerate arguments, where the claim is exactly the evidence, and the justification is extraneous.

Consider a slightly more sophisticated example: that map of home prices laid over a transit map. For this map, more expensive homes read as brighter blocks. The claim is that transit access is associated with higher prices. The evidence is the map itself, ready to be looked at. The justification might be termed visual inspection. By looking at the areas where there are highly priced homes, and seeing that transit stops are nearby, we are making a somewhat convincing argument that the two are related. Not all arguments are solid ones.

Or consider a regression model, where, for example, average prices are modeled by some function of distance to each transit line. The evidence is the model. One possible subclaim is that the model is accurate enough to be used; then the justification is cross-validation, with accuracy measured in dollars.

Another subclaim is that distance from the Lexington Avenue line has the largest effect on prices; then the justification might a single-variable validation procedure[3]. This illustrates a crucial point: the same data can be used to support a variety of claims, depending on what justifications we draw out.

There are always reasons why a justification won't hold in a particular case, even if it is sound in general. Those reasons are called the *rebuttals*. A rebuttal is the yes-but-what-if question that naturally arises in any but the most self-evident arguments.

Consider an attempt to show that a medication is effective at treating warts. If our claim is that the medication cures warts; that our evidence is a randomized controlled trial; and our justification is that randomized controlled trials are evidence of causal relationships, then common rebuttals would be that the randomization may have been done improperly, that the sample size may have been too small, or that there may be other confounding factors attached to the treatment. It pays to be highly aware of what the rebuttals to our arguments are, because those are the things we will need to answer when we need to make a water-tight argument.

In the case that we are using visual inspection to justify our claim that there is a relationship between apartment prices and transit lines, the rebuttal is that visual inspection may not be particularly clear, given that the data will be noisy. There will be highly priced places that are not near public transit lines, and places that have low prices and are on public transit lines.

For a justification of cross-validation, a rebuttal might be that the data is outdated, that the error function we chose is not relevant, or that the sample size is too small. There are always some things that render even the best techniques incorrect.

Finally, all justifications provide some degree of certainty in their conclusions, ranging from possible, to probable, to very likely, to definite. This is known as the degree of *qualification* of an argument. Deductive logic (Tim O'Reilly is a man; All men are mortal; Therefore, Tim O'Reilly is mortal) provides definite certainty in its conclusions, but its use in practice is limited. Having some sense of how strong our result is keeps us from making fools of ourselves.

3. Such as bootstrapping or a t-test.

Adding Justifications, Qualifications, and Rebuttals

A 5% reduction in average travel time to the rest of the city results in a 10% increase in apartment prices, with the reverse true as well, an effect which persists. We know this is true because we have looked back at historical apartment prices and transit access and seen that the effect persists. The graph *(Justification)* shows the predictive power of a model trained on each year's data in predicting the following year (based on 20 years of raw price data from the City), and demonstrates that a relationship is *(Qualification)* very likely.

More importantly, when there have been big shocks in transit access, like the opening of a new train station or the closing of a bus route, there were effects on apartment prices visible within a year. *(Justification)* Because the changes are so tightly coupled and happen more frequently together by chance than other changes, we can conclude that the changes in transit access are causing the drop in apartment prices. *(Qualification)* This leads us to believe in a very strong probability of a relationship.

Average prices of apartments for each of the following five blocks, with lines indicating the addition or closure of a new train route within two blocks of that street, demonstrate the rapid change. On average, closing a train stop results in a 10% decline in apartment prices for apartments within two blocks away over the next year, a relationship which roughly holds in at least *(Qualification)* 70% of cases.

(Rebuttal) There are three possible confounding factors. First, there may be nicer apartments being built closer to train lines. *(Prior knowledge)* This is clearly not at issue, because new construction or widespread renovation of apartments (which would raise their value) or letting apartments decline (which would lower their value) all take place over longer time scales than the price changes take place in. Second, there may be large price swings, even in the absence of changing transit access. *(Prior knowledge)* This is also not an issue, because the average price change between successive years for all apartments is actually only 5%. Third, it might be the case that transit improvements or reductions and changes in apartment price are both caused by some external change, like a general decline in neighborhood quality. *(Prior knowledge)* This is impossible to rule out, but generally speaking, the city in question has added transit options spottily and rent often goes up regardless of these transit changes.

Deep Dive: Improving College Graduation Rates

Another extended example is in order, this one slightly more technical. Suppose that a university is interested in starting a pilot program to offer special assistance to incoming college students who are at risk of failing out of school (the context). What it needs is a way to identify who is at risk and find effective interventions (the need). We propose to create a predictive model of failure rates and to assist in the design of an experiment to test several interventions (the vision). If we can make a case for our ideas, the administration will pay for the experiments to test it; if they are successful, the program will be scaled up soon after (the outcome).

There are several parts to our work. The first major part is to build a model to predict the chance of failure. The second is to make a case for an experiment to test if the typical interventions (for example, guidance, free books on study habits, providing free tutoring) are effective for at-risk students. Just identifying students at risk of failing out isn't enough.

When we think about building the model and designing the experiment, we need to put our findings in terms that the decision makers will understand. In this case, that's probably a well-calibrated failure probability rather than a black-box failure predictor. It also entails presenting the experiments in terms of the range of cost per student retained, or the expected lift in graduation rates, not Type I and Type II errors.[4]

In order to make the experiment easy to run, we can restrict ourselves to students who fail out after the first year. If we're lucky, the majority of failouts will happen in the first year, making the experiment even more meaningful. Performing a quick analysis of dropout rates will be enlightening, as evidence to support the short time span of our experiment (or counter-evidence to support a longer experiment, if the dropouts tend to happen at the end of a degree program).

It would be prudent to interview the decision makers and ask them what they think a reasonable range of expense would be for raising the graduate rate by 1%. By setting goals, we can assess whether our models and experiments are plausibly useful. It is important to find the current and historical graduation rate for the university, the general cost of each intervention, and to play around with those numbers. Having a general sense of what success would look like for a model will help set the bar for the remaining work. It is also important to recognize that not

4. For more information on modeling and measurement in a business context, and other in-depth discussions of topics raised throughout this book, see *Data Science for Business* by Foster Provost and Tom Fawcett (O'Reilly Media, 2013).

all students who drop out should or could have been helped enough to stay enrolled. Some admitted students were actually false positives to begin with.

To make this project useful, we need to define "failout" and differentiate a failout from a dropout, which might not be caused by poor academic performance. If the university already has a good definition of failout, we should use that. If not, we have to justify a definition. Something like: a failout is a student who drops out either when he is forced to do so by failing grades, or when he drops out on his own accord but was in the bottom quartile of grades for freshmen across the university. To justify this definition, we can argue that students who drop out on their own and had poor grades are probably dropping out in part because they performed poorly.

Now we can turn our attention to the modeling, trying to predict the probability that someone will fail out in her first year. Based on some brainstorming, we identify some potentially predictive facts about each student, plan our next steps, and collect data. If we have access to student records and applications, we might start with 10,000 records, including each student's age, high school GPA, family history, initial courseload, declared major, and so on. This step, acquiring and cleaning the data, will probably account for half or three quarters of the time we spend, so it pays to make a series of mockups along the way to show what we expect to get out of our transformations.

Consider for a moment what we will find if we get a good classifier. Suppose that about 10% of first-year undergraduates fail out. A successful classifier will take that 10% and spread it among some people with a large risk and some with a small risk, retaining the average of 10% across all students. A good classifier will be as spread out as possible while simultaneously having a high correspondence with reality. So in the ideal scenario, for example, we can identify students with 50–60% risk of failing out according to the model, for whom 50–60% do actually fail out in reality.

Our argument will be the following: *(Claim)* We have the ability to forecast which students will fail out accurately enough to try to intervene. *(Subclaim)* Our definition of failout, students who drop out either because they are asked to leave due to poor grades or who are in the bottom quartile of their freshman year, is consistent with what the administration is interested in. *(Justification for the first claim)* We can group students we did not fit the model on and see that the predictions closely match the reality. The error in our predictions is only a few percent on average, accurate enough to be acceptable. *(Justification for the subclaim)* Students

who dropped out on their own but had poor grades have essentially failed out; this idea is consistent with how the term is used in other university contexts.

Let's return to our second need, which was to design an experiment. Depending on how much money we have, how many students we have to work with, and how expensive each intervention might be, there are different experimental designs we can pursue. At a high level, we can imagine splitting students into groups by their risk level, as well as by intervention or combinations of interventions. If money is tight, splitting students into just high- and low-risk groups and choosing students at random from there would be fine. Based on the expectations laid out in discussions with decision makers, we can set minimum thresholds for detection and choose our sample sizes appropriately. And using our risk model, we can set appropriate prior probabilities for each category.

