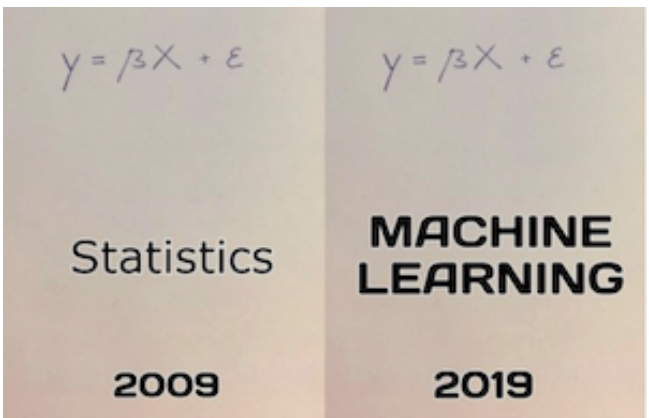
Causal Machine Learning in Practice

Estimating Average and Heterogeneous Effects for Personalized Treatment

Jacob Pieniazek

Lead Data Scientist (84.51°)
Presentation reflects my personal views and not those of my employer.

Miami University Economics Brown Bag
Spring 2026



#10yearchallenge

Bio

A bit about myself

- BS, Economics and Mathematics at University of Dayton (2021)
- MA, Economics at Miami University (2022)
- Presently, Lead Data Scientist at 84.51° (Kroger's data science and analytics subsidiary)
- At 84.51°, I primarily:
  - Lead a team of data scientists developing a production causal inference / experimentation ("measurement") platform for internal marketing use across the organization
  - Partner with business stakeholders upstream to make campaign design more *measurement-aware* / causal-first when treatment assignment is still in our control
  - Lead R&D for CATE/Uplift modeling for personalized/optimized treatment prescription in marketing applications

BS Econ + Math → MA Economics → Lead Data Scientist

Roadmap

Theory first, Practical Interludes, then Application

Running question: Did the treatment work, for whom did it work, and how should that change the next decision?

Part I: Foundations

- Motivation and causal setup
- Why machine learning for causal inference?
- Average effects and inference

Part II: Heterogeneity

- Modeling heterogeneous treatment effects
- Personalized treatment decisions

Part III: Application

- Stylized, simulated examples inspired by real-world marketing applications
- Highlight practical challenges and considerations in real-world applications

Measure average effects → Learn who benefits most → Use estimates for decisions

IN PRACTICE Interludes on *assumptions in the wild*, *scaling ATE measurement*, and *productionalizing heterogeneity modeling and personalization*.

Main Goal

Combine flexible, modern machine learning with the causal framing and inferential discipline economists care about, and show why that combination is especially useful in industry and marketing settings.

Part I

From Classical Regression to Double/Debiased Machine Learning for Average Effects

A Stylized Running Example

Marketing campaign effectiveness as a marketing measurement problem

Incremental Effect of Exposure

- Y: customer spend in the next 30 days
- D: exposure to a marketing campaign (ad, offer, etc.)
- X: high-dimensional customer context

Causal targets

$$\text{ATE} = \theta_0 = \mathbb{E}[Y(1) - Y(0)]$$
$$\text{CATE}(x) = \tau_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Why this can be difficult

- Ad exposures are sometimes non-randomized & highly targeted, so traditional methods inherit strong confounding bias.
- The propensity score $m_0(X)$, outcome surface $g_0(X)$, and CATE function $\tau_0(X)$ can be complex and nonlinear.

What can make X complex? The confounding and heterogeneity structures often live in a mix of structured variables, high-dimensional features, and learned representations.

Structured

- RFM & other aggregated features
- Customer segmentations & pre-engineered features

High-Dim Tabular

- Time-Series of past purchases, spend, and visits
- SKU-level purchase vectors
- Digital behavior tables (e.g., clicks, searches, etc.)

Representations

- Image embeddings of viewed products & ads
- Sequence embeddings from clickstream sessions

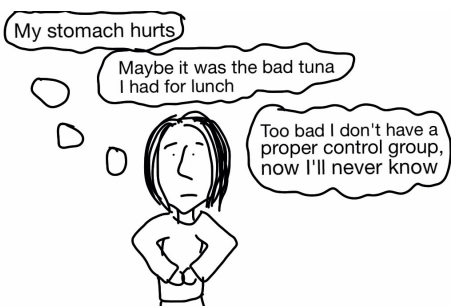
Key Questions

- Did the campaign work on average?
- Did it work better for some customers than others?
- How should we use that information next time?

Potential Outcomes and Notation

Some common language for the causal setup within cross-sectional regime

- For each unit i, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control.
- We observe $Y_i = Y_i(D_i)$, treatment D_i, and covariates X_i.
- Shorthand: $W_i = (Y_i, D_i, X_i)$.
- Fundamental problem: for any given unit, we never observe both $Y_i(1)$ and $Y_i(0)$.



Potential outcomes framing: Rubin (1974).

Estimands and Assumptions

Key objects we'll be working with

Core estimands

Average Treatment Effect (ATE):

$$\theta_0 = \mathbb{E}[Y(1) - Y(0)]$$

Conditional Average Treatment Effect (CATE):

$$\tau_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Core assumptions

- Well-defined treatment / no interference
- Consistency: $Y = Y(D)$
- Unconfoundedness: $(Y(1), Y(0)) \perp D \mid X$
- Overlap: $0 < m_0(X) = P(D = 1 \mid X) < 1$

Important reminder

The machine learning techniques discussed can help us estimate complicated nuisance (to be defined) structure, but it does **NOT** not create identification out of thin air.

In observational settings, unconfoundedness is a substantive assumption (aka, ignorability, exogeneity, etc.). No interference means one customer's treatment does not affect another's outcome. We use ATE generically for simplicity in place of other estimands, such as ATT.

Assumptions In The Wild

How this can show up in industry

IN PRACTICE

When assignment is still in play

- We often have some influence upstream before intervention, so we push hard for a **measurement-aware** / **causal-first** design, implementing RCT-style designs wherever possible.
- IVs, imperfect randomization, or simple rule-based variation is usually much better than trying to rescue a weak design after the fact.
- Many times true randomization is not feasible or practical, or simply not considered in advance.

When the design is already fixed or past

- Then assumptions require investigation & the search for identification begins: who got treated, why, and what the business knew at assignment time.
- Overlap failures, spillovers, contamination, and delivery / measurement issues can dominate the analysis.
- “Not identifiable” rarely ends the conversation; it challenges automation of measurement & changes the identification strategy.

Econometric Rigor vs. Business Practicality. Many times, the ideal assumptions for clean identification are not met, and you have to make do with what you have within reasonable timelines, with reasonably explainable models. The business must often make decisions based on imperfect information. *This is a key tension in "real-world" causal inference work.*

A First Pass: Regression Adjustment for the ATE

What many of us would do first in practice

$$Y = \alpha + \theta_0 D + X'\beta_0 + \varepsilon$$

Interpretation

- Regress the outcome on treatment plus observed controls.
- Read $\hat{\theta}$ as the ATE.
- In a randomized experiment, this is generally a sufficient approach.

What has to go right?

- Confounding is addressed by conditioning on X.
- The linear/additive specification is good enough to control for any complex, non-linear confounding.
- The control set is small enough to estimate reliably ($p \ll n$)

Why this is a useful benchmark

This is the familiar econometrics baseline: transparent, interpretable, and often effective in low-dimensional settings.

Why Move Beyond Classical Low-Dimensional Regression?

Why causal ML enters the picture

Classical regression is comfortable when...

- there are relatively few controls relative to n,
- functional forms and confounding processes are simple,
- controls are mostly tabular and of low-dimensionality,
- interactions are known in advance,
- a hand-specified model feels credible.

But practice often looks more like...

- high-dimensional and complex confounding (possibly $p \gg n$),
- nonlinear response surfaces,
- unstructured data (e.g., text, images, etc.),
- many plausible interactions,
- unknown functional forms of heterogeneity (more on this later).

We want flexible prediction tools for confounding adjustment, but we still want causal effect estimation and valid inference.

