WOMBIR-2
Industry / Government / Academia Collaboration

Matt Calder, Microsoft / Columbia
Avi Freedman, Kentik

Topics

- Matt
  - Alternative Industry / Academia Partnership Models
  - Don’t be afraid to ask!
- Avi
  - Industry data sharing/analysis constraints
  - Getting started with federated analysis - proposal
  - Getting started - call to action

Alternative Industry / Academia Partnership Models

- Why care about Industry?
  - View of the Internet
    - Lots of PoPs, Peers
    - Large client populations
  - Large systems
  - Lots of data
- But access can be sensitive
- Current gap bridged by students; guarded by lawyers
  - Short-term internships

Alternative Industry / Academia Partnership Models

- Ask: NSF to facilitate the creation of joint industry-academic positions
- Why NSF? This is an education and science problem
  - Binary academic / industry choice
  - Future scientists deprived of data and training
- Candidates?
  - Scientists on engineering teams
- Why would anyone do this?
  - Research opportunity
  - Like mentoring / working with students

Alternative Industry / Academia Partnership Models

- Pros for Industry:
  - Better recruiting. Direct and personal relationships with students over years, not months.
  - Network operations is hard; Improvements underfunded
  - Good PR. Helping NSF foster the next generation of scientists
- Pros for Academia:
  - More mentoring for students
  - Medium-long term collaborations better serves students
  - Diverse perspective on problems, faculty and PhD candidates
  - Access to data/validation

Don’t be afraid to ask!

- (Some) networks and ops are more transparent recently
- Content provider and CDN space
  - 10 years ago everything was secret
  - Cloud: now everything is a selling point
  - ZMap + Cloud VMs: difficult to hide
- This trend is going to continue
  - Peers and capacity?
- Reach out for feedback and validate your results!

Don’t be afraid to ask! Challenges

- Responsiveness
- Assumptions about what is sensitive

- Finding the right person
- How to ask
  - Rocketfuel survey
  - FP/FN makes things easier to answer
  - Send scripts
- Last minute asks
  - Get feedback early: “We are thinking of doing this…”
  - Validation should be end-result

Topics

- Matt
  - Alternative Industry / Academia Partnership Models
  - Don’t be afraid to ask!
- Avi
  - Industry data sharing/analysis constraints
  - Getting started with federated analysis - proposal
  - Getting started - call to action

Industry Data Sharing/Access Constraints

Potential issues:

- Legal (MSA/NDA)
- Regulatory (GDPC/CCPA)
- Reputational (direct and indirect)
- Summary vs. detail ->
  is archive + code enough if you
  can’t see the data?
- Need to review
- Obfuscation level

… + who has access today internally

Industry Data Sharing/Access Constraints

Potential issues:

- Legal (MSA/NDA)
- Regulatory (GDPC/CCPA)
- Reputational (direct and indirect)
- Summary vs. detail -> is archive + code enough if you can't see the data?
- Need to review
- Obfuscation level

… + who has access today internally

Why do we care to share?

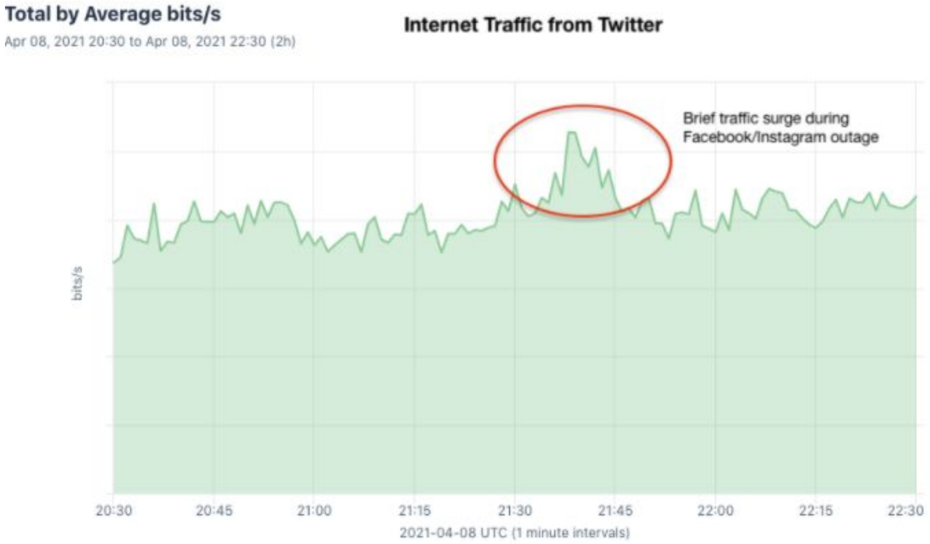
- Want to help (particularly re: traffic weightings)
- Marketing for customers
- Marketing for hiring

So what do we do (Kentik, Akamai)?

