keynote

Uplift Modeling for Digital Advertising

Eustache Diemert
Criteo AI Lab

Agenda

1

Why Uplift Modeling in AdTech ?

2

Instances of the UM problem

3

Challenges of UM in AdTech

4

Beyond Uplift Modeling

Source:

Agenda

1

Why Uplift Modeling in AdTech ?

2

Instances of the UM problem

3

Challenges of UM in AdTech

4

Beyond Uplift Modeling

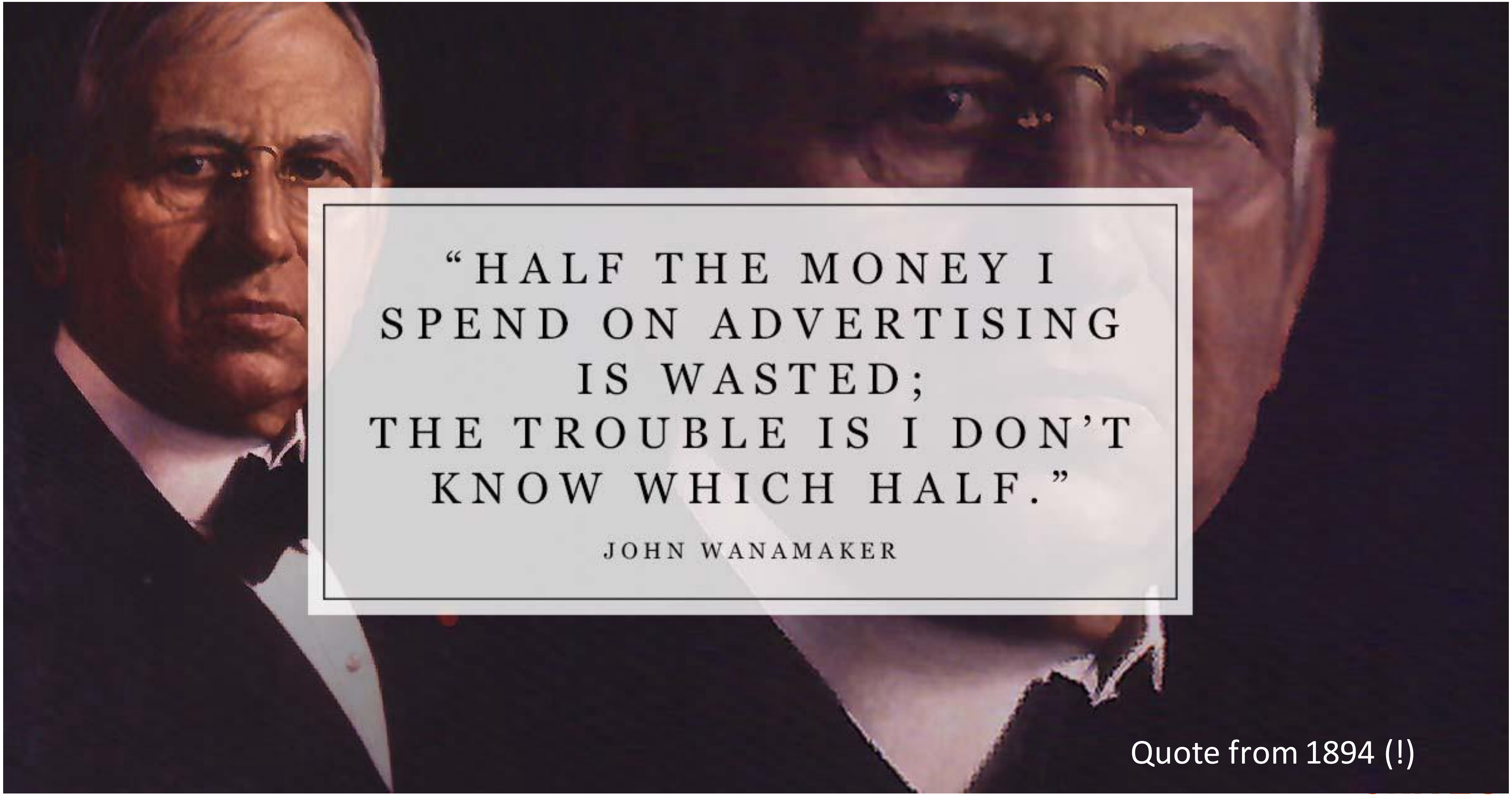
~~Most~~ all of this talk is based on work from the Criteo AI Lab

Source:

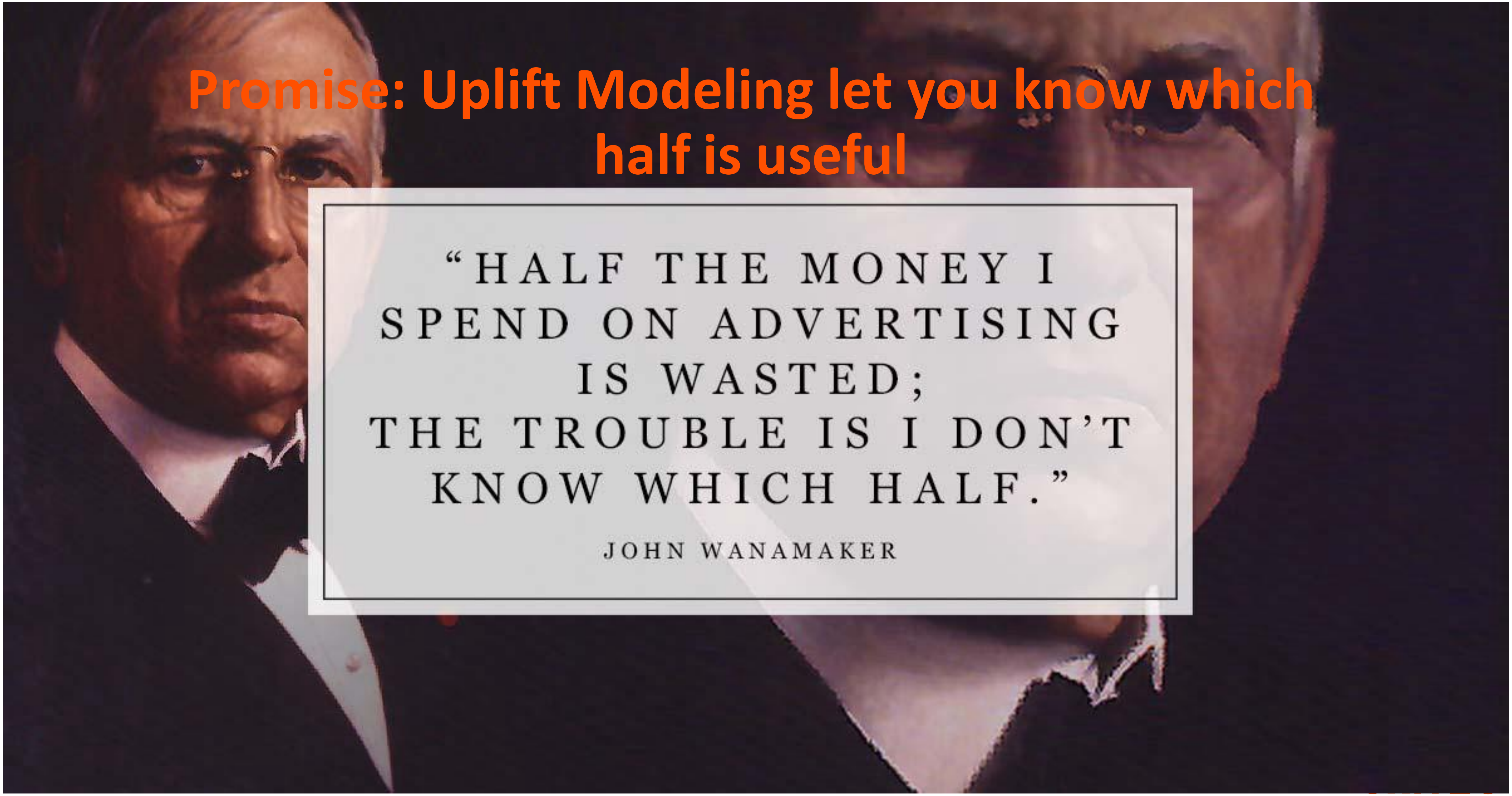
CRITEO

Why Uplift Modeling in AdTech ?

