No User Left Behind

Mohit Suley, Bing Live Site Engineering / mosuley@microsoft.com

A few months ago in a country far away...

...

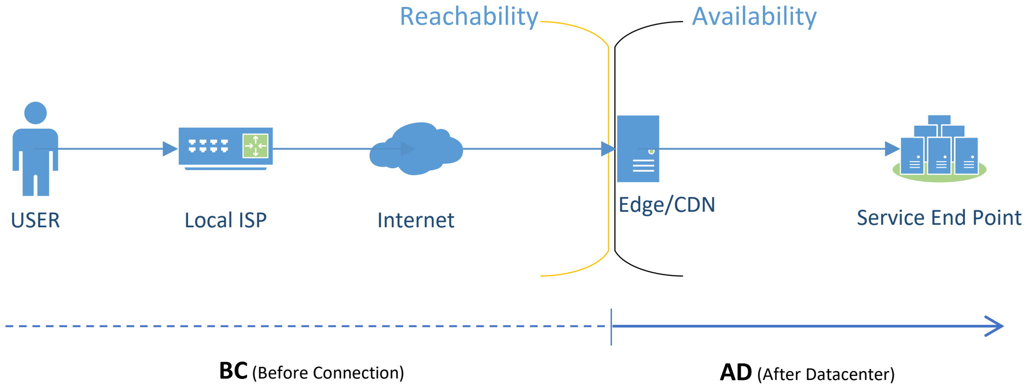
What do we (as engineers) want from Monitoring?

Ensure users get the experience we intended for them to see

Available
Reachable

Fast
High Quality

Reachability



Reachability can be explained as a measure of successful *network* path connectivity between the user and the service end-point

Why?

Why?

Microsoft

bing

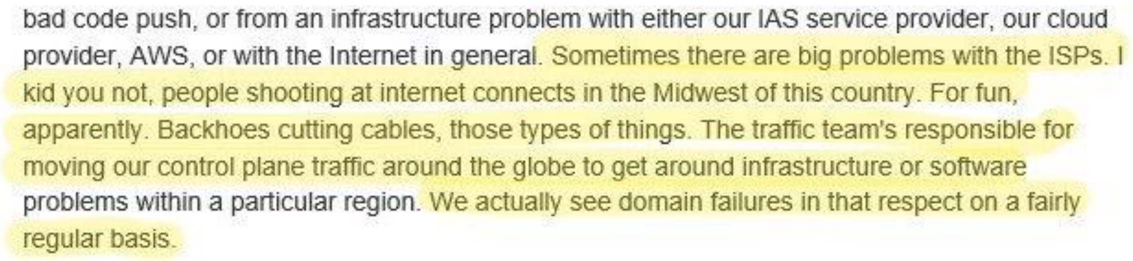
Is Reachability a big problem?

Not Really…

Just like a glacier with its constant fissures and fractures, parts of the Internet go through a constant breakdown/healing lifecycle

It's not *that* big, …but it exists…

Casey Rosenthal (Netflix)

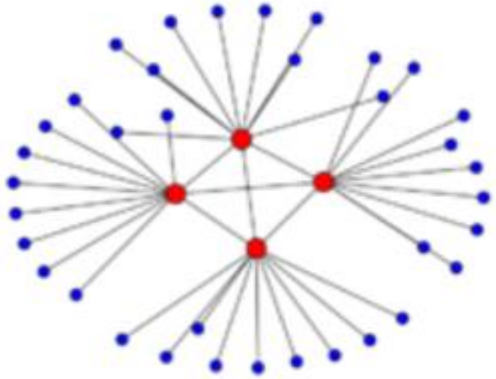


bad code push, or from an infrastructure problem with either our IAS service provider, our cloud provider, AWS, or with the Internet in general. Sometimes there are big problems with the ISPs. I kid you not, people shooting at internet connects in the Midwest of this country. For fun, apparently. Backhoes cutting cables, those types of things. The traffic team's responsible for moving our control plane traffic around the globe to get around infrastructure or software problems within a particular region. We actually see domain failures in that respect on a fairly regular basis.

https://www.stickyminds.com/interview/how-netflix-embraces-complexity-without-sacrificing-speed-interview-casey-rosenthal

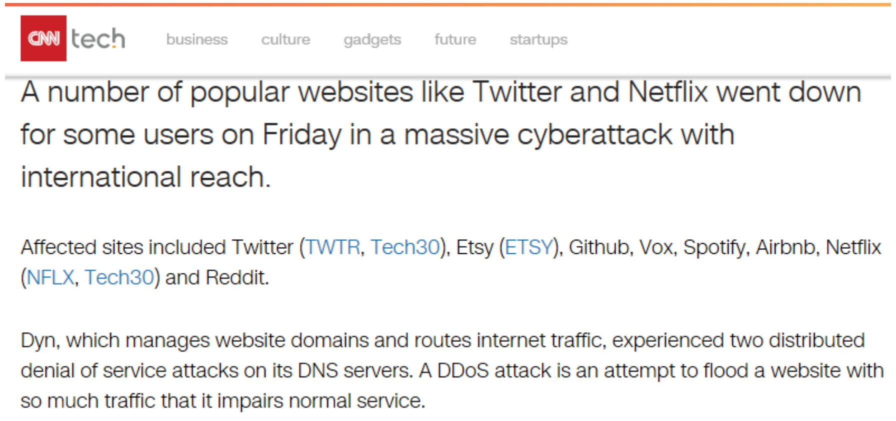
Why should we care?

We are a ‘state-less’ service



Why should we care?

Users expect *us* to own the E2E experience



A number of popular websites like Twitter and Netflix went down for some users on Friday in a massive cyberattack with international reach.

Affected sites included Twitter (TWTR, Tech30), Etsy (ETSY), Github, Vox, Spotify, Airbnb, Netflix (NFLX, Tech30) and Reddit.

Dyn, which manages website domains and routes internet traffic, experienced two distributed denial of service attacks on its DNS servers. A DDoS attack is an attempt to flood a website with so much traffic that it impairs normal service.

Availability and Reachability have direct effects on CAC, ARPU, LTV and Churn

Measuring Reachability is a Hard Problem

Traditionally done with External Synthetic Monitoring

User Perspective →

Monitoring Perspective

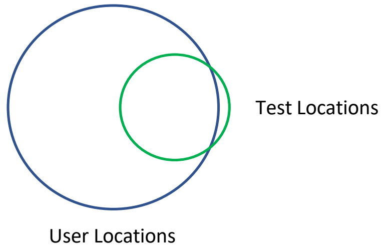
	Known *{users know}*	**Unknown** *{users don't know}*
Known *{we know}*	Eg. "External Tests alerting on an outage"	Eg. "Tests alerting for a datacenter failover"
Unknown *{we don't know}*	Eg. "Quarter of the African continent cannot access your services because of DNS issues."	Eg. "A nation-state sponsored competitor is siphoning your traffic away."

External Testing Alone as a Measure

Prone to the False Positive Paradox

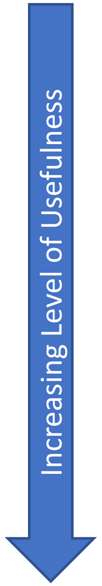
- "For a test that is highly reliable at 99.9%, only about 50% of actual alerts are real problems."
- See Appendix for math

Low coverage of user base



Not Actionable for a single failure. If a test fails, and you get called, what do you do?

What does ‘Actionable’ mean?

