# ECPR Methods Summer School: Automated Collection of Web and Social Data

**Pablo Barberá**

London School of Economics
pablobarbera.com

Course website:
pablobarbera.com/ECPR-SC104

# Course logistics

ECTS credits:

- Attendance: 2 credits (pass/fail grade)
- Submission of at least 3 coding challenges: +1 credit
- Submission of class project: +1 credit
  - Due by August 20th via email to P.Barbera@lse.ac.uk
  - Goal: collect and analyze data from the web or social media
  - Examples:
    - Scrape a Parliament website and do a descriptive analysis of speeches
    - Scrape a site with election results and plot evolution of party vote share over time
    - Collect tweets about a particular topic and identify most central actors
    - ...anything that is useful for your research!
  - 5 pages max (including code) in Rmarkdown format
  - Graded on a 100-point scale

If you wish to obtain more than 2 credits, please indicate so in the attendance sheet

# Encoding issues

# Character encodings

- **Encoding:** how digital binary signals are translated into human-readable characters.

→ e.g. `0100100` is displayed as 'd'

- This also includes characters such as á, ç, ü, etc.
- **Problem:** many different translation tables, sometimes hard to know which one is used
- R works with the default encoding scheme in your system:

```
> Sys.getlocale(category = "LC_CTYPE")
[1] "en_US.UTF-8"
```

- For English Mac and Linux systems, generally UTF-8. For Windows systems, Windows-1252.
- UTF-8 (part of Unicode standard) is most popular scheme and used on many websites.

# Some final reminders...

1. You can download all your code, challenges, and data from RStudio Server:
   - $\rightarrow$ Export $>$ download as .zip file
     - Server will be deactivated tonight at 10pm
2. Materials (but not solutions) will remain on course website
3. How you can contact me after the course:
   - `P.Barbera@lse.ac.uk`
   - `www.pablobarbera.com`
   - `@p_barbera`