Building 21st Century Multi-Platform Audience Measurement Systems: Integrating Multiple Data Sets into a Single Currency

Josh Chasin, Chief Research Officer, comScore
Cameron Meierhoefer, Chief Operating Officer, comScore
Frank Pecjak, Senior Vice President, Statistical Services, comScore
Dr. Michael J. Vinson, Senior Vice President, Operation Management, comScore

Executive Summary

In the media audience measurement field, practitioners are familiar with the problem of limitations on individual data sets, and thus the need to integrate multiple, disparate data sets into a single dataset for manipulation and projection. Fusion is one popular technique that accomplishes this, establishing a host (or perhaps more accurately, "recipient") dataset and a donor dataset, and then appending individual respondent records from donor to host based on matching variables present in both data sets.

However, as media consumption continues to fragment across platforms, and big data sets (e.g. digital census data, Return Path Data from TV Set Top Boxes) become increasingly essential to creating audience measurement solutions, media researchers must consider new techniques to combine multiple data sets into a single planning, buying, selling, and audience analysis tool. The era of the single "magical panel" that can measure all relevant media exposure is gone; we must forge systems that amass best-in-class data assets, that are intelligent about duplication ("the first casualty of fusion"), and, that combine disparate data sets (human panels, census or cookie panels, big data assets, and universe estimates) into a single user-facing dataset—one that also reflects corrections for the biases inherent in individual donor datasets.

In the US market, and triggered by the integration of comScore and Rentrak in early 2016, comScore is developing a new, patent-pending projection system that combines the concepts of Agent-Based Modeling and fusion, to create massive, unified/integrated "synthetic respondent level data sets" (perhaps at a 1:10 ratio of all households in the market.) The methodology involves creating a pool of "Representative Household Units" (RHUs), demographically and behaviorally balanced to represent the market. These RHUs become the recipient dataset into which disparate donor datasets are ported. Person rosters and device rosters are created for each RHU, again representative of the market. Then real data from different datasets may be incorporated at the RHU level. For example, STB data at the household level is assigned to each RHU—actual tuning from a real market household is assigned to a demographically, behaviorally matched RHU. Then tuning data within the household may be assigned to each household member on the RHU's household roster.

The result is a massive respondent-level dataset that projects back to universe, matches individual currency measures from component data sets, and—to the extent that overlap datasets are incorporated—does a better job than traditional fusion of preserving real duplications across platforms and vehicles.

This system has been prototyped in all local US TV markets, and is targeted to go live as beta data in the national US TV marketplace by the end of 2017. The roadmap is to begin reporting TV data (first at the household level), then to layer in digital data. However, it is worth noting for the PDRF audience that the opportunity exists to integrate print datasets into such a system—both from readership studies, and, perhaps more exciting, subscriber lists. Doing so would result in a massive single source data set of TV tuning, digital usage across devices, and print readership and subscription status across tens of millions of synthetic panelists.

Introduction

The consumer media landscape grows increasingly complex. Where once consumer exposure to advertising was generally limited to broadcast TV, radio, newspapers, magazines, and outdoor/place-based media (billboards, etc.), today consumers may choose among these media, as well as cable TV, satellite TV, and Over the Top (OTT) TV; satellite radio and streaming audio; websites, the mobile web, apps, gaming consoles, digital video, digital editions of print publications, digital facsimile editions, and more.

Historically, the audience measurement practitioner was at core a sampling statistician, developing representative single-medium panels, fielding surveys, or placing written instruments, in order to collect media exposure data from a representative and projectable sample of the universe. But increasingly, the traditional audience measurement toolkit is proving insufficient for embracing today's audience measurement challenges. And even in cases where the deployment of a traditional panel or sample may remain appropriate, response and cooperation rates continue to decline, eroding the benefits of so-called random design.

Conversely, while new (mostly digital) technologies are rendering the media landscape largely unmeasurable using traditional toolkits, so are digital technologies emerging to provide practitioners with a new toolkit (Chasin, Harris; 2006). These technologies include census and big data "naturalistic" data sets (i.e. data naturally occurring in the distribution and consumption of digital media, like digital tagged census data or Set Top Box data). They also include platforms for warehousing, accessing, and integrating data sets.

