L2: BECOMING SELF-SUFFICIENT IN STATA

Getting started with Stata

Angela Ambroz

May 2015

Today

- Homework review and questions
- Writing our first .do file
- Commands, commands, commands
- Data cleaning
- The dreaded `error message`

Review

- Homework 1 solution video is up: www.angelaambroz.com/stata.html
- Lecture 1's key messages:
  - .do files are your friend
  - Stata syntax is similar to programming
  - The only command you need to know is `help`
- Questions?

Writing our first .do file

- Today, let's walk through creating our first .do file
- We'll learn **best practices** on how to organize things
- We'll learn the basics of:
  1. Importing data
  2. Cleaning data
  3. Exploring data
  4. Outputting basic summary statistics
- Once this is done, the world is your oyster!

The .do file lifestyle and philosophy

- Before we begin, let's talk **art and beauty**
- Your .do file is both the analysis **and the presentation of the analysis**
- Making it clear – even **beautiful** – will save you lots of confusion later

Think: Who will read my .do file?

An example of something beautiful: "Conversion on the Way to Damascus", by Caravaggio (1601)

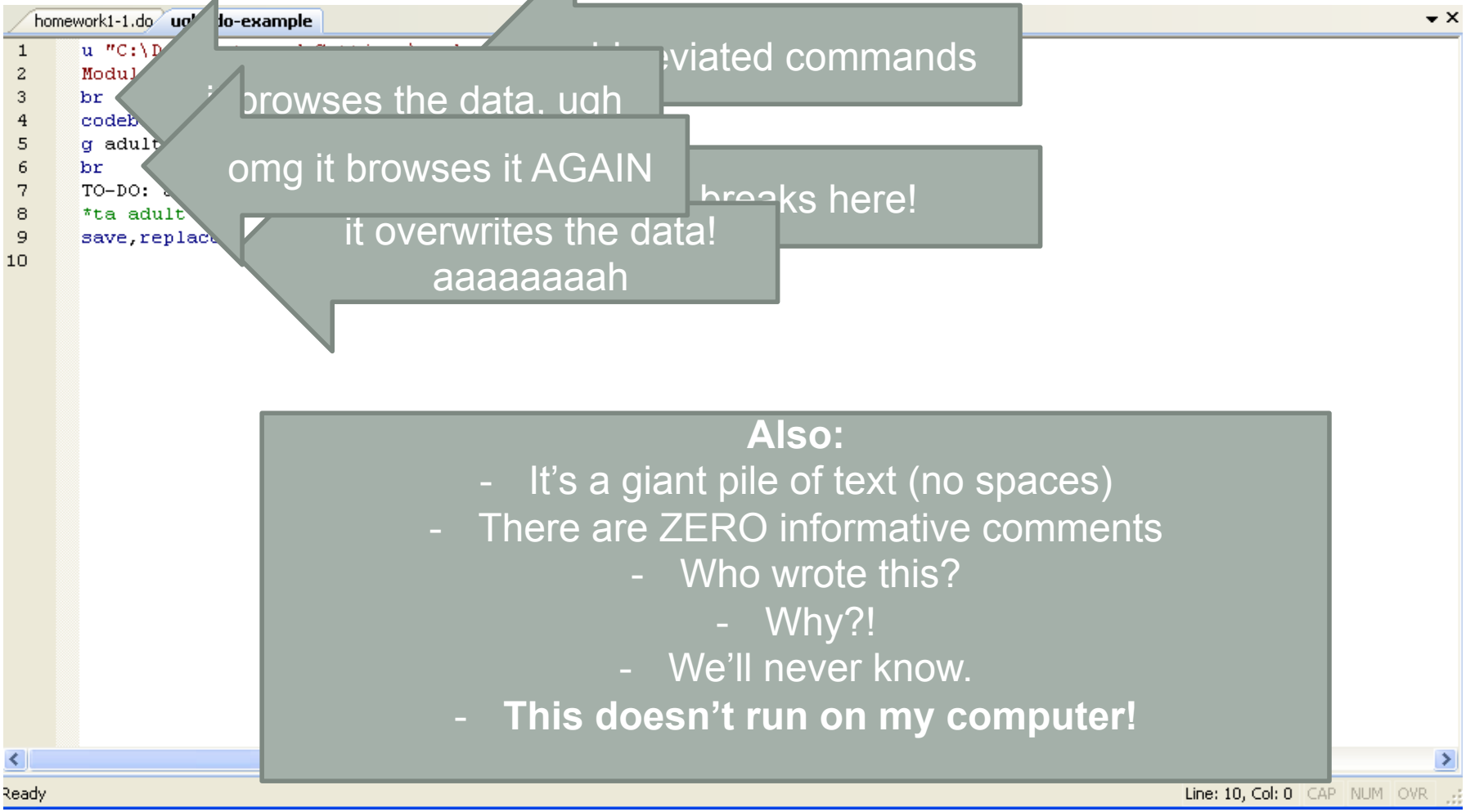
The .do file lifestyle and philosophy

- Before we begin, let's talk **art and beauty**
- Your .do file is both the analysis **and the presentation of the analysis**
- Making it clear – even **beautiful** – will save you lots of confusion later

