Beyond Data Points: Regionalizing Crowdsourced Latency Measurements

Taveesh Sharma, Paul Schmitt, Francesco Bronzino, Nick Feamster, Nicole P. Marwell

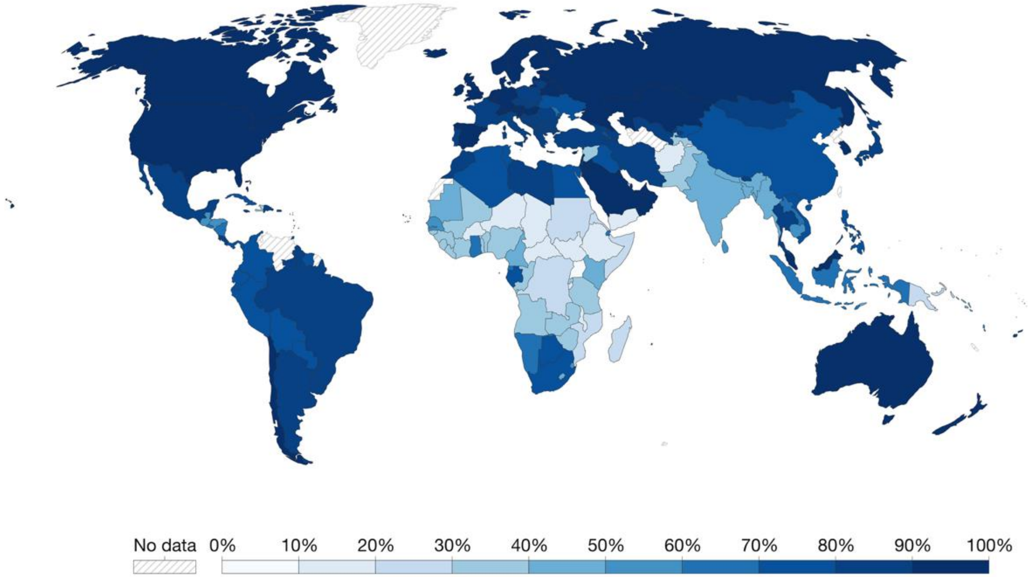
June 11, 2025

The Digital Divide Still Persists

- In high-income countries, **93%** of the population uses the Internet, compared to only **27%** in low-income countries.
- Globally, **83%** of urban residents have Internet access, while only **48%** of rural residents are connected.

Share of the population using the Internet, 2023

Share of the population who used the Internet[1] in the last three months.



Data source: International Telecommunication Union (via World Bank) (2025)

OurWorldinData.org/internet | CC BY

[1] **Internet user** An internet user is defined by the International Telecommunication Union as anyone who has accessed the internet from any location in the last three months.
This can be from any type of device, including a computer, mobile phone, personal digital assistant, games machine, digital TV, and other technological devices.

It’s not just about *being* online — it’s about *how well* you can connect.

Crowdsourced Data: A Powerful Tool

- Platforms like M-Lab and Ookla collect user-generated data
- Millions of real-world latency and speed measurements
- Location-tagged data exposes local performance disparities

OOKLA®

MLAB

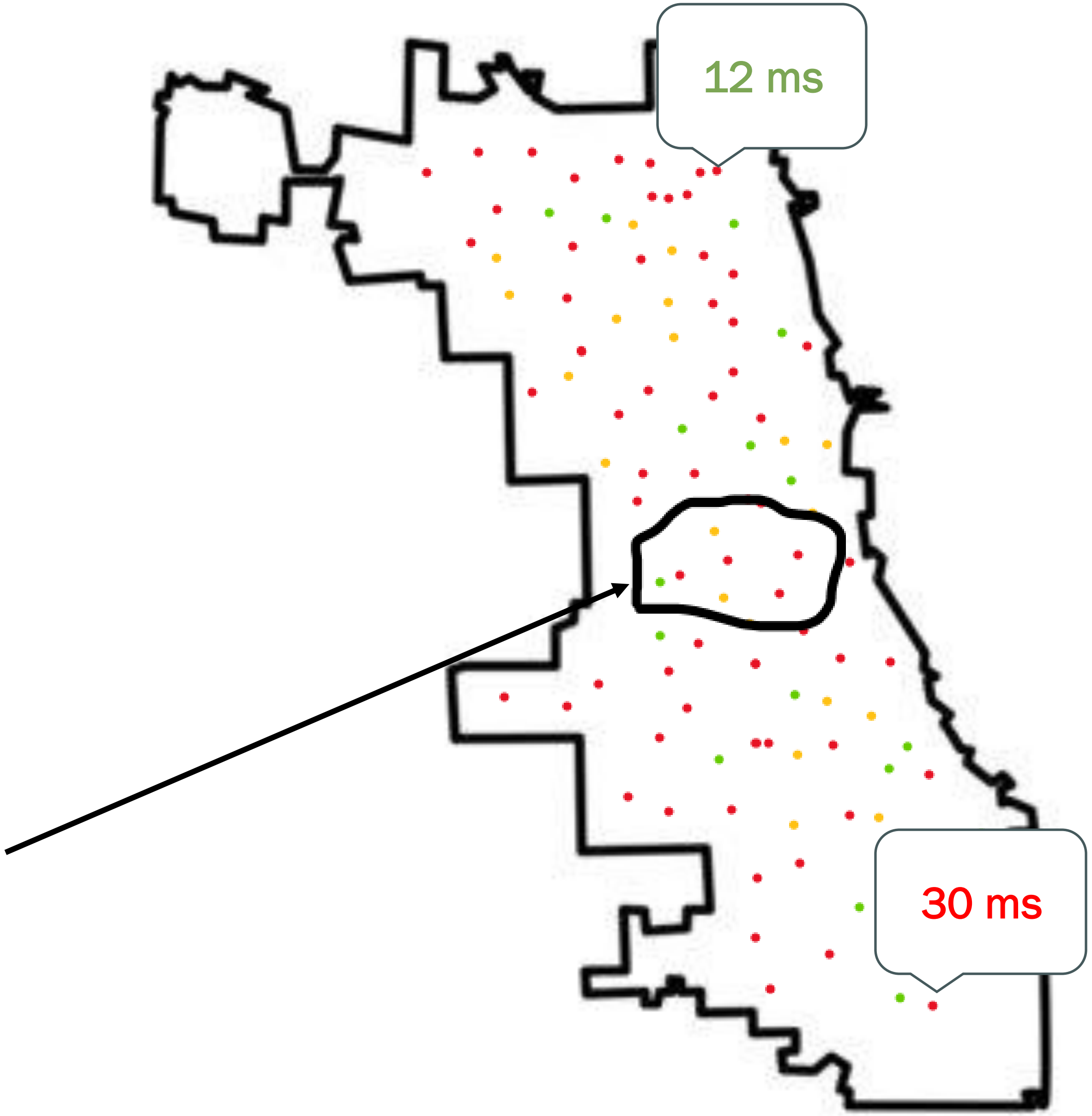
FAST

Speed Test

Why Just Data Points Aren’t Enough?

