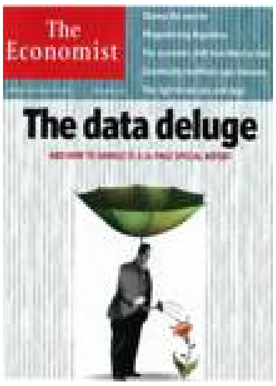
DIGITAL STRATEGIES FOR THE SOCIAL SCIENCES

Étienne Ollion (ollion@unistra.fr)

Master's Program in Computational Social Science
Linköping University
Spring 2019



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)? How can we explore the data (and what do the different methods reveal ?). From databases to deontology, 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 will take place during **2 intensive, 5 days a week session**. Morning sessions will generally be dedicated to lectures, while afternoon will be mostly applied.

Organized around concrete research questions (week 1: collecting data from the web, automating extraction, storage, legal and ethical aspects ; week 2: analyzing massive data with several tools), **the course will serve as an introduction to the techniques and the questions they raise**.

* *
*

Week 1 (May 6th – May 10th 2018)

Lecture: 10:15-12:00
Lab: 1:15-3:00 (extra lab session 3:15-5:00 on Tu/Thu)

Course 1: Beyond « Big Data »
This session will present the current debates about digital methods for social sciences, and a panorama of the tools available for conducting digital research. We will also evoke legal and ethical matters.
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? (1)
Data often comes in mass! This session will present various tools that allow precise data extraction from vast ensembles using some properties of markup languages.
Lab: Using Xpath

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: How to select data? (2)
The course will continue the presentation of data selection techniques, by introducing regular expressions which allow for selection in full-text (regex).
Lab: Using regex

*   *
*

Week 2 (May 20th – May 24th 2018)

Lecture: 10:15-12:00
Lab: 1:15-3:00 (extra lab session 3:15-5:00 on Tu/Thu)

Social scientists often face the challenge of interpreting quantitative data, or qualitative data that could be quantified. Organizational and administrative registers, bio- and bibliographic information, and data concerning practices or responses to ad-hoc surveys have long been exploited by researchers. The ongoing digitization of everyday life makes such data, “big” or “small,” much more readily available for social-scientific analysis.

The goal of this course is to offer an introduction to a form of social-scientific investigation currently on the rise, quantitative description. Located between inferential statistics and qualitative approaches, this type of quantification is increasingly used for exploratory as well as confirmatory purposes. The course gives an overview of some of the most important techniques in this area, before delving into some of them. Throughout th course, students will learn how to operationalize a research question and create a new dataset, and how to analyze this dataset through a range of different methods.

Course 1: An introduction to quantitative description
This session presents the history, merits, and challenges of quantitative description in the social sciences. In so doing, it will cover the main definitions of the terms needed for statistical analysis this week.
Lab: Descriptive statistics and some remarks on visualization

Course 2: Clustering
How can observations be grouped together, and according to which principles? This session will introduce various clustering techniques, classic and more novel ones.
Lab: Implementation (k-means, hierarchical clustering, partition around medoids, etc)

Course 3: Dimensionality Reduction (1) – GDA
This session will offer an introduction to geometric data analysis. Born half a century ago, it has since then gained international recognition following its use by Pierre Bourdieu to map out social spaces. In addition to this possible use, we will see two other applications of the method: data exploration, and feature extraction.
Lab: Implementation: doing Principal component analysis on R

Course 4: Dimensionality Reduction (2) – GDA, advanced topics
This session will continue the presentation of GDA by focusing on specific tools (confidence ellipses, specific CA) and discussing classic issues.
Lab: Implementation: using FactoMineR and related packages

Course 5: Non-supervised machine learning
This session will serve as a brief introduction to machine learning (ML), a new class of statistical techniques. In terms of origins, goals, criteria of validation, ML and classical statistics differ widely. Specifically, we will compare GDA with a number of machine-learning algorithms in an attempt to assess their respective merits and limitations.
Lab: Implementation: t-SNE and Self Organizing Maps.