The Partially Linear Regression Model

A useful backbone for a low-dimensional causal target

$$Y = D\theta_0 + g_0(X) + \zeta, \qquad \mathbb{E}[\zeta \mid D, X] = 0$$
$$D = m_0(X) + V, \qquad \mathbb{E}[V \mid X] = 0$$

- θ_0 is the constant treatment-effect (ATE) in the PLR model.
- $g_0(X)$ captures how observables shape the outcome.
- $m_0(X)$ captures how observables shape treatment assignment.
- The nuisance functions, g_0 and m_0, can be complicated even when the target parameter is simple.
- This is the semiparametric backbone we'll build from: a low-dimensional target parameter with high-dimensional, flexible nuisance structure.
- The previous OLS regression with controls is a special case of this model when $g_0(X) = X'\beta_0$ and $m_0(X) = X'\pi_0$.

The term "nuisance functions" exists to capture that estimation of these components is not of primary concern. They exist as a means to an end (θ_0), rather than as ends in themselves.

Semiparametric backbone: Robinson (1988); DML treatment: Chernozhukov et al. (2018).

Why Naive ML Plug-In Fails

Good prediction is not enough

Regularization bias

If nuisance estimates are biased because the learner shrinks, selects, or smooths aggressively, that bias can spill into the treatment effect estimate (e.g., it can shrink in a "confounder-unaware" fashion).

Overfitting bias

Traditional ML is notoriously prone to overfitting, in which a learner captures noise in the sample and generalization becomes hindered. This overfitting bias can introduce bias into the treatment effect estimate.

Good prediction alone does not guarantee root-n valid causal inference for causal targets. We need to ensure:

1. We sufficiently model complexity in the nuisance functions to control for confounding, but
2. We don't regularize away key confounding variation, and
3. We don't allow our learner to be too complex such that it overfits the training data

Bias transmission: Chernozhukov et al. (2018).

Mitigating Regularization Bias

via Neyman Orthogonality

Classical Bridge: Frisch-Waugh-Lovell (FWL) Partialling-Out

When the nuisance models are linear regressions

$$Y = \alpha_0 + \theta_0 D + X'\beta_0 + \varepsilon$$

$$\ell_0(X) = X'\gamma_0, \qquad m_0(X) = X'\pi_0$$

$$\tilde{Y} = Y - \ell_0(X), \qquad \tilde{D} = D - m_0(X)$$

$$\tilde{Y} = \theta_0 \tilde{D} + \varepsilon$$

$$\hat{\theta}_{\text{OLS with controls}} = \hat{\theta}_{\text{OLS of } \tilde{Y} \text{ on } \tilde{D}}$$

- Regress Y on X, keep the residuals.
- Regress D on X, keep the residuals.
- Regress residualized Y on residualized D.
- You get exactly the same coefficient and variance estimates as full OLS with controls.

Takeaway

FWL is the familiar econometrics bridge. We can generalize this partialling-out intuition.

A Brief Detour: Generalized Method of Moments & Scores

A more general estimation framework

After partialling-out:

$$\tilde{Y} = \theta_0 \tilde{D} + \varepsilon$$

Define the *score* or *moment function*:

$$\psi(W; \theta_0) = \tilde{D}(\tilde{Y} - \theta_0 \tilde{D})$$

Define the moment condition (OLS orthogonality condition):

$$M(\theta_0) = \mathbb{E}[\tilde{D}\varepsilon] = \mathbb{E}[\psi(W; \theta_0)] = 0$$
$$\text{where} \quad \varepsilon = \tilde{Y} - \theta_0 \tilde{D}$$

And solve for the parameter (closed-form in this case of OLS):

$$\mathbb{E}[\psi(W; \theta_0)] = 0 \quad \Rightarrow \quad \theta_0 = \frac{\mathbb{E}[\tilde{D}\tilde{Y}]}{\mathbb{E}[\tilde{D}^2]}$$

Generalized Method of Moments

Generalized estimation technique commonly applied in the context of semi-parametric models, where the parameter of interest is finite-dimensional and the full shape of the data's distribution function may not be known, and therefore maximum likelihood estimation is not applicable.

Solving the moment condition

Choose $\hat{\theta}$ such that sample analog of the moment equals zero. This can be done via closed-form solution, numerical optimization, or other methods depending on the structure of the moment condition.

This framework will be carried forward in our discussion as we construct scores ψ for more complex models.

Residual-on-Residual Intuition

Within the PLR model

Define nuisance functions and residuals:

$$\ell_0(X) = \mathbb{E}[Y \mid X], \qquad m_0(X) = \mathbb{E}[D \mid X]$$

$$R_Y = Y - \ell_0(X), \qquad R_D = D - m_0(X)$$

Define the score:

$$\psi(W; \theta_0, \eta_0) = R_D(R_Y - \theta_0 R_D)$$
$$\text{where} \quad \eta_0 = (\ell_0, m_0)$$

Define population moment condition:

$$M(\theta_0, \eta_0) = \mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0$$

Similarly, solve for the parameter:

$$\mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0 \quad \Rightarrow \quad \theta_0 = \frac{\mathbb{E}\left[R_D R_Y\right]}{\mathbb{E}\left[R_D^2\right]}$$

- First partial out the variation explained by X using flexible learners for $\ell_0(X)$ and $m_0(X)$.
- Then estimate the effect using the residualized treatment and outcome.
- This is a generalization of the FWL idea to the semi-parametric PLR model; see Robinson (1988).
- The score ψ is *Neyman Orthogonal*; see Chernozhukov et al. (2018). Adapted proof in Appendix.

Neyman Orthogonality

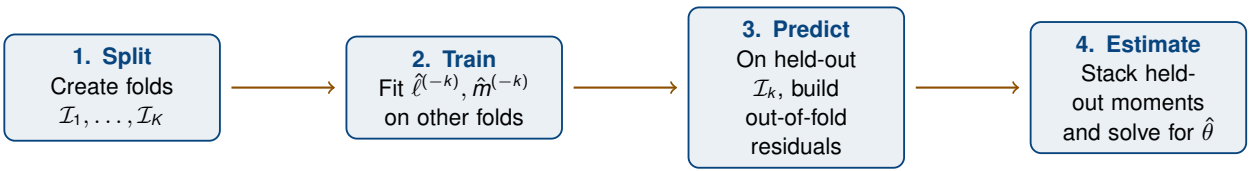
Neyman orthogonality ensures that small errors in estimating nuisance functions η_0 have minimal impact on the target parameter. In other words, the estimate of θ_0 is robust to first-order bias in the nuisance estimation.

Mitigating Overfitting Bias

via Cross-Fitting

Cross-Fitting

Separating training from evaluation



$$\hat{R}_{Y,i} = Y_i - \hat{\ell}^{(-k)}(X_i), \qquad \hat{R}_{D,i} = D_i - \hat{m}^{(-k)}(X_i), \qquad i \in \mathcal{I}_k$$

If we do not cross-fit

- **Data reuse:** the same data train and evaluate the nuisance models.
- **Self-influence:** observation *i* helps fit the model used to residualize itself.
- **Result:** flexible ML can make in-sample residuals look too optimistic, distorting the moment equation.

Why it helps

- **Out of fold:** each observation is residualized using models trained on other folds only.
- **Overfitting bias mitigated:** the learner cannot capture idiosyncratic noise in the fold it is evaluated on.
- **Payoff:** Overfitting bias is pushed to higher order, supporting unbiased estimation & valid inference for θ_0.

Double/Debiased Machine Learning

Putting the pieces together

The Double/Debiased Machine Learning Framework

Three key ingredients from generic DML

1 Neyman Orthogonality

Using a score function $\psi(W; \theta, \eta)$ such that

1. $M(\theta, \eta) = \mathbb{E}[\psi(W; \theta, \eta)]$ identifies θ when $\eta = \eta_0$, and
2. the Neyman orthogonality condition is satisfied.

$$\partial_\eta M(\theta_0, \eta)|_{\eta=\eta_0} = 0$$

Eliminates first-order bias from nuisance estimation.