CRITEO

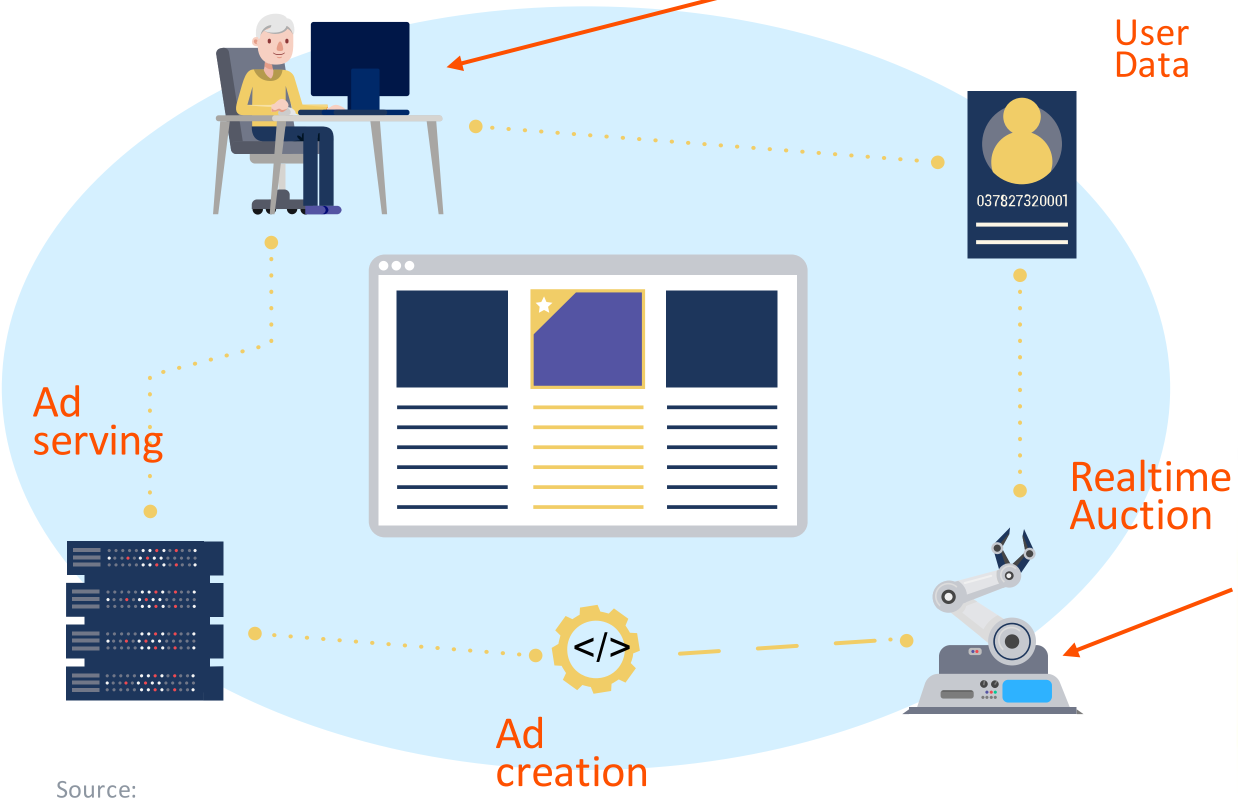

"HALF THE MONEY I SPEND ON ADVERTISING IS WASTED;
THE TROUBLE IS I DON'T KNOW WHICH HALF."
JOHN WANAMAKER
Quote from 1894 (!)

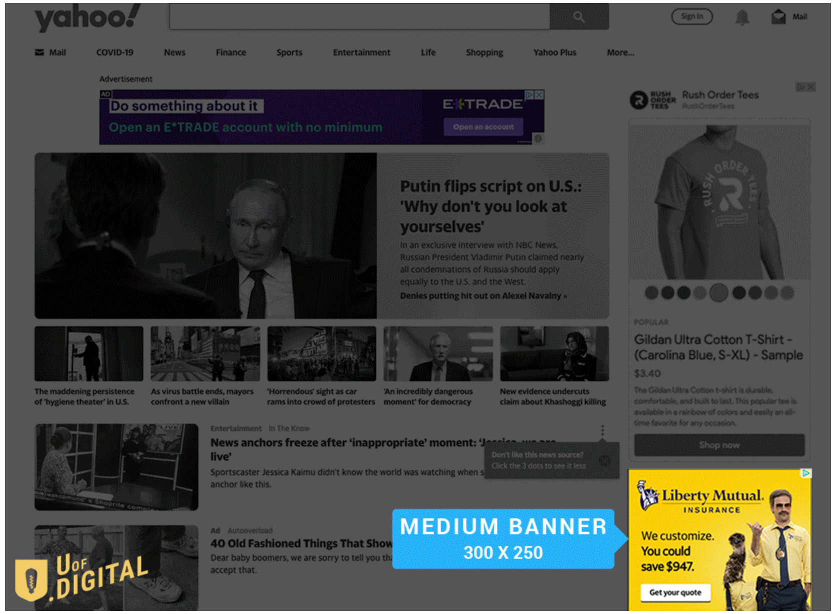
Promise: Uplift Modeling let you know which half is useful

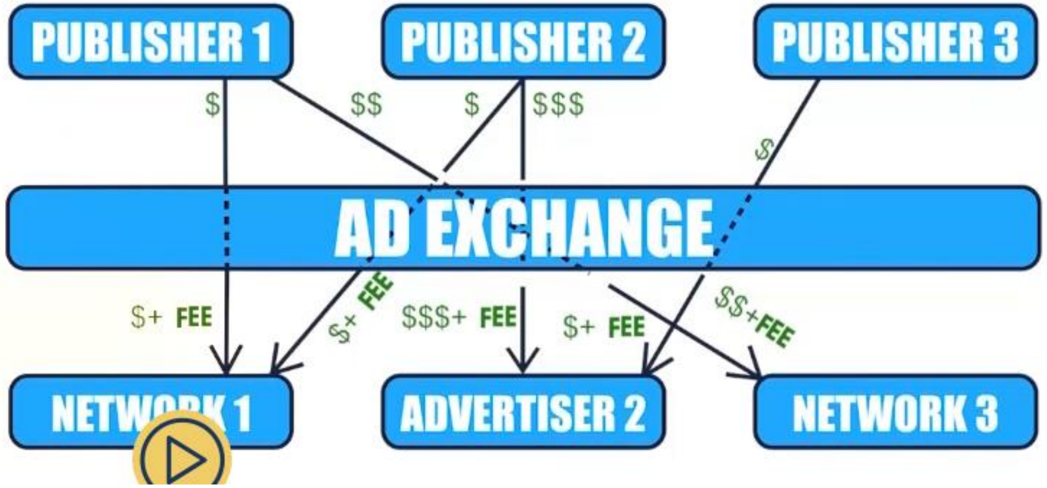


AdTech 101

How ads are run today







Source:

The problem with clicks

the industry standard is (was) for marketers to pay providers per sale *after a click*

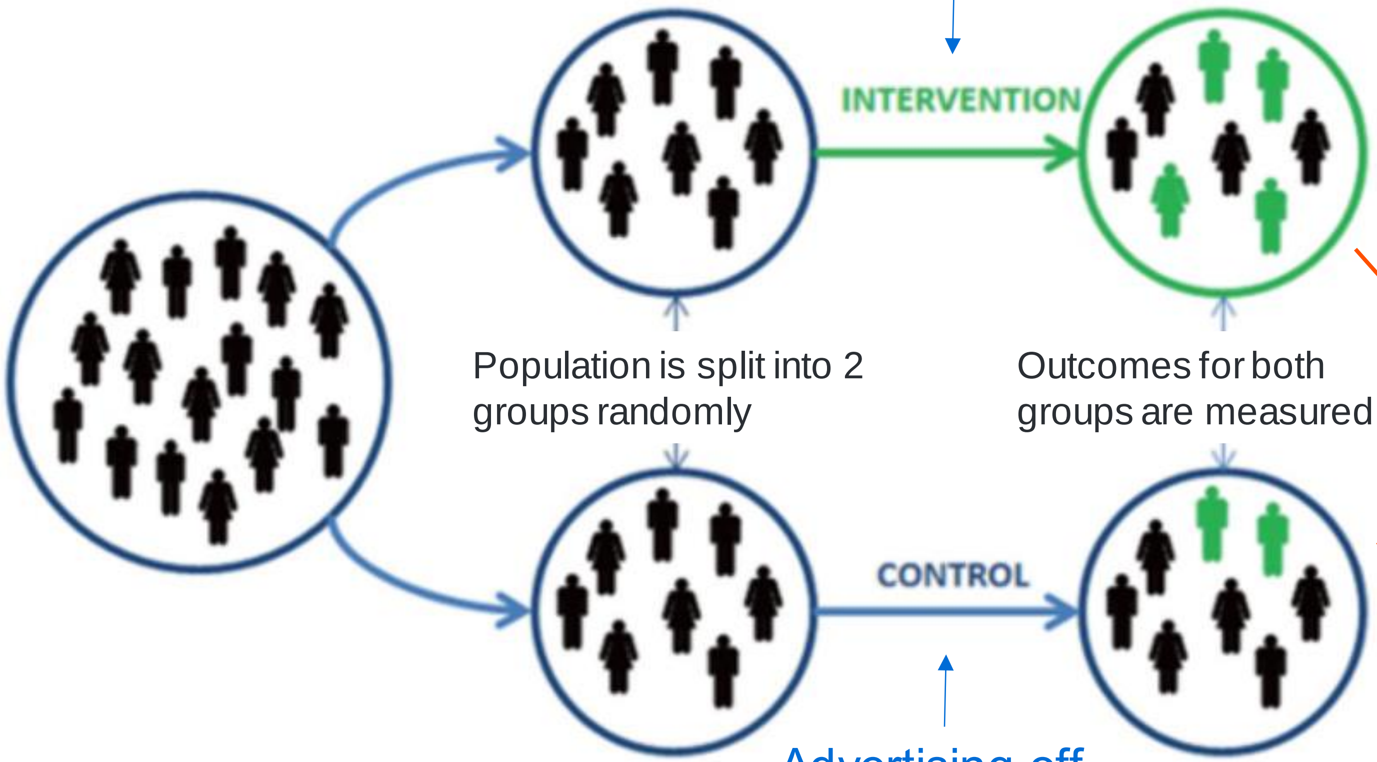
Post-click attribution assumes all the causal effect is happening through clicks

Source:

The new way to measure Ad efficacy

Incrementality = average causal effect of ads

Advertising on



Population is split into 2 groups randomly

Outcomes for both groups are measured

Incrementality is the difference between the two

Advertising off

Incrementality measures the amount of sales explained by ads w/o further assumptions

