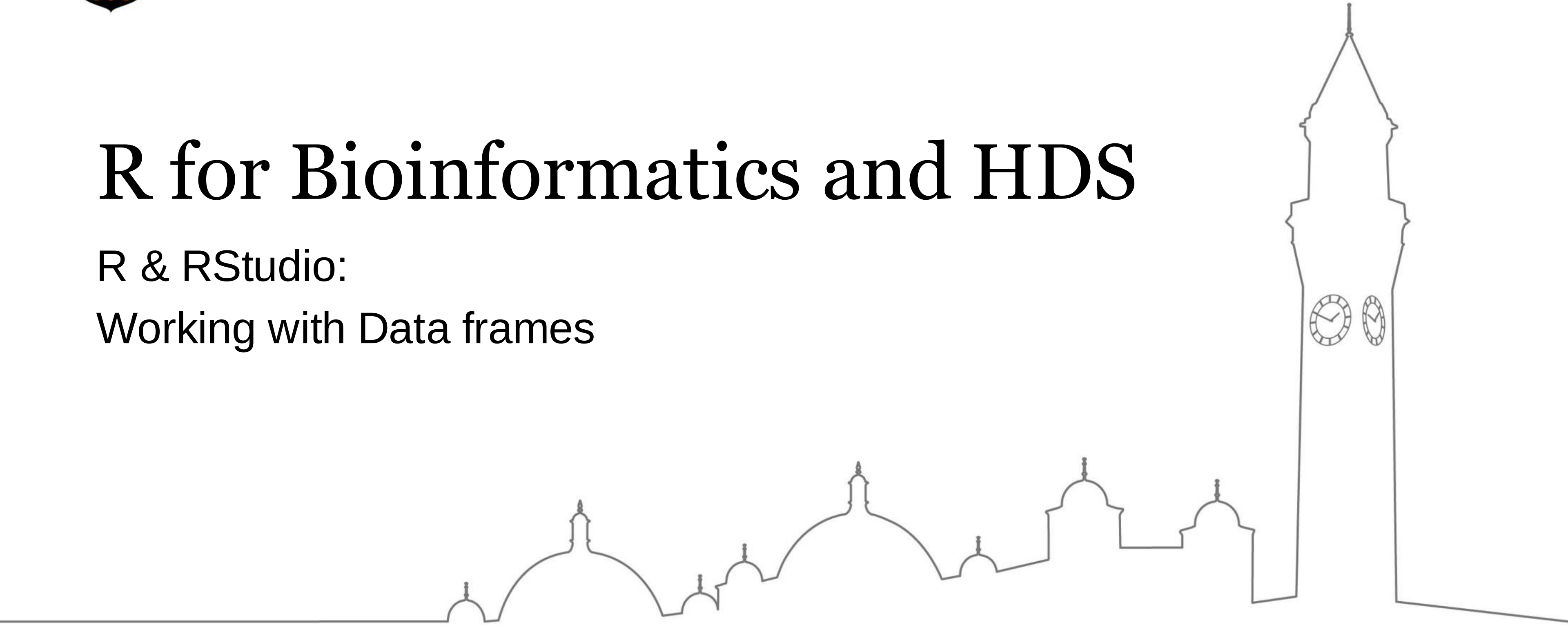
UNIVERSITY OF BIRMINGHAM

R for Bioinformatics and HDS

R & RStudio:
Working with Data frames



Vasileios Panagiotis Lenis
Laura Bravo Merodio

Course Structure

- Introduction to R & RStudio
- Syntax, Comments, Variables, Data Types and Operators
- Conditions, Loops, Functions and Data Structures
- Working with Data frames

