flight

flight

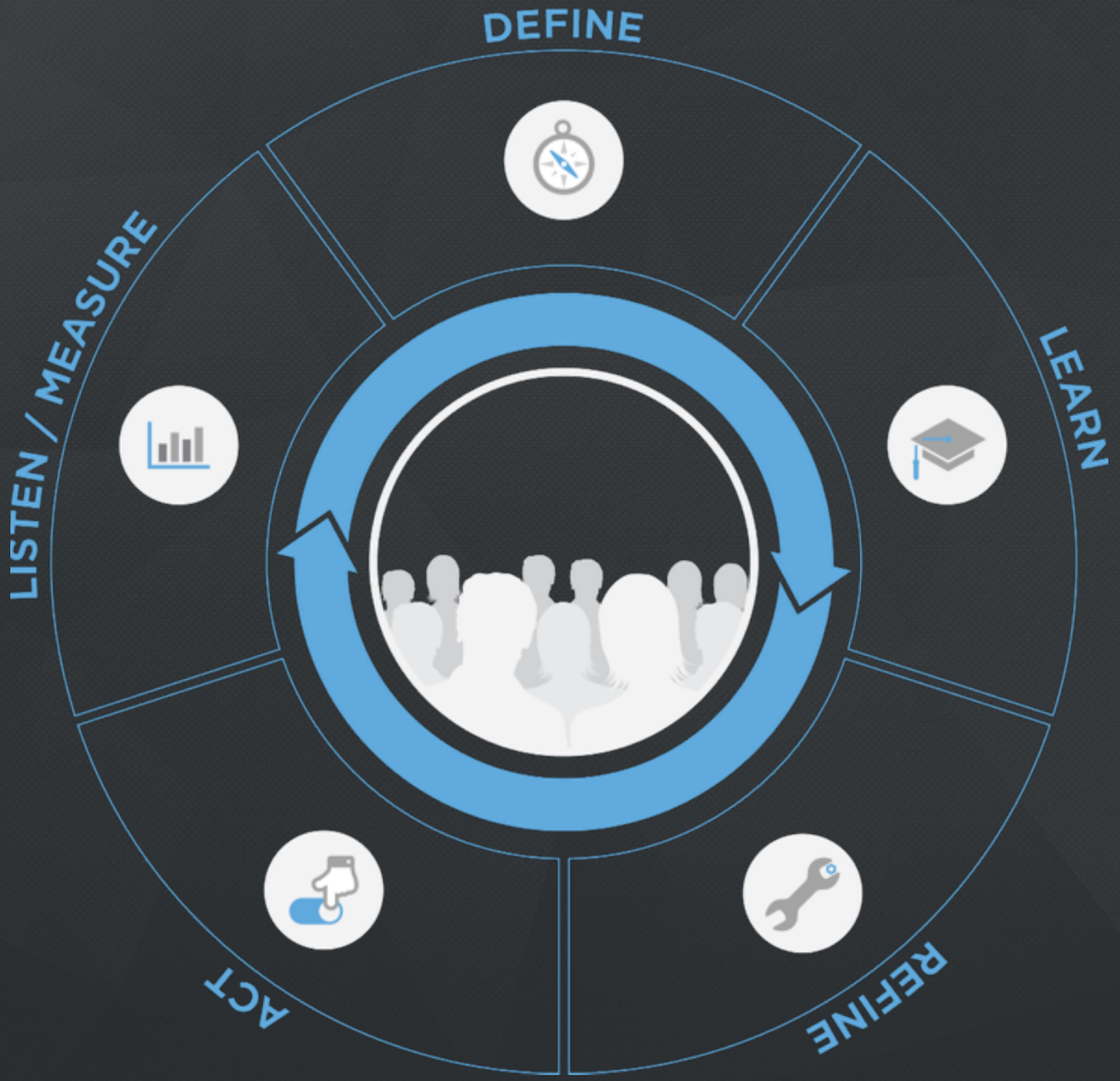
Understanding Your Audience: Using Probabilistic Data Aggregation

