Measuring Advertising Quality on Television

Deriving Meaningful Metrics from Audience Retention Data

DAN ZIGMOND
Google, Inc.
djz@google.com

SUNDAR DORAI-RAJ
Google, Inc.
sdorairaj@google.com

YANNET INTERIAN
Google, Inc.
yannet@google.com

IGOR NAVERNIOUK
Google, Inc.
abednego@google.com

This article introduces a measure of television ad quality based on audience retention using logistic regression techniques to normalize such scores against expected audience behavior. By adjusting for features such as time of day, network, recent user behavior, and household demographics, we are able to isolate ad quality from these extraneous factors. We introduce the model used in the current Google TV Ads product and two new competing models that show some improvement. We also devise metrics for calculating a model's predictive power and variance, allowing us to determine which of our models performs best. We conclude with discussions of retention score applications for advertisers to evaluate their ad strategies and as a potential aid in future ad pricing.

INTRODUCTION

In recent years, there has been an explosion of interest in collecting and analyzing television set-top box (STB) data (also called "return-path" data) (Bachman 2009). As U.S. television moves from analog to digital signals, digital STBs increasingly are common in American homes. Where these are attached to some sort of return path (as is the case in many homes subscribing to cable or satellite TV services), these data can be aggregated and licensed to companies wishing to measure television viewership.

Advances in distributed computing make it feasible to analyze these data on a massive scale. Whereas previous television measurement relied on panels consisting of thousands of households, data can now be collected and analyzed for millions of households. This holds the promise of providing accurate measurement for much (and perhaps all) of the niche TV content that eludes current panel-based methods in many countries.

In addition to using these data for raw audience measurement, it should be possible to make more qualitative judgments about the content—and specifically the advertising—on television. In much the same way that online advertisers frequently measure their success through user-response metrics such as click-through rate (CTR), conversion rate (Richardson, Dominowski, and Ragno, 2007), and bounce rate (Sculley, Malkin, Basu, and Bayardo, 2009), Google has been exploring how to use STB measurement to design equivalent measures for TV.

Past attempts to provide quality scores for TV ads have typically relied on smaller constructed panels and focused on programming with very large audiences. For example, for the 2009 Super Bowl, Nielsen published likeability and recall scores for the top ads (Nielsen Inc., 2009). The scores were computed using 11,466 surveys, and they reported on the five best-liked ads and the five most-recalled ads.

In this article, we define a rigorous measure of audience retention for TV ads that can be used to predict future audience response for a much larger range of ads. The primary challenge in designing such a measure is that many factors appear to impact STB tuning during ads, making it difficult to isolate the effect of the specific ad itself on the probability that a STB will tune away. We propose several ways of modeling such a probability. To the best of our knowledge, this is the first to attempt

DOI: 10.2501/S0021849909091090

to derive a measure of TV ad quality from large-scale STB data.

SECOND-BY-SECOND MEASUREMENT

Google aggregates data—collected and anonymized by DISH Network LLC—describing the precise second-by-second tuning behavior television STBs in millions of U. S. households.[1] These data can be combined with detailed airing logs for thousands of TV ads to estimate second-by-second fluctuations in audience during TV commercials everyday.

For example, audiences fluctuate during a typical commercial break on a major U. S. cable television network (as shown in Figure 1). The total estimated audience drops by approximately 5 percent soon after the ads begin at 8:19 AM (shown by a hollow dot). Google inserted ads at approximately 1 minute into this break (shown by the shaded area), during which there was a slight net increase in the total audience. After Google's ads, the regular programming resumed, and the audience size gradually returned to nearly the prior levels within the first two minutes.

The lower plot shows the level of tuning activity across this same timeline. Tune-away events (solid line) peak at the start of the break, whereas tune-in (dashed line) is strongest once the programming resumes. Smaller peaks of tune-away events also occur at the start of the Google-inserted ads.

TUNING METRICS

These raw data can be used to create more refined metrics of audience retention, which in turn can be used to gauge how appealing and relevant commercials appear to be to TV viewers. One such measure is the percentage initial audience retained (IAR): how much of the audience that was tuned to an ad when it began airing remained tuned to the same channel when the ad completed.

In many respects, IAR is the inverse of online measures like CTR. For online ads, passivity is negative: Advertisers want users to click through. This is somewhat reversed in television advertising, in which the primary action a user can take is a negative one: to change the channel. We see broad similarities, however, in the propensity of users to take action in response to both types of advertising (see Figure 2). In January 2009, the tune-away rates (the additive inverse of IAR) for 182,801 TV ads distribution was broadly similar to the distribution of CTR for a comparable number of randomly selected paid search ads that also ran that month. Although the actions being taken are quite different in the two media, the two measures show a comparable range and variance.

