Summary Statistics (SumStats framework)

What is SumStats?

- **Summary Statistics**
- From Wikipedia:
  - Summary statistics are used to summarize a set of observations, in order to communicate the largest amount as simply as possible

Motivations

- Load balancing made previous techniques all fail
  - SumStats framework hides cluster abstraction
- Better and more repeatable interface and approach for measurement and thresholding
- Give more people the ability to write real world deployable measurement scripts

Approach

- Discrete time slices (epochs)
- Only streaming algorithms allowed
- Every measurement must be merge-able for cluster support
- Probabilistic data structures
  - HyperLogLog & Top-K now

Why do any of this?

Measurement is fun!

SumStats Based Notices

200.29.31.26 had 349 failed logins on 2 FTP servers in 14m47s

92.253.122.14 scanned at least 29 unique hosts on port 445/tcp in 1m4s

88.124.212.10 scanned at least 41 unique hosts on port 445/tcp in 1m13s

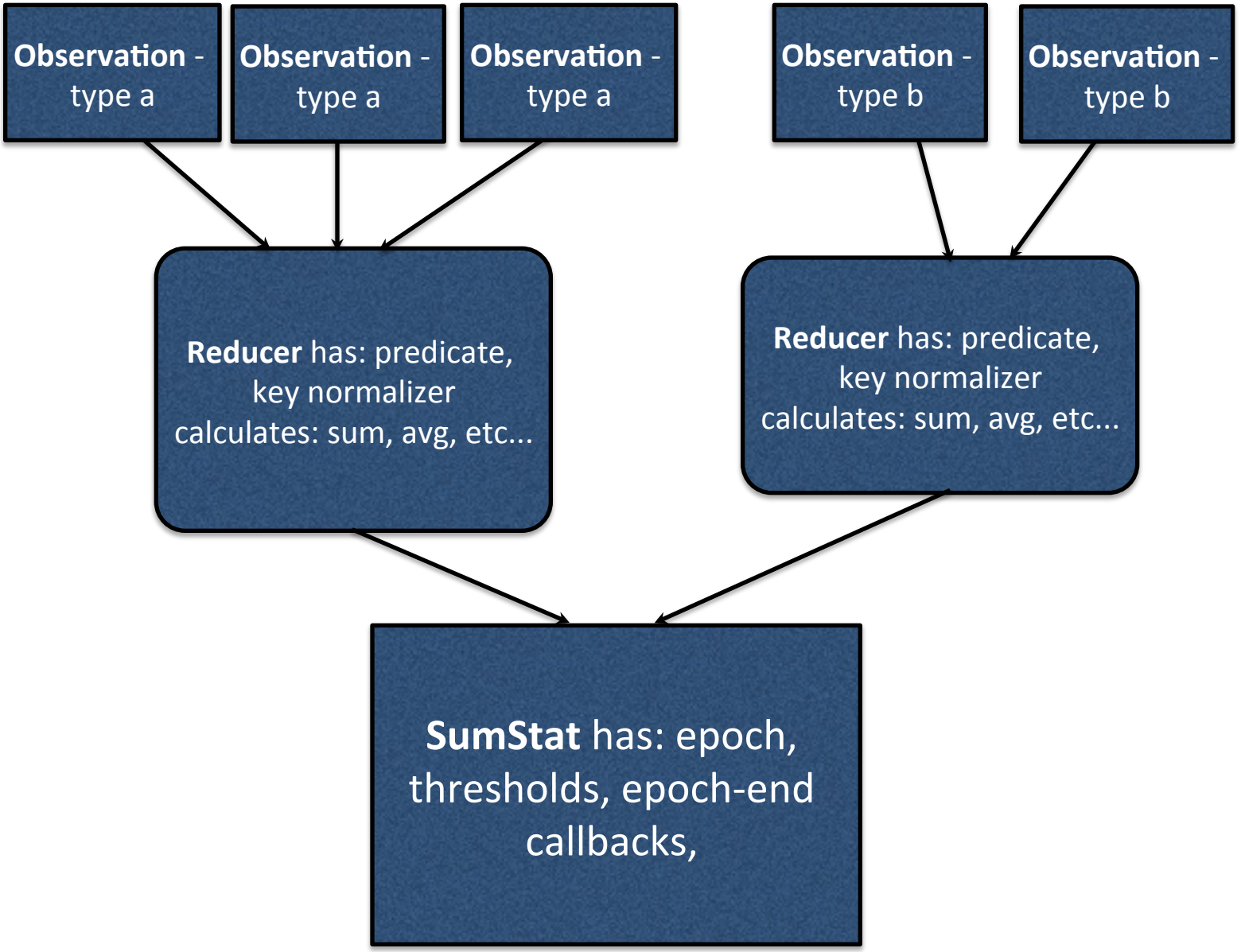
212.55.8.177 scanned at least 75 unique hosts on port 5900/tcp in 0m36s

200.30.130.101 scanned at least 66 unique hosts on port 445/tcp in 1m20s

107.22.92.186 scanned at least 64 unique hosts on port 443/tcp in 0m1s

5.254.140.123 scanned at least 29 unique hosts on port 102/tcp in 4m1s

122.211.164.196 scanned 15 unique ports of host 75.89.37.60 in 0m5s



...

