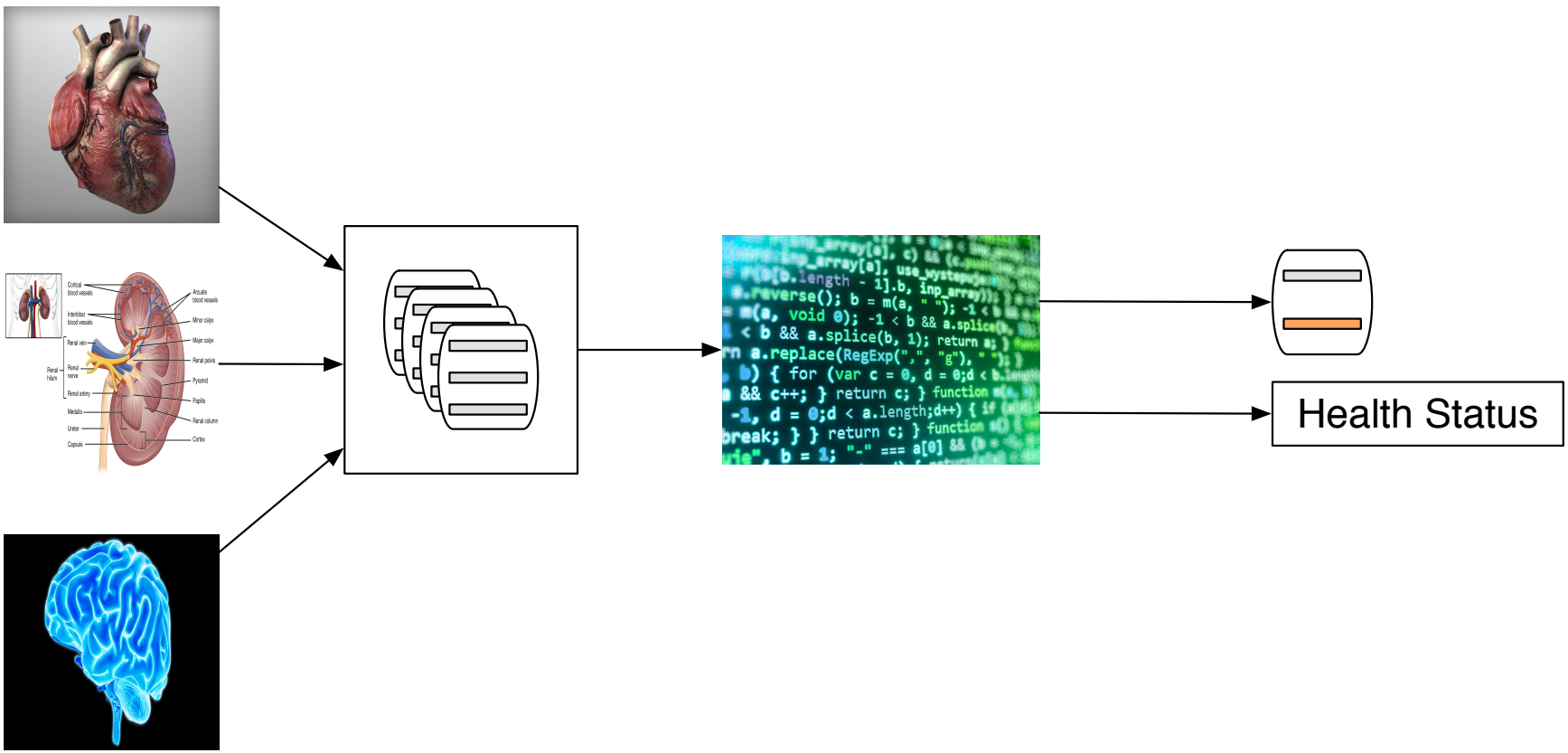
Machine Learning
Causal Inference

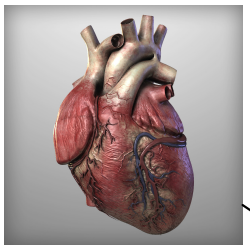
Rajesh Ranganath

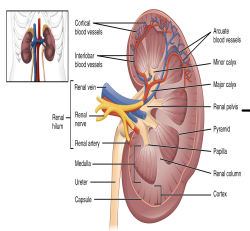
- We have features $\mathbf{t}$
- We have an outcome y
- We build a model $p(y \mid \mathbf{t})$
- Does this model tell us if we change $\mathbf{t}$, we see a change in y?

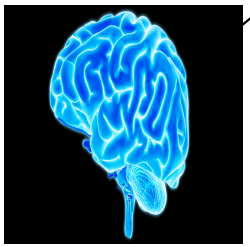
Predictions for Healthcare

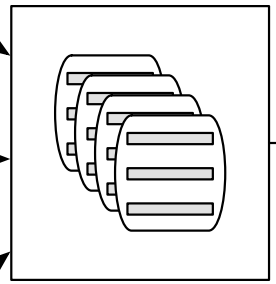


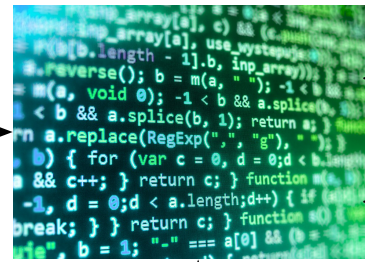
Interventions in Healthcare

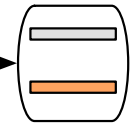








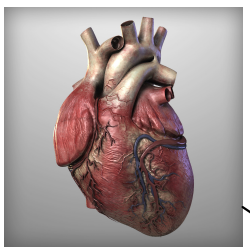


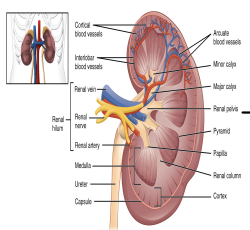


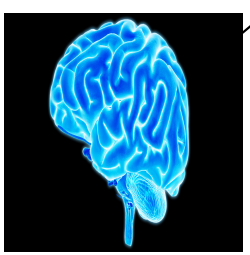
Health Status

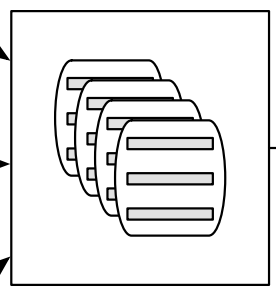
Medication A?

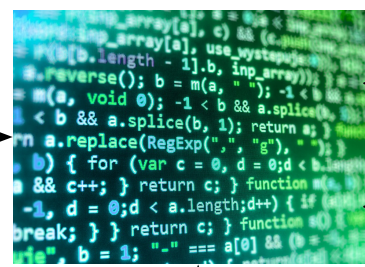
Interventions in Healthcare

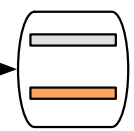








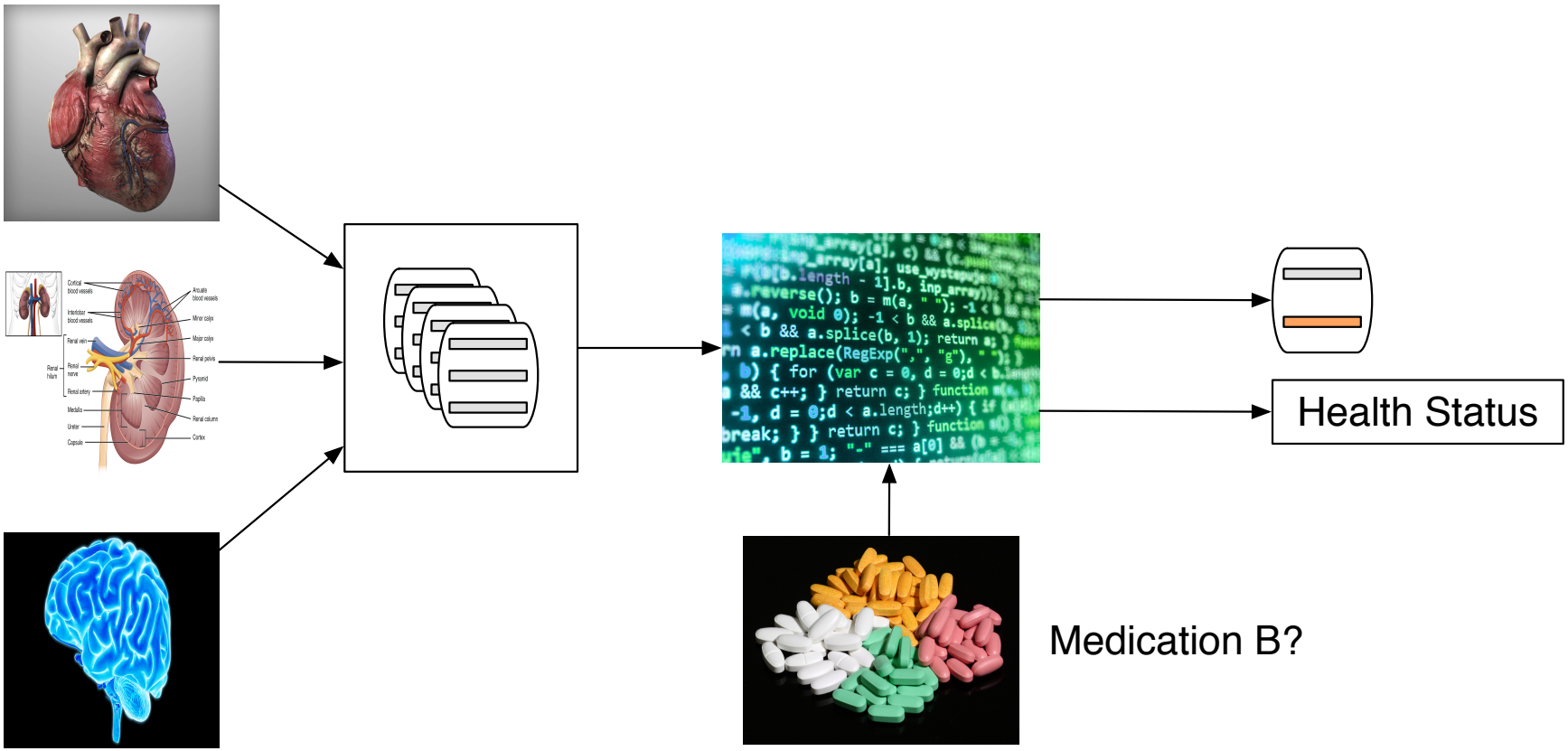




Health Status

Medication B?