Jason Carey

Software Engineer, Data Products

@jmcarey

The Vision



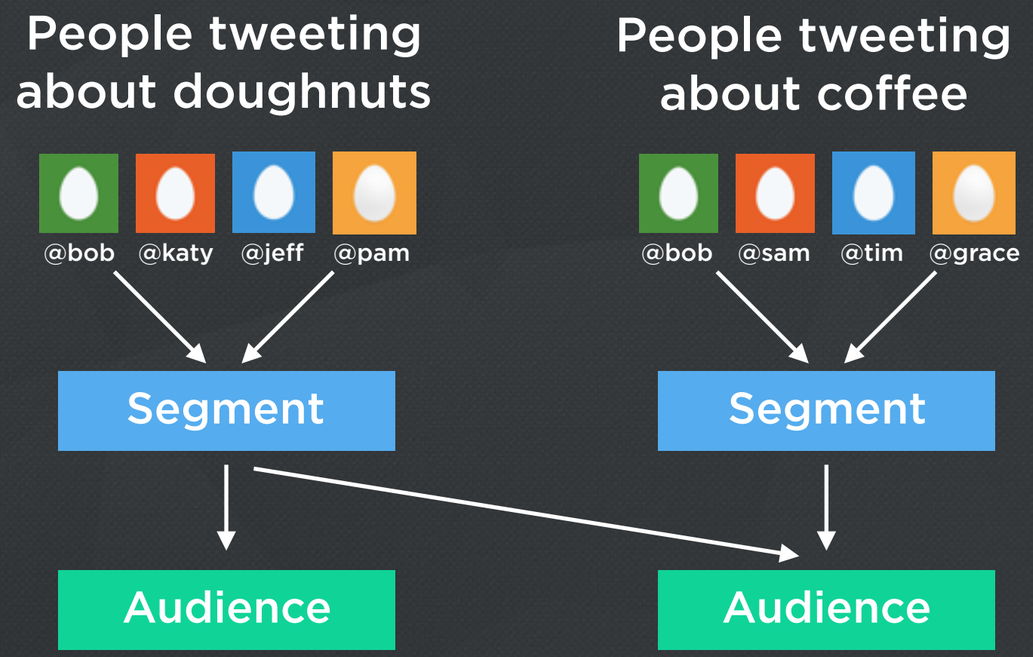
flight

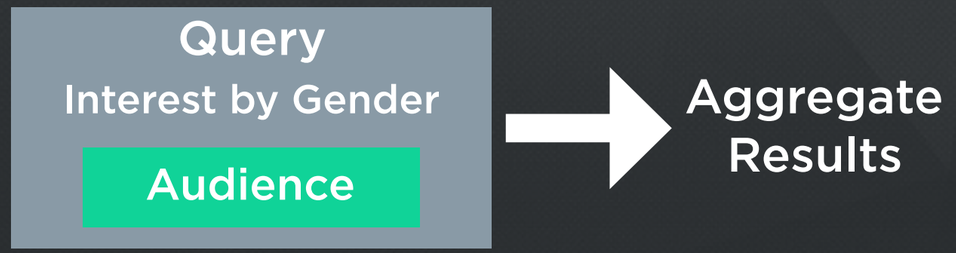
Insights Audience API

- Fast, ad hoc aggregate queries
- 10+ proprietary demographic models
  - Device, Gender, Interest, Language, Location, TV
- Audience size up to tens of millions of users
- User privacy is of paramount importance

Birdseye View

- **User:** a Twitter user
- **Segment:** a collection of users used to build audiences
- **Audience:** a collection of segments that can be queried
- **Query**: an audience and a collection of demographics

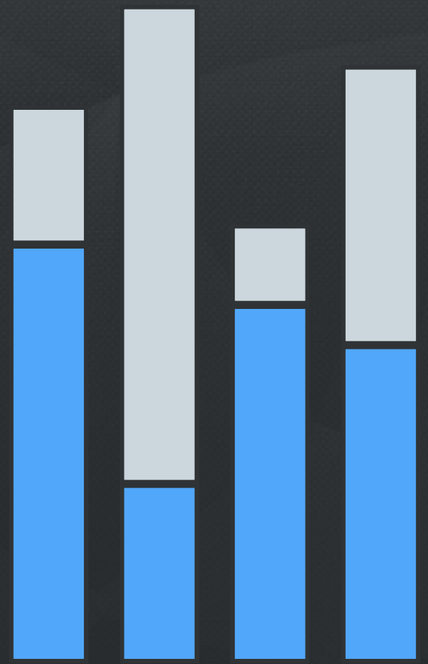




flight

Example Query

For people tweeting about my brand, tell me the top TV shows by gender:



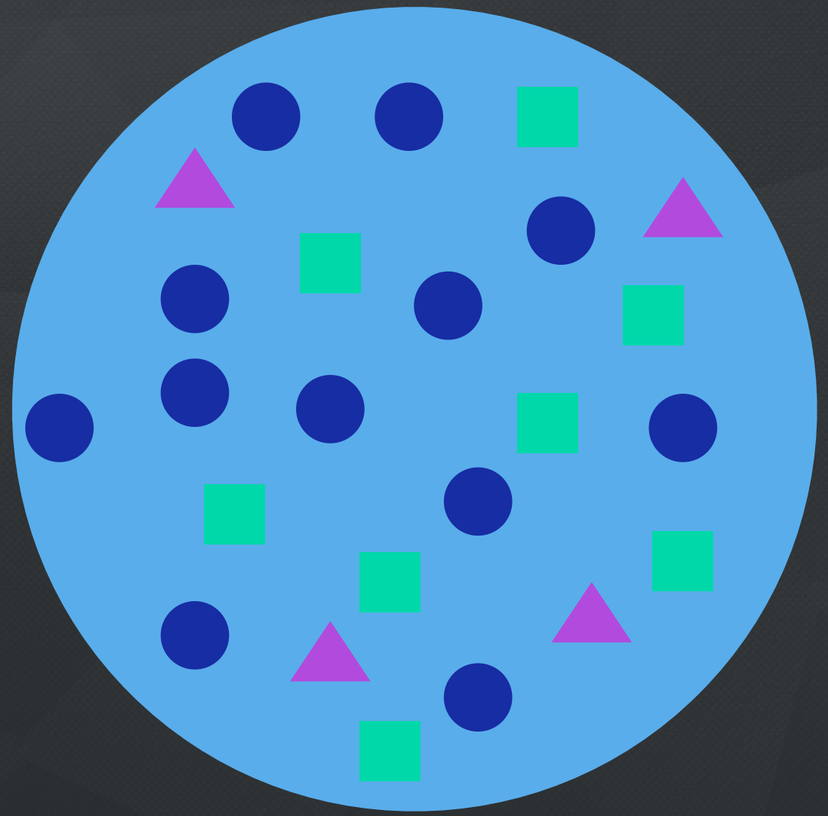
```
{
  “Female”: {
    “Game of Thrones”: “7”,
    “House”: “24”,
    “The Daily Show”: “4”,
    “Wimbledon”: “14”
  },
  “Male”: {
    “Game of Thrones”: “21”,
    “House”: “9”,
    “The Daily Show”: “18”,
    “Wimbledon”: “16”,
  }
}
```

