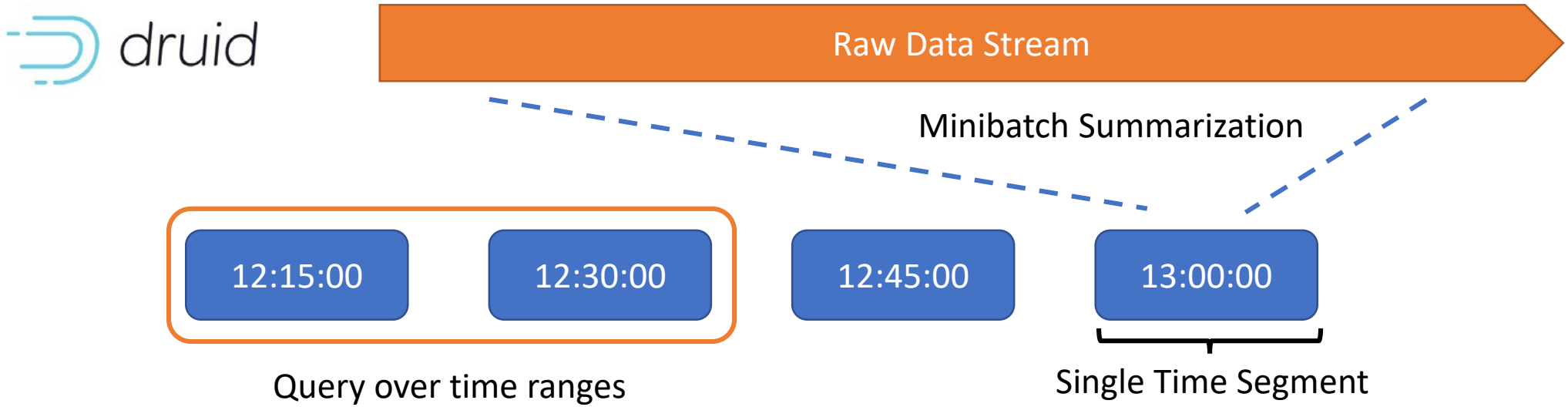
Data Summaries for Large Scale Aggregation

Edward Gan, Moses Charikar, Peter Bailis

Email: edgan8@gmail.com

Data summaries enable scalable queries



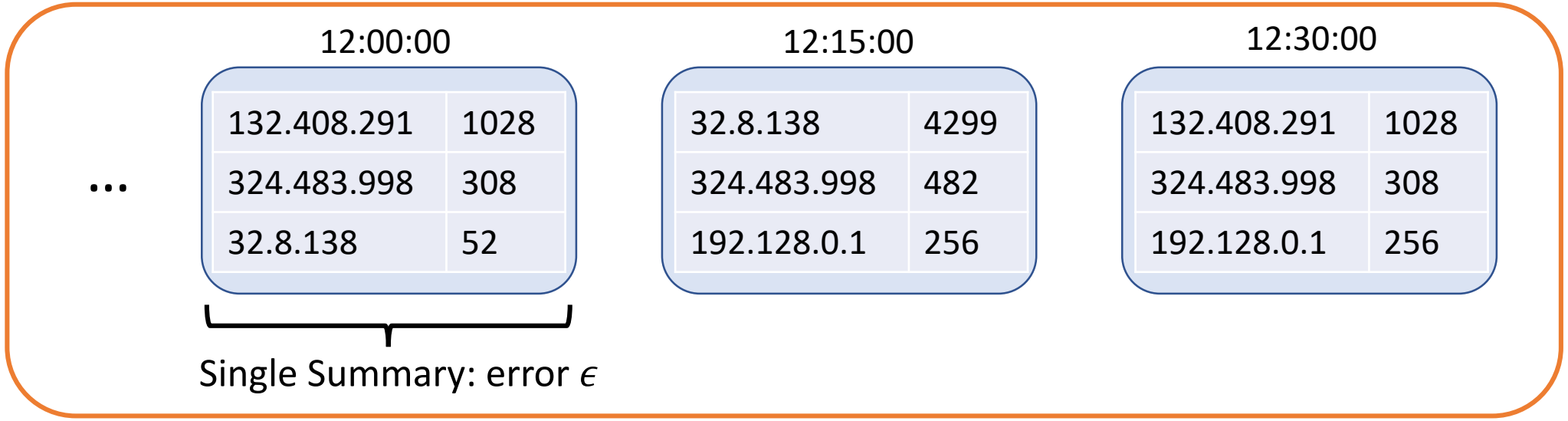
Summaries constructed for each data segment:

Counts, Sums, Samples, HyperLogLog, CountMin, etc…

Queries can be served directly using (approximate) summaries

Challenge: error accumulation

Query: Top 10 ip addresses by request count in last 7 hours?

