Machine Learning Workshop

Iasmin Goes & Daniel Weitzel

Department of Political Science

Colorado State University

22-23 October 2024

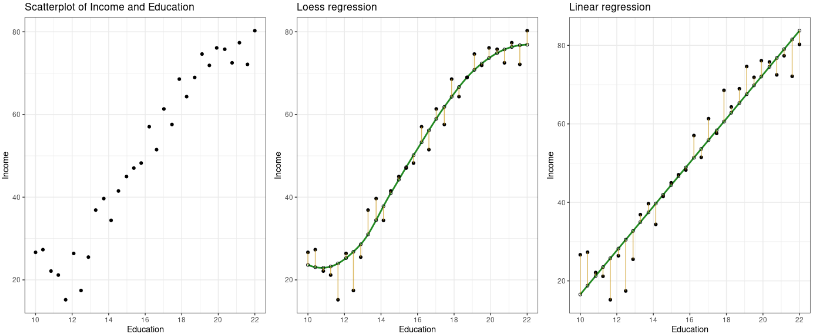
Slides

https://t.ly/x9-el

Why we analyze data

- **Goal:** understand the *systematic* relationship between an outcome (dependent variable) and predictors (independent variables)
- We want to know the function f in $Y = f(X) + \epsilon$
  - f is the unknown form with which X provides systematic information about Y

Why we analyze data



Difference between “traditional” statistics and ML

- Goal
  - **TS:** test hypotheses, make inferences about population parameters
  - **ML:** build predictive models for new data
- Approach
  - **TS:** pre-defined assumptions about the population distribution; these assumptions, in turn, inform model selection
  - **ML:** no assumptions about distribution or functional form; models learn patterns and relationships from the data
- Data
  - **TS:** good for smaller datasets that are a representative sample of a larger population
  - **ML:** good for larger and more complex data sets, also unstructured data (images, text, audio)

Difference between "traditional" statistics and ML

- Interpretability
  - **TS:** a set of pre-defined assumptions allows for higher interpretability; results are easier to understand
  - **ML:** can be "black boxes", models can be difficult to interpret
- Training
  - **TS:** usually no split between training and test data: the analysis is done on one full dataset
  - **ML:** the dataset is split into training, (cross-)validation, and test set to prevent overfitting of models to idiosyncratic features of the data

When ML makes sense

- Large data sets (number of observations)
- Large number of predictors and/or no theory about f
- Accurate predictions are more valuable than causal inference
- Complex non-linearity in the data
- Unstructured data
- Goal is feature generation (making new variables)

Types of ML

- **Supervised learning:** the computer is trained on a labeled dataset and learns to make predictions or classifications based on new data
- **Unsupervised learning:** the computer is given an unlabeled dataset and is tasked with finding patterns or relationships within the data. The model might also generate its own labels (self-supervised learning)
- **Semi-supervised learning:** the computer is given a dataset with labeled *and* unlabeled examples and uses the unlabeled data to improve its performance
- **Reinforcement learning:** the computer learns to make decisions based on a reward signal, which is provided when it performs an action that leads to a positive outcome

Example 1: supervised learning

This paper uses random forests to predict the "critical mass" of women legislators needed to affect health, education, and defense spending

Political Science Research and Methods (2021), page 1 of 19
doi:10.1017/psrm.2021.51

ORIGINAL ARTICLE

Point break: using machine learning to uncover a critical mass in women's representation†

Kendall D. Funk[1], Hannah L. Paul[2*] and Andrew Q. Philips[2]

[1]School of Social and Behavioral Sciences, Arizona State University, Tempe, AZ, USA and [2]Department of Political Science, University of Colorado Boulder, Boulder, CO, USA
*Corresponding author. Email: hannah.paul@colorado.edu

(Received 6 February 2020; revised 9 June 2021; accepted 20 June 2021)

Abstract
Decades of research has debated whether women first need to reach a "critical mass" in the legislature before they can effectively influence legislative outcomes. This study contributes to the debate using supervised tree-based machine learning to study the relationship between increasing variation in women's legislative representation and the allocation of government expenditures in three policy areas: education, healthcare, and defense. We find that women's representation predicts spending in all three areas. We also find evidence of critical mass effects as the relationships between women's representation and government spending are nonlinear. However, beyond critical mass, our research points to a potential critical mass interval or critical limit point in women's representation. We offer guidance on how these results can inform future research using standard parametric models.

