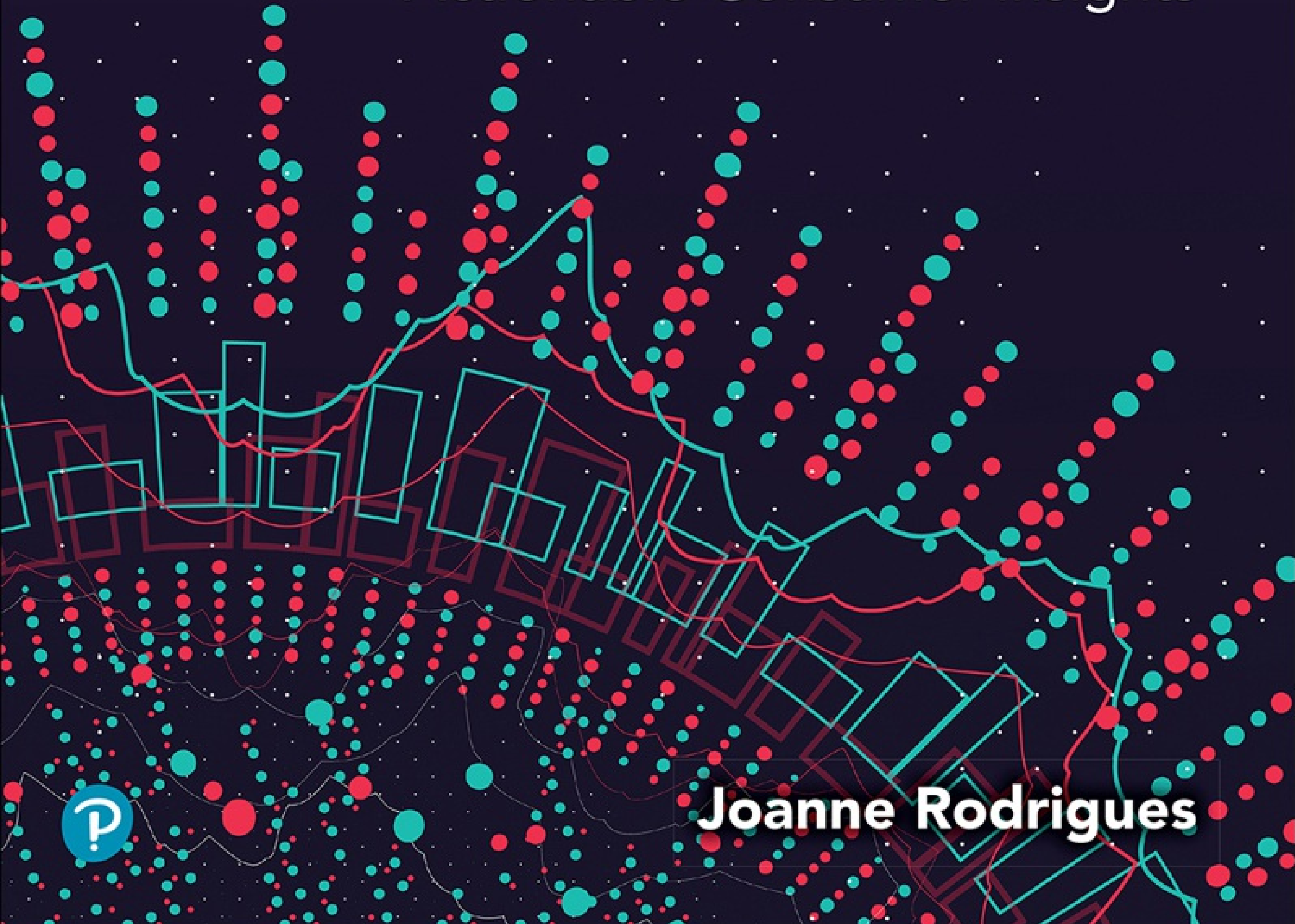
PEARSON BUSINESS ANALYTICS SERIES

Product Analytics

Applied Data Science Techniques for Actionable Consumer Insights

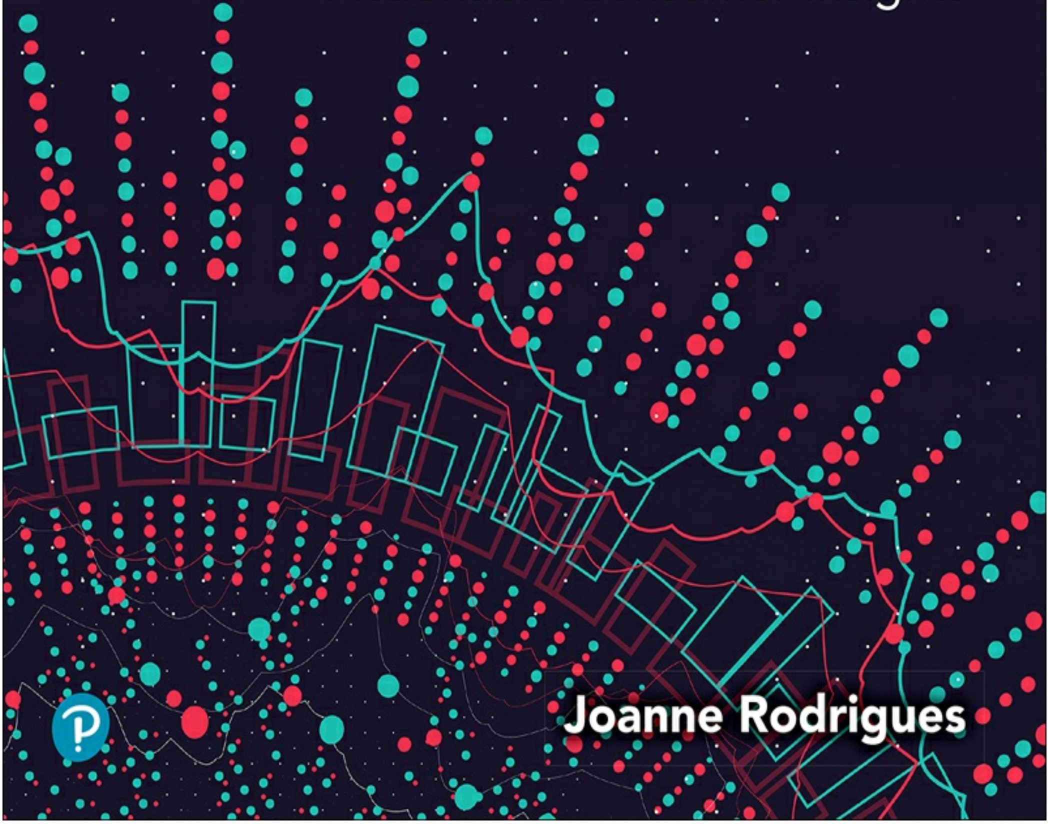


Joanne Rodrigues

PEARSON BUSINESS ANALYTICS SERIES

Product Analytics

Applied Data Science Techniques for Actionable Consumer Insights



Joanne Rodrigues

About This eBook

ePUB is an open, industry-standard format for eBooks. However, support of ePUB and its many features varies across reading devices and applications. Use your device or app settings to customize the presentation to your liking. Settings that you can customize often include font, font size, single or double column, landscape or portrait mode, and figures that you can click or tap to enlarge. For additional information about the settings and features on your reading device or app, visit the device manufacturer's Web site.

Many titles include programming code or configuration examples. To optimize the presentation of these elements, view the eBook in single-column, landscape mode and adjust the font size to the smallest setting. In addition to presenting code and configurations in the reflowable text format, we have included images of the code that mimic the presentation found in the print book; therefore, where the reflowable format may compromise the presentation of the code listing, you will see a "Click here to view code image" link. Click the link to view the print-fidelity code image. To return to the previous page viewed, click the Back button on your device or app.

Product Analytics

Product Analytics

Applied Data Science Techniques for Actionable Consumer Insights

---

Joanne Rodrigues

Addison-Wesley

Boston • Columbus • New York • San Francisco • Amsterdam • Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto • Delhi •
Mexico City
São Paulo • Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

The author and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearson.com.

