Self-Validating Uncertainty

Or: how to do measurement when you cannot afford ground truth

Fabián E. Bustamante
Northwestern University
Evanston, USA

Santiago Klein
Northwestern University
Evanston, USA

Caleb Wang
Northwestern University
Evanston, USA

Kedar Thiagarajan
Northwestern University
Evanston, USA

Ying Zhang
Northwestern University
Evanston, USA

Abstract

Many measurement systems operate where ground truth is scarce, expensive, and systematically biased toward what can be observed. Yet validation remains tied to individual correctness, creating a bottleneck when inferences cannot be independently verified at scale.

We propose *self-validating uncertainty*: measurement systems should expose reliability signals whose meaning and limitations can be independently assessed. Rather than relying exclusively on individual correctness, validation can include *empirical reliability estimation*, assessing how correctness varies with observable properties of an inference.

Internet infrastructure mapping illustrates why the source of uncertainty matters. Multiple submarine cables may underlie a single observed network link, while incomplete evidence may prevent us from determining which cables are involved. Distinguishing physical multiplicity from incomplete knowledge changes what can meaningfully be validated.

A ground-truth-withholding experiment demonstrates the *validation dividend*: limited ground truth can support group-level reliability estimates covering far more inferences than can be individually verified. However, we also show that ground-truth availability is systematically biased, and that statistical precision alone does not guarantee that reliability estimates transfer to the broader population.

CCS Concepts

• **Networks → Network measurement**.



This work is licensed under a Creative Commons Attribution 4.0 International License.

HotNets '26, Salt Lake City, UT, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-3029-0/2026/11
https://doi.org/10.1145/3847290.3859966

Keywords

Internet measurement, partial observability, self-validating uncertainty, reliability estimation

ACM Reference Format:
Fabián E. Bustamante, Santiago Klein, Caleb Wang, Kedar Thiagarajan, and Ying Zhang. 2026. Self-Validating Uncertainty: Or: how to do measurement when you cannot afford ground truth. In *The 25th ACM Workshop on Hot Topics in Networks (HotNets '26), November 16–17, 2026, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3847290.3859966

1 Introduction

Measurement is limited not by data collection but by validation. Measurement systems can gather observations at Internet scale, yet establishing whether the resulting inferences are correct requires ground truth that is sparse, expensive, or fundamentally unavailable [16–18], since operators rarely disclose the internal state we seek to infer. Worse, ground-truth availability is systematically shaped by what can be observed [5]. As a result, many measurement studies face a dilemma. They can either make claims that extend beyond what can be verified, or restrict themselves to the small, potentially unrepresentative subset where ground truth exists.

This problem is especially acute in Internet infrastructure mapping, but it extends more broadly across Internet measurement. Studies routinely infer locations, dependencies, ownership relationships, and routing behavior that cannot be directly observed [7, 8, 14–16, 19, 20]. Yet validation typically targets individual correctness, requiring ground truth aligned with each claim. We can often generate far more inferences than we can confidently validate.

We argue that this bottleneck reflects a broader problem: *partial observability*. The systems we study expose only fragments of their internal state, and different realities may be consistent with the same observations. Ground truth may be scarce, biased, or inaccessible, but even when available, it need not resolve the ambiguity left by those observations.

This calls for broadening what we validate. Correctness is not the only thing that can be assessed. A complementary objective is *empirical reliability estimation*: determining how correctness varies with observable properties of an inference, and identifying where available evidence cannot support such estimates. Statistical calibration provides one way to assess reliability, by testing whether reported probabilities agree with observed correctness frequencies [9]. More generally, reliability signals need not represent probabilities; their relationship to correctness can be estimated empirically, provided the labeled observations support inference to the population of interest.