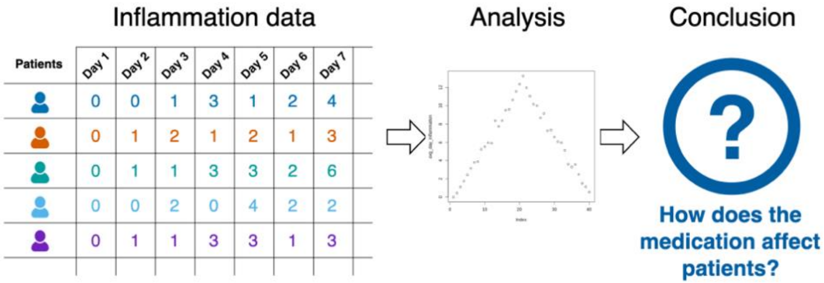
Getting started

- The best way to learn how to work with data frames is to do something useful, so this session is built around a common scientific task: data analysis.

- To do that, we will use the following case scenario, borrowed by the Carpentries:

Arthritis Inflammation

- We are studying **inflammation in patients** who have been given a new treatment for arthritis.
- There are 60 patients, who had their inflammation levels recorded for 40 days. We want to analyse these recordings to study the effect of the new arthritis treatment.
- To see how the treatment is affecting the patients in general, we would like to:
  1. Calculate the average inflammation per day across all patients.
  2. Plot the result to discuss and share with colleagues.



Inflammation data

Patients	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
●	0	0	1	3	1	2	4
●	0	1	2	1	2	1	3
●	0	1	1	3	3	2	6
●	0	0	2	0	4	2	2
●	0	1	1	3	3	1	3

Analysis

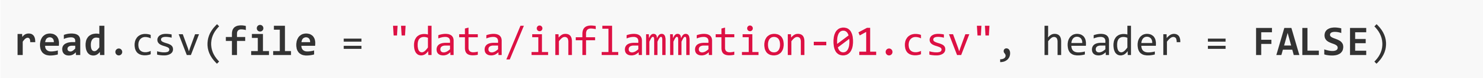
Conclusion

How does the medication affect patients?

Loading the inflammation data

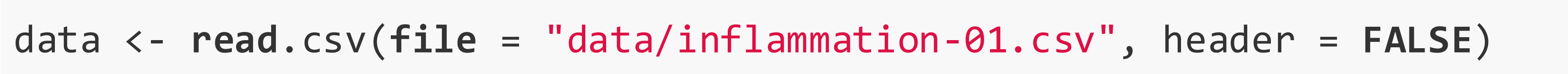
- To begin processing data, we need to load them.

- Loading our inflammation data



```
read.csv(file = "data/inflammation-01.csv", header = FALSE)
```

- Don’t forget! Store them into a variable to reuse them



```
data <- read.csv(file = "data/inflammation-01.csv", header = FALSE)
```

What is the structure of the ”data” object?

- Data frame:
- Think of this structure as a spreadsheet in MS Excel that many of us are familiar with.
- Data frames are very useful for storing data and you will use them frequently when programming in R.
- A typical data frame of experimental data contains individual observations in rows and variables in columns.

Data Inspection

- Let’s see the data

```
data
```

- What’s the type of our data

```
class(data)
```

- What’s the shape? (How many days? How many patients?)

```
dim(data)
```

Data Navigation

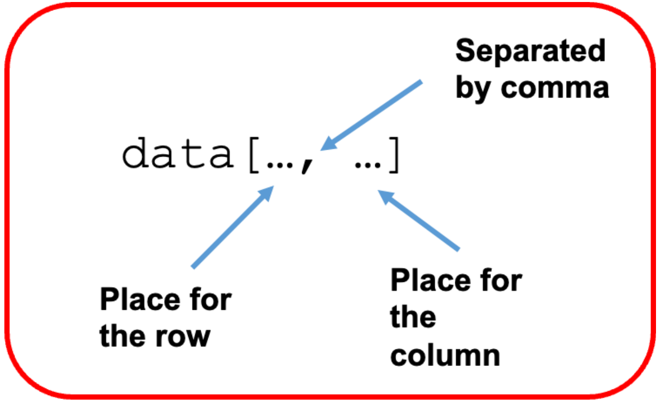
- What is the inflammation value of the first patient in the first day?

