BUILDING BETTER THEORIES THROUGH PREDICTION AND EXPLANATION

Chiara Binelli

Department of Political and Social Sciences
University of Bologna & CeRSP

SICSS-Lake Como

2 July 2025

Josh Peterson
Boston University

Stanley Huang
Boston University

Bartu Tamer
University of Bologna

Building Better Theories
Through Prediction and Explanation

SICSS-Lake Como

2 July 2025

OUTLINE

1. To **Explain** and to **Predict**.
2. ML for Theory Development.
3. Controlled Environments Versus Observational Data.
4. **Prediction** in the Research Design.
5. Examples and Applications.

TO **EXPLAIN**

- The core goal of Social Science is **to explain**. Model's specification:

$$y = f(X) + \varepsilon$$

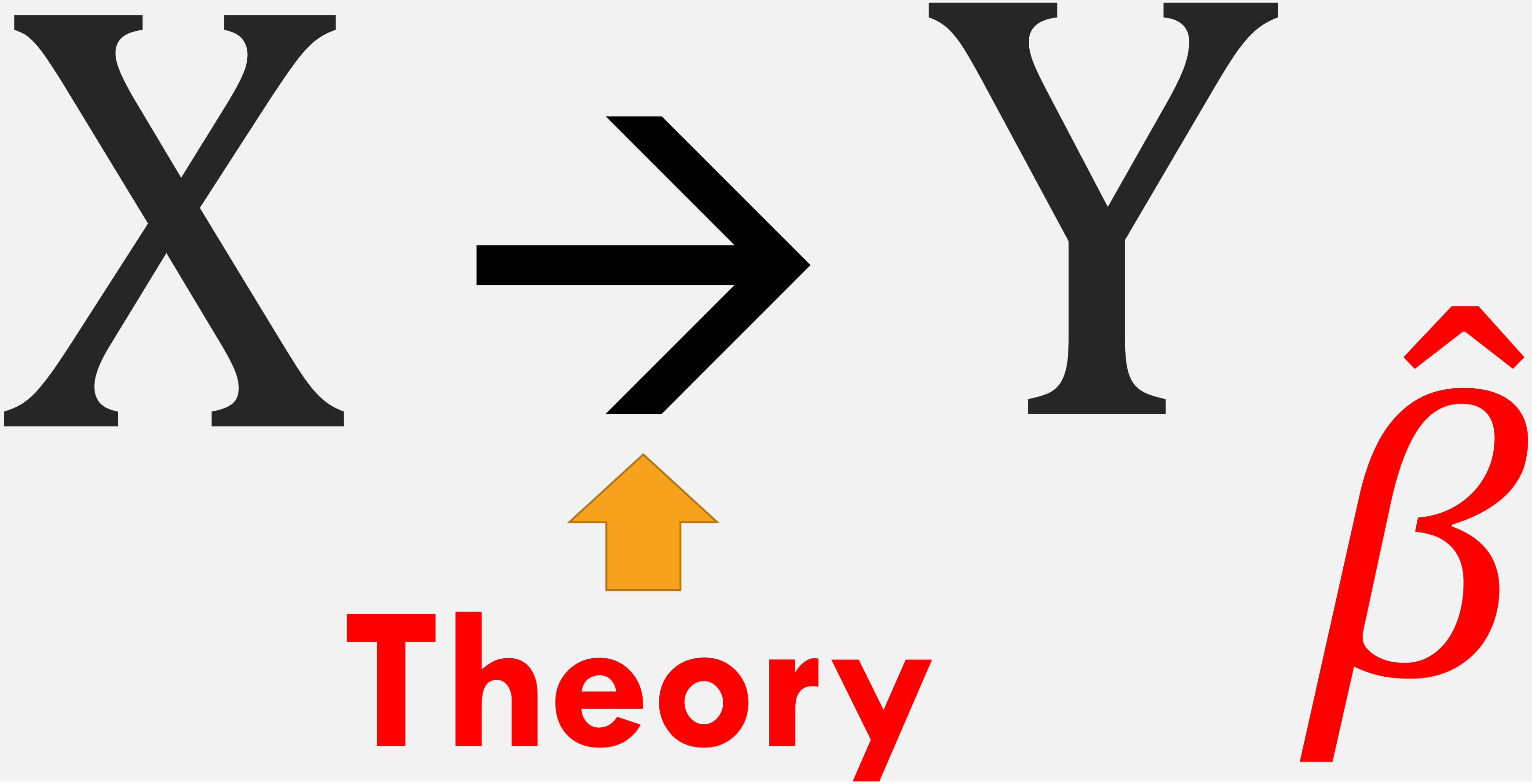
- Assumptions on:

1. Which *f*() (function or program mapping) turns inputs into outcome.
2. Which X (inputs or factors) and how affect *y* (outcome or observed behaviour).

➢ (1) and (2) affect the assumptions on the distribution of the error term (ε) and on which factors are assumed to remain unobserved either because they are unmeasurable and/or very poorly measured and/or too difficult to conceptualize.

➢ **Goal: estimate the marginal change $\left(\frac{\partial y}{\partial x}\right)$ in *y* for one unit change in X.**

THE CENTRAL ROLE OF THEORY



$X \rightarrow Y$

Theory

$\hat{\beta}$

CAUSALITY AND PREDICTION

$$y = f(X) + \varepsilon$$

$\left(\frac{\partial y}{\partial x}\right)$ is the **causal impact** of X on *y* when it measures the difference between the value of *y* given the actual change in X and the counterfactual value of *y* if X had not changed.

- If a model can provide an explanation of how X 'causes' *y*, it should also predict how a change in the causal determinant of *y* translates into a future change in *y*.
  - A causal explanation is necessarily a prediction since any phenomenon that can be explained must – by the definitional implication of 'to cause [something to happen]' - be sufficiently predictable (Buchholz and Grote 2023). **Causality implies prediction**.
  - **By knowing why *y* (explanation), we will know where *y* will be (prediction).**

BIAS-VARIANCE TRADE-OFF

- While theoretically causality implies prediction, empirically the two concepts clash.

- **Bias-variance trade-off**: estimating the **unbiased impact of X on *y*** conflicts with the model's ability to **predict *y*** (in the future or out of sample).

  - If a fitted model does not change its shape significantly in response to a new observation, it will not predict well. **High interpretability-low prediction accuracy trade off**.

    - **Linear Regression**: only two parameters to fit (intercept and slope). Good for causality, bad for prediction.

    - **Neural Network**: there exists at least one network capable of approximating any function. Good for prediction, bad for causality.

CAUSALITY AND PREDICTION

Measure the unbiased and causal impact of X on y $\left(\frac{\partial y}{\partial x}\right)$

High interpretability and descriptive

EX: Develop a theory of the multiple factors/mechanisms that could explain why Kenya experiences high poverty and build explicit models

Learn a function that accurately predicts an outcome in the future or from unseen observations

Less interpretable but more predictive

EX: Build predictive algorithms to forecasted poverty rates in Kenya 5 years from now

HOW TO **EXPLAIN**?

