A Hard-ability "Theorem" on Measuring Advertising Effectiveness*

Randall A. Lewis
Yahoo! Research
ralewis@yahoo-inc.com

Justin M. Rao
Yahoo! Research
jmrao@yahoo-inc.com

October 7, 2011

Abstract

Twenty- ve display advertising eld experiments run at Yahoo!, amounting to over $2.8M worth of impressions, give insight into the volume of data needed to form reliable conclusions concerning advertising e ectiveness. Relatively speaking, individual-level sales are typically volatile, and only \small" impacts from advertising are required for a positive ROI. Using data from major U.S. retailers, we present a statistical argument to show the required sample size for a randomized experiment to generate su ciently informative con dence intervals for a given campaign is typically millions of individual users exposed to hundreds of thousands of dollars of advertising. The argument also shows that sources of heterogeneity bias unaccounted for by observational methods only need to explain a tiny fraction of the variation in sales to severely bias estimates. Measuring advertising e ectiveness is thus a situation with low-powered experiments and faulty observational methods | precisely where we would expect poorly calibrated beliefs in the market.

Keywords: *advertising effectiveness, field experiments, causal inference, electronic commerce*
JEL Codes: *L10, M37, C93*

*Working title. Alternate title: Advertising Effectiveness? Statistically Small Signals in a Sea of Noise or A "Near-impossibility Theorem..." We especially thank David Reiley for his contributions to this work. Ned Augenblick, Arun Chandrasekhar, Garrett Johnson, Clara Lewis, Preston McAfee and Michael Schwarz gave us valuable feedback as well. We also thank countless engineers, sales people and product managers at Yahoo! Inc.

1 Introduction

Whether we like to admit it or not, economists often operate under the assumption that \markets generally get it right," sometimes without regard for how di cult \getting it right" is. When a lot of money is at stake, the reasoning goes, it is natural that beliefs in the market are fairly well calibrated. Yet in some cases forming accurate beliefs is fundamentally di cult. In this paper we argue that measuring advertising e ectiveness (adfx) falls in this category. Perhaps tellingly, papers in this literature often a starting point of \Do ads have any e ect?" For example, in their important paper on adfx, Abraham et. al (1990) *open* with the line, \Until recently, believing in the e ectiveness of advertising and promotion was largely a matter of faith" | a rst sentence that might otherwise seem a bit peculiar given that before they penned it, approximately 4 trillion dollars had been spent on advertising.[1]

This paper uses 25 advertising eld experiments run at Yahoo!, which amount to over $2.8M in spending, to provide insight into the size of data needed to form reliable conclusions concerning the return on ad spending using randomized eld experiments and to show that only tiny amount of omitted selection or heterogeneity is required to generate severe bias in observational methods. Let's start with a simple observation: the e ect of ads should be \small" in equilibrium. Ads are relatively cheap, consumers see many ads each day, and only a small fraction of people need to be converted for a campaign to be pro table. As an example, one of the best known (and expensive) advertising venues in the United States is the Super Bowl. A 30-second Super Bowl television commercial costs between 1.5{2.5 cents per viewer.[2] So if a Super Bowl spot has an impact of 7 cents per viewer in pro t, it is wildly cost-e ective, while if it has an impact of 1 cent per viewer, it loses the company approximately $1M. The line between boom and bust is narrow, and we show in this paper that matters are further complicated by the fact that the standard deviation of sales, on the individual level, is typically 10 times the mean, making it di cult to reliably estimate per-capita e ects of a small magnitude even with a large amount of data.

With this in mind, we make a second observation: ads are *not delivered at random* because *firms do not pay marketing executives to randomly distribute ads*. So the true e ects should be relatively small while selection e ects due to timing (advertising when it is most e ective based on product launches and demand seasonality) and consumer targeting (advertising to consumers most prone to respond) can be quite large, especially if marketers are doing their jobs. So we have a case of a *relatively small* true e ect in a sea of *relatively large* selection and heterogeneity biases, which is unfortunate because the observational techniques of the standard economics toolkit were

[1] This figure ($4.6 trillion) encompasses total ad spending from 1919 through 1990 and is denominated in real 2005 US dollars. The ad data was taken from the Coen Structured Advertising Dataset, and GDP figures were taken from the US Bureau of Economic Analysis.

[2] This figure is not perfectly precise, but definitely in the ballpark. See for instance: http://money.cnn.com/2007/01/03/news/funny/superbowl_ads/index.htm. A 30-second Super Bowl TV spot is priced at $2.5M reaching an estimated audience approximately 90 million viewers according to Nielsen TV ratings.

designed for precisely the converse circumstance.

The 25 eld experiments used in this study involved large campaigns (all had over 500,000 unique users, most had well over 1,000,000) in which we randomly held out eligible users from receiving an advertiser's online display ad.[3] Sales tracking (both online and o ine, through data sharing agreements) allow us to estimate the underlying variability and trends in sales. Power calculations show that even an experiment that has test and control groups of over 500,000 people still cannot reliably detect any impact of advertising. Nine of the 25 experiments fall into this category. They do not possess that statistical power to reliably evaluate the (rather extreme) null hypothesis of a -100% ROI (no impact) of the campaign.

The remaining 16 experiments were powerful enough to reject the null hypothesis of -100% ROI with 90% power. However, even though it is di cult to do statistically, rejecting -100% ROI is hardly the goal of a marketing executive. Thus, we examine a more realistic set of hypotheses. Supposing the campaign was very e ective (+50% ROI), we ask how big the experiment would have to be to reject 0% ROI (one could similarly think of this as assuming 0% ROI and rejecting -50%). Only 3 of the 25 campaigns could reliably distinguish between these two disparate hypotheses (break-even vs. wildly successful). The median campaign would have to be nine times larger to have adequate power to evaluate this hypothesis set, which is perhaps surprising, since many of these underpowered campaigns already reached millions of users and cost hundreds of thousands of dollars. In fact, certain retailers with relatively high average sales and high standard deviation would have to run a campaign more than 100 times larger to have adequate power | in which case, the population of the United States would literally be a binding constraint. Other retailers would have bene ted from more modest increases in campaign size. Five campaign experiments would have had adequate power to test this hypothesis set had they been 2.3 times as large.

The power calculations that use an alternative hypothesis of a very successful campaign present an arti cially favorable view of the inference problem. Just as a marketing executive does not want to simply reject -100% ROI, she would not want to repeatedly run campaigns that have otherwise extraordinary returns of 50%. To optimize, the expected ROI of the last dollar spent should be set somewhere in the 5-10% range. The median standard error on ROI for the 25 campaigns is a staggering 51%. The median sales campaign would have to be $62\ times\ larger$ (mean 421x) to reliably distuinguish beween 10% and 0% ROI, two values that are considered vastly di erent when discussing a typical nancial investment, like returns to an asset on an exchange. For campaigns designed to acquire new account sign-ups, the situation is even worse; the median campaign would have 1241 times as large. For most sales campaigns, it would take a long-term, concerted e ort by the rm to ever hope to answer this sort of question; for account sign-ups, the problem goes even

[3] An example display ad is shown in the appendix. Unlike search ads, these involve creatives that are larger and include picture and potentially motion (Adobe Flash animation). They are typically paid per impression, as opposed to per click, to incentivize producing a high quality creative. In search advertising, link-based ads are text based, so the problem is lessened significantly and further mitigated by using "clickability" in adjusting the effective bid. For a more detailed exposition on price search adveritising, see Edelman, Ostrovsky and Schwarz (2007).

deeper, due to the \winner-take-all" feature of client aquisition for subscription services.

Experiments have the virtue of being unbiased, but as our arguments show, the data landscape is such that they can be underpowered in this setting, especially when taken in isolation. The temptation is to turn to observational methods, which can be applied to scores of historical campaigns. Using our estimates of sales volatility, we dispatch this approach with a concise statistical argument. We show that a tiny degree of model misspeci cation or heterogeneity/selection bias (ex. omitted variables in a correctly speci ed model) totally swamps the ability to accurately estimate adfx using observational methods. For example, consider a regression of sales volume per individual ($) on whether or not she saw advertising. The R^2 for a campaign with a positive rate of return is on the order of 0.000002 (this is not a typo, we are trying to estimate a 15-20 cent e ect on a variable with a mean of around $8 and a standard deviation about $80).[4] If we use an observational method to estimate this e ect, we have to make sure we have not omitted any variables that would generate an R^2 of this order or more. Since ads are, by design, not delivered randomly, this seems to be an impossible feat to accomplish. The very same features of the data that make experimental estimates noisy, render observational estimates unreliable.

Given the di culty in employing observational methods, adfx may be $fundamentally$ di cult to measure for media that cannot accommodate experimentation, because it is di cult to randomize exposure on the individual level. Examples of such media include out-of-home (billboards) and event sponsorship.[5] The alternative is to use geo-based randomization, as in Eastlack Jr. and Rao (1989), but this approach is hampered by e ectively small samples. But it's not all bad news. Individual-level experiments using formats such as internet, and to a lesser extent television, are not only possible, but getting easier. In addition, our statistical analysis implicitly imposes uniform priors. If a rm conducted a series of experiments, new campaigns could be evaluated with informative priors, which could tremendously improve estimate con dence. Overall, we argue strongly for the use of experiments to become the industry standard in measuring adfx and encourage the development of technology and methods to improve their statistical power. While experiments are not a magic bullet, our estimates indicate that reliable information would accumulate over time about the general e ectiveness of the advertising spend. Retrospective analysis of experiments, as done by Abraham et al. (1995) for TV commercials, can also be used to isolate (broadly) what works and what does not. Measuring adfx is not impossible, it is just hard; it is thus rather hopeful and perhaps naive to expect the advertising market to have well calibrated or precise beliefs on the e ectiveness of advertising campaigns.

