Uncertainty and Assessment
Volume 1 · Number 1 · 2026

Pixel Coverage Is Not Region Coverage: Residual Calibration for a Masked Character Reconstruction

Patrick Warren[1] · Spencer Allen[2] · Gavin Bell[3]

[1] Germany · [2] France · [3] United Kingdom

Original publication: October 9, 2026 | Article 665

Abstract

A restoration interval can cover most concealed pixels while missing something in most restored regions. With a fixed linear reconstructor on Oracle-MNIST, a pooled-pixel residual radius covered 89.3% of hidden test pixels but only 17.2% of complete sixty-four-pixel regions. Calibrating the maximum error per image raised region coverage to 86.2%, while average interval width increased from 0.720 to 0.971 on a zero-to-one intensity scale. The comparison exposes a substantial coverage-width tradeoff and keeps numerical reconstruction uncertainty separate from confidence in reading a character.

Exploratory research draft using an internal split of a released character-image test archive. No untouched benchmark, semantic restoration, or submission-ready claim.

Pixel Coverage Is Not Region Coverage: Residual Calibration for a Masked Character Reconstruction

> Exploratory research draft using an internal split of a released character-image test archive. No untouched benchmark, semantic restoration, or submission-ready claim.

Abstract

A restoration interval can cover most concealed pixels while missing something in most restored regions. With a fixed linear reconstructor on Oracle-MNIST, a pooled-pixel residual radius covered 89.3% of hidden test pixels but only 17.2% of complete sixty-four-pixel regions. Calibrating the maximum error per image raised region coverage to 86.2%, while average interval width increased from 0.720 to 0.971 on a zero-to-one intensity scale. The comparison exposes a substantial coverage-width tradeoff and keeps numerical reconstruction uncertainty separate from confidence in reading a character.

1. Which object does a restoration interval cover?

An inpainting output can be evaluated one pixel at a time or as a complete reconstructed region. These are different objects. A system may give reasonable uncertainty for a typical concealed pixel while leaving at least one unsupported pixel in many images. If the intended claim concerns an entire missing stroke region, an average pixel-level coverage statement does not answer it. The distinction is easy to lose when a report presents one nominal confidence percentage without defining the unit that must be covered.

Our research question asks whether calibrating individual concealed-pixel errors yields the same whole-region coverage as calibrating the maximum concealed error per image. We evaluate two residual-radius constructions around the same deterministic linear prediction. The public oracle-character resource described by Wang and Deng, 2024 supplies the numerical images. Its class labels are not used in fitting or calibration. The study concerns artificial central concealment in existing complete images, rather than reconstruction of genuinely missing historical material whose pixel values cannot be observed.

Meng et al., 2025 develop zero-shot inpainting for oracle-bone inscription rubbings using a visual Mamba component. That setting motivates asking what a restoration uncertainty statement actually promises. We do not execute its architecture or assign our interval results to it. Our deliberately limited reconstructor makes the two calibration units transparent. It is useful for identifying an evaluation distinction before a more expressive generative model introduces multiple plausible outputs, semantic ambiguity, and additional sources of uncertainty.

The analytical proposition is that calibrating the imagewise maximum residual should yield wider intervals and higher empirical whole-region coverage than pooling all residual pixels under the same nominal level. This is not a claim that wider intervals are automatically useful. A band approaching the full intensity range may achieve coverage while offering little information about the concealed content. We therefore evaluate coverage and width together, preserving the possibility that a mathematically appropriate region-level construction is practically uninformative for a weak predictor.

2. Prediction sets and the unit of exchangeability

Conformal prediction provides a framework for constructing prediction sets from calibration experience Shafer and Vovk, 2008. Distribution-free predictive inference for regression Lei et al., 2018 supplies related methodological context. Their relevance here is the need to define one score for the unit whose future coverage is intended. We use one imagewise maximum residual as that score for a whole concealed region. The pooled comparator instead collects individual pixel residuals from the same images, which are spatially dependent and positionally heterogeneous.

A nominal level of ninety percent does not mean that at least ninety percent of every finite held batch must be covered. Coverage statements depend on their sampling interpretation and assumptions, including appropriate calibration-test exchangeability. They do not give a guarantee conditional on every realized calibration sample or every character class. Our experiment reports what happened in one internal split and retains the finite calibration construction. It neither declares a formal guarantee for a new archive nor treats a shortfall in one realized batch as a refutation of the general framework.

Probability calibration is a nearby but different object. The neural-network calibration work of Guo et al., 2017 concerns predicted confidence values, whereas our reconstruction rule supplies numerical intensities and residual bands. Dataset-shift research Ovadia et al., 2019 further cautions against assuming that uncertainty behavior survives a change of data source. We do not estimate a probability that a reconstructed character is readable or correct. The intervals address the released hidden pixel intensities under the fixed artificial mask and the selected image distribution.

