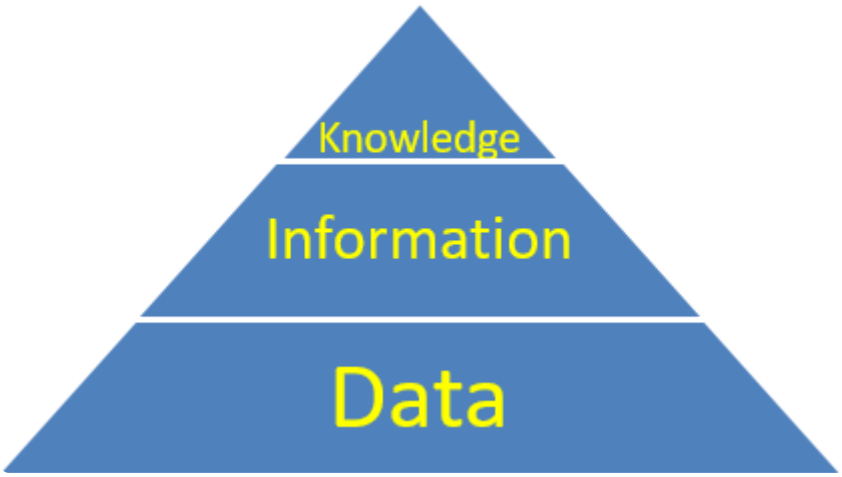
UNIVERSITY OF MARYLAND
University College

Learning Resource

Data and Information



Data Pyramid

The Power of Data

What Is Data?

Data is a fact or set of facts that have been gathered about an object, idea, place, person, etc. The facts are stored or represented in the form of numbers, measurements, words, descriptions, or observations. Note that a single fact is actually a **datum**, and data is the plural form. Conventionally, the plural form (data) is used for singular and plural purposes, and we will do that here.

As noted in the definition above, data can be represented in different forms. Here are several types of data (Pierce, 2017):

- **Qualitative data:** descriptive data that includes such facts as color, texture, feel, description of an experience, perceptions of strengths or weaknesses, etc. This is data to which numbers are not normally assigned.
- **Quantitative data:** facts which are presented as numbers such as test scores, number of students in the class, number of words on a page, capacity of a hard drive. Quantitative data can also be subdivided into discrete and continuous data.

- **Discrete data** can only be assigned a certain value, such as whole numbers. For example, there are 32 students in the class, the hard drive can store eight Gigabytes, the test score was 89 percent.
- **Continuous data** reflects a range into which the values may fall. Optimal tire pressures may fall anywhere between 30 PSI (pounds per square inch) and 33 PSI, including any fractional pressure in between these values.

- **Categorical data:** groups the facts into a category such as "new" or "used" or "for sale" or "not for sale," etc.

Why Is Data Collected?

Typically, data is collected to tell a story or solve a problem. Beginning with a question that needs to be addressed focuses both the type of data to be collected and the follow-on review of what is gathered, perhaps providing an answer to the question, or revealing patterns, or uncovering unusual results that were not expected. There may be interesting results hidden in the facts gathered. But the question that drives the collection of data also helps identify the audience who will be the recipients of your findings or the story you want to tell (School of Data, 2013).

How Is Data Collected?

There are many ways for collecting data—direct observation (counting people in the coffee shop), a census (all items or individuals in the group are measured), or a sample (selected items or individuals in the group are measured), physical measurements taken by persons or machines, interviews, etc.

In much broader terms, the basic data sources are:

- collecting data yourself
- finding data that has already been collected and released for others to use
- getting additional data by asking sources for updates, or by getting access to data that is typically hidden from public use

This last list of "hidden" data sources includes the government, organizations, and scientific projects and institutions. Two great places where individuals can find data are projects such as Open Access Directory's data repository (http://oad.simmons.edu/oadwiki/Data_repositories) and Open Knowledge Foundation's datahub.io (https://datahub.io/dataset).

In What Format Is Data Collected?

