# Identification of Causal Effect in the Presence of Selection Bias

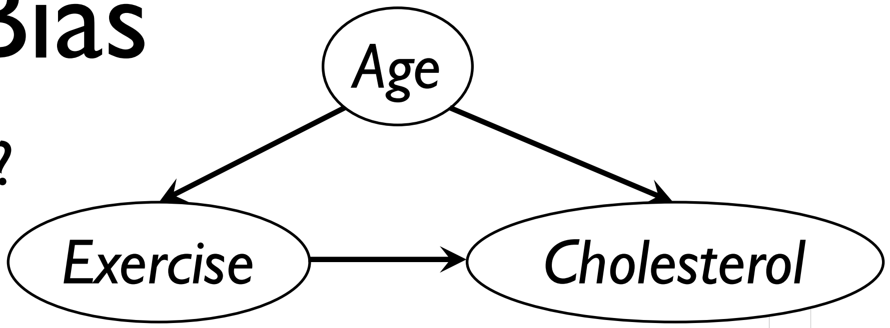
**Juan D. Correa** Jin Tian Elias Bareinboim

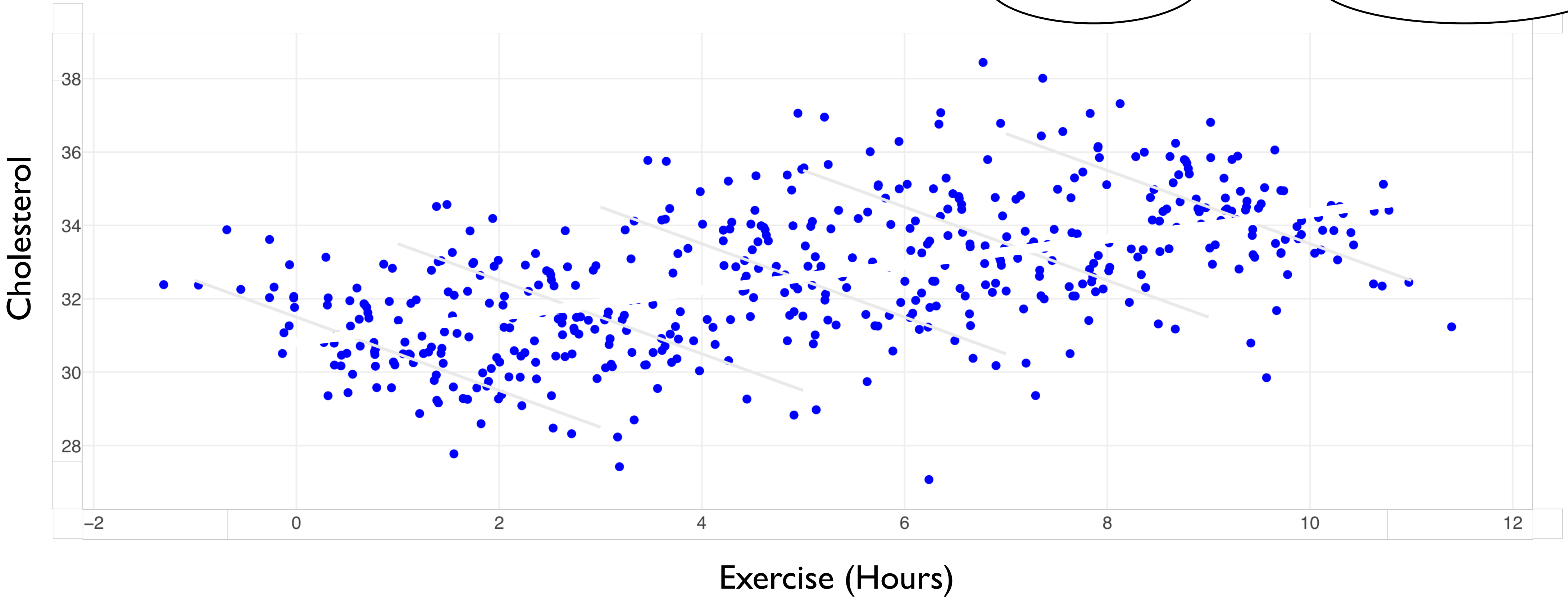
AAAI
Honolulu, 2019

