Transforming Medicine and Healthcare through Machine Learning and AI

Mihaela van der Schaar

John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine

University of Cambridge

Alan Turing Institute

UNIVERSITY OF CAMBRIDGE

The Alan Turing Institute

ML-AIM Group aims to transform medicine and healthcare by *developing new methods* in Machine Learning & Artificial Intelligence

The 5 Challenges of Personalized Medicine and Healthcare

1. Lifestyle optimization and disease prevention
2. Disease detection and prediction of disease progression (longitudinal)
3. Best interventions and treatments

4. State-of-the-art tools for clinicians & healthcare professionals to deliver high-quality care

5. Optimization of healthcare systems (quality, efficiency, cost effectiveness, robustness, scalability)

Why ML-AIM can solve these challenges?

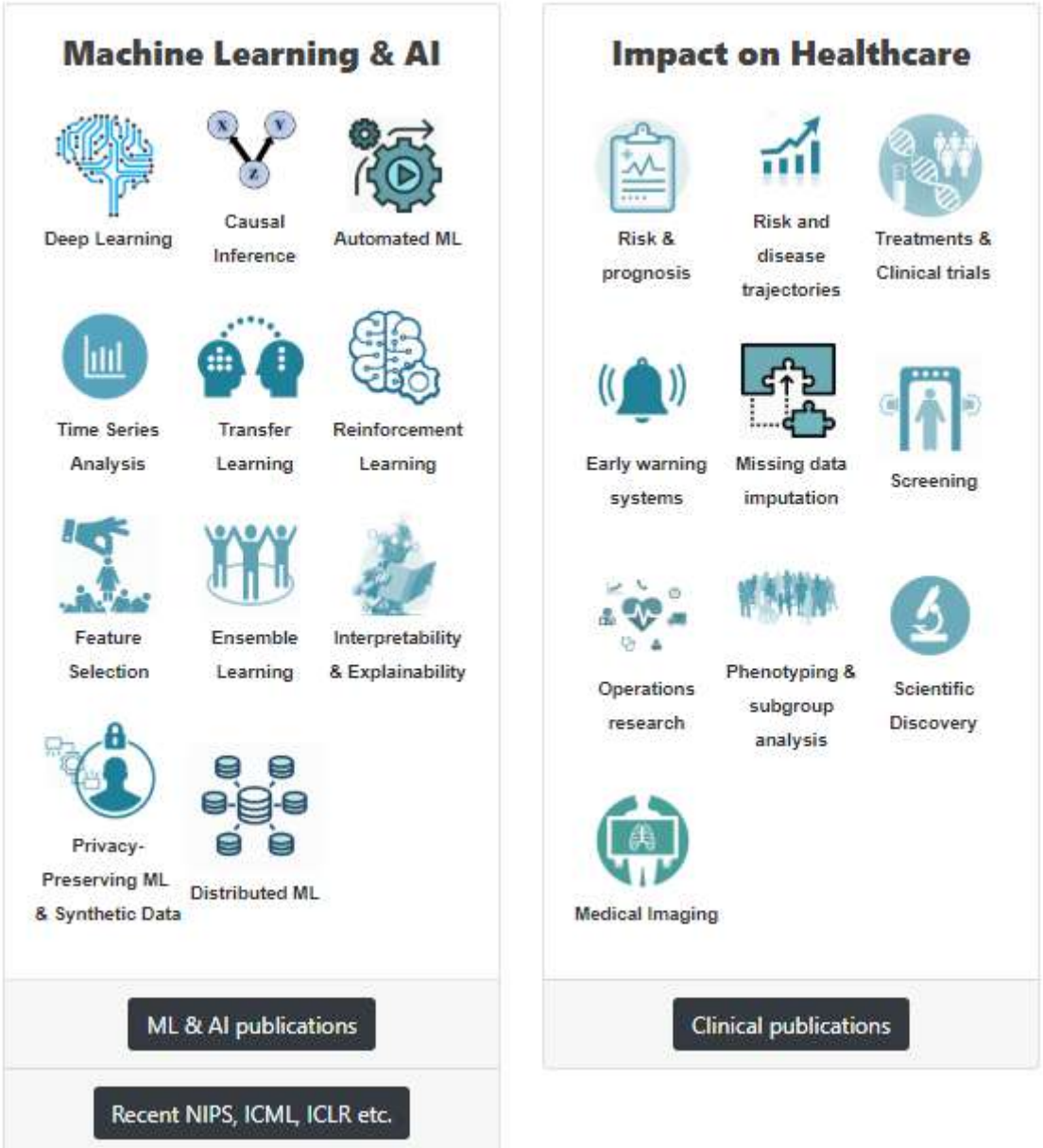
Unique expertise

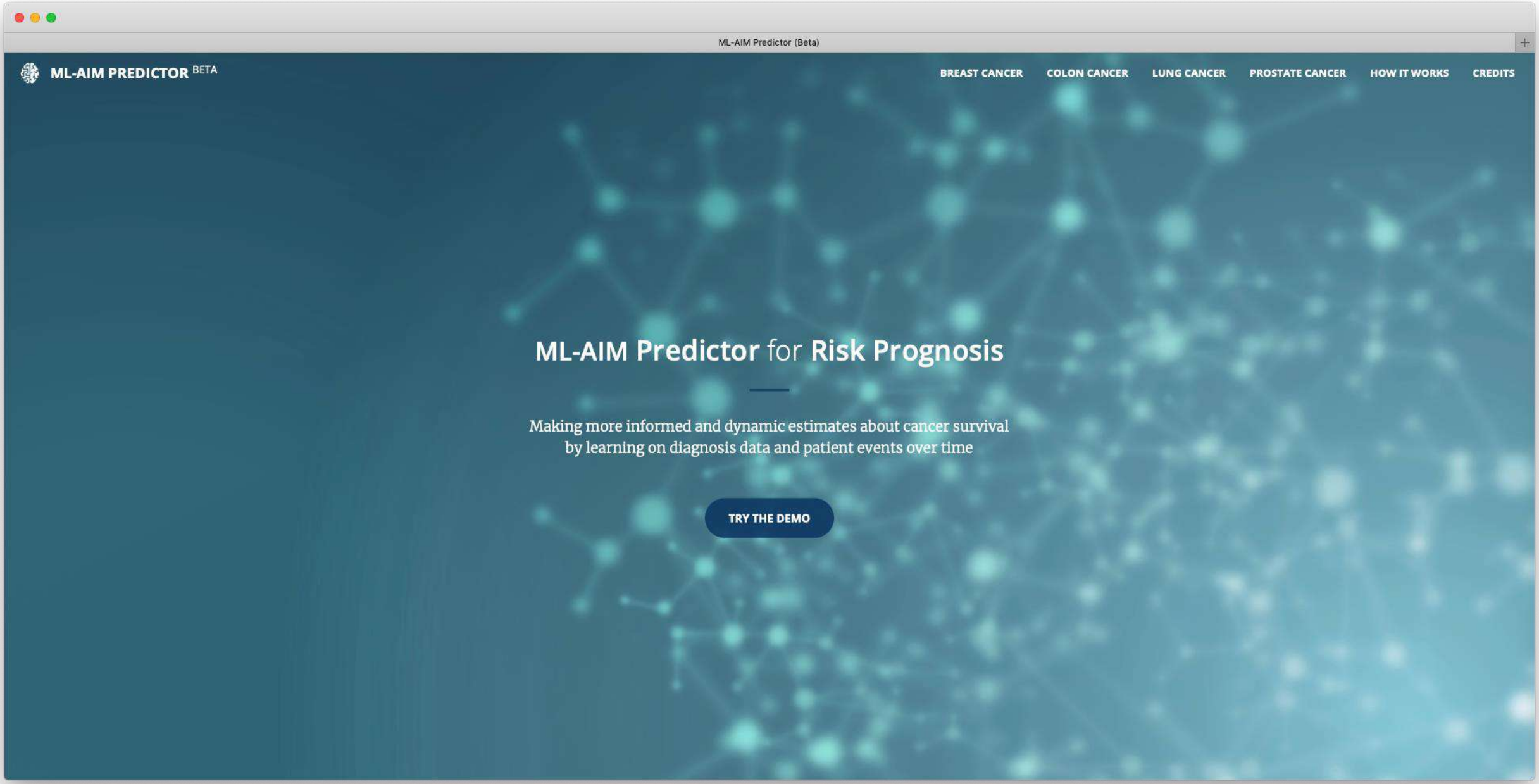
Developing and combing new methods in

- Machine Learning and Artificial Intelligence
- Applied Mathematics and Statistics
- Operations Research
- Engineering, incl. distributed computing

Working with numerous clinical and medical collaborators to make an impact on medicine and healthcare

ML-AIM group: http://www.vanderschaar-lab.com




ML-AIM Predictor (Beta)
ML-AIM PREDICTOR BETA
BREAST CANCER
COLON CANCER
LUNG CANCER
PROSTATE CANCER
HOW IT WORKS
CREDITS
ML-AIM Predictor for Risk Prognosis
Making more informed and dynamic estimates about cancer survival
by learning on diagnosis data and patient events over time
TRY THE DEMO

Turing Lecture: Transforming medicine through AI-enabled healthcare pathways

https://www.youtube.com/watch?v=TWI-WIoWvfk

Part 1: Automate the process of designing Clinical Predictive Analytics at Scale

Hospital care

Cardiovascular disease
- Risk of CVD events
- Mortality risk after heart-failure
- Mortality risk – Cardiac transplantation

