Show me the data!

Week01: Introduction

Big Data & Social Analysis

R

Instructors：Chung-pei Pien

ZU1942001/266868001/Z23937001/ZM1941001

International College of INNOVATION
National Chengchi University
國立政治大學創新國際學院

CONTENTS

Introduction

01 Introduction

The Level of This Course

Domain Knowledge

- Sociology
- Economy
- Politics
- Data Science
- Statistics

This Course

- Basic R Review
- Coding Application
- Project-driven Process

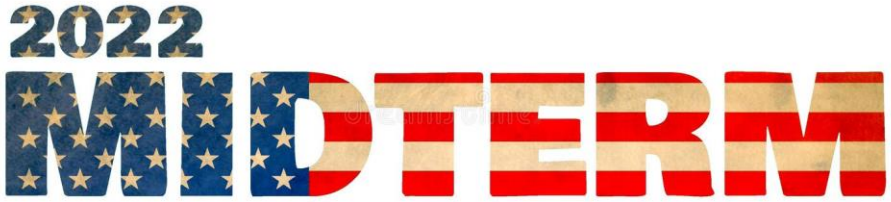
Other GTIM

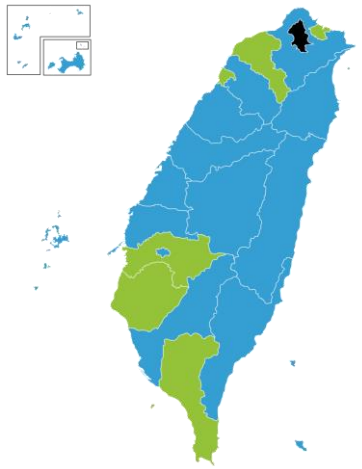
- Capstone
- Data Science
- Business Data Analysis

01 Introduction

Capstone

I will negotiate with two media for the capstone course in Fall 2022.





There are two crucial elections in this November: US midterm election and TW major election.

The participants of the capstone course can produce these two elections' data journalism and publish them in these media.

01 Introduction

Capstone

English Media: Taiwan Major Election

Chinese Media: US Midterm Election

To prepare the capstone course, I will use my previous election data projects to teach you coding in this course.

01 Introduction

The Structure of This Course

Election Data and R

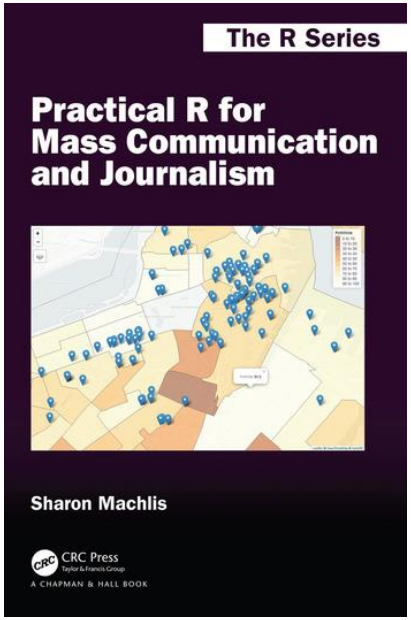
Data Exploration	Regular Expressions	Loop	Data Cleaning	Web Crawler	Data Visualization

Advanced Data Projects

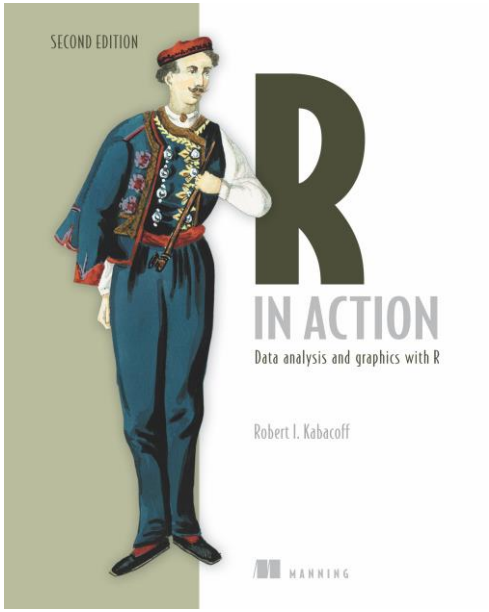
Taiwan Presidential Election Data Project	US Midterm Election Data Project	Taiwan Mayor Election Data Project	Twitter and US Presidential Election

01 Introduction

Reading



Machlis, Sharon. 2019. Practical R for Mass Communication and Journalism. CRC Pres



Kabacoff, Robert. 2015. R in Action, Data Analysis and Graphics with R. Manning Publications; 2 Edition

01 Introduction

Reading

R Reference Card

by Tom Short, EPRI PEAC, tshort@epri-peac.com 2004-11-07
Granted to the public domain. See www.Rpad.org for the source and latest version. Includes material from *R for Beginners* by Emmanuel Paradis (with permission).

Getting help

Most R functions have online documentation.

`help(topic)` documentation on `topic`
`?topic` id.
`help.search("topic")` search the help system
`apropos("topic")` the names of all objects in the search list matching the regular expression "topic"
`help.start()` start the HTML version of help
`str(a)` display the internal *str*ucture of an R object
`summary(a)` gives a "summary" of `a`, usually a statistical summary but it is *generic* meaning it has different operations for different classes of `a`
`ls()` show objects in the search path; specify `pat="pat"` to search on a pattern
`ls.str()` `str()` for each variable in the search path
`dir()` show files in the current directory
`methods(a)` shows S3 methods of `a`
`methods(class=class(a))` lists all the methods to handle objects of class `a`

Input and output

`load()` load the datasets written with `save`
`data(x)` loads specified data sets
`library(x)` load add-on packages
`read.table(file)` reads a file in table format and creates a data frame from it; the default separator `sep=""` is any whitespace; use `header=TRUE` to read the first line as a header of column names; use `as.is=TRUE` to prevent character vectors from being converted to factors; use `comment.char=""` to prevent "#" from being interpreted as a comment; use `skip=n` to skip `n` lines before reading data; see the help for options on row naming, NA treatment, and others
`read.csv("filename",header=TRUE)` id. but with defaults set for reading comma-delimited files
`read.delim("filename",header=TRUE)` id. but with defaults set for reading tab-delimited files
`read.fwf(file,widths,header=FALSE,sep="",as.is=FALSE)` read a table of *f*ixed *w*idth *f*ormatted data into a 'data.frame'; `widths` is an integer vector, giving the widths of the fixed-width fields
`save(file,...)` saves the specified objects (...) in the XDR platform-independent binary format
`save.image(file)` saves all objects
`cat(..., file="", sep=" ")` prints the arguments after coercing to character; `sep` is the character separator between arguments
`print(a, ...)` prints its arguments; generic, meaning it can have different methods for different objects
`format(x,...)` format an R object for pretty printing
`write.table(x,file="",row.names=TRUE,col.names=TRUE, sep=" ")` prints `x` after converting to a data frame; if `quote` is `TRUE`, character or factor columns are surrounded by quotes ("); `sep` is the field separator; `eol` is the end-of-line separator; `na` is the string for missing values; use `col.names=NA` to add a blank column header to get the column headers aligned correctly for spreadsheet input
`sink(file)` output to `file`, until `sink()`

Most of the I/O functions have a `file` argument. This can often be a character string naming a file or a connection. `file=""` means the standard input or output. Connections can include files, pipes, zipped files, and R variables.

On windows, the file connection can also be used with `description = "clipboard"`. To read a table copied from Excel, use
`x <- read.delim("clipboard")`
To write a table to the clipboard for Excel, use
`write.table(x,"clipboard",sep="\t",col.names=NA)`
For database interaction, see packages `RODBC`, `DBI`, `RMySQL`, `RPgSQL`, and `ROracle`. See packages `XML`, `hdf5`, `netCDF` for reading other file formats.

