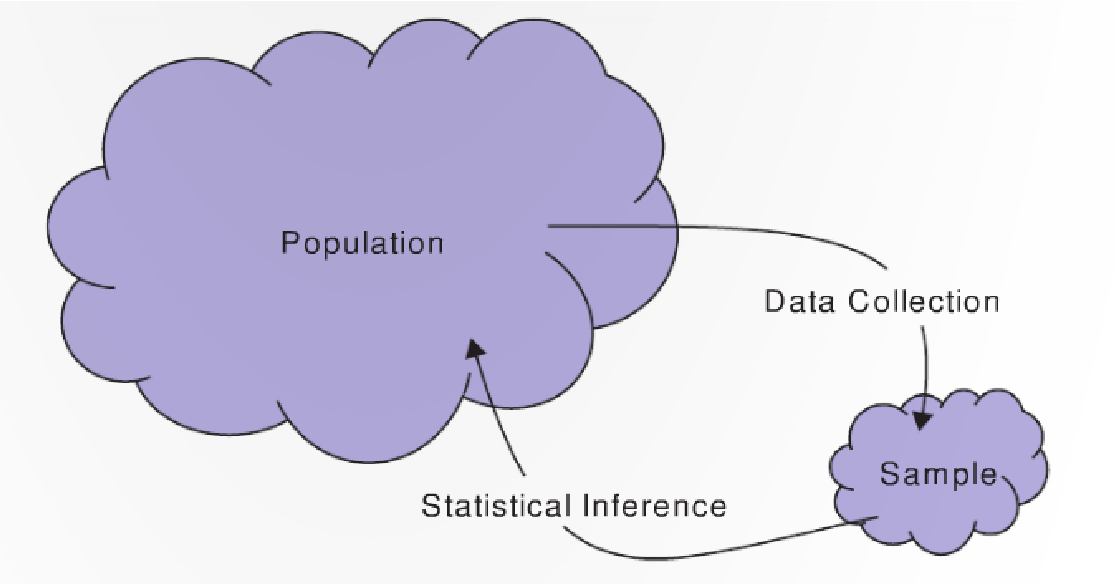
S&DS 100/500
Introductory Statistics



Ethan Meyers

Overview

Course overview

- Introductions
- Syllabus and logistics

What is Statistics?

Samples and Populations (shadows and truth)

Structured data: quantitative and categorical variables

R Studio

Contact Information

Email: ethan.meyers@yale.edu

Office: 24 Hillhouse Ave, Room 206

Planned office hours: Mon, Tues, Thurs 2-3pm

(no office hours this Thursday)

About me

Visiting assistant professor at Yale for this year

Assistant professor of Statistics Hampshire College

Research Fellow at the Center for Brains, Minds and Machines at MIT

Research: Machine learning methods to analyze neural data

Teaching Assistants

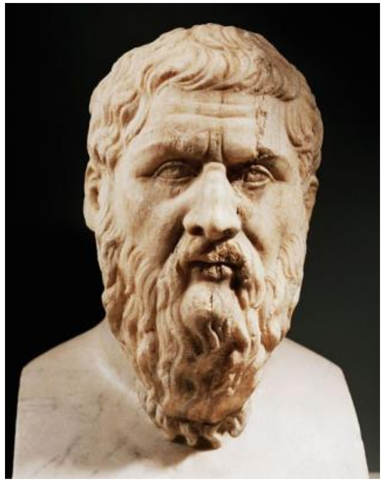
Teaching Fellows:

- Ryan Murphy: r.murphy@yale.edu

Undergraduate Learning Assistants

- Kelsey Evans
- Lu Zheng
- Michael Zhou

Learning goals



1. Understand the key concepts in Statistics

2. To learn how to analyze real data
    - We will use the R programming language
        - do not fear, this will make our life easier!

If you own a laptop, please bring it to class

Plan for the semester

Exploring data/descriptive statistics (weeks 1-4)

Sampling, categorical and quantitative data

Measures of central tendency and spread

- Mean, median, standard deviation

Relationships between variables

- Correlation and regression

Plan for the semester

Inferential Statistics

Sampling distributions

Confidence intervals

- The bootstrap

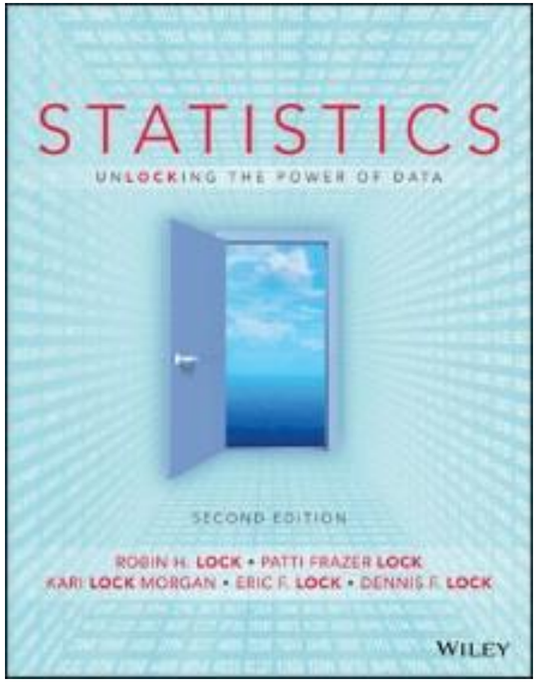
Randomization methods for hypothesis tests