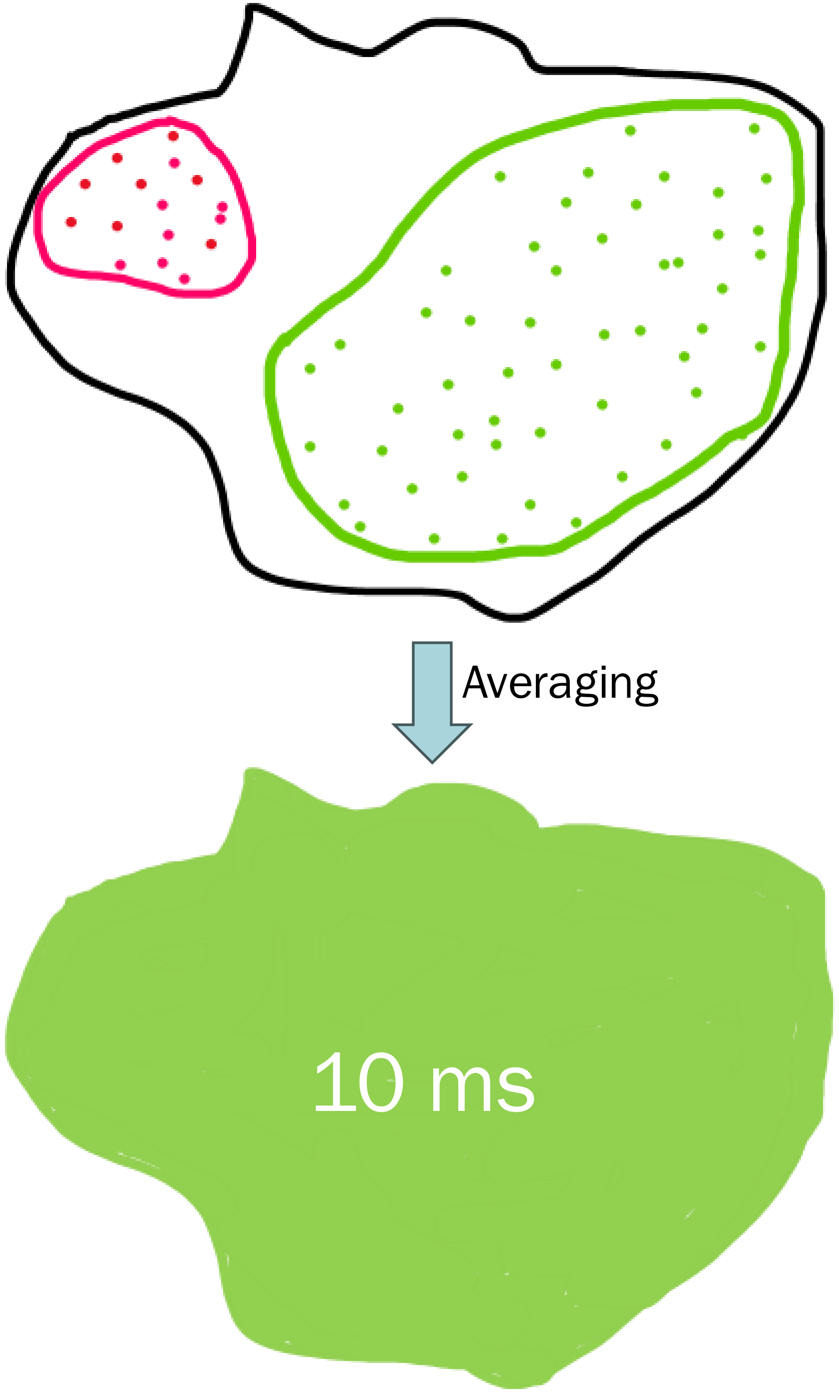
- Where exactly is the Internet slow?

- Has this area improved over time?

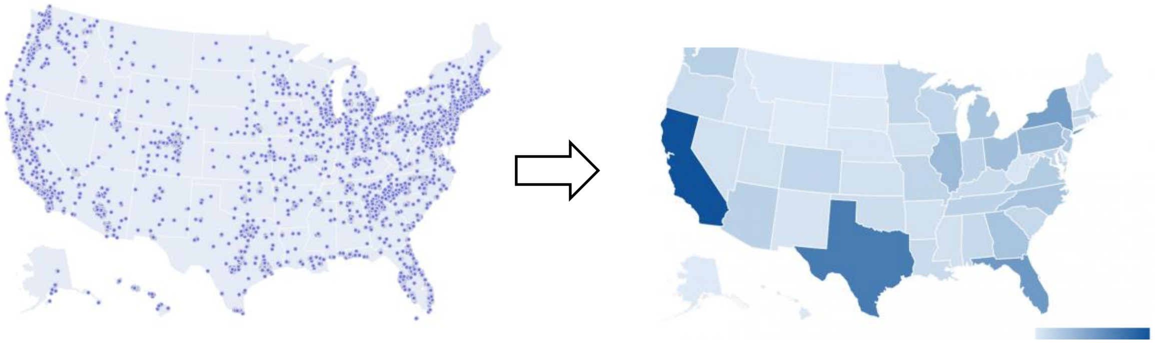


The Problem with Direct Aggregation

Plain aggregation over a region tends to over-represent densely sampled subregions.



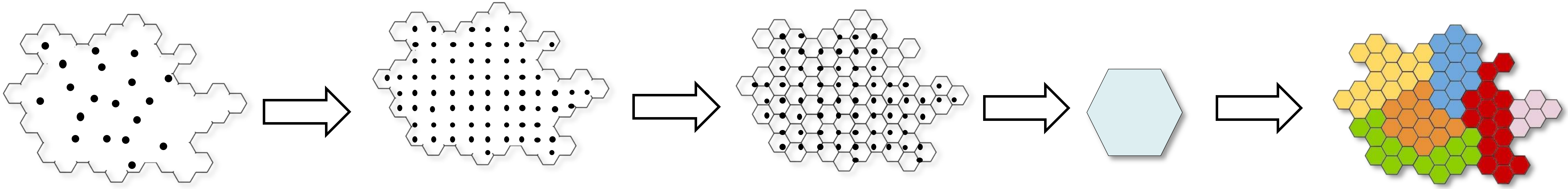
From Points to Boundaries



Key research questions:

- What is the right **spatial granularity** for sampling Internet performance?
- What are the right **metrics** for aggregating Internet performance over regions?

