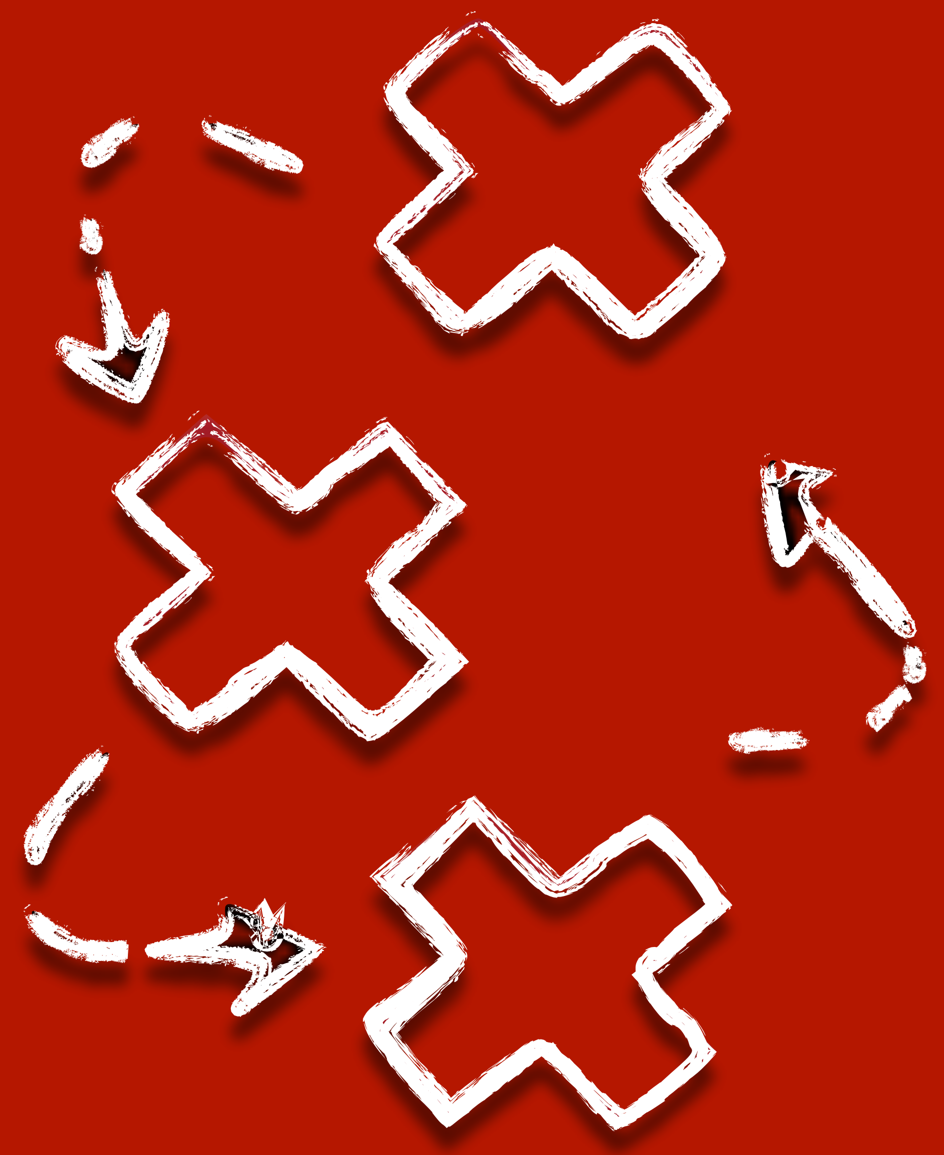


Causal Data Science

Lecture 8.1: Estimation methods 2

Lecturer: Sara Magliacane

UvA - Spring 2023

Last class: Estimands for binary treatments

- We generally cannot estimate **unit-level causal effect:** $Y_i(t=1) - Y_i(t=0)$
- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}[Y \mid \text{do}(T=1)] - \mathbb{E}[Y \mid \text{do}(T=0)]$$

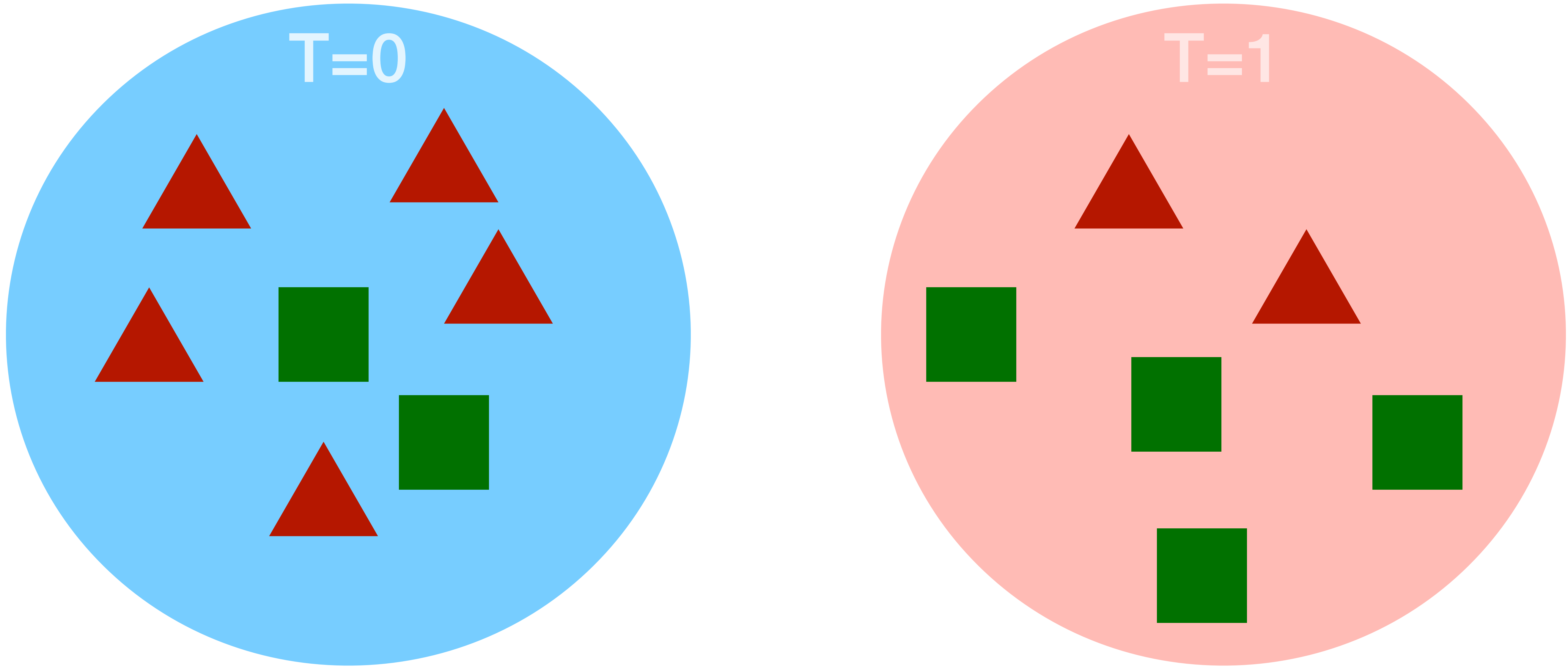
- We can estimate the **average causal effect of treatment on the treated**:

$$\text{ATT} = \mathbb{E}[Y(t=1) - Y(t=0) \mid T=1]$$

- For all, we assume that our covariates $\mathbf{X}$ form a valid adjustment set (e.g. we can check them/filter them with backdoor criterion)

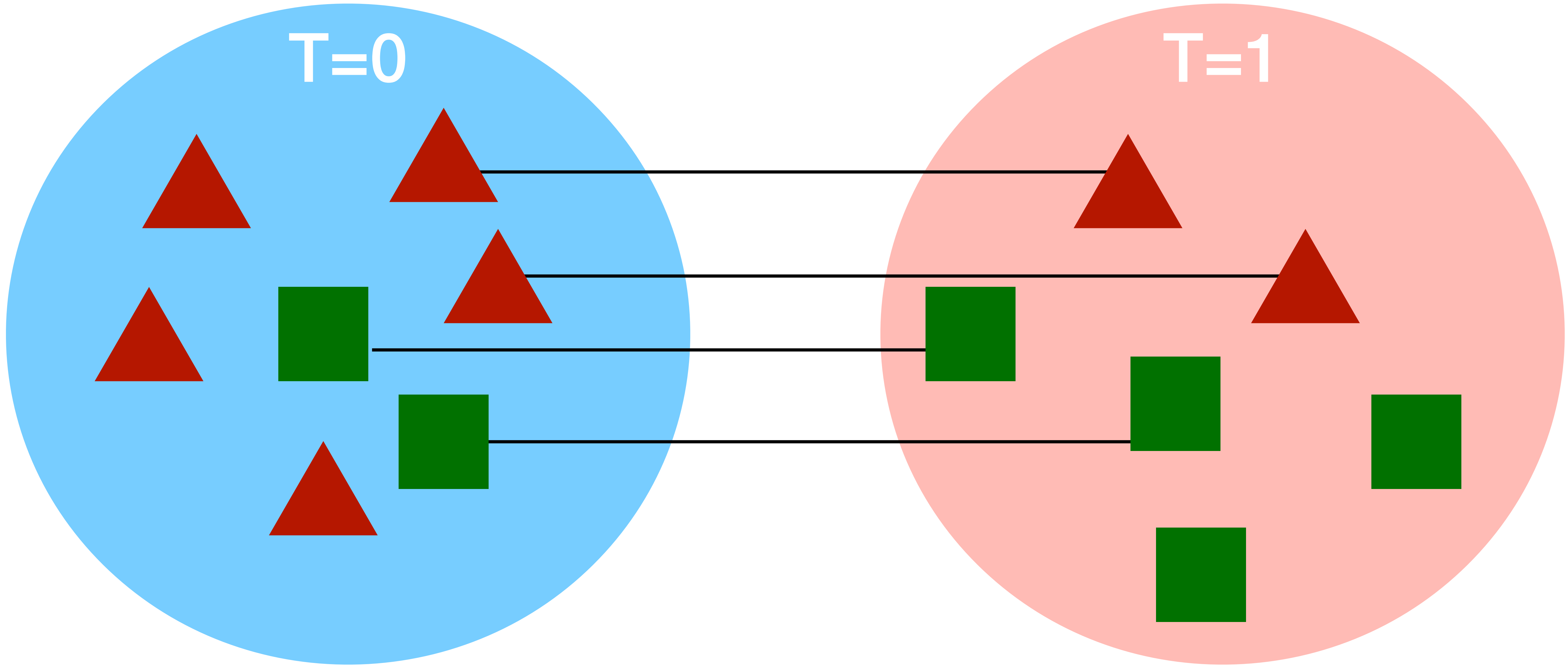
Last class: Exact matching (simplified)

- **Usually for ATT,** sometimes for ATE
- **Intution:** find the most similar couple of units in terms of covariates $\mathbf{X}$, such that one is in the treatment and the other in the control group



Last class: Exact matching (simplified)

- **Usually for ATT,** sometimes for ATE
- **Intution:** find the most similar couple of units in terms of covariates $\mathbf{X}$, such that one is in the treatment and the other in the control group