This shift requires a different kind of output. Rather than emitting only an answer, a system must also expose information about its reliability. Similar ideas have emerged independently in distributed systems and machine learning. The ϕ-accrual failure detector reports a suspicion level rather than a binary verdict [11]. TrueTime returns a bounded interval rather than a timestamp [4]. Conformal prediction provides prediction sets with coverage guarantees under specified assumptions [2]. These examples make different reliability claims, but share a principle: systems should expose information whose meaning and limitations can be independently assessed. We call this design principle *self-validating uncertainty*: measurement systems should expose not only their inferences, but also the evidence, assumptions, and uncertainty needed to assess and challenge their reliability claims. This does not mean that systems can certify their own correctness or reveal every limitation.

Internet measurement adds a requirement when uncertainty has different sources, particularly in infrastructure mapping. Virtualization can hide multiple physical hops or submarine cables behind a single observed network link. A set of candidate cables may therefore represent multiple physical relationships, uncertainty about which cables are involved, or both. The distinction matters because validating multiple actual relationships is different from resolving an unknown assignment. An output should therefore distinguish physical multiplicity from limitations of the available evidence when that distinction can be established.

This paper makes three contributions. First, we propose self-validating uncertainty as a design principle for measurement systems operating under partial observability: outputs should expose reliability signals whose meaning and limitations can be independently assessed. Second, using traceroute geolocation and submarine-cable mapping, we illustrate how reliability signals and the source of uncertainty determine what can meaningfully be validated. Third, through a ground-truth-withholding experiment, we quantify the *validation dividend*: how limited ground truth can support group-level reliability estimates for substantially more inferences than can be individually verified. We also examine how path-length-dependent validation introduces selection bias and assess whether reliability estimates transfer across path lengths, illustrating why statistical precision alone does not establish reliability for the broader population.

Table 1: Examples of systems that expose information about the reliability of their outputs. Each supports a different claim and requires different evidence to assess it.

Example	Output	What can be checked
ϕ-accrual	Suspicion score based on heartbeat behavior	Whether observed delays agree with the assumed model
TrueTime	Interval bounding the true time	Whether the clock system maintains the conditions required for the bound
Geolocation	Physical consistency and path–model alignment scores	Relationship between scores and observed geolocation errors
Cable mapping	Set of candidate cables	Candidate relationships against independent evidence; limits of what the set establishes

2 Forms of Self-Validating Output

Self-validating uncertainty does not prescribe a single measure of confidence or a common validation procedure. Instead, it asks what information about an output's reliability can be exposed, what that information means, and how its meaning can be independently assessed. We consider four examples that make different reliability claims, each with its own assumptions and validation requirements.

Table 1 summarizes these examples. Their mechanisms differ, but each exposes information that can be assessed without independently verifying every output.

A suspicion score (failure detection). A failure detector decides whether a remote process is still alive, and the usual answer is a bare verdict—up or down. The ϕ-accrual detector instead reports a suspicion level derived from the distribution of heartbeat arrivals, leaving the application responsible for choosing when to act [11]. The score has a probabilistic interpretation: it measures how unlikely the observed heartbeat delay would be under the model of normal operation.

That interpretation is checkable. One can test whether observed heartbeat delays follow the assumed distribution, without knowing whether every delayed heartbeat indicates an actual failure. This does not establish the probability that a process has failed, but it gives the score a meaning that can be evaluated independently of individual failure verdicts.

A sound bound (distributed time). TrueTime replaces a timestamp with an interval guaranteed to contain the true time, provided its clock-system assumptions hold [4]. Correctness comes from respecting the width of the interval rather than trusting its midpoint. Unlike the probabilistic interpretation of ϕ-accrual, this is a conditional bound: the true instant lies within the reported interval as long as the underlying guarantees are maintained.

This changes what must be established. Spanner need not know the exact time to reason correctly about the interval. Instead, the system must maintain the conditions under which the bound is valid. Reliability is expressed through the interval and the assumptions supporting it.

A gated score (geolocation). Router geolocation traditionally returns a single location inferred from databases or hostname information [8, 14, 19]. Recent work instead evaluates whether a candidate path is physically consistent with observed latencies [12]. Adjacent hops must be reachable within the latency separating them; candidate paths that violate basic physical constraints score poorly.

