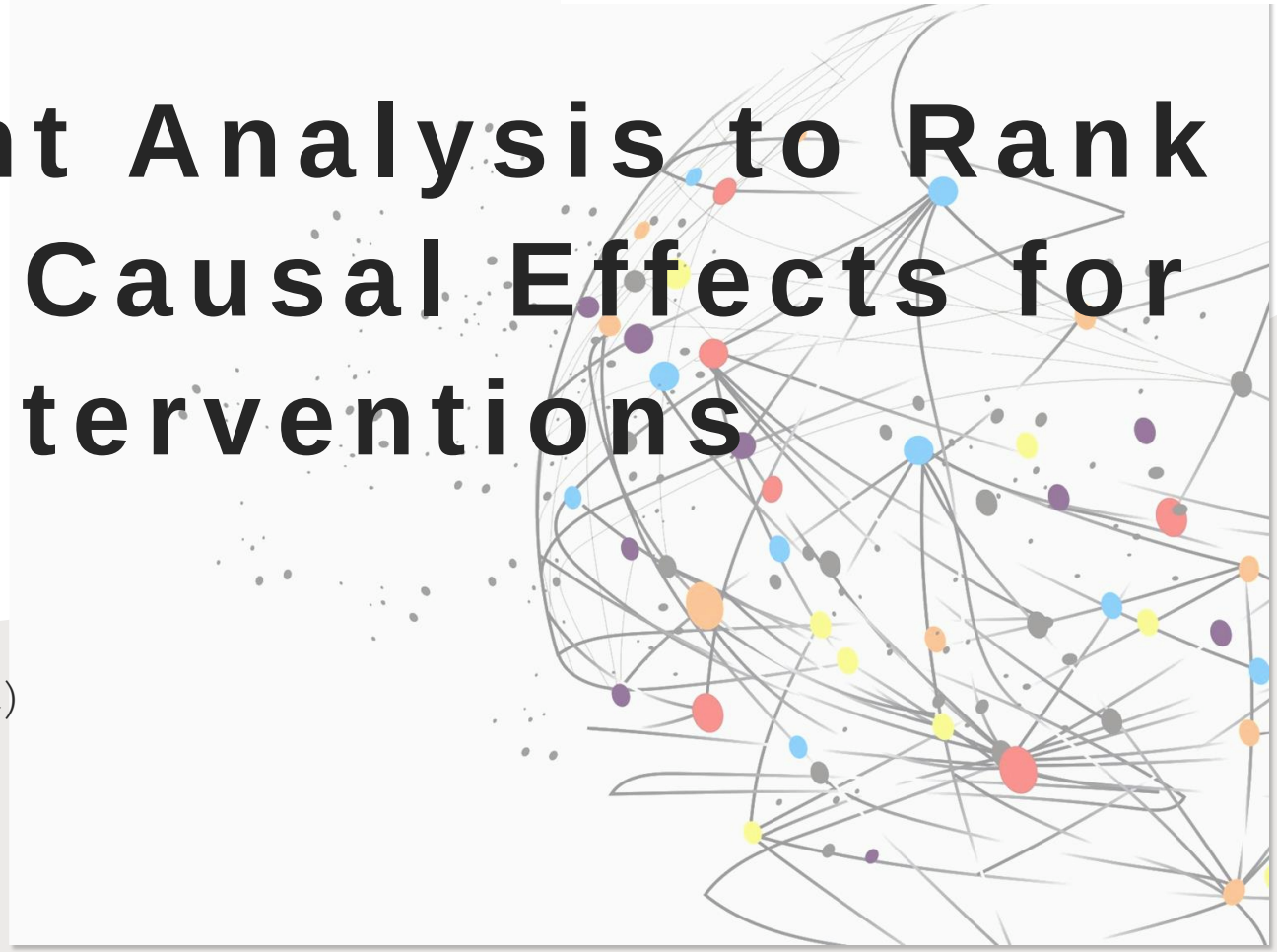
MARCH 8-12, 2021 • JERUSALEM, ISRAEL

Split-Treatment Analysis to Rank Heterogenous Causal Effects for Prospective Interventions

