nielsen

AN UNCOMMON SENSE
OF THE CONSUMER™

NIELSEN RETAIL MEASUREMENT DATA FOR SYSTEMS INTEGRATORS

INTRODUCTION

This document provides IT professionals with an introduction to the unique features of Nielsen's retail measurement data. Although intended for a technical audience it does not require an IT background, and can be read by anyone with an interest in the topic.

NIELSEN'S RETAIL MEASUREMENT DATA

Nielsen offers three services which measure the retail sales of fast moving consumer goods:

- Retail Measurement Services (RMS) – a measure of sales based on purchases from stores RMS data is usually sourced from retailers' electronic point of sales systems.
- Homescan/Panel – a measure of sales based on purchases by consumers.
- Retail Direct Data – a retailer's own sales data, reported to their own definitions of products, categories and so on, supplied by Nielsen.

Nielsen supplements sales data with in store observations, for example of product promotions.

RMS data gives excellent breadth and depth of coverage, and allows sales to be compared across different store types, geographic regions, and retailers.

Panel data is based on purchases made by individuals and links sales to demographics in a way that anonymous electronic point of sale data cannot.

Both RMS and Panel services measure sales in a defined population, known as a *universe*. In the case of RMS the universe is a population of stores, and in the case of Panel it is a population of shoppers. Because universes can be huge and expensive to measure (millions of people and hundreds of thousands of stores), RMS and Panel services are often based on representative samples of the universe being measured. Sample data is statistically expanded to represent the universe.

There are many more shoppers than stores, and it is hard to recruit and maintain a representative sample of shoppers. It is also not yet possible to fully automate the collection of shopper data. For these reasons RMS, especially when based on electronic point of sales data, has a greater breadth and depth than Panel data - which is necessarily based on a proportionately smaller sample. Nevertheless Nielsen ensures that Panel sales are always reported at a level of aggregation which is accurate and representative.

Retail direct data is the retailer's own electronic point of sale data. Its unique selling point is that it shows the retailer's view of the world at store level. It does not provide a view of the universe, only the retailer's own stores.

Combining direct data across retailers is difficult because of the unique-to-retailer definitions used to render each data set, added to which direct data is not available from all retailers. It is much easier to obtain a market read from Nielsen than it is to patch disparate direct data sets together to make one.

RMS, Panel and Retail Direct services have complementary strengths and weaknesses, and between them are able to support the full range of market research use cases. Retail measurement data is also a key part of many systems supporting business functions outside marketing, such as demand management.

DATA OWNERSHIP AND LICENSING

Nielsen does not sell its data to clients; it licences data for their use. The licence establishes guidelines for use.

Some data that Nielsen licenses, for example Panel data, is collected and curated by Nielsen, in which case the licensing agreement is simple.

Nielsen obtains a significant amount of data from third parties, most usually from retailers. In such cases the data owner usually makes stipulations about how it may be used, such as the mandatory masking of sensitive data (for example the sales of own label goods).

If a client wishes to use data in a warehouse, an appropriate agreement is required. Agreement is also required from third party data owners, usually retailers, for the use of the data in advanced analytics.

Nielsen expects that any capabilities a client creates on top of Nielsen licensed data are for their own, exclusive use, and will not be shared or resold to others.

THE DATA NIELSEN COLLECTS

ELECTRONIC POINT OF SALE DATA

The bulk of Nielsen's RMS data is sourced from retailers' electronic point of sale systems.

Electronic point of sale transactions are aggregated to a common level of granularity (usually store, product and week) so that they can be combined across retailers.

Retailers often provide Nielsen with data from all of their stores, so called census data. As long as census data is present and correct, simple summation can be used to calculate sales figures for any aggregation of products, shops and periods *within the census data set.*

Some retailers only provide Nielsen with data from a representative sample of shops. Nielsen statistically expands this data to provide a projection of the sales in all of the retailer's stores.