Example 1: supervised learning

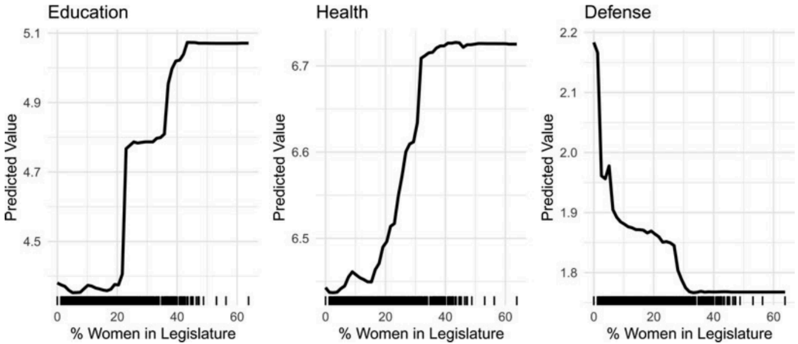
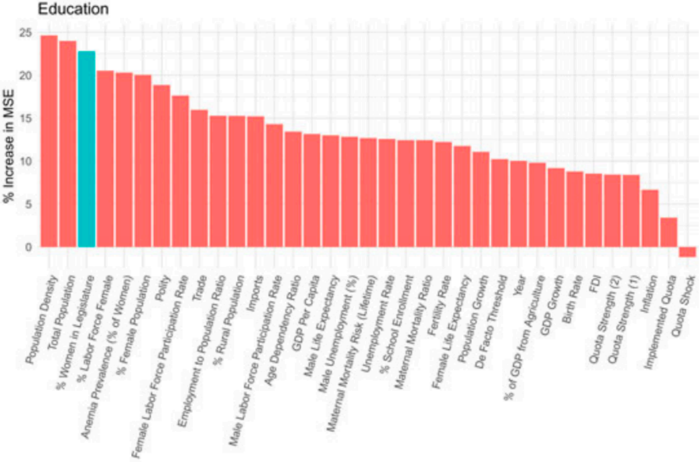


Fig. 4. PDPs of women's representation support critical mass.

Example 1: supervised learning



Example 2: unsupervised learning

This paper uses topic models to classify the content of reports:

Political Science Research and Methods Vol 6, No. 4, 661–677 October 2018

doi:10.1017/psrm.2016.44

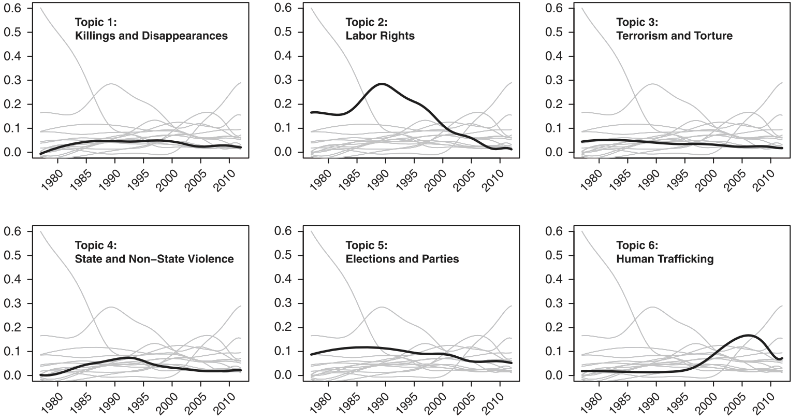
The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports*

BENJAMIN E. BAGOZZI AND DANIEL BERLINER

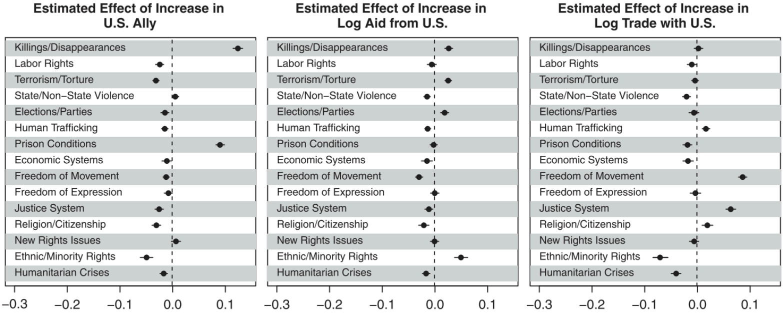
Human rights monitoring reports play important roles both in the international human rights regime and in productions of human rights data. However, human rights reports are produced by organizations subject to formal and informal pressures that may influence the topics considered salient for attention and scrutiny. We study this potential using structural topic models (STMs), a method used for identifying the latent topical dimensions of texts and assessing the effects of covariates on these dimensions. We apply STMs to a corpus of 6298 State Department Country Reports on Human Rights Practices (1977–2012), identifying a plausible set of topics including killings and disappearances, freedoms of expression and movement, and labor rights, among others. We find that these topics vary markedly both over time and space. We also find that while US domestic politics play no systematic role in shaping topic prevalence, US allies tend to receive more *attention to violations of physical integrity rights. These results challenge extant research, and illustrate the usefulness of STM methods for future study of foreign policy documents. Our findings also highlight the importance of* topical attention shifts *in documents that monitor and evaluate countries.*