Data creation

`c(...)` generic function to combine arguments with the default forming a vector; with `recursive=TRUE` descends through lists combining all elements into one vector
`from:to` generates a sequence; ":" has operator priority; `1:4 + 1` is "2,3,4,5"
`seq(from,to)` generates a sequence by= specifies increment; `length=` specifies desired length
`seq(along=x)` generates `1, 2, ..., length(along)`; useful for for loops
`rep(x,times)` replicate `x` `times`; use `each=` to repeat "each" element of `x` `each` times; `rep(c(1,2,3),2)` is `1 2 3 1 2 3`; `rep(c(1,2,3),each=2)` is `1 1 2 2 3 3`
`data.frame(...)` create a data frame of the named or unnamed arguments; `data.frame(v=1:4,ch=c("a","B","c","d"),n=10)`; shorter vectors are recycled to the length of the longest
`list(...)` create a list of the named or unnamed arguments; `list(a=c(1,2),b="hi",c=3i)`;
`array(x,dim=)` array with data `x`; specify dimensions like `dim=c(3,4,2)`; elements of `x` recycle if `x` is not long enough
`matrix(x,nrow=,ncol=)` matrix; elements of `x` recycle
`factor(x,levels=)` encodes a vector `x` as a factor
`gl(n,k,length=n*k,labels=1:n)` generate levels (factors) by specifying the pattern of their levels; `k` is the number of levels, and `n` is the number of replications
`expand.grid()` a data frame from all combinations of the supplied vectors or factors
`rbind(...)` combine arguments by rows for matrices, data frames, and others
`cbind(...)` id. by columns

Slicing and extracting data

Indexing vectors

`x[n]`	n^{th} element
`x[-n]`	*all but* the n^{th} element
`x[1:n]`	first `n` elements
`x[-(1:n)]`	elements from `n+1` to the end
`x[c(1,4,2)]`	specific elements
`x["name"]`	element named "name"
`x[x > 3]`	all elements greater than 3
`x[x > 3 & x < 5]`	all elements between 3 and 5
`x[x %in% c("a","and","the")]`	elements in the given set

Indexing lists

`x[n]`	list with elements `n`
`x[[n]]`	n^{th} element of the list
`x[["name"]]`	element of the list named "name"
`x$name`	id.

Indexing matrices

`x[i,j]`	element at row `i`, column `j`
`x[i,]`	row `i`
`x[,j]`	column `j`
`x[,c(1,3)]`	columns 1 and 3
`x["name",]`	row named "name"

Indexing data frames (matrix indexing plus the following)

`x[["name"]]`	column named "name"
`x$name`	id.

Variable conversion

`as.array(x)`, `as.data.frame(x)`, `as.numeric(x)`, `as.logical(x)`, `as.complex(x)`, `as.character(x)`, ... convert type; for a complete list, use `methods(as)`

Variable information

`is.na(x)`, `is.null(x)`, `is.array(x)`, `is.data.frame(x)`, `is.numeric(x)`, `is.complex(x)`, `is.character(x)`, ... test for type; for a complete list, use `methods(is)`
`length(x)` number of elements in `x`
`dim(x)` Retrieve or set the dimension of an object; `dim(x) <- c(3,2)`
`dimnames(x)` Retrieve or set the dimension names of an object
`nrow(x)` number of rows; `NROW(x)` is the same but treats a vector as a one-row matrix
`ncol(x)` and `NCOL(x)` id. for columns
`class(x)` get or set the class of `x`; `class(x) <- "myclass"`
`unclass(x)` remove the class attribute of `x`
`attr(x,which)` get or set the attribute which of `x`
`attributes(obj)` get or set the list of attributes of `obj`

Data selection and manipulation

`which.max(x)` returns the index of the greatest element of `x`
`which.min(x)` returns the index of the smallest element of `x`
`rev(x)` reverses the elements of `x`
`sort(x)` sorts the elements of `x` in increasing order; to sort in decreasing order: `rev(sort(x))`
`cut(x,breaks)` divides `x` into intervals (factors); `breaks` is the number of cut intervals or a vector of cut points
`match(x, y)` returns a vector of the same length than `x` with the elements of `x` which are in `y` (`NA` otherwise)
`which(x == a)` returns a vector of the indices of `x` if the comparison operation is true (`TRUE`), in this example the values of `i` for which `x[i] == a` (the argument of this function must be a variable of mode logical)
`choose(n, k)` computes the combinations of k events among n repetitions $= n!/[(n-k)!k!]$
`na.omit(x)` suppresses the observations with missing data (`NA`) (suppresses the corresponding line if `x` is a matrix or a data frame)
`na.fail(x)` returns an error message if `x` contains at least one `NA`

Evaluation

02 Evaluation

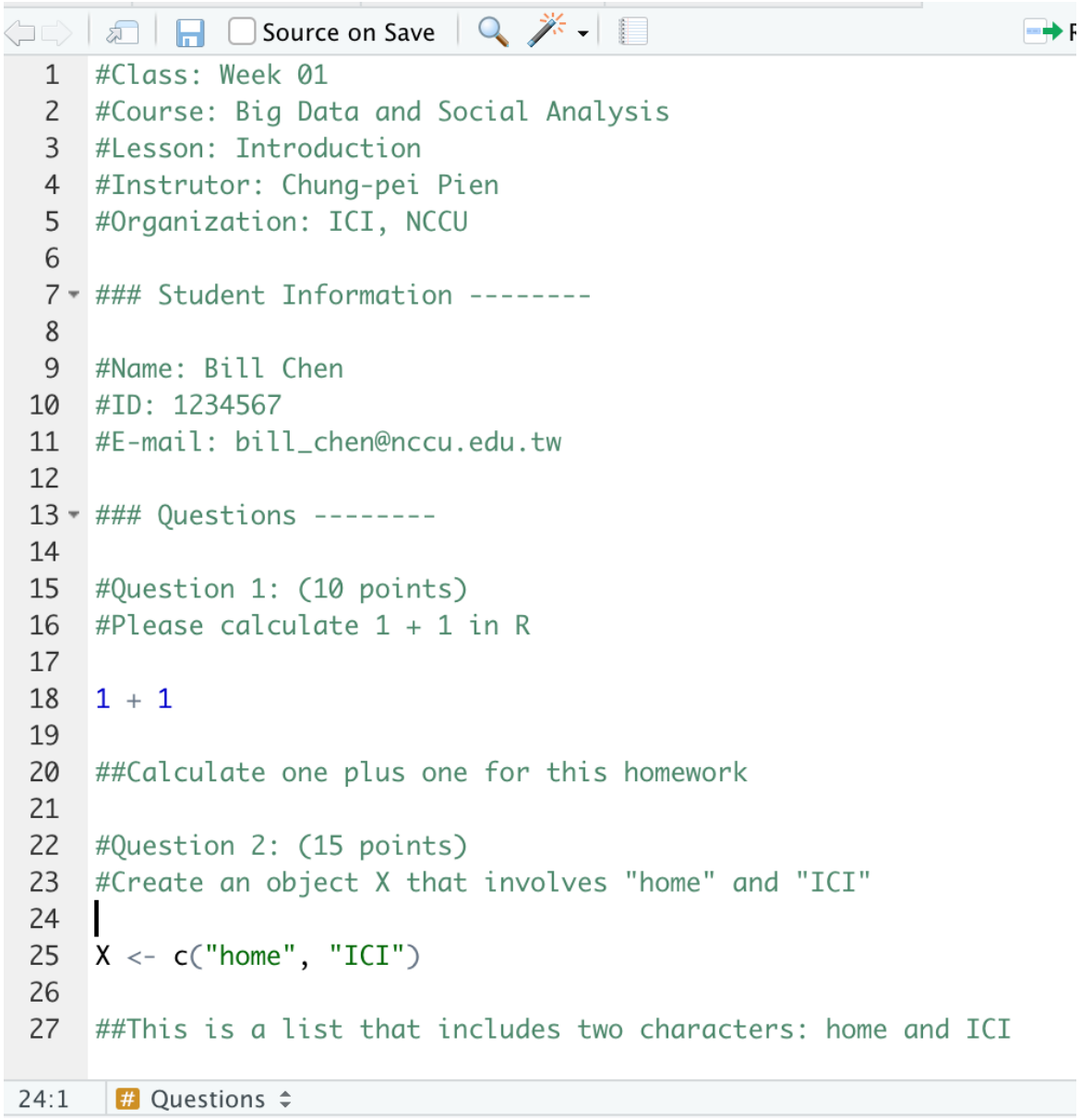
Items	Points	%
Assignments	10 x 30 = 300	23.08%
To-do List	100	7.69%
Midterm Exam	200	15.38%
Project Proposal	150	11.54%
Final Presentation	200	15.38%
Final Paper	350	26.92%
Total	1300	100%

02 Evaluation

Items	Points	%
Assignments	10 x 30 = 300	23.08%

- **10 weeks have assignments. I will upload them (R script file) to Moodle.**
- **Print and submit them in the beginning of the next weeks' classes**

02 Evaluation



02 Evaluation

Items	Points	%
Assignments	10 x 30 = 300	23.08%

- **No comments no points**
- **To meet the ICI dean's attendance request, assignments will link to your attendance: Students who do not ask for leave to skip the class can't submit assignments.**

02 Evaluation

Items	Points	%
To-do List	100	7.69%

- **In week 3, your should decide your teammates and email me your team members' lists.**
- **I will provide you a to-do list guideline.**
- **Your team can follow my suggestions to design your data projects.**
- **In Week 6, your team should submit your to-do lists. I will return them in Week 7.**

02 Evaluation

Items	Points	%
To-do List	100	7.69%

	Team Points	PM bonus	Writers bonus
A+	95	3	2
A	92	3	2
A-	90	2	2
B+	85	2	1
B	82	2	1
B-	80	1	1
C+	75	0	0
C	70	0	0

