Causal Inference

→ So far, purely predictive questions!

→ If there are signs that there is correlation between features and target of interest, that's good enough

→ Causal Directionality is irrelevant. (Not completely true)

→ When there is dataset shift, or non-stationary data, causality matters. Understanding the data deeply is very helpful!

→ In healthcare especially, causal questions are important to answer.

→ For example, it is more important to prevent Type 2 diabetes (causal) vs early diagnosis of Type 2 diabetes (predictive).

→ Naive way of inferring causality:

- Let's say we trained a DL model to predict onset of Type 2 diabetes
- Look at most negative feature (lowest weight), let's say it is Gastrogenic Bypass Surgery (yes or no)
- Then does that mean that if a patient underwent this surgery, he/she won't get diabetes?

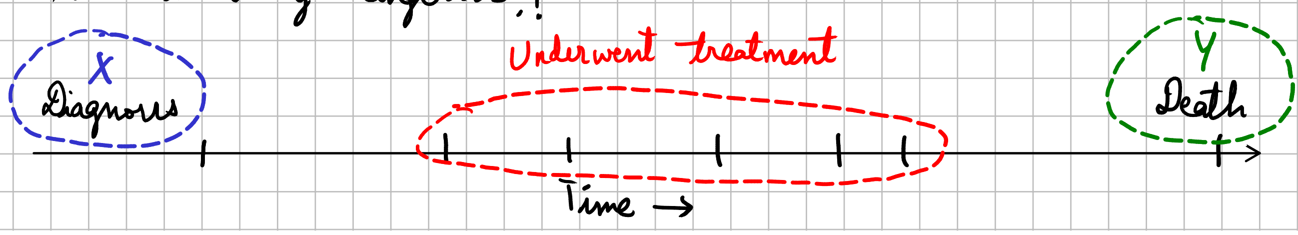
→ Look at predictive weights is not enough.

→ We need to come up with a mathematical model for causality

→ Another example:

- Let's say we train a DL model for predicting survival of breast cancer patient based on radiological mammogram and histopathological slides.
- Let's say one patient diagnosed with breast cancer survived for longer than 5 years.
- When a new patient, with similar diagnosis as the model's examples is given to the model, a higher survival is predicted.
- Does this mean we shouldn't treat the patient?

→ This is very dangerous!!



→ A longer survival time maybe because of treatment! Not solely because of diagnosis

→ But the model only learns the X (diagnosis) → Y (survival) mapping

Guiding Treatment Decisions

→ Another question that needs to be answered is: How do we guide treatment decisions?

→ How do we tell who is likely to be benefitted by a given treatment?

→ But people respond differently to treatments?

→ Also, data used to guide treatments is biased on existing treatment guidelines.

→ Naive way to guide treatment decisions:

Train a predictive model that learns to predict treatment decisions.

David → Treatment A

John → Treatment B

Juana → Treatment A

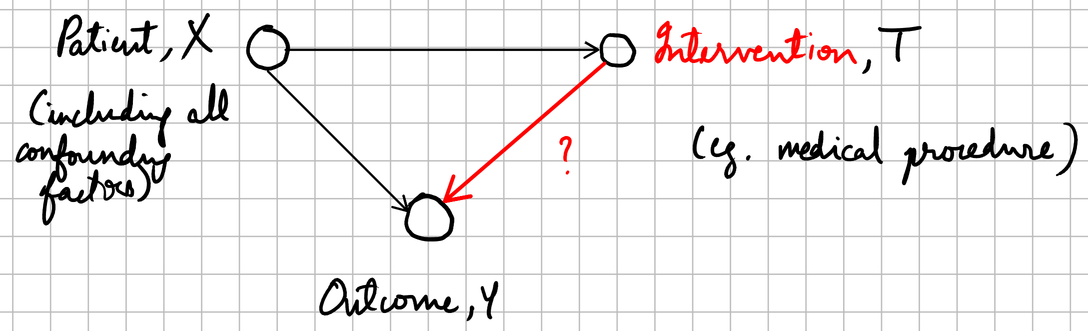
→ Best this can do is match current medical guidelines!

