Computational Social Science Methods (PA397CIINF385T)

Course description

- Instructor: Ji Ma (maji@austin.utexas.edu)

- Time and location: Spring 2024, Tuesday 9:00AM-12:00PM, SRH 3.314.
- Office hour: Tuesday 2:40pm-4:40pm.

This course is academic and research oriented, it introduces and contextualizes computational methods from a social science research design perspective. The first part of this course (w1-w2) gives you an overview of this course and how to analyze computational methods from a research design perspective, namely data management, concept representation, data analysis, and scientific communication. The second part (w3-w14) will analyze and practice different computational methods according to their roles in social science research. Bilingual or multilingual language ability is a plus. Programming is an essential part of this course but not the purpose and will not be taught. We will be coding for social science purposes.

Reading materials/ e-books

Required readings:

This class will use a draft book manuscript I'm currently working on:

- [CSSPrimer] "Computational Social Science Methods: A Research Design Primer"

Additionally, each week is complemented with readings from various other sources. See details on Schedule page.

Recommended readings:

These books give you a good theoretical understanding and are very useful in research design.

- [GRS] Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton, New Jersey Oxford: Princeton University Press.
- [SJ] Scott, John. 2017. Social Network Analysis. Fourth edition. Thousand Oaks, CA: SAGE Publications Ltd. (different versions are fine)

These books/sources introduce more technical and hands-on details.

- [GS] Gentzkow, Matthew, and Jesse M. Shapiro. 2014. Code and Data for the Social Sciences: A Practitioner's Guide. https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf.
- [JM] Jurafsky, Daniel, and James H. Martin. 2022. Speech and Language Processing. 3rd draft. https://web.stanford.edu/~jurafsky/slp3/. (the authors generously made their book publicly available, check their website and use the latest version)
- NetworkX (the package's documentation and the references cited are the best place to start in terms of technical details)

Presentations and final projects from previous semesters

- 2023 spring: students' final projects
- 2019 fall

Grading

Assignments (TBD)

- A>= 95%, A- >= 90
- B+ >= 87%, B >= 83%, B- >= 80%
- C+ >= 77%, C >= 73%, C- >= 70%

- D+ >= 67%, D >= 63%, D- >= 60%

Policies

- Mental health: The instructor urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out if they feel comfortable. This will allow the instructor to provide any possible resources or accommodations. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515. You may also contact Bryce Moffett, LCSW (LBJ CARE counselor) at 512-232-4449 or stop by her office hours-Wednesday 1-2 pm SRH 3.119. Outside CMHC business hours (8a.m.-5p.m., Monday-Friday), contact the CMHC 24/7 Crisis Line at 512-471-2255.
- University Policies
- By taking this course (either for credit or auditing), you automatically authorize the instructor to use or cite the contents created by you for this course in the instructor's working book project. Appropriate academic principles of attribution and integrity will be followed.
- **License for Open Education:** This syllabus and all course content on this public website created by the instructor, TA, and students are licensed under the Creative Commons Attribution-Noncommercial 4.0 International License.
- **License of assignments:** For the assignment examples submitted by creators to OSF, the creators of submissions are the owners of their submissions. The owners grant CC BY 4.0 DEED to their submissions.

Acknowledgements

- 2022: The course is partly supported by the Teaching Innovation Grants 2022-23 from the Center for Teaching and Learning.
- 2019: The special events were supported by UT Austin Graduate School's Academic Enrichment Fund and RGK Center Special Funds for Data Science Speaker Series at the LBJ School of Public Affairs. Co-sponsors also include Center for East Asian Studies, UT Library Research Data

Services. The computing resource for the one-day data hackathon was supported by the XSEDE Educational Resources.

Computational Social Science Methods (PA397CIINF385T)

Prerequisites

Good with numbers

The way to happiness is simple, you just "Be good with numbers, and be good with peaple." - The Pursuit of Happyness (2006)

College level statistics. For example, be confident to test the difference between means, run hypothesis testing with probability theories, understand OLS and fixed effect models. Understanding how to flexibly use probability theories to do hypothesis testing is especially important.

