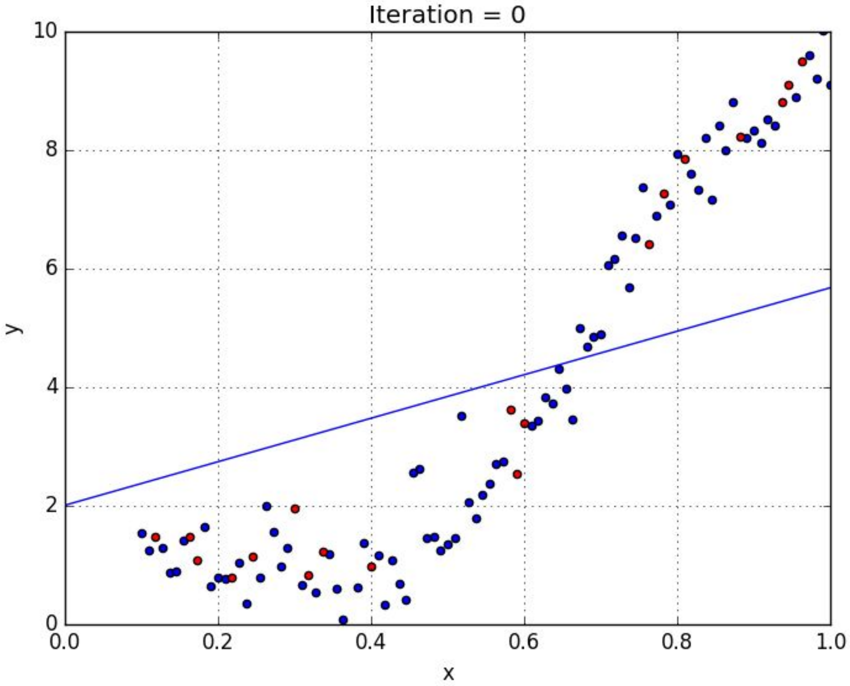
Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport

Adarsh Subbaswamy
Johns Hopkins University

Peter Schulam
Johns Hopkins University

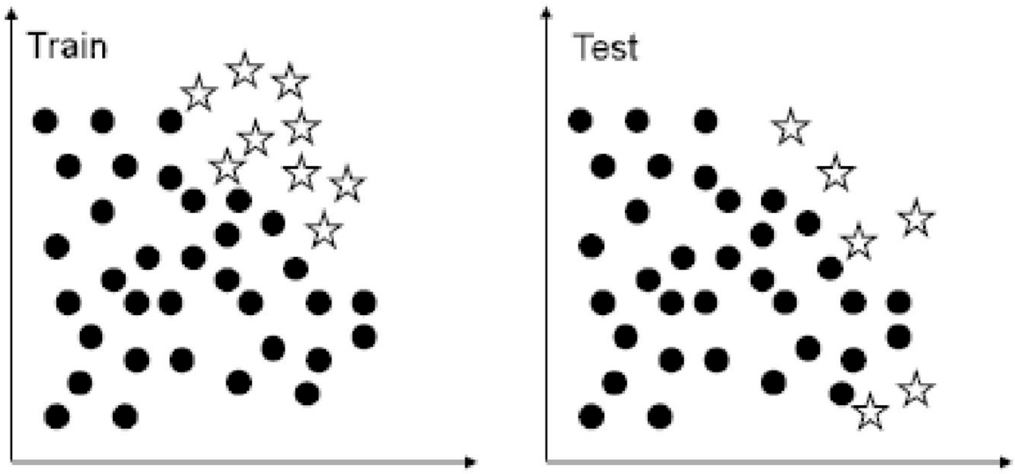
Suchi Saria
Johns Hopkins University

- **Goal**: ML models learn a functions
- **Train**: based on a bunch of example
- **Test:** evaluate the learnt function based on the generalization capabilities (test set accuracy)[1]



[1] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. "Machine learning basics." *Deep learning* 1.7 (2016): 98-164