- Use judgement
- Not scientifically

The Good News: We Can Share

3/10/21 (Twitter to talk about Instagram)



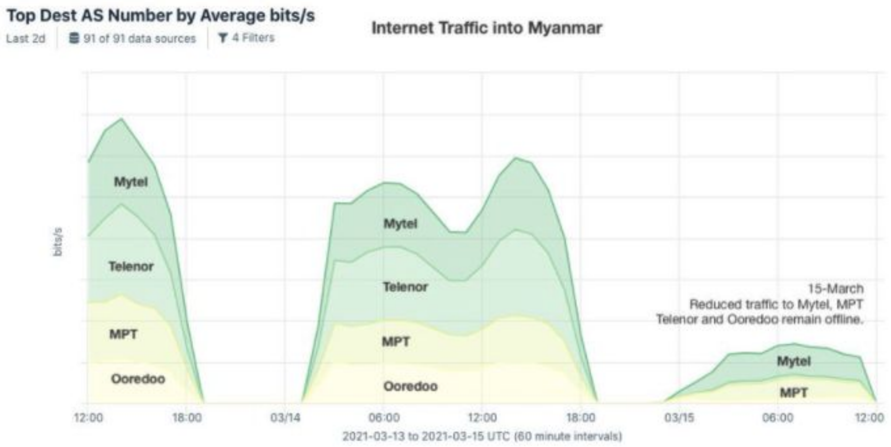
Total by Average bits/s

Apr 08, 2021 20:30 to Apr 08, 2021 22:30 (2h)

Internet Traffic from Twitter

Brief traffic surge during Facebook/Instagram outage

3/15/21 (Myanmar mobile)



Top Dest AS Number by Average bits/s

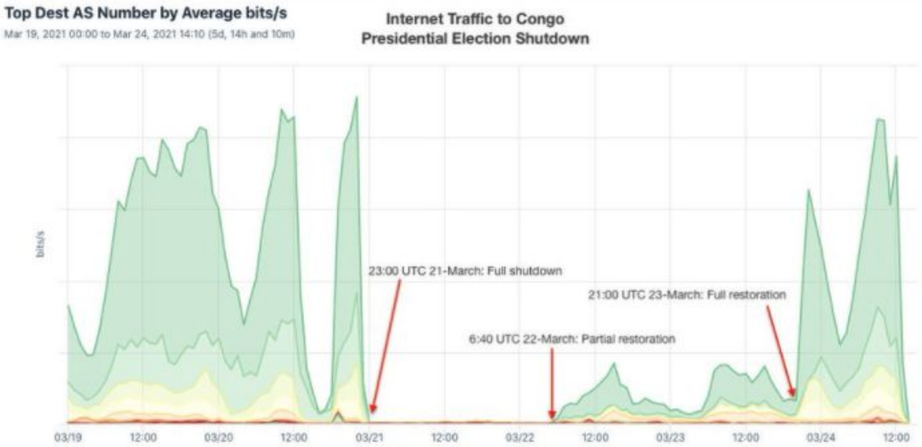
Last 2d | 91 of 91 data sources | 4 Filters

Internet Traffic into Myanmar

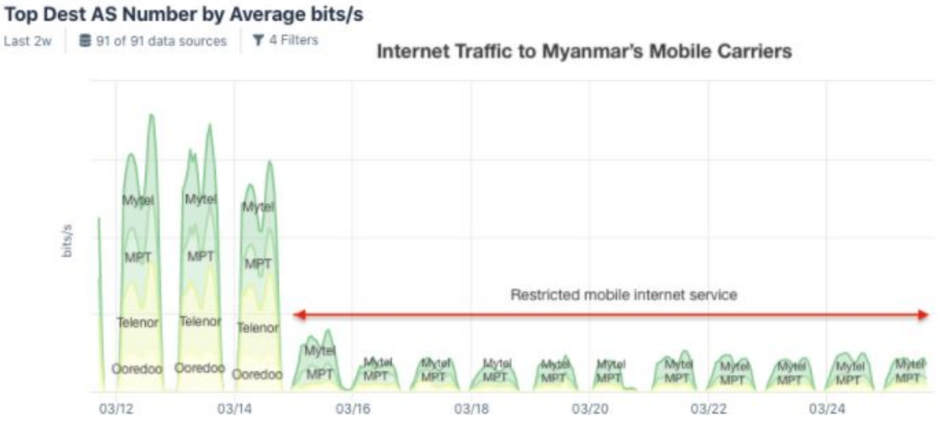
15-March
Reduced traffic to Mytel, MPT
Telenor and Ooredoo remain offline.