Working knowledge of linear algebra. For example, you should know what is a vector and how to calculate the distance between two vectors.

Intermediate to advanced programming skills

Programming may be intimidating, but remember you are coding for social good and "Sometimes it's the people no one imagines anything of who do the things that no one can imagine." - The Imitation Game (2014)

Start coding today.

The class is Python based, but you can use R or any other programming language as long as you can complete the assignments and final project. R has its own advantages for sure, but I personally recommend Python because most of the state-of-the-art NLP implementations are in Python. Example Python packages used in this course: Pandas, Requests, regular expression, NetworkX, NLTK, TensorFlow, Keras, Transformers, and Gensim, etc.

Programming is an essential part of this course but not the purpose and will not be taught in this class. You are expected to have an intermediate to advanced level of programming skills before entering the class. At the minimum, you need to pass the following courses before registering this course (or you are confident that these modules are too easy):

Required fundamentals (no particular order)

You can write an email to me to request a free license for DataCamp. After completing the below modules, you should be familiar with all the topics listed in this tutorial.

- Programming:
  - Introduction to Python (4 hours)

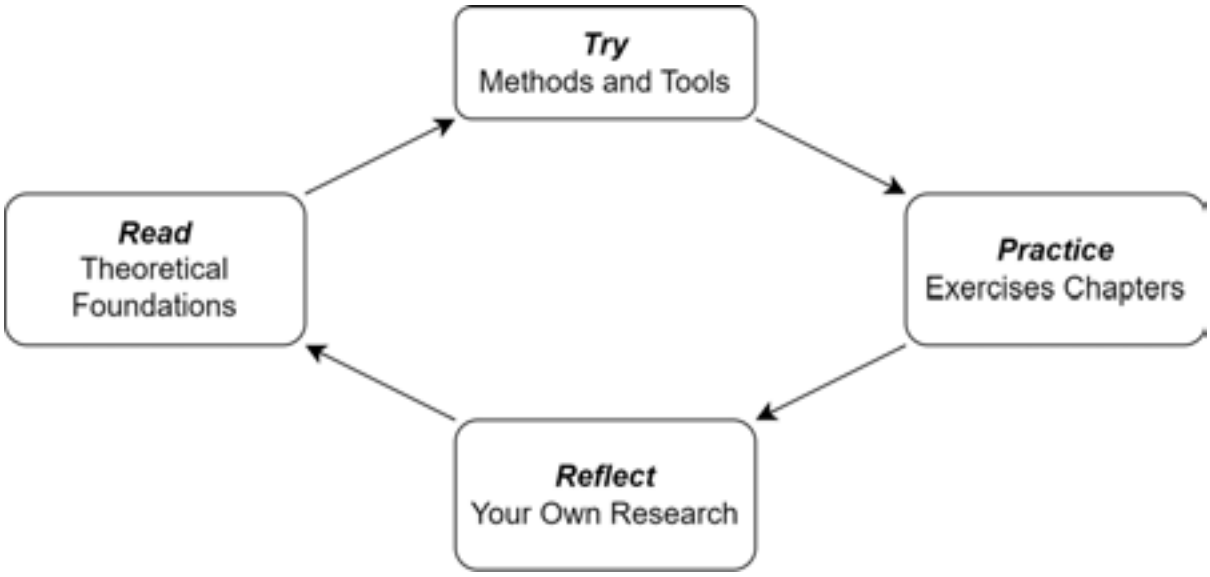
    - Intermediate Python (4 hours)
    - Introduction to Shell for Data Science (4 hours)
    - Writing Functions in Python
    - Writing Efficient Python Code
- Data preprocessing and exploratory analysis:
    - Exploratory Data Analysis in Python
    - Introduction to Importing Data in Python
    - Intermediate Importing Data in Python
    - Cleaning Data in Python
    - Joining Data with pandas
    - Data Manipulation with pandas

Recommended modules

- Introduction to Natural Language Processing in Python
- Web Scraping in Python
- Regular Expressions in Python
- Introduction to Data Visualization with Seaborn