2 High-Quality ML Models

The use of high-quality machine learning estimators of the nuisance parameters, with a sufficient convergence rate of $o_p(n^{-1/4})$, which is achievable by many modern ML methods under reasonable conditions.

Renders second-order bias from nuisance estimation negligible.

3 Cross-Fitting

Use sample splitting, where nuisance functions are estimated on different data than are used in their evaluation when producing the estimator of the main parameter θ_0.

Mitigates overfitting bias from flexible nuisance function estimation.

What does this give us?

An unbiased and root-n consistent estimator $\hat{\theta}$ of the target parameter θ_0, with an asymptotically normal distribution that allows for valid inference, even when using flexible machine learning methods to estimate the nuisance functions!!

Generic DML framing: Chernozhukov et al. (2024, Sec. 9.4); formal reference: Chernozhukov et al. (2018).

A Quick Note: DML Is a Generalized Score-Based Framework

Same workflow, different orthogonal scores

Once you have an orthogonal score, the rest of the workflow is portable. Change the score to match the target, then reuse ML nuisance estimation and cross-fitting.

Examples of Neyman Orthogonal Scores

Non-exhaustive selection

PLR	Partially Linear Regression	Chernozhukov et al. (2018)
PLIV	Partially Linear IV Regression	Chernozhukov et al. (2018)
IRM	Interactive Regression Model	Chernozhukov et al. (2018)
IIVM	Interactive IV Model / LATE	Chernozhukov et al. (2018)
LPLR	Logistic Partially Linear Regression	Liu et al. (2021)
PLPR	Partially Linear Panel Regression	Clarke and Polselli (2026)
DiD	Difference-in-differences extensions	Callaway and SantAnna (2021) and Chang (2020)

From One ATE to Repeatable Measurement

Scaling ATE estimation

IN PRACTICE Interlude

The productionalization challenge

- Thousands of marketing campaigns can be executed throughout the year and quick, repeatable measurement is critical for learning and iteration.
- The business cares about efficiency, scale, and automated reporting, not just one-off estimates.
- Distributed data prep, pipelining, & orchestration in python via `PySpark` / Databricks, then estimation via open source packages, such as `statsmodels`, `DoubleML`, or `EconML`.
- This requires significant coordination across teams & systems, from data engineering to data science to business stakeholders.

The dangers of automated econometrics

- Automation standardizes execution; it does **not** create identification.
- One-size-fits-all pipelines require consistency in experimental design, known patterns, and data contracts/quality assurances.
- Guardrails still matter: overlap, effective sample size, contamination, and implausibly noisy estimates.
- Non-standard quasi-experimental designs typically require bespoke measurement approaches and cannot always be fully automated. We look for common, repeatable patterns.

Operational lesson. The estimator is typically the easy part with proper design; the hard part is standardizing the designs and building the pipelines to execute them at scale with high data quality.

From PLR to IRM

A bridge to fully heterogeneous effects

- In many applications, treatment is binary: email vs. no email, coupon vs. no coupon, eligibility vs. no eligibility.
- The PLR is a useful backbone, but it imposes a constant/additive treatment effect.
- Why move? PLR treats the effect as common across units; IRM lets treatment effects vary with observables.
- The interactive regression model (IRM) lets outcome surfaces differ flexibly across heterogenous effects.

$$g_0(d, x) = \mathbb{E}[Y \mid D = d, X = x], \qquad m_0(X) = P(D = 1 \mid X)$$
$$\tau_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = g_0(1, x) - g_0(0, x), \qquad \theta_0 = \mathbb{E}[\tau_0(X)]$$

Interpretation

IRM is a natural bridge from average effects to heterogeneous effects. That said, PLR is not as restricted as it appears, as one can include flexible transformations of X and interactions between D and X.

DML treatment of IRM: Chernozhukov et al. (2018).

A Moment for IRM?

Finding Neyman orthogonality in the fully heterogeneous case

Given:

$$g_0(d, x) = \mathbb{E}[Y \mid D = d, X = x], \qquad m_0(X) = P(D = 1 \mid X)$$
$$\tau_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = g_0(1, x) - g_0(0, x), \qquad \theta_0 = \mathbb{E}[\tau_0(X)]$$

It is tempting to construct our score as follows:

$$\psi(W; g_0, \theta_0) = g_0(1, X) - g_0(0, X) - \theta_0$$

with corresponding moment condition $\mathbb{E}[\psi(W; g_0, \theta_0)] = 0$.

But, as we have discussed, this naive plug-in approach is not robust to regularization bias in the estimation of g_0. One can verify this score is **not** Neyman Orthogonal. Adapted proof in Appendix.

So, how can we construct a Neyman Orthogonal score for the IRM? We will see that the AIPW/doubly robust signal provides a solution.

AIPW / Doubly Robust Intuition

A signal we will see again later

$$\phi(W) = \underbrace{g_0(1,X) - g_0(0,X)}_{\text{outcome model}} + \underbrace{\frac{D\,(Y - g_0(1,X))}{m_0(X)} - \frac{(1-D)\,(Y - g_0(0,X))}{1 - m_0(X)}}_{\text{IPW correction (propensity)}}$$

- This combines **outcome modeling** and **inverse-probability weighting**.
- Under unconfoundedness and overlap, $\mathbb{E}[\phi(W) \mid X] = \tau_0(X)$ and $\mathbb{E}[\phi(W)] = \theta_0$.
- It is doubly robust: consistency survives if either the outcome model or the propensity model is consistently estimated.

Given $\mathbb{E}[\phi(W)] = \theta_0$, we can construct a Neyman Orthogonal score as follows (adapted proof in appendix):

$$\psi(W; \theta_0, \eta_0) = \phi(W) - \theta_0, \qquad \eta_0 = (g_0, m_0)$$

and corresponding moment condition $M(\theta_0, \eta_0) = \mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0$.

See Robins et al. (1994) for the original AIPW estimator, and Chernozhukov et al. (2018) for the DML treatment of IRM.

Part II

Heterogeneity, Validation, and Personalized Decisions

Why Move Beyond the ATE?

Average effects can hide a lot

- A positive ATE does not imply the treatment helps everyone.
- A near-zero ATE does not imply the treatment helps no one.
- In many applications, the policy question is really: who should receive treatment, and why?

Practical interpretation

Heterogeneity estimation is valuable for segmentation, ranking, and personalized treatment decisions.

CATE motivation and policy relevance: Semenova and Chernozhukov (2021); Chernozhukov et al. (2024).

Heterogeneity Starts with Interactions

The classical regression baseline

$$Y = \alpha + X'\gamma_0 + \theta_0 D + \delta_0 (D \times Z) + u$$

$$\tau_0(z) = \theta_0 + \delta_0 z$$

- This is the familiar econometrics move: let the treatment effect vary with observables through interaction terms.
- If Z is a subgroup indicator or one selected covariate from X, we get an interpretable heterogeneous effect pattern.
- Stacking several interactions, $D \times p(X)$, gives richer heterogeneity but also more variance and more room for functional-form mistakes.
- In observational data, causal interpretation still requires unconfoundedness and overlap.

Bridge to causal ML

Causal ML keeps this interaction intuition, but lets the heterogeneity surface and the confounding adjustment become much more flexible.

Precision in Estimation Heterogenous Effects (PEHE)

The ideal oracle loss for pointwise CATE accuracy

$$\mathrm{PEHE}(\tau) = \mathbb{E}\left[(\tau(X) - \tau_0(X))^2\right]$$

- This is simply the MSE if we observed the true causal label for each unit, and would be the natural way to score CATE accuracy.
- In simulations, PEHE is a clean benchmark for pointwise effect recovery.
- In applications, it is better viewed as an oracle benchmark rather than a directly computable test loss.
- That gap is what makes both CATE estimation and CATE validation harder than ordinary prediction.

Where we are headed

The rest of the section is about feasible stand-ins for this infeasible objective.

Why CATE Validation Is Harder

PEHE is natural in theory, but unavailable in real data

Standard prediction

- Observe the label Y.
- Compute test MSE, AUC, calibration, and so on.
- Estimation and validation use the same observed target.