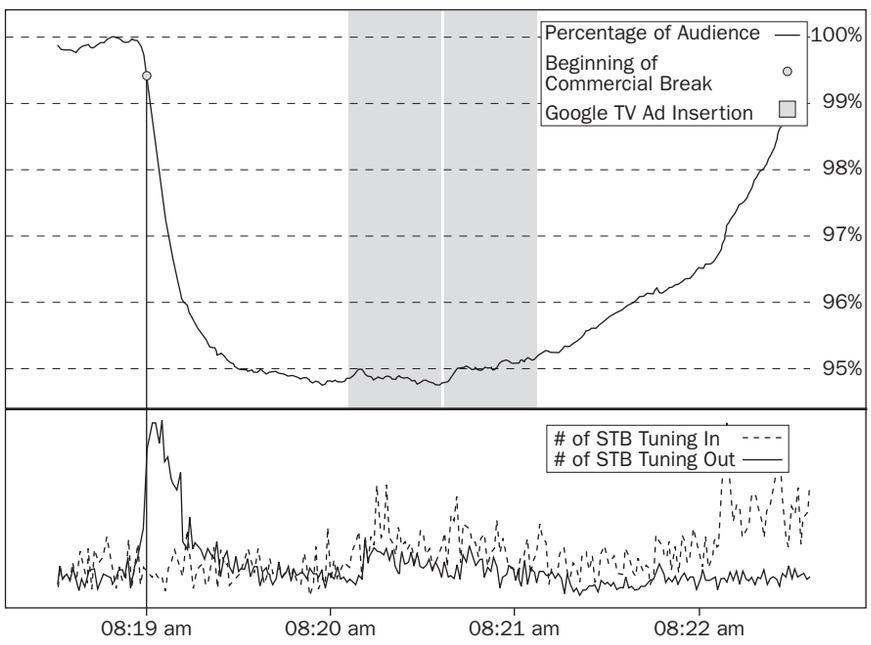
THE BASIC MODEL

Tuning metrics, like IAR, can be useful in evaluating TV ads. We have found, however, that these metrics are highly influenced by extraneous factors such as the time of day, the day of the week, and the network on which the ads were aired. These are nuisance variables and make direct comparison of IAR scores very difficult. Rather than using these scores directly, we have developed a model for normalizing the scores relative to expected tuning behavior.

Definition

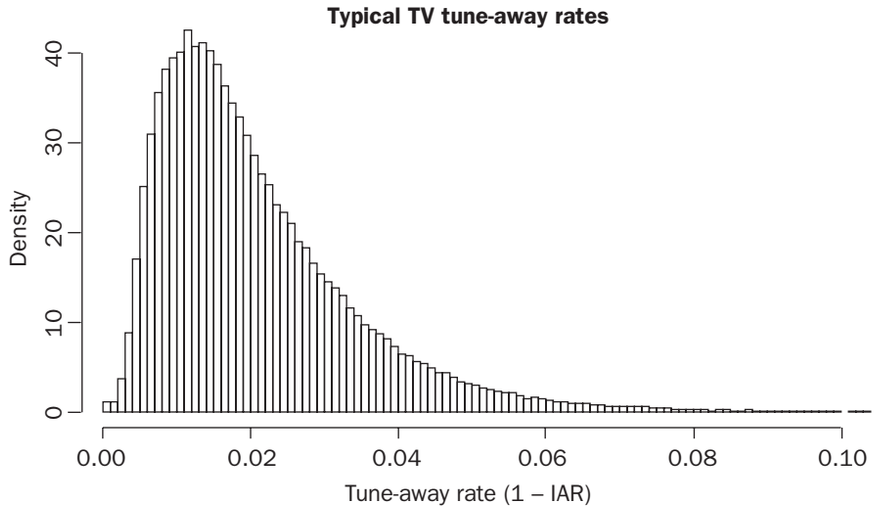
We calculate per airing the fraction of IAR during a commercial. This is calculated by taking the number of TVs tuned to an ad when it began and then remained tuned throughout the ad airing (see Equation 1).

[1]*These anonymous STB data were provided to Google under a license by the DISH Network LLC.*



Note: The number of viewers drops roughly 5% after the advertising break starts (top plot). The number of tune-out events (solid line; bottom plot) is strongly correlated with the beginning of the pod (i.e., advertising break). Toward the end of the pod, we also see an increase of tune-in events (dashed-line; bottom plot).

Figure 1 Pod Graph of STB Tune-In/Out Events on a Major Network



182,801 Google TV ads with at least 1,000 impression, January 2009
Note: For most ads, roughly 1% to 3% of the viewers at the beginning of the ad tune away before the end of the ad.

Figure 2 Tune-Away Rate Distribution for TV Ads

When an ad does not appeal to a certain audience, those viewers will vote against it by changing the channel. By including only those viewers who were present when the commercial started, we hope to exclude some who may simply be channel surfing.

$$\text{IAR} = \frac{\text{Audience that viewed whole ad}}{\text{Audience at beginning of the ad}} \quad (1)$$

We can interpret IAR as a probability of tuning out from an ad. However, as explained, raw, per-airing IAR values are difficult to work with because they are affected by the network, day part, and day of the week, among other factors. To isolate these factors from the creative (ad), we define *Expected IAR* of an airing (see Equation 2):

$$\hat{\text{IAR}} = \text{E}(\text{IAR} \mid \hat{\theta}), \quad (2)$$

where $\hat{\theta}$ is a vector of features extracted from an airing, which exclude any features that identify the creative itself; for example, hour of the day and TV network are included but not the specific campaign or customer. We then define the *IAR residual* to be a measure of the creative effect (see Equation 3).

$$\text{IAR residual} = \text{IAR} - \hat{\text{IAR}} \quad (3)$$

There are a number of ways to estimate (2), several of which will be discussed in this article.

Using equation 3, we can define *underperforming airings* as the airings with IAR residual below the median. Now that we have a notion of underperforming airings, we can formally define the *retention score* (RS) for each creative as one minus the fraction of airings that are underperforming in Equation 3 (see Equation 4).

$$\text{RS} = 1 - \frac{\text{Number of underperforming airings}}{\text{Total number of airings}} \quad (4)$$

The Basic Model

The basic model we currently use to predict expected IAR (IÂR) is a logistic regression of the following form:

$$\ln\left(\frac{\text{IAR}}{1-\text{IAR}}\right) \sim \text{“Network”} + \text{“Ad Duration”} + \text{“WeekDay”} + \text{“DayPart”} \quad (5)$$

where IAR is given by (1) and each feature on the right hand side is a collection of parameters. Here, “Network”, “WeekDay”, and “DayPart” are categorical variables, whereas “Ad Duration” is treated as numeric.

Parameter estimates for (1) are obtained using the glmnet package in R (Friedman, Hastie, and Tibshirani, 2009). The glmnet algorithm shrinks insignificant or correlated parameters to zero using an L1 penalty on the parameter estimates. This avoids the pitfalls of classic variable selection, such as stepwise regression.

