Women in Data Science

Pomona College, Claremont, CA, USA
April 1, 2022
Mitra Akhtari, Airbnb

Data Science at Airbnb

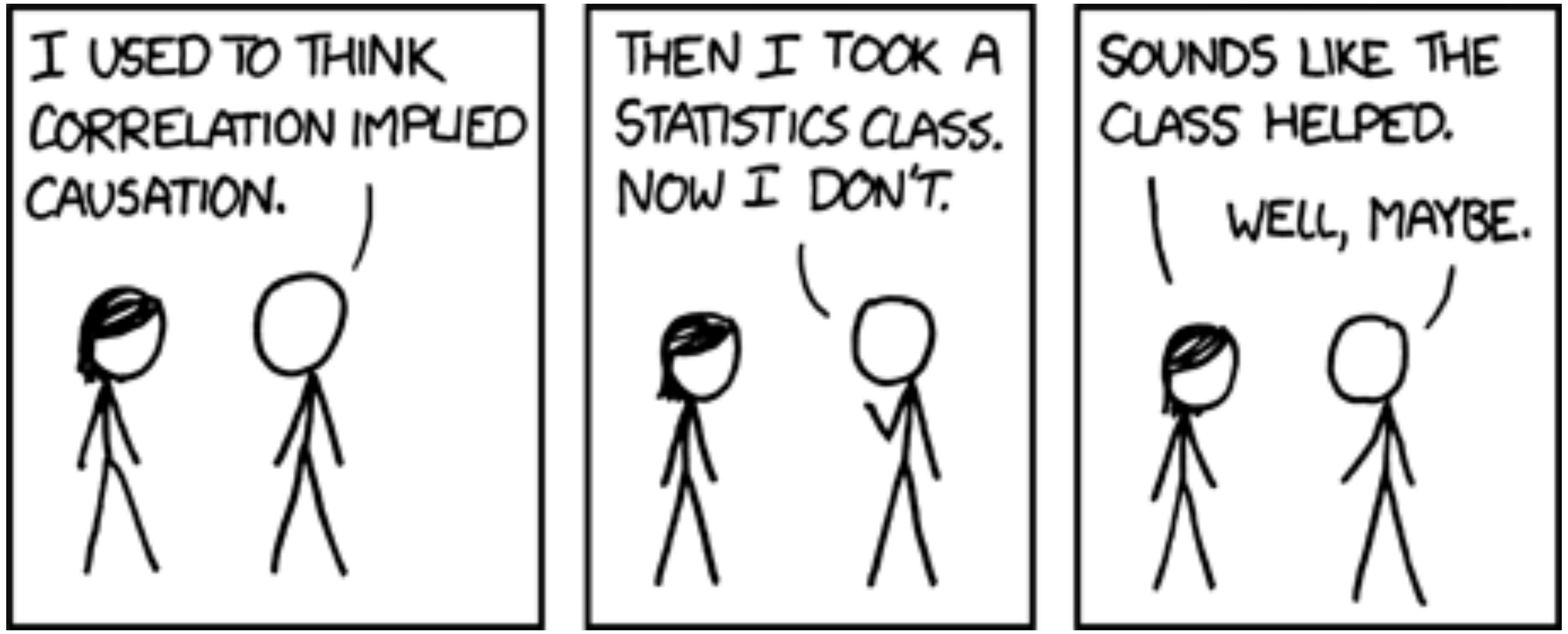
- >200 people
- Embedded in teams but organized centrally as a function within engineering
- 4 tracks: Analytics Engineering, Analytics, Inference, Algo
- 2 career paths: manager and individual contributor

Me

- Joined Airbnb ~5 years ago
- Inference track, growing as an individual contributor
- My first career
  - When I joined: I didn't know R, Python, SQL, GIS, what is a "job" … (you can learn anything & everything)
- Used to work on: Airbnb + Cities
- Now work on: Using methods to build data products
  - Today: "**No Experimentation, No Problem: Using Quasi-experimental Methods in Business Settings**"

CAUSAL INFERENCE

- Interested in the causal impact of **X** on **Y**
  - Impact of **compensation** on **employee retention**
  - Impact of **marketing spending** on **host acquisition**
- The gold standard: **Experimentation** or A/B testing