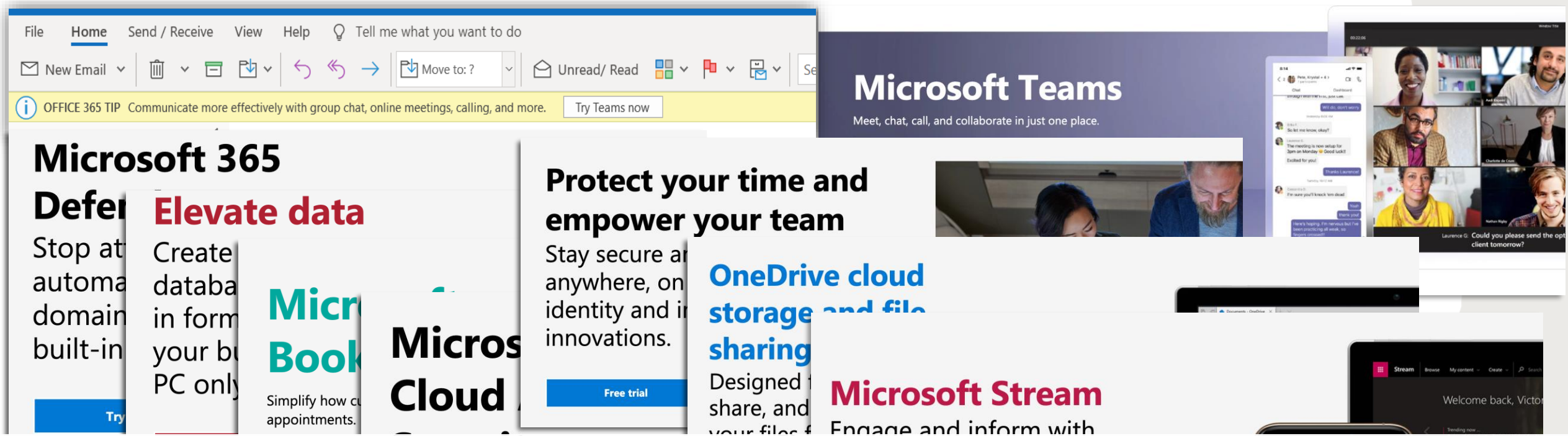


Yanbo Xu (Georgia Tech)
Divyat Mahajan (Microsoft Research India)
Liz Manrao (Microsoft)
Amit Sharma (Microsoft Research India)
Emre Kıcıman (Microsoft Research AI)

Outline

- Motivation
- Identification by **Split-Treatment**
- An Analysis Pipeline using **Split-Treatment**
- Experiment and Results

Motivation



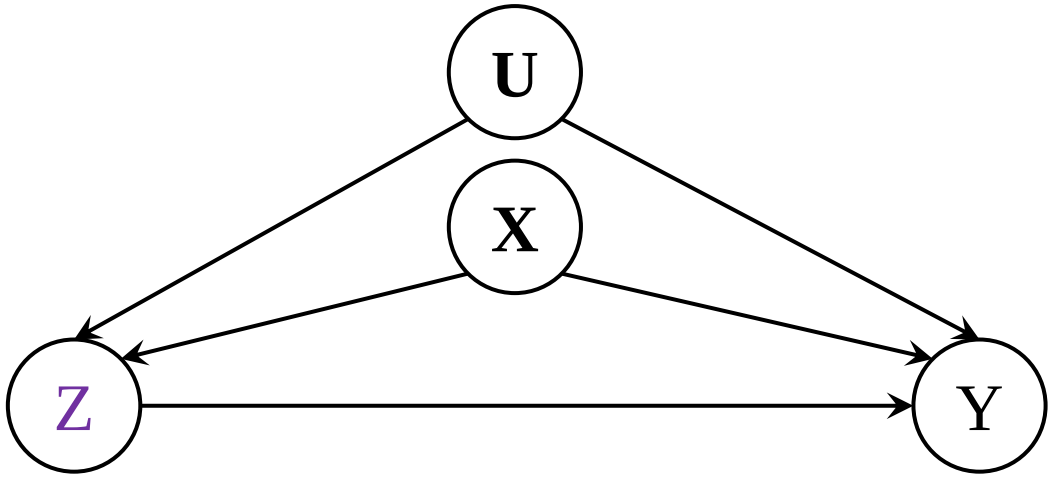
- ❑ So many features to recommend!
- ❑ Not all such messages are useful for every individual!
- ❑ Unaffordable or detrimental to run active experiments on all of them!

➤ **Split-Treatment!**

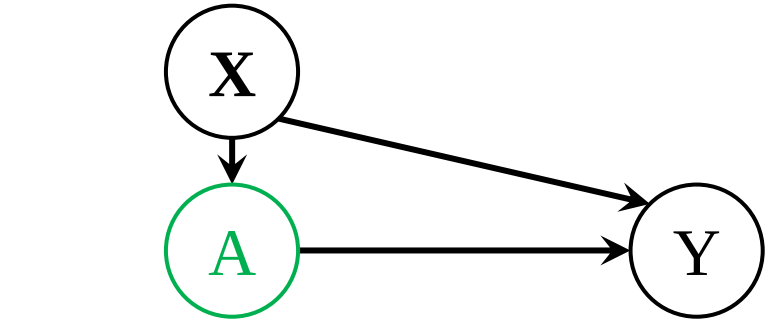
Use logged behavioral data to identify who are likely to benefit from a novel intervention.