Medication A or Medication B?



Question: Should we give medication A or medication B?

Causal Inference

Causal inference seeks to estimate the effect of an intervention

- Which statin to give to a patient with hyperlipidemia?
- Which medication to give to a depressed person?
- Which patients to give hospice care to?

Solutions?

- Find important features in a regression?
  Why should this work?

Solutions?

- Find important features in a regression?
  Why should this work?

- What about other interventions after feature collection?

Need a mathematical definition of causal inference

Potential Outcomes

Define

- Response if treatment 0 given to person i: $Y_{0,i}$
- Response if treatment 1 given to person i: $Y_{1,i}$

Example

- $Y_{0,i}$ PHQ-9 score after 2 months after starting medication for depression
- $Y_{1,i}$ PHQ-9 score after 2 months after starting congnitive behavioral therapy for depression

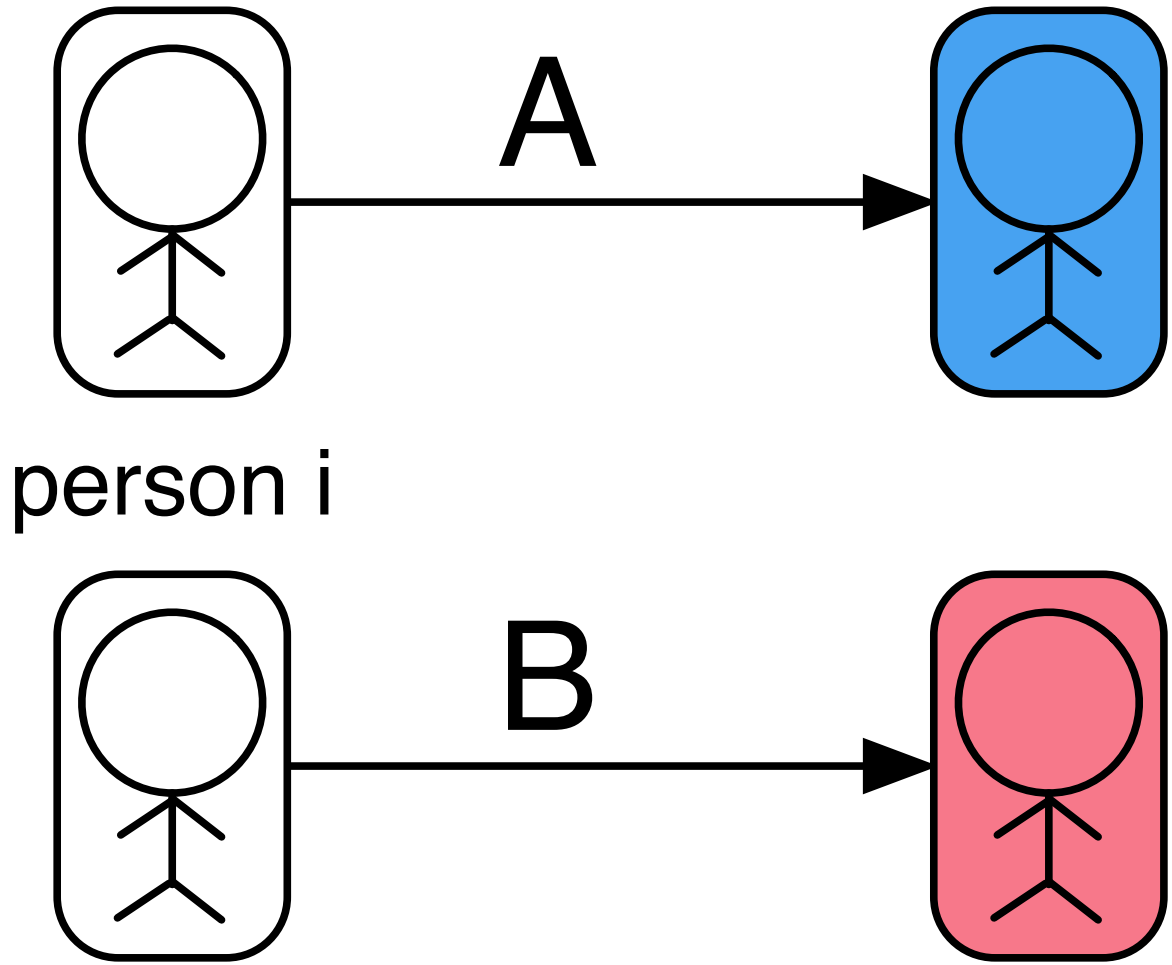
Potential Outcomes

Causal effects are function of the potential outcomes

- Is A better than B for person i: $Y_{0,i} > Y_{1,i}$
- What's the difference between A and B? $Y_{0,i} - Y_{1,i}$

Solution

- Observe $Y_{0,i}$, $Y_{1,i}$
- People identified by covariates $\mathbf{x}_i$
- Build a model for $Y_{0,i}$, $Y_{1,i}$ and $\mathbf{x}_i$
- Make predictions!



How do we observe the effect of both on the same person?

Move to population level questions

Average Treatment Effect

$$ATE = \mathbb{E}_i[y_{1,i}] - \mathbb{E}_i[y_{0,i}]$$

Individualized Treatment Effect

$$ITE(\mathbf{x}_i) = \mathbb{E}_i[y_{1,i} \mid \mathbf{x}_i] - \mathbb{E}_i[y_{0,i} \mid \mathbf{x}_i]$$

Equivalence by averaging over $p(\mathbf{x}_i)$

$$ATE = \mathbb{E}_{p(\mathbf{x}_i)}[ITE(\mathbf{x}_i)]$$

Use different individuals to get population level estimates

Terminology

- Treatment: t
- Controls: Received $t = 0$
- Treated: Received $t = 1$
- Counterfactuals: The unobserved outcome $y_{1-t_i,i}$

What we observe?

- Treatment: t_i
- Features: $\mathbf{x}_i$
- Outcome: $y_{t,i}$

Compute ATE by taking the i that were treated and untreated?

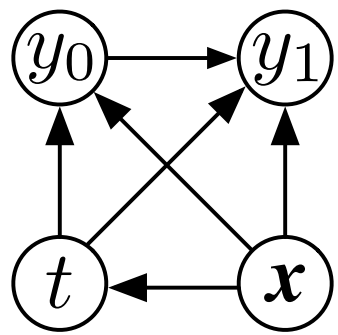
Problem?

What we observe?

- Treatment: t_i
- Features: $\mathbf{x}_i$
- Outcome: $y_{t,i}$

Compute ATE by taking the i that were treated and untreated?

Problem?

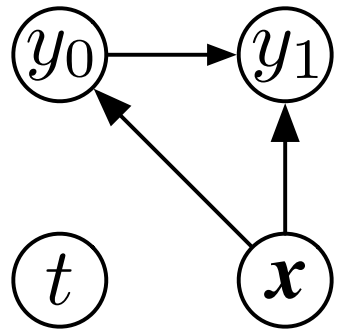


$$\mathbb{E}[y_1 \mid T = 1] - \mathbb{E}[y_0 \mid T = 0] \neq \mathbb{E}[y_1] - \mathbb{E}[y_0]$$

Treated people could be sicker, older, different

Randomization

Assign treatments independent of the rest of the variables



Randomization breaks dependence

$$\mathbb{E}[y_1 \mid T = 1] - \mathbb{E}[y_0 \mid T = 0] = \mathbb{E}[y_1] - \mathbb{E}[y_0]$$

Can estimate ATE from observed data (also ITE)

Randomization

$$\mathbb{E}[y_1 \mid T = 1] - \mathbb{E}[y_0 \mid T = 0] = \mathbb{E}[y_1] - \mathbb{E}[y_0]$$

- Does smoking cause cancer?
- What are the effects of multiple drugs for coronary artery disease?
- What's the effect of a job treatment program?

Many randomized trials are hard or unethical to run

Observational Causal Inference

Goal: To estimate causal effects from non randomized data

$$\mathbb{E}[y_1] - \mathbb{E}[y_0]$$

Problem: Outcomes may depend on the treatment

$$y_1, y_0 \not\perp\!\!\!\perp t$$

Thoughts on what to do?

Observational Causal Inference

Goal: To estimate causal effects from non randomized data

$$\mathbb{E}[y_1] - \mathbb{E}[y_0]$$

Problem: Outcomes may depend on the treatment

$$y_1, y_0 \not\perp\!\!\!\perp t$$

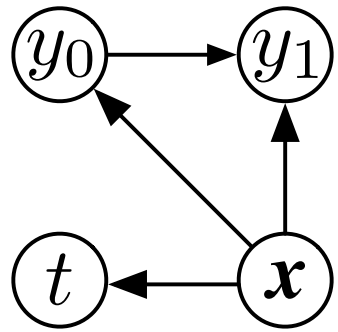
Thoughts on what to do?

Impossible without assumptions. Ideas for assumptions?