Retail direct data usually consists of a daily feed of transactions which identifies the items bought together as a basket of goods, and any discounts, for example coupons, used. This data can be aggregated to the standard granularity (usually store, product and week), and used as a replacement for the retailer's standard electronic point of sale feed.

Some retailers also make available anonymised information about purchasers derived from their loyalty card systems.

Transaction logs and loyalty data contain information not present in standard RMS, and are therefore able to support additional types of analyses.

AUDIT SALES DATA

Some retailers do not have electronic point of sale systems, in which case Nielsen will impute their sales by taking the difference between the amount of stock on hand in a store at two different points in time (allowing for new deliveries in the interim) – the difference being the amount of product sold. This is a clerically intensive task that it is only cost effective to perform for a sample of stores. Audit data is a sample of the universe and must be statistically expanded before it can be reported.

STORE OBSERVATION DATA

In addition to sales transaction data, Nielsen collects observational data about product promotions:

- Special displays of a product in store which give it prominence over its competitors.
- In store offers such as price promotions.
- Advertisements for products in flyers distributed by the retailer (sometimes called features).

Store observations are clerically intensive, and are only carried out on a sample of stores for cost reasons. As before, sample observations are statistically expanded to represent the universe.

PANEL DATA

A Panel is a sample of shoppers recruited to be representative of a universe. Shopper purchases are recorded by a variety of means, from collecting used packaging, to having the shopper barcode scan their purchases, to performing optical character recognition on till receipts.

The end result of Panel data capture is an inventory of purchases by store, week and demographic. As with sample RMS data, Panel data has to be statistically expanded to represent the universe.

The great strength of Panel data is that it tells the analyst who bought a product (or rather their demographic profile: no personally identifiable data is released). Its weakness is that a Panel cannot match the breadth and depth of coverage of electronic point of sale data. The large, but limited, number of data points in Panel data is the reason why sales for slower selling products, or products with poor or localised distribution, sometimes need to be aggregated to a level at which they are representative before they can be reported.

Panel data combines both consumer insights and a total market read. RMS and loyalty data have more breadth and depth than Panel, but Panel's combination of both views in a single product is unique.

Panel also provides some unique measures of shopper activity:

- Penetration – the percentage of shoppers that purchased a product in a time period.
- Purchase size – how much of a product a shopper bought in one transaction.
- Purchase frequency – how often a shopper bought a product in given period of time.
- Buying rate – the average amount of a product purchased by a shopper in a given period of time.

DIMENSIONALITY

As would be expected of data constructed for business intelligence and analytic purposes, Nielsen's data is inherently dimensional.

Retail measurement data is usually provided as a set of data marts, each mart containing the data for a particular product category, such as carbonated beverages or hard surface cleaners.

Because retail measurement data is provided as a set of category marts, a mart may represent several implicit entities which are the same for all data points. Implicit entities may include the source of the data (RMS, Panel), and the category itself, but also country (most marts contain data from a single country), and others. Invariant, implicit dimensions only become important if there is a need to combine marts.

The dimensions invariably found in Nielsen data marts are:

- Market – a segmentation of the stores in which products were sold. Segmentations can be based on a plethora of criteria including geographic location, turnover, and type of store. A mart may contain more than one market segmentation.

- Product – the items sold.

- Period – the sampling points on the time axis. The frequency of measurement is usually weekly, but monthly, four weekly and four-four-five weeks are also common. It is possible to have a mart with a mixed periodicity (for example weekly and monthly data) but this is not common: many tools and systems assume a consistent periodicity.

Additional dimensions may be found in some data sets:

- Promotions – some Nielsen data models split out price reductions, in store displays and other promotions into a separate dimension.

- Demographics – usually only seen in Panel data sets.

For reasons of simplicity and usability, Nielsen tends to standardise and minimise the dimensionality of its data by collapsing dimensions, or by representing dimensions in facts such as "sales with promotion". Combining dimensions is a well-recognised solution to the **"curse of dimensionality"** and data sparsity.