Computational Social Science Methods (PA397CIINF385T)

Course Schedule



Reading Materials by Week

Introduction to the course

- Week 1 1/16: Course introduction
- Week 2 1/23: Computational Social Science: Why Research Design Approach

Data Management

- Week 3 1/30: Data Management: Methods and tools
- Week 4 2/6: Data Management: Background and Purposes (group presentation)
- Week 5 2/13: Data Management Exercise: Gathering Literature in Your Field

Concept Representation

- Week 6 2/20: Concept Representation: Background and Purposes (group presentation)
- Week 7 2/27: Concept Representation: Methods and tools
  - (Large) Language Models as concept representation tools
  - Topic modeling and classification
- Week 8 3/5: Concept Representation Exercise: Automated Coding

Data Analysis

- Week 9 3/19: Data Analysis: Background and Purposes (group presentation)
- Week 10 3/26: Data Analysis: Methods and tools
  - Network analysis as a representation and analysis method
  - Process of network analysis
  - Visualization tool: Gephi
- Week 11 4/2: Data Analysis Exercise: Simulation and Regression

Scientific Communication

- Week 12 4/9: Scientific Communication: Background and Purposes (group presentation)
- Week 13 4/16: Scientific Communication: Methods and tools
  - Programming language-based tools: Plotly and Dash for Python, Shiny for R
  - Off-the-shelf tools: Tableau, PowerBI, Excel.
- Week 14 4/23: Scientific Communication Exercise: Data Dashboard

---

Weekly Details

Week 1: Course introduction *Back2Top*

In class

- Course overview:
  - Context of this course.
  - Course sites: Syllabus website, Open Science Framework, Canvas.

- Helpful resources:
  - Open source communities (e.g., Stack Overflow)
  - ChatGPT. Discussion: How to effectively and responsibly use it? Your best practices.
  - CSS Empirical Studies Database. Discussion: Pick 2 studies of your interests, discuss with neighbors.

After class

- Complete readings for the upcoming week.
- Register accounts:
  - Open Science Framework
  - GitHub / Free GitHub Pro (GitHub Education)
  - Chameleon Cloud
- How to use Chameleon Cloud computing resources:
  - Review "Getting started with Chameleon Cloud''
- Assignment 2 sign up due on upcoming Monday.

Week 2: Computational Social Science: Why Research Design Approach *Back2Top*

Before class

- Required readings:
  - CSSPrimer: Chapter 1 and 2.

In class

- Discussion and lecture on readings. Key points:
  - Philosophical and epistemological fundamentals, research design overview, comparison between CSS and conventional approaches
  - Data management, concept representation, data analysis, and scientific communication
- In-class review and prepare:
  - Group presentations.
  - Empirical studies for analysis.

If time allows: High-performance cloud computing with Chameleon

- Start an instance on Chameleon Cloud
- Install Anaconda Python and Jupyter Notebook.
- Snapshot the instance as an image.
- You can also watch the video recordings below:
  - How to use Chameleon Cloud: Set up a new instance

After class

- Assignment 1 due on upcoming Monday.

- Review tools and platforms for upcoming week, prepare to discuss how you plan to use them.

Week 3: Data Management: Methods and tools *Back2Top*

Before class

- Review Assignment 3: Gathering Literature in Your Field

In class

- File and data format: API, JSON, and relational database.
- Efficiency and automation.
- Tools review:
  - OpenAlex
  - Draw.io
  - MySQL Workbench
- Prepare Assignment 3: Gathering Literature in Your Field

After class

- Group presentation slides and annotated bibliography due upcoming Monday.

Week 4: Data Management: Background and Purposes (group presentation) *Back2Top*

Before class

- Required readings
  - Leonelli, Sabina. "Scientific Research and Big Data." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.
  - Wickham, Hadley. "Tidy Data." The Journal of Statistical Software 59, no. 10 (2014). http://www.jstatsoft.org/v59/i10/.