In the discussion section, we use data from other industries to show that the rms we study are fairly representative of advertisers generally. We also address the concern that the our campaign

[4] $R^2 \quad \frac{.20^2 \cdot .5(1-.5)}{80^2} \quad .00000156$

[5] For example digital billboard could be changed often, but it would be (nearly) impossible to link differential exposure to the individual. Technological breakthroughs could help, such as linking automated toll RFID chips in cars to billboards.

experiments were too small. Our \Super Bowl `Impossibility' Theorem" bounds the set rms, by annual revenue, that can both a ord a Super Bowl ad and reliably detect ROI even if individual-level randomization for a Super Bowl ad was possible. These bands are tight to vanishing for realistic hypothesis sets such as 10% vs. 0% ROI. We also discuss the importance of incentives in an industry with (potentially) low evidentiary standards and examine an industry with a similar inference challenge facing market participants (vitamins and supplements).

The paper proceeds as follows. In Section 2 we present an overview of the statistical inference problem facing the advertiser and calibrate it with data from multiple experiments run at Yahoo!, in Section 3 we present a discussion of what these ndings mean for the advertising market, and in Section 4 we conclude.

2 The Statistical Problem

In this section we rst brie y discuss the magnitudes of campaign in uence that are needed for the campaign to be cost-e ective (note: throughout we will use the term \in uence" to refer to exerting a causal in uence on sales (exceeding zero sales impact, -100% ROI) and \cost-e ective" to refer to achieving a speci ed ROI target). We then present a simple statistical argument to show how detecting such magnitudes is di cult, even with a large amount of data, and relatedly, how a tiny degree of endogeneity or model misspeci cation in an observational method can lead to serious bias. Summary statistics from 25 large-scale advertising eld experiments run at Yahoo! allow us to calibrate the statistical argument using real-world data. These gures include sales volatility, advertising spend and intensity, required in uence for ROI=0% based on margin, size of experiment, and an estimate of the standard error on the ROI from the advertising spend. We use this information to conduct a power analysis of the campaigns and set out what one could reasonably expect to learn from an experiment of a given size.

2.1 Influence and Profitability

Advertising is ubiquitous in Western society. On a daily basis, the average American sees 25{45 minutes of television commercials (Wilbur, 2008), many billboards, and internet ads. Industry reports place annual media advertising revenue in the U.S. in the range of $173B,[6] or about $500 per American per year. So to break even, the universe of advertisers needs to net about $1.35 in pro ts per-person per-day. Given the margins of rms that advertise, our educated guess is that this roughly corresponds to about $4-6 in sales. It is an interesting exercise to think to oneself,

[6]This figure, while not perfect, is consistent with published market reports. We obtained it from http://www.plunkettresearch.com/ which aggregates a few reputable sources. In Appendix Figure 2, we use a another data source, the Coen Structured Advertising Dataset, to plot advertising spending since World War 1. During this period spending as a percent of GDP was fairly stable, 1.5–2%

\Do ads in uence me $5 per day?" Intuitively: ads are cheap, so the impact of each should small. We see many ads per day and \know" a only minority are relevant to us and impact our behavior.

When an advertiser enters this fray, it must compete with many rms for consumers' attention. As mentioned in the introduction, the cost per person of a typical campaign is quite low. A Super Bowl ad has a high price tag, but per person, the cost is only 1{2 cents. Online display ad campaigns with less reach still cost 1{2 cents per person per day, but typically run over a period of about two weeks, cumulating to a cost between 15 and 40 cents (supportive evidence can be found in Table 1).[7] The high-side estimate would be that an intense campaign captures about 2% of a targeted person's total attention to advertising in the campaign window. The relatively modest spend per person, in turn, makes it di cult to assess cost-e ectiveness. Further complicating matters is that individual-level sales are quite volatile for many advertisers. An extreme example is automobiles | the impact is either tens of thousands of dollars, or it is $0. While not as extreme, many other heavily advertised categories including consumer electronics, clothing and apparel, jewelry, and air travel also have volatile consumption patterns. Homogeneous food stu s have more stable expenditure, but their very homogeneity likely reduces own- rm returns to and equilibrium levels of advertising within industry as a result of positive advertising spillovers to competitor rms (Hummel et al., 2011).

In the following two subsections, we quantify how individual expenditure volatility impacts the power of adfx experiments and show that, in general, the signal-to-noise ratio is much lower than we typically encounter in economics.

2.2 Power, Data Size, and Endogeneity

Consider an outcome variable y (sales), an indicator variable x equal to 1 if the person was exposed, and an estimate $\hat{\beta}$, which gives the average di erence between the exposed (E) and unexposed (U) groups. In an experiment, exposure is exogenous. In an observational study, one would also condition on covariates W, which could include individual xed e ects, and the following notation would use $y\|W$. All the following results go through with the usual \conditional upon" caveat (see the appendix for this more general representation).

$$t_\beta = \frac{\bar{y}}{S.E.(\bar{y})} \tag{1}$$

where $\bar{y} \equiv \bar{y}_E - \bar{y}_U$, $\bar{y}_E = \sum_{i \in E} y_i$ and S.E.($\bar{y}$) $\equiv \sqrt{\frac{\hat{s}_E^2}{N_E} + \frac{\hat{s}_U^2}{N_U}}$. Assuming balanced exposed and unexposed samples, $N_E = N_U = N$, we obtain

[7] In terms of statistical power, the Super Bowl ads reach a much larger population–perhaps 15x the typical display ad campaign in our sample–but the longer-lived display ad campaign is analogous to running a lower-reach TV ad once a day over two weeks.

$$R^2 = \frac{\sum_i \left(\frac{1}{2} \quad y\right)^2}{\sum_i (y_i - y)^2} = \frac{\frac{1}{4}N \cdot (\quad y)^2}{2N \cdot s^2} = \frac{1}{2N}\left(\frac{y}{\sqrt{2}s/\sqrt{N}}\right)^2 = \frac{t^2}{2N}. \tag{2}$$

(1) is the standard formula for the t statistic. By substituting this into relationship (2), we get:

$$R^2 \approx \left(\frac{y}{s}\right)^2 \times \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \tag{3}$$

which links the change in the dependent variable due to exposure and its variance to R^2.

A natural application of this formula is to let y equal the expected impact of a campaign. We do so in the following example, which uses (approximate) median values from the 19 retail sales campaigns summarized in Tables 1 and 2. The hypothetical campaign goal, again calibrated from the experiments in Table 1, is a 5% increase in sales in the two weeks following the campaign. Based on data-sharing arrangements, we measure the weekly standard deviation in sales to be \$75 and average sales are a little more than \$7 during the 14 day campaign period. Spanning the range of discount to high-end multibrand retailers, the standard deviation of sales is about 10 times the mean on average, for the campaign period. Customers purchase goods relatively infrequently, but when they do, the purchases tend to be quite large relative to the mean. The campaign costs \$0.14 per customer, which amounted to delivering 20{100 display ads at \$1-\$5 CPM,[8]and the gross margin is assumed to be about 50%.[9] Five percent of sales amounts to \$0.35 per person. Hence, the goal was for the campaign to deliver a 25% ROI:

$$\frac{\$0.35 * 50\% - \$0.14}{\$0.14} = 25\%.$$

The estimation challenge facing the advertiser is to detect a \$0.35 di erence in sales between the treatment and control groups amid the noise of a \$75 standard deviation in sales. We can apply relationship (3):

$$R^2 = \frac{\$0.35^2}{\$75^2} \times \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = 0.0000054 \tag{4}$$

to show that the implied R^2 is 0.0000054 for a *successful campaign* | certainly not the material for scatterplots! A campaign with a *relatively large* internal rate of return, has an *exceedingly small* R^2, and thus requires a large N to identify any in uence at all, let alone the campaign's target level with any meaningful statistical signi cance. In this case, 2M unique users, evenly split between test and control in a fully randomized experiment would generate an expected t-stat of 3.30 (from relationship (2)). This corresponds to a test with power of about 95% at the 10% (5% one-sided)

[8]CPM is the standard for impression-based pricing for online display advertising. It stands for "cost per mille" or "cost per thousand" (M is the roman numeral for 1000).

[9]We base this on our conversations with retailers in the industry and our knowledge of the industry. It is not meant to be an exact figure. Note that if we set ROI=0%, this implies a gross margin of 40%.

signi cance level, as about 5% of the time, the t-stat will be less than 1.65. With 200,000 unique customers, the expected t-stat is 1.04, indicating the test is hopelessly underpowered to reliably detect any impact at all, failing to reject 74% of the time.

The true R^2 = 0.0000054 implied by the treatment variable x in a randomized trial implies that a small amount of endogeneity in an observational method, such as regression with controls, di erence-in-di erences, and propensity score matching, would severely bias estimates of adfx. Any omitted variable, misspeci ed functional form, or slight amount of endogenous exposure that would generate R^2 on the order of 0.00001 is a $full\ order\ of\ magnitude$ larger than the true treatment e ect. Compare this to a classic economic example such as the Mincer wage/schooling regression, in which the endogeneity is on the order of 1/8 the treatment e ect (Card, 1999). For observational studies, it is always important to ask, \What is the partial R^2 of the treatment variable?" If it is very small, as in the case of adfx, clean identi cation becomes more important, as a small amount of bias has a relatively large impact on the coe cient estimates. In Appendix Figure 3, we show a scatterplot for the above example; its resemblance to complete noise explains why you never see scatterplots in adfx papers.

The example shows that minute amounts of endogeneity can seriously bias estimates; even in the ideal experimental case, the volatility of sales necessitates large samples. Table 1 gives an overview of 25 experiments and shows that this example is representative of large display advertising campaigns. We augment the information in Table 1 with detailed estimation statistics in Table 2. Tables 1 and 2 are the centerpiece of the contribution of this paper. The individual experiments are taken from a number of papers from Yahoo! Labs: Lewis and Reiley (2010); Lewis and Schreiner (2010); Johnson, Lewis, and Reiley (2011); and Lewis, Rao, and Reiley (2011). We express sincere gratitude to the non-overlapping authors and encourage readers to examine these papers in more detail.