- **Causality implies prediction** *but* empirically **to explain** clashes with **to predict**.

$$y = f(X) + \varepsilon$$

Assumptions on:

1. Which *f*() (function or program mapping) turns inputs into outcome.
2. Which X (inputs or factors) and how affect *y* (outcome or observed behaviour).

- The bias-variance trade-off is **not** an obstacle to using ML for theory development.
- **Goal**: ML to guide on ass. 1 and 2 to develop a model that can better explain **and** predict.

USING ML FOR THEORY DEVELOPMENT

- In Psychology, ML has been used extensively for theory building. Two main techniques:
  - **Scientific Regret Minimization (SRM) (Agrawal, Peterson, Griffiths 2020 *PNAS*)**: ML to identify the "explainable variance", that is the variance that is signal and not noise, and can thus be used as a benchmark to improve a theory-based model of behaviour.
    - **ML to help with Ass. 2: to identify the X inputs and their interactions**.
  - **Differentiable theories (Peterson *et al.* 2021 *Science*)**: ML to compare the predictive performance of different theories of decision-making.
    - **ML to help with Ass. 1: to identify the *f*() mapping that turns the X factors into y**.

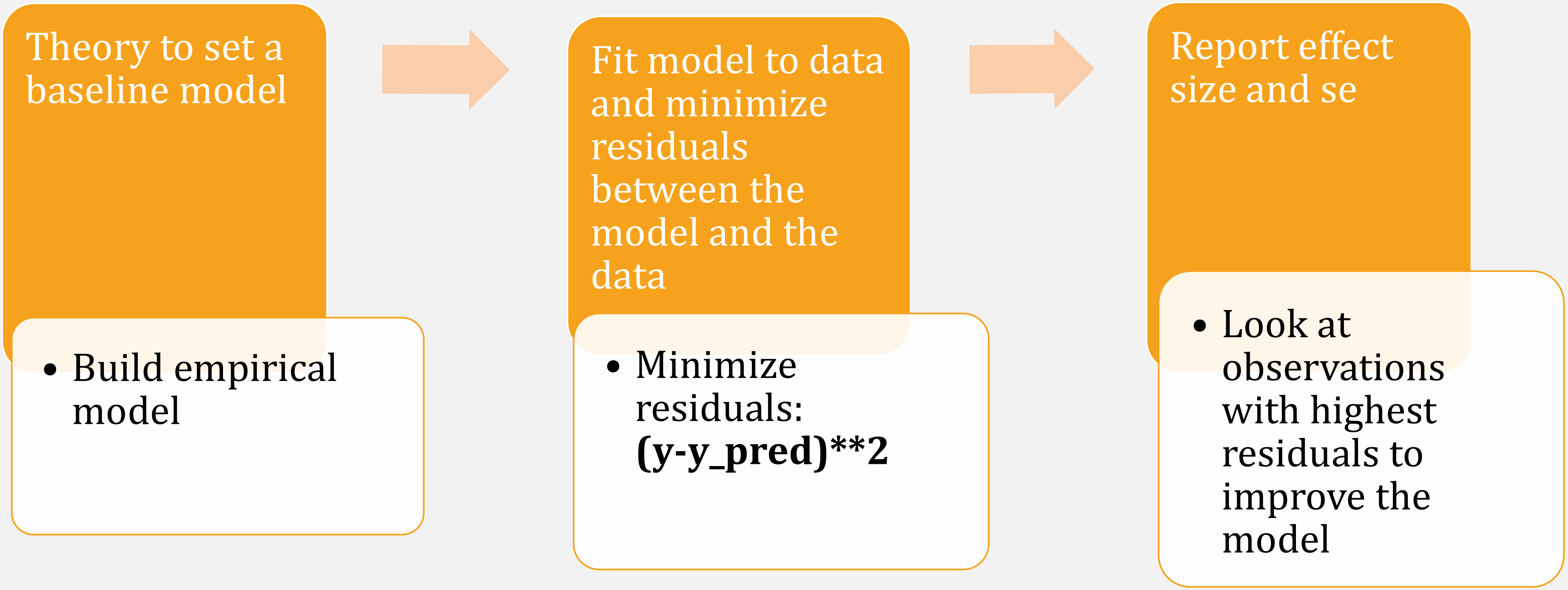
SCIENTIFIC REGRET MINIMIZATION (AGRAWAL, PETERSON, GRIFFITHS 2020 *PNAS*)

- **SRM**: rather than minimizing the expected squared residuals between the model and the data, minimize the expected squared residuals between the model and the prediction from a data-driven ML algorithm such as a NN.
  - Benchmark given by a theoretically unconstrained ML model to identify the explainable variance, that is what is predictable. Kuperwajs, Schütt and Ma (2023) use the same technique to study sequential decision making.

- The prediction error of a ML algorithm decreases when the size of the data increases.

- Interpretable model that jointly maximizes explanatory and predictive power.

TRADITIONAL APPROACH IN SOCIAL SCIENCE



Theory to set a baseline model

- Build empirical model

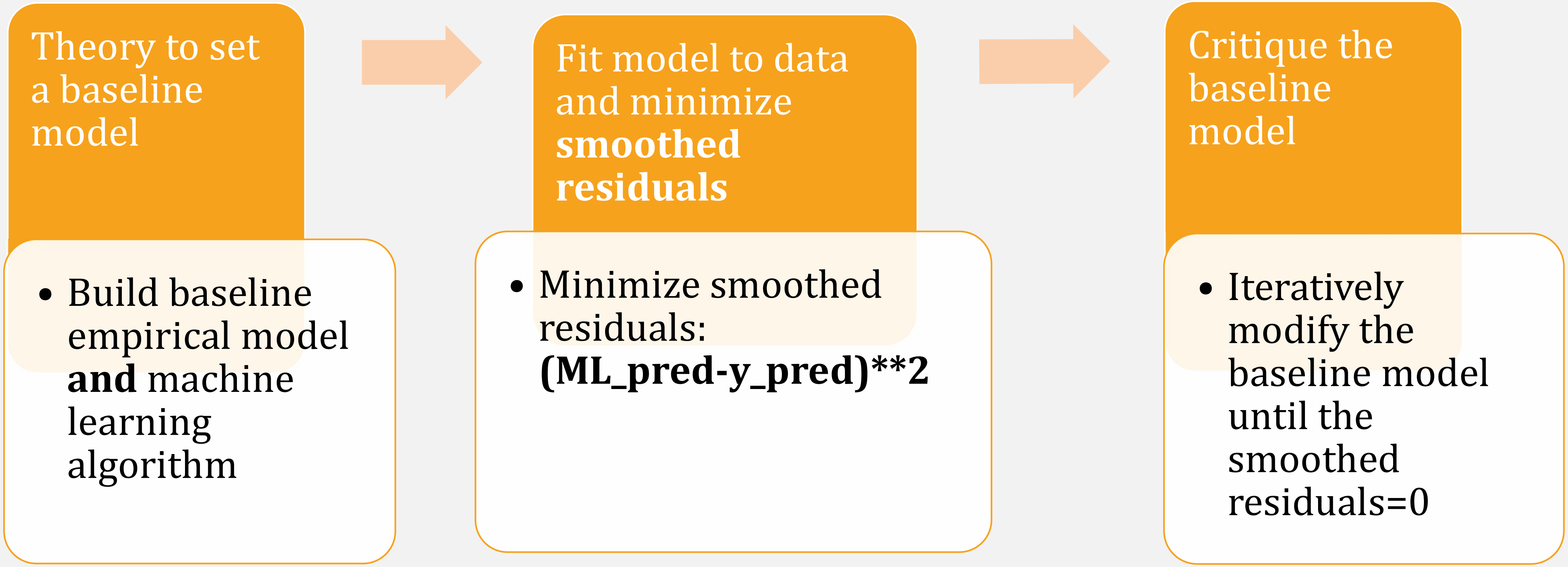
Fit model to data and minimize residuals between the model and the data

