GENERAL ASSEMBLY

DATA SCIENCE
10 WEEK PART TIME COURSE

Week 8 – Causality
Wednesday 30th November 2016

1. Project presentation by Sriram Rajagopalan
2. What is correlation
3. What is causality
4. What kinds of causal analysis are there
5. Experimental Design
6. Uplift
7. Uplift modeling
8. Lab
9. Review

DATA SCIENCE PART TIME COURSE

WHAT IS CORRELATION

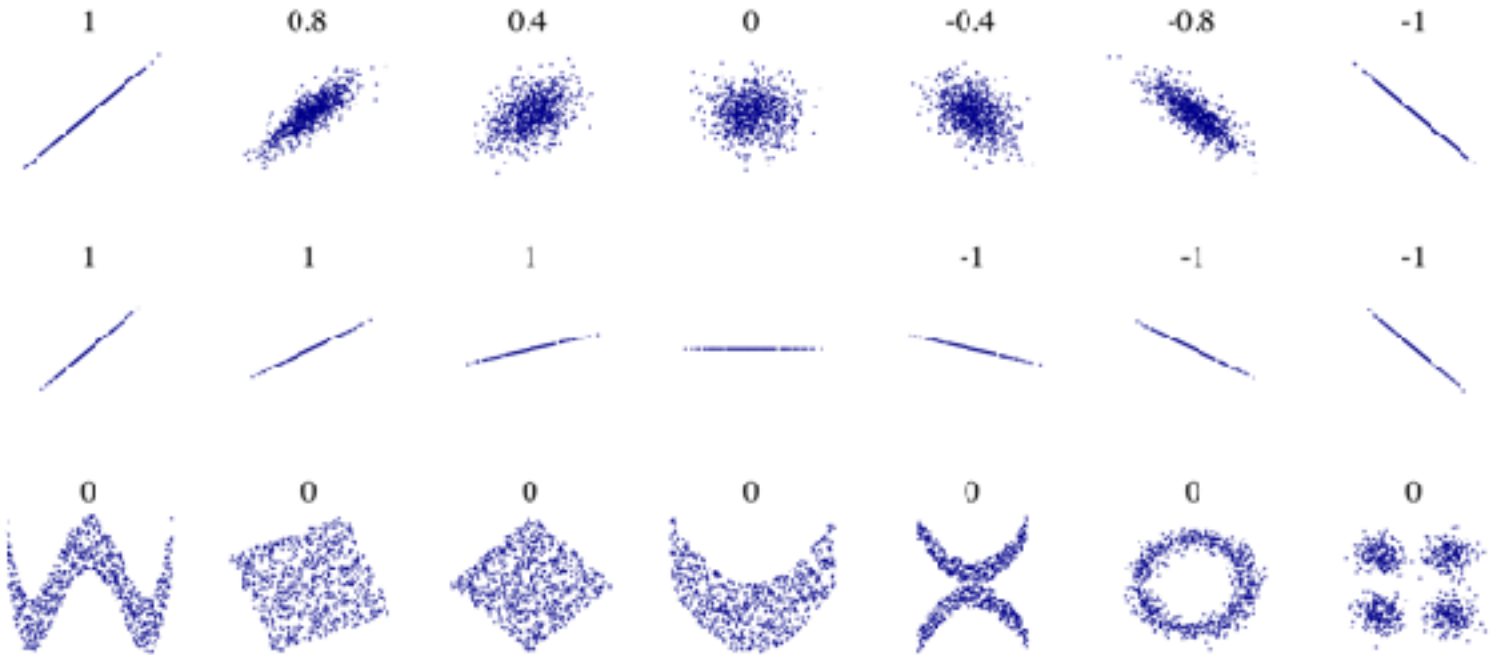
Are predictive models based on correlation or causation?

Are predictive models based on correlation or causation?

The traditional aim of machine learning methods is to infer meaningful features of an underlying probability distribution from samples drawn of that distribution. With the help of such features, one can infer associations of interest and predict or classify yet unobserved samples.

$$Correlation = \rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\boldsymbol{COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}}$$


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

DATA SCIENCE PART TIME COURSE

WHAT IS CAUSALITY

CORRELATION DOES NOT IMPLY CAUSATION

Correlation is a sign of a potential causal connection, and we can use it as a guide to further investigation (for example, trying to understand what the causal chain might be).

