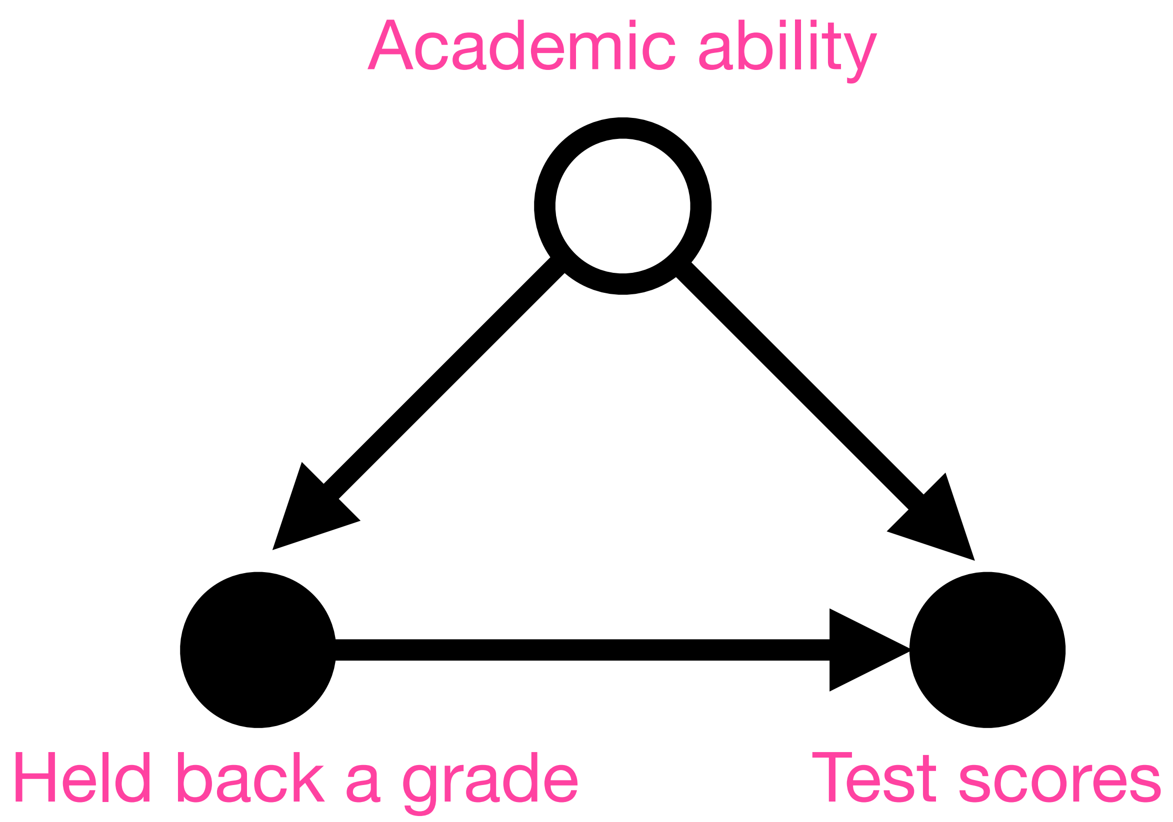
# Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction

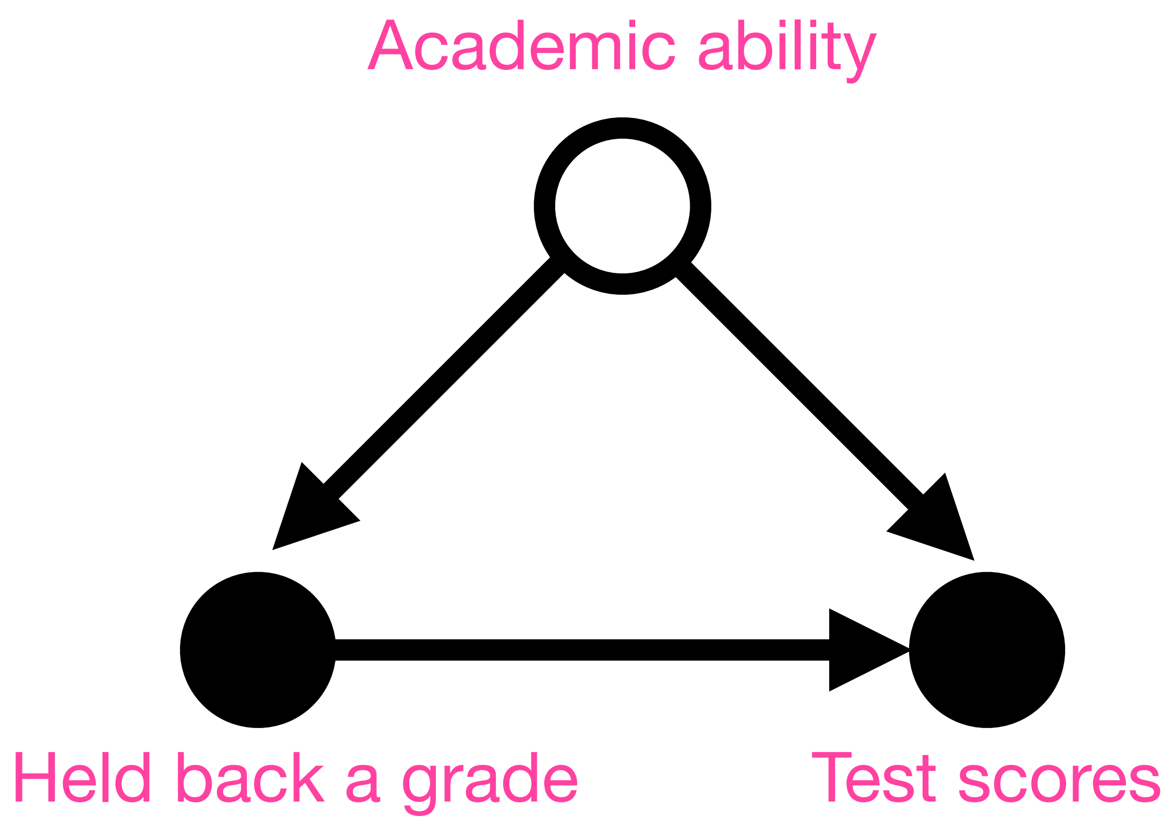
Afsaneh Mastouri*, Yuchen Zhu*, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner
Arthur Gretton▲, Krikamol Muandet▲

UCL
Max Planck Institute for Intelligent Systems
CREST Center for Research in Economics and Statistics
ENSAE ParisTech École Nationale de la Statistique et de l'Administration Économique
GATSBY
The Alan Turing Institute
University of Oxford

# Reliable Decision Making

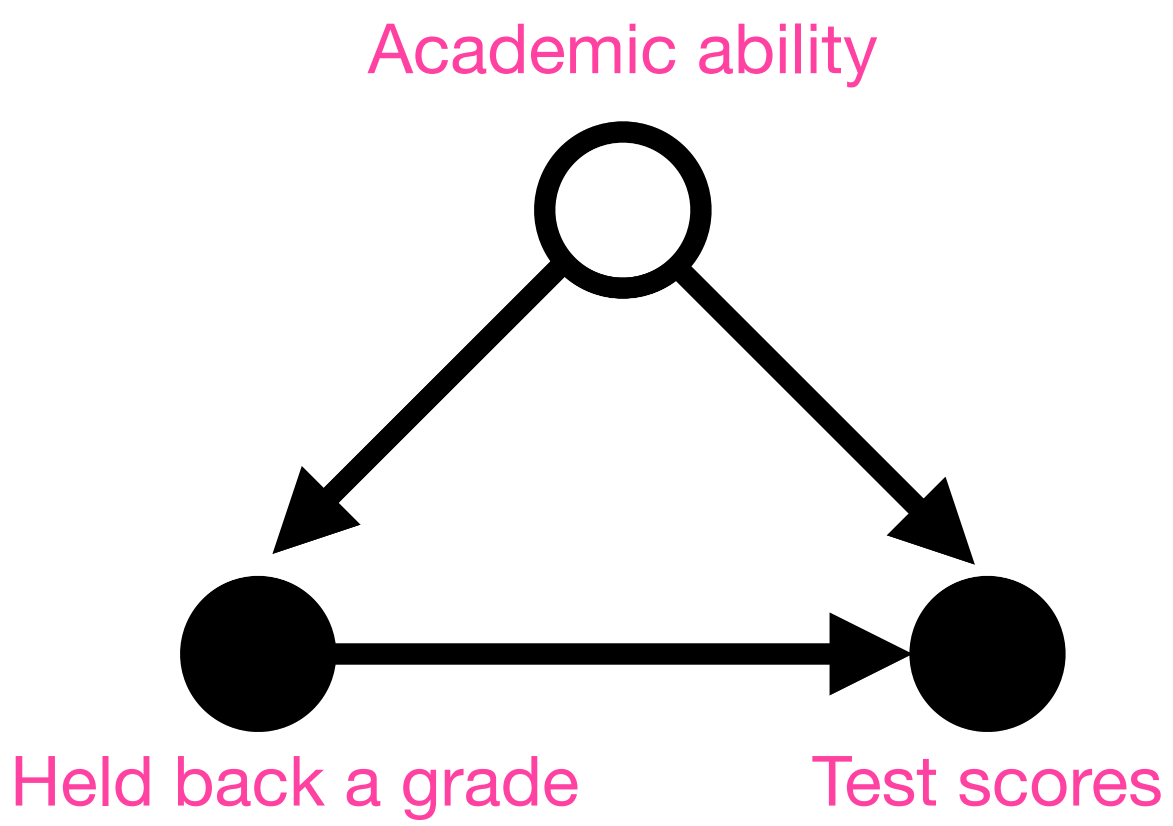



# Reliable Decision Making



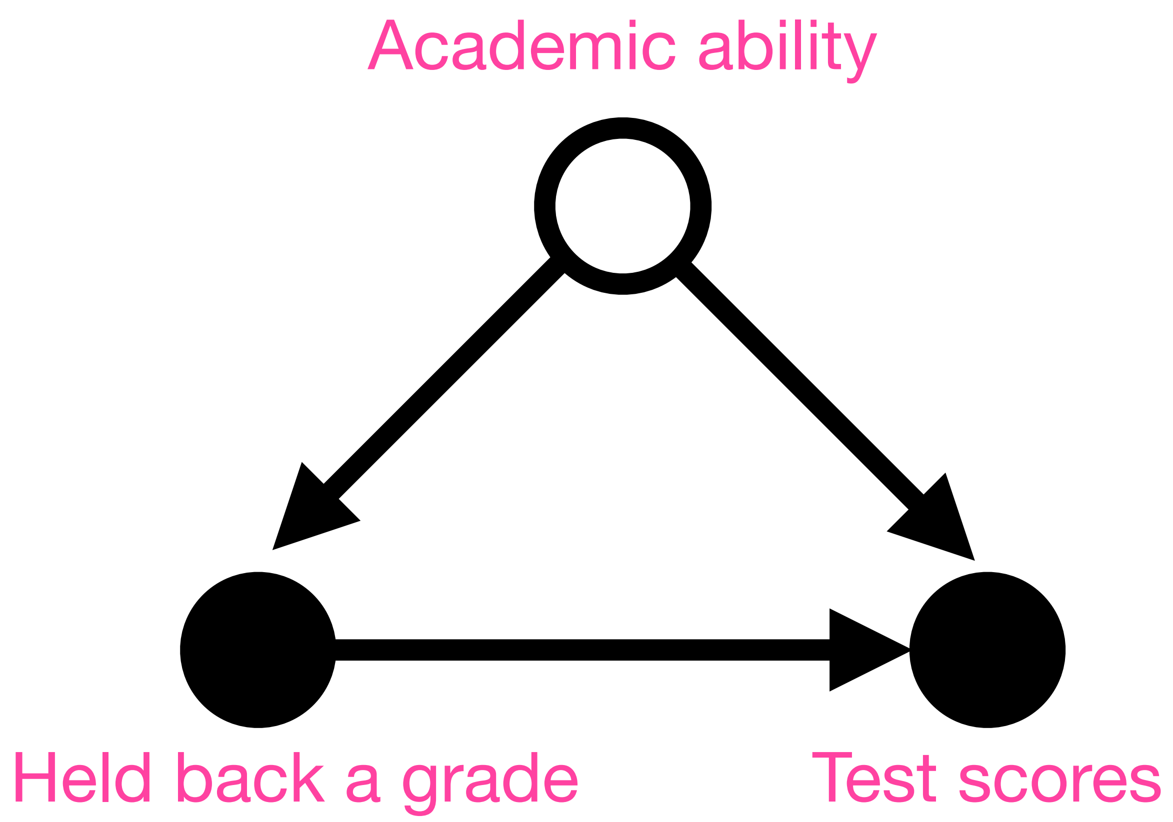

- Machine learning allows us to create models that excel at making **prediction**.