Method



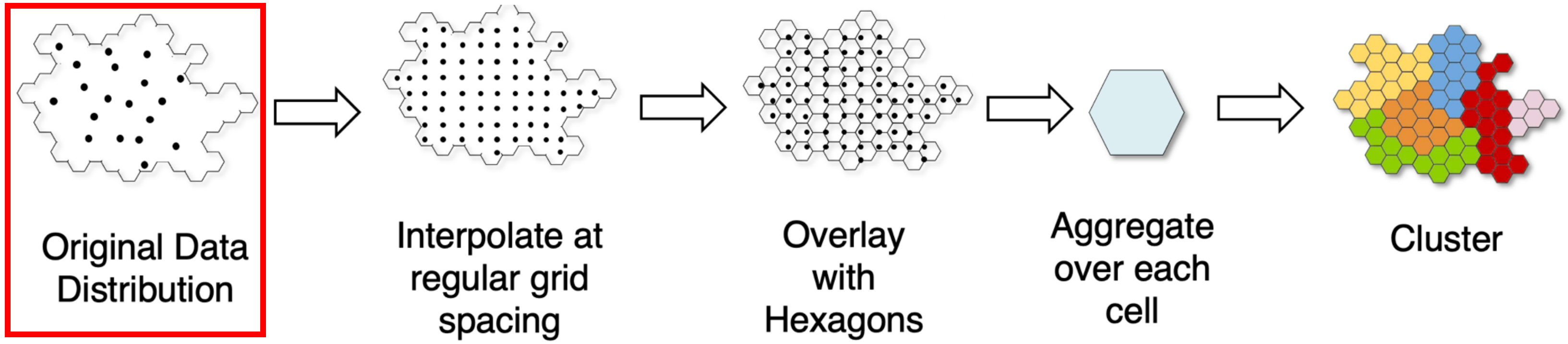
Original Data Distribution

Interpolate at regular grid spacing

Overlay with Hexagons

Aggregate over each cell

Cluster



Original Data Distribution

Interpolate at regular grid spacing

Overlay with Hexagons

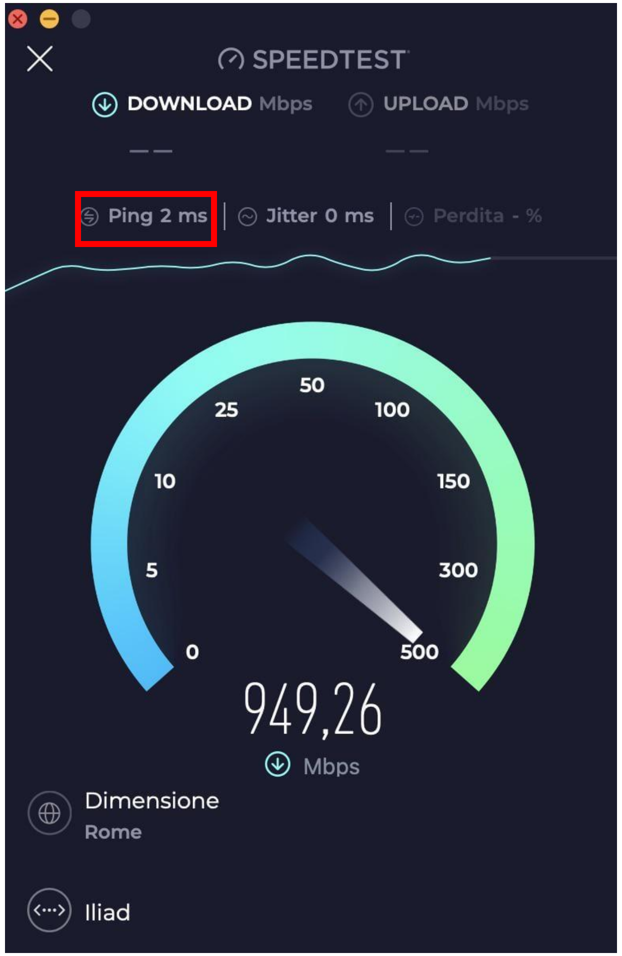
Aggregate over each cell

Cluster

Data

Filter out measurements with:

- VPN connections
- Self-selected servers
- IP geolocations



SPEEDTEST

DOWNLOAD Mbps — —

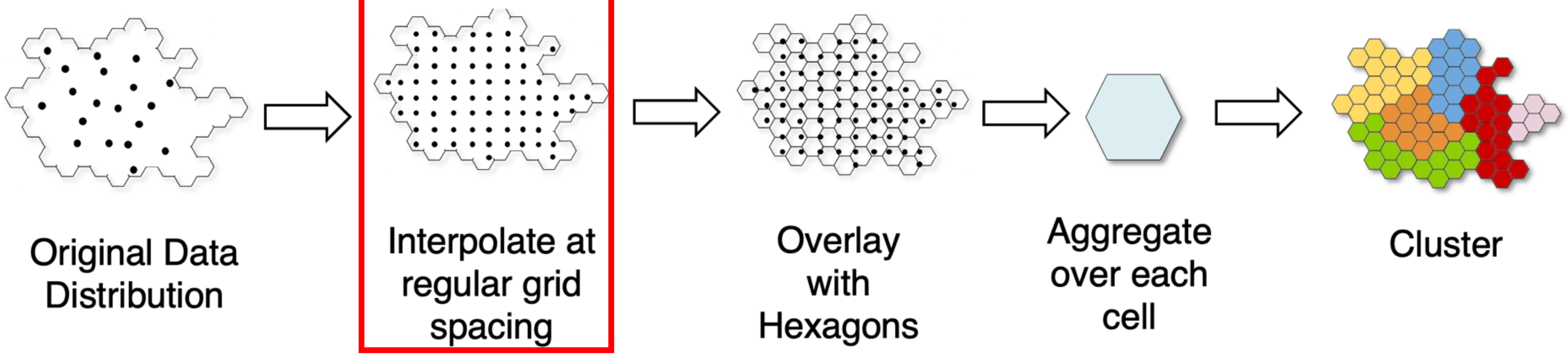
UPLOAD Mbps — —

Ping 2 ms | Jitter 0 ms | Perdita - %

949,26 Mbps

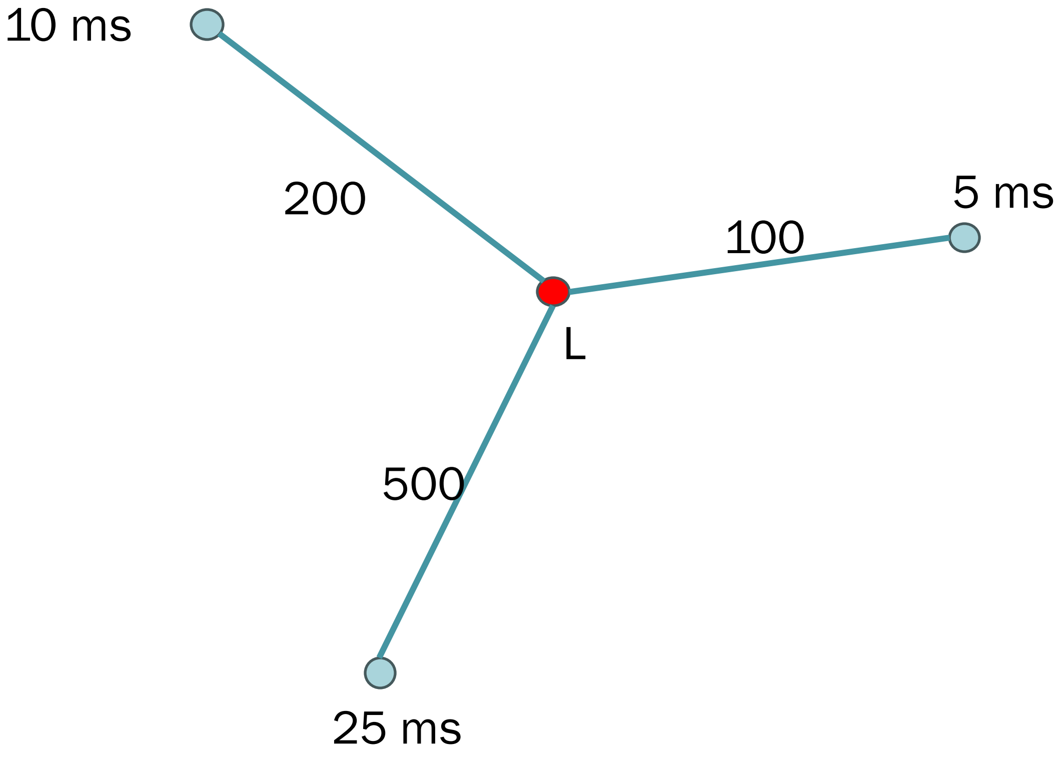
Dimensione
Rome

Iliad

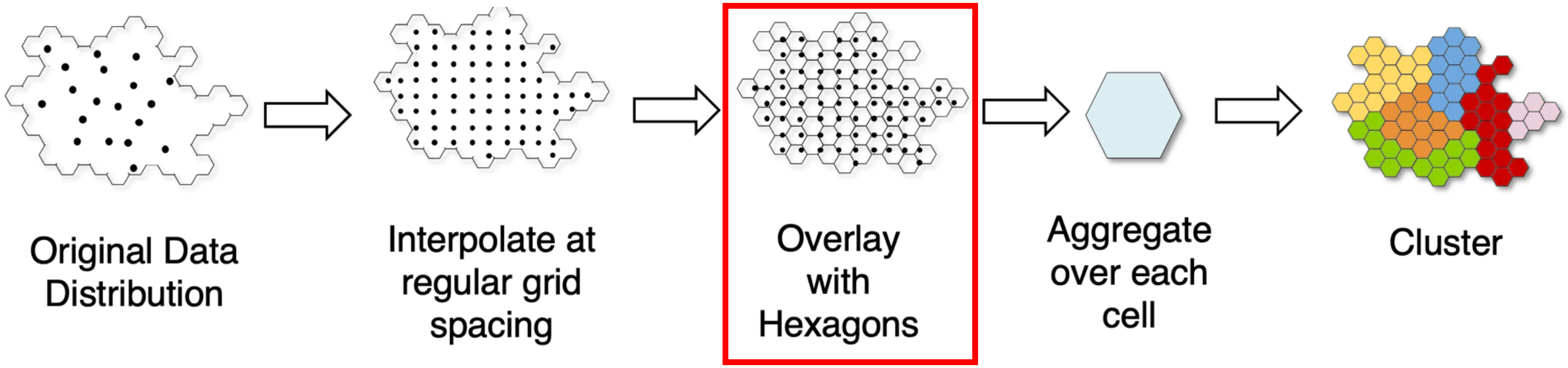

Original Data Distribution
Interpolate at regular grid spacing
Overlay with Hexagons
Aggregate over each cell
Cluster

Interpolation

- Used Inverse Distance Weighting for filling data gaps.
- *“Everything is related to everything else, but near things are more related than distant things.”*



$$L = \frac{10 / 200\boxed{^2} + 5 / 100\boxed{^2} + 25 / 500\boxed{^2}}{1 / 200\boxed{^2} + 1 / 100\boxed{^2} + 1 / 500\boxed{^2}}$$



