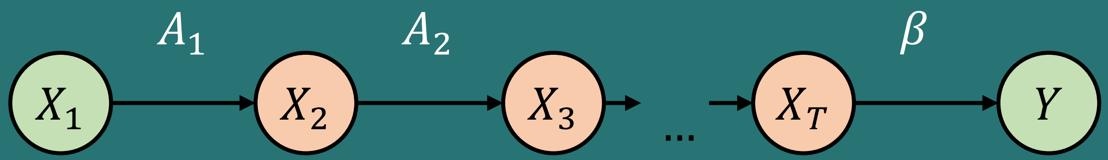
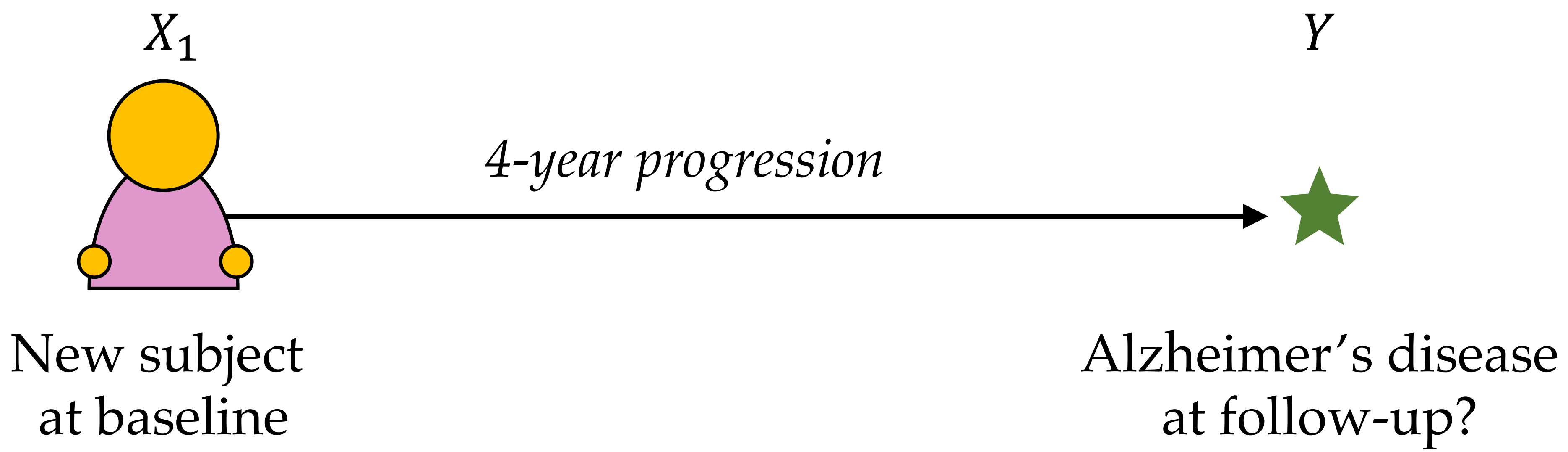


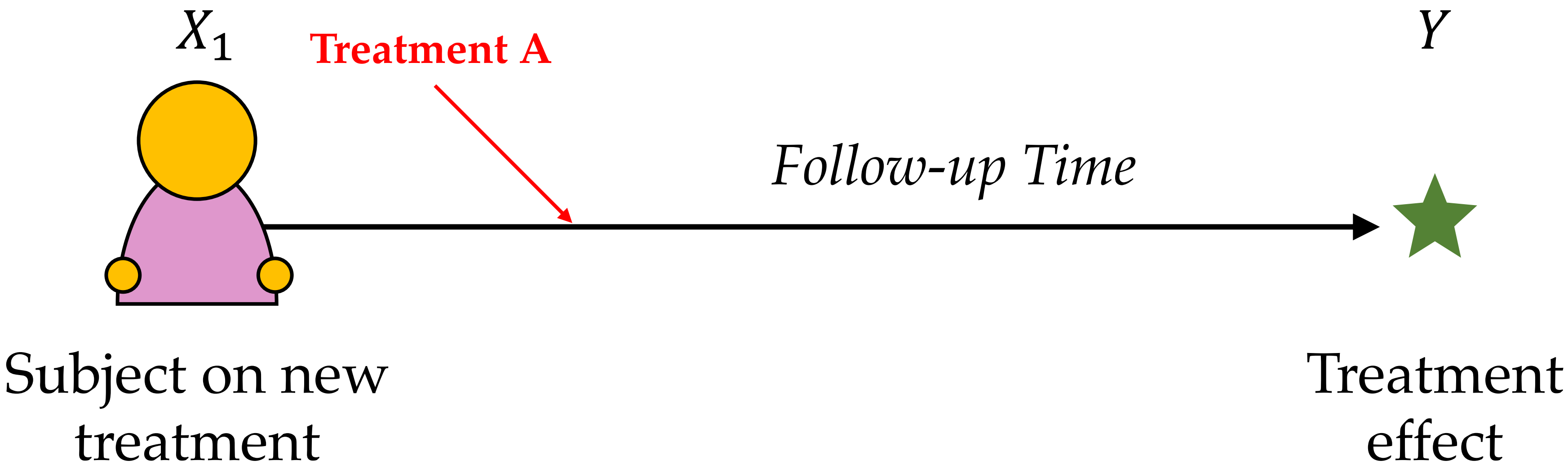
Efficient learning using privileged information with known causal structure

Fredrik D. Johansson
Sep 18, 2022

CHALMERS
UNIVERSITY OF TECHNOLOGY



Dansson, Stempfle, Egilsdottir, …, **J.**, *Alzheimer's Research & Therapy*, 2021

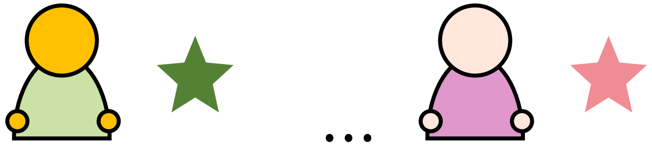

X_1
Treatment A
Follow-up Time
Y
Subject on new treatment
Treatment effect

We can minimize the *empirical prediction risk* over a data set D

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{R}_D(h), \qquad \hat{R}_D(h) \coloneqq \frac{1}{\boldsymbol{m}} \sum_{i=1}^{m} L\big(h\big(x_1^i\big), y^i\big)$$

Empirical risk

$D = \{$



$\}$

$$(x_1^1, y^1), \ldots, (x_1^m, y^m)$$

If $\boldsymbol{m}$ is large and drawn from p, $\hat{R}_D(h) \approx R(h) \coloneqq \mathbb{E}_p[L(h(X), Y)]$