Visit us on the Web: informit.com/aw

Library of Congress Control Number: 2020940464

Cover image: Mad Dog/Shutterstock

Page 6: “By 2018, the U.S. . . . deep analytic talent.” McKinsey & Company.

Page 21: Introduction to Bayes’ Rule, Thomas Bayes.

Page 51: “if people are . . . adversely and resist.” Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven, Conn.: Yale University Press.

Pages 256-257: “1. Strength of the . . . you could easily test?” Hill, Austin Bradford (1965). “The Environment and Disease: Association or Causation?”. Proceedings of the Royal Society of Medicine. 58 (5): 295–300.

Page 271: “Fewer variables are better: There is . . . number of chimneys swept in a day to determine dosage effects.” Radcliffe, Nicholas, and Surry, Patrick. “Real-world uplift modelling with significance based uplift trees.” White Paper TR-2011-1, Stochastic Solutions, 2011.

Page 281, Figure 14.1: Screenshot of R Studio window with four panels © 2020 Apple Inc.

ISBN-13: 978-0-13-525852-1
ISBN-10: 0-13-525852-9

02 22

To my children Sahana and Ronak, whose infectious laughter kept me focused.

Contents

Preface

> *A point of view can be a dangerous luxury when substituted for insight and understanding.*
>
> —Marshall McLuhan

This book is a practitioner's guide to generating *actionable* insights from consumer data. *Actionable* in this context refers to *extracted* insights used to drive change in a web or mobile product or within a broader organization. Many organizations have terabytes of user-generated data from their web products or internal organizations. However, much of the data goes unused. How should they use this data to make changes that will foster user growth, increase revenue, improve engagement, and engender efficiency in an organization?

Product Analytics will take you step-by-step on the journey to extract insight from user data. The reader will traverse the peaks and valleys of theory building, navigate the waters of designing experiments, drive the meandering roads of developing models, and finally embark on translating these results into actionable insights. This book is a primer on the product data science toolkit. Data science is a multidisciplinary field whose goal is to extract insights from data. Product data science is focused on harnessing user data to drive product and organizational changes to meet core business goals. It emphasizes the use of advanced analytics to understand and change user behavior to help start-ups and large companies alike to build engaging products and exceed revenue targets. As a side note, this book does not address other data science workflows, such as building scalable recommendation systems, computer vision, and image recognition, or other types of applications.

The analyzed data can come from a variety of sources. While it's often user data from web products, it could also be data from emails or mailing campaigns, survey data, internal company data, or integrated data from marketing channels, demographic, or census data, and a variety of other types of data.

The Audience

The core audience for this book consists of entrepreneurs, data scientists, analysts, or any other practitioners who are using user data to drive growth, revenue, efficiency, or engagement in their web or mobile products. This book is useful if you are or want to become a product data scientist, an analyst, or an entrepreneur building a website or web product, or if you just have an interest in working with the terabytes of behavioral data available on the web.

The book is not written for an academic audience, but rather with the practitioner in mind. If you're looking to understand real-world product data, look no further than this book.

Product data science relies on multiple disciplines to extract insight from human behavior. While the analytical toolkit is somewhat more modern, it relies on computing and statistical methods, based on xviiisome of the latest machine learning and causal inference techniques. Social scientists have been studying human behavior for the last 400 years. Social science methods and analytical tools need to be adequately integrated to drive “actionable insights.”

Often, practitioners work with one toolkit, but not others. Many data scientists are well versed in the latest machine learning techniques, but lack the user expertise and qualitative skills to apply those techniques to extracting insights from human data. They often get stuck while developing sufficient theories of social processes and operationalizing conceptual ideas into measurable quantities.

In contrast, many user experts with a sufficient understanding of human behavior lack the statistical and machine learning rigor to adequately test their ideas and model data. The goal of this book is to bridge the gap between the subject-matter expert and the machine-learning guru. Merging the contextual insights of the subject-matter expert and the sophisticated methods of the machine-learning guru can generate useful insights in the web or mobile analytics space.