→ How do go we beyond this? We need to capture heterogeneity in treatment response. We need to change how we ask our question.

→ One last example:

- Traditional, does X cause Y?
- Does smoking cause cancer?
- Doing a randomized controlled trial is unethical.
- Could we just compare P(lung cancer | smoker) vs P(lung cancer | non smoker)?
- No because of confounding factors (see below)

→ To properly answer, we need to formulate as causal questions.



High dimensional Observational data

Causal Graphs

→ Instead of just $D \in \{x^{(i)}, y^{(i)}\}$, we need to think in terms of triplets : $D \in \{x^{(i)}, T^{(i)}, y^{(i)}\}$ Interventions

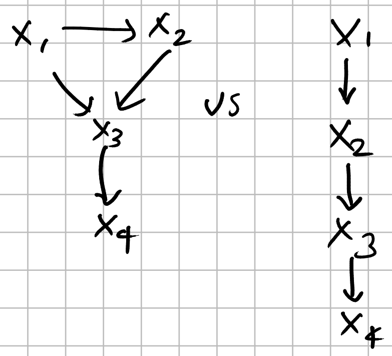
---

→ Earlier Causal Inference might take a form like so: (Can Skip! Won't be discussed further)

Data

x_1	x_2	x_3	x_4
0	1	0	0
0	0	1	0
⋮			
1	1	1	0

we try to find right graph →



What is the underlying causal graph?

For Two random variables and X_1 and X_2

$X_1 \rightarrow X_2$ i.e., $P(X_1)\, P(X_2 | X_1)$

$X_2 \rightarrow X_1$ i.e, $P(X_2)\, P(X_1 | X_2)$

Are indistinguible (because of conditional probability and bayes theorem)

Then we might wanna add interventions to $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$ to disentangle the causalities.

→ This is the simplest possible case!

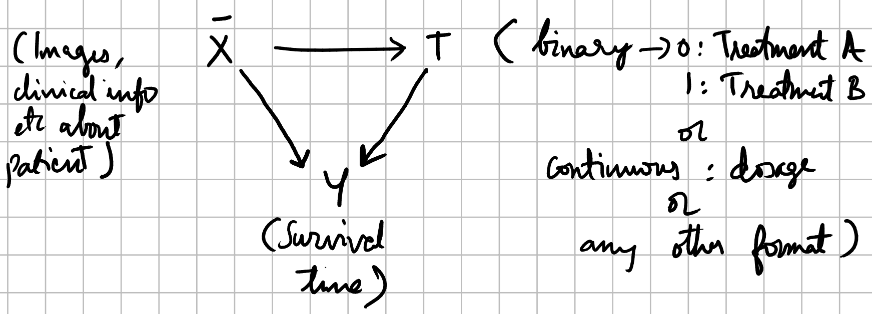
→ X is highdimensional, T and Y and single random variable.

→ The causal graph is assumed here.

→ We just don't know the connection weights' strength!

Modelling Breast Cancer example as a causal graph

X → Patient information (features) at Diagnosis
T → choose between treatment A or treatment B. Also treatment plans only depend on what we know about the patient at diagnosis
Y → Survival time



→ We will not change treatment in between (Dynamic treatment regimes or off policy reinforcement learning)

→ Causal Inference questions have been asked for decades in political science, economics and statistics but we can't intervene (can only use observational data)

→ In tradional statistics, domain knowledge was required to determine factors that determine the treatment to be taken. These factors are called confounding factors.

→ Now causal inference questions are asked on high dimensional data (images) and research on leveraging machine learning algorithms, by mapping a causal inference problem down to a machine learning model.

