Good Practice on Data Management

Enoch Yi-Tung Chen

Department of Medical Epidemiology and Biostatistics, Karolinska Insitutet

09 Feb 2024

Download materials from:
https://enochytchen.com/talks/2024-dataman

About me

- Born and raised in Taipei, Taiwan
- MSc in PHS—Epidemiology (2018-2020)
- PhD student at KI MEB Biostatistics Group (2021-current)
- Key words: survival analysis, biostatistics, health economics modelling, chronic myeloid leukemia

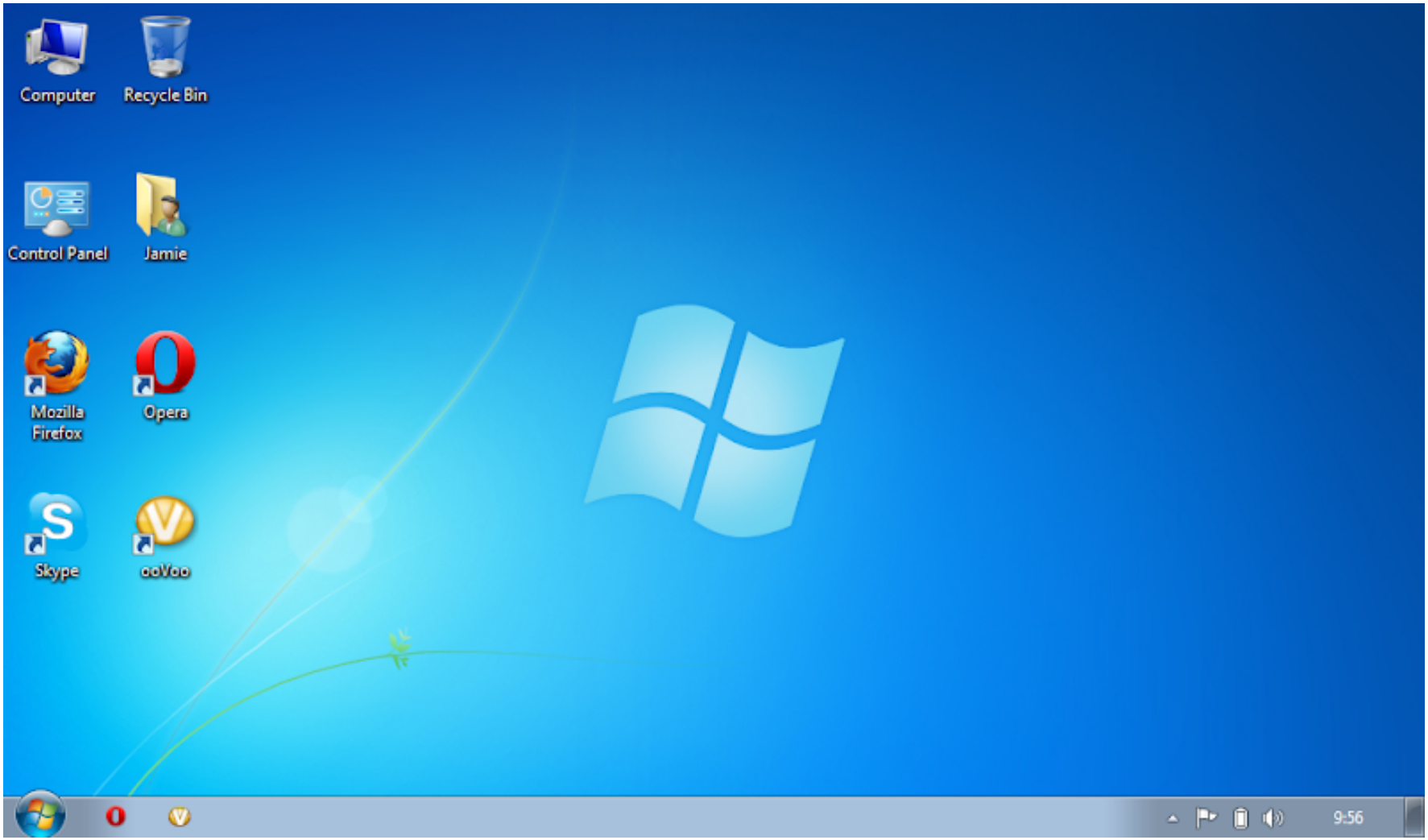
Where are you now?