02 Evaluation

Items	Points	%
Midterm Exam	200	15.38%

- **Midterm exam will be held on April 21 (week 10)**
- **Open books and reference cards**
- **Only can open R-studio or R-studio cloud on your laptop**

02 Evaluation

Items	Points	%
Project Proposal	150	11.54%

- **Every team will be required to present your project proposal in 5-10 slides in the Week 13 for the final presentation and report.**

- **I will hand out a project proposal guideline in Week 7. You should follow the guideline to discuss your project with your teammates and me.**

02 Evaluation

Items	Points	%
Project Proposal	150	11.54%

	Team Points	PM bonus	Writers bonus
A+	143	5	3
A	138	4	3
A-	135	3	3
B+	127	3	1.5
B	123	3	1.5
B-	120	1.5	1.5
C+	112	0	0
C	105	0	0

02 Evaluation

Items	Points	%
Final Presentation	200	15.38%

- **Week 18, this course and other GTIM courses have a joint presentation party on June 15 (Wed.) 12:00pm.**
- **Every team has a booth (a table and a poster frames) and use a poster or a laptop's ppt to present your final project.**
- **The location may be at the 3F lobby of International building.**

02 Evaluation

Items	Points	%
Final Presentation	200	15.38%

	Team Points	PM bonus	Writers bonus
A+	190	6	4
A	184	6	4
A-	180	4	4
B+	170	4	2
B	164	4	2
B-	160	2	2
C+	150	0	0
C	140	0	0

02 Evaluation

Items	Points	%
Final Paper	350	26.92%

- **The final reports should be submitted to my mailbox or emailed before the deadline.**

1. **A4 10 pages (not include reference)**
2. **Times New Roman 12pt**
3. **Double spaces**
4. **Margin 1 inch**

02 Evaluation

Items	Points	%
Final Paper	350	26.92%

	Team Points	PM bonus	Writers bonus
A+	332.5	9	6
A	322	9	6
A-	315	6	6
B+	297.5	6	3
B	287	6	3
B-	280	3	3
C+	262.5	0	0
C	245	0	0

Data Project

Data Scientists: The Sexiest Jo[illegible] 21st Century

Davenport and Patil 2012

03 Data Project

Data analysis process:
Six independent but related steps

1. Identify Goals and Plans

2.Data Collection

3.Data Cleaning

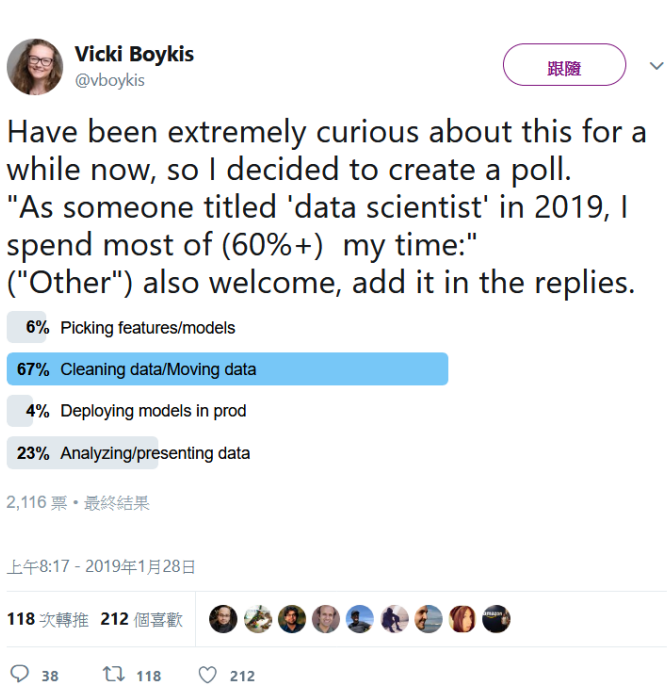
4.Data Analysis

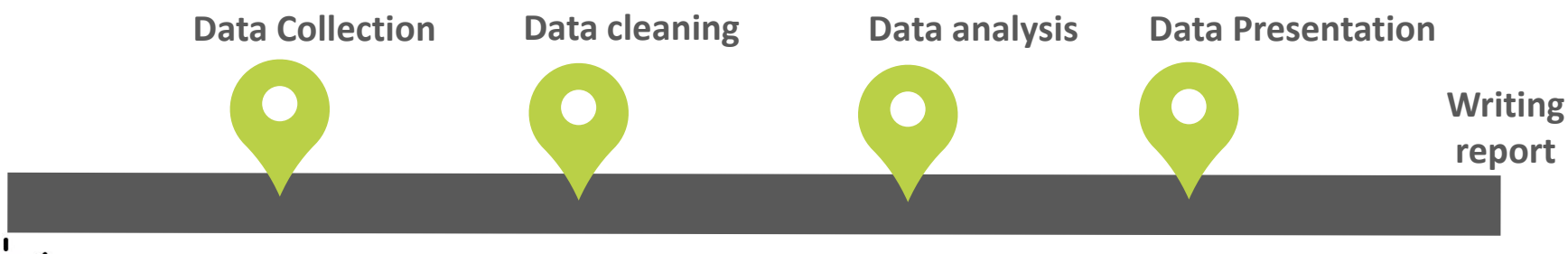
6.Writing Report

5.Data Presentation

3 Data Project

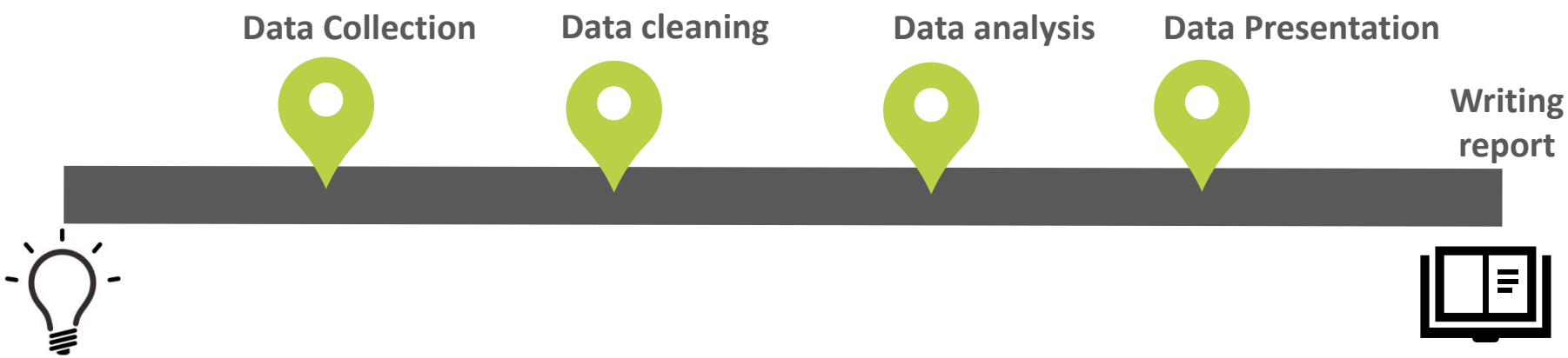
1. Data cleaning spends 80% project time: tired and frustrated
2. There is no data or model to achieve your goals
3. Examining your budgets and ability
4. Team work





3 Data Project

1. Domain Knowledge: Social issues
2. Coding: R (this courses), Python, and etc
3. Models: Data science and statistics courses
4. Graphics: Art designers/adobe illustrators
5. Presentation: Writing



