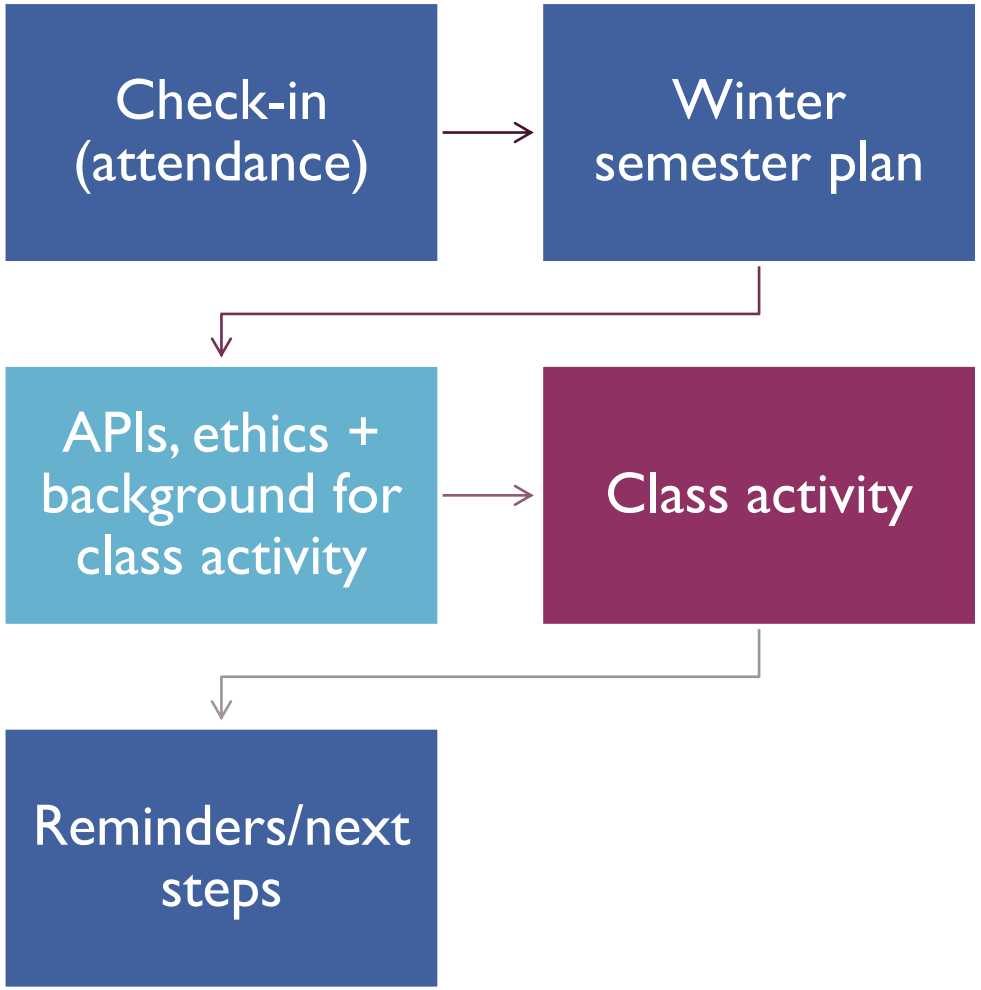
STA490: STATISTICAL CONSULTATION, COMMUNICATION AND COLLABORATION

DO GOOD WORK, WITH OTHERS, AND TALK ABOUT IT WELL

November 26, 2020

PLAN FOR TODAY



Key	Breakouts	Class discussions
Breaks	Information	Admin

CHECK-IN (ATTENDANCE)

- Go to pollev.com/bolton and choose your mood for today

One-word check-in (attendance) → Breakout: Littering case study → Class discussion: Littering case study → Asynchronous writing activity → Reminders/next steps

