General Transportability of Soft Interventions: Completeness Results

Juan D. Correa <jdcorrea@cs.columbia.edu>
Elias Bareinboim <eb@cs.columbia.edu>

Columbia University
IN THE CITY OF NEW YORK

Motivation

- Establishing the effect of new interventions/policies from data is a pervasive task across the empirical sciences.
- Controlled experimentation is considered the gold standard to learn such causal effects. However, experiments rarely generalize to domains outside where it was originally performed. This problem is studied in the literature under the rubric of transportability.
- In this work, we investigate the conditions under which transportability of policies (or soft interventions) can be justified from a collection of heterogenous data.

Observation: All data is not created equal…

- Heterogenous datasets are pervasive. They could …

(1) have different **experimental conditions**, → Surrogate Experiments

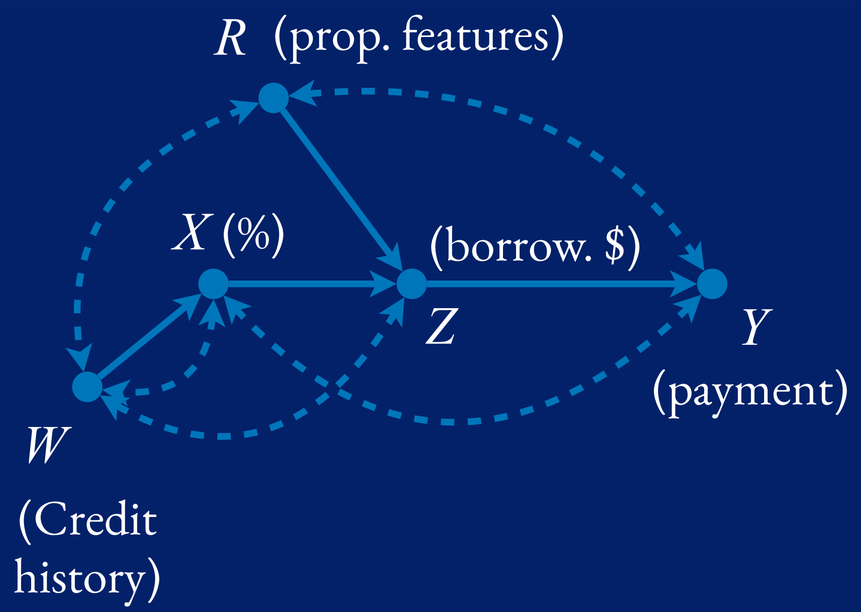
(2) come from different underlying **populations**, → Transportability

(3) suffer from non-random **sampling mechanisms**, → Sample Selection Bias

(4) **measure** different sets of variables. → Partial Observability

Government-backed loan program example

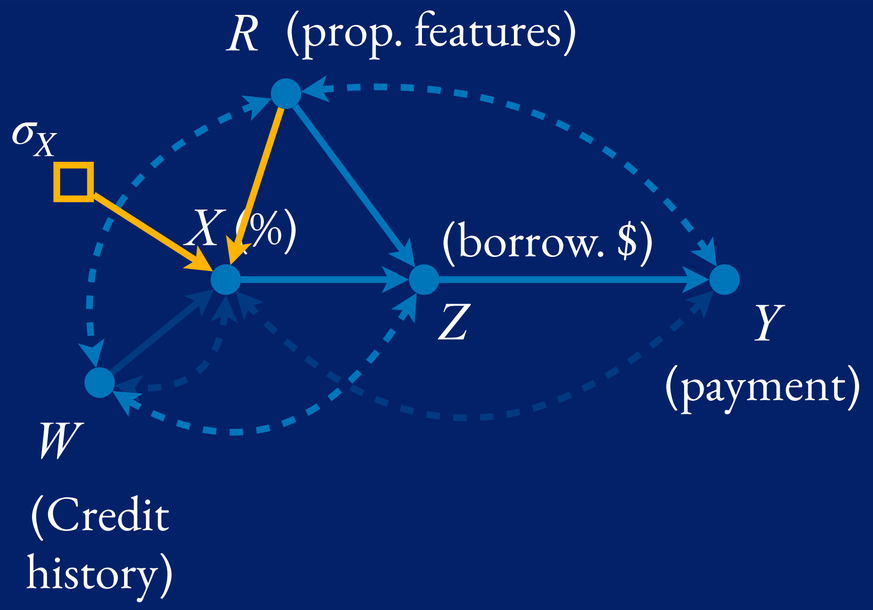
- Depending on credit history (W), loaners are allowed to borrow a maximum percentage (X) of the value of the property the want to purchase.
- Based on the property's features (R), the loan may be approved or not for amount (Z).
- Loan is deemed satisfactory (Y) if it is being paid back diligently.