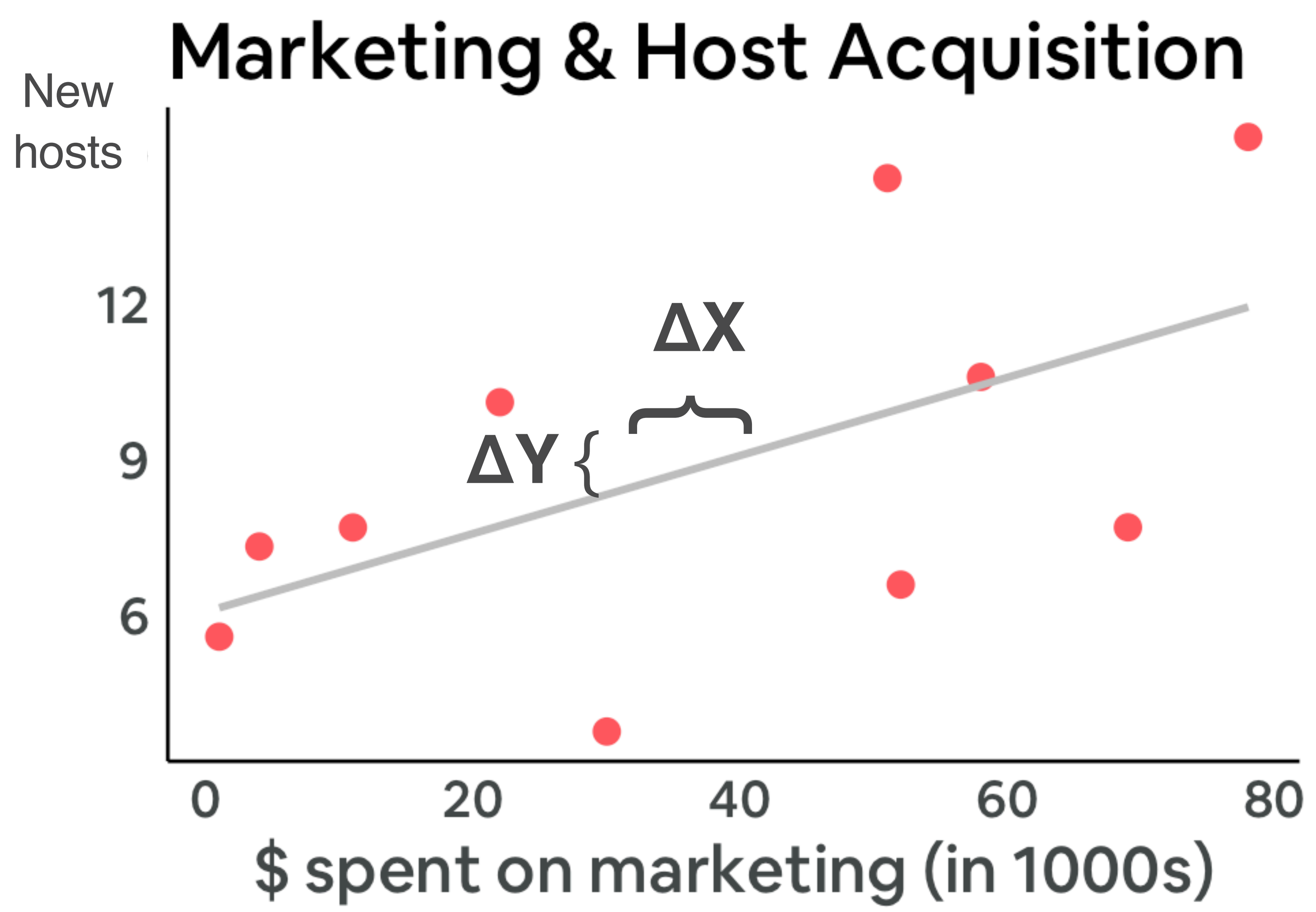
ORDINARY LEAST SQUARE (OLS) REGRESSION

Good place to start

$$Y_i = \alpha + \beta \times X_i + \epsilon_i$$

"Regress X on Y" to isolate how the variation in money spent on marketing

changes the number of new hosts in a market.