Metimeter

Outline

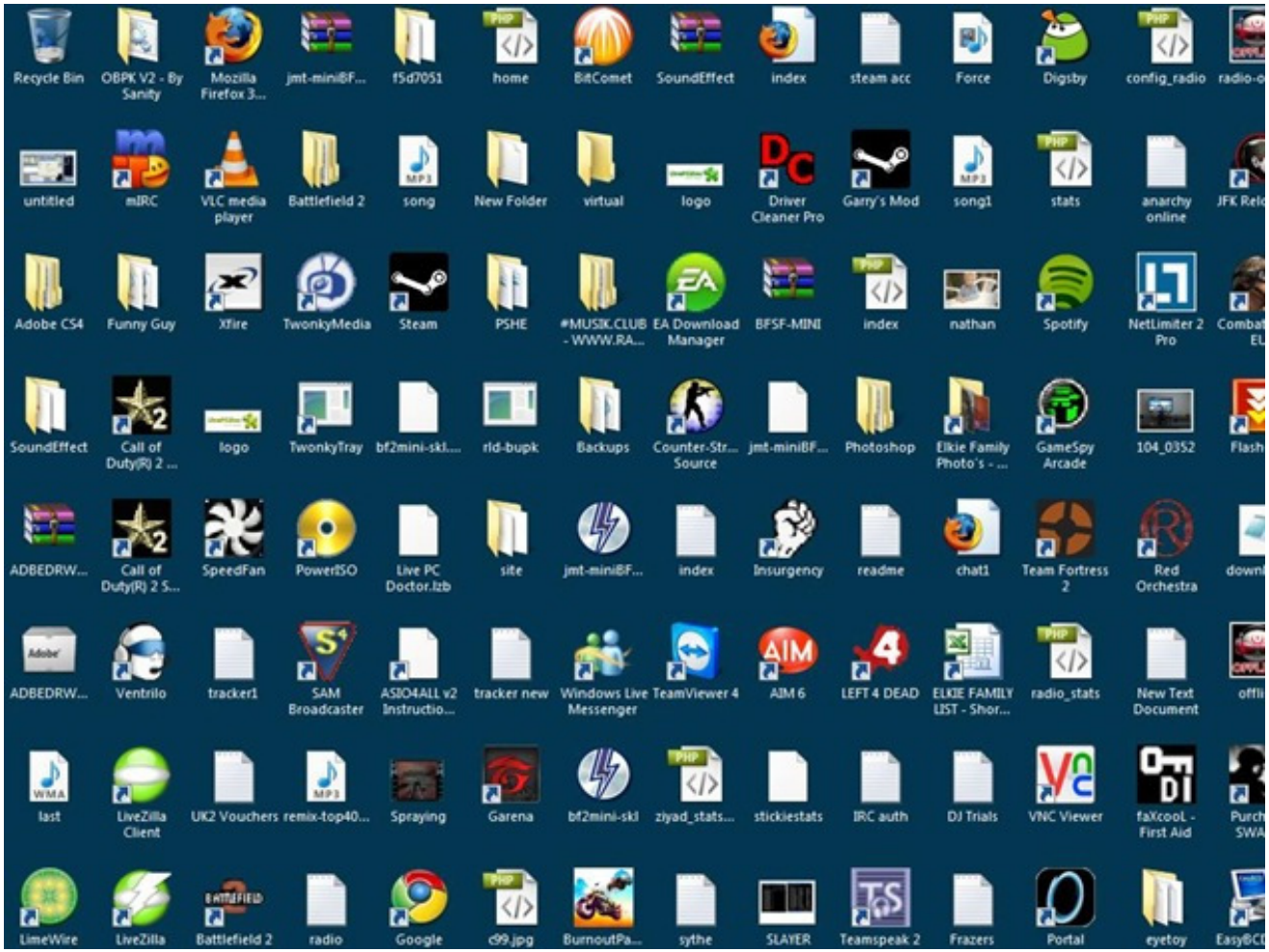
Aims of data management

In the beginning,



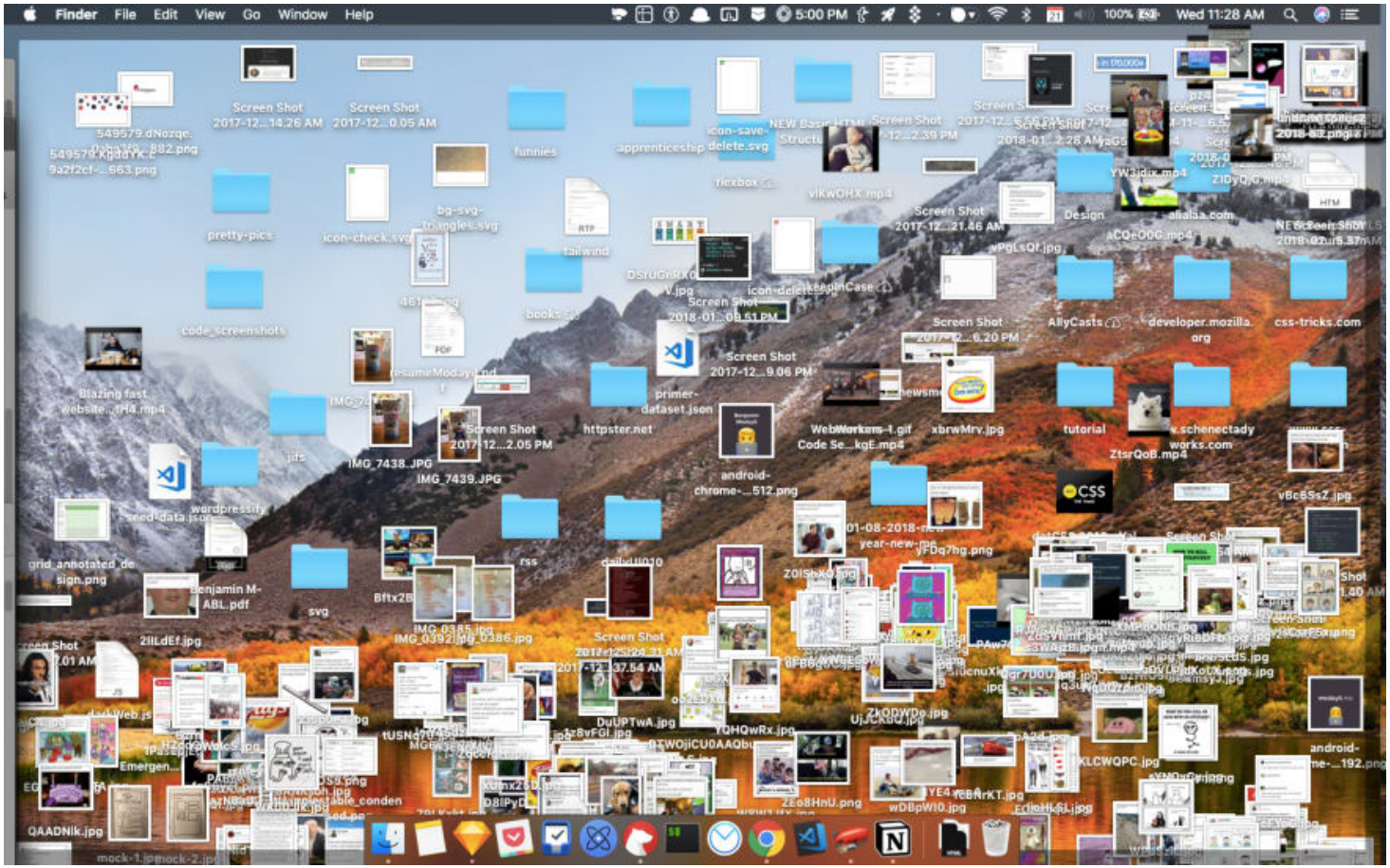
Aims of data management

On the half-way of the research,



Aims of data management

And eventually...



Aims of data management

Situations like:

- if you want to correct Table I, where is the do file for descriptive analysis?