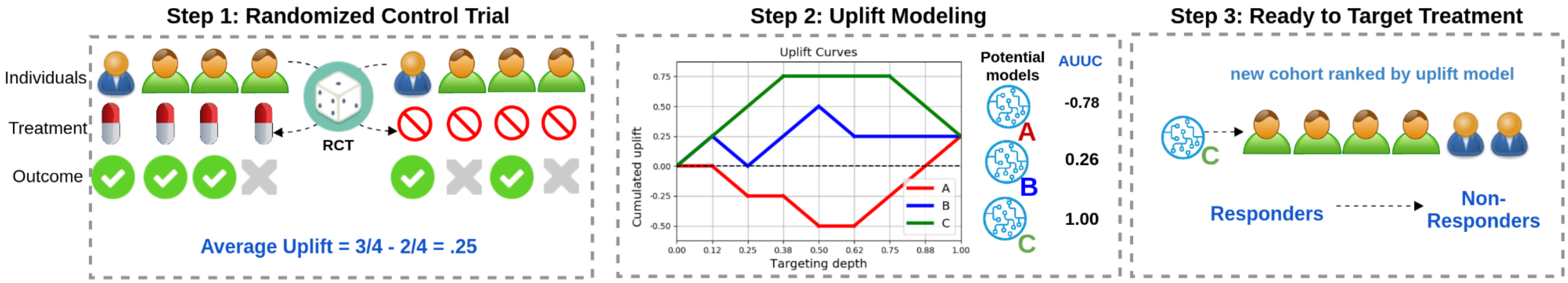
Instances of the UM problem in AdTech

Outline: how to use Uplift Models

From intervention to data, to learning UMs, to using them in production

1. Data collection
   - decide intervention
   - run the system, collect data
2. Learn models
3. Use predictions to improve production

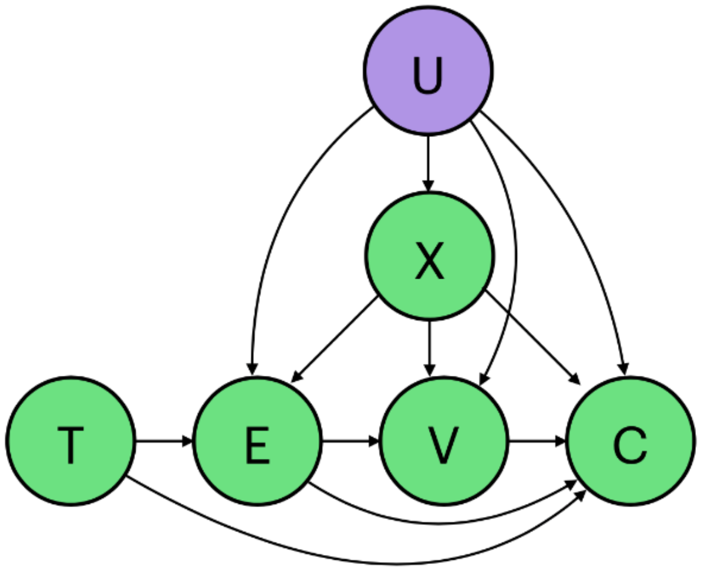
Try to expose users that are responsive to ads



A plausible causal model of advertising

Notations:

- **T (treatment)**: binary, intent to treat (bid or not)
- **E (exposure)**: binary, won the auction (ad displayed to user)
- **V (visit)**: binary, user visited website
- **C (conversion)**: binary, user converted (bought something)
- **X (context)**: multi-dimensional, observable context
- **U (unobserved)**: multi-dimensional++, un-observed confounders



Assumptions

- T = 0 implies E = 0 (no bid implies no ads)
- V = 0 implies C = 0 (no conversion w/o visit)

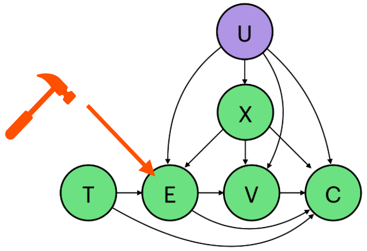
Specificity:

- T=1 does not imply E=1: bidding doesn't imply exposure (because of competition, floor prices etc) aka "one sided non-compliance to prescription"

Source:

CRITEO

1st idea: intervene on the display



Outline

1. Bid in the auction as usual
2. When auction is won, decide to treat randomly
3. If user assigned to control population, display a blank ad (or a charity ad)

Interpretation: U(x) = P(C=1|X=x, do(E=1)) - P(C=1|X=x, do(E=0))

Problems

- Winning the auction and not displaying ad = you lose money
- Winning the auction prevents other competitors to display their ads = you under-estimate the causal effect
    - E=0 means "no display for the advertiser", but competitors can place ads and persuade customers to buy their product instead !

