Best Practices in Matching Databases to Set-Top Box Data

Prepared By:
Myles Glenn Megdal

June 2011

Table of Contents

Introduction…… 3
Study Methodology……..... 4
Executive Summary…….. 7
Advanced and Addressable TV Applications……... 11
Issues and Concerns about Ancillary Databases…….. 13
Data and Databases……... 14
Introduction to Data……. 14
Realities of Data……..…….. 14
Types and Categories of Data and Databases…….. 16
Third Party Data Sources…… 17
Compiled Databases…… 20
Demographic Data…….. 20
Psychographic Data…… 22
Behavioral Data……... 23
Geographic Data…….. 23
Cluster Data…….. 24
Pros, Cons, and Issues…..…….. 25
Shopper Databases……... 34
Pros, Cons, and Issues……..…….. 34
Scope and Composition…….. 35
Miscellaneous Databases……. 38
Credit Data……... 38
Automotive Data……. 39
Geographic Data…….. 40
Advertiser Supplied Data…..…… 40
Database Matching Approaches and Privacy……... 42
Introduction to Database Matching…….. 42
Individual vs. Household Level Matching…….. 43
Matching Technology……... 44
Blind Matching……. 46
Consumer Privacy…… 47
Assuring Database Accuracy and Projectability……... 50
Geographic and Demographic Adjustments……..…… 51
Shopper Data Adjustments…..……... 51
STB Data Adjustments…..……... 52
Observations and Conclusions…… 54
Appendix…… 55

Introduction

This whitepaper is a logical extension of CIMM's prior "Roadmap for Set-Top Box Data" study that focused on the state of STB data as it is currently being used to support a wide range of applications from marketing and carriage negotiations, granular analyses of programming and advertising, to enabling various types of geographic and household level addressable advertising, and granular local and national media planning. For virtually all of these STB applications, various external databases are integrated, resulting in the creation of composite datasets that support various STB related products and services. These databases – linking viewership, demographics, and product/purchase information - are both compliant with consumer privacy laws and guidelines and meet the data integrity and accuracy requirements of users of these advanced and addressable advertising products and services.

As such, this study focused on the various categories and types of external databases that are currently being used with STB data to support advanced and addressable advertising applications. The emphasis of this study is on providing an understanding of the nature of the various databases that either are currently being used, or can be used, to support STB related applications. The advantages and limitations of these databases will be discussed, as well as the processes and procedures that are employed to assure their accuracy and representativeness – especially for media planning applications, where uniformity and conformity with existing measurement metrics are critical.

Focus will also be given to the matching procedures used to combine STB data with other databases to assure both accuracy and compliance with consumer privacy regulations and guidelines.

Ultimately, it is hoped that the information provided in the study will both educate and enlighten STB users on the strengths and limitations of using granular STB data with equally granular external data to augment and enhance target audience definition, media planning, addressable advertising deployment, and campaign measurement.

Study Methodology

As with prior CIMM STB studies, much of the input and insights have been gathered as the result of interviews with advanced and addressable advertising users and service providers. The focus of these interviews was on identifying how these companies and individuals use STB and other databases, and on specific concerns and issues that they face in using the data. Additionally, those companies that are currently integrating several databases with STB data were asked to provide their input on the processes and procedures they employ to assure the accuracy, consistency, and reliability of the composite databases that result from matching other third party databases to STB data.

We are grateful for the support we received from the companies and individuals that were interviewed from this STB user and supplier audience – especially, since many of these companies and individuals were gracious enough to give their time and energy by participating in prior CIMM studies.

In addition to these companies, interviews were also performed with companies that are suppliers of the third party data that is matched to STB data, and with companies that are actually performing the database matching. The focus of these interviews was on the composition and nature of the various databases, and the processes and procedures performed to assure the accuracy of the matching, as well as adherence to consumer privacy regulations and guidelines.

Given that the focus of this study was primarily on data and data matching techniques, the interviews were done sequentially, with end users and STB service providers interviewed first to obtain their perceptions and concerns, and the data and database matching companies interviewed last to make certain that all user issues were addressed.

All interviewees were engaged in this study, and gave their time and input without reservation. We are grateful for their participation.

In total, 38 individuals from 20 companies were interviewed. The composition of this audience was sufficiently broad to obtain an excellent representation of issues and concerns, and assure the anonymity of the participants. Interview guides were tailored for each of the various groups of companies interviewed, and can be reviewed in the appendix.

The composition of the interview audience was as follows:

- ***End users*** – advertisers, agencies, and media networks – this group of companies and individuals are direct users of the advanced or addressable advertising products and services that utilize STB data combined with other databases, and are the primary clients of the STB service providers

- ***STB service providers*** – addressable technology partners, and research and measurement companies – this group of companies and individuals are directly involved in integrating third party databases with STB data for specific applications ranging from addressable advertising deployment, to measurement of long tail networks and local markets, to ROI base planning and buying, and provide a range of planning and addressable solutions to the end user community

- ***Database and database matching suppliers*** – database compilers/suppliers and marketing services companies that provide data and data processing services in support of advanced and addressable advertising applications, with these services primarily provided to the STB service providers

	Companies	***Interviewees***
End Users	**6**	**8**
STB Service Providers	**7**	**15**
Database & Matching Suppliers	**7**	**15**
	20	**38**

In addition to the interviews, information was obtained from web resources for some of the participating companies. This was done as a means of filling in some of the "blanks" uncovered as part of the interview process. Most of the blanks related to topics or issues that were either not addressed, or incompletely addressed.

It should be noted that many of the issues and concerns that were voiced during this study are not unique…just unique their application to television advertising. The use of direct and database marketing targeting approaches and data has been a common occurrence in direct mail, magazine advertising (selective binding), and email and web ad targeting for years. The compiled multisource databases that were used 30 years ago for direct mail are basically the same databases being used for advanced and addressable advertising applications today. Concerns for

consumer privacy are not unique to these television applications, and the “blind matching” techniques currently being used to integrate STB data with other databases in a privacy compliant manner are the same as those used in the ‘80’s to match magazine subscriber bases to compiled databases to identify specific subscriber segments for targeted magazine advertising (selective binding Each time a new media becomes addressable (or targetable) at precise levels, issues concerning consumer privacy and the accuracy of the information arise. This study will attempt to clarify and resolve many of these questions and issues.

Executive Summary

All of the individuals and companies interviewed for this study are strong advocates for the use of ancillary databases that are matched to STB data. Although the primary uses for these matched databases differed vastly, each one required that STB data be either matched or bridged to one or more additional databases – with the most common databases providing insights into consumer demographics, lifestyles, and category and brand preferences.

Those television planning applications that focused on smaller markets, networks, and niche audiences have found that these databases provide demographics and insights into consumers in a permissible, non-personally identifiable (non-PII), manner. Agencies and advertisers seeking to define target audiences in terms of category or brand purchasing behavior rather than demographics, also have a range of options from suppliers that marry known shopping behavior to STB and compiled demographic data.

In every instance, planning applications using these databases provide precision that is not currently available with traditional media planning bases. In the most simplistic sense, information derived from millions of households should be more precise and stable than that derived from thousands of households. For precise planning of niche audiences, small markets, and niche networks, size does matter.

Scale is also important for addressable applications, and while the same demo and shopper enhanced STB data can be used to target addressable ad campaigns to a desired household, currently there is limited household addressable inventory. This is expected to change as additional MVPDs commence household addressability in the next 12 months. Geographic level addressability – cable zone or zip code level – is fairly widespread, and while not providing the precision of household level addressability, still provides sufficient benefits to be cost justified. In virtually all geo-addressable programs, one or more databases are being used to plan and deploy the campaigns.

The specific applications discussed during the interviews encompassed:

1. More precise audience definitions – using all available data (demographics, shopping, behavior, lifestyles, etc.)
2. Planning and deployment of household and geographic level addressable advertising campaigns
3. Local market measurement

4. Measurement of long tail and niche networks
5. Campaign monitoring and measurement

Whether a supplier or end user of STB data products and services, all interviewees were confident that the increased precision and tonality that these demographic and shopper databases provide will become increasingly important in the future as issues are resolved and the user base for these products and services grows.

As with any new technology or approach, several issues exist. Many of these are common to all categories of companies interviewed, while others are more specific to a specific group. Among those noted are:

1. Third Party Database Issues

These common issues for all end users and providers of STB products and services relate to the nature of the demographic and shopper databases that are being used for advanced and addressable support. A number of issues were identified including:

- How accurate is the data, and what sources are used to obtain the data?
- Is the data explicit or modeled, and how accurate are the models?
- Better data transparency is needed. What does a dog-owning household mean? Does a dog live there now, or did one once live there?
- How consistent is the data? If not always accurate, are the inaccuracies consistent, so that adjustments or weighting can be performed?

This study delved deeply into these issues, and, for the most part, the companies supplying and processing this data are moving rapidly to resolve these concerns.

2. Costs

The subject of costs and pricing was difficult to discuss in the context of this study, and not a dollar amount was quoted by a participant. That said, pricing did surface as an issue from each of the types of companies. In general, the costs are perceived to be high. There are a number of reasons for this, with just the sheer volume of data processing and analyses required to match and develop the databases used to support these applications being first and foremost. Another big cost factor is the fact that – with just a few exceptions – most of the STB applications are custom projects that are labor intensive.

Costs should come down over time as more consistent STB data comes on line, and as more of the data matching and products become more automated.

3. Database Matching Concerns

These are concerns related to the database matching process, and focus on two specific issues:

- Assuring compliance with consumer privacy regulations and guidelines
- Assuring that the matched database is not biased with regard to demographics, geographic coverage, viewership, and shopping behavior

Once again, these issues were an area of focus for this study, and it can be stated that consumer anonymity is being maintained, and significant resources are being expended to assure the accuracy and projectability of the aggregated databases.

4. Lack of More Sophisticated Tools

Mention was made of the lack of more sophisticated tools for accessing and using the enhanced STB databases. Most audience segmentation utilizes just a few of the thousands of characteristics that are available. With few exceptions, most advanced planning is performed on a project basis, as are most addressable campaigns.

The automated analysis and planning tools that are available for online media are not yet available for these applications.

One interesting point is that all of the interviewees felt comfortable that complete compliance with consumer privacy regulations and guidelines are being strictly adhered to. At no time in the extensive processing that is performed are name and address, viewing behavior, frequent shopper data, and demographics all resident within the same company. Only completely compliant, depersonalized data is available to support advanced and addressable applications.

With regard to the data focus of this study, several key positives surfaced from the interviews:

- There is a tremendous amount of information available to support advanced and addressable products and services
- Every STB service provider is performing extensive weighting and balancing of their enhanced STB databases to minimize any biases that may exist.
- The compiled third party databases being used for demographic, psychographic, and behavioral data are highly consistent and provide accurate data across a wide range of characteristics

- Testing of advanced and addressable products and services have proven successful

In order to reach the potential benefits that STB data offers, the following enhancements were voiced:

– ***Better data transparency***

All data suppliers need to provide users with the definitions for all available database characteristics, with the associated information on sources of data, accuracy percentages, and means of identifying explicit versus modeled data. This will provide users with the confidence to use the data without reservations.