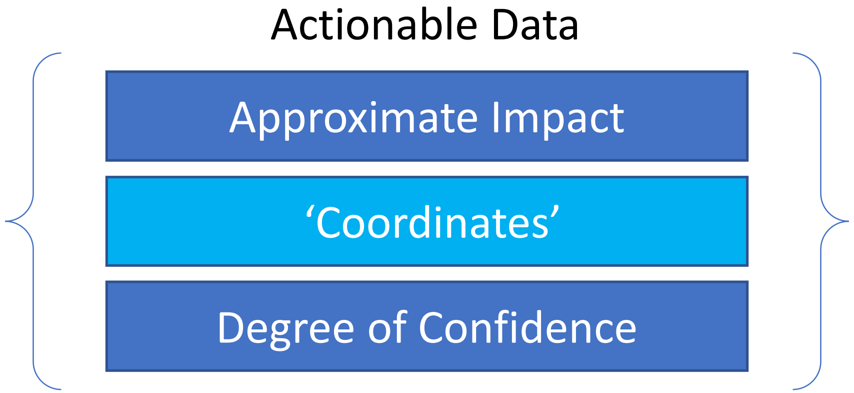


- ASN X saw a 30% traffic drop since 11 AM

- ASN X saw a 40% traffic drop in North-West US since 11 AM

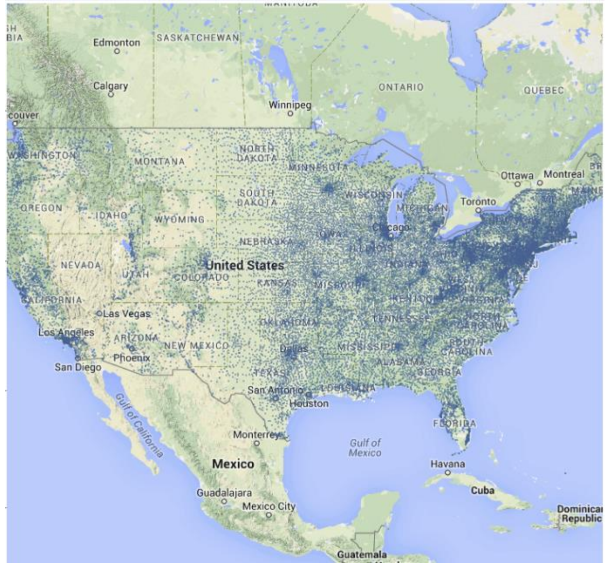
- ASN X saw a 40% traffic drop in Washington State since 11 AM

- ASN X in the **Greater Seattle Area** saw a **70%** traffic drop since 11 AM. **Users affected**: 10-15K. Confidence: **High**. Prefixes seeing issues: 1.2.3.4/22, 2.3.4.5/23

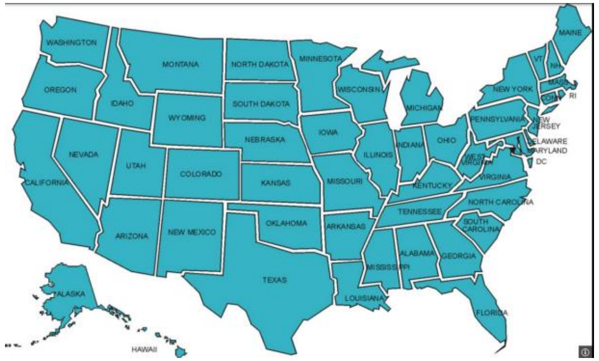


Reachability Coordinate System

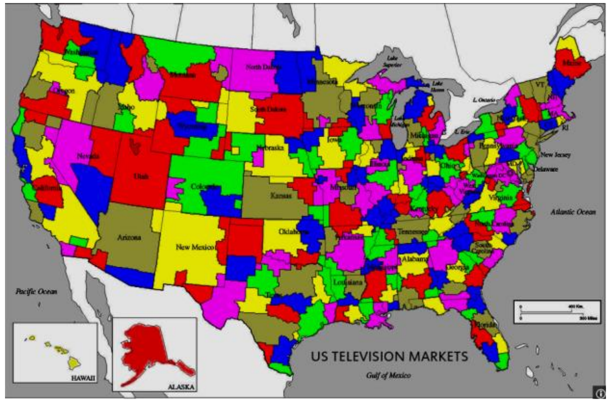
{x,y} == {Location, ISP}



Cities – Too Small



States – Too Big



Metro Areas – Just Right

Don’t these exist already?

- *DMA/Media Market/TMA, Counties, Districts, Prefectures, Provinces*

Metro Areas – Where our users are

- Practical dimension to data that represents geography *and* population distribution

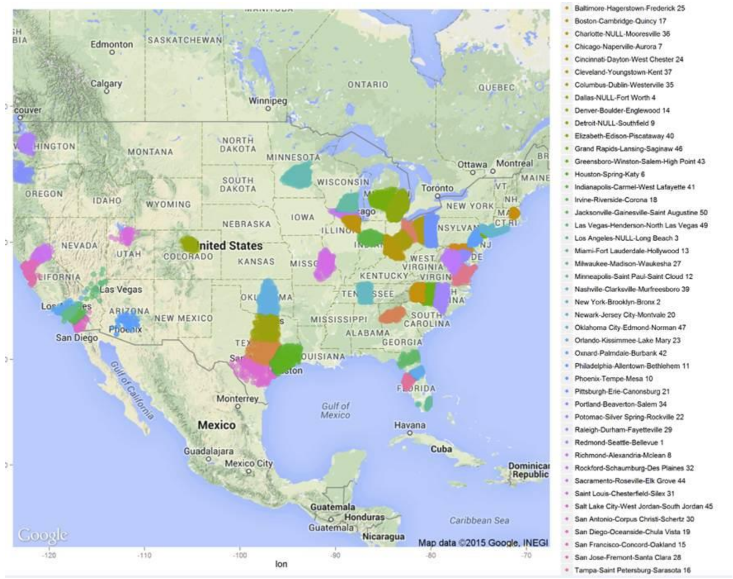
- City-regions based on proximity to each other, and weighted by Query Volume from them.

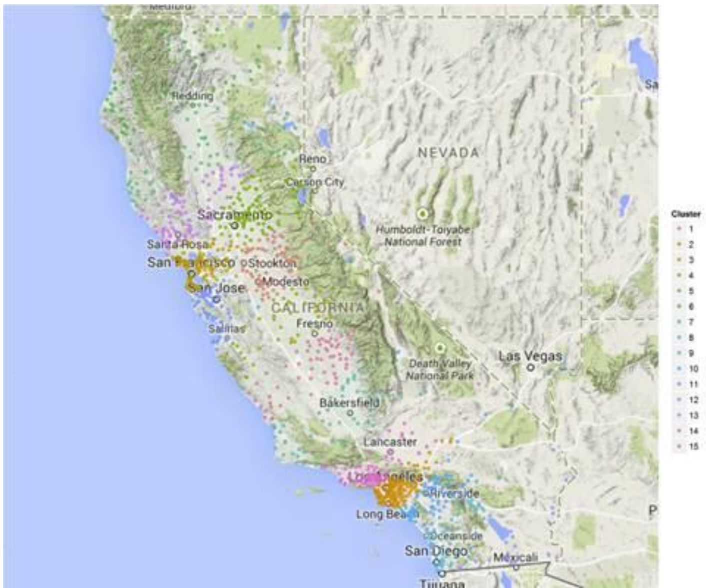
- Created using a form of K-means clustering

- Can change as traffic patterns change. New additions through a *k-NN* Classifier

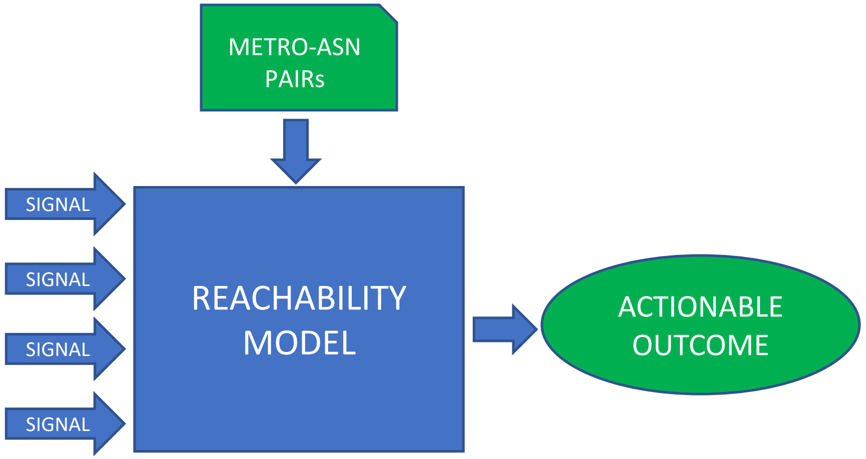
- Scales internationally; allows larger areas in sparse regions

