1
Expanding Your Data Mining Toolbox

When faced with sensory information, human beings naturally want to find patterns to explain, differentiate, categorize, and predict. This process of looking for patterns all around us is a fundamental human activity, and the human brain is quite good at it. With this skill, our ancient ancestors became better at hunting, gathering, cooking, and organizing. It is no wonder that pattern recognition and pattern prediction were some of the first tasks humans set out to computerize, and this desire continues in earnest today. Depending on the goals of a given project, finding patterns in data using computers nowadays involves database systems, artificial intelligence, statistics, information retrieval, computer vision, and any number of other various subfields of computer science, information systems, mathematics, or business, just to name a few. No matter what we call this activity – knowledge discovery in databases, data mining, data science – its primary mission is always to find interesting patterns.

Despite this humble-sounding mission, data mining has existed for long enough and has built up enough variation in how it is implemented that it has now become a large and complicated field to master. We can think of a cooking school, where every beginner chef is first taught how to boil water and how to use a knife before moving to more advanced skills, such as making puff pastry or deboning a raw chicken. In data mining, we also have common techniques that even the newest data miners will learn: How to build a classifier and how to find clusters in data. The title of this book, however, is *Mastering Data Mining with Python*, and so, as a *mastering*-level book, the aim is to teach you some of the techniques you may not have seen in earlier data mining projects.

In this first chapter, we will cover the following topics:

- **What is data mining?** We will situate data mining in the growing field of other similar concepts, and we will learn a bit about the history of how this discipline has grown and changed.
- **How do we do data mining?** Here, we compare several processes or methodologies commonly used in data mining projects.
- **What are the techniques used in data mining?** In this section, we will summarize each of the data analysis techniques that are typically included in a definition of data mining, and we will highlight the more exotic or underappreciated techniques that we will be covering in this mastering-level book.
- **How do we set up a data mining work environment?** Finally, we will walk through setting up a Python-based development environment that we will use to complete the projects in the rest of this book.

What is data mining?

We explained earlier that the goal of data mining is to find patterns in data, but this oversimplification falls apart quickly under scrutiny. After all, could we not also say that finding patterns is the goal of classical statistics, or business analytics, or machine learning, or even the newer practices of **data science** or **big data**? What is the difference between data mining and all of these other fields, anyway? And while we are at it, why is it called **data mining** if what we are really doing is mining for patterns? Don't we already have the data?

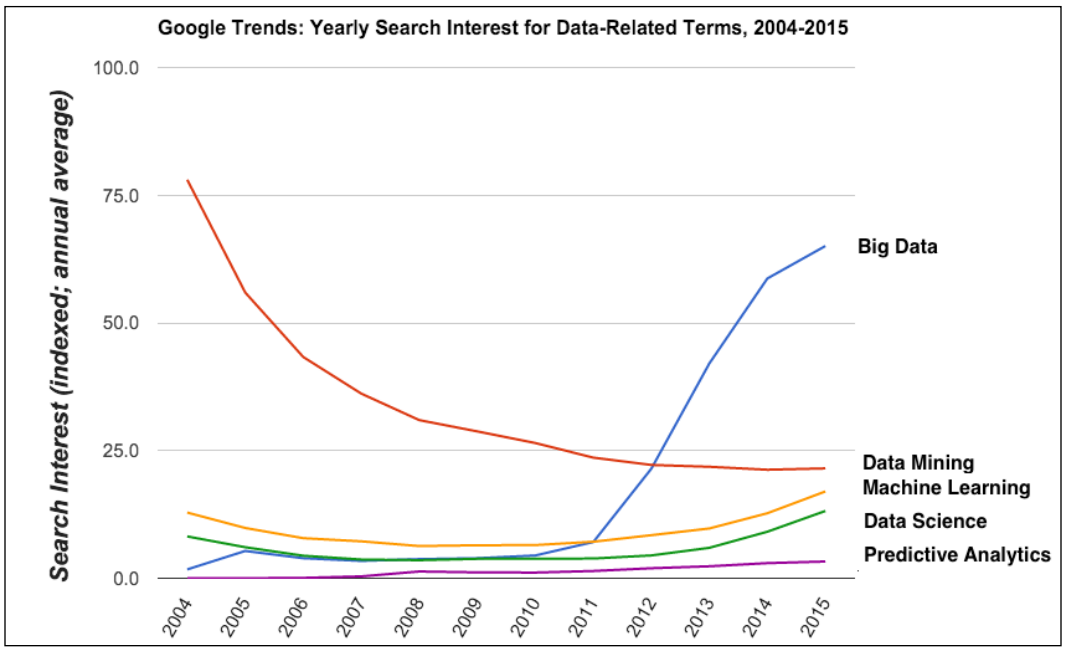
It was apparent from the beginning that the term data mining is indeed fraught with many problems. The term was originally used as something of a pejorative by statisticians who cautioned against going on *fishing expeditions*, where a data analyst is casting about for patterns in data without forming proper hypotheses first. Nonetheless, the term rose to prominence in the 1990s, as the popular press caught wind of exciting research that was marrying the mature field of database management systems with the best algorithms from machine learning and artificial intelligence. The inclusion of the word *mining* inspires visions of a modern-day Gold Rush, in which the persistent and intrepid miner will discover (and perhaps profit from) previously hidden gems. The idea that data itself could be a rare and precious commodity was immediately appealing to the business and technology press, despite efforts by early pioneers to promote the holistic term **knowledge discovery in databases** (**KDD**).