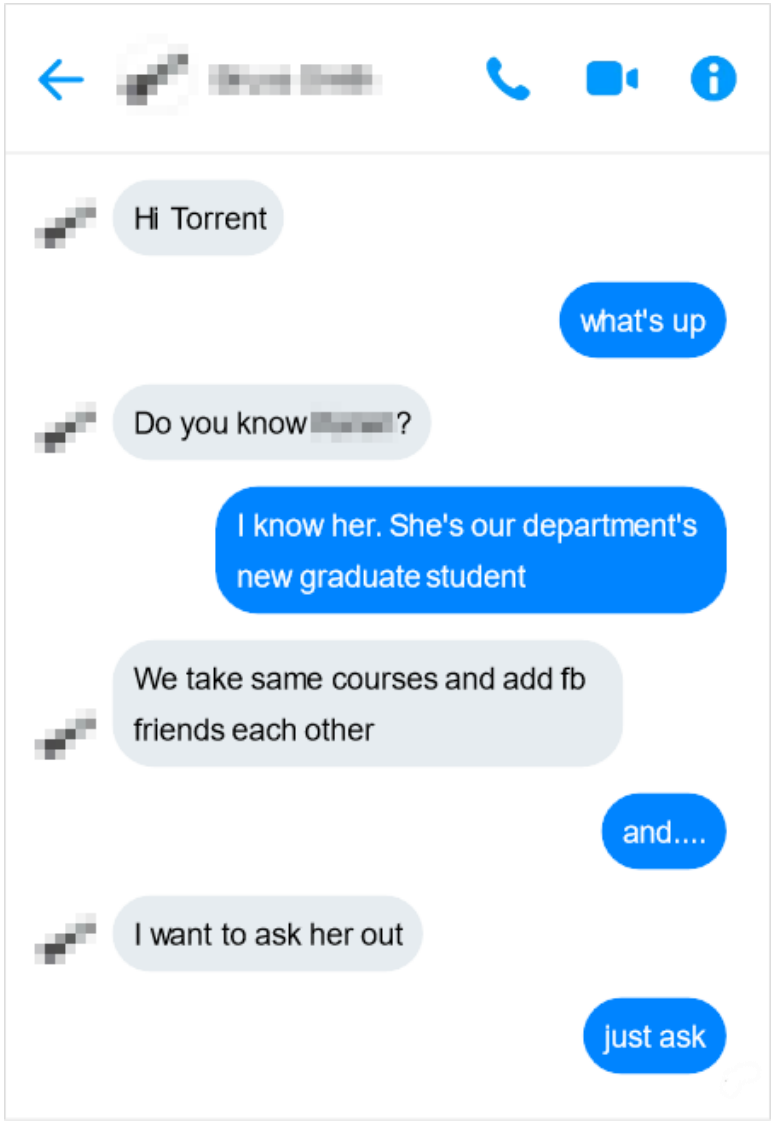
3 Data Project

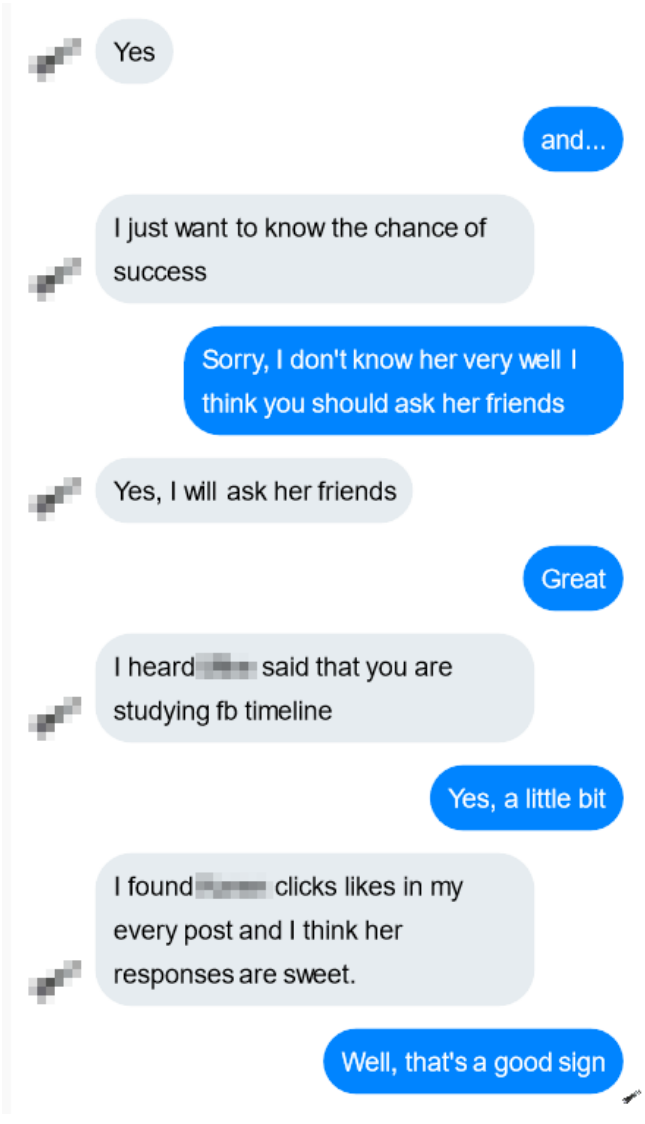
Facebook Analysis of User Activities

One day, a friend of mine raised a request in Facebook messenger……

3 Data Project

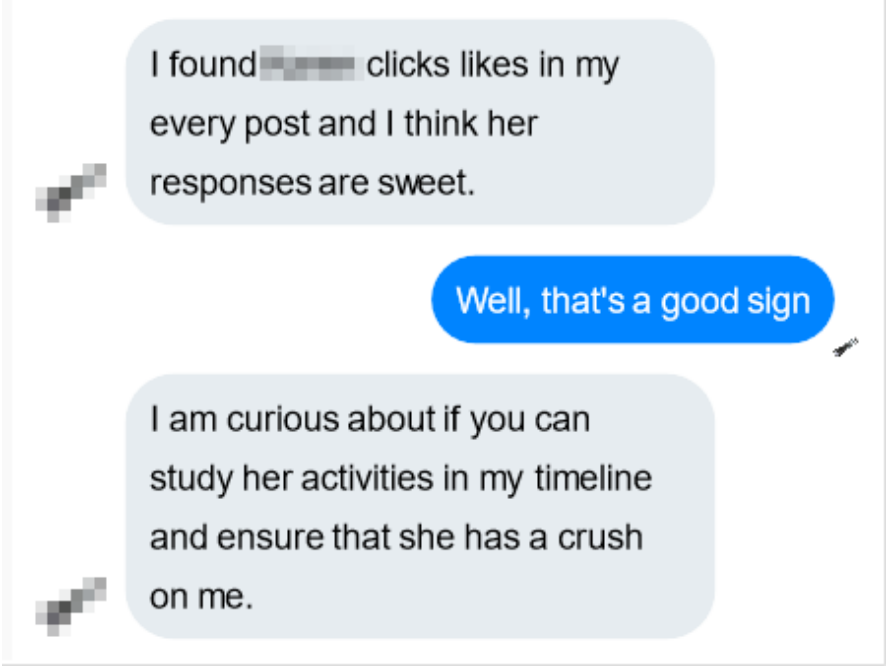
Facebook Analysis of User Activities





3 Data Project

Facebook Analysis of User Activities

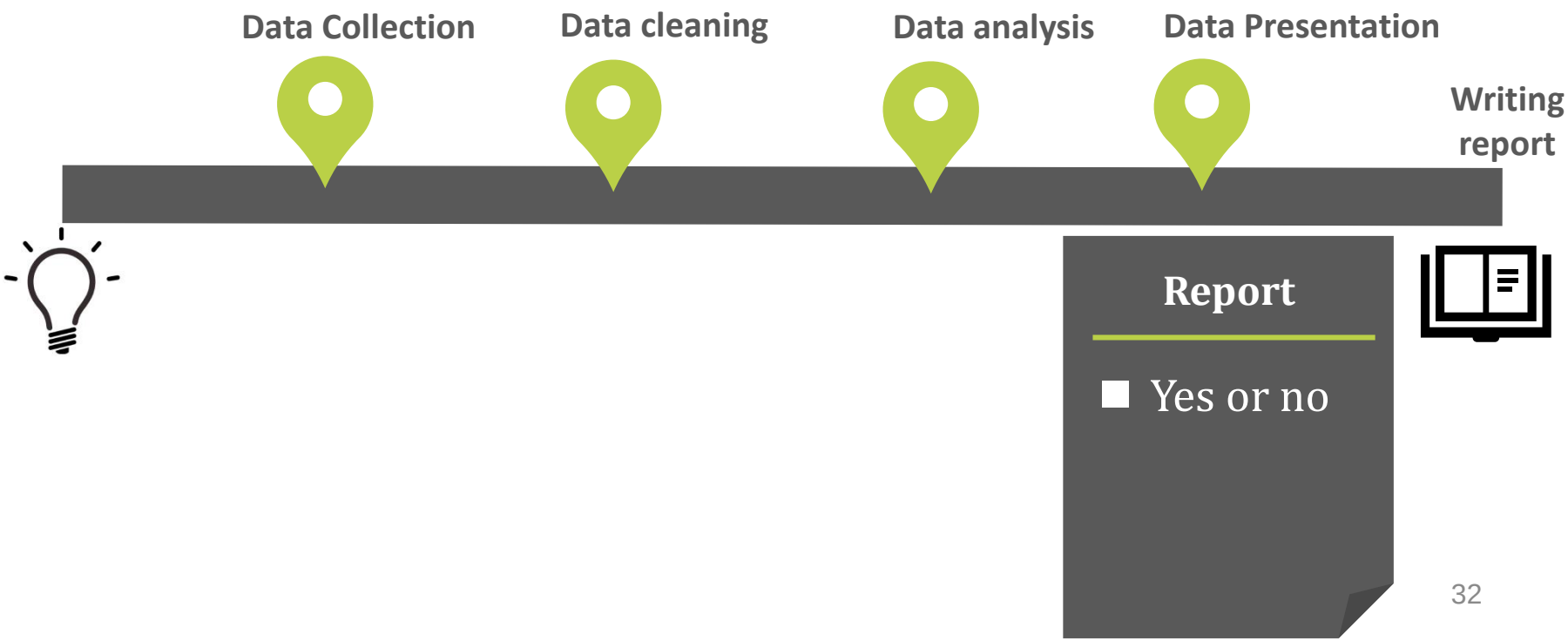


Do you think his demand – ensuring a girl likes him – is technically possible?

