Lecture 4: Let's get data!

Prof. Esther Duflo

Where can we find data?

1. Existing data Libraries
2. Collecting your own data
3. Extracting data from the internet

Existing data libraries

- A Great resource for MIT students and others: http://libguides.mit.edu/ssds
- Popular sources of data
  - Data.gov: Datasets generated by the executive branch of the US government http://www.data.gov/
  - IPUMS : censuses from the US and many more! https://www.ipums.org/
  - International IPUMS https://international.ipums.org/international/
  - ICPSR http://www.icpsr.umich.edu/icpsrweb/ICPSR/: A data repository with many data sets on lots of subjects
  - Harvard-MIT Data center and Harvard Data verse https://dataverse.harvard.edu/ where many researchers archive their data
  - Amazon dataverse http://aws.amazon.com/public-data-sets/

International household survey data

- Demographic and Health surveys
  http://www.dhsprogram.com/
- World bank http://data.worldbank.org/
- LSMS (search for LSMS on the world bank data page)
- Rand public-use databases
  http://www.rand.org/labor/data.html

Replication data from researchers

- Randomized control trials
  - https://dataverse.harvard.edu/dataverse/jpal
  - https://dataverse.harvard.edu/dataverse/socialsciencercts
- The American Economic Association journals require posting of any data used for research: there is lots of data on the AEA website

The internet!

- Many websites that are data intensive are making that data directly available to people
  - 538
  - Yahoo data dump : "a sample of anonymized user interactions on the news feeds of several Yahoo properties". http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75
  - Uber movement (to come) https://movement.uber.com/cities
- Some sites are specializing in aggregating data
  - Sports data sets http://www.opensourcesports.com/, http://nbasavant.com/
  - Web pages: Way back machine https://archive.org/web/
- There is much more... search in library catalog, google, etc.

But what if this is not what I am looking for?

- Sometimes what you are looking for is available, but not free: your library may be able to purchase it… or you may need to get an agreement.
- Sometimes it is available but the access is restricted (for example for confidentiality reasons).
  - For administrative data, see https://www.povertyactionlab.org/admindata
  - The entity that owns the data may be interested in sharing it with you if this is part of a research project (prospective or retrospective)
  - You will then have to comply with the partner's requirements for data security, and also go through an Human Research Board Review (IRB) at your institution.
- And sometimes you will have to harvest it yourself!

Harvesting data

- Scraping data from the internet.
- Collecting your own data.

Scraping data from the internet

- Use an API
- Use a web page

What is Web Scraping?

- Pull data form one page
- Crawl an entire web page
- A set of forms running in the background
- Any of the above in an ongoing fashion

Using an API

- API (Application Programme interface) are programs that help a particular program to communicate with other programs
- Some web sites provide and invest in API (Twitter, Facebook, google map, etc.) and you will typically use those to harvest the data from those sites, some time in conjunction with python

