truth{set}

Household Identity Accuracy Project

Q3 2023

Confidential

coalition for innovative media measurement

truth{set}

CIMM collaborated with industry partners to support a new initiative, focused on improving the accuracy of **Householding Identity**

PROJECT OBJECTIVE

Develop confidence scores that quantify **the accuracy of the linkages** between person-level identifiers (e.g. hashed e-mails) and households (as represented by postal address)

BENEFITS TO ADVERTISING ECOSYSTEM

1. Better precision and performance where HH data is used
2. Measurement of hidden waste
3. More relevant ads to receptive, in-market audiences
4. Higher engagement and ROI

➔ **A new service to validate the accuracy of identity linkages**

Definitions

HEM	Hashed email address; the primary ID for which we score attributes.
Postal Address	The full physical address of a consumer: Street name and number, unit number, city, state, zip, zip+4.
Truthscore™	A numeric value between 0% - 100% that quantifies the likelihood that a given HEM : Postal address pair is accurate.
Data Providers	15 large-scale providers of US consumer data. Sent Truthset BOTH HEMs and postal addresses.
Validation Sets	Providers of verified, highly accurate consumer data. The ground truth about HEMs, their postals, and their demographics. 20M+ distinct records, dating back to 2015.

Household Identity Underpins Streaming TV

truth{set}

In order to target and measure audiences, we rely on linkages to combine data sets covering **devices**, **households**, **people**, and their **audience attributes**.

Devices

:

Attributes

- ❑ Age/ Gender
- ❑ HHI, HH comp
- ❑ Ethnicity
- ❑ Pets, auto, etc

Household Identity Underpins Streaming TV

truth{set}

To assign attributes to an IP, the IP must first be assigned to an address, which then is assigned to people, and finally assign the attributes about the people.

Devices : Households : People

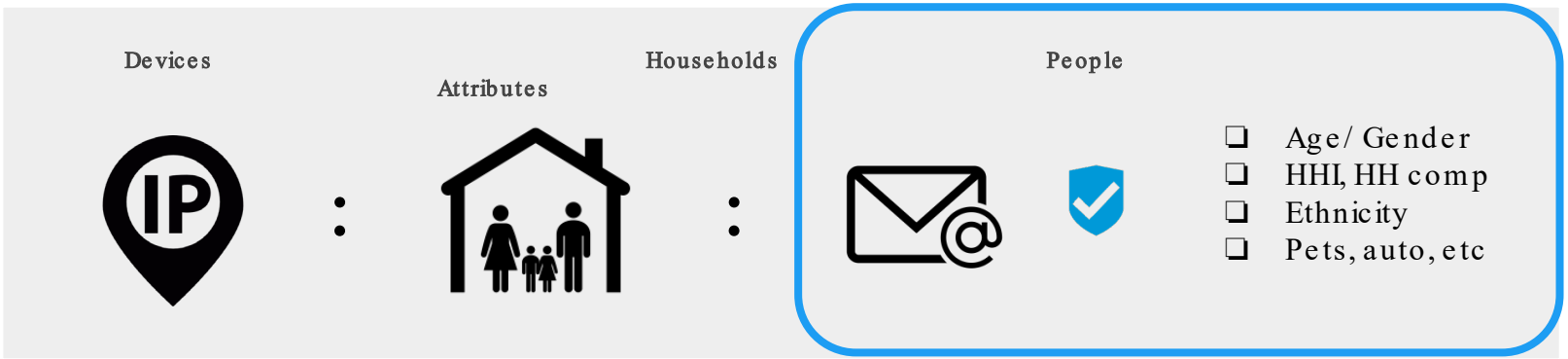
Attributes

- ❑ Age/ Gender
- ❑ HHI, HH comp
- ❑ Ethnicity
- ❑ Pets, auto, etc

Attributes are Primarily Associated with People

truth{set}

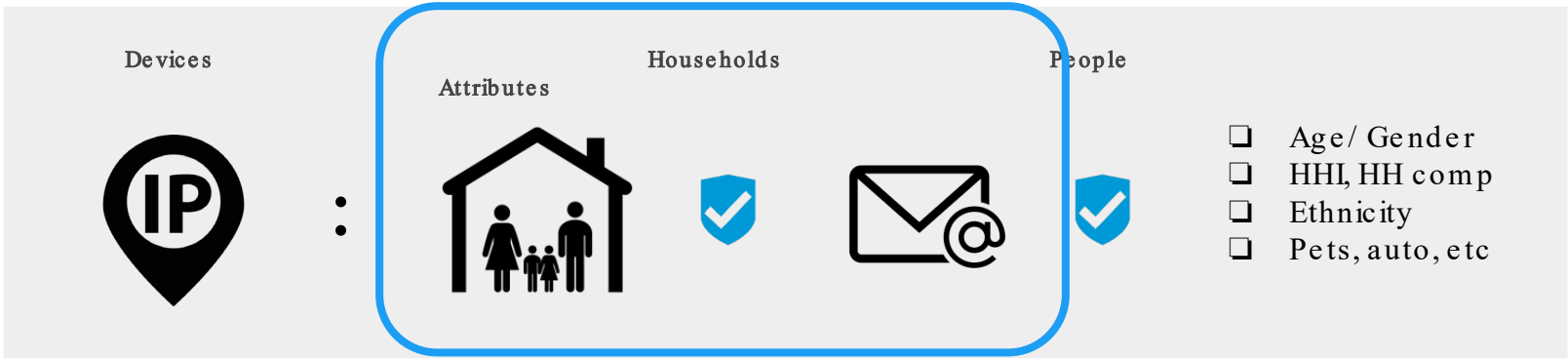
Average error in demographic datasets, when linked to email, is **20-60%**, depending on the attribute.



People Are Associated With Households

truth{set}

People use **email addresses** to authenticate media consumption and, often, purchases.

