SOSC 4300 / SOSC 5500: Computational Social Science

Spring 2025

This version is prepared on January 20th, 2025

Lecture Time: Monday 18:00 - 20:50 **Seminar** Time: Tuesday 18:00 – 18:50

Lecture Room: 6602 **Seminar** Room: LSK1033

Course homepage: Canvas

Instructor

WEI, Jinlin, Room 2363, Academic Building, jinlinwei@ust.hk, Office Hour: Monday 15:00 – 16:00

Teaching Assistants

DENG, Xianglong, Room 3001, Academic Building, xdengaw@connect.ust.hk, Office Hour: TBD

LI, Jiajun, Room 3001, Academic Building, jliko@connect.ust.hk, Office Hour: TBD

Course Description

This course mainly focuses on two topics, text analysis and network analysis. It covers key topics such as supervised machine learning, unsupervised machine learning, centrality, and network formation. This course aims to provide introductions to new fields for senior undergraduate students and postgraduate students. Students will be expected to learn to apply basic text analysis and network analysis techniques to social science phenomena and focus on specific topics that they select to solve social science questions in which they have interests. The course employs a range of assessment methods, including coding assignments and group projects, to foster both collaborative learning and personal growth.

Prerequisites

• Students are expected to be familiar with the materials covered in basic statistics (e.g., SOSC 2400 for UG students and SOSC 5090 for PG students). Students with statistics knowledge but do not meet prerequisite can seek instructor's approval for enrollment.

• Students should also have basic literacy in at least one statistical programming language. We will use R and Python in tutorials. You can also use other programming languages such as Matlab, Julia, etc., as long as you can finish course assignments and projects with the

codes.

Intended Learning Outcomes (ILOs)

Upon finishing the course, students are expected to:

1. Describe the opportunities and challenges of social research in the age of big data.

2. Evaluate research on social phenomena from different fields, including social sciences and computer science/data science.

3. Practice the essential techniques to analyze social big data.

4. Propose research questions that are suited to be examined by computational methods with big data

5. Write a research article that utilizes the techniques and methods of computational social science to address social science problems, or design a project that use computational social science to address some real-world problems.

Assessment and Grading

- Your score will be assessed based on the following components (no mid-term and final exams):

i) Attendence and participation: 10%.

ii) Homework assignments: 10%×3. Due dates: March 3rd , March 31st , April 28th .

iii) Literature review: Report 10% Presentation 5% (5 min). In early April.

iv) Final Paper/Project Presentation 15% (15 - 20 min). Due date: April 28th.

v) Meet specific deadlines: Submit your proposal on March 10th (2%). Submit your preliminary results on April 14th (2%). Submit your final paper/ report on May 19th (1%).

vi) Final Paper/Project Write-up: 25%. Due date: May 19th.

Attendance and participation in class activities

- Based on class attendance and involvement in lectures. You are expected to show your understandings in the materials covered in the lectures or ask questions about the parts you did not understand.

Homework assignments

- There will be 3 assignments to test your knowledge of applying and evaluating basic computational social science algorithms.
- Each exercise is due in two weeks after the release of assignment.

- Discussing with other classmates and the usage of generative AIs are allowed. Do not commit plagiarism.
- We mark based on what is submitted by the due date.

Group work

- You can finish Literature Review and the Final Project in groups

– If there is any MPhil or PhD student in a group: max group size is 2

– Otherwise: 3 to 4 in a group (e.g., 4 UG in a group)

- Fix the Grouping by February 24th .

Literature review

Select a research topic and summarize how past researchers have used computational methods and/or big data to study this particular research area.