Original Data Distribution

Interpolate at regular grid spacing

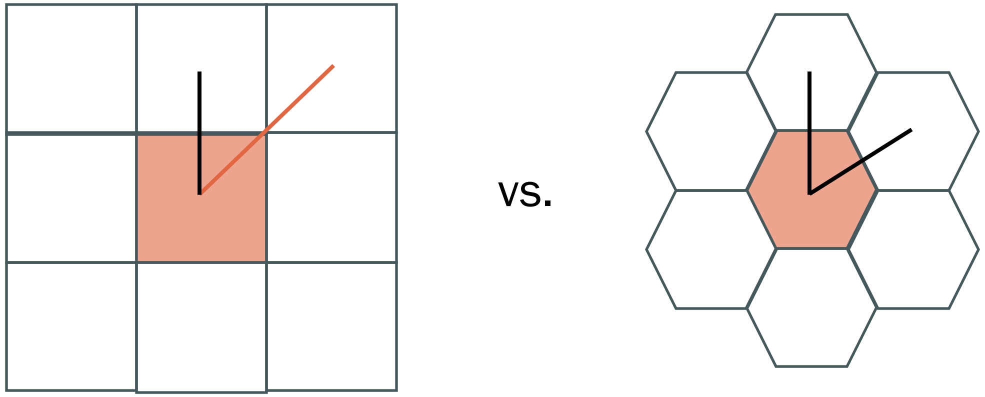
Overlay with Hexagons

Aggregate over each cell

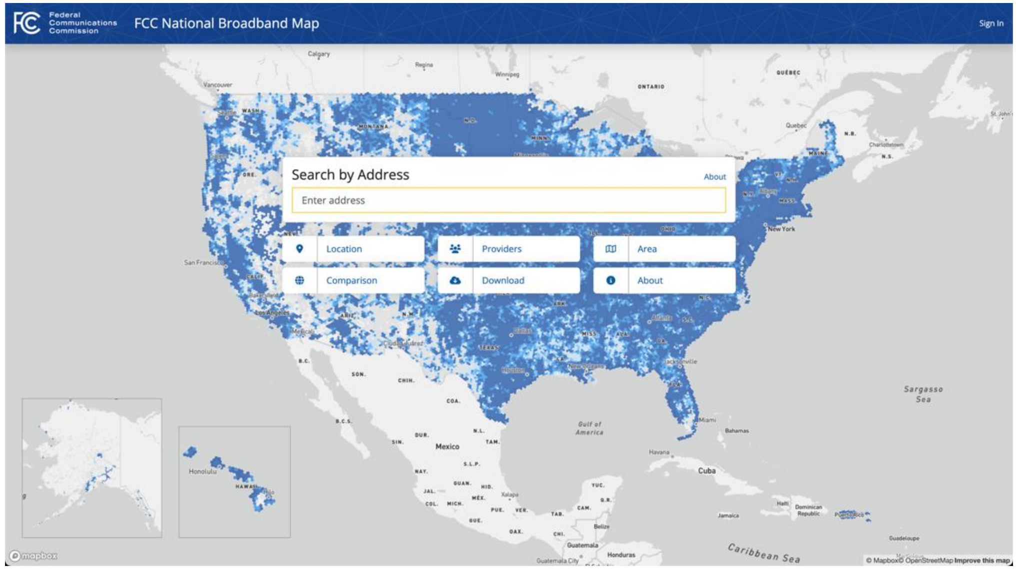
Cluster

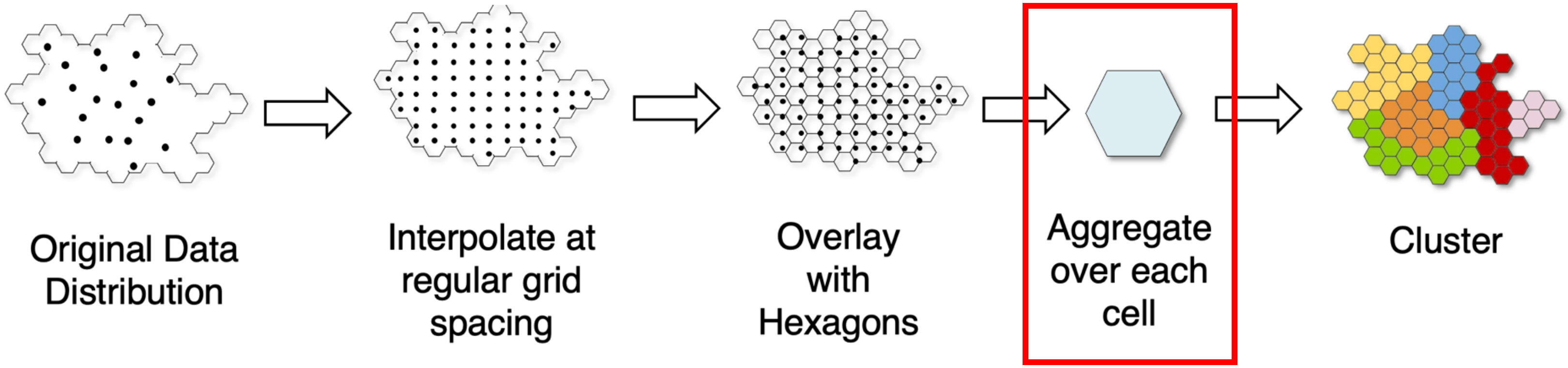
Hexagon Overlay

- Used hexagons because they tile a geography better than any other shape
- Federal Communications Commission (FCC) uses hexagons of resolution 8 for the national broadband map



vs.





Original Data Distribution

Interpolate at regular grid spacing

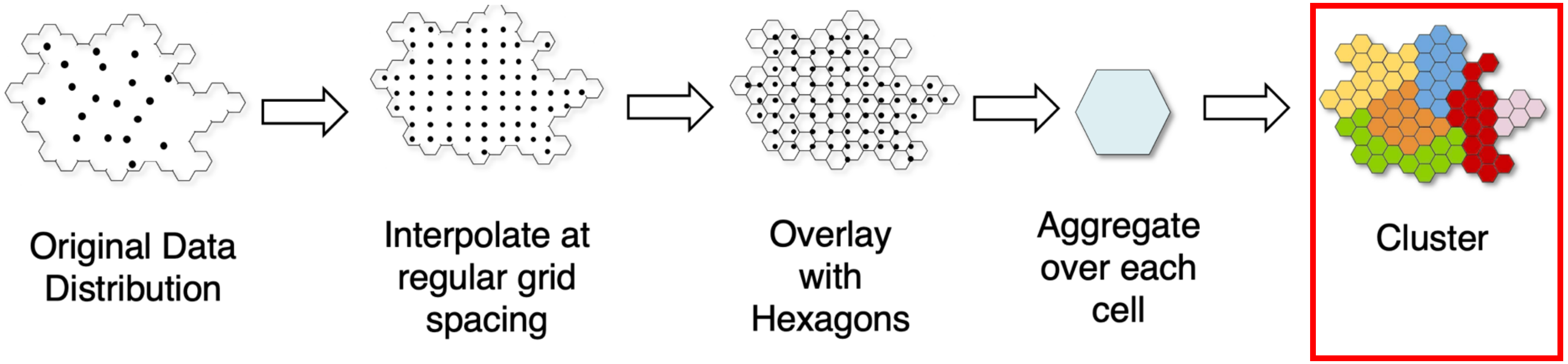
Overlay with Hexagons

Aggregate over each cell

Cluster

Aggregation

- Mean
- Percentiles
- Standard Deviation
- Inequality Ratio (p90 / p10)
- Latency Reduction (p90 – p10)



Original Data Distribution

Interpolate at regular grid spacing

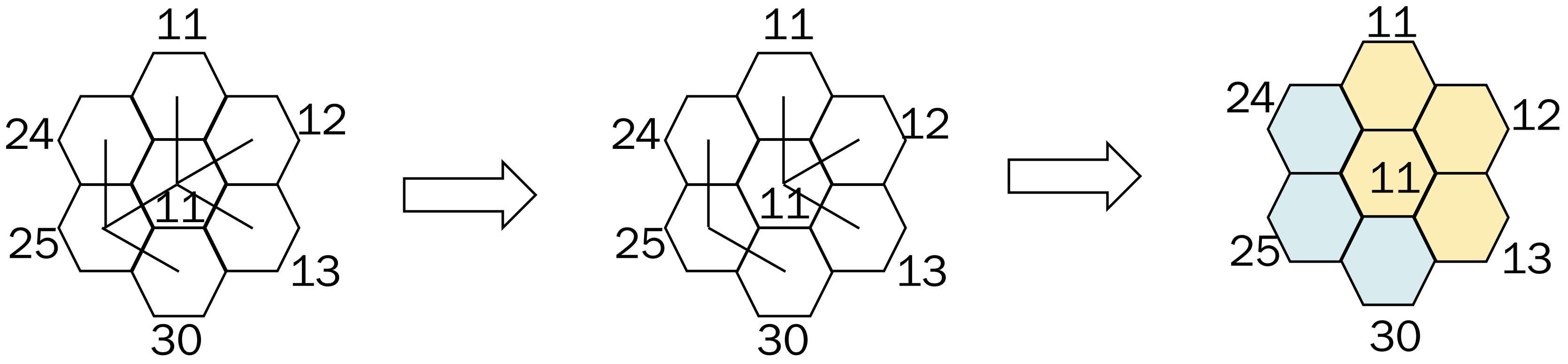
Overlay with Hexagons

Aggregate over each cell

Cluster

Clustering to Obtain Boundaries

Spatial ‘K’luster Analysis by Tree Edge Removal (SKATER)