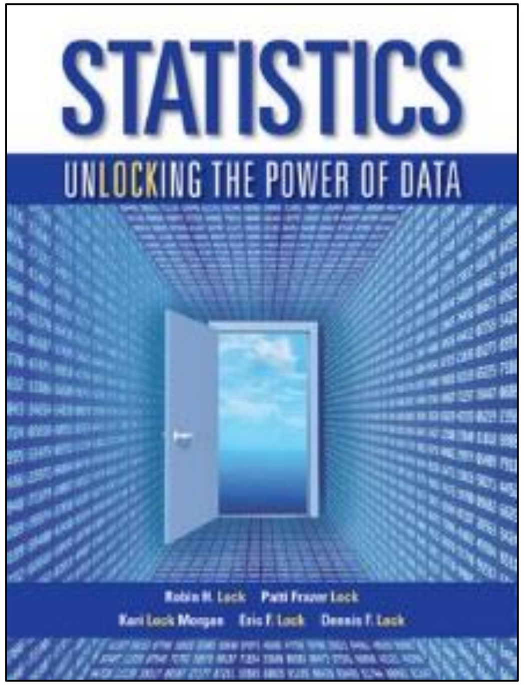
- Permutation tests

Parametric methods for hypothesis tests

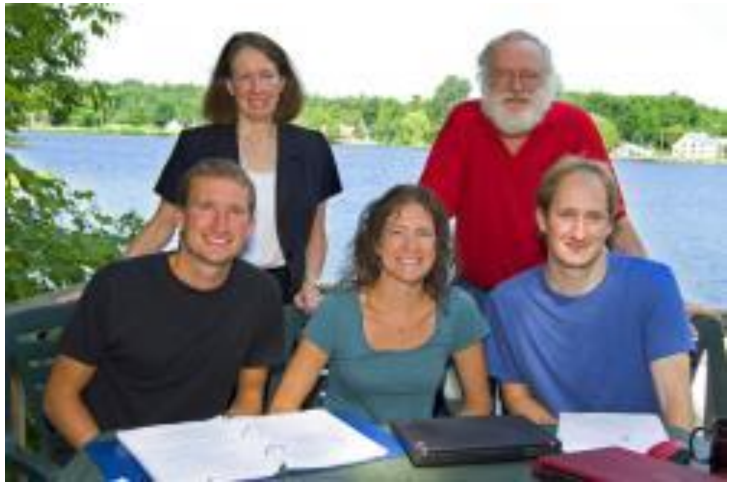
- T-tests, etc.

Textbook: Lock5


STATISTICS
UNLOCKING THE POWER OF DATA
SECOND EDITION
ROBIN H. LOCK • PATTI FRAZER LOCK
KARI LOCK MORGAN • ERIC F. LOCK • DENNIS F. LOCK
WILEY


STATISTICS
UNLOCKING THE POWER OF DATA
Robin H. Lock Patti Frazer Lock
Kari Lock Morgan Eric F. Lock Dennis F. Lock

Does anyone know why the textbooks is called Lock5?



Addition reading and other resources will be posted to Canvas: https://yale.instructure.com/courses/55573

Assignments and grades

1. Homework problem sets (54%)
    - Exploring concepts and analyzing data using R
    - Weekly: 10 total

Worksheet policies

- You may discuss questions with other but the work you turn in must be your own
- Worksheets assigned on Tuesdays and are due at 11:30pm on Sundays
- Late worksheets (90%) credit if turned in by 11:30pm on Monday
    - For any other extension a deans letter is needed
- Lowest scoring worksheet will be dropped

Assignments and grades

2. Final project (8%)

- Similar in length to a homework assignment, but you will analyze data of your own choosing based on your interests using methods discussed in the class
  - (might make this into a regular homework)

3. Exams (35% total)

- Midterm: March 5th (15%)
- Final: May 1st (20%)

4. Participation (3%)

- Active asking and answering questions on Piazza

Policies

Accommodation: please let me know if you have accommodations for homework and/or exams

Academic dishonesty: Don’t do it!

- You can work with others on the homework but the work you turn in needs to be your own
  - i.e., you need to understand the concepts and be able to produce the results yourself
- You can’t talk with others on exam, etc.

Examples of questions/analyses we will look at…

Z-scores: What is most impressive about LeBron James?

Sampling: How can insights from the Swedish chef help us avoid bias?

Confidence intervals: How can we pick ourselves up from the bootstrap to estimate a plausible range of values?

Randomization tests: Is it possible to smell whether someone has Parkinson’s disease?

Class survey

In order for me to get to know you and to better adjust the class to your interests, please fill out the class survey on canvas

- Under the Quizzes link on the left

Any questions about the class logistics???

What is Statistics? (capital S)

“Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world” (De Veaux et al. 2006, p. 2)

“Statistics is a body of methods for making wise decisions in the face of uncertainty” (Wallis & Roberts 1962, p. 11)

Fienberg, S. (2014). What is Statistics? The *Annual Review of Statistics and Its Application,* 1:1-9

My thoughts

Statistics is a way to use data to answer questions:

- Often we use a small amount of data to answer questions about a larger underlying phenomenon
- We want to know the truth, and not be fooled by randomness
    - Quantify uncertainty and randomness

It's <u>part</u> of an argument

- Don't blindly trust statistical tests, think about the results!
    - Do you really believe them?
- Be your own worst critic and try to prove yourself wrong

A warning about terminology

“Boy, those French: They have a different word for everything!”

- Steve Martin

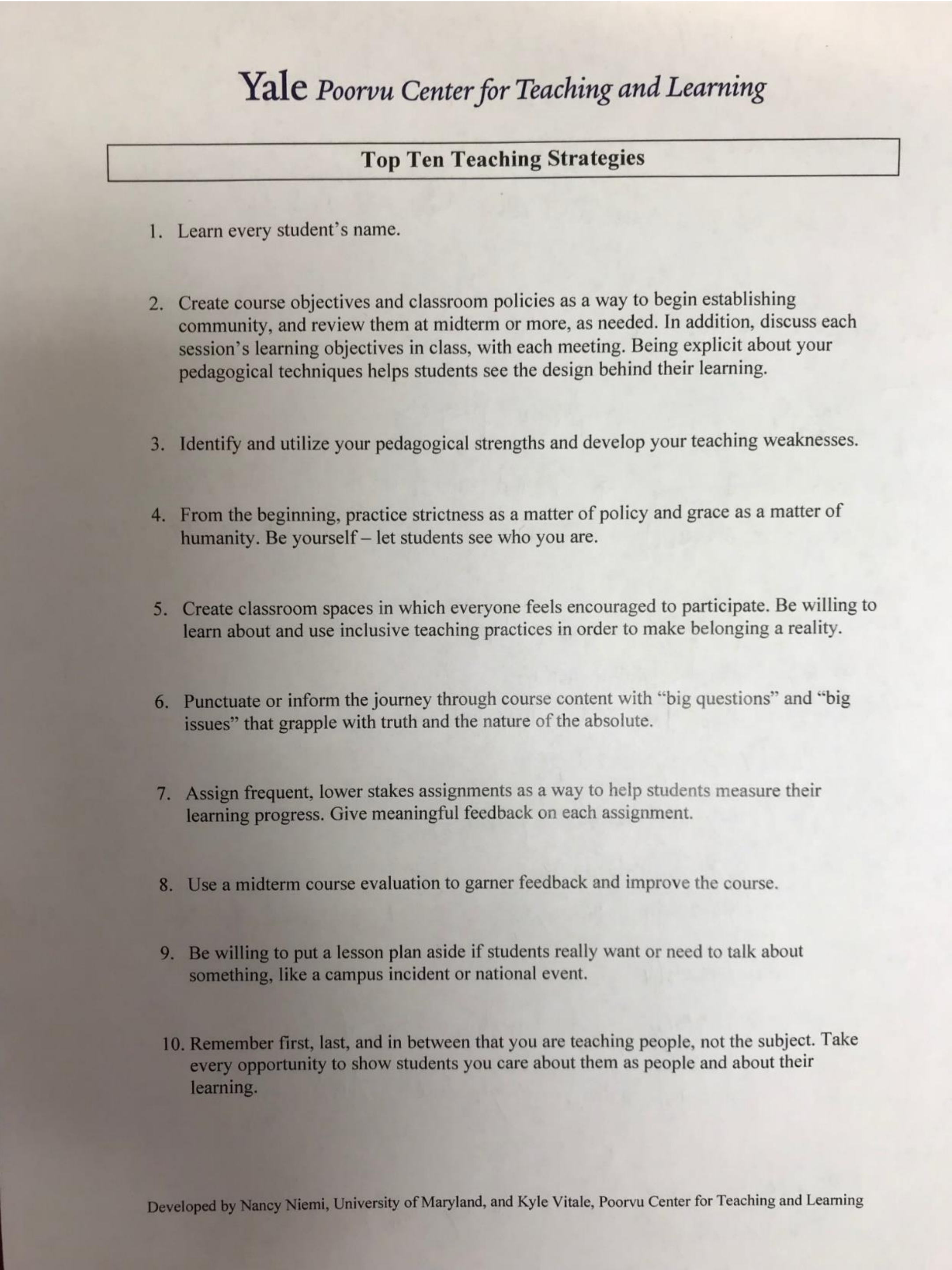
Boy, those Statisticians: They use common words to mean something different!