- Minimize residuals: **(y-y_pred)**2**

Report effect size and se

- Look at observations with highest residuals to improve the model

SRM



Theory to set a baseline model

- Build baseline empirical model **and** machine learning algorithm

Fit model to data and minimize **smoothed residuals**

- Minimize smoothed residuals: **(ML_pred-y_pred)**2**

Critique the baseline model

- Iteratively modify the baseline model until the smoothed residuals=0

WEAKNESSES OF SRM

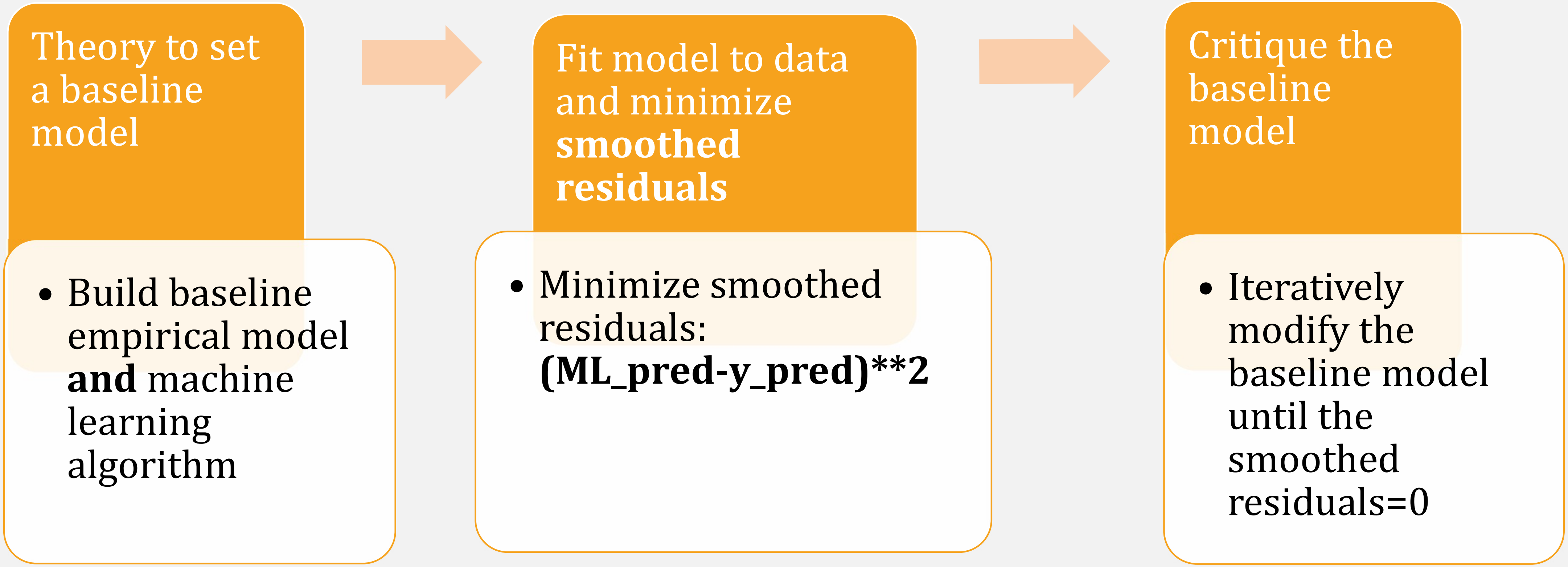
- Internal validity:
  - SRM is used in a lab setting under specific sampling assumptions (random sample and random variation of preferences, etc.): biased sampling and limited meaningful variation among subjects.
    - Online collection tools to collect big datasets improve on results obtained using lab data.
- External validity:
  - The findings can apply only to the sample used.
    - The moral machine data has lots of information on countries and demographics, and using this information makes any difference in predictive performance (Agrawal, Peterson and Griffiths 2020).
- **More fundamental question: is ML useful for theory development when we are not in a controlled environment?**

CONTROLLED ENVIRONMENT VERSUS OBSERVATIONAL DATA

- In Psychology, research is mainly conducted in controlled environments with a design that includes an experimental setup where a stimulus affects behaviour:
  - ➢Agrawal, Peterson and Griffiths (2020) use data from a large-scale experiment.
  - ➢Peterson et *al*. (2021) (as well as Zhu et *al*. 2025 and many others) use lab data.

- Often, research is conducted using ***observational data***, i.e. data collected ***not*** with an experiment (e.g. surveys, online sources such as web scraping and google searches). Dataset where for each unit of analysis *i* there is information on several variables/features (NxK dataset, where N>K or reverse N<K).
  - ➢**How to use ML to develop theories and generate novel explanations of human decision making when observational data are available?**

HOW TO OPERATIONALIZE SRM WITH OBSERVATIONAL DATA?



Theory to set a baseline model

- Build baseline empirical model **and** machine learning algorithm

Fit model to data and minimize **smoothed residuals**

- Minimize smoothed residuals: **(ML_pred-y_pred)**2**

Critique the baseline model

- Iteratively modify the baseline model until the smoothed residuals=0

PREDICTION IN THE RESEARCH DESIGN

1. Theoretical model that identifies the factors that determine *y*: baseline model.

2. ML algorithm to identify the explainable variance (e.g. NN) or ensemble method like a Super Leaner (Verhagen 2024) and model's fit.

➢The fit of the baseline model and the ML algorithm is evaluated out-of-sample, through cross-validation or with a validation set (Verhagen 2024).

➢Possible to include an exploration of the function mapping (Peterson *et al.* 2021).

3. Critique the baseline model in step 1 using the ML algorithm in step 2.

➢The ML algorithm is the benchmark. Penalize only for mis-predicting what is predictable.

4. Based on the results in step 3, update and re-estimate the baseline model in step 1.

PREDICTION IN THE RESEARCH DESIGN: STEP 3

Step 3: critique the baseline model in step 1 using the ML algorithm in step 2.

- Look at the cases where **(ML_pred - baseline_model_pred)**2** are the largest. The larger the difference, the larger the explainable variance that is missed by the baseline model because of limited model's complexity.

PREDICTION IN THE RESEARCH DESIGN: STEP 4

Step 4: based on the results in step 3, update and re-estimate the model in step 1.