# Reliable Decision Making



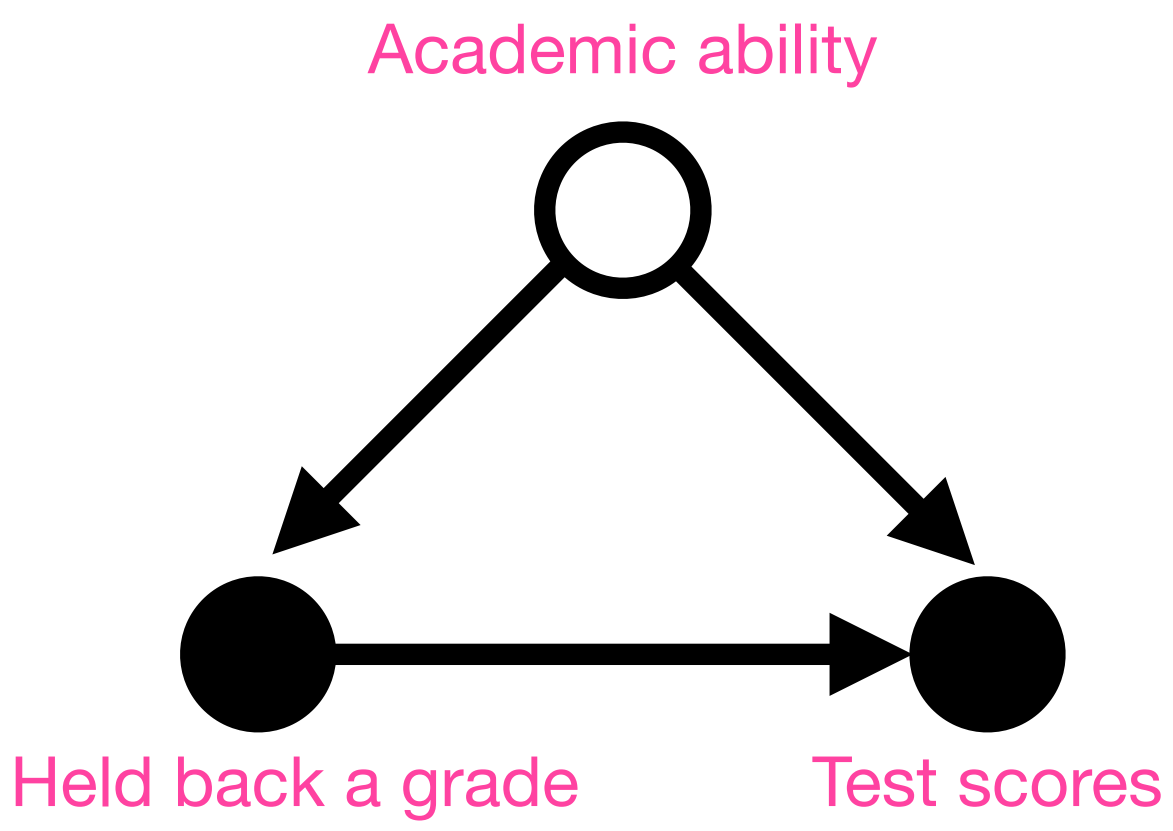

- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.

# Reliable Decision Making



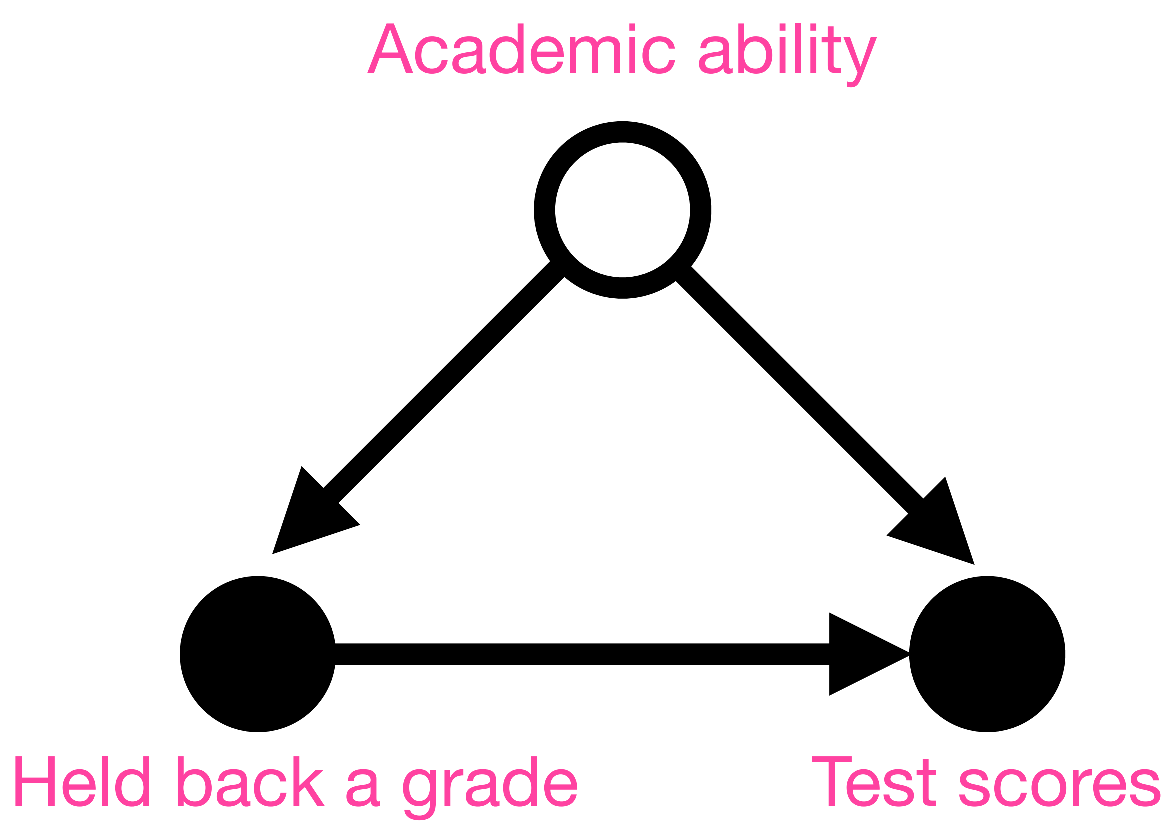

- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
  - Will holding back a grade, improve the test scores of students?

# Reliable Decision Making

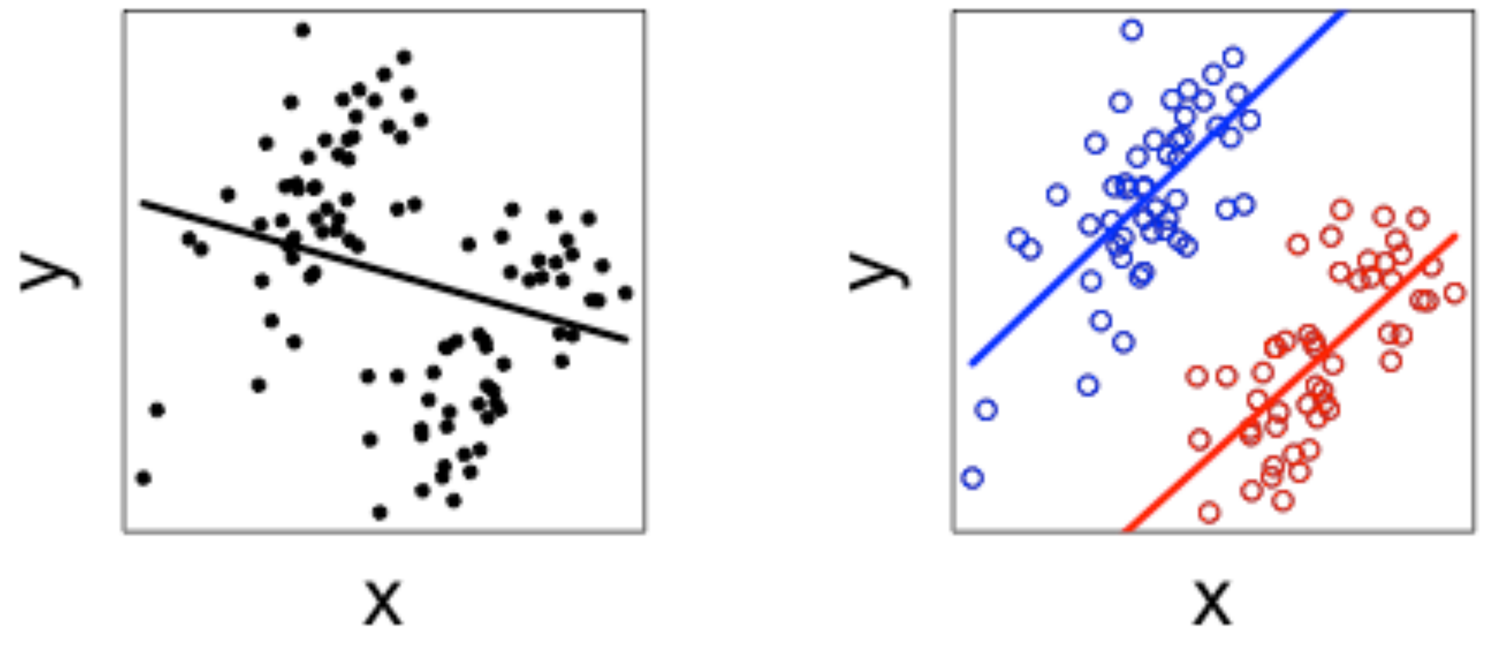



- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
    - Will holding back a grade, improve the test scores of students?
- **Experimental** data are not available. Only **Observational**.

# Reliable Decision Making

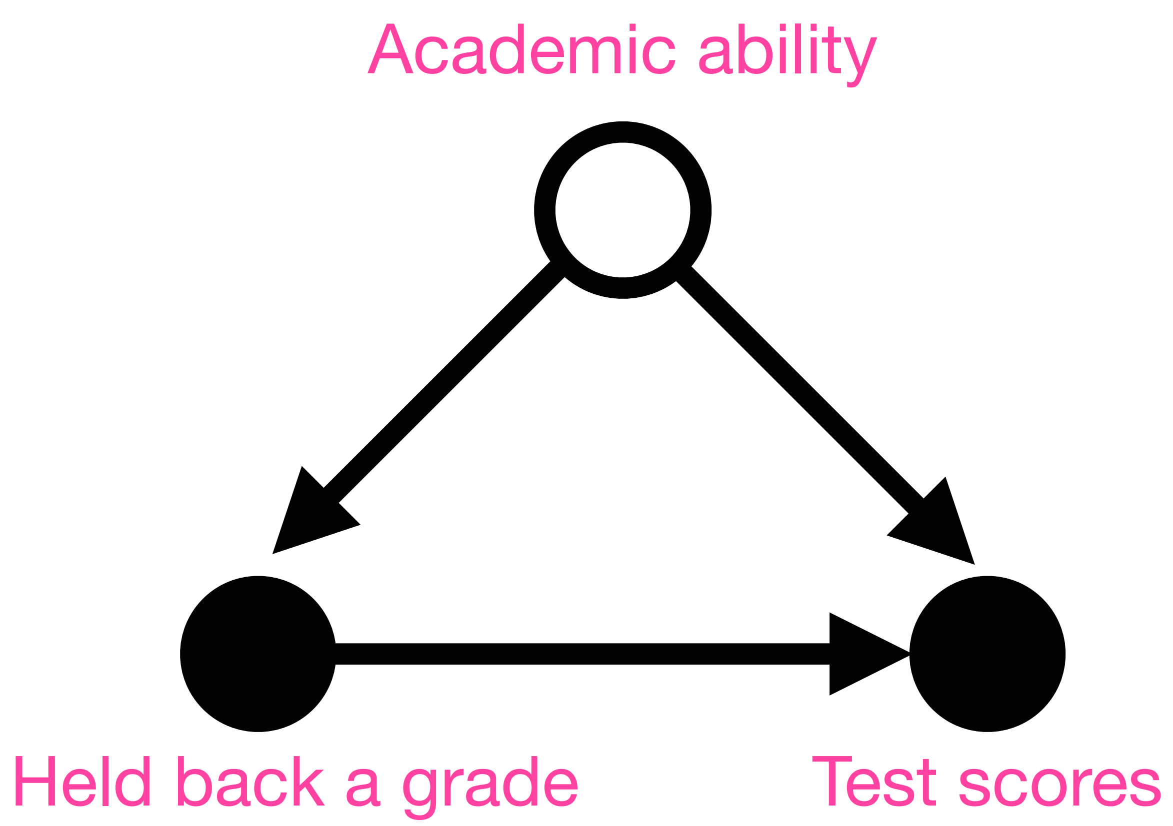



- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
  - Will holding back a grade, improve the test scores of students?

