SOCS0055: Advanced Computational Techniques for Data Science

Overview

This module teaches students the central skills of modern computational social science. The first half of the course covers the necessary programming skills, including data preparation, producing outputs from statistical analyses (tables, text, and figures), writing functions and loops, and working with data from the internet. The second half of the course covers the statistical techniques and conceptual underpinnings of computational social science, including optimising for prediction (compared with inference), bias-variance trade-off, cross-validation and parameter tuning, machine learning (e.g., random forests) and deep learning (neural networks) algorithms, and ensemble methods.

General Information

Module code	SOCS0055
Lecturers	Liam Wright (liam.wright@ucl.ac.uk) Tobias Rüttenauer (t.ruttenauer@ucl.ac.uk)
Lecture	Tue 4pm–6pm: IOE - Bedford Way (20) 102 - Punnett Hall

Main Readings

Weeks 1-5 of the course will follow Hadley Wickham et al.'s textbook R for Data Science (2nd Edition). Weeks 6-10 of the course will follow Boehmke & Greenwell's Hands-On Machine Learning with R. Both are available for free online in the provided links. References to the relevant chapters for a given week are provided in Section 7 (Course Outline) and on Moodle. Lectures will also include references to further reading. This reading is optional but will cement and extend students' knowledge.

- Week 1: **Introduction to the Course**
  - Overview to the content and motivation of the course.
  - Introduction to the R programming language and the tidyverse ecosystem.
  - Introduction to Understanding Society, the dataset used throughout the course

  Literature:

  - An Introduction to R (Wright, 2020)

  See also:

  - Grolemund, Garrett. 2014. Hands-On Programming with R.
  - Wickham, Hadley. 2017. Advanced R.

- Week 2: **Wrangling Data for Analysis**
  - Overview of the steps to import, explore, clean, and reshape code to prepare for statistical analysis
  - Background on (ir)reproducibility in scientific research and advice on how to avoid mistakes in analysis

  Literature:

  - Chapters 3 and 5 of R for Data Science

  See also:

  - Crüwell, S., Van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., and Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science: An Annotated Reading List. *Zeitschrift für Psychologie*, 227(4): 237–248.
  - The tidyverse team. 2025. Tidyverse style guide.
  - Vable, A. M., Diehl, S. F., and Glymour, M. M. (2021). Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research. *American Journal of Epidemiology*, 190(10): 2172–2177.

- Week 3: **Creating Publication-Ready Statistical Outputs**
  - Discussion of the main ways that statistical results can be presented and how, with good design, these can aid understanding and interpretation
  - Introduction to tidy data and R packages and functions which extract results from statistical analysis (e.g., regression coefficients) or present these in formats (tables, figures, text) ready for publication

- An introduction to the layered grammar of graphics as implemented in the tidyverse package ggplot2

Literature:

- Chapter 9 of R for Data Science. Chapter 3 of Keiran Healy,Äôs Data Visualization.

See also:

- Gohel, David. 2025. Using the flextable R package.
- Wilke, Claus. 2019. Fundamentals of Data Visualisation.
- Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen. 2025. ggplot2: Elegant Graphics for Data Analysis (3rd Edition).

- Week 4: **Repeating Yourself – Functions and Iterative Programming**
  - Introduction to writing functions in R and using iterative programming (loops and the purrr::map() family of functions) to repeat repetitive tasks with minimal code.
  - Discussion of Specification Curve Analysis and other ,Äòmany model,Äô type approaches

  Literature:

  - Chapters 25 and 26 of R for Data Science.

  See also:

  - Grolemund, Garrett. 2014. Hands-On Programming with R.
  - Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11): 1208–1214.
  - Wright, L. (2023). Many Models in R: A Tutorial.

- Week 5: **Working with Data from the Internet**
  - Background on modern data sources used in computational social science
  - Introduction to web-scraping and accessing data through application programming interfaces (APIs)
  - Introduction to working with structured text formats
  - Discussion of approaches for working with big (larger-than-memory) datasets

  Literature:

  - Chapters 23 and 24 of R for Data Science.

- Week 6: **Inference vs. Prediction**
  - Inference vs. Prediction
  - Supervised vs. unsupervised learning
  - Parametric vs. non-parametric methods
  - Bias-variance trade-off

  Literature:

  - Chapter 1 of Hands-On Machine Learning.

  See also:

  - Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24(1): 395–419.
  - Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, 105(5): 491–495.
  - Molina, M. D., Chau, N., Rodewald, A. D., and Garip, F. (2023). How to model the weather-migration link: A machine-learning approach to variable selection in the Mexico-U.S. context. *Journal of Ethnic and Migration Studies*, 49(2): 465–491.

- Week 7: **Feature Selection and Automated Regression**
  - Continuous predictions and loss functions
  - Regularization (Lasso, Ridge, Elastic Net)
  - Cross-validation
  - Re-sampling and bootstrapping
  - Multiverse analysis

  Literature:

  - Chapters 2 and 6 of Hands-On Machine Learning.

  See also:

  - Engzell, P. and Mood, C. (2023). Understanding Patterns and Trends in Income Mobility through Multiverse Analysis. *American Sociological Review*, 88(4): 600–626.
  - James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer International Publishing, Cham, Chapter 6.
  - Muñoz, J. and Young, C. (2018). We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness. *Sociological Methodology*, 48(1): 1–33.
  - Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5): 702–712.

- Week 8: **Tree-Based Methods and Ensembles**
  - Decision Trees
  - Random Forests
  - Bagging and Boosting
  - Hyperparameter Tuning

  Literature:
  - Chapters 9 and 11 of Hands-On Machine Learning. Chapters 10 and 12 for Bagging and Boosting.

  See also:
  - Chen, T. and Guestrin, C. (2016). XGBoost. In Krishnapuram, B., Shah, M., Smola, A., Aggarwal, C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794, New York, NY, USA. ACM.
  - Hare, C. and Kutsuris, M. (2023). Measuring Swing Voters with a Supervised Machine Learning Ensemble. *Political Analysis*, 31(4): 537–553.
  - James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer International Publishing, Cham, Chapter 8.
  - Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, - Chapters 9 and 10.
  - Molina, M. D., Chau, N., Rodewald, A. D., and Garip, F. (2023). How to model the weather-migration link: A machine-learning approach to variable selection in the Mexico-U.S. context. *Journal of Ethnic and Migration Studies*, 49(2): 465–491.

- Week 9: **Neural Networks**
  - Nural networks
  - Deep learning
  - Stacked models & Superlearners

  Literature:
  - Chapters 13 and 15 of Hands-On Machine Learning.

  See also:
  - James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer International Publishing, Cham, Chapter 10.
  - Gaskin, T. and Abel, G. J. (2025). Deep learning four decades of human migration. *SocArXiv*.
  - Muchlinski, D., Yang, X., Birch, S., Macdonald, C., and Ounis, I. (2021). We need to go deeper: Measuring electoral violence using convolutional neural networks and social media. *Political Science Research and Methods*, 9(1): 122–139.

- Week 10: **Getting the Explanation out of the Prediction**
  - Causal machine learning
  - Feature importance
  - Shapley explanation values

  Literature:

  - Chapter 16 of Hands-On Machine Learning.

  See also:

  - Brand, J. E., Xu, J., Koch, B., and Geraldo, P. (2021). Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning. *Sociological Methodology*, 51(2): 189–223.
  - Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., and Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866): 181–188.
  - Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30: 1–10.
  - Verhagen, M. D. (2024). Incorporating Machine Learning into Sociological Model-Building. *Sociological Methodology*, 54(2): 217–268.