Example 2: unsupervised learning

It is an unsupervised method because the authors use an unlabeled dataset. They let the model create labels of its own



Example 2: unsupervised learning



Example 3: semi-supervised learning

This paper uses neural networks to classify images and identify election fraud:

American Political Science Review (2019) 113, 3, 710–726
doi:10.1017/S0003055419000285

The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election

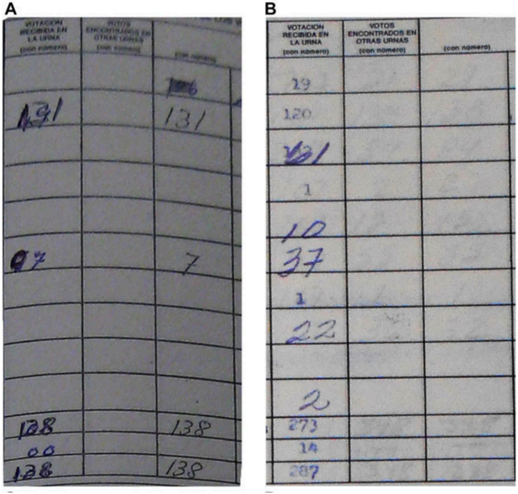
FRANCISCO CANTÚ *University of Houston*

This paper investigates the opportunities for non-democratic regimes to rely on fraud by documenting the alteration of vote tallies during the 1988 presidential election in Mexico. In particular, I study how the alteration of vote returns came after an electoral reform that centralized the vote-counting process. Using an original image database of the vote-tally sheets for that election and applying Convolutional Neural Networks (CNN) to analyze the sheets, I find evidence of blatant alterations in about a third of the tallies in the country. This empirical analysis shows that altered tallies were more prevalent in polling stations where the opposition was not present and in states controlled by governors with grassroots experience of managing the electoral operation. This research has implications for understanding the ways in which autocrats control elections as well as for introducing a new methodology to audit the integrity of vote tallies.

The author manually labeled 1,050 images as "with alterations" or "without alterations," then asked the model to make predictions for ~52,300 unlabeled images

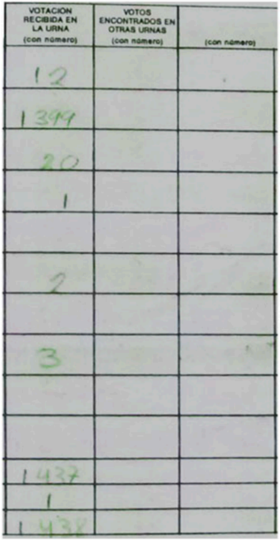
Example 3: semi-supervised learning

FIGURE 1. Examples of Vote Tallies with Alteration in Their Numbers. Mexico, 1988

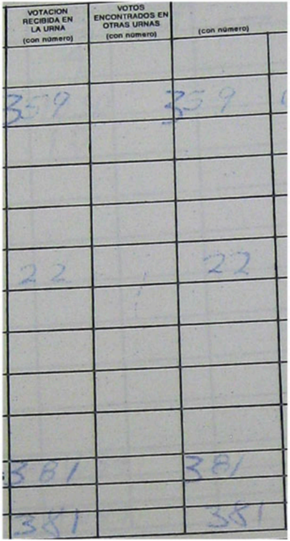


Example 3: semi-supervised learning

C

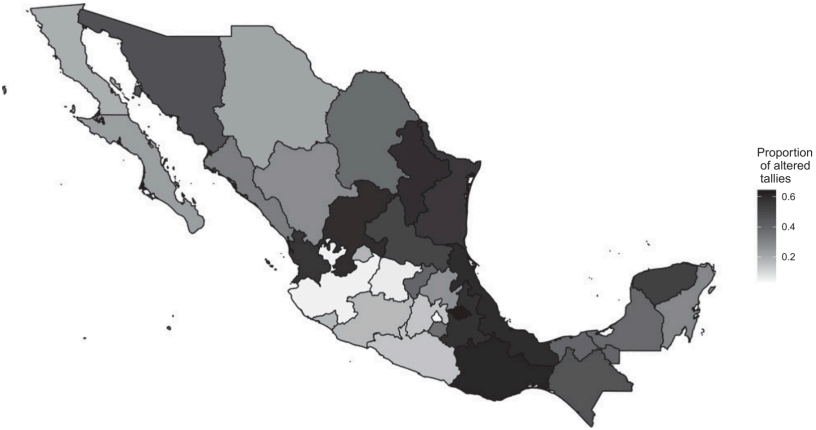


D



Example 3: semi-supervised learning

FIGURE 3. Rates of Tallies Classified as Altered by State



Notes: This figure shows the proportion of tallies in every state classified by the CNN as altered.

