We must acquire supercomputers and a staff of talented programmers who can analyze all our data and tell us what it all means!

The NIH leadership believed, much as my high school's coach believed, that if you have a really big computer and you feed it a huge amount of information then you can answer almost any question.

(Data) Science is a reductive process, moving from complex, descriptive data sets to simplified generalizations. The idea of developing an expensive supercomputer facility to work with increasing quantities of biological data, at higher and higher levels of complexity, seemed impractical and unnecessary (see Glossary item, Supercomputer).

The diagnostic supercomputer facility was never built. The primary diagnostic tool used in hospital laboratories is still the microscope, a tool invented circa 1590. Today, we learn from magazines and newspapers that scientists can make important diagnoses by inspecting the full sequence of the DNA that composes our genes.

Nonetheless, physicians rarely order whole genome scans; nobody understands how to use the data effectively. You can find lots of computers in hospitals and medical offices, but the computers do not calculate your diagnosis. Computers in the medical workplace are largely relegated to the prosaic tasks of collecting, storing, retrieving, and delivering medical records.

Before we can take advantage of large and complex data sources, we need to think deeply about the meaning and destiny of Big Data.

DEFINITION OF BIG DATA

Big Data is defined by the **three V's:**

1. Volume—large amounts of data

2. Variety—the data comes in different forms, including traditional databases, images, documents, and complex records

3. Velocity—the content of the data is constantly changing, through the absorption of complementary data collections, through the introduction of previously archived data or legacy collections, and from streamed data arriving from multiple sources

It is important to distinguish Big Data from "lotsa data" or "massive data." In a Big Data Resource, **all three V's** must apply. It is the size, complexity, and restlessness of Big Data resources that account for the methods by which these resources are designed, operated, and analyzed.

The term "lotsa data" is often applied to enormous collections of simple-format records, for example, every observed star, its magnitude and its location; every person living in the United Stated and their telephone numbers; every cataloged living species and its phylogenetic lineage; and so on.

Some “lotsa data” collections are spreadsheets (two-dimensional tables of columns and rows), mathematically equivalent to an immense matrix. For scientific purposes, it is sometimes necessary to analyze all of the data in a matrix, all at once.

This kind of global analysis on large matrices is not the subject of this book.

Big Data resources are not equivalent to a large spreadsheet, and a Big Data resource is not analyzed in its totality.

Big Data resource

*A Big Data collection that is **accessible for analysis**. Readers should understand that there are collections of Big Data (i.e., data sources that are large, complex, and actively growing) that are not designed to support analysis; hence, not Big Data resources. Such Big Data collections might include some of the older hospital information systems, which were designed to deliver individual patient records, upon request, but **could not support projects wherein all of the data contained in all of the records was opened** for selection and analysis. Aside from privacy and security issues, opening a hospital information system to these kinds of analyses would place enormous computational stress on the systems (i.e., produce system crashes).*

In the late 1990s and the early 2000s, data warehousing was popular. Large organizations would collect all of the digital information created within their institutions, and these data were stored as Big Data collections, called data warehouses. If an authorized person within the institution needed some specific set of information (e.g., emails sent or received in February, 2003; all of the bills paid in November, 1999), it could be found somewhere within the warehouse. For the most part, these data warehouses were not true Big Data resources because they were not organized to support a full analysis of all of the contained data.

Another type of Big Data collection that may or may not be considered a Big Data resource is compilations of scientific data that are accessible for analysis by private concerns, but closed for analysis by the public. In this case, a scientist may make a discovery, based on her analysis of a private Big Data collection, but the data collection is not open for unauthorized critical review.

In the opinion of some scientists, including myself, if the results of a data analysis are not available for review, the analysis is illegitimate; the Big Data collection is never consummated as a true Big Data resource.

Big Data analysis is a multistep process whereby data is extracted, filtered, and transformed, with analysis often proceeding in a piecemeal, sometimes recursive, fashion.

As you read this book, you will find that the gulf between “lotsa data” and Big Data is profound; the two subjects can seldom be discussed productively within the same venue.

BIG DATA VERSUS SMALL DATA

Big Data is not small data that has become bloated to the point that it can no longer fit on a spreadsheet, nor is it a database that happens to be very large.

One can't simply apply their spreadsheet and database skills directly to Big Data resources without mastering new skills and without adjusting to new analytic paradigms.