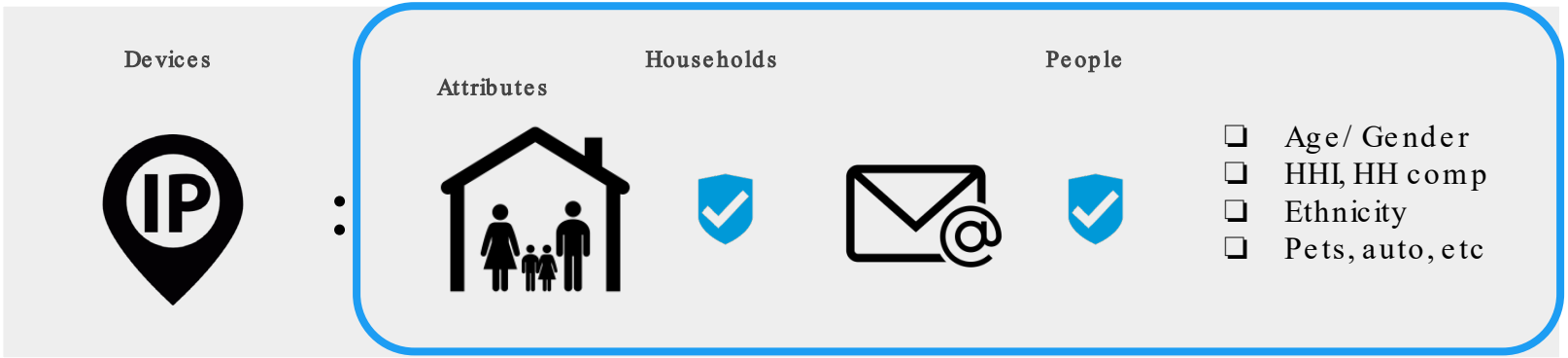


➔ This study measures the accuracy of email-to-postal linkages.

Combined Project Scope

truth{set}

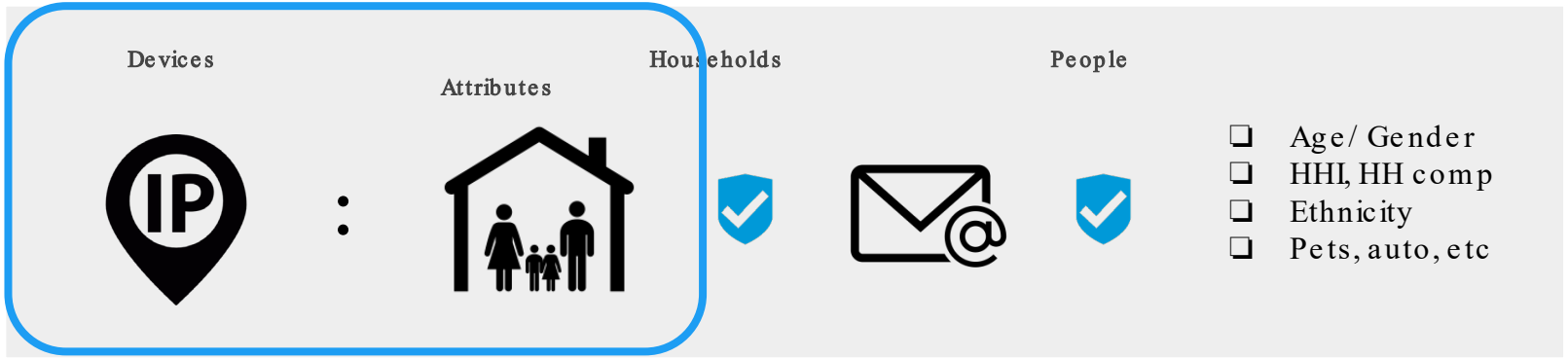
This service allows any stakeholder to identify error in the linkages between Households, People, and Demos, thereby improving accuracy and performance.



Next Phase: IP to Households

truth{set}

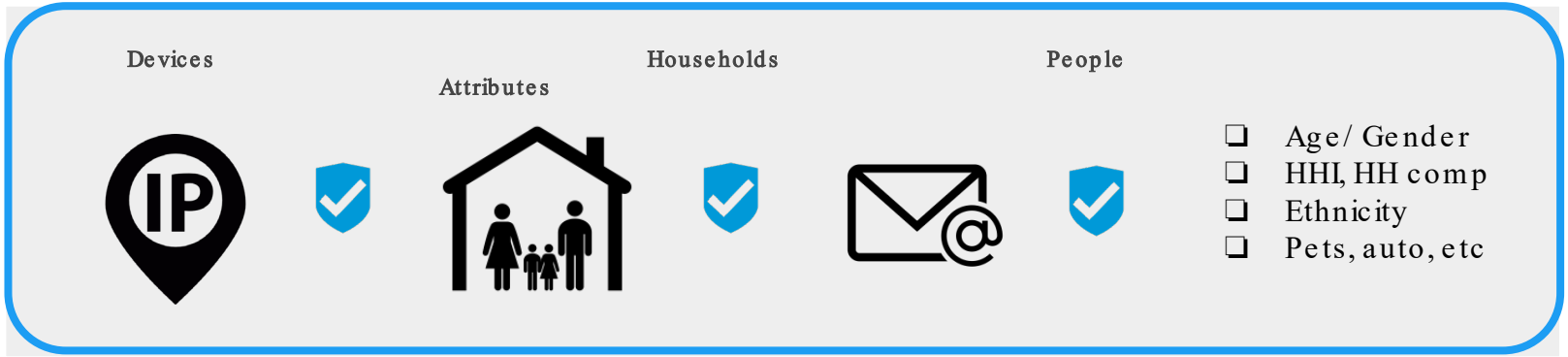
A future phase will measure the accuracy of IP t0 Postal, and provide full control of accuracy when building and measuring IP- and HH-based audiences.



Transparency & Control Across Identity Linkages

truth{set}

The final product will provide transparency into the accuracy of each set of linkages, allowing buyers and sellers to determine exact accuracy thresholds for a given use case.