Expected risk

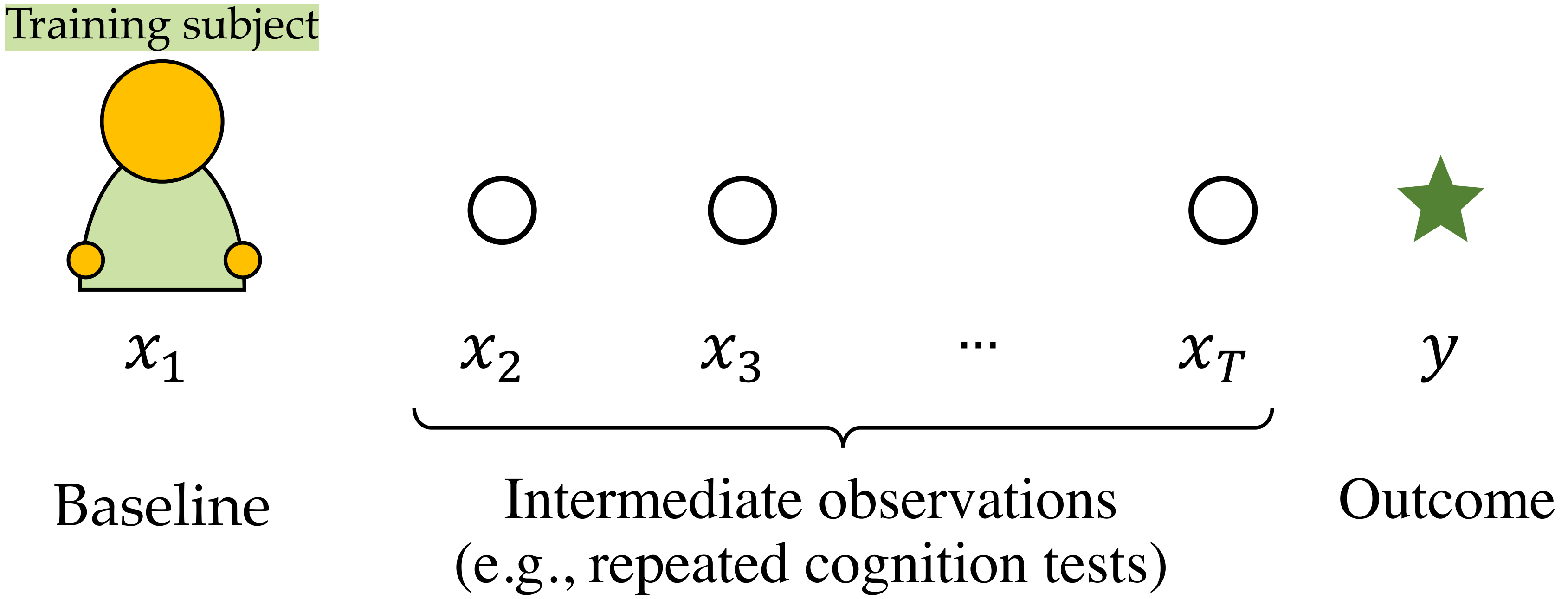
Test time

When we use $\hat{h}$ for new (test) subjects, we only have X_1

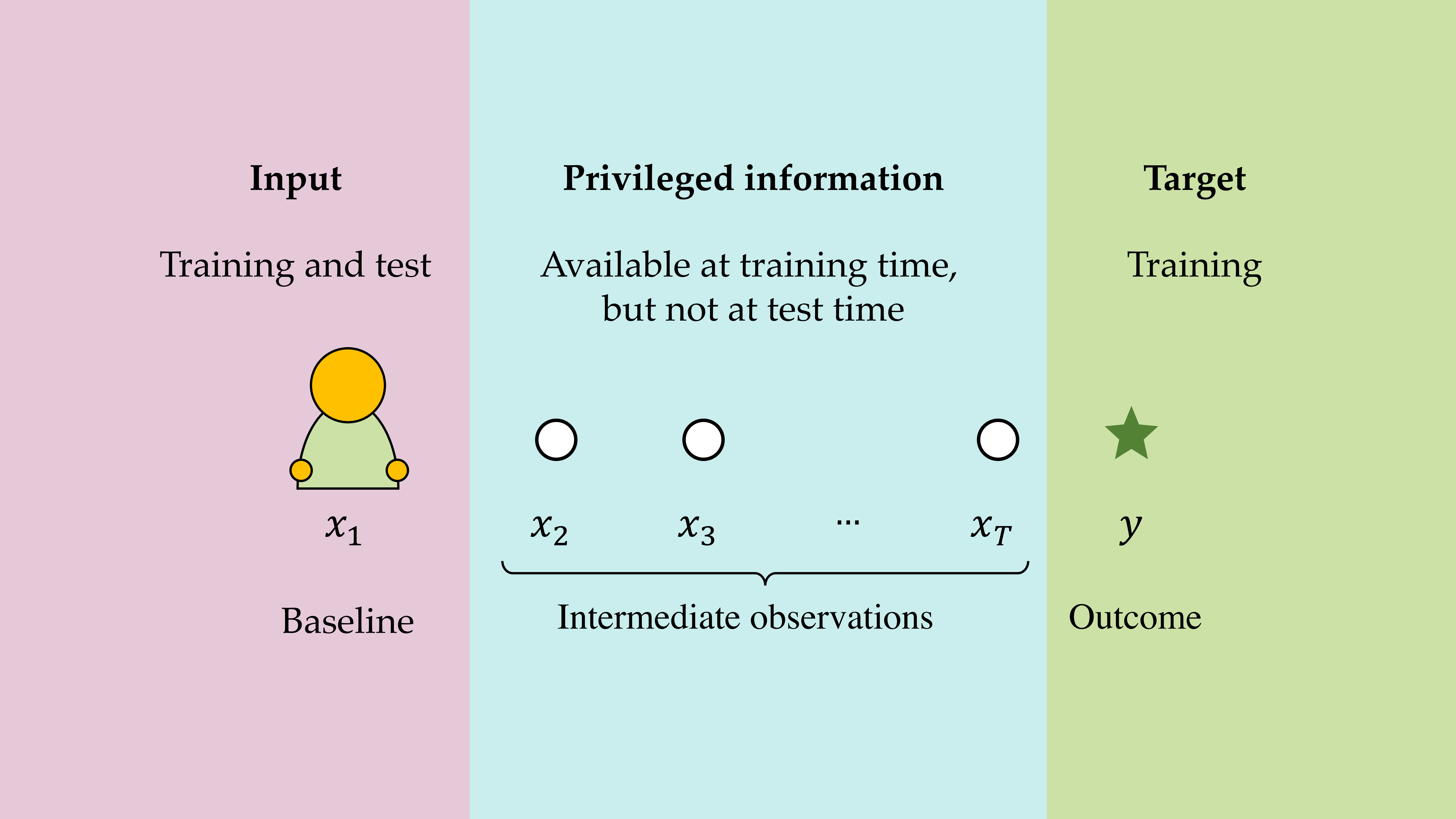
We want to predict their **future** progression based only on X_1

Training time

But we often know more about subjects in training data



* Tons of examples in healthcare and elsewhere: 30-day mortality prediction, user churn prediction, predicting crop yields

