Causal Inference for Social Sciences

Yuchen Zhu

Collaborators









Overview

- What we want to achieve with causality.
- Why is causality suitable for social sciences?
- The characteristics of social science data.
- Algorithms.
  - Proximal causal learning with kernels.
  - Causal inference under treatment measurement error.
  - Generalised Robinson Decomposition.

Overview

- What we want to achieve with causality.
- Why is causality suitable for social sciences?
- The characteristics of social science data.
- Algorithms.
    - Proximal causal learning with kernels.
    - Causal inference under treatment measurement error.
    - Generalised Robinson Decomposition.

What we want to achieve with Causality?

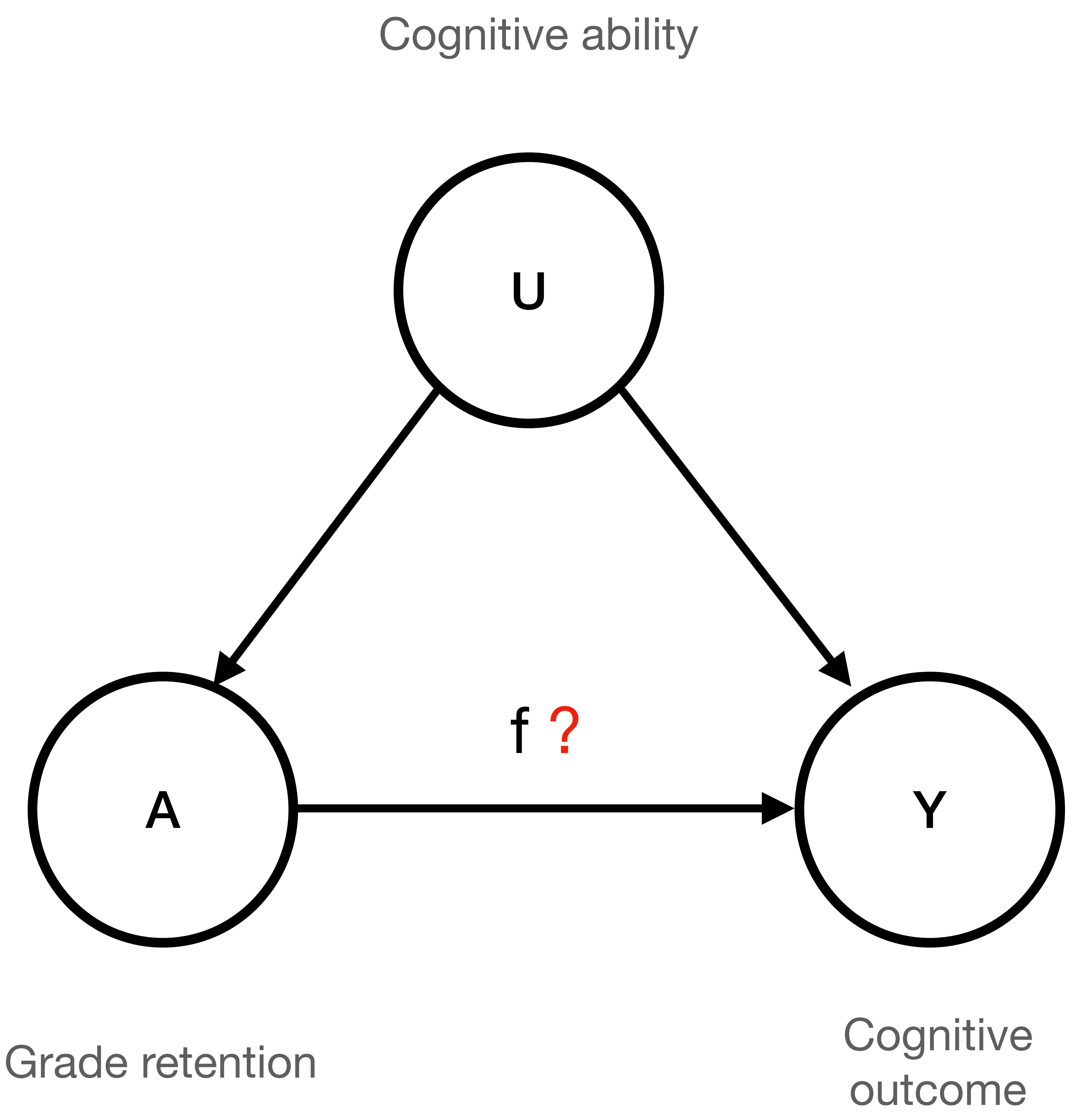
A metric to compare different actions with respect to their effects.

Overview

- What we want to achieve with causality.
- Why is causality suitable for social sciences?
- The characteristics of social science data.
- Algorithms.

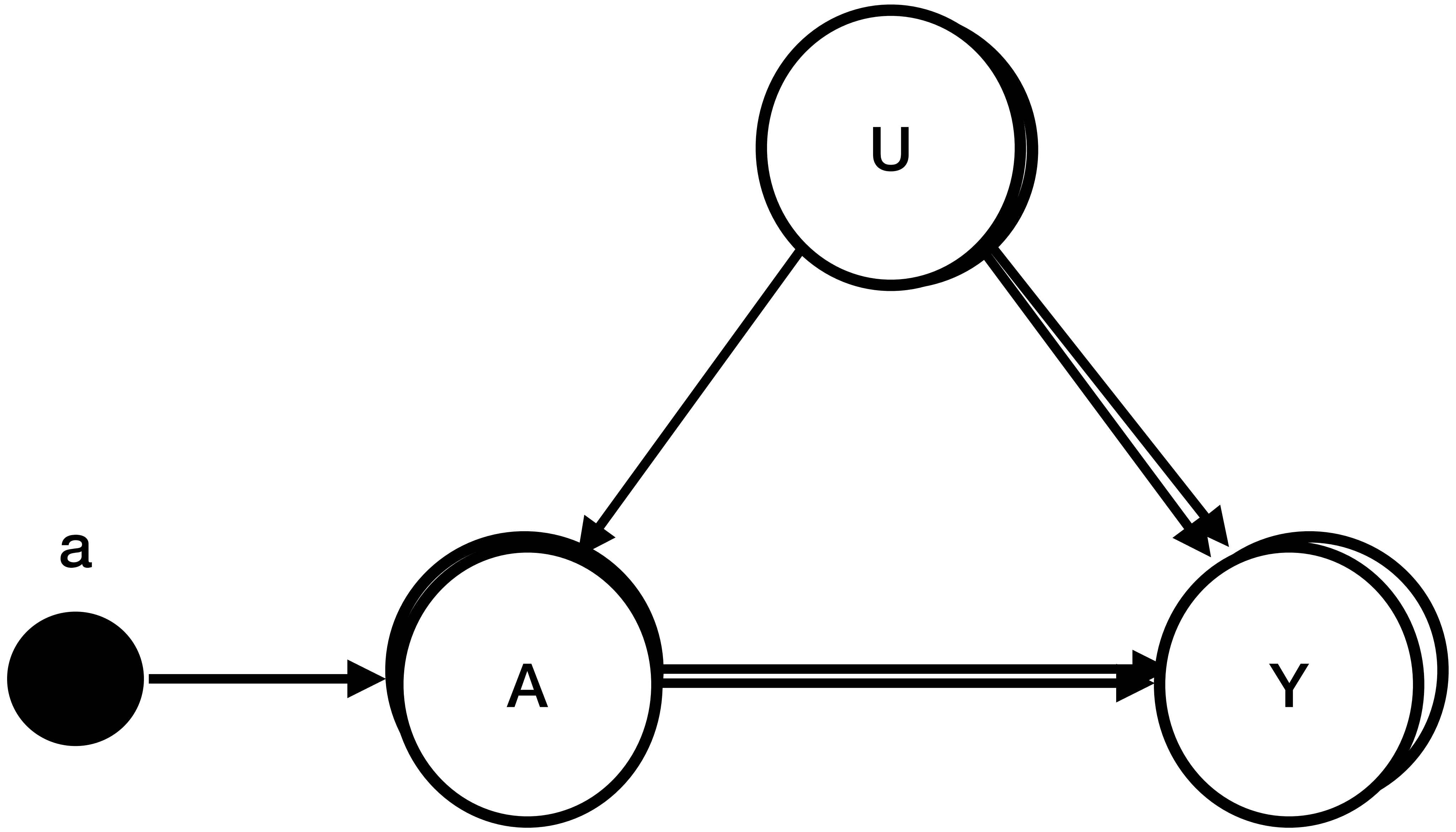
Why causal inference? An example.

Image source: Google image



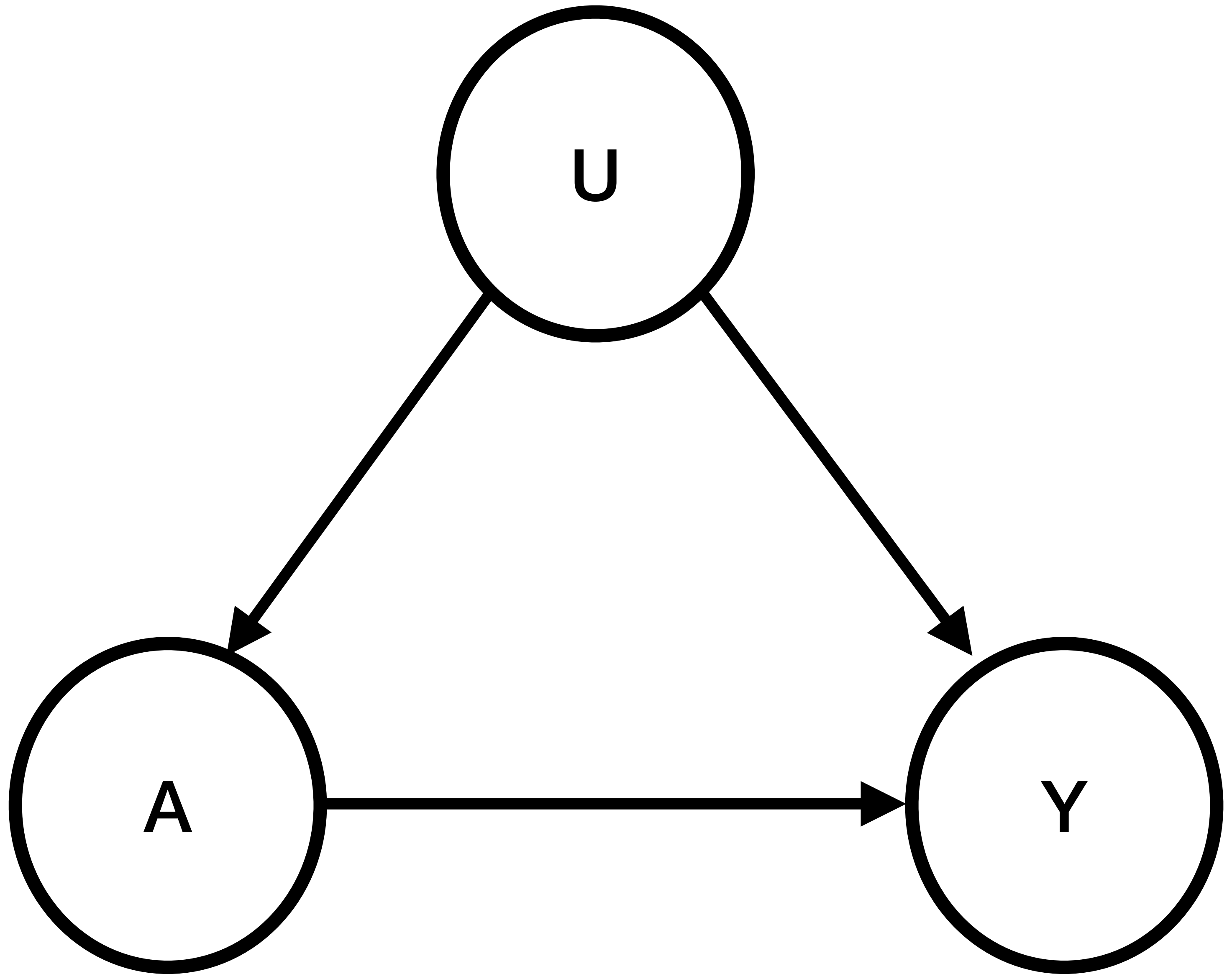
Fruehwirth, Navarro, Takahashi. 2016: How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects.

The target quantity



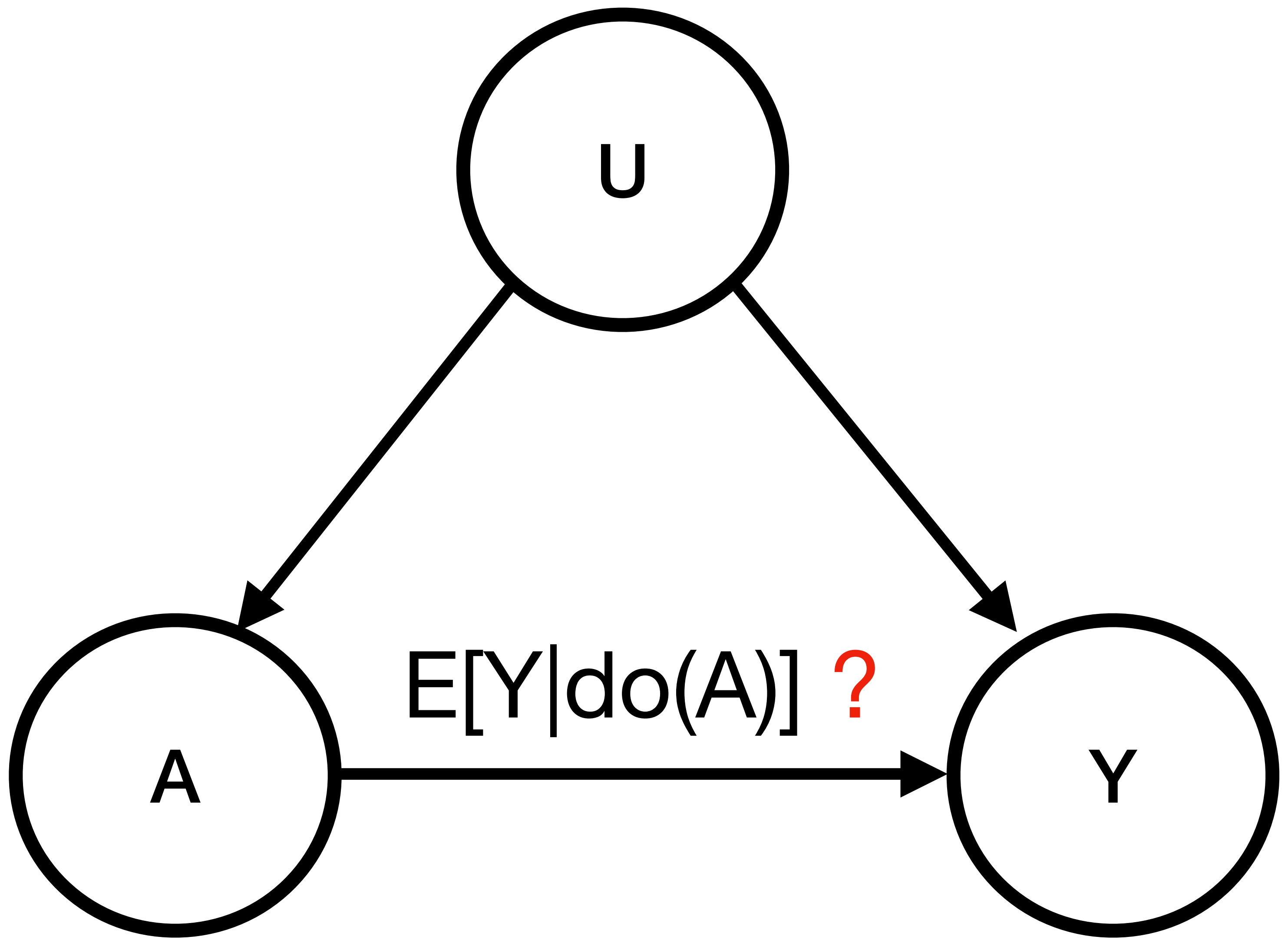
$E[Y|do(A := a)]$

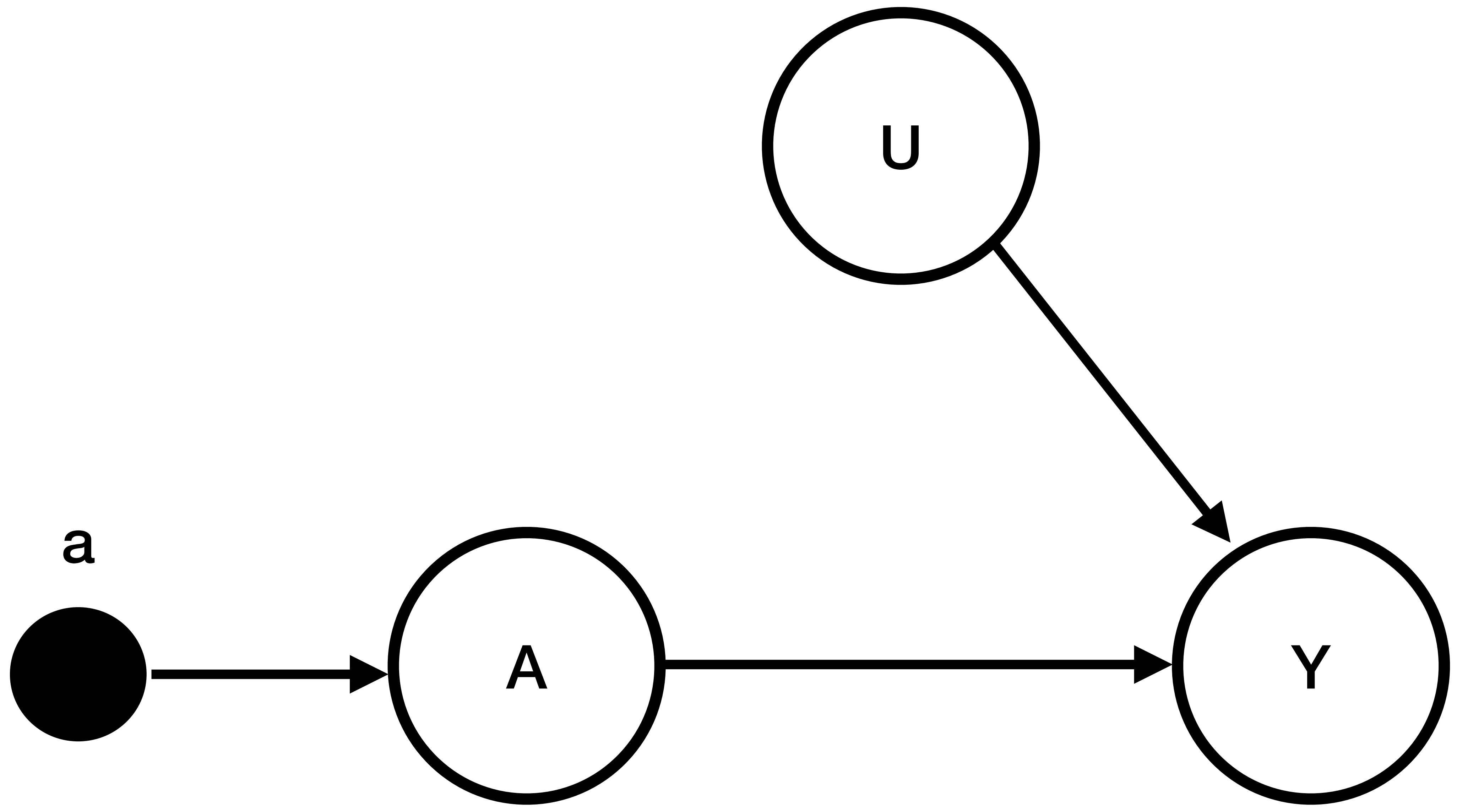
Observed data



Observed data

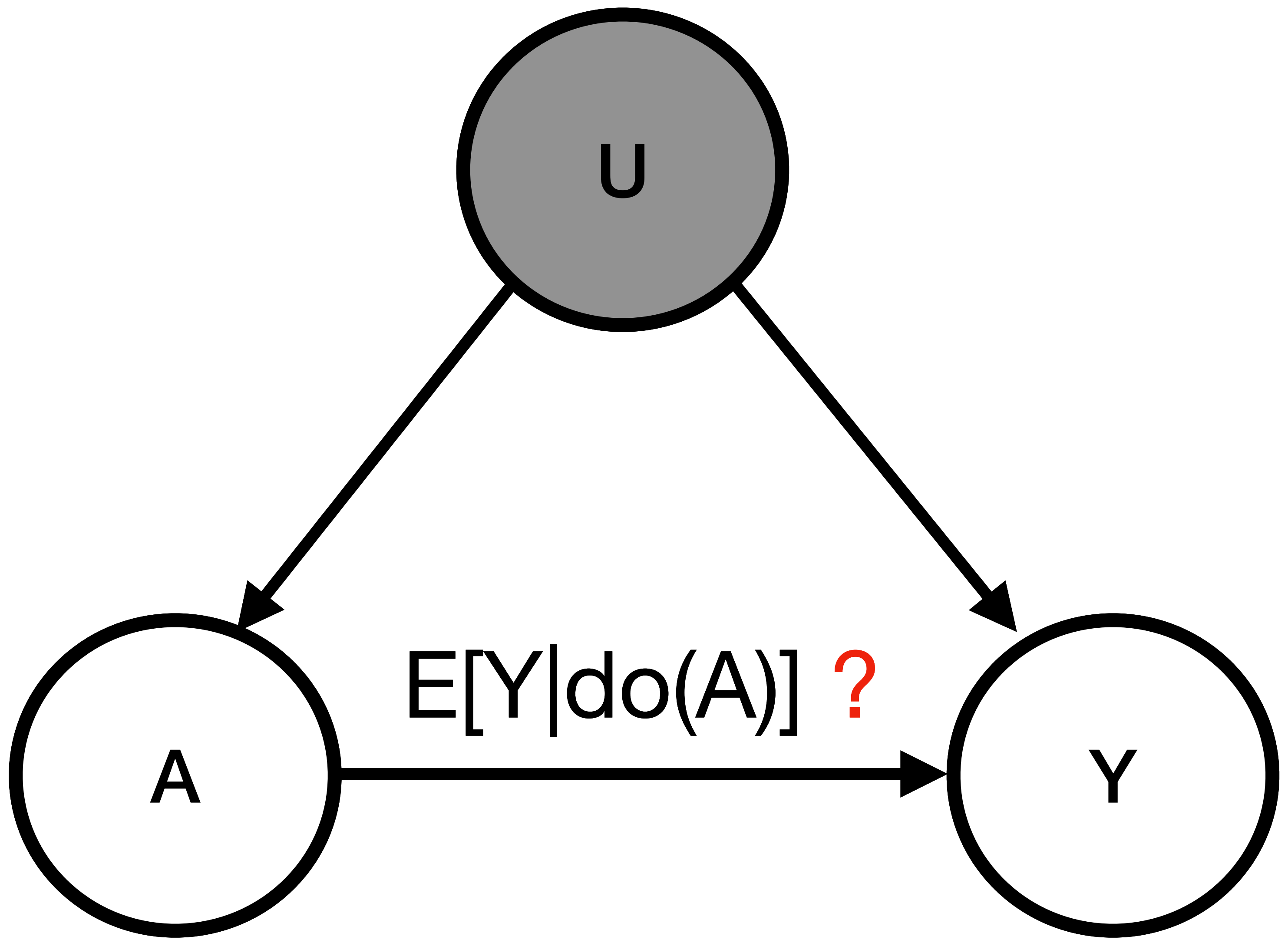
Warm-up: Observed confounders

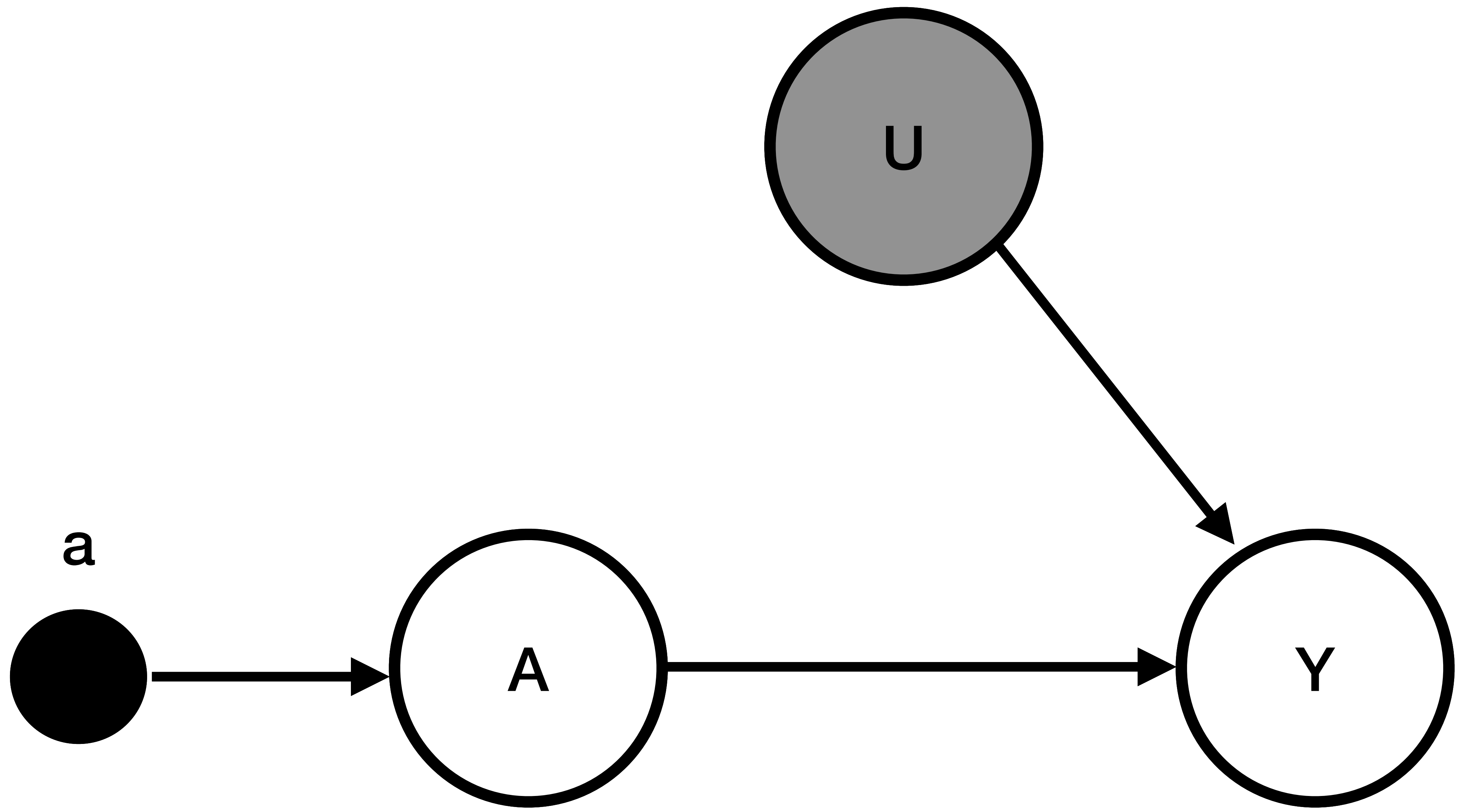




Backdoor adjustment: $\mathbb{E}[Y \,|\, do(a)] = \sum_{i=1}^{n} \mathbb{E}[Y \,|\, A = a, U = i]\mathbb{P}(U = i)$

Warm-up: Observed confounders

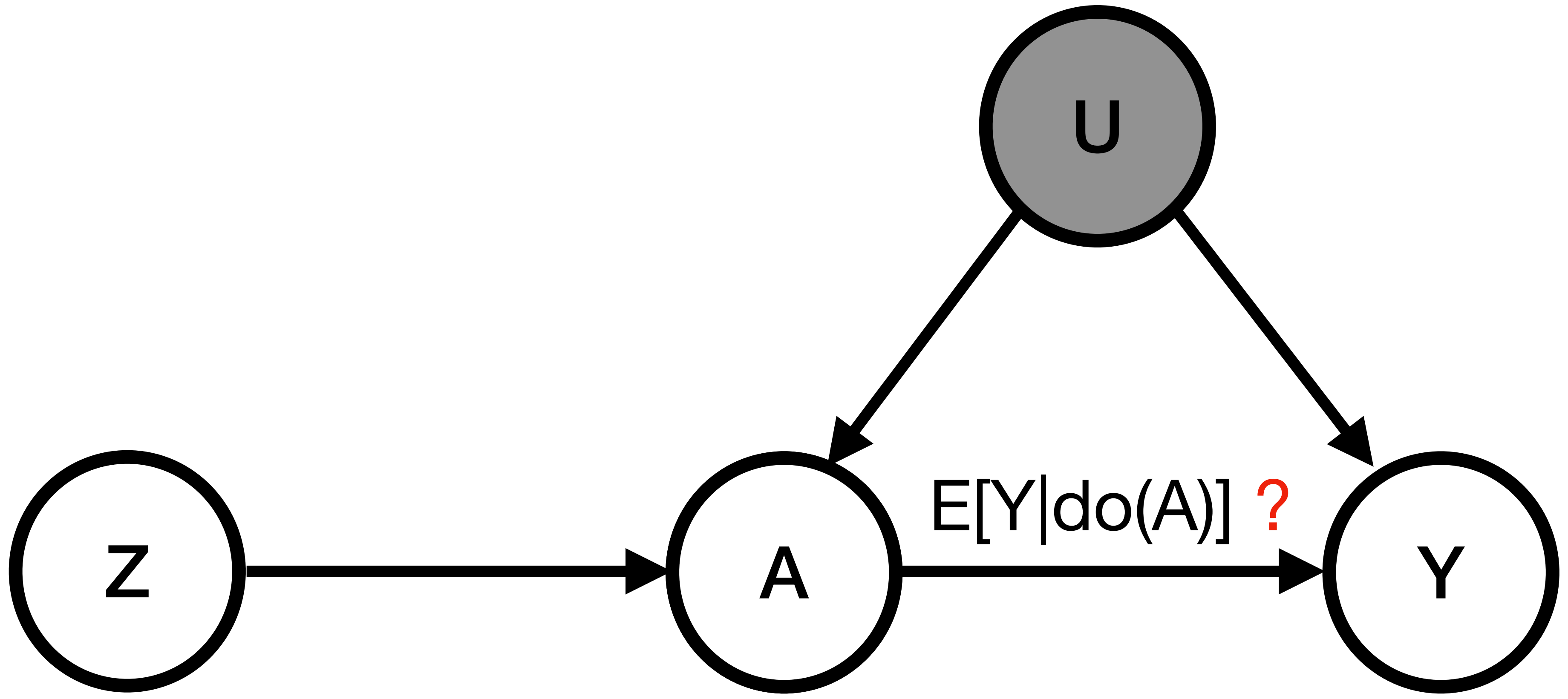




Backdoor adjustment: $\mathbb{E}[Y \mid do(a)] = \sum_{i=1}^{n} \mathbb{E}[Y \mid A = a, U = i]\mathbb{P}(U = i)$

Unobserved confounders?

