Essential Causal Inference Techniques For Data Science

Vinod Bakthavachalam | Data Science at Coursera

Data Science Questions Often Fall into A Standard Format

1. Do free trials increase revenue?
2. Does sales support drive renewals?
3. Why did ABC metric change this month?
4. ...

Data Science Questions Often Fall into A Standard Format

Commonalities:

- Some outcome metric of interest **Y**
- Some variable of interest **X**
- Goal is to estimate ***coefficient of interest***, which is the estimated impact of changing X on Y i.e. **the causal impact**

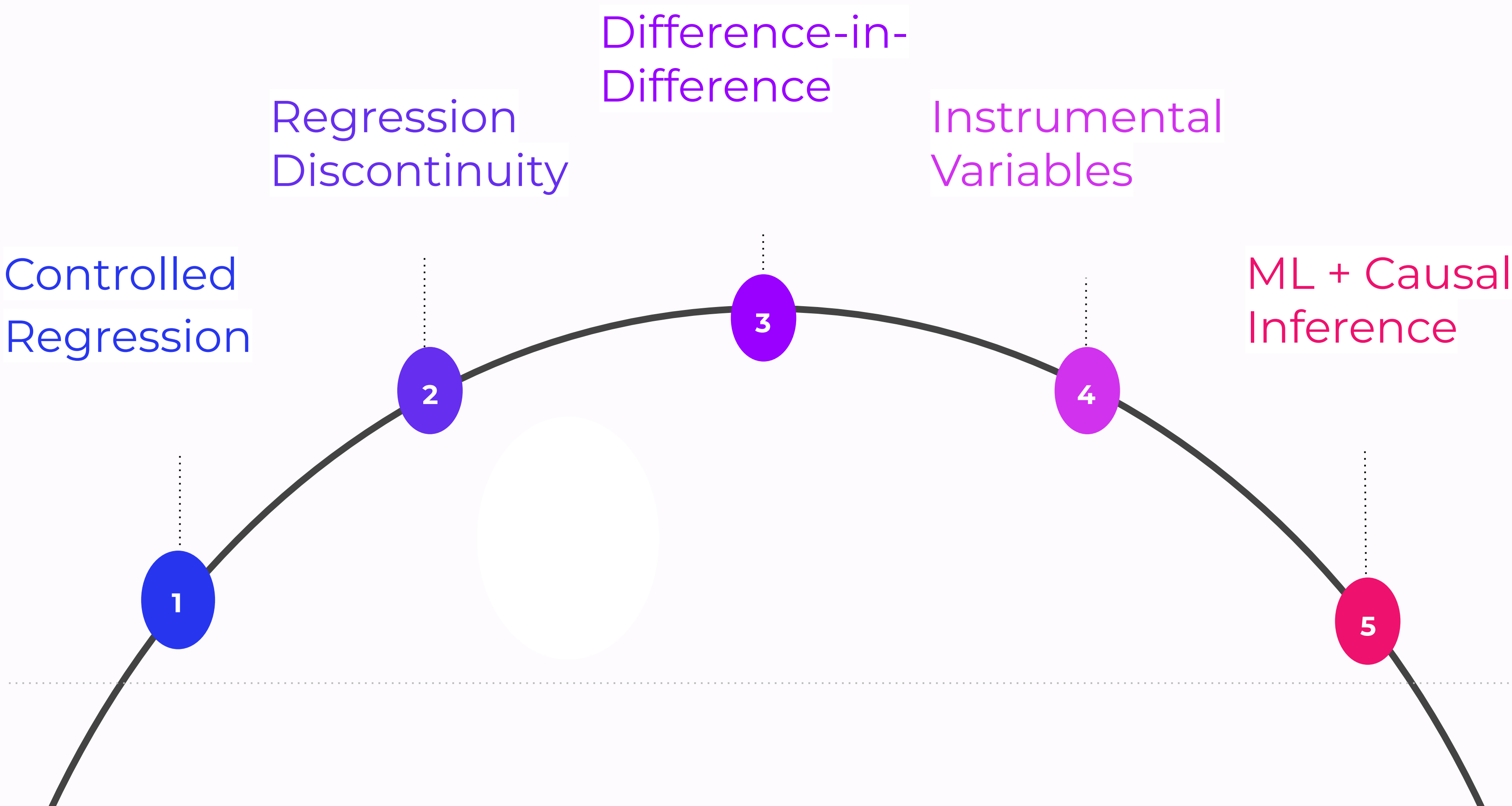
We use causal inference techniques to estimate this coefficient of interest / causal impact

Intuition Behind Causal Inference

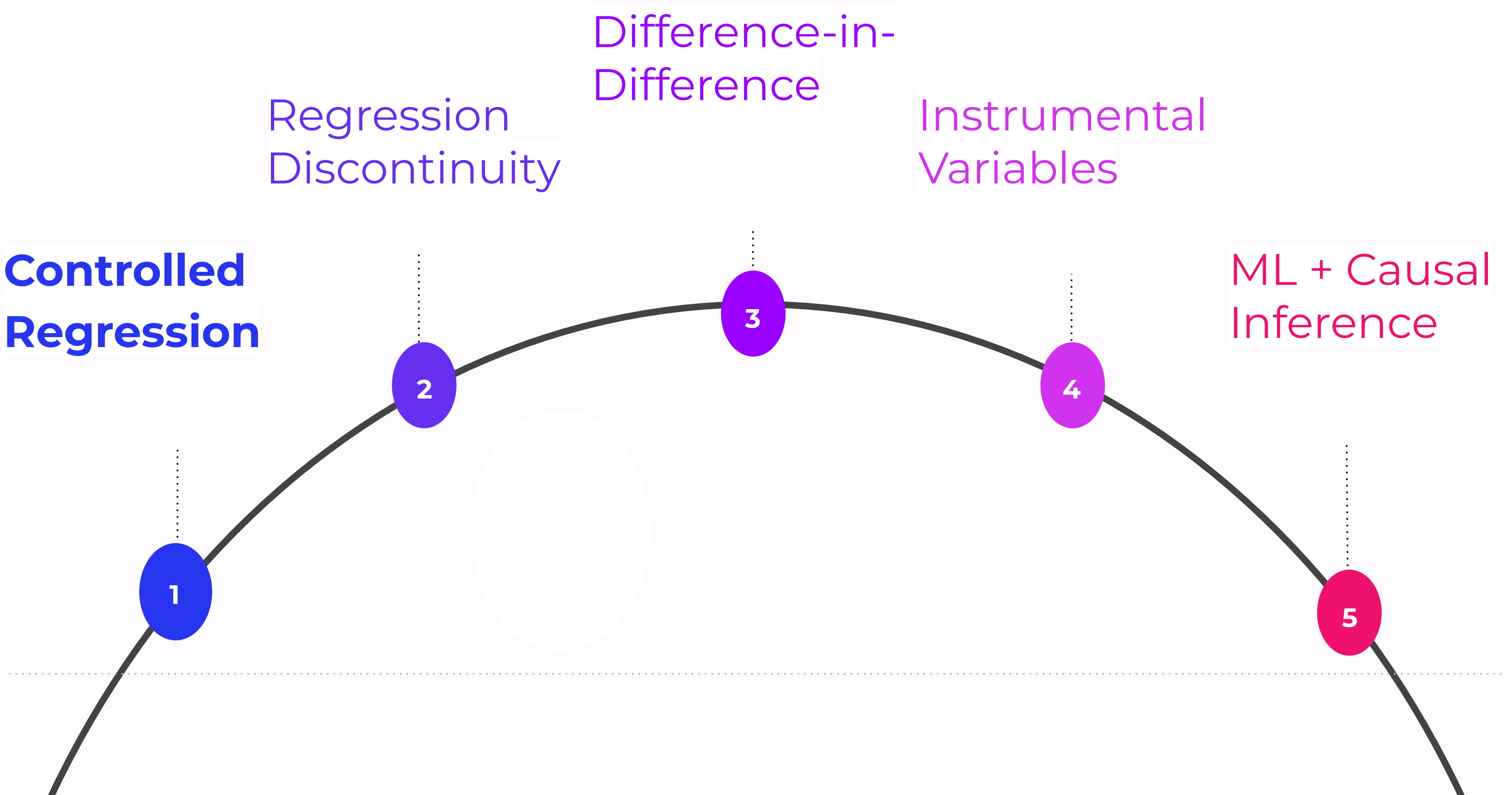
Central Idea:

*Try to control for all possible confounders and look for **"natural sources" of variation** that can split data into quasi-random groups and **mimic the randomization** we would get from AB testing.*

Essential Methods for **Causal Inference**

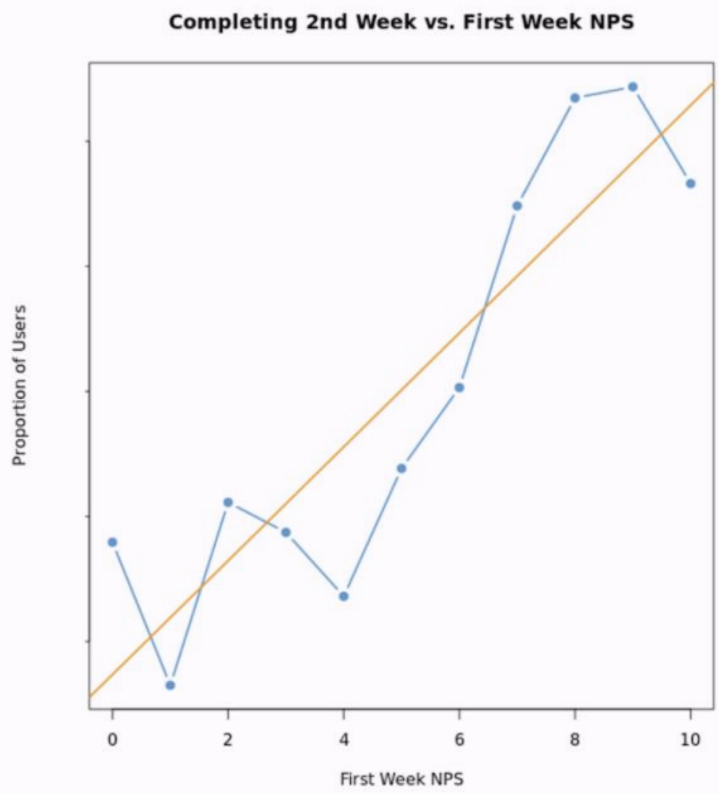


Essential Methods for **Causal Inference**



Suppose we want to measure ***the relationship between product quality and usage***

We have our Y → usage as measured by completing the 2nd week of a course
And our X → First Week NPS (net promoter score) which is a satisfaction rating on a 1-10 scale



Controlled Regression

Steps

(1) Univariate Regression of Y (Usage) on X (Product Quality) only
(2) Multiple regression of Y (Usage) on X (Product Quality) and a set of controls

If…

(1) R squared in second regression is close to 100%
(2) Coefficient on X is similar in the two models

…then by the theory of controlled regression, we can use it as the causal impact.

Example: Product Quality vs. Usage

First column is univariate regression; second column is multiple regression

We see r squared increase a lot when controls are added and coefficient on product quality term (First Week NPS) is fairly stable.

	Complete 2nd Week ~First Week NPS (3)	Complete 2nd Week ~First Week NPS+Controls (4)
First Week NPS	0.0063***	0.0074***
	(0.0005)	(0.0005)
R^2	0.0006	0.1842

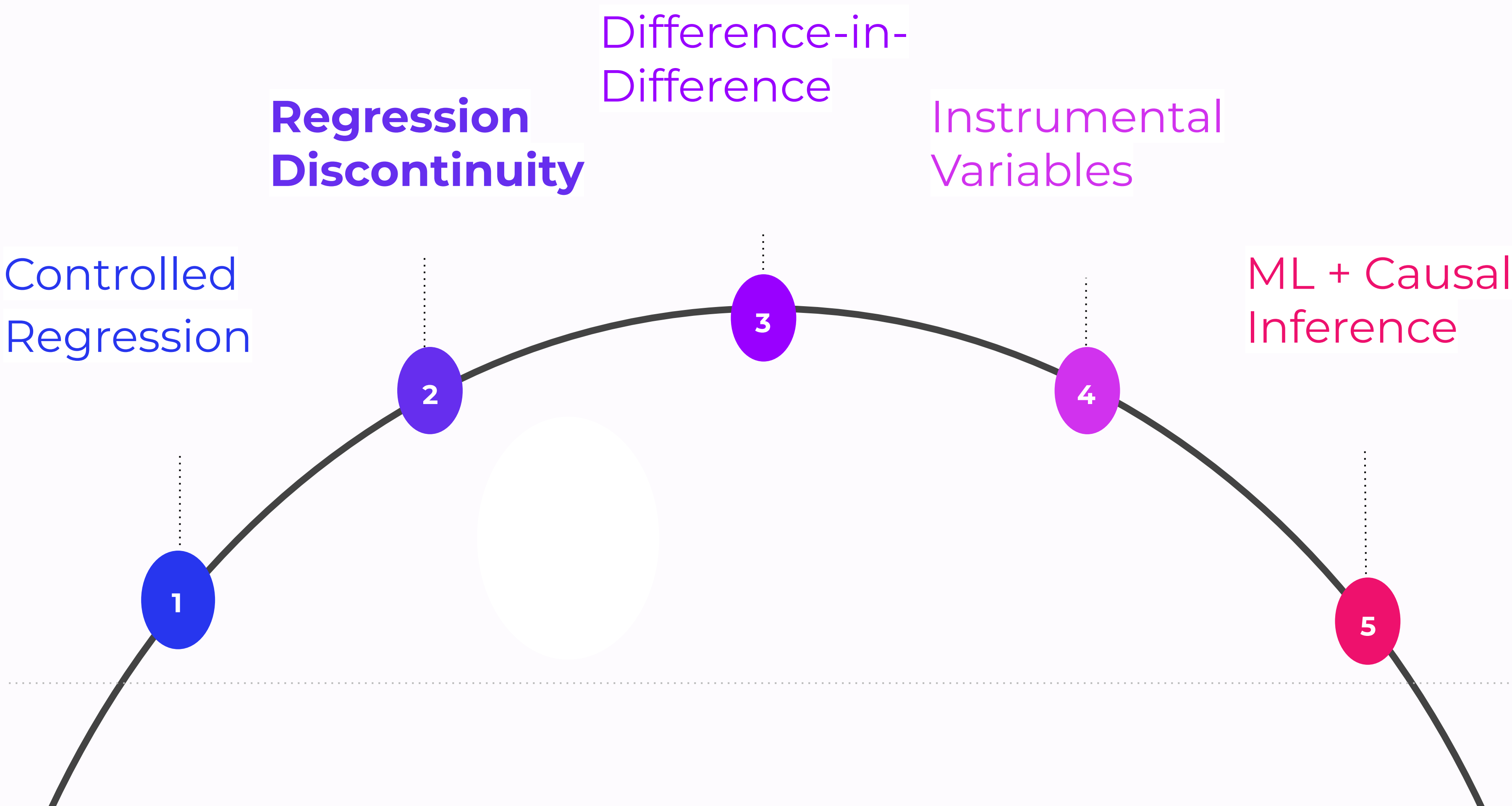
Sources of Error: Omitted Variable Bias

Definition	Omitting control variables that matter from the model
Example	We know completion rates of courses differ by course length, so course length should be included. Leaving out would cause omitted variable bias.
How to Tell?	Look at R Squared in regression with controls and see if close to 100%.