With our model accuracy in hand, we can also estimate the range of effectiveness the various interventions might have, and the associated cost. If, for example, the best our risk model could do was separate out students who were above or below average risk, "at-risk" students could be anywhere from 11% to 100% likely to fail out. With the right distribution of students, it's plausible that almost all of the money will be spent on students who were going to pass anyway. With some thought, we can derive more reasonable categories of high and low risk. With those, we can derive the rough range of cost and benefit in the best- and worst-case post-intervention outcomes.

Our argument for the second need is as follows, assuming that the math works out: *(Claim)* The experiment is worth performing. *(Subclaim)* The high and low ranges of costs for our experiment are low compared to a reasonable range of benefits. *(Subclaim)* Furthermore, their cost is fairly low overall, because we are able to target our interventions with accuracy and we can design the experiment to fit within the budget. *(Claim)* The right way to perform the experiment is to break up students into high- and low-risk groups, and then choose students from the population at random to minimize confounding factors...and so on.

In a fully worked-out project, these arguments would have actual data and graphs to provide evidence. It is also not right to expect that we will arrive at these arguments out of the blue. Like everything else, arguments are best arrived at iteratively. Early versions of arguments will have logical errors, misrepresentations, unclear passages, and so on. An effective countermeasure is to try to inhabit the mind of our friendly but skeptical audience and see where we can find holes. An even better approach is to find a real-life friendly skeptic to try to explain our argument to.

Patterns of Reasoning

One of the great benefits of studying arguments is that we can draw inspiration from patterns that have been noticed and explored by others. Instead of bushwhacking our way through the forest, we have a map to lead us to well-worn trails that take us where we need to go.

We can't simply lift up the patterns that structure arguments in other disciplines and plop them down precisely into data science. There are big differences between a courtroom, a scientific dispute, a national policy debate, and the work that we do with data in a professional setting. Instead, it is possible to take insights from patterns in other fields and mold them to fit our needs.

There are three groups of patterns we will explore. The first group of patterns are called *categories of disputes*, and provide a framework for understanding how to make a coherent argument. The next group of patterns are called *general topics*, which give general strategies for making arguments. The last group is called *special topics*, which are the strategies for making arguments specific to working with data. Causal reasoning, which is a special topic, is so important that it is covered separately in Chapter 5.

Categories of Disputes

A very powerful way to organize our thoughts is by classifying each *point of dispute* in our argument. A point of dispute is the part of an argument where the audience pushes back, the point where we actually need to make a case to win over the skeptical audience. All but the most trivial arguments make at least one point that an audience will be rightfully skeptical of. Such disputes can be classified, and the classification tells us what to do next. Once we identify the kind of dispute we are dealing with, the issues we need to demonstrate follow naturally.

Ancient rhetoricians created a classification system for disputes. It has been adapted by successive generations of rhetoricians to fit modern needs. A point of dispute will fall into one of four categories: *fact*, *definition*, *value*, and *policy*.

Once we have identified what kind of dispute we are dealing with, automatic help arrives in the form of *stock issues*. Stock issues tell us what we need to demonstrate in order to overcome the point of contention. Once we have classified what kind of thing it is that is under dispute, there are specific subclaims we can demonstrate in order to make our case. If some of the stock issues are already believed by the audience, then we can safely ignore those. Stock issues greatly simplify the process of making a coherent argument.

We can also use any of these patterns of argument in the negative. Each list of stock issues also forms a natural list of rebuttals in the event that we want to argue against a policy or particular value judgment. It is a very flexible technique.

FACT

A dispute of fact turns on what is true, or on what has occurred. Such disagreements arise when there are concrete statements that the audience is not likely to believe without an argument. Disputes of fact are often smaller components of a larger argument. A particularly complicated dispute of fact may depend on many smaller disputes of fact to make its case.

Some examples of disputes of fact: Did we have more returning customers this month than the last? Do children who use antibiotics get sick more frequently? What is the favorite color of colorblind men? Is the F1 score of this model higher than that of the other model?

The typical questions of science are disputes of fact. Does this chemical combination in this order produce that reagent? Does the debt-to-GDP ratio predict GDP growth? Does a theorized subatomic particle exist?

What all of these questions have in common is that we can outline the criteria that would convince you to agree to an answer prior to any data being collected. The steps might be simple: count the people in group A, count the people in group B, report if A has more members than B. Or it might be highly complex, involving many parts: verify that a piece of monitoring machinery is working well, correctly perform some pipette work 100 times, correctly take the output of the monitoring machine, and finally, apply a Chi-square test to check the distribution of the results. We can make a case for why meeting such conditions would imply the claim.

There are thus two stock issues for disputes of fact. They are:

- What is a reasonable truth condition?
- Is that truth condition satisfied?

In other words, how would you know this fact was right, and did it turn out the way you expected?

We need to lay out the conditions for holding a fact to be true, and then show that those conditions are satisfied. If the conditions are already obvious and held by the audience, we can skip straight to demonstrating that they are satisfied. If they aren't, then our first task is to make a case that we identified a series of steps or conditions that imply the claim.

Take a famous example, the claim that a debt-to-GDP (gross domestic product) ratio of over 90% results in negative real GDP growth. That is, if a national government had debt equal to 90% or more of its GDP in one year, then on average its GDP in the following year would fall, even adjusting for inflation. For several years, this was taken as a fact in many policy circles, based on a paper by the Harvard economists Reinhart and Rogoff.[1]

Reinhart and Rogoff stipulated the following truth condition: collect debt-to-GDP ratios for a number of years, across dozens of countries. Group each country-year into four buckets by their debt-to-GDP ratio (30%, 30–60%, 60–90%, 90% and above). Calculate the growth between that year and the next year. Average across each country, and then average all countries together. Whatever the average growth was in each bucket is the expected growth rate.

Both their truth condition and claim to satisfy that condition turned out to be flawed. When the result was examined in depth by Herndon, Ash, and Pollin from the University of Massachusetts Amherst,[2] several issues were found.

First, the truth condition was misspecified. Data is not available equally for all countries in all years, so averaging first within each country and then across all averages could weigh one year in one country equally with several decades in another. Specifically, in Reinhart and Rogoff's data, Greece and the UK each had 19 years with a debt-to-GDP ratio over 90% and growth around 2.5%, whereas New Zealand had one year with –7.9% growth. The three numbers were simply averaged.

Second, their execution turned out to be flawed. Excel errors excluded the first five countries (Australia, Austria, Belgium, Canada, and Denmark) entirely from the analysis. Additional country-year pairs were also omitted from the whole data set, the absence of which substantially distorted the result.

1. Reinhart, Carmen M., and Kenneth S. Rogoff. *Growth in a Time of Debt*. NBER Working Paper No. 15639, 2010. *http://www.nber.org/papers/w15639*.
2. Herndon, Thomas, Michael Ash, and Robert Pollin. *Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff*. PERI Working Paper No. 322, 2013. *http://bit.ly/1glDQfN*.

Herndon, Ash, and Pollin made a counter-claim. They declared that there is no sharp drop in GDP growth at a 90% debt-to-GDP ratio, and that in fact the growth slowly falls from an average of 3% per year to around 2%, in the 30% to 120% debt-to-GDP ratio range, beyond which the data volume falls out.

Their truth condition was simply a smoothed graph fit to the original data, without any bucketing. It is worth noting that Herndon, Ash, and Pollin carried out their examination in R, rather than Excel, which provided better statistical tools and far easier replication and debugging.

DEFINITION

Disputes of definition occur when there is a particular way we want to label something, and we expect that that label will be contested. Consider trying to figure out whether antibiotics reduce the incidence of sickness in kids. In the contemporary United States, children are humans under the age of 13 (or 18, in the case of the law). Antibiotics are a known class of drugs. But what does it mean to regard a child as having been sick? Viral load? Doctor's visits? Absence from school? Each of these picks up something essential, but brings its own problems. Quite a lot can be on the line for determining the right definition if you're a drug company. Definitions in a data context are about trying to make precise relationships in an imprecise world. If a definition is already precise and widely shared, there is nothing at issue and we will have nothing to defend.

Words also have prior meanings that are an important part of how we think about them. If humans thought with exact, axiomatic logic, naming things would make no difference. But, for the same reason, our ability to think would be rigidly stuck within narrowly defined categories. We would fall apart the moment we had to think through something outside our axioms.