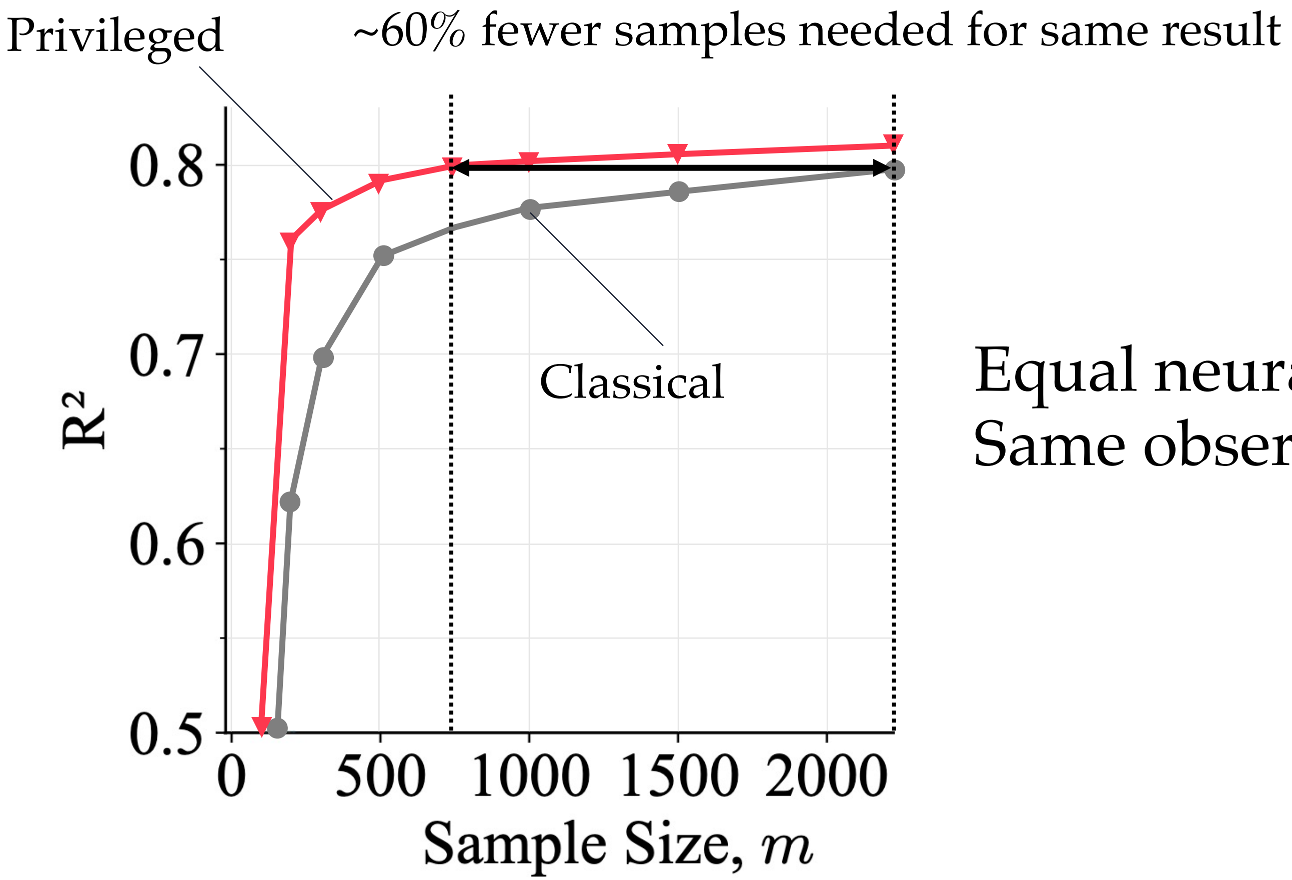

Input
Training and test
Privileged information
Available at training time, but not at test time
Target
Training
x_1
x_2
x_3
...
x_T
y
Baseline
Intermediate observations
Outcome

In standard ML, privileged information $X_2, \ldots, X_T$ is discarded

— Learning from baseline-outcome pairs $(x_1^1, y^1), \ldots, (x_1^m, y^m)$

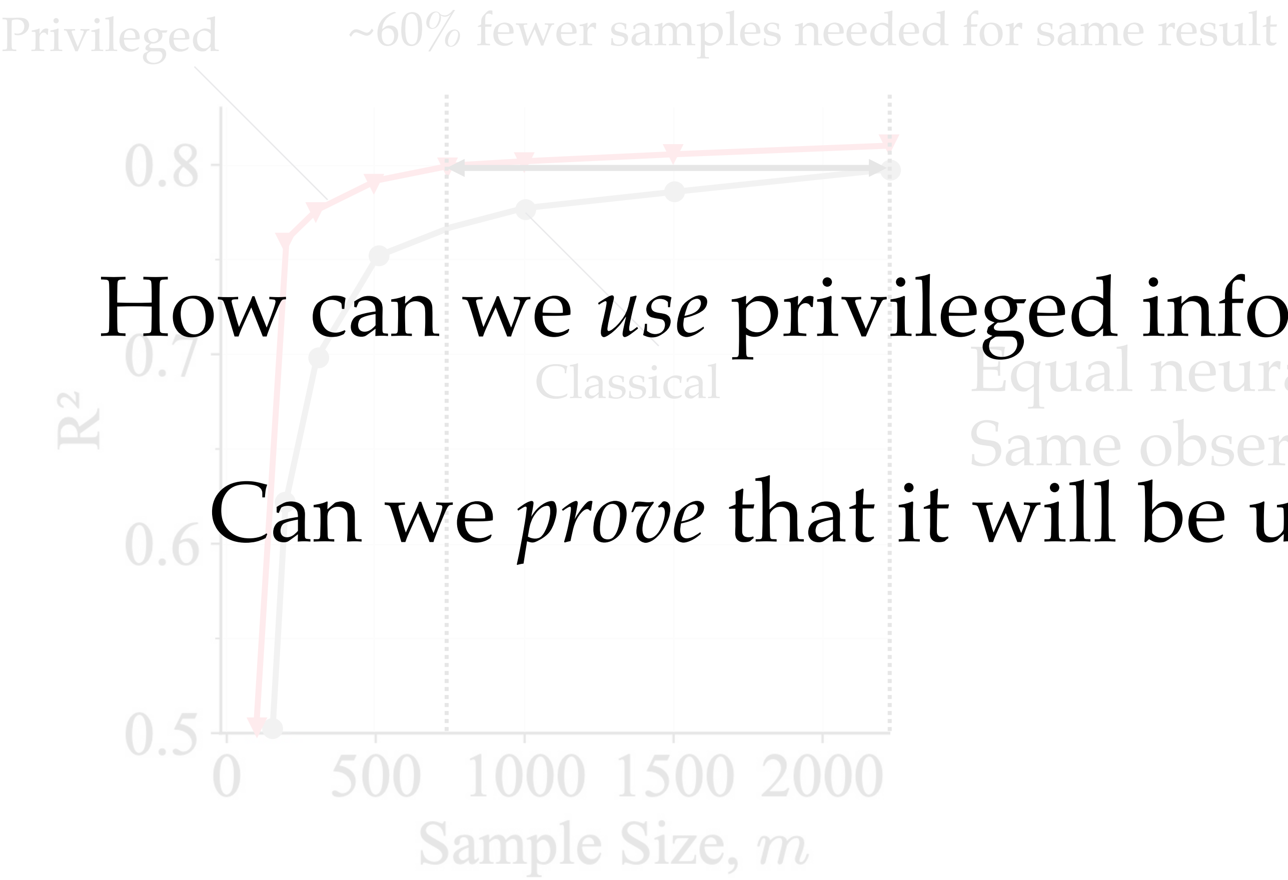
In fact, given enough samples, we can estimate f without PI…

PI is useful in data poor domains!



Equal neural architecture
Same observations w/wo PI

PI is useful in data poor domains!



Equal neural architecture

Same observations w/wo PI

How can we *use* privileged information?

Can we *prove* that it will be useful?