Identification with instrumental variables



Identification:

$$Y = f(A) + U, \mathbb{E}[U] = 0, U \perp Z$$

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

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

Linear case:

$$Y = \beta A + \alpha_Y U \quad U \perp Z$$

$$A = \gamma Z + \alpha_A U \quad U \perp Z$$

$$\implies Y = \beta\gamma Z + (\beta\alpha_A + \alpha_Y)U$$

(Strong) Assumptions:

- **Additive error model**
- $(Z \not\perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

Relax the IV to allow for some dependence with U?

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

Why is causality suitable for social sciences?

- Social sciences often consider decision making for positive impact.
- High-stake domain so we should try to use observational data rather than perform adhoc experiments.
- Spurious correlations need to be corrected by causal algorithms.

But surely the scenarios described are unrealistic?

Overview

- What we want to achieve with causality.
- Why is causality suitable for social sciences?
- The characteristics of social science data.
- Algorithms.

The characteristics of social science data

- Observational data generated from unknown graph.

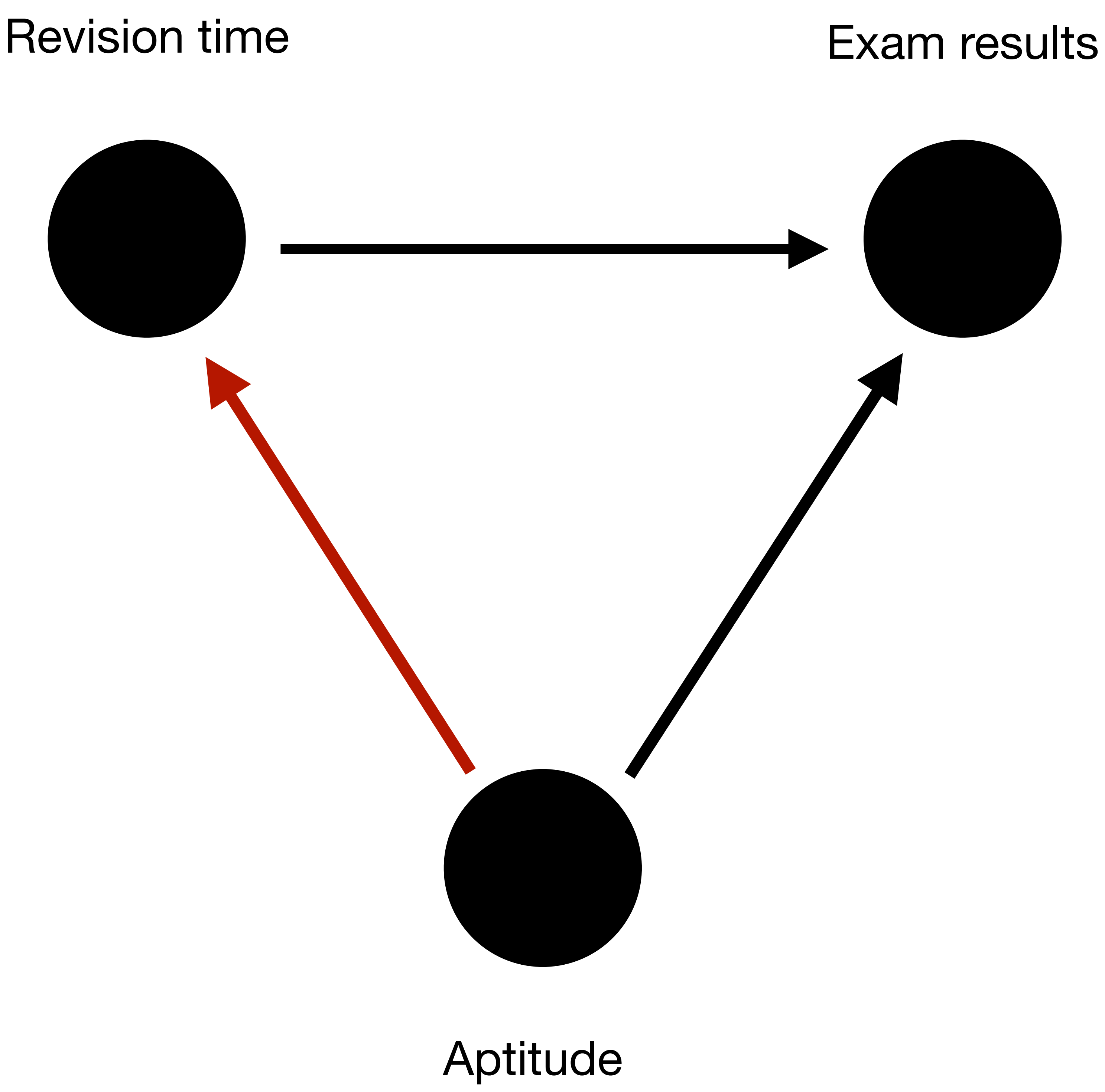
The characteristics of social science data

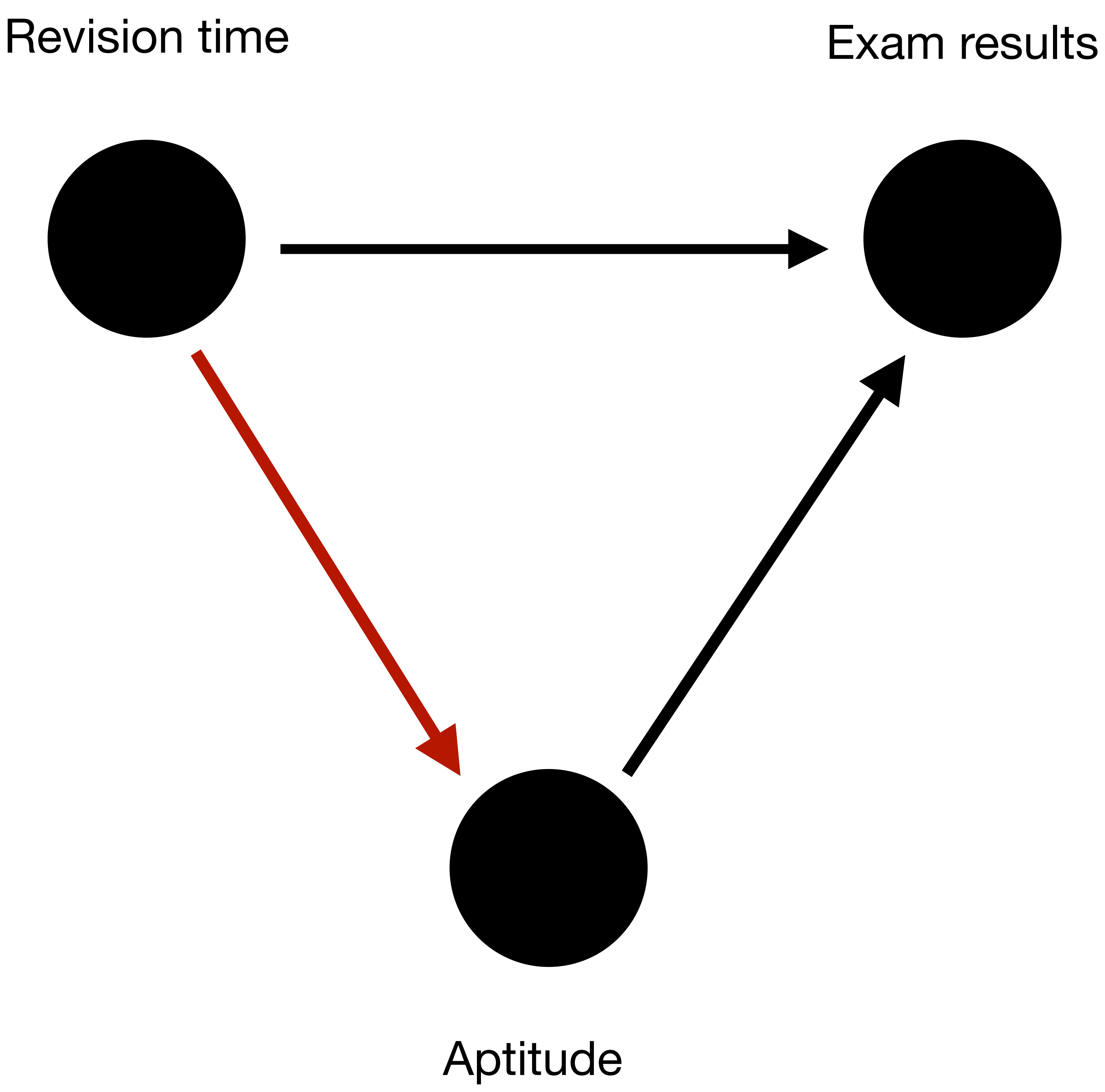
- Observational data generated from unknown graph.
    - Randomised control trial not always possible.

The characteristics of social science data

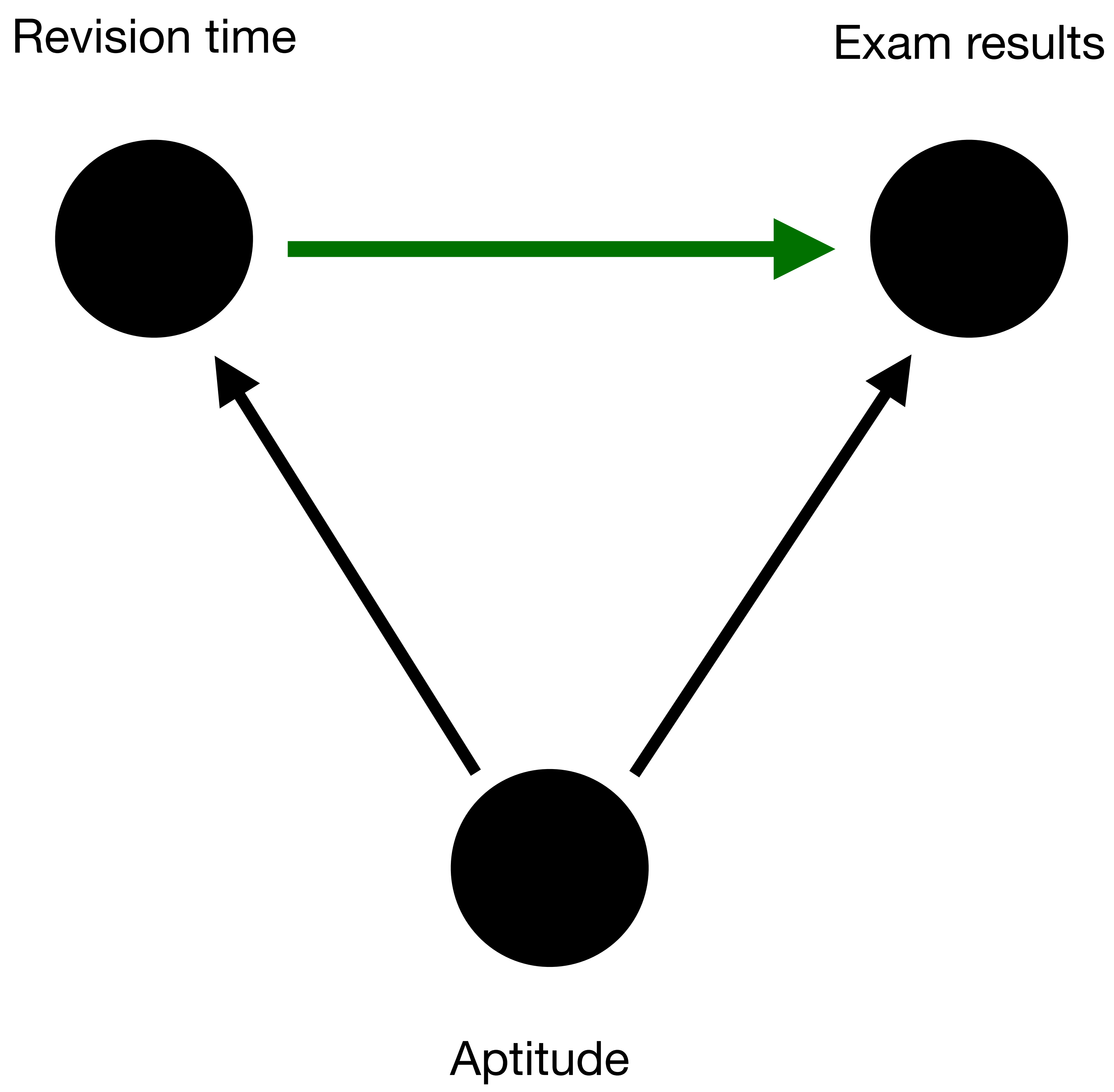
- Observational data generated from unknown graph.
    - Randomised control trial not always possible.
    - Without graph, cannot compute treatment effect.

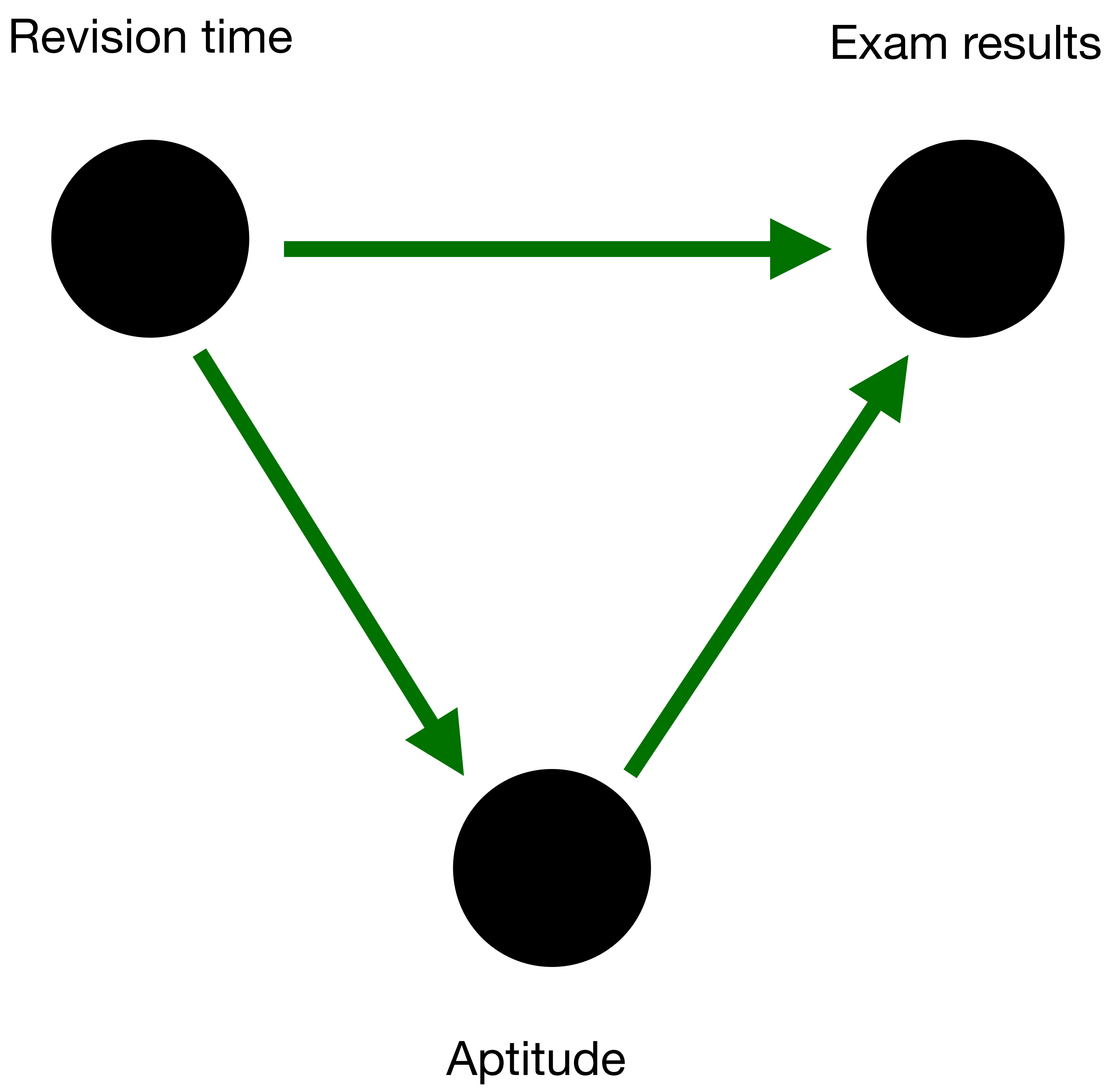
The characteristics of social science data



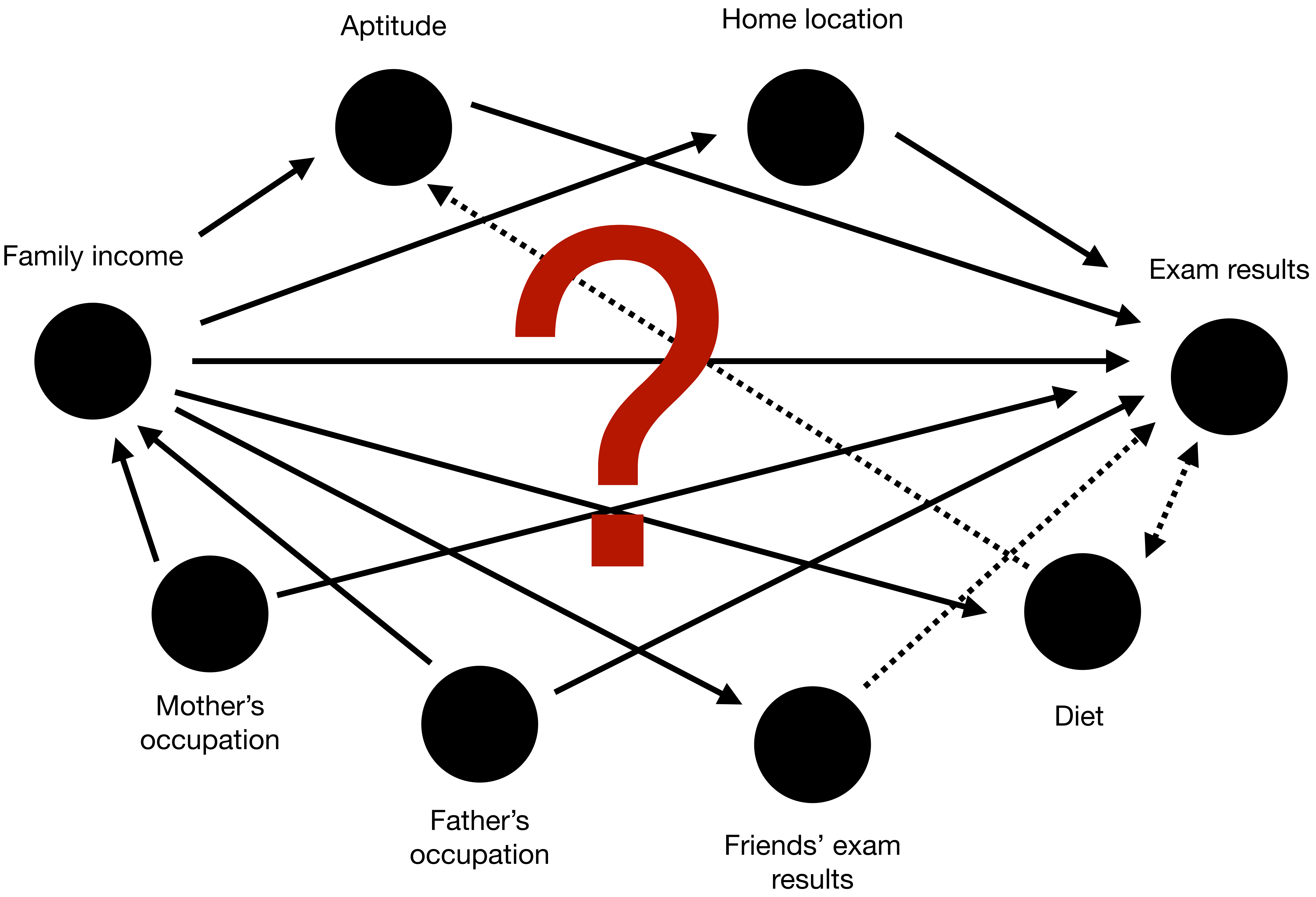


The characteristics of social science data





The characteristics of social science data



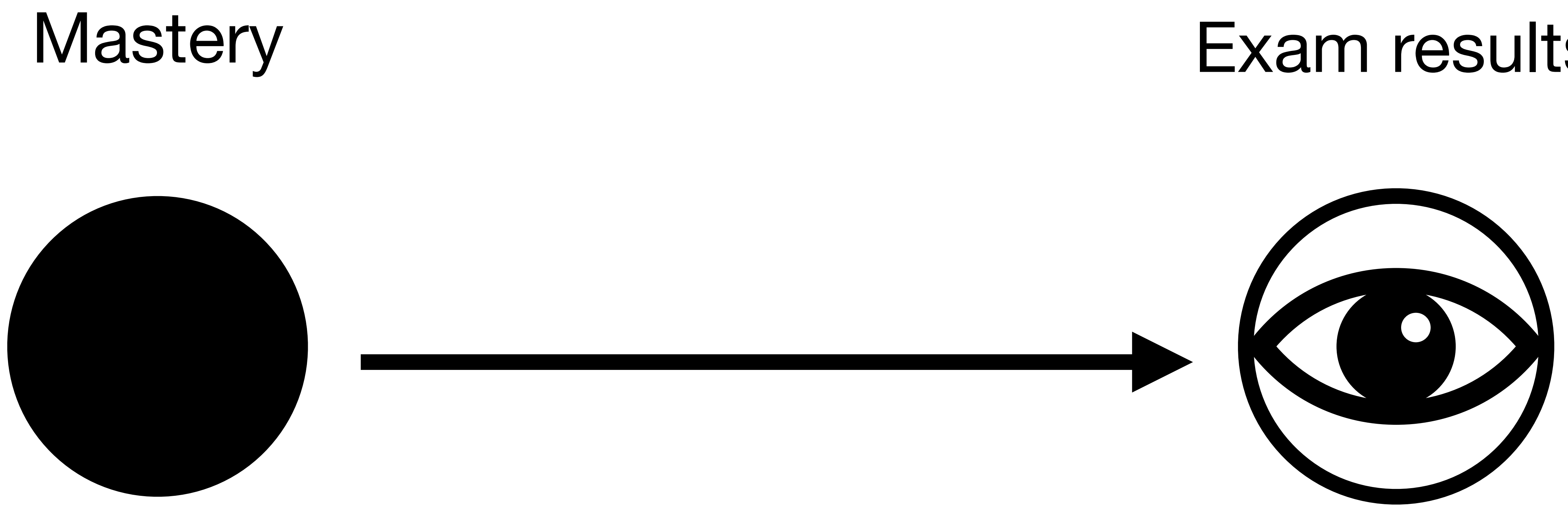
The characteristics of social science data

- Observational data generated from unknown graph.
    - Randomised control trial not always possible.
    - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.

