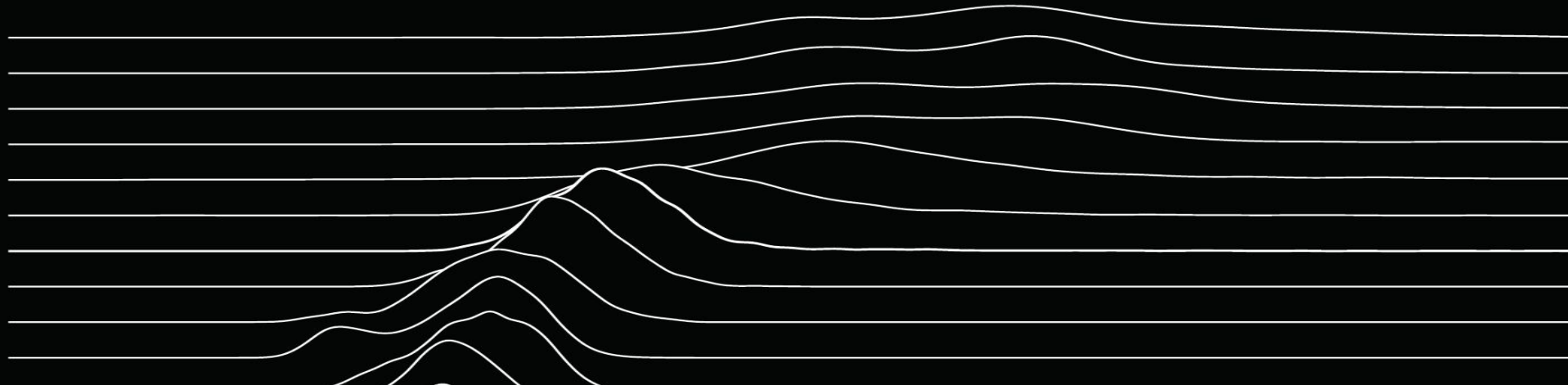
MLAB
10TH ANNIVERSARY
AUGUST 7-8, 2018

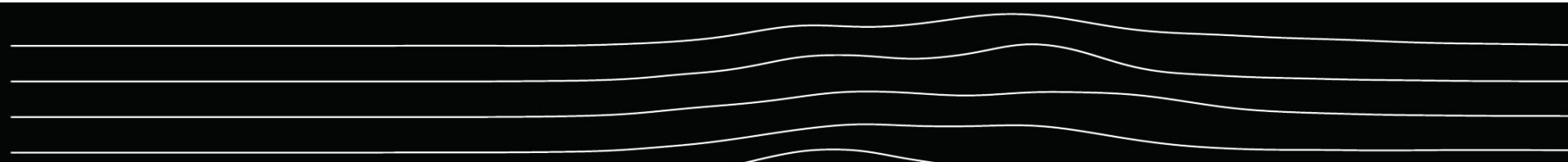
Research with M-Lab Data

Beacons in M-Lab data | Matt Mathis



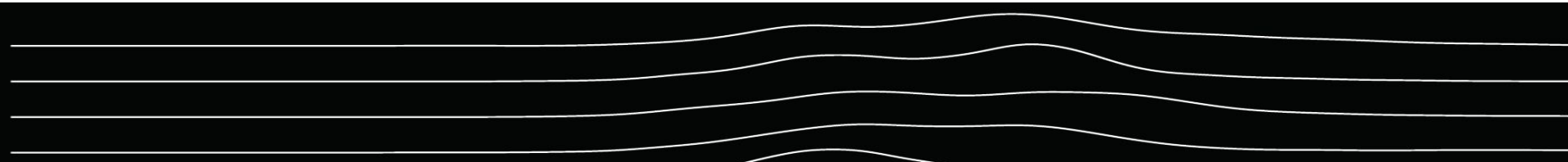
Outline

- Beacons: A gold standard for longitudinal studies
- Methodology
  - A rich and computationally efficient representation
- More results
- Epilog



