From Statistical Transportability to Estimating the Effect of Stochastic Interventions

Juan D. Correa and Elias Bareinboim
{j.d.correa, eliasb}@columbia.edu

COLUMBIA UNIVERSITY

PURDUE UNIVERSITY

Generalization Challenges

Generalization Challenges

- One of the main tasks in ML is to learn/train models of an underlying process using data generated by the same process.

Generalization Challenges

- One of the main tasks in ML is to learn/train models of an underlying process using data generated by the same process.

- In fact, whenever enough data is provided, several approaches are currently capable of learning very accurately the underlying distribution.

Generalization Challenges

- One of the main tasks in ML is to learn/train models of an underlying process using data generated by the same process.

- In fact, whenever enough data is provided, several approaches are currently capable of learning very accurately the underlying distribution.

- In practice, however, the environment in which the data is collected is almost never the same as the one where the model is intended to be used, and will be deployed.

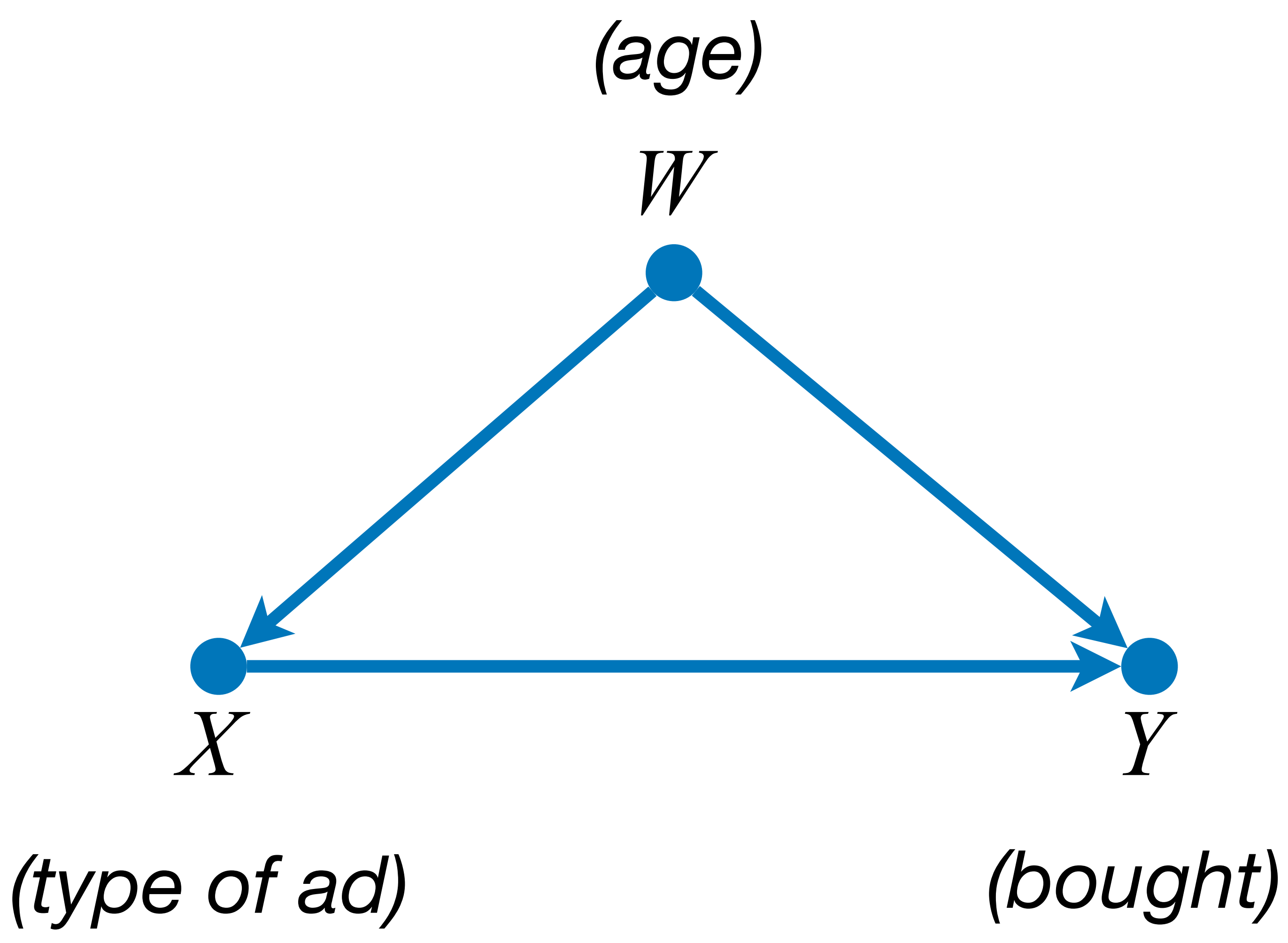
Generalization Challenges

- One of the main tasks in ML is to learn/train models of an underlying process using data generated by the same process.
- In fact, whenever enough data is provided, several approaches are currently capable of learning very accurately the underlying distribution.
- In practice, however, the environment in which the data is collected is almost never the same as the one where the model is intended to be used, and will be deployed.
- Under these constraints, the performance of the model depends on the underlying, structural similarities between training and target environments.

Statistical Transportability

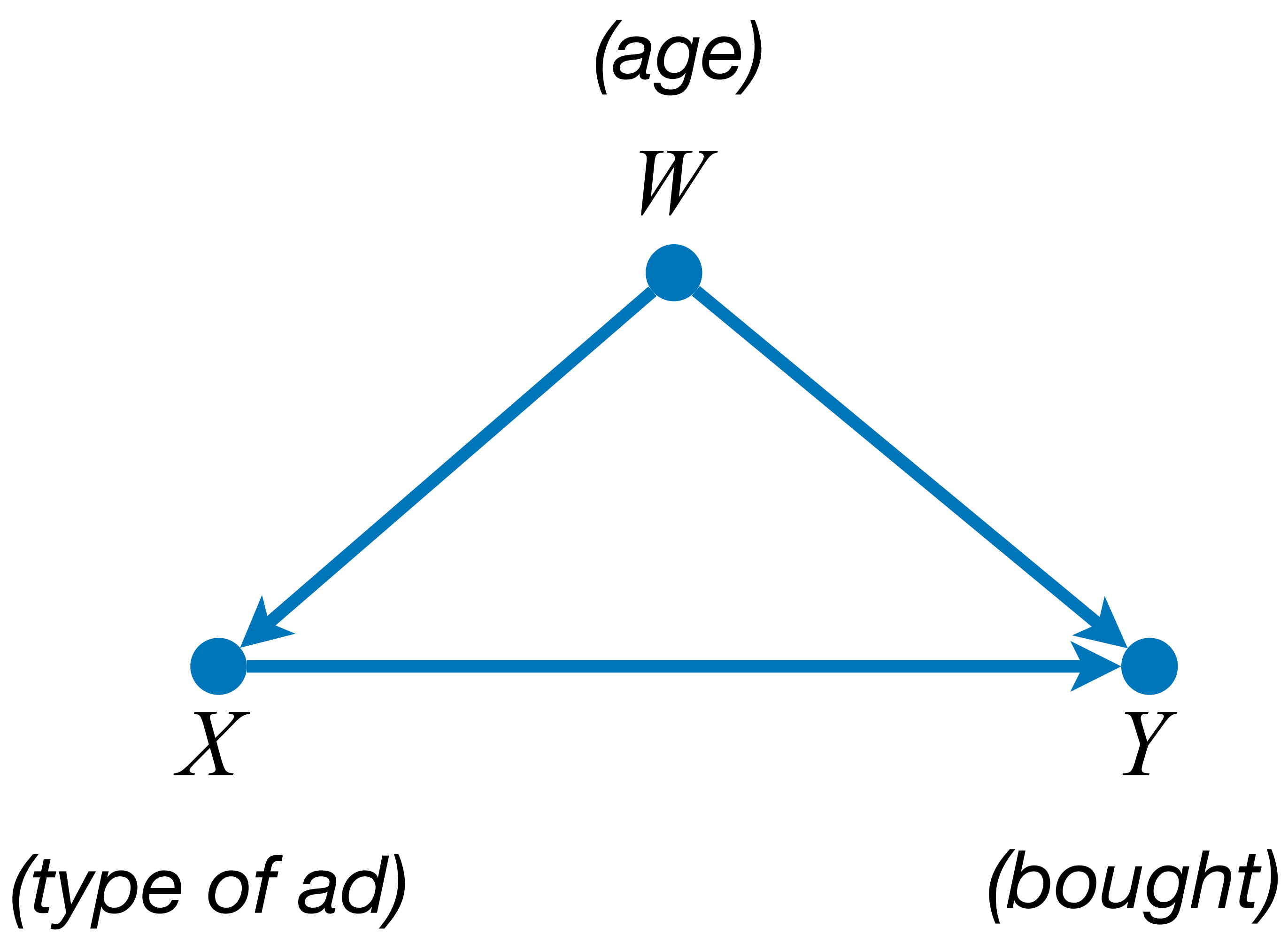
Statistical Transportability

Current Website ($\mathbf{\Pi}$)
(training environment)

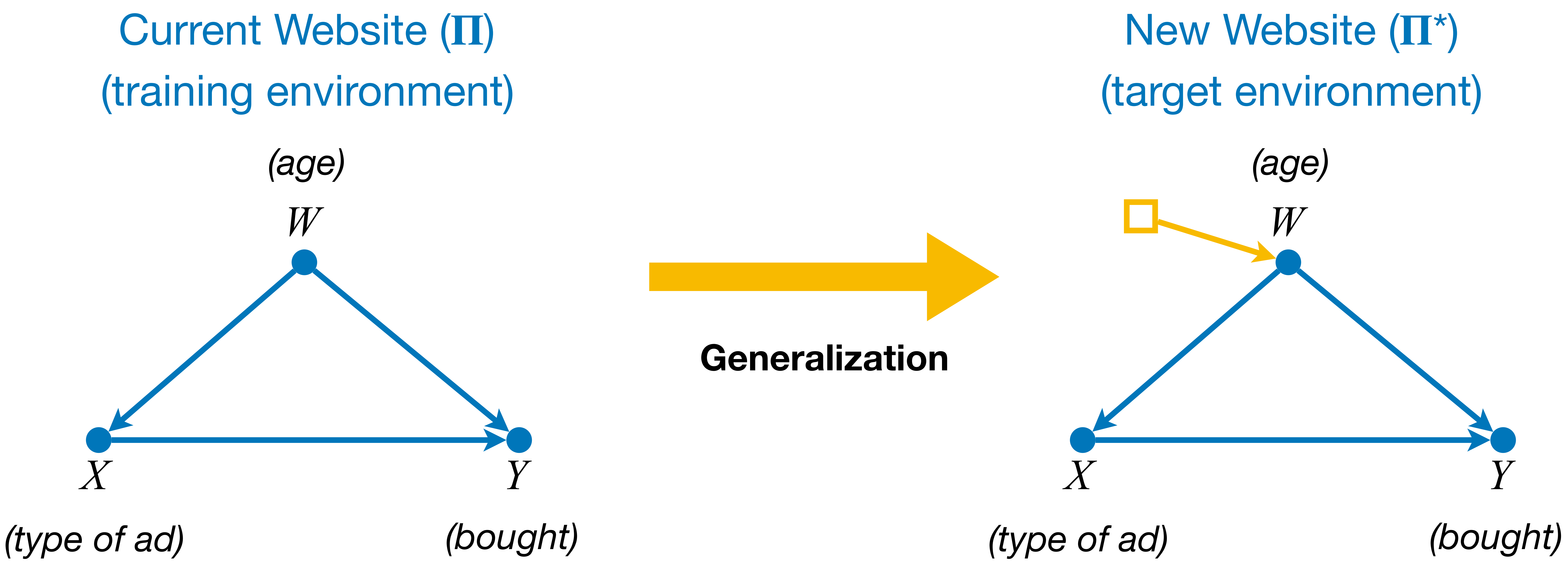


Statistical Transportability

Current Website (Π)
(training environment)

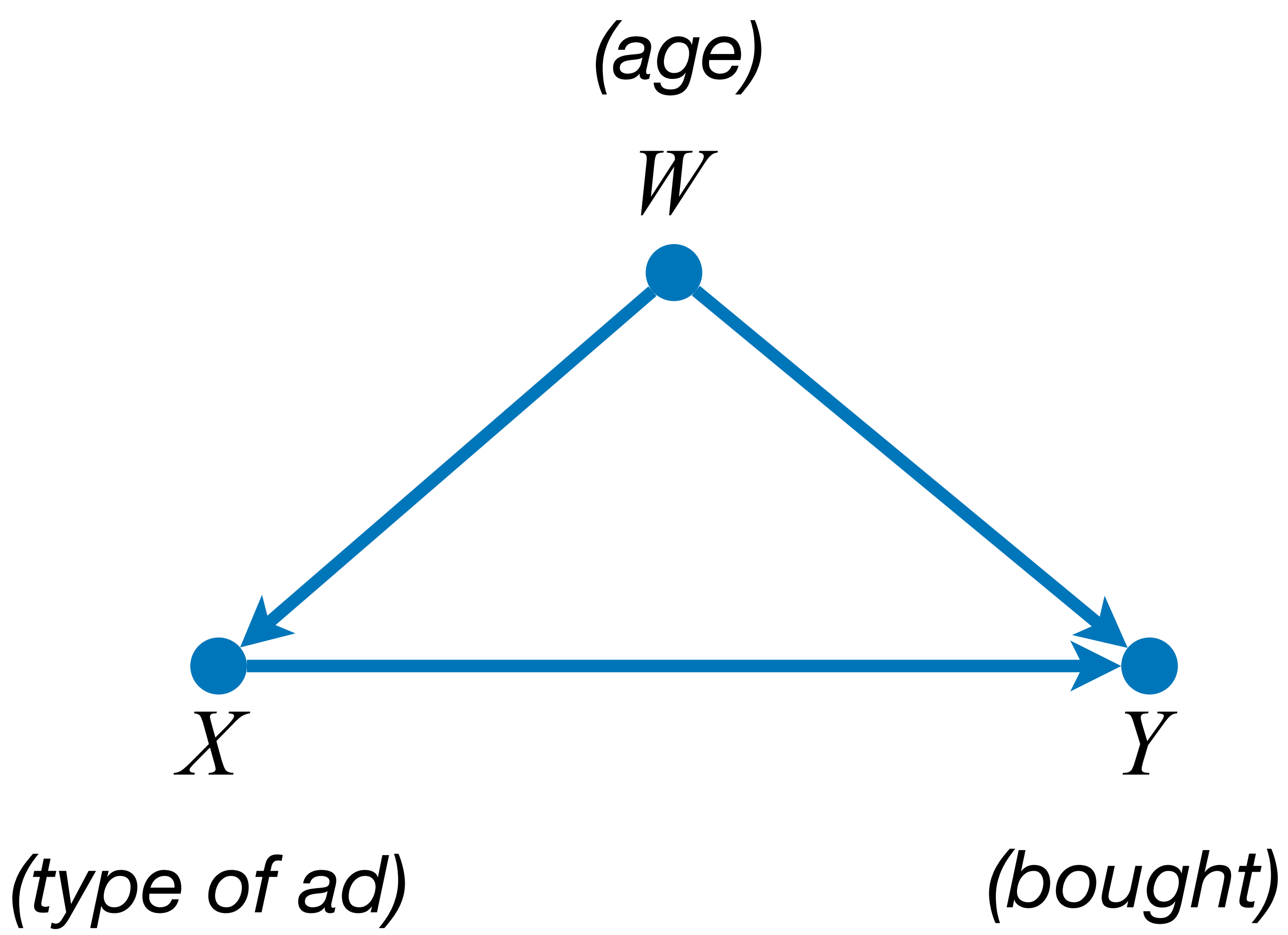


Statistical Transportability



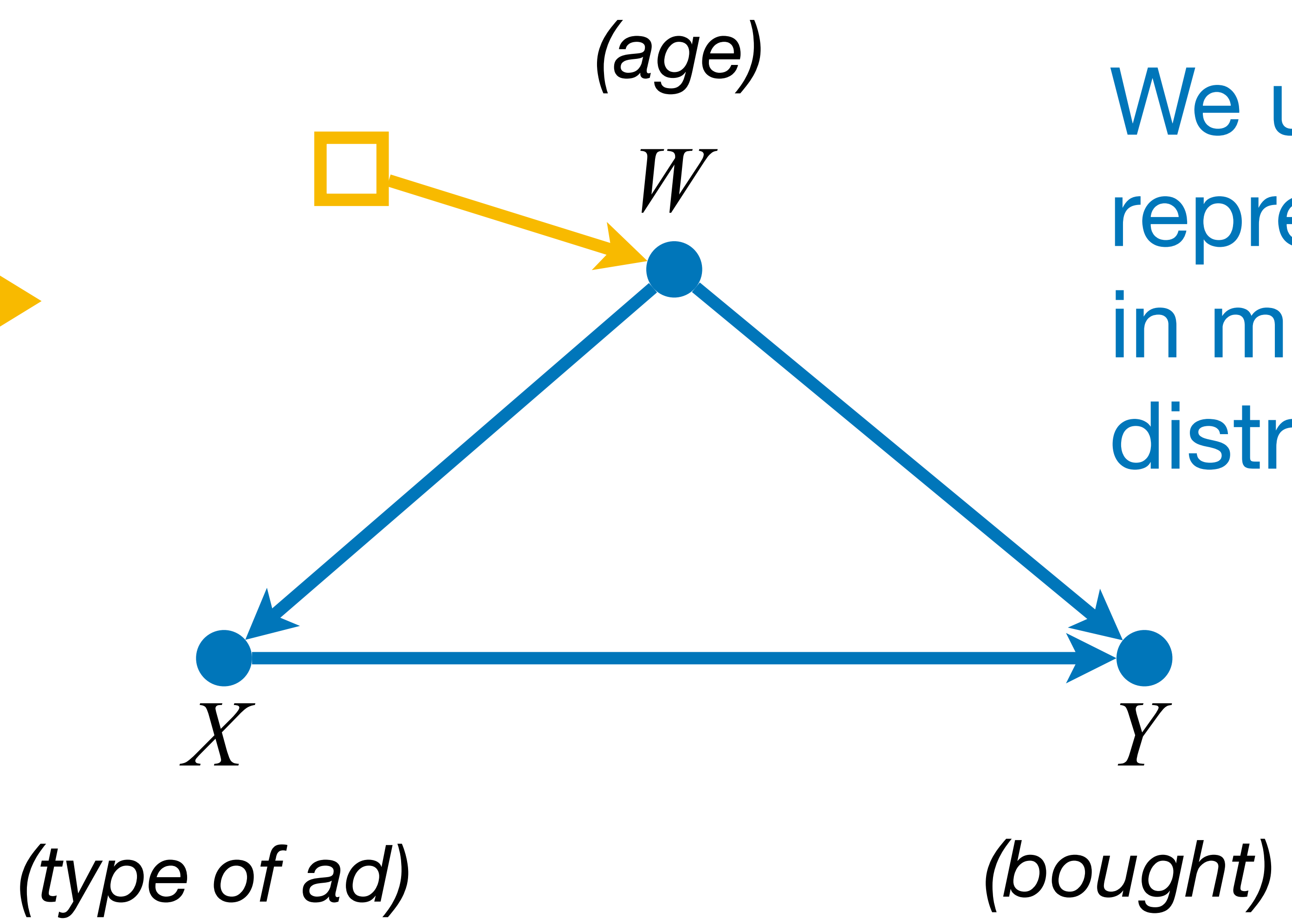
Statistical Transportability

Current Website (Π)
(training environment)



