Data Jujutsu II – PhD Trends[*]

Stefano Allesina & Graham Smith *University of Chicago*

Description of the data

Every year, the National Science Foundation sponsors a very large survey (with almost complete sampling) of the PhD graduates, the *Survey of Earned Doctorates* (SED). They publish statistics on the number of PhDs awarded per year, and report PhD completion by gender, field, ethnic background, etc. In particular, table 16 reports the number of PhDs awarded in total (and divided by sex) for each field of study. We are going to attempt reading the table directly from the `xlsx` files that are published by NSF.

The challenge

1. Read the data The file `urls_and_skip_NSF_SED.csv` reports the location (`url`) of the excel files for the years 2013-1018, as well as the number of lines to skip (`skip`) and the number of lines to read (`read`) for best results. Read the documentation of `read_xlsx` from the library `readxl` to see how to read the file while skipping a few lines and capping the total number of lines to be read.

```
library(tidyverse)
library(readxl)
read_csv("urls_and_skip_NSF_SED.csv")
```

```
## # A tibble: 6 x 4
##    year url                                                     skip  read
##   <dbl> <chr>                                                  <dbl> <dbl>
## 1  2018 https://ncses.nsf.gov/pubs/nsf20301/assets/data-tables/tabl~     3   274
## 2  2017 https://ncses.nsf.gov/pubs/nsf19301/assets/data/tables/sed1~     3   271
## 3  2016 https://nsf.gov/statistics/2018/nsf18304/data/tab16.xlsx         1   270
## 4  2015 https://nsf.gov/statistics/2017/nsf17306/data/tab16.xlsx         1   264
## 5  2014 https://nsf.gov/statistics/2016/nsf16300/data/tab16.xlsx         1   293
## 6  2013 https://nsf.gov/statistics/sed/2013/data/tab16.xlsx              1   284
```

Read all the files, building the tibble `tb_sed` retaining only the field and the total for each year:

```
source("solution_PhD_trends.R") # this is the code you have to write!
tb_sed
```

```
## # A tibble: 1,583 x 3
##    field                                                                total  year
##    <chr>                                                                <dbl> <dbl>
##  1 All fields                                                           55195  2018
```

[*] This document is included as part of the Advanced Computing packet for the U Chicago BSD qBio6 boot camp 2020. **Current version**: August 11, 2020; **Corresponding author**: sallesina@uchicago.edu.

```
##  2 Life sciences                                                     12780  2018
##  3 Agricultural sciences and natural resources                        1445  2018
##  4 Agricultural sciences                                               875  2018
##  5 Agricultural economics                                              108  2018
##  6 Agronomy, horticulture science, plant breeding, plant pathology,~   349  2018
##  7 Animal nutrition, poultry science                                    68  2018
##  8 Animal sciences, other                                              121  2018
##  9 Food science, food technology-other                                 163  2018
## 10 Soil chemistry and microbiology, soil sciences-other                 66  2018
## # ... with 1,573 more rows
```