# Challenge 1: Confounding Bias

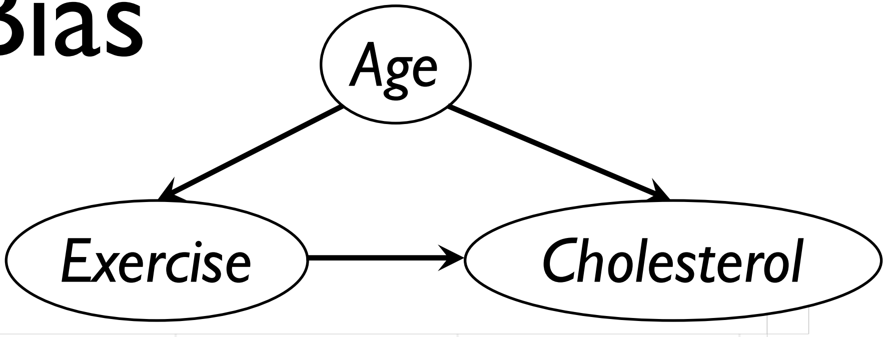
What's the causal effect of *Exercise* on *Cholesterol*?

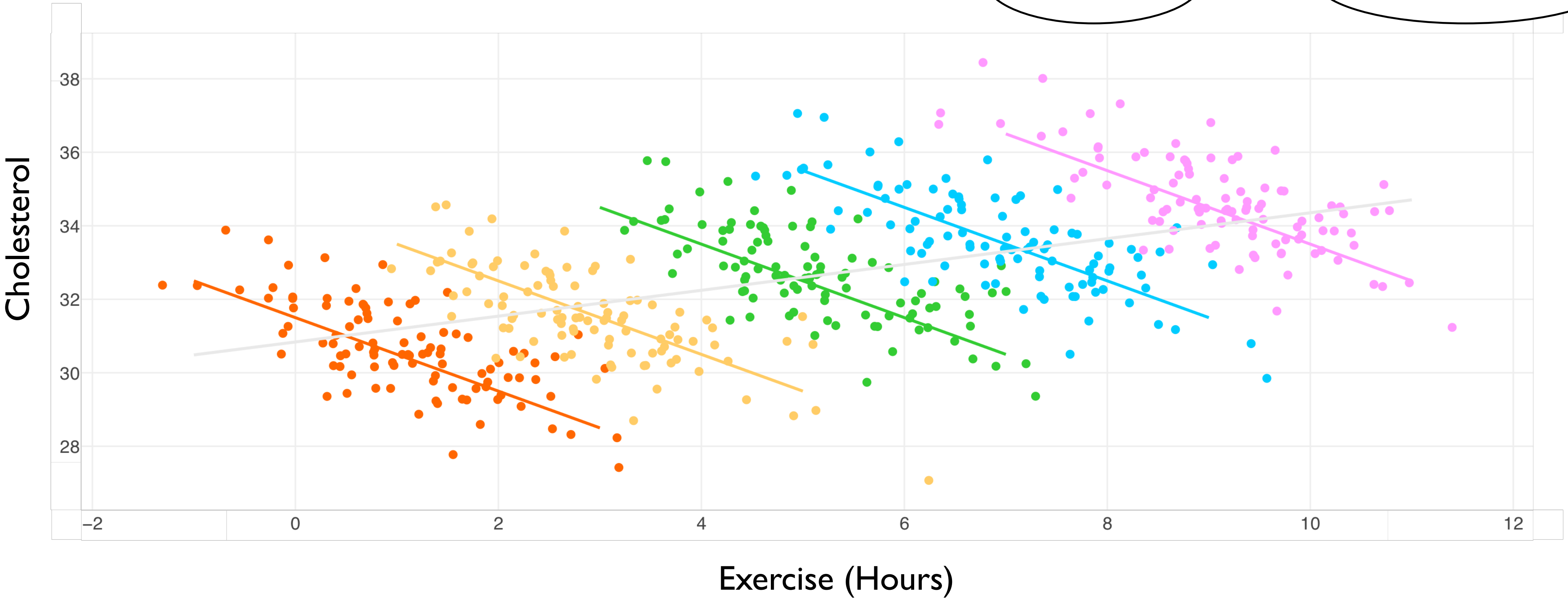
What about $P(cholesterol \mid exercise)$ ?



