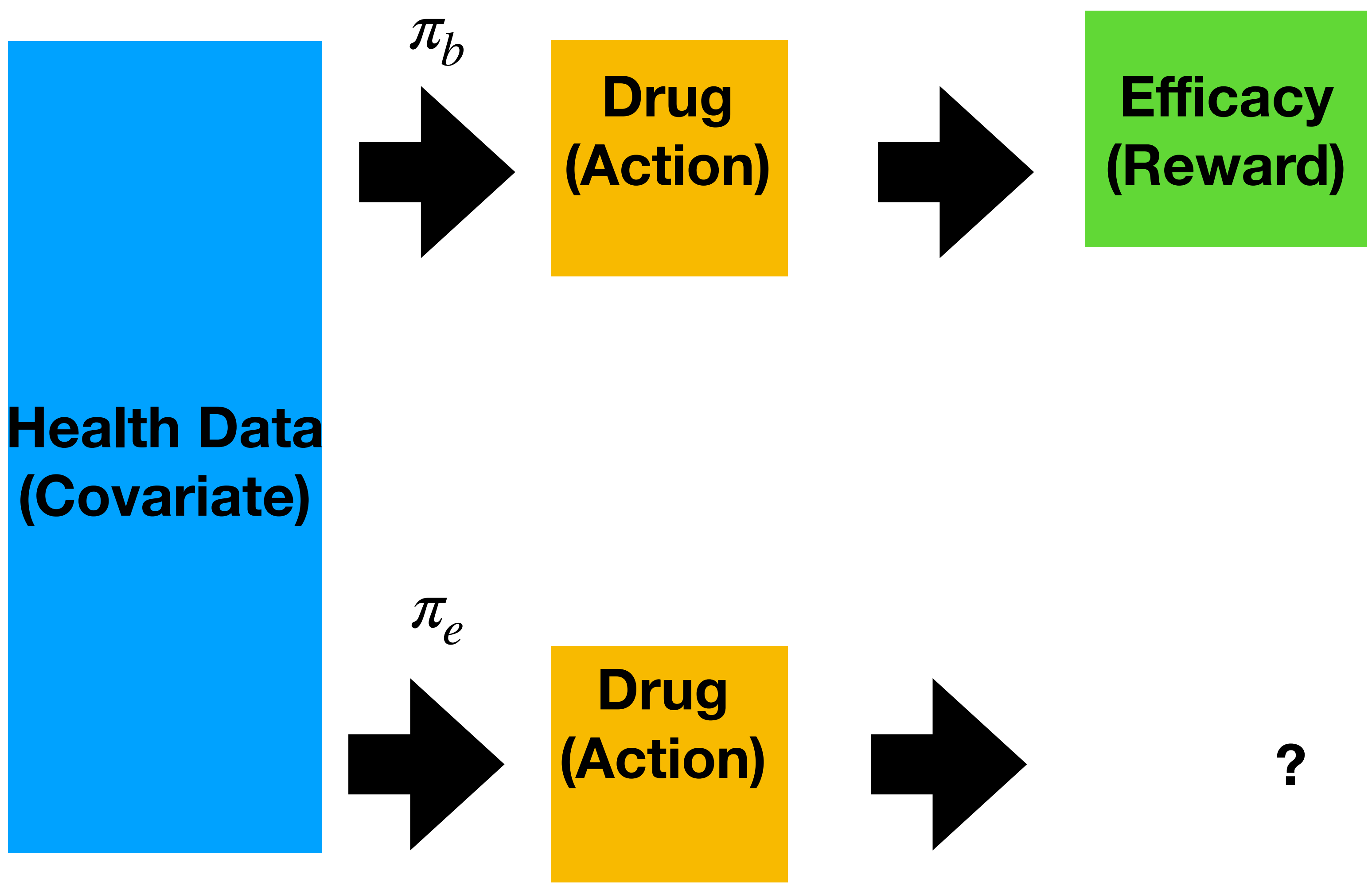
# Optimal Off-Policy Evaluation from Multiple Logging Policies

Nathan Kallus/ Yuta Saito / Masatoshi Uehara



# Off-policy Evaluation

- The goal is estimating a policy value of the evaluation policy from logged data .