As a result, the new audience measurement practitioner paradigm is more aptly one of the data scientist—building, accessing, and integrating various best-in-class datasets in order to create media audience measurement systems. While panels remain vital components in the science of audience measurement, they are no longer sufficient on their own; panels are now an input into complex systems that incorporate, at minimum, panel and census data together. (Meierhoefer, Pellegrini, San Francisco, 2011; Chasin, Nice, 2013). Increasingly, across both syndicated and JIC measurement solutions, the state of the art in audience measurement is moving rapidly toward a multi-sourced approach.

The question that arises, then, is how best to integrate these disparate sources, while correcting for biases inherent in each?

Background

A confluence of media industry developments is driving a multi-sourced approach to building comprehensive audience measurement solutions. These include:

- ***Granularity and fragmentation of media audiences:*** Whereas a print measurement service might report on 200-300 titles, and a TV ratings service might report on a couple of hundred networks, digital measurement services might report each month on tens of thousands of individual entities—and are still criticized for missing the long tail. Fragmentation is occurring by medium; by channel within medium (e.g. the subscriber to HBO now gets 7 different channels); by platforms (iOS, Android), devices (phones, tablets and gaming consoles); delivery systems (cable, broadcast, web, OTT) and players (measurement SDKs must be associated with each content player.)

- ***Demand for cross-platform data***: Magazine and newspaper publishers increasingly need to understand their "total footprint" audience, across physical and digital platforms. TV companies increasingly distribute content digitally, and planners and buyers of advertising on these media must understand the distribution of impressions across all modes of distribution to fully understand audience behavior. Whereas once a linear TV program audience was amassed entirely via traditional TV, today that audience might also be comprised of viewers who saw episodes streamed at the network website, via YouTube, via Hulu, On Demand, or from the Network's Smart TV app. Each of these components must be robustly measured in order to assure full audience credit. Increasingly, these digitally distributed audience components are best measured via census tags and measurement SDKs generating census counts.

- ***The advent of digital census data and TV Return Path Data:*** Instrumentation of websites and apps with tags and SDKs enables third party measurement providers to generate census counts of activity at publisher entities. These counts are imperfect as audience measurement in and of themselves; measurement providers must account for fraud and non-human traffic, viewability, cross-device duplication, co-viewing, and demography. However, census data are invaluable in sizing audiences, and especially granular segments thereof. Similarly, Set Top Box data can provide unprecedented scale, which introduces robustness, granularity of reporting, and stability into TV measurement; but like digital census data, STB data is a similarly imperfect measurement asset on its own.

- ***No single data set suffices for creating holistic granular multi-platform currency-grade audience data***: As noted above, there is no "magical panel" solution anymore that can meet the needs of the cross-platform audience measurement marketplace. It is unfeasible to endeavor to build a single panel that provides currency-grade measurement at scale across media platforms—especially when so much "big data" is readily available. Once we acknowledge that the age of the magical panel is past, we are inevitably left with a need to develop multi-sourced currency solutions.

- ***Data sets to be integrated include very large big data assets as well as panels:*** Census and census-like datasets provide granularity and stability; person-based panel assets provide an understanding of person-level behavior. It is difficult to preserve both sets of benefits with traditional data integration techniques.

The Challenge

The comScore/Rentrak integration created an opportunity to develop a next-generation system of integrating and reporting on TV and digital audiences, at both the local market and national level, in the US. The new comScore had access to numerous panels, digital census data from thousands of publishers, and Set Top Box data from about 22 million (soon to be 35 million) households.

Our challenge was to create a "single-source," persons-level dataset that combines universe estimates, panel data from multiple panels[1], digital census data, and Set Top Box Return Path Data, and that provides household-level and person-level TV viewing, digital usage, demography, and ultimately additional advanced target data; and that does so in a fashion that preserves single-currency results while incorporating an intelligent parsing of duplication, and that plays back to universe estimates.

The resulting dataset should support manipulation by comScore client-facing systems, third party systems, or proprietary systems (via API); and should support fusions and other standard industry database integrations.

[1] comScore's access to the Nielsen Audio PPM panel is restricted to use in the creation of services combining TV and digital data. comScore makes no use whatsoever of the Nielsen PPM panel in any standalone TV service. To date, the system described herein incorporates TV data only, and makes no use of Nielsen Audio PPM data.

