S&DS 230/530:
Data Exploration and Analysis

Ethan Meyers

Overview

Introductions

Overview and logistics of the course

Review of a few central concepts from Intro Stats

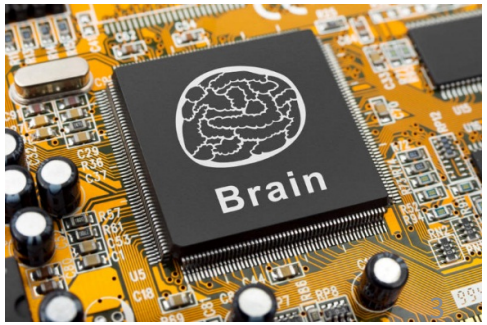
Introduction to R

- R as a calculator
- Objects and vectors
- Installing the class SDS230 package and LaTeX

About me

- Visiting professor at Yale
- Assistant professor of Statistics Hampshire College
- Research Affiliate at the Center for Brains, Minds and Machines at MIT

Research area: Machine learning to analyze neural data



Ethan.Meyers@yale.edu

Teaching Assistants

Teaching Fellows (TF)

- Hayon Michelle Choi: hayonmichelle.choi@yale.edu
- Akshay Surendra: akshay.surendra@yale.edu
- Sam Konstantinov (course manger): Sam.konstantinov@yale.edu

Undergraduate Learning Assistants (ULA)

- Lu Zheng: lu.zheng@yale.edu
- Derek Chen: derek.chen@yale.edu
- Maria (Duda) Eduarda Santana: mariaeduarda.santana@yale.edu
- Stephan Billingslea: stephan.billingslea@yale.edu

Undergraduate Technology Assistant (UTA)

- João Goncalves Cardoso: joao.cardoso@yale.edu
  - Please contact João if you have any questions about how to use Zoom

Introductions

Let’s do some quick introductions

- Your name
- Your major/grad dept (research area)
- Why you are interested in this class
- Anything else you would like to share with your group

If you have questions during today's class

Type them into the chat box

- If possible, add the slide number

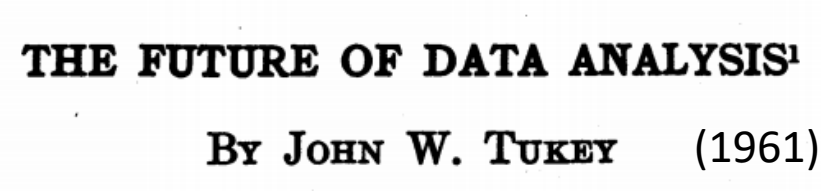
With the help of João, I will try to answer them at fixed points during the class

What is this class about?

What is data analysis?

1. Exploratory data analysis

- Visualize, describe and model data to find **potential** trends and to generate **new** hypotheses
  - Descriptive statistics, data mining


THE FUTURE OF DATA ANALYSIS[1]

BY JOHN W. TUKEY (1961)

2. Confirmatory data analysis

- Confirming or falsifying **existing** hypotheses
  - Inferential statistics, preregistered studies


Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Related concepts:

- Data modeling, predictive analytics
- Data cleaning, transforming, wrangling


Discussion

50 Years of Data Science

David Donoho

Pages 745-766 | Received 01 Aug 2017, Published online: 19 Dec 2017

Course objectives

Gain experience extracting insights from real data

Learn how to use the R programming language to analyze, visualize and wrangle data

Extend methods learned in intro stats

- Non-parametric tests, multiple regression, etc.

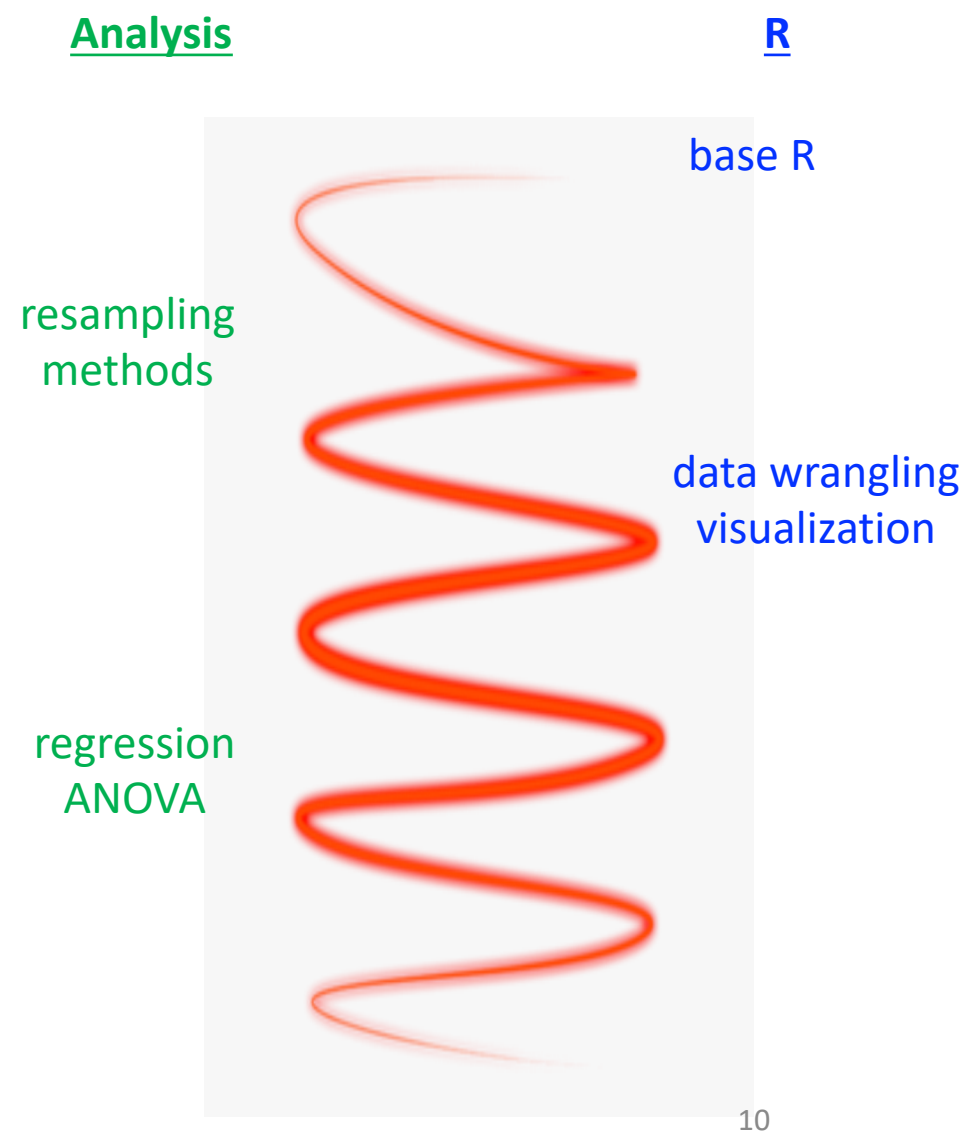
Solidify understanding of statistical concepts

- Focus on insights why methods work rather than proofs

Learn how to find patterns in a large noisy data sets and convincingly convey the results to others!