- We decided to begin with top Metro-ASNs that give >0.5% traffic


Edmonton
SASKATCHEWAN
Calgary
Winnipeg
ONTARIO
QUEBEC
MONTANA
NORTH DAKOTA
MINNESOTA
Ottawa
Montreal
SOUTH DAKOTA
WISCONSIN
Toronto
MAINE
OREGON
IDAHO
WYOMING
IOWA
NEW YORK
NEBRASKA
NEVADA
UTAH
United States
COLORADO
KANSAS
KENTUCKY
VIRGINIA
CALIFORNIA
Las Vegas
TENNESSEE
SOUTH CAROLINA
ARIZONA
NEW MEXICO
MISSISSIPPI
ALABAMA
GEORGIA
San Diego
Phoenix
Houston
FLORIDA
Gulf of California
Monterrey
Gulf of Mexico
Mexico
Havana
Cuba
Guadalajara
Mexico City
Dominican Republic
Guatemala
Honduras
Nicaragua
Caribbean Sea
Google
Map data ©2015 Google, INEGI
-120
-110
-100
-90
-80
-70
lon
Baltimore-Hagerstown-Frederick 25
Boston-Cambridge-Quincy 17
Charlotte-NULL-Mooresville 36
Chicago-Naperville-Aurora 7
Cincinnati-Dayton-West Chester 24
Cleveland-Youngstown-Kent 37
Columbus-Dublin-Westerville 35
Dallas-NULL-Fort Worth 4
Denver-Boulder-Englewood 14
Detroit-NULL-Southfield 9
Elizabeth-Edison-Piscataway 40
Grand Rapids-Lansing-Saginaw 46
Greensboro-Winston-Salem-High Point 43
Houston-Spring-Katy 6
Indianapolis-Carmel-West Lafayette 41
Irvine-Riverside-Corona 18
Jacksonville-Gainesville-Saint Augustine 50
Las Vegas-Henderson-North Las Vegas 49
Los Angeles-NULL-Long Beach 3
Miami-Fort Lauderdale-Hollywood 13
Milwaukee-Madison-Waukesha 27
Minneapolis-Saint Paul-Saint Cloud 12
Nashville-Clarksville-Murfreesboro 39
New York-Brooklyn-Bronx 2
Newark-Jersey City-Montvale 20
Oklahoma City-Edmond-Norman 47
Orlando-Kissimmee-Lake Mary 23
Oxnard-Palmdale-Burbank 42
Philadelphia-Allentown-Bethlehem 11
Phoenix-Tempe-Mesa 10
Pittsburgh-Erie-Canonsburg 21
Portland-Beaverton-Salem 34
Potomac-Silver Spring-Rockville 22
Raleigh-Durham-Fayetteville 29
Redmond-Seattle-Bellevue 1
Richmond-Alexandria-Mclean 8
Rockford-Schaumburg-Des Plaines 32
Sacramento-Roseville-Elk Grove 44
Saint Louis-Chesterfield-Silex 31
Salt Lake City-West Jordan-South Jordan 45
San Antonio-Corpus Christi-Schertz 30
San Diego-Oceanside-Chula Vista 19
San Francisco-Concord-Oakland 15
San Jose-Fremont-Santa Clara 28
Tampa-Saint Petersburg-Sarasota 16



A Model Solution



Signals

- Proxy for Users
- Representative of the State of the Internet
- Temporal
- Aligned to Metro - ASNs

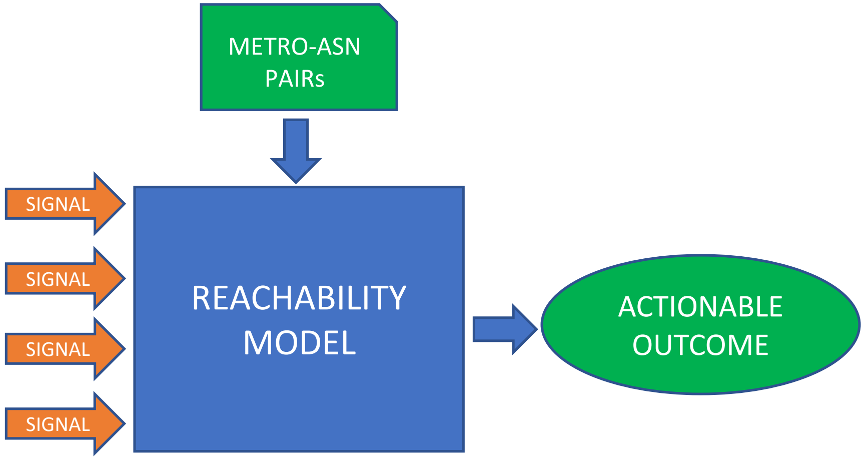
Model

- Statistically defensible
- Capable of learning
- Relatively Simple (KISS)
- Fast

Outcome

- Low False Negative Rate
- Extremely low False Positive Rate

A Model Solution



METRO-ASN PAIRs

SIGNAL

SIGNAL

SIGNAL

SIGNAL

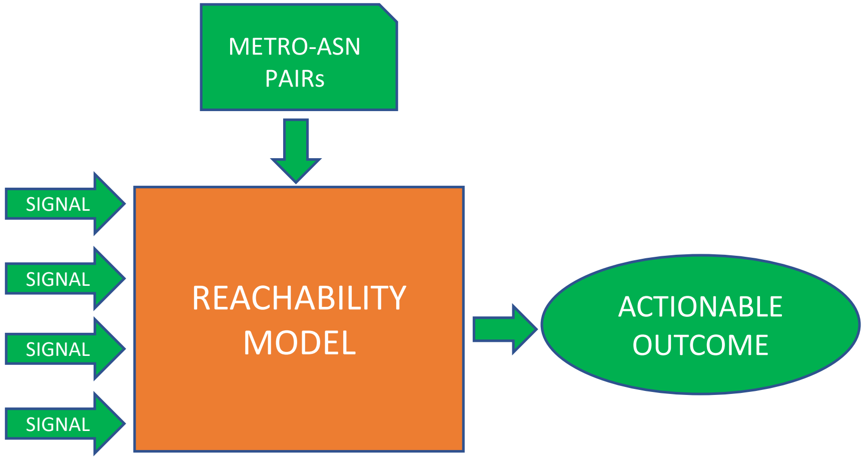
REACHABILITY MODEL

ACTIONABLE OUTCOME

Signals from our space

- Direct
  - Deterministic test to detect failures in Reachability (eg. External Tests)
- Indirect
  - Alternate path telemetry to register failure (eg. HockeyApp, Navigational Error Logging)
- Inferential
  - Measurable from *inside* our datacenters (Traffic loss detection)
- Correlational
  - Social , Crowdsourced (Eg, FB, Twitter, Down Detector)
- Causal
  - BGP, DNS changes (Eg. BGPMon, DNS telemetry)