Does causality exist?

“

Many researchers believe that causality is only a convenient fiction. For example, there is no notion of causality in physical laws. Whether or not causality really exists is a deep philosophical question with no definitive answer in sight, but the practical points for machine learners are two. First, whether or not we call them “causal,” we would like to predict the effects of our actions, not just correlations between observable variables. Second, if you can obtain experimental data (for example by randomly assigning visitors to different versions of a Web site), then by all means do so.

“

Pedro Domingues

The following slides have examples of data insights with suggested explanations.

The left column's discoveries are real, validated by data, but the reasons behind them are unknown. Every explanation put forth, each entry in the rightmost column, is pure conjecture with absolutely no hard facts to back it up.

Insight	Organisation	Suggested Explanation
Online dating: Be cool and unreligious to succeed. Online dating messages that initiate first contact and include the word awesome are more than twice as likely to elicit a response as those with sexy. Messages with “your pretty” get fewer responses than those with “you’re pretty.” “Howdy” is better than “Hey.” “Band” does better than “literature” and “video games.” “Atheist” far surpasses most major religions, but “Zeus” is even better.	OkCupid (online dating website)	There is value in avoiding the overused or trite; video games are not a strong aphrodisiac.

Insight	Organisation	Suggested Explanation
Vegetarians miss fewer flights. Airline customers who preorder a vegetarian meal are more likely to make their flight.	An airline	The knowledge of a personalized or specific meal awaiting the customer provides an incentive or establishes a sense of commitment.

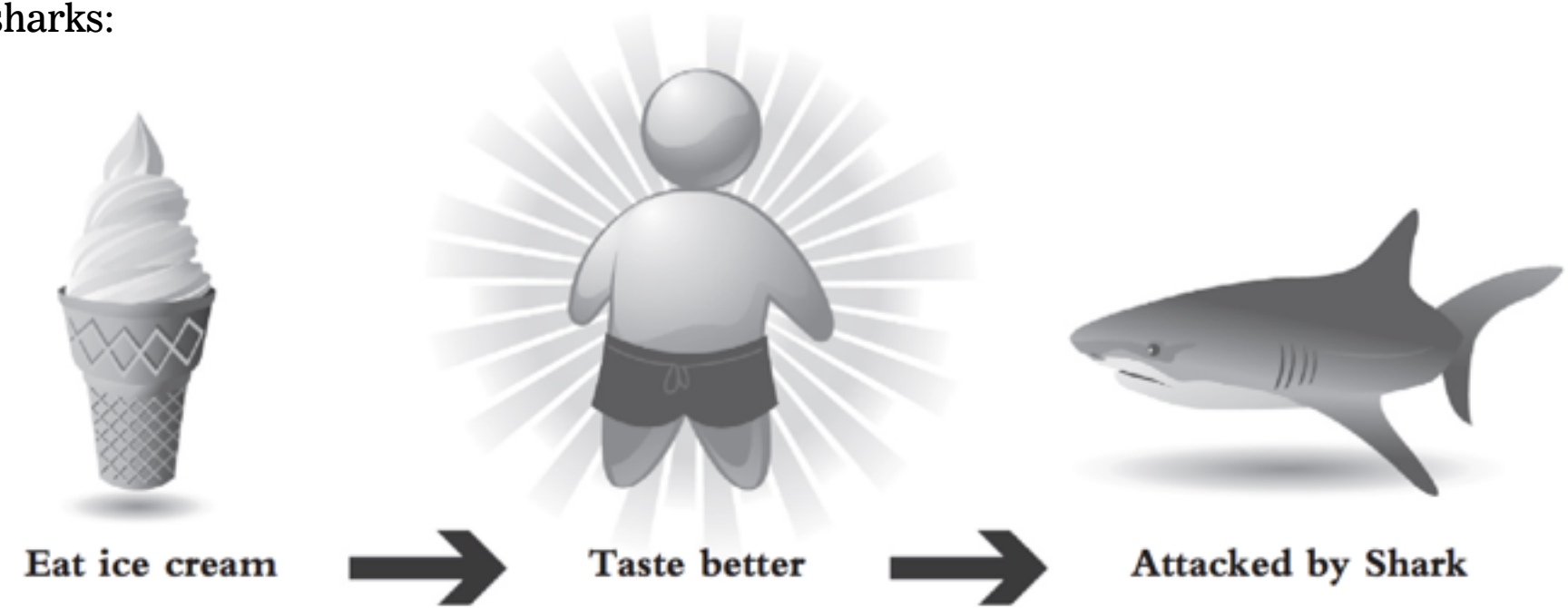
Insight	Organisation	Suggested Explanation
Solo rockers die younger than those in bands. Although all rock stars face higher risk, solo rock stars suffer twice the risk of early death as rock band members.	Public health offices in the UK	Band members benefit from peer support and solo artists exhibit even riskier behaviour.