Retention Score and Viewer Satisfaction

To understand the qualitative meaning of retention scores, we conducted a simple survey of 78 Google employees. We asked each member of this admittedly unrepresentative sample to evaluate 20 television ads on a scale of 1 to 5, where 1 was “annoying” and 5 was “enjoyable.” We chose these 20 test ads such that 10 of them were considered “bad” and the remaining 10 were considered “good” (see Table 1).

Ads that scored at least “somewhat enjoyable” (i.e., mean survey score greater than 3.5) had an average retention score of 0.86 for all creatives (see Table 2). Ads that scored at the other end of the spectrum (mean less than 2.5) had an average

TABLE 1
Using Retention Scores to Categorize Ads as Either “Bad” or “Good”

Ad Quality	Retention Score
“Good”	>0.75
“Bad”	<0.25

These categories were matched empirically with a human evaluation survey

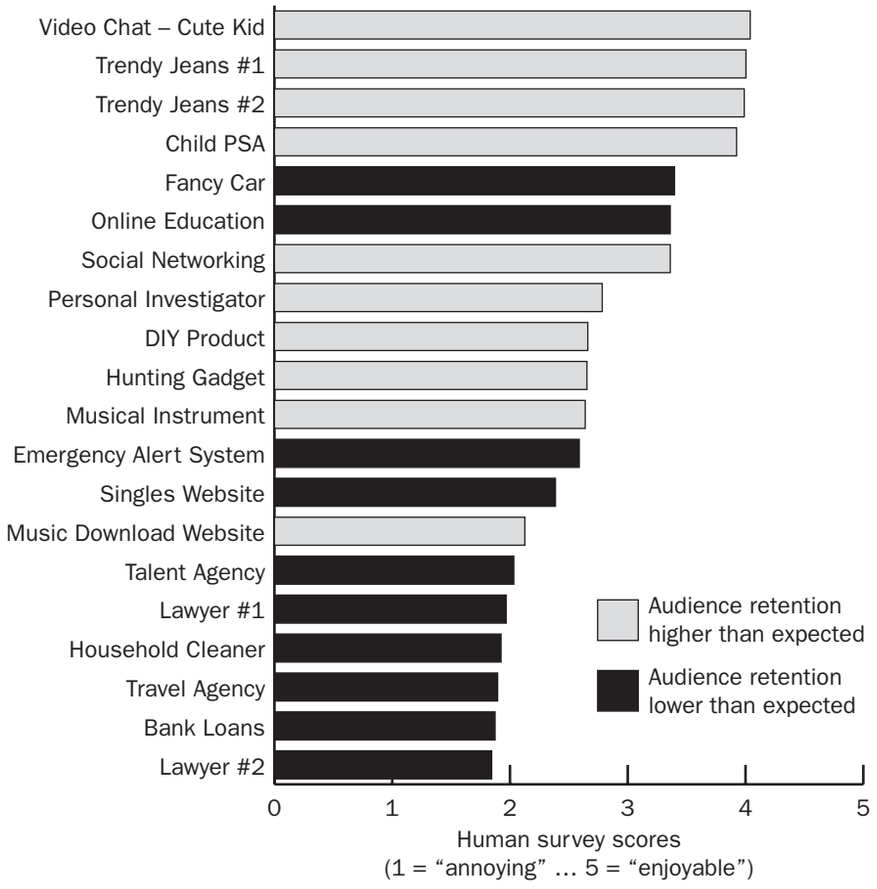
TABLE 2
Correlating Retention Score Rankings with Human Evaluations

Human Evaluation	Mean RS
At least "somewhat engaging"	0.86
"Unremarkable"	0.62
At least "somewhat annoying"	0.30

Survey scores of 3.5 or above (or "somewhat engaging") received retention scores averaging 0.86, whereas survey scores of 2.5 or below (or "somewhat annoying") received retention scores averaging 0.30. These numbers match well with categories defined in Table 1.

retention score of 0.30. Ads with survey scores in between these two had an average retention score of 0.62. These results suggest our scoring algorithm and the categories defined in Table 1 correlate well with how a human being might rank an advertisement.

In another view of these data (see Figure 3), the 20 ads are ranked according to their human evaluation, with the highest-scoring ads on top. The bars are colored according to which set of 10 they belonged, with gray ads coming from the group that consistently outperformed the model and black ads coming from the group that underperformed. Although the correlation is far from perfect, we see fairly good separation of the "good" and "bad" ads, with the highest survey scores tending to go the ads with the best retention scores.



Note: The length of the bar represents the average of the scores given by the 20 respondents. The light gray bars correspond to ads with "good" ad quality, and dark bars correspond to ads with "bad" quality, as determined in Table 1. Though the correlation between retention scores and the human evaluation is not perfect (i.e., black bars receive lower scores than gray), a prominent relationship is very visible in this small study.

