Deterministic vs Probabilistic Data

In our latest Insights Post, we defined 1st, 2nd and 3rd party data. Today we'll walk through the differences between deterministic and probabilistic data methodologies and what really matters.

Deterministic Data

Put simply, deterministic data is 1 to 1 matching of two or more data sets based on unique identifiers. Typically this refers to “declared data,” which means users inputted their own information, such as signing up for a newsletter. Companies then match cookies or mobile IDs to commonly used identifiers such as names, email addresses, birth dates, and phone numbers to build audience targeting data. Additional user information, behaviors and traits are also then linked to these consumers. All personally identifiable information (PII) is encrypted along the way to protect consumer privacy.

Deterministic data can be built from online surveys, e-commerce purchase portals, software platforms and social media channels like Facebook, Instagram and Twitter. A good example of this is Facebook’s “people-based marketing”, where advertisers are able to target Facebook and Instagram users based on user inputted profile information, liked content, interests and more.

Probabilistic Data

Probabilistic data, like its name characterizes, means an audience group with a high probability to be accurately profiled. Probabilistic data has its roots in a subset of deterministic data in which data scientists then model off of to build a larger group of targeted audiences. They do this by observing relational patterns from multiple datasets and segmenting these out as the larger look-a-like groupings.

To give an example of this, a data provider might take anonymized PII like website visitor data, purchasing data, household financial income and device IDs to come up with a persona for in-market Toyota Prius buyers. They then score the entire populous of cookie IDs or device IDs on how closely each profiles matches up with the persona characteristics. And finally, audience groups are built by including the highest scoring profiles that match the persona.

Which is Better?

Deterministic data will always more accurate, because it's user defined and 1 to 1 matched. But while that's true, it's actually really hard to find deterministic data sets large enough to target audiences at scale. As you can imagine, not everyone fills out an online form or inputs their data somewhere. Large deterministic data owners are also reluctant to release their data to outsiders due to privacy concerns and because they want to hold onto the value of their data. And thus, deterministic data often remains only accessible within a data owner’s walled universe.

An important thing to note is that without scale, accuracy won't mean very much. That's why the majority of 3rd party audience data are probabilistic datasets. The question to answer isn't about which is better, but more about what your objectives are and where your audiences can be found.
**Aligning with Objectives**

If your goal is to target only actual buyers of Kraft Instant Mac, then a deterministic data set would be the option of choice. However, if you’re goal is to target people who might buy Kraft Instant Mac, then probabilistic data will not only give you greater scale, but it can potentially provide better conversions too. Note the word “potentially.” A good rule of thumb to understanding your data is not just where the source data comes from, but also where you hope it will take you. Align your data with your objectives and test to see if it actually brings you closer to your actual goal. In the end, choose the dataset that drives the most impact, whether it is for retargeting, brand exposure, conversions, or something else. Data-driven marketing is really only data-driven when the end results actually count.

Related Reading