Reliability terminology should also remain precise. Inter-rater frameworks Shrout and Fleiss, 1979 and guidance on intraclass correlations Koo and Li, 2016 concern repeated measurements and specified raters or instruments. No such response matrix is present here. Agreement methodology Bland and Altman, 1986 and measurement-scale discussion Stevens, 1946 reinforce the need to name the measured quantity. Our outcome is numerical inclusion of original grayscale values, not agreement among archaeologists, reliability of a character transcription, or confidence that an inferred stroke corresponds to an authentic historical marking.

3. A transparent reconstruction experiment

The released Oracle-MNIST test archive contains 3,000 twenty-eight-by-twenty-eight grayscale images. We transparently repurpose that archive for an exploratory internal split: 2,000 images fit the reconstructor, 500 calibrate the residual radii, and 500 form the final internal evaluation set. A single permutation generated with seed 20261107 assigns the rows. This is not reported as performance on an untouched official test set. The original archive and row partition are retained so the distinction is reviewable.

Pixel values are divided by 255 to obtain the zero-to-one scale. The fixed mask conceals an eight-by-eight central block, leaving 720 pixels observed and sixty-four hidden in every image. The fit subset supplies a mean image and the thirty-two leading eigenvectors of its sample covariance. This linear approximation follows the dimensional-reduction tradition of Pearson, 1901, with a latent-variable connection to principal-component modeling Tipping and Bishop, 1999. The implementation performs an eigendecomposition rather than a generative neural training procedure.

For each calibration or evaluation image, visible pixels determine latent coefficients through a ridge solve with penalty 0.01. The role of ridge regularization follows the motivation of Hoerl and Kennard, 1970. Hidden-pixel predictions are reconstructed from those coefficients and clipped to zero through one. Numerical assertions check basis orthonormality and the normal-equation residual. The rank, mask,

penalty, and clipping bounds stay fixed for both interval policies and are not chosen by examining their evaluation coverage.

This is a dense low-rank image model with a spatially concentrated mask. Matrix-completion research Candes and Plan, 2010 supplies context for why low-rank assumptions and observation patterns matter, but we do not invoke a completion theorem for this particular mask or covariance fit. Missing-data theory Little and Rubin, 2002 likewise helps distinguish deliberate concealment from natural damage. The experiment knows the concealed values because it hides complete images; genuine restoration would require other evidence to judge whether a plausible reconstruction is historically justified.

4. Two radii around the same predictions

The pooled rule collects the absolute errors of all sixty-four hidden pixels from each of the five hundred calibration images, producing 32,000 scores. Its radius is the 28,801st ordered score, following the declared ceiling rule applied to ninety percent of one plus the score count. That gives radius 0.4404091. This pooling supplies a numerical comparator, not a formal claim that all calibration pixels are independent or share the same residual distribution. Pixel positions within an image are not separate exchangeable images.

The region rule first takes the maximum absolute hidden-pixel error within each calibration image. It therefore has five hundred scores and uses the 451st ordered maximum, giving radius 0.8078431. For both rules, an interval is centered at the clipped predicted intensity and intersected with the known zero-to-one range. Membership is evaluated by the corresponding inclusive absolute-residual comparison. A numerical diagnostic confirms agreement with direct endpoint membership; adding a one-trillionth tolerance changes no whole-image coverage count.

Intersecting with the intensity range has a particular interpretation. Every recorded target value already lies between zero and one, so removing interval portions outside that support cannot remove a valid target from an otherwise covering interval. It can reduce the measured width substantially. Consequently, twice the residual radius is not the reported average width: the image-specific predictions determine how much each interval is cut off at the lower or upper boundary. The script calculates width from the actual clipped endpoints for every hidden pixel. This distinction prevents a reader from confusing a common residual radius with a constant-width band after respecting the outcome support.

The two policies also share exactly the same prediction error. A coverage gain here cannot be presented as improved reconstruction, because no point prediction changes. Better point estimates could permit narrower calibrated intervals in another experiment, but the present comparison changes only how residual evidence is aggregated. Its question is therefore about the scope of an uncertainty statement, rather than about which restoration method produces a more accurate image.

