A Comparison of Algorithms for TV Ad Targeting

Brendan Kitts and Dyng Au
Adap.tv
Seattle USA
bkitts@adap.tv

Abstract—**Television is the largest advertising category in the United States with 70 billion spent by advertisers per year. We compare a variety of different targeting algorithms, ranging from the traditional age-gender targeting methods employed based on Nielsen ratings, to new approaches that attempt to target high probability buyers using Set Top Box data. We show the performance of these different algorithms on a real television campaign, and discuss the advantages and limitations of each method. In contrast to other theoretical work, all methods presented in this paper are implementable on current television delivery systems.**

Keywords-television; advertising; targeting

I. INTRODUCTION

TELEVISION is the largest advertising category in the United States with around 70 billion per year in revenue [14]. Television consumption is also growing – hours per capita have continued to increase as reported by Nielsen based on their panel – even with the emergence of other video platforms such as mobile [27], [28], [29]. According to Nielsen, approximately 20 times more hours are spent viewing television (TV) as viewing on internet or mobile video [27]. The quality of the TV viewing experience has if anything gotten far better in recent years with high-definition television sets and a flowering of innovative original programs, and many authors have commented that at least for viewers this seems to be a "Golden Age" for television viewing [12], [13], [41], [42].

If there is an area for improvement in television, it is around how advertising technology can continue to remain relevant and effective.

Television presents unique challenges for advertising. In online advertising it is possible to deliver ads to individual persons, via cookies and IP addresses, and to then track the behavior of those persons including whether they convert. In television, advertisements are embedded in a single high definition video stream and broadcast using over-the-air terrestrial transmission towers, satellite and cable. Advertisers are therefore not delivering ads to individuals, but rather to large sets of people referred to as audiences. Thus television advertising has more in common with Contextual Advertising in which ads need to be targeted to the web sites.

However in addition to being constrained to deliver ads to audiences, TV systems also don't typically allow the advertiser to know if individuals saw the ad and if their purchases were related to having seen the ad.

Because of these limitations, since the 1950s this medium has been tracked using a 25,000 person, Nielsen panel with diaries. These individuals could report on what they saw on TV, and then this could be extrapolated to the United States (114,600,000 households). This panel is both expensive to maintain and is also small.

However this situation is changing. In the United States, Set Top Boxes are now present in 91.5% of US homes – more common than computers. More significantly, since 2009 Set Top Boxes with return path capabilities have proliferated in the United States, comprising 30% of households [16]. This means that there are thousands of times more households than the Nielsen panel. This has begun to open up new possibilities for television targeting [4], [16], [27].

This paper will present a current survey of methods for television ad targeting ranging from traditional media buying approaches [8] to new Set Top Box methods [3].

In contrast to some other papers that discuss theoretical methods for TV ad targeting, the present research focuses on methods that are deployable today at full scale using current US data collection and US TV broadcasting capabilities. Our aim is to show the state of the art, and to also help provide a framework for understanding the general TV targeting problem and approaches for solving it that are available today.

The present paper makes four contributions:

1. Describes the data format available for television targeting.
2. Formalizes the TV Ad Targeting problem into a well-defined objective function.
3. Identifies the variables available for Ad targeting which can be used for targeting practical Television campaigns using current television systems.
4. Compares different algorithms on TV data.
5. Discusses a method for combining the different algorithms to get the best effect.

II. The TV Ad Targeting Problem

A. Media Instance

A TV Media Instance M_i (also known as a "spot") is a segment of time on television which can be purchased for advertising. We will define the Media Instance M_i as an element of the Cartesian product of the following:

$$M_i \in S \times P \times D \times H \times T \times G \times POD \times POS \times L$$

where S is Station, P is Program, D is Day-Of-Week, H is Hour-Of-Day, T is Calendar-Time, G is Geography, POD is the Ad-Pod, POS is the Pod-Position, and L is Media-Length.

Stations include Broadcast and Cable stations and are generally identified by their call-letters such as KIRO and CNN. Geography includes National, Direct Market Association Areas such as Miami, FL and Cable Zones such as Comcast Miami Beach. An Ad Pod is a set of advertisements that run contiguously in time during the commercial break for a TV program. Pod position is the sequential order of the ad within its pod. Media Length is the duration of the time segment in seconds – common ad lengths include 30, 15 and 60 second spots.

B. Bids

Advertisers provide a bid $CPI(M_i)$ for each media instance that the advertiser wants to run their ad on. They also provide TV stations with a recording of their television commercial in electronic form which is called the ad copy. If the advertiser's bid clears, the television station then inserts the ad into pod, positions, station, day, hour, date based on the advertiser's instructions.

C. Objective

The ad targeting problem for the advertiser to select a set of media M_i to purchase, and bids for that media $CPI(M_i)$, such that the expected ad response per dollar is maximized:

$$M_i\text{: } \mathbf{max} \sum_i \; R_\Omega(P, M_i) \cdot I(M_i) \qquad (1)$$

subject to $\sum_i \; CPI(M_i) \cdot I(M_i) \leq \text{B}$ and $V(\{M_i\}) = true$

where $R_\Omega(P, M_i)$ is the response (conversion/sales/revenue) per impression for media instance M_i, given an advertiser's product P, $I(M_i)$ are the impressions for media M_i, B is the television campaign budget, V determines if the set of media violates rotation rules (such as running an ad more than once per 60 minutes, having greater than 5% of budget on any one network or day-part, and so on). Rotation rules are defined by television ad buyers and we will not focus on them in this paper.

A greedy strategy for allocating television media is to iteratively select media in order of value per dollar

$$M_i\text{: } \mathbf{max} \frac{R_\Omega(P,M_i)}{CPI(M_i)} \qquad (2)$$

subject to the rotation rule constraints V until the budget is filled. $CPI(M_i)$ and $R_\Omega(M_i)$ are both estimates using historical clearing prices and media observations. Thus our problem reduces to maximizing (2) using machine learning estimates of the price for which the inventory will be listed, and the value the inventory will generate.

The remainder of this paper will discuss the problem of targeting, which amounts to estimating the $R_\Omega(M_i)$ response per impression part of the objective function above. Cost and impression estimation has been discussed elsewhere [20], [43]. In order to estimate the value of buying media, we will define a targeting algorithm using two variables (a) Media asset patterns which represent features for estimating a future airing m_i, and (b) an ad response $R_\Omega(P, m_i)$ measured from those features.