There are two activities involving definitions that can happen in an argument. The first is making a case that a general, imprecise term fits a particular example. If we show some graphs and claim that they demonstrate that a business is "growing," we are saying that this particular business fits into the category of growing businesses. That is something we have to make a case for, but we are not necessarily claiming to have made a precise definition.

The second activity is a stronger claim to make in an argument: that we have made an existing term more precise. If we say that we have developed some rules to determine which programming language is the most popular, or which users should be considered the most influential on Twitter, we are clarifying an existing idea into something that we can count with.

"Popularity," "influence," "engagement," and the like are all loaded terms, which is a good thing. If we shy away from using meaningful language, it becomes difficult to make judgments in new scenarios. Meaningful definitions provide sensible default positions in new situations. There are certain things we expect of a "popular" language beyond it being highly used in open source code repositories, even if that is how we define popularity. And we have some mental model of how "engaged" users should behave, which can be very useful for debugging an argument.

A term that an audience has no prior understanding of, either in the form of examples (real or prototypical) or prior definitions, is not going to be contested. There will be no argument, because the word will be totally new.

There are three stock issues with disputes of definition:

- Does this definition make a meaningful distinction?
- How well does this definition fit with prior ideas?
- What, if any, are the reasonable alternatives, and why is this one better?

We can briefly summarize these as Useful, Consistent, and Best. A good definition should be all three.

First, consider the issue of whether a definition makes a difference *(Useful)*. What would you think of a definition that declared users to be influential on Twitter based on the result of a coin toss? It would add no predictive ability as to how a given tweet would be taken up by the larger network. Or a definition of a growing business that included every business that was not spiraling into bankruptcy? There have to be some useful distinctions made by the definition in order to be worthy of being cared about, and we often have to make a case for the utility of a definition.

Discussions about how well a definition fits with prior ideas can take many forms *(Consistent)*. One is: how well does this definition agree with previously accepted definitions? Another: how well does this definition capture good examples? To justify a definition of influence on Twitter, it helps to both cite existing definitions of influence (even if they aren't mathematically precise) and to pick out particular people that are known to be influential to show that they are accounted for. Our definition should capture those people. And if it does not, we should make a clear case for why their exclusion does not create a problem with fitting into precedent.

Finally, it makes sense to consider any alternatives *(Best)*, lest our audience do it for us. If there are obvious definitions or classifications that we are missing or not using, it behooves us to explain why our particular definition is superior (or why they are not in conflict).

Disputes of definition are closely related to the idea of construct validity used in the social sciences, especially in psychology and sociology. A construct is a fancy term for definition, and construct validity refers to the extent to which a definition is reliably useful for causal explanation. When psychologists define neuroticism, psychosis, or depression, they're trying to make a prior idea more precise and justify that definition to others.

Definitions are also where we typically introduce simplifying assumptions into our argument. For example, in our investigation into apartment prices and transit accessibility, we discussed sticking only to apartments that are advertised to the public. On one hand, that could be a failing of the final argument. On the other hand, it greatly simplifies the analysis, and as long as we are up front about our assumptions when a reasonable skeptic could disagree with them, it is better to have provisional knowledge than none at all.

VALUE

When we are concerned with judging something, the dispute is one of value.

For example, is a particular metric good for a business to use? We have to select our criteria of goodness, defend them, and check that they apply. A metric presents a balance of ease of interpretability, precision, predictive validity, elegance, completeness, and so on. Which of these values are the right ones to apply in this situation, and how well do they apply? At some point we may have to choose between this metric and another to guide a decision. Which is more important, customer satisfaction or customer lifetime value? We often have to justify a judgment call.

Consider the decision to pick a certain validation procedure for a statistical model. Different criteria[3] are useful for solving different problems. Which of these criteria are the right ones in a particular situation requires an argument. There are trade-offs involved between validity, interpretability, accuracy, and so on. By what criteria should our model be judged?

3. Such as cross-validation, external validation, analysis of variance, bootstrapping, t-tests, Bayesian evidence, and so on.

What else but by their fruits? For disputes of value, our two stock issues are:

- How do our goals determine which values are the most important for this argument?
- Has the value been properly applied in this situation?

For example, consider a scenario where we are deciding between two models (not validation procedures as before, but separate models), one of which is easy to interpret and another that is more accurate but hard to interpret. For some reason, we are restricted to using a single model.

Which values matter in this case will depend on what our goals are. If our goal is to develop understanding of a phenomenon as part of a longer project, the interpretable model is more important. Likewise, if our goal is to build something we can fit into our heads to reason off of in new situations, concision and elegance are important. Our next step would then be to make a case that the model in question is, in fact, interpretable, concise, elegant, and so on.

On the other hand, if our goal is to build something that is a component of a large, mostly or entirely automated process, concision and elegance are irrelevant. But are accuracy or robustness more important? That is, is it more important to be right often, or to be able to withstand change? When the US Post Office uses its handwriting recognition tools to automatically sort mail by zip code, accuracy is the most important value—handwriting is not likely to change substantially over the next few years. By contrast, when building an autonomous robot, it is more important that it can handle new scenarios than that it always walks in the most efficient way possible. Our values are dictated by our goals. Teasing out the implications of that relationship requires an argument.

POLICY

Disputes of policy occur whenever we want to answer the question, "Is this the right course of action?" or "Is this the right way of doing things?" Recognizing that a dispute is a dispute of policy can greatly simplify the process of using data to convince people of the necessity of making a change in an organization.

Should we be reaching out to paying members more often by email? Should the Parks Department do more tree trimming? Is relying on this predictive model the right way to raise revenue? Is this implementation of an encryption standard good enough to use? Does this nonprofit deserve more funding?

The four stock issues of disputes of policy are:

- Is there a problem?
- Where is credit or blame due?
- Will the proposal solve it?
- Will it be better on balance?

David Zarefsky distills these down into Ill, Blame, Cure, and Cost.[4]

Is there a problem? We need to show that there is something worth fixing. Is revenue growth not keeping up with expectation? Are there known issues with a particular algorithm? Are trees falling down during storms and killing people? In any of these cases, it is necessary to provide an argument as to why the audience should believe that there is a problem in the first place.

Where is credit or blame due? For example, is revenue not keeping up with what is expected because of weaker than normal growth in subscriptions? If the problem is that we have a seasonal product, as opposed to that our marketing emails are poorly targeted, proposing to implement a new targeting algorithm may be beside the point. We have to make the case that we have pinpointed a source of trouble.

Would our proposal solve the problem? Perhaps we have run some randomized tests, or we have compared before and after results for the same users, or we have many strong precedents for a particular action. We need to show that our proposed solution has a good chance of working.

Finally, is it worth it? There are many solutions that are too expensive, unreliable, or hard to implement that would solve a problem but aren't worth doing. If it takes three or four days to implement a new targeting model, and it will likely have clear gains, it is a no-brainer. If it might take weeks, and nobody's ever done something like this before, and the payoff is actually rather low compared to other things that a team could be doing...well, it is hard to say that it would be a good decision.

Many, many sticky problems actually turn out to be policy issues. Having a framework to think through is invaluable.

4. *Argumentation, The Study of Effective Reasoning*, 2nd ed. Audio course, http://www.thegreatcourses.com/tgc/courses/course_detail.aspx?cid=4294.

General Topics

Discussions about patterns in reasoning often center around what Aristotle called general topics. General topics are patterns of argument that he saw repeatedly applied across every field. These are the "classic" varieties of arguments: *specific-to-general*, *comparison*, *comparing things by degree*, *comparing sizes*, *considering the possible as opposed to the impossible*, etc. Undergraduate literature courses often begin and end their discussion of patterns of argument with these.

Though these arguments might seem remote from data work, in fact they occur constantly. Some, like comparing sizes of things and discussing the possible as opposed to the impossible, are straightforward and require no explanation. Others, like specific-to-general reasoning, or reasoning by analogy, require more exposition.

SPECIFIC-TO-GENERAL

A specific-to-general argument is one concerned with reasoning from examples in order to make a point about a larger pattern. The justification for such an argument is that specific examples are good examples of the whole.

A particularly data-focused idea of a specific-to-general argument would be a statistical model. We are arguing from a small number of examples that a pattern will hold for a larger set of examples. This idea comes up repeatedly throughout this book.

Specific-to-general reasoning occurs even when reasoning from anecdotes. Whenever we make productive user experience testing, we are using specific-to-general reasoning. We have observed a relatively small amount of the user base of a given product in great detail, but by reasoning that these users are good examples of the larger user base (at least in some ways), we feel comfortable drawing conclusions based on such research.

Another example of this reasoning pattern is what might be termed an illustration. An illustration is one or more examples that have been selected to build intuition for a topic. Illustrations are extremely useful early in an argument to provide a grounding to the audience about what is relevant and possible for the subject matter.

