Temporal Systems and Forecasting
Volume 1 · Number 1 · 2026

A Reference Count Can Stay Correct While Its Identity Labels Go Wrong

William Hall[1] · Tyler Wells[2] · Ryan King[3]

[1] Switzerland · [2] Finland · [3] Sweden

Original publication: November 4, 2026 | Article 597

Abstract

Five reference photographs per person need not provide five reliable identity labels. We reassigned one of each person's five gallery photographs in a count-preserving cycle and tested two elementary matchers on 200 untouched queries. Nearest-gallery accuracy fell from 87% to 71%; accuracy using class centroids fell from 82.5% to 76%. The centroid rule therefore lost less accuracy and became the better rule under this particular corruption. A fixed reference budget concealed a consequential change in reference integrity. The experiment establishes a controlled sensitivity result, conditional on one historical collection and one reassignment pattern.

Controlled gallery-label experiment on a historical face collection. Photograph labels are changed deliberately; no naturally occurring annotation-error rate or generated-video performance is estimated.

A Reference Count Can Stay Correct While Its Identity Labels Go Wrong

> Controlled gallery-label experiment on a historical face collection. Photograph labels are changed deliberately; no naturally occurring annotation-error rate or generated-video performance is estimated.

Abstract

Five reference photographs per person need not provide five reliable identity labels. We reassigned one of each person's five gallery photographs in a count-preserving cycle and tested two elementary matchers on 200 untouched queries. Nearest-gallery accuracy fell from 87% to 71%; accuracy using class centroids fell from 82.5% to 76%. The centroid rule therefore lost less accuracy and became the better rule under this particular corruption. A fixed reference budget concealed a consequential change in reference integrity. The experiment establishes a controlled sensitivity result, conditional on one historical collection and one reassignment pattern.

1. When an apparently balanced reference set is unreliable

A reference collection may contain the promised number of photographs for every identity while assigning some photographs to the wrong person. Counting files would not reveal that problem. It matters whenever a later system treats a reference label as a trustworthy instruction about whose appearance to preserve. The consequence can depend on how the system combines the reference images, even when the underlying image features stay fixed.

Our research question concerns that dependence: does a small, balanced reference-label intervention change nearest-gallery and class-centroid identity matching differently? The comparison holds all gallery pixels, query photographs, preprocessing, and class counts constant. Only the identity labels attached to selected gallery images change. It therefore isolates a declared reference-integrity stress from adding photographs, changing the representation, or degrading the query images.

Meng et al., 2026 study reference-conditioned subject preservation through stacked multi-view identity mosaic injection. We do not implement that mechanism or infer its robustness from a pixel matcher. The connection is a more elementary requirement for interpreting reference-conditioned results: the identity content and the identity label of a supplied reference should not be conflated. The present experiment measures the consequences of violating that requirement in an inspectable baseline.

The anticipated tradeoff is not that one aggregation rule must dominate. A nearest-reference rule can exploit a particularly suitable example, but a single wrong label can directly redirect a matching query. A centroid spreads each supplied example's influence across a class representation. Averaging may reduce this specific sensitivity while also discarding variation that helps the clean nearest-reference rule. Both parts of that tradeoff are measured on the same queries.

2. Recognition context and the scope of the baseline

The historical stochastic identification study associated with this collection supplies its research context (Samaria and Harter, 1994). Eigenfaces and Fisherfaces illustrate different ways of constructing identity representations (Turk and Pentland, 1991; Belhumeur et al., 1997). Our calculation fits neither projection. It retains the full centered grayscale vector so that the label intervention is not mixed with a new representation-fitting stage or with the transfer of a fitted basis across people.

Illumination preprocessing also changes the matching problem. Work on robust face normalization motivates stating the transformation applied to every photograph (Tan and Triggs, 2010). Here each vector has its own mean removed and is divided by its own Euclidean norm. We apply that operation once before either label condition. There is no query-specific choice of normalization, learned illumination correction, or threshold selected from query labels.

Modern embedding methods pursue a substantially different representation problem. FaceNet, ArcFace, and CosFace formulate learned identity objectives (Schroff et al., 2015; Deng et al., 2018; Wang et al., 2018). AdaFace and MagFace further connect recognition with qualities of the learned representation (Kim et al., 2022; Meng et al., 2021). Those methods contextualize why raw-pixel matching is a limited baseline; their published performance is not entered into our result table or treated as another observation.

