Finding Data for Economics Research

UC Berkeley Library

February 2018

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."

-John Tukey

Before anything else…

Plan your Research with a Literature Review

http://www.lib.berkeley.edu/

http://scholar.google.com

http://guides.lib.berkeley.edu/all-guides

Plan your Research with a Literature Review

Plainly (as best you can) state your variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

Plainly (as best you can) state your variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

$$Salary = \beta_0 + \beta_1 Gender + \beta_2 Race + \ldots \beta_p X_p + \epsilon$$

Structure and availability of data

Unit of Analysis	**Geography**	**Time-Period**	**Frequency**
Aggregated or Microdata? (counties/nations/households vs. individuals)	Is there a geographic component to your topic? (U.S., Sub-Saharan Africa, India)	Do you want a data for a specific time period? (1980-2000, 1930-1960)	How often do you want measures for your variables? (every year, every ten years, monthly, quarterly)

Providers

Researchers	**Government Agencies**	**NGO/IGOs**	**Research Organizations**
Are there people you know who are doing this kind of research?	Think about government agencies - is the request for some official statistics or data that they'd be likely to collect and publish? (Department of Energy, CDC, Census Bureau)	Are there councils or interest organizations devoted to the topic that might collect data independently? (World Bank, OECD)	Would any specific research organizations be interested in the topic? (Pew, Roper, Gallup, ACLU)

The 80/20 “Rule”

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data.

-Dasu and Johnson

Tidy Data

"Happy families are all alike; every unhappy family is unhappy in its own way."

– Leo Tolstoy

"Tidy datasets are all alike, but every messy dataset is messy in its own way."

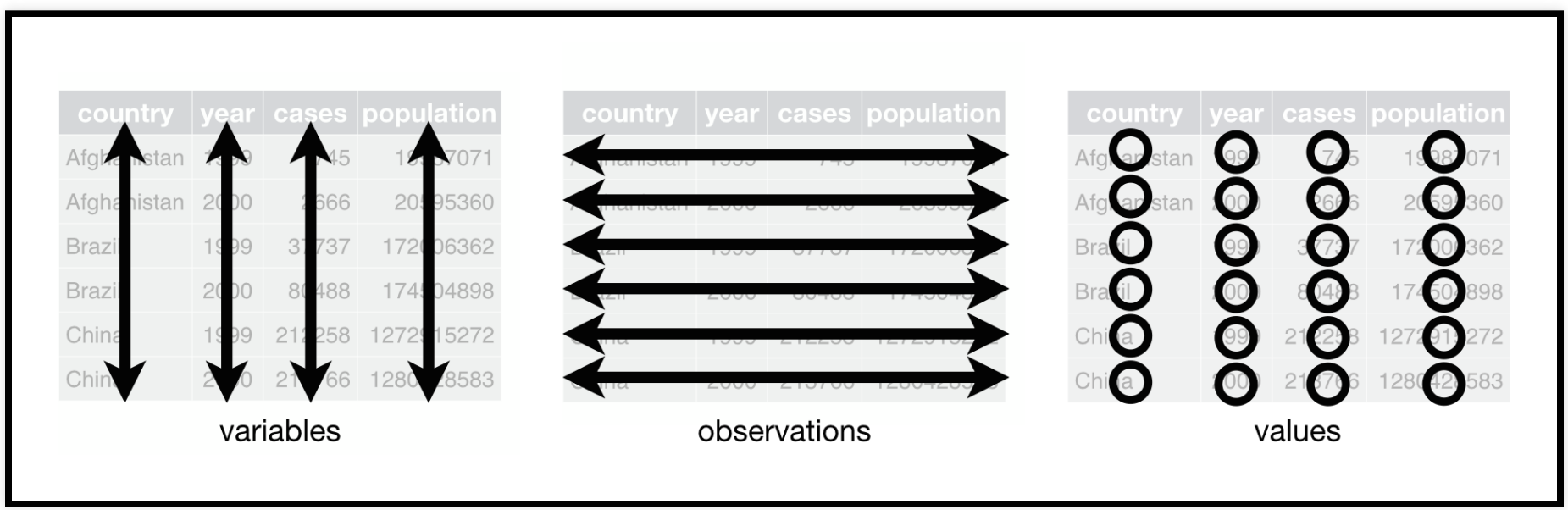
– Hadley Wickham

Tidy Data = Happy Data

Tidy Data has the following attributes:

Each variable forms a column and contains values

Each observation forms a row



Tidy Data = Happy Data

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

Variable Naming

Bad_Variable_Name	**Good_Variable_Name**	**Description**
gnp-2002; gnp#2002		
real interest rate		
1st_score; 2003gnp		
REG; glm; ttest		
xxx; yyy; zmdje;		
gender; race		
Whats_Your_Favorite_Color?		
INCOME; Int_us2003;		
April 20, 2017		

Variable Naming

Bad_Variable_Name	**Good_Variable_Name**	**Description**
gnp-2002; gnp#2002	gnp2010	avoid special characters
real interest rate	real_int	Use underscore
1st_score; 2003gnp	score1; gnp2003	Begin with a character
REG; glm; ttest	reg_out; glm1	Avoid reserved words
xxx; yyy; zmdje;	invest; interest	Use meaningful names
gender; race	male; asian	Use a value of dummy
Whats_Your_Favorite_Color?	fav_color	The shorter, the better
INCOME; Int_us2003;	income; intUS03	Use lower cases
April 20, 2017	2017-04-20	Use common ISO year format

Missing Data

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Library Licensed Data Aggregators

Data Planet

Social Explorer

Data Repositories for Replication Data

Dataverse

ICPSR

Data.gov

American Economics Association

APIs

https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/

Scraping

https://en.wikipedia.org/wiki/UFO_sightings_in_the_United_St

Scraping with Python

```
In [33]: import pandas as pd
         import requests


         url = "https://en.wikipedia.org/wiki/UFO_sightings_in_the_United_States"
         response = requests.get(url)

         df = pd.read_html(response.content)[1]
         print(df)
```

```
                      0                 1                 2  \
0                  Date              City             State
1            April 1941    Cape Girardeau          Missouri
2     February 24, 1942       Los Angeles        California
3         June 21, 1947      Maury Island        Washington
4         June 24, 1947      Maury Island        Washington
5          July 7, 1947            Helena           Montana
6             July 1947           Roswell        New Mexico
7       January 7, 1948         Maysville          Kentucky
8         July 24, 1948        Montgomery           Alabama
9       October 1, 1948             Fargo      North Dakota
10         May 11, 1950       McMinnville            Oregon
11      August 15, 1950       Great Falls           Montana
12      August 25, 1951           Lubbock             Texas
13        July 24, 1952       Carson Sink            Nevada
14         May 24, 1952           Burbank        California
15        July 13, 1952               NaN  Washington, D.C.
16   September 12, 1952         Flatwoods     West Virginia
17       August 5, 1953          Bismarck      North Dakota
```

Scraping with R

```
library(rvest)
library(dplyr)
ufo <- read_html("https://en.wikipedia.org/wiki/UFO_sightings_in_the_Uni

ufo_date <- html_nodes(ufo,'td:nth-child(1)') %>% html_text()
ufo_date <- ufo_date[c(-1, -44)] #remove extra elements
ufo_state <- html_nodes(ufo,'td:nth-child(3)') %>% html_text()
ufo_name <- html_nodes(ufo,'td:nth-child(4)') %>% html_text()
ufo_df<-data.frame(date = ufo_date, name = ufo_name, state = ufo_state)

head(ufo_df, n =5)
```