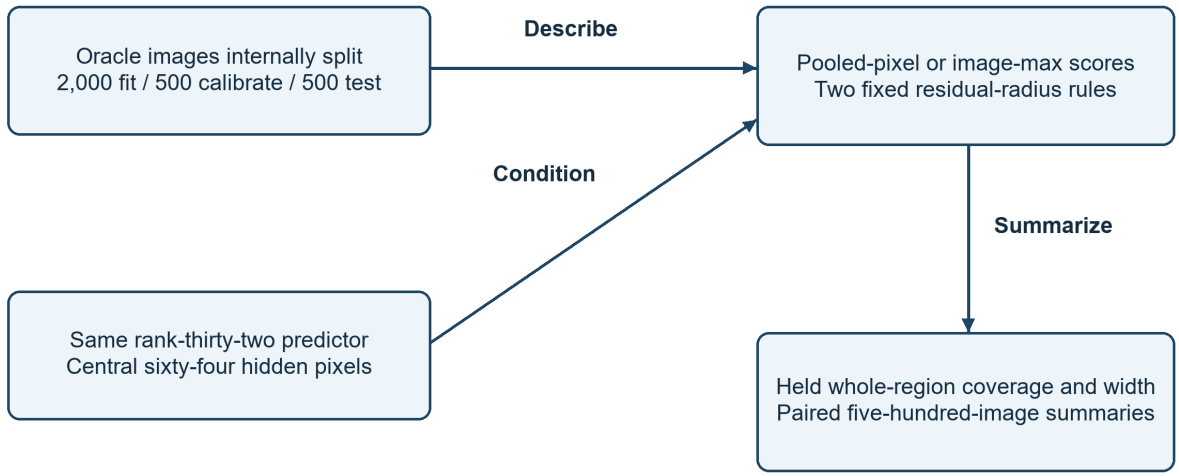

Figure 1. Shared reconstructor, separate calibration units, and held-region coverage-width outcomes.

Figure 1 makes the only changing component explicit: the residual calibration unit. Fit images, prediction rule, test images, and hidden pixels are shared. The pooled rule targets a collection of pixel residuals; the region rule targets an imagewise maximum. Test outcomes include mean covered-pixel share, the indicator that every hidden pixel is covered, and average clipped interval width. None is interpreted as a semantic confidence score or an output quality ranking for a neural restoration model.

5. Estimation and conditional uncertainty

For each held image, we compute the difference between region-rule and pooled-rule whole-region coverage. We also compute their difference in average interval width across its sixty-four hidden pixels. Separate intercept-only OLS models summarize these paired quantities over five hundred images. The models use HC3 standard errors and normal-reference intervals, following robust covariance principles White, 1980. They contain no covariate controls or fixed effects, and their R-squared is zero by construction.

The intervals condition on the fitted representation and the realized calibration sample. They do not include repeated partitioning or uncertainty from choosing another training set. Individual writer identifiers are unavailable, so they also do not provide writer-cluster inference. The guidance in Cameron and Miller, 2015 makes clear that a defensible clustering variable cannot be invented from row order. Structured-validation principles Roberts et al., 2017 instead motivate stating the actual split and the dependence information that the source does not supply.

No rank or radius policy is optimized against evaluation outcomes. This protects the stated comparison from the model-selection problem discussed by Varma and Simon, 2006, although a single frozen split remains limited. Measurement-error reasoning Carroll et al., 2006 further distinguishes calibration of numerical residuals from validity of the grayscale target as a measure of historical truth. The recorded complete image is the evaluation reference for this experiment, not evidence that every scanned feature is an authentic or interpretable stroke.

6. Region coverage improves at a substantial width cost

The pooled radius covers 28,575 of 32,000 hidden evaluation pixels, for pixel coverage 0.8929688. However, only eighty-six of five hundred images have every concealed pixel covered, giving whole-region coverage 0.172. This result directly demonstrates the unit mismatch: a near-ninety-percent covered-pixel share coexists with low complete-region coverage. It does not require an independence approximation among the sixty-four pixels, because both quantities are computed from their actual joint outcomes in each image.

The region radius covers 31,830 hidden pixels, for pixel coverage 0.9946875, and completely covers 431 images, for region coverage 0.862. Its intervals contain those of the pooled rule because its radius is larger. Consequently, 345 additional images become fully covered and none loses coverage in this comparison. Yet the observed 86.2% region coverage remains below the nominal ninety percent in this single realized split. We report that shortfall without reselecting the calibration radius using evaluation labels.

Table 1. Paired held-image coverage and width differences

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.6900 (0.0207)	[0.6494, 0.7306]	500
M1: fit	R-squared = 0.0000	-	-	500
M2	Mean outcome/contrast	0.2502 (0.0018)	[0.2466, 0.2538]	500
M2: fit	R-squared = 0.0000	-	-	500

Model labels: M1 = Whole-region minus pooled-radius image coverage; M2 = Whole-region minus pooled-radius mean interval width.

