



# STOR 320 Introduction to Data Science

Lecture 1

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

# Instructor

- Name: Yao Li
- Email: [yaoli@email.unc.edu](mailto:yaoli@email.unc.edu)
- Office hours: Tuesday, Thursday 1:00PM to 2:00PM
- Personal website: <https://liyao880.github.io/yaoli/>
- Course website: <https://liyao880.github.io/stor320/>
- Research interest: adversarial deep learning, large-scale recommender systems, model compression

# Get to know your instructor

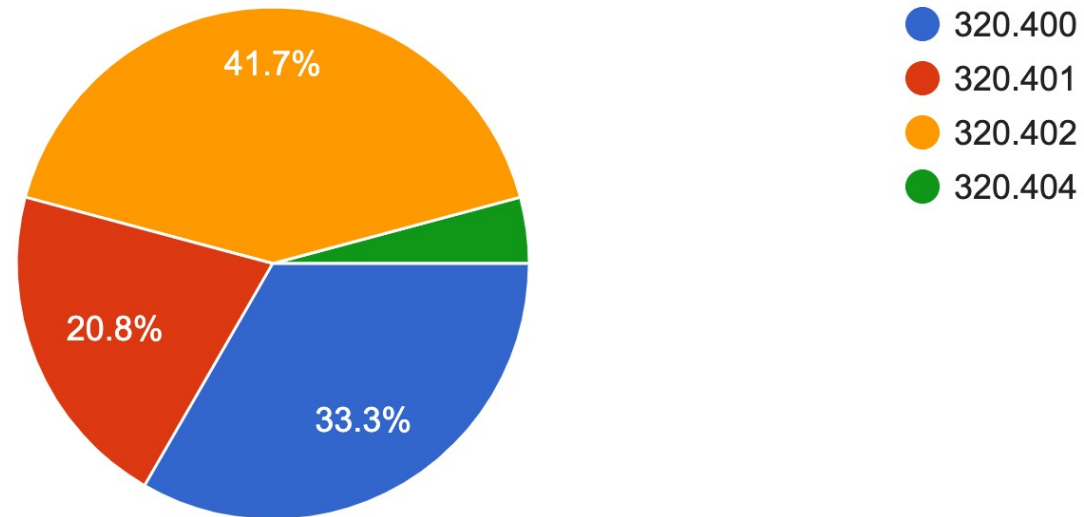
- Join at **[www.kahoot.it](http://www.kahoot.it)**



# Survey Results

In what lab section are you registered?

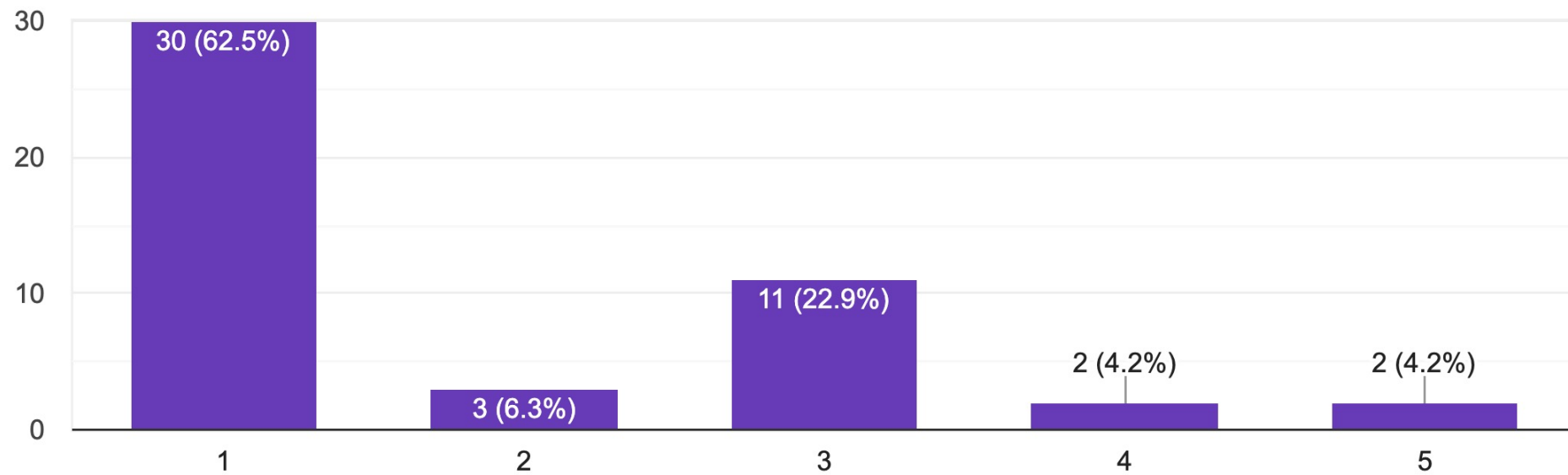
48 responses



# Survey Results

What is your experience with R?