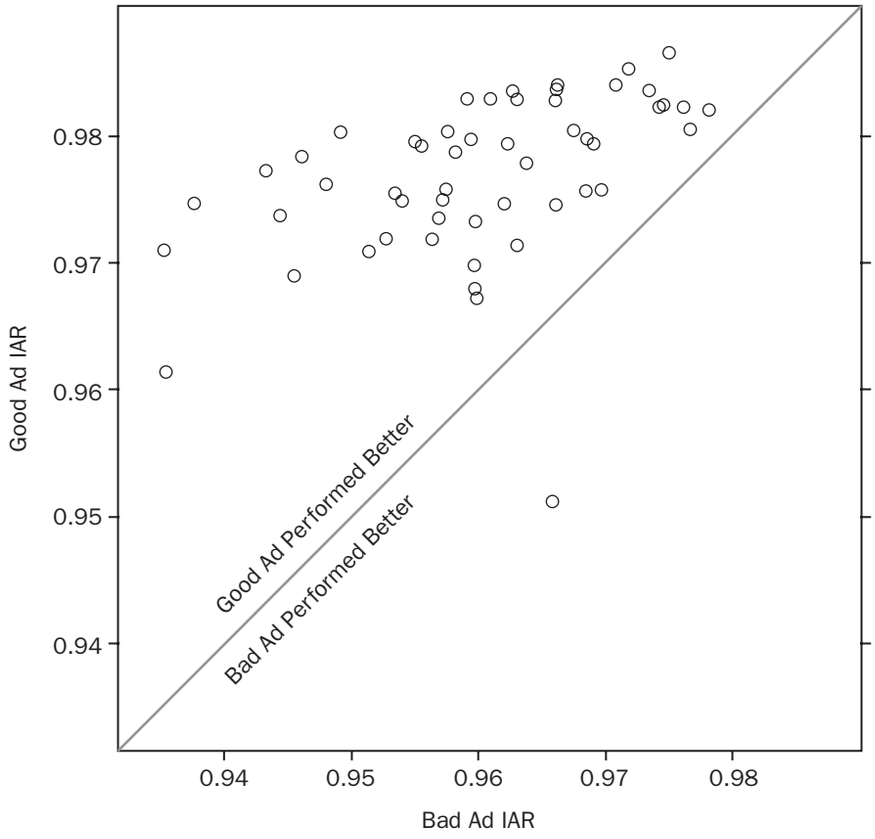
Figure 3 Correlating Retention Score Rankings with Human Evaluations

Live Experiments and Model Validation

To test the validity of our model further, we ran several live experiments. In these experiments, we identified two ads: one with a high retention score and one with a low score. We then placed the two ads side by side, in a randomized order, on several networks. Placing ads in the same commercial break or pod ensured most other known extraneous features (e.g., time of day, network) were neutralized, so comparisons made between the ads would be fair (see Figure 4 for our first such experiment, conducted in 2008).

After running the ad pairs for about a week, we determined whether the retention scores were an accurate predictor of which ad would retain a larger percentage of the audience by observing how often the ad with higher retention score had the larger IAR. In this case, the prediction was nearly perfect, with only one pair incorrectly ordered.

The purpose of running these live experiments was to determine the accuracy of our retention score model. Ad pairs with little difference in retention score (e.g., <0.1) will be virtually indistinguishable in terms of relative audience retention. Conversely, pairs with large differences in retention score (e.g., >0.7) should almost always have higher audience retention associated with the ad with the higher score. To test our retention score's ability to sort a wider range of ads, we produced a plot that relates our predicted retention scores back to the raw data (see Figure 5)—a qualitative method of determining how well our retention scores



Note: Each point represents the IAR of the good ad versus the bad ad. Only one of the pairs had the IAR of the bad ad greater than the IAR of the good ad. We randomized each pair to determine which ad comes first, so there is no pod order bias.

Figure 4 Results from a Live Experiment in 2008

actually sorts creative, both in the structured experiments described earlier and in ordinary airings. As expected, the difference in retention score is proportional to the likelihood of the higher-scoring ad retaining more audience.

We currently have three competing models for obtaining retention scores. All three models use IAR as a response in a logistic regression. They differ, though, either in their lists of features or the type of regularization used to prevent overfitting.

- Basic Model: Estimates IAR using network, weekday, daypart, and ad duration as main effects in a logistic regression model.
- User-Behavior Model: Same as basic model but incorporates behavior of the TV viewer 1 hour prior to an airing. More specifically, we count the number of tune-out events the hour prior to the ad and whether there was a tune-out event in the previous 10 minutes or previous 1 minute before the ad airs. These additional tune-out measures attempt to separate active users (i.e., more likely to tune away) from passive.
- Demographics Model: Same as basic model but splits households according to 113 demographic groups.

As noted in the Basic Model section (previously), we use glmnet to obtain parameter estimates for the basic model. For the user-behavior model, we also apply the same algorithm. For the demographics model, however, we employ a slightly different type of regularization by using principal components logistic regression (PCLR) (Aguilera, Escabias, and Valderrama, 2006). PCLR allows for highly correlated parameters in the model, in this case demographic group and network.

The data we are using to compare the three models are from June 2009. For the sake of brevity, we limit ourselves to the 25 networks with the highest median viewership during that month. This leads to a dataset containing 38,302 ads from which we build our models.

