Chapter 19

Closing Remarks

19.1   WHAT WAS THIS BOOK ABOUT?

Certain types of analytics projects are particularly difficult because they are very dynamic and yet the team is highly constrained. They are dynamic because data and requirements are changing often. They are constrained by tight timelines, limited tooling, and the expectation of testable, traceable analytics outputs. These projects are Guerrilla Analytics projects and their typical workflow is illustrated in Figure 62.

It turns out that much of the confusion, inefficiency, chaos, and frustration of these analytics projects is due to a lack of data provenance. The Guerrilla Analytics principles are a set of guidelines for maintaining data provenance despite the many disruptions of a Guerrilla Analytics project. The principles are as follows.

- **Principle 1:** Space is cheap, confusion is expensive.
- **Principle 2:** Prefer simple, visual project structures over heavily documented and project-specific rules.
- **Principle 3:** Prefer automation with program code over manual graphical methods.
- **Principle 4:** Maintain a link between data on the file system, data in the analytics environment, and data in work products.
- **Principle 5:** Version control changes to data and program code.
- **Principle 6:** Consolidate team knowledge in version-controlled analytics builds.
- **Principle 7:** Prefer analytics code that runs from start to finish.

You may have picked up this book for several reasons. Perhaps you were struggling with the complexity of your analytics projects. Or, you had become demoralized by analytics chaos and were looking for a better way to produce insights from data. Perhaps you wanted a straight-forward perspective on how to approach analytics work with real-world data on real-world projects.

Whatever the reason, hopefully you have found this book to be a useful reference as you work in and manage teams in any stage of the Guerrilla Analytics workflow. The seven Guerrilla Analytics principles above lead to almost 100 practice tips that span the entire Guerrilla Analytics workflow. These can be used as and when needed – different tips will come in handy on different
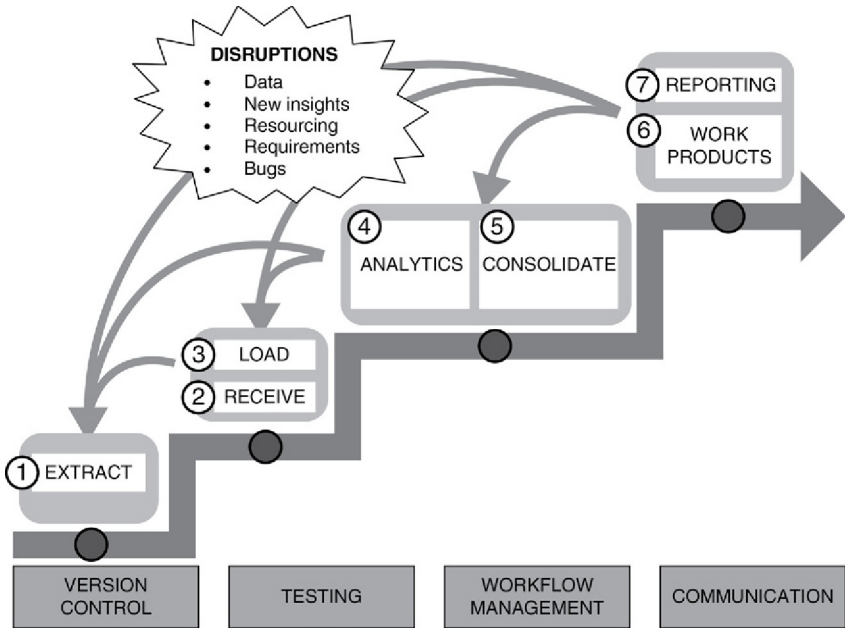
FIGURE 62   The Guerrilla Analytics workflow

projects – and all have been designed to help make your projects run more efficiently and provide you with better results.

For those with a strategic remit, the book also covered the people, process, and technology needed to build a Guerrilla Analytics capability. There is a comprehensive list of skill sets that you should cover in your team if you want to get the most out of your Guerrilla Analytics projects. We also discussed the workflow management needed to manage teams and the technology that you should provide to make the most of their skills.

19.2   NEXT STEPS FOR GUERRILLA ANALYTICS

This book has covered much of what I have learned and implemented over the past 10 years in analytics research, pre-sales, and professional services. In that time, there has been a phenomenal growth in the data volumes and data variety produced around the globe as well as amazing innovations in the technology to do new (and old) kinds of data analytics. Despite or perhaps because of these changes, it still remains complex and difficult to do data analytics. If anything, the number of moving parts on an analytics project has increased. There are more types of data manipulation environment, programming languages, visualizations, web frameworks, etc. than ever before. As long as this is the case, and as long as poorly understood data exists then the Guerrilla Analytics principles will always be required. Here are my thoughts on some of the priorities for Guerrilla Analytics in the coming years.