A Model Solution



METRO-ASN PAIRs

SIGNAL

SIGNAL

SIGNAL

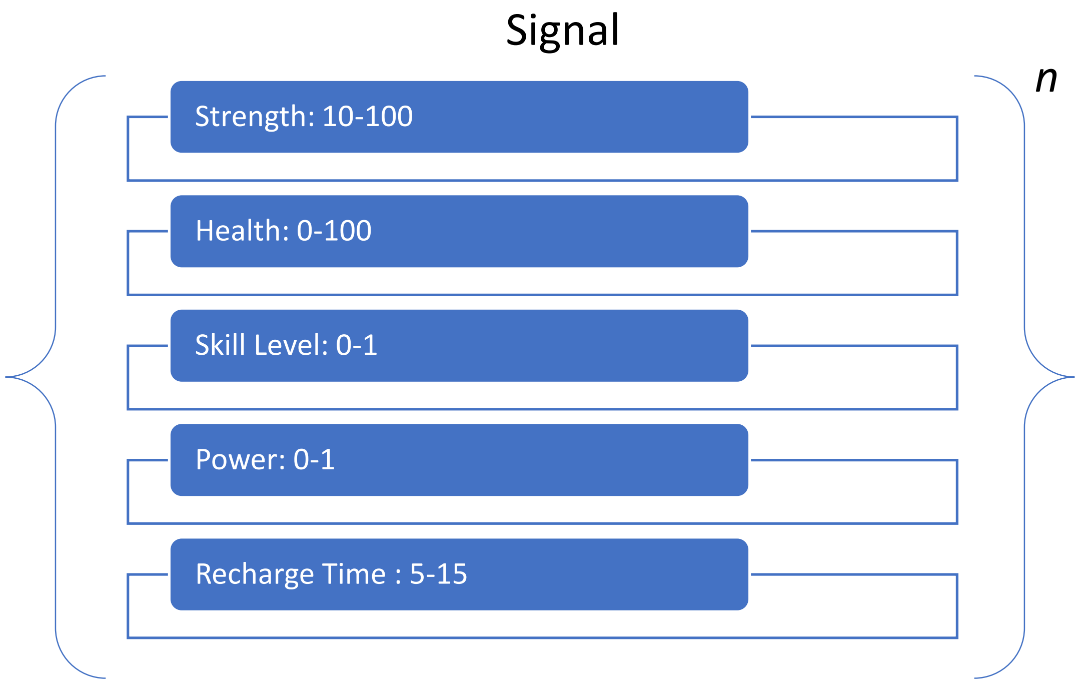
SIGNAL

REACHABILITY MODEL

ACTIONABLE OUTCOME

‘Gaming’ the Model

Bring down a Metro-ASN with most potent signal ‘hits’.



Signal

n

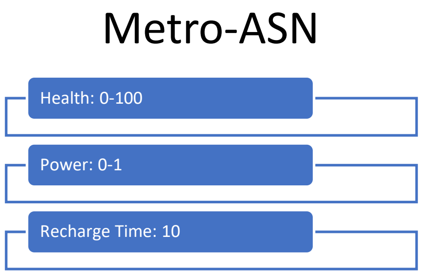
Strength: 10-100

Health: 0-100

Skill Level: 0-1

Power: 0-1

Recharge Time : 5-15



Metro-ASN

Health: 0-100

Power: 0-1

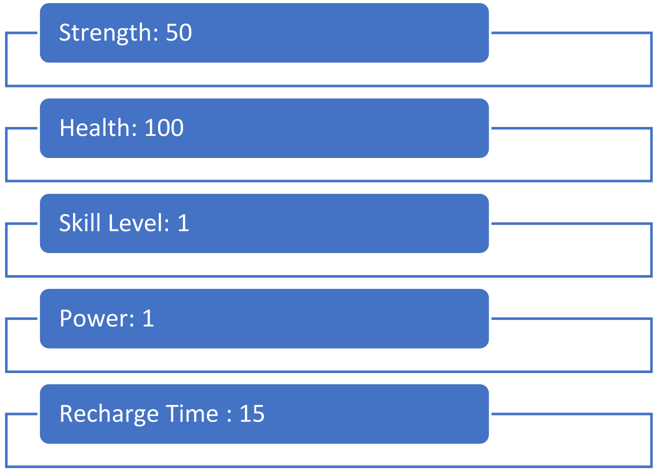
Recharge Time: 10

$$SignalHit = \sum_{t=1}^{t=RechargeTimeSegments} [(Strength \quad \times Skill \quad) - (Health_{max} - Health \quad)]$$

Direct – External Tests

- Synthetic Testing via systems like CatchPoint, Dynatrace et. al.
- No Alerts through these systems. Just API calls. Alerts are through our model.
- Can cover major ISPs and Metro Areas
- ‘Last Mile’ testing increases coverage at a cost
- Top 50 Metro-ASN pairs are easily covered

DT:NewYork-Brooklyn-Bronx:6128



Strength: 50

Health: 100

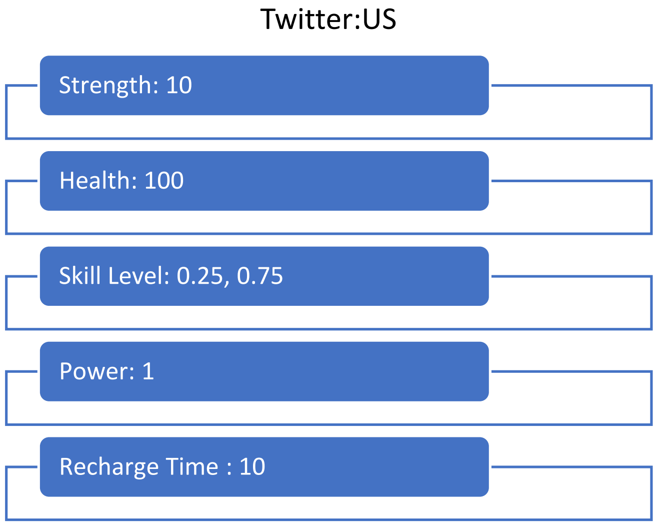
Skill Level: 1

Power: 1

Recharge Time : 15

Correlational - Twitter

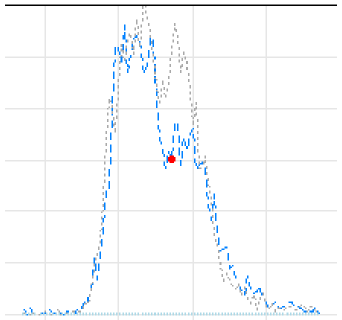
- Simple Sentiment Analysis – looking for known negative words. Also overall mentions.
- Based on Anomaly Detection.
- Low strength; no association Metro-ASN generally speaking
- Prone to feedback/opinion spikes


Twitter:US

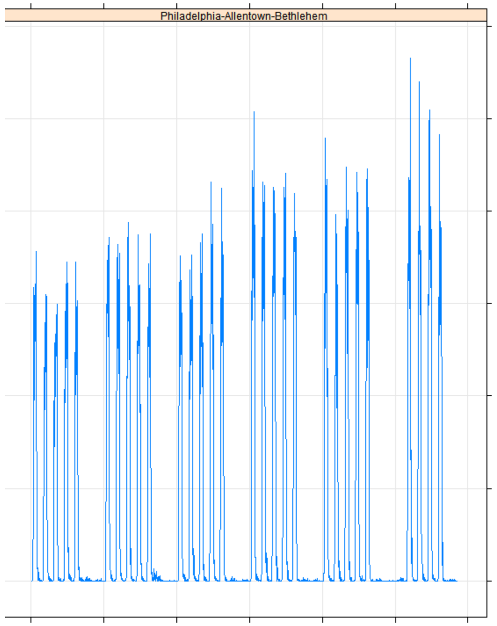
- Strength: 10
- Health: 100
- Skill Level: 0.25, 0.75
- Power: 1
- Recharge Time : 10

Inferential – Traffic Drop

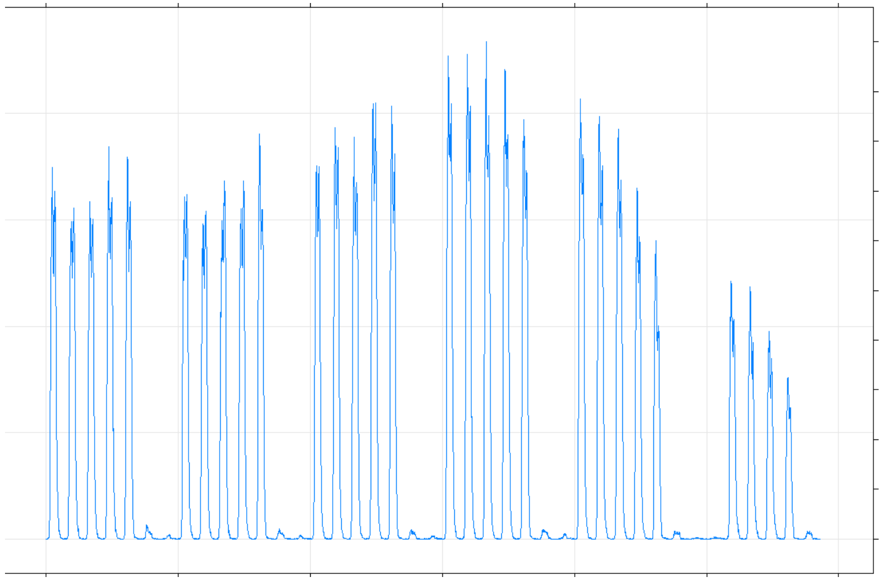
- Uses Anomaly Detection – use Twitter's open-source ESD R library with an outer layer
- Has a damping calendrical signal to account for holidays etc. (power goes low)
- Drops are classified as Egregious, Large or Medium (Level)