Insight	Organisation	Suggested Explanation
Crime rises after elections. In India, crime is lower during an election year and rises soon after elections.	Researchers in India	Incumbent politicians crack down on crime more forcefully when running for reelection.

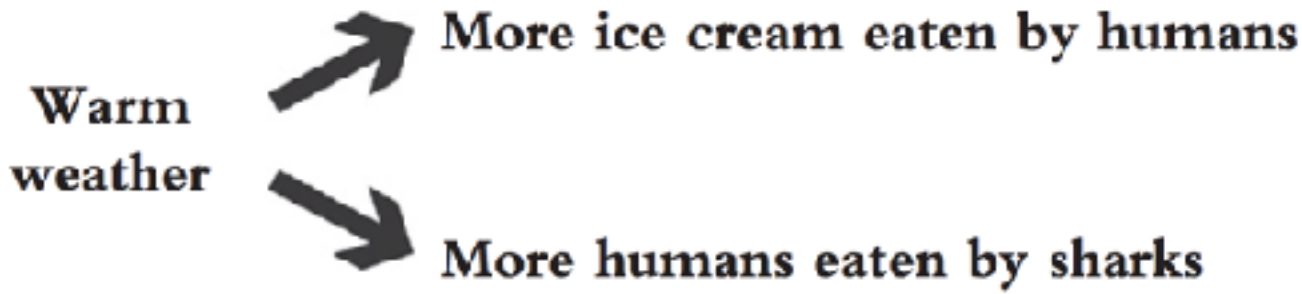
DATA SCIENCE PART TIME COURSE

WHAT IS CAUSAL ANALYSIS

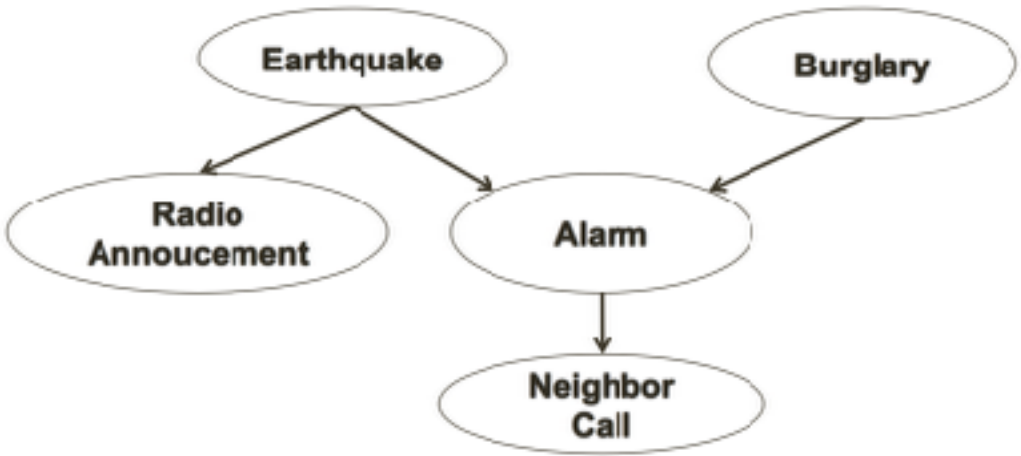
Increased ice cream sales correspond with increased shark attacks. Why do you think that is? A causal explanation could be that eating ice cream makes us taste better to sharks:



Another explanation is that, rather than one being caused by the other, they are both caused by the same thing. On cold days, people eat less ice cream and also swim less; on warm days, they do the opposite:



Bayesian probability models using directed graphs



Instrumental Variables:

Affect one variable in a directed graph, but does not directly affect further downstream viable of interest. All its effects must be due to any change in the variable it directly affects.