Generalization

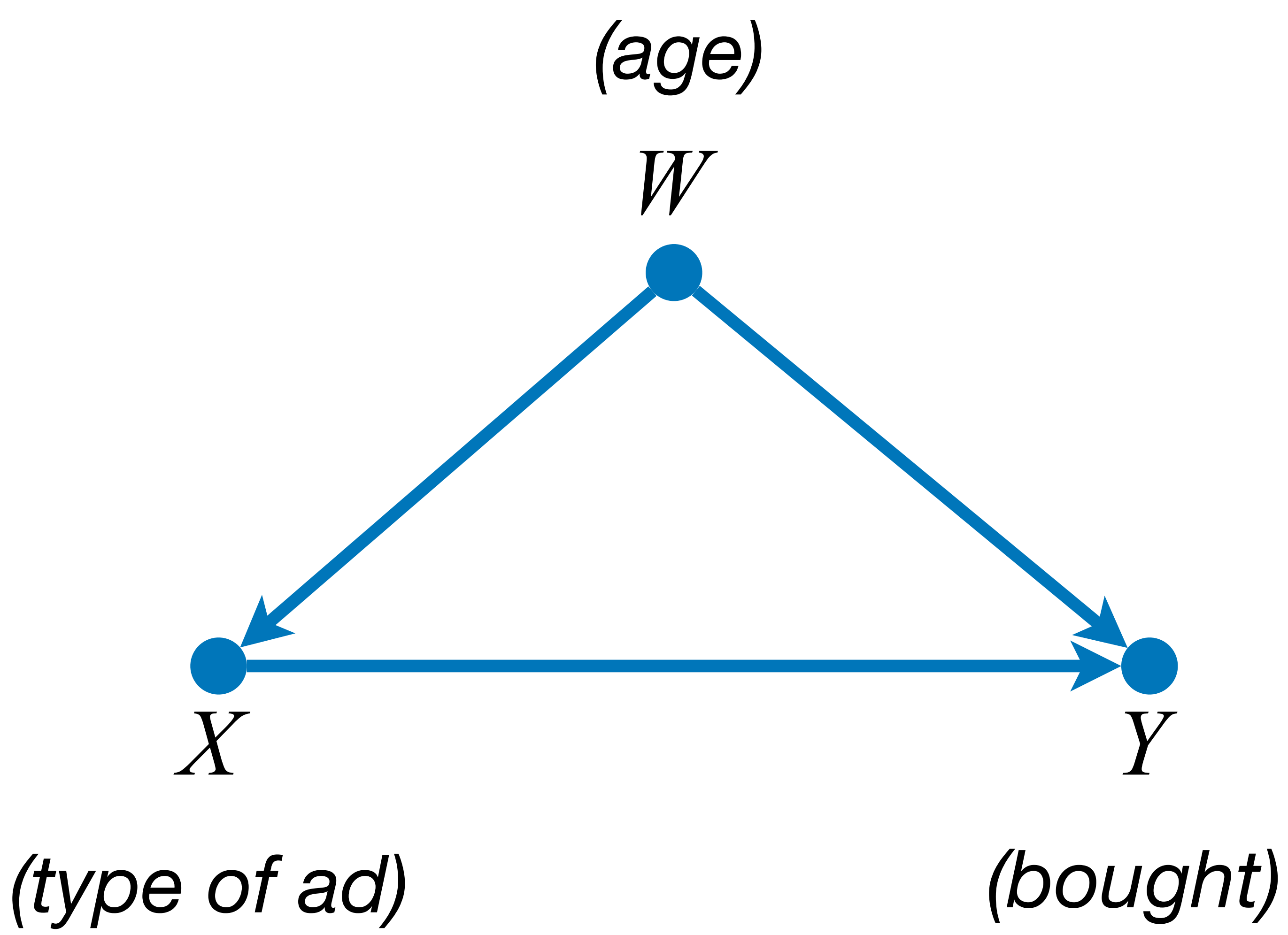
New Website (Π^*)
(target environment)



We use ■ to represent differences in mechanism or distribution

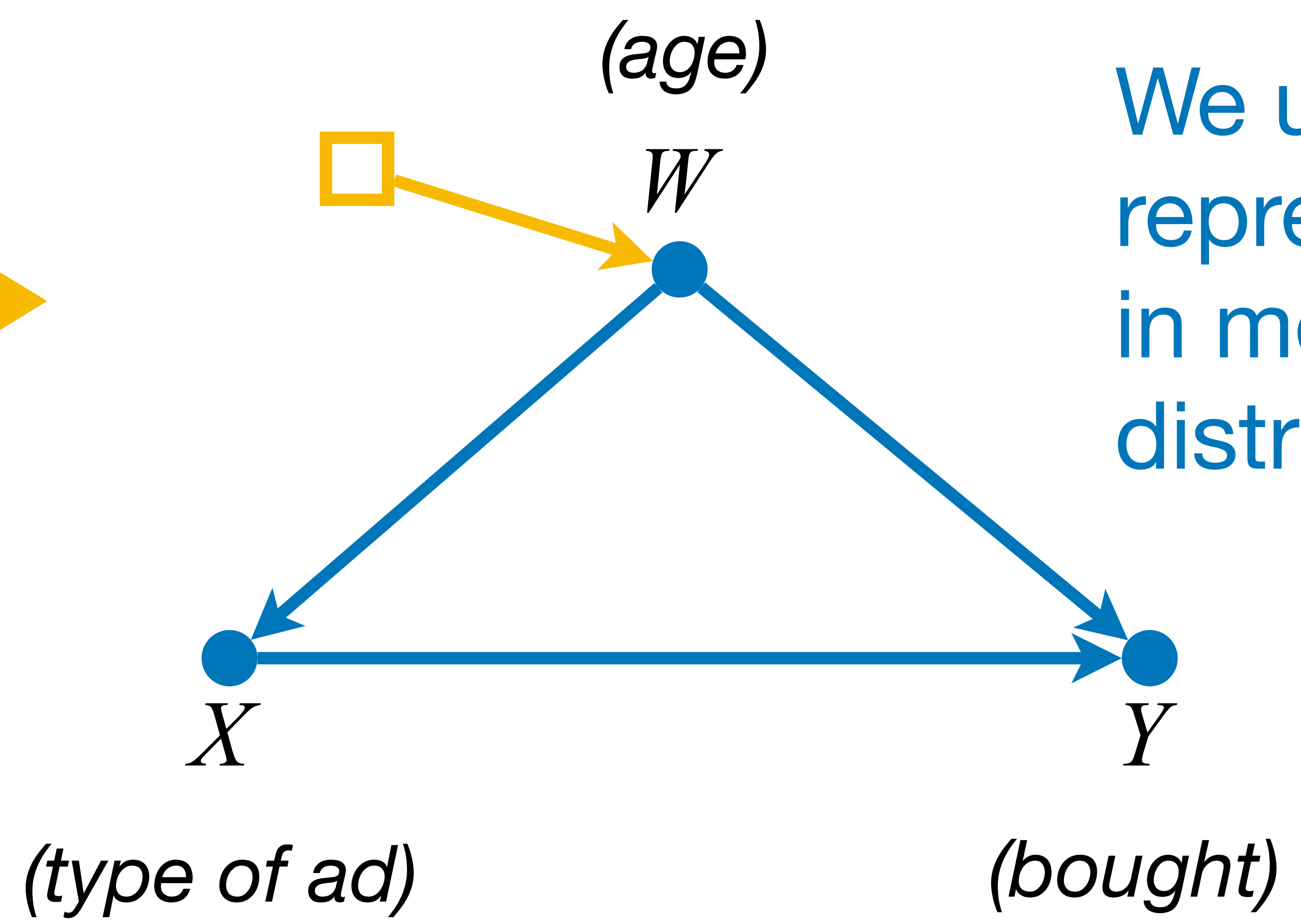
Statistical Transportability

Current Website (Π)
(training environment)



Generalization

New Website (Π^*)
(target environment)



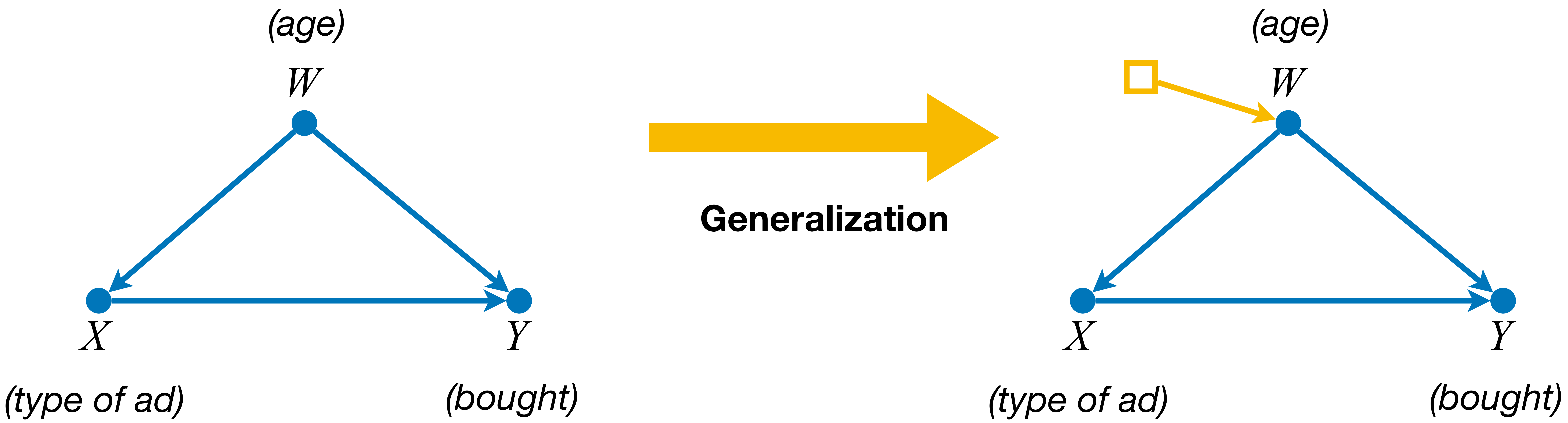
We use ■ to represent differences in mechanism or distribution

$P(W) \neq P^*(W)$ hence $P(y \mid x) \neq P^*(y \mid x)$

Statistical Transportability

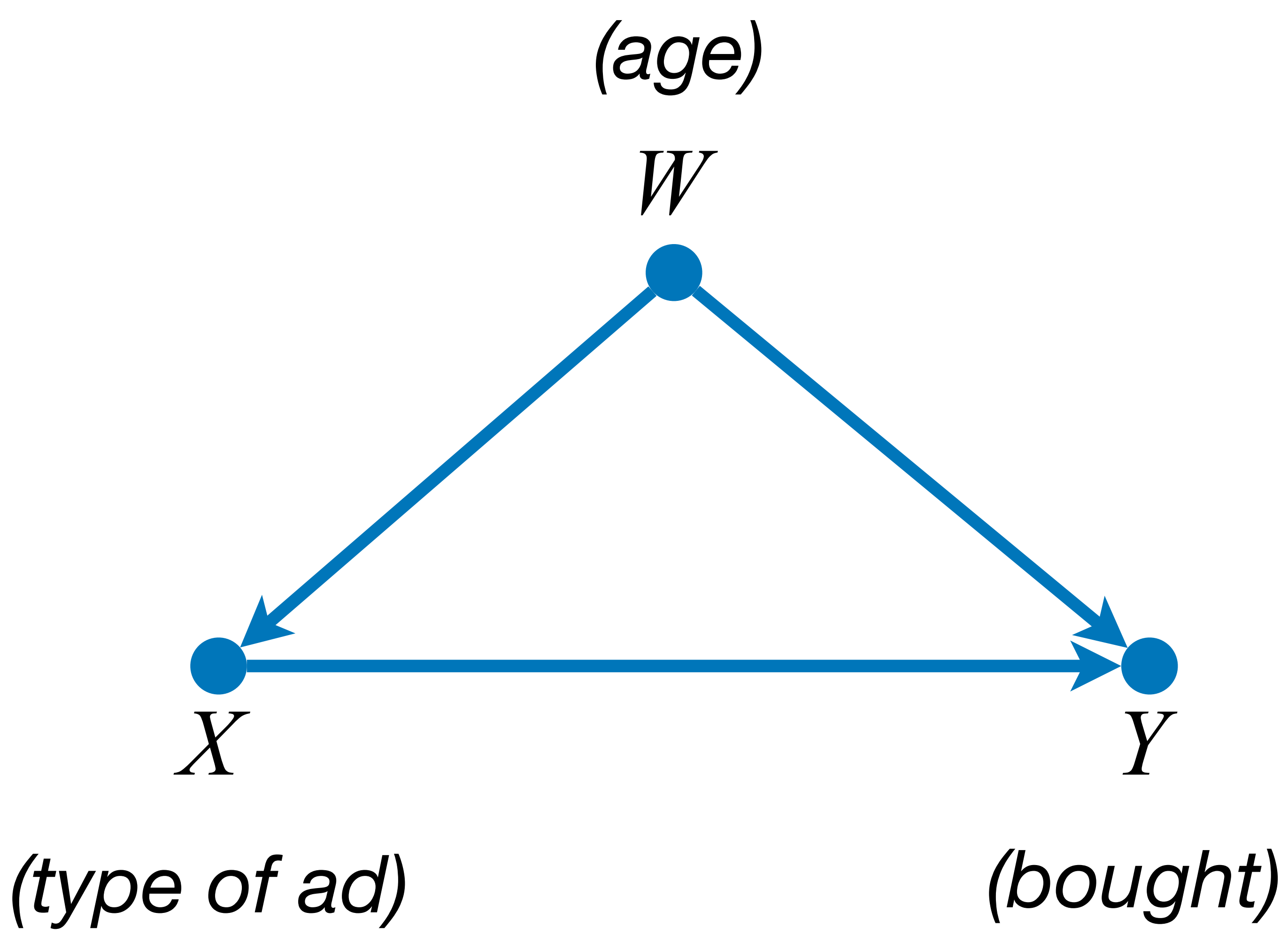
Current Website (Π)
(training environment)

New Website (Π^*)
(target environment)

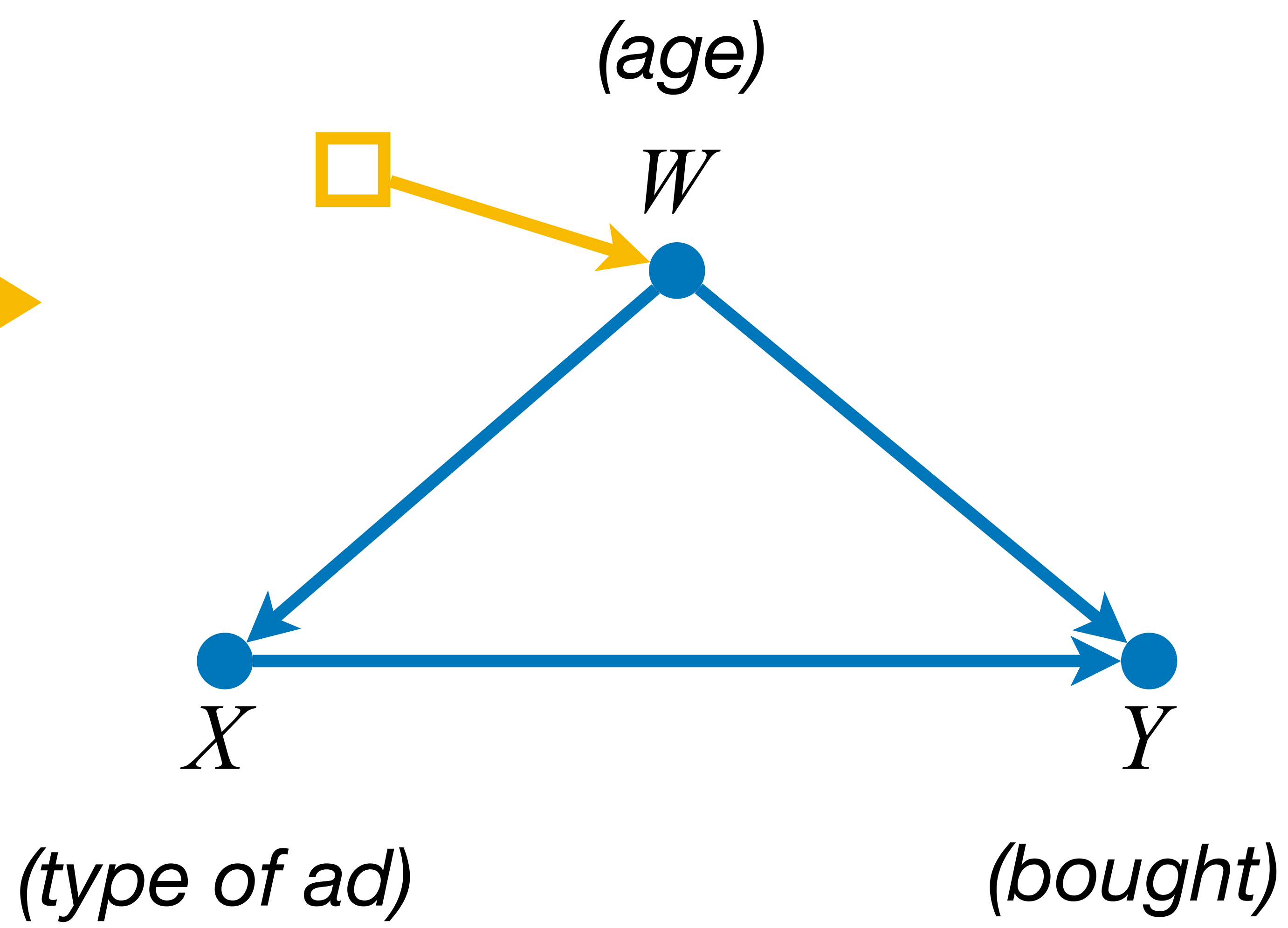


Statistical Transportability

Current Website (Π)
(training environment)



New Website (Π^*)
(target environment)

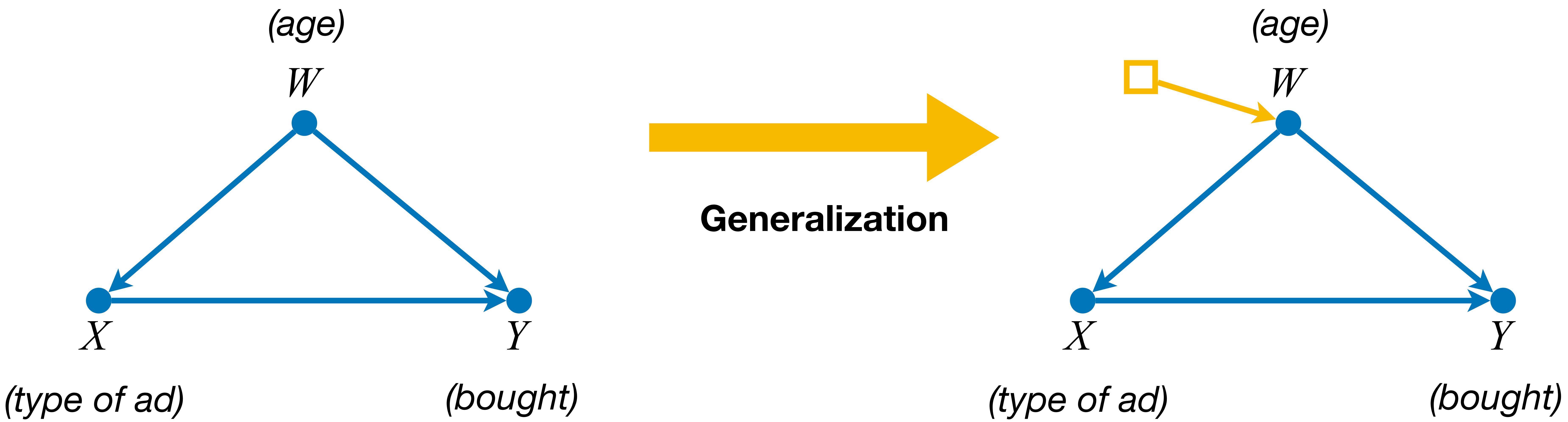


- How to generalize the model learned in the source environment to different (but related) target environments?

