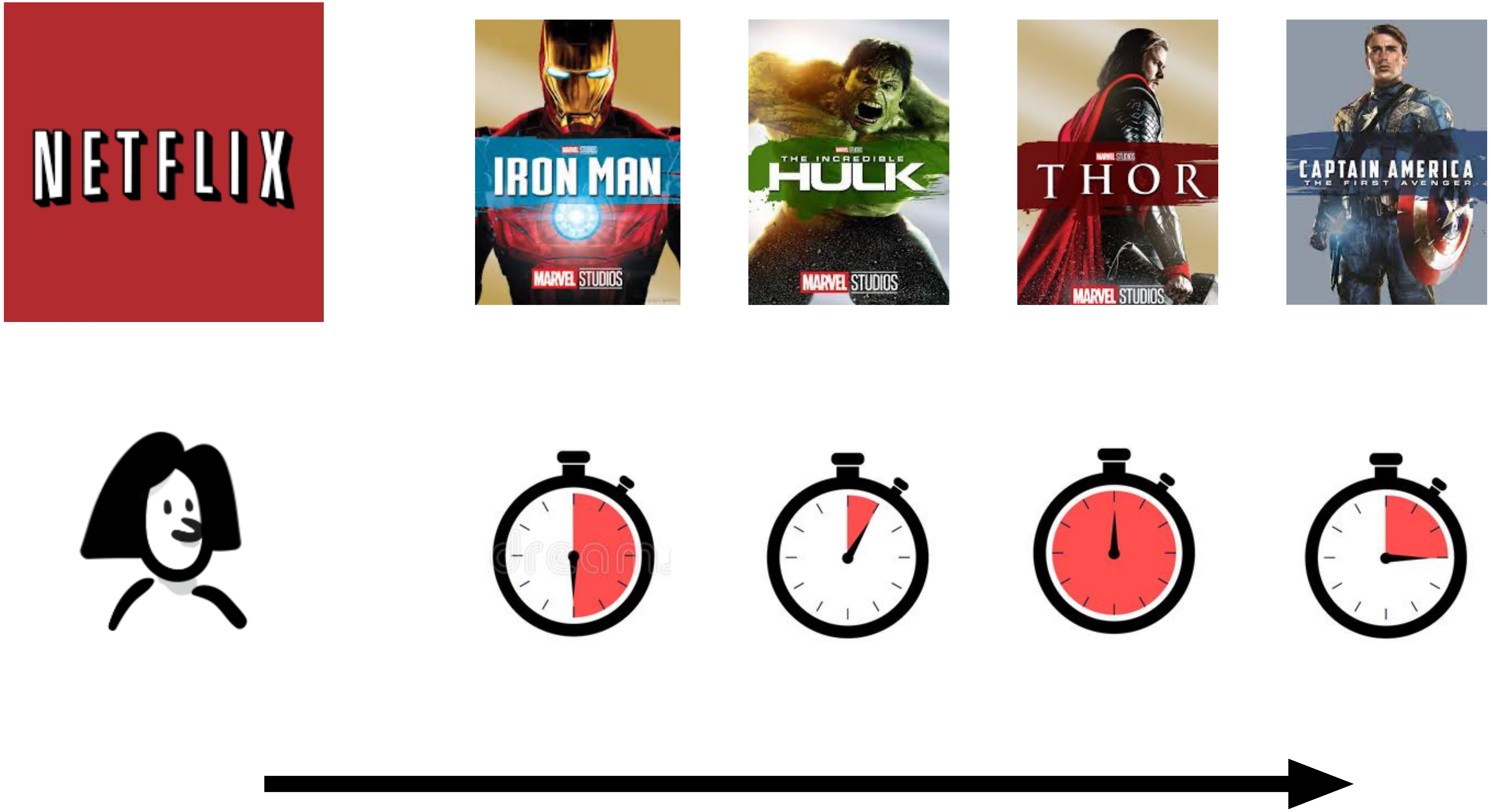
Can we do counterfactual inference in the presence of unobserved confounding with one sample?