We measure *quality* of an algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ by its expected risk

$$\bar{R}(\mathcal{A}) \coloneqq \mathbb{E}_D\left[R\big(\mathcal{A}(D)\big)\right] \quad \text{where} \quad R(h) \coloneqq \mathbb{E}[L(h(X_1), Y)]$$

An **efficient** learner is one that, on average, outputs a hypothesis with smaller risk for the same number of samples $m = |D|$

We consider learners using two types of data sets

Classical learners $\mathcal{A}_{\mathrm{C}}$: (X_1^i, Y^i) — Only baseline time

Privileged learning $\mathcal{A}_{\mathrm{P}}$: $(X_1^i, \ldots, X_T^i, Y_i)$ — Entire time series

When can we prove that PI is **useful** for a fixed sample size?

$$\bar{R}(\mathcal{A}_{\mathrm{P}}) < \bar{R}(\mathcal{A}_{\mathrm{C}})?$$

Learning using privileged information

Pechyony & Vapnik[1] showed that there are cases where privileged information leads to learning rate improvements

$$\left|R(\mathcal{A}_{\mathrm{P}}) - \hat{R}(\mathcal{A}_{\mathrm{P}})\right| \leq O\left(\frac{1}{\boldsymbol{m}}\right) \quad \text{instead of.} \quad \left|R(\mathcal{A}_{\mathrm{C}}) - \hat{R}(\mathcal{A}_{\mathrm{C}})\right| = O\left(\frac{1}{\sqrt{\boldsymbol{m}}}\right)$$

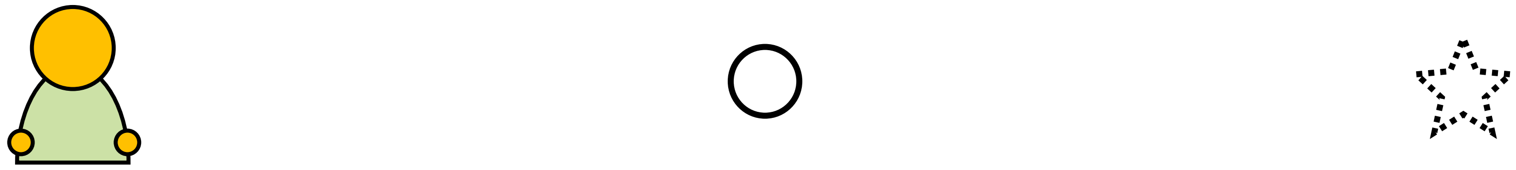
Fast rate *Slow rate*

However, the result is limited to a *highly specialized* data generating process and kicks in only when m is already *large*

[1]Pechyony & Vapnik, *NeurIPS*, 2010

Surrogate learning

Surrogate learning[1, 2] shows that surrogate outcomes (instead of Y) improve asymptotic efficiency when Y is sometimes missing



x_1 $\tilde{y}$ — Surrogate y — Sometimes missing

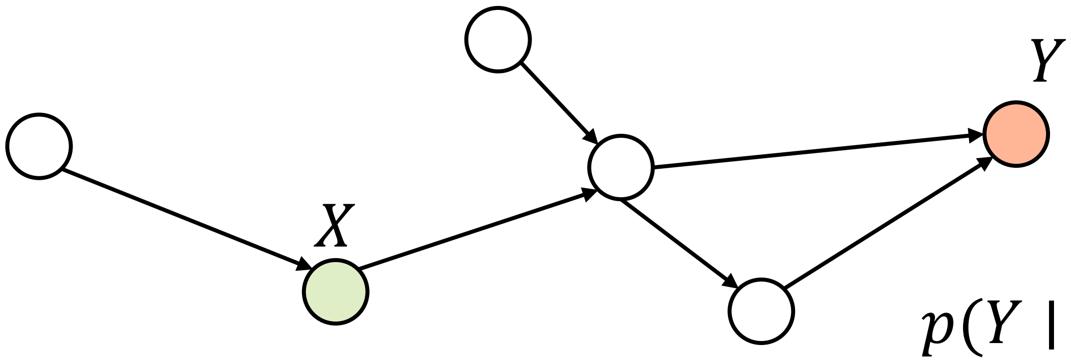
However, the results are uninformative when Y is observed as often as the surrogates (privileged information)*

* *The theory was not developed for our setting*

[1]Kallus & Mao, *On the role of surrogates…*, 2020, [2]Athey et al., *The Surrogate Index*, 2019

Causal effect estimation

Guo & Perkovic[1] showed that *recursive* least-squares is the most efficient regular estimator of **total causal effects** in linear SEMs



$p(Y \mid do(X = x))$?

Using "post-treatment" variables ⇒ higher asymptotic efficiency

[1]Guo & Perkovic, *JMLR*, 2022

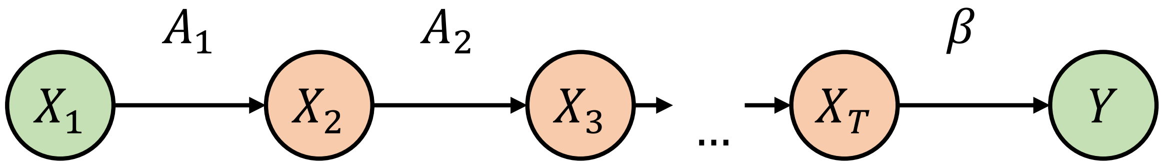
Assumption on causal structure

The privileged information is Markov

Setting 1. The DGP is a linear-Gaussian chain

$$X_{t+1} = X_t A_t + \epsilon_{t+1} \quad \text{where} \quad \epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2 I)$$
$$Y = X_T^\top \beta + \epsilon_Y \quad \text{where} \quad \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$$

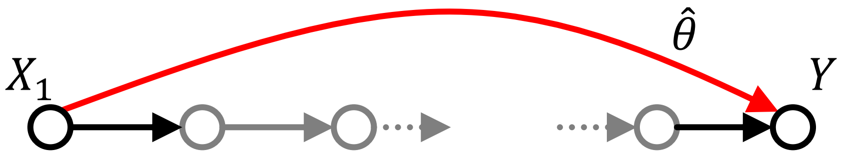
No assumption on X_1!



We don't assume stationarity. I.e., $A_t \neq A_{t'}$ in general.

Classical learner

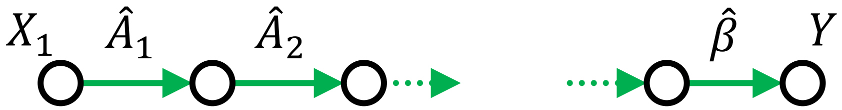
Single-step prediction



$$\mathcal{A}_{\mathrm{C}}(D) = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{Y}$$

Privileged learner

Every-step prediction



$$\mathcal{A}_{\mathrm{P}}(D) = \hat{A}_1 \cdots \hat{A}_{T-1} \hat{\beta}$$

1. Both estimators return linear regressions of X_1
2. Both are unbiased estimators of $\mathbb{E}[Y \mid X_1] = (A_1 \cdots A_{T-1}\beta)^\top X_1$
3. The only difference is variance—sample efficiency

Theorem 1 (informal). Assume that $X_1, \dots, X_T, Y$ is a linear-Gaussian chain with isotropic noise. Then,

$$\bar{R}(\mathcal{A}_{\mathrm{P}}) \leq \bar{R}(\mathcal{A}_{\mathrm{C}}) - \mathbb{E}_{\hat{h}_{\mathrm{P}}, X_1}\left[\mathrm{Var}_D\left(\hat{h}_{\mathrm{C}}(X_1) \mid \hat{h}_{\mathrm{P}}\right)\right]$$

Remaining variance in $\hat{h}_C$ when $\hat{h}_P$ is fixed

Since $\mathrm{Var}(\cdot) \geq 0$, learning using privileged information is never worse under the conditions of Theorem 1.

*The variance in the classical estimator is larger—despite the privileged learner fitting $(T-1)d^2 + d$ parameters!

