DataEngConf 2016

Causal inference in data science: From Prediction to Causation

Amit Sharma
Postdoctoral Researcher, Microsoft Research
amshar@microsoft.com
@amt_shrma

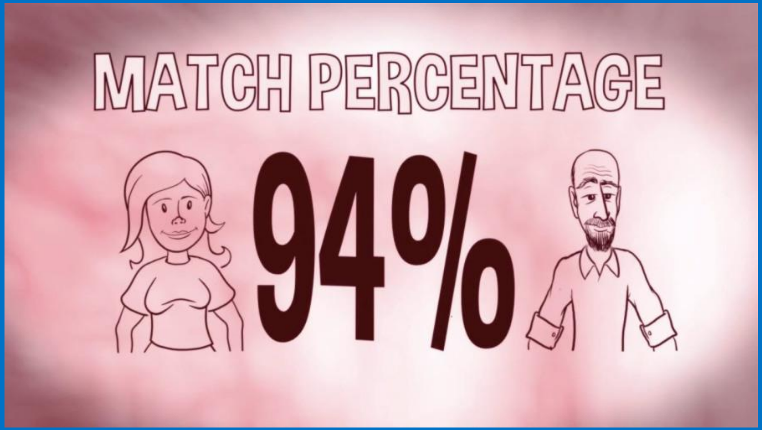
http://www.amitsharma.in

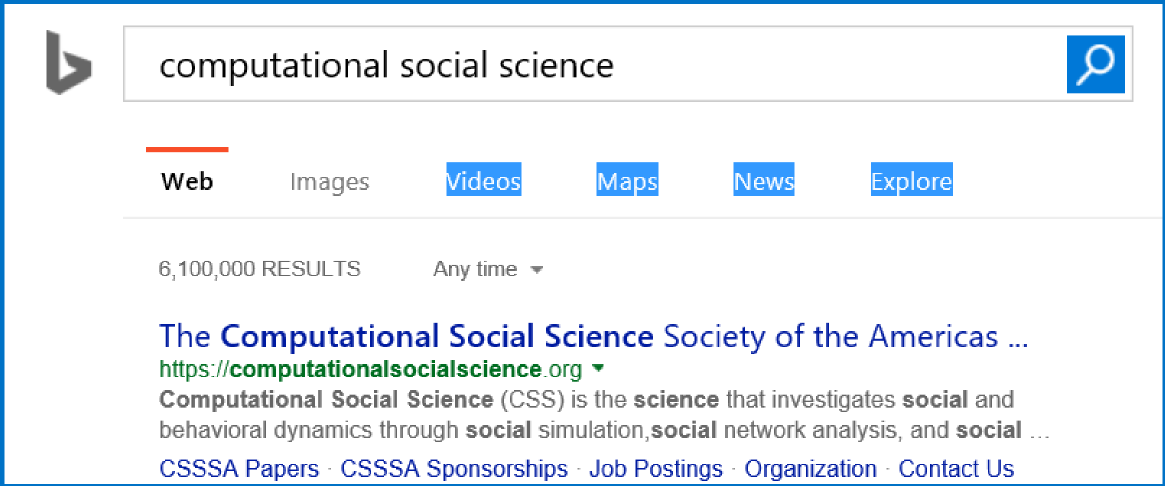
Objectives and Takeaways

- Why should we care about causal inference?
  - Most machine learning algorithms depend on correlations.
  - Correlations alone are a dangerous path to actionable insights.
- Causal inference can help evaluate impact of systems
  - What is the additional revenue if we build a recommender system?
- Causal inference can make prediction models more robust
  - Ensure assumptions more robust to changes in data.

I. We have increasing amounts of data and highly accurate predictions. How is causal inference useful?

Predictive systems are everywhere





How do predictive systems work?

Aim: Predict future activity for a user.


• • •

We see data about their user profile and past activity.

E.g., for any user, we might see their age, gender, past activity and their social network.

From data to prediction

Higher Activity Lower Activity

Use these correlations to make a predictive model.

Future Activity ->

f(number of friends, logins in past month)

From data to "actionable insights"

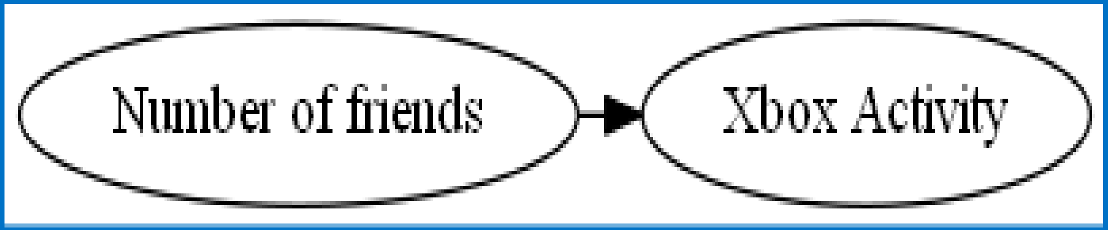
Number of friends can predict activity with high accuracy.

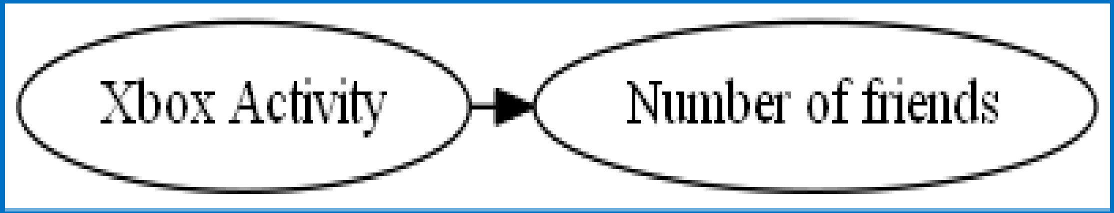
How do we increase activity of users?

Would increasing the number of friends increase people's activity on our system?

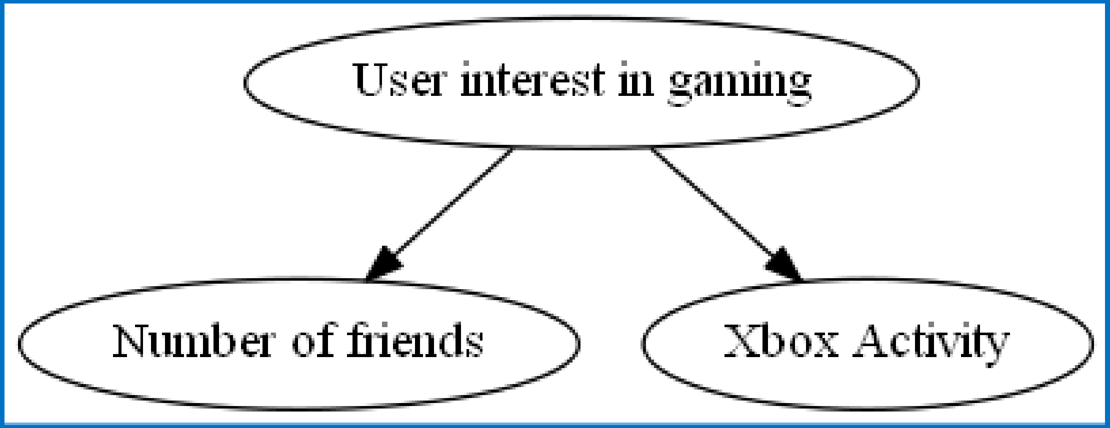
Maybe, may be not (!)

Different explanations are possible





How do we know what causes what?



Decision: To increase activity, would it make sense to launch a campaign to increase friends?

