ELICSIR FOUNDATION

What if

From Counterfactual Worlds to Semantic Kernels

Jim Kazanjian, untitled

Marvellini, Goldrake Cilindro

“We do not know a truth without knowing its **cause**.”
Aristotle

Student: Cieri Manuel

Mentor: Prof. De Prisco Roberto

What we'll see

Sections:

1. Causal Inference Framework
2. Estimating CATE: Propensity Score and ML
3. Role of LLMs

1. Causal Inference Framework

What is Causal Inference?

Causal Inference:

Methodological Approach to define **cause-effect relations** among variables

We'll use as statistics **foundation** and **terminology:**

Rosenbaum and Rubin Framework (1983)

Causes of nature have been always **investigated** throughout **history**

Where is It Applied?

Medicine
(Treatment Decisions)

Economics

C.I.

Public **Policy**

Recommender
Systems /
Personalization

Causal inference asks: *what would happen if we acted differently?*

Treatment Effect

Machine learning predicts *what is likely.*
Causal inference predicts *what would happen if*.

- Beyond correlation
- Central to **decision-making**

Given an action/intervention that we call “**treatment”,** what is its effect on a defined **outcome**?

Effect of “Treatment” T $= Y^1 - Y^0$

T : **Intervention** or Treatment
Y^0 : Outcome if X in **Control** Group
Y^1 : Outcome if X in **Treatment** Group

Predicting outcomes under **hypothetical** actions

The Missing World

- Each unit has **two potential outcomes**
- We **observe only one**
- The **other** is **counterfactual**

Observed Outcome

$$Y_i^{obs} = T_i \cdot Y_i^{(1)} + (1 - T_i) \cdot Y_i^{(0)}$$

$$T_i \in \{0, 1\}$$

Individual Treatment Effect (ITE)

$$\tau(x) = Y^1 - Y^0$$

A Simple Example

Goal: *“Elevating Computer Science in Italy”* - **Idea**: Help CS students.
How? We need to define a specific treatment.

Let’s consider **Bruno Jr:**

Treatment:

Partecipating in “Scuola Ortogonale”

Y^1

Brescia Award

for Scientific Research

For each student we want to know if **policy** is **good or not**

Outcome:

CS Achievements in X years

We can only observe one!

Control:

Not Partecipating in “Scuola Ortogonale”

Y^0

WebDev **lost job** to Claude,
got rejected from art school also

The Fundamental Problem

- **Potential** outcomes: Y^0, Y^1
- **Observed** outcome: $Y = Y^T$
- Fundamental problem: ***we never observe both***

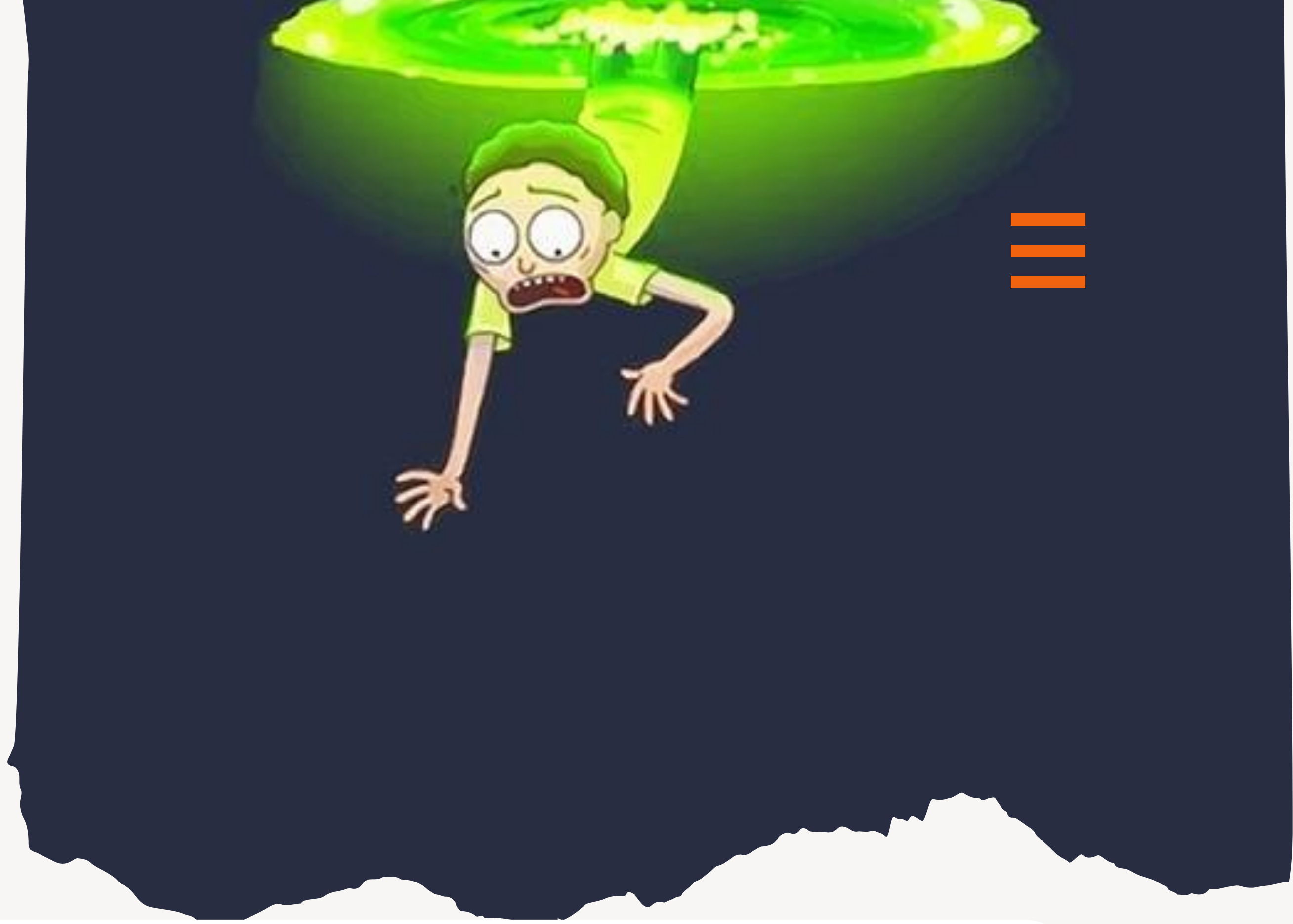
Note!

We'll focus on **Causal Inference in Medicine**

Solutions?

1. We could observe a **parallel counterfactual universe**...

Why Average Effects Are Not Enough



1. ...or we can **estimate** it through approximate metrics

Average Treatment Effect

$$ATE = \mathbb{E}\left[Y^1 - Y^0\right]$$

But:

- Decisions are individual
- Treatment effects are heterogeneous
- **Policy requires personalization**

How Find Counterfactual

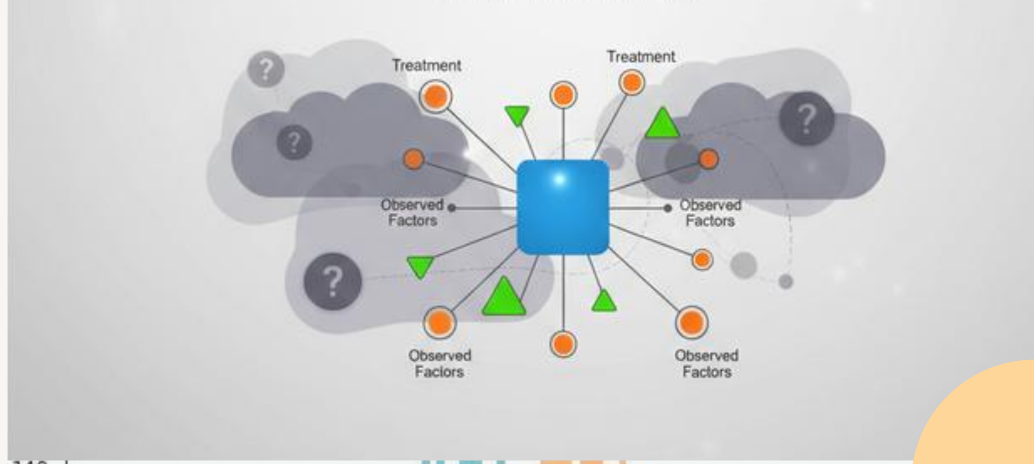
When Data is good: we can identify a "***counterfactual twin***" in the other group with equal/most similar covariates

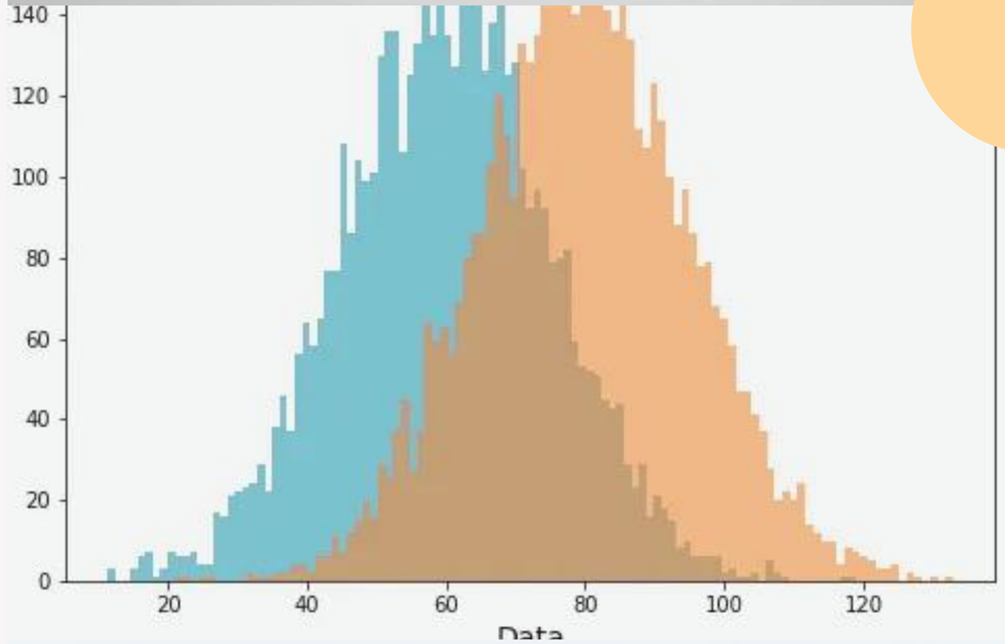
Conditional Average Treatment Effect

$$CATE = \mathbb{E}\left[Y^1 - Y^0 \mid X = x\right]$$

- **Personalization** requires $\tau(x)$
- Harder: **needs stronger assumptions**

When Is CATE Identifiable?



