Evaluating and Optimizing Online Advertising: Forget the click, but there are good proxies

Brian Dalessandro, Rod Hook, Claudia Perlich
m6d research

Foster Provost
NYU/Stern School of Business and m6d research

A main goal of online display advertising is to drive purchases (etc.) following ad engagement. However, there often are too few purchase conversions for campaign evaluation and optimization, due to low conversion rates, cold start periods, and long purchase cycles (e.g., with brand advertising). This paper presents results across dozens of experiments within individual online display advertising campaigns, each comparing different "proxies" for measuring success. Measuring success is critical both for evaluating and comparing different targeting strategies, and for designing and optimizing the strategies in the first place (for example, via predictive modeling). Proxies are necessary because data on the actual goals of advertising (e.g., purchasing, increased brand affinity, etc.) often are scarce, missing, or fundamentally difficult or impossible to observe. The paper presents bad news and good news. The most commonly cited and used proxy for success is a click on an advertisement. The bad news is that across a large number of campaigns, clicks are not good proxies for evaluation nor for optimization: buyers do not resemble clickers. The good news is that an alternative sort of proxy performs remarkably well: observed visits to the brand's website. Specifically, predictive models built based on brand site visits do a remarkably good job of predicting which browsers will purchase. The practical bottom line: evaluating campaigns and optimizing based on clicks seems wrongheaded; however, there is an easy and attractive alternative—use a well-chosen site visit proxy instead.

1. Introduction

One of the grand promises of online advertising is that measurement and optimization can be conducted much more easily than through many traditional advertising channels, because the targeting is embedded in large-scale, real-time information systems. Despite these promises, measurement and optimization can still be particularly challenging, owing to the complicated structure of the online advertising ecosystem, and to the fact that too often the actions worth measuring are in short supply or not measureable at all. This paper focuses on measurement and optimization within online display advertising (rather than search advertising), which recently has been seeing renewed interest both (a) because of the consolidation (via exchanges and real-time bidding systems) of the previously extremely fragmented display-ad targeting and delivery ecosystem, and (b) because browsers are spending increasing amounts of work and leisure time online, conducting activities other than searching and shopping.

Actually measuring the success of a display advertising campaign is stymied by several factors. Conducting a campaign online does not change the fact that the ultimate goal of advertising may be difficult or impossible to observe, as for instance with increased brand recognition or brand affinity. In this paper, we will focus on campaigns where the goal is a relatively short-horizon, post-view "conversion", such as a purchase within a 7-day time horizon after having been shown an ad. At the end we will also discuss why the conclusions ought to generalize to brand advertising as well. Even such short-term conversions may be difficult to observe because (i) they are not online, (ii) because the complicated ecosystem is not instrumented well enough to allow the conversions to be tracked effectively (e.g., cookies are being deleted before the conversion even occurs), and (iii) simply because conversions are so infrequent that making statistically valid conclusions is not possible. Beyond measurement and evaluation, the attempt of campaign optimization[1] and sophisticated

[1] The distinction between evaluation and optimization is not as relevant as it seems. Optimization is conceptually just a long sequence of evaluation steps selecting the better over the worse.

targeting further increases the necessary amount of conversions and (iv) is exacerbated in the beginning of the campaign where simply no conversions are available (the "cold start" problem). The former condition (iii) applies to many products such as cars, cruises, etc., that are the subject of online advertising. Condition (iv) applies to every new campaign.

In all these cases, all players in the ecosystem (who are genuinely interested in improving the effectiveness of the advertising) would benefit from an observable and common "proxy" for the ultimate effectiveness of a campaign. Currently, the most common such proxy is *clicks* on advertisements. Campaigns are both evaluated based on "click through rate" (CTR) and as a result also are optimized towards increased CTR. Indeed, the vast majority of academic research papers on improving the effectiveness of online advertising focus on CTR as the measure of effectiveness. We show, across a wide variety of campaigns, that this is wrongheaded: Clicks are a poor proxy for conversions—and are no better than randomly guessing in a surprisingly large percentage of campaigns.

However, clicks are not the only possible proxy, they are just the proverbial streetlight under which the drunk searches for his keys. In this paper we show through empirical analysis that some proxies are measurably better than others. In particular, well-chosen brand related actions—actions that indicate brand affinity or product interest, and are not purchases—make good proxies of the true metric of interest. In our analysis we define what it means to be a "good" or "better" proxy and we do a comparative study on how our different proxies serve to optimize a campaign towards more purchases. Although we only evaluate two specific types of proxy outcomes, we propose that in general a good proxy should satisfy the following criteria: they often are much more frequent than conversions; they already are observed throughout the ad ecosystem, and they can be argued to be well correlated with purchase behavior. Despite the reasons why one proxy may be better than another, an empirical study is often sufficient (and generally more practical) to make managerial decisions around proxy-based optimization and evaluation. This study also serves as an example on how to conduct such a study.

This paper presents empirical results across dozens of massive-scale experiments involving display advertising campaigns for major brands. As will be described in detail in Section 4, for each campaign we randomly target massive numbers of browsers with display ads and collect the data on which browsers subsequently made a purchase.[2] In addition, we also observe which browsers clicked on the ad, and which visited the chosen "brand action" webpages. Each experiment examines statistical predictive models trained for a given campaign in three different ways: using as the positive training instances (i) the browsers who purchase, (ii) the browsers who visited the chosen pages on the brand's website, and (iii) the browsers who click on the ad. This allow us to ask two related questions for each campaign, and to draw generalizations across the 58 campaigns:

Question 1: Are clicks a good proxy for evaluating and optimizing online display advertising campaigns where the ultimate goal is purchase?

The results provide a clear answer of no. Generally, clicking does not correlate well with purchasing. This complements prior studies showing evidence that few clicks lead to actual product purchases [9, 1, 8] and that most clicks are generated by a very small subset of browsers [5].

