Annals of Science and Inquiry
Vol. 2 No. 6 (2026)

An Incomplete Horizon Can Still Settle a Forecast Ranking: Exact Label-Completion Bounds in Streaming Telemetry

William Murphy[1] · Lucas Walker[2] · Evan Anderson[3]

[1] Austria · [2] Belgium · [3] Italy

Published: June 17, 2026 | Article 856

Abstract

Missing follow-up need not leave every forecast comparison undecidable. For 131,799 later streaming windows, we compare a previously fixed buffer-level probability score with its fixed prevalence benchmark. Complete records supply 128,224 labels; incomplete records contribute 70 observed positives and 3,505 unresolved cells. Allowing every binary assignment to those unresolved cells gives an exact paired Brier difference between −0.000497624 and −0.000397597, favoring the buffer score throughout. Keeping the partial-positive evidence reduces interval width by 60.28%, although the comparison remains favorable when that evidence is withheld. Every single-session and single-hour deletion preserves the direction. These are finite monitoring-label completion results for one archive, without identifying unrecorded physical interruptions or future-day performance.

An Incomplete Horizon Can Still Settle a Forecast Ranking: Exact Label-Completion Bounds in Streaming Telemetry

Abstract

Missing follow-up need not leave every forecast comparison undecidable. For 131,799 later streaming windows, we compare a previously fixed buffer-level probability score with its fixed prevalence benchmark. Complete records supply 128,224 labels; incomplete records contribute 70 observed positives and 3,505 unresolved cells. Allowing every binary assignment to those unresolved cells gives an exact paired Brier difference between -0.000497624 and -0.000397597, favoring the buffer score throughout. Keeping the partial-positive evidence reduces interval width by 60.28%, although the comparison remains favorable when that evidence is withheld. Every single-session and single-hour deletion preserves the direction. These are finite monitoring-label completion results for one archive, without identifying unrecorded physical interruptions or future-day performance.

1. Which comparison survives an unfinished record?

A monitoring record can stop before an evaluation horizon ends. Deleting that window produces a score for the completely observed subset. Calling its outcome negative produces a different score by adding an unverified label. Neither operation answers whether a forecast ranking is robust to the unresolved information. A third possibility is to retain the window and ask whether every declared completion of its label yields the same ordering. That question can sometimes be settled without choosing a probability model for missingness.

We examine this possibility in public streaming telemetry. The Puffer platform provides measurements from real streaming operation and a research setting for examining video delivery (Yan et al., 2020). Our empirical question is: **does a previously fixed buffer-level probability score have lower mean Brier loss than a fixed prevalence benchmark under every binary completion of unresolved later monitoring labels?** A related question asks how much interval width is removed by preserving positive reports found inside incomplete windows. The response remains a recorded technical event, rather than satisfaction or a verified service fault.

Yuan et al. (2026) describe dynamic service-dependency graphs for delay prediction and cause localization in media-consumption workflows. Their registered abstract motivates a richer operational objective than the one supported by endpoint reports alone. Before a client warning contributes to such diagnosis, its evaluation should distinguish available evidence from missing observation. We use that distinction to frame our question; the public telemetry contains neither their service graph nor adjudicated causes. Their private accuracy or localization results are not reproduced, and no cache, permission or comment-service failure is inferred here.

The study also separates forecast evaluation from explanation. Shmueli (2010) discusses the different aims of explaining and predicting, while measurement analysis asks what construct an observable can support (Jacobs and Wallach, 2021). In this archive, a received rebuffer report, a fully followed report interval and a physical playback interruption are related concepts with different evidential requirements. Our contribution concerns a ranking over explicitly completed monitoring labels. It supplies no causal effect of buffer occupancy and no recovery of a hidden physical event sequence.

2. A window has an observation status before it has a usable label

We reuse the fixed official client_buffer release spanning 12 November 2026 at 11:00 UTC to 13 November at 11:00 UTC. The official data description defines its timestamp as the server's receipt time for a client message and documents init, startup, timer, rebuffer and play events (Stanford Puffer project, n.d.-a). Timers are intended to occur periodically, but their spacing can lengthen, including when the browser tab becomes inactive. The reported buffer is measured in seconds. These definitions do not establish an independently synchronized client clock or reliable continuous reporting merely because timestamps are present.

