Welcome to Intro to Data Science

Jacob Smith

Welcome!

What is Data Science?

Data science is an emerging discipline that builds on tools from mathematics, statistics, and computer science to extract knowledge from data.

Course Objectives

- explore, visualize, and analyze data in a reproducible manner
- investigate patterns, model outcomes, and make predictions
- gain experience in data wrangling and munging, exploratory data analysis, predictive modeling, and data visualization
- work on problems and case studies inspired by real-world questions and data
- effectively communicate results

Where to find information

Course Website: https://sta199-spring2022.netlify.app/

- slides and notes
- schedule
- syllabus and course policies
- links to other resources

Sakai: https://sakai.duke.edu

- Gradebook
- Zoom links to live lecture, office hours, and labs
- recorded videos of lecture
- links to other resources like Ed Discussion

GitHub: https://github.com/orgs/sta199-spring2022/

- assignment repos

Activities and Assessments

- The activities and assessments in this course are designed to help you successfully achieve the course learning objectives. They are designed to follow the Prepare, Practice, Perform format.

- **Prepare:** Includes short videos, reading assignments, and a short quiz to introduce new concepts and ensure a basic comprehension of the material. The goal is to help you prepare for the in-class activities during lecture.

- **Practice:** Includes in-class application exercises where you will begin to the concepts and methods introduced in the prepare assignment. the activities will graded for completion, as they are designed for you to gain experience with the statistical and computing techniques before working on graded assignments.

- **Perform:** Includes labs, homework, exams, and the final project. These assignments build upon the prepare and practice assignments and are the opportunity for you to demonstrate your understanding of the course material and how it is applied to analyze real-world data.

Grade Components

- **Video Prep Quizzes (5%):** Short, weekly video quizzes on videos; due before class, will have three attempts, an overall score of 80% will result in full credit.

- **Homework (25%):** Five individual assignments combining conceptual and computational skills.

- **Labs (15):** Nine individual or team assignments focusing on computing. Designed to be completed during the official lab session.

- **Exams (35%):** Two individual take-home exams.

- **Final Project (15%):** Team final project in which you use the data science tools to answer a data-based research question.

- **Participation and Teamwork (5%):** Primarily completion of lecture notes. Due three days following the lecture date.

Course Structure

Lecture

- Focus on concepts behind data analysis
    - TR from either 10:15 to 11:30 or 3:30 to 4:45
    - A lecture notes R Markdown file will be created for you for each lecture.
    - We focus most of our time in class practicing coding.
    - Videos and/or short reading assignments posted in advance for that week.

Lab

- Focus on computing in R `tidyverse` syntax
- Apply concepts from lecture to case study scenarios
- Work on labs individually or in teams of 3-4 after add-drop.
- Designed to be completed during the scheduled lab time; will generally be due on Friday.

Graduate Lab TAs:

- 01L: Alison Reynolds (alison.reynolds@duke.edu)
- 02L: Alison Reynolds (alison.reynolds@duke.edu)
- 03L: Nathan Varberg (nathan.varberg@duke.edu)
- 04L: Kevin Li (kevin.li566@duke.edu)
- 05L: Zoey Liu (zheyuan.liu@duke.edu)
- 06L: Emre Yurtbay (emre.yurtbay@duke.edu)

Some of what you will learn

- Fundamentals of `R`
- Data visualization and wrangling with `ggplot2` and `dplyr` from the `tidyverse`
- Spatial data visualization
- Data types and functions
- Version control with `GitHub`
- Reproducible reports with `R Markdown`
- Regression and classification
- Statistical inference
- Some special topics, as time permits

Textbooks

- Introduction to Modern Statistics
  - Free online
  - Hard copies available for purchase
  - Assigned readings on statistical content
- R for Data Science
  - Free online
  - Hard copies available for purchase
  - Assigned readings on R coding using `tidyverse` syntax
  - A great resource for all things R!
- **Occasional other readings** will be posted on the course webpage

Where to find help in the course

- Attend **office hours** to meet with a member of the teaching team.
- Use **Ed Discussion** for general questions about course content and/or assignments, since other students may benefit from the response.
- Ask questions during lecture.
- Use **email** for questions regarding personal matters and/or grades.