This would be a daunting challenge for the traditional audience measurement practitioner. But the data scientists rolled up their sleeves and got to work.

Integrating Media Datasets

There have been several techniques deployed in the media audience measurement space to combine disparate datasets. Notably, these include:

- ***Meter/diary integration:*** A construct that is particularly relevant to comScore, meter/diary integration has been, and continues to be, used by Television Audience Measurement (TAM) services that use set-based TV meters to create household ratings and tuning levels, and diaries in the same market (but in different households) to create demographic VPVHs (Viewers Per Viewing Household.) The VPVHs are applied to the set meter household ratings to generate persons demographic ratings.

- ***Hybrid digital measurement***: Hybrid digital measurement refers to the integration of panel-side and server-side measurement. Panel measurement provides a greater understanding of the behavior of persons, including demography and cross-entity visitation; server-side, or census measurement, provides an empirical count of all activity at a site or app, but without an accounting of who the persons are behind the activity. An integration of the two enables the integration of the census of activity with the deeper persons insights derived from panels. In comScore's implementation of hybrid measurement, called Unified Digital Measurement, panels are used to provide demography, and to account for cookie deletion and other phenomena that lead to census Unique counts deviating from actual Unique visitors.

- ***Indexing:*** Indexing is a function often performed by third party software for manipulating audience measurement data. The audience size from a currency service serves as the base, and the composition of the audience (typically on advanced target criteria) is derived from a different study. So, for example, if a currency ratings service attributes a raw audience size of a million, and a separate study reports that 35% of the audience owns a luxury vehicle, then the composition of the second study is applied to the audience size of the first, and an inference is made that 350,000 of the vehicle's audience own a luxury vehicle.

- ***Fusion:*** Fusion is a well-known technique, somewhat more complex than those referenced above, that involves designating one pool of respondents as the "host" dataset and a second pool from a different study as the "donor" dataset. Each respondent in the host dataset receives a set of data from a specific respondent in the donor dataset, based on the closest match on a set of variables present in both datasets. For example, TV viewing data may be appended to respondents to a print/readership survey via fusion, based on a common set of variables found in both the TV ratings respondent pool and the readership study respondent pool (behavioral characteristics may become "hook" variables as well; for example, the readership study may contain some TV viewing data that can be used along with demography to provide closest match.) Constraints may be placed upon the fusion to assure that resulting projections from the fused dataset match the projections from each individual dataset, within some tolerance.

- ***Fusion into a hub survey:*** Typically, fusions are between two datasets. In the UK, TouchPoints has pioneered a fusion system wherein a central survey is conducted, designed to serve as a "hub" to which multiple currency datasets are then fused (Wilcox & O'Sullivan, Vienna, 2007).

This hub survey contains extensive variables designed to facilitate multiple fusions to the hub study. The resulting TouchPoints dataset serves as a market-wide, cross-media planning tool.

The Concept

comScore is developing a system based on Representative Housing Units (RHUs), Representative Person Units (RPUs), and Representative Device Units (RDUs.) A set of RHUs is created—currently at a ratio of 1 per 10 households in the market—and assigned demographics and behavioral characteristics reflective of the market at large. Then each RHU is populated with representative RPUs and RDUs, with media exposure assigned to RPUs across the RDUs.

The core data structures within the system are based on the principles of *Agent Based Modeling* (ABM). In this process, individual units (agents) of one or many types are created within a system. These units can be given multiple attributes and “rules” on how they interact with other units or within the system itself. The advantage of this approach is that a comprehensive model or set of assumptions about a large complex ecosystem is not required. Rather, the system is relatively straightforward to set up, assuming the overall population and system can be defined. Propensities and attributes about individual units are generally more easily obtained than probabilities for larger system interactions and through multiple simulations or iterations of observing interactions, behavioral summaries can be observed that include variance or margin of error to assess the certainty of events.