Karlsson, Willbo, Hussein, Krishnan, Sontag, **J.** AISTATS, 2022

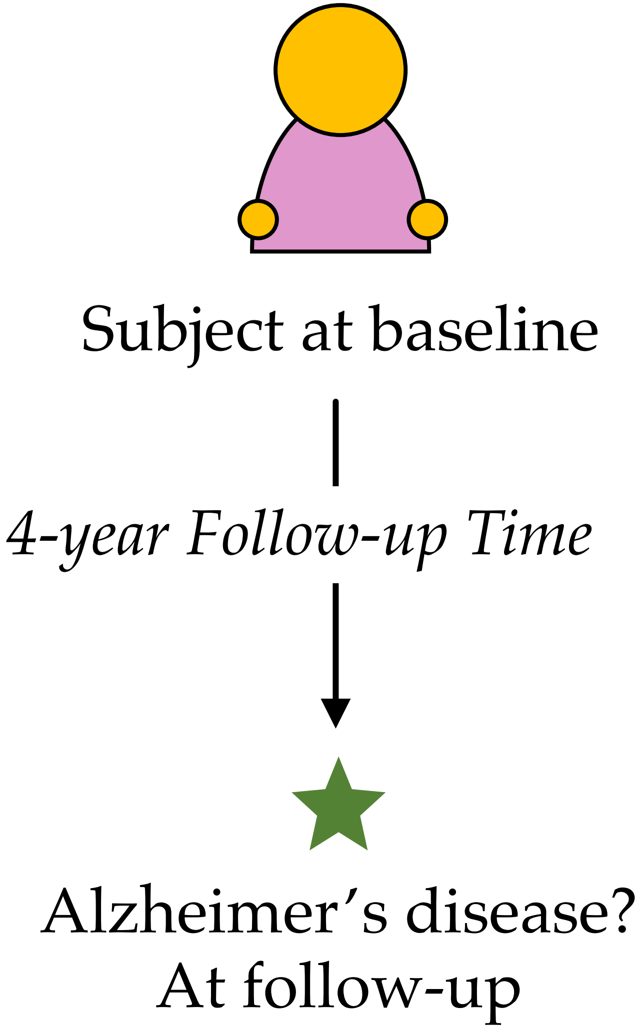
Key step

The privileged learner is both sufficient statistic T *and* estimator δ_1

Lemma. For $\hat{\theta}_{\mathrm{C}}$ and $\hat{\theta}_{\mathrm{P}} = \hat{A}_1 \cdots \hat{A}_T \hat{\beta}$ the classical and privileged estimators, respectively, it holds that

$$\mathbb{E}_D\Big[\underset{\delta}{\hat{\theta}_{\mathrm{C}}} \mid \underbrace{\hat{A}_1, \dots, \hat{A}_T, \hat{\beta}}_{T(D)}\Big] = \hat{A}_1 \cdots \hat{A}_T \hat{\beta} = \underset{\delta_1}{\hat{\theta}_{\mathrm{P}}}$$

An example of a Rao-Blackwell technique



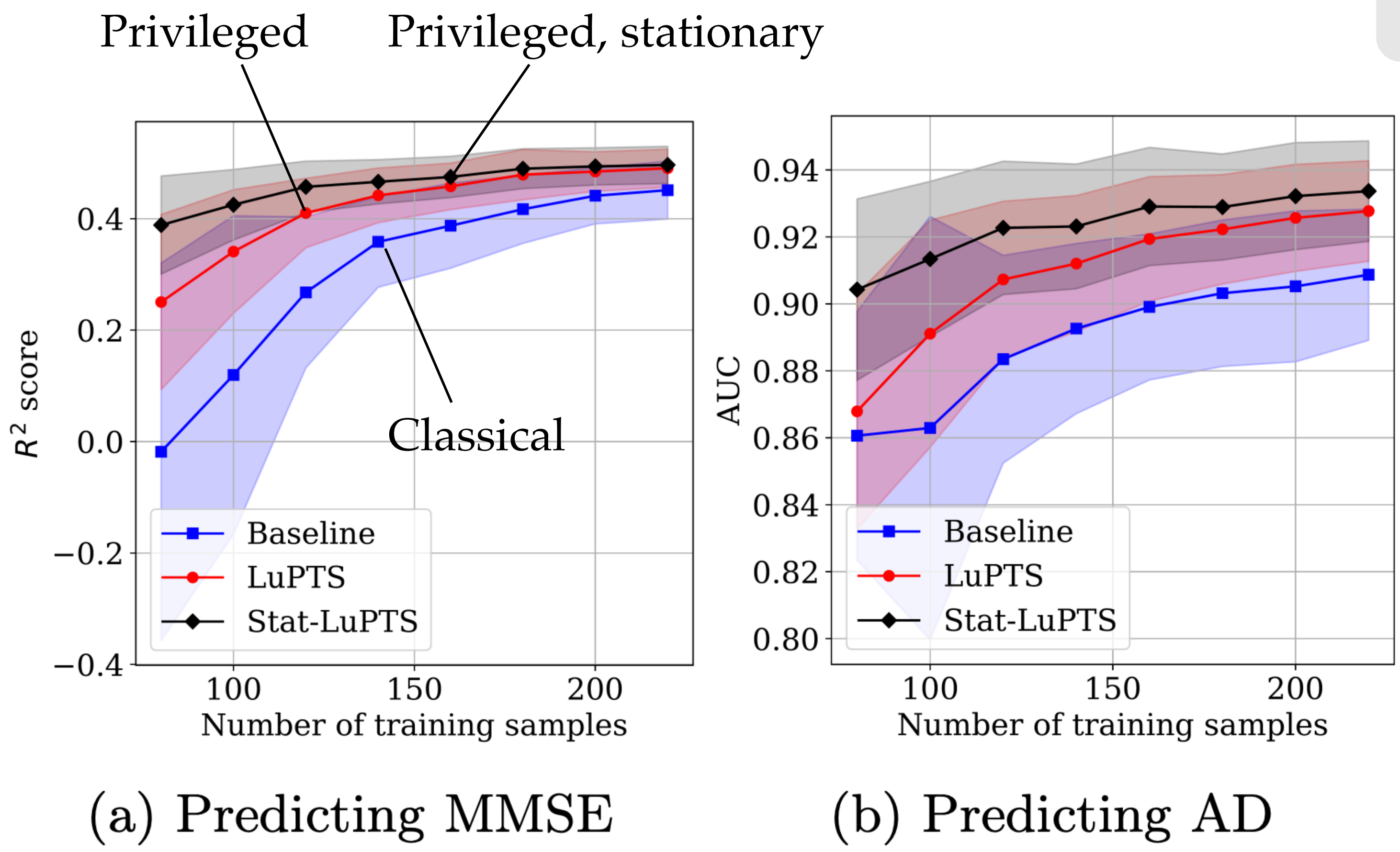
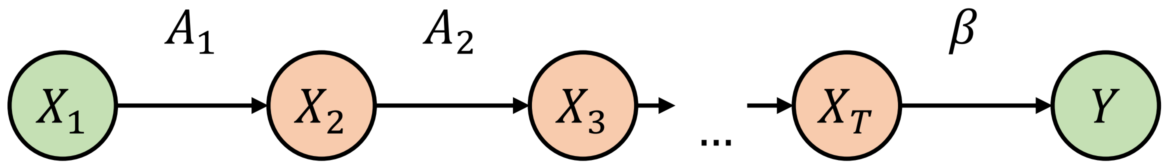