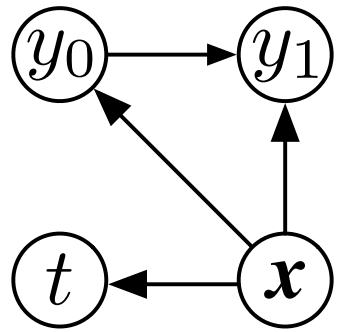
Strong Ignorability

Assume

$$y_1, y_0 \perp\!\!\!\perp t \,|\, \mathbf{x}$$



Strong Ignorability



$$\begin{aligned}\mathbb{E}[y_1] - \mathbb{E}[y_0] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1 \mid \mathbf{x}]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0 \mid \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1 \mid \mathbf{x}, t=1]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0 \mid \mathbf{x}, t=0]]\end{aligned}$$

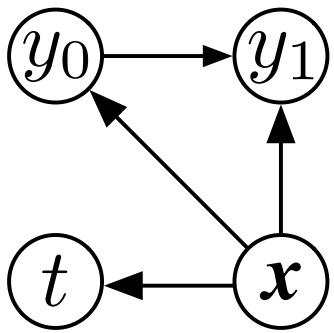
Can estimate from observed data!

For fixed covariates $\mathbf{x}$

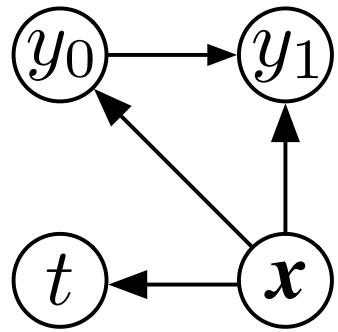
- $\mathbb{E}[y_1 \mid \mathbf{x}, t=1]$ is the average treated outcome on treated
- $\mathbb{E}[y_0 \mid \mathbf{x}, t=0]$ is the average untreated outcome on untreated

Intuition

Strong Ignorability



Conditional Randomization



Strong ignorability is like assuming the “world” consists of conditionally randomized experiments

Can we check strong ignorability?

Another Issue?

What happens when

$$\text{Supp}(p(\mathbf{x}|t=1)) \cap \text{Supp}(p(\mathbf{x}|t=0)) = \emptyset?$$

Another Issue?

What happens when

$$\text{Supp}(p(\mathbf{x}|t=1)) \cap \text{Supp}(p(\mathbf{x}|t=0)) = \emptyset?$$

Consider ATE

$$\mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1|\mathbf{x}, t=1]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0|\mathbf{x}, t=0]]$$

Another Issue?

What happens when

$$\text{Supp}(p(\mathbf{x}|t=1)) \cap \text{Supp}(p(\mathbf{x}|t=0)) = \emptyset?$$

Consider ATE

$$\mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1|\mathbf{x}, t=1]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0|\mathbf{x}, t=0]]$$

Cannot distinguish between $\mathbf{x}$ and t since t determines part of $\mathbf{x}$

Need full support

Computational Tool: Matching

Match treated to similar controls. Units are actually the "same"

$$ATE = \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1 \mid \mathbf{x}, t = 1]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0 \mid \mathbf{x}, t = 0]]$$

- For each unit i
- Find k similar units in the other class using distance function d
- $d(\mathbf{x}_i, \mathbf{x}_j)$ where $t_j = 1 - t_i$

Nice because nonparametric. Drawbacks?

Computational Tool: Regression

Goal: Directly estimate $\mathbb{E}[y \mid \mathbf{x}, t]$

Direct averages may require too many samples. Use regression

$$
\begin{aligned}
&argmin_f \mathbb{E}[(y - f(\mathbf{x}, t))^2] \\
&= argmin_f \int p(\mathbf{x}, t) \int p(y \mid \mathbf{x}, t)(y - f(\mathbf{x}, t))^2 dy dt d\mathbf{x} \\
&= argmin_f \int p(\mathbf{x}, t) \int p(y \mid \mathbf{x}, t)(y^2 - 2yf(\mathbf{x}, t) + f(\mathbf{x}, t)^2) dy dt d\mathbf{x} \\
&= argmin_f \int p(\mathbf{x}, t) \left(\mathbb{E}[y \mid \mathbf{x}, t]^2 + \mathbb{V}\text{ar}(y \mid \mathbf{x}, t) - 2\mathbb{E}[y \mid \mathbf{x}, t] f(\mathbf{x}, t) + f(\mathbf{x}, t)^2 \right) d \\
&= argmin_f \int p(\mathbf{x}, t) \left((\mathbb{E}(y \mid \mathbf{x}, t) - f(\mathbf{x}, t))^2 + \mathbb{V}\text{ar}(y \mid \mathbf{x}, t) \right) dt d\mathbf{x} \\
&= \mathbb{E}[y \mid \mathbf{x}, t]
\end{aligned}
$$

Computational Tool: Regression

Goal: Directly estimate $\mathbb{E}[y \mid \mathbf{x}, t]$

Direct averages may require too many samples. Use regression

$$argmin_f \mathbb{E}[(y - f(\mathbf{x}, t))^2] = \mathbb{E}[y \mid \mathbf{x}, t]$$

Make f parametric $f_{\boldsymbol{\theta}}$ and solve

$$argmin_{\boldsymbol{\theta}} \mathbb{E}[(y - f_{\boldsymbol{\theta}}(\mathbf{x}, t))^2]$$

Substitute back into ATE

$$\begin{aligned} ATE &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_1 \mid \mathbf{x}, t = 1]] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}[y_0 \mid \mathbf{x}, t = 0]] \\ &= \mathbb{E}_{\mathbf{x}}[f_{\boldsymbol{\theta}}(\mathbf{x}, 1)] - \mathbb{E}_{\mathbf{x}}[f_{\boldsymbol{\theta}}(\mathbf{x}, 0)] \end{aligned}$$

Why not two separate regressions?

When is

$$\mathbb{E}[y \mid \mathbf{x}, t]$$

valid?