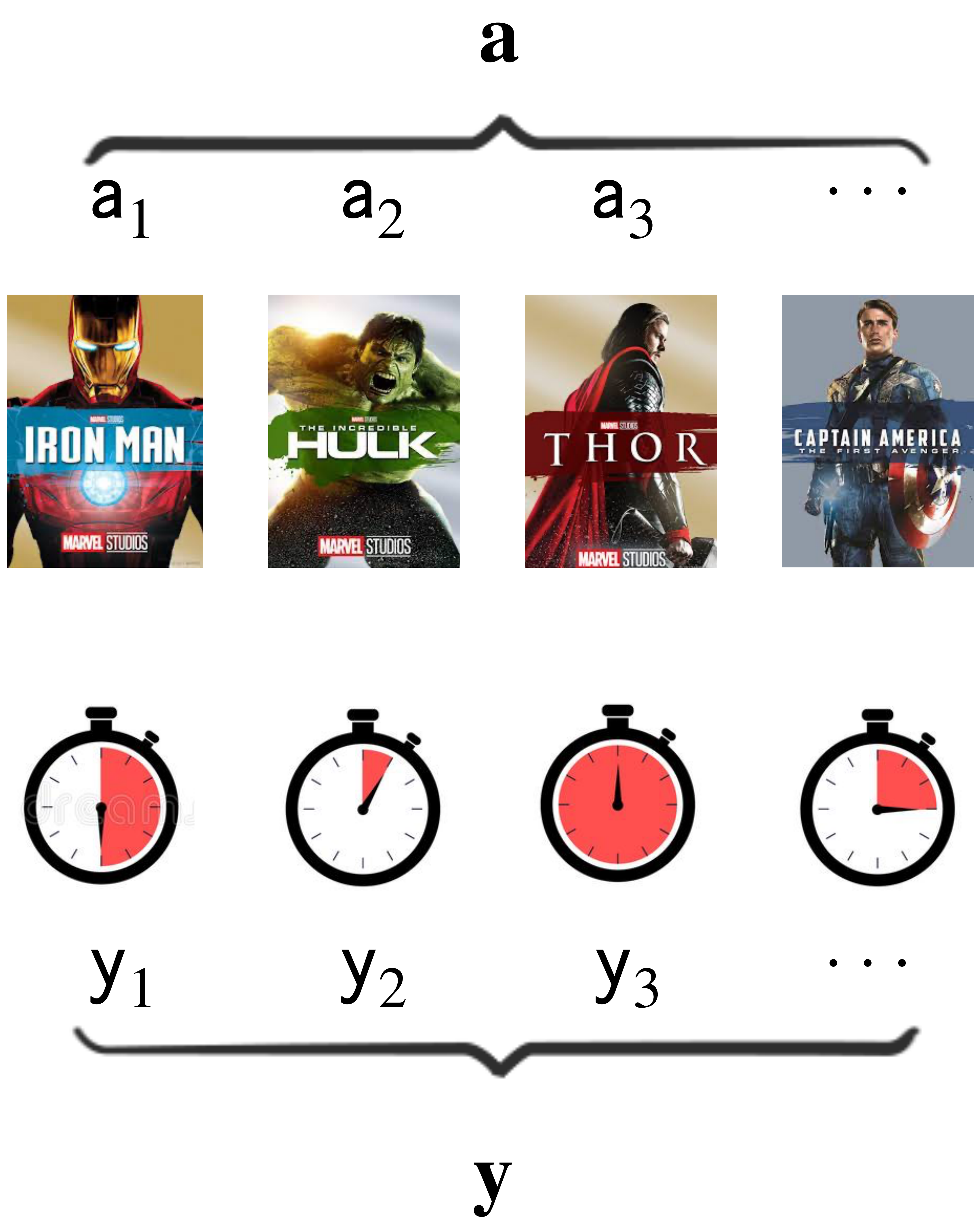
with Raaz Dwivedi, Devavrat Shah, and Greg Wornell

