CHAPTER 2

The Shape of Anomaly Detection

The exciting thing about anomaly detection is the sense of discovery. You need a program that can spot what is unusual, so anomaly-detection models are on the lookout for the outliers. To get a sense of how this works, try a simple human-scale example, such as the one shown in Figure 2-1. Can you spot an outlier?

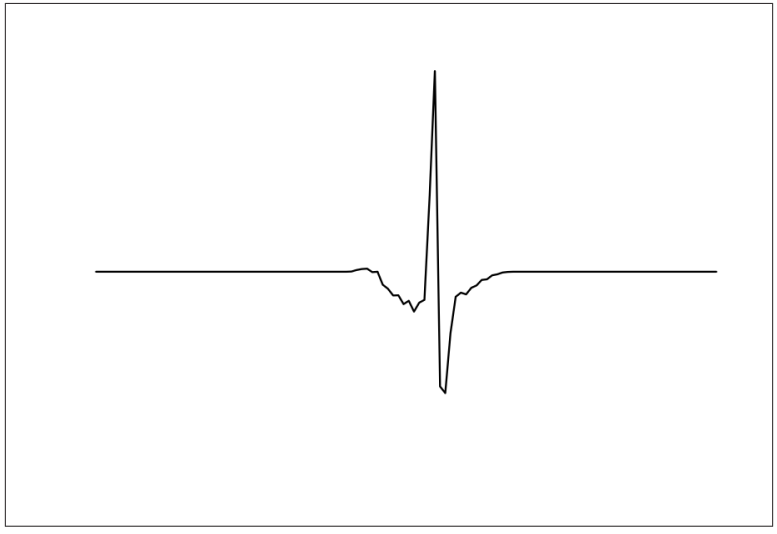

Figure 2-1. Can you spot an anomaly in this data?

Despite the fact that there is apparent noise in the data of the horizontal line shown in Figure 2-1, when you see data like this, it's fairly easy to see that the large spike appears to be an outlier. But is it?

What happens when you have a larger sample of data? Now your perception changes. What had appeared to be an anomaly turns out to be part of a regular and even familiar pattern: in this case, the regular frequency of a normally beating heart, recorded using an EKG, as shown in Figure 2-2.

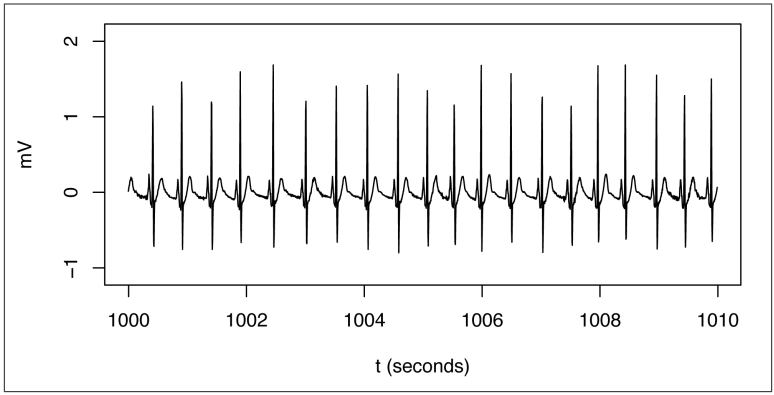

Figure 2-2. Normal heartbeat pattern recorded in an EKG. The spikes that had, in isolation, appeared to be anomalies relative to the horizontal curve are actually a regular and expected part of this normal pattern.

There's an important lesson here, even in this simple small-scale example:

> *Before you can spot an anomaly, you first have to figure out what "normal" is.*

Discovering " normal" is a little more complicated than it sounds, especially in a complex system. Often, to do this, you need a machine-learning model. To do this accurately, you also need a large enough sampling of data to get an accurate representation. Then you must find a way to analyze the data and mathematically define what forms a regular pattern in your training data.

Finding "Normal"

Let's think for a moment about the basic ideas that underlie anomaly detection, including the idea of discovering what is to be considered a normal pattern of behavior. One basic but powerful way to do this is to build a probabilistic model, an idea that we progressively develop