Contact Policy

- Students with questions that focus on class context should begin by posting on Ed Discussions.
- Students with more specific questions that are not of a sensitive nature should first reach out to their Lab TA If a student feels that they would like further elaboration, they may second ask that the TA send the email to the Head TA. If the student still feels that they would like further elaboration they may third ask that the Head TA send the request to me.
- Students should contact me first in the case of a matter that is sensitive (including involving an extension), something involving an accommodation, or asking for their one no excuse late assignment.

Office Hours

- I will hold office hours on Zoom on Tuesdays from Noon to 1 and Thursdays from 2 PM to 3 PM on Zoom. Students interested in meeting on Zoom should sign up at https://calendly.com/jacobfhsmith.

- Appointments are 15 minutes and students may sign up for up to two slots per day. Students are welcome to sign up in groups if they have a similar question or if it is a group assignment.

- TAs will also hold office hours beginning **Thursday January 6.** Zoom links for those office hours are available here

Two other COVID-related policies

- I am willing to answer shorter questions as time permits after class and will try set aside around 10 minutes at the end of class to answer questions. I ask that rather that crowding the front of the room after class that you remain in your seat and raise your hand if you have a question.

- Per Duke policies, you should wear a face mask at all times during class. Please do not eat or drink during class. I understand that this is a long class; if you need a sip of water, please step out of the classroom and then return.

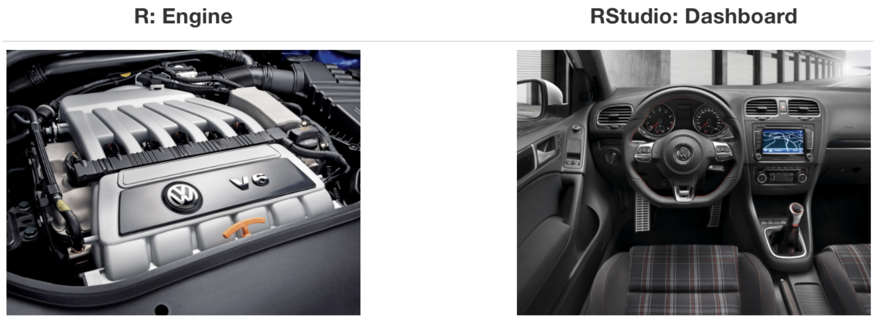
- I understand that there are medical conditions that may require you to take a quick sip or eat a quick bite; in these cases please do so as quickly as possible and please ask SDAO to send documentation to that effect when you have a chance. (I understand that can take time.)

Toolkit

- **R and RStudio**
- **GitHub**

What is R and RStudio?

- R is a statistical programming language
- RStudio is a convenient interface for R (an integrated development environment, IDE)
- At its simplest:[*]
    - R is like a car’s engine
    - RStudio is like a car’s dashboard



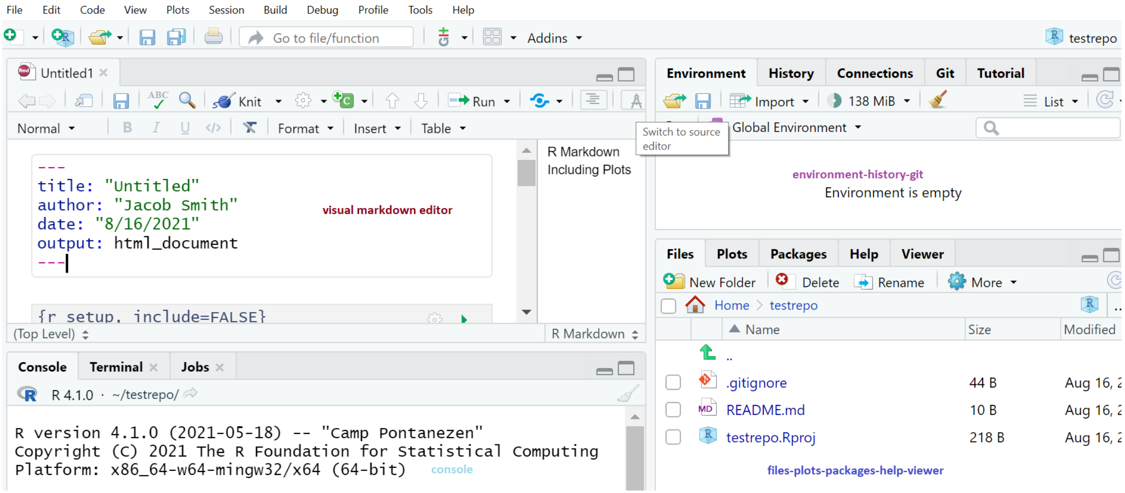
[*] Source: Modern Dive

tidyverse

- The tidyverse is an **opinionated*** collection of R packages designed for data science.
- All packages share an underlying philosophy and a common grammar.