– ***Provide better access to the data***

The database used for advanced and addressable products and services are used in other media, but in those instances more robust tools for accessing and analyzing data are far more prevalent. The implementation of similar tools these applications will both facilitate greater usage, as well as aid in cost control.

– ***Reduce costs***

Costs should become more reasonable as processing automation and data quality improves.

Advanced and Addressable TV Applications

Unlike the prior CIMM whitepaper – ***Roadmap for Set-Top Box Data*** – which looked at all the potential applications for STB data, this current study focuses on a subset of these applications that are related to advanced and addressable advertising support. The common element running through all of these applications is that STB data is not used by itself, but is combined or integrated with other databases to provide a robust base of information with which to support the following key applications:

Addressable Advertising

STB data provides the ability to deploy and measure addressable advertising campaigns – with or without interactivity – at both the geographic and household levels. Third party databases provide the insights into anonymous households necessary to both plan and deploy these campaigns. By using data in this manner, addressable advertising provides television with the same inherent capabilities that direct marketers have used for decades.

The appendix lists a number of definitions for addressable advertising that are part of the CIMM STB Lexicon.

Advanced Advertising

Based on the interviews conducted for this study, the use of ancillary databases matched to STB data is far more prevalent in the area of advanced advertising applications, than for addressable advertising. The primary applications in this area are related to:

- ***Target audience definition***

Third party databases provide the ability to define audiences based on virtually any criteria including category or brand purchasing behavior, desired lifestyle characteristics, life stage, precise demographics, or any combination of these characteristics.

- ***Local market measurement***

Third party data – especially consumer demographics – is extremely valuable in identifying and correcting population bias that may result from using STB data for local market measurement. The demographic composition of the STB viewer sample can be compared – via the appended demographics – to known market demographics (as represented by US Census data).

- ***Representative sample creation***

The availability of household level demographic data from third party databases is an asset in developing STB samples that are representative of a desired population. One application of this approach was noted previously for local market measurement, but a similar approach could be used to create a quasi-national STB sample, or to create any STB sample that closely mirrors a desired market or population segment.

It is interesting to note that suppliers of advanced and addressable advertising services have taken very different approaches in how they develop and utilize their composite bases. For example:

- Compiled demographic databases range from 110MM – 135MM households
- Shopper data households range from 100,000 to 60,000,000
- Some companies are using only 30 key demographic variables, whereas others have access to hundreds of demographic and behavioral characteristics

Despite these variations in approach, all of the companies interviewed have spent significant time and effort in verifying the accuracy and efficacy of their specific techniques and approaches against known reference data sources.

Issues and Concerns about Databases that are Matched to STB Data

There were a number of issues and concerns voiced that relate to the composition and use of demographic and shopper databases that comprise the core information bases that are matched to STB data for virtually all of advanced and addressable advertising products and services.

The primary issues that surfaced with regard to shopper data are related to geographic and demographic biases that result from the nature of the specific retailers that contribute the frequent shopper data. While the relatively large number of shoppers represented in these bases tends to provide good demographic coverage, the nature of the retail chains themselves introduce their own shopper bias. All of the companies utilizing frequent shopper data do an admirable job of adjusting the shopper bases for both geographic and demographic biases. The actual purchase data is transactional in nature, and is highly accurate with regard to category, brand, units, size, etc. Verification and balancing of shopper data is performed via comparisons with generally accepted reference bases (such as IRI/Homescan).

With regard to the multisource compiled databases that are used to provide demographic, psychographic and behavioral data for STB applications, the primary issues relate to the accuracy, consistency, latency, and coverage of the data. As will be seen in the next section of the white paper, multisource compiled databases use a number of different data sources for each characteristic. As such, variations in both data quality and accuracy will exist for different components of data – for example, psychographic data will largely be self-reported, whereas, behavioral data will be primarily transactional.

The single biggest issue voiced by users of this data is getting an accurate assessment, on an individual characteristic basis, of the accuracy and source of each characteristic – i.e., is the data explicit or modeled, and what percent of each characteristic is explicit vs. modeled.

While this issue will be addressed in this study, a few points should be noted now. The multisource compilers take extraordinary care to assure the consistency of their data. While some demographics may not be explicitly accurate, they will be consistently accurate, and thus, perfectly useful in advanced advertising applications where a given demographic is needed in a relative sense – i.e., higher vs. lower income households are desired. On the other hand, if the objective is to mirror a specific demographic for a target audience defined from a generally accepted reference base (Nielsen, for example), then absolute accuracy for that characteristic is required, and the compilers need to be able to provide users with the confidence that some percentage of their data will align perfectly with the desired demographic.

Data and Databases

Introduction to Data

Information is the cornerstone of our world today. Virtually every aspect of consumer and business behavior generates data that is turned into information that is then used to impact our lives. The products we buy, the ads we see, and the services we use are all based on information gathered from our behaviors.

This is especially true for marketing and advertising activities, where the use of information has always been important, but has become even more critical as consumers become more heterogeneous and fragmented, media and communications channels have multiplied and expanded, and products and services have proliferated to meet the needs of every niche consumer segment.

As we have seen from the discussion on advanced and addressable advertising applications, data is the foundation on which these tools and techniques have built. The United States, while still advocating and enforcing consumer privacy, has among the most extensive data and information resources in the world – many of which can be used to support both advanced and addressable applications. The types of databases that are available and how best to utilize them are the focus of this section of the white paper.

Realities of Data

While we are blessed with a wealth of data and databases to be used to support advanced and addressable applications, there are key realities that need to be acknowledged:

- ***Accuracy*** – there is no such thing as a 100% accurate database. Every database has some degree of inaccuracy. These inaccuracies could take the form of simple errors in one or more characteristics on the database (i.e., incorrect age, or income), or in demographic or geographic biases based on the source of the data. Cable operator subscriber bases for example have a geographic bias based on the geography served, and may also have a demographic bias based on the nature of the consumer base within their serviced geography.

- ***Consistency*** – consistency in data is a necessary requirement in order for the data to be useful. Data that may be inaccurate, but is consistently inaccurate, can be used with confidence. Data that is inconsistent is rarely able to be used. For example, if household income is consistently under-reported by $10,000 on a database used for target audience definition, that database can still be used effectively, since income can be adjusted consistently for all households. On the other hand, another database where income may inconsistently vary above or below actual cannot be used with confidence if income is a necessary component of defining the desired target audience.

- ***Source of the Data*** – the source of a particular piece of data is an important factor in determining the data element's accuracy. Dog ownership data derived from pet insurance information is likely to be more accurate than survey response data, since the pet insurance data is the result of a financial transaction. Multiple sources of the same data – common in compiled databases from companies such as Acxiom and Experian – are typically more accurate than data derived solely from a single source. Multiple sources provide multiple verifications of accuracy.

There are a vast number of databases and lists that can be used to support advanced and addressable TV advertising applications. These range from client-supplied lists of known customers or buyers that are typically highly accurate, but limited in the extent of data available, to vast compiled consumer databases that maintain hundreds of characteristics on hundreds of million consumers. In order to evaluate these myriad of databases for advanced TV applications, certain metrics need to be defined including:

- ***Coverage*** – how many individuals or households are represented on the database, and what percent of the population does the coverage represent. In the case of compiled databases, coverage is in terms of the percent of total domestic households. For a cable operator, coverage represents the percent of total households in their markets. In general, higher coverage is desired in order to provide statistically valid representation for small consumer segments and local markets.

- ***Breadth of Data*** – the number of overall data elements, or characteristics, that are available for each individual or household. Compiled databases typically have hundreds of elements available for each individual or household across a wide range of categories (demographics, psychographics, behavioral, geographic, etc.); whereas, shopper databases will contain a more limited number of elements (characteristics) focused strictly on in-store purchase behavior.

- ***Depth of Data*** – the number (or percent) of total individuals or households on a database having a specific data element or characteristic. For example, estimated income is present on 100% of a compiled database, but pet ownership is only available on 20% of the base.

- ***Accuracy*** – the degree to which database characteristics agree with a verifying source. For example, shopper data is compared to either known advertiser data, or to research panels such as Homescan.

- ***Latency*** – the age of a specific characteristic, or the frequency of updating a specific characteristic. Shopper data may be updated daily or weekly; whereas, estimated income on a compiled database may only be updated quarterly at best. Another update concern is how certain characteristics are updated. For example, certain characteristics are only updated via additions. This occurs frequently with survey collected data and could result in situations where "once a dog owner, always a dog owner".

Types and Categories of Data and Databases

In general, a company has access to two types of databases – internal and third party. Internal databases and data consists of information that is collected and maintained as the result of company's core business operations and includes product and service data, customer communications data, and financial data – customer billing, product costs, etc. Another example of internal data is warranty information for auto companies, and offer redemption data for CPG companies.

Internal data is extremely valuable to companies across a wide range of uses, and can be used effectively to define custom target audiences for TV planning and addressable TV campaigns based on information that may only be resident within the advertiser's data resources. Examples of these types of campaigns would be owner loyalty – wherein a manufacturer desires to target their most valuable customers for "defensive" reasons, and the definition of "most valuable" can only be achieved using proprietary internal data.

While valuable for certain applications, internal data has its limitations, most notably in the following areas:

- Non-customers, or potential customers, are not represented. As such, internal data is of limited value in planning and executing prospecting or conquest campaigns.

- Internal databases typically maintain only data that is a byproduct of a company's core business. As such, characteristics that add tonality to understand the nature of consumers is often missing or under-represented. Information on demographics and preferences may be very limited or completely lacking.

Many companies augment their internal databases by incorporating data from third party databases. For example, a credit card issuer such as American Express has a deep understanding of the credit card spending behavior of their customers with regard to American Express usage, but will augment their knowledge of these customers by matching their customer base to other third party databases – notably compiled consumer databases – to provide more information on customer demographics, lifestyles, and interests.

It is this type of database matching that is integral to the advanced and addressable TV, in which various databases – both internal (MVPD subscriber files, STB viewing data) are integrated with third party databases (shopper databases, compiled consumer databases, media databases) to provide a consolidated, privacy compliant, base of information to support audience definition, measurement, and addressable campaign deployment applications that comprise the focus of this white paper.

Third Party Data Sources

Interviewees from the data companies indicated that there are numerous sources of third party data available in the US including the following:

- US Census Bureau
    - Provides detailed demographic information at different geographic levels ranging from census block groups (smallest level) to zip codes to markets. While this data is compiled every 10 years, there are a number of companies that provide interim year projections of the base census data.

- Credit Bureaus
    - The three major credit bureaus – Experian, Trans Union, and Equifax – compile data at the individual level detailing the credit history and performance of virtually every

consumer that has had any type of credit product, or has applied for credit. The data is robust and current, but is highly restricted to uses governed by the Fair Credit Reporting Act. As such, this information is typically used only by credit issuers and insurance companies to make credit and insurance offers and decisions. It should be noted that bureau data – at the geographic level (zip code, for example) – is available from some bureaus, and can be used in a consumer compliant manner for advanced and addressable applications. While not at the individual or household level, this geographic credit data can identify high potential areas or programming for targeted financial services and insurance products.

- Industry Specific Consortium Aggregators
  - There are a number of companies that specialize in aggregating data from different sources for specific applications. Among these are:
    - IMS – focuses on the healthcare industry
    - IXI (division of Equifax) – focuses on the investor and high end financial industry
  - These companies use permissible consumer information at the individual level combined with anonymous confidential information (for example, IXI uses anonymous asset and wealth data) to provide services for companies with highly focused needs.