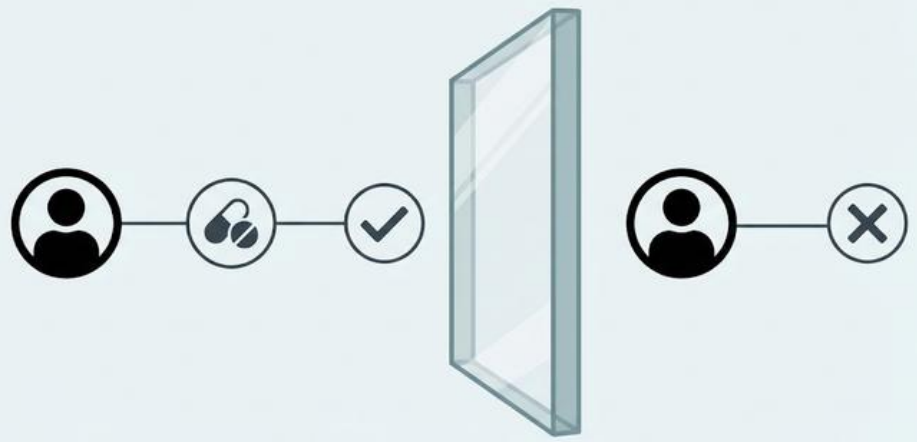


Key Assumptions

- **Stable Unit Treatment Value Assumption:**
  - **Consistency** → observed = realized potential outcome
  - **No Interference:** T=t on subject x doesn’t affect subject y
- **Ignorability** → no hidden confounders
- **Positivity** (overlap) → both treatments possible

$$(Y^0, Y^1) \perp T \mid X$$

$$0 < P(T = 1 \mid X) < 1$$



Stress Testing CATE in Observational ICU Data

Gold Standard:
Randomized Controlled Trials

Randomness guarantees similar distributions

RCTs

Verified **Positivity** and **Exchangeability**

Often Available

Real-world **observational** dataset

Most comparisons failed due to **imbalance**

Observationals

Selection Bias:
Treatment and Outcome both condition Inclusion in test

(T → I ← Y)

Confounding Bias:
external variable common cause of Treatment and Outcome

(T ← X → Y)

We used: **MIMIC-IV** Dataset

In practice, estimates often collapse.

MIMIC-IV

Large Observational Dataset

+200k Hospitalized in Emergency Dpt

+65k Unique ICU Stays

About 30 different tables with:

- **2M+ Chart** Events
- **2M+ Lab** Events
- Patient **Diagnoses**
- **Procedures** and **Pharmacy**
- Outcome and **Output Events** (i.e. in-hosp mortality)

Note: our results are presented when you see this icon

Our Work

1. **Identify subsets** with different bias scale ✓

2. **CATE estimation** with known tools ✓

3. **Bias Tool** Add a quantifiable bias both absolute and "relative"

4. **Use LLMs** for direct **CATE inference**

Lab usi athics

Structural Confounding

Vasopressors in ICU

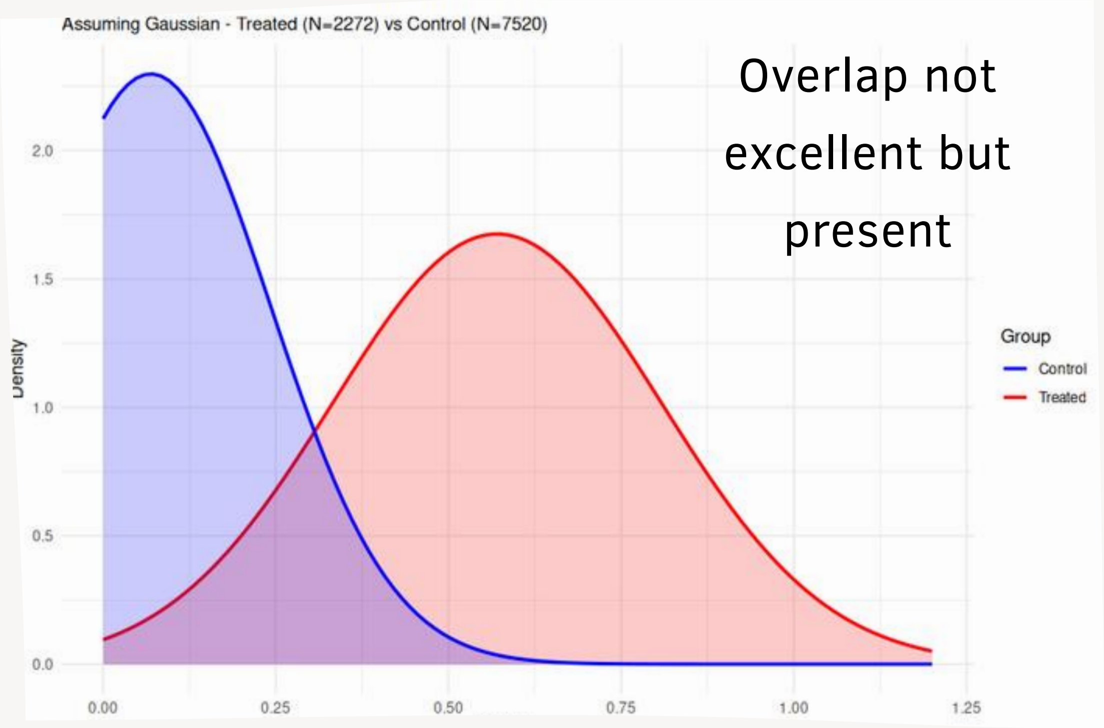
- **Y** = In-hospital Mortality
- **IPTW** ≈ +0.064
- **AIPW** ≈ +0.058
- **ATO** ≈ +0.047

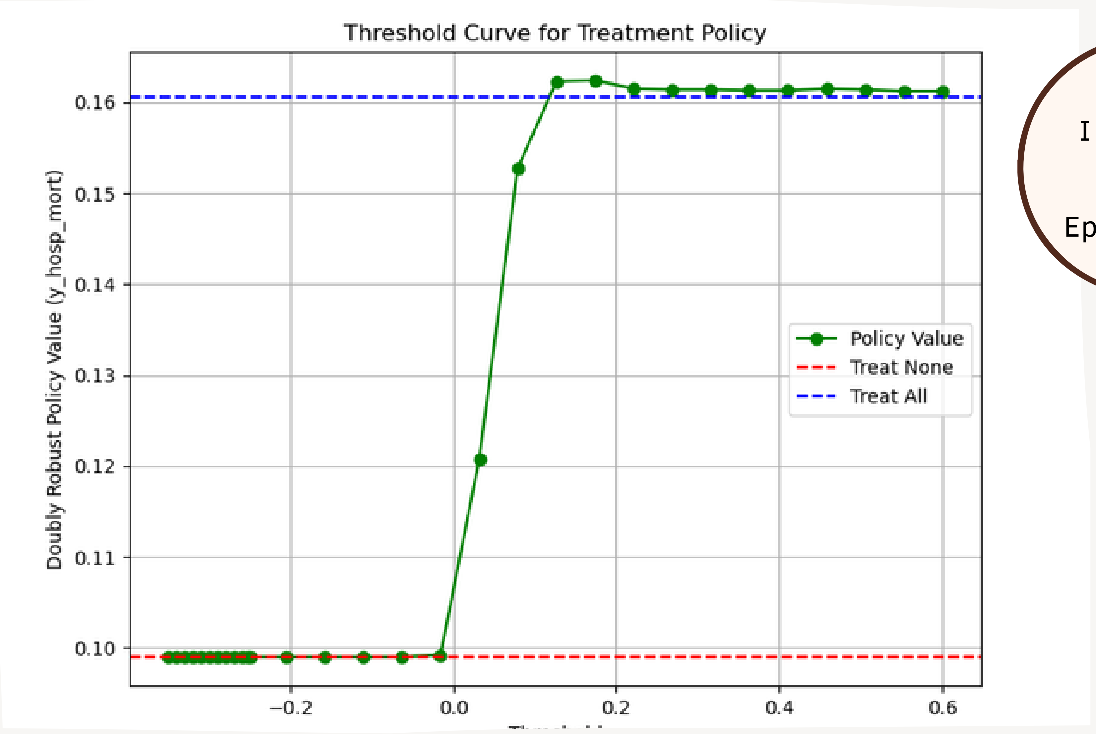
Confounding by Indication

- Familiarizing with MIMIC-IV dataset
- **Treatment appears harmful** even after DR estimation
- **Severity** drives both **treatment** and **outcome**

CATE worst outcome than “treat none” policy:
Vasopressors worsen outcome?!

When X does not capture severity, ignorability fails.





