1.1 What Are Data?

Businesses have always relied on data for planning and to improve efficiency and quality.

Most modern businesses collect information on virtually every transaction performed by the organization, including every item bought or sold.

These data are recorded and stored electronically, in vast digital repositories called data warehouses.

Data collected for recording the companies’ transactions is called transactional data.

The process of using transactional data to make other decisions and predictions, is sometimes called data mining or predictive analytics.

Business analytics describes any use of statistical analysis to drive business decisions from data.

All data have a context.

Data values or observations are information collected regarding some subject. The “Five W’s”: who, what, when, where, and (if possible) why. Often we add how to the list.

Data can be numbers, names, etc., and tells us the “Who and What”.

Data are often organized into a data table like that below.

The rows of a data table correspond to individual cases about whom we record some characteristics.

These characteristics may be collected on or about...

- respondents – individuals who answer a survey
- subjects or participants – people in an experiment
- experimental units – animals, plants, websites, or other inanimate objects

The huge capacity of these warehouses has given rise to the term Big Data to describe data sets so large that traditional methods of storage and analysis are inadequate.

In the past few decades these data warehouses have grown enormously in size, but with the use of powerful computers, the information contained in them is accessible and used to help make decisions.
1.1 What Are Data?

The characteristics recorded about each individual or case are called variables.

These are usually shown as the columns of a data table and identify what has been measured.

Variables

<table>
<thead>
<tr>
<th>Order-Number</th>
<th>Name</th>
<th>Total-Amount</th>
<th>Price</th>
<th>Area Code</th>
<th>Album-Download</th>
<th>Year</th>
<th>Track #</th>
<th>Artist</th>
</tr>
</thead>
<tbody>
<tr>
<td>013-020-0001</td>
<td>John Doe</td>
<td>123-456-7890</td>
<td>$100</td>
<td>123-456-7890</td>
<td>ABCD</td>
<td>2015</td>
<td>123456</td>
<td>ABC</td>
</tr>
</tbody>
</table>

1.1 What Are Data?

Metadata typically contains information about how, when, and where (and possibly why) the data were collected; who each case represents; and the definitions of all the variables.

Data are typically saved in a spreadsheet, where the rows represent cases and the columns represent variables.

1.1 What Are Data?

Data tables are cumbersome for complex data sets, so often two or more separate data tables are linked together in a relational database so that information can be merged across them.

Each data table included in the database is a relation because it is about a specific set of cases with information about each of these cases for all the variables.

1.1 What Are Data?

Example: Credit Card Bank

Carly, a marketing manager at a credit card bank, wants to know if an offer mailed 3 months ago has affected customers’ use of their cards. To answer that, she asks the information technology department to assemble the following information for each customer:

- Total spending on the card during the 3 months before the offer (Pre Spending)
- Total spending for 3 months after the offer (Post Spending)
- The customer’s Age (by category)
- That kind of expenditure they made (Segment)
- If customers are enrolled in the website (Enroll?)
- What offer they were sent (Offer)
- The amount each customer spent on the card in their segment (Segment Spend)

Identify the cases and the variables. Describe as many of the W’s as you can for this data set.

1.1 What Are Data?

Example: Credit Card Bank

The cases are individual customers of the credit card bank.

The data are from the internal records of the credit card bank for the past 6 months (3 months before and 3 months after an offer was sent to the customers).

The variables include the account ID of the customer (Account ID) and the amount charged on the card before (Pre Spending) and after (Post Spending) the offer was sent out.

Also included are the customer’s Age, marketing Segment, whether they enrolled on the website (Enroll?), what offer they were sent (Offer), and how much they charged on the card in their marketing segment (Segment Spend).
1.2 Variable Types

When a variable names categories and answers questions about how cases fall into those categories, it is called a **categorical** or **qualitative variable**.

When a variable has measured numerical values with **units** and the variable tells us about the quantity of what is measured, it is called a **quantitative variable**.

---

1.2 Variable Types

Categorical variables...

- arise from descriptive responses to questions like “What kind of advertising do you use?”.
- may only have two possible values (like “yes” or “no”).
- may be a number like a zip code.

Quantitative variables must have **units**. The units indicate...

- how each value has been measured.
- the corresponding scale of measurement.
- how much of something we have.
- how far apart two values are.

---

Identifiers

An **identifier variable** is a unique identifier assigned to each individual or item in a group.

For example, social security numbers, student ID numbers, tracking numbers, transactions numbers, etc. are all identifier variables for people or items.