Statistical Transportability

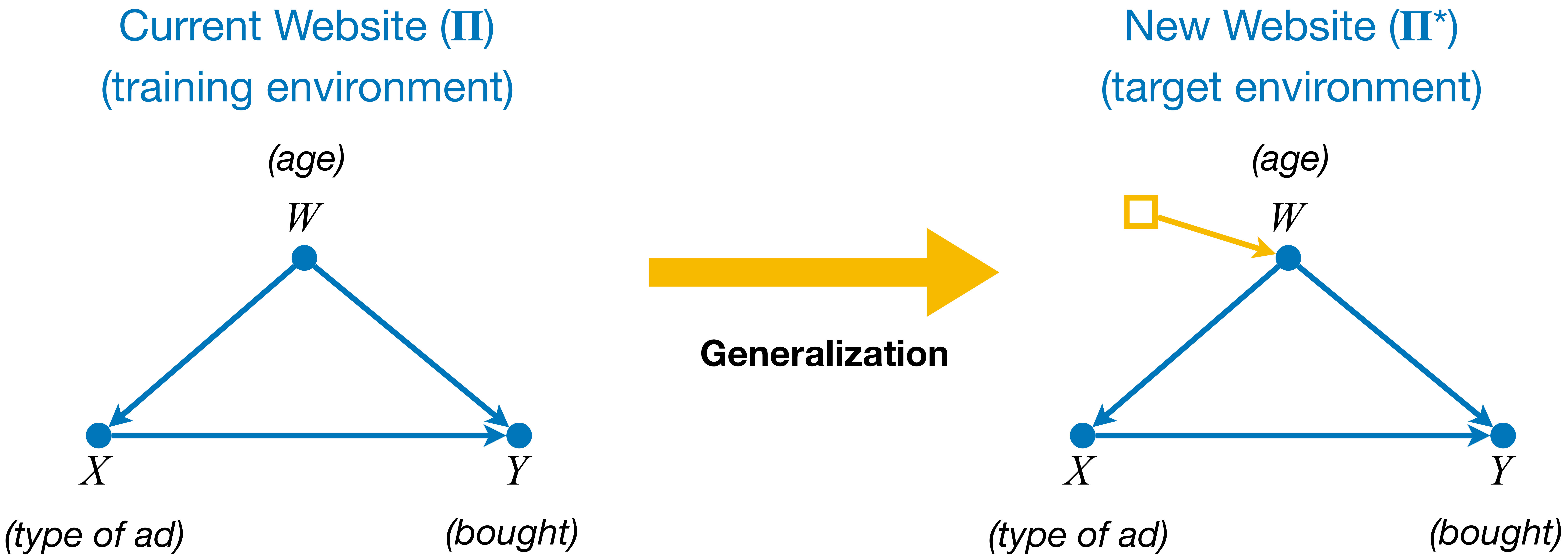
Current Website ($\boldsymbol{\Pi}$)
(training environment)

New Website ($\boldsymbol{\Pi}^*$)
(target environment)



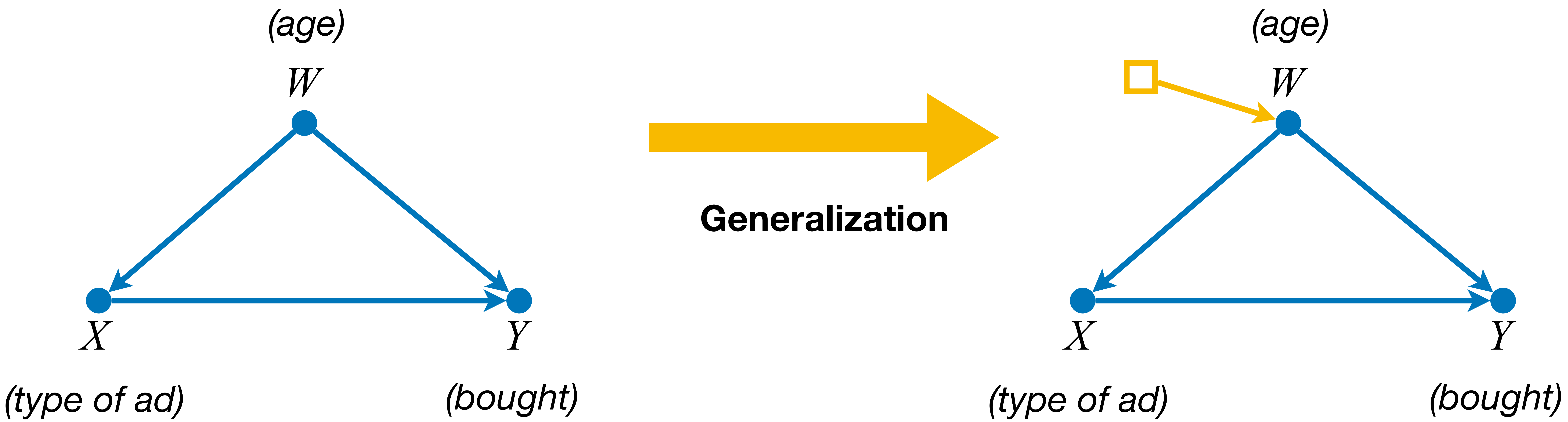
- How to generalize the model learned in the source environment to different (but related) target environments?
- Do we need to obtain samples from $\boldsymbol{\Pi}^*$ and train a new model?

Statistical Transportability



Statistical Transportability

Current Website (**Π**) (training environment) | New Website (**Π***) (target environment)

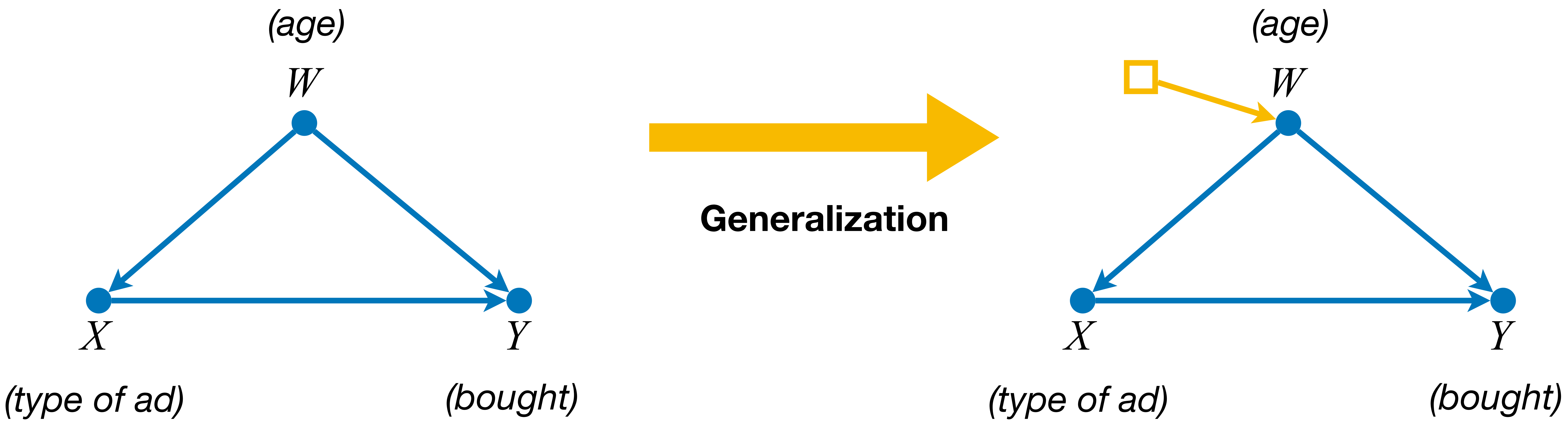


We observe $P(x,y,w)$

Statistical Transportability

Current Website (Π)
(training environment)

New Website (Π^*)
(target environment)



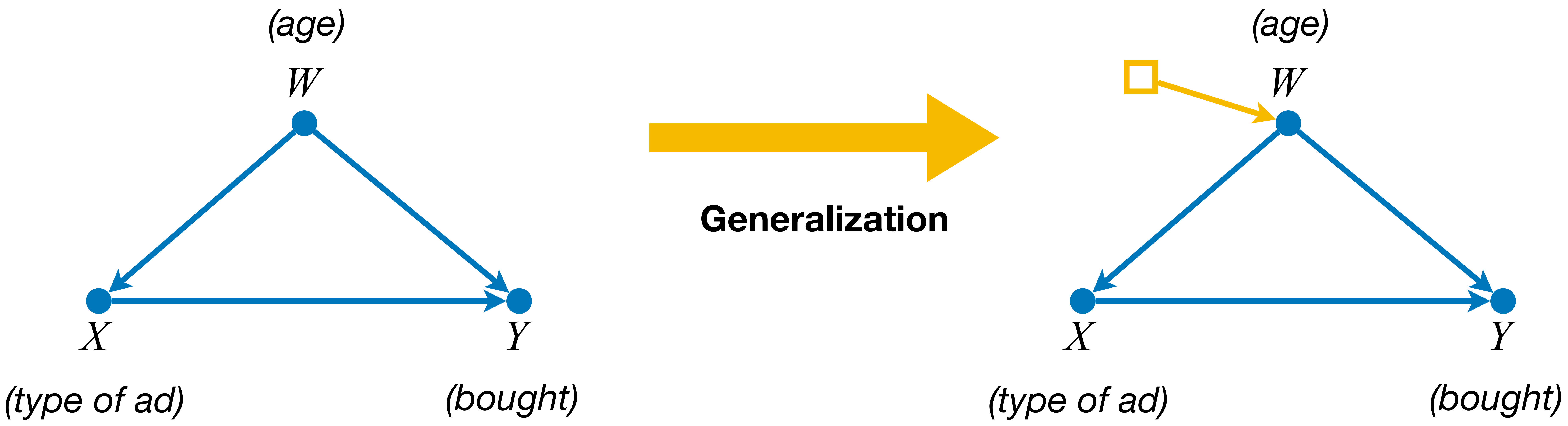
We observe $P(x,y,w)$

We want to say something about $P^*(y|x)$

Statistical Transportability

Current Website (Π)
(training environment)

New Website (Π^*)
(target environment)



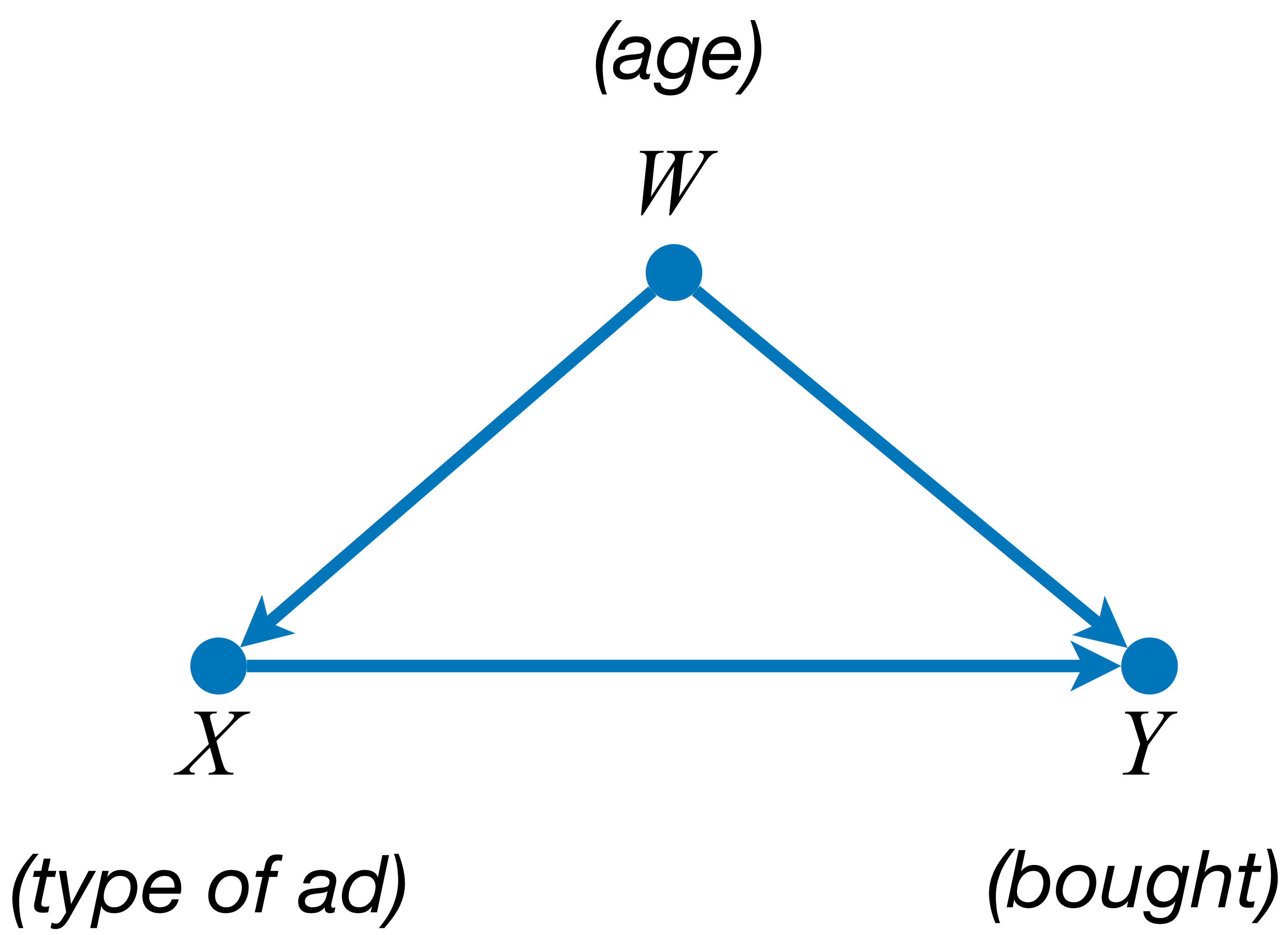
We observe $P(x,y,w)$

$$P(x,y,w)=P(w)\,P(x|w)\,P(y|x,w)$$

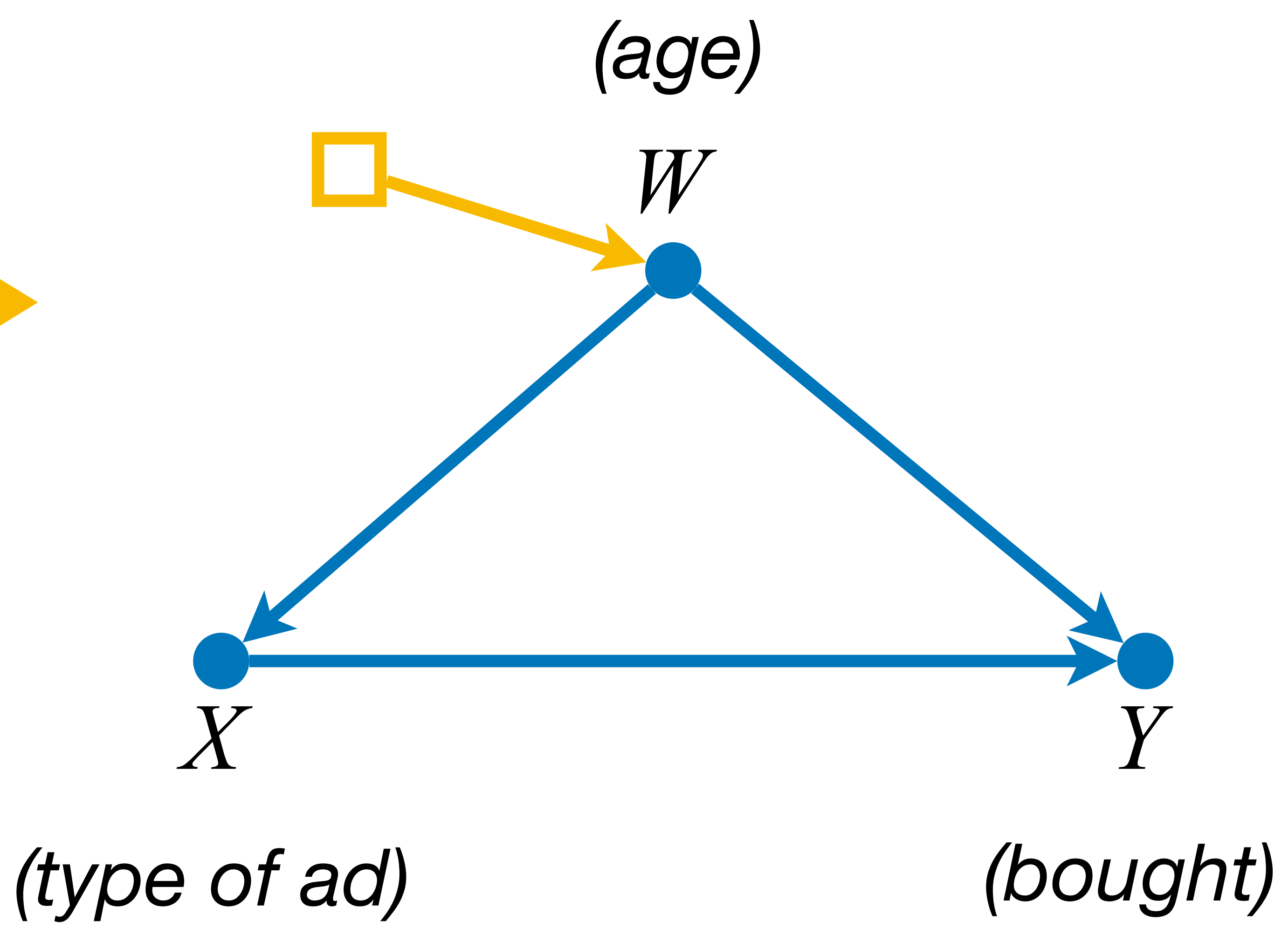
We want to say something about $P^*(y|x)$

Statistical Transportability

Current Website (**Π**)
(training environment)



New Website (**Π***)
(target environment)



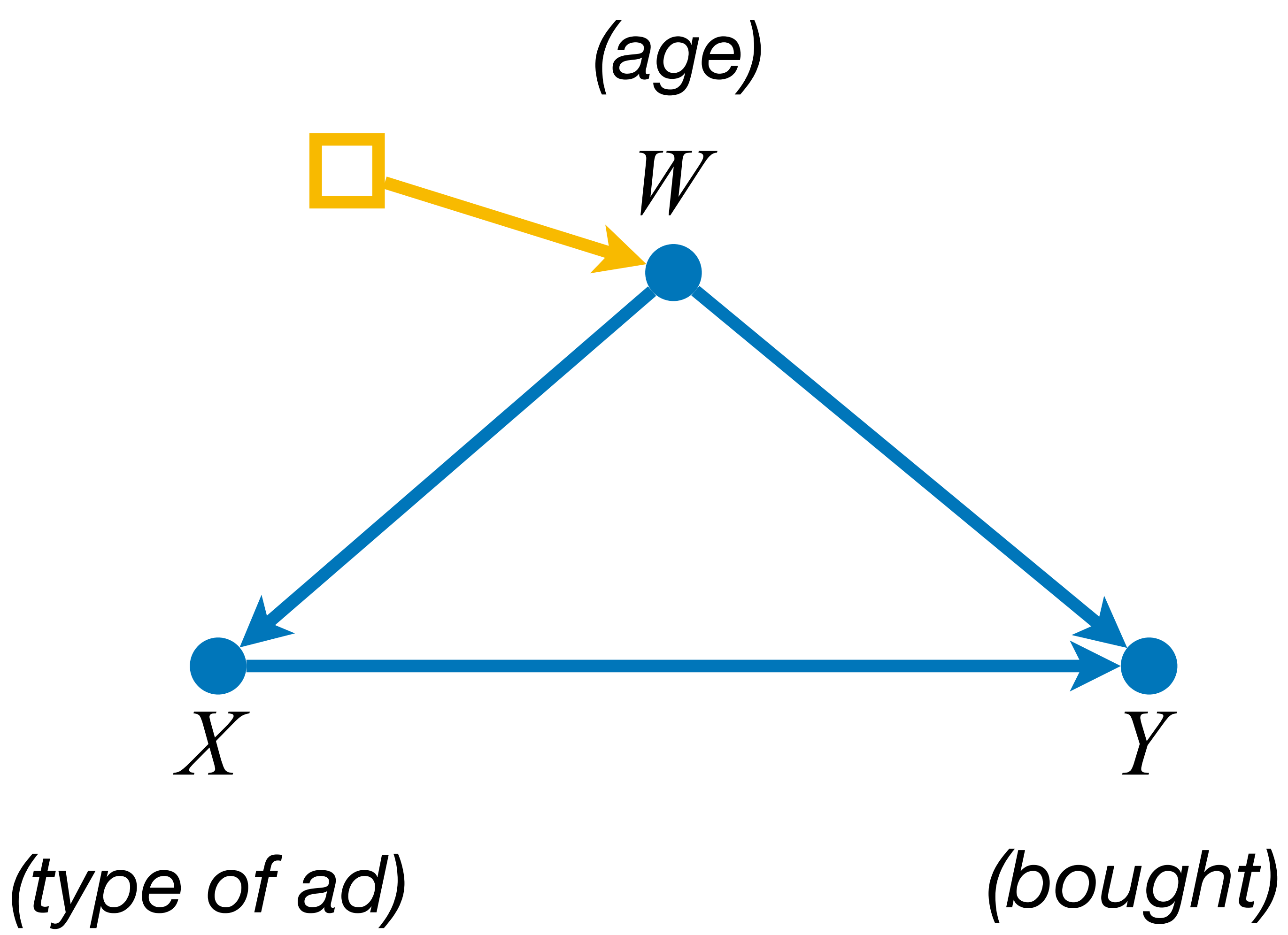
We observe $P(x,y,w)$

We want to say something about $P^*(y|x)$

$P(x,y,w)=P(w)\,P(x|w)\,P(y|x,w)$ **are the same in both environments, which is implied by this causal model.**