3 Data Project

Facebook Analysis of User Activities

- Using FB data to ensure that the girl likes him.

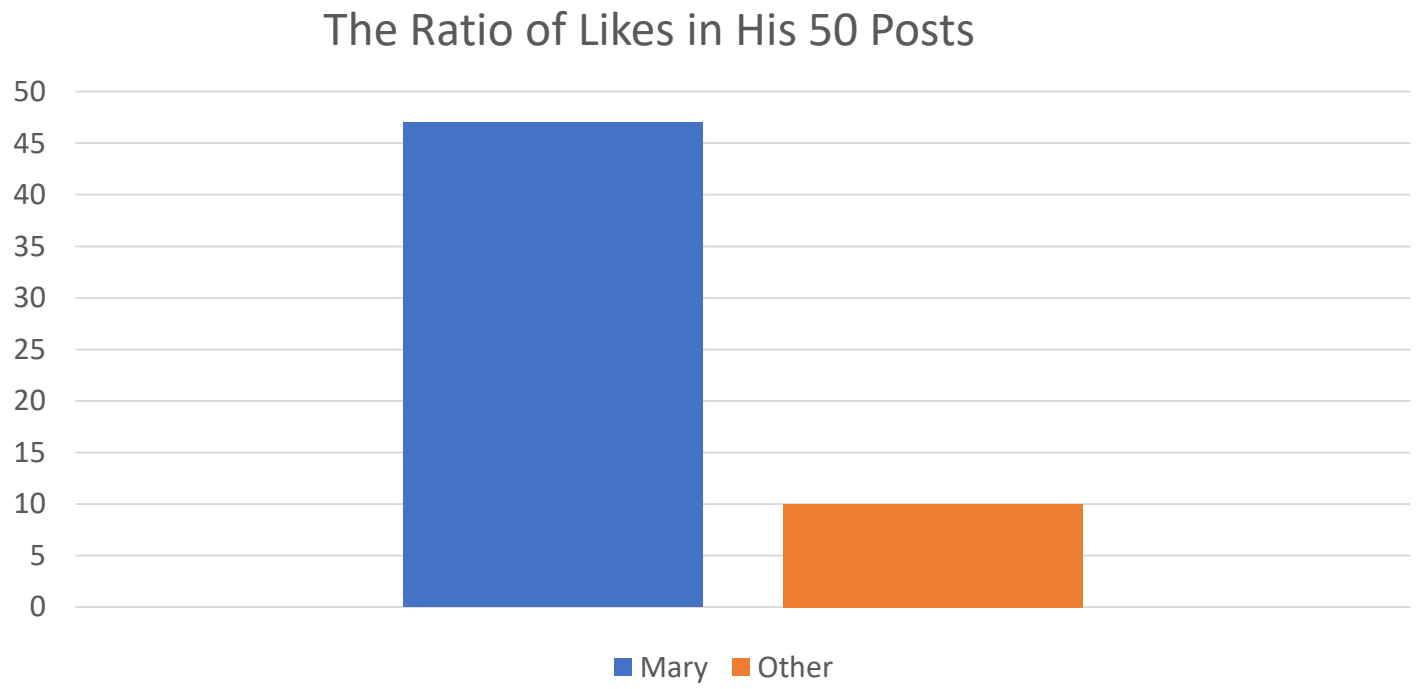


3 Data Project

Facebook Analysis of User Activities

Mary's activities are different than his other friends.

- **The ratio of her likes is larger then other friends, significantly.**



3 Data Project

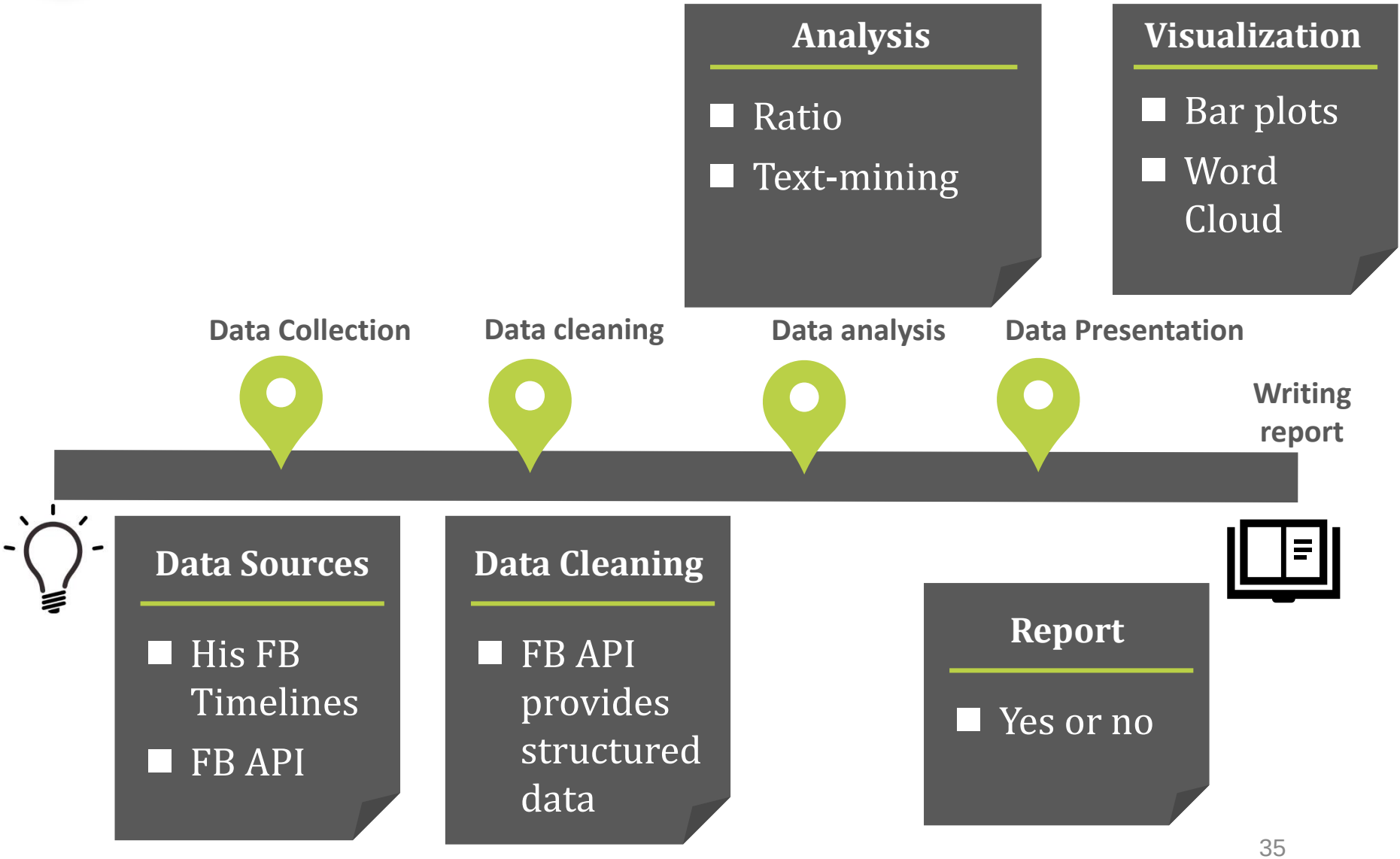
Facebook Analysis of User Activities

Mary's activities are different than his other friends.