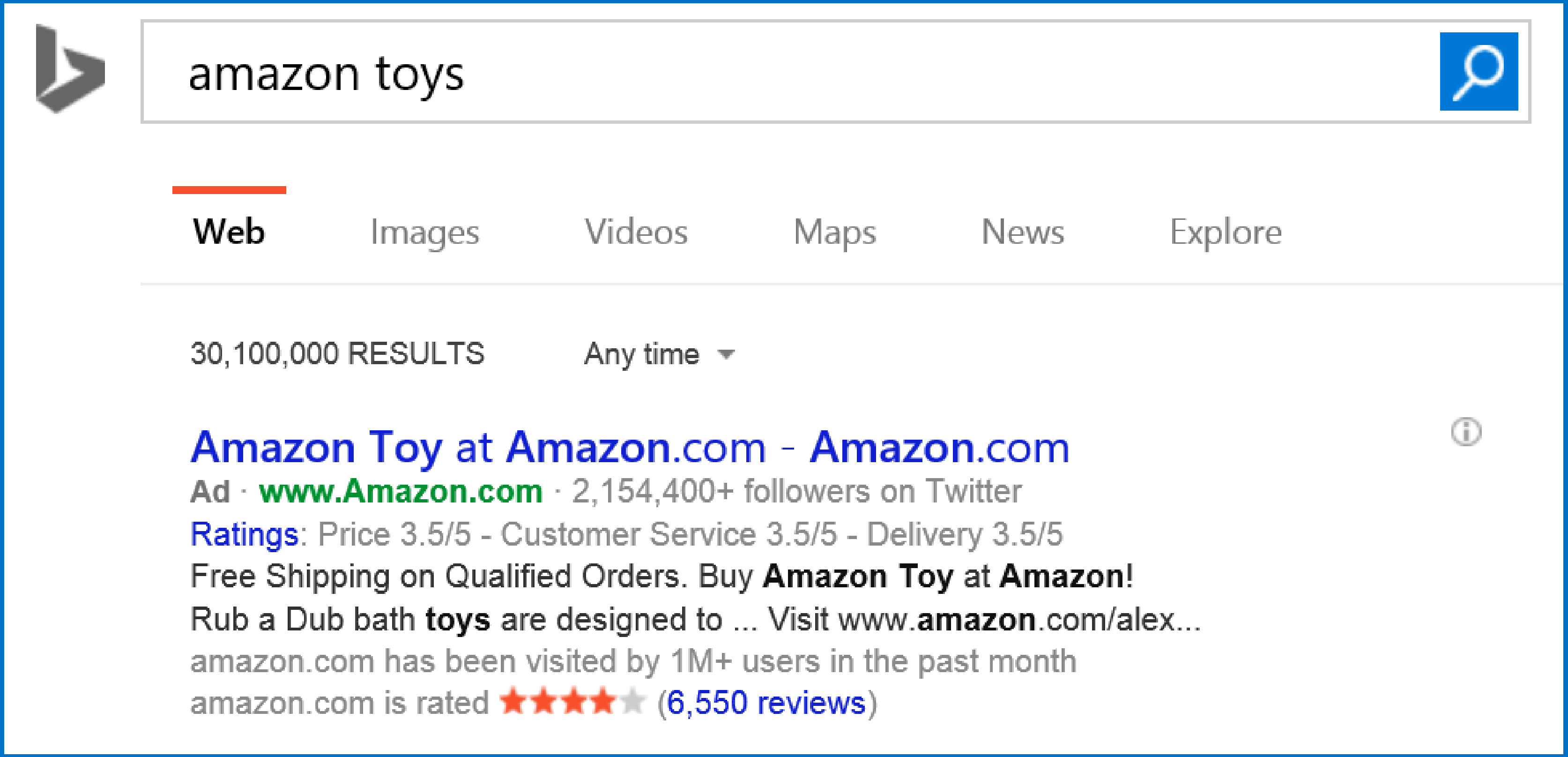
Another example: Search Ads

Search engines uses ad targeting to show relevant ads.

Prediction model based on user’s search query.

Search Ads have the highest click-through rate (CTR) in online ads.

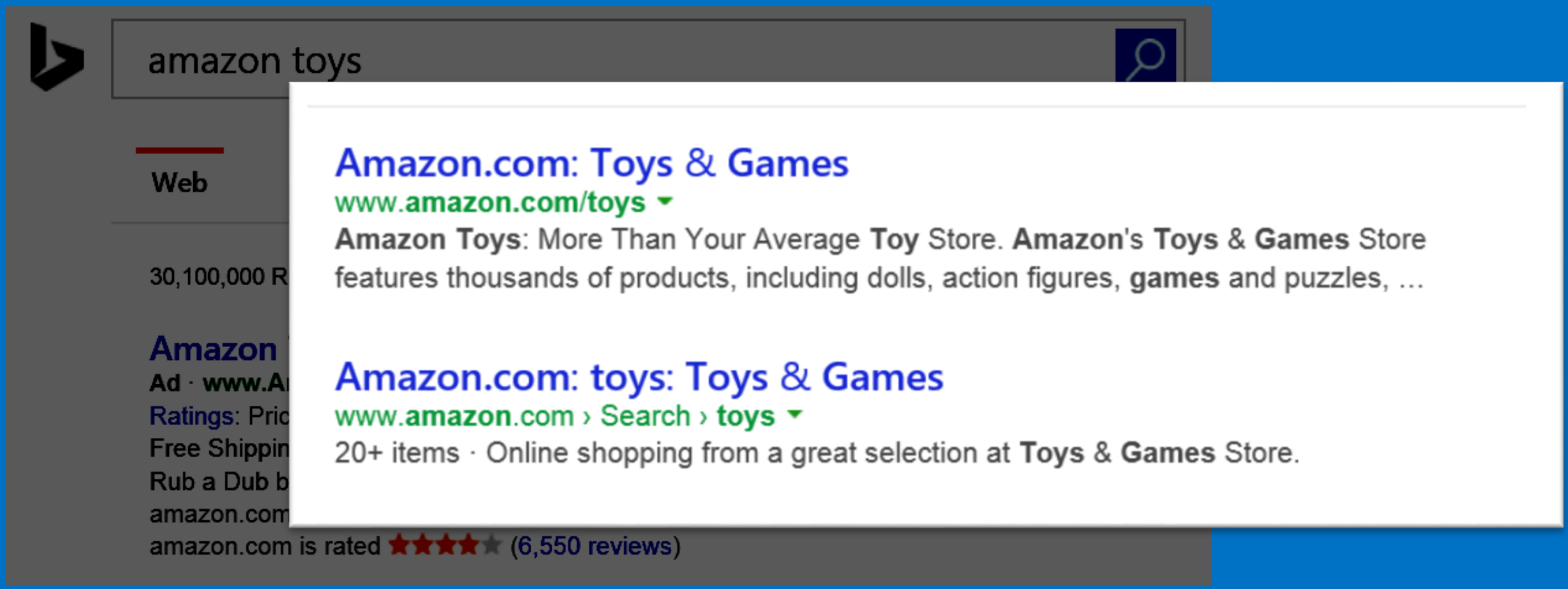
Are search ads really that effective?



Ad targeting was highly accurate.

Blake-Tadelis-Noskos (2014)

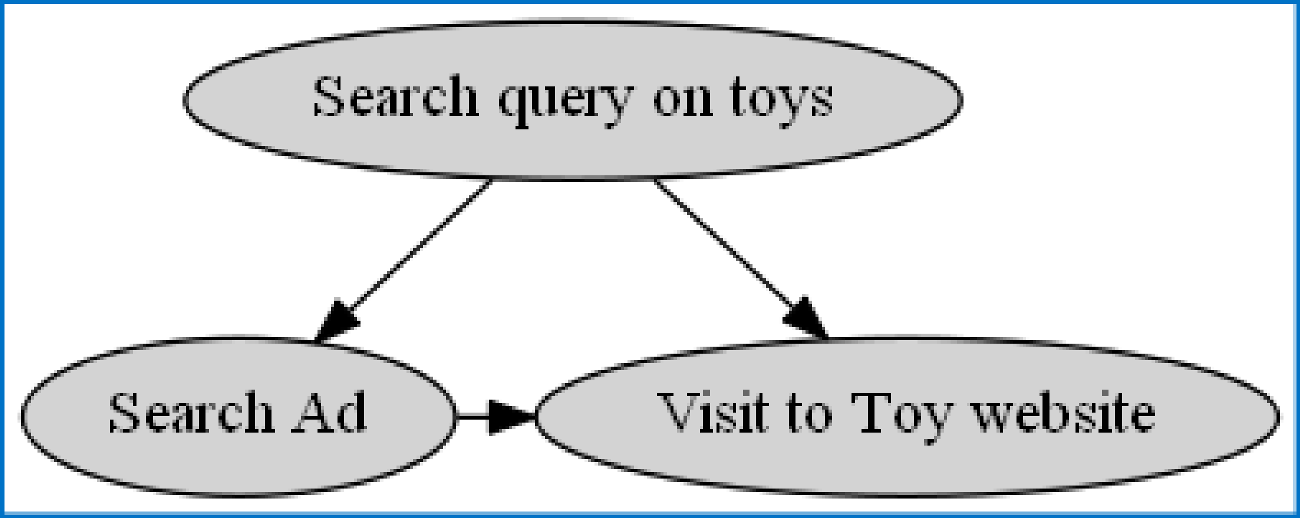
But search results point to the same website



Counterfactual question: Would I have reached Amazon.com anyways, without the ad?