Sources of Error: Included Variable Bias

Definition	This is the opposite of omitted variable bias and involves including too many controls
Example	Time available to take courses is a confounding factor. We could try to use other courses enrolled in as a control, but that is affected by product quality. Adding would cause included variable bias.
How to Tell?	How to Tell? No direct way, but generally leave out controls that are not fixed at time observe X. Analogy: Time traveling in ML feature engineering

Essential Methods for **Causal Inference**



Suppose we want to measure ***the effect of adding subtitles to a course***

Can we run an **AB test**?

Difficult to randomly give some learners to access subtitles given product limitations

Can we do **controlled regression**?

Key unobservables like course popularity → omitted variable bias

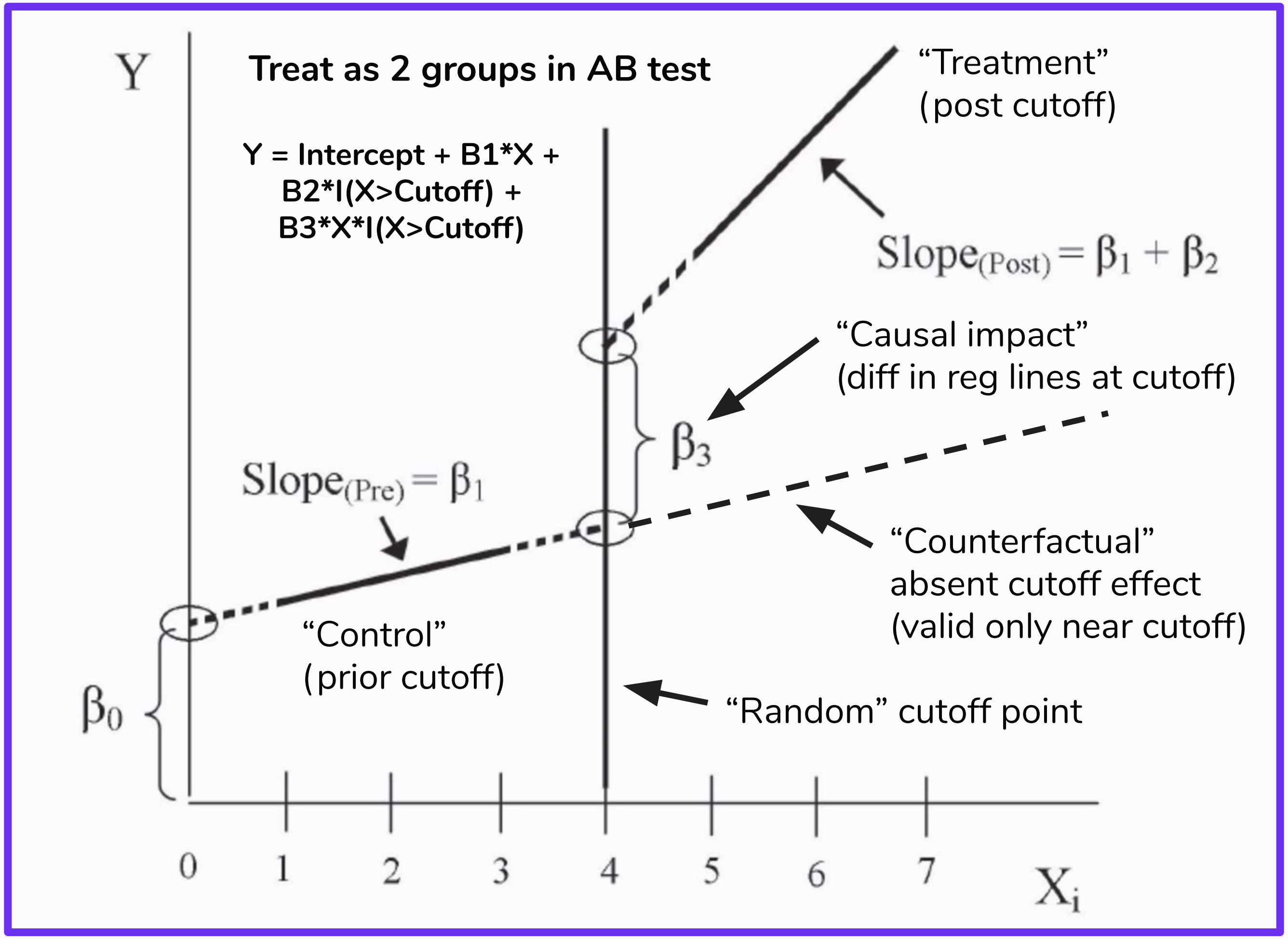
What about a **natural experiment**?

Turns out courses are advertised in a language only when they are at least 80% subtitled

Run a regression discontinuity with a cutoff point of 80% where:
Y variable → Revenue and X variable → % of Course Subtitled

Regression Discontinuity (RDD) Graph

Idea: *Focus on cut-off point that can be thought of as local randomized experiment*

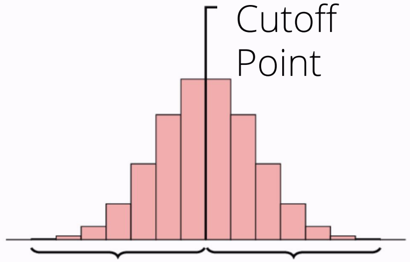


Assumption 1: Sample Similar Above + Below Cutoff

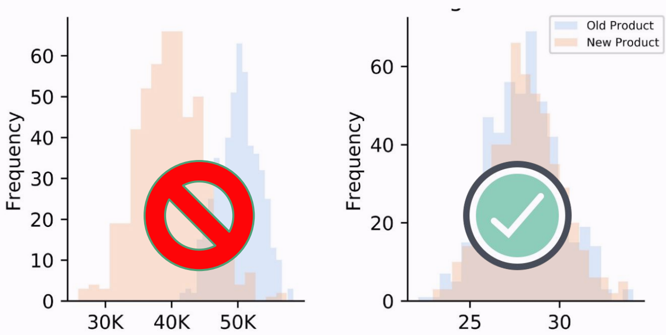
Example: Courses below and above the 80% subtitle threshold are similar to one another, so the discontinuity point effectively randomizes things.