Physical consistency, however, does not establish geographic correctness. Multiple geographically distinct paths may satisfy the same latency constraints. A complementary metric, $\mathcal{R} \in [0, 1]$, measures path–model alignment by comparing the latency–distance residual profiles of the inferred path and the original geolocation-database path. Higher values indicate stronger agreement between these representations, not necessarily more accurate locations: both may agree and still be wrong.

These signals expose different properties of an inference. Physical consistency tests whether a path satisfies latency constraints, while $\mathcal{R}$ characterizes agreement between two representations of its geography. Their relationship to actual geolocation errors can be evaluated using independently validated paths. Such evidence supports empirical reliability estimates for populations to which the validation observations can reasonably generalize, rather than establishing the correctness of every inferred location.

Related ideas appear in selective prediction and out-of-distribution detection, where a system may withhold or qualify a prediction when evidence does not support its use.

A set of candidates (cable mapping). Mapping traceroute paths to submarine cables often produces not a single cable but a set of candidates [16, 20]. Such a set is useful only if its meaning is clear.

Virtualization can conceal multiple physical hops or submarine cables behind a single observed network link. A candidate set may thus include several cables actually traversed by the path, cables that remain possible because the evidence cannot distinguish them, or both. The same set of candidates can consequently represent different underlying physical relationships and different degrees of knowledge about them.

This distinction matters for validation. Corroborating that a path traverses multiple cables is different from determining which of several candidate cables it uses. For example, Calypso uses independently reported cable failures to corroborate traceroute-to-cable relationships [20]. Such evidence can support particular relationships without necessarily resolving every candidate assignment. Determining whether remaining candidates reflect additional physical relationships or unresolved uncertainty may itself require evidence that is unavailable, a question we return to in §3.

From confidence to validation. The examples share a principle, but not a single guarantee. The ϕ-accrual score can be checked against its assumptions about heartbeat behavior. TrueTime provides a bound conditional on the operation of its clock infrastructure. Geolocation scores can be evaluated against observed errors, provided the validation data support inference to the population of interest. Cable candidate sets require clarity about what membership in the set means and what independent evidence can establish about the underlying relationships.

In each case, making reliability information explicit changes what can be checked. It does not eliminate the need for ground truth or guarantee that an individual answer is correct. Rather, it makes possible more limited, defensible claims about reliability when individual correctness cannot be established at scale. The next section examines how empirical reliability estimation can make use of such information.

3 Beyond Individual Correctness

The examples in §2 make different reliability claims, but all expose information that can be assessed without verifying every individual answer. Here we examine three ways this can reduce dependence on ground truth: empirical reliability estimation, distinguishing sources of uncertainty, and falsification through physical constraints.

Traditionally, validation asks whether an output is correct, which requires ground truth aligned with that output. When ground truth is sparse, this limits how many individual inferences can be verified, regardless of how much data can be collected. Empirical reliability estimation offers a complementary target: how often are inferences correct when they share particular observable properties? For example, rather than validating every inferred router location, we can estimate correctness rates for groups of inferences with similar reliability scores.

Statistical calibration is a special case in which the reported scores represent predicted probabilities. Calibration tests whether predictions assigned a given probability are correct at the corresponding frequency. More generally, a

reliability score need not have a probabilistic interpretation. Its relationship to correctness can be characterized empirically, without requiring the score itself to equal the observed correctness rate.

The contribution is not a new statistical technique. *Our claim is that empirical reliability estimation becomes a useful validation target when individual correctness cannot be established at scale.* Rather than establishing correctness one inference at a time, group-level reliability estimation characterizes correctness rates across larger populations without identifying which individual inferences are correct. Those estimates can then inform the interpretation of unlabeled outputs with similar observable properties, provided the relationship established from the sample applies to them.

We call the resulting gain the *validation dividend*: the ability to characterize reliability for a larger population of inferences than can be individually verified with the same ground-truth budget. The dividend depends on how many groups have enough validated observations to support statistically informative estimates, and how many inferences belong to those groups. Crucially, statistical precision does not establish that an estimate generalizes: if validation disproportionately selects certain kinds of paths, the estimated correctness rates may not transfer to other paths, even within the same score group. Nor does a group-level estimate establish the correctness of any particular inference. §4 examines these limits empirically.