Source:

2nd idea: intervene on the bid

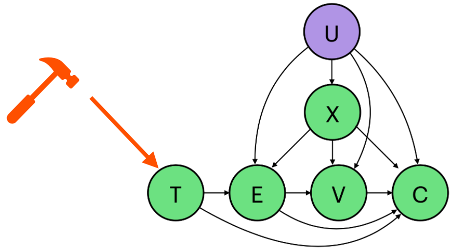
Outline

1. Decide to treat randomly
2. When in control, don't place any bid (conversely: in treatment bid as usual)
3. When auction is won, proceed as usual

Interpretation: U(x) = P(C=1|X=x, do(T=1)) - P(C=1|X=x, do(T=0))

Problem

- Not all bids are successful, so signal is drowned in noise
  - P(E=1|do(T=1)) can be as low as 15%
  - Can be alleviated by zooming on most plausible auction winners
    - using e.g. P^(E=1|do(T=1), X=x) as symmetrical filter/ranker
    - akin to a control variable



Competitor ad

Source:

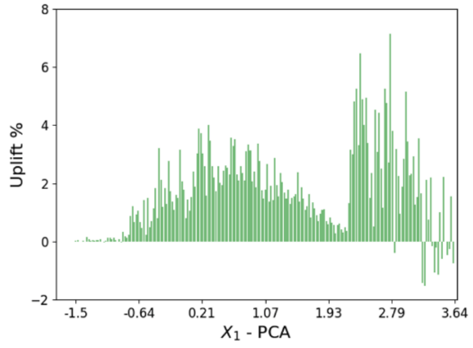
Challenges of UM in AdTech

CRITEO

Challenge #1: noise in uplift signal

Subtitle

- Conversions are noisy by nature
  - P(C=1|E=1) =~ 1e-3/1e-4 --> Expectation and Variance are of the same order
  - May vary widely depending on vertical/advertiser (travel vs retail vs finance vs …)
- Uplift is noisier than conversions
  - E[U] =~ 1e-5
  - Var[U] = Var[E[C=1|do(T=1)]] + Var[E[C=1|do(T=0)]] =~ 2* Var(C) =~ 2e-3



Uplift is heterogeneous

Metric	IHDP (Hill)	JOBS	IHDP (ACIC 2017)	HILLSTROM	CRITEO-UPLIFTv2 (ours)
Size	747	3,212	4,302	42,693	13,979,592
Dimension	25	7	25	8	12
- Continuous	6	3	6	2	4
- Binary	19	4	19	3	0
- Multiple modalities	0	0	0	3	8
Treatment Ratio	.19	.09	-	.50	.85
Avg. positive outcome (Label 1 / Label 2)	-	84.99%	-	12.88% / 0.73%	4.70% / 0.29%
Relative Avg. Uplift (Label 1 / Label 2)	-	-9.7%	-	42.6% / 54.3%	68.7% / 37.2%

Visits

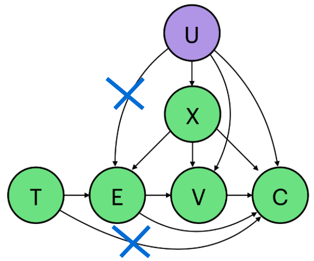
Conversions

(both heavily up-sampled)

Source: Diemert et al, AdKDD'18

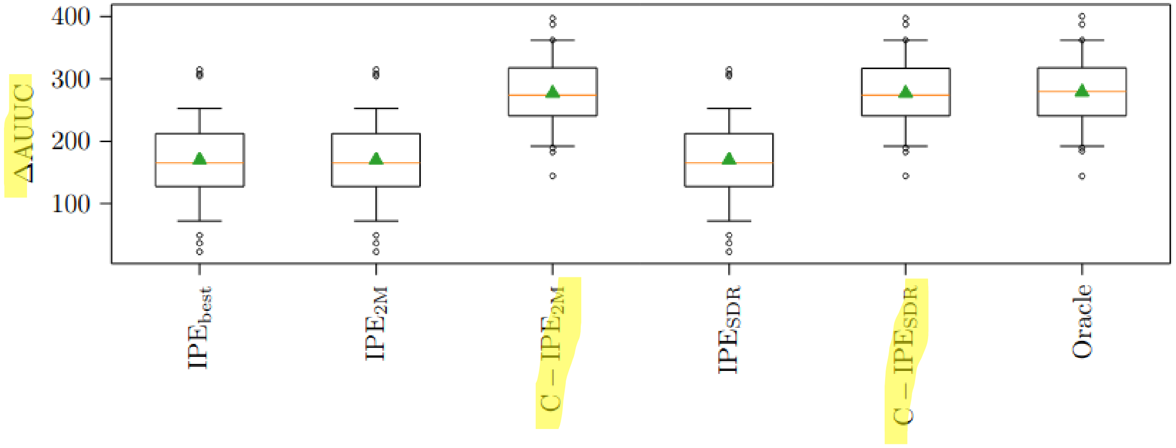
Solutions for noise

Subtitle



- "Zoom in" on the signal
  - Under some additional assumptions on the causal structure:
  - $P(C = 1 | x, T = 1) = [P(C = 1 | x, E = 1) - P(C = 1 | x, E = 0)] \times P(E = 1 | x, T = 1) + P(C = 1 | x, E = 0)$
  - conversion = post-exposure uplift x exposition prob. + "organic" conversion

- Can rewrite "treatment" (causal) uplift as a function of post-exposure uplift
- (+) stronger signal
- (+) compatible with existing models
- (-) assumptions not always verified