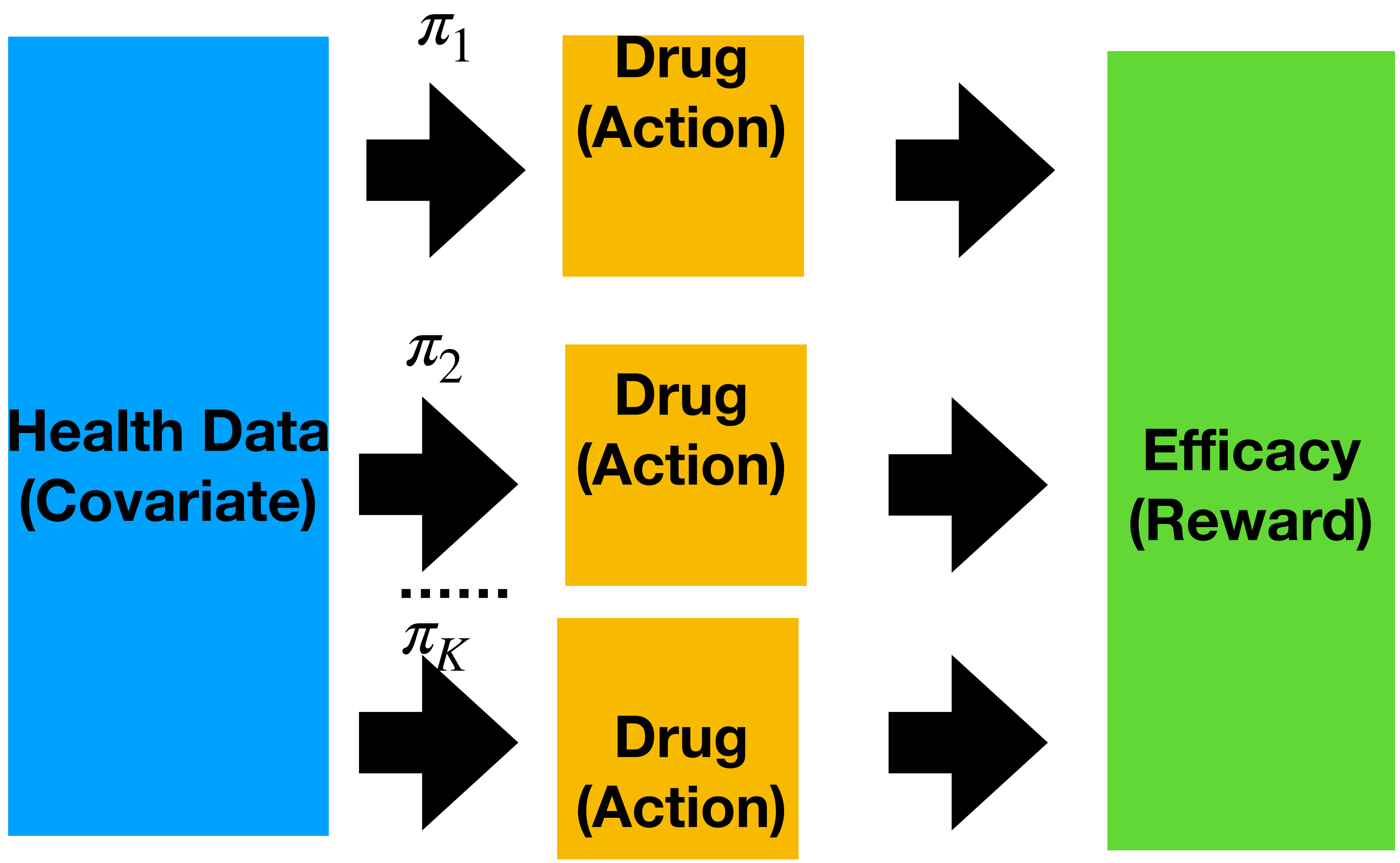



**We want to evaluate** $\mathrm{E}_{a \sim \pi_e(s)}[r]$ **from**

$$\{S_i, A_i, R_i\}_{i=1}^{n} \sim p(s)\pi_b(a \mid s)p(r \mid s, a) \text{ .}$$

# Motivation

- Our goal is estimating the policy value from multiple data sets.




- For each $1 \leq k \leq K$, we have K datasets: $\{(S_i, A_i, R_i)\}_{i=1}^{n_k} \sim p_S(s)\pi_k(a \mid s)p_{R|S,A}(r \mid s, a)$.

**We want to evaluate** $\mathrm{E}_{a \sim \pi_e(s)}[r]$

# Existing Estimators

- Agarwal et. 2017 proposed two estimators.

- IS estimators:

$$\hat{J}_{\text{IS}} = \hat{\mathrm{E}}_{a \sim \pi^*(s)} \left[ \frac{\pi_e(a \mid s) r}{\pi^*(a \mid s)} \right], \pi^*(a \mid s) = \sum_{k=1}^{K} \frac{n_k}{n} \pi_k(a \mid s)$$

- IS estimators 2:

$$\hat{J}_{\text{IS-PW}} = \sum_{k=1}^{K} \lambda_k \hat{\mathrm{E}}_{a \sim \pi_k(s)} \left[ \frac{\pi_e(a \mid s) r}{\pi_k(a \mid s)} \right] \text{ s.t. } \sum_{k} \lambda_k = 1$$

**Which is better?**

# Summary

- Propose a new class including estimators in Agarwal et. 2017. Then, calculate the lower bound of MSEs among the class.