Data & Methodology

Analytics Objectives

1. Measure the accuracy of HEM : Postal Address linkages at scale

1. **Deploy a re-usable algorithm** across an initial cohort of leading, large-scale identity linkage providers

1. **Calculate and assign Truthscores™** – estimated probabilities of HEM : Postal linkage accuracy – for >95% of all providers' HEM : Postal linkages

The output: Record-level scores of the accuracy of any HEM : Postal linkage pair

Truthscore™ Algorithm At-A-Glance

Data Providers

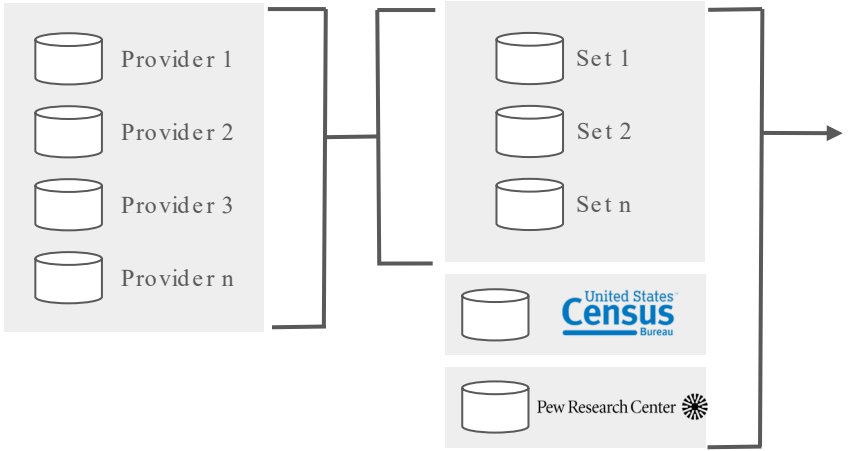
- 15+Leading Companies
- 790 mil+unique HEMs
- 130 mil+unique postals

Validation Sets

- Independent data sources
- Self-reported, declared data
- 20 million+panel and ecommerce records from 2015-2023

Truthscores™

- 0-100% accuracy scores per record
- 1.2 bil+unique HEM : Postal pairs



ID: Hashed Email	Address: 123 Main Street	Address: 456 Other Street	Address: Unknown
abc@gmail.com	82%	12%	6%

Result: Accuracy Scores for All HEM:Postal Linkages

HEM : Postal accuracy scores are ready for release.

User ID	Postal 1	Postal 1 Accuracy	Postal 2	Postal 2 Accuracy	Male	Female
HEM123	123 Main St	82%	789 Cedar St	18%	2%	98%
HEM456	456 Elm St	79%	123 Oak St	15%	45%	55%
HEM789	789 Pine Ln	91%	456 Elm St	5%	8%	92%
792m HEMs …	…	…	…	…	…	…
	Truthscores (0%-100%) for each attribute and value, representing the probability a given HEM has the given attribute					

Findings

Core Stats – Total Scale of Data

Universe of HEMs and postal addresses analyzed

Total HEMs

@ 3.9 billion

Total HEM : Postal Linkages

2.6 billion

Unique HEMs

@ 792 million

Unique Postal Addresses

133 million

Unique HEM : Postal Linkages

@ : 1.2 billion

Estimated US Census Coverage

90+%

On average:

- A postal address corresponded to **9.1 emails**
- @ An email address corresponded to **1.6 postal addresses**

Key Findings

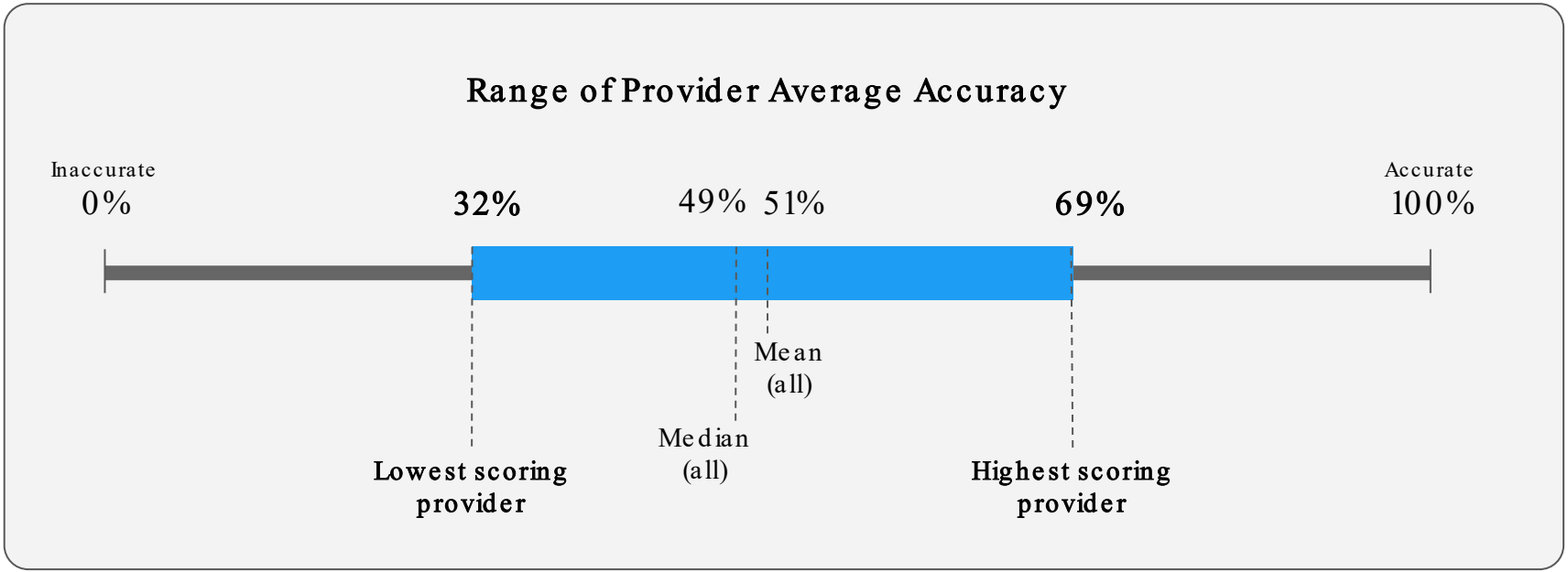
1. On average, **HEM : Postal linkages** are **accurate 51%** of the time

