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

Neural Information Processing Systems

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.

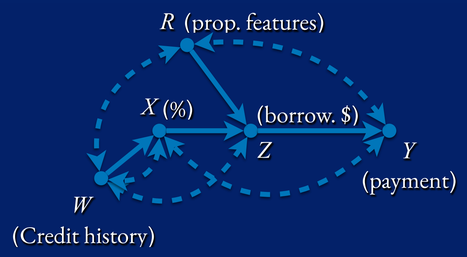
Heterogenous Data

More and more data is collected every day and still taking advantage of multiple and different datasets for any particular application is far from trivial, mainly, due to the heterogenous nature of the data. Different dataset could ...

(1) have different **experimental conditions**, → Surrogate Experiments
(2) come from different underlying **populations**, → Transportability
(3) suffer from non-random **sampling mechanisms**, → Selection Bias
(4) **measure** different sets of variables. → Partial Observability

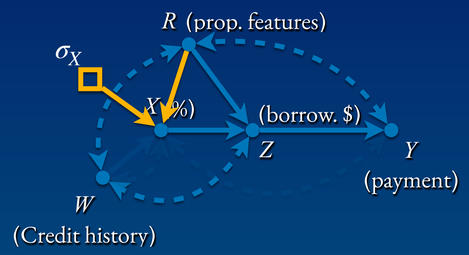
An Example: Government-backed loan program

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



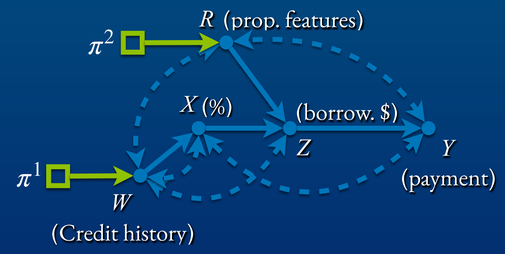
Model of the Data Generating Process

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



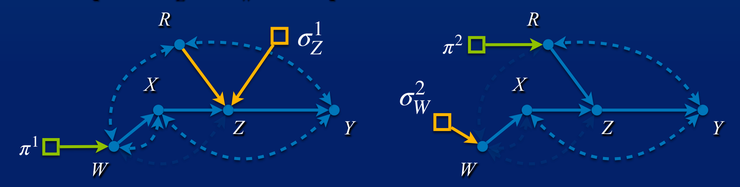
Model of the Data Generating Process **after** intervention

Solution: We will try to leverage data from distinct, but somewhat similar cities.
In city π^1, in average, lenders have a better credit history (W).
In city π^2, properties tend, on average, to be cheaper (e.g., R - price) than in π^*.



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

Similar policies σ_Z^1 and σ_W^2 were implemented in those cities.

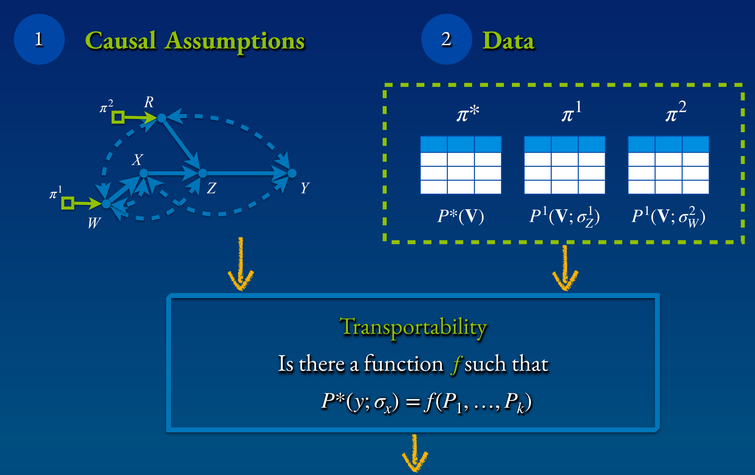


City π^1 implemented a policy σ_Z^1 where the amount Z was allocated based on X, R using a fixed formula.

City π^2 randomized (σ_W^2) credit history W for a group of loaners to study its effect on loan approval.

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.



$$P^*(y;\sigma_X)=\sum_{r,x,z}\Big(\sum_{x'}\underbrace{P^1(y\mid z,x',r;\sigma_Z)}_{\text{experiment }\sigma_Z^1\text{ in }\pi^1}P^1(x'\mid r;\sigma_Z)\Big)\underbrace{P^2(z\mid x,r;\sigma_W)}_{\text{exp. }\sigma_W^2\text{ in }\pi^2}\underbrace{P^*(x\mid r;\sigma_X^*)}_{\text{target }\sigma_X^*\text{ obs. in }\pi^*}P^*(r)$$

Results

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.

For more information, full paper and references visit
https://causalai.net/r68.pdf