- Compiled Database Aggregators
  - These are databases from marketing service providers such as Acxiom, Experian, Epsilon, Merkle, Allant, etc. Typically, these databases are very broad in coverage (in terms of individuals and households), and are very deep in data. In virtually all cases, this companies use multiple sources to compile and verify their data. As such, they tend to produce broader categories and data for larger universes. It is this breadth and depth of data that makes these databases attractive for use in advanced and addressable advertising applications.

- Industry Transactional Aggregators
  - These companies focus on aggregating transactional information specific to a specific industry. Examples of this are companies such as:
    - Catalina – aggregates product purchase data from retailers
    - Dunnhumby – aggregates product purchase data from retailers

        - Abacus – aggregates direct marketing purchases from cataloguers
        - RL Polk – aggregates vehicle purchase and lease information
    - In some instances, this data is available at the individual or household level, but usually only to companies that contribute data into the base. In other instances, the data is restricted at the individual level (automotive data, for example), but can be used when depersonalized and summarized to the geographic level in a manner similar to credit data. Any of these databases can be used to support advanced and addressable advertising applications.

- Research Data Suppliers
    - There are a number of advertising agencies and primary research firms that offer a wide range of information products derived from consumer research panels blended with sophisticated research methodologies. These include companies such as Nielsen, Kantar, Simmons, MRI, Media Audits, IRI/Homescan, and others. These panels differ in the types and categories of data, and data collection approaches, and are used primarily to support core applications for each of these companies, but can also be used for advanced and addressable applications. Some bases – notably Nielsen and Homescan are used to verify data from other databases – STB data in the case of Nielsen, and shopper data in the case of Homescan. Other research bases – notably Simmons and MRI – are often used to define target audiences for either advanced TV planning or addressable advertising.

- Vertical Lists
    - These are interest specific lists, with the primary category being magazine subscriber lists. These lists typically have very limited information – almost always related to a single focus (golf, boating, tennis, health, etc.), with minimal ancillary information other than subscription length, type of subscription, etc. These lists can be used to define a desired target audience – such as golf enthusiasts or potential yacht buyers – however, these lists need to be used with caution. There are large percentages of non-target aspirants on many of these lists. In fact, fully half of the subscribers to one major yachting magazine do not actually own a yacht, or have the means to do so.

Given the number of list and database categories, and the numerous companies with databases within one or more categories, we will focus on those third party databases most often used to

support advanced and addressable TV advertising applications – namely, shopper and compiled databases. In our discussion with advanced and addressable users, it is these two types of databases that are most often integrated with STB data to support audience definition, measurement, and addressability. Other databases – notably, research bases – will be discussed since they often comprise the data used to validate and adjust and balance third party databases.

Categories of Data – Compiled Databases

Most compiled databases encompass four primary categories of data – demographic, psychographic, behavioral, and geographic - each of which is derived from multiple different sources of data. With regard to size and scope, it is not uncommon for a compiled database to maintain information on 95% of the economically viable US household population, with over 1,500 characteristics available for each household or household member.

✧ *Demographic Data*

Demographic data provides descriptive attributes of individuals or households, and is derived from a number of different sources depending on the specific compiler. An average of 5-10 different sources for a specific demographic element is not unusual. Demographic data provides tremendous insights into the nature of individual consumers and their households, and can be used by itself for advanced and addressable TV applications – especially, when the desired target audience does not fit standard demographic definitions. For example, married males, age 25-39, with incomes of $100,000+ could represent a desired audience. Alternatively, demographics can be used to refine a desired target audience that is primarily defined in terms of purchase data – high volume beer drinkers with incomes of $100,000+.

Virtually all companies that are supporting advanced TV advertising applications – either for planning, measurement, or addressability – are using demographic data from a multi-source compiled database as part of their service offerings.

All compiled databases maintain data at both the individual and household levels; however, data accuracy and coverage are significantly better at the household level, than at the individual level. It is because of the higher coverage and accuracy at the household level – as well as the ability to match databases better at the household level than at the individual level – that all of the

advanced and addressable companies interviewed use compiled data strictly at the household level.

One point needs to be made concerning household level data. Just because data is being matched at the household level, doesn't mean that household composition is not known. Compiled databases use a set of business rules to identify a head of household (traditionally, an adult male in a married household with children), but will also maintain information on the age and genders of all household members. As such, advertisers seeking households with teenage girls, will still be able to identify those households that have a high incidence of teenage girls – albeit, with somewhat less accuracy than households with adult males.

This point underscores one of the inherent limitations of multi-source compiled databases – namely, they have the best coverage and the highest degree of accuracy for those households (and household members) having the greatest likelihood of appearing on multiple data sources. Adults – with credit reports, owning phones, cars, homes, making lots of purchases, and participating in surveys and loyalty programs – have extensive public and private domain information histories and are therefore well represented with a high degree of accuracy on compiled databases.

Children, teens, and older seniors (especially those living with their children or in senior living facilities) have significantly less public and private information available and are thus less likely to be accurately represented on these databases.

Additionally, the extremities of the population with regard to affluence are also under-represented. The extremely poor – who own little, and thus have less public and private information are under-represented. The extremely affluent – who own a lot, but do an excellent job of hiding assets – are typically, under-represented (at least with regard to the depth of data).

Demographic data is quite extensive, and a comprehensive list of most of the available characteristics can be found in the appendix. The primary demographic characteristics include:

- Age
- Gender
- Education
- Occupation
- Income
- Ethnicity
- Marital Status
- Household size and composition

- Dwelling unit type and size
- Own or rent
- Credit card ownership
- Life events (derived from updating the demographic data)
    - New Parent
    - New Mover
    - Newlywed
    - Recent Divorce

✧ ***Psychographic Data***

Psychographic data consists of self-reported opinions, attitudes, and intentions of consumers and is primarily acquired from various survey sources. While this data can be potentially very powerful, it is not broadly represented across all households, and is subject to difficulties in verification due to the self-reported nature of the data. Additionally, since this data is the result of consumer surveys, more affluent segments of the population – who typically are not survey responsive – are under-represented in this category of data. As such, users should be aware of a potential demographic bias inherent in this data. Generally, survey response data is derived from fewer sources than demographic data, with 3-5 sources being average.

The primary categories of psychographic (or lifestyle) data available on compiled databases are:

- Interests/hobbies
    - Gourmet dining
    - Boating
    - Golf
    - Tennis
    - Exercise/health enthusiast
    - Upscale living
- Travel preferences
    - Vacation domestic
    - Vacation cruise
    - Vacation European
    - Family travel
- Intentions
    - Intend to buy a new car
    - Intend to move

        - Intend to marry
    - Reading preferences
    - Ailments and medications
    - Contributions

More detail on the extent of this data can be found in the appendix. The Acxiom database maintains well over 100 psychographic data elements (characteristics).

✧ ***Behavioral Data***

Behavioral data is somewhat similar to psychographic data, but is based not on self-reported information, but on actual transactions or observed behaviors. As such, behavioral data tends to be more accurate than psychographic data, and is often used as a verification base for survey response psychographic characteristics. Most behavioral data is focused on buying/purchasing activities by category, including:

- Apparel
- Food/Wine
- Electronics
- Sporting goods
- Music
- Investments
- Online buying
- Mail order buying

While the categories may vary slightly between multisource compiled databases, most of these databases maintain several hundred behavioral characteristics. As is the case with psychographic data, coverage of these characteristics is not broad; however, demographic bias is less than with psychographic data given the transactional nature of the source data. Given the transactional nature of this data, only 1-2 sources per data element or characteristic is common.

✧ ***Geographic Data***

While obviously not at either an individual or household level, geographic data can still be extremely valuable for defining desired target audiences – especially, for geo-level addressable

campaigns. For example, geo-level addressability can be executed at the cable head end (cable zone) level, and provide approximately 10 times the precision of market level targeting (2200 cable zones versus 210 DMA's). Another application for geographic level targeting is the ability to use data that is highly restricted in use at the individual/household level, but unrestricted if depersonalized to the geographic level. Both credit and automotive data are currently being summarized to the small geographic area level, and thus provides a compliant means of associating small units of geography (zip codes or zip+4's) with a desired credit or automotive target audience.

✧ ***Cluster Data***

In addition to the above 4 core categories of data, each of the major compilers also has "cluster" data associated with each household on their databases. While clusters are not new – Claritas' Prizm clustering schema is now over 30 years old – they are finding increased utilization today – especially, in light of concerns over privacy compliant use of data.

Clusters trade off the precision that individual and household level data provide, for cost efficiency and privacy compliance. There are numerous cluster products available today, all of which have some aspects in common:

- They segment the population into discrete segments (or clusters) that have a degree of consumer homogeneity within each cluster
- The clusters are often built from household level data, but are always depersonalized for use
- The number of clusters can vary dramatically, from 40 to well over 100
- Clusters can be applied to specific households or units of geography – As such, they can easily be used for target audience definition, or addressable advertising
- Clusters can be developed to focus on specific behaviors or to conform with specific industry needs:
    - Psycle – financial services focus
    - Spectra – consumer package goods focus
    - PersonicX – product purchase propensity focus
    - Mosaic – demographic segments that can be applied globally
    - Connexion – communications services focus

There are also clustering products that focus on consumer preferences towards media and technology. The appendix will provide greater detail on the range of clustering products that are available for marketing and advertising applications.

✧ ***Compiled Databases – Pros, Cons, and Issues***

If you evaluate multisource compiled databases in terms of the five metrics previously discussed – coverage, breadth of data, depth of data, accuracy, and latency – the databases from all primary suppliers appear to score high; however, issues do exist.

Coverage

All of the major compilations represent between 115MM – 130MM domestic households, 235MM – 250MM adult consumers, and up to eight individuals per household. In general this represents approximately 95% of the economically viable households in the US. While coverage in adults is excellent, coverage on children and teens is less extensive, due in large part to poorer information availability for these segments, and to compliance restrictions on the use child data

Psychographic or lifestyle data is typically only available for approximately half of the household population. Behavioral or transactional data has slightly better coverage at 60% of domestic households.

Geographic level data – while not linked to households, but to the geography in which household reside – still provide significant coverage and excellent granularity. For example, Experian's Auto Market Statistics and Summarized Credit Statistics provide detailed automotive and credit information for 29MM discrete geographic units at the zip+4 and zip code levels.

Given the large size and excellent household coverage provided by compiled databases, they have become the standard base for virtually all companies providing advanced and addressable TV advertising support, with all of the STB service providers interviewed using compiled data from Experian, Acxiom, or Epsilon. Given that many of the input sources used are in common to these bases, decisions as to which base is most appropriate are typically based on pre-existing relationships (especially, if a given MVPD supplying STB data already has a existing compiled data supplier relationship), and costs.