I. Media Asset Patterns

The first concept we will introduce is what we call a media asset pattern. A media asset pattern is a feature set representing a particular set of variable value instantiations of the media instance.

Formally, we define $m_{i,t} \subseteq M_i$ to be a subset of instantiated features from the media instance M_i.

For example consider a future media instance: $M_i =$ (CNN, 8pm, "Piers Morgan", Tuesday, 12/12/2012, Pod1, Pos2, 60s). The following Media Asset Patterns could be used to predict its performance: Station $m_{i1} =$ (CNN) , Station-Hour-Pod $m_{i2} =$ (CNN, 8pm, Pod1) , Geography-Station $m_{i3} =$ (National-CNN), and so on.

We will now enumerate several major Media Asset Patterns that can be scored for an upcoming Media Instance:

A. Program

The distinctive thing about TV are its *programs*. Different programs appeal to different people – for example, viewers of TLC's "I Didn't Know I Was Pregnant" are different to viewers of SYFY's "Continuum".

There are over 450,000 weekpart-daypart-programs available to be purchased on TV. Programs are intuitively what people tune into, and intuitively should be good predictors of ad performance. The most impactful programs are those which have high observed impressions / expected impressions for their station-timeslot $\frac{I(m_P)}{I(m_{SDH})}$. Table I shows a list of the top programs based on the above ranking in 2012. "Super Bowl", "Macy's Thanksgiving Day Parade", and "The Oscars", and others are easy to spot on the list. One of the most interesting programs to appear is a program named "Honey Badgers!". This program became a cultural sensation in 2011. In 2011 a YouTube video was posted on this National Geographic WILD Discovery program, but with an extremely humorous voice-over commentary by a person only identified as "Randall" [44]. The video garnered over 69 million views [45]. A lot of people who saw the YouTube spoof video might have been curious about the original program, which might have sent ratings for the otherwise unassuming WILD TV program through the roof.

TABLE I. IMPRESSIONS OBSERVED OVER EXPECTED

Station-Program	RE
NFLN - NFL Football	20.49714
NBC - Super Bowl XLVI	18.06963
NFLN – Postgame	15.35507
CBS - Super Bowl XLIV	15.2775
ESPN - NFL Football	12.66412
NBCSN - 2012 NHL All-Star Game	10.47042
SPD - NASCAR Sprint Cup	10.39651
FOX - Super Bowl XLV	9.862597
E! - Live from the Red Carpet: The 2012 Grammy Awards	4.467404
NBC - Macy's Thanksgiving Day Parade	4.434626
ABC - Oscars Red Carpet Live	4.288276
BBCA - William & Kate: The First Year	4.135
ABC - Dancing With the Stars	4.126531
VH1 - 2010 MTV Video Music Awards	3.863292
ABC - CMA Awards 2011	3.831977
FUSE - Whitney Houston: A Tribute	3.770582
VH1 - 2011 Video Music Awards	3.423895
E! - Live from the Red Carpet: The Academy Awards	3.30741
NBC – Voice	3.305157
CNN - Arizona Republican Presidential Debate	3.086414
CNN - New Hampshire GOP Debate	3.009244
E! - Live from the Red Carpet: Grammys	2.987157
WILD - Honey Badgers	2.939016

A. *Station-Day-Hour*

Station-Day-Hour (without the program) has an advantage of a large number of observations. Programming changes every few months, but at the same time, stations often run similar programming in the same Station-day-hour timeslots, which adds to the value as a predictor. Thus, the increase in data and signal need to be traded off against the potential error due to changes in programming.

B. *Other*

A variety of other features can also be used to represent media including (a) Run Of Station (average performance for the entire station), (b) Market-Station-Day-Hour which enables local differences to be captured – there are over 200 * 80,000 = 16 million of these features per ad. (c) Most recent Airing: When scoring a program, the most recent airing at the same time can be used – for example, rather than taking an average over several months, this uses the actual buyers per million observed in the last airing. (d) Pod-Station-Day-Hour: Pod position – the sequence in which the ad appears during commercial breaks - also has a large effect on performance as audiences exhibit ad avoidance behavior; however the first pod tends to retain most of its audience.

II. AD EFFECTIVENESS

The second variable that we need to define is an ad effectiveness measure $r_\Omega(P, m_i)$ where P is the advertiser's product and m_j is a media asset pattern. This is a measure which is positive and monotonic with the lift from advertising [39]. Several measures of Ad Effectiveness are possible and we will discuss each in more detail in Section IV.

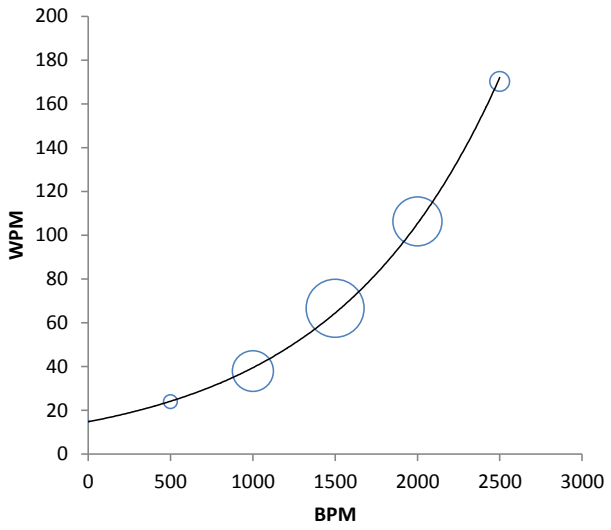
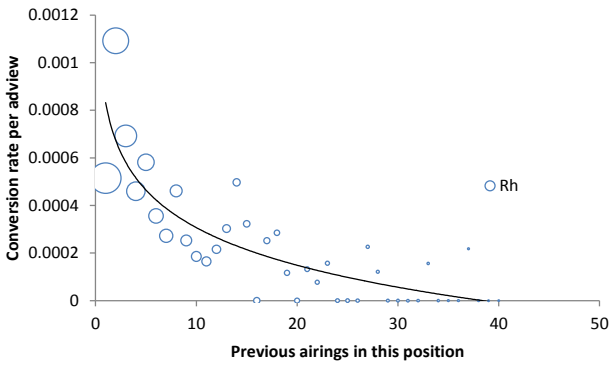


Fig.1. Web spike response per million impressions (WPM) increases with the number of buyers per million (BPM) reached by the television ad.