Table 1: Overview of the 25 Advertising Field Experiments

Retailers: In-Store + Online Sales*

		Estimation Strategies Employed**					Campaign Level Summary					Per Customer	
Adv	Year	#	Y	X	Y&X	W	Days	Cost	Assignment		Exposed		Avg. Sales	σ sales
									Test	Control	Test	Control	(Control)	
R 1	2007	1	1,4	1	-	1,2,3	14	$128,750	1,257,756	300,000	814,052	-	$9.49	$94.28
R 1	2007	2	1,4	1	-	1,2,3	10	$40,234	1,257,756	300,000	686,878	-	$10.50	$111.15
R 1	2007	3	1,4	1	-	1,2,3	10	$68,398	1,257,756	300,000	801,174	-	$4.86	$69.98
R 1	2008	1–6	1,4	1,2,3	-	1,2,3	105	$260,000	957,706	300,000	764,235	238,904	$125.74	$490.28
R 1	2010	1	1,4	1,2	-	1,2,3,4	7	$81,433	2,535,491	300,000	1,159,100	-	$11.47	$111.37
R 1	2010	2–3	1,3,4	1,2,3,4	1	1,2,3	14	$150,000	2,175,855	1,087,924	1,212,042	604,789	$17.62	$132.15
R 2	2009	1a	1,5	1	-	-	35	$191,750	3,145,790	3,146,420	2,229,959	-	$30.77	$147.37
R 2	2009	1b	1,5	1	-	-	35	$191,750	3,146,347	3,146,420	2,258,672	-	$30.77	$147.37
R 2	2009	1c	1,5	1	-	-	35	$191,750	3,145,996	3,146,420	2,245,196	-	$30.77	$147.37
R 3	2010	1	1,3,4	1,2,3	1	1,3	3	$9,964	281,802	161,163	281,802	161,163	$1.27	$18.46
R 3	2010	2	1,3,4	1,2	1	1,3	4	$16,549	483,015	277,751	424,380	-	$1.08	$14.73
R 3	2010	3	1,3,4	1,2,3	1	1,3	2	$25,571	292,459	169,024	292,459	169,024	$1.89	$18.89
R 3	2010	4	1,3,4	1,2,3	1	1,3	3	$18,234	311,566	179,709	311,566	179,709	$1.29	$16.27
R 3	2010	5	1,3,4	1,2	1	1,3	3	$18,042	259,903	452,983	259,903	-	$1.75	$18.60
R 3	2010	6	1,3,4	1,2,3	1	1,3	4	$27,342	355,474	204,034	355,474	204,034	$2.64	$21.60
R 3	2010	7	1,3,4	1,2,3	1	1,3	2	$33,840	314,318	182,223	314,318	182,223	$0.59	$9.77
R 4	2010	1	1,3,4	1,2	1	1	18	$90,000	1,075,828	1,075,827	693,459	-	$0.56	$12.65
R 5	2010	1	1,5	1,2	-	1,3	41	$180,000	2,321,606	244,432	1,583,991	-	$54.77	$170.41
R 5	2011	1	1,3,4	1,2	1	1,3	32	$180,000	600,058	3,555,971	457,968	-	$8.48	$70.20

Financial Services: New Accounts Online Only***

		Estimation Strategies Employed**					Campaign Level Summary						Per Customer	
Adv	Year	#	Y	X	Y&X	W	Days	Cost	Assignment			New	Pr New	SE New
									Test	Control	Exposed	Accts	(Test)	Acct
F 1	2008	1a	2,5	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	794,332	867	0.0011	0.0330
F 1	2008	1b	2,5	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	748,730	762	0.0010	0.0319
F 1	2008	1c	2,5	1,2,4	-	3	42	$75,000	12% of Y!	52% of Y!	1,080,250	1,254	0.0012	0.0341
F 1	2008	1d	2,5	1,2,4	-	3	42	$75,000	12% of Y!	52% of Y!	1,101,638	1,304	0.0012	0.0344
F 2	2009	1	2,3	1,2,3,4	1,2	3	42	$612,693	90% of Y!	10% of Y!	17943572	10,263	0.0006	0.0239
F 2	2011	1	2,5	1,2	-	4	36	$85,942	8,125,910	8,125,909	793,042	1090	0.0014	0.0331

* These retailers do a supermajority of sales via their brick & mortar stores.

** Estimation strategies employed to obtain the standard errors of the ad impact between the test and control groups follow:
"Y" 1:Sales, 2:Sign-ups, 3:Daily, 4:Weekly, 5:Total Campaign Window;
"X" 1:Randomized Control, 2:Active on Y! Network or site where ads were shown 3: Placebo Campaign for Control Group, 4: Multiple Treatments;
"Y&X" 1: Sales filtered post first exposure or first page view, 2: Outcome filtered based on post-exposure time window);
"W" (1: Lagged sales, 2: Demographics, 3: Online behaviors).

*** These financial services advertisers do a supermajority of their business online.

There is simply too much information in the tables to go through each column in detail, but we will try our best to walk the reader through the important features. Columns 1{3 of Table 1 give basic descriptors of the experiment. Based on our con dentiality agreements with advertisers, we are unable to reveal the names of each advertiser. We instead employ a naming convention that gives the sector and an advertiser number. Columns 4{7 outline the outcome measures. The rms in Panel 1 are retailers, such as large department stores. Based on these rms' objectives, sales is the key dependent measure. Column 4 gives the dependent measures and the unit of observation (\3" indicates daily observation, \4" indicates weekly). Panel 2 gives campaigns for online nancial service rms aiming to acquire online account sign-ups. Column 7 gives the control variables we have to reduce noise in the experimental estimates. In column 8, we see that the experiments ranged from 2 to 135 days, with a median of 14 days, which is typical of display campaigns. Column 9 shows the campaign cost varied from relatively small ($9,964) to quite large ($612,693). The mean was $114,083 and the median was a healthy $75,000. Overall, the campaigns represent over $2.8M in advertising spend.

Columns 8{11 show that the median campaign reached over 1M individuals, and all campaigns had hundreds of thousands of individuals in both the test and control cells. The per customer information in the 2nd and 3rd columns from the right show that the average sales per customer varied widely. This is driven by the popularity of the retailer and the targeting level of the campaign (some campaigns targeted existing customers, for instance). The median level of sales per person is $8.48 for the test period. The nal column gives the standard deviation of sales on an individual level. On average, the standard deviation is 9.83 times the mean | one of the two de ning features of the data that makes inference very di cult (the other is the relatively small in uence necessary for pro tability). Examining longer campaigns, we see the standard-deviation-to-mean ratio falls, which is due to negative serial correlation in sales (\if I bought last week, I am less likely to buy next week").

Table 2: Statistical Precision and Power Calculations for the 25 Advertising Field Experiments

In-Store + Online Sales

		Key Statistical Properties of Campaign					HARD: Ads did anything? H0: ROI=-100% Ha: ROI=0%		HARDER: Worth money? H0: ROI=0% Ha: ROI=50%		HARDEST: To optimize? H0: ROI=0% Ha: ROI=10%		CRAZY: Profits maximized? H0: ROI=0% Ha: ROI=5%	
Adv	#	SE β Sales	Radius 95% CI % Sales	Spend Per Exposed	Margin	SE ROI	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3
R 1	1	$ 0.193	4.0%	$0.16	50%	61%	1.64	3.3x	0.82	13.4x	0.16	335x	0.08	1338x
R 1	2	$ 0.226	4.2%	$0.06	50%	193%	0.52	33.5x	0.26	133.8x	0.05	3345x	0.03	13382x
R 1	3	$ 0.143	5.8%	$0.09	50%	84%	1.19	6.3x	0.60	25.2x	0.12	631x	0.06	2524x
R 1	1–6	$ 0.912	1.4%	$0.34	50%	134%	0.75	16.2x	0.37	64.7x	0.07	6939x	0.02	27756x
R 1	1	$ 0.244	4.2%	$0.04	50%	278%	0.36	69.4x	0.37	277.6x	0.07	425x	0.07	1700x
R 1	2–3	$ 0.207	2.3%	$0.12	50%	84%	1.20	6.3x	0.60	25.2x	0.12	629x	0.06	2515x
R 2	1a	$ 0.139	0.9%	$0.09	15%	24%	4.12	0.5x	2.06	2.1x	0.41	53x	0.21	212x
R 2	1b	$ 0.142	0.9%	$0.08	15%	25%	3.99	0.6x	2.00	2.3x	0.40	57x	0.20	226x
R 2	1c	$ 0.131	0.8%	$0.09	15%	23%	4.33	0.5x	2.17	1.9x	0.43	48x	0.22	192x
R 3	1	$ 0.061	9.5%	$0.04	30%	52%	1.92	2.4x	0.96	9.7x	0.19	243x	0.10	972x
R 3	2	$ 0.044	8.0%	$0.04	30%	34%	2.96	1.0x	1.48	4.1x	0.30	103x	0.15	411x
R 3	3	$ 0.065	6.7%	$0.09	30%	22%	4.50	0.4x	2.25	1.8x	0.45	44x	0.23	177x
R 3	4	$ 0.051	7.8%	$0.06	30%	26%	3.82	0.6x	1.91	2.5x	0.38	62x	0.19	247x
R 3	5	$ 0.049	5.5%	$0.07	30%	21%	4.73	0.4x	2.36	1.6x	0.47	40x	0.24	161x
R 3	6	$ 0.064	4.8%	$0.08	30%	25%	3.98	0.6x	1.99	2.3x	0.40	57x	0.20	227x
R 3	7	$ 0.032	10.6%	$0.11	30%	9%	11.32	0.1x	5.66	0.3x	1.13	7x	0.57	28x
R 4	1	$ 0.031	10.9%	$0.13	40%	10%	10.45	0.1x	5.22	0.3x	1.04	8x	0.52	33x
R 5	1	$ 0.215	0.8%	$0.11	30%	57%	1.76	2.9x	0.88	11.6x	0.18	291x	0.09	1165x
R 5	2	$ 0.190	4.4%	$0.39	30%	15%	6.90	0.2x	3.45	0.8x	0.69	19x	0.34	76x