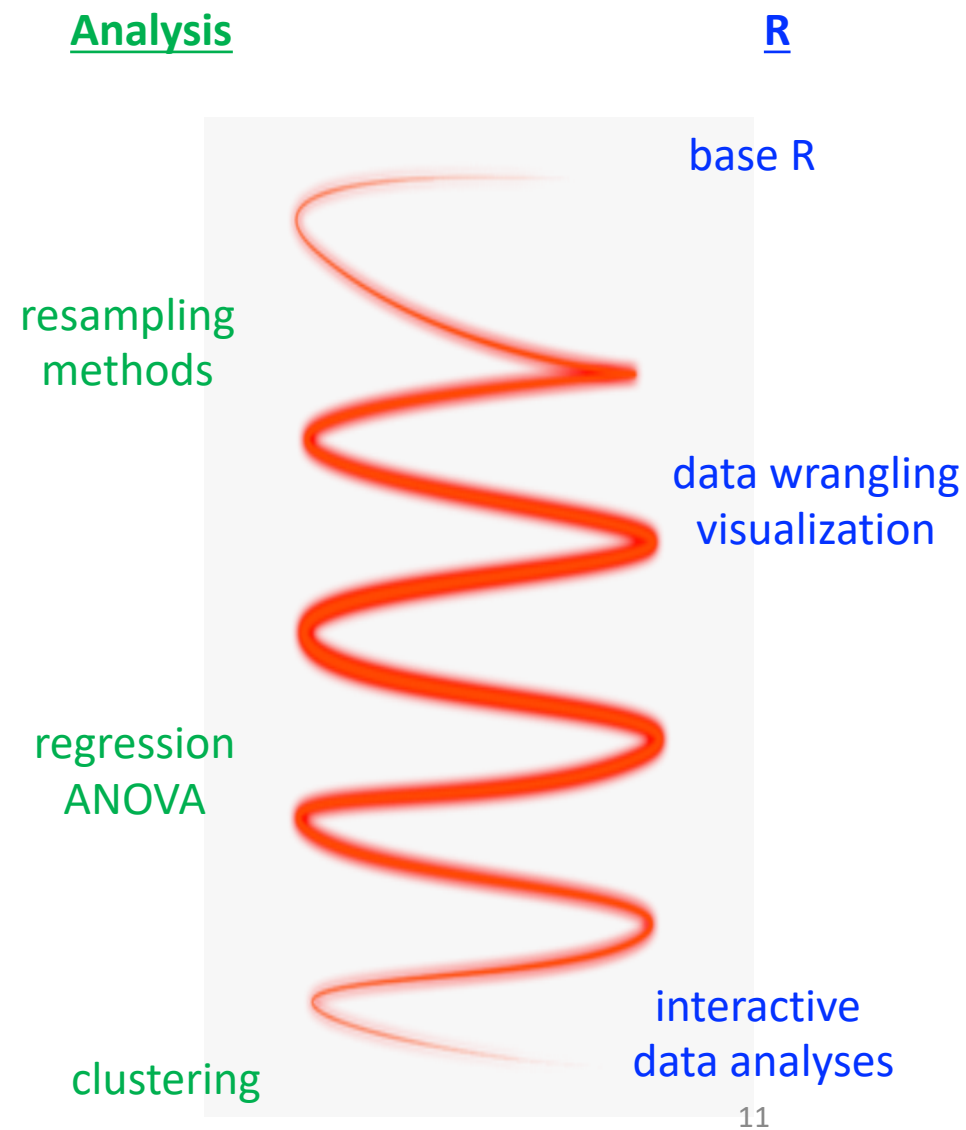
Plan for the semester (subject to change)

1	Sep 1-3	Course overview, introduction to R, descriptive statistics
2	Sep 15-17	Review of central statistical concepts and exploratory analysis using R
3	Sep 22-24	Confidence Intervals and the bootstrap
4	Sep 29-Oct 1	Review of hypothesis tests and permutation tests in R
5	Sep 24-26	Permutation tests continued and parametric tests
6	Oct 6-8	Data wrangling and visualization
7	Oct 13-15	Simple and multiple regression
8	Oct 20-22	Review and midterm exam



Plan for the semester (subject to change)

9	Oct 27-29	Multiple regression continued
10	Nov 3-5	Logistic regression/classification and cross-validation
11	Nov 10-12	Analysis of Variance
12	Nov 17-19	Mapping, joining and interactive graphics
13	Nov 21-29	November break
14	Dec 1-3	Clustering, principal components analysis, wrap up and review
15	Dec 7-10	Online reading period
16	Dec 11-18	Final exam



Topics we will cover

R and descriptive statistics/plots: Base R, fundamental concepts in Statistics

Review confidence intervals: Sampling and bootstrap distributions

Review of hypothesis tests: Permutation and parametric tests, theories of testing

Data wrangling: filtering and summarizing data, joining data sets, reshaping data

Data visualization: grammar of graphics, mapping

Regression: simple/multiple, non-linear terms, logistic regression

ANOVA: one-way/factorial, interactions

Statistical learning: cross-validation, logistic regression, PCA, clustering

Class structure (still figuring it out)

Class time 9-10:15am Tuesdays and Thursdays

- Will introduce some content during class (recorded)
- Class activities, reading discussions, question answering, etc.

Likely will use pre-recorded videos as well

- Potentially will shorten class time
  - e.g., 9:30-10:15am or just Thursdays

For the first two weeks, let's plan on regular class time

- 9-10:15am

Prerequisites

An introductory class in Statistics (AP or 10X)

- We will review Intro Stats concepts using computational methods but we will be going through the material at a fast pace

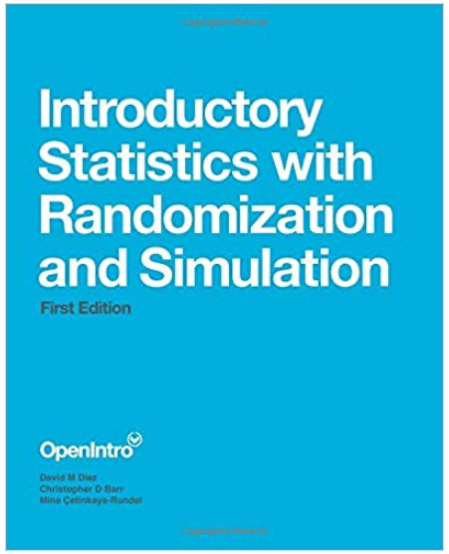
A large component of this class will be using the R programming

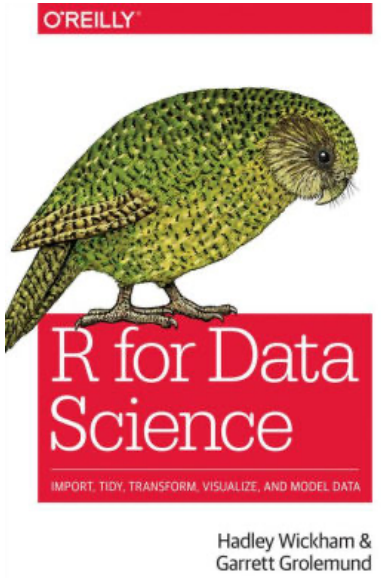
- No prior programming experience needed
  - Everyone in the class has free access to DataCamp

Logistics

Canvas website: https://yale.instructure.com/courses/61201/

No required text, reading resources will be posted to canvas and in the homework assignments





Office hours

My planned office hours (subject to change)

- Mondays and Wednesdays 11-12pm

TA office hours will be posted on calendar on Canvas

- We will try to have consistent office hours, although they might change

For specific questions about content in the class, best to first ask them on Piazza