In general, ABM systems have three phases:

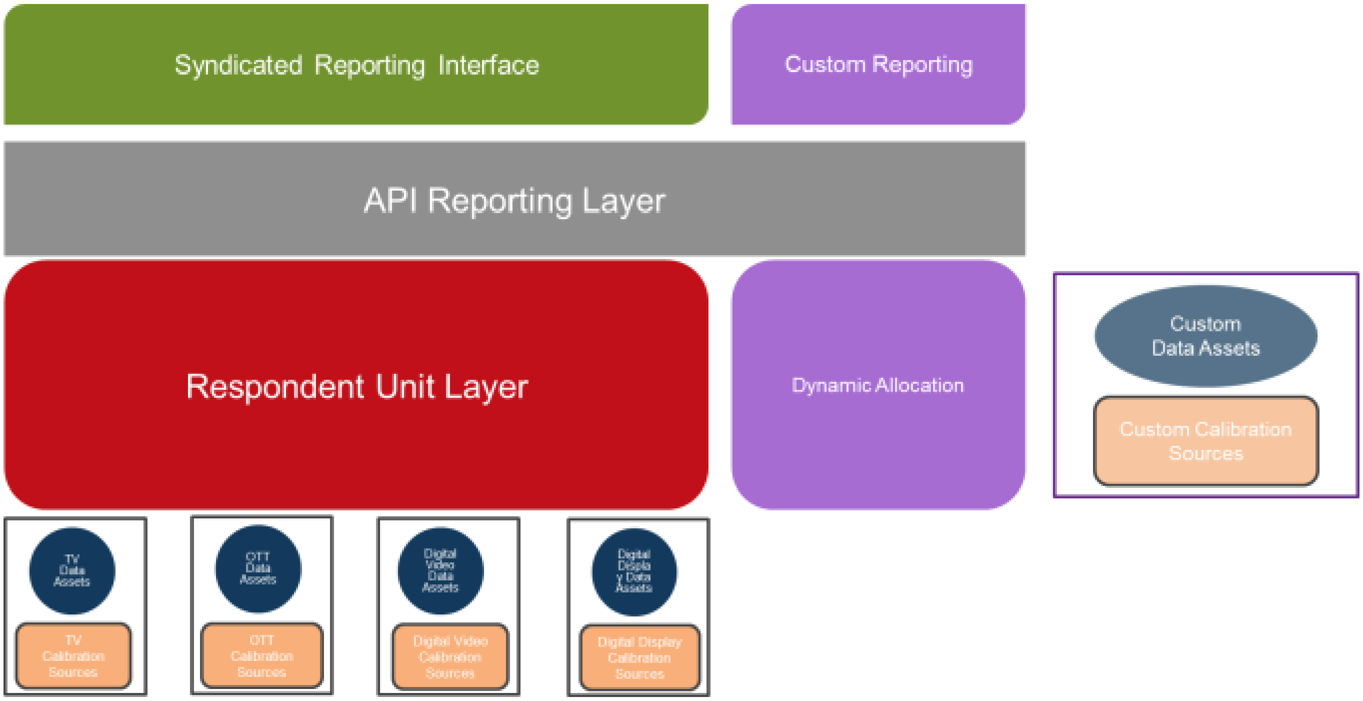
1. The **Initialization Phase** where the system and individual agents are created with attributes and behavioral models and propensities for interactions.

2. The **Simulation Phase** where the agents interact within the ecosystem in a step-wise manner. In many ABM systems, these steps represent the passage of time but the steps can also represent iterations that are evaluated until a convergence criteria are met. The latter case is what the RHU system will use.

3. A **Reporting Phase** where the resulting units, attributes and interactions can be queried and reported.

The diagram below shows an overview of the RHU system and specific points of encapsulation. Within the diagram, the key phases of ABM are represented by the core data ingest processes (**Initialization**, in black boxes), the respondent unit layer and dynamic allocation process (**Simulation**) and the API and reporting interfaces (**Reporting**). The separation of these phases into different processes and systems allows for distributing research and development, flexibility of each platform and pillar in configuring processes based on unique goals and data requirements yet still maintaining a level of standardization and alignment for ease and consistency in reporting and QA.