When is

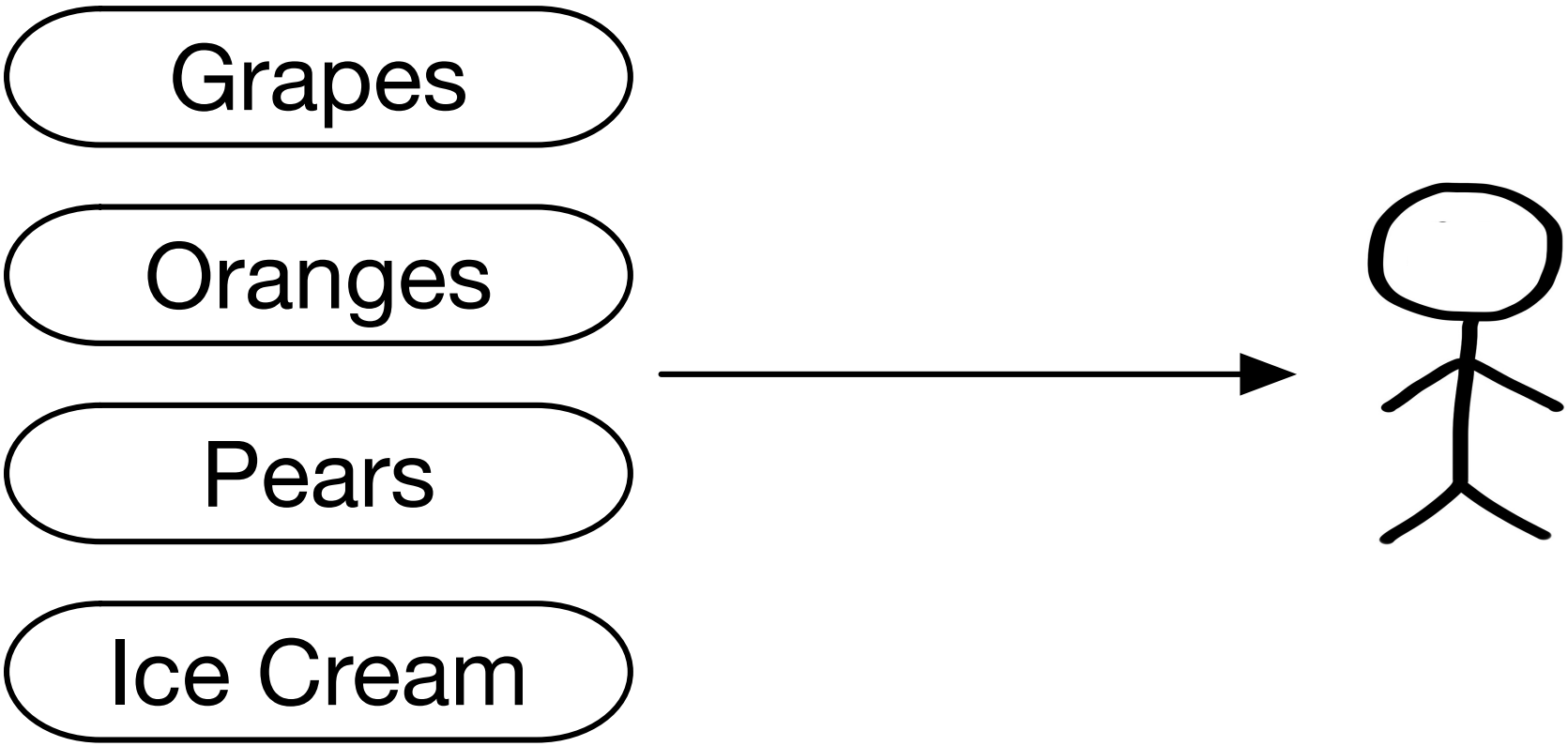
$$\mathbb{E}[y \mid \mathbf{x}, t]$$

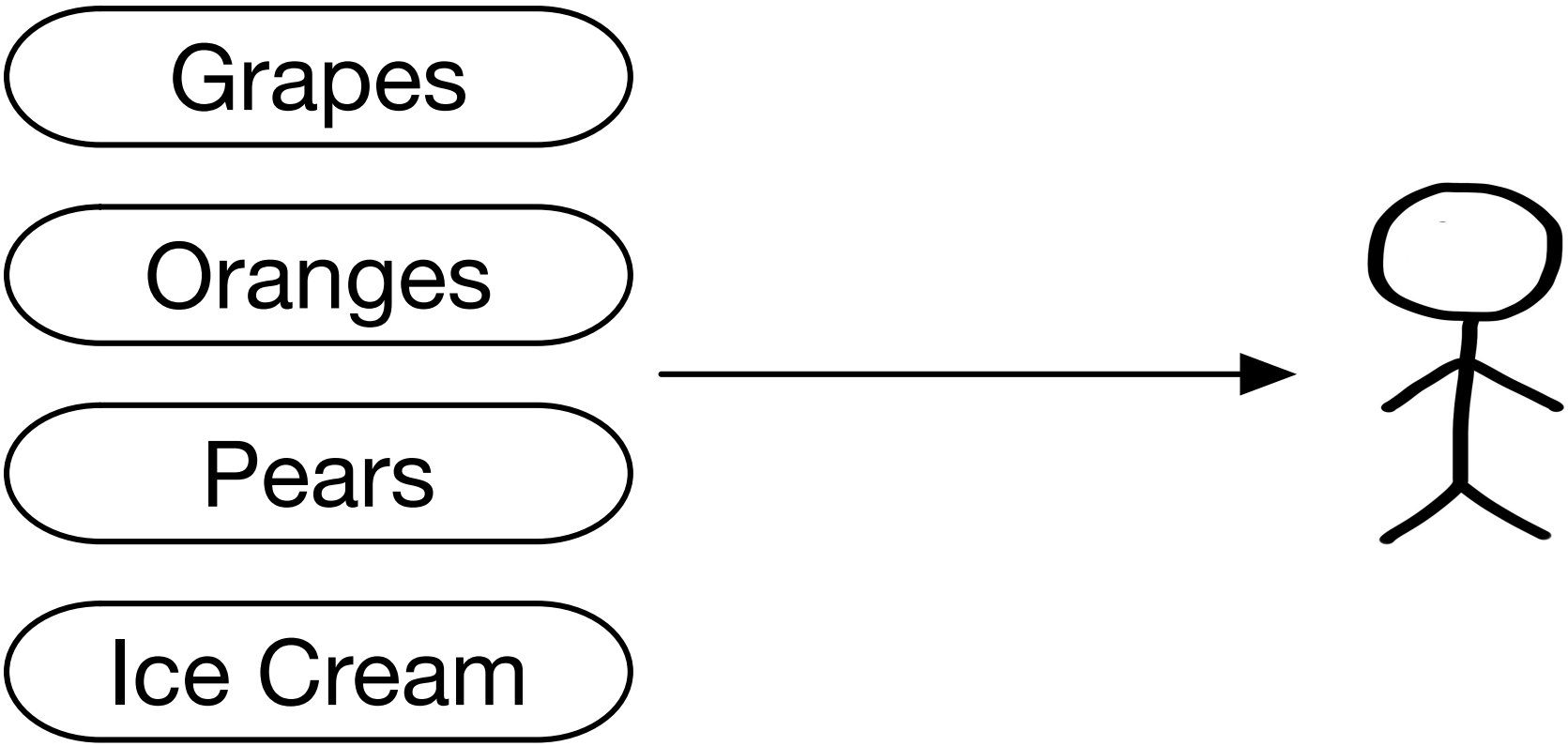
valid?

When $p(\mathbf{x}, t) > 0$

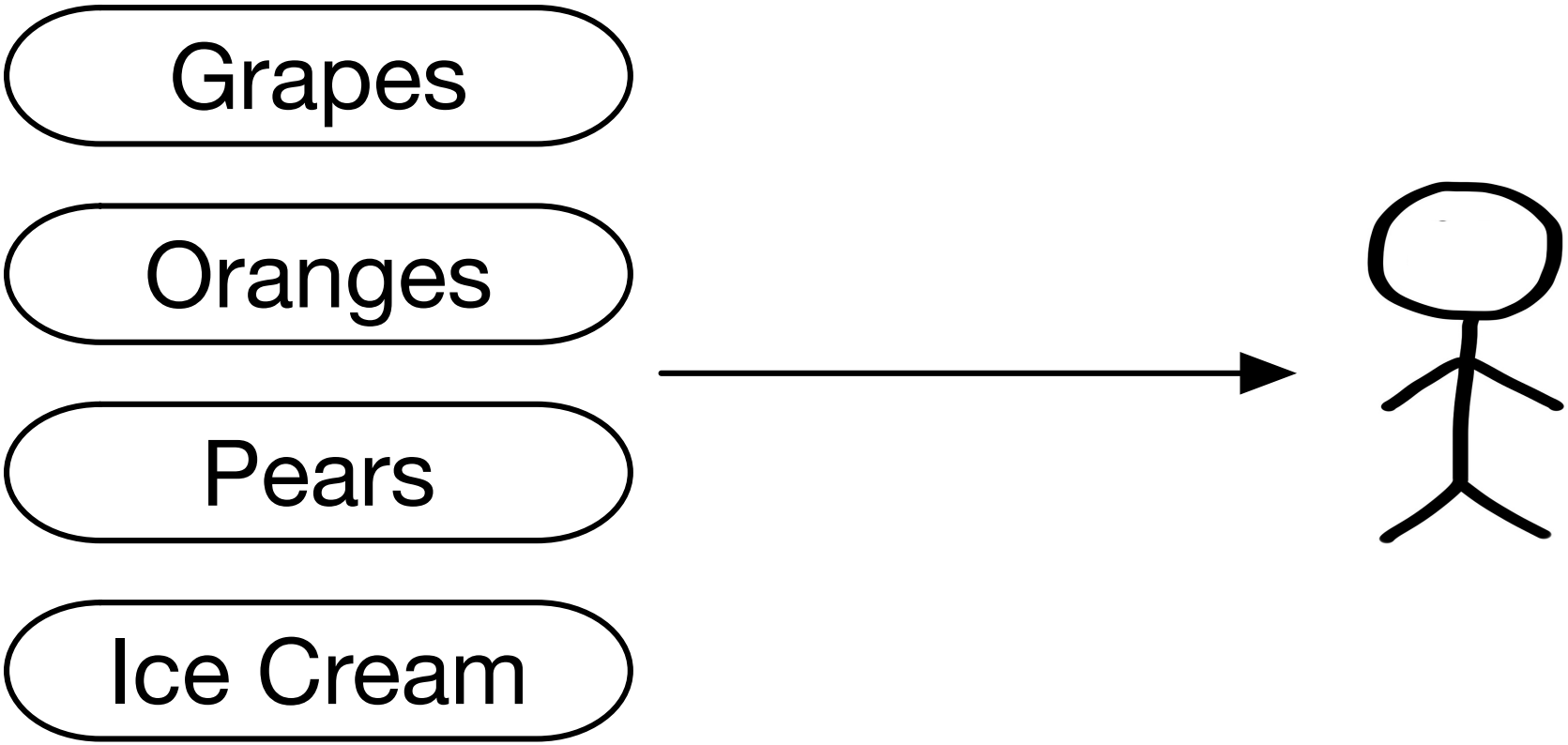
Why does this matter?

