Ethics of Big Data

Peter Danielson
Univ. of British Columbia
COGS 300.002
Mar 26 2015
• Scientific accuracy (vs. hype): Google Flu Trends
• Manipulation & Consent
  – in social science (Facebook)
  – in commerce & politics
• De-identification possible?
According to "The Parable of Google Flu", which of the following was NOT considered a factor of Google Flu Trends (GFT)'s error?

a) GFT's overestimation of flu prevalence
b) Google's use of a static search algorithm
c) GFT's overlooking considerable information by throwing out particular search terms
d) two of the above
e) none of the above

Devin
Rate Quiz Question 1

A. Excellent  
B. Very Good  
C. Good  
D. Acceptable  
E. Poor
Google Flu Trends


• Cheap and very quick!
“Flu” ~ Flu

![Graph showing comparison between Google Flu, Google Flu + CDC, and CDC data. The graph illustrates the percentage of ILI (Illness Likelihood Index) and the error percentage baseline over time. The graph highlights Google estimates more than double CDC estimates in some periods. Google starts estimating high 100 out of 108 weeks.]

Google estimates more than double CDC estimates in some periods. Google starts estimating high 100 out of 108 weeks.
What was the result of removing both positive and negative emotional posts from an individual’s news feed in the study by Kramer et al.

A) Engaging effect – individuals became more active and overall more engaging to compensate for the deficit in emotional posts.
B) Withdrawal Effect – individuals became less expressive over the following days.
C) Positive effect – individuals posted more positive content
D) Negative effect – individuals posted more negative content
Rate Quiz Question 2

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Q3

Which of the following methods increase people's chances of voting according to Duhigg's article?

a) trusting someone you have met before with their suggestion
b) thinking of the logistics you have to prepare before taking action
c) forming a mental image of voting
d) increased exposure of a candidate to the voter
e) all of the above

Ee Faye
Q3 Discussion

a) trusting someone you have met before with their suggestion
b) thinking of the logistics you have to prepare before taking action
c) forming a mental image of voting
d) increased exposure of a candidate to the voter
e) all of the above

If you were exposed to such strategies, do you think that you would realize that you have been primed to vote?

Ee Faye
Rate Quiz Question 3

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
According to political operatives, why would a campaign official ask citizens, before an election, whether they would walk or drive to poll stations?

a) to know whether having a man handing out flyers near the poll station would benefit the campaign.

b) to know how many citizens own a car, so they can change their automobile policies.

c) to get them to think about voting, thereby increasing the chances of actually going to vote.

d) to understand the transportation behaviours of voters, so that locations of poll stations could be changed for future elections.

e) None of the above
Rate Quiz Question 4

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Q5

What is the best way to form a habit like exercising every morning?

a) Choosing a simple cue like laying your workout clothes out the night before
b) A clear reward like the sense of accomplishment from logging it randomly on days
c) A clear reward like a midday treat
d) A and B
e) A and C

Ayeesha
Rate Quiz Question 5

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Q6

According to "How Companies Learn Your Secrets", which of the following is a function of forming habits that matches this description?
"Converting a sequence of actions into an automatic routine"

A) Adaptation
B) Bulking
C) Chunking
D) Habituation
E) None of the above

Sean
Rate Quiz Question 6

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Discussion Question

Do you think we lose a degree of free will when companies target our habit-forming mechanisms to ensure brand loyalty? Is it morally wrong to do this? (Possible lead-in: Think back to Kahneman's explanation of intuition).

Does the greater predictive ability achieved by proper use of big data techniques cause marketing to become "too effective"? Is it possible to delineate a boundary, or can we set up a slippery slope scenario where such techniques would become morally objectionable instead of just being questionable?

Stephen
Rate Discussion Question 1

A. Excellent
B. Very Good
C. Good
D. Acceptable
E. Poor
Ethics Methods Foiled by Big Data

1. “Notice and consent”
   - Dilemma of New unforeseen uses for data
     - Flu from search
   - Not feasible to re-consult
   - Or blanket permission for all uses

2. Opting out:
   - German Street View blur
   - Target for egging!
Get over it:
A New Baseline for Consent?

• ‘We noticed recently that people didn’t like it when Facebook “experimented” with their news feed. Even the FTC is getting involved. But guess what, everybody: if you use the Internet, you’re the subject of hundreds of experiments at any given time, on every site. That’s how websites work.’

• Christian Rudder
  http://blog.okcupid.com/index.php/page/2/
Data, A/B Experiment
Deception & a Justification!

- OKCupid
- “we took pairs of bad matches (actual 30% match) and told them they were exceptionally good for each other (displaying a 90% match.)”
Big Data is Private Data

• Research Methodology
  – Lack of transparency at Google Flue
  – Insider access at Yahoo, Facebook, OKCupid
  – How check or replicate private data?

• To make datasets public need to de-identify
3. “Anonymization vs. Reidentification

- AOL – content of searches alone
  - “60 single men”, “landscapers in Lilburn Ga”
- Netflix contest (and other data)
- “This ... new subspecialty of computer science, reidentification science ... unearths a tension that shakes a foundational belief about data privacy: Data can be either useful or perfectly anonymous but never both.” (Ohm, 2004, p.1703f)
Vs. Yakowitz: “Tragedy of the Data Commons”

– Benefit of Public Data Sets
– Risks: Theoretical?
– Harm: “The risk of privacy harm from re-identification is actually significantly lower than many of the everyday risks we take for granted, such as those attendant on throwing out our trash.”
Big Data Research Dilemma

• “We’re living through a golden age of behavioral research. It’s amazing how much we can figure out about how people think now.” (quoted in Duhigg)
  – Watts to Rudder

• But also undercutting ethical basis of behavioral research?
References


• Rudder, C. (2014) Dataclysm: Who We Are (when we think no one’s looking). Crown,