Introduction to Computational Social Science

Alfredo Hernandez Sanchez

alfredo.sanchez@tspmi.vu.lt

Vilnius University

June 11, 2026

Course Information

Course title: Introduction to Computational Social Science
Level: Undergraduate
Semester: Spring 2027
Office hours: By appointment

This syllabus is a working plan for a project-based undergraduate course. It was produced in the context of the European Union Marie Skłodowska-Curie Postdoctoral Fellowships WIDERA project FIRSA under grant agreement No. 101180601.

Course Description

Computational social science uses data, code, and social theory to study political and social life at scales that were difficult to observe only a few years ago. This course introduces students to that field through substantive questions about politics, public policy, international relations, social media, institutions, and economic governance. Students will learn how to collect, clean, visualize, model, and interpret social data using the R programming language and reproducible Quarto documents.

The course gives special attention to **text as data**. Political life is made of speeches, manifestos, policy documents, news articles, social media posts, parliamentary debates, central bank communications, NGO reports, and other forms of language. We will learn how to transform texts into analyzable data without forgetting that texts are produced by people, institutions, and unequal social contexts. Students will work with methods including corpus construction, tokenization, descriptive text statistics, dictionaries, supervised classification, topic models, scaling, and introductory large language model workflows.

Learning Outcomes

By the end of the course, students should be able to:

- Explain what makes computational social science distinct from, and similar to, other forms of empirical social inquiry.
- Use open-source analytical tools like R, Python, and Quarto to create reproducible analyses.
- Collect, clean, summarize, analyze, and visualize common forms of social data.
- Read basic statistical output and explain uncertainty, association, prediction, and model limita-

tions in plain language.
- Build a robust original dataset and document choices about sources, units of analysis, metadata, preprocessing, and ethics.
- Apply and evaluate introductory text-as-data methods, including dictionaries, document-feature matrices, topic models, supervised classification, and scaling.
- Connect computational results to substantive social science theory.
- Communicate findings through a polished research report, clear visualizations, and a short presentation.

Course Materials

There is no single required textbook. Core readings will come from articles, book chapters, online books, and technical tutorials. The most frequently used general references are:

- Grimmer, Roberts, and Stewart (2022), for the conceptual logic of text as data in the social sciences.
- Wickham (2014) and the broader tidyverse ecosystem, for organizing data analysis around clear data structures.
- Silge and Robinson (2016), for the tidy approach to text mining in R.

Students will also use open online documentation for R, RStudio, Quarto, tidyverse, ggplot2, tidytext, quanteda, and related packages.

Software

We will mostly use:

- R and RStudio.
- Quarto.
- Python.

No prior programming experience is assumed, but students should expect to practice regularly. Early labs are designed to reduce setup friction and build confidence with basic workflows.

Course Structure

Most weeks follow the same pattern:

- **Lecture:** core concepts, research design, and examples from published work.
- **Discussion:** one empirical paper or methodological debate.
- **Lab:** guided work in R, usually connected to the final project.
- **Project step:** a short deliverable that moves students toward the final analysis.

Students should leave each week with something concrete: a script, a figure, a cleaned dataset, a coding scheme, a validation plan, or a paragraph of interpretation.

Assessment

Assignment	Weight	Purpose
Participation and discussion leadership	15%	Practice reading empirical work, asking useful questions, and explaining methods.
Weekly labs and check-ins	20%	Build fluency with R, Quarto, data wrangling, visualization, and text workflows.
Mini project 1: Data story	15%	Produce a reproducible visualization-based analysis of a social or political dataset.
Mini project 2: Text measurement memo	15%	Build a small corpus and compare two ways of measuring a concept in text.
Final project proposal and annotated corpus	10%	Define a research question, identify data, and justify the corpus and measurement strategy.
Final project report and presentation	25%	Complete an original computational social science analysis.