Example 4: semi-supervised learning

This paper uses keyword-assisted topic models to classify IMF conditionality

DOI: 10.1111/ecpo.12214

ORIGINAL ARTICLE

ECONOMICS & POLITICS WILEY

Examining the effect of IMF conditionality on natural resource policy

Iasmin Goes

The author provides some labels and keywords, but allows the model to find additional words and labels

Example 4: semi-supervised learning

Word
prices
percent
oil
petroleum
price
increase
products
gas
electricity
tariffs

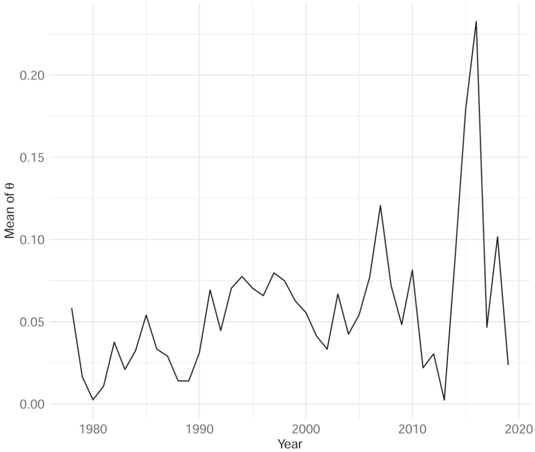
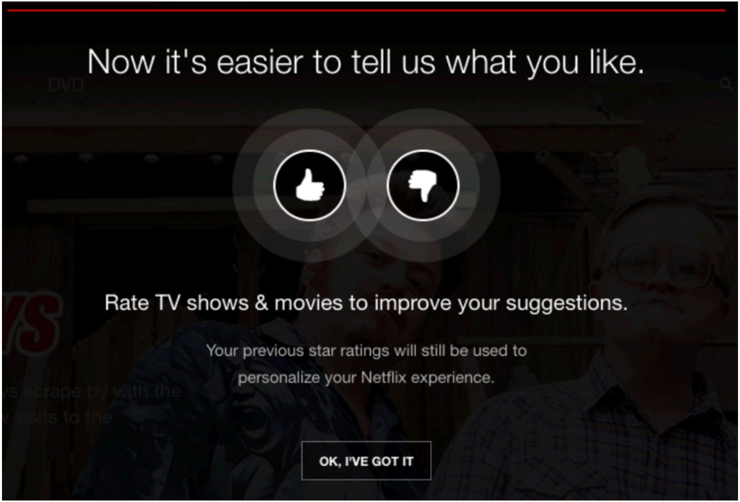


FIGURE 4 Topic prevalence over time, 1980–2019. This plot displays the prevalence of the natural resource topic over time, among all binding conditions, based on the year of program initiation (as indicated by the *x*-axis). The *y*-axis represents the relative proportion θ of this topic in each condition, averaged for all conditions over a year

Example 5: reinforcement learning

We couldn't think of any academic example. But Netflix does this



Supervised learning

Supervised tasks

- **Classification (categorical outcome)**
  - Naive Bayes
  - Clustering algorithms (kNN)
  - Logistic regression
  - Random forest
  - Gradient boosting machine
  - Support vector machine
  - Neural networks
- **Regression (continuous outcome)**
  - Lasso, ridge, and linear regression
  - Random forest
  - Gradient boosting machine
  - Support vector machine
  - Neural networks

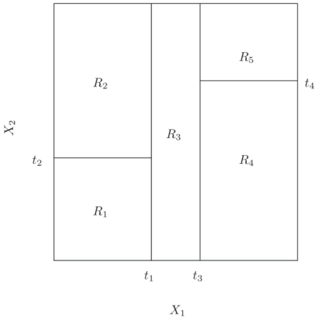
Supervised learning

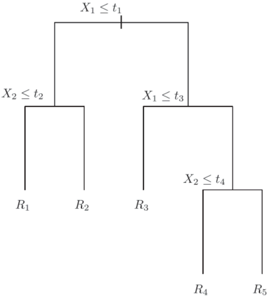
- Need to have a **labeled dataset**
- Split the data into training/test set (and ideally also validation set)
  - **Training set** (60-80% of the data): fit the model, let it learn patterns from the data
  - **Validation set** (10-20%, if you have enough observations): tune hyperparameters and evaluate model performance
  - **Test set** (10-20%): once the model is trained and tuned, make out-of-sample predictions for independent observations