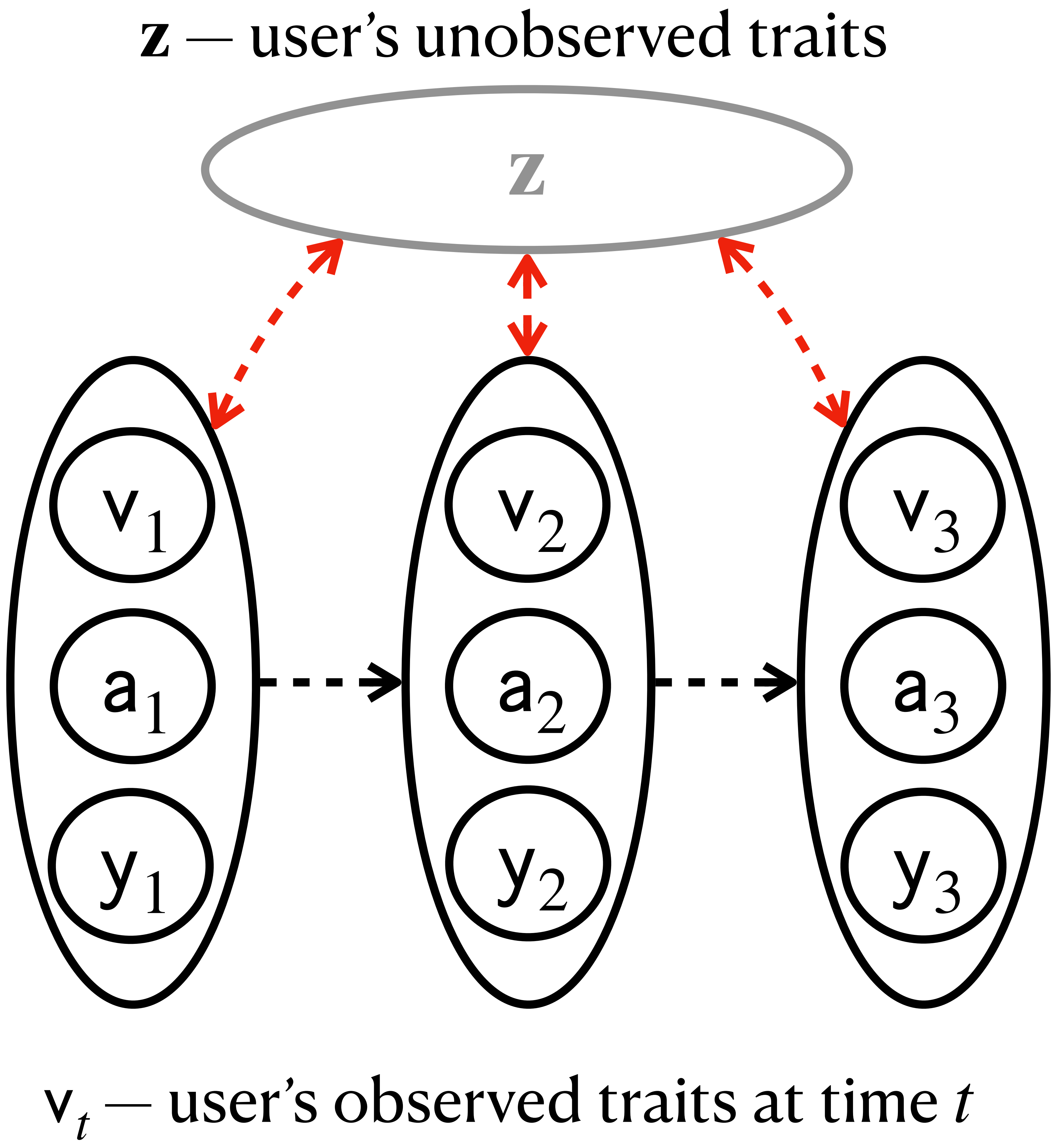
Sequential Recommender System



Sequential Recommender System

A graphical model





Sequential Recommender System

Data

