Introduction to Data Science

Grinnell College

January 20, 2026

format, syllabus, attendance, etc.,

What is Data Science?

Computer science is more than just programming; it is the creation of appropriate abstractions to express computational structures and the development of algorithms that operate on those structures. Similarly, statistics is more than just collections of estimators and tests; it is the interplay of general notions of sampling, models, distributions, and decision-making. [Data science] is based on the idea that these styles of thinking support each other (Pierson, 2016)

Data Science Model

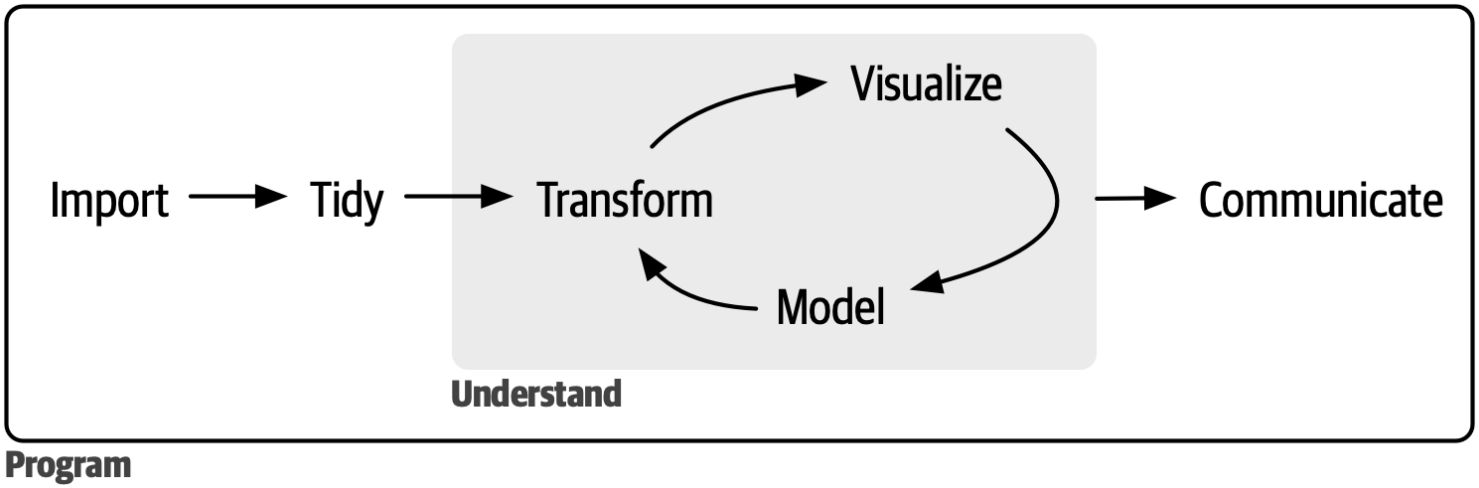


Figure 1: Source: R for Data Science (2e)

What is data?

Data represents the raw, unprocessed facts or observations that become meaningful information once collected, organized, and analyzed

Generally speaking for this class, data will have four attributes:

1. **Variables**
2. **Values**
3. **Observations**
4. **Tabular arrangement**

Data in Practice

We often uses a tabular form to store observations (rows) and variables (columns). This makes it simple to add or remove observations and variables with relative ease

Total Bill	Tip	Sex	Smoker	Day	Time	Size
13.42	1.58	Male	Yes	Fri	Lunch	2
16.27	2.50	Female	Yes	Fri	Lunch	2
10.09	2.00	Female	Yes	Fri	Lunch	2
20.45	3.00	Male	No	Sat	Dinner	4
13.28	2.72	Male	No	Sat	Dinner	2
22.12	2.88	Female	Yes	Sat	Dinner	2
24.01	2.00	Male	Yes	Sat	Dinner	4
15.69	3.00	Male	Yes	Sat	Dinner	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Data in Practice