When the data gets bigger, thinking only the computer must adjust (by getting faster, acquiring more volatile memory, and increasing its storage capabilities) is false;

Big Data does pose special problems that a supercomputer could not solve.

Lack of above understanding among database managers, programmers, and statisticians, is highly counterproductive. It leads to slow and ineffective software, huge investment losses, bad analyses, and the production of useless and irreversibly defective Big Data resources.

General differences that can help distinguish Big Data and small data

1. Goals
   small data—Usually designed to answer a specific question or serve a particular goal.

   Big Data—Usually designed with a goal in mind, but the goal is flexible and the questions posed are tentative.

   Here is a short, imaginary funding announcement for Big Data grants designed "to combine high-quality data from fisheries, Coast Guard, commercial shipping, and coastal management agencies for a growing data collection that can be used to support a variety of governmental and commercial management studies in the lower peninsula."

   No way to completely specify what the Big Data resource will contain and how the various types of data held in the resource will be organized, connected to other data resources, or usefully analyzed.

   Nobody can specify, with any degree of confidence, the ultimate destiny of any Big Data project; it usually comes as a surprise.

2. Location
   small data—Typically, small data is contained within one institution, often on one computer, sometimes in one file.
   Big Data—Typically spread throughout electronic space, typically parceled onto multiple Internet servers, located anywhere on earth.
3. Data structure and content

small data—Ordinarily contains highly structured data. The data domain is restricted to a **single discipline or subdiscipline**. The data often comes in the form of uniform records in an ordered spreadsheet.

Big Data—Must be capable of absorbing unstructured data (e.g., such as free-text documents, images, motion pictures, sound recordings, physical objects). The subject matter of the resource may **cross multiple disciplines**, and the individual data objects in the resource may link to data contained in other, seemingly unrelated, Big Data resources.

4. Data preparation

   small data—In many cases, the data user prepares her own data, for her own purposes.

   Big Data—The data comes from many diverse sources, and it is prepared by many people. **People who use the data are seldom the people who have prepared the data.**

5. Longevity

   small data—When the data project ends, the data is kept for a limited time (seldom longer than 7 years, the traditional academic life span for research data) and then discarded.

   Big Data—Big Data projects typically contain data that must be stored in **perpetuity**. Ideally, data stored in a Big Data resource will be absorbed into another resource when the original resource terminates. Many Big Data projects extend into the future and the past (e.g., legacy data), accruing data prospectively and retrospectively.

6. Measurements

   small data—Typically, the data is measured using one experimental protocol, and the data can be represented using one set of standard units (see Glossary item, Protocol).

   Big Data—Many different types of data are delivered in many different electronic formats. Measurements, when present, may be obtained by many different protocols. **Verifying the quality of Big Data** is one of the most difficult tasks for data managers.

7. Reproducibility

   small data—Projects are typically repeatable. If there is some question about the quality of the data, reproducibility of the data, or validity of the conclusions drawn from the data, the entire project can be repeated, yielding a new data set.

   Big Data—Replication of a Big Data project is seldom feasible. In most instances, all that anyone can hope for is that bad data in a Big Data resource will be found and flagged as such.

8. Stakes

   small data—Project costs are limited. Laboratories and institutions can usually recover from the occasional small data failure.

   Big Data—Big Data projects can be obscenely expensive. A failed Big Data effort can lead to bankruptcy, institutional collapse, mass firings, and the sudden disintegration of all the data held in the resource.

As an example, an NIH Big Data project known as the "NCI cancer Biomedical Informatics Grid" cost at least $350 million for fiscal years 2004 to 2010 (see Glossary item, Grid). An ad hoc committee reviewing the resource found that despite the intense efforts of hundreds of cancer researchers and information specialists, it had accomplished so little and at so great an expense that a project moratorium was called.3 Soon thereafter, the resource was terminated.4 Though

the costs of failure can be high in terms of money, time, and labor, Big Data failures may have some redeeming value.

Each failed effort lives on as intellectual remnants consumed by the next Big Data effort.