Cancer: Breast, Prostate, Colon

Cystic Fibrosis

Asthma

Alzheimer's disease

Machine Learning in Clinical Research

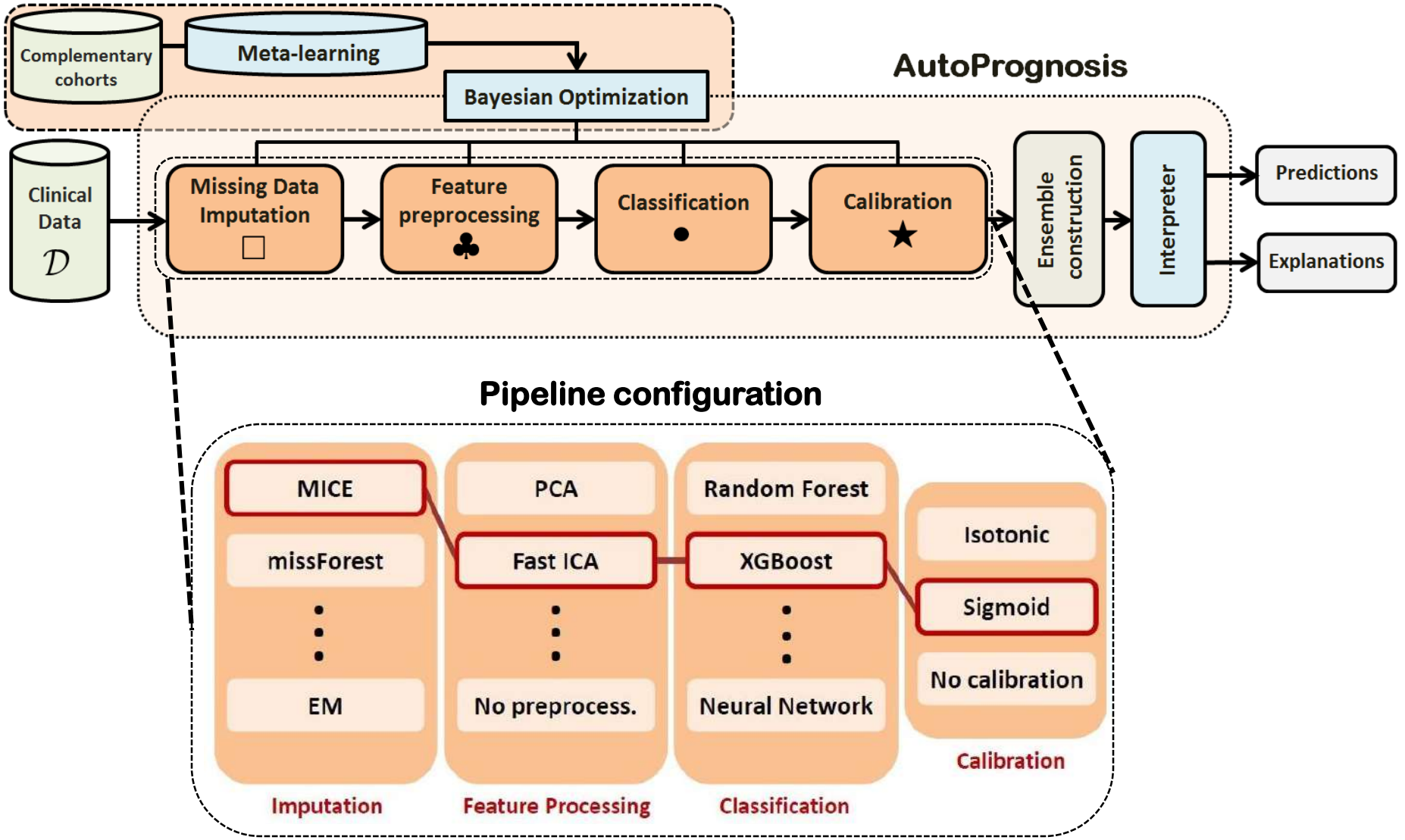
+ High predictive accuracy (for some diseases)
+ Data-driven, few assumptions
- Many ML algorithms: Which one to choose?
- Many hyper-parameters: Need expertise in data science

AUROC	MAGGIC	UK Biobank	UNOS-I	UNOS-II
Best ML algorithm	0.80 ± 0.004	0.76 ± 0.002	0.78 ± 0.002	0.65 ± 0.001
	NN	GradientBoost	ToPs	ToPs
Best Clinical Score	0.70 ± 0.007	0.70 ± 0.003	0.62 ± 0.001	0.56 ± 0.001
Cox PH	0.75 ± 0.005	0.74 ± 0.002	0.70 ± 0.001	0.59 ± 0.001

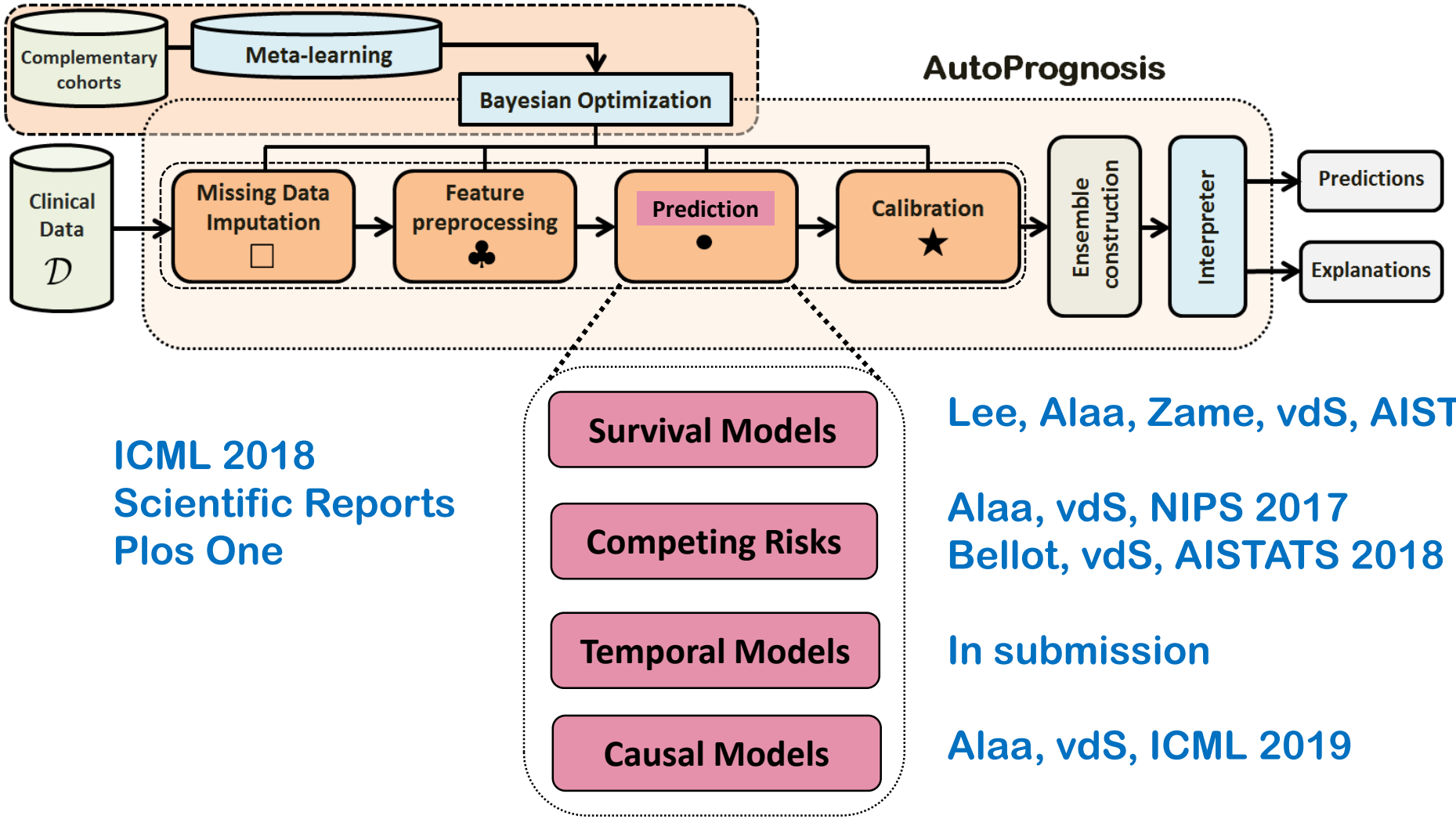
- Can we predict in advance which method is best?
- Can we do better than any individual method?
- Many metrics of performance (AUROC, AUPRC, C-index, quality of well-being)

AutoPrognosis [Alaa & vdS, ICML 2018]:

A tool for crafting Clinical Scores



Automated ML for clinical analytics (beyond predictions)



ICML 2018
Scientific Reports
Plos One

Lee, Alaa, Zame, vdS, AISTATS 2019

Alaa, vdS, NIPS 2017
Bellot, vdS, AISTATS 2018

In submission

Alaa, vdS, ICML 2019

AutoPrognosis: Exemplary technology in Topol Review

9.

Predictive analytics:

Future technology

Risk assessment and prognosis are crucial in many areas of medical practice. Predictive analytics, based on machine learning, have recently been shown to provide more accurate predictions than clinical risk scores. An important recent advance is the AutoPrognosis[103] framework, for risk score development in varied clinical settings. It can automatically discover the relevant risk factors and automatically makes design choices on which algorithms to use. This framework will provide medical clinicians and researchers, with little or no expertise in machine learning, the ability to develop the risk scores needed for their particular situations,

Solution

IPredictive analytic[104] based on AutoPrognosis have shown a 35% improvement in prediction accuracy, compared to existing statistical methods or clinical risk scores, for determining whether a cystic fibrosis (CF) patient should be referred for a lung transplant.

The same AutoPrognosis framework was shown to estimate cardiovascular risk more accurately than current risk scores, especially for patients with co-morbidities such as diabetes.

Roles/functions change

- As predictive analytics are increasingly used and embedded in the electronic patient record, their use will become more ubiquitous. They can be used by clinicians and nurses to better diagnose the patient at hand and by healthcare policy makers to enhance and individualise screening programmes, leading to better allocation of clinical resources.

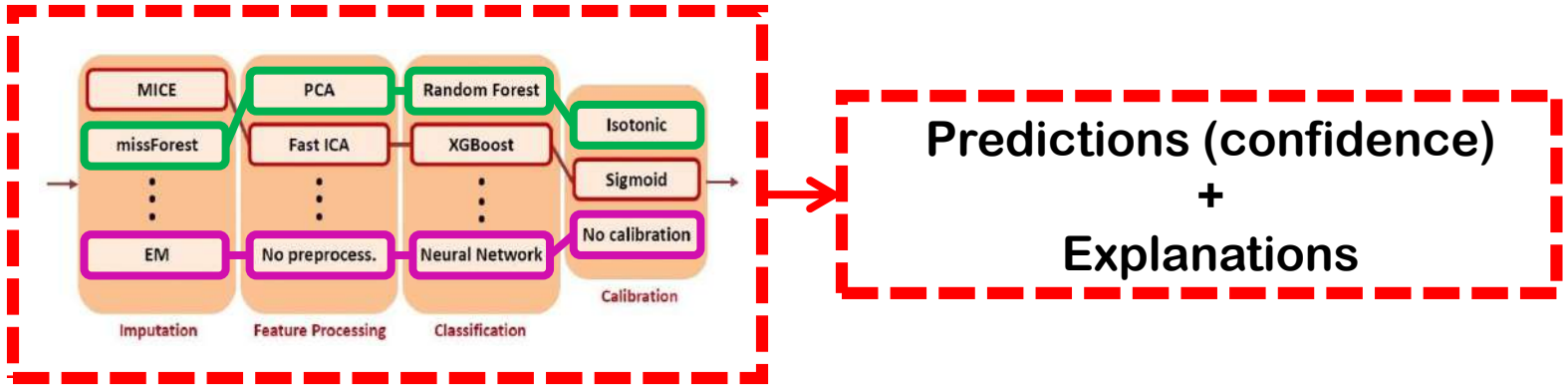
Education/training requirements

- Learn how to integrate predictive analytics into the care and diagnosis pathway, and interpret predictive results.
- Educate/train clinicians and scientists to use frameworks like AutoPrognosis in order to design new predictive analytics, which may be useful for a specific clinician or healthcare organisation.