2. **HEM : Postal accuracy** **varies up to 37 percentage points** between providers (between 32% and 69% accuracy)

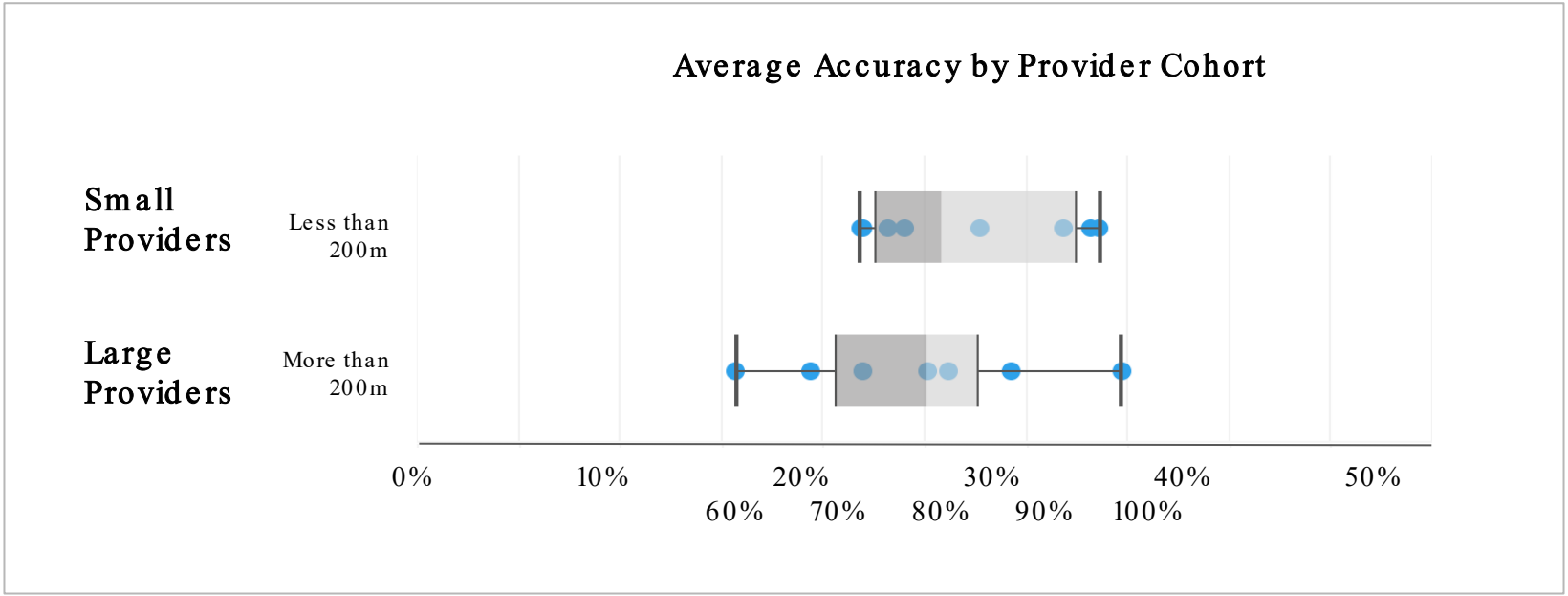
3. A provider's **scale of IDs** **is not predictive** of the accuracy of their HEM : Postal pairs

4. **Every provider** **has variability in their accuracy**. A provider with 69% accuracy has both highly accurate and highly inaccurate records.

There is Significant Variation Between Providers



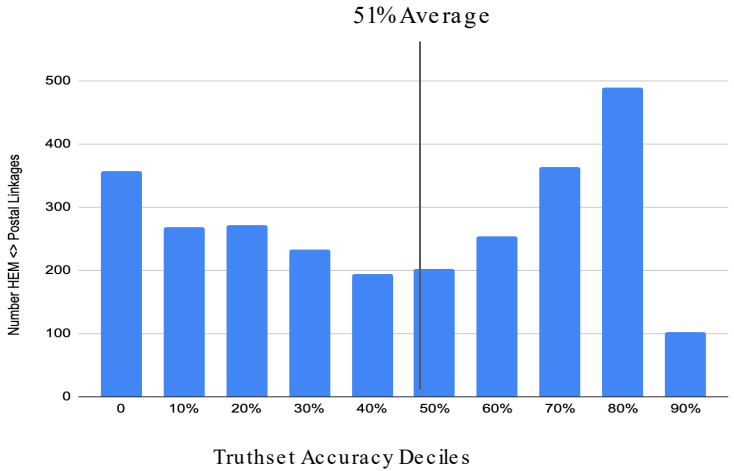
There is Considerable Variation Among Providers…