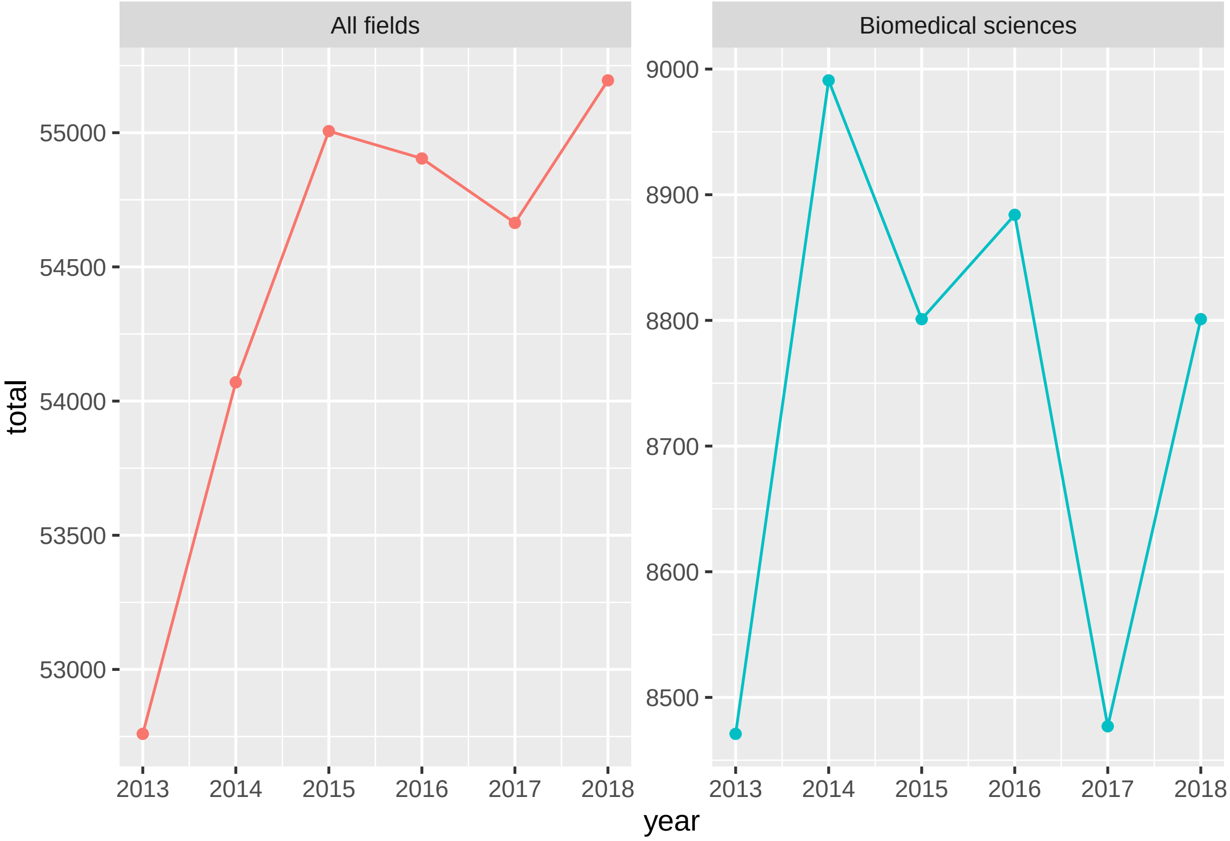
2. Standardize names and filter Notice that there are very many fields, and that the names of some fields have changed through the years (e.g., `Neurosciences, neurobiology`, `Neurosciences and neurobiology`). The file `lookup_fields_filter.csv` contains two columns: retain all the records for the fields specified in the table, and use the column `name_to_use` to standardize the names of the fields. You should end up with 18 fields (all well-represented at U of C) as well as the data for `All fields` and `Biomedical sciences`.

```r
tb_sed
```

```
## # A tibble: 120 x 4
##    field                                    total  year name_to_use
##    <chr>                                    <dbl> <dbl> <chr>
##  1 All fields                               55195  2018 All fields
##  2 Biological and biomedical sciences        8801  2018 Biomedical sciences
##  3 Anatomy, developmental biology             158  2018 Developmental biology
##  4 Biochemistry (biological sciences)         811  2018 Biochemistry
##  5 Bioinformatics                             203  2018 Bioinformatics
##  6 Biometrics and biostatistics               233  2018 Biostatistics
##  7 Biophysics (biological sciences)           152  2018 Biophysics
##  8 Cancer biology                             355  2018 Cancer biology
##  9 Cell, cellular biology, and histology      218  2018 Cell biology
## 10 Computational biology                      146  2018 Computational biology
## # ... with 110 more rows
```