In more detail: tree-based models

A decision tree

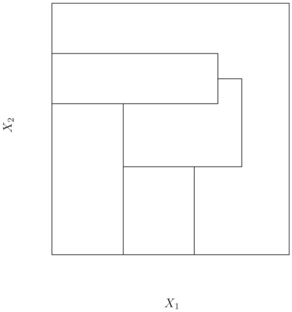
- **Goal:** split covariate space into regions, with each region corresponding to a unique covariate combination
  - The model then makes one prediction for all observations within this region

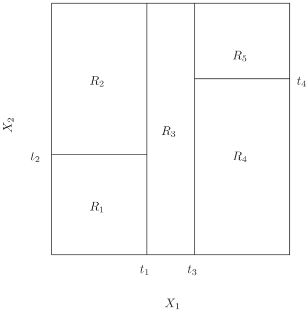




In theory, regions could have any shape

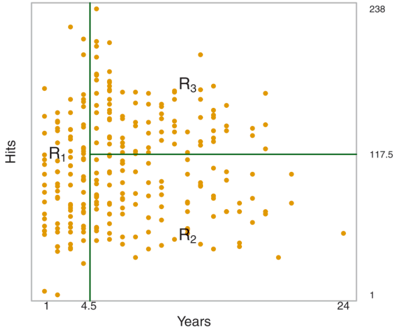
But it is computationally infeasible to consider every possible partition (left), which is why the model works with high-dimensional rectangles (right). This is called *recursive binary splitting*

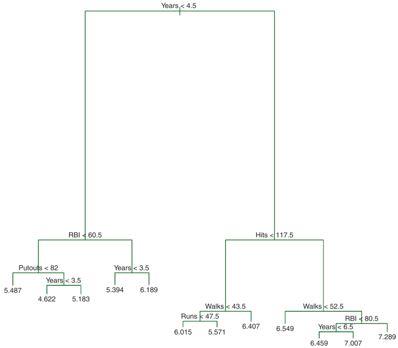




For example, the salary of baseball players

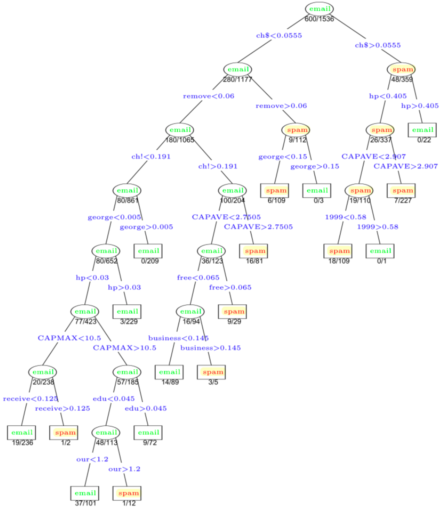
This is a *regression* tree: the outcome is quantitative





Another example: legit emails vs. spam

This is a *classification* tree: the outcome is categorical/qualitative



Multiple models are better than one model

- **Ensemble methods** combine the predictions of multiple models to create a stronger model
  - One model can make errors. Multiple models will make different errors that balance each other out
- Two common types of ensemble methods: bagging ad boosting
  - **Bagging:** train each model on a different random subset (with replacement), then aggregate them
  - **Boosting:** train models sequentially, with one model correcting the errors of the previous one

An example of bagging: random forest

- **Why forest?** Because one tree can be sensitive to data changes
- **Why random?** Because each binary split only considers a random sample of predictors
  - If there is a very strong predictor in the dataset, we don't want *all* trees to use this predictor in the first split
- The model then aggregates the results based on the predictions of most trees

Things you can explain with tree-based models

- Civil war onset (Muchlinski et al 2016)
- Supreme Court rulings (Kaufman, Kraft and Sen 2019)
- Women's legislative representation and the allocation of government expenditures (Funk, Paul and Philips 2022)
- Negative campaigning and voting (Montgomery and Olivella 2016)
- Democracy (Weitzel et al 2024a) and democratic backsliding (Weitzel et al 2024b)
- Variation in GDP data reported across different sources (Goes 2024)

Advantages of tree-based models

- No need to develop theoretical expectations
- No need to make assumptions about...
  - predictor variables
  - functional form of predictors (linear, log, squared)
- More honesty about the lack of causality
  - Regressions are not causal either, but people often interpret them causally

