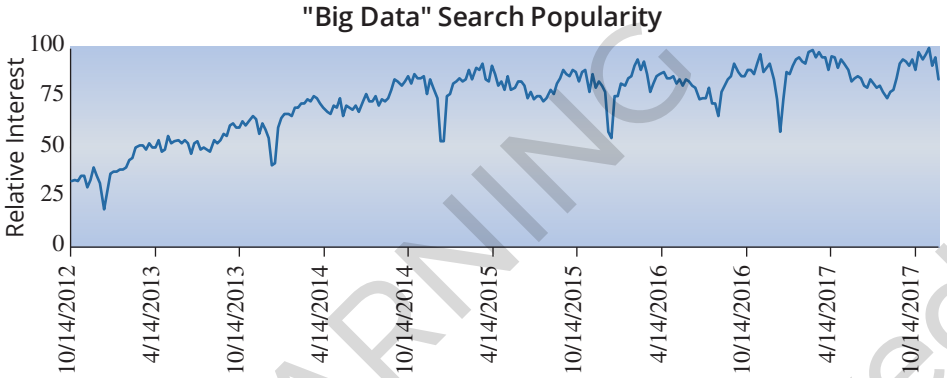
Flash Boys: Data Velocity

Flash Boys is a book written by Michael Lewis about high frequency stock trading. Part of the book describes the great lengths a firm went to reduce the time it takes to send a buy or sell order between New York and Chicago. The best available time in 2008 was 14.65 milliseconds (14.65 thousands of a second). But theoretically it should be possible to communicate between the two cities over a fiber line in 12 milliseconds. The book tells the story of what it took to build and market a "direct" fiber run between New York and Chicago at a cost of 300 million dollars. Note, that is 300 million dollars spent to improve data velocity by slightly more than 2 milliseconds.

1.5 Big Data

Big Data is a loosely defined concept used to describe data sets produced by our globally networked, internet-driven, sensor-laden world. Interest in Big Data has been steadily increasing as evidenced below by the number of Google searches of the term.[3]



While there isn't broad agreement on exactly what Big Data is, there is broad agreement that Big Data will accelerate the pace of discovery in science, as well as innovation in commerce. In fact, that has already happened.

There have been numerous ideas about what makes data "Big Data." Most experts would agree that it is a large volume of data, structured or unstructured. Beyond that, opinions differ. One criteria that is appealing is any data set that is too large to process on commonly available computer systems. More recently the term has been used to refer to a set of analytical models that are used on data sets, regardless of their size. At the moment, the most common meaning of Big Data is a set of data sufficiently large to be challenging to analyze at a typical data center. As a frame of reference, the minimum for a large data center would be on the order of tens of thousands of servers and thousands of data storage arrays.

Another characteristic of Big Data is that it requires teams of programmers, database programmers, statisticians, and machine-learning experts to analyze the data. The Big Data team will usually be using highly scalable cloud computing resources.

Data has many attributes. However, Big Data seems to have four attributes that make it different: volume, variety, velocity, and veracity. These characteristics constitute the four Vs of Big Data.

- **Volume** is the scale of the data, and Big Data implies large volumes of data. According to IBM, most companies in the U.S. have at least 100 terabytes of stored data. But some companies have exabytes of data and are receiving 100's of terabytes of new data every day.
- **Variety** is the different forms data can take—from traditional data elements in a structured database to highly unstructured images, twitter feeds, movies, and audio.
- **Velocity** is how fast data is being sent to the data processing and data management infrastructure. There is a technical term called "streaming data" which is data generated continuously and usually in small amounts by thousands of data sources. Streaming data would be common in e-commerce, gaming, social networks, stock trading, and telemetry data from monitoring

systems. One aspect of Big Data is that the data streams have substantial velocity. YouTube, for example, has an amazingly large data stream.

→ 300 hours of video uploaded every minute

→ 5 billion videos watched every day

→ 30 million visitors per day

It is quite a technical challenge to neatly place high velocity data into the appropriate data repository every minute of every hour of every day without fail.

- **Veracity** is the trustworthiness of the data. Data is a major asset of any company, institution, or government agency. Uncertainty, bias, or inaccuracies in the data make the information less valuable for meaningful analysis and decision–making.