Without reasoning about causality, may overestimate effectiveness of ads

x% of ads shown are effective



<x% of ads shown are effective

Okay, search ads have an explicit intent. Display ads should be fine?

Probably not.

There can be many hidden causes for an action, some of which may be hard to quantify.

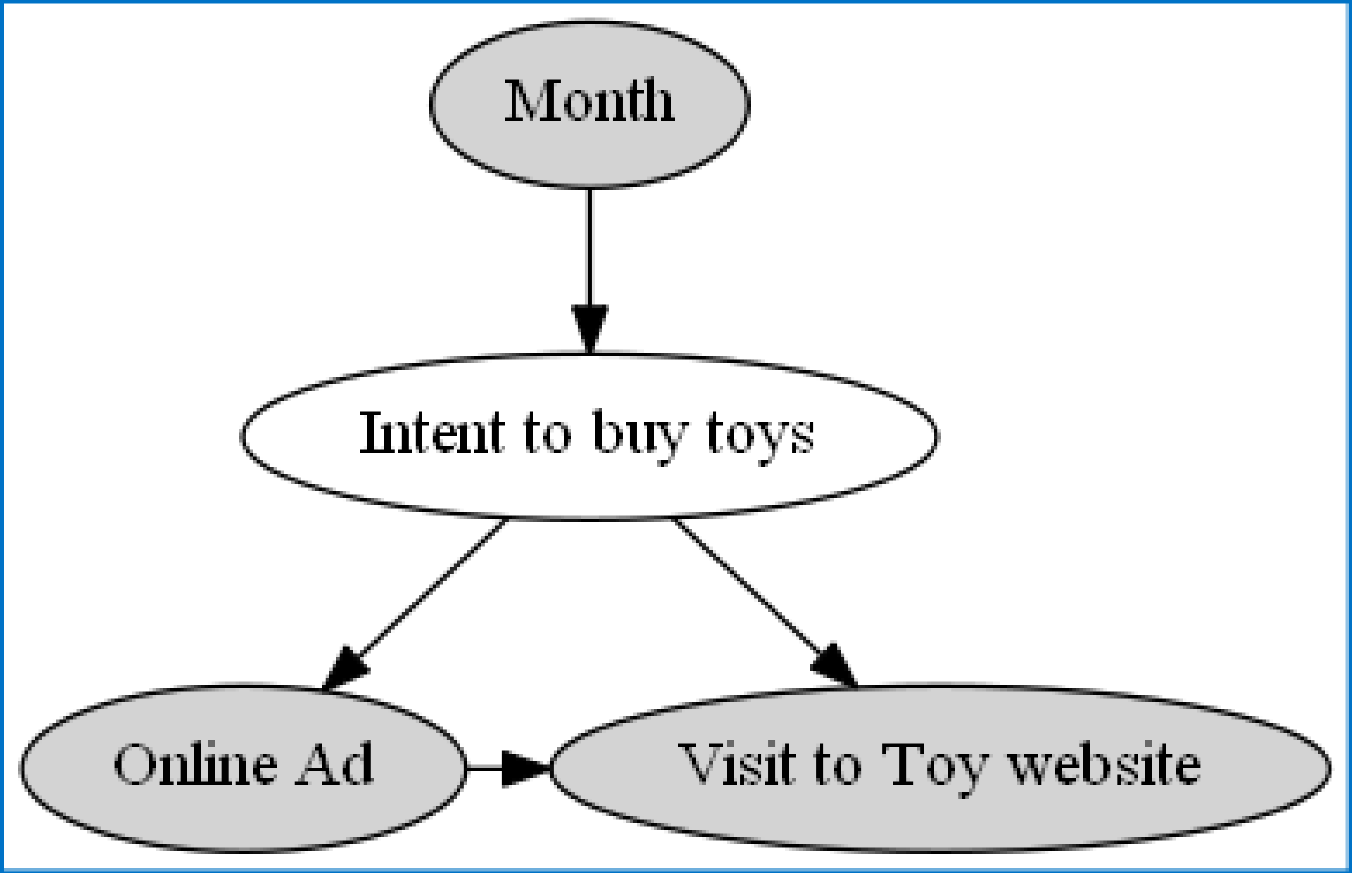
Estimating the impact of ads

Toys R Us designs new ads.

Big jump in clicks to their ads compared to past campaigns.

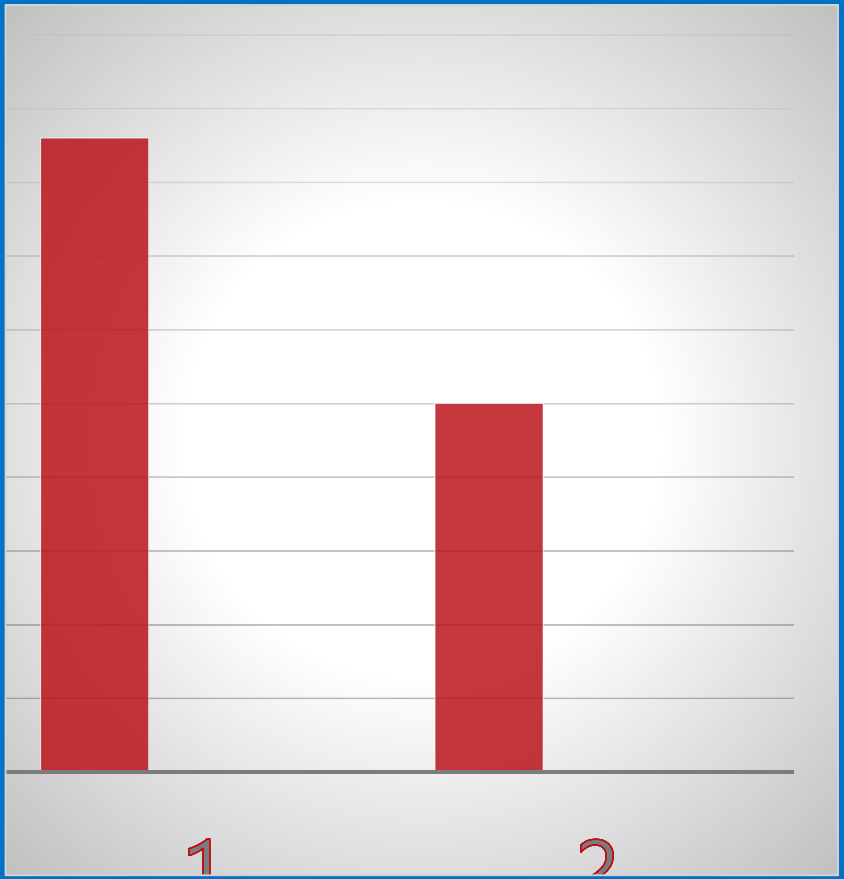
Were these ads more effective?