III. FATIGUE

If an ad is run in the same program every day, its effectiveness should decrease. Meta-studies of hundreds of publications have concluded that advertising response shows diminishing returns at all levels of frequency greater than 1 [15], [17], [18], [33], [38]. Jones [18] writing in Journal of Advertising Research summarizes these findings as "The preponderance of diminishing returns is by now widely accepted by the research community, and the facts do not need to be discussed further." Our own campaign data in television has verified the same using both set top box conversion rate as well as phone response to television airings. Indeed we show that the decline increases as a log of the number of airings (Fig. 2).



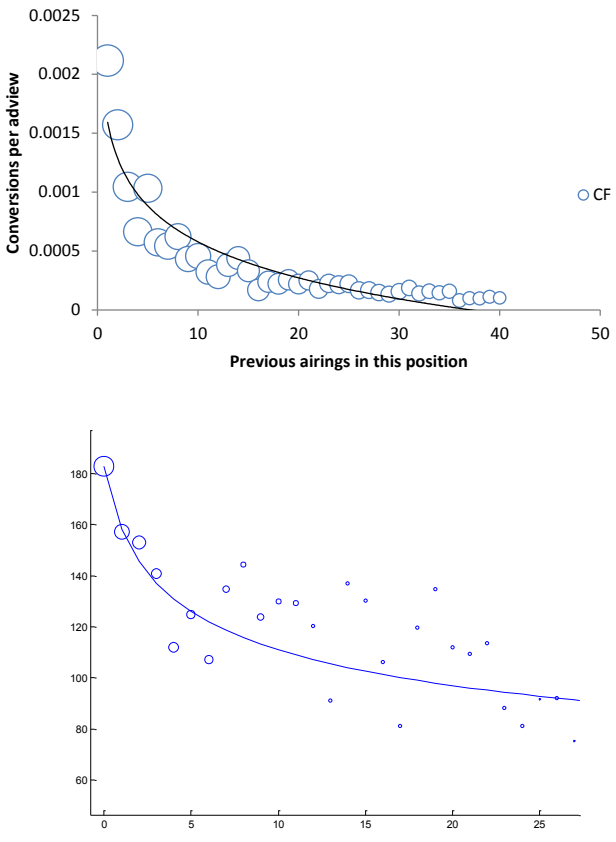


Fig.2. Top: and Middle: Person-level conversions per adview for two different products – conversion rate declines as a function of the log of airings **Bottom:** Phone calls per million impressions in response to an embedded phone number in a TV ad observed after placing an ad in the same station-day-hour 1, 2, 3, ..., 20 times. The number of phone calls declines as a log of the number of previous airings.

We've shown that frequency has a big effect on response per impression. In order account for this we divide our measure of ad effectiveness by a function of the log of airings. We still need to estimate $r_\Omega(P, m_i)$, and we will turn to that next.

$$R_\Omega(\bar{P}, m_i) = \frac{r_\Omega(P, m_i)}{\mathrm{a} * \ln(A(m_i) + 1)}$$

IV. TV Ad Targeting Algorithms

We can now define several classic TV Ad Targeting algorithms by describing their choice of media asset pattern m and ad effectiveness metric $r_\Omega(P, m_i)$.

A. Target Rating Points (TRPs) on Age-Gender (Nielsen, 1950):

Age-gender Target Rating Points are arguably the most widely used form of targeting. This form of targeting defines a Target Rating Point as the number of persons who match the advertiser's target demographics divided by total viewing persons [28].

$$r_A(P, m_i) = 100 \cdot \frac{\tau(m_i, P)}{\#Q(m_i)}$$

where $Q(M_i)$ is a set of viewers who are watching TV media instance M_i and where this viewing activity recorded by Nielsen panel and $q_k \in Q(M_i)$. Let # be the cardinality of a set, $\#r_T$ be persons that match on all demographics.

For example 50% means that 50% of the people are a match to the desired demographics. Age-gender TRPs are defined using standard Nielsen Market Breaks – gender=male|female, age=18-24, 25-34, 35-44, 45-54, 55-64, 65+. The target age and gender breaks are defined as the highest.

B. Tellis' Phone Response Per Impression (Tellis, et. al., 2005):

In cases where a TV ad has been run which included a 1800 number, it is possible to match the phone responses on specific 1800 numbers back to the ad that was placed. We can then use this data to unambiguously track sales due to the TV ad [11].

Tellis' specific method used a series of hour lag terms to predict the number of phone-calls that would be generated on a given hour [34]. We have implemented a Tellis-like method by exposing hour and day-lag terms for historical phone response

$$r_B(P, m_i) = \frac{CALL(m_i)}{I(m_i)}$$

where $CALL(m_i)$ are the number of calls from media m_i.

C. Buyer Ratings (Canning, et. al., 2009):

Buyer targeting looks for media that has a high rate of observed buyers per impression, and targets those programs. The algorithm isn't "trained" per se – it simply scores the percent of buyers observed in each media and so can be thought of as "buyer ratings" [3].

$$r_C(P, m_i) = \frac{B(m_i)}{I(m_i)}$$

D. Balakrishnan (2012)'s Reach Maximizer:

Balakrishnan presented a method based on Set Top Box data for selecting the maximum reach media plan [2]. We can model this counting the number of unique persons reached for each media. They used the same program, previous *N* airings to predict the next airing's reach for the program:

$$r_D(P, m_i) = \frac{\mathrm{U}(m_i)}{I(m_i)}$$

where *U* is the number of unique persons who viewed an airing on media *m*.

E. High Dimensional Demographic Matching (Kitts, et. al., 2013):

This method calculates the demographic match across 3,000 variables between the ad product buyer and each media asset pattern. It is like age-gender matching, but uses a thousand times more variables and a match function that works in high dimensional vector space [20]. We define the demographic match between an ad product and media to be as follows:

$$r_E(P, m_i) = \frac{\bar{P}^+ \cdot \bar{m}_i{}^+}{|\bar{P}^+| \cdot |\bar{m}_i{}^+|}$$

where $\bar{P}$ is a vector of demographics representing the average buyer demographic readings, and M is a vector of demographics for the media placement.

