In Praise of Small Data

George Markowsky
School of Computing and Information Science
University of Maine
Orono, ME 04469-5711
markovmaine.edu

Abstract—Big Data tools can give “explanations” of complex and elaborate data sets. There is the danger that we might be content with the explanations that these tools produce. It is important to bear in mind that truly understanding something often requires simplifying the initial explanation. Indeed, the well-known “Occam’s Razor” states roughly that the simplest explanation is the best. This paper examines two cases, Copernicus’s Solar System Model and a “Big Data” search for Paul Revere, where the initial models were too complex. It shows that simpler models are possible, and more fruitful for further research. Some general techniques for finding simpler explanations are discussed.

I. INTRODUCTION

There is a lot of excitement about Big Data these days and its ability to spot trends and identify important principles. Often overlooked in this excitement is the fact that humans best understand situations that can be explained by “small data.” By small data we mean examples that can be easily grasped by the human mind and easily visualized by the human eye. This paper will analyze two examples in detail and demonstrate some general mathematical techniques that can be used to transition from Big Data to small data. The two examples that we will discuss are Copernicus’s model of the Solar System and the intriguing post by Kieran Healy entitled “Using Metadata to find Paul Revere” [8]. In both cases we will describe simpler models that have all of the analytical power of the original models, but are more suited to being well understood by people and more suited to being a base for further discoveries. We shall describe techniques that can be used for discovering small data models buried in big data models.

It should be noted that it has been a long held belief among scientists and philosophers that the simplest explanation is the best.

II. OCCAM’S RAZOR

A widely quoted principle in philosophy and science is Occam’s Razor, so called in honor of the of William Occam (also written Ockham) a philosopher who lived from 1285 until 1347/49, who used it in many of his writings. Briefly it states that entities are not to be multiplied without necessity [1]. This principle is often stated as the simplest explanation of some observation is to be preferred over more complex explanations. Science has made great progress by starting with simple explanations and gradually producing more complex explanations only when facts suggest that the simpler explanations are not adequate.

Occam’s Razor was not original to Occam, but he used it extensively and to great effect. It has been traced at least as far back as Aristotle. For example, from Aristotle’s Posterior Analytics [2] Book I, Part 25 Item (1) we have:

We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypotheses-in short from fewer premises; for, given that all these are equally well known, where they are fewer knowledge will be more speedily acquired, and that is a desideratum.

With computers and Big Data it is very easy to develop complex models. We shall consider two examples where initial complex models can be replaced with simpler models that lead to greater understanding.

III. COPERNICUS’S MODEL

In 1543 Nicolaus Copernicus (1473-1543) [3] published his great opus On the Revolutions of Heavenly Spheres setting off the Copernican Revolution by arguing that the sun and not the earth was the center of the solar system. Actually, the truth is more complicated since according to Koestler [4, p. 193] the Earth and other planets orbit around an imaginary point displaced by three solar diameters from the position of the Sun. Furthermore, the system proposed by Copernicus was quite complicated and based on circles. In short, the Copernican model called for each planet traveling on a small circle whose center was itself traveling on a larger circle that was centered on a point about 3 solar diameters away from the Sun. Furthermore, Copernicus did not have the physical theory worked out that would support his claims that the Earth moved and rotated on its axis.

The impact of Copernicus and his work can be viewed in many ways [4], [5] and we will not discuss this topic further in this paper. It is clear that the Copernican model is very complicated and based on Aristotelian physics with circular motion being the preferred sort of motion. Copernicus was able to achieve only limited precision with his model because he did not have computers and had to do all the complicated calculations of planetary orbits manually. As we shall see in Sections IV and V it is possible to model any orbit as accurately as you want using epicycles. Fortunately for astronomy, Copernicus did not have access to powerful
computers and was only able to produce a limited model of the Solar System.

IV. FOURIER SERIES

In the early 19th Century the mathematician Joseph Fourier (1768-1830) showed how to use trigonometric functions to solve difficult problems in mathematical physics. Since his time, mathematicians have shown that most “reasonable” functions can be expressed as an infinite sum of sine and cosine functions of various types. We call these representations of functions Fourier series. It turns out that some types of problems are easier to solve by working with the Fourier series associated with the function than by working with the function directly. For more details see Körner [6].

