

# Python Data Collection and Management for Public Policy Research Fudan IGPP Summer School

Last Revised: June 23, 2024

**Instructor:** Dr. Blake Miller

**Course Schedule:** July 1-5, 8-12 14:00-16:30

- 14:00-15:00: Lecture
- 15:00-15:10: Quiz (2, 4, 9, 11 July) or Q&A
- 15:10-15:20: Break
- 15:20-16:30: Applied coding exercises

**Email:** `b.a.miller@lse.ac.uk`

**Course Description:** The massive amount of data available online continues to increase the bounds of social scientific inquiry. Researchers in both academia and the private sector can gain a greater understanding of human behavior by analyzing the abundant social data stored online. To make use of these data, one must first master technical skills necessary to gather and process these data, which can be quite challenging to do properly.





**Course Goals:** The main goal of this course is to provide students with the necessary tools for the construction, processing, and cleaning of data found online. After taking this course, students will have mastered the requisite tools needed to construct datasets out of unstructured, semi-structured, and structured online data.

**Learning Outcomes:**

1. to introduce students to important concepts and methodologies related to the management, collection, processing, and cleaning of data for social science and public policy research




2. to teach students practical concerns and best practices for data management and data collection
3. to build foundational skills necessary to construct useful datasets for their research from unstructured, semi-structured, and secondary data
4. to build a roadmap for continued learning through promoting awareness of more advanced and specialized tools and where to look for problem-solving/reference.




### Textbooks and Course Materials:

- Readings for each session are detailed below.
- Textbooks:
  - **Think Python: How to Think Like a Computer Scientist**, 2nd Edition ([English](#) , [Chinese](#) )
- Other resources:
  - [Unix Shell Tutorial from Software Carpentry](#) 
  - [Version Control with Git Tutorial from Software Carpentry](#) 

### Required Software:

This course is taught in Python, using Python 3. You will need to have Python 3 installed on your computer and bring it to class each session. If you have not yet installed Python 3, you will need to do so. Please use the following resources for installing Python 3 on your machine:




- For Windows users:
  1. Install Sublime Text from the [official Sublime Text website](#) .
  2. [Install Git Bash](#) .
  3. [Install Python 3 \(click “Latest Python3 Release”\)](#) .
  4. Open Git Bash and run `python3 --version` to verify Python 3 is accessible. If that does not work, try `python --version`.
- For Mac users:









1. Open Terminal on Mac (Applications → Utilities → Terminal).
2. Install Xcode Command Line Tools by running `xcode-select --install` in Terminal.
3. [Install homebrew](#)  by pasting the installation command from the website into Terminal. Alternatively download the installer [here](#) .
4. Install Sublime Text from the [official Sublime Text website](#)  and follow the installation instructions.
5. Use `homebrew` to install Python 3 by running `brew install python` in Terminal.
6. Run `python3 --version` in Terminal to verify Python 3 is accessible.



### Grading:

1. Quizzes (60%, in class): There will four in-class quizzes. The quizzes will test knowledge of the material covered in the previous classes and readings.
2. Final Problem Set (40%, due July 19): You will have one final problem set to apply all of the things we learned. For the problem set, you will analyze a dataset we will discuss in class in the final lectures.

### Schedule and Readings:

- *Day 1: Course Intro, Computational Social Science*
  - Edelman, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46, 61-81. ([Download Paper](#) )
  - Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721-723. ([Download Paper](#) )
  - Optional readings:
    - \* Lazer, David, and Jason Radford. 2017. “Data Ex Machina: Introduction to Big Data.” *Annual Review of Sociology* 43(1): 19–39. ([Download Paper](#) )

- \* Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060-1062. ([Download Paper](#) )
- *Day 2: Intro to the Command Line, Sublime Text, Setting Up Git*
  - [Unix Shell Tutorial from Software Carpentry](#) 
- *Day 3: git/Github, regular expressions*
  - [Version Control with Git Tutorial from Software Carpentry](#) 
- *Day 4: Basic Python (Part 1)*
  - Think Python chapters 1-3
- *Day 5: Basic Python (Part 2)*
  - Think Python chapters 5, 7-9
- *Day 6: Basic Python (Part 3)*
  - Think Python chapters 10-14
- *Day 7: Intro to Data and Data Ethics*
  - Willis, D. (2014). Professors' research project stirs political outrage in Montana. *New York Times*. ([Download Article](#) )
  - Optional readings:
    - \* Leslie, D. (2023). The ethics of computational social science. In *Handbook of Computational Social Science for Policy*, pages 57–104. Springer International Publishing Cham. ([Download Book](#) )
    - \* Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24):8788. ([Download Paper](#) )
    - \* Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook.com. *Social networks*, 30(4):330–342 ([Download Paper](#) )
    - \* Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39:629–649. ([Download Paper](#) )

- *Day 8: **pandas** (Part 1)*
  - Chugh, Vidhi. [Python pandas tutorial: The ultimate guide for beginners](#) . (2023). DataCamp.
- *Day 9: **pandas** (Part 2)*
  - Willems, Karlijn. [Pandas Tutorial: DataFrames in Python](#) . (2022). DataCamp.
- *Lecture 10: Obtaining Data from the Web*
  - *no readings.*