Image from Teaching in the Tidyverse 2020

RStudio



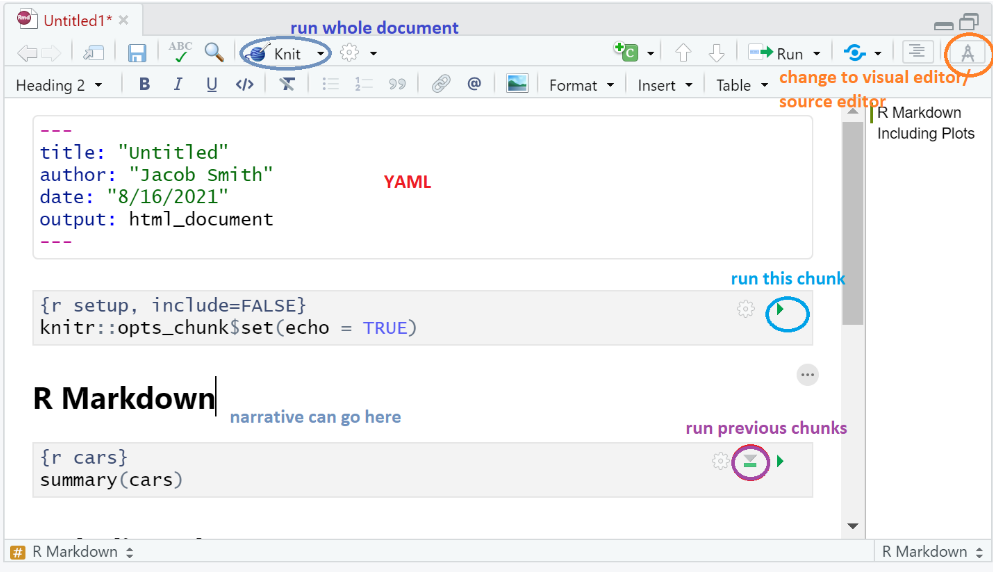
```
---
title: "Untitled"
author: "Jacob Smith"
date: "8/16/2021"
output: html_document
---

{r setup, include=FALSE}
```

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

Name	Size	Modified
..		
.gitignore	44 B	Aug 16, 2
README.md	10 B	Aug 16, 2
testrepo.Rproj	218 B	Aug 16, 2

Visual Editor



run whole document

change to visual editor/ source editor

YAML

```
---
title: "Untitled"
author: "Jacob Smith"
date: "8/16/2021"
output: html_document
---
```

run this chunk

```
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

narrative can go here

run previous chunks