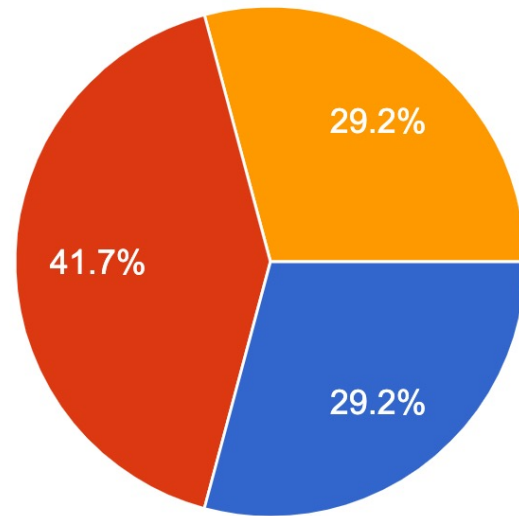
48 responses



# Survey Results

What do you plan to pursue after graduation?

48 responses



- Job Not in Field Related to Data Science
- Job in Field Related to Data Science
- Master's Degree
- Doctoral Degree
- Medical, Pharmaceutical, Dental, Nursing, or Veterinary Degree

# Instructional Assistant

- Shaleni Kovach (400, 401)
  - Email: [shaleni@email.unc.edu](mailto:shaleni@email.unc.edu)
  - Office Hours: Wednesday, Friday 3:30PM – 4:30PM
- Zichao Li (402, 404)
  - Email: [lizichao@live.unc.edu](mailto:lizichao@live.unc.edu)
  - Office Hours: Wednesday 2:30PM-4:30PM



# Lectures and Labs

- Lectures TTH 2:00 PM – 3:15 PM
- Labs
  - 400: W 4:40PM – 5:30PM H130
  - 401: F 4:40PM – 5:30PM H130
  - 402: W 1:25PM – 2:15PM H130
  - 404: W 4:40PM – 5:30PM H107
- Email Christine ([crikeat@email.unc.edu](mailto:crikeat@email.unc.edu))



# Outline

- Administrative details
- What's the course about?
- Introduction to R

# Instruction

- This will be an in-person course:
  - a) No chat in class;
  - b) lectures and labs will be held on campus;
  - c) face mask is required in class;
  - d) office hours will be held online;
  - e) laptop is required.

# Questions and Class Attendance

- Three ways to ask questions:
  - ask questions in class;
  - come to the virtual office hours on Zoom;
  - send an email to the instructor or the IAs.
- Class participation:
  - answer questions to get class participation grades;
  - 2.5 points each time.

# Grading

Class Participation	5%
Lab Attendance	5%
Labs	15%
Homework	45%
Final Project	30%

A	94 to 100	B	83 to 86.99	C	73 to 76.99	D	60 to 66.99
A-	90 to 93.99	B-	80 to 82.99	C-	70 to 72.99	F	0 to 59.99
B+	87 to 89.99	C+	77 to 79.99	D+	67 to 69.99		

# Homework and Labs

- Around 7 homework assignments and 4 data analysis assignments. They will be posted on Sakai and there will be about one week to complete the homework and about two weeks to complete data analysis assignments.
- Lab assignment:
  - Due 30 minutes after the lab ends.
  - No late submission will be accepted.
  - will be based on the topics discussed in lecture or related to your final project.

# Project

- For the final project, each section of STOR 320 will be divided into research groups of size 4 or 5. To ensure fairness, students will be assigned randomly based on lab session.
- The groups will be assigned by August 31, 2021 (Tuesday) and you can find your group on the course website.