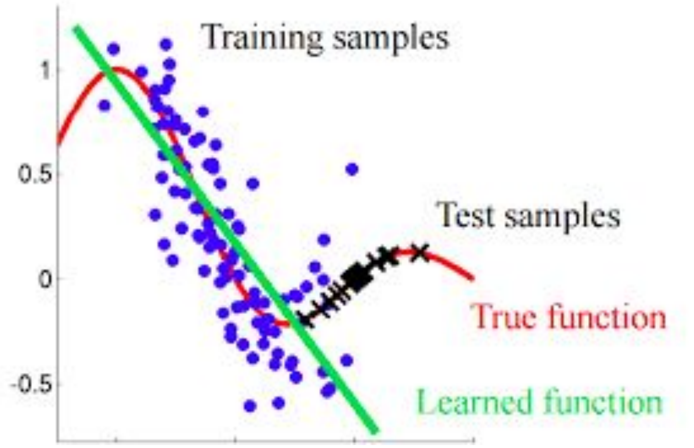
Definition 1. ***Dataset shift appears when training and test joint distributions are different.*** That is, when $P_{tra}(y, x) \neq P_{tst}(y, x)$



Why is DS a relevant problem?

DS affect the ***learned function so it will be different from True function*** because of the partial subset with different distribution

a) If the test set is not available is difficult to obtain good generalisation capability
b) If DS shift is present at least a relationship between train and test should be assumed to be sure that the model generalize.



Different Dataset Shifts[1]

- **Covariate shift**: distribution of input data shifts between the training environment and test environment $P(x_{train})$ not equal to $P(x_{test})$

- **Target shift:** the conditional distribution is the same but the marginal distribution $P(y)$ shifts between training and test

[1]Quinonero-Candela, Joaquin, et al. "Dataset shift in machine learning. Neural Information Processing." (2008).

Data generating process (DGP)

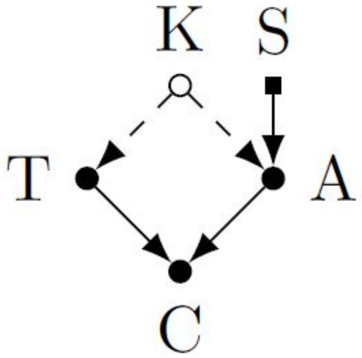
```
this is not
the classic
DAG !!!
```

Predict the **target** (T) based on the chest pain (C) and the possibility of not that patient take aspirin (A).

Moreover **Smoking** (unrecorded variable K) causes lung cancer but also heart disease (aspirin needed

In the picture on the right this situation is represented.

- **Why would the DGP vary?** → multiple prescription policies for aspirin to smokers

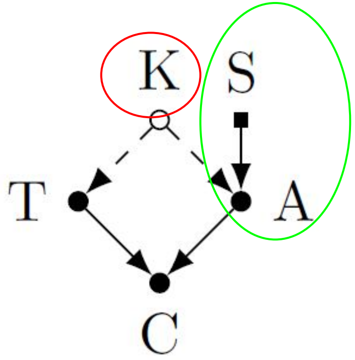



```
Unobserved
(K) and
mutable
variables(S)
```

From DAG to Selection Diagram

1) DAG: **relevant variables** observed or not, like smoking (K in slide 5)

2) **Auxiliary Selection variables**: variables that originates the uncertainty. For example **S=PRESCRIPTION POLICY** → **P(A|K) = uncertain (it varies) A** is a mutable variable cause we expect that the underlying process is not always the same



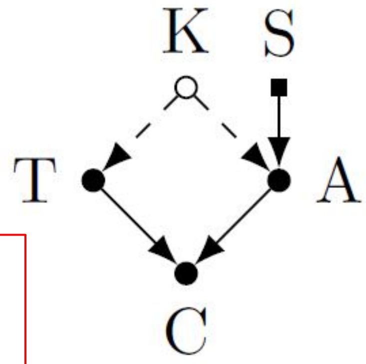
Selection Variables and model stability

- In other words we need predictions independent by the selection variables

- In our DAG: *P(T | A;C)* **NOT EQUAL** *P(T | A;C; S)*

All recorded features taken into account

Selection Variables affect the distribution of the target Variable if S is taken into account



Bounded distributional robustness Methods

a. **Domain adaptation:** i.e. reweight learning process to optimize in the target domain. These methods assumes a relationship between training and test data shift:
    i. Distributions Centered on the training distribution
    ii. Shift is Bounded in magnitude

→ *Absence of robustness guarantees on perturbations that are beyond the prespecified magnitude used during training*

→ *Need for proactive solutions: DS shift anticipated!!!*

RQ
How can we find a stable model?