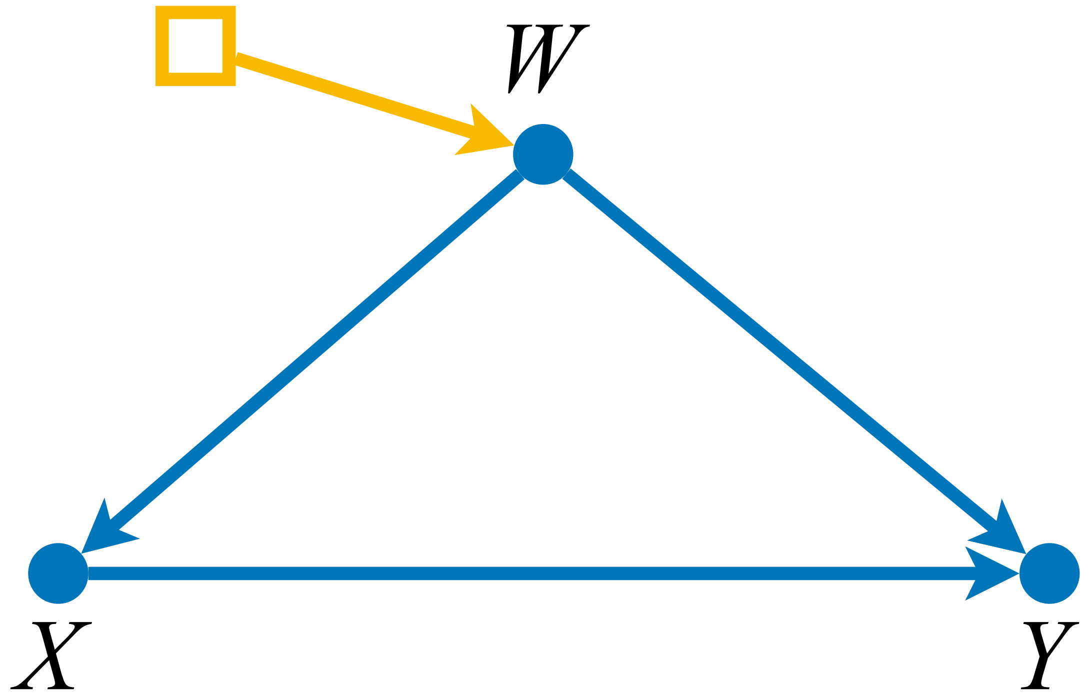
Statistical Transportability

New Website (Π^*)
(target environment)



Statistical Transportability

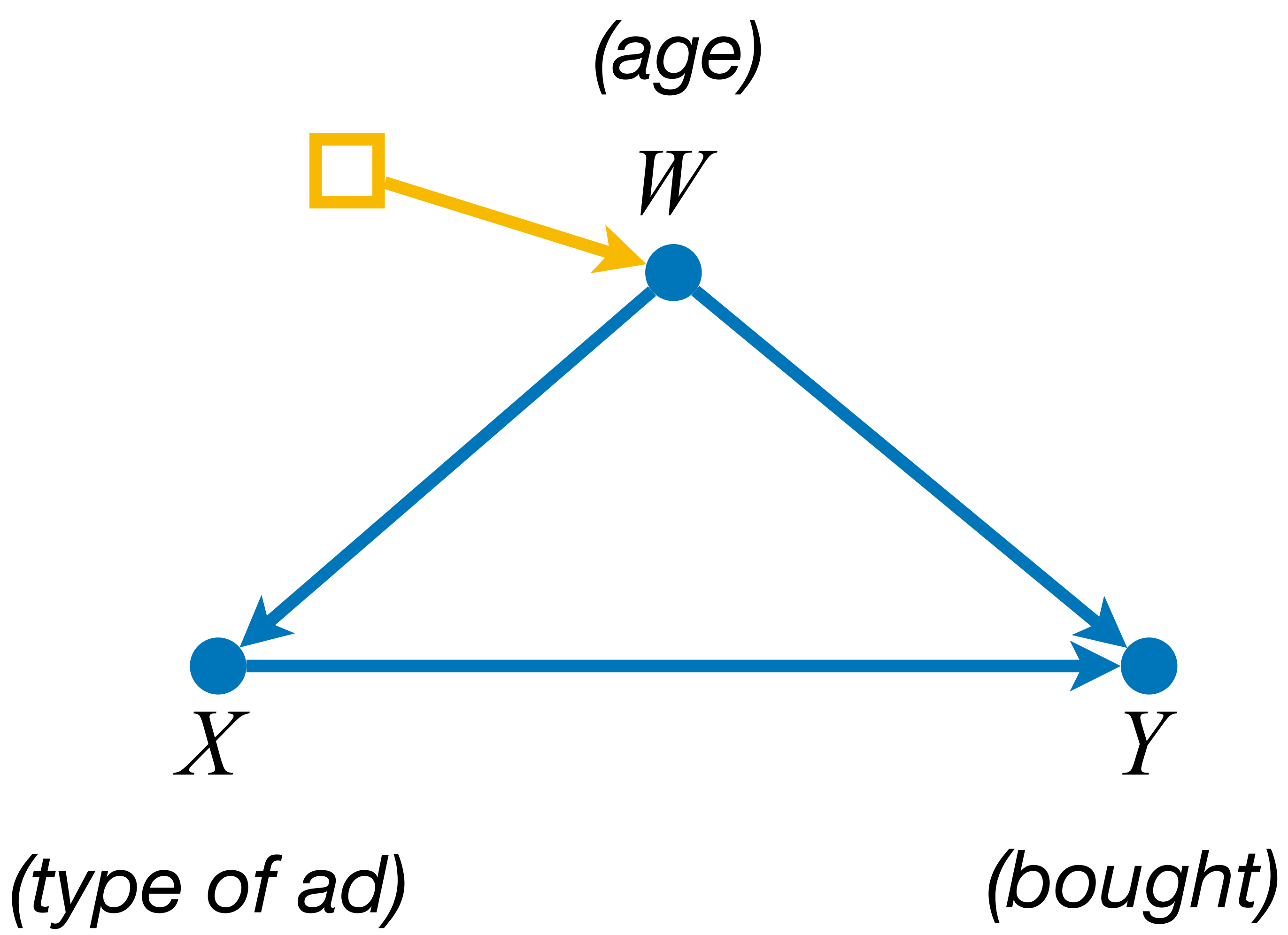
New Website ($\mathbf{\Pi}^*$)
(target environment)



- The target distribution $P^*(y|x)$ can be expressed as:

Statistical Transportability

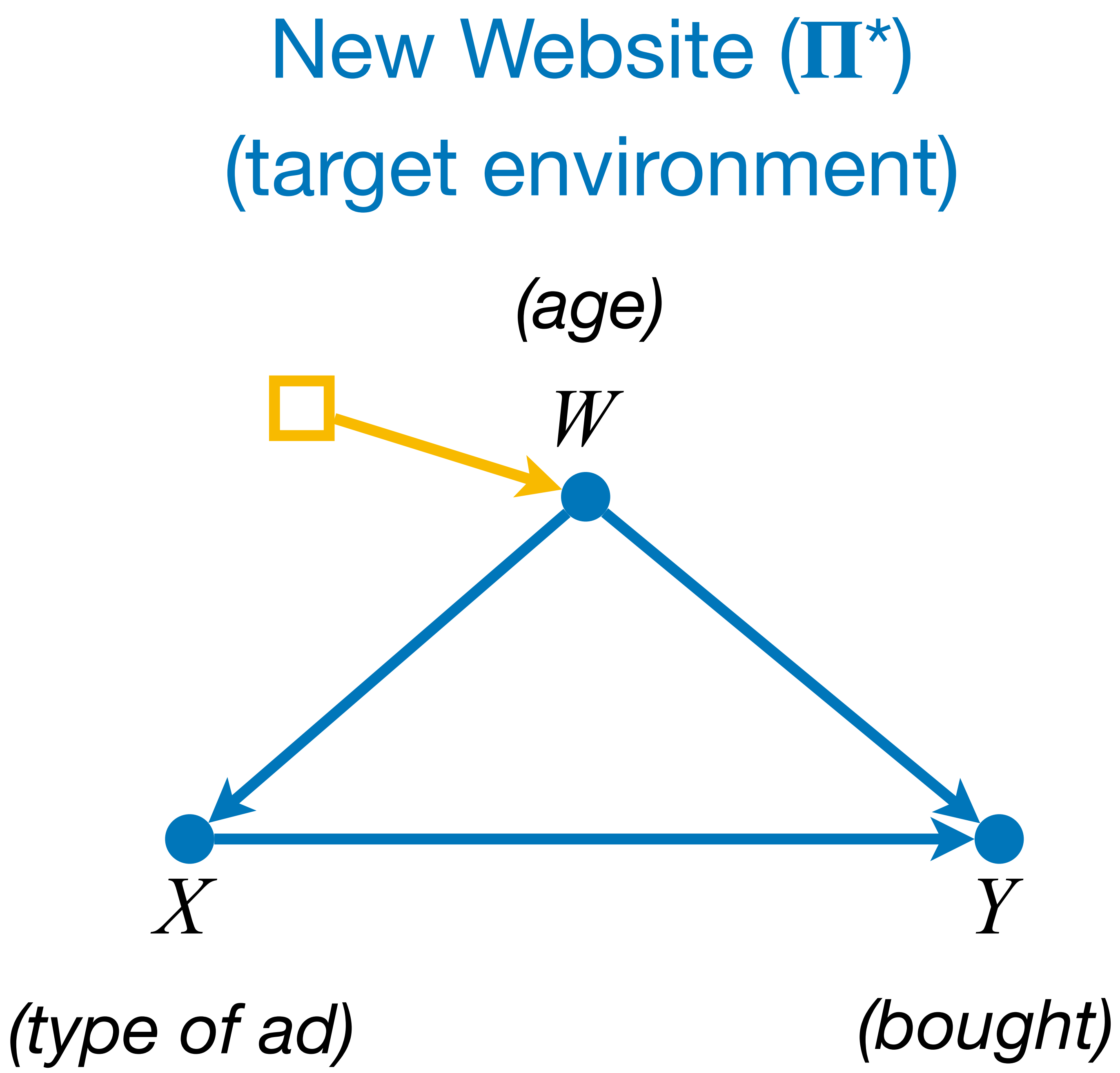
New Website (Π^*)
(target environment)



- The target distribution $P^*(y|x)$ can be expressed as:

$$P^*(y\,|\,x) = \frac{P^*(y, x)}{P^*(x)} = \frac{\sum_w P^*(y\,|\,x, w)P^*(x\,|\,w)P^*(w)}{\sum_w P^*(x\,|\,w)P^*(w)}$$

Statistical Transportability



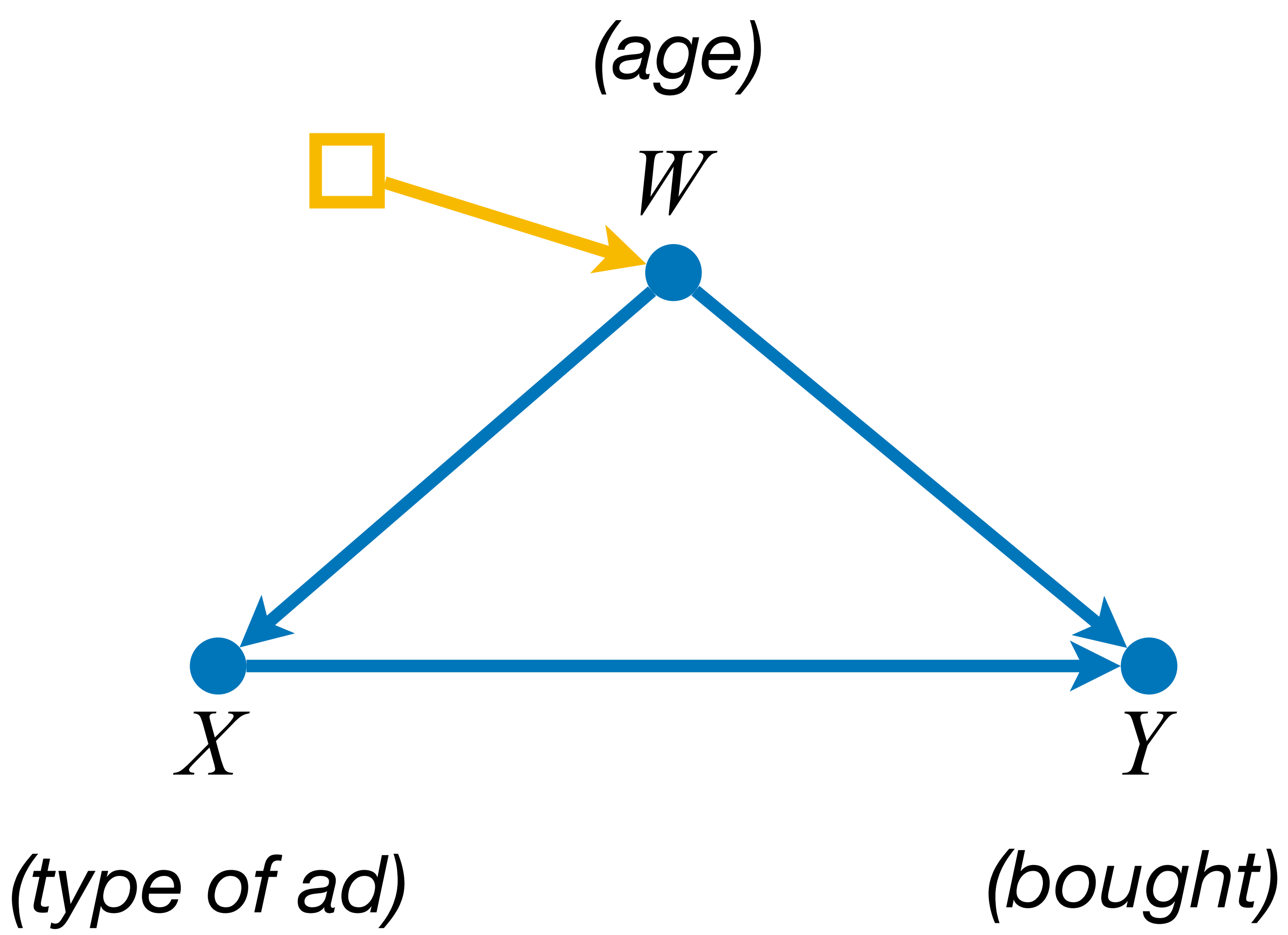
- The target distribution $P^*(y|x)$ can be expressed as:

are the same in source and target

$$P^*(y \mid x) = \frac{P^*(y,x)}{P^*(x)} = \frac{\sum_w P^*(y \mid x, w) P^*(x \mid w) P^*(w)}{\sum_w P^*(x \mid w) P^*(w)}$$

Statistical Transportability

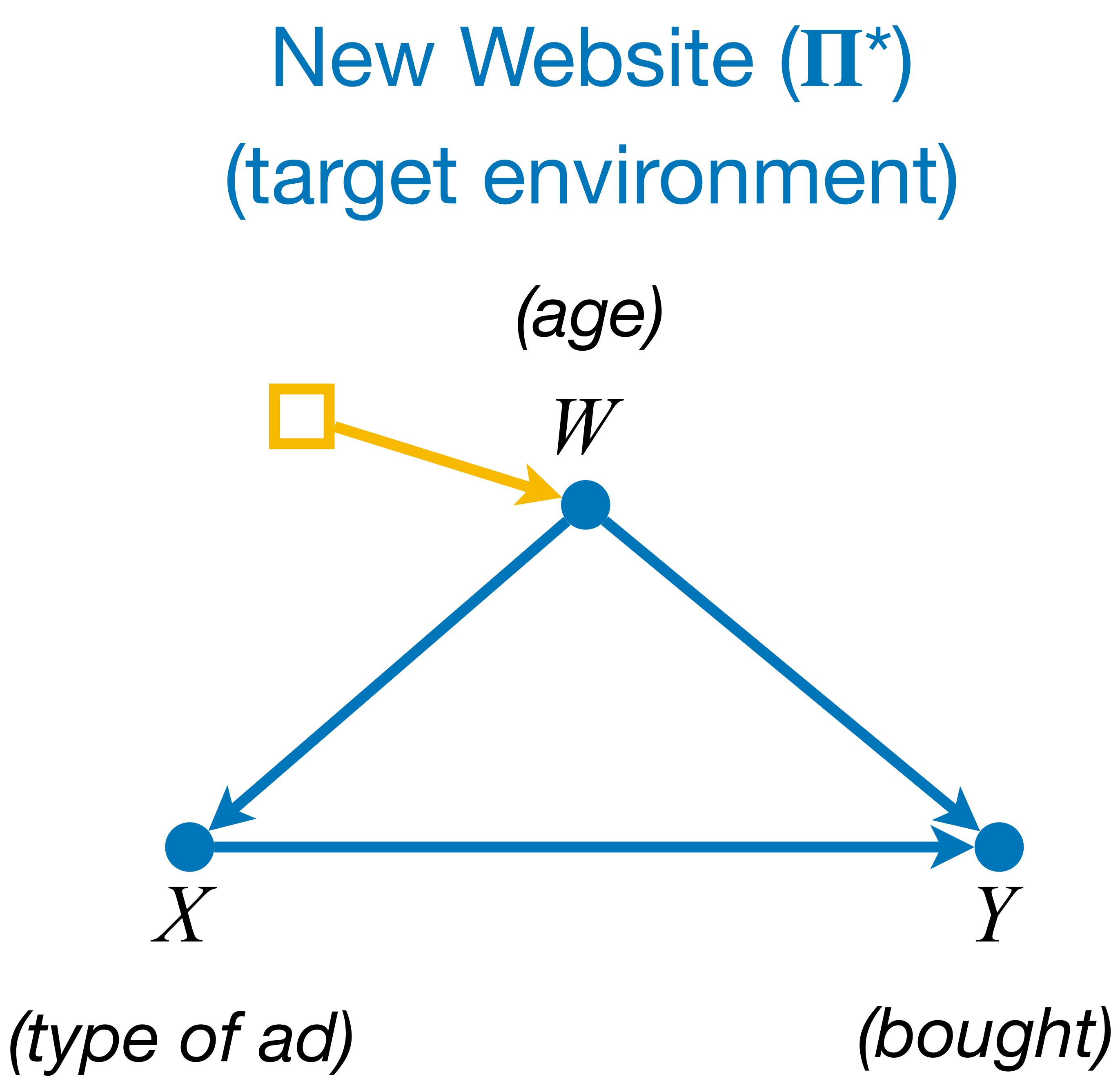
New Website (Π^*)
(target environment)



- The target distribution $P^*(y|x)$ can be expressed as:

$$P^*(y|x) = \frac{P^*(y,x)}{P^*(x)} = \frac{\sum_w P^*(y|x,w) P^*(x|w) P^*(w)}{\sum_w P^*(x|w) P^*(w)}$$