The preserved acquisition manifest describes 40,456,730 source rows from 3,267 local session keys and 22,888 stream keys, including 2,062 raw rebuffer-report rows. Original compressed length, provider checksum and gzip integrity were verified during that acquisition. Disposable original-byte files were subsequently removed. The present study freshly checks the retained technical projection and derived-window hashes; it does not claim to have downloaded or reverified those removed original bytes. Content channels have already been discarded, and no video, profile or network address enters the analysis.

The release exporter and its pinned source revision help explain why file position cannot substitute for chronology (Stanford Puffer project, n.d.-b). The retained audit records 342 global timestamp reversals, while individual stream histories are strictly ordered. Stream-specific construction therefore precedes evaluation. Public session equality keys define grouping within this source, without identifying people or guaranteeing anonymity. Documenting those meanings and processing boundaries follows the dataset-description concerns emphasized by Gebru et al. (2021), rather than treating a publicly accessible table as context-free evidence.

Candidate windows inherit an earlier archived-data construction. An observed init must precede startup, and the latest named playback state must be startup or resumed play rather than rebuffering. Anchors begin after sixty seconds of startup history and occur at least thirty seconds apart. An anchor timer may be at most two seconds later than scheduled. Five preceding seconds require a report sequence without any adjacent gap above two seconds. Current reported buffer is available at the anchor; later coverage never enters the score.

Complete future support requires reports spanning the thirty-second horizon, no intervening report gap above two seconds, and an endpoint no more than two seconds late. A positive label means at least one explicit rebuffer report received in the open-closed interval from the anchor through thirty seconds later. A report beyond that horizon can establish coverage but cannot create a positive inside it. Complete windows without an in-horizon report receive zero. This is an operational monitoring definition; it is not an assurance that every physical playback event was detected.

Incomplete windows are divided into two statuses. A partial positive contains an in-horizon positive report despite failing complete support, and its binary monitoring label is fixed at one. An unknown window lacks both complete support and a positive report in the available horizon. Its label remains blank in the analytical data. Source endings, channel switches and report gaps are not converted into observed negative responses or voluntary exits. Little and Rubin (2002) explain why incomplete-data analysis requires attention to the observation process. A small unresolved fraction alone does not establish the ignorability conditions considered by Rubin (1976) or clarify the distinct meanings of missing at random discussed by Seaman et al. (2013).

3. Frozen scores and a finite completion class

The principal cohort includes all eligible anchors at or after the fixed 23:00 UTC boundary whose session keys were first observed at or after that boundary. It contains 131,799 windows from 1,150 session keys and 2,660 streams. There are 128,224 complete windows, including 189 positives; another seventy are partial positives, and 3,505 are unknown. A separate diagnostic retains 54,958 later windows from 140 continuing session keys. Its corresponding counts are 53,843 complete windows, 31 complete positives, twelve partial positives and 1,103 unknowns. The principal keys are absent from earlier fitting records, which does not establish that the people or real-world visits are new.

Both probabilities were fixed in the earlier local analysis, before this completion question was evaluated. The buffer score uses the logistic transformation of -7.6990590103 minus 0.4827650569 times the standardized value of log(1 + buffer seconds). Its training mean and scale are 2.7933861035 and 0.1073213332. The fixed comparator is 116/109256, or approximately 0.001061727, obtained from the earlier training prevalence with one positive and one negative pseudocount. The earlier fitting cohort comprised 109,254 complete windows with 115 positives. We change neither coefficients nor training data and make no additional fit.

Freezing the prediction object keeps this experiment focused on evaluation information. Dawid's prequential perspective places future observable forecasts at the center of statistical assessment (Dawid, 1984). Here the scoring object already exists; additional reporting knowledge changes the admissible evaluation labels. It does not become a new predictor. The protocol was frozen after source support and prior results were available, but before the present loss bounds were computed. This is a secondary follow-up using shared source data and a shared frozen score, without a claim of preregistration, outcome blindness or independent replication.

