Best Practices
in Matching Databases to
Set Top Box Data

cimm
coalition for innovative
media measurement

Presentation by:

Myles Megdal

MCG Consulting Group

July 2011

Study Overview

- Define and describe the range of databases that can be matched with STB data to support advanced and addressable TV advertising applications

- Understand the pros and cons associated with each database

- Define best practices in matching databases and assuring consumer privacy

- Address issues and concerns voiced by STB data processors and end users

Study Approach

- Confidential interviews with executives from companies that are either users or suppliers of STB data products and services

- Questions tailored by type of company and job responsibility

- All responses aggregated and anonymized

- Interview responses supplemented with quantitative data from various sources

Company Types

- **_STB End Users_** – agencies, advertisers, and media networks that are end users of STB products and services

- **_STB Service Providers_** – research/measurement companies and addressable technology partners that are directly involved in creating STB-driven addressable and advanced advertising products and services

- **_Database and Matching Service Providers_** – companies that supply data and provide database matching services in support of STB-driven products and services

Interviewed 20 Companies and 37 Executives

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

Questions Asked

Questions varied somewhat by type of company – especially with regard to emphasis:

- *<u>End users</u>* **–** questions focused on the extent of use of advanced and addressable services and key issues encountered with third party databases

- *<u>Service providers</u>* **–** questions focused on how third party databases are matched to STB data, and the processes and procedures used to assure data accuracy and projectability

- *<u>Database and matching suppliers</u>* **–** questions focused on technical issues concerning database accuracy, matching techniques, and consumer privacy

Questions Asked

- Types and extent of advanced and addressable advertising product and service usage
- Key issues encountered with either applications or specific types of databases
- What benefits have been observed from using advanced and addressable products and services?
- What benefits are derived from the third party databases?
- How is consumer privacy assured?
- What enhancements would be welcome?

Key Findings

- *There are a number of large consumer databases that, when matched to STB data, provide insights and tonality to all advanced and addressable applications*

- *In general, these demographic, lifestyle, and purchase behavior databases provide accurate and consistent information on consumer households*

- *Third party data suppliers need to do a better job of making their data more transparent and understandable to users in the television industry*

Database Applications

- More precise audience definition
- Support of both household and geographic addressable programs
- Enhanced TV measurement at the individual or household levels (broadcast networks to local markets)
- Creation of representative samples from STB data
- Enhanced precision for campaign planning and buying

Database Issues Noted

- Lack of transparency in data sources and data definitions

- Understanding the degree of data accuracy consistency – explicit vs. modeled data

- Understanding and accommodating inherent database biases

- Cost – building and using aggregated databases is expensive

- Lack of standard approaches to segmenting and using the aggregated databases

Database Evaluation Criteria

Five criteria were used to evaluate databases that can be combined with STB data

- _**Coverage**_ **-** **#** of individuals or households
- _**Breadth of data**_ **-** **#** of unique characteristics
- _**Depth of data**_ **—** coverage for a specific characteristic
- _**Accuracy**_ **—** the degree to which characteristics agree with verifying sources
- _**Latency**_ — the age of the data or update cycle frequency

Databases Studied

- **_Compiled databases_** – maintain diverse data with broad population coverage

- **_Transaction databases (shopper, pharma)_** – maintain detailed transaction (purchase) data specific to product categories and retailers

- **_Geographic databases_** – maintain diverse data at the small area geographic level – especially valuable for utilizing restricted data (credit, automotive)

All STB service providers interviewed use all of the above data as part of their STB product and service offerings

Global Data Realities

- *There is no 100% accurate database – all databases have some degree of inaccuracy*

- *Consistency of data is critical and is more important than overall accuracy – Data that is consistent is useful – Data that is inconsistent is difficult to work with*

- *Source of data is critical – data derived from transactional or purchase data is typically more accurate and consistent than data derived from surveys*

Compiled Databases

Four primary categories of data:

- ***Demographic*** – descriptive attributes of individuals and households
- ***Psychographic*** – self-reported opinions, attitudes, and intentions typically acquired from surveys
- ***Behavioral*** – preferences and purchases based on transactional information
- ***Geographic*** – characteristics – demographics, credit, and automotive – associated with consumer neighborhoods

Compiled Databases Pros and Cons

Pros

- ***Broad coverage*** – over 90% of the population represented

- ***Excellent breadth of data*** – generally, over 2,000 characteristics available

- ***Excellent – depth of data*** – average coverage of each characteristic is 60%