F. Web Spike Per Impression

If TV broadcasts are aligned in time and geography with web traffic, it is possible to calculate the difference in web visits due to each broadcast by comparing web activity a few minutes before and after the broadcast. These web spike effects are strongest about 13 seconds after the airing. Details on calculation of web spike per impression are provided in [21]. We can use

$$r_F(P, m_i) = \frac{\Delta W(m_i)}{I(m_i)}$$

where $\Delta W(m_i) = W(m_i, t_1, g) - W(m_i, t_2, g)$ is the difference in web activity at time t_1 vs t_2 and coming from same geographic area g and where the time and geography matches the media asset pattern m_i and the airings match crtieria described in more detail in [21].

G. Other Methods

There are many other methods, but our objective is to highlight major classes of approach and to see what we can learn about each method. Methods A through E have had anecdotal data published on their effectiveness [28], [34], [2], [20]. However in no publication to our knowledge, is there a comprehensive comparison of all of these techniques using a significant volume of live campaign data.

V. Comparison Of Targeting Algorithms

In order to measure each method, we used data collected from three live TV campaigns. These campaigns comprised 18,476 TV media instances, run between 1/30/2012 and 2/17/2013, representing $2.6 million in live advertising spend. The TV ads included embedded phone numbers and so could use phone response data. These TV ads did not include web data feeds, and so the web spike method could not be tested.

The variables tested with this live TV campaign data included:

(a) STBHeadMatch: High Dimensional targeting
(b) Telesales: Tellis' phone response methods
(c) Age-Gender: Nielsen Age-Gender TRPs
(d) Reach: Balakrishnan's reach maximize
(e) Sale: Canning et. al.'s Buyer ratings
(f) US Census: Sales per capita in a geographic area.

The ad targeting algorithms are each a combination of the (i) Ad Effectiveness Metric R_Ω and (ii) Media Asset Pattern Type m. For example, STBHeadMatch-Station-Day-Hour refers to a High Dimensional Match with Set Top Box data using statistics on Station-Day-Hours.

In order to assess each algorithm, we measured the correlation coefficient between each algorithm's ad effectiveness estimate, and the number of buyers per million in the program in an upcoming airing.

We note as a caveat that algorithm performance was influenced by the commercials that we used for evaluation. We used a product that appeals to an older demographic and that tends to watch Daytime television, and as a result many station-day-hour features performed quite well. In Prime Time the station-program features tend to be more predictive. Thus the results shown here can vary with the mix of commercials [43].

The results are shown in Table II. The top performing targeting algorithm is STBHead-Station-Day-Hour. The method has a correlation with buyers per million of 0.8471 and is also present 93.9% of the time, so can be used for a large number of airings.

Telesale-Station-Day-Hour-Local has a correlation of 0.8245 – which is also quite high – but is present only 48% of the time.

The worst performer was sales per capita of the geographic broadcast (0.0162).

In general most of the features below provided some value in predicting airings that would have high buyers per impression.

TABLE II. TV AD TARGETING ALGORITHMS

TV Ad Targeting Algorithm	R	Present
32-STBHead-Station - Day – Hour	0.8471	0.9391
40-Telesale-Station - Day - Hour-Local	0.8245	0.4775
60-STBHead-Station - Program Authority	0.7585	0.2385
39-Telesale-Station-Local	0.7498	0.7451
65-AgeGender-SpecialEvent-Station - Program Authority	0.6964	0.0081
118-Reach-Station - Day – Hour	0.6597	0.2688
45-Sale-Station - Day – Hour	0.6471	0.8938
31-STBHead-Station – Rotation	0.6102	0.9391
59-AgeGender-Station - Program Authority	0.4901	0.2037
28-STBHead-Program	0.4801	0.5162
124-Reach-Program Authority	0.4544	0.465
30-STBHead-Hour of Day	0.4424	1
27-STBHead-Station	0.3886	0.9391
55-AgeGender-Program Authority	0.3771	0.5985
53-AgeGender-Station – Program	0.3262	0.153
58-Telesale-Station - Day – Hour	0.2793	0.802
46-Sale-Station	0.26	0.9087
51-AgeGender-Station - Day – Hour	0.2478	0.8313
29-STBHead-Day of Week	0.1601	1
52-AgeGender-Station	0.1099	0.9009
57-Telesale-Station	0.1079	0.8702
33-USCensus-DMA	0.0162	0.8073

A. Behavior of TV Ad Targeting Algorithms

One critical element affecting each algorithm is sparsity. Set Top Box buyer data on persons who have bought the advertiser's product, and were also detected watching a particular program – will have the greatest problems with sparsity. The probability of detection of these customers is small. For given media the number of buyers that we can expect to observe viewing the media is equal to:

$$B(m_i) = I(m_i) \cdot \frac{A}{TVHH} \cdot \frac{S}{TVHH}$$

Given S=1 million Set Top Boxes (out of TVHH=114 million households) and A=10,000 advertiser customers, this would mean that only for programs with more than 100,000 impressions, we could expect 1 buyer to be detected, which are very small statistics. The key reason for sparsity is because each person must be matched in both STB data and Advertiser data. This in turn means that direct Buyer Rating

counting algorithms are likely to have problems with all but the highest TV programs.

High Dimensional Demographic matching is not as impacted by sparsity. It aggregates all STB data into a demographic vector and then matches using this vector. Let's take the same A=10,000 advertiser conversions and then enrich with 3,000 demographics. We do the same with our STB persons (1 million people as before).

By converting to a demographic vector we have now eliminated the need for "cross-domain" person-to-person linkage. Rather than 1 person matching in 100,000 impressions, the entire STB population can now used for targeting. This is certainly orders of magnitude more data, although the profile may lose information. Which method is better?

We investigated this question by dividing all airings into 5 quartiles based on impression volume. We then analyzed the performance of each metric in predicting future phone responses on airings in this impression quartile.

Figure 3 shows an analysis of the three major classes of Ad Effectiveness metric (a) Demographic match, (b) RPI: Phone response per impression, and (c) BPI: Buyers per impression, versus the size of media being scored. The y-axis is the correlation coefficient between the predicted phone responses and actual phone responses in the future. The x-axis is the number of impressions generated by the media that is being scored. Each dot is a quartiled set of airings, with their correlation coefficient for predicting future phone response. A linear fit has been added to each set of points to give an idea of the accuracy trend for that ad effectiveness metric versus impressions.

Phone RPI tends to perform very well and is sloped upwards. That means that as an airing has more impressions, prediction improves. For large airings around 50,000 impressions in size, the correlation coefficient averages 0.6. For programs with fewer than 1,100 impression, RPI prediction performance goes to random.