Problem

- Unreliability of classical ML models when train and test distributions differ
- Minimize loss without assumption about the DS (distributional robustness assumptions)
- critical environments require to find a stable estimator when different policies are applied to some of the variable involved in the DGP

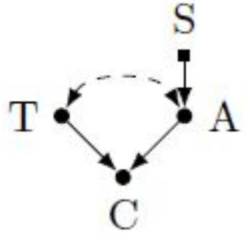
Solution =
Surgery Estimator

- Convert the **DAG** into an **ADMG** (acyclic direct mixed graph) to model unobserved confounding using prior knowledge about selection variables (NOT always identifiable)
- Surgery estimator Allow the target to be explicitly be generated by varying mechanisms
- The algorithm searches all possible interventional distributions (which intervene on S) for the optimal identifiable distribution

3. Methods

From DAG to ADMG

a) bi-directed Graph : are acyclic in the sense that they not contain purely directed cycles.
b) Will be causal DAGs whose nodes can be partitioned into sets
   i) **O** of observed variables,
   ii) **U** of unobserved variables, and S of selection variables.
c) **O** and **U** consist of variables in the DGP
d) **S** are auxiliary variables that denote mechanisms of the DGP that vary across environments



(c)

Any hidden variable DAG can be converted to an ADMG by taking its latent projection onto O. (bi) directed edges are created based on the fact that internal nodes in direct (divergent) path are observed (unobserved) nodes.

Main Components of the surgery estimator algorithm

Our **DAG** could be affected by dataset shift. Need to find stable (independence from **M**) estimator for the target variable

ADMG: is the tool that allows to model selection variables.
Algo perform `do(X)-> remove all edges out of` **`S`** `→` **`disconnected and d-separeted from target`**

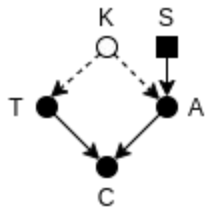
`Interventional Distribution algo`[1]`:`
Determines the identifiability of interventional distribution, searches all possible ID (which intervene on M), with respect to the optimal loss.

[1]Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In 22nd Conference on Uncertainty in Articial Intelligence, UAI 2006

3.2 Graph Surgery Estimator

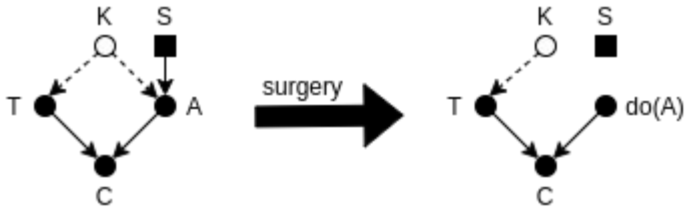
Assumes the data modeler has constructed or been given a causal **DAG** of the **DGP** with target prediction variabile **T**, observed variables **O**, and unobserved variables **U** that has been augmented with selection variables **S** using prior knowledge about mechanisms that are expected to differ across environments (e.g. prescription policy)

3.2 Graph Surgery

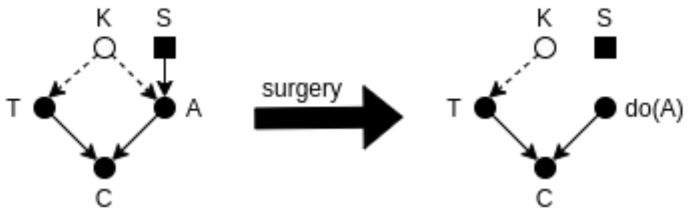


Goal: predict T using observed features A and C

- Naively using all features is **unstable**
  - $P(T|A,C) \neq P(T|A,C,S)$
- Only stable feature set is empty set!
  - $P(T) = P(T|S)$



3.2 Graph Surgery



Solution: Use **interventional** not observational distribution to predict T

- Hypothetical intervention in which we set **A** to observed value for every individual
- Resulting conditional interventional distribution is stable

$$P(T|C, do(A)) \propto P(T, C|do(A)))$$
$$= \underbrace{\underbrace{P(T)}_{\text{Feature selection}} P(C|T, A)}_{\text{Graph Surgery}}$$

3.2 Graph Surgery Estimator: Algorithm

An overview of the procedure is as follows:

- The selection **DAG** is converted to a selection **ADMG**
- Children of **S** in the selection **ADMG** form the set of mutable variables **M**
- The proposed algorithm then searches all possibles interventional distributions (which intervene on **M**) for the optimal identifiable distribution, which is normalized and returned as the surgery estimator

3.3 Graph Surgery Estimator: Algorithm

Graph Surgery Estimator Algorithm is sound in that only returns stable estimators and complete in that it finds

Theorem 1 (Soundness): When Algorithm returns an estimator, the estimator is stable

Theorem 2 (Completeness): If Algorithm fails,then there exists no stable surgery estimator for predicting T

Pseudocode

Algorithm 2: Graph Surgery Estimator

input : ADMG $\mathcal{G}$, mutable variables $\mathbf{M}$, target T
output: Expression for the surgery estimator or `FAIL` if there is no stable estimator.