7% of audience are females following Game of Thrones

flight

Technical Approach

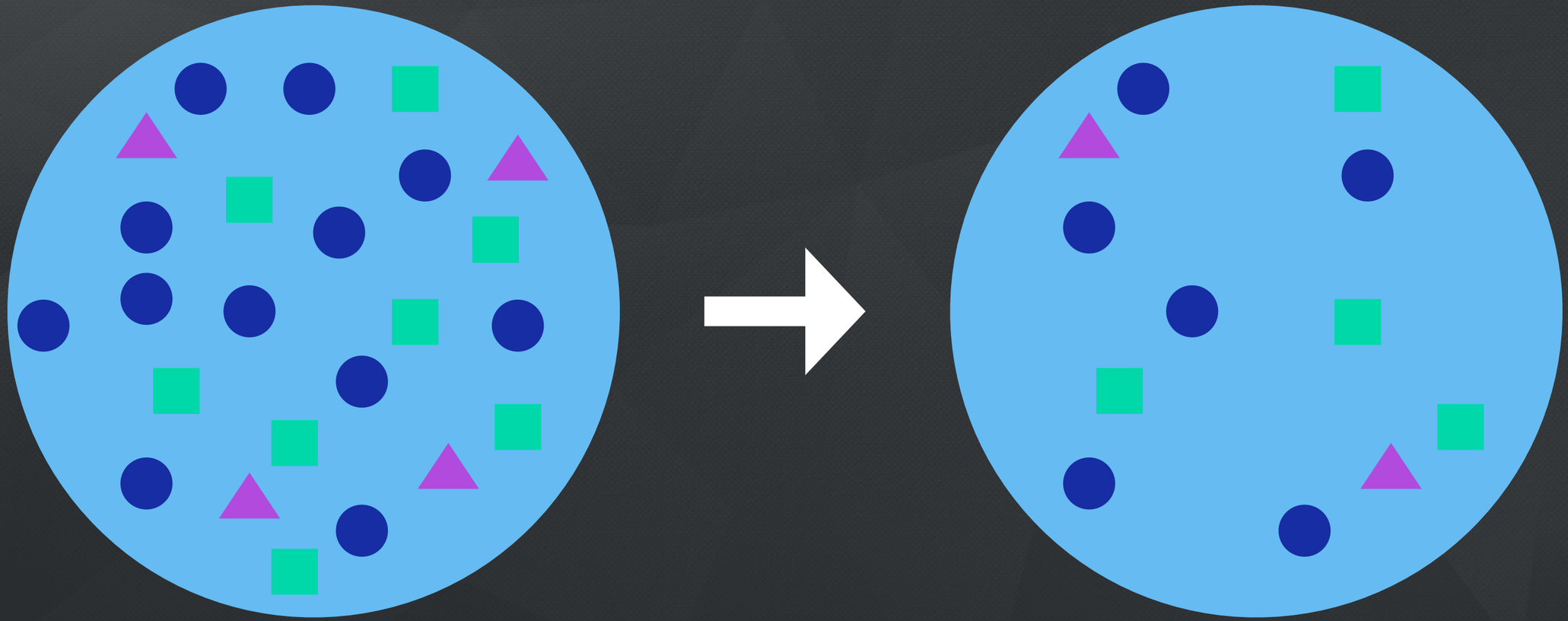
Technical Approach



Users in a Segment

flight

Technical Approach



Users in a Segment

Users in a Random Sample

flight

Technical Approach

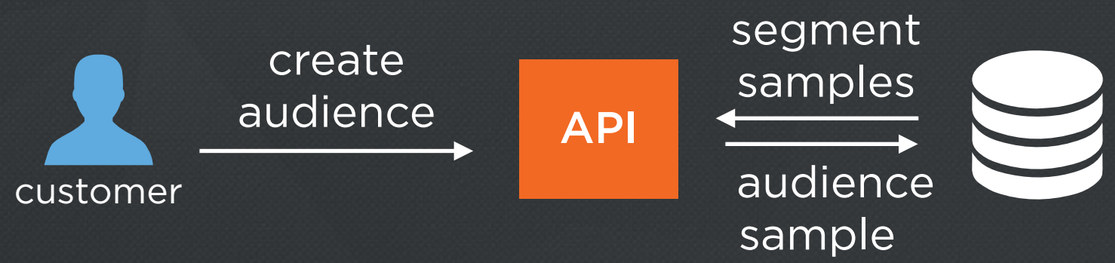
When adding users to a segment, maintain a random sample