CATE estimation

- The individual causal label $Y(1) - Y(0)$ is never observed.
- PEHE is not directly computable outside simulations.
- Estimation needs surrogate losses, and validation must be indirect.

This is one of the biggest conceptual differences between causal ML and standard ML.

Overlap, nuisance estimation, and causal assumptions all matter here, so validation is not as assumption-light as test MSE or AUC.

Validation framing follows Chernozhukov et al. (2024).

Estimating Heterogeneity in the PLR

Start with a structured approximation using $p(X)$

$$Y = g_0(X) + D\,p(X)'\beta_0 + \zeta, \qquad \mathbb{E}[\zeta \mid D, X] = 0$$

$$\hat{\beta} = \arg\min_{\beta} \mathbb{E}_n \left[\left(\hat{R}_Y - \hat{R}_D\, p(X)'\beta \right)^2 \right]$$

- One feasible answer to the infeasible PEHE problem is to restrict $\tau_0(x) \approx p(x)'\beta$ where $p(x)$ is specified basis for linear projection.
- Here $\hat{R}_Y = Y - \hat{\ell}(X)$ and $\hat{R}_D = D - \hat{m}(X)$ are cross-fitted residuals.
- If $p(X)$:
  - $= 1$, we get the constant-effect PLR
  - contains group indicators, we get GATE-style summaries
  - contains bins, splines, or selected interactions, we get an interpretable low-dimensional heterogeneity map.

Interpretation

If the interacted PLR is correctly specified, $p(X)'\beta_0 = \tau_0(X)$. Otherwise, it is an interpretable approximation.

Debiased low-dimensional heterogeneity summaries: Semenova and Chernozhukov (2021).

Relaxing $p(X)$: The R-Learner

A feasible surrogate objective for flexible CATEs

$$\hat{\tau}_R \in \arg\min_{\tau \in \mathcal{T}} \mathbb{E}_n \left[\left(\hat{R}_Y - \hat{R}_D \tau(X) \right)^2 \right] = \arg\min_{\tau \in \mathcal{T}} \mathbb{E}_n \left[\hat{R}_D^2 \left(\frac{\hat{R}_Y}{\hat{R}_D} - \tau(X) \right)^2 \right]$$

- Now the low-dimensional restriction $\tau(x) = p(x)'\beta$ is replaced by a flexible learner over $\tau \in \mathcal{T}$.
- The loss here amounts to a weighted regression problem with target $\hat{R}_Y / \hat{R}_D$ and weights $\hat{R}_D^2$ that can be used with any regression method / flexible learner.
- This is not PEHE itself, but an estimable surrogate loss whose population minimizer targets τ_0 under unconfoundedness and overlap.
- Valid statistical inference is much harder here than for low-dimensional ATE or GATE targets.

Quasi-oracle R-learner framing: Nie and Wager (2021).

Estimating Heterogeneity in the IRM

One target for both parametric and nonparametric CATE models

$$\widehat{\phi}_i = \hat{g}_1(X_i) - \hat{g}_0(X_i) + \frac{D_i\,(Y_i - \hat{g}_1(X_i))}{\hat{m}(X_i)} - \frac{(1 - D_i)\,(Y_i - \hat{g}_0(X_i))}{1 - \hat{m}(X_i)}$$

Parametric summary

$$\hat{\beta}_{DR} = \arg\min_{\beta} \mathbb{E}_n \left[(\hat{\phi} - p(X)'\beta)^2 \right]$$

Flexible CATE

$$\hat{\tau}_{DR} \in \arg\min_{\tau \in \mathcal{T}} \mathbb{E}_n \left[(\hat{\phi} - \tau(X))^2 \right]$$

- Recall the DR signal satisfies $\mathbb{E}[\phi(W) \mid X] = \tau_0(X)$, so we can regress it on either a structured $p(X)$ or a flexible learner directly.
- This is a clean bridge between interpretable BLP / GATE summaries and fully nonparametric CATE models.
- The same held-out DR signal will also power our validation step as it is an unbiased signal for the true CATE, albeit a noisy one, thereby providing observation-level validation.

DR signal CATE use: Semenova and Chernozhukov (2021).

Validating CATEs with Held-Out DR Signals

What we do instead of PEHE in practice

Split into train $\mathcal{T}$ and validation $\mathcal{V}$: fit candidate CATE model τ on $\mathcal{T}$; on $\mathcal{V}$ compute DR pseudo-outcomes $\hat{\phi}_i^{val}$ and model predictions $\hat{\tau}_i$.

DR-Loss & BLP Validation

Doubly-Robust Loss (DR-Loss):

$$\hat{\mathcal{L}}_{DR}(\hat{\tau}) = \mathbb{E}_{\mathcal{V}}\left[\left(\hat{\phi}^{val} - \hat{\tau}\right)^2\right]$$

- Direct feasible surrogate for PEHE on held-out data, but can mislead in policy settings.

Best Linear Predictor (BLP):

$$\hat{\phi}^{val} = \alpha_0 + \alpha_1 \hat{\tau} + u$$

- $\alpha_1 > 0$: ranked predictions align with held-out causal signals
- $\alpha_1 \approx 1$: scale calibration

TOC & RATE

Targeting Operator Characteristic (TOC): let S_q = top-q fraction ranked by $\hat{\tau}_i$.

$$\widehat{\mathrm{TOC}}(q) = \mathbb{E}_{\mathcal{V}}\left[\hat{\phi}^{val} \mid S_q\right] - \mathbb{E}_{\mathcal{V}}\left[\hat{\phi}^{val}\right] = \widehat{\mathrm{GATE}}_{S_q} - \widehat{\mathrm{ATE}}_{\mathcal{V}}$$

- Gain over random targeting at budget level q

Rank-Weighted ATE (RATE):

$$\widehat{\mathrm{RATE}} = \int_0^1 \alpha(q)\, \widehat{\mathrm{TOC}}(q)\, dq$$

- $\alpha(q)$ encodes policy/budget emphasis
- AUTOC ($\alpha(q) = 1$): uniform average over all budget levels; our main ranking diagnostic

BLP/GATE: Semenova and Chernozhukov (2021); RATE: Yadlowsky et al. (2025); Overview: Chernozhukov et al. (2024).

$$\pi(x) = 1\{\hat{\tau}(x) > c(x)\}$$

- Under linear welfare and a common cost scale, treat when estimated incremental net benefit exceeds zero or a budget shadow price.
- If treatment cost is roughly constant, ranking units by $\hat{\tau}(x)$ may be enough.
- Good policy learning usually depends on both effect estimation and honest out-of-sample evaluation.
- With budget, fairness, or capacity constraints, the optimal rule need not be a simple threshold.

Interpretation

The move from treatment effects to decisions is where the theory starts to matter operationally.

Productionalizing CATE Models

The system behind personalization

IN PRACTICE

Automating the CATE modeling pipeline

- MLOps for causal inference: data extraction, feature engineering, and model training pipelines that are robust to the extra complexity of CATE estimation.
- AutoML for candidate learners: Model selection & validation with DR signals, BLP, and TOC / RATE (and more) rather than plain predictive metrics.
- Serving model endpoints and monitoring: use in future personalized treatment decisions (batch and real-time), and monitor policy value over time.
- ⋆ Bias, fairness, and ethical personalization: avoidance of bias amplification, fairness constraints, and guardrails against unintended, learned unethical patterns.

Practical & theoretical challenges

- The true causal label is never observed, so validation and monitoring are non-trivial, indirect, and continue to be an active area of academic research.
- Translating business objective functions into standardized policy learning objectives is as much art as it is science.
- Distributed tooling, monitoring, and governance for production causal systems are still much less mature and lacking in many areas.
- ⋆ Ethical and bias concerns are amplified in personalized treatment settings, and there is still much work to be done on best practices and guardrails here.