The purpose for gathering data is to tell that story or answer a question. But in order to do that, the data has to be in a format that allows the data to be analyzed. Outside of simply eyeballing the data or using paper and pencil, the best format to use with computer analysis tools (such as Excel) is to obtain the data in machine-readable form—that is, in a form such that the data can be imported into a computer program. The most common format for exchanging or importing data is in comma separated values (CSV). The data pieces, whether words or numbers, are separated by commas, and the data can be read directly into a spreadsheet program.

Big Data

The term "big data" became mainstream in the early 2000s via the work of industry analyst Doug Laney. His definition of the term incorporates the following (SAS, n.d.):

- **Volume:** Organizations collect data from a variety of sources, including business transactions, social media, and information from sensor or machine-to-machine data. In the past, storage would have been an issue, but new technologies have helped..
- **Velocity:** Data streams into the data center at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors, and smart metering drive the need to deal with torrents of data in near-real time.
- **Variety:** Data comes in all types of formats—from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

The SAS Institute Inc.(SAS) adds two additional dimensions when it comes to big data (SAS, n.d.):

- **Variability:** In addition to the increasing velocities and varieties of data, data flows can be inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal, and event-triggered peak data loads can be challenging to manage.
- **Complexity:** Data comes from multiple sources, which makes it difficult to link, match, cleanse, and transform data across systems. Connecting relationships, hierarchies, and multiple data linkages is important.

What Is the Importance of Big Data?

It is actually not the amount of data that is important but what is done with that data. Particularly in the business environment, analysis of that data can find answers to questions about potential reductions in cost and time, and help in making smart decisions about new product development (SAS, n.d.). Other critical business tasks can be supported by using the gathered data to determine what has caused failures, issues, and/or defects or

detecting and mitigating fraudulent behavior before the organization is severely affected (SAS, n.d.). If done properly, data collection (and analysis) will allow a business to focus time, personnel, and resources on the issues that will generate the greatest returns.

What Are the Sources for Big Data?

Where Does Big Data Come From?

Source: IBM Big Data and Analytics Hub

The three most prevalent sources for large amounts of data are the following:

- Streaming data that comes from the information infrastructure within an organization. For example, all transactions accomplished via the IT systems within the business are captured on a daily or even hourly basis. This includes logs of daily activities and email or other types of messages received from internal sources.
- Social media data, including audio, photo, and video files that are retrieved from watching activity on Facebook, the business's website, or websites of related businesses or competitors. This can aid marketing, sales, and customer support functions.

- Publicly available sources, such as data.gov, the CIA World Factbook, or the European Union Open Data Portal. A browser search of "sources for data sets" will provide a long list of sources for data that addresses many areas of interest.

Sampling, Surveys, and Polls

When researchers are looking to collect data about a particular topic that affects a large group or even the entire population, it is not cost-effective or even practical to contact every member of the group or the population for data input. Instead, most such studies are based on gathering responses from a **sample,** or a subset of the entire population. Although everyone in the population is not individually contacted, the results of sampling are considered to be representative of the population.

In order for the sampling to be truly representative, it is critical that the sample subset be representative of the large group. Randomly selecting the subset of participants is the primary way of guaranteeing that anyone could have been selected. "The basic principle: If selected correctly, a randomly selected small sample of a population of people can represent the attitudes, opinions, or projected behavior of all of the people from which the sample is obtained" (Newport, Saad, & Moore, 1997).

Whereas sampling is the method for creating the pool of persons to be contacted, surveys and polls are the means by which the data is collected from the sample.

What is the difference between a survey and a poll? Both may use sample sets of participants that represent the group that is being surveyed or polled. A poll typically asks one question while a survey is generally used to ask a range of questions.

Here is an example of a poll question—one multiple-choice question and a list of answers from which the participant selects one or more answers (including "Other," which allows the participant to enter an answer not in the list).

What is your favorite color?	
O	Red
O	Orange
O	Yellow
O	Green
O	Blue
O	Indigo
O	Violet
O	Other _____

Sample Poll

A survey, on the other hand, allows for asking more than one question that covers a wider area of interest. And it allows for different types of questions and/or responses, including such things as age, address, as well as multiple-choice questions. Customer satisfaction surveys are one common form of a survey. And students are asked to complete a course evaluation survey in the latter weeks of each course at UMUC.