Let's start from the beginning again


Grapes
Oranges
Pears
Ice Cream

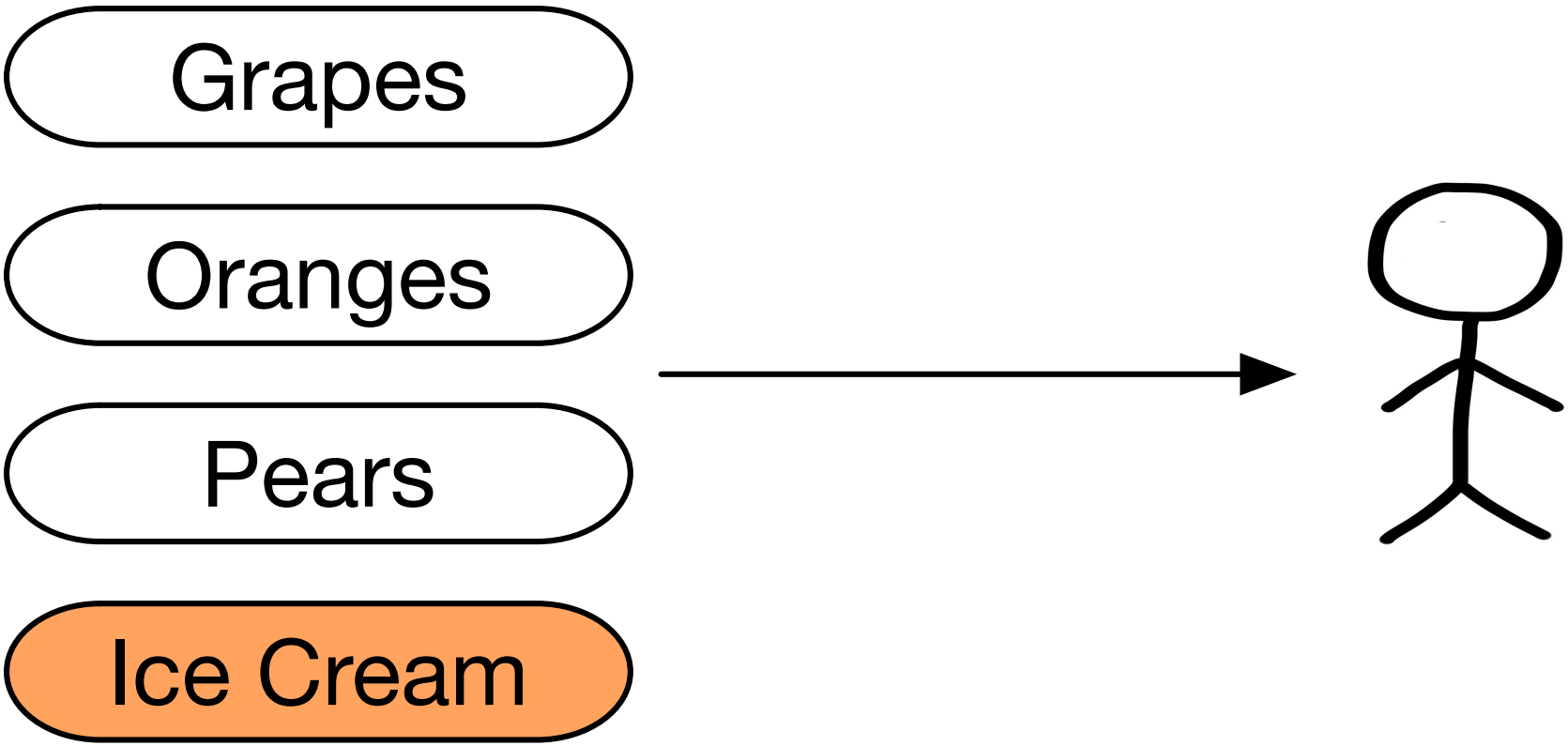


Does ice cream at age 3 make you live longer?

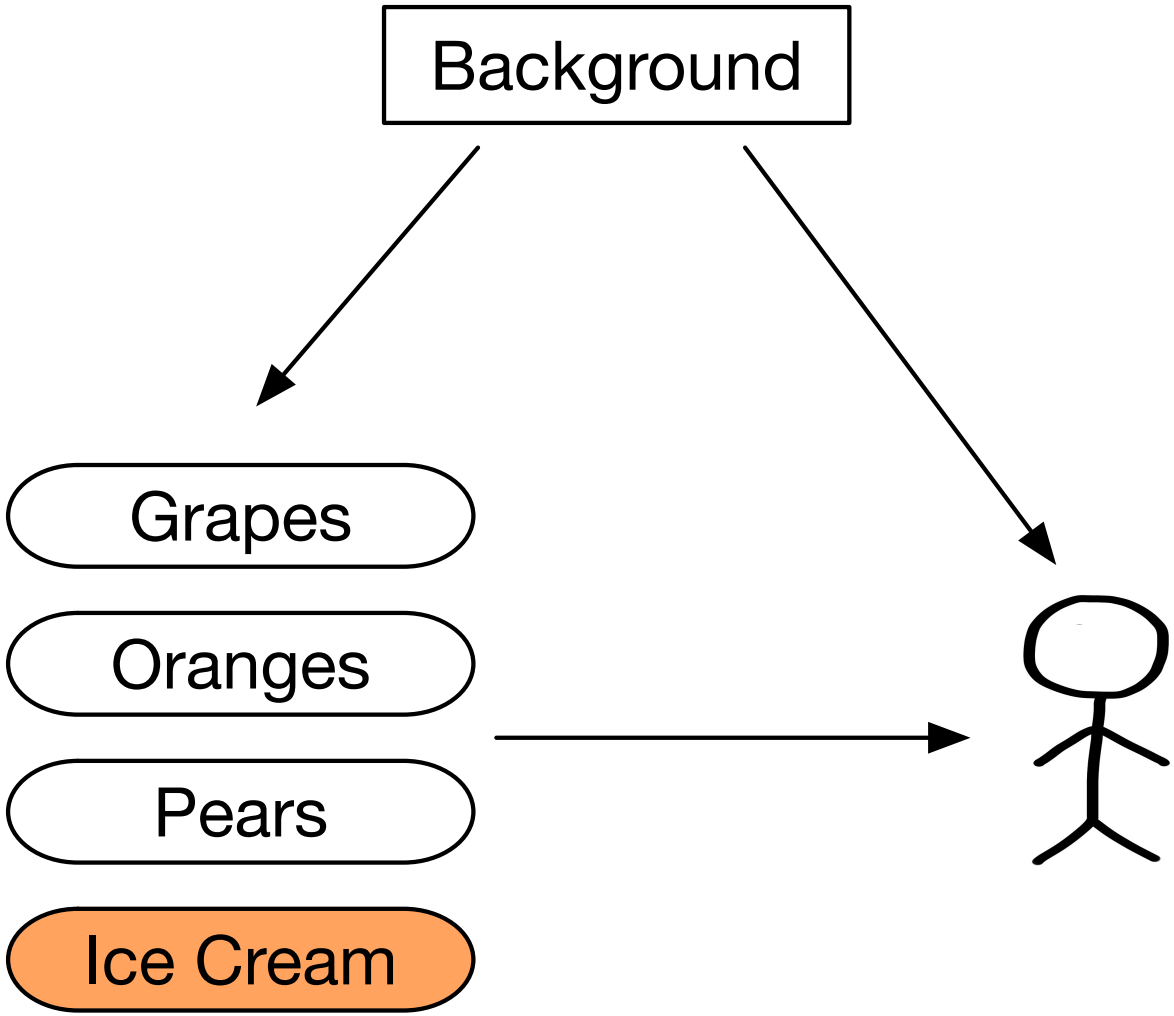


Does ice cream at age 3 make you live longer?

$$\mathbb{E}[\text{life-span} \mid do(\text{ice-cream-at-age-3} = \text{5-scoops})]$$

