ENRICHING MEDIA DATA: Quality is Key Requisite for Maximizing ROI

Advisor:
Gerard Broussard, Pre-Meditated Media

June 2015

FOREWARD

"Buying and selling TV advertising based on a richer set of audience metrics is now a reality made possible by integrating audience data and TV viewership from multiple sources. This is an exciting advancement that provides a leap in value to advertisers and should lead to higher ROI for TV campaigns; however, we must be vigilant about the quality of the data. With scores of data suppliers out there, CIMM's research will empower data users to ask the right questions of data suppliers so that we can evaluate its power to inform decision-making." ***Howard Shimmel, Chief Research Officer, Turner Broadcasting***

"Data is fueling the advertising industry as we can see more clearly than ever the pathway to addressability, the holy grail for our marketing spend. Precision is driving ROI higher, at a time when improvements in marketing productivity are being demanded by senior management and shareholders alike. But the data revolution is just beginning and continued improvements in returns will only come from improvements in data quality. We've seen time and again the promise of data-driven precision being undermined by poor data integrity, lack of standards in data hygiene, over-reaches in modeling, and a lack of standardization. CIMM's research and marketplace feedback is important in providing guidance to our industry, clearly articulating the need for process controls and improved data quality across the multitude of data sources. Improvements in data will be a key driver of improved ROI across for all marketers, regardless of industry." ***Julie Fleisher, Senior Director, Data + Content + Media, Kraft***

"Being able to explain the provenance of the consumer data to our media partners is the only way to provide a sustainable decision-making process for them. Transparency and control are the common themes we hear when working with both the media partner and the advertiser. The CIMM report is a leap forward for the industry because sharing the vocabulary and criteria for what successful consumer marketing data is, does, and comes from will help everyone work better together and have the data to prove it." ***Rick Erwin, President, Audience Solutions, Axciom***

"Merkle is pleased that CIMM has undertaken this valuable study to help understand and improve the quality of data, especially since the number and breadth of consumer data sources is accelerating at such a rapid clip. While the effort of combining data from multiple sources is well worth the improvement in advertising performance, it is essential that there is transparency and credibility in that process. And it's also critical to the industry that we have clear detail on the techniques and methods used for consumer profile models, lookalike segments and data validation. The CIMM study provides critical questions to ask about data quality as well as a high-level view of how some data suppliers are providing quality assurance." ***Andy Fisher, Chief Analytics Officer, Merkle***

Contents

INTRODUCTION

In recent years the Marketing and Advertising industry has accelerated the integration of large data sets to provide richer insights for linking advertising and promotional efforts with consumer response. These applications have reached critical mass for digital media placement and are gaining traction in the TV advertising arena. The sheer volume and appearance of different types of data will likely increase in the future while the industry fast tracks Big Data capabilities. While "fast" is the operative word, there is concern among many researchers and analytic experts about the completeness, quality and transparency of these data as well as their applicability for consumer segmentation and marketing and media investment decisions. Furthermore, considering the expanding number of data sources, there's a need to examine the notion of data harmonization and standardization of nomenclature to make the data integration process simpler and more accurate for end-users.

The primary purpose of this document is to help inform CIMM members and the general media community about the quality, recency, consistency and representative aspects of the data being offered up by third-party data providers. The secondary purpose is to provide feedback on the industry's appetite for master data and reporting standardization.

APPROACH

Two modes of data gathering were deployed: 1) a literature search and 2) interview format discussions with Subject Matter Experts (SMEs) on the topic of third-party data quality:

Literature search

- Search criteria: Third-party Big Data and marketing, advertising, media placement, data quality
- Sources: Google searches and WARC literature search

SME discussions/recruitment – 22 interviews were conducted among third-party end users and supplier companies (see Appendix for listing):

- 10 end users – combination of senior research personnel and media placement experts at agencies, advertisers, media organizations and data intermediaries*.
- 12 data supplier companies – senior personnel in third-party data supplier firms

* Data intermediaries are organizations that integrate and process data on behalf of end users but are not considered the primary sources of underlying data

CIMM Big Data Quality Study: Research Snapshot and Action Steps*

Data Integration Frenzy: Helping Marketers Improve ROI! The Marketing and Advertising industry has been frenetically stitching together consumer data from a potpourri of sources that include TV set top boxes, tablets, smart phones, desktop PCs, magazine circulation, retail stores, online stores and warehouses. The goal of combining all these data is to create a sandbox of insights for improving marketing and advertising ROI through enhanced segmentation and targeting.

CIMM Collects Marketplace Feedback On Data Quality - 10 End User companies tell CIMM their key data quality concerns; 12 Data Supplier firms provide their account of data quality procedures.

Data Quality: Clouds Looming in Data Integration Paradise? CIMM's research found that end users of the data like what they see: an uptick in consumer response from media spend but feel there is room for improving data quality. At the top of the list of their concerns are consumer authentication and the custom creation of modeled targets by data supplier firms. Recent studies have shown that the rate of correctly identifying consumer targets by some data suppliers has been abysmally low, particularly within the realm of digital media.

Transparency: To Be Or Not To Be - Some data suppliers are reticent to reveal the "special sauce" ingredients that go into creation of target segments. This non-disclosure, in turn, fans the flames of discontent among end users, who feel that transparency is the key to understanding the business decision-making power of the data. Larger data supplier firms are perceived to be more open about their processes while smaller and middle-size operations, less so.

Time to Take Action - The growth of data integrations will likely progress at a rapid clip for years to come so CIMM is recommending that: 1) End users ask data suppliers for more transparency and 2) industry data quality guidelines should be created

Ask Questions! End users should insist on data suppliers providing detail on the following:

- Data freshness or recency
- Quality of data source
- Consumer authentication procedures
- Target model creation detail
- Target model testing for effectiveness
- Data integration techniques, matching and keys

This request for greater data transparency will likely raise the consciousness of data quality issues and result in data quality improvements

Get Involved! Take part in future industry engagement on data quality

- Data quality guidelines for prepping and integrating multiple data sources
- End user and data supplier participation
- Details to follow from CIMM

* See Executive Summary and balance of document for full CIMM study details

EXECUTIVE SUMMARY

Readers of this Executive Summary can elect to drill down further into any of the topics within the body of the report for greater detail and illustrative examples.

Defining Data Quality

This document focuses on 1) underlying quality – source credibility, recency, consumer classifications, collection method, representativeness and 2) integration process – quality of techniques/methods used to combine disparate data sets. Data hygiene is a third consideration and refers to the cleansing process that addresses missing fields, formatting errors, etc. Data hygiene is a critical first-step component of overall data quality; however, document emphasis concentrates on where industry scrutiny has been most intense: underlying quality and integration process. CIMM also recognizes that confidentiality and privacy should never be compromised, and that end users and data suppliers must be ever mindful of the importance of cultivating trusting relationships with the most important members of the marketing and advertising chain: consumers.

