DEMYSTIFYING DATA

# Understanding Deterministic vs. Probabilistic Data in Marketing

A practical guide to get you closer to the right data based on your goals

AnalyticsIQ

CONTENTS

# Let's see what's **inside**.


Defining Deterministic vs Probabilistic 3 →

Why & When to Use a Blend of Both 5 →

The Lifecycle of a Data Attribute 8 →

Data Partner Selection Best Practices 9 →

Key Questions for Vendors & Partners 10 →

The Important Role of Privacy 11 →

About AnalyticsIQ 12 →

# Unpacking the industry's misconceptions surrounding **deterministic vs probabilistic** data.

The marketing and ad tech industries often complicate even basic ideas, and the distinction between probabilistic data and deterministic data is no exception. These terms aren't arbitrary; they hold significant influence in data selection.

However, a lack of real comprehension may lead decision makers to unwittingly compromise quality or scale. This whitepaper aims to draw clearer distinctions between probabilistic and deterministic data methodology, as well as offer guidance on when each is most appropriately applied. Let's start by establishing key definitions.

## Deterministic Data

Deterministic data is data that is supplied by people directly or is personally identifiable, and therefore is known to be accurate and true. The most common occurrence of deterministic data is a brand's zero- and first-party data. As an example, if a consumer provides their name and email address to create an account, a brand could make the connection that this email address belongs to this person. That connection is deterministic because it is taken as fact, not theory. This may also be referred to as authenticated data.

These deterministic data points often fuel other deterministic matches as long as one of the identifying data points matches perfectly. As an example, a brand may choose to match their user with a publisher's audience to execute an addressable media campaign. If the email address provided by the brand matches the email address the consumer also uses with the publisher, then a deterministic match is made. For example:

| Brand Supplied Email | Publisher Supplied Email | Match Result |
|---|---|---|
| Johnchen1987@gmail.com | Johnchen1987@gmail.com | Successful match! |

## Probabilistic Data

Probabilistic data is instead based on probabilities, sometimes predictive in nature. Probabilistic data is extremely common in the third-party data ecosystem as most consumers are not actively reporting their interests, future needs or identifiers to advertising platforms.

The overall quality of a probabilistic dataset is based on the quality of the data science team making these determinations. They must assemble and scientifically analyze dozens, hundreds and sometimes thousands of pieces of information together to draw a conclusion.

As an example, let's assume there are two records. One has the name listed as "Jon" and the other has the name "John". However, each record reflects similar information like last name, age and location. By triangulating this information, a data provider may conclude these otherwise separate records are in fact the same person. The result is a probabilistic match, although the data is very rich in nature.

| Record | Gender | Zip Code | Age | Marital Status |
|---|---|---|---|---|
| Jon Chen | M | 30033 | 37 | Single |
| John Chen | M | 30033 | 37 | Single |

## Predictive = Probabilistic

Not only are probabilistic methods using in data hygiene and matching, but it also fuels predictive modeling. And in today's fast-moving and otherwise unpredictable world, predictive data is mission critical. In fact, sophisticated techniques can give brands the competitive edge they need, while outdated data and techniques come at a big expense, costing time, resources and growth.

With the right data, you can avoid these common modeling pitfalls:

- **Predicting the past**: Creating a model that only looks at historical trends and fails to keep up with evolving consumer needs, patterns and cultural shifts.
- **Narrowcasting**: Crafting a high performing model which only works within a narrow dataset or geographic area.
- **Shifting goals:** Having multiple stakeholders involved with competing goals that also change course once modeling has started.
- **Inconsistent definitions:** When building a model, how are responses being defined as either good or bad?

Want to learn more about marketing models? Check out this guide. →

## One Record = Both Deterministic & Probabilistic

Many data decision-makers ask the question, "Is your data deterministic?" while not understanding that the question is over-simplistic. The most advanced data providers, however, should answer the question with a "yes and no." The reason? Because nearly every conclusion made by a data provider can be labeled as either deterministic or probabilistic. And a single audience data set is often comprised of a multitude of deterministic and probabilistic factors.

> "The question of deterministic vs. probabilistic data exists in two areas – the data creation layer and the activation layer."

MIKE HATTUB
CHIEF SOLUTIONS OFFICER
ANALYTICSIQ

## Connecting the Dots Requires Both Approaches

Many data and identity graph providers leverage deterministic matching to draw a conclusion that a consumer identifier (ex: email address) is associated with the same individual or household as another identifier (ex: Mobile ID). This is done by analyzing and matching on multiple data points within their graph to make the link.