Final Project

The main assignment is a semester-long research project. Students will choose a substantive question, collect or curate data, create a reproducible analysis, and present results in a short research report. Projects may focus on political speeches, party manifestos, news coverage, international organization documents, social media posts, central bank communication, public consultations, NGO reports, policy documents, or another approved corpus.

The final report should include:

- A clear research question and motivation.
- A short connection to relevant social science literature.
- A description of the data source, corpus, unit of analysis, metadata, and sampling choices.
- A transparent account of cleaning, preprocessing, and measurement.
- At least two meaningful visualizations.

- One primary computational method and a validation or robustness check.
- A discussion of what the method can and cannot show.
- Reproducible code and a Quarto report.

AI and Academic Integrity

Students may use AI tools for brainstorming, debugging, translation, summarizing documentation, and improving prose. They may not use AI tools to fabricate sources, invent results, conceal missing work, or submit analysis they do not understand. Any substantial use of AI must be briefly disclosed in the assignment. The disclosure should state what tool was used, for what purpose, and how the student checked the output.

In text-as-data work, AI systems can be research instruments as well as writing assistants. When they are used for classification, annotation, or interpretation, they must be treated like any other measurement tool: their prompts, settings, inputs, outputs, and validation strategy must be documented.

> *Code does not need to be elegant at first, but it must be readable, reproducible, and documented well enough that another person could follow the analysis. Students are encouraged to revise. Computational work often improves through iteration.*

Weekly Schedule

Part I: Foundations

Week 1: What Is Computational Social Science?

Theme: Data, scale, interpretation, and the social science question.
Skills: RStudio orientation, Quarto documents, project folders, and reproducible habits.
Discussion reading: Grimmer, Roberts, and Stewart (2022), introduction; Ziems et al. (2024).
Lab: Create a Quarto notebook and reproduce a small figure from a provided dataset.
Project step: Write three possible research questions and identify one possible dataset or corpus for each.

Week 2: Data, Measurement, and Tidy Workflows

Theme: Concepts, measurement validity, units of analysis, and tidy data.
Skills: Importing data, variable types, missingness, joins, and grouped summaries.
Discussion reading: Wickham (2014); Adcock and Collier (2001).
Lab: Clean a messy social science dataset and produce a short data dictionary.
Project step: Define the unit of analysis for a possible final project.

Week 3: Seeing Data Clearly

Theme: Visualization as analysis, not decoration.
Skills: `ggplot2`, visual encodings, comparison, uncertainty, and visual ethics.
Discussion reading: selected chapters from Healy's *Data Visualization* and examples from contemporary political data journalism.

Lab: Build and revise three plots: distribution, comparison, and change over time.
Project step: Submit one exploratory figure with a 150-word interpretation.

Week 4: Basic Statistics for Computational Work

Theme: Description, association, uncertainty, and model humility.
Skills: Correlation, simple regression, confidence intervals, prediction errors, and residual plots.
Discussion reading: a short applied paper selected in class, with attention to how statistical evidence supports the argument.
Lab: Estimate and visualize a simple model; translate coefficients into substantive claims.
Project step: Identify one comparison, pattern, or relationship that your project might test or describe.

Week 5: Data Collection, APIs, Scraping, and Research Ethics

Theme: Where computational data comes from, who is represented, and who is missing.
Skills: Reading files from the web, simple scraping, metadata, sampling, de-identification, and documentation.
Discussion reading: Campbell et al. (2023); Kern and Mustasilta (2023).
Lab: Collect a small public corpus or structured dataset and create a reproducible data log.
Project step: Submit a data source memo.

Part II: Text As Data

Week 6: From Documents to Data

Theme: Texts as social artifacts and measurable objects.
Skills: Corpus construction, tokens, types, stop words, stemming, lemmatization, n-grams, document-feature matrices.
Discussion reading: Grimmer, Roberts, and Stewart (2022), chapters on text representation; Silge and Robinson (2016).
Lab: Build a corpus from political or policy documents and compare tokenization choices.
Project step: Submit a corpus plan with sources, inclusion rules, and unit of analysis.