9. Introspection
   small data—Individual data points are identified by their row and column location within a spreadsheet or database table (see Glossary item, Data point). If you know the row and column headers, you can find and specify all of the data points contained within.
   Big Data—Unless the Big Data resource is exceptionally well designed, the contents and organization of the resource can be inscrutable, even to the data managers (see Glossary item, Data manager). **Complete access to data, information about the data values, and information about the organization of the data is achieved through a technique herein referred to as introspection** (see Glossary item, Introspection).
10. Analysis
    small data—In most instances, all of the data contained in the data project can be analyzed together, and all at once.
    Big Data—With few exceptions, such as those conducted on supercomputers or in parallel on multiple computers, Big Data is ordinarily analyzed in incremental steps (see Glossary items, Parallel computing, MapReduce). The data are extracted, reviewed, reduced, normalized, transformed, visualized, interpreted, and reanalyzed with different methods.

WHENCE COMEST BIG DATA?

Often, **the impetus** for Big Data is entirely ad hoc. Companies and agencies are forced to store and retrieve huge amounts of collected data (whether they want to or not).

Generally, Big Data come into existence through any of several different mechanisms.

1. An entity has collected a lot of data, in the course of its normal activities, and seeks to organize the data so that materials can be retrieved, as needed.
   The Big Data effort is intended to streamline the regular activities of the entity. In this case, the data is just waiting to be used.

   The entity is not looking to discover anything or to do anything new. It simply wants to use the data to do what it has always been doing—only better.

   The typical medical center is a good example of an "accidental" Big Data resource. The day-to-day activities of caring for patients and recording data into hospital information systems results in terabytes of collected data in forms such as laboratory reports, pharmacy orders, clinical encounters, and billing data. Most of this information is generated for a one-time specific use

(e.g., supporting a clinical decision, collecting payment for a procedure). It occurs to the administrative staff that the collected data can be used, in its totality, to achieve mandated goals: improving quality of service, increasing staff efficiency, and reducing operational costs.

2. An entity has collected a lot of data in the course of its normal activities and decides that **there are many new activities that could be supported by their data**. Consider modern corporations—these entities do not restrict themselves to one manufacturing process or one target audience. They are constantly looking for new opportunities. Their collected data may enable them to **develop new products based on the preferences of their loyal customers, to reach new markets, or to market and distribute items via the Web.** These entities will become hybrid Big Data/manufacturing enterprises.

3. An entity plans a business model based on a Big Data resource. Unlike the previous entities, this **entity starts with Big Data and adds a physical component secondarily**. Amazon and FedEx may fall into this category, as they began with a plan for providing a data-intense service (e.g., the Amazon Web catalog and the FedEx package-tracking system). The traditional tasks of warehousing, inventory, pickup, and delivery had been available all along, but lacked the novelty and efficiency afforded by Big Data.

4. An entity is part of a group of entities that have large data resources, all of whom understand that it would be to their mutual advantage to **federate their data resources**.5 An example of a federated Big Data resource would be hospital databases that share electronic medical health records.6

5. An entity with skills and vision develops a project wherein **large amounts of data are collected and organized to the benefit of themselves and their user-clients**. Google, and its many services, is an example (see Glossary items, Page rank, Object rank).

6. An entity has no data and has no particular expertise in Big Data technologies, but it has money and vision. The entity seeks to **fund and coordinate a group of data creators and data holders who will build a Big Data resource that can be used by others**. Government agencies have been the major benefactors. These Big Data projects are justified if they lead to important discoveries that could not be attained at a lesser cost, with smaller data resources.

THE MOST COMMON PURPOSE OF BIG DATA IS TO PRODUCE SMALL DATA

If I had known what it would be like to have it all, I might have been willing to settle for less. Lily Tomlin
Imagine using a restaurant locater on your smartphone. With a few taps, it lists the Italian restaurants located within a 10 block radius of your current location. The database being queried is big and complex (a map database, a collection of all the restaurants in the world, their longitudes and latitudes, their street addresses, and a set of ratings provided by patrons, updated continuously), but the data that it yields is small (e.g., five restaurants, marked on a street map, with pop-ups indicating their exact address, telephone number, and ratings). Your task comes down to selecting one restaurant from among the five and dining thereat.

In this example, your data selection was drawn from a large data set, but your ultimate analysis was confined to a small data set (i.e., five restaurants meeting your search criteria). **The purpose of the Big**

Data resource was to proffer the small data set. No analytic work was performed on the Big Data resource—just search and retrieval. The real labor of the Big Data resource involved **collecting and organizing complex data so that the resource would be ready for your query.**

Along the way, the data creators had many decisions to make (e.g., Should bars be counted as restaurants? What about take-away only shops? What data should be collected? How should missing data be handled? How will data be kept current?).