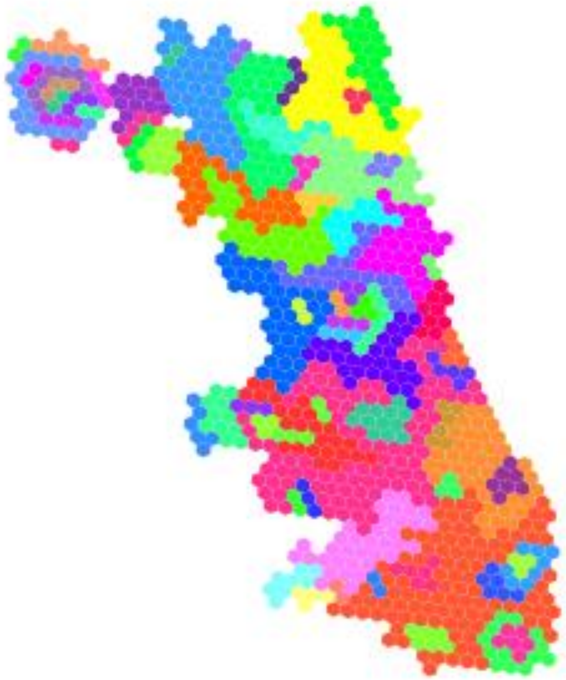


Minimum Spanning Tree Construction

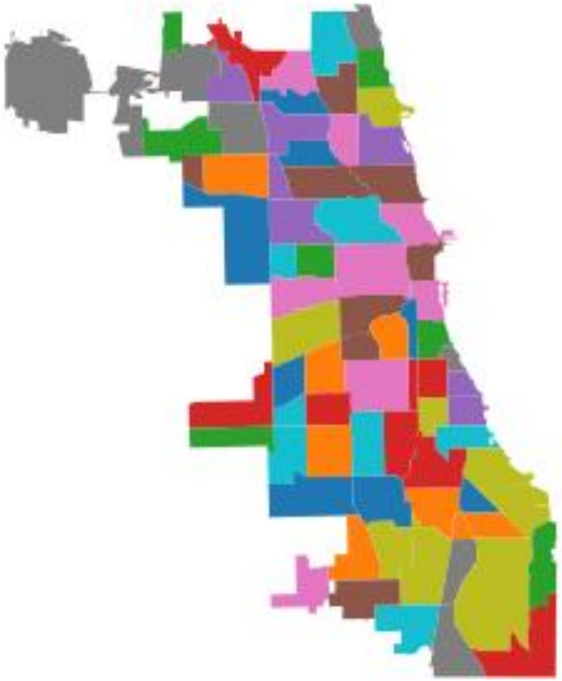
Iterative Edge Removal

Cluster Assignment

Data-driven Boundaries Cut Across Administrative Boundaries

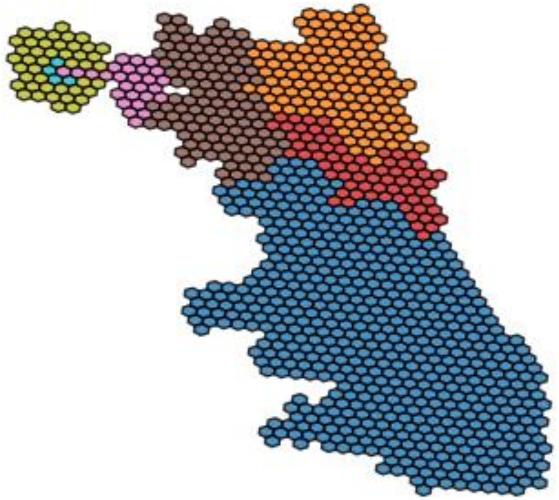


Data-driven boundaries

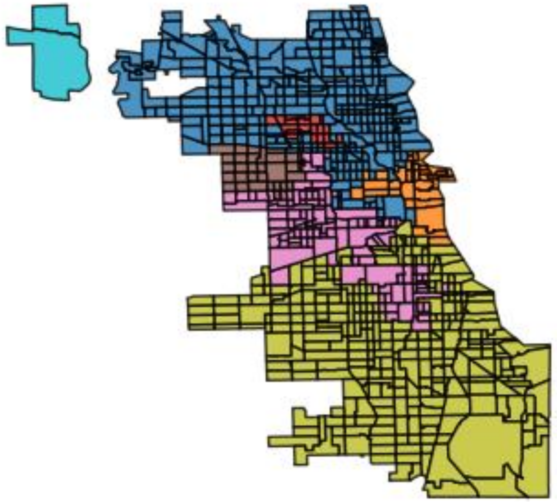


Neighborhood map of Chicago

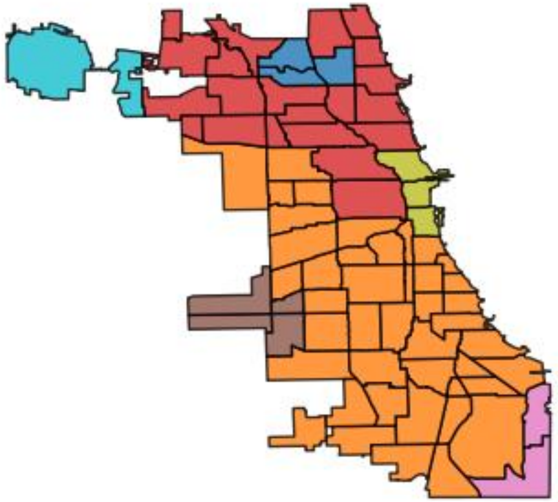
One Dataset, Three Units, Three Perspectives