Here is a sample survey based on a Likert scale (ranking the response using values 1-5). There are four questions, making this a survey and not a poll, which typically consists of a single yes/no/uncertain.

Mythical Unicorns just sold you a T-shirt. Check the response that best matches your satisfaction level with this product.

	Strongly Disagree				Strongly Agree
	5	4	3	2	1
The item was as specified	O	O	O	O	O
The size was as expected	O	O	O	O	O
The material is as expected	O	O	O	O	O
I would recommend this item to others	O	O	O	O	O

Two of the most familiar polling companies are Gallup and Nielsen. Gallup's method of selecting polling participants is to generate a list of all phone numbers (landline and cell phone) in the United States and then use a subset of that list, which covers all geographical areas based on area codes, to call and interview individuals.

Nielsen uses a slightly different approach. Households that participate are selected at random from a predefined sample based on census data. The census data provides critical information on household income, size, age of residents, etc. A certain number of houses from each group is selected.

The sample size in any poll is critical in meeting validity criteria for poll results. However, simply increasing the size of the sample group does not equate to increased validity. Gallup and other major polls use sample sizes of between 1,000 and 1,500 for standard surveys of the US population "because they provide a solid balance of accuracy against the increased economic cost (Newport, Saad, & Moore, 1997).

Nielsen's TV ratings work on the same principle. Nielsen gets around 5,000 households to agree to be part of the representative sample to find out who is watching TV and what those people are watching. To be accurate, that sample set of 5,000 households needs to be representative of all U.S. households with TVs (How Stuff Works, Entertainment, n.d.).

Meters are installed in the home, and these meters can track when TV sets are on and the channels that are being watched. Data gathered by these boxes is then sent to the company each night. Nielsen than compares the data received with the programs that are on TV at any time, and thus determines how many people watch which program.

> *This research is worth billions of dollars. Advertisers pay to air their commercials on TV programs using rates that are based on Nielsen's data. Programmers also use Nielsen's data to decide which shows to keep and which to cancel. A show that has several million viewers may seem popular to us, but a network may need millions more watching that program to make it a financial success. That's why some shows with a loyal following still get canceled* (How Stuff Works, Entertainment, n.d.).

Who Uses Big Data?

Big data plays a role in almost every industry. SAS (n.d.) provides a summary of some of the industries affected by big data.

Banking	Understanding customers and customer satisfaction, minimizing risk and fraud while maintaining regulatory compliance

Education	Impact school systems, students, and curriculum by identifying at-risk students, ensuring adequate student progress, and implementing systems to evaluate and support teachers and principals
Government	Managing utilities, running agencies, dealing with traffic congestion, preventing crime. Governments must also address issues of transparency and privacy.
Health Care	Respecting privacy as it relates to patient records, treatment plans, prescription information while at the same time uncovering insights into improving patient care
Manufacturing	Boost quality and output while minimizing waste; support for more agile business decisions.
Retail	Building customer relationships, marketing, handling transactions, revitalizing business

