Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

Which Label Does a Completion Benchmark Know? Privileged and Observable Digit Prototypes

Ian Scott[1] · Jonathan Wells[2] · Paul Lewis[3]

[1] Portugal · [2] Spain · [3] Austria

Original publication: October 2, 2026 | Article 679

Abstract

A completion rule can appear stronger when the benchmark supplies information unavailable in use. On 1,797 held handwritten digits, a prototype selected with the true digit label had hidden-feature MSE 0.07848. Selecting the prototype from sixty visible features instead gave 0.08549; both improved on the unconditional mean prototype at 0.14362. The observable selector chose the correct class for 1,580 images. The remaining gap illustrates the value of privileged category information in this particular rule, without establishing that true-label selection is an optimal reconstruction or that any existing inpainting benchmark leaks labels.

Controlled comparison of information conditions in a public feature benchmark. True test labels are used only in an explicitly privileged reference condition.

Which Label Does a Completion Benchmark Know? Privileged and Observable Digit Prototypes

> Controlled comparison of information conditions in a public feature benchmark. True test labels are used only in an explicitly privileged reference condition.

Abstract

A completion rule can appear stronger when the benchmark supplies information unavailable in use. On 1,797 held handwritten digits, a prototype selected with the true digit label had hidden-feature MSE 0.07848. Selecting the prototype from sixty visible features instead gave 0.08549; both improved on the unconditional mean prototype at 0.14362. The observable selector chose the correct class for 1,580 images. The remaining gap illustrates the value of privileged category information in this particular rule, without establishing that true-label selection is an optimal reconstruction or that any existing inpainting benchmark leaks labels.

1. The benchmark's information boundary

A reconstruction comparison is defined by more than the pixels or features hidden from a model. It is also defined by the auxiliary information that the model may use. A class label can convey a strong structural prior even when no hidden feature value is shown directly. If an evaluator supplies the true label to one completion rule and expects another to infer it, the apparent reconstruction difference combines model behavior with a difference in information.

We ask how much a simple completion benchmark changes when the prototype is selected using the true class rather than a class inferred from visible features. The comparison uses public handwritten-digit block counts with four central features concealed. It is deliberately elementary: all prototypes are training means, and only the rule choosing among them changes. This makes the auxiliary-information boundary inspectable without relying on a neural network's internal representations.

The problem is relevant to restoration of structured symbols. Meng et al., 2025 presents zero-shot inpainting for oracle-bone inscription rubbings. Our experiment concerns ordinary digit features rather than inscriptions, and it does not run the visual Mamba model. Its contribution is an evaluation distinction: reconstruction conditional on a supplied category is a different task from reconstruction when that category must itself be inferred from the remaining observations.

Our analytical propositions are conditional rather than universal. Class-specific training means may represent hidden digit structure better than one unconditional mean. However, replacing a true class with an inferred class can introduce additional reconstruction error. The magnitude of that penalty depends on the errors the selector makes and on how different the corresponding hidden prototypes are. Classification accuracy alone does not determine the completion penalty.

2. Semantic priors and legitimate task definitions

Context encoders Pathak et al., 2016 organize completion around observed image context. Partial convolutions Liu et al., 2018 and gated convolutions Yu et al., 2018b provide other mechanisms for handling missing regions, while EdgeConnect Nazeri et al., 2019 introduces an additional structural representation. These methods highlight the potential value of information surrounding a gap. The present prototype system does not reproduce any of them; it isolates a different source of information, namely a category supplied to the reconstruction rule.

LaMa Suvorov et al., 2021 and RePaint Lugmayr et al., 2022 illustrate broader approaches to recovering large missing regions. MambaIR Guo et al., 2024 develops another restoration architecture. They motivate distinguishing an architecture's capacity from the conditions under which it is evaluated. A change in supplied labels can alter the task even if the same reconstruction machinery is used. We make no claim that these cited systems use labels improperly or that their reported results are affected by the mechanism tested here.

Class information changes the interpretation of a representation. Belhumeur et al., 1997 contrasts class-specific projections in face recognition, a different task that illustrates why supervised information must be named. Oracle-character data in Wang and Deng, 2024, Japanese character data in Clanuwat et al., 2018, and digit-recognition work in Lecun et al., 1998 provide distinct symbol contexts. None is pooled into this experiment. Their class structures and acquisition histories do not establish that a privileged digit prototype supports interpretation of archaeological symbols.