Identification of **Split-Treatment**

Prospective Experimental data
(Expensive, not available)



Observational data
(Cheap, massive, available)



The ultimate goal is to estimate the effect of a prospective treatment Z, but it is not observed.

➡ Alternatively, **Split-Treatment** estimates the effect of a proxy treatment A.

➡ Under proper assumptions, the ranking of the two heterogeneous effect can be aligned.

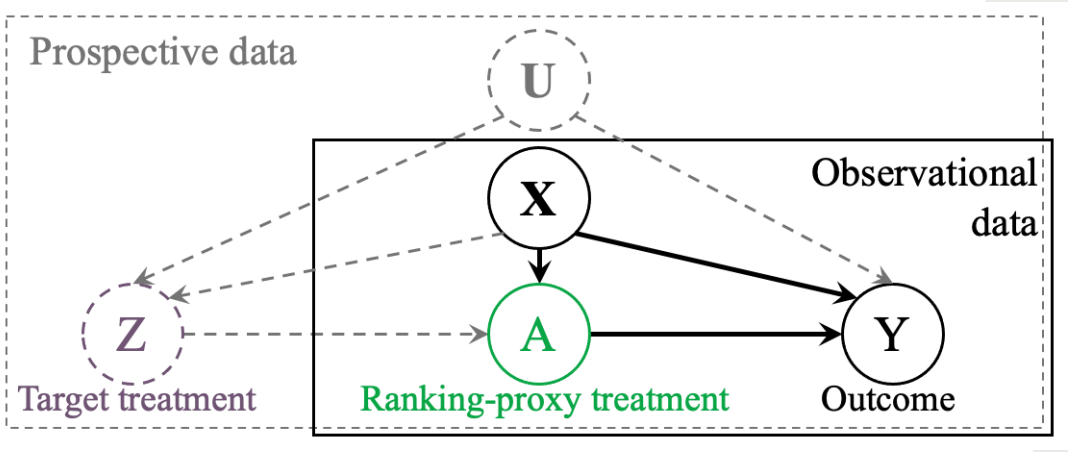
Identification of **Split-Treatment**

- **Assumption1 (Ignorability):** $P(Y|do(a), \boldsymbol{x}) = P(Y|a, \boldsymbol{x})$

$$\begin{aligned}\mathrm{ITE}^{(z)}(\boldsymbol{x}) &= \mathbb{E}[Y|do(z=1), \boldsymbol{x}] - \mathbb{E}[Y|do(z=0), \boldsymbol{x}] \\ &= \underbrace{\big(P(a=1|z=1, \boldsymbol{x}) - P(a=1|z=0, \boldsymbol{x})\big)}_{\mathrm{Compliance}(\boldsymbol{x})} \\ &\quad \cdot \underbrace{\big(\mathbb{E}[Y|a=1, \boldsymbol{x}] - \mathbb{E}[Y|a=0, \boldsymbol{x}]\big)}_{\mathrm{ITE}^{(a)}(\boldsymbol{x})}.\end{aligned}$$

- **Assumption2 (Compliance):** $\mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\big[\mathrm{Compliance}(\boldsymbol{x})\big] > 0$

$$\begin{aligned}\mathrm{CATE}_{\mathcal{G}}^{(z)} &= \mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\big[\mathrm{ITE}^{(z)}(\boldsymbol{x})\big] \\ &\propto \mathbb{E}_{\boldsymbol{x} \in \mathcal{G}}\big[\mathrm{ITE}^{(a)}(\boldsymbol{x})\big]\end{aligned}$$



- Z: Prospective treatment
- A: Proxy treatment
- Y: Outcome
- X: Observed Confounder
- U: No-unmeasured unobserved Confounder