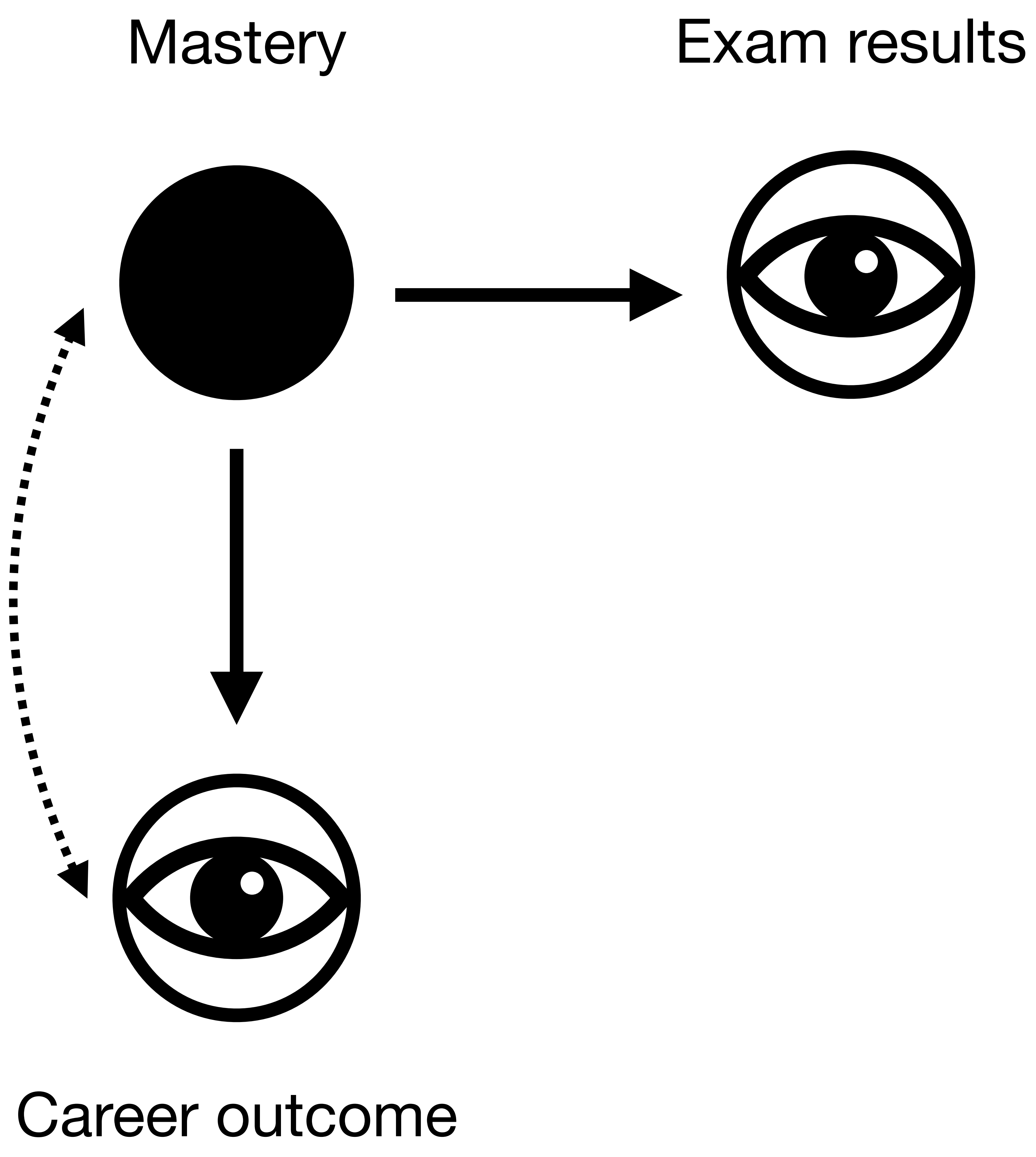
The characteristics of social science data

- Observational data generated from unknown graph.
  - Randomised control trial not always possible.
  - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
  - Masks interesting relationships in data.

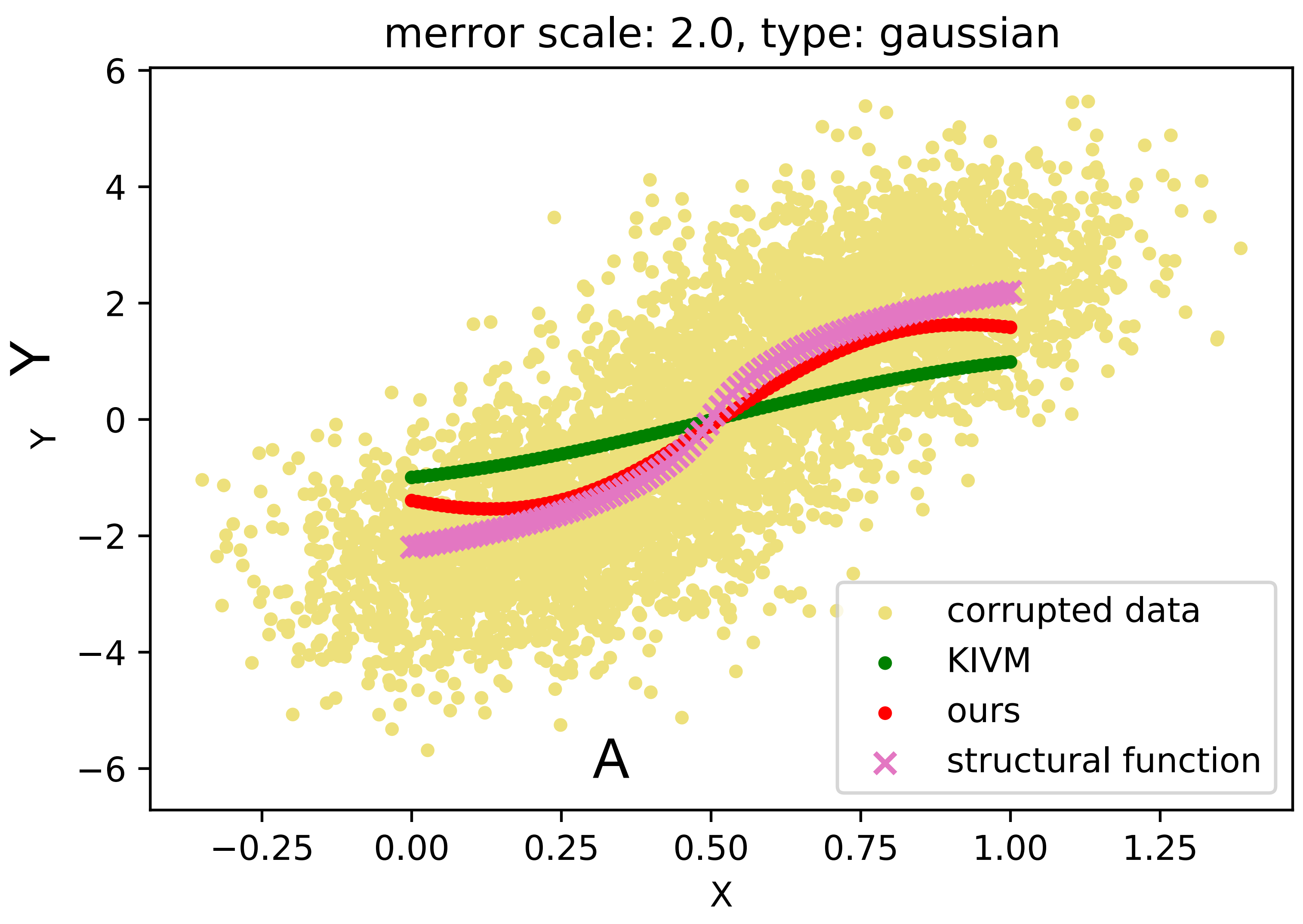
The characteristics of social science data



The characteristics of social science data



Mask interesting relationships:



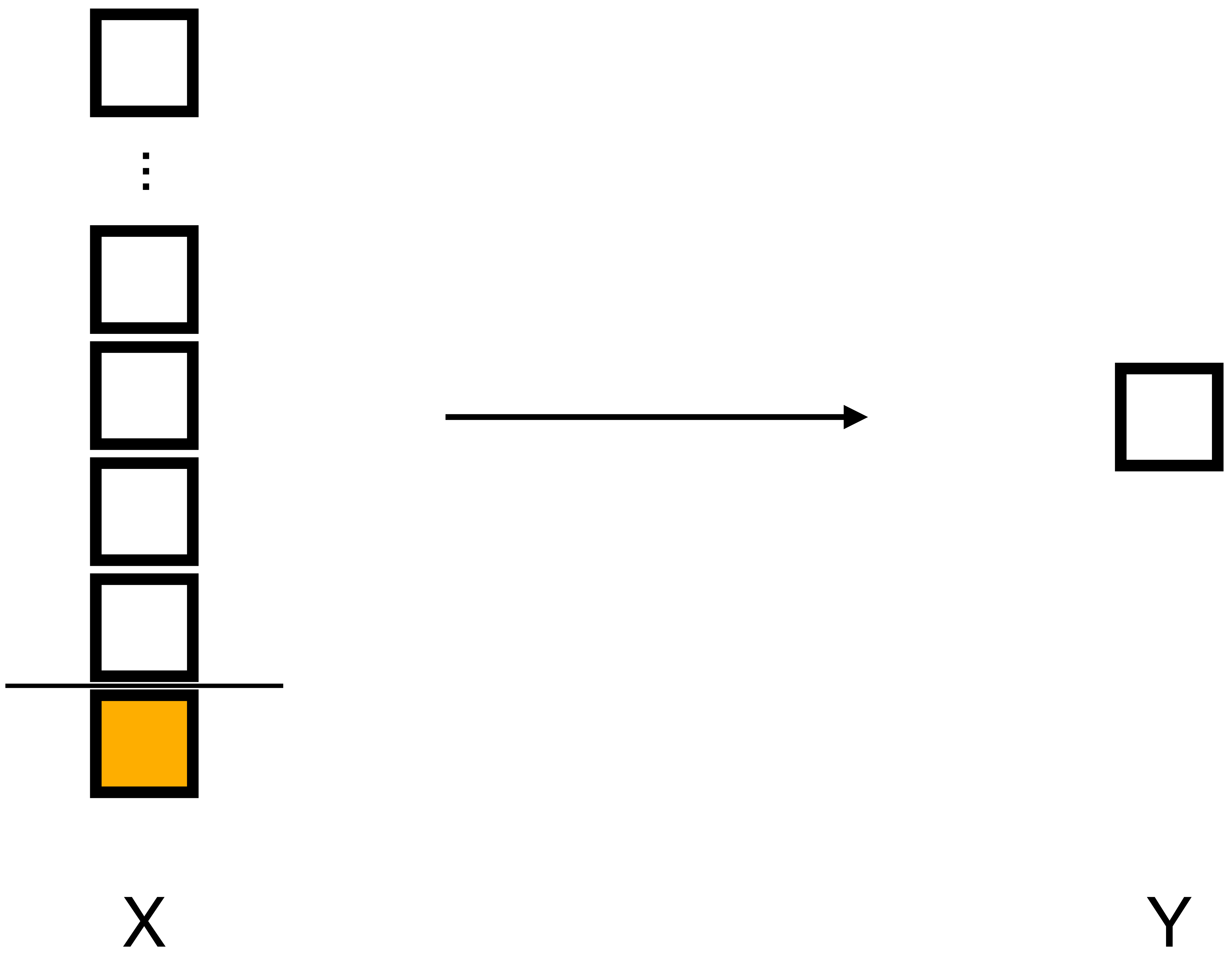
The characteristics of social science data

- Observational data generated from unknown graph.
    - Randomised control trial not always possible.
    - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
    - Masks interesting relationships in data.
- High-dimensional - e.g. text data, video data, many covariates.

The characteristics of social science data

- Observational data generated from unknown graph.
  - Randomised control trial not always possible.
  - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
  - Masks interesting relationships in data.
- High-dimensional - e.g. text data, video data, many covariates.
  - Regularisation bias.

The characteristics of social science data



The characteristics of social science data

- Observational data generated from unknown graph.
  - Randomised control trial not always possible.
  - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
  - Masks interesting relationships in data.
- High-dimensional - e.g. text data, video data, many covariates.
  - Regularisation bias.
- Confounded - e.g. exam outcome and exam preparation is confounded by aptitude.

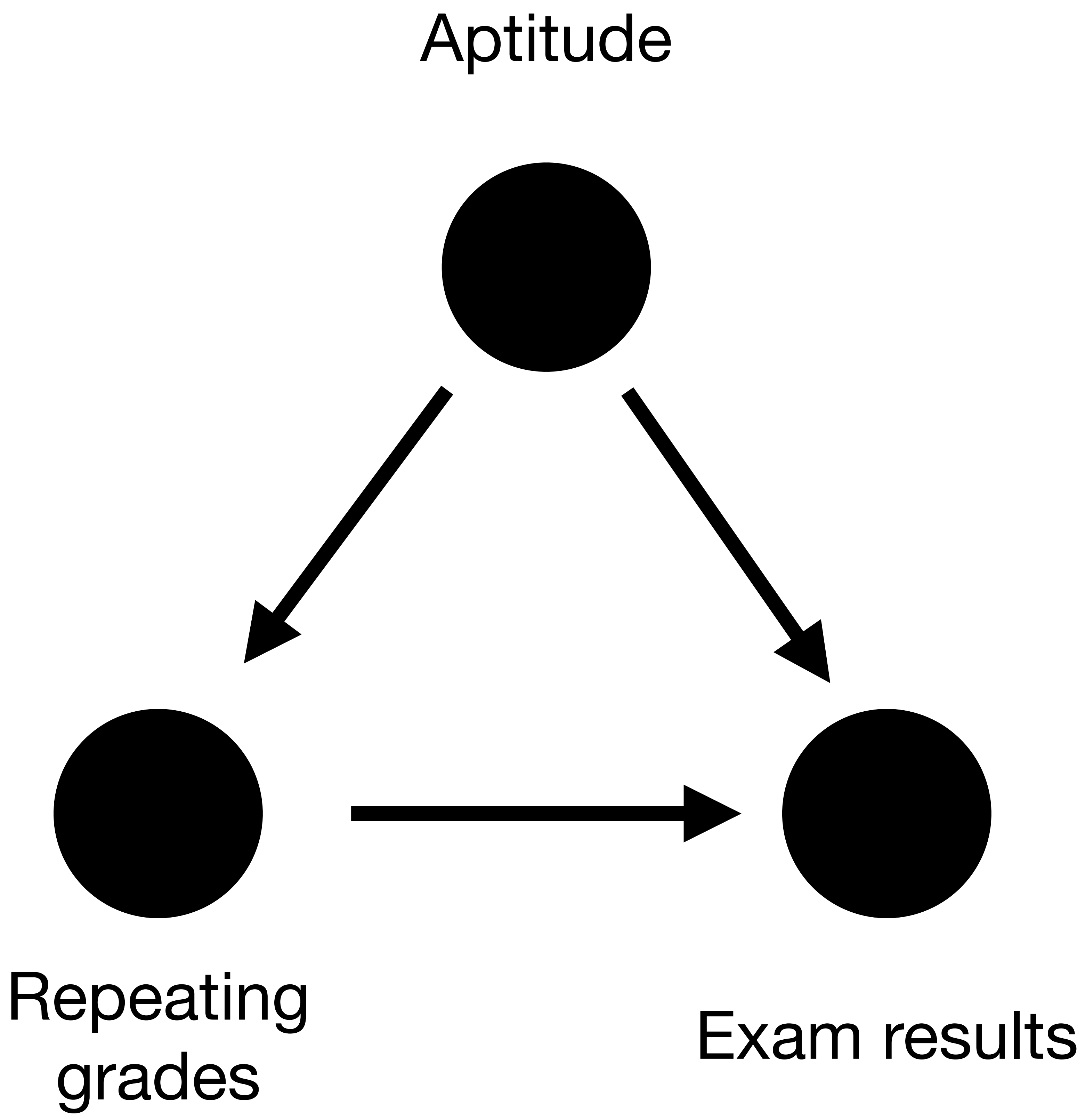
The characteristics of social science data

- Observational data generated from unknown graph.
  - Randomised control trial not always possible.
  - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
  - Masks interesting relationships in data.
- High-dimensional - e.g. text data, video data, many covariates.
  - Regularisation bias.
- Confounded - e.g. exam outcome and exam preparation is confounded by aptitude.
  - Sometimes the confounding is **not observed**.

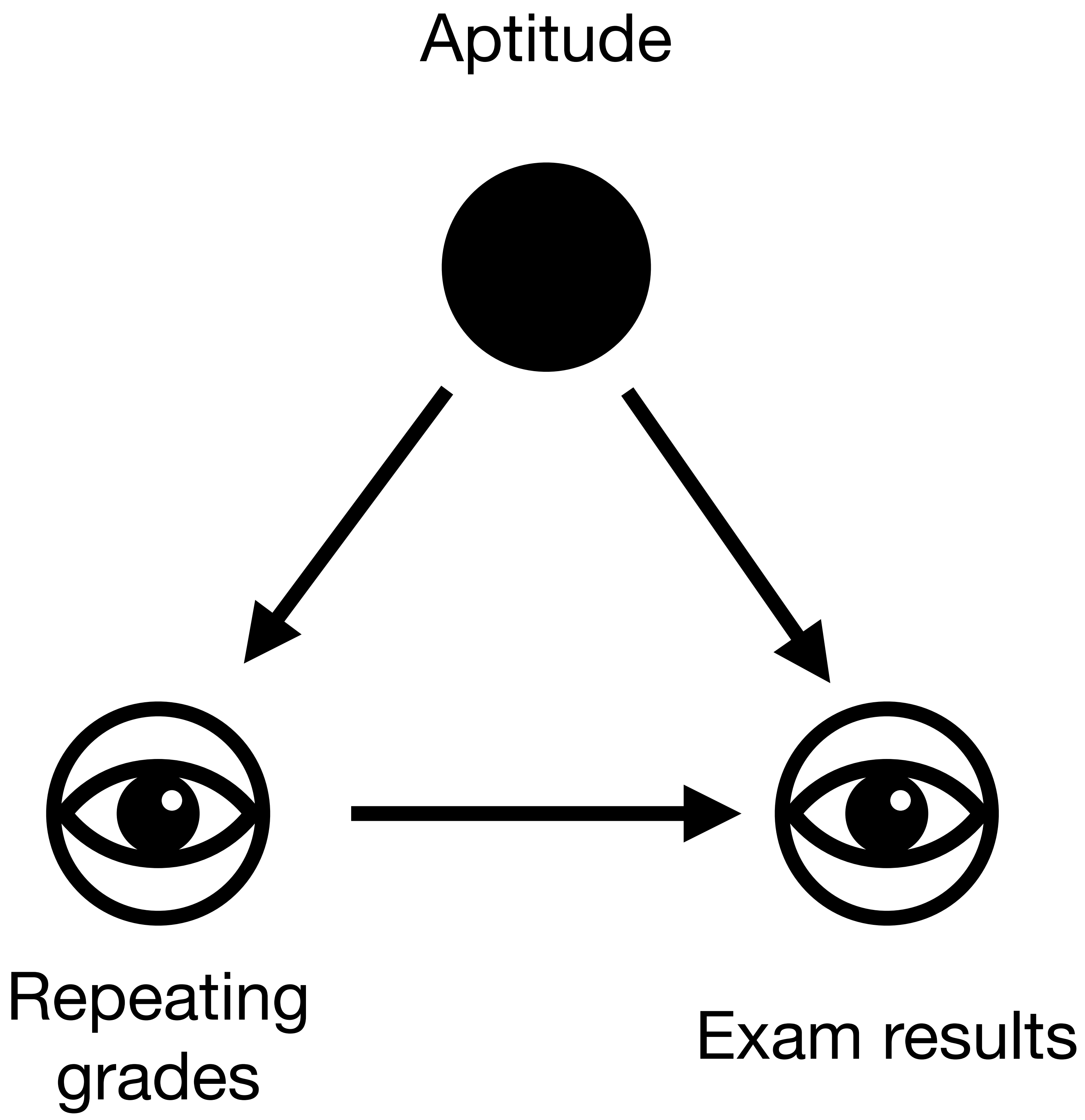
The characteristics of social science data

- Observational data generated from unknown graph.
    - Randomised control trial not always possible.
    - Without graph, cannot compute treatment effect.
- Noisy / Measurement error - on exposure variables, response variables, and potentially other covariates.
    - Masks interesting relationships in data.
- High-dimensional - e.g. text data, video data, many covariates.
    - Regularisation bias.
- Confounded - e.g. exam outcome and exam preparation is confounded by aptitude.
    - Sometimes the confounding is **not observed**.
    - Simpson's paradox.

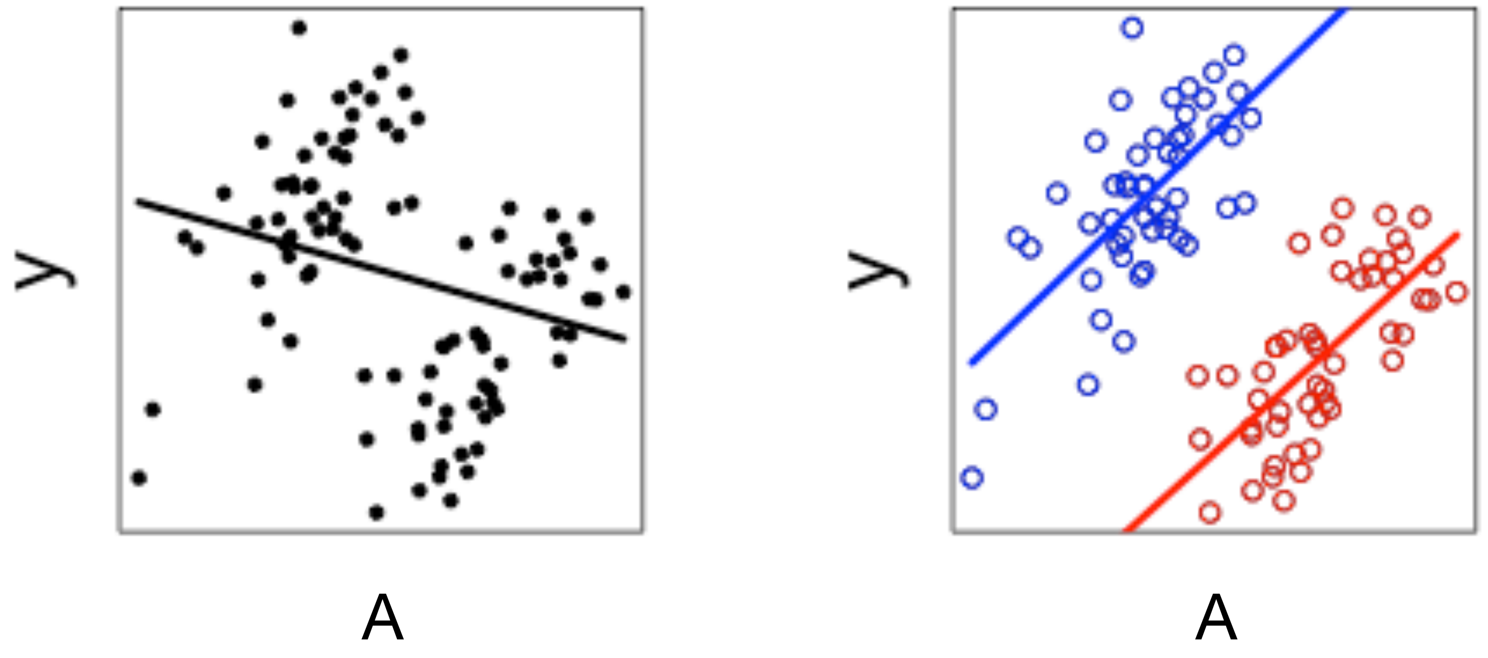
The characteristics of social science data



The characteristics of social science data



Simpson's paradox:



Four problem themes

Unknown graph structure

Measurement noises

Confounding

High dimensional data

Overview

- What we want to achieve with causality.
- Why is causality suitable for social sciences?
- The characteristics of social science data.
- Algorithms.

A causal toolset for social sciences.

Unknown graph structure

Measurement noises

Confounding

High dimensional data

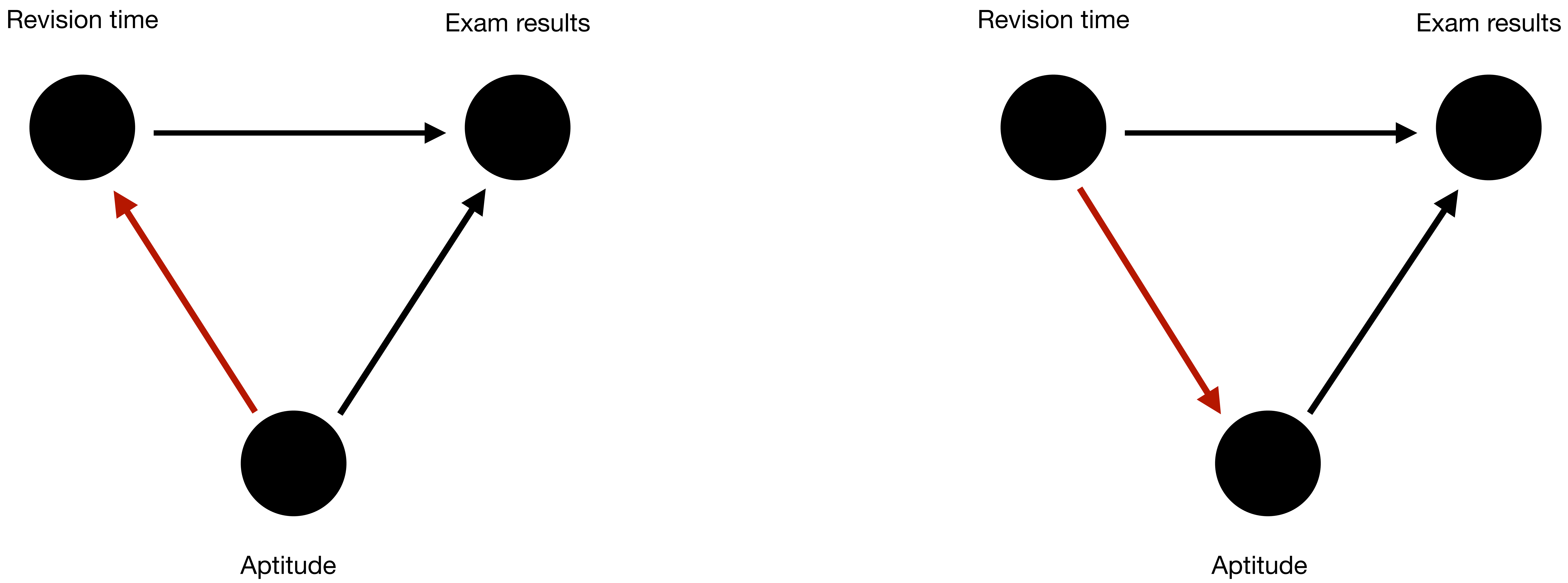
Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
  - Temporal reasoning, expert knowledge, or from some experimentation.

The characteristics of social science data



Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
    - Temporal reasoning, or from some experimentation.
    - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
  - Temporal reasoning, or from some experimentation.
  - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.
  - Just because we are starting from a graph does not mean we are making strong graphical assumptions.

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
  - Temporal reasoning, or from some experimentation.
  - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.
  - Just because we are starting from a graph does not mean we are making strong graphical assumptions.
- It is an assumption; we have to start from *some* assumptions.

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
    - Temporal reasoning, or from some experimentation.
    - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.
    - Just because we are starting from a graph does not mean we are making strong graphical assumptions.
- It is an assumption; we have to start from *some* assumptions.
    - Structural learning algorithm does not mean we are suddenly making no assumptions - rather the structural learning algorithms also depends on their meta-assumptions.

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
    - Temporal reasoning, or from some experimentation.
    - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.
    - Just because we are starting from a graph does not mean we are making strong graphical assumptions.
- It is an assumption; we have to start from *some* assumptions.
    - Structural learning algorithm does not mean we are suddenly making no assumptions - rather the structural learning algorithms also depends on their meta-assumptions.
    - The strength of assumptions we need depends on the amount/quality of the data we have.

Unknown graph structure.

- Partial claims about the causal graph can be made with reasonable confidence
    - Temporal reasoning, or from some experimentation.
    - In some situations, partial knowledge of graph structure is enough. E.g. just need to be able to group proxies into ‘treatment-inducing’ and ‘outcome-inducing’ while the structure among themselves don’t matter.
    - Just because we are starting from a graph does not mean we are making strong graphical assumptions.
- It is an assumption; we have to start from *some* assumptions.
    - Structural learning algorithm does not mean we are suddenly making no assumptions - rather the structural learning algorithms also depends on their meta-assumptions.
    - The strength of assumptions we need depends on the amount/quality of the data we have.
    - The quality of results from the structural learning algorithm depends on the quality of data.