Aims of data management

Situations like:

- if you want to correct Table I, where is the do file for descriptive analysis?
- if your supervisor says, "Please summarise how far you've gone in this project." You probably cannot just drop him/her your syntax.

Aims of data management

Situations like:

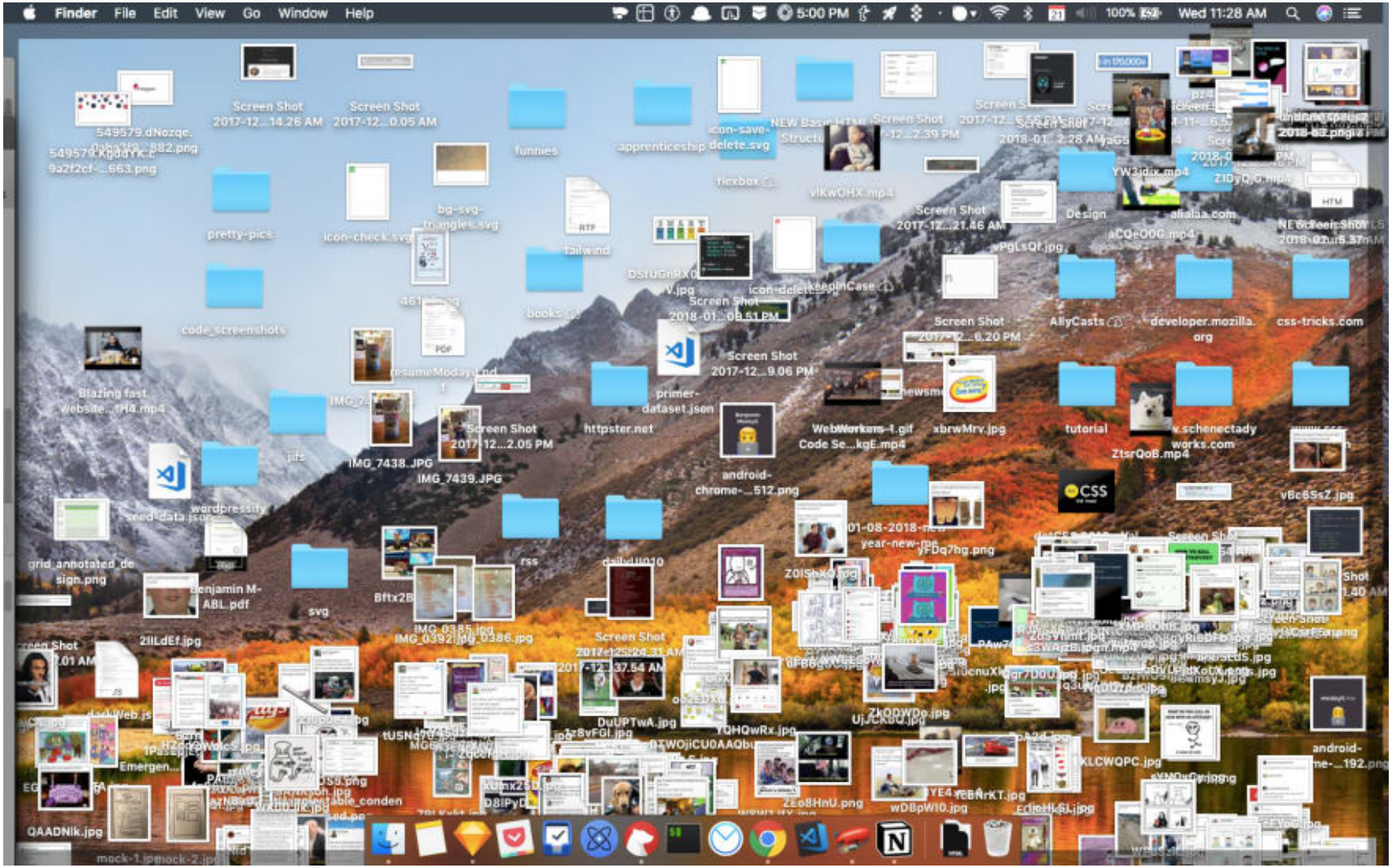
- if you want to correct Table I, where is the do file for descriptive analysis?
- if your supervisor says, "Please summarise how far you've gone in this project." You probably cannot just drop him/her your syntax.
- if your classmate asks you to teach her how to write a certain Stata code, but...where did you put it?