flight

Technical Approach

When adding users to a segment, maintain a random sample

When creating an audience, merge segment samples into an audience sample

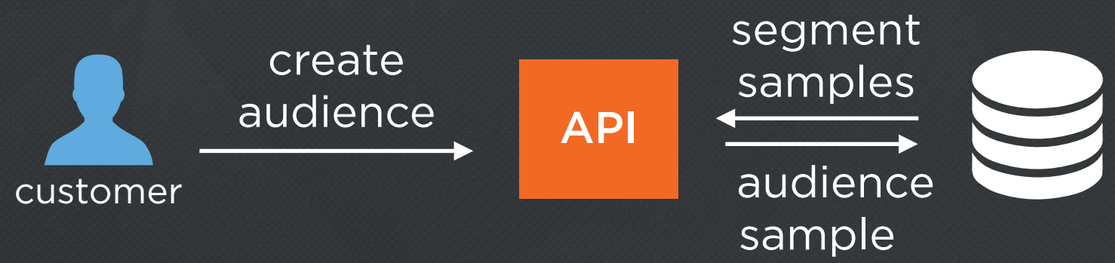


flight

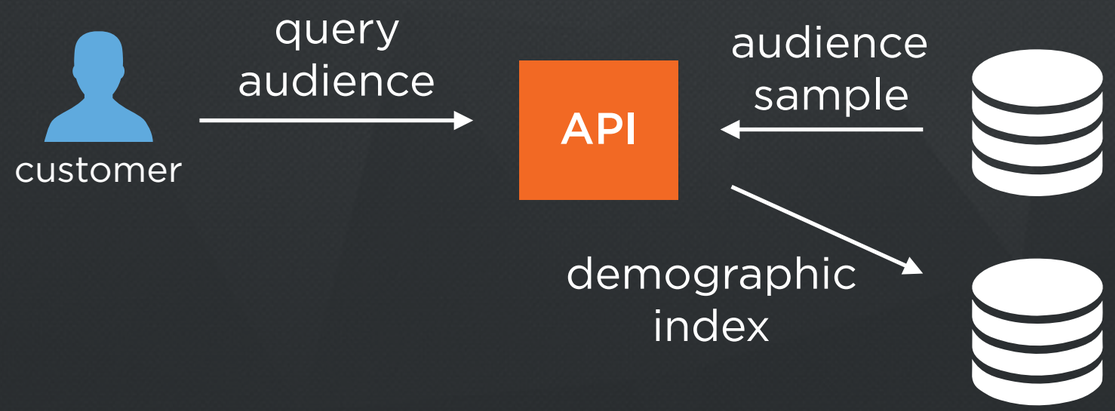
Technical Approach

When adding users to a segment, maintain a random sample

When creating an audience, merge segment samples into an audience sample



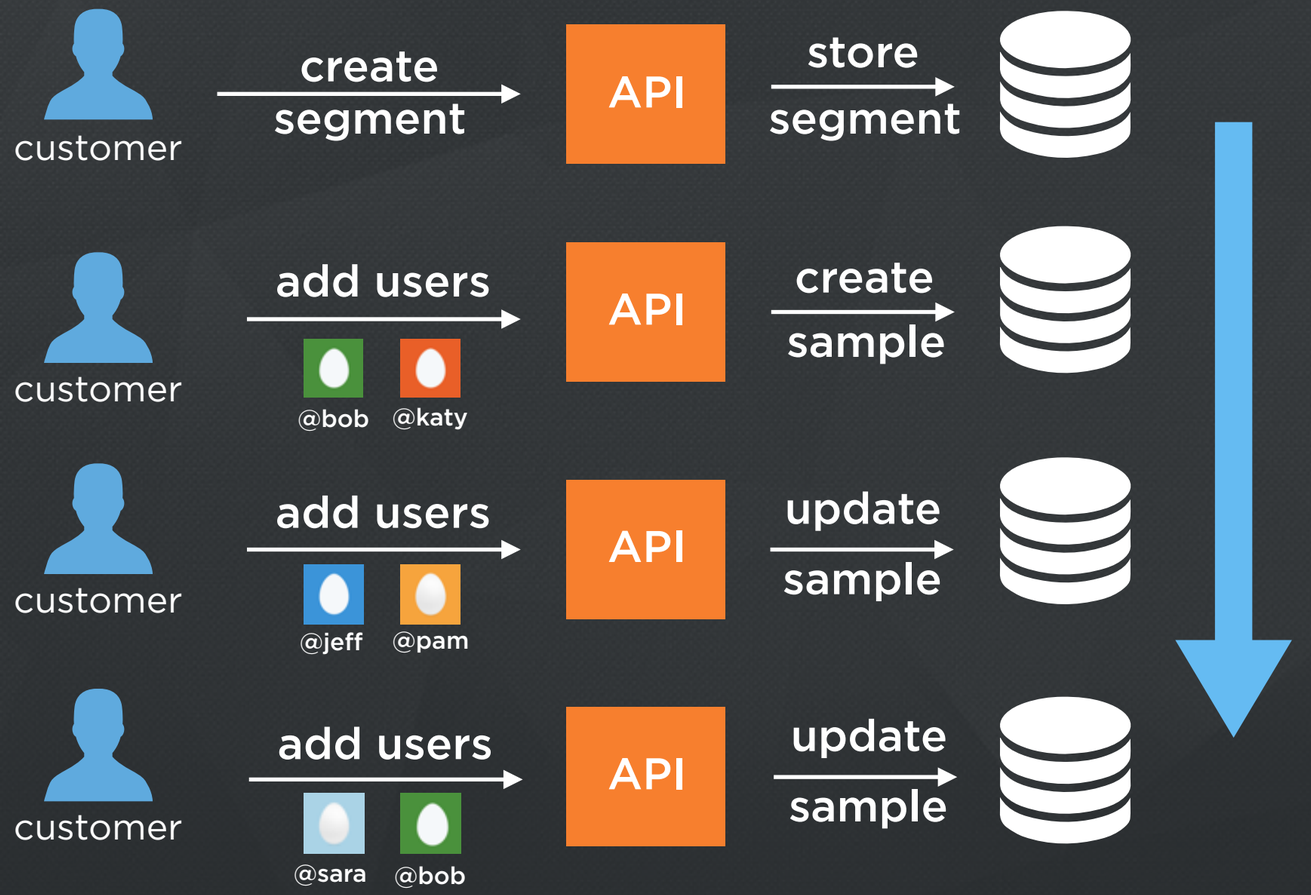
When querying, call demographic index for users in audience sample and aggregate results



flight

Segment Sampling

- Technical Challenges
  - Segment can be updated over multiple requests
  - Duplicate users may be sent in requests