Breadth of Data

Data coverage on compiled databases is excellent, due in large part to the numerous different sources used in developing and maintaining the database. While there are a number of data sources unique to each compilation, the core sources are typically the same and include:

- Telephone directory white pages
- Property/realty records including deeds
- Mail order transactions
- Catalog and retail transactions
- Census data
- Survey response data – online and offline
- Aggregated credit data
- Aggregated vehicle registration data

With literally hundreds of data sources available, compiled databases are rich in available data, with the following coverage for each category of data:

- **Demographic data** – typically 500+ demographic characteristic available for defining consumer segments or target audiences

- **Psychographic/lifestyle data** – on average 1,200+ survey response characteristics across a wide range of consumer interests, attitudes, and purchase intentions

- **Behavioral/transactional data** – on average 150+ transactional characteristics primarily focused on mail order, internet, and retail purchase activity

- **Geographic data** – Experian indicates that both the automotive and credit geo-aggregated data contain in excess of 300 characteristics each. Additionally, all compilations offer US census data that provides approximately 450 demographic characteristics at the small area geographic level.

As can be seen from these statistics, data coverage is extensive across all components of a compiled database. That said, it needs to be noted that not all data is available for all households

Depth of Data

As would be expected on a database potentially maintaining over 2,000 characteristics, not all individuals and households will have all available data. In general, the depth of data is excellent for standard demographics – age, income, marital status, household composition, home ownership – however, coverage in certain lifestyle (psychographic) and behavioral characteristics could be quite sparse. One of the major compilers details coverage for key demographic characteristics as follows:

- Household Income – 100%
- Occupation – 99.8%
- Education – 99.7%
- Homeowner/renter – 91.0%
- Ethnicity – 90.0%
- Age – 86.3%
- Marital Status – 85.5%

If one were to calculate "average coverage" at the data element or characteristic level, the answer would be close to 60%; however, this is misleading. As can be seen from the aforementioned demographic data, some characteristics are extremely well represented, whereas, other characteristics – most notably life style interests and purchase intentions – have very small household level coverage. In general, niche segments will have low coverage on the compiled base. That said, coverage is still usually sufficient to be able to use lifestyle data in advanced advertising applications for planning, if not sufficiently large to support an addressable TV campaign.

Latency

Latency refers to how current the database information is and how frequently it's updated. In the case of the compiled databases used to support advanced and addressable TV applications, there are really three discrete update cycles to be considered.

- **Database updating** – given the number of input sources (often 2,000 or more for a compiled database), and the number of update related processes that are required (change of address processing, model scoring, etc.), most compiled databases are updated continuously on a 24X7 basis. These databases are virtually all maintained in relational database structures that allow for efficient updating of components of the database. In general, data feeds for demographic, psychographic, and behavioral data are updated in this 24X7X365 schedule, with geographic data – summarized credit

statistics, and aggregate auto data - only updated on a quarterly basis. It is important to note that just because the primary database is being updated constantly, not all characteristics are updated at that frequency. Certain data – such as real estate data – may only be updated when a change occurs, or once annually, depending on the source of the data.

- **Updating of application specific databases** – the primary database maintained by all large compilers is updated in a continuous manner; however, this database is used to create extracts that are used to support specific core business functions for the compilers – most notably direct mail list rentals, and list enhancement (overlaying of information onto a client supplied file or database). These application specific bases are created on a frequency that is either weekly or monthly. In most instances, the database extract used for enhancement applications is created monthly.

- **Updating for advanced or addressable TV applications** – the costs associated with matching two databases totaling in excess of 100MM records are significant. As such, all of the companies using STB data with compiled databases perform this matching on a quarterly basis. Assuming a month to perform the matching, validate and verify the results, and load the updated combined base into a production environment, most of the advanced advertising suppliers are using compiled data that is on average 4 months old.

In today's world of rapid data feeds from click stream transactions and people meters, 4 month old demographic data may seem woefully out of date. The reality is that the demographic data does not vary significantly over a 4 month period to overly influence results for campaign planning or target audience definition. In the case of addressable advertising applications, the MVPD subscriber file is probably matched to append compiled data on a more frequent basis, but would still result in compiled data being 1-2 months old at best.

Accuracy

During the interview discussions, most of the compiled data issues that were raised concerned the accuracy of the data. Specifically, concerns were voiced in three specific areas:

- **Demographic Bias** – are there inherent biases in the compiled database, such as under or over represented population segments?

- **Explicit vs. Inferred Data** – what percentage of the data (available characteristics for each individual and household) is based on explicit information versus inferred or modeled data?

- **Data accuracy verification** – what procedures and sources are performed to verify data accuracy for explicit data, and what approaches are used to assure the accuracy of modeled data?

While these issues were voiced by virtually all of the advanced and addressable advertising users and service providers interviewed, the actual degree of accuracy required for advanced and addressable applications differ significantly.

If the use of the compiled data is to target specific households for household level addressable TV campaigns, then high concentrations of accurate, explicit data are required to assure the highest composition audience. Inferred or modeled data can also be used, but only to augment explicit data, and only if the models provide sufficient accuracy to assure high composition audiences. Segment biases, while important, become less important in household level addressable programs, since the specific MVPD households are being targeted. As such, if a desired target audience is over-represented, that works in favor of the addressable campaign. If the desired segment is under-represented for a specific MVPD subscriber base, then the economics of the addressable campaign may not be favorable.

Geographic addressable campaigns to the cable zone level can tolerate slightly less data accuracy than household level addressability. This is due to the fact that geographic areas (cable zones or head ends) are inherently broad and contain heterogeneous populations. As such, demographic biases or data inaccuracies will tend to distribute across all geographies within an MVPD's service area, resulting in a fairly consistent means of ranking zones based on a given target audience incidence.

For advanced advertising applications – target audience definition and granular media planning – where compiled demographic data is integrated with STB data and shopper data (in a non-personally identifiable manner) – segment biases and data accuracy ***inherent*** in the compiled database becomes less critical due to the fact that companies integrating the data to support the various advanced advertising applications perform their own data verification, balancing, and weighting on the integrated demographic/shopper/STB dataset to assure accuracy in multiple dimensions – demographic, geographic, product, and viewership. As such, any issues inherent in

the underlying demographic data are usually taken care of prior to the combined dataset being used for advanced advertising applications.

With the above in mind, the following details how the compilers interviewed address each of these key issues.

Demographic Bias

As discussed previously, no database is 100% accurate, and there are some demographic biases inherent in any multisource compiled database. In general, these biases result in under-representation of population segments that have limited numbers of personally identifiable data sources to contribute to a multisource compiler. Among these biases are:

- **Top tier income households** – this affluent population segment tend to have vehicles and homes that may not be directly linked to a specific household (corporate leases, for example), unlisted phones, assets that held by trusts, and are generally not survey responsive. As such, they may be represented on a compiled database, but the underlying data may not be indicative of their high level of affluence.

- **Low income or impoverished households** – this economically challenged population segment lacks many of the sources of data used by a multisource compiler – no phones, no real estate, no vehicles, do not receive surveys, etc. As such, they tend to be under-represented on compiled databases.

- **Young adults (emerging consumers)** – once again, limited sources of public information tend to under-represent these individuals on the compiled database. They are often present, but as part of their parent's household, and will have limited individual level data available.

- **Institutionalized individuals/group quarters** – these typically older individuals are also under-represented on compiled databases.

Multisource compilers are very cognizant of rules, regulations, guidelines, and philosophies that govern the use of consumer data. The use of data on children is highly regulated. As such, children – predominantly only age and gender information – are typically attributes only associated with households, with virtually no other information maintained on interests or activities. Other

segments of the population may be represented on the compiled base, but key pieces of data will be absent because there are limited information sources for this data. For example, property deeds and tax assessor information does not include renters, leading to limited home and apartment rental information. Unlisted phone numbers are not maintained on white pages directories, and the number of consecutive years listed at the same telephone number/address is used to determine length of residency. As such, length of residency for unlisted phone numbers will be slightly less accurate than that of households with listed phone numbers.

Aside from the demographic segment bias, there will also be data coverage biases. These will always be associated with the sources of data. Transactional information on mail order, internet, and retail purchases, and magazine subscription data that indicates lifestyle interests will always have a bias that reflects the specific sources of data that are contributors to the compiled database.

Another type of bias that can be found in compiled databases relates to financial characteristics that may have significant variations and impacts based on geographic bias. An example of this would be a specific income range that could be considered to be affluent in the rural Southeast, and subsistence in an urban Northeast city. Many of the compilers maintain fields that adjust for these regional differences, and categorize data such as income and home value in terms that relate to the local geography. For example, a specific income range might represent the top 10% of households in a rural community, but be ranked significantly lower in an urban center.

Explicit vs. Inferred or Modeled Data

Most multisource compilers will primarily use known or explicit data, and model or infer data where there are either gaps, or where data is not available, reliable, or deep. One example, of where a characteristic is not available explicitly is “length of residency”. This important tenure variable is inferred from other sources of data such, listed telephone directories, property deeds, magazine subscriptions, purchase transactions, etc. While inferred from these various sources, this information is typically as accurate as explicit data such as age.

Most, but not all, multisource compilers will not blend known and inferred data, but prefer to keep them separate and identifiable in order to provide database users with the greatest flexibility in using the data for their specific applications.

While the composition of each compiled database is different, with regard to explicit data there are universal similarities:

- Psychographic (lifestyle) information is 100% explicit
- Behavioral (transactional) information is 100% explicit
- Geographic information:
    - US Census – 100% explicit for short form data, long form data is projected from 1 in 6 samples
    - Summarized auto data – 100% explicit, but aggregated to the small area geographic level
    - Aggregated credit data – 100% explicit, but aggregated to the small area geographic level

The base demographic data comprising compiled databases is typically 70-80% explicit information, with the balance either inferred or modeled. In many instances, the modeled or inferred data is used to bring the explicit data more in line with known national data or statistics (weighted to conform to census data).

Given the number of characteristics maintained on compiled databases, there will likely be variations in explicit versus inferred or modeled data on a per element basis. Here are just a few key characteristics and there composition (as provided by interviewees from data companies):

- **Hispanic ethnicity** – 100% inferred from consumer name tables and small area geography (census data detailing high concentrations of Hispanic households)
- **Pet Owners** – 76% known data from self-reported surveys, 24% modeled to conform with known number of pet owning households
- **Adults indicating interest in foreign travel** – 52% known data from self-reported surveys, 48% modeled
- **Households with children 2 and under** – 100% known, but older children age ranges are modeled

One issue that did surface is the fact that, in many instances, self-reported data is updated only when new survey responses are acquired for a given household. This means that if a household stated in a prior survey that they were a dog owner, and no subsequent survey information was received, that household would be considered to a dog owning household for a lengthy period of

time. If, on the other hand, a survey is received from this household that indicates that they don't own a dog, their record would be updated accordingly. From the interviews, there is no consistency as to how "aged" self reported data is treated.

Data Accuracy Verification

All of the major data compilers use different verification procedures that typically encompass the following similarities:

- Key demographic characteristics – age, income, home ownership, ethnicity, etc. – are compared to US census statistics and adjustments made using those records that don't have explicit data for these characteristics to adjust the totals.

- For multiple data sources for the same characteristic, each source is weighted based on reliability, and consensus across sources is extremely important. For example, if 4 of 6 sources for age agree, and are high confidence sources (i.e., license or registration data vs. survey responses), age will reflect the consensus sources.

