An Introduction to *Causal Inference in Data Science*

Vinod Bakthavachalam | Data Science at Coursera

Data Scientist @ Coursera

- Focused on building a data driven content strategy and extracting skill development insights from our platform
- See our blog for examples of our work: https://medium.com/coursera-engineering/data/home
- Github Resources: https://github.com/b-vinod/ODSC-2019

Experimental Design
Econometrics
Machine Learning
Causal Inference

Does X Drive Y?

Central Data Science Questions Often Involve Causality

1. Did PR coverage drive sign-ups?
2. Does customer support increase sales?
3. Did improving the recommendations model drive revenue?
4. Why did ABC metric change this month?
5. …

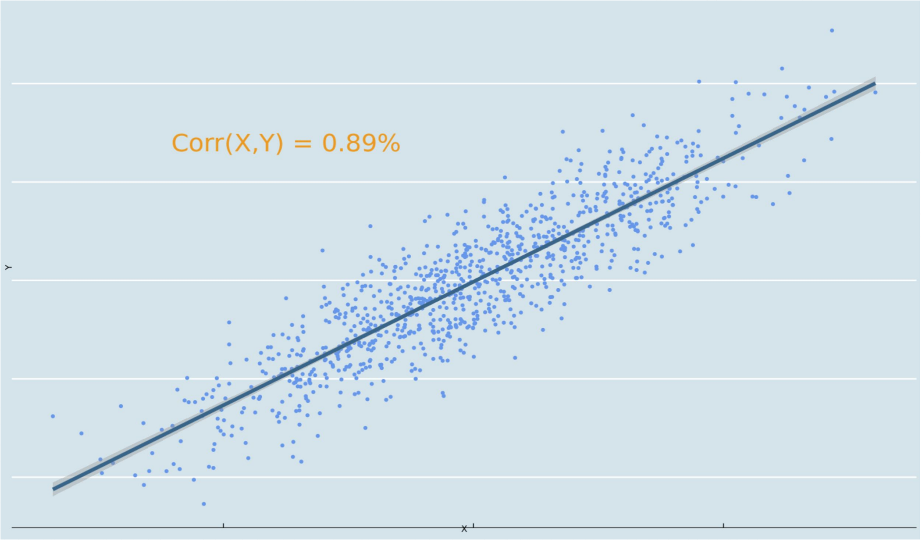
Adapted from previous work by ***Emily Glassberg Sands (Coursera) & Duncan Gilchrist (Uber)***

Does X Drive Y?

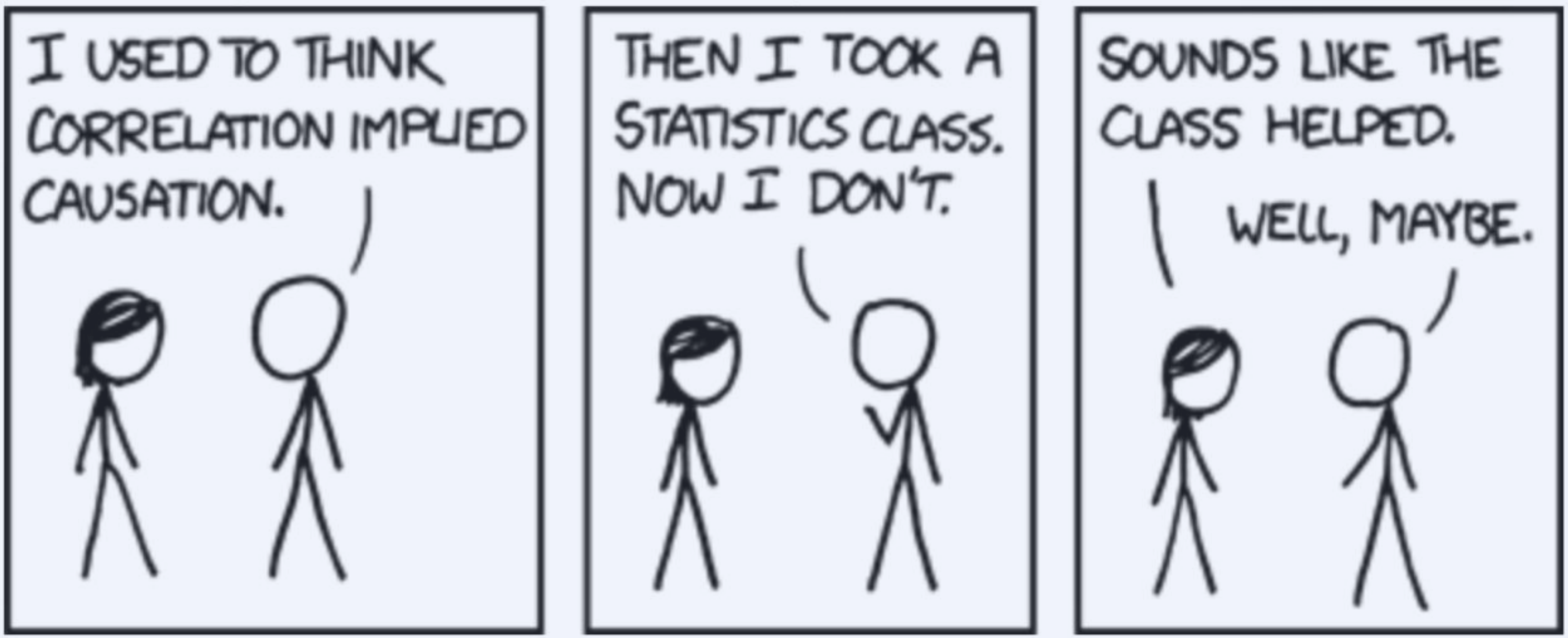
Start with Raw Correlation

Does X associate with increase in Y?

- Plot Y against X in a scatterplot
- Find and test corr(X, Y)



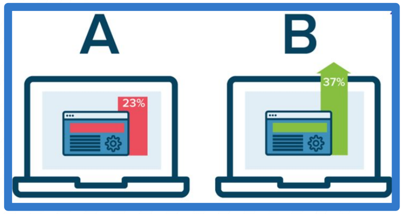
Correlation is not Causation!



http://xkcd.com/552/

Does X Drive Y?

AB Testing / Experimentation



- **Randomly assign** one group of users an experience and another group a different experience
- **Experience is uncorrelated** with any potential confounders
- **Difference between groups** is *causal effect / treatment effect* of the experience (X) on the outcome (Y)

Often best path forward…but not in all cases

Limitations of AB Testing / Experimentation

- *Product experience*
- *Ethics*
- *Trust*
- *Feasibility*

Examples: pricing, mobile app access, PR, etc.

Causal Inference to the Rescue!

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.*

Objectives of the Session

Identify Potential Applications

Learn how to recognize when AB testing is not feasible and how to apply causal inferences in those cases.

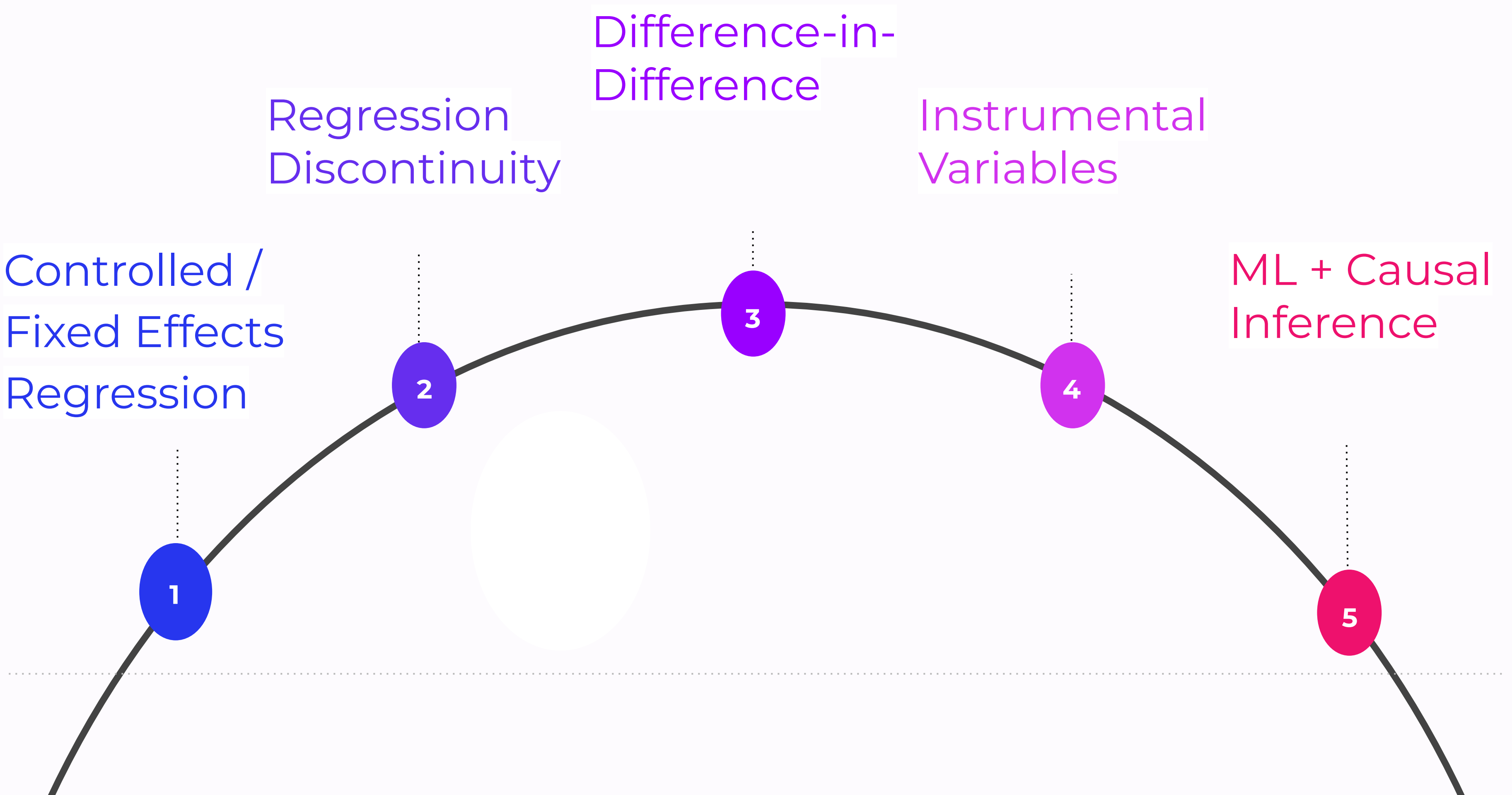
Select The Right Technique

Identify when one causal inference technique is preferred over another and how to match the best technique to the current problem.