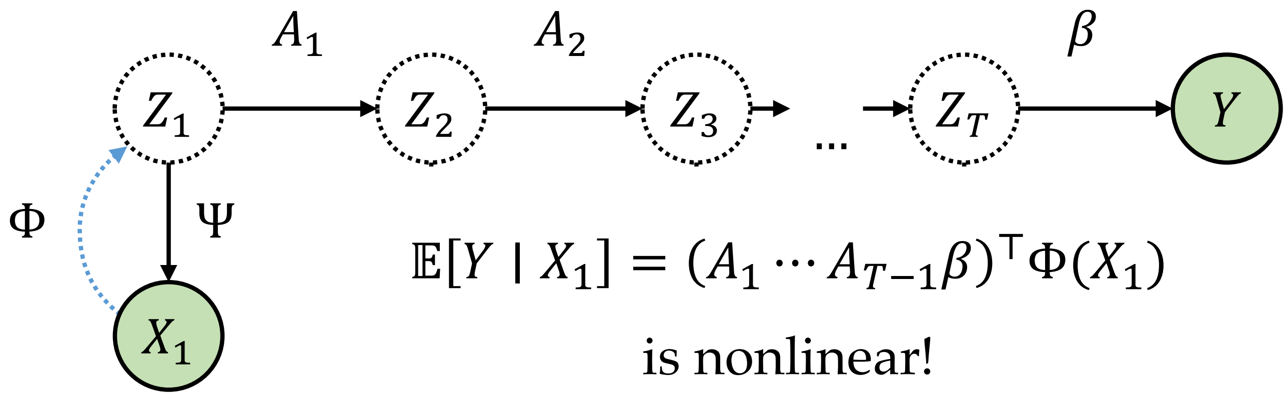
Figure 4: **Alzheimer's disease progression tasks**. Follow-up at 12, 24 and 36 months after baseline as privileged information. Metrics used are R^2/AUC; shaded region corresponds to one standard deviation across 100 iterations.

Nonlinearity through latent dynamics

Setting 1: Linear-Gaussian



Setting 2: Latent system, nonlinear emissions:



$$\mathbb{E}[Y \mid X_1] = (A_1 \cdots A_{T-1}\beta)^\top \Phi(X_1)$$

is nonlinear!

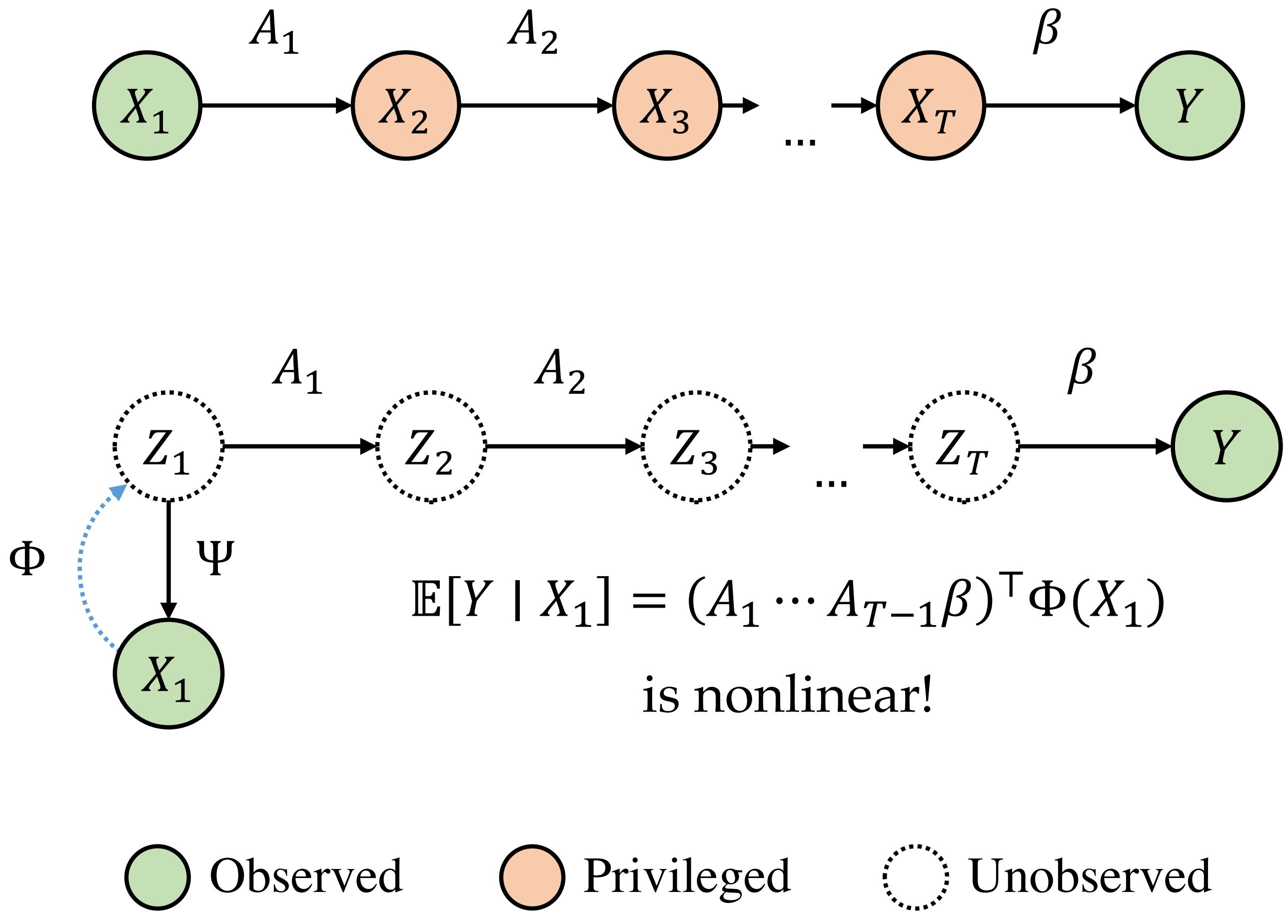
* Jung & Johansson, *Accepted to NeurIPS*, 2022

Nonlinearity through latent dynamics

Setting 1: Linear-Gaussian

Setting 2: Latent system, nonlinear emissions:

$Z_1 \dots Z_T$ is a linear-Gaussian system like before. Now, only observed through nonlinear $X_t = \Psi(X_t)$



$$\mathbb{E}[Y \mid X_1] = (A_1 \cdots A_{T-1}\beta)^\top \Phi(X_1)$$

is nonlinear!