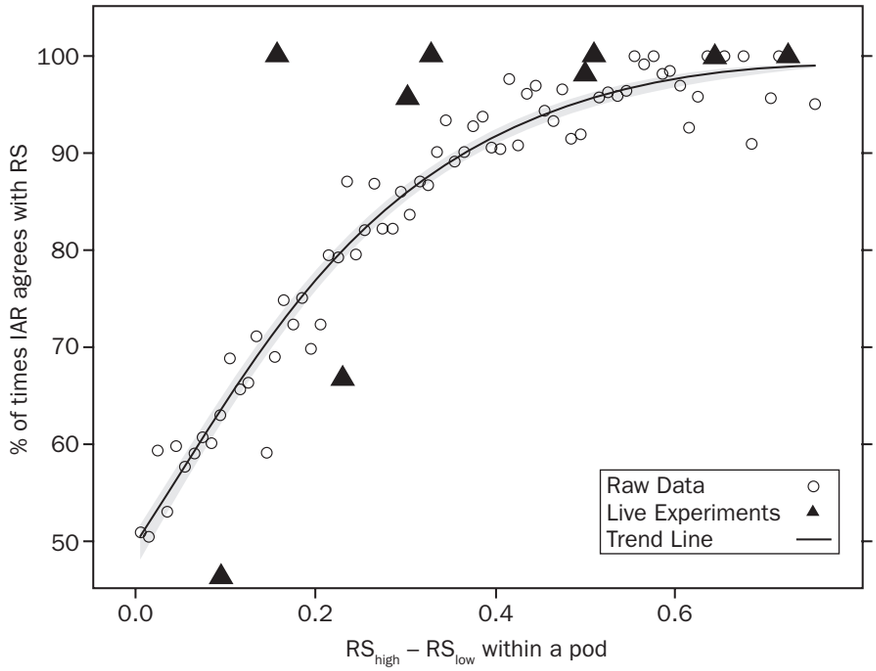
User Behavior

For a typical ad, one to three percent of viewers present at the beginning of the ad tune out before the end of the ad (Interian et al., 2009). The User-Behavior Model adjusts IAR by splitting the audience base into active and passive groups. Our hypothesis is that active users are more likely to tune out from an ad they do not like, whereas passive users will watch anything regardless of the creative. In fact, active users typically have a much lower IAR than passive users (see Figure 6).

By adding to our model parameters that capture recent tuning behavior for every STB, we are able to predict more accurately when a viewer will tune out during an ad. The variance of active users is much higher than passive users, simply because we have observed IAR further from the upper bound of one (see Figure 6, right panel). This increased variance in the response improves our model and provides less noisy predictions of IAR.

Demographic Groups

Like the users behavior model described earlier, we also believe different demographic groups react differently to ads. For



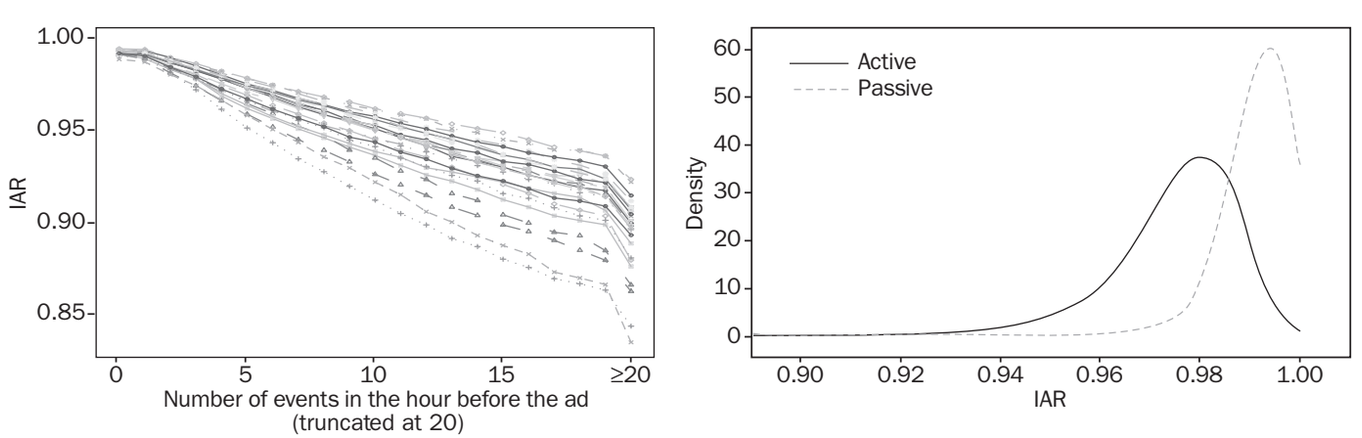
Note: Each point (circle) represents the percentage of times that two ads within the same pod (ad break) agree with their respective retention scores. For example, of all ad pairs that have an approximately 0.2 difference in retention scores, roughly 70% of those pairs have the IAR of the lower-ranked ad smaller than the IAR of the higher-ranked ad. We superimposed our live experiments onto the plot (triangles) to show a general agreement with the trend.

Figure 5 Figure Demonstrating the Predictive Power of Our Retention Score Model

example, in an IAR comparison for single men versus single women, almost regardless of the creative, women tend to tune away less than men (see Figure 7).

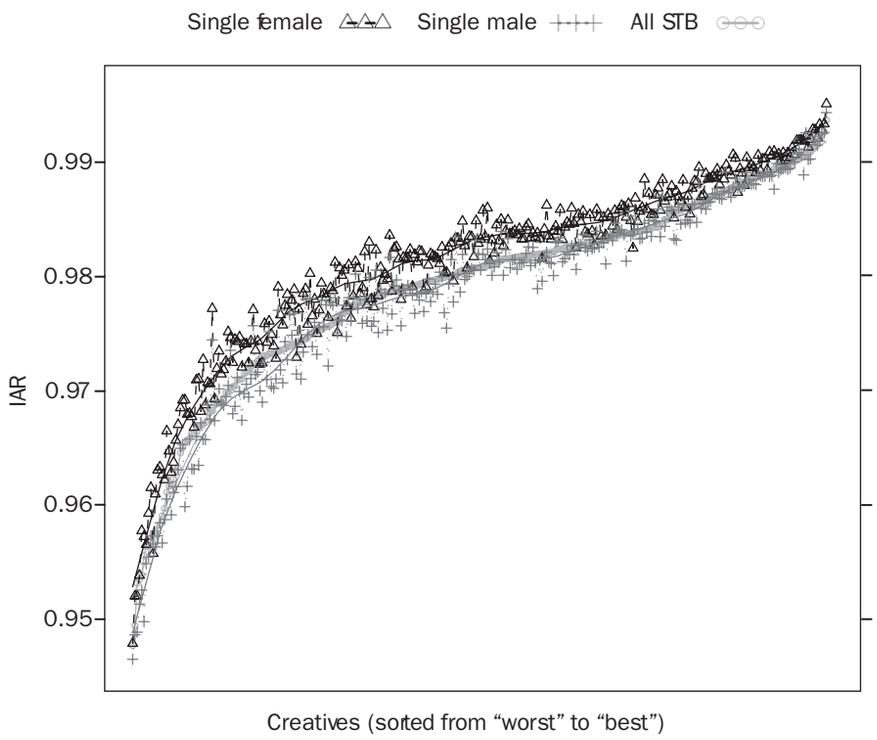
For our demographics model, we include gender of adults, presence of children, marital and cohabitation status, and age of oldest adult as additional features. These categories were identified by an internal data-mining project, which ranked demographic groups according their relative impact on IAR. Other demographics, such as number of adults present and declared interest in sports TV, have promise for improving our model. The make up of the included features is described in Table 3.