3.1 Why the Source of Uncertainty Matters

Not all uncertainty has the same meaning, and treating different sources of uncertainty alike can lead to misleading validation. In particular, a set of candidates may represent multiple actual relationships, uncertainty about which relationships exist, or both.

Recall the cable-mapping example from §2. Virtualization may conceal multiple physical hops and submarine cables behind a single observed network link. Some candidate cables may therefore correspond to relationships the path actually traverses, while others remain candidates only because the available evidence cannot distinguish them. Validating the set requires understanding what membership means: whether a cable is known to be involved, merely possible, or supported by some degree of evidence. Corroborating multiple actual relationships is different from resolving uncertainty about which relationships exist.

The same distinction arises in geolocation. If IPinfo, MaxMind, and rDNS place a router in three different cities and latency measurements cannot distinguish among them, the disagreement reflects limitations of the available evidence. Additional probing may narrow the possibilities, although it need not resolve them. Anycast presents a different situation: a single address may be served from multiple physical locations, so a set of locations can represent genuine multiplicity rather than uncertainty about one fixed location.

This distinction is related to the familiar separation between aleatoric and epistemic uncertainty, although the correspondence is not exact and real measurement systems may combine both. What matters here is its consequence for validation: a procedure appropriate for identifying one unknown answer may not establish the correctness of multiple actual relationships. Recognizing the difference helps determine what evidence is needed. When the source of uncertainty cannot itself be established, that limitation must also be reflected in the reliability claim.

3.2 Falsification Without Ground Truth

Some measurement claims can be rejected without knowing the correct answer. Physical constraints provide a useful example.

A router placement that would require traffic to propagate faster than the speed of light is impossible, regardless of the router's true location. Establishing this violation requires no ground-truth location for the router. The claim can be rejected because it contradicts a known physical constraint.

This form of validation is asymmetric. A violation establishes that a claim is wrong, but satisfying the constraint does not establish that it is correct. Many incorrect locations may remain physically plausible. Physical consistency checking is already common in Internet measurement; its relevance here is that it provides useful negative evidence without requiring the correct location.

The result is knowledge about which answers can be excluded, even when the correct answer remains unknown. Such constraints complement empirical reliability estimation: they eliminate impossible inferences, while labeled observations help characterize the reliability of those that remain.

Together, these approaches reduce different aspects of the dependence on ground truth. Empirical reliability estimation supports group-level assessments, distinguishing sources of uncertainty clarifies what those assessments mean, and physical constraints can reject some claims without ground truth. The next section examines how far reliability estimation can extend validation under a limited ground-truth budget.

4 How Much Ground Truth Is Enough?

We test the central claim directly: how much ground truth is needed to assess the reliability of a large collection of inferences? Validating an inference individually requires ground truth for that inference. Estimating reliability across groups of inferences, by contrast, may require far fewer labels. If so, the two approaches should behave differently as labels become scarce: the number of individually validated

outputs should shrink with the label budget, while group-level reliability estimates should remain informative until too few labels remain to estimate them precisely.

We measure this difference by progressively withholding ground truth and asking what fraction of the corpus can still be characterized. For individual validation, this is the fraction of outputs checked against ground truth. For empirical reliability estimation, it is the fraction belonging to groups whose observed correctness rates can be estimated with useful statistical precision. These are different guarantees; the comparison measures their reach under the same label budget.

Setup. We instantiate the experiment using traceroute geolocation. Given a traceroute path, the system combines multiple geolocation signals—including hostname-derived hints, database matches, and latency-based constraints—to infer router locations [12]. Each geolocated path carries a reliability signal $\mathcal{R} \in [0, 1]$, measuring path–model alignment: the agreement between the latency–distance residual profiles of the inferred path and the original IPinfo geolocation path. Higher values indicate stronger agreement between these representations, not necessarily more accurate locations. The signal is not itself a probability of correctness and can be computed without ground truth.

An independently validated set of paths, obtained through active probing, provides ground truth for a subset of the corpus. We simulate label scarcity by progressively withholding these labels.

