Chapter 1

Introducing Guerrilla Analytics

Having read this chapter, you will understand

- what data analytics is in a very general sense
- the projects in which data analytics is applied
- the type of analytics that is "Guerrilla Analytics"
- examples of Guerrilla Analytics projects

1.1 WHAT IS DATA ANALYTICS?

The last decade has seen phenomenal growth in the creation of data and in the analysis of data to provide insight. Social media and search giants such as Facebook and Google probably spring to mind. These analytics innovators gather immense amounts of data to understand Internet search and social habits so that they can better target online advertising for their customers. Online digital media is generating hours of content and streaming it around the globe for major sporting events such as the FIFA World Cup and the Olympics. In the Financial Services industry, firms process and store billions of financial transactions every day and analyze those transactions to gain an edge in the market over their competitors. Ubiquitous Telco operators store data on our call patterns to analyze it for indicators of customer churn and up-selling opportunities. Every time you book a hotel, flight, or go to the supermarket, loyalty card data is analyzed to better understand customer-purchasing habits and to better target marketing opportunities.

And this growth in data and analytics is not restricted to businesses. Scientific research centers are also creating immense amounts of experimental data in fields such as particle physics, genetics, and pharmacology. Government departments too are not exempt from this trend.

The complexity and pace of change have created a market for data analytics teams in consulting services firms to help their clients both cope with and profit from new data-driven opportunities.

Unsurprisingly, given the growth in data generation, the last decade has also seen a proliferation of the skills and tools needed for extracting value from data. Names for this field include Data Analytics, Data Mining, Quantitative Analysis, Big Data, Machine Learning, Business Intelligence, Artificial Intelligence,

and Data Science. Vendors are frantically racing to provide enterprise grade tools to support work in these fields and to distinguish their offerings from those of their competitors. Universities are trumpeting degree programs that will train a generation of graduates to be conversant in these new technologies and skills.

All of this marketing noise, vendor hype, and pace of change can be confusing and overwhelming for somebody who just wants to get started in data and answer questions to solve problems. Big Data, data velocity, unstructured data, NoSQL, key-value stores, predictive modeling, social network analysis – it is very hard to know where to begin.

Before we get into the details, it will be helpful to step back and think a little about what "data analytics" can mean and agree on what it means to us in this book. Wikipedia, for example, defines "data analytics" as:

> *... a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. (Anon n.d.)*

This definition acknowledges the wide range of activities encompassed by the term data analytics. Tom Davenport's book "Competing on Analytics" offers the following definition.

> *By analytics we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions. The analytics may be input for human decisions or may drive fully automated decisions. (Davenport, 2006)*

Again this is a broad definition. Clearly there are many different opinions on what data analytics is and what it should be called. Let's step back and define data analytics for the purposes of this book.

1.1.1 Data Analytics Definition

First and foremost, this book is a practitioner's book. We, therefore, need a practical definition of data analytics, so we can agree on what is in scope for discussion and what should be left to academic debate.

> *Data analytics is any activity that involves applying an analytical process to data to derive insight from the data.*

Figure 1 illustrates this definition. A customer and/or a third party provides raw data to an analytics team. Analysis is done on the data, producing some modified data output. This output is returned to the customer to provide the customer with insight.

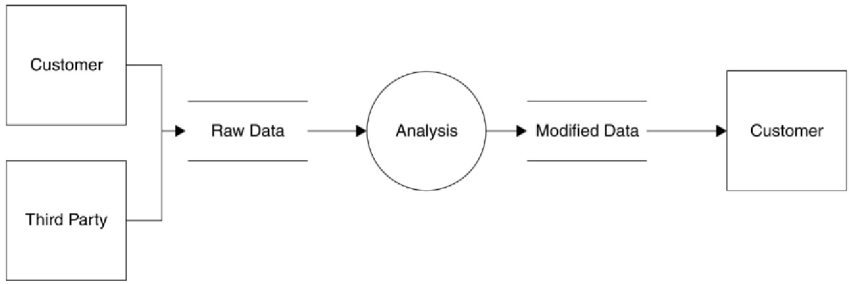

FIGURE 1 Definition of data analytics

1.1.2 Examples of Data Analytics

Such a general definition of data analytics means we can recognize analytics in many scenarios. Here are just a small number of data analytics activities.