Example: using google map API to look at traffic in Delhi

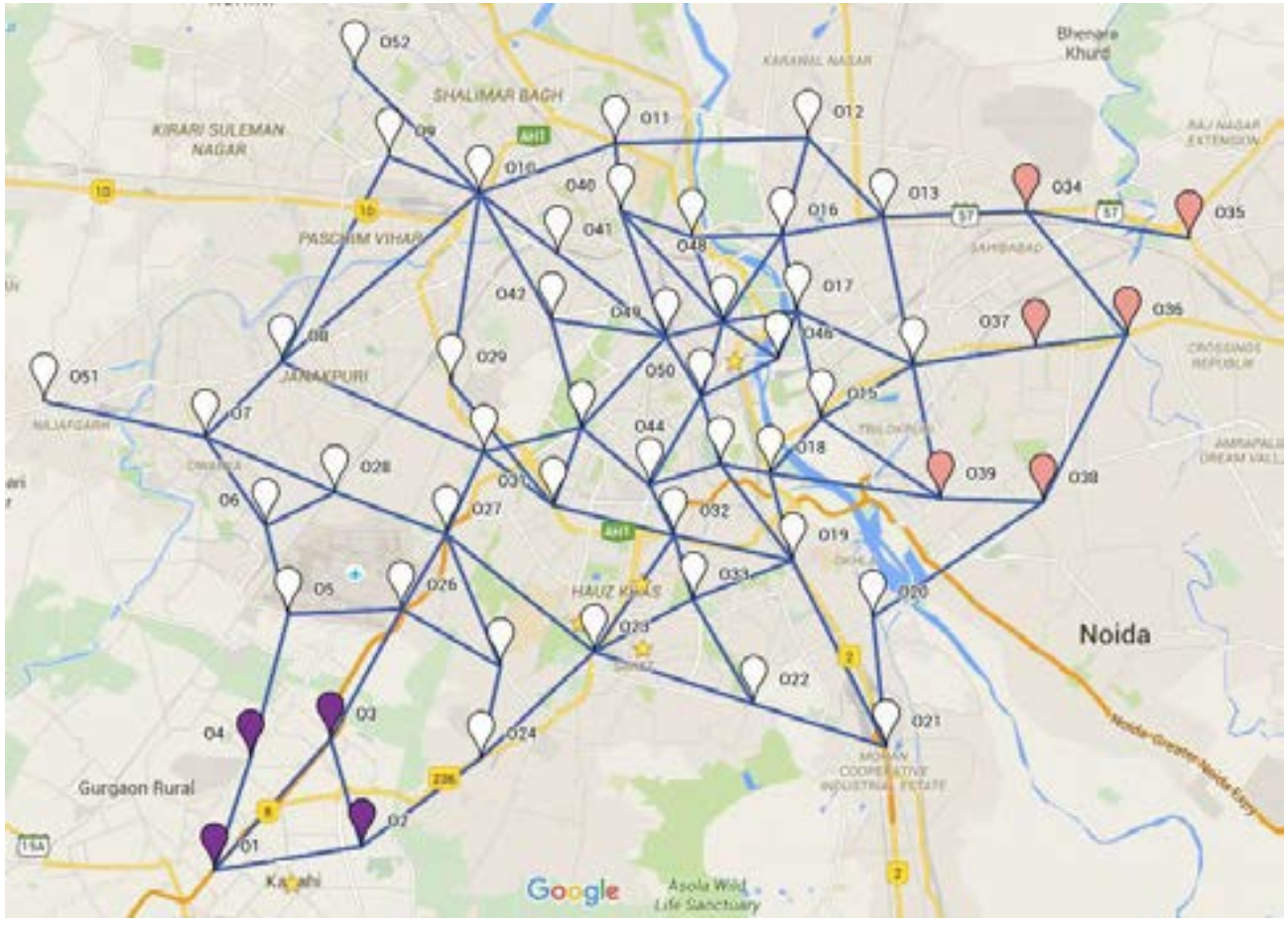
Gabriel Kreindler

- Google has set up an API (application program interface) that allows access to Google Maps.
  - Mostly used by smartphone apps, websites, etc.
  - Google provides documentation and libraries to access.
  - Very simple to access using Python, Java, HTML, etc.
  - No need for complicated scraping!
  - Designed for commercial applications.Cost structure:
    - First 2,500 queries per day free, then 1 USD/2,000 queries.
    - Researchers can also take advantage!

- One example: travel time(distance matrix)
- Takes into account traffic conditions at different hours.
- Two types of queries, depending on departure time:
  - Departure time = now ?prediction based on crowdsourced live data from Google Android users, as well as historical data on that route.
  - Departure time = 6/15/2016 8:34am ?prediction based on historical data.
- No direct access to historical data.

Example: Delhi Odd-Even study

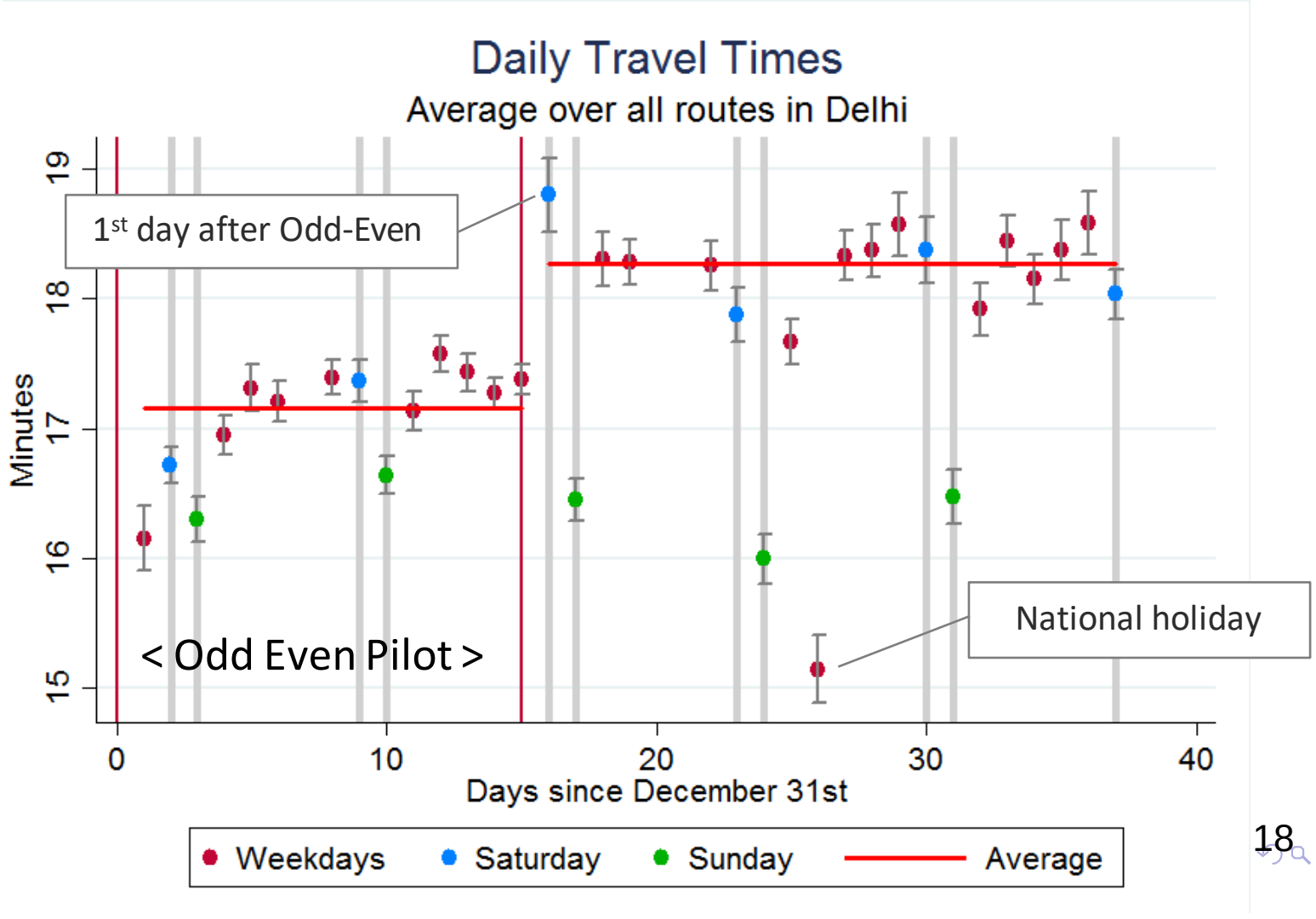
- Delhi, one of the most polluted cities in the world, experimented with a driving restriction policy called Odd-Even (based on license plate numbers)
- Implemented between January 1-15th 2016, and again in April
- The following results are based on queries made every 20 minutes since January 1st2016, on 93 routes across Delhi.
- The input is a API key and .csv file that has a departure point (latitude, longitude) and arrival point
- The out put is many .csv files with a time it takes.
- A python code keeps querying google API
- Code+readme available on the course web site for those interested



```
### General parameters
# Path to input and output files
osdir = 'YOUR PATH HERE'

### Query parameters
delta_time = 30#20*60 # time interval between queries in seconds # 20 minutes
fname = 'input/routes.csv' # input file
oname = 'output/live_travel_times' # output file

### Google API key - create Google Maps API client
mykey='YOUR GOOGLE API KEY HERE'
client = googlemaps.Client(key=mykey)
```