Disease areas: Cystic Fibrosis, Cardiovascular Disease, Breast cancer, Prostate cancer etc.

Not only black-box predictions, also interpretations

- Essential for trustworthiness, transparency etc.



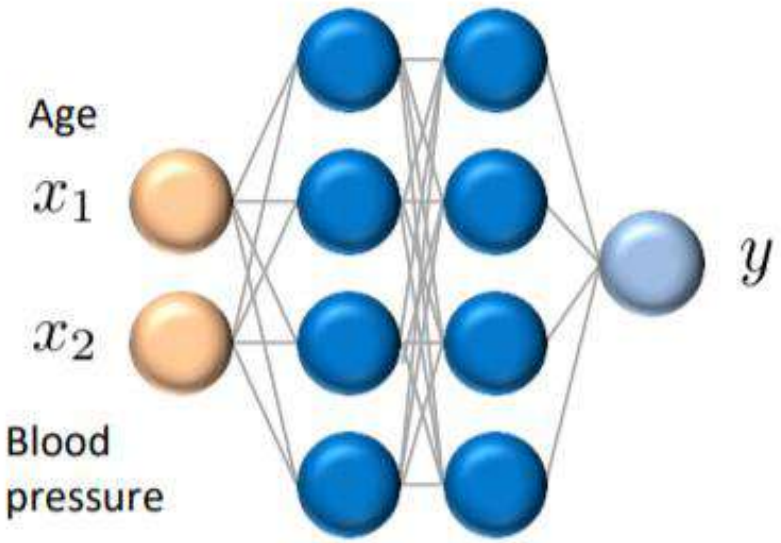
Black-box model

- INVASE: Instance-wise Variable Selection using Deep Learning [Yoon, Jordon, vdS, ICLR 2019]
- Clinician-AI interaction using Reinforcement Learning [Lahav, vdS, NeurIPS workshop 2018]
- Metamodeling [Alaa, vdS, 2019]

Interpretability using symbolic metamodeling
[A. Alaa & vdS, NeurIPS 2019]

From black-box models to white-box functions

Black-box Model $f(\mathbf{x})$



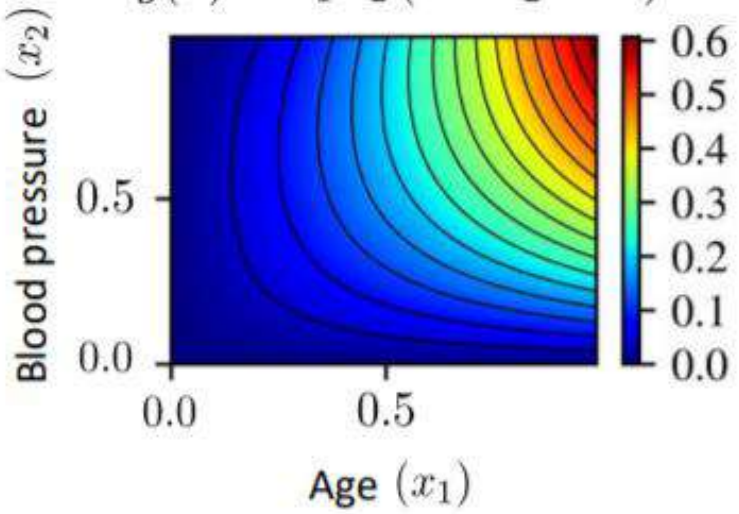
Symbolic Metamodeling

$$g(\mathbf{x}) = G(\mathbf{x}; \theta^*)$$

$$\theta^* = \arg\min_{\theta \in \Theta} \ell(f(\mathbf{x}), G(\mathbf{x}; \theta))$$

Metamodel $g(\mathbf{x})$

$$g(\mathbf{x}) = x_1\, x_2\, (1 - x_2\, e^{-x_1})$$



A symbolic metamodel takes as an input a trained machine learning model and outputs a transparent equation describing the model's prediction surface

Part 2:
From Individualized Predictions to Individualized Treatment Effects

Individualized Treatment Recommendations

Bob

Diagnosed with Disease X

Which treatment is best for Bob?

- **Problem:**
  Estimate the effect of a treatment/intervention on an individual

RCTs do not support Personalized Medicine

Randomized Control Trials:
Average Treatment Effects

Population-level

Non-representative patients
Small sample sizes
Time consuming
Enormous costs

Adaptive Clinical Trials
[Atan, Zame, vdS, AISTATS 2019]
[Shen, van der Schaar, 2019]

Delivering Personalized (Individualized) Treatments

Randomized Control Trials: Average Treatment Effects

Population-level

Non-representative patients
Small sample sizes
Time consuming
Enormous costs

Machine Learning: Individualized Treatment Effects

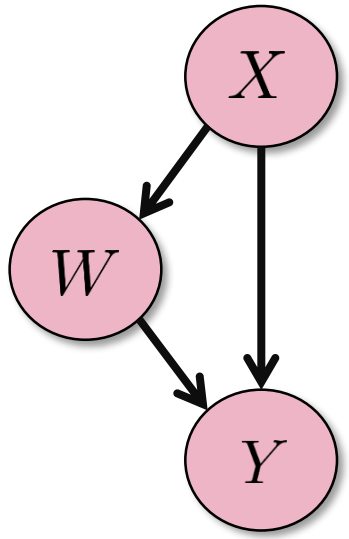
Patient-centric

Real-world observational data
Scalable & adaptive implementation
Fast deployment
Cost-effective

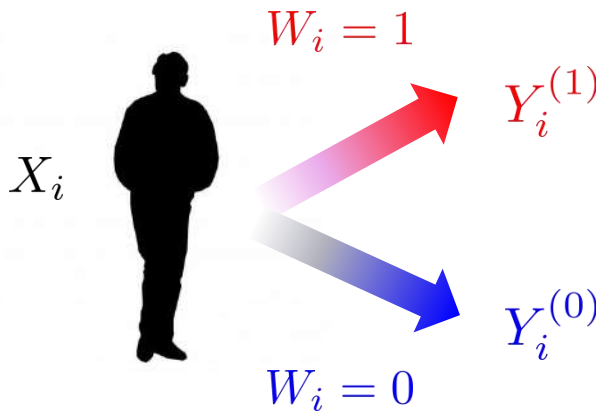
[Atan, vdS, 2015, 2018]
[Alaa, vdS, 2017, 2018, 2019]
[Yoon, Jordon, vdS, 2017]
[Lim, Alaa, vdS, 2018]
[Bica, Alaa, vdS, 2019]

Potential outcomes framework [Neyman, 1923]

Observational data (X_i, W_i, Y_i)



- Each patient i has features $X_i \in \mathcal{X} \subset \mathbb{R}^d$
- Two potential outcomes $Y_i^{(1)}, Y_i^{(0)} \in \mathbb{R}$
- Treatment assignment $W_i \in \{0, 1\}$



Factual outcomes

$$Y_i = W_i\, Y_i^{(1)} + (1 - W_i)\, Y_i^{(0)}$$

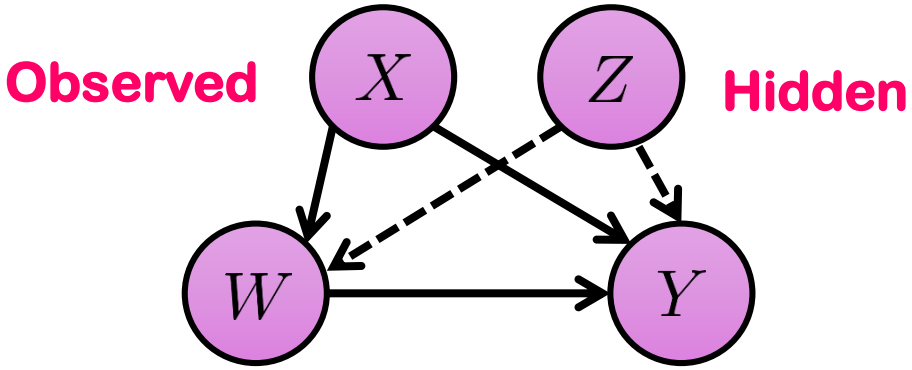
Causal effects

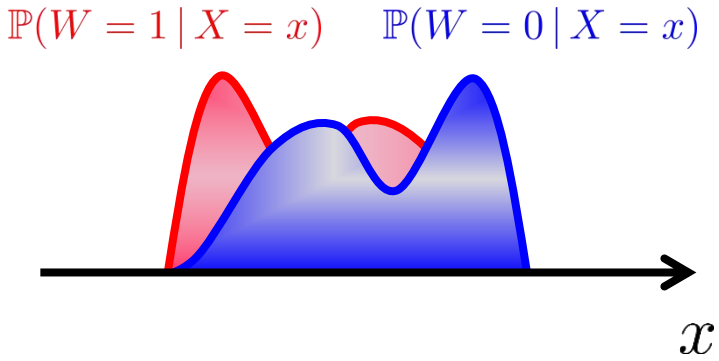
$$T(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \,\middle|\, X_i = x \right]$$

Assumptions

No unmeasured confounders (Ignorability)

Common support





Our work on hidden confounders

[Lee, Mastronarde, van der Schaar, 2018]

[Bica, Alaa, van der Schaar, 2019]

The learning problem

- **Response surfaces**

$$f_1(x) = \mathbb{E}[Y^{(1)} \mid X = x]$$

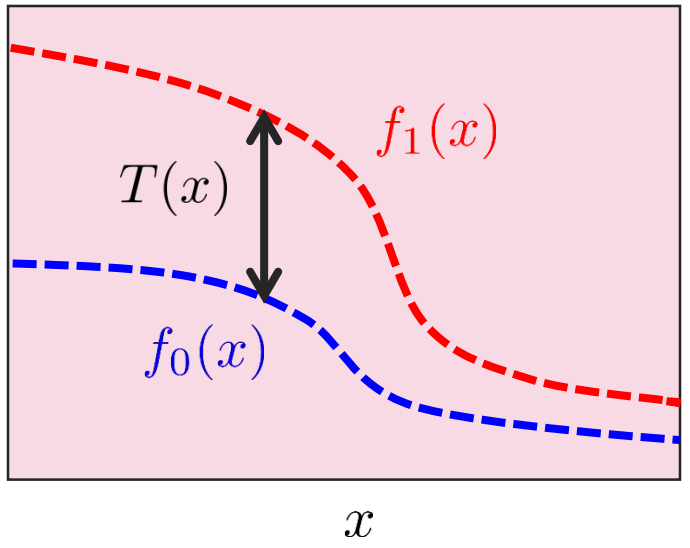
$$f_0(x) = \mathbb{E}[Y^{(0)} \mid X = x]$$