Simple example - randomly assign a customer to receive a piece of marketing

Instrumental Variables - Natural Experiments:

Instrumental Variables - Natural Experiments:

We want to find a variable that:

1. Explains changes in the supply of coffee
2. Does not explain changes in the price of coffee other than through its effect on supply (ie. No effect on demand)

Eg. Plague of locusts in Ethiopia, rainfall in Brazil. These would constitute natural experiments—randomly assign treatment (lower/higher prices) in some periods but not others.

DATA SCIENCE PART TIME COURSE

EXPERIMENTAL DESIGN

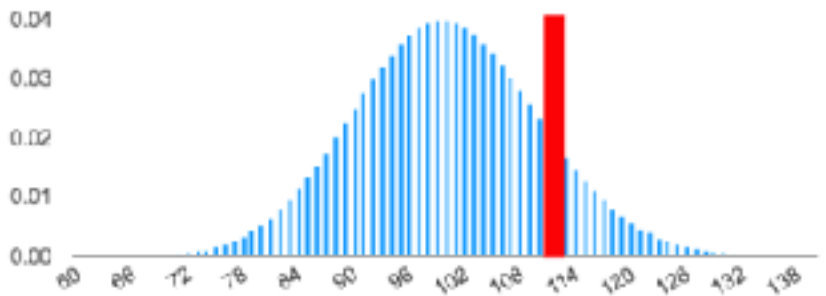
Randomised Clinical Trials

We randomly assign some group of people to receive a "treatment" and others to be in the "control" group —that is, they don't receive the treatment.

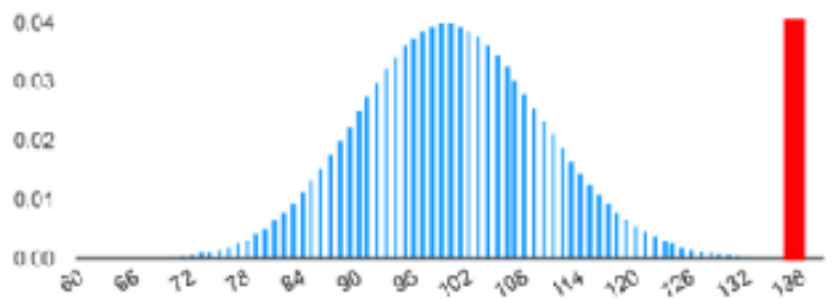
A/B Testing


50 % visitors
see variation A
Variation A
23%
conversion
50 % visitors
see variation B
Variation B
11%
conversion

Check that the result is outside some typical range based on an assumed or derived distribution.



