FOSE1025 — Scientific Computing

Week 6 Lecture 1: Towards Using Scripts for Reproducibility

Diego Mollá

FOSE1025 2022H1

Abstract

In this lecture we will have a very brief introduction to the use of scripts to store and manipulate data. The emphasis will be on how to use scripts for reproducibility, and we will focus in a particular environment: MATLAB.

Update March 24, 2022

Contents

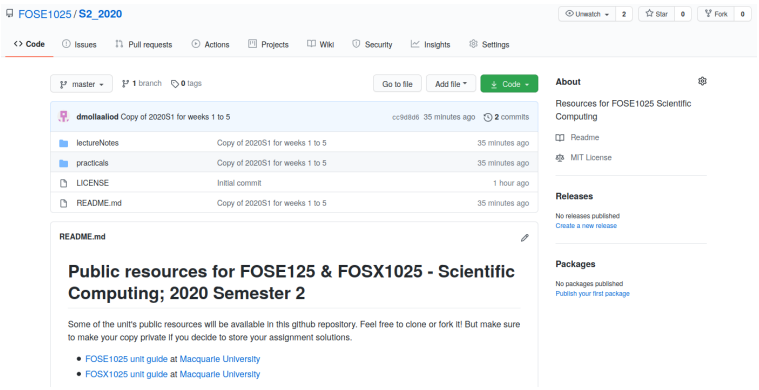
Reading

- These notes
- *https://au.mathworks.com/help/matlab/getting-started-with-matlab.html*
- *https://au.mathworks.com/videos/getting-started-with-matlab-1564521672719.html*

Announcements

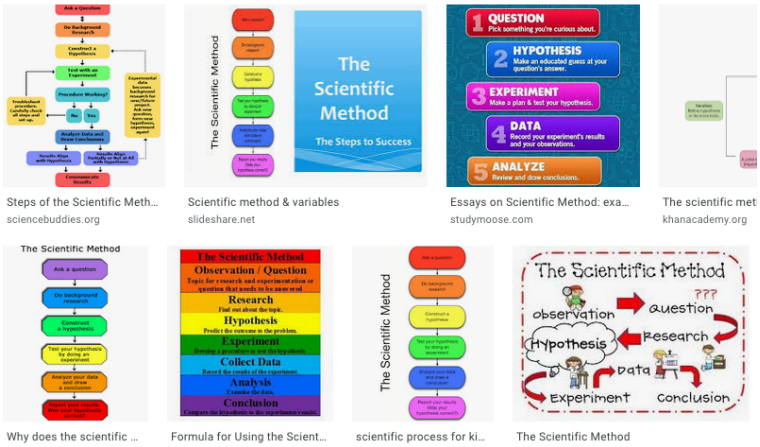
1. In-class test 2 this week (at your scheduled SGTA, or Friday 6-9pm for FOSX students).
2. Lecture notes in Echo360.
3. Material in github (next slide).

FOSE1025's public github page



1 Excel and MATLAB for Science

The Scientific Method



Some results of a Google image search with the words "scientific" and "method" — 1 April 2020.

Excel and MATLAB to Manage Data in Science

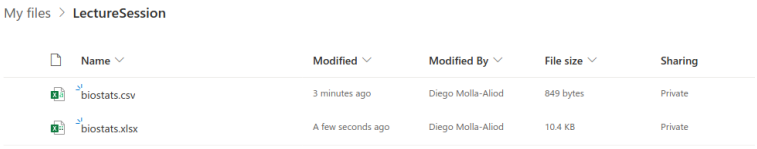
We are covering these aspects in FOSE1025:

- Represent data in Excel — Weeks 2 & 3.
- Represent data in MATLAB — Week 5.
- Explore data in Excel — Weeks 3 & 4.
- Visualise data in Excel — Week 5.
- *(you are here)*
- Import data from external files (e.g. CSV) — Week 6.
- MATLAB scripts for reproducibility — Week 6.
- Clean the data (Excel, MATLAB) — Week 7.
- Preprocess, transform the data (Excel, MATLAB) — Week 8.
- Analyse, summarise, interpret the data (MATLAB) — Week 9.
- Ethics of Data — Week 10.

