Week 1: Welcome to statistics and data

2. Data basics

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Today: Data basics

2. Main ideas
    1. Identify the 5W's
    2. Understanding the data table

3. Summary

Outline

1. Today: Data basics

2. Main ideas
    1. Identify the 5W's
    2. Understanding the data table

3. Summary

Outline

1. **Data** are values, whether numerical or labels, together with their context.
2. **Observational units** are what you take measurements on
3. **Variables** are the characteristics recorded about each individual
4. **Categorical variables** identify a category for each case. They have a limited number of different values, called **levels**. E.g., Maritial status is a categorical variable, and the levels are single, married, divorced, widower, etc.
5. **Quantitative variables** record measurements or amounts of something

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

What are the observational units?

1. All patients in France
2. The French hospital
3. Patients who entered the emergency room in the previous week

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
How long the patient waits to be seen by a medical professional

1. Categorical
2. Numerical
3. Not a variable

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
Whether or not the patient has health insurance

1. Categorical
2. Numerical
3. Not a variable

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
Day of the week on which the patient arrives

1. Categorical
2. Numerical
3. Not a variable

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
Average wait time before a patient is seen by a medical professional

1. Categorical
2. Numerical
3. Not a variable

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
Whether or not wait times tend to be longer on weekends than weekdays

1. Categorical
2. Numerical
3. Not a variable

Suppose we have a data set that consists of patients who entered the emergency room at French Hospital in the previous week.

Poll question

Indicate whether the following is a categorical variable, a numerical variable, or not a variable with regard to these observational units:
Total cost of the emergency room visit

1. Categorical
2. Numerical
3. Not a variable

- What is the research question?
- What is the population of interest?
- What are the observational units?
- Name all the variables.
- Specify for each variable whether its use indicates that it should be treated as categorical or quantitative.

Tutorial exercise: 10 minutes

Finish Topic 1: online shopping
Goal: practice identifying observational units, categorical variables and numerical variables
Put 'raise your hand' button when you are done.
I'm also looking for volunteer to share their answer with the class.

Outline

Data is usually represented by a data matrix

- row: observational units
- column: variables

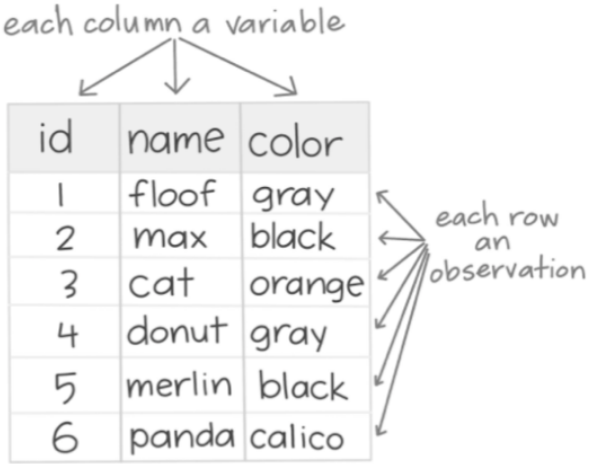
year <int>	month <int>	day <int>	dep_time <int>	dep_delay <dbl>	arr_time <int>
2013	6	30	940	15	1216
2013	5	7	1657	-3	2104
2013	12	8	859	-1	1238
2013	5	14	1841	-4	2122
2013	7	21	1102	-3	1230
2013	1	1	1817	-3	2008

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

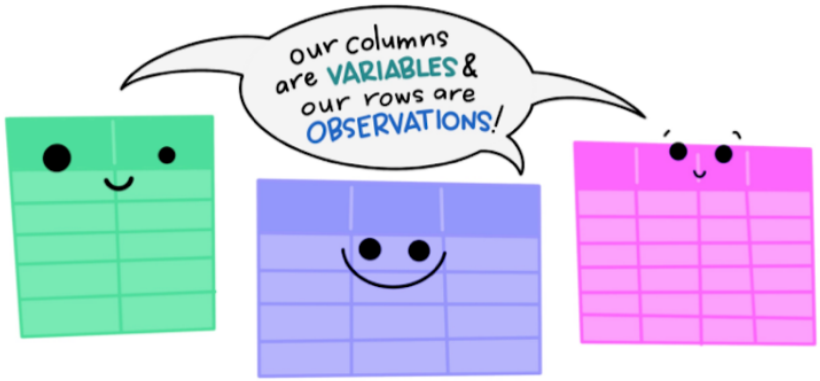
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