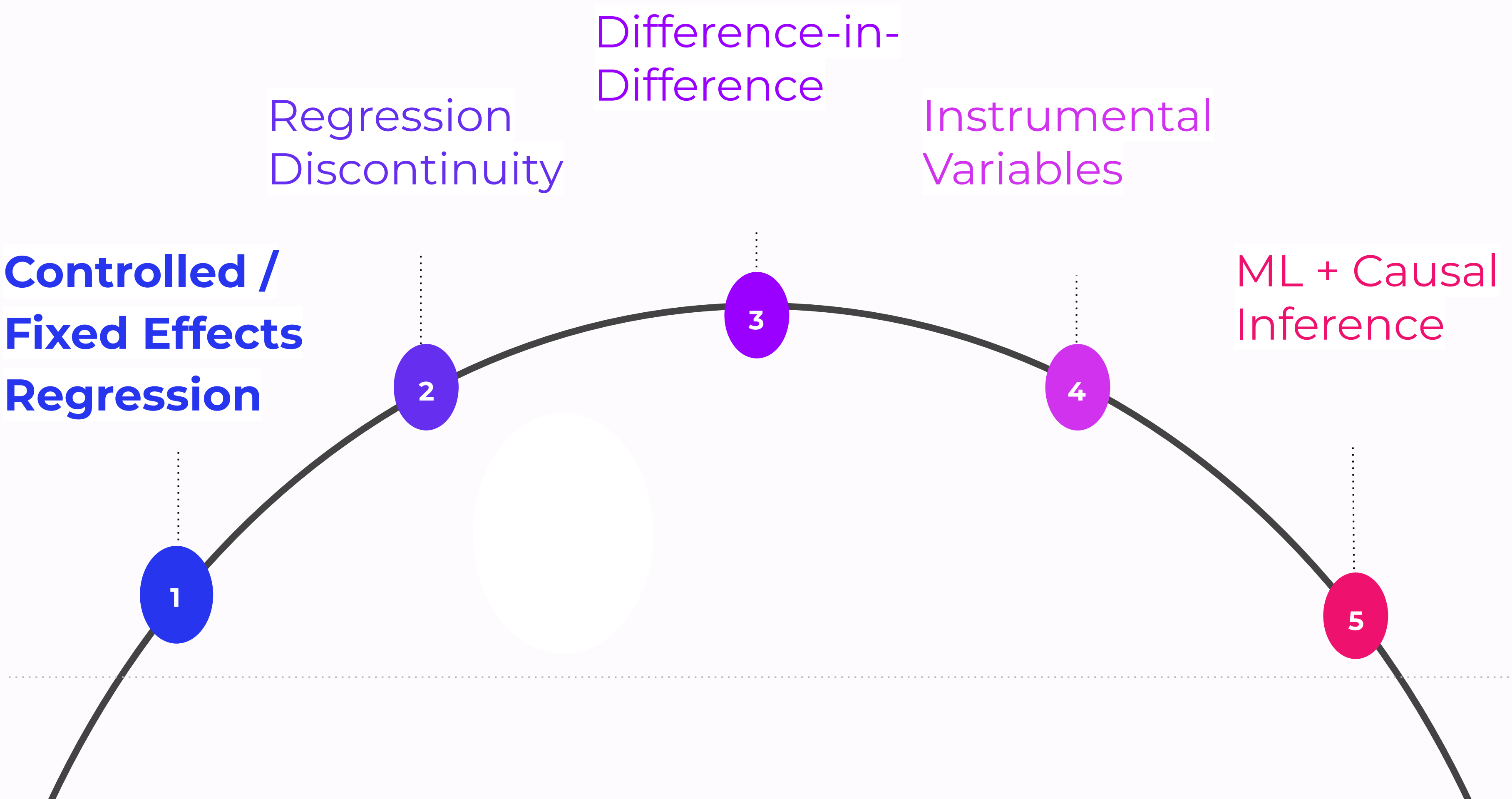
Expand your data science toolkit

Add a valuable skill to your data science toolkit and expand the set of business and product problems you can solve.

Econometric Methods for **Causal Inference**



Econometric Methods for **Causal Inference**



Method 1:

Controlled / Fixed Effects Regression

Idea: Control directly for the confounding variables in a regression of Y on X

Example: Effect of product quality on usage

- Product confounder →
  Demand may differ across product types
- Add controls for product characteristics

In R:

```
fit <- lm(Y ~ X + C, data = …)
summary(fit)
```

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

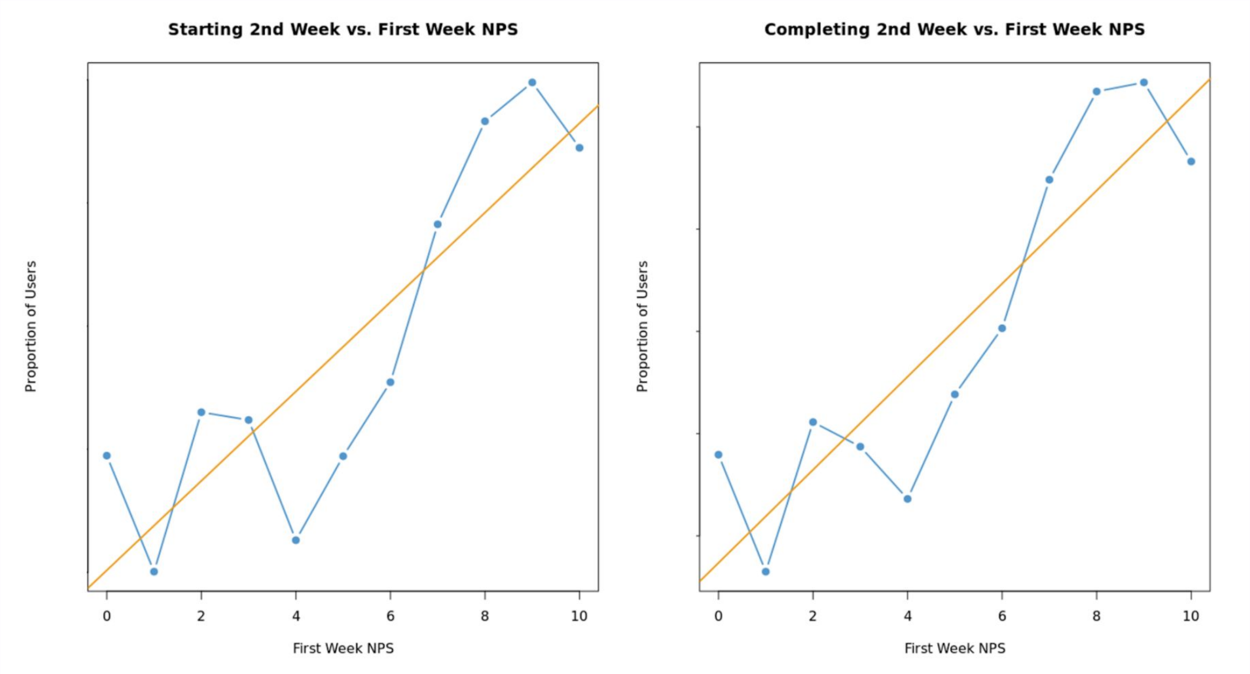
Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 1:
Controlled / Fixed Effects Regression

Example: Effect of course quality on completion

- Course length confounder → Completion rate differs by course length
- Add controls for course characteristics



Method 1:

Controlled / Fixed Effects Regression

Pitfall 1: "Missing" controls →

Omitted Variable Bias

Can we tell how much of a problem?

- If adding proxies increases (adjusted) R-squared without impacting estimate, could be ok…*

*Oster 15 provides a formal treatment.

Method 1:
Controlled /
Fixed Effects
Regression

Method 2:
Regression
Discontinuity

Method 3:
Difference-in-
Difference

Method 4:
Instrumental
Variables

Method 5: ML +
Causal Inference

Method 1:

Controlled / Fixed Effects Regression

Example: Effect of course quality on completion

- Course length confounder →
  Completion rate differs by course length
- Add controls for course characteristics

	Start 2nd Week ~First Week NPS (1)	Start 2nd Week ~First Week NPS+Controls (2)	Complete 2nd Week ~First Week NPS (3)	Complete 2nd Week ~First Week NPS+Controls (4)
First Week NPS	0.0070***	0.0065***	0.0063***	0.0074***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
R^2	0.0009	0.1015	0.0006	0.1842

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 1:

Controlled / Fixed Effects Regression

- ...*but* if adding proxies to regression impacts coefficient on X, regression won't suffice.

💔 **Adding controls DOES change point estimate**

Relationship between Instructor & Enrollee Gender

	Share Enrollments F	
Any Instructor F	.090***	.035***
	(0.0076)	(0.0074)
Controls	NO	YES
Adjusted R-squared	0.07	0.74
Base Group Mean	0.32	0.32

Watch for omitted variables biasing coefficient of interest

Method 1:

Controlled / Fixed Effects Regression

Pitfall 2: “Bad” controls →

Included Variable Bias

Example: Effect of course quality on completion

- Suppose think time available to take courses is a confounding factor.
- Control for other courses enrolled in?

💔 Not if directly impacted by treatment!

Leave out "controls" that are not fixed at the time of treatment (think of time traveling in ML feature engineering)

Method 1:

Controlled / Fixed Effects Regression

Idea: Special type of controlled regression

- most commonly used with panel data
- often to capture heterogeneity / unobserved differences across and products and time

Example: Estimate effect of price on conversion

- 1(pay) = α + β**Price* + X'θ + T'δ
  - X is vector of product fixed effects
  - ϴ is a vector of product-specific intercepts
  - T is vector of time fixed effects (e.g. month)
  - δ is a vector of time-specific intercepts
  - β is coefficient on interest (price sensitivity)

Method 1:

Controlled / Fixed Effects Regression

In R:

```
fit <- lm(Y ~ X + factor(Product)
          + factor(Time) + C,
          data = …)
summary(fit)
```

Note: Requires sufficient number of observations in each group have a fixed effect for i.e each product and time period combination.

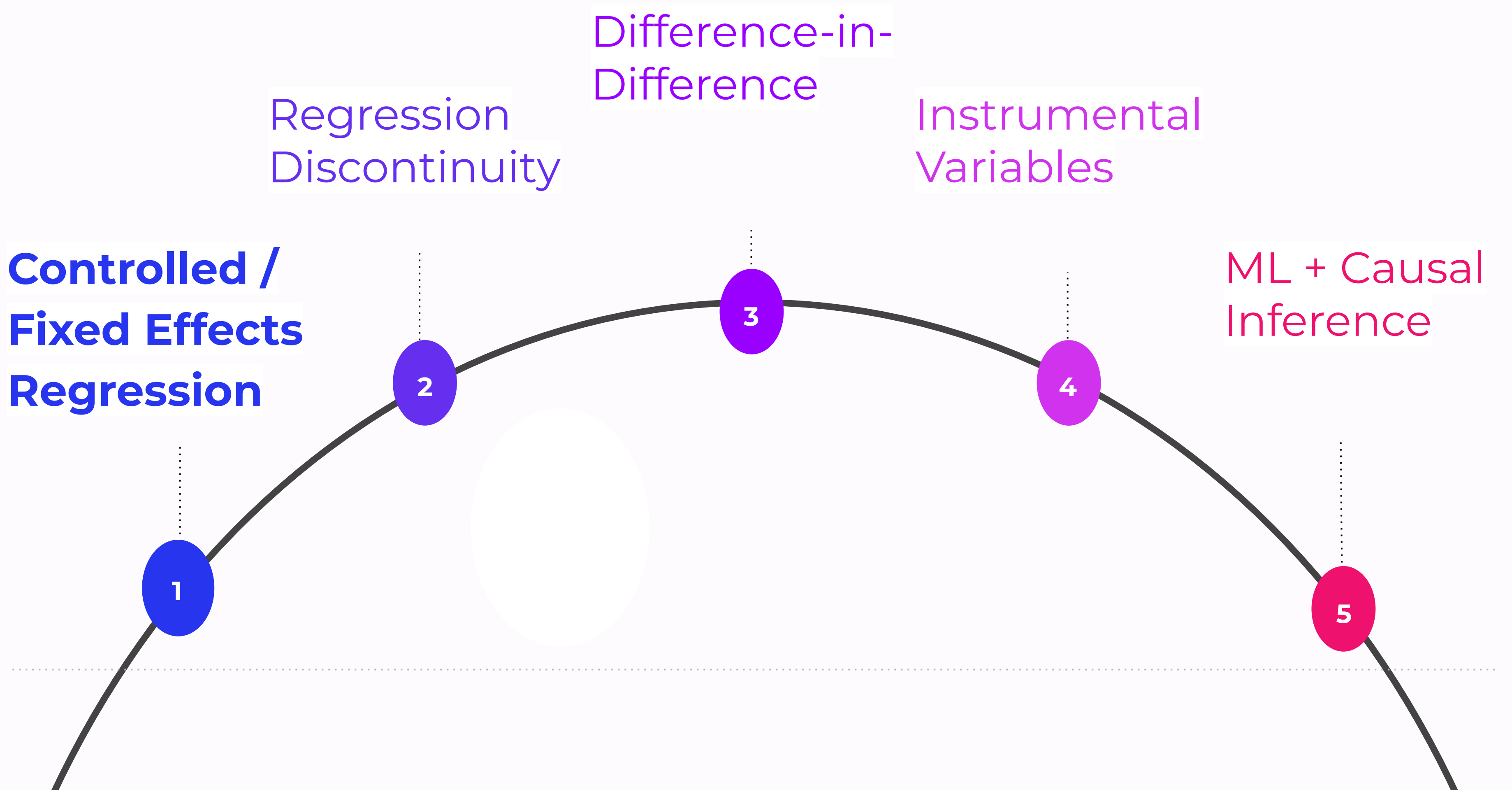
Note on Validity - **A/B testing**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	Randomized correctly, i.e. samples balanced
External validity	Unbiased for full population	Experimental group representative of overall