There is an element of the imagination in every argument. If someone literally cannot imagine an example or the possibilities of the thing under discussion, it is less likely that they will be swayed by the more abstract bits of reasoning. Worse, it is less likely that the argument will actually stick with the person. Practically speaking, it is not enough to convince someone. They need to stay convinced when they

walk away from your argument and not forget it moments later. Concrete examples, sticky graphics, and explanations of the prototypical will help ground arguments in ways that improve the chance of a strong takeaway, even when they are incidental to the body of the argument.

GENERAL-TO-SPECIFIC

General-to-specific arguments occur when we use beliefs about general patterns to infer results for particular examples. While it may not be true that a pattern holds for every case, it is at least plausible enough for us to draw the tentative conclusion that the pattern should hold for a particular example. That is, because a pattern generally holds for a larger category, it is plausible that it should hold for an example.

For example, it is widely believed that companies experiencing rapid revenue growth have an easy time attracting investment. If we demonstrate that a company is experiencing rapid revenue growth, it seems plausible to infer that the company will find it easy to raise money. Of course, that might not be true; the revenue growth may be short-term, or funding may be scarce. That doesn't make it improper to tentatively draw such a conclusion.

This is the opposite of the specific-to-general pattern of reasoning. Using a statistical model to make inferences about new examples is a straightforward instance of general-to-specific results arising from reversing a specific-to-general argument. We use the justification of specific-to-general reasoning to claim that a sample of something can stand in for the whole; then we use general-to-specific reasoning when we go to apply that model to a new example.

The archetypal rebuttal of a general-to-specific argument is that this particular example may not have the properties of the general pattern, and may be an outlier.

Consider a retail clothing company with a number of departments. Menswear departments may overwhelmingly be in the business of supplying men, but a fair number of women shop in menswear departments for aesthetic reasons or on behalf of their partners or children. If we argued that sales of men's dress shirts are indicative of more male customers, we are probably right, but if we argue that because a *particular* customer purchased a men's shirt that the shopper is probably male, we may be wrong.

ARGUMENT BY ANALOGY

Arguments by analogy come in two flavors: literal and figurative. In a literal analogy, two things are actually of similar types. If we have two clients with a similar

purchasing history to date, it seems reasonable to infer that after one client makes a big purchase, the other client may come soon after. The justification for argument by analogy is that if the things are alike in some ways, they will be alike in a new way under discussion.

In a figurative analogy, we have two things that are not of the same type, but we argue that they should still be alike. Traditionally, these kinds of analogies are highly abstract, like comparing justice to fire. This may seem out of place in a book on data analysis, but actually, figurative analogies are constantly present in working with data—just under a different banner.

Every mathematical model is an analogy. Just as justice is not a flame, a physical object moving in space is not an equation. No model is the same as the thing it models. No map is the territory. But behavior in one domain (math) can be helpful in understanding behavior in another domain (like the physical world, or human decision-making).

Whenever we create mathematical models as an explanation, we are making a figurative analogy. It may be a well-supported one, but it is an analogy nonetheless.

The rebuttal for argument by analogy is the same as the rebuttal for general-to-specific arguments—that what may hold for one thing does not necessarily hold for the other. Physical objects experience second-order effects that are not accounted for in the simplified physical model taken from an engineering textbook. People may behave rationally according to microeconomic models, but they might also have grudges that the models didn't account for. Still, when the analogy holds, mathematical modeling is a very powerful way to reason.

Special Arguments

Every discipline has certain argument strategies that it shares with others. The special arguments of data science overlap with those of engineering, machine learning, business intelligence, and the rest of the mathematically inclined disciplines. There are patterns of argument that occur frequently in each of these disciplines that also pop up in settings where we are using data professionally. They can be mixed and matched in a variety of ways.

Optimization, bounding cases, and cost/benefit analysis are three special arguments that deserve particular focus, but careful attention to any case study from data science or related disciplines will reveal many more.

OPTIMIZATION

An argument about optimization is an argument that we have figured out the best way to do something, given certain constraints. When we create a process for recommending movies to someone, assigning people to advertising buckets, or packing boxes full of the right gifts for customers based on their taste, we are engaging in optimization.

Of course, optimization occurs whenever we fit a model through error minimization, but in practice we rarely talk about such activities as optimizations.

This is one of the least argument-laden activities in data science…assuming we already know what we are intending to optimize. If we don't, we first have a dispute of value to deal with. The question of making a value judgment about the right thing to be optimizing is often far more interesting and controversial than the process itself.

BOUNDING CASE

Sometimes an argument is not about making a case for a specific number or model, but about determining what the highest or lowest reasonable values of something might be.

There are two major ways to make an argument about bounding cases. The first is called *sensitivity analysis*. All arguments are predicated on certain assumptions. In sensitivity analysis, we vary the assumptions to best- or worst-case values and see what the resulting answers look like. So if we are making a case that the number of new accounts for a small business could be (based on historical data) as low as two per month and as high as five, and that each account could bring in as little as $2,000 and as much as $10,000 each, then the worst case under these assumptions is a new income of $4,000 a month and a best case is as high as $50,000 a month. That is a huge range, but if we're only concerned that we have enough growth in the next month to pay off a $2,000 loan, then we're already fine.

A more sophisticated approach to determining bounding cases is through *simulation* or *statistical sensitivity analysis*. If we make assumptions about the plausibility of each value in the preceding ranges (based on historical data), say that 2 and 5 clients are equally likely but that 3 or 4 are twice as likely as 2 or 5, and that $2,000 to $10,000 are uniformly likely, then we can start to say things about the lower and upper decile of likely results. The simplest way to do that would be to simulate a thousand months, simulating first a number of clients and then a value per client.

Based on a particular level of risk, we can calculate percentiles of revenue that suit the task at hand. Cost projections for investors might be based on median revenue, whereas cost projections for cash flow analysis may use the lower decile or quartile for safety reasons. Uptime calculations for mission-critical servers, meanwhile, are only considered well-functioning if they can provide service 9,999 seconds out of every 10,000 or better. Bounding cases need to be matched to the relevant risk profile of the audience.

Sensitivity analysis or simulation can provide rough bounds on where important numbers will go. Sometimes just providing orders of magnitude is enough to push a decision forward—for example, by demonstrating that the energy needs for a construction project are a hundred times higher than what is available.

COST/BENEFIT ANALYSIS

A variant on disputes of both value and policy is a cost/benefit analysis. In a cost/benefit analysis, each possible outcome from a decision or group of decisions is put in terms of a common unit, like time, money, or lives saved. The justification is that the right decision is the one that maximizes the benefit (or minimizes the cost). In some sense, such an argument also draws on the idea of optimization, but unlike optimization, there is not necessarily an assertion that the argument takes into account all possible decisions, or a mix of decisions and constraints. A cost/benefit analysis compares some number of decisions against each other, but doesn't necessarily say anything about the space of possible decisions.

For a cost/benefit analysis, there is some agreement on the right things to value, and those things are conveniently countable. It may be quite an achievement to acquire those numbers, but once they are acquired, we can compare the proposed policy against alternatives.

Cost/benefit analyses can also be combined with bounding case analyses. If the lowest plausible benefit from one action is greater than the highest plausible benefit from another, the force of the argument is extremely strong. Even if one is just higher than the other on average, already the evidence is in its favor, but questions of the spread start to matter.

The rebuttals to a cost/benefit analysis are that costs and benefits have been miscalculated, that this is generally not the right method to make such a decision, or that the calculations for cost and benefit do not take into account other costs or benefits that reorder the answers. Next-quarter cash earnings may conflict with long-term profitability, or with legal restrictions that would land a decision maker in jail.

Causality

Causal reasoning is an important and underappreciated part of data science, in part because causal methods are rarely taught in statistics or machine learning courses. Causal arguments are extremely important, because people will interpret claims we make as having a cause-and-effect nature whether we want them to or not. It behooves us to try our best to make our analyses causal. Making such arguments correctly gives far more insight into a problem than simple correlational observations.

Causal reasoning permeates data work, but there are no tools that, on their own, generate defensible causal knowledge. Causation is a perfect example of the necessity of arguments when working with data. Establishing causation stems from making a strong argument. No tool or technique alone is sufficient.

Cause and effect permeates how human think. Causation is powerful. Whenever we think about relationships of variables, we are tempted to think of them causally, even if we know better. All arguments about relationships, even about relationships that we know are only associative and not causal, are positioned in relation to the question of causation.