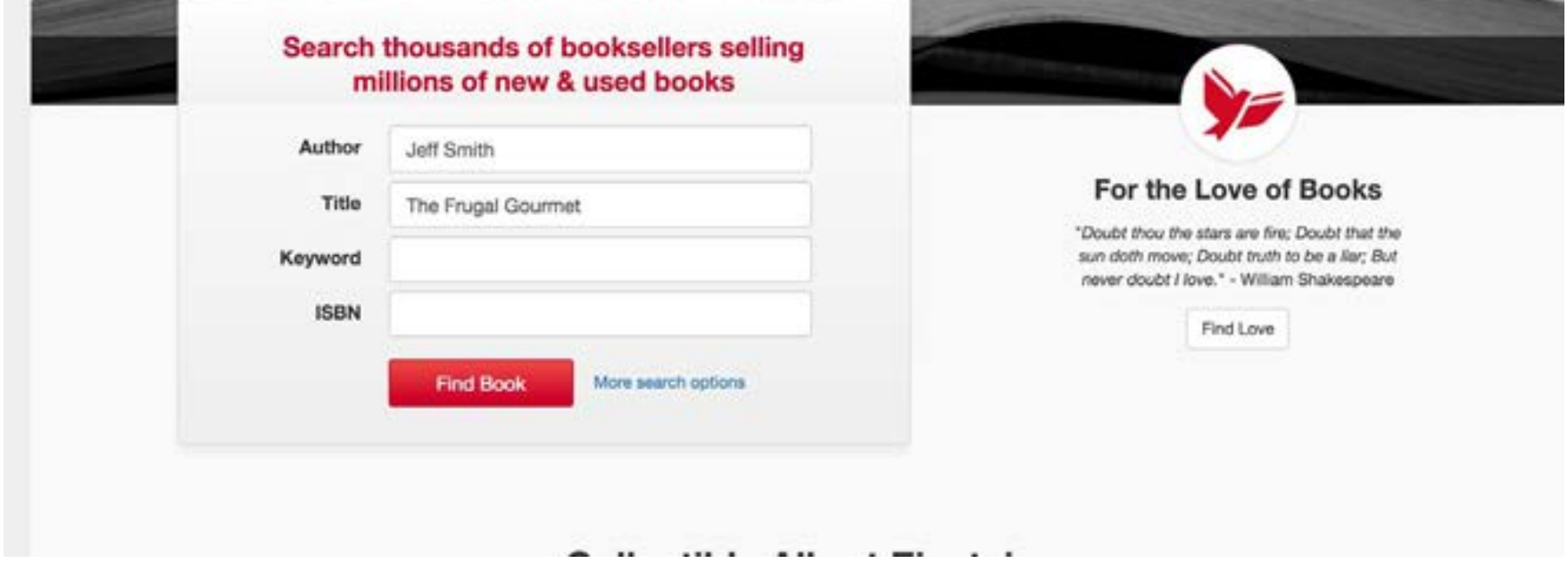
```
            current_dict["dep_time_local"] = time.strftime('%d %b %Y %H:%M:%S', now_loc)

            # request travel time for driving with "best guess" traffic estimate
            local_result = client.distance_matrix((olat, olon), (dlat, dlon), departure_time="now", mode="driving")

            # extract results
            temp = local_result["rows"][0]["elements"][0]

            if temp['status']=='OK':
                current_dict["distance_bg"] = temp["distance"]["value"]
                current_dict["duration_bg"] = temp["duration"]["value"]
                current_dict["duration_in_traffic_bg"] = temp["duration_in_traffic"]["value"]

            ## add local dictionary to the list of results
            results.append(current_dict)

        # write for this time:
        sufix = str(now_loc.tm_mon) + "_" + str(now_loc.tm_mday) + "_" + str(now_loc.tm_hour) + "_" + str(now_loc.tm_min) + "_"
        write_to_csv(results, oname + sufix + ".csv")
except:
    pass
```

Ex1 – Delhi Odd-Even [Live queries]