2. Estimating CATE: Propensity Score and ML

Estimating the Missing World

All these methods rely implicitly on **balancing distributions**

Inverse Probability Weighting

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

ML models

Meta-learners

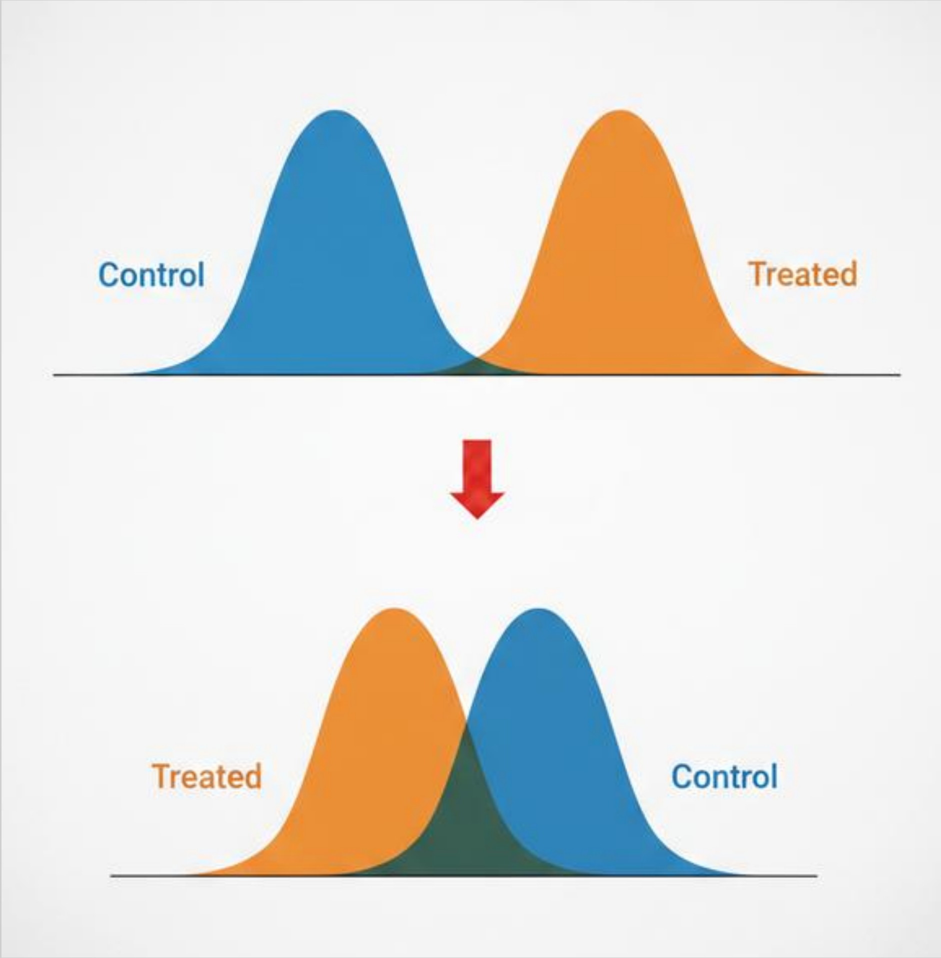
Doubly Robust estimators

We'll focus on the tools and bounds as demonstrated in **Shalit et al.** (2017)

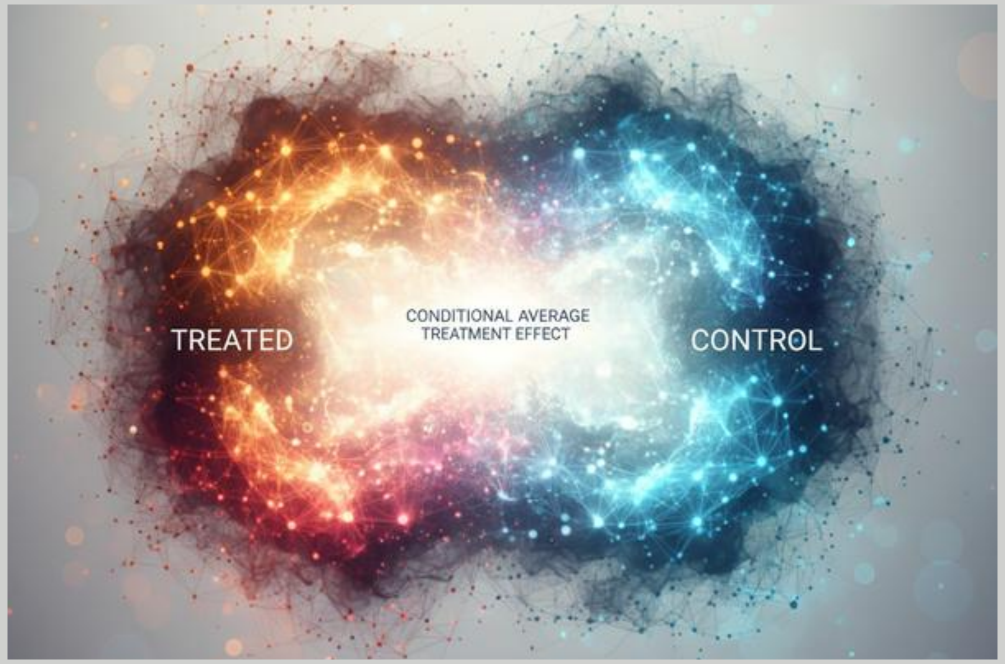
The Real Difficulty: Distribution Shift

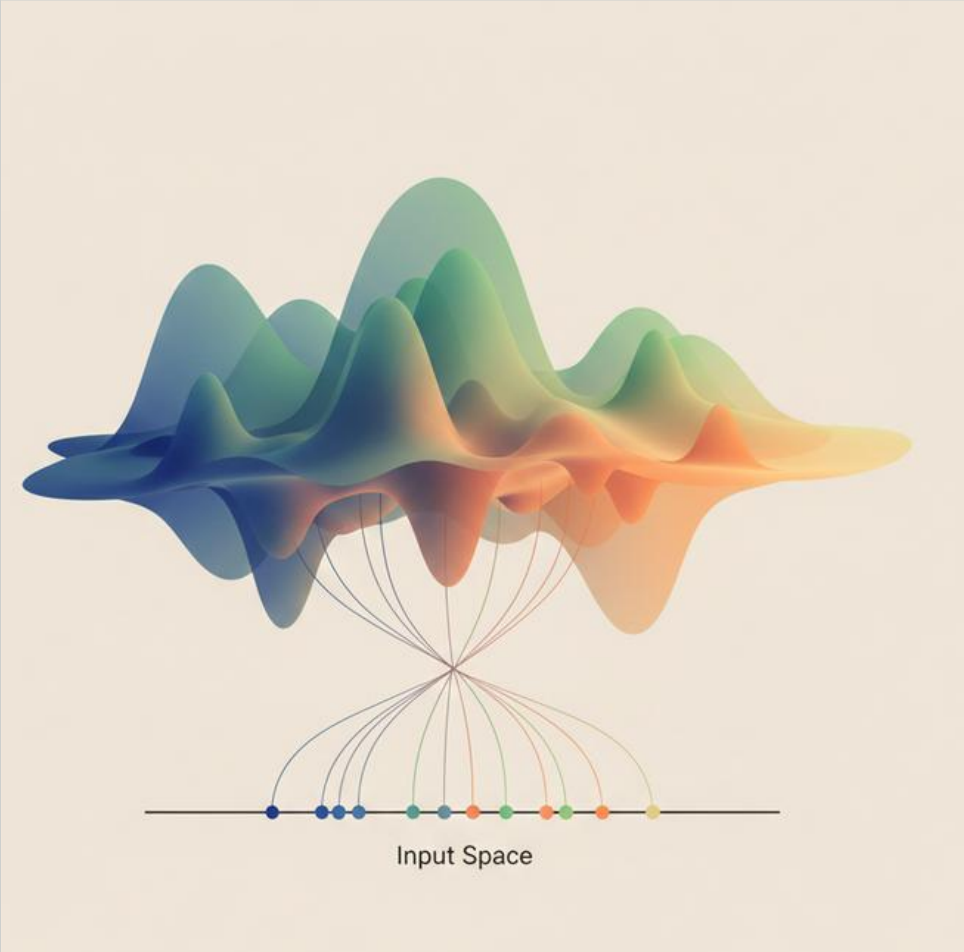
Treated and **control** **differ** in covariates

- **Counterfactual = extrapolation**
- **Poor overlap** → unstable CATE



CATE estimation is fundamentally a **domain adaptation problem.**





Propensity Score as Compression

$$e(X) = P(T = 1 \mid X)$$

Balancing Score

Conditional to b(x), covariate distribution X is **independent** from treatment assignment T:

$$X \perp T \mid b(X)$$

Propensity Score

- Rosenbaum & Rubin (1983)
- **Probability of Treatment given covariates X**
- Weighting and **Stratification** over *e(X)* make observational data **RCT-like**

Compressing Imbalance

- Balancing property
- **Reduces dimensionality**
- Does not solve complex geometry
- **Sufficient** for cofounding adjustment

Propensity score **reduces imbalance to one dimension**, but geometry may still be high-dimensional.

Designing for Overlap

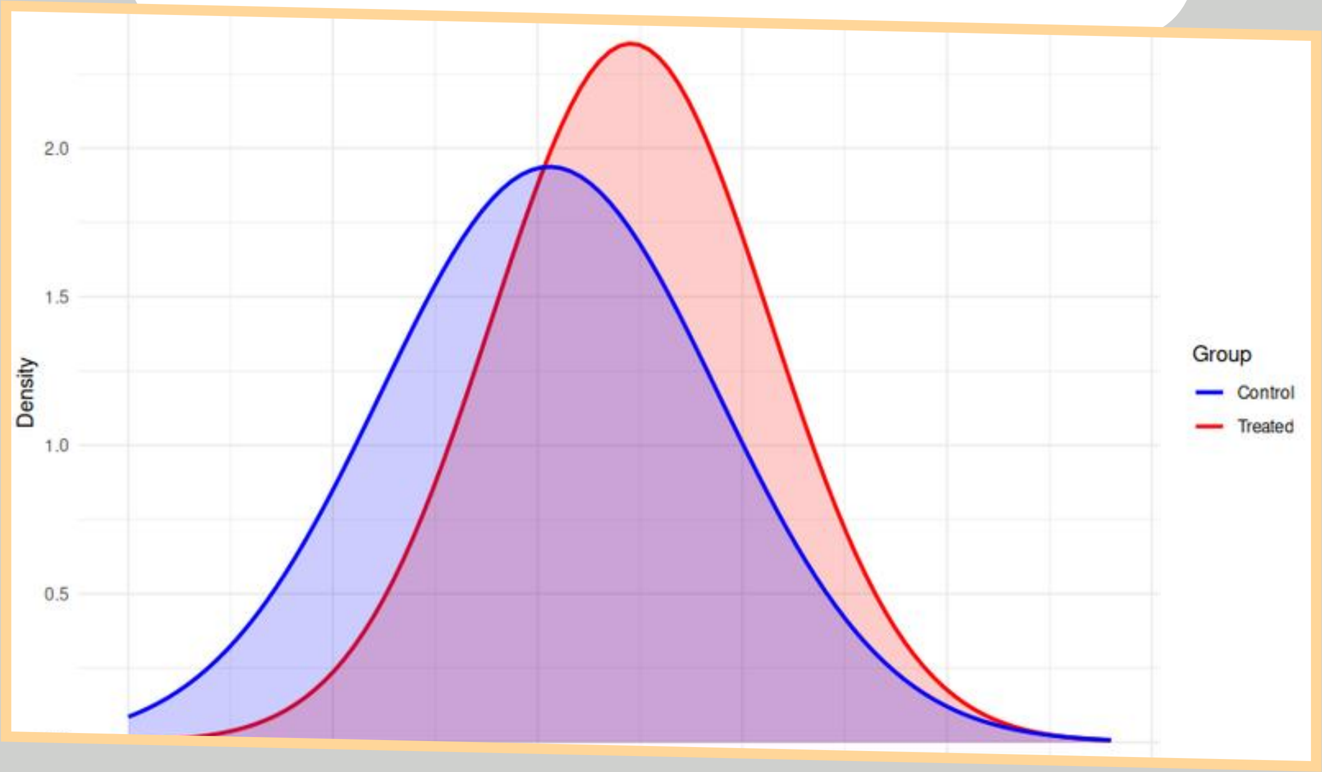
A (Slightly) **Better Design**: Early Diuretics in Sepsis

- N = 6,216 ICU patients
- Landmark design (t_0 = sepsis + 6h)
- Excluded active shock on admission
- Treatment = **early vs delayed diuretics**

CATE is **estimable** but still **inconsistent**!!