How to Check:

1. Sample sizes similar just below and above cutoff i.e. are roughly balanced



2. Sample just below and above cutoff are similar on observables/confounders (other variables that might drive Y variable of interest)

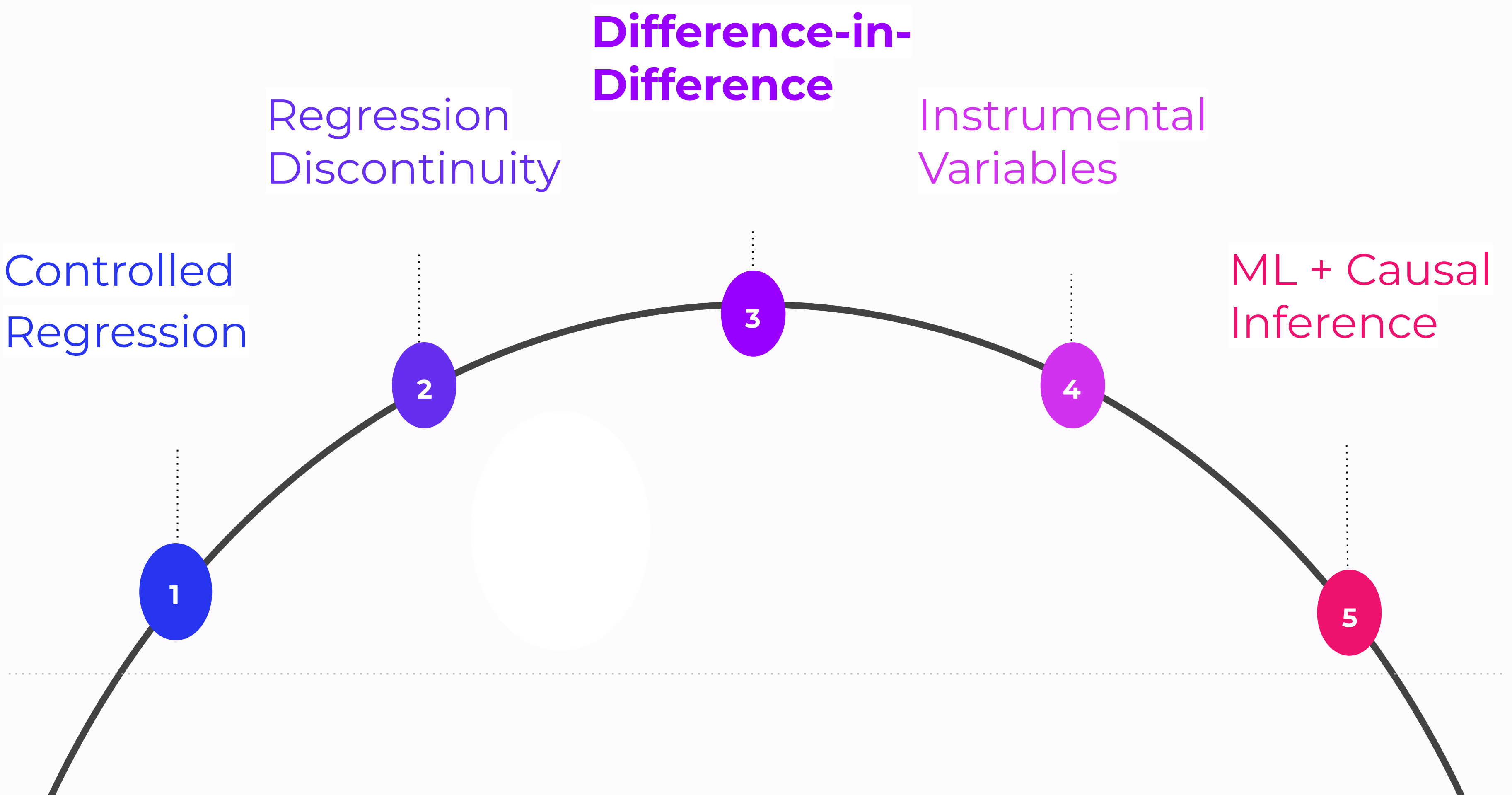


Assumption 2: No Confounding Discontinuities

Example: *For subtitles, assume advertising available or not is the only differentiator between 70% and 90% (for example no emails of content saying this is coming soon, etc.)*

How to Check: Run *placebo tests where run regression discontinuity at points other than the cutoff and check for no effect. → Run Regression Discontinuity at 20%*

Essential Methods for **Causal Inference**



Suppose we want to measure ***the effect of lowering price on revenue***

Can we run an **AB test**?

We could, but customers may complain if only some get lower prices and hear about it

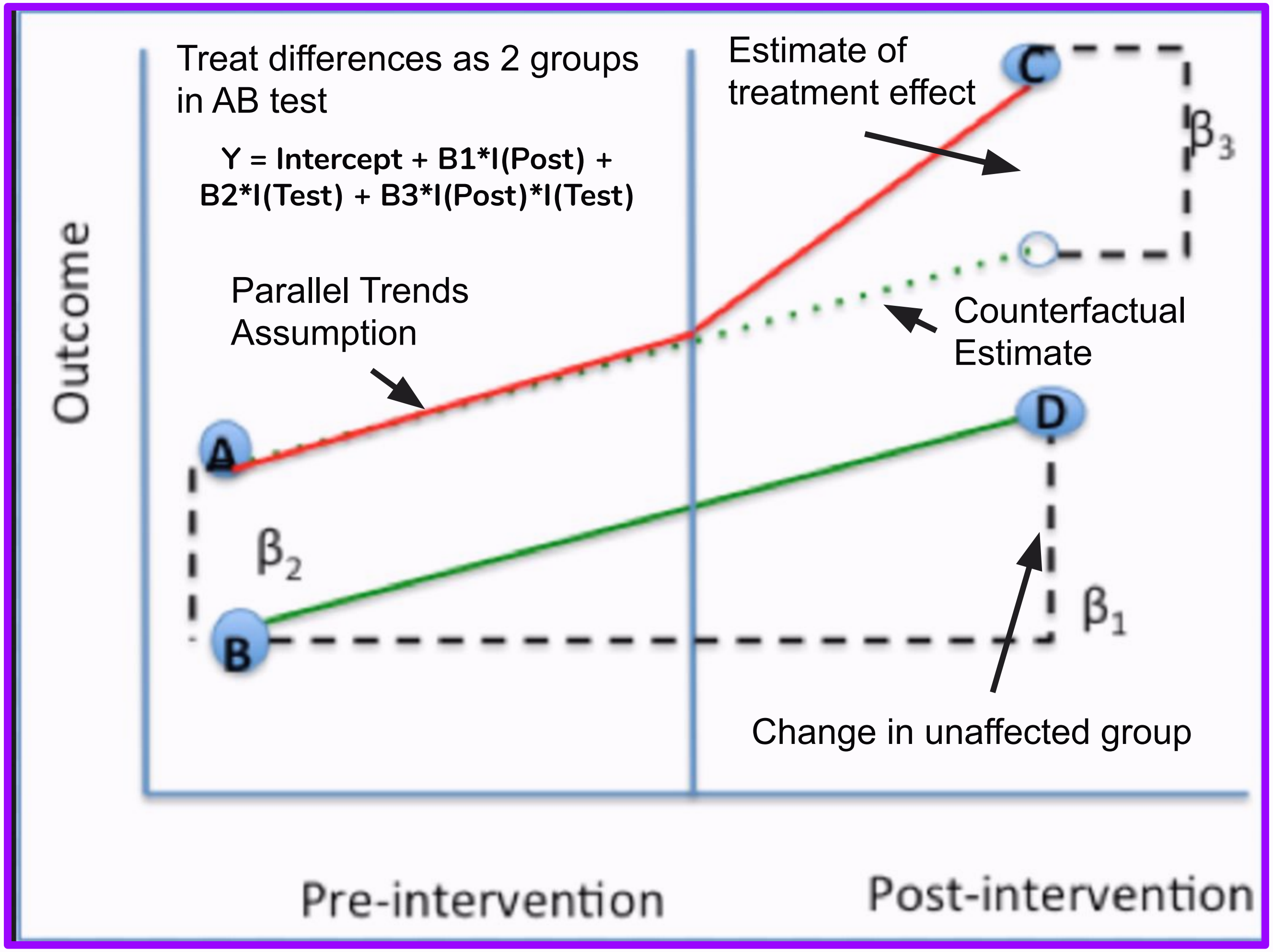
What about a **quasi experiment**?