Statistical Transportability

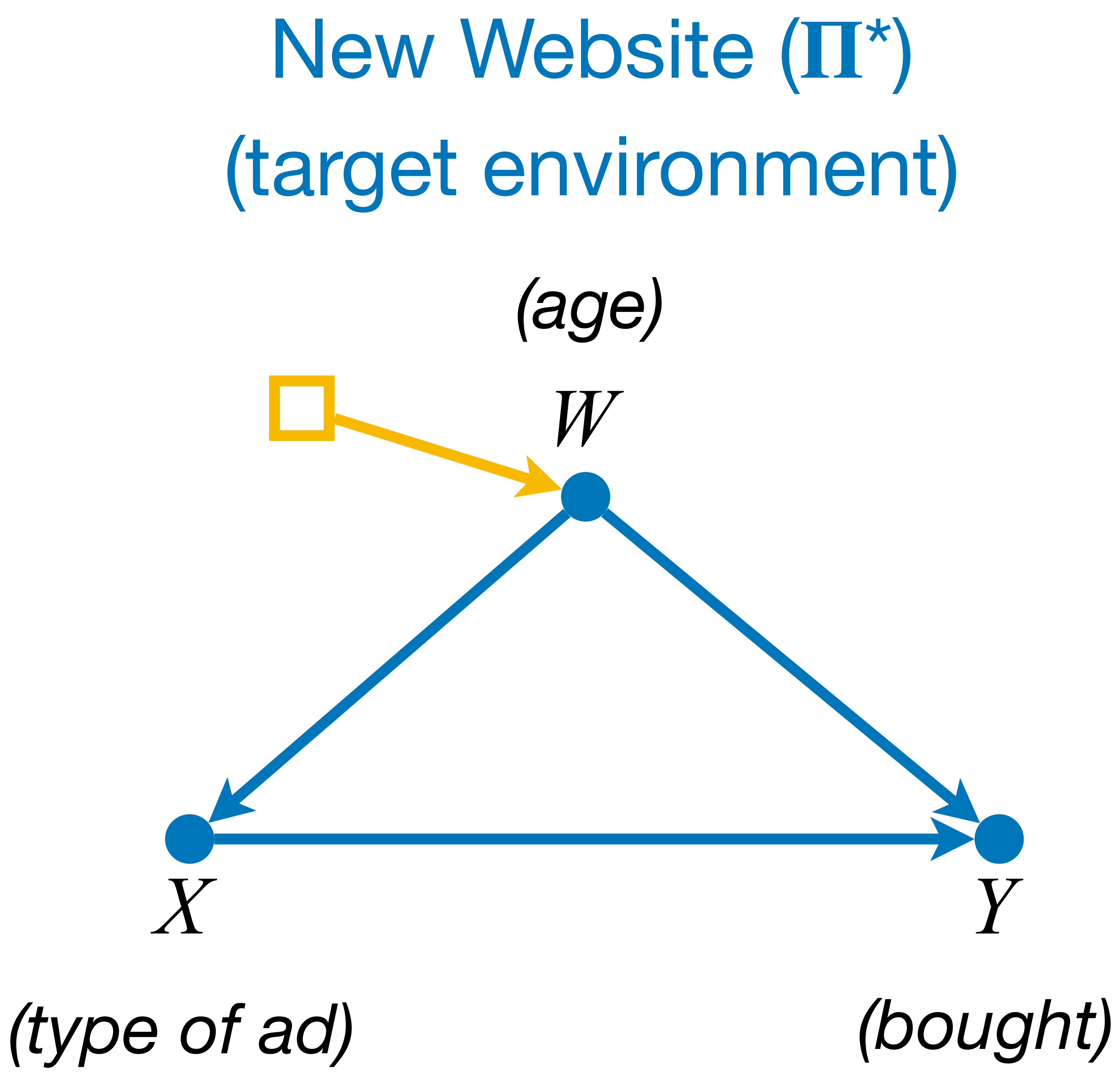


- The target distribution $P^*(y|x)$ can be expressed as:

$$P^*(y|x) = \frac{P^*(y,x)}{P^*(x)} = \frac{\sum_w P^*(y|x,w)P^*(x|w)P^*(w)}{\sum_w P^*(x|w)P^*(w)}$$

$$= \frac{\sum_w P(y|x,w)P(x|w)P^*(w)}{\sum_w P(x|w)P^*(w)}$$

Statistical Transportability



- The target distribution $P^*(y|x)$ can be expressed as:

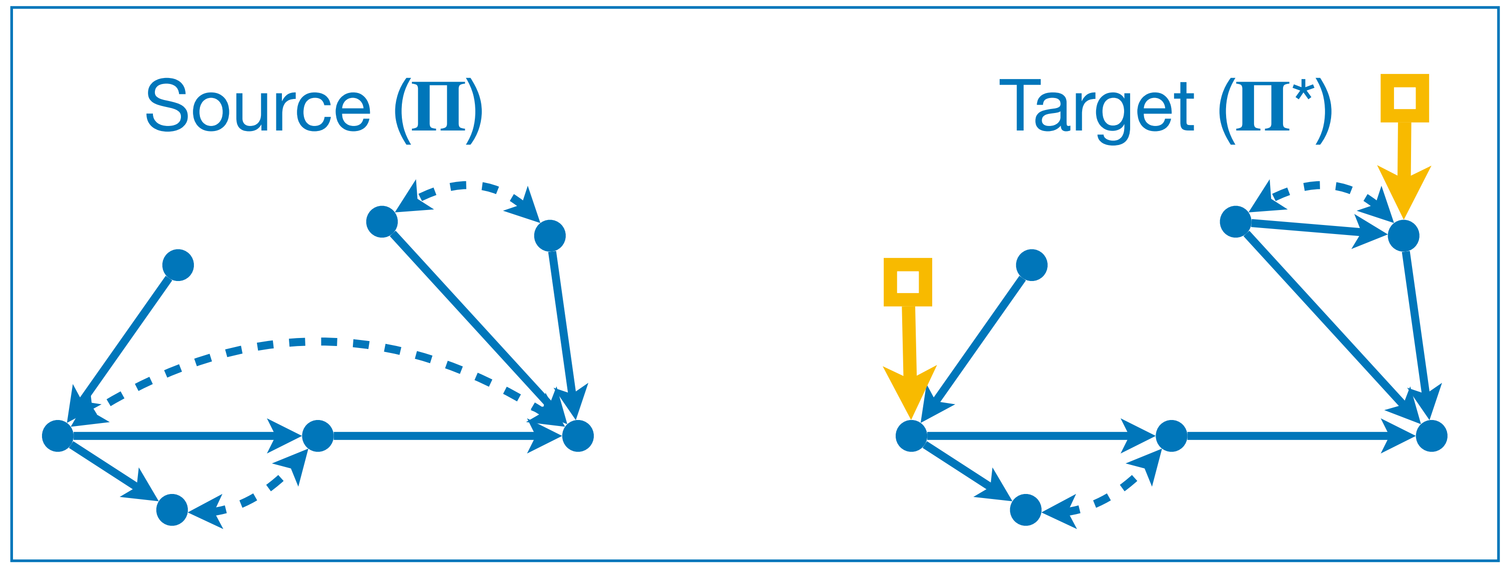
$$P^*(y|x) = \frac{P^*(y,x)}{P^*(x)} = \frac{\sum_w P^*(y|x,w)P^*(x|w)P^*(w)}{\sum_w P^*(x|w)P^*(w)}$$

$$= \frac{\sum_w P(y|x,w)P(x|w)P^*(w)}{\sum_w P(x|w)P^*(w)}$$

- Under the assumptions implied by the diagram, only $P^*(w)$ needs to be measured in the target environment, while the other distributions can be reused from the data collected in the source environment.

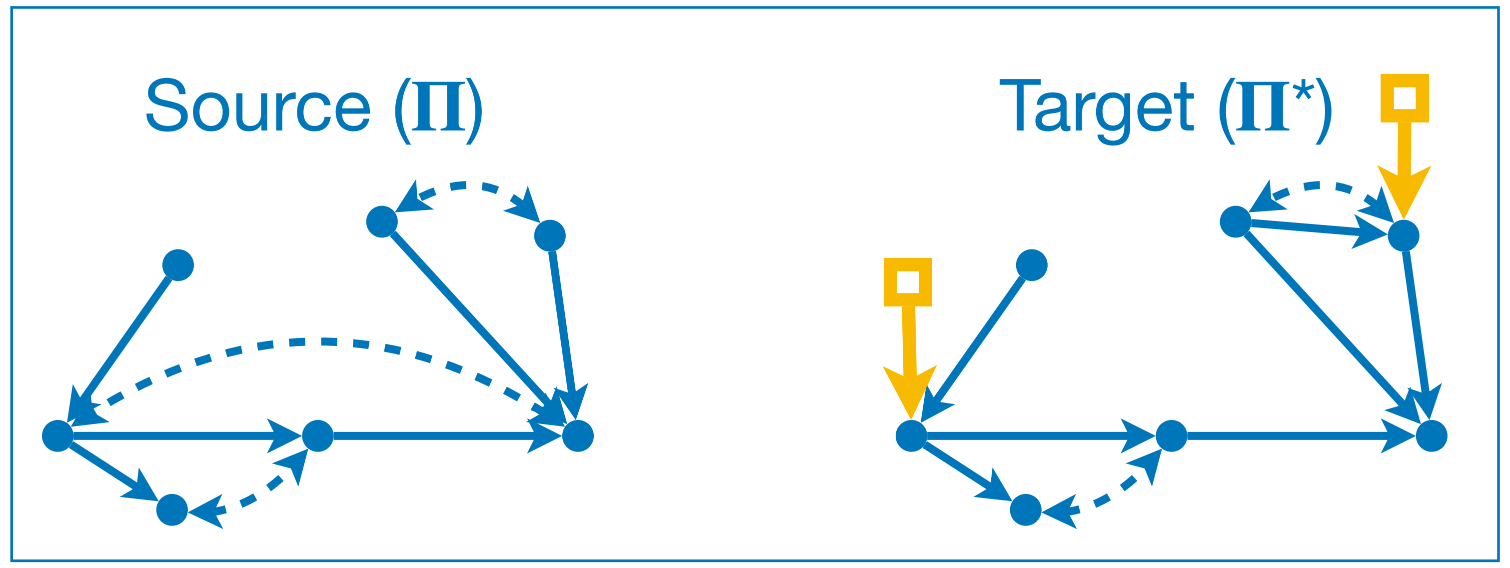
Deciding Transportability

Deciding Transportability



Selection Diagram D

Deciding Transportability

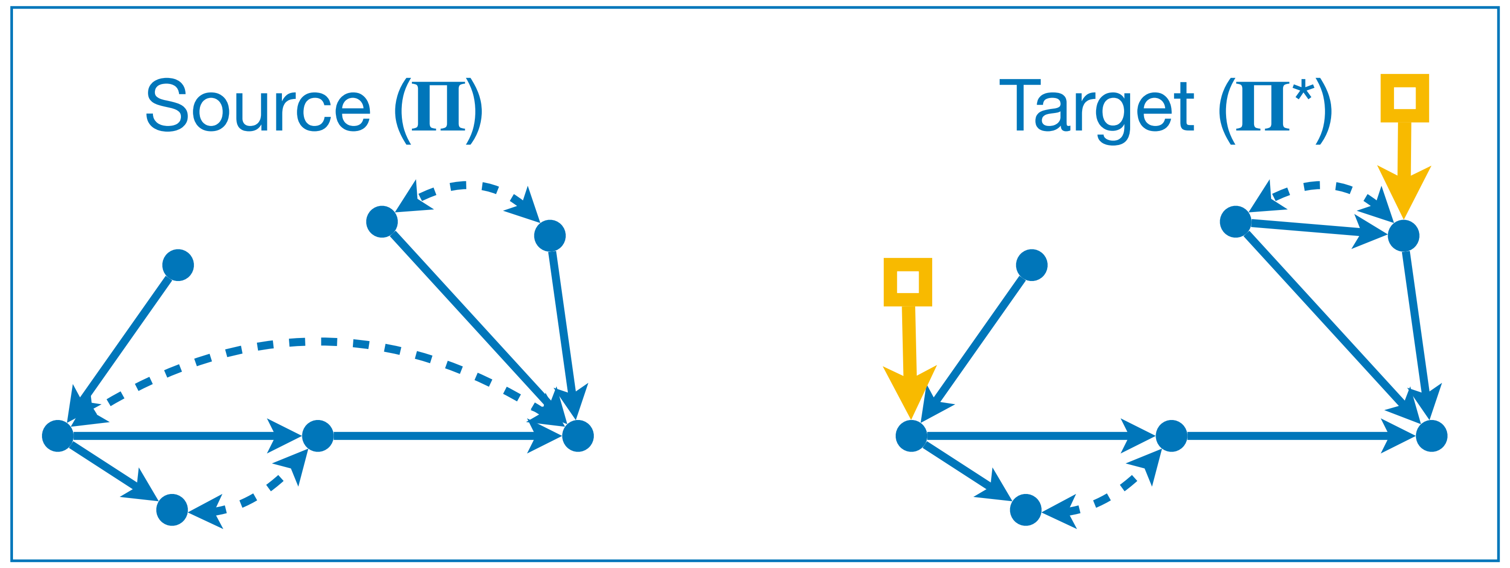


Selection Diagram D

$P(\mathbf{v})$			

Distribution learned from $\boldsymbol{\pi}$

Deciding Transportability



Selection Diagram D

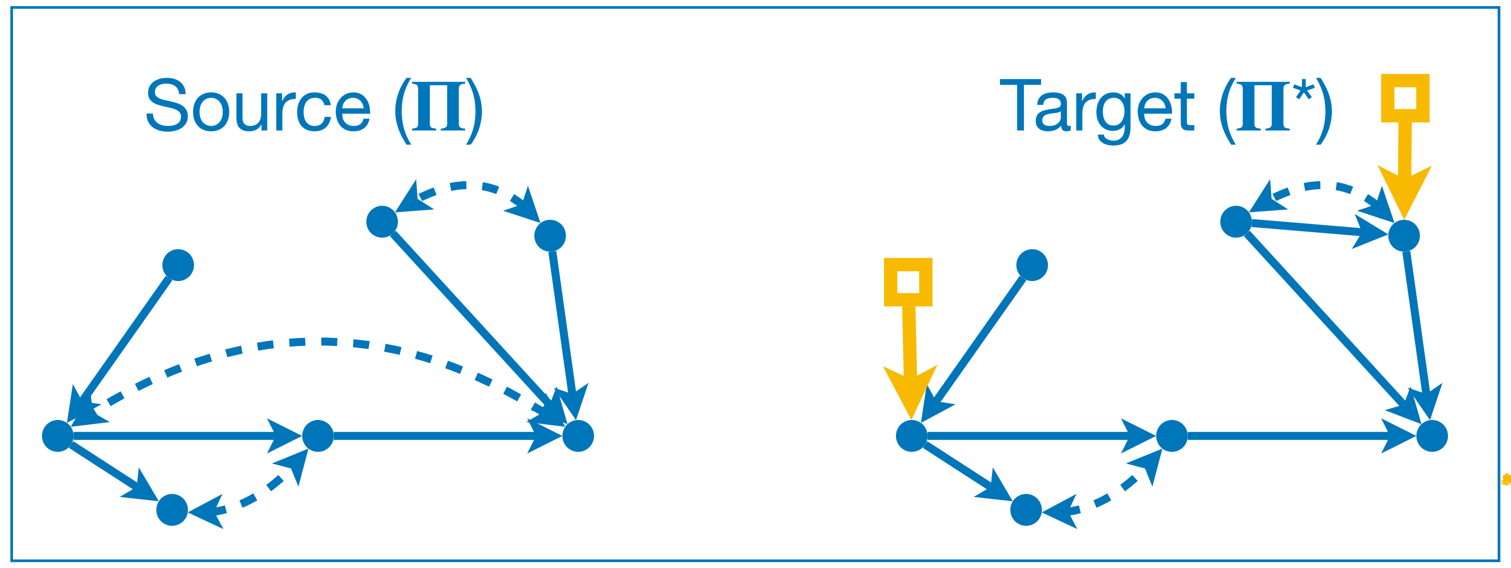
$P(\mathbf{v})$			

Distribution learned from $\boldsymbol{\pi}$

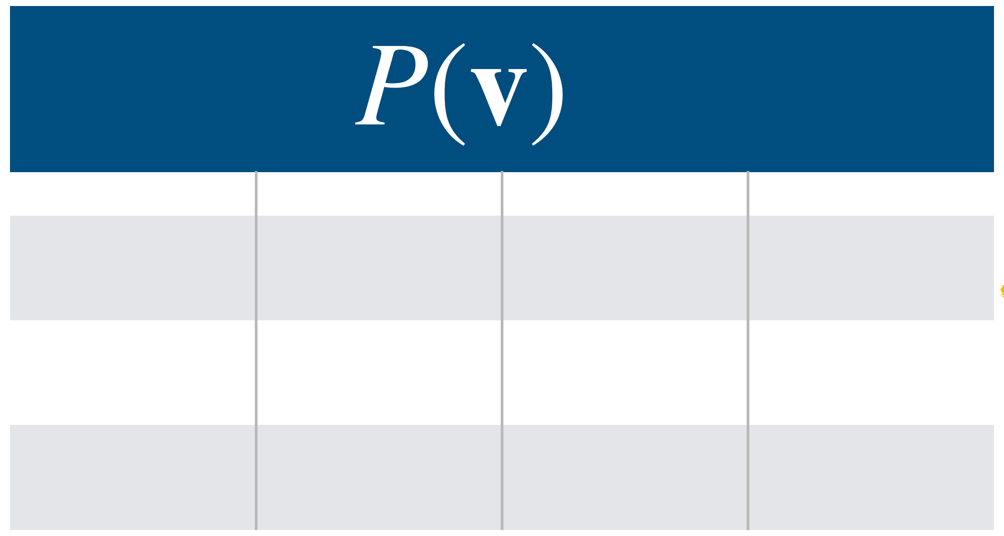
$P^*(\mathbf{w})$		

Partial distribution from $\boldsymbol{\pi}^*$

Deciding Transportability



Selection Diagram D

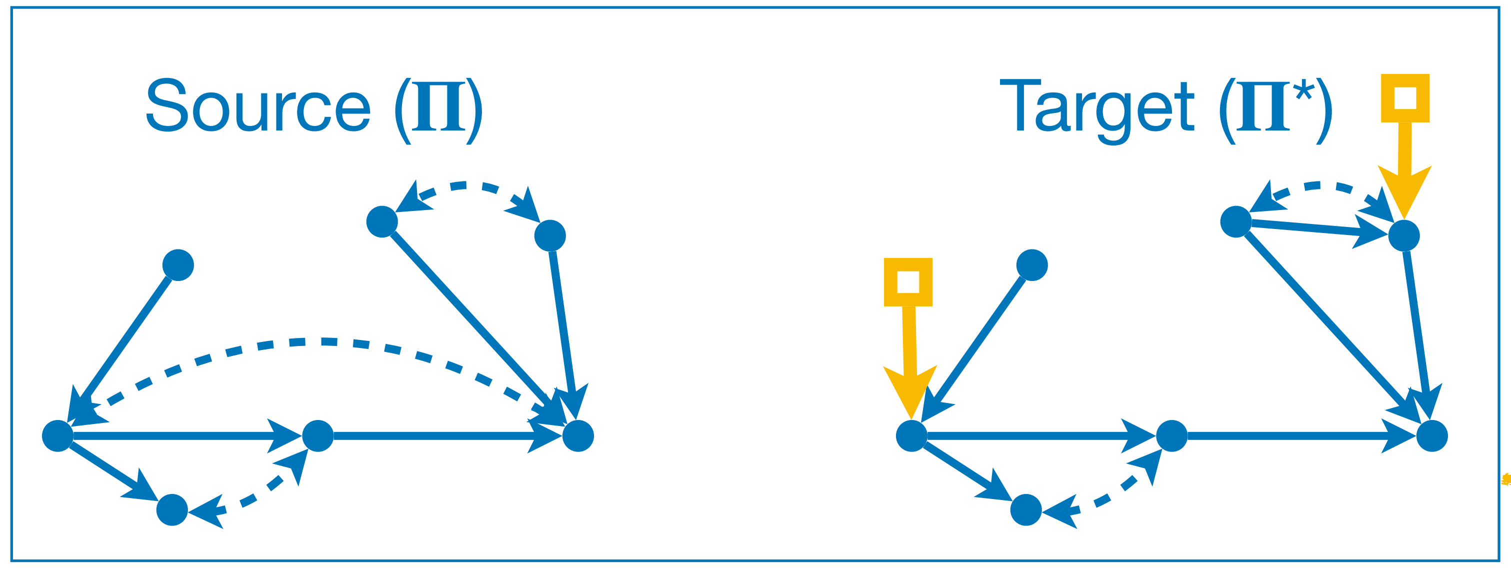


Distribution learned from $\boldsymbol{\pi}$

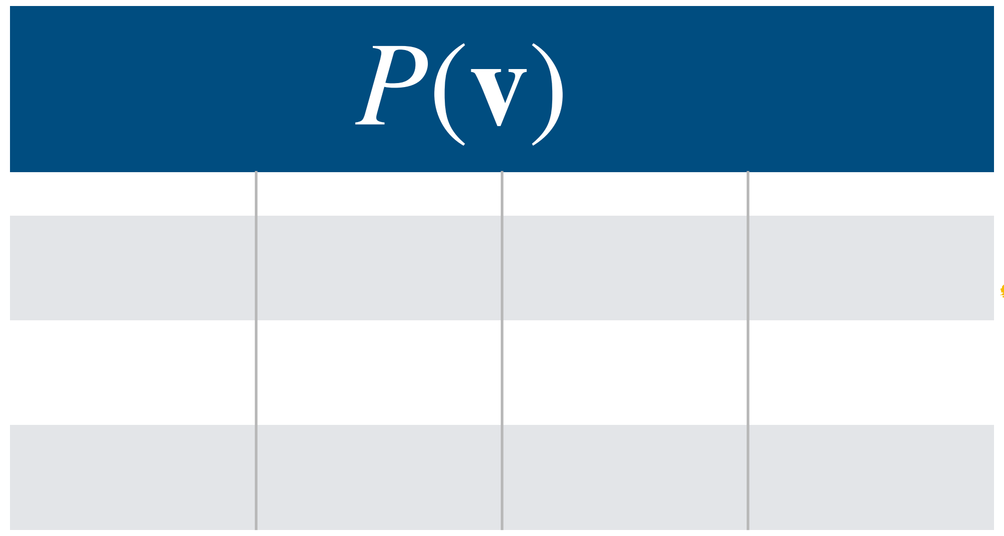
Partial distribution from $\boldsymbol{\pi}^*$

Is there a function f such that $P^*(y \mid x) = f(P(\mathbf{v}), P^*(\mathbf{w}))$?

