Preface

WHY THIS BOOK?

Data analytics involves taking some data and exploring and testing it to produce insights. You can put a variety of names on this process from Business Intelligence to Data Science but fundamentally the approach does not change. Understand a problem, identify the right data, prepare the data appropriately, and run the appropriate analysis on it to find insights and report on them. This is difficult. You are probably seeing this data for the first time. Worse still, the data usually has issues you will only uncover during your journey. Meanwhile, the problem domain must be understood so the data that represents it can be understood. But what is discovered in the data often helps define the problem domain itself.

Faced with this open-ended challenge, many analysts become lost in the data. They explore multiple lines of enquiry. One line of enquiry can invalidate or confirm a previous line. The structure and exceptions in the data are discovered during the process and must be accounted for. Many of the analyses themselves can be executed in a multitude of ways, none of which are categorically correct but instead must be interpreted and justified. Just when you thought you had a handle on the problem, new data arrives and everything you have already done is potentially invalidated. This makes planning, executing, and reproducing data analytics challenging.

If you have ever been in this situation then this book is for you.

WHAT THIS BOOK IS AND WHAT IT IS NOT

First of all, let me cover what this book is not.

- This book is not a prescriptive guide to either specific technologies or analytics techniques. For that you will have to read widely in fields such as machine learning, statistics, database programming, scripting, web development, and data visualization. It is my belief that while technology continues to improve at pace, the fundamental principles of how to do data analytics change little.
- This is not a project management book. I certainly believe project management of analytics needs more attention. Analytics projects are complex and fast-paced and it seems that established project management techniques can struggle to cope with them. This book will help you in areas such as tracking of work but it does not take a project management focus in the presentation of any of its material.
- This book is not about "Big Data." It is also not about little data or medium data. Debates about whether Big Data is something new or indeed something

at all are left to others. As you will see, this book's principles and its practice tips are applicable to all types of data analysis regardless of the scale.
- This book is not about how to build large data warehouses and web-based Business Intelligence platforms. These techniques are also well covered in the literature having been tackled in academia and the software development industry for several decades.

My goal in writing this book is to help people who have been in the same situation as me. I want them to benefit from my experiences and the lessons I have learned, very often the hard way. This book aims to help you in the following three ways.

- **How to do:** This book is a guiding reference for data analysts who must work in dynamic analytics projects. It will help them do high-quality work that is reproducible and testable despite the many disruptions in their project environment and the typically open-ended nature of analytics. It will guide them through each stage of a data analytics job with overarching principles and specific practice tips.
- **How to manage:** This book is a how-to for data analytics managers. It will help them put in place light weight workflows and team conventions that are easy to understand and implement. Teams managed with this book's principles in mind will avoid many of the pain points of analytics. They will be well coordinated, their work will be easily reviewed and their knowledge will be easily shared. The team will become safely independent, freeing up the manager to communicate and sell the team's work instead of being mired in trying to cover every detail of the team's activities.
- **How to build:** Finally, this book is a guide for those with the strategic remit of building and growing an analytics team. Chapters describe the people, processes, and technology that need to be put in place to grow an agile and versatile analytics team.

WHO SHOULD READ THIS BOOK?

Data analytics is a hugely diverse area. Nonetheless, the fundamentals of how to do data analytics, manage analytics teams, and build analytics capability do not change significantly. You will benefit from reading this book if you work in any of the following roles.

- **Data Analyst or Data Scientist:** You are somebody who works directly with data and needs guidance on best practice for doing that work in an agile, controlled, reproducible way. If you have ever experienced been "lost in the data" or losing track of your own analyses and data modifications then this book will help you. If you have ever been frustrated with repeated conversations with your colleagues about where data is stored, what it means, or how your colleague analyzed it then this book will help you both.
- **Analytics Manager:** You are somebody who has several direct reports and you are responsible for guiding and reviewing their analytics work. You have

to jump into many different work products from different team members to review their correctness. You do not have time to waste on facing a different approach, coding convention, data location, or test structure every time you sit down to review a piece of work. Your project resources come and go and you want to facilitate fast transitions and handovers with minimal overhead. You need to be able to explain your team's work with confidence to customers but do not have time to be down in all the details of that work.

- **Senior Manager:** You are somebody who is busy interfacing with a customer and perhaps architecting a high-level approach to a customer's problems. You need to know that your team's work products are reproducible, tested, and traceable. You need to sell the quality, versatility, and speed of mobilization of your team to your customers.
- **Team Director/Chief Information Officer/Chief Data Officer:** You are somebody who wants to build the best analytics team possible to solve customer problems and respond to a wide variety of analytics challenges. To support this ambition, you need a uniformity of skills and methods in your analytics teams for flexibility of resourcing and sharing of knowledge. You want your teams to produce to high standards without suffocating them with rules or requiring they use expensive niche tools. You want your teams to have the right training and toolsets at their fingertips so they can get on with the work they do best.
- **Researcher:** You gather and analyze experimental data for research and publication purposes. This could be algorithm design in computer science, instrumentation data in physics, or any field requiring gathering data to test hypotheses. With such an open-ended exploratory approach to your work, you may struggle to coordinate multiple parallel lines of inquiry, multiple versions of analyses, and experiment result data. This book helps you do all of that so you can focus on reproducible and repeatable publication of results.
- **Research Director:** You are somebody who runs a team of researchers working on multiple concurrent research projects. Your concern is that research is reproducible and sharable among your teams so that the body of knowledge of your team and lab grows over time. You do not have time to be down in the details but you want to know that your team's work is of publication quality in an academic context or can be easily transferred into production in an industry context.

HOW THIS BOOK IS ORGANIZED

I have designed the book so that each chapter is as self-contained as possible and chapters can be read in any order. The book is organized into four parts.

Part 1 Principles introduces Guerrilla Analytics and the Guerrilla Analytics Principles. Begin here if you need an introduction to why analytics is difficult, what can go wrong, and how to mitigate the risks of things going wrong.

- Introducing Guerrilla Analytics
- Guerrilla Analytics: Challenges and Risks
- Guerrilla Analytics Principles

Part 2 Practice covers how to apply the Guerrilla Analytics Principles across the entire analytics workflow. Read any of these chapters if you are working in a particular stage of the Data Analytics Workflow. For example, jump into "Data Load" if you have just received some data from a customer. Look through "Creating Work Products" if you are beginning a piece of work that you will deliver to your customer.

- Data Extraction
- Data Receipt
- Data Load
- Analytics Coding for Ease of Review
- Analytics Coding to Maintain Data Provenance
- Creating Work Products
- Reporting
- Consolidating Knowledge in Builds

Part 3 Testing discusses how to test analytics work to discover defects. Begin with the introduction, if testing is new to you.

- Introduction to Testing
- Testing Data
- Testing Builds
- Testing Work Products

Part 4 Building Guerrilla Analytics Capability is all about the people skills, technology, and processes you need to put in place to establish and grow a Guerrilla Analytics team. Pick up one of these chapters if you are setting up a Guerrilla Analytics environment or are looking to hire and train a team in this book's techniques.

- People
- Process
- Technology

Throughout the book, many of the points will be illustrated with simple examples. "War stories" will describe instances of how things can go badly wrong without the Guerrilla Analytics Principles. The war stories cover a variety of domains to appeal to as many readers as possible.

DISCLAIMER

It is important to state that the examples and war stories from this book are fictional and based on a decade of experiences, conversations, projects, and study. While drawn from real-world experiences, they are not particular to any of my employers or clients, past or present, and should not be interpreted as such.