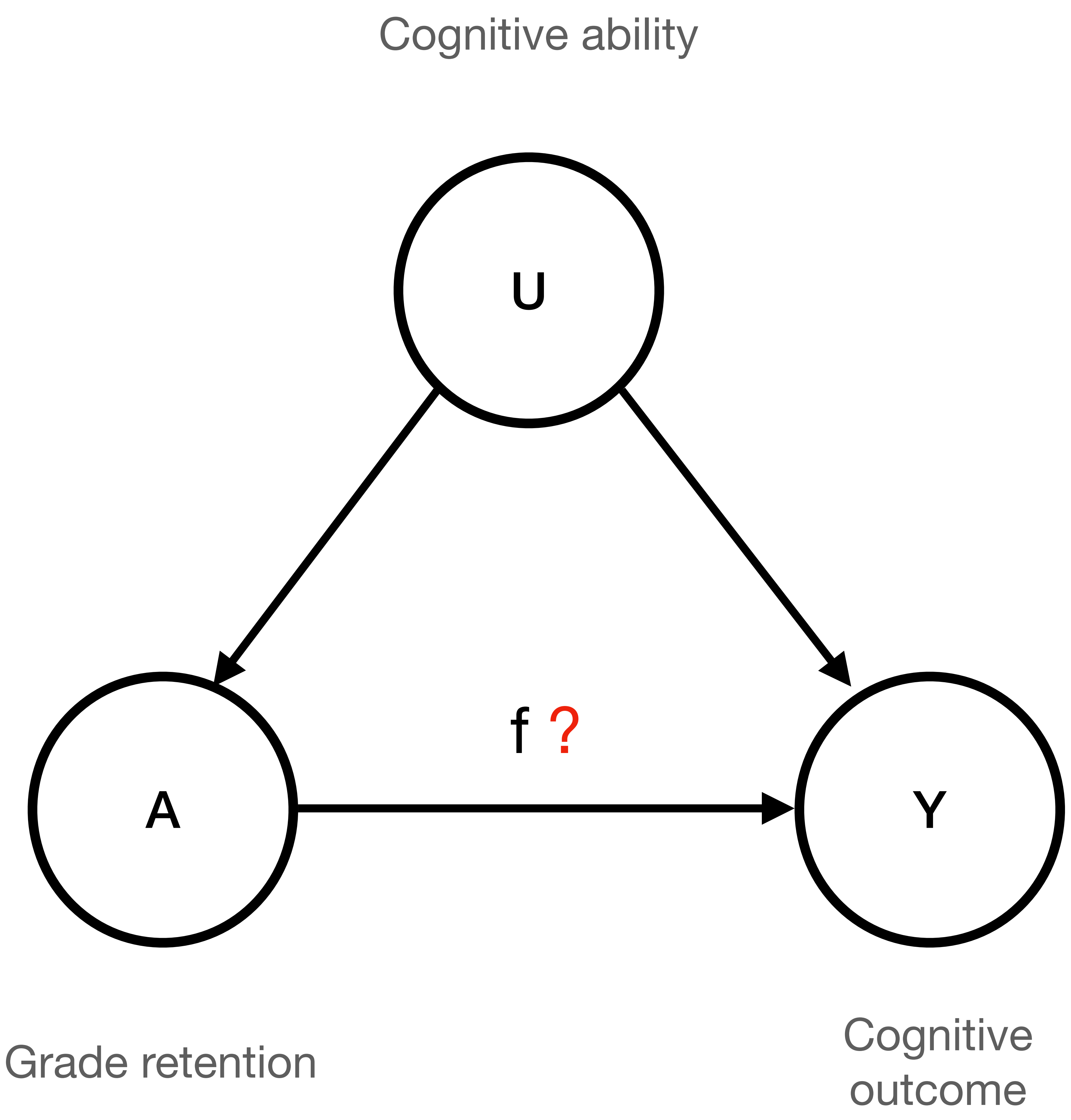
Relaxing observability assumptions in causal inference

Yuchen Zhu. *S-DCE talk 05.2022*

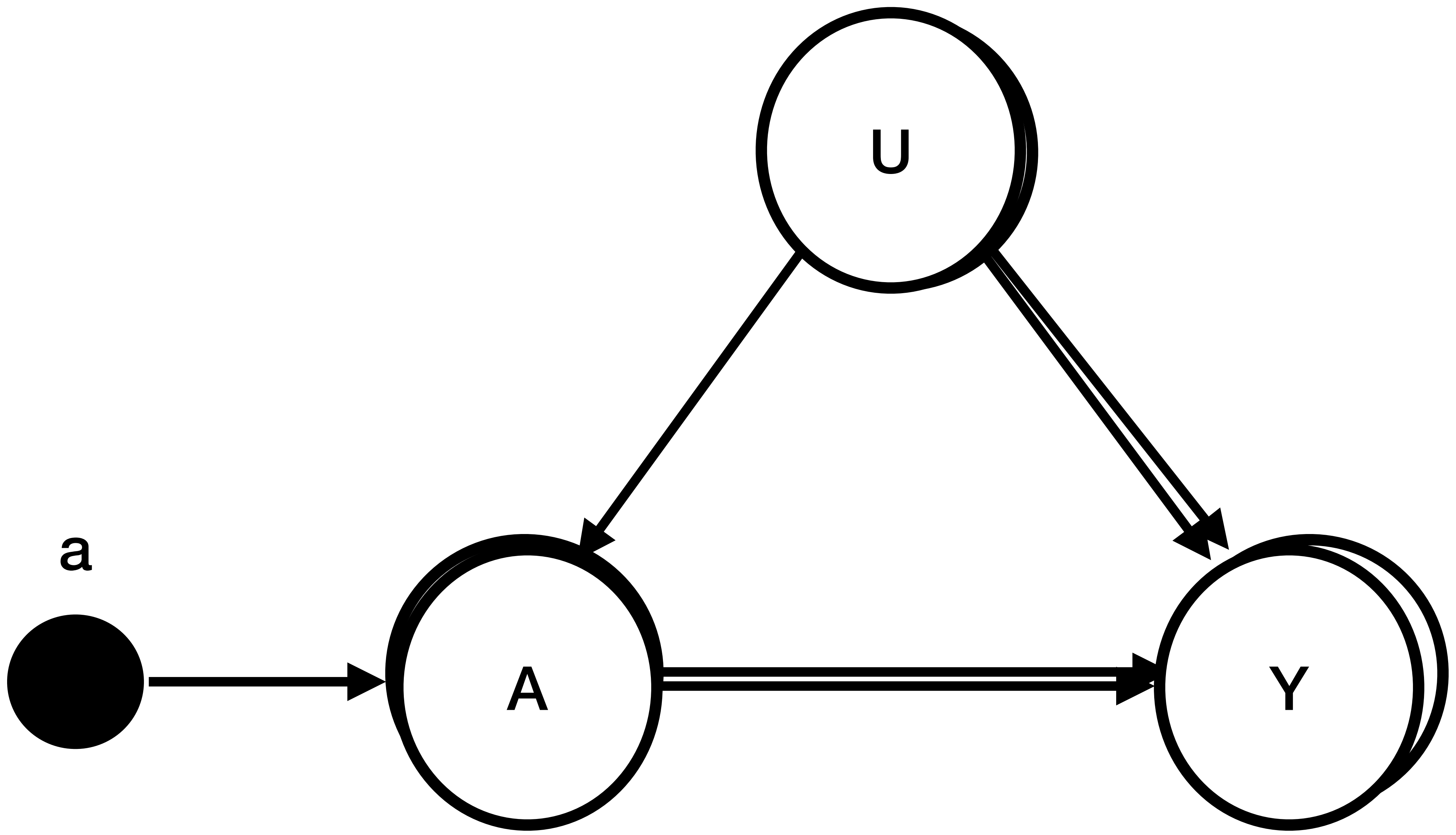
Why causal inference? An example.