People anyways buy more toys in December

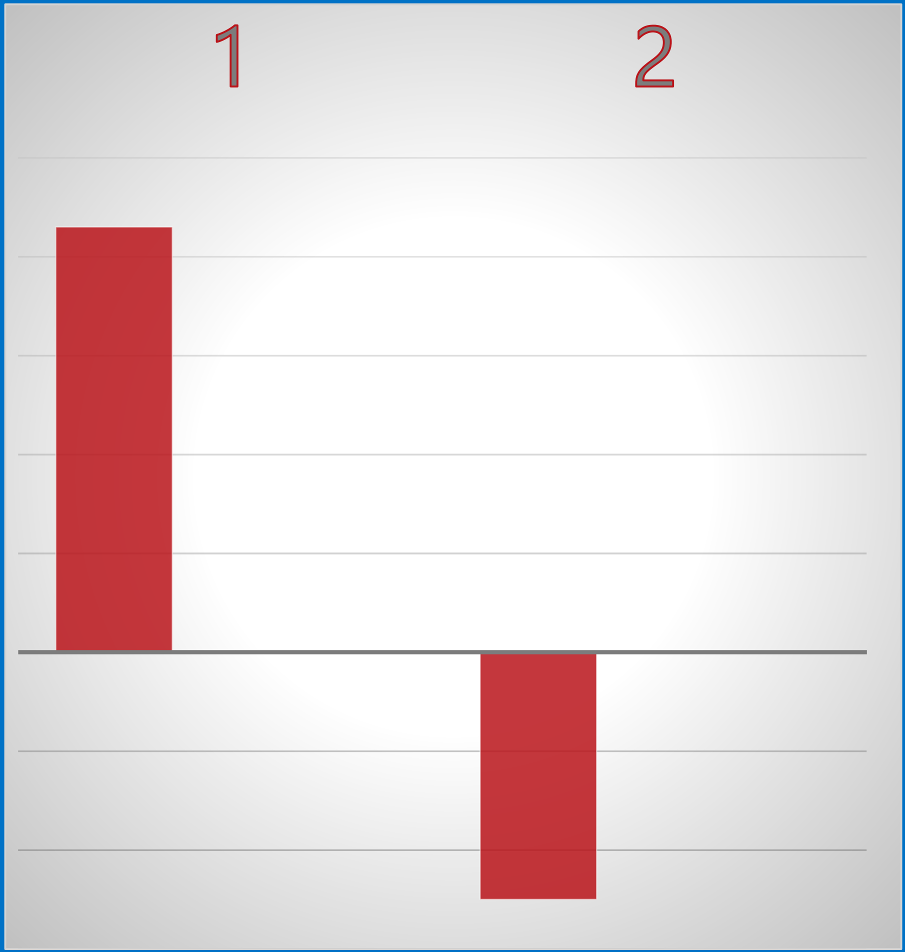


Misleading to compare ad campaigns with changing underlying demand.

So far, so good. Be mindful of hidden causes, or else we might overestimate causal effects.



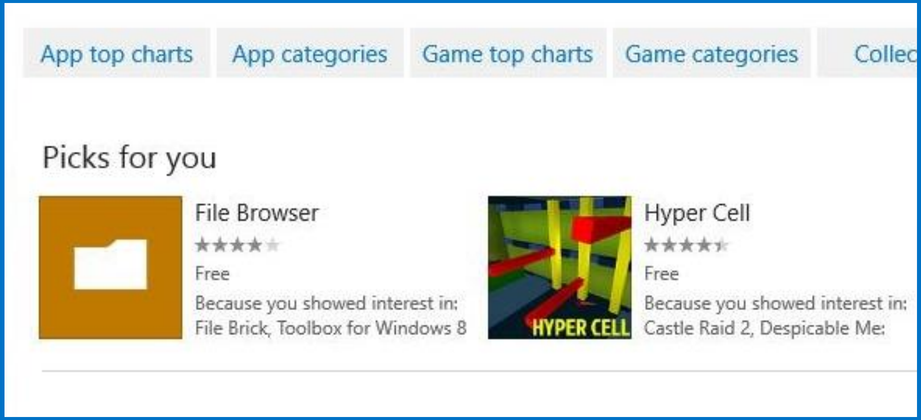
(But)
Ignoring hidden causes can also lead to completely wrong conclusions.



Example: Which algorithm is better?

Have a current production algorithm. Want to test if a new algorithm is better.

Say recommendations on app store.



Algorithm A

?

Algorithm B

Comparing old versus new algorithm

Two algorithms, A (production) and B (new) running on the system.

From system logs, collect data for 1000 sessions for each. Measure CTR.

Old Algorithm (A)	New Algorithm (B)
50/1000 **(5%)**	54/1000 **(5.4%)**

New algorithm is better?

Looking at change in CTR by activity

Suppose we divide users into two groups:

Low-activity

High-activity

Old Algorithm (A)	New Algorithm (B)
10/400 **(2.5%)**	4/200 **(2%)**

Low-activity Users

Old Algorithm (A)	New Algorithm (B)
40/600 **(6.6%)**	50/800 **(6.2%)**

High-activity Users

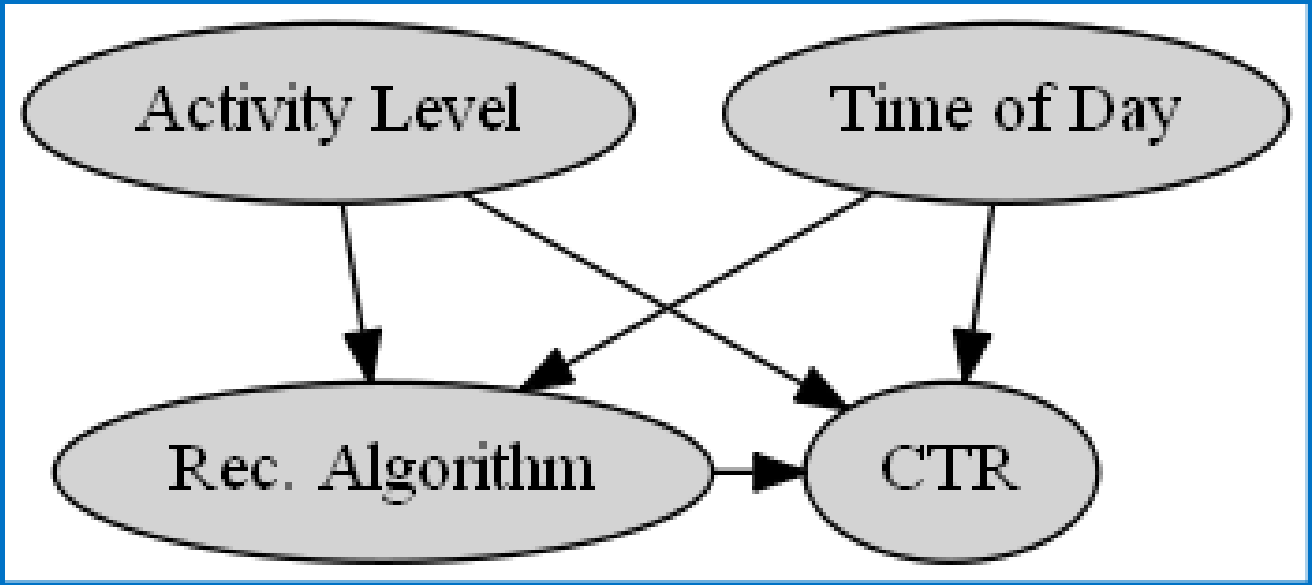
The Simpson's paradox

	Old algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400 (2.5%)	4/200 (2%)
CTR for High-Activity users	40/600 (6.6%)	50/800 (6.2%)
Total CTR	**50/1000 (5%)**	**54/1000 (5.4%)**

Is Algorithm A better?