The Good News: We Can Share

3/24/21 (Congo Shutdown)



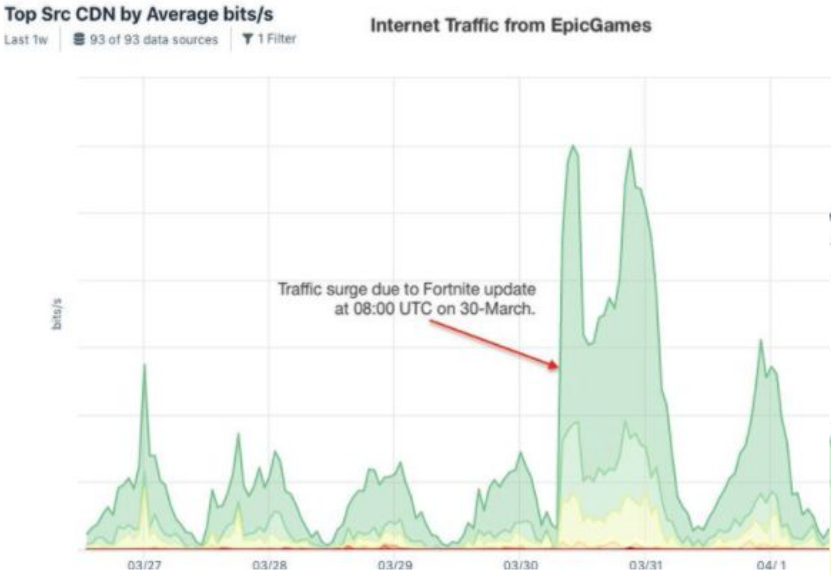
3/25/21 (Myanmar)



The Good News: We Can Share

4/2/21 (Gaming Update)

4/9/21 (Gaming Update)

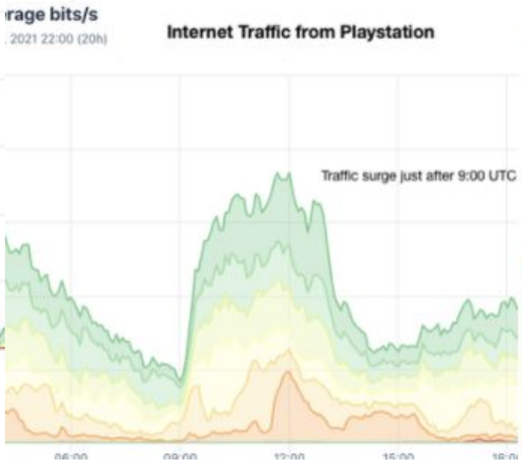


Top Src CDN by Average bits/s

Last 1w | 93 of 93 data sources | 1 Filter

Internet Traffic from EpicGames

Traffic surge due to Fortnite update at 08:00 UTC on 30-March.

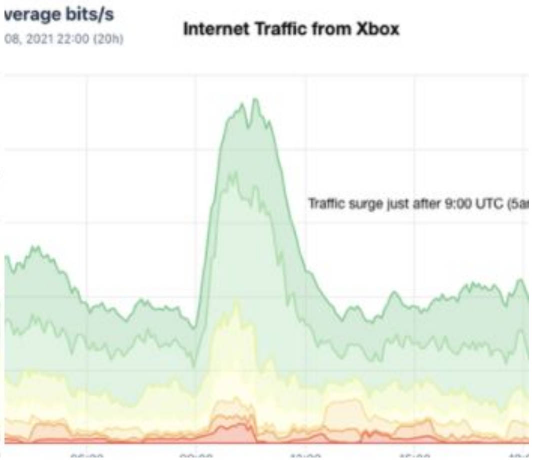


rage bits/s

2021 22:00 (20h)