Table 1 reports intercept-only OLS models with HC3 standard errors and normal-reference intervals, N500 for both outcomes, no controls, and no fixed effects. M1 estimates the region-rule minus pooled-rule whole-region coverage difference; M2 estimates their mean interval-width difference. Width is measured on the normalized intensity scale after intersecting each interval with zero through one. These are conditional paired summaries, not formal repeated-calibration coverage certificates or semantic restoration scores.

The whole-region gain is 0.6900000, with standard error 0.0207248 and interval from 0.6493802 to 0.7306198. Average interval width rises from 0.7204785 to 0.9706735. The increase is 0.2501950, with standard error 0.0018192 and interval from 0.2466295 to 0.2537604. The latter width is close to the entire available intensity range. The coverage gain therefore comes with a major loss of specificity, rather than an unequivocal improvement in a useful uncertainty statement.

7. Consequences for restoration evaluation

The two policies do not change reconstructed point values. They change how uncertainty around those values is summarized. Scoring and estimation principles Gneiting and Raftery, 2007 favor stating the object that a set or score evaluates. An imagewise maximum is appropriate to a complete-region inclusion question, while a pooled residual summary answers a different empirical question. Neither automatically assesses recognition accuracy, stroke topology, or whether a reader can distinguish the restored character from plausible alternatives.

More capable restoration models could alter the tradeoff. Context Encoders Pathak et al., 2016, large-mask Fourier inpainting Suvorov et al., 2021, diffusion-based RePaint Lugmayr et al., 2022, and the deep image prior Ulyanov et al., 2017 organize different reconstruction mechanisms. State-space

restoration and vision work Guo et al., 2024 and Liu et al., 2024 offer further architectural context. None is implemented here. It would be unsupported to infer that any one of these methods would yield the observed coverage, or that the wide intervals represent an unavoidable limit of character restoration rather than a limitation of this simple predictor.

There are also several useful unexecuted checks. Repeating the internal split would reveal how much the radius and coverage move with calibration sampling. Separate masks would test whether the central region is unusually difficult. Character-conditional summaries could identify heterogeneity that an image-marginal statement omits, although they would need adequate calibration support. Influence diagnostics Cook, 1977 could inspect high-residual calibration images without automatically deleting them. These extensions would address distinct uncertainties and should be performed explicitly rather than implied by the present table.

8. Data record and conclusion

The workflow saves the source checksum, internal row partition, fitted-method settings, imagewise calibration maxima, residual radii, per-image test outcomes, and structured regression results. Source attribution and repository terms accompany the images. Dataset documentation Gebru et al., 2021 makes the repurposed official test archive and artificial mask visible, while FAIR stewardship Wilkinson et al., 2016 supports preserving machine-readable provenance. These records enable numerical reconstruction without claiming an external originality or scientific-validity certificate.

The empirical answer is clear within the declared experiment. Pooling concealed-pixel errors produced near-nominal pixel coverage while leaving most whole regions incompletely covered. Calibrating imagewise maxima greatly increased region coverage but produced bands almost as wide as the full intensity scale. A restoration uncertainty report should therefore identify its coverage unit and show its width. Without both, a single confidence percentage can conceal how little has actually been established about the missing region.

References

1. Lei, Jing; G'Sell, Max; Rinaldo, Alessandro; Tibshirani, Ryan J.; Wasserman, Larry. (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2017.1307116.
2. 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.
3. 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.
4. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.
5. Shrout, Patrick E.; Fleiss, Joseph L.. (1979). Intraclass correlations: Uses in assessing rater reliability.. Psychological Bulletin. https://doi.org/10.1037/0033-2909.86.2.420.
6. 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.

7. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.

8. Liu, Yue; Tian, Yunjie; Zhao, Yuzhong; Yu, Hongtian; Xie, Lingxi; Wang, Yaowei; Ye, Qixiang; Jiao, Jianbin; Liu, Yunfan. (2024). VMamba: Visual State Space Model. arXiv:2401.10166. https://arxiv.org/abs/2401.10166.

9. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

10. 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.

11. 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.

12. Shafer, Glenn; Vovk, Vladimir. (2008). A Tutorial on Conformal Prediction. Journal of Machine Learning Research,9(12),371-421. https://www.jmlr.org/papers/v9/shafer08a.html.

13. 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.

14. 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.

15. 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.

16. Candes, Emmanuel J; Plan, Yaniv. (2010). Matrix Completion With Noise. Proceedings of the IEEE. https://doi.org/10.1109/jproc.2009.2035722.

17. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.

18. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

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. 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.

21. 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.

22. 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.

23. 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.

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

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

26. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

27. 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.

28. 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.

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

30. 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.