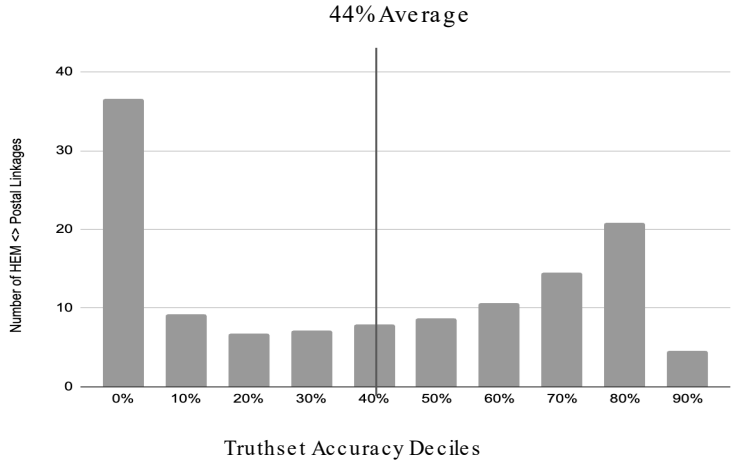
.regardless of size

Variation is Present on Total and by Individual Provider

All Providers

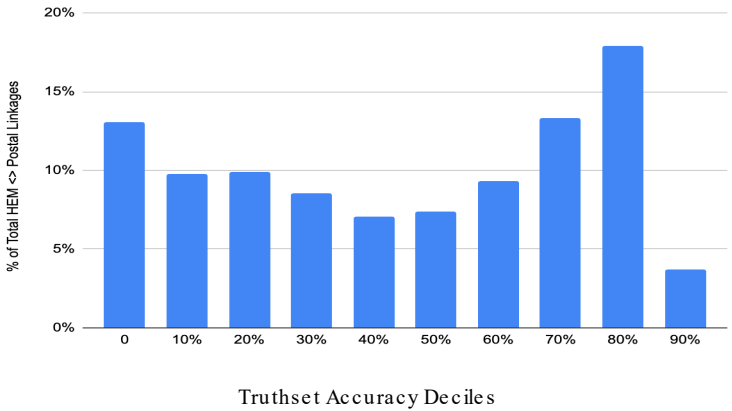


One Provider

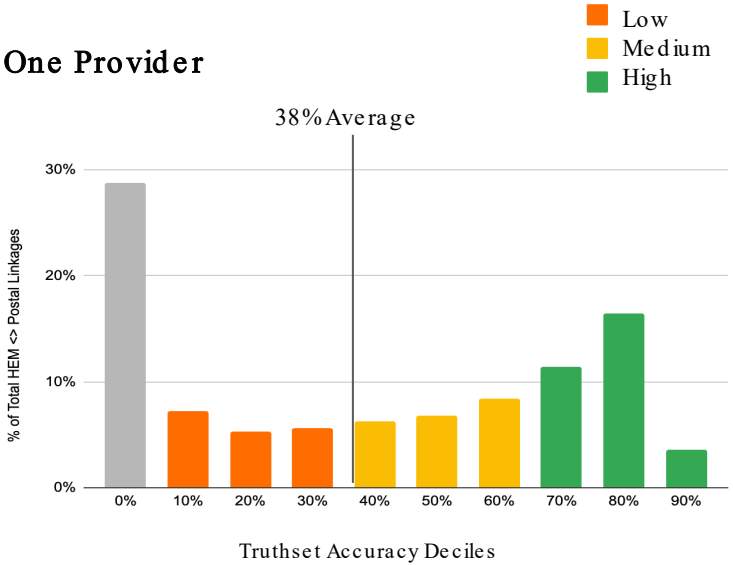


Data Accuracy for Different Use Cases

All Providers

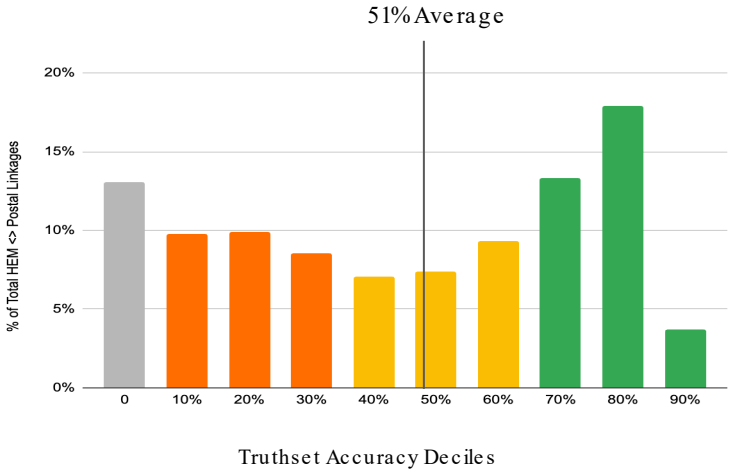


One Provider

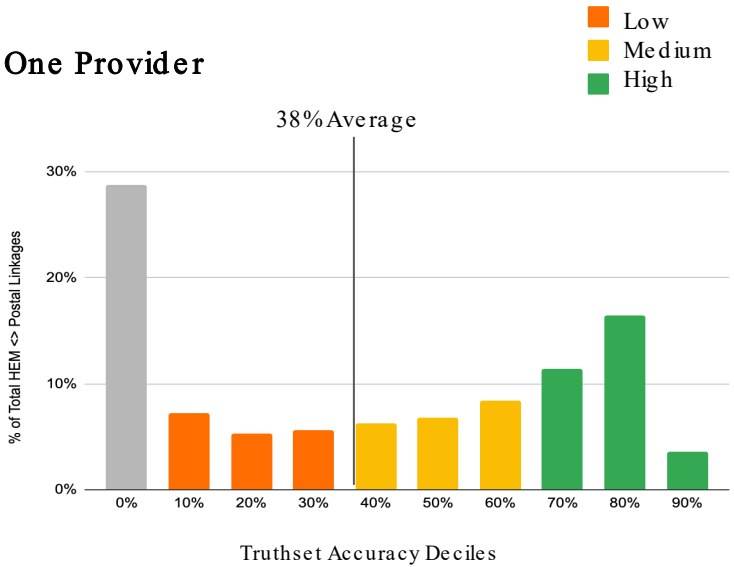


The Future: Multi-sourced Data Ratings