End-User Roundup

Data Quality Generally Considered Good – End users were generally positive about data quality from third-party firms and felt that data integrations enhanced marketing and advertising effectiveness. Despite this endorsement, end users would like to see improvements in data quality. Larger, historically process-rich, data firms were perceived as delivering higher quality data while no one firm was recognized as the leader in the field; data firms tend to specialize and excel in specific categories such as automotive or CPG, for example. *(pg. 13)*

Data Integration: Focus on Segmentation and Targeting – Data integrations come in many shapes and sizes but can be broadly categorized into consumer segmentation and targeting, CRM, media optimization and measurement of marketing and advertising sales impact. CIMM research found that end users were most interested in the data quality aspects of consumer segmentation and targeting, however, the focus on marketing and advertising performance measurement wasn't far behind. *(pg. 14)*

Consumer Modeled Segments and Authentication: Top Data Quality Issues – First and foremost, end users expressed concern about the accuracy of modeled consumer segments created by data partners for use in look-alike and/or behavioral targeting. End users sought to understand what variables were used to build modeled targets and how effective they were in driving consumer action; they want to ensure that modeled consumers are "in-target" as well as "in-market" to minimize reaching the wrong people at the wrong time. Consumer authentication was another area of concern; end users want to know, for example, that males 18-34 have been accurately classified and gender-validated. *(pg. 17)*

Transparency and Decision-Making – The call-out for greater data quality transparency resounded repeatedly during end-user conversations as they sought to better understand techniques and processes used for modeling target consumers, recency of data and quality and reliability of underlying data sources. Disclosure of this information would be helpful to inform the utility of the data for business decision making. Larger data suppliers were perceived to be more revealing of their techniques and processes while smaller firms were sometimes found to be vague or non-responsive, especially in the digital ad tech space. Some end users trust the third-party solutions and tolerate limited transparency while others enforce strict disclosure policies. As one advertiser put it, "If they're not forthcoming in answering our questions, we don't work with them." *(pg. 14)*

Compound Data Integrations Can Mean Compound Error – Data integrations may be layered upon data integrations to the point where original source identification is eclipsed. This situation can exist within digital ad tech targeting tools where the quality of cookies, for example, can vary according to credibility of source, recency,

expected life, original vs. modeled, etc. There may be no quality assurance procedures followed to ensure the underlying data vitality of these thickly-layered solutions. As one data intermediary stated: "The farther the distance from the original data source, the greater the chance of error." *(pg. 16)*

Data Intermediaries - Muscle in the Middle – Data intermediaries, firms that sit between data suppliers and end users, have emerged to fill the end-user resource gap for prepping data. These companies will implement quality assurance measures and build them into the integrated data products they market or provide custom work-for-hire where they will simply prepare and format data for delivery to end users. Visibility into multiple third-party data supplier offerings is their advantage point. *(pg. 17)*

Data Harmonization: Yes, But Quality First – End users support the idea of a common set of names for consumer target segments, to potentially simplify future data integrations, however, only after tackling data quality issues. One naming convention for "Men-25-54," "Males aged 25-54," and "M 25-54", for example, should be made possible. Need was also expressed for consistent target definition meanings which would likely work well for standard age and demographics descriptors but pose a challenge for targets like "auto intenders" and "travel planners" due to the fluid and customized nature of how these consumer segments are created by third-party firms. Lastly, virtually everyone interviewed indicated that their company would be directly involved in any industry effort for data naming standardization. *(pg. 18)*

Data Supplier Roundup

The Data Supplier Ecosystem – Third-party data suppliers in the advertising and marketing space can be described as a checkerboard of companies that all have similar traits but different, nuanced personalities. For example, some firms specialize in tracking and reporting CPG activity, others excel in department and retail store transactions while a few might offer digital behavior as a means for identifying in-market consumers. The older firms' ancestry tracks back to direct marketing with data often sourced from the credit card industry as well as direct mail; they became the CRM stalwarts of marketers. The digital companies grew their business from processing and matching consumers' online activity. Third-party processors of TV viewing data integrate consumer descriptors from other companies within the ecosystem. Net, net: end users can pick from among a wide diversity of suppliers depending on relevance and fit to their business needs. *(pg. 11)*

Consumer Authentication & Targeting – Top Customer Concerns – Data supplier firms indicated that the most frequently asked questions embrace the theme of consumer authentication and modeled targeting. End users want to know, for example, that males 18-34 have been accurately classified and gender-validated. Another prominent area of inquiry is about techniques used for modeled targeting, particularly the portion of actual vs. estimated data to create the targets. Recency of data was also on the most-asked list as it plays a core role in estimating which consumers are in market to purchase. *(pg. 17)*

Addressing Consumer Authentication Issues – Most data suppliers will run consistency checks across multiple consumer data sources to control for misidentification of misclassifications of consumers. For example, the data firms will compare the level of agreement of age, marital status, residential address, change of address, income, etc to determine the correct description and classification of the consumer. These cross checks can be made using gold-standard CRM data, if available to the data supplier. On the digital side, data can be checked for valid domain names, email addresses, search preferences and recency and frequency measures to ensure that electronic activity actually represents live consumers who have recently exhibited desired in-market behavior. *(pg. 18)*

Validating Modeled Target Segments – Most data suppliers cited the use of traditional "hold-out" testing to see how the modeled target will perform versus the benchmark of CRM/1st party or other actual sales data. If the

modeled target performance is subpar then the data supplier will likely test new combinations of variables to improve the model's effectiveness. *(pg. 20)*

Data Consistency Checks – Most data suppliers indicated that they have automated alert systems set up that trigger when unusual swings in key metrics occur across standard time period reporting. After vetting the anomaly to understand whether there was an actual change in consumer behavior versus potential data error, the data firm will adjust its estimates accordingly. *(pg. 21)*

Data Harmonization: Suppliers In Sync with End Users – Like end users, the majority of data suppliers supported the notion of a common set of names and definitions for demographic descriptors and consumer target segments. They also offered their support and willingness to engage with industry associations in a joint effort to address issues with the harmonization process. *(pg. 18)*

Conclusions & Next Steps

The growing number of consumer data sources and complexity of integrations across multiple consumer touch points poses a challenge for end users to assess data quality. The primary issue is gaining a better understanding of underlying quality of data to inform how best to deploy for advertising and marketing decisions. Following are recommended steps to be taken by the marketing and advertising industry *(pg. 23)*:

1. Establish Principles of Data Quality Disclosure
2. Create roster of items to ask of data suppliers
3. Request Transparency - data firm disclosures:
    a. Consumer authentication procedures
    b. Multiple data source descriptions
    c. Model target techniques and testing results
    d. Recency of data
    e. Integration techniques
    f. Data labeling; top line descriptions for all of the above
4. Create Best Practice Standards
    a. Acceptability guidelines for prepping and integrating multiple data sources
    b. Recruit third-party members to participate, shape guidelines

INTRODUCTION:
Data Integration Overview

In recent years the Marketing and Advertising industry has accelerated the integration of large data sets to provide richer insights for linking advertising and promotional efforts to consumer response. These integrations have reached critical mass application for digital media placement and are gaining increased traction in the TV advertising arena. The sheer volume and appearance of different types of data will likely increase in the future while the industry fast tracks Big Data capabilities. While "fast" is the operative word, there is concern among many research/analytic experts about the completeness, quality and transparency of these data for consumer segmentation and marketing/media investment decision-making. Furthermore, with the ever-expanding number of data sources, there's a need to examine the notion of data harmonization, or standardization of nomenclature to make the data integration process simpler for end-users.

Data integrations can come in many shapes and sizes but can be broadly categorized into consumer segmentation and targeting, CRM, media optimization and measurement of marketing and advertising sales impact. It appears that marketers are now attempting to apply the precision of direct marketing CRM segmentation and targeting at larger scale by connecting consumer data across myriad sources for more pinpointed messaging during the customer acquisition phase of advertising and marketing.

The Targeting Accuracy Conundrum

Users of third-party data have recently been very vocal about the quality of consumer target segments created by third-party data suppliers, contending that target accuracy can vary widely from vendor to vendor. In some cases, the additional cost of buying enhanced targets may not be commensurate with the level of improved targeting accuracy. According to Paul Rostowski, president of online ad buying company Varick Media Management: "If you're doubling your costs by adding in data, the performance has to be at least twice as good. Often the cost of the data won't justify itself." [1]

The concept of target accuracy refers to how closely a consumer segment definition reflects a marketer's actual media target. Improving accuracy means reducing waste which results in elevating marketing and advertising ROI. One of the big challenges marketers and third-party data firms face is developing a means to scale precise, verified purchase behavior profiles of existing customers to the set of consumers of competing brands and services. To that end, third-party firms create target models from CRM data that resemble existing customers. In addition to describing target consumers with greater accuracy, marketers often want to identify customers and prospects who are "in-market" to purchase.