  Remember! We start counting from “1”!

```
data[1, 1]
```

- What about the middle one?

```
data[30, 20]
```



Separated by comma

data[…, …]

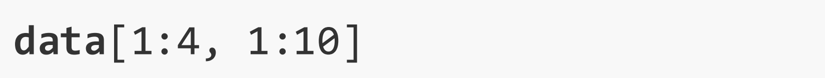
Place for the row

Place for the column

Slicing data

- We can take more than one values from the dataset

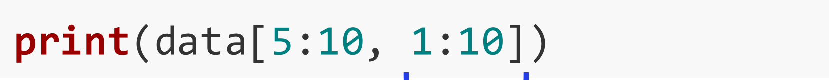
The inflammation scores of the first 4 patients across the first 10 days:



```
data[1:4, 1:10]
```

Or

The inflammation scores of 5 -10 patients across the first 10 days:



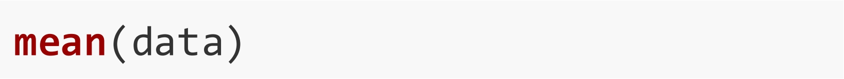
```
print(data[5:10, 1:10])
```

Range

Analyzing Data

- We use functions (“canned scripts” that automate something complicated)
- Functions take inputs as arguments and return values as outputs (not in all cases!)

- How can we find the average of all the inflammation scores:



```
mean(data)
```

We use the function ”mean” and our data as input

Analyzing data – Descriptive functions

- We can apply many descriptive statistics

  maxval, minval, stdval = max(data), min(data), sd(data)

- We can focus on the stats of one patient, or of one day

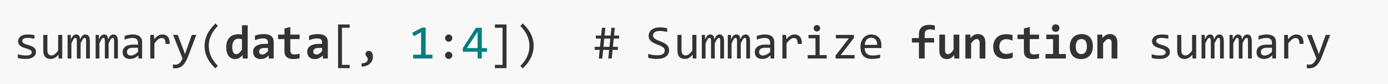
Like:

```
max(data[1,] #The maximum inflammation of patient 1
min(data[,1] #The minimum inflammation of day 1
median(data[,7]) #The median inflammation of day 7
```

UNIVERSITY OF BIRMINGHAM

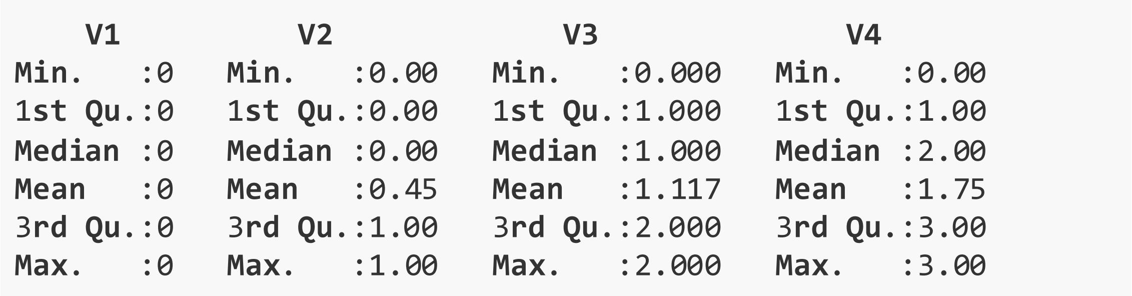
Analysing data - summary

- We can have a summary of the descriptive statistics of all our data



```
summary(data[, 1:4])  # Summarize function summary
```