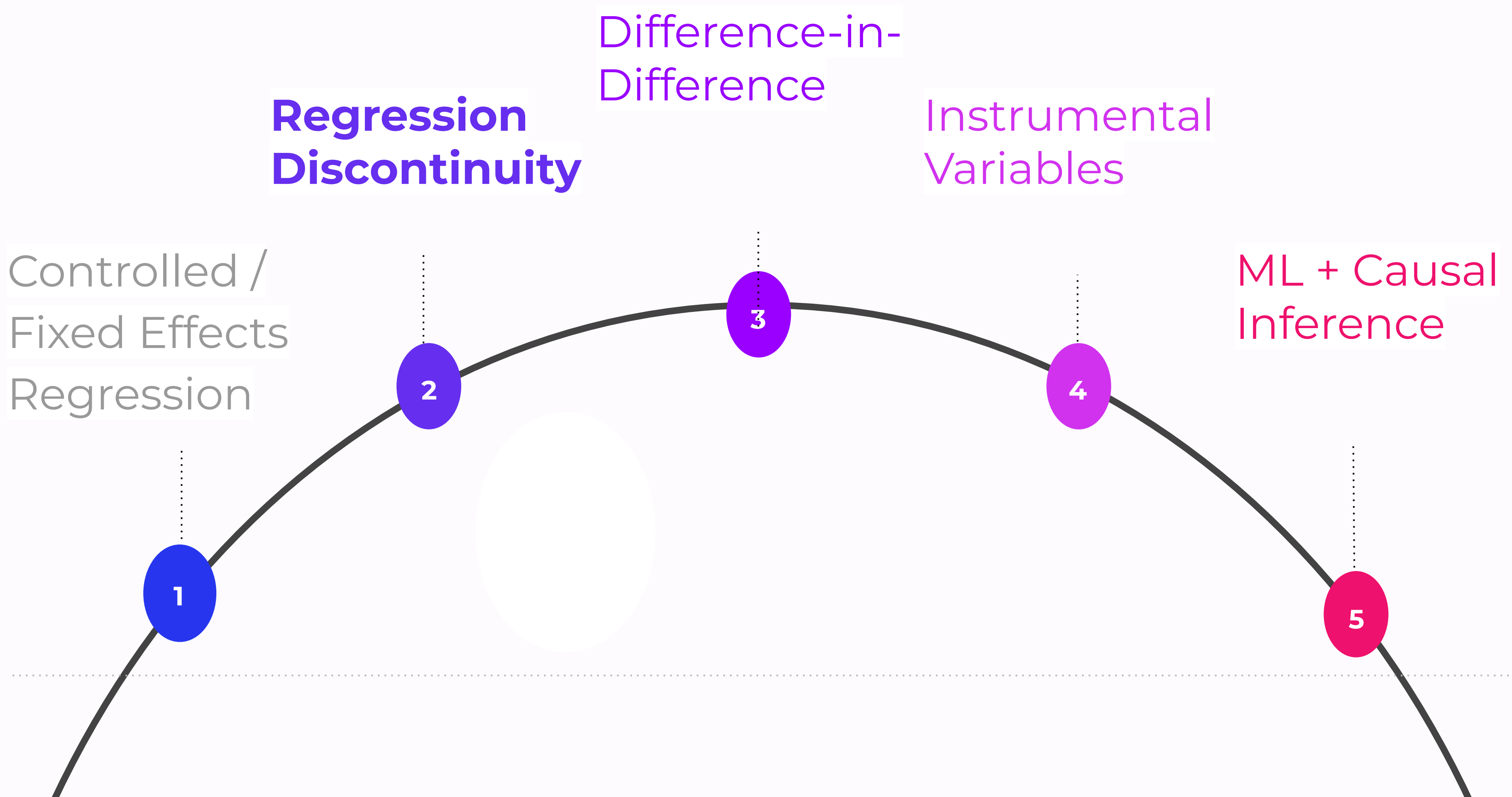
Note on Validity - **Fixed Effects**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	1. Imprecise control of assignment 2. No confounding discontinuities
External validity	Unbiased for full population	Homogeneous treatment effects

Example in R Time



Econometric Methods for **Causal Inference**



Method 2:

Regression Discontinuity Design

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

Example: Effect of adding subtitles to a course?

- A/B test? Randomly give some learners to access subtitles, difficult given product limitations
- Controlled regression? Key unobservables like course popularity

Method 2:

Regression Discontinuity Design

Example cont'd:

Launch cutoff → natural experiment

- Courses are advertised in a language only when they are at least 80% subtitled

In R:

```
fit <- lm(Y ~ X + I(X>Cutoff)
            + X*I(X>Cutoff) + C,
          data = …)
summary(fit)
```

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

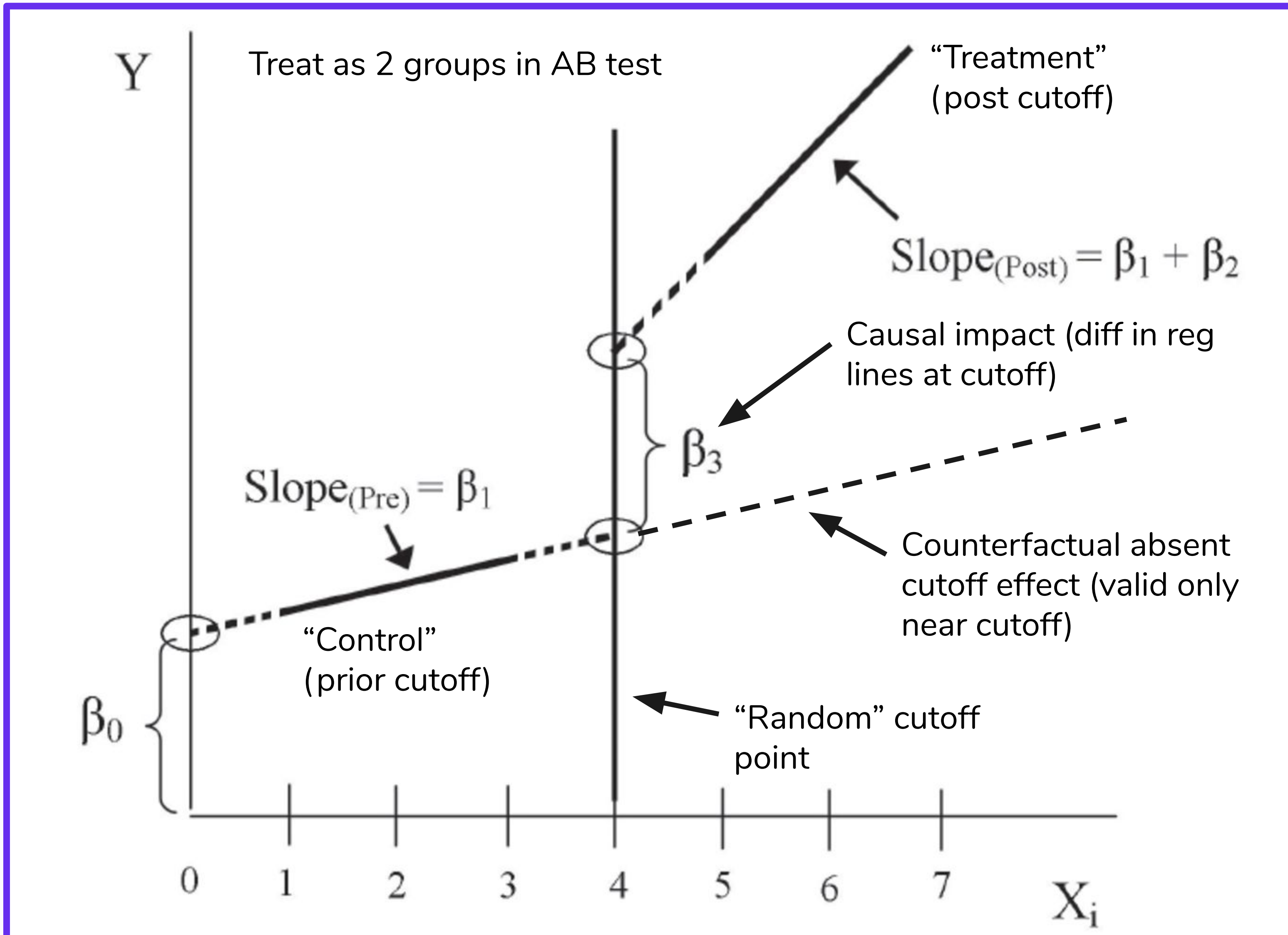
Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 2:

Regression Discontinuity Design



Method 2:

Regression Discontinuity Design

Difference in Regression Lines at cutoff is Causal Impact:

14.77 + 0.8 * (-17.02) ~ 0

	[illegible]	[illegible]	(3)
% Subtitled	[illegible]	[illegible]	1.02
	[illegible]	[illegible]	(7.69)
% Subtitled Above 80%	[illegible]	[illegible]	14.77
	(13.90)	[illegible]	(18.63)
% Subtitled X % Subtitled Above 80%	8.97	9.69	-17.02
	(15.97)	(15.97)	(21.51)
Lagged Enrollments	Yes	Yes	Yes
Language Fixed Effects	Yes	Yes	Yes
Time Trend	No	Yes	Yes
Course Fixed Effects	No	No	Yes

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Note on Validity - **Regression Discontinuity Design**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	1. Imprecise control of assignment 2. No confounding discontinuities
External validity	Unbiased for full population	Homogeneous treatment effects

Method 2:

Internal Validity in RDD

Assumption 1: Imprecise control of assignment, AKA no manipulation at the threshold

- Users cannot control whether just above versus just below the cutoff

In example: Across courses, process of advertising subtitles is the same with the 80% threshold rule i.e. no relationship between course attributes and when advertised.

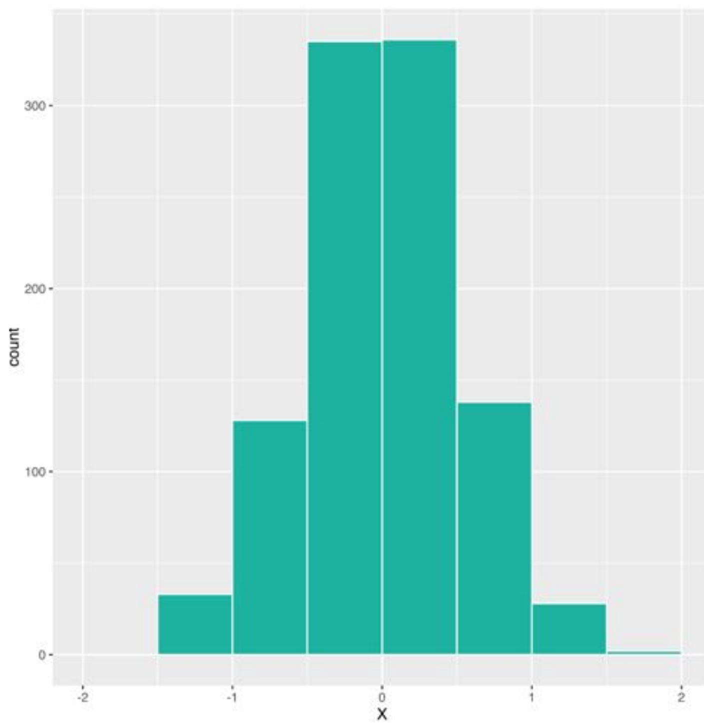
How can we tell?