Importing Data

Excel Files

- Excel saves files in a special format.
- The name of these files ends with .xlsx
- Other programs (e.g. MATLAB) can read Excel files.
- But there are many other formats!



Importing Data

Comma Separated Values

CSV — Comma Separated Values

- CSV is a very simple file format used by many applications.
- Each line represents a table row.
- Different values in the row are separated with a comma.
  - We say that comma is the *delimiter*.
  - Other common delimiters are: tabulator space (tab), semicolon (;).
  - In some file formats, the data fields are determined by the width.

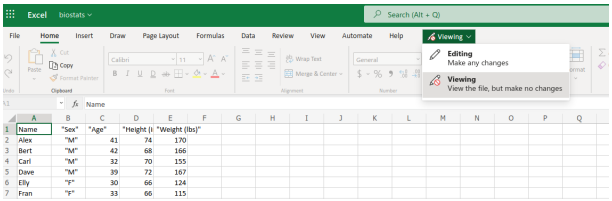
Example of a CSV File

biostats.csv from *https://people.sc.fsu.edu/~jburkardt/data/csv/csv.html*

```
"Name",    "Sex", "Age", "Height (in)", "Weight (lbs)"
"Alex",      "M",    41,          74,         170
"Bert",      "M",    42,          68,         166
"Carl",      "M",    32,          70,         155
"Dave",      "M",    39,          72,         167
"Elly",      "F",    30,          66,         124
"Fran",      "F",    33,          66,         115
"Gwen",      "F",    26,          64,         121
"Hank",      "M",    30,          71,         158
"Ivan",      "M",    53,          72,         175
"Jake",      "M",    32,          69,         143
"Kate",      "F",    47,          69,         139
"Luke",      "M",    34,          72,         163
"Myra",      "F",    23,          62,          98
"Neil",      "M",    36,          75,         160
"Omar",      "M",    38,          70,         145
"Page",      "F",    31,          67,         135
"Quin",      "M",    29,          71,         176
"Ruth",      "F",    28,          65,         131
```

Importing a CSV File into Excel Online
The easy option: upload & double click

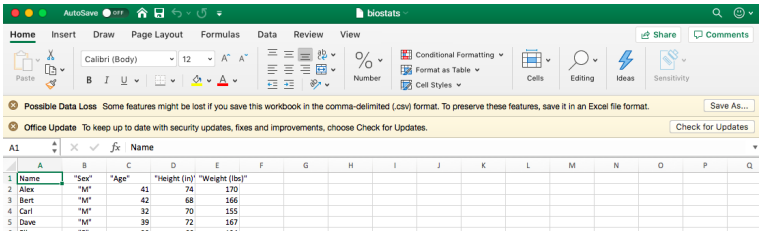
1. Upload the file to OneDrive.
2. Click on the file.
   - Excel Online will open the file in viewing mode.
3. Set the viewing mode to *Editing*.
   - Excel Online will create a copy of the file and save it as an Excel (.xlsx) file.



Importing a CSV File into Excel Online
Option 2: copy & paste

1. Double click on the file.
   - The file will open in your desktop.
   - If you have Excel installed, Excel will open the file.
2. Create a blank workbook in Excel Online.
3. Copy and paste from the desktop application to Excel Online.
4. Save the file.

Careful if you double-click on a CSV file in your desktop!



- If you double-click on a CSV file, Excel will open the file.
- But the file opened is a CSV file, not an Excel (.xlsx) file!
  - Read the warning that you get if you double-click on the CSV file.
- There are many things that you cannot save in a CSV file.
  - Formulas, formatting, charts, etc.

Tables in Excel

- Each row indicates a data sample.
- Each column indicates a type of data.
  - Number, string, date, etc.
  - Categorical data: when there is a pre-determined set of values (more on this in future lectures).