The Content

Using practical examples from the world of web analytics, readers will discover how to

- Think like a social scientist to contextualize individual behavior in social environments, explore how human behavior develops, and establish the conditions for change
- Develop core metrics and effective key performance indicators for user analytics in any web product
- Understand statistical inference, the differences between correlation and causation, and when to apply each technique
- Conduct more effective A/B tests
- Build intuitive predictive models to capture user behavior in product
- Tease out causal effects from observational data, using the latest quasi-experimental design techniques and statistical matching
- Implement sophisticated targeting methods such as uplift modeling for marketing campaigns
- Project business costs/subgroup population changes by using advanced demographic projection methods

Themes of the Book

This book has three subthemes that permeate the text: (1) qualitative tools from sociology, psychology, and demography integrated with quantitative tools from statistics, machine learning, and computer science applied to the domain of web analytics; (2) methods for causal inference (rather than prediction) since causal inference is integral to altering human behavior; and (3) nonmathematical explanations and demonstrated application in the R language for machine learning and causal inference topics, since most texts in these spaces are not written for practitioners.

Theme 1: Qualitative versus Quantitative Techniques

The first subtheme goes to the heart of this text. The goal is not just to provide analytical tools, but also to provide the resources needed to apply these analytical tools and examples where they are best applied for web applications. Many books within the data science or machine learning realm simply cover the underlying algorithms. While algorithms do play an important role, the cliché “Garbage in, garbage out” comes to mind. Without appropriate data, the algorithms themselves are useless. Applying the wrong algorithm to the wrong problem can lead to a whole host of problems.

To properly apply an algorithm or design an experiment, we must go over the full process of theory building, conceptualization, operationalization, metric building, hypothesis testing, falsification, and more. A large number of qualitative tools are available that we can use to model human behavior and social processes accurately. If we fail to use these tools, we lose out on a great deal of information, nuance, and insight. We also might completely misunderstand “why,” “how,” or “what” users are doing in our web products. Chapters 1–3 examine the qualitative tools needed to understand and model behavior in web products.

Obtaining actionable insights requires understanding the context and the information stored in each variable. If one cannot connect broader conceptual ideas to analytical results, we’re not left with much of anything. A good friend who had a PhD in physics and who worked as a data scientist at a women’s clothing company illustrates this disconnect best. He loved physics and loved applying physics algorithms to any data set, but struggled to connect their results to the business context of interest. I would often ask him what insights he had derived about the women’s apparel business. He always answered that he had applied the latest “X” model with “some extremely complex tuning.” While applying complex, well-tuned algorithms to the right context is awesome, those algorithms can also be applied to the wrong set of data or used to hide the lack of true insight into a topic.

In practice, “actionable insights” do not rely on using the latest algorithm. Better algorithms generally will only slightly improve your results, but bad data will destroy any hope of gaining valuable insights. What is even more common than bad data is misinterpretation of accurate data—a surprisingly frequent occurrence in industry.

For this reason, it’s essential to have good qualitative methodologies in place before any data analysis begins, so we don’t end up with “garbage out.” Since raw data is often not well documented, it’s easy to misunderstand what a variable is measuring or counting. It’s imperative to understand exactly which steps users must take to get to a particular variable and what they have done to get a particular variable outcome. If you’re using a variable as a proxy for a conceptually complex idea, what pieces of that idea is this variable actually measuring? Having theories and good qualitative frameworks in place will allow for the most robust interpretation and actionable use of your data.

Theme 2: Causal Inference

The second theme is this book is the preference for causal inference over prediction. Many data science books are focused on predictive algorithms. This book provides a basic predictive toolkit consisting of the following algorithms: *k*-means, principal components analysis (PCA), linear regression, logistic regression, decision trees, support vector machines, and some time-series modeling techniques. The more advanced topics, such as difference-in-difference modeling, statistical matching, and uplift modeling, are related to causal inference.

The only exception is found in Chapter 9, which covers advanced predictive techniques from demography on population forecasting. In Chapter 9, we use predictive modeling techniques in a somewhat novel way to create better core user metrics (e.g., retention), understand subgroup population changes in our web product, and forecast future population. Generally, for the analysis of user behavior, causal inference is preferred to prediction.