The Power of Information

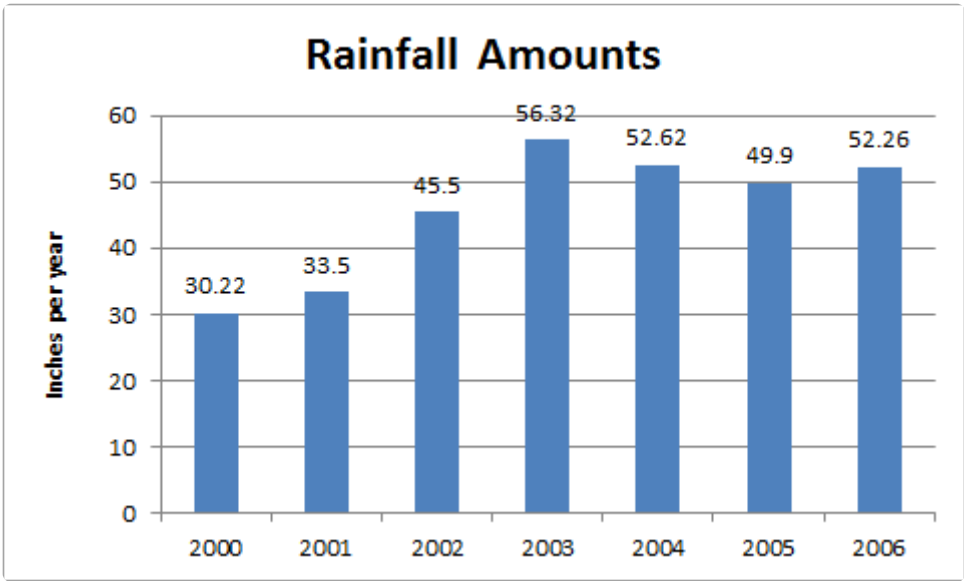
So now you have the data. What do you do with it? If all you have is a set of random numbers, they tell you nothing until you also know the context in which these numbers were gathered. For example, if you were given 32 random numbers, some repeated, between 18 and 95, they would be meaningless until you were also told that these numbers were the ages of the students in your course. Until you know the context, the data by itself only provides you with the foundation for eventually organizing the data in such a way as to provide you with the information needed to find answers to the questions or tell the story. How is that organization done? How is data transformed into information?

Data, as you know, comes in many forms—numbers, words, pictures. In an example, we will use a set of numerical, discrete data.

Here is the data set:

30.22, 35.5, 45.5, 56.32, 52.62, 49.90, 52.26

This raw data is useless as displayed here. However, once the context is included (the question), the data now can result in useful information. The question was: What was the annual precipitation in Reading, Pennsylvania between 2000 and 2006? Now you have data you can work with. The data can be transformed into useful information by importing it into a spreadsheet and displaying results of minimum and maximum values, or creating a picture or graph of the same results.



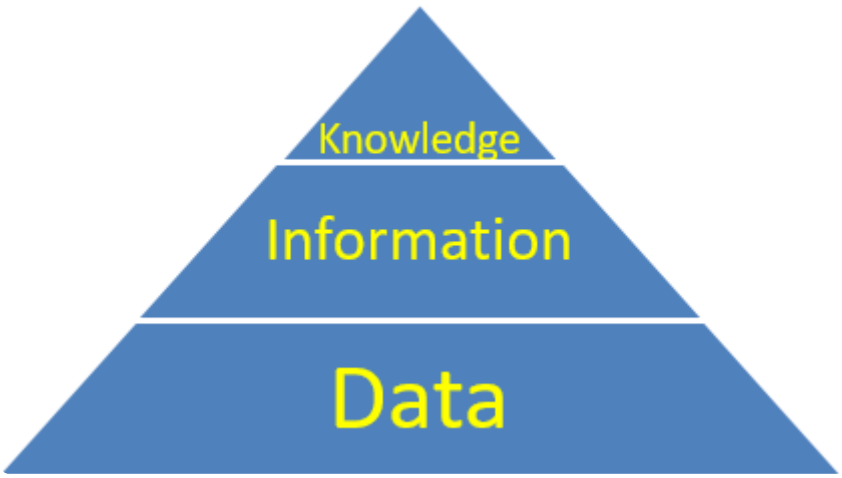
Rainfall Amounts

In such a small data set, you can easily pick out the minimum and maximum rainfall. But in a data set that covers 1863 to 2006 (123 years), it would be more difficult. This is where the analysis functions in spreadsheet programs become helpful in organizing the data and providing you with usable information about the topic or question.

One caveat—the value or correctness of your information is dependent upon the correctness of the data you use to generate that information. An old adage that applied to computer programs applies here, as well: "Garbage in; garbage out." If your raw data is bad, your answer to the question, or the resulting story you tell may be flawed as well.

In the tutorials, you will work with some of the basic functions that spreadsheet applications provide. Although we will work exclusively with Microsoft Office's Excel program, there are other spreadsheet apps available, some created specifically for work with big data and those that provide more analysis functionality than that provided by Excel. See "List of Spreadsheet Software" on Wikipedia for a list of free and proprietary spreadsheet apps. But we will be using Excel for tutorials and exercises (projects) in this course.

How Is Information Used?