- A phone company's customer complaints team keys in 500 poorly scanned customer complaint letters for their data team. The data team reports back on what the common complaint theme is in those letters. They have converted data that was difficult to access into usable data, which was then enriched with complaint keywords. They now have an insight into the common complaint themes from their customers.
- My dad gives me a spreadsheet of household purchases and I tell him how much he spends on groceries per month. I have taken data in the form of dates and purchases and summarized them by month to provide insight into spending patterns.
- Emma, the IT administrator, is concerned about user access controls. She gives Aaron, the data analyst, a year of system log activity. Aaron reports back how users can be grouped based on their activity and what the likely activity is at a particular time on a particular day of the week. Emma now has an insight into who is doing what on the systems she manages.
- Feargus is always looking for new indie bands. An online streaming music website trawls through its user data, mining song plays to make recommendations to Feargus on new artists that he might like.
- A utilities contractor receives its subcontractors' expense claims in hundreds of spreadsheets every month. These spreadsheets are brought together in a database, cleaned, and used to report on subcontractor expenses and search for potentially fraudulent expense claims.
- A financial services firm called OlcBank, having mis-sold financial products to their customers, is tasked with reconstructing the history of its product sales for inspection by a third party and a government regulator.
- A manufacturing plant Widget Inc., suspicious of fraud in its material purchase approvals wants to search its financial and manufacturing data for evidence of fraud.

There are several points to note from these examples of data analytics activities.

- **Technology agnostic:** First, there is no mention of any specific technology involved in the data analytics process. The analyst may be dragging formulas in a spreadsheet. They may be pushing data through the latest parallel streaming data processor. They may be training a troop of analytical monkeys to manipulate the data as required. Our definition is independent of the technology used and should not be confounded with the latest technology trend.
- **Activity agnostic:** Second, there is no differentiation between different types of data analytics activities. Some work is descriptive analytics that creates a summary and profile of data. Other work is data mining that trawls through data looking for patterns. Some work is predictive analytics that builds a model of the data and uses it to make predictions about new data. Some work is combinations of these things. The details of what is done with the data do not matter as long as the data is used to produce insight at some level of sophistication.
- **Scale agnostic:** Third, there is no attempt to comment on the scale or type of the data being analyzed. The work can deal with 100 rows in a spreadsheet table, 10,000 text documents describing insurance claims or some social media data feed approaching scales currently called "Big Data" (Franks, 2012).

This book is aimed at people involved in taking a variety of types of data from a variety of sources, analyzing it with a variety of methods of varying sophistication and returning it to their customers with insight. This insight can be used to make recommendations and take actions.

I cannot emphasize enough the importance of our general data analytics definition. It may surprise you how many activities can be considered as data analytics and how often people fail to recognize that they are working with data and doing analytics!

1.2 TYPES OF DATA ANALYTICS PROJECTS

Data analytics projects exist on a spectrum. At one end of this spectrum we have projects that are close to traditional software engineering projects. By traditional software engineering I mean the production of websites and web applications, desktop software applications, and data warehouses. To develop these analytics applications, a data model is carefully specified, coded, tested, and rolled out through development, user acceptance, and production environments. A presentation layer or application layer is programmed to sit on top of this data and present it to users so they can interact with it. Users may be customers on a website who see recommendations that match their purchasing habits. Users might be online banking customers who see analytics summarizing the performance of their investments or internal business employees who need insights

related to their business's operations. Typical projects are those that manage data feeds, populate data warehouses or implement analytics and management reporting layers on top of warehouses. The development team involved in these projects typically has a variety of roles including database developers, application layer developers, testers as well as data analysts determining how best to extract value from the data. These projects produce software applications in the general sense that we all encounter and use every day on our computers and mobile devices.

At the other end of the spectrum, there are more ad-hoc data analytics projects. These involve taking some sample of data, exploring and analyzing it, and turning around some insight and recommendations based on the analysis. This type of work occurs in many fields. In research, you gather data and analyze it to test hypotheses and ultimately support the publication of research papers. In finance, quantitative analysts gather multiple data sources and mesh them together to present new financial models that give their trading teams an advantage. An organization's internal analytics team is often required to produce one-off ad-hoc support for internal business customers and these analyses often drive key business decisions. In consulting, short-term projects help a client understand the value in their data and inform the client's decision to invest in an analytics platform or an extension to their data warehouse. Inspired by a Harvard Business Review blog (Redman and Sweeney, 2013), I refer to the former projects as **Data Factory** projects and the latter as **Data Lab** projects.