More importantly, clicks are unsuitable as a criterion for designing and optimizing targeting strategies (i.e., 'finding the best browsers to show an ad to'). Targeting models built on clicks do a poor job of identifying browsers who later will purchase. This is very consistent with the results of a recent study [13]. In our study, generalizing across all the campaigns, the targeting

[2] In order to be able to perform the comparison we use campaigns for which conversions are available. These results may generalize to campaigns for which conversions are not available for one or more of the reasons discussed above.

performance of click-driven models is statistically indistinguishable from random guessing! But is it really the 'click' that is at fault? That question is much harder to answer and also relates to the generalizability of our results. Clearly, the correlation between the click and the purchase depends on the creative design. We suspect that we observe here the results of years of the bad habit of click optimization in the industry. Some creatives are clearly designed to entice the click (e.g., flashy, promising to win something, etc.)—at times at the expense of the brand message [19, 11]. After all, if evaluation is by CTR, the 'best' creative is the one that was clicked on the most, no matter by whom.

Question 2: Are site visits a good proxy for evaluating and optimizing online display advertising campaigns?

In contrast to clicks, our results show that site visits turn out generally to be good proxies for purchases. Specifically, site visits do remarkably well as the basis for building models to target browsers who will purchase subsequent to being shown the ad. A very interesting result in our analysis is that in many cases, site visits produce better models than when actually training on purchases. Our results show this to be mostly true in cases where few purchases are available. The fact that the impact of using a good proxy like site visits is more dramatic in cases where purchases are few and far between has very important strategic implications for campaigns advertising big ticket items (like vacations), high consideration purchases (like bank accounts, 401ks) or products where purchases are mostly offline. One limitation on the generalizability of our results is that our study does not include products with only offline purchases (such as automobiles, typical consumer packaged goods), and thus the claims we make vis-a-vis different proxy modeling strategies may not hold to such products.

In the remainder of the paper we discuss the display advertising setting in more detail, and then discuss why *empirical* assessment is necessary to judge how good they are as proxies.

After that, we describe in detail the data, the empirical setting, the design of the experiments, and the empirical results. Finally, the paper will discuss the generalizability and implications of these results more broadly.

2. Background: Display advertising, evaluation, and targeting

In recent years, online advertising has seen major growth in the use of display, or banner, ads for both direct response and branding styles of marketing. This growth has been facilitated by the automation of the ad buying process, where entities called ad exchanges connect ad sellers and ad buyers via real-time bidding (RTB) systems. This machine-automated ad environment enables sophisticated data exchange, audience targeting and ad optimization in a way that has little precedent in offline and more rudimentary online advertising systems. Display advertising is usually contracted via two types of business arrangements: 1) direct buys between an advertising brand and a publisher, 2) exchange-facilitated buys, where advertisers, or the agents thereof, purchase advertisements via ad exchanges who represent publishers. For this study, all experiments were set up via the latter arrangement [16]. The generalizability of the results of this study to the former arrangement should hold to the extent that advertisements and users targeted via the former are similar to those targeted via the latter.

In online display campaigns, it is usually possible to observe whether or not a user makes a purchase after an ad exposure, even when the path to purchase does not include a click. These observed, post-impression, click-less purchases are often called "Post-Views" in the industry, and can be of any time horizon (typically 7 days, but reaching from 1 hour to 90 days). Our analysis assumes that a purchase, whether following a click or not, and whether made online or not, is the major goal of the advertising campaign.

An analysis of post-view purchases usually warrants a discussion of attribution and ad effectiveness. These are extensive and important topics but largely out of the scope of this paper. We refer the interested reader to [4, 19, 14] for a thorough treatment. Nonetheless, one issue that is important to discuss at least briefly is the thorny question of whether to measure purchase rate or *incremental* purchase rate. When evaluating and optimizing online display advertising campaigns, different marketers have different goals. Often the *explicit* goal of the campaign, as described to the campaign targeters and traffickers, is some variant of maximizing the cost-adjusted purchase rate. For example, some campaigns explicitly evaluate and pay advertising companies using a so-called cost-per-acquisition (CPA) model, where the targeting/trafficking firm gets paid for each purchase following an ad impression. Since the targeting firm pays for the actual ad placements ("buying inventory"), and often gets to choose more or less expensive placements, the firm would like to maximize the purchases per dollar spent on buying inventory. In other campaigns, the payment model is cost per impression, or "CPM", for cost per mille—a thousand impressions. However, many savvy brands nevertheless evaluate the campaigns' purchase rates or the "effective CPA" or other similar measures. They will use these evaluations to decide which online advertising contracts to renew or to expand. Therefore, accurate *comparative* evaluation (between similar campaigns) is vital. Let's call this set of goals "the common scenario," for contrast with what follows.

Although the common scenario is by far the most typical evaluation and optimization scenario in online advertising today, there is an important criticism that must be considered in order for a discussion of evaluating and optimizing advertising to be complete. In particular: the common scenario does not consider whether the advertisement actually had any effect on the purchase rate. Perhaps the people targeted with advertisements were the people most likely to purchase the product anyway! Some marketers may prefer to evaluate and optimize campaigns for "incremental" purchases, in order that the advertising show a substantive return on investment. Let's call this the "incremental purchase scenario." In Section 5 we discuss that our results apply both to the common scenario and to the incremental purchase scenario. (The explanation requires some technical detail.)

All campaigns we examine were conducted via standard industry practices for large campaigns. For all campaigns examined the campaign parameters (e.g., the time frame within which to determine purchase, how much to bid, etc.) were determined in advance between the advertising company and the client advertiser, completely separate from any knowledge of this study. The one exception to standard industry practices is that for this paper's experiments browsers were targeted randomly rather than selectively, as described Section 4.

3. Purchases, proxies, and proxy modeling

Before presenting an empirical analysis of proxy modeling and optimization, we will first discuss in more detail the motivation as to why to use proxies in the first place. The key problem for evaluating and optimizing online display advertising is that ultimate conversions are scarce. The term 'ultimate conversion' covers sets of alternative tangible (online) brand- or product-related actions that the advertiser desires; primarily we are interested in purchasing a product online, but other ultimate conversions include registering an offline purchase, filing for a rebate on an offline purchase, registering oneself on a site, joining a loyalty club, down-loading a free version of a product, etc. For this paper, we will call these all "purchases," since "ultimate conversions" is awkward and "conversions" has various meanings in the industry (e.g., in some contexts a click on an ad is regarded as a conversion).