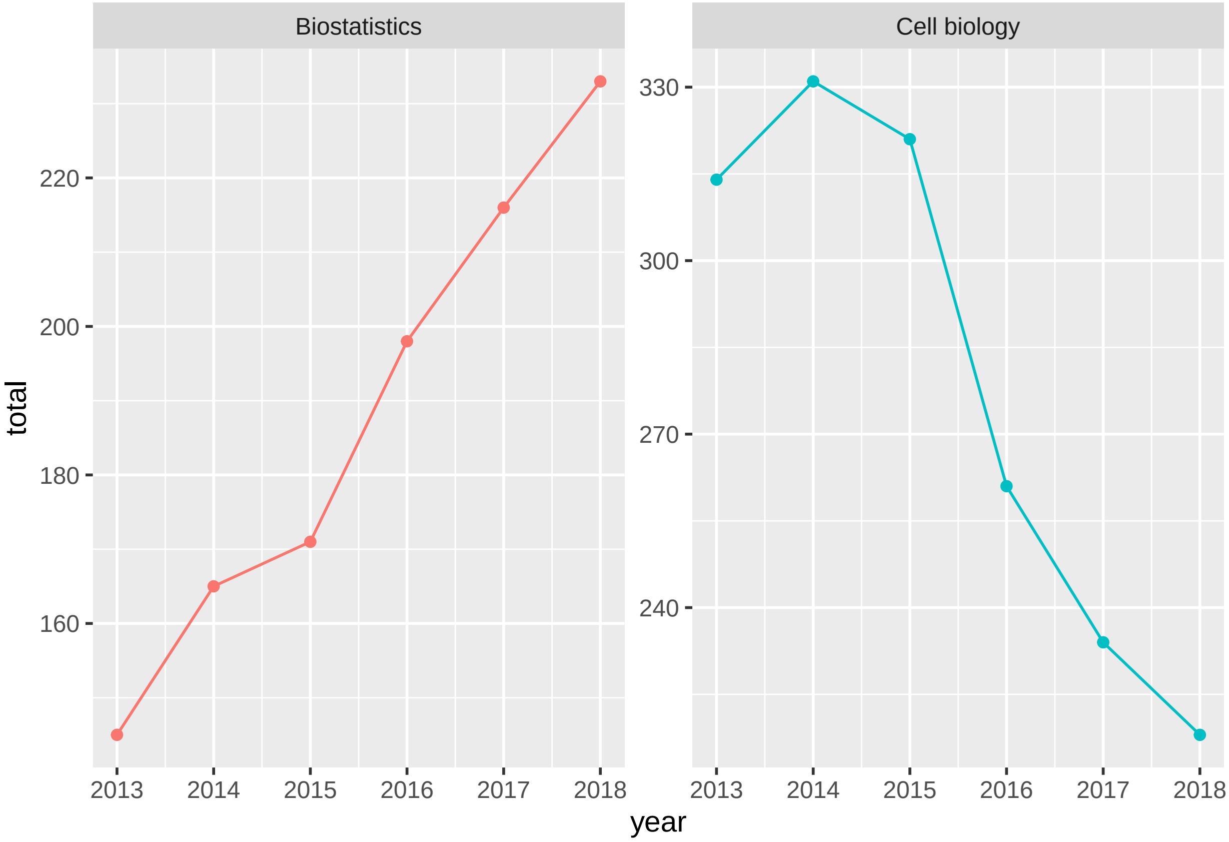
3. Plot the time series Write a generic function for plotting the number of PhDs awarded for all the fields in a tibble. For example, here are the trends for all PhDs and all PhDs in biomedical sciences:

```r
tb_sed %>%
  filter(name_to_use %in% c("All fields", "Biomedical sciences")) %>%
  plot_PhD_in_time()
```



4. Fields that have changed the most Some fields have grown considerably in the past 6 years, while some have shrunk. For example:

```r
tb_sed %>%
  filter(name_to_use %in% c("Cell biology", "Biostatistics")) %>%
  plot_PhD_in_time()
```



Find the fields for which the ratio between the maximum number of PhDs and the minimum number of PhDs for the period considered is the largest.

5. Correlation between time series [Optional] Compute the correlation (using the function `cor`) between the time series of any two fields. Which fields have changed in synchrony? Plot the matrix of correlations using `geom_tile`.

6. Order the matrix [Optional, requires some math] Find a good ordering for the matrix by plotting the field according the the eigenvector of the correlation matrix corresponding to the largest eigenvalue (the function `eigen` computes eigenvalues and eigenvectors of a squared matrix).

Hints & Nifty tricks

- If you don't want to store the downloaded zip file, use a temporary file (it will be deleted by `R` automatically once you call `unlink()`)
- Some lines are empty: use something like `filter(!is.na(field))` to get rid of them.
- For each year, you only need to store the number of PhD awarded.
- To force a certain order for the labels in the graph, transform them to factors, using `factor`: `my_tibble <- my_tibble %>% mutate(my_labs = factor(my_labs, levels = my_order))`