Simpson (1951)

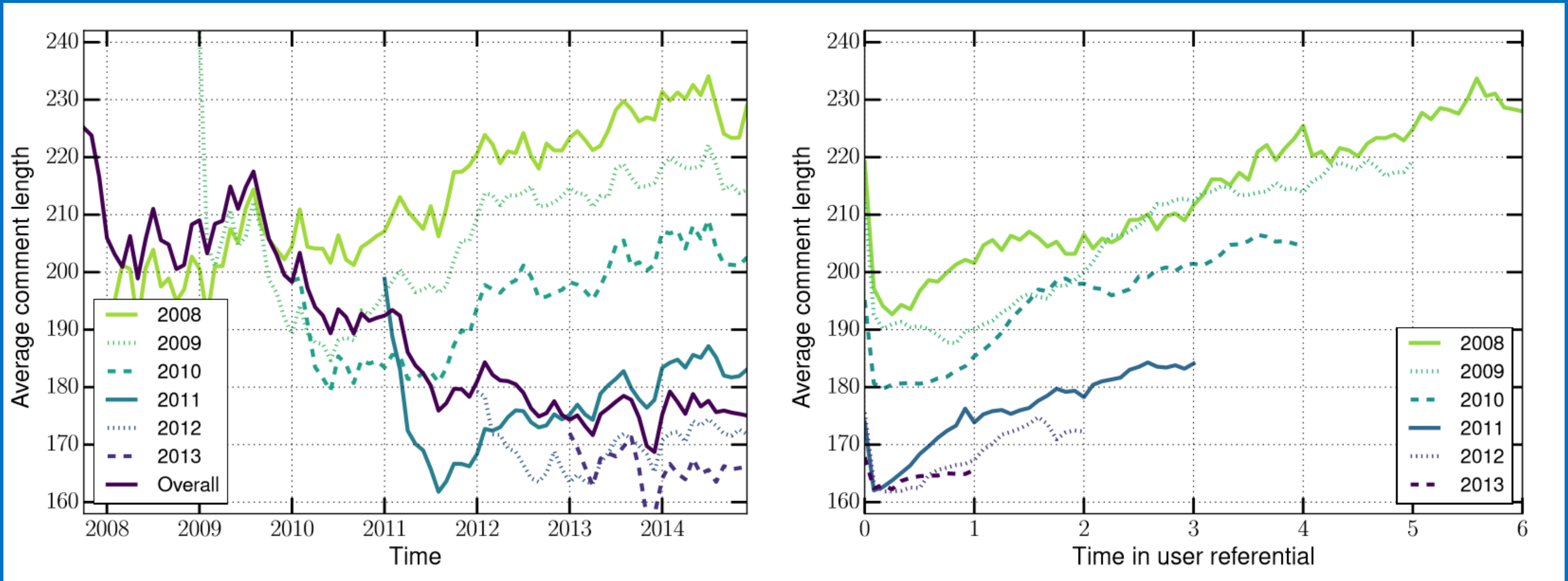
Answer (as usual): May be, may be not.



E.g., Algorithm A could have been shown at different times than B.

There could be other hidden causal variations.

Example: Simpson's paradox in Reddit

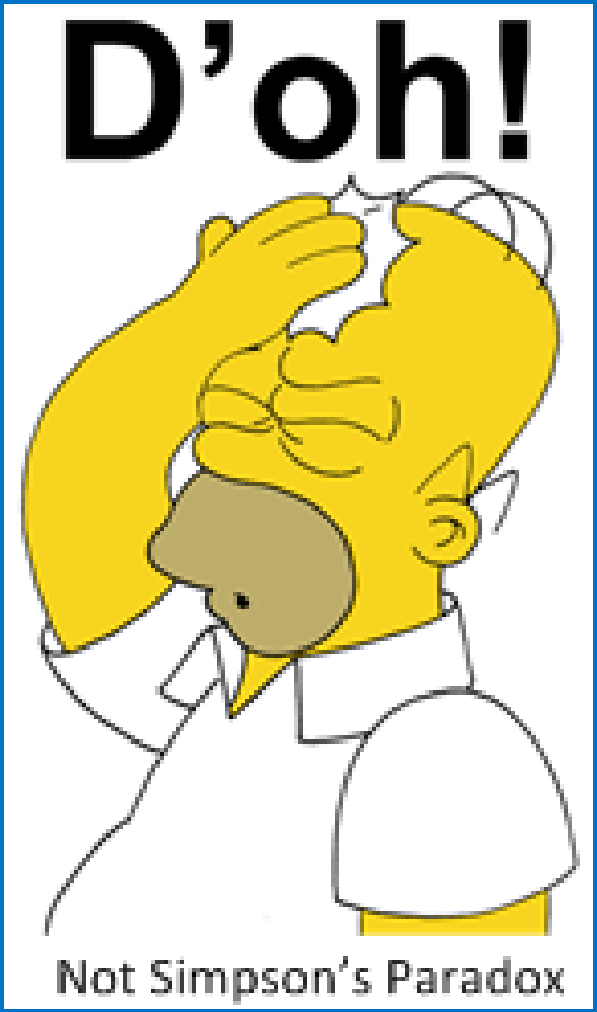


Average comment length decreases over time.

But for each yearly cohort of users, comment length increases over time.

Barbosa-Cosley-Sharma-Cesar (2016)

Making sense of such data can be too complex.



II. How do we systematically reason about and estimate the relationship between effects and their causes?

Formulating causal inference problems

Causal inference: Principled basis for both experimental and non-experimental methods.

Aside: Such questions form the basis of almost all scientific inquiry.

E.g., occur in medicine (drug trials, effect of a drug), social sciences (effect of a certain policy), and genetics (effect of genes on disease).

Frameworks:

- Causal graphical models [Pearl 2009]
- Potential Outcomes Framework [Imbens-Rubin 2016]

What does it mean to *cause*?

A big philosophical debate (since the times of Aristotle, Hume and others).

Practical meaning*: X causes Y iff
changing X leads to a change in Y,
keeping everything else constant.

The **causal effect** is the magnitude by which Y is changed by a unit change in X.

**Interventionist* definition
[http://plato.stanford.edu/entries/causation-mani/]

Need answers to “what if” questions

Basic construct of causal inference.

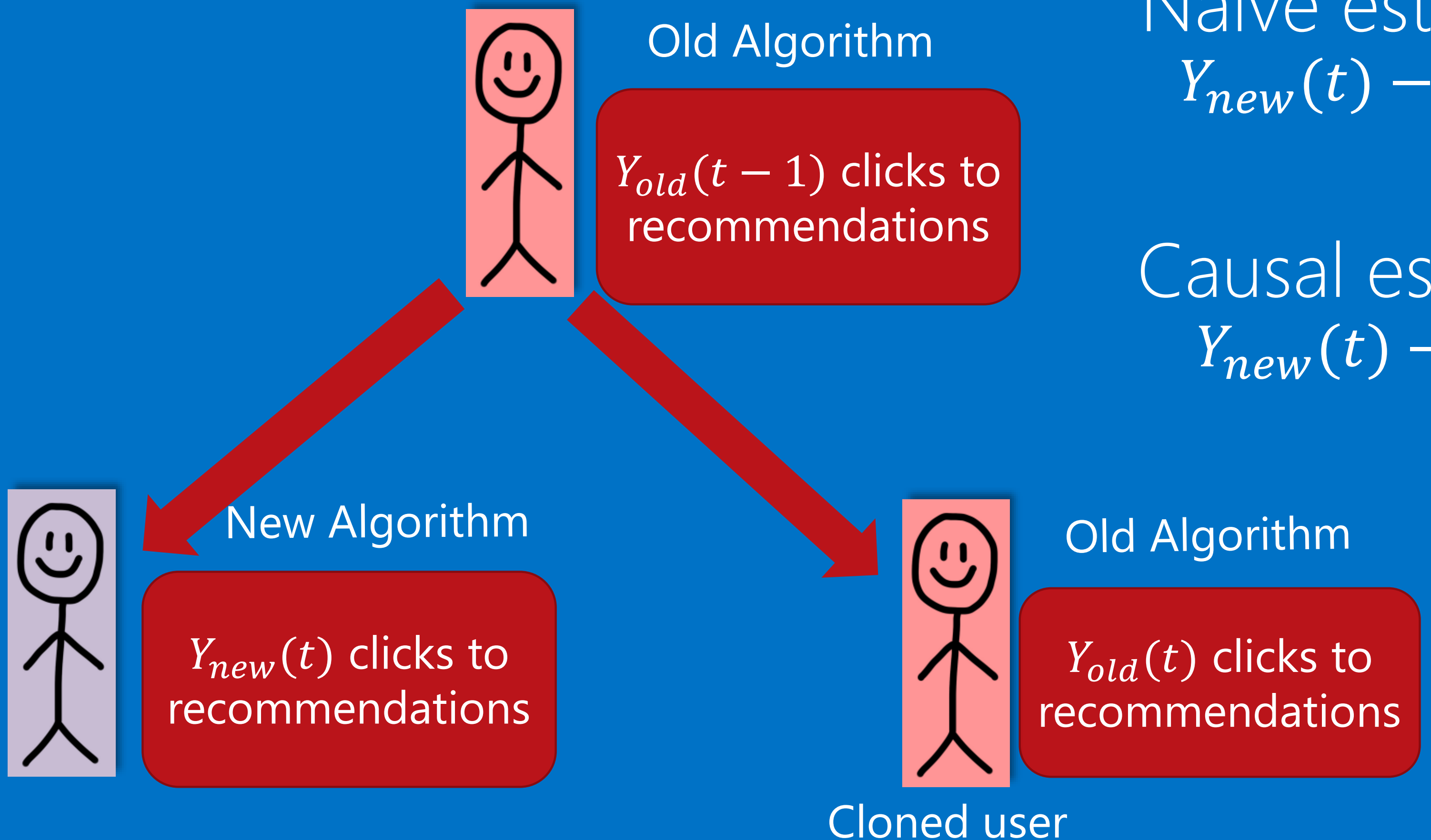
Counterfactual thinking*:
What would have happened if I had changed X?