Week 7: Describing Texts

Theme: What can word counts show, and what do they hide?
Skills: Frequency plots, tf-idf, keywords-in-context, collocations, readability, lexical diversity.
Discussion reading: Baturo, Dasandi, and Mikhaylov (2017).
Lab: Create a descriptive text profile of a corpus and write a measurement caveat.
Project step: Produce two descriptive text visualizations for your corpus.

Week 8: Dictionaries, Categories, and Human Judgment

Theme: Turning concepts into word lists and codes.
Skills: Dictionary construction, validation samples, intercoder agreement, confusion matrices, and transparent coding rules.
Discussion reading: Tausczik and Pennebaker (2010); Benoit et al. (2016).
Lab: Apply an existing dictionary, revise it for a new context, and test what changes.

Project step: Submit mini project 2, the text measurement memo.

Week 9: Supervised Classification

Theme: Using labeled examples to classify new texts.
Skills: Training and test data, feature engineering, naive Bayes or logistic classifiers, accuracy, precision, recall, and error analysis.
Discussion reading: Laurer et al. (2024); Burnham (2024).
Lab: Train a simple classifier and inspect misclassified documents.
Project step: Decide whether supervised classification is useful for your final project and justify the choice.

Week 10: Topic Models and Inductive Discovery

Theme: Finding themes in large corpora without pretending that topics interpret themselves.
Skills: Topic modeling, choosing the number of topics, labeling topics, topic prevalence, and validation.
Discussion reading: Chang et al. (2009); Chen et al. (2023).
Lab: Estimate a topic model and compare model output with close reading of selected documents.
Project step: Submit final project proposal and annotated corpus.

Week 11: Scaling, Positions, and Political Meaning

Theme: Estimating latent positions from text.
Skills: Text scaling intuition, reference texts, unsupervised scaling, interpretation, and sensitivity checks.
Discussion reading: Di Leo et al. (2025); Ornstein, Blasingame, and Truscott (2025).
Lab: Use a small manifesto or speech corpus to compare estimated positions across actors or time.
Project step: Write a short validation plan for your final analysis.

Part III: Extensions, Synthesis, and Projects

Week 12: Networks, Images, and Multimodal Social Data

Theme: Beyond rectangular datasets and bags of words.
Skills: Co-occurrence networks, actor-document networks, basic network plots, and visual/multimodal evidence.
Discussion reading: Kohavi (2026) and one applied network or multimodal text analysis paper selected in class.
Lab: Build a word co-occurrence or actor-document network from a text corpus.
Project step: Submit a draft results figure for peer feedback.

Week 13: Large Language Models in Social Research

Theme: LLMs as assistants, classifiers, annotators, and uncertain measurement instruments.
Skills: Prompt documentation, zero-shot and few-shot classification, model comparison, reproducibility, and ethical limits.
Discussion reading: Ziems et al. (2024); Spirling (2023); Benoit et al. (2026).
Lab: Compare human labels, dictionary scores, and LLM-assisted labels on a small validation

sample.
Project step: Submit a methods and limitations draft.

Week 14: Project Workshop and Research Communication

Theme: Making computational results understandable, defensible, and useful.
Skills: Tables, captions, slide design, narrative structure, executive summaries, and reproducibility checks.
Discussion reading: Birkenmaier, Lechner, and Wagner (2024).
Lab: Peer review of final projects; reproducibility swap.
Project step: Final presentation.

Major Deadlines

Week	Deliverable
3	Exploratory visualization
5	Data source memo
6	Mini project 1: Data story
8	Mini project 2: Text measurement memo
10	Final project proposal and annotated corpus
12	Draft results figure
13	Methods and limitations draft
14	Final presentation
Exam period	Final project report and reproducible Quarto submission

Discussion Leadership

Each week, a student will help lead one discussion of an empirical or methodological paper. The discussion leader should help the class understand:

- What question the authors ask.
- What data they use and how the data were created.
- What method connects the evidence to the claim.
- What assumptions the method requires.
- What the paper teaches us about doing computational social science.