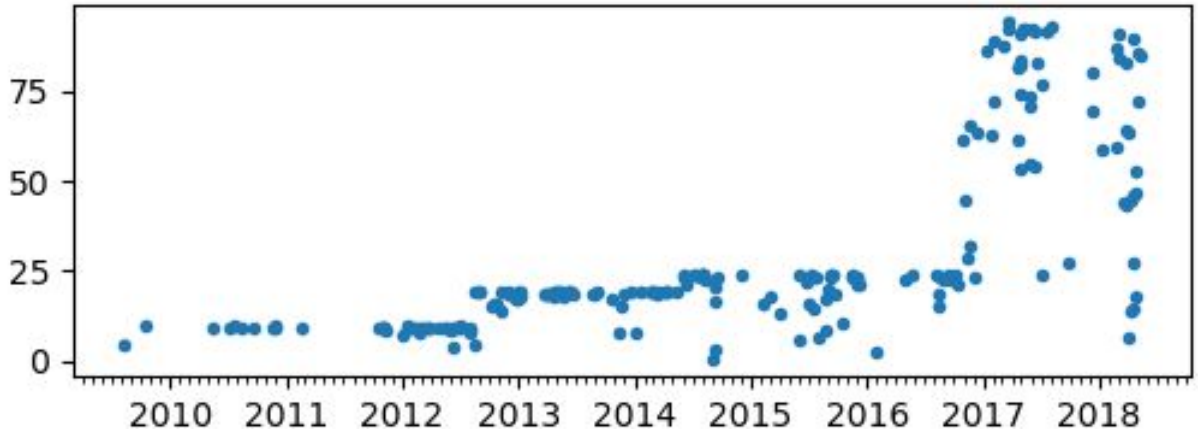
Part 1: Beacons

- A gold standard for longitudinal studies
- Single IP addresses that ran MLab tests for months or years
  - 1.5 M devices active for 1 more than 1 year
  - 600 k devices active for more than 2 years
  - 2000 devices active for more than 6 years
- Self calibrated measurement of network change



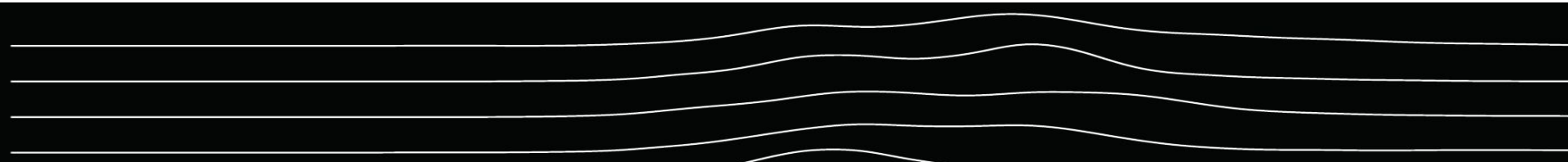
A typical(?) Beacon

- 600+ tests
- Stable IP for almost 10 years
- Small(?) ISP in East Europe
- Probably checking their upstream ISP



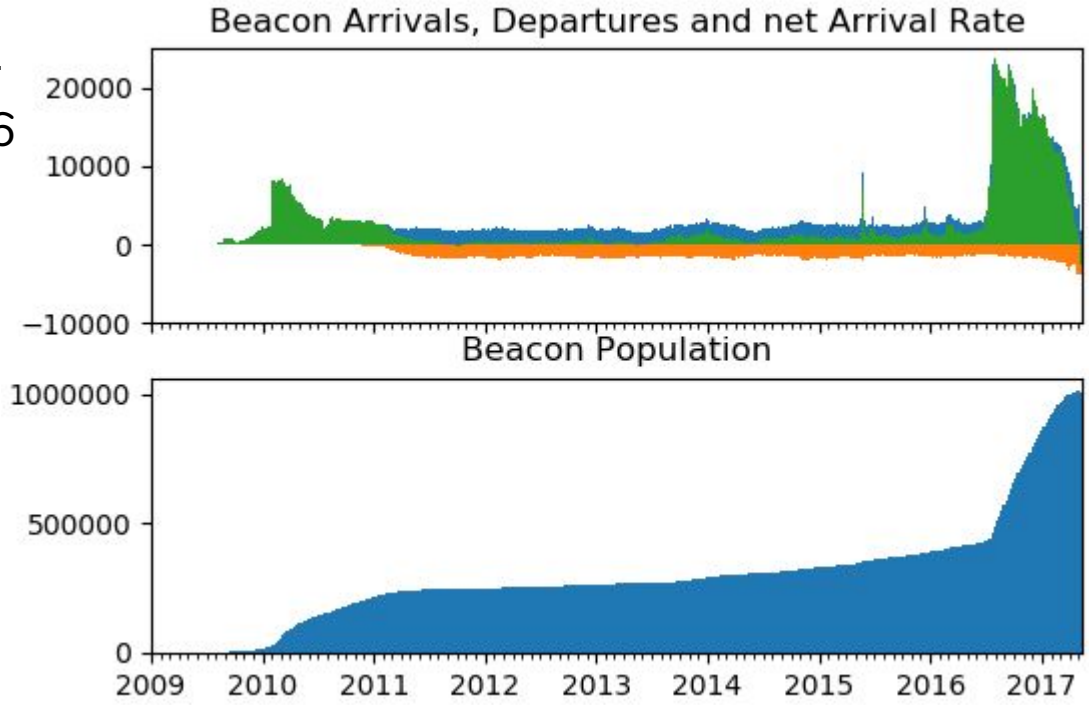
Reasons for beacons

- Small ISPs checking their own upstream connectivity
  - First observed in early 2009
  - Long stable IP addresses
- Autoconfig in applications or devices
  - Applications that need to measure their Intenet connectivity
    - e.g. BitTorrent
  - Likely to be subject to periodic IP reassignment
    - Which appear as non-overlapping sequential beacons
    - Does not affect basic longitudinal studies