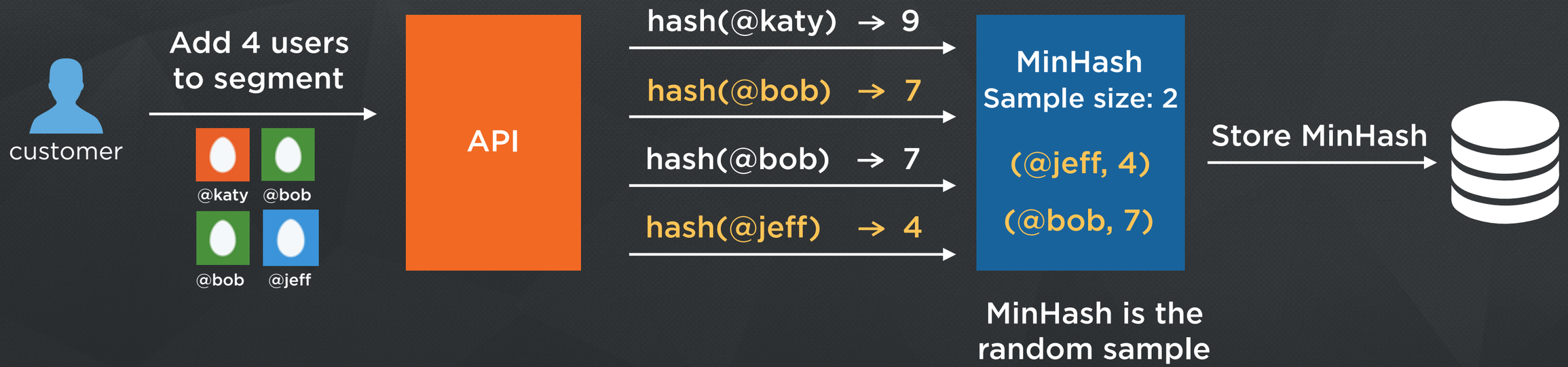


@bob is sent in two requests

flight

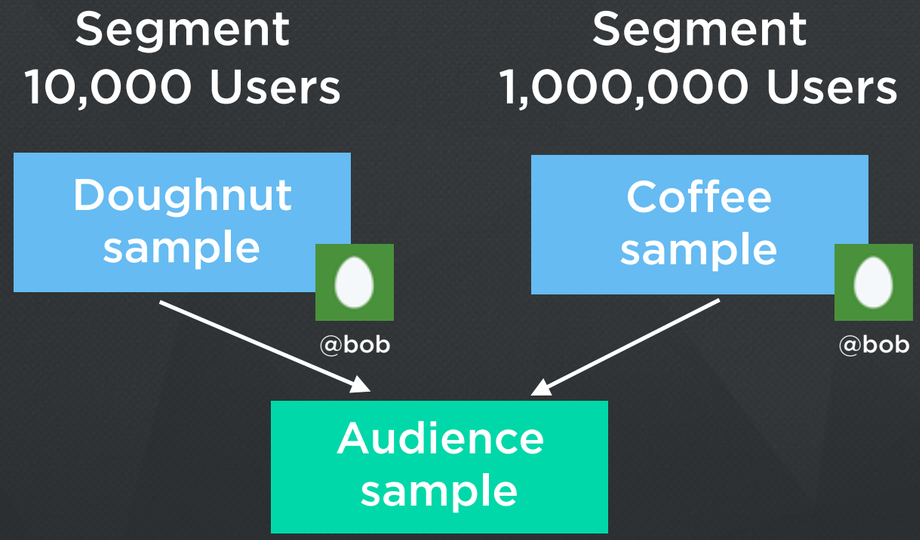
Segment Sampling

Maintain a MinHash for each segment



Audience Sampling

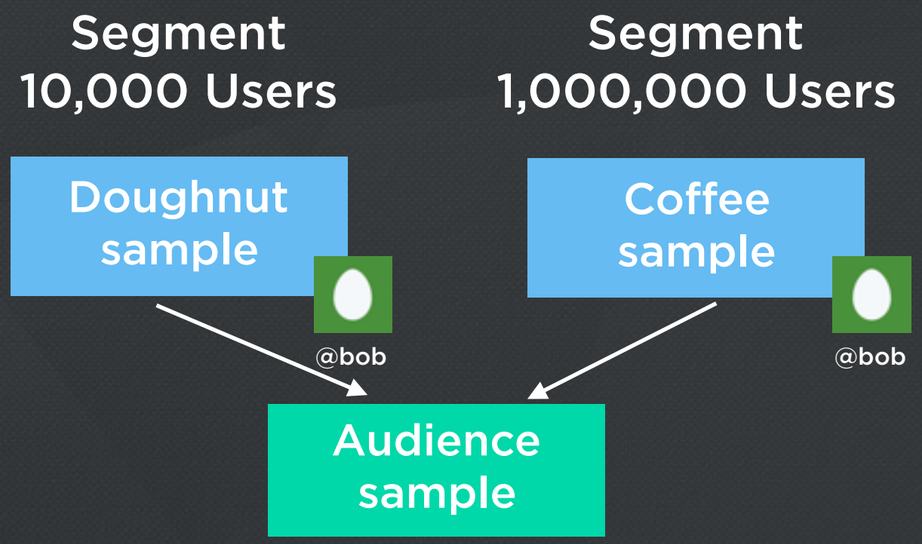
- Technical Challenges
  - Duplicate users across segments
  - Merging segment samples into a single random sample



flight

De-Duping Users In Segment Samples

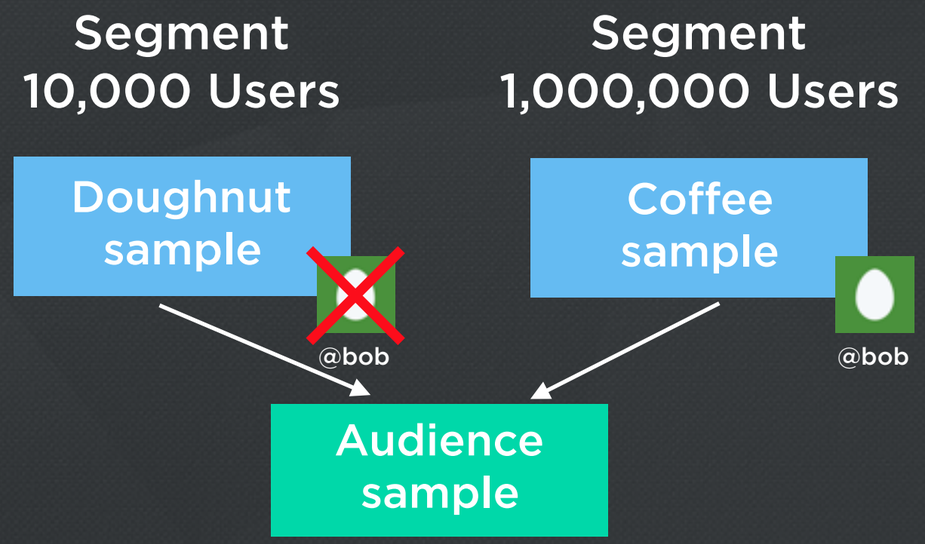
If the same user exists in multiple segment samples, does it matter how you de-dupe?