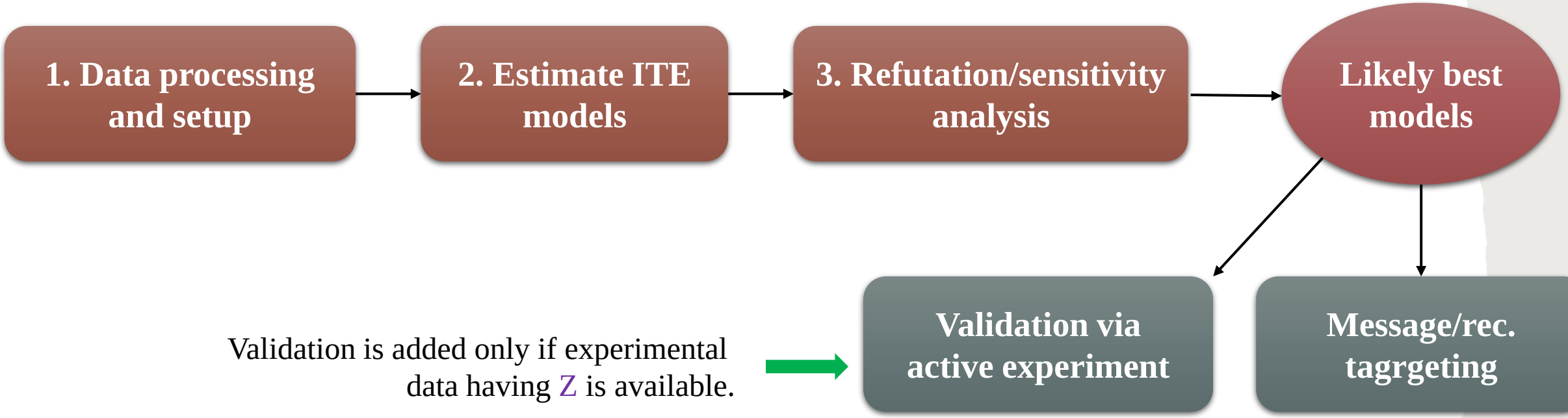
We pick a proxy treatment A such that:

➡ A exists, with some natural variation, in our observational logs.

➡ The effect of Z on Y should be mediated through A.

Estimation using **Split-Treatment**

- An end-to-end analysis pipeline of using **Split-Treatment** in feature/product recommendation.



Estimation using **Split-Treatment**

Inverse probability of treatment weighting (IPTW): Predict propensity scores to reweight individual outcome estimates and obtain unbiased ITE.

- Given the observational data $\mathrm{D}^{\mathrm{n}} = \{(x_i, a_i, y_i)\}$, we learn outcome function f by minimizing the following loss:

$$\sum_{i=1}^{n} w_i \mathcal{L}(y_i, f(x_i, a_i)),$$

with stabilized IPTW: $w_i = a_i \cdot \frac{P(a_i = 1)}{\hat{P}(a_i = 1|\boldsymbol{x}_i)} + (1 - a_i) \cdot \frac{1 - \hat{P}(a_i = 1)}{1 - \hat{P}(a_i = 1|\boldsymbol{x}_i)}.$

$$\hat{\mathrm{ITE}}^{(a)}(\boldsymbol{x}) \equiv f(\boldsymbol{x}, a = 1) - f(\boldsymbol{x}, a = 0).$$

Estimation using **Split-Treatment**

We use sensitivity analyses to eliminate unreliable models in the absence of experimental validation.

- *Placebo* test

  Place a random variables as the treatment A

  ➡ Test if an estimator returns zero causal effect.

  ➡ Prune out those estimators that show significant non-zero causal effect.

- *Unobserved-confounding* test

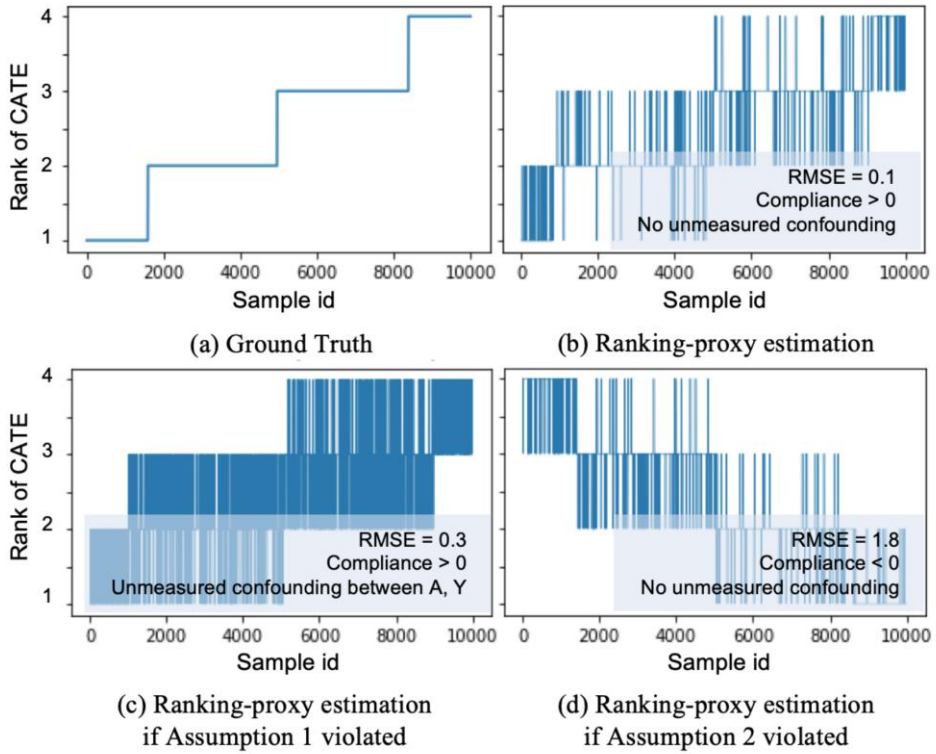
  Add a new confounder to the feature set with varying degrees of its effect on A and Y.