We also have found that certain demographics are a partial proxy for network. For example, older adults watch more cable news networks whereas households with children have higher viewership of children's networks. This observation suggests that network, one of the features in our Basic Model, might offer redundant information provided more succinctly by demographics. Including demographic



Note: The left plot displays the aggregated IAR from 0 events prior to an ad (highest IAR) to 20 or more events prior to an ad (lowest IAR). Each line represents one of 25 networks used in this study. The right plot shows density functions of IAR for airings in June 2009, split by active (solid) and passive (dashed) users. The IAR for active users has a much larger variance because viewers in this group are more likely to change channels during an ad.

Figure 6 Active Users (i.e. Viewers Who Changed Channels Within 10 Minutes Prior to the Ad) Are More Likely Than Passive Users to Tune Away from an Ad



Note: The creatives (x axis) are sorted from "best" (low IAR) to "worst" (high IAR) across all demographics. The IAR for single women with no children (black triangles) tends to be higher than the IAR for single men with no children (gray crosses), with very few exceptions. IAR for all STBs (including single men and women) tends to be between the two.

Figure 7 Average IAR for Creatives in June 2009

TABLE 3

Demographic Groups Measured for Each Household

Gender	Kids	Married	Single	Age
Male	Yes	Yes	Yes	18–24
Female	No	No	No	25–34
Both	Unknown	Unknown		35–44
Unknown				45–54
				55–64
				65–74
				75+

This table describes 113 possible groups, including groups where the demographic was not measured. Note that Single is not the opposite of Married; Single implies no other adult living in the household, so two cohabitating adults are both not Single and not Married.

information in the same feature list as network may, therefore, lead to over-parameterization of the model.

Redundancy in the network viewership and demographic groups lead to colinearities in our model formulation. Fitting a model with known correlations will lead to misleading parameter estimates (Myers, 1990). To overcome these problems, we use PCLR as an alternative to glmnet. With PCLR, we have more control over the model with respect to known correlations.

For the data discussed in this article, the demographics model contains 144 possible parameters, including the intercept, 112 parameters from the demographic groups, and remaining parameters from network, daypart, and weekday differences. In PCLR, we drop the principal components with little variation, in this case the last 44 dimensions. This leaves us estimating only 100 parameters and thus greatly reducing the complexity of the model. All further comparisons of the demographics model in the next section are based on the first 100 principal components.

COMPARING MODELS

We have devised four metrics to describe the quality of the models we described in the previous sections. Although these metrics tend to agree in ranking models, each measures a different and important aspect of a model's performance.

Dispersion

The dispersion parameter in logistic regression acts as a goodness-of-fit measure by comparing the variation in the data to the variation explained by the model (McCullullagh and Nelder, 1989) (see Equation 6). The formula for dispersion is given by:

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i - n_i\hat{y}_i)^2}{n_i\hat{y}_i(1-\hat{y}_i)}, \qquad (6)$$

where N is the number of observations, p is the number of parameters fit in our model, y_i is the observed IAR, $\hat{y}_i$ is the expected IAR from our model, and n_i is the number of viewers at the beginning of an ad. The closer equation 6 is to 1, the better the fit.

Captured Variance

A reasonable model should minimize the variance within a creative while maximizing the variance between creatives. Using the residuals r given by (3), "captured variance" is given by

$$\frac{\mathrm{E}(\mathrm{Var}_c(r))}{\mathrm{Var}(\mathrm{E}_c(r))}, \tag{7}$$

where Var_c and E_c are the variance and expectation of residuals within a creative c. The expression in (7) should be small for better models. Or more specifically, a good model will have small residual variance within creatives (numerator) and a large residual variance between creatives (denominator).

Predictive Strength

Predictive strength compares models through their respective retention scores. Figure 8 shows the predictive strength for the basic model. In this plot, we see that as the differences in retention score increase, the respective ads also agree in terms of IAR. So that comparisons of IAR are fair (i.e., extraneous variables are minimized), each ad pair considered is within a pod.

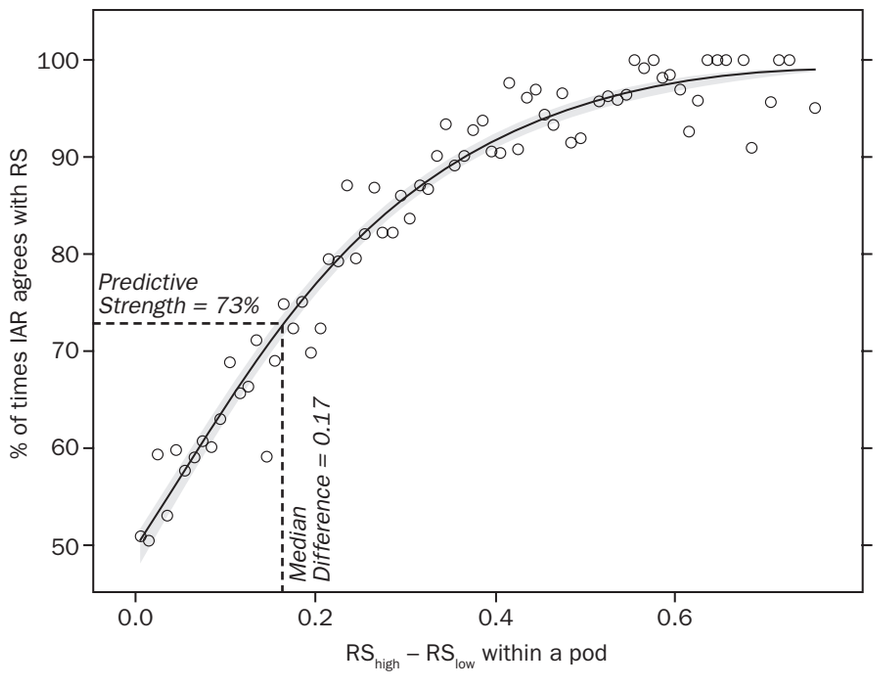
To compute the metric, we draw a curve through the scatter plot using logistic regression. From the fitted line, we determine the point on the y-axis that corresponds to the median of all retention score differences. The larger the predictive strength (i.e., the steeper the curve), the better the model is at sorting ads that are relatively close together in terms of retention scores.