Leakage research Kapoor and Narayanan, 2023 motivates making the information allowed at evaluation explicit. Importantly, a deliberately labeled conditional task is legitimate when that label will also be available in use. The privileged condition becomes misleading only if its performance is presented as that of an unlabeled system. We therefore report it as a reference information condition, alongside an observable route and an unconditional baseline, rather than naming it a deployable inpainting method.

3. Data, features, and three reconstruction routes

We use the official training and test feature files of Optdigits Alpaydin and Kaynak, 1998. They contain 3,823 training images and 1,797 test images. Each digit is represented by sixty-four block-count features on an eight-by-eight grid. Counts range from zero to sixteen and are divided by sixteen. Four central entries, with zero-based flattened indices 27, 28, 35, and 36, are the reconstruction targets. The remaining sixty entries are visible under every condition.

The training file contains all ten digit classes, with between 376 and 389 examples per class. We estimate one sixty-four-feature mean for each class and one unconditional mean over all training examples. Neither the means nor the mask depend on held test errors. The official test labels are kept separate from the observable selector and used only to define the privileged condition and to describe its classification agreement afterward.

The unconditional route assigns every test image the same four hidden-feature means. It ignores visible features and labels, providing a simple baseline for the target's marginal variation. The true-label route selects the training prototype corresponding to the test image's known digit category. It uses no hidden test values, but it has access to category information that is unavailable in the declared unlabeled restoration setting.

The observable route compares the test image's sixty visible entries with the corresponding entries in each training class prototype. It selects the nearest prototype under unweighted squared Euclidean distance, breaking a tie in favor of the smallest digit index. Its hidden prediction is then the selected prototype's four concealed entries. The same class means appear in the true-label and observable routes; only the selection information differs.

No route estimates a new per-image latent representation, adjusts the hidden prototype using a visible residual, or clips a prediction after selection. A mean of bounded source features already lies within their bounds. This is a prototype experiment rather than a claim about the best linear predictor. Principal-component constructions Pearson, 1901 and probabilistic PCA Tipping and Bishop, 1999

provide different representation choices, while ridge regression Hoerl and Kennard, 1970 would allow a continuous mapping from visible to hidden features. They are not additional fitted competitors here.

4. What is compared and why it is paired

For each route, image loss is the mean squared error over the four hidden entries. We calculate three paired differences: true-label minus unconditional loss, observable minus true-label loss, and observable minus unconditional loss. All comparisons use the same 1,797 test images and target entries. Each image contributes one scalar per difference, avoiding an artificial sample size of four times as many independent observations.

Denoising representation learning Vincent et al., 2008 makes clear that a model can learn useful structure from corrupted observations. The current design asks a prior question: what information was supplied before the reconstruction was produced? Measurement-error reasoning Carroll et al., 2006 is relevant because a predicted label is an imperfect proxy for a true label. A class-selection error can propagate into the hidden prediction, even though the prototype estimates themselves remain unchanged.

Figure 1 separates training prototype estimation from the three test information routes. The true-label arrow is explicitly privileged. This label is not used to choose the mask, fit a class mean, tune a distance, or modify a failed observable prediction. The diagram therefore describes a controlled computational comparison with a disclosed information advantage, rather than an accidental contamination of the observable route.