In **Data Factory** projects, the team typically has their own development environment that is well-stocked with the necessary software engineering tools. These projects may be project managed with any number of well-established techniques for software development. The requirements of the project are generally well understood and agreed at a high level. Ultimately, the output is some software application that users will interact with to consume the data analytics insights provided. The team consists of process-oriented engineers who strive for engineering goals such as consistency, testing, and scalability.

In **Data Lab** project types, the data is often completely unknown and the project objective is simply to determine where the data is and what can or should be done with the data. Specification of an analysis or a data model is therefore pointless except on the shortest timescales and needs to be frequently revised as the project progresses. As the data is better understood, requirements change and business rules are revised. As a variety of analyses come into project scope, a wide variety of analytical tools must be applied to the data. The team is composed of "data scientists" whose goals are to find value and insight and to create and test hypotheses in one-off projects.

As Redman and Sweeney (2013) mention, the Data Lab and the Data Factory are complementary entities as illustrated in Figure 2. Innovation and data exploration happen in the lab. Productionized data analytics is rolled out in the factory. This is similar to a typical pharmaceutical company, for example. New drugs are developed and trialed in industrial labs. Then these drugs are

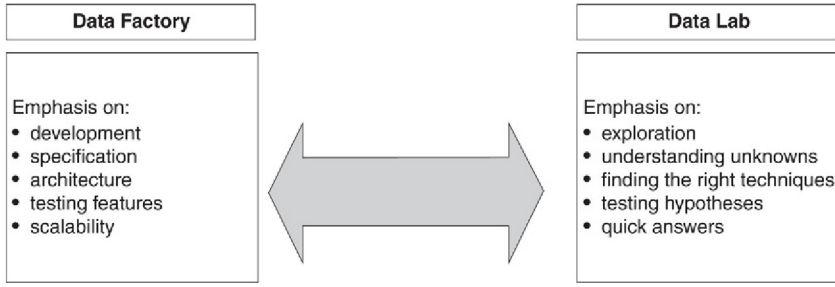

FIGURE 2 Spectrum of data analytics projects

mass-produced in factories and distributed to pharmacies and hospitals. Google advocates a similar complimentary approach to the interplay between their research and engineering (Spector et al., 2012).

1.3 INTRODUCING GUERRILLA ANALYTICS PROJECTS

Data Factory type analytics projects can leverage well-established software engineering approaches because they have so much in common. Collier (2011), for example, gives data warehouse development an analytics perspective. Tools and techniques from software engineering, such as version control, testing, and refactoring, are applicable to data factory projects.

The techniques and tools to use in Data Lab type projects are also well covered by texts on data analytics and machine learning (Witten et al., 2011). Analysts in a data lab have suites of algorithms to choose from and research fields devoted to improving those algorithms and tuning them to work on new data problems.

However, there is a large class of projects where some of the expectations of both the factory and the lab are present. These projects have many of the characteristics of the lab in that the data is not understood and must be explored. They also have many of the requirements of a factory in terms of repeatable and tested analyses that can be easily rolled out to end users. The project environment is extremely dynamic in terms of available resources (both people and technology), changing requirements, and changing data. In these scenarios, the Data Lab approach begins to fail.

- Teams with a range of skill sets and experiences are sharing code and data. There is no longer an individual data scientist coding analytics in isolation.
- The advanced analytics toolkits may not be available if the team is located on the customer's site.
- Clean data is not immediately identifiable or available for experimentation and the data keeps changing.
- Every experiment performed may be subject to external scrutiny and test.
- Every experiment performed may need to be explained to a customer in the context of previous experiment results.

In these scenarios the Data Factory approach also begins to struggle.

- There is no time for detailed data modeling, specifications, and requirements.
- Any requirements that do exist will probably change frequently.
- Helpful engineering tools such as test frameworks and refactoring methods may not be available or may never be made available in a project with short timescales.
- There is little role division – every team member has to be able to contribute to a bit of everything.