As explained later in this document, Nielsen also provides dimensional aggregates, most especially in the product dimension. These aggregates are usually arranged in a logical hierarchy, such as category > manufacturer > brand > product.

FACTS

The handful of facts collected by Nielsen (sales volume, sales value, price, promotions) is used to generate a huge number of derived measures relevant to market research. Some Nielsen RMS products are capable of reporting several thousand facts. This section describes the main kinds of facts Nielsen provides and the fundamental differences between them.

ADDITIVE FACTS

Additive facts are absolute measures such as sales volume and sales value. They can be summed to create a meaningful measure. For example the sales volume of product X can be added to the sales volume of product Y to get an accurate measure of the number of units of both items sold.

The pitfall with additive facts is the potential for double counting, especially in the presence of aggregates. Adding the sales of product X to the sales of brand Z, when product X is part of brand Z, results in an inaccurate number because the sales of product X are counted twice.

NON-ADDITIVE FACTS

Non-additive facts are relative measures, such as percentages, proportions, and averages.

The majority of the facts in a Nielsen data mart are non-additive, and these facts are often the most useful ones for market research purposes.

The base data used to calculate simple non-additive measures, such as percentages, is usually present in the data mart, and it should be possible to calculate an accurate percentage for any aggregate not already present.

Most non-additive facts are complex and cannot be derived from the data in the mart. A good example is *% ACV distribution*. *% ACV distribution* is calculated as the turnover of the stores in which a product actually sold as a percentage of the total turnover of the stores in which it could have sold. Calculating *% ACV* for an aggregate not already present in the mart requires access to the lowest level data (store, product, week), which is not available in the mart, as well as statistical expansion.

In some cases it is possible to derive useful metrics by summing ostensibly non-additive facts. One such a measure is Total Distribution Points (TDP), which is the sum of the % ACV distribution of a set of products. TDP is a measure of how widely a set of products is available, and how many products are in the set, weighted by store turnover. TDP is a useful measure of the performance of a range of products, and can help identify which feature of a product (packaging, flavor etc) is the most important determinant of sales.

Range Distribution is another key non-additive measure. Range distribution measures the distribution of a set (range) of products, counting distribution where *any* of the products is sold. It is a good way of analyzing the sales of substitutable items, such as products available in a bottle, can or carton form, where the shopper is likely to buy whichever packaging is available, regardless of their preference.

DECEPTIVELY SIMPLE FACTS LIKE PRICE

Some of the facts in Nielsen data have colloquial English names, like price. These names suggest that the definition of the fact is simple and obvious, but invariably this is far from the case.

For example Nielsen data sets can contain literally hundreds of facts relating to price, and it is important to know which one is appropriate to a particular analysis. Is *average price* (sales value divided by sales volume) good enough, or does the analysis depend on the actual price that a consumer paid? Does the analysis need the everyday regular price the consumer pays for a product, or the promoted price? And so on.

The take home message is to make no assumptions based on names, and ensure you understand the definition of the facts you are using.
Your Nielsen service team will be able to help.

MODELLED FACTS

Some of the most useful Nielsen facts are calculated by modelling rather than statistical expansion. The most important of these is *baseline.* The baseline algorithm models the sales of an item in the absence of promotional activity. Comparing baseline and actual sales gives an analyst a measure of the success, or otherwise, of a promotional campaign.

The sales of some products, such as Easter eggs, are highly seasonal, while others, such as ice cream, are strongly dependant on external factors like weather. Particularly fast and slow moving products have their own distinctive sales patterns which overlay these factors. The sales of a product can also grow and shrink because the category to which it belongs is growing or falling out of favour (VHS, DVDs, video on demand).

The baseline algorithm needs to model all of these conditions (and more) in order to separate the changes in sales due to promotion from all of the other factors influencing sales.