We can change price in select geos (e.g., countries) but not others and use control markets to compute counterfactual (what would have happened absent price change in the treatment markets).

Run a regression discontinuity with control and treatment markets where:
Y variable → Revenue and X variable → treatment group in the post period

Difference in Difference

Idea: Compare pre and post outcomes between treatment and control groups

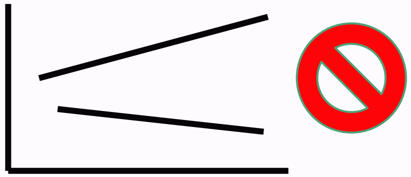


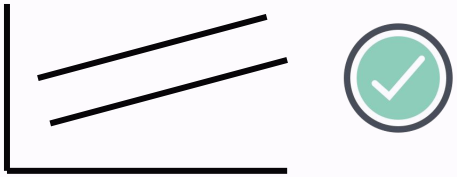
Assumption: Parallel Trends

Example: Check that revenue in control country with no price change was similar and highly correlated with revenue in treatment country with price change (ensures control can serve as counterfactual).

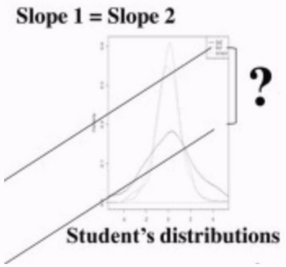
How to Check:

1. Graph control and treatment groups in the pre period and see if highly correlated.





2. Build a regression model to check whether trends are identical (no difference in slopes of two groups).



Extension: Synthetic Control

Problem with regular Diff-in-Diff:
Need to pick a single control group that satisfies parallel trends → can be arbitrary

Synthetic control creates a synthetic control group that is a weighted average of many control groups

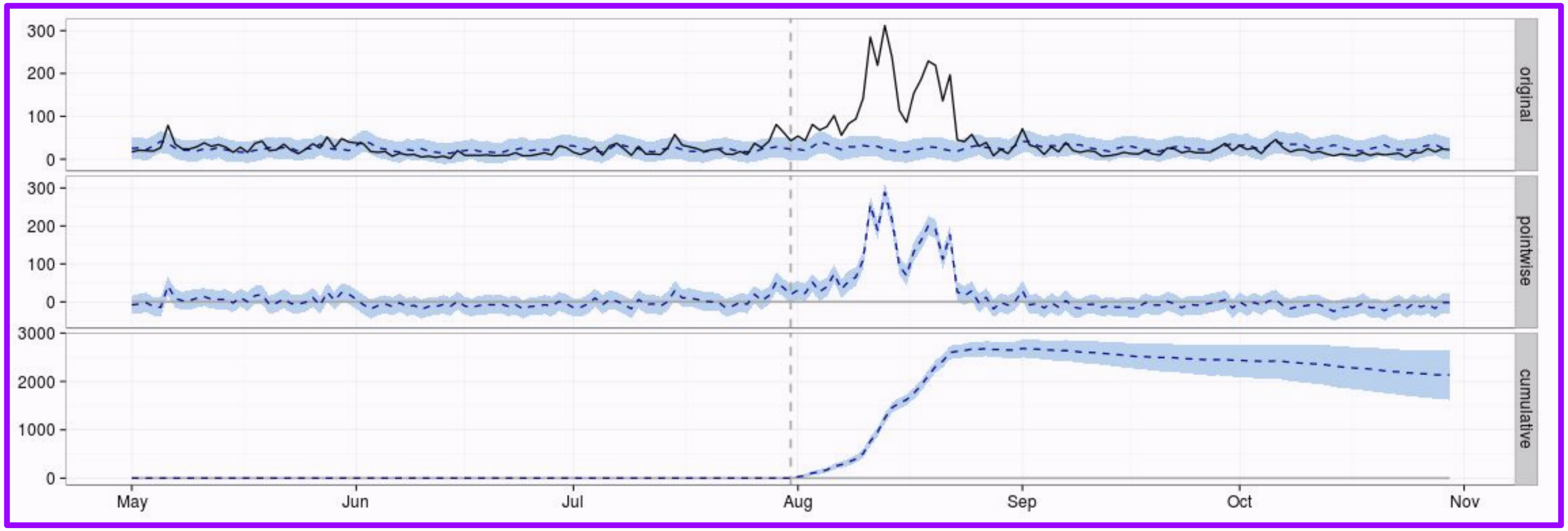
1. Choose weights to minimize tracking error with treatment group pre intervention → auto parallel trends.
2. Casual estimate is difference post intervention between treatment and “synthetic control”.

R Packages: Synth; Causal Impact (Bayesian Version)