Demographic matching has a shallower slope. Its prediction gets better with more impressions, but it is ultimately out-performed on high impression airings by RPI. However a key differentiator of the Demographic match method is that the shallow slope means that it continues to show good prediction performance far down the list of airings, into very low impression airings. This is a critical advantage for the demographic mach method, and means that virtually the entire TV spectrum can be scored and used with this method.

BPI (labeled "abilitec" in Fig. 3) shows the most intriguing performance. Because of the high sparsity associated with it, this method only begins to be useful on airings over 600,000 impressions in size – very large airing sizes. However the slope of BPI is quite steep. It is possible that BPI might out-pace RPI and ultimately be a more predictive variable, with enough Set Top Boxes or the right Advertiser that is generating a lot of purchases.

In terms of usable predictions (scoring airings with impressions such that prediction performance is above 0), Demographic match covered 99% of all airings. RPI covered 57%. BPI covered only 0.5% (Figure 4).

All three methods are needed in practice since were we to rely on RPI, for example, half of all airings would not have any information. It is also clear that all three methods could be combined to produce better prediction performance. Demographic matching beats all methods on low impression airings (<6,000 impressions). However RPI is effective on medium impression sizes. BPI should be incorporated on airings with > 600,000 impressions.

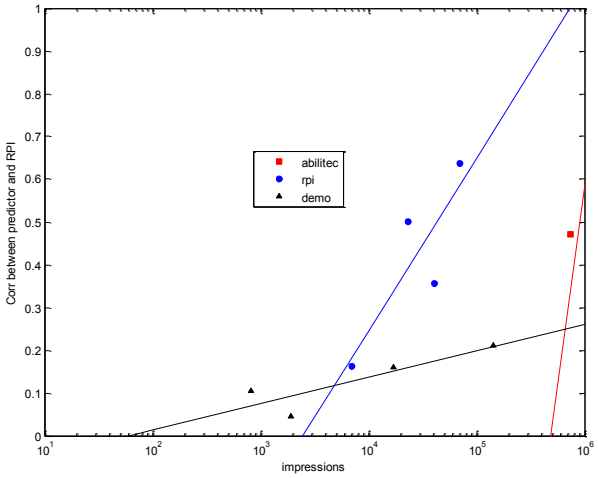


Fig.3. Three classes of ad effectiveness measure, and their performance compared to the size of airing. Some points are below the 0 correlation and are not shown.

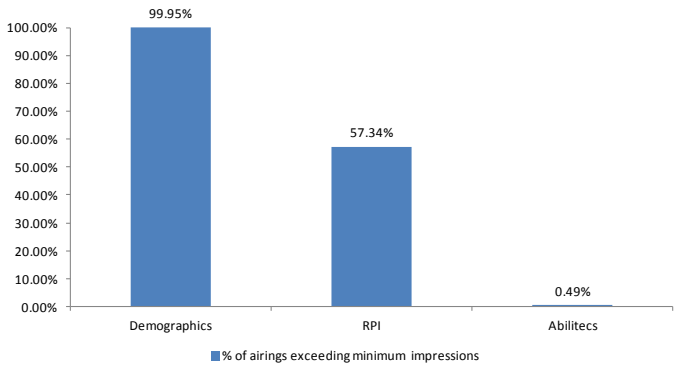


Fig.4. Number of airings in our TV data that were able to be scored by High-Dimensional Demographic match, RPI and BPI ("Abilitecs").

VI. Combined Algorithm

A. Combined Algorithm

Is it possible to use all of the methods that we have described above in order to improve performance?

In order to build a combined algorithm we would need to overcome problems introduced by the different metrics and range of each algorithm, we need to be able to select features that are most predictive, and we need to be able to train the algorithm. We'll describe the procedures we've developed below.

B. Model

The basic idea of the model is to take all of the available media asset patterns $m_{i,t}$ and Ad Effectiveness measures $r_a(m_{i,t})$, and to use them to predict the ad response per impression $R_\Omega(M_i)$. The ad response $R_\Omega(M_i)$ is a special variable and is typically one of the ad effectiveness measures that the advertiser decides is the quantity that they want to maximize in their objective function. For example, the advertiser may want to maximize buyers per impression reached or phone responses per impression generated.

This is fundamentally a supervised learning problem as ad effectiveness information is available for some airings, and so the system can be trained to predict the quantity based on historical examples.

Our model is a form of Stacked Estimator [37] where each ad effectiveness model $r_\alpha(m_{i,t})$ is an expert, and the assembly is trained to predict ad response $R_\Omega(M_i)$.

$$R_\Omega(M_i) = Z^{-1}(y, \mu_\Omega, \sigma_\Omega)$$

$$y = \sum_t w_t x_t$$

$$x_t = Z(r_t(m_t), \mu_t, \sigma_t)$$

The predictors x_t and ad response target $y = Z(R_\Omega, \mu, \sigma)$ are standardized (details below).

C. Variable Standardization

We would like to be able to use different ad effectiveness variables – ranging from telephone response per impression, to buyers per impression and demographic match. However each of these variables has a different set of units. In order to handle these different scales, we standardize all variables using the following transform.

$$x_t = Z(r_t); y = Z(R_\Omega); Z(a) = (a - \mu)/\sigma \quad (1)$$

When we train the system to predict standardized target y for each ad effectiveness predictor x_t, each predictor is effectively measuring the relationship between a change of a unit standard deviation in its distribution, to what that translates into in terms of standard deviations of movement in the target variable (Figure 5). This has several useful properties:

1. No constant term: The constant term is in effect removed and the co-variance is measured. The constant term is "added back" later when the prediction is converted back into target unit.
2. Interpretability: By standardizing the variables in this manner, all variables are on the same scale. When we estimate weights, we can then read off the weights in order of magnitude and clearly see which variables are contributing most to the prediction.
3. Usability: It also makes it easy for users to enter their own weights if they have some domain knowledge. Because of standardization (1), w=0.4 intuitively means that 40% of the decision should be based on this variable.

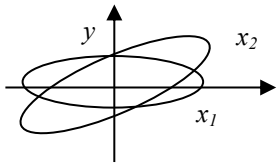


Fig.5. Ad effectiveness measures x_1 and x_2 the target variable y are all transformed into standardized 0 mean unit standard deviation coordinates. The slope of each standardized predictor indicates how effective the standardized predictor is in predicting the target variable.