COMBINATORIAL AND COMPARATIVE FACTS

Nielsen pre-calculates a large number of facts in order to create a well performing, easy to use data mart. A consequence of this strategy is that Nielsen creates a large number of facts which are combinatorial or comparative or both.

The many measures of sales provide an example of combinatorial facts:

- Actual sales.
- Baseline sales.
- Actual sales under the Cartesian product of a limited list of promotional conditions (sales with and without any promotion, with and without price reduction, with and without a display...).

Longitudinal and cross sectional comparative measures, such as year ago *% change in sales* and *% share of market* respectively, also result in the creation of a large number of facts.

The wealth of possible combinations and comparators leads to the explosion in numbers of Nielsen facts which can reach hundreds or even thousands in the data marts created by some systems.

NIELSEN'S VALUE PROPOSITION

The data that Nielsen collects from retailers, panel members and in store cannot be reported as is. Although electronic point of sale data is generally highly accurate, mistakes can happen, and samples (whether of shoppers or stores) need careful maintenance. Even once data is validated, complete and statistically expanded, it still needs to be consistently described and enriched before it is useful.

The final section of this document provides a high level description of the value that Nielsen adds through the curation of its data.

VALIDATION

The first step in the data production process is to validate that the data collected is present and correct. Different data sources have similar but different types of error, and validation algorithms are tailored accordingly. Missing and incorrect data is replaced and corrected.

ESTIMATION AND IMPUTATION

Although failures are very rare, Nielsen collects such a vast volume of data that data which fails validation is a standard feature of the production process. It may not be possible to correct or replace *all* of the missing or incorrect data for a processing period, or to do so without impacting delivery to clients. In circumstances where the technique gives a trustworthy result, Nielsen will use algorithms to estimate missing or incorrect data.

In addition to data validation failures, there may be known voids in the universe. The most usual reason for a void is the unwillingness of a retailer to share their sales data with Nielsen. A recent example was Wal-Mart's decision not to share their electronic point of sale data during the noughties - though the phenomenon is world-wide, and has existed for as long as market research.

Often the retailer comprises a significant portion of the universe being measured. In these cases Nielsen has to find a way of representing the void, and has many techniques for doing so. It may be possible, for example, to use the retailer's sales measured by the Nielsen Panel to estimate electronic point of sale data.

Estimates will never be as accurate, complete and detailed as the real data, but they can provide remarkably useful figures at low to medium levels of detail, and are much better than ignoring a significant portion of the market.

ENTITY ENRICHMENT

Although there are strong identifiers in the data Nielsen collects, none of them is unique or universal in time and space. The effect of this redundancy is felt most in the largest dimension – products. There are several orders of magnitude more products than there are periods, markets and demographics. Products also have many times more attributes than those other entity types, and are much more susceptible to change over time. For this reason the rest of this section on entity enrichment concentrates on products. However the general principles discussed apply to all dimensions.

In theory product barcodes are unique and universal, but barcode reuse and abuse, and the extensive use of codes locally assigned by the retailer, means that Nielsen has a substantial and on-going workload to identify and de-duplicate products. Products also evolve ("new recipe") over time, and these changes need to be recognised and incorporated in the product definition. Nielsen has a battery of techniques to address these problems from pack in hand coding, coding from photographs and textual descriptions, to store visits and clerical and automated product matching. These processes are made more efficient by technologies such as search, machine learning, language and image processing, and optical character recognition.

Once entities have been identified, they are enriched with attributes: for example, products are assigned a manufacturer and a brand. This gives entities a meaningful "description" in addition to a code, and provides the basis for analytic queries against the data.

In order to support analysis within and across data marts attributes and values need to be standardised and harmonised - sometimes across multiple countries, reporting languages and categories. There will almost certainly be subtle differences in the packaging and naming of a product across that span, and in some cases the product may even be owned by different manufacturers in different countries.