Noncausal relationships are still useful. Illustration sparks our imaginations, prediction allows us to plan, extrapolation can fill in gaps in our data, and so on. Sometimes the most important thing is just to understand how an event happened or what examples of something look like. In an ideal scenario, however, our knowledge eventually allows us to intervene and improve the world. Whenever possible, we want to know what we should change to make the world be more like how we want it.

We can think about causal arguments in terms of a spectrum, running from arguments that barely hint at causality to arguments that establish undeniable causality. On the near end of the causal spectrum are arguments that are completely observational and not really causal in nature. At most, they vaguely gesture in the direction of causality. They are very useful for building understanding of what a phenomenon is about. Taxonomies, clustering, and illustrative examples stand at

this end. Again, these are not causal methods, but they can hint at causal relationships that require further investigation.

The next step in the direction of a full causal explanation is an argument claiming to have found a relationship with predictive power. A predictive relationship allows for planning, but if those variables were different, the outcome might still change as well. Suppose that we understand how likely customers are to quit their subscription to a dating site based on how quickly they reply to messages from other users. We aren't claiming that getting people to reply more quickly would change the chance that they drop off, only that people who tend to behave like that tend to drop off sooner. Most statistical models are purely predictive.

Moving further along, we have a range of methods and designs that try to eliminate alternative explanations. A *design* is a method for grouping units of analysis according to certain criteria to try to reduce uncertainty in understanding cause and effect. This is where causal influence actually begins. Techniques like *within-subject designs* and *randomized controlled trials* fall here. These methods make an effort to create defensible causal knowledge of a greater or lesser degree, depending on the number of issues stymieing causal analysis that we can eliminate. Randomized controlled trials are at the far end of this part of the causal spectrum.

On the very far end of the causal spectrum are arguments for which a causal relationship is obvious, such as with closed systems (like components in a factory) or well-established physical laws. We feel no compunction in claiming that dropping a vase is the cause of it breaking. Physical examples provide us with our intuition for what cause and effect mean. They are the paragons of the phenomenon of causation, but unfortunately, outside of some particularly intensive engineering or scientific disciplines, they rarely arise in the practical business of data work.

No matter how many caveats we may put into a report about correlation not implying causation, people will interpret arguments causally. Human beings make stories that will help them decide how to act. It is a sensible instinct. People want analysis with causal power. We can ignore their needs and hide behind the difficulty of causal analysis—or we can come to terms with the fact that causal analysis is necessary, and then figure out how to do it properly and when it truly does not apply.

To be able to make causal statements, we need to investigate causal reasoning further. Causal analysis is a deep topic, and much like statistical theory, does not have a single school that can claim to have a monopoly on methodology. Instead, there are multiple canons of thought, each concerned with slightly different scenarios. Statistics as a whole is concerned with generalizing from old to new data;

causal analysis is concerned with generalizing from old to new scenarios where we have deliberately altered something.

Generally speaking, because data science as a field is primarily concerned with generalizing knowledge only in highly specific domains (such as for one company, one service, or one type of product), it is able to sidestep many of the issues that snarl causal analysis in more scientific domains. As of today, data scientists do little work building theories intended to capture causal relationships in entirely new scenarios. For those that do, especially if their subject matter concerns human behavior, a more thorough grounding in topics such as construct validity and quasi-experimental design is highly recommended.[1]

Defining Causality

Different schools of thought have defined causality differently, but a particularly simple interpretation, suitable to many of the problems that are solvable with data, is the *alternate universe perspective*. These causes are more properly referred to as "manipulable causes," because we are concerned with understanding how things might have been if the world been manipulated a bit.

Not all causal explanations are about examining what might have been different under other circumstances. Chemistry gives causal explanations of how a chemical reaction happens. We could alter the setup by changing the chemicals or the environment, but the actual reasoning as to why that reaction will play out is fixed by the laws of physics.

Suppose there is a tree on a busy street. That particular tree may have been trimmed at some point. In causal parlance, it has either been *exposed* or *not exposed* to a *treatment*. The next year it may have fallen down or not. This is known as the *outcome*. We can imagine two universes, where in one the tree was treated with a trimming and in the other it was not (Table 5-1). Are the outcomes different?

Table 5-1. Consider one tree

	Treatment (trimming)	**Outcome (fall down)**
Universe A	True	False
Universe B	False	True

1. For more information, see *Experimental and Quasi-Experimental Designs* by William Shadish, et al. (Cengage Learning, 2001).

We can see that the trimming caused the tree to not fall down, or alternatively, that the lack of trimming prevented the tree from falling down. The trimming was not sufficient by itself (likely some wind or storm was involved as well) and it was not necessary (a crane could also have brought the tree down), but it was nevertheless a cause of the tree falling down.

In practice, we are more likely to see a collection of examples than we are a single one. It is important to remember, though, that a causal relationship is not the same as a statistical one. Consider 20 trees, half of which were trimmed. These are now all in the same universe (see Table 5-2).

Table 5-2. Looking at many trees

	Fell down	Didn't fall down
Trimmed	0	10
Not trimmed	10	0

It looks at first like we have a clear causal relationship between trimming and falling down. But what if we were to add another variable that mostly explained the difference, as shown in Table 5-3?

Table 5-3. Confounders found

Exposed to wind

	Fell down	Didn't fall down
Trimmed	0	2
Not trimmed	8	0

Not exposed to wind

	Fell down	Didn't fall down
Trimmed	0	8
Not trimmed	2	0

More of the trees that are exposed to the wind were left untrimmed. We say that exposure to the wind *confounded* our potential measured relationship between trimming and falling down.

The goal of a causal analysis is to find and account for as many confounders as possible, observed *and* unobserved. In an ideal world, we would know everything we needed to in order to pin down which states always preceded others. That

knowledge is never available to us, and so we have to avail ourselves of certain ways of grouping and measuring to do the best we can.

Designs

The purpose of causal analysis, at least for problems that don't deal with determining scientific laws, is to deal with the problem of confounders. Without the ability to see into different universes, our best bet is to come up with *designs* to structure how we perform an analysis to try to avoid confounders. Again, a design is a method for grouping units of analysis (like people, trees, pages, and so on) according to certain criteria to try to reduce uncertainty in understanding cause and effect.

A design is a method for grouping units of analysis (like people, trees, pages, and so on) according to certain criteria to try to reduce uncertainty in understanding cause and effect.

Before we stray too far away from multiple universes, one more point is in order. If we had included wind into our investigation of one single tree under different universes, we might see something interesting. After all, there does have to be to be some external agent, like a wind storm, to combine with the trimming or lack thereof to produce a downed tree. This is an illustration of *multiple causation*, where more than one thing has to be true simultaneously for an outcome to come about.

For data about one tree, reference Table 5-4.

Table 5-4. Multiple causation

	Treatment (trimming)	Wind storm	Outcome (fall down)
Universe A1	True	True	False
Universe A2	True	False	False
Universe B1	False	True	True
Universe B2	False	False	False

Typically, our goal in investigating causation is not so much to understand how one particular instance of causation played out (though that is a common pattern in data-driven storytelling), but to understand how we would expect a treatment to change an outcome in a new scenario. That is, we are interested in how well a causal explanation generalizes to new situations.

Intervention Designs

A number of different design strategies exist for controlling for confounders. In a typical randomized controlled trial, we randomly assign the treatment to the subjects. The justification for the claim that we can infer generalizability is that, because the treatment was chosen randomly for each unit, it should be independent of potential confounders.

There are numerous rebuttals for randomized controlled trials, the details of which would take up many books. The randomization may have been done improperly. The proposed treatment might not be the only difference between groups (hence double-bind experiments). The sample size may not have been large enough to deal with natural levels of variation. Crucially, new subjects may be different in novel ways from those included in the experiment, precluding generalization. If the same people are included in multiple experiments over time, they are no longer necessarily similar to a brand-new person.

In the case of *nonrandom intervention*, consider observing the stress level of someone for an hour and applying an electrical shock at the 30-minute mark without warning them. It will show a clear evidence of a relationship, but randomization was nowhere to be found. This is an example of a *within-subject interrupted time series design*; see Figure 5-1.

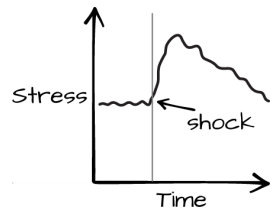


Figure 5-1. A shocking example of within-subject design

The causal relationship between shock and stress is justified on a number of levels. First, if we were carefully observing the subject and didn't see any other things that would have caused stress, such as a bear entering the laboratory, we have eliminated a large class of potential confounders. Second, we have prior intuition that shocks and stress should be related, so this result seems, prima facie, plausible. Third, the exact timing would be unreasonable to line up with another cause of stress, such as an unpleasant thought that lingered with the subject for the rest of the experiment. Finally, the magnitude of the jump was much larger than variation observed for the same subject prior to the shock.