Hypothetically, the ideal marketing and advertising targeting plan calls for communicating only with consumers that meet two criteria:

1. They are in-target
2. They are in-market to buy

The ability to execute an advertising plan against these conditions is challenging from two perspectives. The first is accurately defining the target and addressing the issue of timing the communication when consumers are ready to purchase. The second is the addressable level of the media required to execute against the two target criteria, especially the in-market component. The marketing and advertising industry has looked to electronically captured digital activity as a robust, immediate source of consumer intention that would indicate in-market buyer

[1] From *Marketers Question Quality of Ad-Targeting Data Providers,* by Jack Marshall, February 23, 2015, Wall Street Journal

readiness. So digital ad tech's ability to pinpoint and advertise to in-target, in-market consumers is presumed to be easier than, for example, TV where advanced addressable capability is just beginning to take root.

The preponderance of industry concern about targeting accuracy has been largely centered in the digital media space where third-party data enrichment is highly prevalent among publishers and ad tech firms. Kraft Foods Group, citing its digital experience, quantified the level of accuracy across multiple third-party data suppliers from in-house studies they've conducted. Julie Fleischer, Senior Director, data + content + media at Kraft contends, "With a few exceptions, intransigent publishers are shooting for profit, they are creating opaque systems that defy tracking and measurement, privileging themselves and their operations over their customers." [2] After vetting third-party data across four suppliers on accuracy of fundamental demographics and some behavioral characteristics, Kraft concluded the levels of target accuracy to be grossly underwhelming. One example for targeting owners of Keurig coffee brewers resulted in a target hit rate ranging between 14% and 20% across the data providers examined.

In an analysis independent of Kraft's investigation, media agency MediaSmith conducted a study that assessed targeting accuracy for pro-bono campaigns in the U.K. and U.S. during a three-week period in November, 2014. The analysis compared targeting accuracy for vendor-created segments versus an untargeted control campaign. Chart A shows an enormous chasm between the most- and least-accurate vendors: in five out of 14 cases the targeting accuracy was at least double that of the untargeted control group while improving only 5% to 15% in four instances.

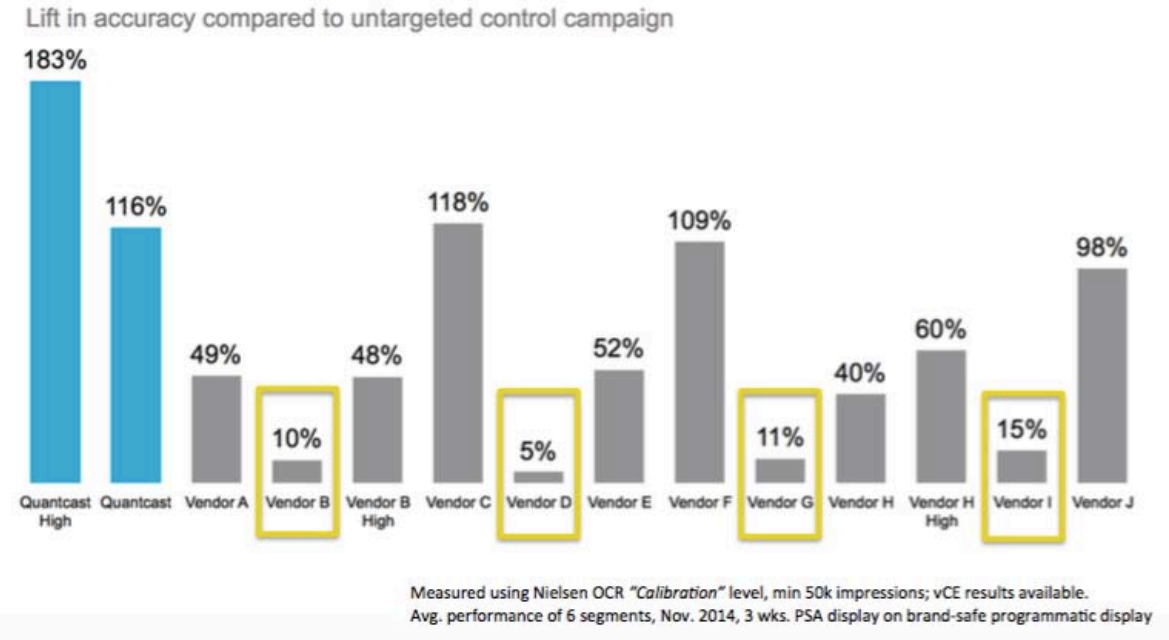

CHART A

2 From *Kraft tackles advertising's data integrity problem,* Steven Whiteside, October 2014, WARC

The Data Supplier Ecosystem and Process

Third-party data suppliers in the advertising and marketing space can be described as a checkerboard of companies that have many similar traits but different, nuanced personalities. At the base of nearly every offering is a description and/or behavioral profile of the consumer but seen through the lens of that particular company's expertise. For example, some firms specialize in tracking and reporting CPG activity, others excel in department and retail store transactions while a few might offer digital behavior as a means for identifying in-market consumers. The older firms' ancestry tracks back to direct marketing with data often sourced from the credit card industry as well as direct mail; they became the CRM stalwarts of marketers. The digital companies grew their business from processing and matching consumers' online activity; their business spawned from web publishers and they've branched out into the offline world. There are also firms that process TV viewing data, integrating it with descriptors from other companies within the ecosystem. Net, net: end users can pick from among a wide diversity of suppliers depending on relevance and fit to their business needs.

Many marketers put data enrichment into action by matching their customer files with consumer transactions and media exposure from external third-party sources. Chart B demonstrates a simplified view of this process which uses the power of an advertiser's CRM data base to identify key customers and their non-customer look-alikes to be targeted in the media universe. Many third-party firms work closely with advertisers in building and maintaining their customer data bases and often have permission to access customer files for authenticating consumer information across multiple sources. This process benefits the marketer and other customers who engage specific third-party firms. This same process depicted below is also used by media companies as a service to advertiser clients and a means to increase the value of their ad inventory.

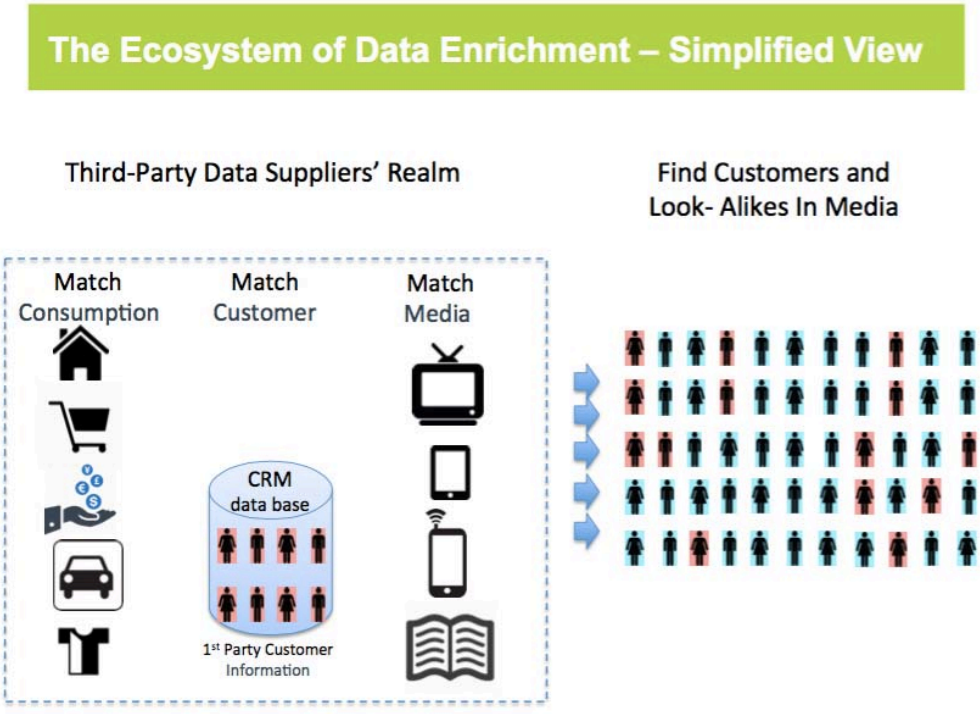

CHART B

Defining Data Quality

The notion of data quality can be viewed in three tiers:

1. Data preparation hygiene – cleansing processes to address missing fields, spelling conventions, typos, value and logic checks
2. Underlying quality – source credibility, recency, consumer classifications, collection method, representativeness
3. Integration process – quality of techniques/methods used to combine disparate data sets

Data preparation hygiene is a critical requisite for promoting overall data quality, since lack of scrubbing can indirectly throw off the underlying data quality as well as introduce error during the integration process. While this document focuses primarily on underlying data quality and integration process procedures that have received prominent industry attention, the importance of rigorous data preparation measures cannot be understated. Gartner reported that poor data quality is the primary reason for 40% of all business initiatives failing to achieve their targeted benefits.[3]

CIMM Study Details

Two modes of data gathering were deployed: 1) a literature search and 2) interview format discussions with Subject Matter Experts (SMEs) on the topic of third-party data quality:

Literature search

- Search criteria: Third-party Big Data and marketing, advertising, media placement, data quality
- Sources: Google queries and WARC literature search

As expected, the search yielded virtually no articles from marketing research trade journals that specifically addressed the issues of consumer authentication and target modeling. There were however, marketing and advertising industry-related articles that voiced concerns and provided quantitative evidence of inaccuracies in identifying target audiences for some third-party supplier modeled solutions.

SME discussions/recruitment – In order to provide a rich account of end user data quality concerns and corresponding feedback from data suppliers, CIMM conducted 22 interviews as follows:

- 10 end users – combination of senior research personnel and media placement experts at agencies, advertisers, media organizations, MVPDs and Data Intermediaries*.
- 12 data supplier companies – senior personnel in third-party data supplier firms broken out by
    - General purpose - Core consumer data sourced from credit transactions and census data; multi-category data offerings; direct marketing legacy; new entrants focus on digital behavior
    - TV audience - 3rd party processors of digital set top box data; match TV data with consumer descriptors from general purpose firms
    - Specialty firms - Focus on one or two categories of data; e.g., Hispanic consumers, movers, medical
    - Data connector – Firms specializing in providing connective infrastructure for multiple data sources, e.g., identifying consumers both online and offline

Chart C provides a schematic overview of the CIMM research

[3] From *Measuring the Business Value of Data Quality Gartner,* October 10, 2011.

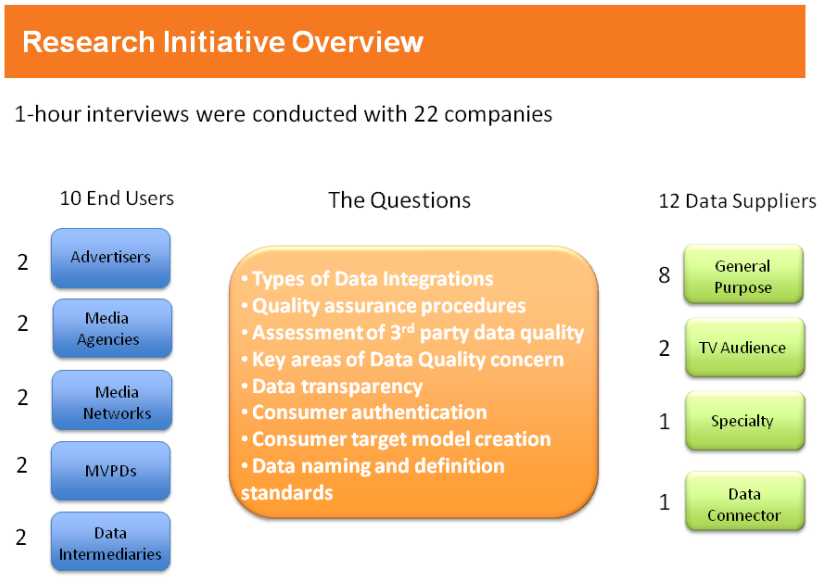

CHART C

END-USERS: FEEDBACK ON DATA QUALITY

This section re-caps key data quality perceptions that were gathered during personal interviews, including anonymous quotes from end-user companies.

Data Quality Generally Considered Good (with an asterisk)

End users were mostly positive about the overall quality of data provided by third-party firms, particularly the larger ones who have been perfecting their standards and processes for decades. There was recognition that some firms tend to excel in delivering data for specific industry sectors like automotive vs. insurance or CPG and that no one firm scores high grades across the entire gamut of marketplace data. Most importantly, end users felt that third-party data is making their marketing and advertising campaigns more effective; however, they would like to see some improvement in data quality.

> ***"'Will never be absolutely satisfied with data quality but feel our company has industry edge in performance through data integrations." –Advertiser***
>
> ***"Significant lift in effectiveness driven by 1st party data but third-party data generates better lift than standard targeting." -MVPD***
>
> ***"Big firms that originate their own third-party assets, they tend to do it very well, use same approach on clients 1st party data." -Data Intermediary***
>
> ***"All do a decent job but occasionally make back-end mistakes (calculations, wrong data) when, for example, matching viewer to transaction, lack of quality control." -Media Agency***

Consumer Segments: Top Data Quality Issue

First and foremost, end users expressed concern about accuracy of modeled consumer segments created by data partners for use in look-alike and/or behavioral targeting. They cited the practice of creating target audience profiles from transactional data sourced from a small base of consumers and extrapolated to the larger population. End users also sought to understand how effective modeled targets were in driving sales; they called into question sales-predictive behavioral indicators such as "visited an automobile site" or "read travel content" as markers for automotive or travel intenders, for example. The feeling was that these consumers may be targeted too late, after they've purchased the car or made their travel plans. Not surprisingly, data recency was cited as another data quality issue that ties back to the ability to identify in-market consumers, as well as capturing timely changes in consumer marketplace dynamics. Finally, it was acknowledged that smaller, low-incidence targets tend to have the highest rate of error when modeling due to the paucity of actual consumer behavior data available to model from.

> ***"We have used firm A for modeling TV show websites visitors and found little agreement in profile with syndicated sources." -TV Network***
>
> ***"Small modeled targets can be land mines with large error." -MVPD***
>
> ***"Recency of information is key, especially for behavioral, in-market aspects." -TV Network***
>
> ***"Want to identify in-market prospects, propensity modeling; some advertisers want to break out actual target vs. modeled." -Media Agency***

Transparency and Decision-Making

End users recognize the fact that there is no perfect data source; there will always be some amount of error inherent in any data set depending on the quality/reliability of source, data preparation procedures, survey response accuracy, etc. But disclosure of the extent of the error would go a long way to inform how the data should be used in making marketing and advertising investment decisions. To that end, the call-out for greater data quality transparency resounded repeatedly during end-user conversations as they sought to understand, for example, techniques used for modeling target consumers, recency of data and quality and reliability of underlying data sources. Insight into key-driver model variables as well as modeled ROI results versus actual consumer ROI were of keen interest (See Chart D).