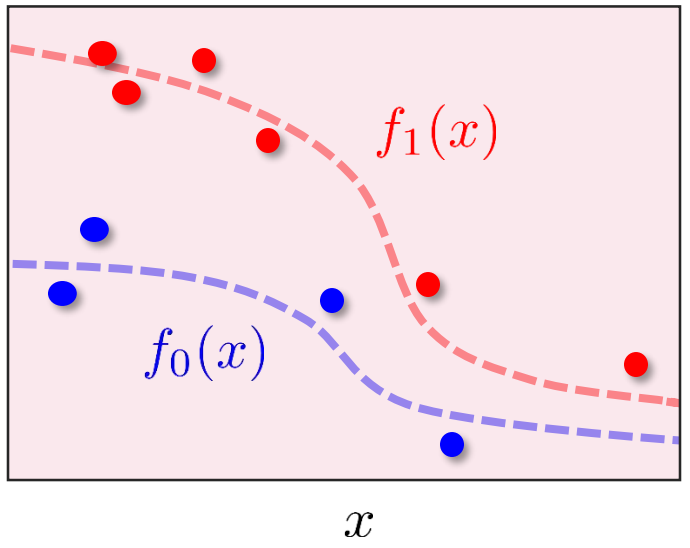
- **Causal effects**

$$T(x) = f_1(x) - f_0(x)$$

- **Observational data**

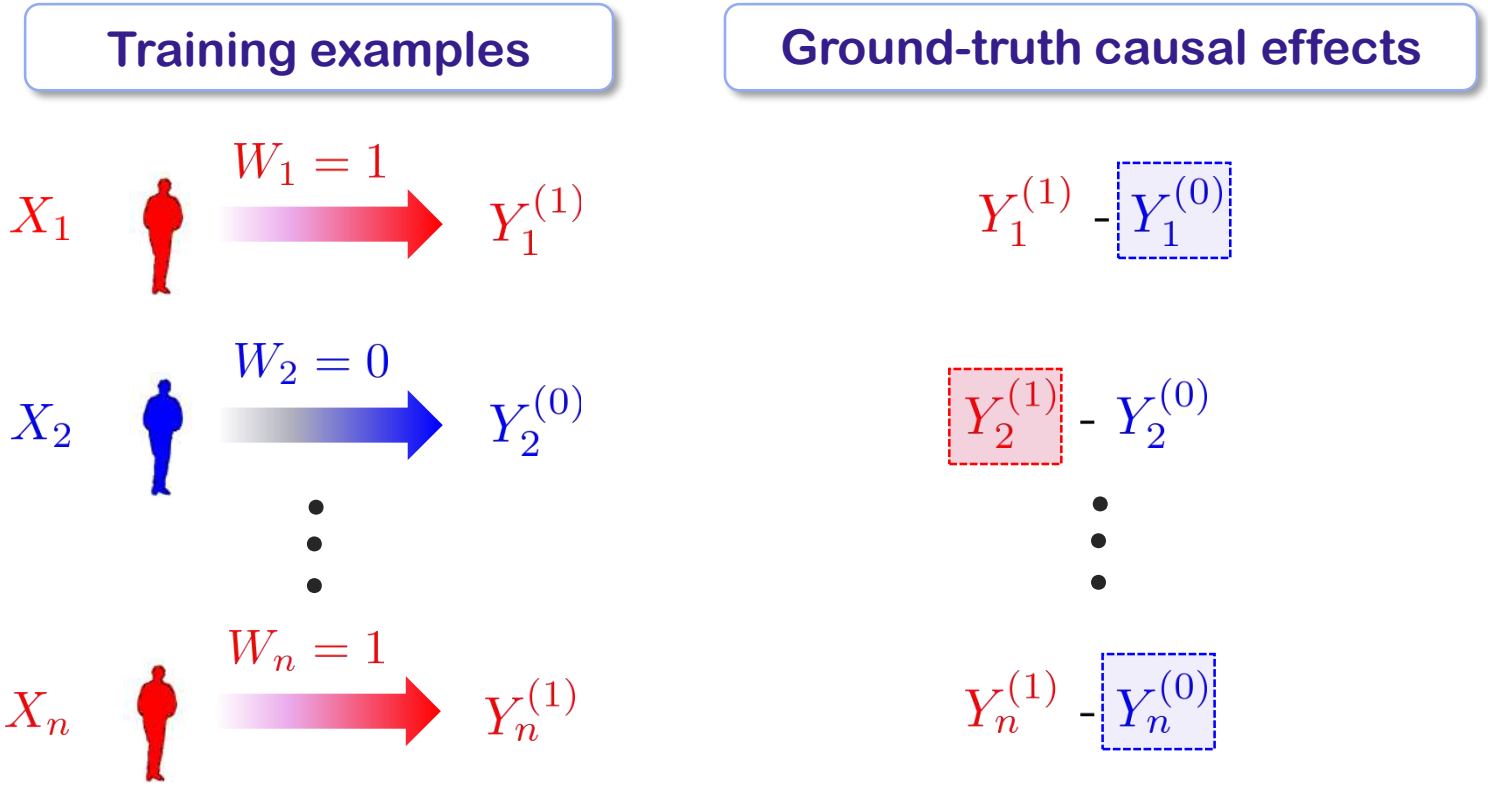
$$(X_i, W_i, Y_i) \begin{cases} W_i = 1 \rightarrow (X_i, Y_i) \\ W_i = 0 \rightarrow (X_i, Y_i) \end{cases}$$





Beyond supervised learning…

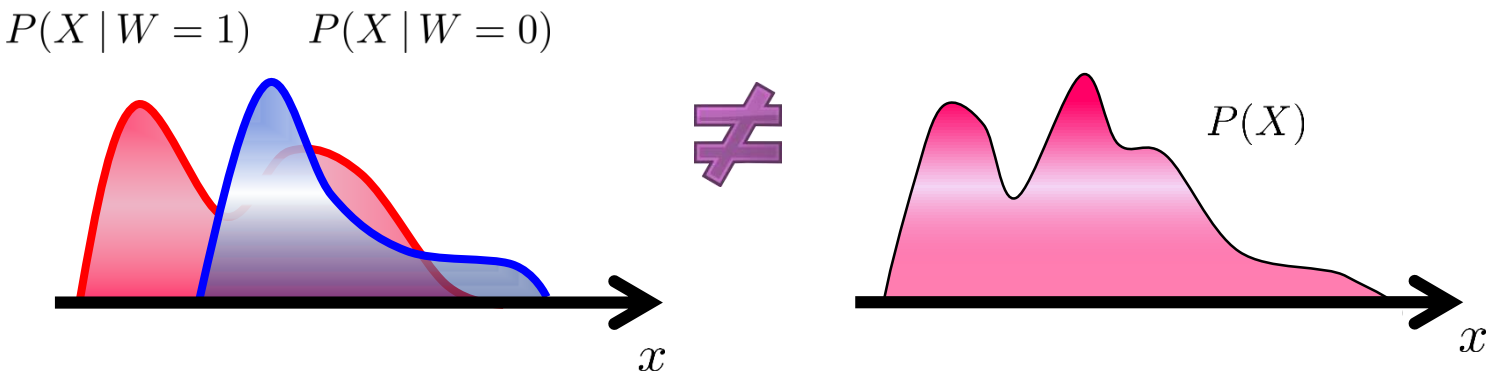
- **"The fundamental problem of causal inference"**
  is that we never observe counterfactual outcomes



Causal modeling ≠ predictive modeling

1- Need to model interventions (X_i, W_i, Y_i)

2- Selection bias → covariate shift:
training distribution ≠ testing distribution



Training distribution

Testing distribution

Previous works on treatment effects

- Bayesian Additive Regression Trees (BART) [Chipman et. al, 2010], [J. Hill, 2011]
- Causal Forests [Wager & Athey, 2016]
- Nearest Neighbor Matching (kNN) [Crump et al., 2008]
- Balancing Neural Networks [Johansson, Shalit and Sontag, 2016]
- Causal MARS [Powers, Qian, Jung, Schuler, N. Shah, T. Hastie, R. Tibshirani, 2017]
- Targeted Maximum Likelihood Estimator (TMLE) [Gruber & van der Laan, 2011]
- Counterfactual regression [Johansson, Shalit and Sontag, 2016]
- CMGP [Alaa & van der Schaar, 2017]

No theory, ad-hoc models

A first theory for causal inference - individualized treatment effects

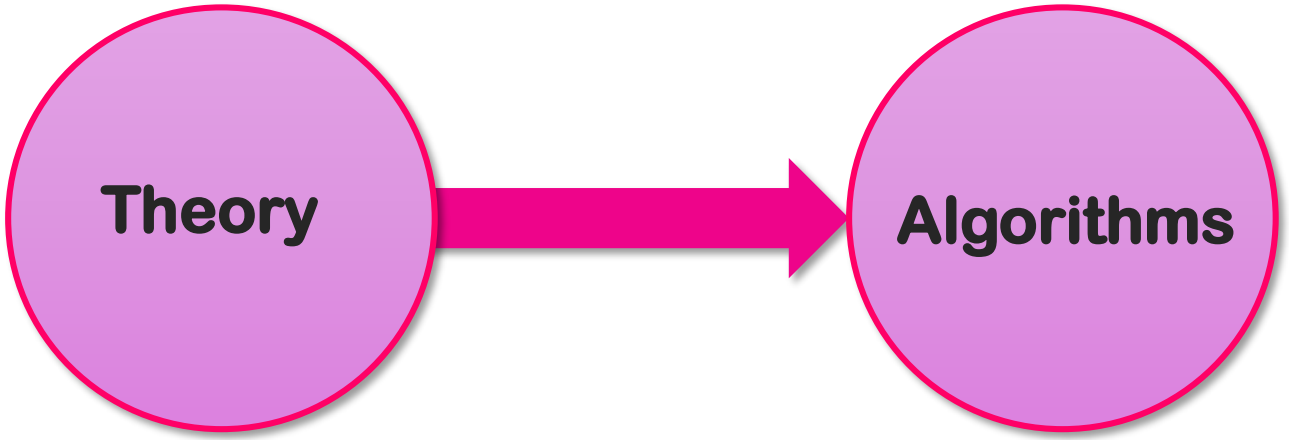
[Alaa, van der Schaar, JSTSP 2017][ICML 2018]

What is possible?