- Geographic coverage for each demographic is also adjusted based on US census statistics

There are also a number of quality control procedures that contribute to overall data accuracy that are periodically performed as part of each compiler's database update procedures. Some compilers use proprietary data sources as part of verification and validation procedures. For example, Experian uses data from the Simmons consumer research panelist responses (explicit responses only) as a means of verifying database composition and accuracy.

Modeling to fill in, or augment, missing data uses a traditional approach to assure accuracy that starts with a universe of individuals where the desired characteristic or behavior is known. This group is then split into representative development and validation samples, each of which is appended with all available database characteristics as candidate variables for the model. Once the final model is developed, it is applied to the validation sample to determine how well the model predicts the desired characteristic or behavior.

As with all models, they are most effective and accurate at the extremes of the model's distribution. High scoring model groups tend to have very high concentrations of the desired audience; whereas, the lowest scoring model groups tend to have extremely low or no

concentration of the desired audience. As such, the compilers only assign modeled values for individuals and households with high model group scores, and a minimum confidence level of 95%.

It should also be noted, that all of the major compiled database suppliers maintain extensive and skilled analytical resources. While variations do exist, in general, users can feel confident in the modeling capabilities of these firms.

Shopper Databases – Pros, Cons, and Issues

Shopper data is gaining importance in advanced and addressable TV advertising applications as advertisers see the value in defining their desired target audiences based on actual product and brand purchasing behavior, rather than on broad demographic definitions.

Conceptually, any manufacturer or retailer that captures information on purchases made by specific consumers can be considered to maintain shopper data. These databases include product warranty bases for electronics, automobiles, and other consumer durable goods, as well as financial services company data encompassing checking, savings, investment, and credit card product information linked to the consumer.

In all of these examples, the data is highly biased because it will typically only represent a single company or brand, and a consumer audience that is likewise biased by the nature of the manufacturer or retailer.

As such, all of the companies interviewed utilize shopper data that is far broader with regard to products, geographies, and consumer demographics. While shopper data can be derived from a number of different sources – product purchase warranties, consumer surveys, etc. – the companies interviewed tend towards using large databases derived from actual retail purchases that have been linked to consumers by loyalty, or frequent shopper programs, and are focused on consumer package goods and over-the-counter pharmaceuticals purchased at grocery chains.

In the US, the two primary sources for broad reaching CPG shopper data are Catalina and Cannondale. Dunnhumby – while having a presence in the US – is more extensively used in overseas markets, although they do handle one major US grocery retailer. In all instances, the core applications that these companies use shopper data for are not specific to media planning. All of these companies primary leverage shopper data to provide a range of consulting, marketing, and promotion services to manufacturers and retailers. The use of shopper data for target audience definition for advanced advertising applications is a relatively new phenomenon.

Each of these shopper databases can be considered to be multisource in nature, but not in the same sense as the multisource compiled databases. In the case of shopper databases, multisource references the fact that data is acquired from numerous retailers and store locations that provides excellent coverage at the product and brand level (consumer shopping basket), but does not provide corroborating sources for a specific characteristic (i.e., compiled databases may have 4-5 sources corroborating a consumer age, whereas, shopper databases may have 4-5 sources that all sum to the total shopping basket for a consumer or household, but don't necessarily corroborate the products purchased).

Scope and Composition of Shopper Databases

Both Cannondale and Catalina are extremely large databases that encompass large numbers of consumers, retailers, and products:

- Household coverage: 50,000,000 – 60,000,000 households for CPG products, 100,000,000+ patients for prescription medications
- Retail store coverage: 15,000 – 25,000 individual grocery locations, 15,000+ individual pharmacy locations
- Volume coverage: approximately 85% of CPG purchases are captured

The data captured and maintained consists of:

- Frequent Shopper Card Transactions including:
  - Date and time stamp of shopping event
  - Store number of shopping event
  - UPC codes for all products purchased
  - Number of items purchased per UPC
  - Price for each item at the UPC level
  - Price category (regular, store promotion, manufacturer promotion)
  - Product category

- Syndicated Panel Data
  - This data – typically IRI/Homescan data, or manufacturer supplied data, is used to provide product, market, and demographic benchmarks necessary to adjust the shopper data to be representative of all geographies, audiences, and product category and brand volumetrics

If we assess shopper bases as we did compiled bases in terms of coverage, breadth of data, depth of data, latency (age of data), and accuracy of data, we arrive at the following conclusions:

Coverage

With coverage of 50 - 60MM households, the primary shopper bases used in advanced advertising support provide excellent coverage across the domestic household population. Given that the underlying purchase data is derived from frequent shopper cards from specific retailers, we would expect some degree of bias in demographic segments represented, and potentially geographic coverage. On the other hand, the sheer size of the household population provides sufficient quantities to weight and adjust the total household universe so that a subset of all households can be derived that is properly weighted with regard to consumer demographics, geographic coverage, and product and brand volumetrics.

One other point of importance is the fact that shopper data is captured at the individual level, but is aggregated to the household level for use in advanced and addressable advertising applications. As such, the data represents the household's shopping behavior, not the individual's. Also, it is probable that not all household purchases are captured, since some purchases may be made without the use of the shopper card, or at non-participating retailers. These gaps in purchase behavior are accommodated during weighting and balancing prior to the shopper data being used for media applications.

Breadth of Data

Unlike compiled databases that encompass a broad range of data across numerous categories, shopper data is basically restricted to only grocery or prescription pharmaceutical data as previously described. It should be noted that all items in the consumer's shopping basket are captured, including the number of units for a specific UPC. In some instances, this information is augmented with basic demographic data about the shopper, but typically these represent less than 30 characteristics. It is also important to note that shopper data is always used in a non-personally identifiable manner when linked to the specific purchase data.

Depth of Data

Although the number of unique data elements related to shopping transactions is limited, there is tremendous consistency to each transaction. Such is the nature of automated data capture at POS – consistency and accuracy is assured, due in large part, to the financial nature of the data being captured.

Latency

Shopper data is collected in real time by the underlying retailers; however, the processing of this data by end users – notably, the research and measurement companies providing advanced advertising solutions – occurs less frequently. While monthly processing would be desirable, most of the user companies incorporate new shopper data on a quarterly basis when performing their compiled data updates. This minimizes processing costs, and assures that both compiled and shopper data are consistent with regard to age of data.

Accuracy

In its raw form as received by end users, shopper bases should be considered as convenience samples, rather than total representative universes. The underlying purchase transaction data is ostensibly accurate with regard to specific items purchased and their associated pricing. On the other hand, the overall composition of the total base – with inherent geographic, demographic, and volumetric biases based on the sources of the data – requires that shopper data be significantly processed, weighted, and balanced on multiple criteria prior to it being used for media planning applications.

It is to the credit of all companies that work with shopper data in support of advanced advertising applications that they readily acknowledge and address these issues before the data is used.

Shopper data is proving to be extremely valuable in support of advanced advertising applications. Target audiences can be defined with desired precision – in many instances, combining category and brand consumption in combination with desired buyer demographics – and then used to plan and deploy advertising campaigns at both the local market and national levels that get impressions to the desired audience with optimal efficiency.

Miscellaneous Databases and Their Applications

As previously discussed, virtually any list or database can be used to support advanced or addressable TV advertising; however, the applications of these databases tend to be highly specialized.

Credit Data

Credit data is provided by the three major credit bureaus – Experian, TransUnion, and Equifax – and is derived from multiple financial services and insurance sources that provide credit information to the bureaus in exchange for being able to use the data for their core businesses. Detailed, individual level credit data from the three major bureaus could, in theory, be used for advanced and addressable applications in a manner similar to that of shopper and compiled data – namely, to define a target audience in terms of desired credit information (credit worthiness, affinity to a specific credit product, etc.) – however, the use of this data is governed by specific regulations that restrict the use of this information to only those instances where a credit product is being offered to a specific consumer. As such, the use of this information at the individual or household level is not permitted for general marketing and advertising applications where credit is not being directly offered. That said, a case could be made that "blind matching" would allow the use of this information in a non-personally identifiable manner analogous to that of STB and shopper data; however, no use of credit in this fashion has yet been identified.

The alternative approach to using credit data is at the anonymous geographic aggregate level – most often the zip+4. In this instance, the underlying credit data is summarized to provide virtually all the same credit information that is available at the individual level (basically, type of loan or credit product, payment performance, delinquency information, balances for each type of loan product from first and second mortgages, to auto and home improvement loads, to all types of credit cards), but in a non-personally identifiable geographic format. At this level, the aggregated geographic credit data is still very applicable for target audience definition, media planning and addressable advertising – especially for addressable applications that are performed at the geographic head end/cable zone level.

In terms of the criteria used to evaluate databases, aggregate credit data has:

- Broad coverage across all credit active households
- Broad range of data elements (characteristics) – albeit, restricted to credit related attributes

- A large number of characteristics for each credit/loan product – by product and in total across all products
- The age of the data is excellent in that the underlying individual credit data is updated daily, and the summarized data is created from the individual data on a monthly basis
- Accuracy of the credit data – either individual or aggregated is excellent since the underlying sources of data are financial institutions and their systems. On the other hand, there is no such thing as 100% accurate data, and identity theft has caused issues with credit information. That said, credit data aggregated to the geographic level may actually be more stable than individual level credit data, because individual data anomalies are less likely to be significant when aggregated with accurate data.

Financial services advertisers are most likely to use aggregated credit data as part of the advanced advertising audience definition. In most instances, the target will be defined in terms of desired demographics (income, life stage, home value) and residency in neighborhoods (zip+4 areas) having the desired credit composition. Additionally, these geographic characteristics can be easily used for geo-level addressability by identifying high potential cable zones or markets for specific credit or insurance products.

Automotive Data

Automotive data is similar to credit data in that its use at the individual level is restricted to specific automotive related applications – recalls, auto insurance, etc. As such, the 500MM+ vehicles in the US are summarized – much like credit data – to the small area geographic level, with zip code and zip+4 being the most granular units.

The summarized automotive data provides broad coverage across all 50 states, with the breadth of data limited to those summary level characteristics that can be used in a compliant manner. These include:

- Vehicle class segments – economy, luxury, midsize, minivan, pickup, SUV, etc.
- Manufacturers
- Bought new, used, leased
- Age ranges for vehicles
- Vehicle price ranges
- Specialty vehicles – hybrid, motorcycle, camper

As with most geographic data, such as census and aggregate credit data, summary automotive data is provided in terms of the percent of the population for a specific characteristic (i.e., percent of population owning luxury cars), as well as the number of specific vehicles in a geographic area. Formatting the data in this manner provides an easy means of ranking geo areas based on a desired characteristic.

There are two primary providers of automotive data domestically – Experian and Polk. Both companies use multiple sources to compile the automotive data including vehicle registration data from the states, vehicle warranty data from manufacturers, and vehicle financing information. As such, the coverage and quality of data is excellent, but the granularity of summarized auto data is deliberately higher level – manufacturer and vehicle class (BMW cars), instead of make and model (BMW 535i), for example. Updates to the automotive databases are frequent as different update sources are applied to the database continuously. The summary geographic base is created on a monthly basis.

The applications of summary automotive data are analogous to those of geo level credit data, but with a focus on vehicle related advertiser needs.