In R

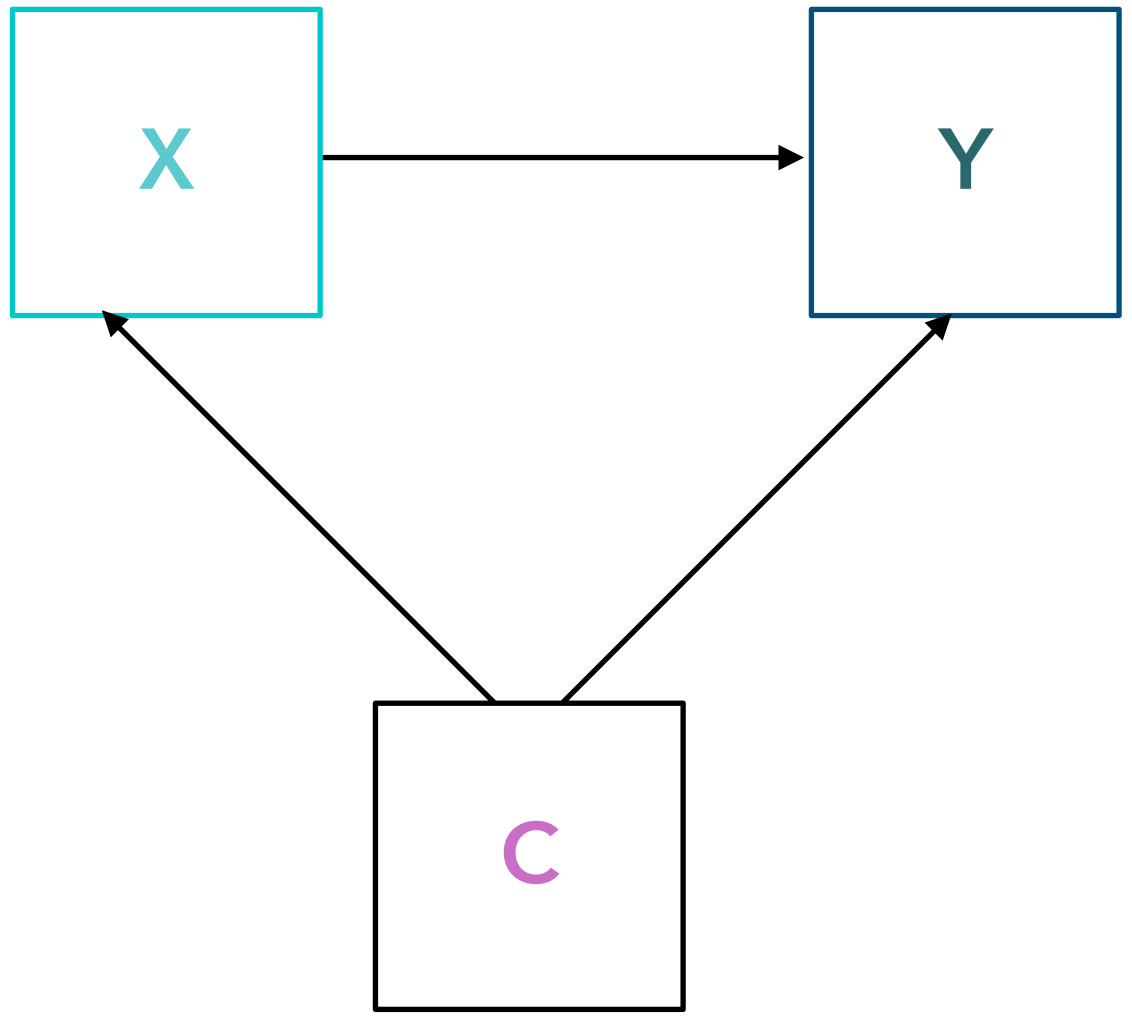
```
ols_model <- lm (Y ~ X, data = dataframe, ...)
summary(ols_model)
```

REGRESSION WITH CONTROLS

$$Y_i = \alpha + \beta \times X_i + \eta \times C_i + \epsilon_i$$

Omitted Variable Bias: the association between **X** and **Y** is driven by an omitted factor, **C**, that drives both.

- Places with more **money spent on marketing** are more **urban geographies** where our user base, including **new hosts**, is growing more quickly.
- If these "confounding" factors can be measured, control for them in a **controlled regression**.