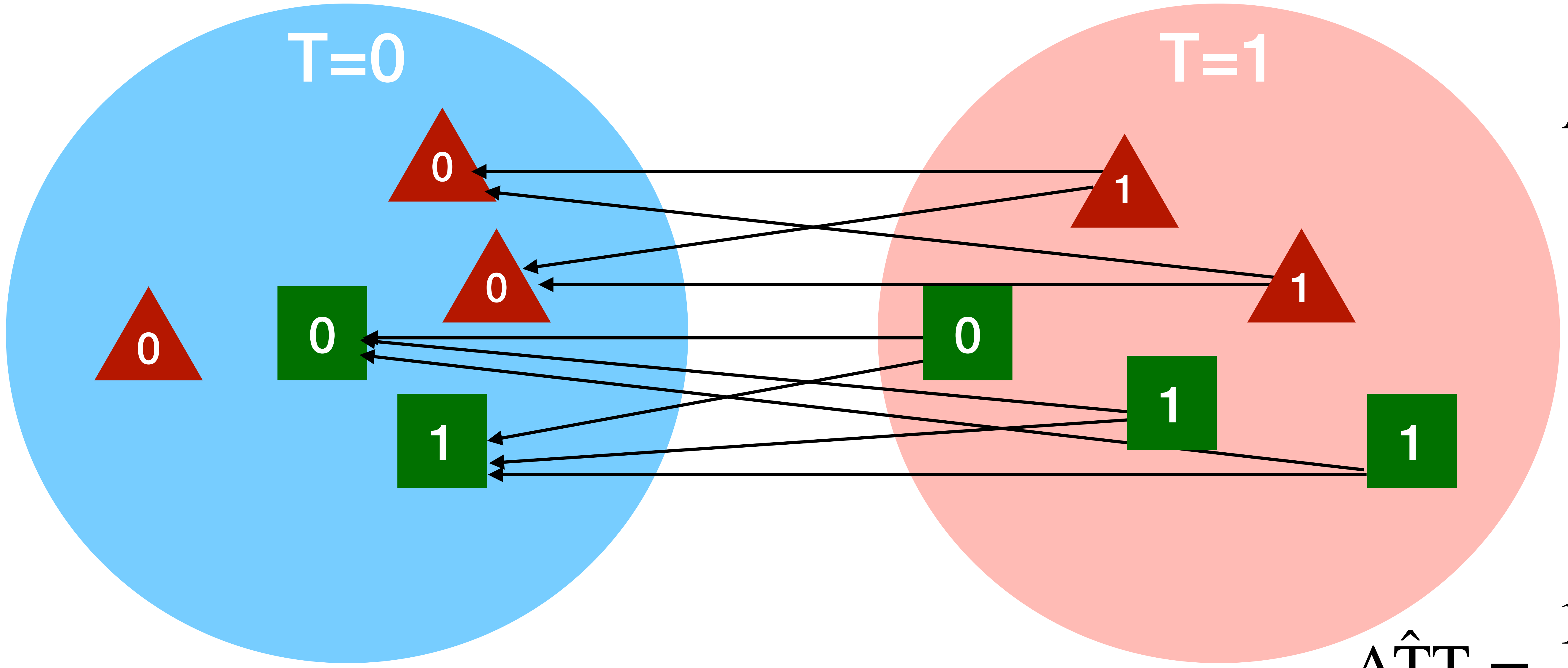


Last class: Exact matching (slightly less simplified)

- **Usually for ATT,** with multiple matches M (e.g. M=2, can be random):

$$\hat{\text{ATT}} = \frac{1}{n_t}\sum_{i=1}^{n_t}(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{mj(i)})$$

$Y_{mj(i)}$ **match m for I**



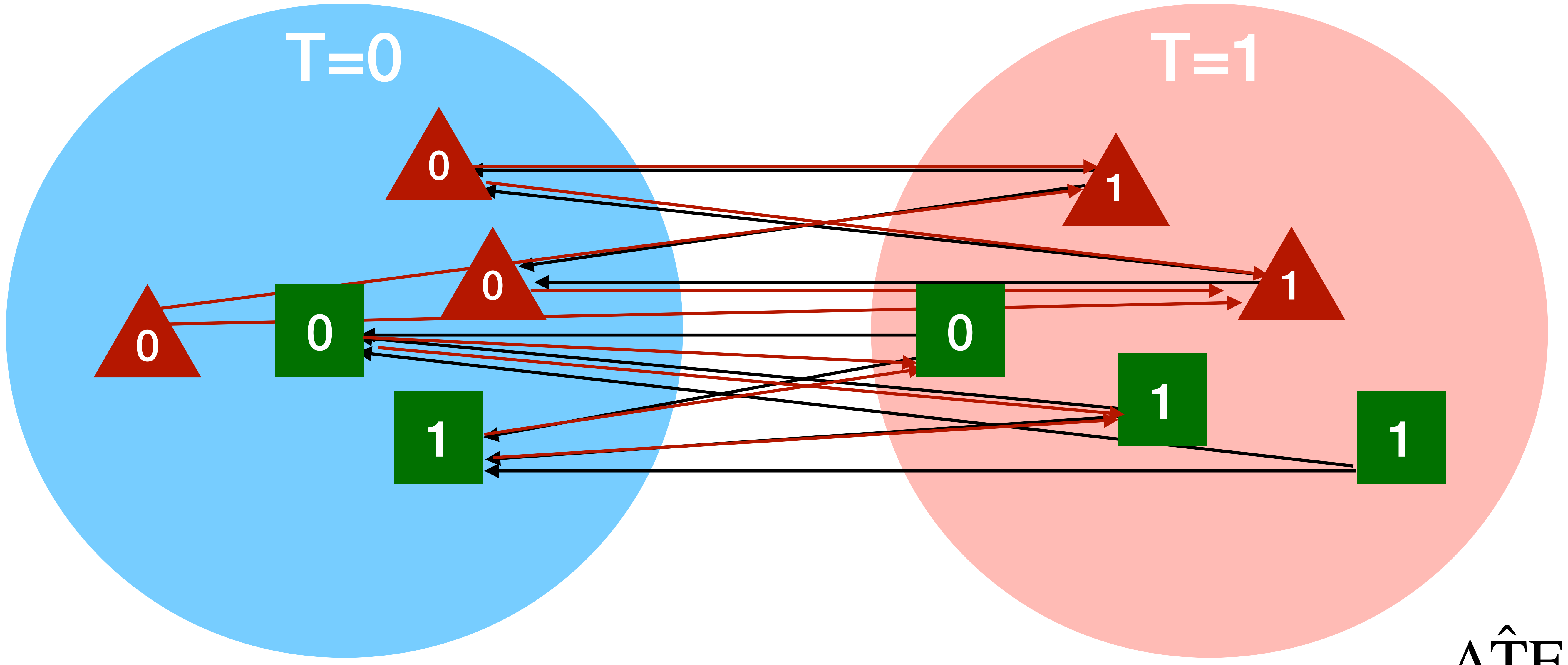
$$\hat{\text{ATT}} = \frac{1}{5}\sum_{i=1}^{5}(Y_i - \frac{1}{2}\sum_{m=1}^{M} Y_{mj(i)})$$

$$\hat{\text{ATT}} = \frac{1}{5}[1 + 1 - \frac{1}{2} + \frac{2}{2}] = \frac{1}{5}\cdot\frac{5}{2} = \frac{1}{2}$$