In R, tabular data is typically stored as a `data.frame`

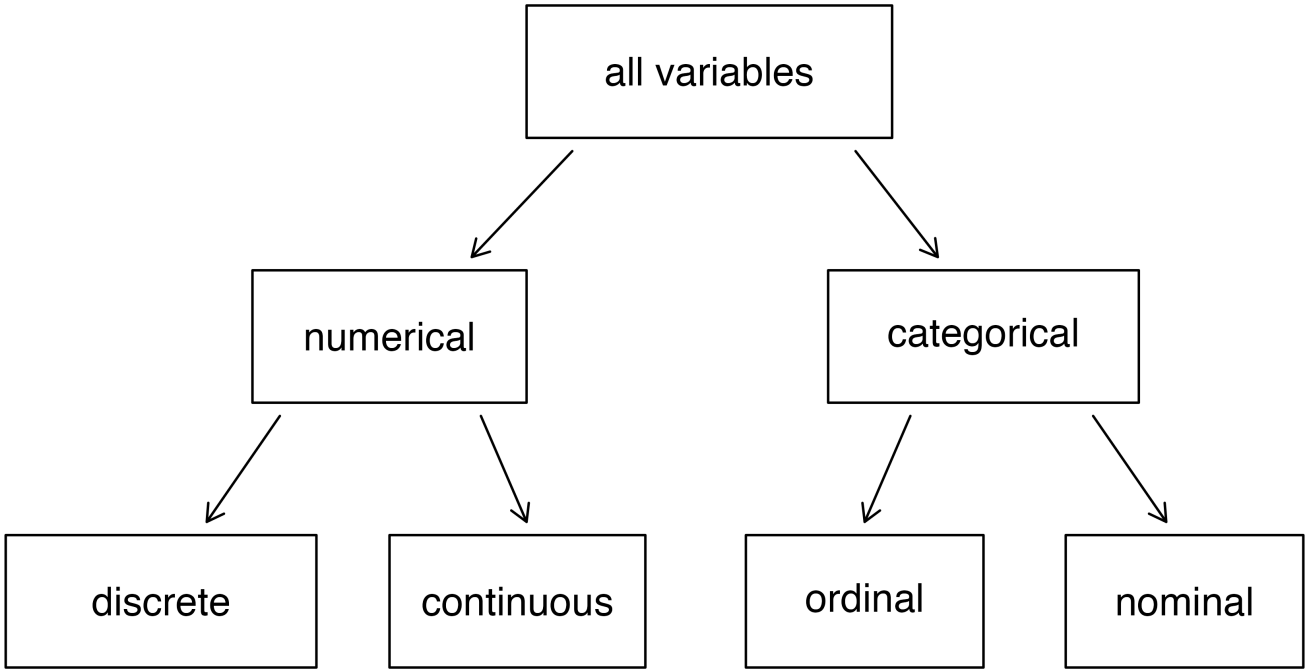
```
> tips
     total_bill  tip    sex smoker  day   time size
  1:      16.99 1.01 Female     No  Sun Dinner    2
  2:      10.34 1.66   Male     No  Sun Dinner    3
  3:      21.01 3.50   Male     No  Sun Dinner    3
  4:      23.68 3.31   Male     No  Sun Dinner    2
  5:      24.59 3.61 Female     No  Sun Dinner    4
 ---
240:      29.03 5.92   Male     No  Sat Dinner    3
241:      27.18 2.00 Female    Yes  Sat Dinner    2
242:      22.67 2.00   Male    Yes  Sat Dinner    2
243:      17.82 1.75   Male     No  Sat Dinner    2
244:      18.78 3.00 Female     No Thur Dinner    2
```

On variables

Variables typically come in one of two types:

1. **Quantitative Variable:** Data that is typically stored in the form of *numbers* and is numerical in nature
    - Continuous data e.g., height and weight
    - Discrete data e.g., points scored in a game
2. **Categorical Variable:** Variables that are naturally divided into *groups*
    - Binary
    - Nominal
    - Ordinal

Variables


all variables
numerical
categorical
discrete
continuous
ordinal
nominal

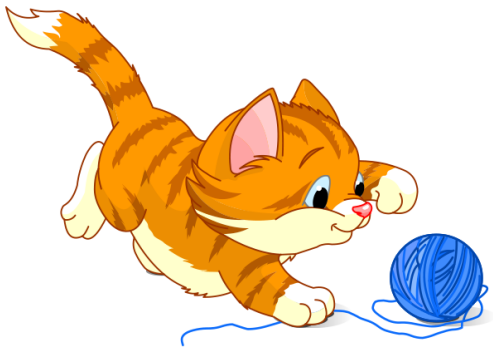
Using R Markdown

R Markdown describes a specific type of file that is used in R (.Rmd)

Uses *markdown* language to easily add headers, or write things in **bold** or *italics*

Alongside written text allows us to write and compute R code

Can be knit into pdf and submitted to gradescope



Go forth and conquer

1. Find lab on course website
2. Do it