GESIS Fall Seminar in Computational Social Science 2023

Syllabus for week 1:
"Big Data and Computation for Social Data Science"

Lecturers:	Akitaka Matsuo	David (Yen-Chieh) Liao
Affiliation:	University of Essex	Aarhus University
Email:	a.matsuo@essex.ac.uk	davidycliao@gmail.com

Date: September 11-15, 2023
Time: 09:00-16:00

About the Lecturers

Akitaka Matsuo is a postdoctoral fellow at the Institute for Analytics and Data Science, the University of Essex. His research interests lie in data science and politics, in particular in the statistical methodology for scaling survey responses, legislative behavior, and natural language processing of political texts. The applications of his methodological expertise include analyses to examine the impacts of gender on election campaigns, sentiment in parliamentary speeches, and the polarization of public opinion on diplomatic issues.

David (Yen-Chieh) Liao is a postdoctoral researcher at the School of Politics and International Relations at University College Dublin. He is also a member of the Connected_Politics Lab at UCD's College of Social Sciences and Law. His main research interests include legislative studies, party competition, and electoral systems. He has a specific interest in the measurement of ideological preferences through legislative voting, expert surveys, and the analysis of parliamentary speeches. His recent research agenda focuses on quantitative text analysis and computational methods to gain a deeper understanding of how political elites position themselves through their political narratives. In addition, he explores how these narratives influence political behavior and shape the attitudes and expectations of the masses concerning the future.

Course Description

This course is intended for social science researchers and practitioners who wish to gain insight by analyzing large data sets ("big data"), teaching them the infrastructure for data manipulation and analysis, and how to use that infrastructure with statistical and programming languages.

The amount of data available to social scientists is increasing every year, and such large amounts of data have the potential to provide novel insights that were previously unavailable. However, as the volume of data increases, it becomes less feasible to load and process them on a personal computer. What is needed in such cases is databases for data storage and parallel processing, and distributed computing systems for data processing and computation. Learning about them is the objective of this course.

With regard to database systems, after learning the basic concepts, participants will learn SQL, the most widely used relational database language, and its management systems. As a more advanced topic, we will overview databases other than SQL, especially MongoDB, which is an excellent non-relational destination for storing large unstructured data (e.g., text data). For data processing and computation, students will learn how to parallelize data processing and analysis and how to use distributed computation systems, such as Apache Spark.

To learn these technologies, both theory and practice are very important, and thus the course will provide both lectures and labs as one set. The primary programming language will be R, as it is a language familiar to most quantitative social scientists but given the increased importance of Python in social data science, the course will show how to use Python to do what we have learned in R, when appropriate.

Keywords

databases; parallel computing; distributed computation; cloud computing; big data

Course Prerequisites

- Experience with data analysis using R including:
  - Manipulate objects (scaler, vector, data.frame)
  - Open/write data files
  - Run and interpret basic statistical models (e.g. OLS regression, Logit/Probit models)
  - Work with packages
- Experience in Python is not required but would be a plus to understand Python examples

Target Group

Participants will find the course useful if:

- they want to work with large datasets and need to perform complex computations and data analysis tasks
- they are interested in using relational databases with SQL and non-relational databases with NoSQL, distributed computation systems such as Apache Spark, and cloud computing for their data analysis
- they have a background in R programming

Course and Learning Objectives

By the end of the course, students should have a good understanding of how to work with SQL as well as NoSQL databases in R, as well as how to leverage distributed computation systems like Spark for large-scale data processing. They should also be able to work with databases and compute clusters in the cloud. To be more concrete:

- Understanding the basics of SQL and NoSQL databases
- Writing SQL queries to retrieve data from a database
- Importing and exporting data from databases using R
- Working with non-relational databases, such as MongoDB, and understanding their data structures and query languages
- Understanding the concept of parallel computing and its advantages in data processing and analysis, including faster processing times and increased scalability
- Working with distributed computing systems such as Apache Spark
- Using R to perform data manipulation and analysis with the tidyverse packages
- Learning how to do the same process as above in Python, thereby understanding the advantages and disadvantages of R and Python in their respective ecosystems
- Understanding the importance and practice of benchmarking in data processing and analysis
- Profiling the code to find the pieces that are causing performance problems in R and Python

Organizational Structure of the Course

Each day of the course will have two 3-hour units. Each unit will include both lectures and labs.

In the lab, students will receive exercise problems to work on. The exercises are essentially given in R, and students answer them in the time allotted by the instructor. Students will work with other students to answer the questions on their own, and the two instructors will both be present in the classroom, so if they have any questions, they can always ask. The instructor will then provide the answer and, if possible, a demonstration of how to do the same thing in Python.

The instructor will also have office hours after class, where students can not only ask questions regarding the lectures and lab but also consult with the instructors on the methodological issues with their own research projects.

Software and Hardware Requirements

In this course, we will access various cloud data and computational environments, mainly using R and RStudio as well as Python and JupyterLab as a client. Participants should bring their own laptops with the following software installed:

- R (preferably latest, minimum 4.1.0)
- RStudio (latest)
- Miniconda (latest)
- Git environment (for Windows users who do not have a Bash environment)

R and RStudio should be installed beforehand, and Windows users should install Git for Windows. For Python, please install Miniconda, but building a conda environment will be done in the lab. Detailed instructions on packages and additional software (e.g., VS Code, MongoDB client) installation will be provided during the lecture and lab.

