GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
EXPERIMENTAL DESIGN

EXPERIMENTAL DESIGN 2

Experiments are designed around three central concepts:

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some dependent variables X and an independent variable y.

EXPERIMENTAL DESIGN

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some independent variables X and an dependent variable y.

2. Control: in order to determine above, we need some method of knowing the result when the independent variables are nil

EXPERIMENTAL DESIGN

Experiments are designed around three central concepts:

1. Causation: we want to be able to determine the causation or relationship between some independent variables X and an dependent variable y.

2. Control: in order to determine above, we need some method of knowing the result when the independent variables are nil

3. Variability: controlling variability in ease to detect reason behind change

DESIGN PATTERNS

Three most commonly used experimental design patterns

DESIGN PATTERNS

Three most commonly used experimental design patterns

Completely Random
+ effectively randomizes your sample completely, strives to eliminate bias
— how do you test randomness? How do you determine bias?

n samples = 1000

DESIGN PATTERNS

Three most commonly used experimental design patterns

Randomized Block
+ Similar to using a control, you now have a random sample blocked around a particular variable (or many)
+ limits variability!
— no longer truly random

n samples = 500 girls, 500 boys.
1000 all together

Three most commonly used experimental design patterns

Matched pairs
+ 1:1 relationship for each testing group = all variance eliminated
— literally impossible in the real world

n samples = all the same on each end
1000 samples

How does this fit in with AB testing?

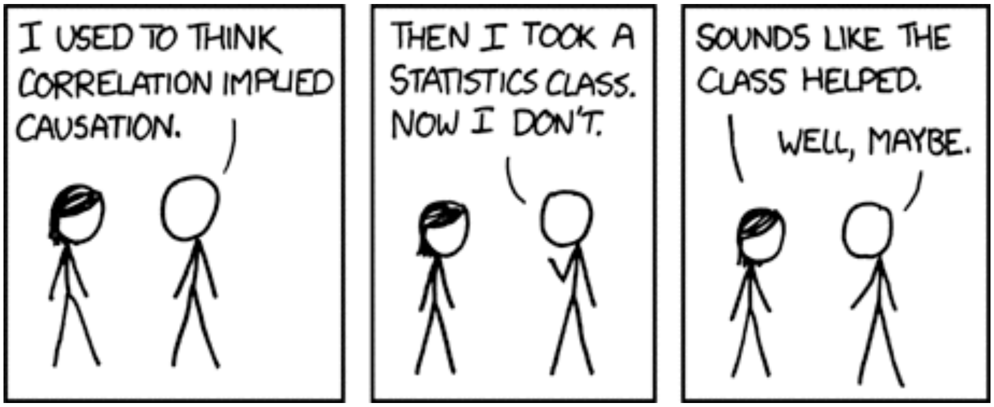
How does this fit in with AB testing?

Sample: Testing how price effects conversion rate of a photography-related product on a web page.

VALIDITY OF DATA

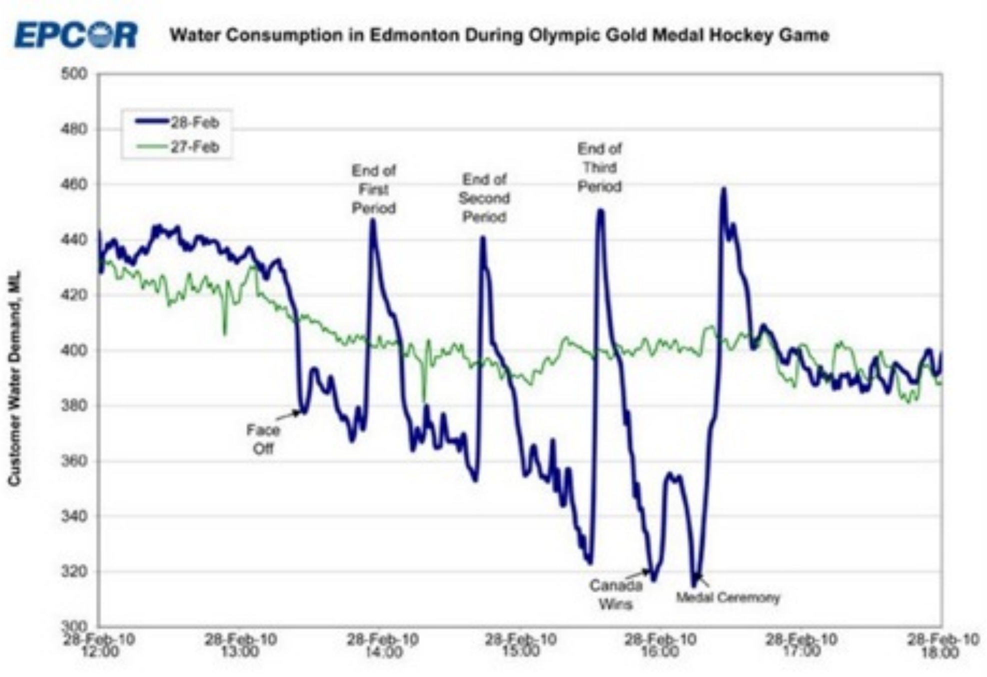
As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