D. Constraints due to Ad Theory

There are certain constraints that we can impose on the model due to experimental findings from Advertising Theory. Ad theory suggests that as the traits of the ad match the product more, response to advertising should increase [1], [9], [35], [36]. This leads us to the following propositions for ad effectiveness metrics:

Proposition 1: Ad Effectiveness $\forall i: x_i y > 0$: Each ad effectiveness metric x_i is positively correlated with ad response y.

Proposition 2: Non-negativity for ad effectiveness: Given a model predicting ad response $y = \sum w_t x_t \ \forall t: w_t \geq 0$. The effect of improved ad effectiveness is always zero or positive on ad response.

E. Minimum Weight Constraints

In order to be consistent with Propositions 1 and 2 we build a positivity constraint into our weights

$$w_t \geq 0 \quad (1)$$

F. Sum of Weight Constraints

For reasons of robustness in production, it is important to ensure that predictions do not extrapolate higher or lower than the range of values that had been observed previously. For example, a weight of 2 might lead to the system predicting outside of the range of the ad response variable.

We ensured this by adding a constraint that all weights sum to 1. As a result of this additional constraint we also have

$$1 \geq w_t \geq 0 \wedge \sum_{t=1}^{T} w_t = 1 \quad (2)$$

A. Low Data Behavior / Variable Participation Thresholds

Each media asset pattern covers a certain number of historical airings. For each MAP m, we sum the number of impressions observed $I(m)$. This gives rise to a problem – the ad effectiveness measures may be unreliable on small amounts of data.

Some authors use Bayesian priors to "fill in" performance when there is less information available, modifying the ad effectiveness score as follows.

$$r = e^{-\alpha \cdot I(m)} \cdot r + (1 - e^{-\alpha \cdot I(m)}) \cdot r_{PRIOR}$$

where α is a parameter which governs how many impressions need to be collected for the posterior estimate to be favored more heavily than the prior. However priors are often incorrect and require creation themselves, and since there are hundreds of thousands of variables per product (not to mention hundreds of products), this introduces a large number of parameters that need to be set. The effect of poorly set priors is quite significant as they cause variables that may have been good predictors to be spoiled, and the training process to be unable to weight them properly. Our production system needs to be able to work reliably with minimal human intervention. We have found it more reliable to train the system using *participation thresholds*. We define a I_{MIN} which are the minimum impressions allowed on a particular media asset pattern in order for it to be used in prediction. If a MAP fails to meet this threshold, it is converted to missing value, and so does not participate further. The prediction formula elegantly handles missing values.

$$\text{if } I(\bar{m}_{i,t}) < I_{MIN} \vee \sigma_t = 0 \text{ then } w_t = 0; x_t = MV$$

B. Missing Value Handling

We may have a situation in which a Media Asset Pattern Type may be missing or otherwise unable to report a value. For example, the system may not have enough data on the Program to be able to provide a prediction. When this happens, it is important that the system *degrade gracefully*. The system will need to use a more general Media Asset Pattern type – such as the Station to provide a prediction.

Missing value handling is fairly graceful in that if a variable is not available, it is zeroed out and the other variables that are present are used to create the prediction.

For production robustness we also ensure that more general – and low missing value - maptypes are defined with small weights, so that if there is a failure then the system will default to one of these more general maptypes. For example, if Station-Day-Hour is undefined, then Station will be defined but at a very low weight. It therefore only exerts a significant weight when there is a failure on the primary features.

A. Transforming into Target Units

Ultimately we want to convert our standardized prediction into original units. We can do that by inverting the z-score transform

$$\mathrm{Z}^{-1}[y] = y\sigma_j + \mu_j$$

where j is the ad effectiveness measure that is being reported. The Z^{-1} transform is like performing a Programming language cast operation into the appropriate units.

A. Training Algorithm

Weight training uses the subspace trust-region method described by Coleman and Li [6] which is specially designed for the 0..1 and sum of weights = 1 constraints below.

$$w_t \ \mathbf{min}\, E = \mathbf{min} \sum_{\mathrm{i}} \left[\left(\frac{\mathbf{1}}{\sum_{t=1}^{T} \boldsymbol{w_t}} \sum_{t=1}^{T} w_t x_t \right) - y^* \right]^2$$

$$1 \geq w_t \geq 0 \wedge \sum_{t=1}^{T} w_t = 1$$

$$\text{If } x_t = \text{MV then } w_t = 0$$

We use a Forward-Backward selection algorithm to select new features to include in the model.

B. Analysis of model

We will next review the behavior of the algorithm to help set up the algorithm for success.

Theorem 1: Model with one variable $y = w_i x_i$ will have positive weight $w_i \geq 0$ (from Proposition 1)

From $y = w_i x_i$ we can show that

$$w_i x_i y = y^2; w_i = \frac{y^2}{x_i y}$$

Since $y^2 \geq 0$ and $x_i y > 0$ from Proposition 1 then $w_i \geq 0$.

Theorem 2: Given any model with multiple variables $= \sum_i w_i x_i$, a new variable x_n will have positive weight if Case I, II, or III below is met $\forall i$.

Consider any new variable x_n along with a set of one or more variables x_i The error from the model is defined below:

$$E(yest, y) = (yest - y)^2 = \left(\sum_i w_i x_i + w_n x_n - y \right)^2$$

$$\frac{\partial E}{dw_n} = 2x_n \left(\sum_i w_i x_i + w_n x_n - y \right)$$

$$= 2x_n(yest - y) = 2(x_n \cdot yest - x_n y) \quad (5)$$

$$= 2x_n err \quad (6)$$

We know that $x_n y > 0$ from Proposition 1. Therefore error will decrease when variable x_n is added in three cases:

Case I: If the variable x_n has higher covariance with the dependent than the current assembly $yest$, ie. $x_n y > x_n \cdot yest$ (from 5)

Case II: If the new variable is negatively correlated with the existing linear combination, ie. $x_n \cdot yest < 0$, yet is still correlated with the dependent $x_n y > 0$. (from 5)

Case III: If the new variable has negative co-variance with the existing error – ie. when yest overshoots y, the variable is producing a low estimate and vica-versa $x_n err < 0$ and $x_n y > 0$. (from 6)

Case III replicates findings by Dietterich that ensemble methods should reduce error if the classifiers are correlated with the dependent but have uncorrelated errors [10].

A corollary of Case II is that the learner will be susceptible to error from variables that are strongly co-linear. This pushes the prediction to heavily reinforce the errors of the consensus variable.