Internet Traffic from Playstation

Traffic surge just after 9:00 UTC



verage bits/s

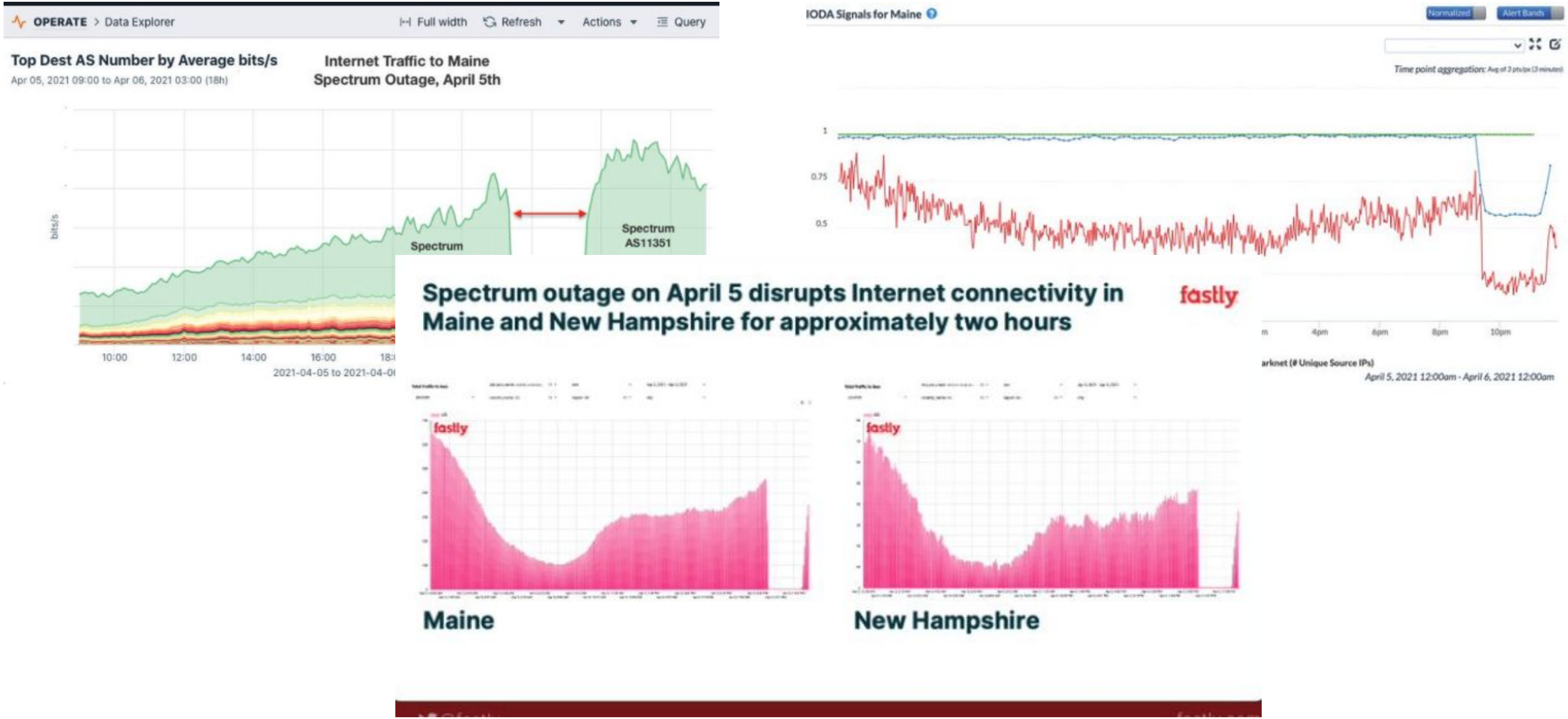
08, 2021 22:00 (20h)

Internet Traffic from Xbox

Traffic surge just after 9:00 UTC (5a

The Good News: Many Can Share

4/5/21 (SP Outage)



OPERATE > Data Explorer

Top Dest AS Number by Average bits/s

Apr 05, 2021 09:00 to Apr 06, 2021 03:00 (18h)

Internet Traffic to Maine
Spectrum Outage, April 5th

IODA Signals for Maine

Time point aggregation: Avg of 3 pts/px (3 minutes)

