# Summary of *Machine Learning as A Tool for Hypothesis Generation*

*2024.05.28*

*Yiming LI*

1. **What are the research questions?**
   - How can machine learning algorithms be used to generate hypotheses that are both novel and testable?
2. **Why are the research questions interesting?**
   - Science exhibits a curious asymmetry.
   - Idea generation is an empirical activity that happens off stage.
   - How can this process be formalized?
     - Machine learning can identify patterns unnoticed by humans.
     - Data on human behavior is proliferating and becoming machine-readable.
     - Machine learning models often act as "black boxes" in terms of their predictors. But science is to achieve generalization, which needs interpretability.
3. **What is the paper's contribution?**
   - Contribute to literature on Hypothesis Generation:
     - Recent: Generates hypotheses based on existing theories or economic intuition.
     - Extend: Proposes a systematic procedure for generating hypotheses using ML.
   - Contributes to literature on Machine Learning in Economic Research:
     - Recent: Focus on new measures, new models, and causal inference tools.
     - Extend: Applies data-driven machine learning algorithms to a novel field.
4. **What hypotheses are tested in the paper?**
   - H1: ML algorithms can be used to generate hypotheses that are both novel and testable.

   **a) Do these hypotheses follow from and answer the research questions?**

   - Yes, simple and clear, counterfactual synthetic images.

   **b) Do these hypotheses follow from theory? Explain logic of the hypotheses?**

   - Yes, this paper's logic is straight forward.
   - Human judgments often contain a significant amount of "noise," while intuition can be associated with "overfitting."
   - Machine learning produces predictions for new (out-of-sample) data but remains a "black box" in its workings.
   - Algorithms uncover new signals and then rely on humans to label those discoveries.
5. **Sample: comment on the appropriateness of the sample selection procedures.**
   - Only use data from Mecklenburg county might exist geographical bias.
6. **Comment on the appropriateness of variable definition and measurement.**
   - Too subjective to choose psychological traits like trustworthiness assessed by HITs.
7. **Comment on the appropriateness of the regress/predict model specification.**
   - High quantity of images for human selection lead to attention problems.

8. **What difficulties arise in drawing inferences from the empirical work?**
    - The accuracy of conclusions relies on the caliber and inclusiveness of the data employed.
9. **Describe at least one publishable and feasible extension of this research.**
    - Explore how machine learning algorithms trained on various types of data (e.g., financial transactions, social media interactions, consumer behavior) could uncover hidden patterns influencing decisions made by professionals in fields such as finance, healthcare, or marketing.