- Class participation grade partially based on questions and answers on Piazza

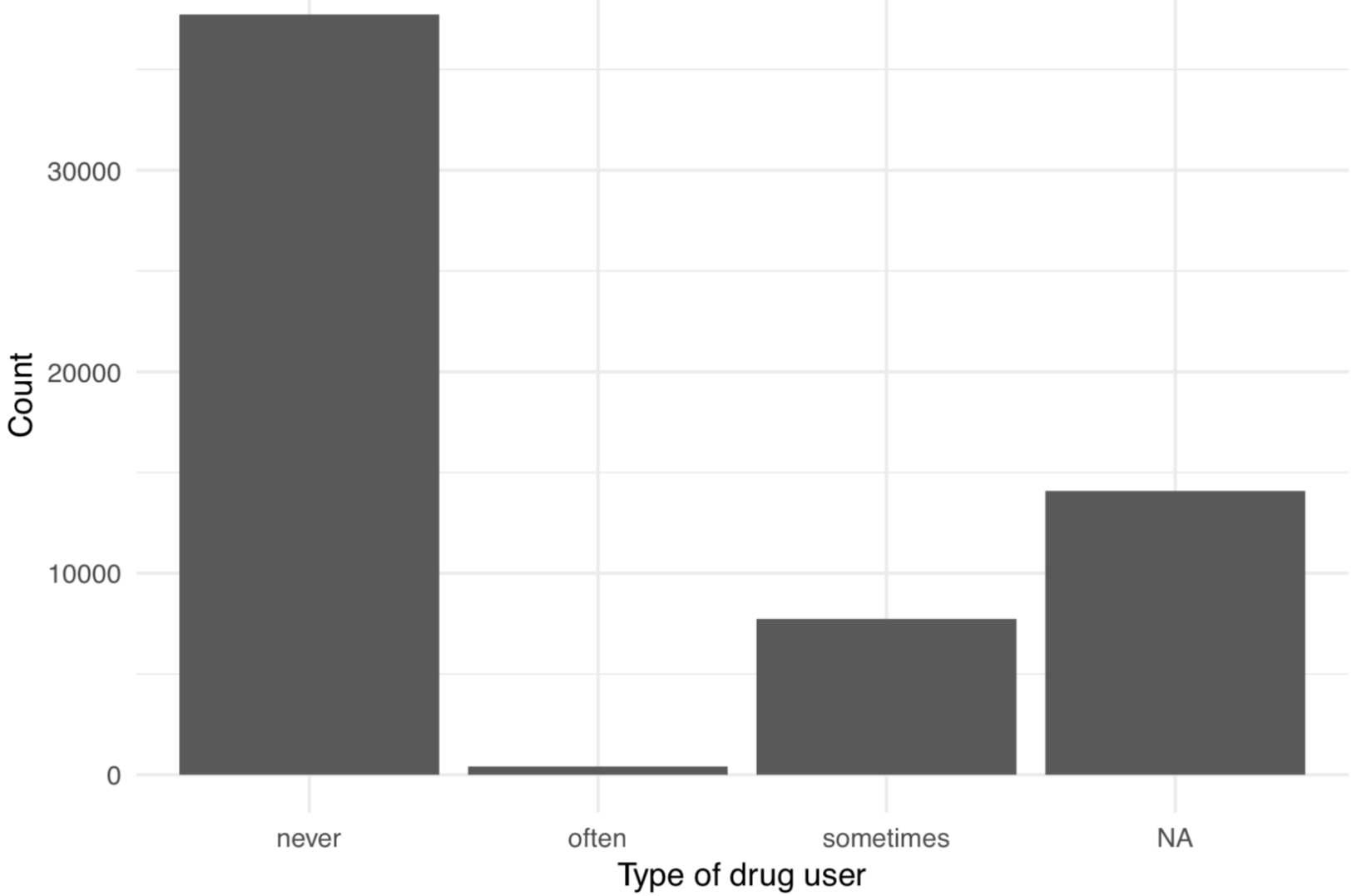
Assignments and grades

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

Worksheet policies

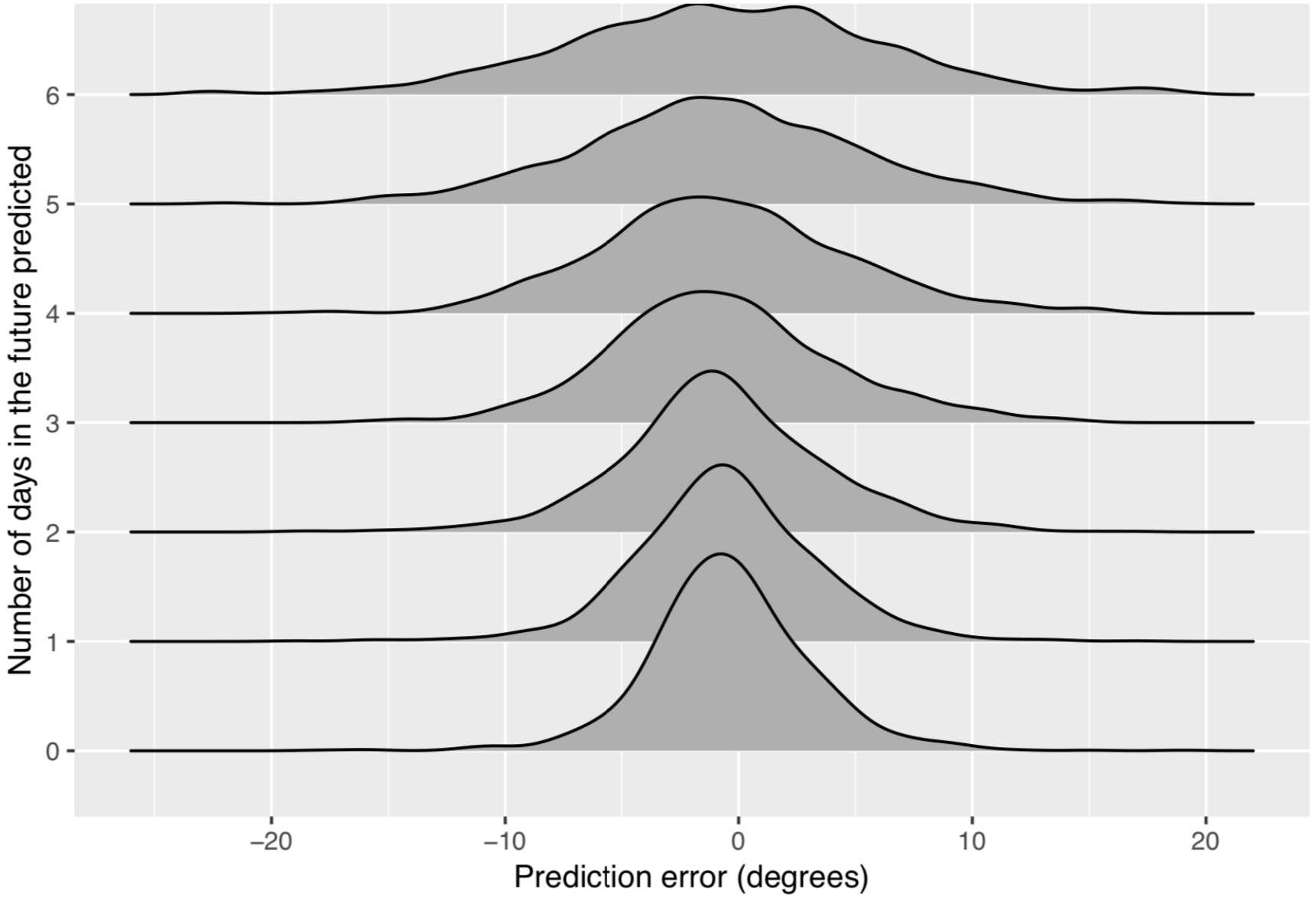
- You may discuss questions with other but the work you turn in must be your own
- Worksheets assigned on Tuesdays and are due at 11:59pm on Sundays
- Late worksheets (90%) credit if turned in by 11:59pm on Monday
  - For any other extension a dean's excuse is needed
- Lowest scoring worksheet will be dropped

Typical homework assignment piece




```
# Bonus: create a pie chart of the self reported frequency of
# drug use and make it look good!
profiles %>% count(drugs) %>% filter(!is.na(drugs)) %>% ggplot(aes(x = "",
    y = n, fill = drugs)) + geom_col(width = 1) + coord_polar(theta = "y") +
    theme_minimal() + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) + xlab("")
```

Typical homework assignment piece



Answers: Personally I like the joy plot best here because it most clearly shows how the distribution becomes more spread out for predictions made further in the future (although all three plots do a reasonable job of showing this).

Assignments and grades

2. Final project (10%)
   - Find a data set and analyze it on your own (~5-7 page report)

3. Exams (30%; 15% each)
   - Midterm Oct 22nd
   - Final Dec 12-18th

4. Participation (6%)
   - Active asking and answering questions on Piazza and in class
   - Engaging with short class readings and discussions

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 on Canvas

Preliminary class survey results

As of 5pm yesterday, 82 people had filled out the class survey

- ~60% of the class is undergraduates,
- ~40% graduate students

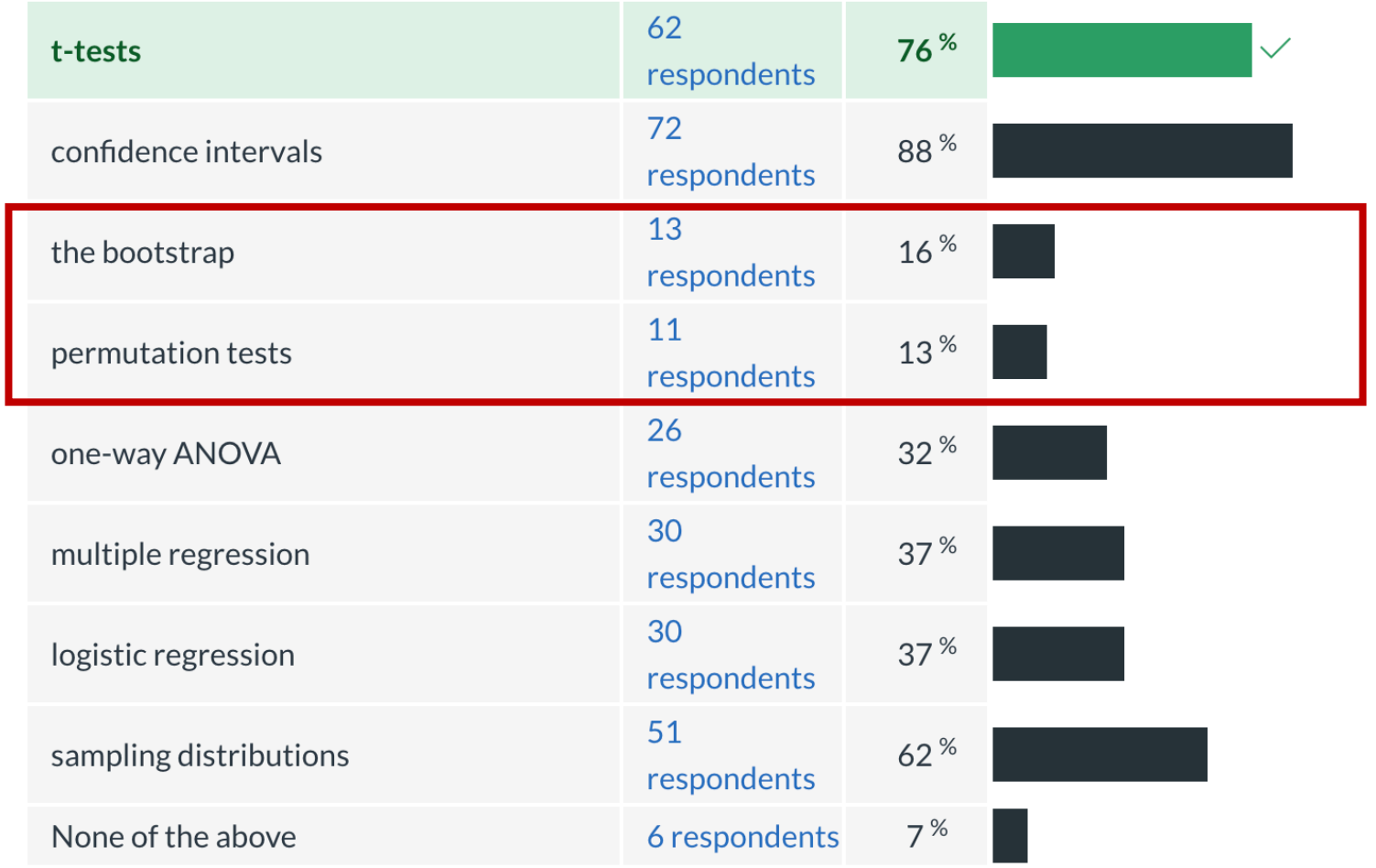
Have you taken an introductory Statistics class before?