Sources of Big Data in Science

Despite the availability of enormous computing power, some areas of science and industry have data sets so large that they overwhelm modern computing systems. In the sciences, particle physics, astronomy, genomics, meteorology, and internet searches have amassed enormous quantities of data.

Science is being profoundly affected by an abundance of measurements. Almost all large natural science data sets grow because their data is being measured and gathered by specially designed automated measurement systems (machines).

- In the case of genomics, the development of very fast and relatively inexpensive DNA sequencers.
- In the case of astronomy and cosmology, it is new telescopes with very large digital camera arrays.
- In the case of meteorology, it is satellite imagery and automated weather sensors.
- In the case of particle physics, it is particle colliders (like the eight-billion-dollar Large Hadron Collider LHC).

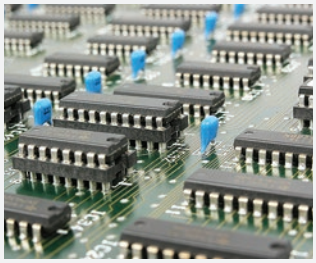
These sensing machines are generating enormous quantities of data from their sensor arrays. The data they are providing offer a huge opportunity to advance our understanding of science and medicine.

Medicine

It used to be that doctors recorded everything they did on your chart. Now, it all goes into a database. This happens every day on every patient. For example, your doctor may order a test such as an MRI of your brain or an echocardiogram of your heart. It would not be unusual for an MRI of your brain to be 220+ megabytes of data. An echocardiogram could be as little as 40 megabytes, and an interventional study (a surgical procedure) could be as much as one gigabyte. Even a chest x-ray would be about 20 megabytes. Medical technology generates an enormous amount of data. Figure 1.5.1 shows the number of echocardiograms paid for just by Medicare from 2007 to 2011. Storing just the annual echocardiography data that Medicare pays for would require roughly 280 petabytes uncompressed.[4]

Data Compression and Big Data

The size of medical imaging files can be greatly reduced by data compression. Compression that eliminates statistical redundancy yet enables the complete restoration of the original data is called "lostless." There are many forms of "lostless" data compression. For example, telescope imagery has many images that are nearly the same, it would not be surprising if the images are stored as difference from a previous image. This is called data differencing. The storage size of some forms of Big Data can be dramatically shrank using compression designed for specific data types. In 2012 a special algorithm was developed for compressing genetic data. The algorithm achieved 20-fold compression (shrunk the file by 95%). Newer algorithms for genetic data have compression rates up to 1200-fold enabling an entire 6 billion pair genome to be stored in 2.5 megabytes.



"The Most Important Master's Thesis of the 20th Century"

In 1948, Claude Shannon wrote a paper entitled "A Mathematical Theory of Communication" which was the foundational work for a field now called information theory. The paper introduced the term "bit" and demonstrated that a series of bits "1s and 0s" (eight of them make a byte) could be used to represent all information. The bit/byte would become the standard unit for data storage and network communication of the future. Shannon's foundational work in information theory was not his only contribution. His masters thesis has been called the most important masters thesis of the 20th century. It showed that electrical switches could be configured to perform Boolean logic functions (i.e. digital logic). Shannon's work became the foundation of digital circuit design. Digital circuits are the fundamental component of all digital computers and without them we would not have modern computers, nor modern statistics.