Bias,

confidence,

significance

Center for Teaching and Learning tips



Yale *Poorvu Center for Teaching and Learning*

Top Ten Teaching Strategies

1. Learn every student's name.

2. Create course objectives and classroom policies as a way to begin establishing community, and review them at midterm or more, as needed. In addition, discuss each session's learning objectives in class, with each meeting. Being explicit about your pedagogical techniques helps students see the design behind their learning.

3. Identify and utilize your pedagogical strengths and develop your teaching weaknesses.

4. From the beginning, practice strictness as a matter of policy and grace as a matter of humanity. Be yourself – let students see who you are.

5. Create classroom spaces in which everyone feels encouraged to participate. Be willing to learn about and use inclusive teaching practices in order to make belonging a reality.

6. Punctuate or inform the journey through course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute.

7. Assign frequent, lower stakes assignments as a way to help students measure their learning progress. Give meaningful feedback on each assignment.

8. Use a midterm course evaluation to garner feedback and improve the course.

9. Be willing to put a lesson plan aside if students really want or need to talk about something, like a campus incident or national event.

10. Remember first, last, and in between that you are teaching people, not the subject. Take every opportunity to show students you care about them as people and about their learning.

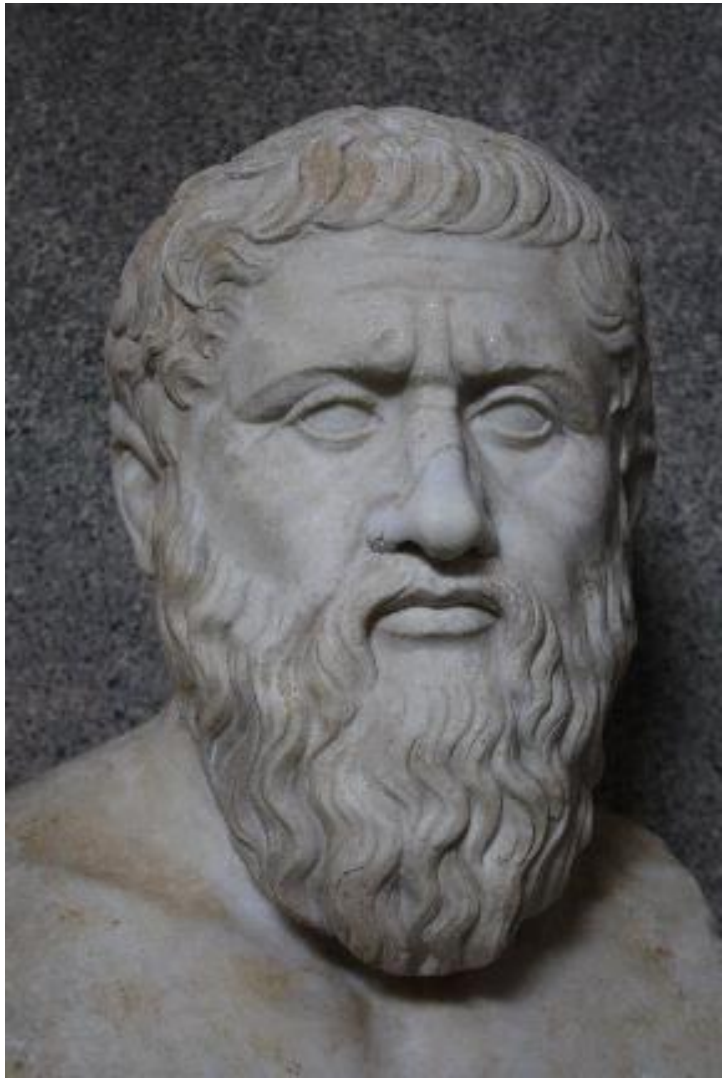
Developed by Nancy Niemi, University of Maryland, and Kyle Vitale, Poorvu Center for Teaching and Learning

Tip 1: Learn every student's name

Tip 6: Punctuate or inform the journey through the course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute

Central concepts in Statistics

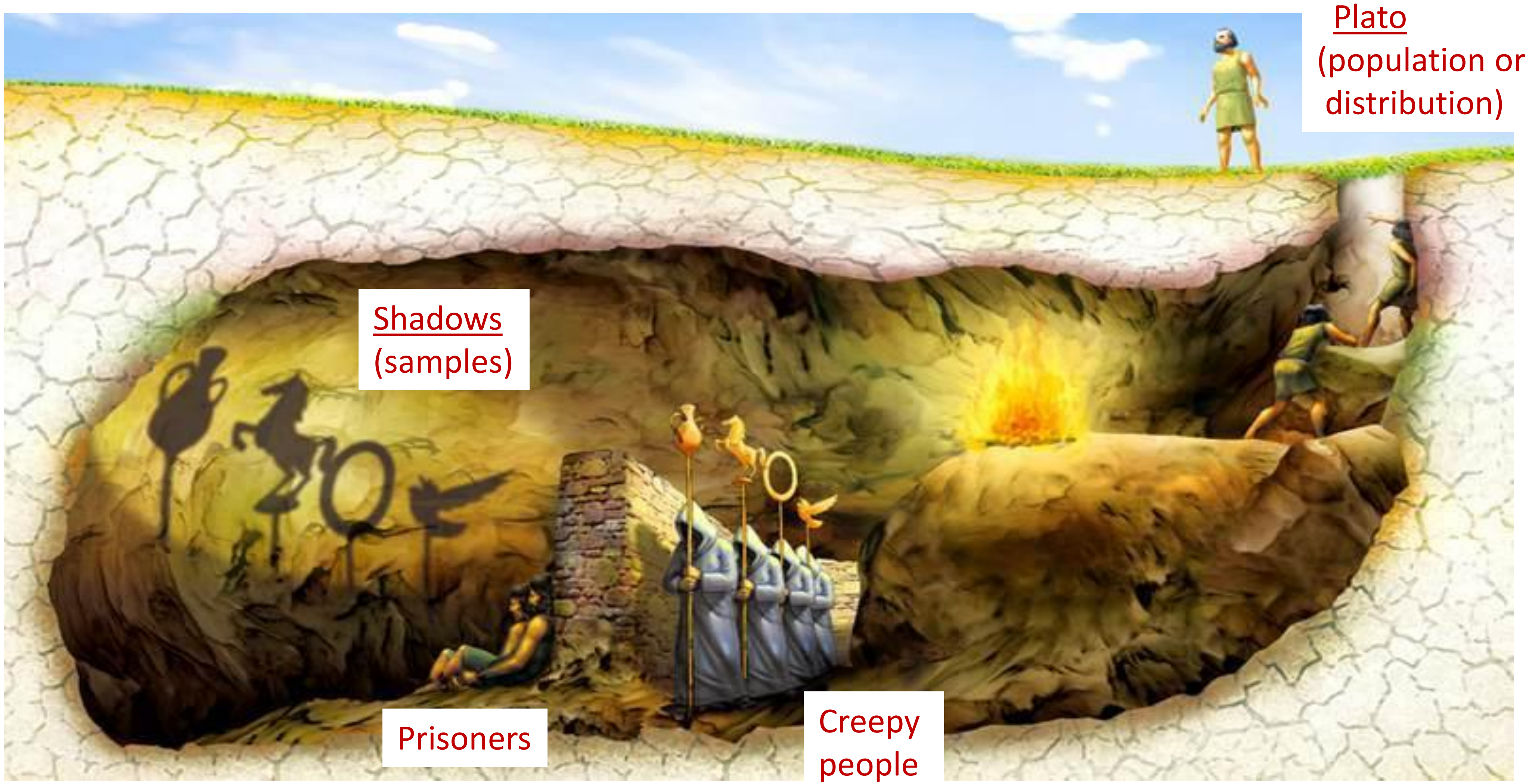
The Truth!



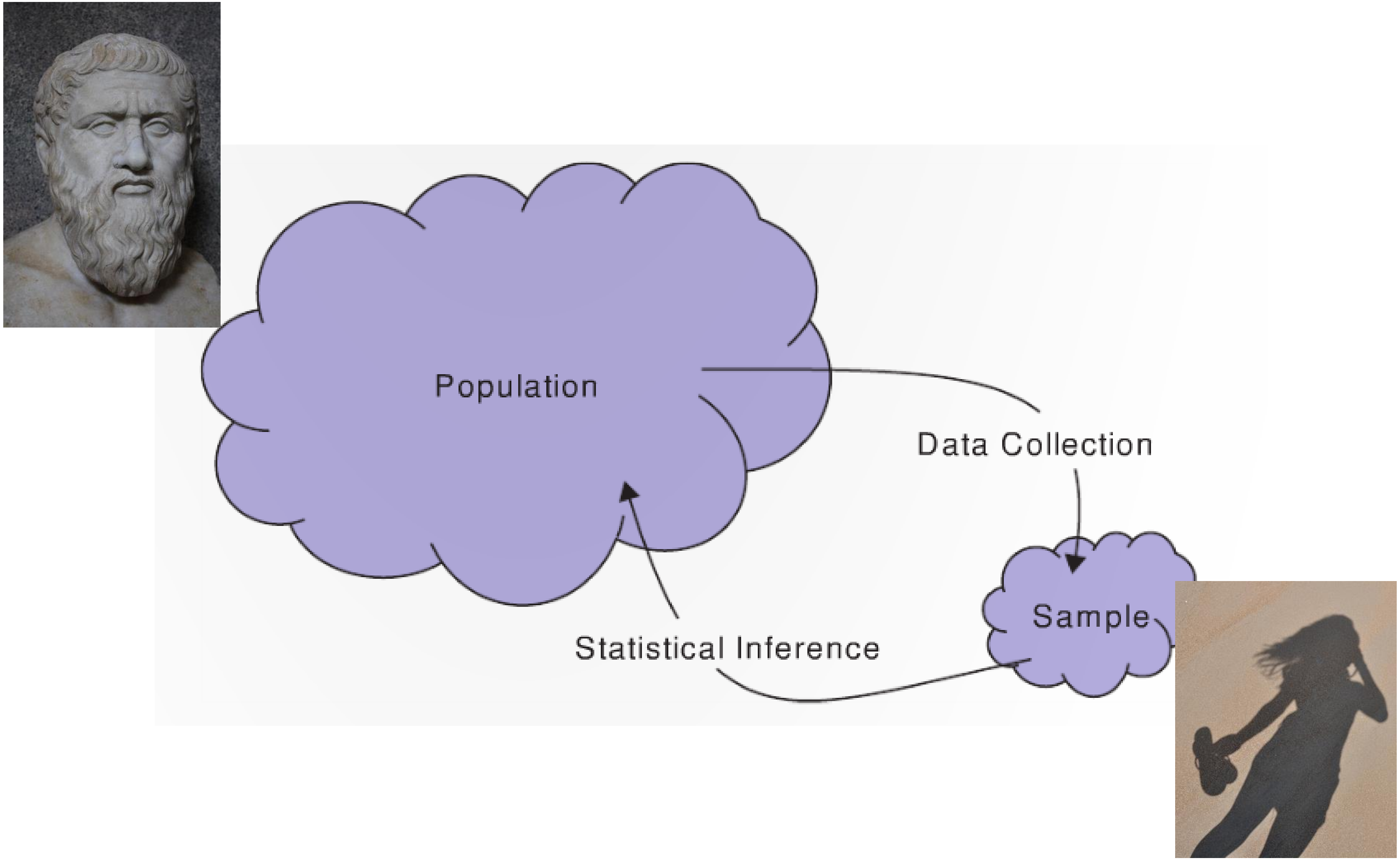
If we could see all the (infinite) data, we would know the Truth®!