  ➡ Test if an estimator is less sensitive to the varying degrees of effect of the new confounder.

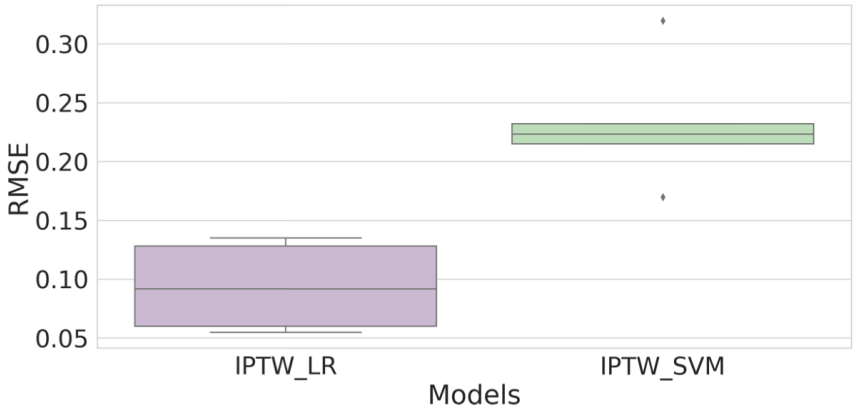
  ➡ Prune out those estimators that are sensitive to such changes

Experiment And Results (Simulation)

- Simulation results



Violation of the two Assumptions: Comparison between the ground-truth rank and the proxy-estimated rank in simulations with or without violation of the assumptions.



IPTW-LR is less sensitive to unobserved confounding. Box plots are for 5 runs with different degrees of confounding.

Unobserved-confounding analysis: Comparison between estimated causal effect with and without unobserved confounding, for two causal models.

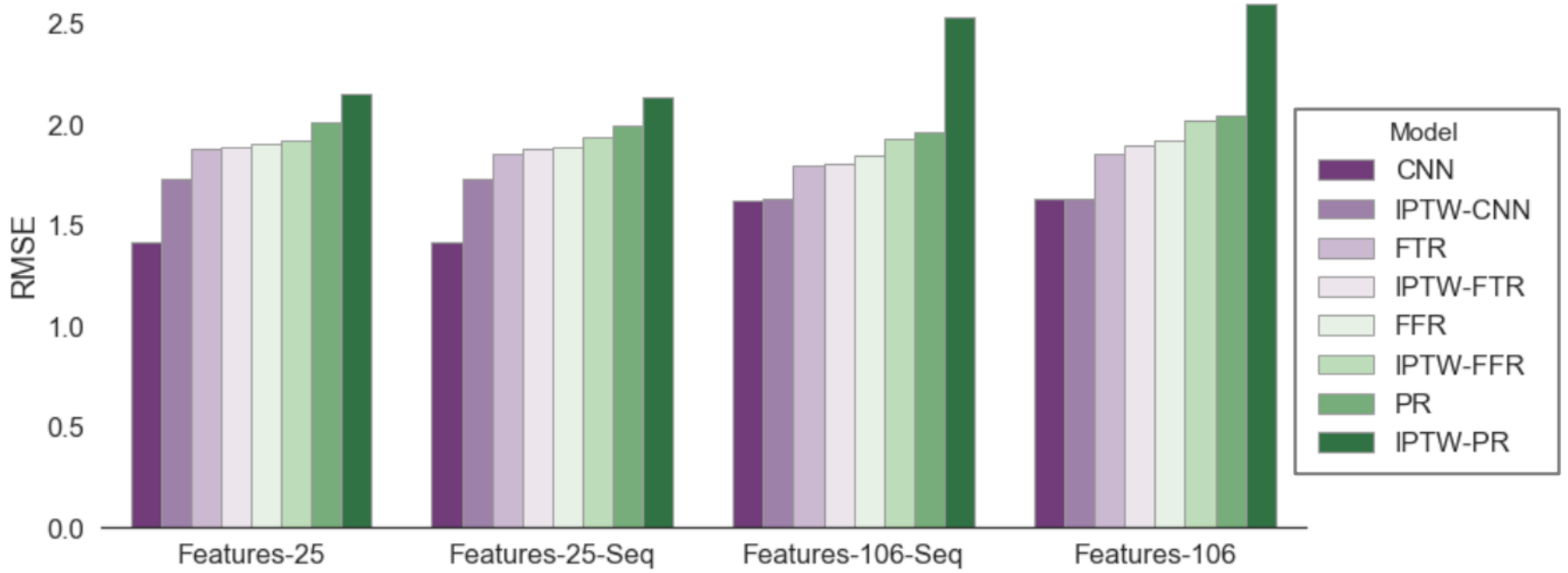
Experiment And Results (Real data)

- Real-world data: Product recommendation in a large software ecosystem.