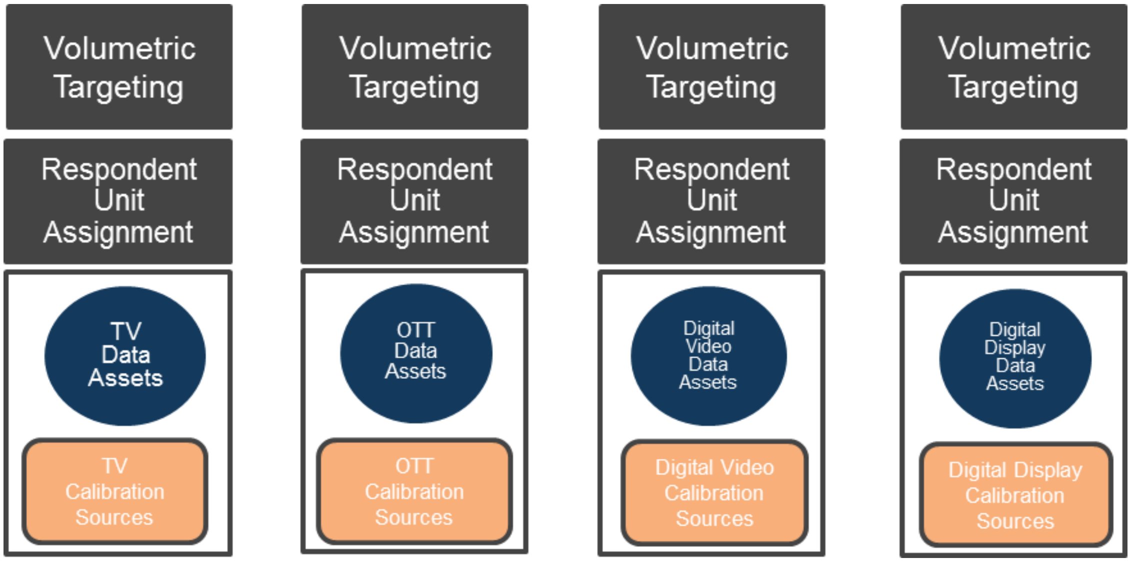
Figure 1


Syndicated Reporting Interface
Custom Reporting
API Reporting Layer
Respondent Unit Layer
Dynamic Allocation
Custom Data Assets
Custom Calibration Sources
TV Data Assets
TV Calibration Sources
OTT Data Assets
OTT Calibration Sources
Digital Video Data Assets
Digital Video Calibration Sources
Digital Displa y Data Assets
Digital Display Calibration Sources

Core Data Ingest

The Core Data Ingest processes provide the framework for the initialization of the household, person and device respondent units. Each data platform provides specific information on the sub-population size, attributes specific to the platform (e.g. TV strata (Cable, DBS, Over the Air) or digital access platform (PC, Smartphone, Tablet)) and behavioral profile such that the projected audience reach and consumption volumetrics (Page Views, Time Spent etc.) reflect accepted targets and expectations for the platform.

Figure 2


Volumetric Targeting
Volumetric Targeting
Volumetric Targeting
Volumetric Targeting
Respondent Unit Assignment
Respondent Unit Assignment
Respondent Unit Assignment
Respondent Unit Assignment
TV Data Assets
TV Calibration Sources
OTT Data Assets
OTT Calibration Sources
Digital Video Data Assets
Digital Video Calibration Sources
Digital Display Data Assets
Digital Display Calibration Sources

A primary assumption, and differentiator for comScore in making this type of approach work, is that the initialization phase as described above can be achieved as an allocation of census or near census behavior (i.e. based on “Big Data” assets as opposed to panels or surveys) rather than something that needs to be simulated with propensities. The individual platform (or pillar) teams utilize domain and data expertise to allocate observed events through the development of business rules. With this allocation, context is applied to the events giving a representative view of platform behavior and demographics that can be reported directly as-is in appropriate platform-specific settings.

In separating the input streams by platforms, the burden of coordinating definitions across platforms-- and having these complications potentially affect the ability to achieve specific constraints or assumptions-- is relieved. Also, research and development can proceed independently for each platform (media data source) at cadences that are appropriate to each use case. Finally, QA on individual platform data can be contained in the core ingest layer, assuring behavior and assignment for a platform are aligned to platform-specific expectations.