Aims of data management

Situations like:

- if you want to correct Table I, where is the do file for descriptive analysis?
- if your supervisor says, "Please summarise how far you've gone in this project." You probably cannot just drop him/her your syntax.
- if your classmate asks you to teach her how to write a certain Stata code, but...where did you put it?
- if your collaborator needs to take over, can he/she understand what you've completed?

Aims of data management



Aims of data management

To make your research

1. **Repeatable**: if you can use your data & code to generate them

Point 1 **Repeatable** is the minimum to aim for your master thesis.

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Aims of data management

To make your research

1. **Repeatable**: if you can use your data & code to generate them
2. **Reproducible**: if others can use your data & methods to generate them

Point 1 **Repeatable** is the minimum to aim for your master thesis.

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Aims of data management

To make your research

1. **Repeatable**: if you can use your data & code to generate them
2. **Reproducible**: if others can use your data & methods to generate them
3. **Replicable**: if others can use their data to generate compatible results

Point 1 **Repeatable** is the minimum to aim for your master thesis.

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Benefits of good data management

Reasons for implementing good data management practice

1. Scientifically (be a serious scientist)

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Benefits of good data management

Reasons for implementing good data management practice

1. Scientifically (be a serious scientist)
2. Legally (deal with sensitive human data)

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Benefits of good data management

Reasons for implementing good data management practice

1. Scientifically (be a serious scientist)
2. Legally (deal with sensitive human data)
3. Morally (hold your accountability)

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Benefits of good data management

Reasons for implementing good data management practice

1. Scientifically (be a serious scientist)
2. Legally (deal with sensitive human data)
3. Morally (hold your accountability)
4. Psychologically (be in peace before going to bed)

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Benefits of good data management

Reasons for implementing good data management practice

1. Scientifically (be a serious scientist)
2. Legally (deal with sensitive human data)
3. Morally (hold your accountability)
4. Psychologically (be in peace before going to bed)
5. Spiritually (who doesn't like "clean and nice"?)

Adapted from: Alexander Ploner. Data management: a brief overview. 2023.

Good folder structure

The core elements of folders are listed below:

- Data
- Documents
- Log
- Output
- Program

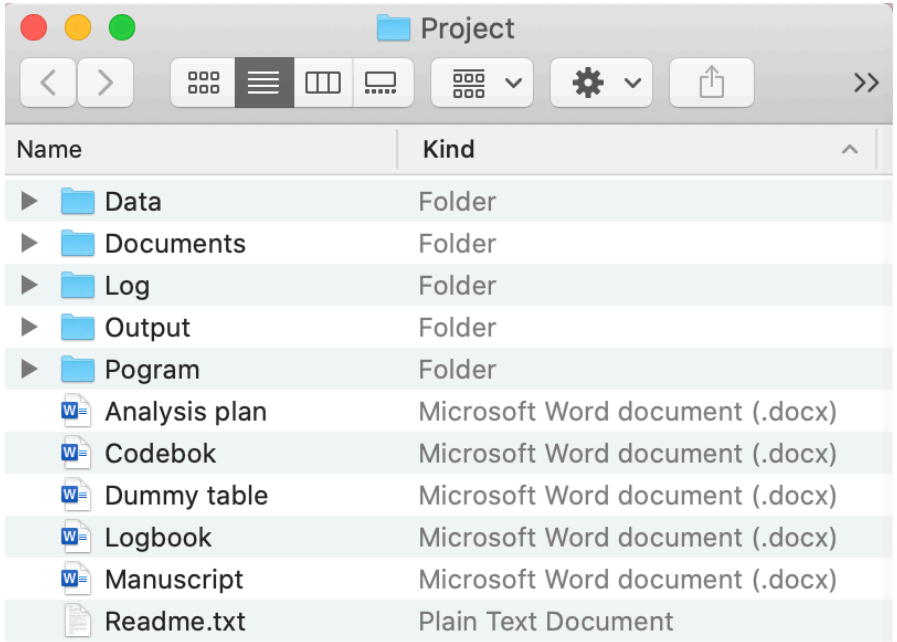


Figure: Project folder structure.