	A	B	C	D	E	F
1	Name	"Sex"	"Age"	"Height (i	"Weight (lbs)"	
2	Alex	"M"	41	74	170	
3	Bert	"M"	42	68	166	
4	Carl	"M"	32	70	155	
5	Dave	"M"	39	72	167	
6	Elly	"F"	30	66	124	
7	Fran	"F"	33	66	115	
8	Gwen	"F"	26	64	121	
9	Hank	"M"	30	71	158	
10	Ivan	"M"	53	72	175	
11	Jake	"M"	32	69	143	
12	Kate	"F"	47	69	139	
13	Luke	"M"	34	72	163	
14	Myra	"F"	23	62	98	
15	Neil	"M"	36	75	160	
16	Omar	"M"	38	70	145	
17	Page	"F"	31	67	135	
18	Quin	"M"	29	71	176	
19	Ruth	"F"	28	65	131	

Question: What are the data types of each column?

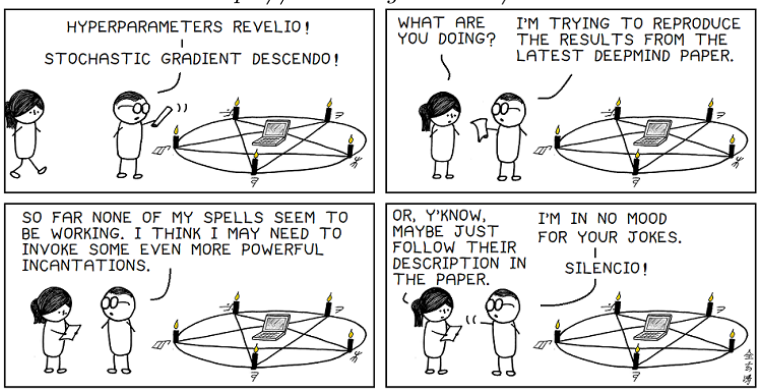
It is not easy to determine the type of data stored in an Excel cell. This is a design decision made by the developers of Excel. If we don't know the type of data stored, sometimes Excel will do something unexpected with the data. We will see examples of this in future lectures, especially when we look into how to store and operate with dates and times.

2 Scripts for Reproducibility

The Problem with Reproducibility

It can be difficult to write clearly enough to allow reproducibility.

https://abstrusegoose.com/588



Reproducibility in Science

- When you conduct science, you need to make sure that others can reproduce what you did.
  - If others can reproduce what you did, then your claims are more likely to be taken as valid.

- Reproducibility means that someone else should be able to do the same as you did by following your instructions.
- When the experiments are performed with computers, reproducibility can mean one of two:
  1. "I can re-implement what you did after I read your report."
  2. "I can run the code that you wrote."
- The employability modules ("Achiever" and "Communicator") touch item 1.
- Here we will touch item 2.

Scripting Languages

- Scripting languages are programming languages designed for *rapid prototyping*.

  ⇒ These languages make it easy to quickly write and execute a program.
- Scripting languages are normally *interpreted languages*.

  ⇒ This means that you can execute instructions one by one using a *run time environment*.

Example: Running Scripts in MATLAB Online

- In this demonstration, we will run MATLAB code *in the cloud*.
- We use a web browser to interact with the runtime.
- Can be done with any computer as long as it has:
  - An internet connection.
  - A modern browser.
- There is no need to install additional software in your computer.

MATLAB Online

- *https://au.mathworks.com/academia/tah-portal/macquarie-university-916052.html*
- Create an account using your student email address.
- *Do not use your student password (create a new one).*

File vectorsMatrices.m

```
data1 = zeros(1, 5) % row vector
data2 = zeros(5, 1) % column vector
mult = 5.4
data3 = ones(10, 1) * mult %scalar multiplication
data4 = data3 + 5 %scalar addition
taxicab1 = [10 70 20 90]
taxicab2 = taxicab1' %converting between row and column vector
myFirstMatrix1 = [10 70; 20 90; 30 80]
myFirstMatrix2 = myFirstMatrix1 + 2.5 %scalar addition
```

```
col1 = [10 70 20 90 30 80]'
col2 = [40 80 20 60 30 10]'
col3 = [20 10 0 -100 -2000 0]'
sumCols = col1 + col2 + col3
%you can add or subtract vector/matrices of the same size

mat1 = [10 70 20; 30 90 80]
mat2 = [50 10 90; 100 30 70]
mat3 = (mat1 + mat2)'

mat3(3,1) %matrix(row number, column number)
```

Running File vectorsMatrices.m in MATLAB Online

In this demonstration we will:

1. Upload a MATLAB script to MATLAB Drive.
2. Open the MATLAB script.
3. Run the MATLAB script.

MATLAB File Format

- We will use the file extension .m for all MATLAB scripts.
- If you open the file with a text file, you will see that it is plain text.
- If you double click on the file and MATLAB is installed in your computer, MATLAB will open the file.

