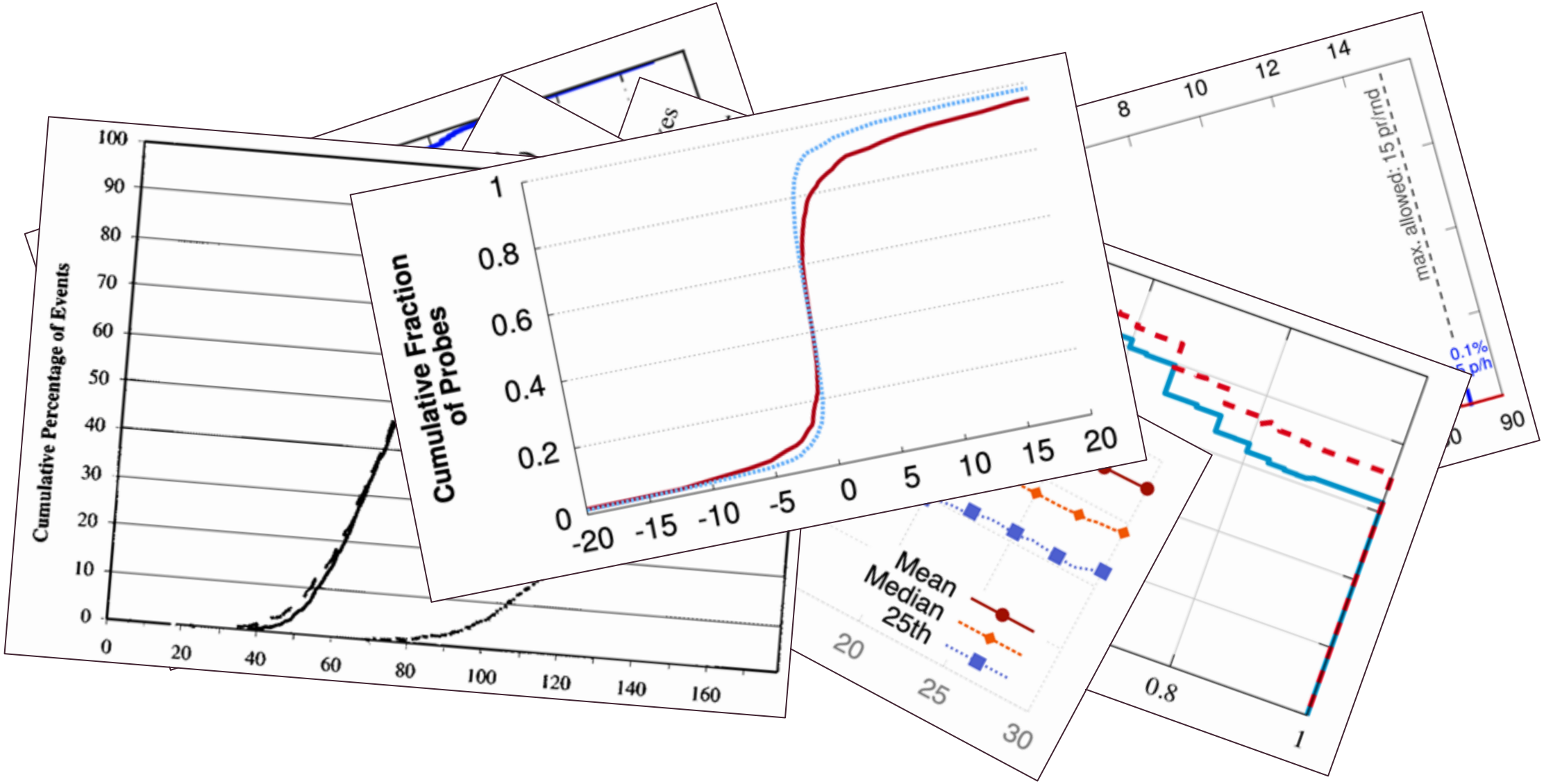
What's in the Dataset? Unboxing the APNIC per AS User Population Dataset

Loqman Salamatian
Calvin Ardi
Matt Calder
Vasilis Giotsas
Ethan Katz-Bassett
Todd Arnold

Madrid, ACM IMC 24

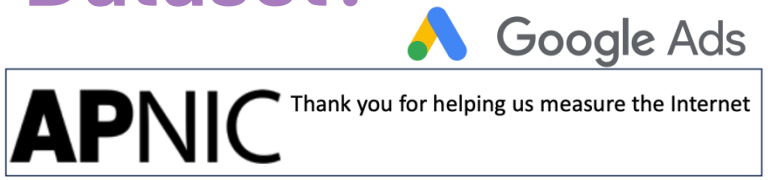


Courtesy of Tom Koch.

APNIC Dataset: Researchers have started using the APNIC per AS User Population dataset as a proxy for (i) user populations and (ii) traffic volume, but its reliability is unknown.

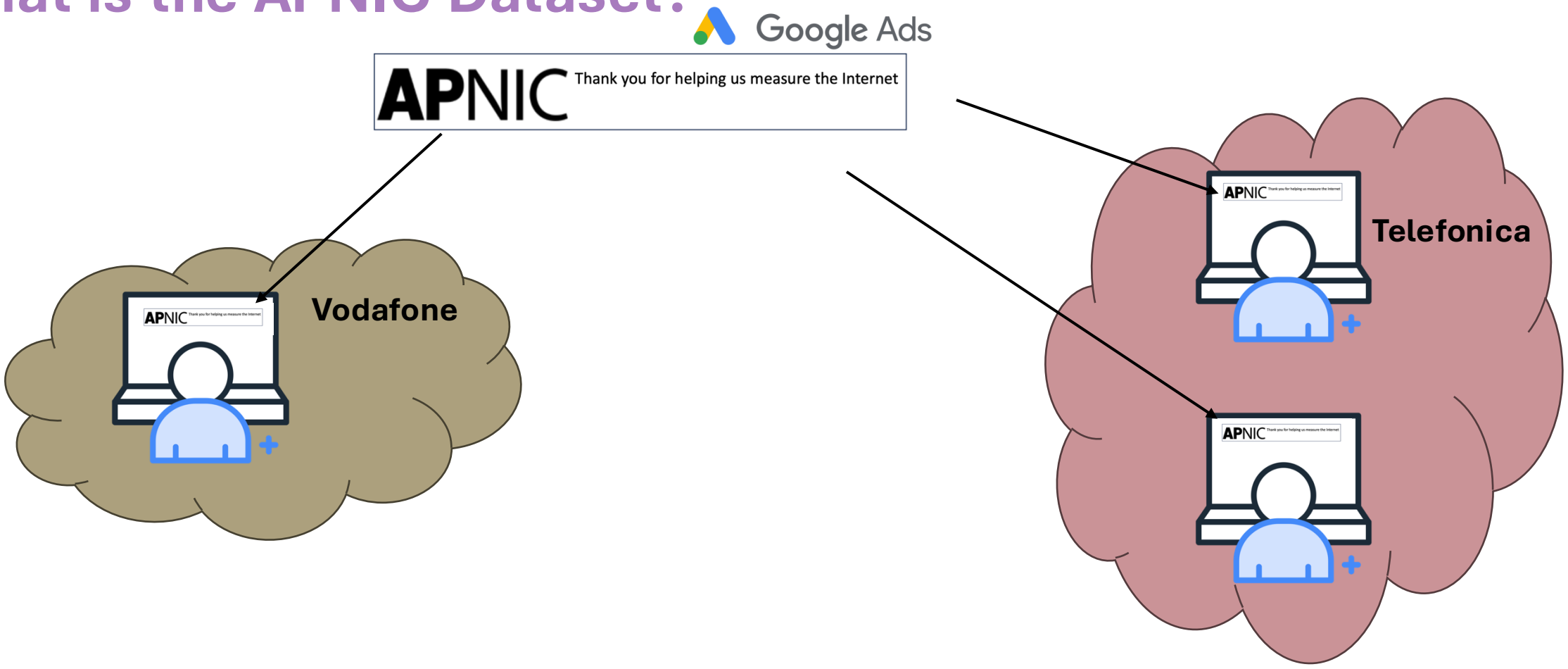
Goal of our paper:
We validate the APNIC dataset, assess its strengths and weaknesses, and improve its usability for Internet measurement research.