- Some examples of research areas:
  - Sociology: internal migration, international migration, social inequality, race and ethnicity relations, happiness,
  - Political science: government performance, government policy (and its effectiveness), election, social movements
  - Economics: measuring economic growth with big data
  - History: historical development of an idea
  - Psychology: measuring personality with big data
  - Communication and information science: content and spread of fake news/hate speeches
- You are recommended to select a research area that is similar to your final research paper. Students can discuss with instructors and TA for possible topics or feasibility.
- Your performances will be assessed by:
  - Presentation of literature review (5 minutes): each student/group needs to present their literature reviews in class.
  - Written Report: limit your report to 6 - 8 pages, 12 points, double space. Spell out clearly contributions of each group member in the first page of your report and sign electronically.
- Your written reports and presentations will be assessed based on the following 5 criteria. Each criterion is worth 1 point in the presentation and 2 points in the literature review.
  - What is the research area you have chosen, and why it's important or interesting
  - How people studied it traditionally (e.g., what data they use, what methods they use), and what are limitations of traditional methods/data?

- What are the advantages of using computational social science methods and data?
- What are the shortcomings of using computational social science methods and data?
- How you communicate with the audience and the reader.

• We will mark based on what is submitted by the due date.

Final paper/project

You can choose to write a research final paper, or a project that analyze a "real-world" social science problem. The differences between two options lie in their intended audience: research final paper should talk to researchers, while project talk to lay audience. The paper/project needs to be performed in the same group for presentation. It's recommended that you discuss your ideas with the instructor in early weeks of the course, during offices hours or through emails.

Research final paper

Choose a research topic and write a research paper using computational social science methods or digital data. This research article should follow the format of a standard research article, with the following components: introduction; review of past studies; research methods and data; results; conclusions. Consider the articles you read in class and for literature review as good examples of research articles.

Project

Focus on a real-world case. Develop a website or mobile app or software. Consider that you want to sell some social science ideas to layman using cool data analysis and visualization. Some ideas of cool demo/projects can be found here:

- https://projects.fivethirtyeight.com/
- https://github.com/matiasmascioto/awesome-soccer-analytics
- https://github.com/academic/awesome-datascience

Presentation

• Every group need to do a presentation (15 minutes): follow a standard presentation style for academic talks.

• Your presentations will be assessed based on the following 5 criteria. Each criterion is worth 3 points in your presentations.

- What is the research question you have chosen, and why it's important or interesting
- How people studied it traditionally (e.g., what data they use, what methods they

use), and what are limitations of traditional methods/data? How you studied it and what are your contributions?
- How you collected your data and the methods you used to solved your proposed question/ problem?
- How you interpret your results?
- How you communicate with the audience and the reader.

Final Paper

• If you are writing a final paper: limit your report to no more than 20 pages, 12 points, double space, including Tables, Figures and References.

• Your final papers will be assessed based on the following 5 criteria. Each criterion is worth 5 points in your final submissions.

- What is the research question you have chosen, and why it's important or interesting
- How people studied it traditionally (e.g., what data they use, what methods they use), and what are limitations of traditional methods/data? How you studied it and what are your contributions?
- How did you collect your data and the methods you used to solved your proposed question/ problem?
- How did you interpret your results?
- How you communicate with the audience and the reader.

Final Project

- If you are doing a project, you will have 5 more minutes in the presentation to receive questions and comments from your classmates.
- Your project will be assessed by your classmates based on the following 5 criteria. Each criterion is worth 3 points in your final grade.
  - What is the research question you have chosen, and why it's important or interesting
  - How people studied it traditionally (e.g., what data they use, what methods they use), and what are limitations of traditional methods/data? How you studied it and what are your contributions?
  - How you collected your data and the methods you used to solved your proposed question/ problem?
  - How you interpret your results?
  - How you communicate with the audience and the reader.

After receiving feedback from your classmates, please write them down and update them before you submit the final submission.

- Your final submission will be assessed by the teaching team based on how you improved your project in response to the comments of your classmates according to the following 5 criteria. Each criterion is worth 2 points in your final grade.
  - What is the research question you have chosen, and why it's important or interesting
  - How people studied it traditionally (e.g., what data they use, what methods they use), and what are limitations of traditional methods/data? How you studied it and what are your contributions?
  - How you collected your data and the methods you used to solved your proposed question/ problem?
  - How you interpret your results?
  - How you communicate with the audience and the reader.