New Accounts Only

		Key Statistical Properties of Campaign					HARD: Ads did anything? H0: ROI=-100% Ha: ROI=0%		HARDER: Worth money? H0: ROI=0% Ha: ROI=50%		HARDEST: To optimize? H0: ROI=0% Ha: ROI=10%		CRAZY: Profits maximized? H0: ROI=0% Ha: ROI=5%	
Adv	#	SE β New Accts	Radius 95% %New Accts	Spend Per Person	Lifetime Value	SE ROI	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3
F 1	1a	69	15.6%	$0.06	$1,000	138%	0.73	17.1x	0.36	68.3x	0.07	1707x	0.04	6828x
F 1	1b	69	17.7%	$0.07	$1,000	137%	0.73	17.0x	0.36	67.9x	0.07	1697x	0.04	6790x
F 1	1c	70	10.9%	$0.07	$1,000	93%	1.07	7.8x	0.54	31.4x	0.11	785x	0.05	3139x
F 1	1d	70	10.5%	$0.07	$1,000	93%	1.08	7.7x	0.54	30.9x	0.11	774x	0.05	3094x
F 2	1	288	5.5%	$0.03	$1,000	47%	2.13	2.0x	1.06	8.0x	0.21	199x	0.11	795x
F 2	1	46	8.3%	$0.02	$1,000	233%	0.43	48.7x	0.21	195.0x	0.04	4874x	0.02	19496x

In Table 2, we delve into this statistical problem more deeply. Column 3 gives the standard error associated with the estimate of β, the test-control sales di erence (in dollars for Panel 1, sign-ups for Panel 2). β is estimated by conditioning upon the control variables outlined in Column 7 of Table 1 in order to get the most precise estimate possible. In column 4, we convert this into the implied radius (+/- window) of the 95% con dence interval for the sales impact, in percentage terms. Column 5 gives the spend per person. These gures can be compared to the standard errors given in column 3 to get an idea for the relative level of noise in estimation. Even in these large experiments, the standard error of in uence $exceeds$ the per-person spend in 12 of the 19 cases given in Panel 1. Using our estimates of gross margins given in column 6, which are based on our conversations with advertisers and SEC lings, we calculate the standard error of the return on investment in column 7. The median standard error for ROI is a staggering 26.1%, and the mean is a rather haunting 61.8% | estimating ROI is far from precise, even with relatively large randomized experiments as the median con dence interval is about 100% wide.

In the nal 8 columns in Table 2, we present various sets of hypotheses the advertiser might be interested in evaluating. We colloquially refer to these as going from \hard" (estimating in uence to be signi cantly greater than 0) to \crazy" (reliably estimating the ROI to be near the cost of capital for the rm). For each hypothesis set, we give the expected t-stat relative to the null hypothesis given that the alternative hypothesis is true. We also give \data multiplier" (how much larger would N have to be) required to expect a t-stat of 3. As we noted earlier, an expected t-stat of 3 provides power of 91% with a one-sided test size of 5% (to see this, note that with an expected t-stat of 3, the distribution is symmetric around 3, meaning that about 9% of the distribution's mass will fall below 1.65 in the left tail). Put another way, the multiplier gives the multiple of N (and advertising expenditure) needed for each experiment to serve as a powerful test of the null hypothesis against the alternative.

We start at the \hard" level. Most papers on adfx end at this level as well, in that the main goal is to measure whether in uence can be measured to signi cantly exceed 0 (Bagwell, 2005). Column 9 gives the expected t for the null hypothesis that the ad had zero e ect (ROI=-100%), assuming that alternative hypothesis holds (the ad broke even, ROI=0%), given margins the break-even sales in uence is 2{7 times the the cost given in column 6 . We see that 9 of 25 experiments lacked su cient power to detect any in uence at all (E[t]$<$1.65). As an aside, we note that these experiments are not meant to represent optimal experimental design. Often the advertisers came to us looking to understand how much can be learned via experimentation, given a number of budgetary and campaign-objective constraints. Nearly half (10 of 25) of the experiments had E[t]$>$3, meaning they possessed more than enough power reliably evaluate if the ads had any in uence. These tests are performed in the multiple papers cited earlier and generally reveal a statistically signi cant impact of advertising (the papers also discuss features of advertising such as the impact of local targeting, interested readers are encouraged to consult these papers) (Lewis

and Reiley, 2010; Johnson et al., 2010; Lewis and Schreiner, 2010). Reliably rejecting zero influence is certainly possible.

While interesting to the psychologist and economist, documenting a non-zero impact of a campaign is not the goal of a marketing executive. For instance, a -50% impact is grossly unacceptable. In the \harder" column we ask a more appropriate question from a business perspective, \Are the ads worth it?" Here we set the null hypothesis as ROI=0% and the alternative to a figure that sometimes comes up in conversations with marketing executives, 50%. By using ROI=50% we are imagining a case in which the ads are expected to be highly effective. Intuitively, it is easier to do statistical estimation in this case, as compared to one in which the effects are hypothesized to be more modest. Column 11 shows that only 3 of 25 campaigns have E[t]$>$3 in this case. In fact, half have E[t]$<$1, which is consistent with the large standard errors on ROI shown in column 8.

The multipliers show that many retailers have sales sufficiently noisy that the experiment would have to be almost impossibly large to reliably answer the \worth the money" question. We find this startling, but it is not universal. Retailer 5's 2nd campaign cost $180,000 and reached 457,968 people. The $\frac{\sigma}{\mu}$ ratio was the standard 9.6. The campaign achieved statistical precision by having a very large control group, 3,505,971 people. This level of precision could be achieved for smaller campaigns by using new \ghost ad" technology, which allows advertisers to run experiments without the need to pay for control impressions. Lewis (2011) presents this technology and related experiments in more detail . We also note that the large experiments run by Retailer 4 and the experiment by Retailer 2 also had nice power. This is largely due the relatively small standard deviation of sales, which in turn is driven by having lower mean sales. In other words, a smaller retailer gets more statistical bang for their buck. Overall, 12 of 25 had E[t]$<$1 (severely underpowered), 4 had E[t]$\in$[1,2], 5 had E[t]$\in$[2,3] (90%$>$power$>$50%) and only 3 had E[t]$>$3. Therefore, of the 25 campaigns, only 3 had sufficient power to reliably conclude that a $very\ cost\text{-}effective$ campaign was worth it, and an additional 5 could reach this mark by increasing the size of the experiment by a factor of about two (those with E[t]$\in$[2,3])

Using an alternative hypothesis of a 50% ROI makes the power calculations artificially strong, or put another way, hypothesis testing is made easier by the stark alternative hypothesis. The \harder" analysis shows that if it were indeed the case that a firm's campaigns had a 50% ROI, it would still be difficult to reject 0% ROI. Running a 0% ROI campaign is not optimal for the firm, because it has a non-zero cost of capital. Similarly, repeatedly running 50% ROI campaigns is not optimal, as the firm ought to advertise more in equilibrium, given the substantial profits it is reaping. In the 3rd and 4th columns from the right, we use an alternative hypothesis closer to that of an optimizing firm, 10% ROI. Strikingly, every experiment is severely underpowered to reject 0% ROI in favor of 10%. E[t]$<$0.5 for 21 of 25 campaigns and even the most powerful experiment (Retailer 3, Experiment 7) would have to be 7 times larger to have sufficient power to distuinguish this difference. Indeed, the median sales experiment would have to be a daunting $61\ times\ larger$ to

reliably detect the difference between an investment that, using conventional standards, would be considered a strong performer (10% ROI) and one that would be considered a dud (0% ROI). For new account sign-ups, the median multiplier is an almost comical 1241x. The statistical properties of sales and stiff competition for attention due to the large quantity of advertising a consumer faces conspire to make it near-impossible for a firm to reliably optimize its advertising spend, at least in the medium-run. In comparison to the noise in the returns to advertising, even the riskiest financial asset looks like a sure-thing.

In the final two columns of Table 2 we push the envelope further, setting the alternative hypothesis roughly to the profit maximizing level of the cost of capital, 5%. The E[t]'s and multipliers for E[t]=3 demonstrate that this is not a question an advertiser could reasonably hope to answer for a specific campaign or in the medium-run across campaigns | in a literal sense, the world's total population and the advertiser's annual advertising budget would be binding constraints for many cases! We think it is important to note that \hardest" and \crazy" are not straw men. These are the real standards we use in textbooks, teach our undergrads/MBA's, and employ for many investment decisions. The fact that it is nearly impossible to apply them here with any precision is one of the key contributions of the paper.

3 Discussion

In this section we discuss the practical implications and extensions of our findings, and comment on the general features of an industry in which mistaken and/or imprecise beliefs (priors) can easily persist.

3.1 Are these online campaigns anomalous?

The natural question to pose is, \How representative of advertising in general are the campaigns we have presented?" Perhaps these retailers and financial firms have abnormally volatile sales or maybe these campaigns are too small, and so on. To combat the sales volatility criticism, we use data from an industry that advertises very heavily and for which data is easily available: American automobile manufacturing (one out of every twelve TV commercial spots advertises pick-up trucks, accounting for roughly $9 billion annually, source: Kantar Media). Automakers turn out to face more volatility than the firms we study. To address the data size objection, we look at the relatively large advertising venture of running a 30-second Super Bowl commercial. Our \Super Bowl `Impossibility' Theorem" bounds the set of firms that could reliably estimate the impact of a Super Bowl ad even if exposure could be randomized (and sales linked to the randomization) on the individual level. The intuition is thats many of the firms large enough to afford a Super Bowl ad have baseline sales so high that making reliable ROI inference is essentially impossible. For small firms, inference is easy, but the Super Bowl ad is too expensive.

3.1.1 Automobile Sales

Advertising is a huge part of the automobile industry. The leading industry trade group[10] estimates that U.S. automakers spend $674 in advertising for each vehicle sold. We'll try to back-out sales volatility based on sales data. It is reasonable to suppose the average American purchases a new car every 5{10 years, but it is hard to get a rm estimate for this gure. We will generously assume it is every 5 years, generous insofar as higher purchase frequency helps the advertiser inference-wise.[11] Suppose that the advertiser has market share similar to many major automakers, 15%. Then the annual probability of purchase is 0.03 (Pr(buy) = .2*.15 = .03). This implies a standard error of $\sqrt{0.03} \approx \frac{1}{6}$.