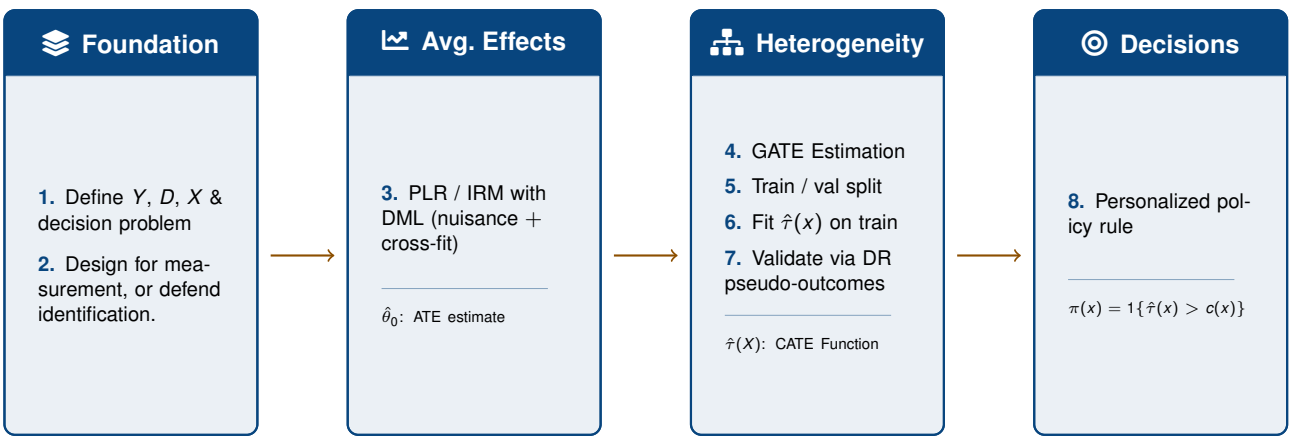
Operational lesson. In production, the goal is usually not a precise individual treatment effect; it is enough ranking and policy value out of sample to justify the extra system complexity.

Part III

Running Simulated Example and Takeaways

End-to-End Workflow in the Stylized Campaign

How the main pieces connect



Synthetic Example Setup

A simple DGP with nonlinear confounding

$$Y_i = g_0(X_i) + \theta_0 D_i + \varepsilon_i$$

$$D_i \sim \text{Bernoulli}(m_0(X_i))$$

$$\theta_0 = 1.2$$

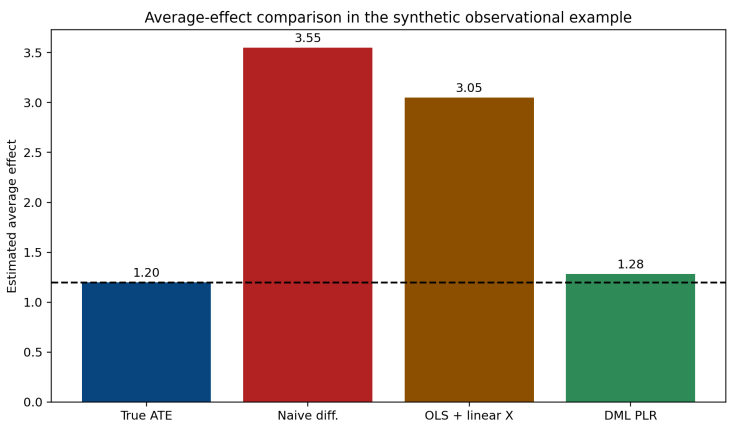
- In the running example, Y is post-treatment spend.
- D is whether the customer was exposed to the marketing campaign.
- X is pre-treatment customer information.
- Both $g_0(X)$ and $m_0(X)$ are deliberately nonlinear.
- That creates confounding that raw linear OLS cannot fully remove.

What this example is designed to show

If confounding runs through nonlinear functions of X, naive comparisons fail badly and linear OLS can still be biased, while PLR with flexible nuisance models can recover the average effect much better.

Average Treatment Effect and Confounding Bias

Classical OLS can struggle with arbitrary, nonlinear confounding



- Here, Y is post-treatment spend and D is marketing initiative exposure.
- Naive difference-in-means is biased because treatmentdepends on customer covariates X.
- OLS with raw controls still misses the nonlinear confounding structure.
- PLR, with Random Forests as nuisance functions, helps when confounding is richer than a hand-specified linear model.

Second Simulation: Heterogeneity-Focused DGP

Now let the treatment effect vary across customers

$$Y_i = g_0(X_i) + \tau_0(X_i)D_i + \varepsilon_i$$

$$D_i \sim \text{Bernoulli}(m_0(X_i))$$

$$\tau_0(X_i) = \tau_0(x_{i1}, x_{i2}, x_{i4})$$

$$\text{ATE} = \mathbb{E}[\tau_0(X)] \approx 0.53$$

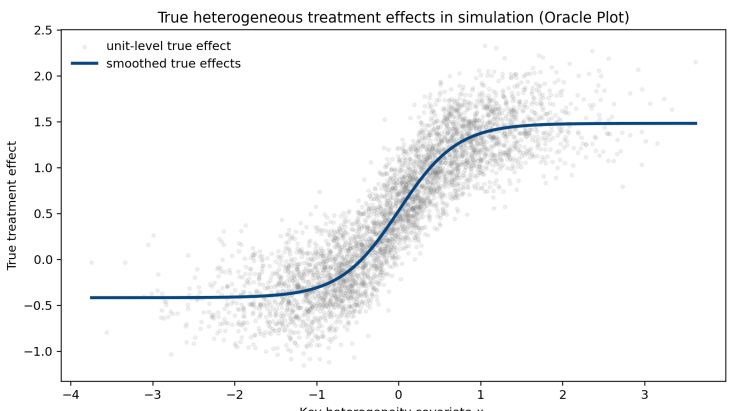
- Same running example: Y is spend, D is exposure, X is customer context.
- Now some customers benefit much more than others.
- This is the DGP used for the CATE, GATE, and policy visuals below.

Interpretation

This second toy example is not about showing OLS bias. It is about showing how heterogeneity can be real and consequential, and how we can use causal ML tools to understand it and make better decisions.

Heterogeneity Is Real

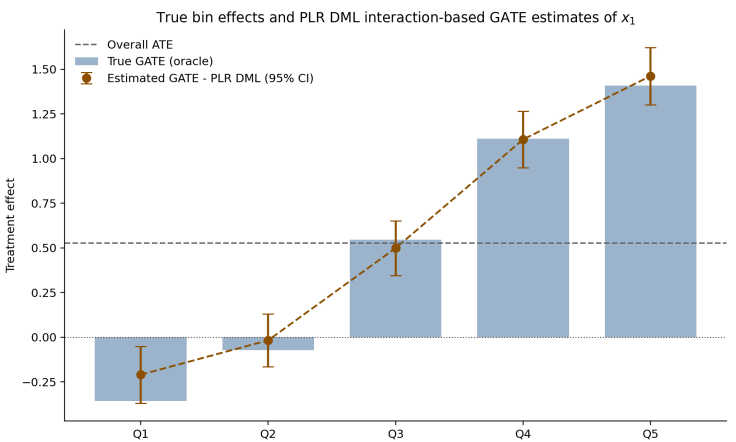
What the true CATE looks like in the DGP



- The horizontal axis is a key customer feature in X.
- The vertical axis is the true individual level treatment effect of exposure.
- Some customers clearly benefit more than others, and the relationship is nonlinear.

An Interpretable First Step

Segment-level heterogeneity before a black box



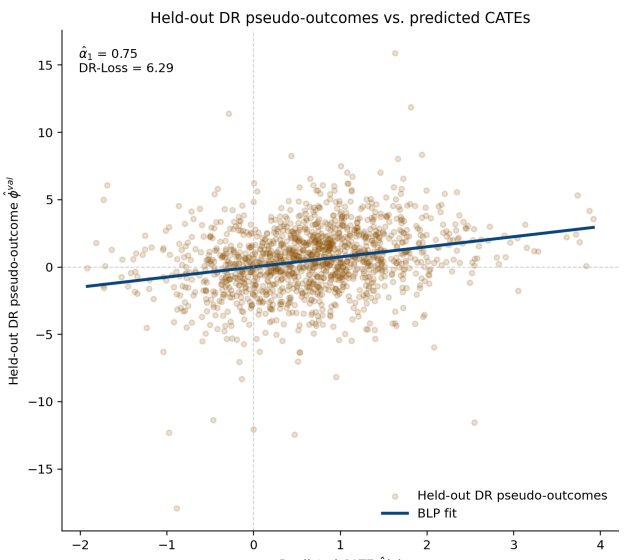
- The blue bars are the true bin-level effects in the simulation.
- The bronze points and whiskers are GATE estimates from PLR using linear projection of treatment interactions with basis $p(X)$.
- Only after this would I reach for a more flexible CATE learner.