Figure 1 shows the purchase rate frequencies across the 58 campaigns in the experiments discussed below. In about half of the campaigns, purchase rates are less than one in a million, and none are more than one in a thousand. This distribution of conversion rates represents a wide variety of product categories in many industries. It is typical to serve at least several million impressions in a given campaign, but despite this high number, the low rate of purchase conversion results in very

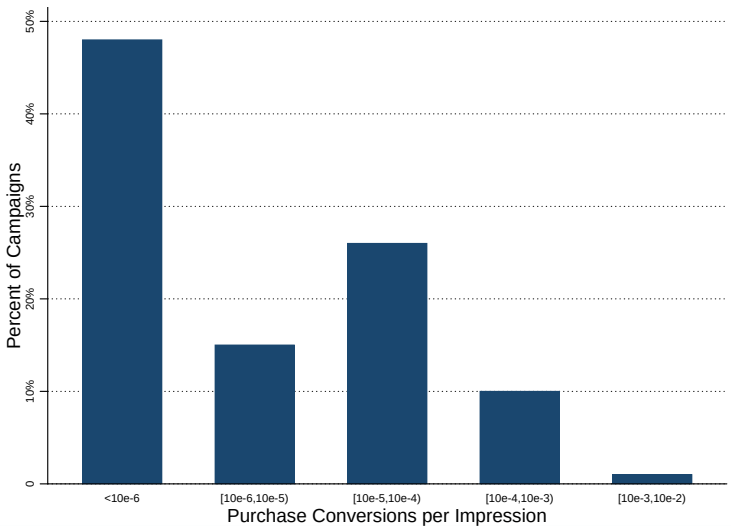

Figure 1 **Even for campaigns with relatively high purchase rates, the purchase rates are quite low. This chart shows the frequency of conversion rates across the 58 campaigns. Almost half of the campaigns see fewer than one purchase per one million impressions. Note that this targeting company is seen as being very good at delivering impressions, and has a high rate of renewal of contracts with its client marketers.**

few purchases in absolute terms. Such low event occurrence presents many challenges in model building and evaluation, and is often a primary driver motivating the development of conversion proxies.

The sparsity (and in many times complete absence) of purchase data has two implications: (i) it is unclear how to evaluate campaign effectiveness, and (ii) it is difficult to add data-driven modeling to campaign targeting. For evaluation, in many campaigns one would have to deliver millions of ads just to begin to get small numbers of purchases. Every time a brand wants to compare several different creatives or several competing targeting firms, millions or tens of millions of ads need to be shown for each. Even then, the numbers of positive cases may be small enough that the evaluation statistics have high variance, and so comparative conclusions cannot be drawn with confidence. Therefore, for many marketers establishing the effectiveness of an online campaign is a leap of faith.

The situation is much worse for optimizing campaign targeting via building data-driven predictive models. The modeling literature includes extensive discussions of the problems of building predictive models with rare target actions [21]; even there it is uncommon to see discussions of base rates as low as 10^{-5} or 10^{-6}. If a modeler were to want just 100 positive cases from which to model,[3] tens or hundreds of millions of ads would have to be shown.

As discussed in the introduction, the main idea of this paper is to assess different proxies for purchases. This assessment is based on the following key insight: If a surrogate action is indeed a good proxy for purchase, then a statistical model built to predict the proxy should do a good job of predicting purchase. In essence, the people who have been observed in the past to take the surrogate action look similar to people who purchase the product, with respect to the covariates used in the modeling. Thus, we will build predictive models to predict whether browsers will take the proxy action (e.g., clicking), and then evaluate whether those proxy-trained models do a good job of predicting purchase.

This line of reasoning introduces two preliminary questions. (Q1) Are purchases predictable in the first place? If predictive models trained using purchases, with large-enough training data sets,

[3] Negative cases—browsers who did not respond to an ad—are in abundance, and are usually downsampled

cannot predict purchases, then it may not be reasonable to expect that proxy-trained models would be able to predict purchases. If purchases can indeed be predicted, then we still might wonder: (Q2) Why might we believe that any surrogate action might be a good proxy for purchase?

We will present the results more formally after we introduce the experimental procedures below, but in summary the answer to Q1 is yes. With enough training data, purchases generally are predictable, where the degree of predictability is a function of the amount of training data and the covariates used to make predictions.

When we do not have sufficient purchases, why might surrogate actions be good proxies for the purpose of evaluating or modeling ad campaigns (Q2)? Clicks have been used as proxies for purchases for years; however, to our knowledge no principled justification previously has been made explicit. We make an empirical argument that clicks are in fact a poor proxy, and use the same empirical line of reasoning to show that a good proxy does indeed exist. While there may be both underlying social and/or statistical theory explaining why one proxy ought to be better than another, a full discussion of this theory is outside the scope of this paper on empirical generalizations. An area for further research is to understand how factors related to a consumer's underlying brand-affinity manifest themselves as observable outcomes (such as clicks, site visits and purchases). If correlations between brand-affinity and observable outcomes can be measured and established, then the strength of the correlation between the different outcomes and the brand-affinity might explain the quality of the proxy. If we accept an intuitive notion of correlated brand actions, we could further explain the strength of a proxy in classic statistical terms. A model built off of a proxy would induce bias (i.e., it is correlated to the target action but still is not exact) but its relative abundance would enable a lower-variance estimation [7]. The exact trade-off between the bias and the variance of a proxy is something that would not be established a priori, so even when supported by theory, the determination of whether a particular proxy is a good one is inherently an empirical question [15].

4. Experiments and Results

The results we have introduced above are based on 58 randomized experiments with different marketers.

4.1. Experimental Design

The focus of our empirical results is the question whether or not clicks and site visits respectively are good proxies for purchases. In particular, how well in terms of predicting the probability of purchases are models/campaigns performing when the optimization is done estimating the probability of either of the two proxies P where $y_P \in 0, 1$ is either a click or a site visit indicator. To build the predictive models that are the basis of the study, we also need features (independent variables) for each (anonymized) browser. The features can be grouped into three types $x \in x_{Basic}, x_{Extended}, x_{Full}$, which correspond to different approaches to predictive model-based targeted advertising.

$$Pr(y_P) = f_P(x) \tag{1}$$

The first two feature types can be thought of as "technographics"—the technical information revealed by the browser when interacting with the advertising system. The third represents preferences revealed by browsers via their webpage visits:

Basic: Use aggregate statistics of the cookies including age, number of webpages visited, number of checkins, connection speed, etc.