Let p denote the buffer probability, q the constant probability and y a completed binary monitoring label. Brier loss is squared probability error (Brier, 1950). Its paired difference is $D(y) = (p-y)^2-(q-y)^2 = p^2-q^2-2y(p-q)$, with negative values favoring p. Proper-scoring-rule theory supports evaluating the probabilities themselves rather than selecting a convenient classification threshold (Gneiting and Raftery, 2007). For a known label we calculate its observed D. For an unknown cell, its lower contribution is the smaller of D(0) and D(1), and its upper contribution is the larger. Summed contributions divided by the same total window count give the exact paired interval.

Every complete label and partial positive stays fixed, while each unknown cell independently takes zero or one. The resulting mathematical class is deliberately explicit. Its endpoint assignments are attainable because the loss sum is separable: choose the minimizing label in each unresolved cell for one endpoint and the maximizing label for the other. Bound analysis makes conclusions conditional on stated information and assumptions (Manski, 2003; Tamer, 2010). This application, however, defines a finite class of monitoring-label completions; it does not assert that all those assignments are physically feasible trajectories or that they identify a population parameter.

P1 asks whether the primary interval remains on one side of zero. P2 compares its width with an information ablation that also treats partial positives as unknown, using the identical windows. P3 deletes each session key and each UTC anchor hour while keeping probabilities fixed. The figure connects these calculations. None of the intervals is a sampling confidence interval. Confidence procedures for partially identified population parameters address an additional inferential task (Imbens and Manski, 2004), which is not carried out here.

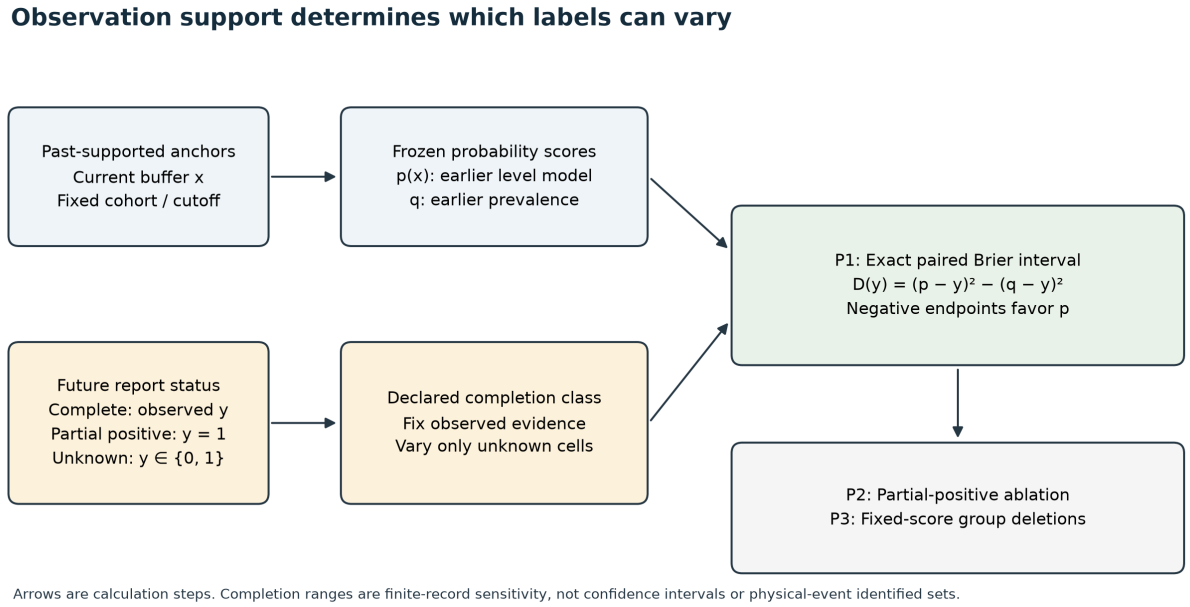

Figure 1. Research model. *P1 evaluates the paired Brier interval over a stated completion class. P2 withholds partial-positive evidence to assess interval narrowing. P3 checks finite group-deletion sensitivity. Arrows denote calculation steps; they do not denote causal effects or reconstructed physical events.*

4. The unresolved cells do not reverse this ranking

Table 1 reports the primary estimate as an interval rather than a selected imputation. In the principal cohort, the paired Brier difference ranges from -0.000497624 to -0.000397597. Its upper endpoint is already below zero, so the frozen buffer score has lower loss than the constant for every completion in the declared class. This answers P1 for this archived comparison. It does not imply that either probability is well calibrated, that all viewers benefit, or that a competing model would receive the same ordering.