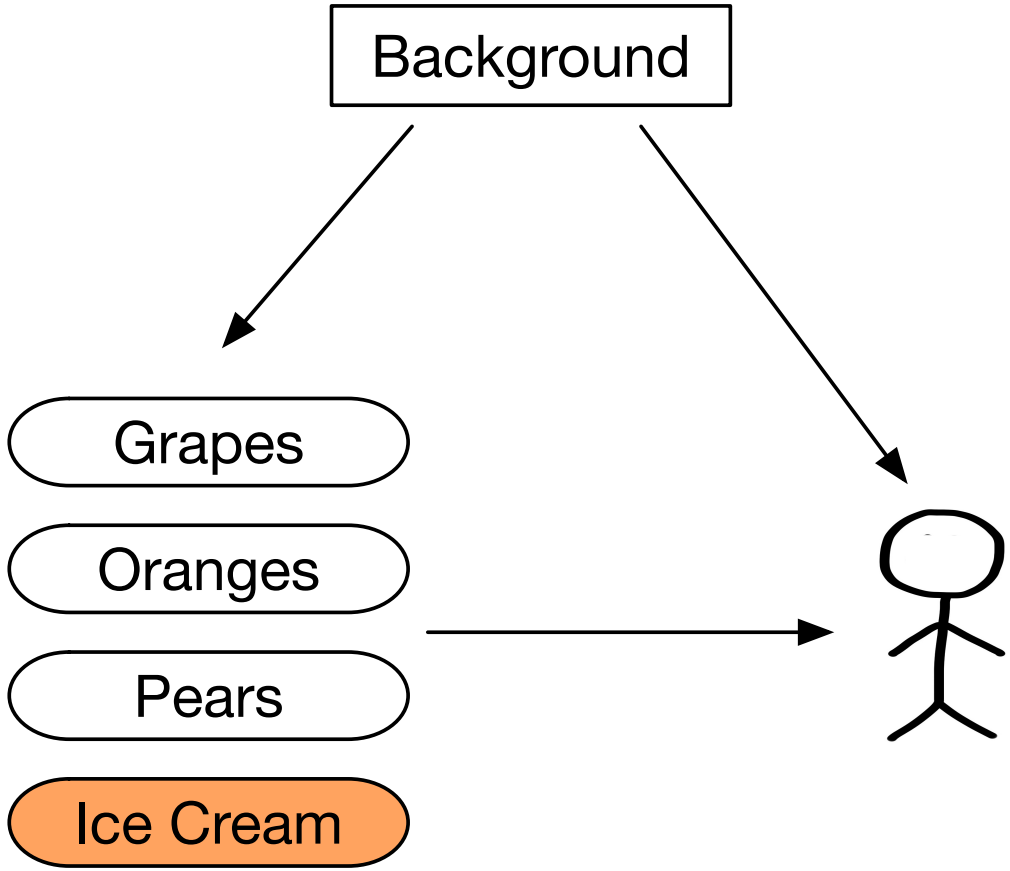


Look at the ice cream given at age 3 paired with lifespan



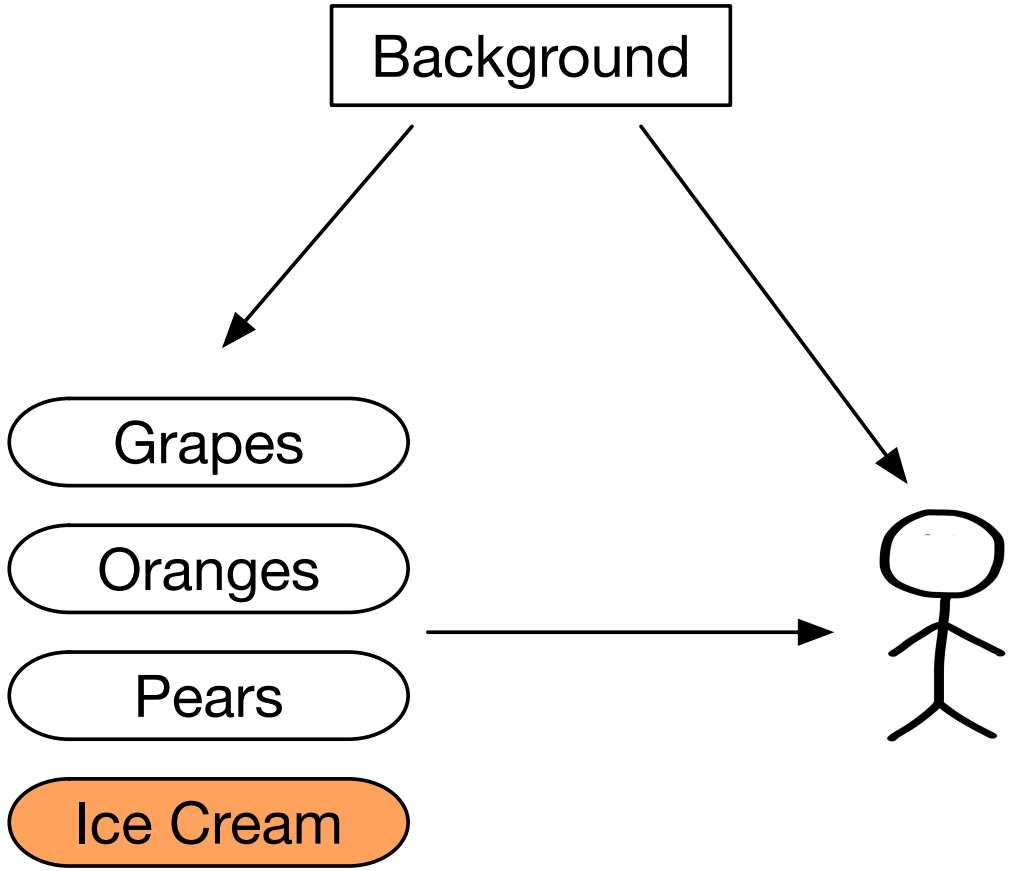
Need to separate out background variation from food variation

Called a confounder



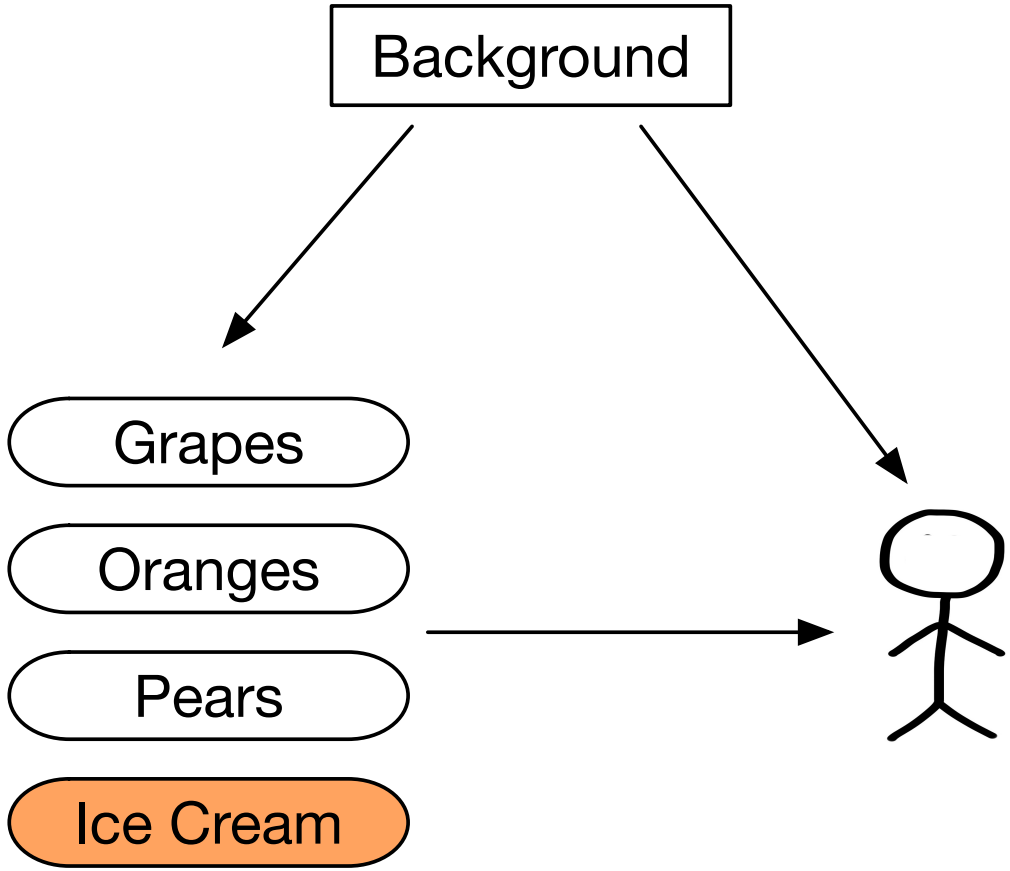
Know all background; With background fixed

an assumption

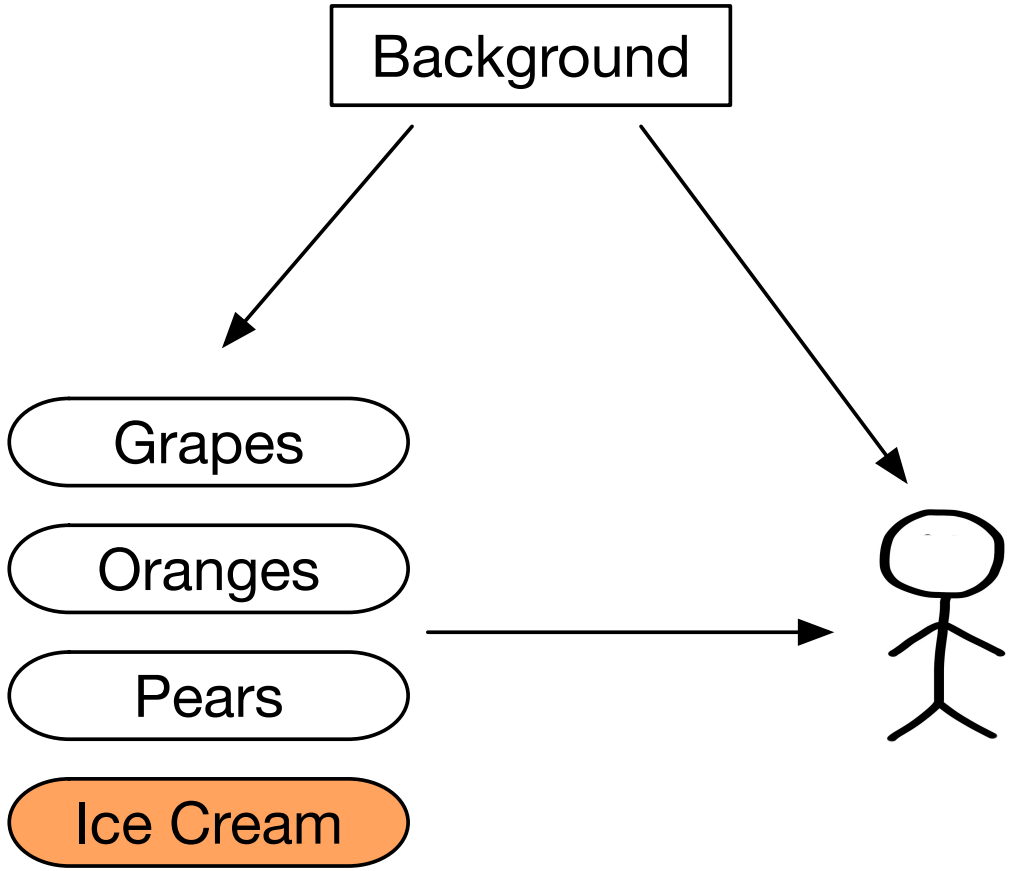


Know all background; With background fixed

- With background fixed
- Leftover variation between ice cream and lifespan
- Treatment needs have some variation given background



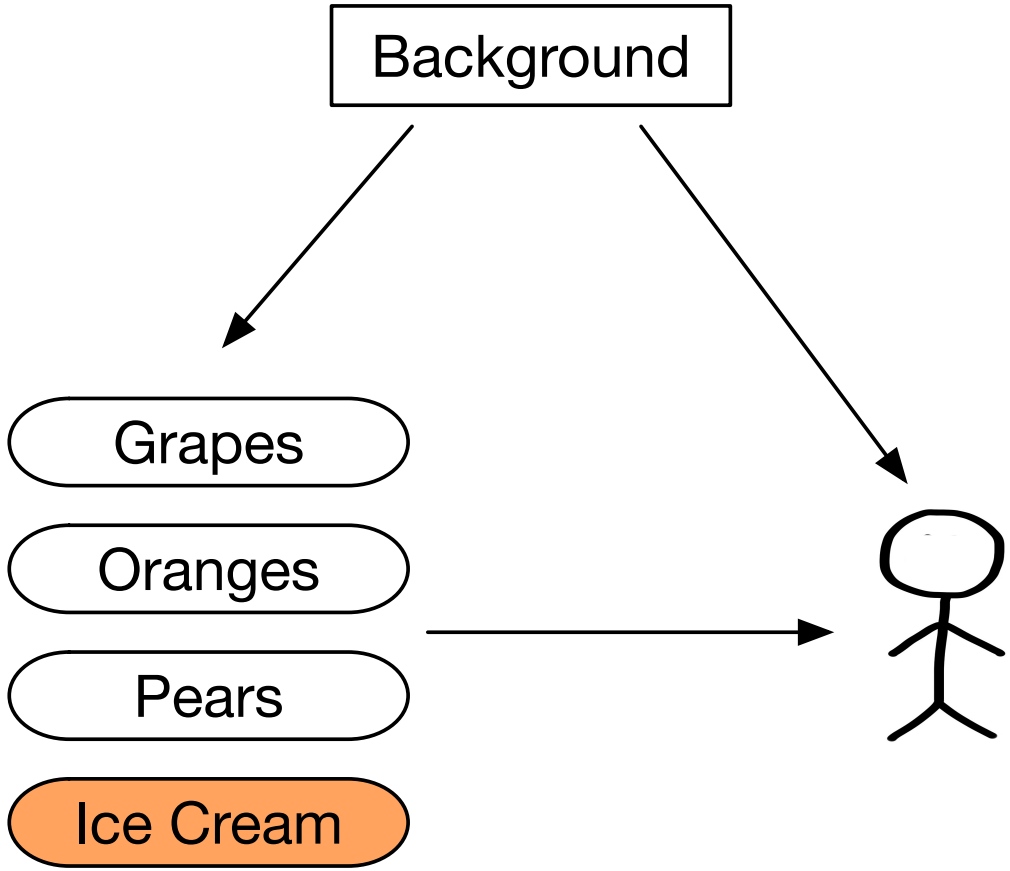
$$\mathbb{E}[\text{life-span} \mid do(\text{ice-cream} = 5)]$$
$$= \mathbb{E}_{\text{background}} \mathbb{E}[\text{life-span} \mid do(\text{ice-cream} = 5), \text{background}]$$
$$= \mathbb{E}_{\text{background}} \mathbb{E}[\text{life-span} \mid \text{ice-cream} = 5, \text{background}]$$



