DIGITAL METHODS FOR SOCIAL RESEARCH

Étienne Ollion (eollion@ens.fr)
CNRS Permanent Research Fellow

The recent data deluge has been met with strong reactions. While some enthusiastically hailed the availability of massive information about individual behaviors, others openly voiced concerns. The relevance for research, the potential transformations in the focus of the disciplines or the ethical dilemmas raised by the use of such data were widely discussed. Empirical bounty vs. meager quality of the data; pathway to scientific breakthroughs vs. threats to public liberties; methodological revolution vs. latest commercial fad…In the public sphere just like in academia, *big data* became the flag under which the proponents of the "data revolution" and their critics waged an unremitting war.

Rather than directly taking sides in these ongoing controversies, this course will concretely assess the relevance of digital methods for social sciences. The class will cover some of the most central tools for conducting data analysis, and the approach will be hands-on so as enable the students to carry out their own project by teaching them the central tools of data analysis. It is our contention that this *detour* through the techniques of data science can effectively shed new lights on these controversies.

The course is then twofold, as it meshes a presentation of data collection and management tools with a reflection on their production and their potential use for research. How is the web written (and how to crawl it)? How to extract relevant information from a file (and what is lost in the process)? How to deal with the mass of data (and how does this abundance impact scientific research)? From databases to visualization, different aspects of these methods will be explored.

REQUIREMENTS

The course requires no preexisting computer science skills. It does not require statistical skills either, since the goal of this course is to produce the data that can then be treated with quantitative methods (or else). Most of the work will be carried out using the R statistical software.

COURSE ORGANIZATION AND GOALS

This intensive course is organized in **5 sessions**. Each will be twofold. Organized around concrete research questions (collecting data from the web, automating extraction, storage, legal and ethical aspects), **the course will serve as an introduction to the techniques and the questions they raise**. The second part of the course will be **dedicated to presenting tools that concretely help students with their own project**.

Course 1: Digital Strategies for the Social Sciences

This session will present the current debates about digital methods for social sciences, and a panorama of the tools available for conducting digital research.
Lab: Getting started with the software: a gentle introduction to R

Course 2: How is the web written (and how to read it)?

Taking a look at the Internet "from the backstage," this session will focus on how the web functions, and in particular on how it is written. Through an emphasis on the techniques to crawl and parse a webpage, the course will serve as an introduction to broader class of markup languages.
Lab: Crawling the web, gathering and restructuring information

Course 3: How to select data?

Data often comes in mass! This session will present various tools that allow precise data extraction from vast ensembles.
Lab: Using Xpath and Regex, two languages for data selection.

Course 4: Automation and storage

This session will focus on two recurrent challenges for digital methods: automation and storage. It will show various methods to repeat a task, and how to save the relevant information in the process.
Lab: How to save data, how to save time (yours and that of your computer)?

Course 5: APIs and pathways for research

The course will continue the presentation of data collection techniques by presenting the APIs, before we introduce various aspects related to digital methods that can be useful for research: visualization forms of quantification (we will emphasize those who deal with vast amount of data –data reduction techniques, random forests, etc), legal and ethical aspects, and cloud computing.