- If the comparison in step 3 identifies new factors, interactions and/or new functional forms (Agrawal, Peterson and Griffiths 2020 and Verhagen 2024), re-estimate the model accordingly.
  - ➢Use RCT to assess the relevance of newly identified factors (Agrawal, Peterson and Griffiths 2020).

EXAMPLE 1:
CLIMATE CHANGE PERPCEPTIONS

- Binelli, Loveless and Schaffner (2023) "Explaining Perceptions of Climate Change in the US".
  - Theory: model of climate change perceptions that includes all exogenous determinants.
  - Data: 2014 wave of the CCES survey merged with county-level data on the climate indexes constructed by Kaufmann *et al.* (2017) using daily high and low temperatures for 18,713 weather stations located in the US. **N=9,407**
  - Main results: local climate change significantly affects perceptions and in the expected direction, but partisanship and political ideology maintain the strongest effect.
- We can build a similar dataset using the 2012 wave of the CCES. **N=52,613**

EXAMPLE 1: STEP 1

STEP 1: Theory-driven baseline model that identifies the factors that determine *y* (Binelli, Loveless and Schaffner 2023):

$$y_{ic} = \beta_0 + \beta_1 TMax_c + \beta_2 RW1_c + \beta_3 RW2_c + \beta_4 RW3_c + \beta_5 RW4_c + \beta_6 PID_{ic} + \beta_7 PolId_{ic} + \gamma \mathbf{X}_{ic} + \alpha_c + \varepsilon_{ic}$$

y_{ic} is the perception of climate change of individual *i* in US county *c* (7 response items)

$TMax_c$ and RWj_c, j=1,2,3,4 are the climate indexes measuring local change in the climate in county *c*

PID_{ic} is party identification

$PolId_{ic}$ is political ideology

$\mathbf{X}_{ic}$ is a vector of characteristics (gender, race, age, education, religiosity) of individual *i* in county *c*

α_c is the county fixed effect

ε_{ic} is a normally distributed error term

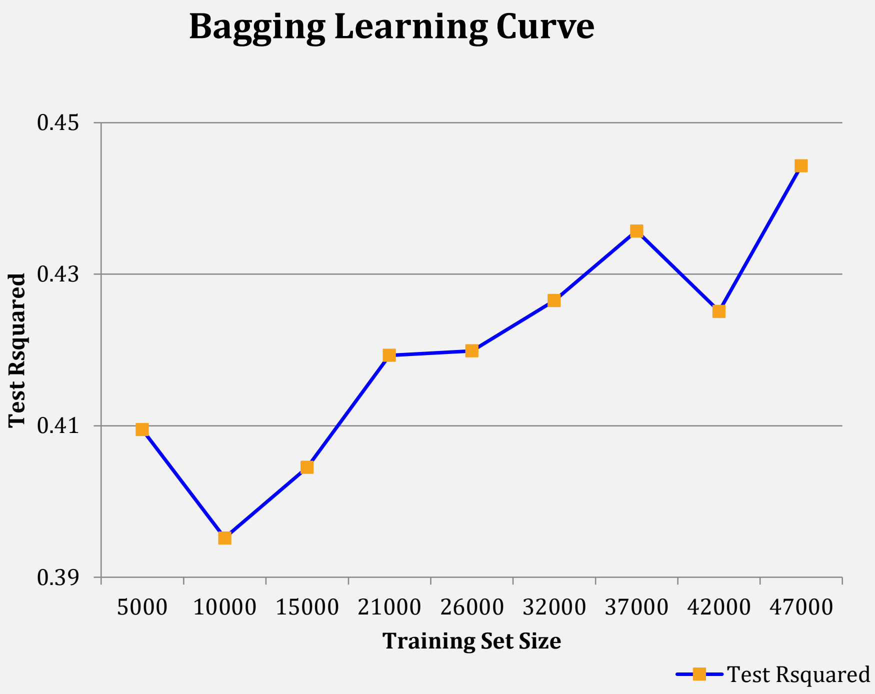
EXAMPLE 1: STEP 2

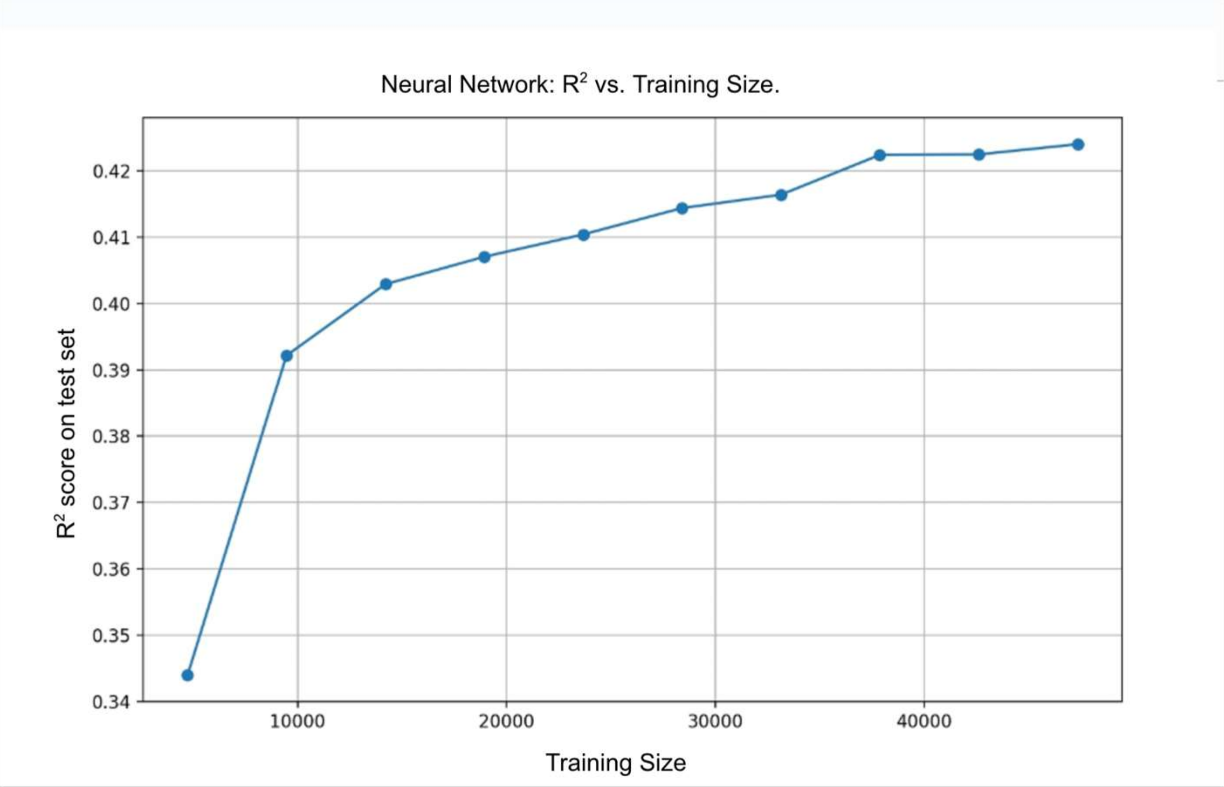
BASELINE MODEL: Rsquared=0.42.