Data Pyramid

Information is the next step up in the information theory pyramid. It is the foundation for knowledge. But in a more practical way, information provides the basis for making decisions. It is the next step in answering the question or telling the story.

Making decisions is a part of everyday life—from what to prepare for dinner, to more life-changing decisions such as where to live, whom to marry, or what to choose as a course of study in college. To make the best decisions, it is important to gather the relevant information. You can delay making a decision if all you do is endlessly search for information without coming to any conclusion. Or you could take a vote, throw a dart at a list, or toss a coin.

However, it is possible that an inability to make a reasoned decision is because there is not enough information or too much information. Even if your information is on target, if you involve too many others in the decision process, the need to include everyone's views and values may end up being too complicated.

If the decision involves change, that potential movement in the status quo may make the solution too difficult to accept. Finally, if you just don't care about the outcome, one way or the other, it may be hard to invest the effort needed to come to a conclusion. Regardless of the outcome, information is gathered for a reason (Skills You Need, n.d.)

Here is another way to look at the uses of information. These uses are tied very closely to a "need" that has been identified (that question or story) for gathering information. As such, the list does not directly indicate **how** the information is used, but **why** it was gathered. The assumption may be made that the information is then **used** to address the issue (Taylor, 1991):

- Enlightenment: context information
- Problem understanding: better comprehension of a specific problem

- Instrumental: what to do and how to do something
- Factual: precise data
- Confirmational: verify a piece of information
- Projective: future oriented
- Motivational: relates to personal involvement
- Personal or political: relationships, statue, reputation, personal fulfillment

Summary

Data does not depend on information, but information depends on data. Raw data by itself has no meaning. Information results when context or meaning is added to the raw data, resulting in at least the first level of understanding the answer to whatever question prompted the gathering of that data. Here are some properties of data:

- Data can be stored, copied, duplicated, modified, and/or moved.
- Data remains static—it does not necessarily improve over time; rather, data can decay as it becomes outdated or is no longer applicable to the question being asked.
- Data has no value until it is converted into usable information ("Value" here only refers to the fact that, standing alone, raw data does not tell a story or answer a question. The data itself may have great "value," financial or otherwise, to the person or entity that seeks to use that data).
- Data that is incorrect or used outside of the context for which it was gathered may result in incorrect information.

Information, on the other hand (Doyle, 2014):

- results when context is added to data—what, when, where, why, how the data was collected
- is data that has been converted into a form that makes understanding of the data useful; it is data with meaning
- becomes the basis for understanding a question, or making inferences, or making decisions; it helps tell the story.

We have begun with an overview of the first two elements in the information pyramid: data and information. Readings in the following weeks will focus on knowledge, knowledge management, and business intelligence.

References

Doyle, M. (2014, August 6). What is the difference between data and information? [Blog post]. Retrieved from https://salespop.pipelinersales.com/sales-management/difference-between-data-and-information/

How Stuff Works - Entertainment. (n.d.) How do television ratings work? Retrieved from http://entertainment.howstuffworks.com/question433.htm

IBM Big Data and Analytics Hub. (n.d.). Where does big data come from? Retrieved from http://www.ibmbigdatahub.com/infographic/where-does-big-data-come

Newport, F., Saad, L., & Moore, D. (1997). How are polls conducted? In M. Golay, *Where America stands*. John Wiley & Sons.

Pierce, R. (2017, February 15). *Data, probability, and statistics*. Retrieved from https://www.mathsisfun.com/data/index.html

Quantitative Environmental Learning Project. (n.d.). DataSet#049; Reading, PA precipitation. Retrieved from http://seattlecentral.edu/qelp/sets/049/049.html

SAS. (n.d.). Big data: What is it and why it matters. Retrieved from http://www.sas.com/en_th/insights/big-data/what-is-big-data.html

School of Data. (2013). What is data? Retrieved from https://schoolofdata.org/handbook/courses/what-is-data/

Skills You Need. (n.d.). Decision making. Retrieved from https://www.skillsyouneed.com/ips/decision-making.html

Taylor, R. (1991). Information use environments. In B. Dervin & M. J. Voight (Eds.), *Progress in communication science.* Norwood, NJ: Ablex.