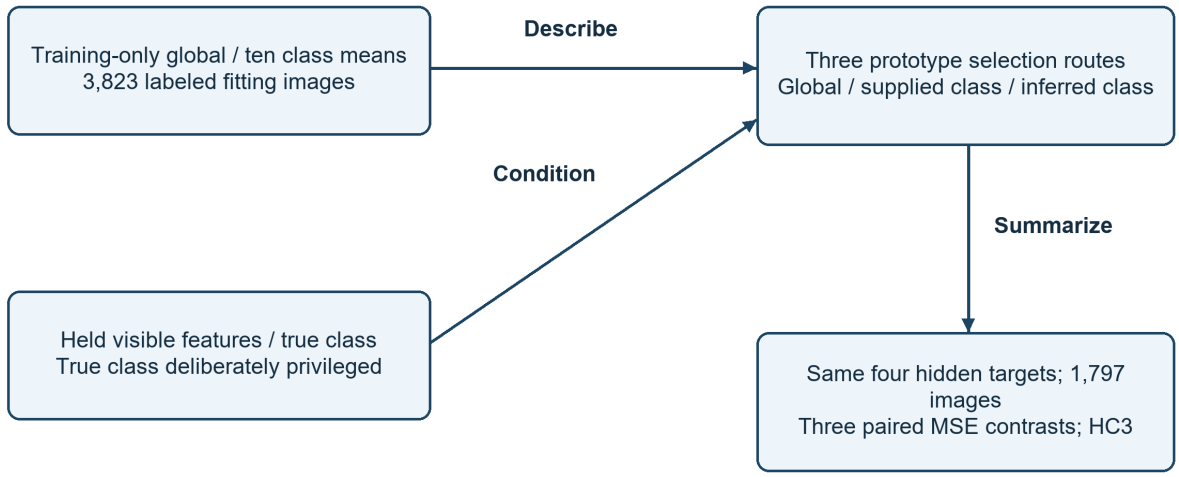

Figure 1. *All prototypes use the official training partition. Only the privileged route reads the held true class; the observable route selects a class from sixty visible features.*

Three intercept-only OLS models summarize the paired image differences. HC3 standard errors White, 1980 and t intervals use 1,796 residual degrees of freedom. The data documentation describes different training and test writers, but individual test-writer identifiers are not present in the feature file. Cluster guidance Cameron and Miller, 2015 therefore motivates a limitation rather than an invented correction: the reported intervals do not account for within-writer dependence or uncertainty from estimating the training means.

5. Results: class information helps, and inference has a cost

The unconditional hidden-feature MSE is 0.143619. Supplying the true category reduces it to 0.078478, a paired difference of -0.065141. The observable route has MSE 0.085487. Thus, inferring the category from visible features retains much of the improvement over the unconditional prototype but introduces an average penalty of 0.007009 relative to true-label selection.

Table 1. Paired changes in hidden-feature MSE across information conditions.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.065141 (0.002529)	[-0.070100, -0.060181]	1797
M1: fit	R-squared = 0.0000	-	-	1797
M2	Mean outcome/contrast	0.007009 (0.001372)	[0.004317, 0.009700]	1797
M2: fit	R-squared = 0.0000	-	-	1797
M3	Mean outcome/contrast	-0.058132 (0.002688)	[-0.063403, -0.052861]	1797
M3: fit	R-squared = 0.0000	-	-	1797

Model labels: M1 = True-label prototype minus global prototype; M2 = Visible-inferred label minus true-label prototype; M3 = Visible-inferred label minus global prototype.

The models contain one intercept and no controls. They describe mean paired loss differences over 1,797 images, using HC3 uncertainty and t-based intervals. Model 1's interval is [-0.070100, -0.060181]. Model 2's interval is [0.004317, 0.009700]. Model 3 estimates observable-minus-unconditional loss at -0.058132, with interval [-0.063403, -0.052861]. R-squared is effectively zero in these mean-only models.

The observable class selector is correct on 1,580 images, or 87.92 percent. For those images it chooses exactly the same prototype as the true-label route, so their reconstruction losses are identical by construction. The remaining 217 images account for the entire average gap between those two routes. Within that selected error subgroup, true-label MSE is 0.113412 and observable-route MSE is 0.171451, a mean difference of 0.058039.

This subgroup comparison is descriptive and uses selector correctness defined on the same images. It is not an independent test or a causal subgroup effect. It nonetheless clarifies the mechanism inside the declared pipeline: a changed class index changes the hidden prototype. It also explains why a high classification accuracy does not imply that the small remaining set of classification errors is harmless for completion.

The true-label route is not necessarily the lower-loss route for every individual image. A coarse class mean ignores within-class handwriting differences. An incorrectly selected neighboring class might occasionally have hidden means closer to a particular image's target entries. We do not suppress such possibilities or reinterpret the privileged average as a mathematical lower bound for all completion systems. The measured advantage belongs to these fitted prototypes and this held cohort.

6. Separating information value from benchmark quality

A squared-error objective Gneiting and Raftery, 2007 evaluates numerical prediction under its declared target, but it does not establish readability or preservation of a character's semantic identity. Agreement methods Bland and Altman, 1986 motivate keeping numerical agreement separate from broader claims about interchangeable measurements. A lower error over four block counts can coexist with a visually or linguistically unimportant change. No human reader or character-recognition system evaluates the completed features in this experiment.

Validation guidance Roberts et al., 2017 and distribution-shift analysis Ovadia et al., 2019 also limit extrapolation from the official split. The feature files separate source writers, but the present uncertainty conditions on one training fit and the available test sample. It does not estimate a distribution of prototype performance across alternative writing populations. Feature scaling, mask position, and class granularity are additional fixed conditions that a future replication should vary deliberately.