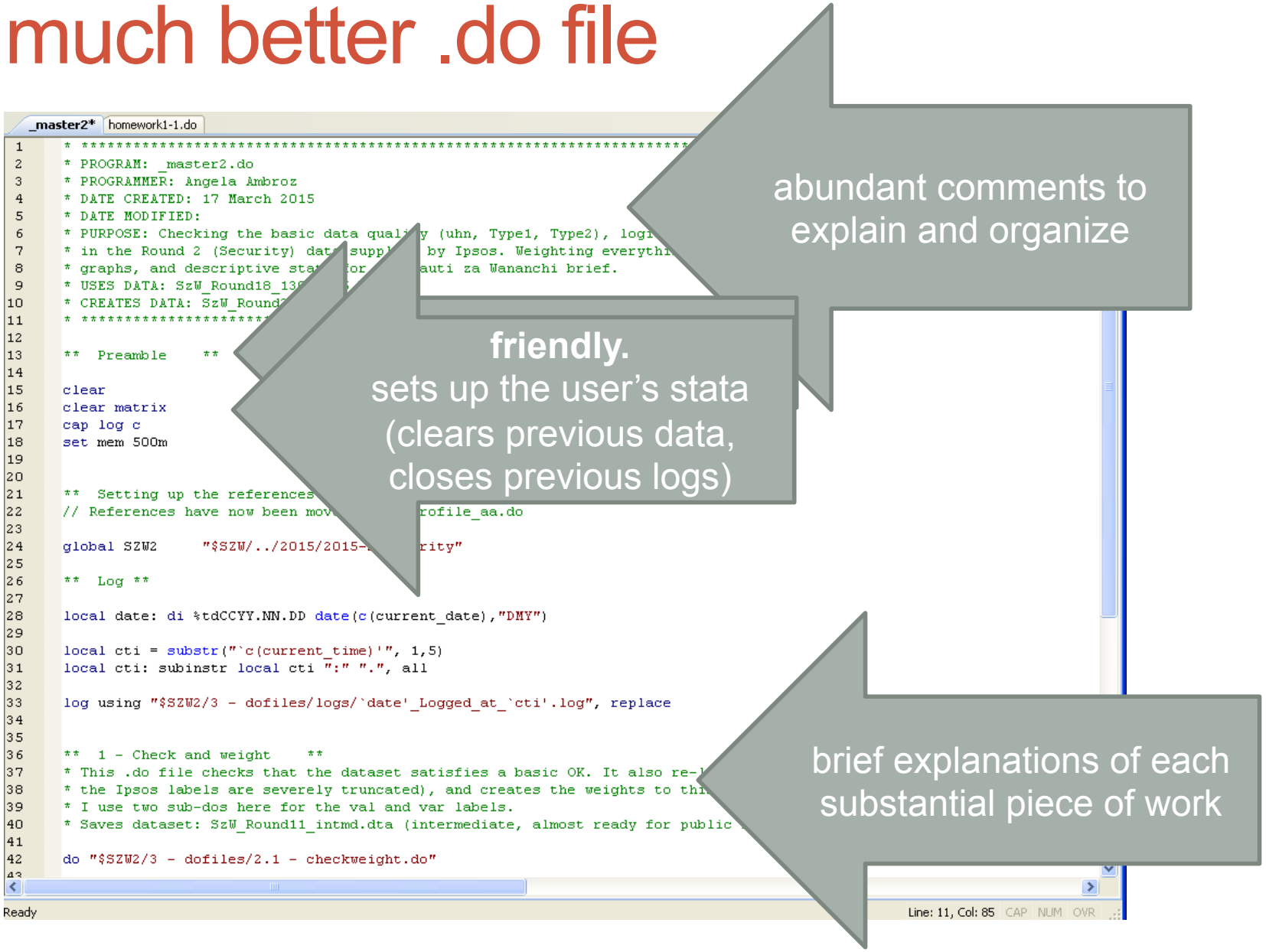
Think: Will my .do file run on ANY computer, as-is?