Source: Rahier et al, KDD'21

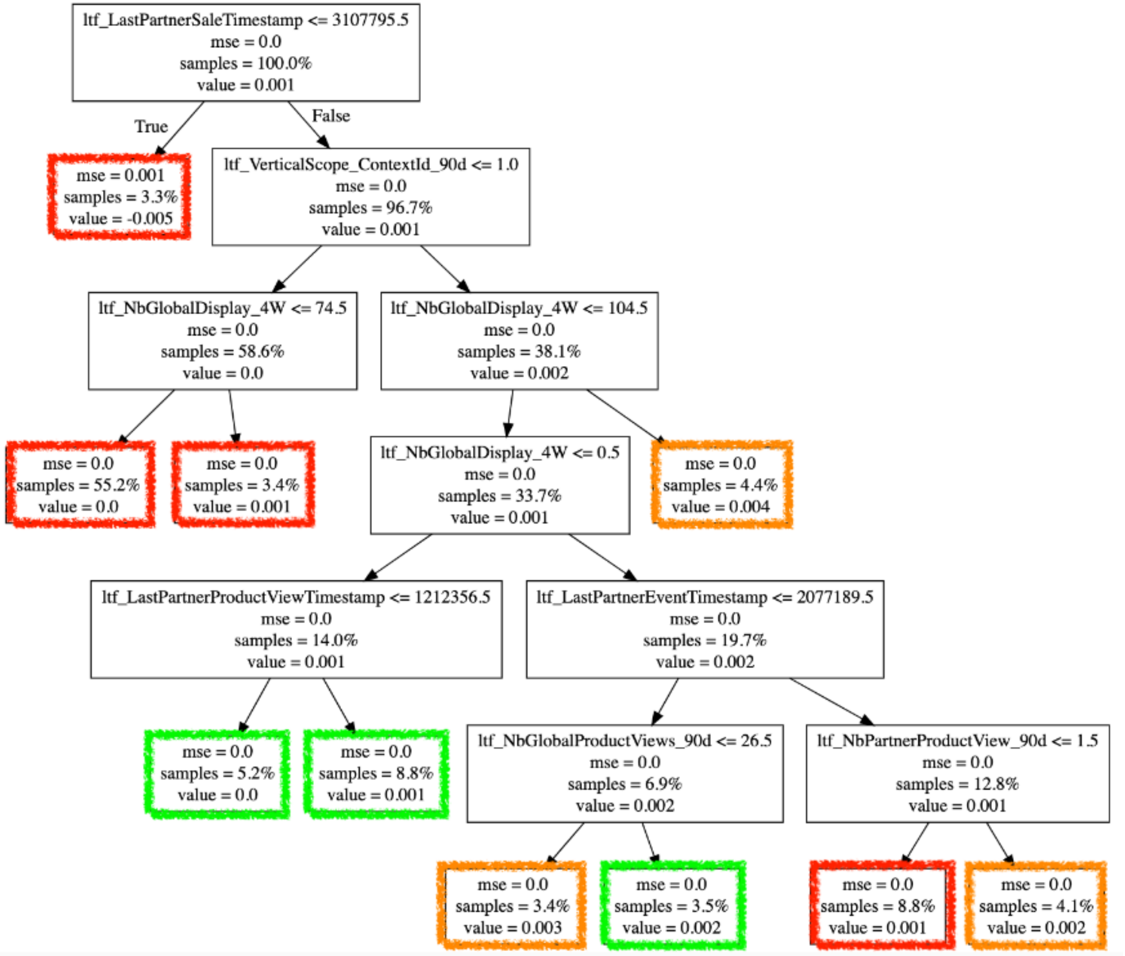
Solutions for noise (2)

Use simple, heavily regularized models: Trees

- Target can be:
  - Reverse label (CVT)
  - Predicted uplift (from another model)
  - Post-exposure Uplift
- Prediction is average uplift within leaf

- (+) (relatively) robust to noise
- (+) useful in practice
- (-) not sure we can do better in difficult cases

- Probably other targets are possible



Source: unpublished work

Challenge #2: Privacy constraints

How to learn uplift models from aggregated, differentially private data ?

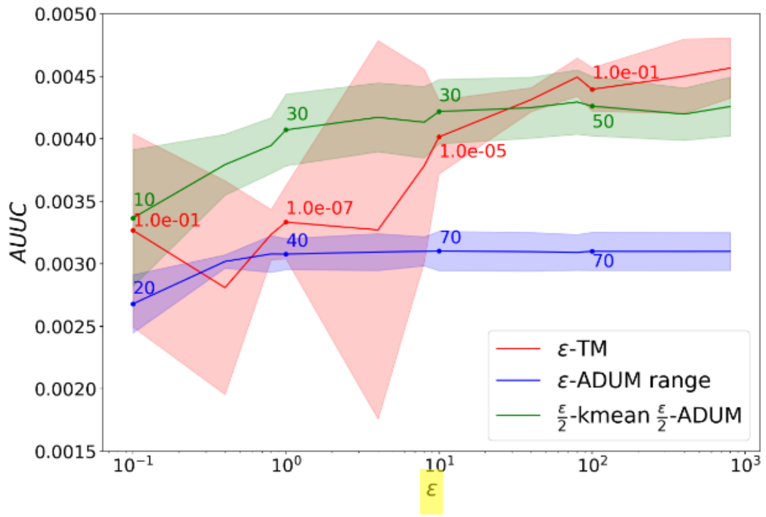
- Setup
  - Labels and features are aggregated
  - Summary statistics (count, sum) are noised with differential privacy
  - ε - DP := $\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S],$
- Proposal
  - Learn a piece-wise constant, ε-DP model