Residual Permutation

For the last metric, we randomly reorder the residuals from our model and recalculate the retention scores according to (4). We then measure the area between the distributions of the new retention scores and the observed retention scores. The result is interpreted as the difference between determining scores using our model and selecting scores at random. The greater the difference, the better our model is at producing scores that do not look random (see Figure 9).

Model Comparisons

Using the four defined metrics from the previous sections, we compute the relative differences between our three models (see Table 4). The user model is the best according to the metrics we have defined, followed by the demographics model and the basic model. The greatest improvements over the basic model thus far have been in dispersion, whereas predictive strength has only slightly improved.



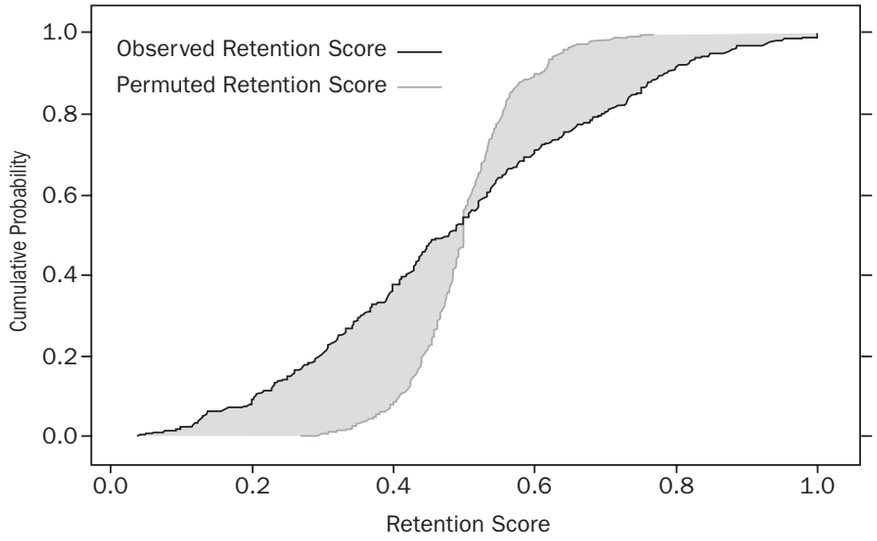
Note: The curve is identical to Figure 5 (with the live experiments removed). We first use the curve to determine the median difference in retention scores of all ad pairs. From this difference, and using the logistic regression trend line, we estimate the percentage of ad pairs for which the aggregated IAR agrees with the retention score difference. For the basic model (pictured), the median difference is 0.17, of which 73% of the ad pairs agree with the retention score ordering.

Figure 8 Predictive Strength Is Computed from the Curve Above

TABLE 4 Comparison Metrics for Each Model

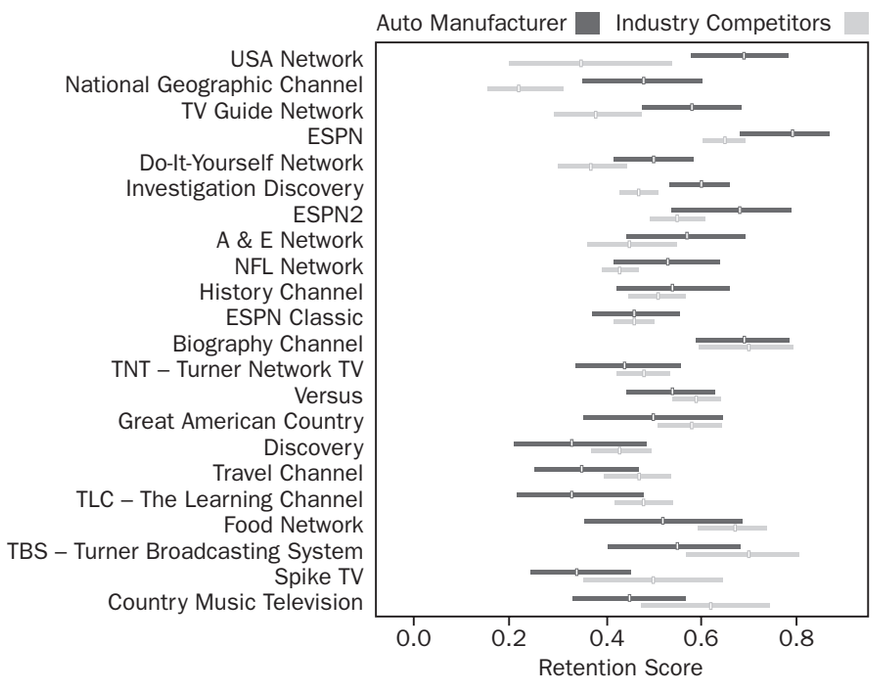
	Basic	User	Demographics
Dispersion	41.8	3.2	7.5
Captured variance	5.2	4.1	3.9
Predictive strength	73%	75%	70%
Permutation	43.9	53.6	50.8

The model with the best metric is shaded in gray. For three of the four models, the user behavior model wins the comparisons. The demographics model is second or first for three of the four comparisons. The basic model fairs the worst among the three.