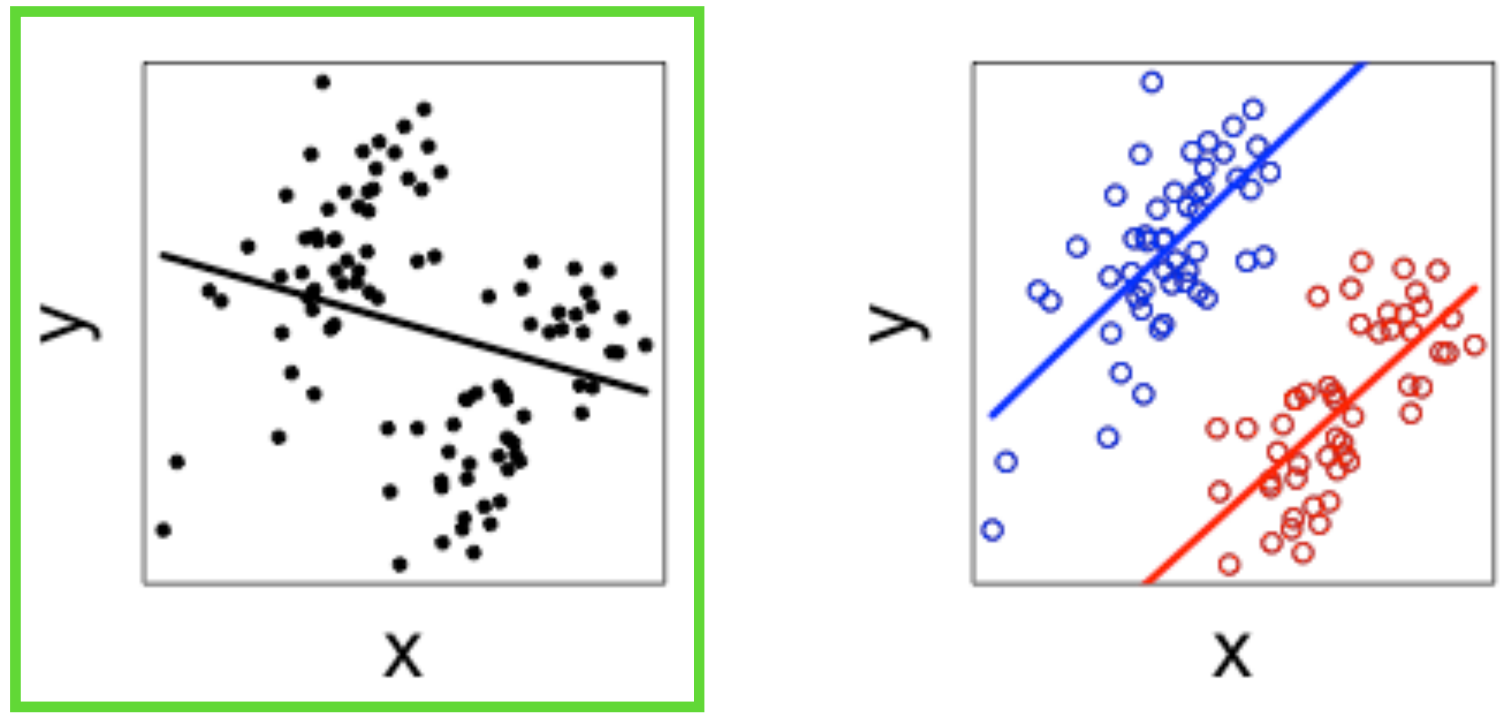



- **Experimental** data are not available. Only **Observational**.
- We cannot rule out the effect of **unobserved confounders** (Simpson’s paradox)

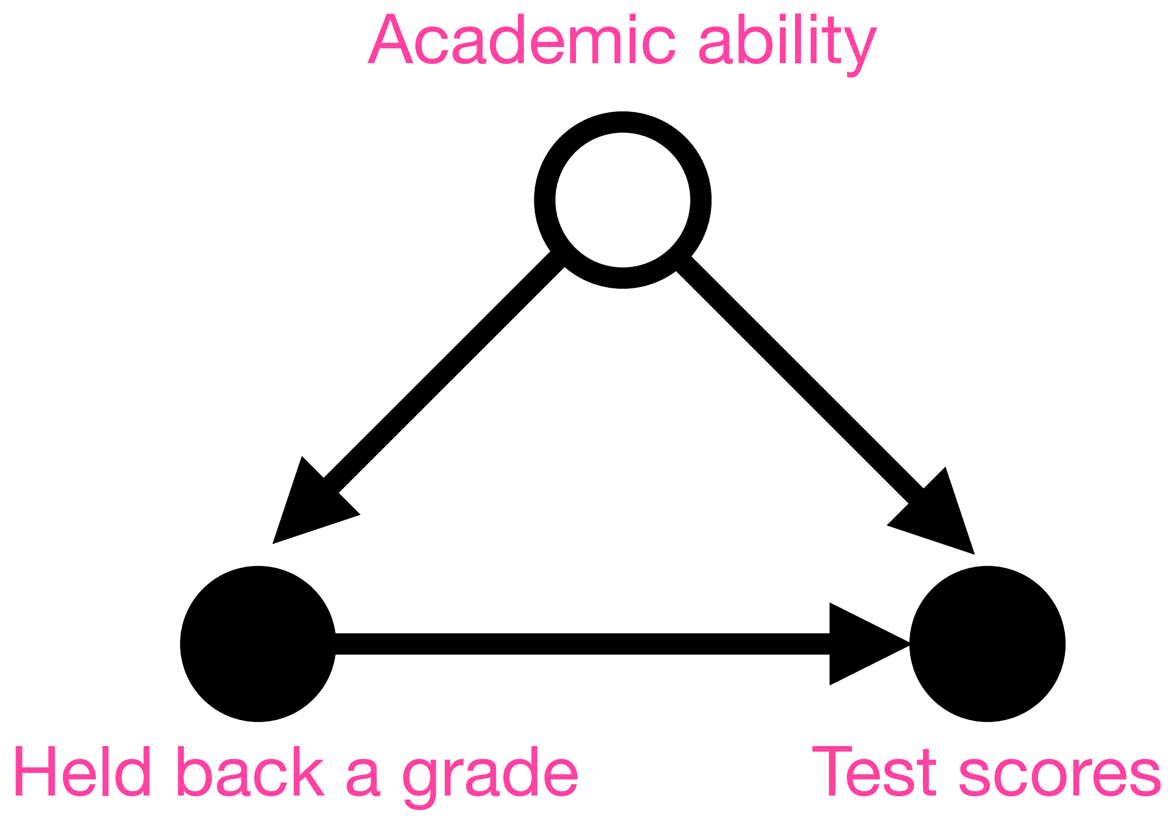
# Reliable Decision Making



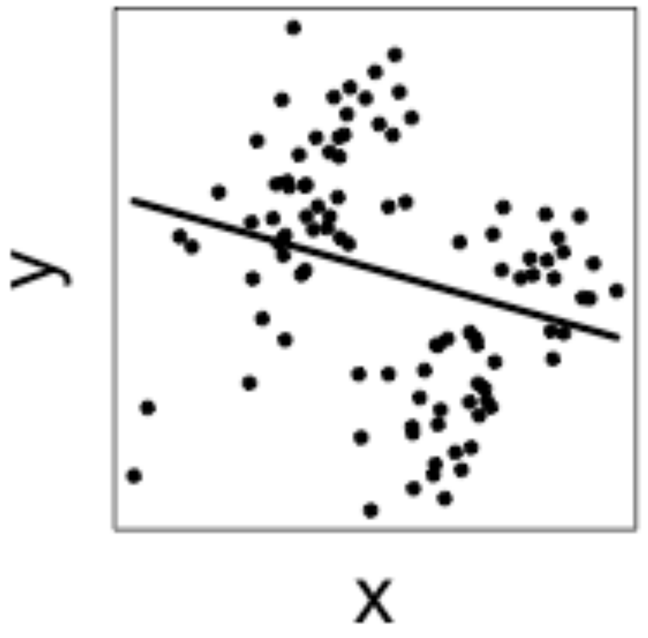

- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
  - Will holding back a grade, improve the test scores of students?
- **Experimental** data are not available. Only **Observational**.
- We cannot rule out the effect of **unobserved confounders** (Simpson's paradox)

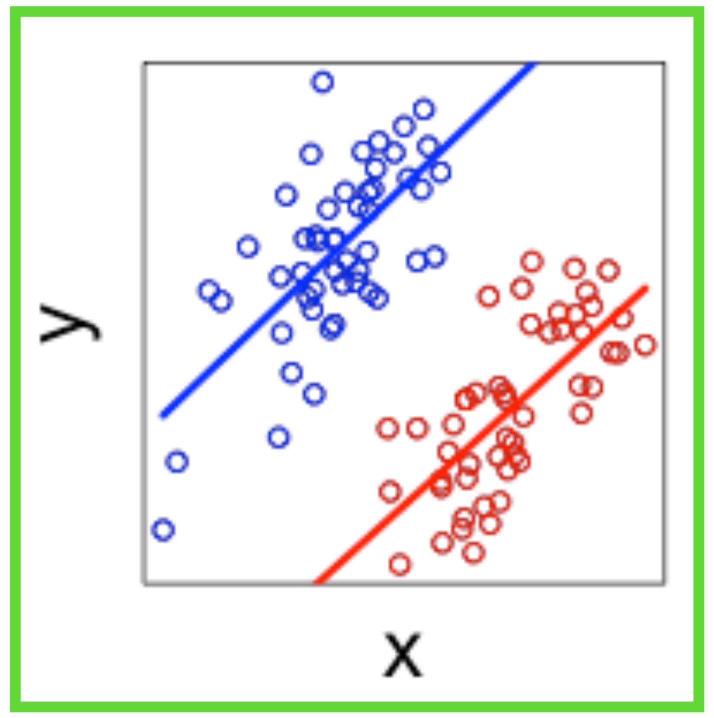


# Reliable Decision Making

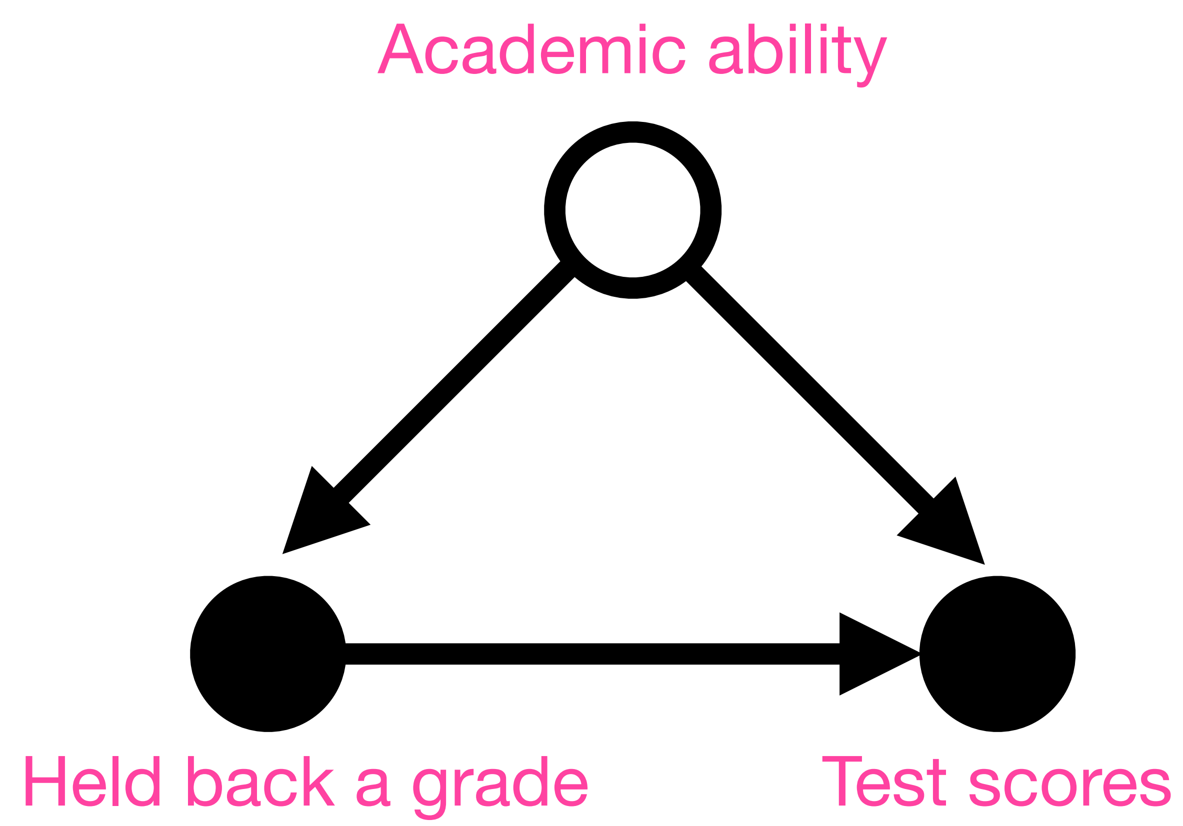



- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
  - Will holding back a grade, improve the test scores of students?
- **Experimental** data are not available. Only **Observational**.
- We cannot rule out the effect of **unobserved confounders** (Simpson’s paradox)

