# Machine Learning for Social Science

## 2-hour seminar

Maël Lecoursonnais

2025-12-02

# Overview

- Duration: **2 hours**
- Structure:
  1. Assign roles
  2. Short summaries
  3. Small-group discussion
  4. Feedback
  5. (If we have time) Mini-debate
  6. Wrap-up

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.
- Group responsibilities:

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.
- Group responsibilities:
  - Small-group discussion (30-35 min)

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.
- Group responsibilities:
  + Small-group discussion (30-35 min)
  + Prepare for take-aways (5-10 min)

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.
- Group responsibilities:
  + Small-group discussion (30-35 min)
  + Prepare for take-aways (5-10 min)
  + Discussion (10-15 min)

# Logistics & roles (0–5 min)

- **3/4 groups** (4-5 students each).
- Each group must include **at least one** reader of each paper.
- Group responsibilities:
  - Small-group discussion (30-35 min)
  - Prepare for take-aways (5-10 min)
  - Discussion (10-15 min)
  - (Optional) Mini-debate (15 min)

# Readings

- Salganik et al. (2020), "Measuring the predictability of life outcomes with a scientific mass collaboration" (*PNAS*)
- Torres & Cant (2022), "Learning to see: Convolutional neural networks for the analysis of social science data" (*Political Analysis*)
- Brand et al. (2021), "Uncovering sociological effect heterogeneity using tree-based machine learning" (*Sociological Methodology*)

# Short summaries

## Salganik et al. (2020)

- 160 teams predicted 6 life outcomes using rich longitudinal data
- Even optimized ML models showed low predictive accuracy
- Predictions only slightly outperformed simple benchmarks
- Errors varied more by *families* than modeling techniques
- Highlights value of mass scientific collaboration

# Short summaries

## Torres & Cant (2022)

- Introduces CNNs for visual data classification in social sciences
- Automates tedious image coding tasks
- Example: handwriting classification in vote tallies
- Shows usefulness for researchers and policy practitioners
- Discusses implementation challenges and limitations

# Short summaries

## Brand et al. (2021)

- Individuals respond differently to treatments (effect heterogeneity)
- Traditional subgrouping often biased or limited by priors
- Causal trees uncover previously unconsidered subgroups
- Case study: heterogeneity in college effects on wages
- Uses causal trees with confounding adjustments (IPW, matching, doubly robust)
- Encourages systematic data-driven exploration of heterogeneity

# Small-group work — Instructions

1. (Optional but recommended) Appoint a note-taker to summarize key points.

# Small-group work — Instructions

1. (Optional but recommended) Appoint a note-taker to summarize key points.
2. Discuss the set of *general* questions.

# Small-group work — Instructions

1. (Optional but recommended) Appoint a note-taker to summarize key points.
2. Discuss the set of *general* questions.
3. Prepare to share **take-aways** with the larger group.

# Small-group guiding questions

0. **Framing & goals**:
   - What is the main research question or problem each paper addresses?
   - How does ML help tackle that question compared to traditional methods?

# Small-group guiding questions

0. **Framing & goals**:
   - What is the main research question or problem each paper addresses?
   - How does ML help tackle that question compared to traditional methods?

1. **Prediction vs. explanation**: All three papers use ML but for different goals.
   - Would you classify each paper as primarily focused on $\hat{y}$ (getting the best prediction) or on $\hat{\beta}$ (estimating effects)? Why?
   - What is "good prediction" in social science? Is it the same as a "good model," or does it risk shifting our attention away from causal inference?
   - What are the implications of prioritizing prediction over explanation for social science research more broadly?

# Small-group guiding questions

2. **Data, complexity, and social theory**: ML & CSS methods often assume that social patterns can (only) be captured with very big data.
   - How do we reconcile ML's data-driven complexity with social science's focus on meaning and structure?
   - To what extent is this assumption compatible with theories that emphasize contingency, particularism, and context-dependence?

# Small-group guiding questions

2. **Data, complexity, and social theory**: ML & CSS methods often assume that social patterns can (only) be captured with very big data.
   - How do we reconcile ML's data-driven complexity with social science's focus on meaning and structure?
   - To what extent is this assumption compatible with theories that emphasize contingency, particularism, and context-dependence?

3. **Model transparency and interpretability**:
   - How do the 3 papers approaches attempt to make ML interpretable for social scientists, and where do they fall short?
   - Is interpretability necessary for social scientific usefulness? Or can "black box" models be acceptable if they yield insight or prediction?
   - How should we balance methodological sophistication with theory-driven inquiry when using ML for social science?

# Small-group guiding questions

4. **Causality and heterogeneity**: Brand et al. demonstrate that ML can be a powerful tool for discovering causal effects.
   - How do data-driven approaches to social scientific explanation–including heterogeneity (e.g., causal forests)–complement or challenge traditional approaches to causal inference in social science?

# Small-group guiding questions

4. **Causality and heterogeneity**: Brand et al. demonstrate that ML can be a powerful tool for discovering causal effects.
   + How do data-driven approaches to social scientific explanation–including heterogeneity (e.g., causal forests)–complement or challenge traditional approaches to causal inference in social science?

5. **Ethics, validity, and the social consequences of ML**:
   + What are the ethical stakes of prediction in socio-economic outcomes (e.g., child well-being, poverty, education)?
   + What are the risks of when using CNNs and transfer learning to analyze images of people or communities?

# Small-group guiding questions

6. **Synthesis**:
   - What do these papers collectively tell us about the strengths and weaknesses of machine learning in social science?
   - If you were to design a new project combining elements of the papers—large-scale prediction, CNNs for new forms of data, and heterogeneous treatment effects—what would the research question look like?

# Small-group guiding questions

6. **Synthesis**:
   - What do these papers collectively tell us about the strengths and weaknesses of machine learning in social science?
   - If you were to design a new project combining elements of the papers—large-scale prediction, CNNs for new forms of data, and heterogeneous treatment effects—what would the research question look like?

7. **Implications**
   - Salganik et al. claim that many life outcomes are only weakly predictables. Does this suggest fundamental limits to quantitative social science?
   - How do you think these methods can be applied to your own research interests?

# Mini-debate — Motion & format (15 min)

**Motion:** "Machine learning is best suited for prediction and measurement tasks, but not for causal explanation in social science."

# Mini-debate — Motion & format (15 min)

**Motion:** "Machine learning is best suited for prediction and measurement tasks, but not for causal explanation in social science."

Format:

- Form two teams (pro / con).
- Preparation: **5 minutes**
- Debate!

# References


- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., … & McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *PNAS*, 117(15), 8398–8403.
- Torres, M., & Cant, F. (2022). Learning to see: Convolutional neural networks for the analysis of social science data. *Political Analysis*, 30(1), 113–131.
- Brand, J. E., Xu, J., Koch, B., & Geraldo, P. (2021). Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology*, 51(2), 189–223.