What is the APNIC Dataset?

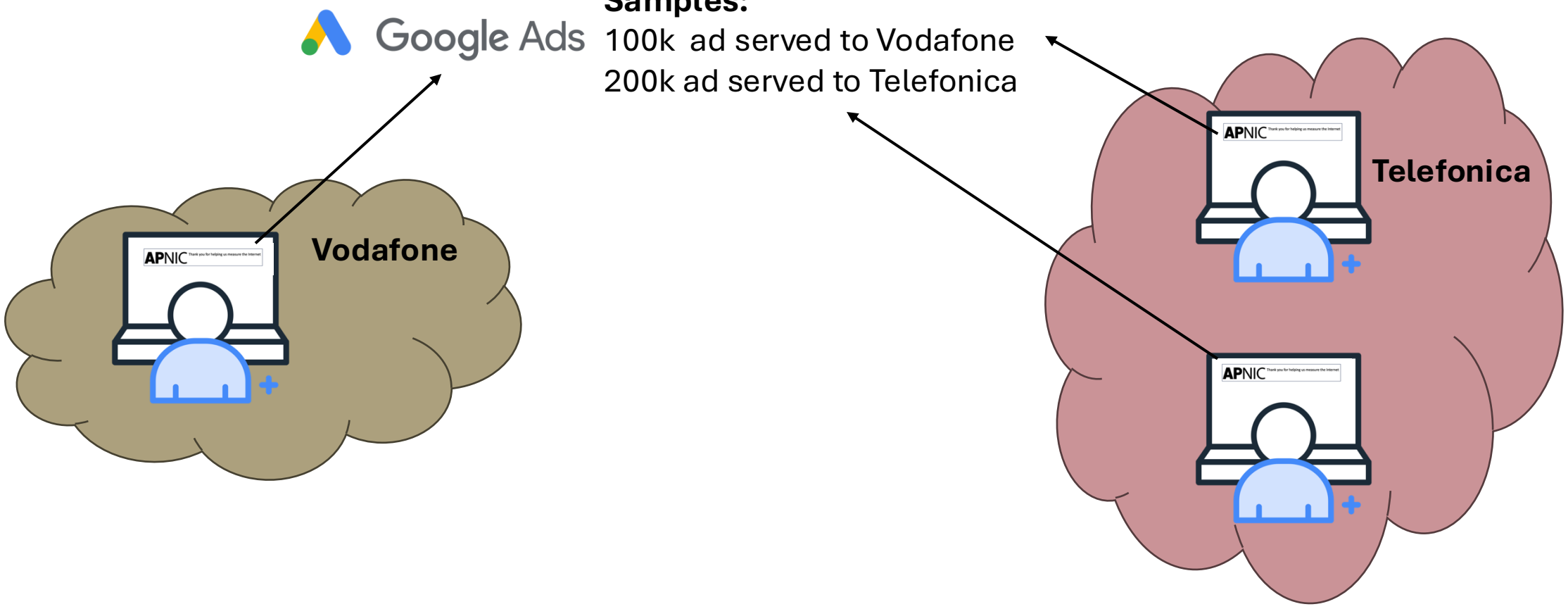


The ad being served.

What is the APNIC Dataset?



What is the APNIC Dataset?



What is the APNIC Dataset?

Samples:
100k ad served to Vodafone
200k ad served to Telefonica

Fraction of samples:
1/3 of ads were served by Vodafone.
2/3 of ads were served by Telefonica.

Total Internet population in Spain (ITU):
45.6M

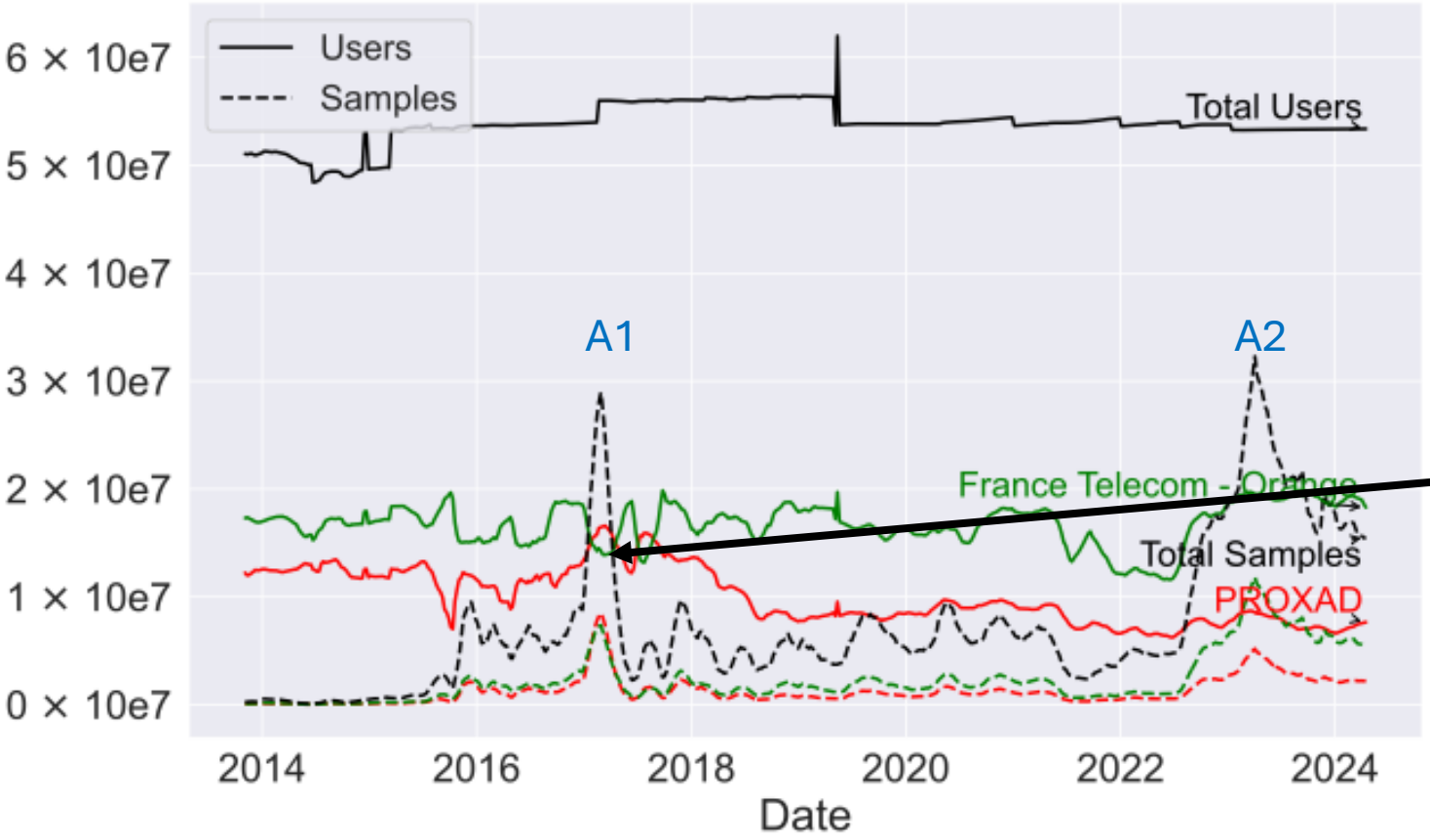
Final output:
User estimates
per AS in Spain

Vodafone: 1/3 of ads * 45.6M users in Spain = 15.2M
Telefonica: 30.4M

Some clear biases in the process

Non-Uniform Ad Placement: Google Ads has different levels of penetration in different countries or different types of networks, potentially leading to inaccurate user estimates where it is less prevalent.
Accuracy of ITU-T Estimates: Fluctuations in ITU-T's Internet user estimates can impact the APNIC dataset's accuracy.