Yes, in college	62 respondents	**76 %**	
Yes, in high school (AP stats)	22 respondents	27 %	
No	7 respondents	9 %	

Class survey results

Which Statistics methods/concepts are you comfortable with?

t-tests	62 respondents	**76 %** ✓
confidence intervals	72 respondents	88 %
the bootstrap	13 respondents	16 %
permutation tests	11 respondents	13 %
one-way ANOVA	26 respondents	32 %
multiple regression	30 respondents	37 %
logistic regression	30 respondents	37 %
sampling distributions	51 respondents	62 %
None of the above	6 respondents	7 %



Class survey results

Have you taken linear algebra and/or calculus?

- These are not prerequisites
- Programming experience is also not a prerequisite

Yes, linear algebra	16 respondents	**20 %**	
Yes, calculus I/II	29 respondents	35 %	
Yes, multivariable calculus	29 respondents	35 %	
I have not taken any of these classes	8 respondents	10 %	

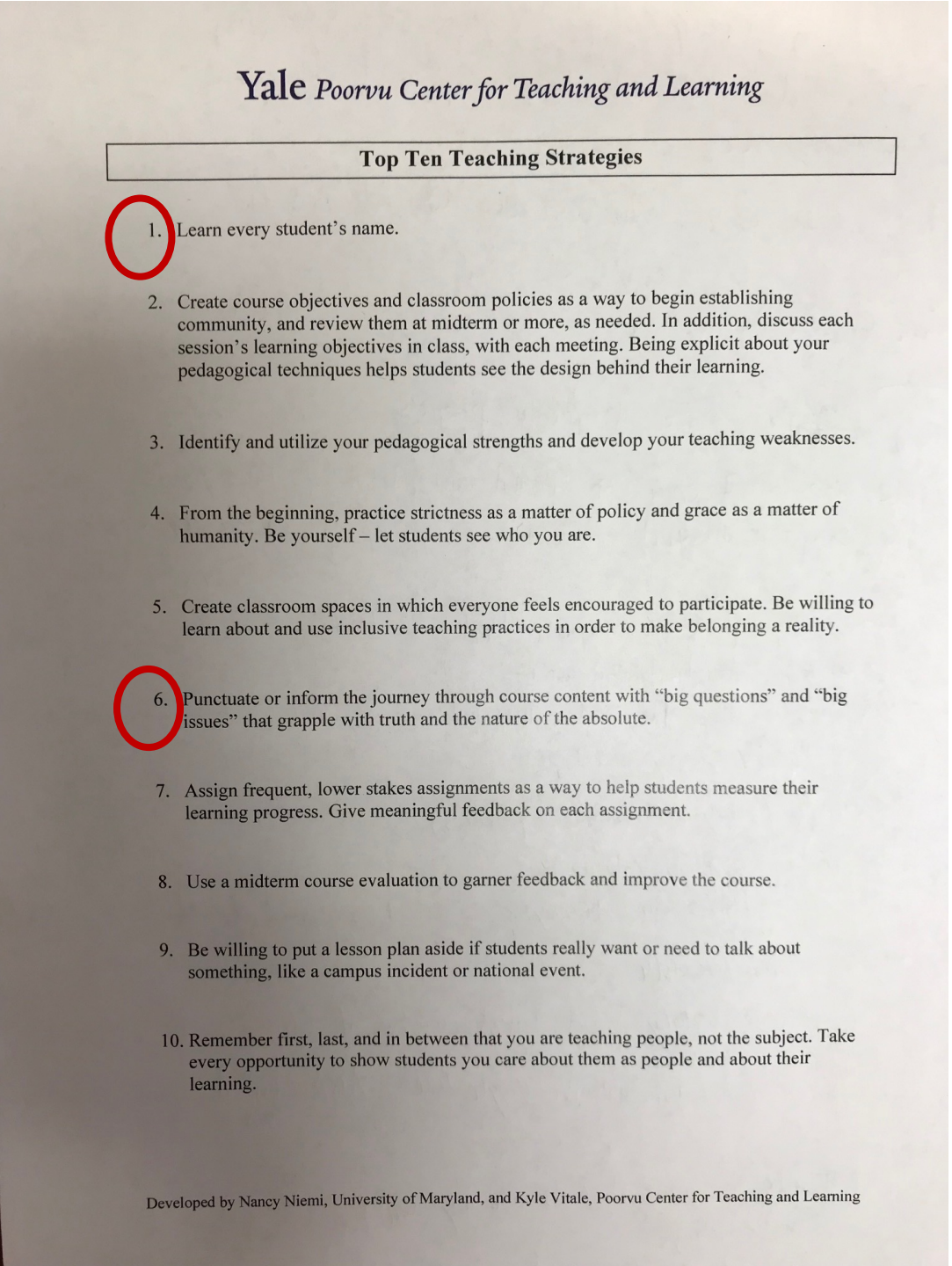
Class survey results

Will you be able to attend class during the regularly schedule class time?

Yes, I can attend at that time	51 respondents	**62 %**	
I can attend > 90% of the regular class time	29 respondents	35 %	
No, I cannot attend the regular class time	2 respondents	2 %	

Questions?

Quick Review of central concepts in Intro Statistics



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.

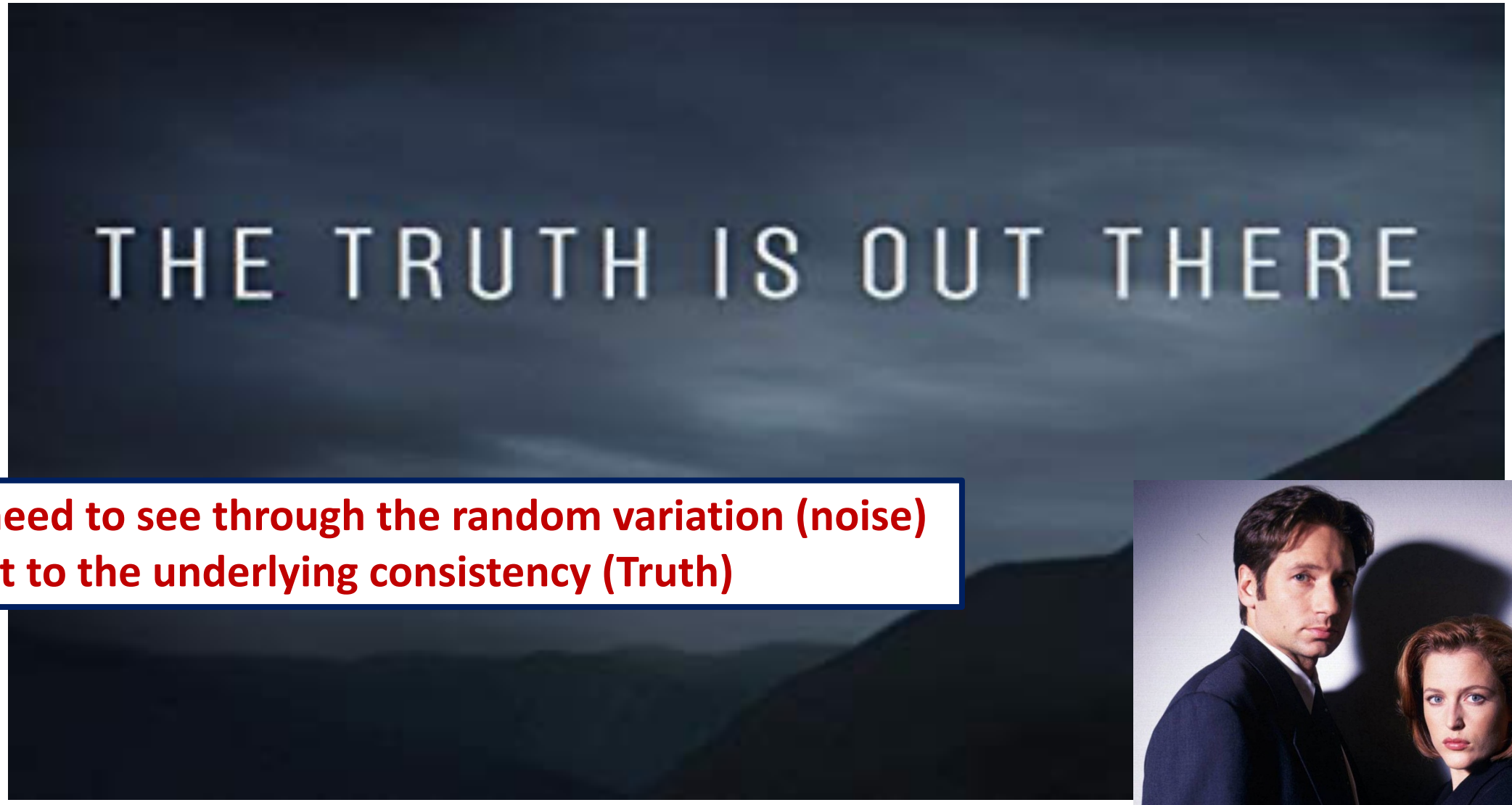
Developed by Nancy Niemi, University of Maryland, and Kyle Vitale, Poorvu Center for Teaching and Learning