Collections such as VGGFace2, LFW, and MegaFace support broader questions about variation and scale (Cao et al., 2017; Huang et al., 2008; Kemelmacher-Shlizerman et al., 2015). They are not merged into the current resource. FairFace additionally emphasizes the importance of the evaluated population (Kärkkäinen and Joo, 2019). No demographic attributes are available in this analysis, and the forty source identities cannot substantiate demographic fairness or an estimate of operational recognition accuracy.

3. Constructing an error without changing the reference budget

We use the captured modified Olivetti MATLAB archive, whose file identity and identity-block ordering are recorded locally (AT&T Laboratories Cambridge, 2026). Its access-year citation refers to the captured archive, not a newly collected cohort. The matrix contains 400 photographs, arranged as ten photographs for each of forty source identities. The experiment uses all forty identities within a known roster.

A pseudorandom generator with seed 20261109 permutes the ten image indices within every identity. The first five become gallery images; the remaining five become queries. This yields 200 gallery photographs and 200 disjoint query photographs. The five-query count is the same for every person, so averaging query correctness and averaging identity-specific accuracy produce the same point estimate in this particular design.

For the error condition, the same generator defines a permutation of the forty identities. One selected gallery photograph from each identity is reassigned to the next identity in that cycle, with the final identity returning to the first. Every reassigned label differs from its source label. Exactly forty of the 200 gallery labels change, representing a twenty-percent intervention on the gallery labels.

The reassignment preserves five gallery photographs under every assigned identity. Each identity loses one of its own images and receives one image from another identity. Neither an image-count audit nor an imbalance statistic would distinguish the clean and perturbed galleries. Their image matrices are identical; the label array and the label-defined centroid memberships are the changing inputs.

This construction is deliberately structured. It is not an estimate of how annotators make mistakes, and the forty label changes are not independent flips. No probability of real annotation error is inferred from the chosen twenty-percent fraction. The source identity labels are treated as operational truth for constructing and scoring the intervention, without claiming to have conducted an independent identity-verification exercise.

Label-noise research demonstrates why the corruption mechanism belongs in the problem definition (Natarajan et al., 2013). Our cyclic, multiclass gallery reassignment differs from an independent binary class-conditional noise model. We use that literature to motivate explicit disclosure of the mechanism, rather than importing a robustness theorem whose assumptions the present experiment does not satisfy.

4. Two matchers and one shared query set

The nearest-gallery matcher selects the gallery photograph with minimum squared Euclidean distance from a query's normalized vector. It then returns that photograph's assigned identity. Since the image vectors do not change between conditions, the selected nearest photograph stays the same. Its returned label can change when that photograph belongs to the reassigned subset.

The centroid matcher first averages the five normalized gallery vectors bearing each assigned identity. It chooses the closest of the forty resulting centroids. Centroids are not normalized again after averaging. Their lengths can therefore vary with within-class photographic spread, and that property is part of the fixed rule. The error condition recomputes centroid membership from the perturbed labels while retaining the same averaging operation.

Neither matcher has a fitted network, classification threshold, rejection option, or calibration set. All queries belong to the supplied identity roster. Query labels enter only when correctness is scored, and no query outcome is used to select a gallery photograph for reassignment. The random seed and reassignment map were fixed before calculating the accuracy contrasts.

Averaging is not itself a robust estimator with a general contamination guarantee. Huber's location-estimation framework and work on robust multivariate estimation make that distinction useful (Huber, 1964; Rousseeuw, 1984). We do not fit either robust procedure. We compare two ordinary rules under one supplied error mechanism and retain the possibility that another mechanism could favor a different rule.