Deciding Transportability



Selection Diagram D



Distribution learned from $\boldsymbol{\pi}$

Partial distribution from $\boldsymbol{\pi}^*$

Is there a function f such that $P^*(y \mid x) = f(P(\mathbf{v}), P^*(\mathbf{w}))$?

→ **yes (f) / no**

Proposed Strategy

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments.

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments. → Selection diagrams (with □)

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments. → Selection diagrams (with □)

2. Identify the stable mechanisms across environments.

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments. → Selection diagrams (with □)

2. Identify the stable mechanisms across environments.

3. Determine the variables that need to be re-measured.

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments. → Selection diagrams (with □)
2. Identify the stable mechanisms across environments.
3. Determine the variables that need to be re-measured.
4. Construct an estimator from the available data.

Proposed Strategy

1. Encode the assumptions about the differences and commonalities across environments. → Selection diagrams (with □)
2. Identify the stable mechanisms across environments.
3. Determine the variables that need to be re-measured.
4. Construct an estimator from the available data.

(Steps 2–4:) Exploit Causality Theory

Results

Results

1 We introduce a novel graphical decomposition of the observed/learned distribution into factors that take into account the latent structure, which generalizes C-components (Tian & Pearl 2002), and is suitable to reason about distributions with different sets of measured variables.

Results

1. We introduce a novel graphical decomposition of the observed/learned distribution into factors that take into account the latent structure, which generalizes C-components (Tian & Pearl 2002), and is suitable to reason about distributions with different sets of measured variables.

2. We derive a complete algorithm that determines if a distribution $P^*(\boldsymbol{y}|\boldsymbol{x})$ can be uniquely identified from distributions $P(\boldsymbol{v})$ and $P^*(\boldsymbol{w})$ ($\boldsymbol{W} \subseteq \boldsymbol{V}$) based on the assumptions encoded in graphs corresponding to the source and target domains.

Results

1. We introduce a novel graphical decomposition of the observed/learned distribution into factors that take into account the latent structure, which generalizes C-components (Tian & Pearl 2002), and is suitable to reason about distributions with different sets of measured variables.

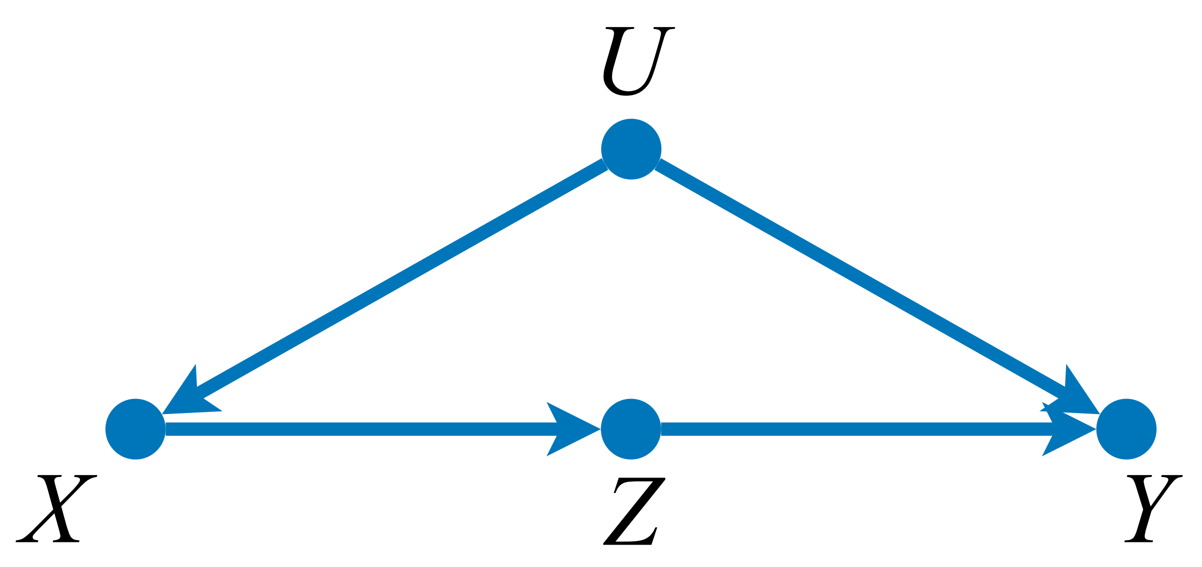
2. We derive a complete algorithm that determines if a distribution $P^*(\mathbf{y}|\mathbf{x})$ can be uniquely identified from distributions $P(\mathbf{v})$ and $P^*(\mathbf{w})$ ($\mathbf{W} \subseteq \mathbf{V}$) based on the assumptions encoded in graphs corresponding to the source and target domains.

3. We connect this problem with the problem of identifying the effect of stochastic plans and how it reduces to the former problem.

Factorization of Observed Distributions

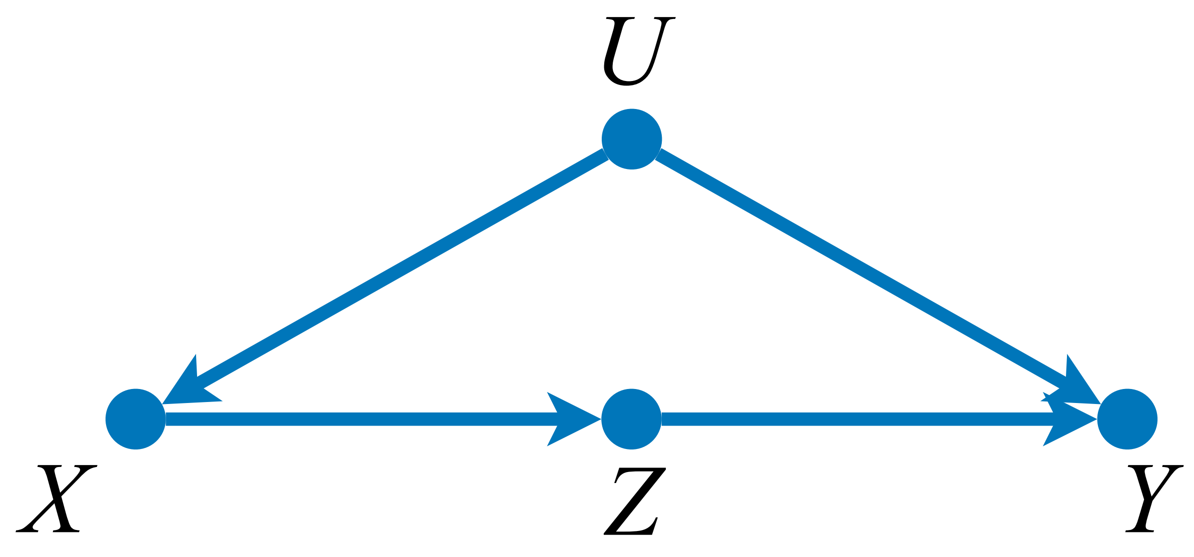
Factorization of Observed Distributions

- The Markov property leads to a natural factorization when all variables are observed, ie:



Factorization of Observed Distributions

- The Markov property leads to a natural factorization when all variables are observed, ie:

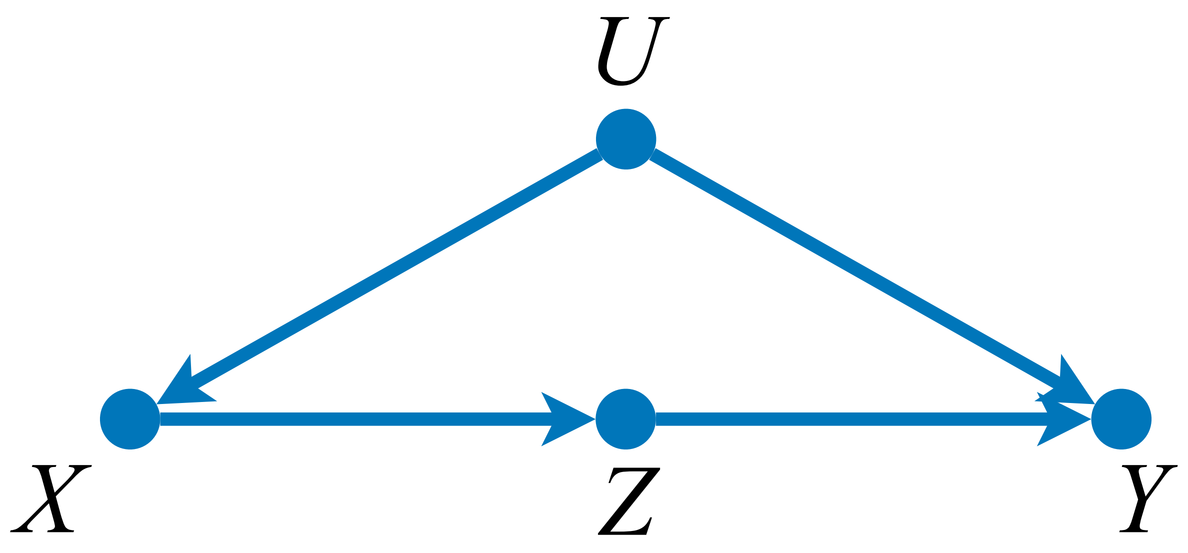


$$P(\mathbf{v}) = \prod_i P(v_i \mid pa_i) = P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

(where V is the set of all observable variables)

Factorization of Observed Distributions

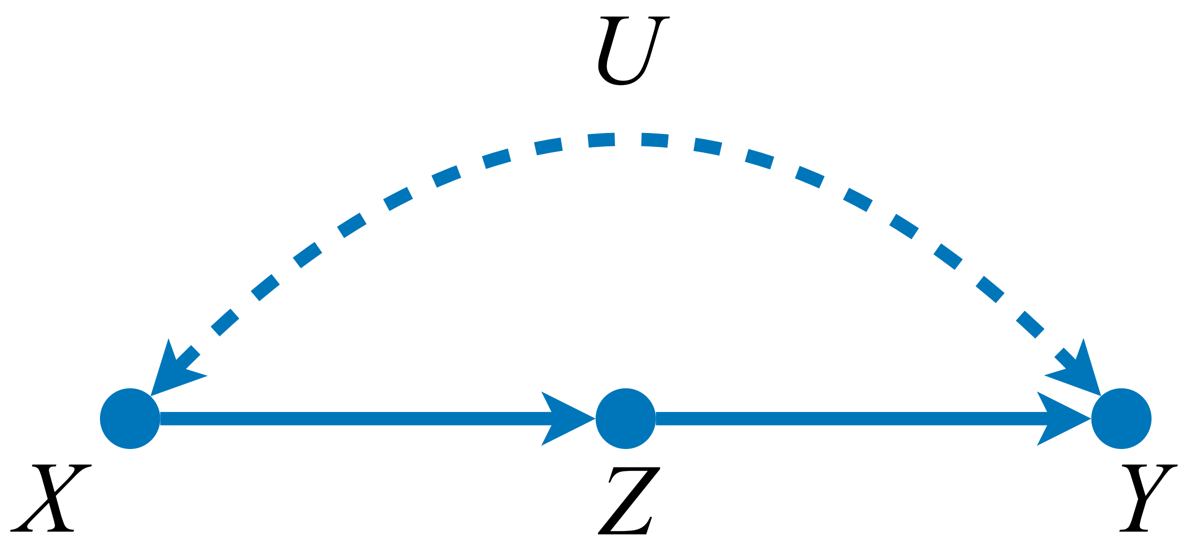
- The Markov property leads to a natural factorization when all variables are observed, ie:



$$P(\mathbf{v}) = \prod_i P(v_i \mid pa_i) = P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

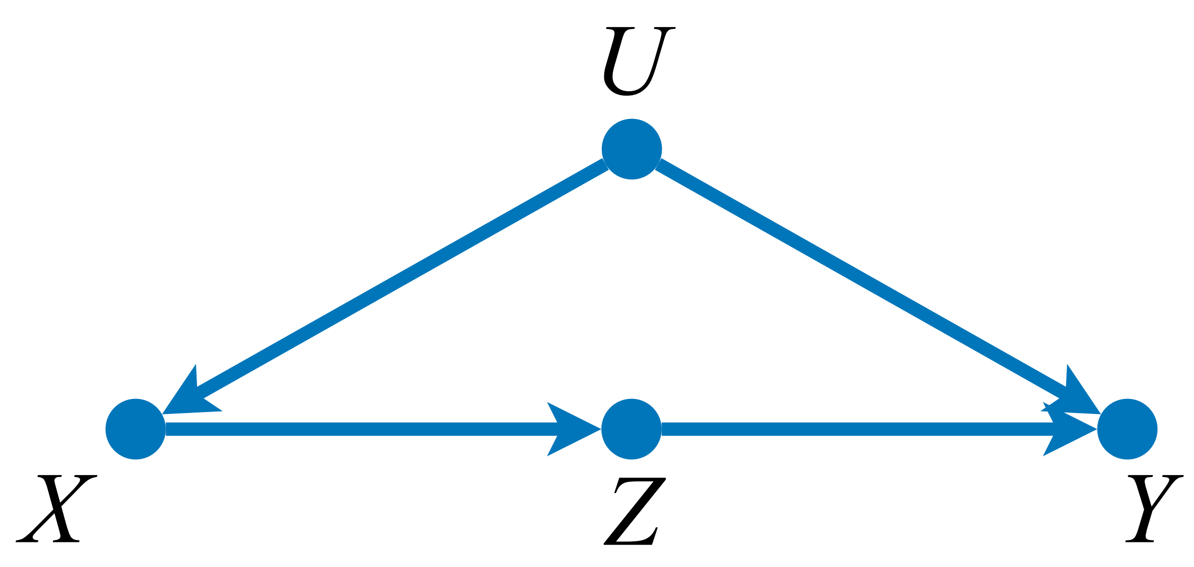
(where V is the set of all observable variables)

- How to factorize the observed distribution in the presence of latent variables?



Factorization of Observed Distributions

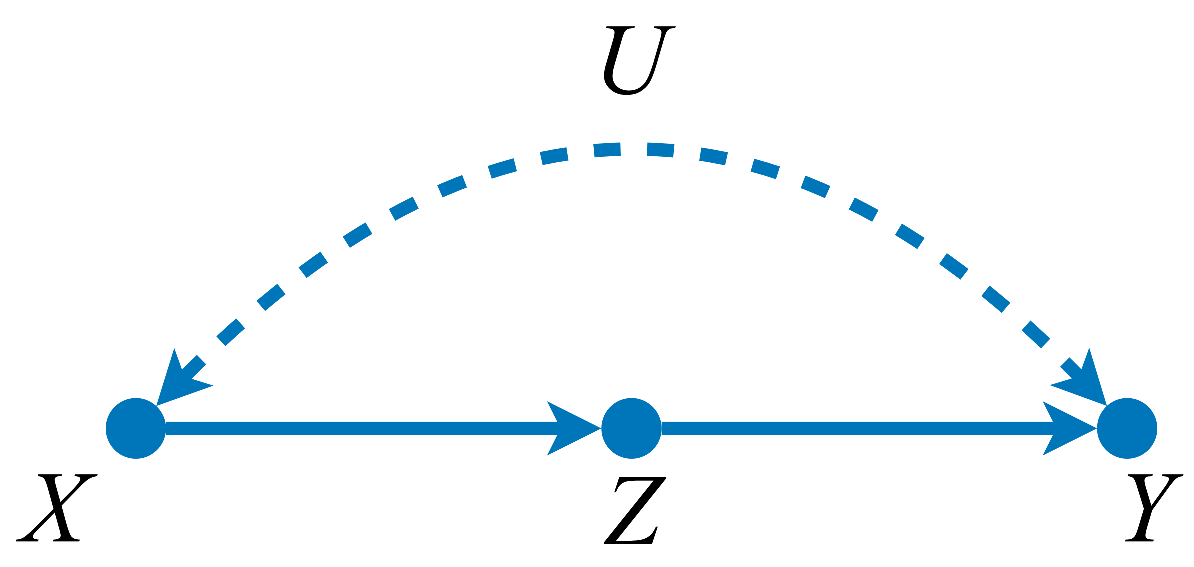
- The Markov property leads to a natural factorization when all variables are observed, ie:



$$P(\mathbf{v}) = \prod_i P(v_i | pa_i) = P(x | u)P(z | x)P(y | z, u)P(u)$$

(where V is the set of all observable variables)

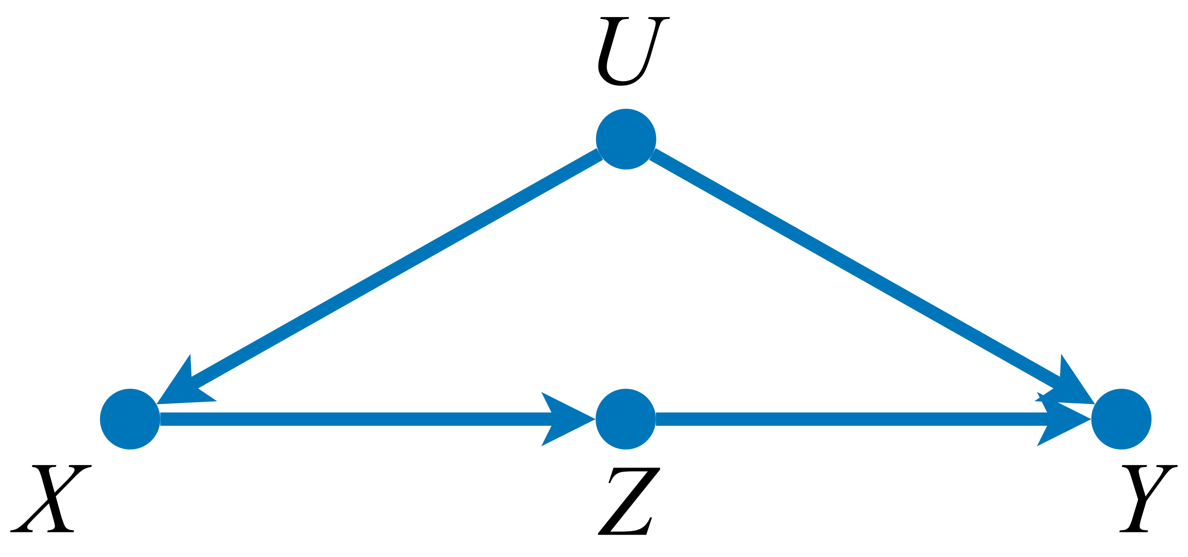
- How to factorize the observed distribution in the presence of latent variables?



$$P(\mathbf{v}) = \sum_u P(x, z, y, u) = \sum_u P(x | u)P(z | x)P(y | z, u)P(u)$$

Factorization of Observed Distributions

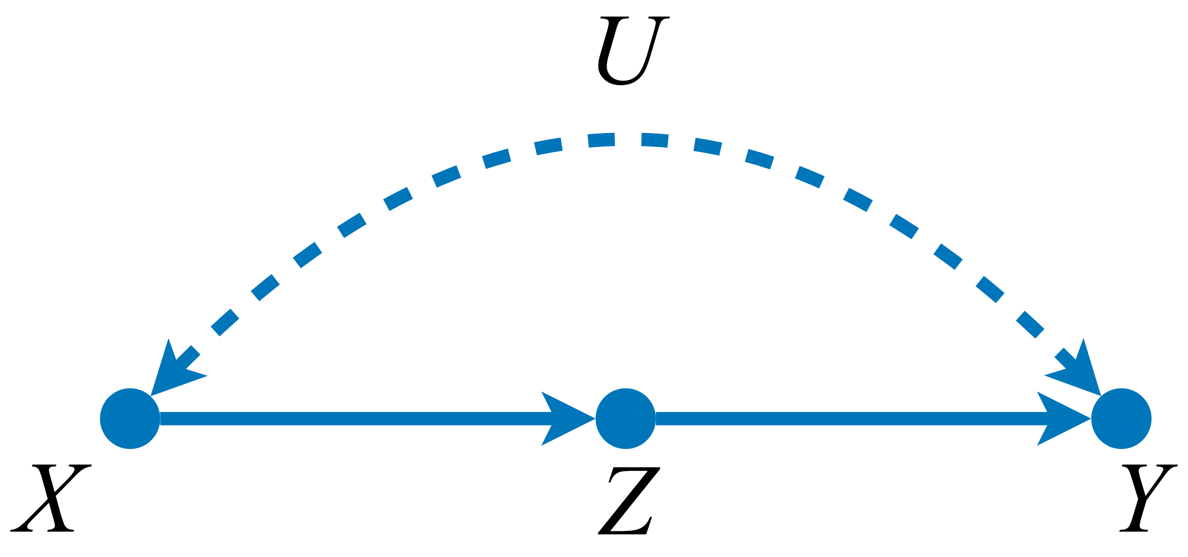
- The Markov property leads to a natural factorization when all variables are observed, ie:



$$P(\mathbf{v}) = \prod_i P(v_i \mid pa_i) = P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

(where V is the set of all observable variables)

- How to factorize the observed distribution in the presence of latent variables?