Metro-ASN: Detected



Complete Metro Area – Not Detected



Complete ASN – Not Detected

Drop:NewYork-Brooklyn-Bronx:6128

Strength: 100

Health: 100

Skill Level: 0.25, 0.75, 1.0

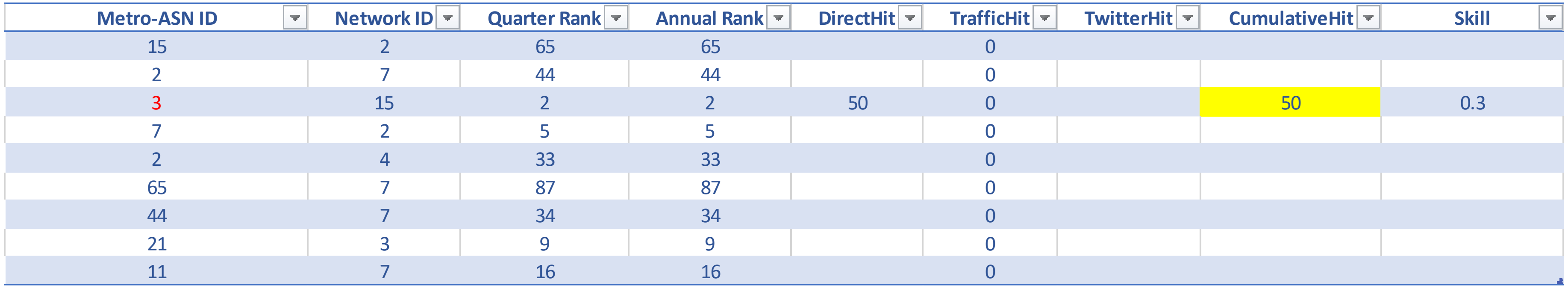
Power: 1

Recharge Time : 5

Other Signals in Appendix

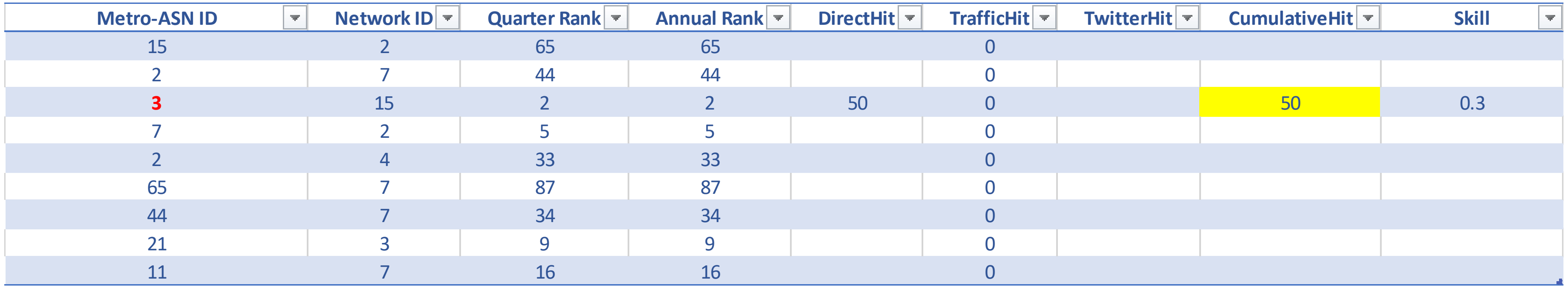
Simulations – Scenario A

External Test in Metro-ASN 3 Fails



Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0			
2	7	44	44		0			
3	15	2	2	50	0		50	0.3
7	2	5	5		0			
2	4	33	33		0			
65	7	87	87		0			
44	7	34	34		0			
21	3	9	9		0			
11	7	16	16		0			

Test Fails again, 15 minutes later

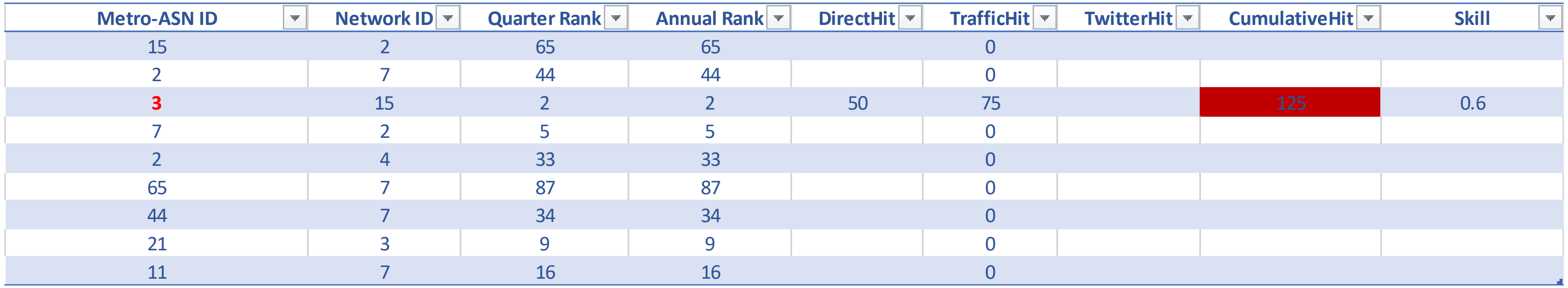


Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0			
2	7	44	44		0			
3	15	2	2	50	0		50	0.3
7	2	5	5		0			
2	4	33	33		0			
65	7	87	87		0			
44	7	34	34		0			
21	3	9	9		0			
11	7	16	16		0			

- Single node failures with no supporting signal don’t cause false positives
- RechargeTime(Metro-ASN) < RechargeTime(External Tests): Allows failing node to not exceed threshold

Simulations – Scenario B

External Test in Metro-ASN 3 Fails *AND* traffic drops by 50%

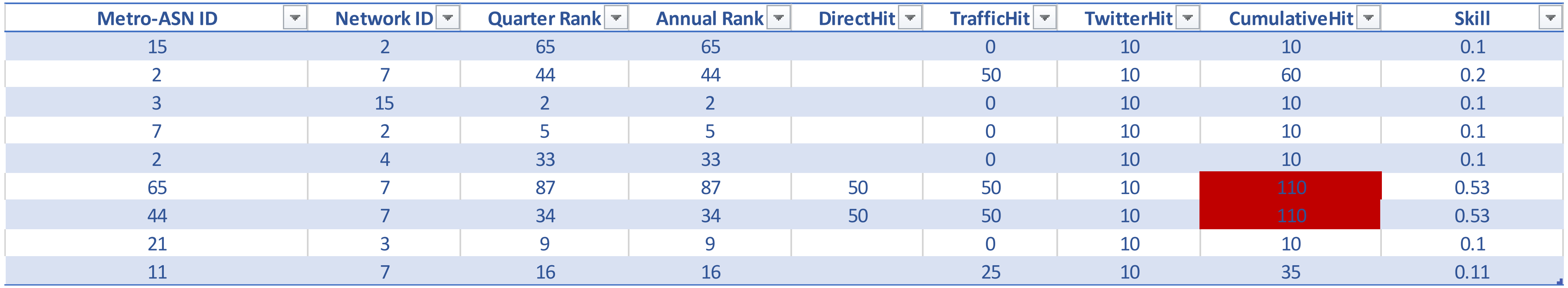


Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0			
2	7	44	44		0			
3	15	2	2	50	75		125	0.6
7	2	5	5		0			
2	4	33	33		0			
65	7	87	87		0			
44	7	34	34		0			
21	3	9	9		0			
11	7	16	16		0			

- Two supporting signals reach threshold easily
- In real-life, chances are that there are more than one Metro-ASN ID failures

Simulations – Scenario C

Network ID 7 has failures in a lot of Metro-Areas. Someone complained on Twitter as well