Result is within typical values



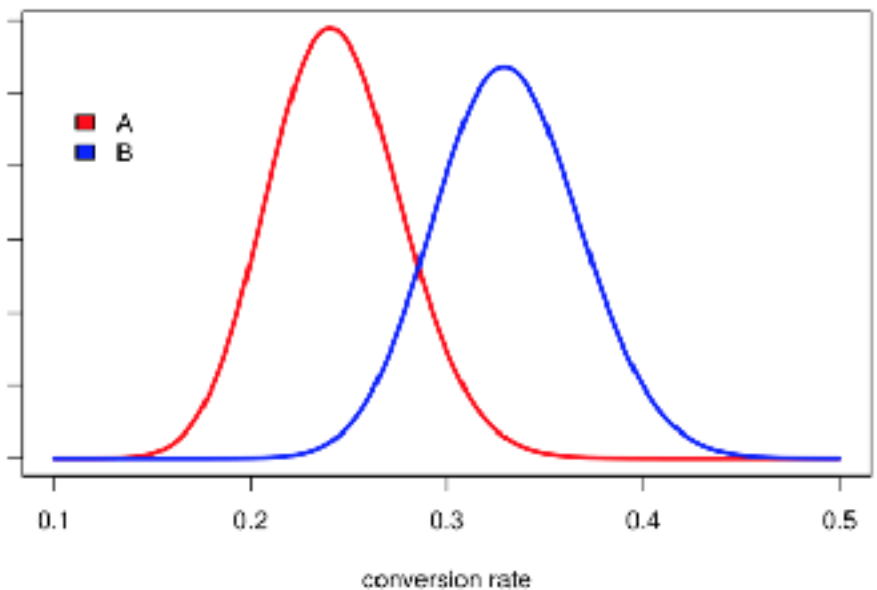
Result is outside typical values

We can run many experiments at the same time

DATA SCIENCE PART TIME COURSE

INCREMENTAL IMPACT OR UPLIFT

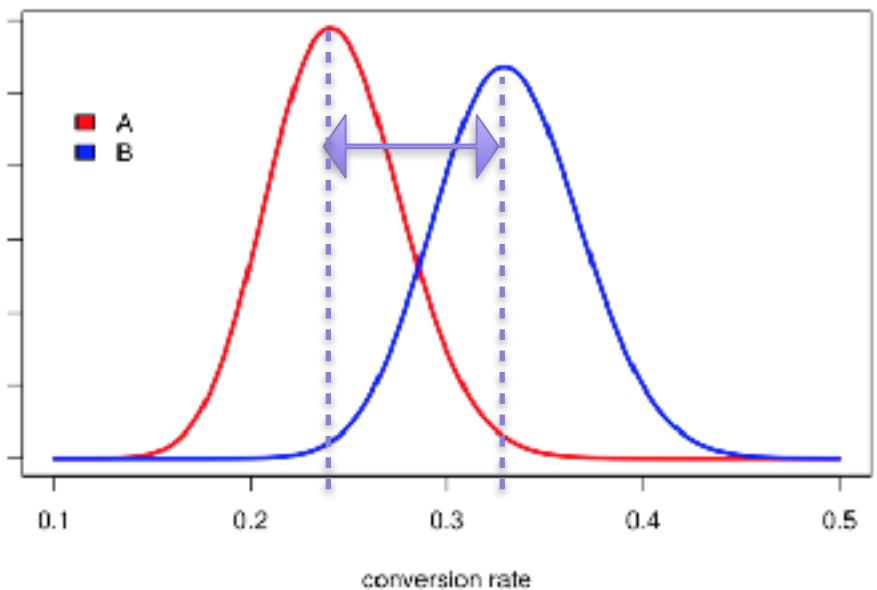
The difference in response rate between a treated group and a randomised control group



A – Treatment

B – Control

The difference in response rate between a treated group and a randomised control group



A – Treatment

B – Control

A treatment may be from:

- A science or medical experiment
- A marketing campaign
- A natural experiment

A control is a hold out group that is statistically equivalent to the treated group; allowing for statistical inference regarding differences.

A control is a hold out group that is statistically equivalent to the treated group; allowing for statistical inference regarding differences.

A control allows for causal inference:

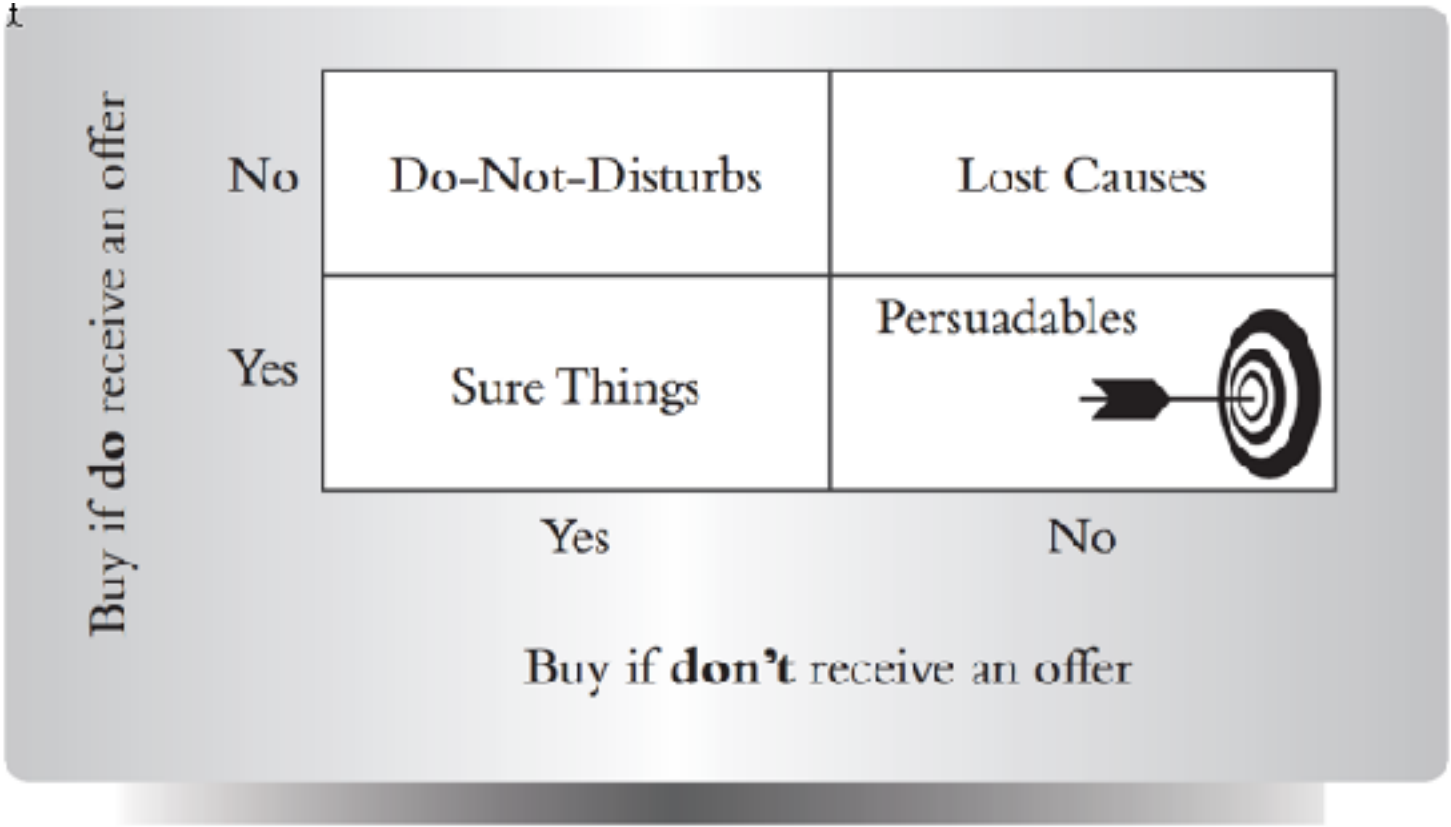
If the only difference is the treatment, then we can make inferences relating the treatment to observed effects.

Calculate uplift rate as:

conversion rate (treatment) - conversion rate (control)

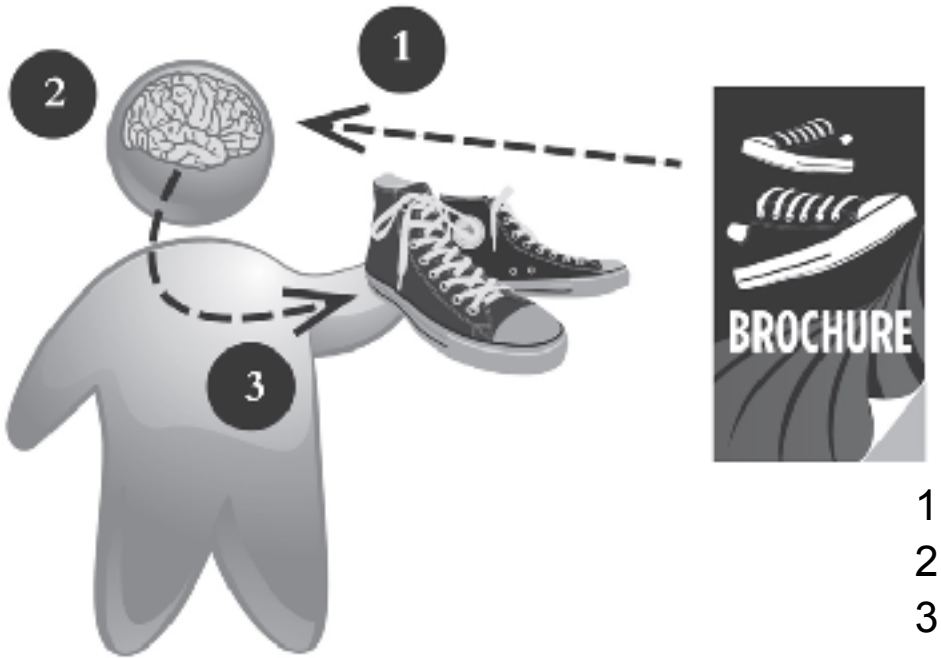
Incremental impact = uplift rate (%) x conversions