All Providers



One Provider

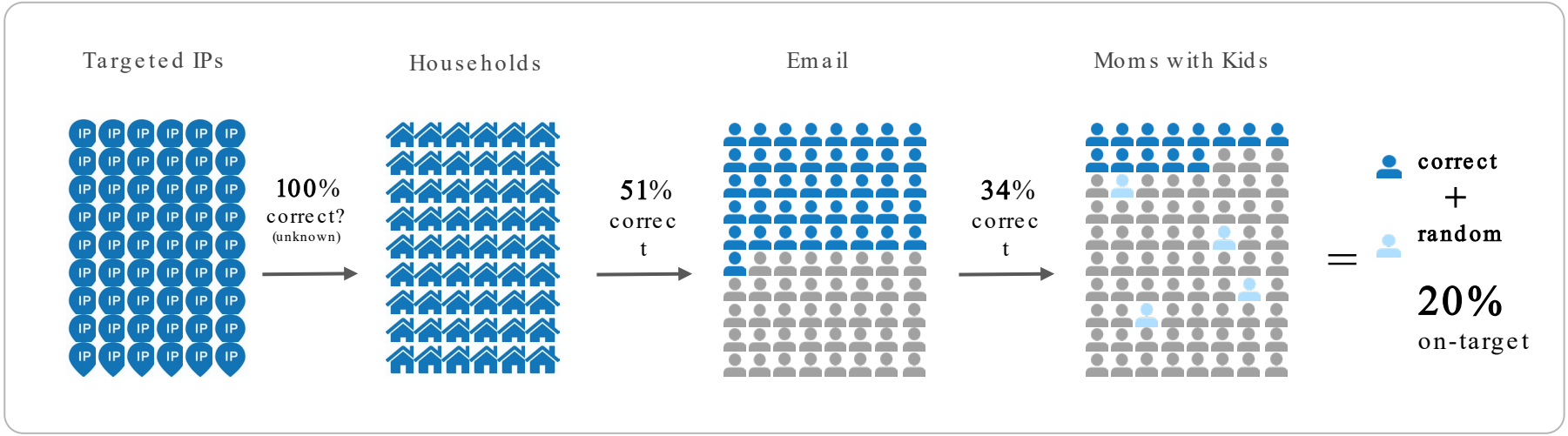


Validate Data Before You Use It

Business Implications

Data Validation Identifies Waste

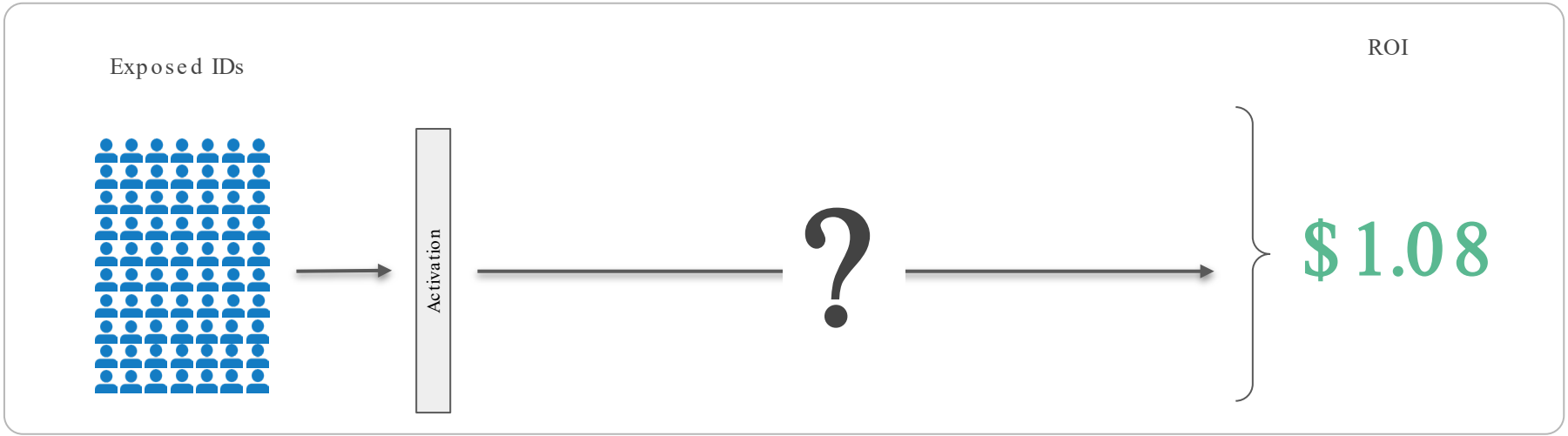
Target: Moms with children in the home



Opportunity: Increase on-target-% by 3-4x

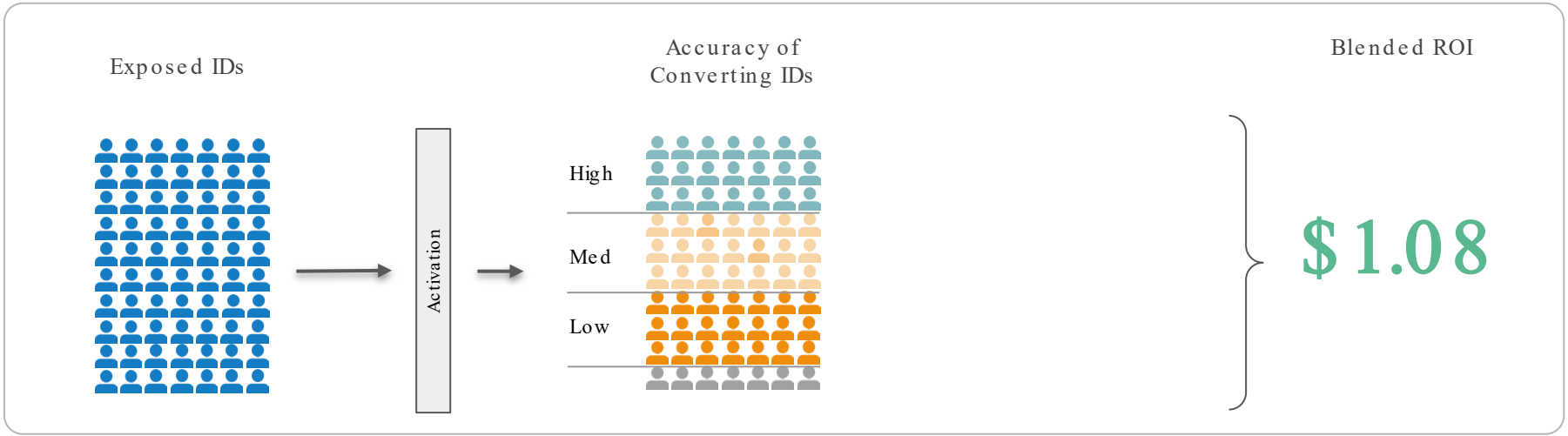
Data Validation Increases ROI

Example: Targeting may produce a positive ROI, but how can it be improved?