(Fundamental limits)

How can it be achieved?

(Practical implementation)



Fundamental limits

$$Z = (X, W, Y) \sim \mathbb{P}_\theta$$

- $\widehat{T}$: estimated causal effect
- Precision in estimating heterogeneous effects (PEHE) [Hill, 2011]

$$\ell_\theta(\widehat{T}) = \|T(X) - \widehat{T}(X)\|_\theta^2$$

Minimax estimation loss: $\min_{\widehat{T}} \max_{f_0, f_1} \ell_\theta(\widehat{T})$

Best estimate

Most "difficult" response surfaces

Minimax loss = information-theoretic quantity, independent of the model.

Theoretical Foundations

- **Theorem [Alaa & van der Schaar, JSTSP 2017]**

  $f_0(x)$ **has** d_0 **relevant dimensions in a Hölder space** H^{α_0}
  $f_1(x)$ **has** d_1 **relevant dimensions in a Hölder space** H^{α_1}

  If $d_w \leq \min\{d, n\},\ w \in \{0, 1\}$**, then**

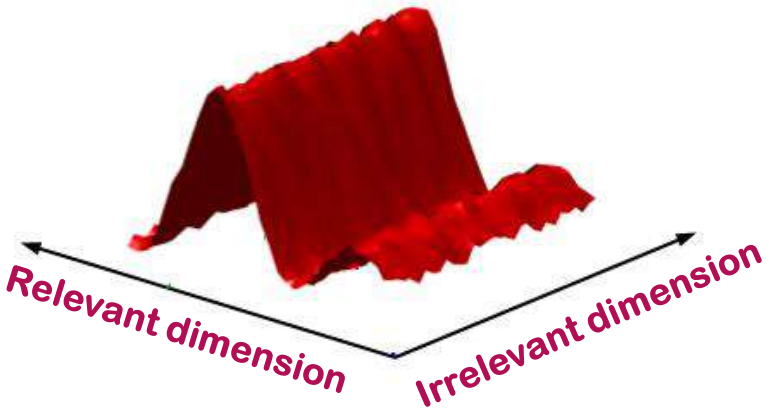
$$\min_{\widehat{T}} \max_{f_0, f_1} \ell_\theta(\widehat{T}) = \Theta\left(n^{-\left(1+\frac{1}{2}\left(\frac{d_0}{\alpha_0} \vee \frac{d_1}{\alpha_1}\right)\right)^{-1}}\right)$$

Characterizing response surfaces

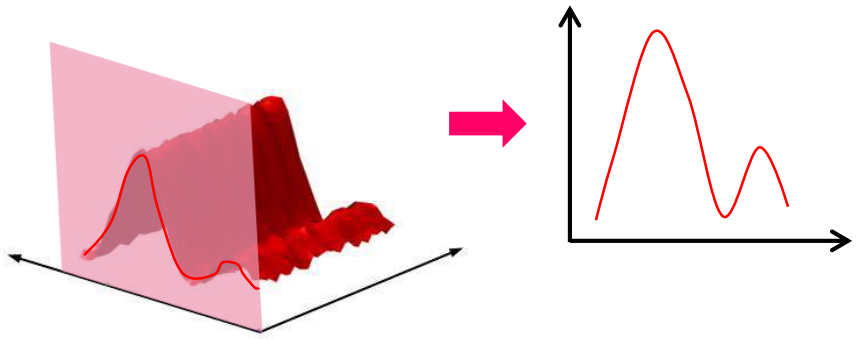
- **We prove that the minimax estimation loss:**
  - **Depends on the complexity of $f_0(x)$ and $f_1(x)$**

$f_0(x)$ **has** d_0 **relevant dimensions in a Hölder space** H^{α_0}
$f_1(x)$ **has** d_1 **relevant dimensions in a Hölder space** H^{α_1}

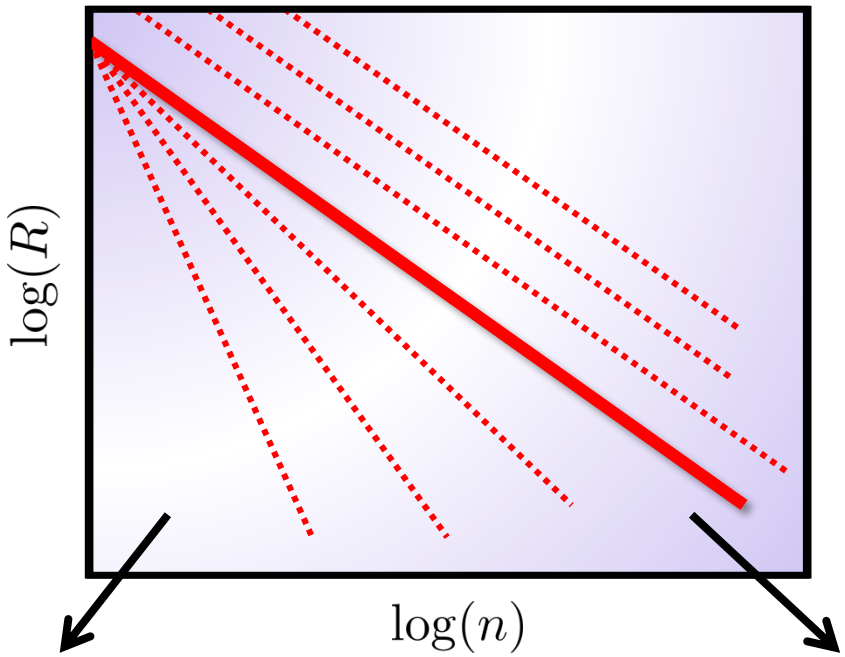
Sparsity d



Smoothness α



Theory – what have we learned?



Small sample regime

- Handling selection bias
- Sharing training data between response surfaces

Large sample regime

- ML model and hyperparameter tuning

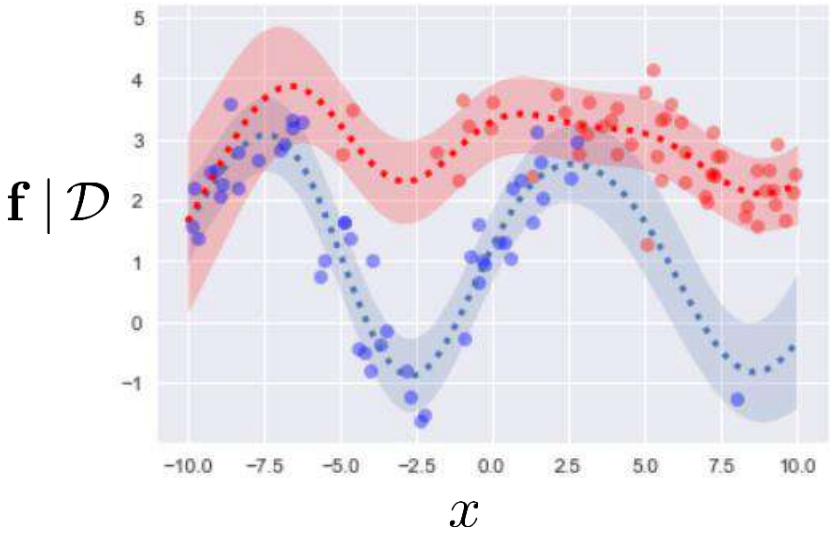
Multi-task Gaussian Processes [Alaa & van der Schaar, NIPS 2017]

- **Prior on vvRKHS = Multi-task Gaussian Process**

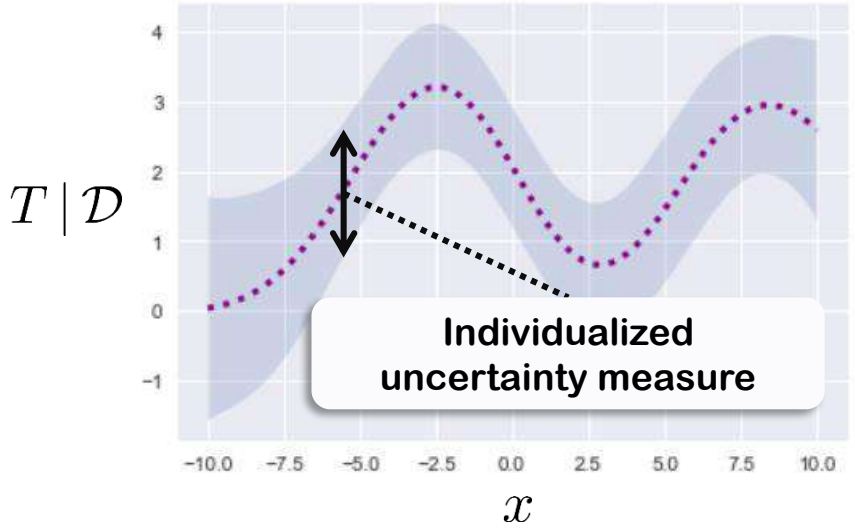
$$f_0, f_1 \sim \mathcal{GP}(0, \mathbf{K}_{\beta_0,\beta_1}) \qquad \text{Matern kernel = Prior over } H^{\beta_0} \times H^{\beta_1}$$

$$\mathbf{K}_\theta(x, x') = \mathbf{A}_0\, k_{\beta_0}(x, x') + \mathbf{A}_1\, k_{\beta_1}(x, x')$$