Let $S_{ID} = \emptyset$; Let $Loss = \emptyset$;
for $\mathbf{Z} \in \mathcal{P}(\mathbf{O} \setminus (\mathbf{M} \cup \{\mathbf{T}\}))$ **do**
  if $T \notin \mathbf{M}$ **then**
    Let $\mathbf{X}, \mathbf{Y} = \text{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G})$;
    try
      $P = \text{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
      $P_s = P / \sum_T P$;
      Compute validation loss $\ell(P_s)$;
      $S_{ID}.\text{append}(P_s)$; $Loss.\text{append}(\ell(P_s))$;
    catch
      pass;
  Let $\mathbf{X}, \mathbf{Y} = \text{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G}_{\overline{T}})$;
  $\mathbf{X} = \mathbf{X} \cup \{T\}$; $\mathbf{Y} = \mathbf{Y} \setminus \{T\}$;
  if $\mathbf{Y} \cap (T \cup ch(T)) = \emptyset$ **then**
    continue;
  try
    $P = \text{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
    $P_s = P / \sum_T P$;
    Compute validation loss $\ell(P_s)$;
    $S_{ID}.\text{append}(P_s)$; $Loss.\text{append}(\ell(P_s))$;
  catch
    continue;
if $S_{ID} = \emptyset$ **then**
  return `FAIL`;
return $P_s \in S_{ID}$ with lowest corresponding $Loss$;

The **input** are:

- graph ADMG G,
- mutable variables M
- target T

the possible **output** are:

- Expression for the surgery estimator
- FAIL if there is no stalbe

If the set S_{ID} is empty set then the algorithm return FAIL
else return Ps belong to S_{ID} with lowest corresponding Loss

propose an exhaustive search over possible conditioning sets

Algorithm 1: Unconditional Query: $\mathrm{UQ}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}; \mathcal{G})$

input : ADMG $\mathcal{G}$, disjoint variable sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{O}$
output: Unconditional query $\propto P_{\mathbf{X}}(\mathbf{Y}|\mathbf{Z})$.
$\mathbf{X}' = \mathbf{X}$; $\mathbf{Y}' = \mathbf{Y}$; $\mathbf{Z}' = \mathbf{Z}$;
while $\exists Z \in \mathbf{Z}$ *s.t.* $(\mathbf{Y} \perp\!\!\!\perp Z|\mathbf{X}, \mathbf{Z} \setminus \{Z\})_{\mathcal{G}_{\overline{\mathbf{X}},\underline{Z}}}$, **do**
    $\mathbf{X}' = \mathbf{X}' \cup Z$;
    $\mathbf{Z}' = \mathbf{Z}' \setminus \{Z\}$;
$\mathbf{Y}' = \mathbf{Y} \cup \mathbf{Z}'$;
return $\mathbf{X}', \mathbf{Y}'$ of unconditional query $P_{\mathbf{X}'}(\mathbf{Y}')$

Algorithm 2: Graph Surgery Estimator

input : ADMG $\mathcal{G}$, mutable variables $\mathbf{M}$, target T
output: Expression for the surgery estimator or `FAIL` if there is no stable estimator.
Let $S_{ID} = \emptyset$; Let $Loss = \emptyset$;
for $\mathbf{Z} \in \mathcal{P}(\mathbf{O} \setminus (\mathbf{M} \cup \{\mathbf{T}\}))$ **do**
    if $T \notin \mathbf{M}$ **then**
        Let $\mathbf{X}, \mathbf{Y} = \mathrm{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G})$;
        try
            $P = \mathrm{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
            $P_s = P / \sum_T P$;
            Compute validation loss $\ell(P_s)$;
            $S_{ID}.\mathrm{append}(P_s)$; $Loss.\mathrm{append}(\ell(P_s))$;
        catch
            pass;
    Let $\mathbf{X}, \mathbf{Y} = \mathrm{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G}_{\overline{T}})$;
    $\mathbf{X} = \mathbf{X} \cup \{T\}$; $\mathbf{Y} = \mathbf{Y} \setminus \{T\}$;
    if $\mathbf{Y} \cap (T \cup ch(T)) = \emptyset$ **then**
        continue;
    try
        $P = \mathrm{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
        $P_s = P / \sum_T P$;
        Compute validation loss $\ell(P_s)$;
        $S_{ID}.\mathrm{append}(P_s)$; $Loss.\mathrm{append}(\ell(P_s))$;
    catch
        continue;
if $S_{ID} = \emptyset$ **then**
    return `FAIL`;
return $P_s \in S_{ID}$ with lowest corresponding $Loss$;