However, matches can also be made probabilistically. This approach is used to connect pieces of data that may contain misspellings or variations of structure. As an example, one dataset may only contain first initial and last name, while another dataset has a full name, and yet another dataset is using a shortened version of the consumer's full name (nickname). Data science teams calculate the probability that certain signals indicate a match with varying levels of success. Beyond deterministic identity resolution, probabilistic identity resolution increases reach and also helps resolve identity resolution issues.

Like many things in business, brands should not consider their approach to identity as an "either/or" scenario. The combination of both deterministic and probabilistic approaches is often best for most business cases when it comes to identity and linkage.

## Data In Action

Let's take a closer look at the identity resolution process.

| Record | Address |
|---|---|
| J. Chen | 123 Main St, Atlanta, GA |
| John Chen | 123 Main St, Atlanta, GA |
| Johnathan Chen | 123 Main St, Atlanta, GA |
| Jennifer Rose | 123 Main St, Atlanta, GA |
| Jon Chen | 123 Main St, Atlanta, GA |


> "Data within the activation layer still depends on the media and linkage providers, but it's becoming more deterministic everyday. That is because more and more media sources are including user authentication."

MIKE HATTUB
CHIEF SOLUTIONS OFFICER
ANALYTICSIQ

## Data Attributes

Now that we've focused on identifiers and linkage, let's take a closer look at how probabilistic and deterministic data play into the creation of individual and household attributes.

For instance, did you know that many demographics or psychographics are more commonly probabilistic? Although this data can feel very factual, many consumers do not typically provide hundreds of lifestyle data points openly and accurately, especially with brands they don't have a direct relationship with. As an example, a consumer may indicate an interest in gluten-free products when logged in and browsing their grocery store app. This same info, however, is not likely shared with other brands the consumer interacts with, let alone third-party data providers.

Secondly, just because one piece of information arrives as “deterministic”, smart data scientists understand that relying on multiple signals and pieces of information can lead to a probabilistic conclusion that is, in fact, more accurate. As expressed by Mike Hattub, the Chief Solutions Officer at AnalyticsIQ, “Most of the deterministic data that brands rely on is either first-party data, is highly regulated or geographically thin. It really ends up being difficult to scale. That’s where probabilistic data comes in to save the day. **The reality is that most data within the data creation layer is probabilistic, but buyers don’t realize it**.”

PRO TIP

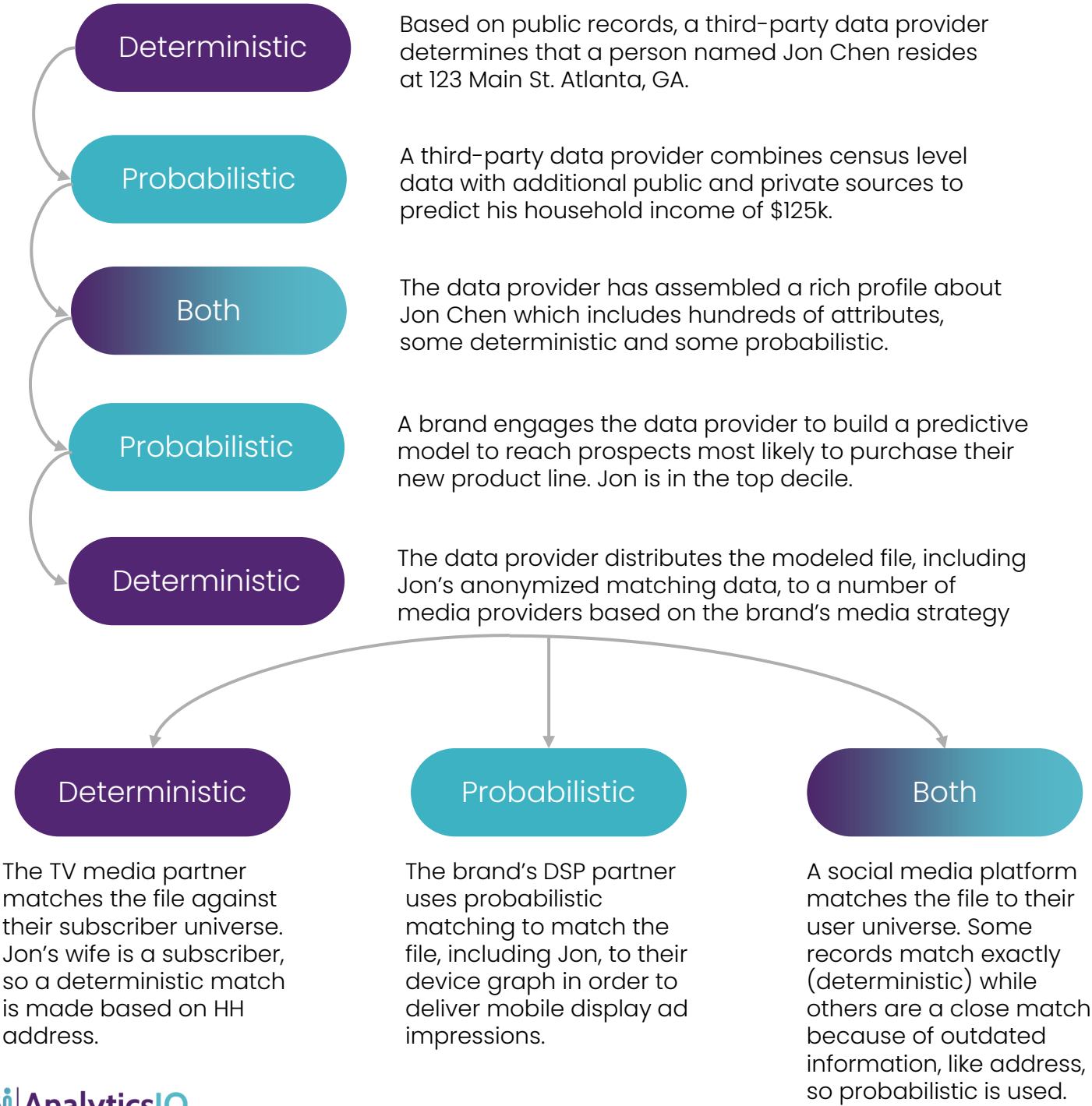
Beware of vendors that claim to supply a scalable target based on a purely deterministic methodology. Vendors that answer conclusively that their data is deterministic may actually:

- Sell targets at a geographically aggregated level only, potentially resulting in media inefficiency
- Work in a field where their data is truly deterministic but is limited in application due to privacy concerns or a lack of scale. (ex; HIPAA data)
- Not understand the nuances of data creation
- Are being deliberately misleading to win a deal

THE LIFECYCLE OF A DATA POINT

# **How and where** data is being used can impact the type of data being utilized.

Whether a data point is deterministic or probabilistic can also change during the advertising life cycle depending on the vendors participating in the campaign. Explore this example.



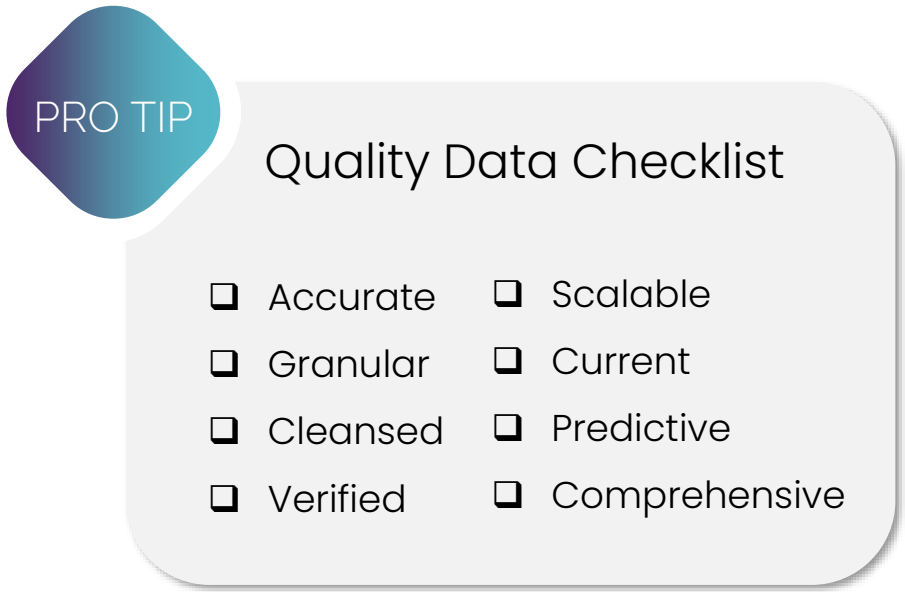
## There is No "One-Size-Fits-All" Approach to Data Selection

Platinum metal is highly valued for its rarity, density, and beauty. However, compared to other metals, it may not always be the best choice. Platinum is a relatively soft metal, is not magnetic, and is less conductive than other metals. Similarly, while deterministic data can appear to be a safe bet, it typically suffers from major scale, fragmentation and privacy issues and therefore struggles to perform across certain marketing channels.

## Vetting Potential Data Providers Requires a Deeper Look

Data buyers should not only understand the fundamentals of how the data they are evaluating is being created, but just as important, they must also ensure it performs consistently, and across channels and use cases. Data buyers should also closely analyze audience size to ensure It reflects an accurate, real-world estimation. Unfortunately, far too many brands appreciate "scale" so much that they end up selecting data vendors whose audience sizes are overinflated, ultimately diluting the overall accuracy, performance and ROI of the data investment.