Image source: Google image



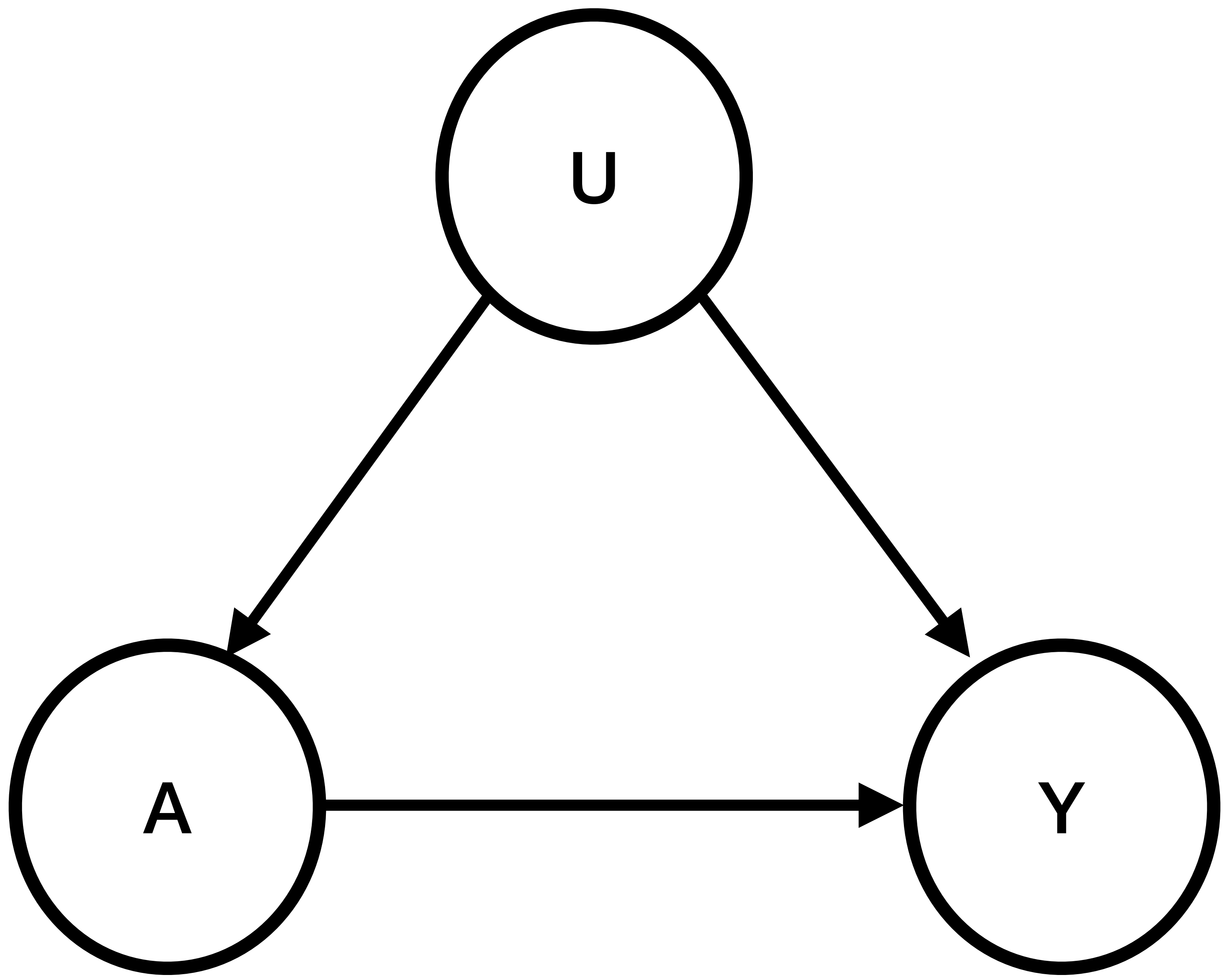
Fruehwirth, Navarro, Takahashi. 2016: How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects.