Estimated CATE model

$$\hat{\beta} = \arg\min_{\beta} \mathbb{E}_n \left[\left(\hat{R}_Y - \hat{R}_D \, p(X)'\beta \right)^2 \right]$$

BLP WITH HELD-OUT DR SIGNALS

A CALIBRATION-STYLE CHECK USING NOISY CAUSAL PROXIES



- CATE model is estimated on the training fold via R-Learner with Random Forests for nusiance functions and final CATE learner.
- On the held-out fold, each point is a DR pseudo-outcome rather than a true individual effect label.
- The blue line is the BLP fit:

$$\hat{\phi}_i^{val} = \alpha_0 + \alpha_1 \hat{\tau}_i + u_i$$

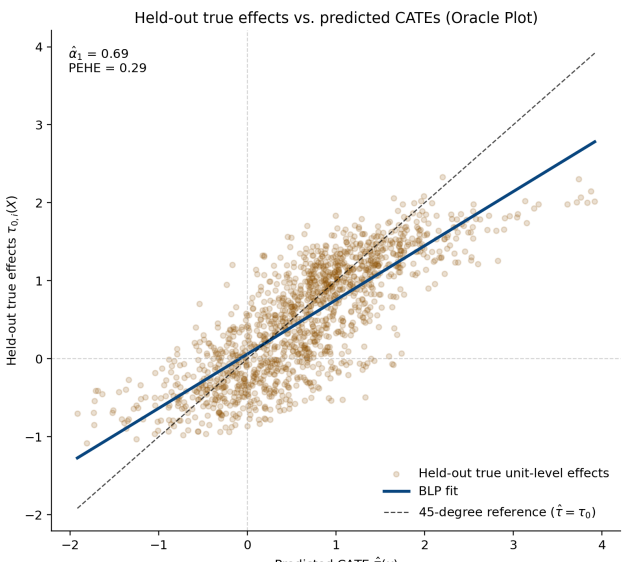
- The trend matters more than point-by-point fit because the pseudo-outcomes are noisy.

Estimated CATE model

$$\hat{\tau}_R \in \arg\min_{\tau \in \mathcal{T}} \mathbb{E}_n \left[\hat{R}_D^2 \left(\frac{\hat{R}_Y}{\hat{R}_D} - \tau(X) \right)^2 \right]$$

BLP WITH GROUND TRUTH

A CALIBRATION-STYLE ORACLE CHECK USING TRUE VALUES



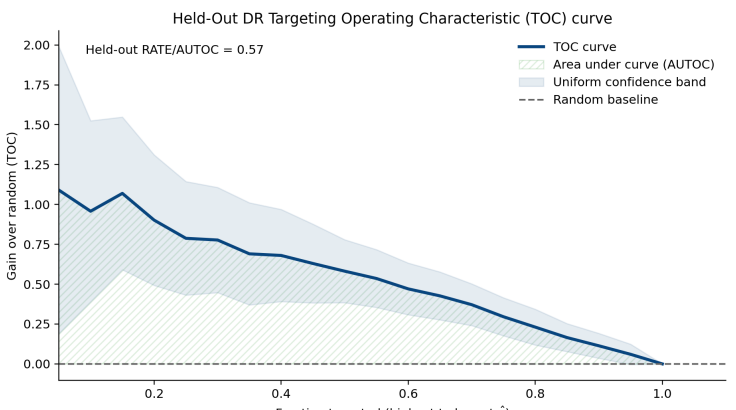
- Same BLP setup as before, but now each point is the true individual effect from the DGP rather than a DR pseudo-outcome.
- The blue line is the BLP fit:

$$\tau_{0,i} = \alpha_0 + \alpha_1 \hat{\tau}_i + u_i$$

- The Oracle metric $\text{PEHE} = \mathbb{E}_n[(\hat{\tau} - \tau_0)^2]$ is very small at 0.29
- Note, the Y-axis scale is significantly different here because the true effects are noiseless, so the point cloud is much tighter.
- This shows how noisy the DR pseudo-outcomes are as proxies for true effects.

TOC WITH HELD-OUT DR SIGNALS

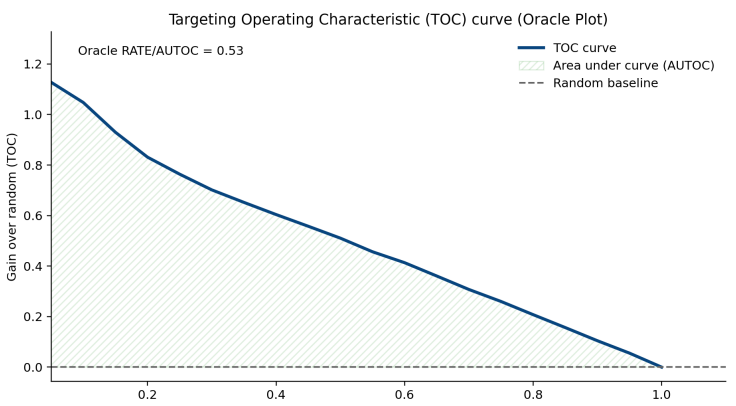
A RANKING, GAIN-OVER-RANDOM CHECK



- Before prescribing a policy, we want evidence that the ranking by $\hat{\tau}(x)$ is useful out of sample.
- The blue curve is a held-out gain-over-random / TOC curve.
- The area under the curve is the RATE metric (AUTOC in this case).
- The gray horizontal line is the random benchmark; moving above it means the ranking is capturing useful heterogeneity.

TOC WITH GROUND TRUTH

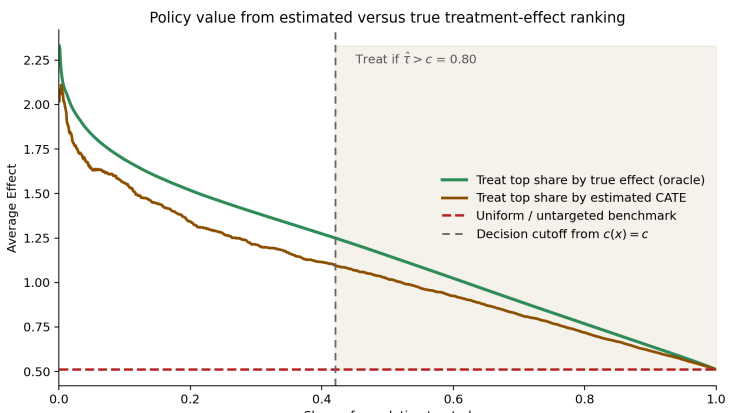
AN ORACLE RANKING, GAIN-OVER-RANDOM CHECK



- Same TOC setup as before, but now the vertical axis is the true average effect of the estimated CATE ranking rather than a DR pseudo-outcome.
- Note the Oracle ranking aligns with the DR ranking much more than the pointwise BLP fits did.
- This check is thus more directly relevant to decision quality than the BLP because it focuses on the ranking rather than pointwise fit.