These projects are best described with phrases such as "extremely agile," "highly dynamic," "having many disruptions". These projects are where you need **Guerrilla Analytics**.

Think about guerrilla warfare. It is fought with small independent teams having limited weapons at their disposal. Guerrilla fighters are agile and move through a landscape making attacks on their larger enemy. They do not conduct battle in accordance with the conventional rules of engagement.

Our data analysts in these types of projects are similar to guerrillas. They often have limited tools. They do not have the time to produce detailed analytics plans and specifications. Instead they must be agile and go for quick wins under their tight timelines. They must deploy a wide variety of available analytical weapons against their foe – complex data that refuses to yield insights.

1.4 GUERRILLA ANALYTICS DEFINITION

Guerrilla Analytics is data analytics performed in a very dynamic project environment that presents the team with varied and frequent disruptions and constrains the team in terms of the resources they can bring to bear on their analytics problem.

The project environment can be dynamic for the following reasons.

- **Data changes** because of updates, corrections, and discovery of a requirement for new data sources.
- **Requirements change** because as the project progresses, the team and customer's understanding of the problem evolves.
- **Resources change** because these are real-world projects where staff go on leave or change roles and teams are composed of individuals with a wide range of experience and skills.

The project can be constrained in several ways too.

- Time is usually limited and so a "good enough" answer has to be reached quickly and justified.
- Toolsets will often be limited either because of circumstances at a customer's site or because it is simply impossible to anticipate the required tools for an almost infinite number of analytical scenarios.

The next sections elaborate on the typical characteristics of a project requiring Guerrilla Analytics.

1.4.1 Changing Data

Data in real-world projects is always subject to change. The data provided to the team can be replaced, appended to, or updated at a fairly high frequency. For example, in a dynamic data environment you may receive a delivery of several datasets that are critical to the project. After working on those datasets for several days, you could receive another delivery of those same datasets containing the very latest data. Alternatively, perhaps the earlier datasets were incomplete in some way or contained errors. Perhaps new data fields were discovered and added to the project scope.

1.4.2 Changing Requirements

Requirements in real-world analytics projects change at a high frequency. This is common in projects where the data is poorly understood. It is only as the first analyses and data explorations are completed that the analytics team and their customers better understand what can be done with the data. This presents a challenge for the guerrilla analyst who has to develop their analytics in a flexible and agile manner that can accommodate changing requirements while respecting the need for backwards compatibility with previous analyses.

1.4.3 Changing Resource

Highly dynamic resourcing means that you cannot guarantee who your team members will be and the team composition may change during the course of the project. This is very common in professional services and pre-sales functions where the work pipeline is usually quite "lumpy" and hard to predict. Your team is often dictated by who is available at the office. Similarly, in research, collaborative teams are very often composed of researchers from a variety of institutions and departments. The researchers available may be on relatively short-term contracts as they build their cases for academic tenure. You therefore have teams with a moving composition during the lifetime of the project and your projects need to be able to cope with the challenges they present.

1.4.4 Limited Time

Projects requiring Guerrilla Analytics are typically subject to tight timelines, particularly at the start of the project. Analyses can be due within a day or even an afternoon. Progress needs to be demonstrated within days of the project commencing. The guerrilla analyst therefore faces the challenge of developing and releasing work products in a staged manner so they can be interrupted and delivered at multiple time points.

1.4.5 Limited Toolsets

The toolsets available to a Guerrilla Analytics team are often restricted. This can be due to the team being located on a customer's site and subject to their customer's IT policies and available software licenses. It can also be due to tight project timelines where it can often take IT days or weeks to provision software. Since analytics projects are so varied, the right tool for the job is often not known in advance. A Guerrilla Analytics team must be prepared to do the best with what they have available.

1.4.6 Analytics Results Must be Reproducible

It is imperative that despite the obstacles presented above, the work products of the analytics team are reproducible. That is, when the team releases any given dataset, analysis, report, or other work product then the team must be capable of recreating that work product. While reproducibility is desirable in most work, in Guerrilla Analytics projects it is usually a critical requirement.

1.4.7 Work Products must be easily explained

It is one thing to be able to reproduce work products. It is another to be able to easily explain how that work product was derived. Explaining the derivation of a result entails three things.