Disadvantages of tree-based models (and ML more broadly)

- Increased computational cost
- Need to have a large dataset
- No causal inference — ML is mostly concerned with making **predictions**
- Generally no confidence intervals or p-values to quantify uncertainty
  - **Exception:** conformal prediction
- Complexity: ML models are black boxes and difficult to interpret
  - Most social scientists likely won't know what you're doing

ML steps

- Data cleaning and preparation
  - Missing values, feature engineering (one-hot encoding, ranging, standardizing)
- Choose a model and train it
  - Select appropriate ML algorithm
  - Split the data into training, (cross-)validation, and testing sets to evaluate the model's performance
  - Train the algorithm on the prepared training set
  - Be aware of *data leakage*
- Evaluate the model
  - Evaluate the model's performance on the (cross-)validation set
  - Fine-tune it as necessary to improve accuracy

What is Causality?

- Definition:
  - Causality is the study of cause and effect relationships. It helps us answer questions like, "What would happen to Y (outcome) if we changed X (treatment)?"

What is Causality?

- Definition:
  - Causality is the study of cause and effect relationships. It helps us answer questions like, "What would happen to Y (outcome) if we changed X (treatment)?"
- Key Concept:
  - Counterfactual: The hypothetical scenario of what would happen to Y if X had not occurred.

What is Causality?

- Definition:
  - Causality is the study of cause and effect relationships. It helps us answer questions like, "What would happen to Y (outcome) if we changed X (treatment)?"
- Key Concept:
  - Counterfactual: The hypothetical scenario of what would happen to Y if X had not occurred.
- Example:
  - Imagine a government increases minimum wage (X). The causal question is: "What is the effect of increasing the minimum wage on employment rates (Y)?"
  - The challenge is determining whether observed changes in employment are due to the policy or other factors.

Why is Causality Important?

- Prediction vs. Causation:
  - Prediction: Estimates outcomes based on patterns (e.g., predicting heart disease risk based on age, cholesterol).

Why is Causality Important?

- Prediction vs. Causation:
  - Prediction: Estimates outcomes based on patterns (e.g., predicting heart disease risk based on age, cholesterol).
- Causation: Explains the effect of an intervention or change (e.g., what is the impact of lowering cholesterol on heart disease risk?).

Why is Causality Important?

- Prediction vs. Causation:
  - Prediction: Estimates outcomes based on patterns (e.g., predicting heart disease risk based on age, cholesterol).
- Causation: Explains the effect of an intervention or change (e.g., what is the impact of lowering cholesterol on heart disease risk?).
- Example:
  - In medicine, identifying the effect of a new drug (causation) is far more critical than just predicting who will get sick (prediction).

Challenges in Identifying Causal Effects

- Confounding Variables: These are variables that affect both the cause (X) and the effect (Y), creating bias.
  - Example: If we want to study the effect of exercise (X) on weight loss (Y), diet could be a confounder that influences both.

Challenges in Identifying Causal Effects

- Confounding Variables: These are variables that affect both the cause (X) and the effect (Y), creating bias.
  - Example: If we want to study the effect of exercise (X) on weight loss (Y), diet could be a confounder that influences both.
- Selection Bias: Occurs when the participants or data points selected are not representative of the target population.
  - Example: Studying the effect of a new teaching method on students' grades, but only using top-performing schools as your sample.

Challenges in Identifying Causal Effects

- Confounding Variables: These are variables that affect both the cause (X) and the effect (Y), creating bias.
  - Example: If we want to study the effect of exercise (X) on weight loss (Y), diet could be a confounder that influences both.
- Selection Bias: Occurs when the participants or data points selected are not representative of the target population.
  - Example: Studying the effect of a new teaching method on students' grades, but only using top-performing schools as your sample.
- Endogeneity: This occurs when explanatory variables are correlated with the error term.
  - Example: Studying the effect of income on education, where education itself can influence income (creating a two-way relationship).