An example of something beautiful: "Conversion on the Way to Damascus", by Caravaggio (1601)

The worst .do file I could think of

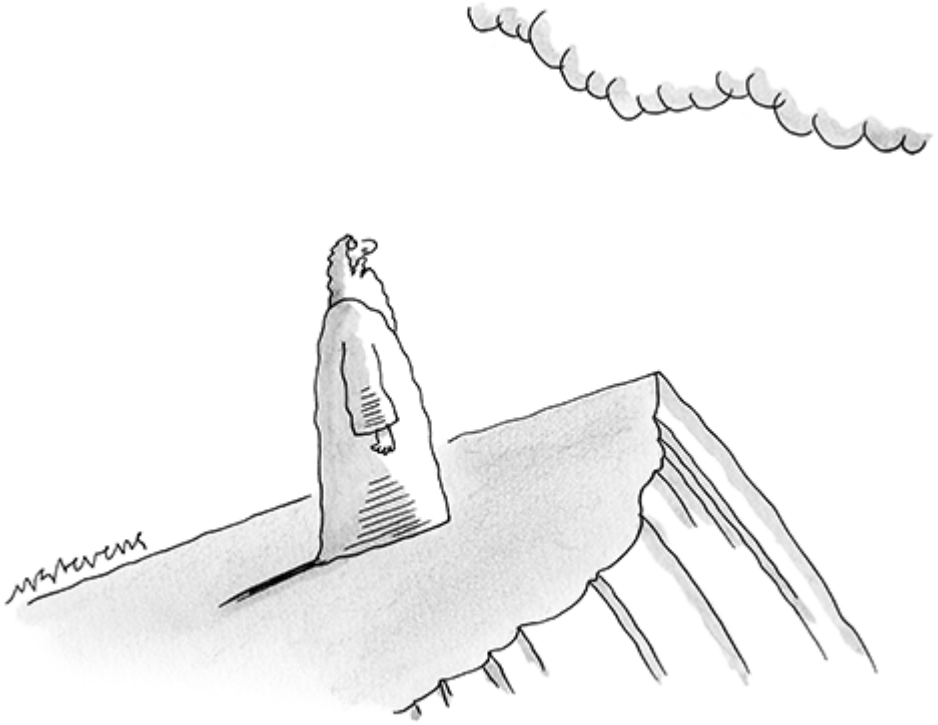


A much better .do file



Good .do file structure

- LOTS of **comments**
- LOTS of **space**
- Include information about the .do file's author, origin date, purpose, input and output data
- Make sure it can run on other people's computers
- If at all possible, do your analysis and ***do not overwrite your data with it***



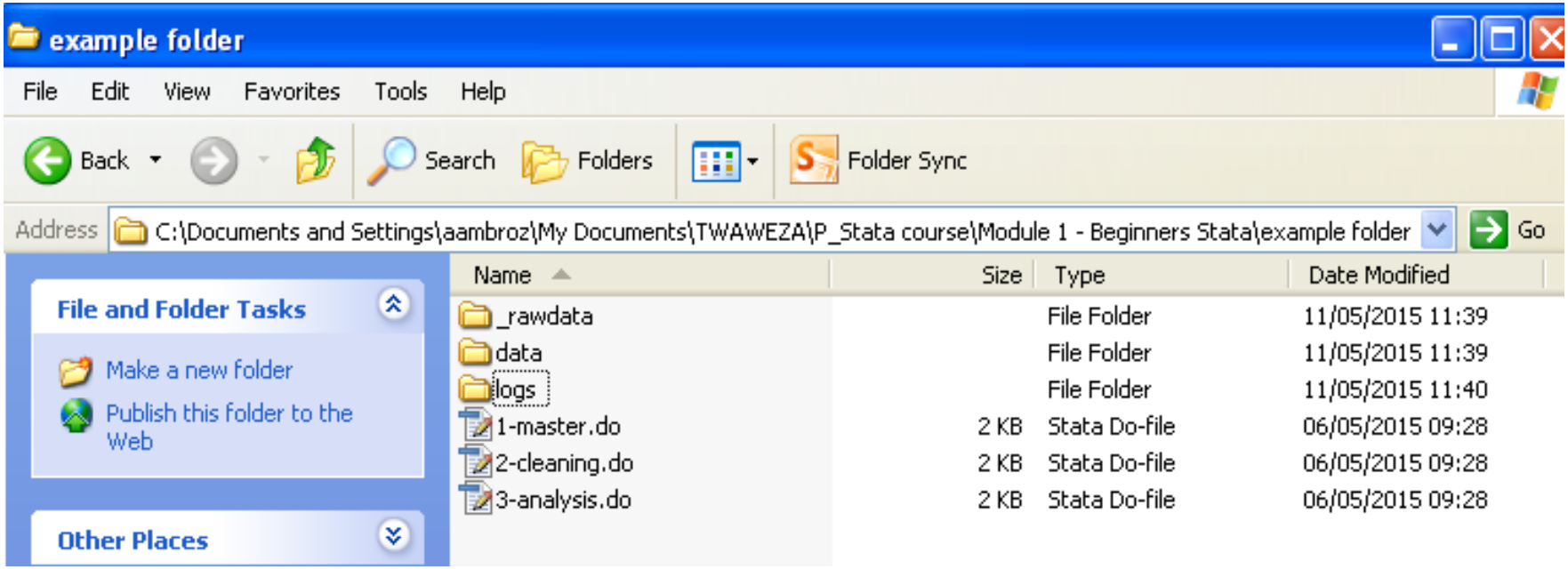
"They broke all the Commandments. Can they have some more?"