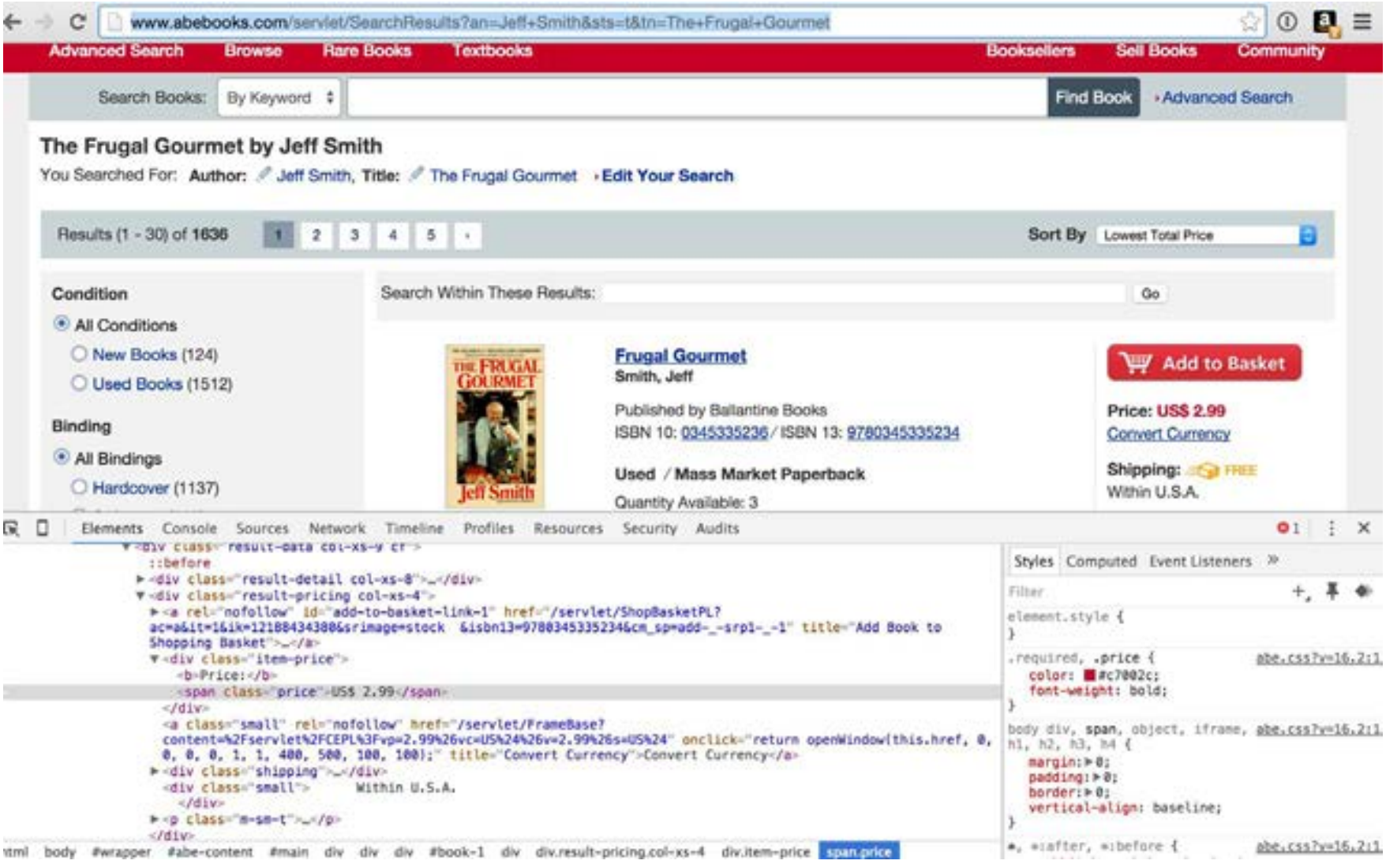


Scraping web sites

- Some providers will not have an API
- Then you need to extract the information from the page.
- Example: Ellison and Ellison: Did the internet change the price of used books?
  - Want to compare the price of the same used book in stores and online
  - Need to collect data at regular interval on the prices of a bunch of used books.
  - From the web site http://www.abebooks.com/