Geographic Data Pros and Cons

Geographic level credit and automotive data are valuable assets to have available for advanced and addressable TV use, especially for advertisers where this type of information is valuable in defining desired target audiences. Given the highly regulated nature of both credit and vehicle/drivers' data, aggregated or summarized credit and vehicle data is the only permissible means of using this information with broad coverage in both geography and consumer segments.

On the other hand, limitations do exist with summarized data. In order to depersonalize the data via geographic summaries, both these geographic databases lack granularity in key characteristics. For example, is it sufficient to know only that a zip code has a high concentration of luxury car owners, or do you need to know that they're owners of Mercedes 550s, BMW 750is, Audi A8s, and Jaguar XJs? The lack of precision, or granularity, is the most glaring short-coming of geographic data. That said, this geographic level data can easily be used with most advanced planning approaches in combination with other data – demographics and STB data – to add desired tonality and precision to target audience definition and media planning.

The bottom line is that all geo-aggregated data provides the ability to target high potential areas, but not specific individuals.

Advertiser Supplied Data

One means of using highly accurate and proprietary data in an advanced or addressable application is to utilize data supplied by an advertiser or their agency. A client supplied list can easily be accommodated in the "blind matching process" that is performed by trusted third party processors to derive a non-personally identifiable (non-PII) composite database consisting of demographics, STB viewership data, and information from the client base. The client supplied data can then be used to precisely define, or refine, target audience definitions, and then use the resulting targets for planning or addressable advertising deployments (household or geographic). The client list could be of known customers, or the results of research projects. The only criteria is that name and address is provided to assure the highest degree of accuracy and precision in the matching process, and that the quantities are sufficient to assure that the resulting matched universe is of sufficient size to provide statistical validity.

The precision offered by client supplied data is very beneficial, but not without costs. There will additional costs incurred to the advertiser in incorporating this data into the blind match process, and additional costs to properly integrate this data with the balance of the composite database. It is likely that a client supplied file will be highly biased – geographically or demographically. As such, some weighting and balancing will be required in order to use client data within an advanced planning environment to assure accuracy and validity.

Database Matching Approaches and Privacy

Introduction to Database Matching

Whether supporting target audience definition, media planning, addressable campaign deployment, or post campaign measurement, the ability to accurately match disparate databases in a consumer privacy compliant manner is integral to both advanced and addressable TV advertising applications.

Depending on the specific use, two or more databases will be matched to result in a consolidated base of information that will then be used to support advanced and addressable TV advertising applications. Matching can be performed at the individual, household, or geographic level depending on the specific need and application. Additionally, some applications require matching at multiple levels concurrently in order to achieve the desired results.

The companies providing the database matching services are, in large part, the same as those providing the multisource compiled databases – Experian, Acxiom, and Epsilon – to name three; however, any company with experience in supporting direct and database marketing services will have sufficient experience and expertise to perform database matching.

While database matching can be performed on any number of databases, the most common databases that are matched consist of matching MVPD subscriber bases to compiled databases and one or more ancillary bases (shopper data, research panels, etc.) to develop a consolidated base of information that links – in a non-personally identifiable manner – consumer demographics, purchase behavior, and television viewing for the purpose of supporting advanced advertising applications.

It is important to note that, in many instances, the choice of the matching vendor is made by the MVPD or STB processor – not the advertiser or agency. MVPD's are most likely to have established relationships with companies that perform the matching, wherein, their subscriber bases are appended with third party data to support core marketing programs such as up selling and cross selling. As such, it is logical – for both data security and cost reasons – that these established relationships be leveraged for advanced and addressable applications.

Individual versus Household Level Matching

Matching at the individual level is the ideal scenario that would allow advertisers to absolutely reach the right person with their message; however, the reality is that this goal is not achievable, nor viable, due to a number of reasons including:

- Compiled databases only maintain a limited number of individual level characteristics (gender, age, occupation, etc.), with the majority of the data related to the household (home ownership, income, etc.). Additionally, compiled databases designate a head of household – typically, the primary wage earner – with all other household members linked to this individual.

- MVPD subscriber files – to which compiled databases are matched for access to STB level data – only maintain a single individual as the subscriber account holder. In many instances, this individual may or may not be the same "head of household" on the compiled base. As such, household level matching allows for a "match" to occur, even if different individuals are present on two or more of the databases being matched.

- From a matching technology standpoint, the difference between individual and household level matching is the presence of the first name as part of the name and address matching process. Given the number of first name variations that could be found on various lists and databases – nicknames, initials, misspellings – the inclusion of first name – or individual level matching – significantly reduces overall match rates.

For these reasons, all of the interviewees participating in this study database confirmed that matching was performed at the household level. This assures the highest match rates between files, and since the compiled database represents data on multiple individuals within each household, the precision necessary to analyze and plan based on one or more members within a household is still present.

Matching Technology

Name and address matching is a mature technology that has been refined based on over 35+ years of commercial use, predominantly in the direct marketing industry. The most sophisticated approaches utilize all data fields present in a consumer name and address record, and extensive business rules to assure the highest match rates between files or bases being matched, without introducing inaccuracies by matching records that do not represent the same households.

Each supplier of matching services will use different approaches to optimize match rates, with customization required to accommodate variations in name and address fields across the various databases – MVPD subscriber, compiled, shopper, research, and client – that could be part of a match process. Generally, the results of the match processing will be provided in terms of the number of records – on each of the files – that were matched to one of more of the other files. The match statistics are ranked by the type of match:

- Exact name and address match – this is the highest degree of matching, wherein, both the surname and all address components match between files

- Close name and address match – there are usually a number of close name and address match categories, all of which accept a certain degree of variation between name and address data between matched records. In many instances, the variation might only be the presence or absence of an apartment number, or a slight variation in name spelling or street address.

- Address only match – this is the lowest level of matching and considers records matched if only the address matches. While not highly accurate, this level of matching is sometime used in high mobility areas – such as college housing neighborhoods. Matches at this level are rarely used in advanced and addressable applications.

In addition to household level matching, geographic matching is also performed. The geographic data is typically matched at the lowest geographic level, with postal zip+4 being the most granular unit. Zip code and NCC cable zone appends are also performed.

Match rates between files can vary significantly, with geographic matching having the highest match rates – nearly 100% at the zip code and cable zone levels.

Household match rates can vary significantly based on the nature and size of the files, and it is difficult to provide average match rates. A typical blind matching process for a provider of advanced advertising services might include:

- 125MM compiled database households
- 10MM MVPD subscriber households
- 30MM grocery shopper households
- 10MM pharmaceutical households

The intersection of this bases – where households are in common to all 4 – is the obvious sweet spot, since all information from all data sources is available. This matched universe cannot exceed the size of the smallest file – in this instance, less than 10MM households – and is likely to be less than that (7-8MM).

In general, when working with databases that are large – millions of records – household level matching will result in a composite universe that will be sufficiently large to provide for balancing and weighting, and still provide statistically valid quantities to support target audience definition and media planning at very granular levels.

Aside from the software and business rules used for household and geographic level matching, accuracy and match rates are also subject to variations in address quality – basically, how addresses are formatted and whether they are up to date. Many of the companies providing matching services will perform extensive address standardization and updating prior to performing the match. The primary processes that are performed are:

- National Change of Address Processing (NCOA) – updates addresses
- Locatable Address Conversion System (LACS) – standardizes address
- Delivery Sequence File (DSF) – identifies residential/commercial addresses and verifies deliverability (actual address)
- Deceased Identification – eliminates deceased individuals

Files that are non-transactional in nature, and may not have been subject to recent updating, should always be processed through these hygiene procedures prior to matching. Client supplied files such as warranty data should be processed in this manner.

Blind Matching

Blind matching is the term used to describe the matching of personally identifiable databases to result in a non-personally identifiable composite base that incorporates data from all input sources.

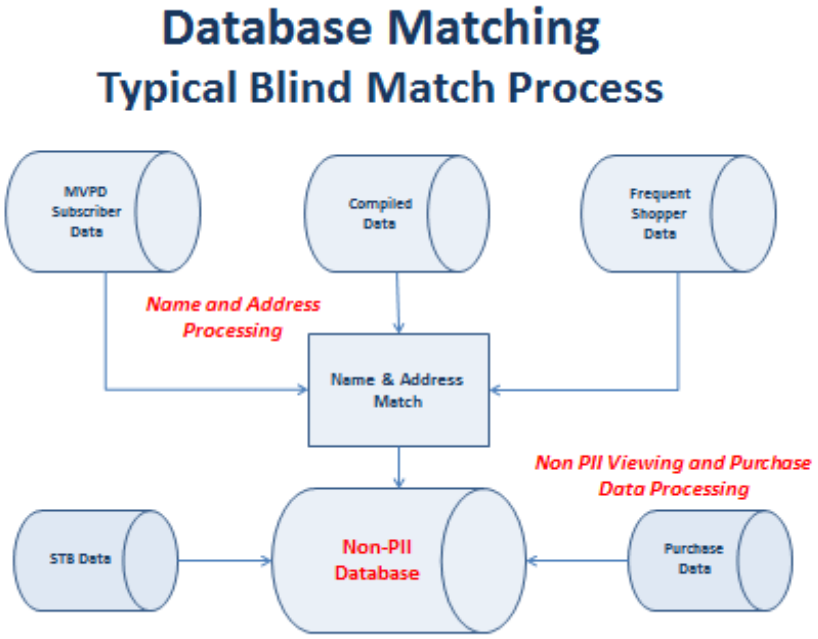
This matching is typically done at the household level, by a trusted third party processor, and will minimally include the following types of input databases:

- Compiled database – provides demographic, psychographic, behavioral, and geographic data
- Shopper database – provides product purchase data
- MVPD subscriber database – provides audience data and links to STB viewing behavior though STB ID's linked to each subscriber household

Once the match is completed, all personally identifiable name and address data is deleted and the resulting composite base contains all information necessary to support advanced and addressable advertising applications in complete compliance with consumer privacy regulations.

The choice of which third party processor performs the blind match processing is often made by the MVPD that is contributing the subscriber and viewing behavior data. Most MVPD's have existing relationships with database services companies for up sell/cross sell applications for their core services, and using the same company for blind match processing minimizes the number of companies that have access to their subscriber bases.

Another point that should be kept in mind is that blind matching is always used to integrate STB viewership data with other categories of data (demographics, shopper data); however, the actual viewership data is never directly associated with an individual. The actual STB transaction data is never personally identifiable, but is referenced by a STB identification number, and these identification numbers – not the transactional viewing numbers – are related to subscribers on the subscriber account file. It is only the subscriber file that contains name and address, and it is only this file that is processed for blind matching. Once the match procedure is completed, and the resulting composite database has been depersonalized, transactional STB viewing data can be incorporated into this composite base in a completely anonymous and compliant manner.



Consumer Privacy

All of the companies interviewed for this study are acutely aware of consumer privacy, and the need to adhere to strict processes and procedures to assure the proper and anonymous use of consumer behavioral data.

The database suppliers and processors are at the forefront of the processes and procedures designed to assure consumer privacy, but all users of consumer data (measurement and research companies, addressable technology partners) are also cognizant of the need to keep consumer data completely anonymous.