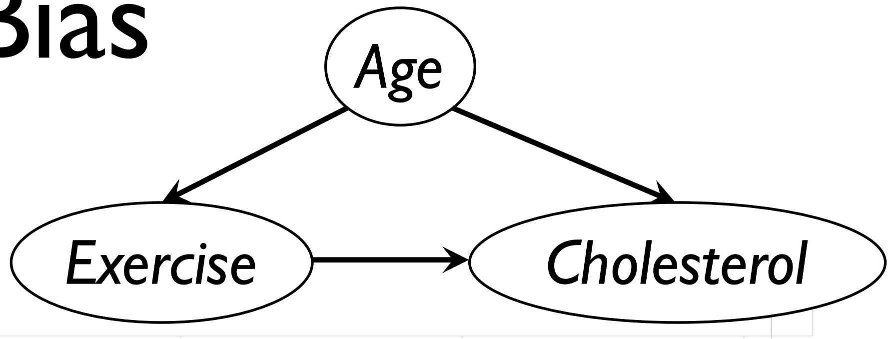



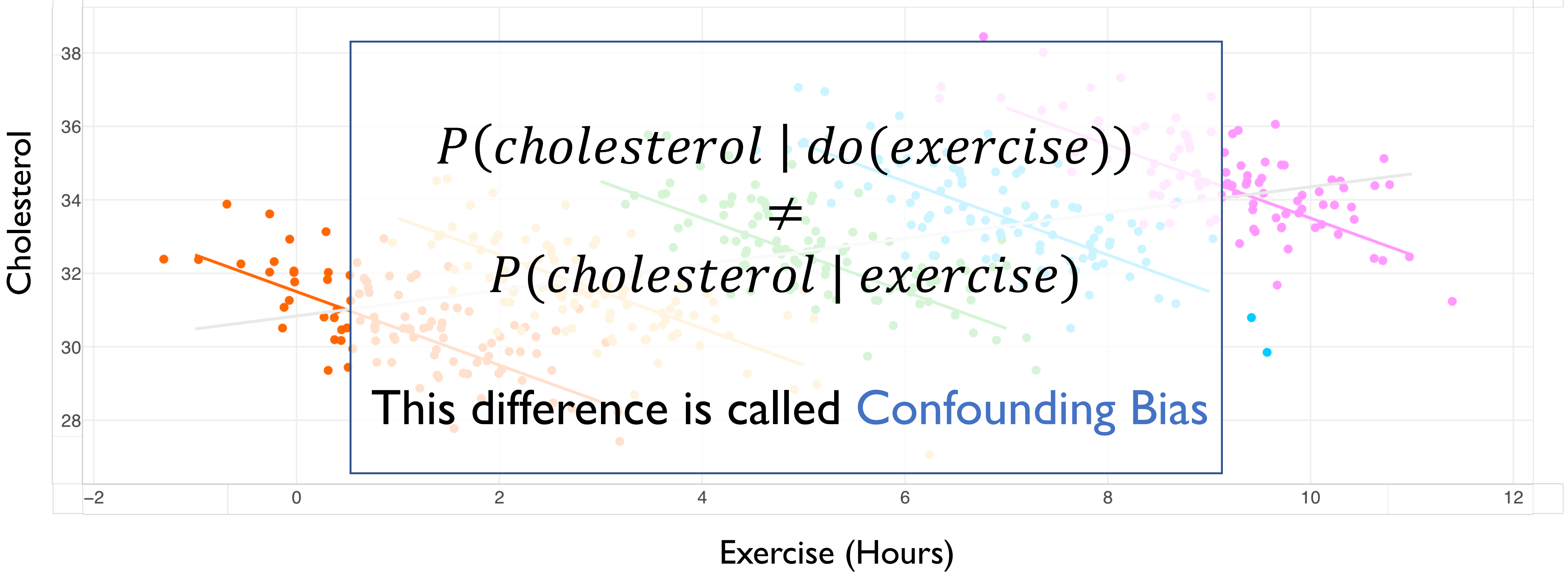
# Challenge 1: Confounding Bias



# Challenge 1: Confounding Bias

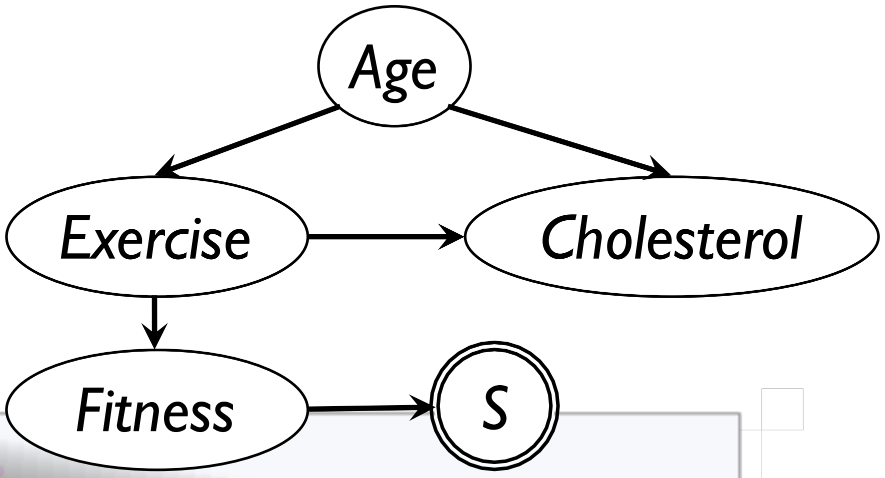


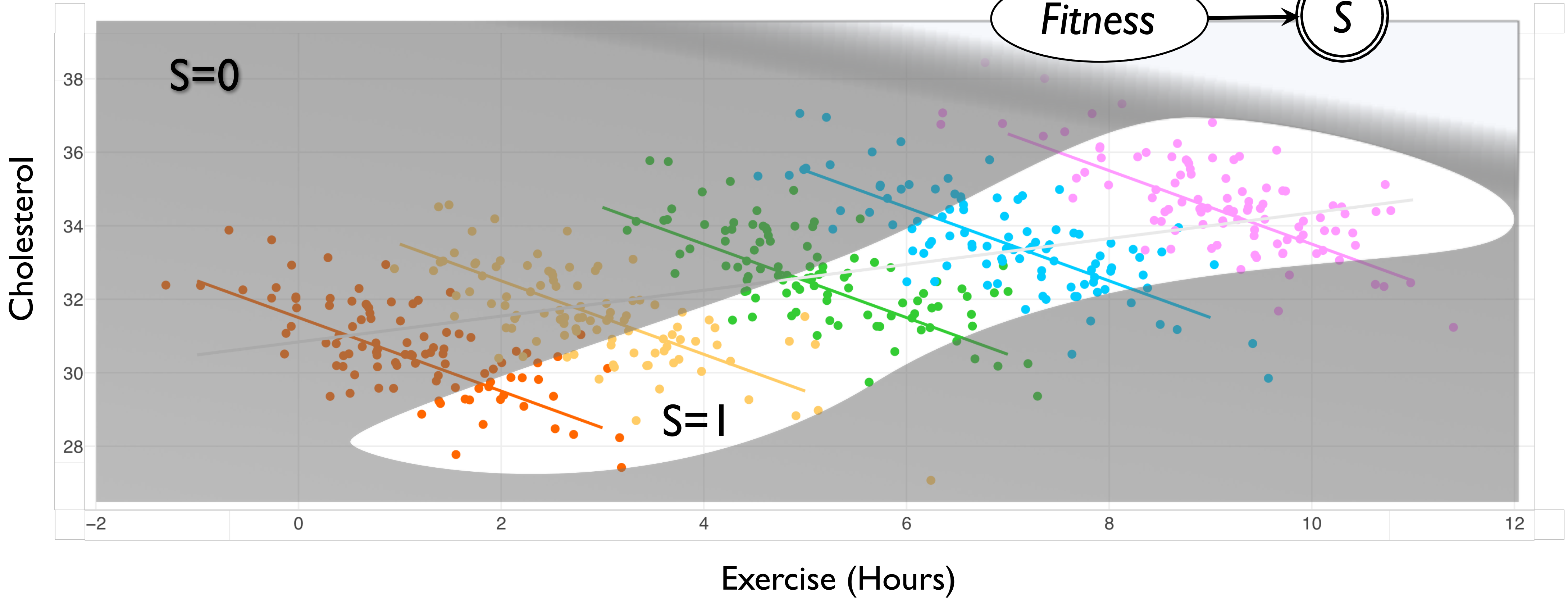

$$P(cholesterol \mid do(exercise)) \neq P(cholesterol \mid exercise)$$

This difference is called Confounding Bias

# Challenge 2: Selection Bias

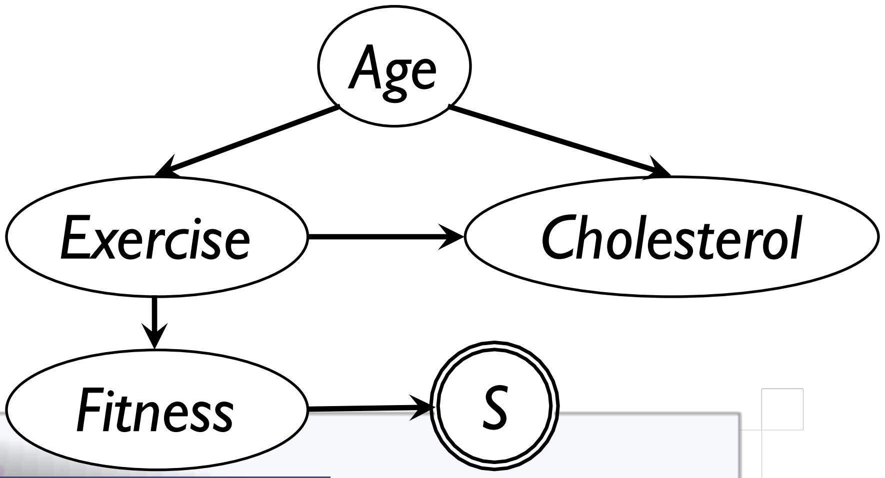