Allocation in each platform can be broken into two phases:

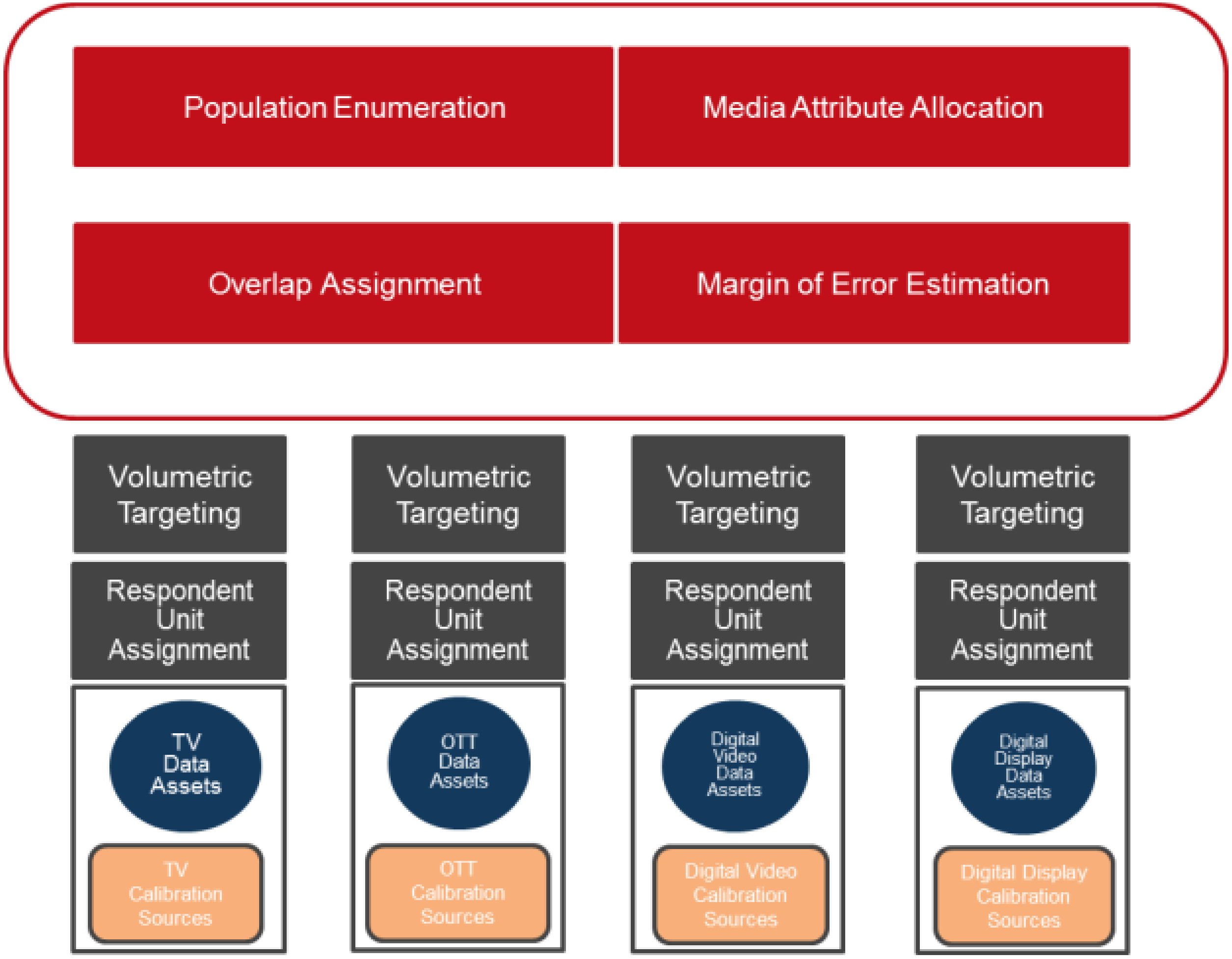
- **Assignment:** Where whole units (HH, Person, Device) are associated with the appropriate representative unit with little to no inference. This requires some subset of the raw data have demographic attributes, so that they may be matched into the representative units. The attributes of match for allocation would include core demographics as well as platform specific attributes that are key to describing behavior (e.g. TV strata and DVR status or digital device(PC/Mobile)). Some degree of inference and ascription is required during the assignment phase, in cases where either (1) there are not enough exact matches for the units available; or (2) the raw unit assigned is not behaviorally complete (i.e. the donor household doesn’t have as many set top box devices as have been designated for the RHU; or if DVR records are absent). In this case, the match is constrained to as many demographics and attributes as available and the best remaining raw record is assigned based on closest distance behaviorally. Assignments made to a particular unit will be maintained if the raw input is still available (i.e. not attritted in the data set). This maintains longitudinal consistency in behavioral profiles. The goal of the assignment process is to achieve most of the target unique visitation and additive totals directly without additional allocations or adjustments (the target is 85% of market-network-day hours for TV).