- **Is 42% Rsquared good or bad? Very bad for some problems (explaining global temperature), very good for other problems (explaining the root factors of altruism).**

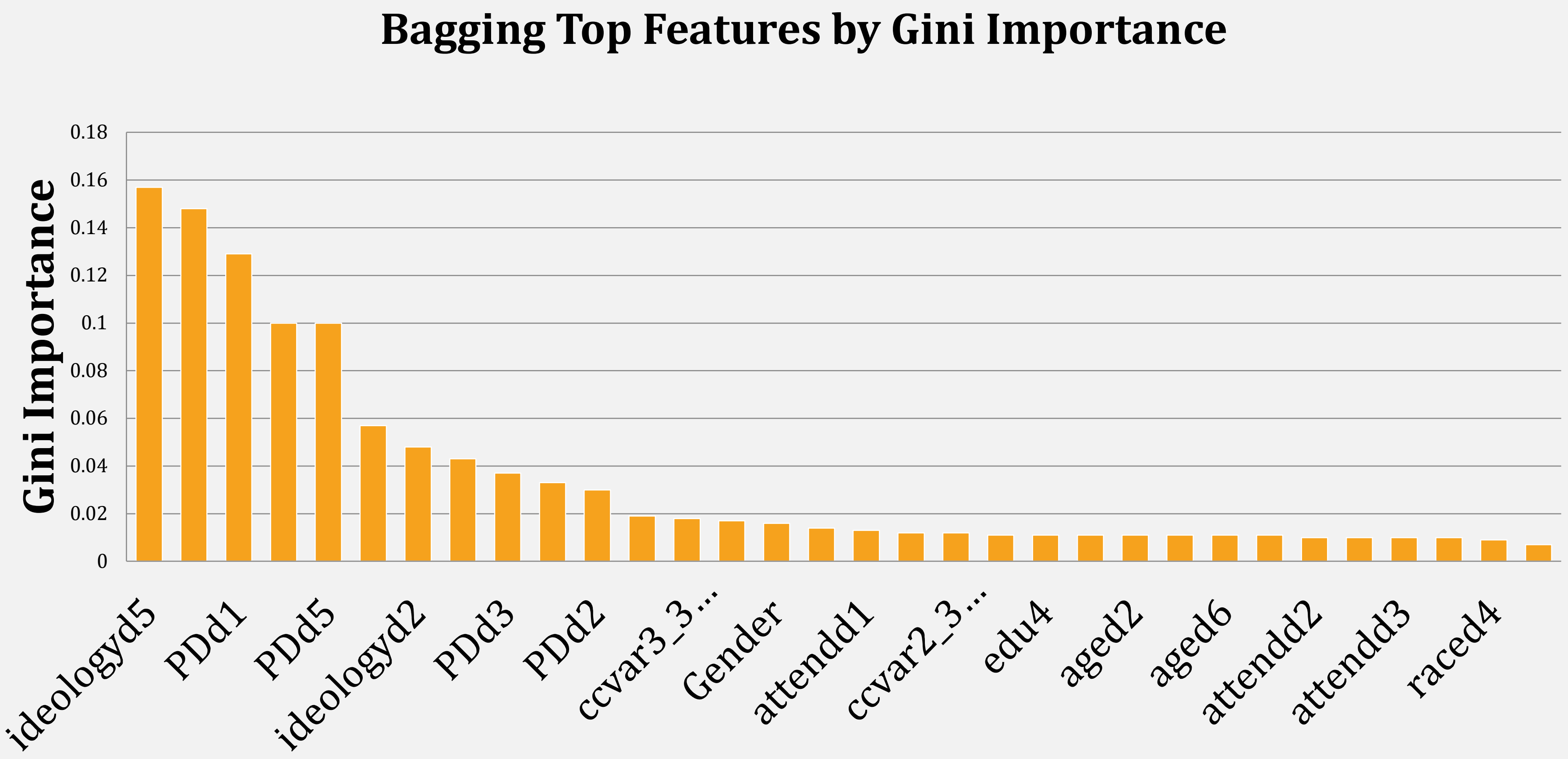
ML Algorithm	Train MSE	Test MSE	Train Rsquared	Test Rsquared
XgBoosting	0.703	0.919	0.51	0.364
Random Forest	0.413	0.8	0.72	0.43
Bagging	**0.431**	**0.822**	**0.703**	**0.44**
Neural Network	**0.79**	**0.821**	**-**	**0.439**
XgBoosting with county dummies	0.716	0.932	0.504	0.368
Random Forest with county dummies	0.466	0.834	0.677	0.435
Bagging with county dummies	0.54	0.83	0.621	0.438
Neural Network with county dummies	0.78	0.82	-	0.433