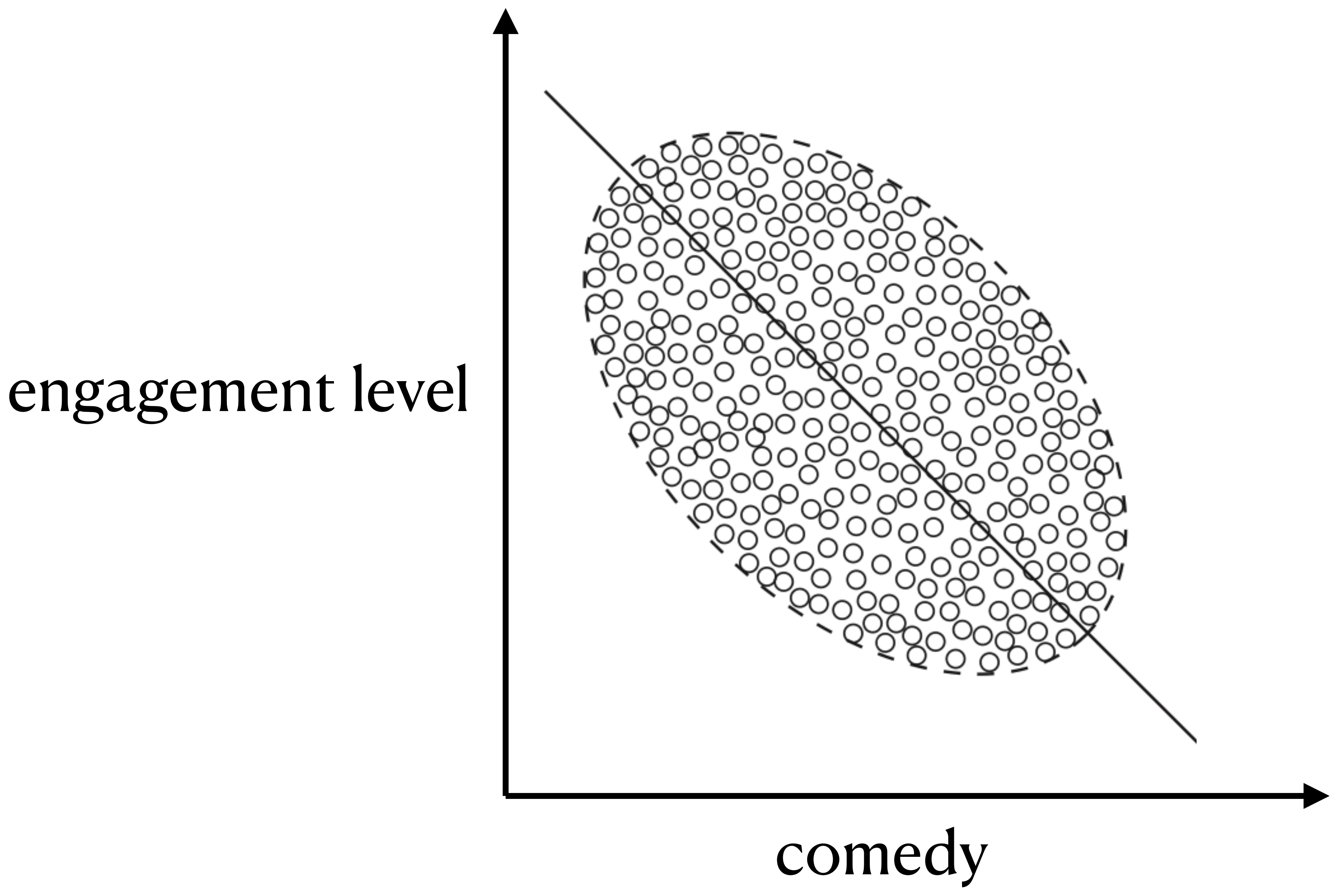
Sequential Recommender System

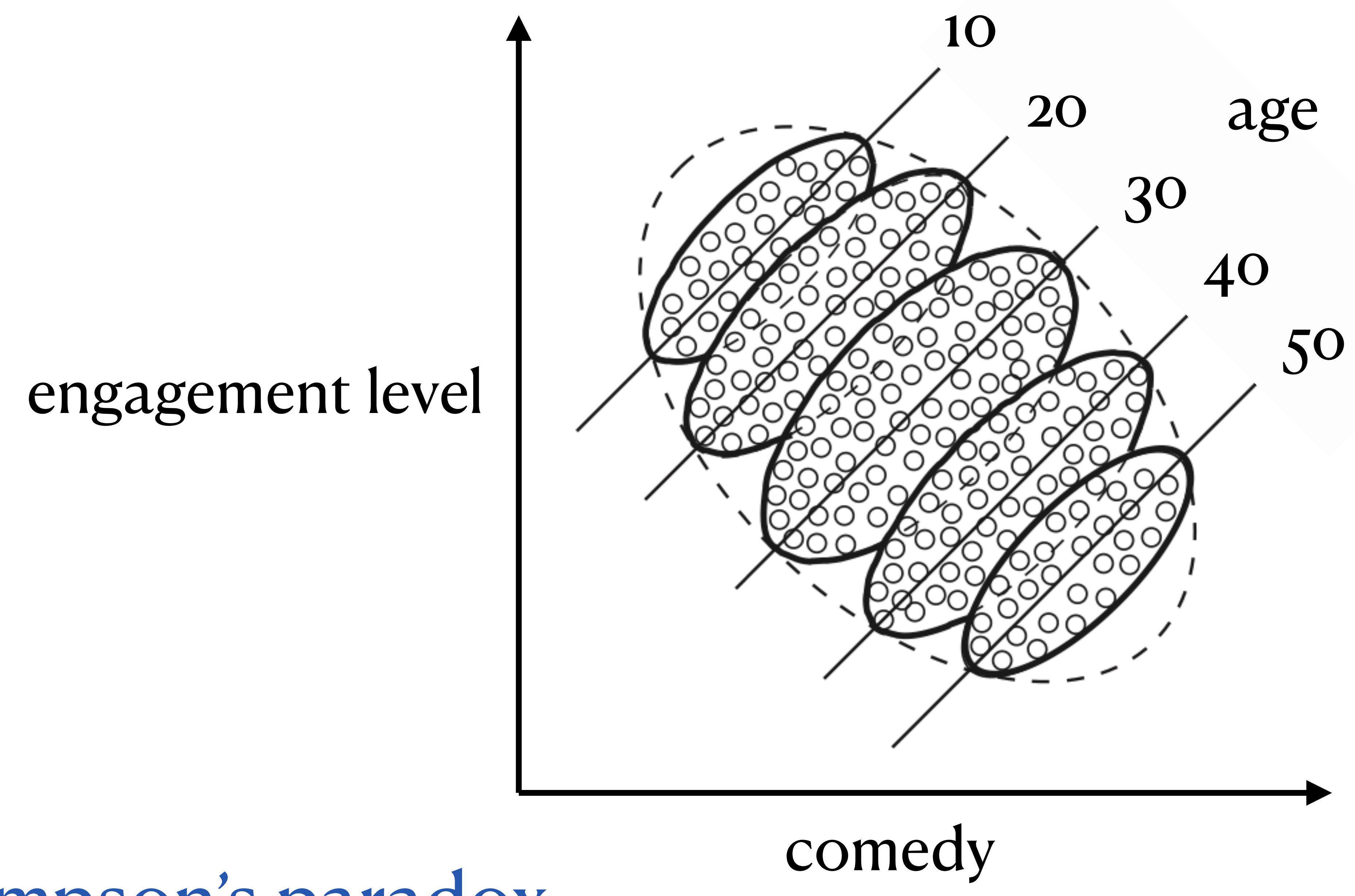
Goal

Sequential Recommender System

Challenges