Table 1. Exact paired Brier differences under the frozen label-completion design. *The fixed earlier ridge-logistic buffer score is compared with fixed earlier prevalence; no model is refitted. All differences are multiplied by 10^6; negative values favor the buffer score. Completion endpoints describe deterministic uncertainty over binary unknown cells, not confidence intervals. The complete-only reference uses the separate complete-window denominator.*

Cohort	All windows / sessions	Complete / partial positive / unknown	Complete-only ΔBrier × 10^6	Completion lower × 10^6	Completion upper × 10^6
Later first-observed keys	131799 / 1150	128224 / 70 / 3505	-350.0038	-497.6236	-397.5967
Later continuing keys	54958 / 140	53843 / 12 / 1103	-101.2523	-192.4418	-114.5572

The complete-window reference difference is -0.000350004, obtained from buffer and prevalence Brier losses of 0.001121977 and 0.001471980. That reference concerns 128,224 windows. The completion interval concerns all 131,799 eligible principal anchors. Comparing their numerical values as though one were a corrected version of an identical estimand would obscure both the changed denominator and the retained partial-positive evidence. The relevant finding is the sign throughout the declared completion class, rather than a claim that missingness necessarily strengthens physical predictive performance.

The continuing-session diagnostic also remains favorable throughout, from -0.000192442 to -0.000114557. It contains fewer known positives and is not a second day or independent platform replication. Its role is to expose a separately defined history condition, not to estimate a causal effect of session age. In particular, its apparently narrower interval does not establish that continuing sessions are

easier to monitor. The number and score distribution of their unresolved cells differ from those of the principal cohort.

Forecast comparison has several dimensions. DeGroot and Fienberg (1983) distinguish calibration and refinement, and Murphy (1973) analyzes components of the probability score. We estimate neither component separately. Murphy (1993) also distinguishes aspects of forecast goodness that cannot be reduced to one universal number. The present result is a paired mean-loss ordering against a fixed, simple benchmark. It does not establish alarm usefulness at a chosen budget, the cost of intervention or satisfactory calibration across operating conditions.

5. Seventy incomplete windows carry substantial usable information

An incomplete horizon need not be wholly uninformative. In the principal cohort, retaining the seventy partial positives reduces interval width from 0.000251858 to 0.000100027, a reduction of 0.000151831, or 60.28%. Table 2 shows the endpoints before and after that evidence is used. Both ablation endpoints remain negative. Thus the positive reports sharpen this comparison substantially, but they are not necessary to certify its direction. Describing them as the reason the score wins would overstate the result.

Table 2. Retaining positive evidence within incomplete horizons. *Endpoints and widths are paired Brier differences multiplied by 10^6. The ablation frees partial-positive labels as well as unknown labels; it changes the permitted information and preserves the same denominator and scores.*

Cohort	All windows	Partial-positive labels retained	Ablation interval × 10^6	Evidence-retaining interval × 10^6	Width reduction × 10^6
Later first-observed keys	131799	70	[-497.7670, -245.9091]	[-497.6236, -397.5967]	151.8310
Later continuing keys	54958	12	[-192.5857, -36.4908]	[-192.4418, -114.5572]	78.2103

The contribution of a newly known cell depends on its two forecast probabilities, not simply its presence. For an unresolved cell, the paired-loss width is 2|p-q|. Consequently, a small number of cells with large score disagreement can account for more uncertainty than many cells where the scores nearly agree. The seventy principal partial positives have mean buffer probability 0.143729, compared with 0.001753 among unknown windows. Those finite descriptive values help explain the observed narrowing without asserting that reporting status caused a score or that partial positives represent every unrecorded interruption.