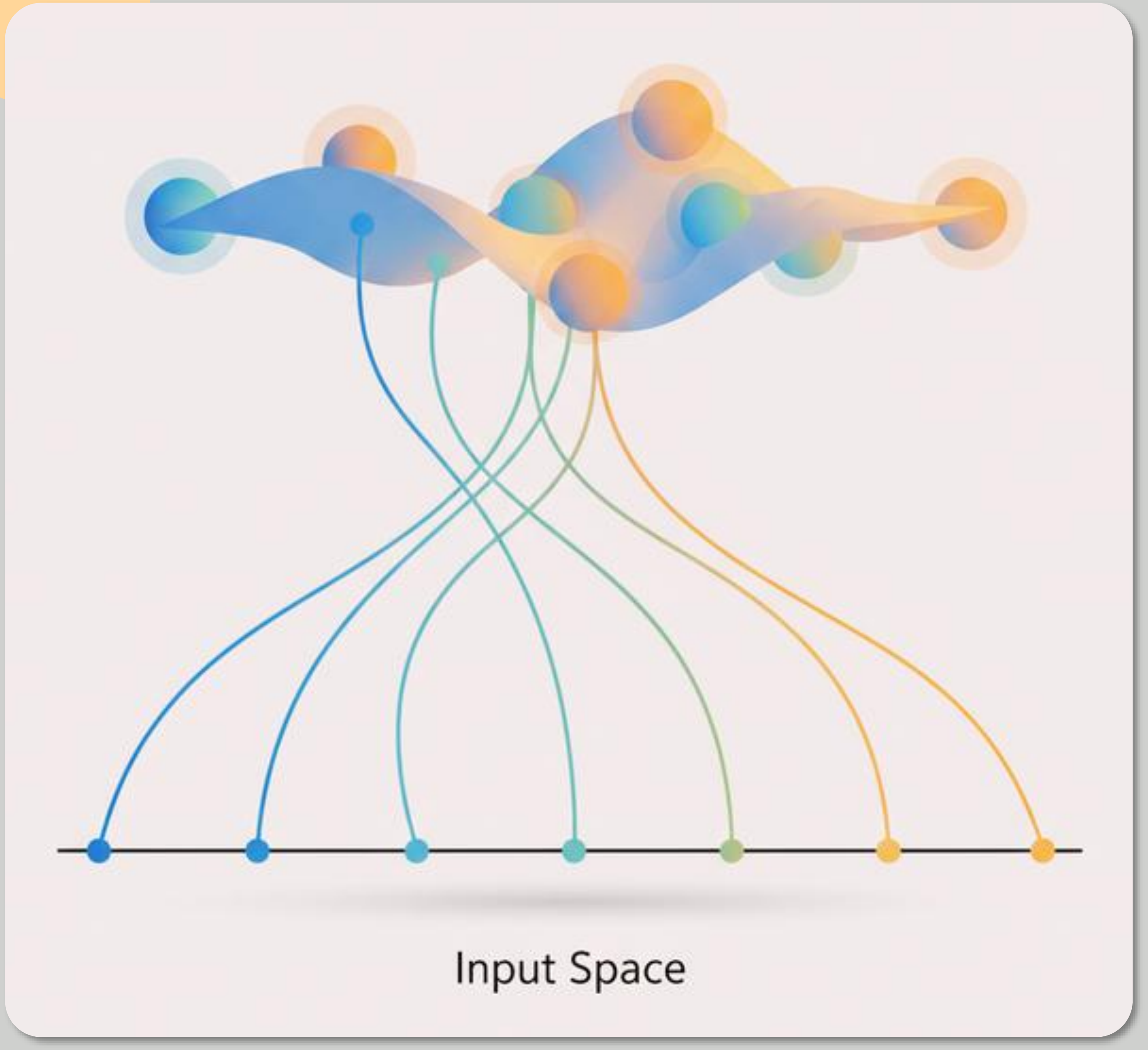
Overlap Enables Identification

- No dominant strata
- **Balanced covariates**
- **Bias still present!**



Careful temporal design reduces confounding,
but sometimes not enough.

Beyond Means: Comparing Distributions



- Means are not enough
- **PS** for balance but **ignores covariates** for **outcome** estimates
- Need **distribution-level comparison**
- Introduce **similarity function**

$$k(x, x') := \langle K_x, K_u \rangle_{\mathcal{H}}$$

A Kernel defines here a notion of **local similarity**

If we define similarity, we can compare groups structurally.

Kernel-Based Balancing in Practice

Overlap is not absolute: it depends on how we encode the data.

Kernels in Representation Learning

$$k(x, x') = exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right)$$

- **Kernel defines similarity** in feature space
- Common choices: Radial Basis Function (RBF), linear, polynomial,...
- **Geometry** is **fixed** by chosen kernel

Classical kernels impose a predefined geometry.

Why Balancing Improves CATE

$$\varepsilon_{PEHE} \leq 2(\varepsilon_F^{(1)} + \varepsilon_F^{(0)}) + 2B_\Phi \cdot IPM(p_\Phi^{(1)}, p_\Phi^{(0)})$$

Theoretical Justification (Shalit et al., 2017)

Precision in Estimation of Heterogeneous Effect

$$\varepsilon_{PEHE} = \mathbb{E}[(\hat{\tau}(x) - \tau(x))^2]$$

Basically, **Mean Squared Error** (MSE) between estimation and real outcome

Representation learning is not heuristic but it has theory behind it.

Integral Probability Metrics (IPM)

Error decomposes into:

- **Factual** prediction **error**
- **Distribution imbalance**
- Reducing IPM reduces counterfactual error
- **Geometry** directly **affects identifiability**

This is essentially a domain adaptation upper bound applied to causal inference.

MMD: Measuring Imbalance

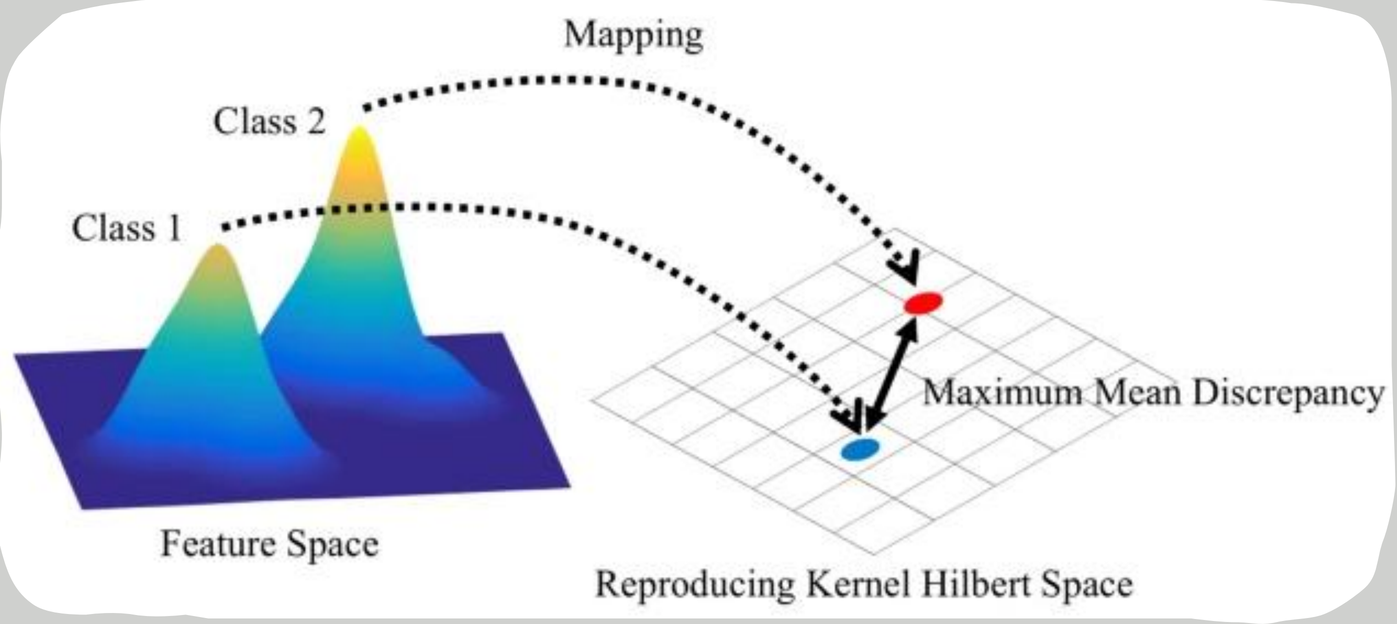
Maximum Mean Discrepancy

$$MMD^2 = \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\mathbb{E}[k(X, Y)]$$