The Book of Why

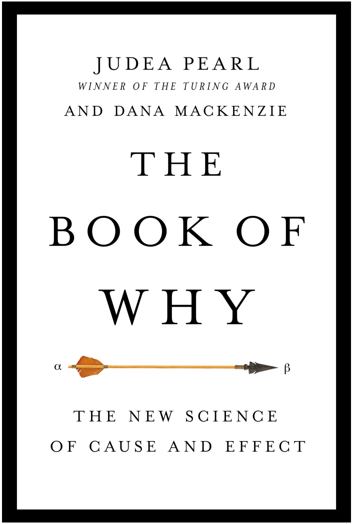


Figure 1: The Book of Why

Counterfactuals and Causal Inference

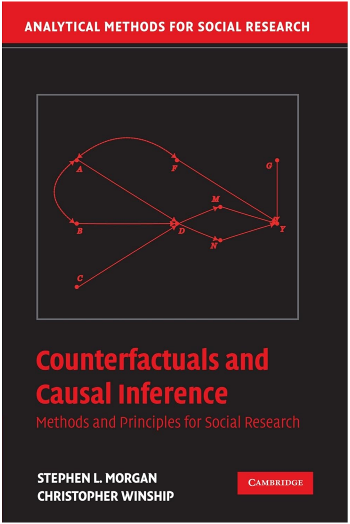


Figure 2: Counterfactuals and Causal Inference

What Machine Learning Does Well

- Pattern Recognition: ML excels at finding patterns in large, high-dimensional datasets (e.g., predicting housing prices based on location, size, amenities).

What Machine Learning Does Well

- Pattern Recognition: ML excels at finding patterns in large, high-dimensional datasets (e.g., predicting housing prices based on location, size, amenities).
- Prediction Power: Algorithms like random forests, gradient boosting, and neural networks can capture complex, non-linear relationships.
  - Example: Using ML to predict customer churn in a telecom company based on usage data, payment history, and customer service interactions.

What Machine Learning Does Well

- Pattern Recognition: ML excels at finding patterns in large, high-dimensional datasets (e.g., predicting housing prices based on location, size, amenities).
- Prediction Power: Algorithms like random forests, gradient boosting, and neural networks can capture complex, non-linear relationships.
  - Example: Using ML to predict customer churn in a telecom company based on usage data, payment history, and customer service interactions.
- Handling High Dimensionality: ML can incorporate thousands of variables (features) and model complex interactions.
  - Example: Predicting disease outcomes from genetic data involving thousands of genes.

Where Machine Learning Falls Short in Causality

- Correlation != Causation:
  - Example: A ML model may predict that people who buy umbrellas also buy raincoats. However, this doesn't mean buying umbrellas causes raincoats to be purchased. The true cause could be rainy weather, a confounding factor.

Where Machine Learning Falls Short in Causality

- Correlation != Causation:
  - Example: A ML model may predict that people who buy umbrellas also buy raincoats. However, this doesn't mean buying umbrellas causes raincoats to be purchased. The true cause could be rainy weather, a confounding factor.
- Omitted Variables: ML models may leave out variables that are crucial for understanding causal effects.
  - Example: A model predicting wage increases based on experience may omit education, a critical confounder.

Where Machine Learning Falls Short in Causality

- Correlation != Causation:
  - Example: A ML model may predict that people who buy umbrellas also buy raincoats. However, this doesn't mean buying umbrellas causes raincoats to be purchased. The true cause could be rainy weather, a confounding factor.
- Omitted Variables: ML models may leave out variables that are crucial for understanding causal effects.
  - Example: A model predicting wage increases based on experience may omit education, a critical confounder.
- Interpretability: Many ML models are "black boxes," making it difficult to interpret causal pathways.
  - Example: A neural network predicting housing prices may perform well, but it's challenging to understand which factors (location, crime rates, etc.) directly influence the prices.

How Causal Machine Learning Tackles Causality

- Counterfactual Predictions:
  - Example: A company tests a new advertising campaign. Using uplift modeling, we can predict the difference in customer behavior if they hadn't been exposed to the ad.

How Causal Machine Learning Tackles Causality

- Counterfactual Predictions:
  - Example: A company tests a new advertising campaign. Using uplift modeling, we can predict the difference in customer behavior if they hadn't been exposed to the ad.
- Treatment Effect Estimation:

How Causal Machine Learning Tackles Causality

- Counterfactual Predictions:
  - Example: A company tests a new advertising campaign. Using uplift modeling, we can predict the difference in customer behavior if they hadn't been exposed to the ad.
- Treatment Effect Estimation:
- Causal Forests: A method to estimate how the treatment effect varies across individuals (heterogeneous treatment effects).
  - Example: A healthcare intervention might work better for older patients than younger ones, and causal forests help us understand this variation.

How Causal Machine Learning Tackles Causality

- Counterfactual Predictions:
  - Example: A company tests a new advertising campaign. Using uplift modeling, we can predict the difference in customer behavior if they hadn't been exposed to the ad.
- Treatment Effect Estimation:
- Causal Forests: A method to estimate how the treatment effect varies across individuals (heterogeneous treatment effects).
  - Example: A healthcare intervention might work better for older patients than younger ones, and causal forests help us understand this variation.
- Propensity Score Methods: Use ML to estimate the likelihood of receiving treatment based on observable characteristics, then adjust to mimic a randomized experiment.
  - Example: Estimating the effect of a job training program by comparing participants with similar non-participants after adjusting for demographics and prior job experience.