Two ways to use ground truth. Individual validation establishes correctness only for paths with ground truth. With a budget of B labels over a corpus C, its coverage is thus $B/|C|$. Empirical reliability estimation uses those same labels differently: it estimates the fraction of correct paths within each decile of $\mathcal{R}$. Once a decile has enough labels for a sufficiently precise estimate, the estimate can be associated with all corpus paths in that decile without validating them individually. The resulting coverage depends on where labels fall across deciles, rather than on the number of corpus paths alone.

We compare the fraction of paths individually validated with the fraction assigned to groups having sufficiently precise reliability estimates. This measures the reach of the available evidence, not whether the estimated correctness rates necessarily generalize to every path in those groups.

Protocol. The labeled pool L is the actively validated set decoded against IPinfo ($|L| = 4{,}678$ paths), each carrying its reliability signal $\mathcal{R}$ and a correctness label y (1 iff mean path error ≤ 200 km, the threshold of [12]). The corpus C is the full IPinfo decode ($|C| \approx 408$k), each path carrying $\mathcal{R}$ computed identically, without ground truth.

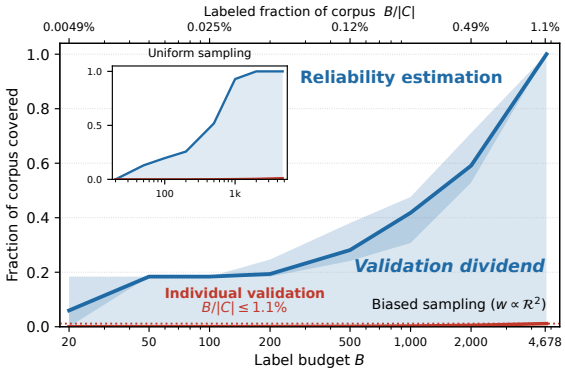


Figure 1: The validation dividend. Corpus coverage as labels are withheld: individual validation covers labeled paths; empirical reliability estimation covers groups with at least ten labels and a Wilson 95% interval half-width ≤ 0.10. Main: biased sampling ($w \propto \mathcal{R}^2$); inset: uniform sampling. Bands: 5–95% range over 200 repetitions. The shaded gap implies neither equivalent guarantees nor generalization to the corpus.

We sweep the budget $B \in \{4{,}678, \ldots, 20\}$ and, at each B, draw B labels under two sampling regimes: uniformly at random from L, and biased toward high-$\mathcal{R}$ paths ($w \propto \mathcal{R}^2$). Individual validation coverage is $B/|C|$. For empirical reliability estimation, we estimate the correctness rate in each $\mathcal{R}$-decile from the sampled labels. We count a decile as covered when it contains at least ten labeled paths and its Wilson 95% confidence interval has a half-width of at most 0.10.[1] Coverage is the fraction of corpus paths belonging to covered deciles. Decile edges are held fixed across budgets; we repeat the sampling 200 times per budget and sampling regime and report the 5–95% bands.

Result. The two coverage curves separate sharply (Fig. 1), illustrating the validation dividend. Under biased sampling, individual validation covers only $B/|C|$ of the corpus: even at the full budget, approximately 1.1%. Empirical reliability estimation covers substantially more—the entire corpus at the full budget, 0.42 at $B = 1{,}000$, and 0.28 at $B = 500$.

Between roughly fifty and a hundred labels, coverage plateaus near 0.18, corresponding to the high-$\mathcal{R}$ corpus mass. Biased sampling leaves too few labels in lower-alignment

[1]The underlying measurement study takes several steps to limit dependence and bias: it retains one traceroute per source–destination anchor pair, separates the data used to construct latency priors from the evaluation period, and obtains validation labels through independent active probing. These safeguards do not guarantee statistical independence: paths may still share routers, links, or geolocation evidence. Wilson intervals therefore provide approximate measures of precision, potentially optimistic when observations are correlated.

deciles to estimate their correctness rates with sufficient precision; at $B = 20$, even the high-alignment decile falls short. Rather than extrapolating beyond the available evidence, the procedure leaves these groups uncharacterized. The advantage is not confined to biased sampling: under uniform sampling (inset), reliability estimates cover the entire corpus down to approximately $B = 2{,}000$ and half of it at $B = 500$. Across much of the range, the same labels support reliability estimates for two to three orders of magnitude more paths than can be individually validated.