De-Duping Users In Segment Samples

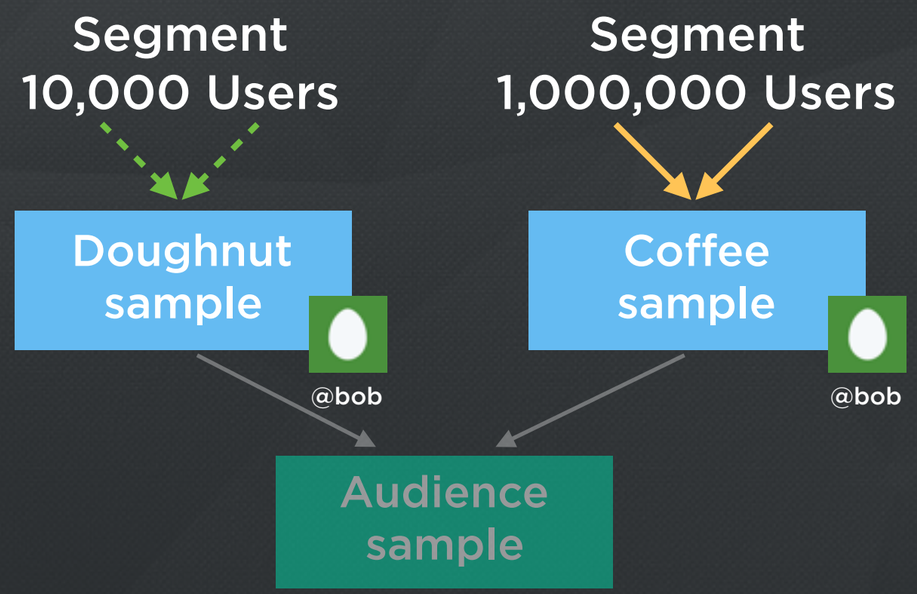
If the same user exists in multiple segment samples, does it matter how you de-dupe?

No, it does not matter *as long as the audience sample is constructed using a weighted selection based on the segment cardinalities.*