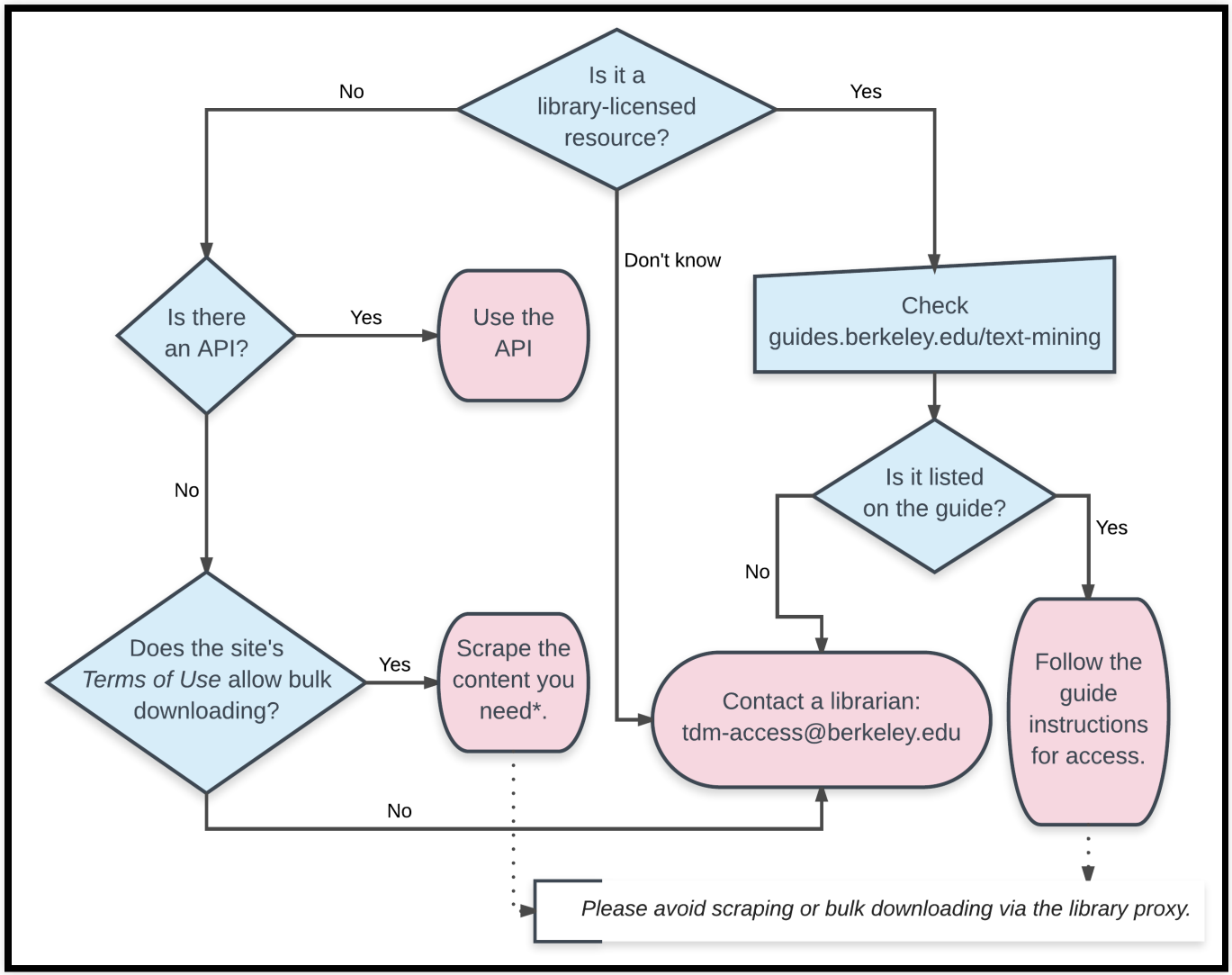
```
##                date                          name      state
## 1        April 1941    Cape Girardeau UFO crash   Missouri
## 2 February 24, 1942       Battle of Los Angeles California
## 3     June 21, 1947       Maury Island incident Washington
## 4     June 24, 1947 Kenneth Arnold UFO sighting Washington
## 5      July 7, 1947                                Montana
```

Miscellaneous Collections

https://vincentarelbundock.github.io/Rdatasets/datasets.html

https://github.com/caesar0301/awesome-public-datasets

Text-mining


Is it a library-licensed resource?
No
Yes
Don't know
Is there an API?
Yes
Use the API
No
Does the site's *Terms of Use* allow bulk downloading?
Yes
Scrape the content you need*.
No
Check guides.berkeley.edu/text-mining
Is it listed on the guide?
No
Yes
Contact a librarian: tdm-access@berkeley.edu
Follow the guide instructions for access.
Please avoid scraping or bulk downloading via the library proxy.