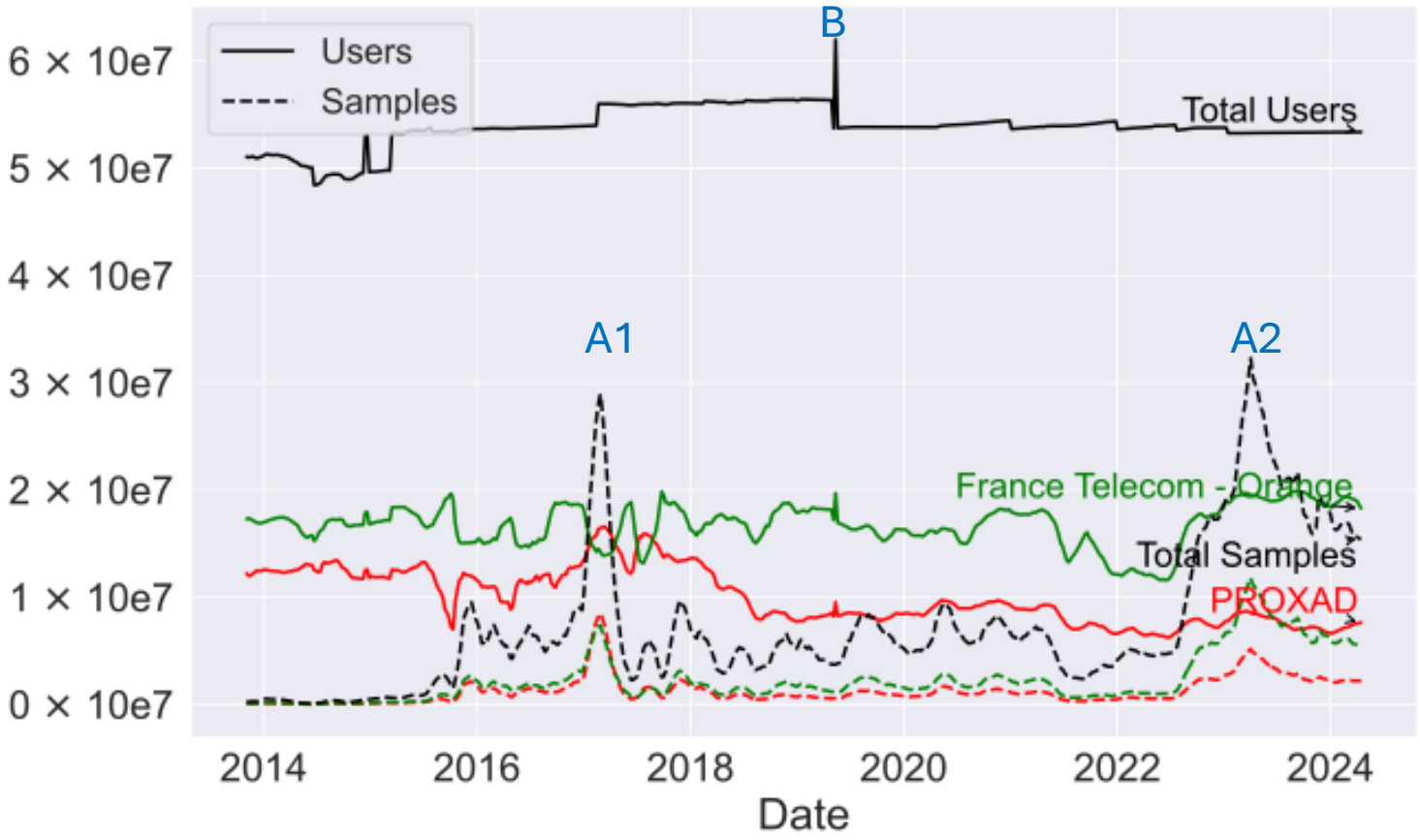
Highlighting Existing Biases and Variability



A) Sudden increases of all the Samples in France.

PROXAD appears to be the largest ISP in France for a couple of month

Highlighting Existing Biases and Variability



A) Sudden increases of all the Samples in France.
B) Two weeks increase of Total Users in the Country according to ITU

To understand the APNIC dataset accuracy, we collected four data sources:

Provider	Data	Availability
Global CDN	Distinct User Agents per ASN	Proprietary
	HTTP traffic per ASN	Proprietary
Manual survey	Broadband subscribers	Public
PeeringDB	Cumulative IXP peering capacity	Public
M-Lab	Speed tests per ASN	Public

Not discussed today

Validating APNIC with CDN User-Agents and Traffic Datasets

Validating APNIC with CDN User Agents and Traffic Datasets

Do the datasets agree on which organizations serve end-users?

Only **40% of Total Org** is seen in both datasets.

Validating APNIC with CDN User Agents and Traffic Datasets

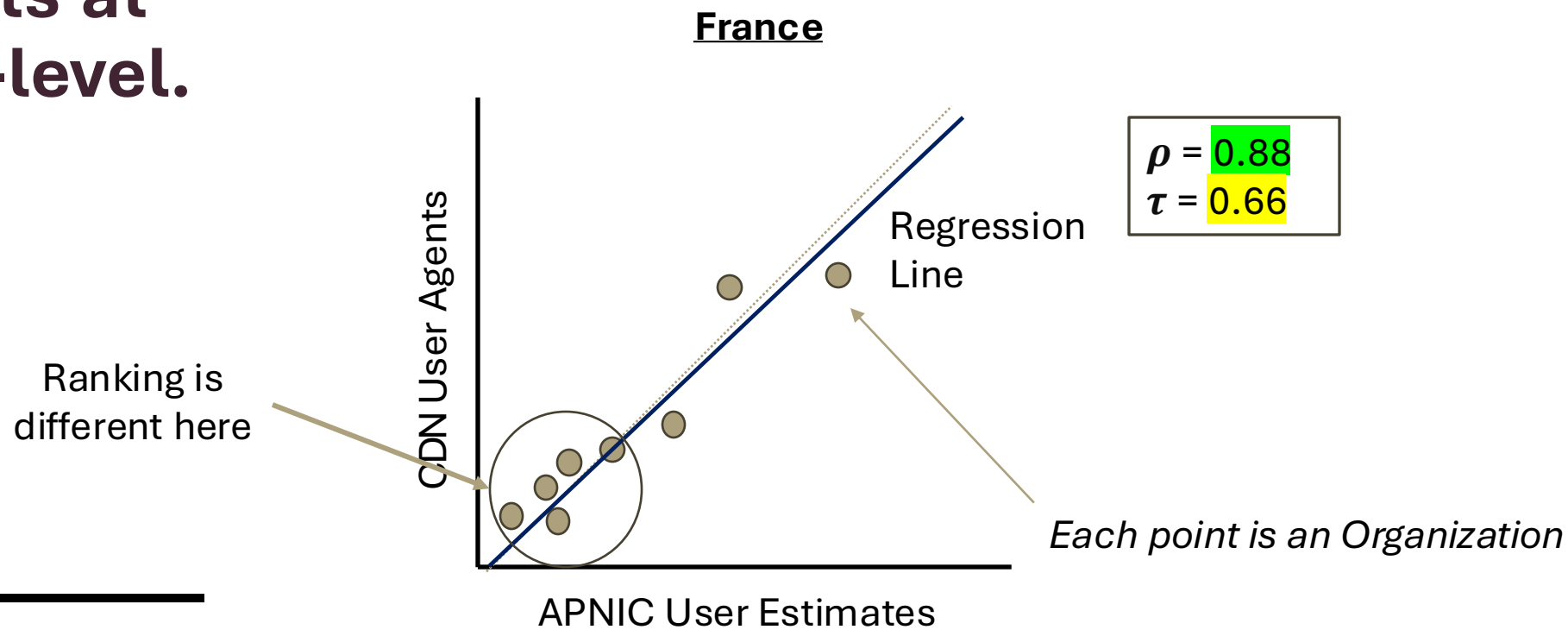
Do the datasets agree on which Orgs serve end-users?