Method 2:

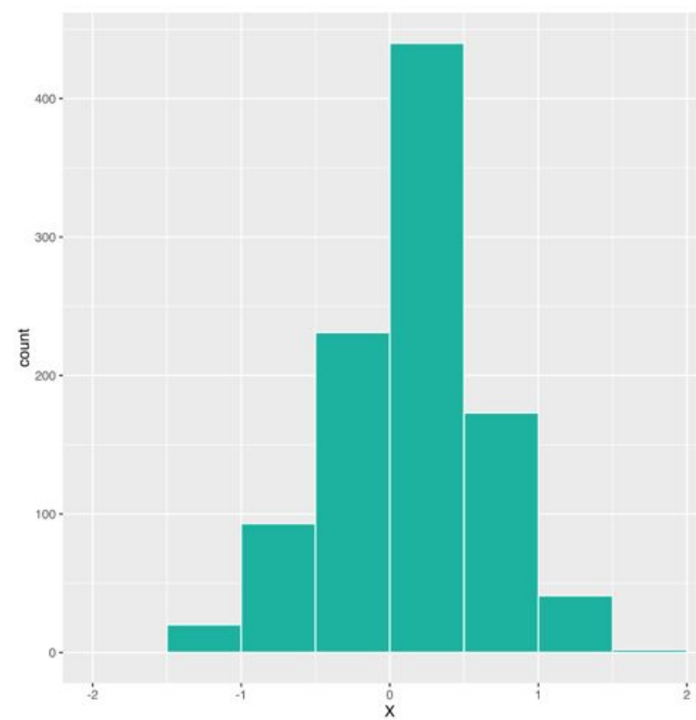
Internal Validity in RDD

Check 1: Mass just below ~= Mass just above

✓ **Even mass around cut-off**



💔 **Agency over assignment**



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

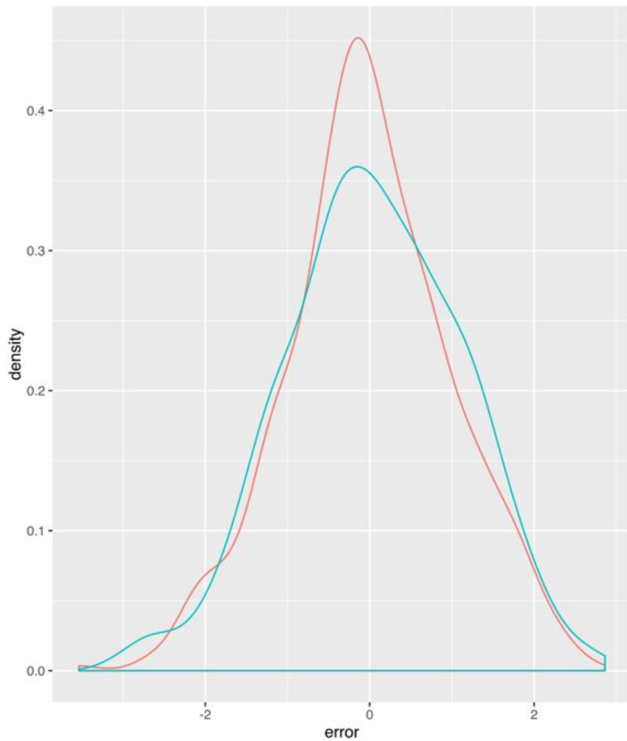
Method 5: ML + Causal Inference

Method 2:

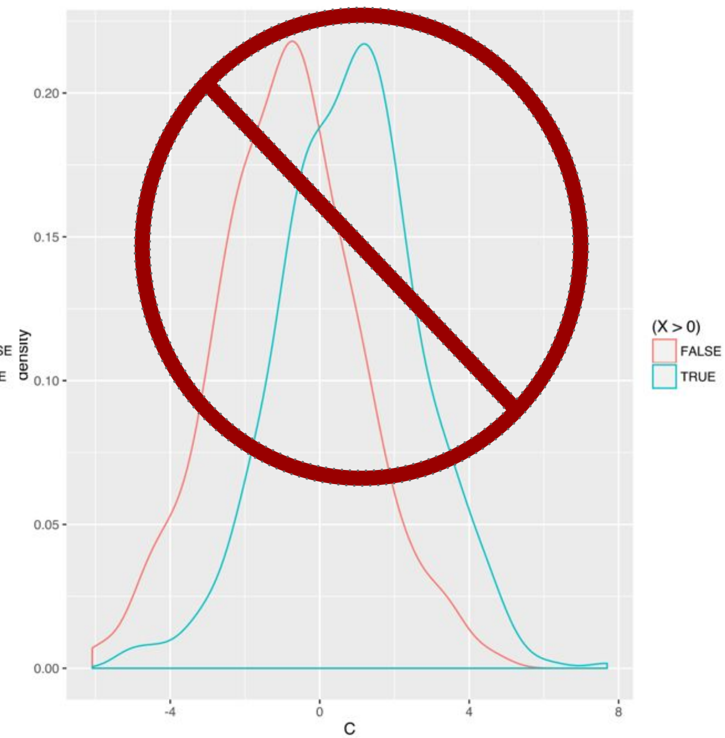
Internal Validity in RDD

Check 2: Composition of users in two buckets similar along key observable dimension(s)

✓ **Similar on observable**



Different on observable



Check for manipulation at the threshold

1. Mass just below ~= Mass just above?
2. Just below vs. just above similar on key observables?

Method 2:

Internal Validity in RDD

Assumption 2: No confounding discontinuities

- Being just above (versus just below) the cutoff should not influence other features

In example: Assumes advertising of subtitles is the only differentiator between 70% and 90% (for example no emails of content saying this is coming soon, etc.)

Placebo tests where run regression discontinuity at points other than the cutoff and check for no effect

Note on Validity - **Regression Discontinuity Design**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	1. Imprecise control of assignment 2. No confounding discontinuities
External validity	Unbiased for full population	Homogeneous treatment effects

Method 2:

External Validity in RDD

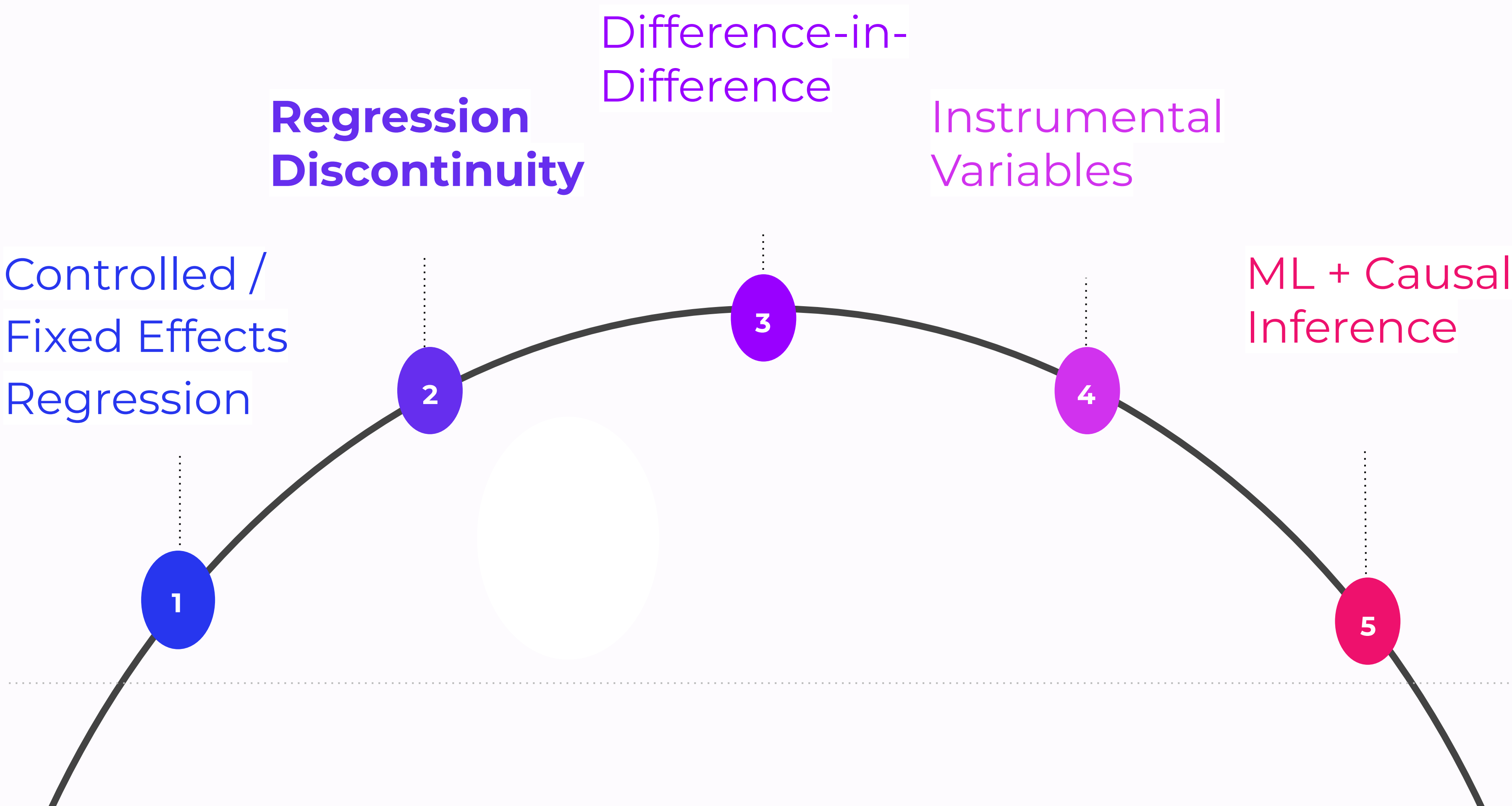
LATE: RDD estimates Local Average Treatment Effect (LATE)