VALIDITY OF DATA


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

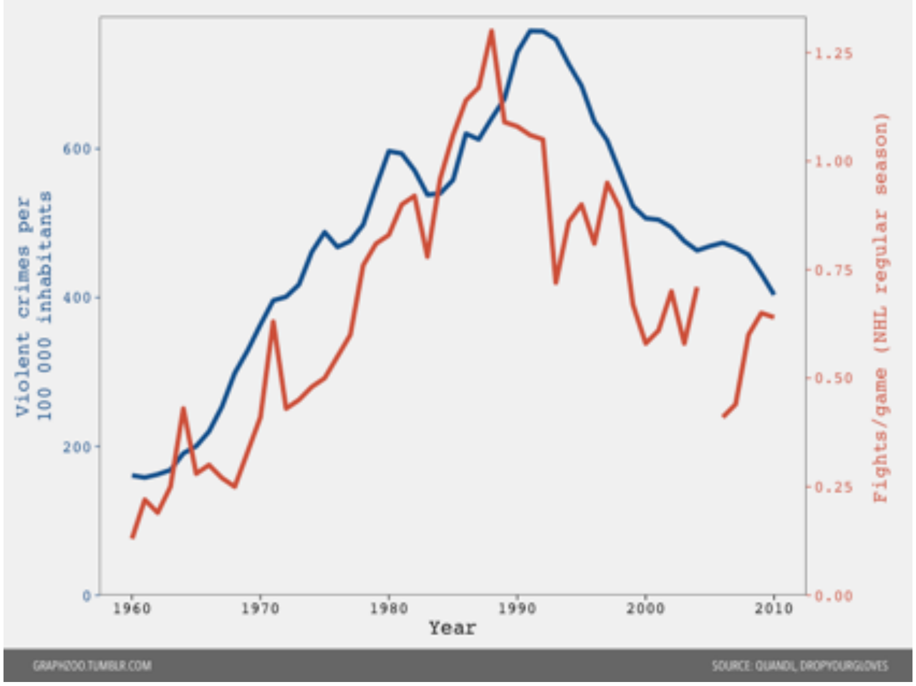
http://xkcd.com/552/

DOES THIS CORRELATION SUGGESTION CAUSATION?



http://i.imgur.com/npieo.jpg

HOW ABOUT THIS ONE?



http://graphzoo.tumblr.com/post/84722935157/data-source-http-www-quandl-com

VALIDITY OF DATA

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

Short answer: We don’t

VALIDITY OF DATA

As we start controlling for more features, how do you know each independent feature is causing a dependent feature?

Short answer: We don’t

Long answer: Start counting “correlation does not equal causation”’s before going to bed

ASSOCIATIONS

Smoking :: Cancer

Running the ball :: Winning games

Facebook fan :: Makes purchase

Shown ad :: Makes purchase

Toothbrushing :: Heart Disease

SAT scores :: Success in College

RELATIONSHIPS

Independence: Pr(Y | X) = Pr (Y)

Dependence: Pr(Y | X) != Pr (Y)

Dependence is useful. it’s how we build predictive models!