Center for Teaching and Learning tips

Tip 1: Learn every student's name

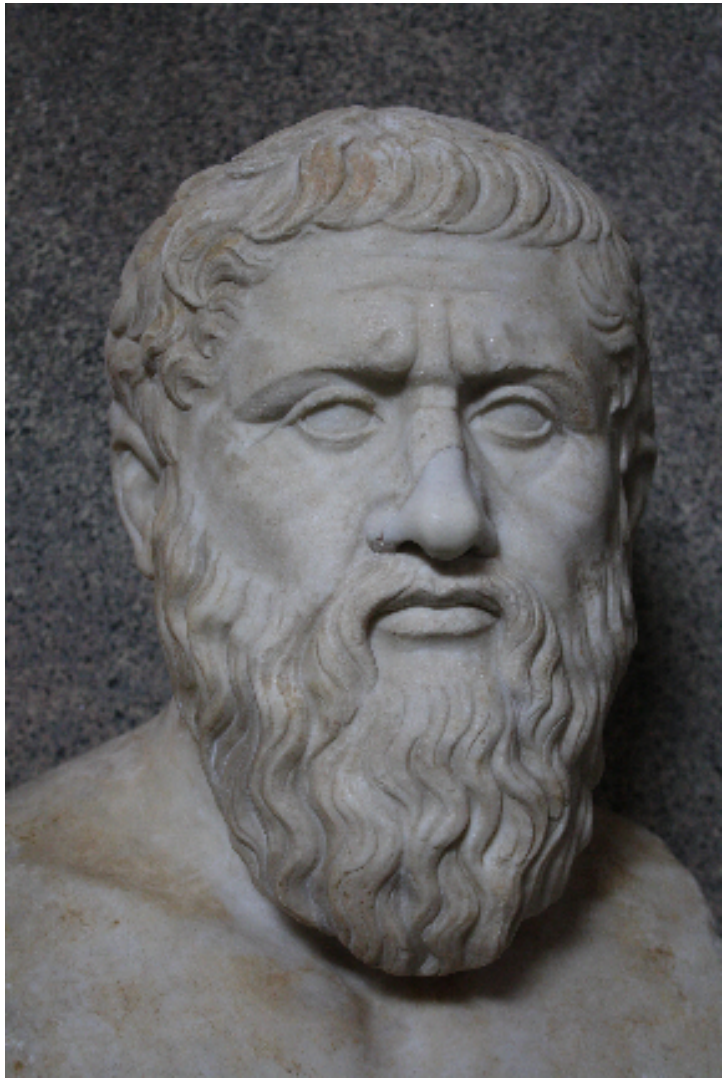
Tip 6: Punctuate or inform the journey through the course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute

Quick Review of central concepts in Intro Statistics



We need to see through the random variation (noise) to get to the underlying consistency (Truth)

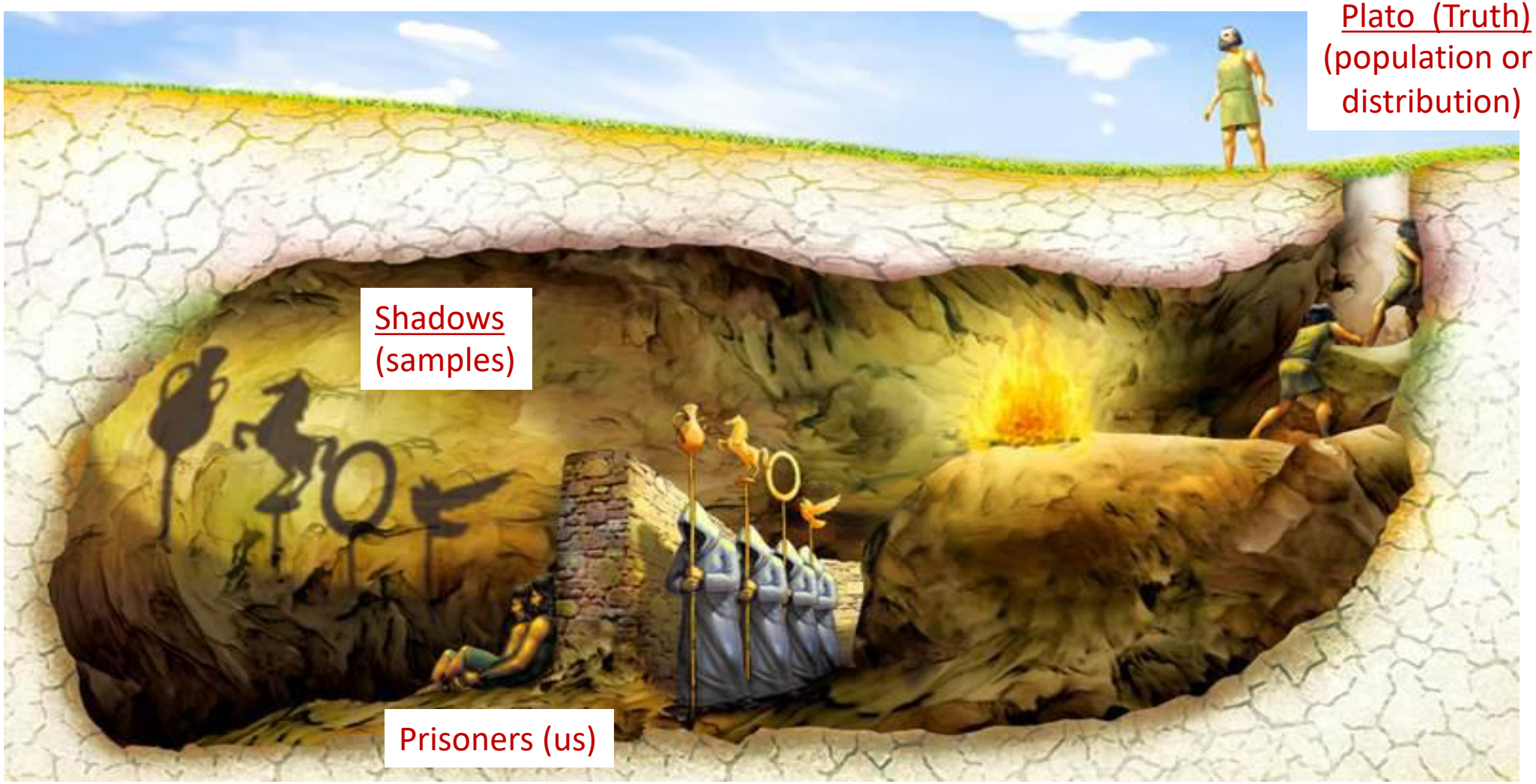
The Truth®!



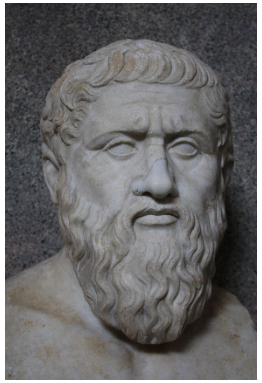
If we could see all the (infinite) data, we would know the Truth®!