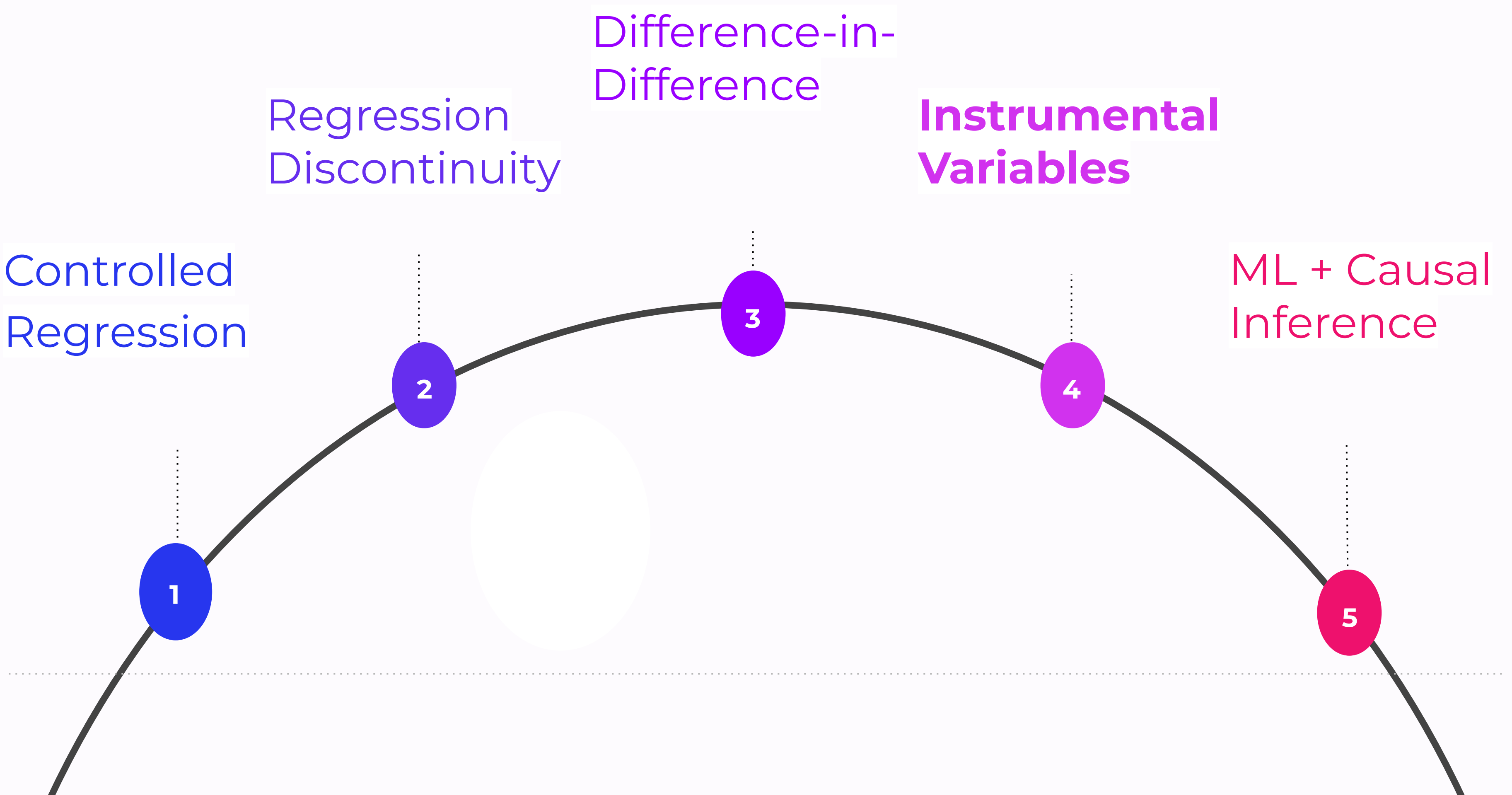
Extension: Bayesian Approach

Example: Discrete shock in given market, e.g.,

- PR announcement in India
- New partnership with Singaporean government
- A/B testing infeasible



Essential Methods for **Causal Inference**



Suppose we want to measure ***the effect of using the mobile app on course completion***

Can we run an **AB test**?

Difficult to randomly give some learners access to the mobile app

Can we do **controlled regression**?

Key unobservables like learner motivation → omitted variable bias

What about a **natural experiment**?

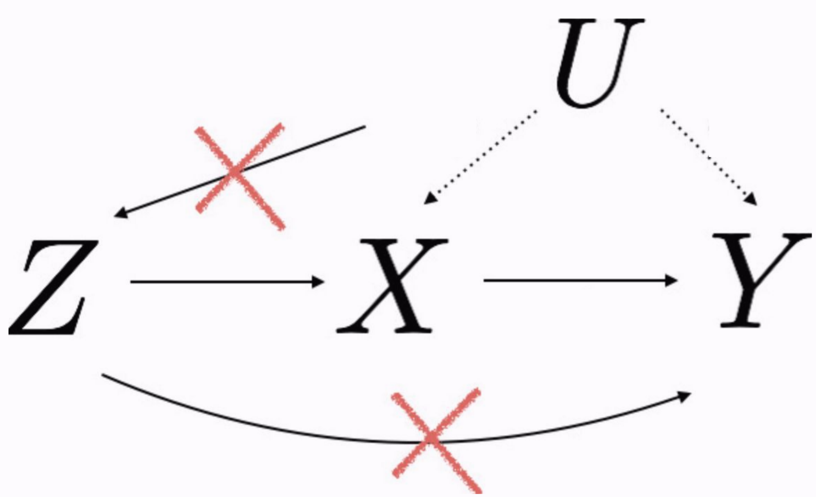
We can randomly nudge learners to download the mobile app in a randomized controlled trial (the nudge here is what is known as an instrument that we use to measure the relationship between mobile app usage and course completion)

Run instrumental variables where:
Y variable → Completion, X variable → Use mobile app,
Z (instrument) variable → Received random nudge

Instrumental Variables

General Problem: Unobserved variable(s) C affect both X and Y; can't use controlled regression because of omitted variable bias with no proxy variable that can be used as control

Idea: "Instrument" for X of interest with some feature, Z, that drives Y only through its effect on X → use to indirectly measure impact of Y on X



Assumption 1: Strong First Stage

Example: Study the impact of using the mobile app on course completion. Use an instrument created from a randomized nudge to download the mobile app, so **need it to predict mobile app usage strongly**.

How to Check:
Regress X variable of interest (mobile app usage) on instrument Z (nudge to download mobile app) and check that F statistic of regression is above 11 (rough rule of thumb) or perform other hypothesis tests for weak instruments.

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{df_{MSR}}}{\frac{SSE}{df_{MSE}}}$$