What we start with



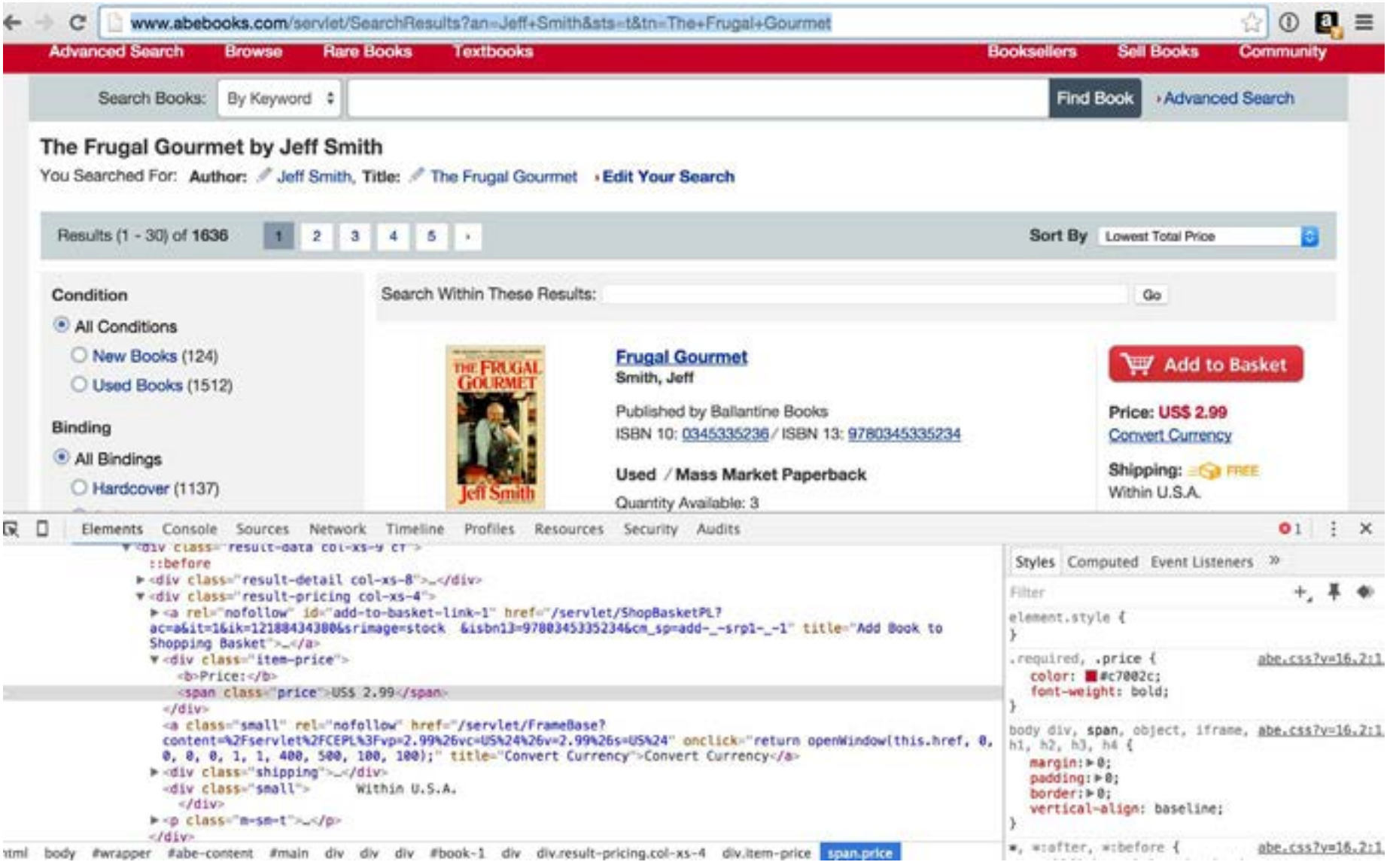
What we want to get

A nice table we can import in R

- Name of title (for lots of titles)
- Date
- Price

Web scraping with Python

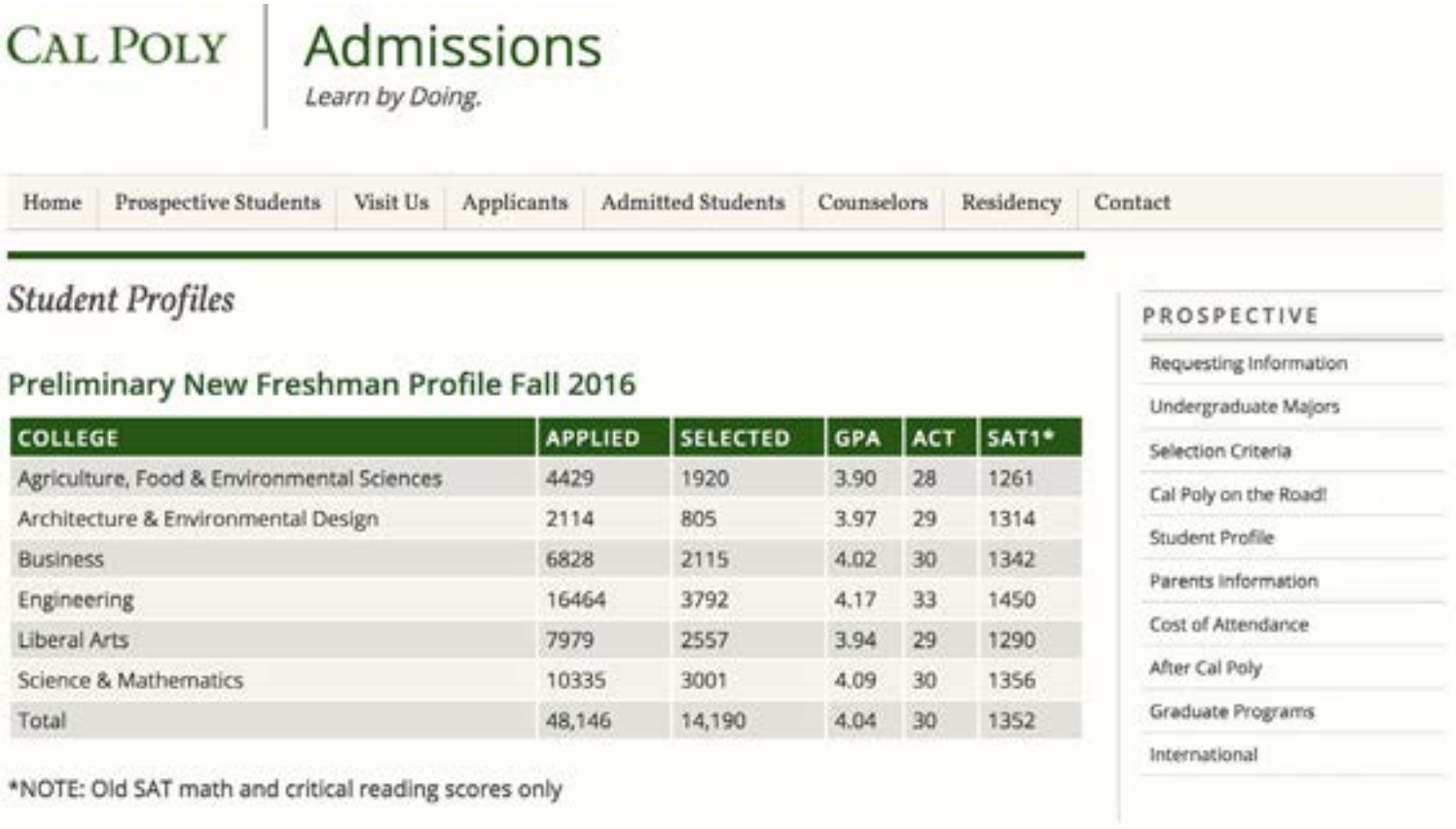
- Most conventional way to do it... the internet is full of tutorials
- You will work using the request library and the BeautifulSoup library
- With those you will write simple routine that will extract what you are looking for.
- In the used book example, we need to pull up the page for each book at specified date, and instruct python to search for the price (which is nicely identified to as a class).
- And export them into a table.



Web scraping in R

- R has a web scrapping package built by Hadley Wickham (same person who wrote the R for data science book, ggplot2, tidyverse,), called rvest.
- See: http://blog.rstudio.org/2014/11/24/rvest-easy-web-scraping-with-r/
- Works well in conjunction with a google chrome plugin selectorgadget.com/
- It has the ability to submit forms and search web pages, etc.
- See demo(package = ”rvest”) for demonstrations

Harvesting a table



CAL POLY | Admissions

Learn by Doing.