flight

De-Duping Users In Segment Samples



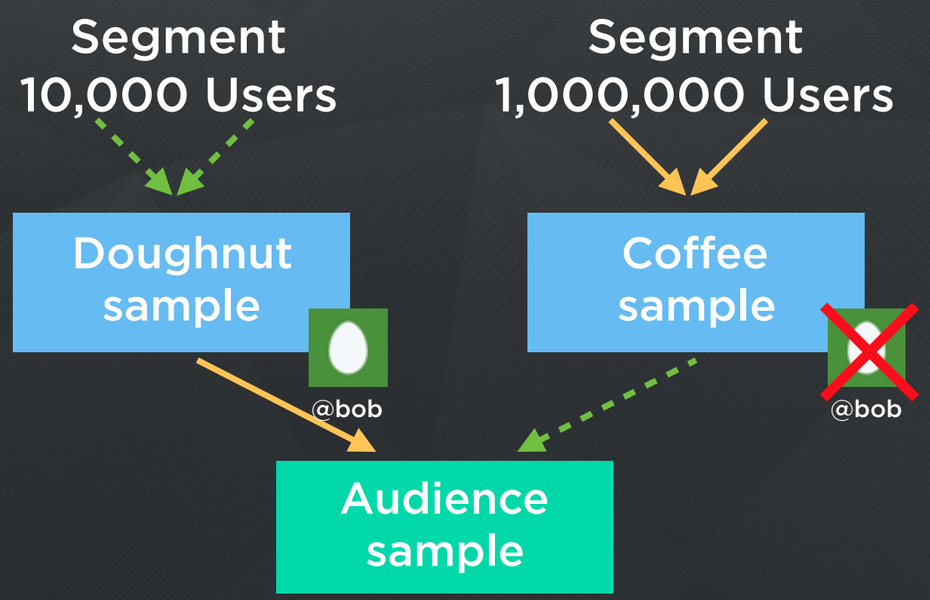
---- with higher probability

—— with lower probability

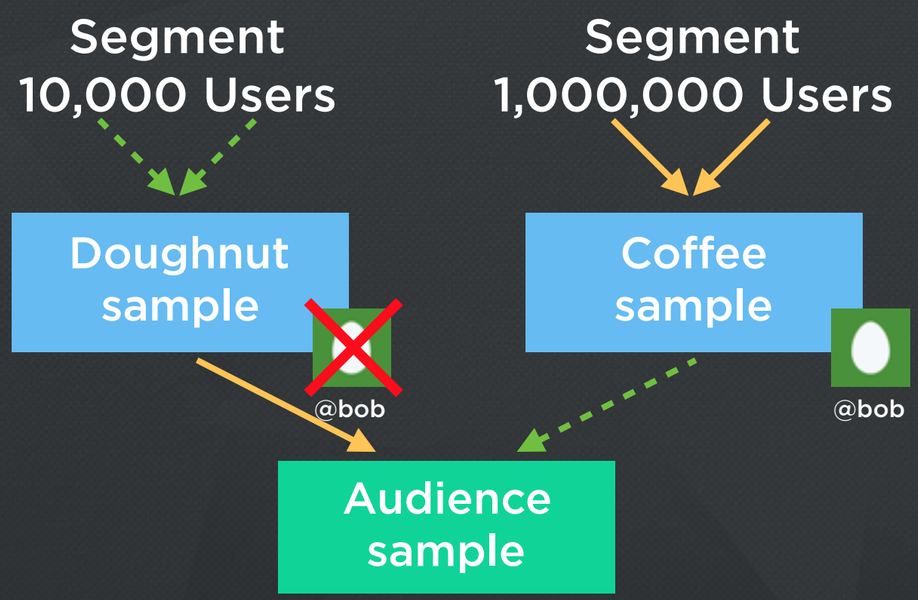
flight

De-Duping Users In Segment Samples

Option 1
@bob in Doughnut sample



Option 2
@bob in Coffee sample

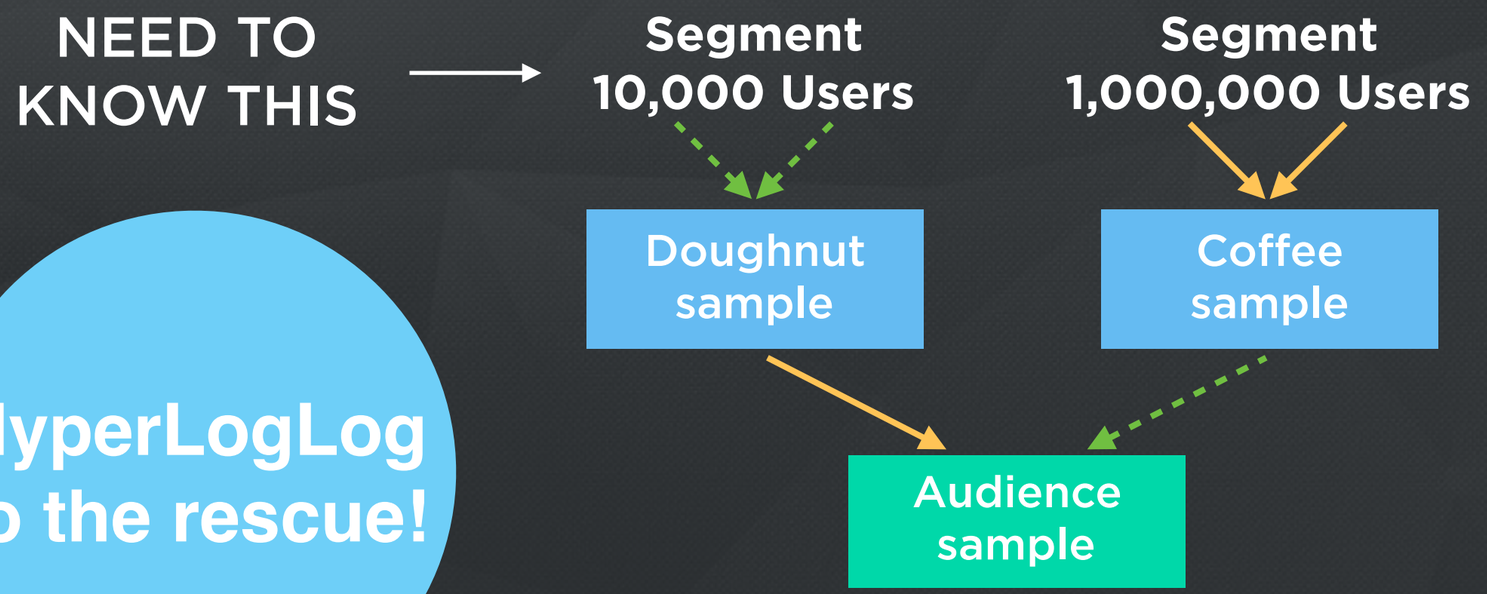


---- with higher probability

—— with lower probability

flight

De-Duping Users In Segment Samples



flight

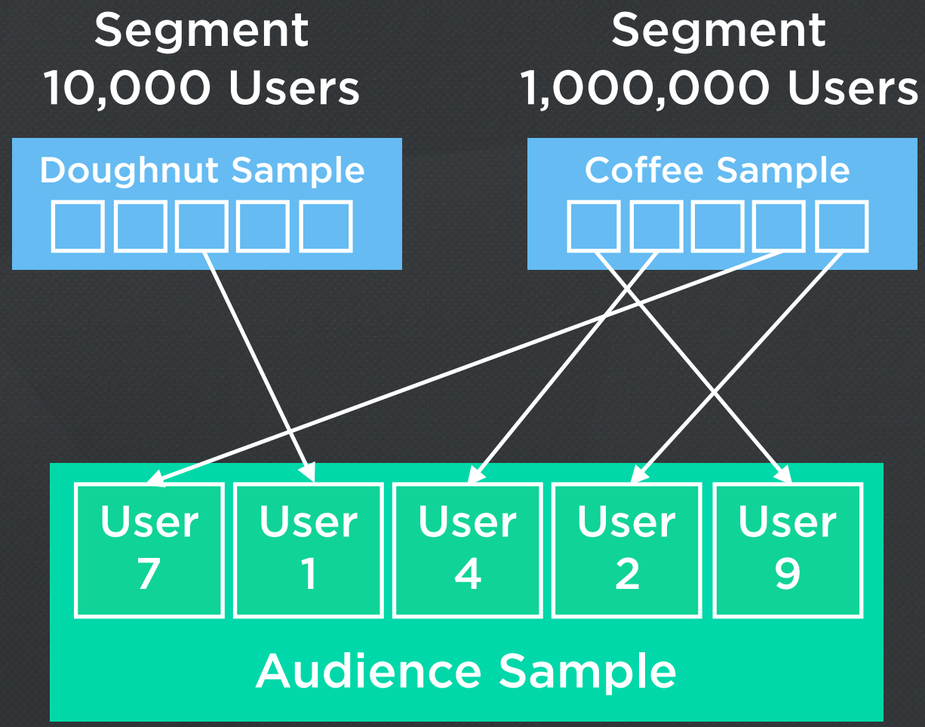
Did Somebody Say HyperLogLog?