> ***"Data not perfect but want to know how imperfect it is." -Media Agency***
>
> ***"Need to understand sources of error and how they're are mitigated." -Advertiser***
>
> ***"Integrations introduce error, for example, how wrong are they getting personal data?" -Media Agency***

End Users: Modeled Target Transparency

What % of total target are actual consumers vs. modeled?
What are key model variables?
What is the modeled-target ROI vs. the actual consumer target?

CHART D

To Be, or Not To Be Transparent

Inconsistent transparency was a sore spot for many end users when it came to answers received from third-party firms regarding data quality. Some data companies were found to be vague or non-responsive about the techniques and methods used to create propensity or look-alike models, a key area of concern. Another transparency issue that surfaced was the lack of information about original data sources, especially culled from digital media behavior metrics. In general, larger third-party companies were considered to be more revealing of their approaches to consumer target creation then their intermediate-sized counterparts. Some end users trust the third-party solutions and tolerate limited transparency while others enforce strict disclosure policies. As one advertiser put it, "If they're not forthcoming in answering our questions, we don't work with them."

"'Insufficient transparency, the more we dig the more inconsistencies and gaps are uncovered." - Advertiser

"Vague about modeling techniques and the sources their models are built on; can be defensive." – MVPD

"Big companies = lots of transparency. Smaller can be elusive. Companies within AdTech space are not very transparent." -Data Intermediary

"Firm A won't reveal retail partners. Firm B and Firm C won't break out known vs. modeled results." -Media Agency

Compound Data Integrations Can Mean Compound Error

When the mortgage crisis hit the U.S. during the last decade, it was discovered that consumer debt was sold and re-sold so often that it became nearly impossible to trace records back to the point of origination. Similar situations sometimes exist with data integrations, particularly within digital ad tech targeting tools, where integration may be layered upon integration to the point where original source identification is eclipsed. Furthermore there may be no quality assurance procedures followed to ensure that these thickly-layered solutions deliver superior results; the implication is that data error may be compounded with introduction of each additional data source. And the lack of transparency within and across these layers makes it nearly impossible for end users to assess the sources as well as the data error. It would be valuable to know, for example, which data elements are key drivers of developing a media target, how recent the data are, where they are sourced from and how accurate they are at capturing the true advertiser target.

> ***"The farther the distance from the original data source, the greater the chance of error; there's no opportunity truly to understand the underlying target elements and their impact on media decision making." -Data Intermediary***
>
> ***"I'm very suspicious when companies sell pre-segmented packages; data could come from any number of sources." -Media Agency***
>
> ***"There tends to be transparency for single, independent metric sources but not for multiple sources." -TV Network***

Each data source within an integration comes with an inherent amount of error that end users should know about in order to gauge impact on marketing and advertising decision making. In the digital space, for example, the practice of high-volume cookie pooling can pose a data quality mystery as end users seek to understand information about cookies' point of origination, active lifespan, recency, whether from a registered vs. non-registered site, original vs. look-alike, model selection criteria, etc. (See Chart E) There are currently no industry audits or standards to ensure the quality of cookies or at least facilitate disclosure about their level of business vitality.

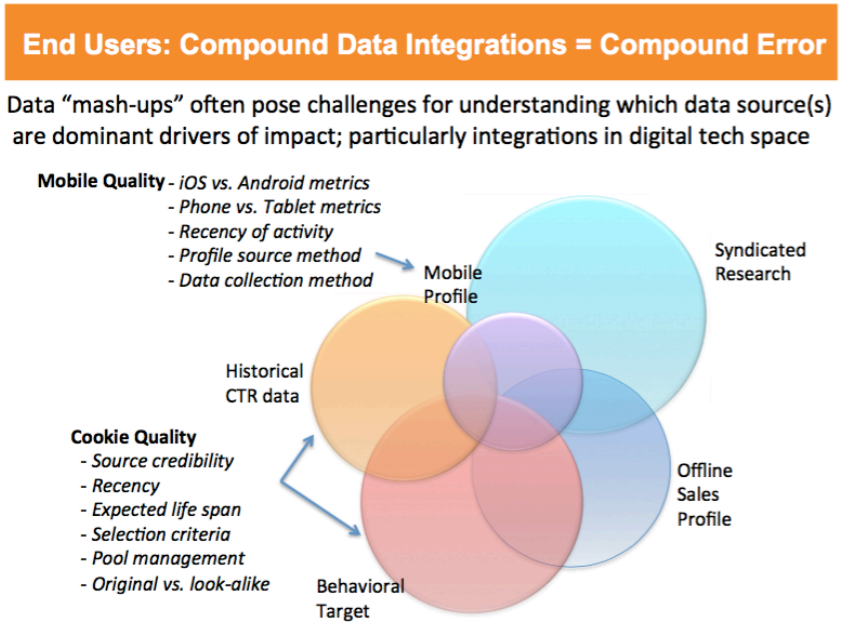

CHART E

Data Intermediaries - Muscle in the Middle

The task of ingesting, cleaning and formatting data across multiple sources can be daunting for many end users who do not possess the in-house staff or expertise to handle complex data integrations. Data intermediaries, firms that sit between data suppliers and end users, often fill the end-user resource gap for prepping data. These companies will implement quality assurance measures and build them into integrated data products they market to the industry or provide custom work-for-hire where they will simply prepare and format data for delivery to end users. Data intermediaries have a unique vantage point in that they have visibility into multiple third-party data supplier offerings and capabilities.

Data Harmonization: Yes, But Quality First – Bearing in mind the proliferation of data integration and onset of automation in the marketing and advertising industry, end users support the idea of a common set of names for consumer target segments. But many mentioned that the industry should first focus on data quality issues before standardization, citing the current concerns around consumer authentication, modeled targeting and transparency. Once these issues are addressed, they felt that consistent nomenclature would make life simpler by helping to mitigate differences across data coming from multiple third-party firms. For example, one naming convention for "Men-25-54," "Males aged 25-54," and "M 25-54" should be made possible. In addition to common names, it was felt there should be consistent target definition meanings which would likely work well for standard age and demographics descriptors but be a bit more challenging for targets like "auto intenders" and "travel planners" due to the fluid and customized nature of how these consumer segments are created by third-party firms. Lastly, virtually everyone interviewed indicated that their company would be directly involved in any industry effort for data naming standardization, citing the 4A's, ANA, IAB and ARF as industry associations to help steer the process.

DATA SUPPLIERS: FEEDBACK ON DATA QUALITY PROCEDURES

This section covers how data suppliers address the data quality issues posed by end users by describing quality assurance procedures they can practice.

Consumer Authentication & Targeting – Top Customer Concerns

Data suppliers were asked to recount the core questions posed by end users regarding data quality. Their responses were consistent with feedback gathered during the end user interviews: consumer authentication and modeled targeting were high up on the short list of data quality questions asked of third-party firm customers. According to data suppliers, end users want to know, for example, that males 18-34 have been accurately classified and gender-validated and about techniques used for modeled targeting, particularly the portion of actual vs. estimated data to create the targets. Recency of data was also on the most-asked list as it plays a core role in estimating which consumers are in market to purchase.

> ***"Data source, validation. Is it representative, how do you know its Men 18-34?" -Data Firm F***
>
> ***"What is underlying data used to build the modeled solutions and how are the models built?" -Data Firm B***
>
> ***"What goes into the modeled projections?" -Data Firm I***
>
> ***"How recent is the data? Where is it sourced from?" -Data Firm B***

Consumer Cross Checking to Address Authentication Issues

Consumer authentication is a core aspect of data quality. Misidentification or misclassifications of consumers at the very beginning of the integration process could have a rippling effect of error that compounds as each data source is combined with the next. For this reason, most data suppliers will run consistency checks to determine whether the same type of information about a consumer, for example age or marital status, is in agreement across multiple sources; they will also validate residential address by frequently checking various sources to determine if there has been a recent move. Cross checks will also be made against gold-standard CRM data, if available to the data supplier.