- Understanding what data was used in creating the work product and where that data came from.
- Knowing how the data was filtered, cleaned, augmented, or any other modifications.
- Quantifying the impact of data modifications on analyses and populations.

1.5 EXAMPLE GUERRILLA ANALYTICS PROJECTS

You are probably wondering where such demanding and difficult sounding projects could occur. Not all projects will have all of the characteristics and requirements laid out in the Guerrilla Analytics definition of the previous section. Nonetheless, projects that have some combination of these characteristics and requirements are actually very common. Here are some examples, which you may recognize in your own work.

- **Forensic Data Analytics:** A financial event needs to be investigated. This could be an accounting fraud involving manipulation of accounts. It could be the circumstances leading up to a bankruptcy. It could be instances of bribery or price fixing. In all cases, legal pressures and scrutiny from one or more parties will require that all data analytics results are clearly derived, tested for correctness, and verifiably complete. However, the sensitivity of the data may mean that much of the analytics work has to be done in the

unfamiliar territory of a client site. Timescales for the data analytics will often be very tight with teams usually arriving to analyze data within days of the alleged events. Because the need for such investigations is unpredictable, the analytics team is usually assembled from available resources. The team may not have worked together before. There will not be established team processes and procuring tools may take too much time under these timescales.

- **Data Analytics for Research:** Research is the ultimate unknown. Its very aim is to better understand some phenomenon. This is usually done by preparing hypotheses, designing an experiment to test the hypotheses, and gathering data from the experiment execution. The reality of modern research both in industry and academia is that it is a business. Like any business, its research outputs are measured and their quality and reputation leads to further funding from sponsors. In this competitive environment, successful research directors build labs that follow a particular line of research. There will be research contributions from summer undergraduate interns, graduate students, postgraduate researchers on short-term contracts, and collaborators from other research labs. In such an environment, a research director's team must produce reproducible analytics under tight publication timelines in such a way that knowledge and analyses can be handed off to other team members and the lab's capability can be grown.
- **Data Journalism:** Data journalism is a relatively new field (Rogers, 2012) pioneered by news publications such as the *Guardian* and the *New York Times*. Simply put, data journalism is about using available data sources to drive or support a compelling news article. The analytics to support data journalism faces many of the Guerrilla Analytics challenges. As news articles are released to a huge public audience, data journalists must ensure that their data, analyses, and conclusions are traceable, reproducible and that the data sources contributing to an article have good provenance. As publication deadlines are tight and breaking news is difficult to predict, data journalists often find themselves facing dynamic requirements, resourcing, and data.
- **Business Analytics and Management Information (MI):** Many organizations have their own internal business analytics team. One of the roles of this team is to provide ad-hoc analyses. These help answer business questions, drive strategy decisions, and provide insights and MI that are not yet available in productionized reports from the organization's data warehouse. We see this in many areas. A loyalty card provider may want to better understand its customer segmentation in light of new products launched by its rival. In banking, a retail bank is embarking on rationalizing its branches nationwide and wants to better understand customer profiles and activities at various branches. In all cases, the requirements are dynamic as they are driven by business exploration. In manufacture, an industrial engineer considering a change in a production process may first want to understand where there are bottlenecks and the current inefficiencies of the manufacturing process.

Data provenance and analytics traceability and reproducibility are critical because the analytics are usually reported to key business stakeholders and are the basis for important business decisions. These internal analytics teams are often extremely busy and have many internal customers to service. They need to be agile in their resourcing and produce analytics that are easily shared and swapped between team members. Since any of their work products may be further developed or may become critical in the boardroom, they also need to maintain their provenance, traceability, and reproducibility despite the very dynamic environment.

- **Quantitative Analytics:** The majority of modern financial trading functions are supported by quantitative analytics teams. These teams gather market data and other relevant third-party data, and use advanced analytics techniques to produce statistical models and recommendations that are relied on by traders. As markets move quickly, so must the quantitative analytics and underlying data. Analyses in these environments may have to stand up to regulatory scrutiny and internal audit despite the very dynamic nature of the work and the challenging demands of traders.
- **Analytics Pre-Sales and Proof of Concept:** In many scenarios, the value of analytics must be established before a customer commits to a sale or an internal stakeholder can secure a budget. This process is called pre-sales and a typical approach is to produce a proof of concept on a sample of the customer's data so they can justify investment in analytics to their business and stakeholders. Since funding is usually limited for a pre-sale, the analytics pre-sales team will face challenges of quickly and cheaply producing analytics. Having won a pre-sale however, there is often an expectation of a quick transition from the pre-sale "lab" to the production "data factory". Pre-sales teams therefore need to be able to mobilize quickly, explore several analyses in parallel, and consolidate the knowledge they acquire so it can be passed on to an implementation team if the sale is successful.