Good .do file structure

Be considerate to your two main audiences:

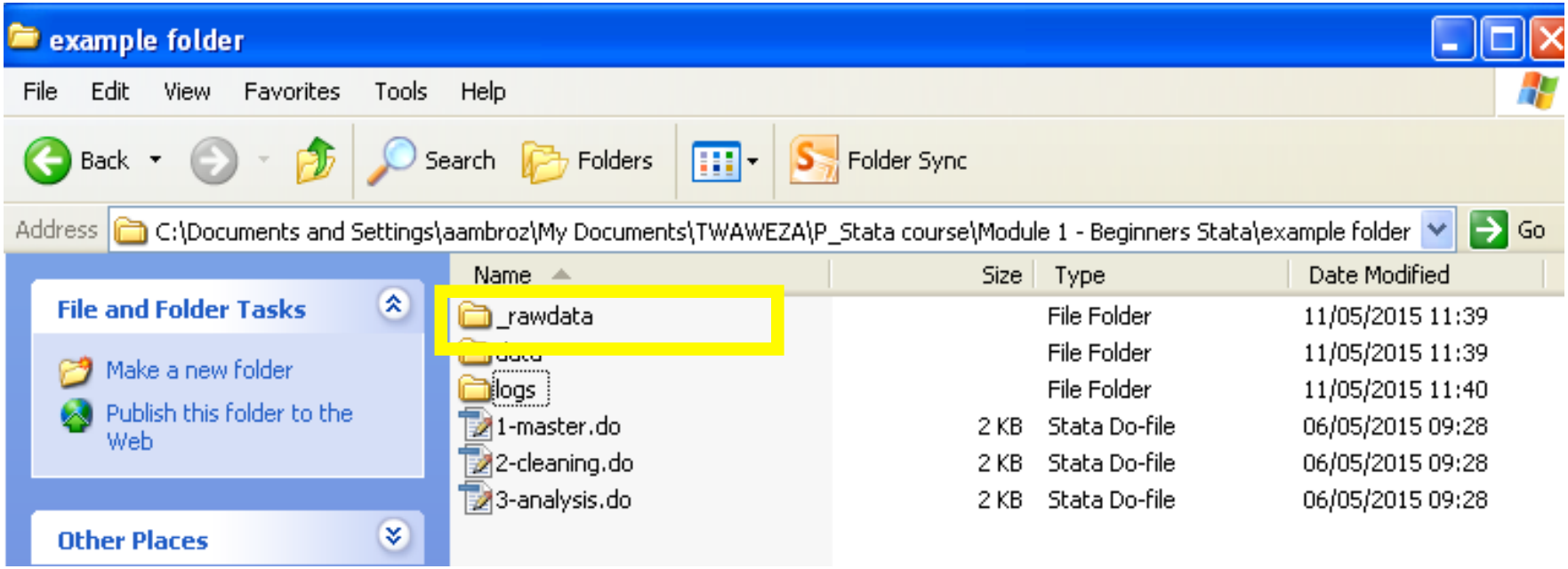
1. A forgetful you in the future
2. Other people with different machines

Good folder structure



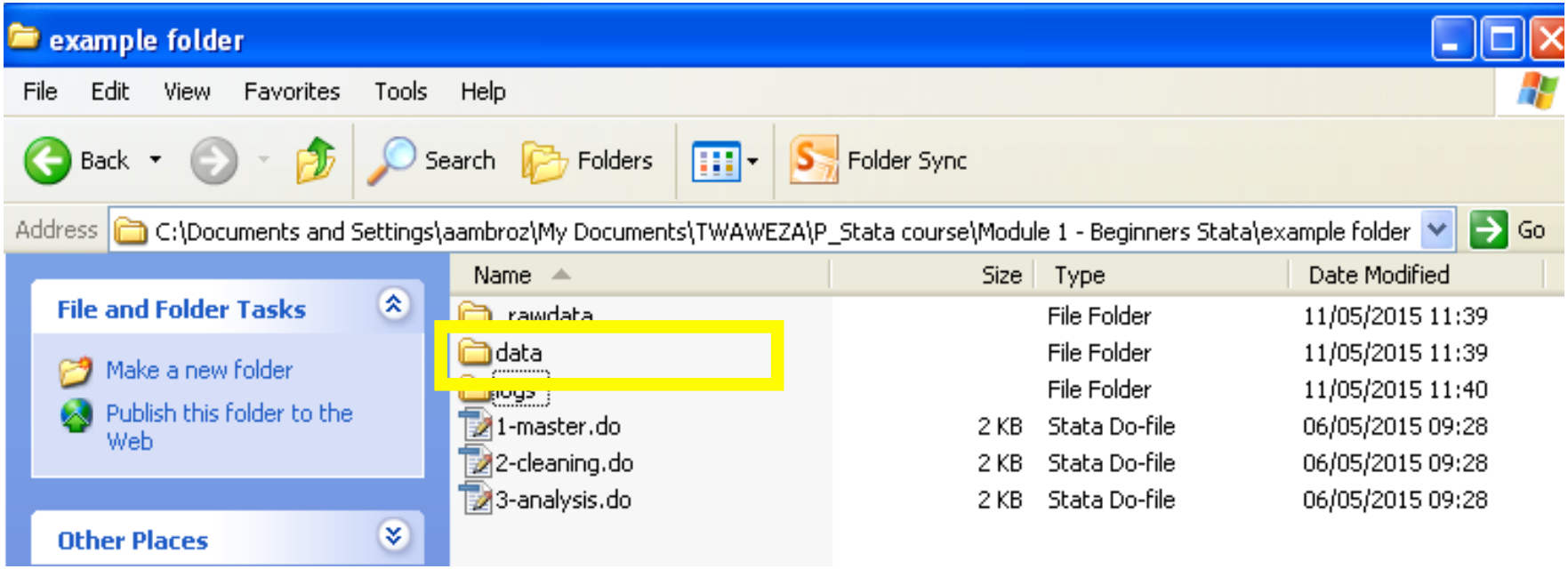
Organizing your folder well will also save you **lots of time** in the future!