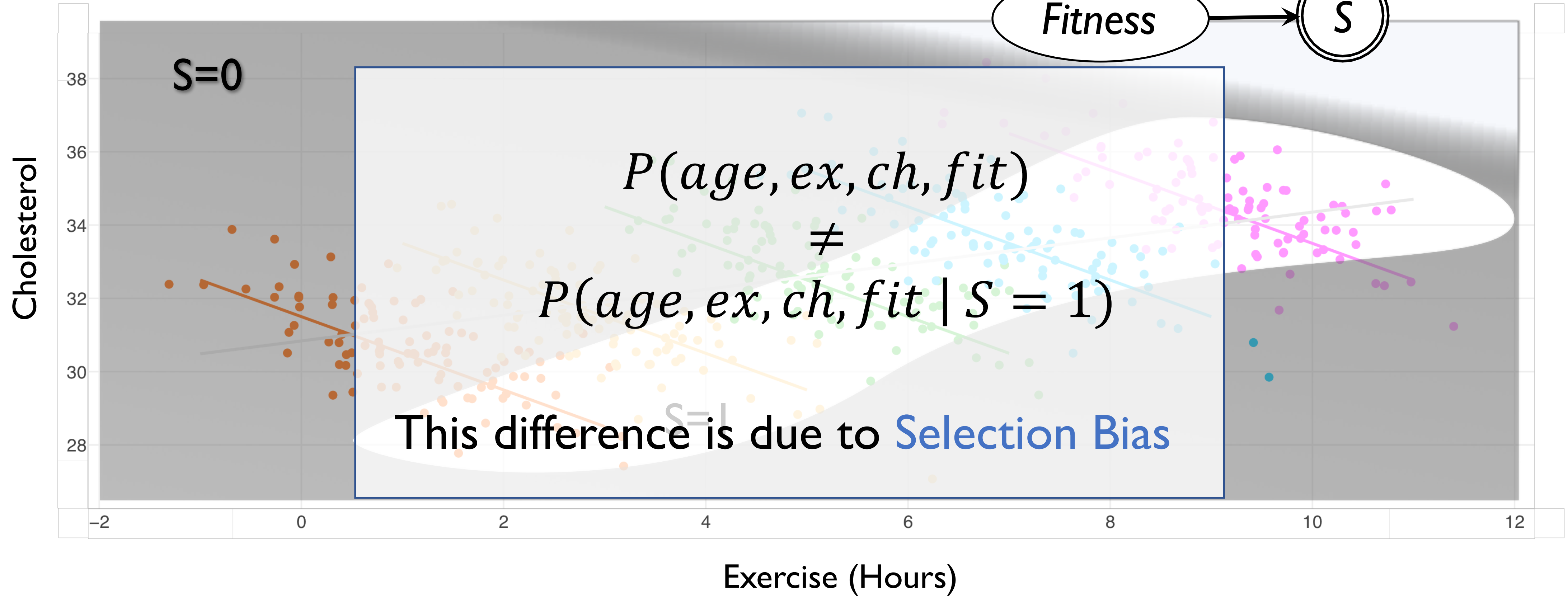

Variables in the system affect the inclusion of units in the sample



# Challenge 2: Selection Bias

Variables in the system affect the inclusion of units in the sample




$$P(age, ex, ch, fit) \neq P(age, ex, ch, fit \mid S = 1)$$

This difference is due to Selection Bias

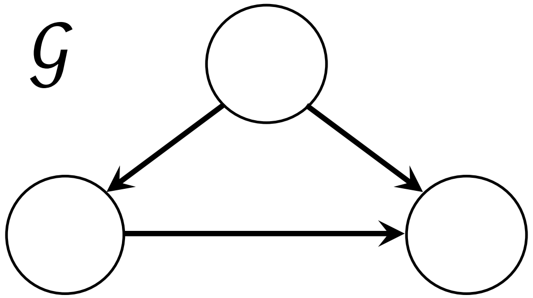
# Current literature

| | No Confounding | Confounding |
| --- | --- | --- |
| No Selection | Association = Causation<br>No control | Complete Algorithms<br>[Tian and Pearl '02; Huang and Valtorta '06; Shpitser and Pearl '06; Bareinboim and Pearl '12] |
| Selection | Controlling Selection Bias<br>[Bareinboim and Pearl '12]<br>Recovering from Selection Bias in Causal and Statistical Inference<br>[Bareinboim, Tian, Pearl '14] | RCE<br>[Bareinboim, Tian, Pearl '15]<br>Generalized Adjustment<br>[Correa, Tian, Bareinboim '18]<br>**IDSB**<br>[Correa, Tian, Bareinboim '19] |

# Problem I

Given:

$\mathcal{G}$



Variables

$\boldsymbol{X}, \boldsymbol{Y}$

| | | | $S$ | $P(v \mid S=1)$ |
|---|---|---|---|---|
| | | | 1 | … |
| | | | 1 | … |
| | | | 1 | … |

$P$

Is there a function $f$ such that

$$P(\boldsymbol{y} \mid do(\boldsymbol{x})) = f(P_1)$$

?