- Show how to construct an estimator achieving this bound asymptotically under mild assumptions. This estimator has a doubly-robust property.

# New Class

- We use weights $h_k(s, a)$ depending on the strata so that it satisfies $J = \mathrm{E}[\hat{J}]$:
$$\hat{J} = \sum_{k=1}^{K} \hat{\mathrm{E}}_{a \sim \pi^*}[h_k(s, a)\pi_e(a \mid s)\{r - g(s, a)\} + g(s, \pi_e)].$$

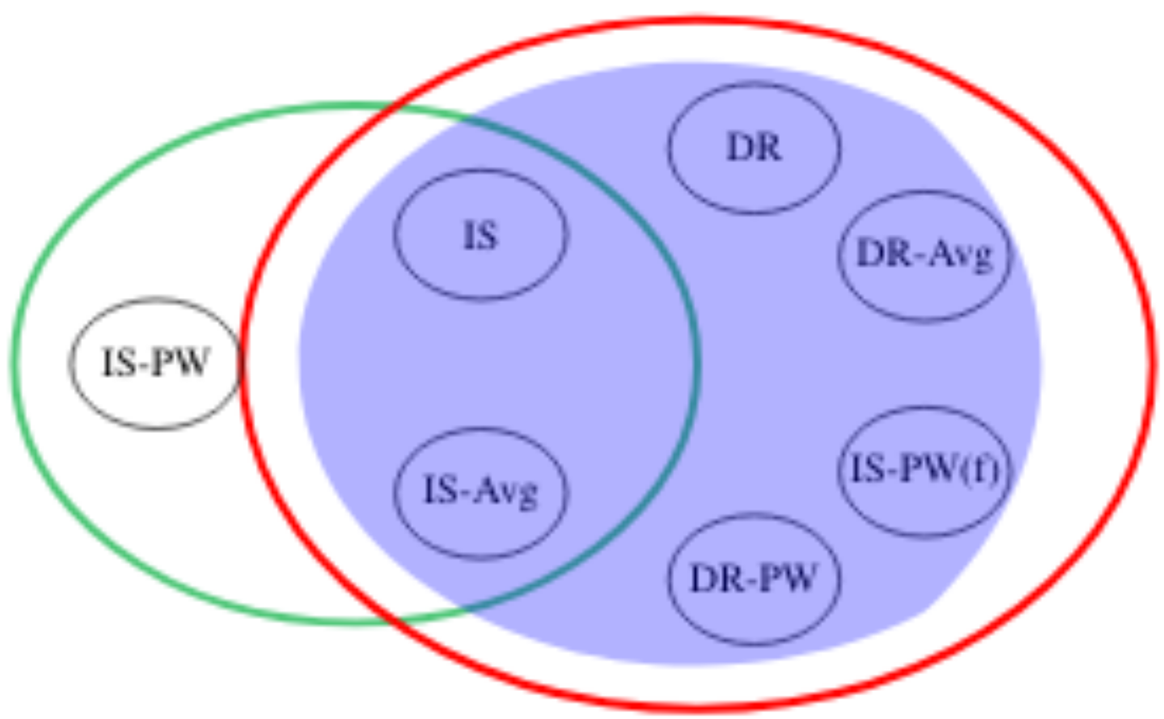
- Optimal among the above class when $h_k = 1/\pi^*, g = q, q(s, a) = \mathrm{E}[r \mid s, a]$.

- I.E. $$\hat{J}_{\mathrm{DR}} = \hat{\mathrm{E}}_{a \sim \pi^*(s)}\left[\frac{\pi_e(a \mid s)(r - \hat{q}(s, a))}{\hat{\pi}^*(a \mid s)} + \hat{q}(s, \pi_e)\right].$$

- The above has a DR property.

# Optimality

- Reg: Regular estimators.
- Blue: A new class.




-

# I.I.D vs Stratified

**Is the case with multiple logging polices different from the case with a single logger?**

**Let $n_1, \cdots, n_K$ be each sample size in K data sets.**

- $n_1, n_2, \cdots, n_K$ are fixed.
* $n_1, n_2, \cdots, n_K$ are random.



**Stratified (Our case)**

**I.I.D sampling forom a mixture policy**

# Other topics

- Extension to More Robust Doubly Robust estimators: Improved version of DR estimator.
- Extension to Reinforcement Learning Cases