EXAMPLE 1: STEP 2





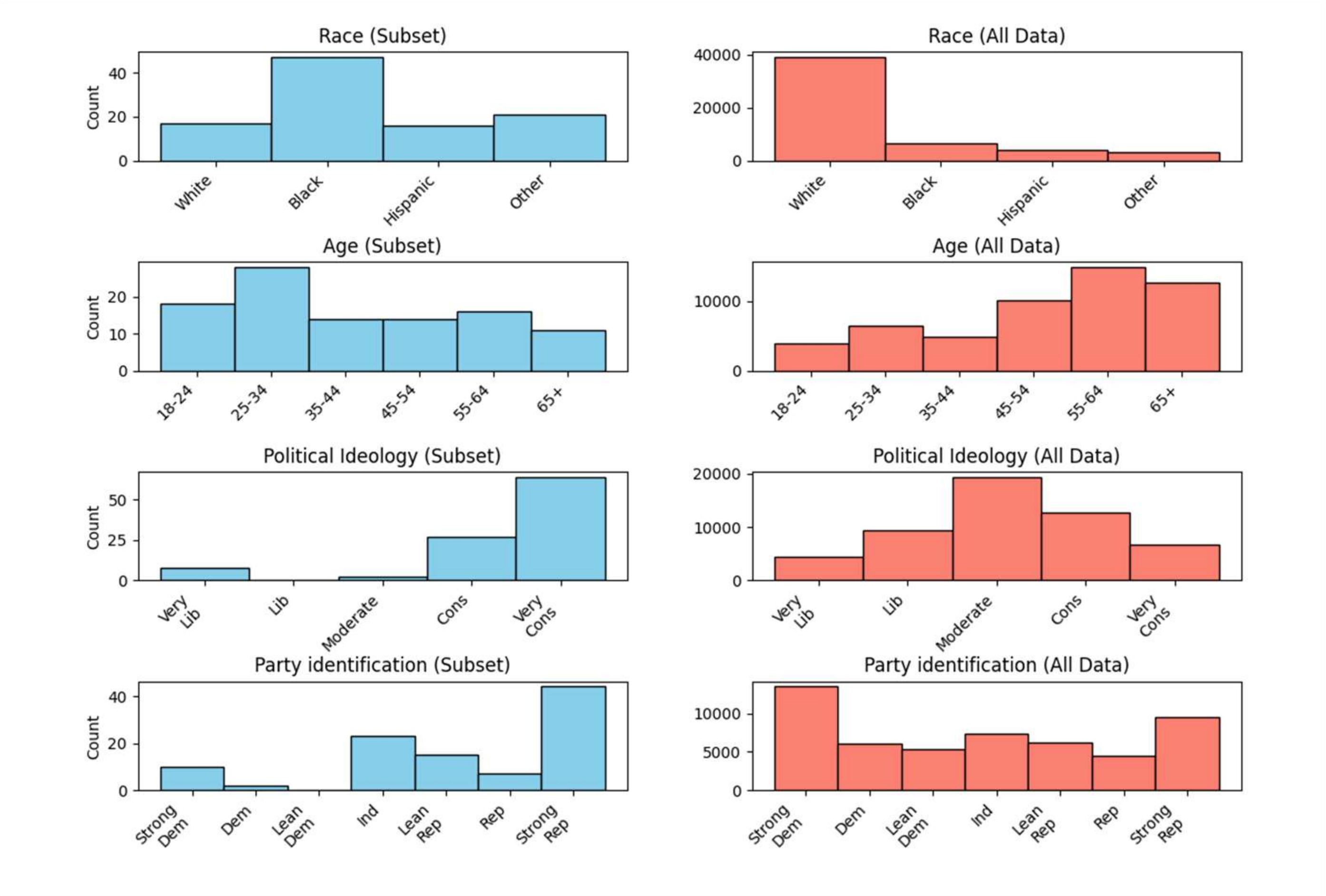
EXAMPLE 1: STEP 2



EXAMPLE 1: STEP 3

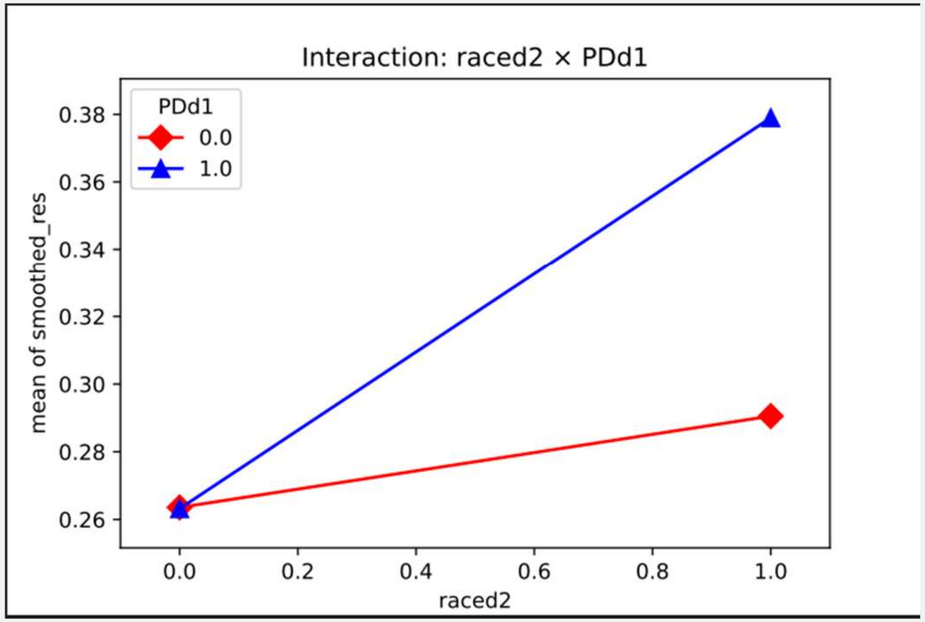
- STEP 3: Critique the baseline model using ML.
  - Identify the top 100 observations for which **(ML_pred-baseline_model_pred)**2** are the largest.

EXAMPLE 1: STEP 3



EXAMPLE 1: STEP 3

- **Bagging:** the analysis by features' importance in the baseline model shows that:
  - ➢the baseline model mis-predicts for Very Conservative, Strong Democrats and Black.
  - ➢Interaction between Strong Democrats and Black, and, reversely, between Strong Republicans and White.



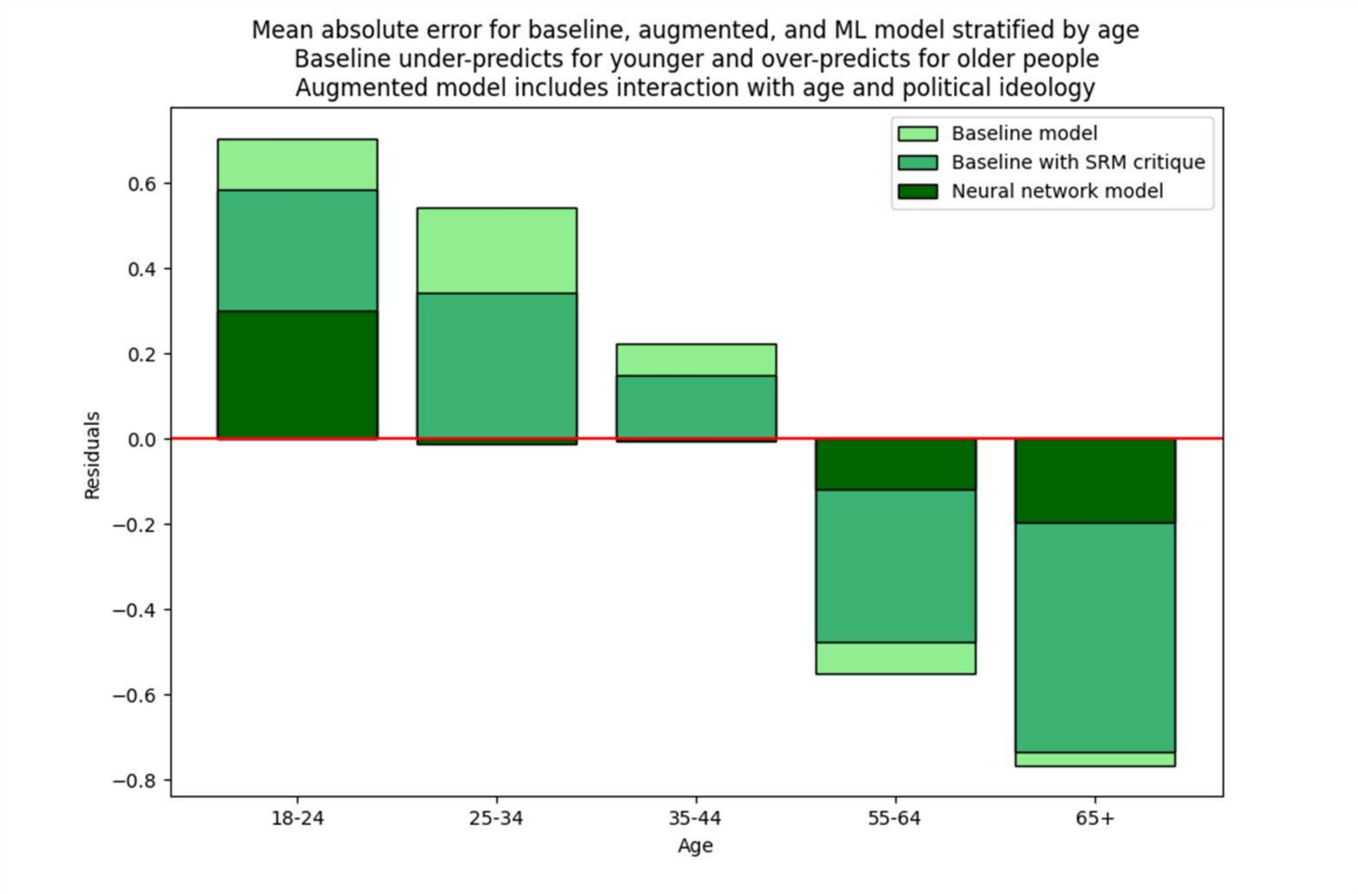
EXAMPLE 1: STEP 4

- STEP 4: Based on the results in step 3, update and re-estimate the baseline model.

