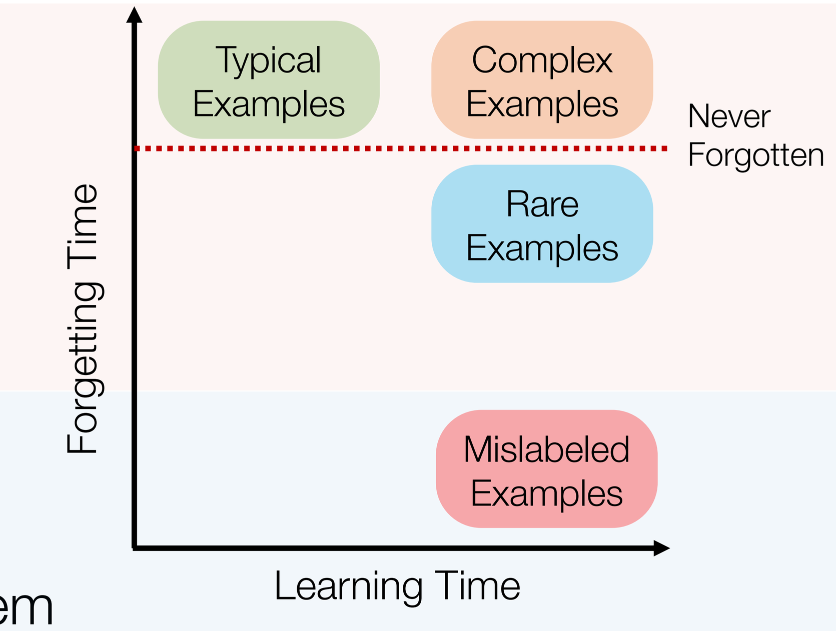
Carnegie Mellon University

BOSCH
Invented for life
Bosch Center for Artificial Intelligence

Characterizing datapoints via Second-Split Forgetting

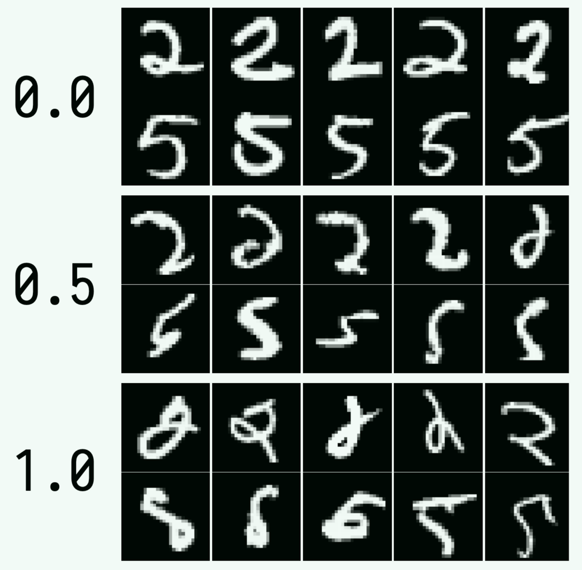
Pratyush Maini †, Saurabh Garg †, Zachary C. Lipton †, Zico Kolter † ✦

† Carnegie Mellon University | ✦ Bosch Center for AI



Motivation

- Recent works have shown that large fractions of benchmark datasets contain atypical examples (see Figure) [1].
- Memorization of atypical examples by deep nets improves generalization, but that of mislabeled examples hurts. How can we separate them?