- **Comparing other Facebook friends, her responses' key words are sweet.**

3 Data Project

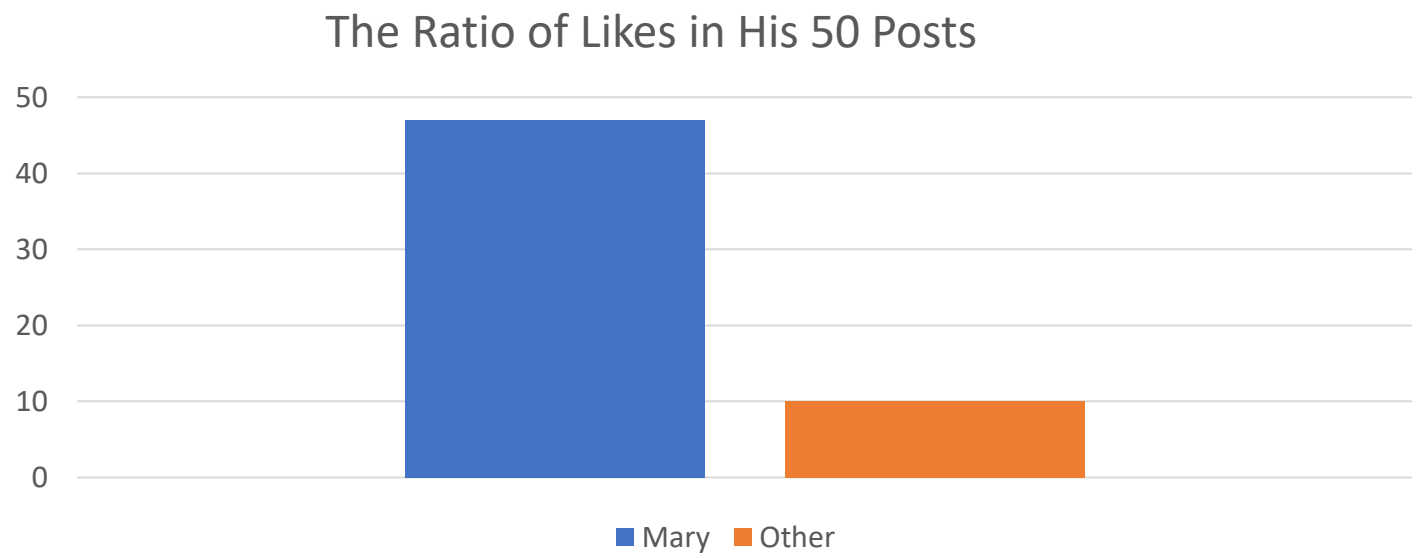
Facebook Analysis of User Activities



3 Data Project

Facebook Analysis of User Activities

The ratio of Mary's likes is larger than other friends, significantly.



- **Maybe she also has a high clicking like rate in her friends' timelines.**
- **We can't (or should not) get her friends' timelines.**

3 Data Project

Facebook Analysis of User Activities

Comparing other Facebook friends, her responses' key words are sweet.

- There are online dictionaries to check a word's positive or negative.
- Positive words do not mean "LIKE."
- You can do your dictionary, but you need a lot of data to do machine learning.

3 Data Project

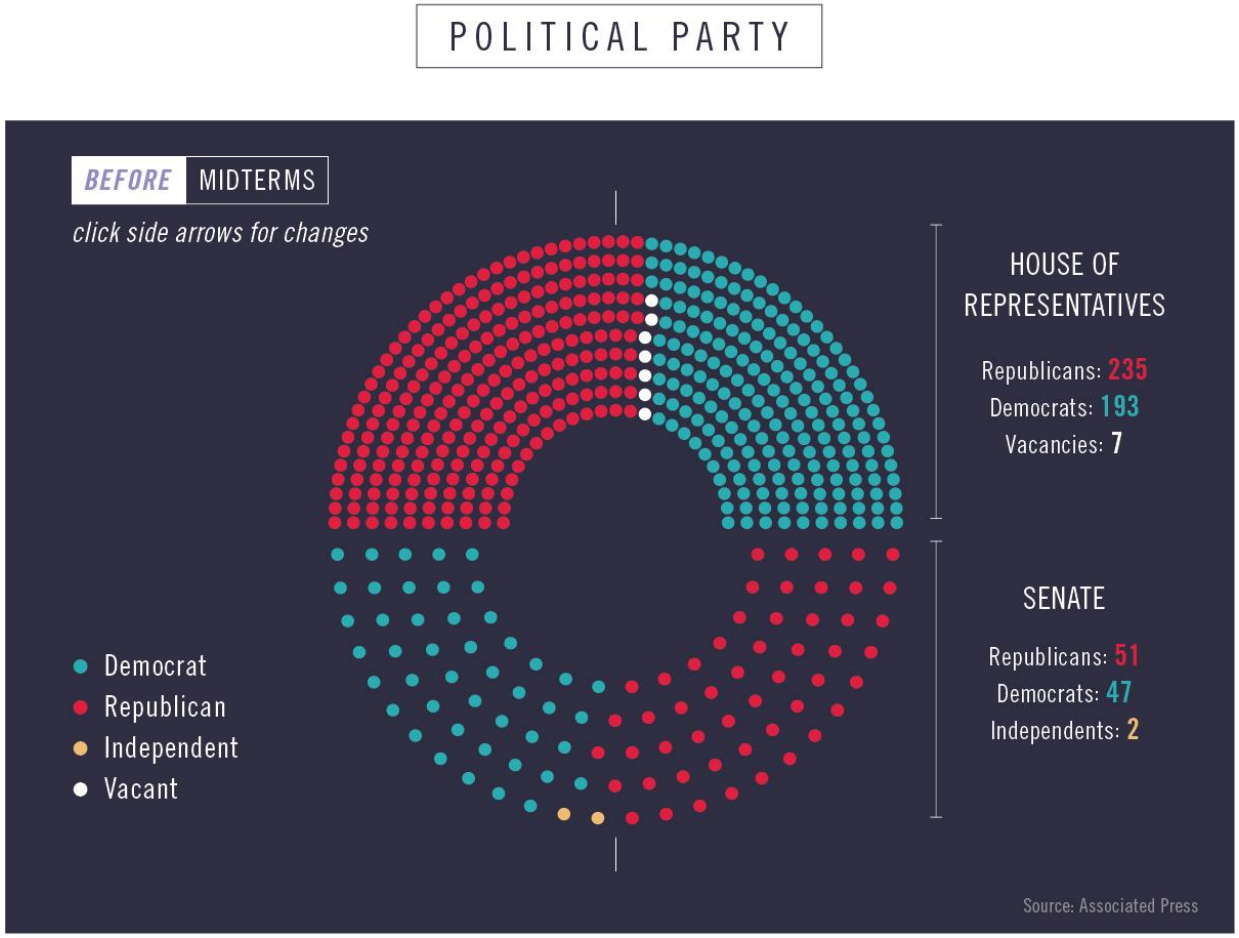
US Congressmen and Taiwan

In 2018, after Trump's 2 years' presidency, Democrats might flip the control of congress in the US midterm election.

Pro-Taiwan lawmakers would be replaced? New congress would support Taiwan?

3 Data Project

US Congressmen and Taiwan



2018

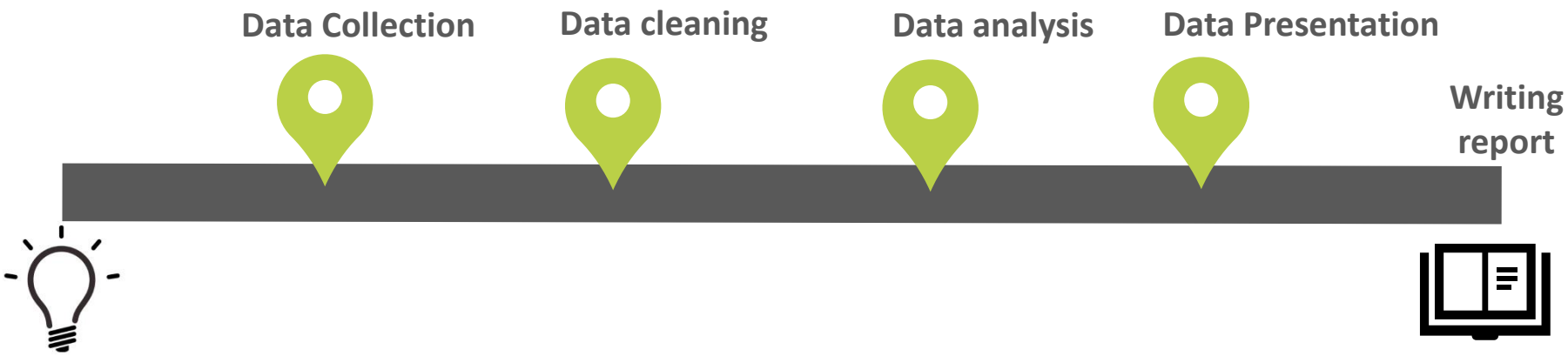
House: 435

Senate: 100

3 Data Project

US Congressmen and Taiwan

- All US lawmakers' attitude toward Taiwan before and after the midterm election



3 Data Project

US Congressmen and Taiwan

The primary difficulty of this project is data:

What kinds of data can represent a lawmaker to support Taiwan or not?????

News stories?
Twitters?
Congressional records?