Only **40% of total organization** is seen in both datasets, but the remaining ones are very small:

1. **APNIC** estimates **96% of Internet users** are in Org that host clients seen by the CDN.
2. **99% of CDN clients** are in organizations that APNIC says have Internet users.

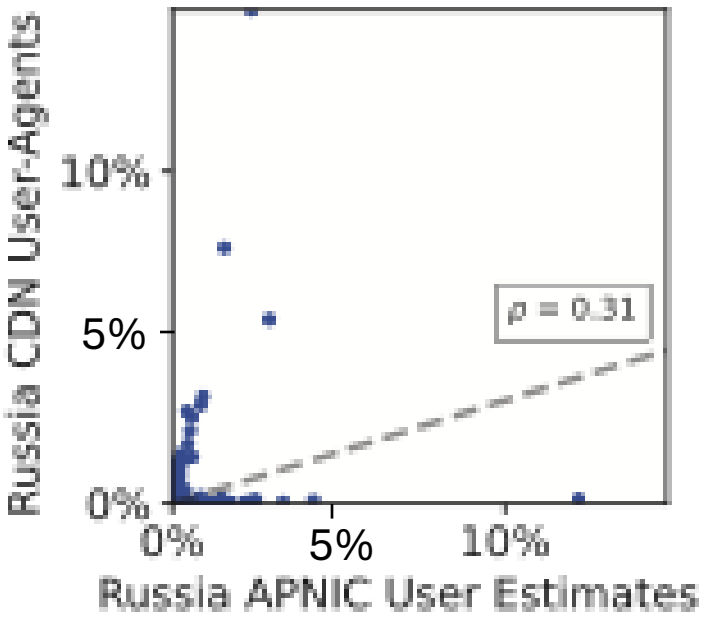
Evaluate the level of agreement between the APNIC and the CDN datasets at the country-level.

- **Pearson Correlation ρ:**
Measures the linear relationship between datasets, showing how well changes in one dataset correspond to proportional changes in the other.
- **Kendall-Tau Correlation τ:**
Assesses how well the two datasets align in ranking organizations.



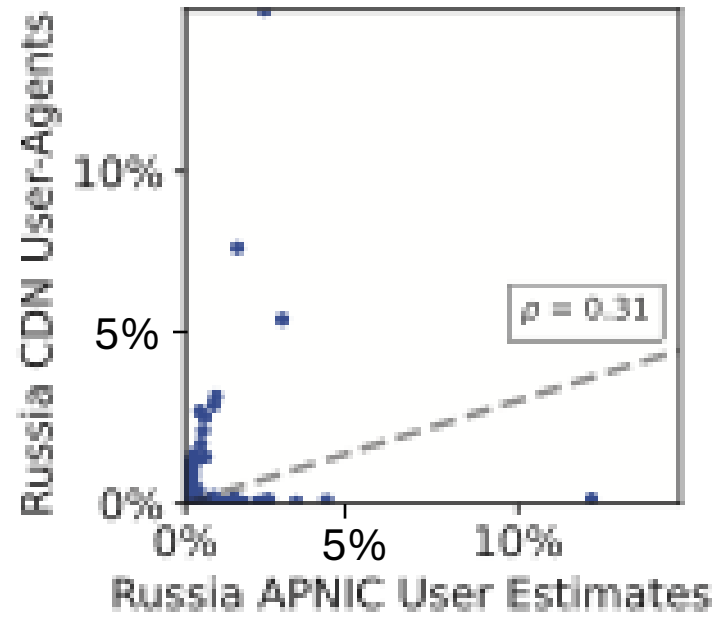
Understanding the Outliers

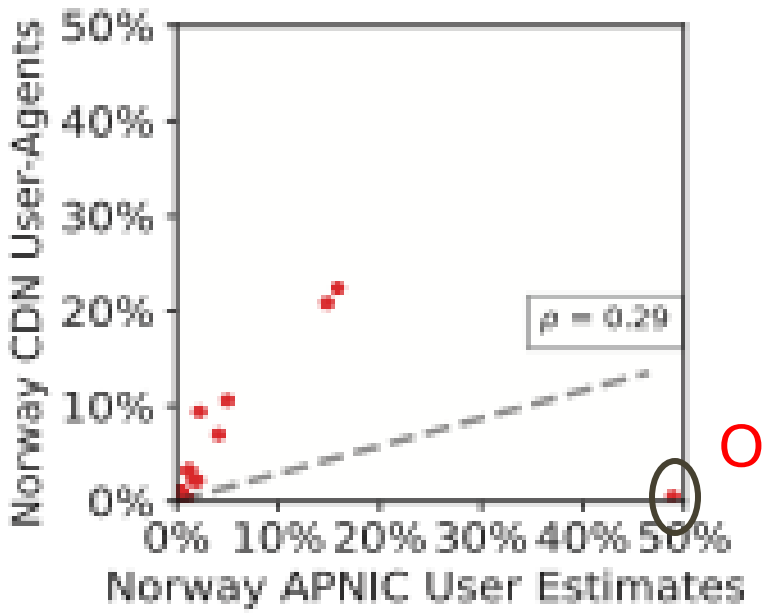
- **Russia:** Discrepancies are due to Yandex's market dominance, Russia's isolated Internet efforts, and Google's reduced presence following the Ukraine conflict.



Understanding the Outliers

- **Russia:** Discrepancies are due to Yandex's market dominance, Russia's isolated Internet efforts, and Google's reduced presence following the Ukraine conflict.
- **Norway:** Overrepresentation is caused by VPN traffic routing through a few IP addresses in Norway, leading to misinterpretation in the APNIC data.