Last class: Exact matching (slightly less simplified)

- **Usually for ATT,** with multiple matches M (e.g. M=2, can be random):

$$\hat{ATE} = \frac{1}{n_t + n_c}[\sum_{i=1}^{n_t}(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{mj(i)}) + \sum_{j=1}^{n_c}(\frac{1}{M}\sum_{m=1}^{M} Y_{mi(j)} - Y_j)]$$



$$\hat{ATE} = \frac{1}{10}[\frac{5}{2} + 3 + \frac{1}{2} - \frac{1}{2}] = \frac{11}{20}$$

Last class: Distance matching

- If exact matching on the value is not possible, e.g. because we have continuous covariates, we can use any **distance**, e.g Mahalanobis distance
  - For example, kNN
- Need to check **covariate balancing** after matching (e.g. std mean difference)

$$T \perp\!\!\!\perp \mathbf{X} \equiv P(\mathbf{X} \mid T = 0) = P(\mathbf{X} \mid T = 1)$$

- **Potential issue:** $\mathbf{X}$ is high-dimensional (has many dimensions) and it's difficult to find good matches -> can we find a single number to match?

Last class: Propensity score matching (PSM)

- **Assumptions**: binary treatment T, $\mathbf{X}$ is valid adjustment set
- **Propensity score:** the probability of getting assigned the treatment

$$e(x) \quad \pi(x) := P(T = 1 \mid \mathbf{X} = x)$$

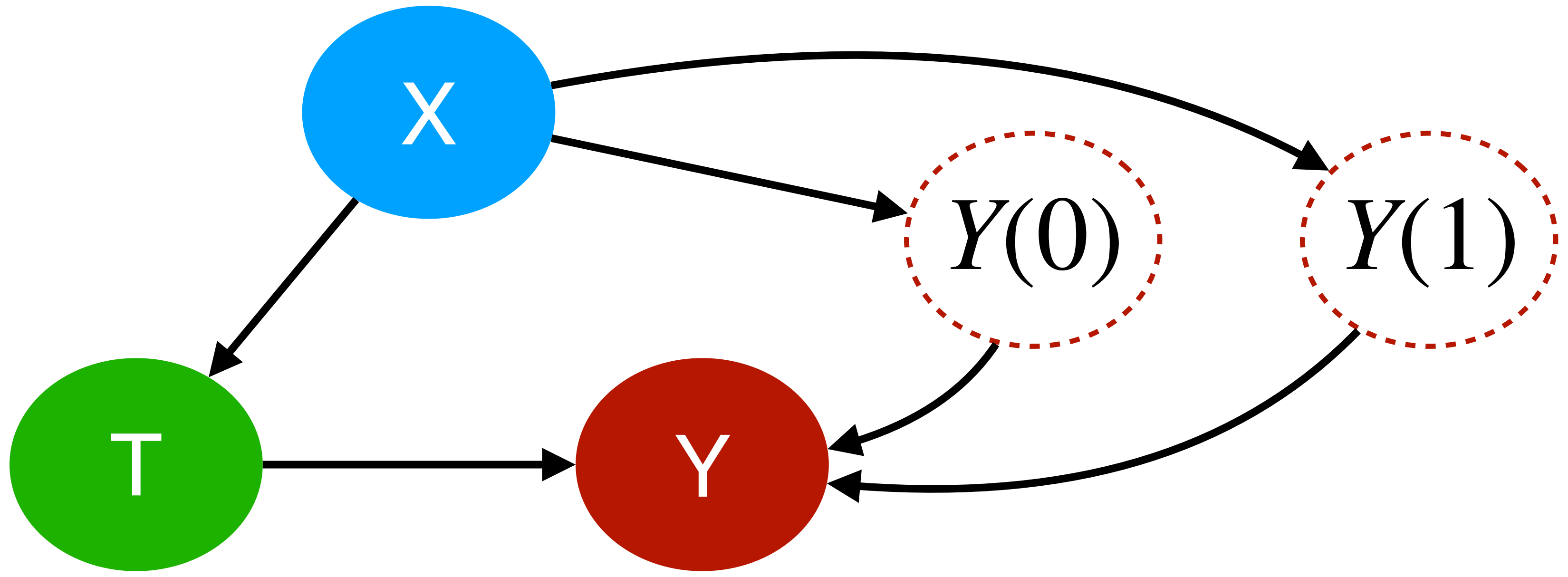
Last class: Propensity score matching (PSM)

- **Assumptions**: binary treatment T, $\mathbf{X}$ is valid adjustment set
- **Propensity score:** the probability of getting assigned the treatment

$$e(x) \quad \pi(x) := P(T = 1 \mid \mathbf{X} = x)$$

Conditional ignorability/No unmeasured confounding

- We can show that $T \perp\!\!\!\perp \mathbf{X} \mid \pi(\mathbf{X})$ and that if $Y(0), Y(1) \perp\!\!\!\perp T \mid \mathbf{X}$ then