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

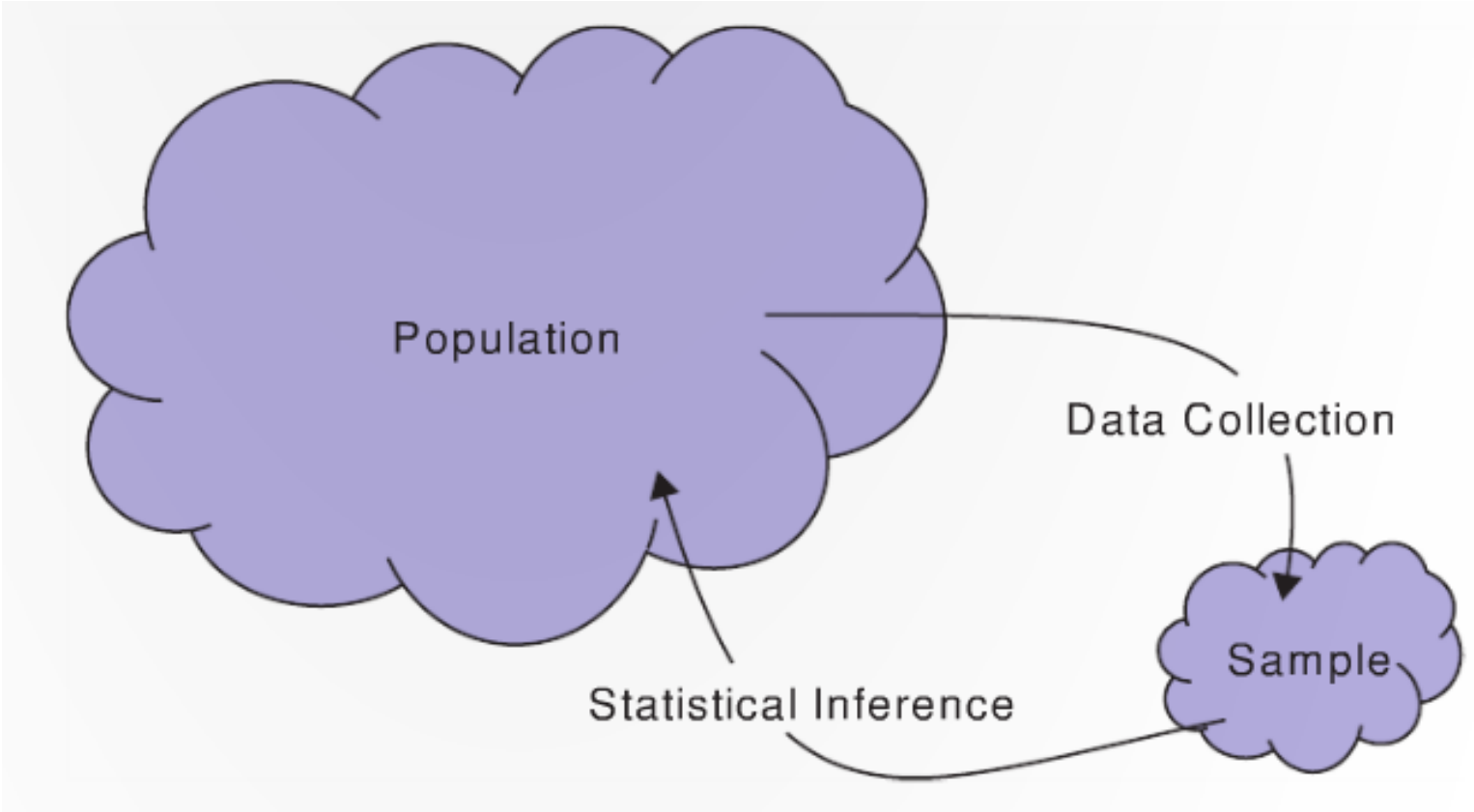
Plato's cave

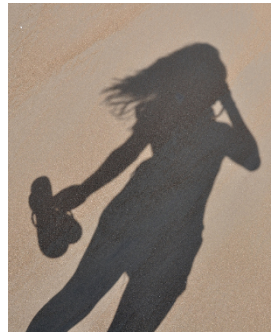


From The Republic (~ 380 BCE)