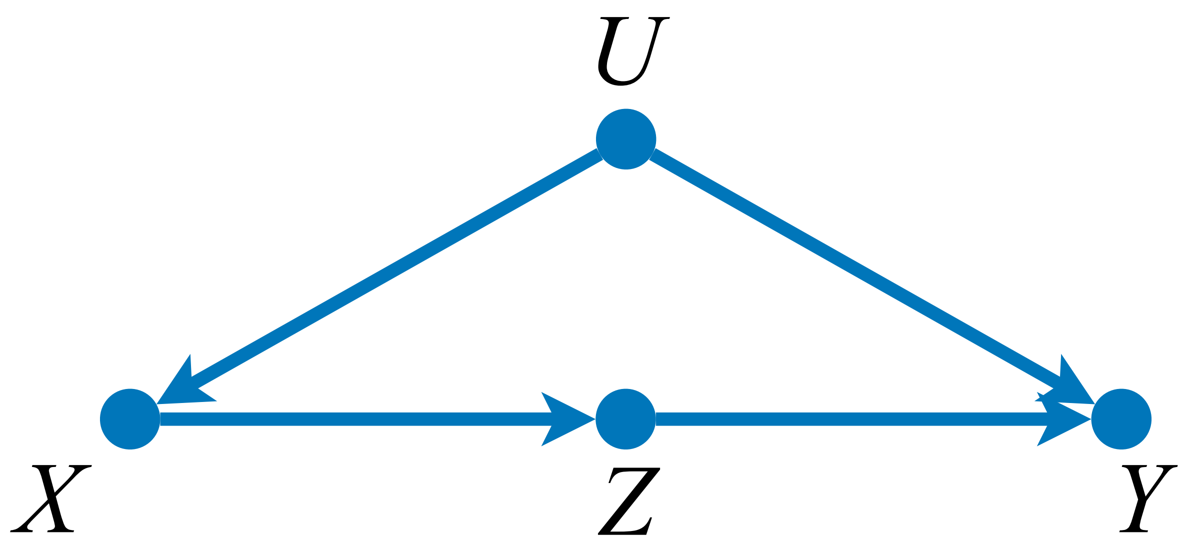


$$P(\mathbf{v}) = \sum_u P(x, z, y, u) = \sum_u P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

$$= P(z \mid x)\left(\sum_u P(x \mid u)P(y \mid z, u)P(u)\right)$$

Factorization of Observed Distributions

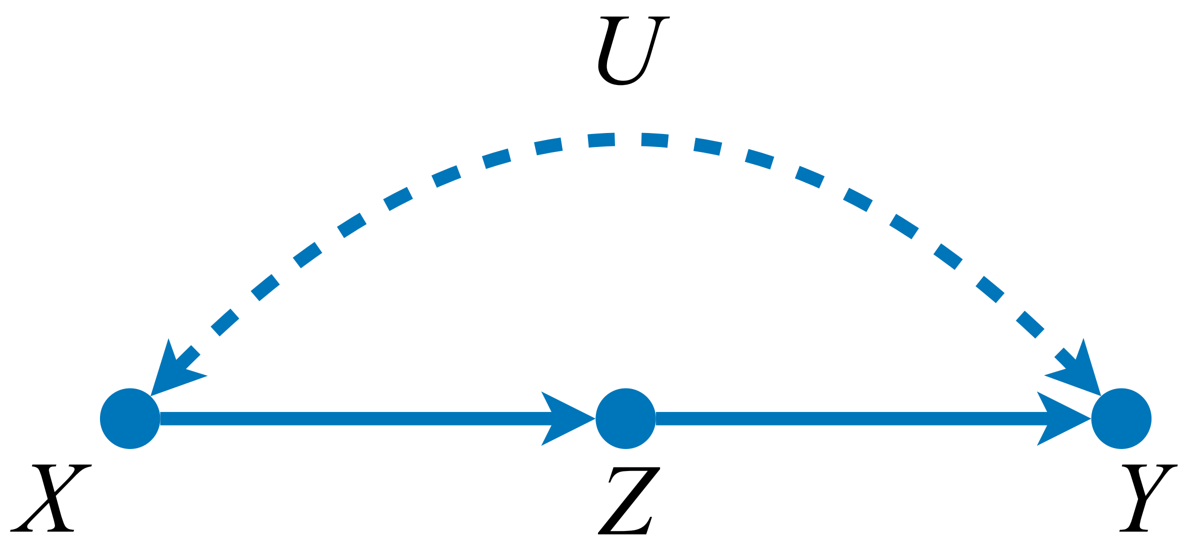
- The Markov property leads to a natural factorization when all variables are observed, ie:



$$P(\mathbf{v}) = \prod_i P(v_i \mid pa_i) = P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

(where V is the set of all observable variables)

- How to factorize the observed distribution in the presence of latent variables?



$$P(\mathbf{v}) = \sum_u P(x, z, y, u) = \sum_u P(x \mid u)P(z \mid x)P(y \mid z, u)P(u)$$

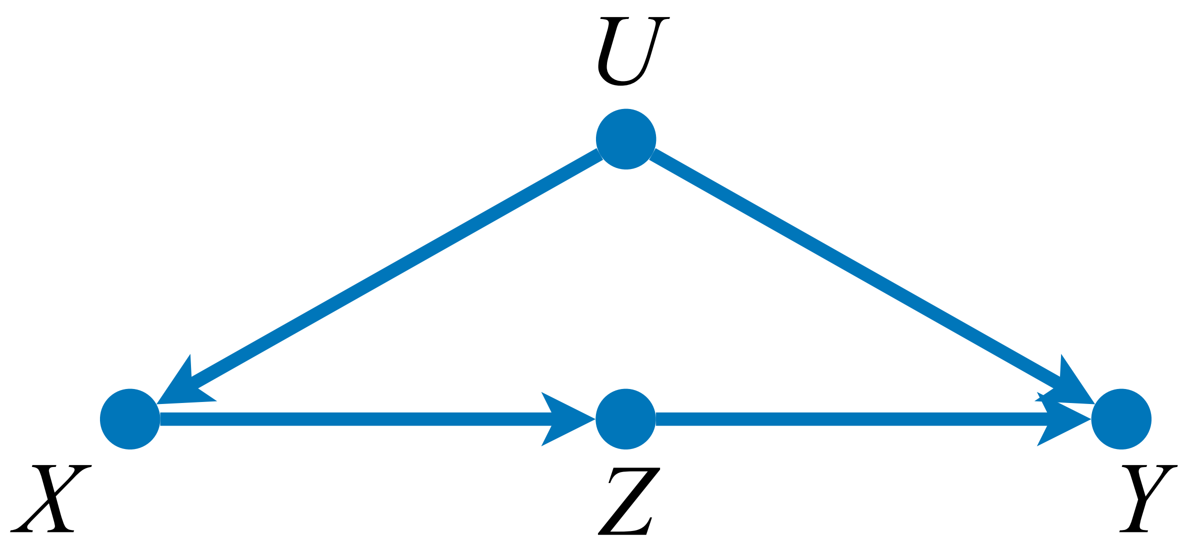
$$= P(z \mid x)\left(\sum_u P(x \mid u)P(y \mid z, u)P(u) \right)$$

$$= P(x)P(z \mid x)\left(\sum_{x'} P(y \mid z, x')P(x') \right)$$

Causal Inference tools give us the means to identify some factors involving latent variables from observed distributions.

Factorization of Observed Distributions

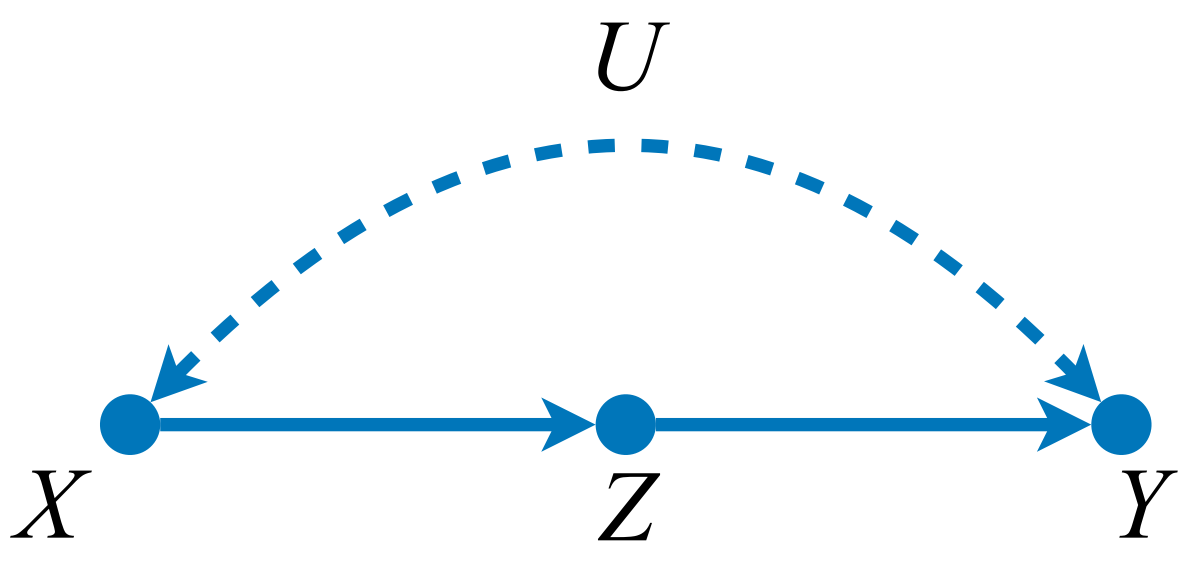
- The Markov property leads to a natural factorization when all variables are observed, ie:



$$P(\mathbf{v}) = \prod_i P(v_i \,|\, pa_i) = P(x \,|\, u)P(z \,|\, x)P(y \,|\, z, u)P(u)$$

(where V is the set of all observable variables)

- How to factorize the observed distribution in the presence of latent variables?



$$P(\mathbf{v}) = \sum_u P(x, z, y, u) = \sum_u P(x \,|\, u)P(z \,|\, x)P(y \,|\, z, u)P(u)$$

$$= P(z \,|\, x)\left(\sum_u P(x \,|\, u)P(y \,|\, z, u)P(u) \right)$$

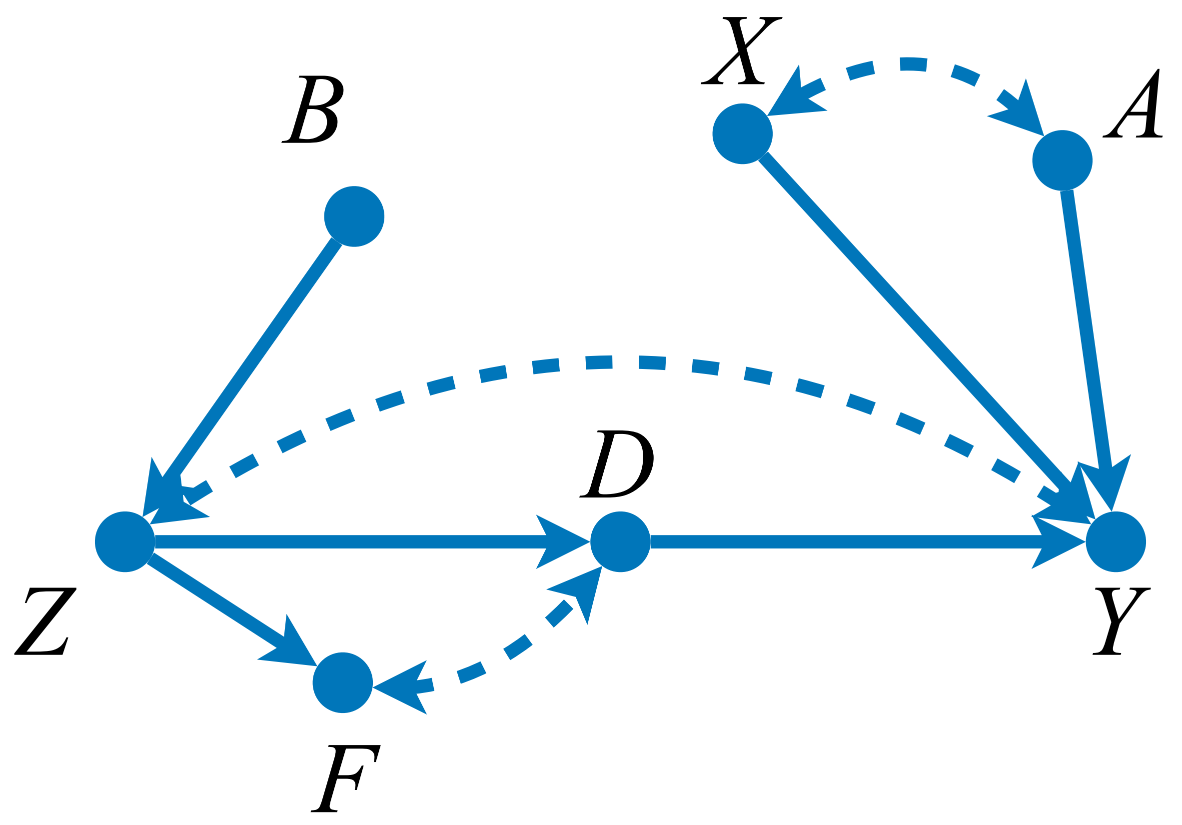
$$= P(x)P(z \,|\, x)\left(\qquad Q[Y] \qquad \right)$$

Causal Inference tools give us the means to identify some factors involving latent variables from observed distributions.

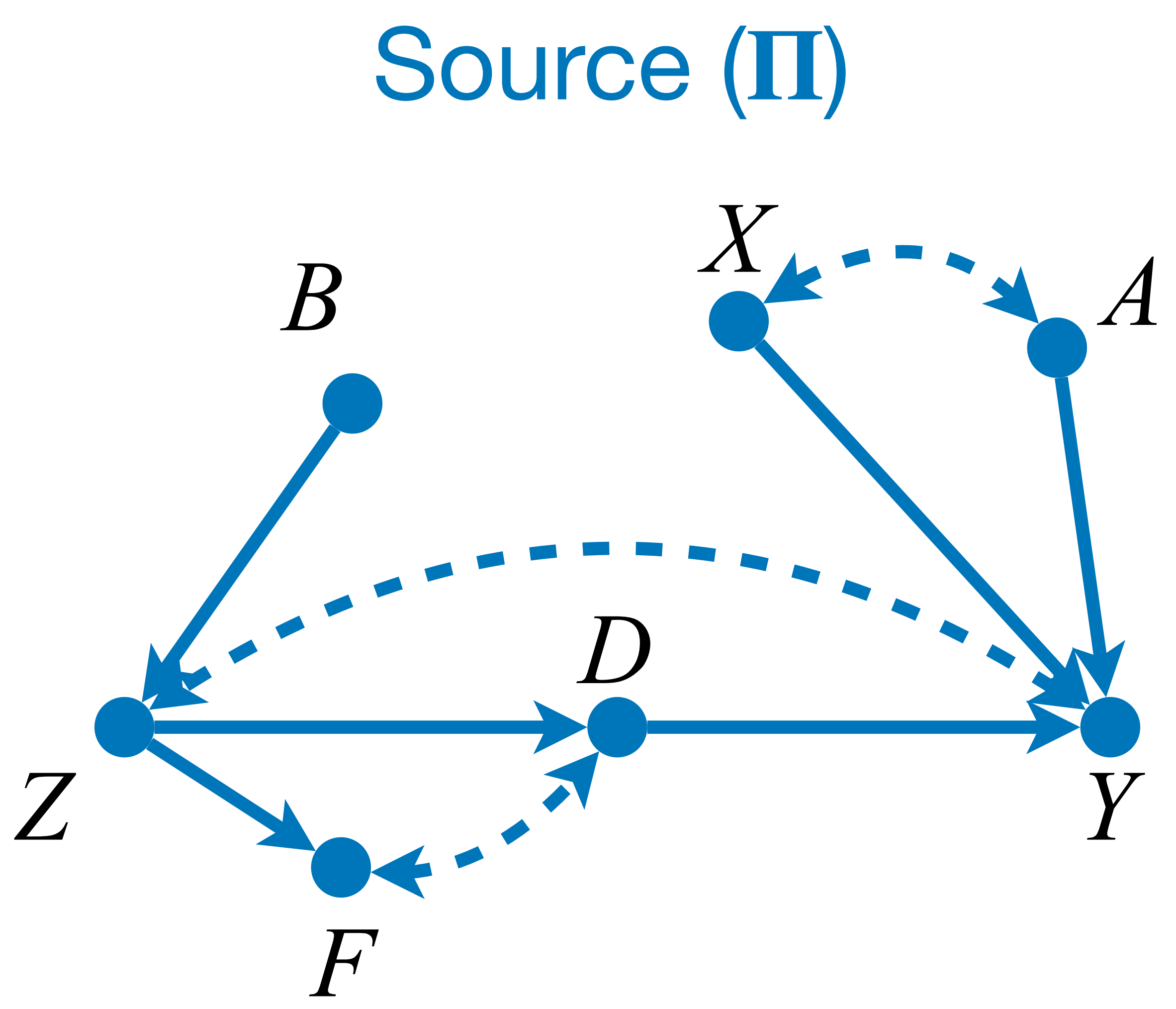
A slightly more complicated example

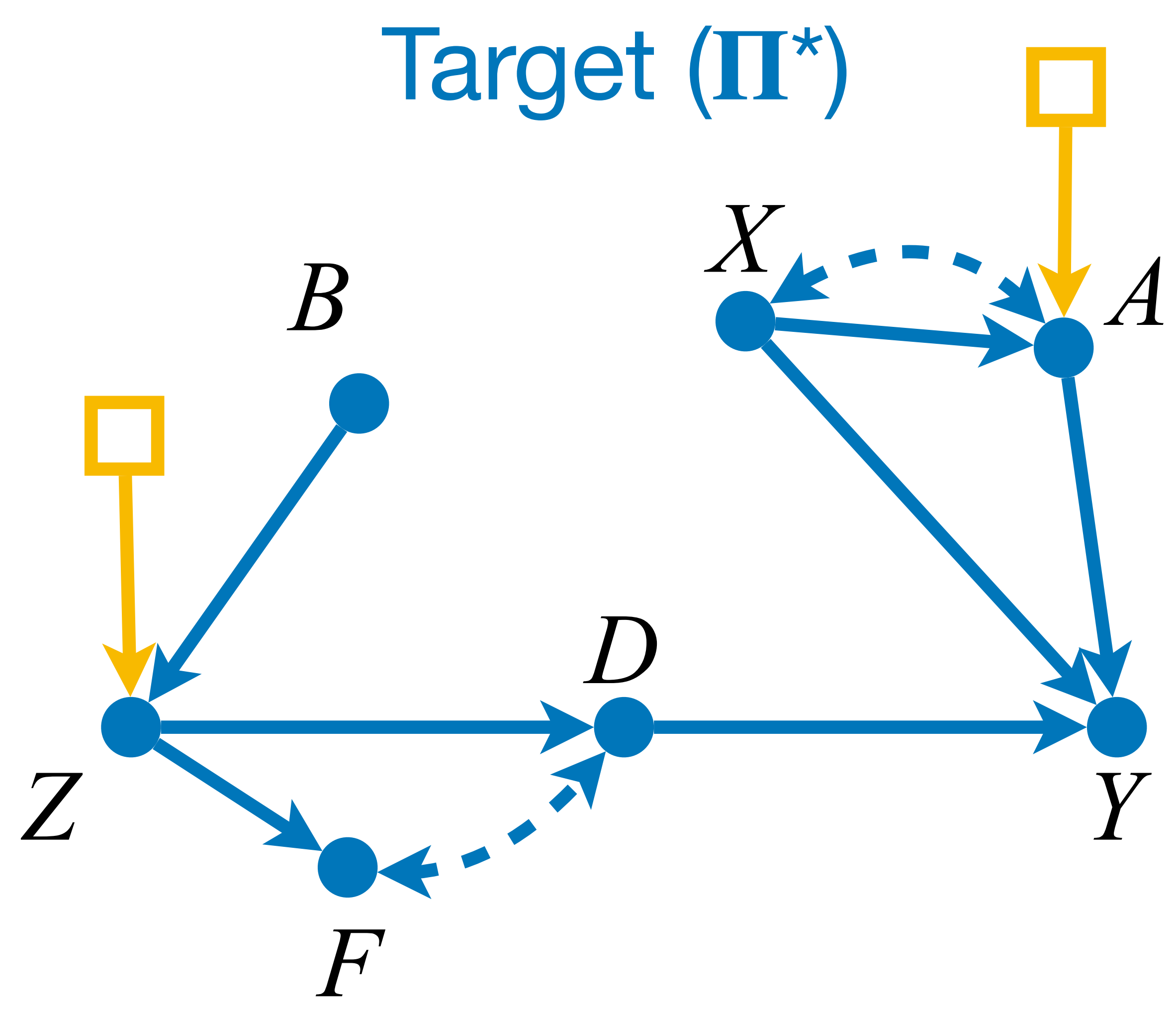
A slightly more complicated example

Source (Π)

