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
- Office hours: Tuesday, Thursday 10:00AM to 11:00AM
- 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**

Lectures and Labs

- Lectures TTH 11:30 AM - 12:45 PM

- Labs
  - 400 Friday 10:40AM – 11:30AM (FF/Hy)
  - 401 Friday 12:00PM – 12:50PM (Remote only)
  - 402 Friday 1:20PM – 2:10PM (FF/Hy)

- Email Christine (crikeat@email.unc.edu)

Instructional Assistant

- Kevin O Connor (401)
  - Email: koconn@live.unc.edu
  - Office Hours: TH 4:00 PM-5:00 PM; F 1:00 PM-2:00 PM
- Pavlos Zoubouloglou (402)
  - Email: pavlos@live.unc.edu
  - Office Hours: M 9:00 AM-10:00 AM; F 2:25 PM-3:25 PM
- Sam Booth (400)
  - Email: slbooth@live.unc.edu
  - Office Hours: W 3:00 PM to 5:00 PM

Outline

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

Ask Questions in Class

- By default, your microphone will be muted.
- If you have a question, feel free to unmute yourself and ask questions.
- Also, you can type your question in the in-meeting chat window.

Remote Instruction

- This will be a hybrid course:
  a) lectures will be held live online during the scheduled time and recorded so that you can watch them later;
  b) lab session 401 will be online and recorded, the other two labs will be held face-to-face in classrooms;
  c) some of the lectures might be prerecorded if there are connection issues and livestreaming is not possible;
  d) office hours will be held online but not recorded;
  e) all assignments will be done remotely.

Questions

- Three ways to ask questions:
  - post questions on Sakai forum;
  - come to the virtual office hours on Zoom;
  - send an email to the instructor or the IAs.

Grading

Lab Attendance	10%
Labs	30%
Homework	30%
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 4 homework assignments. They will be posted on Sakai and there will be about one week to complete the assignment.

- 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 (ideally) 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 28, 2020 (Friday) and you can find your group on shared via google sheet.