Posterior potential outcomes distribution

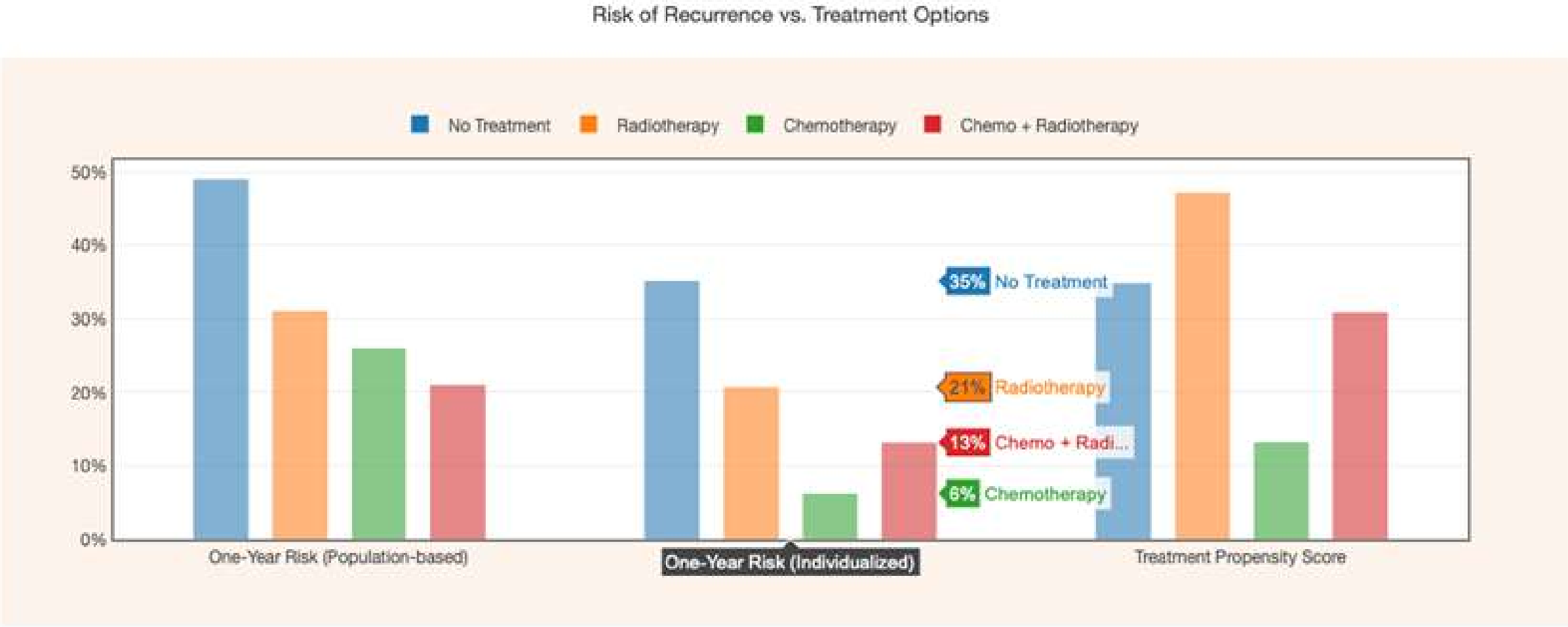


Posterior ITE distribution



Multiple Treatments: GANITE [Yoon, Jordon, vdS, ICLR 2018]

Estimation of Individualized Treatment Effects using Generative Adversarial Nets

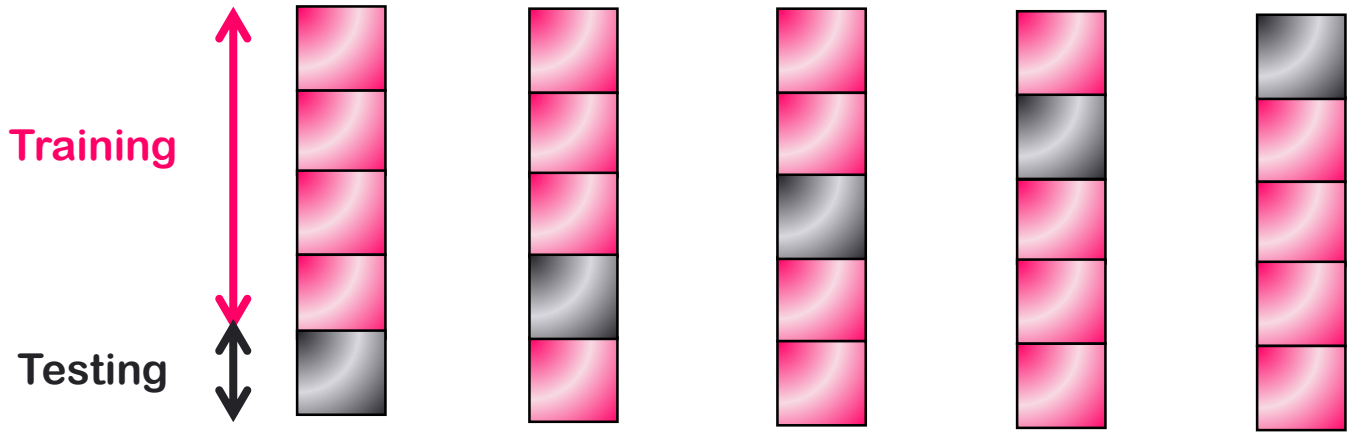


But how can we know how to select a model?

- Precision in estimating heterogeneous effects (PEHE) [Hill, 2011]

$$\ell_\theta(\widehat{T}) = \|T(X) - \widehat{T}(X)\|_\theta^2$$

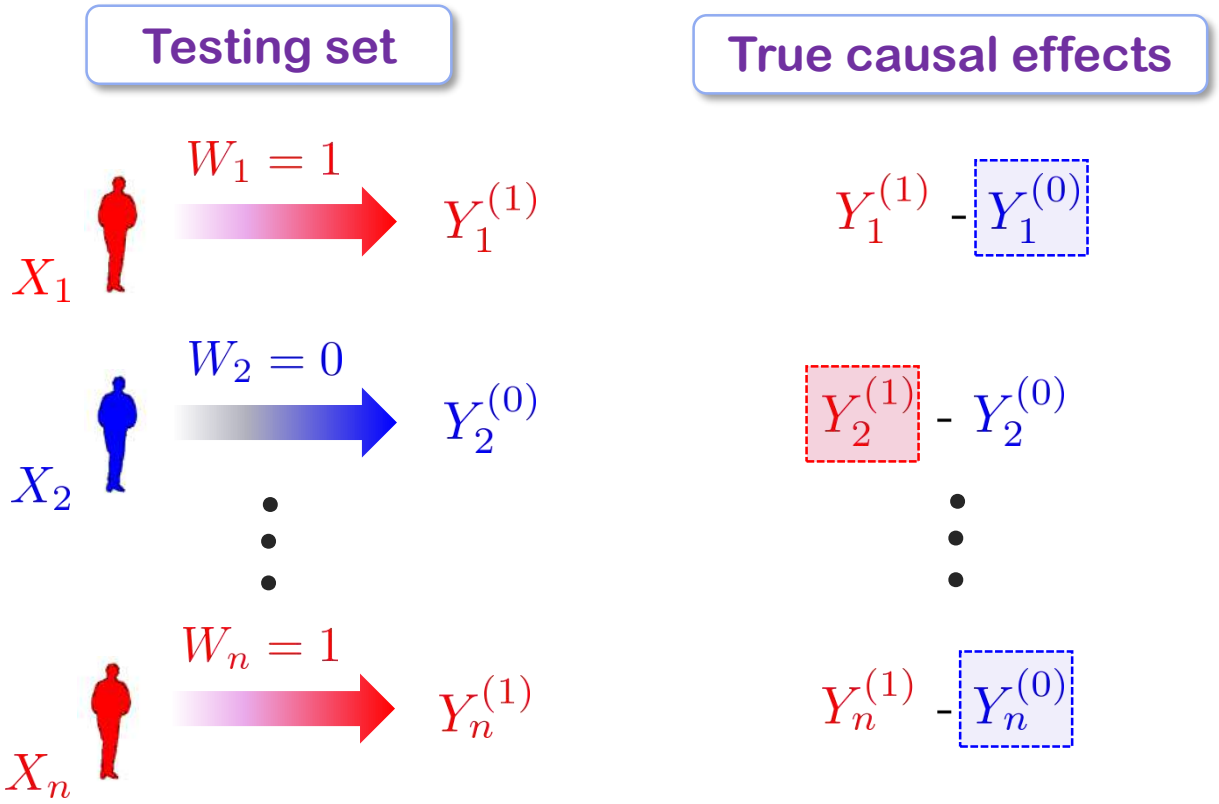
- Supervised learning → cross-validation!



$$\frac{1}{n}\sum_i(\widehat{T}(X_i) - (Y_i^{(1)} - Y_i^{(0)}))$$

Validating causal inference models

- **No explicit label:** cannot apply supervised cross-validation.



- **Goal:** developing a similar procedure for causal inference

Solution: Alaa and van der Schaar, ICML 2019

A performance metric is a statistical functional

- A functional is a function of a function.
- A statistical functional is a function of a distribution.

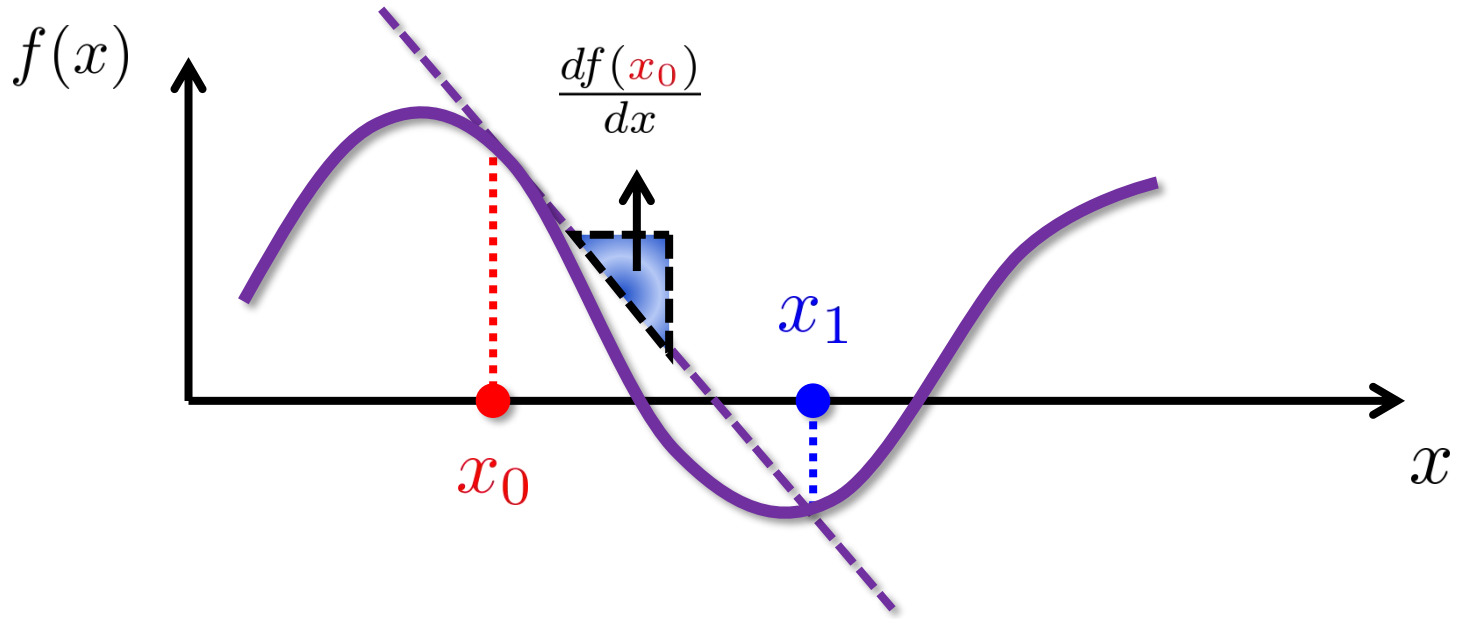
PEHE $\ell_\theta(\widehat{T}) = \|T(X) - \widehat{T}(X)\|_\theta^2$

Statistical functional $f(\mathbb{P}_\theta)$ $Z = (X, W, Y^{(0)}, Y^{(1)}) \sim \mathbb{P}_\theta$

Empirical measure $\frac{1}{n}\sum_{i=1}^n f(Z_i) = \frac{1}{n}\sum_i (\widehat{T}(X_i) - (Y_i^{(1)} - Y_i^{(0)}))$

Taylor series approximation

- The value of a function at a given input can be predicted using its value and (higher-order derivatives) at a proximal input.