- **Adaptive:** A daily process which compares the projected assigned reach and additives to the individual platform targets (typically available as either census totals or a combination of census and enumeration) and assigns individual events into the appropriate units to hit the targets. This assures that the projected results in the system match individual-currency reported results. The events added are actual events from the pool of previously unassigned activity. The specific rules and targets for triggering the adaptive process will be determined by the individual platforms but generally an incremental event will be added to a unit that shows a high propensity for the type of event and has a gap in activity that can accept the event. The result of the adaptive process is an individual respondent level profile (that is empirically valid) and the aggregation of those profiles achieves the core platform targets (e.g. market-network-day hours for TV). This process is run daily and does not guarantee longitudinal consistency of assignment of events across units or raw households (that is, a representative unit is not guaranteed to get an adaptive event every day, nor will it get an event from the same source household/person/device every day).

Respondent Unit Layer and Dynamic Allocation

The Respondent Unit Layer is the final data service layer supplying all data to downstream API and Reporting systems. It is responsible for integrating data from the platform-specific core ingest systems into a single set of profiles. The integrations take place such that the platform specific totals and populations allocated through the core ingest process stay intact and are combined to represent the overall population and overlaps. There are four main components in the Respondent Unit Layer.

Figure 3


Population Enumeration
Media Attribute Allocation
Overlap Assignment
Margin of Error Estimation
Volumetric Targeting
Volumetric Targeting
Volumetric Targeting
Volumetric Targeting
Respondent Unit Assignment
Respondent Unit Assignment
Respondent Unit Assignment
Respondent Unit Assignment
TV Data Assets
OTT Data Assets
Digital Video Data Assets
Digital Display Data Assets
TV Calibration Sources
OTT Calibration Sources
Digital Video Calibration Sources
Digital Display Calibration Sources

Population Enumeration

The population cells that make up the individual unit types (Household, Person, Device) are defined using the total population estimates available from Census data, establishment surveys, and other universe estimate sources. The key definition of the cell, weight are selected to match granularity and reporting requirements from the API, independent of the platform attributes. This results in cells defined by demography, RHU roster composition, and geographic location. The attributes used in the population enumeration process are considered fixed and key values for identifying a cell.

Media Attribute Selection

Population composition and universe size for individual media platforms are enumerated by assigning appropriate attributes, such as TV strata or Internet access status, to units sourced from population cells. The combination of the demography from the population enumeration step and the mix of platform specific attributes assigned the representative household/person/device constitutes the key for that unit and the core values and constraints that will be explicitly controlled for in the overall assignment processes. The targets for these attributes come from the individual platform enumeration targets and assignment may shift over time to account for changing compositions (e.g. as the PC audience shifts more towards tablets, or households in the market shift from cable or DBS to OTT and streaming for consumption of TV program content).

Overlap Assignment

The process by which platforms are combined within each respondent unit to build a complete behavioral profile. The targets for the overlaps come from the calibration panels and enumeration sources available to measure cross-platform. Generally, these targets are not at the event level granularity or are not as detailed as the platform specific attribute targets so they cannot be assigned as deterministically. The assignment of the overlaps will occur as a stochastic process (probability based assignment with context derived from the targets) where the propensities are constrained by the cell demographics attributes and the individual platform attributes available for units within the cells. To maintain longitudinal consistency of these assignments, the overlaps will remain in place for a period of time (to be determined but likely no less than weekly).

It is worth noting though that comScore's measurement philosophy is to amass best-in-class single-medium datasets, and to amass (or build) best-in-class duplication data sets. Duplication across platforms and media is a vital component in the development of campaign reach. comScore endeavors to incorporate empirical measures of media and platform duplication, as opposed to letting duplication patterns emerge organically from the integration of individual-medium datasets.