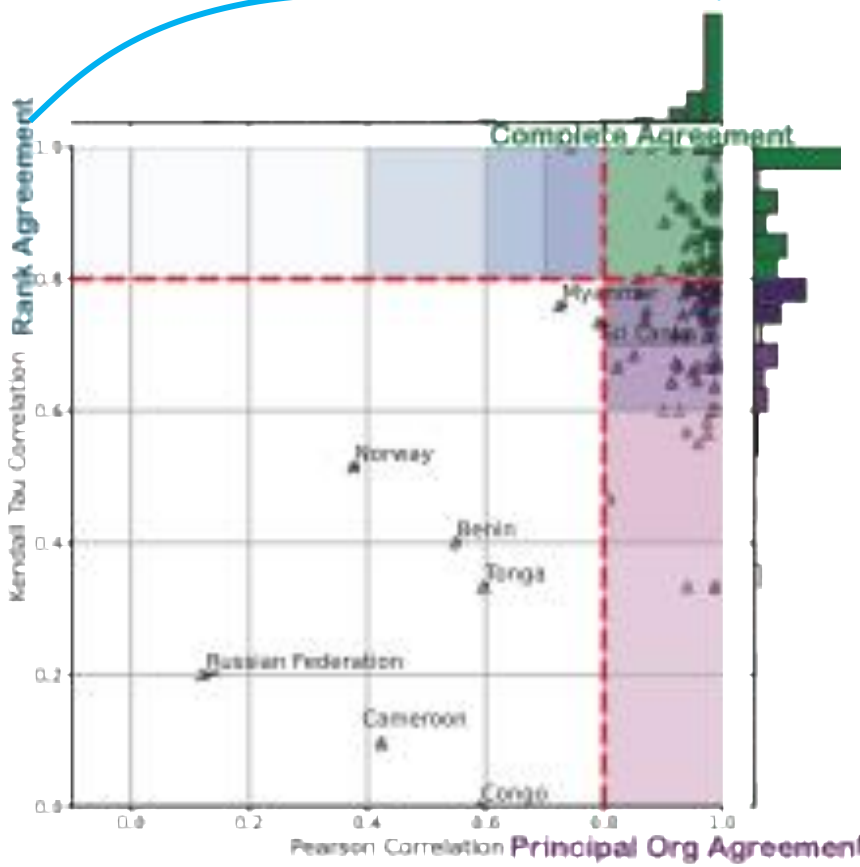




Opera VPN

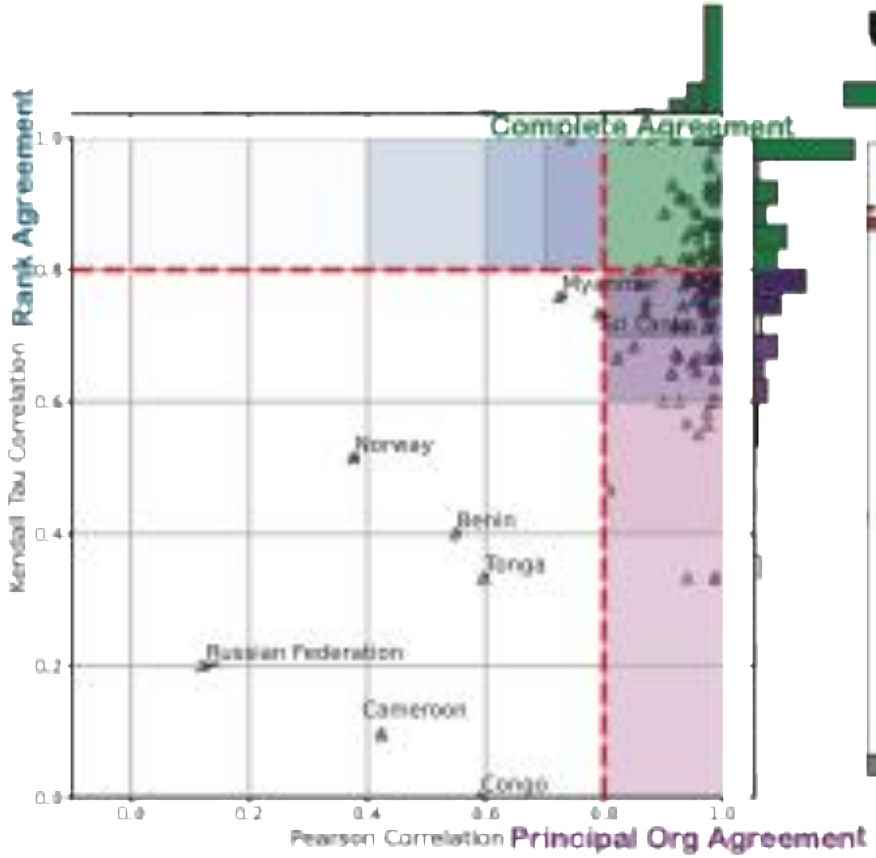
Validating APNIC with CDN's User at the Country-Level

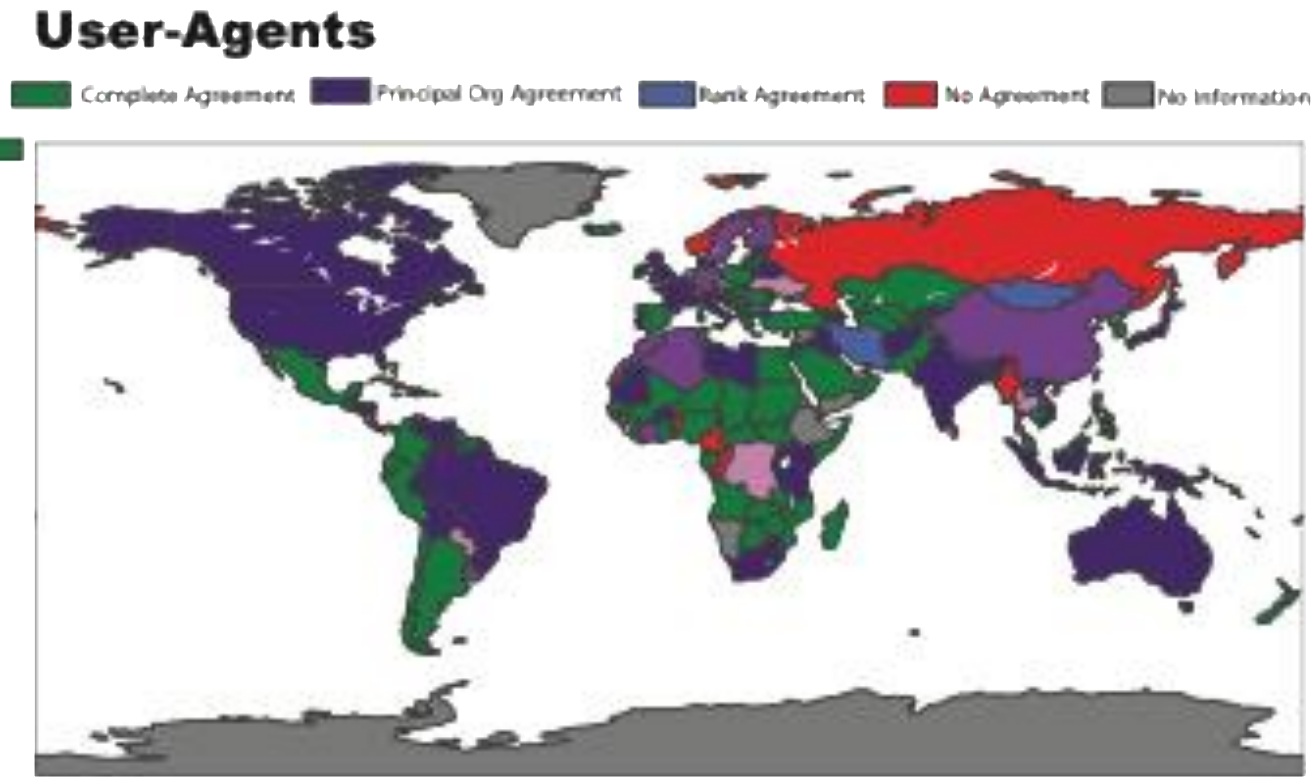
Both datasets identify similar organization order, even if their specific user estimates differ.