However the negative effects of co-linearity are limited. Because of the weight constraint (1), the next theorem shows that the model has the useful property that the prediction error will be no worse than the prediction error from any one of the estimators. In other words, co-linear variables will just lead to a reinforcement of one of the predictors. This is an important property for robustness.

Theorem 3: Error for the model will be less than or equal to the error for the worst predictor x_n. $E(f(x_{1..i}, x_n), y) \leq E(f(x_n), y)$

$$f(x_{1..i}, x_n) = w_n x_n + \sum_i w_i x_i$$

If $(w_n x_n - y)^2 \leq (w_n x_n + \sum_i w_i x_i - y)^2$ then set $\forall i: w_i = 0$. Therefore we can at least equal the error for $f(x_n)$, and we may reduce the error further.

C. Discussion

Theorem 3 suggests that it is important to (1) limit the number of variables that are allowed into the ensemble since our bound on error is the worst variable we allow into the ensemble. Theorem 2 suggests that it is important to (2) combine variables that are not co-linear.

Colinearity (2) is reduced by using the Stacked Estimator framework to train the features. Variable participation (1) is limited due to (a) Participation thresholds which remove variables, (b) Missing value handling, which enables the system to elegantly operate with missing features, and (c) forward-backward selection which aggressively removes variables that do not make a significant contribution to the model.

VII. Experimental Results

The above system which we internally call "Scoring Service" has been progressively refined and improved from 2010 to 2014. As of 2014 the system has been used to purchase over a quarter of a million premium US television spots. The features used by the system have been progressively expanded, and in 2014 there are now about 1.2 million features being used for each advertiser (Figure 6). Experimental results on the performance of parts of the system in live television campaigns can be found in [20] and the system has been involved in advertising industry awards [19], [23], [26]. We would also note that not only ad response, but also many other variables needed for media targeting including Impressions, CPI and other quantities, are estimated using the same ensemble media asset pattern framework as presented here (each is referred to as a "target type").

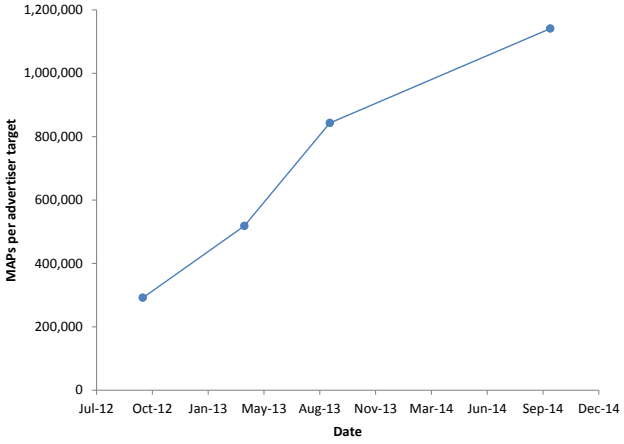


Fig.6. Features used by the system have grown over time.

VIII. Conclusion

We have provided a framework for describing TV ad targeting algorithms. We note that most algorithms can be defined as a choice of media asset pattern and ad effectiveness metric. This makes it straight-forward to incorporate different representations of television media when predicting performance of an upcoming airing. We have noted some limitations for some ad targeting algorithms such as Canning et. al.'s Buyer Ratings, specifically that match-rates can be very low. We have also shown that different ad targeting algorithms each carry particular weaknesses, yet they can be usefully combined to offset weaknesses in each method. The ensemble method presented here is designed to elegantly work with missing values and features with different intrinsic degrees of sparsity, and not only provides for improved prediction accuracy, but also enables a greater degree of robustness for real-world conditions.

IX. References

[1] J. Aaker, A. Brumbaugh, S. Grier, "Nontarget markets and Viewer Distinctiveness: The Impact of Target Marketing on Advertising Attitudes", Journal of Consumer Psychology, Vol. 9, No. 3, 2000. pp. 127-146. Lawrence Erlbaum Associates Inc.

[2] S. Balakrishnan, S. Chopra, D. Applegate, S. Urbanek, "Computational Television Advertising", Twelfth IEEE International Conference on Data Mining, 2012. pp. 71-80.

[3] B. Canning, P. Bochman and Z. Faro, "Using Consumer Purchase Behavior For Television Targeting", US Patent 8,060,398 B2, 2011. http://www.google.com/patents/US8060398

[4] R. Chang, R. Kauffman and I. Son, “Consumer micro-behavior and TV viewership patterns: data analytics for the two-way set-top box”, Proceedings of the 14th Annual International ACM Conference on Electronic Commerce, 2012. pp. 272-273, ACM New York

[5] D. Clack Authority Control: Principles, Applications, and Instructions, American Library Association, Chicago, 1990.

[6] T.F. Coleman and Y. Li, "A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables," SIAM Journal on Optimization, Vol. 6, Number 4, pp. 1040-1058, 1996.

[7] A. Cota and K. Dion, “Salience of gender and sex composition of adhoc groups: An experimental test of distinctiveness theory”, Journal of Personality and Social Psychology, Vol. 50, No. 4, 1986. pp. 770-776.

[8] C. Dawson, “Television Advertising: In Need of Reinvention?”, International Journal of Marketing, Vol. 16, No. 4. 1996.

[9] R. Deshpande and D. Stayman, “A tale of two cities: Distinctiveness theory and advertising effectiveness”, Journal of Marketing Research, Vol. 31, 1994. pp. 57-64.

[10] Dietterich, Ensemble methods in machine learning, 2000.

[11] Direct Response Academy, (2008), Advanced DRTV Media: Direct Response Television Media Management, Training Course materials on Direct Television media buying, http://www.directresponseacademy.com

[12] “In Praise of Television: The Great Survivor”, The Economist, April 29, 2010.

[13] N. Garramone, M. Papuga, H. Kent, J. Bergeron, “Reaching Audiences in an Increasingly Fragmented TV World”, Audience Measurement 7, NCC Media, 2012.

[14] D. Hallerman, “US Television Ad Spend: Factors Shaping Today’s Television Market”, March 2014, eMarketer.

[15] D. Hanssens, L. Parsons, R. Schultz, “Market Response Models: Econometric and Time Series Analysis”, Kluwer Academic Press, Boston. 2001.

[16] Internet Advertising Bureau, IAB Platform Status Report: Interactive Television Advertising, 2012.