Selection bias and transfer. The validation dividend measures how widely group-level estimates can be assigned, but their applicability depends on whether the labeled paths represent the corresponding corpus groups. In our data, the labeled pool contains a substantially larger fraction of high-$\mathcal{R}$ paths than the corpus (approximately 0.37 versus 0.18). This difference arises largely from path-length-dependent validation: obtaining ground truth requires sufficiently close active-probing vantage points along a path, a condition much easier to satisfy for shorter paths. The labeled paths have a median of six hops, compared with nine in the corpus.

Path length also helps explain the difference in $\mathcal{R}$ distributions. Shorter paths have fewer latency increments over which the two representations can disagree. Reweighting the labeled pool to match the corpus path-length distribution reduces its high-$\mathcal{R}$ fraction from approximately 0.37 to 0.19, close to the corpus value of 0.18.

We also examine how estimated correctness rates transfer across path lengths. Estimates derived from paths with at most six hops remain close for paths with at least nine hops in the highest-$\mathcal{R}$ decile (0.990 versus 0.985 correctness), but differences reach 0.14 in some lower-alignment deciles. Thus, similar reliability scores do not always imply interchangeable populations. The experiment demonstrates a substantial potential reduction in the ground-truth budget, while also showing why validation selection and population transfer must be examined separately from statistical precision.

The result is a different form of validation, not a replacement for individual correctness. Limited ground truth can support group-level reliability estimates covering many more inferences than can be individually verified, but neither precise estimates nor broad coverage alone establish their validity for the unlabeled population.

5 Limits of the Pattern

Estimating reliability from limited ground truth can extend the reach of validation, but its usefulness depends on the evidence behind those estimates. Three limitations deserve particular attention.

Correlated evidence. Multiple signals do not necessarily provide independent evidence. Geolocation databases may inherit coordinates from the same upstream source, while multiple hostname hints may reflect a single operator's naming convention. Their agreement can therefore reinforce a shared error rather than corroborate an inference. Similar concerns arise in co-training [3], where the value of agreement depends on assumptions about the relationship between views. The problem is one of dependence: treating correlated observations as independent can overstate the evidence supporting an inference. A practical response is to track provenance and avoid counting multiple observations derived from the same source as independent support. Reliability signals should make these dependencies visible; agreement alone cannot establish independent support.

Selection bias. Ground truth is rarely available uniformly. In traceroute geolocation, active validation favors paths with sufficient measurement coverage, which may differ systematically from paths for which ground truth is unavailable. In §4, we identify path length as a major source of this bias: shorter paths are much more likely to be validated, and they also tend to have higher path–model alignment scores. Accounting for path length largely explains the difference in the fraction of high-alignment paths between the labeled pool and the full corpus.

However, score distributions do not guarantee that correctness rates transfer. Our cross-length comparison finds close agreement for high-alignment paths but larger differences in some lower-alignment groups. Thus, withholding labels demonstrates how widely group-level estimates can be assigned, not that their correctness rates necessarily apply throughout the corpus. Validation effort should prioritize underrepresented path lengths, regions, and conditions, rather than collecting more labels where they are already plentiful.

Systematic errors. Even independent observations can produce systematically misleading results when the measurement process has a shared blind spot. Unlike correlated evidence, the problem here need not arise from multiple signals relying on the same source. Instead, the observations themselves may be insufficient to distinguish correct inferences from incorrect ones. For example, a router placement may satisfy observed latency constraints and agree with neighboring hops while remaining geographically wrong: many different placements may be consistent with the same measurements. Physical consistency can establish plausibility without establishing correctness.