Good folder structure



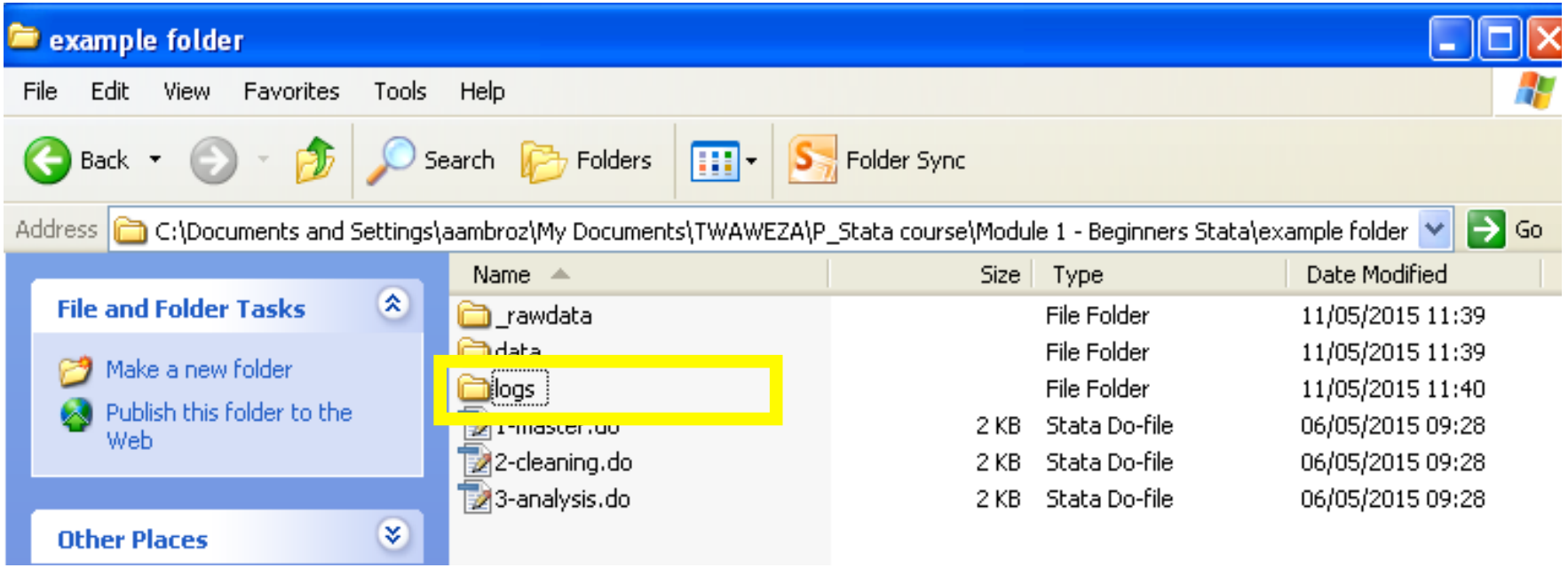
A "raw data" folder – This contains a copy of the data, as you received it from the survey company. This data is never altered.