Positive-and-unlabeled learning provides a useful comparison because an unlabeled example need not be a negative one (Bekker and Davis, 2020). Our task additionally includes complete negative windows and does not train a positive-and-unlabeled classifier. Elkan and Noto (2008) derive a correction under a random selection assumption for labeled positives. No such assumption is established for these reporting gaps, so their correction is not applied. The completion analysis retains uncertainty instead of estimating a label-selection probability from unavailable evidence.

A corresponding continuing-session ablation reduces width from 0.000156095 to 0.000077885 when twelve partial positives are fixed. Its favorable sign also predates that narrowing. In both cohorts, the information comparison preserves exactly the same anchors and probabilities. Only which labels are permitted to vary changes. This isolates the mathematical value of already observed positive reports from changes in the sample or predictive specification; it does not identify why particular horizons became incomplete.

6. Finite deletions and the reach of the result

All 1,150 single-session deletions and twelve single-hour deletions in the main cohort retain a negative upper endpoint. Table 3 gives extrema across these deletions. The continuing diagnostic also retains the direction for its 140 session deletions and twelve hour deletions, although its least favorable hourly upper endpoint approaches zero more closely. No group is removed from the primary result because its contribution is large. Scores remain fixed throughout, and there is no refitting, bootstrap or population standard error hidden in these ranges.

Table 3. Fixed-score group-deletion sensitivity. *Entries are extrema of completion endpoints across deletions, multiplied by 10^6. They are finite deletion ranges, without a sampling-confidence interpretation. Every retained interval remains below zero.*

Cohort	Omitted group	Deletions	Lower-endpoint range × 10^6	Upper-endpoint range × 10^6
Later first-observed keys	Session key	1150	[-503.8962, -346.8031]	[-405.9933, -278.8964]
Later first-observed keys	UTC anchor hour	12	[-545.7461, -430.4518]	[-444.9845, -328.6787]
Later continuing keys	Session key	140	[-200.5916, -94.1423]	[-122.1957, -16.2435]
Later continuing keys	UTC anchor hour	12	[-218.3470, -103.4655]	[-137.4018, -7.9773]

Structured validation depends on what is withheld, particularly with temporal and hierarchical records (Roberts et al., 2017). A public session key is one available grouping unit; it is not a person-level sampling unit or evidence of independence across the day. The deletion calculation answers whether the finite sign conclusion relies on one recorded session or hour. It leaves common operating conditions, reporting delays and cross-stream dependence outside a population model. The result therefore supports a finite-source robustness statement with narrower reach than a deployment guarantee.

Information boundaries also constrain the study's adaptation. Leakage and reproducibility concerns require stating what was available before performance assessment (Kapoor and Narayanan, 2023). Prior complete-window results and source support were known when this new question was designed. We froze the comparator, cohort, completion class and diagnostic family before calculating these bounds. We searched no alternative day, horizon, penalty or model after obtaining their signs. The experiment investigates a particular unresolved-label problem, rather than discovering the best streaming predictor through repeated evaluation.

7. Reusable evidence and limits

The analytical table keeps one row per window and separates status, known label, frozen probabilities and endpoint contributions. This organization makes the distinction between a missing label and a numeric zero inspectable, consistent with the data-organization principles in Wickham (2014). NumPy performs vectorized probability and endpoint calculations (Harris et al., 2020); SciPy supplies the logistic transformation (Virtanen et al., 2020); pandas implements filtering, grouping and exports (McKinney, 2010). Matplotlib renders the calculation diagram (Hunter, 2007). These software references describe tools actually used, rather than methods claimed to have been run.

An independent scalar implementation reconstructs all 186,757 later probabilities and endpoint contributions. It compares 182,067 complete-window probabilities with the prior saved object, directly evaluates squared losses under endpoint-attaining assignments, and recomputes all 1,314 group deletions. A separate scan of the retained 40,456,730-row projection independently checks thirty-two actual windows, including partial positives, unknowns and complete labels. Those checks cover anchor location,

received report timing, support and source-local first observation. Inspectable numerical objects and transformations serve computational reproducibility (Peng, 2011; Sandve et al., 2013); they do not substitute for accountable scientific review.