5. Identity-level contrasts

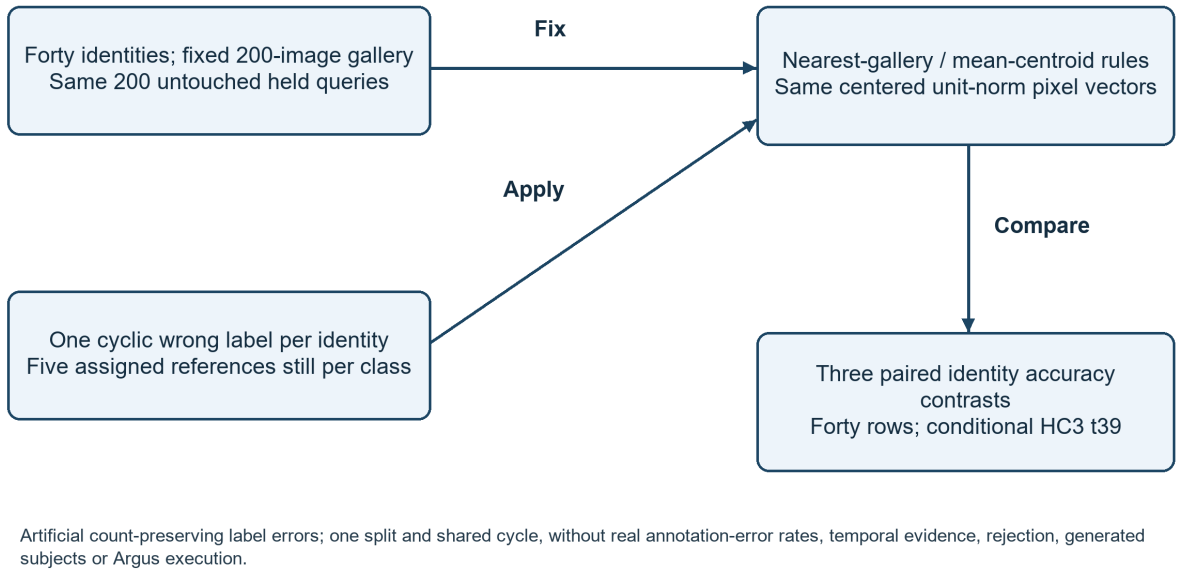


Figure 1. *Clean and count-preserving mislabelled galleries are evaluated on the same queries. The reported interaction compares changes in correctness; gallery pixels and the known identity roster remain fixed.*

For each identity, we calculate the mean correctness of its five queries under both rules and both label conditions. Model 1 summarizes the nearest-gallery change, perturbed minus clean. Model 2 summarizes the centroid change on the same scale. Model 3 takes the centroid change minus the nearest-gallery change within each identity, preserving the pairing rather than subtracting two independently estimated standard errors.

Each regression contains an intercept and forty identity rows, with HC3 covariance and t39 intervals. There are no additional covariates or identity fixed effects. The coefficient is an equal-identity accuracy change. Heteroskedasticity-consistent variance provides a declared working calculation (White, 1980); it does not make the identities independent under a shared cyclic reassignment or account for uncertainty from selecting another gallery split.

Table 1. *Accuracy changes in proportion units. Each model uses N=40 identity summaries, an intercept only, HC3 standard errors and t39 confidence intervals. R-squared is zero apart from numerical rounding. Intervals condition on the fixed photographs, split and label cycle.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.1600 (0.0243)	[-0.2091, -0.1109]	40
M1: fit	R-squared = 0.0000	-	-	40
M2	Mean outcome/contrast	-0.0650 (0.0168)	[-0.0990, -0.0310]	40
M2: fit	R-squared = 0.0000	-	-	40
M3	Mean outcome/contrast	0.0950 (0.0229)	[0.0486, 0.1414]	40
M3: fit	R-squared = 0.0000	-	-	40

Model labels: M1 = Nearest-gallery accuracy change under label error; M2 = Centroid accuracy change under label error; M3 = Centroid minus nearest-gallery change.

The pairing is central to the interpretation. Agreement-oriented methods distinguish correspondence from mere association (Bland and Altman, 1986; Lin, 1989). Our outcome is direct identity correctness, and our contrast follows each person's queries through both conditions. We do not report a correlation between the two matchers as evidence that their accuracy losses are equivalent or that their error patterns coincide.

Measurement-error reasoning also separates a perturbed observed label from the underlying construct of identity (Carroll et al., 2006). The intervention is known in this experiment because we created it. In a natural reference collection, that distinction would require additional verification evidence. A system's successful match to an incorrect supplied label would not automatically establish successful preservation of the intended person.

6. The clean ranking reverses under the declared intervention

Nearest-gallery accuracy is 0.870 with clean labels and 0.710 after reassignment. Its change is -0.160, with standard error 0.024271 and interval [-0.209093, -0.110907]. The aggregate loss is sixteen percentage points. Twenty-five identities have a lower five-query accuracy, fifteen are unchanged, and none has a higher identity-average accuracy under this rule.

The centroid rule starts at 0.825 and ends at 0.760. Its change is -0.065, with standard error 0.016833 and interval [-0.099049, -0.030951]. Fourteen identities lose accuracy, one gains, and twenty-five are unchanged. The positive change for one identity is retained; the result is not defined by selecting only identities whose outcomes deteriorate.

The difference between the two changes is 0.095, with standard error 0.022920 and interval [0.048641, 0.141359]. Under clean labels, nearest-gallery matching leads by 4.5 percentage points. Under the perturbed labels, centroid matching leads by five points. This reversal reflects a difference in sensitivity to this label intervention rather than a universal superiority of the centroid representation.