Among the privacy protection processes and procedures employed are:

- **Physical Data Security**
    - Secure physical locations that maintain or process consumer data
    - Extensive network security procedures designed to identify and prevent unauthorized user access to computer maintained data
    - Disaster-proof and, in many instances, redundant data centers

- Systems compliance with ISO 27002 compliance and data security standards
- Mandatory employee training in data security and privacy compliance

- **Regulatory Compliance**

  - Fair Credit Reporting Act (FCRA) – governs the use of credit information
  - Driver's Privacy Protection Act (DPPA) – governs the use of driver's license and automotive data
  - State and Federal Do Not Call, Mail, Email lists – govern channel contact with consumers
  - Gramm-Leach-Bliley Act (GLBA) – governs the use of all consumer financial data
  - Children's Online Privacy Protection Act (COPPA) – governs how information is collected and used for children under age 13
  - CAN-SPAM – governs consumer email communications

- **Consumer Privacy**

  - Implement secure "blind matching" processing for all datasets containing personally identifiable information
  - Assure that all personally identifiable data is destroyed following blind match processing
  - Maintain the output of blind matches in an encrypted format to protect the security of the non-personally identifiable information
  - Adherence to the Direct Marketing Associations guidelines for Consumer Privacy Promise, Ethical Business Practices, and Email Service Provider Coalition guidelines
  - Federal Trade Commission (FTC) guidelines on consumer protection

In addition to the above, the multisource data compilers also adhere to the Fair Information Values Assessment (FIVA) requirements for data acquisition that requires data only be acquired for inclusion in a multisource compilation from ethical data sources that give consumers notice and/or choice about the use of their personal information for marketing purposes. It should be noted that FIVA notice and choice are terms only associated with information collected as part of the marketing process, and do not apply to public records, telephone directories, and other sources of data that are not collected as the result of marketing efforts.

Miscellaneous Privacy Points

The database matching companies need to have access to personally identifiable information – name and address – in order to perform the matching necessary to develop the composite dataset that will be ultimately be used by companies to support advanced and addressable applications. On the other hand, the matching companies never take possession of any personally identifiable shopping data or viewing data. All they process are the names and addresses of consumers that are on the shopper base or the MVPD subscriber base. The output of their processing consists of a completely depersonalized database with keys that link to transaction data for shopping behavior and television viewing.

On the other hand, the advanced and addressable advertising companies that maintain these composite databases never take possession or maintain any personally identifiable information, but have the ability (and need) to incorporate shopping and viewership data to support their product and service offerings. This division in responsibilities assures that consumer privacy and anonymity is maintained through processing and use of consumer data.

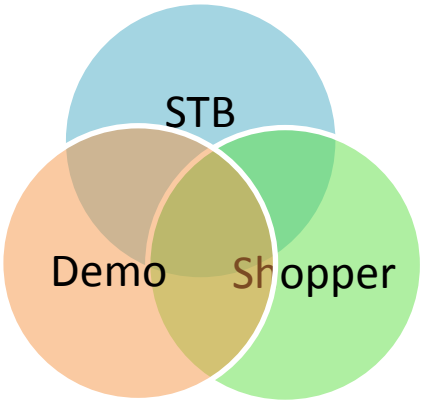
Another consumer privacy safeguard is that small samples – less than 30 households – are precluded in virtually all advanced advertising systems and addressable advertising platforms. Aside from statistical reliability concerns for planning, small samples start to encroach on privacy with regard to targeted advertising.

One other point concerning privacy that is not a regulation or guideline, but is becoming increasingly important as identify theft increases, and breaches in email and other consumer databases become more frequent. The term “behavior” has an extremely negative connotation – especially to those individuals (notably legislators) that do not understand data and database terminology. To these individuals, the mere mention of “behavioral” data – whether it is shopping behavior, catalog purchase behavior, or television viewing behavior – raises eyebrows and concerns about misuse of consumer information. On the other hand, if “behavioral data” is replaced with the phrase consumer “profile information”, concerns are significantly reduced. Profile is assumed to be depersonalized; whereas, behavioral is assumed to be personally identifiable.

Assuring Database Accuracy and Projectability

As noted previously, all of the data sources that are primarily used to support advanced and addressable applications – notably, STB, demographic, and shopper data – do not, in their raw forms, stand the rigorous requirements of researchers and media buyers and planners. As such, all of the suppliers in this area perform significant adjustments prior to using a subset of the composite data that best provides an accurate representation of the US population with regard to demographics, category and brand purchasing, and television viewing.

All of the STB data processors interviewed for this study create composite universes at the household level using blind matching procedures performed at trusted third parties that result in totally anonymous bases at the intersection of all the input sources.



The larger the size (in households) of each of the input bases, the greater the potential to have higher match rates between all three bases, and the larger the composite universe formed at the intersection of these bases. Unfortunately, the biases inherent in each of these bases result in a highly biased…or, unrepresentative…universe of matched records (anonymous households).

As such, the matched composite universes are manipulated to result in subsets that are far more representative of the "true" population with regard to the aforementioned key metrics.

While each of the STB processors uses different approaches to adjust the composite universe – many of which are proprietary in nature – there are common aspects to each.

Geographic and Demographic Adjustments

Given the large size and broad coverage inherent in the demographic and shopper databases used for advanced advertising applications, the first level of adjustments – typically, weighting and rebalancing – occurs at the geographic and demographic segment levels.

Several companies perform this balancing procedure by creating a number of demographic/geographic segments that encompass characteristics such as: age, income, household size, ethnicity, and county size, and then balance these multivariate groups geographically to represent both the total US population, as well as key DMA markets.

Following this balancing procedure, other adjustments (weighting) are performed to assure that each of the balanced segments provides accurate representations of viewership and product purchases.

Other companies attempt to perform the geographic and demographic adjustments concurrently with television subscriber data by creating groups that encompass demography, geography, and key television related data such as HD vs. non HD, presence of DVR, subscriber tenure, etc.

For all of these companies, the number of unique cells or segments that are balanced is typically greater than 100; thus, providing a fairly high degree of granularity to the balancing process.

The base to be used for the geographic and demographic balancing is the US census at the small geographic area level (typically, zip codes).

Shopper Data Adjustments

In all instances, adjustments required to provide accuracy and projectability of shopping behavior are performed after the geographic and demographic balancing has been completed. As noted in previous sections of the whitepaper, shopper data has inherent biases in terms of demography and geography based on the nature and location of specific retailers that permit linkage of shopping transactions to their customers. Balancing shopper data – in terms of category and overall spending behavior – is best done once representative geographic/demographic segments have been defined.

In this procedure, the shopper data for the geo/demo balanced segments are compared to reference data, and adjustments (in the form of weighting or rebalancing) are performed on an individual segment basis. Three sources of reference data were identified:

- IRI/Homescan
- Proprietary shopper research data
- Client (advertiser) supplied purchase data

Regardless of which reference base was used, category and overall purchase data are adjusted by segment to be analogous to data from the reference sources.

STB Data Adjustments

Whereas the adjustments made on the geographic, demographic, and shopper levels are important components of assuring the accuracy of the advanced and addressable services provided by companies in this sector, adjustments to assure the accuracy and representativeness of viewership data are of critical importance.

As previously discussed, the STB data used by all companies in this sector is highly biased based on the geographic coverage of the underlying MVPD sources (especially, cable operators), or demographic coverage of the subscriber bases (especially, the satellite companies).

In all instances, the reference base that is used is the currency of television – namely, the Nielsen AMLPD data – and the balancing metric is GRPs.

While the reference base is consistent, the approaches employed to achieve the proper weighting or balancing differ greatly. In many instances, viewership data is adjusted so the data for each of the geo/demo segments mirrors that reported by Nielsen. In many case, these adjustments are highly consistent with Nielsen, whereas, in other instances, the reported viewing data is not consistent with Nielsen, but is consistently off (either consistently under or over reported). While not ideal, consistent misrepresentation can be accommodated for planning applications.

Some STB suppliers go a step further by attempting to adjust the STB data in terms of representation of television media strata – cable, satco, telco, over the air In this application, some of this channel allocation data will be accurate based on explicit STB data sources, whereas, other allocation data will be modeled.

Miscellaneous Adjustments

All of the companies in this sector are staffed with researchers and fully understand the nature of the disciplines required to support accurate television planning applications. As can be seen from the discussion on accuracy and representation, significant efforts are undertaken to assure the most accurate data and platforms for advanced advertising applications.

In addition to the above, several other tasks are performed including:

- **Set on/set off adjustments**
    - All STB processors have a mechanism to determine whether a TV is on or off, though the mechanics vary by processor. Some use explicit information as the reference base (Nielsen People Meters, IPTV boxes), whereas, others have developed sophisticated models that look at STB data by program type, day of week, and daypart, and calculate the probability of a set being on or off. In either case, these adjustments are factored into the viewership data reporting.
- **Over the Air broadcast adjustments**
    - The adjustment of over-the-air households is done on an assumptive basis by most STB suppliers with the viewership adjustments typically made by market and network.

In addition to these processes and procedures, an entire plethora of ancillary tasks are performed just to assure the consistency and accuracy of STB data prior to its being subject to the aforementioned procedures. These procedures are basically data processing quality assurance procedures and are rigorously adhered to by all STBV processors. One should not discount the amount of effort required to process STB data. An average STB daily feed is approximately 2 gigabytes of data, and needs to be processed and validated each day.

Observations and Conclusions

The use of ancillary databases that are matched to STB data provides significant benefits to the television industry. In the broadest sense, it is this data that gives anonymous STB data the consumer insights and tonality necessary support advanced and addressable applications. Without it, STB data is merely a base of transactions devoid of personality.

The databases that are being matched to STB data are generally very accurate, and subject to minimal biases that can easily be accommodated.

The matching processes that are used to construct these integrated bases are focused on assuring consumer privacy, while providing highly accurate matching at the consumer household level.

Another benefit of these third party databases is that it provides consistency with other media in the area of target audience definition. The same databases that are being married to STB data for television applications are also being used to define and target audiences in direct mail, email, and online advertising. As such, the potential exists to plan and deploy integrated multimedia campaigns that use a consistent audience definition across all media…not a translation, but the same exact criteria applied at the same anonymous individual/household level.

All of the data, processing capabilities, and analytical skills necessary to fully leverage the benefits of databases matched to STB exists today. The one significant issue that needs to be addressed is cost. There is a high cost associated with processing and using the vast quantities of data that result from matching very large databases, and it is still to be seen whether these costs can be justified through improved campaign efficiency and effectiveness.

Appendix

CIMM Definitions

Advanced Advertising

See also: Addressable Advertising, Interactive Advertising, Customized Advertising, Dynamic Advertising, In-Navigation Video Ads

CIMM DEFINITION: A range of advertising solutions designed to leverage the interactive nature of digital Set-Top Boxes and enhance the value of TV by offering, for example, request for information, polling and trivia , Telescoping, Ad-Versioning Dynamic Advertising and T-commerce applications via the television through the use of the Remote Control.