where P(...) denotes the power set. In the interest of identifiability

we may want to consider intervening on T

For example, $P_X(T|Y)$ and $P_X(T)$ are not identifiable, but $P_{X,T}(Y)$ is. Thus, we should consider the unconditional query returned by Algorithm 1

Note: that it returns the estimator that performs the best on held out source environment validation data with respect to some loss function

Graph Surgery Overview

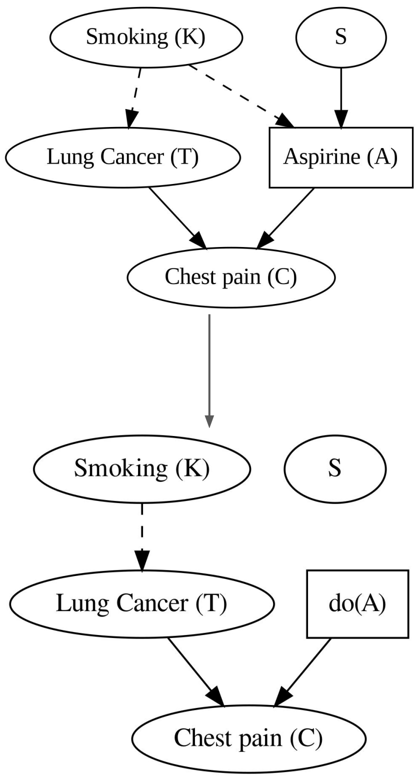
main point of view:

- Graph surgery strictly generalizes feature selection-based methods for achieving stability in proactive transfer learning
  - Stable cond distributions are special case of stable interventional distributions
  - Surgery can capture more stable paths
  - Surgery yields stable predictions in scenarios in which conditioning cannot
  - Requires the interventional distribution to be identified
- Graph surgery achieves (unbounded) distributional robustness for family of distributions defined by selection diagrams

Relationship with Graph Pruning

Graph pruning is a special case of surgery.

For this reason, there **exists** a **problem** for which **graph pruning cannot find** a non-empty stable **conditioning set** but for which **graph surgery does not fail**.



Experiments

Goals

The goal of these experiments are:

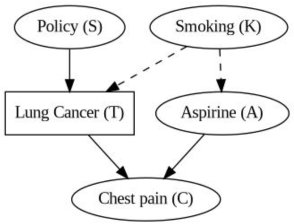
- Evaluate the stability of the algorithm
- Evaluate the trade of between stability and performance
- Compare the graph surgery estimator against 3 algorithms:
  - **Ordinary least squares (OLS):** a baseline approach which doesn't take into account the variation between training set and test set.
  - **Causal Transfer Learning (CT):** a state of the art pruning approach [1]
  - **Anchor Regression (AR):** a distributionally robust method for bounded magnitude shift interventions [2]

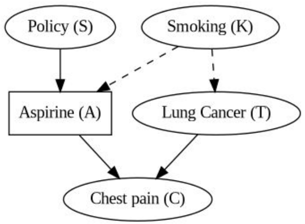
[1] Rojas-Carulla, Mateo, et al. "Invariant models for causal transfer learning." The Journal of Machine Learning Research 19.1 (2018): 1309-1342.
[2] Rothenhäusler, Dominik, et al. "Anchor regression: Heterogeneous data meet causality." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83.2 (2021): 215-246.

Datasets

Simulated Data

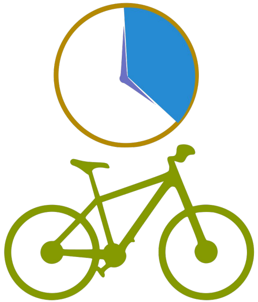
Starting from a selection diagram the authors generated two synthetic datasets





Real Data: Bike Rentals

The authors used the UCI Bike Sharing dataset in which the goal is to predict the number of hourly bike rental



Simulated Data

The authors simulated data from zero-mean linear Gaussian systems using the the DAGs in Figure A and Figure B. In both cases the task require to compute the **posterior probability of T**.