The problem also arises in aggregate measurements. Early traceroute studies suggested that Internet topology followed a power-law degree distribution [6]. Subsequent work showed how traceroute sampling can produce apparent power-law behavior even when the underlying topology does not have

that structure [1, 13]. The problem was not necessarily an incorrect individual traceroute, but a systematic distortion introduced by the measurement process. Even independently collected traceroutes can share this sampling bias. Checking consistency among the observed paths would not, by itself, reveal what the sampling process had missed.

There is a further question of whether the measured quantity is the right one. As Willinger et al. argue, degree distributions alone can be a poor abstraction for understanding network structure and behavior [21]. Reliability estimation can help assess whether an inference is supported by available evidence; it cannot determine whether the chosen abstraction answers the question we care about.

These limits do not erase the benefit of using limited ground truth to estimate reliability. They define what the estimates can support, and what remains uncertain. The goal is not to eliminate the need for external validation, but to use it more effectively while making its limits visible.

6 An Agenda

Partial observability should shape not only how measurement systems produce inferences, but also what they report and what conclusions those reports support. The validation dividend illustrates the potential of this shift: even when individual correctness cannot be established at scale, systems may expose evidence that supports more limited, empirically assessable reliability claims.

Design for auditable reliability. Measurement systems are typically evaluated by the quality of the answers they produce. We argue for an additional design objective: outputs should expose information that helps assess their reliability. Confidence scores, bounds, consistency measures, and conditions under which an inference applies can all contribute, provided their meaning and assumptions are explicit. The challenge is not simply to produce a score, but to establish what that score reveals about correctness, what evidence supports that interpretation, and where it ceases to be informative. A signal that describes a model's internal behavior is not necessarily a reliable indicator of correctness.

Use ground truth where it matters most. Once a system exposes a reliability signal, scarce validation effort can be directed toward groups whose correctness rates remain poorly understood. Even high-scoring outputs deserve scrutiny when their scores lack ground-truth validation. Our geolocation experiment illustrates another challenge: validation can disproportionately select shorter paths, leaving longer paths underrepresented even when they have similar reliability scores. Validation should therefore target not only poorly characterized score groups, but also regions, networks, and conditions for which estimates may not transfer. The objective is to collect the ground truth that most improves our understanding of where reliability claims hold.

Extend the approach to learned systems. Network foundation models and other learned representations increasingly derive inferences from large unlabeled corpora [10]. Synthetic traces and embedding-based similarity can extend their reach, but do not remove the validation problem. They introduce new signals whose relationship to correctness must be established. An important direction is to determine which properties of learned representations can support empirical reliability estimates, and under what conditions those estimates transfer to previously unseen networks or measurement settings.

Change what we reward. Measurement studies are often judged by the number of inferences they produce and the ground truth they collect. Under severe ground-truth scarcity, these criteria can favor broad coverage without corresponding evidence of reliability. We should also evaluate how well a system characterizes its limitations: whether its reliability estimates are informative, how much ground truth they require, whether their assumptions are supported, and where they cease to hold. Concretely, measurement studies should report not only inference coverage and accuracy on available ground truth, but also *reliability coverage*: the fraction of inferences for which an empirically supported reliability characterization is available, together with the assumptions and limitations of that characterization. Reporting uncertainty should be evidence of rigor, not a weakness.

Several questions remain open. How broadly does the validation dividend apply beyond traceroute geolocation, particularly to inferences that combine multiple uncertain inputs, such as submarine-cable mapping? How can we estimate reliability when ground truth is unevenly distributed or the evidence itself is correlated? And how can we detect systematic errors that remain consistent with every signal a system observes? These questions call for experiments that measure not only inference accuracy, but also the cost and limits of assessing that accuracy.

Ground truth remains essential, but its role expands: beyond checking individual outputs, it can characterize reliability across groups, test transfer to new settings, and guide what to measure next.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-2107392 and CNS-2246475.