[1] Feldman and Zhang: What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation

Main Result

- Mislabeled Examples: learnt late, forgotten fast
- Rare Examples: learnt late, forgotten late
- Complex Examples: learnt late, never forgotten
- Typical Examples: learnt early, never forgotten

Method

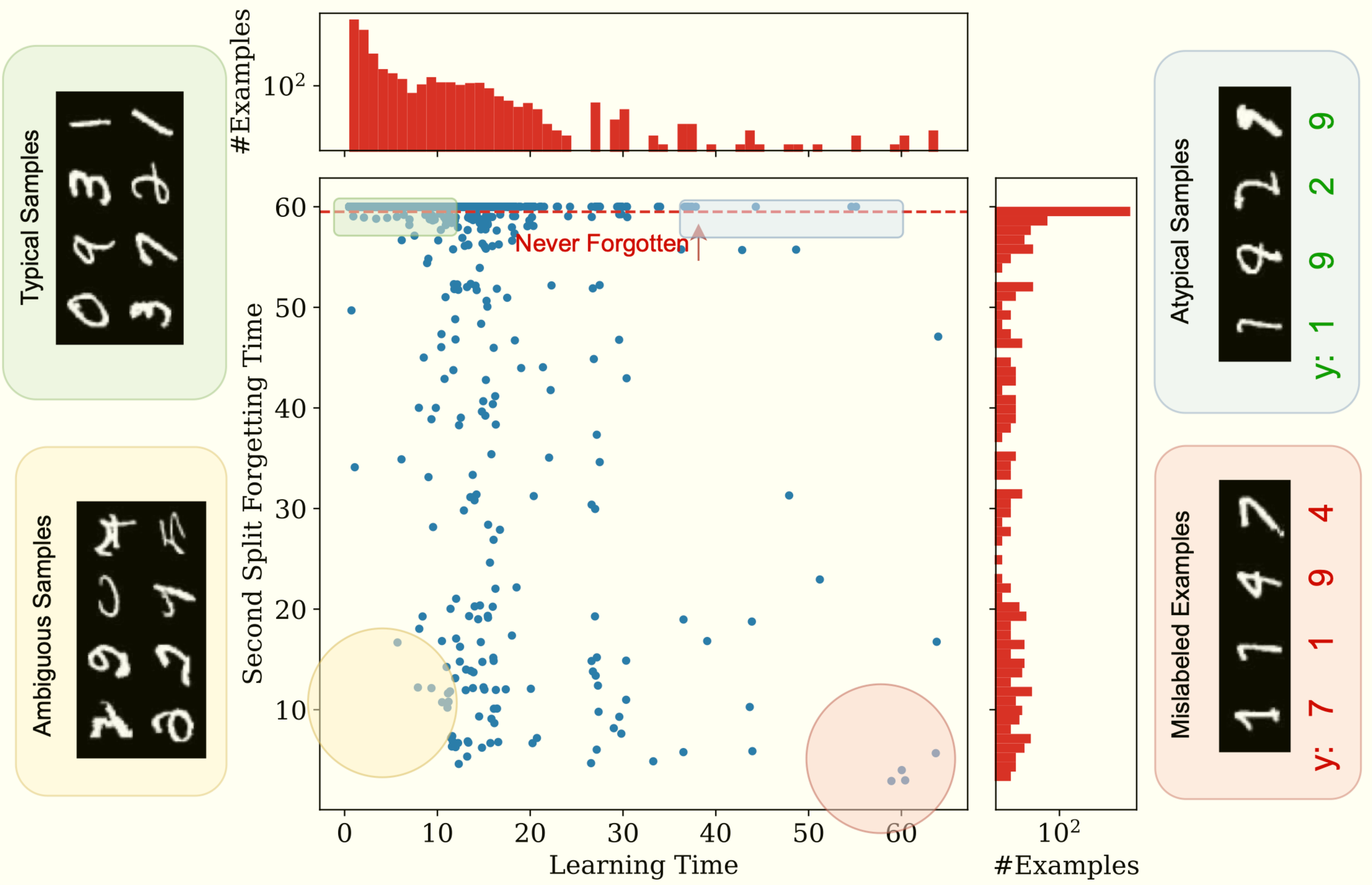
- Split your dataset into two halves
- Train on the 1st split till 100% accuracy
- Continue fine-tuning on the 2nd split
- Track accuracy of examples from 1st split as we continue training on 2nd

Learning Time: Earliest epoch during 1st split training after which an example is always predicted correctly.

