SOSC 4300 / SOSC 5500: Computational Social Science

Spring 2023

Lecture Time: Monday 9:00 - 11:50

Lecture Room: LSK 1033

This version prepared on Jan 27, 2023

	Instructor	Teaching Assistant
	ZHANG, Han	**CHEN, Pei**
Office	Room 2379, Academic Building, Lift 15	Room 3001, Academic Building, Lift 4
Email	zhangh@ust.hk	pchenam@connect.ust.hk
Office Hour	Monday 16:00 - 17:00	TBD

- Course homepage is: `https://github.com/HKUST-SOSC4300-5500/`
- Syllabus will be regularly updated at `https://github.com/HKUST-SOSC4300-5500/Lecture-Material`

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 assignment and projects with the codes.

Goals

Upon finishing the course, students should be able 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.

Grading

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

	%	Due
Attendence and participation	10%	
Homework assignments	30%	Two weeks
Literature review		
Report	5%	
Presentation	5% (5 min)	mid April
Final Paper/Project		
Presentation	15% (15 min)	May 8
Write-up	35%	May 22

Homework, literature review, and final paper/project need to be submitted on Github. We will cover the steps in the first lecture.

Grouping

- You should finish all tasks 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)
- Finish grouping by **Week 3**

Attendance and participation in class activities

- Based on class attendance and involvement in lecture and tutorial. you are expected to be either able to answer questions about the assigned readings or ask questions about the parts you did not understand. If you are uncomfortable speaking up in class, send the question in Zoom's chat window, post them on Canvas's discussion forums, come to my office hours, or send your questions via e-mail.

Homework assignments

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

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 areas that are similar to your final research paper. Students can discuss with instructors and TA for possible topics or feasibility.
- Your performances will be accessed by:
  - **Written Report**: limit your report to **8 pages, 12 points, double space**. Spell out clearly contributions of each group member in the first page of your report.
  - **Presentation of literature review** (**5 minutes**): each student/group needs to present their literature reviews in class.
- Include the following items in written report and presentations:
  - 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?

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. You can also brower the GitHub repo and see the projects of previous year's students.

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`

Evaluation

- Every group need to do a presentation (**15 minutes**): follow a standard presentation style for academic talks.
- If you are writing a final paper: limit your report to **20 pages, 12 points, double space, including Tables, Figures and References**. You can write fewer pages if you feel necessary.
- If you are doing a project: based on how your classmates like your website/app/software.

Grading policies

- **Late delivery** of due items will be marked down 75% if received within 1 day of the due date, and 50% if received within 3 days of the due date; you will receive zero credit if the due item is not delivered within 3 days of the due date. Contact the instructor if there are rare unforeseen circumstances.
- If you want to dispute a grade, please submit your argument in writing along with your assignment. We will evaluate the merit of your argument as well as perform a full reassessment of your entire assignment. This means that your grade may end up lower than it was originally.
- Final papers are 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.

Schedule (Tentative)

Week	Date	Topic
1	*[2023-02-06 Mon]*	Introduction; big data
2	*[2023-02-13 Mon]*	Prediction;
3	*[2023-02-20 Mon]*	Prediction; Evaluation
4	*[2023-02-27 Mon]*	Text (I)
5	*[2023-03-06 Mon]*	Text (II); supervised
6	*[2023-03-13 Mon]*	Text (III); embedding
7	*[2023-03-20 Mon]*	Text (IV); unsupervised
8	*[2023-03-27 Mon]*	Network; basics
9	*[2023-04-03 Mon]*	Network; small worlds
10	*[2023-04-10 Mon]*	**NO CLASS (mid-term break)**
11	*[2023-04-17 Mon]*	Causal Inference and Big Data: network as example
12	*[2023-04-24 Mon]*	Image data (or other elective topics)
13	*[2023-05-01 Mon]*	**NO CLASS (holiday)**
14	*[2023-05-08 Mon]*	Presentation

Weekly reading material

The course materials will be drawn from lecture slides and assigned readings. Readings are available at Canvas. You are **required to read the readings before the start of each class**. Optional readings are for students who are interested to read more on the topic.

Lecture 1: Digital Traces

- 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

Lecture 2: Prediction; algorithms

- 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

Lecture 3: Prediction; evaluation

- 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

Lecture 4: Text (I); basics

- Grimmer, J. & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(03), 267–297

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
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535–574
- Goldberg, Y. (2015). A Primer on Neural Network Models for Natural Language Processing. *arXiv:1510.00726 [cs]*

Lecture 5: Text (II); Dictionary and Supervised

- 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)

Lecture 6: Text (III) Word Embeddings

- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5), 905–949
- 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

Lecture 7: Text (IV); Unsupervised Methods and Topic Models

- 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

Lecture 8: Network, small world, and agent-based modeling

- Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational Social Science. *Science*, 323(5915), 721–723

Optional readings

- Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press
- Schelling, T. C. (2006). *Micromotives and Macrobehavior*. W. W. Norton & Company

Lecture 9: Network, strength of ties, and diffusion

- Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996), 1194–1197

Lecture 10: Causal inference with big data

- Christakis, N. A. & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4), 370–379
- Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1), 68–72
- Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 (pp. 519–528). New York, NY, USA: ACM

Lecture 11: Image data

- 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