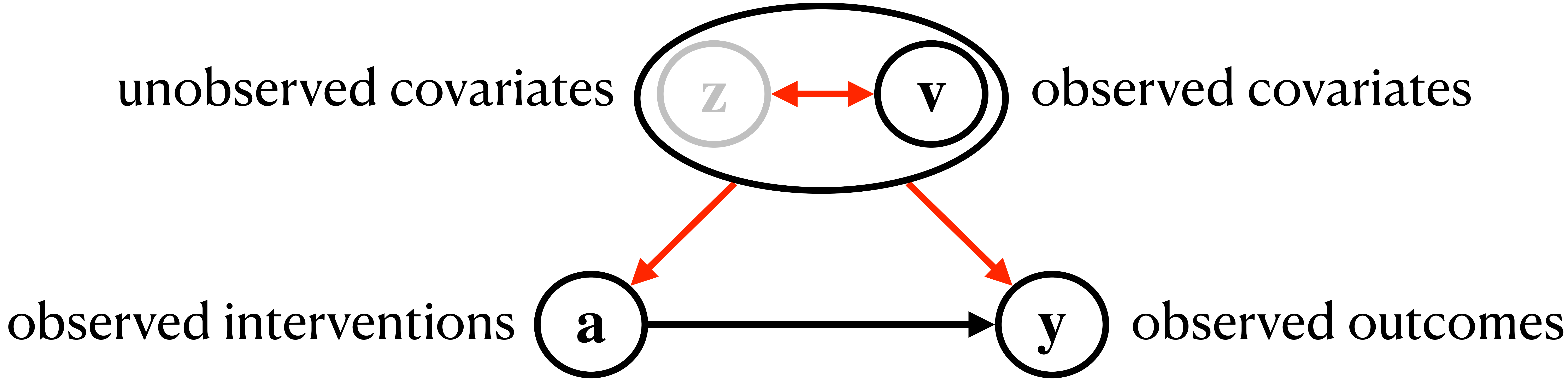
- unobserved factors → spurious associations
- users could be heterogeneous
- each user → a single interaction trajectory