E.g. What would have been the CTR *had we not* shifted to the new algorithm?

**Counterfactual theories of causation*
http://plato.stanford.edu/entries/causation-counterfactual/

III. Evaluating systems for their causal impact

A hard problem.



Naïve estimate

$$Y_{new}(t) - Y_{old}(t-1)$$

Causal estimate

$$Y_{new}(t) - Y_{old}(t)$$

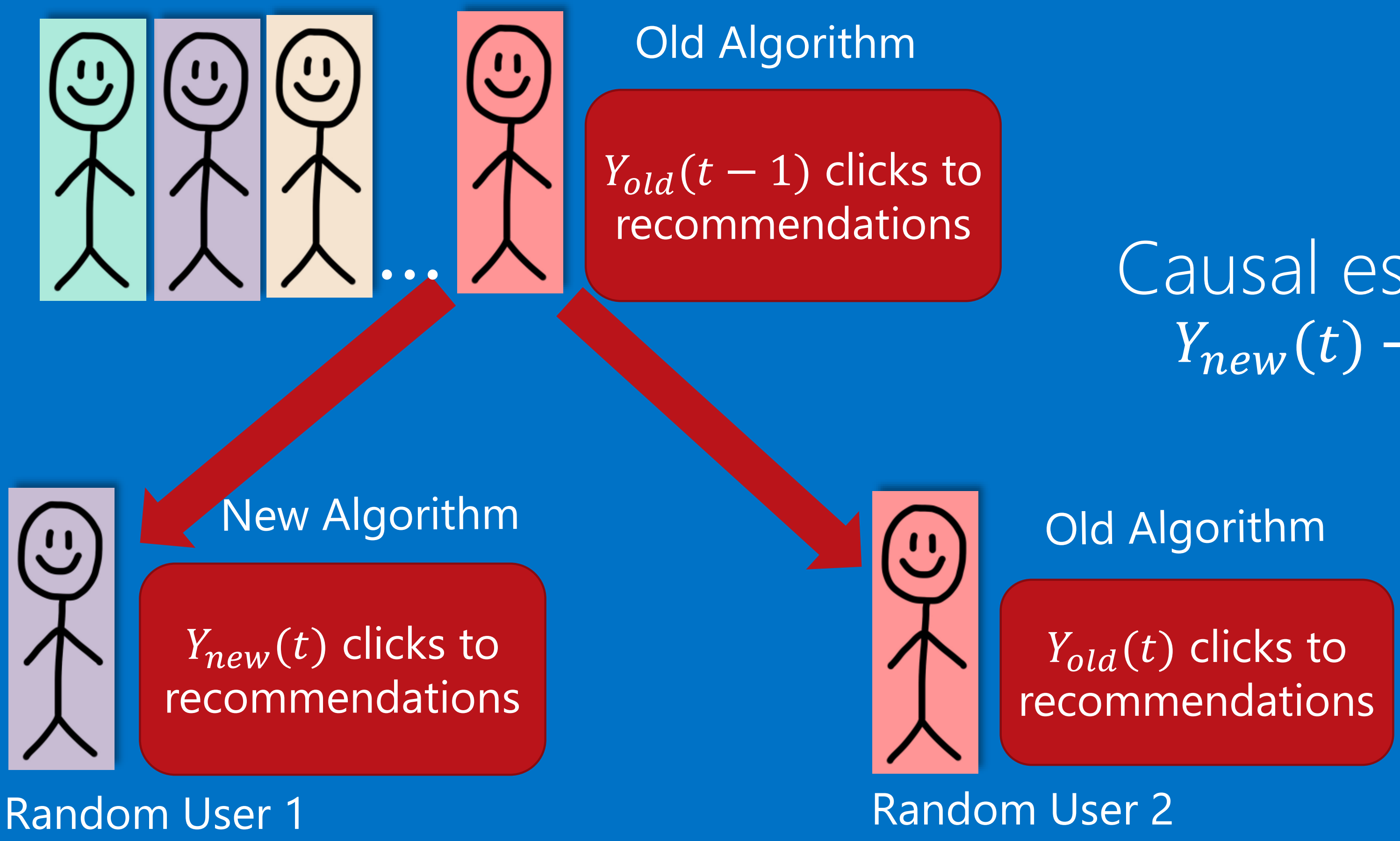
Ideally, requires creation of multiple worlds.

Randomizing algorithm assignment: A/B test

We cannot clone users.

Next best alternative: Randomly assign which users see new Algorithm’s recommendations and which see the old algorithm’s.

Randomization removes hidden variation



Causal estimate
$Y_{new}(t) - Y_{old}(t)$

May be infeasible, unethical or possibly bad experience for many users

Experiment to determine if becoming a subscriber makes you shop more.

Experiment that shows different subscription price to different users to find price elasticity.

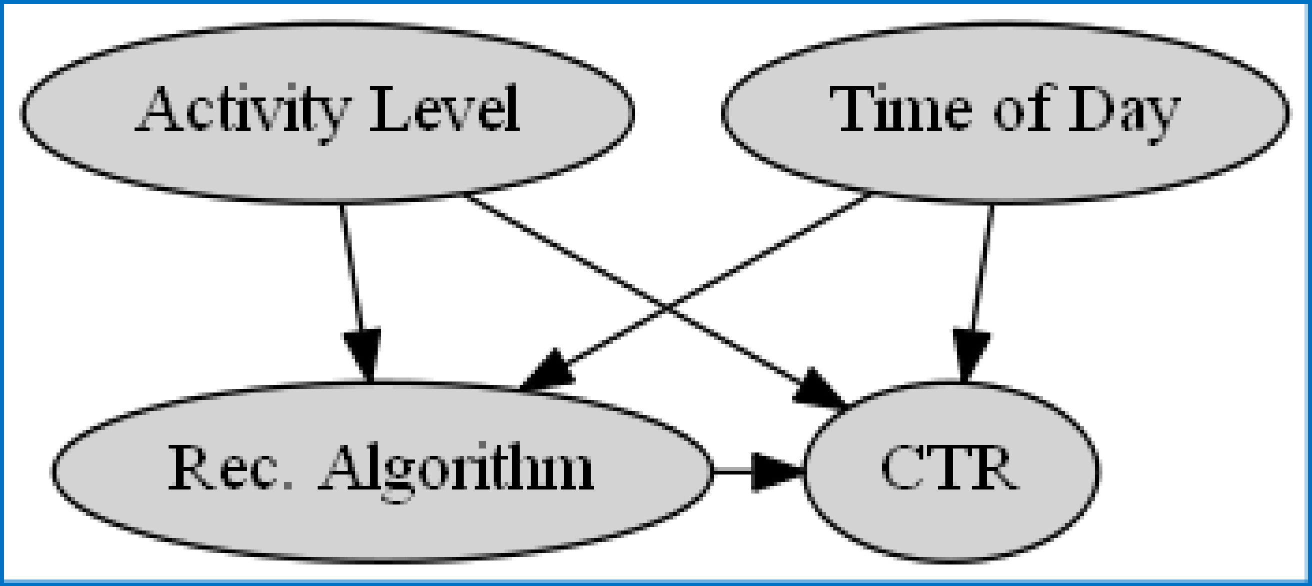
Experiment that changes a familiar UI component.

