WHITE PAPER / TRAINING DATA

The class that disappeared in the average

A repeated labeling mistake can be rare across a whole dataset and common within one class. We measured both views of the same fault.

The annotation error with a direction

An indexing team discovers that one batch of documents went through an outdated mapping. Every affected item moved to the same neighboring category. An overall quality score tells the team how often the model is right. It gives little indication of which queue now receives the wrong work.

ICLR research into robustness under label noise motivated our test of a concentrated error. We kept the model, training images and evaluation set fixed. The sole change was to the assigned class of some handwritten eights in the training data.

Change the labels and keep the images

We used 1,797 images from the public handwritten digits dataset and split them into 1,347 training images and 450 held-out images. The split was stratified by digit with seed 42. The held-out set included 43 eights. We divided the pixel intensities by 16 and fitted a logistic regression classifier.

For the second fit, a seeded random draw selected 65 training eights, half the training examples for that class rounded down. We relabeled those images as threes. This changed 4.8% of all training labels. We evaluated both classifiers against the original held-out labels.

A six-point drop hid a much larger failure

Held-out results after changing 65 training labels

Measure	Original labels	Altered labels
Correct overall	433 / 450	406 / 450
Eights correctly identified	38 / 43	9 / 43
Eights read as threes	0 / 43	29 / 43

Overall accuracy moved from 96.2% to 90.2%. Recall for eights fell from 88.4% to 20.9%. The destination of the errors was especially informative. Twenty-nine held-out eights now went to the same class that had received the altered training labels.

Trace both ends of a handoff

For a new label export, compare the old and new mappings on a trusted set. Record both the true class and the predicted destination. If one route changes sharply, inspect examples from the source class and the class receiving the extra work.

The injected fault deliberately targets one pair of digits. Its size gives a reproducible example of an operational blind spot. A production review should repeat the analysis across data suppliers, annotation batches and collection dates, using labels checked independently of the training export.

Research & data

- When Optimizing f-Divergence is Robust with Label Noise · ICLR 2021
- Optical recognition of handwritten digits · scikit-learn