A causal toolset for social sciences.

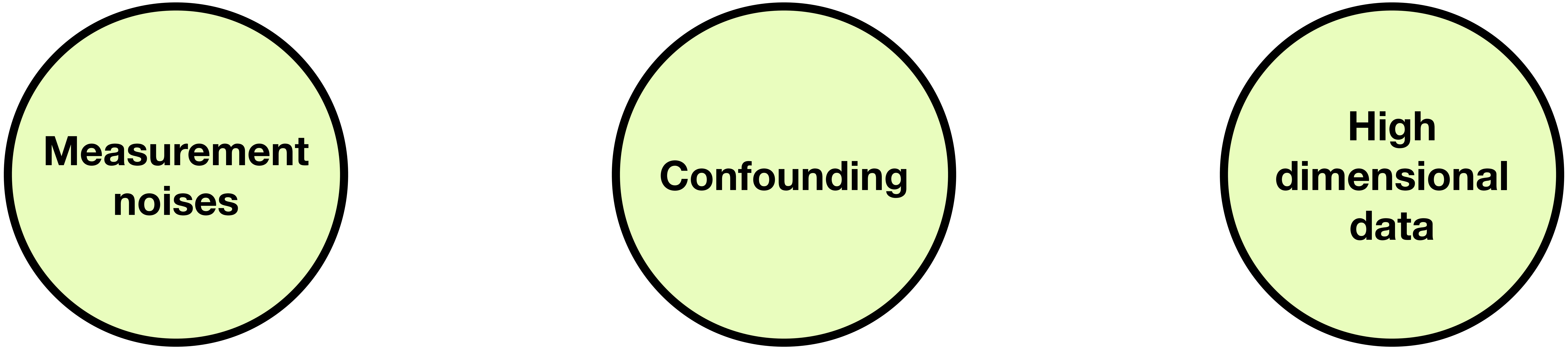
Unknown graph structure

Measurement noises

Confounding

High dimensional data

A causal toolset for social sciences.



A causal toolset for social sciences.

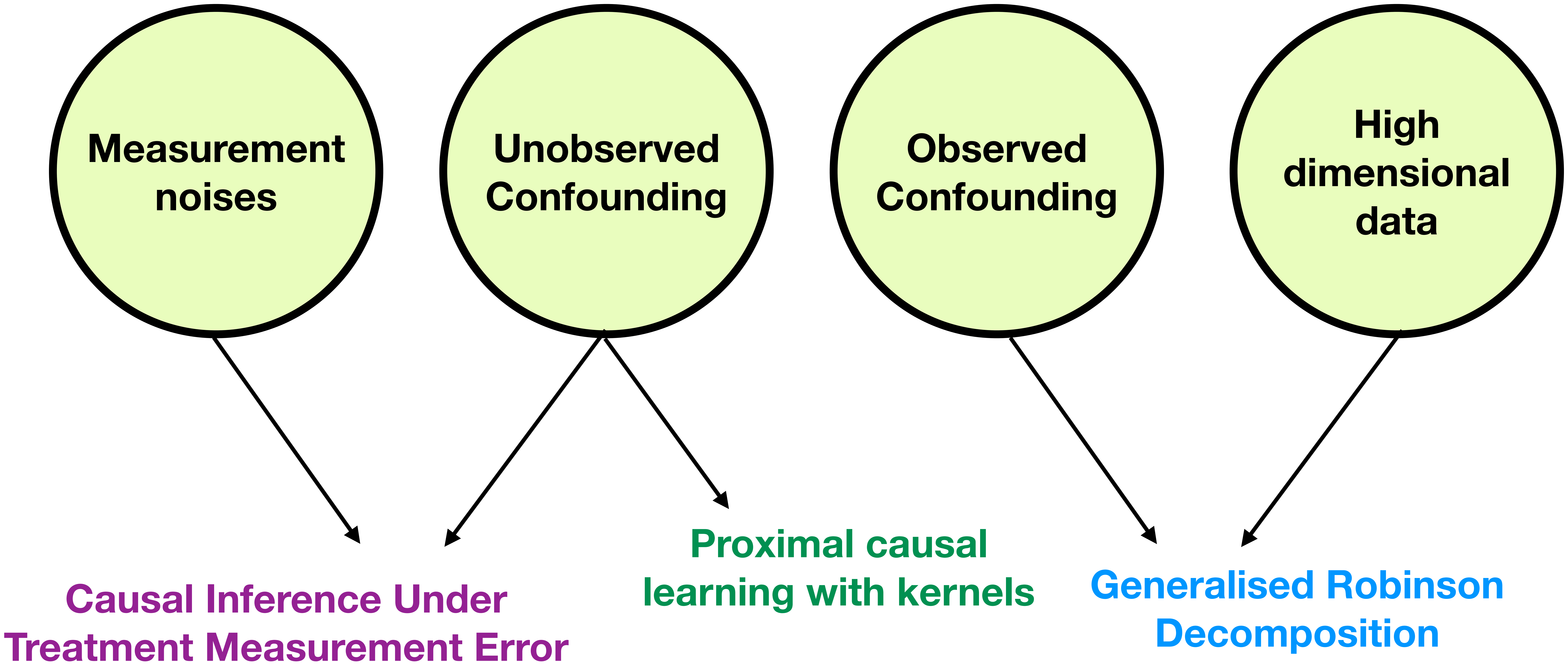
Measurement noises

Unobserved Confounding

Observed Confounding

High dimensional data

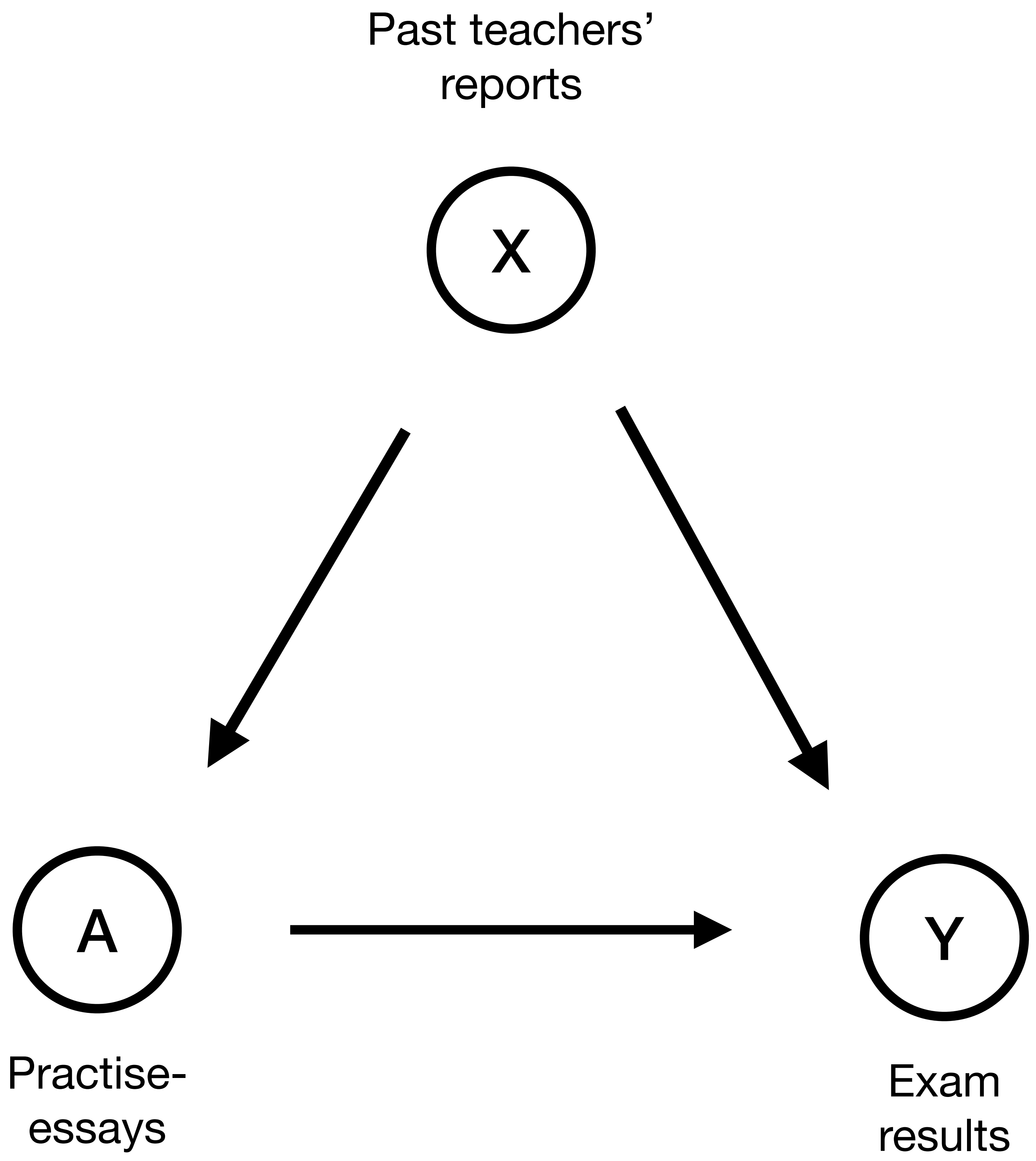
A causal toolset for social sciences.



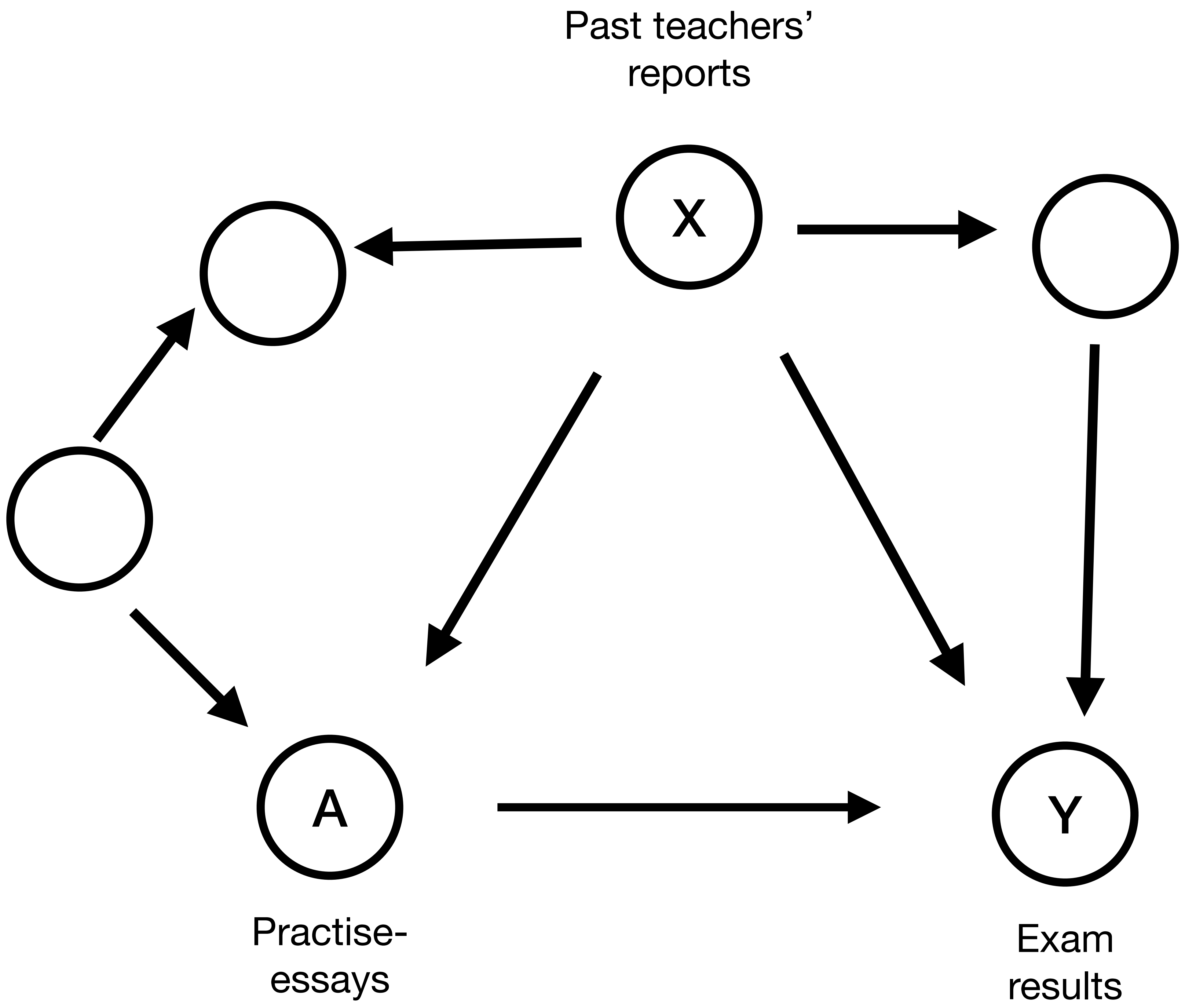
Generalised Robinson Decomposition

Kaddour, **Z.**, Liu, Kusner, Silva. Causal Inference for Structured Treatments. NeurIPS 2021.

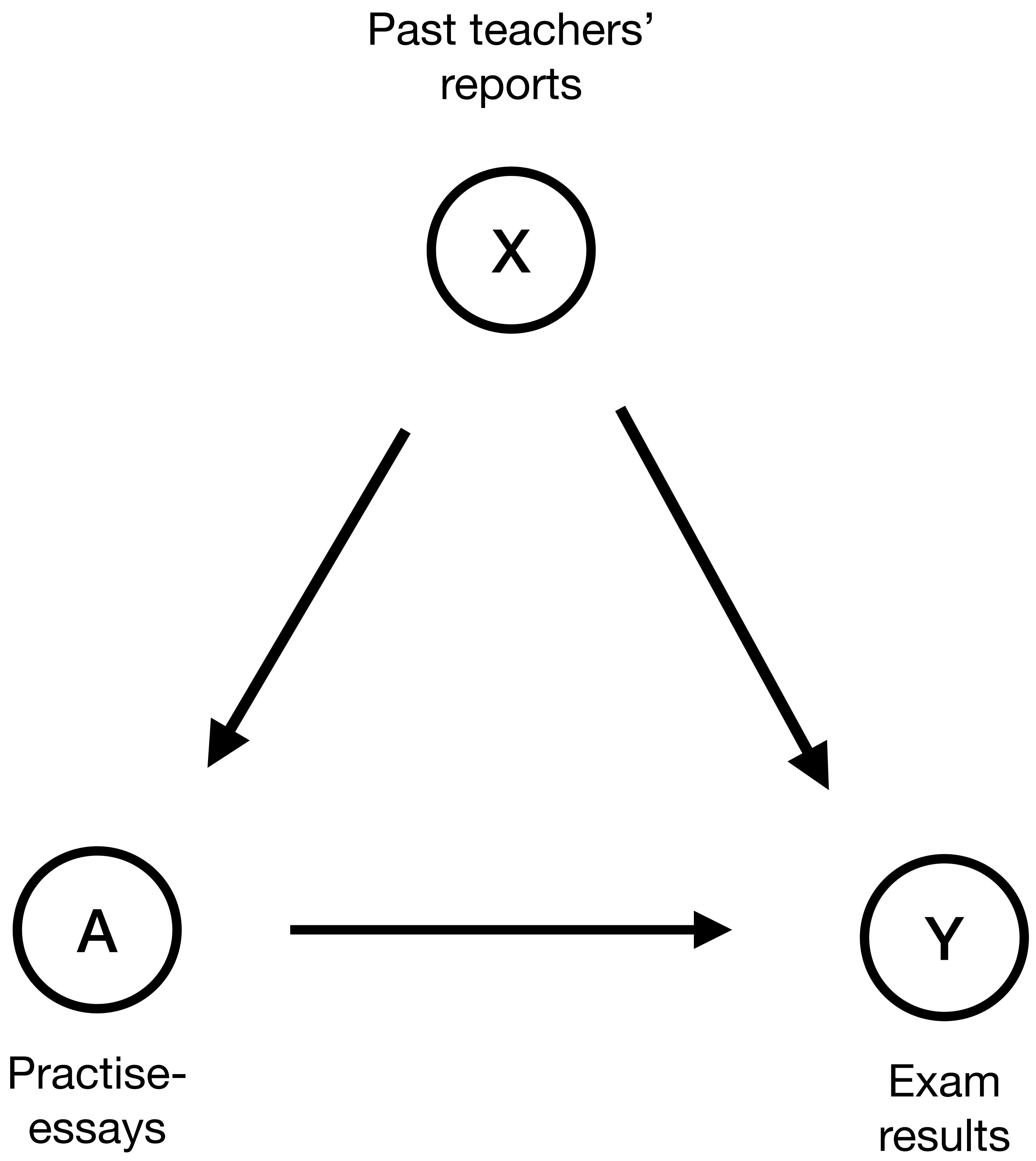
Assumptions and usage contexts



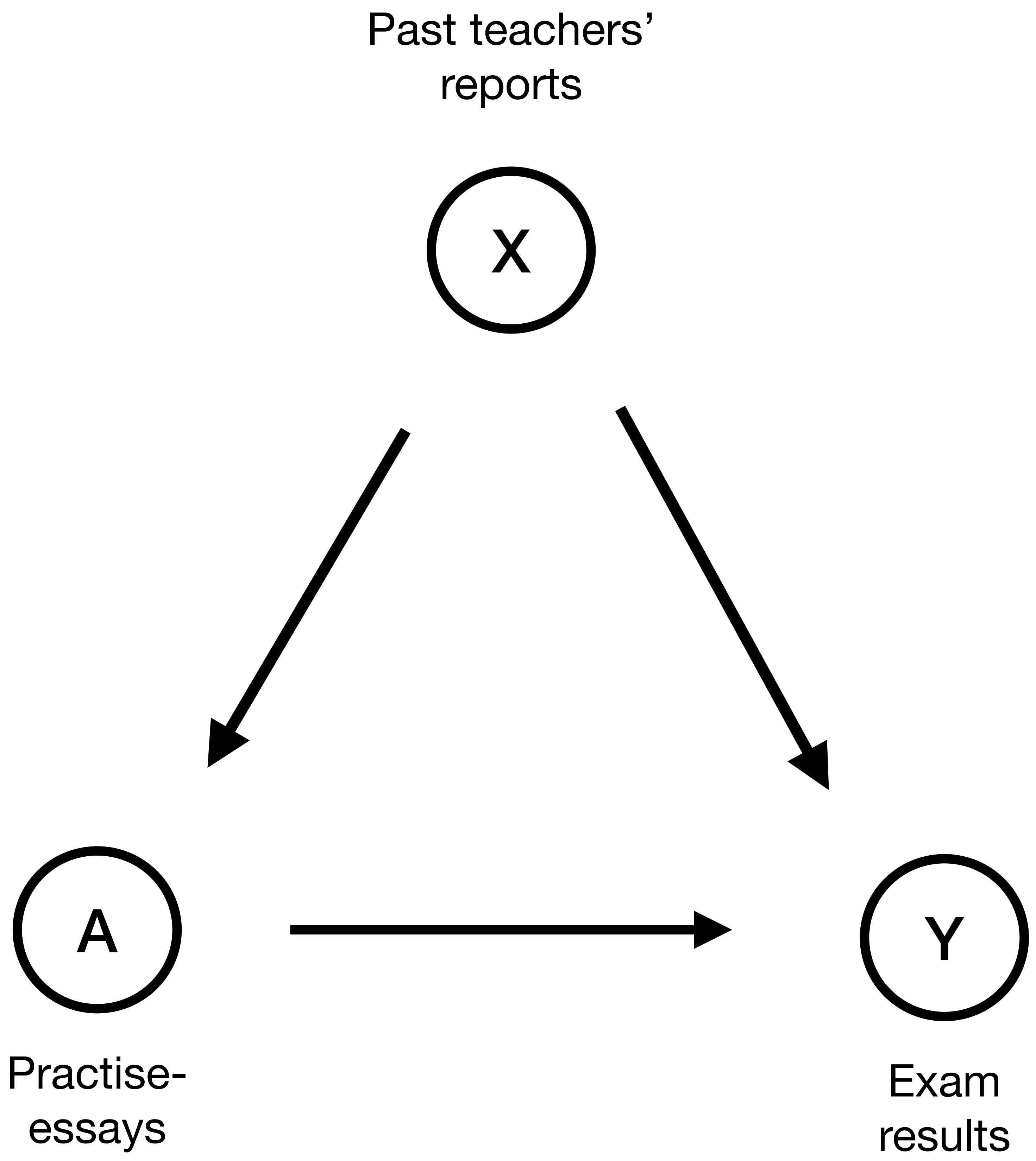
Assumptions and usage contexts



Assumptions and usage contexts

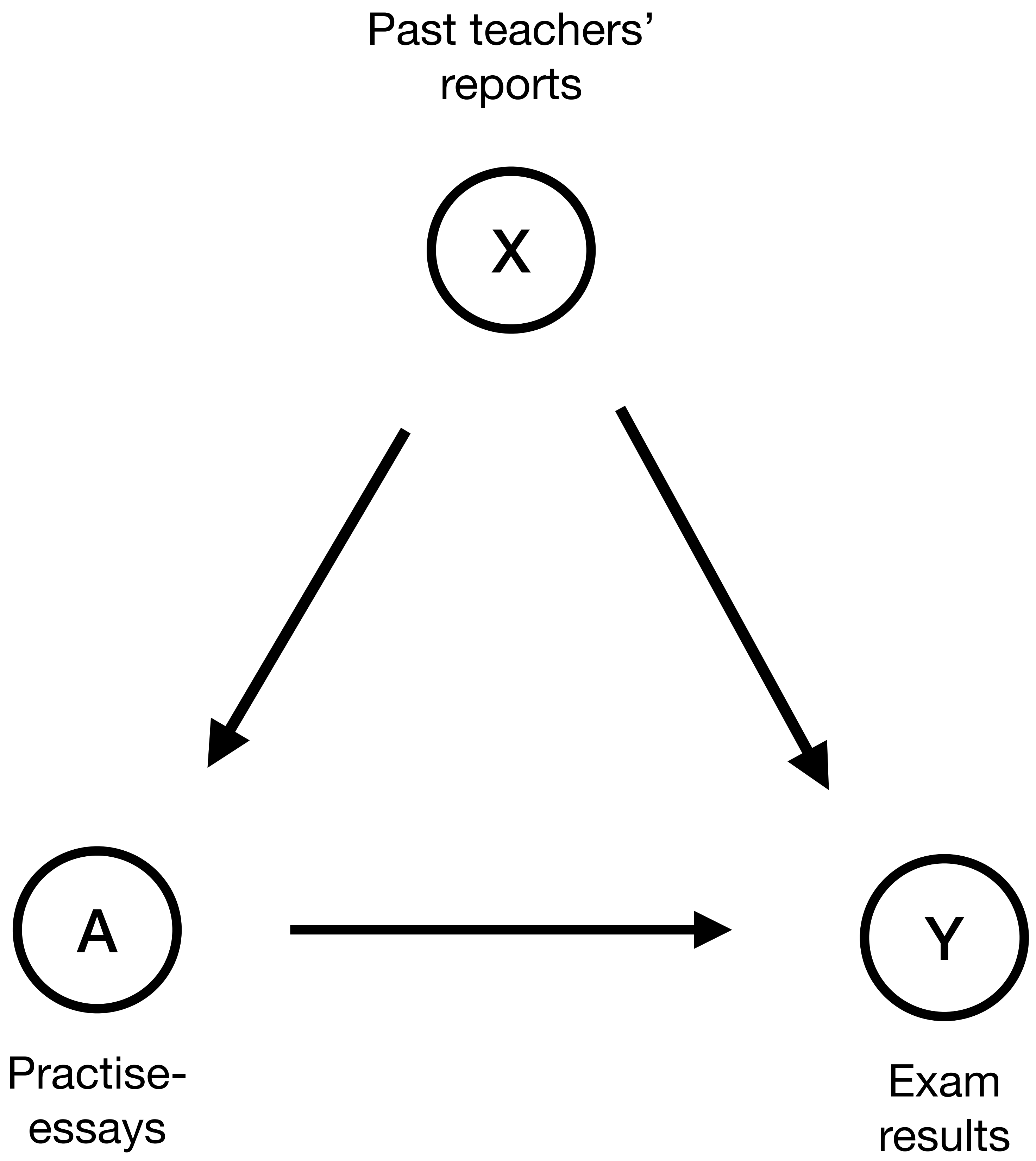


Assumptions and usage contexts



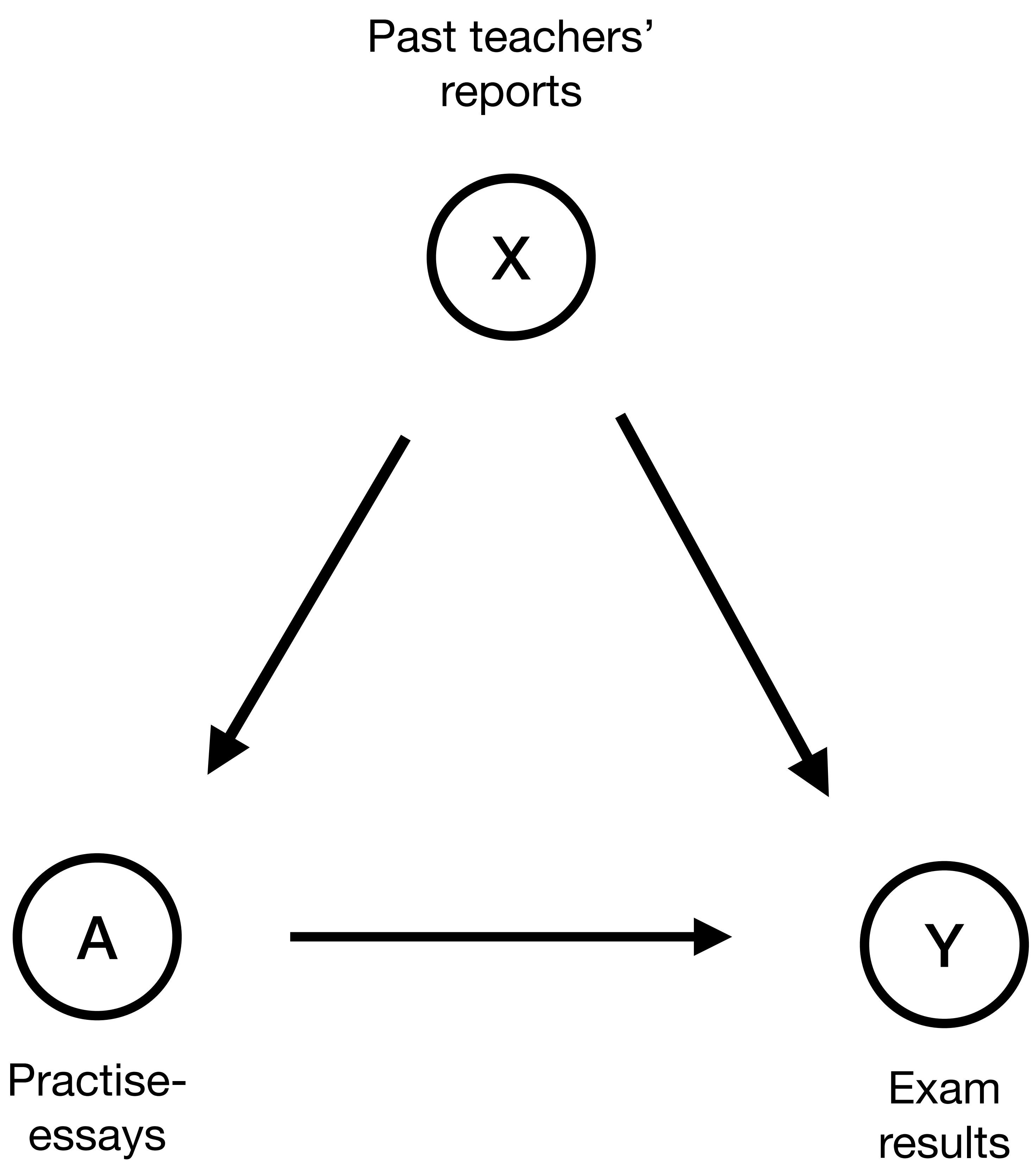
- Data in arbitrary forms - e.g. text, images; low- or high-dimensional data.