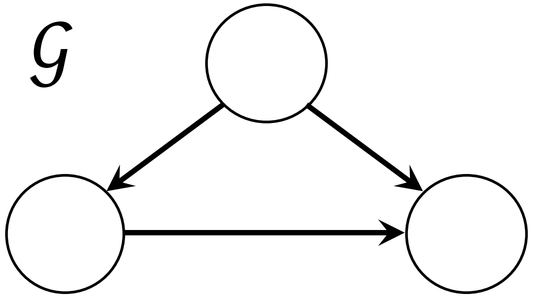
# Result 1

Theorem 1:
Let $\boldsymbol{X}, \boldsymbol{Y} \subset \boldsymbol{V}$ be two disjoint sets of variables and $\mathcal{G}$ a causal diagram over $\boldsymbol{V}$ and $S$. If $(\boldsymbol{Y} \perp S)_{\mathcal{G}^{pbd}_{\boldsymbol{XY}}}$, then $P_{\boldsymbol{x}}(\boldsymbol{y})$ is not recoverable from $P(\boldsymbol{v} \mid S = 1)$ in $\mathcal{G}$.

# Problem II

Given:

$\mathcal{G}$



Variables
$\boldsymbol{X}, \boldsymbol{Y}$

| | | | $S$ | $P(\boldsymbol{v} \mid S = 1)$ |
|---|---|---|---|---|
| | | | 1 | … |
| | | | 1 | … |
| | | | 1 | … |

$P_1$

| | $P(\boldsymbol{t})$ |
|---|---|
| | … |
| | … |
| | … |

$P_2$

Is there a function $f$ such that

$$P(\boldsymbol{y} \mid do(\boldsymbol{x})) = f(P_1, P_2)$$

?

# Result II

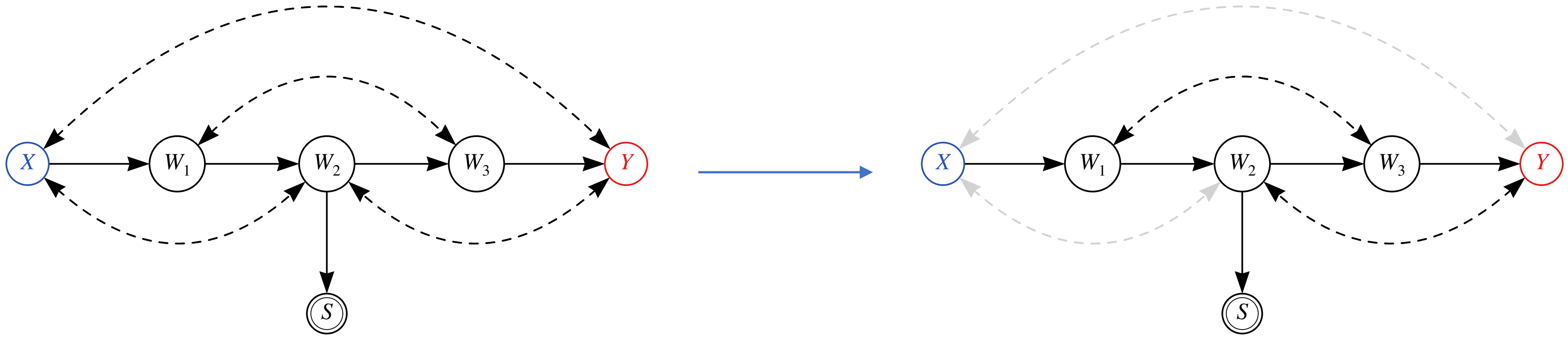
Algorithm **IDSB**

Given a causal diagram, a selection-biased distribution and external data over a subset of the variables and the variables of interest $(\boldsymbol{X}, \boldsymbol{Y})$; returns an expression for $P_{\boldsymbol{x}}(\boldsymbol{y})$ in terms of the input or *failure*.