On the digital side, data will be checked for valid domain names, email addresses, search preferences and recency and frequency measures; the goal is to ensure that electronic activity actually represents live consumers who have recently exhibited the desired in-market behavior. Perhaps the most common bridge for identifying the same consumer in the online, mobile and offline world is a validated email address, preferably checked against 1st party CRM data and/or registrations from trusted web sites. A repository of credible email addresses enables data suppliers and data connectors to tie smart phones and tablets, for example, to HHs and consumers through a digital footprint like hashed email.

Chart F provides a simplified example of how a data supplier might validate descriptive information about consumers. In this case, the data firm cross checks Consumer A's age, gender, residential address and income against six sources of information. The firm found that three of their twenty client CRM databases contain information about Consumer A and these were supplemented with externally-sourced data from syndicated survey research, automotive registrations, digital publisher registrations and TV viewing data. While some data disparities surfaced regarding age, address and income, there was enough agreement across the sources to conclude that Consumer A is a 47-year-old male living at 125 Smith Street with a household income over $125,000.

Data suppliers: Authenticating Consumers

Many data suppliers use multiple data sources to confirm the correct information about consumers

CRM Client DBs | Survey Research | Residential Change of Address | Automotive Registrations | Digital Publisher Registrations | TV Viewing Data

Consumer A is classified as a 47 year old male residing at 125 Smith St. with a HH income of $125K+ . . .

Consumer A	CRM 2	CRM 3	CRM 6	Survey	Change of Address	Auto	Digital Pub
Age	47	47	46	47	NA	47	42
Address	123 Smith St.	123 Smith St.	123 Smith. St.	456 Jackson St	123 Smith St.	123 Smith St.	456 Jackson St.
Gender	Male	Male	Male	Male	Male	Male	Female
Income	$125K+	$125K+	$110K+	$125K+	NA	NA	$125K+

CHART F

General purpose third party data firms, those offering core consumer data across multiple-categories, sometimes conduct studies that benchmark the quality of their data versus that of competitors. Chart G below demonstrates the reported data accuracy of five companies across more than a dozen attributes, among them: household size, education, home market value, household Income, length of residence, gender and age. The range of accuracy varies from a low of 60% to a high of 78% after removing records with missing data (see Appendix for calculation detail). If accuracy rates were provided for each individual attribute, some scores would likely be much lower than the average range reported here. Typically household income and gender can present accuracy challenges; income because of its personal nature and gender due to much of the data being collected at the household level which must be attributed to household members.

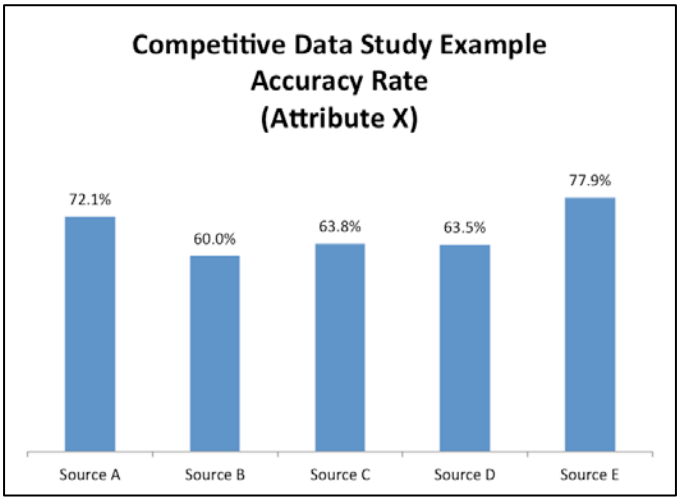

CHART G

Modeled Target Creation

Data suppliers were asked about the procedures used for developing segmentation/targeting models for deployment in look-alike targeting. Some data firms resisted this question citing the need to maintain confidentiality due to the proprietary nature of the modeling techniques. Those who disclosed details did so in a very generic way. In general, the most common approach was to identify the demographic and behavioral variables in their databases that exhibited the strongest relationship with the desired consumer activity. For example, a target description for those likely to take a European vacation during the next year might look something like this:

- Traveled to Europe in past 18 months
- Visited European travel sites in past 3 months
- Searched online for European travel destinations in past 3 months
- Read any travel magazine publications in past month
- Has frequent flyer membership with European airline
- Downloaded mobile travel apps
- Searched mobile apps for European travel destinations in past 3 months
- College education
- HHI $125,000 plus
- Attended graduate school

- Own non-domestic automobile
- Reside in North East, Pacific or Mid-Central census regions
- Speak a non-English European language

Data suppliers generally have access to thousands of behavioral and demographic variables that can be used to generate a target descriptor list that resembles the one above. An example of the process for combing the database for target descriptors is illustrated in Chart H below. The procedure begins with screening database variables with a sparsity* check to remove all zero values and a Weight of Evidence check to assign values to the variables. The next step is to identify the combinations of demographic and behavioral traits that hold the strongest relationships with the expected consumer behavior, European travel, for example. In most situations, the data supplier generally uncovers 25-50 variables that hold a strong relationship with European travel from an original set of 10,000.

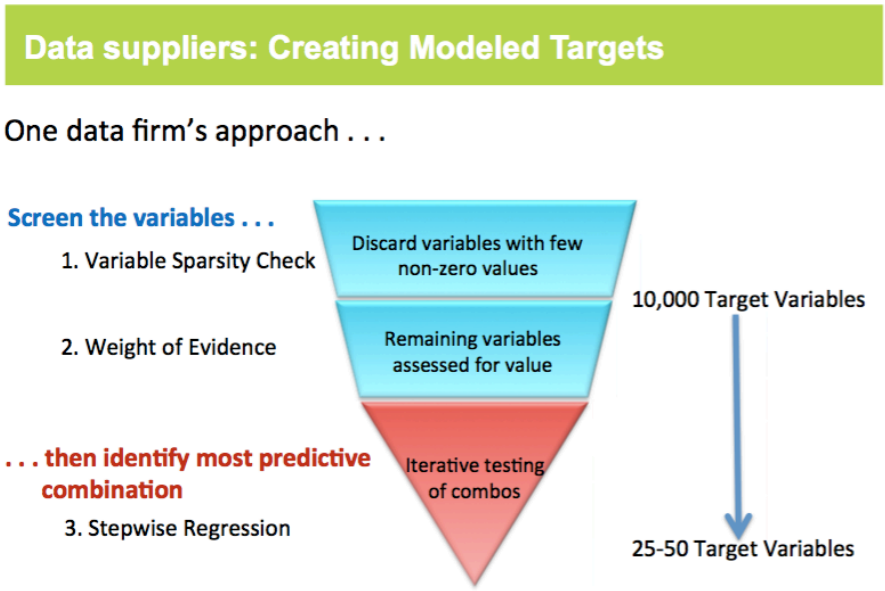

CHART H

Validating Modeled Target Segments

Drawing from multiple streams of consumer touch point data, data suppliers have the opportunity to develop propensity models that identify consumers who are likely to buy. Most data suppliers cited the use of traditional "holdout" testing to determine how well the modeled target will perform versus the benchmark of CRM/1st party or other actual sales data. If the modeled target does not meet the performance standards of the 1st party benchmarks then the data supplier is likely to test new combinations of variables to improve the model's effectiveness. It's important to note that end users would like to see more transparency of the holdout test results so they can be confident that the modeled targets are driving results.

Chart I demonstrates the use of holdout testing to determine modeled target strength. The holdout consumer group is a representative sub-sample of the modeled target group and receives no campaign messages. As the campaign runs, modeled target (exposed) product/service consumption is compared to that of the holdout (not exposed) group as a benchmark to assess the strength of the target model.