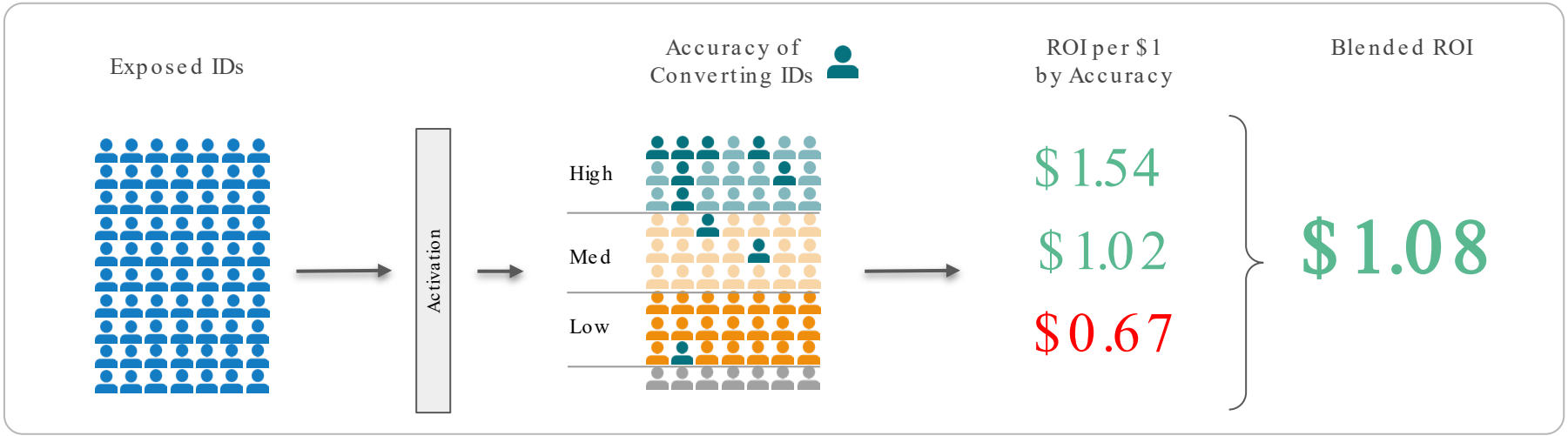
Audience Data Can Be Stratified by Accuracy

ROI is lowered by off-target exposures. Profitability comes from ads reaching receptive audiences who are actually in-market.



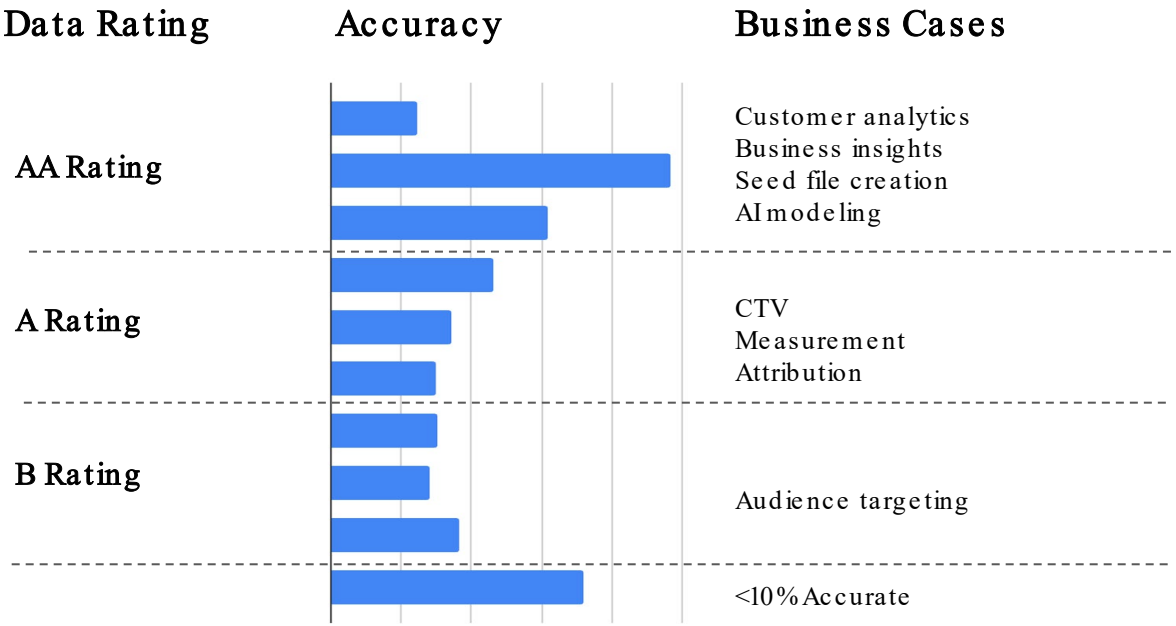
Accurate Targets Convert Better

ROI is lowered by off-target exposures. Profitability comes from ads reaching receptive audiences who are actually in-market.



Opportunity: Target validated IDs to increase precision and ROI

Different Data is Suited for Different Purposes



All data has value. But different grades of data will perform more profitably than others, depending on the use case.