Within-group similarities

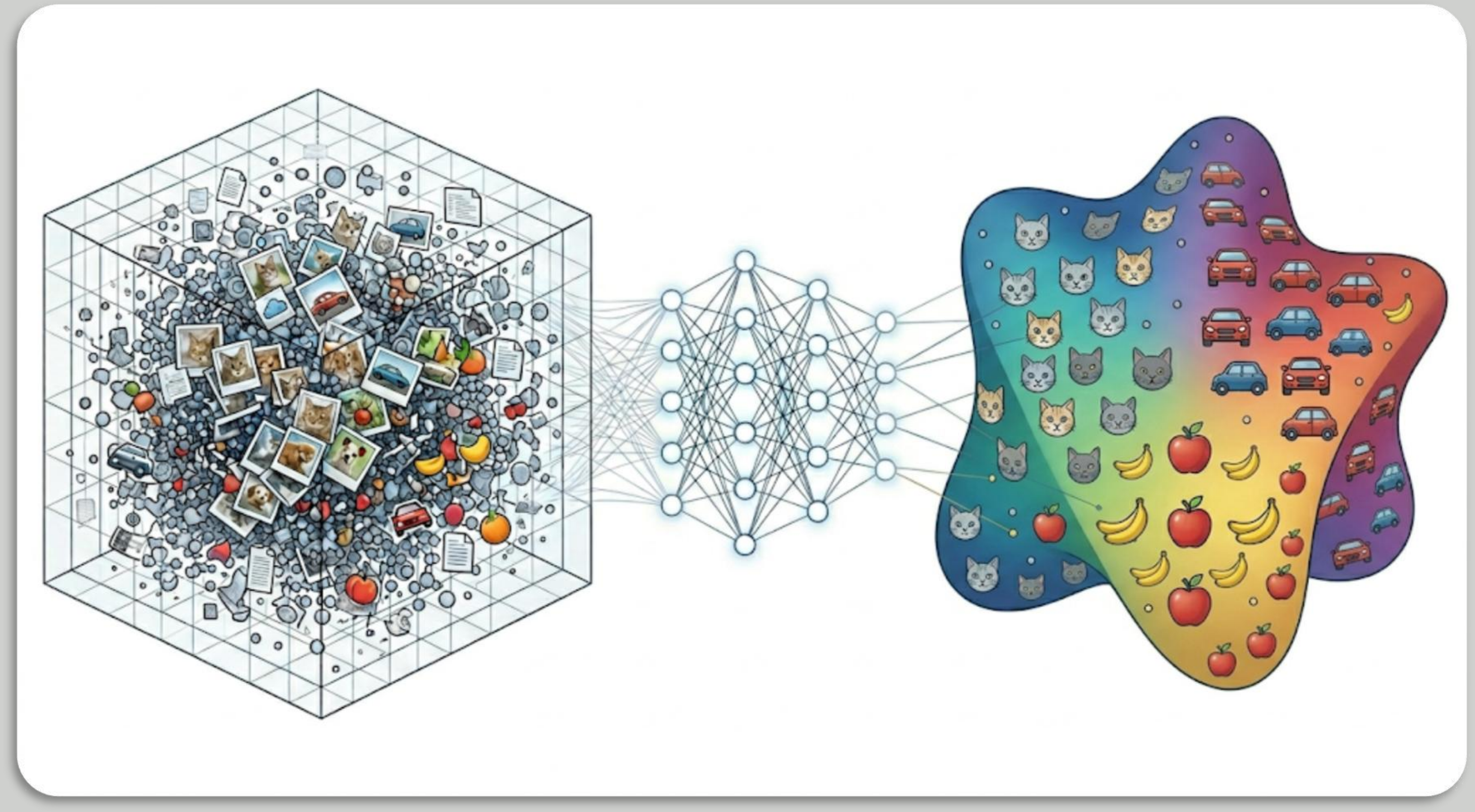
Between-group similarity

Measures distribution gap



We now have a geometric distance metric for distributions too

Learning a Better Geometry



- Learn **representation Φ(X)**
- **Reduce imbalance** in latent space
- **Improve counterfactual** stability

$$\Phi(X) \to \text{induced geometry}$$

$$\mathcal{L} = \text{Outcome Loss} + \alpha \cdot IPM$$

$$p(\Phi(1)) \approx p(\Phi(0))$$

Better representations may improve balance and reduce extrapolation error.

Representation Learning for CATE

$$\Phi(X) \rightarrow \text{induced geometry}$$

1
- **Shared representation** Φ(X)
- **Separate** outcome **heads**
- Explicit **balancing regularization**

2
- **TARNet**: shared representation + two heads
- **CFRNet**: TARNet + IPM penalty
- **DragonNet**: (Shi et al.) adds propensity head + targeted regularization

3
Representation learning turns **balancing** into **optimization.**

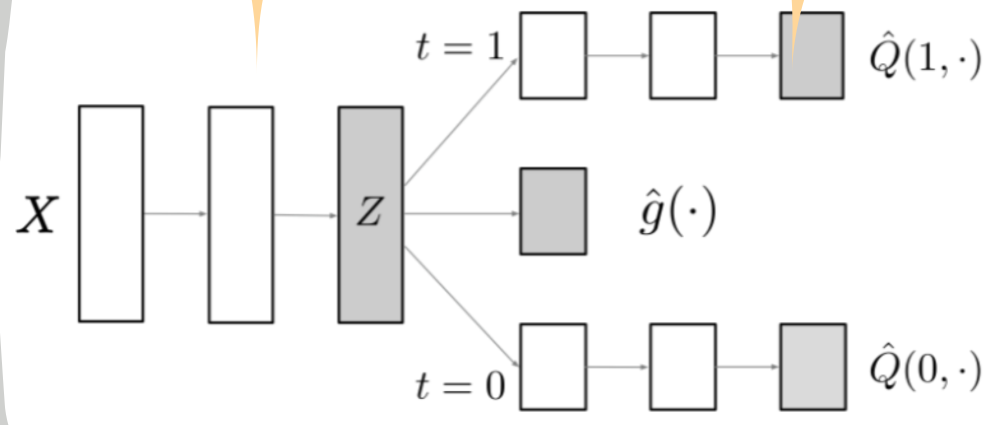


Figure 1: Dragonnet architecture.

$$\mathcal{L} = \text{Outcome Loss} + \alpha \cdot IPM + \mathcal{L}_{targeted}$$

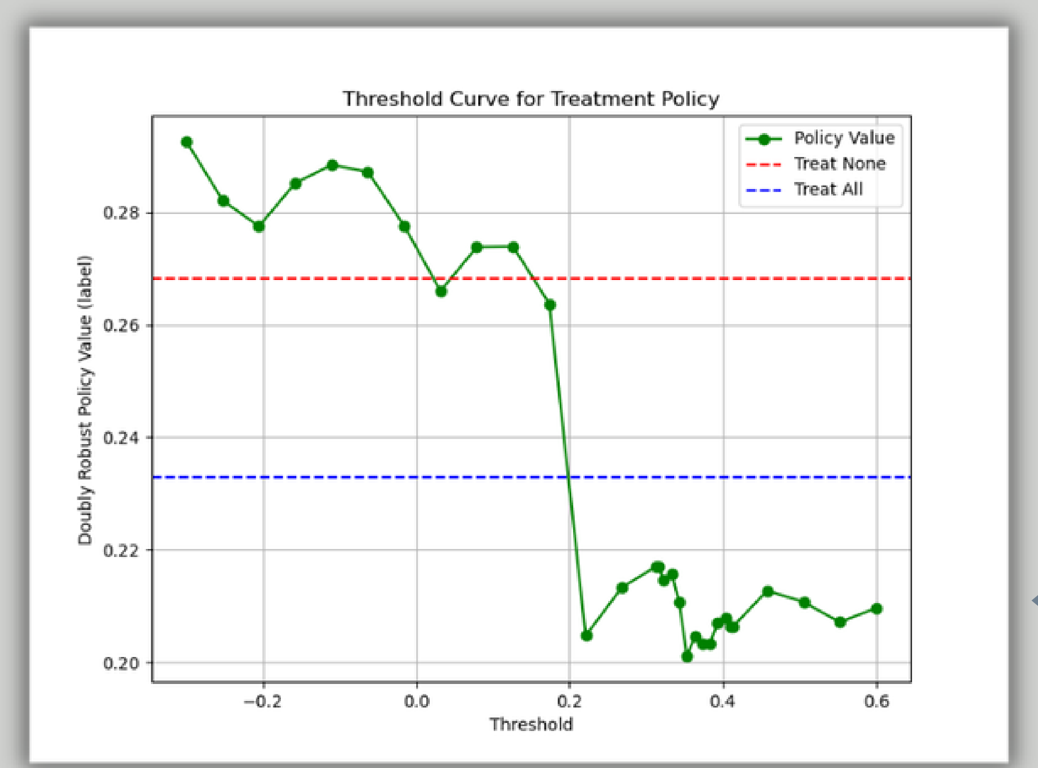
$$p(\Phi(1)) \approx p(\Phi(0))$$

From Average Effect to Decision Rule

aka Policy

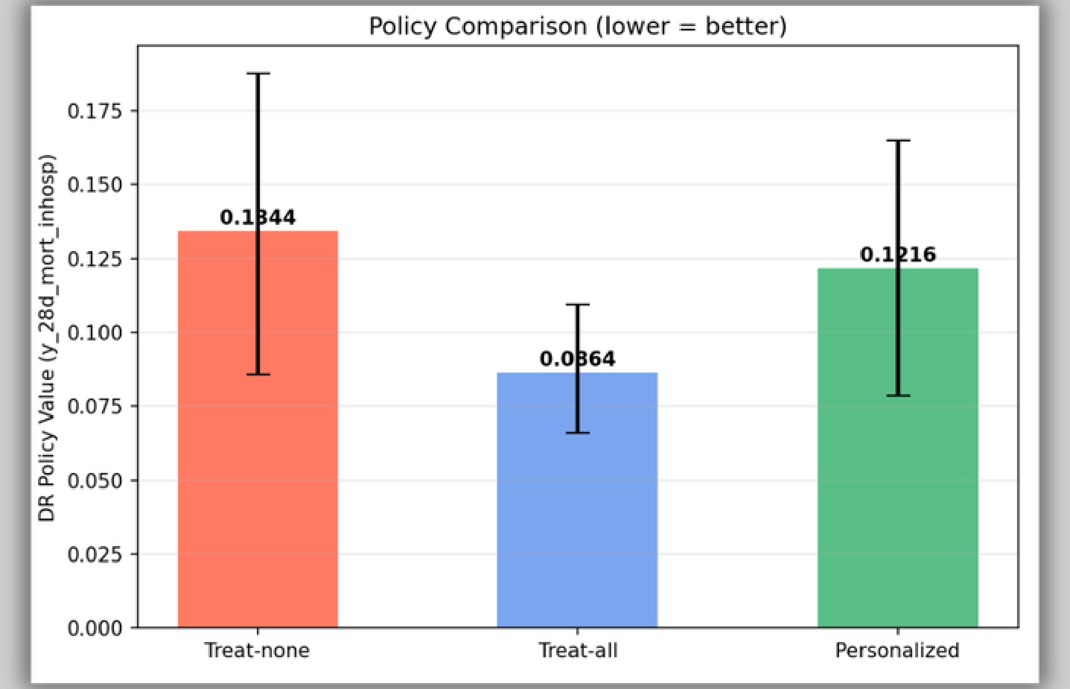
Policy

- After estimates we can define a **treatment policy**: e.g. on threshold
- Ideally if CATE is perfect: treat only if t(x) < 0



From CATE to Policy

- **Treat-all globally optimal**
- Personalized **policy > treat-none**
- Strong heterogeneity present
- Still lot of **noise** and **bias**



But **CATE** stills **reveals structure** beyond ATE.