→ To formalize these notions concretely, we need a mathematical model

Potential Outcomes Framework (Rubin-Neyman Causal Model)

→ Each unit (individual) x_i has two potential outcomes:

- $Y_0(x_i)$ is the potential outcome had the unit not been treated: "control outcome"

- $Y_1(x_i)$ is the potential outcome had the unit been treated: "treated outcome"

→ Conditional average treatment effect for unit i:

$$CATE(x_i) = E_{Y_1 \sim P(Y_1|x_i)}[Y_1 | x_i] - E_{Y_0 \sim P(Y_0|x_i)}[Y_0 | x_i]$$

Given x_i (such as age = 56), mean all outcomes where treatment is given

given x_i (such as age = 56) mean all outcomes where treatment is not given.

also called Individual Treatment effect

→ Average treatment effect:

$$ATE : E[Y_1 - Y_0] = E_{x \sim p(x)}[CATE(x)]$$

→ Observed factual Outcome (You can only Y_0 or Y_1 at a time but not both):

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$$

→ Unobserved counterfactual outcome (What if the opposite treatment was given?)

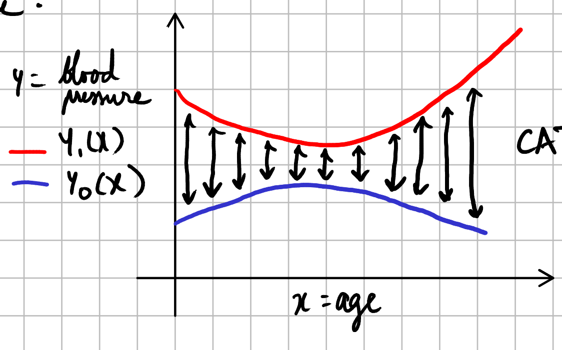
$$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

→ Since we observe only factual outcomes, we need to impute the counter factual outcomes.

"We only ever observe one of the two outcomes"

→ Fundamental problem in Causal Inference.

Example:



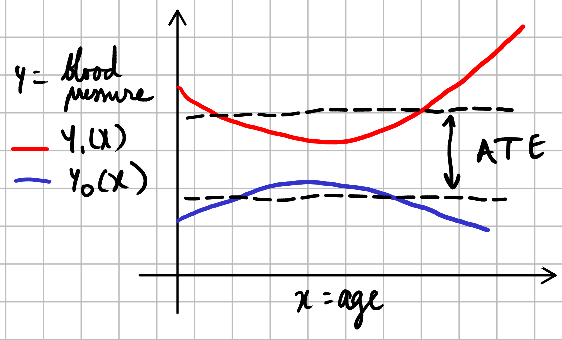
For middle aged people, difference between treatments 0 and 1 is significantly lower than younger and old aged people.

If treatment 1 is much cheaper than 0, we could use it for middle aged people.

→ We wanna predict CATE value using our data.

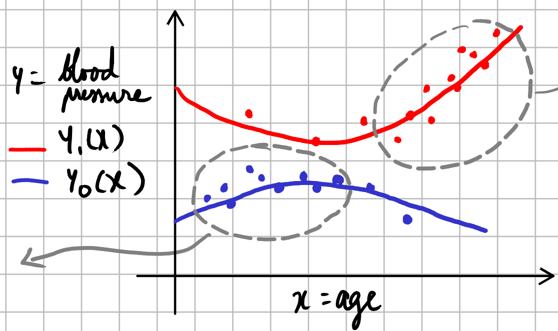
→ Sometimes we don't have the luxury of having personalized treatment effects (CATE). For example, the govt may start a policy that all men above 50 should receive prostate cancer screening.

→ Then you would wanna see the amount of decrease in prostate cancer related deaths. Such a policy is very broad. Mathematically captured by ATE.



Treatment 0 on average is better than treatment 1.

→ But IRL we don't observe curves, rather we observe data points.



Most people given treatment 1 are older

Most people given treatment 0 are younger.

Can be due to a variety of reasons. This is modelled by $X \longrightarrow T$ in causal graph

→ What if the other treatment had been given to the patient?