On the cost side, we'll assume the national average of a new vehicle, $29,793.[12] Mean annual sales per person are thus μ=$893 and SE($\mu$)= 1/6*$29,793 = $4,700. This gives a $\frac{\sigma}{\mu}$ ratio of roughly 10, similar to our standard nding. However, this is $yearly$, as opposed to the ner granularity used in our study. To convert this gure to monthly we multiply by $(1/\sqrt{12})/(1/12) = \sqrt{12}$ or about 3.5.to get ratio of 20:1, which is double that of our median online display advertising experiment for retailers. What this means is that an automobile advertiser would have to run a year-long experiment to gain the same insights, all else equal, that our median advertiser would gain in a month. We think this example demonstrates that our advertisers have sales volatility well within the bounds of many major advertisers in the market and are within the proverbial bounds of \representative."

3.1.2 Super Bowl "Impossibility" Theorem

In this section we address the concern that these experiments are unrepresentative because they are too small. Our rst remark is that the average cost was over $100,000, not exactly small potatoes. Here we bolster our point through a thought experiment concerning one of the most expensive advertising venues in the U.S., the NFL Super Bowl. First, we imagine the case that a Super Bowl ad can be randomized on the individual level. Given current television technology, this is not technically feasible; since many people watch the game at parties, it would be di cult to link ad exposure to individuals even if television sets were a viable unit of randomization.[13] Current technology would allow geographic randomization, so one can think of our approximation as the \best case geo-randomization" in which cross-geo correlation is zero and advertisers do not pay for control ads.

[10] The National Automobile Dealers Association (NADA).

[11] Average mileage is approximately 12,000 miles per year–making a reasonable estimate of the expected lifespan of a vehicle 10-15 years; a growing population and increasing wealth will cause the average sales rate of new vehicles to exceed the mortality rate of old cars.

[12] Source: http://www.nada.org/Publications/NADADATA/2011/default

[13] Technology that gets around these issues does not, to our knowledge, currently exist. But it is conceivable that there will be a point in time where the television recognizes viewers because a smart phone is near the TV and records who was exposed to what commercial.

The formal argument we present bounds the set of advertisers that can both a ord a Super Bowl ad, which we call the a ordability constraint, and detect the return on investment, which we call the detectability constraint. The a ordability constraint is a straightforward accounting exercise to determine the rm size necessary to have the advertising budget for such a large expenditure. To build intuition on the detectability constraint, recall that ROI is the percentage return on the $ad\ cost$; speci cally it is the the sales lift times the margin minus the cost, all divided by the cost. Notice this does not depend on the level of sales. Since Super Bowl ads cost roughly the same amount for all advertisers, this means that for a small rm, the sales $level$ lift that nets a positive ROI is a much larger $percentage$ lift than it is for larger rms. The smaller rm will have an easier time identifying the sales change because it represents a much larger percentage of revenue. This is intuitive and can also be seen from the cost perspective. For a small rm, the Super Bowl ad represents a very large expenditure relative to revenue | as we have seen, this gives them more statistical bang for the buck. The detectability constraint gives the largest rm that can meaningfully evaluate a given ROI hypothesis set.

We will now present the formal argument and calibrate it with data from our experiments and publicly available information on Super Bowl advertising. We need to de ne some terms. Let N_{Total} be the total adult population, N is the total adult audience and we de ne $\rho = \frac{N}{N_{Total}}$ as the reach of the Super Bowl. N_E gives the number of reached (exposed) individuals, which we set equal to $N/2$ to maximize power. On the cost side, C the total cost of the ad, with c the cost per exposed person. Let μ equal the mean purchase amount for all customers during the campaign window and σ be the standard deviation of purchases for customers during the campaign window. We will use $\frac{\sigma}{\mu}$, the coe cient of variation, which we have noted is typically 10 for advertisers in our sample and greater than 10 in other industries, to calibrate the argument. m is the gross margin for the advertiser's business

We also need to de ne a few terms to describe the advertiser's budget. Let w be the number of weeks covered by the campaign's analysis (and the advertising expense), b gives the fraction of revenue devoted to advertising (% advertising budget) and R the total annual revenue. To get the a ordability bound, we de ne γ_C as the fraction of the ad budget in the campaign window devoted to the Super Bowl ad. For instance, if $\gamma_C = 1$, this means the rm spends all advertising dollars for the period in question on the Super Bowl.

We now present the argument, which an algebraic exercise with the one key step, substituting for the coe cient of variation and solving for the revenue bounds.

First let's construct the a ordability bound. To a ord the ad, it must be the case that it costs less than the ad budget, which is the revenue for the time period in question, $R \cdot \frac{w}{52}$, times b, the percentage of the revenue devoted to advertising, times γ_c, the fraction of the budget that can be devoted to one media outlet:

$$C \leq (R \cdot \frac{w}{52}) \cdot b \cdot \gamma_c$$

Solving this equation for revenue we get the a ordability limit:

$$R \geq \frac{C}{\gamma_C b \cdot \frac{w}{52}} \tag{5}$$

For the detectability limit, let r and r_0 be the target ROI and null hypothesis ROI respectively. The t statistic is given by:

$$\begin{aligned} t_{ROI} &\leq \frac{r - r_0}{\sqrt{\frac{2}{N}} \times \sigma_{ROI}} \\ t_{ROI} &\leq \frac{(r - r_0)}{\sqrt{\frac{2}{N}} \left(\frac{m\sigma}{\tilde{c}}\right)} \\ t_{ROI} &\leq \frac{(r - r_0)}{\sqrt{\frac{2}{N}} \left(\frac{\sigma}{\mu}\right) / \frac{\tilde{c}}{m\mu}} \end{aligned}$$

The rst equation is just the de nition of the test statistic. The second equation follows from substituting in the standard deviation of ROI, which is a linear function of the sales standard deviation, per-capita cost and gross margin. The nal equation simply multiplies the denominator by $\frac{\mu}{\mu}$. We do this so we can substitute in a constant for the coe cient of variation, $\frac{\sigma}{\mu}$ and solve for μ, as given below:

$$\mu \leq \frac{(r - r_0)\, c}{\sqrt{\frac{2}{N}} \left(\frac{\sigma}{\mu}\right) m \cdot t_{ROI}} = \mu$$

The right-most equality is for notational purposes. We can also relate mean sales during the campaign period to total revenue.

$$\mu = R \cdot \frac{\frac{w}{52}}{N_{Total}} \tag{6}$$

We can solve for revenue in the above equation and substitute in μ for μ to get the detectability limit:

$$R \leq N_{Total} \cdot \mu / \frac{w}{52} \tag{7}$$

Examining the detectability limit, referring back to μ where necessary, we see that it decreases with $\frac{\sigma}{\mu}$. This is intuitive, as the noise to signal ratio increases, inference becomes more di cult. It also falls with the required t and gross margin. To understand why the bound rises as margin falls, consider two companies, one with a high margin, one with a low margin. All else equal, the low

margin firm is experiencing a larger change in sales for a given ROI change. Naturally the bound also rises with the gap between the null hypothesis and target ROI.

Putting both limits together, we obtain the interval for detectability and affordability:

$$\frac{C}{\gamma_C b \cdot \frac{w}{52}} \leq R \leq \frac{N_{Total} \cdot \mu}{\frac{w}{52}} \tag{8}$$

.

We use the following parameter set to calibrate the model. We set w = 2 (weeks) to match most of the analysis of this paper; it is a reasonable window to measure the Super Bowl ad impact. t_{ROI} = 3 to match our standard power requirement and $\frac{\sigma}{\mu}$ = 10 to match the value we see strong evidence for in our study, even though it will understate volatility for advertisers such as automakers. We use ρ = .5 to match the empirical viewing share for adults for the Super Bowl. For the advertising budget we choose what we think is a rather high value of 5% of revenue, in particular with regards to large companies. We set the fraction of ad spend for the given period that is devoted to the Super Bowl ad at $\gamma_c = \frac{1}{w}\rho$. It says that a company would be willing to devote all advertising resources for a given week if the Super Bowl ad reached all potential consumers.

The toughest parameter to pin down is gross margin, as it varies across industry and firm. We report bounds for two values of gross margin, 0.25 and 0.50. The "low" margin, 0.25, corresponds to the gross margin, according to SEC Filings, of automobile companies and electronics retailers. For instance, Super Bowl advertiser Honda Motor Company reported a gross margin of 0.27 in 2010, Ford reported 0.18 and Best Buy reported 0.24. Roughly 40% of 2011 Super Bowl ads were in this category. The "high" margin, 0.50, corresponds to the gross margin of consumer goods such as beer and processed foods. 2011 Super Bowl advertisers Imbev (Budweiser) and Pepsi Co. reported 2010 margins of 0.55 and 0.51 respectively. This category accounts for roughly 35% of Super Bowl ads. In general we have tried to choose fair, converative values to calibrate our argument.

The final step is to calibrate pricing and audience. We use the following parameters: N_E is 50 million (1/2 the viewers), the cost of the ad is 1/2 the market rate, C = \$1,000,000. As a reminder, the thought experiment here is for geographic randomization of so that you pay half the cost, but you only get half the reach.

Table 3 gives the upper and lower bounds on R for both margin values we consider. We first note that a fully randomized Super Bowl ad experiment would be much larger than any of the campaigns we used, typically about 10 times as large. Recall that if the median campaign in our study was about 9 times larger, it would posses sufficient power to evaluate the "hard" hypothesis set. Consistent with this reasoning, we see in Table 3 that a wide range of companies could reliably test the "hard" hypothesis set using a Super Bowl experiment. Indeed most of our retailers fall within the (high margin) range of revenue.