Good documents

Besides good folder structure, also consider keeping good documents

- Analysis plan
- Codebook
- Dummy table
- Logbook
- Manuscript

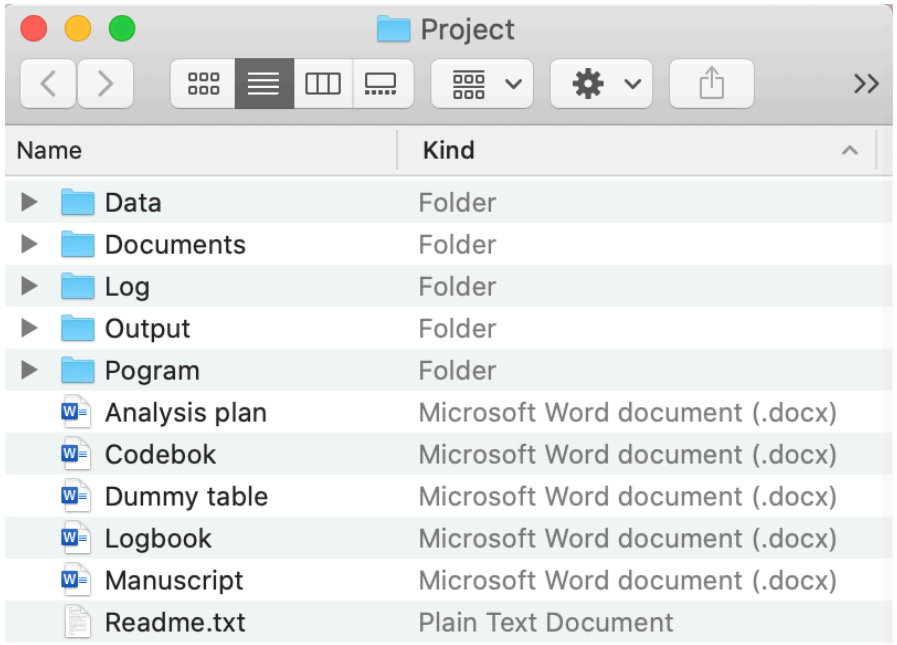


Figure: Project folder structure.

[3] can be included in analysis plan as well

Good documents

- You should illustrate how to use these documents/folders in the Readme.txt.
- A good Readme.txt is a good tourist guide in this project folder.

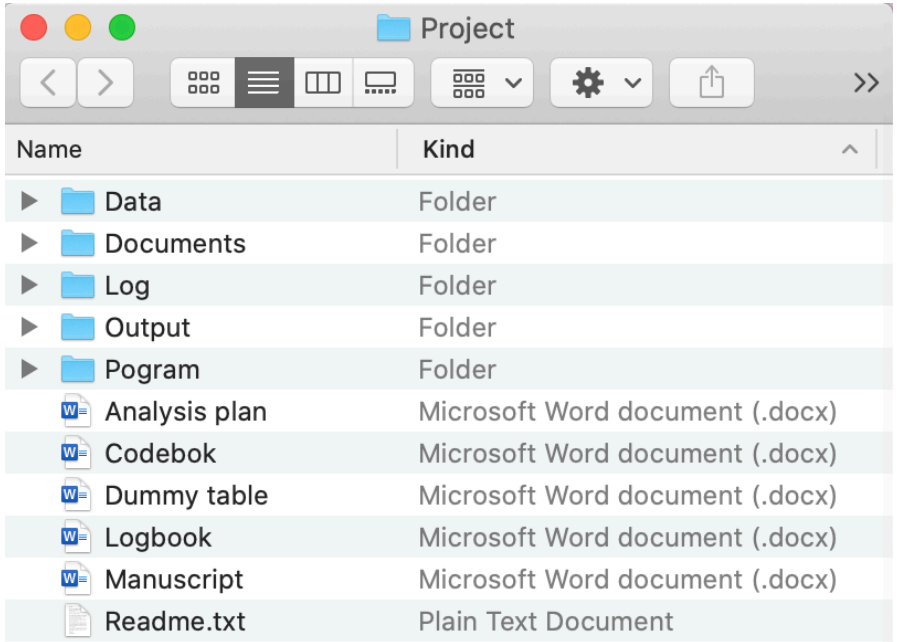


Figure: Project folder structure.