3 Data Project

US Congressmen and Taiwan

	News Stories	Twitters	Congressional Records
Easy to get?	Is it possible to get all lawmakers' Taiwan opinions in media?	Does all lawmakers have Twitter account. Twitter's	US congress website has csv file or API?
Easy to clean?	News is unstructured data. How to evaluate pro-Taiwan or anti-Taiwan?	How to evaluate pro-Taiwan or anti-Taiwan	How to evaluate pro-Taiwan or anti-Taiwan?
The data can be counted or modeled?	It's very difficult to evaluate the level like or dislike TW	It's very difficult to evaluate the level like or dislike TW	Pro-Taiwan acts
The result can be explained?	Difficult	Difficult	Much better

3 Data Project

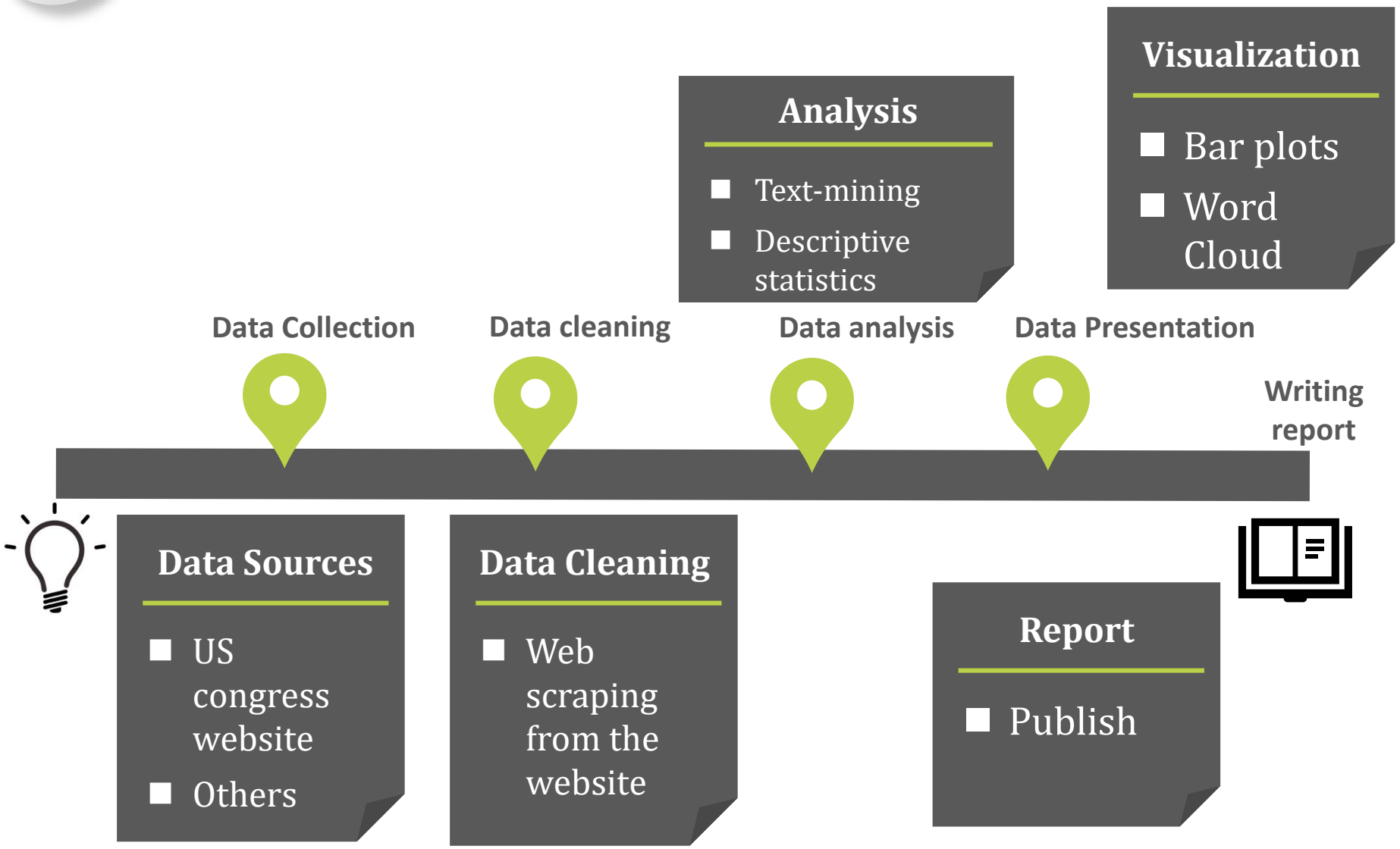
US Congressmen and Taiwan

I decided to use congressional records:

1. Pro-Taiwan acts: sponsors and voters
2. Statements
3. Lobby

3 Data Project

US Congressmen and Taiwan

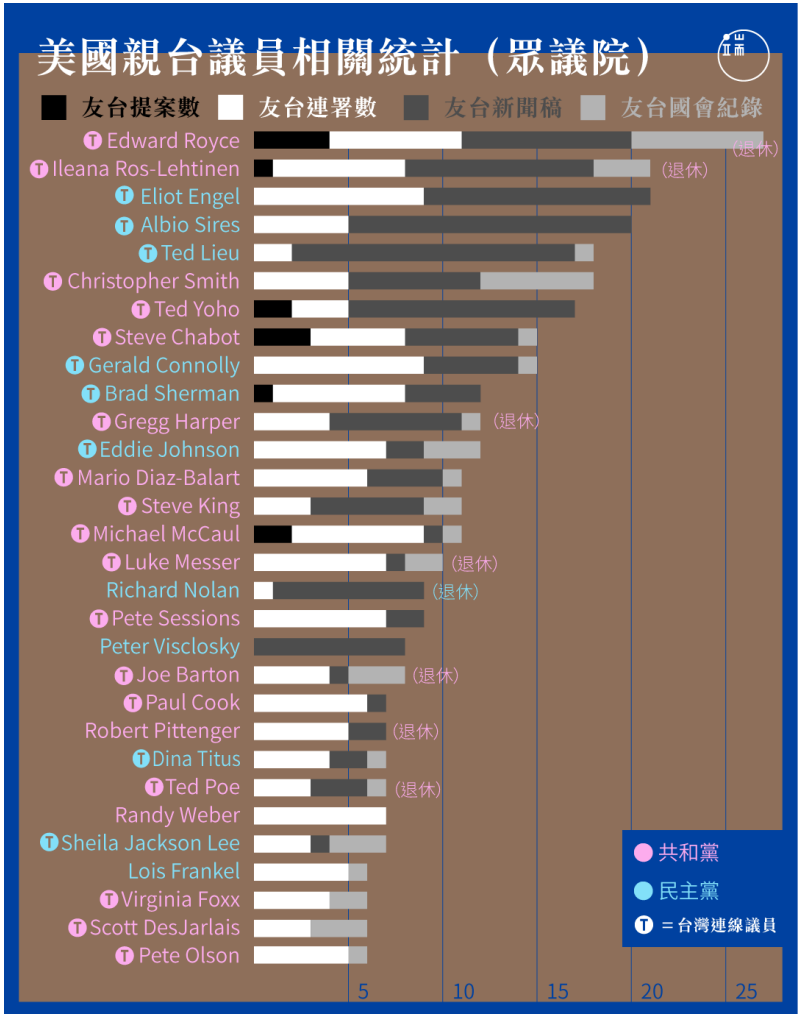


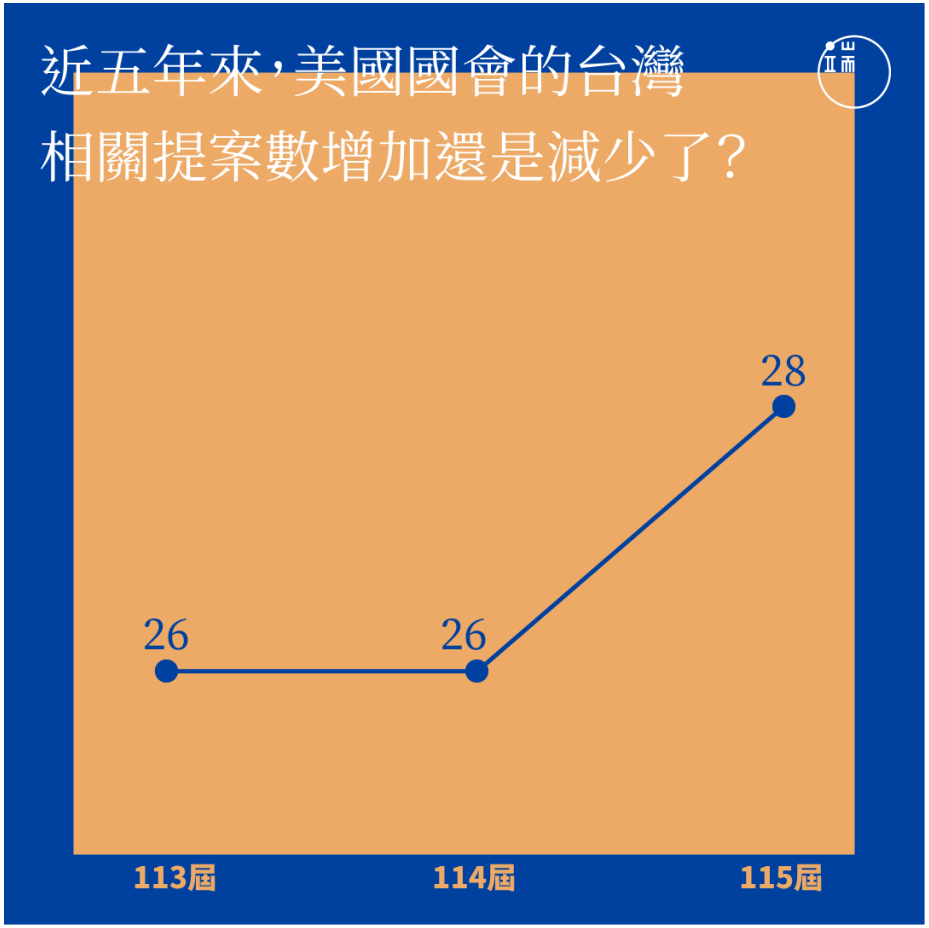
3 Data Project

US Congressmen and Taiwan

3 Data Project

US Congressmen and Taiwan





3 Data Project

Mazu and Election

In 2018, KMT candidate Han Kuo-yu (韓國瑜) won the Kaohsiung mayor election.

