# Introduction to Data

## Raw data-unorganized



## Who, What, When, Where, Why, and How

**Who**: The rows of a data table contains information about whom we record some characteristics. The who can be described by the following vocabulary:

 *Cases, respondents, participants, experimental units, records, or observations.*

**What and Why**: The characteristics recorded about each individual are called *variables.* Sometimes it is easier to ask “*What* has been measured?” The variables can be classified into two main groups: Categorical (those that classify by a category) and quantitative (those that have a numerical value). The *what* usually has units like feet or seconds.

*Why* isn’t always known. It will be given within the context of the problem. Although to make sense of the results, we usually need to know the why of the study.

**When, Where, and How**: If possible knowing the *when* and the *where* of data can help to make complete sense of your results.

Example: Area Codes

Location (*where*) of the data collection can reveal important facts about the *who* and *what*. Data from New Mexico and data from New York may have different meaning depending on the variables collected.

The method of collecting data (*How*) can reveal bias within data, although this is another piece of the puzzle that isn’t always known.

## Organized Data



## Assignment

In the summer of 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fasted Tour de France in its 100-year history. In 2004 he made history again by winning the race for an unprecedented sixth time. In 2005 he became the only 7-time winner and once again set a new record for the fastest average time. The following are a selection of the times of the Tour de France.

1. List as many of the W’s as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative give the units.