The term data mining persisted, however, and ultimately some definitions of the field attempted to re-imagine the term data mining to refer to just one of the steps in a longer, more comprehensive **knowledge discovery process**. Today, data mining and KDD are considered very similar, closely related terms.

What about other related terms, such as machine learning, predictive analytics, big data, and data science? Are these the same as data mining or KDD? Let's draw some comparisons between each of these terms:

- **Machine learning** is a very specific subfield of computer science that focuses on developing algorithms that can learn from data in order to make predictions. Many data mining solutions will use techniques from machine learning, but not all data mining is trying to make predictions or learn from data. Sometimes we just want to find a pattern in the data. In fact, in this book we will be exploring a few data mining solutions that do use machine learning techniques, and many more that do not.
- **Predictive analytics**, sometimes just called analytics, is a general term for computational solutions that attempt to make predictions from data in a variety of domains. We can think of the terms business analytics, media analytics, and so on. Some, but not all, predictive analytics solutions will use machine learning techniques to perform their predictions. But again, in data mining, we are not always interested in prediction.
- **Big data** is a term that refers to the problems and solutions of dealing with very large sets of data, irrespective of whether we are searching for patterns in that data, or simply storing it. In terms of comparing big data to data mining, many data mining problems are made more interesting when the data sets are large, so solutions discovered for dealing with big data might come in handy to solve a data mining problem. Nonetheless, these two terms are merely complementary, not interchangeable.
- **Data science** is the closest of these terms to being interchangeable with the KDD process, of which data mining is one step. Because data science is an extremely popular buzzword at this time, its meaning will continue to evolve and change as the field continues to mature.

To show the relative search interest for these various terms over time, we can look at Google Trends. This tool shows how frequently people are searching for various keywords over time. In the following figure, the newcomer term data science is currently the hot buzzword, with data mining pulling into second place, followed by machine learning, data science, and predictive analytics. (I tried to include the search term *knowledge discovery in databases* as well, but the results were so close to zero that the line was invisible.) The y-axis shows the popularity of that particular search term as a 0-100 indexed value. In addition, I combined the weekly index values that Google Trends gives into a monthly average for each month in the period 2004-2015.



Google Trends search results for five common data-related terms

How do we do data mining?

Since data mining is traditionally seen as one of the steps in the overall KDD process, and increasingly in the data science process, in this section we get acquainted with the steps involved. There are several popular methodologies for doing the work of data mining. Here we highlight four methodologies: Two that are taken from textbook introductions to the theory of data mining, one taken from a very practical process used in industry, and one designed for teaching beginners.

The Fayyad et al. KDD process

One early version of the knowledge discovery and data mining process was defined by Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth in a 1996 article (*The KDD Process for Extracting Useful Knowledge from Volumes of Data*). This article was important at the time for refining the rapidly changing KDD methodology into a concrete set of steps. The following steps lead from raw data at the beginning to knowledge at the end:

- **Data selection**: The input to this step is raw data, and the output of this selection step is a smaller subset of the data, called the **target data**.
- **Data pre-processing**: The target data is cleaned, oddities and outliers are removed, and missing data is accounted for. The output of this step is **pre-processed data**, or **cleaned data**.
- **Data transformation**: The cleaned data is organized into a format appropriate for the mining step, and the number of features or variables is reduced if need be. The output of this step is **transformed data**.
- **Data mining**: The transformed data is mined for patterns using one or more data mining algorithms appropriate to the problem at hand. The output of this step is the **discovered patterns**.
- **Data interpretation/evaluation**: The discovered patterns are evaluated for their ability to solve the problem at hand. The output of this step is **knowledge**.