- Experimental setup:
  - Treatment window: 3/29/2019 – 4/27/2019 (1 month) ➡ Extract proxy treatment assignments.
  - Pre-Treatment window: 3/22/2019 – 3/28/2019 (1 week) ➡ Extract confounding factors.
  - Post-Treatment window: 4/28/2019 – 5/24/2019 (4 weeks) ➡ Extract outcome measures.
- Split treatment:
  - Proxy treatment A: 1st use of the product in Treatment window.
  - Outcome Y: Sustained usage of the recommended product in Post-treatment window
- Data description:
  - Observational data: 2.2M
    - 2.3% used the product in the Treatment window
  - Experimental data (Timeline aligned): 1.1M
    - 66.1% exposed to Z, 5.7% vs. 5.3% used the product from the exposed vs. unexposed group

Experiment And Results (Real data)

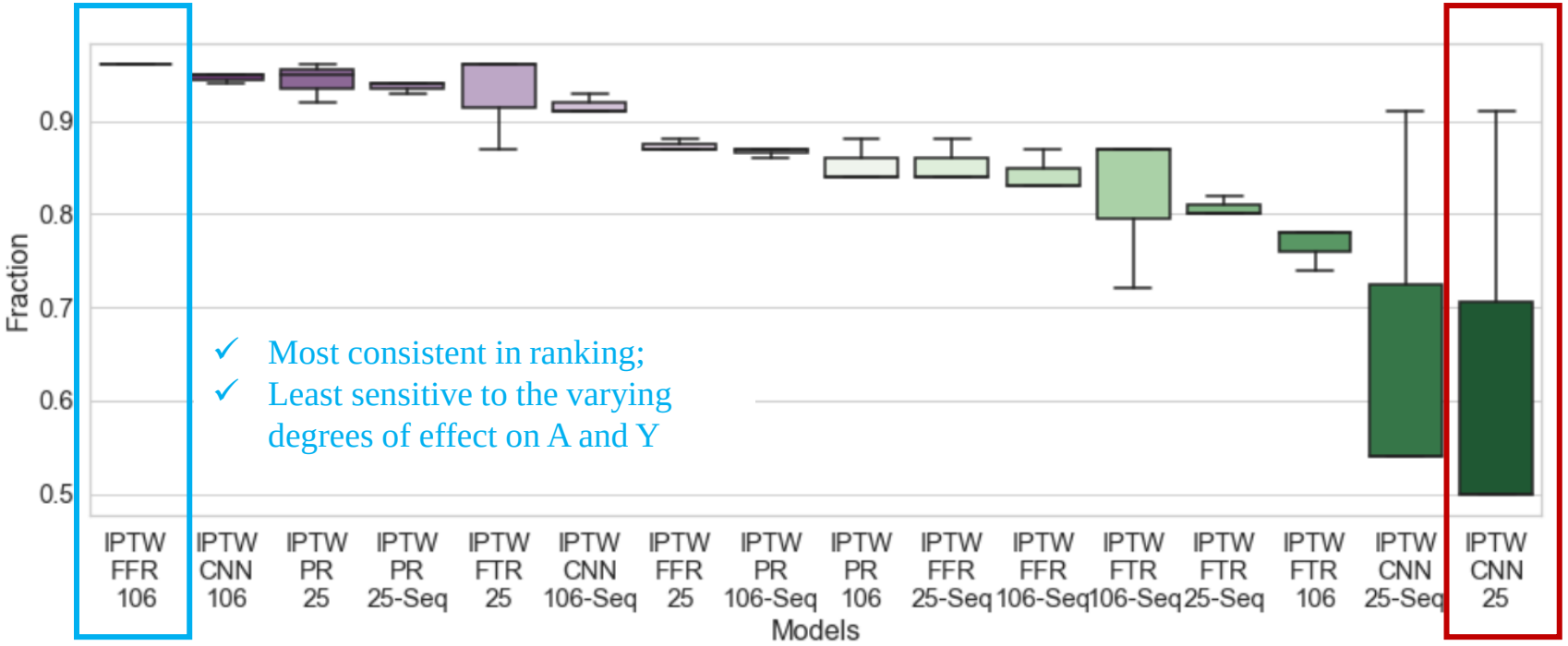
- RMSE of outcome prediction from the baseline models.