Extended = Basic + User Agent: Includes the basic features, plus browser "user agent" information, such as operating system (Vista, OS X, Linux, etc.), operating system version,

browser type (Safari, IE, Google Chrome, etc.), browser options installed, language preferences, etc.

Full = Basic + User Agent + Browsing History: In addition to the previous features we also include (anonymized) indicators of visits to commonly visited web sites. Web site visitation is widely believed to reveal user preferences, and to improve targeting [14, 17]. For the sake of keeping the results of the experiments general, we select the 100 most common websites in our data pool and not the full 50 Million websites that a specialized targeting company like m6d uses.

We estimate $\hat{f}$ [4] using logistic regression with L1 regularization [2]. Although this paper does not focus on how to build the best targeting models for each brand, we subsequently built and evaluated various alternative models/modeling techniques; these (less rigorous) empirical evaluations came to the same qualitative conclusions as the paper's main results.

We generate predictions of the models using 10-fold cross validation with pooled evaluation.[5]

$$\hat{Pr}(y_P) = \hat{f}(x) \tag{2}$$

and evaluate the predictions against purchase indicators: $L(y_{Purchase}, \hat{Pr}(y_P))$. We consider two ranking-focused measures that are most relevant for marketing: the area under the ROC curve (AUC), which measures how well the model ranks candidates for targeting and is equivalent to the Mann-Whitney-Wilcoxon statistic (and essentially to the Gini coefficient) [3, 10] and the lift at 20% of the population targeted, the ratio of the number of purchases at the top of the list ranked by the model to the number of purchases that would be expected to be there by chance [12] .

4.2. Data

Our empirical results are based on data generated from the day-to-day operations of a medium-sized online display advertising firm.

Sampling Campaigns

We initially selected 58 suitable campaigns from the set of all active campaigns during 2011—mostly based on duration/size and on having (comparatively) high purchase conversion rates (recall Figure 1). Both are necessary to ensure sufficient numbers of purchase events to derive statistically reliable conclusions. All campaigns operate under a 7-day view-through window and have no other specific properties (no specific geographic targeting, etc.). The campaigns cover well-known brands, across a wide variety of industries, including airlines, apparel, banking, car rental, consumer electronics, consumer services, credit cards, drugs and fitness, energy, home, hotels, insurance, manufacturing, mobile/cellular, movies/tv, non-profit, reservations, resorts, restaurants, shoes, software, telecommunications, and others. All campaigns in this set ran for between 1 and 5 months (some longer, but we only consider up to 5 months of data), using standard industry practices—specifically, bidding for the browsers in real-time bidding systems using the system parameters chosen beforehand (outside the context of this experiment) for each marketer.

[4] We use ˆ to denote predictions/estimates.

[5] The choice of cross validation is dictated by the occasionally low number of purchases. In this way we optimally utilize the full dataset while or each campaign, training and test data are always separated and each purchase is used exactly once in the evaluation of generalization (predictive) performance. With pooled evaluation, all the predictions across all 10 folds are pooled before the two ranking metrics are calculated; this allows evaluation even with fewer positive examples than folds.

Sampling Browsers/Cookies

The data are based on showing ads to stable and active browsers who had not previously been observed to have taken a brand-related action (click, purchase, site-visit). A stable browser is a browser whose cookie was created at least 2 days previously. An active browser is one that has been observed to appear in ad exchanges, generating bid requests and as a result allows the delivery of an ad impression. An example of a non-stable browser would be one where the cookies are cleared very often.[6]. The experiments in this paper were conducted as a supplement to the standard targeting done by the advertising firm. For each experiment, approximately 5000 randomly selected browsers (non-targeted) were served ads per day for a period of 1 to 5 months (the experiments were run concurrently with the targeted advertising campaigns and thus were subject to the same business cycles). They represent a random sample of 100,000 to 4,000,000 total browsers having seen an impression for the campaign. These browsers were removed from the population eligible for normal targeting. Ultimately, the events of interest are the ad impression to a specific browser and the potential engagement actions observed subsequently. However, to avoid double counting, we de-duplicate browsers and only consider the first impression event. As explained above, for each browser/impression we recorded three engagement outcomes: clicks on ads, post-impression purchases, and post-impression brand site visits. All predictor features are recorded as of the time of the impression.

4.3. Are Purchases at all Predictable?

To review the main idea of the experimental comparison, we will assess the predictions of click-trained models and site-visit-trained models to see whether they are effective at predicting which browsers will purchase. We can compare the predictive ability of these models to the predictive ability of models actually trained using purchases. It may be that the complicated nature of the online advertising ecosystem, and the vagaries of browser cookie behavior, introduce so much noise that browser purchase is not effectively predictable. So (i) to give us a baseline for our proxy performance comparison and (ii) to assess whether purchases are predictable in the first place, given the predictor features, we also conducted experiments in the more traditional setting of estimating and evaluating the predictive model on the same dependent variable: purchases.

Figure 2 shows that indeed it is possible to predict purchases using purchase-trained models when there are sufficient purchases for training. The figure plots each campaign's AUC as a function of the number of purchases available in the campaign's data. The average AUC across all 58 campaigns using the full feature set is 0.53 and is 0.55 on campaigns with at least 10 purchases. The average lift-at-20% is 1.15 across all campaigns and 1.3 on campaigns with at least 10 purchases. All 4 results are statistically significant at the 1% level using a t-test to compare to random targeting. Recall that a random model would have a lift of 1 and an AUC of 0.5.

Looking more deeply at the distribution of results shown in Figure 2, the average predictive performance is clearly affected by the very high variance when there are few purchases available (the left side of the chart). There are two sources for the high variance. First, having very few purchases in the testing data will yield high variance in the estimation of the AUC or lift. Second, and likely more critical, having few purchases for model training will yield overfitting and high variance in the resulting predictive models. Indeed we can see that as we move toward the right-hand side of the chart, the variance in the AUC decreases; the few campaigns with substantially more than 100 purchases all have AUC around 0.6.