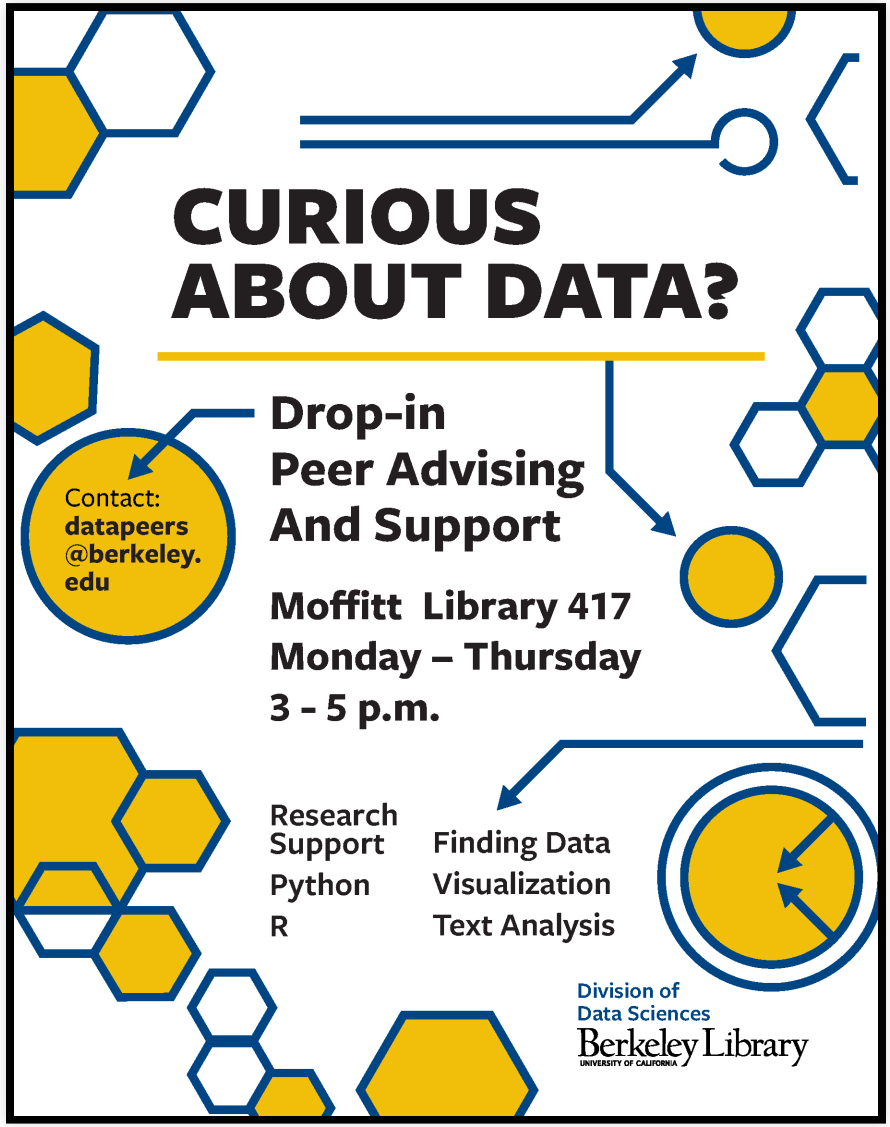
Text-mining

http://guides.lib.berkeley.edu/text-mining

D-Lab, Library Data Lab, Statistics Department, Student Learning Center

- http://dlab.berkeley.edu/
- http://www.lib.berkeley.edu/libraries/data-lab
- https://statistics.berkeley.edu/consulting
- http://slc.berkeley.edu/econ

Peer Advising at Moffitt



https://data.berkeley.edu/education/data-peers-consulting

Reaching out

http://www.lib.berkeley.edu/libraries/data-lab