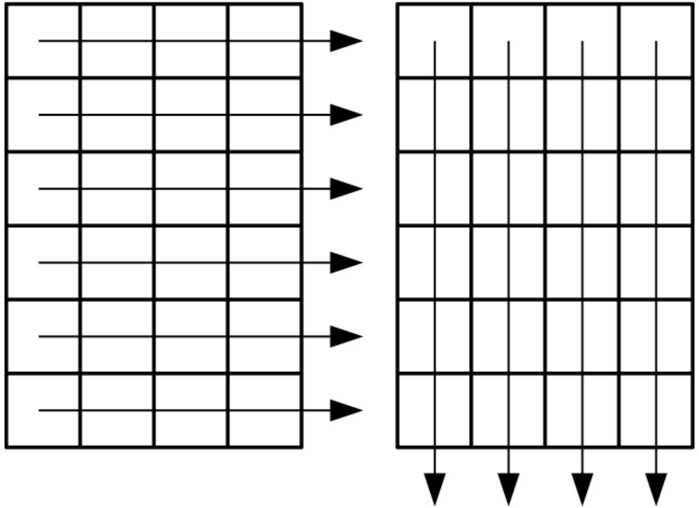
OUTPUT<>



```
     V1          V2               V3                V4
Min.   :0   Min.   :0.00   Min.   :0.000   Min.   :0.00
1st Qu.:0   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:1.00
Median :0   Median :0.00   Median :1.000   Median :2.00
Mean   :0   Mean   :0.45   Mean   :1.117   Mean   :1.75
3rd Qu.:0   3rd Qu.:1.00   3rd Qu.:2.000   3rd Qu.:3.00
Max.   :0   Max.   :1.00   Max.   :2.000   Max.   :3.00
```

A little bit more advanced calculations….

- What if we need the maximum inflammation for each patient over all days or the average inflammation score for each day?



across rows (1)

across columns (2)

The *apply()* function

- If we want to calculate the average across the rows

```
apply(data, 1, mean)
```

- If we want to calculate the average across the columns

```
apply(data, 2, mean)
```

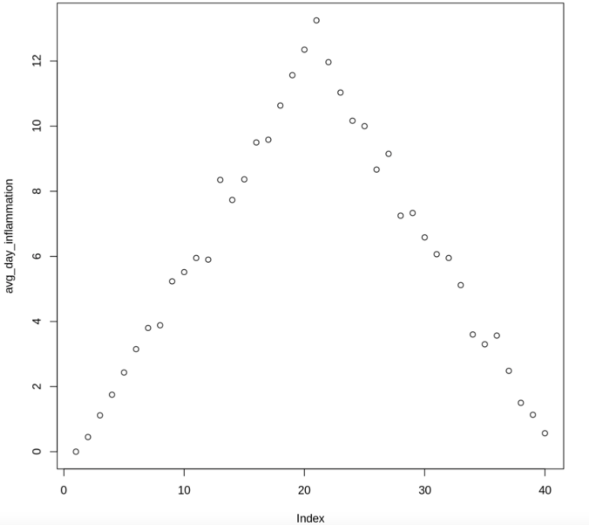
- And we can store them to a variable to reuse them…