Theme 3: Layman’s Explanations

This book was written because most books about data science, statistical causal inference, or demography are extremely academic and proof-laden. While that is necessary in some contexts, mathematically heavy texts are inaccessible to the common person. Most of these tools don’t need mathematically heavy explanations and can be extremely easy to apply with a minimal understanding of R. Statistical data science and causal inference tools are useful in many business contexts, but are rarely applied in those settings due to their inaccessibility.

The goal of this book is to make all of this information accessible to anyone who has completed high school–level mathematics and statistics. This is a little bit optimistic, since some of the topics—such as statistical matching, uplift modeling, and population forecasting—are extremely mathematically complex. The goal is to make them conceptually understandable first. Those readers with a minimal math background should get a general idea of how the algorithm works and when to apply it. After reading the book, readers should be able to find the right design and/or model to apply to their own specific use-cases. After determining the right setup and algorithm, they should be able to run their analysis in R. The core goal of the book is to teach readers how those algorithms generally work, in which situations they should apply particular algorithms in the user or web analytics context, and which tools in R they can apply to get the answers that they're looking for.

In this book, we'll sparsely use mathematical notation as it turn's away non-mathematically inclined readers. Chapter's from 1-6 will use as little mathematical notation as possible and we'll verbally describe equations. After Chapter 6, the material becomes too mathematically intensive to not rely on not using mathematical notation and later chapters will occassionally use mathematical notation in the text.

Organization of the Book

The goal of this book is to better model, understand, and change user behavior in web and mobile products. The book is organized in the following way:

- Chapters 1–3 explain qualitative tools and theories to model user behavior.
- Chapters 4–6 cover introductory statistical methods in product analytics.
- Chapters 7–9 explore predictive modeling and forecasting methods.
- Chapters 10–13 cover causal inference methods for real-world data.
- Chapters 14–16 implement the methods explained in the quantitative chapters in R.

Chapter 1, "Data in Action: A Model of a Dinner Party," is an introductory chapter, which uses the metaphor of a dinner party to showcase common pitfalls that hinder understanding of user behavior. These pitfalls include that social data is often viewed as a "process," rather than a problem. Social data often has xxino clear outcomes, has rampant problems of incomplete information, has large numbers of variables that are strongly interconnected, is a system that can be easily perturbed, and prevents us from easily inferring causality.

Chapter 2, "Building a Theory of the Social Universe," reviews the scientific method and walks you through sociological tools of quantifying human behavior. Exploring ideas of conceptualization forces us to spend time thinking about "quantifying"—both what that means and what is lost in the process. Today, everything is moving toward metrics. The difficulty with replacing complex qualitative metrics with a few quantitative measures is that these measures can rarely capture the level of sophistication of the original human heuristics or the sophistication that a human expert would expect. Practitioners rarely delve deeply into the shortcomings of their metrics, which leads to even more misguided strategies.

Chapter 3, "The Coveted Goalpost: How to Change Human Behavior," is about human behavior change. User analytics has shifted from demographic profiling to sophisticated methods of targeting and altering user behavior in your web product. What features are most likely to change user behavior? This chapter explores current theories of behavior change, the factors that are most likely to cause change, and the magnitude of a given change.

Chapter 4, "Distributions in User Analytics," takes you through basic statistical tools to start working with user data. In Chapter 5, "Retained? Metric Creation and Interpretation," we explore the nitty-gritty of developing quantitative measures of key ideas. This chapter uses demographic ideas of period, age, and cohort to inform our metric development and expands our toolkit for measuring populations. In addition, Chapter 5 explores the benefits and shortfalls of working with commonly used metrics by working through examples from the four key areas in user analytics: acquisition, retention, engagement, and revenue.

Chapter 6, "Why Are My Users Leaving? The Ins and Outs of A/B Testing," is a practical how-to guide to A/B testing. What is an A/B test? How do you set one up? How do you analyze the results? This chapter also goes through statistical testing and simple power analysis. Finally, it explores the complexities of A/B testing, such as best courses of action for conflicting results between short- and long-run indicators.

Chapter 7, "Modeling the User Space: *k*-Means and PCA," and Chapter 8, "Predicting User Behavior: Regression, Decision Trees, and Support Vector Machines," explore the basics of supervised and unsupervised learning. This introduction to pattern recognition focuses on graphical descriptions and examples to drive understanding. It's a basic toolkit to help you with everyday explanatory or predictive analysis. It also underlies the more sophisticated statistical techniques in Chapters 10–13. Topics covered include *k*-means, PCA, linear regression, logistic regression, decision trees, and support vector machines.