Figure 1 shows what is often called a “sawtooth” function. Its equation is given by $y = x$ if $0 \leq x \leq 2$ and $y = 4 - x$ if $2 \leq x \leq 4$. It seems counterintuitive that a sawtooth function can be represented by a series of sines and cosines. Figure 2 shows a first approximation to this function. Its equation is given by

$$y = 1 - \frac{8}{\pi^2} \cos \frac{\pi x}{2}$$

where $0 \leq x \leq 4$. A more accurate approximation to the sawtooth function is shown in Figure 3. Its equation is given by

$$y = 1 - \frac{8}{\pi^2} \left( \cos \frac{\pi x}{2} + \frac{1}{3^2} \cos \frac{3\pi x}{2} \right)$$

where $0 \leq x \leq 4$. An even more accurate approximation to the sawtooth function is shown in Figure 4. Its equation is given by

$$y = 1 - \frac{8}{\pi^2} \left( \cos \frac{\pi x}{2} + \frac{1}{3^2} \cos \frac{3\pi x}{2} + \frac{1}{5^2} \cos \frac{5\pi x}{2} \right)$$

where $0 \leq x \leq 4$.

Equation 3 shows the general pattern that can be followed to produce the infinite series that is equal to the sawtooth function. Note that each succeeding term makes a smaller contribution.

V. VISUALIZING EPICYCLES

In Copernicus’s physics ideal motion would be given by uniform circular motion. Such an orbit is shown in Figure 5. The equation for such motion is given by the following parametric equation

$$(x, y) = (3 \cos t, 3 \sin t)$$
where $0 \leq t \leq 2\pi$. Notice how sine and cosine generate circular motion.

The observations that ancient astronomers were able to make showed that pure circular motion could not explain the motion of the planets. This led to the idea of epicycles. These are circular orbits whose centers travel around a larger circular motion. One possible such orbit is shown in Figure 6 which is given by the parametric equation

$$ (x, y) = (3 \cos t + \cos \frac{4t}{2}, 3 \sin t + \sin \frac{4t}{2}) $$  \hspace{1cm} (5)

where $0 \leq t \leq 2\pi$. The actual representation of epicycles was a bit more complicated than we are presenting here, but the additional details would only obscure the main point we are trying to make.

By adding epicycles to the orbits it is possible to get some very complex motions as shown Figure 7 which is given by the following parametric equation

$$ (x, y) = (3 \cos t + \frac{\cos 4t}{2} + \frac{\cos 16t}{5}, 3 \sin t + \frac{\sin 4t}{2} + \frac{\sin 16t}{5}) $$  \hspace{1cm} (6)

where $0 \leq t \leq 2\pi$.

From results in Fourier theory we know that if we add enough epicycles and play with various parameters we can eventually produce equations that would reproduce the orbits of the planets as accurately as we wanted. Fortunately, Copernicus and his contemporaries did not have powerful computers and were unable to complete this process of explaining planetary orbits. There were significant deviations between their models and the observed motions of the planets. The next section will show why their technical limitations turned out to be a blessing for science.
VI. KEPLER’S SIMPLIFICATION

Johannes Kepler (1571-1630) was a fascinating individual who put astronomy on the right track and provided one of the foundations for the Newtonian revolution. For a very entertaining biography of Kepler see [4]. For a variety of reasons, Kepler was led to realize that the orbits of the planets were ellipses. An ellipse is shown in Figure 8. This curve is given by the following parametric equation

\[(x, y) = (3 \cos t, 4 \sin t)\]  

where \(0 \leq t \leq 2\pi\). It does not require much mathematical experience to realize that the preceding equation is much simpler than all the equations in Section V except for equation 4 which is the equation of a circle.

Ellipses are generalizations of circles. Whereas a circle has a center and a radius, an ellipse in general has two points called foci which are set off from the center an equal distance. Each ellipse is the set of points such that the sum of the distances to the foci is a constant. From this perspective, we can see that circles are ellipses where the two foci are located at the same point. While a circle has a single radius, people think of ellipses as having two radii which are called the semi-major axis and the semi-minor axis. The semi-major axis is the larger of the two radii. In Figure 8 it would be the vertical radius.

Because Kepler was able to come up with a simpler model for planetary orbits, he was able to discover three very fundamental laws of planetary motion which are listed below.

1) Each planetary orbit is an ellipse with the Sun at one of its foci.
2) A line segment drawn between a planet and the Sun sweeps out equal areas in equal time.
3) For all planets, \(p^2/r^3\) is a constant where \(p\) is the orbital period of the planet and \(r\) is the semi-major axis of its orbit.

For their time, these laws were an astonishing achievement both because they were so accurate and because they were so simple.

VII. FURTHER DISCOVERIES

The laws of Kepler and the laws of Galileo laid the foundation for Isaac Newton’s (1642-1727) astonishing achievements in mathematics and physics. In working with these laws and trying to explain the motion of the Moon around the Earth, Newton was led to discover [4, pp. 504-509] his Universal Law of Gravitation which is given by the following equation.

\[F = \frac{GMm}{r^2}\]  

where \(M\) and \(m\) are the masses of the two objects involved, \(r\) is the distance between them and \(G\) is a constant called the Gravitational Constant. One of the validations of the Universal Law of Gravitation is that Newton was able to show that Kepler’s Laws of Planetary Motion followed from it. The result was astonishing: that the motions of the planets that had puzzled astronomers for thousands of years could be explained by a simple quadratic formula. This discovery provided the impetus for modern science. It also did not hurt that Newton discovered the Calculus while working on this problem.