But is it causal? X —> Y OR y —> X

Dependency doesn’t say which direction the relationship runs, only that there is correlation

CORRELATION WITHOUT CAUSATION

Let’s talk about Z

Chain: X —> Z —> Y OR Y —> Z —> X

Fork: X <— Z —> Y

Collider: X —> Z <— Y

CHAINS

X doesn’t cause Y, it Causes the cause…Z

A McDonald’s opening doesn’t cause obesity, it causes overeating…which causes obesity

FORKS

Z causes both X and Y

High natural intelligence causes both good SAT scores and success in college

Colliders

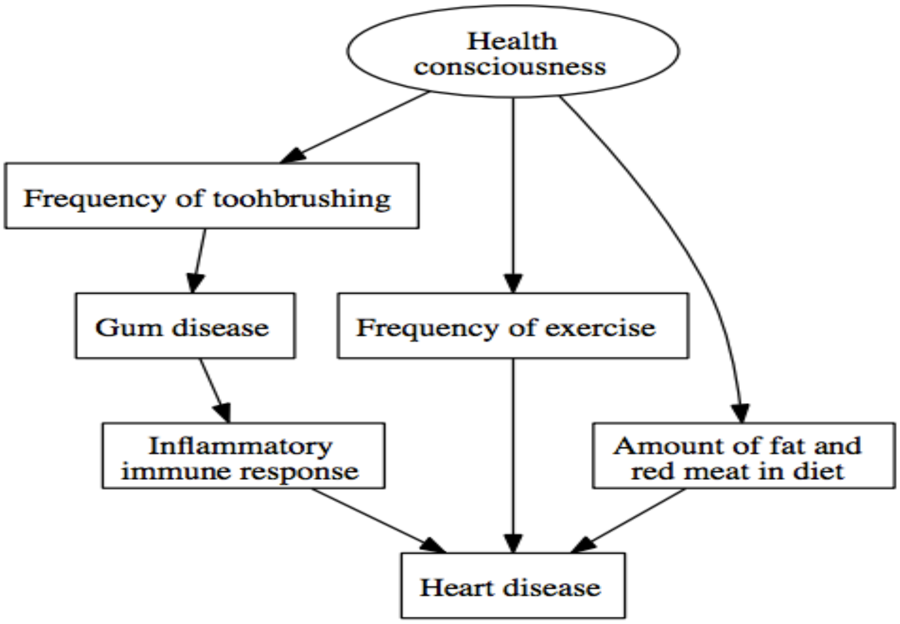
X & Y are independent, but not conditioned on Z

Your car won't start because you ran out of gas or the battery is dead (or 50 other things). These are independent events, but if you know the car isn't starting, they are negatively correlated.

ASSOCIATIONS

Smoking :: Cancer —> Genetics
Running the ball :: Winning games —> Having a lead
Facebook fan :: Makes purchase—> Brand loyalty
Shown ad :: Makes purchase —> Intent to buy
Toothbrushing :: Heart Disease —> Health consciousness
SAT scores :: Success in College —> Genetic intelligence

EXAMPLE LOGIC


Health consciousness
Frequency of toohbrushing
Gum disease
Frequency of exercise
Inflammatory immune response
Amount of fat and red meat in diet
Heart disease

WHY DO WE CARE?

Because Science!

1: Associations are always more interesting when they’re causal

2: Understanding a phenomena is different than predicting that phenomena

Because Decisions!

Smoking :: Cancer —> Quit smoking?

Running the ball :: Winning games —> Run the ball more?

Facebook fan :: Makes purchase —> Recruit FB fans?

Shown ad :: Makes purchase —> Purchase ads?

Toothbrushing :: Heart Disease —> Health consciousness?

SAT scores :: Success in College —> Take SAT class?