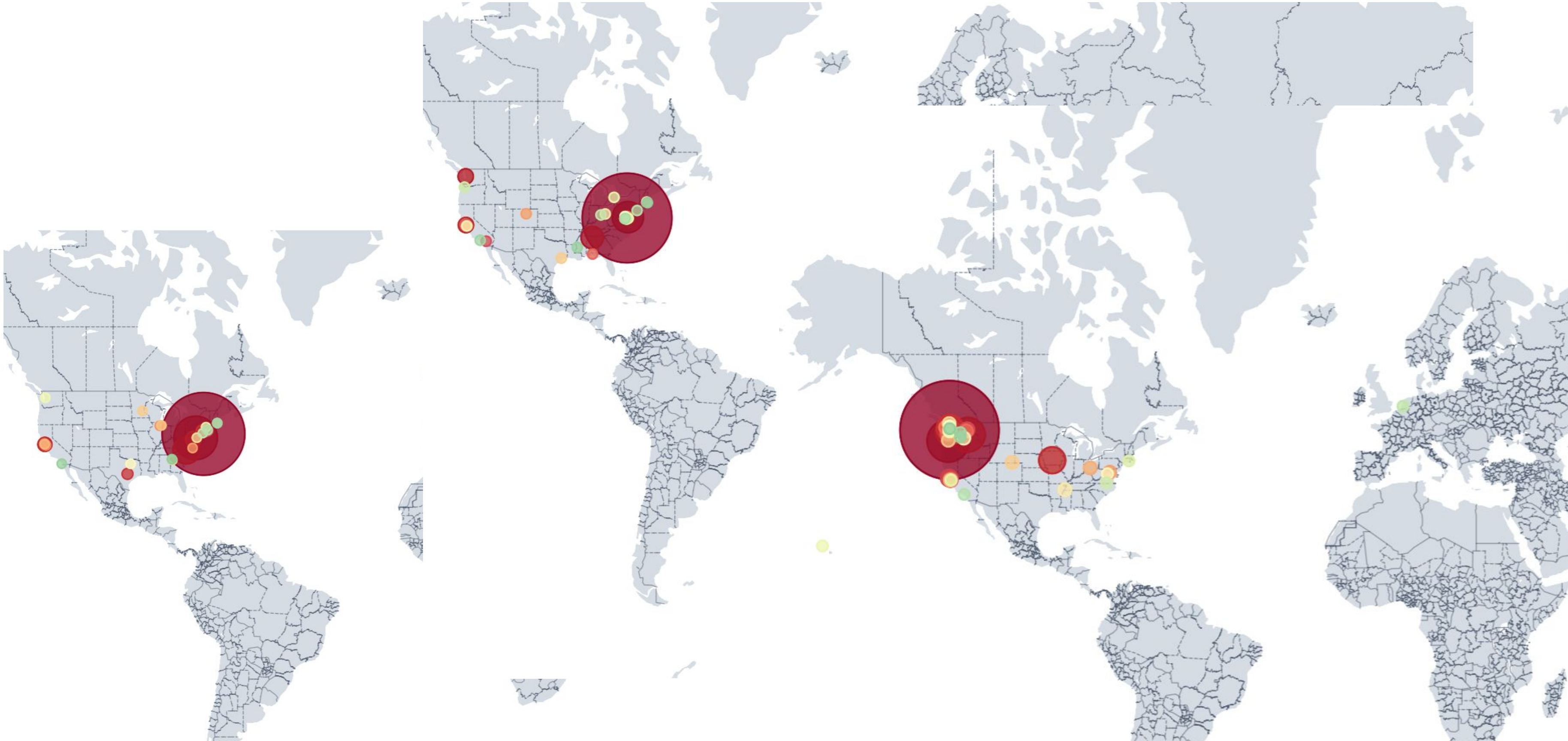
April 5, 2021 12:00am - April 6, 2021 12:00am

Spectrum outage on April 5 disrupts Internet connectivity in Maine and New Hampshire for approximately two hours

Maine

New Hampshire

By City… Maybe Not (3 Customer Sets)



And Even More by ASN+City

Getting started with federated analysis: Proposal + Going 0->1

(But not to be doom and gloom!)

What if we had:

- A way of describing telemetry, features of telemetry, and computations to run over features?
- A way of running those specs on data?
- That multiple parties could run?
- And could be published and re-used/tested by others?
- Maybe even that data holders could be incented to archive and allow re-computation over?
- I say feature because it could evolve to model training over data without access

Getting started with federated analysis: Proposal

(But not to be doom and gloom!)