- Race, Age, Political Ideology and Political Identification.
  - Augment the baseline model with respect to the interactions between these features.
  - Assess improvement in model performance and changes in estimated coefficients.

EXAMPLE 1: STEP 4

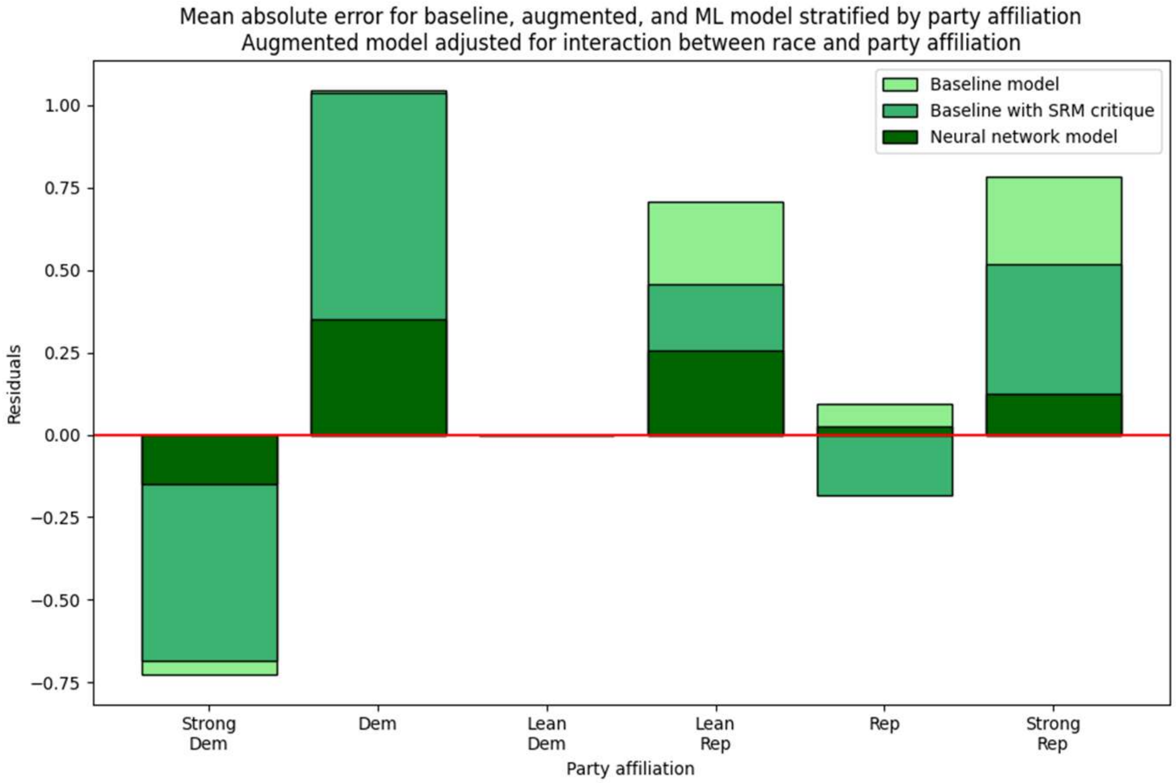
Improvement in model performance when adding interactions between Age and Political Ideology



- ➢ **Controlling for the interaction between age and political ideology improves the baseline model more for the young than for the old.**

EXAMPLE 1: STEP 4

Improvement in model performance when adding interactions between Race and Political Identification



- **Controlling for the interaction between race and party affiliation improves the baseline model for Republicans.**

EXAMPLE 1: STEP 4

- Assess changes in estimated coefficients.
  - ➢Augment the model by adding interaction terms between Black and Strong Democrat and White and Strong Republican and between Black and Very Liberal and White and Very Conservative, for a total of 4 new interactions terms.
  - ➢In the baseline model the negative impact of Strong Rep and Very Conservative are overestimated by 3% points each, and the positive impact of Strong Dem and Very Liberal are underestimated by, respectively, 1 and 3 % points.

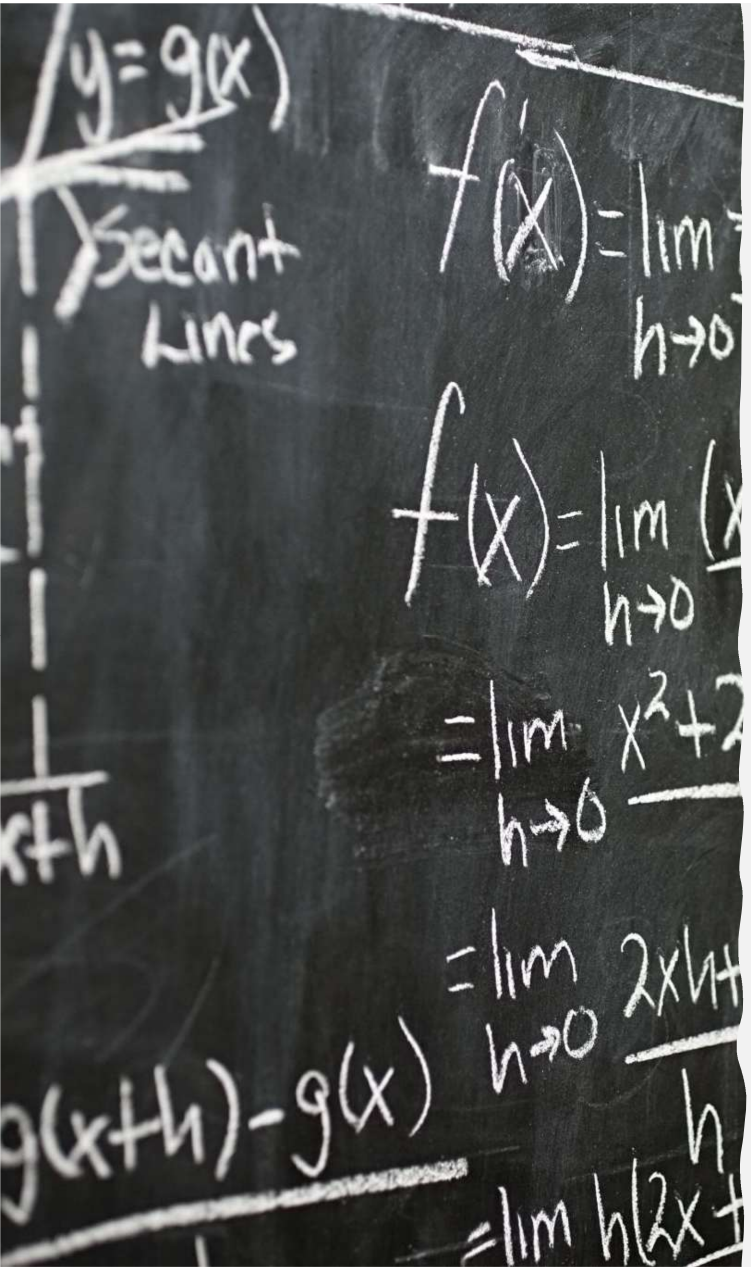
EXAMPLE 2:
INDIVIDUALS' PRO-ENVIRONMENTAL ACTIONS