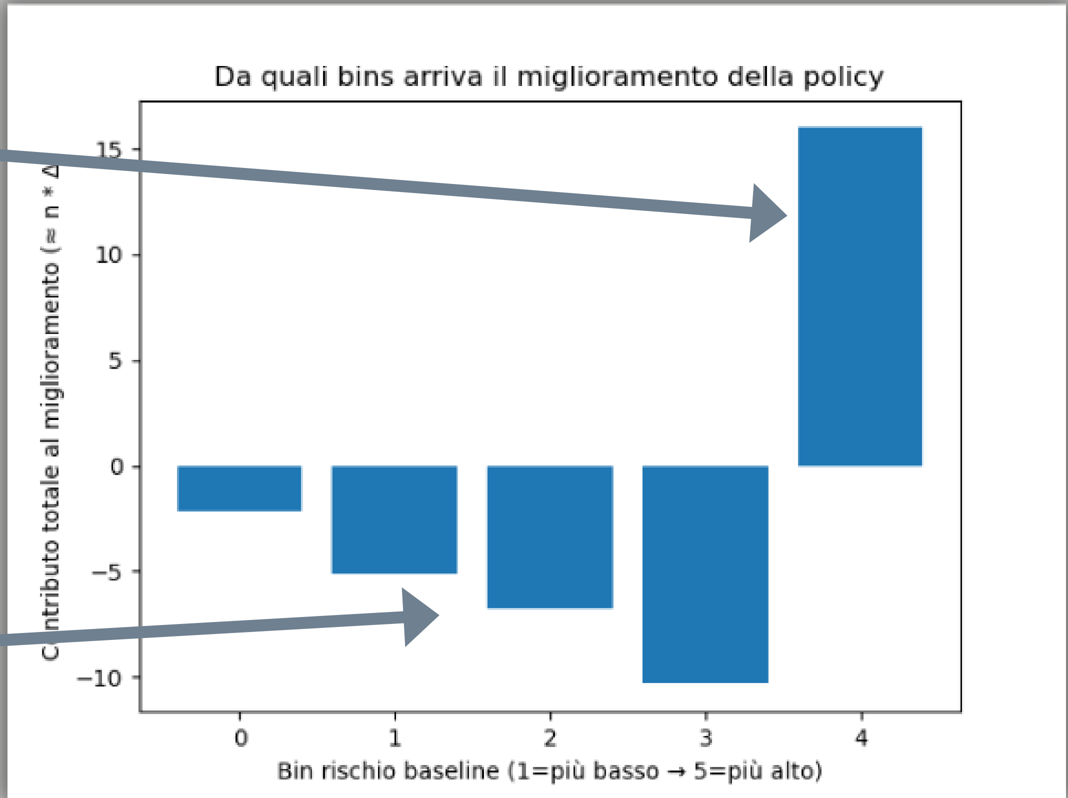
Treatment Effect Heterogeneity

Back to **Early vs Delayed Diuretics** in Sepsis

High-risk patients
benefit substantially

Clinical coherence
with severity

Low-risk
unhelpful or slighly
harmful **(?)***



Unable to define a correct policy.

- **Still bias**: $\tau(x) > 0$ for treatment group even if mortality descrease
- **Inconsistencies** in policy and CATE estimates

* Only for a **tiny subgroup**: for most is still helpful.
The rest is noise.

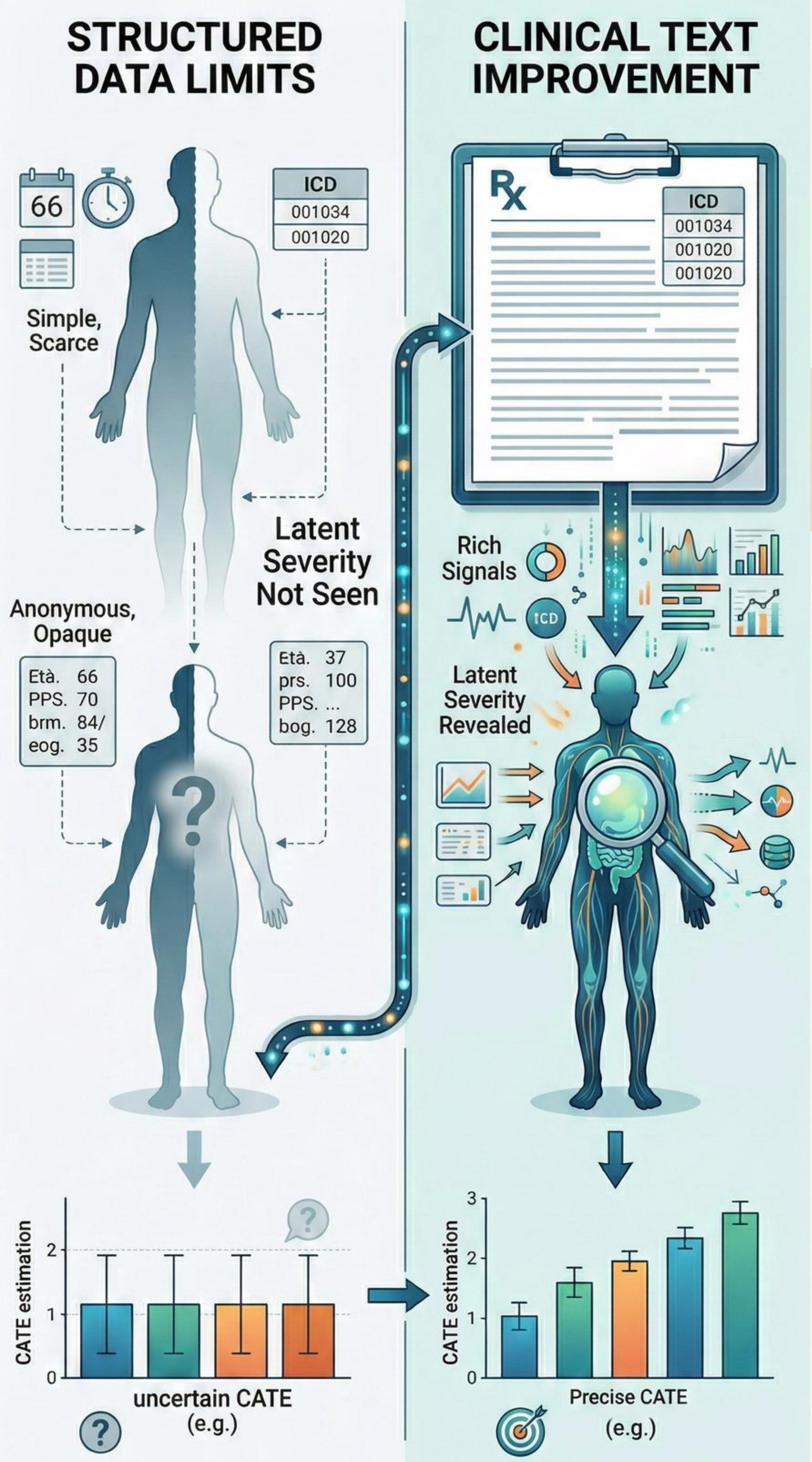
3. Role of LLMs

When Covariates are Incomplete

Limits of Tabular Covariates

- **Clinical text** contains **rich** signals
- Severity is partially **latent**
- Structured variables are **coarse summaries**
- **Ignorability** depends on **feature richness**

If relevant **confounders** are **missing** from **data**,
no estimator can fix it.



LLM Embeddings as Semantic Kernels

Foundation Models are being used as Representation Engines

- Pretrained **semantic embeddings**
- Implicit **feature construction**
- Potentially **richer** confounder **encoding**

$$\Phi_{LLM}(x) \in \mathbb{R}^d$$

$$k_{LLM}(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

- Use of **unstructured data** like medical reports

Semantic Similarity:
e.g. Idiopatic - Primary
- **Classical Learner:** 2 separate categories
- **LLMs:** semantically the same

LLMs may define a new geometry for causal inference.

Open Research Questions

Human-in-the-loop

Expert: Valides or correts priors (Domain Expertise).

Causal for LLMs

Improve **Reasoning Capacity** of LLMs helping understand **cause-effects** relationships

- Can **semantic embeddings improve overlap**?
- Can they **reduce hidden confounding**?
- How to **integrate** with **DR estimators**?
- How to **validate causal consistency**?

Active Causal Learning

LLMs now **passive:** they only analyze data.