```
{r cars}
summary(cars)
```

Accessing RStudio

- Link: https://cmgr.oit.duke.edu/containers/sta198-199
  - Link also on the top of the course webpage

Let's try!

Reproducibility checklist

What does it mean for a data analysis to be "reproducible"?

Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear **why** it was done?

Long-term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

R Markdown

R Markdown

- Fully reproducible reports -- the analysis is run from the beginning each time you knit
- Simple Markdown syntax for text
- Code goes in chunks, defined by three backticks, narrative goes outside of chunks

How will we use R Markdown?

- Every assignment / lab / project / etc. is an R Markdown document
- You'll always have a template R Markdown document to start with
- The amount of scaffolding in the template will decrease over the semester

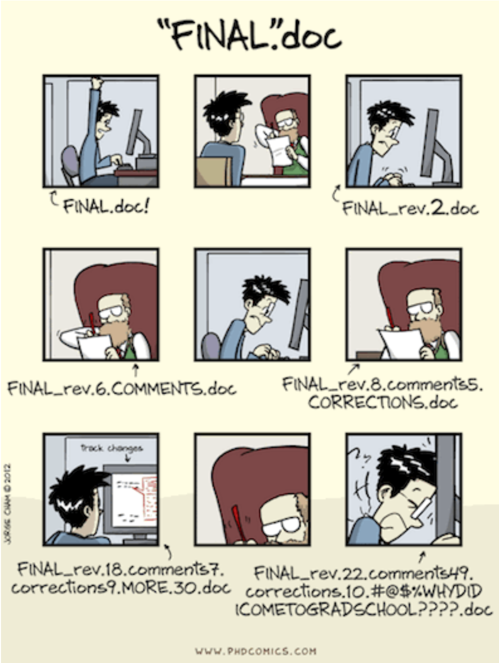
Let's try!

Git and GitHub

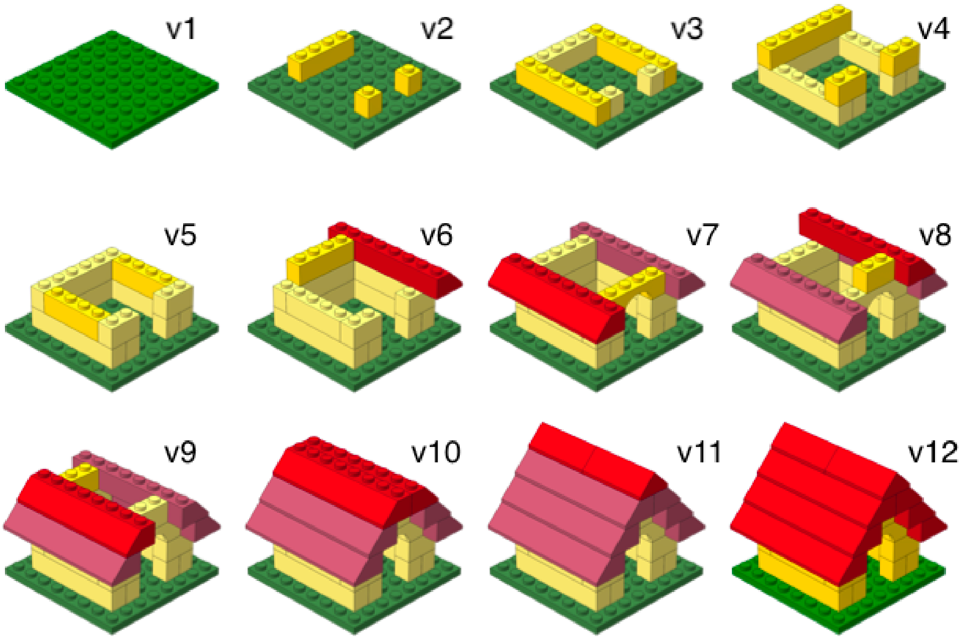
Version control

We will use GitHub as a platform for collaboration and version control.

Why do we need version control?

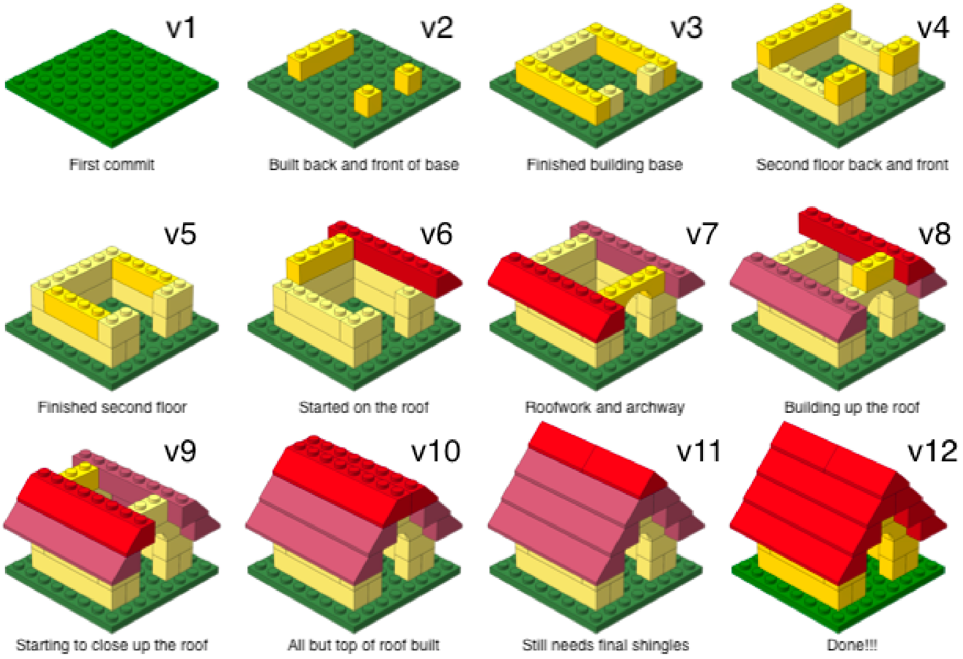


What is versioning?



What is versioning?

with human readable messages



- **Git** is a version control system -- like “Track Changes” features from Microsoft Word.
- **GitHub** is the home for your Git-based projects on the internet (like DropBox but much better).
- There are a lot of Git commands and very few people know them all. 99% of the time you will use git to stage, commit, push, and pull.

Let's try!

Step #1. Create a repository on GitHub

repository: contains files associated with a particular project and each file's version history

(a) Click the link below to create the repository for the test assignment (you do not have to turn anything in from this).

- https://classroom.github.com/a/EI_jAChV
- You will be prompted to create a Github account if you did not already for lab. This is free and you should create one with both your first and last name (not NET ID).

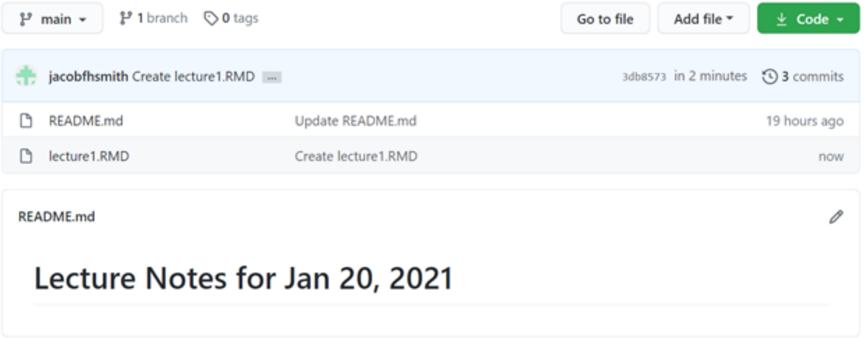
(b) When prompted, select "Accept this assignment".

(c) Refresh the page.

(d) Click the link following "Your assignment repository has been created:".

A private repository was just created for you.