References

[1] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. 2009. On the Bias of Traceroute Sampling: or, Power-Law

Degree Distributions in Regular Graphs. *J. ACM* 56, 4, Article 21 (2009), 28 pages. doi:10.1145/1538902.1538905

[2] Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* 16, 4 (March 2023), 494–591. doi:10.1561/2200000101

[3] Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proc. of COLT*. doi:10.1145/279943.279962

[4] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. 2012. Spanner: Google's Globally-Distributed Database. In *Proc. of USENIX OSDI*. https://www.usenix.org/conference/osdi12/technical-sessions/presentation/corbett

[5] Omar Darwich, Hugo Rimlinger, Milo Dreyfus, Matthieu Gouel, and Kevin Vermeulen. 2023. Replication: Towards a Publicly Available Internet Scale IP Geolocation Dataset. In *Proc. of IMC*. doi:10.1145/3618257.3624801

[6] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On Power-Law Relationships of the Internet Topology. In *Proc. of ACM SIGCOMM*. doi:10.1145/316194.316229

[7] Lixin Gao. 2001. On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on Networking* 9, 6 (December 2001). doi:10.1109/90.974527

[8] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. 2017. A Look at Router Geolocation in Public and Commercial Databases. In *Proc. of IMC*. 463–469. doi:10.1145/3131365.3131380

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proc. of ICML*. https://proceedings.mlr.press/v70/guo17a.html

[10] Satyandra Guthula, Navya Battula, Roman Beltiukov, Wenbo Guo, and Arpit Gupta. 2023. netFound: Foundation Model for Network Security. *arXiv preprint arXiv:2310.17025* (2023).

[11] Naohiro Hayashibara, Xavier Défago, Rami Yared, and Takuya Katayama. 2004. The φ Accrual Failure Detector. In *Proc. of IEEE SRDS*. 66–78. doi:10.1109/RELDIS.2004.1353004

[12] Santiago Klein, Caleb J. Wang, and Fabián E. Bustamante. 2026. Overconfident Coordinates: Quantifying Confidence in Traceroute Geolocation. arXiv:2606.24027 [cs.NI] doi:10.48550/arXiv.2606.24027

[13] Anukool Lakhina, John W. Byers, Mark Crovella, and Peng Xie. 2003. Sampling Biases in IP Topology Measurements. In *Proc. of IEEE INFOCOM*. doi:10.1109/INFCOM.2003.1208685

[14] Matthew Luckie, Bradley Huffaker, Alexander Marder, Zachary Bischof, Marianne Fletcher, and year = 2021 booktitle = conext doi = 10.1145/3485983.3494869 kc Claffy, title = Learning to extract geographic information from internet router hostnames. [n. d.].

[15] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review* 41, 2 (2011), 53–56. doi:10.1145/1971162.1971171

[16] Alagappan Ramanathan and Sangeetha Abdu Jyothi. 2024. Nautilus: A Framework for Cross-Layer Cartography of Submarine Cables and IP Links. In *Proc. of ACM SIGMETRICS*.

[17] Mario A. Sanchez, Fabián E. Bustamante, Balachander Krishnamurthy, Walter Willinger, Georgios Smaragdakis, and Jeffrey Erman. 2014. Inter-Domain Traffic Estimation for the Outsider. In *Proc. of IMC*. doi:10.1145/2663716.2663740

[18] Neil Spring, Ratul Mahajan, and David Wetherall. 2002. Measuring ISP Topologies with Rocketfuel. In *Proc. of ACM SIGCOMM*. doi:10.1145/633025.633039

[19] Kedar Thiagarajan, Esteban Carisimo, and Fabiń E. Bustamante. 2025. The Aleph: Decoding DNS PTR Records With Large Language Models. In *ACM CoNEXT*. doi:10.1145/3709374

[20] Caleb J. Wang, Ying Zhang, Qianli Dong, Esteban Carisimo, Ramakrishnan Durairajan, and Fabián E. Bustamante. 2026. Calypso's Voyage: Charting Traceroute Paths Through Submarine Cables. In *Proc. of IMC*.

[21] Walter Willinger, David Alderson, and John C. Doyle. 2009. Mathematics and the Internet: A Source of Enormous Confusion and Great Potential. *Notices of the American Mathematical Society* 56, 5 (2009), 586–599.