Chapter 1. Introduction

Let's begin with the most important question: why should you read this book? The answer is simple: you want more value from your data. To put a little more meat on that statement, our objective in writing this book is to help the variety of people who manage the analysis or application of data in their organizations. The data might or might not be "yours," in the strict sense of ownership. But the pains in extracting value from this data are.

We're focused on two kinds of readers. First are people who manage the analysis and application of data indirectly—the managers of teams or directors of data projects. Second are people who work with data directly—the analysts, engineers, architects, statisticians, and scientists.

If you're reading this book, you're interested in extracting value from data. We can categorize this value into two types along a temporal dimension: near-term value and long-term value. In the near term, you likely have a sizable list of questions that you want to answer using your data. Some of these questions might be vague; for example, "Are people really shifting toward interacting with us through their mobile devices?" Other questions might be more specific: "When will our customers' interactions primarily originate from mobile devices instead of from desktops or laptops?"

What is stopping you from answering these questions? The most common answer we hear is "time." You know the questions, you know how to answer them, but you just don't have enough hours in the day to wrangle your data into the right form.

Beyond the list of known questions related to the near-term value of your data is the optimism that your data has greater potential long-term value. Can you use it to forecast important seasonal changes? What about risks in your supply chain due to weather or geopolitical shifts? Can you understand how the move to mobile is affecting your customers' purchasing patterns? Organizations generally hire data scientists to take on these longer-term, exploratory analyses. But even if you have the requisite skills to tackle these kinds of analyses, you might still struggle to be allocated sufficient time and resources. After all,

exploratory analytics projects can take months, and often contain a nontrivial risk of producing primarily negative or ambiguous results.

As we've seen, the primary impediment to realizing both the short-term and long-term value of your data is time: your limited time and your organization's limited time. In this book, we describe how improving your data wrangling efforts can create the time required to get more near-term and long-term value from your data. In Chapters 1-3, we describe a workflow framework that links activities focused on both kinds of value, and explain how data wrangling factors into those activities and into the overall workflow framework. We introduce the basic building blocks for a data wrangling project: data flow, data wrangling activities, roles, and responsibilities. These are all elements that you will want to consider, at a high level, when embarking on a project that involves data wrangling. Our goal is to provide some helpful guidance and tips on how to coordinate your data wrangling efforts, both across multiple projects by making sure your wrangling efforts are constructive as opposed to redundant or conflicting, and within a single project by taking advantage of some standard language and operations to increase productivity and consistency.

There's more to effective data wrangling than just clearly defined workflows and processes; to most effectively wrangle your data, you should also understand which transformation actions constitute data wrangling, and, most important, how you can use those transformations to produce the best datasets for your analytic activities.

Those nitty-gritty transformations constitute our discussion in Chapters 4-7. You can think of those chapters as a rough "how-to" guide for data wrangling. That said, we do not intend this book to provide a comprehensive tutorial on all possible data wrangling methods. Instead, we want to give you a collection of techniques that you can use when moving through the stages of the data workflow framework.

As we introduce each of the key transformation and profiling activities that comprise data wrangling, we will walk through a theoretical data project involving a publicly available dataset containing US campaign finance information. You can walk through the project along with us in your data wrangling tool of choice.

Finally, we end by discussing roles and responsibilities in a data wrangling project in Chapter 8, and exploring a selection of data wrangling tools in Chapter 9.

Throughout the book, we ground our discussion in example data, transformations of that data, and various visual and statistical views of that data. Along those lines, we open with a story about Facebook.

Magic Thresholds, PYMK, and User Growth at Facebook

Growth is about tapping and delivering value to the yet unserved part of your market. Facebook stands as a quintessential example of how to drive growth. Toward the end of 2015, Facebook reported more than one billion daily active users with a year-over-year growth around 17 percent.[1] There are, of course, many factors that have contributed to this growth. We'll focus here on a series of data-driven insights that armed Facebook with strategies to deliver robust growth, year over year over year.