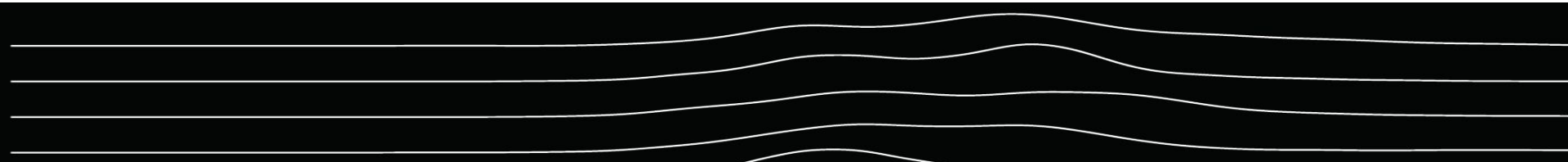
Beacons

- 2-400k through mid 2016
- Mid 2016
  - Google One Box
  - New embedded clients
- Todo: study IP reassignment



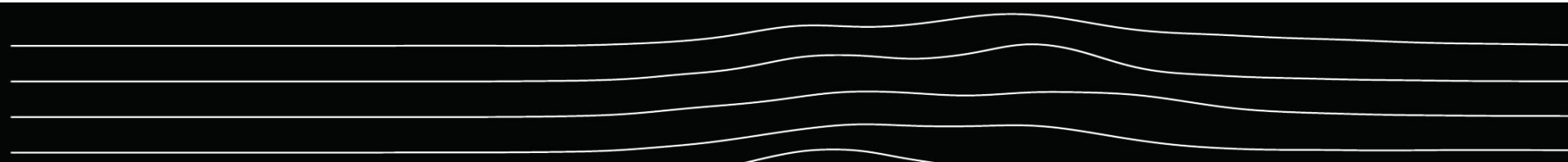
Methodology

- It started as a computational shortcut….
  … because it was quick and easy ….

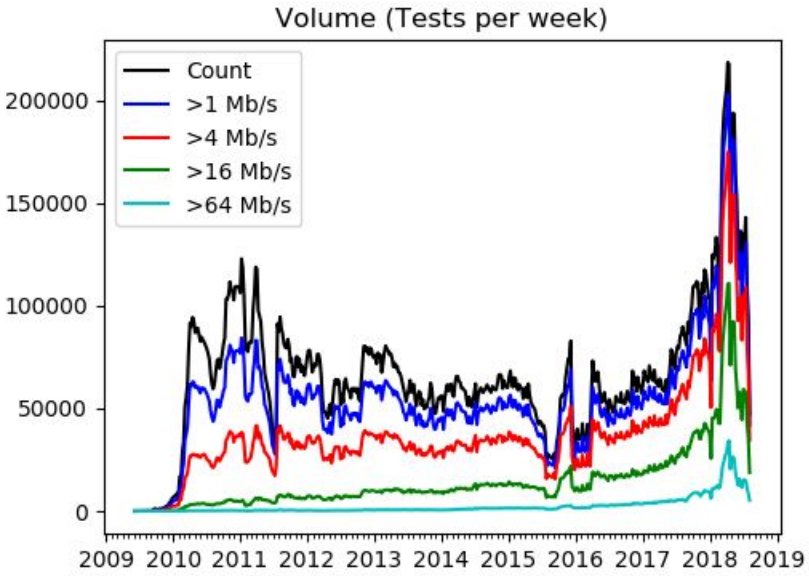


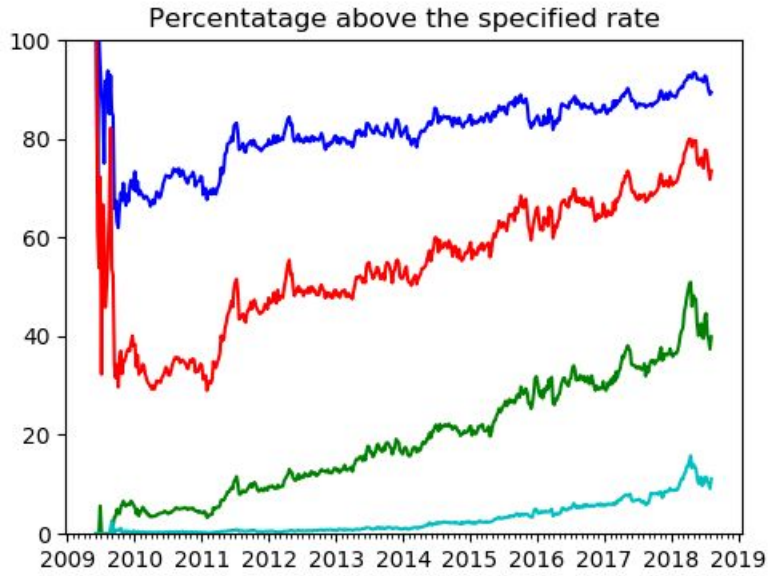
Methodology details

- Count the entire M-Lab corpus into multidimensional arrays
  - Tabulate 1.4 B rows into about 500k counters
  - Typical axises:
    - Test date or time
    - Selected M-Lab servers, pods or metros
    - Powers of 2 performance bins from 1 Mbit/s to 512 Mbit/s
    - .....
  - Extremely efficient in BigQuery (~40 seconds)
- Plotting phase aggregates (sums) bins
  - Collapses some of the dimensions