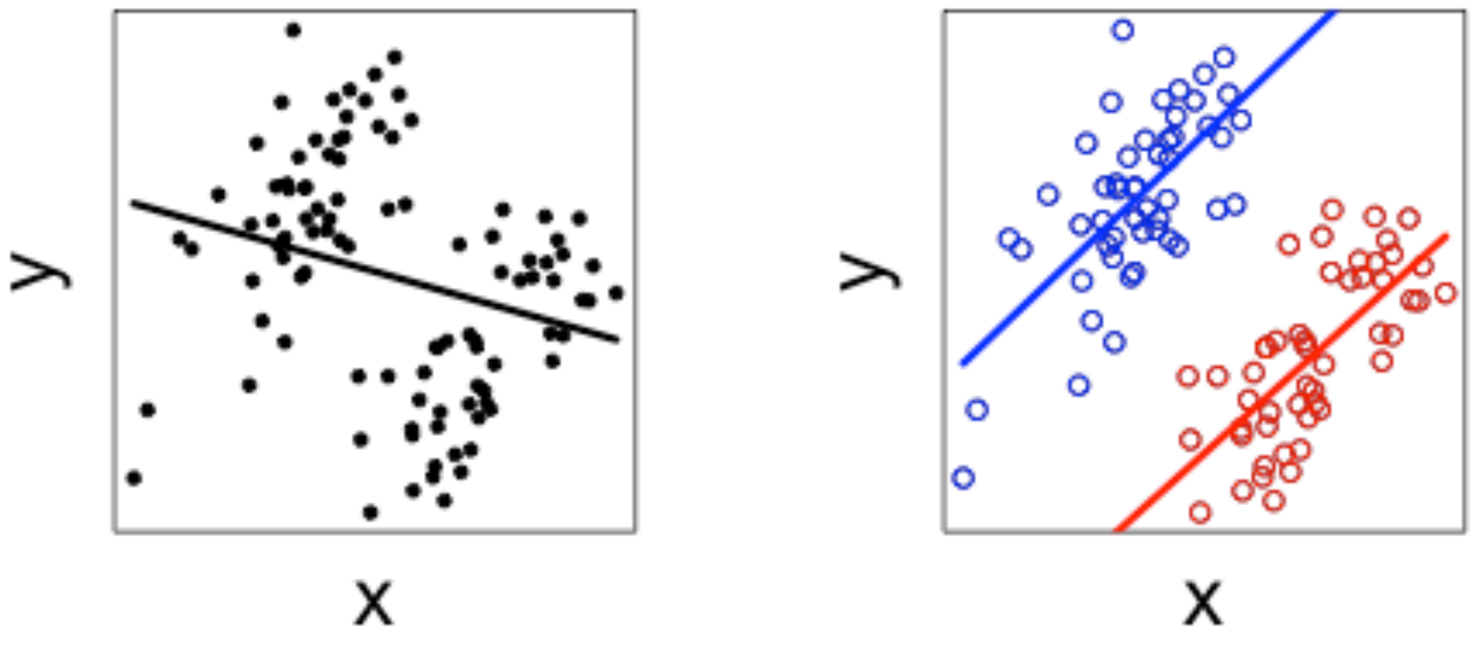


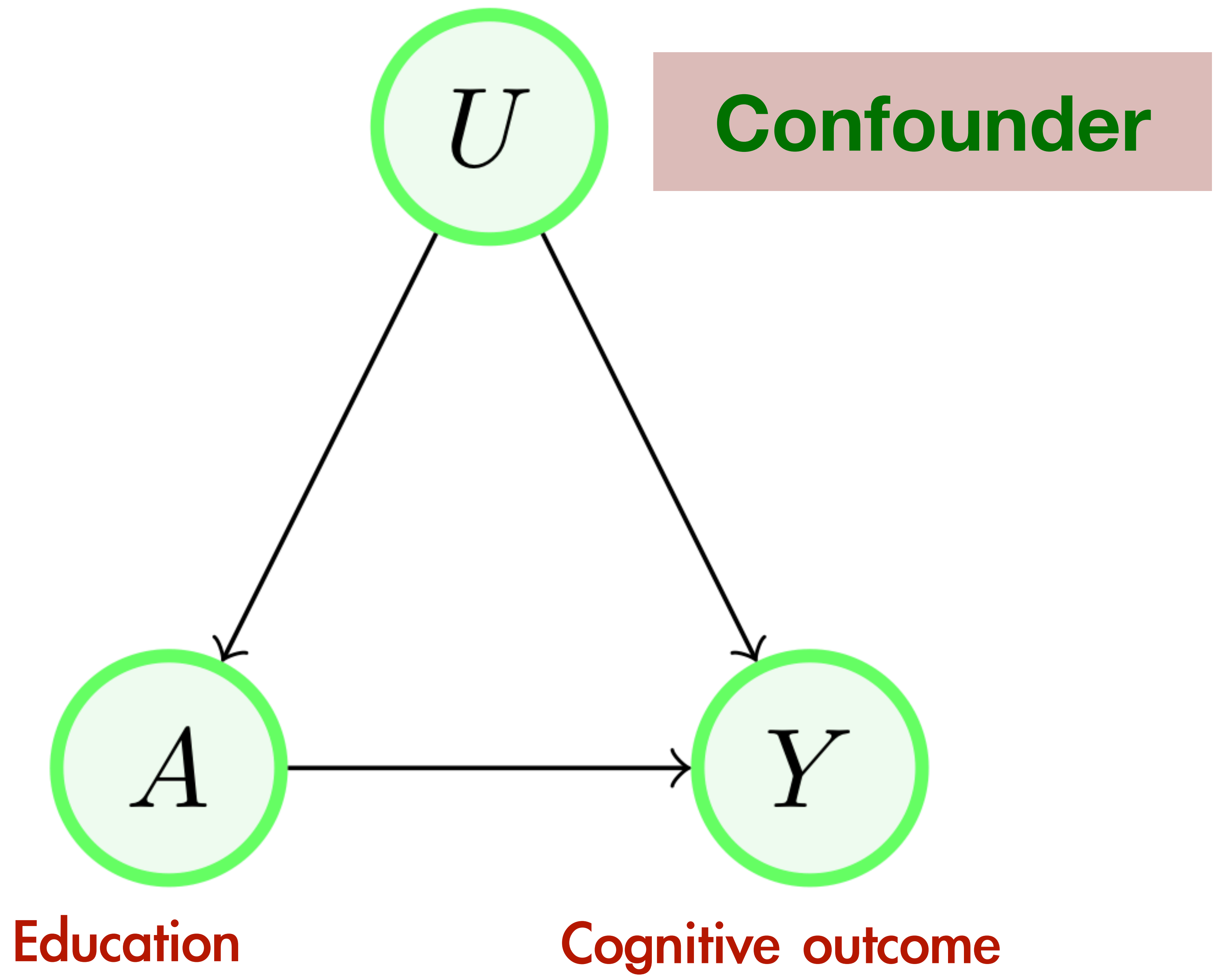
# Reliable Decision Making



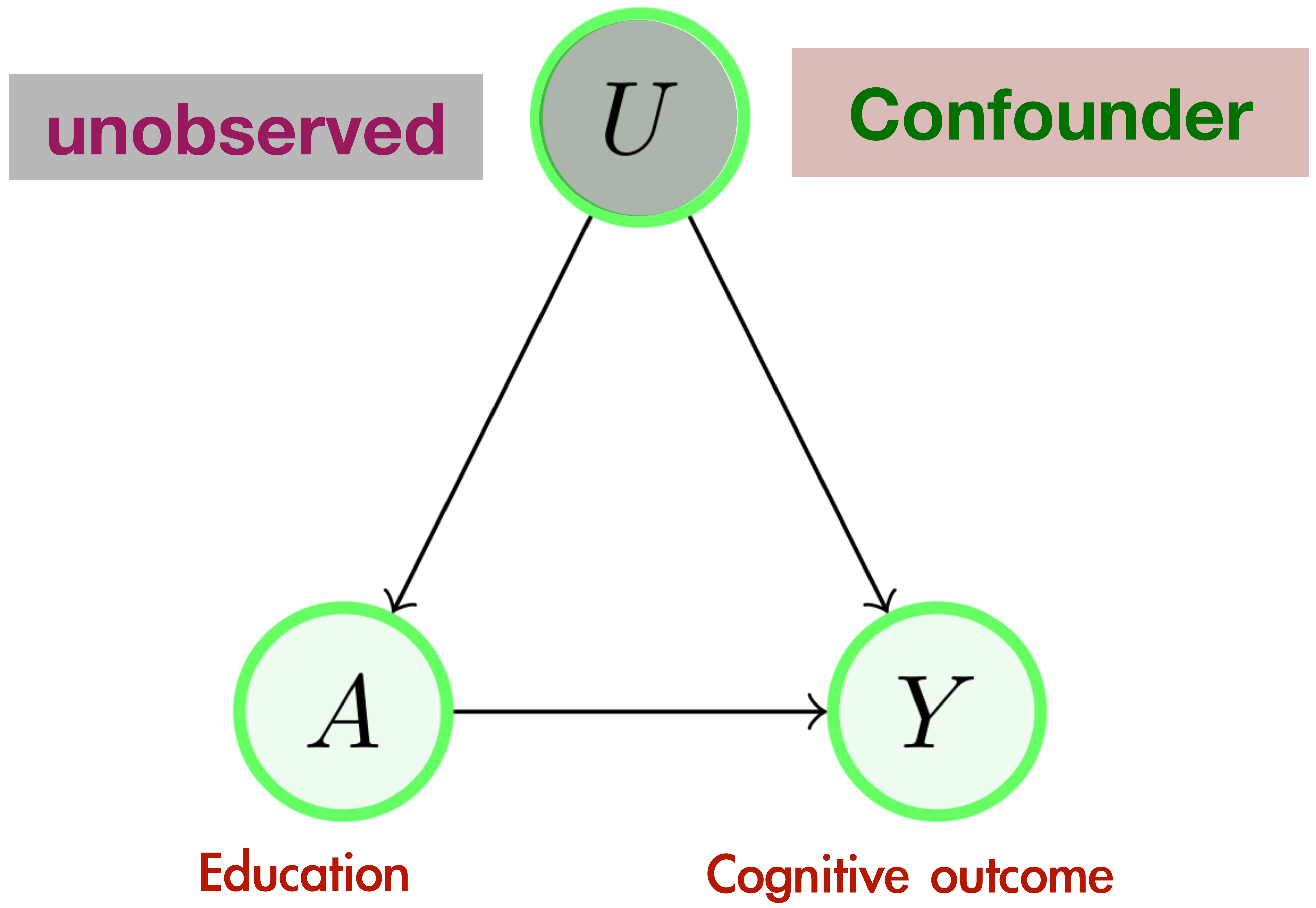

- Machine learning allows us to create models that excel at making **prediction**.
- We aim to predict an outcome of some **intervention**.
    - Will holding back a grade, improve the test scores of students?
- **Experimental** data are not available. Only **Observational**.
- We cannot rule out the effect of **unobserved confounders** (Simpson's paradox)



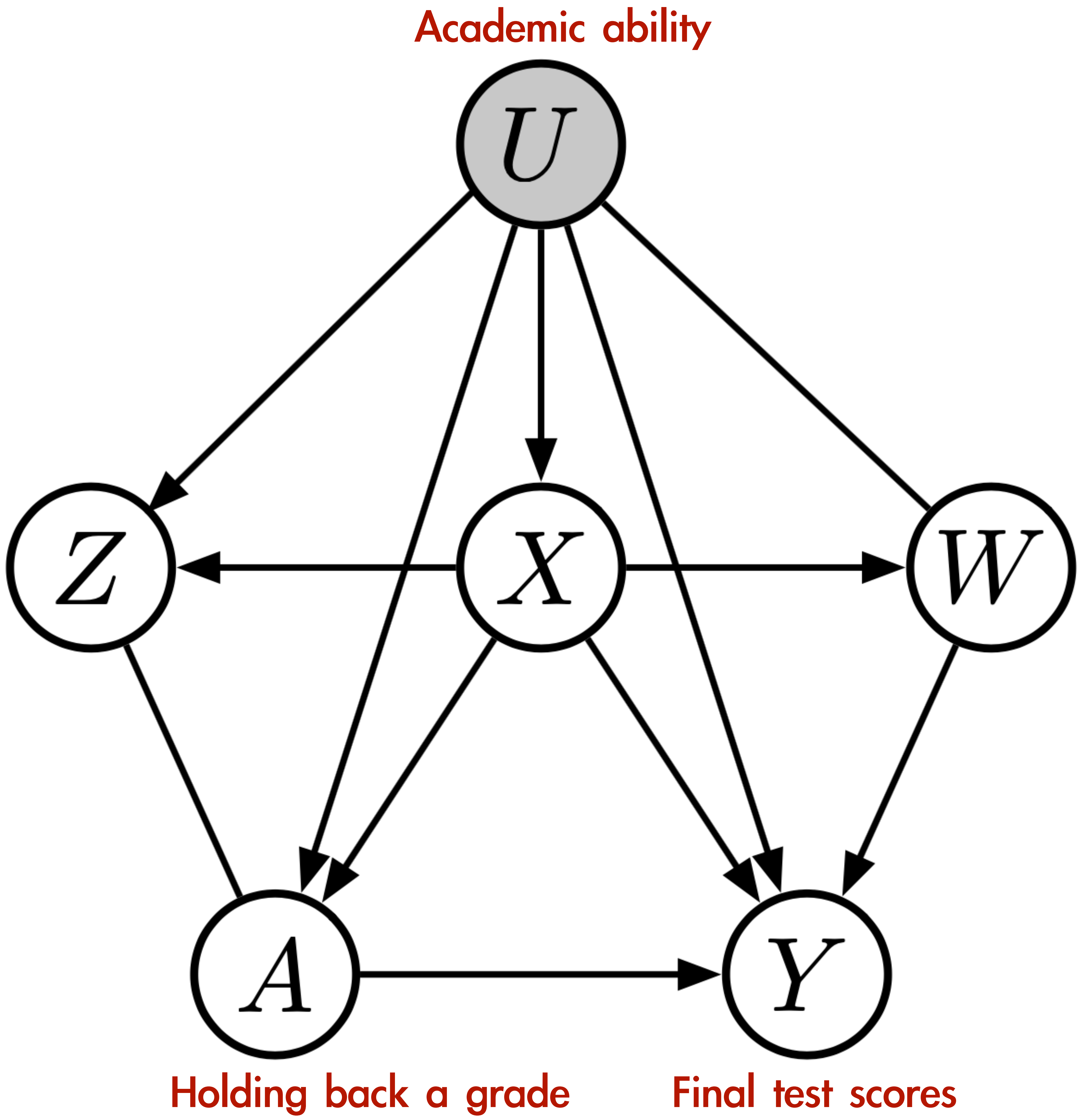
# How to resolve unobserved confounding?