Causal Inference Frameworks in Machine Learning

- A/B Testing (Randomized Experiments):
  - Example: A tech company tests two versions of a webpage to see which one leads to more purchases. Randomization ensures that any difference in outcomes is likely due to the webpage design.

Causal Inference Frameworks in Machine Learning

- A/B Testing (Randomized Experiments):
  - Example: A tech company tests two versions of a webpage to see which one leads to more purchases. Randomization ensures that any difference in outcomes is likely due to the webpage design.
- Instrumental Variables (IV): An external variable (instrument) influences the cause (X) but not the outcome (Y), allowing isolation of causal effects.
  - Example: Using distance to school as an IV for the effect of schooling on earnings, since distance affects education but not directly income.

Causal Inference Frameworks in Machine Learning

- A/B Testing (Randomized Experiments):
  - Example: A tech company tests two versions of a webpage to see which one leads to more purchases. Randomization ensures that any difference in outcomes is likely due to the webpage design.
- Instrumental Variables (IV): An external variable (instrument) influences the cause (X) but not the outcome (Y), allowing isolation of causal effects.
  - Example: Using distance to school as an IV for the effect of schooling on earnings, since distance affects education but not directly income.
- Difference-in-Differences (DiD): Compare changes over time between a treatment and control group to isolate the causal effect of an intervention.
  - Example: Studying the effect of a new tax law by comparing business profits before and after the law in states that adopted it versus those that didn't.

Causal Machine Learning Methods

- Causal Trees and Causal Forests:
  - Example: Suppose a health intervention is given to different subgroups of patients. A causal forest can estimate treatment effects for each subgroup (e.g., older vs. younger patients), allowing for personalized treatment recommendations.

Causal Machine Learning Methods

- Causal Trees and Causal Forests:
  - Example: Suppose a health intervention is given to different subgroups of patients. A causal forest can estimate treatment effects for each subgroup (e.g., older vs. younger patients), allowing for personalized treatment recommendations.
- Double/Debiased ML:
  - Combines ML models for the outcome and treatment to reduce bias in estimating causal effects.
  - Example: Estimating the effect of college attendance on future earnings by using ML to predict both the probability of attending college and the resulting earnings.

Causal Machine Learning Methods cont.

- Uplift Modeling:
  - Predicts the incremental effect of a treatment on individual outcomes.
  - Example: A marketing campaign to increase product sales — uplift modeling predicts which customers are most likely to increase their spending because of the campaign.

Trade-offs in Causal Machine Learning

- Accuracy vs. Interpretability:
  - Example: Neural networks may outperform simpler models in accuracy but are hard to interpret, making it difficult to understand how the model arrived at its causal conclusions.

Trade-offs in Causal Machine Learning

- Accuracy vs. Interpretability:
  - Example: Neural networks may outperform simpler models in accuracy but are hard to interpret, making it difficult to understand how the model arrived at its causal conclusions.
- Assumptions:
  - Example: Many causal methods (e.g., propensity scores, IV) assume no unobserved confounders, which can be hard to justify in practice. Violating this assumption can lead to biased results.

Trade-offs in Causal Machine Learning

- Accuracy vs. Interpretability:
  - Example: Neural networks may outperform simpler models in accuracy but are hard to interpret, making it difficult to understand how the model arrived at its causal conclusions.
- Assumptions:
  - Example: Many causal methods (e.g., propensity scores, IV) assume no unobserved confounders, which can be hard to justify in practice. Violating this assumption can lead to biased results.
- Generalizability:
  - Example: A model predicting the effect of job training in one region may not generalize well to another region with a different economic context.

Conclusion

- Key Takeaways:
  - Causal inference seeks to understand "why" something happens, going beyond predictive models.
  - Machine learning offers powerful tools to address causal questions but requires careful consideration of confounders and assumptions.
  - Methods like causal forests and double ML are helping bring predictive accuracy and causal insights together.

Books, packages, and articles

- Introduction to Statistical Learning, great introduction
- Elements of Statistical Learning, more advanced
- List of R packages for Machine Learning, CRAN Task View List
- Machine Learning Methods That Economists Should Know About, academic article with an overview

An applied example

Potential R packages for machine learning

```r
## Libraries
library(caret)        # classif and reg training
library(tidymodels)   # ML using the tidyverse
library(randomForest) # random forests
library(gbm)          # gradient boosting machines
library(e1071)        # support vector machines
library(nnet)         # neural networks
library(H2O)          # EVERYTHING
```

Our code for today

https://t.ly/ZivLy