Causal-graph reasoning Greenland et al., 1999 helps resist a different overinterpretation: the category's usefulness here does not prove that semantic information is the cause of successful archaeological restoration. The experiment alters which training prototype is selected. It observes no human semantic reasoning, no archaeological capture process, and no independent restoration outcome beyond block-count error. The mechanism demonstrated is the computational effect of an information condition inside a specified rule.

Missing-data principles Little and Rubin, 2002 remind us that the four hidden entries are deliberately concealed, rather than naturally missing. Their positions are fixed for every image. The study cannot infer what happens when damage follows strokes, when category labels are themselves uncertain, or when damaged regions contain the information needed to distinguish classes. Those changes could alter both the observable selector and the relative usefulness of a supplied label.

7. Reporting the allowed information

A clear reconstruction report should identify whether labels, text descriptions, reference images, or semantic annotations are available during fitting, evaluation, and eventual use. Nested-validation concerns Varma and Simon, 2006 explain why repeated test-based tuning would add another information advantage; no such tuning is part of this calculation. The true-label condition is retained openly so that its result can be interpreted as a comparison of task definitions, rather than silently entering the observable pipeline.

Dataset documentation Gebru et al., 2021 supports this explicit account of partitions, features, labels, and mask construction. FAIR principles Wilkinson et al., 2016 motivate retaining the original archive and reproducible processed losses. Neither practice proves that a benchmark covers all relevant restoration phenomena. They make the evidence inspectable and allow a subsequent study to change one information condition without accidentally changing several others.

The substantive implication is modest but useful. A supplied category substantially improves a class-prototype baseline, and most of that improvement remains when a crude visible-feature selector chooses the category. The residual penalty is real in the held feature sample and concentrated among classification disagreements. Reporting both routes reveals an information advantage that a single reconstruction score would conceal.

The numerical comparison also separates two kinds of supervision. Training labels are used to estimate the class means under both prototype-selection routes. The observable route is therefore not an

unsupervised completion system. Its distinction from the privileged route is narrower: it does not read the true label of the held image. Calling it label-free without this qualification would conceal a training-information requirement that the implementation plainly has.

The unconditional prototype is intentionally simple, so its large gap should not become a general estimate of the value of semantics. A stronger continuous predictor using the same sixty visible features might reduce that gap without ever predicting a digit class. Such a comparison would answer another modeling question. The present experiment holds the prototype family fixed to isolate the information used for selection, and it reports the resulting scope rather than promoting the simplest baseline into a universal restoration standard.

8. Conclusion

The same fitted prototype collection produces different completion errors depending on how its class index is obtained. True-label MSE is 0.07848, compared with 0.08549 for visible-inferred selection and 0.14362 for an unconditional mean. These figures characterize an explicitly bounded information experiment. They support declaring auxiliary information in restoration benchmarks while leaving optimal reconstruction, semantic fidelity, and the behavior of specialized inpainting systems untested.

References

1. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
2. Vincent, Pascal; Larochelle, Hugo; Bengio, Yoshua; Manzagol, Pierre-Antoine. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390294.
3. 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.
4. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
5. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.
6. 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.
7. 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.
8. Nazeri, Kamyar; Ng, Eric; Joseph, Tony; Qureshi, Faisal Z.; Ebrahimi, Mehran. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv:1901.00212. https://arxiv.org/abs/1901.00212.
9. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.
10. 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.
11. 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.
12. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.
13. Clanuwat, Tarin; Bober-Irizar, Mikel; Kitamoto, Asanobu; Lamb, Alex; Yamamoto, Kazuaki; Ha, David. (2018). Deep Learning for Classical Japanese Literature. https://arxiv.org/abs/1812.01718.
14. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.

15. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.

16. 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.

17. Guo, Hang; Li, Jinmin; Dai, Tao; Ouyang, Zhihao; Ren, Xudong; Xia, Shu-Tao. (2024). MambaIR: A Simple Baseline for Image Restoration with State-Space Model. arXiv:2402.15648. https://arxiv.org/abs/2402.15648.

18. Alpaydin, E.; Kaynak, C. (1998). Optical Recognition of Handwritten Digits [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.

19. Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. https://doi.org/10.1080/14786440109462720.

20. 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.

21. Wang, Mei; Deng, Weihong. (2024). A dataset of oracle characters for benchmarking machine learning algorithms. Scientific Data, 11, 87. https://doi.org/10.1038/s41597-024-02933-w.

22. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.

23. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.

24. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. https://doi.org/10.1109/5.726791.

25. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

26. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

27. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.

28. 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.

29. 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.

30. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.