This is a small sample of scenarios from which you can see how prevalent Guerrilla Analytics projects are in research, industry, and consulting and professional services.

1.6 SOME TERMINOLOGY

This book is deliberately not prescriptive in its recommendations. The field of data analytics is too varied and too fast moving to make this a book about a specific programming language or technology. The book is also deliberately general so that data analysts and analytics and senior managers can benefit from its recommendations regardless of their particular industry sector. Before we progress, I want to lay out some common terminology used in the book.

- **Data Manipulation Environment (DME):** This is any environment in which data is modified and analyzed. The term is deliberately general as it

covers relational databases, NoSQL databases, statistical environments such as R and SAS, and quite possibly scripts being run on a file system with a language such as AWK (Dougherty and Robbins, 1997) or Perl (Christiansen et al., 2000). This book's definition of data analytics focuses on the manipulation of data to provide insight. The DME is where this happens.

- **Data Analytics Environment:** The data analytics environment is everywhere that the data analytics team works. This is primarily two places. It is their project folder on a file system and it is their DME(s) as described above.
- **Dataset:** This is a general data structure that is manipulated in a DME. In a relational database, it would be a "table." In R it could be a "data frame." In a NoSQL document database it would be a JSON or XML document. Again, we wish to keep this term as general as possible as the book's principles apply to all data structures.
- **Data Field:** A dataset in the form of a table contains columns of data. A dataset in the form of a NoSQL document contains arbitrary attributes. For example, a "person" document could have first name, last name, and age attributes. In the general sense, this book refers to table columns and document attributes as data fields. This allows us to explore Guerrilla Analytics without being distracted by the underlying data modeling paradigm.
- **The Team:** When we refer to the team, we mean the data analytics team as opposed to any broader project team that they might be embedded in or collaborating with.
- **The Customer:** The customer is anybody who uses the insights created by the data analytics team. Examples include a business analyst or forensic accountant who is working with our analytics team, a client for whom we are engaged to do work or even the reader of a publication. Again, we keep things general and simple to avoid the distraction of internal and external stakeholders, team members, and clients.
- **Work Product:** A work product is a self-contained piece of analytics work. It does not necessarily get delivered to a customer. Typical analytics work products can encompass one or more program code files, spreadsheets, dashboards, data samples, presentations, and reports. When we speak of a work product we are referring to all the components that combine to define an atomic analytics output.
- **Business Rule:** This is a rule about the data that has been agreed between the team and the customer. Business rules are where the data and business understanding interface. Example business rules include "the field called EXP_DATE is the expiry date of an item" or "All financial product description records must have at least one corresponding financial product detail record."
- **Data Flow:** A data flow is one or more data manipulations executed in the production of a work product. A typical data flow takes some raw data, modifies it in one or more ways into derived data, and finally turns it into a presentation format for delivery to the customer.

- **Code:** Code or program code refers to the program commands written to manipulate and visualize data. Again, in the interest of generality, this does not refer to any specific programming language.

1.7 WRAP UP

This chapter introduced Guerrilla Analytics. You should have an understanding of the following.

- There is a wide variety of fields involved in what can be considered data analytics.
- To ensure that the scope of this book is clear, we defined data analytics very generally as "any activity that involves applying an analytical process to data to derive insight from the data."
- We discussed the spectrum of projects that involve data analytics. This spectrum ranges from the "data lab" where data is explored and analyses are trialed through to the "data factory" where the outputs of the lab are scaled up and rolled out through software applications.
- We introduced a type of analytics project that has many of the characteristics and challenges of the lab but is expected to produce outputs with the traceability, reproducibility, and provenance of the factory. These are the projects that require Guerrilla Analytics.
- We gave examples of common Guerrilla Analytics projects and explained the guerrilla warfare metaphor.
- We introduced some common general terminology that will be used throughout the book.