Good folder structure



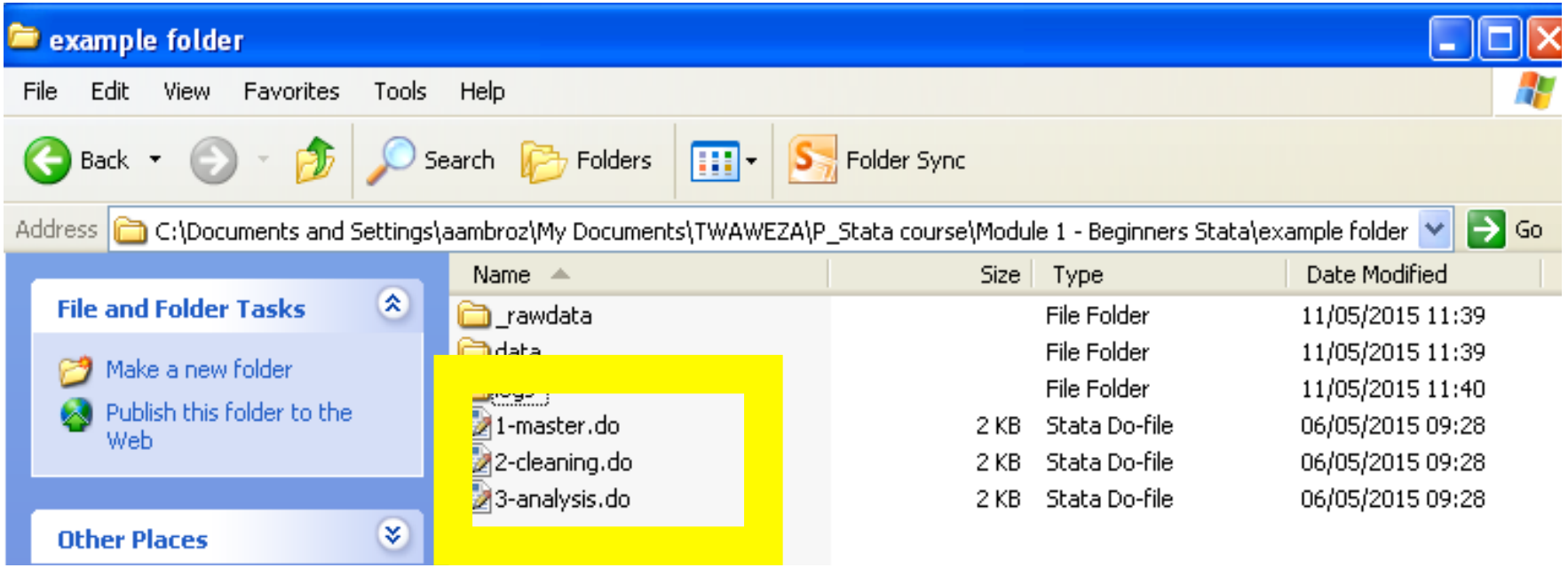
A "data" folder – This contains the data that you use for analysis (and may alter).

Good folder structure



A "logs" folder – Here, you keep logs of all of your Stata sessions. This is good to go back and see the input/output of your work.

Good folder structure



Your .do files. You can number them in the order that they are meant to be called. A "master" .do file can include lots of information about the project and analysis.

Sketch our .do file on paper

GOAL: Conduct summary statistics on the Sauti za Wananchi constitutional round data (`szw-constitution.dta`). Put Angela out of a job.

TO DO

1. Import the Sauti data.
2. Make sure it's clean.
3. Find out the percentage of people that are:
   - Aware of the constitutional draft process.
   - Planning to vote for the new constitution.

Making a .do file

TO DO

1. Import the Sauti data.
2. Make sure it's clean.
3. Find out the percentage of people that are:
   - Aware of the constitutional draft process.
   - Planning to vote for the new constitution.

1. Importing data

- To import Stata datasets (`.dta`):

```
use "file.dta", clear
```

- To import Excel files (`.xls`):

```
import excel "file.xls", clear
```

- To import Excel files (`.csv`):

```
insheet using "file.csv", clear
```

File paths

- **Note**: You need to explicitly tell Stata in which folder to look for your file.

Quick version:

```
1 use "C:/MyDocs/file.dta", clear
```

Better version:

```
1 cd "C:/MyDocuments/"
2 use "file.dta", clear
```

Best version:

```
1 global myfolder "C:/MyDocuments"
2 use "$myfolder/file.dta", clear
```

Making a .do file

TO DO

1. ~~Import the Sauti data.~~
2. Make sure it's clean.
3. Find out the percentage of people that are:
   - Aware of the constitutional draft process.
   - Planning to vote for the new constitution.

Checking the data out

- Some good commands to just get to know the data:

```
browse
describe
describe, short
summarize
codebook
```

Cleaning the data

- 70%* of all data “analysis” is actually just data cleaning
- Data cleaning means **preparing the data for analysis**
- This can include:
  - Converting strings to numerics (e.g. `“1,756”-->1756`)
  - Creating new variables based on old ones (e.g.
    ```
    avg_weekly_mobile_credit =
    avg_daily_mobile_credit * 7 )
    ```
  - Checking for any weird observations (duplicates, all missing, etc.)
  - Deciding how to treat outliers.
  - Whatever the data needs!
- Data cleaning is more **art** than **science**.

* 80% of all statistics are made up.