- "Local" around the cut-off

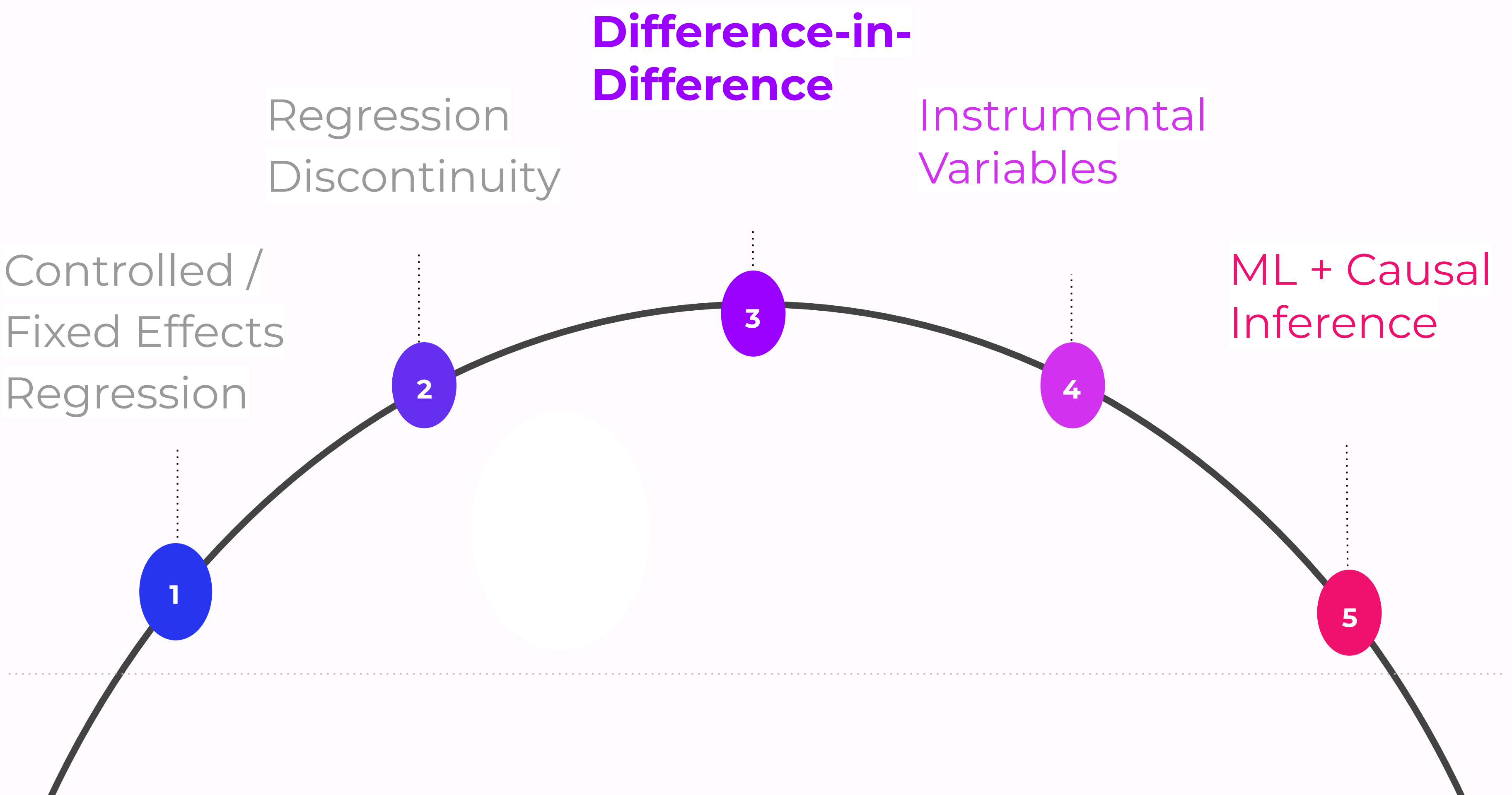
If heterogeneous treatment effects may not be applicable to the full group.

But interventions we'd consider would often occur on margin anyway

Example in R Time



Econometric Methods for **Causal Inference**



Method 3:

Difference-in-Differences

Idea: Comparison of pre and post outcomes between treatment and control groups

Example: Effect of lowering price on revenue?

- A/B test? Could, but may be perceived as unfair
- Alternative: Quasi-experimental design + DD

DD design: Change price in some geos (e.g., countries) but not others. Use control markets to compute counterfactual in treatment markets.

DD more robust than RDD so design for DD where feasible; controls for contemporaneous shocks

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

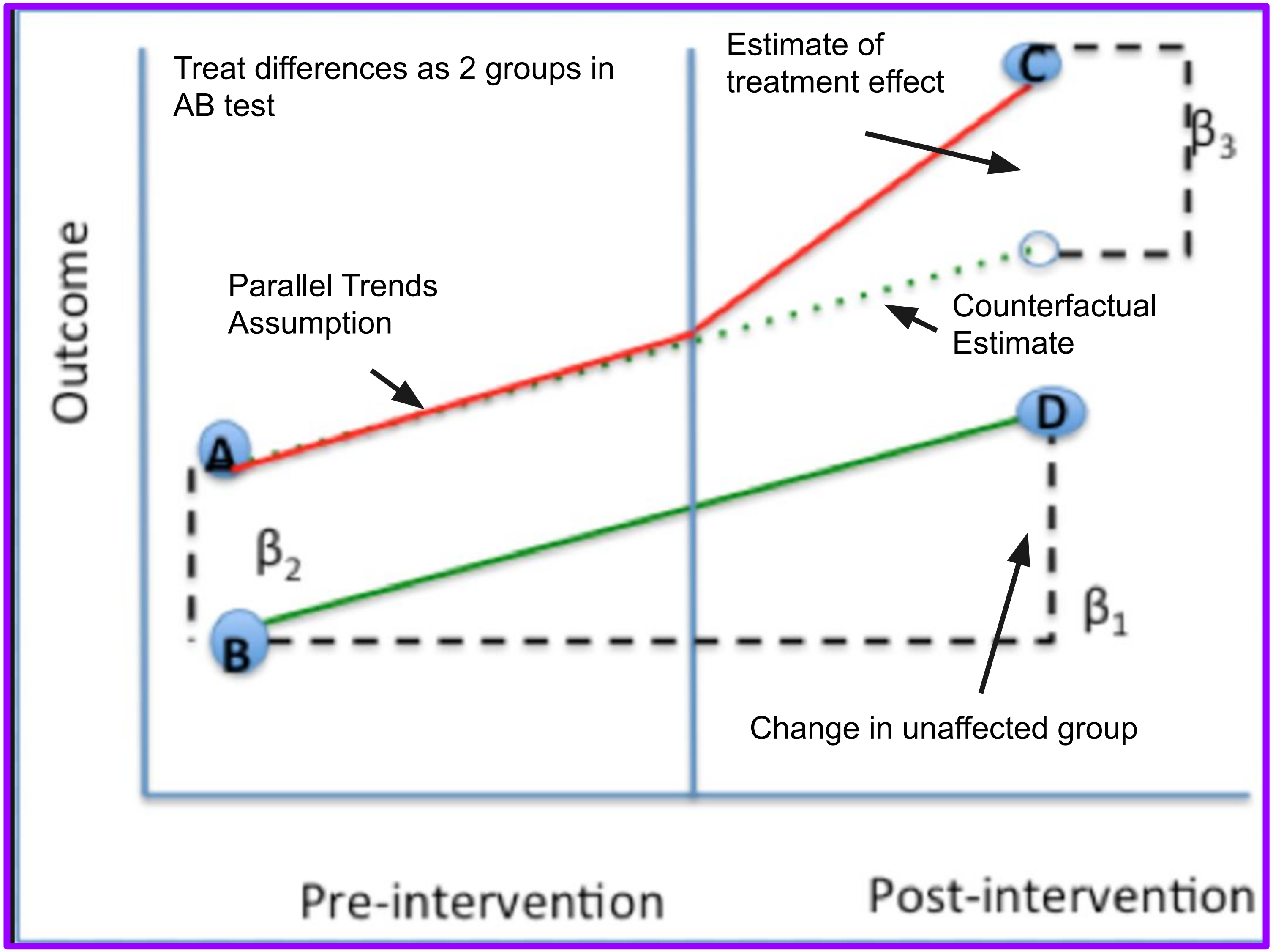
Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 3:

Difference-in-Differences



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 3:

Difference-in-Differences

In R:

```
fit <- lm(Y ~ treatment + post +
              treatment : post + C,
          data = … )
summary(fit)
```

Note on Validity - **Difference-in-Differences**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	Parallel trends
External validity	Unbiased for full population	Homogeneous treatment effect

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 3:

Internal Validity in DD

Assumption: Parallel trends

- Absent treatment, same trends

In example: Treatment and control markets would have followed same trends if no price change

How can we tell?

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

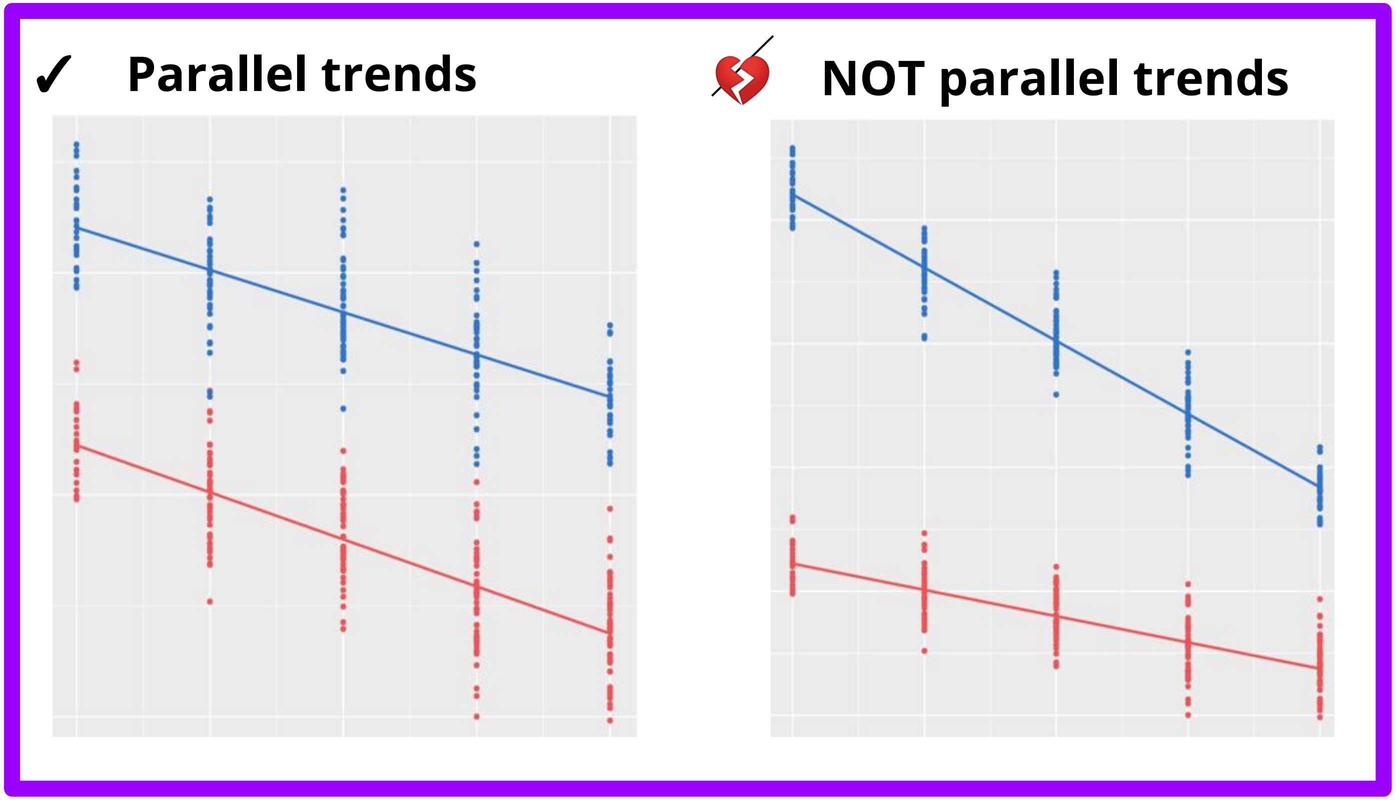
Method 5: ML + Causal Inference

Method 3:

Internal Validity in DD

Pre-experiment (cont):