Simpson's paradox

Problem Setup



The micro-level graphical is consistent with this macro-level graphical model

- n heterogenous and independent units
- only one observation per unit - $\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$

Goal

Counterfactual questions for these n units

For every unit $i \in [n]$, what would the potential outcome $\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)})$ be while $\mathbf{z} = \boldsymbol{z}^{(i)}$ and $\mathbf{v} = \boldsymbol{v}^{(i)}$?

Under SUTVA, learning unit-level counterfactual distributions is equivalent to learning unit-level conditional distributions:

$$p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot\,, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)}) \qquad \text{for all } i \in [n]$$

Challenges

1. Unobserved confounding — $\mathbf{z}$ introduces statistical dependence between $\mathbf{a}$ and $\mathbf{y}$
2. Heterogeneity — $(\boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$ could be different for different units

➡ we only observe one realization that is consistent with $p(\mathbf{y} = \cdot \mid \boldsymbol{a}, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$

Is it possible to learn n heterogeneous distributions
with only one sample per distribution?

Our approach

- Model the joint distribution of $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$ as an exponential family

$$p(\boldsymbol{w}; \phi, \Phi) \propto \exp\left(\phi^\top \boldsymbol{w} + \boldsymbol{w}^\top \Phi \boldsymbol{w}\right)$$

- The conditional distribution of $\mathbf{y}$ given $\mathbf{a}, \mathbf{z}, \mathbf{v}$ can be written as

$$p(\boldsymbol{y} \,|\, \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\underbrace{\phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y}}_{\text{different for different units}} + 2\boldsymbol{a}^\top \underbrace{\Phi_{a,y}}\right]\boldsymbol{y} + \boldsymbol{y}^\top \underbrace{\Phi_{y,y}}\boldsymbol{y}\right)$$

n heterogeneous conditional distributions →

n distributions from the same exponential family
but with parameters that vary across units

Our approach

$$p(\boldsymbol{y} \mid \boldsymbol{a}, z = z^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\underbrace{\phi_y + 2z^{(i)\top}\Phi_{z,y}}_{=\theta(z^{(i)})} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top\Phi_{a,y} \right]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y} \right)$$

$$\Theta$$

$$\phi, \Phi \xrightarrow{\text{Reparameterize}} \theta(z), \Theta \xrightarrow{n \text{ units}} \begin{matrix} \theta(z^{(1)}), \Theta \\ \vdots \\ \theta(z^{(n)}), \Theta \end{matrix} \xrightarrow{\text{shared parameters}} \theta(z^{(1)}), \cdots, \theta(z^{(n)}), \Theta$$

1. If $z^{(1)} = \cdots = z^{(n)} \rightarrow$ a single exponential family with n samples
2. If $n = 1 \rightarrow$ a single exponential family with one sample (assume Θ is known)

Inference Tasks

1. Parameters
   A. Unit-level — $\theta^\star(z^{(i)})$ for all $i \in [n]$
   B. Population-level — $\Theta^\star$
2. Expected potential outcomes — $\mathbf{E}\big[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \mid \mathbf{z} = z^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\big]$

Parameter estimation

$\theta(z^{(1)}), \cdots, \theta(z^{(n)}), \Theta \longrightarrow$  $\longrightarrow \hat{\theta}(z^{(1)}), \cdots, \hat{\theta}(z^{(n)}), \hat{\Theta}$

Loss function

For all row j, $\|\Theta_j^\star - \hat{\Theta}_j\|_2 \le \epsilon$ whenever $n \ge O\left(\frac{\log(\mathrm{dim})}{\epsilon^4}\right)$

For all unit i, $\|\theta^\star(z^{(i)}) - \hat{\theta}(z^{(i)})\|_2 \le \max\left\{\epsilon, \mathrm{Comp}\right\}$ whenever $n \ge O\left(\frac{\mathrm{dim}^2\mathrm{Comp}^2}{\epsilon^4}\right)$