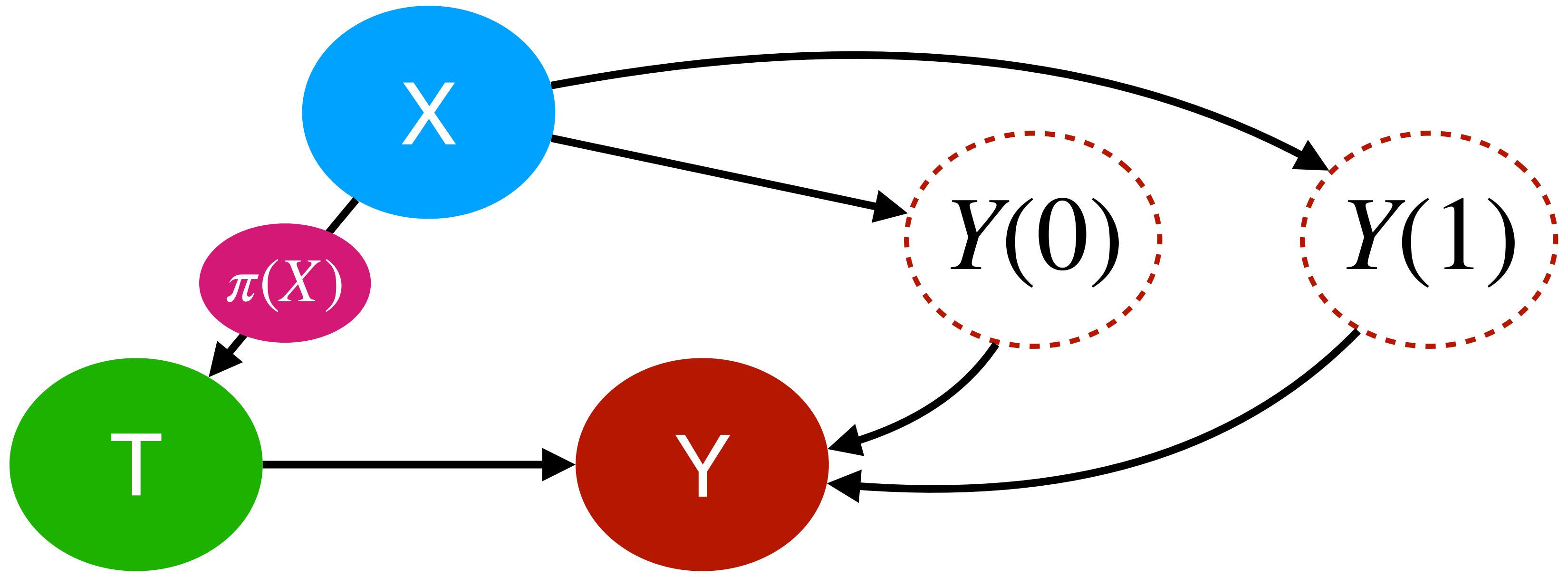
Last class: Propensity score matching (PSM)

- **Assumptions**: binary treatment T, $\mathbf{X}$ is valid adjustment set
- **Propensity score:** the probability of getting assigned the treatment

$$e(x) \quad \pi(x) := P(T = 1 \,|\, \mathbf{X} = x)$$

- We can show that $T \perp\!\!\!\perp \mathbf{X} \,|\, \pi(\mathbf{X})$ and that if $Y(0), Y(1) \perp\!\!\!\perp T \,|\, \mathbf{X}$ then

$$Y(0), Y(1) \perp\!\!\!\perp T \,|\, \pi(\mathbf{X})$$



- We can estimate π from data and use it to match
  - If $\mathbf{X}$ has a lot of covariates, it is easier to match since it's a single number

Last class: Inverse probability weighting (IPW)

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}[Y \mid \text{do}(T=1)] - \mathbb{E}[Y \mid \text{do}(T=0)]$$

- $\mathbf{X}$ is a valid adjustment set for the causal effect of T on Y, so:

$$P(Y=y \mid \text{do}(T=1)) = \sum_{\mathbf{x}} P(Y=y \mid \mathbf{X}=\mathbf{x}, T=1) P(\mathbf{X}=\mathbf{x})$$

Last class: Inverse probability weighting (IPW)

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}[Y \mid \text{do}(T=1)] - \mathbb{E}[Y \mid \text{do}(T=0)]$$

- $\mathbf{X}$ is a valid adjustment set for the causal effect of T on Y, so:

$$P(Y=y \mid \text{do}(T=t)) = \sum_{\mathbf{x}} P(Y=y \mid \mathbf{X}=\mathbf{x}, T=t) P(\mathbf{X}=\mathbf{x})$$

$$\mathbb{E}[Y \mid \text{do}(T=t)] = \sum_{y} y \sum_{\mathbf{x}} P(Y=y \mid \mathbf{X}=\mathbf{x}, T=t) P(\mathbf{X}=\mathbf{x})$$

Last class: Inverse probability weighting (IPW)

$$\mathbb{E}[Y|\mathrm{do}(T=t)] = \sum_y \sum_{\mathbf{x}} y \cdot P(Y=y|\mathbf{X}=\mathbf{x}, T=t)P(\mathbf{X}=\mathbf{x})$$

$P(T=t|\mathbf{X}=\mathbf{x}) \neq 0$

$$= \sum_y \sum_{\mathbf{x}} y \cdot P(Y=y|\mathbf{X}=\mathbf{x}, T=t)P(\mathbf{X}=\mathbf{x})\frac{P(T=t|\mathbf{X}=\mathbf{x})}{P(T=t|\mathbf{X}=\mathbf{x})}$$

$P(Y=y, \mathbf{X}=\mathbf{x}, T=t)$

$$= \sum_y \sum_{\mathbf{x}} y \cdot P(Y=y|\mathbf{X}=\mathbf{x}, T=t)P(\mathbf{X}=\mathbf{x})\frac{P(T=t|\mathbf{X}=\mathbf{x})}{P(T=t|\mathbf{X}=\mathbf{x})}$$

$$= \sum_y \sum_{\mathbf{x}} \frac{y \cdot P(Y=y, \mathbf{X}=\mathbf{x}, T=t)}{P(T=t|\mathbf{X}=\mathbf{x})}$$

π for t = 1, $(1-\pi)$ for t = 0

Estimation method: Inverse probability weighting (IPW)