In R

```
ols_w_controls_model <- lm (Y ~ X + C, data = dataframe, ...)
summary(ols_w_controls_model)
```

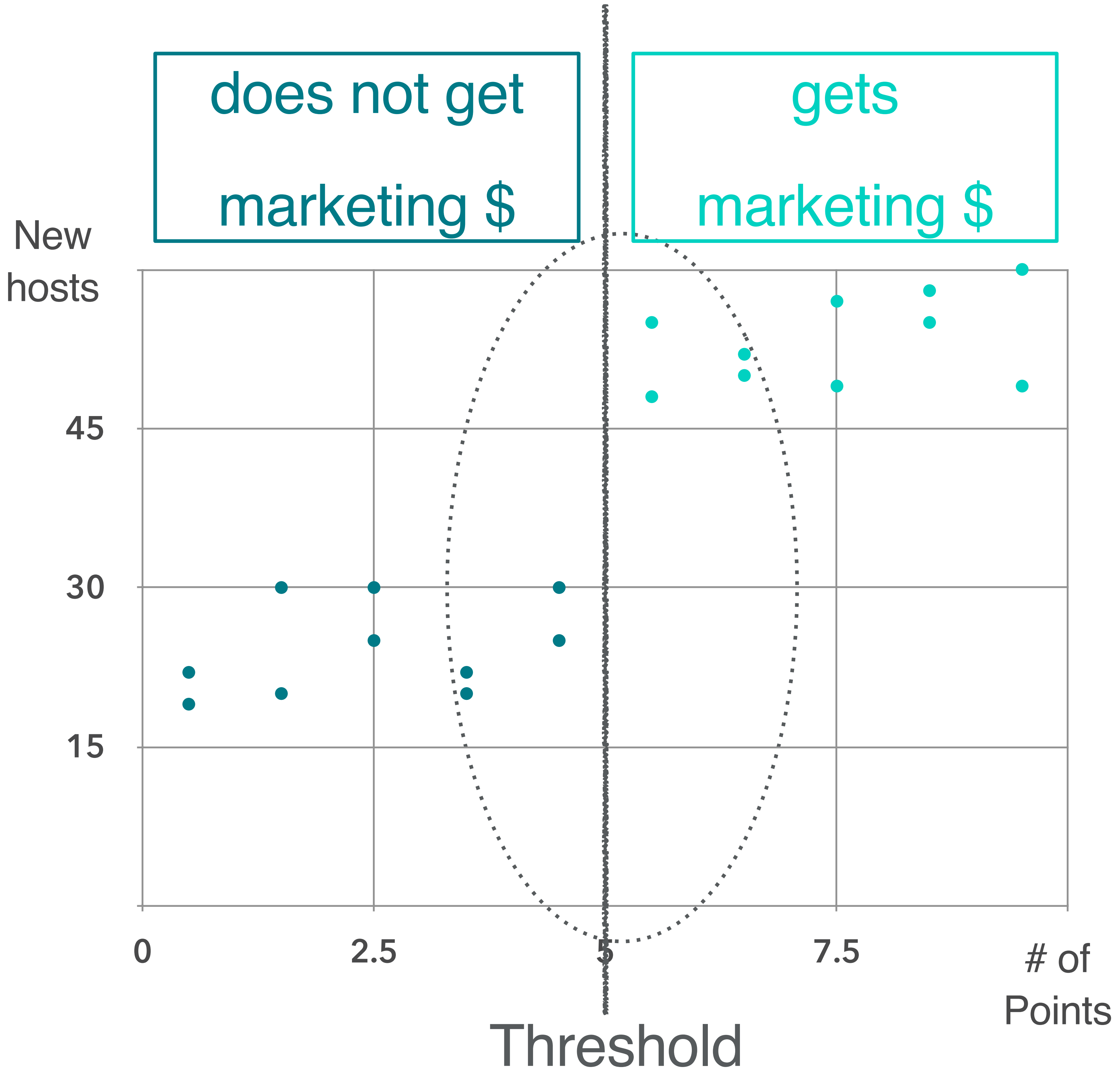
Quasi-experimental Methods

REGRESSION DISCONTINUITY DESIGN

Especially relevant when treatment is based on a "point" system

Imagine the business decides **marketing funding** based on a "formulaic point system:" current size of the market, future growth forecasts, # of upcoming events, etc.

- Threshold determines whether or not a place received marketing funding.
- Similar places fall on different sides of the threshold: some receive funding the others do not.