- Estimates the number of unique items in a data stream
- A billion items requires only KBs of storage for an accuracy of 2%
- Intuition: I flip a coin a bunch of times and tell you the largest number of heads flipped in a row. I then ask you how many times I flipped the coin.

Flajolet, P.; Fusy, E.; Gandouet, O.; Meunier, F. (2007). "HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm" (PDF). AOFA '07: Proceedings of the 2007 International Conference on the Analysis of Algorithms.

Merging Segment Samples

- Repeat until sample size is reached
  - Weighted selection of segment sample based on segment's cardinality
  - Randomly choose a user from selected segment sample



flight

Sampling Recap

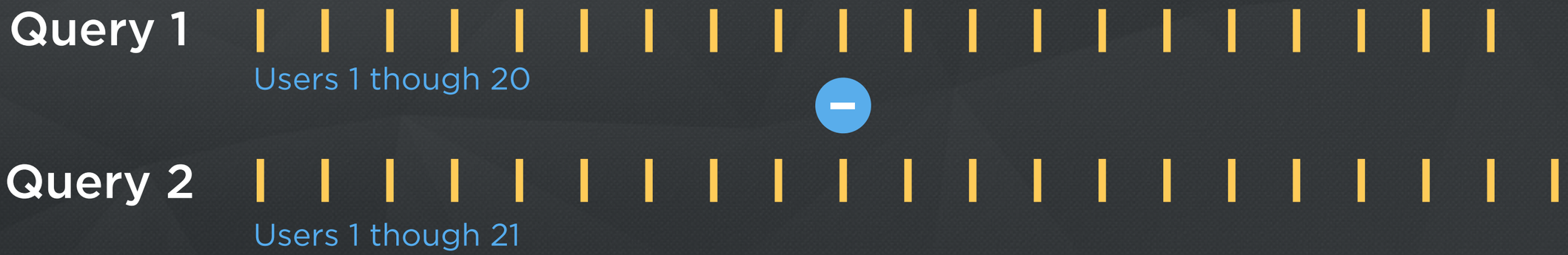
- MinHash for sampling users in a segment
- HyperLogLog for estimating segment cardinality
- Segment samples de-duped in any order
- Weighted selection of segment samples for audience sample

User Privacy

Don’t Forget About User Privacy

- Make it extremely difficult to reverse engineer a user’s demographics
- Attack vectors
  - Set balancing
  - Side channel
  - Homogenous group search

Set Balancing (Simple)



Is information leaked?

flight

Set Balancing (Simple)

Is information leaked?

flight

Set Balancing (Simple)

User 21

Is information leaked?

Yes! Q2 - Q1 leaks info about User 21

flight

Set Balancing (Complex)

Execute a series of queries

Query 1

Users 1 though 10

Query 2

User 1 plus users 11 through 20

Query 3

Odd users between 1 and 20

Query 4

Even users between 1 and 20

Is information leaked?

flight

Set Balancing (Complex)

Execute a series of queries

Query 1

Users 1 though 10

Query 2

User 1 plus users 11 through 20

Query 3

Odd users between 1 and 20

Query 4

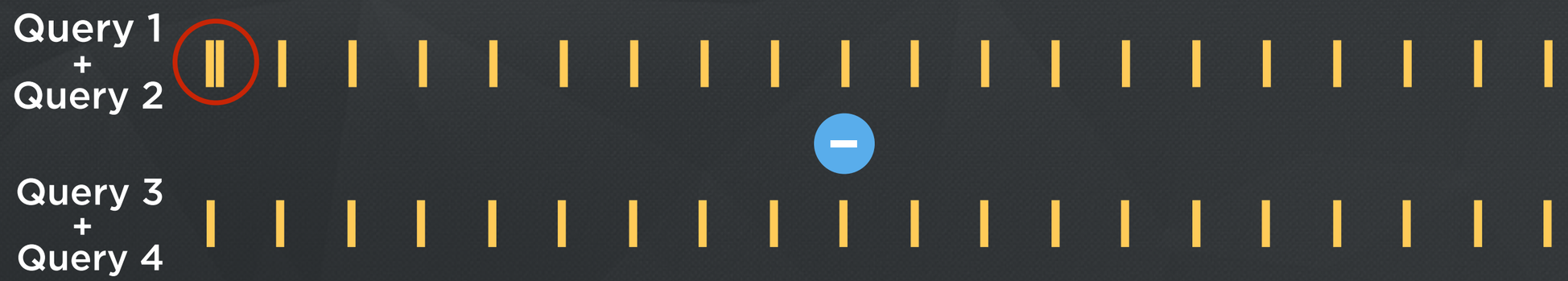
Even users between 1 and 20

Is information leaked?

flight

Set Balancing (Complex)

Execute a series of queries



Is information leaked?

flight

Set Balancing (Complex)

Execute a series of queries

Is information leaked?

Yes! Q1 + Q2 - Q3 - Q4 leaks info about User 1

flight

Mitigating Attacks

- Minimum audience size
- Random sampling
- Reporting thresholds
- Query auditing

Architectural Highlights

- Valuable insights while protecting user privacy
- Small storage footprint
- Query latency does not increase with audience size

Thank You

@jmcarey