Margin of Error Assignment

Because overlap is assigned in a probabilistic manner, there will be variance in the assignment commensurate with the relative uncertainty of the relationships. The value of utilizing an agent based approach is these uncertainties can be simulated multiple times to show the uncertainties and report margins of error for events. The respondent unit layer will simulate assignment processes multiple times and report the average outcomes for overlaps. The confidence intervals based on these simulations will also be reportable to account for the likelihood of specific behavior overlapping across platforms. These margins will be available for reporting.

Dynamic Allocation

The details above outline the attributes and behaviors that serve as explicit constraints on the population represented by the respondent unit layer. Controlling for these attributes ensure platform and cross-platform consistency with targets and behavior over time. However, there are needs for additional attributes and behaviors to be applied into the units to augment reporting. Examples include advanced audience segments (media targets beyond traditional demography) or ecommerce behavior. The dynamic allocation process manages applying the attributes as overlays onto the units without directly impacting the core reporting attributes and behaviors.

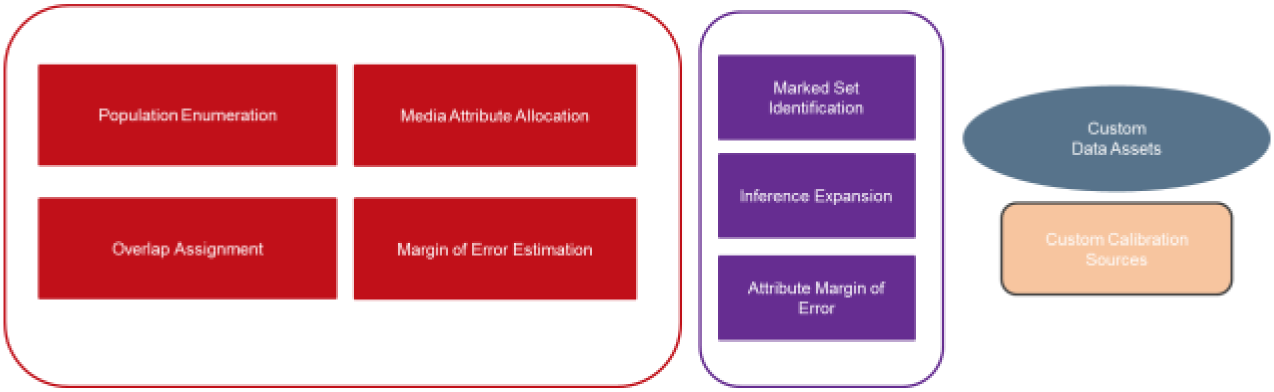
In general, data that is to be overlaid will come from a secondary data asset and calibration or enumeration source. With this set, for each individual overlay, a three-step process is established:

- A marked set is identified by establishing individual units that can be deterministically assigned the attributed or behavior. This is typically done by joining through STB household identifiers or individual digital identifiers (cookie, device ID, IP address).

- Developing propensity for the specific attribute or behavior based on the marked set that can be constrained by the unit keys and utilizing any enumeration or calibration data to set population targets. Once established, the attributes can be expanded throughout the respondent unit layer in a process similar to the overlap evaluation step described above.

- A secondary margin of error calculation that measures the variance of assigning a particular attribute or behavior is calculated in the same way as the margin of error calculation form above and will also be surfaced to the reporting and API layers.

Figure 4


Population Enumeration
Media Attribute Allocation
Overlap Assignment
Margin of Error Estimation
Marked Set Identification
Inference Expansion
Attribute Margin of Error
Custom Data Assets
Custom Calibration Sources

In Conclusion

By deploying a database integration technique derived from Agent Based Modeling—a new construct to audience measurement, but, the authors believe, a necessary and inevitable evolution—comScore is solving the problem of how to integrate both panels and Big Data assets, across multiple media measurement systems, into a single, massive, unified dataset for end-user manipulation. This new end-user dataset preserves the single-currency values at a granular level; preserves the duplication patterns across platforms derived from duplication datasets; and represents the universe without the requirement of sample-balancing. The resulting dataset may be used in comScore's and third-party systems, supports fusion with other media datasets, and is well-suited (due to size and granularity) to fuel emerging platforms and exchanges for buying media inventory at the impression level, and based on either traditional demography or advanced targets.