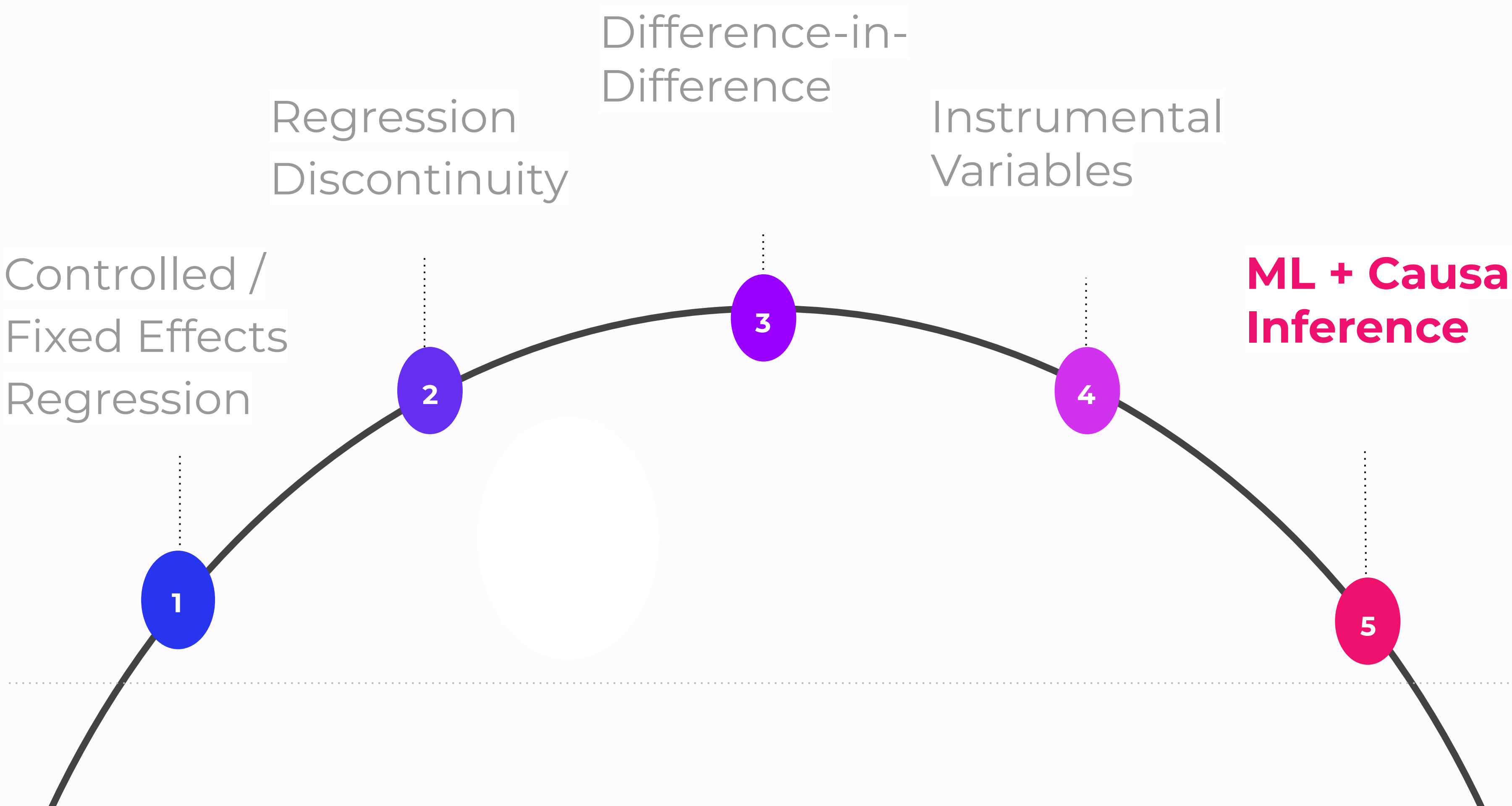
Assumption 2: Exclusion Restriction

Example: Study the impact of using the mobile app on course completion. Use an instrument created from a randomized nudge to download the mobile app, **so need it to have an impact on course completion only through its impact on mobile app usage**.

How to Check:
No test generally, so we need to use logic. **But, if construct a randomized encouragement trial where create instrument as a randomly assigned nudge that prompts X variable of interest, we can ensure exclusion restriction through random assignment (and like a strong first stage as well!)**.

Therefore, randomized encouragement trials are great in companies where can nudge customers to take the action we care about measuring impact of.

Econometric Methods for **Causal Inference**



Machine Learning + Causal Inference

Weaknesses of classic causal approaches:

- Fail with many covariates
- Model selection unprincipled
- Generally assumes linear relationships and no interactions

Benefits of ML:

+ Can handle high dimensionality
+ Principled ways to choose model
+ Many nonlinear models that implicitly use higher order features

Machine Learning + Causal Inference

Idea: Use variables or reasonable proxies to isolate causal relationship of variable of interest by controlling for other factors

Standard Steps

- Regress Y on X and a set of controls C to identify coefficient of interest on X
- Be wary of omitted and included variable biases

Machine Learning + Causal Inference

Idea: Use variables or reasonable proxies to isolate causal relationship of variable of interest by controlling for other factors

ML Flavor

- Use ML Models to control for many potential confounders and/or nonlinear effects
- Two types (note theory mostly developed for binary treatment but should generalize):
    - Double Selection (Lasso)
    - Double Debiased (Generic ML models)

Machine Learning + Causal Inference

Steps

- Have Y and treatment indicator X, high dimensional set of controls C
- Split data into two sets: Tr, Te*
- Fit two Lassos of X~C and Y~C on Tr
- Take fitted models and apply to Te
- Get all nonzero variables in C and use as controls in controlled regression of Y on X

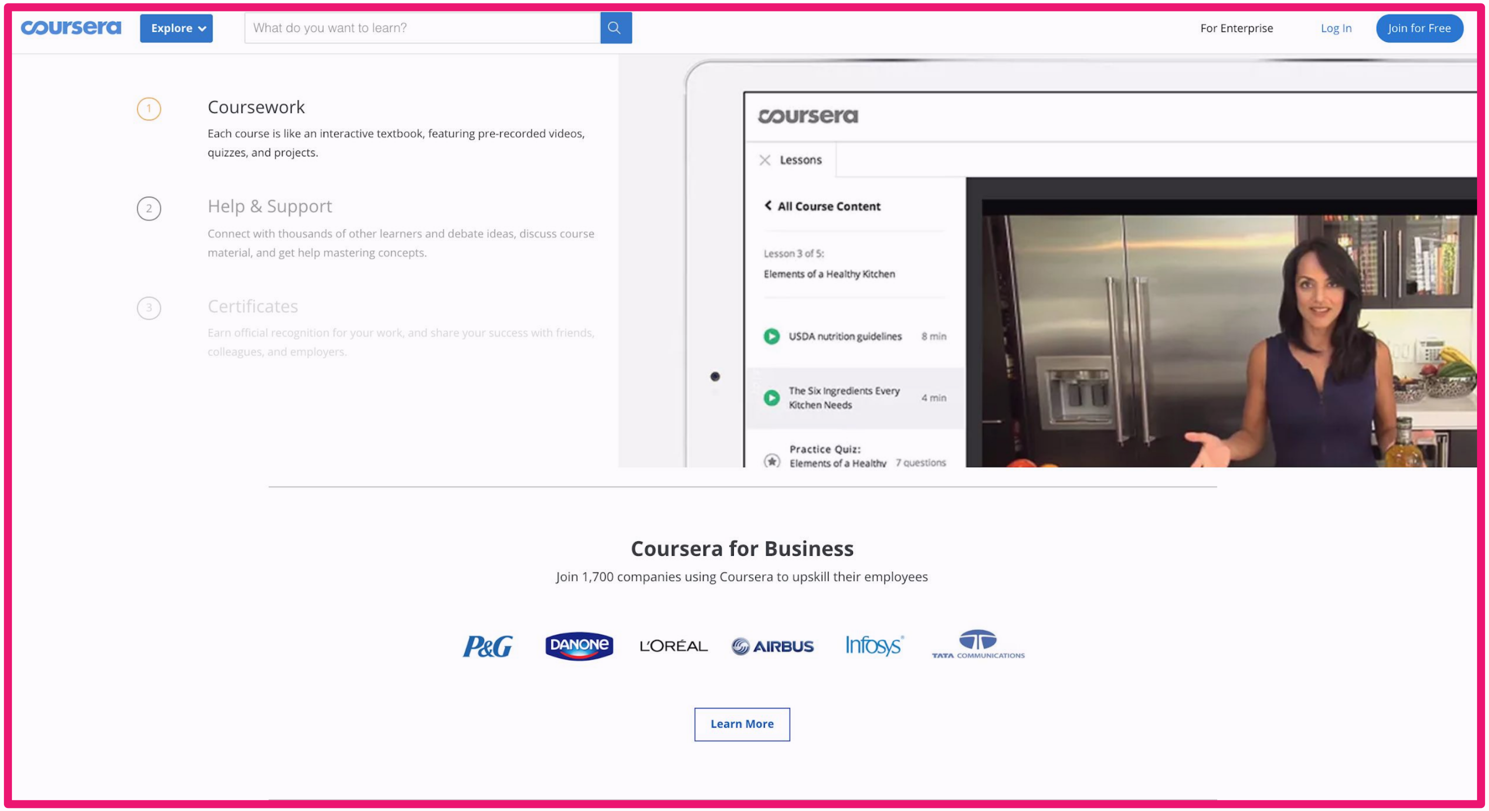
*Can generalize to K-folds

ML + Causal Inf: AB Testing

Idea: Perform Double Selection on AB test data with treatment assignment and large set of controls (that were fixed at beginning of experiment)