Thus, the predictor features do provide information that allows for purchase predictions better than random. The statistically savvy reader might observe the close relationship between the

[6] To the extent that stable browsers are systematically different from non-stable browsers in behavior, this choice would limit the generalizability of our results to stable-browser users. We estimate the percentage of stable browsers to be between 60% (conservatively) and 90% (optimistically)

Figure 2 **Purchase-based model performance (AUC) as a function of the number of purchases in the training set using 10-Fold CV.**

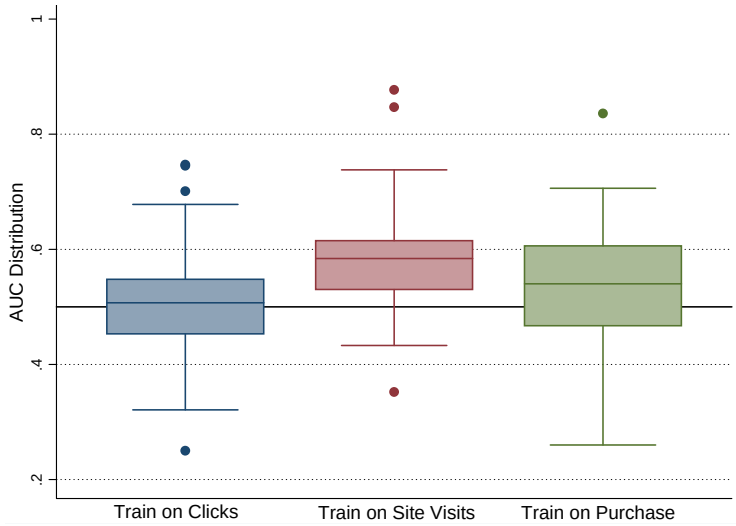

Figure 3 **AUC performance distribution in with respect to purchase prediction of the models trained on clicks, site visits and purchases respectively.**

proxy results (and motivation) and the classical bias-variance tradeoff[7]. Purchases are rare and create variance errors in the targeting models, whereas clicks and site visits are more common and therefore should yield smaller variance errors. However, the proxies are likely to be biased compared to purchases: there is no reason to believe that $f_{Purchase}(x) = f_{Click}(x) = f_{SiteVisit}(c)$. The tradeoff between the bias inherent in the proxy and the reduction in variance due to having more data will determine whether a proxy is a good proxy. Elaboration is beyond the score of this empirically focused paper; the interested reader is referred to an associated technical paper[15].

4.4. Comparing Different Proxies for Purchases

Now we are ready to discuss the main results in detail. Are clicks good proxies for purchases? Are site visits good proxies for purchases?

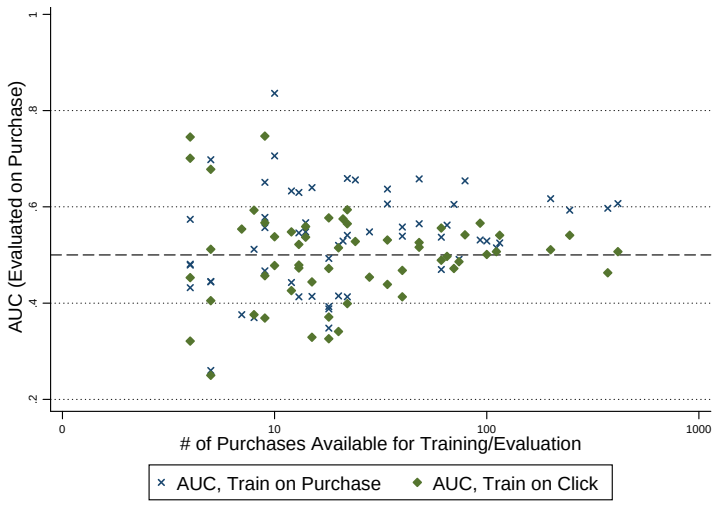

Figure 4 **Click- and Purchase-based model performance (AUC) as a function of the total number of purchases using cross-validation as described in the text.**

Figure 3 shows the distributions of the AUCs across the 58 experiments. Recall that each experiment assessed the ability of three different models (click-trained, site-visit-trained, and purchase-trained) to predict purchases. The right-hand box-whisker plot depicts the distribution of the AUCs for the purchase-trained models just discussed, shown in Figure 2. The line in the box shows the median AUC. The box delineates the interquartile range (IQR); the whiskers delineate the total range, excluding outliers, and the outliers (circles) are those points beyond 1.5·IQR from the IQR boundaries [20]. Despite their high variance, across the 58 experiments, as mentioned above the performance of the purchase-trained models is statistically significantly better than one would expect by chance (AUC$= 0.5$).

On the far left of the box-whisker plot is the distribution for models trained using clicks as a proxy for purchases (i.e., training on clicks, predicting purchases). Disturbingly for anyone using clicks to evaluate or optimize a campaign, the clicks distribution is centered just about exactly at what you would expect if targets were selected purely by chance, AUC$= 0.5$; the mean AUC is 0.49 and the median is 0.51. Click-trained models are statistically significantly worse predictors of purchases than purchase-trained models. More strikingly, click-trained models are not statistically significantly better than random targeting. In individual experiments, some of the click-models seem to be better than random, but when we look across experiments we see that an equal number of experiments show models that are worse than random! The better performances therefore may simply be due to the high variance (see below). Thus, in general, clicks cannot be viewed to be a good proxy for purchases.

Figure 4 shows the distribution of the 58 click-trained predictive performances overlaid on the purchase-trained performances shown above in Figure 2. We see that across range of numbers of purchases, the distribution of the click-model AUCs is centered around 0.5; when the evaluation variance is reduced by having a large number of purchases, the click-AUC distribution simply converges toward 0.5. This suggests that we should not interpret even a very large AUC in some particular campaign as being due to anything but chance.