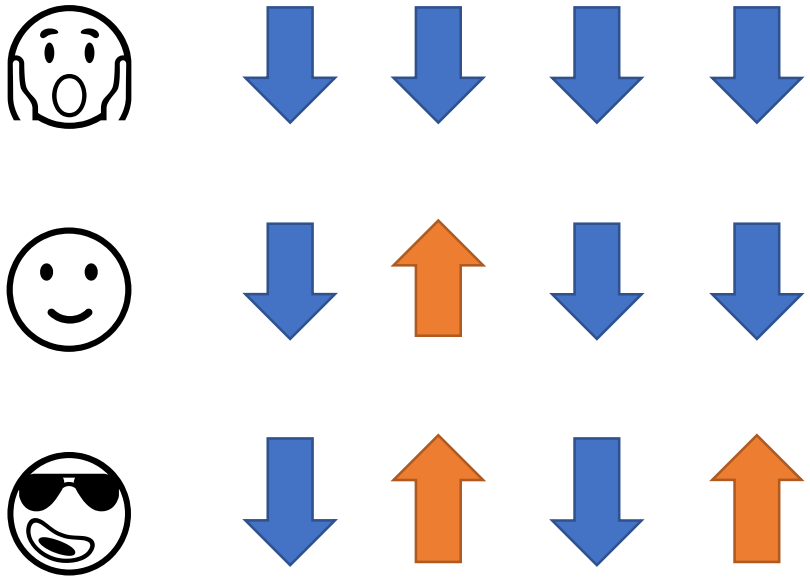


Aggregating k summaries: error $k\epsilon$

Query accuracy degrades *linearly* with aggregation

Opportunity: error cancellation

IP Address counts can be either overestimates or underestimates



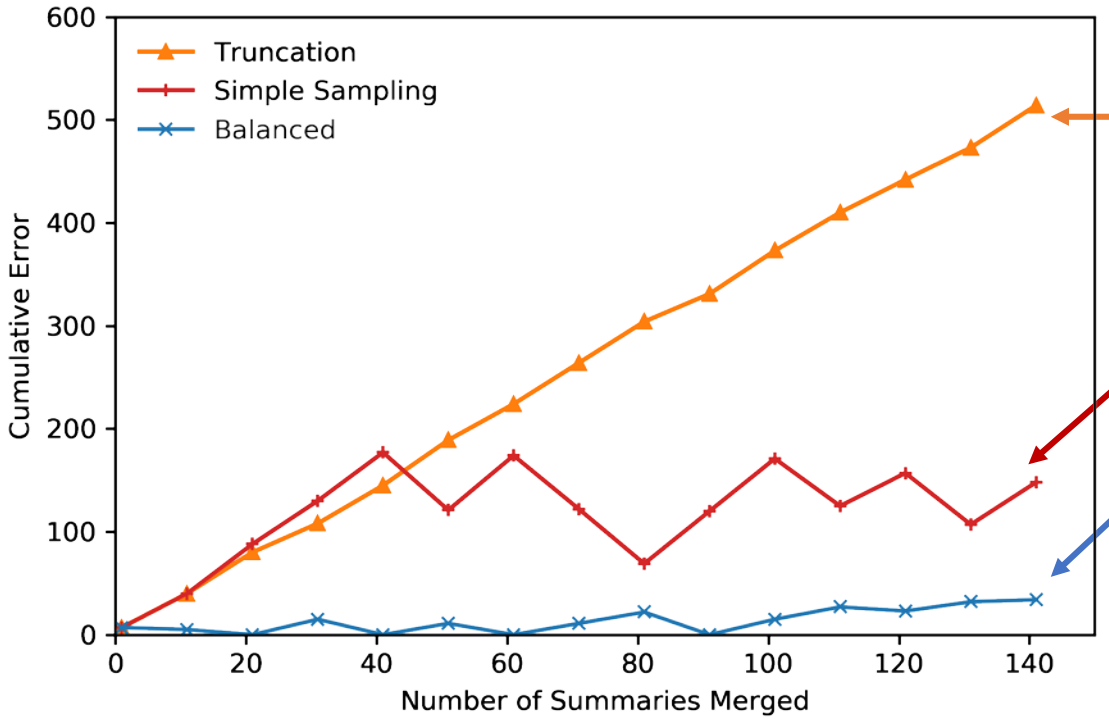
Consistent Bias

Independent Errors

Perfect Cancellation

Summaries for range aggregations

Time-series summaries aggregated over **k contiguous windows**



Truncate Top: $O\left(\frac{n}{s}k\right)$

Simple Sampling: $O\left(\frac{n}{\sqrt{s}}\sqrt{k}\right)$

Balanced: $O\left(\frac{n}{s}\log k\right)$

Balanced Summarization: bias summaries to cancel out errors of **previous** consecutive summaries.

Designing summaries for error cancellation

Problem: Summary approximation error grows with aggregation

Goal: Design summaries for **error cancellation** as an **ensemble**

- For contiguous ranges, error can be controlled incrementally
- Other **aggregation patterns**: 2d ranges, hierarchical, sliding window
- Other **queries:** quantiles, sums

Love to hear feedback and more use cases for data summaries!

Email: edgan8@gmail.com