Identifying a highly sophisticated vendor requires a much deeper analysis than simply asking, "Is the data probabilistic or deterministic?". Evaluators must dig deeper into the data creation methodology, modeling integrity and activation techniques, alongside an actual calculated test of the data.




Want to dig in deeper?
Check out "The Marketer's Guide to Evaluating Data Quality" →

QUESTIONS TO CONSIDER

# Better applications of data start with better, more informed **conversations**.

- **Questions for third-party data providers**
  - How do you construct your database?
  - How do you validate the accuracy of your database, and specific attributes within it?
  - How do you make "householding" decisions? Is it simply street address, or do you factor last name, or local attributes (EX: the presence of a University)?
  - How are you validating that your predictive models will perform?
  - At what level should I be creating my targets? Individual, household, or zip+4?
  - What is your modeling logic, and what specific drivers are contributing to these variables?

- **Questions for all partners working with data**
  - Can you explain your matching logic?
  - How do you determine that an identifier, such as an IP address, matches a household or person?
  - Which identifiers do you consider persistent?
  - Which third parties do you use to for matching and/or onboarding?
  - How do you manage individual and household level updates?
  - Can you share materials on your approach to consumer privacy?

- **Questions for media providers & platforms**
  - How are you preventing bot traffic from viewing impressions for my campaign?
  - How are you preventing users beyond my target from viewing impressions for my campaign?
  - If a third party determines that some impressions were served outside of target, what processes are in place to provide a makegood?
  - At what level will my media be targeted? Individual, household, or zip+4?

PRO TIP

If the concern is "reaching real people", the question of probabilistic vs. deterministic won't necessarily lead you to an answer. Most providers apply a blend of both. A better approach would be to explore how the media provider is blocking poor quality inventory and bot fraud.

## Deterministic Data Attributes Trigger a Deeper Privacy Review

Pure deterministic datasets can also face dramatically higher privacy scrutiny. Most large-scale providers use a blend of deterministic and probabilistic data. For example, age and gender are two data points that are more commonly known, while household income is modeled.

This is because household income is only maintained at scale by an institution like the IRS. Clearly that data is not available to outside parties. Even if it were, it would not be permitted for marketing use cases. If a vendor claimed they had deterministic data on household income for the entire population, it would be a major red flag. The buyer should pursue deeper diligence to ensure consumer privacy is respected, and use cases are compliant.

"Striking a balance with privacy and personalization necessitates trust in vendors. Unfortunately, a single organization's data misstep can cause a domino effect that impacts others, even those with their internal data house in order. Amidst mounting privacy scrutiny, buyers must prioritize working with ethical, knowledgeable partners more than ever."

SCARLETT SHIPP
CHIEF EXECUTIVE OFFICER
ANALYTICSIQ

## Solid Strategies are Built Upon Solid Data, Both Probabilistic and Deterministic

At AnalyticsIQ, we understand the root of the question, "Is the data probabilistic or deterministic?". Often, it's much less about the creation methodology, and more so driven by a desire to ensure the data is accurate. Sophisticated data practitioners understand that highly accurate datasets are typically the result of using a mix of both deterministic and probabilistic methods. Our company's mission has been to create highly accurate data, and continually improve both the deterministic and probabilistic, predictive data we create.

As shared by AnalyticsIQ's Chief Data Scientist Gregg Weldon, "A high quality data model is created using high quality ingredients, many deterministic. However, you also need processes in place to validate accuracy, ensure your models are stable and aren't overly correlated to an existing attribute." This perfectly highlights the common need for a balanced approach.

## Ready To Go Deeper?

The difference between good and bad data for a strategy is complex and nuanced. AnalyticsIQ loves to partner with organizations that want to go deeper, and craft audiences designed to exceed expectations and benchmarks. Whether you need a reliable and consistent database to craft your own predictive models, or quick audiences available for digital media targeting, let's strategize.

Contact us at sales@analyticsiq.com.

## About AnalyticsIQ

AnalyticsIQ is the leading people-based marketing data creator and predictive analytics innovator. Our mission is to fuel better outcomes for all by creating reliable and predictive people-based data by blending cognitive psychology with data science. We help B2C and B2B organizations across industries understand who people are, what they do, and why they make decisions – regardless of whether at home as consumers or at work as professionals.

Our PeopleCore consumer data, BusinessCore B2B data, and Connection+ B2B2C linkages provide insight into individuals that empower organizations to achieve better marketing results. Our fast and flexible approach makes it easy to get started using sophisticated data to grow your business.

Whether you're looking to improve your marketing results across channels, build predictive models, power research, or drive better outcomes, AnalyticsIQ can be your partner. For more information, visit https://analytics-iq.com and follow us on Twitter @AnalyticsIQ and LinkedIn.