- (+) works better than ε-DP protected gradient methods like ε-2Models
- (+) can be combined with ε-DP k-means
- (+) we have theoretical guarantees (PEHE bounds)
- (-) performance varies based on strata definition

Algorithm 1 ϵ-ADUM

function TRAIN($(x_i, t_i, y_i)_{i\in[1,n]}, \pi \in \Pi_p(\mathcal{K}), D_y > 0, \epsilon > 0$):
    for $k \in [1, p]$ **do**
        for $t \in \{0, 1\}$ **do**
            $E_{k,t} = (y_i \mid \pi(x_i) = k,\ t_i = t)$
            $C_{k,t} = \mathrm{COUNT}(E_{k,t}) + \mathrm{Lap}(\frac{2}{\epsilon})$
            $S_{k,t} = \mathrm{SUM}(E_{k,t}) + \mathrm{Lap}(2\frac{D_y}{\epsilon})$
            $\widehat{y}_{k,t} = \frac{S_{k,t}}{C_{k,t}}$
        end for
        $\widehat{u}_k = \widehat{y}_{k,1} - \widehat{y}_{k,0}$
    end for
    return $(\widehat{u}_k)_{k\in[1,p]}$
end function

function PREDICT($x_{new} \in \mathcal{K}$):
    return $\widehat{u}_{\pi(x_{new})}$
end function



Source: Betlei et al, PPML'21

Beyond Uplift Modeling

How to use uplift predictions ?

Setup:

- A given advertiser has a fixed budget
- A typical "value based bidder" bids b(x) =~ P(C=1|T=1,Click=1,X=x)
- Assume U(X) predictions available
- How to "act" with this additional information ? i.e. how to change our bids ?
- Note: we don't control the order of exposure opportunities

Advertising is a dynamic budget allocation problem

First idea:

- if U^(x) <= 0 then bid 0; bid as usual otherwise
- (-) predictions are noisy and very small on average – we make errors
- (-) you can buy lots of cheap, sometimes useful inventory when bidding just above 0 – we loose opportunities
- (-) re-investing saved budget on non-responders is not always useful – we over-expose some groups

Source:

How to use uplift predictions ? (2)

Important remarks:

- U(X) = E[C=1|X, do(Bid=prod)] - E[C=1|X, do(Bid=0)] : this is a "prod vs nothing" uplift
- U(X) is not predictive (in theory) of uplift when varying the bid level:
    - U(X) <> E[C=1|X, do(Bid=b)] - E[C=1|X, do(Bid=prod)]

Also, we can assume:

- Spend is convex wrt to Bid (with some jumps – increased bids make you win more auctions)
- Uplift is concave wrt to Bid (diminishing returns)

So we need to devise a much more complex system than just ranking individuals by uplift !

Source:

How to use uplift predictions ? (3)

In theory we should solve that with RL...

But exploration is utterly costly !

Second idea:

- Work at population level (by strata / leaves)
- if uplift is high and exposure is low => over-bid (wrt to prod)
- Elif uplift is low or negative => under-bid (wrt to prod)
- Else => bid as usual
- Equalize costs between under- and over-bid

- (-) a bit ad-hoc... need a solver to equalize costs
- (+) approximately robust predictions: U(x, do(Bid=b)) =~ U(x, do(Bid=b')), when b' close to b
- (+) exposing more where current exposure is low
- (+) counter-factual techniques can predict effect of changes *within uplift prediction strata*
- (++) excellent practical results: up to 3x more incremental sales vs prod in AB test !

Source: unpublished work

References

- **Individual Treatment Prescription Effect Estimation in a Low Compliance Setting** - Thibaud Rahier, Amélie Héliou, Matthieu Martin, Christophe Renaudin and Eustache Diemert – KDD'21

- **Uplift Modeling with Generalization Guarantees** - Artem Betlei, Eustache Diemert, Massih-Reza Amini – KDD'21

- **Differentially Private Individual Treatment Effect Estimation from Aggregated Data** - Artem Betlei, Théophane Gregoir, Thibaud Rahier, Aloïs Bissuel, Eustache Diemert, Massih-Reza Amini – PPML'21

- **Uplift prediction with dependent feature representation in imbalanced treatment and control conditions** - Artem Betlei, Eustache Diemert, and Massih-Reza Amini – ICONIP 2018

Source:

CRITEO

Come see us at the Criteo AI Lab booth :)

Thank you!