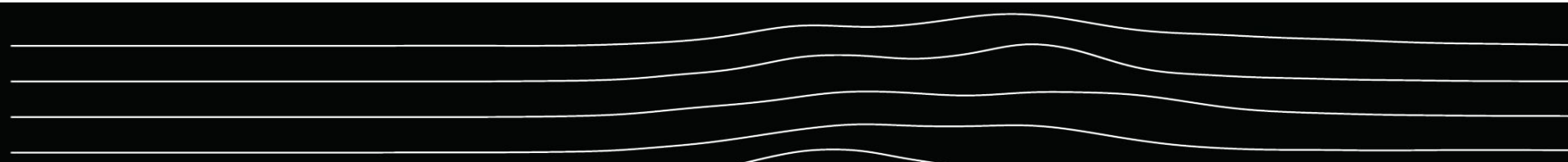
Europe revisited





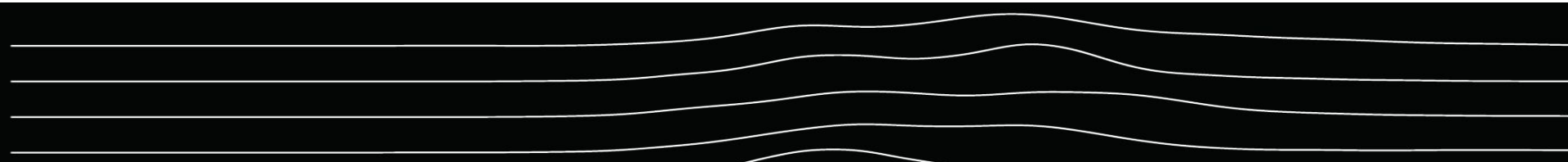
Directly infer user experience

- Some users experience can be noted directly from the graph
  - In 2010 about 30% of the users could run an application requiring 4 Mb/s
  - By 2018, that had risen to about 65%
  - Other performance levels (e.g. HD video) can be interpolated
- Contrast this to conventional summary statistics
  - Mean, variance, quartiles, percentiles, etc
  - None easily predicts if users are happy



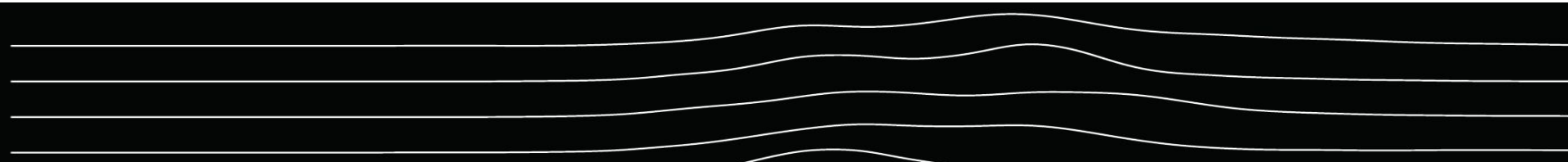
Algebra on metrics

- Arrays of counters can be added or compared
  - e.g. Compute US statistics by subtracting Canada and Mexico from North America
  - Dynamically aggregate small geographical areas into larger areas
  - Or to ask the net numbers of people who are better off
- This might have a profound impact on policy conversations
  - Recent strong encouragement on this point
  - Seeking opportunities to collaborate
- Again, all of these are nearly impossible with conventional summary statistics



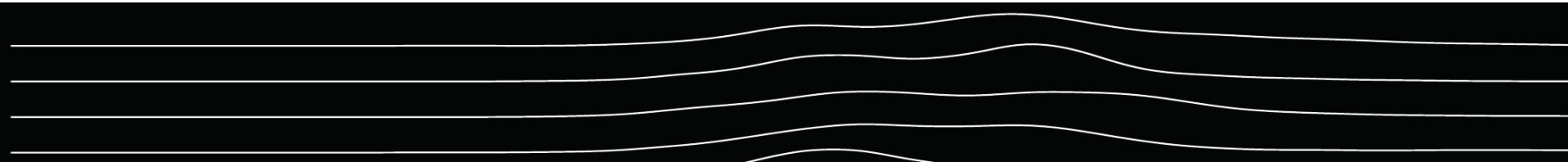
Side discussion: The need for algebra

- For most metrics (e.g. milk fat) they can be predicted from other measures
  - e.g. Mixing equal parts 4% and skim milk yields 2% milk
  - Also under pins properties such as vantage independence, repeatability, etc
  - Similar concepts apply to nearly all metrics
  - Implicitly provides ways to cross check other people's measurements
- But not Internet performance
- RFC 2330 [1998] positis an "Analytical Framework" for Internet metrics
  - Twenty years later, this still remains a dream
  - Can not predict performance from any other metrics
- However arrays of counts might be able to predict many-to-many