Mike Kelly Mp
Gets the job done
VOTE
STABLE POPULATION PARTY
1
GREENS
HENDY
TTER,
OT
GER


Buy if do receive an offer
No
Yes
Do-Not-Disturbs
Lost Causes
Sure Things
Persuadables
Yes
No
Buy if don't receive an offer

Each action risks backfiring:

- The customer cancels,
- The patient suffers an adverse reaction,
- The beneficiary becomes dependent on assistance.



The brain is a black box. Influence cannot be observed.

We can never witness an individual case of persuasion with certainty.

1. The individual perceives the sales brochure.
2. Something happens inside the brain.
3. The individual buys the product.

How do we know the brochure made a difference?
Perhaps the individual would have purchased anyway.

DATA SCIENCE PART TIME COURSE

UPLIFT MODELING

What is uplift modelling?

Supervised learning without real labels

(problematic)

So we approximate over groups

What is uplift modelling?

Supervised learning without real labels

(problematic)

So we approximate over groups

Uplift Modelling combines previously separate paradigms:

1. Comparing treated and control results.
2. Predictive modelling

Probability of response given treatment

$$P(O = 1 \mid x;T)$$

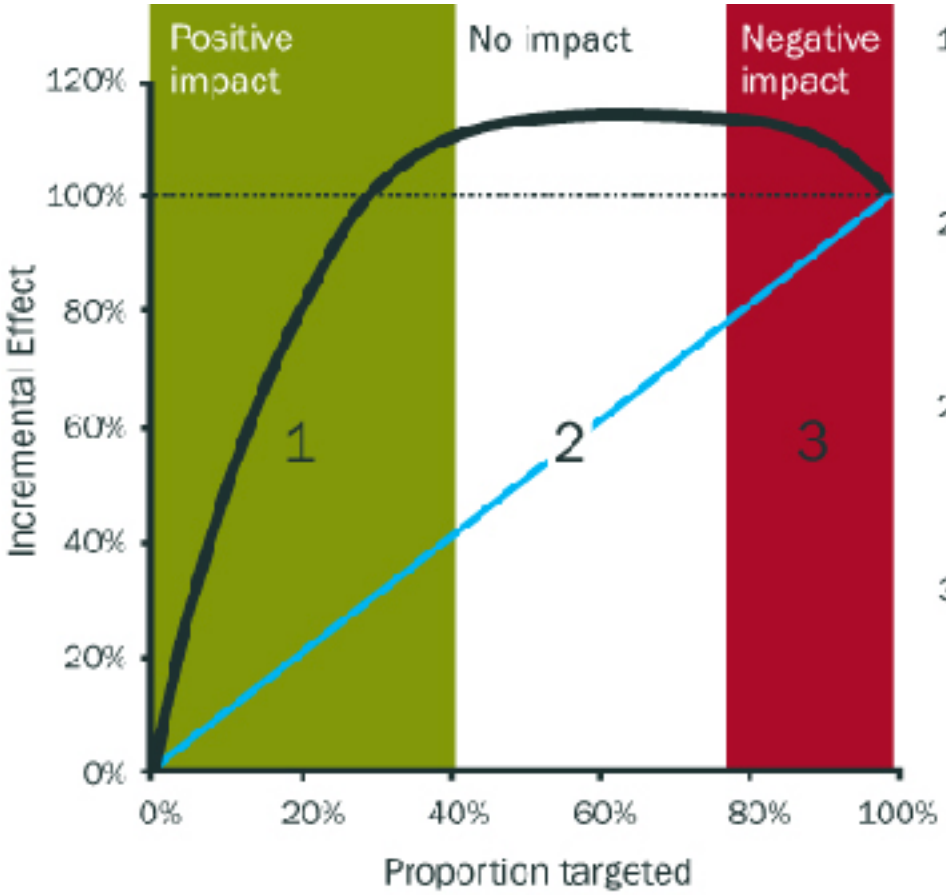
-

Probability of response given control

$$P(O = 1 \mid x;C)$$

Types of uplift models

1. Subtract two response models (one trained on treatment, one trained on control)
2. Leave the treatment variable as a feature. Predict results for treatment = 1, subtract results for treatment = 0
3. Matched pairs. Assign equivalent subjects to treatment and control. Then study.
4. Specialist uplift models
5. R - Uplift package
6. SAS - Net Incremental Lift model