Growth is ultimately about increasing the number of actively engaged users and customers. It follows a simple equation:[2]

active users = new users + returning users + resurrected users

A critical aspect of growth is bringing new users and customers to your product or service. But just as critical is delivering value to new users so that they stay engaged. Ideally, users are "returning" (i.e., active from one period to the next). However, depending on how you are tracking engagement, you might see blips of inactivity followed by reengagement (placing these users in the "resurrected" group in the aforementioned equation). We'll focus on this second critical aspect of growth—delivering value to new users quickly so that they are motivated to stay engaged.

As Alex Schultz, vice president of growth at Facebook, points out, the primary value for Facebook users revolves around connecting people to the content from their friends.[3] Obviously, for this to work, users need friends on Facebook. But is this the only thing that matters—any content

from any friend? Common sense would tell you that this can't be true, and that people engage with some content more than other content. So here we have a set of near-term questions to answer:

- How many friends does a new Facebook user need to be *X*-percent likely to return as a user in 30 days? In 60 days? In 180 days?
- For new users, what characteristics of their friends stand out to differentiate between new users who churn (leave the platform and don't come back) versus those who remain active?
- Do the preceding findings change by user cohort (groups of users that initially joined Facebook at around the same time)?

Answering questions like these is the purview of the Growth and Analytics team at Facebook. Interestingly, the team found a magic threshold that captured a key predictor of long-term user engagement: new users should connect to 10 friends within 14 days. Magic thresholds have two key characteristics: first, they should correspond to a concise Key Performance Indicator (KPI) target that predicts (and if you are lucky, drives) the impact you want; and second, they should be actionable. KPI targets are standard across industries and departments, but what sets a magic threshold apart is that it exposes the core dynamic of the system and provides a lever for achieving a desired outcome. In the case of Facebook, connecting to friends quickly is a critical driver of value for new users, and if Facebook can find ways to reach that threshold for more new users, more new users should stay engaged over the long term.

This magic threshold has the advantage of encoding the core value proposition of Facebook: users connecting to their friends. It also has the advantage of coordinating a number of product decisions to help satisfy this threshold for as many new users as possible.

So, how does Facebook find friends for new users? There are simple, manual mechanisms that allow new users to import their email contact lists (which Facebook then triangulates with its known list of users). This provides short-term value. Facebook also utilizes more sophisticated mechanisms to link users to friends. We consider these mechanisms to fall into the realm of long-term value, in part because the depth of analyses and experimentation that are required to robustly

expose this value take months to years. But more importantly, these in-depth analyses give rise to data-driven services that automatically perform the desired operations.

In Facebook's case, one of the core systems used to drive growth, by helping new users connect to friends within Facebook, is known as PYMK, or People You May Know. PYMK is a recommender system, not unlike Amazon's product recommendation system or Netflix's movie/show recommendation system. It employs a well-known and often-used user experience rule: recognition is better than recall. In other words, it's easier and more enjoyable for users to say "yes" or "no" to a series of suggestions than it is for them to generate the content of the suggestions through search or a menu-driven builder experience.

PYMK uses a number of features about the new users and, more important, about the first few friends to whom they have connected. In its most basic form, you can think of PYMK as collecting all the friends of a user's friends to whom they are not currently connected. Then, based on metrics like the number of mutual friends, age similarity, education similarity, and so on, it ranks this list and presents it back to the user as recommendations.

So, with a little bootstrapping from an important contact list or a few manual friend searches, new users on Facebook begin receiving recommendations on who to connect with. The PYMK system that enables these connections has been critical to Facebook's continuous growth.

But the story becomes even more interesting. After some long-running analyses and experimentation, Facebook found that a more effective use of PYMK for user growth was not to focus on recommendations for new users (because bootstrapping is difficult and the early recommendations can come with low-confidence scores), but rather to focus on recommendations to heavy, long-time users of Facebook with vast and diverse connections. Specifically, the key is to recommend new users to the heavy Facebook users. This primes a new user with all sorts of interesting content and the friend network of the heavy user can provide better estimates on friend recommendations directed to the new user.

Although certainly unique in many ways, Facebook's use of data stands as a repeatable process that many other organizations can follow. Starting with a clear motivation—driving user growth—a number of explicit, near-term questions can provide critical insights to improve the business. Over the long term, these insights can blossom into data services that automate and optimize the earlier insights for deeper and additional value.

In Chapter 2, we describe our workflow framework that links near-term and long-term value from data with the variety of activities involved in working with data.