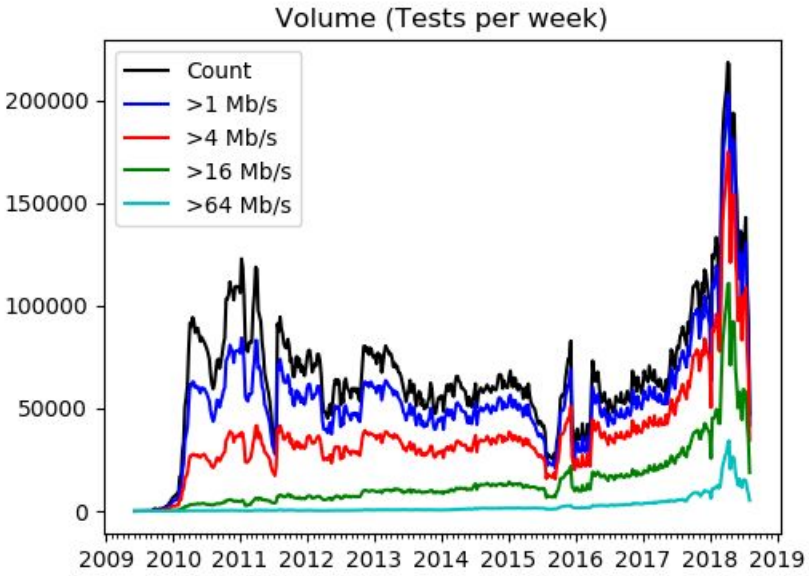


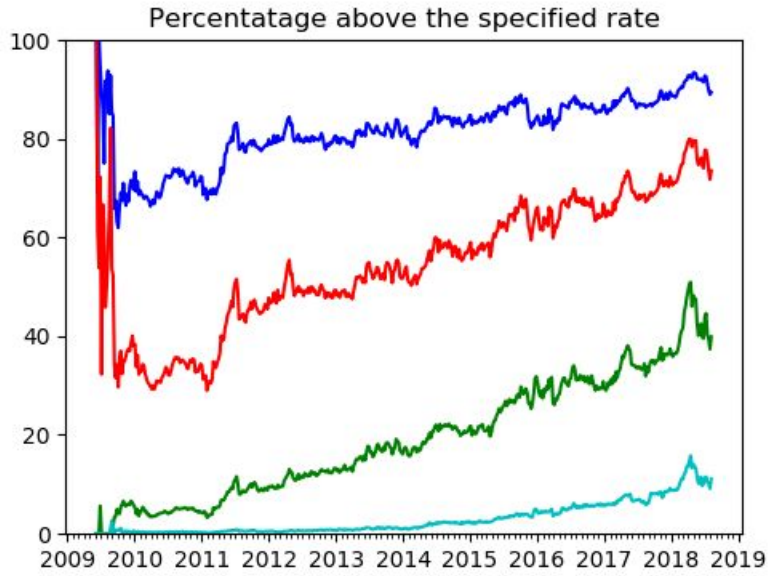
More results: The good, bad and ugly

-
-
-

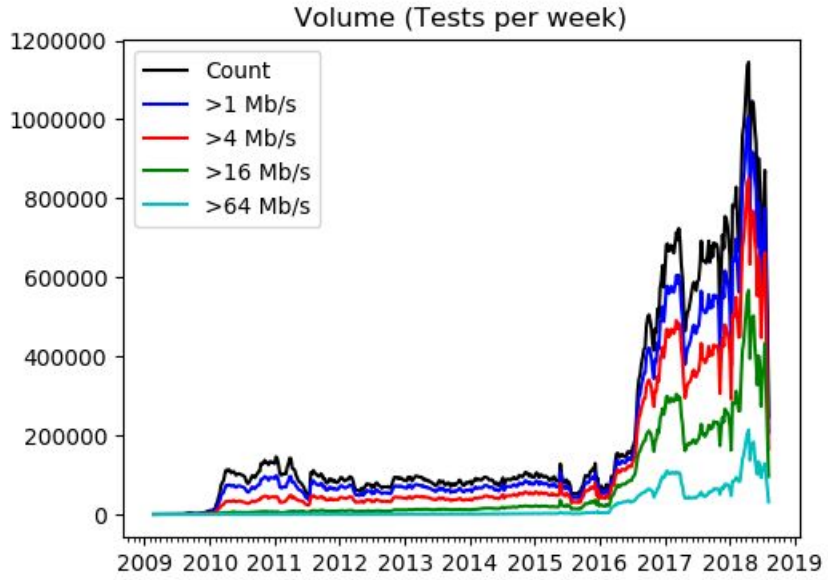