Examining row 1, we see that a fairly wide range of companies would be able to reliably measure if the ad had influence. However, high margins companies over \$35B in revenue are unable to meet

Table 3: Super Bowl \Impossibility" Theorem Bounds

	H_A: ROI	H_0: ROI	A ordability Annual Rev.	Detectability, m=.50 Annual Rev.	Detectability, m=.25 Annual Rev.
Hard	0%	-100%	$2.08B	$34.47B	$63.3B
Harder	50%	0%	$2.08B	$17.33B	$34.6B
Hardest	10%	0%	$2.08B	$3.47B	$6.9B
Crazy	5%	0%	$2.08B	$1.73B	$3.4B

even this minimal goal. $35B sounds like a lot, but some rms operating at these margins exceed this mark. For instance, Pepsi Co. had annual revenue of $62.4B in 2010. Automakers fall in the low margin category, meaning the relevant upper bound is $63.3B, which they often soar over. Honda pulled in $107.8B in 2010, Ford $135.1B. Even when accounting for product category (cars vs. trucks, say), major automobile manufacturers are often dangerously close to this limit. Recall as well that we have assumed a $\frac{\sigma}{\mu}$ ratio of 10, which is probably half the true value for car sales, meaning the correct limit is probably double the one reported (however if the advertiser is able to advertise speci cally by model, this pushes back against the problem).

Examining row 2, we see that the upper limits are cut in half for the \hard" hypothesis set. Still though, medium-sized brands within large rms, for example \Doritos" within Pepsi Co. or a new model of automobile, and medium-size rms fall well within this constraint. For the \hardest" and \crazy" cases, the bands are tight to vanishing. Very few companies are both large enough to a ord the ad, but small enough to reliably detect relatively small di erences in ROI. The only example from last year's Super Bowl of a company (plausibly) in this range is GoDaddy.com (which is privately held). This means that the vast majority of Super Bowl advertisers would be unable to tell if the advertising investment netted more or less than the other investments they routinely make, which typically require 5-10% of well-measured ROI.

The Super Bowl \Impossibility" Theorem illustrates two important points. The rst is that even exceedingly large (hypothetical) experiments could be uninformative for many companies that advertise at such a scale. Indeed many actual Super Bowl advertisers, notably brand ads for car manufacturers, would be unable to test if the ad had any in uence on consumer behavior. We think this is a striking example of inference challenges in the advertising market. The second point highlights the importance of campaign cost relative to rm size. Since a Super Bowl ad costs the same for all advertisers, it is much easier for small rms to measure ROI changes. The reason is that for a small rm the ROI change represents a much larger percentage change in sales. For example, both GoDaddy.com and Ford need to net a similar change in the level of sales to achieve a positive return on the ad | estimating this change is far easier for the much smaller GoDaddy.com.

3.2 Experiments: Reliable but underpowered in isolation

We start with a telling quote from Eastlack Jr. and Rao (1989), who report on multiple experiments run by the Campbell Soup Company.

> The record of implementation of experimental results is mixed, however. For example, summer advertising on Condensed Soup was never implemented [estimates indicated e ectiveness]. Again, when (as in Experiments 2 and 17) it was found that reductions in spending [in regions where spend was already heavy] did not adversely impact sales, the action was not to cut budgets, but rather, to search for alternative and hopefully better creative.

Campbell devoted serious time and e ort to run a number of intelligent, geo-randomized experiments. The results were surprisingly underwhelming for rm managers (who had not read the paper you are currently reading, which is forgivable since both authors were yet to be born at the time of their experiments). In aggregate, the evidence indicated that the advertising did stimulate sales, but for individual campaigns, the estimates were imprecise. The authors had 3 major conclusions, two of which seemed to have been ignored. Based on the strength (or lack thereof) of the statistical evidence, this is rational with even somewhat informative priors.

A point we have continually made is that experiments are unbiased, but each experiment needs a large sample, and the act of experimentation may be costly itself (withholding ads to customers, for instance, could prove to be costly). The experiments we document demonstrate that through repeated experimentation a rm can build knowledge about overall characteristics and e ectiveness of its advertising spend, but evaluating a campaign or creative in real-time may be impossible. Even though this is a far-cry from the stylized textbook example of a rm advertising until the marginal dollar spent nets a dollar in pro ts, knowledge can still be accumulated. As shown in Table 2, designing an experiment and accompanying analysis that can powerfully reject ROI=-100% (\ad does something") is a realistic goal. To a lesser extent, so is distuinguishing between a very cost-

Here we'll abstract from the problem of who the rm should be advertising to in the rst place and instead focus on how targeting can be used to increase experimental power. The idea is that rms can perhaps more powerfully assess their advertising stock by performing on experiments on the particularly susceptible portion of the population. The trade-o is that targeting reduces the size of the experiment, which works against power at a rate of $\sqrt{N}$, but increases the in uence, making it easier to detect.

Suppose there are N individuals in the population the rm would consider advertising to. We assume that the rm does not know the in uence of a campaign on these individuals, but can order them by relative rank. This order is not assumed to be perfect, we only assume that errors are mean 0, so it is not the case that in expectation someone lower in the ordering has a higher in uence. The rm wants to design an experiment using M of the possible N individuals, split evenly between test and control. The question is, \When is the rm better o choosing $M < N$?" Let's de ne the following functions $\mu(M)$, $\sigma(M)$, and $C(M)$ as the mean sales, standard deviation and average cost as a function of advertising to the rst M people. First let's look at the t-statistic against the null hypothesis of -100% ROI.

$$t = \sqrt{M}\frac{\mu(M)}{\sigma(M)} \tag{9}$$

To build some intuition, if the ad has a constant e ect on the population, then $\mu(M)$ and $\sigma(M)$ are constants, meaning we get the standard results that t increase at $\sqrt{M}$. More generally, this function is increasing in M as long as the signal-to-noise ratio decreases at a rate less than $\frac{1}{\sqrt{M}}$. Stated another way, each additional observation added to the average has to add in at least $O(1/\sqrt{M})$ e ect to the mean or have a decreasing variance at $O(1/\sqrt{M})$.

Now let's calculate the σ_{ROI}.

$$\begin{aligned} ROI &= \frac{M \cdot \mu(M) - M \cdot C(M)}{M \cdot C(M)} = \frac{\cdot\mu(M)}{\cdot C(M)} - 1 \\ \sigma^2_{ROI} &= Var\left(\frac{\cdot\mu(M)}{\cdot C(M)}\right) = \frac{\cdot\sigma^2(M)}{(M \cdot C(M))^2} = \frac{\sigma^2(M)}{M \cdot (C(M))^2} \end{aligned}$$

which implies:

$$\sigma_{ROI} = \frac{\sigma(M)}{\sqrt{M} \cdot C(M)} \tag{10}$$

Notice that this formula does not rely upon the actual impact of the ads, except that we calibrate the expected e ect against the cost (in reality, costs will be correlated with ad impact). It only incorporates the average volatility of the M observations. The standard error of our estimate of the ROI is decreasing in M as long as the ratio $\sigma(M)/C(M)$ does not increase faster than $\sqrt{M}$.

For the special case of a constant variance, the standard error of the ROI can be more precisely estimated as long as the average costs do not decline faster than $\frac{1}{\sqrt{M}}$. Note average costs cannot decline faster than $\frac{1}{M}$ unless the advertiser is actually paid to take extra impressions, which seems unlikely. Another special case is constant average cost. Here as long as $\sigma(M)$ does not increase faster than $\sqrt{M}$, more precision is gain by expanding reach.

Overall, the question of whether targeting helps or hurts inference is an empirical one. If in uence is concentrated on a certain portion of the population, one is better o sample size to gain a higher signal-to-noise ratio. Conversely, if in uence is spread rather evenly across the population, targeting damages power.

3.4 Average vs. Marginal: The "Right" Interpretation and Reaction to Experimental Findings

In textbooks, the distinction between average and marginal is unambiguous. \Average" is just the total sales increase divided by total spend (over a given time period, say) and \marginal" is the impact of that \last little bit" of advertising, divided by its cost. In the interest of simplicity, we have until now remained agnostic about the average-marginal interpretation. In our experience, rms typically have ROI goals by campaign so we focused on the statistical challenges of reliably measuring ROI in relation to a speci ed campaign goal. But a deeper problem is identifying which ads are \marginal" across the various media used by the rm.

In addition to advertising online, most of our retailers were actively advertising on television, out-of-home (billboards, etc.), at major sporting events and through direct mailings. Exactly what part of this spend is marginal, from the perspective of the rm's decisions, is entirely unclear. Mechanically, the online experiments we report here measure the impact of the \marginal campaign," because the experimental randomization holds some users out from seeing a particular online campaign. These users still see the same billboards, television commercials, online ads on other websites and so forth.

Suppose a rm runs a series of online experiments and eventually rejects 0% ROI in favor of the most likely alternative, say -50% ROI. What is the appropriate response? While it's readily apparent that doing nothing is dominated by cutting online spend,[14] it is not at all obvious what the right thing to do is. Spending should be cut, yes, but where? Perhaps the rm saturated consumers with television ads, so the online campaign had little room for in uence. Alternatively, the online ads could have been duds. Or maybe the rm is advertising too much across the board, and all spending should be cut equally. In this sense, the online experiment could just signal the e ectiveness of marketing levels generally. Suppose instead that the rm measures the ROI of online spend to be +50%. Here the problem is even more di cult. Increasing online spend *does not* clearly dominate doing nothing, because the high average might bely a low marginal impact of

[14] Here we implicitly assume concavity of influence. For supportive evidence see Lewis (2011) .

extra impressions per user. Expanding reach to previously unexposed consumers seems like a safe bet, but perhaps the right response is to increase spend on all media evenly or it could be the firm should just focus on a particular media, say television.

Even with reliable estimates, producing actionable insights is hampered by these sorts of average-marginal uncertainty. Ideally the firm would experiment with all forms of media simultaneously, but delivery technology for out-of-home, television and radio makes randomizing exposure difficult or impossible. As such, insights gleaned from online experiments must be carefully incorporated into the media plan. The science of this cross-media incorporation will likely evolve as online and television experimentation becomes more common.

3.5 Advertisers in the Dark? Incentives and Statistics

Up until now we have focused on the informational and statistical challenges facing the advertiser by assuming the firm processes information rationally. This simplification might gloss over important features of the market. Given that reliable estimates are hard to come by, it is potentially a situation where incentive problems can lead to large distortions in firm behavior. In this subsection we briefly discuss the within-firm incentives to accurately measure the effectiveness of advertising.