* Sparsity defined as the degree to which variables in a data set exhibit a scattered or weak relationship

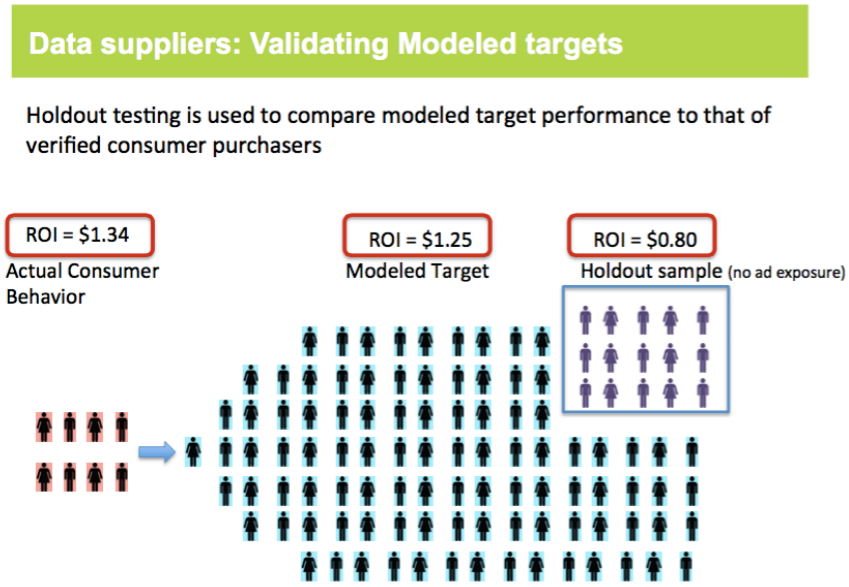

CHART I

Monitoring Data Fluctuations

The blending of multiple data sources requires keeping an eye out for unusual fluctuations from one reporting period to the next to ensure that integrated data set metrics are sound and stable. To that end, most data suppliers indicated that they have alert systems set up that trigger when unusual swings in key metrics occur across standard time period reporting. These alerts are made operational in quality assurance processing software and are activated when a metric falls outside the acceptable range of established normative benchmarks. The highlighted household income and small business owner composition metrics in Chart J below provide an example of how estimates might fall outside a normal range within a five-month period. Data suppliers stated that once a flag is triggered through their automated systems, a manual investigation takes place to determine the reason for the variation. After vetting the anomaly to understand whether there was an actual change in consumer description or behavior versus some type of error, the data firm will adjust its estimates accordingly.

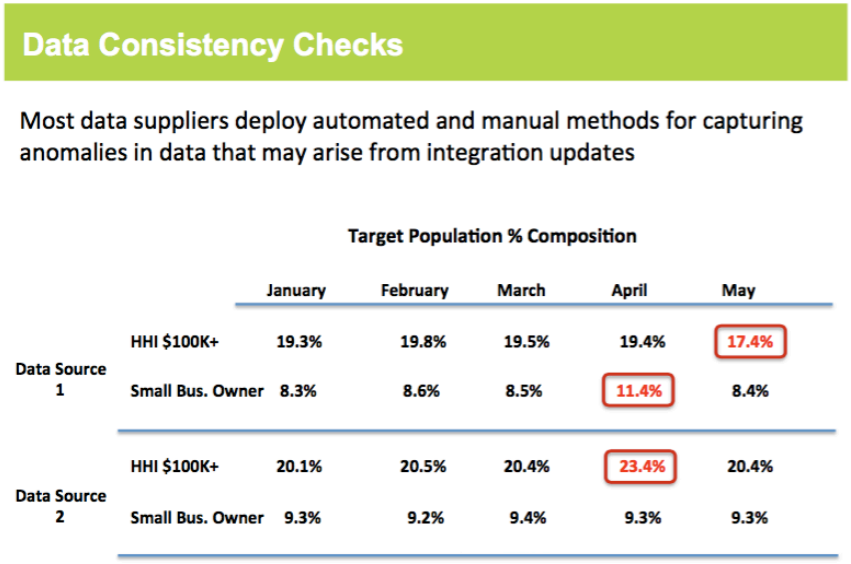

CHART J

Independent Testing of Data Supplier Solutions

In general, data suppliers indicated that they test their modeled targeting and segmentation solutions, using techniques such as holdout samples, to ensure that modeled target performance exceeds the norm. End users who want to assess performance differences across multiple data suppliers have the opportunity to test and compare the various solutions. One data intermediary interviewed for the study stated that media agencies are in a unique position to objectively test data supplier offerings: "Agencies can look across the providers with an impartial eye, have the best vantage point for evaluation." While media agencies hold a distinct position for evaluating data supplier offerings, they must also meet certain logistics and resource requirements before they can take advantage of their vantage point:

- **Analytic resources** – staff time must be available to design and steward the testing process and results evaluation
- **Number of data suppliers** – usually dictated by advertiser client preference and user history, likely to number no more than two data supplier firms
- **Synchronized campaign timing** – ideally, testing across multiple providers should occur at the same time to control for variations in media weight, media mix, seasonality of response, special events, etc.
- **Matched testing technique** – design and execution of test should be the same across all data supplier solutions
- **Multiple testing occasions** – tests should be repeated to ensure help understand changes in data supplier targeting or segmentation approach that may impact testing results

Chart K provides an example of the timing, technique matching and testing frequency across two data suppliers.

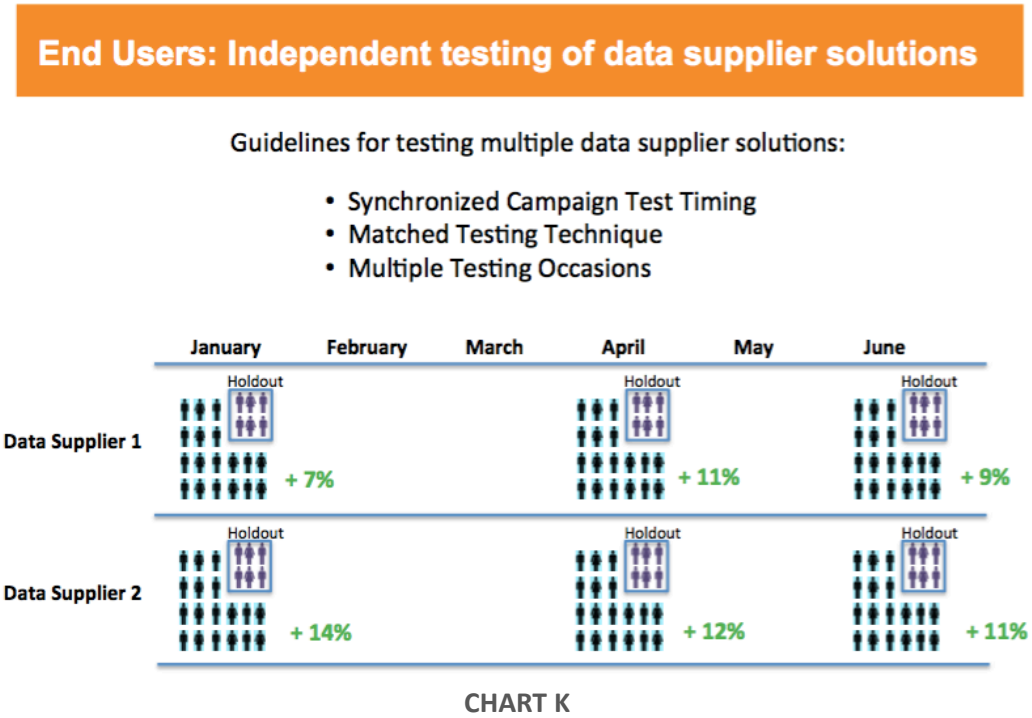

CHART K

CONCLUSIONS & NEXT STEPS

The growing number of consumer data sources and complexity of integrations across multiple consumer touch points poses a challenge for end users to evaluate data quality. The primary issue is gaining a better understanding of underlying quality of data to inform how best to deploy for advertising and marketing decisions. Following are recommended steps to be taken by the marketing and advertising industry:

1. Establish Principles of Data Quality Disclosure

Following are guidelines for gathering the information necessary to understand the quality level of data sets and data integrations. The purpose of these principles is not to prescribe any specific sources or approaches to data suppliers but rather to encourage richer transparency so that end users can make informed decisions on which data products and services meet their standards requirements. Tolerance of data precision and quality may vary across end users and within specific marketing and advertising questions they're looking to address, so clear transparency is needed to ascertain how close or far afield a data source or integration may be from fulfilling end-user needs.