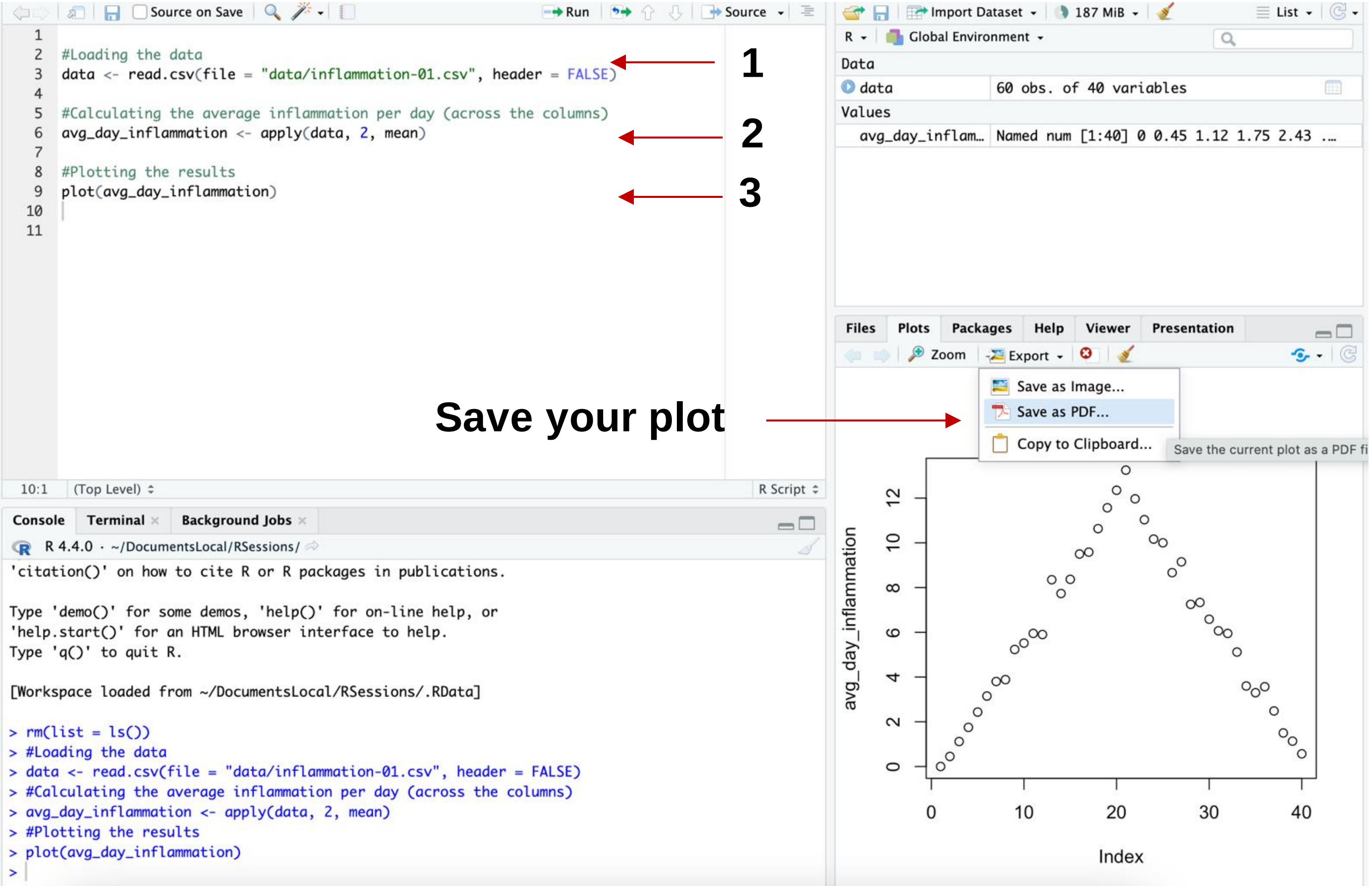
```
avg_patient_inflammation <- apply(data, 1, mean)
avg_day_inflammation <- apply(data, 2, mean)
```

Let’s plot!

- Let’s have a look at the average inflammation over time:

```
avg_day_inflammation <- apply(data, 2, mean)
plot(avg_day_inflammation)
```





```
#Loading the data
data <- read.csv(file = "data/inflammation-01.csv", header = FALSE)

#Calculating the average inflammation per day (across the columns)
avg_day_inflammation <- apply(data, 2, mean)

#Plotting the results
plot(avg_day_inflammation)
```