Step #1. Create a repository on GitHub



jacobfhsmith Create lecture1.RMD ... 3db8573 in 2 minutes 3 commits

README.md	Update README.md	19 hours ago
lecture1.RMD	Create lecture1.RMD	now

README.md

Lecture Notes for Jan 20, 2021

Step #2: Configuring SSH and GitHub

Until recently, you could use a user name and password to log into GitHub.

GitHub has deprecated using a password in that way. Instead, we will be authenticating GitHub using public/private based keys.

This is a short overview for how to do authenticate in this way.

Before doing this, save and close any existing projects (e.g., the first lab).

1. First, type `credentials::ssh_setup_github()` into your console.

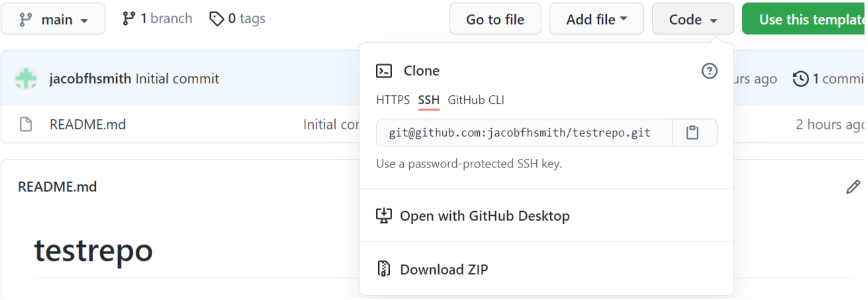
2. Second, R will ask "No SSH key found. Generate one now?" You should click 1 for yes.

3. Third, you will generate a key. It will begin with "ssh-rsa...." R will then ask "Would you like to open a browser now?" You should click 1 for yes.

4. Fourth, you may be asked to provide your username and password to log into GitHub. This would be the ones associated with your account that you set up. After entering this information, you should paste the key in and give it a name.

Step #3. Clone a GitHub repo & Make a new RStudio project

(a) In the repository that was just created, click the green "CODE" button, select "Use SSH" and click the clipboard icon to copy the repo URL.



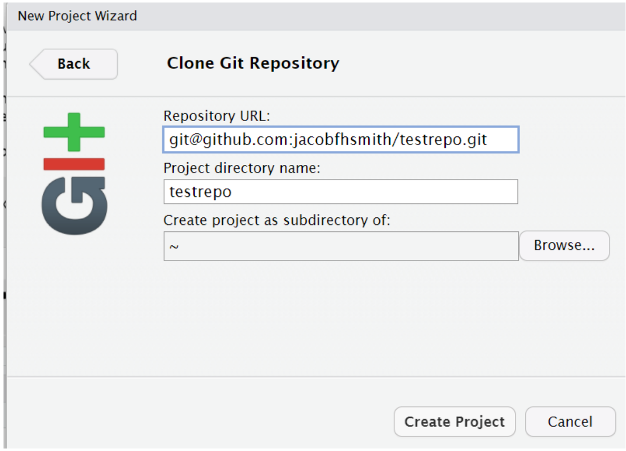
```
git@github.com:jacobfhsmith/testrepo.git
```

Use a password-protected SSH key.

Step #3, cont. Clone a GitHub repo & Make a new RStudio project

(b) In RStudio, click "File" --> "New Project" --> "Version Control" --> "Git".

You should see something like this...



Clone Git Repository

Repository URL: git@github.com:jacobfhsmith/testrepo.git

Project directory name: testrepo

Create project as subdirectory of: ~

(c) Copy and paste the URL of your assignment repo in the dialog box "Repository URL:".

(d) Click "Create Project" and enter your GitHub username and password when prompted.

(e) The files from your GitHub repo should now be displayed in the "Files" pane in RStudio.

The "Project" drop-down menu in the upper-right should show the project you are currently working on.

Step #3, cont.

In the **terminal**, you should type