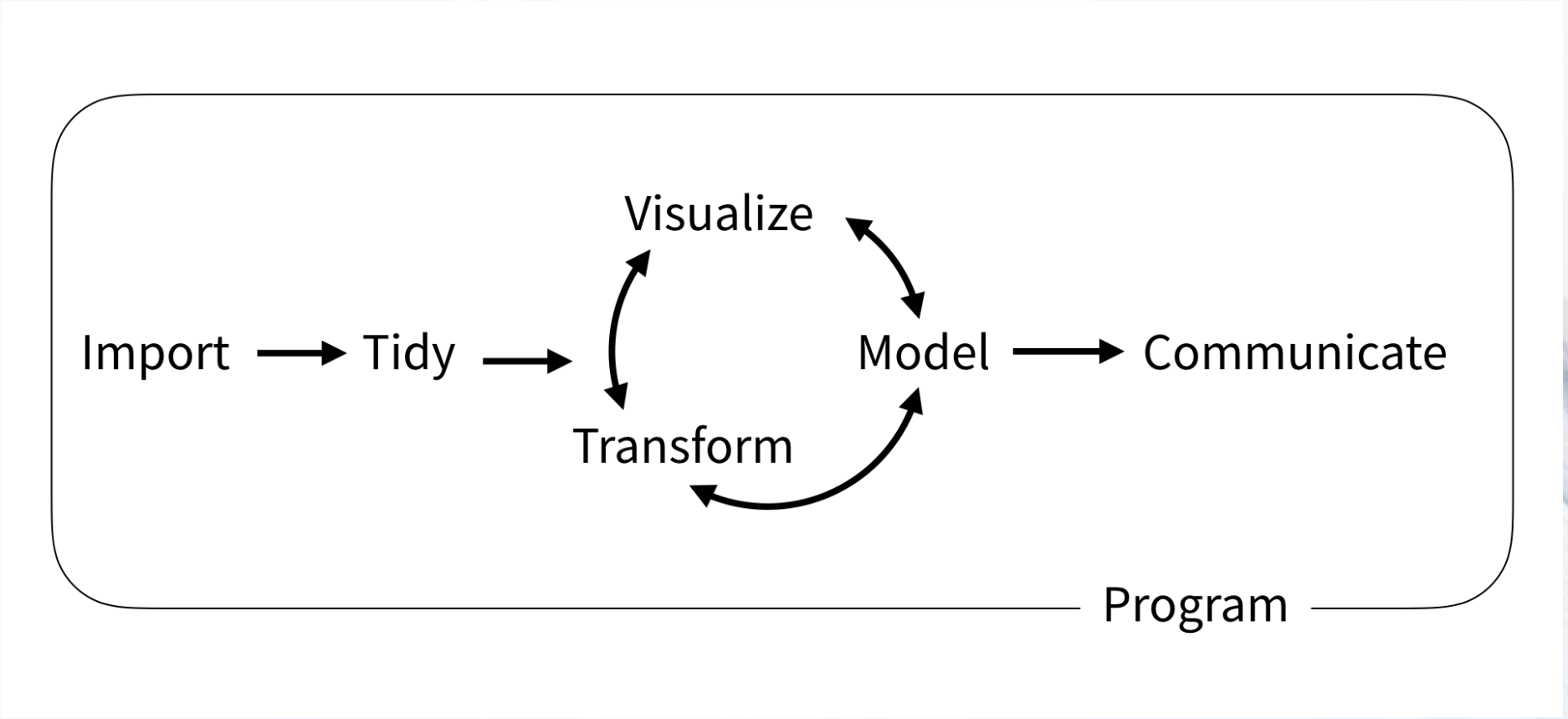
Project

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

Important dates

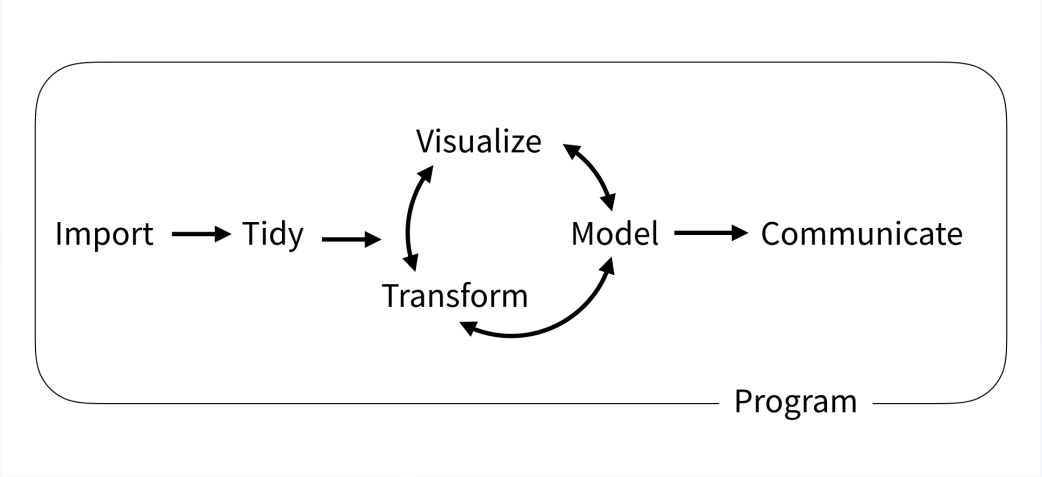
Project proposal	September 15
Exploratory data analysis	October 9
Final report	November 13
Final Presentation	November 12 or November 17

What is data science?



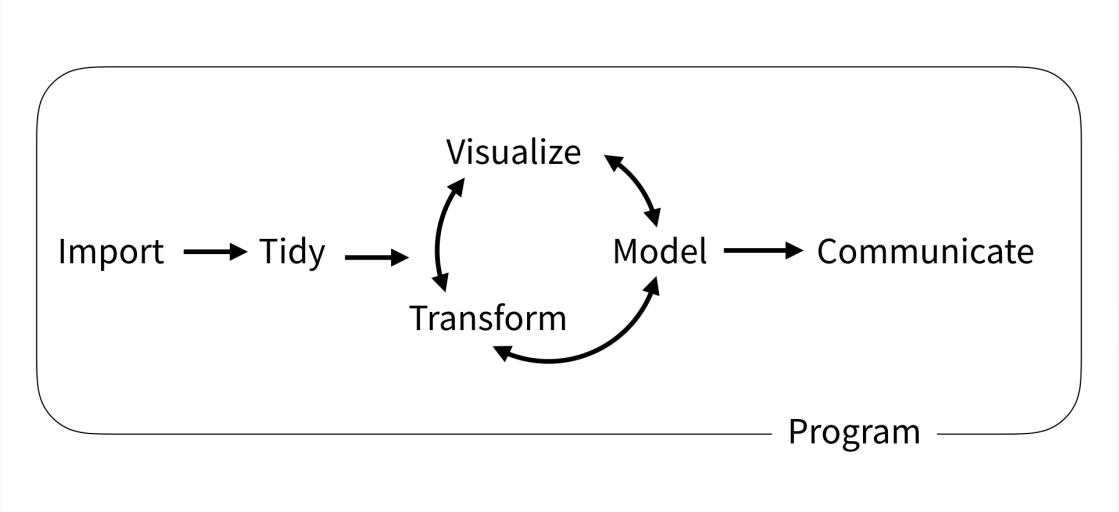
Wickham and Grolemund (2017)