It is not clear that the Newtonian Revolution would have happened if there were enough computing power available at the time to put in enough epicycles to explain planetary orbits. Because people were able to find simpler explanations they were able to advance much further.

VIII. DEALING WITH CONTINUOUS DATA

In this section we wish to give some guidance when looking for “explanations” of continuous data. To be sure there is value in being able to model data accurately even if the explanation is complex. Fortunately, since the time of Copernicus mathematics has advanced tremendously. There are now many techniques for approximating data. A summary of these techniques can be found in the book by Cheney [7], which describes many techniques that can be used for approximating data. It is a good principle to be guided by Occam’s Razor and look for the simplest explanation. Kepler found great success by considering ellipses which are generalizations of circles. So with other data, considering variations of first explanations. There is more hope for making grand discoveries if one can start with a relatively simple and comprehensible explanation. It is good to bear this in mind since we now have the capability of having large numbers of epicycles in our models.

IX. BIG DATA AND THE SEARCH FOR PAUL REVERE

The previous sections dealt with continuous data. This and the following sections will deal with discrete data. The inspiration for this section is a whimsical piece written by a Kieran Healy [8], a Sociology Professor at Duke University. The piece is titled “Using Metadata to Find Paul Revere” and
it takes the form of a letter being written by an member of the Royal Security Administration in London in 1772 in which he analyzes the “Bigge Data” produced by “agent” David Hackett Fischer. In reality, David Hackett Fischer is a Professor of History at Brandeis University who wrote a book entitled *Paul Revere’s Ride* [9] which among other things contains a list of which Revolutionary Era figures belonged to which groups. The groups include organizations and participants in events such as the Boston Tea Party.

The data derived from Fischer’s book [9] can be represented as a matrix with 8 columns and 255 rows. Prof. Healy kindly made the data and his images available for download [10] so everyone can analyze the data personally without having to type it in. The first column lists the name of the person, while the the other 7 columns each correspond to one of 7 groups or events mentioned before. The first row contains the names of the groups. A sample view of the dataset is shown in Figure 9.

If you ignore the first row and first column you get a $254 \times 7$ 0-1 matrix, which we call $A$. Computing $AA^T$ where $A^T$ is the transpose of $A$ gives a $254 \times 254$ matrix that shows the relationships between people. This is pictured in Figure 10 [8] where a line is drawn between people if they belong to some group in common. It is very difficult to read this diagram in detail, but it is clear that there are certain important clusters. Among the conclusions that can be drawn from this analysis is that Paul Revere is a person who bears watching.

At the same time, computing $A^TA$ gives us a $7 \times 7$ matrix that shows the relationships between groups. Figure 11 [8] pictures these relationships in a graph where the thickness of the lines indicates the number of people who belong to both groups.

The analysis presented in [8] is very interesting and the post is well-written so that everyone is encouraged to read the blog entry. In the next section we will revisit this data from the perspective of small data and show that we can learn just as much by keeping the objects we work with as small as possible.

### X. Small Data and the Search for Paul Revere

In this section we will analyze the data discussed in Section IX with an emphasis on keeping the data as small as possible and show how the data can be handled with less computational power. The first thing to note is that the 7 groups presented originally are not all equal. As pointed out by Shin-Kap Han [11, p. 150] 2 of the 7 columns are not membership organizations. In particular, one column shows who participated in the
Boston Tea Party and another column shows who was listed on the Enemies of London list. We will begin our analysis using 5 columns instead of 7 which gives us an almost 29% reduction in the amount of data that we need to handle. We will then revisit the analysis using all 7 columns and compare the results. We will show how to analyze the data using a minimal amount of computation.

First note that if you have rows having 5 cells and each cell can have only 0 or 1, there are at most $2^5 = 32$ different rows possible. This means that the 254 rows can represent at most 32 different types of rows. It turns out that there only 14 different types of rows. These are listed in Figure 12. The number in parentheses at the beginning of each line tells how often each type appears. If you had to do all this work manually without the aid of a computer, you could write each row on a card and sort the cards using some rapid sorting algorithm such as radix sort.

Before proceeding let’s discuss a relatively efficient way of producing the results in Figure 12. If you are using a spreadsheet you can just sort the table by rows. This will make identical rows group together and you can identify the different rows by just reading through the table and noting where the rows change. If you mark where the rows change you can easily determine the number of different types of rows and how often each type appears. If you had to do all this work manually without the aid of a computer, you could write each person’s membership information on a index card and then sort the cards using some rapid sorting algorithm such as radix sort.