# Project

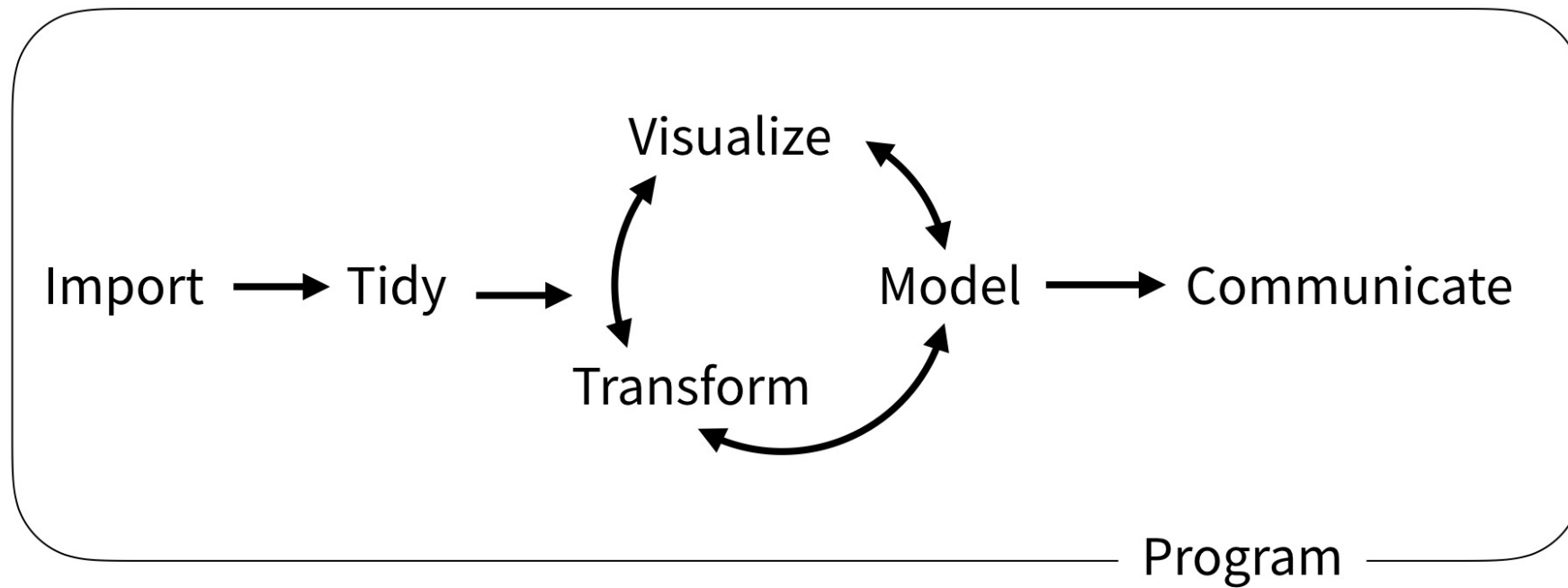
Project proposal	10%
Exploratory data analysis	20%
Final report	40%
Final presentation	30%



# Important dates

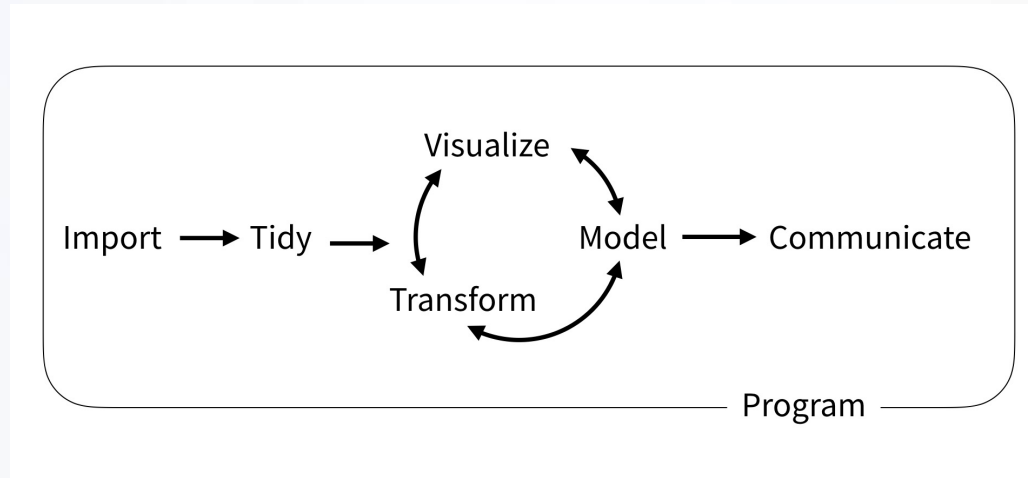
Project proposal	September 13
Exploratory data analysis	Oct 15
Final report	Nov 30
Final Presentation	Nov 23, Nov 30