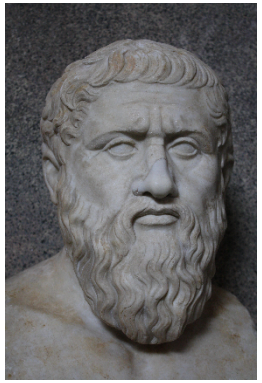


Population: all individuals/objects of interest



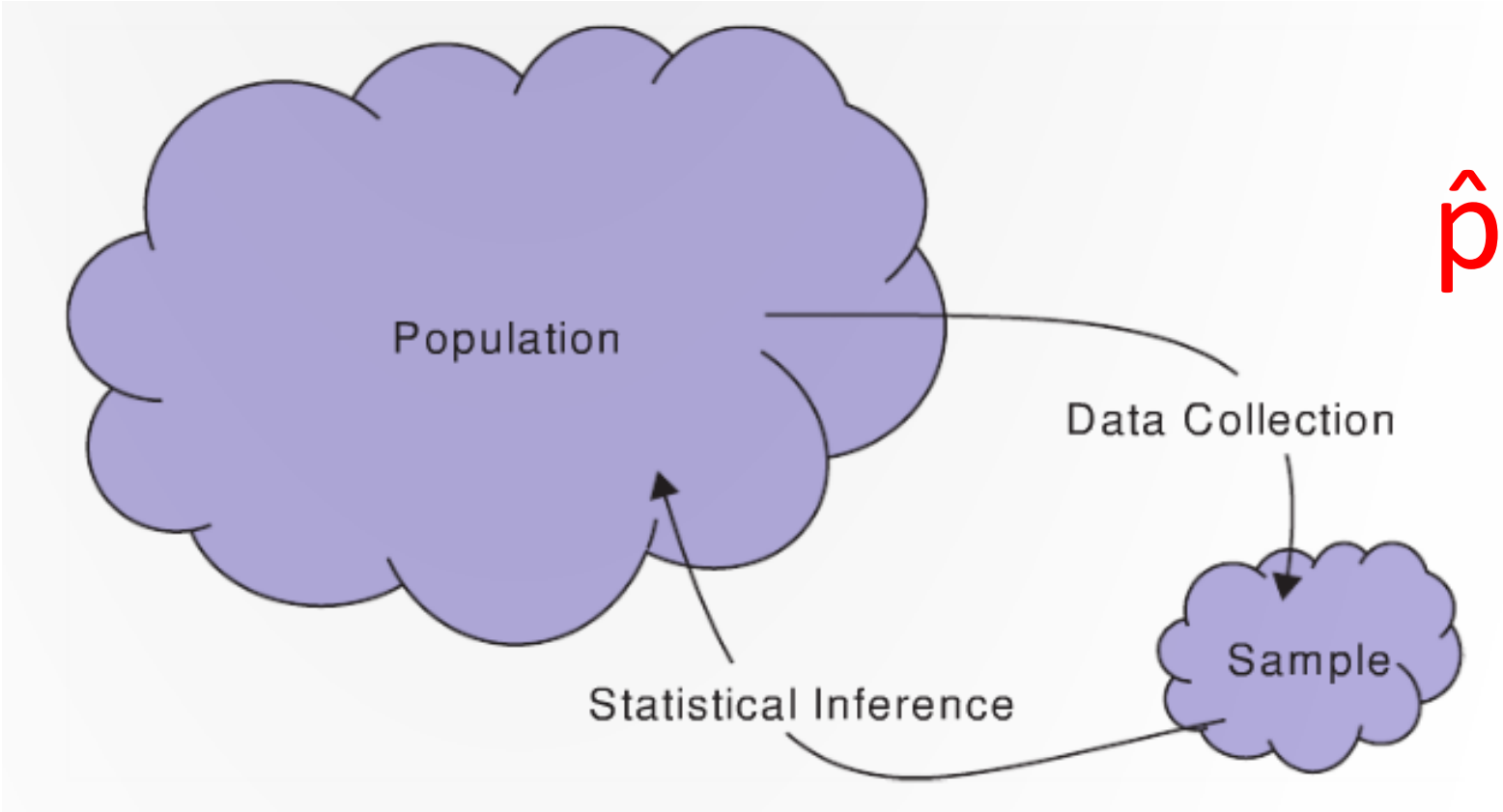


Sample: A subset of the population

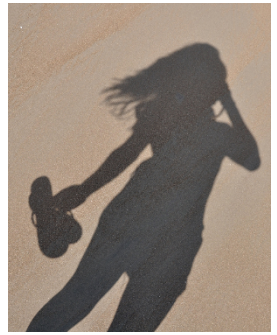


$\pi, \mu, \sigma, \rho, \beta$

Parameter: a number characterizing a property of a population

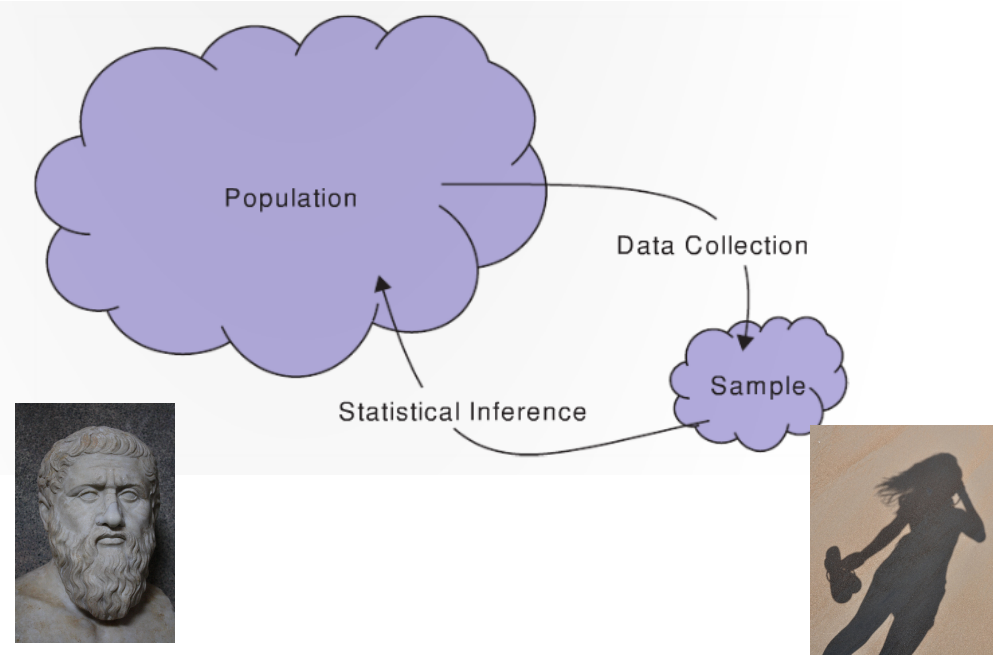


$\hat{p}, \bar{x}, s, r, b$



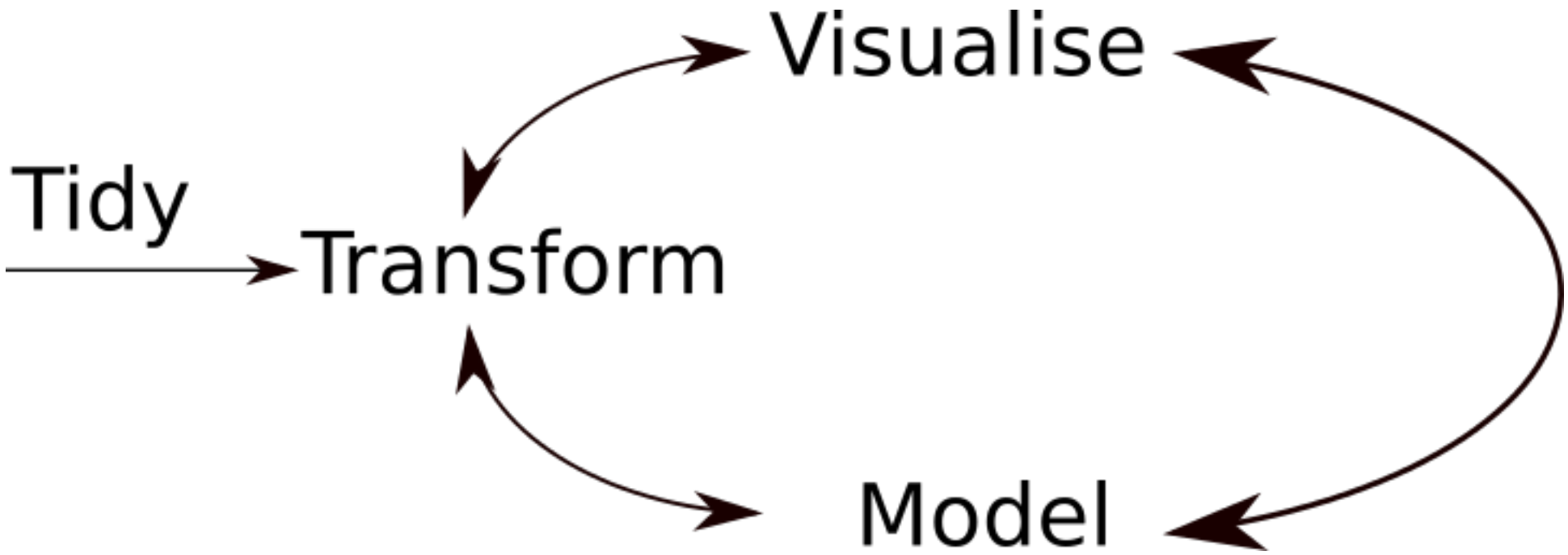
Statistic: A number computed from a sample

Parameters and statistics commonly used symbols



	Population parameter (Plato)	**Sample statistic (shadow)**
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	π	$\hat{p}$
Correlation	ρ	r
Regression slope	β	b

Sometimes the Truth is more complicated…



Questions?

Question

Q: What programming language do pirates use?

A: Arrrr

Q: Worst joke of the semester?

A: Wait and see…

R Basics

Does everyone have R and R Studio installed?

- Instructions and a video are on Canvas

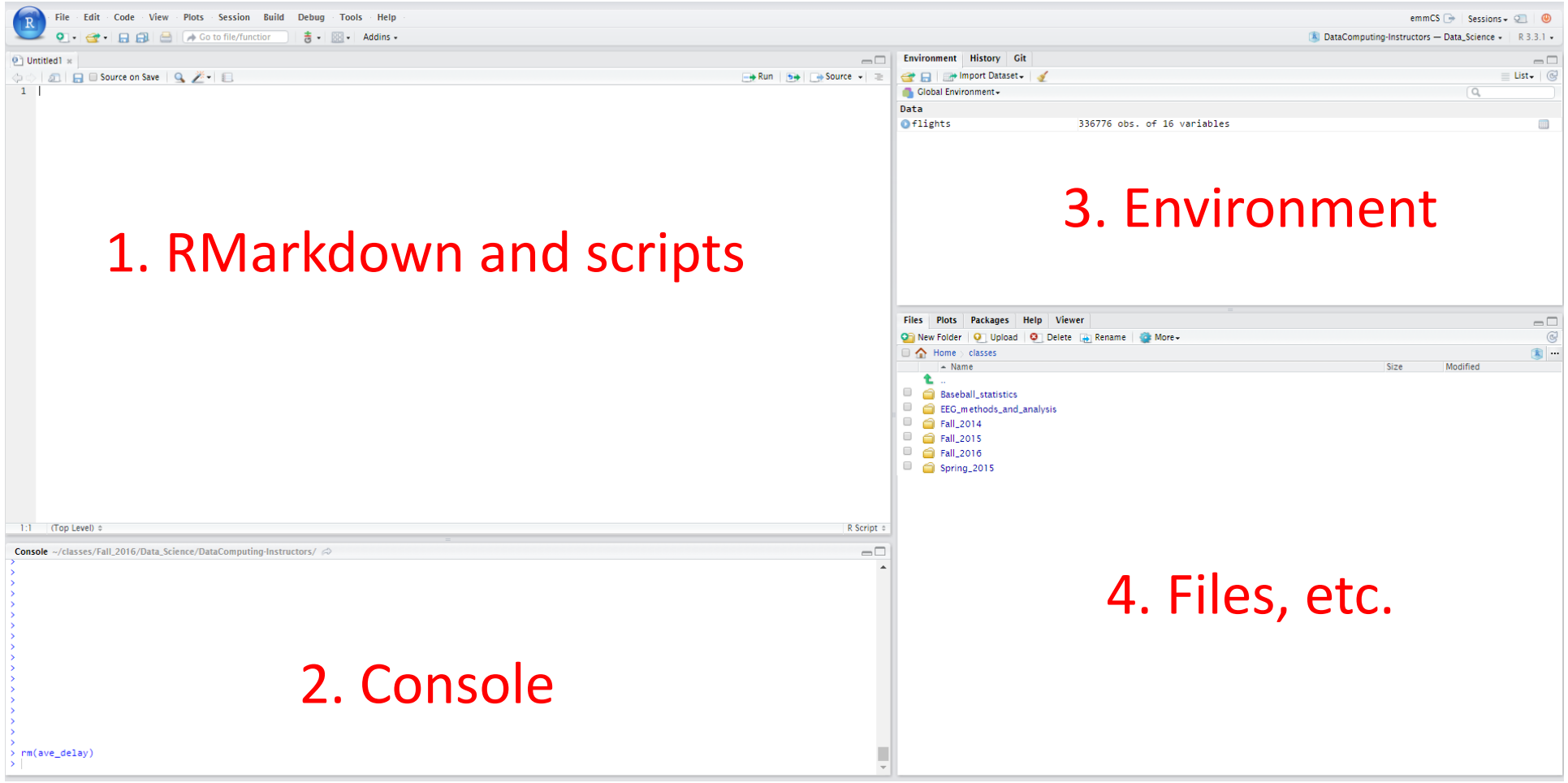
Let’s take a 5 minute break, and then open R Studio and follow along…