Since 112 of the people do not belong to any of the organizations, then we can just drop them from further consideration since we do not have any reason to suspect them of any mischief. Now we are left with a $142 \times 5$ matrix instead of a $254 \times 7$ matrix. This reduces the dataset by 60%. Let’s suppose that our task is to find the people who have the largest number of contacts. We can do so without too much trouble as follows.


Second, observe that for two patterns A and B, if $A \leq B$, but $A \neq B$, in other words, $A < B$, then people with membership pattern B know more people than people with membership pattern A since they belong to at least one more group and that group is not empty since it includes people with membership pattern B. By quickly scanning Figure 12 we can see that membership patterns (01), (02), (04), (07), and (09) are the only ones that need to be considered. All the other patterns are $<$ than some other pattern in the table. For example, (10) $<$ (07).

To calculate the number of people who belong to the groups that appear in a given pattern, we simply add together the number of people having a pattern that has a 1 in common with the pattern we are interested in. For example, to determine how many people are in groups associated with pattern (04) we add together the last numbers of all lines in Figure 12 that have a 1 in some position that pattern (04) has. To calculate the number of people who belong to the groups that appear in a given pattern, we simply add together the number of people having a pattern that has a 1 in common with the pattern we are interested in. For example, to determine how many people are in groups associated with pattern (04) we add together the last numbers of all lines in Figure 12 that have a 1 in some position that pattern (04) has.

### Table 1: The Unique Rows For 5 Columns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Total Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>121</td>
</tr>
<tr>
<td>02</td>
<td>65</td>
</tr>
<tr>
<td>03</td>
<td>32</td>
</tr>
<tr>
<td>04</td>
<td>126</td>
</tr>
<tr>
<td>05</td>
<td>13</td>
</tr>
<tr>
<td>06</td>
<td>74</td>
</tr>
<tr>
<td>07</td>
<td>85</td>
</tr>
<tr>
<td>08</td>
<td>11</td>
</tr>
<tr>
<td>09</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>69</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
</tr>
</tbody>
</table>

### Table 2: The Numbers Known by Pattern

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Total Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>121</td>
</tr>
<tr>
<td>02</td>
<td>65</td>
</tr>
<tr>
<td>03</td>
<td>32</td>
</tr>
<tr>
<td>04</td>
<td>126</td>
</tr>
<tr>
<td>05</td>
<td>13</td>
</tr>
<tr>
<td>06</td>
<td>74</td>
</tr>
<tr>
<td>07</td>
<td>85</td>
</tr>
<tr>
<td>08</td>
<td>11</td>
</tr>
<tr>
<td>09</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>69</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
</tr>
</tbody>
</table>
that people having any of the other patterns of membership are acquainted with many fewer people than people whose membership patterns (01) and (04).

Let’s quickly revisit this approach using the full 7 columns. In this case there would be a maximum of $2^7 = 128$ different patterns, but again we are lucky because there are only 25 distinct patterns. The resulting patterns are given in Figure 14. We can redo the membership calculations and the results are shown in Figure 15. This time pattern (05) stands out markedly from the other patterns with a membership number of 249, while no other pattern has a membership number greater than 201. From Figure 14 we see that there is only one person having membership pattern (05): Paul Revere!

**XI. Dealing with Discrete Data**

In the previous section we showed how we could obtain useful results by reducing the data. Not only was there less data to handle, but we were able to find relatively efficient algorithms for finding the most suspicious person or persons out of a set of people. These techniques have broad applicability, but the exact way to apply them depends on the problem at hand and the questions that need to be answered. In general, when doing data mining it is helpful to think clearly about what you are trying to do. The dream is to get answers fully automatically, but at this time high quality results generally require some human thought and intervention.

**XII. Conclusions**

Occam’s Razor is alive and well in the 21st Century. We have examined two cases, one dealing with continuous data and one dealing with discrete data. In the continuous case, we showed that being flexible with various parameters allowed us to come up with a much simpler explanation for planetary orbits. This simpler explanation facilitated discovering the Law of Gravitation. Discovering this law is of greater significance than having a simpler explanation for planetary orbits. In the discrete case, we showed that some simple data manipulations allowed us to reduce the size of data drastically and to still come to the same conclusions. In both cases we demonstrated that focusing on shrinking the data and complexity of our models can pay big dividends not only because algorithms are more efficient on smaller inputs, but also because models that are easier to comprehend are more likely to lead to great discoveries than models that are overly complex.

**References**

[10] https://github.com/kjhealy/revere