Cleaning the data: An example

```
. ta q3, m
```

Q3. Can you identify the remaining step(s) in the constitutional review proce	Freq.	Percent	Cum.
Don't know	106	7.58	7.58
Yes, Mentioned correct steps	339	24.23	31.81
Yes, But mentioned wrong steps	116	8.29	40.10
No	838	59.90	100.00
Total	1,399	100.00	

GOAL: Let's create a variable out of `q3` which just calculates whether a person knows the steps or not.

Cleaning the data: An example

```
. codebook q3

-------------------------------------------------------------------------------
q3                        Q3.    Can you identify the remaining step(s) in the constitutional review proce
-------------------------------------------------------------------------------

                  type:  numeric (int)
                 label:  q3

                 range:  [-888,3]                     units:  1
         unique values:  4                        missing .:  0/1399

            tabulation:  Freq.   Numeric  Label
                           106      -888  Don't know
                           339         1  Yes, Mentioned correct  steps
                           116         2  Yes, But mentioned wrong steps
                           838         3  No
```

```
generate knows_process = .
replace knows_process = 1 if q3==1
replace knows_process = 0 if q3!=1 & q3!=.
label def knowledge 1 "yes" 0 "no"
label val knows_process knowledge
```

Basic summary statistics

TO DO

1. ~~Import the Sauti data.~~
2. ~~Make sure it's clean.~~
3. Find out the percentage of people that are: aware of the constitutional draft process, how they would vote in the referendum.

`q1` – Awareness of the constitutional draft process

`q12` – How people plan to vote in the referendum

Many different ways to find frequency: we'll use the popular `tabulate`

Basic summary statistics

Input: `tab q1, m`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	Percent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

Basic summary statistics

Input: `tab q1, m`

Command: tabulate the frequencies of each answer for variable `q1`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	Percent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

Basic summary statistics

Input: `tab q1, m`

Option: tell me how many observations are missing

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	Percent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

Basic summary statistics

Input: `tab q1, m`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	Percent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

The variable's label

Basic summary statistics

Input: `tab q1, m`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	ercent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

Number of people in each category

Basic summary statistics

Input: `tab q1, m`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq	Percent	Cum.
Yes	1,07	76.70	76.70
No	32	23.30	100.00
Total	1,39	100.00	

Percentage of people in each category

Basic summary statistics

Input: `tab q1, m`

Output:

Q1. Are you aware that that the Constituent Assembly passed a final draft of	Freq.	Percent	Cum.
Yes	1,073	76.70	76.70
No	326	23.30	100.00
Total	1,399	100.00	