Alas, we can only see a small subset of the data (a sample) so we merely see a shadow of the truth

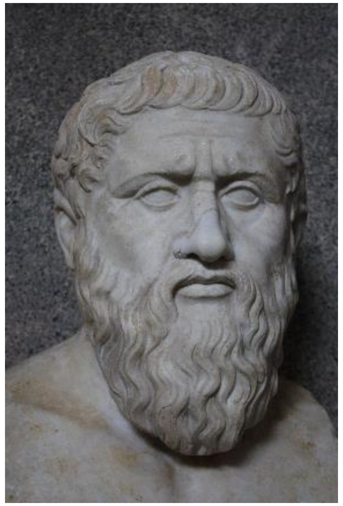
Plato’s cave



From The Republic (~ 380 BCE)


Population
Data Collection
Sample
Statistical Inference

Sample from a Population



Population: all individuals/objects of interest

Sample: A subset of the population

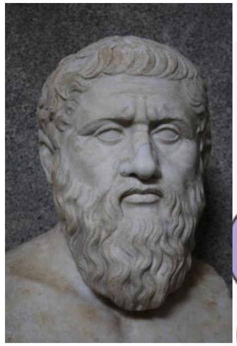
Descriptive and inferential statistics

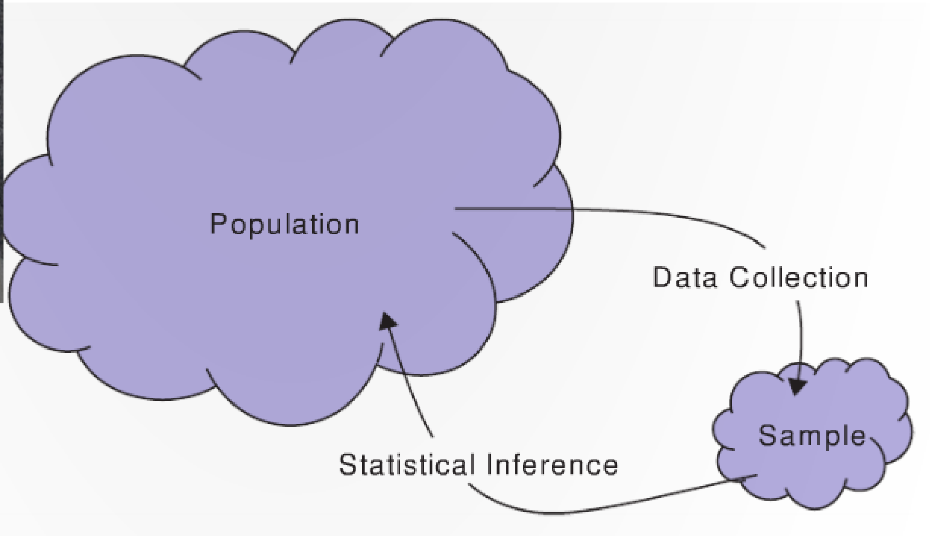
Descriptive Statistics: describe the sample of data we have

- i.e., describe the shadows

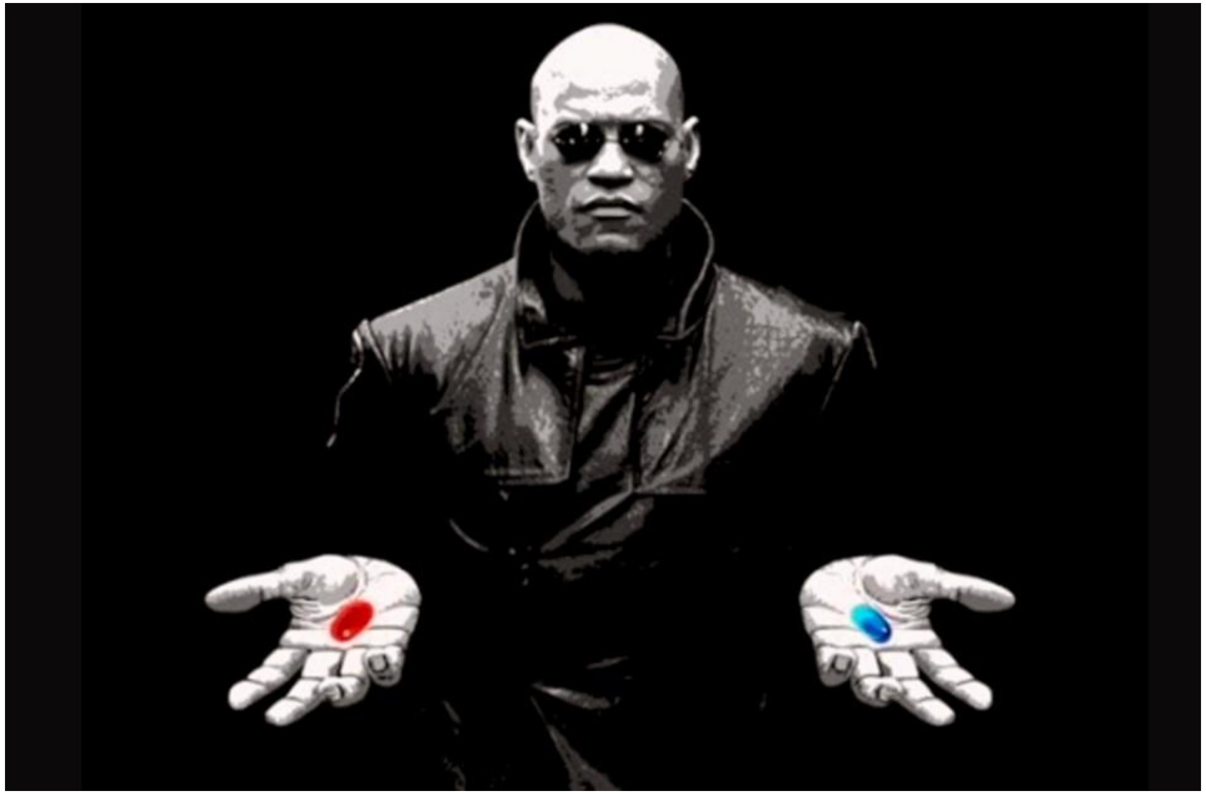
Inferential Statistics: use the sample to make claims about properties of the population/process

- i.e., try to use the data to get at the truth





Can you handle the Truth?



