Day 1

Brain Storm w/ class what Data Science Is?

- Data Gathering
- Data Storage
- Data Analysis
- Data Presentation

Syllabus & Schedule overview

What is HTML?

- Makes webpages along with css

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>
<p>This is a paragraph.</p>
<a href="https://www.google.com/"> Google</a>


</body>
</html>
```

Save in a .html file. Now when you open it you have a very basic webpage! Hosted on your computer, no one else can see it but you could, if you wanted to host it online and project it out into the world.

Let's look at another HTML page. If you go to a browser and a website, two finger/left click and select Inspect Element you can see the HTML and then what is being coded for on the page. What are some other common tags that you see on this website?

<!--...-->	Defines a comment
<!DOCTYPE>	Defines the document type
<a>	Defines a hyperlink
<b>	Defines bold text

<big>	Not supported in HTML5. Use CSS instead. Defines big text
<body>	Defines the document's body
 	Defines a single line break
<dir>	Not supported in HTML5. Use instead. Defines a directory list
	Defines a section in a document
<font>	Not supported in HTML5. Use CSS instead. Defines font, color, and size for text
<h1> to <h6>	Defines HTML headings
<head>	Contains metadata/information for the document
<img>	Defines an image
<p>	Defines a paragraph
<q>	Defines a short quotation
<script>	Defines a client-side script
<style>	Defines style information for a document
<table>	Defines a table
<title>	Defines a title for the document
<video>	Defines embedded video content

There are many thousands of tags more you can learn about html. If you are interested in those at this point I encourage you to go to w3school.

BY NEXT CLASS DOWNLOAD Anaconda so you can use Jupyter Notebooks to run your python.

Day 2

To read in a file:

```
f = open("ex.txt")
print(f.read())
```

Or

```
with open("ex.txt") as file:
        lines = [line.rstrip() for line in file]
        print (lines)
```

Make an example txt file and play around with how to print stuff out.

Work on Project.

Day 3

Review Questions from EL

Watch this video on the basics of Web Scraping and try to answer the following questions:

- What is web scraping?
    - Extracting Data from websites, typically in an automated process.
- Is there only one way to web scrape?
    - **Human copy-and-paste**
    - **Database**
    - **HTML parsing**
    - **Vertical aggregation -** News Aggrigator, RSS Reader
- How might web scraping be prevented?
    - Firewall
    - CSS Sprite - Accessibility concerns
    - Adding characters to HTML
    - CAPTCHA
    - Traffic limitations

Read this article on the legal system and Web Scraping while keeping in mind the following questions:

- In the United States, website owners can use four major legal claims to prevent undesired web scraping, what are they?
    - (1) copyright infringement - while outright duplication of original expression will in many cases be illegal, in the United States the courts ruled in *Feist Publications v. Rural Telephone Service* that duplication of facts is allowable.
    - (2) violation of the Computer Fraud and Abuse Act ("CFAA") - “prohibits intentionally accessing a computer without authorization or in excess of authorization.”
    - (3) trespass to chattel - Lack of consent, actual harm, intent. intentional interference with another’s personal property. In order for it to be enforceable, there must be proof of damage.
    - Terms of service - While violating the terms of service is not a crime, it may be a breach of contract and result in the removal of any content used from the site from your project.
- Was the original purpose of these legal structures to prevent Web Scraping? How does the original intent effect how the system works now?
    - Terms of Service:
        - When the first wave of digital goods in the 1970’s, businesses devised a legal hack to prevent users from copying their products, End-User

License Agreements (EULAs), b/c unlike other kinds of property, software could be copied instantly with almost no effort.
        - Decades have passed since the first software licenses were stuck onto floppy disks, but the actual law remains largely the same.
        - TOS are essentially very one-sided contracts written by the company selling the digital goods.
    - (1) copyright infringement - 1700s originally for taking rights away from publishers and giving rights to authors. All seemed very simple till radio and then the internet when we are no longer making physical copies of everything.
    - (2) violation of the Computer Fraud and Abuse Act ("CFAA") - has widened since original law
    - (3) trespass to chattel - origins in the 1200s but became more clear in the 1500s with Queen Elizabeth I. Dealt with if you lost something and someone else found it or if a third party (the bailee) was taking care of your stuff and it got damaged while it was with them - usually this was livestock. The idea in the original law was that the other person had temporarily taken ownership of the stuff and must compensate the owner for it, and if it was damaged the owner wasn't obligated to take it back instead there would be a forced sale.

Early scraping decisions are not uniform

Pattern = courts protect proprietary content on commercial sites from uses which owners of sites don't want.

The degree of protection for such content is not settled and will depend on:

- type of access
- amount of information
- degree to which the access adversely affects the site owner's system
- types and manner of prohibitions on such conduct.

Internet Archive collects and distributes a significant number of publicly available web pages without being considered to be in violation of copyright laws.

Difference between law and morality

In groups: Think of examples when web scraping is good.

Discuss as class

In groups: Think of examples when web scraping is bad.

Discuss as class

Examples - weather data, pad mapper v. craigs list, research scientists,

In groups: Think of examples who are the actors and interested parties in both of the previous discussions

Discuss as class

In groups: What laws would you write if you could to govern web scraping

Discuss as class

Day 4

For each of the following people list what you are and are not comfortable with them knowing about you:

Best Friend since forever
Friend you made last month but really like
Your date tomorrow night
Your professor (you don't need specifics just categories)
Your classmate who you don't hang out with
The President of the school
Your state senator
The FBI
The Police
The grocery store
A casino

Source: https://www.coursera.org/learn/data-science-ethics/lecture/hCwR8/data-ownership

If Data is about you do you own in?

- Generally no.
- **Biography** if I wrote the book then I own the rights to it even though it is about you.
- Limitations: Liable - incorrect information that is demeaning and could harm somebody's career or reputation, then they might sue
- **Photos**, same deal
- Limitations: Liable, can't take photos in certain places, can't pretend you are promoting something you aren't
- **Electronics**, same deal
- If I record something about you then my records are my records
- **Translations**
- Limitations: You have to get permission to write the translation but after that the translated work is yours.
- **Copyright and Patents**
- Credit or compensate owner by law if you use part of their work.
- We make you do this through citations in the work you turn in for class. Why?
    - They worked hard to create the original work and we want to acknowledge this
    - It absolve you of some responsibility if there are mistakes or falsehoods in the data because it was there responsibility to find those.

- If data is used technically people should use the same crediting system
- But if it is just one piece of data that is then aggregate with many other data points it is less meaningful to do that. Giving credit can be hard.

- **Library** - collects lots of books and curate a collection. Curation takes a lot of time and effort, they compensate the publishers of the books by buying the books and at the end of the day the collection belongs to the library to do with as they will. - for example put a book about climate change in the fiction section
- **Wikipedia** You don't need to be a creator to claim ownership. Wikipedia and Rotten Tomatoes both own the data on their servers even though it was put their through crowd sourcing. The majority of the work was done by people who were not compensated and while Wikipedia chooses to provide their material for free they don't have to because they own it. Rotten Tomatoes is free to put what ever ads they like on their site and keep the profits without distributing them to the users who rated the movies.

Limits on Recording Data

Expectation of Privacy.

Are there any special places where you would expect more privacy? In the physical world? Online?

- Bathroom, changing room
- your phone company shouldn't listen in on your calls (Apple is pushing this limit)
- There are **cameras in stores** is it ok that they
    - Record for security purposes? - usually the answer is yes in our social contract
    - Analyze customer movements for product placement?
    - Post videos of people walking around their store on the internet? - usually no
- **Phone** companies have GPS on their phones is it ok for them to
    - Use the GPS to find where you are to give you cell service? - yes or what is the point
    - Keep and track your location 24/7? - this would represent a huge loss of privacy, still coming to a societal understanding here
- Limit the use of your data not the recording of your data (there is reason for the store to record for security purposes but if they use that data in other ways it would not be good)
- **Police Body Cameras** - we want this data to make sure police officers act ethically and to have a record when they don't but we would be mad if they posted the videos to the web every day because the people they are interacting with might not be at their best and should not be unfairly judged.
- **Government Surveillance** - Suck in a lot of data without looking at it because there is too much to look at and they don't know what they will need. Then when there is a need, they (hopefully) get a warrant to look at the data they have collected.

What are possible consequences of just sucking up lots of data?

Hitler's Willing Business Partners
"The Nazis ordered censuses in both countries [Holland and France] soon after they were occupied. In Holland, a country with "a well-entrenched Hollerith infrastructure," out of "an

estimated 140,000 Dutch Jews, more than 107,000 were deported, and of those 102,000 were murdered—a death ratio of approximately 73 percent." In France, where the "punch card infrastructure was in complete disarray," of the estimated 300,000 to 350,000 Jews in both German-occupied and Vichy zones, 85,000 were deported, of whom around 3,000 survived. "The death ratio for France was approximately 25 percent."

“Black [author of book] gives evidence to qualify the implied claim that the Hollerith technology made the decisive difference. In Holland the Nazis installed a zealous bureaucrat to take the census. France had a moral hero in charge who frustrated German efforts to find Jews—and paid with his life. Holland had a long and innocent tradition of recording religion on all manner of official documents. France "lacked a tradition of census taking that identified religion."

Discussion

It is generally considered inappropriate to record a conversation without permission. When police informants wear a wire, there is usually pre-authorization by a judge. Twelve states in the US require all parties to consent to recording before a conversation can be recorded. The remaining states require at least one party to consent. So, in the US, it is never OK for a third party, such as law enforcement, to record a conversation without consent, or warrant. However, it is generally considered appropriate to record metadata about the conversation: who participated, at what time it took place, how long it lasted, and so on.

Now consider a recent case involving Amazon's Echo: https://techcrunch.com/2016/12/27/an-amazon-echo-may-be-the-key-to-solving-a-murder-case/.

There are multiple parties involved: (1) law enforcement (2) Amazon (3) the owner of the Echo (4) Various people who may have been recorded by the Echo (including the owner).

There are two different types of data involved: (1) The recorded conversations and other sounds, and (2) Metadata about the time, duration, etc. of these conversations.

Who owns what data? and who may have access to what data WITHOUT a judicial warrant?

(Let us assume that if a crime has been committed, such as the murder in this example, that a judge will sign a warrant that can override normal privacy and ownership expectations. So this question is about what you could/should do if no warrant can be obtained).

Extend your analysis of the preceding paragraph to data regarding your interaction with your electronic merchant. There are two parties to this interaction: you and the merchant. Who may record the data, who may publish/share the data? Make sure you also consider scenarios where you, as the customer, may wish to publish a review of your (unsatisfactory) interaction with an online merchant.

- Is just sucking up data Ok? Is it different from warrants wire tap someone's phone? Or a warrant to search someone's home? Would your answer change if you were in a different country?

Data | Philosophy Tube

Day 5

Data Visualization

Day 6 - Work Day

Day 7

Data preparation: loading, cleaning, transforming, and rearranging. Such tasks are often reported to take up 80% or more of an analyst's time.

Data Cleaning Activity

What data irregularities are you encountering? How is each messing up your data? What could you do about each?

- Missing Data
- Irrelevant Data - people who don't live on campus or who didn't fill out the survey right - outliers - be careful when removing these you have to have a good reason
- Unnecessary Data — Duplicates
- Inconsistent Data — Capitalization