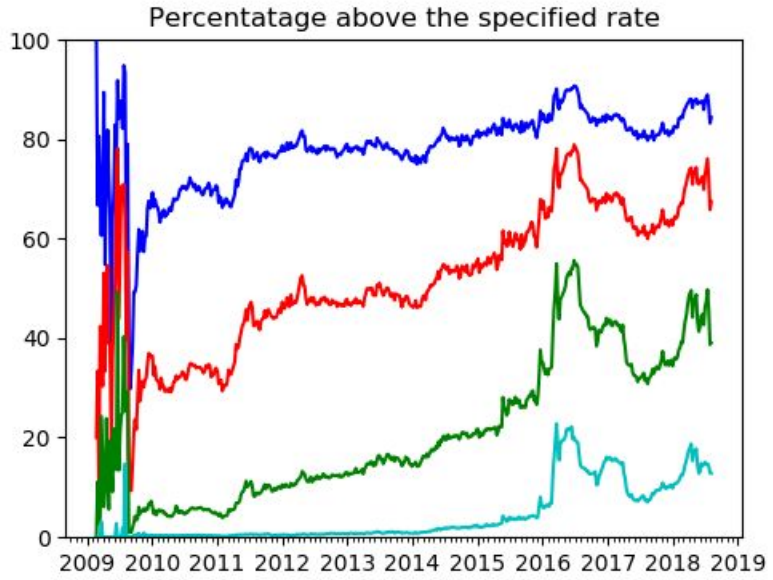
Europe revisited



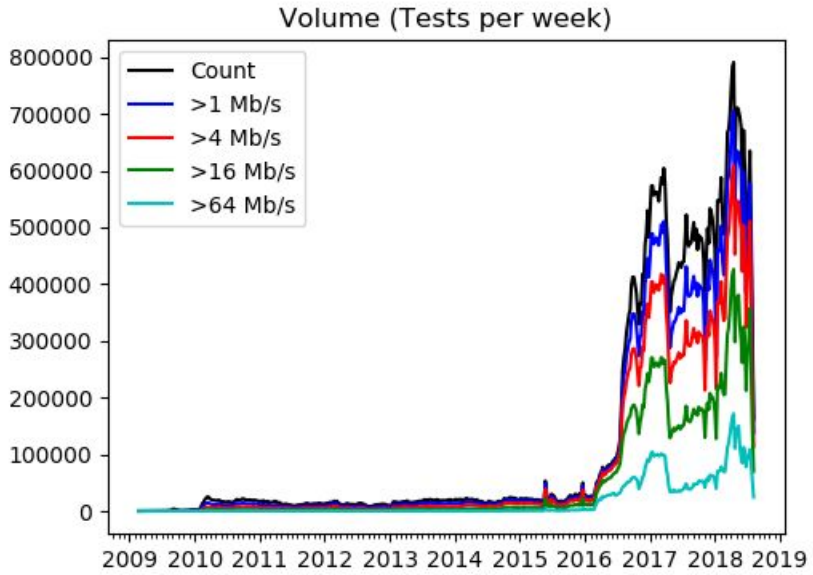


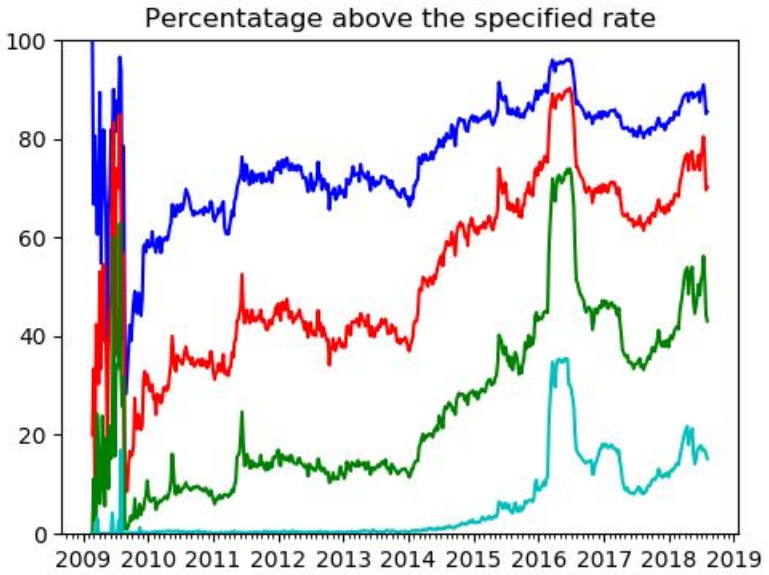
Global Internet Performance