How does

$$\mathbb{E}[\text{life-span} \mid do(\text{ice-cream} = 5)]$$

relate to potential outcomes?

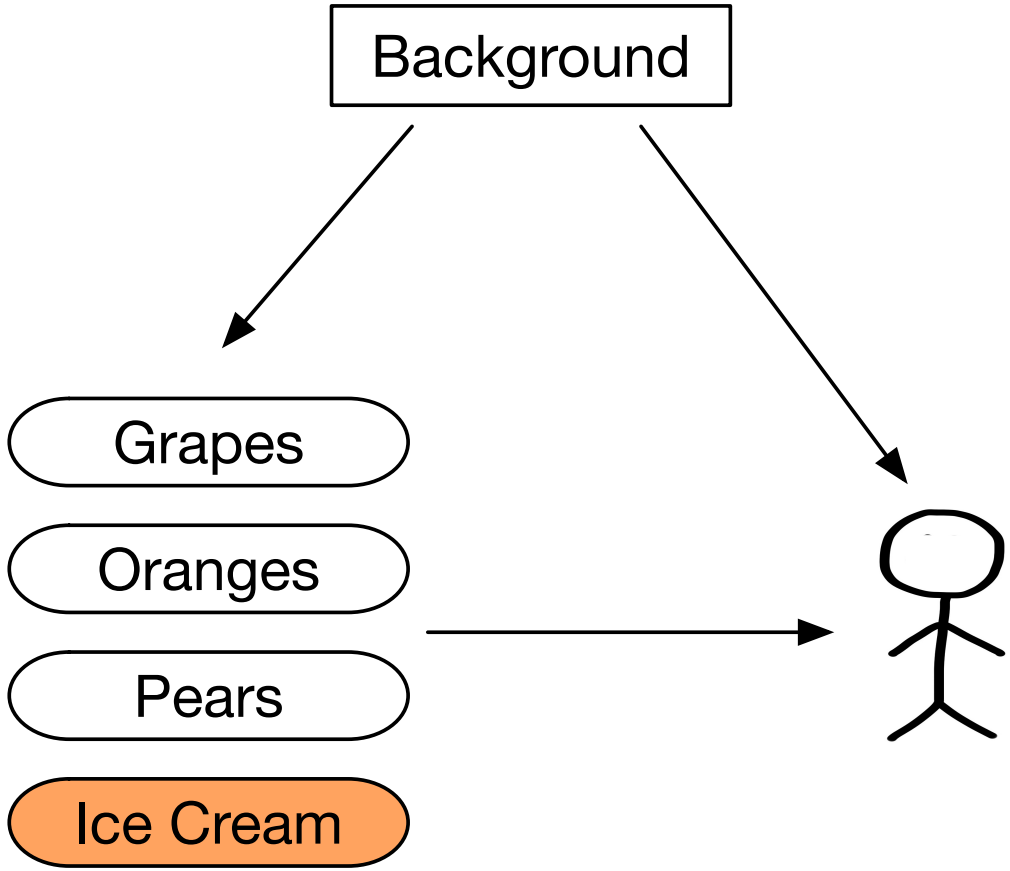


How does

$$\mathbb{E}[\text{life-span} \mid do(\text{ice-cream} = 5)]$$

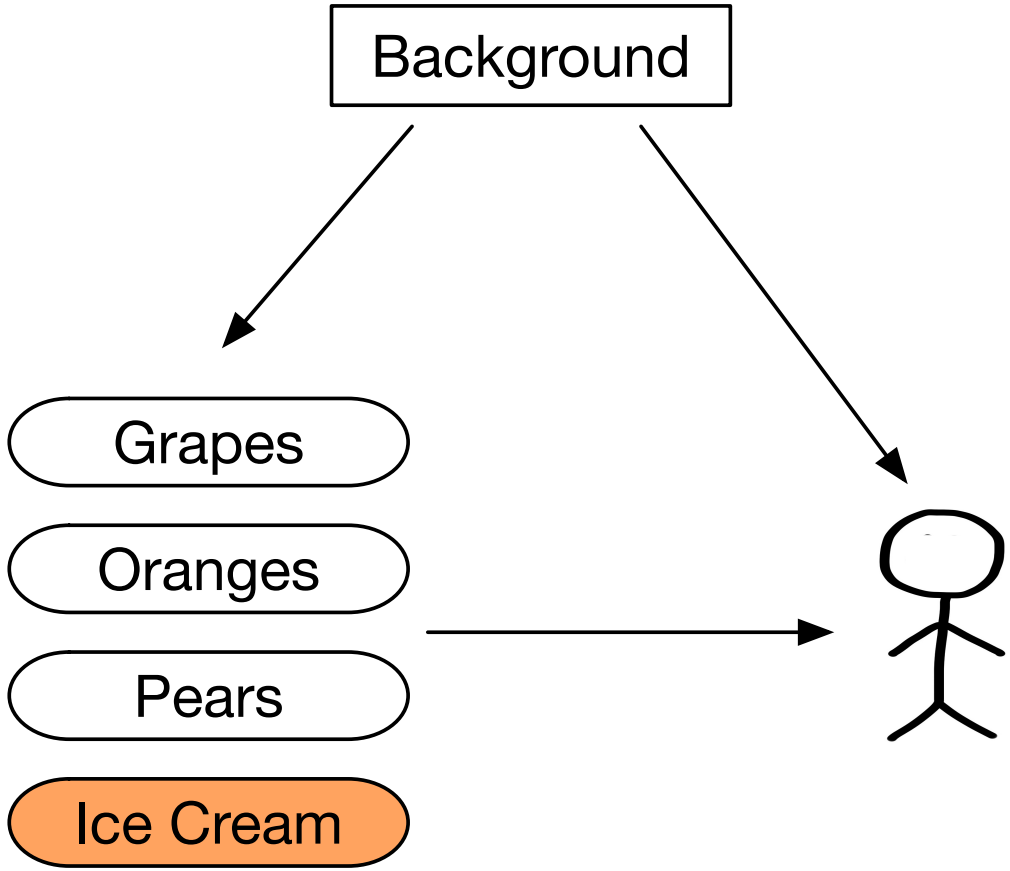
relate to potential outcomes?

$$\mathbb{E}[\text{life-span} \mid do(\text{ice-cream} = 5)] = \mathbb{E}[y_1 = \text{life-span}_{\text{ice-cream}=5}]$$

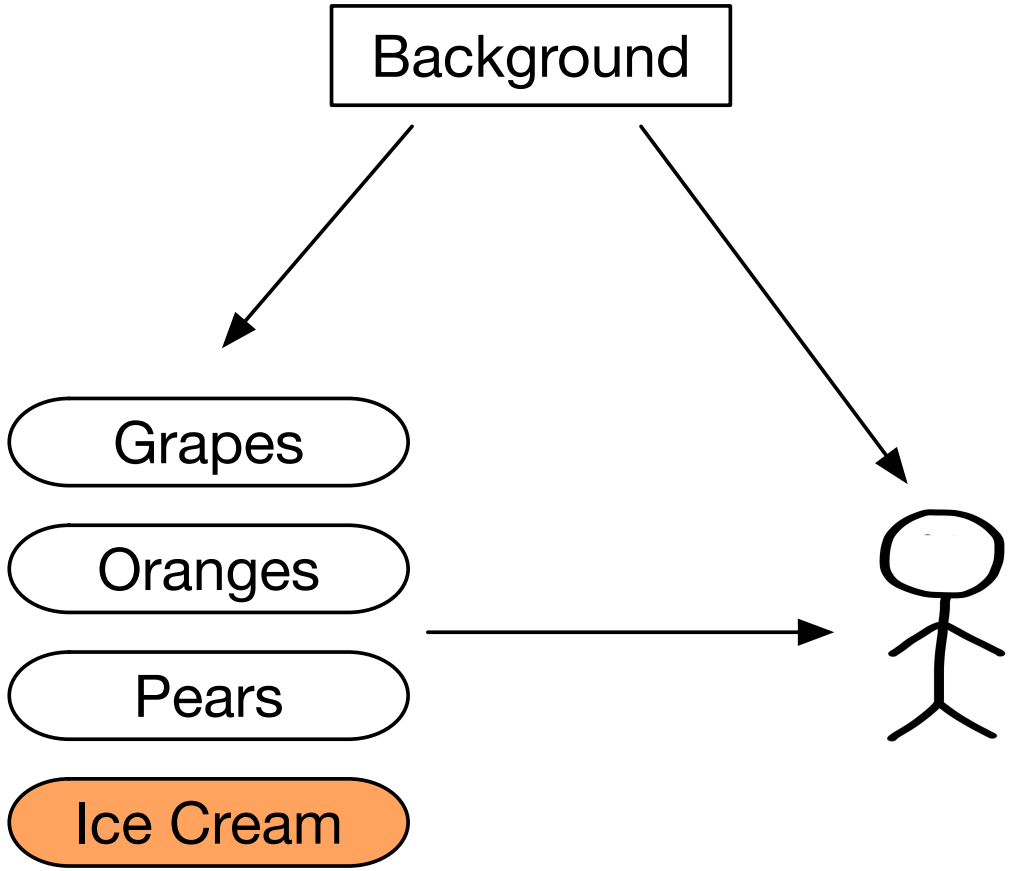


Notation

- y: outcome or response — lifespan
- $\mathbf{x}$: confounder — background
- t: treatment — ice cream at age 3