II. EVALUATION METHODOLOGIES

RANDOM ASSIGNMENT

Randomly assign people to groups, treatments, variants…

Eliminates impact of any confounding factors

Gold standard in clinical trials and policy experiments

MORE REALISTICALLY...THE QUASI-EXPERIMENT

Quasi-Experimental design tackles opt-in bias problem

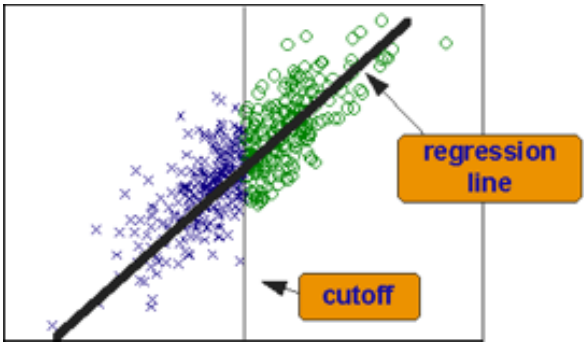
Matching

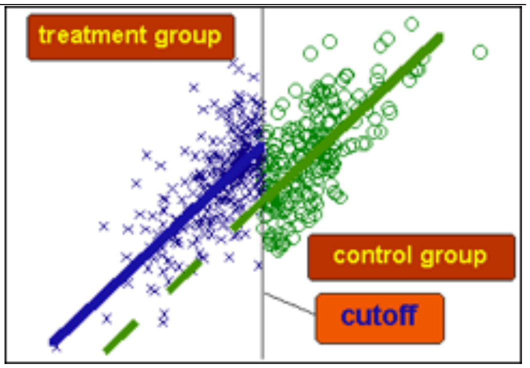
- For every user, find another adopter with as similar characteristics as possible on every measure

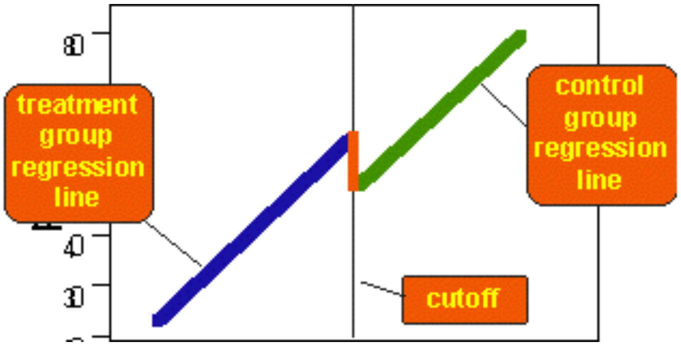
- Exclude unmatched users

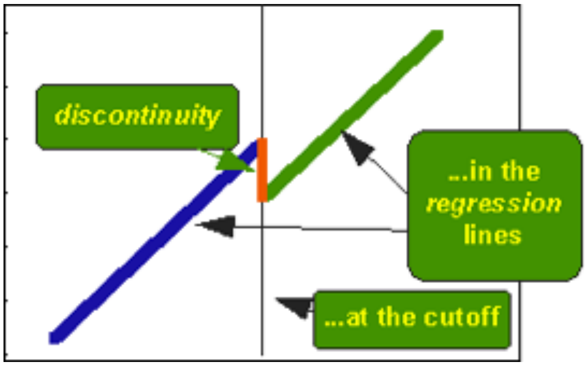
- Be sure to find characteristics which are good proxies for confounding variables