In terms of attribute values, Nielsen is a splitter rather than a lumper. Nielsen characterises a product to the maximum level of difference justified by its packaging. Nielsen does not try to collapse "zero calorie", "no calorie", and "less than 10 calorie" drinks into a common attribution of "low calorie". By preserving detail as a matter of principle, Nielsen allows clients to make their own decisions about if and how they wish to aggregate.

To give an impression of the size of the task of maintaining product attributes, as of 2015 Nielsen has 61 million unique attribute values, and adds 5 million new attribute values per year across 35 countries, mostly in Europe and North America. In addition Nielsen maintains assignments of these values to tens of millions of items, each item having on average 30 assigned values - though the number of values can vary by an order of magnitude depending on the category to which the product belongs. In the course of a year between 10 and 20% of all products evolve in a way that requires at least one of their attribute values to be updated.

The number of entities in other dimensions (market, period, demographic etc) is far less than that in the product dimension, as is the number of facts. Nevertheless all of these entities need to be uniquely identified and described in a similar way to products, in order to support standardised, comparable reporting even within a category. For cross category reporting standardisation and harmonisation is critical.

SAMPLING & PROJECTION

Creating and maintaining a sample which accurately represents a defined universe is not a trivial task. The sample needs to be large and diverse enough to answer all of the questions that will be asked of it - but no more, as that would add cost without value.

Also critical to the quality of the reported data are the statistical algorithms used to expand sample data to universe level: in a complex environment, simple, proportional, multiplication of sample data does not give an accurate result.

The projection process is made more complex by the need to decompose the universe in multiple ways, for example by geography and store type. It is further complicated by the need to support combinatorial breakdowns, such as sales by store type within geographic region.

Although a sample will always be a good representation of the universe at a high level, the world is not homogeneous. Geographies and store types may have different expansion characteristics that need to be reconciled. Stores can carry hundreds of different product categories and a sample may not be uniformly representative across every category.

Nielsen's projection factors and statistical expansion models are tuned to cope with this diversity.

Most data marts are created from a mixture of census and sample data. Store observation data is collected on a sample of stores, and needs to be apportioned to the total market, which is comprised of both sample and census data. Projection applies an alignment factor which appropriately scales observation-based metrics to the total market.

This is by no means an exhaustive description of the nuances of sampling and projection, or of the exceptions dealt with by Nielsen. The more diverse and richer the base data, and the greater the level of sophistication required in the reportable data, the more complex the projection process needs to be in order to deliver accurate results.

AGGREGATION, CONFIDENTIALITY & MASKING

Sample data has to be expanded to universe level before it can be used for market research. And as explained above, it is only possible to perform simple calculations on expanded data. Non-additive measures for subtotals and line items must be calculated during the expansion process: the base data needed to calculate complex non-additive measures for ad hoc subtotals is not present in Nielsen data marts – it is aggregated out of existence in the creation of the mart.

The size of a sample may mean that reporting rare events, such as the sales of a poorly distributed, slow selling product, may give an inaccurate reading. Nielsen takes responsibility for ensuring that all data reported is meaningful and representative, aggregating to an appropriate level as necessary. Aggregation is rarely if ever required with sales data, but is used with Panel data when appropriate.

The retailers who provide Nielsen with data, whether from their electronic point of sale systems, audits, or in store observations, often place stringent conditions on the data that Nielsen is allowed to release to its clients. Retailers often require Nielsen to mask information that they consider to be competitively sensitive, such as the sales of their own label products. These confidentiality rules are implemented in the Nielsen expansion engine so that the resulting data marts never disclose sensitive information.

Census data does not require statistical expansion, and is trivial to aggregate. However only a minority of Nielsen's data is census, and census data has to be combined with sample data to create a holistic picture of a universe. Combining census and sample data in a meaningful way is algorithmically complex, and performing sample expansion across a wide range of categories and store types is even more complex. As before retailers are concerned about revealing competitive information and usually make stipulations about how much detail can be disclosed.