- **Inverse probability (of treatment) weighting:** weight by inverse of probability of treatment **received**:
  - For treated $T = 1$: weight by the inverse of $\pi = P(T = 1 \mid \mathbf{X})$
  - For untreated $T = 0$: weight by the inverse of $1 - \pi = P(T = 0 \mid \mathbf{X})$

$$\hat{\mathbb{E}}(Y(t=1)) = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot 1\{T=1\} \cdot \frac{1}{\boxed{P(T=1 \mid X_i)}} \qquad \pi$$

$$\hat{\mathbb{E}}(Y(t=0)) = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot 1\{T=0\} \cdot \frac{1}{\boxed{P(T=0 \mid X_i)}} \qquad (1-\pi)$$

IPW Example

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ T \arrow[rr] & & Y \end{tikzcd}$$

Population:

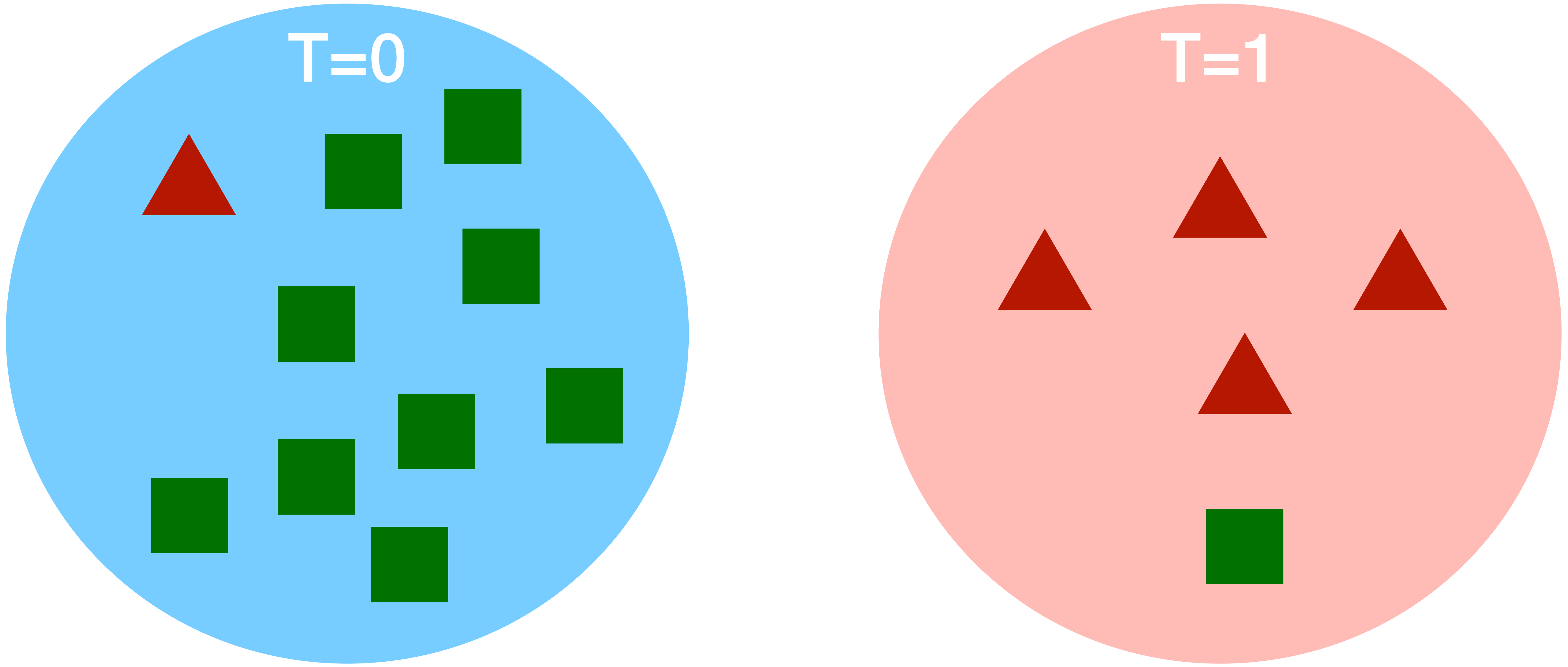
	X=0	X=1
T=0	1	9
T=1	4	1

$P(T=1 \mid X=1) = 0.1$

$P(T=1 \mid X=0) = 0.8$

$P(T=0 \mid X=1) = 0.9$

$P(T=0 \mid X=0) = 0.2$



IPW Example

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ T \arrow[rr] & & Y \end{tikzcd}$$