Home | Prospective Students | Visit Us | Applicants | Admitted Students | Counselors | Residency | Contact

Student Profiles

Preliminary New Freshman Profile Fall 2016

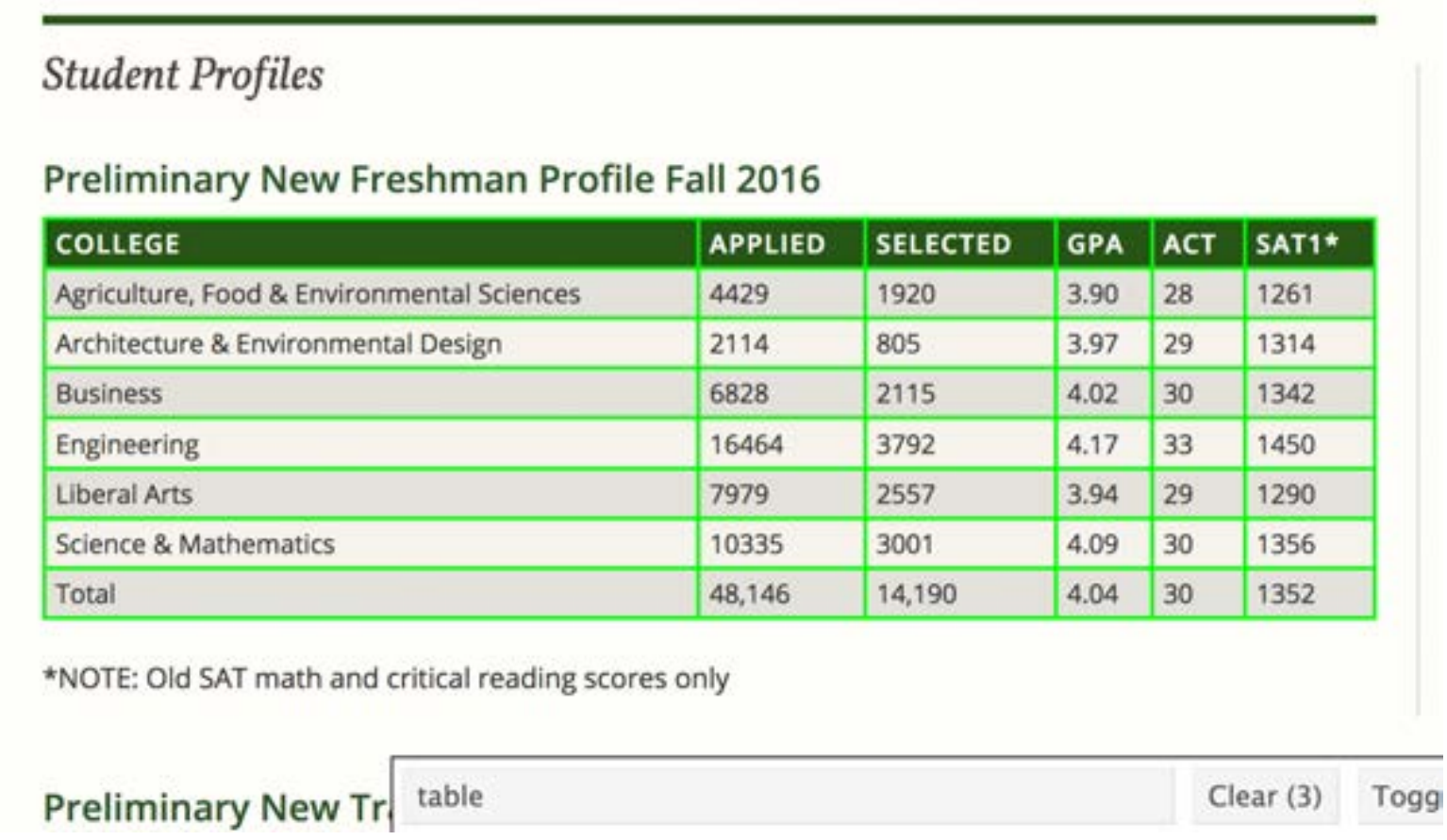
COLLEGE	APPLIED	SELECTED	GPA	ACT	SAT1*
Agriculture, Food & Environmental Sciences	4429	1920	3.90	28	1261
Architecture & Environmental Design	2114	805	3.97	29	1314
Business	6828	2115	4.02	30	1342
Engineering	16464	3792	4.17	33	1450
Liberal Arts	7979	2557	3.94	29	1290
Science & Mathematics	10335	3001	4.09	30	1356
Total	48,146	14,190	4.04	30	1352

*NOTE: Old SAT math and critical reading scores only

PROSPECTIVE

- Requesting Information
- Undergraduate Majors
- Selection Criteria
- Cal Poly on the Road!
- Student Profile
- Parents Information
- Cost of Attendance
- After Cal Poly
- Graduate Programs
- International

Harvesting a table



Student Profiles

Preliminary New Freshman Profile Fall 2016

COLLEGE	APPLIED	SELECTED	GPA	ACT	SAT1*
Agriculture, Food & Environmental Sciences	4429	1920	3.90	28	1261
Architecture & Environmental Design	2114	805	3.97	29	1314
Business	6828	2115	4.02	30	1342
Engineering	16464	3792	4.17	33	1450
Liberal Arts	7979	2557	3.94	29	1290
Science & Mathematics	10335	3001	4.09	30	1356
Total	48,146	14,190	4.04	30	1352