The target quantity



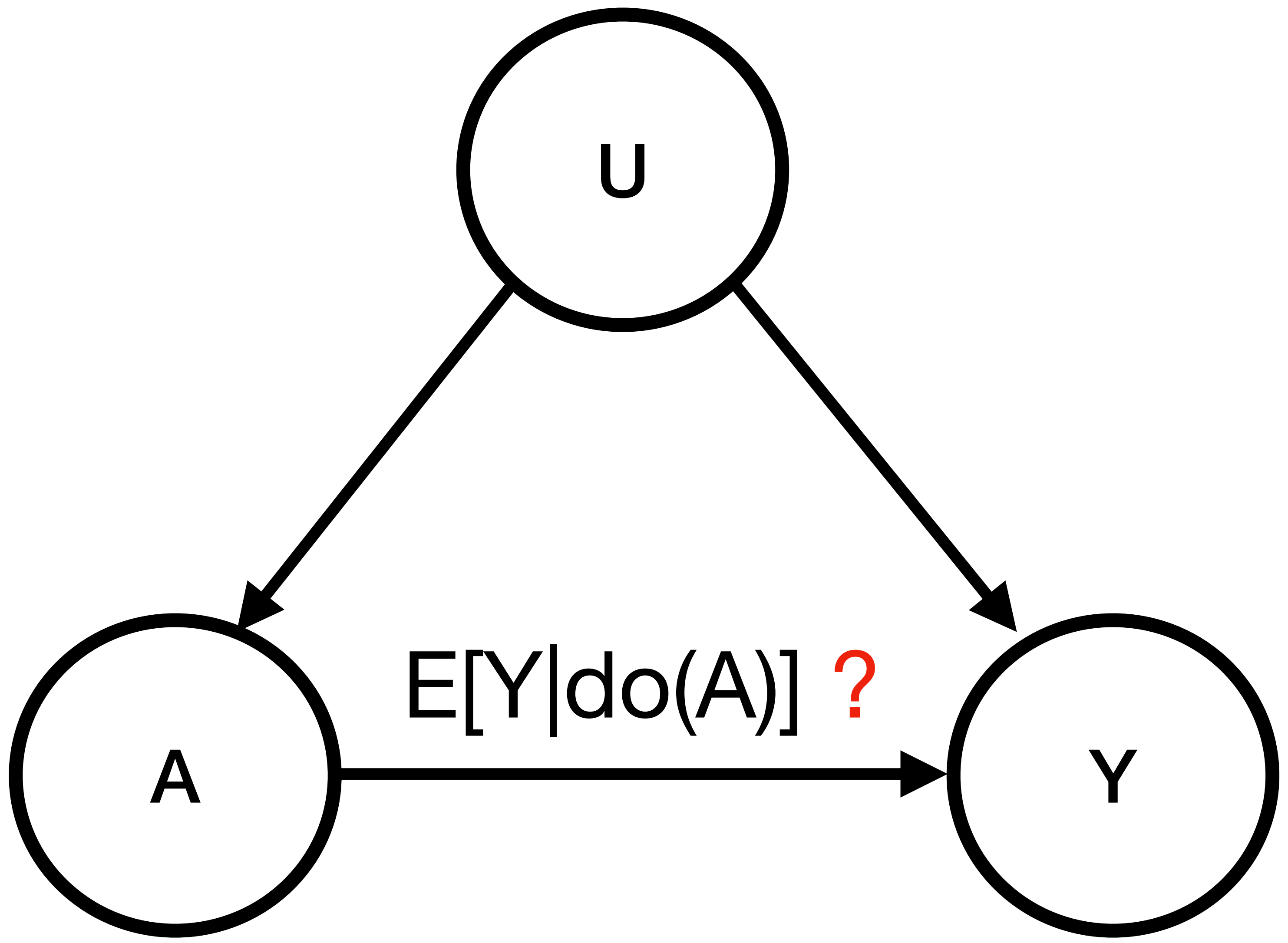
$E[Y|do(A := a)]$

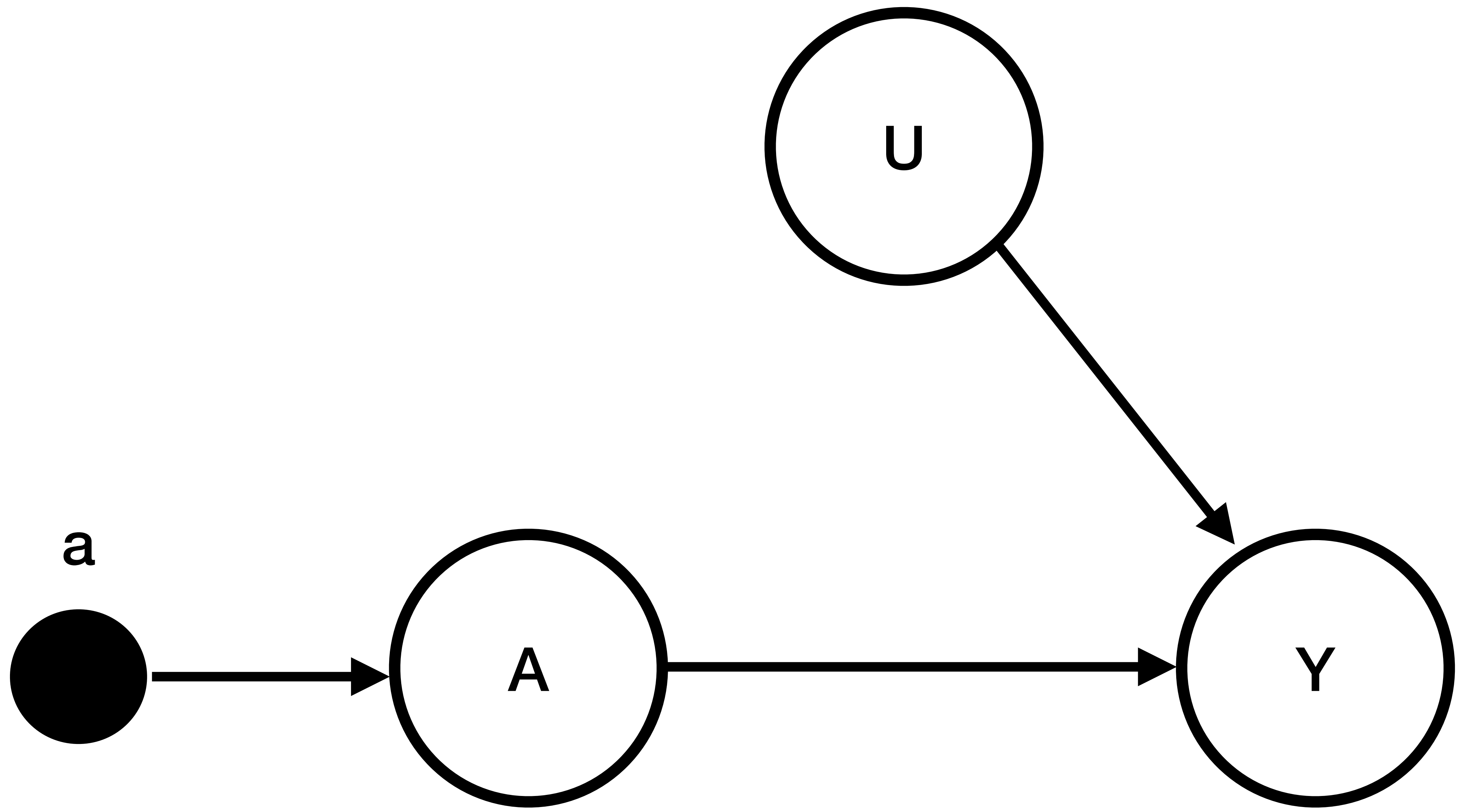
Observed data



Observed data

Warm-up: Observed confounders



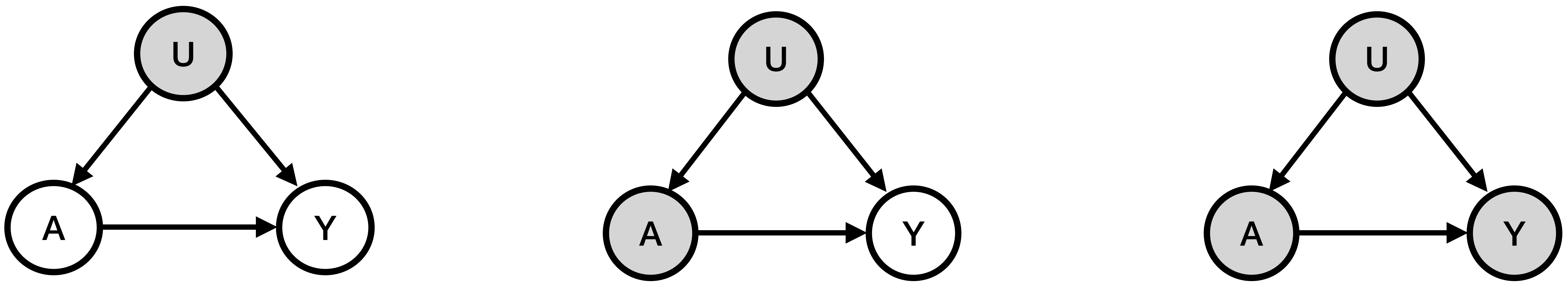


Backdoor adjustment: $\mathbb{E}[Y \,|\, do(a)] = \sum_{i=1}^{n} \mathbb{E}[Y \,|\, A = a, U = i] \mathbb{P}(U = i)$

What are the obstacles?

- Disentanglement issues from U.
    - e.g. U might lead to imbalanced datasets.
- Non-identifiability from latent variables.

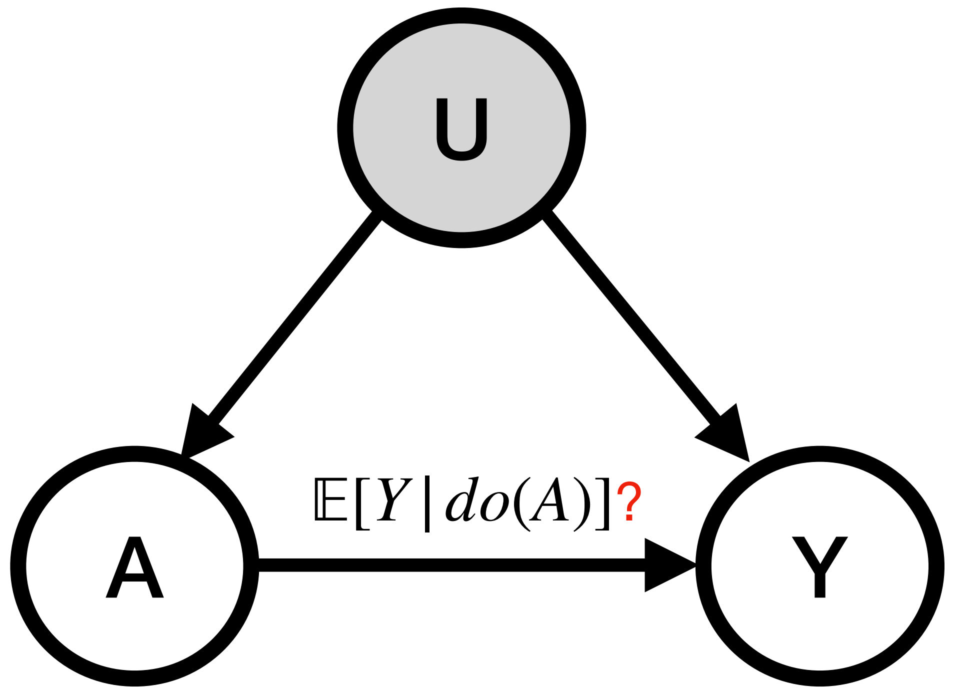
<— More on this next



Why relax observability assumptions?

Unobserved confounders:

Academic ability

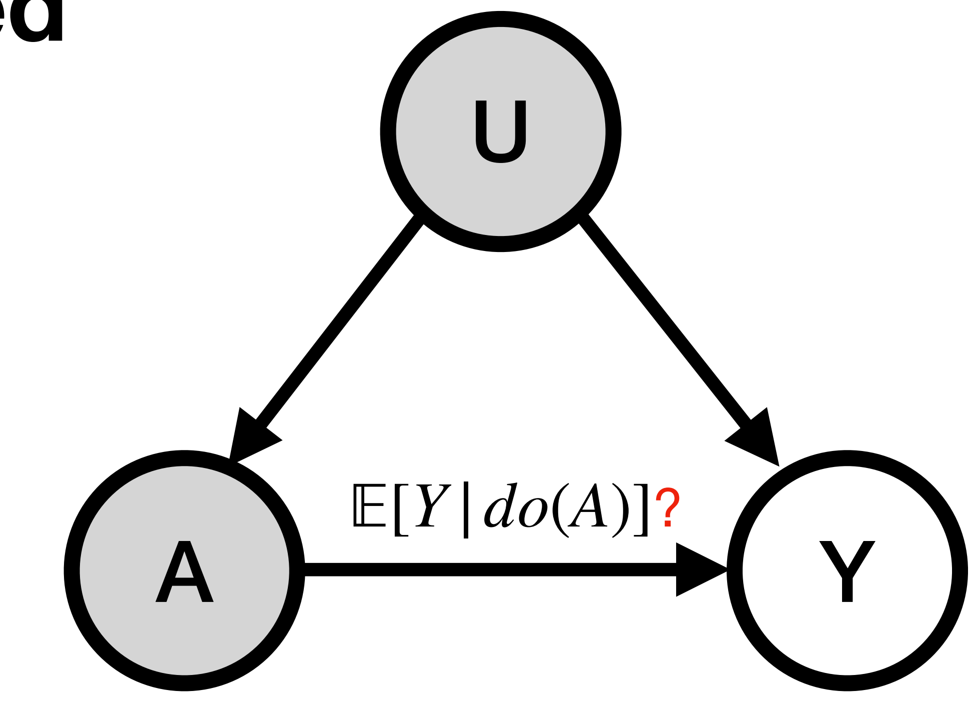


Grade retention

Study outcome

Action observed with error:

Genetics

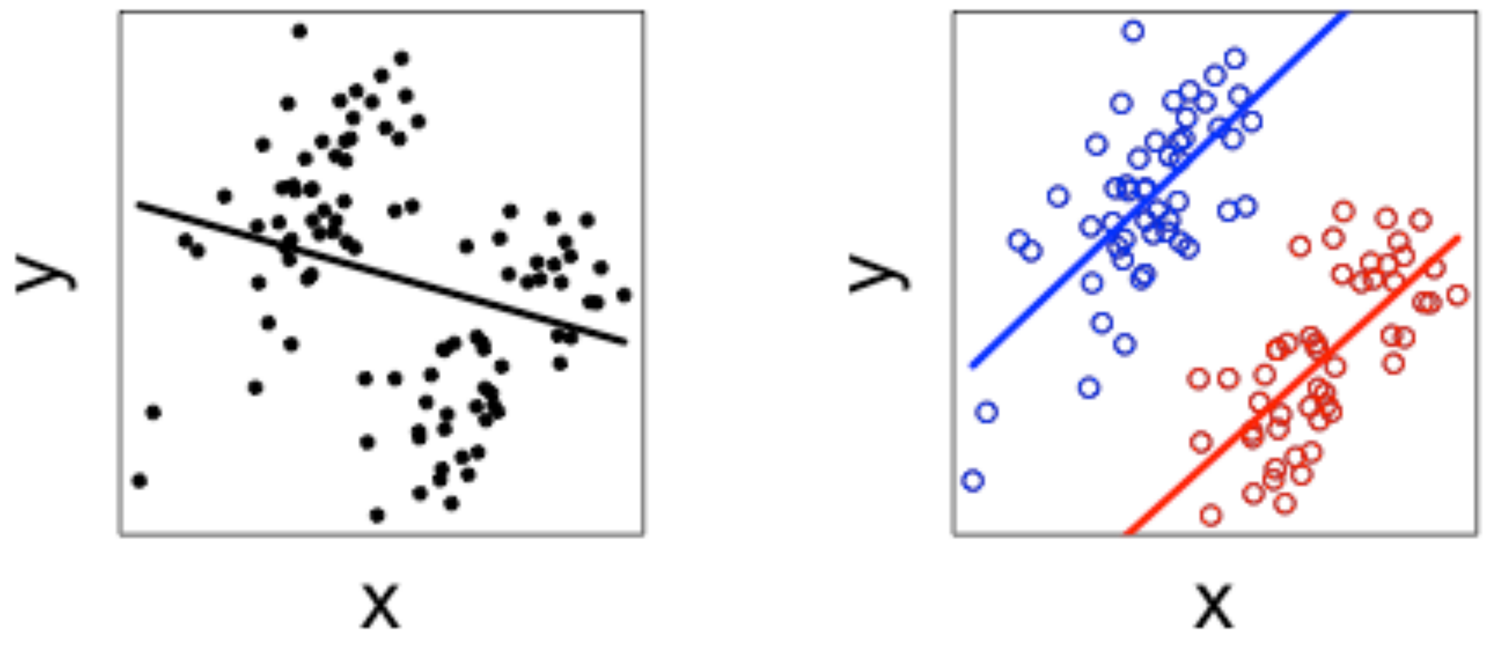


OPEN study: Subar, Thompson, Kipnis, et al. 2001

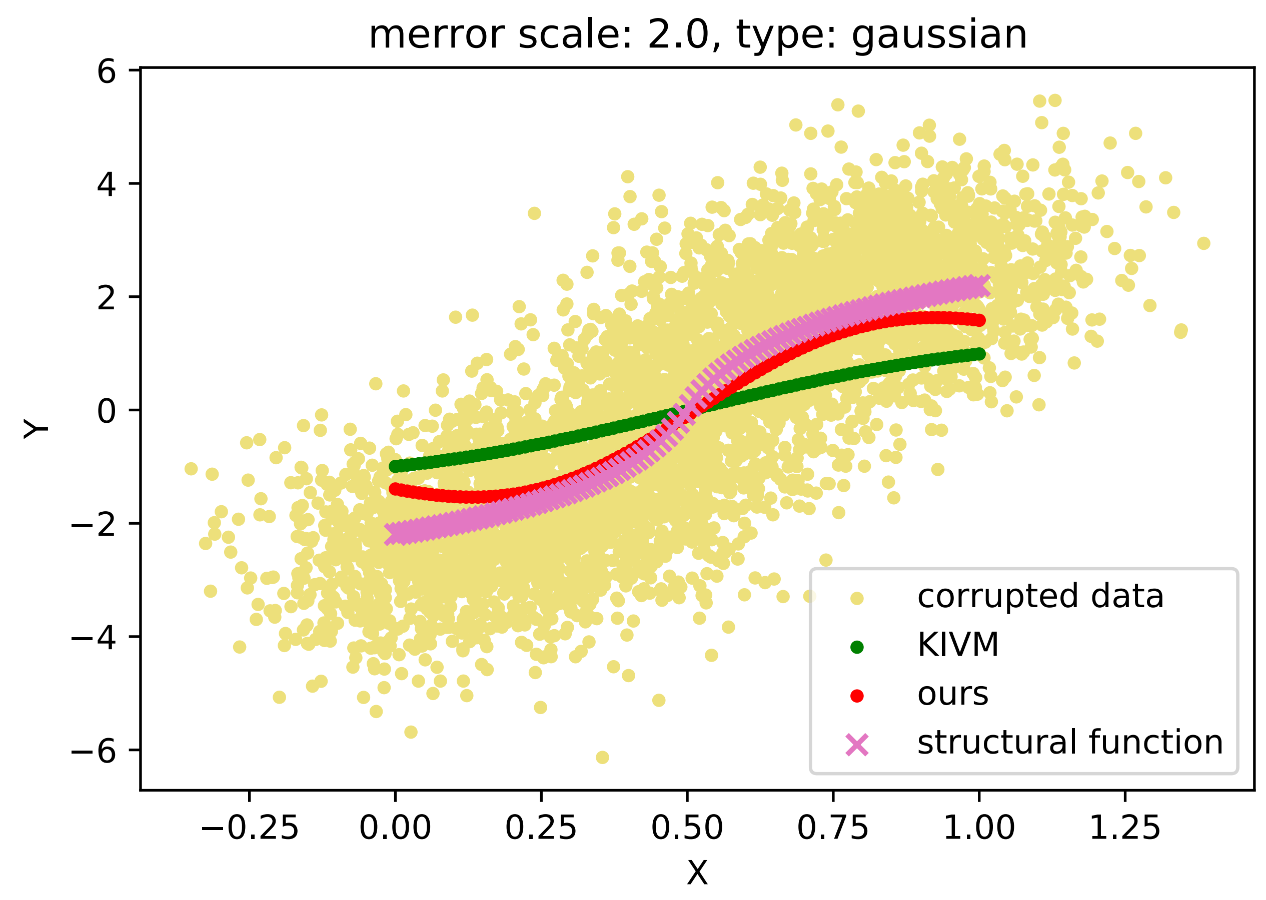
Long term saturated fat intake