Course Policies

Participation

Participation means preparation, attention, and constructive engagement. Some students participate best by speaking in class; others do so through careful peer feedback, lab collaboration, written questions, or office hour conversations. All of these count when they contribute to the learning environment.

Late Work

Short extensions are possible when requested in advance. Work submitted late without prior arrangement may receive a penalty. The main priority is to keep students moving through the project sequence.

Accessibility

Students who need accommodations should contact the instructor as early as possible.

References

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46. https://doi.org/10.1017/S0003055401003100.

Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov. 2017. "Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus." *Research & Politics* 4 (2): 1–9. https://doi.org/10.1177/2053168017712821.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. https://doi.org/10.1017/S0003055416000058.

Benoit, Kenneth, Scott De Marchi, Conor Laver, Michael Laver, and Jinshuai Ma. 2026. "Using Large Language Models to Analyze Political Texts Through Natural Language Understanding." *American Journal of Political Science*, March, ajps.70050. https://doi.org/10.1111/ajps.70050.

Birkenmaier, Lukas, Clemens M. Lechner, and Claudia Wagner. 2024. "The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation." *Communication Methods and Measures* 18 (3): 249–77. https://doi.org/10.1080/19312458.2023.2285765.

Burnham, Michael. 2024. "Stance Detection: A Practical Guide to Classifying Political Beliefs in Text." *Political Science Research and Methods* 12 (3): 597–604. https://doi.org/10.1017/psrm.2024.35.

Campbell, Rebecca, McKenzie Javorka, Jasmine Engleton, Kathryn Fishwick, Katie Gregory, and Rachael Goodman-Williams. 2023. "Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for de-Identifying Sensitive Narrative Data." *Advances in Methods and Practices in Psychological Science* 6 (4): 25152459231205832. https://doi.org/10.1177/25152459231205832.

Chang, Johnathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In. Vancouver. http://cs.umd.edu/~jbg//docs/nips2009-rtl.pdf.

Chen, Yingying, Zhao Peng, Sei-Hill Kim, and Chang Won Choi. 2023. "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review." *Communication Methods and Measures* 17 (2): 111–30. https://doi.org/10.1080/19312458.2023.2167965.

Di Leo, Riccardo, Chen Zeng, Elias Dinas, and Reda Tamtam. 2025. "Mapping (a)ideology: A Taxonomy of European Parties Using Generative LLMs as Zero-Shot Learners." *Political Analysis*, April, 1–8. https://doi.org/10.1017/pan.2025.7.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton; Oxford: Princeton University Press.

Kern, Florian G, and Katariina Mustasilta. 2023. "Beyond Replication: Secondary Qualitative Data Analysis in Political Science." *Comparative Political Studies* 56 (8): 1224–56. https://doi.org/10.1177/00104140221139388.

Kohavi, Noya. 2026. "Making Sense of Memes: How to Study Distributed Knowledge as Lay Economics." *Review of International Political Economy*, June, 1–33. https://doi.org/10.1080/09692290.2026.2681767.

Laurer, Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." *Political Analysis* 32 (1): 1–16. https://doi.org/10.1017/pan.2022.15.

Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13 (2): 264–81. https://doi.org/10.1017/psrm.2024.64.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in r." *The Journal of Open Source Software* 1 (3): 37. https://doi.org/10.21105/joss.00037.

Spirling, Arthur. 2023. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." *Nature* 616 (7957): 413. https://doi.org/10.1038/d41586-023-01207-9.

Tausczik, Yla R., and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54. https://doi.org/10.1177/0261927X09351676.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). https://doi.org/10.18637/jss.v059.i10.

Ziems, Caleb, William Held, Omar Shaikh, Jianshu Chen, Zeyu Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50 (1): 237–91. https://doi.org/10.1162/coli_a_00464.