The model of data science



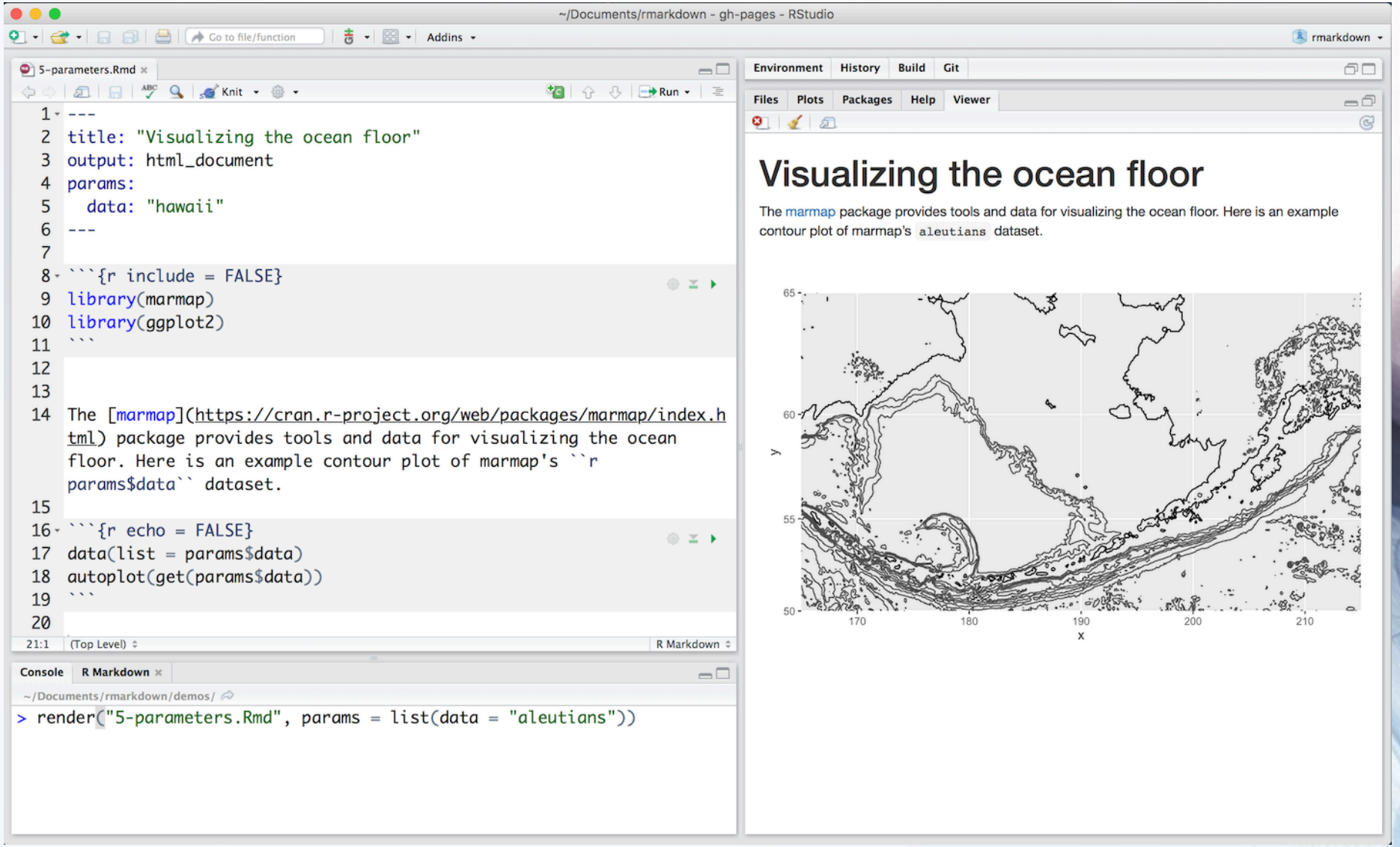
- First we must *import* our data.
- *Tidy* data → consistent structure
- Transformation:
  - narrowing in on observations of interest
  - creating new variables
  - calculating a set of summary statistics

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



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.