Looking back at the box-whisker plot in Figure 3, in the middle we see the distribution of predictive performances for models trained using brand site-visits as a proxy for purchasing. Here we see two striking and encouraging trends. First, the distribution is shifted upward, even from the purchase-trained distribution. Site visits are significantly *better* for training purchase-prediction models than purchases themselves with a p-value of 0.001 in a paired t-test! Second, the variance

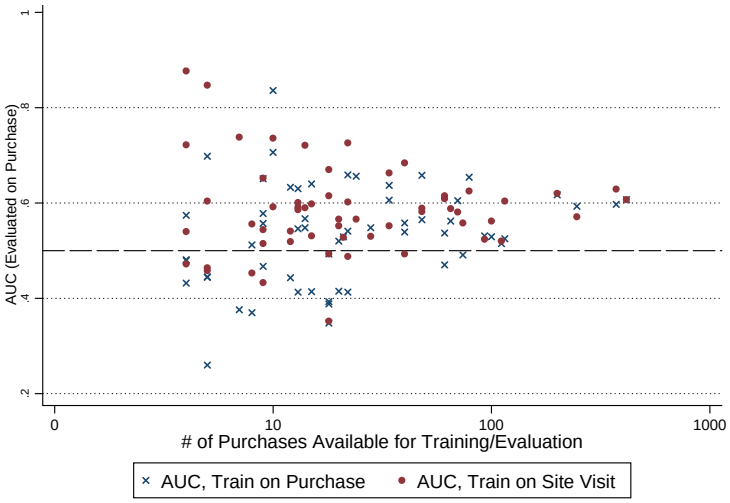

Figure 5 **Site VIsit and Purchase-based model performance (AUC) as a function of the number of purchases in the training set using cross-validation as described in the text.**

Feature Set	Click	SV	Purchase
Basic	0.84*	1.54*	1.00
Extended	0.80*	1.53*	1.25
Full	0.91*	1.61*	1.21

Table 1 **Average lifts in performance for predicting purchases, trained on the proxy stated in the column heading; every pairwise comparison against the purchase-trained model yielded a significant difference (*), (better or worse as appropriate, based on a paired t-test.**

in the distribution of the site-trained AUCs is substantially smaller than the variance in the distribution of purchase-trained AUCs. Using site visits as a proxy appears to be more consistent than training models using purchasing itself. In particular, it is much rarer for the model in some experiment to (seem to) be substantially worse than random.

Figure 5 helps to explain these results, overlaying the site-visit-trained AUCs on the purchase-trained AUCs from Figure 2. As shown in the box-whisker plot, almost all of the points are above the AUC= 0.5 (random targeting) line. However, we still see large variance when there are few purchases for *evaluation*. When we look to the far right of the figure, again the variance gradually reduces (as expected). Although there are too few points there to draw conclusions with statistical confidence, clearly there is no evidence that site-visit-training is better when there are a large number of purchases for both training and testing, which is intuitively satisfying. On the other hand, it is promising that purchase-training does not seem to be better than site-visit training either. This suggests that marketers may be well advised to use site-visits as a proxy by default: they work well when there are few purchases, and they seem to also work well even when there are large numbers of purchases. (In individual instances, marketers may run experiments to show that purchases actually perform substantially better, in which case they may override the default.)

Figure 6 compares the performance of the two different proxies (clicks and site visits) directly on the 58 different campaigns. Here it becomes even more obvious how much better the site visit proxy performs relative to the click-based models. Almost all campaigns are above the identity line, indicating that the click model is worse. Another relevant observation is how many models perform worse than random. In the case of clicks, a large proportion is left of the vertical 0.5 line, whereas only a few site visit models are below the horizontal 0.5 line.

In the remainder of this sectoin we will extend these results by taking a brief look at (i) the lift performances and (ii) the impact of using different feature sets.

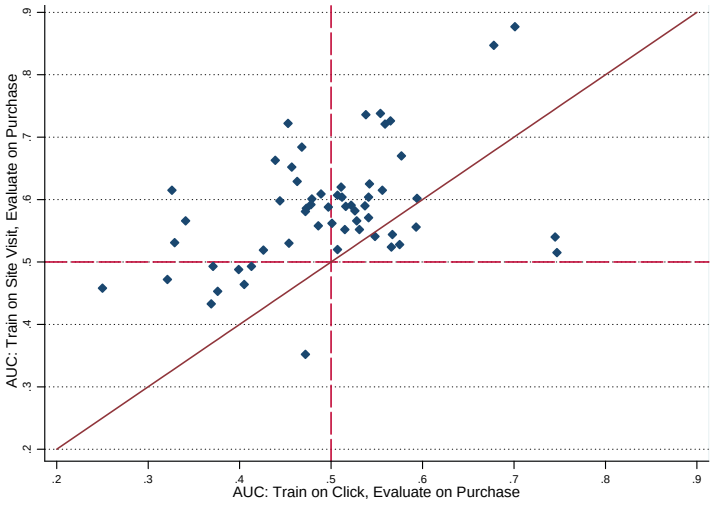

Figure 6 **AUC performance comparison between site visit models and click-based models using cross-validation as described in the text.**

The Lift results summarized in Table 1 are conceptually identical, if not more impressive than the AUC results from the previous figures. Again, clicks perform worse than both purchases and site visits, and on average clicks perform notably worse than random (lifts below 1). Site visits on the other hand perform better than both clicks and purchases, well above random.[7]

The feature set has some notable impact on the model performance. Overall, the user agent information does not seem to help any of the proxy models, but benefits the purchase-based model. The most interesting observation from our perspective is that adding the browsing history (even in this very limited form of only 100 most common websites) improves the performance for both proxy models (although not statistically significantly). We believe that the browsing information is predictive, but requires more positive instances to estimate a reliable model (larger feature sets introduce larger training variance).

5. Toward evaluating and optimizing for incremental purchases

The results we have presented in this paper apply to both the common scenario and the incremental purchase scenario, introduced and described in Section 2. Examining exactly why requires some technical details.

Let us first consider the common scenario. In the common scenario, evaluating campaigns requires the estimation of a quantity related to the likelihood of purchase conditioned on being shown an ad. For example, let's call the event of a purchase $y=1$ and the event of having seen an ad $a=1$. One may estimate $\mathbb{E}\left[Pr(y=1|a=1)\right]$—essentially the average purchase rate—for different ad targeters, different creatives, different subchannels (e.g., real-time bidding systems), different targeting strategies, etc. The cost-adjusted purchase rate (or profit) could be estimated, for example, as the expected profit: $\mathbb{E}\left[Pr(y=1|a=1)b(y=1)-c\right]$, where $b(y=1)$ is the benefit of a purchase and c is the cost of targeting (whether or not a purchase is made). Other formulations are possible; the critical factor is that any reasonable formulation will incorporate $\mathbb{E}\left[Pr(y=1|a=1)\right]$ or some derivative thereof. For optimization, one wants to target specific individuals who are more likely to convert,

[7] Please note that the lifts as well as the AUCs reported here should not be taken as representative of regular campaign performances by targeting companies such as m6d. In a typical campaign, only about 1% rather than 20% of browsers would be targeted—lifts for a good model would be much higher. However, lift at 1% is far too volatile to support generalizable results at the scale of these experiments.

based on the characteristics of the individuals. Let's say those characteristics are represented by vector x, then the optimization formulations will incorporate some variant of the conditional probability $\mathbb{E}[Pr(y=1|x,a=1)]$ (and the cost-adjusted formulations follow directly—e.g., target those for whom you expect the highest profit).