Big Data is seldom, if ever, analyzed in toto. There is almost always a **drastic filtering process that reduces Big Data into smaller data.** This rule applies to scientific analyses. The Australian Square Kilometre Array of radio telescopes,7 WorldWide Telescope, CERN's Large Hadron Collider, and the Panoramic Survey Telescope and Rapid Response System array of telescopes produce petabytes of data every day (see Glossary items, Square Kilometer Array, Large Hadron Collider, WorldWide Telescope). **Researchers use these raw data sources to produce much smaller data sets for analysis.**8

OPPORTUNITIES

Make no mistake. Despite the obstacles and the risks, the potential value of Big Data is inestimable. A hint at future gains from Big Data comes from the National Science Foundations (NSF) 2012 solicitation for grants in core techniques for Big Data (BIGDATA NSF12499). The NSF aims to

advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life. The new knowledge, tools, practices, and infrastructures produced will enable breakthrough discoveries and innovation in science, engineering, medicine, commerce, education, and national security.10

The NSF envisions a Big Data future with the following pay-offs:

Responses to disaster recovery empower rescue workers and individuals to make timely and effective decisions and provide resources where they are most needed; Complete health/disease/genome/environmental knowledge bases enable biomedical discovery and patient-centered therapy; the full complement of health and medical information is available at the point of care for clinical decision-making; Accurate high-resolution models support forecasting and management of increasingly stressed watersheds and eco-systems; Access to data and software in an easy-to-use format are available to everyone around the globe; Consumers can purchase wearable products using materials with novel and unique properties that prevent injuries; The transition to use of sustainable chemistry and manufacturing materials has been accelerated to the point that the US leads in advanced manufacturing; Consumers have the information they need to make optimal energy consumption decisions in their homes and cars; Civil engineers can continuously monitor and identify at-risk man-made structures like bridges, moderate the impact of failures, and avoid disaster; Students and

researchers have intuitive realtime tools to view, understand, and learn from publicly available large scientific data sets on everything from genome sequences to astronomical star surveys, from public health databases to particle accelerator simulations and their teachers and professors use student performance analytics to improve that learning; and Accurate predictions of natural disasters, such as earthquakes, hurricanes, and tornadoes, enable life-saving and cost-saving preventative actions.10

Many of these hopes for the future may come true if we manage our Big Data resources wisely.

BIG DATA MOVES TO THE CENTER OF THE INFORMATION UNIVERSE

Physics is the universe's operating system. Steven R. Garman

In prior times, scientists followed a well trodden path towards truth: hypothesis, then experiment, then data, then analysis, then publication. The manner in which a scientist analyzed his or her data was crucial because other scientists would not have access to the same data and could not reanalyze the data for themselves. Basically, the final manuscript was the scientific product. Scientific knowledge was built on trust.

In the **Big data paradigm**, the concept of a final manuscript has little meaning. **Big Data resources are permanent, and the data within the resource is immutable (see Chapter 6). Any scientist's analysis of the data does not need to be the final word; another scientist can access and reanalyze the same data over and over again.**

Today, hundreds or thousands of individuals might contribute to a Big Data resource. The data in the resource might inspire dozens of major scientific projects, hundreds of manuscripts, thousands of analytic efforts, or billions of search and retrieval operations. The Big Data resource has become the central, massive object around which universities, research laboratories, corporations, and federal agencies orbit. These orbiting objects draw information from the Big Data resource, and they use the information to support analytic studies and to publish manuscripts. Because Big Data resources are permanent, any analysis can be critically examined, with the same set of data, or reanalyzed anytime in the future. Because Big Data resources are constantly growing forward in time (i.e., accruing new information) and backward in time (i.e., absorbing legacy data sets), the value of the data is constantly increasing.

Big Data resources are the stars of the modern information universe. All matter in the physical universe comes from heavy elements created inside stars, from lighter elements. **All data in the informational universe is complex data built from simple data.** Just as stars can exhaust themselves, explode, or even collapse under their own weight to become black holes, Big Data resources can lose funding and die, release their contents and burst into nothingness, or collapse under their own weight, sucking everything around them into a dark void. It's an interesting metaphor. The following chapters show how a Big Data resource can be designed and operated to ensure stability, utility, growth, and permanence; features you might expect to find in a massive object located in the center of the information universe.