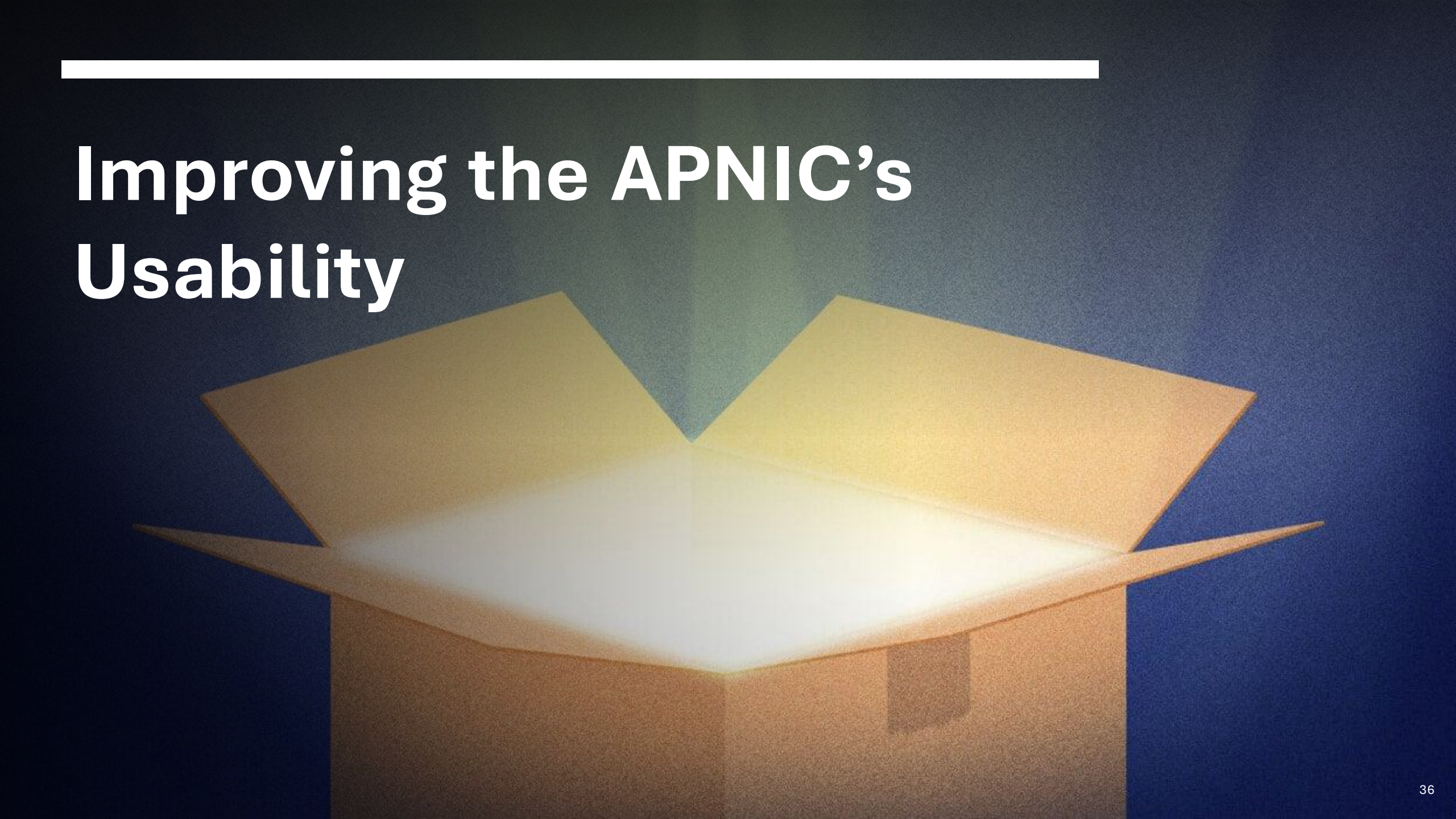
Agreement between the two datasets regarding the most significant networks within each country

Validating APNIC with CDN's User at the Country-Level





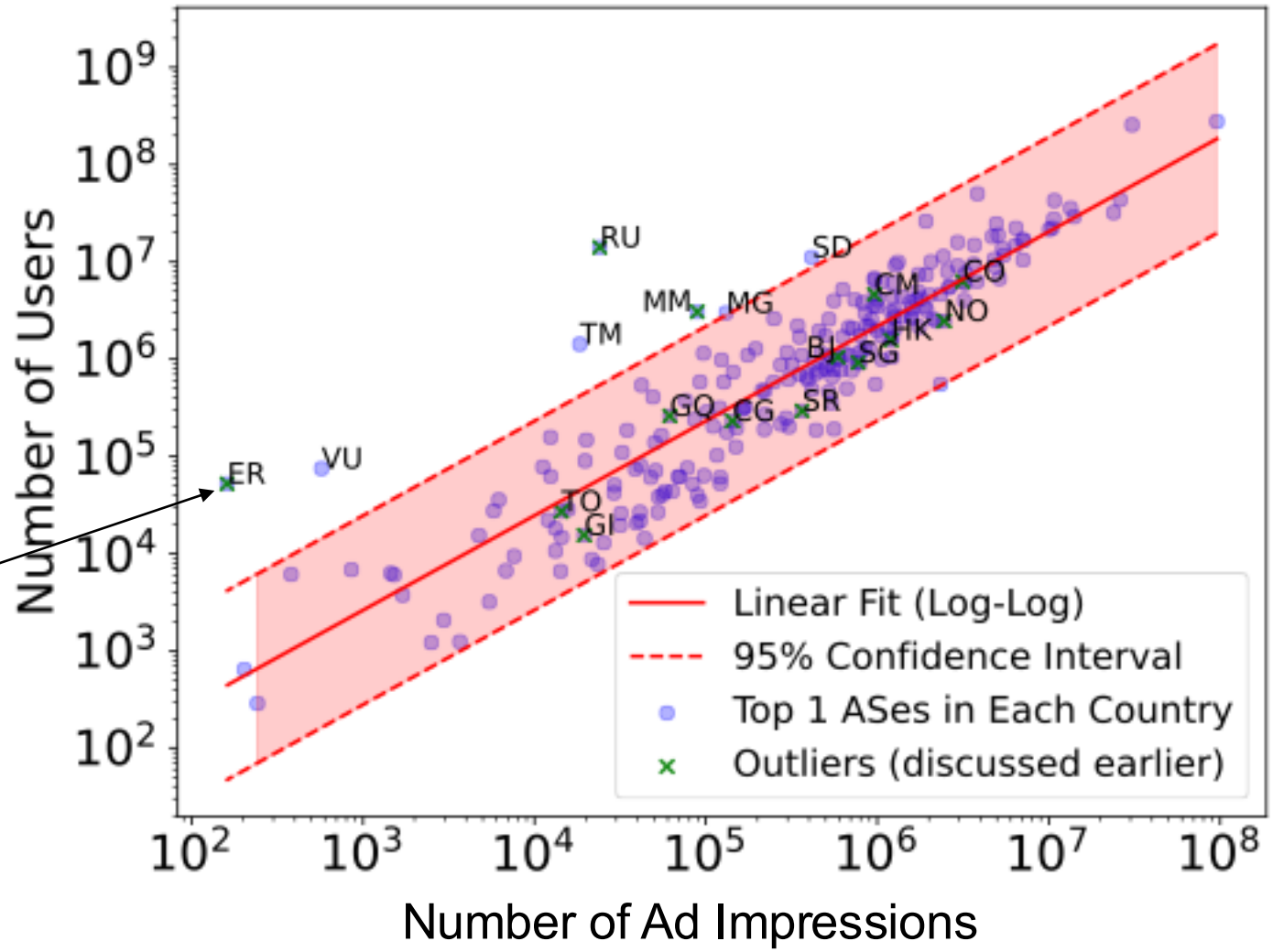
Improving the APNIC's Usability



Dissecting How User Estimates Are Computed

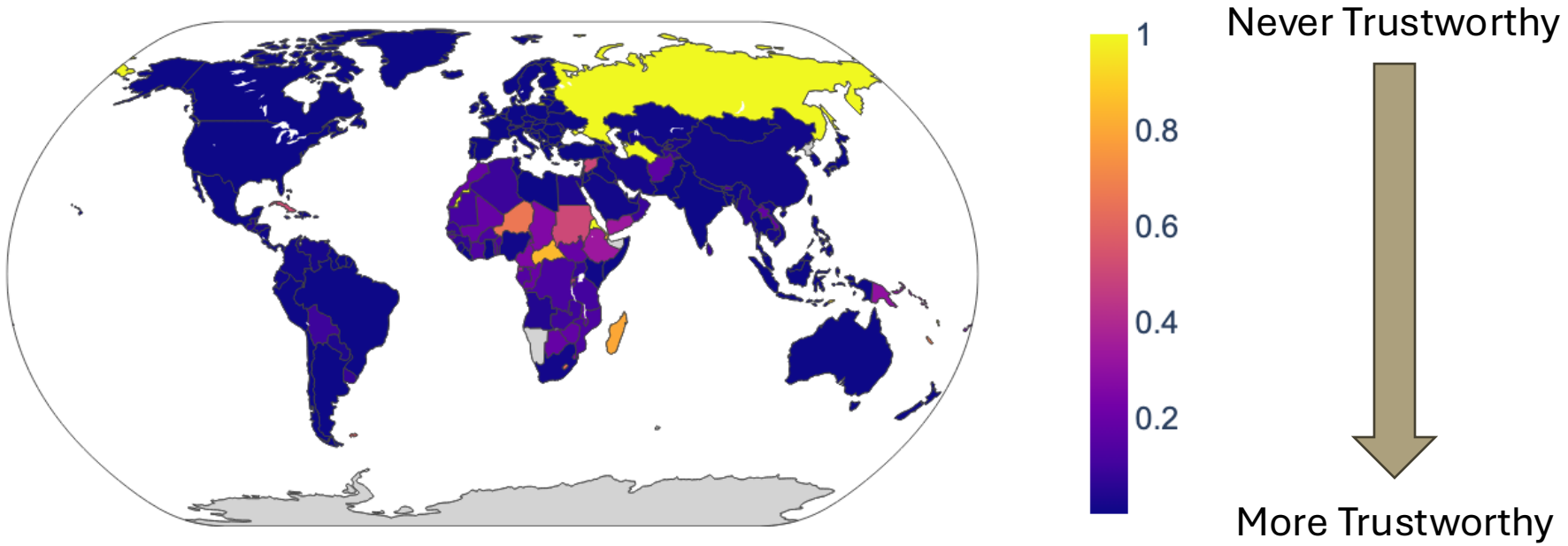
The elasticity coefficient β is the log-log linear coefficient.
Here: β = 0.98