---

Some variables can be both categorical and quantitative. How data are classified depends on **Why** we are collecting the data.

For example, variable **Age** is obviously the quantitative value, measured in years, that may be used for finding the average age of customers.

But, **Age** can also be the categorical value, e.g. child, teen, adult, or senior, used to classify books for an internet store.
1.2 Variable Types

Other Data Types

Categorical variables used only to name categories (that don’t have order) are sometimes called *nominal variables*.

When data values can be ordered, we say that the variable has *ordinal* values. For example, employees can be ranked according to the number of months employed.

Example: Credit Card Bank

Account ID – categorical (nominal, identifier)
Pre Spending – quantitative (units $)
Post Spending – quantitative (units $)
Age – categorical (ordinal). Could be quantitative if we had more precise information
Segment – categorical (nominal)
Enroll? – categorical (nominal)
Offer – categorical (nominal)
Segment Spend – quantitative (units $)

The data are cross-sectional. We do not have successive values over time.

1.2 Variable Types

Cross-Sectional and Time Series Data

Variables that are measured at regular intervals over time are called a *time series*. Typical measuring points are months, quarters, or years.

When several variables are all measured at the same time point, the data is called *cross-sectional data*. For example, data on sales revenue, number of customers, and expenses for last month at each Starbucks (more than 20,000 locations as of 2012) at one point in time would be cross-sectional data.

Example: Credit Card Bank

Before she can continue with her analysis, Carly must classify each variable as being quantitative or categorical (or possibly both), and whether the data are a time series or cross-sectional.

For quantitative variables, what are the units?
For categorical variables, are they nominal or ordinal?

1.3 Data Sources: Where, How, and When

When data are collected can be important. Data that are decades old may mean something different than similar values recorded last year.

How the data are collected can make the difference between insight and nonsense. To make inferences from the data you have at hand to the world at large, you need to ensure that the data you have are representative of the larger group.

Where data are collected can be important. Data collected in Mexico may differ in meaning than data collected in the United States.

Data can be found …

- by performing an experiment and actively manipulating variables.
- in information collected by public or private agencies.
- on internet sites.
Example: Credit Card Bank

On the basis of her initial analysis, Carly asks her colleague Ying Mei to e-mail a sample of customers from the Travel and Entertainment segment and ask about their card use and household demographics.

Carly asks another colleague, Gregg, to design a study about their double miles offer.

In this study, a random sample of customers receives one of three offers:

- The standard double miles offer
- A double miles offer good on any airline
- No offer

Question:

For each of the three data sets—Carly’s original data set and Ying Mei’s and Gregg’s sets—state whether they come from a designed survey or a designed experiment or are collected in another way.

Answer:

Carly’s data set was derived from transactional data, not part of a survey or experiment.

Ying Mei’s data come from a designed survey.

Gregg’s data come from a designed experiment.

What Have We Learned?

Understand that data are values, whether numerical or labels, together with their context:

- who, what, why, when, where, when (and how)—the W’s—help nail down the context of the data.
- We must know who, what, and why to be able to say anything useful based on the data. The who are the cases. The what are the variables. A variable gives information about each of the cases. The why helps us decide which way to treat the variables.
- Stop and identify the W’s whenever you have data, and be sure you can identify the cases and the variables.

What Can Go Wrong?

Don’t label a variable as categorical or quantitative without thinking about the data and what they represent. The same variable can sometimes take on different roles.

Don’t assume that a variable is quantitative just because its values are numbers. Categories are often given numerical labels. Don’t let that fool you into thinking they have quantitative meaning. Look at the context.

Always be skeptical. One reason to analyze data is to discover the truth.

Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different.

The context colors our interpretation of the data, so those who want to influence what you think may slant the context.

A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website.

The question that respondents answered may be posed in a way that influences responses.
What Have We Learned?

Identify whether a variable is being used as categorical or quantitative.

- Categorical variables identify a category for each case. Usually we think about the counts of cases that fall in each category. (An exception is an identifier variable that just names each case.)
- Quantitative variables record measurements or amounts of something; they must have units.
- Sometimes we may treat the same variable as categorical or quantitative depending on what we want to learn from it, which means some variables can’t be pigeonholed as one type or the other.

What Have We Learned?

Consider the source of your data and the reasons the data were collected.

That can help you understand what you might be able to learn from the data.

The five W’s (Who, What, Why, Where, and When, and How) help nail down the context of the data.