- Binelli, Loveless and Schaffner (2023) and Binelli and Loveless (2025): information on local future climate change positively affects individuals' pro-environmental actions and support for green policies.
- Y: pro-environmental actions and support for green policies.
  - Experiment: provide information on future climate change (how climate change will affect changes in temperature in 30 years in the city of residence). to assess how it affects Y compared to a control group that did not receive the information.
  - Finding: holding an individualistic view of climate change is key for the treatment to affect pro-environmental actions.
- ML to predict Y: holding an individualistic view of climate change is key to predict. New experiment to affect individualist view of climate change.
- **SRM to improve model building <u>and</u> guide in choosing which experiments to run.**

A NEW RESEARCH DESIGN (I)

The Research Design outlines the path of scientific research.
Prediction becomes an integral part of the Research Design.

- *Step One*: Research Question
  - RQ: **To Predict**: we must test **how the model predicts *y***
  - RQ: **To Explain**: *implies prediction*, thus we must test **both the impact of X on *y* and how the model predicts *y***.
- *Step Two:* Validating Research Plausibility and Theoretical Model
  - Theory and Literature Review
  - Testable hypothesis (impact of X on *y* **and** predicted *y*)
    - Any question of "**to explain**" implies a prediction, thus we must test **both the impact of X on *y* and how the model predicts *y***.



A NEW RESEARCH DESIGN (II)

- *Step Three*: Methods of Discovery: **Estimation and empirical tests/Theory to the data**
  - Regressions *and* ML
  - SRM to compute the explainable variance and improve the theoretical model.
  - We can use an interpretable ML to improve the transparency of the ML benchmark.
- *Step Four*: Model Fit/Robustness
- *Step Five:* Interpretation

RELATED APPROACHES AND APPLICATIONS

1. ML for theory development
   - Ludwig and Mullainathan (2024, QJE): use ML to generate novel hypotheses from high-dimensional data sets. Concrete application of judge decisions about whom to jail.
   - Mullainathan and Rambachan (2024): ML prediction algorithm to generate theory's anomalies.
   - Fudenberg, Liang and coauthors: ML to measure model's completeness as the achievable level of prediction (2022, JPE) and model's restrictiveness (2023, RES).
   - Verhagen (2024) "Incorporating Machine Learning into Sociological Model-Building", Soc. Meth.
2. Predicting play in games/choice theory and general methodology
   - Hartform, Leyton-Brown and Wright (2017): deep learning to predict play in games.
   - Zhao, Ke, Wang, and Hsieh (2020): "behavioral neural network" to predict choices over lotteries (choices with uncertain outcomes) by mimic the properties of existing economic theories.
   - Fessler and Kasy (2018): "regularizing towards the theory". ML as an unrestricted search in the domain and the theory as a tool to impose structure on the functions in the domain.

DISCUSSION AND WAYS FORWARD (I)

- The complexity of human behaviour is greater than we can even imagine. With more data, we have more variance to explain, and more complex models can get us closer to an unbiased estimate of the underlying DGP.
    - ➢ As we move to larger and larger data sets, we are going to need to become more comfortable with complexity (Griffiths 2015).
    - ➢ To predict and to explain have to be combined (Shmueli 2010; **Hofman et *al.* 2021 *Nature* "Integrating explanation and prediction in computational social science"**).

- Assessing how much explanatory models can predict is necessary to show how much variance a given explanation is actually explaining.
    - ➢ Salganik *et al.* (2020): unpredictability may provide insights on how to construct better theories borrowing from fields where unpredictability is the object of study such as civil conflicts (Cederman and Weidmann 2017) or weather forecasts (Alley, Emanuel, and Zhang 2019).

DISCUSSION AND WAYS FORWARD (II)

- Embedding prediction in the research design and using it for theory development advances the ability to explain.
  - A merge between predicting and explaining changes the way we do science - from our research designs to our empirical analyses, and allows to better understand behaviour.

- Standard ML methods can improve theory building and explain behaviour.
  - To build better models and reduce the possibility of P-hacking.
  - To identify new phenomena that we can empirically validate and replicate.

- Frontier approach to develop theory-based ML:
  - Bourgin et al. (2019): ML algorithms with appropriate inductive biases for capturing human behaviour (HB) by pretraining neural networks with synthetic data generated by cognitive models.
  - Jian-Qiao et al. (2025, Nature HB): deep NN predict choices better than theories of strategic behaviour, and modified network to produce a new and interpretable model that help generate novel explanations of HB.

CHALLENGES AND OPEN QUESTIONS

- When is the data "big enough"? Which level of complexity of theory and data (number of features, K vs N) require ML? Which amount of improvement justifies re-estimating the model? The answer depends on the true DGP that remains unobserved.

- Peterson *et al.* (2021 *Science*) use ML to test which theory best explains risky choices.
  - ➢Gambles characterized by payoffs and probability of realizations; the value of a gamble can be modelled using different theories (expected UT, prospect theory). Utility function modelled as a NN, and comparison of the predictive performance of the learned utility function with utility functions chosen by researchers.
    - The learned models outperform the base models when enough data are available.
    - The best performed model is an unconstrained NN that maps the information about the gamble to a decision probability.

➢Can we operationalize and use it to guide ass. 1 on which f() turns inputs into outcome?

NEXT STEPS

- Use ML for theory building to study core social science questions when very rich data are available.
  - ➢Leverage on the power of ML to capture complexity.

- The Millenium Cohort Study: https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/
  - ➢UK longitudinal birth study that started in 2001 and has followed the lives of around 19,000 people born across England, Scotland, Wales and Northern Ireland from birth to 2023. Info on physical, socio-emotional, cognitive and behavioural development, economic circumstances, parenting, relationships and family life across the life course.
  - ➢Example: study educational choices, starting from the simple college/no college choice for which we can use ML to guide on both ass. 1 (which function mapping turns X into y) and on ass. 2 (which X and how affect y).

BUILDING BETTER THEORIES THROUGH PREDICTION AND EXPLANATION

Thank you and we welcome your input!

EMAIL: chiara.binelli@gmail.com

SICSS-Lake Como

2 July 2025