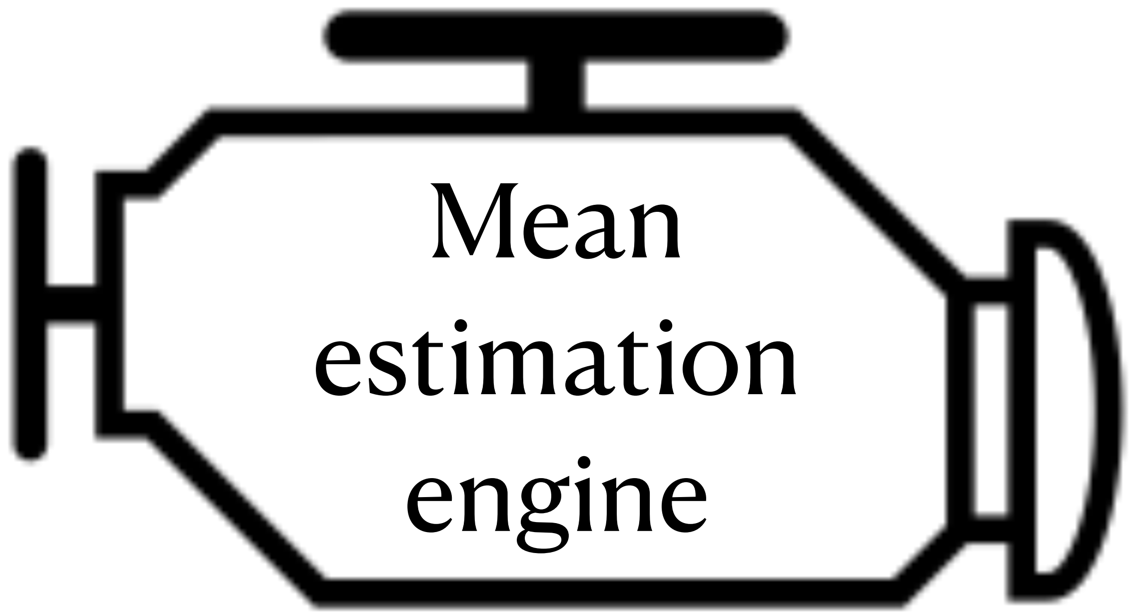
When the true parameters are s-sparse linear combination of k known vectors, $\mathrm{Comp} = O\big(s\log(k \cdot \mathrm{dim})\big)$

Outcome estimation

Expected potential outcomes — $\mu^{(i)} \triangleq \mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \,|\, \mathbf{z} = z^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\right]$

$(\boldsymbol{v}^{(i)}, \tilde{\boldsymbol{a}}^{(i)})$

$\hat{\theta}(z^{(i)}), \hat{\Theta}$



$\hat{\mu}^{(i)}$

When the true parameters are s-sparse linear combination of k known vectors, for any $\{\tilde{\boldsymbol{a}}^{(i)} \in \mathscr{A}\}_{i=1}^{n}$

For all unit i, $MSE(\mu^{(i)}, \hat{\mu}^{(i)}) \leq \dfrac{s \log(k \cdot \text{dim}) + \epsilon^2}{\text{dim}}$ whenever $n \geq O\left(\dfrac{sp^2 \log(k \cdot \text{dim})}{\epsilon^4}\right)$

Loss function

Condition on z

- Recall the joint distribution of $\mathbf{w} = (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

$$p(\boldsymbol{w}; \phi, \Phi) \propto \exp\left(\phi^\top \boldsymbol{w} + \boldsymbol{w}^\top \Phi \boldsymbol{w}\right)$$

- Letting $\mathbf{x} \triangleq (\mathbf{v}, \mathbf{a}, \mathbf{y})$, the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ can be written as

$$p(\boldsymbol{x} \,|\, z; \theta(z), \Theta) \propto \exp\left(\left[\theta(z)\right]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x}\right)$$

$$p(\boldsymbol{y} \,|\, \boldsymbol{a}, z = z^{(\boldsymbol{i})}, \boldsymbol{v} = \boldsymbol{v}^{(\boldsymbol{i})}) \propto \exp\left(\left[\phi_y + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top \Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top \Phi_{y,y} \boldsymbol{y}\right)$$