Recommended Literature to Look at in Advance

Programming

- Lander, J.P. 2014. *R for everyone: Advanced analytics and graphics*. Pearson Education. (a good refresher for R. Especially chapter 1-15)
- Sweigart, A. 2019. *Automate the boring stuff with Python: practical programming for total beginners*. No Starch Press. (a gentle introduction to Python, this is not necessary but to understand Python examples, you might want to have a look at chapters 1-6, https://automatetheboringstuff.com/)

Overview of Data Science and Big Data in Social Science

- Grimmer, J. 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(1), pp. 80-83.
- Grimmer, J., Stewart, B.M. and Roberts, M.E. 2021. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*, pp. 395-419.
- Molina, M. and Garip, F. 2019. Machine learning for sociology. *Annual Review of Sociology*, *45*, pp. 27-45.

Day-to-day Schedule and Literature

Day 1: Overview and Introductions to the Computational Environments

Session 1 (morning):

- Introduction to big data management and computation
  - Infrastructure and distributed computing
- R: Introduction
  - Tidyverse : dplyr, purr, readr, ggplot2
  - data.table
- Python: Introduction
  - Numpy, pandas, and Matplotlib

Session 2 (afternoon): Infrastructures, Environments, and Necessary Installations

- R infrastructures
  - R and RStudio
- Python
  - Anaconda/miniconda and conda environments
  - IDE (Jupyter Notebook, Colab, and Visual Studio Code)
- Lab time

Literature

Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G. 2023. *R for Data Science, 2nd Edition*. Chapters 1-7, 21. https://r4ds.hadley.nz/

VanderPlas, J. 2022. Python Data Science Handbook. Chapters 1-3. https://jakevdp.github.io/PythonDataScienceHandbook/

Day 2: Processing Big Data

Session 3 (morning): Parallelization

- Introduction to parallel processing (hardware, memory, and performance)
- Parallel strategies and tools
- Benchmarking and code optimization
- Lab time

Session 4 (afternoon): Typical Tasks of Big Data Processing in the Social Sciences

- Dealing with NLP tasks in parallelization (part of speech tagging and named entities recognition)
- Data storage
    - From R/Python environment to locally-hosted database formats
    - Database storage in the cloud (AWS S3, Azure, etc.)
- Lab time

Literature:

Gillespie, C. and Lovelace, R. 2016. *Efficient R programming: a practical guide to smarter programming*. O'Reilly. Chapter 7. https://csgillespie.github.io/efficientR/

Matloff, N. 2015. *Parallel Computing for Data Science: with Examples in R, C++ and CUDA*. CRC Press. Chapters 2, 4.

Day 3: Databases

Session 5 (morning):

- Introduction to databases and SQL
- Relational database model
- Creating and managing databases
- Basic SQL queries
- Lab time

Session 6 (afternoon):

- More on SQL queries
    - GROUP BY
    - ORDER BY
    - SUM and other aggregation
- How to use dbplyr
- Lab time

Literature:

IBM. *What's a relational database?* https://www.ibm.com/topics/relational-databases

Teate R.M.P. 2021. *SQL for Data Scientists: A Beginner's Guide for Building Datasets for Analysis.* Wiley. Chapters 1-3, 6.

Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G. 2023. *R for Data Science, 2nd Edition*. Chapter 22. https://r4ds.hadley.nz/

Day 4: Advanced SQL and noSQL Databases

Session 7 (morning):

- Advanced SQL topics
  - JOIN, VIEW
  - Subqueries and derived tables
- noSQL databases overview: MongoDB
- Lab time

Session 8 (afternoon):

- NoSQL database and MongoDB basic
- Schema and relation in MongoDB
- MongoDB queries
- Lab time

Literature:

Teate R.M.P. 2021. *SQL for Data Scientists: A Beginner's Guide for Building Datasets for Analysis.* Weily. Chapters 5-7, 12.

Phaltankar, A., Ahsan, J., Harrison, M., and Nedov, L. 2020. *MongoDB Fundamentals.* Packt, Chapters 1, 2, 4 (5).

Day 5: Distributed Computation and Apache Spark

Session 9 (morning):

- Introduction to distributed computation systems and Apache Spark
- Sparklyr and Sparkr
- Spark data wrangling
- Lab time

Session 10 (afternoon):

- Data analysis with Apache Spark
- PySpark
- Lab time

Literature:

Luraschi, J., Kuo, K. and Ruiz, E. 2019. *Mastering Spark with R: the complete guide to large-scale analysis and modeling.* O'Reilly Media. Chapters 1-5, 8. https://therinspark.com/

Additional Recommended Literature

Bealieu, A. 2020. *Learning SQL*, 3rd edition, O'Reilly (a bit more in-depth coverage of SQL, including how to work with Big Data)

Kakarla, R, Krishnan, S, and Alla, S. 2020. *Applied Data Science Using PySpark.* Apress (how to use PySpark, a Python frontend for Spark)