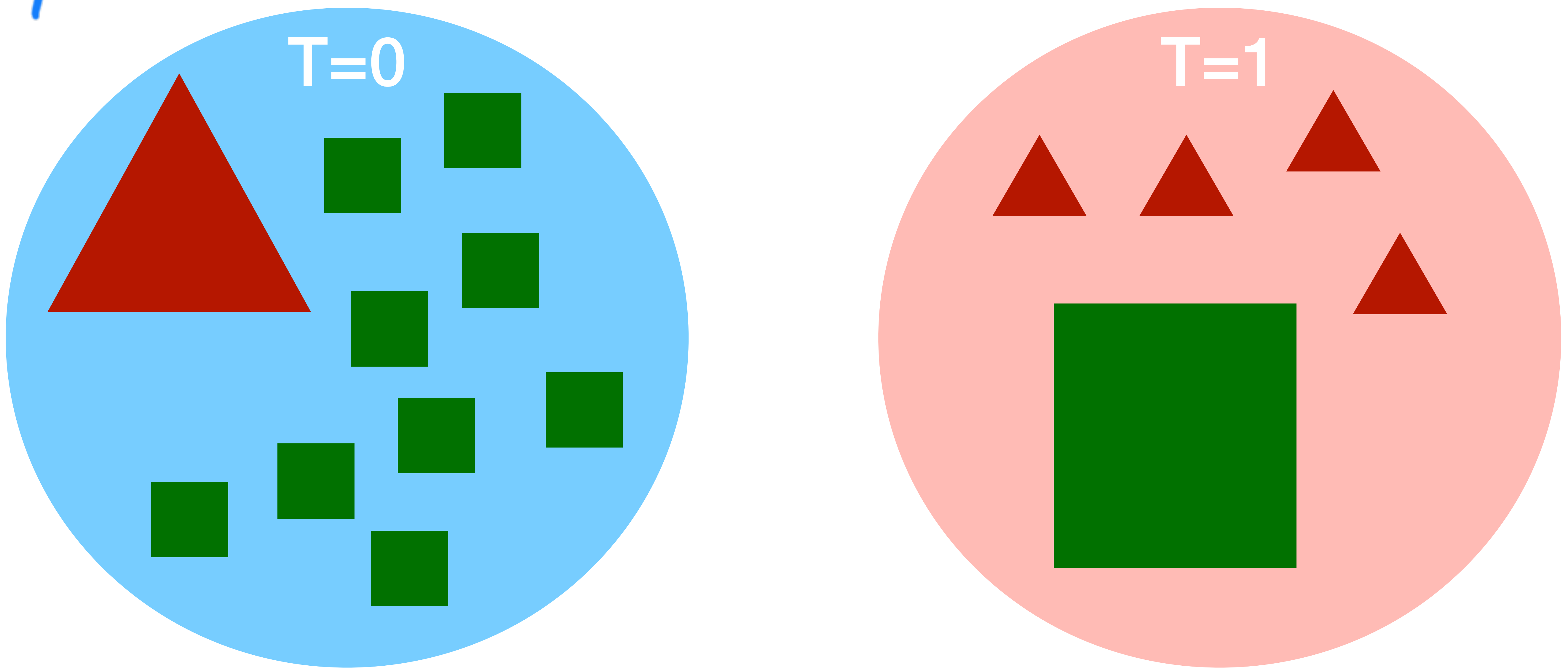
Population:

$P(T=1 \mid X=1) = 0.1$

$P(T=1 \mid X=0) = 0.8$

$P(T=0 \mid X=1) = 0.9$

$P(T=0 \mid X=0) = 0.2$



	X=0	X=1
T=0	1/0.2	9/0.9
T=1	4/0.8	1/0.1

	X=0	X=1
T=0	5	10
T=1	5	10

Reweight by $\dfrac{1}{P(T_i \mid X_i)}$

Estimation method: Inverse probability weighting (IPW)

- **Inverse probability (of treatment) weighting:** weight by inverse of estimated probability of treatment **received**:
  - For treated $T = 1$: weight by the inverse of $\hat{\pi}(X_i)$
  - For untreated $T = 0$: weight by the inverse of $1 - \hat{\pi}(X_i)$

$$\hat{\mathbb{E}}(Y(t=1)) = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot T_i \cdot \frac{1}{\hat{\pi}(X_i)}$$

$$\hat{\mathbb{E}}(Y(t=0)) = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot (1 - T_i) \cdot \frac{1}{1 - \hat{\pi}(X_i)}$$

For example with logistic regression

What if the estimated $\hat{\pi}(X_i)$ is biased?

Estimating (conditional) average treatment effects

- We can estimate the average causal effect/**average treatment effect**

$$\mathrm{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y(t=1) \mid \mathbf{X}] - \mathbb{E}[Y(t=0) \mid \mathbf{X}]]$$

We still assume $\mathbf{X}$ is a valid adjustment set!

Estimating (conditional) average treatment effects

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}_X[\underbrace{\mathbb{E}[Y(t=1)\,|\,X]}_{\hat{\mu}(1,\mathbf{X})} - \underbrace{\mathbb{E}[Y(t=0)\,|\,X]}_{\hat{\mu}(0,\mathbf{X})}]$$

We still assume $\mathbf{X}$ is a valid adjustment set!

Estimating (conditional) average treatment effects

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}_X[\underbrace{\mathbb{E}[Y(t=1) \mid X]}_{\hat{\mu}(1,\mathbf{X})} - \underbrace{\mathbb{E}[Y(t=0) \mid X]}_{\hat{\mu}(0,\mathbf{X})}]$$

$$\hat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1,\mathbf{x}_i) - \hat{\mu}(0,\mathbf{x}_i)$$

We still assume $\mathbf{X}$ is a valid adjustment set!

Estimating (conditional) average treatment effects

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}_{\mathrm{X}}[\underbrace{\mathbb{E}[Y(t=1)\,|\,\mathrm{X}]}_{\hat{\mu}(1,\mathbf{X})} - \underbrace{\mathbb{E}[Y(t=0)\,|\,\mathrm{X}]}_{\hat{\mu}(0,\mathbf{X})}]$$

$$\hat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1,\mathbf{x}_i) - \hat{\mu}(0,\mathbf{x}_i)$$

We still assume $\mathbf{X}$ is a valid adjustment set!

- We can also estimate the **conditional average treatment effect:**

$$\text{CATE}(\mathrm{w}) = \mathbb{E}[Y(t=1) - Y(t=0)\,|\,\mathrm{W} = \mathrm{w}]$$

Estimating (conditional) average treatment effects

- We can estimate the average causal effect/**average treatment effect**

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}_{\text{X}}[\underbrace{\mathbb{E}[Y(t=1) \mid \text{X}]}_{\hat{\mu}(1,\mathbf{X})} - \underbrace{\mathbb{E}[Y(t=0) \mid \text{X}]}_{\hat{\mu}(0,\mathbf{X})}]$$

$$\hat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1,\mathbf{x}_i) - \hat{\mu}(0,\mathbf{x}_i)$$

We assume $\mathbf{X} \cup W$ is a valid adjustment set!