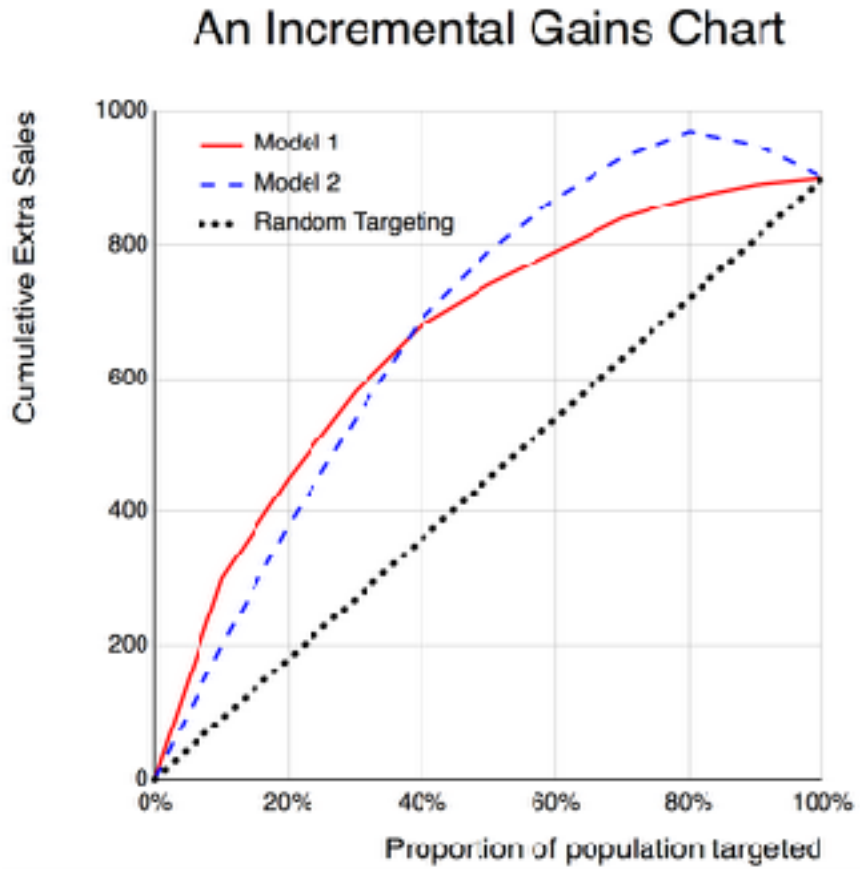


1 Persuadables:
The marketing works.
These people buy only
when treated

2a Sure Things:
The marketing has no impact:
They buy whether or not
they are treated

2b Lost Causes:
Again, the marketing has no
effect, but these people do
not buy either way

3 Do Not Disturbs:
The marketing backfires.
These customers don't buy
when treated but do when
left undisturbed


An Incremental Gains Chart
Cumulative Extra Sales
1000
800
600
400
200
0
Model 1
Model 2
Random Targeting
0%
20%
40%
60%
80%
100%
Proportion of population targeted

Uplift model complications:

- Uplift is a second-order effect, with large errors typical for the estimates.
- The single biggest challenge with uplift modelling tends to be producing stable models
- Overall uplift is often small compared to the background effect

LAB

Hillstrom email campaign dataset

Segment	Treated Records	Control Records
Mens E-Mail	21,307	0
Womens E-Mail	21,387	0
No E-Mail	0	21,306

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m "descriptive label for the commit"
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

HOMEWORK

Homework

- Homework 3 - Due Monday 28th of November (most questions are based around your project)

Reading

- blog post: https://yanirseroussi.com/2016/02/14/why-you-should-stop-worrying-about-deep-learning-and-deepen-your-understanding-of-causality-instead/'
- Recommended text: Causal Inference in Statistics_ A Primer - Judea Pearl, Madelyn Glymour, Nicholas P. Jewell-Wiley (2016)