In large firms, advertising spend is typically handled by the marketing division. Consider the plight the head of marketing (CMO). The CMO hires advertising firms to generate ad copy, media buyers to negotiate with publishers, and analysts (third party or internal), to monitor and direct the spend. What are the personal incentives of the CMO and everyone who reports to and conducts business with her? In the case that the marginal dollar of advertising has a positive ROI (the firm is under-advertising), it seems clear that these people are quite happy reporting the truth. More contracts will be given to the advertising firm, larger budgets go to the media buyers and smiles will great the analysts in internal presentations. But in the converse case, these incentives flip. News that a series of campaigns were ineffective is bad for the agency (business will go elsewhere), bad for the media buyers (who might be to blame for the poorly performing ad placements they purchased or face layoffs), bad for the analysts who have to report back on the failed campaign, which may have been a product of faulty decisions on their behalf and bad for the CMO, although to a lesser degree, whose organization might face budget cuts.

Starting from the CEO, the further one goes down the organizational chart, the less benefit one gets from promoting the correct estimate of adfx (when it is low), and the more benefit one gets from erring on the side of interpreting campaigns as cost-effective. These mistakes could be "honest," in that people tend to naturally put more weight on good news versus bad news (Eil and Rao, 2011), or more cunning. The "honest" mistake version of the story leads to everyone taking a favorable view of evidence and methods that produce overestimates. We have shown that observational methods tend to induce positive bias | the incentives help show why flaws in these methods can go unnoticed. We view this area as a fertile ground for future research.

3.6 How Unusual is this Market?

The best example of a market that we think shares the property that the information structure makes it easy for mistaken beliefs to persist is the vitamin and supplement market. In the United States, the vitamin and supplement market is estimated to gross around $20B annually.[15] Based on this lofty gure, one might nd it surprising that it is quite contentious in the medical community whether supplements do *anything* for a healthy individual. This is not to say that vitamins are not important to health | without Vitamin C one will develop scurvy, without Vitamin D rickets, etc. But supplements are not touted to prevent these types of disease, because manufacturers know that in the developed world one gets enough of these vitamins through even the unhealthiest of diets.[16] Rather supplements are supposed to improve health for a *healthy person*, and therein lies the inference challenge. The e ect is supposed to be subtle, making it di cult for an individual to detect, and across people, medical outcomes that are easily and accurately quanti able, such as illness requiring hospitalization, are noisy. Observational methods are insu cient because people who take supplements are likely more health conscious than average (selection e ects) or have recently experienced poor health, which might rebound naturally (\Ashenfelter dip" e ects[17]).

Here a standard-sized randomized medical trial of a few hundred subjects is utterly useless. Recognizing this, two large, long-running experiments were commissioned 18 years ago to examine the impact of Vitamin E, selenium, and beta-carotene on disease prevention. The Physicians Health Study II, results published in Lee et al. (2005), followed 39,876 healthy women over 12 years.[18] Half of the women received Vitamin E through a supplement pill, and the other half took a placebo pill. The dependent measures were cardiovascular disease and cancer, both of which had a per-person incidence rate of less than 10% over the span of the study. The authors state their results strongly in the abstract:

> The data from this large trial indicated that 600 IU of natural-source vitamin E taken every other day provided no overall bene t for major cardiovascular events or cancer... These data do not support recommending vitamin E supplementation for cardiovascular disease or cancer prevention among healthy women.

However, the strength of this statement belies the uncertainty of the author's estimates. For example, the 95% con dence interval on the impact on heart attacks ranged from a 23% risk

[15] For comparison, according to the Internet Advertising Board's annual report for 2009, advertisers spent $22.7 billion online in 2009 with display ads accounting for 22% or $5.1 billion.

[16] In a survey of Canadian pediatricians, researchers estimated that the overall annual incidence rate for vitamin D deficiency rickets was 2.9 cases per 100,000 people (Ward et al., 2007). In comparison, cancer incidence in Canada is 410.5 cases per 100,000 people (Marrett et al., 2008).

[17] We use this term as a nod to the classic finding that observational estimates from job training programs overstate the true effects because participants often had a negative wage shock prior to training that dissipates in the absence of training as well (Ashenfelter and Card, 1985).

[18] To consider the economic magnitude of the test, consider that 300 Costco Multivitamin costs $13.99. The pills were taken every other day, so we estimate the economic cost of the treatment roughly as $12 \frac{365/2}{300} \$13.99\ 1.05 \frac{39{,}876}{2} \$2.2M$, a little less than the cost of the 25 online advertising campaigns we examined.

reduction to an 18% risk increase (in raw terms, there were 482 myocardial infarctions in the supplement condition and 517 in the placebo). The total economic cost of a heart attack is in the neighborhood of $1M, according to recent estimates (Shaw et al., 2006). This gure places the con dence interval of the economic beni ts/costs in the neighborhood of $\pm 0.20 \times 482 \times \$1M = \pm 96M$ against the $2.1M of vitamin expenditure. The con dence interval is a staggering 50 times the cost![19] Cancer incidence had a tighter con dence interval but still ranged from a 7% risk reduction to an 8% increase. A related large-scale experiment, The Selenium and Vitamin E Cancer Prevention Trial, followed 35,533 men from 2001{2008, using a similar design. The results, reported in Lippman et al. (2009), are similar to Lee et al. (2005). No signi cant bene ts are found | the point estimates are close to zero | but the con dence bounds are wide.

These two experiments were costly and time consuming. If one's prior was these supplements provided a 1-5% improvement in health outcomes, medically important magnitudes indeed, then the data *should* do little to dissuade. The authors' strong conclusions are in fact based more on the lack of power of the experiments than the actual evidence.[20] While inconclusive in terms of medical impact, the studies do conclusively show that the supplement market is such that a new supplement can enter and make essentially untestable claims, provided it does not contain compounds known to cause harm.

We draw the comparison to the supplement market to show that it is not \crazy" to suppose that market beliefs could be severely mistaken for purely statistical inference reasons.[21] The example also illustrates that too often focus is placed on rejecting a null hypothesis or not, rather than the con dence interval of the estimate.

4 Conclusion

In this paper we quantitatively assessed the di culty of the statistical problem facing an advertiser. The challenge is driven by two key facts. First, since campaigns typically involve a modest spend per person, especially in comparison to the total amount of advertising the average person sees, the implied break-even per-capita e ect of an advertising campaign is small. Second, on the individual level, the ratio of the standard deviation of sales to the mean is about 10:1 for the majority of advertisers we study across a variety of industries. These two properties mean that estimating advertising cost-e ectiveness is akin to measuring a relatively weak signal in a sea of noise.

Using data from 25 large eld experiments run at Yahoo!, accounting for $2.8M in advertising spend, we show that even large experiments can be underpowered, given the noise in sales. A well

[19] While this is by no means a rigorous estimate of the economic impact of multivitamins, it does inform us regarding the economic uncertainty of perhaps the best study performed assessing multivitamins.

[20] For a lay summary, see http://health.usnews.com/health-news/diet-fitness/diet/articles/2008/12/09/vitamins-and-supplements-do-they-work.

[21] The supplement market stands in stark contrast to other pharmaceuticals in which the signal-to-noise ratio is much more favorable, such as antibiotics, where there is little disagreement in what works and what does not.

designed experiment can be informative, but is by no means perfect, and some questions, such as \is my ROI near 10%" are shown to be nearly impossible to answer. Even if true e ect of the campaign is hugely successfully, such as ROI=50%, we show it is di cult, but not impossible, to reliably reject that the campaign merely broke even. Given the underwhelming power of experiments, the temptatation is to turn to observational methods. It turns out that the data features that make experiments underpowered, severely bias observational methods in this setting. Observational methods that fail to account for sources of endogeneity that only explains a tiny fraction of variation in sales, R^2 on the order of 0.000005, would severely bias estimates, typically upwards.

Our conclusion is twofold. The rst is that we argue strongly for the use of experiments, given the severe biases of observational methods. The second is that experiments are not magic bullets. The information structure of the market means that advertising e ectiveness is very di cult to measure, and thus we should not be surprised if imprecise or mistaken beliefs proliferate.

References

Abraham, M. and Lodish, L. (1990). Getting the most out of advertising and promotion. *Harvard Business Review*, 68(3):50.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the e ect of training programs. *The Review of Economics and Statistics*, pages 648{660.

Bagwell, K. (2005). The economic analysis of advertising. *Handbook of Industrial Organization*, 3.

Card, D. (1999). The causal e ect of education on earnings. *Handbook of labor economics*, 3:1801{1863.

Carroll, V., Rao, A., Lee, H., Shapiro, A., and Bayus, B. (1985). The navy enlistment marketing experiment. *Marketing Science*, 4(4):352{374.

Eastlack Jr, J. and Rao, A. (1989). Advertising experiments at the campbell soup company. *Marketing Science*, pages 57{71.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American Economic Review*, 97(1):242{259.

Eil, D. and Rao, J. M. (2011). The good-news bad-news e ect:Asymmetric proccessing of objective information about yourself. *American Economic Journal: Microeconomics*, pages 529{544.

Fulgoni, G. and Morn, M. (2008). How online advertising works: Whither the click. *Comscore.com Whitepaper*.

Hummel, P., Lewis, R., and Nguyen, D. (2011). Positive spillovers from advertising on search: Empirical evidence and theoretical implications. In *Working paper*.

Johnson, G., Lewis, R., and Reiley, D. (2010). The impact of hyper-local advertising. In *Working paper*.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604{620.

Lee, I. e. a. (2005). Vitamin e in the primary prevention of cardiovascular disease and cancer. *The Journal of the American Medical Association*, 294(1):56.

Lewis, R. (2010). *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*. PhD thesis, MIT PhD Dissertation.

Lewis, R., Rao, J., and Reiley, D. (2011). Here, there, and everywhere: correlated online behaviors can lead to overestimates of the e ects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157{166. ACM.

Lewis, R. and Reiley, D. (2010). Does retail advertising work: Measuring the e ects of advertising on sales via a controlled experiment on Yahoo! In *Working paper*.

Lewis, R. and Schreiner, T. (2010). *Can Online Display Advertising Attract New Customers?* PhD thesis, MIT Dept of Economics.

Lippman, S. e. a. (2009). E ect of selenium and vitamin e on risk of prostate cancer and other cancers. *The Journal of the American Medical Association*, 301(1):39.