In

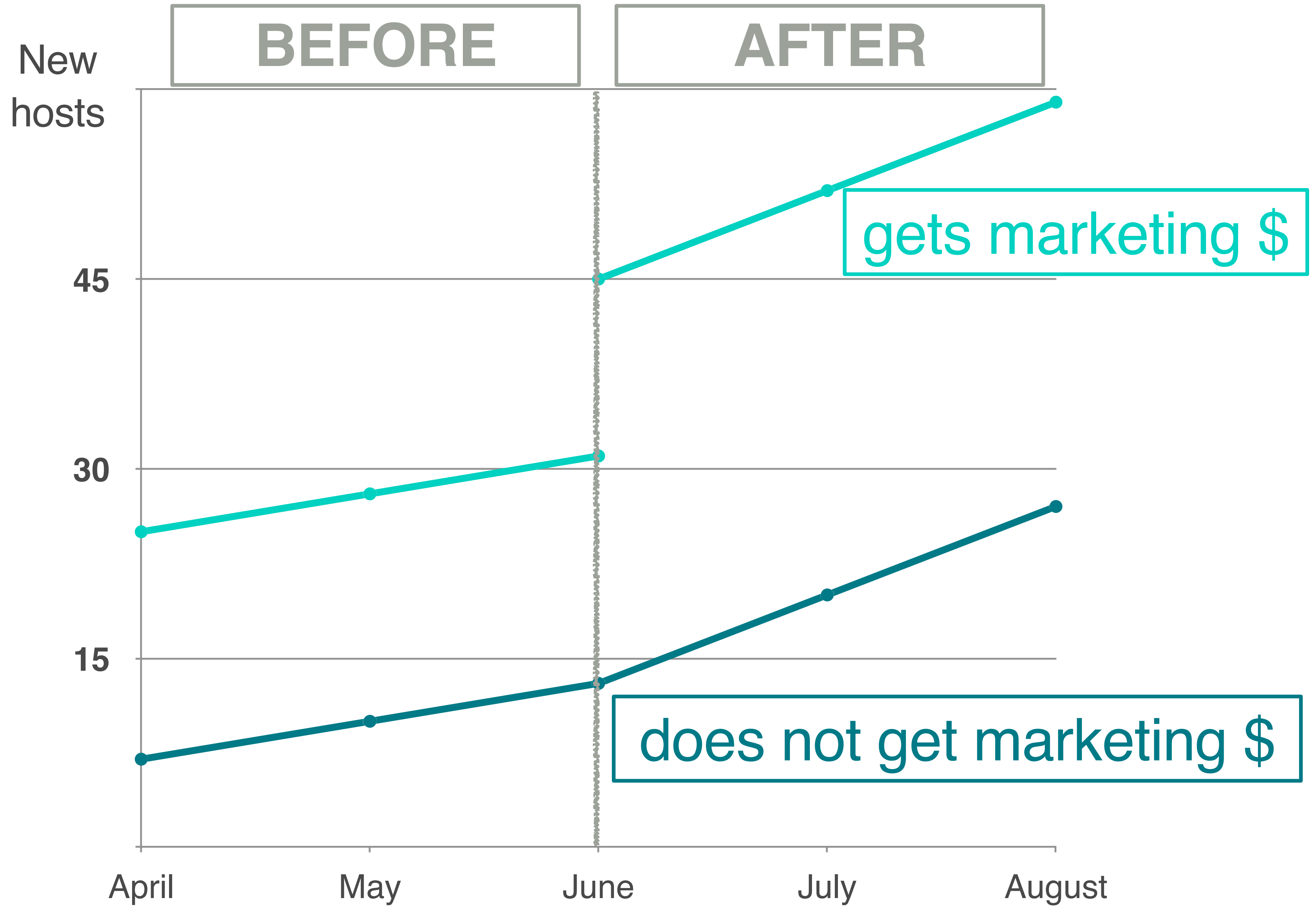
```
library(rdd)
rdd_model <- RDestimate(Y ~ X, data = dataframe,
                        cutpoint = 5, ...)
```

DIFFERENCES-IN-DIFFERENCES

Especially relevant for outcomes within a group and over time

Imagine the business decides to give **marketing funding** to a subset of markets.