Chapter 9, "Forecasting Population Changes in Product: Demographic Projections," covers ways to forecast general and subgroup population changes in your web product. It relies on tools of demographic population prediction to model user behavior in a multidimensional and
unique way.

Most data produced is observational, meaning that we must tease out causal relationships. Chapter 10, "In Pursuit of the Experiment: Natural Experiments and the Difference-in-Difference Modeling," and Chapter 11, "In Pursuit of the Experiment, Continued," go through some elementary techniques for deriving causal insights from observational data. These techniques include natural experiments, the difference-in-difference design, and regression discontinuity—all of which can help us derive actionable insights from real-world data. Chapter 12, "Developing Heuristics in Practice," explores statistical matching and situations where causal inference is not possible or is not easy.

Predictive modeling with A/B testing can be a powerful combination. Chapter 13, "Uplift Modeling," explores uplift modeling, a technique that combines the two and leads to improved user targeting.

The final section of the book implements all these techniques in R. Chapter 14, "Metrics in R," runs through statistical distributions and metric calculation in R. Chapter 15, "A/B Testing, Predictive Modeling, and Population Projection in R," discusses A/B testing, predictive modeling techniques, and population projection techniques in R. Chapter 16, "Regression Discontinuity, Matching, and Uplift in R," introduces difference-in-difference modeling, statistical matching, and uplift modeling in R.

Final Thoughts

This book provides an intermediate guide to user analytics and relies on both causal and predictive inference. After reading this book, you should be able to build theories about user behavior, test those theories, and generate actionable insights to improve your product. The tools and practical advice from this book can be used in almost any role—from marketing and project management to business analytics and entrepreneur.

Register your copy of *Product Analytics* on the InformIT site for convenient access to updates and/or corrections as they become available. To start the registration process, go to informit.com/register and log in or create an account. Enter the product ISBN (9780135258521) and click Submit. Look on the Registered Products tab for an Access Bonus Content link next to this product, and follow that link to access any available bonus materials. If you would like to be notified of exclusive offers on new editions and updates, please check the box to receive email from us.

Acknowledgments

I’d like to acknowledge the following people. Debra Williams Cauley, the book’s acquisitions editor, guided the book’s development and greatly improved its content. I’d also like to acknowledge the help and excellent feedback of the reviewers, particularly Lawrence Rodrigues, Jared Lander, Luda Janda and Nick Cohron, who read entire sections of the book word-for-word. All of these reviewers provided rich feedback that greatly improved or transformed the book in their own way. Finally, I’d like to acknowledge the general feedback of Sohini Sircar, Cloy Rodrigues, Hubert Lee, Rajesh Mascarenhas, and Paul Chung.

About the Author

Joanne Rodrigues is an experienced data scientist and enterprise manager with master's degrees in mathematics (London School of Economics), political science (University of California, Berkeley), and demography (University of California, Berkeley), and a bachelor's degree in international economics (Georgetown University). Her passion is to analyze large sets of structured, semi-structured, and unstructured data to solve real-world problems. She has six years of experience applying machine learning/statistical algorithms to derive business insights (in health care and gaming). She pioneered new analytics techniques at Sony PlayStation, and led all of MeYou Health's data science efforts. She is also the founder of ClinicPriceCheck.com, a health technology company.

I: Qualitative Methodology

1. Data in Action: A Model of a Dinner Party

Never in human history have we had access to terabytes of social data in a variety of settings. Such data can lead to amazing, *revolutionary* insights into human behavior. The goal of this book is to take you on a *journey* to gain the skills needed to generate those insights to grow your business and improve your products.

Let’s start with the basics. This book is about analyzing customer behavior in a web or mobile product. In the context of this book, a web or mobile product is defined as any product or service available online or through your phone. It includes real products sold online, like a website selling snowmobiles, as well as web and mobile products like social networks. The Internet has changed the game: It allows us to collect lots of data on customer behavior—more than was ever possible before—based on everything from what they clicked on to what they told their friends.