- FTR (Fast-Tree Regression): An efficient tree regression with gradient boosting
- FFR (Fast-Forest Regression): An efficient random forest regression using the Fast-Tree learners.
- PR (Poisson Regression): A linear regression with respect to minimizing Poisson loss instead of mean squared errors.
- CNN (Convolutional Neural Network): A 2-layer 1-D convolutional network with Poisson loss.
- Features-25 and Features-106: two static feature sets
- Features-25-Seq and Features-106-Seq: two aggregated sequential feature sets

Experiment And Results (Real data)

- Sensitivity analysis (unobserved confounding)



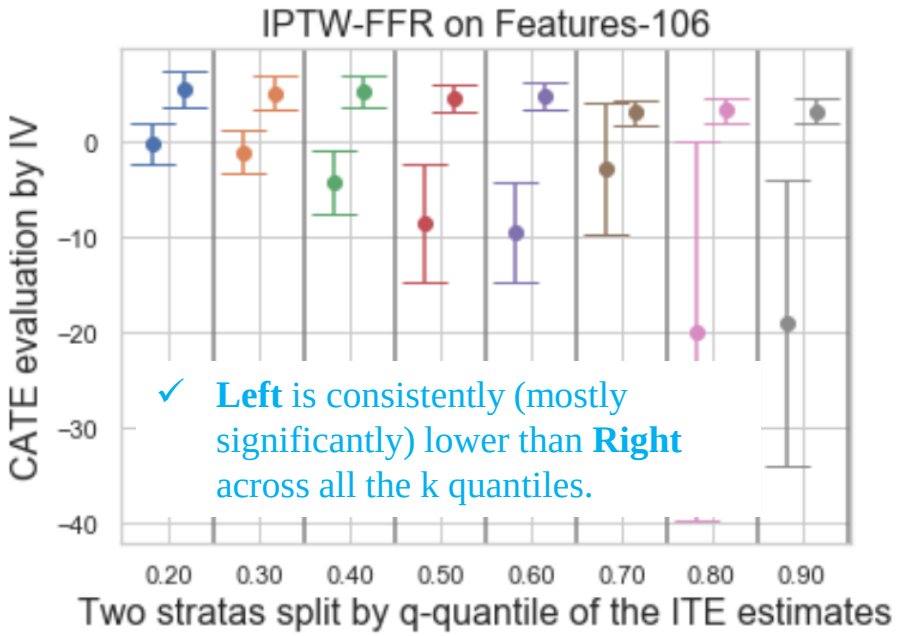
- Least consistent in ranking;
- Most sensitive to the varying degrees of effect on A and Y

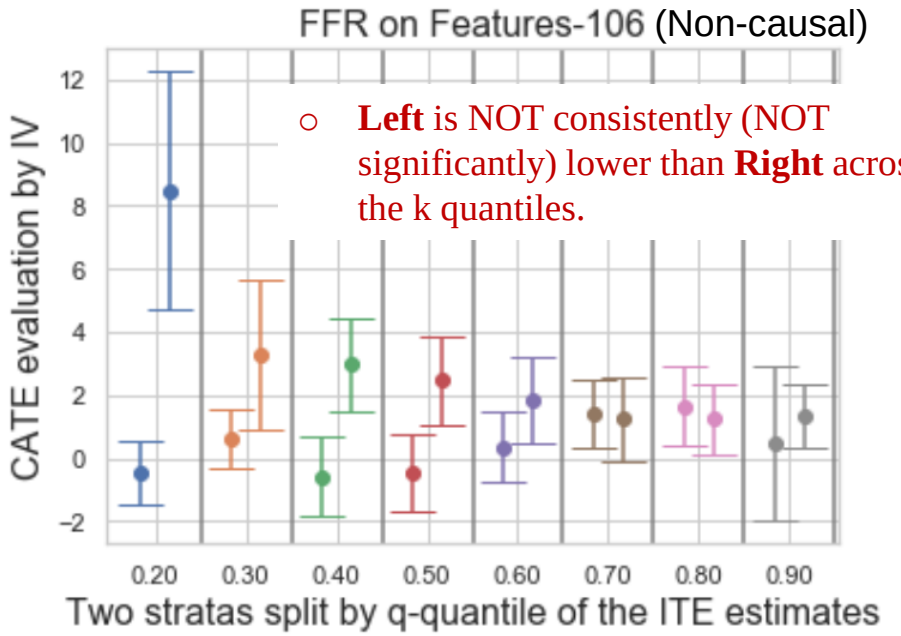
Fraction of the top 50-percentile individuals that remain in the top 50-percentile after adding an observed confounder. Box plots are for 3 runs with different degrees of confounding.