WINTER SEMESTER PLAN

CHANGES TO WINTER SEMESTER PLAN

As we hope you've heard, winter semester is starting one week later than originally planned. We've made some changes to the due dates for some of the the coursework in line with this. You can find the updated schedules

- L0101 link: https://q.utoronto.ca/courses/183070/pages/weekly-materials-winter-updated
- L0201 link: https://q.utoronto.ca/courses/183086/pages/weekly-materials-winter-updated

- If you have any concerns about these changes, please reach out to Prof Bolton (L0101) or Prof Moon (L0201) before our Thursday class meeting next week and we will work with you to find a solution. If we don't hear from anyone, these changes will be official as of next Thursday.

"EXPLAIN LIKE I'M A…" PRESENTATIONS

- For these presentations, the goal is to practice explaining a concept at a range of levels, using appropriate language, analogies, examples, etc. This activity was inspired by WIRED's '5 levels (Links to an external site.)' videos (these are available on YouTube, but in no way required watching for this assessment).
- Each group will have 4 members and produce a pre-recorded video. There are three dates these videos will be due and we'll have a 'film festival' in class that week. Attendance still mandatory, there will be participation activities required. Your grade will be awarded as a group.
- **You will have an opportunity to self-sign up for a topic and date. Self-sign up will be available from 10:10 a.m. ET, Friday, 27 November and will be first-come, first-serve.**

[L0101 link] [L0201 link]

"EXPLAIN LIKE I'M A…" PRESENTATIONS

Video

You will need to cover the following 4 prompts in your video.

1. Using only words from the list of the 1000 most common English words, describe your assigned concept. This 'checker (Links to an external site.)' from XKCD will help you. (This is the comic that inspired it all (Links to an external site.)). For example, this can be displayed on a slide and read through. Briefly comment on any particular challenges or ways you solved communicating this concept with these words.
2. Explain this concept to a 10-year old child.
3. Explain this concept to a first-year statistics student who likes mathematics. (i.e. demonstrate or draw on some of the mathematics related to the concept).
4. Explain this concept to a researcher with an advanced degree in a *non*-statistics subject (someone like your project collaborators!).

Group evaluation

All students will have an evaluation task to do after their presentations about how their group worked together to create the submission. **While ungraded, there is a 5 percentage point penalty for not completing this.** A link will be made available after your presentation. Please let us know as early as possible if your group is having any challenges working together.

"EXPLAIN LIKE I'M A…" PRESENTATIONS

- These videos will be **pre-recorded** but with an opportunity for live Q&A when we have our 'film festival' in class
  - You can use Zoom and have one person share the slides and record that meeting or if you like to edit videos you can do however else you see fit. You won't lose marks for not being a superstar video editor, we care most about your content).
- **Length**: These videos should be no *less* than 12 minutes long and no *more* than 15 minutes long.
- Use some of the time to introduce yourselves. Everyone in your group should **speak** in the video for approximately the *same* amount of type. Not making a meaningful speaking contribution will affect your score. That said, you don't need to assign every person a specific bullet point above. E.g. bullet 1 might be quite short and that person might also have some dialogue helping with point 4, etc.
- You have lots of freedom when it comes to your visual aids. You can use slides, draw illustrations/annotations, act things out with props, or anything else you like. Do make sure there is both a compelling visual and oral aspect to the video.
- You can upload the video to MyMedia. and then provide the link to me in the assignment dropbox.