```
git config -- global user.name 'username'
git config -- global user.email ‘useremail’
```

For example, mine would be

```
git config -- global user.name 'jacobfhsmith'
git config -- global user.email ‘jacob.f.smith@duke.edu’
```

Step #4. Make a change locally

(a) Click the R Markdown file named example_day1.Rmd in the lower-right pane.

(b) Change the author name to your name.

(c) Knit the document.

Examine this file. What changed?

Step #5. Stage and Commit

stage: prepare file for commit

commit: save changes to local repository

(a) Open the "Git" pane in the top-right panel.

(b) Click on the .Rmd file

(c) Click "Diff" to see the difference between the last commit and the current state

(d) If you're happy with the changes, stage them by checking the box next to the file.

(e) Write a meaningful commit message "changed author" in the commit message box.

(f) Click commit

Step #6. Push these changes to the repo on GitHub.

- push: upload files to remote repository (gitHub)

(a) Click "Push".

- Now go to the repo on GitHub. What do you notice?

Vocabulary

- **pull**: update a local repository from a remote repo (GitHub)
- **stage**: prepare file for commit
- **commit**: save changes to local repository
- **push**: upload files to remote repository (gitHub)

As you work on a data science project, you should periodically knit, stage, commit, and push.

If you are working on a team, there may be updates to your GitHub repo that aren't in your local repo. To make sure you are starting with the most up-to-date files, click "Pull" to update your local repo before adding any new work.

Review of steps

- Step #1. Create / navigate to the repository on GitHub
- Step #2. Configuring SSH and GitHub
- Step #3. Clone the GitHub repo & make a new RStudio project
- Step #4. Make a change locally
- Step #5. Stage and commit
- Step #6. Push these changes to the repo on GitHub.

For next class:

- Several videos to watch under "Prepare" on Syllabus page of website. There are also some optional but not required readings.