Assumptions and usage contexts



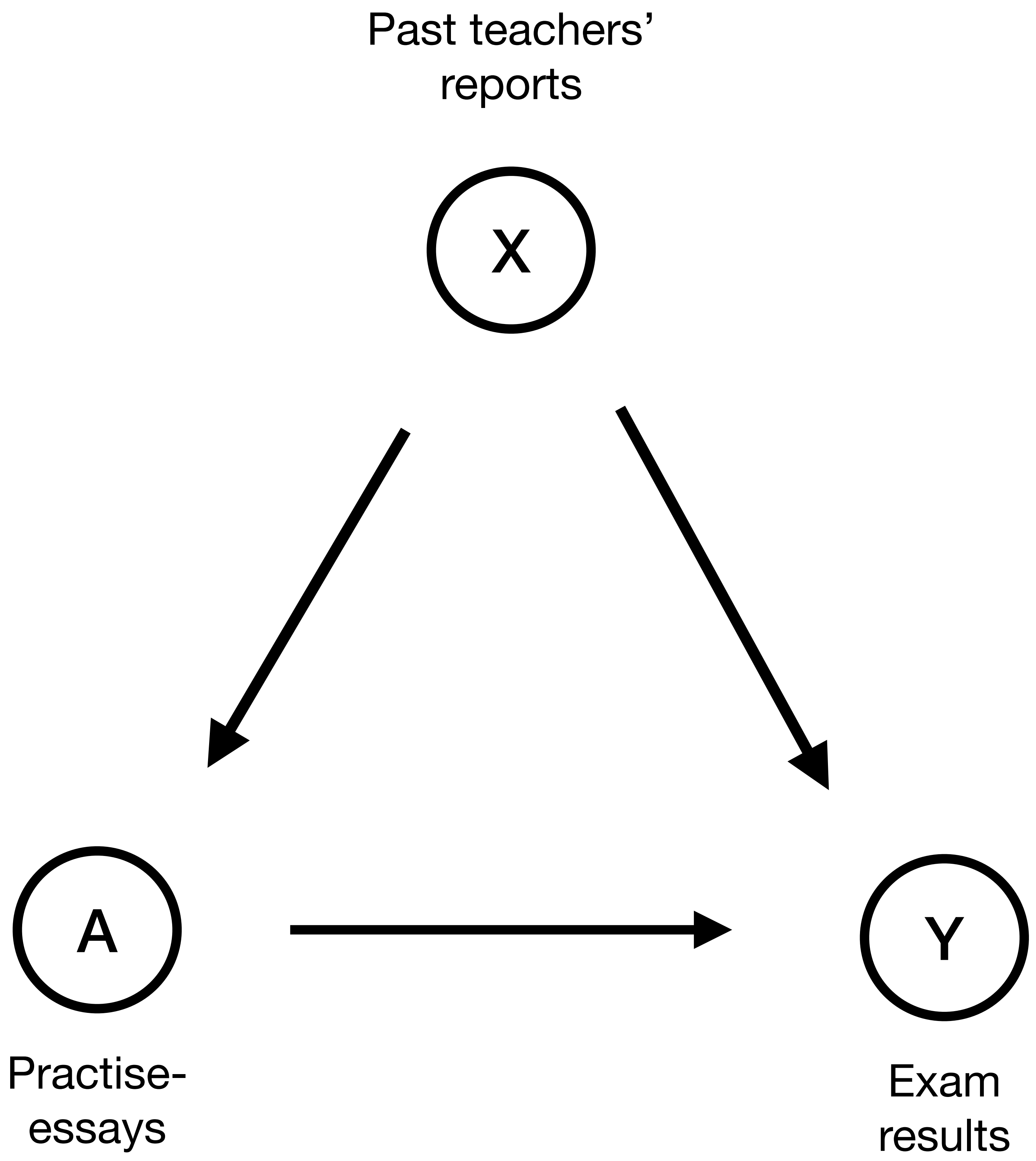
- Data in arbitrary forms - e.g. text, images; low- or high-dimensional data.
- Continuous data or data with many categories - e.g. different intensities of revision.

Assumptions and usage contexts



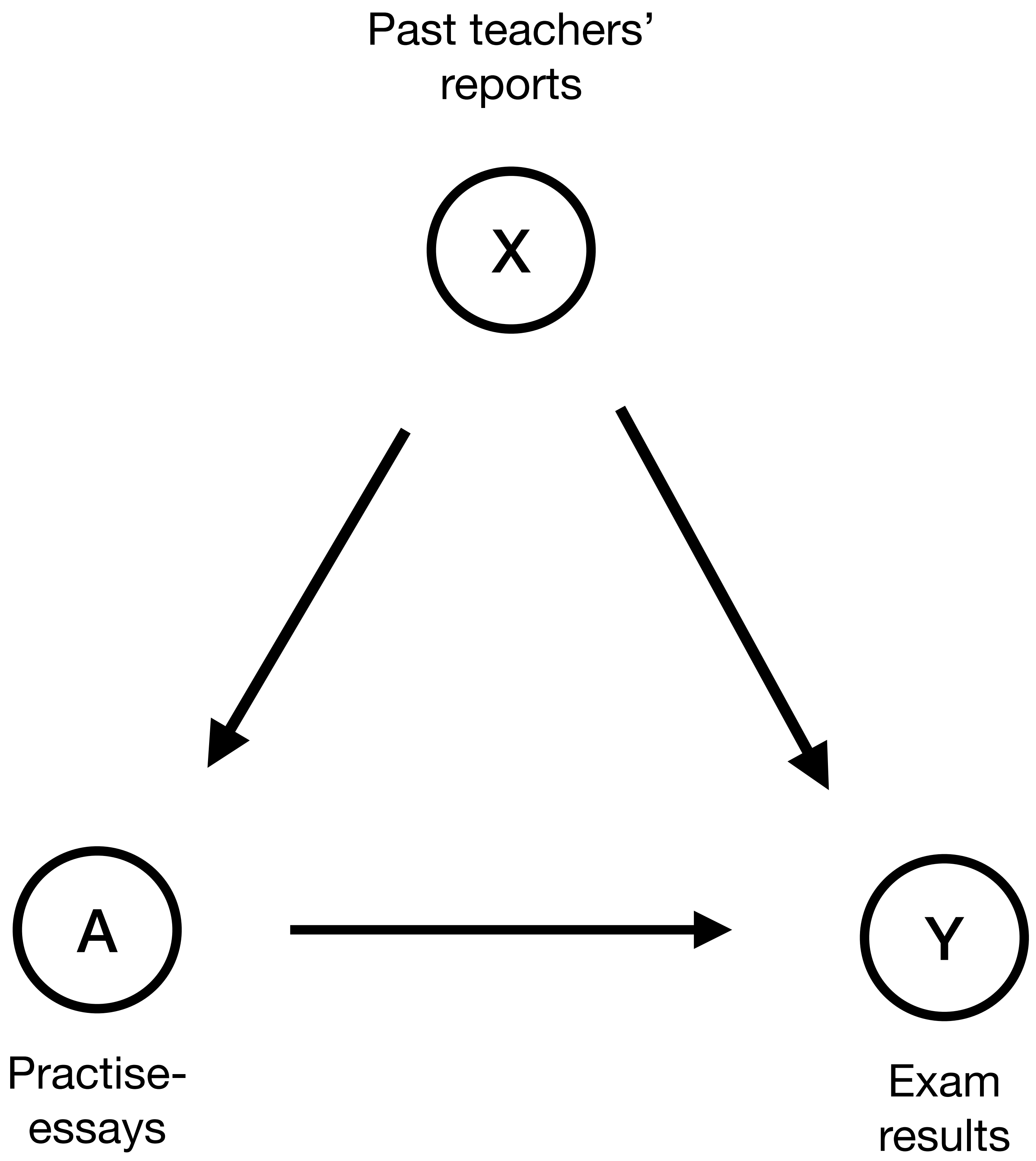
- Data in arbitrary forms - e.g. text, images; low- or high-dimensional data.
- Continuous data or data with many categories - e.g. different length of revision time.
  - *More realistic examples from the audience are appreciated!*

Assumptions and usage contexts

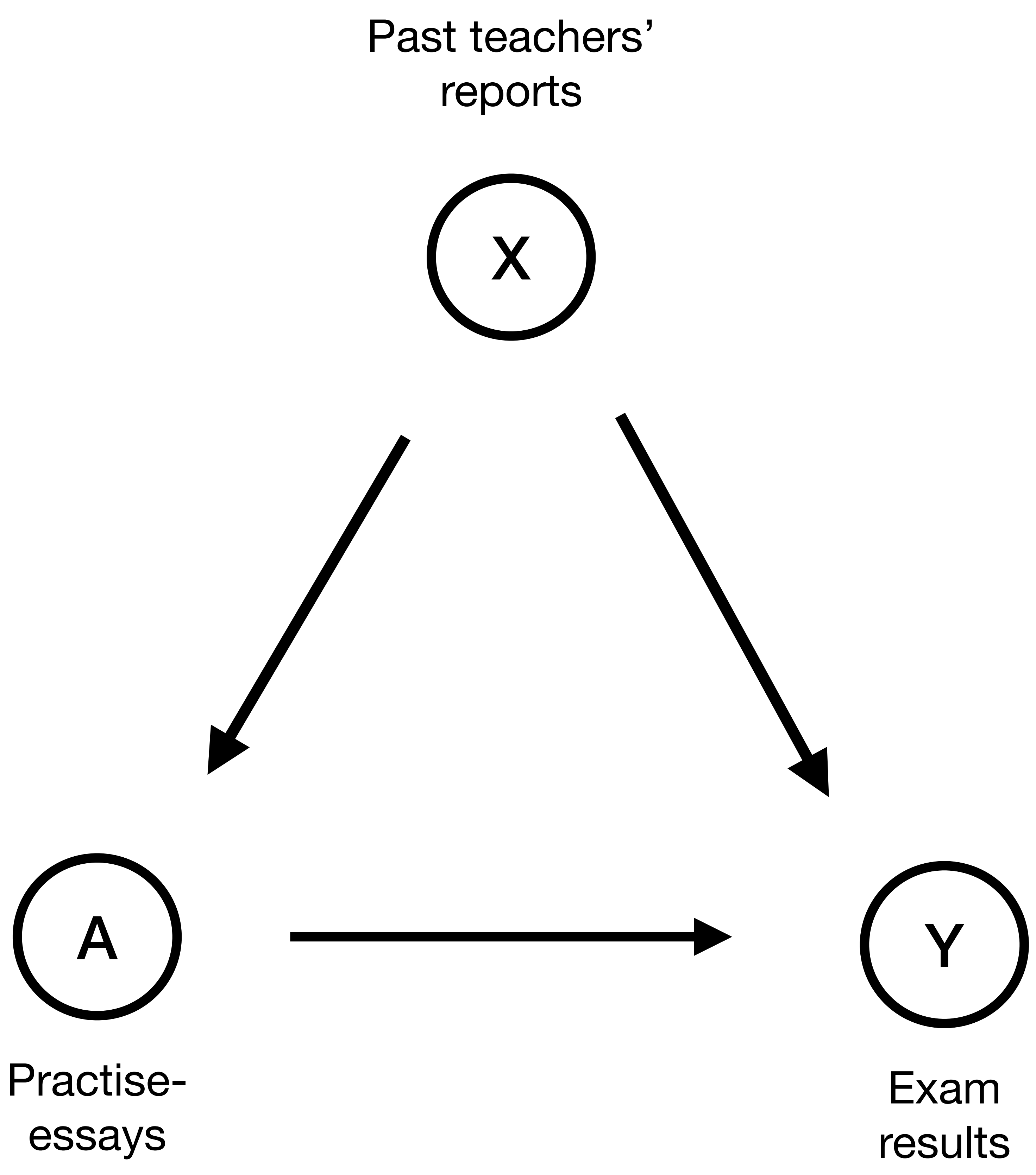


- Data in arbitrary forms - e.g. text, images; low- or high-dimensional data.
- Continuous data or data with many categories - e.g. different length of revision time.
  - *More realistic examples from the audience are appreciated!*
- Overlap condition required on features of A x features of X.

Assumptions and usage contexts

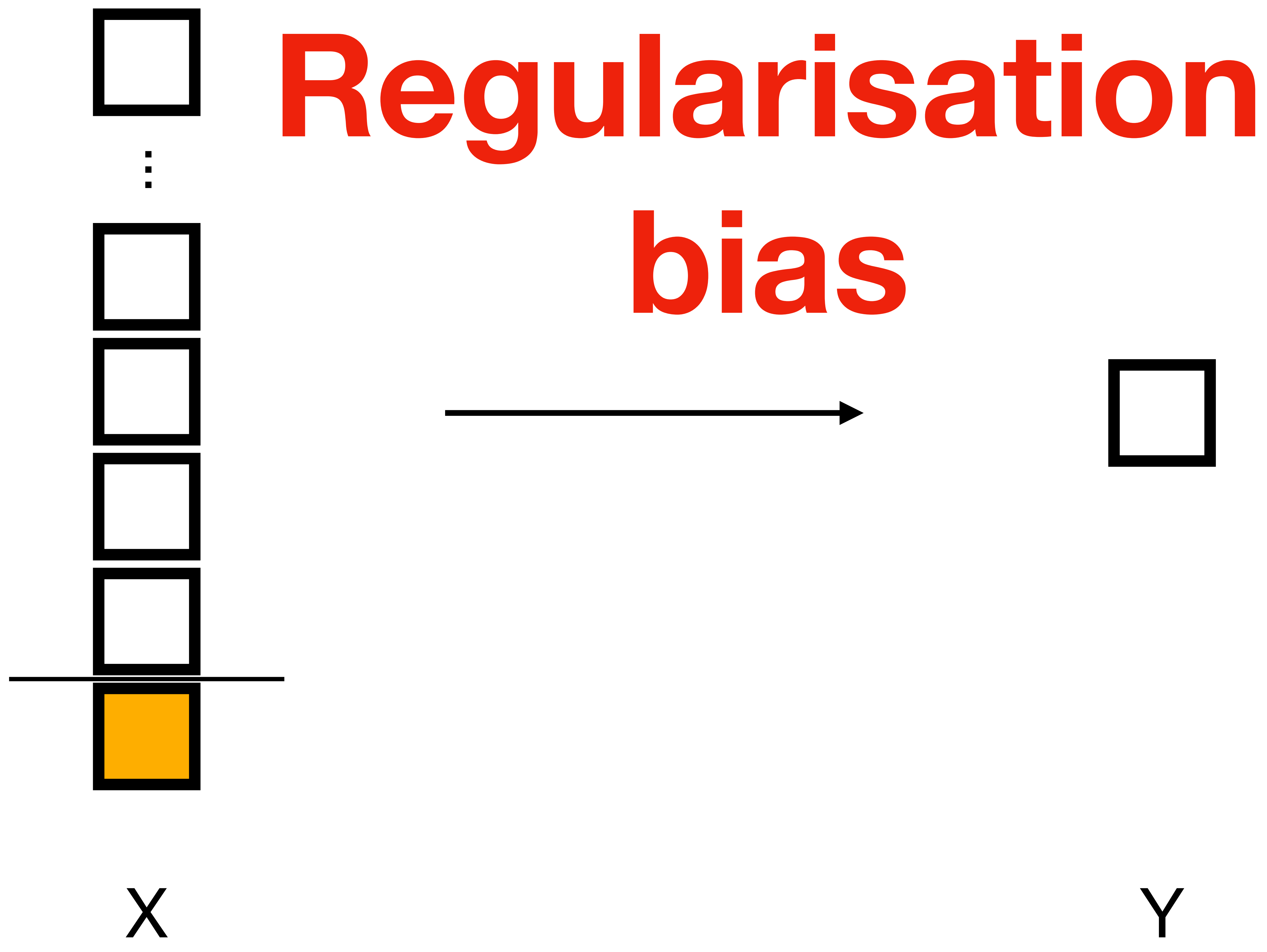


- Data in arbitrary forms - e.g. text, images; low- or high-dimensional data.
- Continuous data or data with many categories - e.g. different intensities of revision.
  - *More realistic examples from the audience are appreciated!*
- Overlap condition required on features of A x features of X.
- Conditional average treatment effect.



Why not just do regression?

The characteristics of social science data



Robinson Decomposition

- **Allows us to construct a learnable objective of the binary CATE.**
- **Define the propensity score** $e(x) := p(A = 1 \,|\, \mathbf{x})$**.**
- **Define the conditional mean outcome** $m(x) := \mathbb{E}[Y \,|\, \mathbf{x}]$**.**
- **Define** $\tilde{y}_i := y_i - \hat{m}(\mathbf{x}_i)$ **and** $\tilde{a}_i := a_i - \hat{e}(\mathbf{x}_i)$ **we yield the objective**

$$\tilde{\tau}_b(\cdot) = \arg\min_{\tau_b} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - \tilde{a}_i \times \tau_b(\mathbf{x}_i))^2 + \Lambda(\tau_b(\cdot)) \right\}$$

- **We call** $\hat{m}(\mathbf{x})$ **and** $\hat{e}(\mathbf{x})$ **the (estimated) nuisance components.**

- Xinkun Nie and Stefan Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. Biometrika, 2021.
- Thanks to Jean Kaddour for providing the original slides the current slide is based on.

Generalised Robinson Decomposition

- **Product Effect Assumption: Re-parameterise the outcome surface as $Y = g(\mathbf{X})^\top h(\mathbf{A}) + \epsilon$ where $g : \mathscr{X} \to \mathbb{R}^d, h : \mathscr{A} \to \mathbb{R}^d$ are feature maps.**
- **Universality property: As we let the dimensionality of $g(\cdot)$ and $h(\cdot)$ grow, we may approximate any bounded function in $\mathscr{C}(\mathscr{X} \times \mathscr{A})$.**
- **So the conditional average treatment effect is $\tau(\mathbf{a}', \mathbf{a}, \mathbf{x}) = g(\mathbf{x})^\top (h(\mathbf{a}') - h(\mathbf{a}))$**

Kaddour, **Z.**, Liu, Kusner, Silva. Causal Inference for Structured Treatments. NeurIPS 2021.

Thanks to Jean Kaddour for providing the original slides the current slide is based on.

Generalised Robinson Decomposition

- **Define propensity features $e^h(\mathbf{x}) := \mathbb{E}[h(\mathbf{A}) \,|\, \mathbf{x}]$.**
- **Recall $m(\mathbf{x}) := \mathbb{E}[Y \,|\, \mathbf{x}] = g(\mathbf{x})^\top e^h(\mathbf{x})$.**
- **Following the same steps as for the binary treatment case, we yield**

$$Y - m(\mathbf{X}) = g(\mathbf{X})^\top (h(\mathbf{A}) - e^h(\mathbf{X})) + \epsilon$$

- **Solution: For a fixed $h(\,\cdot\,)$ a generalisation to structured treatment is**

$$\hat{g}(\,\cdot\,) = \arg\min_g \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}(\mathbf{X}_i) - g(\mathbf{X}_i)^\top (h(\mathbf{A}_i) - \hat{e}^h(\mathbf{X}_i)))^2 \right\}$$

Why is the decomposition useful?

$$\hat{f}(\mathbf{x}, \mathbf{a}) := \Psi(\mathbf{x})^\top \Theta \Phi(\mathbf{a})$$

$$f^*(\mathbf{x}, \mathbf{a}) := \mathbb{E}[Y \mid \mathbf{x}, \mathbf{a}]$$

Why is the decomposition useful?

$$\hat{f}(\cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}) := \Psi(\cdot_{\mathbf{x}})^{\top}\Theta\Phi(\cdot_{\mathbf{a}})$$

$$\Big\downarrow \; \tilde{O}(n^{-\frac{1}{2(1+p)}})$$

$$f^{*}(\cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}) := \mathbb{E}[Y \,|\, \cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}]$$

$$\hat{m}(\cdot_{\mathbf{x}}) \to m(\cdot_{\mathbf{x}})$$

$$\hat{e}^{h}(\cdot_{\mathbf{x}}) \to e^{h}(\cdot_{\mathbf{x}})$$

** Main statement in Theorem 2 of paper.*

Why is the decomposition useful?

$$\hat{f}(\cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}) := \Psi(\cdot_{\mathbf{x}})^{\top}\Theta\Phi(\cdot_{\mathbf{a}})$$

$$\Big\downarrow \; \tilde{O}(n^{-\frac{1}{2(1+p)}})$$

$$f^{*}(\cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}) := \mathbb{E}[Y | \cdot_{\mathbf{x}}, \cdot_{\mathbf{a}}]$$

$$\hat{m}(\cdot_{\mathbf{x}}) \rightarrow_{O(n^{-1/4})} m(\cdot_{\mathbf{x}})$$

$$\hat{e}^{h}(\cdot_{\mathbf{x}}) \rightarrow_{O(n^{-1/4})} e^{h}(\cdot_{\mathbf{x}})$$

Overlap: $\mathscr{P}_{\Psi(\mathbf{X})\times\Phi(\mathbf{T})} > 0$

** Main statement in Theorem 2 of paper.*

Why does this mean?

- **The target or nuisance functions never converge faster than $O(n^{-1/2})$.**
- **Usually this rate caps the rate of the target function - see the discussion in e.g. *Chernozhukov et al.*, 2018 (Double Machine Learning).**

We show that in the fixed features setting, the target function converges at almost $n^{-\frac{1}{2(1+p)}}$ rate as long as the nuisance functions converge at $n^{-1/4}$ rate.

Practical algorithm

- Stage 1: Learn parameters of $\hat{m}_\theta(\mathbf{X})$

- Stage 2: Alternate between optimizing $\hat{g}_\psi(\mathbf{X}),\ \hat{h}_\phi(\mathbf{A})$ and $\hat{e}^h_\eta(\mathbf{X})$

  A. Freeze $\hat{m}_\theta(\mathbf{X})$ and $\hat{e}^h_\eta(\mathbf{X})$ to optimize $\hat{g}_\psi(\mathbf{X}), \hat{h}_\phi(\mathbf{A})$

  B. Freeze $\hat{m}_\theta(\mathbf{X})$ and $\hat{g}_\psi(\mathbf{X}), \hat{h}_\phi(\mathbf{T})$ to optimize $\hat{e}^h_\eta(\mathbf{X})$

Slide credit: Jean Kaddour

Experimental Setup

- **Data**: Two semi-synthetic datasets involving graph-treatments

Small-World (SW)

X: Samples from multivar. uniform dist.

T: Watts–Strogatz small-world graphs

The Cancer Genomic Atlas (TCGA)[1]

X: Gene expression data of cancer patients

1 | Data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

2 | L. Ruddigkeit, et al., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, 2012.

3 | Harada & Kashima, GraphITE: Estimating Individual Effects of Graph-structured Treatments, 2020.

Slide credit: Jean Kaddour

Experimental Setup

- **Data**: Two semi-synthetic datasets involving graph-treatments

Small-World (SW)
X: Samples from multivar. uniform dist.
T: Watts–Strogatz small-world graphs

The Cancer Genomic Atlas (TCGA)[1]
X: Gene expression data of cancer patients

- **Tasks**: Predicting in-sample/out-sample CATEs

1 | Data generated by the TCGA Research Network: https://www.cancer.gov/tcga.
2 | L. Ruddigkeit, et al., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, 2012.
3 | Harada & Kashima, GraphITE: Estimating Individual Effects of Graph-structured Treatments, 2020.

Slide credit: Jean Kaddour

Experimental Setup

- **Data**: Two semi-synthetic datasets involving graph-treatments

Small-World (SW)
X: Samples from multivar. uniform dist.
T: Watts–Strogatz small-world graphs

The Cancer Genomic Atlas (TCGA)[1]
X: Gene expression data of cancer patients

- **Tasks**: Predicting in-sample/out-sample CATEs

- **Baselines:** GraphITE[3], **Vanilla Regression (**GNN/CAT), Zero

1 | Data generated by the TCGA Research Network: https://www.cancer.gov/tcga.
2 | L. Ruddigkeit, et al., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, 2012.
3 | Harada & Kashima, GraphITE: Estimating Individual Effects of Graph-structured Treatments, 2020.

Slide credit: Jean Kaddour

Experimental Setup

- **Data**: Two semi-synthetic datasets involving graph-treatments

Small-World (SW)
X: Samples from multivar. uniform dist.
T: Watts–Strogatz small-world graphs

The Cancer Genomic Atlas (TCGA)[1]
X: Gene expression data of cancer patients

- **Tasks**: Predicting in-sample/out-sample CATEs

- **Baselines:** GraphITE[3], **Vanilla Regression (**GNN/CAT), Zero

- **Metric**: (Un-)Weighted expected Precision in Estimation of Het. Effects

$$\epsilon_{\mathrm{UPEHE}(\mathrm{WPEHE})} \triangleq \int_{\mathcal{X}} \left(\widehat{\tau}\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right) - \tau\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right)\right)^2 p(\mathbf{t} \mid \mathbf{x}) p\left(\mathbf{t}' \mid \mathbf{x}\right) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

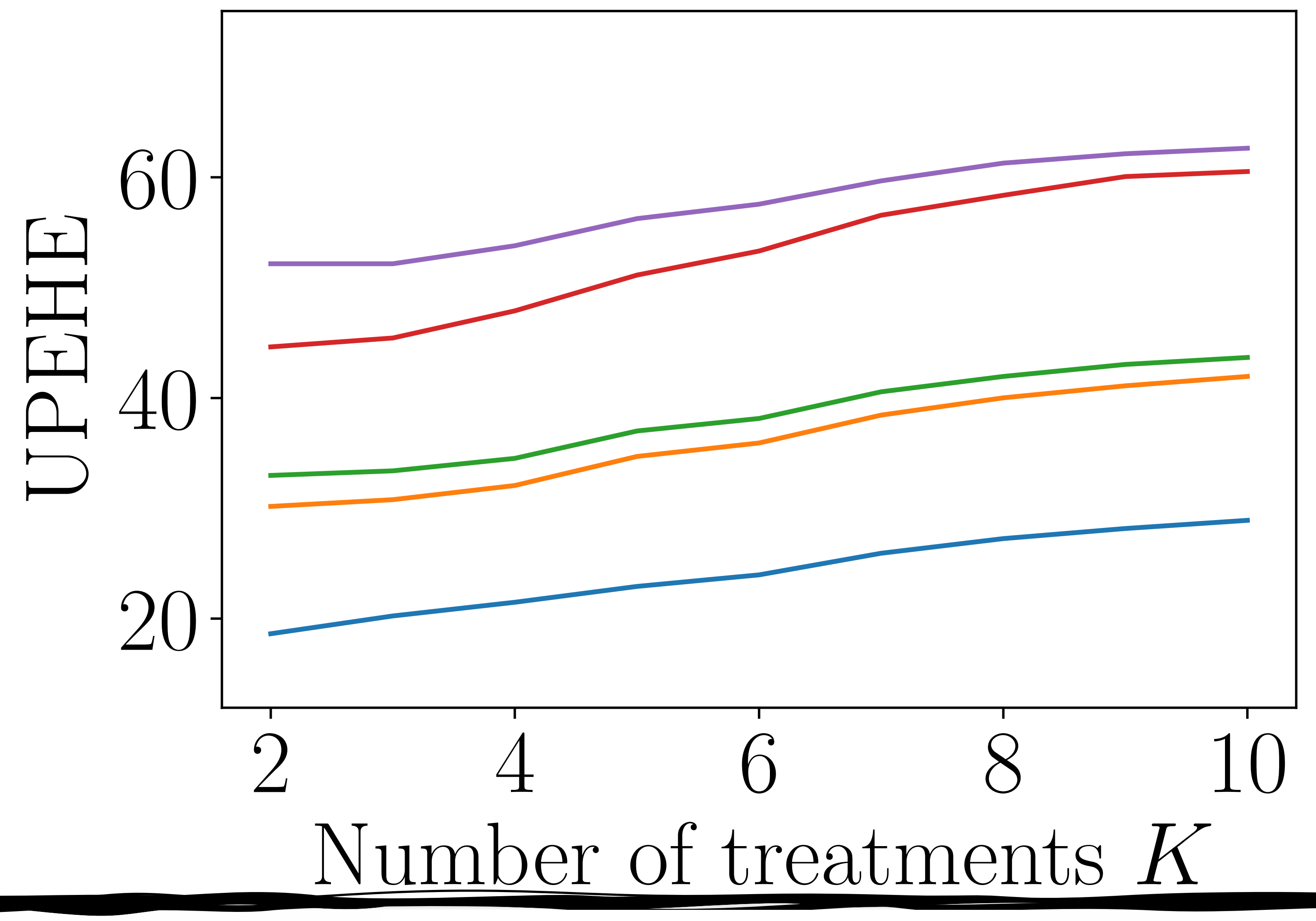
1 | Data generated by the TCGA Research Network: https://www.cancer.gov/tcga.
2 | L. Ruddigkeit, et al., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, 2012.
3 | Harada & Kashima, GraphITE: Estimating Individual Effects of Graph-structured Treatments, 2020.