- Compute effects for each background
- Average over them



Variation in treatment given backgrounded make well-defined

$$\mathbb{E}[y \mid t,x] = \int y p(y \mid t,x) = \int y \frac{p(y,t,x)}{p(t,x)}$$

Otherwise, divide by zero

$$p(t \mid x)p(x) = p(t,x) = 0 \text{ if } p(t \mid x) = 0$$

Digging a bit deeper

The world has a true function form

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 10t_{\text{ice-cream}} + 10x_{\text{background}} + \epsilon_y$$

ϵ_y independent random noise

Digging a bit deeper

The world has a true function form

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 10t_{\text{ice-cream}} + 10x_{\text{background}} + \epsilon_y$$

ϵ_y independent random noise

If background perfectly predicts ice cream consumption

$$t_{\text{ice-cream}} = x_{\text{background}}$$

Then the response function is non-identifiable from data

Digging a bit deeper

The world has a true function form

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 10t_{\text{ice-cream}} + 10x_{\text{background}} + \epsilon_y$$

ϵ_y independent random noise

Could instead be

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 20x_{\text{background}} + \epsilon_y$$

Digging a bit deeper

The world has a true function form

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 10t_{\text{ice-cream}} + 10x_{\text{background}} + \epsilon_y$$

ϵ_y independent random noise

Or could instead be

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 20t_{\text{ice-cream}} + \epsilon_y$$

Digging a bit deeper

The world has a true function form

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 10t_{\text{ice-cream}} + 10x_{\text{background}} + \epsilon_y$$

ϵ_y independent random noise

Or could instead be

$$y_{\text{life-span}}(t_{\text{ice-cream}}, x_{\text{background}}, \epsilon_y) = 20t_{\text{ice-cream}} + \epsilon_y$$

All functions look the same relative to the observed data

View 3: The functional view

The functional view

The world has a true function form

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= g(\mathbf{x}, \boldsymbol{\epsilon}_t)\\ y &= h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

$\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_y$ are jointly independent

The functional view

The world has a true function form

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= g(\mathbf{x}, \boldsymbol{\epsilon}_t)\\ y &= h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

$\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_y$ are jointly independent

What does it mean to *intervene* on t by setting it to 5?

The functional view

The world has a true function form

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= g(\mathbf{x}, \boldsymbol{\epsilon}_t)\\ y &= h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

$\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_y$ are jointly independent

What does it mean to *intervene* on t by setting it to 5?

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= 5\\ y &= h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

The functional view

Observational Distribution

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= g(\mathbf{x}, \boldsymbol{\epsilon}_t)\\ y &= h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

Conditional expectation:

$$\begin{aligned}&\mathbb{E}[y \mid t = 5]\\ &\quad = \mathbb{E}_{p(\boldsymbol{\epsilon}_y, \mathbf{x} \mid t=5)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\\ &\quad = \mathbb{E}_{p(\mathbf{x} \mid t=5)p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

Intervened Distribution

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= 5\\ y &= h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

Conditional expectation:

$$\begin{aligned}&\mathbb{E}[y \mid t = 5]\\ &\quad = \mathbb{E}_{p(\boldsymbol{\epsilon}_y, \mathbf{x} \mid t=5)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\\ &\quad = \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

The functional view

Randomized Distribution

$$\mathbf{x} = f(\boldsymbol{\epsilon}_x)$$
$$t = g(\boldsymbol{\epsilon}_t)$$
$$y = h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)$$

Conditional expectation:

$$\begin{aligned}\mathbb{E}[y \mid t = 5] &= \mathbb{E}_{p(\boldsymbol{\epsilon}_y, \mathbf{x} \mid t=5)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)] \\ &= \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

Intervened Distribution

$$\mathbf{x} = f(\boldsymbol{\epsilon}_x)$$
$$t = 5$$
$$y = h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)$$

Conditional expectation:

$$\begin{aligned}\mathbb{E}[y \mid t = 5] &= \mathbb{E}_{p(\boldsymbol{\epsilon}_y, \mathbf{x} \mid t=5)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)] \\ &= \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

The functional view

Adjustment

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= g(\mathbf{x}, \boldsymbol{\epsilon}_t)\\ y &= h(t, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

Conditional expectation:

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}}[\mathbb{E}[y \mid t=5, \mathbf{x}]]\\ &\quad= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\boldsymbol{\epsilon}_y \mid t=5, \mathbf{x})}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\\ &\quad= \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

Intervened Distribution

$$\begin{aligned}\mathbf{x} &= f(\boldsymbol{\epsilon}_x)\\ t &= 5\\ y &= h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)\end{aligned}$$

Conditional expectation:

$$\begin{aligned}&\mathbb{E}[y \mid t=5]\\ &\quad= \mathbb{E}_{p(\boldsymbol{\epsilon}_y, \mathbf{x} \mid t=5)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\\ &\quad= \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon}_y)}[h(5, \mathbf{x}, \boldsymbol{\epsilon}_y)]\end{aligned}$$

Back to estimating with confounders

Computational Tool: Propensity Scores

- From strong ignorability $\mathbf{x}$ predicts treatment chance
- Treatment chance has enough information to break dependence

Build a treatment chance model with regression

$$p_{\phi}(t = 1 \mid \mathbf{x})$$

Called the *propensity score*.

Use importance sampling to compute expectations

Computational Tool: Propensity Scores and Regression

$$
\begin{aligned}
ATE &= \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[y_1 \mid \mathbf{x}]] - \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}[y_0 \mid \mathbf{x}]] \\
&= \int p(\mathbf{x}) \frac{p(t=1 \mid \mathbf{x})}{p(t=1 \mid \mathbf{x})} \mathbb{E}[y_1 \mid \mathbf{x}] d\mathbf{x} - \int p(\mathbf{x}) \frac{1 - p(t=1 \mid \mathbf{x})}{1 - p(t=1 \mid \mathbf{x})} \mathbb{E}[y_0 \mid \mathbf{x}] d\mathbf{x} \\
&= \int p(\mathbf{x}) \frac{\mathbb{E}[t \mid \mathbf{x}]}{p(t=1 \mid \mathbf{x})} \mathbb{E}[y_1 \mid \mathbf{x}] d\mathbf{x} - \int p(\mathbf{x}) \frac{\mathbb{E}[1-t \mid \mathbf{x}]}{1 - p(t=1 \mid \mathbf{x})} \mathbb{E}[y_0 \mid \mathbf{x}] d\mathbf{x} \\
&= \int p(\mathbf{x}) \frac{\mathbb{E}[t y_1 \mid \mathbf{x}]}{p(t=1 \mid \mathbf{x})} d\mathbf{x} - \int p(\mathbf{x}) \frac{\mathbb{E}[y_0(1-t) \mid \mathbf{x}]}{1 - p(t=1 \mid \mathbf{x})} d\mathbf{x} \\
&= \int p(\mathbf{x}) \frac{\mathbb{E}[t y \mid \mathbf{x}]}{p(t=1 \mid \mathbf{x})} d\mathbf{x} - \int p(\mathbf{x}) \frac{\mathbb{E}[y(1-t) \mid \mathbf{x}]}{1 - p(t=1 \mid \mathbf{x})} d\mathbf{x}
\end{aligned}
$$

How do we compute this?

Computational Tool: Propensity Scores and Regression

$$
\begin{aligned}
ATE &= \int p(\mathbf{x}) \frac{\mathbb{E}[ty \mid \mathbf{x}]}{p(t=1 \mid \mathbf{x})} d\mathbf{x} - \int p(\mathbf{x}) \frac{\mathbb{E}[y(1-t) \mid \mathbf{x}]}{1-p(t=1 \mid \mathbf{x})} d\mathbf{x} \\
&= \int p(\mathbf{x}) \frac{\int typ(t,y \mid \mathbf{x}) dtdy}{p(t=1 \mid \mathbf{x})} d\mathbf{x} - \int p(\mathbf{x}) \frac{\int y(1-t)p(t,y \mid \mathbf{x}) dtdy}{1-p(t=1 \mid \mathbf{x})} d\mathbf{x} \\
&= \int \frac{p(\mathbf{x})p(t,y \mid \mathbf{x})ty}{p(t=1 \mid \mathbf{x})} dtdyd\mathbf{x} - \int \frac{p(\mathbf{x})p(t,y \mid \mathbf{x})y(1-t)}{1-p(t=1 \mid \mathbf{x})} dtdyd\mathbf{x} \\
&= \int \frac{p(\mathbf{x},t,y)ty}{p(t=1 \mid \mathbf{x})} dtdyd\mathbf{x} - \int \frac{p(\mathbf{x},t,y)y(1-t)}{1-p(t=1 \mid \mathbf{x})} dtdyd\mathbf{x}
\end{aligned}
$$

How do we compute this?

Combine Propensity Scores and Regression?

Evaluation?