The most consequential boundary is the completion class itself. An unknown cell is varied as a monitoring label, without an assertion about events after viewing stopped, reports that were never sent, or the response under continued observation. The calculation does not recover those counterfactual processes. Even complete support is a stated reporting rule rather than proof of perfect physical ascertainment. Unknown transport delays can separate server receipt from client occurrence. A new service architecture or reporting cadence could change both the meaning of the label and the amount of unresolved information.

The public research-release context and participation terms are retained locally with provenance. No unrestricted redistribution license is invented, no identity linkage is attempted, and no restricted manuscript or raw source document is transferred to an external analysis service. Institutional ethics determination, accountable source-claim verification, author attribution and final scientific approval remain pending. This artifact is an empirical draft prepared with AI assistance and local numerical checks, without a claim of human approval or submission readiness.

Conclusion

For these fixed streaming forecasts, unresolved monitoring labels do not prevent a definite paired Brier ordering: the buffer-level score outperforms its prevalence comparator throughout the declared finite completion class. Partial-positive reports cut the interval width substantially while leaving an already favorable sign unchanged. The result shows how an incomplete record can still answer a narrow evaluation question when known evidence and unresolved cells are kept separate. Its scope remains one source, two frozen scores and explicitly completed monitoring labels; stronger claims about physical interruption, service causes or later deployment require additional evidence.

References

1. Imbens, Guido W.; Manski, Charles F.. (2004). Confidence Intervals for Partially Identified Parameters. Econometrica. https://doi.org/10.1111/j.1468-0262.2004.00555.x
2. DeGroot, Morris H.; Fienberg, Stephen E.. (1983). The Comparison and Evaluation of Forecasters. The Statistician. https://doi.org/10.2307/2987588.
3. Stanford Puffer project. (n.d.-b). Puffer-statistics: public release pipeline, pinned source revision 59afb3149b6249dcbedca4cfd9dcd7025ed8fd00. Accessed 17 November 2026. https://github.com/StanfordSNR/puffer-statistics/tree/59afb3149b6249dcbedca4cfd9dcd7025ed8fd00
4. RUBIN, DONALD B.. (1976). Inference and missing data. Biometrika. https://doi.org/10.1093/biomet/63.3.581
5. Dawid, A. P.. (1984). Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. Journal of the Royal Statistical Society. Series A (General). https://doi.org/10.2307/2981683.
6. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723.
7. Wickham, Hadley. (2014). Tidy Data. Journal of Statistical Software. https://doi.org/10.18637/jss.v059.i10
8. Murphy, Allan H.. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting. https://doi.org/10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2
9. Little, Roderick J. A.; Rubin, Donald B.. (2002). Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9781119013563

10. Virtanen, Pauli; Gommers, Ralf; Oliphant, Travis E.; Haberland, Matt; Reddy, Tyler; Cournapeau, David; Burovski, Evgeni; Peterson, Pearu; Weckesser, Warren; Bright, Jonathan; van der Walt, Stéfan J.; Brett, Matthew; Wilson, Joshua; Millman, K. Jarrod; Mayorov, Nikolay; Nelson, Andrew R. J.; Jones, Eric; Kern, Robert; Larson, Eric; Carey, C J; Polat, İlhan; Feng, Yu; Moore, Eric W.; VanderPlas, Jake; Laxalde, Denis; Perktold, Josef; Cimrman, Robert; Henriksen, Ian; Quintero, E. A.; Harris, Charles R.; Archibald, Anne M.; Ribeiro, Antônio H.; Pedregosa, Fabian; van Mulbregt, Paul; SciPy 1.0 Contributors; Vijaykumar, Aditya; Bardelli, Alessandro Pietro; Rothberg, Alex; Hilboll, Andreas; Kloeckner, Andreas; Scopatz, Anthony; Lee, Antony; Rokem, Ariel; Woods, C. Nathan; Fulton, Chad; Masson, Charles; Häggström, Christian; Fitzgerald, Clark; Nicholson, David A.; Hagen, David R.; Pasechnik, Dmitrii V.; Olivetti, Emanuele; Martin, Eric; Wieser, Eric; Silva, Fabrice; Lenders, Felix; Wilhelm, Florian; Young, G.; Price, Gavin A.; Ingold, Gert-Ludwig; Allen, Gregory E.; Lee, Gregory R.; Audren, Hervé; Probst, Irvin; Dietrich, Jörg P.; Silterra, Jacob; Webber, James T; Slavič, Janko; Nothman, Joel; Buchner, Johannes; Kulick, Johannes; Schönberger, Johannes L.; de Miranda Cardoso, José Vinícius; Reimer, Joscha; Harrington, Joseph; Rodríguez, Juan Luis Cano; Nunez-Iglesias, Juan; Kuczynski, Justin; Tritz, Kevin; Thoma, Martin; Newville, Matthew; Kümmerer, Matthias; Bolingbroke, Maximilian; Tartre, Michael; Pak, Mikhail; Smith, Nathaniel J.; Nowaczyk, Nikolai; Shebanov, Nikolay; Pavlyk, Oleksandr; Brodtkorb, Per A.; Lee, Perry; McGibbon, Robert T.; Feldbauer, Roman; Lewis, Sam; Tygier, Sam; Sievert, Scott; Vigna, Sebastiano; Peterson, Stefan; More, Surhud; Pudlik, Tadeusz; Oshima, Takuya; Pingel, Thomas J.; Robitaille, Thomas P.; Spura, Thomas; Jones, Thouis R.; Cera, Tim; Leslie, Tim; Zito, Tiziano; Krauss, Tom; Upadhyay, Utkarsh; Halchenko, Yaroslav O.; Vázquez-Baeza, Yoshiki. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. https://doi.org/10.1038/s41592-019-0686-2.