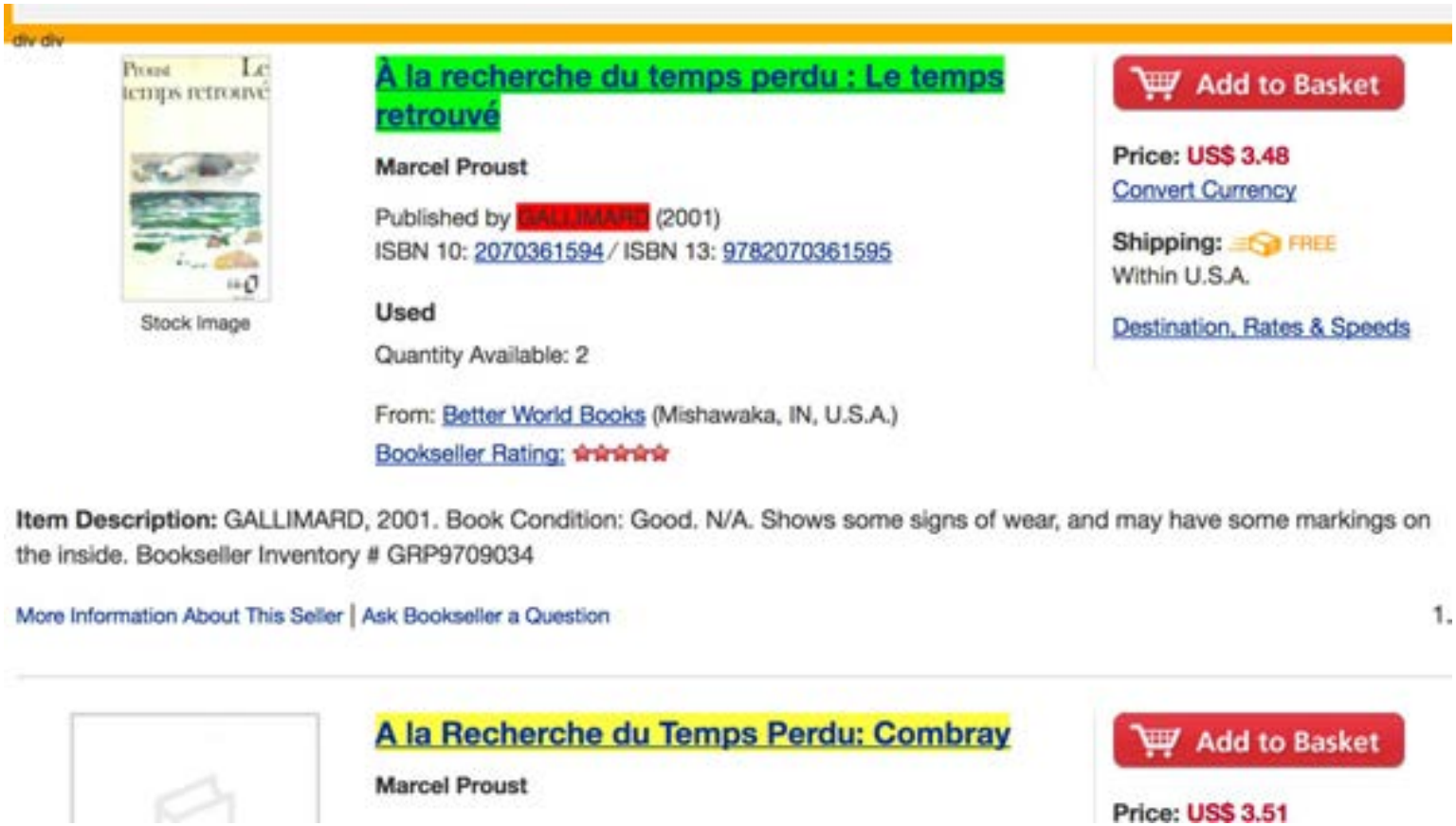
*NOTE: Old SAT math and critical reading scores only

Preliminary New Tr table Clear (3) Togg

Harvesting a table

```
1  library(rvest)
2  CPadmission <- read_html("https://admissions.calpoly.edu/prospective/profile.html")
3  CPadmission %>%
4   html_nodes("table") %>%
5  .[[1]]%>%
6  html_table()
7
8  admission_1<-html_table(CPadmission)
9
10 CPadmission %>%
11   html_nodes("table") %>%
12   .[[2]]%>%
13   html_table()
14
15 admission_2<-html_table(CPadmission)
```

Harvesting specific items



Harvesting specific items

```
1  library(rvest)
2  larecherche <- read_html("https://www.abebooks.com/servlet/SearchRe
3  titlehtml <- html_nodes(larecherche, ".col-xs-8 a span")
4  titletext <-html_text(titlehtml)
5  pricehtml<-html_nodes(larecherche, ".item-price .price")
6  pricetext<-html_text(pricehtml)
```

Harvesting specific items

```
> pricetext
 [1] "US$ 3.48" "US$ 3.51" "US$ 3.51" "US$ 3.81" "US$ 3.81"
 [6] "US$ 3.91" "US$ 3.91" "US$ 4.35" "US$ 4.62" "US$ 1.00"
[11] "US$ 1.00" "US$ 1.00" "US$ 2.15" "US$ 3.05" "US$ 3.13"
[16] "US$ 6.86" "US$ 3.66" "US$ 3.62" "US$ 4.00" "US$ 3.50"
[21] "US$ 3.28" "US$ 3.28" "US$ 8.18" "US$ 4.36" "US$ 5.46"
[26] "US$ 4.99" "US$ 5.00" "US$ 5.00" "US$ 5.36" "US$ 4.37"
> titletext
 [1] "À la recherche du temps perdu : Le temps retrouvé"
 [2] "A la Recherche du Temps Perdu: Combray"
 [3] "A la Recherche du Temps Perdu: Combray"
 [4] "Ã\u0080 la recherche du temps perdu"
 [5] "Ã\u0080 la recherche du temps perdu"
 [6] "Ã\u0083? la recherche du temps perdu : Le temps retrouvÃ\u0083Â"
 [7] "Ã\u0083? la recherche du temps perdu : Le temps retrouvÃ\u0083Â"
 [8] "A la recherche du temps perdu, A l'ombre des jeunes filles en fleurs, volume 2"
 [9] "A la recherche du temps perdu, Sodome et Gomorrhe, volume 2"
[10] "A la recherche du temps perdu III Le cote de Guermantes"
[11] "A la recherche du temps perdu III Le cote de Guermantes"
[12] "Combray A La Recherche Du Temps Perdu"
[13] "la recherche du temps perdu, tome 1 : Le c?t? de Guermantes"
[14] "A la recherche du temps perdu, Tome 1 : Un amour de Swann : Deuxième partie"
[15] "Broch? - A la recherche du temps perdu viii - le c?t? de guermantes ***"
[16] "La Prisonniere (A La Recherche Du Temps Perdu Tome VI)"
[17] "A la recherche du temps perdu (French Edition)"
[18] "LE COTE DE GUERMANTES TOME 1 - A LA RECHERCHE DU TEMPS PERDU 3"
[19] "Sodome Et Gomorrhe (a la Recherche Du Temps Perdu)"
[20] "a la Recherche Du Temps Perdu (1) (French Edition)"
[21] "A la recherche du temps perdu IV- Le côté de guermantes II"
[22] "A la recherche du temps perdu VIII - Le côté de Guermantes ***"
[23] "À la recherche du temps perdu : Le temps retrouvé"
[24] "À La Recherche Du Temps Perdu - Tome 3 - À L'ombre Des Jeunes Filles En Fleurs"
[25] "A la recherche du temps perdu, tome 4 : Un amour de Swann"
[26] "A La Recherche Du Temps Perdu Tome V: Sodome et Gomorrhe FRENCH"
[27] "Combray A La Recherche du Temps Perdu"
[28] "A la recherche du temps perdu (French Edition)"
[29] "A la recherche du temps perdu, A l'ombre des jeunes filles en fleurs, volume 2"
[30] "A la recherche du temps perdu IV- LecÃ té de guermantes II"
>
```