- Compare outcome **Y** before and after the **marketing campaign** between control and treatment group.
- Check "parallel trends"



In R

```
DiD_model <- lm(Y ~ treatment + post + treatment*post,
                data = dataframe)
summary(DiD_model)
```

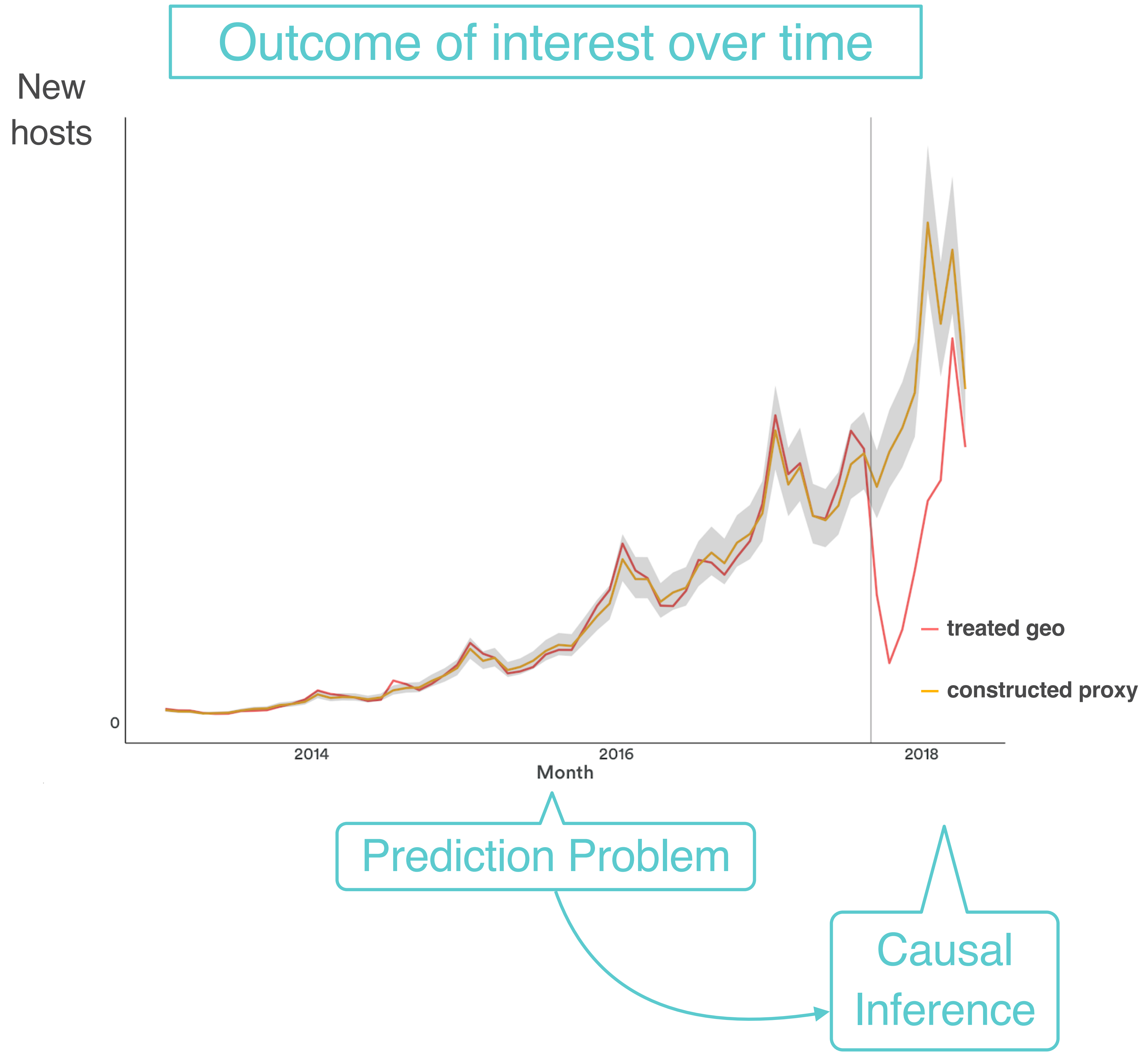
SYNTHETIC CONTROL

Causal inference + prediction

- So far, **causal inference**: Does X cause Y?
- What about **prediction**: Does X predict Y?

→ Use prediction to construct a refined "synthetic" control group. Then compare outcome for treated observation with outcome in synthetic contr

to identify causal impact of an external shock.



In R

```
library()
cvfit <- cv.glmnet(x, y, type.measure = "mse", nfolds = 20)
```

Sometimes, don’t even need causal methods

A FRAMEWORK

How to structure thinking about a problem

Prices have increased. Is that good or bad? Temporary or permanent?