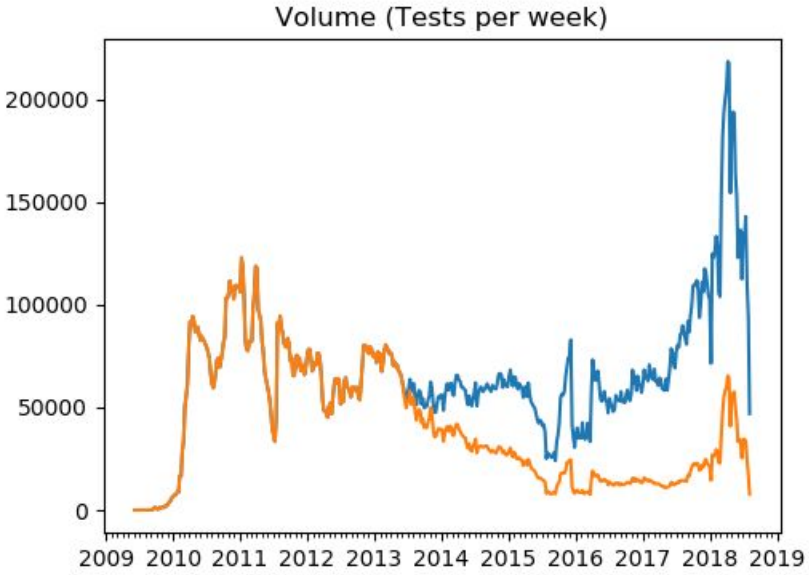


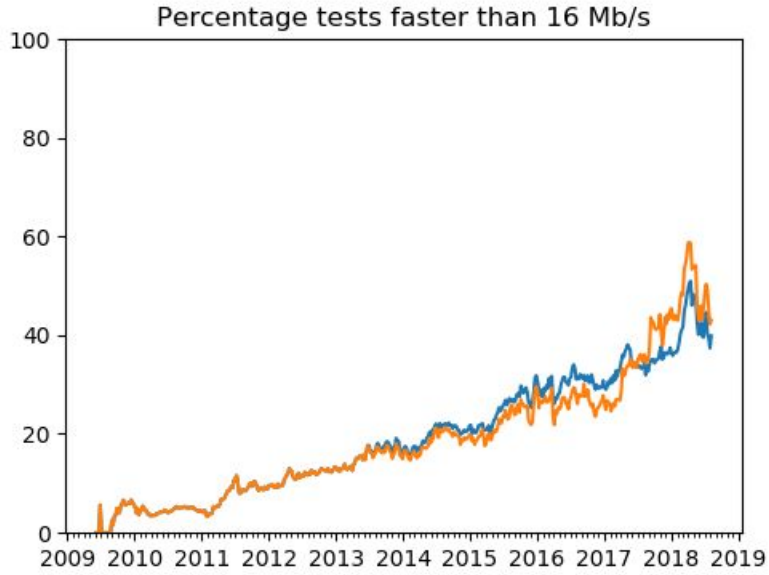
North America





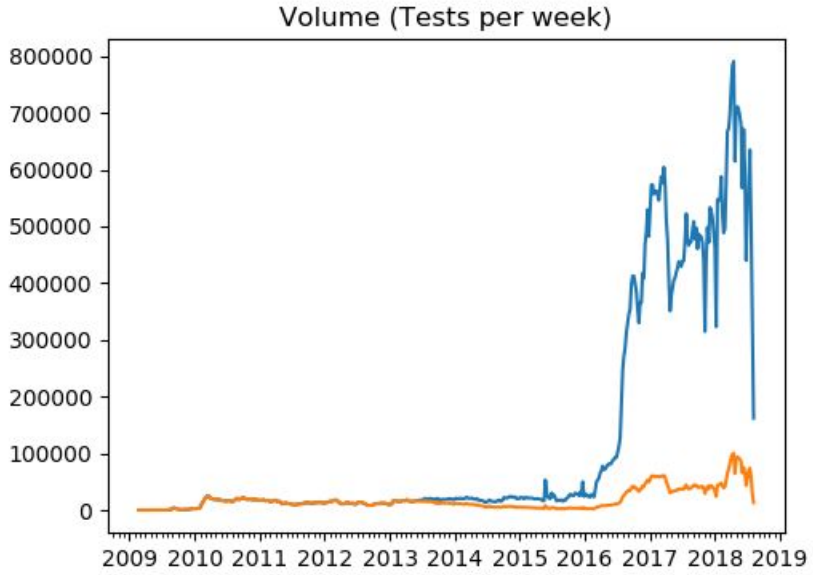
Europe Cohorts (all, before 7/2014)