We showed above that site-visit trained models rank browsers well (and click-trained models do not) by their likelihood of purchasing (they have comparable areas under the ROC curve, or equivalently Mann-Whitney-Wilcoxon statistics) to purchase-trained models. Our results also hold for other related measures of likelihood-based scoring performance, e.g., the purchase lift when targeting the top-ranked individuals. This is not quite the same as producing accurate estimates of $\mathbb{E}[Pr(y=1|a=1)]$ or $\mathbb{E}[Pr(y=1|x,a=1)]$; it's not quite the same because the probability estimates likely will not be well calibrated [18] when the model is trained with a proxy. However, if even a small number of purchases are available, it is possible to recalibrate the scores to true estimates of $\mathbb{E}[Pr(y=1|a=1)]$ or $\mathbb{E}[Pr(y=1|x,a=1)]$ [6]. Even if no purchases are available, since most evaluations are comparative—caring not about the exact purchase rate but in how purchase rates compare across different possible choices—calibration may well be irrelevant practically. Comparing or optimizing based on site visits will yield the same final results as doing so using purchases.[8]

The foregoing analysis not only elucidates some technical details of using proxies for evaluation and optimization, it prepares us now to discuss the incremental purchase scenario.

In the incremental purchase scenario, the marketer's goal is not merely to have purchases follow ad impressions, but to generate more purchases (or more profit) than they would have had without the advertisements (generating "incremental" purchases). The basic case corresponds to augmenting the estimation from the common scenario above: we do not want simply to estimate $\mathbb{E}[Pr(y=1|a=1)]$ or $\mathbb{E}[Pr(y=1|x,a=1)]$ (for simplicity of presentation let's just consider the latter—the former is completely analogous). Instead, we want to estimate the change in the probability of purchase based on showing the ad:

$$\mathbb{E}[Pr(y=1|x,a=1)] - \mathbb{E}[Pr(y=1|x,a=0)] \tag{3}$$

where $a=0$ denotes not having been shown an ad. (The cost-adjusted and expected-profit cases are directly analogous.)

We can see now why the results of this paper apply to the incremental purchase scenario as well. For evaluation and optimization of incremental purchases (or profit), as illustrated by Equation 3 we need to estimate $\mathbb{E}[Pr(y=1|x,a=1)]$, just as in the common case. We show that site visits are a good proxy for this estimation. In the incremental case, we also need to estimate $\mathbb{E}[Pr(y=1|x,a=0)]$. However, estimating this quantity is almost identical to estimating $\mathbb{E}[Pr(y=1|x,a=1)]$—the estimation is simply done over a population or sample of browsers who have *not* been shown an ad. We have not presented empirical results here showing that site visits are a good proxy this case. Informally it makes intuitive sense that the results would generalize: if browsers who purchase after seeing an ad are quite similar to browsers who visit the brand's web site after seeing an ad, the same ought to be true for browsers who have not seen an ad. Separately, we have conducted experimental studies toward supporting this conclusion. They are not as rigorous as the studies performed for this paper, but generally show the same qualitative results.

The situation is more striking when we turn to the use of clicks as a proxy for purchase. Consider again Equation 3. We have shown that clicks are not a good proxy for the left-hand term. Clicks *cannot* help to estimate the right-hand term—if the browser was not shown an advertisement, then she cannot possibly click on it. (The estimate of $\mathbb{E}[Pr(y=1|x,a=0)]$ based on clicks would

[8] The main exception is for a sophisticated analysis where the costs differ significantly across choices. In practice, this often can be dealt with by comparing strategies or targeters based on a fixed targeting budget.

be zero.) Thus, clicks are by their very nature inapplicable as proxies in the incremental purchase scenario.

6. Conclusions and discussion of generalizability

This paper examined a set of 58 large-scale experiments assessing whether clicks and/or brand site visits are good proxies for purchases, for the purpose of evaluating and optimizing online display advertising campaigns. The results show that site visits are quite good proxies and that clicks are not.

Clicks not being good proxies has far-reaching implications. Technically, it means that clickers on ads in general do not resemble purchasers of the products/services being marketed. It is clearly not a good idea in general to use click-through to built or optimize targeting models for display advertising. More generally, if clickers do not resemble buyers, it is a dubious practice to use click-through to evaluate display advertising campaigns, or to conduct research on the effectiveness of display advertising campaigns!

Why might click-through be such a poor proxy? Presumably at least some clickers later buy? There are several possible reasons.[9] First, a common reaction to our presentation of these results to those outside of the industry is, "who clicks on online display ads anyway?" As discussed in the introduction, others have noted that many clicks are generated by a small number of browsers [5]. Second, industry insiders believe that click-fraud is rampant, which would tend to render clicks actually worse than random (since click-fraudsters generally do not buy). Third, there also may be insidious effects of the history of payment for campaigns based on click-through rate: it is possible that click-inducing creatives do indeed produce clicks, but from a different population from the true targets of the marketing.

The paper's results also show very good news. Site visitors *do* tend to look like purchasers. Models built to predict who will visit the brand's site after having seen an advertisement do a remarkably good job of predicting who will purchase. Site visits therefore should be considered for evaluating and optimizing display ads in cases where there are not enough purchases to do so reliably (or at all). We should note again that by necessity the experiments were conducted on campaigns where conversions were observed. However, examination of the distribution of results as a function of the number of purchases available shows no dependency on the number of purchases available—site visits seem to be good proxies even when very few purchases are available. This suggests (although our results do not show) that they also should be good proxies for many campaigns where purchases are not observed at all for one of the common reasons: (i) the purchases are not online, (ii) because the complicated ecosystem is not instrumented well enough to allow the purchases to be tracked effectively, (iii) simply because conversions are extremely infrequent, and (iv) because it is the beginning of the campaign where simply no conversions are available ("cold start").