Scripting Languages and Reproducibility

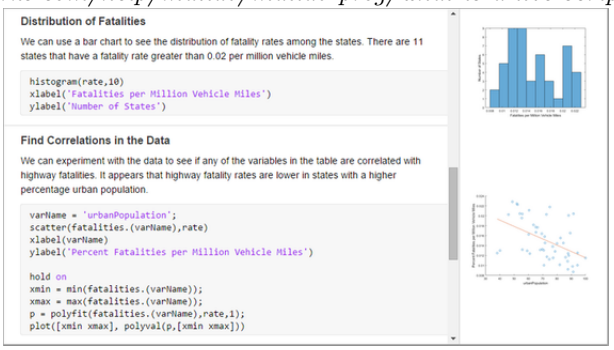
The Problem with Regular Scripts

- Regular scripts (e.g. MATLAB's .m files) are good if we want to run code.
- But what if we can to *keep record of a scientific experiment?*
- We will want to supplement the code with comments and explanations.
- We will also want to show the output of some of the code, e.g. plots.

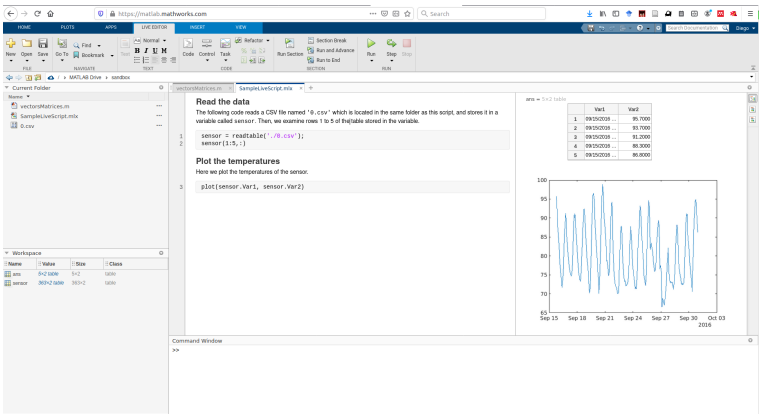
Notebooks for Reproducibility

- Some run time environments allow the creation of notebooks.
  - Called *live scripts* in MATLAB.
- These notebooks are the digital equivalent of lab notebooks.
- Notebooks contain sections that can be executed.
- The results of execution appear in the notebook.
- Notebooks also contain formatted text for documentation and explanations.

https://au.mathworks.com/help/matlab/matlab_prog/what-is-a-live-script-or-function.html



Demonstration of a Live Script
File: SampleLiveScript.mlx



Running SampleLiveScript.mlx in MATLAB Online

In this demonstration we will:

1. Upload a MATLAB Live Script to MATLAB Drive.
2. Open the MATLAB Live Script by double-clicking on the uploaded file.
3. Run the MATLAB Live Script.

MATLAB File Format

- MATLAB Live Scripts use the file extension .mlx.
- This is a special file format. If you open the file with a text editor, you will see garbage.
- If you double click on the file and MATLAB is installed in your computer, MATLAB will open the file.

3 MATLAB

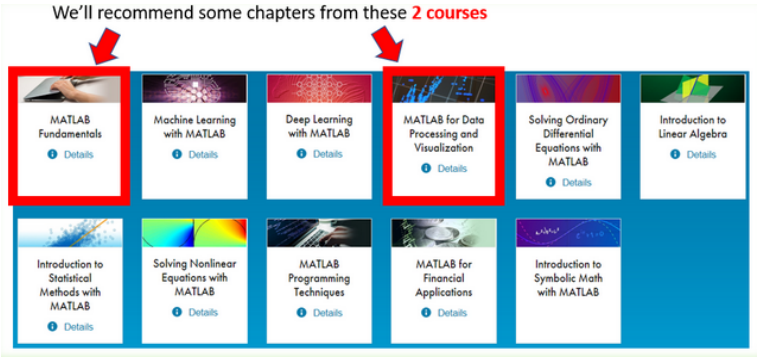
What is MATLAB?