We can use this data to improve our product. Social data is powerful because it allows us to analyze thousands, often millions, of simple behaviors that were unknowable even ten years ago and to then work to change and alter that behavior.

Why is user analytics important? User analytics is the lifeblood of the modern economy. Understanding user interactions with web products often determines whether a company will succeed or fail, even when that firm sells real products. As everything moves online, from retail transactions to doctor’s visits, one must understand why users do what they do and rapidly work to improve user experience.

In the last ten years, the data explosion has allowed us to rapidly iterate and improve social and traditional product marketing, sales, and delivery.

This chapter explores the user analytics space, identifying six core ways in which applications in user analytics differ from traditional data science and statistical applications. Using the example of a dinner party, this chapter will illustrate some of the unique aspects of working with user data. Few technical skills will be learned in this chapter; instead, it primarily lays the framework for how to approach theory building and experimental and/or model design for testing hypotheses in web and mobile products.

The following sections delve into this divide between intuitive insight and advanced analytics, which currently exists in many industries. The rest of the book takes you on a journey to bridge that expansive divide.

1.1 The User Data Disruption

Understanding user behavior in a variety of contexts can lead to better-targeted campaigns, increased revenue, and greater user satisfaction and engagement for any product. A myriad of professionals, from data scientists to product managers, are tasked with understanding, altering, and predicting user behavior. However, even with generous investment, most organizations have difficulty effectively utilizing their data. Leveraging data is difficult, and many analysts do not ask the right questions, utilize the appropriate context, or employ the best tools to make inferences about human behavior. This book will show you the most effective tools in the social sciences and statistics to derive *useful insights* from your users.

In this chapter, we'll go over common problems with analyzing user data, an example of a social process (in this case, a dinner party), and common pitfalls when drawing conclusions from complex social processes. At the end of this chapter, you will have a better sense of some of the difficulties of user analytics. Later chapters will help you solve or work through some of these problems.

1.1.1 Don't Leave the Users out of the Model

Most modern web products have embedded social components that make them microcosms of society. Social hierarchy, friendship, culture, and a litany of interesting interactions and behaviors drive the lives of these products. The complexity of the human behavior involved makes social products incredibly difficult to analyze without the correct toolkit. Even simple purchasing websites can have gigantic stores of behavioral data and complex behavioral processes that can be analyzed, such as user clicks, sessions, purchasing behavior, and churn.

Clickstream data is the path of clicks through a website or model product ordered by time. User sessions comprise a pattern of consistent use from the first to last interaction on a site. Churn is the number of users or rate of leaving the site over a particular period. All of this data is present in almost all web products, and it can be extremely useful, when combined with the right context, for understanding what your users are doing in your product.

To add context, all this clickstream and web data is behavioral data. Sometimes this reality is lost because it's present in a web context, where you cannot view your users engaging with your product.

Why is understanding the behavioral aspect of this data useful? Having a model of human behavior will help us to organize, derive insights, and change user behavior. From a business perspective, when you understand who your users are, what they are doing with your product, and what drives them to purchase and engage with your product, then you can try to modify their engagement and revenue behavior.

For instance, as an analyst, you might ask the following questions: How can I make this web product *sticky*—that is, increase user retention? What causes my users to buy? If we make changes to the product, will users adapt?

First, let's understand how many people initially work with data. A very easy way to look at your data is to focus on *description*. Most analysts stop at this level. Description simply means that you collect data about what people are doing with your product. For instance, suppose the average person visits your site three times in the first month. Users spend 30 seconds looking at listings during the average session. Only 10% of users progress past the homepage. There are a large number of potential descriptive tidbits that you can collect from a web product. Most people’s first inclination is to weave a story together from disparate descriptive facts or to create a story and then search for descriptive facts that support it.

> **Spoiler Alert**
>
> ***Weaving a story together from disparate facts is not a good way to understand your product.*** Why? We don’t have a holistic picture of what’s happening with our product and our inferences are probably not correct. This book will walk you through the process of moving from description to real statistical inferences that will improve your web product. Before, we start this process, let’s explore some common problems with analyzing user data.