Grading policies for the final paper/project

- We will mark based on what is submitted by the due date. Spell out clearly contributions of each group member and sign electronically. Each group member should submit the same copy via Canvas.
- Final papers will be checked by anti-plagiarism software. Students should take steps to avoid plagiarism and copying. For confirmed cases of plagiarism, severe sanctions including but not limited to a failure grade may be imposed.

Mapping of Course ILOs to Assessment Tasks

Assessed Task	Mapped ILOs	Explanation
Participation	ILO1, ILO2, ILO4	This task assesses students' understandings (ILO1) and abilities to evaluate (ILO2) the application of big data in social science researches and in solving real world problems. It also assesses students' ability to come up with potential solutions to social science problems. (ILO4).
Assignments	ILO1, ILO3	This task assesses students' understandings (ILO1) and abilities to use statistical software (ILO3) of applying big data in social science researches
Literature Review	ILO1, ILO2	This task assesses students' understandings (ILO1) of applying big data in social science researches in specific topics in which students have interests. It

		also assesses their ability to evaluate the importance of specific reseaches in corresponding topics (ILO2).
Group Paper/ Project	ILO1, ILO2, ILO3, ILO4, ILO5	This task assesses students' understandings (ILO1) of applying big data in social science researches in specific topics in which students have interests. It assesses their ability to evaluate the importance of specific reseaches in corresponding topics (ILO2) and propose a meaningful research question or a practical social problem to solve (ILO4). It assesses students' ability to collect data, coduct analyses (ILO3), and finally write a research article or design a project to address problems (ILO5).

Final Grade Descriptors

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates a deep understanding of the fundamental concepts and the application of methods in text analysis and network analysis. Exhibits exceptional critical thinking skills in evaluating the application of big data in solving social science research questions or practical social problems. Effectively communicates complex ideas about the development of the research using big data and consistently apply computational social science methods to enhance problem-solving.
B	Good Performance	Shows a solid grasp of the fundamental concepts and the application of methods in text analysis and network analysis. Demonstrates good critical thinking skills in evaluating the application of big data in solving social science research questions or practical social problems. Effectively communicates about the development of the research using big data and apply computational social science methods to support problem-solving.
C	Satisfactory Performance	Shows an adequate understanding of the fundamental concepts and the application of methods in text analysis and network analysis. Displays satisfactory critical thinking skills in evaluating the application of big data in solving social science research questions or practical social problems. Communicates about the development of the research using big data and apply computational social science methods to aid problem-solving, but may lack depth in analysis and application.

D	Marginal Pass	Has basic knowledge of the fundamental concepts and the application of methods in text analysis and network analysis. Shows limited critical thinking skills in evaluating the application of big data in solving social science research questions or practical social problems. Communicates about the development of the research using big data and apply computational social science methods for problem-solving, but with minimal effectiveness.
F	Fail	Demonstrate inufficient understanding of the fundamental concepts and the application of methods in text analysis and network analysis. Lack critical thinking skills in evaluating the application of big data in solving social science research questions or practical social problems. Struggles to communicate about the development of the research using big data and does not effectively apply computational social science methods for problem-solving.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Feedback on assignments will include comments on strengths and areas for improvement.

Students who have further questions about the feedback including marks should consult the instructor within five working days after the feedback is received.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST's Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Academic Integrity | HKUST -Academic Registry for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

Reading materials

All materials are suggestive. Suggestive readings makes it easier to follow the lectures. Optional readings are for students who are interested to read more on the topic. Search engines and Generative AIs are very useful if you would like to master specific topics.

Reference Books

1. Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.
2. Jackson, M. O. (2008). *Social and economic networks*. Princeton University Press.