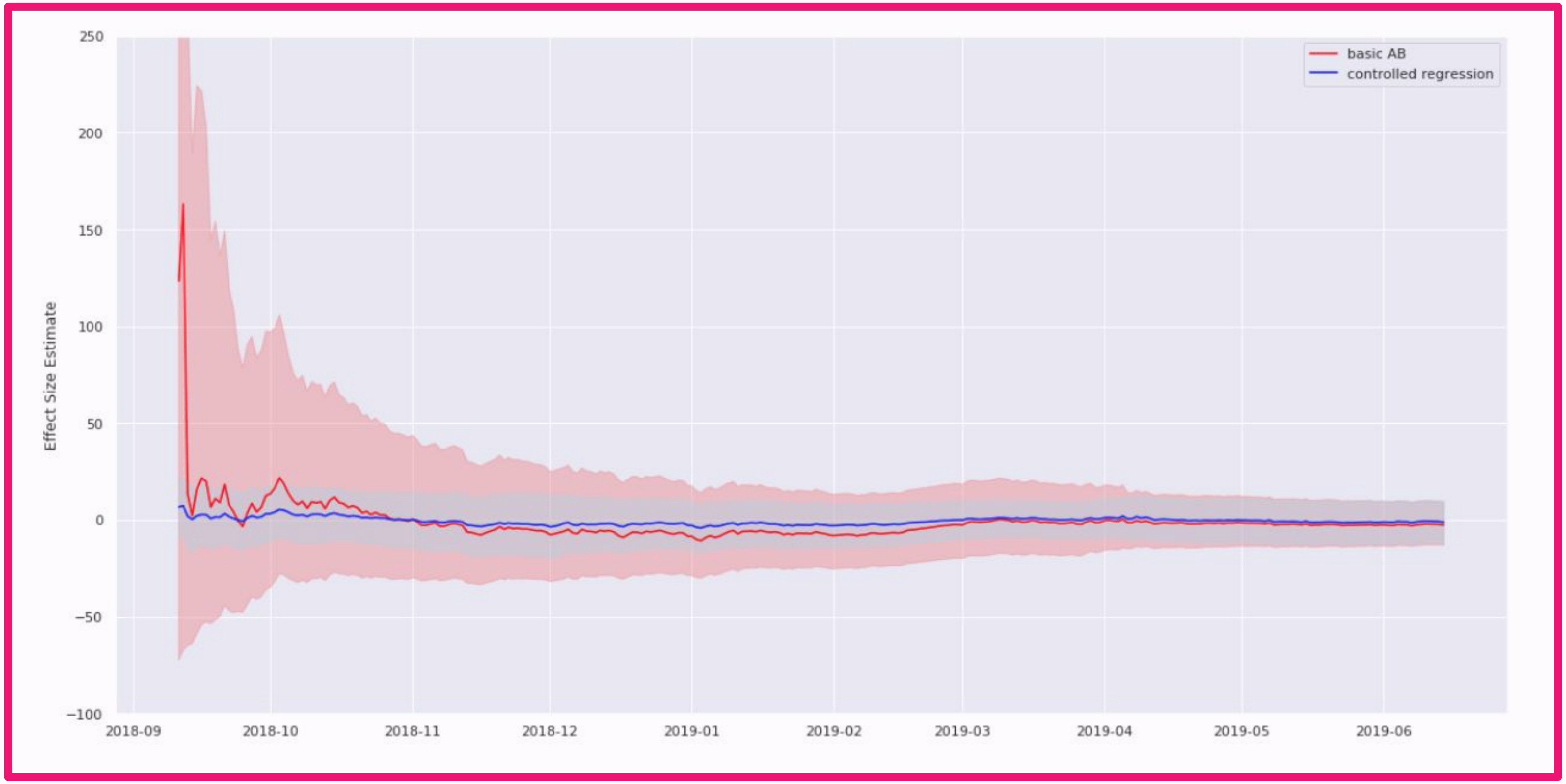
ML + Causal Inf: AB Testing

Example: Testing advertising of Coursera for Business; less traffic and small conversion rate



ML + Causal Inf: AB Testing

Benefits: Increased statistical power gives smaller confidence intervals and increased time to resolution; good for small samples and effect sizes



ML + Causal Inf: Causal Trees/Forests

Idea: Everything previously assumed homogeneous treatment effects. Causal trees/forests estimates heterogeneous treatment effects where impact differs on observed criteria.

Use trees (or forests) to identify partition of the space that maximizes observed difference of Y between treatment and control while balancing overfitting.

ML + Causal Inf: Causal Trees/Forests

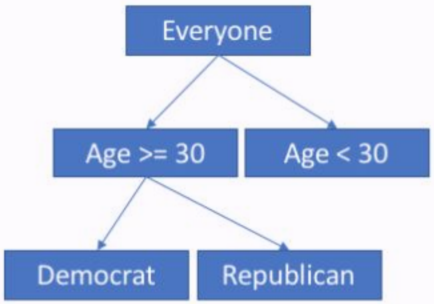
Steps:

- Split data into two halves
- Fit tree/forest on one half and apply to second half to estimate treatment effects
- Heterogeneous treatment effects from difference in Y in leaf nodes i.e. effect conditioned on C attributes in leaf nodes
- Optimization criteria set up to find best fit given the data splitting
- Forest is just average of a bunch of trees with sampling

ML + Causal Inf: Causal Trees/Forests

Sample from Randomized Experiment

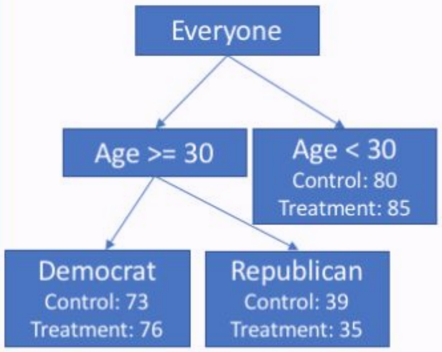
Splitting Subsample



Using the splitting criteria for a causal tree on this subsample, we find three groups in the data:

- People under 30
- Democrats 30 or older
- Republicans 30 or older

Estimating Subsample



We drop everyone in this subsample down the tree and find the percent favorable toward our candidate in each condition in each node. The differences are treatment effects:

- People under 30 = +5 points
- Democrats, 30 and older: +3 points
- Republicans, 30 and older: -4 points

Actual People We Are Trying to Target

We can only afford to target two of these people:

1. 19 year-old Republican
2. 25 year-old Democrat
3. 64 year-old Republican
4. 31 year-old Democrat

↓

Using tree fit by splitting subsample and treatment effects from estimating subsample, we predict the following effects on these people:

1. +5 points
2. +5 points
3. -4 points
4. +3 points

↓

Target people 1 and 2

Thank **you**

Additional Resources:

- **Mostly Harmless Econometrics**
- **Econometrics by Greene**
- **Econometrics & Causal Inference Online Courses**