Model of the Data Generating Process

Government-backed loan program example

- To increase the number of home owners, city π^* is considering a new policy σ_X that relaxes credit history requirements (W) on the allowed percentage (X), for homes purchased in certain areas of the city (R).
- Based on the causal model (to the right), the effect of policy σ_X on Y can be proved not identifiable from the observational dataset.

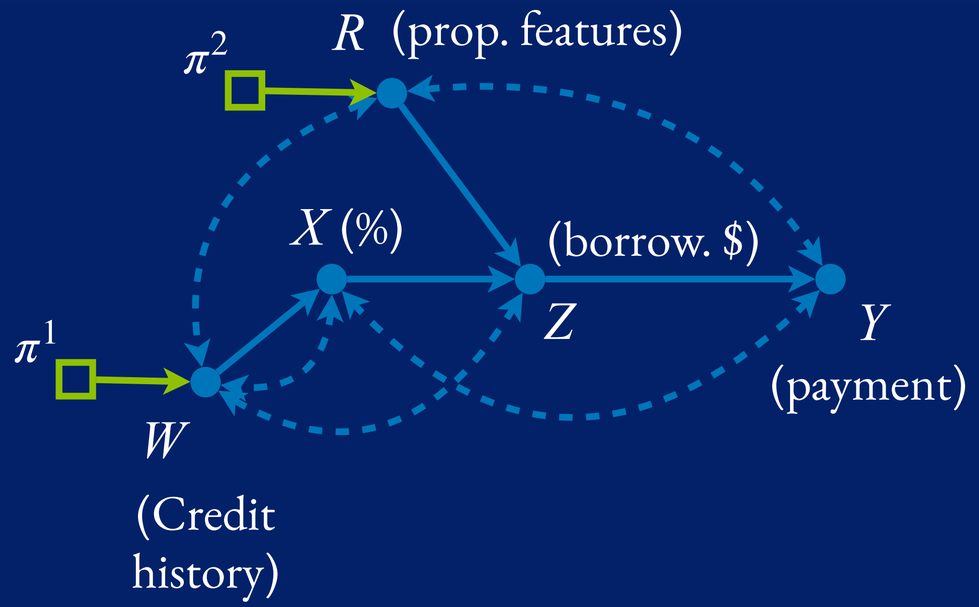


Model of the Data Generating Process **after** intervention

Government-backed loan program example

Solution: We'll try to leverage data from distinct, but somewhat similar cities.

- City π^1 is different in the sense that the credit history (W) of the average lender is better.
- City π^2 has properties that tend, on average, to be cheaper (e.g., R - price) than π^*.
- Similar policies σ_Z^1 and σ_W^2 were implemented in those cities.

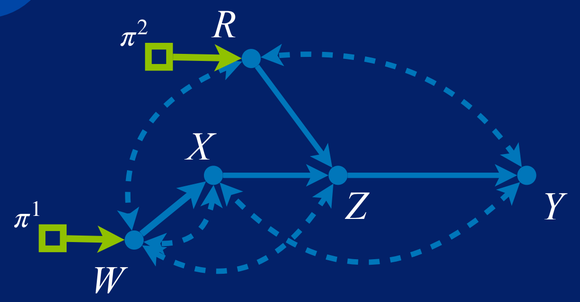


Model of the **differences** between Data Generating Processes in different populations

Transportability Task

- Based on the causal model, determine whether (and how) the available datasets can be combined to compute the effect of a new policy σ_X.

1 **Causal Assumptions**



Transportability

Is there a function f such that

$P^*(y; \sigma_x) = f(P_1, \ldots, P_k)$

→ yes (f) / no

2 **Data**

$$P^*(y; \sigma_X) = \sum_{r,x,z} \left(\sum_{x'} P^1(y \mid z, x', r; \sigma_Z) P^1(x' \mid r; \sigma_Z) \right) P^2(z \mid x, r; \sigma_W) P^*(x \mid r; \sigma_X^*) P^*(r)$$

experiment σ_Z^1 in π^1 | exp. σ_W^2 in π^2 | target σ_X^* | obs. in π^*

π^*

$P^*(\mathbf{V})$ $P^1(\mathbf{V}; \sigma_Z^1)$ $P^1(\mathbf{V}; \sigma_W^2)$

Contributions

1. Provided a necessary and sufficient graphical condition that characterizes the existence of an unbiased estimator for the effect of a target policy (possibly stochastic) given assumption in the form of a diagram and heterogeneous datasets.

2. Developed a complete algorithm to efficiently determine whether the transport formula exists, and an unbiased estimator of the corresponding transport formula (whenever it exists).

3. Prove that a set of inference rules, known as σ-calculus, are complete for this task.

Thank you!