Experiment And Results (Real data)

- Validation on experimental data

Best model (IPTW-FFR) picked by the sensitivity analysis





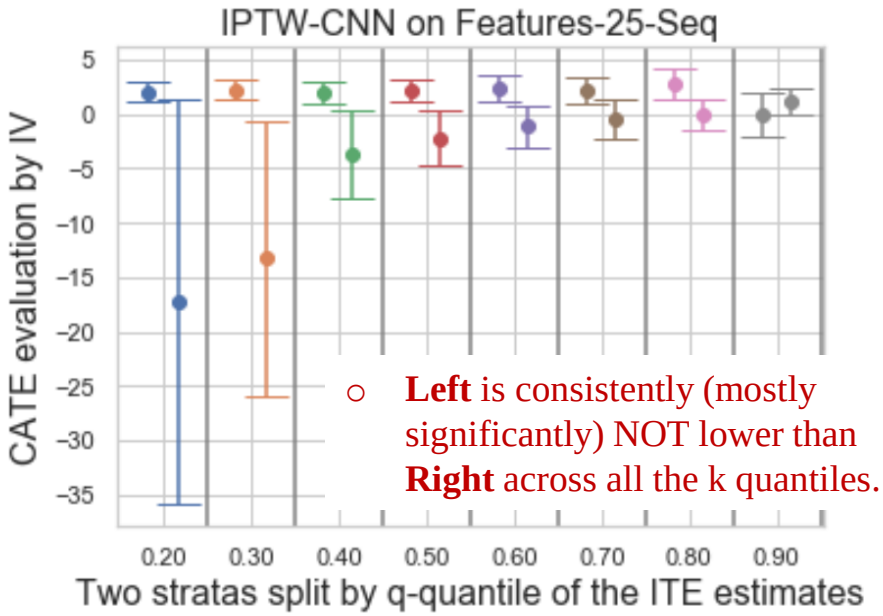
- Group ITE estimates by k quantiles and stratify each group into *low* (**left**) and *high* (**right**) subgroups.
- Given the subgroup assignment, run IV analysis on the experimental data.
- Use the ground-truth CATE (from IV analysis) to validate:

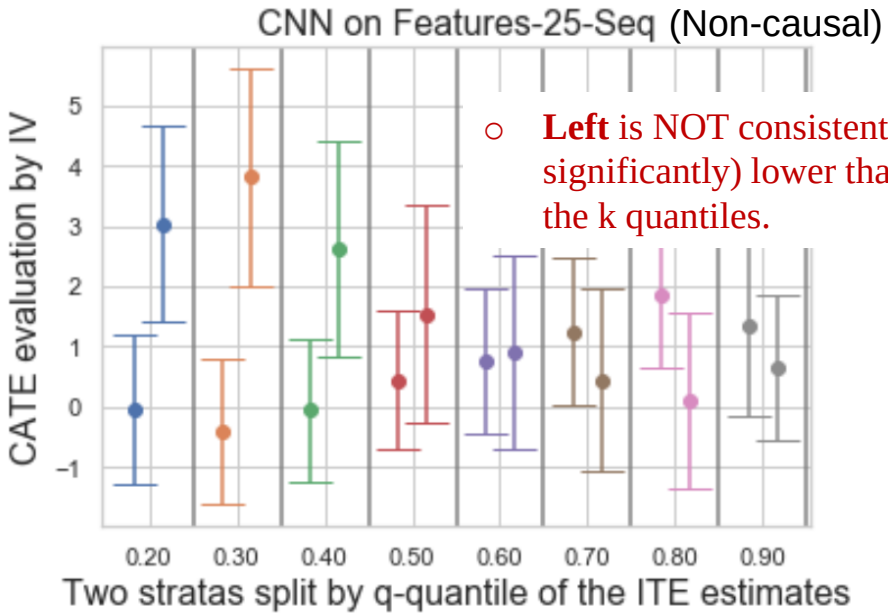
  the *low* (**left**) should be consistently **lower** than the *high* (**right**) subgroup across all the k groups!

Experiment And Results (Real data)

- Validation on experimental data

Worst model (IPTW-CNN) picked by the sensitivity analysis





- Group ITE estimates by k quantiles and stratify each group into *low* (**left**) and *high* (**right**) subgroups.
- Given the subgroup assignment, run IV analysis on the experimental data.
- Use the ground-truth CATE (from IV analysis) to validate:

  the *low* (**left**) should be consistently **lower** than the *high* (**right**) subgroup across all the k groups!

Conclusion

- We presented a practical, observational analysis pipeline for
  - Identifying individuals likely to benefit from a novel treatment Z
  - Using proper causal analysis of existing logs that contain proxy treatment A
- A key contribution:
  - Refutation tests and sensitivity analyses enable a principled a priori identification of the feature selection and elimination of unreliable algorithmic design.
- We validated our analysis with an A/B experiment in a large real-world setting.

THANK YOU!