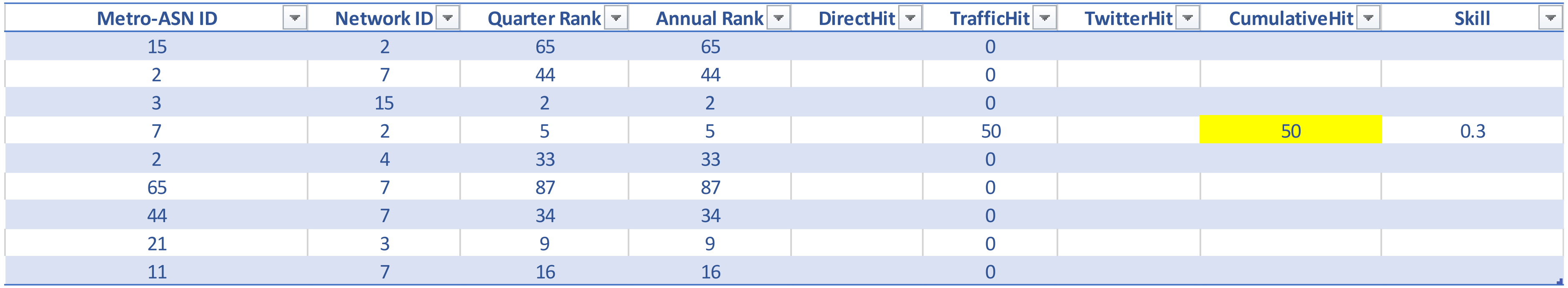


Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0	10	10	0.1
2	7	44	44		50	10	60	0.2
3	15	2	2		0	10	10	0.1
7	2	5	5		0	10	10	0.1
2	4	33	33		0	10	10	0.1
65	7	87	87	50	50	10	110	0.53
44	7	34	34	50	50	10	110	0.53
21	3	9	9		0	10	10	0.1
11	7	16	16		25	10	35	0.11

- For now, Twitter feed is location agnostic, so hits add everywhere.
- Only 2 locations happened to have External Tests running – that was the reason this became a True Positive
- What if we didn't have any External Tests at all?

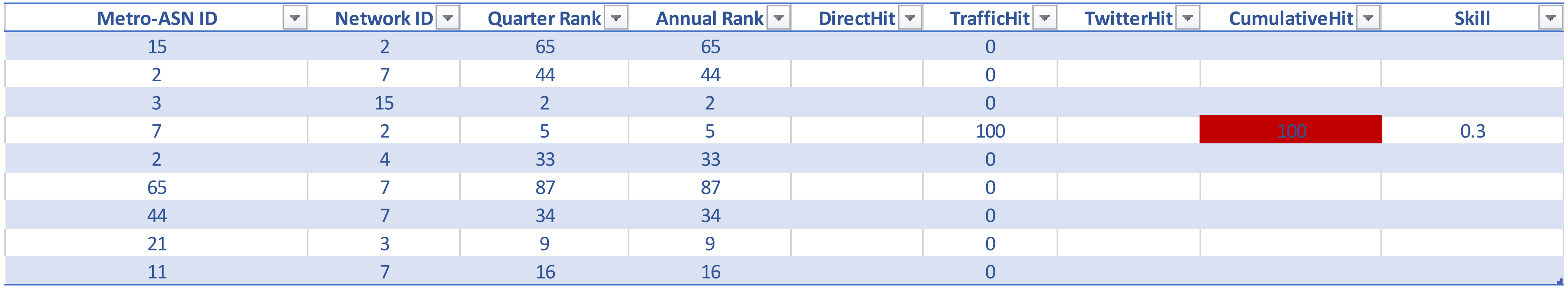
Simulations – Scenario D

Metro-ASN ID 7 has a ‘medium’ drop in traffic. No External Tests or Reports



Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0			
2	7	44	44		0			
3	15	2	2		0			
7	2	5	5		50		50	0.3
2	4	33	33		0			
65	7	87	87		0			
44	7	34	34		0			
21	3	9	9		0			
11	7	16	16		0			

15 minutes later, Traffic Hit accumulates to trigger alert



Metro-ASN ID	Network ID	Quarter Rank	Annual Rank	DirectHit	TrafficHit	TwitterHit	CumulativeHit	Skill
15	2	65	65		0			
2	7	44	44		0			
3	15	2	2		0			
7	2	5	5		100		100	0.3
2	4	33	33		0			
65	7	87	87		0			
44	7	34	34		0			
21	3	9	9		0			
11	7	16	16		0			

- A *sustained* traffic drop is different than a single node sustained failure.
- Longer Time To Detect at the cost of reducing False Positives

What Did We *Really* Find?

- Multiple DNS Hijacks across ISPs (even in the US) – we fixed a few
- Reachability impact due to fiber cuts
- Accidental school district blockade that lasted a week
- Another BGP hijack event (much smaller one) in South America
- An unusual day when a government agency took half the day off ;)

What Doesn’t Work

- Existing Issues part of baseline
- No Modeling for a Calendrical System
- Bing Outage or Internet Outage?
- Low Signal-To-Noise ratio for small MetroASNs
  - Experimenting with extending recharge time based on recall
- Social: Feature DSAT can spike pretty hard
- Reverse IP data change
- Escalation. So yeah, TTD improved. What about TTM?

SWOT ANALYSIS

Helpful
to achieving the objective

Harmful
to achieving the objective

Internal origin
(attributes of the organization)

External origin
(attributes of the environment)

Strengths

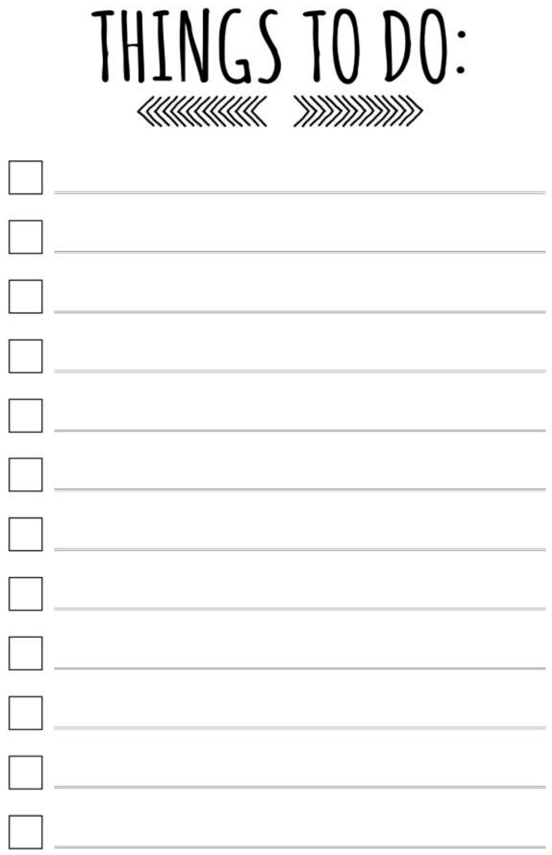
Weaknesses

Opportunities

Threats

Wishlist

- Indirect Signals – W3C Network Error Logging
    - While app-based mechanisms exist, the W3C NEL draft is a great standard to support for the web.
- Fix ‘Bing Outage or ISP Outage’
    - Have a plan in mind
- Make plugin-based system for more signals
- Stronger inferential signal with other products
- Build better learning in the model for seasonality
- Cover edge or unusual cases
    - What if traffic drops, but Test succeeds?
    - Not good at catching intermittent issues.