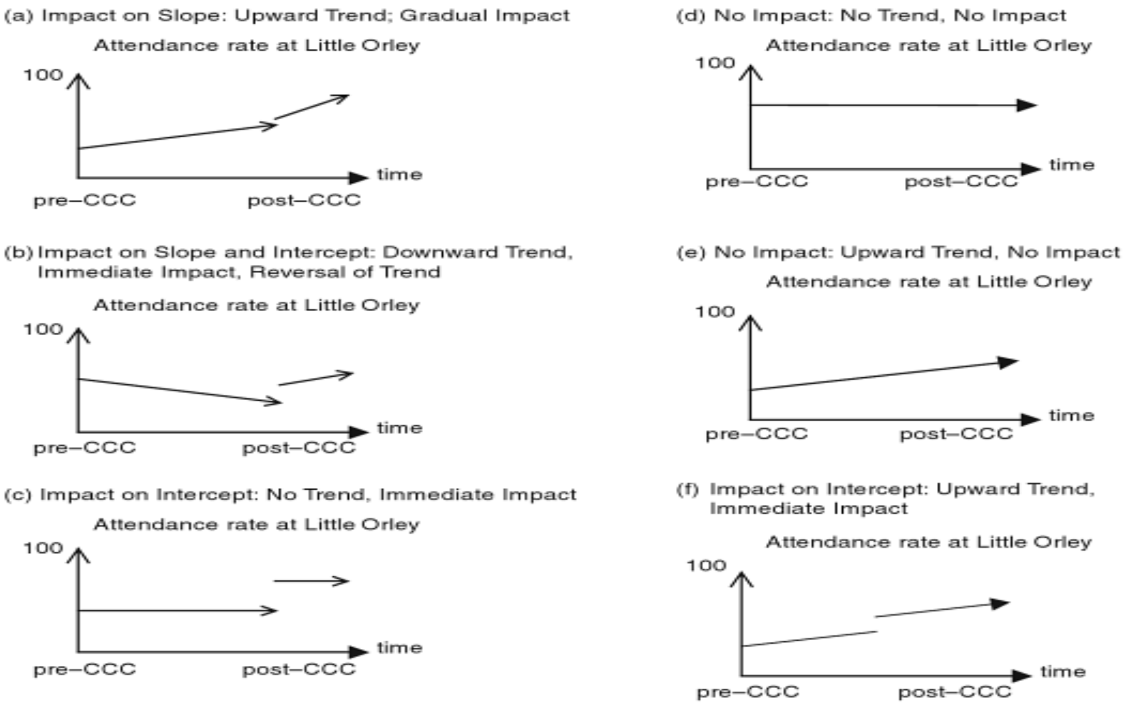
REGRESSION DISCONTINUITY






80
40
30
treatment group regression line
control group regression line
cutoff





(a) Impact on Slope: Upward Trend; Gradual Impact

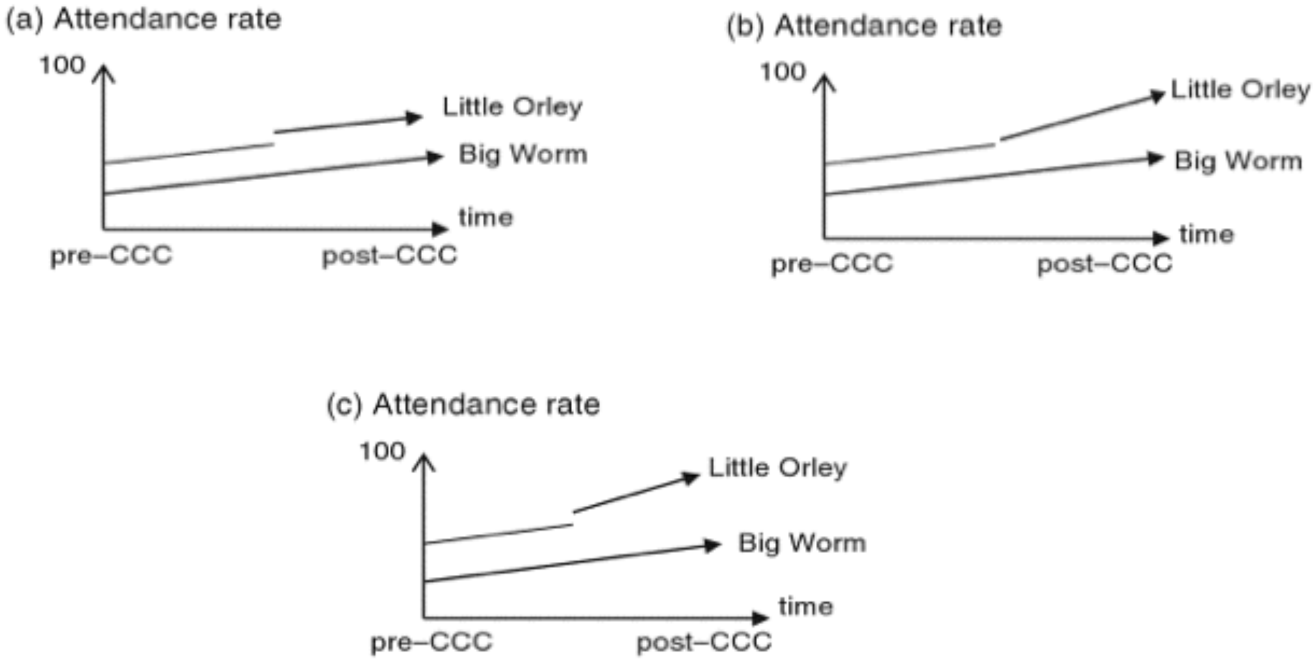
(b) Impact on Slope and Intercept: Downward Trend, Immediate Impact, Reversal of Trend

(c) Impact on Intercept: No Trend, Immediate Impact

(d) No Impact: No Trend, No Impact

(e) No Impact: Upward Trend, No Impact

(f) Impact on Intercept: Upward Trend, Immediate Impact


(a) Attendance rate
100
Little Orley
Big Worm
time
pre–CCC
post–CCC
(b) Attendance rate
100
Little Orley
Big Worm
time
pre–CCC
post–CCC
(c) Attendance rate
100
Little Orley
Big Worm
time
pre–CCC
post–CCC