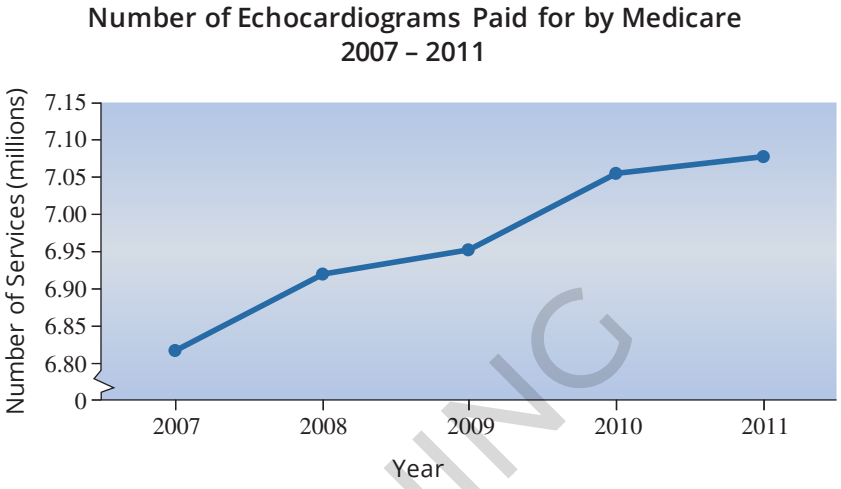


Figure 1.5.1

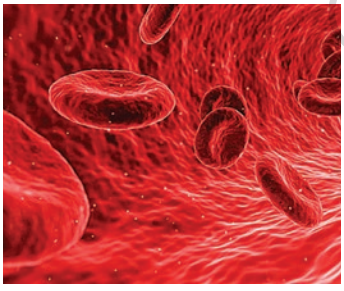
Once a patient's data is anonymized, it can be combined and aggregated. Looking at disease from a broad perspective of aggregated patient health data can provide new insight in a disease process. It can reveal biomarkers that were unknown, and more easily predict the trajectory of a disease, and perhaps, offer an intervention.

For example, there are two large cancer databases that have followed hundreds of thousands of cancer victims for 15 years or more. Once an oncologist diagnoses the specific cancer, they need to develop a treatment plan. A good oncologist will usually recall 6–8 similar cases to help formulate the plan. Now, the oncologist can call upon a cancer database that will have thousands of similar cases, and the information system attached to the database can make recommendations for the treatment plan.

Also, the oncologist might use the cancer-genome atlas—which classifies cancers by their genome—looking for treatments against a specific cancer genome. The oncologist might utilize the new field of proteomics, which is a study of the proteins in a patient's blood. One drop of blood passed through a superconducting magnet can generate 40+ gigabytes of data on all the proteins in the blood, which is the environment that the cancer cells are growing in.

Genomics

Genomics is a field that maps and studies the DNA (genomes) of biological entities. Every plant, animal, bacteria, and virus has a design that is contained in its genetic material (DNA) stored in each cell. The DNA is a blueprint for the organism. It determines whether an organism will produce leaves or legs, and of course many other things. In 1995 the genomes of two bacteria, homopholus influenza and microplasm genitalia, were sequenced, meaning the letters of their DNA were read and stored. The influenza genome is 2,000,000 base pairs long. Once there were large volumes of data, it wasn't long before computer scientists and statisticians began entering the field of biology.



The DNA strand is made up of four chemical building blocks, called nucleotides [adenine (A), thymine (T), guanine (G) and cytosine (C)]. Essentially, DNA encodes information. Human DNA has about 3 billion base pairs of nucleotides. Sequencing a human genome means to determine the specific base pairs for nearly all 3 billion pairs associated with the individual's genome. So, one entry into a human genomic database contains various combinations of ATGC for the individual's 3 billion base pairs.

Genomics is around 20 years old. The Human Genome Project originally took 10 years to process one human genome; now this can be achieved in less than a week. As of 2015 all genomic data represented approximately 25 petabytes. The amount of data being produced in genomics daily is currently doubling every seven months. Within the next decade, genomics is looking at generating somewhere between 2 and 40 exabytes a year, depending upon whether the data doubles every seven months or every 18 months as shown in Figure 1.5.2.[5]

It is estimated that 1 billion people will have their DNA sequenced by 2025. If this happens, genomic databases will likely be the largest databases in existence.

There are databases that contain large numbers of completely sequenced human genomes. There are databases with completely sequenced genes for people with autism, cancer, muscular dystrophy, heart disease, and virtually any other disease that might have a genetic component. There are DNA databases that specialize in specific mammals, insects, bacteria, viruses, and plants.