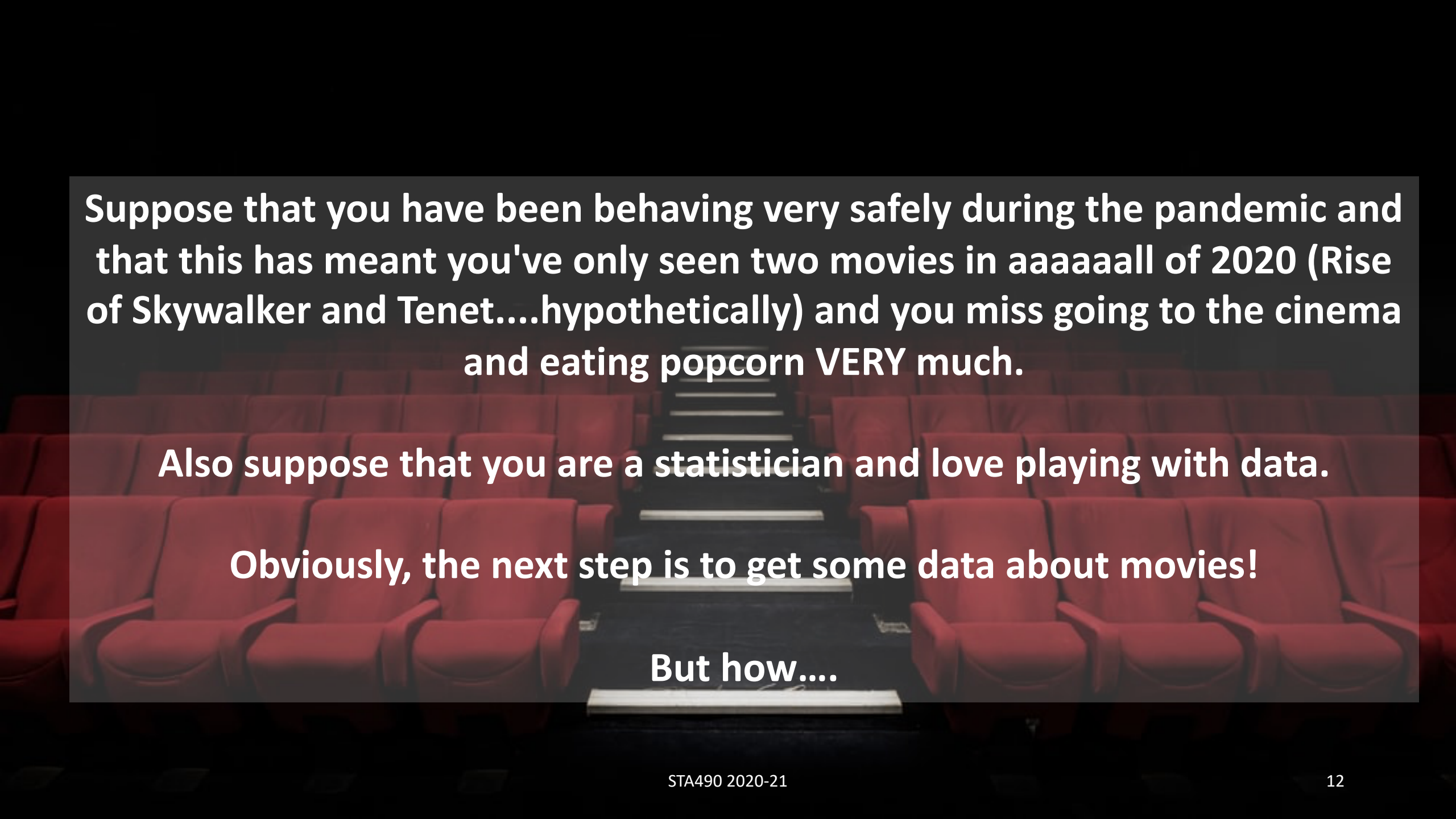
APIS, WEB SCARPING AND THE ETHICS OF GETTING DATA (PART 1)

The ‘big picture’ for today

We’re going to end up with some data about movies to play with…but you know they say the journey is just as important as the destination

GETTING DATA FROM THE INTERNET

- Weather, sports, stock prices, house prices, social media sentiment, transcripts of legislative meetings… you name it, there is data about it some where on the internet.
- BUT! Rarely is it nicely formatted in a .csv file that you can download and start working with in R right away.