Regular Hexagons



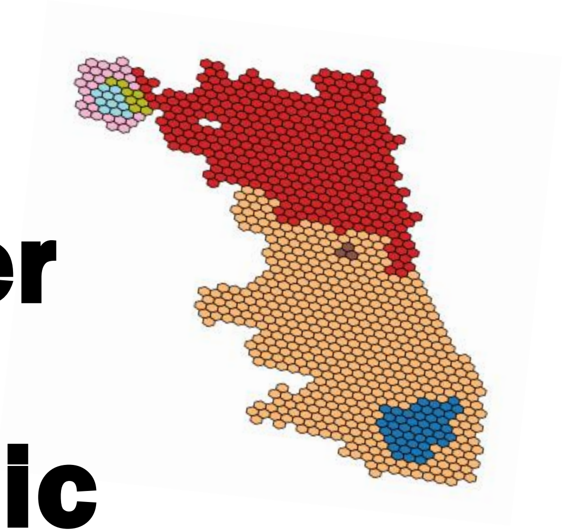
Census Tracts



Neighborhoods

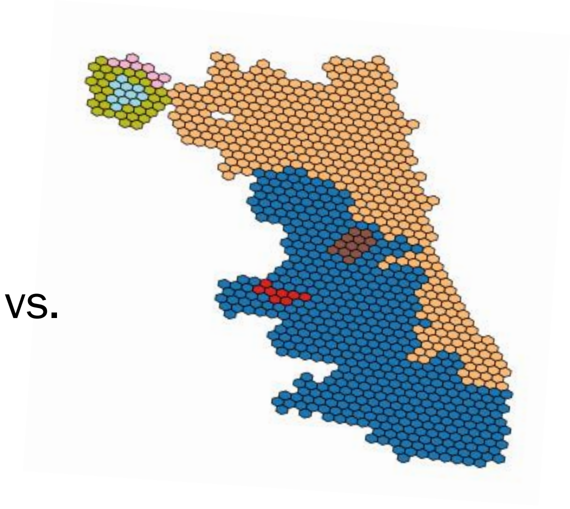
Temporal Cluster Stability as an Evaluation Metric

Stable clusters over time are likely to reflect meaningful, persistent patterns — not just noise.

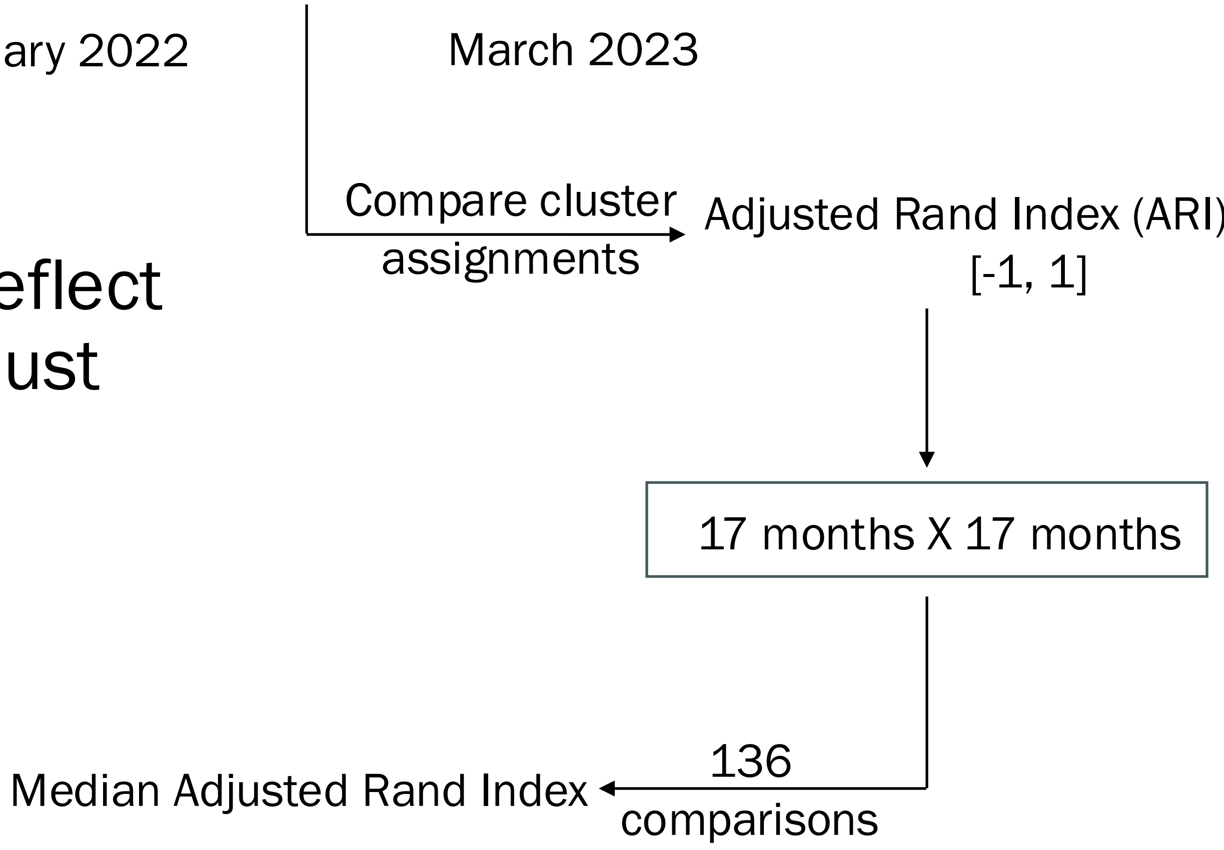


January 2022

vs.



March 2023



Compare cluster assignments

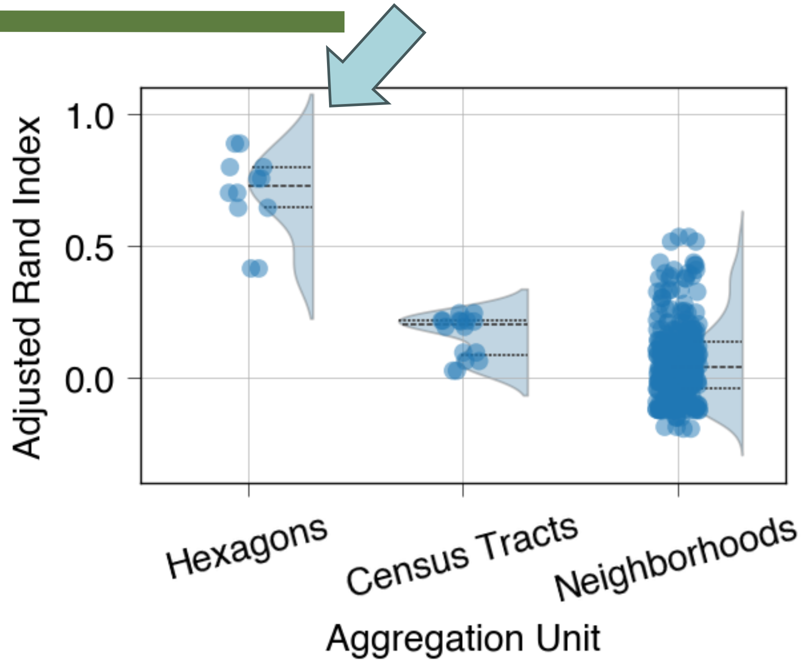
Adjusted Rand Index (ARI)
[-1, 1]