Exercise I: Create your project folder

1. Download the templates for analysis plan, codebook, README, logbook from: `https://enochytchen.com/talks/2024-dataman`
2. Create your own project folder

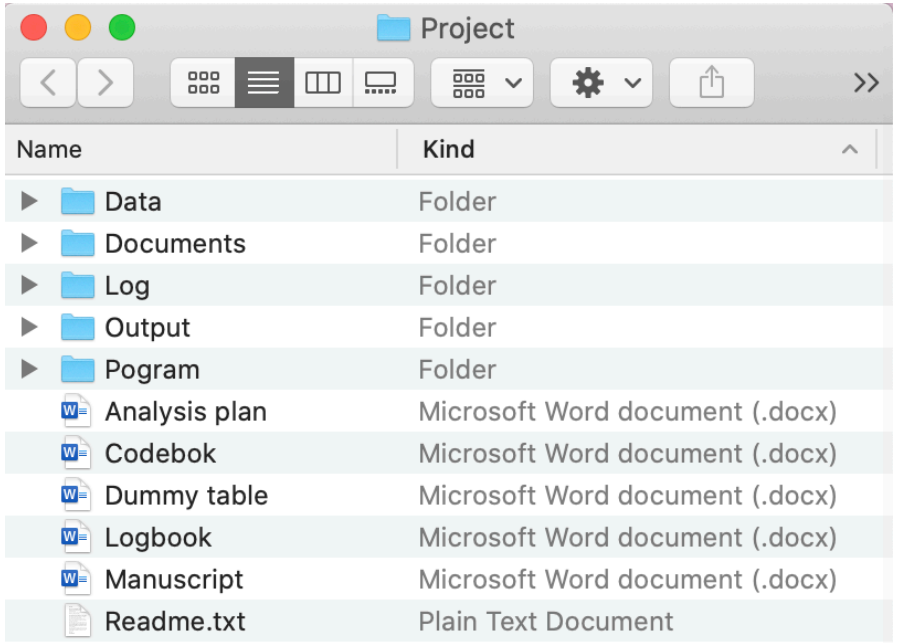


Figure: Project folder structure.