A cleaner code for a cleaner output

```
1  library(rvest)
2  library(tidyr)
3  library(dplyr)
4
5  larecherche <- read_html("https://www.abebooks.com/servlet/SearchResults?s
6
7  price <- larecherche %>%
8       html_nodes(".item-price .price") %>%
9       html_text() %>%
10      readr::parse_number()
11
12 title<-larecherche %>%
13   html_nodes(".col-xs-8 a span") %>%
14   html_text() %>%
15   readr::parse_character()
16
17 combined <- data_frame(title, `date and time` = Sys.time(), price)
18
19
```

Collecting your own data

- It is not as infeasible as it sounds!
- Survey tool on the internet (survey monkey, amazon mturk)
- Install Apps on willing participants that will track their movements (or other things) with https://moves-app.com/
- Sit in the science center and administer questionnaires
- Set up some A/B testing
- And of course if you have more money, organize a data collection team to collect whatever you would like!

Steps for collecting your own data

- Obtain the funding you may need
- Prepare a data management plan : how will you keep the data safe? will you share it?
- Obtain Human Subjects Approval
- Design your data collection instrument
- Pilot your data collection instrument
- Implement!

Protecting Human Subjects

The research governing human subjects is regulated, to ensure the protection of the participants.

The background, Nazi research, Tuskegee Syphilis trials

- American medical research project conducted by the U.S. Public Health Service from 1932 to 1972, examined the natural course of untreated syphilis in black American men.
- The subjects, all impoverished sharecroppers from Macon county, Alabama, were unknowing participants in the study; they were not told that they had syphilis, nor were they offered effective treatment.
- By the end of the experiment, 28 of the men had died directly of syphilis, 100 were dead of related complications, 40 of their wives had been infected, and 19 of their children had been born with congenital syphilis.
- People were also lured to come for tests with publicity of free treatment.

- In 1972 the whistle was blown, and the men finally won a $10 million class action trial against the PHS. The scientific merit of the study was also shoddy: apparently not much was ever learnt from it about how to treat the disease
- President Clinton apologized in the name of the Nation in 1997.

What is wrong here?

Protection of Human Subject

The research involving human subjects is governed by federal regulation . HHS Regulations for the Protection of Human Subjects at Title 45 Code of Federal Regulations Part 46.
The HHS regulations are intended to implement the basic ethical principles governing the conduct of human subjects research. These ethical principles are set forth in the report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research entitled: Ethical Principles and Guidelines for the Protection of Human Subjects of Research (the ”Belmont Report”).

Human subject research

Research - A systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge. Activities that meet this definition constitute research for purposes of the HHS regulations, whether or not they are conducted or supported under a program, which is considered research for other purposes. For example, some demonstration and service programs may include research activities.
Human Subject - A living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information.

Human subject research

Research - A systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge. **This means that Facebook or Amazon can experiment as much as they want on you unless they publish; but if you are working with them with the goal of publishing you need to go through an IRB**
Human Subject - A living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information.

Key principles of the Belmont report

1. Respect for persons
   - Respect individual autonomy
   - Protect individuals with reduced autonomy
2. Beneficence
   - Maximize benefits and minimize harms
3. Justice
   - Equitable distribution of research burdens and benefits

Related requirements

Application of the general ethical principles to the conduct of human subjects research leads to the following requirements:

- Respect for Persons
  - Informed consent
  - Protecting privacy and maintaining confidentiality
  - Additional safeguards for protection of subjects likely to be vulnerable to coercion or undue influence
- Beneficence
  - Assessment of risk/benefit analysis including study design
  - Ensure that risks to subjects are minimized
  - Risk justified by benefits of the research
- Justice
  - Ensure that selection of subjects is equitable.

How this works in practice

- For any research by an MIT, you need to be trained in human subject
- You need to submit to MIT COUHES a form describing your research.
- You need to follow the appropriate deadlines; and get the authorization BEFORE you start.
- You need to submit your informed consent forms as well, unless you request a waiver
- The form is used to asses if there is risk to the subjects or others.
- A committee assess the risks and the benefits, and if necessary asks you for changes to protect the subjects better.
- When work is conducted abroad, typically you also need to obtain human subject permission from that country.
- Some research which has minimum risk and does not involve individually linked data is considered exempt.

14.310x Data Analysis for Social Scientists
Spring 2023