From Effects to Decisions

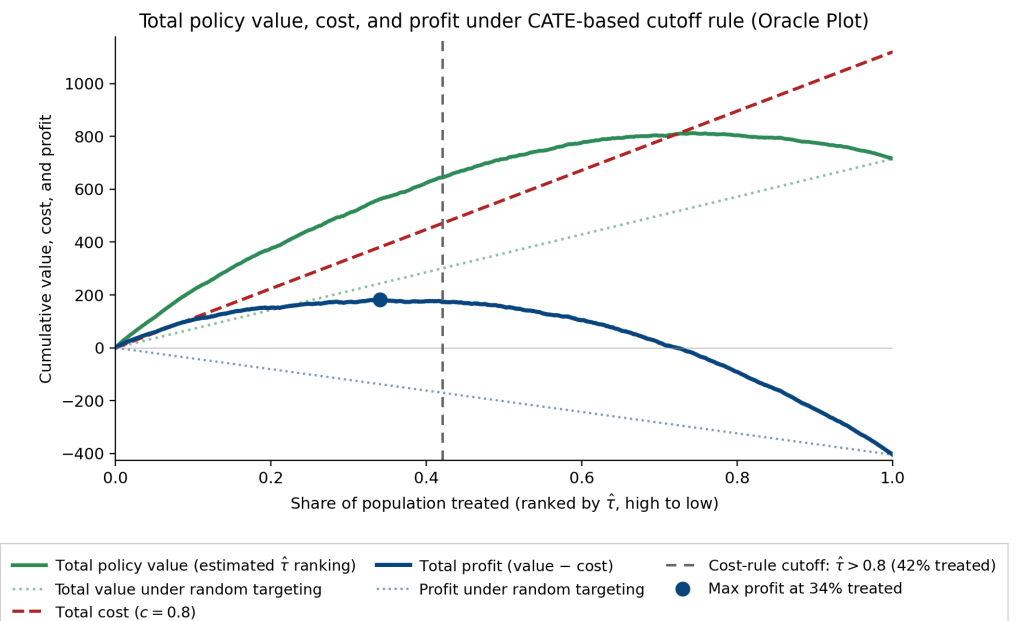
A constant-cost decision rule



- The green line is the oracle policy that ranks customers by the true effect.
- The bronze line ranks customers by the estimated CATE $\hat{\tau}(X)$.
- The dashed vertical line shows the cutoff implied by a constant treatment cost $c(x) = c = 0.80$.
- If treatment is costly, even an imperfect heterogeneity model can improve value relative to an untargeted benchmark.

Policy Ranking Matters

It can be the difference between profitability and loss



Limits and Open Problems

What causal ML still does not solve

- Identification assumptions still do the heavy lifting.
- Poor overlap and measurement problems can dominate the analysis.
- Flexible CATE estimation is easier to overfit than standard prediction.
- Validation, model selection, and inference for rich heterogeneity models remain active research areas.
- Distributed tooling, monitoring, and governance for production causal systems are still much less mature than in standard predictive ML.

Honest framing

Causal ML is powerful, but it is not magic; it works best when theory, design, and empirical discipline all line up.

Four Key Takeaways

If you take nothing else from this talk, remember these four points

1. Causal ML is about combining identification logic and semi-parametric specifications with flexible learning, not replacing one with the other; in practice, the highest-leverage move is often upstream design.
2. Average effects are only part of the story; heterogeneity matters when decisions are personalized and is often of more policy relevance to the business.
3. Validation is fundamentally harder than in standard ML because individual treatment effects are never directly observed, which is one reason production CATE systems are genuinely hard.
4. These techniques are powerful but not magic; they can help us learn more from data and make better decisions, but they do not obviate the need for good theory, design, and empirical discipline.

QUESTIONS?

Jacob Pieniazek

Lead Data Scientist, 84.51°

jacob-pieniazek.com

jacob@pieniazek.me

github.com/jakepenzak

in/japieniazek

jacob-pieniazek.com

Presentation reflects personal views and not those of my employer.

Part IV

APPENDIX

Neyman Orthogonality

A key property for robust estimation

- A moment function $M(\theta_0, \eta_0) = \mathbb{E}\left[\psi(W; \theta_0, \eta_0)\right]$ is Neyman orthogonal at (θ_0, η_0) if the Gateaux derivative of the mapping $\eta \mapsto \mathbb{E}[\psi(W; \theta_0, \eta)]$ at η_0 is zero.
- Intuitively, this means that small errors in estimating the nuisance functions η have only a second-order effect on the bias of the estimator for θ_0.
- This property is crucial for achieving root-n consistency and asymptotic normality when using flexible machine learning methods to estimate nuisance functions.

For subsequent proofs, identification of θ is straightforward by construction of the score and we proceed to show Neyman orthogonality by calculating the Gateaux derivative of the moment function at the true parameter and nuisance values in the direction of arbitrary perturbations h. $\overset{\text{LIE}}{=}$ denotes the use of the law of iterated expectations to simplify the expression by conditioning on X. Please excuse any notational abuse for the sake of clarity and brevity in the calculations.

For more details, see Chernozhukov et al. (2018, Chap. 3).

Proof: Neyman Orthogonality in the PLR

Gateaux derivative at the truth

Proof.

$$\psi(W;\theta,\eta) = (D - m(X))\big(Y - \ell(X) - \theta(D - m(X))\big), \qquad \eta = (\ell, m)$$

$$\ell_0(X) = \mathbb{E}[Y \mid X], \qquad m_0(X) = \mathbb{E}[D \mid X]$$

$$\ell_r(X) = \ell_0(X) + r h_\ell(X), \qquad m_r(X) = m_0(X) + r h_m(X), \qquad \eta_r = (\ell_r, m_r)$$

$$\begin{aligned}
\partial_r \mathbb{E}[\psi(W;\theta_0,\eta_r)]|_{r=0} &= \partial_r \mathbb{E}\left[(D - m_0(X) - r h_m(X))\left(Y - \ell_0(X) - r h_\ell(X) - \theta_0\left(D - m_0(X) - r h_m(X)\right)\right)\right]|_{r=0} \\
&= \partial_r \mathbb{E}\left[(D - m_0(X) - r h_m(X))\left(Y - \ell_0(X) - r h_\ell(X)\right) - \theta_0\left(D - m_0(X) - r h_m(X)\right)^2\right]\Big|_{r=0} \\
&= -\mathbb{E}\left[h_\ell(X)\left(D - m_0(X)\right)\right] - \mathbb{E}\left[h_m(X)\left(Y - \ell_0(X)\right)\right] - \mathbb{E}\left[2\theta_0 h_m(X)\left(D - m_0(X)\right)\right] \\
&\overset{\text{LIE}}{=} -h_\ell(X)\left(\mathbb{E}\left[D \mid X\right] - m_0(X)\right) - h_m(X)\left(\mathbb{E}\left[Y \mid X\right] - \ell_0(X)\right) \\
&\quad - 2\theta_0 h_m(X)\left(\mathbb{E}\left[D \mid X\right] - m_0(X)\right) \\
&= 0.
\end{aligned}$$

since $\mathbb{E}[D \mid X] = m_0(X)$ and $\mathbb{E}[Y \mid X] = \ell_0(X)$. □

Illustrative Neyman orthogonality proof adapted from Chernozhukov et al. (2024, Chap. 9, Sec. 9.B).