Suppose that you have been behaving very safely during the pandemic and that this has meant you've only seen two movies in aaaaaall of 2020 (Rise of Skywalker and Tenet....hypothetically) and you miss going to the cinema and eating popcorn VERY much.

Also suppose that you are a statistician and love playing with data.

Obviously, the next step is to get some data about movies!

But how….

OPTION 1: WEB SCRAPING

Web scraping (also known as web harvesting, web crawling or web data extraction) is any method of copying data from a webpage, usually to then store it in a spreadsheet or database.

Downloading a ready-made .csv file hosted by a site *wouldn't* be considered web scraping. (Although you might find a programmatic way to download many of these could be.)

MORE THAN ONE WAY TO SCRAPE A SITE

Copying and pasting	Chrome extensions and other user-friendly tools	Code it yourself
Pros: • Familiarity with data • Less to worry about when reading T&Cs	**Pros:** • Faster • Works for large datasets • Can set scrape times	**Pros:** • Control • Customizability • Smooth workflow
Cons: • Lots of time and manual labour • Easy to make mistakes	**Cons:** • Less customisable • Choppy workflow • Data must be organised	**Cons:** • Need to learn to code • Data must be organised

IF YOU'RE CODING IT YOURSELF, YOU'LL NEED:

- Some knowledge of URLs, HTML and CSS
  URL - Universal Resource Locator
  HTML - HyperText Markup Language
  CSS - Cascading Style Sheets

- The rvest package in R (or Rcrawler, there may be others too) or the Beautiful Soup package in Python

- Professional ethics!

Just because you CAN do something, should you?

THE ETHICAL SCRAPER

I, the web scraper will live by the following principles:

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need it OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the **purpose of creating new value from the data**, not to duplicate it.

Source: James Densmore, https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

T&CS AND ROBOTS.TXT

Many sites give instructions about what you're allowed and not allowed to do on them. One way is through the Terms and Conditions and another is through a file called robots.txt.

T&Cs

Ideally, we should all be reading all the Terms and Conditions of all the websites we use…and of course I'm sure you dooooo.

But when in a hurry, search (CTRL+F or CMD+F) "scrape", "harvest" "crawl" and if none of those come up then "data" and "copied" more generally and that can give you a sense if they prohibit certain uses.

Robots.txt

Most large websites have use a robots.txt page to give instructions about what 'robots' are and aren't allowed to visit on the page. This is most often used for search engines, but we can check them too. Bad bots can still do what they want.

http://www.robotstxt.org/robotstxt.html

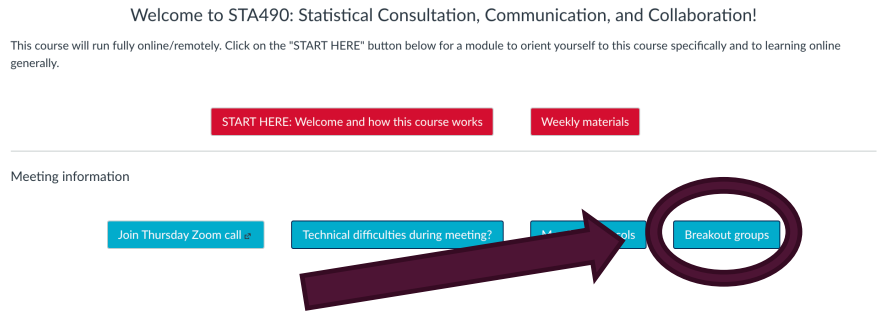
AN EXAMPLE OF A ROBOTS.TXT

Note: Websites are set up much like the files on your computer might be. The slashes indicate a subfolder, and you can nest them.

BREAKOUT ROOM DISCUSSION: T&CS + ROBOTS.TXT

BREAKOUT GROUP SYSTEM

- Check the Fall breakout group assignments role designation key (and which group you're supposed to be in). This week: 🧣**s are facilitators,** ☕**s are notetakers,** 🎃**s are reporters and** 🍂**s are timekeepers**. If your group is missing a member, combine the timekeeper and reporter roles and adjust. The notetaker should make note of who is filling each role in the grey area of the document.