If not, perhaps you should not take this class

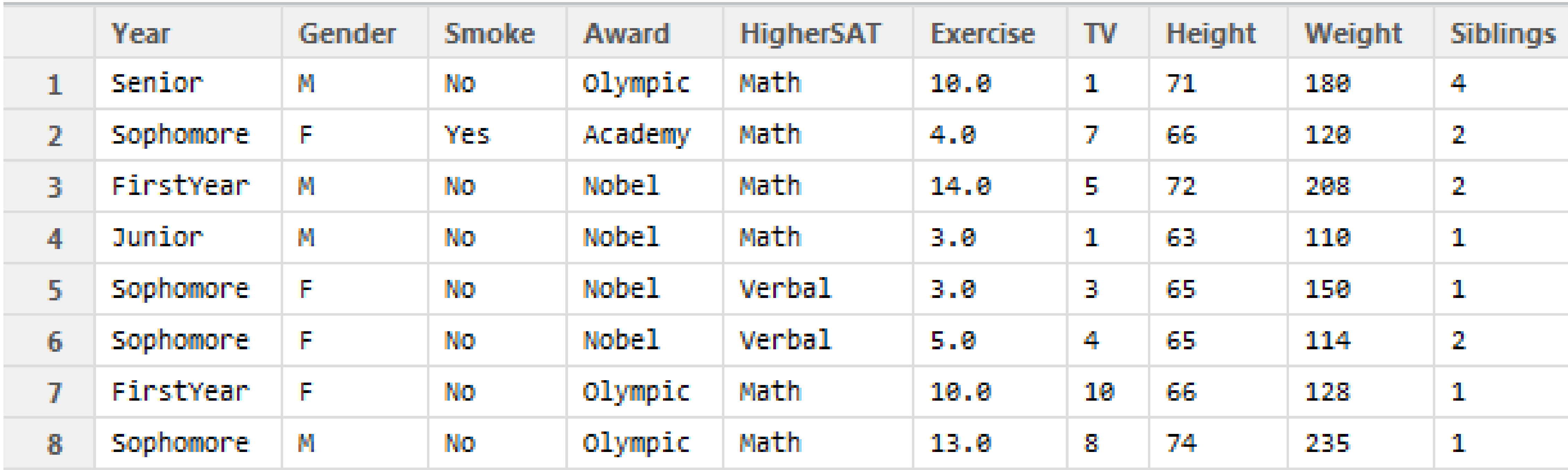
You’ve been warned…

Structured data – exploring the shadows

An Example Dataset (Shadows)

Variables

Cases



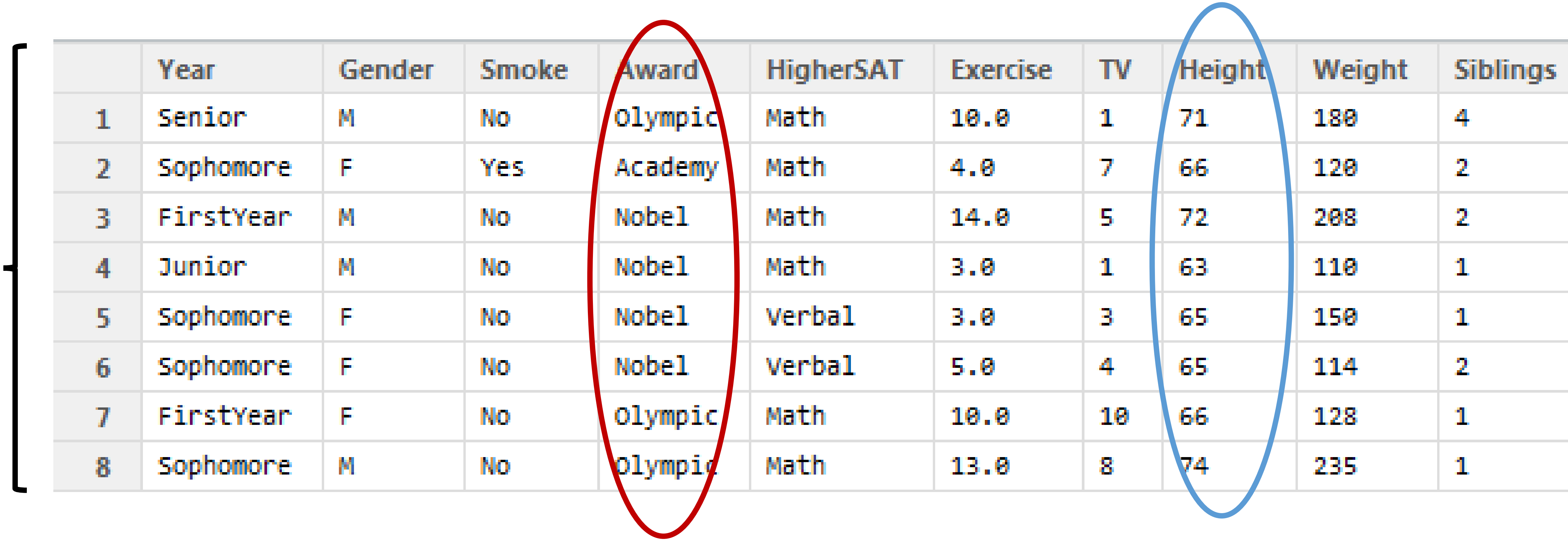
	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

An Example Dataset (Shadows)

Categorical Variable

Quantitative Variable

Cases (observational units)

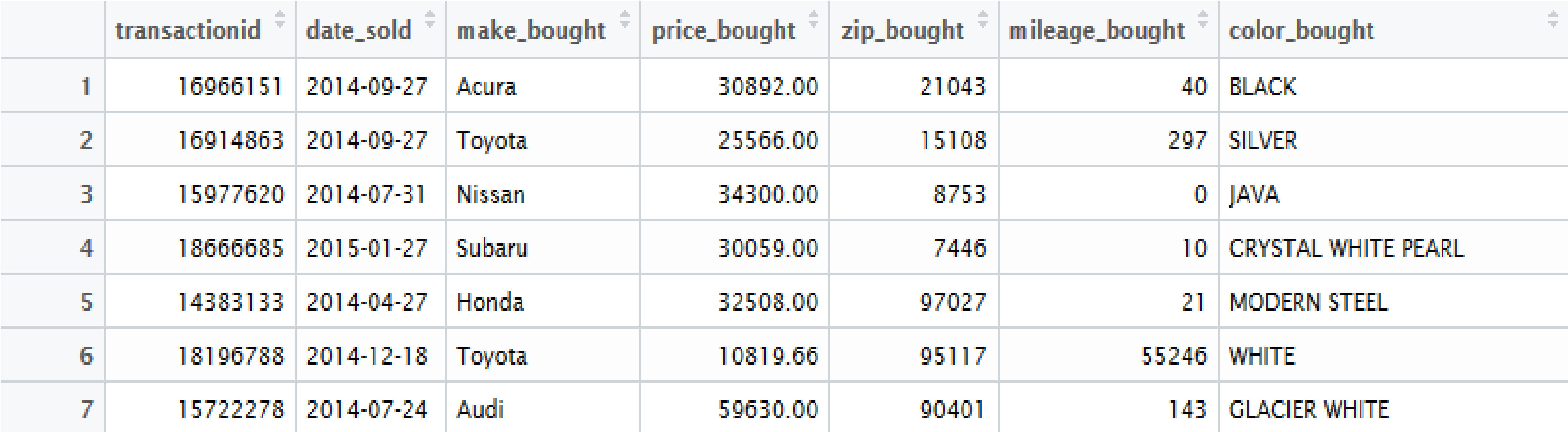


	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

Edmunds transaction data

- What are the observational units (cases)?
- Which variables are: quantitative or categorical?

Discuss!



	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

Question

Q: What programming language do pirates use?

A: Arrrr

Q: Worst joke of the semester?