The increasing number of product attributes means that creating aggregates is a strategy of diminishing returns: it is not feasible to create and analyse aggregates for all attributes and all possible attribute combinations and hierarchies. Nielsen provides a core set of the most useful aggregates and hierarchies, plus information at the lowest reportable level on the product dimension. The lowest level product data can be aggregated on the fly to create any desired subtotals in this dimension. As explained previously measures such as total distribution points can be used to create useful distribution numbers for such ad hoc aggregates, even though it is not possible to go back to the base data to create a full set of non-additive measures. Analytically critical non-additive measures need to be included in the core set of pre-aggregated data.

Pre-aggregation may mean that a client does not have the subtotals they need for a particular analysis in their mart – most usually because they are performing some sort of data discovery. Nielsen systems such as Answers On Demand™ address this issue by allowing users to interactively define ad hoc data sets and aggregates, which are created semi-synchronously by a Nielsen statistical expansion engine – filling in any void in the provided data set.

OUTPUT DEFINITION & CUSTOMISATION

The input data to a Nielsen retail measurement service, and the sample design are givens: these are too costly to customise on a per client basis – Nielsen configures both based on the aggregate needs of clients.

The periodicity of data in the data mart is determined by the common periodicity of data capture, and is generally not under client control, other than to ask for time aggregations.

The market breakdowns provided are standard, and determined by the sample design.

Nielsen provides standard sets of facts which address most market research use cases, but it is possible for clients to request unique subsets of facts, or to ask for new facts to be calculated.

Outside of these inherent restrictions, the dimensional design of the data mart provided to a client is under the client's control - so long as they are willing to pay for any incremental work involved in creating it. The way that entities are described and attributed, entity hierarchies and the placement of entities within hierarchies, the facts and periods in the output data set can all be customised to meet precise client requirements.

Nielsen has its own standards for identifying and attributing dimensional entities such as products. It also constructs logical dimension hierarchies, such as manufacturer > brand > UPC, to make data easier to navigate and analyse.

These Nielsen definitions and hierarchies may not meet the needs of clients straight out of the box. At its simplest clients may want to remove irrelevant detail – for example by combining the zero calorie, no calorie, and less than 10 calorie drinks into a single classification of low calorie.

At the other extreme, some clients have complex models of their business, products and market, which give them competitive advantage. These clients want the data they receive from Nielsen to be organised consistently with their model, thereby allowing the easy integration of market research data with their business processes.

Nielsen has a set of tools and techniques for implementing client views on top of Nielsen data, which can include deriving a client view dynamically through a complex set of rules, or by implementing the client view based on machine learning, or, in extreme cases by hand. These views can include custom product dimension hierarchies, if desired.

Whether simple, complex or standard, the client view is fed into the Nielsen expansion system and a data mart is created to the client's specification.

MANAGING CHANGE

Inevitably the data Nielsen provides to its clients will change. There are two key sources of change: changes in the external environment, and changes in client requirements.

Nielsen's dimensional model allows entities and entity attributes to be updated and added to a mart with no disruption to the client (for example without schema changes, or changes to unique identifiers). This absorbs a lot of routine environmental change, such as product evolution.

Large environmental changes such as the merger of manufacturers or retailers, or major divestitures are harder or impossible to accommodate in this way, especially if the "old state" has been embedded in the structure of the data delivered to the client - say as elements in a dimensional hierarchy. Imagine that manufacturer is a level in a client's product dimension hierarchy. A major merger takes place. Many products have to be re-attributed to the merged manufacturer. A client may wish to see data for the merged manufacturer going forward, but to see historic data for the individual companies. Alternatively the client might wish to see history restated on the current basis. However the issue is handled, there is a change which needs to be managed, and potentially a significant restatement of historic data in line with current reality.