Reference Articles

Introduction

Suggested readings

• Salganik, M. (2019). Bit by Bit: Social Research in the Digital Age. Princeton University Press. Chapter 1 and 2. This book can be freely accessible at https://www.bitbybitbook.com/en/1st-ed/preface/

• Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062

Optional readings

• Lazer, D. & Radford, J. (2017). Data ex Machina: Introduction to Big Data. Annual Review of Sociology, 43(1), 19–39

• Golder, S. A. & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. Annual Review of Sociology, 40(1), 129–152

Prediction: Algorithms

Suggested readings

- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. Science, 355(6324), 486–488
- Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., & Spector, T. D. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature Medicine, 26(7), 1037–1040

Optional Readings

- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 133(1), 237–293

Prediction: evaluation

Suggested reading

- Salganik, M. (2019). Bit by Bit: Social Research in the Digital Age. Princeton University Press. Chapter 3.

Optional readings:

- Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The Generalizability of Survey Experiments. Journal of Experimental Political Science, 2(02), 109–138
- Beauchamp, N. (2017). Predicting and Interpolating State-Level Polls Using Twitter Textual Data. American Journal of Political Science, 61(2), 490–503

Text Analysis: Basics

Suggested readings

- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. Annual Review of Political Science, 24(1)
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574

Optional readings:

- Wilkerson, J. & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annual Review of Political Science, 20(1), 529–544
- Denny, M. J. & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), 168–189
- Benoit, Kenneth (2020). Text as Data : An Overview. In L. Curini & Franzese, Robert (Eds.), The SAGE Handbook of Research Methods in Political Science and International Relations. SAGE Publications Ltd

Text Analysis: Dictionary and Supervised

Suggested readings

- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-

sourced Text Analysis: Reproducible and Agile Production of Political Data. American Political Science Review, 110(2), 278–295

- Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). Automated Text Classification of News Articles: A Practical Guide. Political Analysis, (pp. 1–24)

Text Analysis: Word Embeddings

Suggested reading

- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644

Optional readings

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13 (pp. 3111–3119). USA: Curran Associates Inc.
- Levy, O. & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems (pp. 2177–2185)
- Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in London's Old Bailey. Proceedings of the National Academy of Sciences, 111(26), 9419–9424

Text Analysis: Unsupervised Methods and Topic Models

Suggested reading

- Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. American Political Science Review, 113(4), 883–901

Optional readings

- Blei, D. M. (2012). Probabilistic Topic Models. Commun. ACM, 55(4), 77–84
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064–1082
- Blaydes, L., Grimmer, J., & McQueen, A. (2018). Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds. The Journal of Politics, 80(4), 1150–1167
- Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. Proceedings of the

National Academy of Sciences, 112(35), 10837–10844

- Bearman, P. (2015). Big Data and Historical Social Science. Big Data & Society, 2(2), 2053951715612497

Network: Basics

Suggested readings

- Jackson, Matthew O., Tomas Rodriguez-Barraquer, and Xu Tan. 2012. "Social Capital and Social Quilts: Network Patterns of Favor Exchange." *American Economic Review*, 102 (5): 1857–97**.**

Network: Centrality and Influence Measures

Suggested readings

- Bloch, F., Jackson, M. O., & Tebaldi, P. (2023). Centrality measures in networks. *Social Choice and Welfare*, *61*(2), 413-453.
- Gallea, Q., Morelli, M., & Rohner, D. (2022). Power in the Pipeline. *arXiv preprint arXiv:2210.03572*.

Network: Network Formation

Suggested readings

- Ruiz-García, M., Ozaita, J., Pereda, M., Alfonso, A., Brañas-Garza, P., Cuesta, J. A., & Sánchez, A. (2023). Triadic influence as a proxy for compatibility in social relationships. *Proceedings of the National Academy of Sciences*, *120*(13), e2215041120.
- Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, *10*(1), 1017.

Image data

Suggested readings

- Zhang, H. & Pan, J. (2019). CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. Sociological Methodology, 49(1), 1–57
- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences, 114(50), 13108–13113