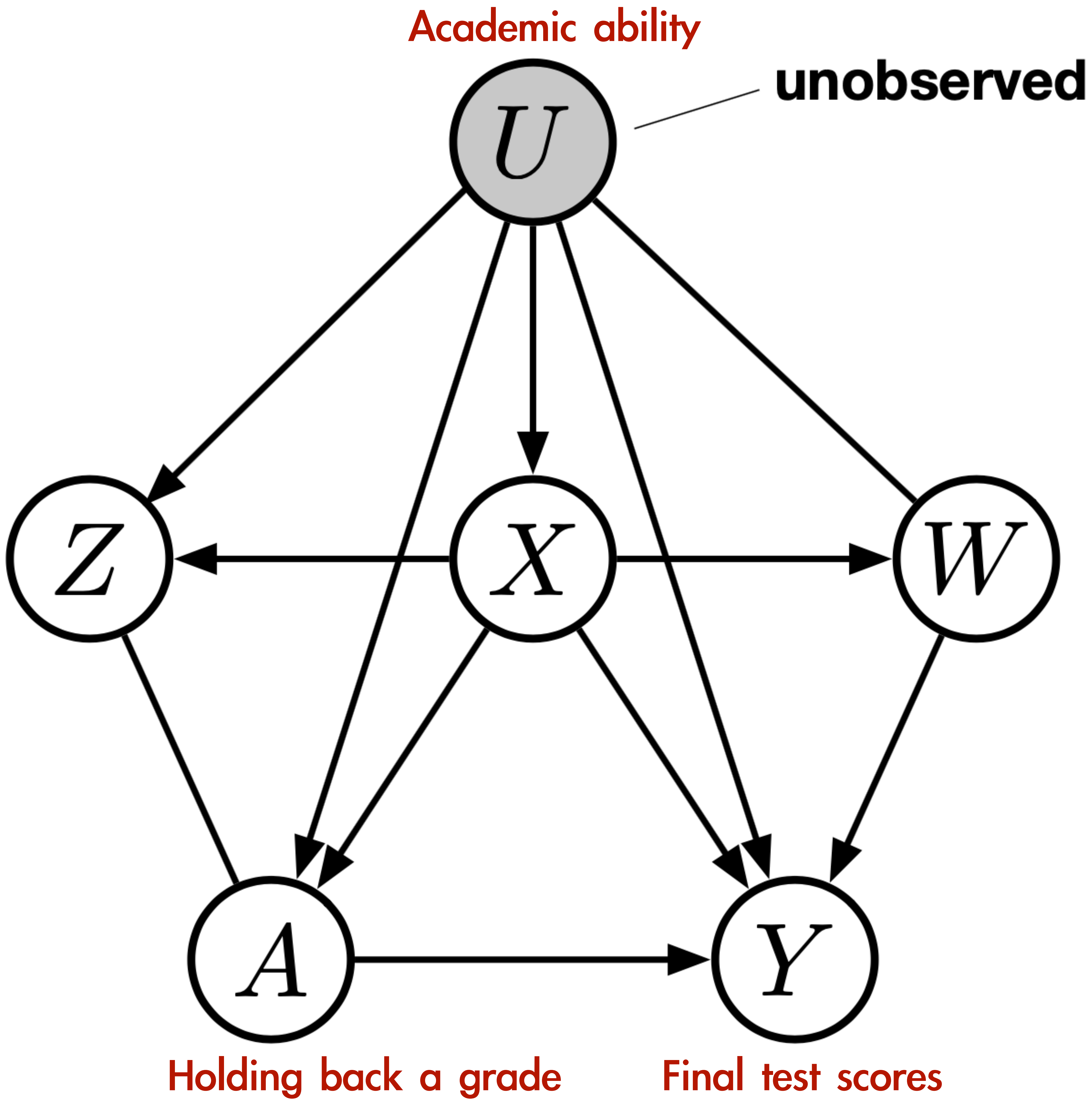
# How to resolve unobserved confounding?



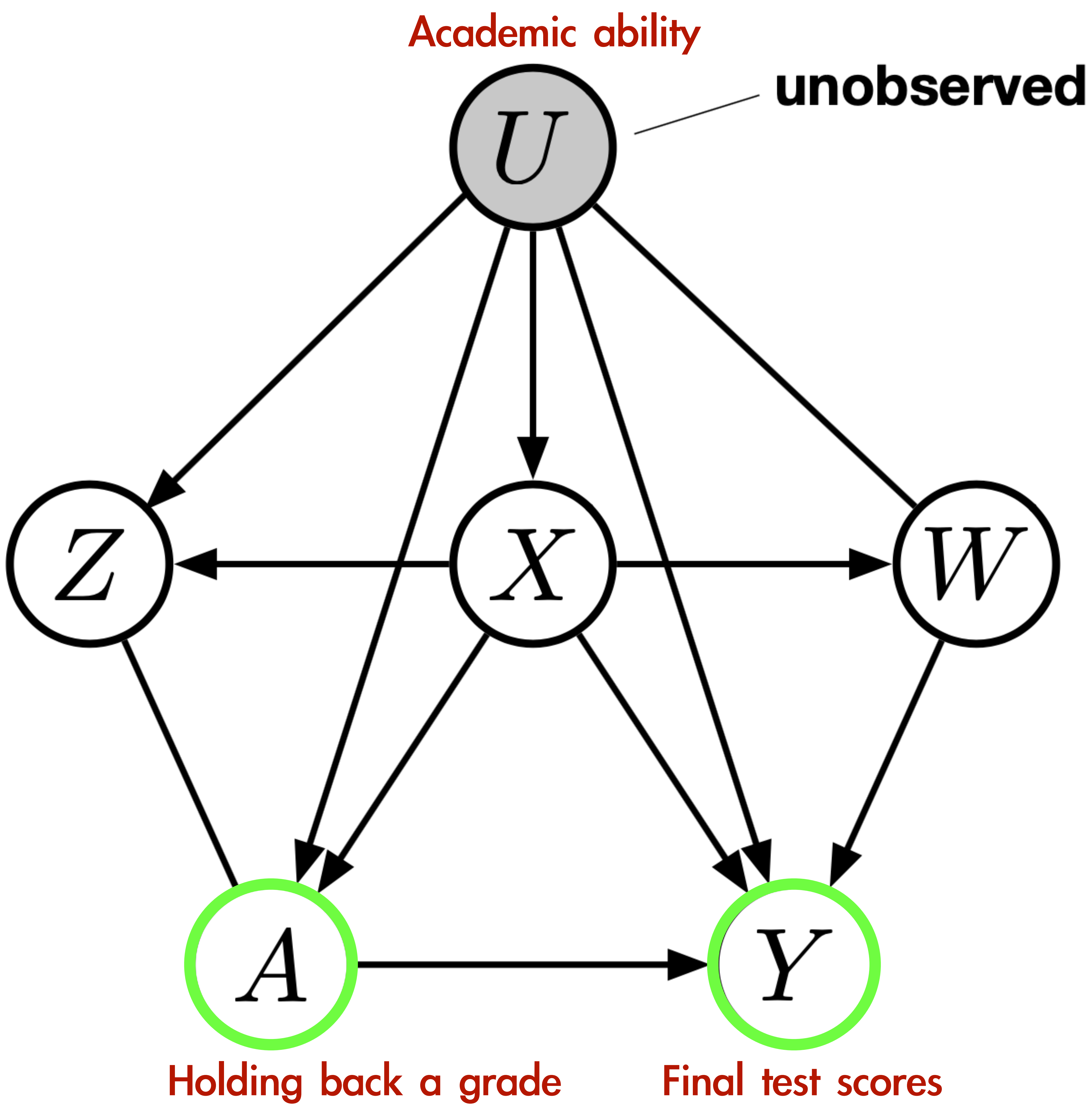
# How to resolve unobserved confounding?



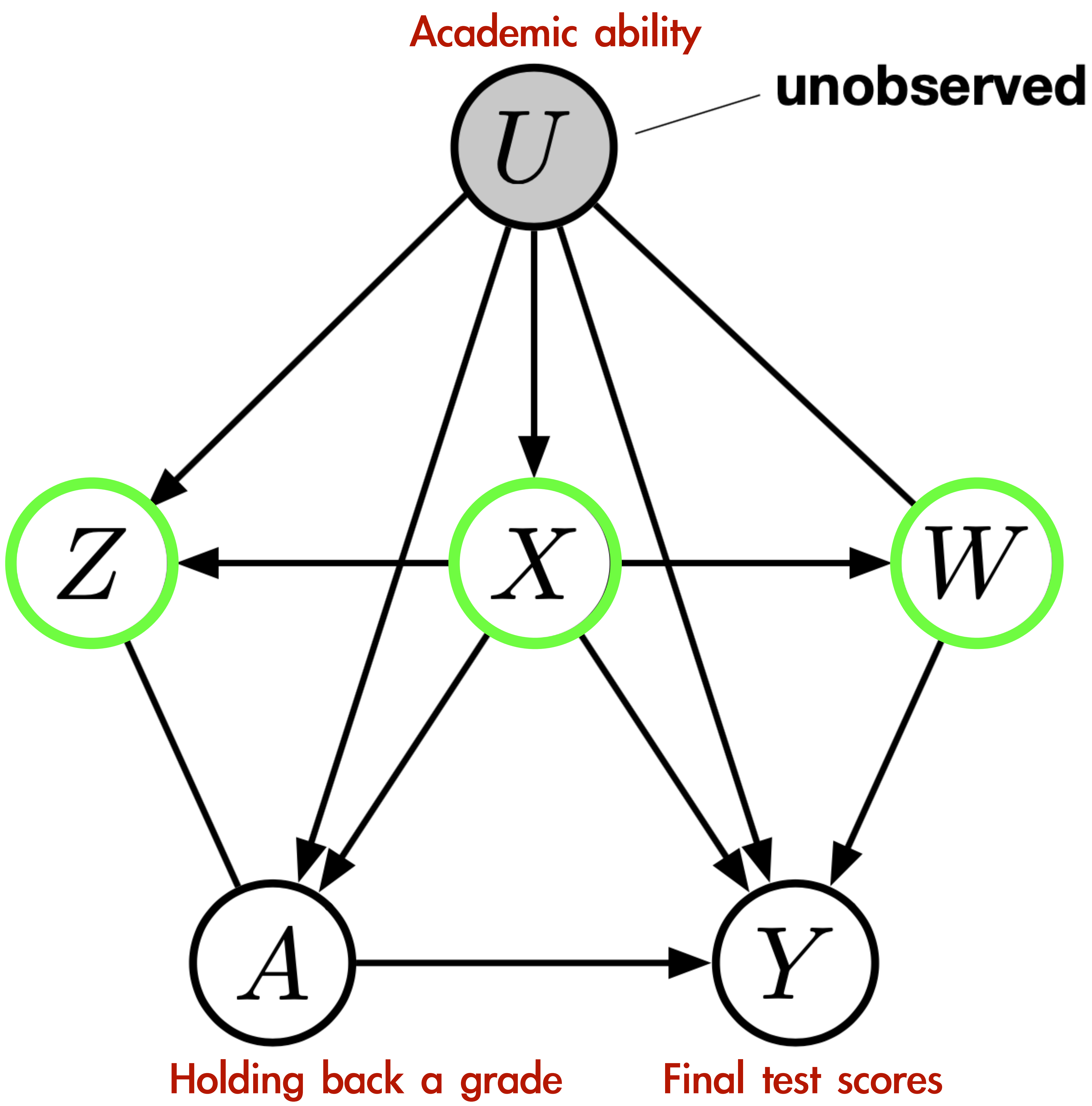
# How to resolve unobserved confounding?