One-word check-in (attendance) → **Breakout: Littering case study** → Class discussion: Littering case study → Asynchronous writing activity → Reminders/next steps

BREAKOUT ROOM DISCUSSION ROLES

Facilitator

Responsible for seeking out opinions from all group members and ensuring that everyone has the opportunity to contribute.

Notetaker

Responsible for recording the key points of your group's discussion in the collaborative document (I'll share the link in Zoom and it is also on the Weekly resources page).

Reporter

Responsible for reporting back to the class on behalf of the group. Seek consensus from the group about your most important aspects of your discussion to share.

Timekeeper

Responsible for managing time for the group, making sure you stay on track and spend equal time on each question. If there are only three members of the group, the time keeping should be done by the person in the **Reporter** role.

One-word check-in (attendance) → 1-1 EDA meeting (Tuesday) → **Breakout: Consulting case study (part 1)** → Class discussion: Consulting case study (part 1) → 10-minute break → Breakout: Consulting case study (part 2) → Class discussion: Consulting case study (part 2) → EDA troubleshooting → Reminders/next steps

BREAKOUT ROOM DISCUSSION 1: CONSULTING CASE STUDY PART 1 (5 MINUTES)

Check out the following sites:

1. Their Terms and Conditions (if it exists)
2. Their robots.txt page (if it exists)

Can determine you what their scraping rules are. Is there a specified crawl delay? Are there a lot of restrictions? Or very few?

Write down anything you notice that is interesting.

- www.kijiji.ca/ (www.kijiji.ca/robots.txt)
- www.utoronto.ca (www.utoronto.ca/robots.txt)
- One additional site of your choice

The notetaker should open the collaboration document from the weekly materials page (I will also try to share in Zoom).

Record your group members roles in the grey area, discuss the questions and keep notes in the white area.

💄 **Facilitator:** Make sure every has a turn to speak on each question

☕ **Notetaker:** Record team member roles and take notes.

🎃 **Reporter:** Prepare to feedback to the class.

🍂 **Timekeeper:** Make note of the total breakout time and the number of questions. Divide your time so you can spend about the same time on each question.

One-word check-in (attendance) → 1-1 EDA meeting (Tuesday) → **Breakout: Consulting case study (part 1)** → Class discussion: Consulting case study (part 1) → 10-minute break → Breakout: Consulting case study (part 2) → Class discussion: Consulting case study (part 2) → EDA troubleshooting → Reminders/next steps

QUICK REPORT BACK

THE ETHICAL SCRAPER

- Follows the site's terms and conditions and/or robots.txt
  - Uses an API when provided
- Rate limits their requests
- Credits sources

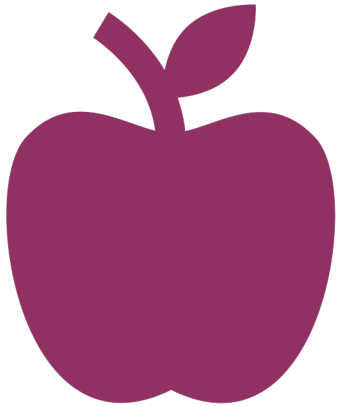
Packages that help with this:

- Polite package
  - The goal of `polite` is to promote responsible web etiquette.
- Rvest package
  - `Rvest` is a package that enables the webscraping


A NETFLIX FILM
The Princess
SWITCH
NOVEMBER 16 | NETFLIX


N
The Princess Switch
SWITCHED
AGAIN


To the Jupyter Hub!



5-minute drink/snack stretch break

APIS, WEB SCARPING AND THE ETHICS OF GETTING DATA (PART II)

OPTION 11: USING AN API

API stands for **a**pplication **p**rogramming **i**nterface.

It is a structured way for data (broadly) requests to be made and fulfilled with computers.