17 months X 17 months

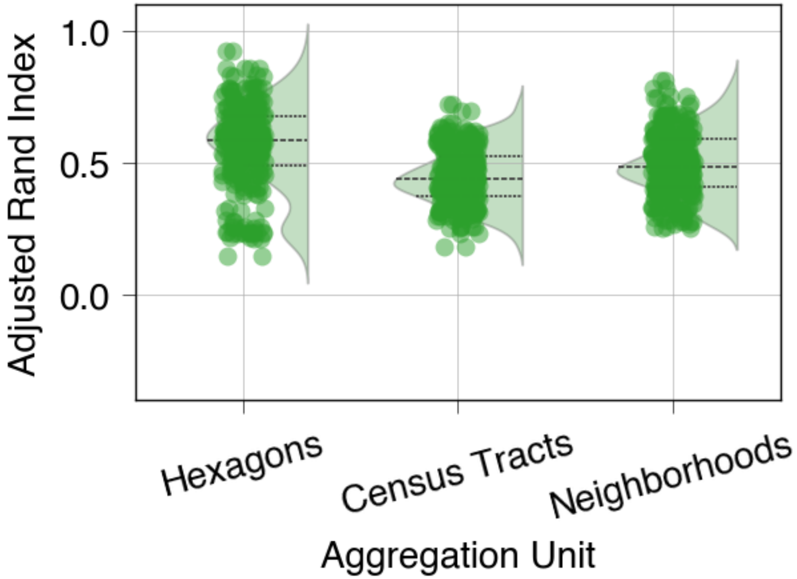
136 comparisons

Median Adjusted Rand Index

Interpolation Reduces Sensitivity Towards Spatial Unit Choice



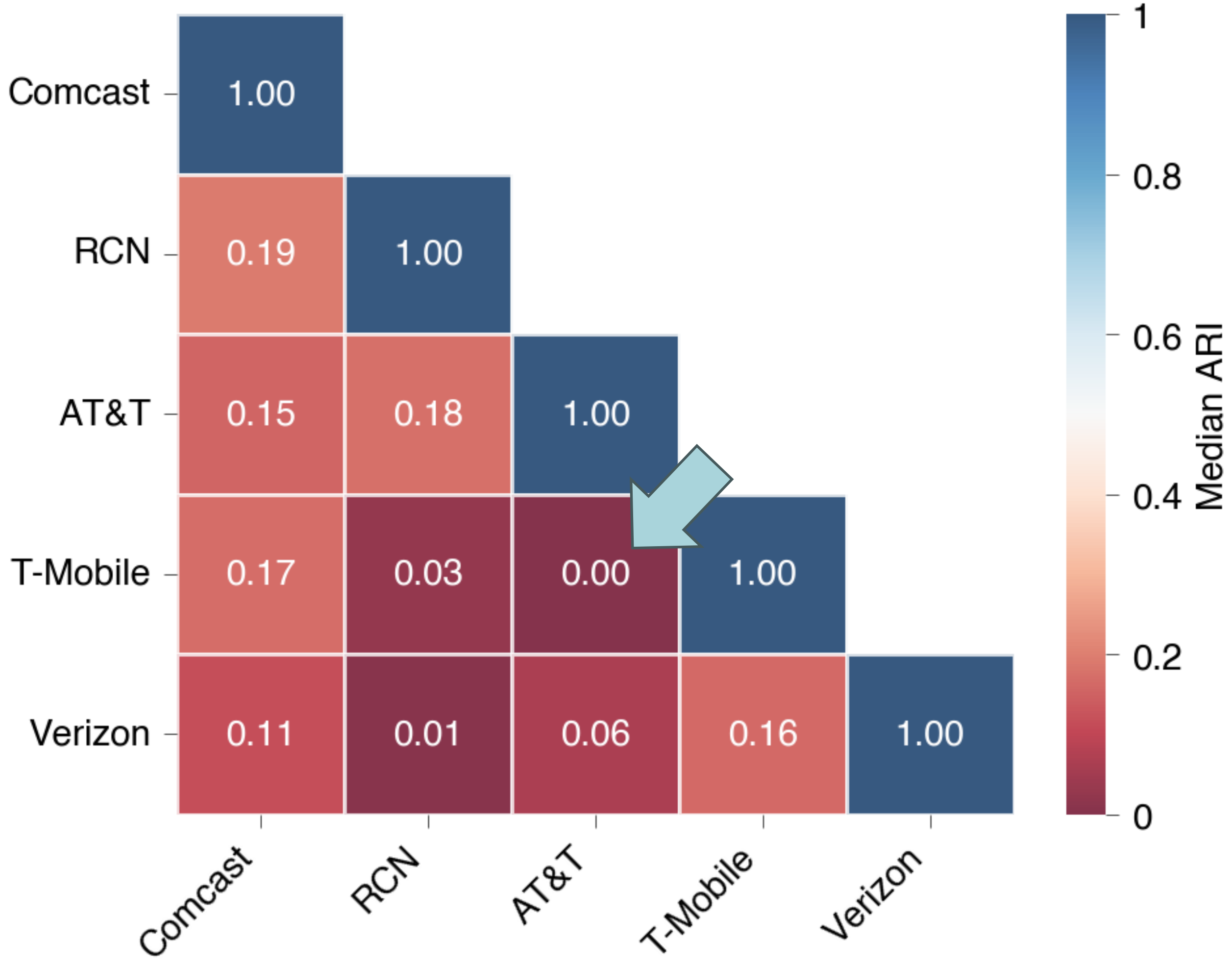
Raw Averaging



Interpolated Averaging

ISPs Disagree on Spatial Boundaries

Individually interpolated maps per ISP may be a more sensible sampling approach



	Comcast	RCN	AT&T	T-Mobile	Verizon
Comcast	1.00				
RCN	0.19	1.00			
AT&T	0.15	0.18	1.00		
T-Mobile	0.17	0.03	0.00	1.00	
Verizon	0.11	0.01	0.06	0.16	1.00

Takeaways

- Spatial interpolation methods can be extended to draw sampling boundaries for Internet latency
- Our approach allows for an adjusted rand index of 0.59, indicating a moderate to high stability between the boundaries
- Aggregating latency directly over administrative boundaries may not be the best approach
- Our approach can be used by ISPs for infrastructure planning and optimization