- MATLAB is a scripting language.

- Includes types designed to store and manipulate data.
  - Vectors and matrices (MATLAB = MATrix LABoratory)
  - *Tables* (our focus in this unit)
- Includes a large library of functions for data analysis, manipulation, and visualisation.
- Has extensive documentation and on-line courses.
- Easy to use
- Others programming languages have attempted to integrate some of MATLAB's features.
  - Matrices, tables
  - Plots
  - Interactive notebooks

Accessing MATLAB and MATLAB Online

- Macquarie University has a license for students: *https://au.mathworks.com/academia/tah-portal/macquarie-university-916052.html*
- MATLAB Online here: *https://matlab.mathworks.com/*
- Getting started: *https://au.mathworks.com/help/matlab/getting-started-with-matlab.html*
- Self-paced courses: *https://matlabacademy.mathworks.com/*



Importing CSV Files in MATLAB

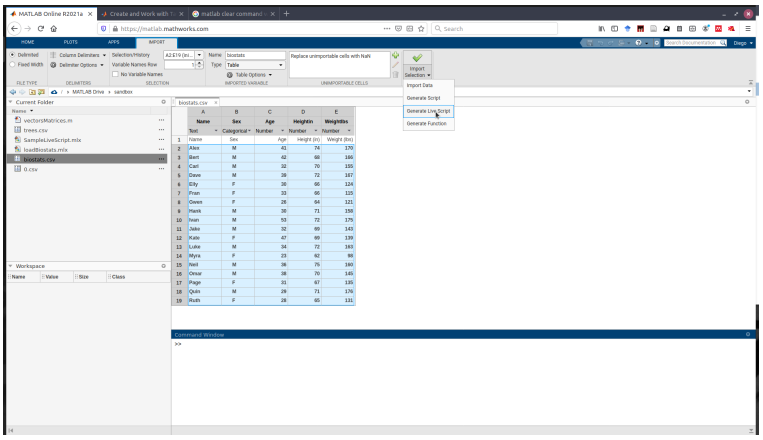
- MATLAB Fundamentals, Chapter 10, "Tables of Data"
- *https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/ create-a-table.html*
- MATLAB can store tables into variables.
- You can use the MATLAB "Import Data" wizard to load CSV files (next slide).
- Or you can use the readtable instruction.
  - trees = readtable("biostats.csv");

  (The "Import Data" tool will generate a MATLAB script that ultimately executes readtable("biostats.csv", opts), where opts specifies options that override readtable's defaults.)

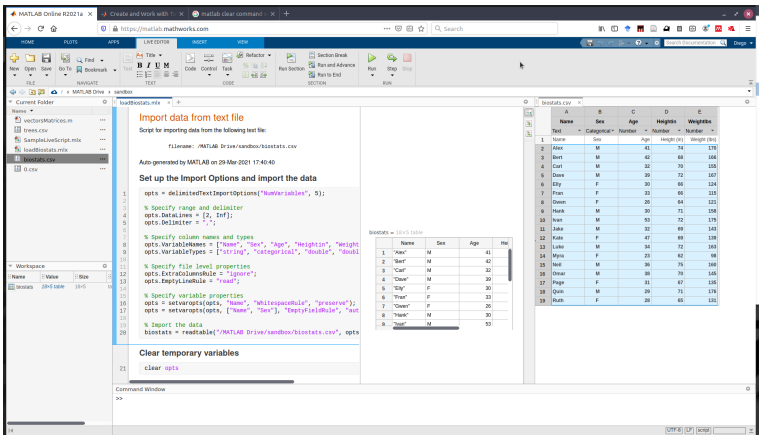
Generate a live script to import file biostats.csv

1. Upload file biostats.csv to MATLAB Drive.
2. Double-click on biostats.csv. The “Import Data” tool will open.
3. Select the correct options (the defaults will not always work):
   - Name of the MATLAB variable where the table will be stored.
   - Names of the columns.
   - Data types of the columns.
   - The range of the data to import.
4. Select the correct import option in the ”Import Selection” dropbox.
   - Select ”Generate Live Script”
5. Save the live script and execute it (click on ”Run”).
   - The live script will be saved into a file with extension .mlx
   - This file is saved in MATLAB Drive “in the cloud”, not in your desktop computer.