1.1.2 The Junior Analyst

When starting out, many analysts try to answer questions about user behavior without exploring the larger context. They silo a certain behavior and then try to explain why that behavior is occurring. However, siloing a certain set of behaviors often does not work, because human behavior and web processes are often complex and deeply interconnected. We also have no “causal linkages,” making any connections proposed very dubious in nature.

For instance, let’s say you have a website that sells snowmobiles. You have great products on your site for snowmobile lovers, but most users never seem to progress past the homepage. Your bounce rate is very high. The bounce rate is the proportion of users who leave the site after viewing only one page. An analyst might spend weeks trying to answer why there is so little progression deeper into the site.

To address this question, the analyst might pull a variety of descriptive statistics. Let’s call her Ana. Ana latches onto one descriptive statistic: 40% of users coming from the search engine Google progress past the homepage compared to only 30% of users coming from the search engine Yahoo. She might then start crafting a story to fit this descriptive data. Google users are richer and worldlier than Yahoo users, and richer, worldlier users are more interested in our snowmobiles, so they progress past the homepage at higher rates. Ana, proud of her work, then shares this nicely crafted story with you.

Okay, you think, this story makes sense. So, to sell more snowmobiles, we simply need to figure out how to better target rich users. But before you invest in better targeting, you ask your analyst what evidence she has that supports this story. She mumbles an answer, “Err … well, we know Google users are richer on average.” Okay, you say, but we do not know what subset of Google users make it to our website. Are they richer than the average Google user? Google users are also more likely to be international, male, and millennials, so how did Ana deal with these potentially conflicting attributes? Any of these attributes might explain some of the difference in rates.

So you ask Ana how she settled on wealth as the driving factor. Ana looks worried, and you realize that she has not thought about this issue. This scenario is not atypical in many industry settings. Without a sound background in statistical inference, it's often difficult to make sense of conflicting descriptive information.

Cherry-picking descriptive statistics also often leads to frustration. Exploring a single behavior can feel like a game of whack-a-mole, where you try to explain away one fact, only to have another conflicting fact pop up in its place. When we focus on only a couple of descriptive facts, we do not get a holistic picture of what is happening in the product—and eventually we just get lost in the details.

Even worse, we may find that our conclusions are incorrect. Since we do not know why more users are coming to our product from Google than Yahoo, we make up a story: *Here they are progressing because they are supposedly richer*. This reasoning makes it seem like these users would be more likely to purchase our snowmobiles.

However, we never even looked at whether they are *actually* purchasing at a higher rate. This comes to the second problem. As we figure out the reasoning for one behavior, another measure of user behavior emerges that fails to make sense in this context.

Let's say that Google users are less likely to purchase snowmobiles than Yahoo users. This information seems to contradict our analyst's story. This often happens in industry when descriptive tidbits are cherry-picked to explain a phenomenon. We do not know why this descriptive feature exists, so we need to do more analysis to figure that out. As you can tell, this is not a good way of exploring user behavior in products. However, coming up with crafted stories around descriptive information is a common practice in many companies and industries.

Executives often want stories that clearly elucidate and explain phenomena. While storytelling is useful when we know what is happening with our products and we can justify those conclusions, faux storytelling can be destructive, leading to the misallocation of resources and failure to effect real change within a web product.

This book explains how to build a model of user behavior, test out that model, and make accurate inferences about what causes that behavior—to help you tell accurate stories about your users. What is *actually leading* to higher purchasing of snowmobiles?

Going back to our example, another question that you may want to ask your analyst is about the size of the effect. Often, this question gets asked in industry, but without understanding the broader relationships between variables, the answer is often wrong.

Let's keep the numbers simple. You have 1,000 organic (or without ads) users daily who come to the site, 50% from Google and 50% from Yahoo. You note that 30% of Yahoo users make a purchase, while 40% of Google users make a purchase. Your analyst theorizes that if all users came from Google and none from Yahoo, it would increase the number of new purchases. On an annual basis, she suggests, you'd see 36,500 more purchases. You tell her great work—let's invest in buying Google ads!

Let's say your ads are so effective that now 1,000 people come directly from Google daily (you meet the criteria for your theory). However, you see only 150 more