Strictly more powerful than the best known algorithm that accepts both biased and unbiased data.

# Decomposing the Problem

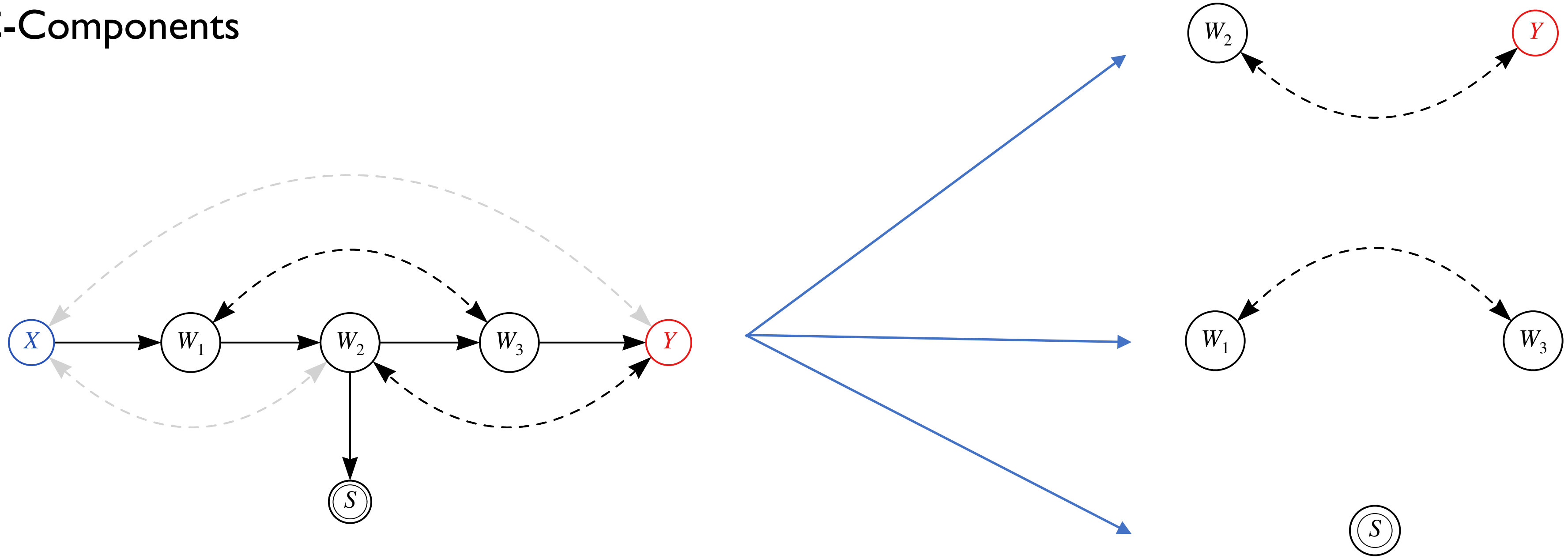
Intervention




$$P_x(y) = \sum_{w_1,w_2,w_3} P_x(y, w_3, w_2, w_1)$$

# Decomposing the Problem

C-Components




$P_{x,w_1,w_3}(y, w_2)$

$P_{w_2,y}(w_3, w_1)$

$$P_x(y) = \sum_{w_1,w_2,w_3} P_x(y, w_3, w_2, w_1) = \sum_{w_1,w_2,w_3} P_{x,w_1,w_3}(y, w_2)\, P_{w_2,y}(w_3, w_1)$$

# Summary

1. Complete characterization recoverable causal effects from the causal diagram and a selection-biased probability distribution.
2. Sufficient procedure to recover causal effects from a causal diagram, selection-biased distributions and auxiliary unbiased data which is strictly more powerful than state-of-the-art procedure.

Thanks!