Breast cancer

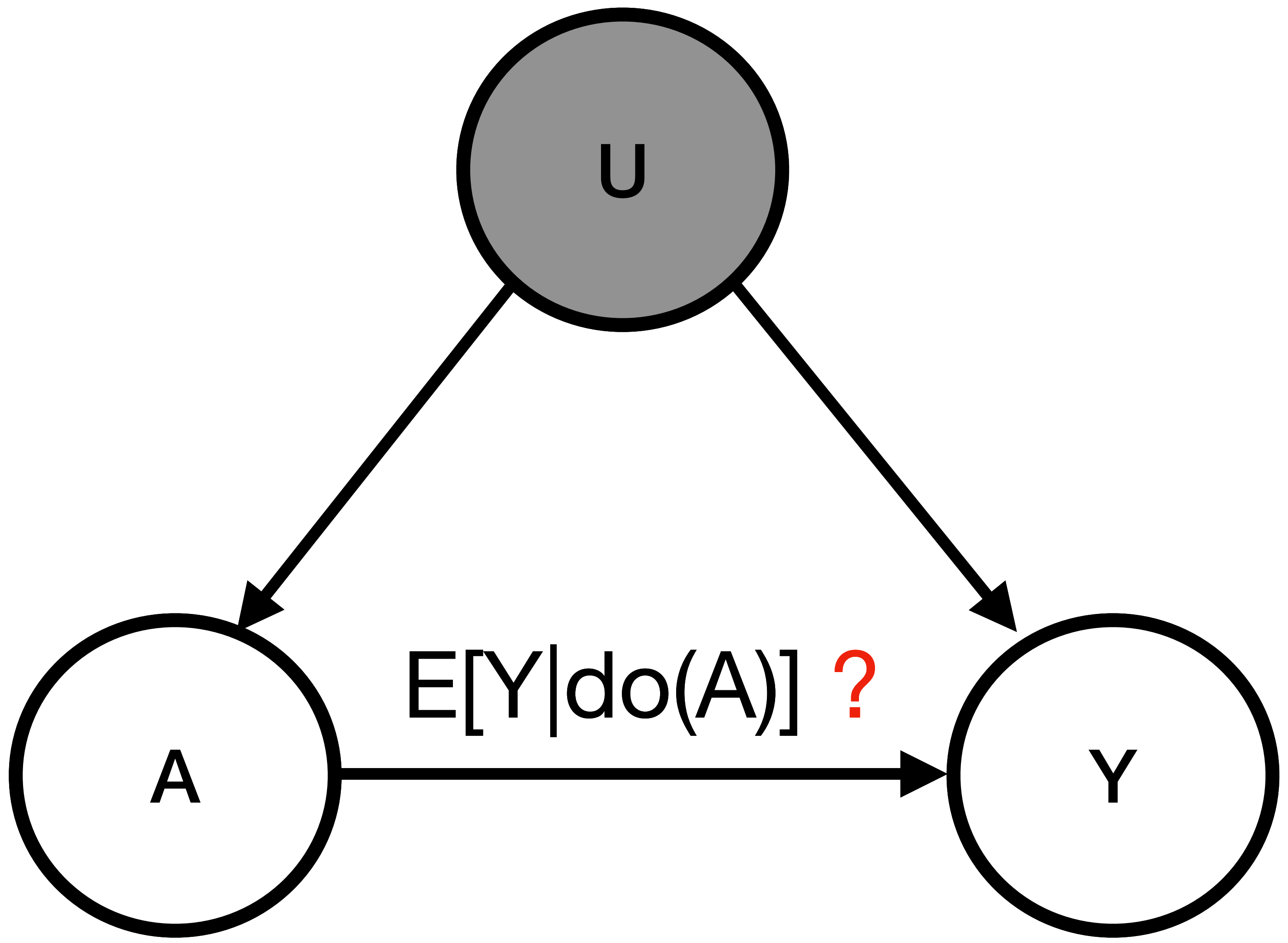
Simpson's paradox:

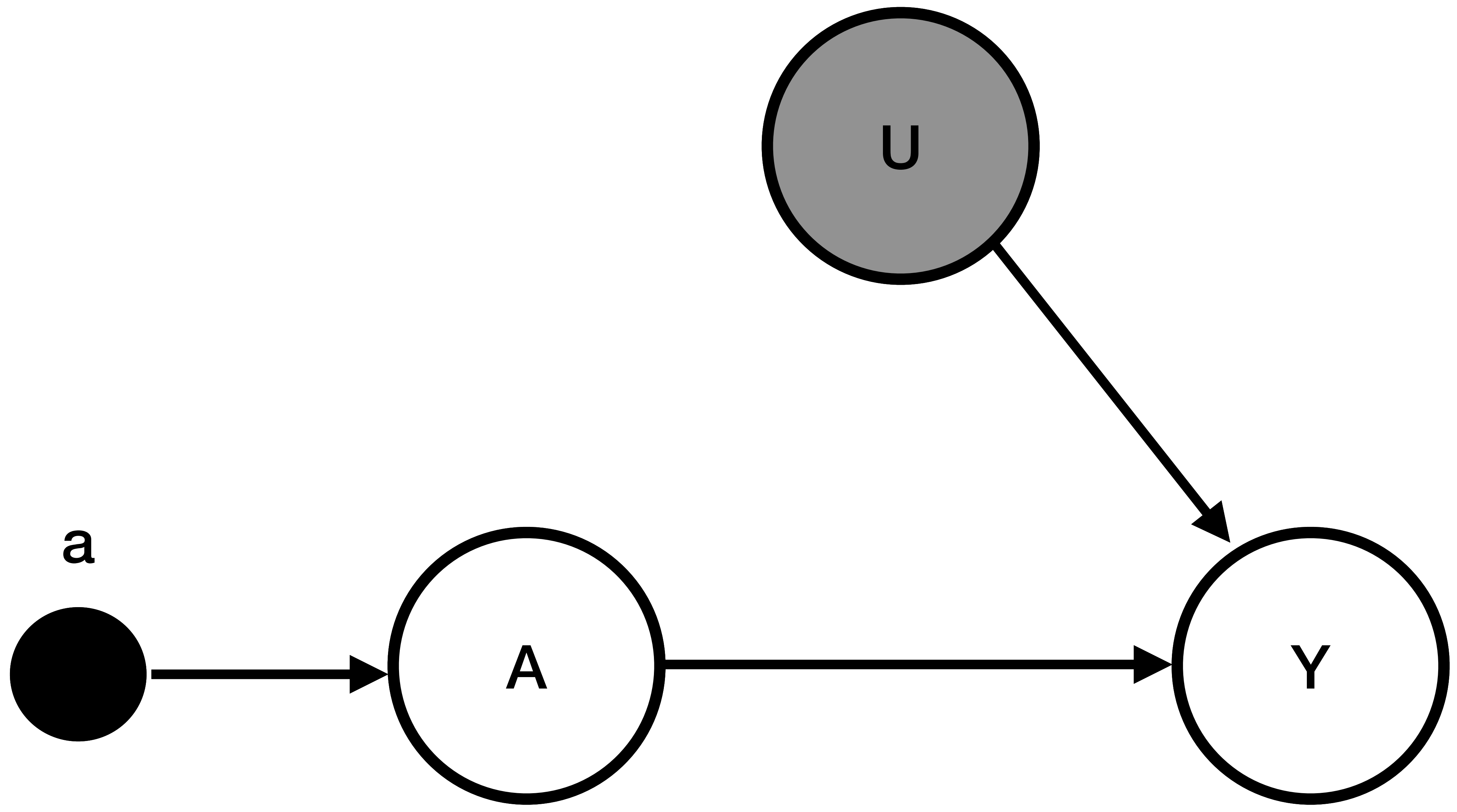


Mask interesting relationships:



Warm-up: Observed confounders

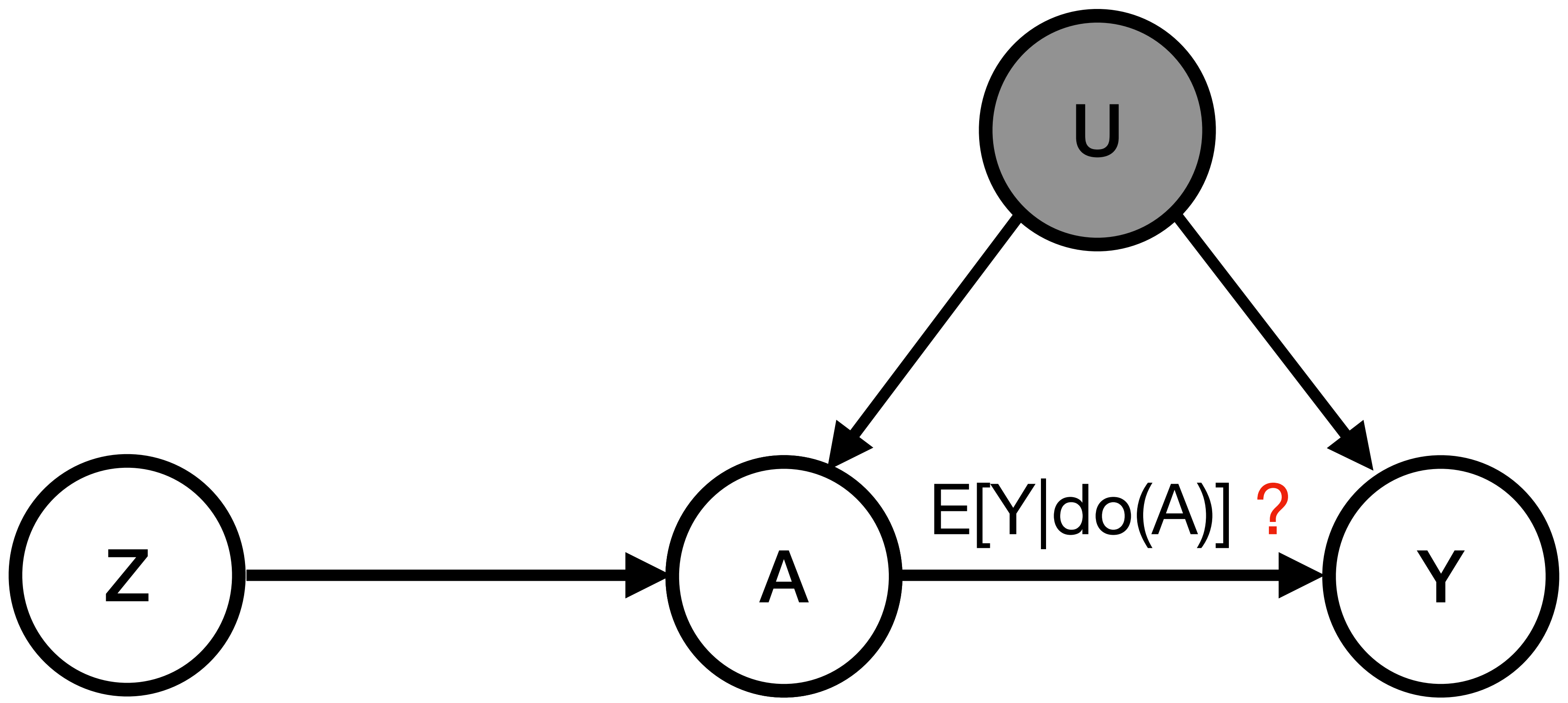




Backdoor adjustment: $\mathbb{E}[Y \,|\, do(a)] = \sum_{i=1}^{n} \mathbb{E}[Y \,|\, A = a, U = i]\mathbb{P}(U = i)$

Unobserved confounders?

Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \epsilon \perp Z$$
$$f(A) = \mathbb{E}[Y | do(A)]$$
$$\mathbb{E}[Y | Z] = \int_{\mathcal{A}} f(a)p(a | Z)da$$

Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$
$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$
$$\implies Y = \beta\gamma Z + \beta\epsilon_A + \epsilon_Y$$

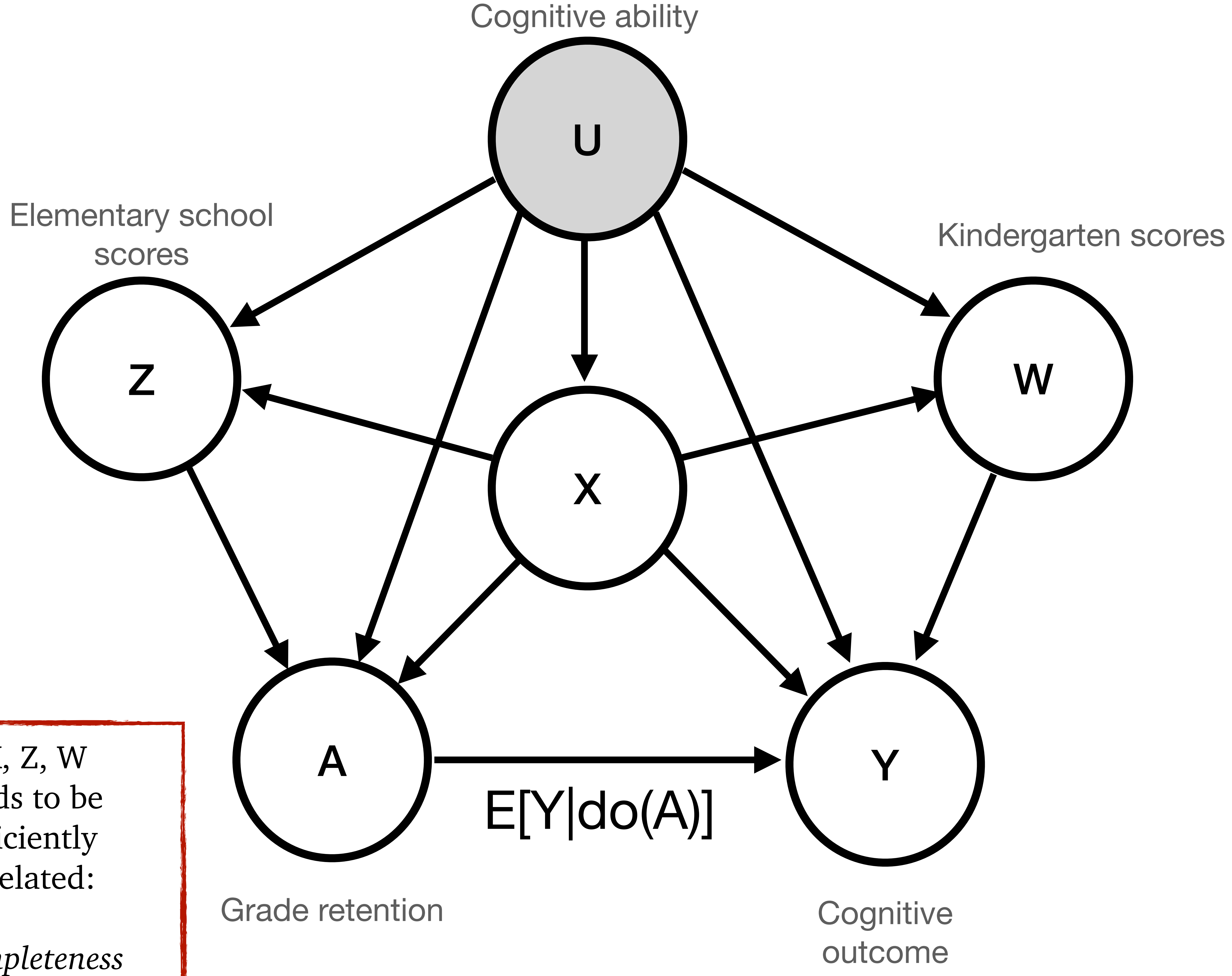
(Strong) Assumptions:

- **Additive error model**
- $(Z ! \perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

Relax the IV to allow for some dependence with U?

False IV: using same 'IV' for several different actions.

Proximal Causal Learning Background



U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Average causal effect estimation:

$$\mathbb{E}[Y \mid do(A=a)] = \int_{XW} h(a,w,x)p(w,x)dxdw$$

How to get h?

- **Expectation operator:** $\mathbb{E}[g(\cdot_U) \mid A, Z, X]$
- $\mathbb{E}[Y \mid A, U, X] = \int h(A, w, x)p(w, x \mid U, X)dxdw$

$$\mathbb{E}[Y - h(A, W, X) \mid A, Z, X] = 0 \quad \text{a.s. } P_{AZX}$$

- **Normal regression equation:** "$\mathbb{E}[Y - h(A, Z, X) \mid A, Z, X] = 0$ a.s. P_{AZX}"
- **Here we also need to take the expectation over $P_{W|AZX}$.**

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \,|\, A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

$$\Updownarrow$$

- If E[A|B] = 0,
- Then (for g measurable):
- E[Ag(B)] = E[E[Ag(B)|B]]
- = E[E[A|B]g(B)] = 0

CMR

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \quad \textbf{a.s. } P_{AXZ}$$

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$$\Updownarrow$$

- Restrict g to $\mathcal{H}_{\mathcal{AXZ}}$

PMMR surrogate loss $R_k(h)$ k indexes the kernel.

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

Precursor loss:

$$R(h) = \sup_{g} (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

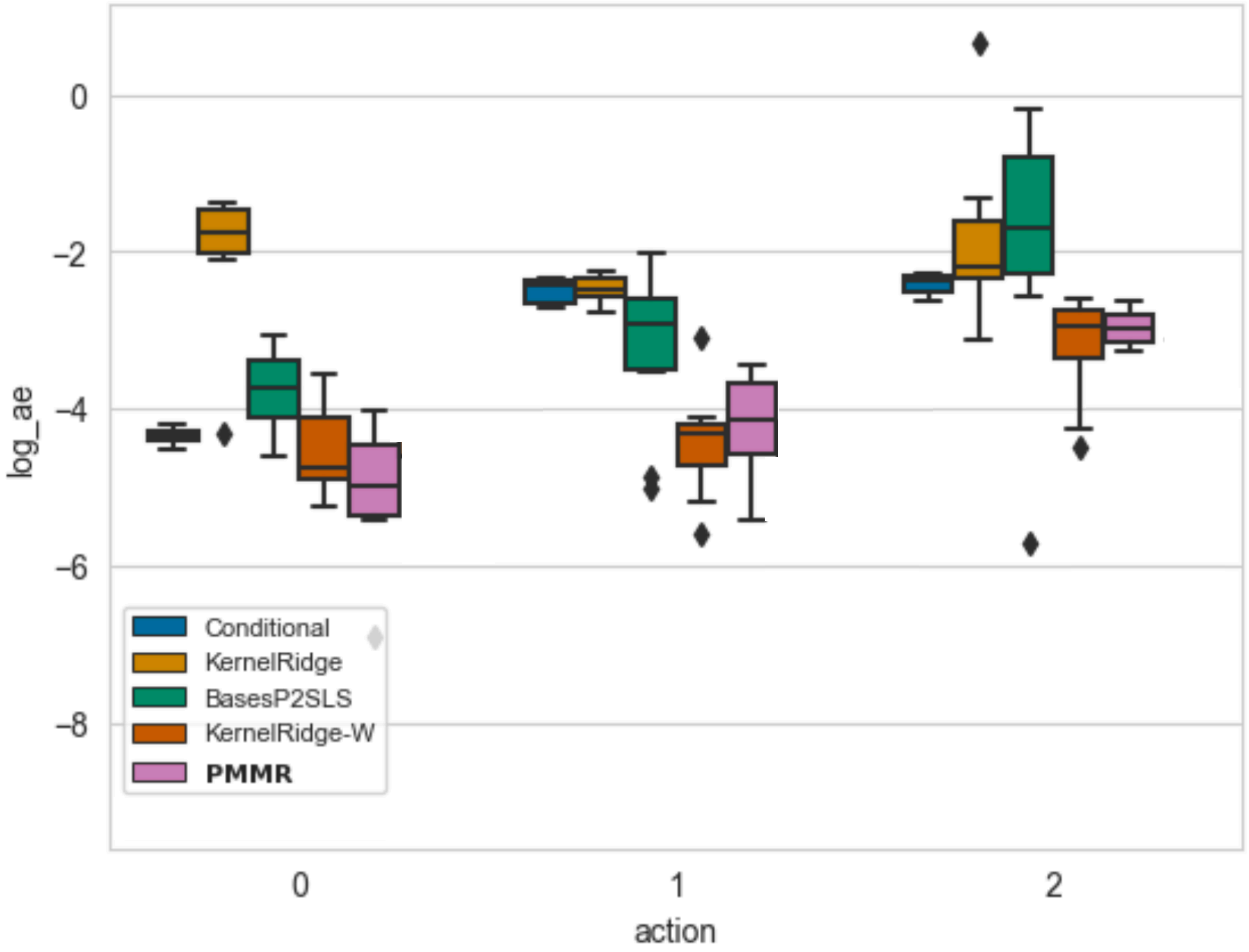
$$\Downarrow$$

$$R_k(h) = \sup_{g \in \mathcal{H}_{\mathcal{A}\mathcal{Z}\mathcal{X}}, \quad \|g\| \leq 1} (\mathbb{E}[(Y - h(A, W, X))\langle g, k((A, Z, X), \cdot)\rangle])^2$$

$$= \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X'))k((A, Z, X), (A', Z', X'))]$$

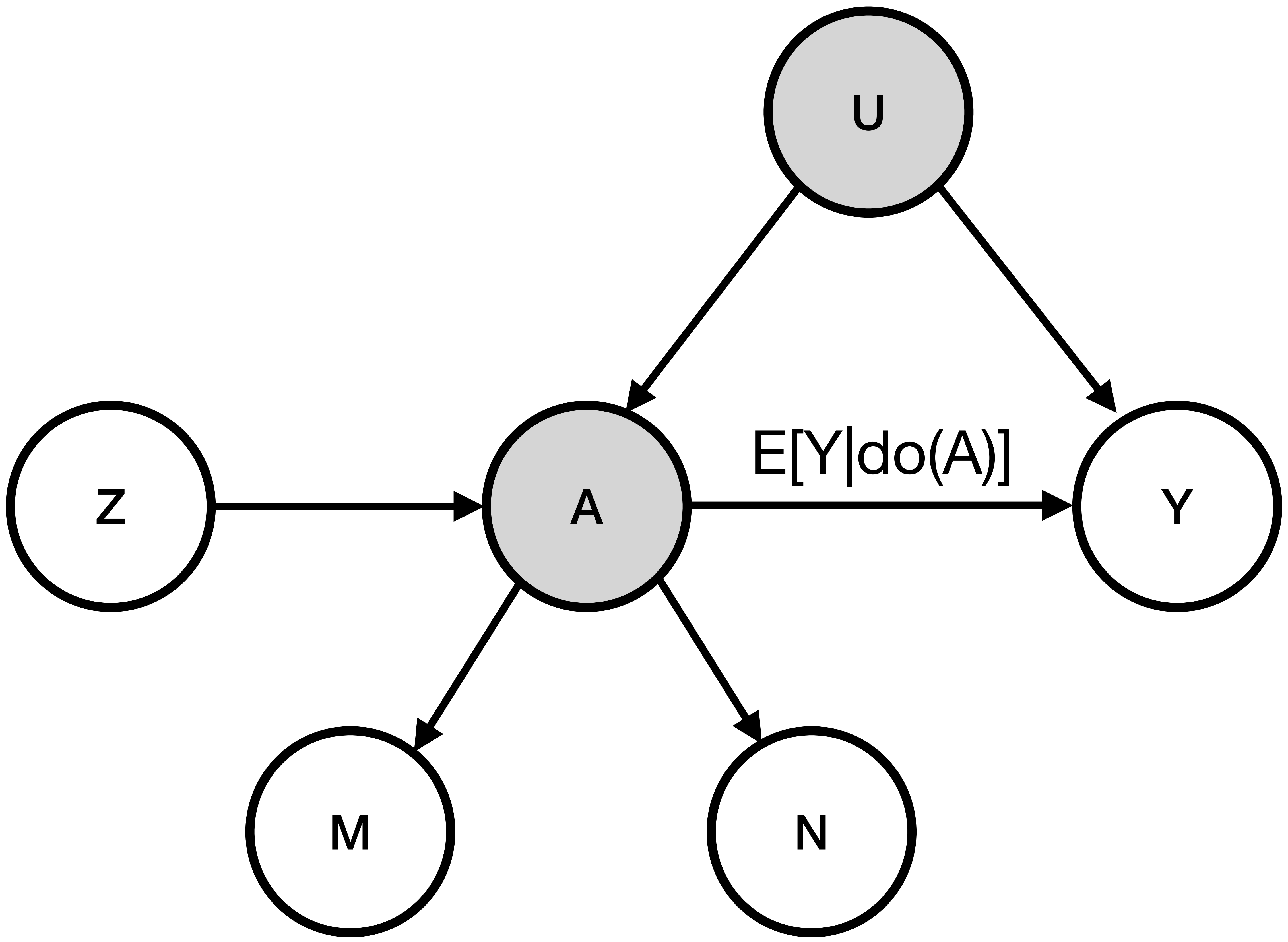
V-statistic: $R_V(h) := \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - h_i)(y_j - h_j)k_{ij}$ (reweighed ERM!)