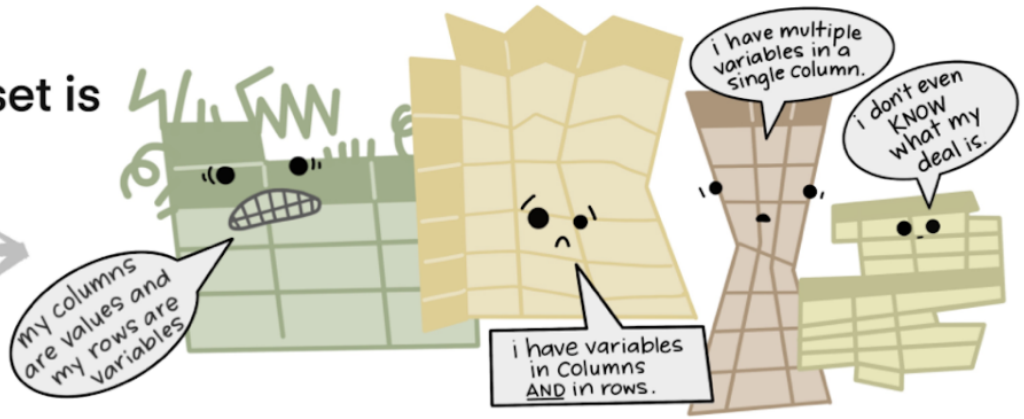
Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst

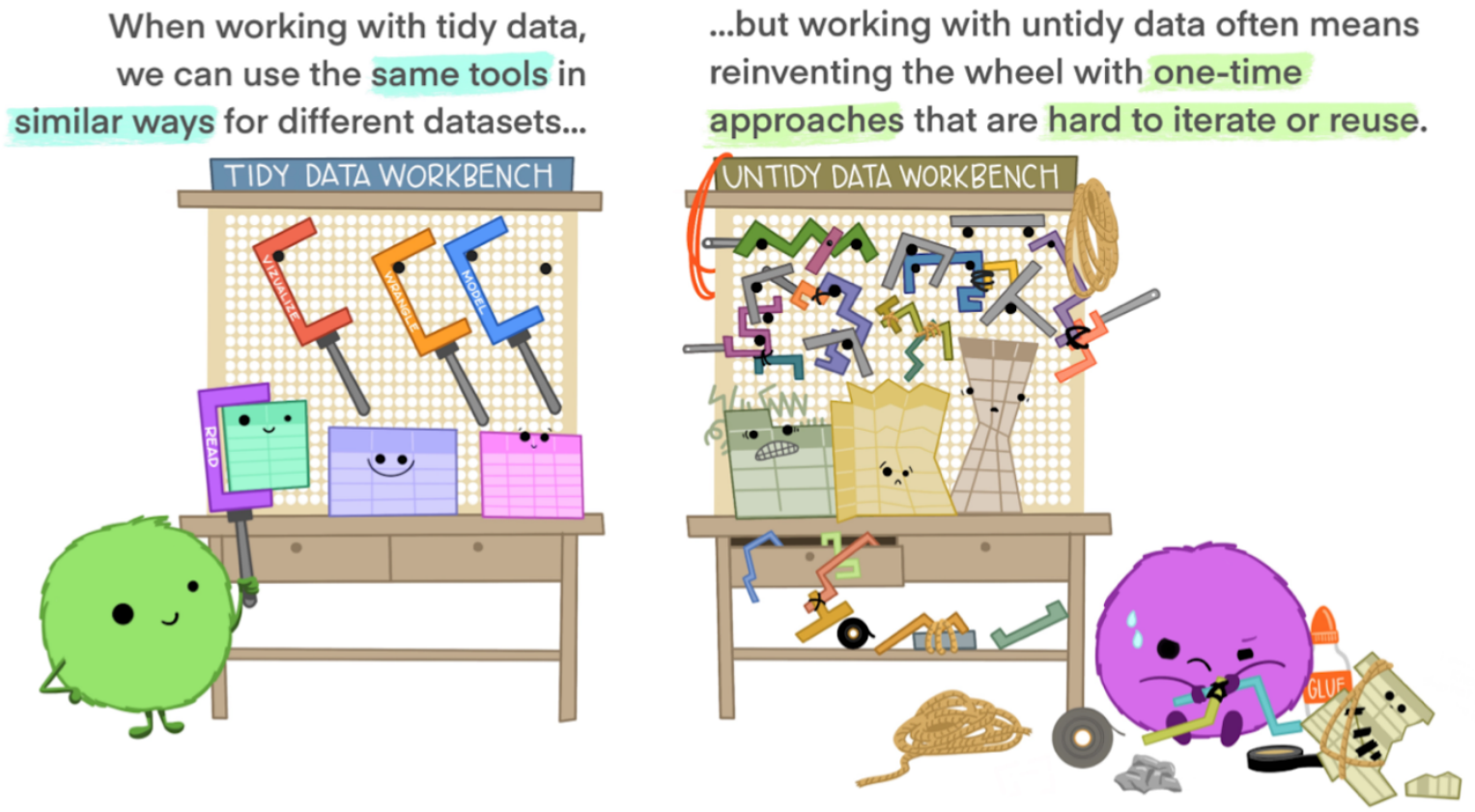
The standard structure of tidy data means that "tidy datasets are all alike..."