Were we to perform this same experiment a dozen times, and each time see a similar jump in stress levels, then it would be reasonable to conclude that, for new subjects similar to the ones that we tested on, an unexpected shock would cause a jump in stress levels. That is, we could argue that the relationship is, to some extent, generalizable.

There may still be unmeasured confounders (if all of the experiment subjects were college students, they may have a lower pain threshold than, say, a construction worker) that would stymie generalization. But it is very important to note again that such unmeasured confounders can exist even in a randomized experiment. There may be important, unmeasured aspects of the subjects that can't be accounted for because all subjects are drawn from the same pool.

A truly randomized experiment would be randomized not only in the treatment, but also in the subjects, and in the pool of potential people or tools that would be carrying out an intervention. That is, we would be picking at random from the potential units (i.e., selecting people from the population at random) and executors of the treatment (i.e., if the treatment is applying a drug, choosing random doctors) in addition to applying the treatment at random. Random subjects and treaters are very rarely applied designs.

Observational Designs

In the case of purely *observational designs*, we are tasked with employing clever strategies for inferring generalizablity without the ability to intervene.

Observational designs are necessary when it is either impractical, costly, or unethical to perform interventional experiments. Most socially interesting causal estimation problems are of this kind. They frequently occur in business too; it may be advantageous for certainty's sake to take down a part of a business to see how it affects profit, but the downstream implications are almost always too great to consider doing so. Instead, we have to take advantage of natural differences in exposure or temporary outages to determine the causal effect of critical components.

Natural Experiments

Our best hope in setting up observational designs is to try to find natural experiments. Consider an effort to determine how being convicted of a crime in high school affects earnings from age 25 to 40. It isn't enough to just compare earning differences between those who were convicted and those who weren't, because clearly there will be strong confounding variables. A clever strategy would be to compare earnings between those who were arrested but not convicted of a crime,

and those who were convicted. Though there are obvious differences in the groups, they are likely to be similar on a number of other variables that serve as a natural control. This is what is known as a natural experiment: we try to argue that we have made a choice about what to compare that accounts for most alternative explanations.

When natural between-subject controls are not available, we can sometimes use methods such as *within-subject interrupted time series designs*, analogous to the electrical shock example. If sign-ups on a website are rising steadily, and when we introduce a new design for the sign-up page, sign-ups begin to drop immediately, it seems reasonable to assume that the redesign is at fault. But to claim so in good faith (and to make a strong argument), we have to eliminate other contenders. Did the site get slower? Is there a holiday somewhere? Did a competitor get a lot of coverage? By looking at bar charts and maps comparing the sign-ups for different groups of people in different regions, and confirming that the drop is fairly universal, we can strongly justify the redesign as the culprit. We are looking at the same subject and trying to account for alternatives.

The same methodology can also deal with fuzzier problems where we have no natural control group. If we are interested in whether changes in nutrition standards at schools improve test scores, we can go find test score data for many schools where we know that nutrition overhauls were put into place. If test scores (or test score ranking within a given region, to account for changes in the tests themselves) tend to be relatively flat before a nutrition change, then rise, and we have enough examples of this, it starts to be reasonable to implicate the nutrition changes. The justification for the claim of a causal relationship is that there are few reasonable alternative explanations, especially given the proximity in timing, so a causal relationship is the most logical explanation.

What we would have to do to make a good-faith effort is figure out all of the reasonable confounding factors and do our best to account for them. Our conclusion will necessarily be of a less definite kind than if we had somehow been able to randomly give individual kids more nutritious lunches—but such experiments would be not only unethical, but also self-defeating. How would you possibly enforce the restrictions of the experiment without unduly influencing the rest of the environment?

By comparison, just calculating the current test scores of schools that have booted nutrition standards against schools that have not would not be a particularly effective design, because there are likely to be a great number of confounding factors.

This, generally, paints the way for how we do causal reasoning in the absence of the ability to set interactions. We try to gather as much information as possible to find highly similar situations, some of which have experienced a treatment and some of which have not, in order to try to make a statement about the effect of the treatment on the outcome. Sometimes there are confounding factors that we can tease out with better data collection, such as demographic information, detailed behavioral studies, or pre- or post-intervention surveys. These methods can be harder to scale, but when they're appropriate, they provide us with much stronger tools to reason about the world than we are given otherwise.

Statistical Methods

If all else fails, we can turn to a number of statistical methods for establishing causal relationships. They can be roughly broken down into those based on causal graphs and those based on matching. If there have been enough natural experiments in a large data set, we can use statistical tools to tease out whether changes in some variables appear to be causally connected to others.

The topic of causal graphs is beyond the scope of this book, but the rough idea is that, by assuming a plausible series of relationships that would provide a causal explanation, we can identify what kinds of relationships we should *not* see. For example, you should never see a correlation between patient age and their treatment group in a randomized clinical trial. Because the assignment was random, group and age should be uncorrelated. In general, given a plausible causal explanation and a favorable pattern of correlations and absences of correlation, we have at least some measure of support for our argument of causality.[2]

The other variety of statistical causal estimation is matching. Of matching, there are two kinds: deterministic and probabilistic. In deterministic matching, we try to find similar units across some number of variables. Say we are interested in the effect drinking has on male fertility. We survey 1,000 men on their history and background, and measure their sperm count. Simply checking alcohol consumption history and comparing light and heavy drinkers in their sperm count is not sufficient. There will be other confounding factors like age, smoking history, and diet. If there are a small number of variables and a large number of subjects, we can reduce some confounding by finding *pairs* of men who match along many or all of the variables, but wherein only one of the two is a heavy drinker. Then we can

2. For more information on this topic, please see Judea Pearl's book *Causality* (Cambridge University Press, 2009).

compare the sperm count of the two men—and hopefully, if we have measured the right controls, it will be as if we had discovered a natural experiment.

If there are a large number of variables (say that diet is determined by a 100-question questionnaire, or we are introducing genetic data), it is more reasonable to use probabilistic matching. The most famous probabilistic matching methodology is *propensity score matching*. In propensity score matching, we build a model that tries to account for the probability that a subject will have for being treated, also called the propensity. In the alcohol example, we would model the probability of being a heavy drinker given age, smoking history, diet, genetics, and so on. Then, like in the deterministic matching example, we would again pair up similar subjects (this time, those who had roughly the same probability of becoming heavy drinkers) wherein one was a heavy drinker and one was not. We are seeking to create what might be termed an artificial natural experiment.

There are good theoretical reasons to prefer propensity score matching even in the case of a small number of variables, but it can sometimes be worthwhile to skip the added difficulty of fitting the intermediate model.

Putting It All Together

We should look at some extended examples to see the method of full problem thinking in action. By looking at the scoping process, the structure of the arguments, and some of the exploratory steps (as well as the wobbles inevitably encountered), we can bring together the ideas we have discussed into a coherent whole.

The goal of this chapter is not to try to use everything in every case, but instead to use these techniques to help structure our thoughts and give us room to think through each part of a problem. These examples are composites, lightly based on real projects.

Deep Dive: Predictive Model for Conversion Probability

Consider a consumer product company that provides a service that is free for the first 30 days. Its business model is to provide such a useful service that after 30 days, as many users will sign up for the continued service as possible.

To bring potential customers in to try its product, the company runs targeted advertisements online. These ads are focused on groups defined by age, gender, interests, and other factors. It runs a variety of ads, with different ad copy and images, and is already optimizing ads based on who tends to click them, with more money going toward ads with a higher click rate. Unfortunately, it takes 30 days or so to see whether a new ad has borne fruit. In the meantime, the company is spending very large amounts of money on those ads, many of which may have been pointlessly displayed. The company is interested in shrinking this feedback loop, and so asks a data scientist to find a way to shrink it. What can we suggest?

First, let us think a bit about what actions the company can take. It constantly has new users coming in, and after some amount of time, it can evaluate the quality of the users it has been given and choose whether to pull the plug on the advertisement. It needs to be able to judge quality sooner, and compare that quality to the cost of running the ad.

Another way of phrasing this is that the company needs to know the quality of a user based on information gathered in just the first few days after a user has started using the service. We can imagine some kind of black box that takes in user behavior and demographic information from the first few days and spits out a quality metric.

For the next step, we can start to ask what kind of behavior and what kind of quality metric would be appropriate. We explore and build experience to get intuition. Suppose that by either clicking around, talking to the decision makers, or already being familiar with the service, we find that there are a dozen or so actions that a user can take with this service. We can clearly count those, and break them down by time or platform. This is a reasonable first stab at behavior.