19.2.1   Better Education in Software Engineering

If there were one course of study that would produce a better Guerrilla Analyst who is ready to work on real-world data, it would probably be software engineering. Many Guerrilla Analytics challenges have already been encountered in a different guise in software engineering. Version control, migration of data, data quality, testing, workflow tracking are well understood and supported by the tools and training of software engineering. Analytics has not yet reached that level of maturity. The approaches in this book are the bare minimum you should do to better control disruptions in the typical analytics project. Much more thinking, training, and tooling needs to be devoted to the equivalent of software engineering for Guerrilla Analytics and analytics in general.

19.2.2   Better Analytics Workflows

Many of the challenges in Guerrilla Analytics are due to the wide variety of activities and tools required to extract insight from ever-changing data. A significant proportion of this book was devoted to overcoming those challenges with existing technology adapted for Guerrilla Analytics and simple conventions to preserve data provenance. This is good but not enough. The analytics workflow needs to be better understood and supported with tools that embrace its dynamism and disruptions. There is a need for better ways to trace data provenance that do not get in the way of doing analytics.

19.2.3   Better Analytics Testing

Data is fundamental to everything that is done in data analytics. In spite of this, the ability to test data and analytics work with powerful tools and well-grounded thinking is almost nonexistent. What is needed is the equivalent of software test frameworks for the five Cs of Completeness, Correctness, Coherence, Consistency, and aCcountability. Yes, these are supported in the fields of data quality and master data management but they do not quite work in the reality of a Guerrilla Analytics project.

19.2.4   What About Big Data?

I have deliberately avoided mentioning Big Data throughout this book. I made this decision because the definition remains muddled and influenced by both technology vendors and by sales pitches. Perhaps that will change in the near future. Nonetheless, there has certainly been a change in the technologies that permit analytics on large volumes and varieties of data. That will not disappear. This begs the question of whether and how Guerrilla Analytics should change to support Big Data. What I suspect will happen is that the convergence of software engineering and analytics will accelerate as more analytics is done on large powerful software platforms, rather than with a mixed bag of tools

borrowed from other fields. The principles of Guerrilla Analytics can certainly help with this, as this scale of data also needs to be explored, analyses need to be version controlled and data provenance will remain of primary importance.

19.3 KEEP IN TOUCH

That brings this book to an end. The principles, practice tips, and war stories described in this book are abstracted from many real-world analytics projects. The hope is that my experiences can help you with your analytics projects in consulting, industry, research, or elsewhere. It has certainly helped me clarify and consolidate my own thinking and experiences over the years. I am always happy to answer questions, provide further detail, and receive feedback about this book.

Also, I am always looking for opportunities to help organizations understand their data and find value in it. This may be through both tactical consulting work and the strategic development of Guerrilla Analytics teams and technology. I welcome any opportunities to share ideas and seek out ways to work together.

You can find the latest news on Guerrilla Analytics at my personal website that I set up for this book - www.guerrilla-analytics.net. Please contact me on Twitter @enda_ridge or via email at guerrillaanalytics@gmail.com.

ACKNOWLEDGMENTS

This book would not have happened without my friend and colleague Edward Curry. The early thoughts on formalizing what became Guerrilla Analytics were shaped through many conversations and several conference presentations with Ed. My heartfelt thanks go to Ed for his early encouragement.

Guerrilla Analytics gathers together all my lessons learned in doing data analytics over many years. My industry and research work was not done in isolation. I thank all my colleagues present and past that supported, challenged, and encouraged me and those who were the guinea pigs and contributors to some of what eventually became the principles of Guerrilla Analytics.

I am very grateful to Andrea Dierna, Acquisitions Editor at Elsevier for guiding me through the book proposal, Kaitlin Herbert and Steve Elliot for their project management support throughout, the copyeditors Priya and Ritu, and other Elsevier staff who made this possible.

I wish to thank my parents and family for their support (and home-cooked meals during some of my most productive writing sessions at home in Galway). Finally, a thank you to Sarah, my patient and supportive sounding board whose proof reading provided a different non-data perspective to help round out the book.

Go raibh míle maith agaibh go léir.