Proof: Neyman Orthogonality Fails in the Naive IRM Case

The simple plugin estimator is **not** orthogonal

Proof.

$$\psi(W;\theta,\eta) = g(1,X) - g(0,X) - \theta, \qquad \eta = (g(0,\cdot), g(1,\cdot))$$

$$g_0(1,X) = \mathbb{E}[Y \mid D=1, X], \qquad g_0(0,X) = \mathbb{E}[Y \mid D=0, X]$$

$$g_{1,r}(X) = g_0(1,X) + rh_1(X), \qquad g_{0,r}(X) = g_0(0,X) + rh_0(X), \qquad \eta_r = (g_{1,r}, g_{0,r})$$

$$\begin{aligned}\partial_r \mathbb{E}[\psi(W;\theta_0,\eta_r)]|_{r=0} &= \partial_r \mathbb{E}\left[g_0(1,X) + rh_1(X) - g_0(0,X) - rh_0(X) - \theta\right]|_{r=0} \\ &= \mathbb{E}\left[h_1(X) - h_0(X)\right] \neq 0.\end{aligned}$$

□

Proof: Neyman Orthogonality in the Doubly Robust IRM

Gateaux derivative at the truth

Proof.

$$\psi(W;\theta,\eta) = g(1,X) - g(0,X) + \frac{D\{Y - g(1,X)\}}{m(X)} - \frac{(1-D)\{Y - g(0,X)\}}{1-m(X)} - \theta, \qquad \eta = (g(0,\cdot), g(1,\cdot), m), \qquad 0 < m(X) < 1$$

$$g_0(1,X) = \mathbb{E}[Y \mid D=1, X], \qquad g_0(0,X) = \mathbb{E}[Y \mid D=0, X], \qquad m_0(X) = \mathbb{E}[D \mid X]$$

$$g_{1,r}(X) = g_0(1,X) + r h_1(X), \qquad g_{0,r}(X) = g_0(0,X) + r h_0(X), \qquad m_r(X) = m_0(X) + r h_m(X), \qquad \eta_r = (g_{1,r}, g_{0,r}, m_r)$$

$$\begin{aligned}
\partial_r \mathbb{E}[\psi(W;\theta_0,\eta_r)]|_{r=0} &= \partial_r \mathbb{E}\left[g_0(1,X) + r h_1(X) - g_0(0,X) - r h_0(X) + \frac{D\{Y - g_0(1,X) - r h_1(X)\}}{m_0(X) + r h_m(X)} - \frac{(1-D)\{Y - g_0(0,X) - r h_0(X)\}}{1 - m_0(X) - r h_m(X)} - \theta_0 \right]_{r=0} \\
&= \mathbb{E}\left[h_1(X) - h_0(X)\right] + \mathbb{E}\left[\frac{D\{-h_1(X)\}}{m_0(X)} - \frac{(1-D)\{-h_0(X)\}}{1-m_0(X)}\right] + \mathbb{E}\left[\frac{D\{Y - g_0(1,X)\}h_m(X)}{m_0(X)^2} + \frac{(1-D)\{Y - g_0(0,X)\}h_m(X)}{(1-m_0(X))^2}\right] \\
&\overset{\text{LIE}}{=} h_1(X) - h_0(X) + \frac{\mathbb{E}[D|X]\{-h_1(X)\}}{m_0(X)} - \frac{(1 - \mathbb{E}[D|X])\{-h_0(X)\}}{1-m_0(X)} + \frac{\mathbb{E}[D\{Y - g_0(1,X)\}|X]h_m(X)}{m_0(X)^2} + \frac{\mathbb{E}[(1-D)\{Y - g_0(0,X)\}|X]h_m(X)}{(1-m_0(X))^2} \\
&= h_1(X) - h_0(X) - h_1(X) + h_0(X) + 0 + 0 \\
&= 0.
\end{aligned}$$

since $\mathbb{E}[D \mid X] = m_0(X)$ and $\mathbb{E}[Y - g_0(d,X) \mid D = d, X] = 0$ for $d \in \{0,1\}$. □

Illustrative orthogonality calculation adapted from Chernozhukov et al. (2024, Chap. 9, Sec. 9.B).

Synthetic Example 1: Bias DGP

Used for the naive vs OLS vs PLR comparison

$$X_1, X_2, X_5 \sim \mathcal{N}(0,1), \qquad X_3 \sim \mathrm{Unif}[-2,2], \qquad X_4 \sim \mathrm{Bernoulli}(0.4)$$

$$c(X) = (X_1^2 - 1) + 0.8(X_2^2 - 1) + 0.6\sin(1.3X_3) + 0.4X_1X_2$$

$$g_0(X) = 1 + 0.8X_1 - 0.5X_2 + 0.5X_3 + 0.25X_4 + c(X)$$

$$m_0(X) = \Lambda\left(-0.2 + 0.4X_1 + 0.3X_3 + 0.15X_4 + 0.9c(X)\right)$$

$$D \sim \mathrm{Bernoulli}(m_0(X)), \qquad \theta_0 = 1.2, \qquad Y = g_0(X) + \theta_0 D + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0,1), \qquad m_0(X) \in [0.05, 0.95] \text{ after clipping}$$

Why this DGP is useful pedagogically

Both treatment assignment and outcomes depend on the same nonlinear confounding score $c(X)$. A linear OLS adjustment with only raw controls misses that curvature, while flexible nuisance learners in PLR can learn it.

Synthetic Example 2: Heterogeneity DGP

Used for CATE, GATE, and policy visuals

$$X_1, X_2, X_5 \sim \mathcal{N}(0,1), \qquad X_3 \sim \mathrm{Unif}[-2,2], \qquad X_4 \sim \mathrm{Bernoulli}(0.4)$$

$$g_0(X) = 1 + 0.7X_1 - 0.45X_2 + 0.35X_3 + 0.25X_4 + 0.25X_1X_5 + 0.35\sin(1.2X_2)$$

$$m_0(X) = \Lambda\left(-0.1 + 0.55X_1 - 0.35X_2 + 0.25X_3 + 0.3X_4 - 0.2X_1X_2\right)$$

$$\tau_0(X) = 0.4 + 0.95\tanh(1.4X_1) + 0.35X_4 - 0.25X_2$$

$$D \sim \mathrm{Bernoulli}(m_0(X)), \qquad Y = g_0(X) + \tau_0(X)D + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1)$$

Why this DGP is useful pedagogically

It creates meaningful treatment-effect heterogeneity with reasonable overlap, so the CATE curve, GATE summaries, and policy-value plot are visually informative and easy to interpret.

Meta-Learners at a Glance

A Non-Comprehensive Summary of Additional CATE Estimators

Method	Strength	Watch-out
S-learner	Simple and stable baseline	Can regularize away small treatment effects
T-learner	Flexible treated / control modeling	Can be noisy when one arm is small
X-learner	Often strong with imbalance	Still leans heavily on outcome modeling
DR-learner	Uses doubly robust pseudo-outcomes	Can become unstable with extreme propensities
R-learner	Residualization logic is elegant and practical	Targets overlap-weighted objects under misspecification

References: Künzel et al. (2019); Nie and Wager (2021); Chernozhukov et al. (2024).

References I

- Callaway, B., & SantAnna, P. H. (2021). **Difference-in-differences with multiple time periods [Themed Issue: Treatment Effect 1].** *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.12.001
- Chang, N.-C. (2020). **Double/debiased machine learning for difference-in-differences models.** *The Econometrics Journal*, *23*(2), 177–191. https://doi.org/10.1093/ectj/utaa001
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). **Double/debiased machine learning for treatment and structural parameters.** *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097
- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). **Applied causal inference powered by machine learning and ai [Available at https://www.causalml-book.org].**
- Clarke, P. S., & Polselli, A. (2026). **Double machine learning for static panel models with fixed effects.** *The Econometrics Journal*, *29*(1), 69–86. https://doi.org/10.1093/ectj/utaf011
- Künzel, S., Sekhon, J., Bickel, P., & Yu, B. (2019). **Metalearners for estimating heterogeneous treatment effects using machine learning.** *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

References II

- Liu, M., Zhang, Y., & Zhou, D. (2021). **Double/debiased machine learning for logistic partially linear model.** *The Econometrics Journal*, *24*(3), 559–588. https://doi.org/10.1093/ectj/utab019
- Nie, X., & Wager, S. (2021). **Quasi-oracle estimation of heterogeneous treatment effects.** *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). **Estimation of regression coefficients when some regressors are not always observed.** *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818
- Robinson, P. M. (1988). **Root-*N*-consistent semiparametric regression.** *Econometrica*, *56*(4), 931–954. https://doi.org/10.2307/1912705
- Rubin, D. B. (1974). **Estimating causal effects of treatments in randomized and nonrandomized studies.** *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
- Semenova, V., & Chernozhukov, V. (2021). **Debiased machine learning of conditional average treatment effects and other causal functions.** *The Econometrics Journal*, *24*(2), 264–289. https://doi.org/10.1093/ectj/utaa027
- Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2025). **Evaluating treatment prioritization rules via rank-weighted average treatment effects.** *Journal of the American Statistical Association*, *120*(549), 38–51. https://doi.org/10.1080/01621459.2024.2393466