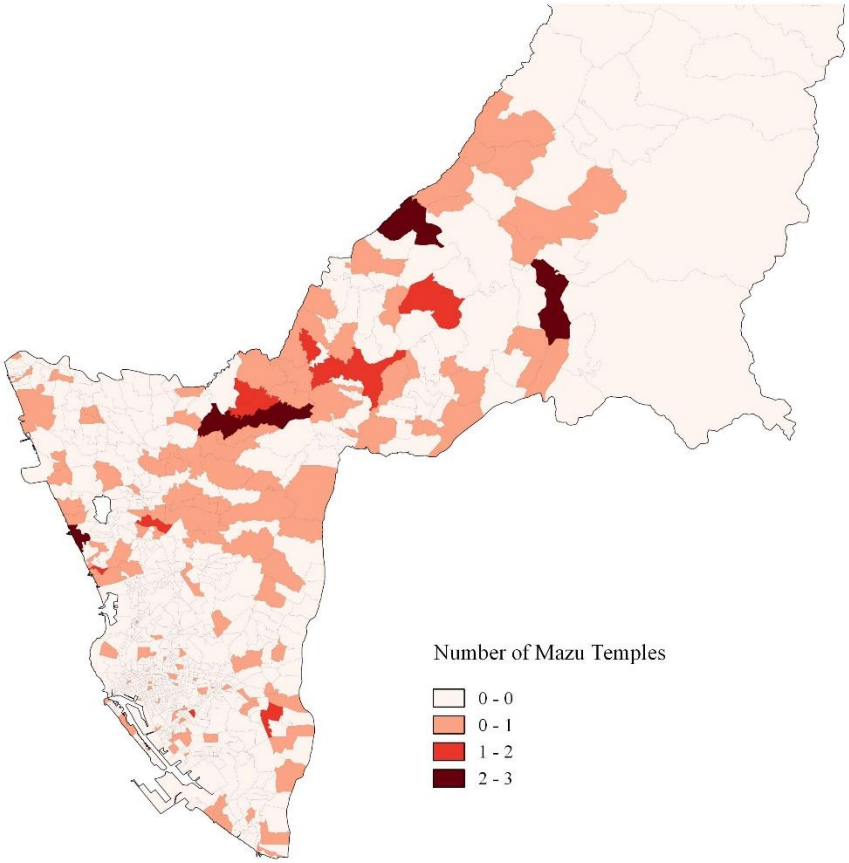
There are many arguments to explain his win.

One important argument is:

> **Many Mazu temples organized pilgrimages to the birthplace of Mazu. The Chinese government could influence their political attitudes and they tended to elect KMT.**

3 Data Project

Mazu and Election



Can we use data analysis to prove this hypothesis?

> **The Kaohsiung citizens who are influenced by Mazu temples are more likely to elect Han.**

Chien-yuan Sher, Chung-pei Pien, and Yu-hsi Liu have attempted to examine this hypothesis and write a research paper form 2018.

3 Data Project

Mazu and Election

- KH li which have more Mazu temples voted for Han?

Dependent variable: Voting rate for KMT
Independent variable: The number of Mazu temple

Analysis

- Aggregate multi-level nested logit model

Visualization

- GIS
- Regression Tables

Data Collection → Data cleaning → Data analysis → Data Presentation → Writing report

Data Sources

- Ministry of the Interior
- Central Election Commission

Data Cleaning

- 6 Months to merge data

Report

- Research Paper
- Data Journalism

3 Data Project

Mazu and Election

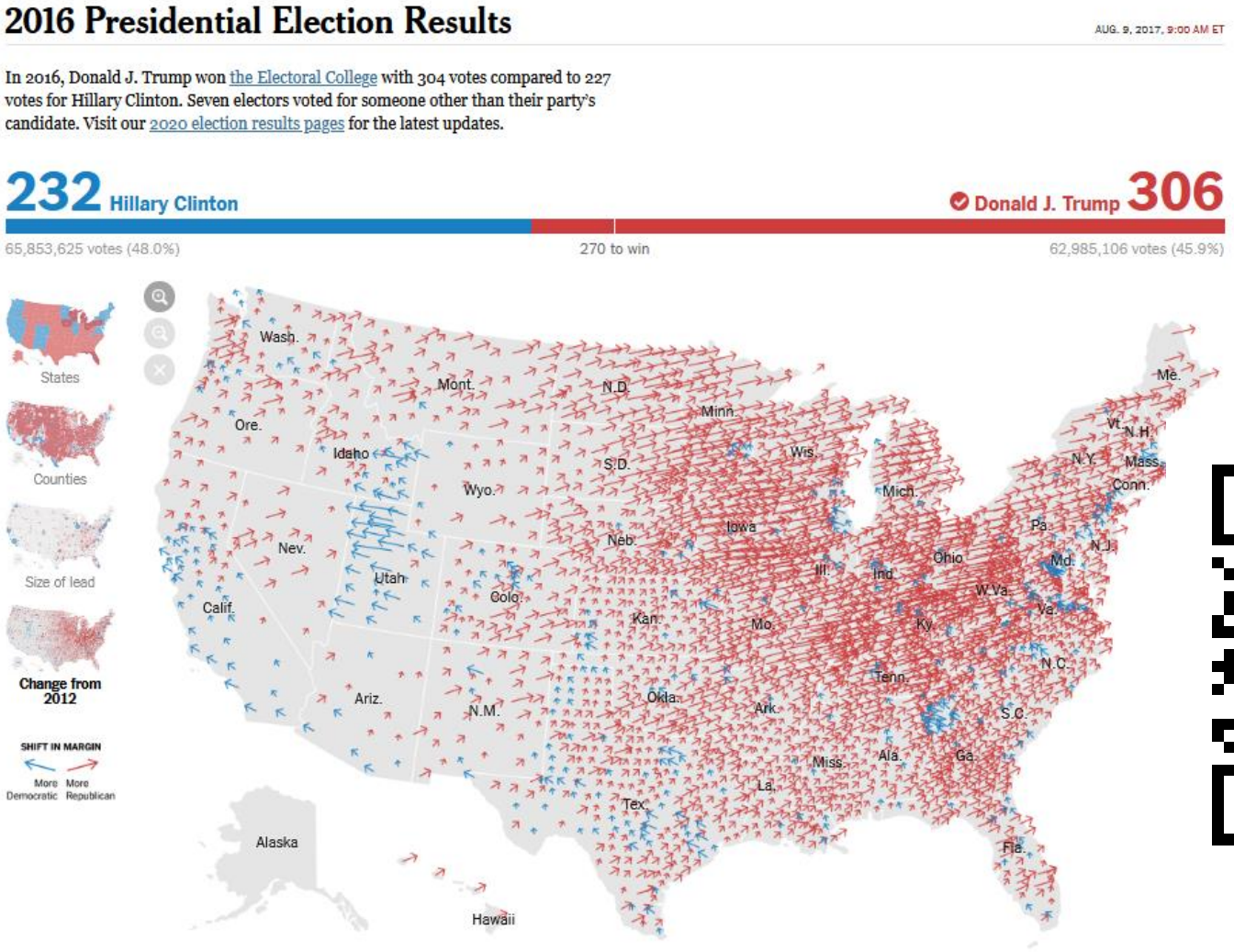
3 Data Project

Your Final Project

1. You can do election data projects to prepare the capstone course

2. Or you can choose any topics for your final project.

Exercise: Learn from classic projects



3 Data Project

Classic Projects

- Think about how to imitate the same plot:

Analysis

Visualization

Data Collection

Data cleaning

Data analysis

Data Presentation

Writing report

Data Sources

Data Cleaning

Report

3 Data Project

Classic Projects

Do some ~~google~~ research and answer the following questions in Moodle:

1. **Project aim:**
2. **Data sources and levels:**
3. **The skills or codes to clean the data in R:**
4. **The way to calculate to change from 2012-2016:**
5. **The way to create a similar plot in R:**

R

4 R and R-Studio

Any problems to install R and R-Studio?

4 R and R-Studio

I may create R-Studio Cloud for this course in the next week.