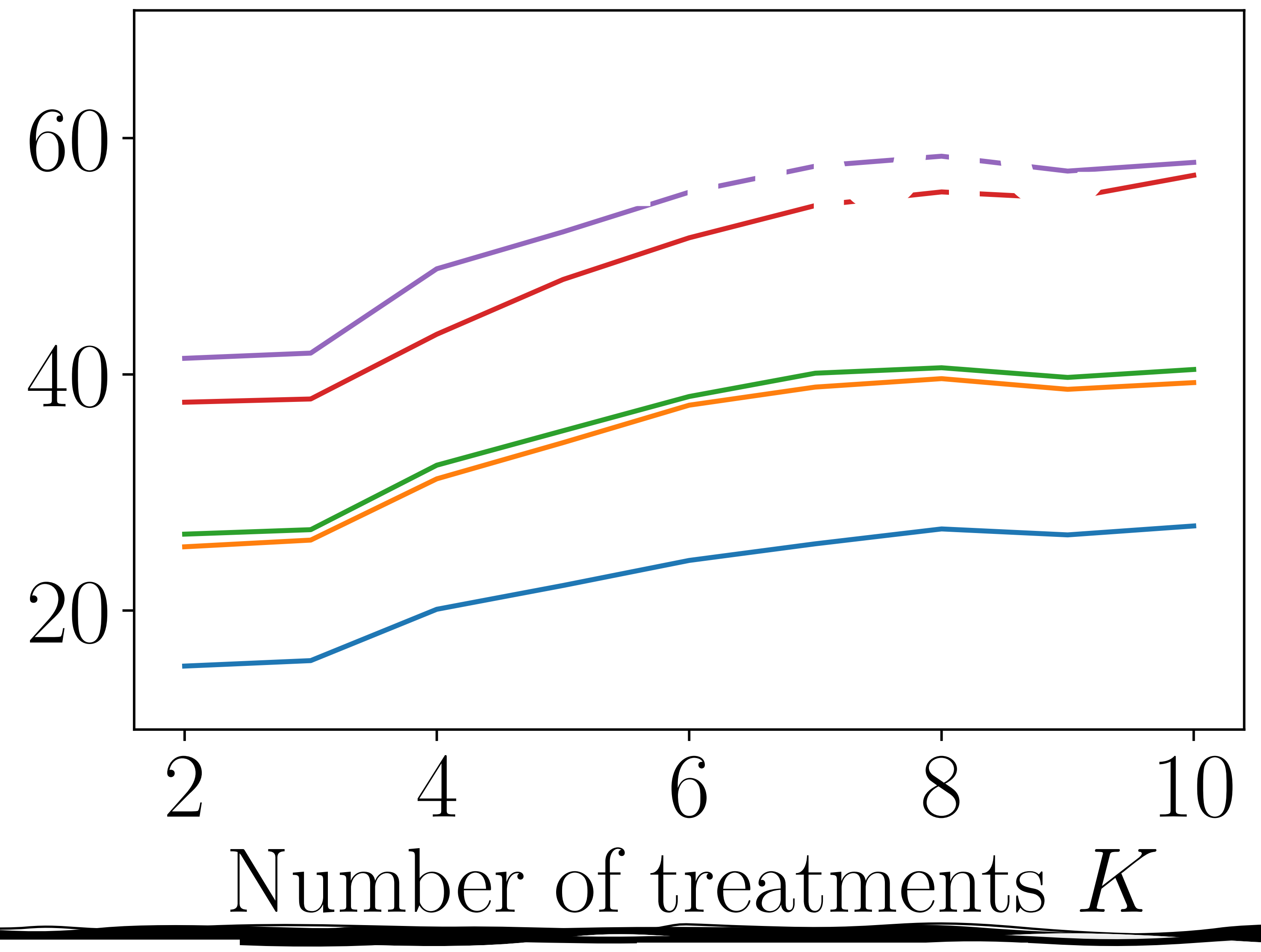
Slide credit: Jean Kaddour

Results: **In-Sample** **Out-Sample**

Results: **In-Sample** **Out-**

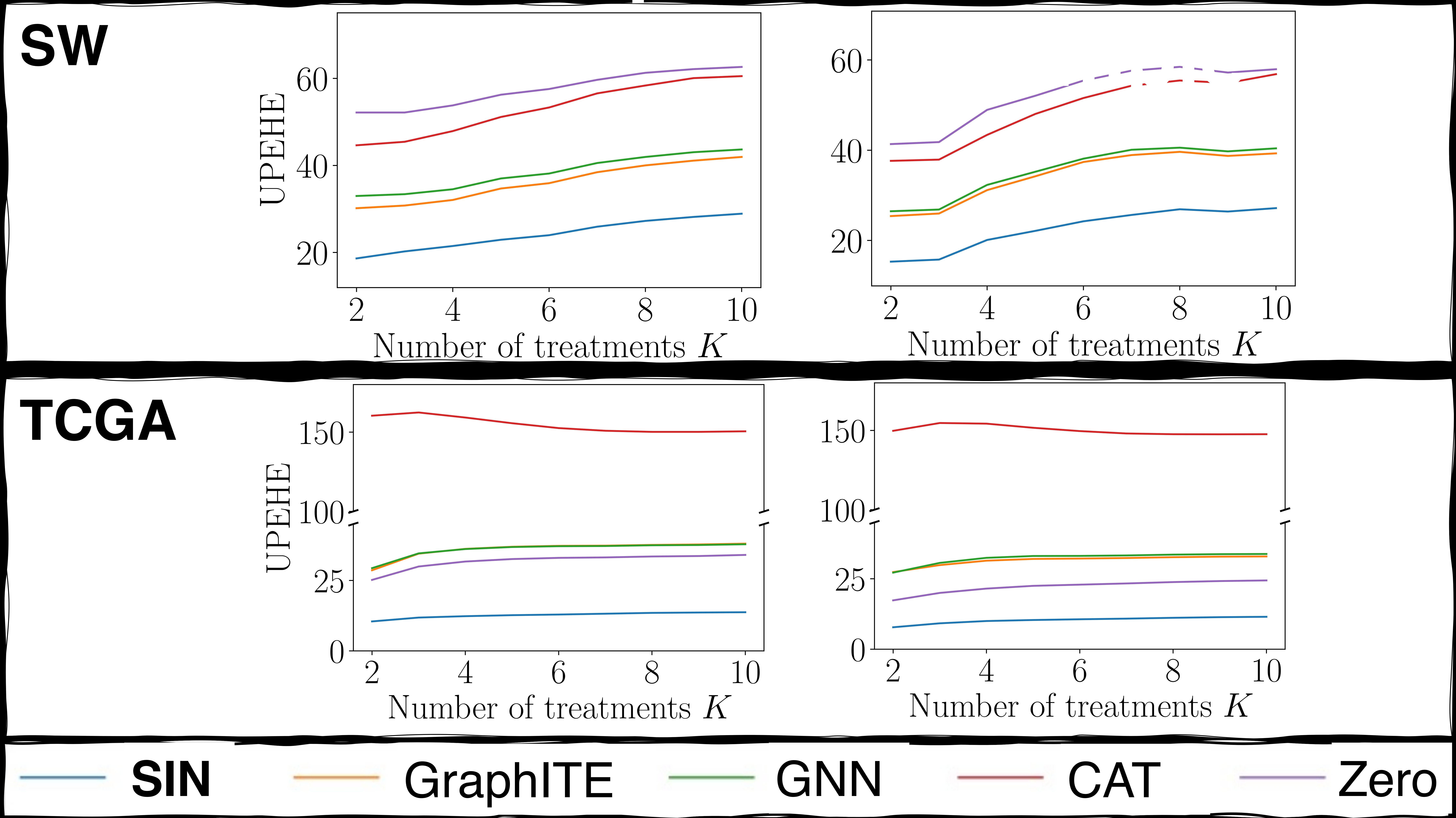
SW





Slide credit: Jean Kaddour

Results: In-Sample Out-



Slide credit: Jean Kaddour

WPEHE for most likely K=6 treatments

Method	SW		TCGA	
	In-sample	Out-sample	In-sample	Out-sample
Zero	56.26 ± 8.12	53.77 ± 8.93	26.63 ± 7.55	17.94 ± 4.86
CAT	51.75 ± 8.85	49.76 ± 9.73	155.88 ± 52.82	146.62 ± 42.32
GNN	37.10 ± 6.84	36.74 ± 7.42	30.67 ± 8.29	27.57 ± 7.95
GraphITE	34.81 ± 6.70	35.94 ± 8.07	30.31 ± 8.96	27.48 ± 8.95
SIN	$\mathbf{23.00 \pm 4.56}$	$\mathbf{23.19 \pm 5.56}$	$\mathbf{10.98 \pm 3.45}$	$\mathbf{8.15 \pm 1.46}$

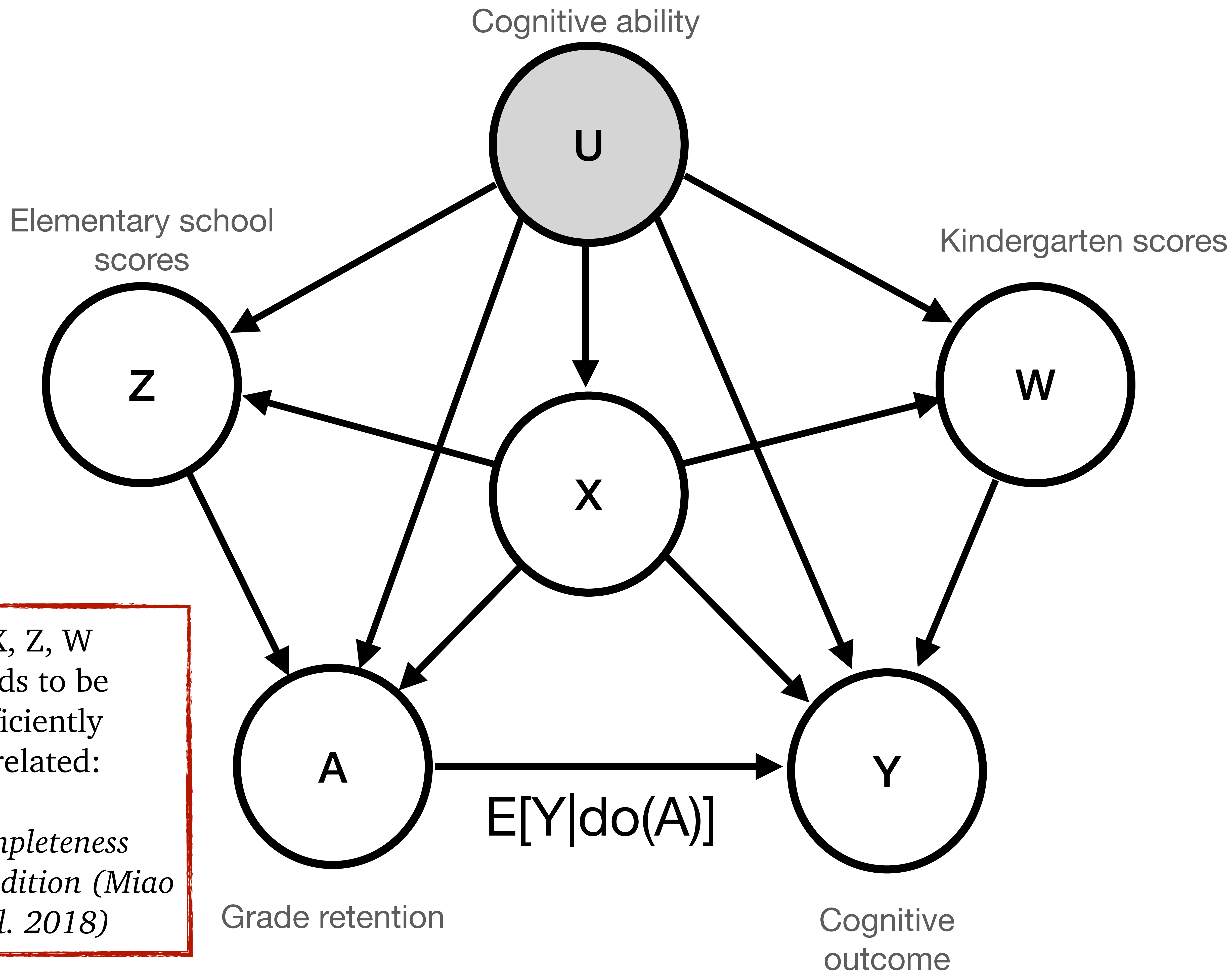
Take home messages

- This is an algorithm that can take arbitrary treatments: categorical, continuous, structural….
- The structures in the ‘structural’ treatments do NOT have to be causal!
  - Only needs to model causal relationships when we need to ask about interventions on it.
- Fast rates from partially out the nuisance parameters.

Proximal Causal Learning with Kernels

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

Assumptions



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

Completeness Condition (Miao et al. 2018)

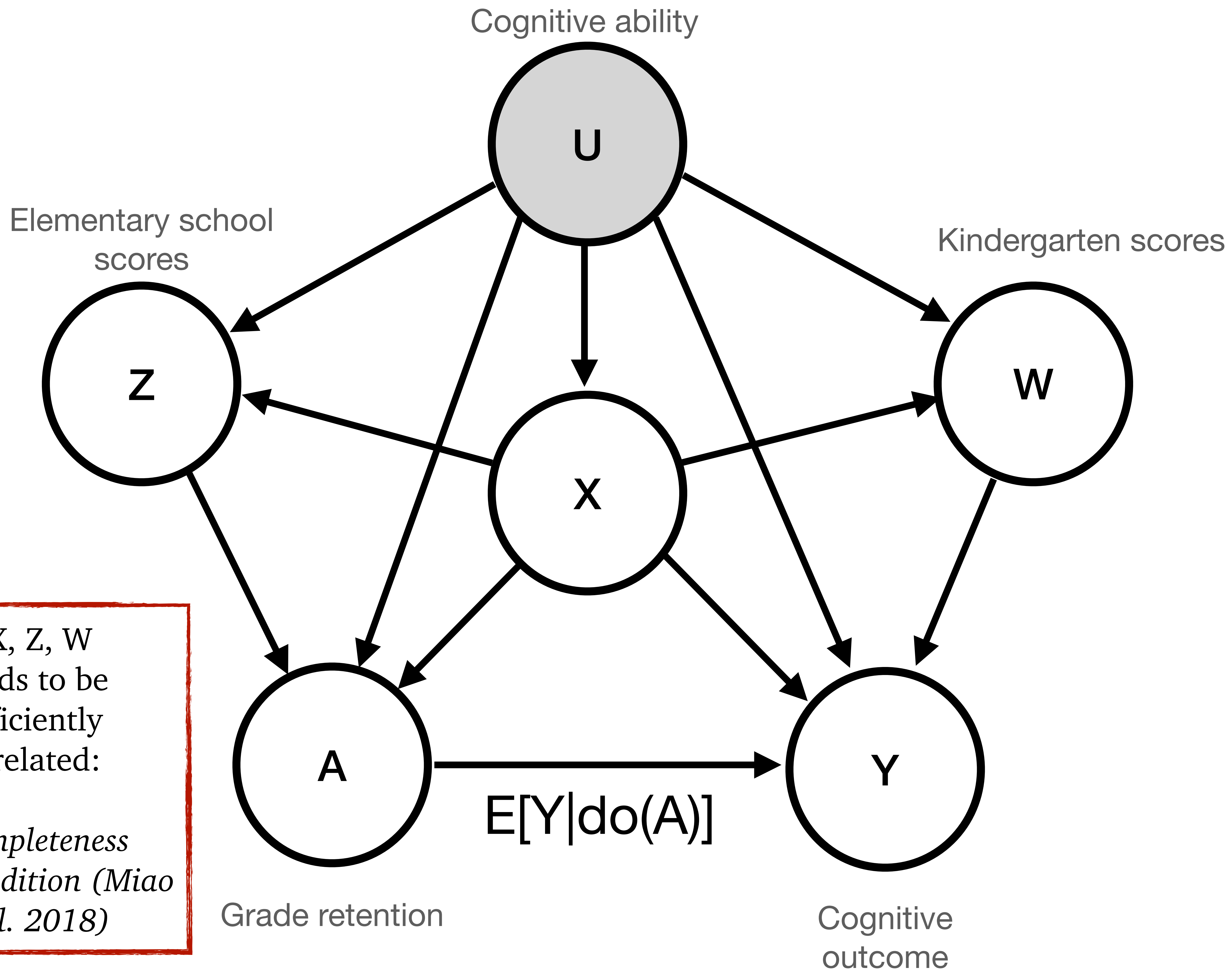
U and X contains all the confounders between A and Y.

$$Y \perp Z \mid A, U, X$$

$$W \perp (A, Z) \mid U, X$$

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

Proximal Causal Learning Background



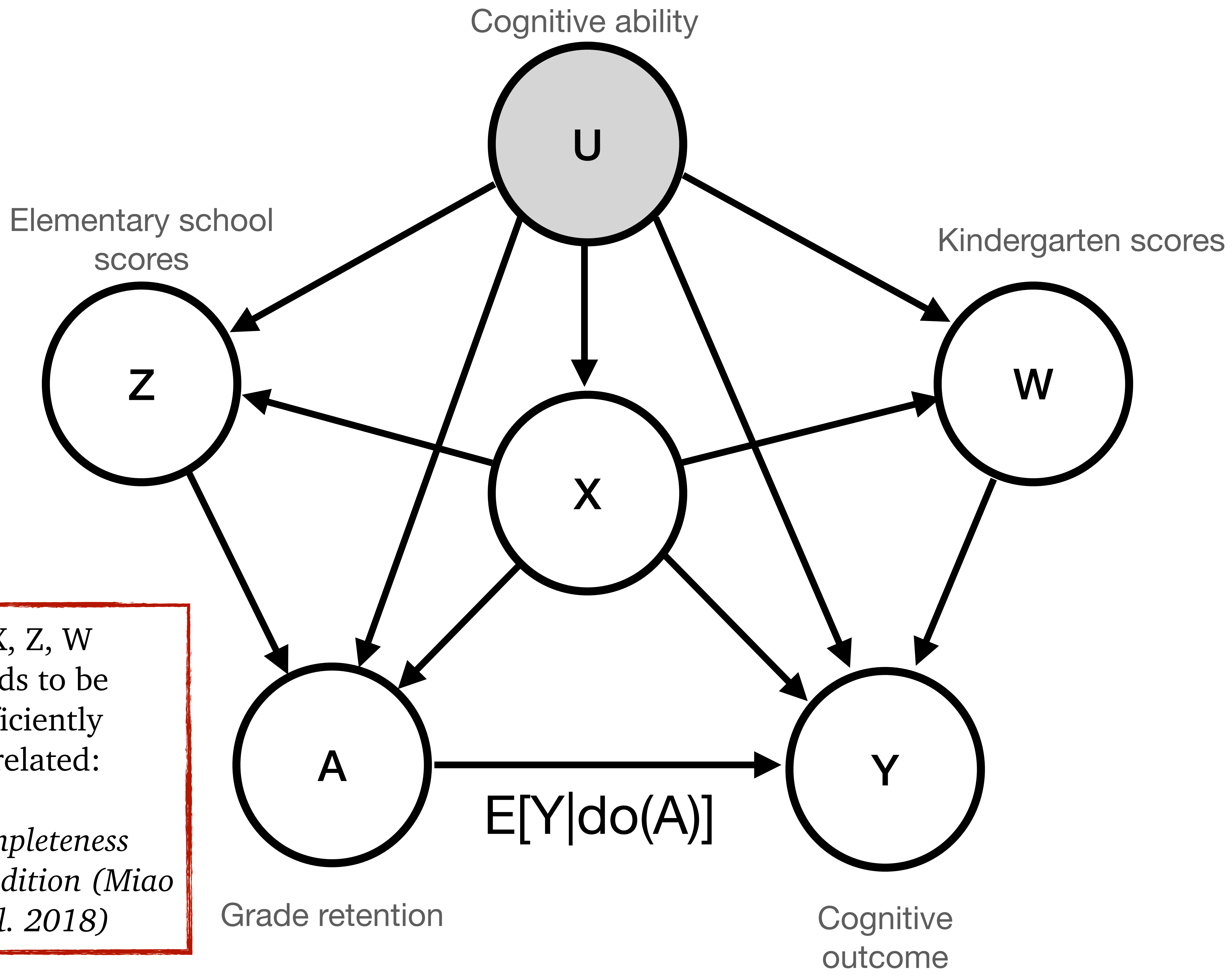
U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

$$Y = \beta_0 + \beta_a A + \beta_u U + \beta_x' X + \epsilon_y$$
$$W = \eta_0 + \eta_u U + \eta_x' X + \epsilon_w$$

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

Proximal Causal Learning Background



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

Completeness Condition (Miao et al. 2018)

$$Y = \beta_0 + \beta_a A + \beta_u U + \beta_x' X + \epsilon_y$$

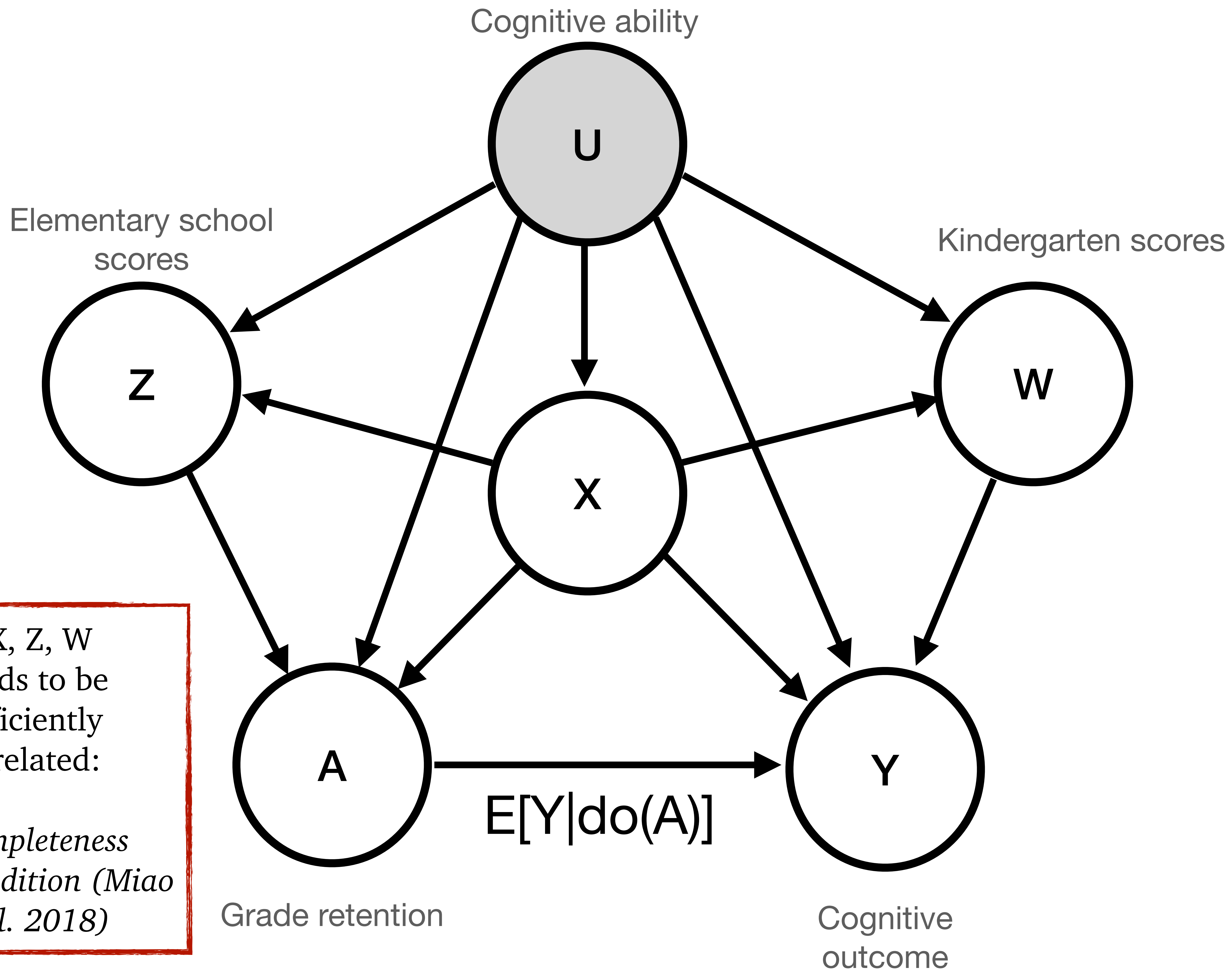
$$W = \eta_0 + \eta_u U + \eta_x' X + \epsilon_w$$

$$\mathbb{E}[Y | A, Z, X] = \beta_0 + \beta_a A + \beta_u \mathbb{E}[U | A, Z, X] + \beta_x' X$$

$$\mathbb{E}[W | A, Z, X] = \eta_0 + \eta_u \mathbb{E}[U | A, Z, X] + \eta_x' X$$