Results

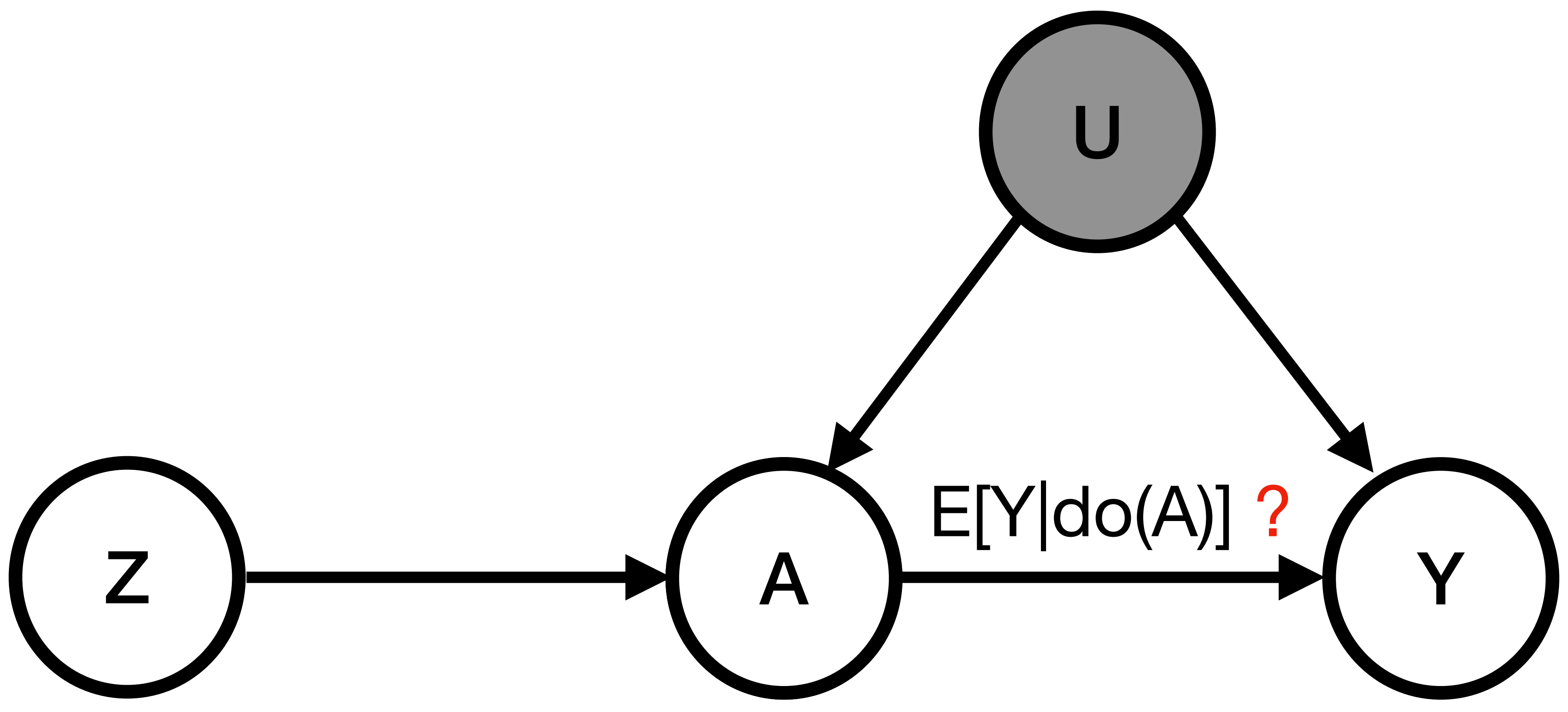


Y: maths score

Measurement error on action variables - overview



Recall: Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \epsilon \perp Z$$

$$f(A) = \mathbb{E}[Y \mid do(A)]$$

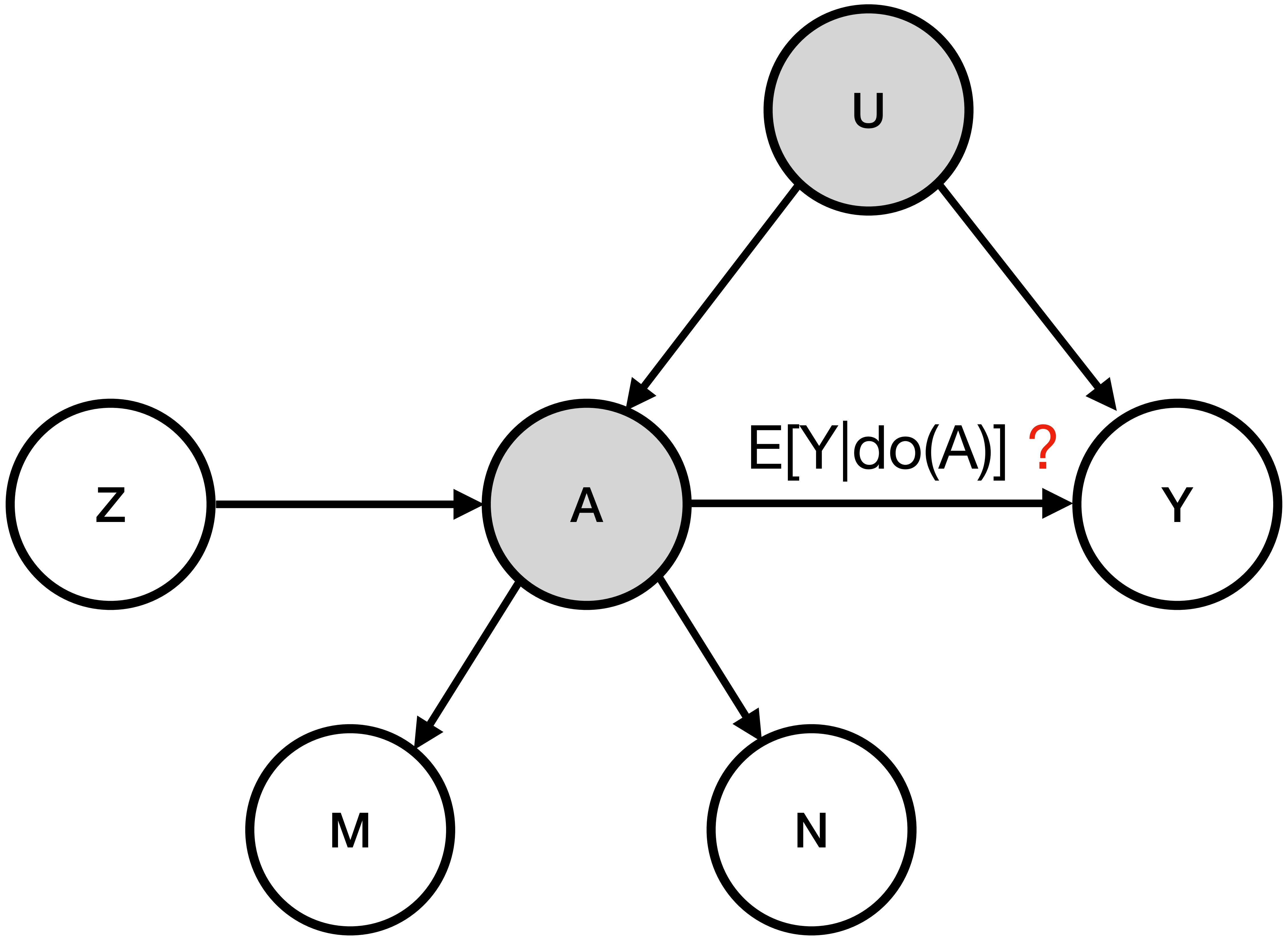
$$\boxed{\mathbb{E}[Y \mid Z] = \int_{\mathscr{A}} f(a) p(a \mid Z) da}$$

???

But if $f(a) = \theta^T \phi(a)$, then simplies to

$$\mathbb{E}[Y \mid Z] = \theta^T \mathbb{E}[\phi(A) \mid Z]$$

Measurement error on action variables - overview



Schennach 2004:

$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}]}^{\psi_{A|Z}(\alpha)} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}[Me^{i\nu N}|z]}{\mathbb{E}[e^{i\nu N}|z]}d\nu\right) \quad (1)$$

Use $\psi_{A|Z}$ to get expectations of basis functions under a parametric assumption (Convolve FT of basis function and characteristic function).

Can we do better? Yes!

From $\hat{\psi}^n_{A|z}(\alpha)$ to $\hat{\mu}^n_{A|z}(y) :='' \mathbb{E}[\phi(A) \mid z]''$

$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}](\alpha)}^{\psi_{A|Z}(\alpha)} = \exp\left(\int_0^{\alpha} i\frac{\mathbb{E}[Me^{i\nu N} \mid z]}{\mathbb{E}[e^{i\nu N} \mid z]} d\nu\right) \quad (1)$$

Have $\hat{\mu}^n_{A|z}(y) = \sum_{j=1}^{n} \hat{\gamma}^n_j(z) k(a_j, y)$.

Let $\hat{\psi}^n_{A|z}(\alpha) := \sum_{j=1}^{n} \hat{\gamma}^n_j(z) e^{i\alpha a_j}$.

Where $\hat{\gamma}^n_j(z) = (K_{ZZ} + n\hat{\lambda}^n I)^{-1} K_{Zz}$.

Theorem 1. With translation-invariant, characteristic kernel:
$\hat{\mu}^n_{A|Z} \to^n \mu_{A|Z}$ iff $\hat{\psi}^n_{A|Z} \to^n \psi_{A|Z}$ in IFT of kernel.

Assume $f \in \mathscr{H}_A$:

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[f(A) \mid Z] = \langle f, \mu_{A|Z} \rangle_{\mathscr{H}_A}$$

How to get $\mu_{A|Z}$?

Approximate $\hat{\mu}_{A|Z}$ via $\hat{\psi}^n_{A|z}$ (eq.1):

Only depend on $\hat{\gamma}^n_j(z)$ and $\{a_j\}^n_{j=1}$!

Z. et al. 2022: Causal Inference with Treatment Measurement Error: A Nonparametric Instrumental Variable Approach. *Under review at UAI 2022.*

Method overview

- The goal is to get the kernel mean embedding of P(A|Z): an infinitely long vector of conditional expected moments of A.
- This can be got from the characteristic function of P(A|Z), and vice versa! They are identical up to suitable measures.
- If A is observed - the embedding turns out to be only a function of the A samples and a regularisation parameter.
- So when A is unobserved, we can optimise to *hallucinate* the ‘samples’ back.
- To avoid integration - differentiate.

MEKIV results