For this experiment the author took into account the algorithms OLS and CT. AR was excluded because there isn't any anchor variable (an observable variable without parents)

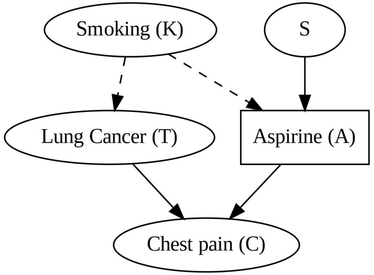


Figure A

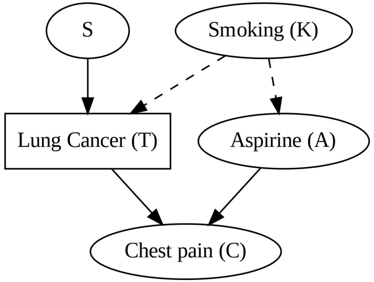
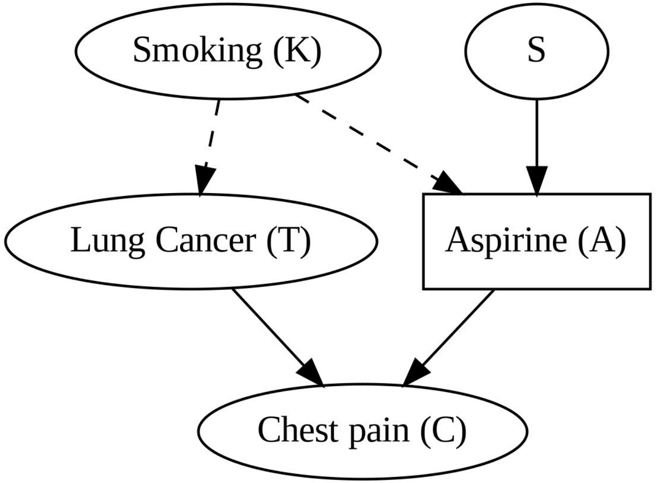


Figure B

Simulated Data - Aspirine

For this dataset the stable models CT and Surgery are able to generalize beyond the training environments.

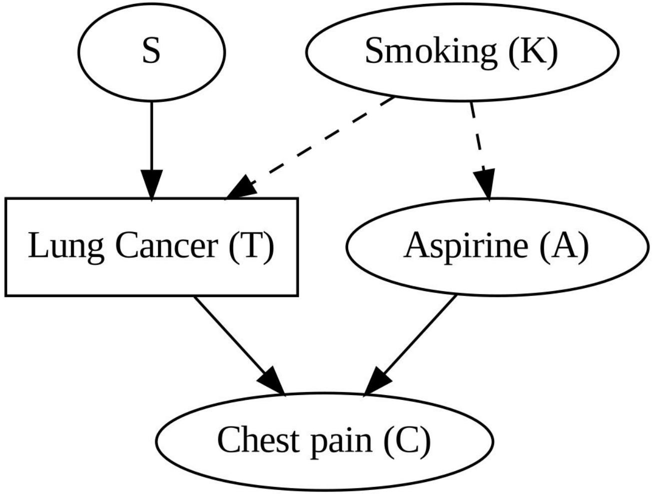
However, for small deviations from the training environment, OLS outperforms the stable methods which shows that there is a **tradeoff** between stability and performance.



Simulated Data - Lung Cancer

In this dataset the authors consider the target shift scenario is which T is the mutable variable.

This DGP violates the assumption of CT. For this reason the only stable algorithm is Surgery.



Real Data: Bike Rentals

The dataset contains: **hourly date rentals**, **weather data** (temperature, wind speed, humidity) and **temporal informations** (season and year).

The authors partition the data by season then, iteratively, they select **one season as the target** using the other three seasons as source.

The **Surgery estimator performs competitively,** achieving the best results i 3 of 8 test cases.

On the contrary **CT struggle** in this settings because no stable pruning estimator exists.

AR achieve very good performance. However, it requires tuning of a hyperparameter. For this reason, when the target environment is unknown, the **surgery algorithm is a safer option**.

Conclusion

Dataset shift is a common problem which **negatively affects** the **performance** of ML models.

The authors developed the **surgery estimation** to address this problem.

This framework finds a **stable** and **identifiable interventional distribution**.

PROS	**CONS**
• Superset of Graph Pruning • No anchor required • Parameter tuning not required • Stability	• Outperformed by more specific methods (i.e. AR) • Trade Off between stability and performances