Each sample point in Eritrea maps to ≈ 1000 users



A New Aggregation Mechanism to Improve APNIC's Accuracy

Fraction of days across 2024 where the User-to-Sample ratio did not fall in the confidence interval estimated

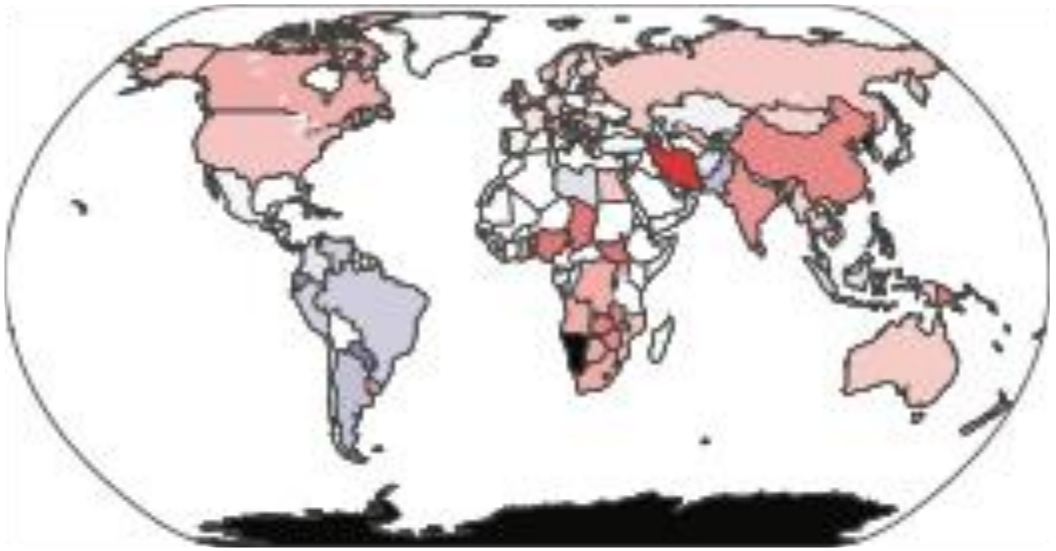


A New Aggregation Mechanism to Improve APNIC's Accuracy

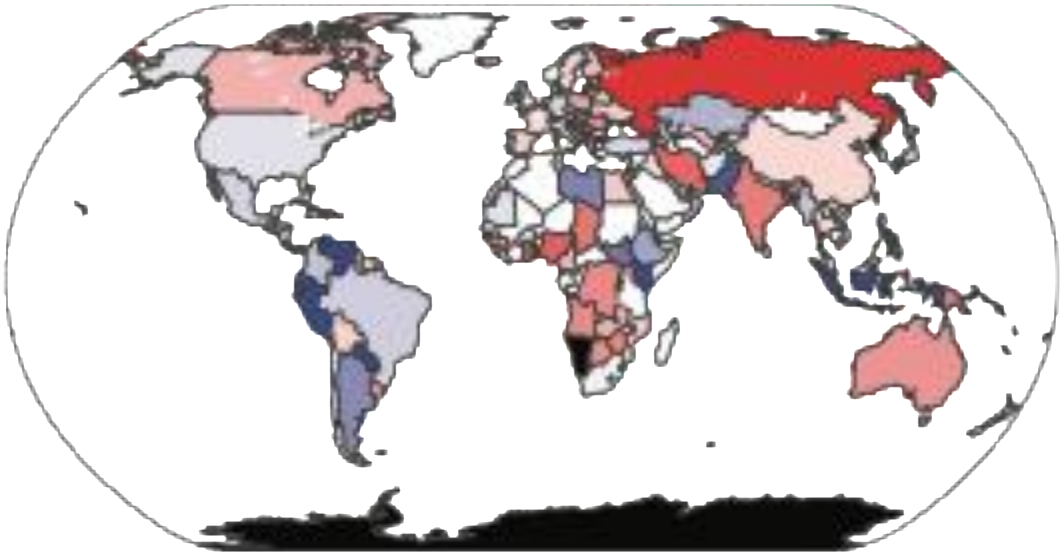
Recommendation:

For each country, select the day with the lowest users-to-samples ratio within a 60-day window.

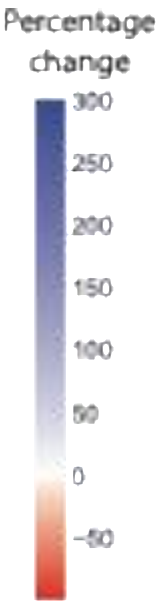
A new kind of consolidation never measured before!



(a) 2019 to 2021

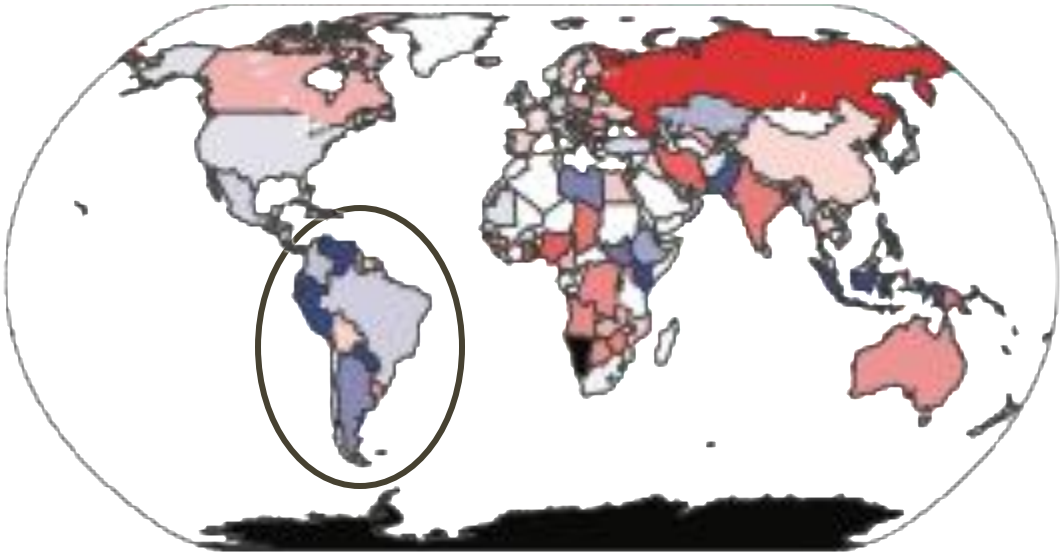


(d) 2019 to 2024

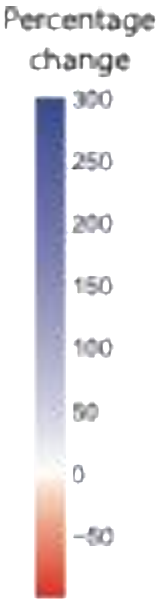


A new kind of consolidation never measured before!