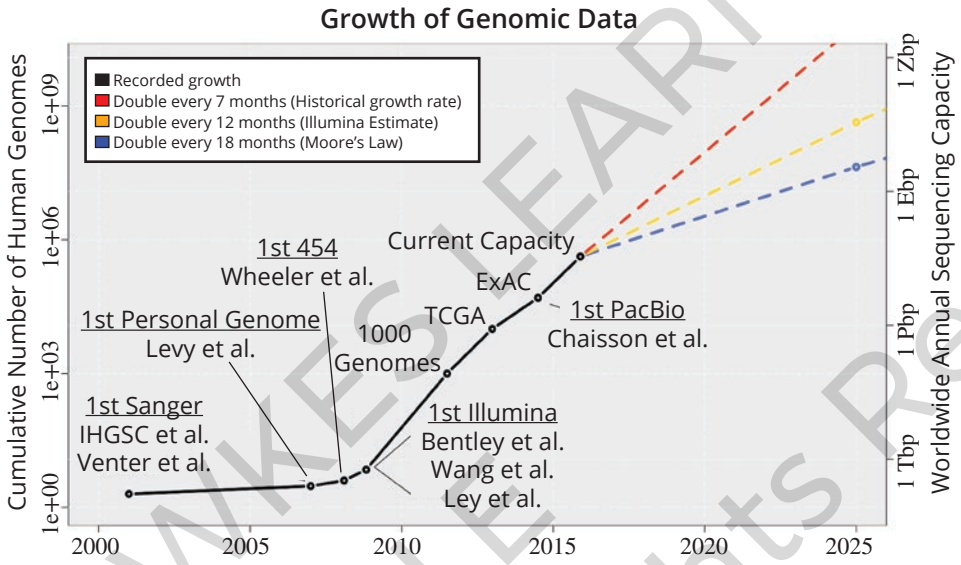


Figure 1.5.2

Astronomy and Cosmology

The Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000. The telescope collects data at the rate of about 200 gigabytes per night. In its first few weeks of operation, it collected more data than all the data collected in the history of astronomy.

The Large Synoptic Survey Telescope (LSST) being built in Chile is the successor to SDSS. When it goes fully online in 2018, it will acquire 15 terabytes of data per night or 1.28 petabytes annually. That is, each night it acquires 75 times more data than the SDSS. The data will be images and will be analyzed by computer programs that require massive amounts of computing power.

The James Webb telescope is scheduled to be launched in 2018. It will operate 1,000,000 miles from earth and will be 5 times more powerful than the Hubble telescope. The James Webb telescope will be able to directly image exoplanets of nearby stars. If the downlinks on the satellite work perfectly for 10 years, it will generate 209 terabytes of data over its life.

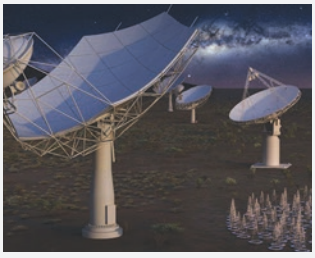
There are plant and animal species that have genomes that dwarf the human genome in size.

The marbled lungfish genome contains 133 billion base pairs. It is the largest animal genome. The largest plant genome is a rare Japanese flower named Paris japonica. It has 149 billion base pairs, making it 50 times the size of a human genome—and the largest genome ever found.

As of 2017, one of the largest genomes ever sequenced is the loblolly pine. Its genome contains 22 billion base pairs, although much of it is repetitive.

In October of 2017 a group of six scientists sequenced the bread wheat genome. Because the genome contains six copies of each chromosome, the genome has approximately 16 billion base pairs.

https://commons.wikimedia.org



The Square Kilometer Array

The Square Kilometer Array, or SKA, is the name of what will soon be the world's largest radio telescope. It is a global scientific endeavor that will generate a tremendous amount of data. When fully operational in the 2030s, SKA will produce several petabytes of data per second. In terms of annual data production, this is anywhere from 50–100 times the annual global internet traffic in 2016. In order for data of this quantity and magnitude to be collected and processed, the central supercomputer requires the processing power of 100,000,000 average 2017 personal computers. Getting the data to the supercomputer in a timely and efficient manner requires enough fiber optic cable to wrap around the equator twice. You may be asking yourself, "What do we get for all of this effort?" SKA will provide us with the ability to detect electromagnetic output from extreme distances. The telescope will be so sensitive that it will be able to detect an airport's radar system from tens of light years away. This level of detail will help test our understanding of fundamental physics, find other habitable planets, and search for extraterrestrial life. The quest for knowledge, and the desire to drive science ever forward is increasingly being driven by data.