Create Your Plan of Attack

Run a Diagnosis

1. How are you managing identity & demo data today?
2. Validate your 1st and 3rd party data for accuracy (in as little as a week)
3. Quantify the impact of better data on your organization

Implement Solutions

1. Run an A/ B pilot for a segment of your business, create a case study for internal cohorts
2. Understand right mix of accuracy for various use cases
3. Engage with partners and providers to establish accountable systems

Remember…

1. Increasing data accuracy doesn't have to be expensive or time consuming
2. Change is inevitable
3. Don't wait for your partners to figure this out
4. Data accuracy will be a key deciding factor of success in programmatic/ AI future
5. Accurate data leads to better quality products & outcomes

Accurate Data Is Better for Everyone

truth{set}

How you can benefit from validation of 1st party, 3rd party, and identity data:

Brands & Agencies	→ More precision, greater ROI
Networks & Publishers	→ Less waste, higher CPM, better yield
Measurement Suppliers	→ Credible currency-grade data; accurate projections
Data Providers	→ Differentiated pricing based on accuracy
Consumers	→ Greater ad relevance

Appendix

Methodology First Principles

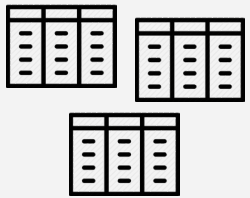
truth{set}

- Our HEM : Postal Truthscores **are built upon the same first principles** as the HEM : Demographic Truthscores:
    - HEM : Postal Truthscores are numbers between 0-100 (i.e., they are estimated probabilities).
    - HEM : Postal Truthscores are assigned for every unique combination of HEM : Postal (i.e., at the record level, every HEM receives Truthscore for every postal address it is linked to)
    - Smaller, research grade validation sets establish the ground truth about HEMs and their postals.
    - We unify all available information about a given HEM : Postal combination from across all Data Providers, and feed all this information into an algorithm.
- Our algorithm is tasked with estimating how likely is it that a certain provider, or group of providers, has made a correct HEM : Postal assignment. The algorithm has the option to choose **that no provider is correct.**
- Whenever possible, we use our HEM : Demographic Truthscores as features in the HEM : Postal Truthscore algorithm.
- We check the performance of HEM : Postal algorithm against a 20% holdout of our validation set to ensure the model is both accurate and well calibrated.

Validation Sets Overview

Identify & Vet Validation Sets

Must be committed to **data hygiene & data recency**

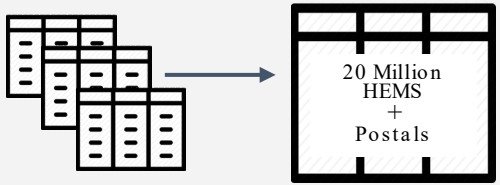


Perform Data Hygiene & Cleansing

Truthset Data Science team performs **additional hygiene**

Merge Multiple Validation Sets

A combined asset of over **20 million HEMs and postals**



Example Validation Set: E-commerce data with 100m+ logged-in users providing both shipping & billing addresses. Encapsulates all customer orders placed +shipped since 2015.

The Truthscore Algorithm

- This (combined, cleansed) validation set serves as ground truth → this is a set of HEMs : postal linkages that we know are accurate.

- HEM ; Postal linkages from the Data Collective are **the inference set** → this is a set of HEM : Postal linkages with unknown accuracy.

- **The intersection of HEMs from the Data Collective and the validation set is used to train and test the HEM : Postal algorithm.** For this set of HEMs, we know:
  - The HEM's true address, according to the validation set,
  - What each member of the Data Collective asserts as the HEM's address
  - The likely demographics (demographic Truthscores) of that HEM

- The HEM : Postal Truthscore algorithm analyzes this set of HEMs to uncover patterns that define **accurate** HEM : Postal pairs.

- These patterns are then applied to ALL HEM : Postal pairs in the Data Collective (the inference set). The result? 0-1 Truthscores, representing likelihood of accuracy, for 1.1B linkages.

How Error May be Introduced

- **Stale linkages**
  - Theory: How recently a linkage was last updated is directly related to the accuracy of that linkage.
  - HEM : Postal linkages with some recent signal, or that have been updated more recently, are more likely to be accurate

- **Frequent Changes of Physical Address, Correlated to Demographics**
  - Theory: Certain segments of the population (e.g., younger folks) move around more often. These consumers may have many physical addresses.
  - Therefore, the accuracy of a linkage is tied to the demographics of the underlying consumer

- **HEMs (and people) may have more than one "accurate" postal address**
  - Theory: Many-to-many linkages make it harder to tease out signal and noise, identify accurate linkages.

- **Error compounds over multiple linkages**
  - Theory: There is built-in error when accurately linking people to households
  - Hardly any methodology assigns a HEM directly to a postal address. Instead, people are clustered into households and then device IDs/ HEMs are clustered into people
  - If an email is correctly assigned to a person, but that person is incorrectly assigned to a household, then the Email : Postal address linkage is automatically inaccurate.

Potential Analytics Next Steps

1. **Repeat analysis quarterly**, track industry HEM : Postal linkage accuracy QoQ
2. **Add more metadata** on HEM : Postal pairs
   - For example, one theory is that less accurate linkages have been updated less recently (i.e., are stale)
   - Adding metadata (like a timestamp) to this analysis will allow us to identify further predictors of overall partner accuracy
3. **Add more identifiers** to pinpoint where error is introduced
   - Examine person-level IDs (first and last name) and quantify HEM : Person accuracy then Person : Postal accuracy separately
   - This will allow us to tease out if the majority of the error comes into grouping IDs to people, or grouping people into households.
4. **Layer in Truthset's demographic accuracy Truthscores** on top of linkage accuracy Truthscores, to quantify how error compounds across household identity and household attribution
5. **Run experiments** to assess IMPACT of linkage accuracy on outcomes
   - For example: Case studies targeting low, medium, and high accuracy households; tracking how accuracy impacts outcomes (ROAS, conversions, brand metrics)