Even when feasible, randomization methods need a limited set of "good" alternatives to test.

- How do we identify a good set of algorithms or a good set of parameters?

Need **causal metrics.**

What can we do with only observational data (such as log data)?



"Natural" experiments: exploit variation in observed data

Can exploit naturally occurring **close-to-random** variation in data.

Since data is not randomized, need assumptions about the data-generating process.

If there is sufficient reason to believe the assumptions, we can estimate causal effects.

Dunning (2002), Rosenzweig-Wolpin (2000)

Example: Effect of Store recommendations

Suppose instead of comparing recommendation algorithms, we want to estimate the causal effect of showing *any* algorithmic recommendation.

Can be used to benchmark how much revenue a recommendation system brings, and allocate resources accordingly.

(and perhaps help analyze the tradeoff with users' privacy)

A natural experiment: Instrumental Variables

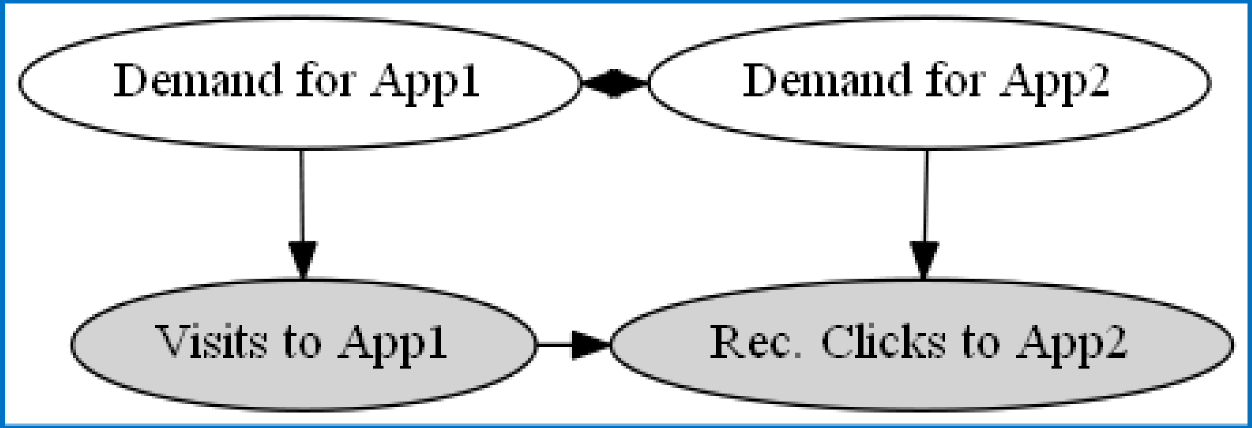
Can look at *as-if random* variations due to external events.

E.g. Featuring on the Today show may lead to a sudden spike in installs for an app.

Such external *shocks* can be used to determine the causal effect of showing recommendations.

Angrist-Pischke (2008)

Cont. example: Effect of store recommendations

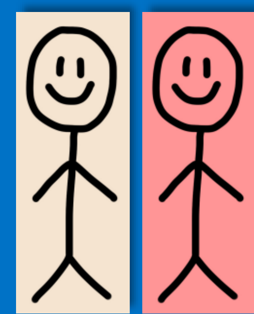


How many new visits are *caused* by the recommender system?

Demand for App 1 is correlated with demand for App 2.

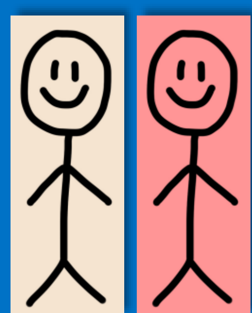
⇒ Users would most likely have visited App 2 even without recommendations.

Traffic on normal days to App 1



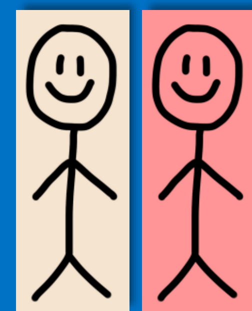
$Y_{old}(t-1)$ click-throughs from App 1 to App 2

Cannot say much about the causal effect of recommendations from App 1.



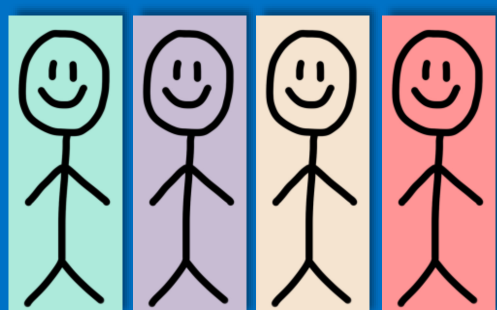
$Y_{new}(t)$ click-throughs from App 1 to App 2

External shock brings as-if random users to App1



$Y_{old}(t-1)$ click-throughs from App 1 to App 2

Spike in visits to App 1



$Y'_{new}(t)$ click-throughs from App 1 to App 2

If demand for App 2 remains **constant**, additional views to App 2 would not have happened had these new users not visited App 1.

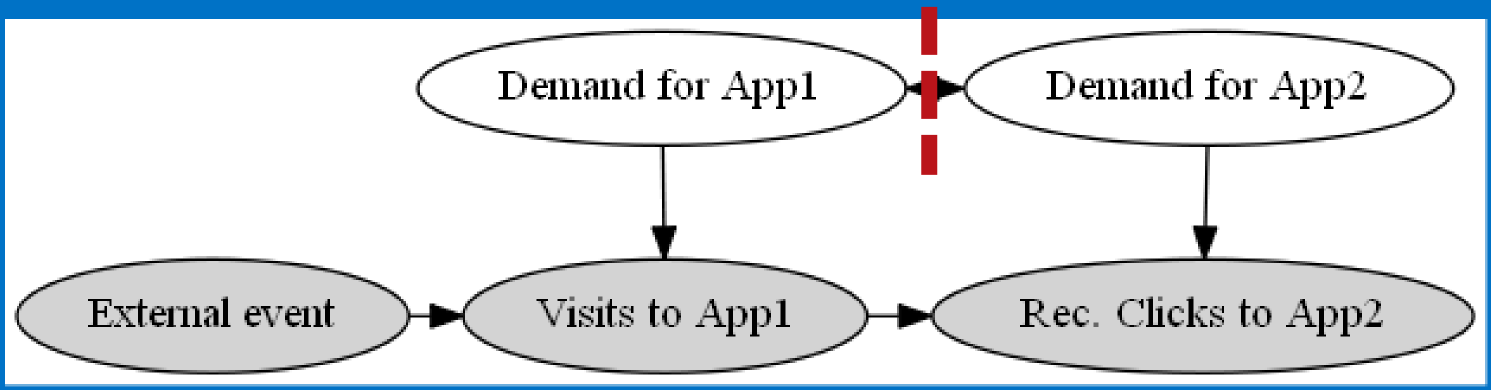
$$Causal\ clicks = Y'_{new}(t) - Y_{old}(t-1)$$

Sharma-Hofman-Watts (2015)

Exploiting sudden variation in traffic to App 1

To compute Causal CTR of Visits to App1 on Visits to App2:

- Compare observed effect of external event separately on Visits to App1, and on Rec. Clicks to App2.
- Causal click-through rate $= \dfrac{\Delta(\text{Rec. Click-throughs from App1 to App2})}{\Delta(\text{Visits to App1})}$



Caveat: Natural experiments are hard to find

Estimates may not be generalizable to all products.

Whenever possible, use randomization.

If number of output items low, consider using contextual bandits.

If randomization is not feasible, consider exploiting natural experiments.

Better to consider multiple sources of natural experiments.