1

2

3

Save your plot

Data

data 60 obs. of 40 variables

Values

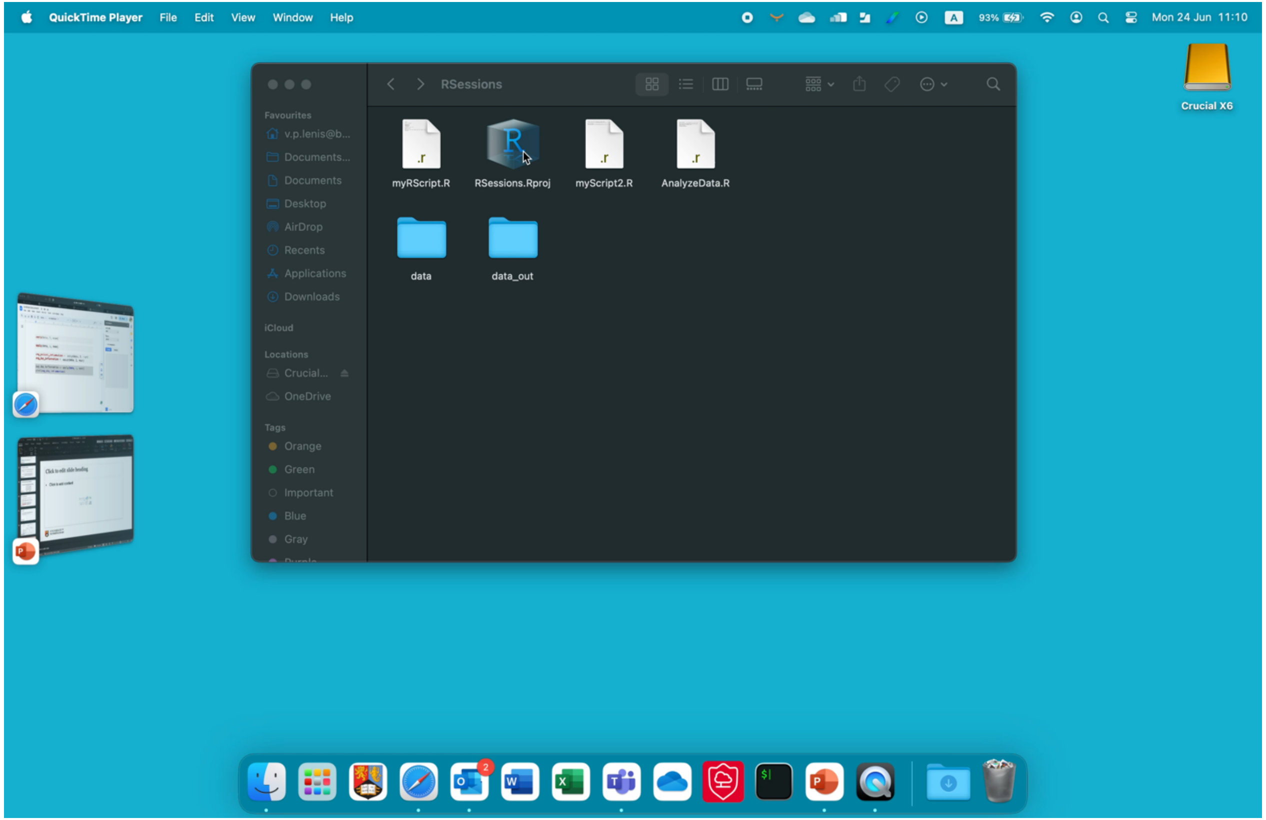
avg_day_inflam… Named num [1:40] 0 0.45 1.12 1.75 2.43 ...

```
R 4.4.0 · ~/DocumentsLocal/RSessions/
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/DocumentsLocal/RSessions/.RData]

> rm(list = ls())
> #Loading the data
> data <- read.csv(file = "data/inflammation-01.csv", header = FALSE)
> #Calculating the average inflammation per day (across the columns)
> avg_day_inflammation <- apply(data, 2, mean)
> #Plotting the results
> plot(avg_day_inflammation)
>
```


QuickTime Player
File
Edit
View
Window
Help
93%
Mon 24 Jun 11:10
Crucial X6
RSessions
Favourites
v.p.lenis@b...
Documents...
Documents
Desktop
AirDrop
Recents
Applications
Downloads
iCloud
Locations
Crucial...
OneDrive
Tags
Orange
Green
Important
Blue
Gray
myRScript.R
RSessions.Rproj
myScript2.R
AnalyzeData.R
data
data_out

v.p.lenis@bham.ac.uk (Vasilis)

l.bravo@bham.ac.uk (Laura)