Since this process leads from raw data to knowledge, it is appropriate that these authors were the ones who were really committed to the term *knowledge discovery in databases* rather than simply data mining.

The Han et al. KDD process

Another version of the knowledge discovery process is described in the popular data mining textbook *Data Mining: Concepts and Techniques* by Jiawei Han, Micheline Kamber, and Jian Pei as the following steps, which also lead from raw data to knowledge at the end:

- **Data cleaning**: The input to this step is raw data, and the output is **cleaned data**.
- **Data integration**: In this step, the cleaned data is integrated (if it came from multiple sources). The output of this step is **integrated data**.
- **Data selection**: The data set is reduced to only the data needed for the problem at hand. The output of this step is a **smaller data set**.

- **Data transformation**: The smaller data set is consolidated into a form that will work with the upcoming data mining step. This is called **transformed data**.
- **Data mining**: The transformed data is processed by intelligent algorithms that are designed to discover patterns in that data. The output of this step is one or more patterns.
- **Pattern evaluation**: The discovered patterns are evaluated for their interestingness and their ability to solve the problem at hand. The output of this step is an interestingness measure applied to each pattern, representing knowledge.
- **Knowledge representation**: In this step, the knowledge is communicated to users through various means, including visualization.

In both the Fayyad and Han methodologies, it is expected that the process will iterate multiple times over the steps, if such iteration is needed. For example, if, during the transformation step the person doing the analysis realized that another data cleaning or pre-processing step, is needed, both of these methodologies specify that the analyst should double back and complete a second iteration of the incomplete earlier step.

The CRISP-DM process

A third popular version of the KDD process that is used in many business and applied domains is called **CRISP-DM**, which stands for **CRoss-Industry Standard Process for Data Mining**. It consists of the following steps:

1. **Business understanding**: In this step, the analyst spends time understanding the **reasons** for the data mining project from a business perspective.
2. **Data understanding**: In this step, the analyst becomes familiar with the data and its potential promises and shortcomings, and begins to generate **hypotheses**. The analyst is tasked to reassess the business understanding (*step 1*) if needed.
3. **Data preparation**: This step includes all the data selection, integration, transformation, and **pre-processing** steps that are enumerated as separate steps in the other models. The CRISP-DM model has no expectation of what order these tasks will be done in.
4. **Modeling**: This is the step in which the algorithms are applied to the data to discover the **patterns**. This step is closest to the actual data mining steps in the other KDD models. The analyst is tasked to reassess the data preparation step (*step 3*) if the modeling and mining step requires it.

5. **Evaluation**: The model and discovered patterns are evaluated for their value in **answering the business problem** at hand. The analyst is tasked with revisiting the business understanding (*step 1*) if necessary.
6. **Deployment**: The discovered knowledge and models are **presented** and put into production to solve the original problem at hand.

One of the strengths of this methodology is that iteration is built in. Between specific steps, it is expected that the analyst will check that the current step is still in agreement with certain previous steps. Another strength of this method is that the analyst is explicitly reminded to keep the business problem front and center in the project, even down in the evaluation steps.

The Six Steps process

When I teach the introductory data science course at my university, I use a hybrid methodology of my own creation. This methodology is called the Six Steps, and I designed it to be especially friendly for teaching. My Six Steps methodology removes some of the ambiguity that inexperienced students may have with open-ended tasks from CRISP-DM, such as *Business Understanding*, or a corporate-focused task such as *Deployment*. In addition, the Six Steps method keeps the focus on developing students' critical thinking skills by requiring them to answer *Why are we doing this?* and *What does it mean?* at the beginning and end of the process. My Six Steps method looks like this:

1. **Problem statement**: In this step, the students identify what the problem is that they are trying to solve. Ideally, they motivate the case for why they are doing all this work.
2. **Data collection and storage**: In this step, students locate data and plan their storage for the data needed for this problem. They also provide information about where the data that is helping them answer their motivating question came from, as well as what format it is in and what all the fields mean.
3. **Data cleaning**: In this phase, students carefully select only the data they really need, and pre-process the data into the format required for the mining step.
4. **Data mining**: In this step, students formalize their chosen data mining methodology. They describe what algorithms they used and why.
   The output of this step is a **model** and **discovered patterns**.
5. **Representation and visualization**: In this step, the students show the results of their work visually. The outputs of this step can be **tables**, **drawings**, **graphs**, **charts**, **network diagrams**, **maps**, and so on.