Data Source

- Very first point of origination, including intermediary specialty firms sub-contracted to provide feeds into commercial-ready data products

Data Collection Technique

- Method of data gathering, for example, pixel tracking, scanning, sensor, internet survey, telephone survey and other monitoring

Data Integration Matching / Aggregation Method and Key(s)

- Matching methodology – e.g. direct household-/device-/individual match, geographic cluster aggregation (ZIP+4, PUMS), etc.
- Specific data fields used as matching keys to combine multiple datasets – e.g. name, physical address, email address, IP address, device MAC address, software device identifier, ad identifier, etc.

Initial Cleansing, Processing & Formatting

- Correction and/or removal of inaccuracies, misspellings, inconsistencies, contradictions, disparities, data entry mistakes, missing fields, etc.

Recency/Frequency of Data

- Freshness of data - e.g., purchases made during the last week, month, quarter, etc.
- Depth or activity – e.g., how often purchases were made during the last week, month, quarter, etc.

Consistency Over Time

- Monitoring for significant changes over time that might surface due to new data sources, collection techniques, etc.

Consumer Authentication

- "Truth" data source(s) to validate description of the current and correct household and/or household members (minus PII) in the data set
- Cross-checking systems that help establish the accurate description of homes and persons (minus PII) through multiple source validation; how they work, how often they are implemented and at what level of data granularity

Propensity or Look-alike Modeling Techniques

- Methodology description of modeling approach
- Underlying data sources
- Underlying metrics or variables

- Recency of data sources, metrics and variables
- Sample size composition of modeled purchaser target vs. actual purchaser target

Propensity or Look-alike Modeling Validation

- Methodology description of validation approach
- Effectiveness of modeled target purchaser vs. actual purchaser target
- Frequency of model validation

2. Request Transparency –

End users should ask for disclosure of a list of items from the Principles of Data Quality Disclosure roster that meets the needs of their specific marketing and advertising scenarios.

3. Create Best Practice Standards –

a. The advertising and marketing industry should draft acceptability guidelines for prepping and integrating multiple data sources.

b. Involvement of data suppliers is crucial for shaping guidelines

APPENDIX:

DATA SUPPLIER ACCURACY RATES:

Competitive Data Study Example

	Source A	Source B	Source C	Source D	Source E
Attribute X					
# Reported	149,417	155,806	156,792	131,591	106,893
# Null	7,383	994	8	25,209	49,907
Total Possible	156,800	156,800	156,800	156,800	156,800
# Correct	107,758	93,469	99,979	83,543	83,266
Accuracy Rate	72.10%	60.00%	63.80%	63.50%	77.90%
Balanced Data Score	68.70%	59.60%	63.80%	53.30%	53.10%

ACKNOWLEDGEMENTS

CIMM would like to thank all companies that participated in this very valuable initiative as the marketing and advertising industry strives to paint a more complete picture of target consumers and their behavior across multiple platforms. Below is a list of participating Data Supplier firms and End User companies who were interviewed for this study to which CIMM expresses its profound appreciation:

DATA SUPPLIERS:

LiveRamp
Nielsen Catalina SOLUTIONS
ACXIOM
TiVo Research
datalogix
bluekai
Experian
Ethnic TECHNOLOGIES
RENTRAK
MasterCard Advisors
neustar
IRi

END USERS:

THIRD-PARTY DATA PROVIDERS:

(Please note that the "Data Type" description is a high-level indication of the various suppliers' expertise. Please proceed to their website or call them directly to obtain a more complete profile of their offerings)

Data Provider	Data Type
ABIS	*CEO at Home Data*
Accudata	*Various Data Sources*
AccuTrend	*New Business Data*
AcquireWeb	*Email Data*
Action International	*International Data*
Acxiom	*Consumer Data, Propensity Data*
AGS (Applied Geographic Solutions)	*Consumer Expenditure and Retail*
ALC	*Specialty Data*
Amacai Business	*New Business Data*
Anchor	*Consumer Data, Phone Data*
Andrew Wharton	*Specialty Data*
Bizo	*Special Business Data*

Brilig	*Special Business Data*
Budco	*Mover Data*
Businessweek	*SMB Business Data*
Catalina	*Shopping Data*
Compass Marketing Solutions	*Compiled Business data*
Corelogic	*Mover Data*
Costal Lead Services	*Real Estate Data*
CoStar property Data	*Real Estate Data*
D&B	*Business Data*
DataPartners	*Specialty Data*
DataQuick Information System Inc.	*Property Data*
Digital List	*Apartment/Complex Rollup*
Dunnhumby	*Shopping Data*
Epsilon	*Consumer Data*
ESRI	*Geo Data*
E-Tech	*Ethnic Technology software*
Equifax	*Business Data*
Experian	*Consumer, Auto and Financial Data*
Firmgraphix	*Business Data*
FIS	*New Mover*
Forrester Data	*Research Data*
Fresh Address	*Email Data*
FYI	*TV Guide Data*
GeoResults	*Telco Data*
Geoscape	*Hispanic Targeting Data*
Harte Hanks	*Business IT Data*
HDS	*Medical Data*
Ibehavior	*CPG Modeled Data*
iLeads	*Mover Data*
Infogroup	*Consumer, Email, Business Data*

Information Refinery	*Specialty Data*
Infutor	*Directory Assistance Data, Consumer Data*
Iridium	*Specialty Business Data*
IXI	*Credit Trigger Data*
Jigsaw	*Business Contact Data*
Kantar Media	*Shopcom Data*
KBM	*Consumer Data*
Labels and Lists	*Political Data*
Latin-Pak	*Hispanic Targeting Data*
List Services Direct	*Specialty Data*
LiveRamp	*Online/Offline Integration*
LSSI	*Directory Assistance Data*
Mapping Analytics	*DMA Mapping Data*
Market Reason	*Specialty Data*
Marketforce	*New Business Data*
MasterCard Advisers Data Insights	*Credit Card Transaction Data Insights*
MCH	*Specialty Data*
Media 1	*Research Data*
MeritDirect	*Business Data*
MerryGold	*Business SMB Data*
Gfk/MRI	*Research Data*
NCC	*Zip-zone data (Cabletrak)*
Neilsen Claritas	*Consumer, research and Segmentation*
Neustar	*Consumer and Business Data Assets*
Polk	*Auto Data*
Rapleaf	*Consumer and Email Data*
ReachForce	*Business Contact Data*
Relevate	*Specialty Data*
Rentrak	*Research Data*
Scarborough	*Research Data*

Scorelogic	*Propensity Data*
SMA Data	*Political Data*
SNL Data	*Geo Data*
Speedeon	*Various Data Sources*
Transunion	*Various Providers*
Stirista	*Propensity Data*
Target Data	*Premover*
TargusInfo	*Specialty Data*
TNS	*Business and Research Data*
Tower Data	*Email Data*
Transunion	*Credit Trigger Data*
TruEffect	*First Party Cookie Data*
Valassis Direct Mail	*Property Data*
Zoom Info	*Business Contract Data*