Observations

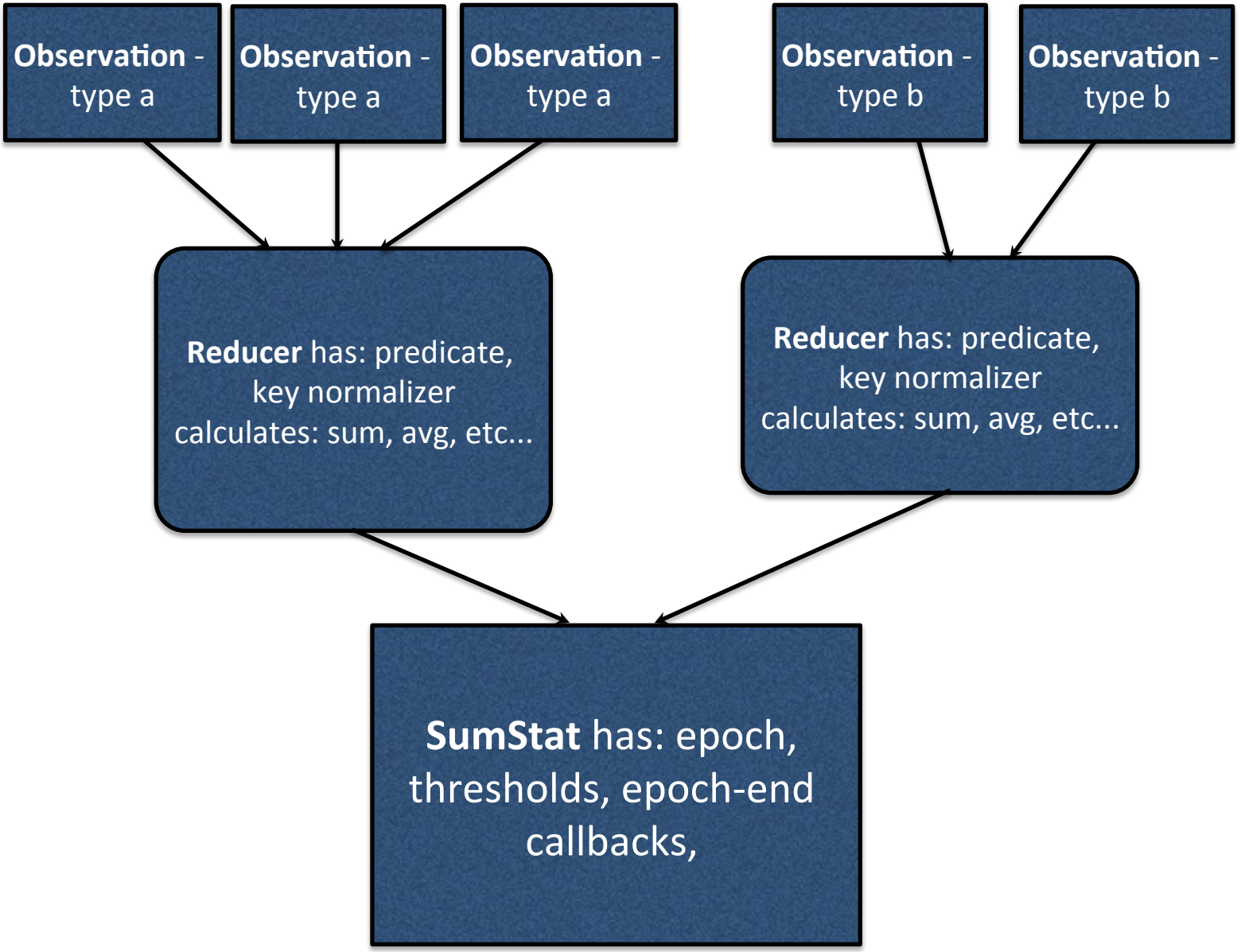
- Observations observe a single point of data
  - An HTTP request
  - A DNS lookup
  - An ICMP message

Reducers

- Reducers collect observations and apply calculations to them
  - Sum of Content-Length headers
  - Unique number of DNS requests

SumStat

- A SumStat collates multiple reducers
  - Set thresholds at ratios between reducers
    - e.g. Ratio of unique DNS requests and unique HOST names seen in HTTP traffic
  - Handle results from Reducers and do something


Observation - type a
Observation - type a
Observation - type a
Observation - type b
Observation - type b
Reducer has: predicate, key normalizer calculates: sum, avg, etc...
Reducer has: predicate, key normalizer calculates: sum, avg, etc...
SumStat has: epoch, thresholds, epoch-end callbacks,