What about a quality metric? We are interested in how many of the users will convert to paid customers, so if possible, we should go directly for a probability of conversion. But recall that the action the company can take is to decide whether to pull the plug on an advertisement, so what we are actually interested in is the expected value of each new user, a combination of the probability of conversion and the lifetime value of a new conversion. Then we can make a cost/benefit decision about whether to keep the ad. In all, we are looking to build a predictive model of some kind, taking in data about behavior and demographics and putting out a dollar figure. Then, we need to compare that dollar figure against the cost of running the ad in the first place.

What will happen after we put the model out? The company will need to evaluate users either once or periodically between 1 and 30 days to judge the value of each user, and then will need some way to compare that value information to the cost of running the advertisement. It will need a pipeline for calculating the cost of each ad per person that the ad is shown to. Typical decisions would be to continue running an advertisement, to stop running one, or to ramp up spending on one that is performing exceptionally well.

It is also important to measure the causal efficacy of the model on improving revenue. We would like to ensure that the advertisements that are being targeted to be cut actually deserve it. By selecting some advertisements at random to be spared from cutting, we can check back in 30 days or so to see how accurately we have predicted the conversion to paid users. If the model is accurate, the conversion probabilities should be roughly similar to what was predicted, and the short-term or estimated lifetime value should be similar as well.

Context

A consumer product company with a free-to-try model. It wants people to pay to continue to use its product after the free trial.

Need

The company runs a number of tightly targeted ads, but it is not clear until around 30 days in whether the ads are successful. In the meantime, it's been spending tons of money to run ads that might have been pointless. How can it tighten up the feedback loop and decide which ads to cut?

Vision

We will make a predictive model based on behavior and demographics that uses information available in the first few days to predict the lifetime value of each incoming user. Its output would be something like, "This user is 50% less likely than baseline to convert to being a paid user. This user is 10% more likely to convert to being a paid user. This user....etc. In aggregate, all thousand users are 5% less likely than baseline to convert. Therefore, it would make sense to end this advertisement campaign early, because it is not attracting the right people."

Outcome

Deliver the model to the engineers, ensuring that they understand it. Put into place a pipeline for aggregating the cost of running each advertisement. After engineering has implemented the model, check back once after five days to see if the proportions of different predicted groups match those from the analysis. Select some advertisements to not be disrupted, and check back in one month to see if the predicted percentages or dollar values align with those of the model.

What is the argument here? It is a policy argument. The main claim is that the model should be used to predict the quality of advertisements after only a few days of running them. The "Ill" is that it takes 30 days to get an answer about the quality of an ad. The "Blame" is that installation probability (remember that we were already tracking this) is not a sufficient predictor of conversion probability. The "Cure" is a cost-predictive model and a way of calculating the cost of running a particular advertisement. And the "Cost" (or rather, the benefit) is that, by cutting out advertisements at five days, we will not spend 25 days worth of money on unhelpful advertisements.

To demonstrate that the Cure is likely to be as we say it is, we need to provisionally check the quality of the model against held-out data. In the longer term, we want to see the quality of the model for advertisements that are left to run. In

this particular case, the normal model quality checks (ROC curves, precision and recall) are poorly suited for models that have only 1–2% positive rates. Instead, we have to turn to an empirical/predicted probability plot (Figure 6-1).

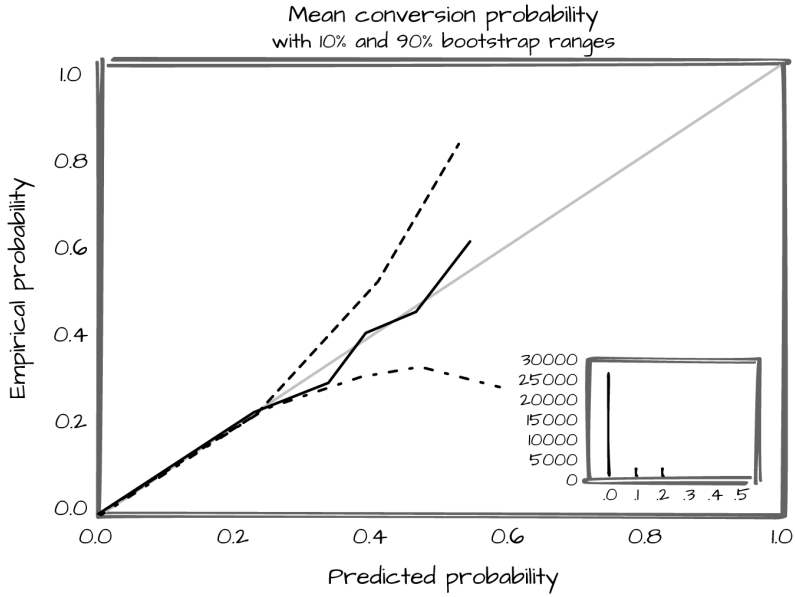


Figure 6-1. Predicted probability plot

To demonstrate the Cost, we need some sense of the reliability of the model compared to the cost range of running the ads. How does our predicted lifetime value compare to the genuine lifetime value, and how often will we overshoot or undershoot? Finally, is the volume of money saved still positive when we include the time cost of developing the model, implementing it, and running it? If the model is any good, the answer is almost certainly yes, especially if we can get a high-quality answer in the first few days. The more automated this process is, the more time it will take up front—but the more time it will save in the long run. With even reasonable investment, it should save far more than is spent.

In the end, what is the audience (in this case, the decision makers who will decide whether to proceed with this project and whether to approve the build-out) actually going to dispute? The Ill, Blame, and Cost may already be apparent, so the discussion may center on the Cure (how good is the model?). But if we were unaware of the possibility that there could be other things to discuss (besides the quality of the model), it would be easy to be caught unaware and not be prepared

to coherently explain the project when pointed questions are asked by, for example, higher levels of management.

Deep Dive: Calculating Access to Microfinance

Microfinance is the provision of traditional bank services (loans, lines of credit, savings accounts, investments) to poor populations. These populations have much smaller quantities of money than typical bank customers. The most common form of microfinance is microloans, where small loans are provided as startup capital for a business. In poorer countries, the average microloan size is under $500. Most microloan recipients are women, and in countries with well-run microfinance sectors, the vast majority of loans are repaid (the most widely admired microfinance programs average over 97% repayment).

There is a nonprofit that focuses on tracking microfinance around the world. It has a relationship with the government of South Africa, which is interested in learning how access to microfinance varies throughout their country. At the same time, the nonprofit is interested in how contemporary tools could be brought to bear to answer questions like this.

From talking to the organization, it is clear that the final outcome will be some kind of report or visualization that will be delivered to the South African government, potentially on a regular basis. Having some summary information would also be ideal.

Context

There has been an explosion of access to credit in poor countries in the past generation. There is a nonprofit that tracks information about microfinance across the world and advises governments on how they can improve their microfinance offerings.

Needs

The South African government is interested in where there are gaps in microloan coverage. The nonprofit is interested in how new data sets can be brought to bear on answering questions like this.

Vision

We will create a map that demonstrates where access is lacking, which could be used to track success and drive policy. It will include one or more summary statistics that could more concisely demonstrate success. There would be bright spots around remote areas that were heavily populated. Readers of the map should be able to conclude where the highest priority places are, in order to

place microfinance offices (assuming they were familiar with or were given access to a map displaying areas of high poverty in South Africa).

Outcome

Deliver the maps to the nonprofit, which will take them to the South African government. Potentially work with the South African government to receive regularly updated maps and statistics.

Some immediate challenges present themselves. What does access mean? If a loan office is more than a half-day's journey away, it will be difficult for a lendee to take advantage of the service. Walking in rural areas for several hours probably progresses at around 3 kilometers per hour (about 1.86 miles per hour). If we figure that three or four hours is a reasonable maximum distance for a walk in each direction, we get about 10 kilometers as a good maximum distance for access to a microfinance office.

What do we mean when we say microfinance offices? In this particular case, the microfinance tracking organization has already collected information on all of the registered microfinance offices across South Africa. These include private groups, post office branches, and nonprofit microfinance offices. For each of these, we start with an address; it will be necessary to geocode them into latitude and longitude pairs.

What about population? A little digging online reveals that there are approximate population maps available for South Africa (using a 1 km scale). They are derived from small-scale census information. Without these maps, the overall project would be much more difficult—we would need to define access relative to entire populated areas (like a town or village) that we had population and location information from. This would add a tremendous amount of overhead to the project, so thankfully such maps can easily be acquired. But keep in mind that their degree of trustworthiness, especially at the lowest scale, is suspect, and any work we do should acknowledge that fact.