6. **Problem resolution**: This is an important step for beginner data miners. This step explicitly encourages the student to evaluate whether the patterns they showed in *step 5* are really an answer to the question or problem they posed in *step 1*. Students are asked to state the limitations of their model or results, and to identify parts of the motivating question that they could not answer with this method.

Which data mining methodology is the best?

A 2014 survey of the subscribers of Gregory Piatetsky-Shapiro's very popular data mining email newsletter KDNuggets included the question *What main methodology are you using for your analytics, data mining, or data science projects?*

- 43% of the poll respondents indicated that they were using the CRISP-DM methodology
- 27% of the respondents were using their own methodology or a hybrid
- 7% were using the traditional KDD methodology
- The remaining respondents chose another KDD method

These results are generally similar to the 2007 results from the same newsletter asking the same question.

My best advice is that it does not matter too much which methodology you use for a data mining project, as long as you just pick one. If you do not have any methodology at all, then you run the risk of forgetting important steps. Choose one of the methods that seems like it might work for your project and your needs, and then just do your best to follow the steps.

For this book, we will vary our data mining methodology depending on which technique we are looking at in a given chapter. For example, even though the focus of the book as a whole is on the data mining step, we still need to motivate each chapter-length project with a healthy dose of *Business Understanding* (CRISP-DM) or *Problem Statement* (Six Steps) so that we understand why we are doing the tasks and what the results mean. In addition, in order to learn a particular data mining method, we may also have to do some pre-processing, whether we call that data cleaning, integration, or transformation. But in general, we will try to keep these tasks to a minimum so that our focus on data mining remains clear. One prominent exception will be in the final chapter, where we will show specific methods for dealing with missing data and anomalies. Finally, even though data visualization is typically very important for representing the results of your data mining process to your audience, we will also keep these tasks to a minimum so that we can remain focused on the primary job at hand: Data mining.

What are the techniques used in data mining?

Now that we have a sense of where data mining fits in our overall KDD or data science process, we can start to discuss the details of how to get it done.

Since the early days of attempting to define data mining, several broad classes of relevant problems consistently show up again and again. Fayyad et al. name six classes of problems in another important 1996 paper (*From Data Mining to Knowledge Discovery in Databases*), which we can summarize as follows:

- **Classification problems**: Here, we have data that needs to be divided into predefined classes, based on some features of the data. We need an algorithm that can use previously classified data to learn how to put unknown data into the correct class.
- **Clustering problems**: With these problems, we have data that needs to be divided into classes based on its features, but we do not know what the classes are in advance. We need an algorithm that can measure the similarity between data points and automatically divide the data up based on these similarities.
- **Regression problems**: We have data that needs to be mapped onto a predictor variable, so we need to learn a function that can do this mapping.
- **Summarization problems**: Suppose we have data that needs to be shortened or summarized in some way. This could be as simple as calculating basic statistics from data, or as complex as learning how to summarize text or finding a topic model for text.
- **Dependency modeling problems**: For these problems, we have data that might be connected in some way, and we need to develop an algorithm that can calculate the probability of connection or describe the structure of connected data.
- **Change and deviation detection problems**: In another case, we have data that has changed significantly or where some subset of the data deviates from normative values. To solve these problems, we need an algorithm that can detect these issues automatically.

In a different paper written that same year, those same authors also included a few additional categories:

- **Link analysis problems**: Here we have data points with relationships between them, and we need to discover and describe these relationships in terms of how much support they have in the data set and how confident we are in the relationship.
- **Sequence analysis problems**: Imagine that we have data points that follow a sequence, such as a time series or a genome, and we must discover trends or deviations in the sequence, or discover what is causing the sequence or how it will evolve.

Han, Kamber, and Pei, in the textbook we discussed earlier, describe four classes of problems that data mining can help solve, and further, they divide them into descriptive and predictive categories. Descriptive data mining means we are finding patterns that help us understand the data we have. Predictive data mining means we are finding patterns that can help us make predictions about data we do not yet have.

In the descriptive category, they list the following data mining problems:

- Data characterization and data discrimination problems, including data summarization or concept characterization or description.
- Frequency mining, including finding frequent patterns, association rules, and correlations in data.

In the predictive category, they list the following:

- Classification, regression
- Clustering
- Outlier detection and anomaly detection

It is easy to see that there are many similarities between the Fayyad *et al*. list and the Han *et al*. list, but that they have just grouped the items differently. Indeed, the items that show up on both lists are exactly the types of data mining problems you are probably already familiar with by now if you have completed earlier data mining projects. Classification, regression, and clustering are very popular, foundational data mining techniques, so they are covered in nearly every data mining book designed for practitioners.