Introduction to the Mathematics of Big Data

Philippe B. Laval

Fall 2017
Introduction

In recent years, Big Data has become more than just a buzz word. Every major field of science, engineering, business, and finance produces huge amount of data [7]. This data, has to be generated, acquired, stored, then repeatedly processed and analyzed. Because Big Data has affected so many disciplines, some even talk about the Big Data revolution. In fact, Big Data is mentioned not only in books, [1], [2], [4], [5], [7], and [12], but also in various magazines [3], [6], [8], [9], [10], and [11]. But what is Big Data and how big is Big Data? This class will give a short overview of Big Data, will discuss the issues associated with Big Data, and will provide some answers. Big Data is a new area which covers many disciplines and technologies. This class will focus on scientific data and on the mathematical techniques used to address some of the issues which arise with Big Data. The computer science and statistical techniques used to address the same issues are equally important but will not be addressed in this class.
Part I

Basic Concepts
Chapter 1

Mathematics and Big Data

1.1 Definitions and Overview of the Issues with Big Data

1.1.1 What is Big Data?

Big Data is still a new area, evolving quickly. Many of the definitions and techniques are still evolving rapidly and are therefore not formal. As of the writing of these notes, here is how big data is defined.

Definition 1.1.1 Big Data refers to a data set that is so large and/or complex that it cannot be perceived, acquired, managed, and processed by traditional Information Technology (IT) and software/hardware tools within a tolerable time.

Of course, this definition is not absolute. As technology improves, the amount of data that can be processed within a tolerable time also increases. With this definition, Big Data, ten years ago, would have been much smaller than today. In addition, the notion of tolerable time is also relative. Medical doctors are known to be impatient; they do not want to have to wait a lot between the time an image is captured and the time it is available for them to analyze. In contrast, scientists studying outer space wait years before receiving images taken and sent by some probe.

Big Data can be characterized by the three V’s.

1. Volume: refers to the huge volume of data that is generated and has to be stored and analyzed.

2. Variety: refers to the different types of data (text, images, sound). Data can also be structured or unstructured. The complexity of the data can also drastically change.

3. Velocity: refers to the fact that more and more data is being generated at a faster and faster pace.
Some researchers have now started to add a fourth V, Value. It refers to the value that could be saved if Big Data techniques were creatively and effectively used to improve efficiency and quality.

Big Data has given rise to several new and related technologies. These would be studied in a class approaching this subject from the computer science point of view. These related technologies include:

1. Cloud computing.
2. Internet of Things (IoT).
3. Data Centers.
4. Hadoop.

1.1.2 How Big is Big Data?

The smallest unit of measurement for data is a bit (b). A bit can be either 0 or 1 and is therefore not large enough to hold any data. A byte (B), which is 8 bits, is used as the fundamental unit of measurement for data. A byte can hold $2^8 = 256$ different values, which is enough to represent the standard ASCII characters, such as letters, numbers and some basic symbols. Following the tradition of the metric system, terms to measure large quantities of data can be formed using SI prefixes as shown in the table below. These prefixes are often used for multiple of bytes. For example, a kilobyte is $10^3$ bytes since kilo means $10^3$. Because Big Data is so large, the prefixes used for it are not known to the public. We review them in the table below.

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Unit Name</th>
<th>Symbol</th>
<th>SI Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilo</td>
<td>kilobyte</td>
<td>kB or KB</td>
<td>$10^3$</td>
</tr>
<tr>
<td>mega</td>
<td>megabyte</td>
<td>MB</td>
<td>$10^6 = (10^3)^2$</td>
</tr>
<tr>
<td>giga</td>
<td>gigabyte</td>
<td>GB</td>
<td>$10^9 = (10^3)^3$</td>
</tr>
<tr>
<td>tera</td>
<td>terabyte</td>
<td>TB</td>
<td>$10^{12} = (10^3)^4$</td>
</tr>
<tr>
<td>peta</td>
<td>petabyte</td>
<td>PB</td>
<td>$10^{15} = (10^3)^5$</td>
</tr>
<tr>
<td>exa</td>
<td>exabyte</td>
<td>EB</td>
<td>$10^{18} = (10^3)^6$</td>
</tr>
<tr>
<td>zetta</td>
<td>zettabyte</td>
<td>ZB</td>
<td>$10^{21} = (10^3)^7$</td>
</tr>
<tr>
<td>yotta</td>
<td>yottabyte</td>
<td>YB</td>
<td>$10^{24} = (10^3)^8$</td>
</tr>
</tbody>
</table>

It is hard to measure how big Big Data is, as the size of Big Data is constantly increasing, and increasing at a faster and faster rate. More than a specific number, what is important is an order of magnitude. Here are some figures though.

Every day, humanity tweets 500 million times [3].
Every day, humanity shares 70 million photos on Instagram [3].
Every day, humanity watches 4 billion videos on Facebook [3].
Every minute, we upload 300 hours of new content on YouTube [3].
A 2014 study by the market-research firm IDC estimated that the world of digital data would grow by a factor of 10 from 2013 to 2020, to 44 zettabytes [3].

1.1.3 Issues with Big Data
With new technologies, as more and more objects are connected, more and more data is being produced every day. This data has to be generated, acquired, stored, transmitted, and analyzed. There are issues to deal with at each stage of this data pipeline.

- Acquisition/storage: the total data generated is larger than the total storage capacity. Hence, we have to find ways to store data more efficiently.
- Transmission: the increase in the generation rate of this data is far greater than the increase in communication rate.
- The analysis of a data set can be very complex. Even if it is feasible, it may take a long time.

There are additional issues to consider.

- Data can be noisy, hence it has to be processed to remove as much noise as possible, without modifying the data.
- Data is also often unstructured. Think of the data we get from social networks. It is a mixture of text, voice, images, videos. What someone is looking for in a data set is not likely to come from a table in which data is nicely organized, as it was the case in the past.
- Data can also be dynamic, hence it has to be processed in real time.

Mathematics plays an important role in addressing these issues, as we will see. More specifically, mathematics ...

- allows us to formalize both the data and the problem. In other words, it allows us to transform a complex problem into a mathematical problem on which we can use all the tools mathematics has at its disposal.
- provides a big "chest" of tools or methodologies. Mathematics does have a very large chest of tools.
• ...allows validation of the methodologies (proof of functionality). Think, for example, about a noisy image that we clean to see what it has to show us. It may be a medical image. How do we know that the procedure used to remove noise is not going to insert features which were not in the original image?

1.1.4 Exercises

1. Write at least one paragraph for each question below:

(a) What is cloud computing is and how does it relates to Big Data?

(b) What is the Internet of Things (IoT) is and how does it relates to Big Data?

(c) What are Data Centers are and how do they relate to Big Data?

(d) What is Hadoop is and how does it relates to Big Data?

(e) Give examples in which Big Data plays and will play an important role.

(f) Does Big Data play a role in your life, in your career? How?
Bibliography