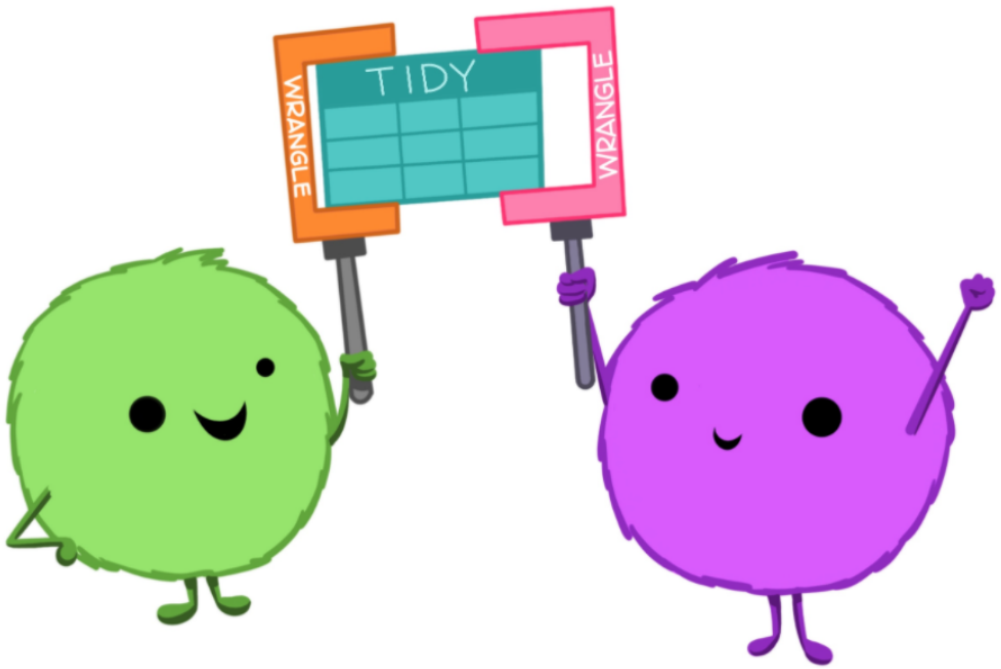
"...but every messy dataset is messy in its own way."
–HADLEY WICKHAM



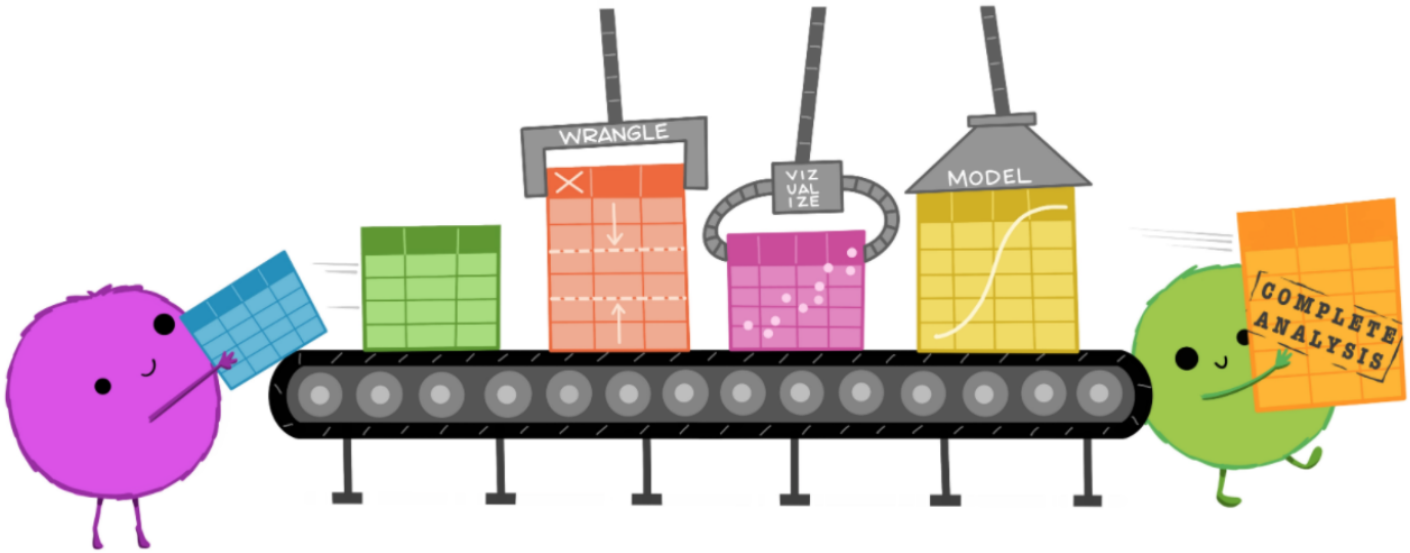
Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst



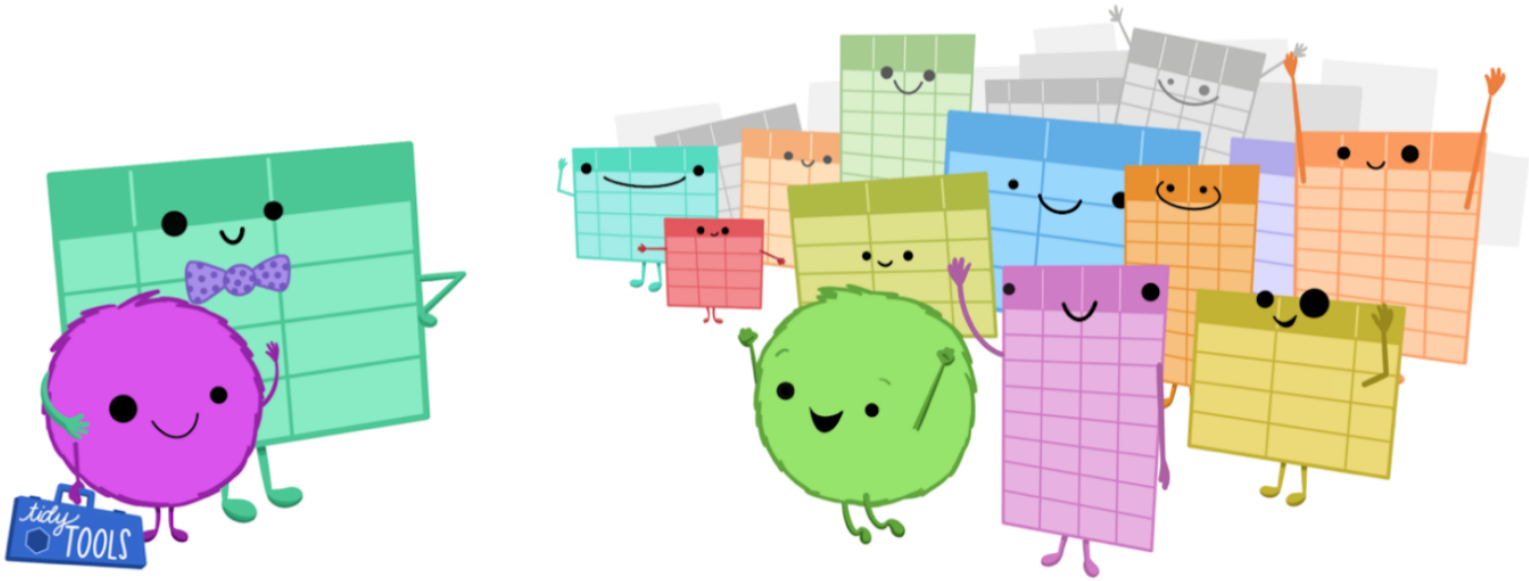
Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst



Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst



Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst



Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst

The tidyverse is a collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Tidy Data Illustrated Series
CC By Julie Lowndes Allison Horst

Some helpful R commands to have a first look of your data matrix

- head
- str
- dim
- nrow (or ncol)
- names
- $

Tutorial exercise: 15 minutes

Finish topic 2
Goal: practice using R command for data matrix/frame
Let me know if you have any questions
You are allowed to leave once you are done.

Outline

1. Today: Data basics

2. Main ideas
    1. Identify the 5W's
    2. Understanding the data table

3. Summary

1. Identify the 5W's
2. Understanding the data table