Possible Future:

1. Agent identifies ambiguities
2. Self-presents an intervention
3. Analyzes results and updates its prior

Neuro-Symbolic Systems

LLM (Semantics): Priors

Symbolic Solver (Math): Estimating Effect

Deep Learning (Geometry): MMD and Representation Learning

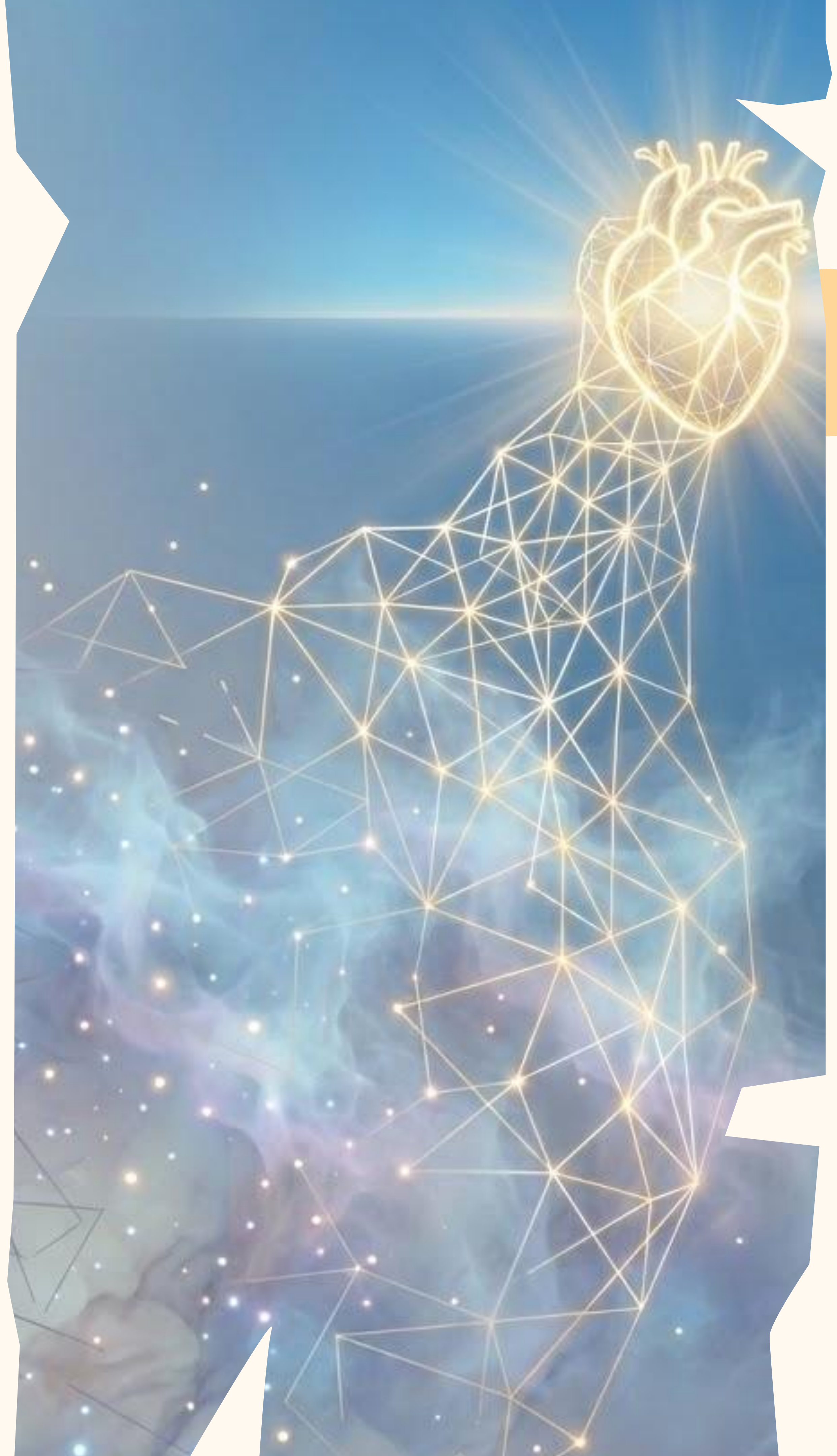
Takeaways

Causale Inference and CATE

- **CATE estimation** is fundamentally **geometric**.
- **Overlap** is **essential** for identifiability.
- **Design matters** more than algorithm.
- **Representation learning** may be the key.
- **LLMs** are being integrated in the loop

We cannot access the counterfactual world, but **we can engineer better geometries**.

Causa latet: vis est notissima. *Ovid, Metamorphoses, IV. 287.*



References

Theory and ML

- **Rosenbaum, P. R., & Rubin, D. B. (1983)**. The central role of the propensity score in observational studies for causal effects. Biometrika.
- **Gretton, A., et al. (2012)**. A kernel two-sample test. JMLR. (MMD Paper).
- **Shalit, U., Johansson, F. D., & Sontag, D. (2017)**. Estimating individual treatment effect: generalization bounds and representation learning. ICML.
- **Shi, C., Blei, D. M., & Veitch, V. (2019)**. Adapting neural networks for the estimation oftreatment effects. NeurIPS. (Dragonnet).
- **Hernan, M. A., & Robins, J. M. (2025)**. Causal inference: What if. CRC Press.

LLMs and Causality

- **Kiciman, E., et al. (2023)**. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. ArXiv. (Microsoft Research).
- **Zhang, C., et al. (2023)**. Understanding the causal reasoning capabilities of large language models. ArXiv.

elicsir
Elevating Italian Computer Science
Innovation and Research

Thanks!

Student: Cieri Manuel
Mentor: Prof. De Prisco Roberto

Bound Expanded (Q&A)

$$\varepsilon_{PEHE} \leq 2(\varepsilon_F^{(1)} + \varepsilon_F^{(0)}) + 2B_\Phi \cdot IPM(p_\Phi^{(1)}, p_\Phi^{(0)}) - 2\sigma_Y^2$$

- Factual risk $\varepsilon_F^{(t)}$
- Lipschitz constant B_Φ
- IPM term controls domain shift
- Irreducible error $-2\sigma_Y^2$

- PEHE is limited by **sum of errors on each component** (Y(1), Y(0))
- We'll use triangular inequality

$$\varepsilon_{PEHE} \leq 2(\epsilon_{Y(1)} + \epsilon_{Y(0)})$$

$\epsilon_{Y(1)}$

Expected error on prediction of Y(1) on entire population

- Total error for **Y(1)** is composed by **2 parts**, weighted on **treatment probability u** (for Y(0) is similar):

$$\epsilon_{Y(1)} = u \cdot \underbrace{\mathbb{E}_{x \sim P(X|T=1)}[\text{Loss}]}_{\text{Factual (observed)}} + (1-u) \cdot \underbrace{\mathbb{E}_{x \sim P(X|T=0)}[\text{Loss}]}_{\text{Counterfactual (unknown)}}$$

- We know the observed error, but we need to **estimate** the **counterfactual**, via IPM:

$$\mathbb{E}_Q[\text{Loss}] \leq \mathbb{E}_P[\text{Loss}] + IPM(P, Q)$$

Estimated loss is less than or equal to observed loss plus IPM distance

$$PM_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)]|$$

This inequality comes from the IPM math definition above

- **Replacing** all values:

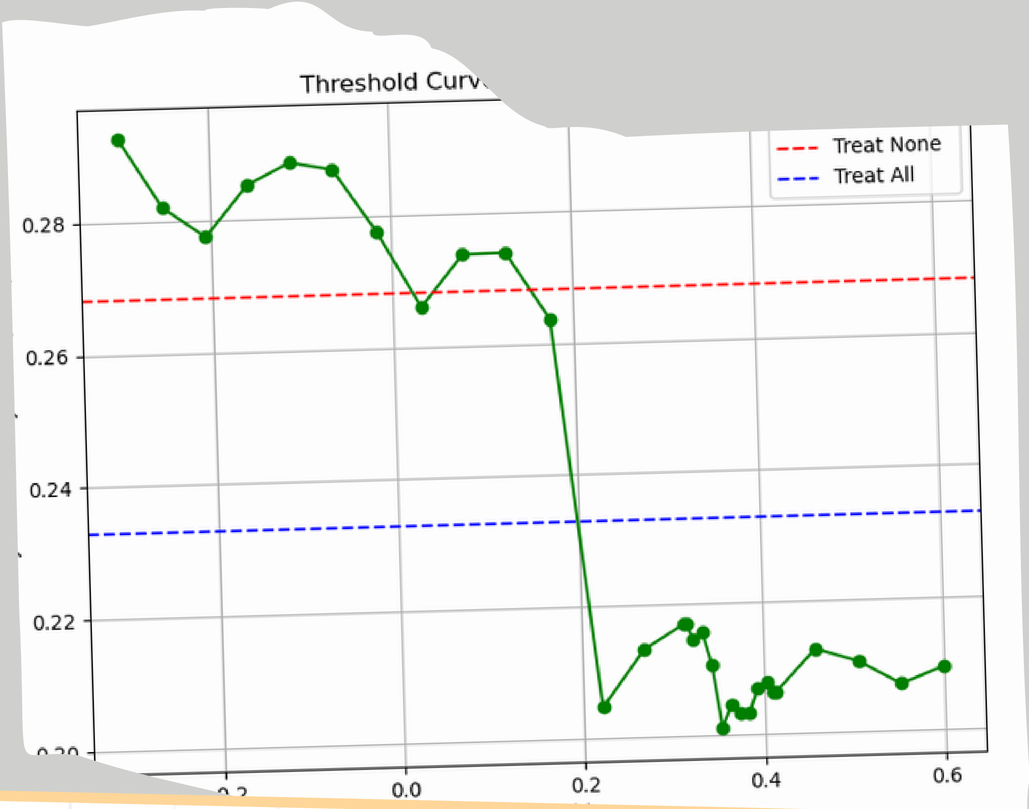
$$\epsilon_{PEHE} \leq 2 \cdot (\underbrace{\epsilon_{Factual}}_{\text{Observed Loss}} + \underbrace{B_\Phi \cdot IPM_{\mathcal{F}}(\Phi)}_{\text{Distribution Distance}} + C)$$

When we use an RCT instead (Q&A)