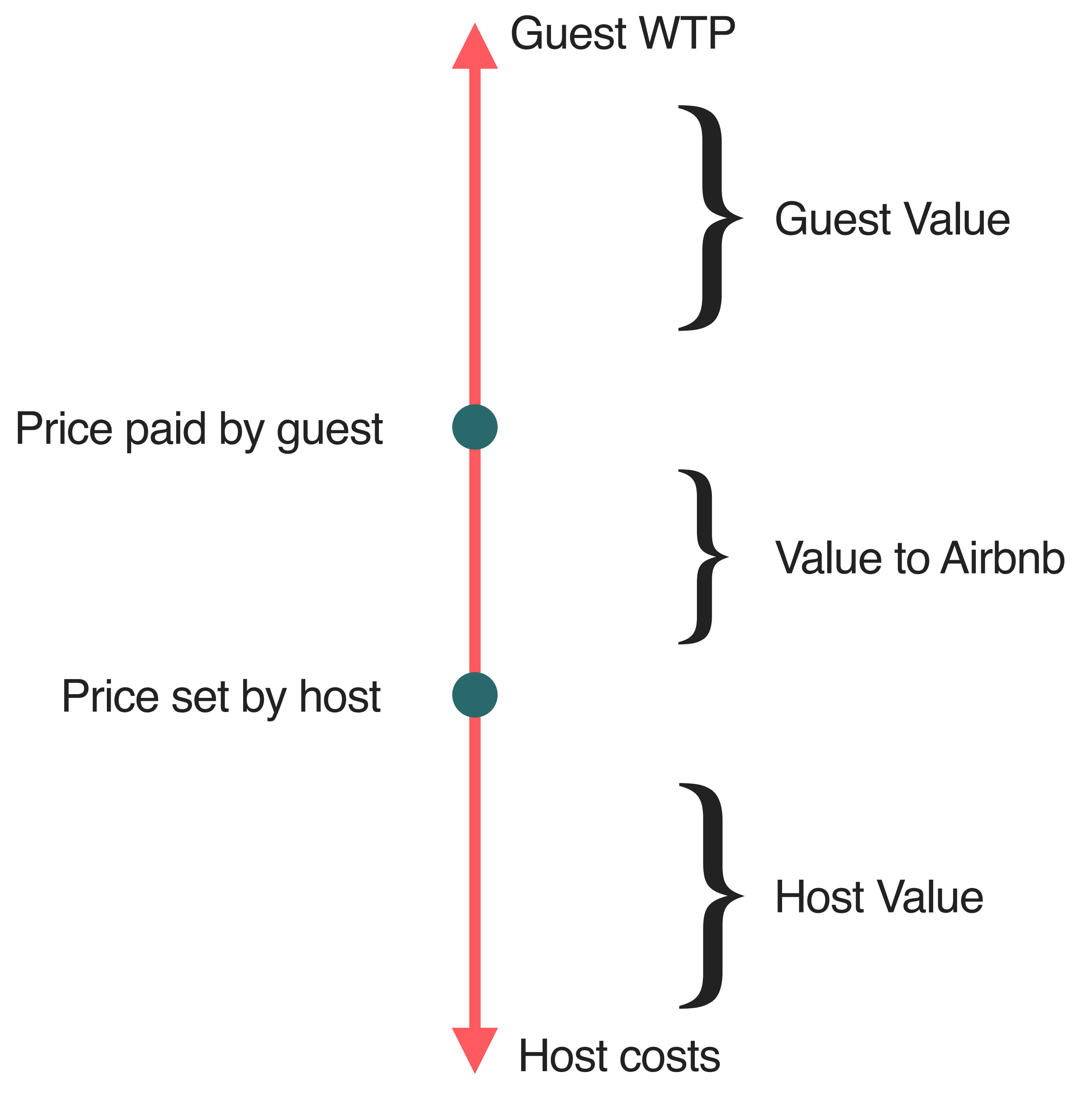
Average daily rates

ADR averaged $154 in Q4 2021, representing a 20% increase compared to the same prior year period, and a 36% increase compared to the same period in 2019. Q4 ADR remained elevated and outperformed our expectations that it would be stable relative to Q3. The sequential increase in ADR from the prior

If we think "value is bang for your buck:"

- Is value really down ~30%?
- What about hosts?
- What about willingness to pay?

→ "**<u>Value Sticks</u>**:" For each transaction, total value is difference between guest willingness to pay (WTP) and host costs.