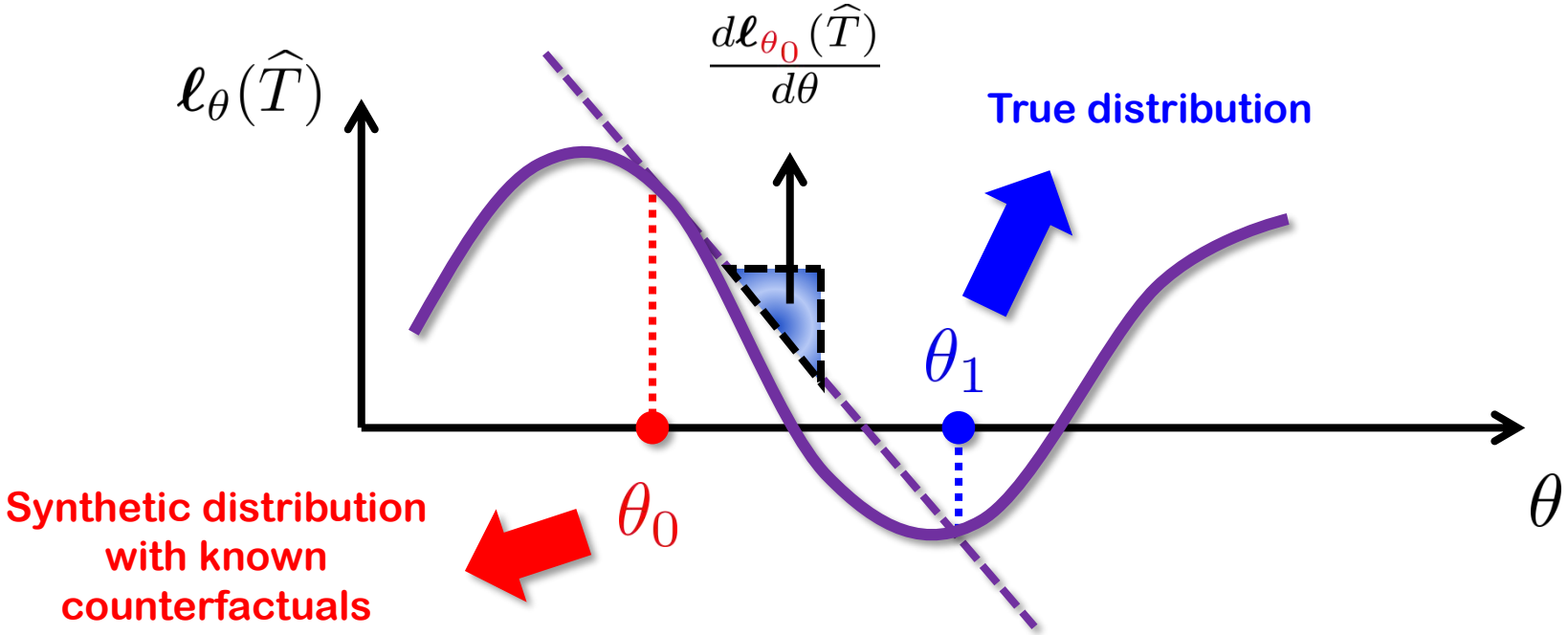


$$f(x_1) = f(x_0) + (x_1 - x_0)\frac{df(x_0)}{dx} + \frac{1}{2!}(x_1 - x_0)^2\frac{d^2f(x_0)}{dx^2} + \ldots$$

Analogy with Taylor series approximation

- The performance of a causal inference model is a functional of the data-generating distribution $\mathbb{P}_\theta$.

- Functional = a function of a function.



Functional calculus: von-Mises expansion (VME)

- A distributional analog of Taylor expansion [Fernholz, 1983]

$$\ell_{\theta_1}(\widehat{T}) = \ell_{\theta_0}(\widehat{T}) + \int \dot{\ell}^{(1)}_{\theta_0}(z;\widehat{T}) d(\mathbb{P}_{\theta_1} - \mathbb{P}_{\theta_0}) + \frac{1}{2!}\int \dot{\ell}^{(2)}_{\theta_0}(z;\widehat{T}) d(\mathbb{P}_{\theta_1} - \mathbb{P}_{\theta_0})^2 + \dots$$

- Influence functions ↔ Derivatives

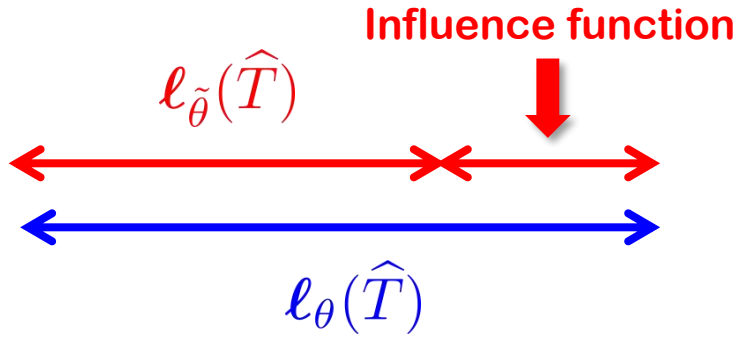
We can predict the performance of a causal inference model using the influence functions of its loss on a "similar" synthetic dataset.

Estimating a model's performance

- First-order "Taylor approximation"

Inaccessible empirical measure

True distribution



$$\ell_{\theta}(\widehat{T}) \approx \ell_{\tilde{\theta}}(\widehat{T}) + \int \dot{\ell}^{(1)}_{\tilde{\theta}}(z;\widehat{T}) d(\mathbb{P}_{\theta} - \mathbb{P}_{\tilde{\theta}})$$

Synthetic (Plug-in) distribution

Influence function

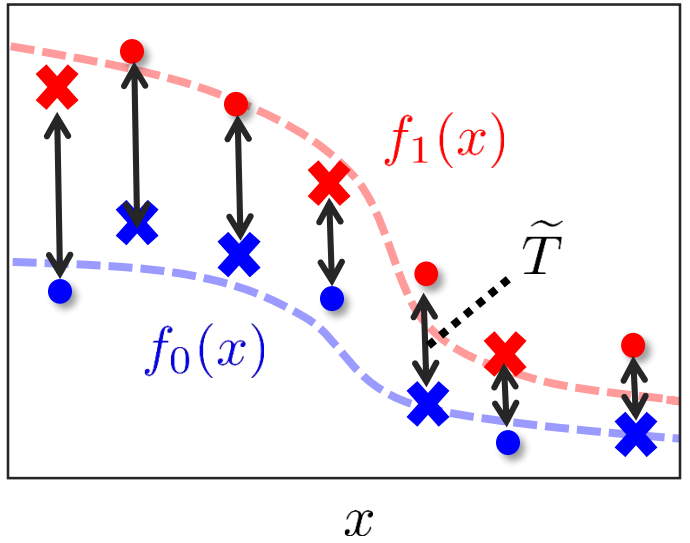
Accessible empirical measure

Estimating a model's performance

- **No need to simulate an entire observational dataset:** just synthesize counterfactuals!

Step 1: Plug-in estimation

- Plug-in model $\widetilde{T}$
- Plug-in PEHE loss $\ell_{\tilde{\theta}}(\widehat{T})$



Step 2: Bias correction

$$\ell_\theta(\widehat{T}) = \ell_{\tilde{\theta}}(\widehat{T}) + \int \dot{\ell}^{(1)}_{\tilde{\theta}}(z;\widehat{T}) d\mathbb{P}_\theta$$

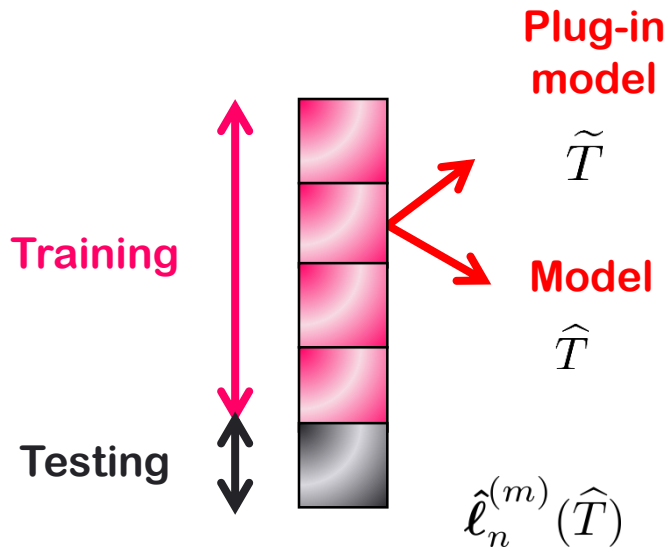
Consistency and efficiency

- **Theorem**

  Let $\hat{\ell}_n^{(m)}(\widehat{T})$ be an IF-based estimator using truncated *m*-term VME.