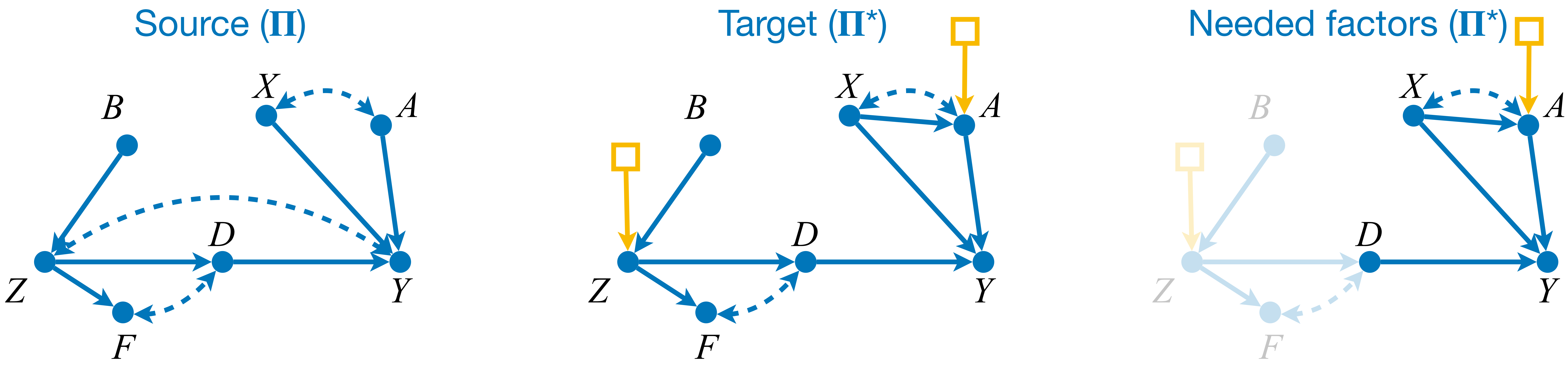


A slightly more complicated example



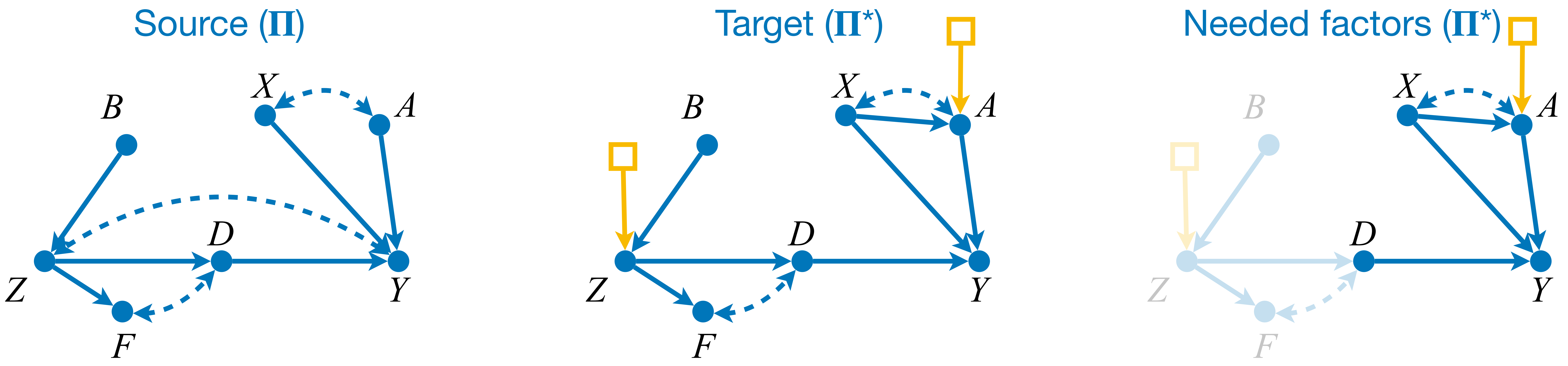


A slightly more complicated example



- Suppose the inferential target is $P^*(y|x,z)$. After some algebra, one can show that given $P(b,z,f,d,x,a,y)$ and $P^*(x,a)$, it can be written as

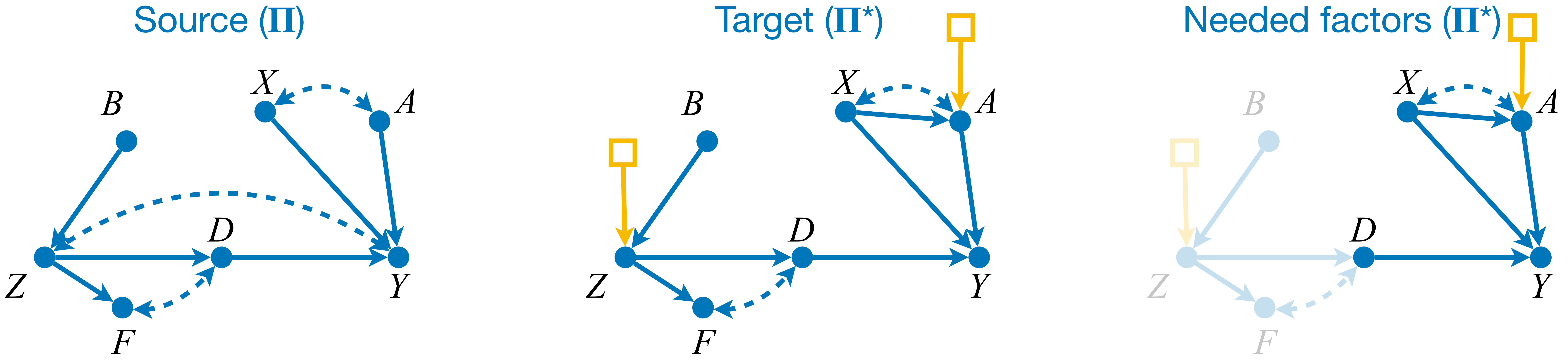
A slightly more complicated example



- Suppose the inferential target is $P^*(y|x,z)$. After some algebra, one can show that given $P(b,z,f,d,x,a,y)$ and $P^*(x,a)$, it can be written as

$$P^*(y \mid x, z) = \sum_{a,d} Q^*[A,X]Q[D]Q[Y] \Big/ \sum_{a,d,y} Q^*[A,X]Q[D]Q[Y]$$

A slightly more complicated example



- Suppose the inferential target is $P^*(y|x,z)$. After some algebra, one can show that given $P(b,z,f,d,x,a,y)$ and $P^*(x,a)$, it can be written as

$$P^*(y\,|\,x,z) = \sum_{a,d} Q^*[A,X]Q[D]Q[Y] \Big/ \sum_{a,d,y} Q^*[A,X]Q[D]Q[Y]$$

$$P^*(y\,|\,x,z) = \sum_{a} P^*(a\,|\,x) \sum_{d} P(d\,|\,z) \sum_{z'} P(y\,|\,x,z',d,a)P(z')$$

Dynamic Plan Identification reduces to Statistical Transportability

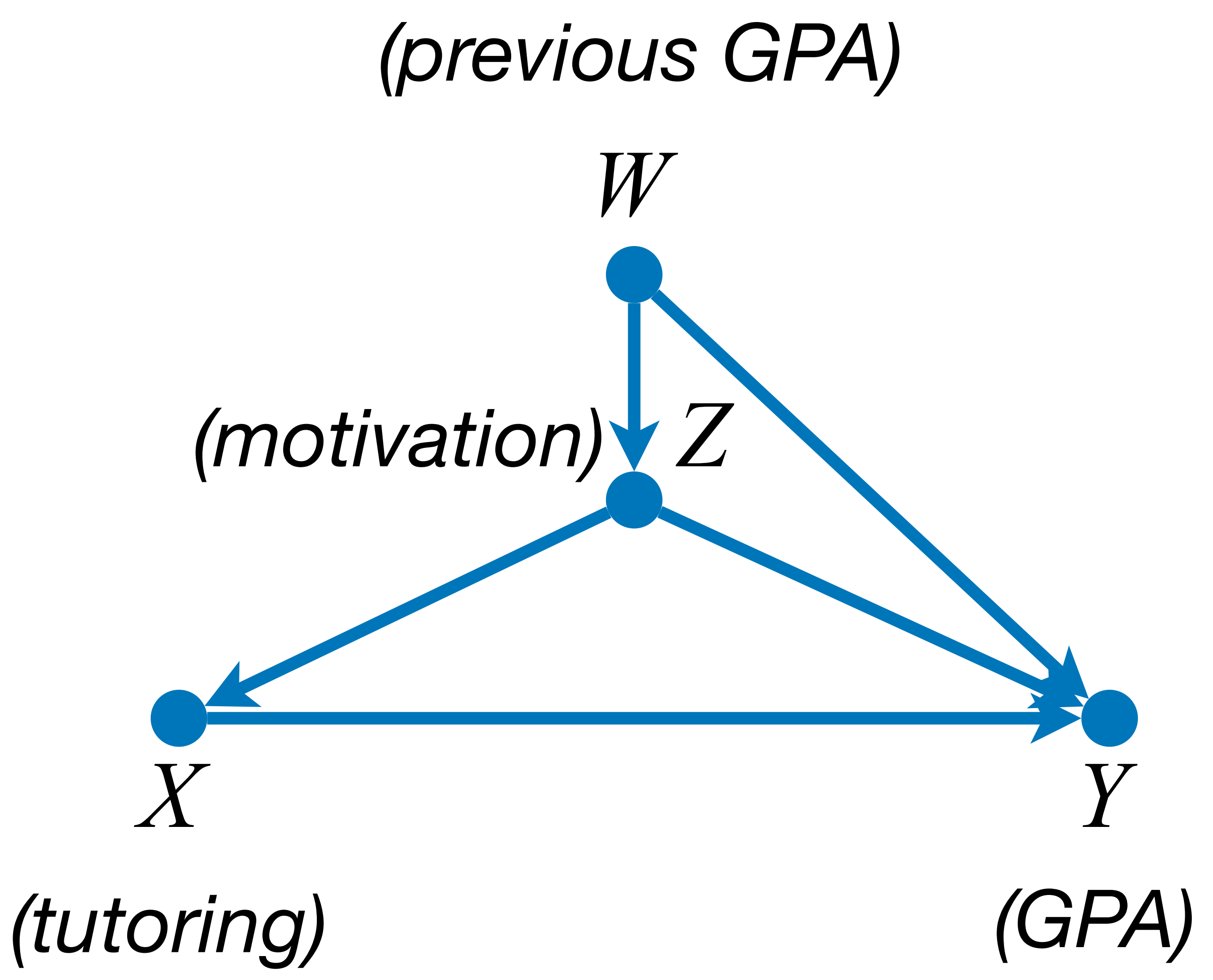
Dynamic Plan Identification reduces to Statistical Transportability

Key observation. If the source environment corresponds to the current system, and the target environment corresponds to the source after an intervention, then transporting the distribution $P^*(\mathbf{y})$ is the same as identifying the effect of the intervention on an outcome $\boldsymbol{Y}$.

Dynamic Plan Identification reduces to Statistical Transportability

Key observation. If the source environment corresponds to the current system, and the target environment corresponds to the source after an intervention, then transporting the distribution $P^*(\mathbf{y})$ is the same as identifying the effect of the intervention on an outcome $\mathbf{Y}$.

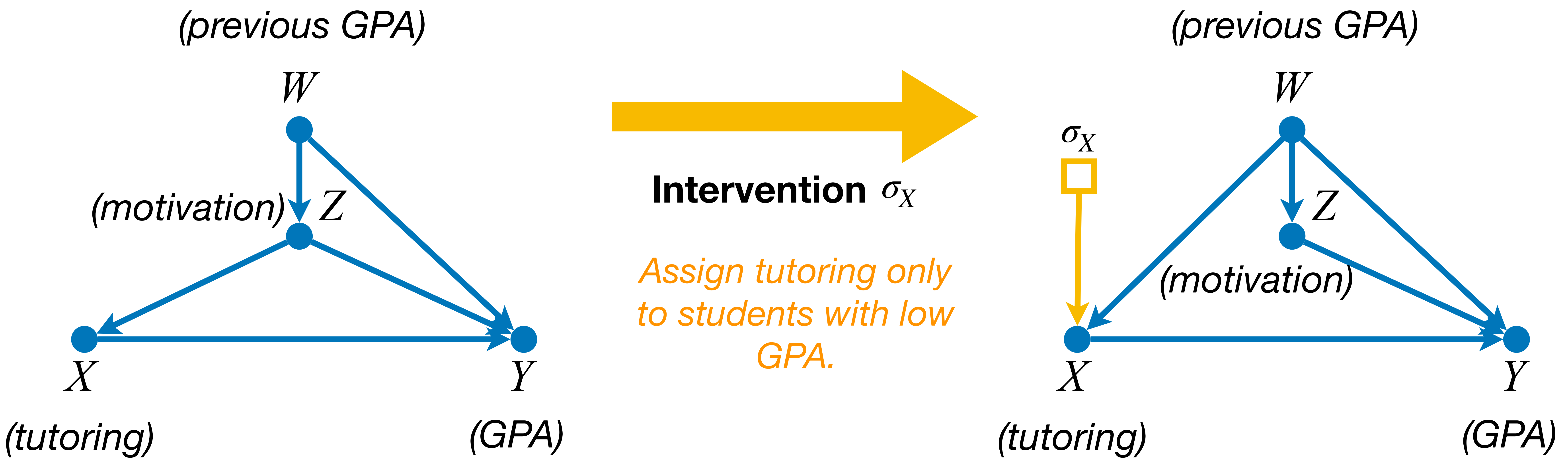
Students get tutoring on their own volition based on their motivation.



Dynamic Plan Identification reduces to Statistical Transportability

Key observation. If the source environment corresponds to the current system, and the target environment corresponds to the source after an intervention, then transporting the distribution $P^*(y)$ is the same as identifying the effect of the intervention on an outcome Y.

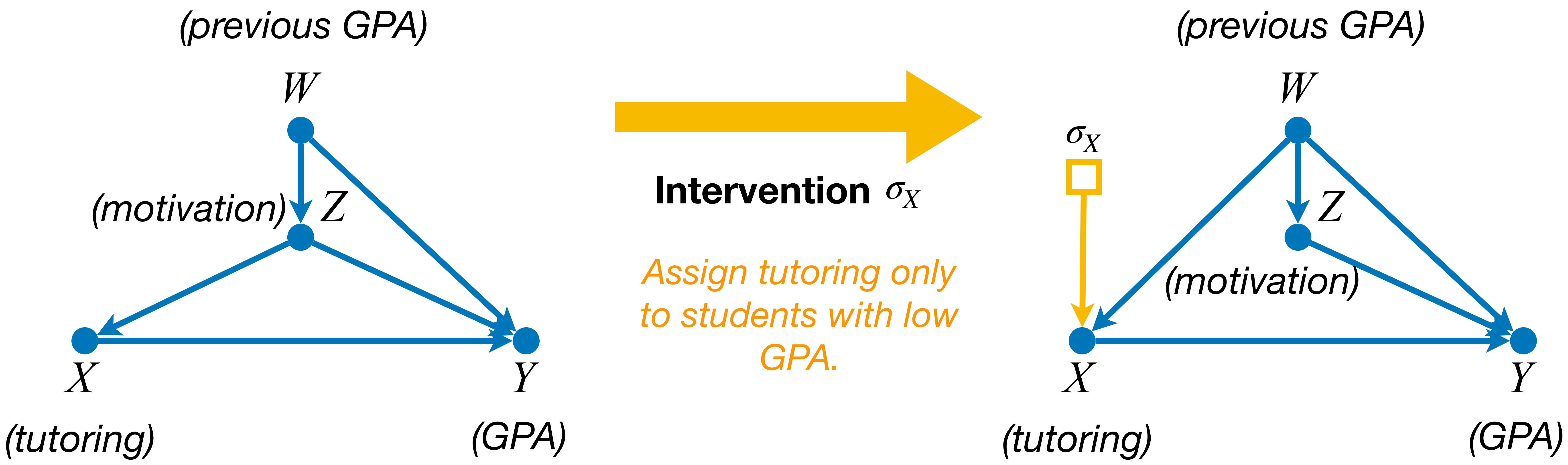
Students get tutoring on their own volition based on their motivation.



Dynamic Plan Identification reduces to Statistical Transportability

Key observation. If the source environment corresponds to the current system, and the target environment corresponds to the source after an intervention, then transporting the distribution $P^*(\mathbf{y})$ is the same as identifying the effect of the intervention on an outcome $\mathbf{Y}$.

Students get tutoring on their own volition based on their motivation.



$P^*(y)$ represents the effect of σ_X on Y.

Conclusions

Conclusions

- Leveraging causal inference tools, we solved the problem of generalizability of probability distributions across different, but related environments.

Conclusions

- Leveraging causal inference tools, we solved the problem of generalizability of probability distributions across different, but related environments.
- We proposed a sound and complete procedure to decide whether a target distribution is transportable from observations in a source domain and partial measurements in the target domain, following the assumptions encoded in graphical models representing the data generating process in the domains.

Conclusions

- Leveraging causal inference tools, we solved the problem of generalizability of probability distributions across different, but related environments.
- We proposed a sound and complete procedure to decide whether a target distribution is transportable from observations in a source domain and partial measurements in the target domain, following the assumptions encoded in graphical models representing the data generating process in the domains.
- Leveraging these results, we solved the problem of identification of stochastic interventions.

Thank you!