A: Wait and see…

Log in to R Studio Cloud

http://bit.ly/SDS100

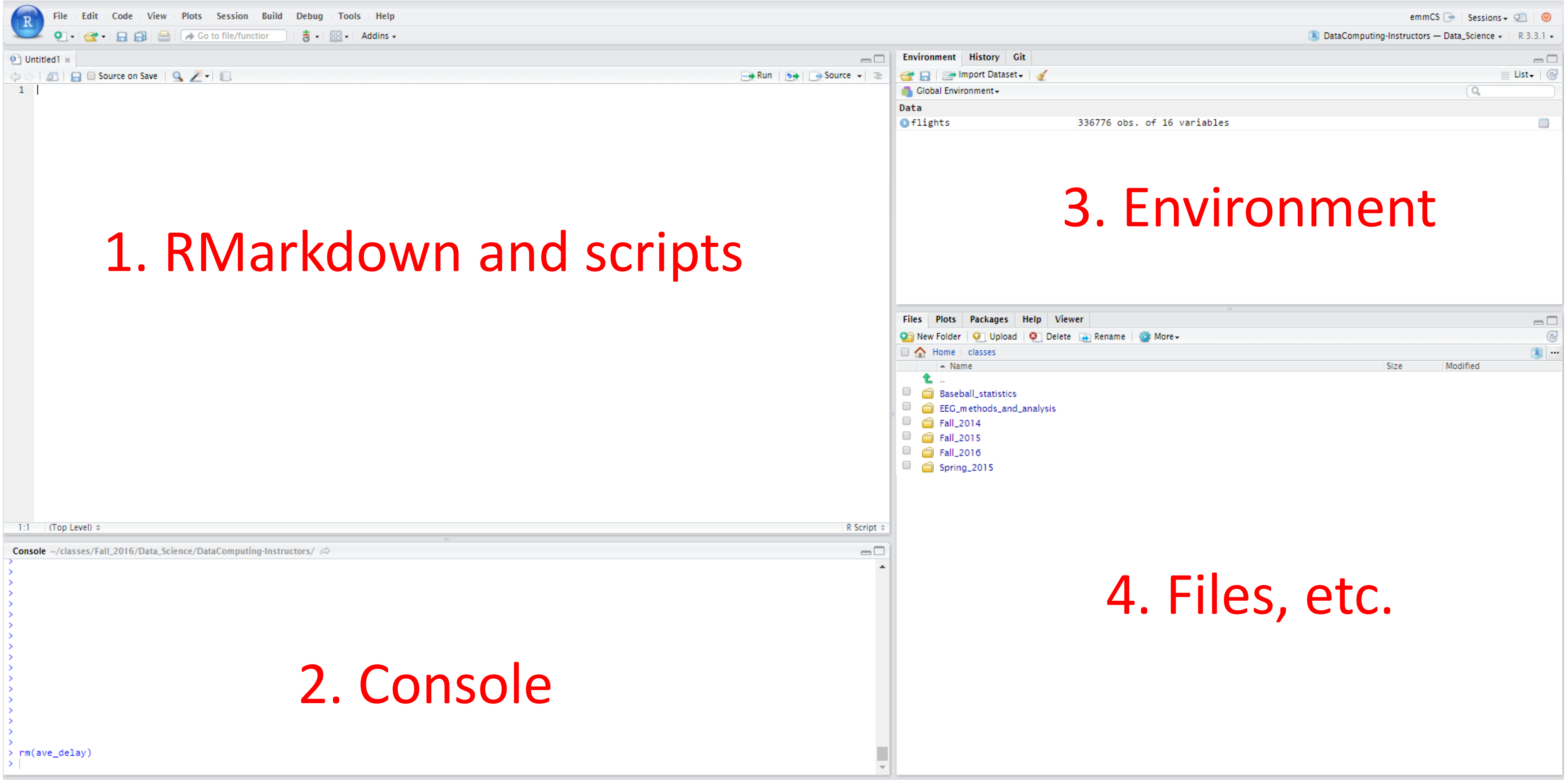
Or download and install R Studio

Talk to your neighbor while code is loading…

R and R Studio

R: Engine

RStudio layout



1. RMarkdown and scripts

2. Console

3. Environment

4. Files, etc.

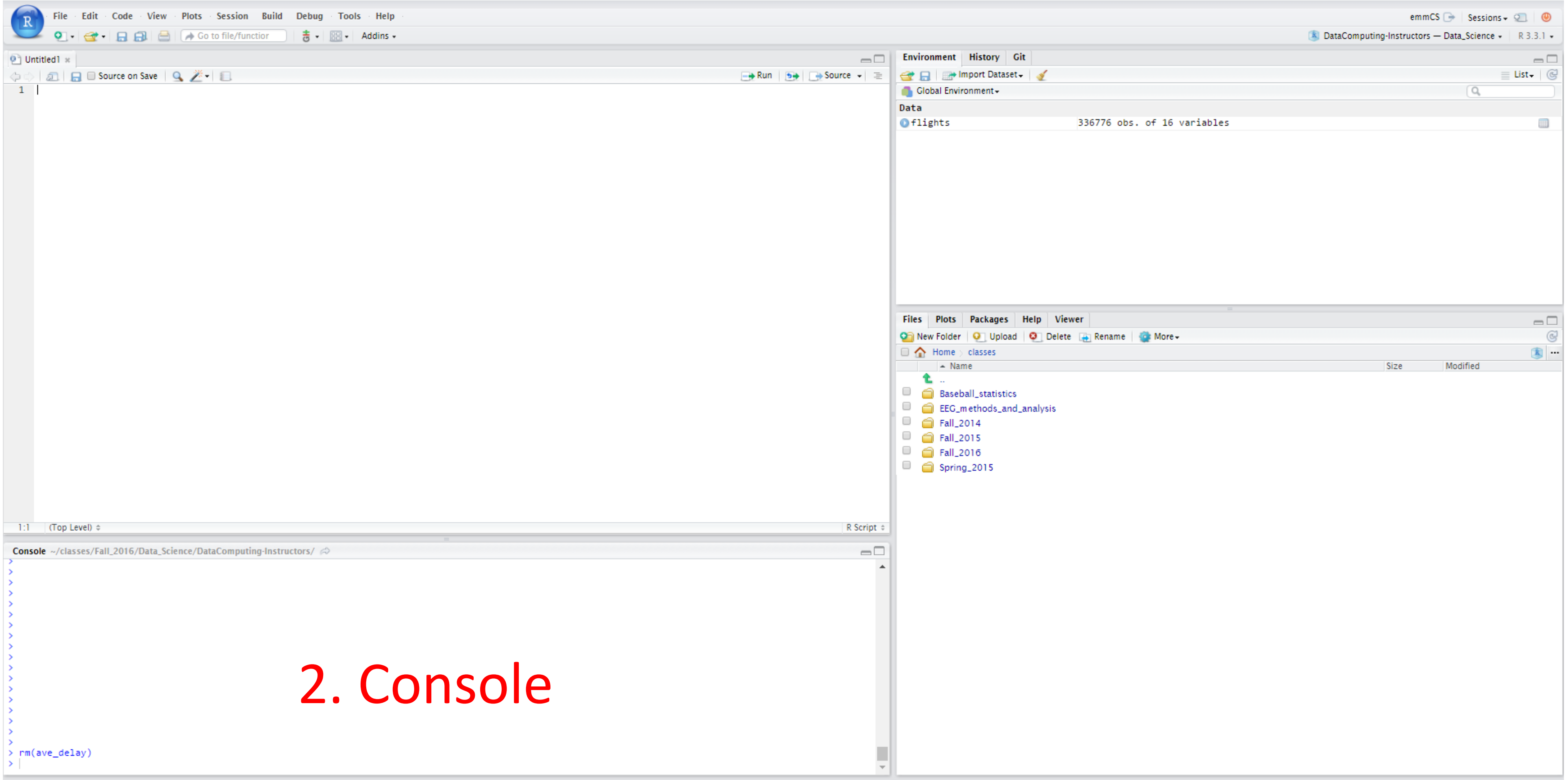
```
Data
flights                    336776 obs. of 16 variables
```

```
..
Baseball_statistics
EEG_methods_and_analysis
Fall_2014
Fall_2015
Fall_2016
Spring_2015
```