NOTE - "Advanced TV Advertising capabilities should include Addressable Advertising, Interactive Advertising, Customized Advertising, Dynamic Advertising, and Measurement. Addressable Advertising would include TV ad targeting based on geographic, viewer or household segment attributes. Interactive Advertising would include things like the use of polls and voting mechanisms but these do not have to be addressable. Customized Advertising entails the ability to efficiently and automatically customize video in real or near-real-time so that the ads can be made more relevant to each of the targeted segment(s) of viewers Dynamic Advertising would entail the ability to update the content of an ad in real-time or near-real-time basis based on automated data feeds (e.g. changes in local TV ads based on local weather conditions or inventory data). Measurement would include the ability to obtain census level campaign metrics based on STB data." (Source: Visible World)

Customized Advertising

See also: Advanced Advertising

CIMM DEFINITION: The ability to efficiently and automatically customize video in real or near-real-time so that the ads can be made more relevant to each of the targeted segment(s) of viewers. (Source: Visible World)

NOTE - Just as on the Internet, efficient customization has become a critical component of dynamic web-pages, online video and TV ads can leverage efficient video customization to enhance message relevance. (Source: Visible World)

CIMM Definitions

Addressable Advertising

See also: Advanced Advertising, Versioning

CIMM DEFINITION: Advertising that is directed to specific geographies or audiences to increase its relevance.

2: "An advertisement sent to a specific home, Set-Top Box or geography." (Source: Nielsen Media Research)

3: "Specific video advertisements that target a set of audiences, homes, or Set-Top Boxes. Such targeting can be based on viewer information including thematic, geography, demographic, and /or behavioral data. Such targeting techniques can be applied to various video services including broadcast, SDV, DVR, and /or VOD program channels." (Source: BigBand Networks)

4: "An advertisement or interactive enhancement that is presented to a specific subset of STBs in the universe/footprint. Alternatively, a collection of advertisements or enhancements that are broadcast to the universe/footprint, from which a single advertisement and/or enhancement is individually selected and presented to each STB." (Source: FourthWall Media)

NOTE – "There is a broad spectrum of addressability mechanisms. On one end is the Canoe CAAS architecture that selects a specific advertisement for every STB at each placement opportunity and inserts the chosen ad or enhancement into a custom stream for each STB. At the other end is the AdWidgets system from FourthWall Media, which embeds (binds) EBIF enhancements into spot ads, which are then broadcast to an entire footprint or zone, and once executing on the STB the EBIF enhancement makes a decision about whether or not to present itself on the current STB. (Source: FourthWall Media)

NOTE - We define addressable advertising as the use of data sets to enable more targeted matches between messages and audiences than takes place in the current environment. Addressable advertising can therefore occur at the national, regional, market, neighborhood, household, or individual level." (Source: Visible World)

NOTE – Addressable Advertising allows for multi-advertiser spots and is the foundation for interactivity. (Source: Invidi)

Please refer to the CIMM Lexicon online at http://www.cimm-us.org/lexicon.htm for additional information on these and other terms.

Interview Questions

Advertisers and Agencies

1. Do you use external databases such as Experian or Acxiom to support advanced advertising applications such as media planning, audience segmentation or addressable campaign deployment? Could you give a brief overview of how you've used these databases?

2. Have you deployed addressable advertising programs that utilized STB data that has been matched to one or more external databases or a client's own database?

3. What was the objective of each addressable advertising program...i.e., target specific households, support alternative creative testing, advertise multiple brands for a given advertiser, etc?

4. What was the duration of each addressable campaign?

5. Was the database matching done to define specific individuals, households, or geographic areas to receive the addressable ads?

6. For advanced advertising applications other than addressable campaigns, how was the data used, and at what level was the data matched to other data (STB, subscriber, other)?

7. For addressable campaigns, how large was the size of your addressable target(s) as a percentage of total household population?

8. What company performed the database matching, and were the results of the matching satisfactory in terms of the quality of service provided and the quality of the database matching?

9. What procedures did you, or your database matching company, perform to assure compliance with privacy regulations?

10. How complex was the overall database matching and addressable deployment process? Did the processing require significant staff resources and time to manage the process?

11. Were you able to measure the impact of the addressable program, and, if so, how was measurement performed – sales data, surveys, post campaign database matching, duration of targeted ads viewed?

12. Did you perform any test versus control comparisons between homes addressed or not?

13. Did you calculate the cost savings or efficiency of the addressable advertising campaign? If so, can you share your approach?

14. Did you pay a premium for using addressable advertising? Was the program still cost justified?

15. In assessing the addressable campaign, was the viewing behavior (STB data), or the third party data more important to planning the campaign, targeting the proper audience and deploying the campaign?

16. What other companies or types of companies, participated in deploying the addressable campaign...i.e., content management, measurement, analytics.

17. How many different distribution partners have you worked with on addressable campaigns – i.e., the number of cable operators, satellite operators, telcos

18. What improvements would you like to see in the use of STB data combined with other data sources to either make the processes simpler, or provide better capabilities and results?

19. Would you use addressable advertising again? Why or why not?

Interview Questions

Media Companies

1. Do you use external databases such as Experian, Acxiom, or client/advertiser purchase databases to support advanced advertising applications such as media planning, audience segmentation or addressable campaign deployment?

2. For media planning or advertising sales applications, which types of databases are most useful? – compiled databases (Experian, Acxiom) that provide richer consumer demographics and psychographics, purchase databases (Cannondale, Dunnhumby, or client) that provide either brand or category purchase behavior, or other databases?

3. For media planning or advertising sales, how are these databases used? – to provide better definitions of desired target audiences, to highlight high incidence programming, highlight high incidence segments and markets, other?

4. Have you participated addressable advertising programs that utilized STB data that has been matched to one or more external databases or a client's own database?

5. What was the objective of each addressable advertising program...i.e., target specific households, support alternative creative testing, advertise multiple brands for a given advertiser, etc?

6. What was the duration of each addressable campaign?

7. Was the database matching – for either enhanced segmentation or addressable advertising - done to define specific individuals, households, or geographic areas?

8. For advanced advertising applications other than addressable campaigns, how was the data used, and at what level was the data matched to other data (STB, subscriber, other)?

9. What company performed the database matching, and were the results of the matching satisfactory in terms of the quality of service provided and the quality of the database matching?

10. What procedures did you, or your database matching company, perform to assure compliance with privacy regulations?

11. How complex was the overall database matching process? Did the processing require significant staff resources and time to manage the process?

12. Were you able to measure the impact of the enhanced segmentation planning or addressable program, and, if so, how was measurement performed – sales data, surveys, post campaign database matching, duration of targeted ads viewed?

13. Did you perform any test versus control comparisons between homes addressed or not?

14. Did you calculate the cost efficiency of the advanced segmentation or addressable advertising campaign? If so, can you share your approach?

15. If an addressable campaign, was the viewing behavior (STB data), or the third party data more important to planning the campaign, targeting the proper audience and deploying the campaign?

16. What other companies or types of companies, participated in deploying the addressable campaign...i.e., content management, measurement, analytics.

17. How many different distribution partners have you worked with on addressable campaigns – i.e., the number of cable operators, satellite operators, telcos

18. What improvements would you like to see in the use of multiple databases to either make the processes simpler, or provide better capabilities and results?

19. Would you use these types of advanced or addressable advertising again? Why or why not?

Interview Questions

STB Measurement and Research Companies

1. Describe the types of services you provide that utilize STB data – campaign planning, target audience definition/targeting, post campaign measurement?

2. Who are your typical clients? – advertisers, agencies, media companies

3. Do you utilize external databases matched to STB data, and for what applications...addressable advertising measurement, enhanced (more precise) advertising measurement, addressable advertising deployment, enhanced campaign planning, etc.

4. How many addressable advertising campaigns have you worked on?

5. What types of databases do you use and how do you use them?

6. Who performs your database matching (database owner, marketing services provider, etc.), and are the results satisfactory for your needs? What improvement would you like to see to make the process easier or more accurate?

7. In performing the database matching, what level is used to match the databases – individual, household, geographic, other?

8. What issues have you faced in dealing with STB data (data inconsistencies, geographic coverage, etc.), and which issue presented the great challenge? How have you resolved these issues?

9. What issues have you faced in dealing with other databases (data quality, match rates, data coverage, etc.)? How have you resolved these issues?

10. Please describe some of your most successful uses of STB data combined with other databases?

11. Have you matched STB data to client-specific databases in a complaint manner, and how was this accomplished (i.e., blind matching, etc.)?

12. Based on your experience, is there a minimum penetration level (minimum number or percent of total households) to justify addressability?

13. Have you been involved in analyses to validate whether addressable ads work? If so, what criteria were used to validate the programs, and what approach was used?

14. What trends do see in using STB data with other databases?

Interview Questions

Addressable Advertising Technology Partners

1. Describe the types of services your company provides in supporting advanced and addressable TV advertising?

2. Do you utilize STB data in your services? If so, how is the data used – target specific viewers based on viewing behavior, target specific content based on viewing behavior, other?

3. What issues do you face in using STB data, and how do you resolve them?

4. Who are your typical clients – advertisers, MVPD's, agencies?

5. Do you utilize other databases – consumer or geographic – to support your services? If so, which databases are used and how are they used?

6. At what level of precision – individual, household, geographic - does your technology deliver addressable ads?

7. Does your technology for delivering addressable ads work on all STB's?

8. Has your technology for addressable ads been deployed – not just tested – on all operator platforms (cable, satellite, telco)?

9. How do you verify addressable ad delivery? Is it done in-house or by a third party?

10. What trends do you see in using STB data for addressable TV support?

Interview Questions

Database and Matching Service Providers

1. Describe the types of databases maintained by your company, and indicate which have been used to support advanced TV advertising (enhanced audience segmentation, addressable advertising program deployment, and post campaign measurement). Please indicate at what level the data is maintained – individual, household, geographic, population cluster levels.

2. For each of the databases, please provide the following information:

    a. Size of the database
    b. The categories of data maintained – demographic, behavioral, psychographic, purchasing, etc.
    c. For each category of data, what are the sources of the data
    d. Approximately how many data fields (pieces of information) are available for each category of data
    e. Please provide coverage across the database for key fields on the database (age, income, etc.) as well as the average coverage for all fields.
    f. How many data fields are explicit information vs. modeled of inferred data
    g. For modeled data, what sources are used to construct the models, and how are the models validated
    h. For both explicit and modeled data, what level of statistical confidence do you have in the category of data?
    i. Are there inherent biases in the databases, such as under or over represented population segments? If so, please identify the biases, and what is being done to rectify them.
    j. How frequently is the data updated?
    k. For these specific categories, what percentage of the population do you KNOW to have these characteristics? And what percent do you model?
        i. Hispanic HHs
        ii. Pet owners
        iii. Adults interested in foreign travel
        iv. Households with children 2 and under

3. Does your company provide data processing support of advanced or addressable TV advertising – for campaign planning, target audience definition, deployment or post campaign measurement? Please describe the services you provide.

4. Has your company provided third party matching services that integrate data from your databases with that of MVPD subscriber and STB data? If so, what level is the matching performed (individual, household, segment, geographic)?

5. Please describe the pros and cons of matching at the individual versus household level in terms of accuracy and match rates.

6. How do you assure consumer privacy throughout the matching process?

7. Could you share case studies, or provide process flows, for how you've performed addressable third party processing for your clients?

8. Who are your typical clients for advanced and addressable applications?...advertisers, agencies, MVPD's, research companies, etc.

9. What key issues have you faced in dealing with STB data and third party matching, and how have you been able to resolve them?

10. Describe how your company envisions supporting the growth in advanced and addressable TV advertising, especially with regard to processing and enhancing STB data? Are there specific projects underway?