# What is data science?



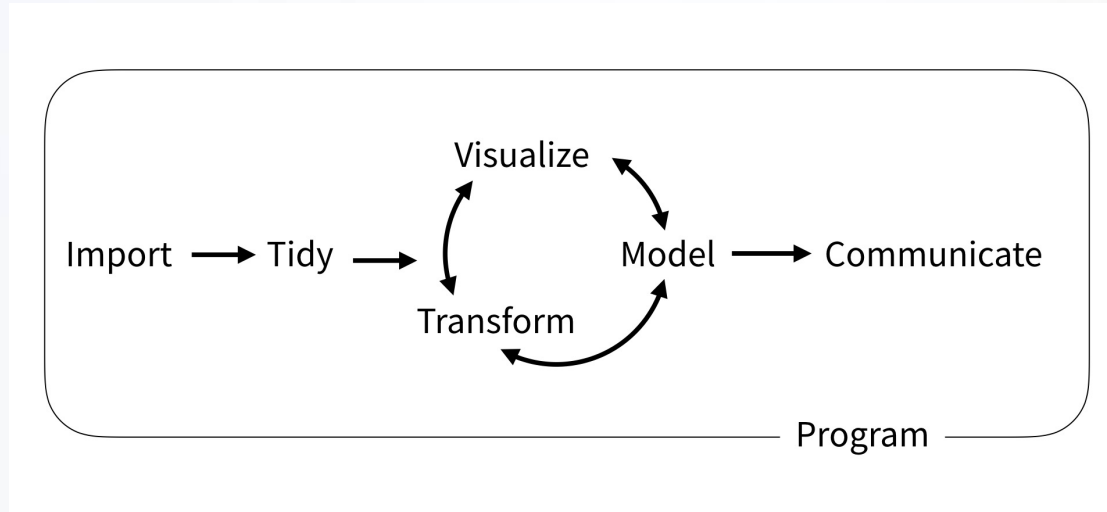
Wickham and Grolemund (2017)

# The model of data science



- First we must *import* our data.
- *Tidy* data → consistent structure
- Transformation:
  - narrowing in on observations of interest
  - creating new variables
  - calculating a set of summary statistics

# The model of data science



- *Visualization*: show you things that you did not expect or raise new questions about the data.
- Use a *model* to answer your questions
- *Communication*: an absolutely critical part of any data analysis project.
- Surrounding all these tools is programming.

# R and RStudio



~/Documents/rmarkdown - gh-pages - RStudio

5-parameters.Rmd x

```
1 ---
2 title: "Visualizing the ocean floor"
3 output: html_document
4 params:
5   data: "hawaii"
6 ---
7
8 ```{r include = FALSE}
9 library(marmap)
10 library(ggplot2)
11 ```
12
13
14 The [marmap](https://cran.r-project.org/web/packages/marmap/index.html) package provides tools and data for visualizing the ocean floor. Here is an example contour plot of marmap's ``r
15 params$data`` dataset.
16
17 ```{r echo = FALSE}
18 data(list = params$data)
19 autoplot(get(params$data))
20 ```
21
22 21:1 (Top Level) R Markdown
```

Environment History Build Git

Files Plots Packages Help Viewer

## Visualizing the ocean floor

The `marmap` package provides tools and data for visualizing the ocean floor. Here is an example contour plot of marmap's `aleutians` dataset.

A contour plot of the ocean floor showing the Aleutian Islands. The x-axis is labeled 'X' and ranges from 170 to 210. The y-axis is labeled 'Y' and ranges from 50 to 65. The plot shows a series of contour lines representing the depth of the ocean floor, with the Aleutian Islands visible as a series of peaks and valleys.

Console R Markdown x

```
~/Documents/rmarkdown/demos/
> render("5-parameters.Rmd", params = list(data = "aleutians"))
```

# Why R?

- Easy to learn and easy to use.
- Very popular and one of the standard languages for statistics, data science, computational biology, finance, industry, etc.
- Free and open-source.
- A lot of high-quality packages.
- New technology and ideas often appear first in R.
- Supported by a vast community that maintains and updates R.
- Runs on basically any platform.

# Learning Programming

- Transfer the concepts to other languages
- How you approach a computational task and reason about the computations is similar
- Learning another programming language will be much easier in the future



# Statistical Learning

- Linear regression.
- Classification (logistic regression, LDA, K-nearest neighbors).
- Cross-validation and bootstrap.
- Principal component analysis.
- Clustering methods (K-means clustering and hierarchical clustering).
- Recommender systems.
- Neural networks.

# Textbooks

- *R for Data Science*. Hadley Wickham. Legally free online, but can be purchased for less than \$40 on Amazon. Additional suggested texts are provided on the website. All texts used in this course are free and downloadable from course website.
- *The elements of statistical learning: data mining, inference, and prediction*. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.