Note: The shaded area between the two distributions is our permutation metric, which should be large for better models. The distribution of retention scores shown in the figure is from the basic model.

Figure 9 Empirical Distribution of Observed Retention Scores Versus the Scores Determined from Permuting the Model Residuals



Note: Some networks have better scores than others, which provide important feedback to the advertisers. The length of the bar represents a 90% confidence interval on the score.

Figure 10 Retention Scores for Ads Run by an Auto Manufacturer to Their Competitors' Ads

POSSIBLE APPLICATIONS

We have started using retention scores for a variety of applications at Google. These scores are made available to advertisers, who can use them to evaluate how well their campaigns are retaining audience. This may be a useful proxy for the relevance of their ads in specific settings.

For example, Figure 10 shows the retention scores for an automotive advertiser, compared with the average scores for other automotive companies advertising on television. Separate scores were calculated for each network on which this advertiser aired. We can see not only significant differences in the retention scores for these ads but differences in the relative scores compared against the industry average. On the National Geographic Channel, for example, this advertiser's retention scores exceed those of the industry average by a significant margin. On County Music Channel, this advertiser's scores are lower than the industry average, although there is substantial overlap of the 90 percent confidence intervals. This sort of analysis can be used to suggest ad placements where viewers seem to be more receptive to a given ad.

Audience loss during an ad can also be treated as an economic externality, because it denies viewers to later advertisers and potentially annoys viewers. Taking this factor into account might yield a more efficient allocation of inventory to advertisers (Kempe and Wilbur, 2009), and might create a more enjoyable experience for TV viewers.

CONCLUSIONS AND FUTURE WORK

The availability of tuning data from millions of STBs, combined with advances in distributed computing that make analysis of such data commercially feasible, allows us to understand for the first time the factors that influence television tuning behavior. By analyzing the tuning behavior of millions of individuals across many

A reasonable model should minimize the variance within a creative while maximizing the variance between creatives.

thousands of ads, we can model specific factors and derive an estimate of the tuning attributable to a specific creative. This work confirms that creatives themselves do influence audience viewing behavior in a measurable way.

We have shown three possible models for estimating this creative effect. The resulting scores—the deviation of an ad audience from the expected behavior—can be used to rank ads by their appeal, and perhaps relevance, to viewers and could ultimately allow us to target advertising to a receptive audience much more precisely. We have developed metrics for comparing the models themselves, which should help ensure a steady improvement as we continue experimenting with additional data and new statistical techniques. We hope in the future to incorporate data from additional television service operators and to apply similar techniques to other methods of video ad delivery. We would also like to expand the small internal survey we conducted into a more robust human evaluation of our scoring results.

In the long run, we hope this new style of metric will inspire and encourage better and more relevant advertising on television. Advertisers can use retention scores to evaluate how campaigns are resonating with customers. Networks and other programmers can use these same scores to inform ad placement and pricing. Most important, viewers can continue voting their ad preferences with ordinary remote controls—and using these statistical techniques, we can finally count their votes and use the results to create a more rewarding viewing experience. JAR

ACKNOWLEDGEMENTS

The authors thank Dish Network for providing the raw data that made this work possible and in particular Steve Lanning, Vice President for Analytics, for his helpful feedback and support. They also thank P. J. Opalinski, who helped us obtain the data disc used in this paper. Finally, they thank Kaustuv who inspired much of this work when he was part of the Google TV Ads team.

Dan Zigmond is manager of Google's TV Ad Effectiveness and Pricing group and the founder of the Google TV Ads engineering team. He holds a BA in computational neuroscience from the University of Pennsylvania.

Sundar Dorai-Raj is a senior quantitative analyst at Google. His areas of interests include linear models and statistical computing. He has a Ph.D. in statistics from Virginia Tech.

Yannet Interian is a quantitative analyst at Google specializing in data mining. She has a Ph.D. in applied mathematics from Cornell.

Igor Navernioux is a software engineer at Google. His work includes distributed computing and machine learning. He has an MSc in computer science from the University of British Columbia.

REFERENCES

Aguilera, A. M., M. Escabias and M. J. Valderrama. "Using principal components for estimating logistic regression with high-dimensional multicollinear data." *Computational Statistics & Data Analysis* 50 (2006): 1905–24.

Bachman, K. "Cracking the Set-Top Box Code." 2009. [URL: http://www.mediaweek.com/mw/content_display/news/media-agencies-research/e3i8fb28a31928f66a5484f8ea330401421].

Friedman, J., T. Hastie, and R.Tibshirani. "glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.1-3", 2009. [URL: http://www-stat.stanford.edu/~hastie/Papers/glmnet.pdf]

Interian, Y., S. Dorai-Raj, I. Navernioux, et al. Ad quality on TV: predicting television audience retention. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, Paris: Association for Computing Machinery, 2009.

Kempe, D. and W. C. Wilbur. "What can television networks learn from search engines? How to select, price and order ads to maximize advertiser welfare." 2009 [URL: http://ssrn.com/abstract=1423702]

McCullagh, P and J. A. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 1989.

Myers, R. H. *Classical and Modern Regression with Applications*, 2nd ed. Belmont, CA: Duxbury Press, 1990.

Nielsen Inc. "Nielsen Says Bud Light Lime and Godaddy.Com Are Most-Viewed Ads During Super Bowl XLIII." 2009. [URL: http://en-us.nielsen.com/main/news/news_releases/2009/February/nielsen_says_bud_light]

Richardson, M., E. Dominowska and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, New York: Association for Computing Machinery, 2007.

Sculley, D., R. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris: Association for Computing Machinery, 2009.