[17] J.K. Johansson, “Advertising and the S-Curve: A New Approach”, Journal of Marketing Research, Vol. 16, No. 3, 1979. pp. 346-354. American Marketing Association.

[18] J. Jones, “What Does Effective Frequency Mean in 1997?”, Journal of Advertising Research, July, 1997. pp. 14-20.

[19] B. Kitts, R. Brooks, T. Roberts, J. Abdo, T. Duncan, J. Perry, R. Tate, C. Taylor, P. Tellefsen, C. Krebs, S. Vacanti, S. Zlomek, S. Giusti, T. Box, A. Chun, D. Au, J. Wadsworth-Drake, "Rockwell Tools: Breaking Sales Records at Rockwell Tools", Ogilvy Award Winning Case Study, (etail/retail category - Silver winner), 2012, The Audience Research Foundation

[20] B. Kitts, D. Au, B. Burdick, “A High-Dimensional Set Top Box Ad Targeting Algorithm including Experimental Comparisons to Traditional TV Algorithms”, Proceedings of the Thirteenth IEEE International Conference on Data Mining (ICDM 2013), December, 2013. Dallas, TX, IEEE Press.

[21] B. Kitts, M. Bardaro, D. Au, A. Lee, S. Lee, J. Borchardt, C. Schwartz, J. Sobieski, J. Wadsworth-Drake, “Can Television Advertising Impact Be Measured on the Web? Web Spike Response as a Possible Conversion Tracking System for Television”, Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, Aug 24, 2014, New York City, held at the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Aug 24-27, 2014, New York City.

[22] B. Kitts, D. Au, S. Ulger, “How Well Does Addressable Targeting Work on Television?”, First Workshop on Recommender Systems for Television and online Video (RecSysTV), held at the Eighth ACM Conference on Recommender Systems 2014, Foster City, Silicon Valley, USA from 6th-10th October 2014

[23] B. Kitts, J. Wadsworth-Drake, W. Vollmann, I. Ross, G. Martin, D. Tjen, D. Au, S. Zlomek, A. Chun, J. Sobieski, S. Giusti, M. Lyons, J. Harris, I. Kovalik, B. Perkins, S. Smith, M. Hill, A. Boyarsky, E. Morse, "Art.com: Find Your Art. Love Your Space.", 2014 David Ogilvy Award Winning Case Study, 2014. The Advertising Research Foundation.

[24] H. Krugman, “Why Three Exposures May Be Enough”, Journal of Advertising Research, December, 1972. pp. 11-14.

[25] C.L. Lawson, R.J. Hanson, Solving Least-Squares Problems, Prentice-Hall, Chapter 23, p. 161, 1974.

[26] A. Lee, S. Lee, S. Giusti, B. Kitts, D. Au, J. Shepard, L. Moore, S. Zlomek, C. Schwartz, J. Crim, M. Nelson, K. Lesinski, N. Rongisch, M. Carstens, J. Bangs, B. Pyle, P. Nygard, M. Ebeling, "Physicians Mutual Insurance Company: Insurance For All Of Us”, David Ogilvy Award Winning Case Study 2014, The Advertising Research Foundation.

[27] Nielsen Corporation, “A2/M2 Three Screen Report: Television, Internet and Mobile Usage in the United States”, 4th Quarter 2008.

[28] Nielsen Corporation, “Americans watching more TV than ever”, Nielsen Wire, May, 2009. http://blog.nielsen.com/nielsenwire/online_mobile/americans-watching-more-tv-than-ever/

[29] Nielsen Corporation, “The Cross-Platform Report”, 2014. http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2014%20Reports/nielsen-cross-platform-report-june-2014.pdf

[30] M. Schneider, “Fox wants answers from Nielsen”, Variety, May 18, 2009. http://www.variety.com/article/VR1118003924?refCatId=14 ,

[31] A. Segal, “Nielsen Ratings: An Inaccurate Truth Out of date television ratings system exposed”, The Cornell Daily Sun, April 26, 2007, http://cornellsun.com/node/23180.

[32] H. Simon, “Price Management”, North-Holland Publishing Company, Amsterdam. 1989.

[33] J.L. Simon and J. Arndt, “The Shape of the Advertising Response Function”, Journal of Advertising Research, Volume 20, Number 4, 2002. pp. 767.

[34] G. Tellis, R. Chandy, R, D. MacInnis, P. Thaivanich. “Modeling the Microeffects of Television Advertising: Which Ad Works, When, Where, for How Long, and Why?”, Marketing Science, Marketing Science 24(3), 2005. pp. 351–366, INFORMS. http://www-rcf.usc.edu/~tellis/AdMicro.pdf

[35] T. Whittler, “The effects of actors’ race in commercial advertising: Review and extension”, Journal of Advertising, Vol. 20, 1989. pp. 54-60.

[36] J. Williams and W. Qualls, “Middle-class Black consumers and intensity of ethnic identification”, Psychology and Marketing, Vol. 6, 1989. pp. 263-286.

[37] D.H. Wolpert, “Stacked Generalization”, Neural Networks, 5, 241-259, 1992.

[38] D. Vakratsas, F. Feinberg, F. Bass and G. Kalyanaram, Advertising Response Functions Revisited, Marketing Science, 23(1). 2004. pp. 109–119.

[39] P. Kotler, G. Armstrong and R. Starr, Principles of marketing, Englewood Cliffs, NJ: Prentice Hall. 1991.

[40] H.E. Krugman, “Why Three Exposures May Be Enough.” Journal of Advertising Research 12, 6. 1972, pp. 11-14.

[41] L. Cowen, “Welcome to TV’s Second Golden Age”, CBS News, October 1, 2014.

[42] A. Kirsch and M. Hamid, “Are the new ‘Golden Age’ TV Shows the New Novels?”, New York Times, February 25, 2014.

[43] B. Kitts, D. Au, M. Bardaro, S. Lee, “A Machine Learning Approach to Forecasting Television Viewing”, Working Paper, Unpublished.

[44] Randall, “The Crazy Nastyass Honey Badger”, You Tube video, https://www.youtube.com/watch?v=4r7wHMg5Yjg video derived from original program on WILD Discovery, Honeybadgers.

[45] Wikipedia, “The Crazy Nastyass Honey Badger”, Wikipedia entry accessed Oct 27, 2014. http://en.wikipedia.org/wiki/The_Crazy_Nastyass_Honey_Badger