11. Yan, Francis Y.; Ayers, Hudson; Zhu, Chenzhi; Fouladi, Sadjad; Hong, James; Zhang, Keyi; Levis, Philip; Winstein, Keith. (2020). Learning in situ: a randomized experiment in video streaming. 17th USENIX Symposium on Networked Systems Design and Implementation. https://www.usenix.org/conference/nsdi20/presentation/yan

12. Stanford Puffer project. (n.d.-a). Puffer: public streaming measurements and data description. Accessed 17 November 2026. https://puffer.stanford.edu/data-description/

13. Elkan, Charles; Noto, Keith. (2008). Learning classifiers from only positive and unlabeled data. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. https://doi.org/10.1145/1401890.1401920.

14. Murphy, Allan H.. (1973). A New Vector Partition of the Probability Score. Journal of Applied Meteorology. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

15. Hunter, John D.. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55

16. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847

17. McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-00a

18. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804

19. Roberts, David R.; Bahn, Volker; Ciuti, Simone; Boyce, Mark S.; Elith, Jane; Guillera-Arroita, Gurutzeta; Hauenstein, Severin; Lahoz-Monfort, José J.; Schröder, Boris; Thuiller, Wilfried; Warton, David I.; Wintle, Brendan A.; Hartig, Florian; Dormann, Carsten F.. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. https://doi.org/10.1111/ecog.02881.

20. Sandve, Geir Kjetil; Nekrutenko, Anton; Taylor, James; Hovig, Eivind. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1003285

21. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.

22. Yuan, Yuewei; Xu, Tianyi; Yin, Jiayang; Huang, Jialei. (2026). Modeling Loading Anomalies and Identifying Root Causes in Media Consumption Workflows on Large Social Media Platforms. SSRN working paper 7393318. https://doi.org/10.2139/ssrn.7393318.

23. Manski, Charles F.. (2003). Partial Identification of Probability Distributions. Springer Series in Statistics. https://doi.org/10.1007/b97478

24. Tamer, Elie. (2010). Partial Identification in Econometrics. Annual Review of Economics. https://doi.org/10.1146/annurev.economics.050708.143401

25. Bekker, Jessa; Davis, Jesse. (2020). Learning from positive and unlabeled data: a survey. Machine Learning. https://doi.org/10.1007/s10994-020-05877-5.

26. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901

27. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

28. Shmueli, Galit. (2010). To Explain or to Predict?. Statistical Science. https://doi.org/10.1214/10-sts330

29. Brier, Glenn W.. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

30. Seaman, Shaun; Galati, John; Jackson, Dan; Carlin, John. (2013). What Is Meant by “Missing at Random”?. Statistical Science. https://doi.org/10.1214/13-STS415.