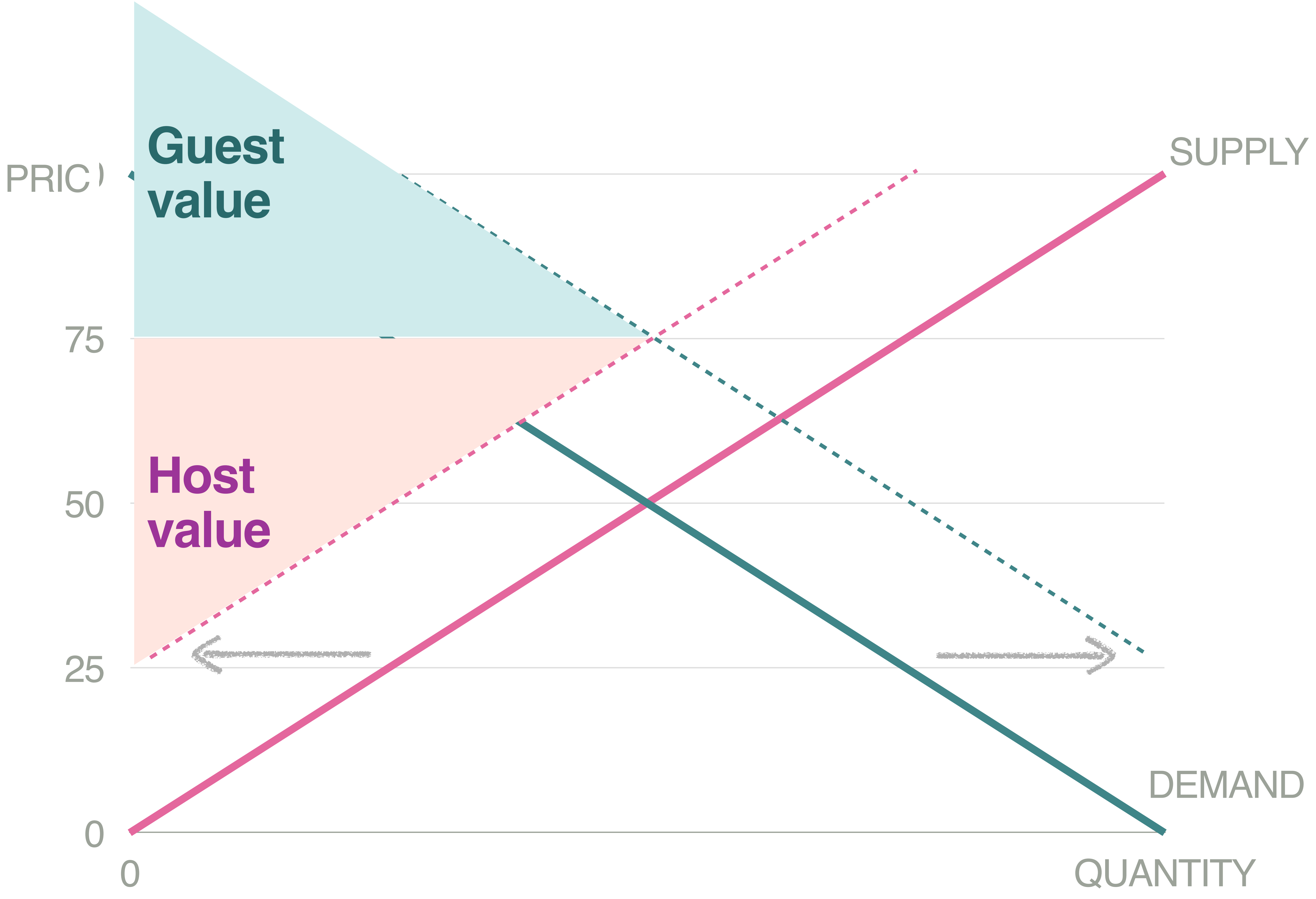


A FRAMEWORK

How to structure thinking about a problem

- Prices go up when demand curve shifts out (good for value) : more demand, increased WTP
- Prices go up when supply curve shifts in (bad for value): host churn, higher cost of hosting

→ We can test which factors are at play empirically.



Some Closing Thoughts

- **Machine Learning**: Regressions get you very far!
- **Feedback**: It's mostly a **gift**, but not always
- **Areas of growth**: Important, but not the only way to **progress** in your career
- Most of all: **Be yourself** !