- Goble, Carole, and David De Roure. "The Impact of Workflow Tools on Data-Centric Research." In The Fourth Paradigm: Data-Intensive Scientific Discovery, edited by Tony Hey, Stewart Tansley, Kristin Tolle, and Jim Gray. Microsoft Research, 2009. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.
- Fidler, Fiona, and John Wilcox. "Reproducibility of Scientific Results." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University, 2021. https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/.

In class

- Group presentation.
- Group discussion on annotated bibliography.
- Prepare Assignment 3: Gathering Literature in Your Field.

After class

- Assignment 3: Gathering Literature in Your Field due upcoming Monday.

Week 5: Data Management Exercise: Gathering Literature 1n Your Field *Back2Top*

Before class

- Complete the Assignment and submit to OSF.

In class

- Presentation and discussion of Assignment.
- Review peer assignments and provide feedback.
- Preview next exercise.

After class

- Revise assignments according to feedback.

- Research design analysis:
  - Prepare group presentation for next session.
  - Write annotated bibliography for next session.

Computational Social Science Methods (PA397CIINF385T)

Assignments

- Assignment 1: Plagiarism Test (5 points)
- Assignment 2: Group presentation and annotated bibliography (15 points)
- Assignment 3: Gathering Literature in Your Field (15 points)
- Assignment 4: Automated Coding (20 points)
- Assignment 5: Simulation and Regression (20 points)
- Assignment 6: Data Dashboards (15 points)
- Assignment 7: Participation and in-class assignments (10 points)

Late submissions are not accepted. Check due dates on Canvas.

Assignment 1: Plagiarisms test

The first assignment of this course is to pass the plagiarism test and obtain a certificate at the master and doctoral level. Plagiarism is a serious academic misconduct. You will receive zero grade on plagiarized work and there may be other consequences. We have been told not to do this maybe since primary school, and we are always assuming we know what plagiarism is. However, we may assume we know too much (e.g., famous cases of plagiarism).

You do not need to take this test if you have a comparable certification or you took this test before, but the validity of your certification needs to be approved by the instructor.

"All assignments in this course may be processed by Turnltln, a tool that compares submitted material to an archived database of published work to check for potential plagiarism. Other methods may also be used to determine if a paper is the student's original work. Regardless of the results of any

Turnltln submission, the faculty member will make the final determination as to whether or not a paper has been plagiarized" (Statement from the *Faculty Writing Committee: Guidelines for Preventing Plagiarism).*

For this assignment, please submit your certificate as a file on Canvas.

Assignment 2: Group presentation and annotated bibliography

(1) Group presentation

Sign up your presentation here.

Students will lead four short presentations on each of the research design themes: data management, concept representation, data analysis, and scientific communication.

I will provide some key definitions, frameworks, and references for you to start, you are expected to select and analyze empirical studies from the CSS Empirical Studies Database, and prepare a 30-minutes group presentation in class.

The presentation should respond to the following points:

[1] How to define this function of CSS methods from a research design perspective, and why it is necessary?

[2] What are the common technical methods or practices, how do they complement existing research approaches, and how are they unique?

[3] How do existing empirical studies apply specific methods, and how can these applications be improved? Select at least 2 articles from the CSS Empirical Studies Database and 2 articles of your own selection (add your article to the database).

[4] What are the general patterns or rationales you can abstract from your analysis?

For this assignment, submit your presentation slides to OSF.

- Analysis: Data Management
- Analysis: Concept Representation
- Analysis: Data Analysis
- Analysis: Scientific Communication

(2) Annotated bibliography

For the rest of the class who are not presenting, should complete a one-page annotated bibliography analyzing two empirical studies of your choice from the CSS Empirical Studies Database. The structure and purpose generally follow those of the presentation, with the expectation that the presentation and annotated bibliography may resonate with each other.

For this assignment, send your annotated bibliography to the presentation group.

Assignment 7: Participation

You are expected to present your Assignments 3-6 in class (no need to prepare slides). There may be small writing or review exercises helping you keep up with the course content. These are counted as participation. Details of these small exercises are listed in the schedule of each week.

No submission is required for this assignment.