R and R Studio

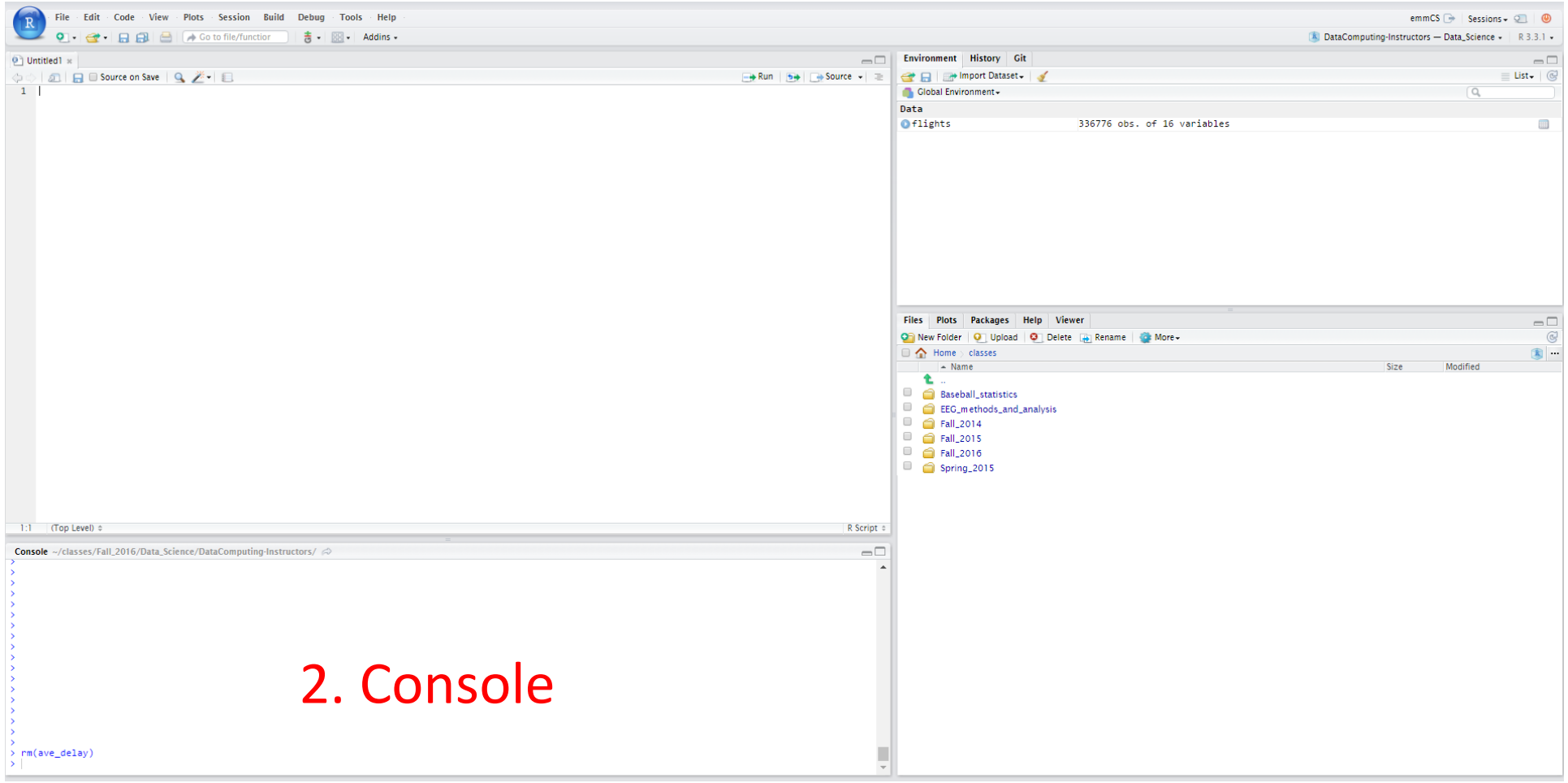
R: Engine

RStudio: Dashboard

RStudio layout



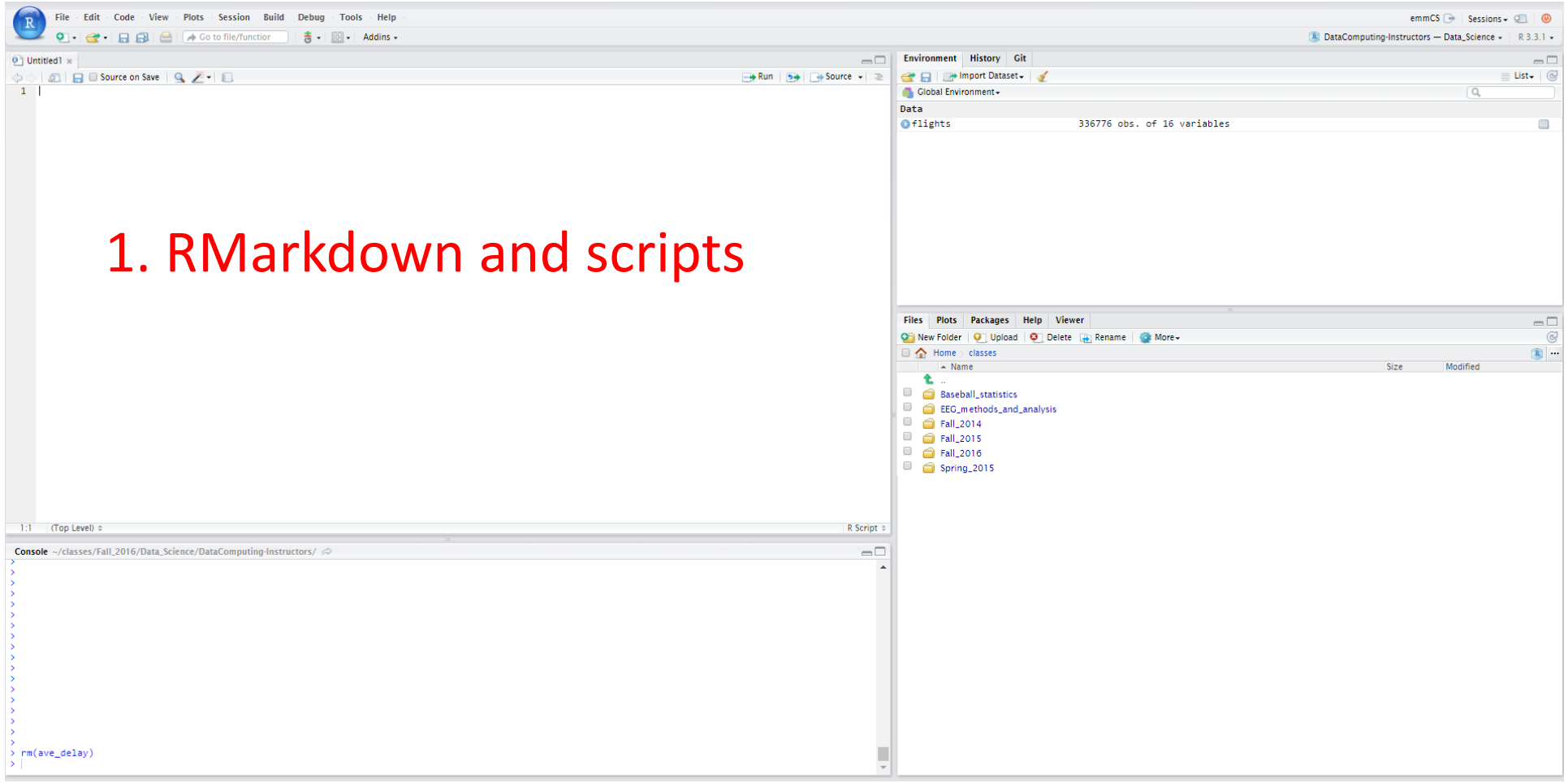
RStudio layout



R as a calculator

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

RStudio layout



Create a new script

File -> New File -> R Script

Save the script with a reasonable name, e.g., week1_notes.R

R Basics

Arithmetic:

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

Assignment of values to ***objects***:

```
> 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

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

R packages

Packages add additional functionality to R

We will use many additional packages in this class

- gplyr, ggplot2, tidyr, etc.

There is also a class specific package (SDS230) I wrote that you can use to download homework and other files

- All class materials are also on GitHub: https://github.com/emeyers/SDS230

Installing SDS230 package and LaTeX

To install the SDS230 package you first need to install the devtools package which can be done using:

```
install.packages("devtools")
```

You can then install the class SDS230 package using the function:

```
devtools::install_github("emeyers/SDS230")
```

Installing SDS230 package and LaTeX

Finally, after you have installed the SDS package, there is a function in the SDS package that installs LaTeX on your computer

- (this function uses the tinytex package)

To install LaTeX use:

```
SDS230::initial_setup()      # will install LaTeX via tinytex package
```

Test that the installation worked

```
tinytex:::is_tinytex()    # will return TRUE if it works  (note: 3 colons)
```

For next class

1. If you have not done so already
    - Fill out class survey on Canvas under the Quizzes link
    - Install R and RStudio if you have not done so already

2. Install the SDS230 class package and LaTeX