CATE is **finally < 0 → Better Outcome**

RCT: AIDS Treatment

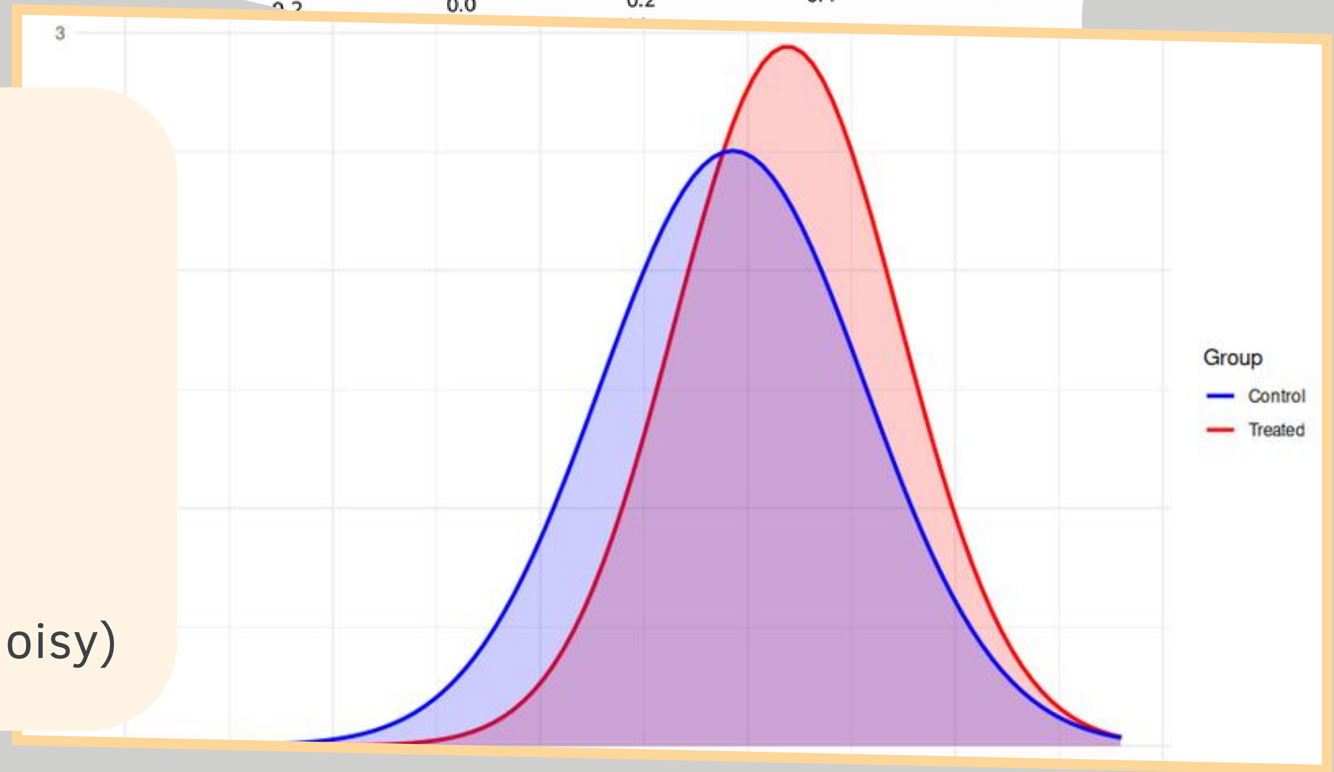
- N = 2139 patients
- **Zidovudine vs Alternative Therapies**
- Outcome: 1:failure, 0:censoring



Improved data quality

- No dominant strata
- **Balanced covariates**
- **No Bias**

- **Naive ATE**: -0.0859
- IPW ATE: -0.0696 (propensity weighted)
- DR ATE : -0.0353
- Bias corrected by DR: -0.0507
- **Mean CATE** (tau_hat): -0.1123 (SD=0.6448)
- **Frac predicted benefit**: 0.545 (tau < 0)
- Heterogeneity SNR: 18.28 (SD/|ATE|, >> 1 = noisy)



RCT is gold standard, bias is strongly reduced

Representation Models in Detail (Q&A)

TARNet

- **Shared Latent Space**
- 2 **Separate Heads:**
  - Treat Outcome
  - Control Outcome

$$\mathcal{L}_{\text{TARNet}} = \sum_i \left(Y_i - \hat{y}_{T_i}(X_i)\right)^2$$

No Explicit Causal Term:

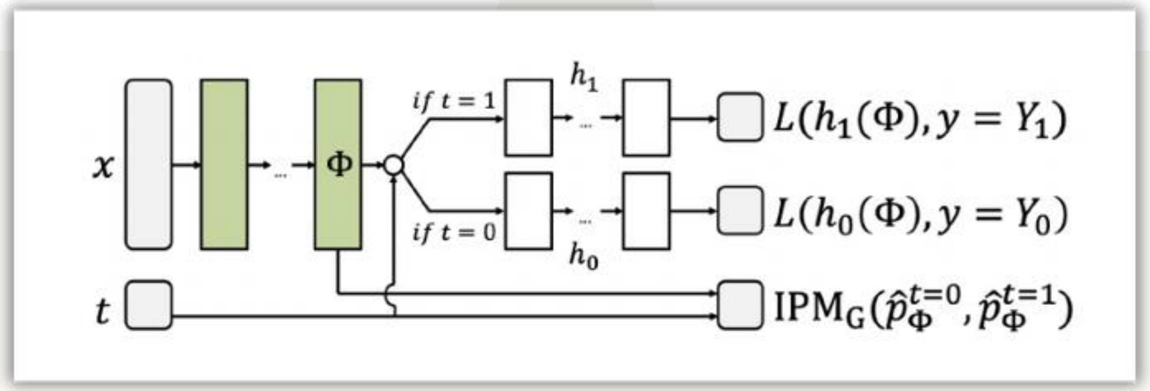
Shares weights but it doesn't balance populations

CFRNet

- Same structure as TARNet
- Adds a **balancing loss** via IPM

$$L = \mathcal{L}_{\text{TARNet}} + \underbrace{\lambda \cdot \text{IPM}\big(P(\Phi(X) \mid T=1), P(\Phi(X) \mid T=0)\big)}_{\text{CFRNet: Adds IPM}}$$

- **Global Balance** and outcome independent
- **No** use of **Propensity Score**



Dragonnet

- Adds a 3rd Head to CFRNet: **Propensity Score Head**
- **Composed Loss**

$$\mathcal{L} = \mathcal{L}_Y + \alpha\,\mathcal{L}_T + \beta\,\mathcal{L}_{\text{targeted}}$$

$\mathcal{L}_Y$ Same TARNet prediction loss

$$\mathcal{L}_T = -\sum_i \left[\; T_i \log \hat{e}(X_i) + (1 - T_i)\log(1 - \hat{e}(X_i))\right]$$

Propensity-supervised Treatment Loss

$\mathcal{L}_{targeted}$ **Targeted Learning Loss**

It's a little parametric measurement to ensure that loss correction is doubly-robust: each update must be coherent with a causal update.

Doubly Robust Learner (Q&A)

Pseudo-outcome

- If PS is wrong but outcome is correct → still **ok**
- If outcome wrong but PS correct → still **ok**

This is **double robustness** extended to CATE.

DR estimator follows this schema:

Baseline outcome model +

propensity-weighted residual correction

$$\mathbb{E}\left[\frac{T}{\tilde{e}(X)}(Y - m_1(X)) \mid X\right] = \frac{e(X)}{\tilde{e}(X)}\mathbb{E}[Y - m_1(X) \mid X, T = 1]$$

Let's assume estimated PS is wrong:

- If its too big → under-correction
- If its too small → over-correction

Influence-function-based Pseudo-outcome

$$\psi_i = \underbrace{\hat{m}_1(X_i) - \hat{m}_0(X_i)}_{\text{Estimates Delta}} + \frac{T_i}{\hat{e}(X_i)}\underbrace{(Y_i - \hat{m}_1(X_i))}_{\text{Error on Treated}} - \frac{1 - T_i}{1 - \hat{e}(X_i)}\underbrace{(Y_i - \hat{m}_0(X_i))}_{\text{Error on Control}}$$

- If m_t(X) correct, the residues are zero → IPW terms ignorable (on average)
- If e(X) is correct, IPW weights correct m_t(X) error

The Geometry (Q&A)

Hilbert Space

Hilbert Space (HS) extends Euclidean geometry to functions :

- **Dot product**, compute angles and orthogonality: $\langle f, g \rangle = \sum_{i=1}^{n} u_i v_i$
- **Norm** (L2), "amplitude/energy": $\|f\| = \sqrt{\langle f, f \rangle}$
- **Completeness**: limits of sequences tends to values inside the space. Needed for Optimization (e.g. gradient descent)

Usually we use, *L squared*, were the integral of the square of the function must be finite: $\int f(x)^2 dx < \infty$

Dot product, here is: $\langle f, g \rangle = \int f(x) g(x) dx$

RKHS

Problem

- Two functions, f and g, may have distance equal to zero but have different values in a specific point in HS due to punctual evaluation not being a continuous functional.

RKHS: Reproducing Kernel Hilbert Space

- "Special" space where we impose that an *evaluation functional* must be continous and limited:

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}} \quad \forall x \in \mathcal{X}$$

Riesz Representation Theorem

In a Hilbert space, if a linear functional is continous there exits only one vector g in the space, such that the functional can be expressed as a dot product with g.

Reproducing Property is the direct conseguence:

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}$$

With K point defined as **Feature Map**, uniquely mapping each raw point x in the Hilbert space.

We can now define the Kernel Function k(x, y):

$$k(x, y) := \langle K_x, K_y \rangle_{\mathcal{H}} = K_x(y)$$

Function Classes: L^2 vs RKHS (Q&A)

Class of Functions

Why we need a function class $\mathcal{F}$:

$$IPM(p,q) = sup_{f\in\mathcal{F}} \mid E_p[f(X)] - E_q[f(X)] \mid$$

• To avoid degenerate solutions, we must **restrict** $\mathcal{F}$;

• This class choice defines the **notion of distance**.

L^2 space

$$\| f \|_{L^2}^2 = \int f(x)^2 dx < \infty$$

- Controls **global magnitude** of functions;
- Based on **integrals** (average behavior).

- Does **not control pointwise** values;
- **Unstable** for comparing individuals.

RKHS

$$f(x) = \langle f, k(x,\cdot)\rangle$$

- HS with **kernel-induced geometry**;
- **Pointwise evaluation** is continuous;
- **Norm** controls function **everywhere**;
- Enables **kernel methods**;

Pipeline

Step 1- Data: $x \in \mathbb{R}^d$ →patient features

Step 2 - Feature map: $\Phi(x) = k(x,\cdot)$ →each patient becomes a similarity function

Step 3 - Distribution embedding: $\mu_p = \mathbb{E}[\Phi(X)]$ →distributions become mean functions

Step 4 – Distance: MMD →compare treated vs control

Geometry (kernel) + **Averaging** (expectation) → **Distribution Comparison**

SUTVA Violations (Q&A)

Consistency Example

- **Dataset:** Autistic Children in a RCT Study
- **Treatment:** Robot-assisted social skills intervention
- **Control group**: Physician-assisted Therapy
- **Outcome:** Improvements in social skills

- **Problem:** In control group, different children were followed by different doctors whose approach and outcome differs from each other

- **Conseguence:** No consistency in control group

No Interference Example

- **Dataset**: Adults in a community
- **Treatment:** Vaccine for a viral disease
- **Control:** No vaccine
- **Outcome:** Health improvement/Disease Incidence

- **Problem:** Control group is still “protected” from the disease due to Herd Immunity.

- **Conseguence:** Assignment to subjects on Treatment Group influence outcome of Control Group → No interference violated.

Often SUTVA violations derive from erroneous test definitions

Thanks!

(Again)