Take-Aways

- For large consumer services, someone is having a problem, just with your service, as you sit here.
- For a large, ‘stateless’, high-churn service like ours, reachability is important
- External Tests are not the complete story on reachability
- There’s power in unity – leverage different signals to make a composite
- Anomaly Detection and Machine Learning can play a larger role in monitoring

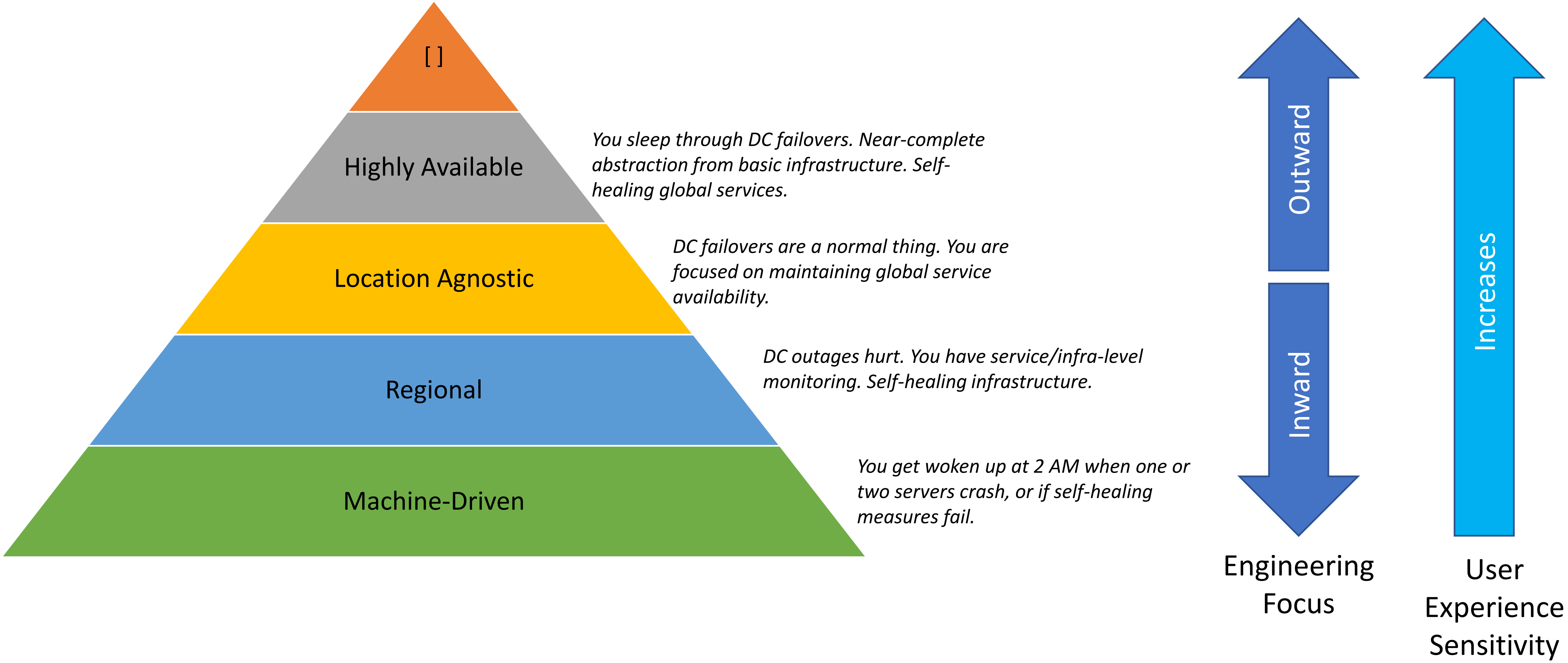
Thank you for listening!

mosuley@microsoft.com

@mohitsuley

Appendix

Obligatory Maturity Model (OMM)



[ ]

Highly Available

You sleep through DC failovers. Near-complete abstraction from basic infrastructure. Self-healing global services.

Location Agnostic

DC failovers are a normal thing. You are focused on maintaining global service availability.

Regional

DC outages hurt. You have service/infra-level monitoring. Self-healing infrastructure.

Machine-Driven

You get woken up at 2 AM when one or two servers crash, or if self-healing measures fail.

Outward

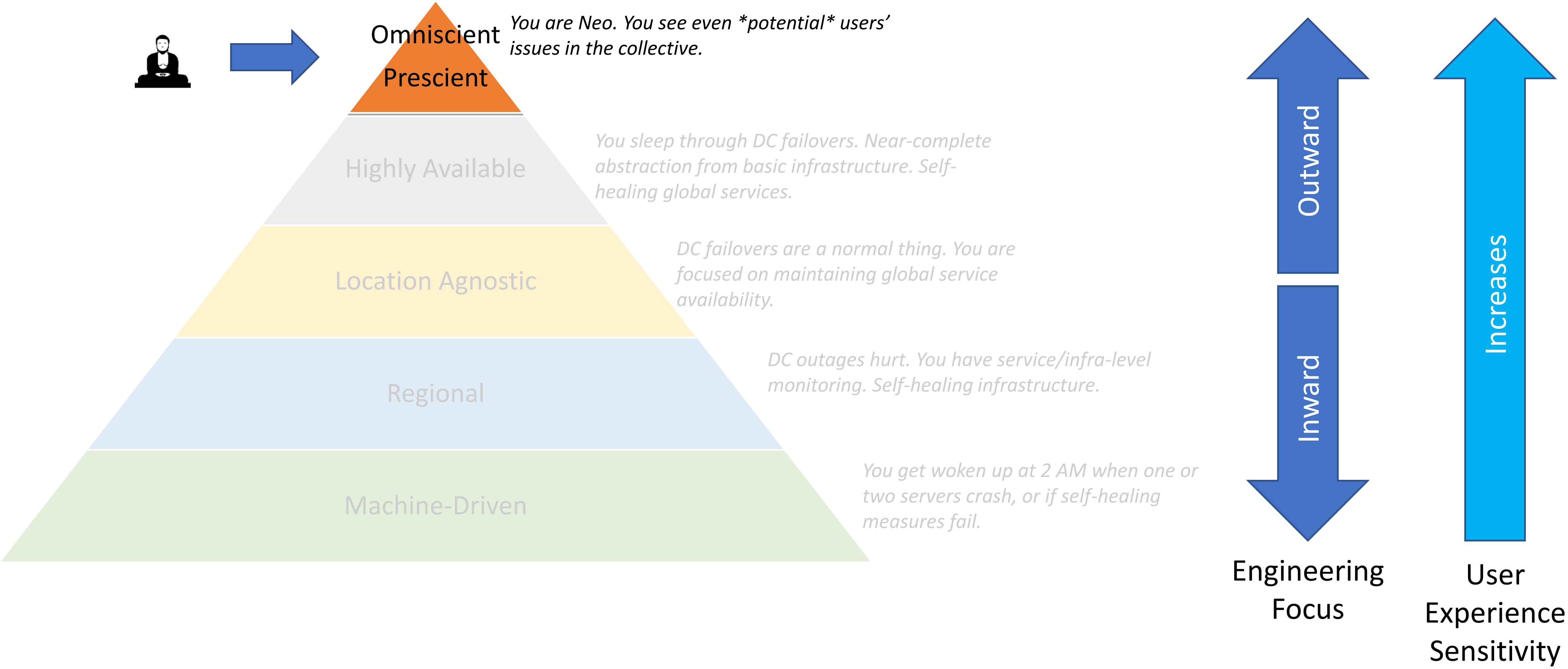
Inward

Engineering Focus

Increases

User Experience Sensitivity

Obligatory Maturity Model (OMM)



Omniscient
Prescient

*You are Neo. You see even *potential* users' issues in the collective.*

Highly Available

You sleep through DC failovers. Near-complete abstraction from basic infrastructure. Self-healing global services.

Location Agnostic

DC failovers are a normal thing. You are focused on maintaining global service availability.

Regional

DC outages hurt. You have service/infra-level monitoring. Self-healing infrastructure.

Machine-Driven

You get woken up at 2 AM when one or two servers crash, or if self-healing measures fail.

Outward

Inward

Engineering Focus

Increases

User Experience Sensitivity

False Positive Paradox Explained

Assume

- External test has false positive rate of 0.001
- 1000 tests run in an hour
- 0.1% tests are actual failures

$$1000\ X\ \frac{1}{100} = 1\ Actual\ Failure$$

- With a false positive rate of 0.1%

$$1000\ X\ \frac{100 - 0.1}{100}\ X\ 0.001 = 0.9\ False\ Positives$$

- Therefore, probability of being a real problem:

$$\frac{1}{1 + 0.9} = 52\%$$

- For a test that is highly reliable at 99.9% only about 52% of actual alerts are real problems.

Social – Down Detector

- Plan to use this soon.
- Provides ISP-level information. More reliable and useful than just a Twitter feed, historically

DD:7922

Strength: 25

Health: 100

Skill Level: 0.25, 0.75, 1.0

Power: 1

Recharge Time : 10

Causal – BGP Stream

- BGPMon stream
- No Metro Areas.
- If an announcement removal is detected, causal.