Lodish, L., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., and Stevens, M. (1995). How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments. *Journal of Marketing Research*, 32(2):125{139.

Marrett, L., De, P., Airia, P., and Dryer, D. (2008). Cancer in canada in 2008. *Canadian Medical Association Journal*, 179(11):1163.

Montgomery, A. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science*, pages 315{337.

PriceWaterhouseCoopers, L. (2010). Iab internet advertising revenue report 2009. *www.iab.net/insights_research/1357*.

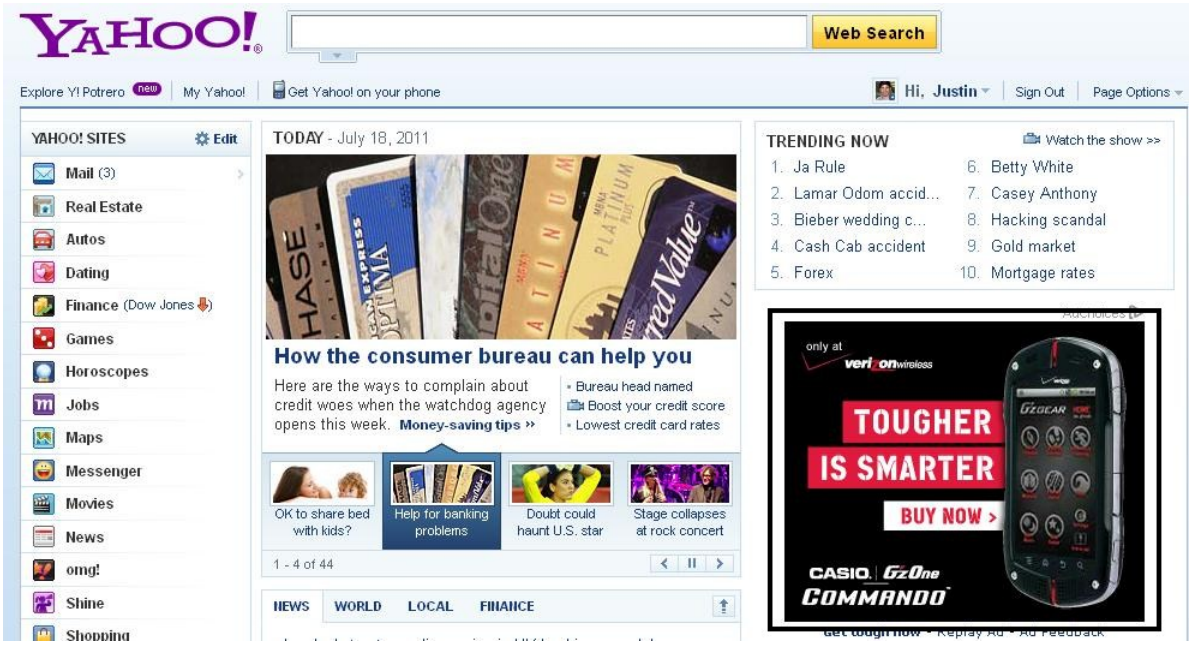
Rossi, P., McCulloch, R., and Allenby, G. (1996). The value of purchase history data in target marketing. *Marketing Science*, pages 321{340.

Shaw, L., Merz, C., Pepine, C., and et al. (2006). The wise study. the economic burden of angina in women with suspected ischemic heart disease: Results from the national institutes of health-national heart, lung, and blood institute-sponsored women's ischemia syndrome evaluation. *Circulation*, 114(9):894{904.

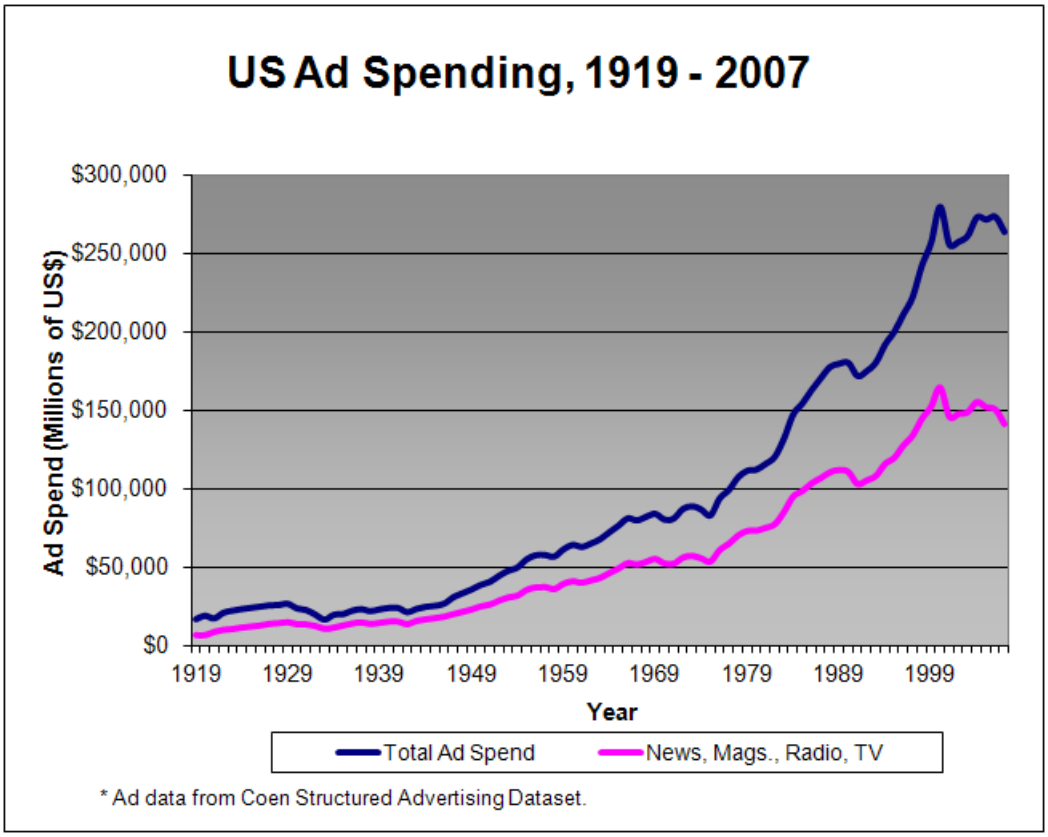
Ward, L. M., Gaboury, I., Ladhani, M., and Zlotkin, S. (2007). Vitamin dde ciency rickets among children in canada. *CMAJ*, 177(2):161{166.

Wilbur, K. (2008). How the digital video recorder (dvr) changes traditional television advertising. *Journal of Advertising*, 37(1):143{149.

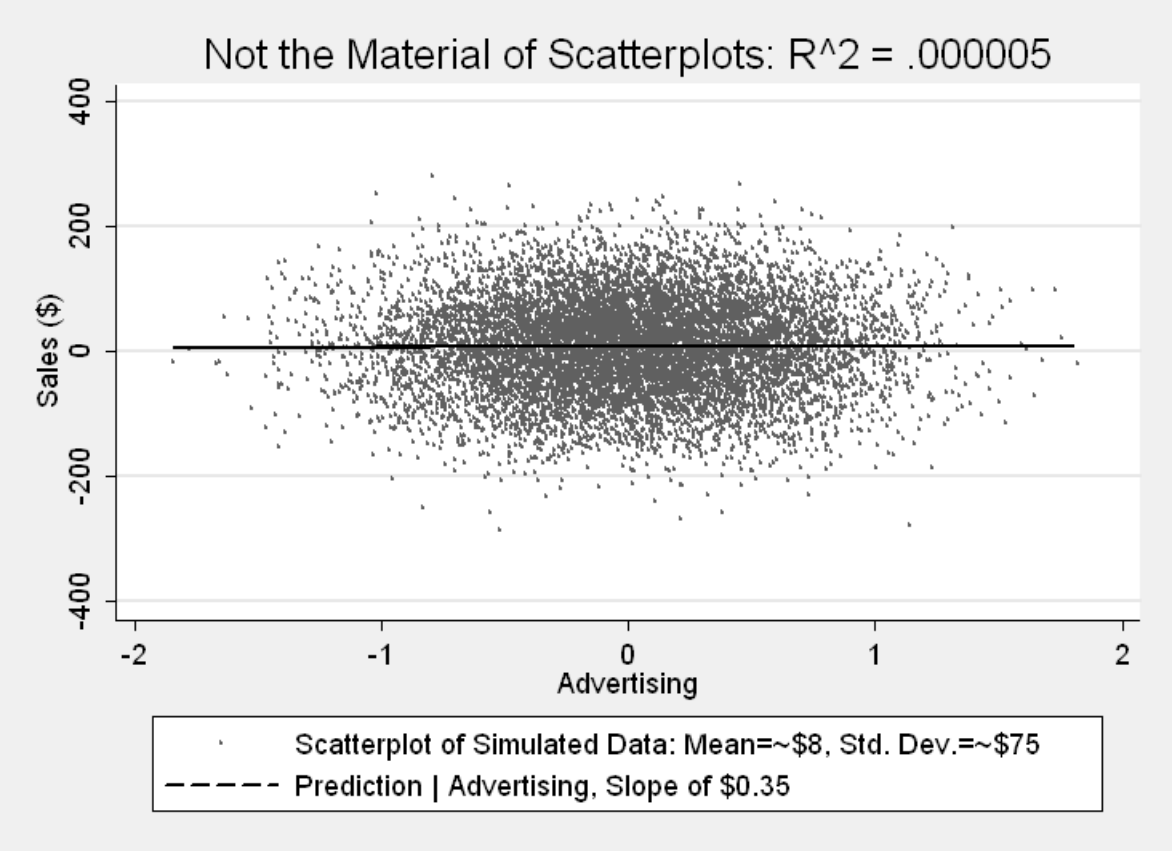
5 Appendix



Appendix Figure 1: Example of display ad.



Appendix Figure 2: U.S. Ad Spending 1919{2007.



Appendix Figure 3: Visual example of what an e ective advertisement looks like. Data calibrated with median values from Table 1. We use a normal distribution for illustration purposes only.