What if we had:

- A way of describing telemetry, features of telemetry, and computations to run over features?
- A way of running those specs on data?
- That multiple parties could run?
- And could be published and re-used/tested by others?
- Maybe even that data holders could be incented to archive and allow re-computation over?
- I say feature because it could evolve to model training over data without access

A proposal:

- We pick some simple questions
- Find multiple parties with data
- Assume human review of input and output
- Run the computations and archive the features used (raw data too if possible, even if not sharable)
- Let's do this collaboratively for 1 or 2 questions

Problems

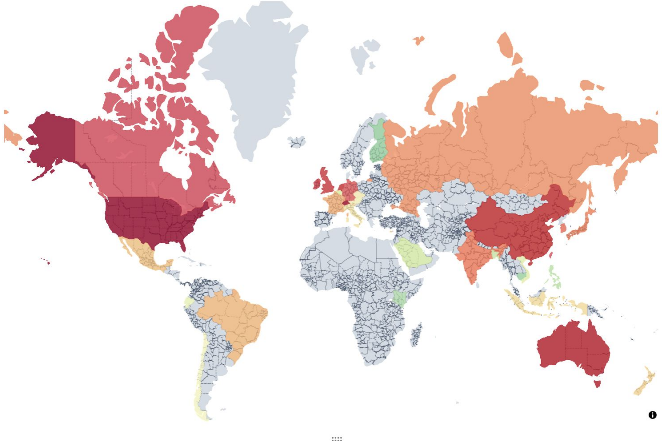
- Generally focused on weighting but open

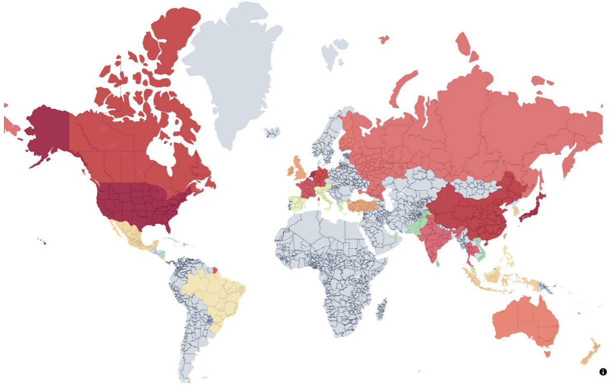
Existence Proof / Starter: Dst Geo/Country

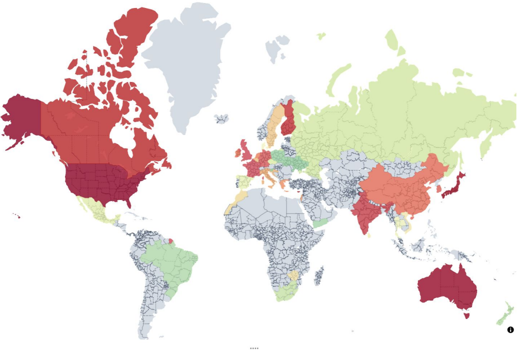
Example (fair bit of spec incompleteness/hand waving here):

- Feature: lookup(dstip, github/foo/maxmin-free-2021-04-05, country)
- Bucket: 5 minute width, simple sum
- Computation: avg of buckets
- Time period: 2021-04-15-00:00:00 UTC to 2021-04-15-00:01:00:00 UTC
- Known problems
  - Describing source
    (sFlow different than NetFlow/Juniper MX diff than NetFlow/ASR9K …)
  - Many more

Results: (Same) 3 Customer Sets, by Country







Challenge… Anyone Want to Join + Try?

- Ideally at least a few others with data
- Then we
  - Iterate to feature, computation, and data source descriptions
  - And whatever else we find is needed!
- Maybe on-campus (not just industry!)
- What questions?
  - Including not only traffic data

Challenge… Anyone Want to Join + Try?

- Ideally at least a few others with data
- Then we
  - Iterate to feature, computation, and data source descriptions
  - And whatever else we find is needed!
- Maybe on-campus (not just industry!)
- What questions?
  - Including not only traffic data

Or Flame?

- Can people use results over data they can’t ‘have’?
- Is it science if you can’t have the data to validate yourself and/or publish?