In a similar way the client's preferred fact set, dimensional hierarchies and aggregates may change as their business and their mental model of it, evolves. Again these changes need to be reflected in the Nielsen data, and historical data may need to be restated in line with current perceptions.

Restating data in line with changes in the real world, or with changes in a client's mental model of their business is a fundamentally sound thing to do. However continually restating data in an uncontrolled fashion can damage the usefulness of the data, especially as a baseline against which to measure performance.

In all cases, from large to small, Nielsen provides careful and controlled change management in order to ensure the maximum benefit and minimum disruption from any changes.

Acknowledgement: I would like to thank the following people for reviewing and contributing to this document: Bob Blake, Art Dirik, John Naduvathusseril, and Ken Rabolt.

Author: Ian Dudley, Enterprise Architect

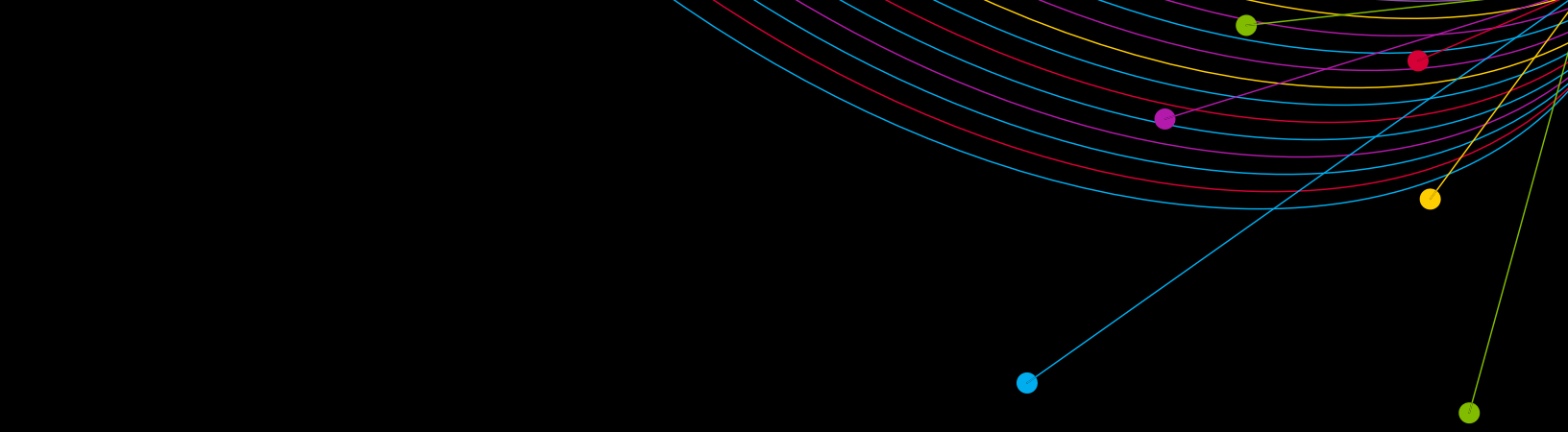
Acknowledgement: Bob Blake, Art Dirik, John Naduvathusseril, and Ken Rabolt.

ABOUT NIELSEN

Nielsen Holdings plc (NYSE: NLSN) is a global performance management company that provides a comprehensive understanding of what consumers watch and buy. Nielsen's Watch segment provides media and advertising clients with Total Audience measurement services for all devices on which content — video, audio and text — is consumed. The Buy segment offers consumer packaged goods manufacturers and retailers the industry's only global view of retail performance measurement. By integrating information from its Watch and Buy segments and other data sources, Nielsen also provides its clients with analytics that help improve performance. Nielsen, an S&P 500 company, has operations in over 100 countries, covering more than 90% of the world's population.

For more information, visit www.nielsen.com.

nielsen AN UNCOMMON SENSE OF THE CONSUMER™



nielsen

AN UNCOMMON SENSE
OF THE CONSUMER™