- We can also estimate the **conditional average treatment effect:**

$$\text{CATE(w)} = \mathbb{E}[Y(t=1) - Y(t=0) \mid \text{W} = \text{w}]$$

$$= \mathbb{E}_{\text{X}}[\underbrace{\mathbb{E}[Y(t=1) \mid \text{X}, \text{W} = \text{w}]}_{\hat{\mu}(1,\mathbf{x}_i, w)} - \underbrace{\mathbb{E}[Y(t=0) \mid \text{X}, \text{W} = \text{w}]}_{\hat{\mu}(1,\mathbf{x}_i, w)}]$$

S-learners [Küntzel et al 2019]

- We learn a single model to predict the both potential outcomes $Y_i(0), Y_i(1)$

$$\hat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, \mathbf{x}_i) - \hat{\mu}(0, \mathbf{x}_i)$$

$$\hat{\text{CATE}}(\text{w}) = \frac{1}{n_w}\sum_{i=1}^{n} 1(W = w)[\hat{\mu}(1, \mathbf{x}_i, w) - \hat{\mu}(0, \mathbf{x}_i, w)]$$

- **Issue:** for high-dimensional $\mathbf{X}$, S-learners can ignore the treatment

https://www.pnas.org/doi/10.1073/pnas.1804597116

X-learners [Küntzel et al 2019]

1. Learn two separate models $\hat{\mu}_1(\mathbf{x}_i)$ (only treated) and $\hat{\mu}_0(\mathbf{x}_i)$ (only control)
2. We impute the treatment effect per unit (*individual treatment effect)*

$Y(0)$

Treatment group

$$\hat{\tau}_{i,1} = Y_i - \hat{\mu}_0(\mathbf{x}_i)$$

Estimated from control

Control group

$$\hat{\tau}_{i,0} = \hat{\mu}_1(\mathbf{x}_i) - Y_i$$

Estimated from treated

Unit	Y(0)	Y(1)	T	X
1	?	1	1	1
2	1	?	0	1
3	?	0	1	0
4	0	?	0	0
5	?	1	1	1
6	?	0	1	0

X-learners [Küntzel et al 2019]

1. Learn two separate models $\hat{\mu}_1(\mathbf{x}_i)$ (only treated) and $\hat{\mu}_0(\mathbf{x}_i)$ (only control)
2. We impute the treatment effect per unit (*individual treatment effect)*

$Y(0)$

Treatment group

$$\hat{\tau}_{i,1} = Y_i - \boxed{\hat{\mu}_0(\mathbf{x}_i)}$$

Estimated from control

Control group

$$\hat{\tau}_{i,0} = \boxed{\hat{\mu}_1(\mathbf{x}_i)} - Y_i$$

Estimated from treated

Unit	Y(0)	Y(1)	T	X
1	?	1	1	1
2	1	?	0	1
3	?	0	1	0
4	0	?	0	0
5	?	1	1	1
6	?	0	1	0

Unit	Y(0)	Y(1)	T	X
1	1	1	1	1
2	1	0	0	1
3	1	0	1	0
4	0	0	0	0
5	1	1	1	1
6	1	0	1	0

X-learners [Küntzel et al 2019]

1. Learn two separate models $\hat{\mu}_1(\mathbf{x}_i)$ (only treated) and $\hat{\mu}_0(\mathbf{x}_i)$ (only control)
2. We impute the treatment effect per unit (*individual treatment effect)*

$Y(0)$

Treatment group

$$\hat{\tau}_{i,1} = Y_i - \boxed{\hat{\mu}_0(\mathbf{x}_i)}$$

Estimated from control

Control group

$$\hat{\tau}_{i,0} = \boxed{\hat{\mu}_1(\mathbf{x}_i)} - Y_i$$

Estimated from treated

3. Learn two separate models $\hat{\tau}_1(\mathbf{x}_i)$ (only treated) and $\hat{\tau}_0(\mathbf{x}_i)$ (only control)
4. The final estimator is a weighted average where $g(\mathbf{x}) : \mathcal{X} \rightarrow [0,1]$

$$\hat{\tau}(\mathbf{x}) = g(\mathbf{x}_i)\hat{\tau}_1(\mathbf{x}_i) + (1 - g(\mathbf{x}_i))\hat{\tau}_0(\mathbf{x}_i)$$

Doubly robust methods: Augmented Inverse probability weighting (AIPW)

- Doubly robust, we can either estimate in an **unbiased way:**
  - Propensity scores $\hat{\pi}(\mathbf{x}_i)$
  - S-learner (outcome model) $\hat{\mu}(t_i, \mathbf{x}_i) \approx y_i$

$$\hat{\text{ATE}}_{S-learn} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1,\mathbf{x}_i) - \hat{\mu}(0,\mathbf{x}_i)$$

$$\hat{\text{Adj}}_{S-learm} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\hat{\pi}(x_i)}(Y_i - \hat{\mu}(1,x_i)) - \frac{1-T_i}{1-\hat{\pi}(x_i)}(Y_i - \hat{\mu}(0,x_i))$$

$$\hat{\text{ATE}}_{AIPW} = \hat{\text{ATE}}_{S-learn} + \hat{\text{Adj}}_{S-learn}$$

Missing data (very briefly)

- **Typical approaches in practice (depending on the assumptions):**
  - Remove all samples with a missing feature (listwise deletion)
  - Ignore the problem and use the non-missing features of all samples
  - Impute the missing values

Missing data (very briefly)

- **Typical approaches in practice (depending on the assumptions):**
  - Remove all samples with a missing feature (listwise deletion)
  - Ignore the problem and use the non-missing features of all samples
  - Impute the missing values

- **Typical assumptions:**
  - R_X is an indicator variable that is 0 if X is missing and 1 otherwise
  - Missing completely at random (MCAR): $R_X \perp\!\!\!\perp \mathbf{X_V}$
  - Missing at random (MAR): $R_X \perp\!\!\!\perp X \,|\, \mathbf{X_V} \setminus \{X\}$
  - Missing not at random (MNAR) - anything else

Missing at random (MAR)

- Missing completely at random (MCAR): coin toss, quite unrealistic

- **Missing at random (MAR):** missing at random given the completely observed (not missing) variables
  - Similar to **ignorability/unconfoundedness**
  - Imputation with EM
  - Multiple imputation (Rubin 1987) - impute m datasets, analyse, combine
    - (Augmented) IPW can be used to analyse/estimate ATE of each dataset

  - See https://scikit-learn.org/stable/modules/impute.html#impute, http://juliejosse.com/wp-content/uploads/2018/07/LectureNotesMissing.html