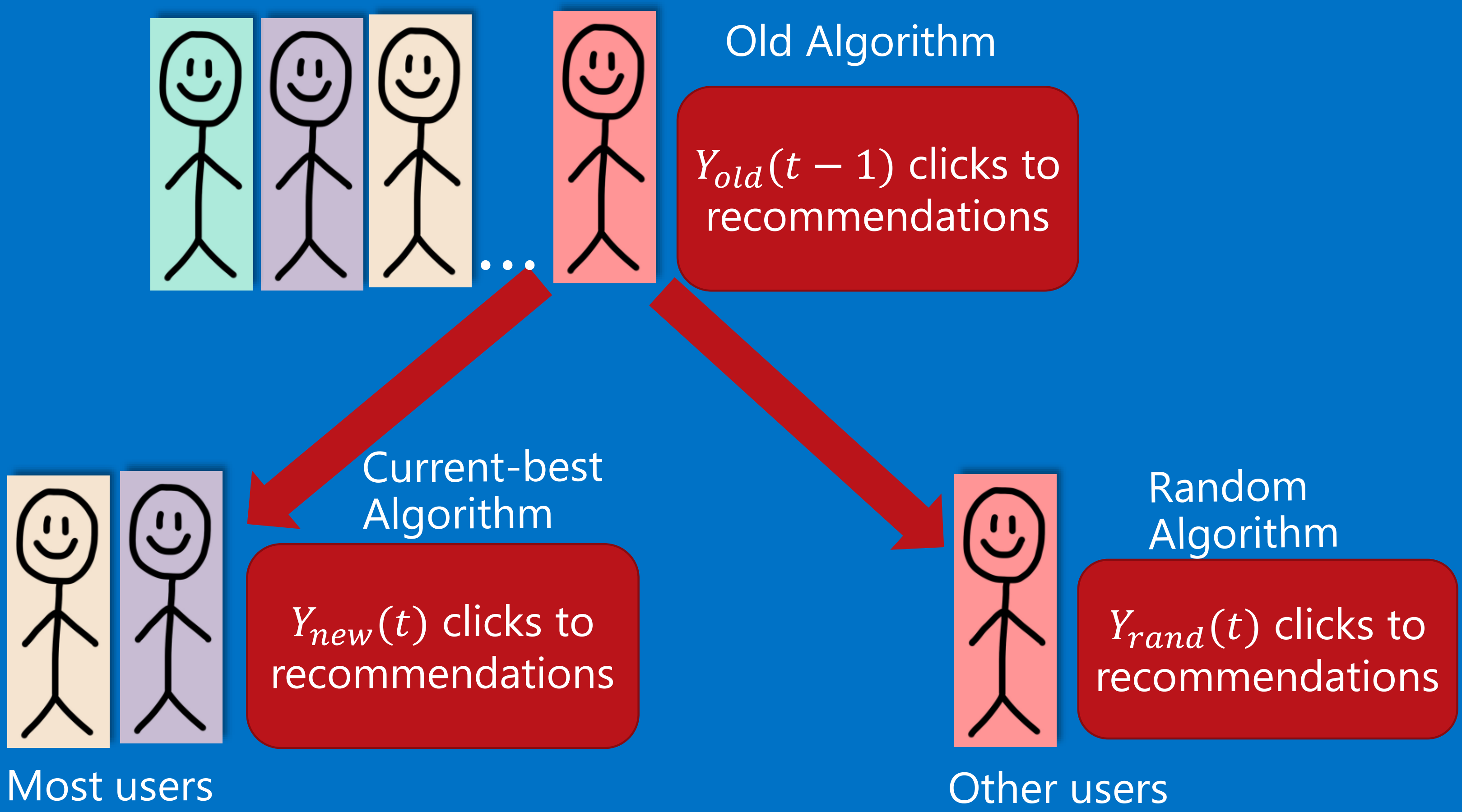
IV. Developing robust prediction algorithms with causal inference

Continuous experimentation: Multi-armed bandits

Two goals:

1. Show the best known algorithm to most users.
2. Keep randomizing to update knowledge about competing algorithms.

Bandits: The right mix of explore and exploit



Algorithm: ε-greedy multi-armed bandits

Repeat:

(Explore) With low probability ε, choose an output item randomly.

(Exploit) Otherwise, show the current-best algorithm.

Use CTR results for Random output items to train new algorithms offline.

Practical Example: Contextual bandits on Yahoo! News

Actions: Different news articles to display

A/B tests using all articles inefficient.

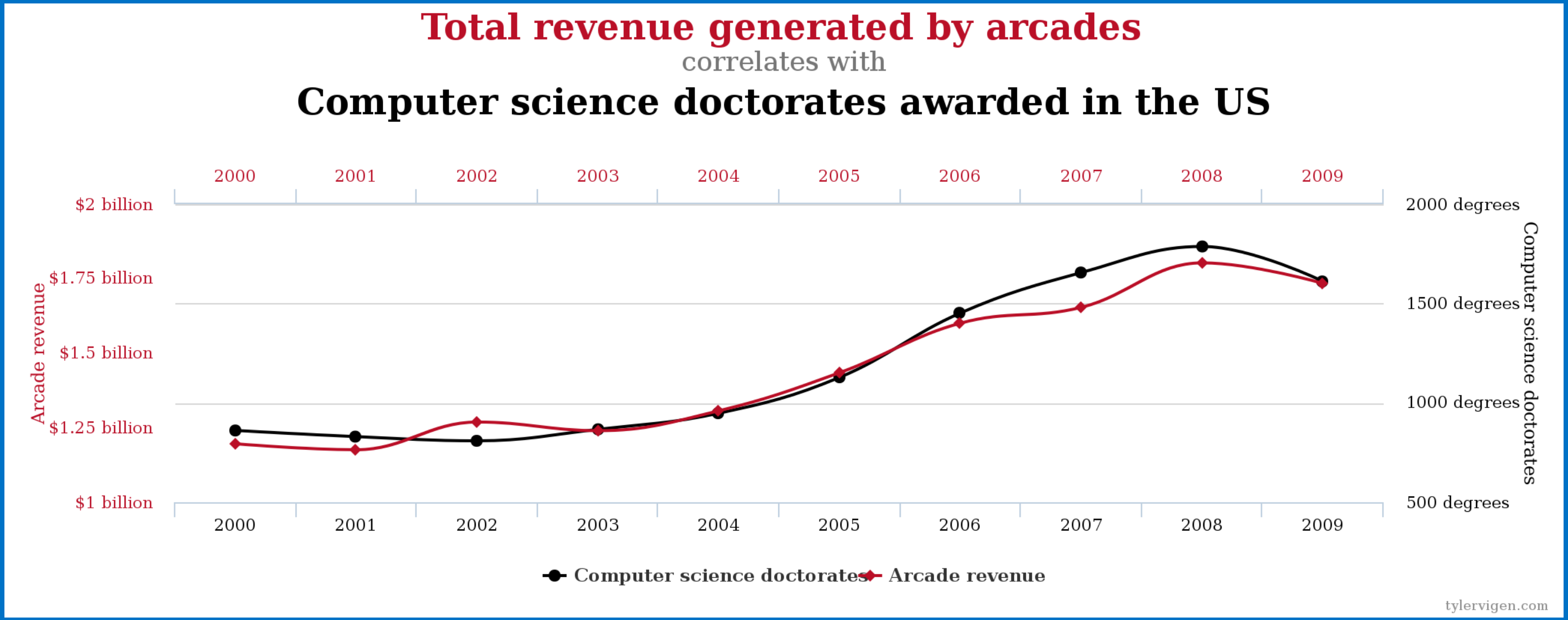
Randomize the articles shown using ε-greedy policy.

Better: Use context of visit (user, browser, time, etc.) to have different current-best algorithms for different contexts.

Li-Chu-Langford-Schapire (2010)

Causal inference is tricky

Correlations are seldom enough. And sometimes horribly misleading.



Always be skeptical of causal claims from ~~observational~~ any data.

More data does not automatically lead to better causal estimates.

http://tylervigen.com/spurious-correlations

For more with R code and a practical example, check out the Github repo:
http://www.github.com/amit-sharma/causal-inference-tutorial

Microsoft

thank you!

@amt_shrma
amshar@microsoft.com

© 2015 Microsoft Corporation. All rights reserved.