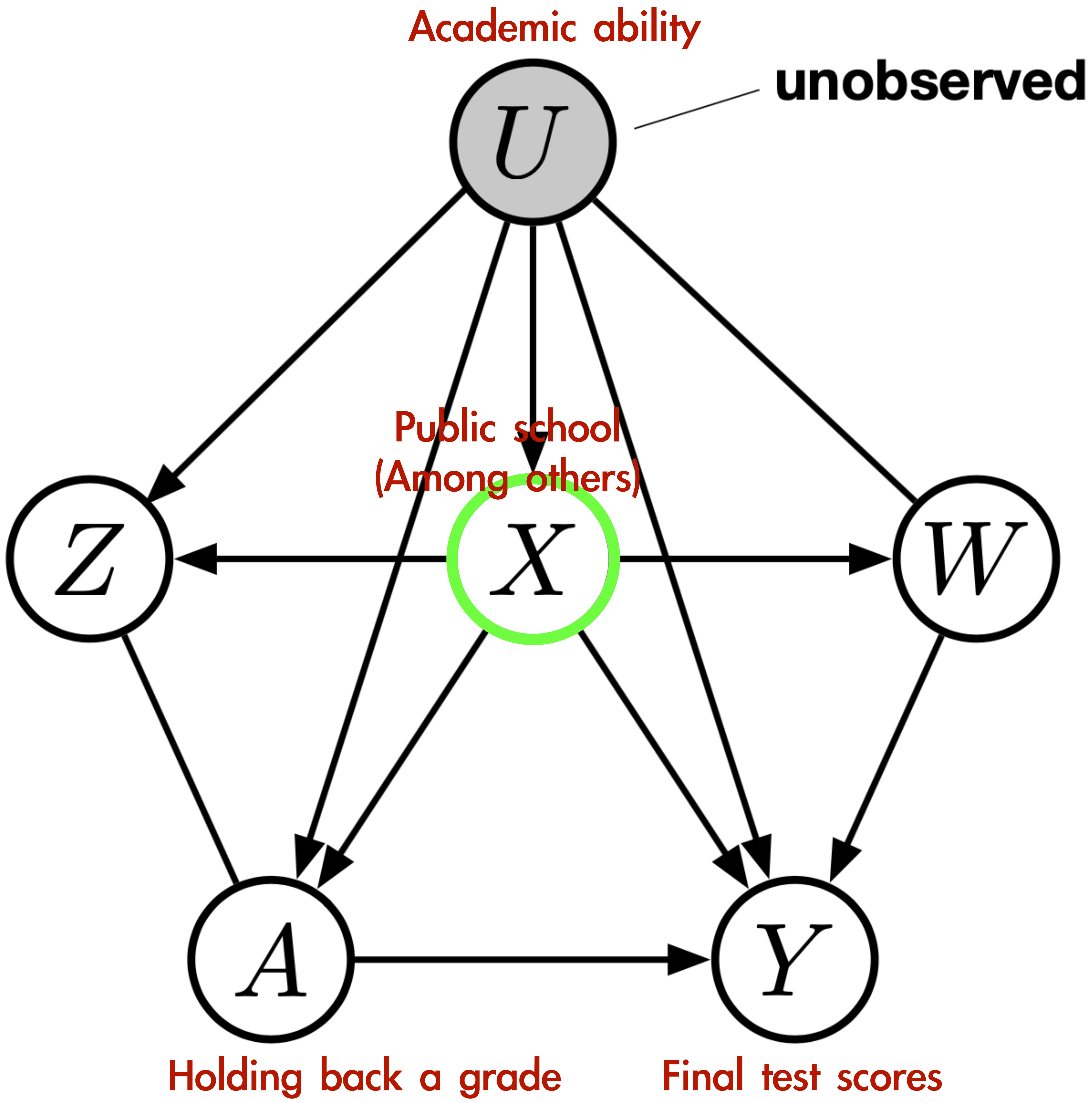
# How to resolve unobserved confounding?



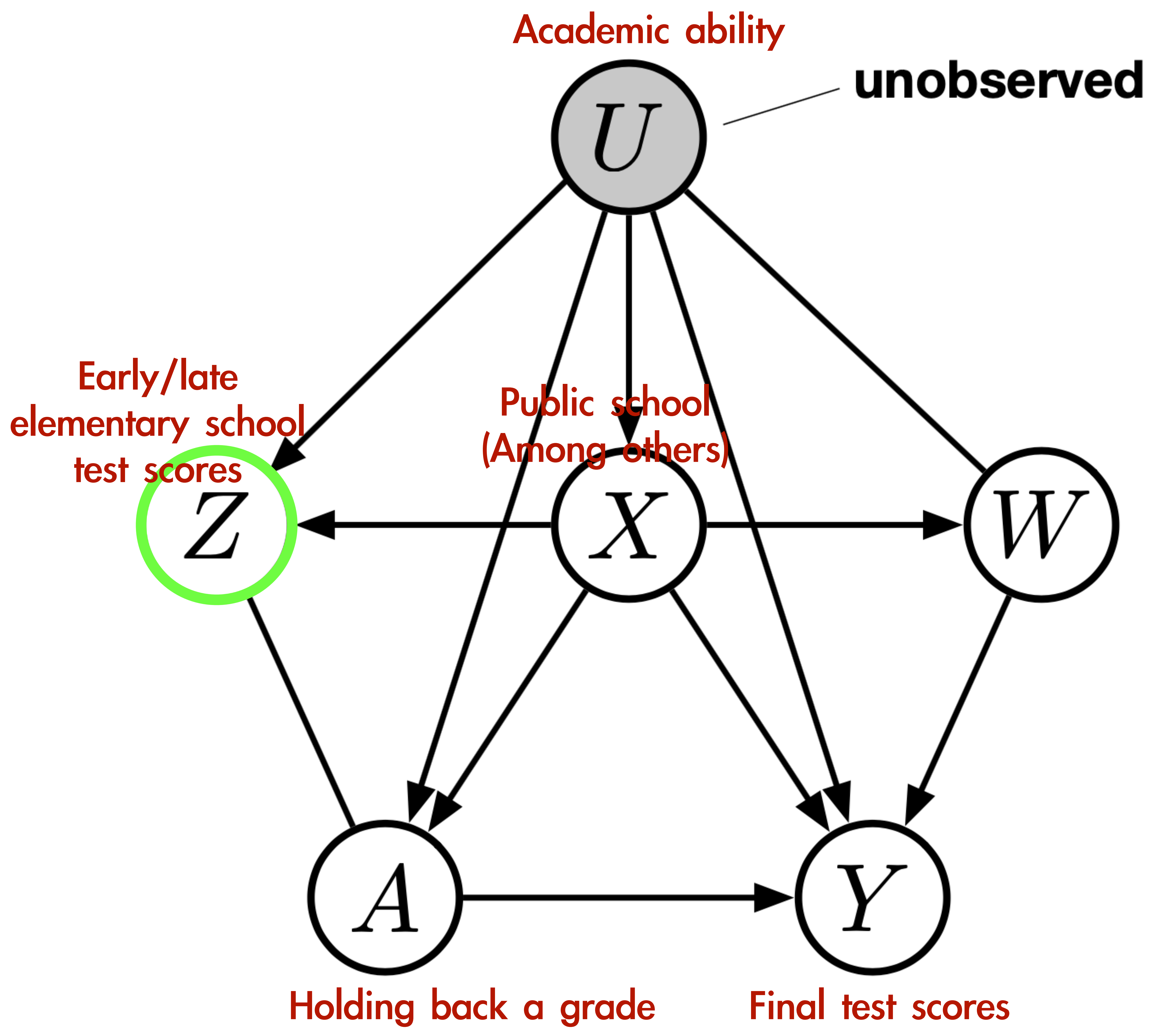
# How to resolve unobserved confounding?



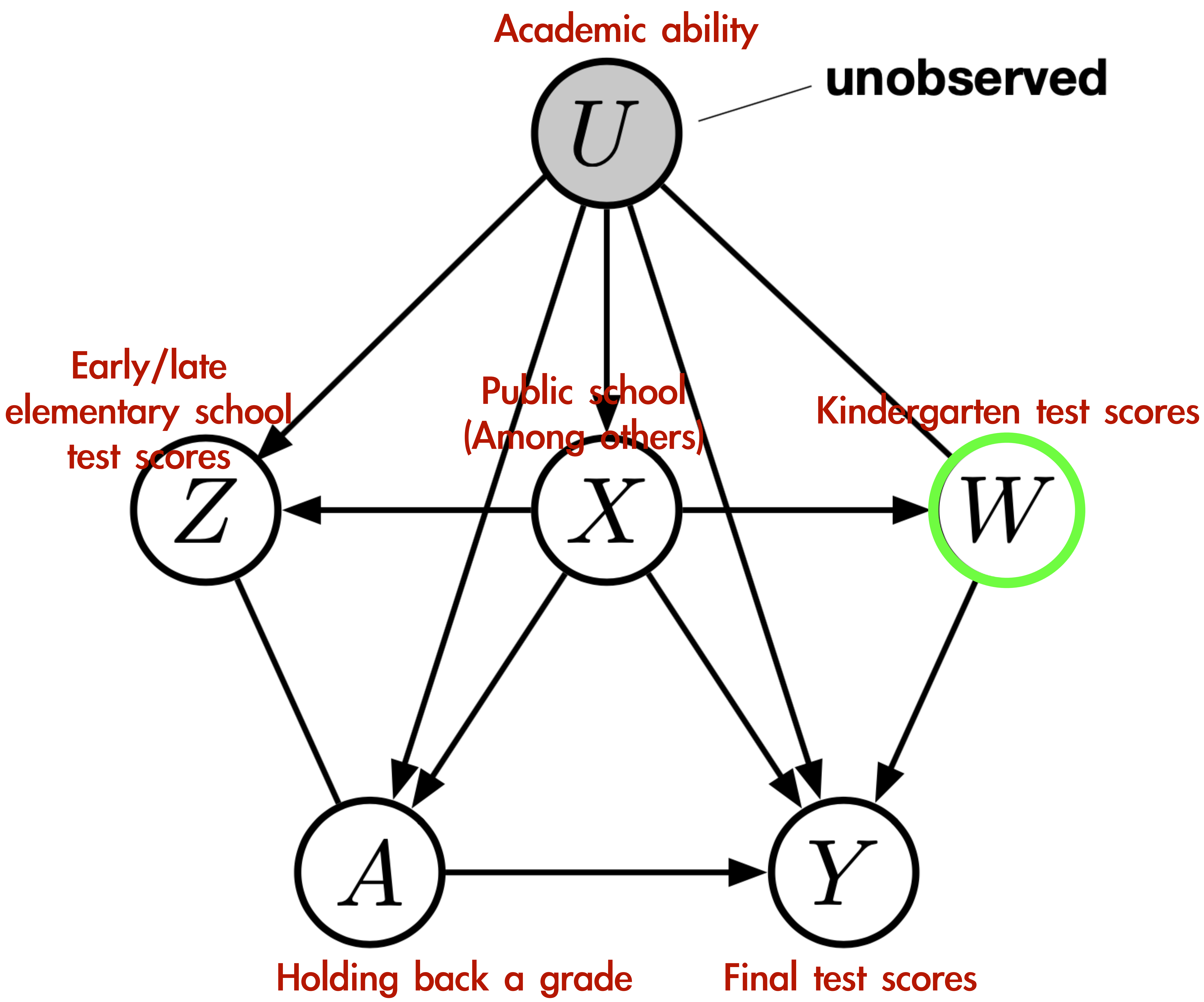
# How to resolve unobserved confounding?



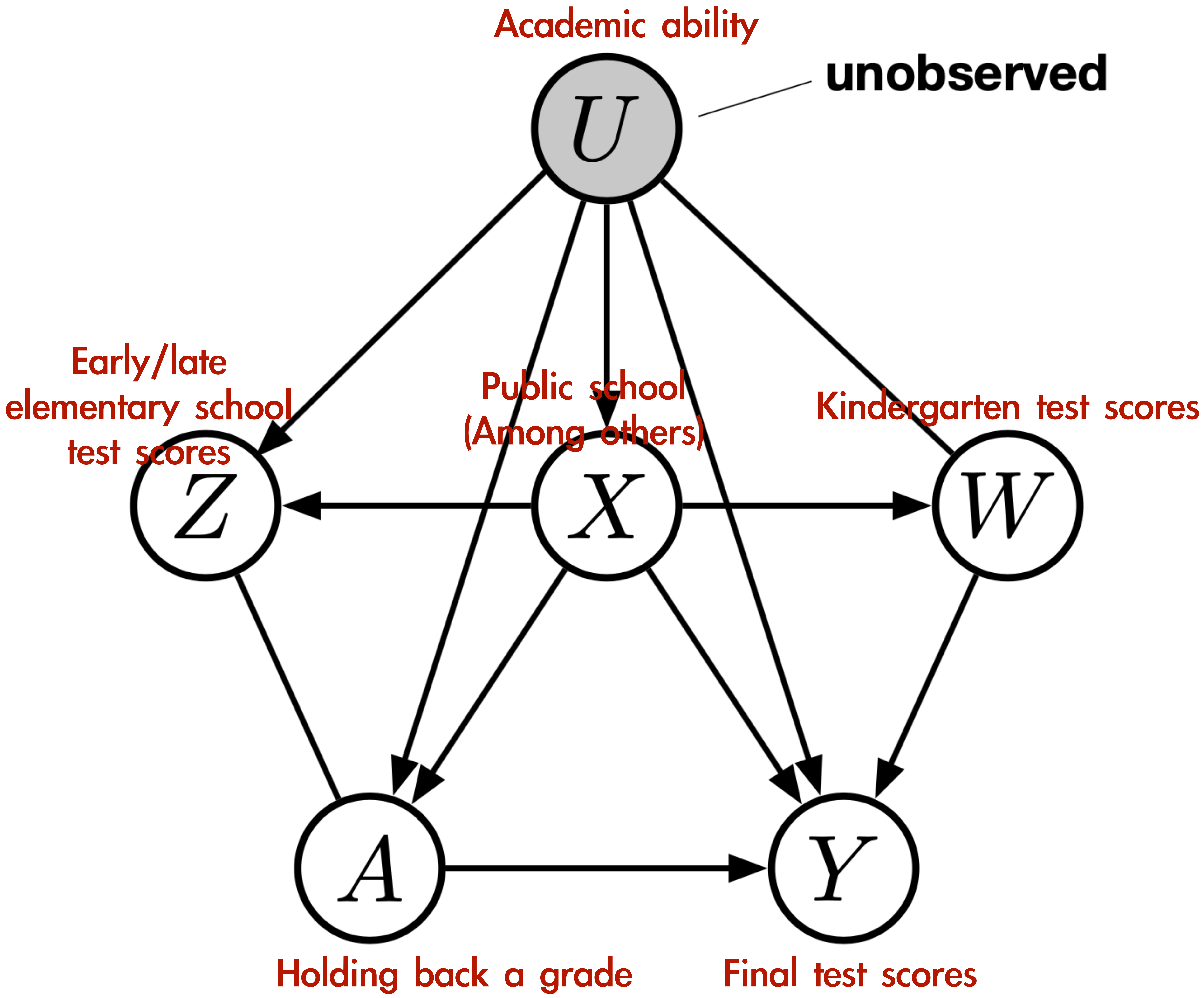
# How to resolve unobserved confounding?