- Check graphically & statistically that pre-experiment trends parallel



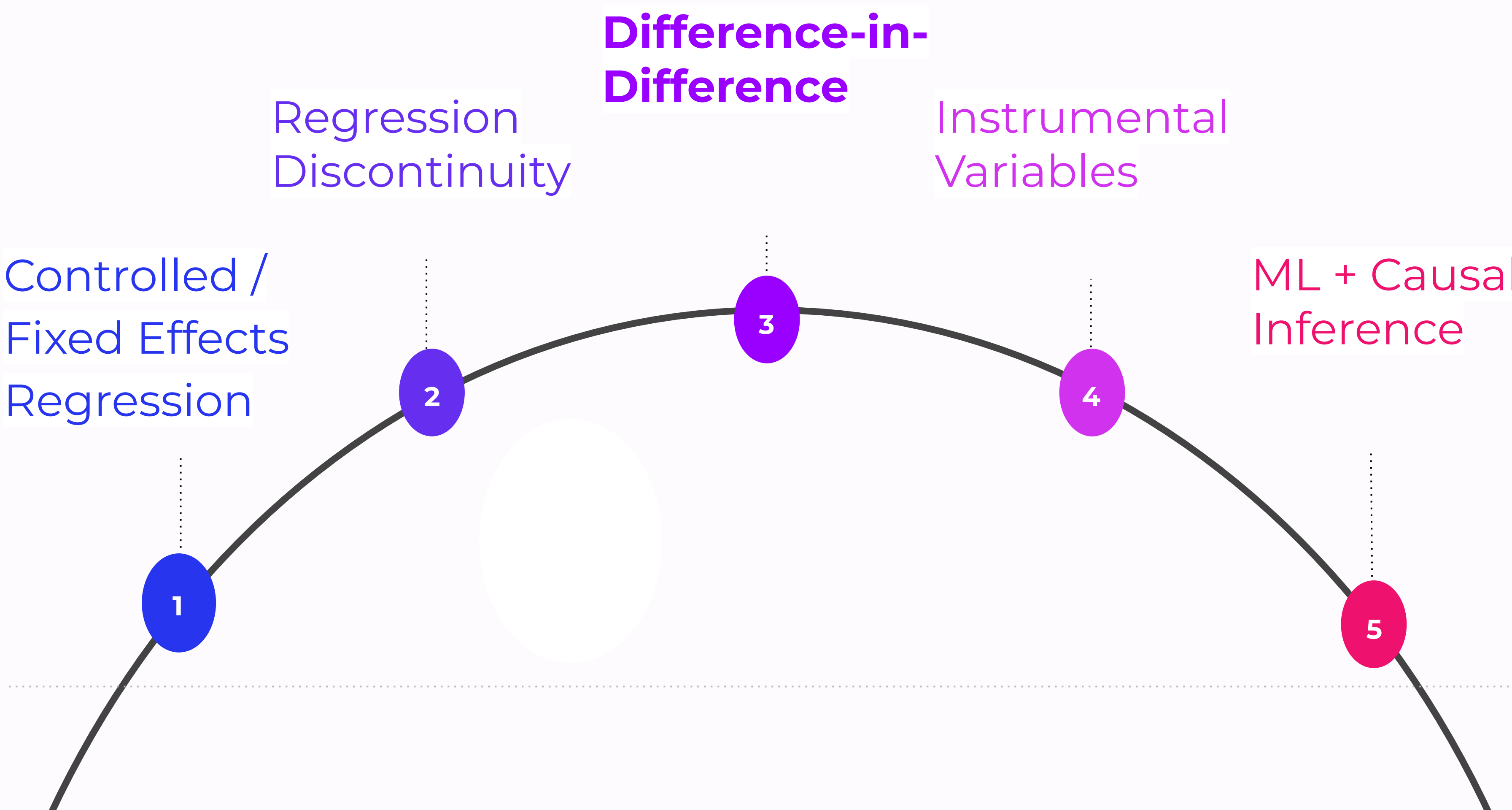
Design DD for parallel trends

1. Check: parallel trends ex ante
2. Stratify into/create groups that expect to be similar (such as through propensity score balancing/sampling)
3. Placebo tests:
   a. Run DD for two markets without treatment and see if no effect + parallel trends
   b. Run DD for two markets at time point prior to intervention and see if no effect

Note on Validity - **Difference-in-Differences**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	Parallel trends
External validity	Unbiased for full population	Homogeneous treatment effect

Example in R Time



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 3:

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

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

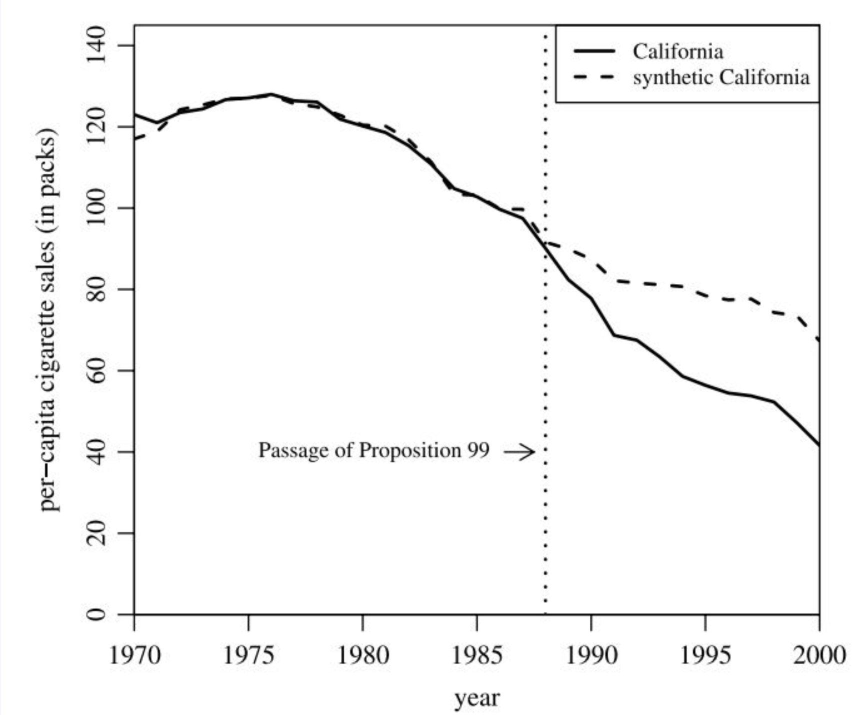
R Package: Synth

Method 3:

Extension: Synthetic Control

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0



Method 3:

Extension: Bayesian Approach

Bayesian structural time-series model

- Similar to synthetic control methodology where have control markets and infer post trend on treated group from a weighted average.
- Build a Bayesian prior and likelihood to dictate model instead as a Bayesian time series.

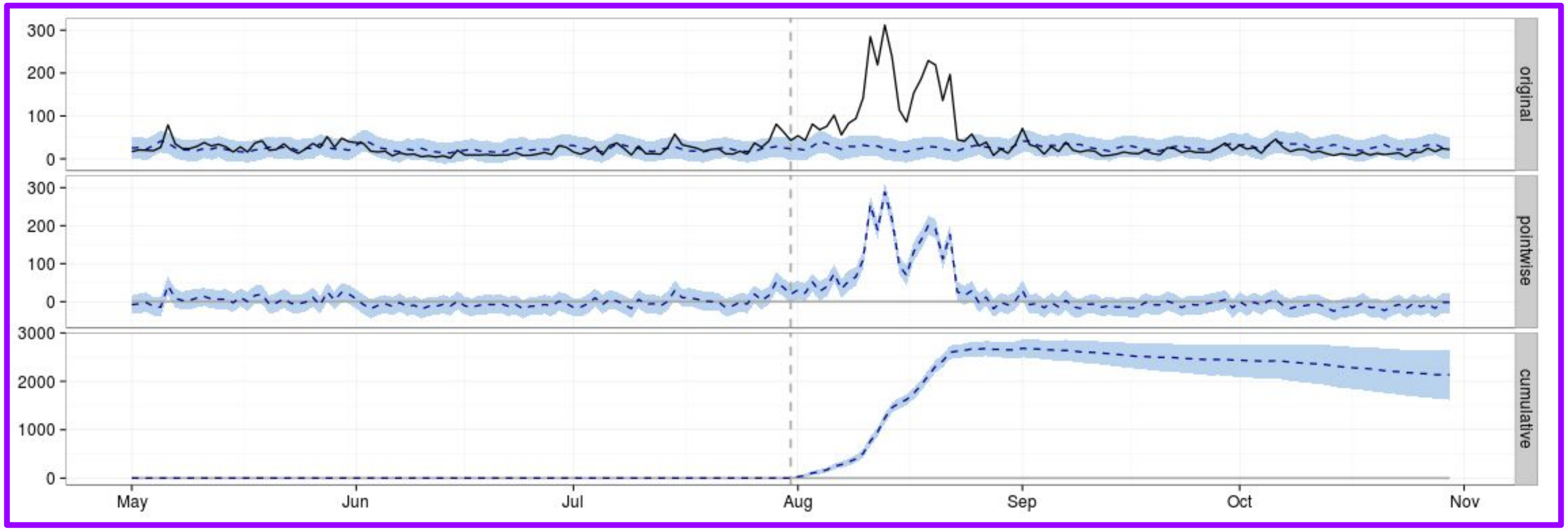
R Package: Causal Impact

Method 3:

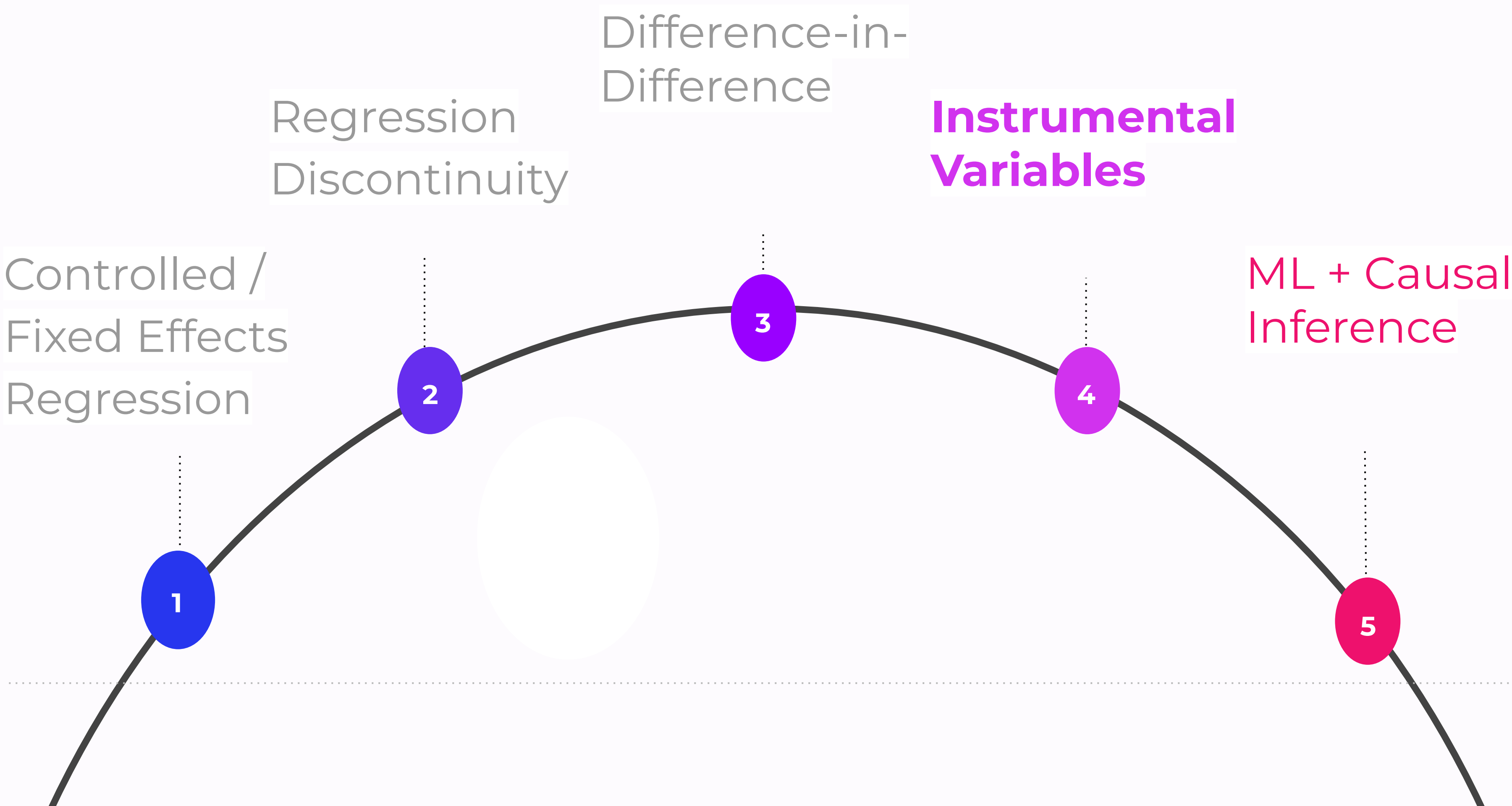
Extension: Bayesian Approach

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

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



Econometric Methods for **Causal Inference**



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 4:

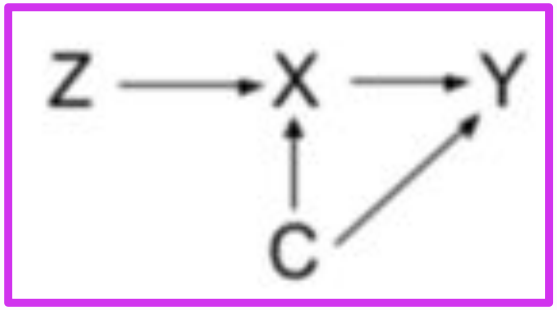
Instrumental Variables

Problem: Unobserved variable(s) C affect both X and Y; can't use controlled regression → Omitted variable bias with no proxy can use 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

Requirements:

- **Strong first stage**: Z meaningfully affects X
- **Exclusion restriction**: Z affects Y only through its effect on X



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 4:

Instrumental Variables

Implementation (Two Stage Least Squares):

1. Instrument for X with Z
   a. Regress X on Z and get fitted values Xhat
2. Estimate the effect of (instrumented) X on Y
   a. Regress Y on Xhat

In R:

```
library(aer)
fit <- ivreg(Y ~ X | Z, data = …)
summary(fit, vcov = sandwich,
        df = Inf, diagnostics = TRUE)
```

Method 4:

Instrumental Variables

Instruments in real world? Often look to policies

Y	X	Instrument	Economist(s)
Earnings	Education	Vietnam Draft lottery	Angrist
		Compulsory schooling laws	Angrist & Krueger
		Quarter of birth	Angrist & Krueger
Crime	Prison populations	Prison overcrowding litigation	Levitt
	Police	Electoral cycles	Levitt

Method 4:

Instrumental Variables

Instruments in tech? Everywhere! Especially old A/B tests → Useful for measuring long term metrics

Y	X	Instrument	Data Scientist
Platform retention	Having friends on the platform	Referral test 1	You!
		Referral test 2	You!
		Referral test 3	You!
		...	...

Method 4:

Instrumental Variables

Instruments in tech? Everywhere! Especially old A/B tests → Set up proxies for long term metrics

Y	X	Instrument	Data Scientist
Platform retention	Having friends on the platform	Referral test 1	You!
		Referral test 2	You!
		Referral test 3	You!
		...	...

Note on Validity - **Instrumental Variables**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	1. Strong first stage 2. Exclusion restriction
External validity	Unbiased for full population	Homogeneous treatment effect

Method 4:

Internal Validity in IV

Assumption 1: Strong first stage

- Experiment we chose “successful” at driving X

Why matters: If Z not strong predictor of X, *second stage estimate will be biased.*

How can we tell? Check F-statistic on the first stage regression; should be **> 11** (rule-of-thumb)

- `Diagnostics = TRUE in AER package’ in R will include test of weak instruments

```
Diagnostic tests:
                  df1 df2 statistic p-value
Weak instruments    1 998     839.9  <2e-16 ***
```

Method 4:

Internal Validity in IV

Assumption 2: Exclusion restriction

- Z affects Y only through X

How can we tell? No test; have to go on logic

In the example:

✓ Control group got otherwise equivalent email

💔 Control group got no email

Note on Validity - **Instrumental Variables**

Type	Definition	Assumptions
Internal validity	Unbiased for subpopulation studied	1. Strong first stage 2. Exclusion restriction
External validity	Unbiased for full population	Homogeneous treatment effects

Example in R Time

1. Controlled / Fixed Effects Regression
2. Regression Discontinuity
3. Difference-in-Difference
4. **Instrumental Variables**
5. ML + Causal Inference

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

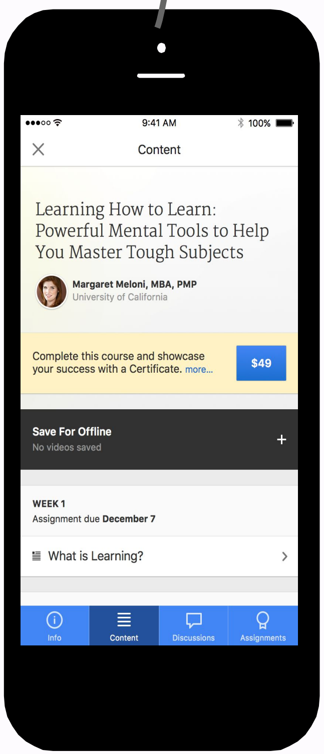
Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 4:

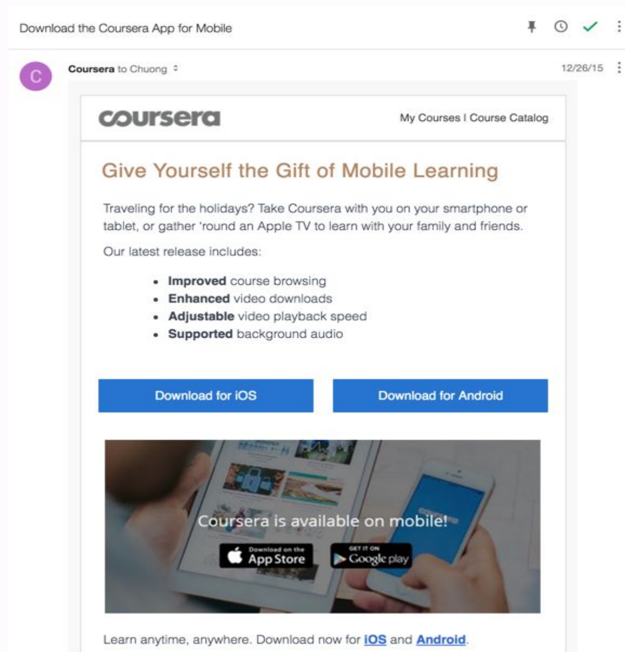
Make-Your-Own-Instrument!



Mobile Usage?	MoM Retention
No	35%
Yes	40%

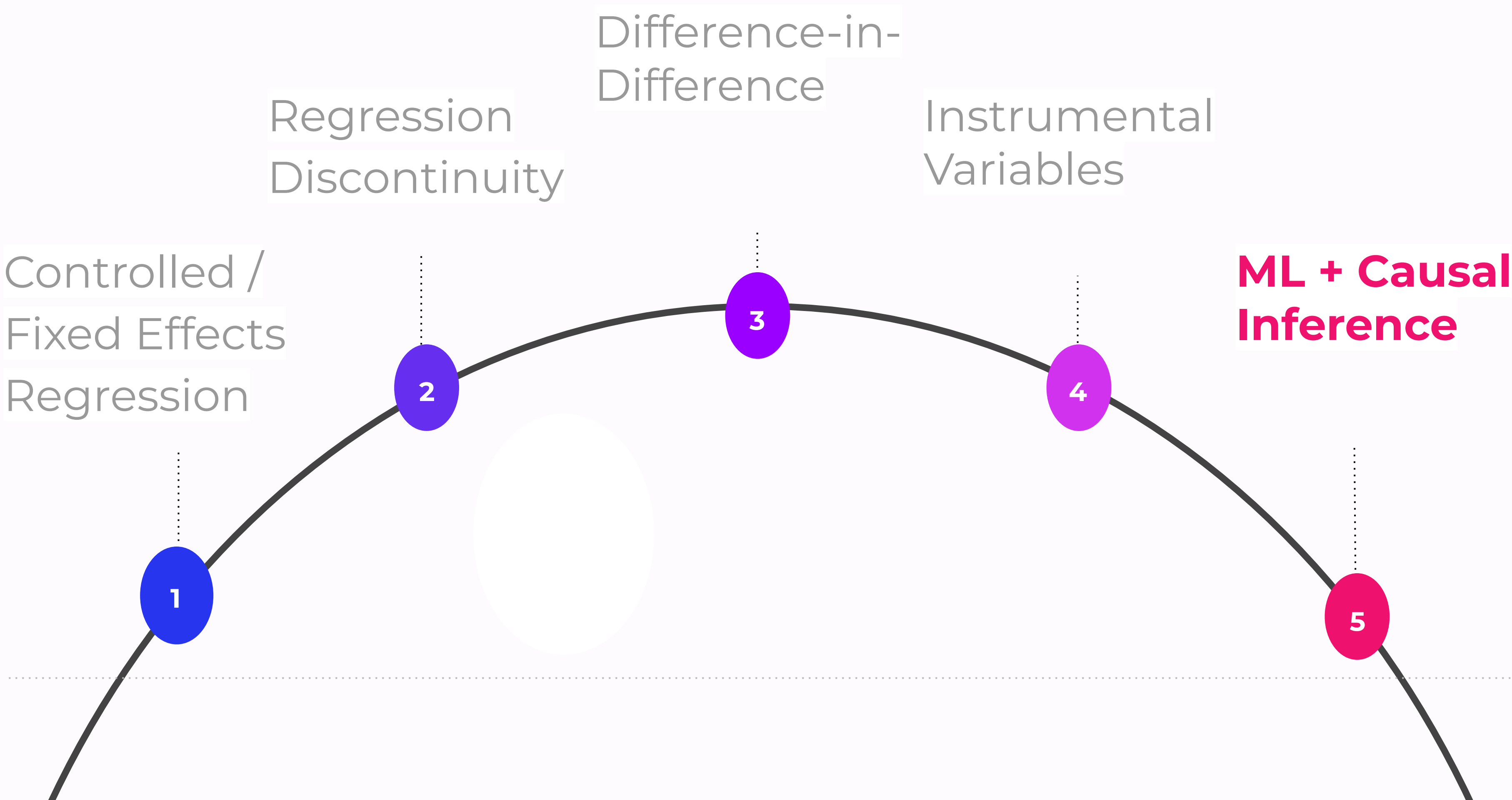
Selection Bias?

+



Instrumental variables via *randomized encouragement*

Econometric Methods for **Causal Inference**



Method 5:

ML + 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

Method 5:

ML + Causal Inf: Controlled Reg

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

Method 5:

ML + Causal Inf: Controlled Reg

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)

Method 5:

ML + Causal Inf: Double Selection

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

Method 5:

ML + Causal Inf: Double ML

Steps

- Have Y and treatment indicator X, high dimensional set of controls C
- Split data into two sets: Tr, Te*
- Fit two models (Rf, etc) for X~C and Y~C on Tr
- Take fitted models and apply to Te, find residuals
- Regress residuals on each other to estimate causal effect
- Reverse roles of Tr and Te sets, repeat
- Average resulting coefficients for final estimate

*Can generalize to K-folds

Method 5:

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)

Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

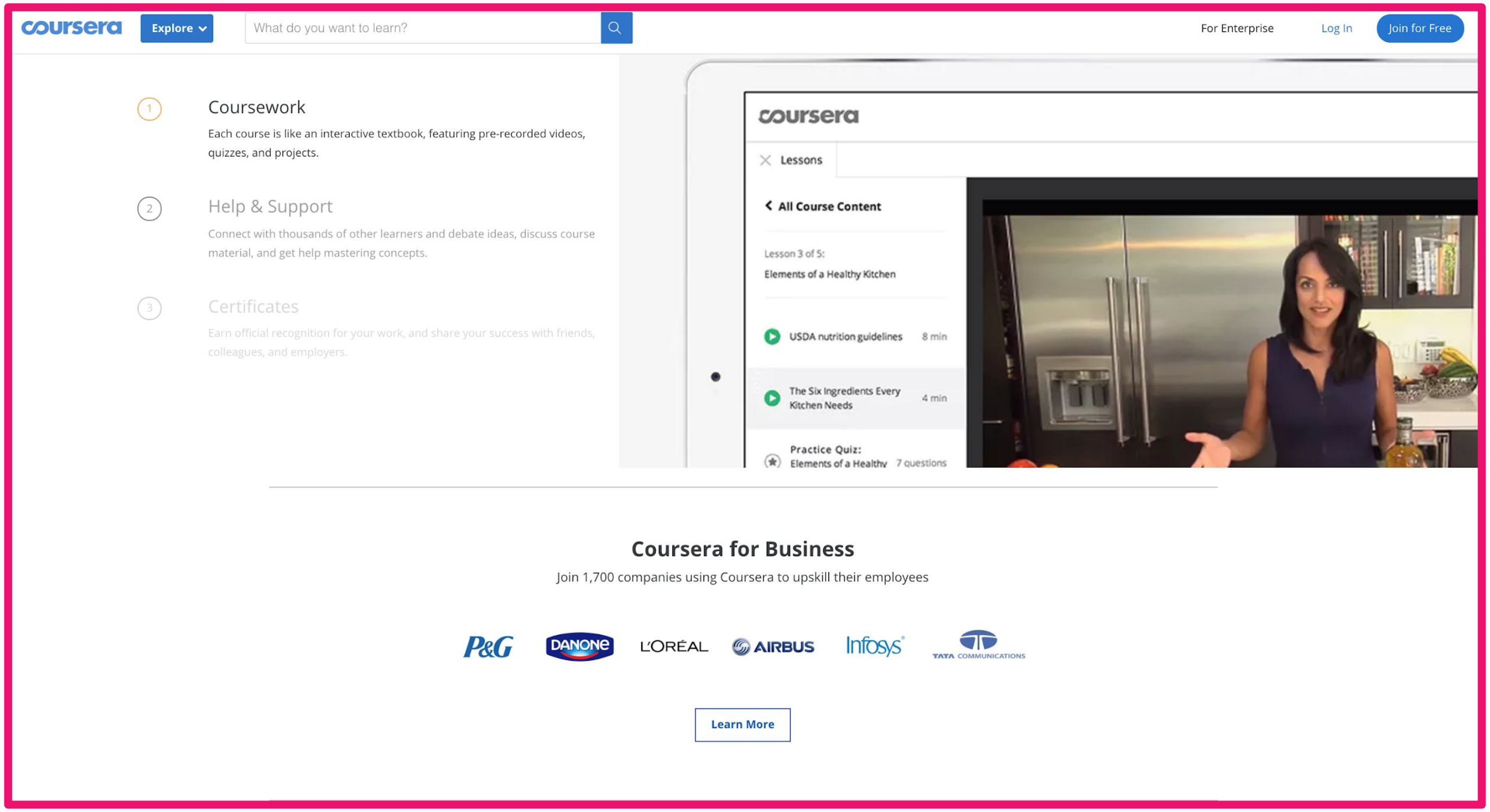
Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 5:
ML + Causal Inf: AB Testing App

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



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

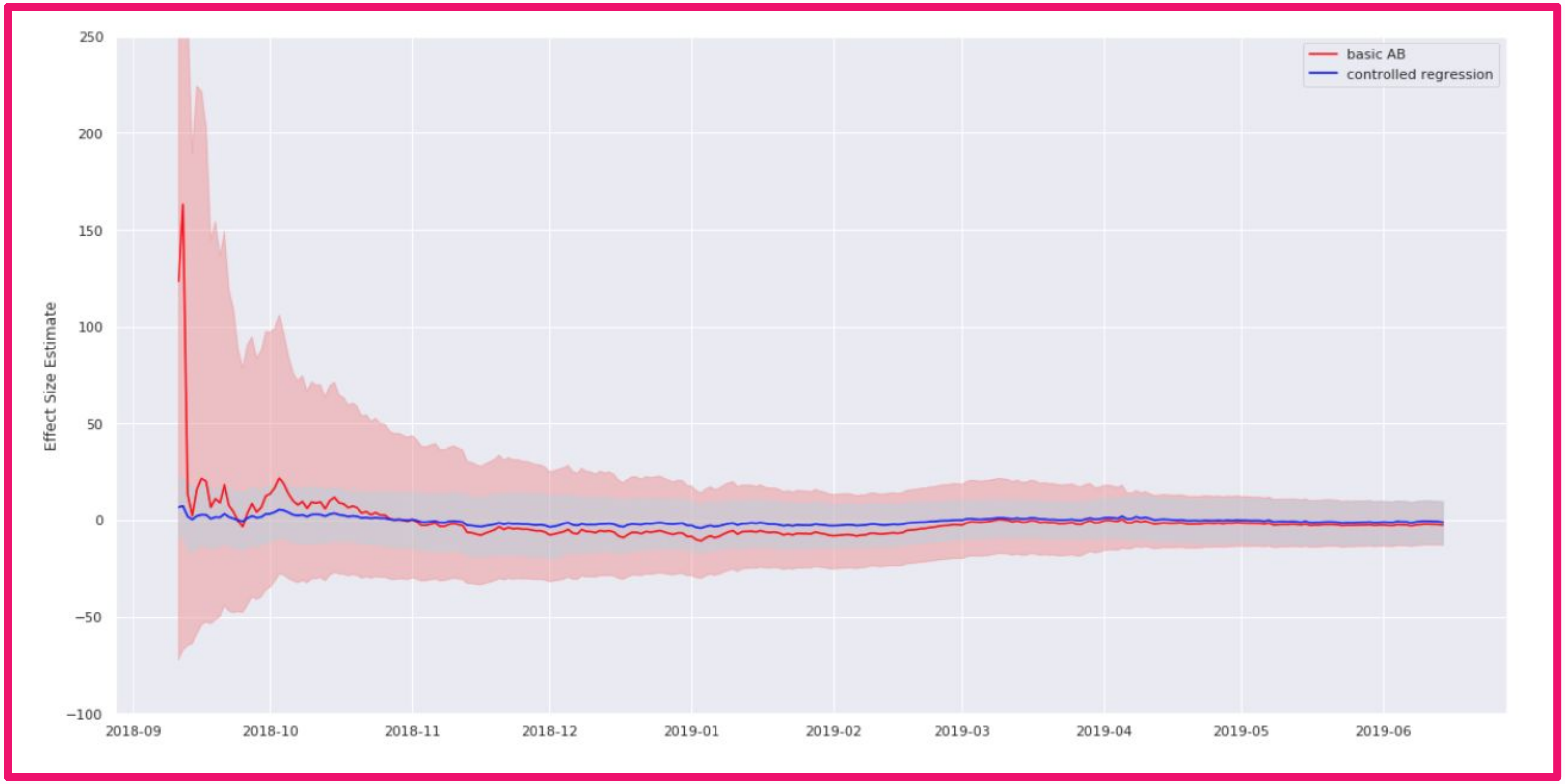
Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 5:

ML + Causal Inf: AB Testing App

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



Method 1: Controlled / Fixed Effects Regression

Method 2: Regression Discontinuity

Method 3: Difference-in-Difference

Method 4: Instrumental Variables

Method 5: ML + Causal Inference

Method 5:

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.

Method 5:

ML + Causal Inf: Causal Trees/Forests

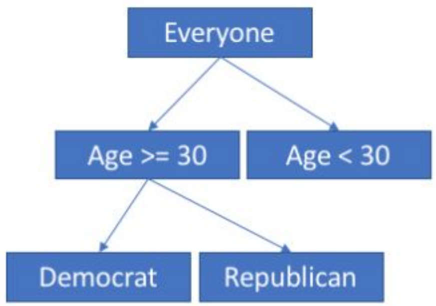
Steps:

- 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

Method 5:
ML + Causal Inf: Causal Trees/Forests

Sample from Randomized Experiment

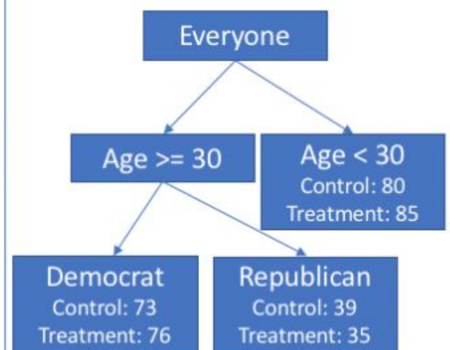
Splitting Subsample



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:

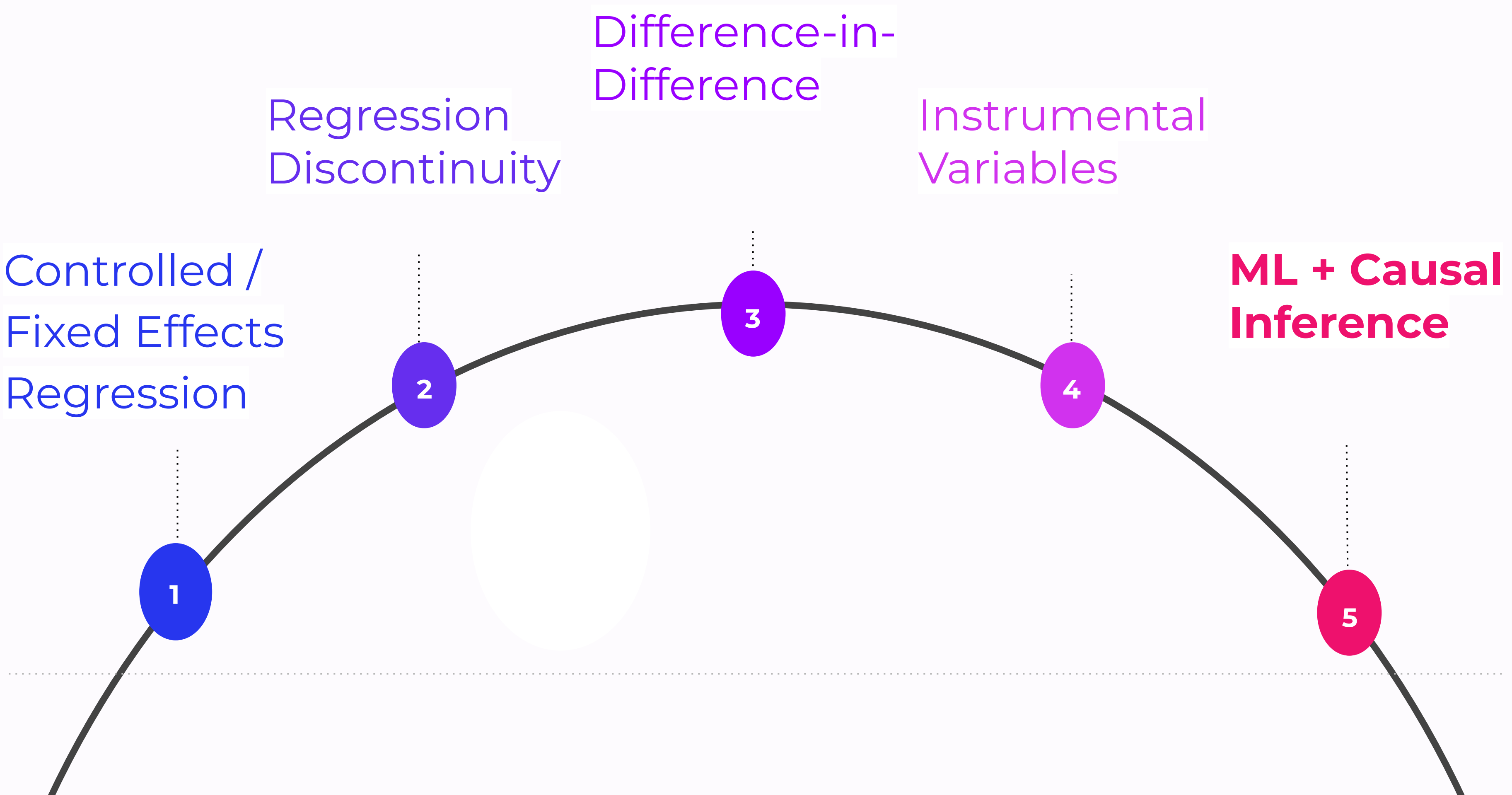
- People under 30 = +5 points
- Democrats, 30 and older: +3 points
- Republicans, 30 and older: -4 points

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

Example in R Time



Thank **you**

vinod@coursera.org

Additional Resources:

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