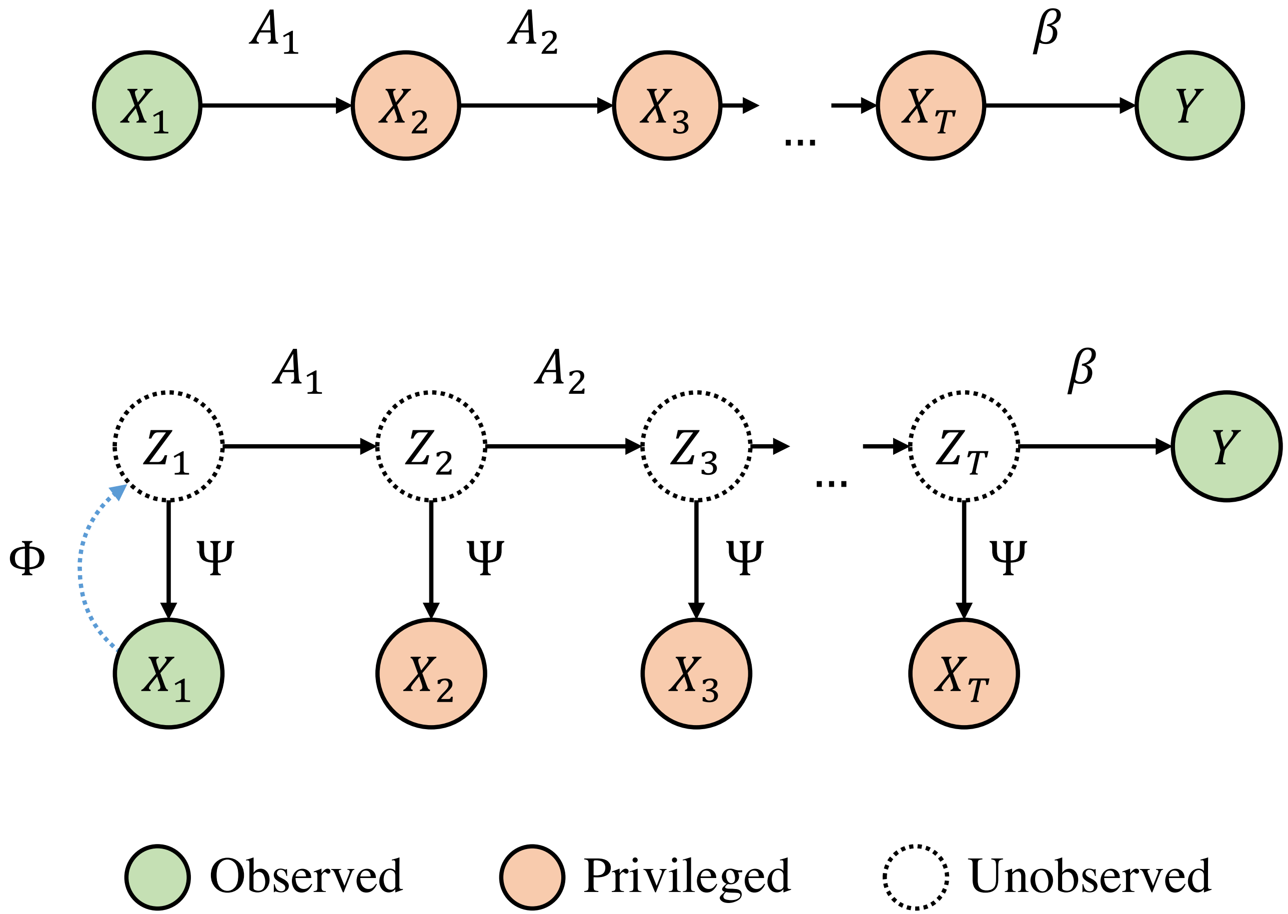
* Jung & Johansson, *Accepted to NeurIPS*, 2022

Nonlinearity through latent dynamics

Setting 1: Linear-Gaussian

Setting 2: Latent system, nonlinear emissions:

$Z_1 \dots Z_T$ is a linear-Gaussian system like before. Now, only observed through nonlinear $X_t = \Psi(X_t)$

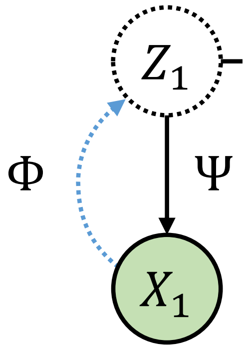


* Jung & Johansson, *Accepted to NeurIPS*, 2022

Latent linear-Gaussian system

Theorem 2 (informal). Assume that $Z_1, \ldots, Z_T, Y$ is an isotropic linear-Gaussian chain and $X_t = \Psi(Z_t)$ with **$\boldsymbol{\Phi} = \boldsymbol{\Psi}^{-1}$ known up to linear transform**, explicitly or as a kernel $k(x, x') = \langle \phi, \phi' \rangle$. Then,

$$\bar{R}(\mathcal{A}_{\mathrm{P}}) \leq \bar{R}(\mathcal{A}_{\mathrm{C}}) - \mathbb{E}_{\hat{h}_{\mathrm{P}}, X_1}\left[\mathrm{Var}_D\left(\hat{h}_{\mathrm{C}}(X_1) \mid \hat{h}_{\mathrm{P}}\right)\right]$$



1. Implication is the same as before, but the setup is generalized
2. Limited to partial knowledge of Φ

* Jung & Johansson, *Accepted to NeurIPS*, 2022

Random feature representations

If the representation Φ is unknown, *random feature embeddings* can be used consistently for both classical and privilieged learners.

Random features are $\widehat{\Phi} = \sigma(WX)$ w. nonlinearity σ, random W.

— For example, random single-layer ReLU NN

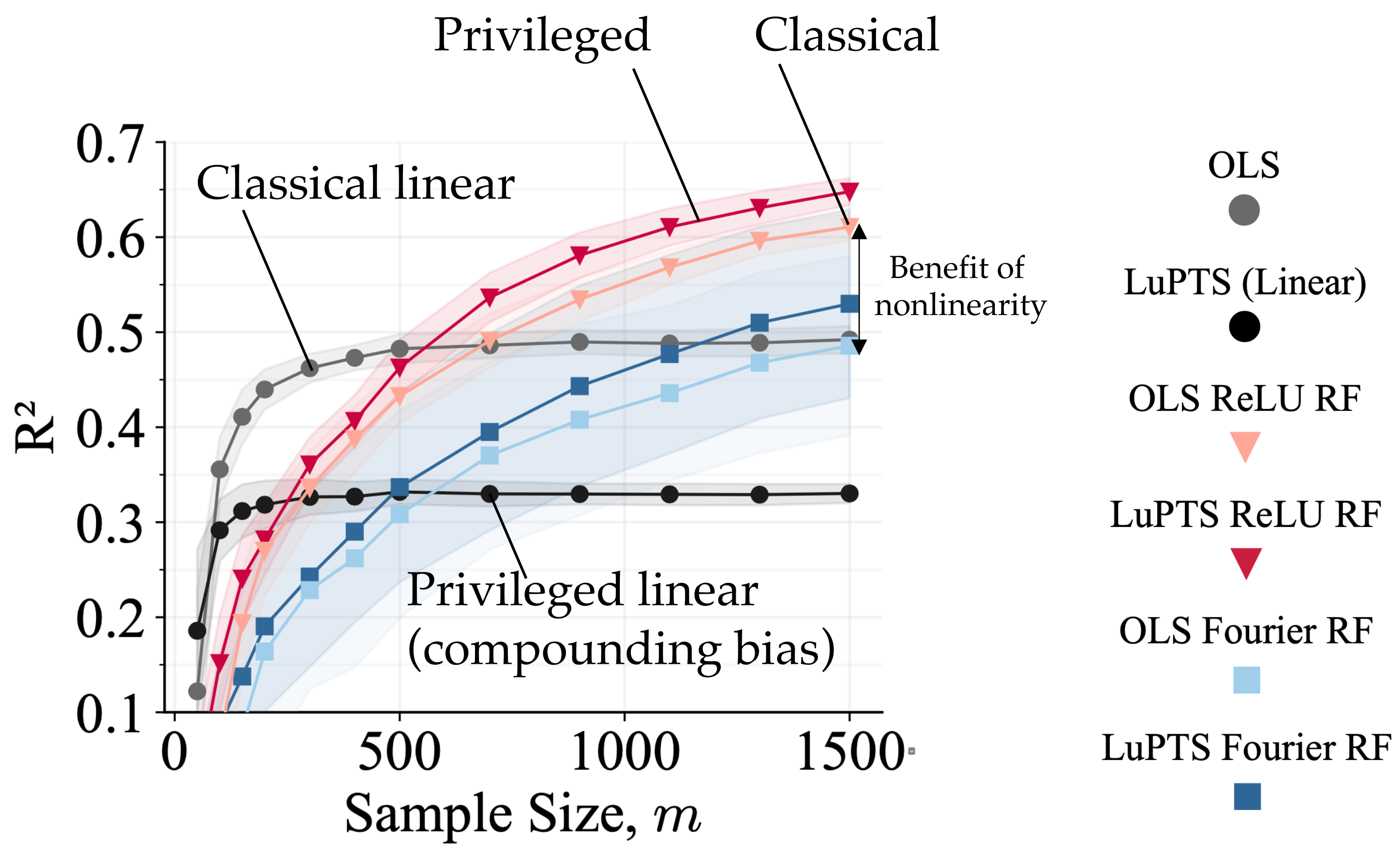
Proposition. if either learner uses $\hat{d}$ random features for $\widehat{\Phi}$

$$\mathcal{A}_{\mathrm{P/C}}(D) \to f \quad \text{as} \quad m > \hat{d} \to \infty$$