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

Proximal Causal Learning Background



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

Completeness Condition (Miao et al. 2018)

$$Y = \beta_0 + \beta_a A + \beta_u U + \beta_x' X + \epsilon_y$$

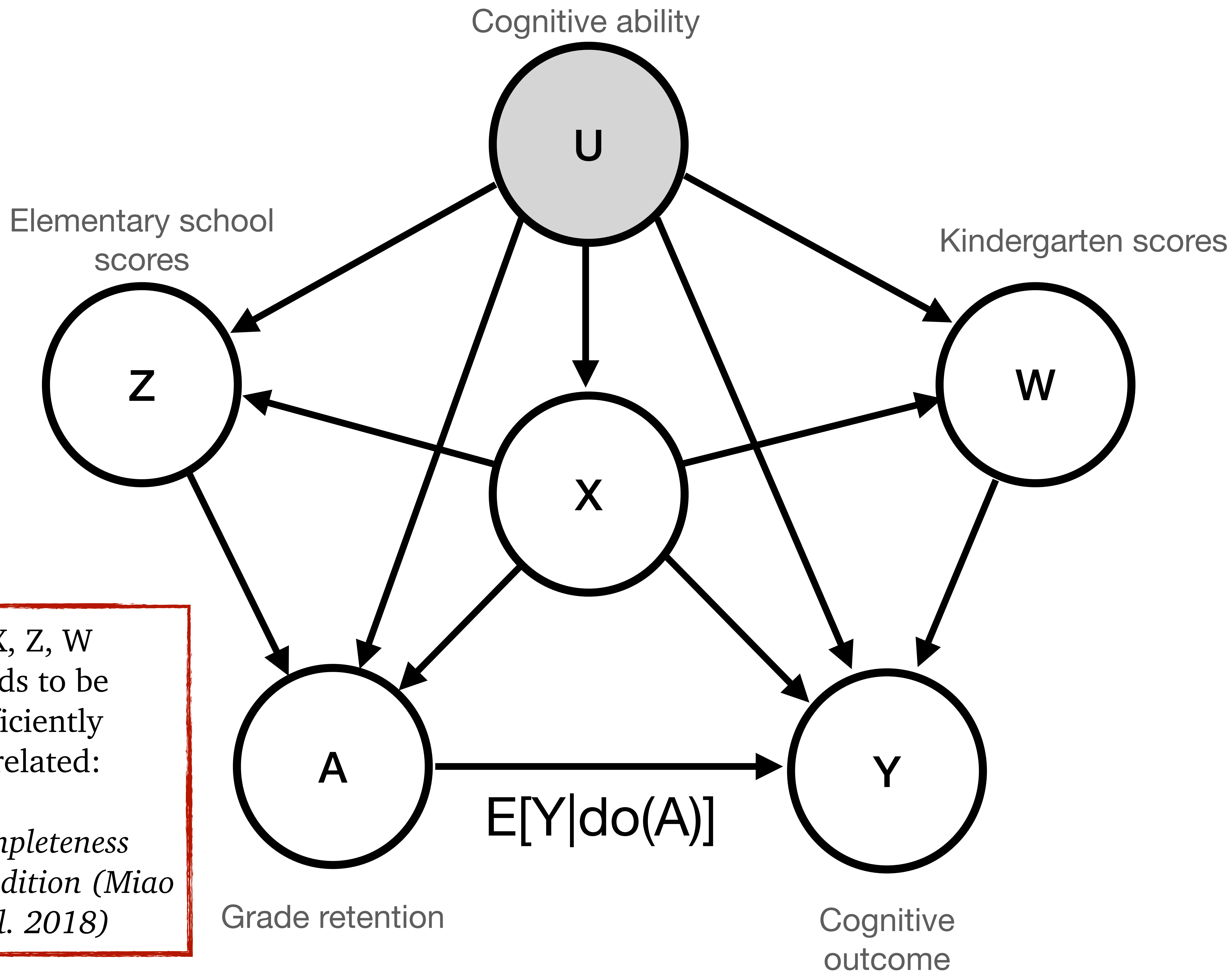
$$W = \eta_0 + \eta_u U + \eta_x' X + \epsilon_w$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0 + \beta_a A + \beta_u \mathbb{E}[U|A,Z,X] + \beta_x' X$$

$$\mathbb{E}[W|A,Z,X] = \eta_0 + \eta_u \mathbb{E}[U|A,Z,X] + \eta_x' X$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0^* + \beta_a A + \beta_u^* \mathbb{E}[W|A,Z,X] + (\beta_x^*)' X$$

Proximal Causal Learning Background



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

Completeness Condition (Miao et al. 2018)

$$Y = \beta_0 + \beta_a A + \beta_u U + \beta_x' X + \epsilon_y$$

$$W = \eta_0 + \eta_u U + \eta_x' X + \epsilon_w$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0 + \beta_a A + \beta_u \mathbb{E}[U|A,Z,X] + \beta_x' X$$

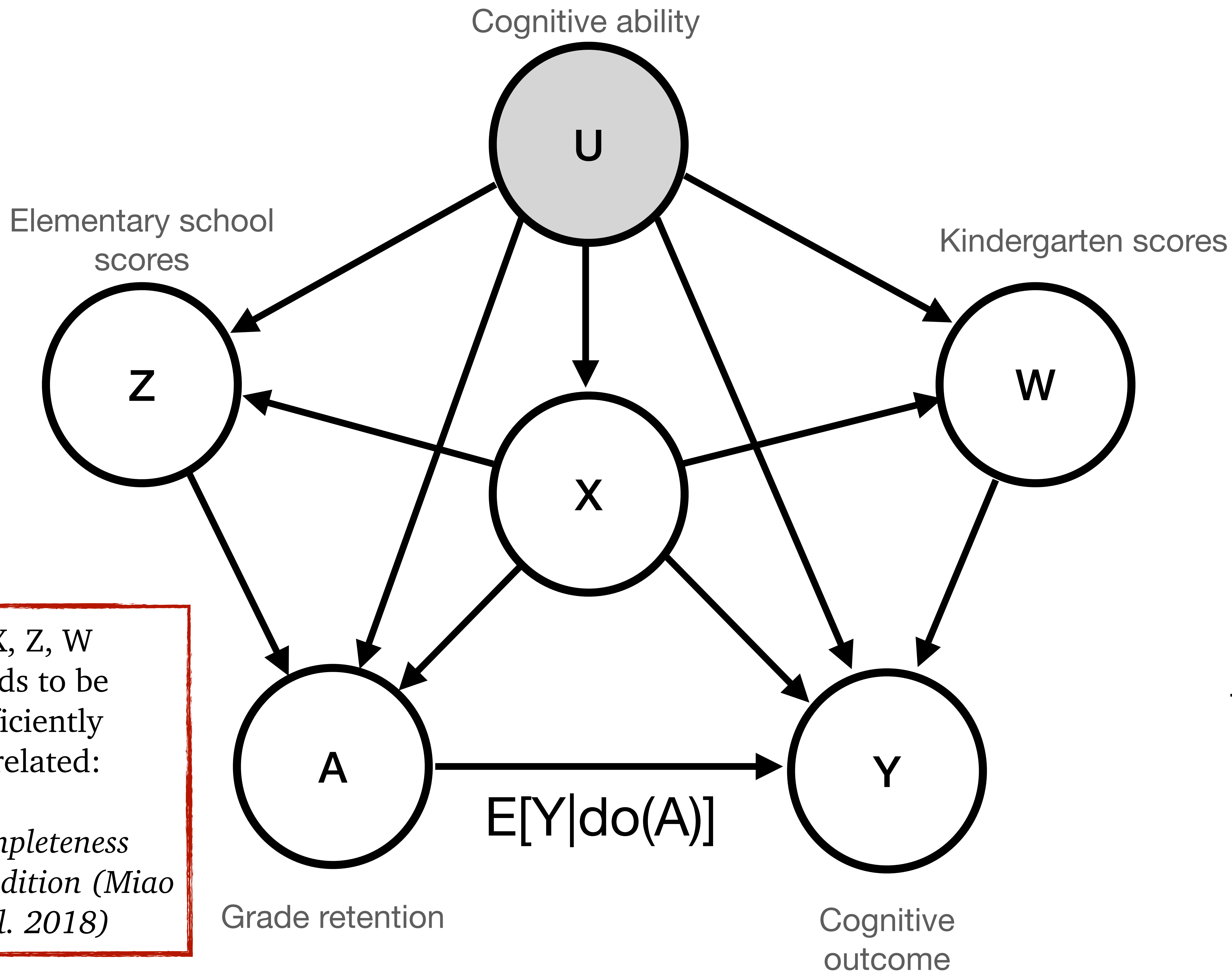
$$\mathbb{E}[W|A,Z,X] = \eta_0 + \eta_u \mathbb{E}[U|A,Z,X] + \eta_x' X$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0^* + \beta_a A + \beta_u^* \mathbb{E}[W|A,Z,X] + (\beta_x^*)' X$$

$$Y = \underbrace{\beta_0^* + \beta_a A + \beta_u^* W + (\beta_x^*)' X}_{h_{linear}} + \epsilon^* \qquad \mathbb{E}[\epsilon^* | A, Z, X] = 0$$

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

Proximal Causal Learning Background



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

Completeness Condition (Miao et al. 2018)

$$Y = \beta_0 + \beta_a A + \beta_u U + \beta_x' X + \epsilon_y$$

$$W = \eta_0 + \eta_u U + \eta_x' X + \epsilon_w$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0 + \beta_a A + \beta_u \mathbb{E}[U|A,Z,X] + \beta_x' X$$

$$\mathbb{E}[W|A,Z,X] = \eta_0 + \eta_u \mathbb{E}[U|A,Z,X] + \eta_x' X$$

$$\mathbb{E}[Y|A,Z,X] = \beta_0^* + \beta_a A + \beta_u^* \mathbb{E}[W|A,Z,X] + (\beta_x^*)' X$$

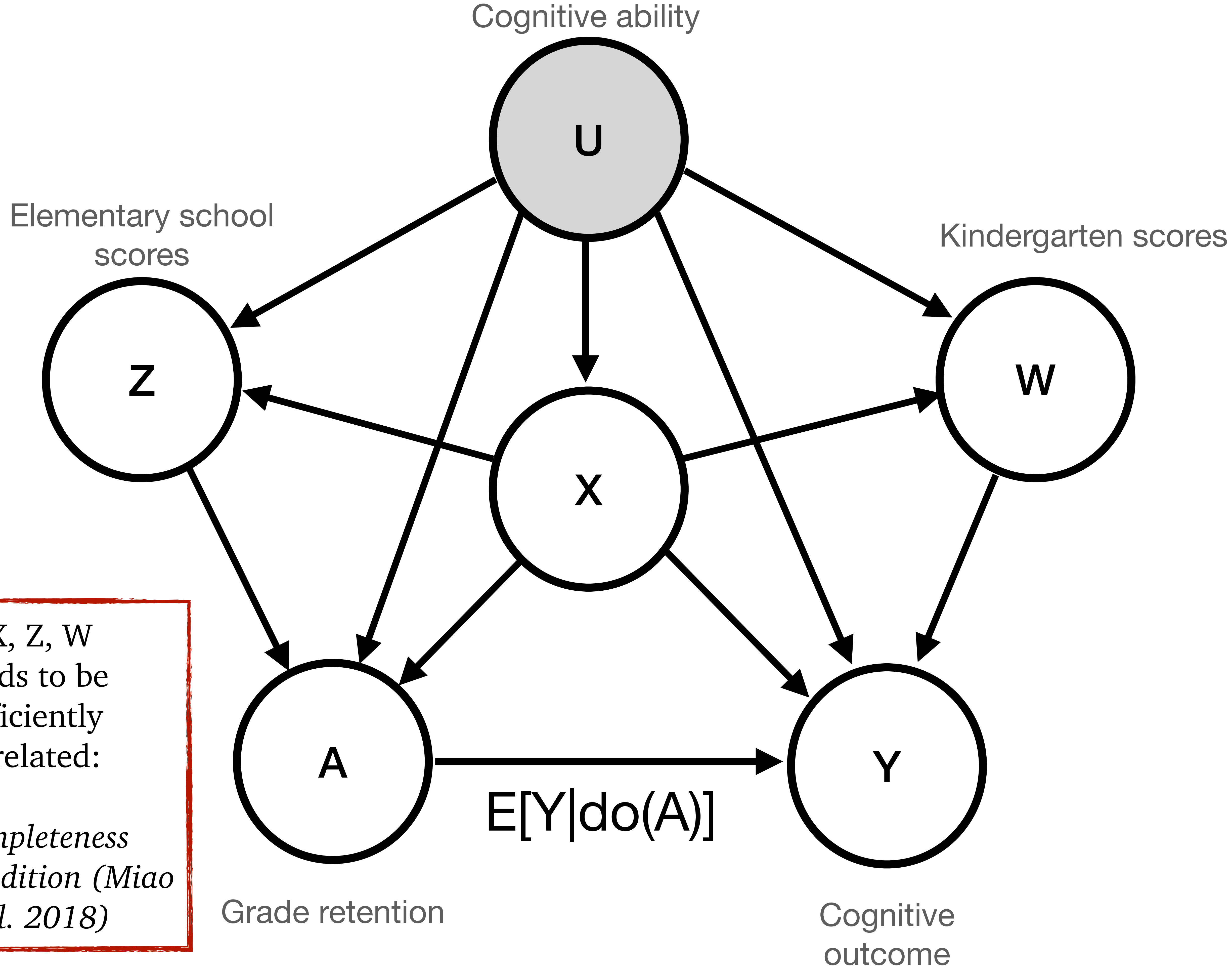
$$Y = \underbrace{\beta_0^* + \beta_a A + \beta_u^* W + (\beta_x^*)' X}_{h_{linear}} + \epsilon^* \qquad \mathbb{E}[\epsilon^*|A,Z,X] = 0$$

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

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

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

Proximal Causal Learning Background



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

Completeness Condition (Miao et al. 2018)

Average causal effect estimation:

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

Where h is from:

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

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

Introduction to kernel ridge regression

Finite-basis / Featurised regression

$$f(x) = \theta^\top \phi(x), \;\; \phi(x) \in \mathbb{R}^D$$

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^D} \left(\sum_{i=1}^{n} \left(y_i - \phi(x_i)^\top \theta\right)^2 + \lambda \|\theta\|^2 \right)$$

Gretton lecture slides on Kernel Methods - lecture 1. http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/Slides4A.pdf

Introduction to kernel ridge regression

Finite-basis / Featurised regression	$f(x) = \theta^\top \phi(x), \;\; \phi(x) \in \mathbb{R}^D$ $\theta^* = \arg\min_{\theta \in \mathbb{R}^D} \left(\sum_{i=1}^{n} \left(y_i - \phi(x_i)^\top \theta \right)^2 + \lambda \|\theta\|^2 \right)$
Reproducing Kernel Hilbert Space (RKHS)	$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \;\; \phi(x) \in \mathcal{H}, \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$ $f^* = \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \phi(x_i), f \rangle_{\mathcal{H}} \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right)$

Gretton lecture slides on Kernel Methods - lecture 1. http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/Slides4A.pdf

Proximal Causal Learning Background

Solve for h:

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

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

Proximal Causal Learning Background

Solve for h:

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

$$h \in \mathcal{H}_{AWX}$$

$$h(A, W, X) = \langle h, \phi(A) \otimes \phi(W) \otimes \phi(X) \rangle_{\mathcal{H}_{AWX}}$$

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

Proximal Causal Learning Background

Solve for h:

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

$$h \in \mathscr{H}_{AWX}$$

$$h(A, W, X) = \langle h, \phi(A) \otimes \phi(W) \otimes \phi(X) \rangle_{\mathscr{H}_{AWX}}$$

$$\mathbb{E}[h(A, W, X) \mid A, Z, X] = \langle h, \phi(A) \otimes \underbrace{\mathbb{E}[\phi(W) \mid A, Z, X]}_{\mu_{W|A,Z,X}} \otimes \phi(X) \rangle_{\mathscr{H}_{AWX}}$$

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

Introduction to kernel ridge regression

	Definition	Learning
Finite basis:	$f(x) = \theta^\top \phi(x)$ $\mathbb{E}[f(X)\,\vert\,Z] = \theta^\top \mathbb{E}[\phi(x)\,\vert\,Z]$	$\mathbb{E}[\phi(x)\,\vert\,Z] = \Theta^\top \psi(Z)$ $\Theta^* = \arg\min_{\Theta \in \mathbb{R}^{D_Z \times D_X}} \left(\sum_{i=1}^{n} \left\Vert \phi(x_i) - \Theta^\top \psi(z_i) \right\Vert^2 + \lambda \Vert\Theta\Vert_2^2 \right)$
RKHS basis:	$f(x) = < f, \phi(x) >_{\mathscr{H}_X}$ $\mathbb{E}[f(X)\,\vert\,Z] = < f, \underbrace{\mathbb{E}[\phi(X)\,\vert\,Z]}_{\mu_{X\vert Z}} >_{\mathscr{H}_X}$	$\mu_{X\vert Z} = E_\lambda^* \psi(Z)$ $E_\lambda = \arg\min_{E \in L_2(\mathscr{H}_\mathcal{X}, \mathscr{H}_\mathcal{Z})} \left(\sum_{i=1}^{n} \left\Vert \phi(x_i) - E^* \psi(z_i) \right\Vert_{\mathscr{H}_\mathcal{X}}^2 + \lambda \Vert E \Vert_{L_2(\mathscr{H}_\mathcal{X}, \mathscr{H}_\mathcal{Z})}^2 \right)$

[1] Gretton lecture slides on Kernel Methods - lecture 4. http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture5_distribEmbed_1.pdf
[2] Singh et al 2019. Kernel Instrumental Variable Regression.

Kernel Proxy Variables

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

Kernel Proxy Variable (KPV)

$$Stage1.\ KRR : \phi(A) \otimes \phi(X) \otimes \phi(Z) \rightarrow \phi(W)$$

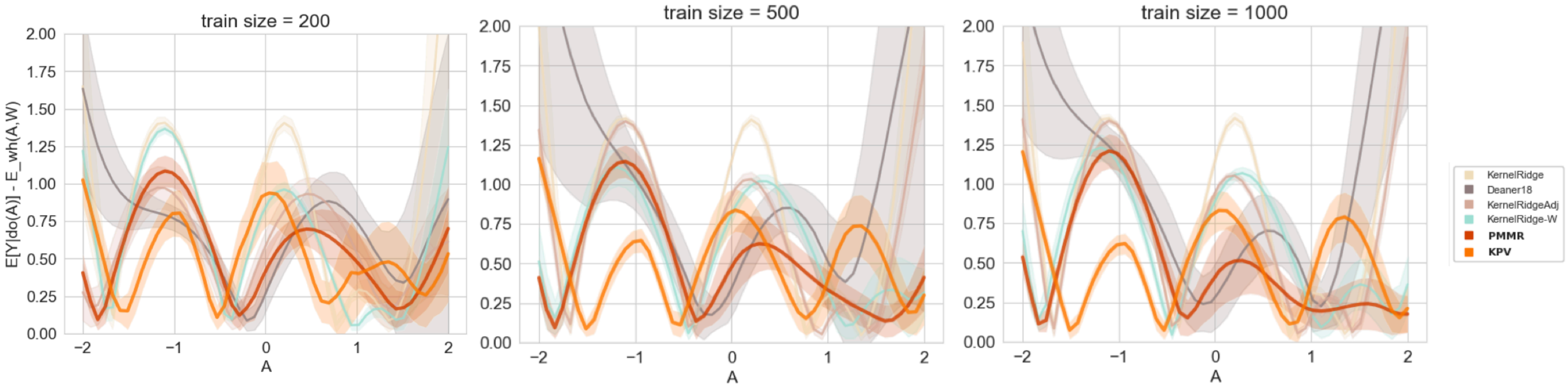
$$Stage2.\ KRR : \phi(A) \otimes \phi(X) \otimes \hat{\mu}_{W|A,X,Z} \rightarrow Y$$

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

Results

Under suitable conditions specified in the paper, KPV provably converges.

Synthetic experiments



However, empirically it might be better to learn adaptive features rather than using fixed kernel features.

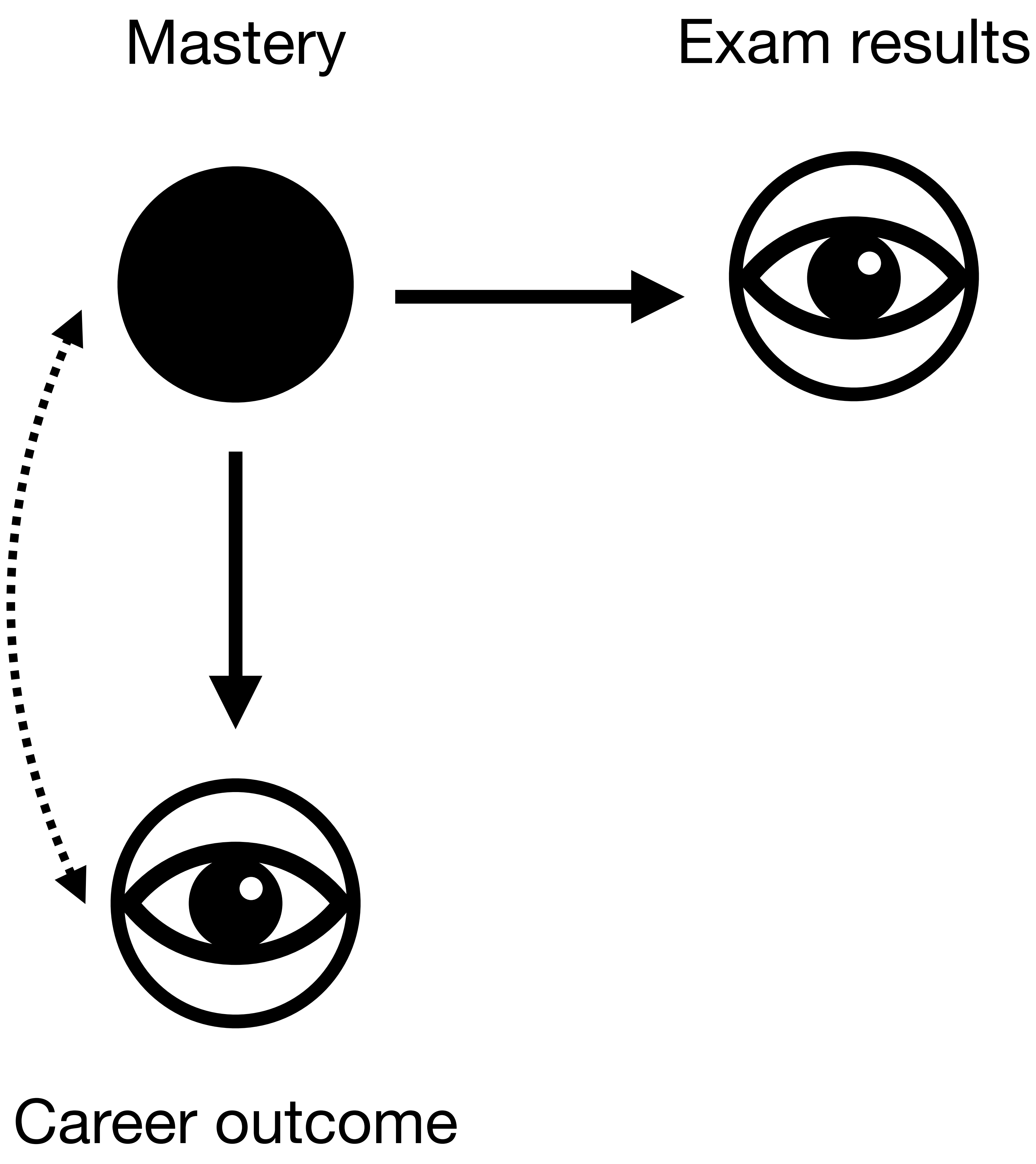
[1] Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-stage Estimation and Moment Restriction. ICML 2021.
[2] Xu, et al. Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation. NeurIPS 2021.