Several aspects of the experiments suggest other specific dimensions of empirical generalization. The experiments are for quite varied brands/marketers. These marketers span all the industries that currently are heavily engaged in display advertising. Although all the campaigns have some purchases, they have widely different conversion rates, spanning several orders of magnitude. The campaigns have different bidding "strategies", which include in this case different parameters in the bidding systems (these were all selected by business agreements between the marketers and the advertising company, prior to and completely separate from this study). Taken as a whole, the experiments span all the top bidding systems currently in use in the display advertising ecosystem.

[9] Note that out of the 58 campaigns from the study—those specifically with larger numbers of purchases—in 42 campaigns there were in fact no purchases at all from clickers. It is possible that the lack of purchases from clickers is simply due to the very low purchase rates and click rates, even though the click rates are at least an order of magnitude higher than the purchase rates.

In addition, the experiments span all internet users with "active and stable" cookies (as we describe above).

It is not clear that there is sufficient evidence for a confident scientific conclusion that the results do *not* hold under some conditions. However, although it is clear that site-visit training is a good proxy across the span of numbers of purchases in the experiments, site-visit training does not seem to be *better* than purchase training when there are a very large number of purchases. This is suggestive at this point, because it is based on the four campaigns with hundreds of purchases (from which we cannot draw conclusions with statistical confidence). Nevertheless it makes sense, as described above, and should be verifiable with further experimentation with large campaigns. To summarize, the effect of site-visit training being better on average than purchase training is due to the poor performance of purchase training for campaigns with fewer purchases.

On individual experiments, click training does *seem* to "work" in certain cases. However, when we look across the experiments, we see no evidence that this seeming better performance is due to anything but chance. Looking at the chart depicting the number purchases vs. click-AUC, across the spectrum of numbers of purchases, the click AUCs seem to be well centered around 0.5, with a spread that most simply could be attributed to high variance due to the small number of purchases available for evaluation. As we get more and more purchases, the AUCs converge toward 0.5 in a well-behaved fashion.

Looking beyond the specific results presented in the paper, the general results raise the intriguing question as to whether site visits also would be a good proxy for *brand* advertising. Currently, online advertising is dominated by short-horizon conversion-driven advertising. However, the change in online consumer behavior—to be increasingly dominated by social and leisure activities—implies that the Web ought to be an increasingly attractive place for brand advertising with no short- or medium-term purchase outcome. We conjecture that site visits may be a good proxy for aspects of deeper brand affinity as well. If this is borne out by (future) research, for example by examining the correlation between online observable brand actions and traditional measures of brand affinity (e.g., via surveys), then proxies such as brand site visits could provide a systematic way to evaluate brand campaigns at a much lower cost and in real-time, which in turn could allow the optimization of brand campaigns using the same technology as for optimizing conversion-driven campaigns.

Finally, the general results suggest an important new area for practice and for research. What are good proxy brand actions? This paper was based the site visits chosen through the business practices of the advertising firm. We are aware of no published guidelines, let alone scientific research, helping marketers to understand what brand actions are good proxies and under what conditions.

Acknowledgements

Foster Provost thanks NEC for a Faculty Fellowship.

References

[1] Research shows link between online brand metrics and offline sales. NielsenWire, 2011. http://blog.nielsen.com/nielsenwire/consumer/research-shows-link-between-online-brand-metrics-and-offline-sales/.

[2] S. Balakrishnan and D. Madigan. Algorithms for sparse linear classifiers in the massive data setting. *Journal of Machine Learning Research*, 9:313–337, June 2008.

[3] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[4] B. Dalessandro, O. Stitelman, C. Perlich, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the Workshop on Data Mining and Audience Intelligence for Online Advertising at KDD*, 2012.

[5] J. Durham, E. Hunter, K. McCarthy, and G. Rogers. Natural born clickers. Presented at the 2008 iMedia Brand Summit.

[6] T. Fawcett and A. Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1):97–106, 2007.

[7] J. Friedman. On bias, variance, 0/1 – loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[8] G. Fulgoni and M. Moern. Whither the click?: How online advertising works. *Journal of Advertising Research*, 49(2), 2009.

[9] A. Ghose and S. Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622, Oct 2009.

[10] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic ROC curve. *Radiology*, 143(1):29–36, 1982.

[11] E. Hershberger, R. Lohtia, and N. Donthu. The impact of content and design elements on banner advertising click-through rates. *Journal of Advertising Research*, 43(4):410–418, 2003.

[12] G. Linoff and M. Berry. *Data mining techniques: for marketing, sales, and customer relationship management.* * Wiley Computer Publishing, 2011.

[13] S. Pandey, M. Aly, A. Bagherjeiran, A. Hatch, P. Ciccolo, A. Ratnaparkhi, and M. Zinkevich. Learning to target: What works for behavioral targeting. In *Proceedings of the 20th ACM International Conference on Information and knowledge management*, CIKM '11, pages 1805–1814, 2011.

[14] P. Papadimitriou, H. Garcia-Molina, P. Krishnamurthy, R. Lewis, and D. H. Reiley. Display advertising impact: Search lift and social influence. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[15] C. Perlich, B. Dalessandro, and F.Provost. Bias and variance considerations in proxy modeling for online display advertising. Technical report, m6d research, 2012.

[16] C. Perlich, B. Dalessandro, R. Hook, O.Stitelman, T. Raeder, and F. Provost. Bid optimizing and inventory scoring in targeted online advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[17] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge discovery and data mining*, 2009.

[18] R. Stein. Benchmarking default prediction models: Pitfalls and remedies in model validation. *Journal of Risk Model Validation*, 1(1), 2007.

[19] O. Stitelman, B. Dalesandro, C. Perlich, and F. Provost. Estimating the effect of online display advertising on browser conversion. In *Proceedings of the The 5th International Workshop on Data Mining and Audience Intelligence for Online Advertising at ACM SIGKDD*, 2011.

[20] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[21] G. M. Weiss. Mining with rarity: A unifying framework. *SIGKDD Explorations Newsletter*, 6(1):7–19, June 2004.