Random feature regression*

We see both benefits of nonlinearity and of privileged information

For linear estimators, bias compounds in the privileged learner

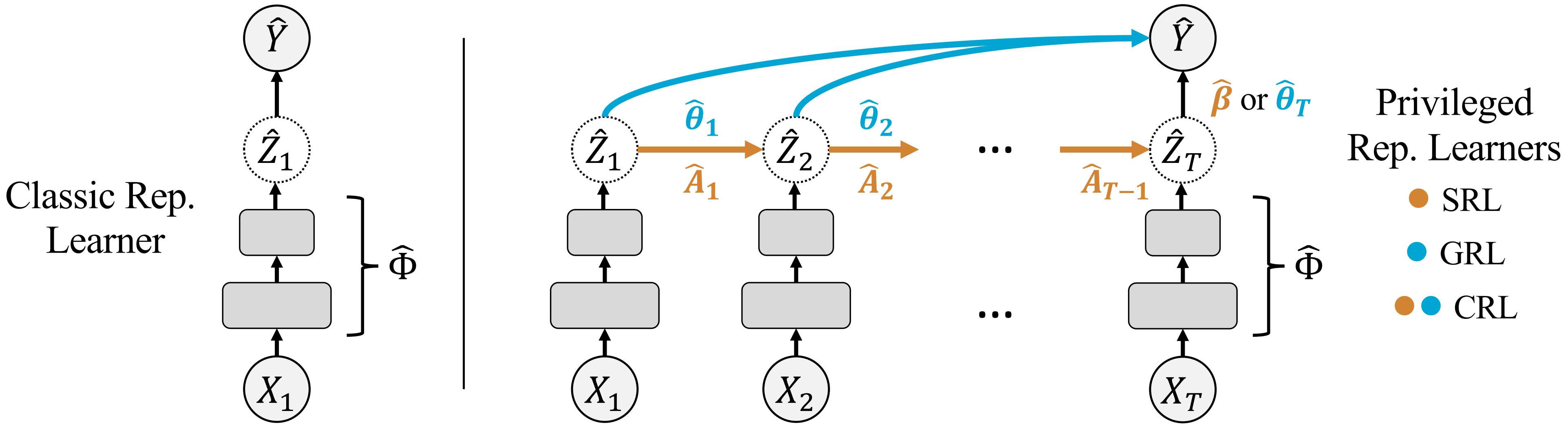


(b) **Square-Sign**, $T = 5$, $d = 10$.

Map all X_t using random ReLU/Fourier features, fit OLS estimators as before

Privileged representation learners

We can construct multiple representation learning architectures $\hat{\Phi}$



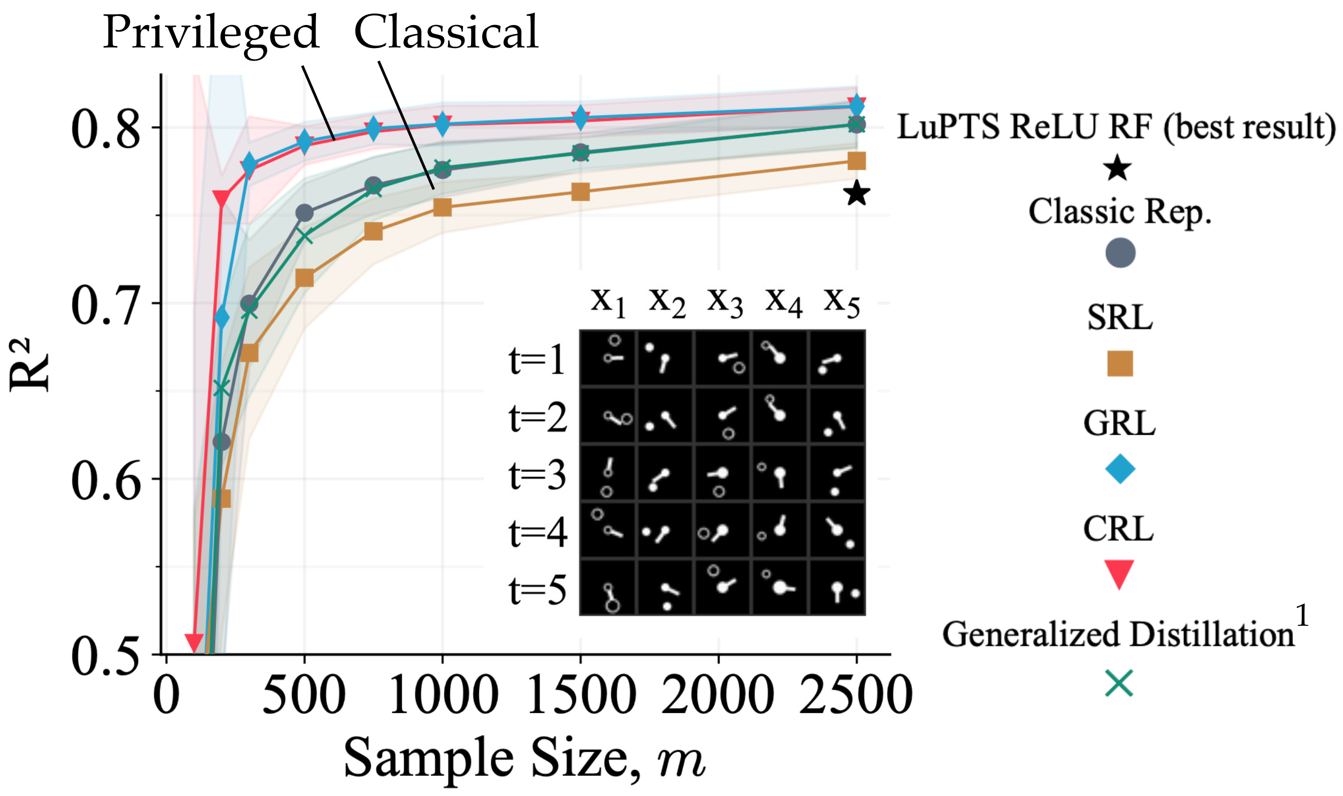
Neural network regression

Data generated from a latent linear-Gaussian system,

$Z_1, \dots, Z_T \in \mathbb{R}^2$.

Observed variables X are image representations of 2D coords., like "clock faces"

Outcome is linear in Z



(c) **Clocks-LGS**, $T = 6$, $q = 1$.

[1]Lopez-Paz, Bottou, Schölkopf, *abs/1511.03643*, 2015

Take-aways

- Preference for privileged learners is independent of sample size (finite regime) — the gap varies with m
- Privileged information explains part of the variance in Y
- Random features and learned representations both perform better empirically *with* privileged information

Open questions

- Results in the biased / regularized case?
  - (E.g., finite sample random features)
- Causal structures beyond chain graphs (arbitrary DAGs?)
- Finite-sample preference guarantees beyond Rao-Blackwell
  - The theorem requires being able to characterize the predictions made by $\mathcal{A}_\text{C}$ conditioned on $\mathcal{A}_\text{P}$. Possible for OLS but not in general

Fredrik D. Johansson

fredrik.johansson@chalmers.se

These slides presented joint work with

Bastian Jung, Rickard Karlsson, Martin Willbo,
Zeshan Hussain, Rahul Krishnan and David Sontag