Generating the live script



After running the live script



Anatomy of a MATLAB Table

- A MATLAB table has *rows* and *columns*.
- All cells in the column are of *the same data type*.
  - For example, the "Name" column of the "biostats" table has cells of type string.
- Cells in different columns of the same table may have different data types.
- When you load a CSV file and store it into a MATLAB variable, you can determine the data type associated with each column.
  - MATLAB will guess the best data type for each column but sometimes you want to override these guesses.

biostats = 18×5 table

	Name	Sex	Age	Heightin	Weightlbs
6	"Fran"	F	33	66	115
7	"Gwen"	F	26	64	121
8	"Hank"	M	30	71	158
9	"Ivan"	M	53	72	175
10	"Jake"	M	32	69	143
11	"Kate"	F	47	69	139
12	"Luke"	M	34	72	163
13	"Myra"	F	23	62	98
14	"Neil"	M	36	75	160

Accessing and Modiying Data in MATLAB Tables

We will see how we can do the following in MATLAB:

- Accessing *all* the values of a column.
- Accessing *some* values of a column.
- Accessing *all* the columns of a row.
- Accessing *some* of the columns of a row.
- Modifying the data of the table.

Accesing data in MATLAB tables

Example: biostats.csv

https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/ access-data-in-a-table.html

- Accessing all values of a column: names = biostats.Name
  - This will create a *column vector* and store it in variable names.
- Accessing some values of a column: names1to5 = biostats.Name(1:5)
  - This will create a *column vector* with the first 5 values of the Name column and store it in variable names1to5.
- Accessing the second row:
  - secondrow1 = biostats(2,:) as table (will create a *table* with one row only, and store it in variable secondrow.)
  - secondrow2 = biostats{2,4:5} as an array (will extract the data from row 2 and columns 4 and 5 (Height_in and Weight_lbs) and create a *row vector* which will be stored in variable secondrow.)

- secondrow3 = biostats{2,1:5} will not work because the result cannot be an array! (there are multiple data types)

- Accessing columns: heightcm = biostats.Heightin ∗ 2.54
  - This will convert all values of column Heightin to cm and store them in a new variable heightcm.

Modifying data in MATLAB tables

Example: biostats.csv

https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/ access-data-in-a-table.html

- Creating a new column: biostats.Heightcm = heightcm
  - This will create a new column in table biostats. The name of the column is Heightcm. The values of variable heightcm will be copied (assigned to) this new column.
- Modifying part of a column: biostats.Heightcm(1:3) = [0;0;0] (Q: why the semicolons?)
  - This will set the first 3 elements of column Heightcm to zero. Note that we must assign a *column vector*.
- Modifying a row. The following two are equivalent:
  - biostats{7,["Age" "Weight_lbs]} = [31 110]: indexing by column name.
  - biostats{7,[3 5]} = [31 110]: indexing by column number.

Take-home Messages

- Excel as a tool to manage data in science.
- Excel tables.
- Scripting languages are powerful means to allow reproducibility.
- Scripting languages can be executed by a computer.
- Some environments allow the use of interactive notebooks for better reproducibility.
- MATLAB is a powerful scripting language designed for data analysis.
- Importing data in MATLAB.
- Accessing table rows and columns in MATLAB.

What's Next

- Week 6, SGTA: Quiz 2
  - *Attend the SGTA class you registered to*
  - *Friday 1 April 6-9pm for FOSX students*
- Week 7 lecture: Cleaning data
- Week 7, Friday 8 April 5pm: Communicator hurdle
- *2 weeks without class after week 7*