Percentage points and relative percentage losses would express different quantities (Stevens, 1946). We retain proportion changes in the table and percentage-point descriptions in the prose. No relative-error denominator is needed to compare the two losses. The primary interaction is calculated from paired identity rows, rather than inferred from whether the two separate confidence intervals overlap.

Because nearest-gallery matching always selects the same image, its sensitivity comes directly from how those selected images are labelled. The centroid rule changes the representative vector for two assigned classes per reassignment, so its response is less local. That difference helps explain why equal class counts do not imply equal robustness. It does not reveal which erroneous photograph would be most damaging in another split.

7. What an operational reference audit would still require

The geometry of the centroid change is also inspectable without fitting another model. Each perturbed centroid equals its clean value plus one fifth of the incoming vector minus one fifth of the outgoing vector. Every input vector has unit norm, so the centroid displacement is bounded above by 0.4 in this normalized coordinate space. That upper bound follows from the triangle inequality and the five-image class size; it is not a measured bound on identification error. A much smaller displacement could cross a nearby decision boundary, while a larger displacement might leave every query decision unchanged.

This calculation explains why preserving class counts does not preserve the class representative. The two exchanged vectors can point toward different appearance conditions or different people. The centroid is affected by their difference, whereas the nearest-gallery rule attaches an identity to an individual reference image. Neither rule can infer from the balanced counts that an exchange happened. The relevant information has been removed from the supplied labels, not added to the visual features.

A practical audit could compare the effect of removing a suspicious image with replacing its label, but these would be separate interventions. Removal changes class counts and potentially nearest-reference coverage. Relabelling preserves the image inventory but changes its meaning for the matcher. Our experiment only evaluates the latter. It would be inappropriate to claim that a reference-screening procedure recovered the clean accuracy, since no detector of the deliberately wrong labels was trained or tested.

The working intervals should not be read as coverage guarantees over future populations. A cycle links the reference contents of different identities, and all queries share one gallery. Influence analysis encourages examining how a few units affect a reported contrast (Cook, 1977); a stronger follow-up would repeat prespecified splits and error maps rather than treating the current forty rows as forty independently generated reference systems.

Distribution-shift studies show why an internal stress result cannot substitute for transport evidence (Ovadia et al., 2019). Validation design must match the intended deployment boundary (Roberts et al., 2017), and evaluation choices must remain outside outcome-driven tuning (Varma and Simon, 2006). Here that boundary includes a known identity roster, historical grayscale images and an artificial label cycle. Unknown people, diverse capture conditions and adversarial reference replacement remain outside the evidence.

The saved gallery table records source rows, source identities and assigned identities; the query table records correctness in all four conditions. Dataset documentation makes those meanings explicit (Gebru et al., 2021), and provenance supports locating the exact captured input (Wilkinson et al., 2016). Maintaining the label map as an input artifact also addresses the data dependencies highlighted in work on

technical debt (Sculley et al., 2015). A balanced reference count is therefore a useful inventory fact, but reference integrity needs its own evidence and its own sensitivity test.

References

1. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.
2. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.
3. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.
4. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.
5. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
6. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.
7. Kärkkäinen, Kimmo and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age" *arXiv preprint arXiv:1908.04913*, 2019. https://arxiv.org/abs/1908.04913.
8. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
9. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.
10. Samaria, F.S.; Harter, A.C.. (1994). Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. https://doi.org/10.1109/acv.1994.341300.
11. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.
12. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
13. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
14. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
15. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
16. Cao, Qiong, et al. "VGGFace2: A dataset for recognising faces across pose and age" *arXiv preprint arXiv:1710.08092*, 2017. https://arxiv.org/abs/1710.08092.
17. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
18. AT&T Laboratories Cambridge. (2026). Modified Olivetti faces: Sam Roweis MATLAB version [Data archive; mirror accessed 7 November 2026]. File identity and row-label layout verified against the scikit-learn maintainer loader. https://ndownloader.figshare.com/files/5976027.
19. Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.
20. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.
21. Kemelmacher-Shlizerman, Ira, et al. "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale" *arXiv preprint arXiv:1512.00596*, 2015. https://arxiv.org/abs/1512.00596.
22. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

23. Natarajan, Nagarajan; Dhillon, Inderjit S.; Ravikumar, Pradeep; Tewari, Ambuj. (2013). Learning with Noisy Labels. Advances in Neural Information Processing Systems 26. https://papers.nips.cc/paper_files/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html.

24. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

25. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

26. Xiaoyang Tan; Triggs, Bill. (2010). Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2010.2042645.

27. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

28. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

29. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

30. Wang, Hao, et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.