Compiled Databases Pros and Cons

Cons

- *<u>Demographic biases</u>* - for some population segments – affluents, seniors, children, lower income households

- *<u>Age of data</u>* – some components of compiled databases are updated infrequently (annually)

- *<u>Data variability</u>* – some data is explicit while other data is modeled – sources for specific characteristics can differ significantly

Shopper Databases

The key data captured and maintained consists of:

- 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

Shopper Databases Pros and Cons

Pros

- ***Broad household coverage*** – 50MM+ households represented
- ***Highly accurate purchase data*** – data provided on virtually all CPG, OTC, and prescription pharma products
- ***Frequent updating*** – results in data that is relatively current

Shopper Databases Pros and Cons

Cons

- *__Limited breadth of data__* – data is typically only related to product level purchases

- *__Limited depth of data__* – not all households are represented for all product categories – total purchases only reflect those done at a specific retailer

- *__Demographic/geographic bias__* – not all population segments or geographies are represented due to the nature of the retailer's audience and store locations

Geo-Aggregated Databases

- *Census databases* – provide consumer demographics and ethnicity at the small area geography level

- *Aggregated credit databases* – provide detailed consumer credit information (loan types, balances, and creditworthiness) at the small area geography level

- *Summary auto databases* – provide detailed vehicle ownership information (make, vehicle class, bought new, old, or leased) at the small area geography level

Geo-Aggregated Databases Pros and Cons

Pros

- **_Broad geographic coverage_** – national coverage provided

- **_Permits the use of restricted data_** – provides the ability to use restricted credit and automotive data for financial and automotive advertising

- **_High degree of accuracy_** – data is derived from transactional data – credit data from issuers, vehicle registration and financing data

# Geo-Aggregated Databases Pros and Cons

Cons

- ***<u>Data not linked to specific individuals</u>*** – only small area geography (zip+4)

- ***<u>Limited granularity</u>*** – credit products do not indicate type of credit card; automotive data does not indicate vehicle model
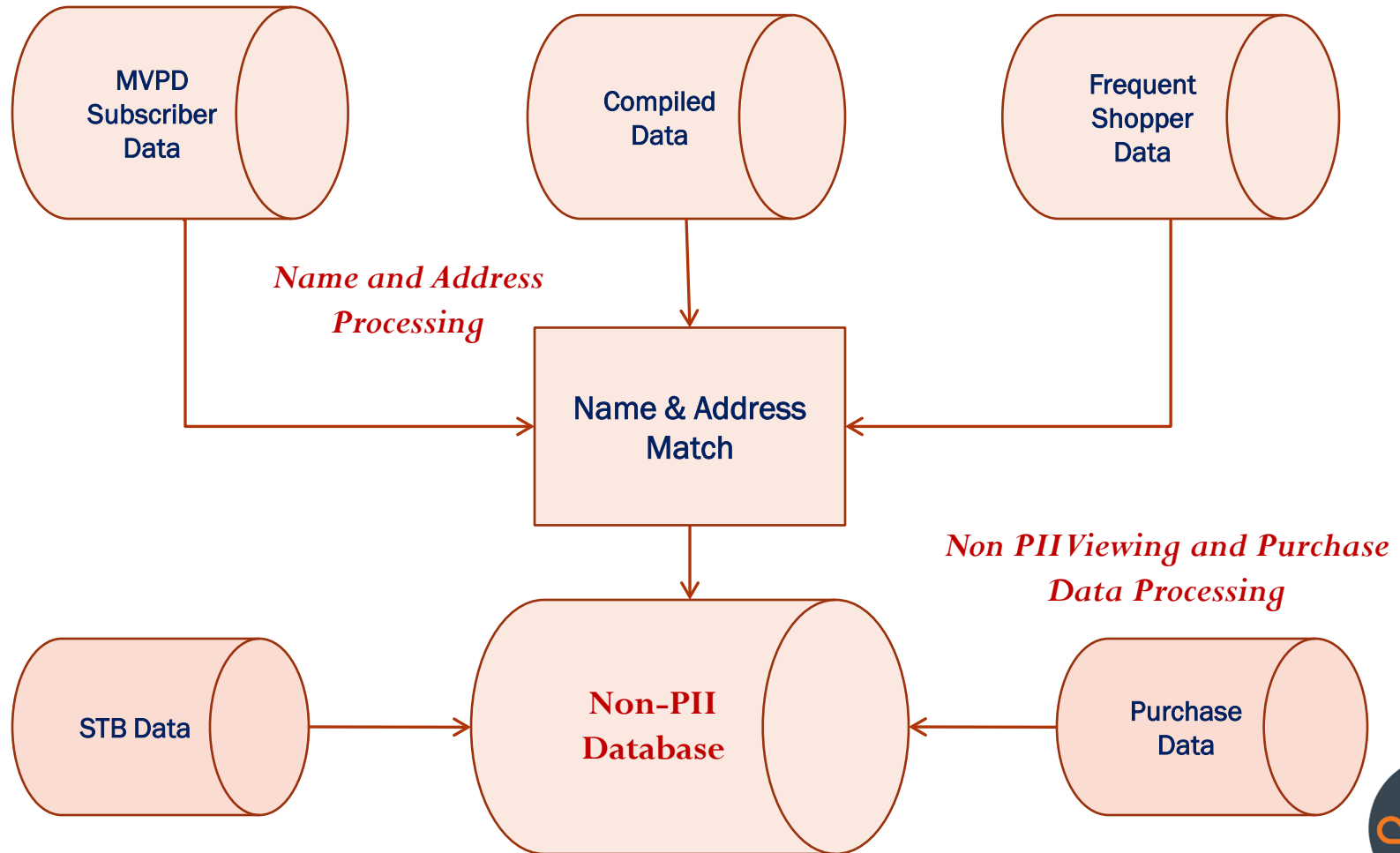
Database Summary