We are also faced with some choices about what to include on such a map. In practice, only a single quantity can be mapped with color on a given map. Is it more important to show gradations in access or the number of people without access? Would some hybrid of people-kilometers be a valid metric? After some consideration, demonstrating the number of people is the smarter decision. It makes prioritization simpler.

The overall argument is as follows. We claim that "has access to microfinance" can be reasonably calculated by seeing, for each square kilometer, whether that

square kilometer is within 10 kilometers of a microloan office as the crow flies. This is a claim of definition. To justify it, we need to relate it to the understanding about access and microfinance already shared by the audience. It is reasonable to restrict "access" to mean foot access at worst, given the level of economic development of the loan recipients. Using the list of microfinance institutions kept by the microfinance tracking nonprofit is also reasonable, given that they will be the ones initially using this map and that they have spent years perfecting the data set.

This definition is superior to the alternative of showing degrees of access, because there is not much difference between a day's round-trip travel and a half-day's round-trip travel. Only a much smaller travel time, such as an hour or so, would be a major improvement over a day's round-trip travel. However, such density is not achievable at present, nor is it going to provide a major discontinuity from mere half-day accessibility. As such, for our purposes, 10 kilometer distance to a microloan office is a sufficient cutoff.

We claim that a map of South Africa, colored by population, masked to only those areas outside of 10 kilometers distance to a microloan office, is a good visual metric of access. This is a claim of value. The possible competing criteria are legibility, actionability, concision, and accuracy. A colored map is actionable; by encouraging more locations to open where the map is currently bright (and thus more people are deprived of access to credit), the intensity of the map will go down. It is a bit less legible than it is actionable, because it requires some expertise to interpret. It is fairly accurate, because we are smoothing down issues like actual travel distance by using bird's-eye distance, but is otherwise reasonably reliable on a small scale. It is also a concise way to demonstrate accessibility, though not as concise as per-province summaries or, at a smaller level of organization (trade-off of accuracy for concision!), per-district and per-metropolitan area summaries.

To remedy the last issue, we can join our map with some summary statistics. Per-area summary statistics, like a per-district or per-metropolitan percentage of population that is within 10 kilometers of a microloan office, would be concise and actionable and a good complement to the maps. To achieve this, we need district-level administrative boundaries and some way to mash those boundaries up with the population and office location maps.

With this preliminary argument in mind, we can chat with the decision makers to ensure that what we are planning to do will be useful. A quick mockup drawing, perhaps shading in areas on a printout of a map of South Africa, could be a useful focal point. If this makes sense to everyone, more serious work can begin.

From a scaffolding perspective, it pays to start by geocoding the microloan offices, because without that information we will have to fall back on a completely different notion of access (such as one based on town-to-town distances). It pays to plot the geocoded microloan offices on a map alongside the population density map to get a sense of what a reasonable final map will look like. It is probably wise to work out the logic for assigning kilometer squares to the nearest microloan office, and foolish to use any technique other than brute force, given the small number of offices and the lack of time constraints on map generation.

After much transformation and alignment, we have something useful. At this point the map itself can be generated, and shared in a draft form with some of the decision makers. If everyone is still on the same page, then the next priority should be calculating the summary statistics and checking those again with the substantive experts. At this point, generating a more readable map (including appropriate boundaries and cities to make it interpretable) is wise, as is either plotting the summary statistics on a choropleth map or arranging them into tables separated by district.

Final copies in hand, we can talk again with the decision makers, this time with one or more documents that lay out the relevant points in detail. Even if our work is in the form of a presentation, if the work is genuinely important, there should be a written record of the important decisions that went into making the map and summary statistics. If the work is more exploratory and temporary, a verbal exchange or brief email exchange is fine—but if people will be making actual decisions based on the work we have done, it is vitally important to leave behind a comprehensive written record. Edward Tufte has written eloquently about how a lack of genuine technical reports, eclipsed instead by endless PowerPoints, was a strong contributing factor to the destruction of the space shuttle Columbia.

Wrapping Up

Data science, as a field, is overly concerned with the technical tools for executing problems and not nearly concerned enough with asking the right questions. It is very tempting, given how pleasurable it can be to lose oneself in data science work, to just grab the first or most interesting data set and go to town. Other disciplines have successfully built up techniques for asking good questions and ensuring that, once started, work continues on a productive path. We have much to gain from adapting their techniques to our field.

We covered a variety of techniques appropriate to working professionally with data. The two main groups were techniques for telling a good story about a project, and techniques for making sure that we are making good points with our data.

The first involved the scoping process. We looked at the context, need, vision, and outcome (or CoNVO) of a project. We discussed the usefulness of brief mock-ups and argument sketches. Next, we looked at additional steps for refining the questions we are asking, such as planning out the scaffolding for our project and engaging in rapid exploration in a variety of ways. What each of these ideas have in common is that they are techniques designed to keep us focused on two goals that are in constant tension and yet mutually support each other: diving deep into figuring out what our goals are and getting lost in the process of working with data.

Next, we looked at techniques for structuring arguments. Arguments are a powerful theme in working with data, because we make them all the time whether we are aware of them or not. Data science is the application of math and computers to solve problems of knowledge creation; and to create knowledge, we have to show how what is already known and what is already plausible can be marshaled to make new ideas believable.

We looked at the main components of arguments: the audience, prior beliefs, claims, justifications, and so on. Each of these helps us to clarify and improve the process of making arguments. We explored how explicitly writing down arguments can be a very powerful way to explore ideas. We looked at how techniques of transformation turn data into evidence that can serve to make a point.

We next explored varieties of arguments that are common across data science. We looked at classifying the nature of a dispute (fact, definition, value, and policy) and how each of those disputes can be addressed with the right claims. We also looked at specific argument strategies that are used across all of the data-focused disciplines, such as optimization, cost/benefit analysis, and casual reasoning. We looked at causal reasoning in depth, which is fitting given its prominent place in data science. We looked at how causal arguments are made and what some of the techniques are for doing so, such as randomization and within-subject studies. Finally, we explored some more in-depth examples.

Data science is an evolving discipline. But hopefully in several years, this material will seem obvious to every practitioner, and a clear place to start for every beginner.

Further Reading

Paul, Richard and Linda Elder. *The Miniature Guide to Critical Thinking*. Foundation for Critical Thinking, 2009.

A brief introduction to structures for thinking.

Wright, Larry. *Critical Thinking: An Introduction to Analytical Reading and Reasoning*. 2nd ed. Oxford University Press, 2012.

Readable, useful textbook on finding the essence of arguments.

Papert, Seymour. *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books, 1993.

A classic on how mental models open up the possibility of understanding new ideas.

Jones, Morgan D. *The Thinker's Toolkit: 14 Powerful Techniques for Problem Solving*. Crown Business, 1998.

A compendium of brainstorming and decision structuring techniques.

Moore, David T. *Critical Thinking and Intelligence Analysis*. CreateSpace, 2007.

Applications of argument and critical thinking with data in a wide-ranging and adversarial situation: national intelligence.

Toulmin, Stephen E. *The Uses of Argument*. Cambridge University Press, 2003.

Philosophical treatment of the foundations of argumentation.

Croll, Alistair and Benjamin Yoskovitz. *Lean Analytics*. O'Reilly, 2013.

In-depth guide to choosing the right metrics for a given organization at a given time.

Hubbard, Douglas W. *How to Measure Anything: Finding the Value of Intangibles in Business,*. Wiley, 2010.

Guide to measuring and acting on anything, including "intangibles" like security, knowledge, and employee satisfaction.

Provost, Foster and Tom Fawcett. *Data Science for Business*. O'Reilly Media, 2013.
In-depth look at many of the same topics in this book, with a greater focus on the high-level technical ideas.

Tufte, Edward. *Envisioning Information*. Graphics Press, 1990.
A classic in structuring visual thinking for both exploration and communication.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning, 2001.
Very readable textbook on causal designs.

Jaynes, E.T., and G. Larry Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
A book about the connection between classical logic and probability theory.

About the Author

Max Shron runs a small data strategy consultancy in New York, working with many organizations to help them get the most out of their data. His analyses of transit, public health, and housing markets have been featured in the *New York Times*, *Chicago Tribune*, Huffington Post, WNYC, and more. Prior to becoming a data strategy consultant, he was the data scientist for OkCupid.

Colophon

The cover font is BentonSans Compressed, the body font is ScalaPro, the heading font is BentonSans, and the code font is TheSansMonoCd.