Loss function

Assumptions

$$p(\boldsymbol{x} \,|\, z; \theta(z), \Theta) \propto \exp\left(\left[\theta(z)\right]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x} \right)$$

(A) Every element of $\theta^\star(z^{(1)}), \cdots, \theta^\star(z^{(n)})$, and $\Theta^\star$ is bounded

(B) Every row of $\Theta^\star$ is sparse

(C) Every diagonal entry of $\Theta^\star$ is zero

- $\Lambda_\theta \triangleq \{\theta : \theta$ is consistent with (A) + low complexity$\}$
- $\Lambda_\Theta \triangleq \{\Theta : \Theta$ is consistent with (A), (B), and (C)$\}$

Loss function

$$p(\boldsymbol{x} \mid z; \theta(z), \Theta) \propto \exp\Big(\big[\theta(z)\big]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x} \Big)$$

- inspired by the conditional distribution of x_t given $\mathbf{x}_{-t}$ and $\mathbf{z}$ —

$$p(x_t \mid \boldsymbol{x}_{-t}, z; \theta_t(z), \Theta_t) \propto \exp\Big(\big[\theta_t(z) + 2\Theta_t^\top \boldsymbol{x}\big] x_t \Big)$$

- maps the parameters $\theta^{(1)}, \cdots, \theta^{(n)},$ and Θ to $\mathscr{L}$

$$\mathscr{L}(\underline{\Theta}) = \frac{1}{n} \sum_{t \in [\dim]} \sum_{i \in [n]} \exp\Big(-\big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

Loss function

Estimate

$$\mathscr{L}(\underline{\Theta}) = \frac{1}{n} \sum_{t \in [\dim]} \sum_{i \in [n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

$$\widehat{\underline{\Theta}} = \arg\min_{\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta} \mathscr{L}(\underline{\Theta})$$

- convex optimization problem
- strictly proper loss function — $\underline{\Theta}^\star$ uniquely maximizes $\mathbb{E}[\mathscr{L}(\,\cdot\,)]$

Loss function

Decomposition

$$\mathscr{L}(\underline{\Theta}) = \frac{1}{n} \sum_{t \in [\dim]} \sum_{i \in [n]} \exp\left(- \left[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\right] x_t^{(i)} \right) \text{ where } \underline{\Theta} \triangleq \left[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\right]$$

$$\widehat{\underline{\Theta}} = \arg\min_{\underline{\Theta} \in \Lambda_\theta^n \times \Lambda_\Theta} \mathscr{L}(\underline{\Theta})$$

$$\min_{a,b} f(a, b) = \min_a \min_b f(a, b)$$

1. minimize w.r.t Θ
2. minimize w.r.t $\theta^{(1)}, \cdots, \theta^{(z)}$

Loss function

Learning population-level parameter

$$\mathscr{L}(\underline{\Theta}) = \frac{1}{n} \sum_{t \in [\dim]} \sum_{i \in [n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

$\Lambda_\Theta \rightarrow$ (A) bounded elements, (B) sparse rows (C) zero diagonals

Λ_Θ places independent constraints on the rows of Θ

p independent convex optimization problems

$$\mathscr{L}_t(\underline{\Theta}_t) = \frac{1}{n} \sum_{i \in [n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ for all } t \in [\dim] \longrightarrow \widehat{\Theta}_t$$

Loss function

Learning unit-level parameter

$$\mathscr{L}(\underline{\Theta}) = \frac{1}{n} \sum_{t\in[\dim]} \sum_{i\in[n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

$\Lambda_\theta \rightarrow$ (A) bounded elements, (B) low complexity

$\theta^{(1)}, \cdots, \theta^{(n)} \in \Lambda_\theta^n$ places independent constraints on units, i.e., $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$

n independent convex optimization problems

$$\mathscr{L}^{(i)}(\theta^{(i)}) = \sum_{t\in[\dim]} \exp\Big(- \big[\theta_t^{(i)} + 2\widehat{\Theta}_t^{\top} \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ for all } t \in [\dim] \longrightarrow \widehat{\theta}^{(i)}$$

Can we do counterfactual inference in the presence of unobserved confounding with one sample?

For every user, Netflix can estimate the expected potential outcomes with MSE scaling as 1/dimension

Exponential family to the rescue!

Social network setting