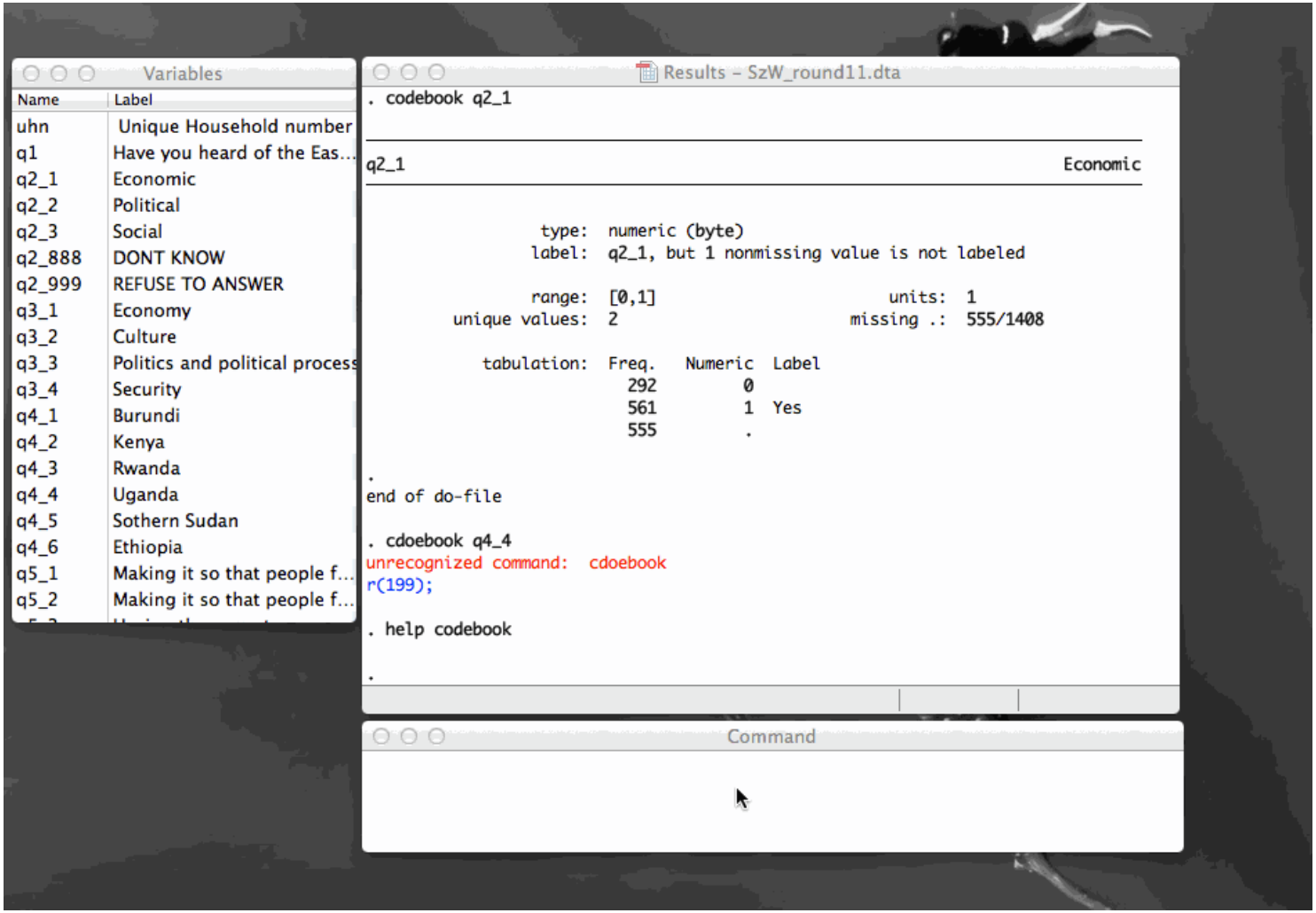
The cumulative frequency

Now what?

Options to output from Stata:

- Copy + paste
- Generate an Excel of summary statistics: `tabout`
- Generate a new dataset: `save, outsheet, export`
- Generate a Word document of estimation results: `outreg2`
- Create visualizations? `graph bar chart`

The error message :(

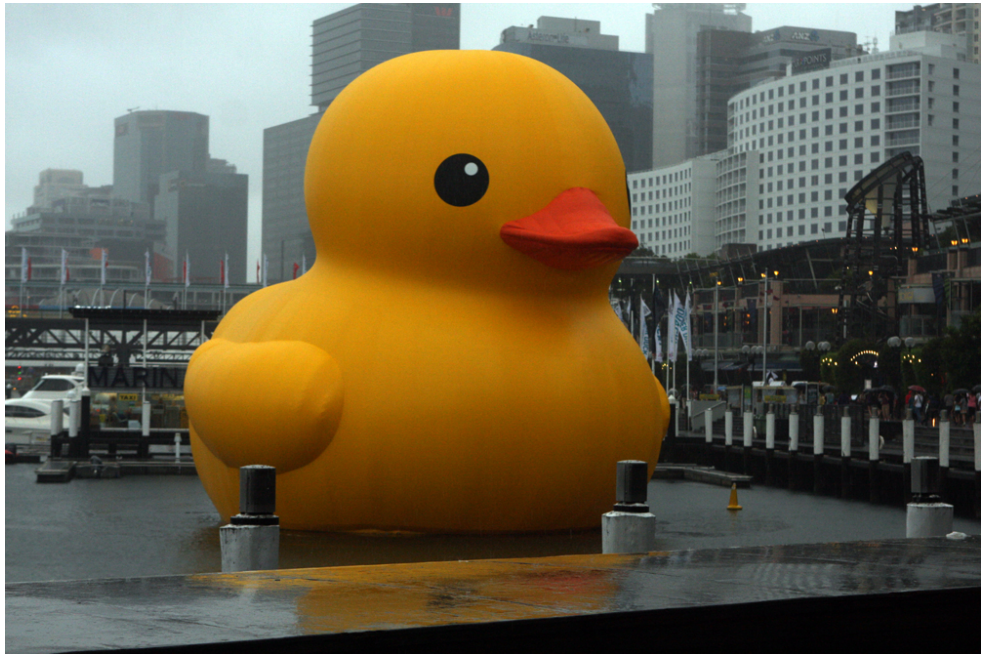


The error message :(

- Sometimes Stata gives you an error.
- Don't panic!
- Sometimes the error explains what went wrong (this is rare...)
- Error-checking:
  - Try to isolate your error: run each line, line by line
  - Run it from your `.do file` (highlight the line and `CTRL+D`)
  - Re-type it in the Command Editor
  - Is there a typo somewhere in your code?
  - Did you try `help`?
  - The rubber duck technique
  - Google!

The rubber duck technique

- Used in programming
- You explain what you're doing, out loud, to something at your desk (like a rubber duck)
- Often, as you explain, ***the solution reveals itself***



Here to help.

Homework

- Homework 2 – Write your first `.do file`.
- Instructions are in the Google form.
- Stuck? E-mail/Skype.
- Finished? E-mail your completed `.do file` to me. Fill out the Google form.
- **DEADLINE: COB Friday, 22 May 2015**