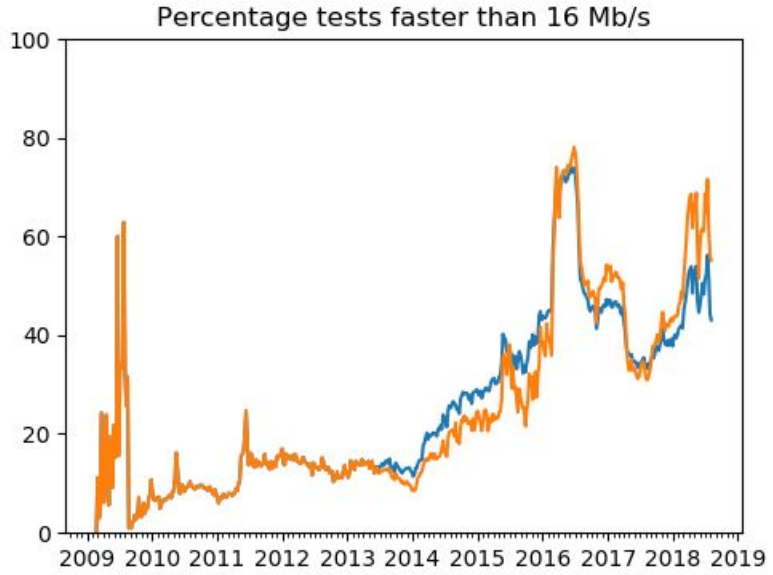




North American Cohorts (all, before 7/2013)

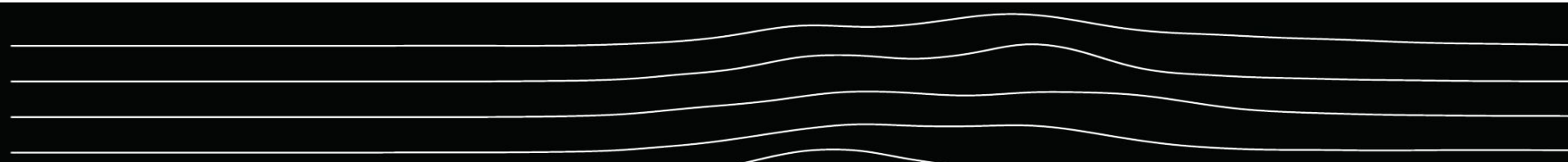
MLAB





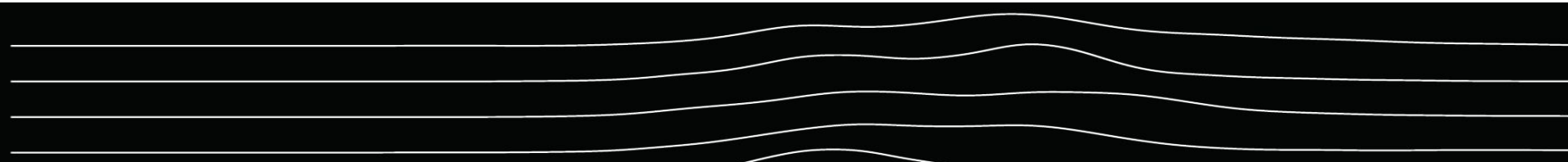
Epilog

- Things I learned on the way...



Observations about beacons

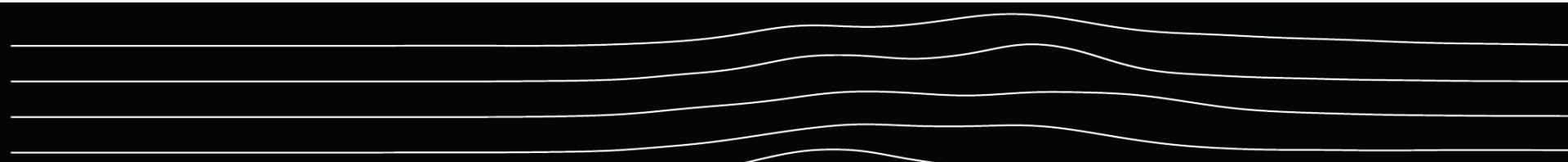
- Repeated tests from single clients ("A Beacon") help a lot.
  - Can compare beacons: Why are outliers different?
  - Beacons that share patterns share properties
    - Properties specific to beacons help identify "Beacon Swarms"
      - Beacon swarms that share code and or deployment



More observations...

- Beacons mostly eliminate the hard problems
  - Bias due to irregular testing
    - Each beacon is "self calibrated"
    - This property is preserved in aggregate, even if not individually identified
  - There is path to understanding self selection bias

    By fingerprinting "swarms of beacons", and comparing different swarms
    - Are the users representative? Does it matter?
    - Is the test schedule representative? Does it matter?
    - Is testing correlated with network problems? ("testing in anger")



Some observations about big data

- With enough data, extremely subtle patterns are exposed
  - In particular, any changes to the network appear in the data
  - When looking for changes, the data is self calibrated
- Deliberate manipulation is hard
  - Any one source is a minority of the data, deliberate manipulation causes it to look different than other data