Console ~/classes/Fall_2016/Data_Science/DataComputing-Instructors/

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> rm(ave_delay)
>
```

RStudio layout



```
1 |
1:1
```

Data

flights	336776 obs. of 16 variables

Home › classes

Name	Size	Modified
..		
Baseball_statistics		
EEG_methods_and_analysis		
Fall_2014		
Fall_2015		
Fall_2016		
Spring_2015		

Console ~/classes/Fall_2016/Data_Science/DataComputing-Instructors/

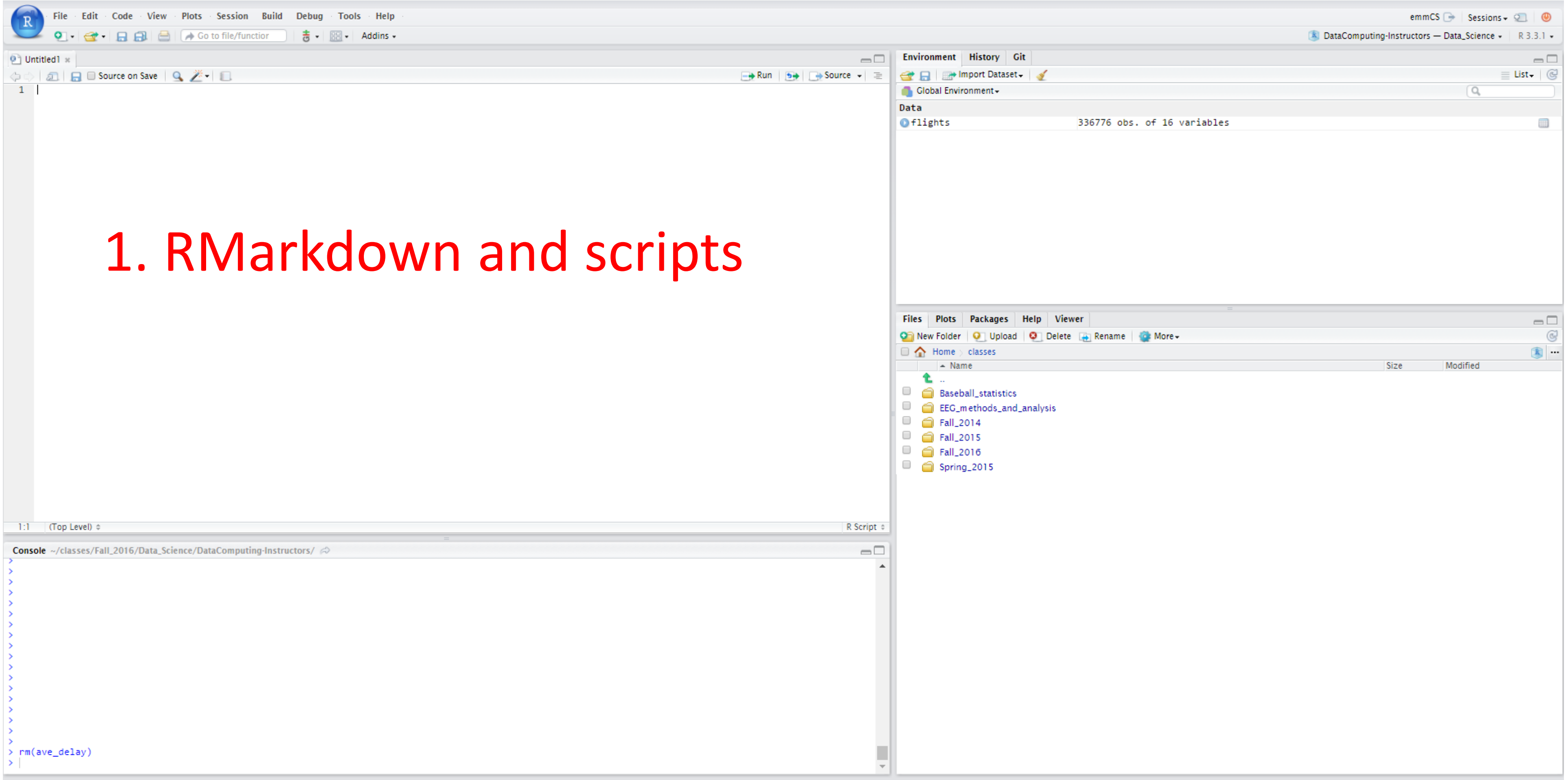
```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> rm(ave_delay)
> |
```

2. Console

R as a calculator

```
>   2 + 2
>   7 * 5
```

RStudio layout



1. RMarkdown and scripts

```
flights          336776 obs. of 16 variables
```

```
Console ~/classes/Fall_2016/Data_Science/DataComputing-Instructors/
> rm(ave_delay)
>
```

R Basics

Arithmetic:

```
>   2 + 2
>   7 * 5
```

Assignment:

```
>  a <- 4
>  b <- 7
>  z  <- a + b
>  z
[1]  11
```

Number journey…

Number journey

```
> a <- 7
> b <- 52
> d <- a * b
> d
[1]  364
```

Character strings and booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE

> class(a)
[1] numeric

> class(s)
[1] character
```

Summary of concepts

1. Population: all individuals/objects of interest (truth)

2. Sample: A subset of the population (shadows)

3. Statistical inference: Making judgments about the population using data from the sample

4. Structured data has

- Cases/observational units: rows in a data set
- Variables: columns in a data set

5. Variables can be

- Categorical: fall into discrete categories
- Quantitative: are numbers

For next class

Please fill out the class survey on Canvas under quiz

Practice problems from Lock 5, first edition:

1.1, 1.3, 1.5, 1.11, 1.25, 1.26

Chapter 1 is posted on Canvas