NASA has 100 active missions. In the time it took to read the previous sentence NASA downloaded 1.73 gigabytes of data from its missions. The rate of NASA's data gathering is growing exponentially. NASA has plans for missions that will stream 24 terabytes a day.

But the mother of all telescopes is the forthcoming $2.1 billion Square Kilometre Array (SKA) radio telescope. When it is completed in 2020, its designers believe it could generate more data in one day than the entire internet in one year. Further, it will be 10,000 times more powerful than any other telescope. Given the amount of data it produces, SKA will require three times the computing power of the world's largest supercomputer in 2017.

Physics

The Large Hadron Collider (LHC) is a 17-mile ring filled with superconducting magnets that send protons in opposite directions at nearly the speed of light, only to have them smash into one another. It has an annual budget of over one billion dollars and cost many billions to build. The LHC is regarded as the pinnacle of modern science.

The biggest finding from the LHC thus far has been the discovery of the Higgs Boson, a particle predicted by the standard model of physics but never shown to exist. The LHC has an amazing 150 million sensors that deliver data at an incredible 14 million times per second. Inside the accelerator there are 600 million collisions per second. Since only a few of the collisions are of interest, the LHC "only" stores about 25 petabytes of data per years. As of 2017 the analysis of the LHC data is done on a computing grid with 500,000 processors and 500 petabytes of storage.

Sources of Big Data in Business and Industry

In business and industry, most large machines have sensors that monitor system components many times per second:

- General Electric (GE) manufactures a gas turbine with 200 embedded sensors which generate about 600 gigabytes of data per day. One gas turbine would generate 219 terabytes of data in a year. GE is the largest producer of gas turbines in the world with more than 10,000 gas and steam turbines operating throughout the world. Assuming all these turbines have the same number of embedded sensors, the data generated by all these machines would be on the order of 2 exabytes annually.
- GE also produces aircraft engines that generate 10 data points per second on 1000 parameters (sensors). On a flight from New York to London one of these engines would generate about 8 gigabytes of data. One model of GE aircraft engine is used on approximately 2000 Boeing 737 aircrafts. Most commercial aircraft fly about 3000 hours per year. At two engines per aircraft, the Boeing 737 aircraft fleet generates about 92 petabytes of data per year.
- A modern commuter train's sensors will collect and send 9,000,000 data points per hour.
- A smart energy meter could send 35 gigabytes of data per day.
- Modern buildings are full of sensors that monitor sound, temperature, humidity, and motion.

- Computer logs monitor and diagnose computer system problems. Even a 50 server data center will generate 100 gigabytes of log data per day.
- Worldwide there are about 100 billion credit card transactions per year. Building fraud prevention models for credit cards has become a Big Data problem.
- By the year 2020, 50 billion machines are expected to be connected to the internet.

In industrial applications—like gas turbines, aircraft engines, and train motors—sensor data is used to determine an optimal operations strategy and to detect the root cause of failures and defects in near real-time. Companies also use sensor data to look for correlations among variables that may signal a design improvement in the system.

In 2016, a company developed a model to forecast corn yield per acre. It was an unusual model because the model used one petabyte of satellite imagery data that was run through a cluster of 30,000 computer processors to predict average corn yields per acre for 2016. Interestingly, the satellite image model predicted an average corn yield of 169 bushels per acre yield for 2016. The US Department of Agriculture (USDA) predicted 175.1 bushels per acre and the actual corn yield for 2016 was 178 bushels per acre. This is a case where Big Data modeling is not always good. But the accuracy of this kind of modeling will undoubtedly improve.

Satellite image data is also being used to produce revenue forecasts for large box retailers (e.g., Walmart, Target). Using satellite imagery, computer programs count cars in these retailers' parking lots every day and connect this data with quarterly revenue estimates.

1.5 Exercises

Basic Concepts

1. What is Big Data?
2. What are the four attributes of Big Data?
3. List two sources of Big Data in science.
4. List two sources of Big Data in business and industry.

1.6 Introduction to Statistical Thinking

What methods do statisticians use to make predictions? The most difficult part of the process is finding a sample that accurately reflects the larger group under study. In statistics the group we wish to study is called the **population**.