BGP:6128

Strength: 10

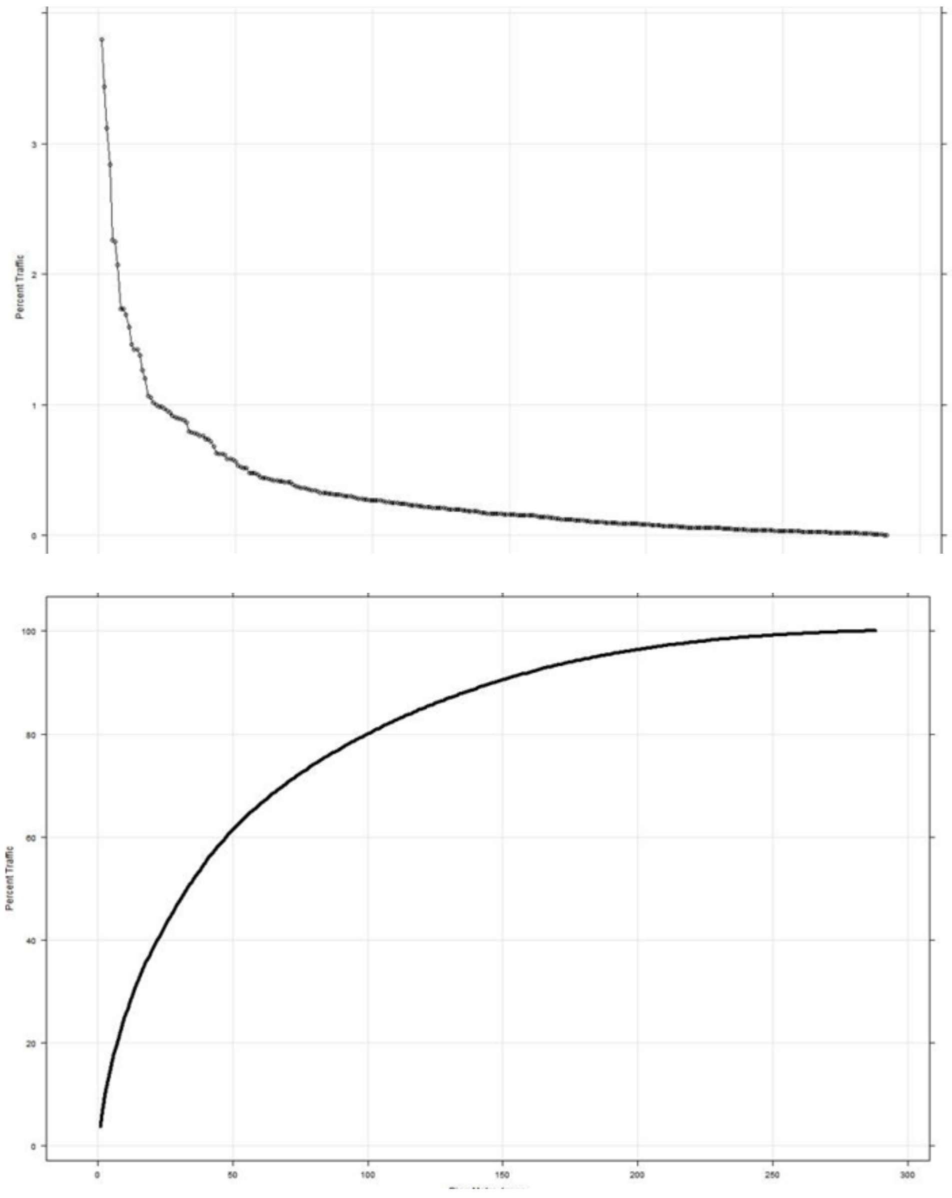
Health: 100

Skill Level: 0.25, 0.75, 1.0

Power: 1

Recharge Time : 5

CDFs for Metro Areas



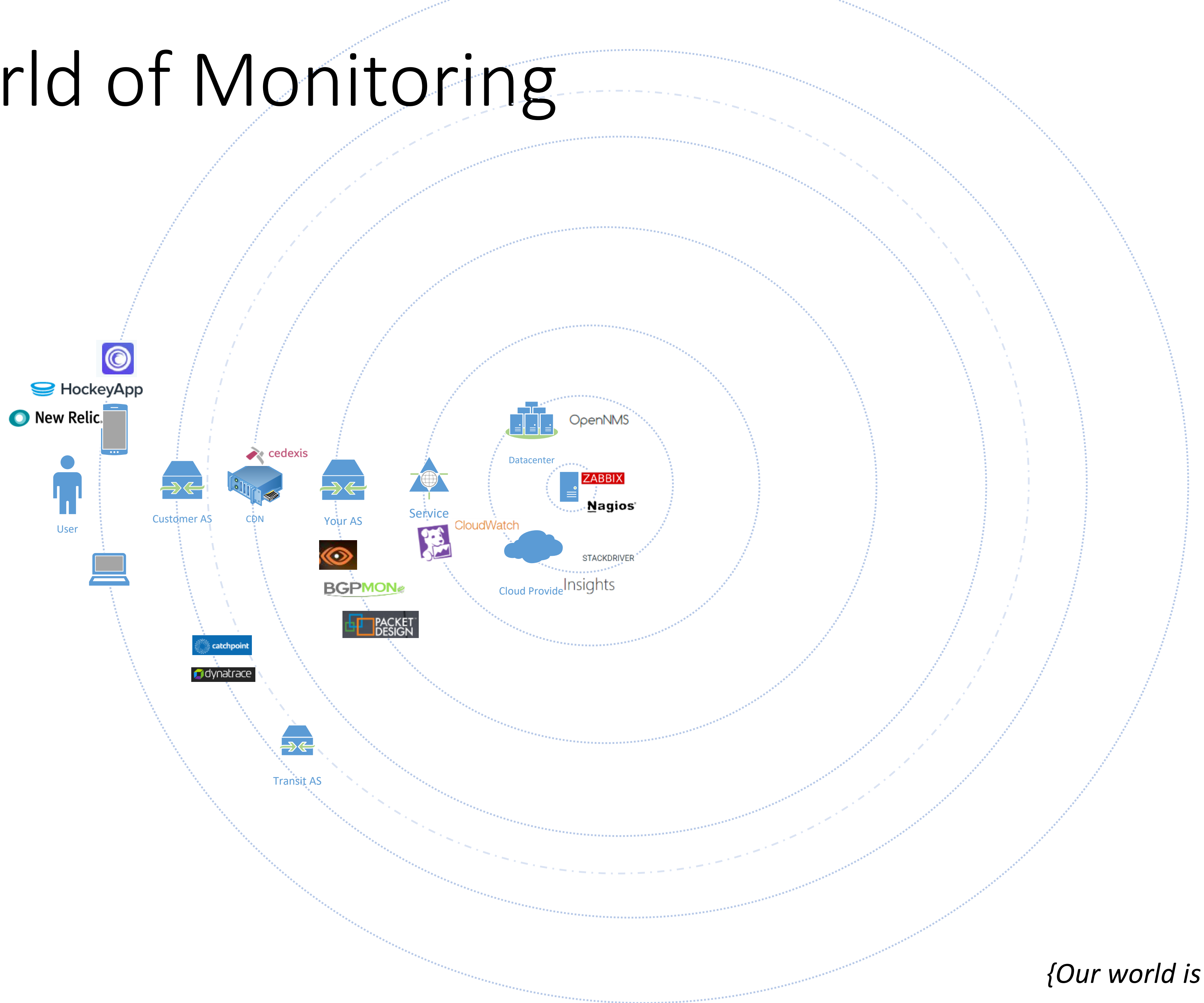
Simplified Model Representation

$$SignalHit = \sum_{t=1}^{t=RechargeTimeSegments} [(Strength \quad \times Skill \quad) - (Health_{max} - Health \quad)]$$

$$Health_{MetroASN} = [Health_{MetroASN} \times Power_{MetroASN}] - \sum_{n=1}^{n=signals_{max}} SignalHit_n$$

$$OverallSkill = \frac{\sum Skill_{signal}}{\sum Signals}$$

The World of Monitoring



{Our world is a small subset of this}