Good habit on coding

- **log on**
- Filename
- Study
- Created
- Updated
- Purpose
- Note

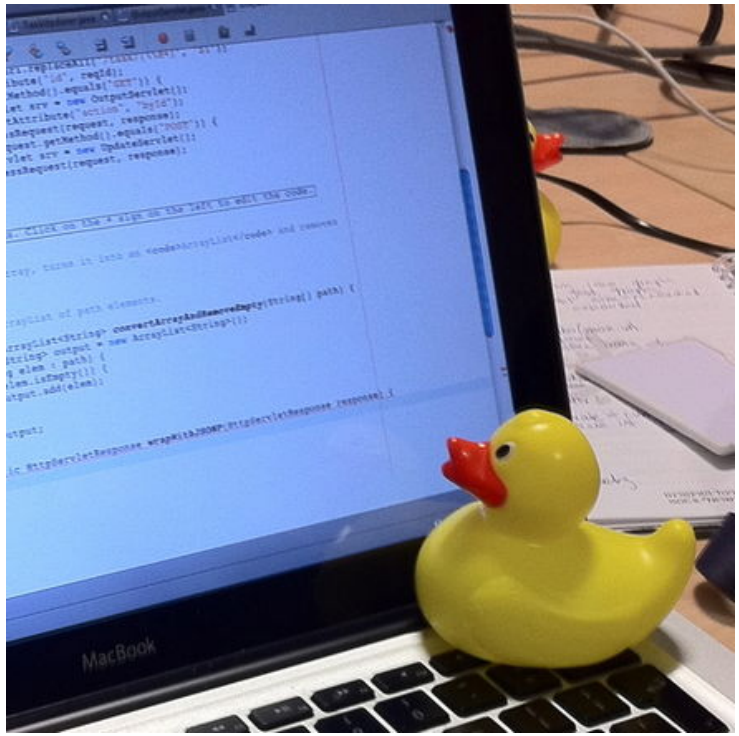
- **Program**

- log close

```
local todaydate: di %tdCYND date(c(current_date),"DMY")
capture log close
log using "your log folder route\do file name_`todaydate'.log",

/*==================================================================
Filename: make_analysis_data.do
Study:    Colon cancer patient survival, Sweden, 2010-2015

Created:  20201015 Enoch Yi-Tung Chen
Updated:  20201017 Enoch Yi-Tung Chen

Purpose:  Conduct data clearance for the project
Note:     Well, this is just an example.
==================================================================
// Start of Stata code



// End of Stata code

log close
```

Good habit on coding

- Talk to yourself what you are doing.
- Rubber duck debugging



Exercise II: Create your do-file header

1. Use the templates from Project/Program downloaded from Templates (`https://enochytchen.com/talks/2024-dataman`)
2. Replace the content with one of your own do/R files.

```
local todaydate: di %tdCYND date(c(current_date),"DMY")
capture log close
log using "your log folder route\do file name_`todaydate'.log", replace

/*===================================================================
Filename: make_analysis_data.do
Study:    Colon cancer patient survival, Sweden, 2010-2015

Created:  20201015 Enoch Yi-Tung Chen
Updated:  20201017 Enoch Yi-Tung Chen

Purpose:  Conduct data clearance for the project
Note:     Well, this is just an example.
===================================================================*/
// Start of Stata code




// End of Stata code

log close
```

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)
   - For binomial variables, $= 1$ implies yes, and $= 0$ implies no. (E.g., name your sex variable as female/male 1/0)

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)
   - For binomial variables, $= 1$ implies yes, and $= 0$ implies no. (E.g., name your sex variable as female/male 1/0)
   - Label your variables, please!

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)
   - For binomial variables, $= 1$ implies yes, and $= 0$ implies no. (E.g., name your sex variable as female/male 1/0)
   - Label your variables, please!
3. Same names for linking files (.do .r .sas $\rightarrow$ .log $\rightarrow$ .doc)

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)
   - For binomial variables, $= 1$ implies yes, and $= 0$ implies no. (E.g., name your sex variable as female/male 1/0)
   - Label your variables, please!
3. Same names for linking files (.do .r .sas $\rightarrow$ .log $\rightarrow$ .doc)
4. Don't replace the original files or variables. Keep a folder called "old" or at least the log files.

Other do's and don'ts

1. Use a shared drive/project server. Real-life stories (theft/coffee) (Required to do that because of data privacy.)
2. Give appropriate names to your files and variables.
   - No stupid names, such as new1, new2, new3, final1, final2, final3, latest1
   - No space, special character, dots (E.g., variable names should NOT start with numbers in Stata, underscore in R.)
   - For binomial variables, $= 1$ implies yes, and $= 0$ implies no. (E.g., name your sex variable as female/male 1/0)
   - Label your variables, please!
3. Same names for linking files (.do .r .sas $\rightarrow$ .log $\rightarrow$ .doc)
4. Don't replace the original files or variables. Keep a folder called "old" or at least the log files.
5. Don't edit the data directly. Please write syntax.

Wrap it up

- In summary, a good data management
  1. aims for 3R (Repeatable, Reproducible, Replicable)
  2. with GOOD folder structure, GOOD documents, and GOOD habits
- Further readings: The Department of Medical Epidemiology and Biostatistics, Karolinska Institutet. MEB Guidelines for Documentation and Archiving Version 6. 2018.

Wrap it up

- In summary, a good data management
  1. aims for 3R (Repeatable, Reproducible, Replicable)
  2. with GOOD folder structure, GOOD documents, and GOOD habits
- Further readings: The Department of Medical Epidemiology and Biostatistics, Karolinska Institutet. MEB Guidelines for Documentation and Archiving Version 6. 2018.
- All the files in this seminar can be found at `https://enochytchen.com/talks/2024-dataman`

Good data management saves your life.

Happy thesis life $< 3 < 3 < 3$