Take home messages

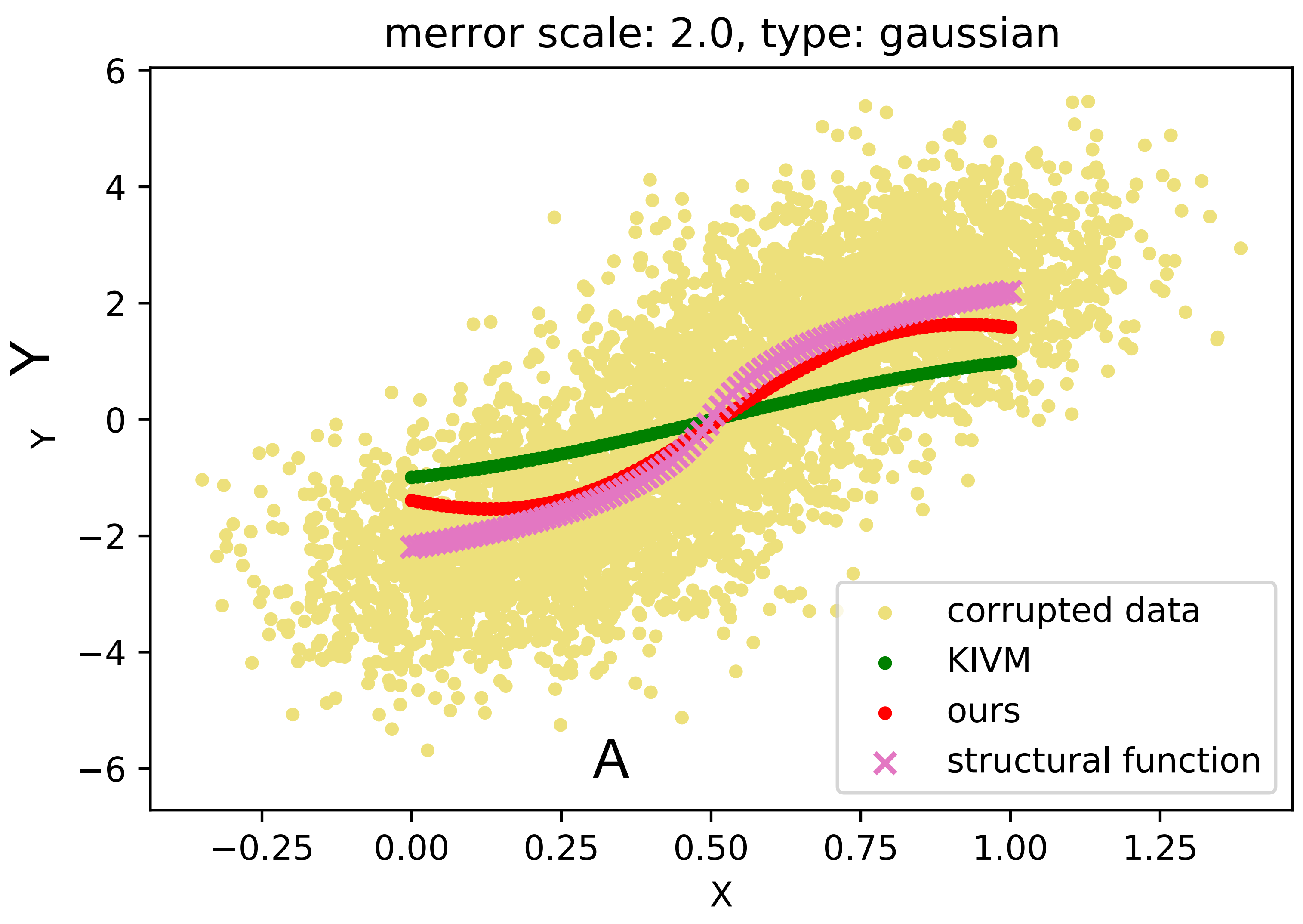
- **This is an algorithm allowing nonlinear treatment effect estimation under unobserved confounding, with theoretical convergence rates.**
- **The conditions are weak because only partial knowledge of the graph is needed.**
  - **Only need to categorise the proxies, do not need to know their own causal structures.**

Causal Inference Under Treatment Measurement Error

Flash back: The characteristics of social science data



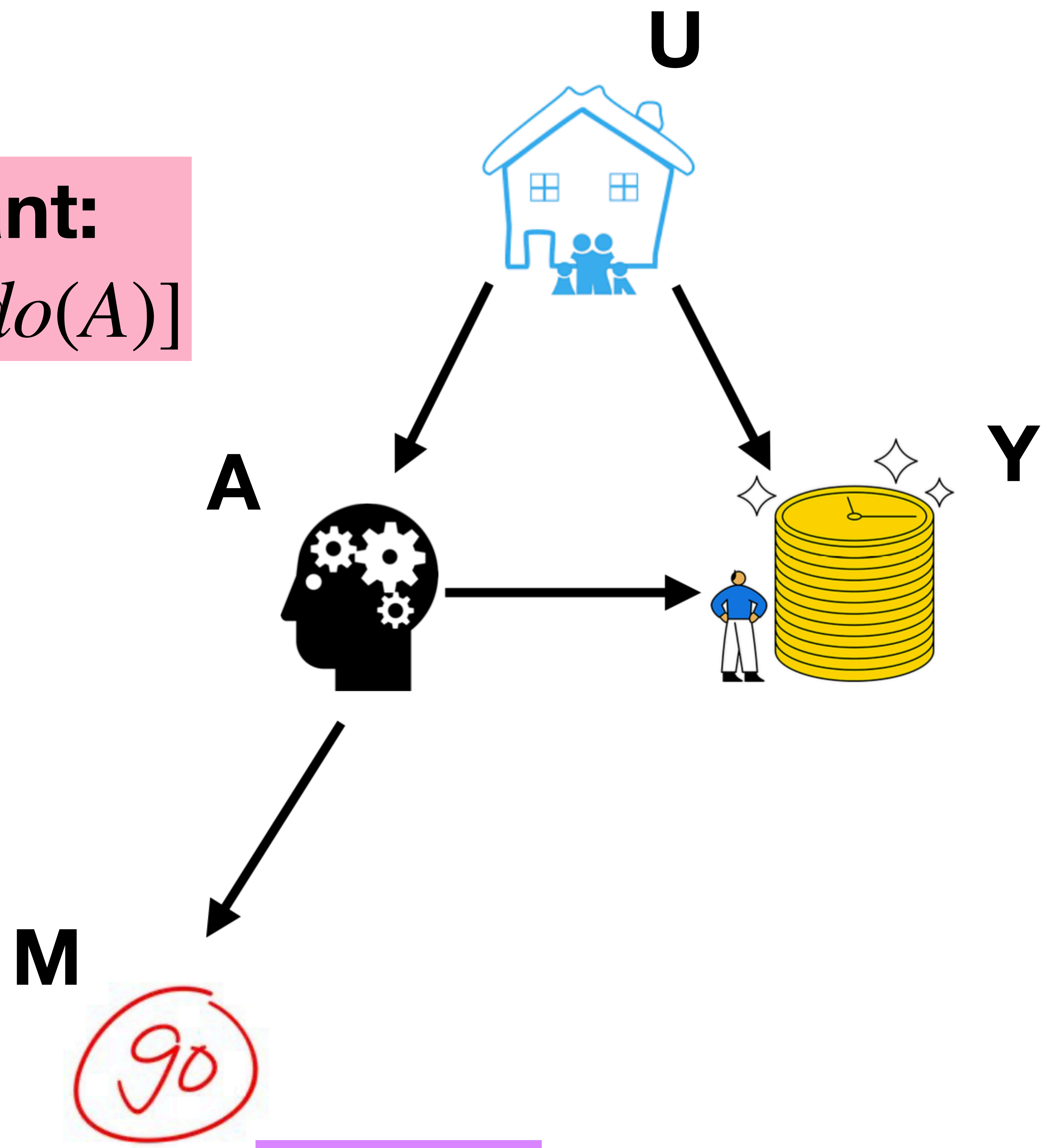
Mask interesting relationships:



Measurement error on action variables - overview

Want:

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

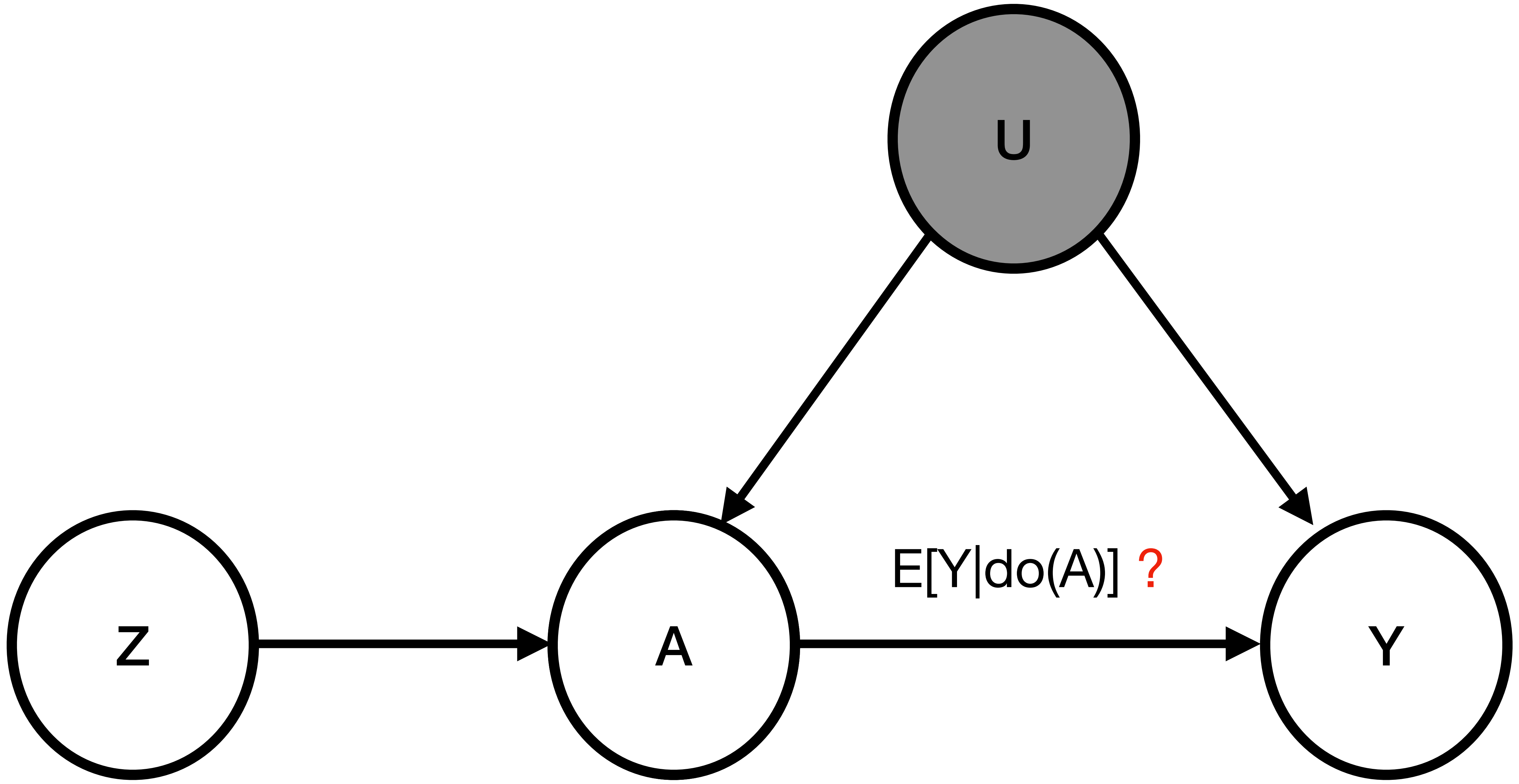


$$M = A + \Delta M$$
$$N = A + \Delta N$$

Kotlarski's:

$$\overbrace{\mathbb{E}_{\mathscr{P}_A}[e^{i\alpha A}]}^{\psi_A(\alpha)} = \exp\left(\int_0^\alpha i \frac{\mathbb{E}[M e^{i\nu N}]}{\mathbb{E}[e^{i\nu N}]} d\nu \right)$$

Recap: Identification with instrumental variables



Identification:

$$Y = f(A) + U \quad \mathbb{E}[U \,|\, Z] = 0$$

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

$$\mathbb{E}[Y \,|\, Z] = \int_{\mathcal{A}} f(a) p(a \,|\, Z) da$$

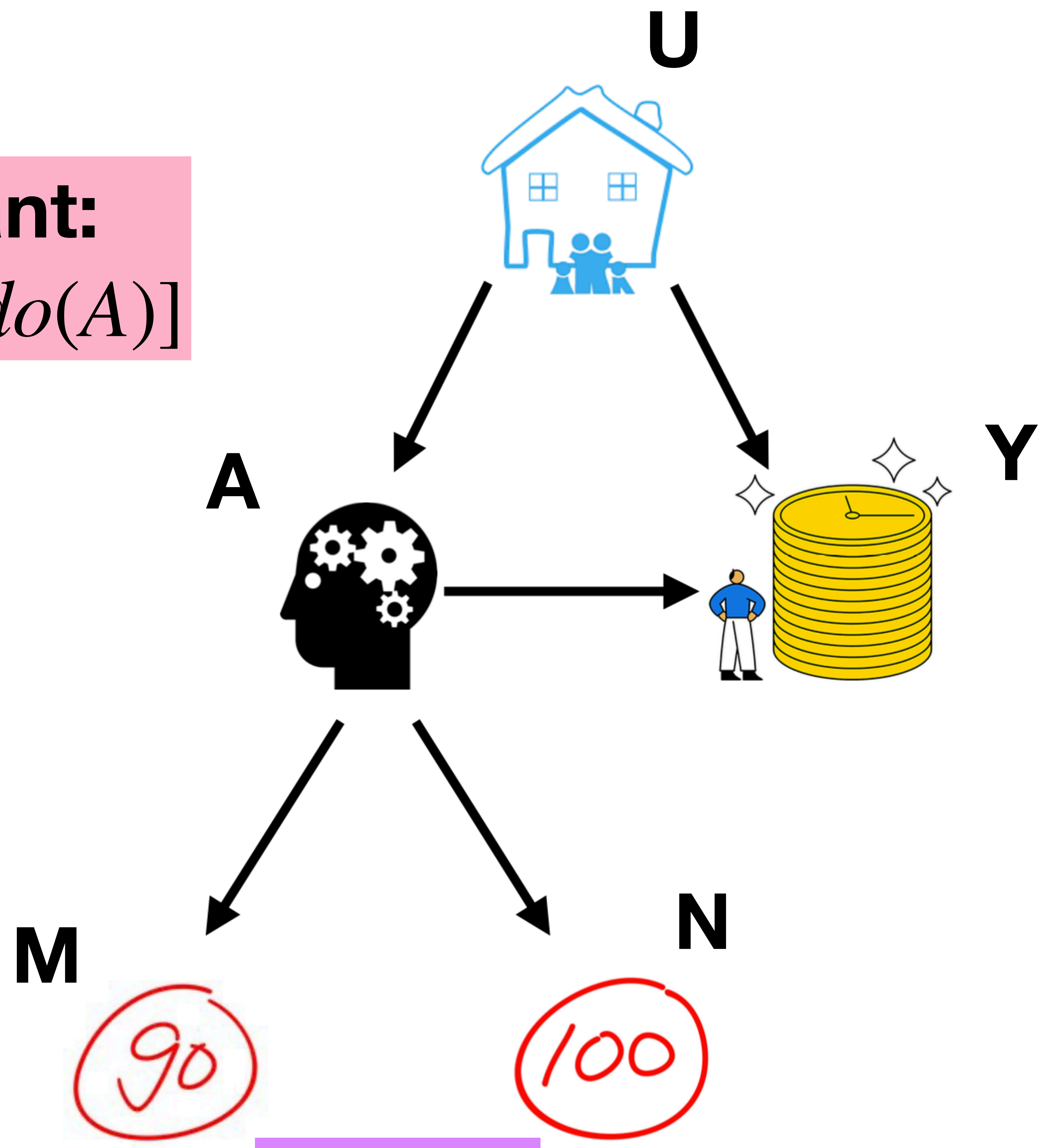
???

But if $f(a) = \langle f, \phi(a) \rangle_{\mathcal{H}_A}$, then rhs simplies to

$$\mathbb{E}[Y \,|\, Z] = \langle f, \underbrace{\mathbb{E}[\phi(A) \,|\, Z]}_{\mu_{A|Z}} \rangle_{\mathcal{H}_A}$$

Measurement error on action variables - overview

Want:
$\mathbb{E}[Y \mid do(A)]$



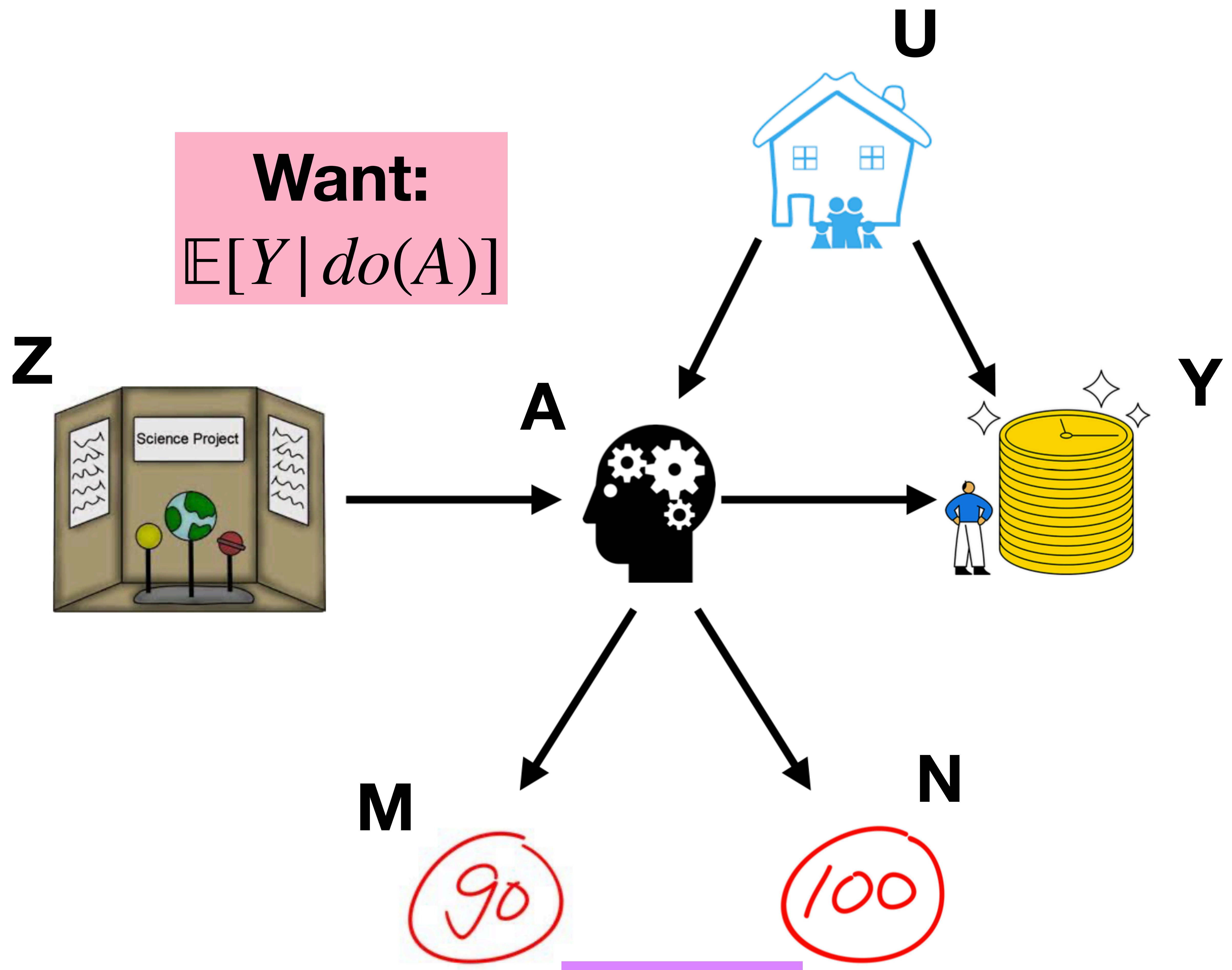
$M = A + \Delta M$
$N = A + \Delta N$

Kotlarski's:

$$\overbrace{\mathbb{E}_{\mathscr{P}_A}[e^{i\alpha A}]}^{\psi_A(\alpha)} = \exp\left(\int_0^\alpha i \frac{\mathbb{E}[Me^{i\nu N}]}{\mathbb{E}[e^{i\nu N}]} d\nu\right)$$

Measurement error on action variables - overview

Want:
$\mathbb{E}[Y \mid do(A)]$



$M = A + \Delta M$
$N = A + \Delta N$

Kotlarski's**:**

$$\overbrace{\mathbb{E}[e^{i\alpha A} \mid z]}^{\psi_{A|z}(\alpha)} = \exp\left(\int_0^\alpha i \frac{\mathbb{E}[M e^{i\nu N} \mid z]}{\mathbb{E}[e^{i\nu N} \mid z]} d\nu \right)$$

What about
$\mu_{A|z} := \mathbb{E}[\phi(A) \mid z]$**?**

From $\hat{\psi}^n_{X|z}(\alpha)$ to $\hat{\mu}^n_{X|z}(y)$ (= $\mathbb{E}[\phi(X)\,|\,z]$)

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

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

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

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

$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}](\alpha)}^{\psi_{A|z}(\alpha)} = \exp\left(\int_0^\alpha i \frac{\overbrace{\mathbb{E}[Me^{i\nu N}\,|\,z]}^{\frac{\partial}{\partial v}\psi_{M,N|z}(v,\nu)\big|_{v=0}}}{\underbrace{\mathbb{E}[e^{i\nu N}\,|\,z]}_{\psi_{N|z}(\nu)}} d\nu \right)$$

Measurement Error KIV

To obtain $\hat{\psi}^n_{A|z}$:

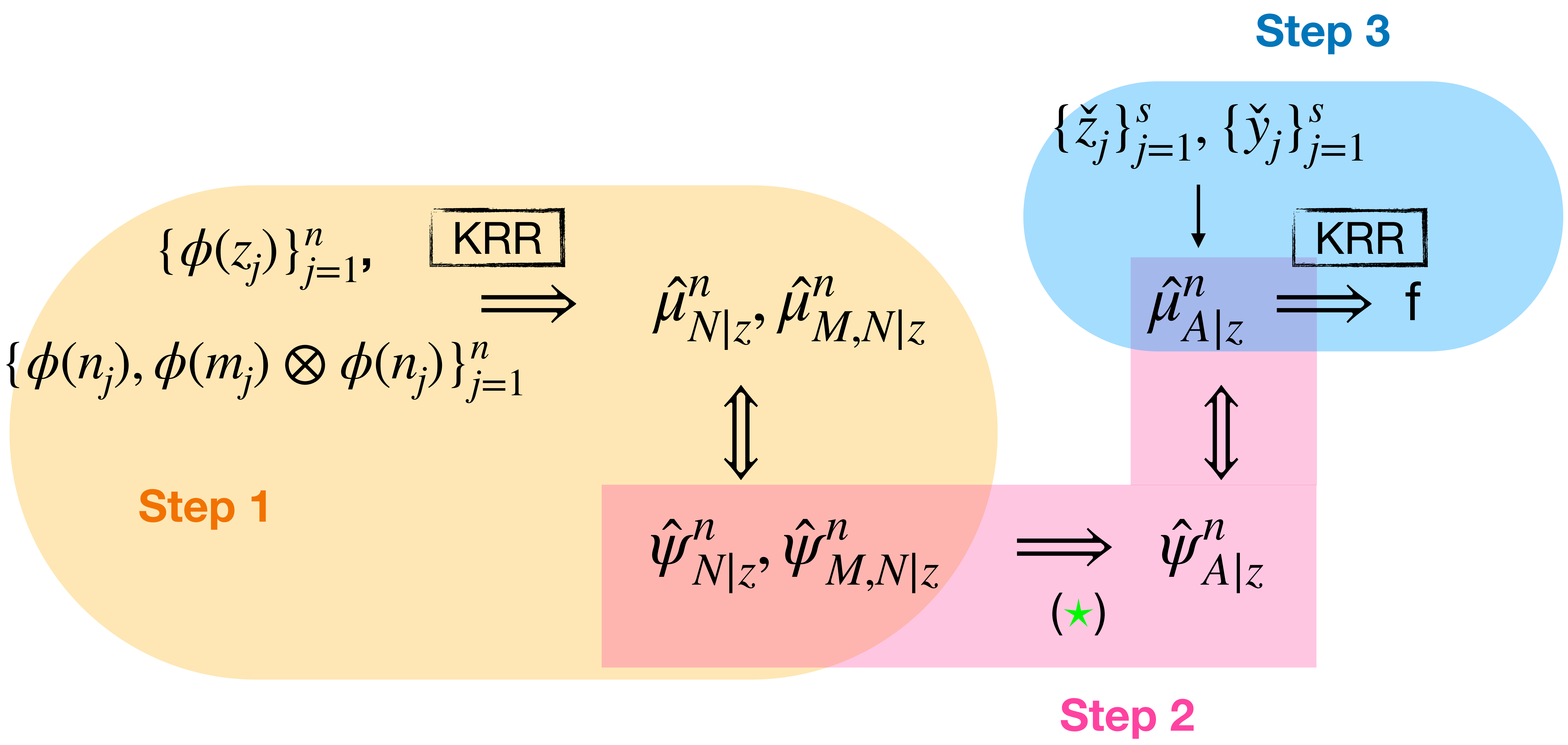
$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}](\alpha)}^{\psi_{A|z}(\alpha)} = \exp\left(\int_0^\alpha i \frac{\overbrace{\mathbb{E}[Me^{i\nu N} \,|\, z]}^{\frac{\partial}{\partial \upsilon}\psi_{M,N|z}(\upsilon,\nu)\big|_{\upsilon=0}}}{\underbrace{\mathbb{E}[e^{i\nu N} \,|\, z]}_{\psi_{N|z}(\nu)}} d\nu \right) \qquad (1)$$

Differentiate wrt α to remove integral.

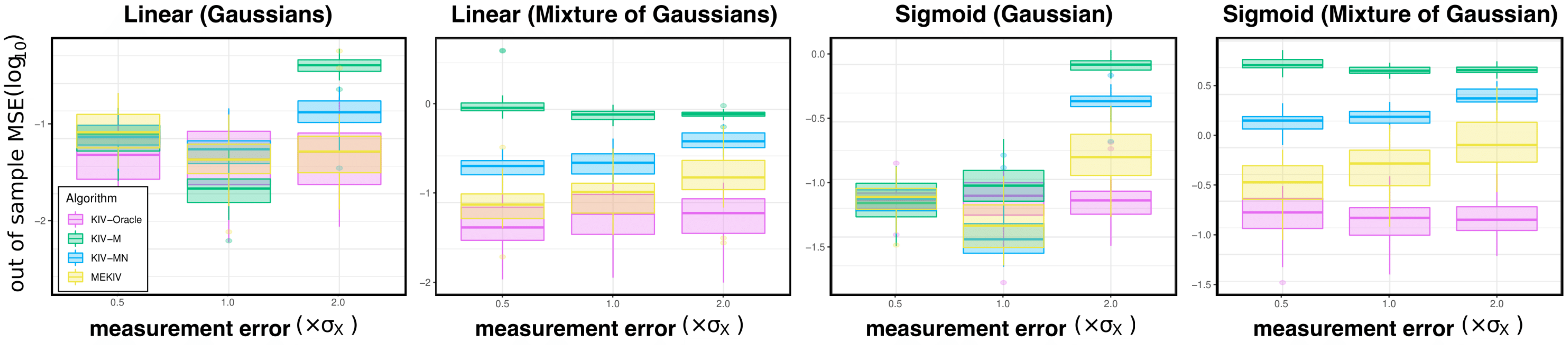
$$\frac{\frac{d}{d\alpha}\hat{\psi}^n_{A|z}(\alpha)}{\hat{\psi}^n_{A|z}(\alpha)} = \frac{\frac{\partial}{\partial \upsilon}\hat{\psi}^n_{M,N|z}(\upsilon, \alpha)\Big|_{\upsilon=0}}{\hat{\psi}^n_{N|z}(\alpha)} \qquad (2)$$

(Replace with sample estimates.)

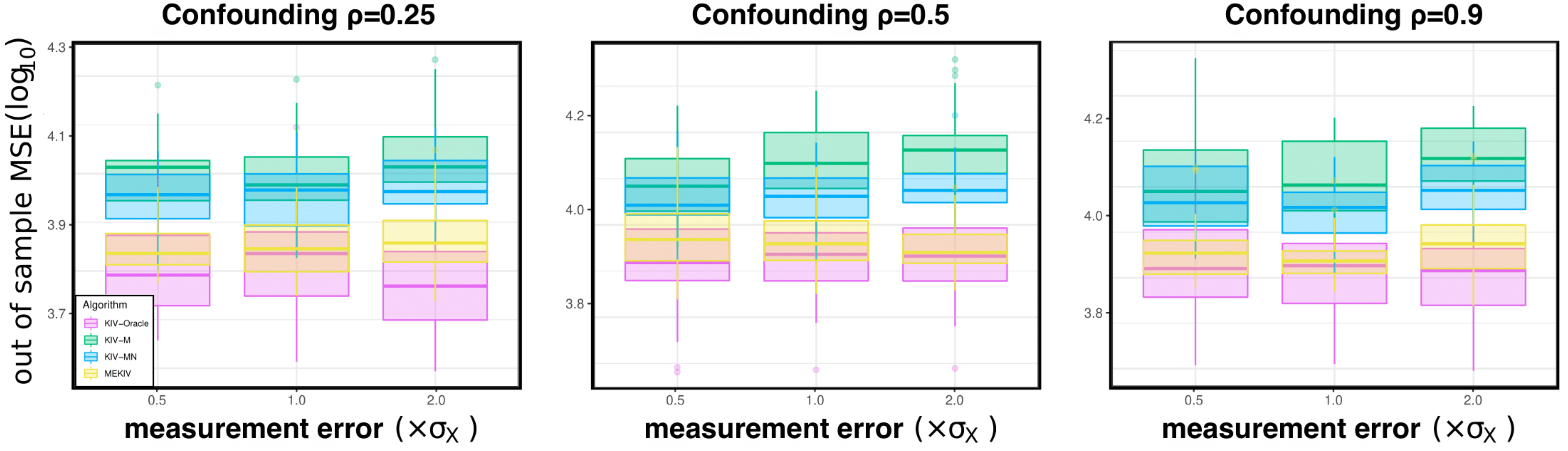
Measurement Error KIV



MEKIV results



Demand Design (Mixture of Gaussians)



Open questions

- Relax the measurement error assumption and IV assumption.
- Extend to sequential settings.

Take home messages

- Nonparametric features can be learned even using corrupted measurements.
- This algorithm relaxes observability from confounding to treatments.
- IV is a restrictive assumption for observational studies, but can work for studies with an experimental component.

Conclusion

- Causality for social sciences from a high-level perspective:
  - Decision making, exploiting observational data, spurious correlation correlation.
- Causal graph can be viewed as a way to encode expert knowledge which can be hard to learn with pure data.
  - Graphs can have a spectrum of restrictiveness.
- Observability assumptions can be relaxed at various degrees.