  $f_0(x)$ is in a Hölder space H^{α_0}

  $f_1(x)$ is in a Hölder space H^{α_1}



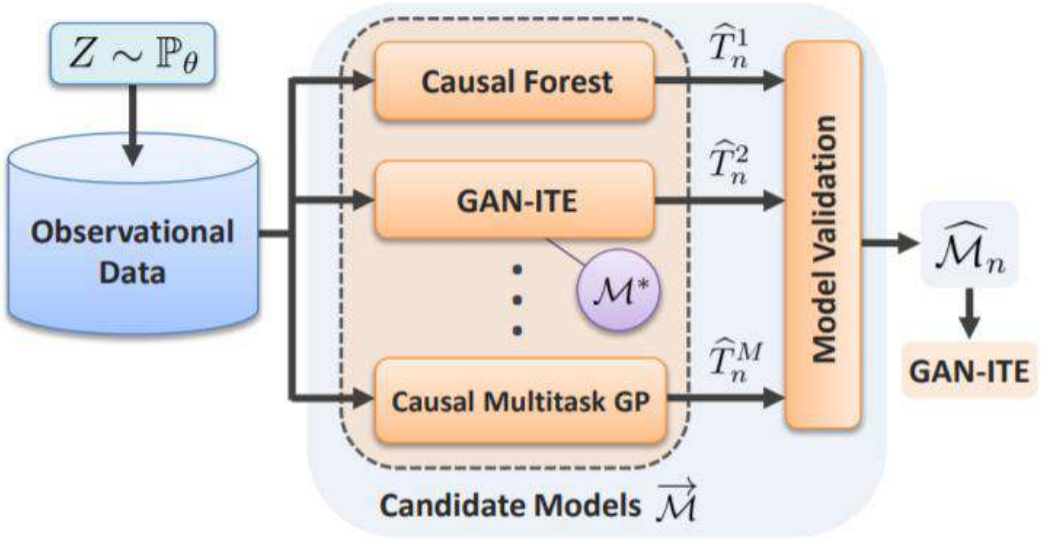
If plug-in model is minimax optimal:

$$\hat{\ell}_n^{(m)}(\widehat{T}) - \ell_\theta(\widehat{T}) = O\left(\frac{1}{\sqrt{n}} \vee n^{-\frac{2(\alpha_0 \wedge \alpha_1)(m+1)}{2(\alpha_0 \wedge \alpha_1)+d}}\right)$$

When enough number of VME included: $\sqrt{n}$ - consistent!

Automating causal inference!

- Selecting the right model for the right observational study.
- Collection of all models published in ICML, NeurIPS and ICLR between 2016 and 2018.



BNN	ICML 2016
CMGP	NIPS 2017
TARNet	ICML 2017
CFR Wass.	ICML 2017
CFR MMD	ICML 2017
NSGP	ICML 2018
GAN-ITE	ICLR 2018
SITE	NIPS 2018
BART	
Causal Forest	

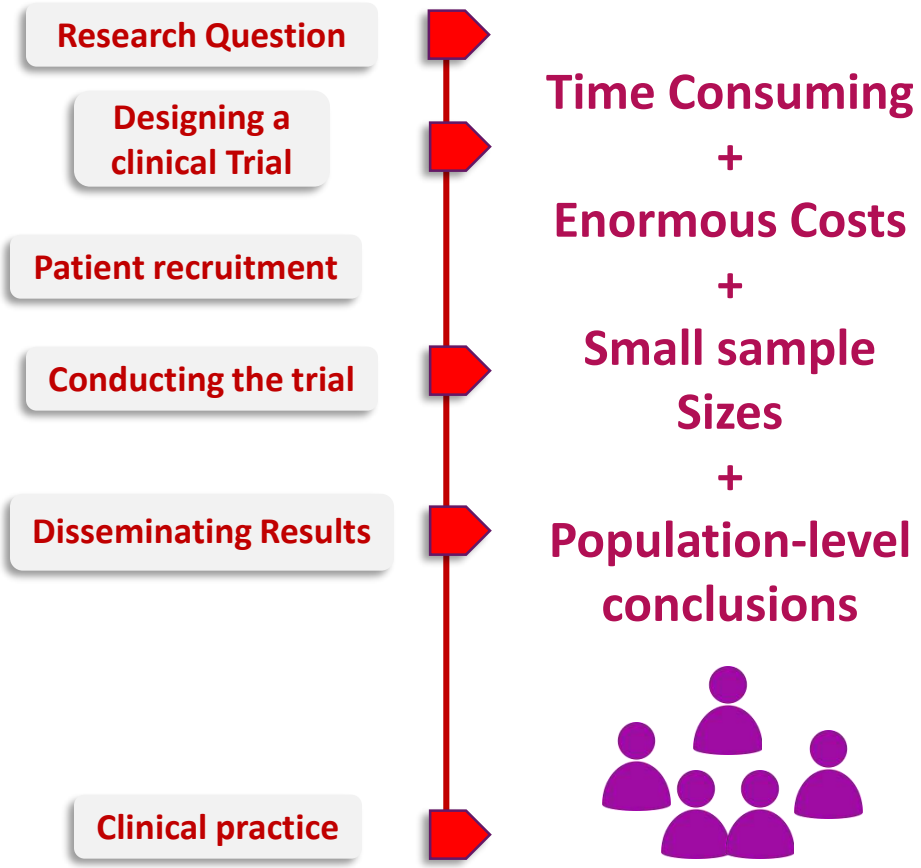
Results

- Average performance on the 77 benchmark datasets.
- No absolute winner on all datasets.
- IF-based selection is better than any single model.
- Factual selection is vulnerable to selection bias.

Method	% Winner
BNN	3%
CMGP	12%
NSGP	17%
TARNet	8%
CFR Wass.	9%
CFR MMD	12%
GAN-ITE	7%
SITE	7%
BART	15%
C. Forest	7%
Random	10%
Factual	53%
IF-based	72%
Supervised	84%

Machine Learning and Clinical Trials

Randomized Control Trials

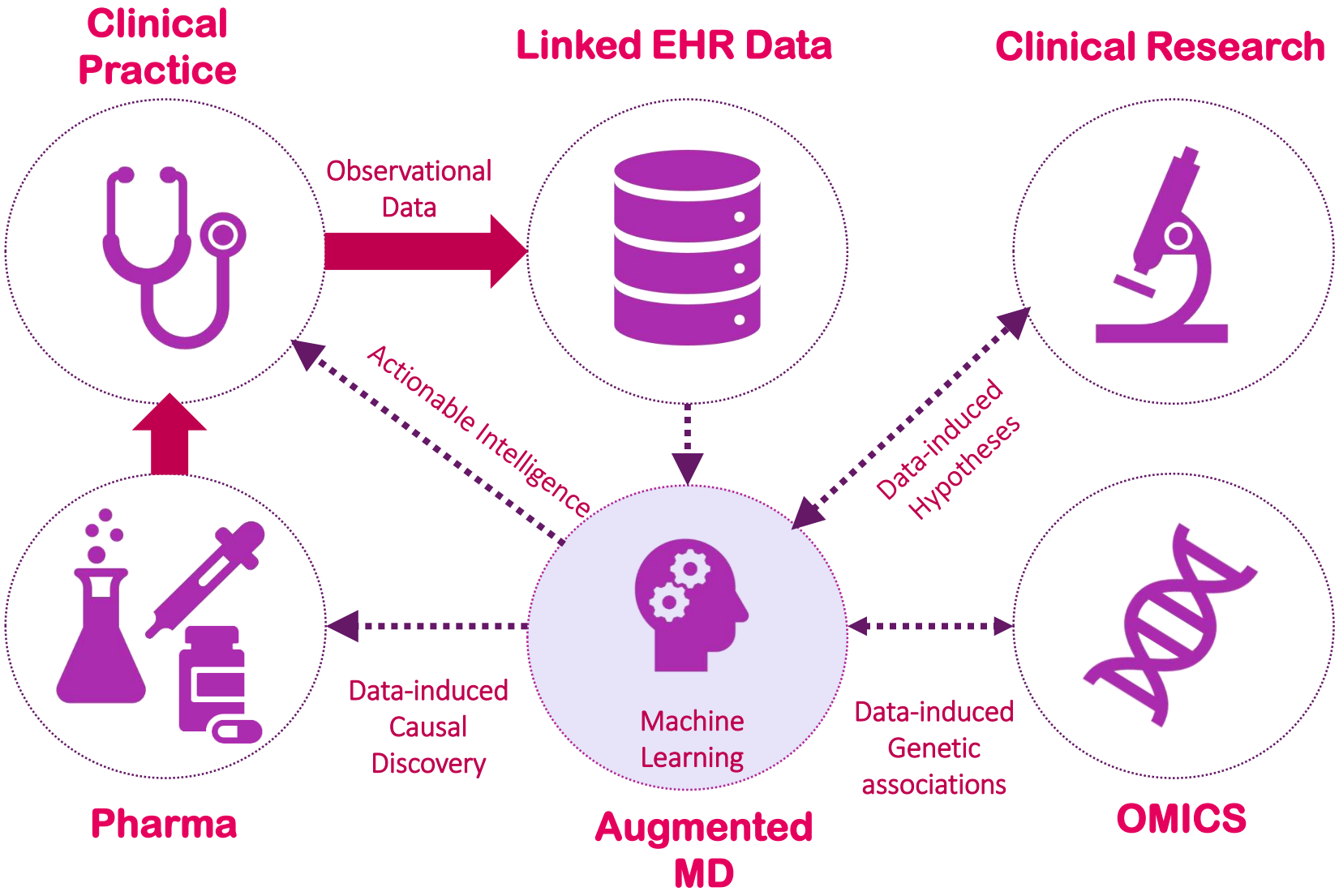


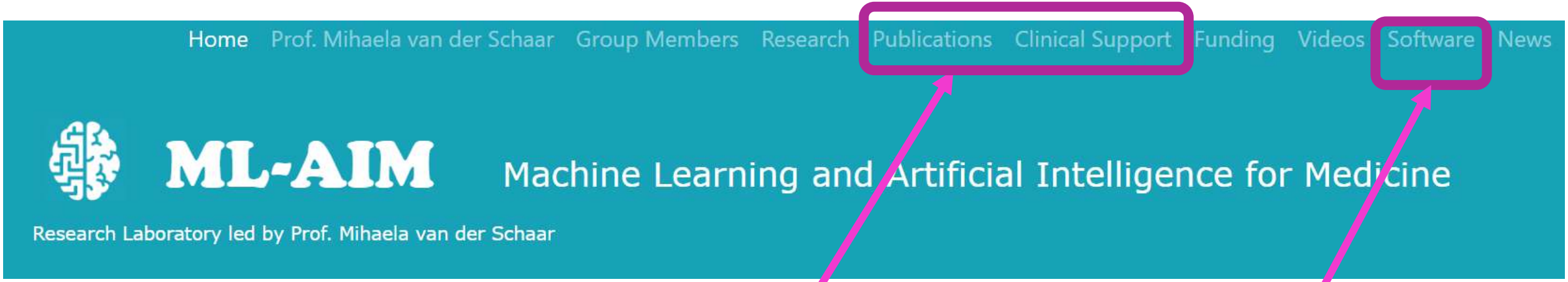
Machine Learning can Transform RCTs

- **Post-hoc subgroup analysis** for previously conducted clinical trials.
- **Recommender systems** for individualized treatment planning.
- **Designing clinical trials** for new drugs using data for similar drugs.

Patient-centric, cheap, big data, quick

Machine Learning & Medicine: Vision





Details about our algorithms:

http://www.vanderschaar-lab.com

Details about our software:

http://www.vanderschaar-lab.com