Class Reading Assignment 7: Methodological Pitfalls
Course: CS 6474 / CS 4803 Social Computing

Grade: 4% of overall course grade (40 points total)

Due Date: the last class of instructional period on April 27 | *Earlier submissions encouraged* | *Earlier submissions will be graded sooner and feedback provided ~2-3 weeks from the time of submission*
What to hand in: Submit as a single PDF on Canvas

Formatting Guidelines:
• Length: approximately 3 pages single-spaced, 1-inch margins
• Font: at least 11pt, readable serif or sans-serif

Grading Emphasis:
- Accurate understanding of key concepts
- Use of specific examples from the readings (not vague summaries)
- Clear distinction between related methodological issues
- Depth of reasoning
- Effective integration across papers where required
- Critical interpretation of predictive metrics

Collaboration Policy:
This is an individual assignment. You may discuss high-level ideas with classmates, but all submitted work must be your own. You may not share written responses.

AI Use Policy:
You may use AI-based tools only for proofreading or improving clarity. You **may not** use AI tools to generate ideas, arguments, or structure. Responses should reflect your own reasoning and engagement with the readings and lectures.

This assignment builds directly on Week 13's lectures and discussions on "Methodological Pitfalls (I and II)". Papers included:

- The parable of Google Flu: traps in big data analysis [pdf]
- Exploring Limits to Prediction in Complex Social Systems [pdf]
- Private traits and attributes are predictable from digital records of human behavior [link]

Question 1: Methodological Pitfalls in Big Data and Prediction

(a) (7 pts) The Google Flu Trends (GFT) paper introduces the concept of "big data hubris." Explain this concept and describe one specific methodological issue in GFT's modeling approach that illustrates it (e.g., how search terms were selected or how overfitting occurred).

(b) (7 pts) The GFT paper argues that its predictions suffered from both measurement issues and data-generating process changes. Identify one example of each from the paper and briefly explain how each contributed to prediction error. Your answer should clearly distinguish between (i) issues with what is being measured and (ii) instability in how the data are produced over time.

(c) (6 pts) Drawing specifically on both the GFT paper (Lazer et al.) and the Martin et al. paper, explain why simply increasing the volume of data (e.g., more search logs or more social media features) may not substantially improve predictive performance. Your answer should reference one limitation from each paper.

Question 2: Interpreting Predictive Power and Its Risks

(a) (8 pts) The Kosinski et al. paper reports high accuracy in predicting sensitive personal attributes from Facebook Likes. Describe one methodological concern and one ethical concern raised by this approach. Your answer should refer to specific aspects of the data, modeling approach, or evaluation (e.g., use of self-reported labels, dimensionality reduction, or prediction of latent traits).

(b) (6 pts) In the Martin et al. paper, the authors show that even their best-performing models explain less than half of the variance in cascade size. Explain two distinct methodological reasons discussed in the paper for this limitation (e.g., feature limitations vs. intrinsic randomness, or the "luck versus skill" exercise we went through during lecture on April 6).

(c) (6 pts) Across the Kosinski et al. and Martin et al. papers, explain why high predictive accuracy (e.g., AUC or R^2) can be misleading when interpreting model performance. Provide one example from each paper to support your answer.