	Compiled	Shopper	Geo-aggregated
Coverage	120MM+ HHs 230MM+ Ind.	50MM+ HHs 15,000 Retailers	29MM geo units
Breadth of Data	2,000+ discrete characteristics	Limited to product and some demos	Limited to demos, credit, vehicles
Depth of Data	Avg. of 60% of data is represented	Variable by category and brand	100% of the data is represented
Accuracy	Highly consistent to Highly Accurate	Highly Accurate	Highly Accurate
Latency (Update Cycle)	Monthly-Quarterly	Monthly	Monthly

Database Matching
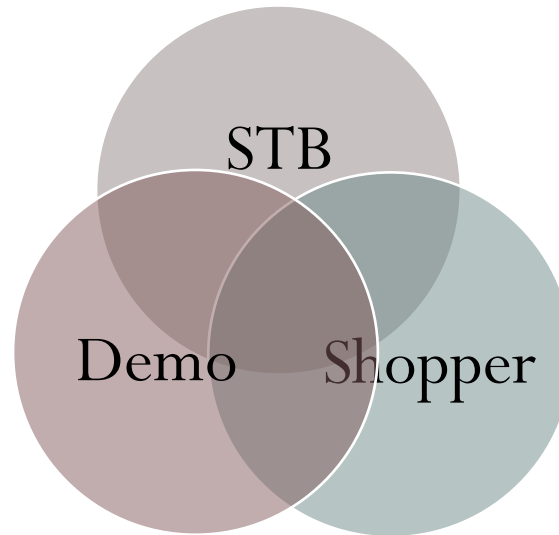
- All name and address matching is performed at the HH level
- Most STB service providers use only the intersect between the three sets of data to support their products and services
- All matching is performed so that viewing and purchasing behavior *is never associated* with an individual or household

Database Matching Typical Blind Match Process

MVPD Subscriber Data

Compiled Data

Frequent Shopper Data

Name and Address Processing

Name & Address Match

Non PII Viewing and Purchase Data Processing

STB Data

Non-PII Database

Purchase Data

Assuring Database Accuracy and Projectability

- Large size of databases makes balancing and other adjustments easier

- Geographic and demographic balancing use US Census data

- Shopper databases balanced versus reference source: IRI/Homescan, Proprietary Research, Advertiser data

- Only verifiable explicit data is used to develop models to "fill in" missing data

Interviewee Wish List

- More data transparency
- Standard variable definitions
- Lower costs for advanced and addressable
- Simplified logistics – traffic, reporting, distribution
- More sophisticated tools for accessing and using the aggregated databases
- More addressable inventory – nationally, broader program selection

"A single system that provides accurate audience measurement for in and out of home, in terms of ROI, for every product and service"

Observations

- There is a far greater base of available accurate data than is currently being used for advanced and addressable use

- More process automation is needed to reduce costs associated with implementing and maintaining these large databases

- Database suppliers and television advertising users need to better understand each others' terminology and needs

- Business models for advanced and addressable need to be defined (refined)

- The use of the same databases that are used in other media (direct mail, email, online, and mobile) will facilitate consistent audience definitions and media planning integration

Questions?