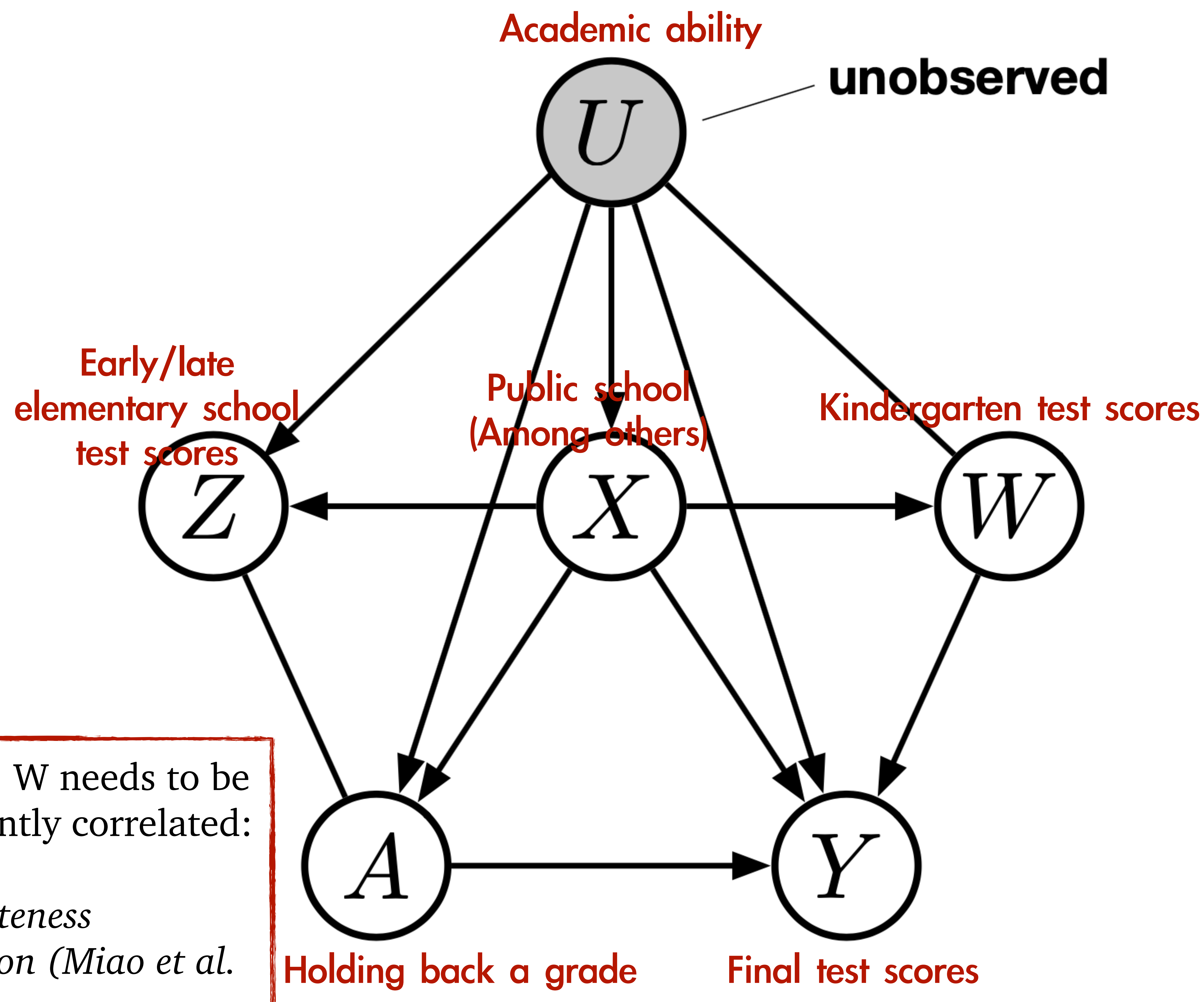
# How to resolve unobserved confounding?



# How to resolve unobserved confounding?



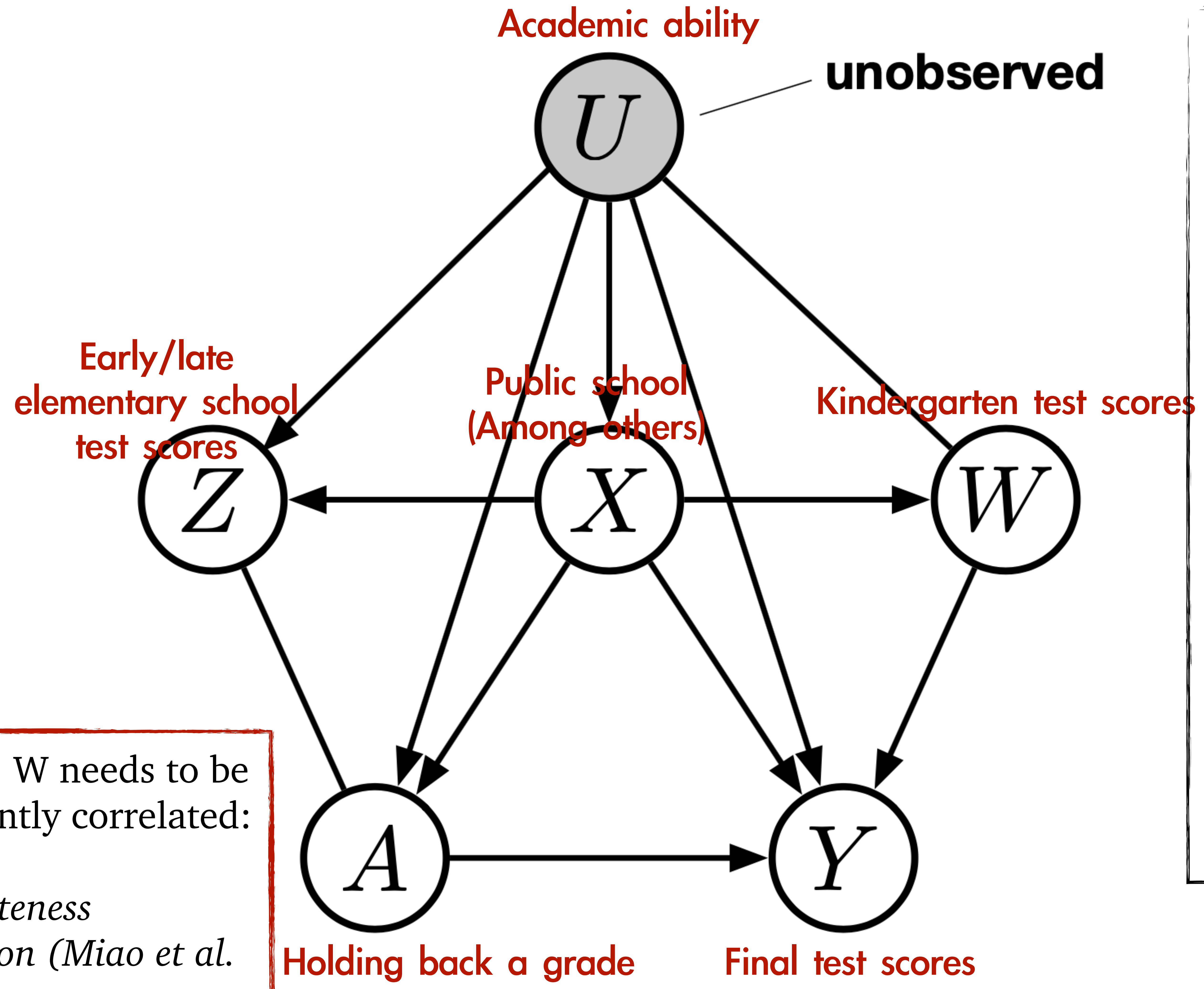
# How to resolve unobserved confounding?




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

*Completeness Condition (Miao et al. 2018)*

# How to resolve unobserved confounding?



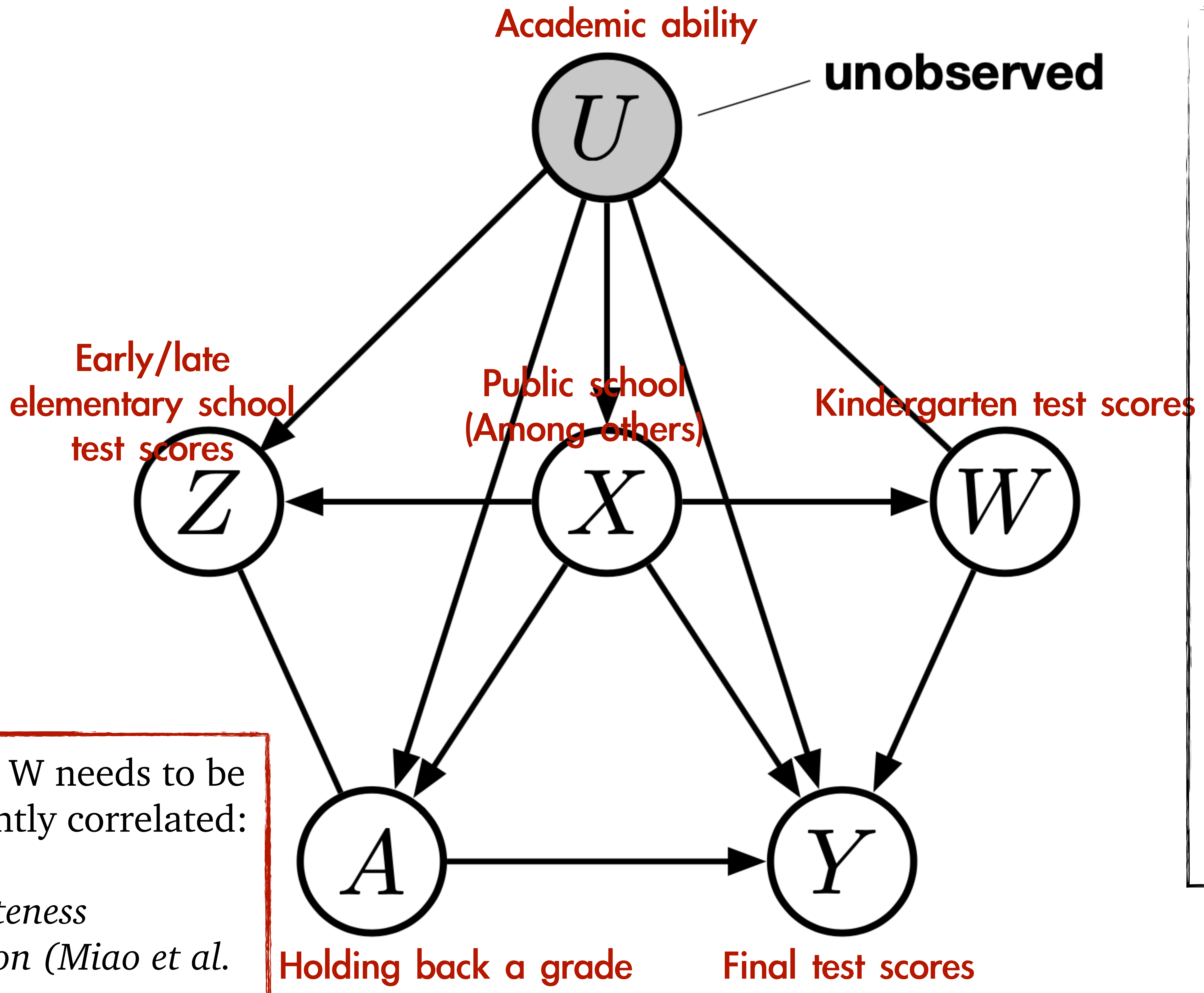

U, X, Z, W needs to be sufficiently correlated:

*Completeness Condition (Miao et al. 2018)*

**Average causal effect estimation:**

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

# How to resolve unobserved confounding?




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

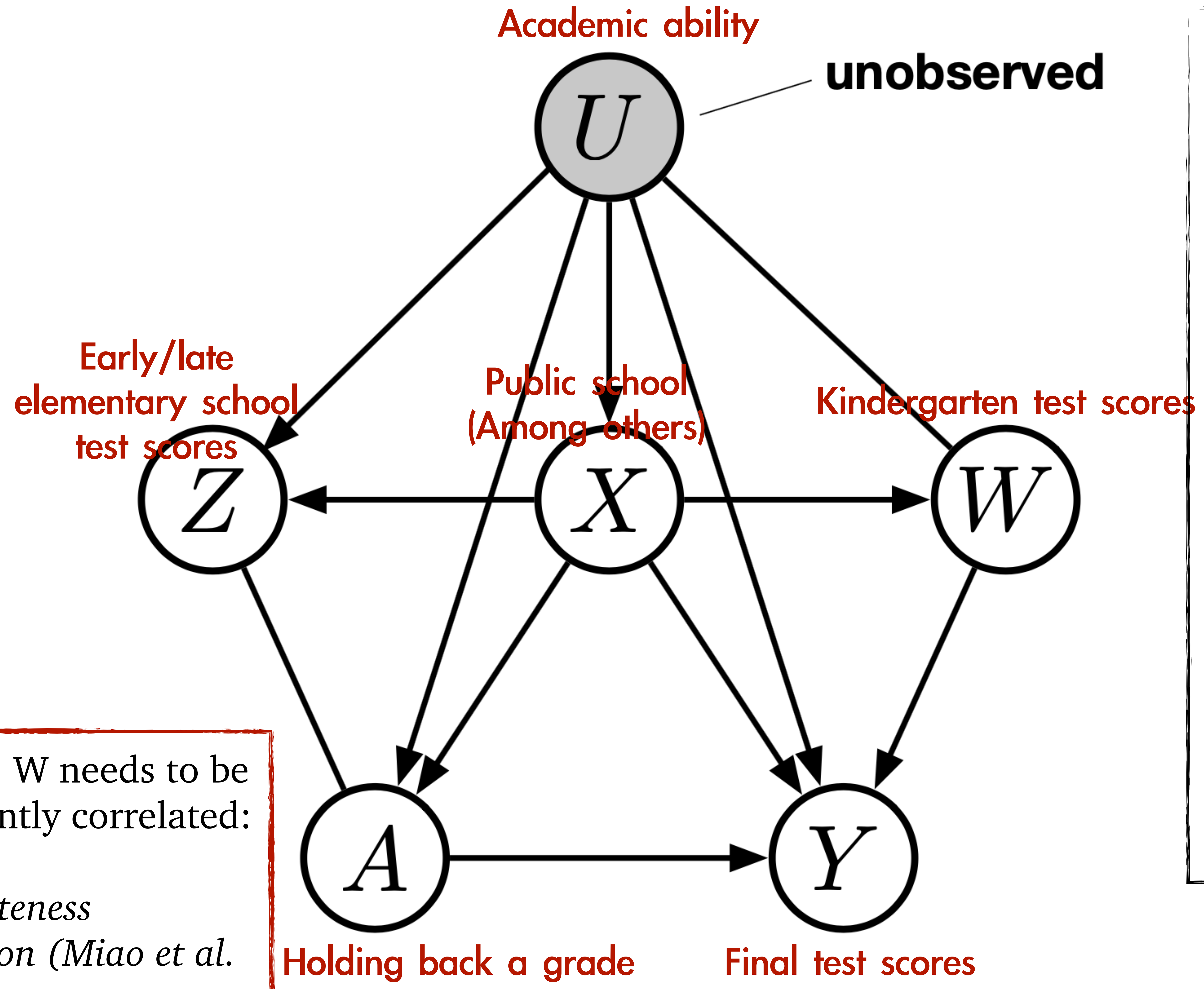
*Completeness Condition (Miao et al. 2018)*

**Average causal effect estimation:**

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

**How to get h?**

# How to resolve unobserved confounding?




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

*Completeness Condition (Miao et al. 2018)*

**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?**

$$\Downarrow$$

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

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

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

**Proximal setting characteristic equation**

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

**Kernel Proxy Variable (KPV)**

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

**Kernel Proxy Variable (KPV)**

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)}h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A,X,W) \mid A,X,Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A,X,W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A,X,Z) \xrightarrow{f} \phi(W)$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

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

Proximal setting characteristic equation

True Loss

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

characteristic equation

$$\Updownarrow$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A,X,W) \mid A,X,Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A,X,W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A,X,Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A,X,W) = \gamma(A,X,Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

$$\Updownarrow$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

$$\Updownarrow$$

**CMR**

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

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

$$\Updownarrow$$

**CMR**

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

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A, X, W) \mid A, X, Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A, X, W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A, X, Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A, X, W) = \gamma(A, X, Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

$$\Updownarrow$$

**CMR**

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

PMMR surrogate loss $R_k(h)$

$$R_k(h) = \| \mathbb{E}[(Y - h(A, W, X))k((A, Z, X), .)] \|^2_{\mathcal{H}_{AXZ}}$$

$$\arg\min R(h) = \arg\min R_k(h)$$

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

**Proximal setting characteristic equation**

**True Loss**

$$\min_h \; R(h) := \mathbb{E}_{AXZ}[(\mathbb{E}[Y - h(A,X,W) \mid A,X,Z])^2]$$

## Kernel Proxy Variable (KPV)

KPV surrogate loss

$$\tilde{R}(h) := \mathbb{E}_{AXYZ}[(Y - \mathbb{E}_{W|(A,X,Z)} h(A,X,W))^2]$$

$$R(h) \leq \tilde{R}(h)$$

$$Stage1 : (A,X,Z) \xrightarrow{f} \phi(W)$$

$$Stage2 : \mathbb{E}_{W|A,X,Z} h(A,X,W) = \gamma(A,X,Z)$$

## Proxy Maximum Moment Restriction(PMMR)

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

**characteristic equation**

$$\Updownarrow$$

**CMR**

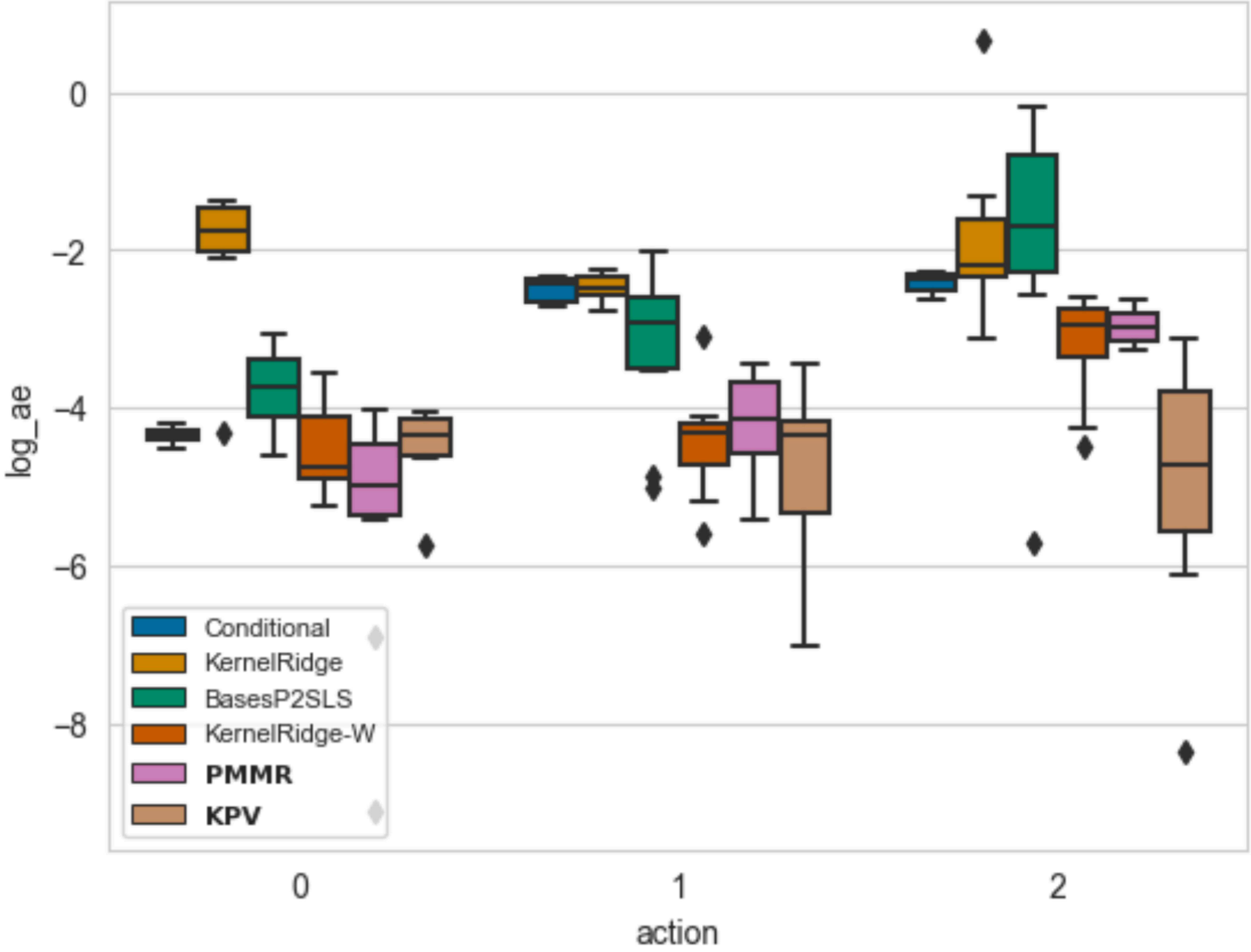
$$\mathbb{E}[(Y - h(A,X,W))k((A,X,Z),.)] = 0 \;\; \textbf{a.s. } P_{AXZ}$$

PMMR surrogate loss $R_k(h)$

$$R_k(h) = \| \, \mathbb{E}[(Y - h(A,W,X))k((A,Z,X),.)] \, \|^2_{\mathscr{H}_{AXZ}}$$

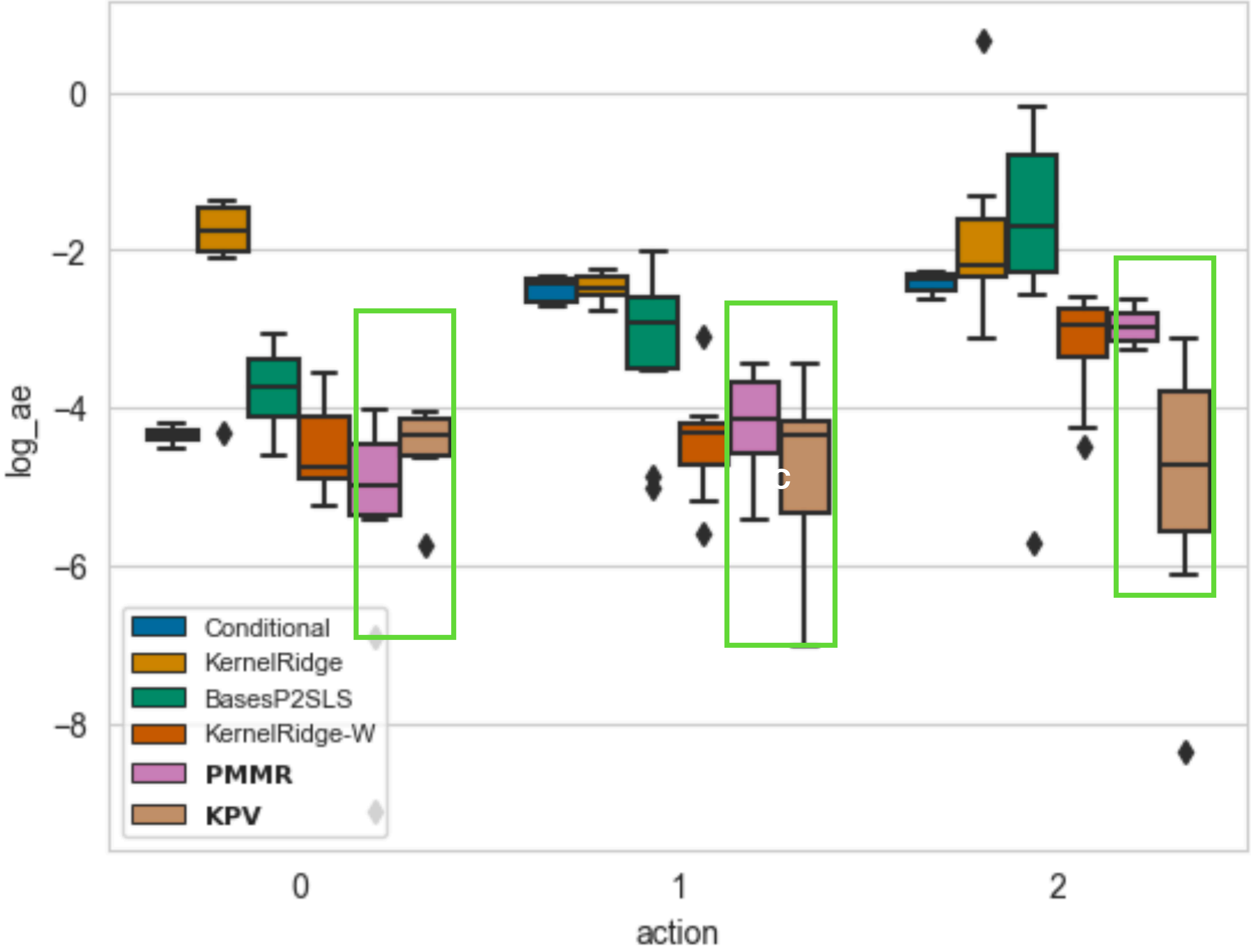
$$\arg\min R(h) = \arg\min R_k(h)$$

# Results




**Y: maths score**

# Results




Y: maths score