I like this comparison to a restaurant menu. You don't need to know HOW to make crème brûlée to be able to know you WANT it. ☺


To the Jupyter Hub!

REMINDERS/NEXT STEPS

MID-YEAR OBJECTIVE REFLECTION DUE DEC 9 AT 11:59AM

Prompt

In the syllabus for this course, you will find the eight learning objectives for STA490. Choose TWO of these learning objectives that are particularly relevant/important to you and write a reflection that covers the following topics:

- Explain each learning objective in your own words (i.e. paraphrase what is written in the syllabus).
- Explain WHY each objective is important/relevant to you. You may wish to make reference to your Focus in your Specialist program, your career goals or goals for future education.
- Evaluate your current progress towards this goal. What can you already do? What are your next steps for improvement?
- Discuss how you will personally know you have succeeded in this objective. I.e. don't say "an A in STA490", think about how you will personally evaluate your success, independent of grades. What *will* you be able to do?
- Make sure you **introduce** your topic/context at the beginning (don't just dive straight in, prepare your reader!) and close with a summary/concluding statements (don't just end abruptly).

Other details

- **File type:** pdf
- **Word count:** 500 words ± 75 words (i.e about one full page, 12 pt font, single-spaced)
- **Audience**: Imagine a future employer or grad school supervisor is your reader. Assume they have no idea what STA490 is and that while they have some general statistical knowledge, you should not be throwing lots of technical details or R functions at them.

[L0101 Link] [L0201 link]

UPCOMING

- TOMORROW (starting at 10:10 a.m.): Sign up for your 'Explain like I'm a…' presentation group
- NEXT Class: Presentations from Library 1A, Library 2A, Turtles 1B, Turtles 2A, Fish 1B, Fish 2A, Masks 1B, Masks 2A (also new Zoom links)
- Dec 9: Mid-year objective reflection
- **Project**
  - Continue working on your analyses
  - Pod-based meetings next week. You will need to submit pre- and post- project logs.

SAMUEL BEATTY IN-COURSE SCHOLARSHIP [DEADLINE DEC 4]

- **Who is eligible?**
  - Any student in the **second, third** or **fourth** year and taking a **Specialist** Program offered by the Departments of **Computer Science**, **Mathematics**, **Physics** or **Statistics.**
- **What are the criteria of selection?**
  - Awards will be given on the basis of academic performance during the previous year and on the basis of need.
- **Who will select the winners?**
  - A committee composed of the Undergraduate Associate Chairs (or their representatives) of the Departments of Mathematics, Physics, Statistics and Computer Science.
  - Apply here by Dec 4: https://www.physics.utoronto.ca/undergraduate/samuel-beatty-in-course-scholarship

RA JOB WITH PROFS MOON, BOLTON & CO! [DEADLINE DEC 4]

Research Assistants (2) – ISSC (Undergraduate)

To apply for this position, please complete this form by 5:00 p.m. ET, Friday, December 4, 2020.

Only applicants selected for an interview will be contacted.

These positions can also be found on the Career & Co-Curricular Learning Network

- https://clnx.utoronto.ca/staff-faculty/oncampus.htm

WE'LL BE AROUND AFTER CLASS FOR ANY QUESTIONS AND TO CHAT FURTHER

IMAGE CREDITS

- Web: https://phys.org/news/2018-10-illuminating-dark-web.html
- Paint scraper: https://www.totalcarepainting.com/paint-scraping-gone-easy-tools/
- Eye art: https://unsplash.com/photos/xb0wLfZH9Zo
- Princess Switched poster: https://en.wikipedia.org/wiki/The_Princess_Switch
- Princess Switched Again: https://popcornandtequila.com/the-princess-switch-2-trailer/
- Jupiter: https://en.wikipedia.org/wiki/Jupiter
- Crème brûlée: https://www.simplyrecipes.com/recipes/how_to_make_creme_brulee/