(a) 2019 to 2021

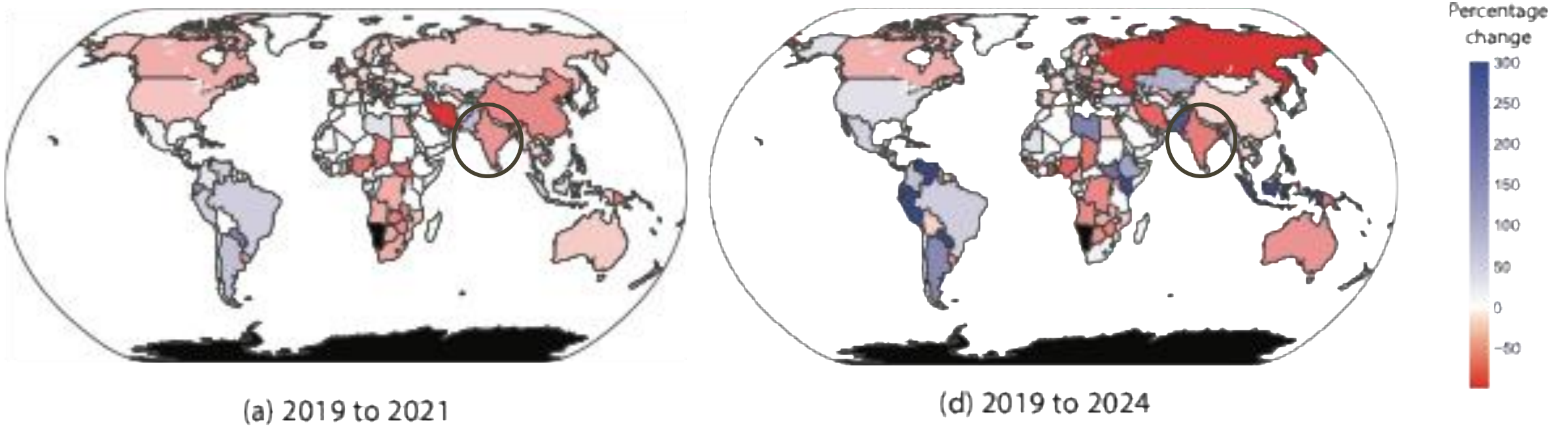


(d) 2019 to 2024



Latin America: Steady increase since 2019

A new kind of consolidation never measured before!

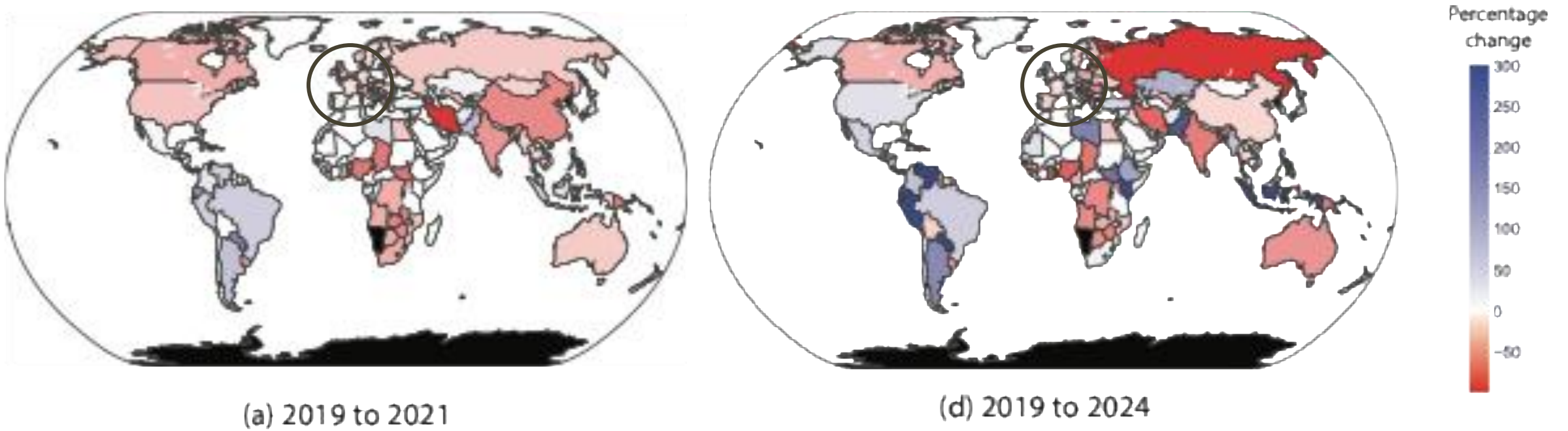


(a) 2019 to 2021

(d) 2019 to 2024

Latin America: Steady increase since 2019
India: Stronger concentration from 2021 onwards.

A new kind of consolidation never measured before!

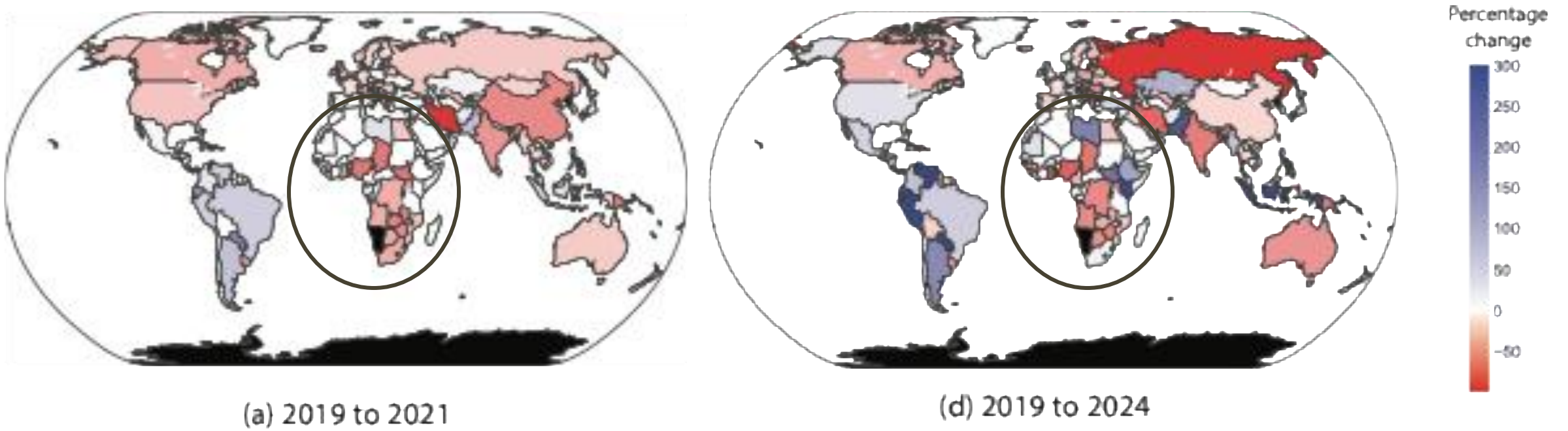


(a) 2019 to 2021

(d) 2019 to 2024

Latin America: Steady increase since 2019
India: Stronger concentration from 2021 onwards.
Europe: Small decline since 2019.

A new kind of consolidation never measured before!



(a) 2019 to 2021

(d) 2019 to 2024

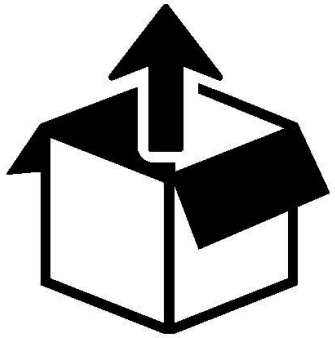
Latin America: Steady increase since 2019
India: Stronger concentration from 2021 onwards.
Europe: Small decline since 2019.
Africa: Split between increase (West Africa) and decrease (Sub-Saharian)

Conclusion

APNIC dataset works well in countries with sufficient Google Ads data. We share the list of countries where we found these results to be trustworthy.

Accuracy improves by verifying the user-to-sample ratio (and ensuring it aligns with external datasets *discussed in the paper*).

Feel free to look at our repo!



Back-up Slides