Second-Split Forgetting Time (SSFT): Earliest epoch during 2nd split training after which an example from the 1st split is always predicted incorrectly.

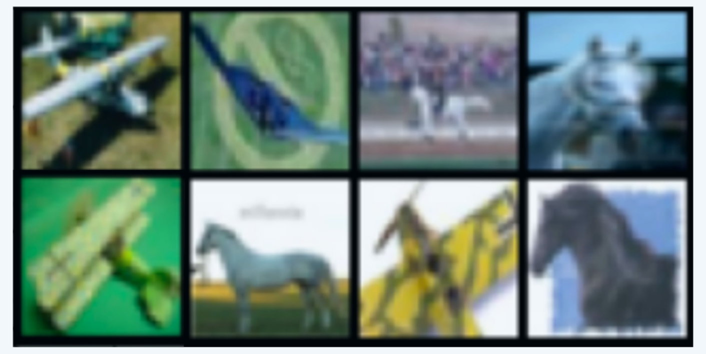
Learning-Forgetting Spectrum

- Examples that were learnt last and forgotten fastest were mislabeled (4th quadrant). The ones learnt early and never forgotten were characteristic typical (2nd quadrant) examples of the MNIST dataset.
- Examples in the 1st and 3rd quadrant are *seemingly* atypical, and ambiguous.
- Similar observations hold across different modalities and datasets.



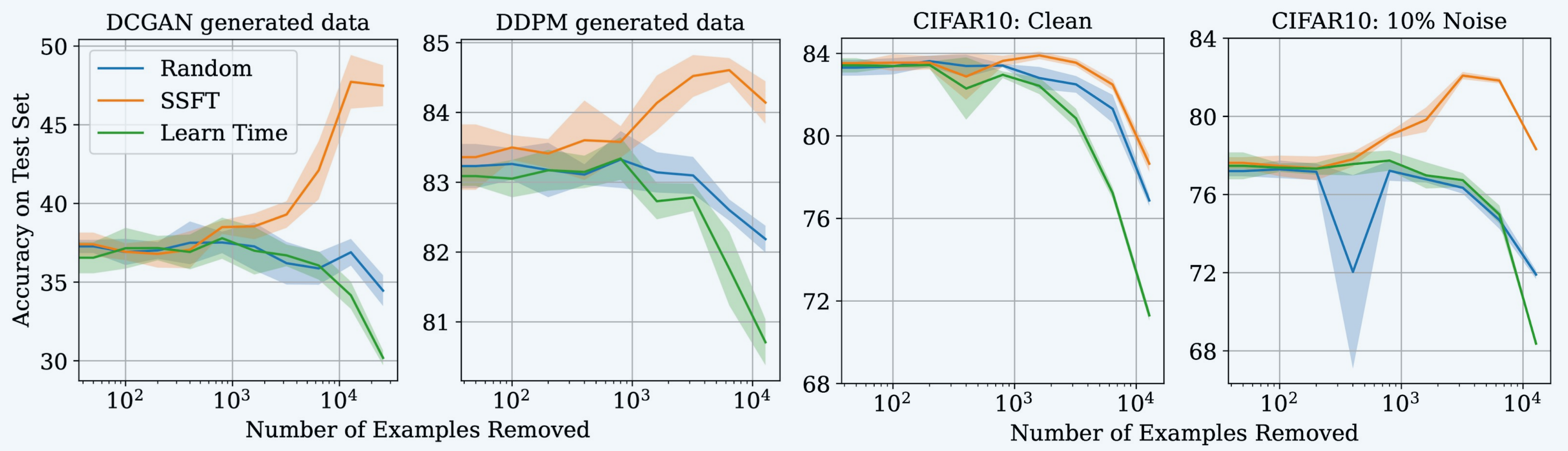
Spurious Correlations

- **Setup:** Create a 2-class classification problem from the horse and airplane class of CIFAR-10.
- **Observation:** The model quickly forgets planes with green backgrounds and horses on blue backgrounds. This suggests that the classifier may have used background as a (spurious) feature during first split training.



Applications

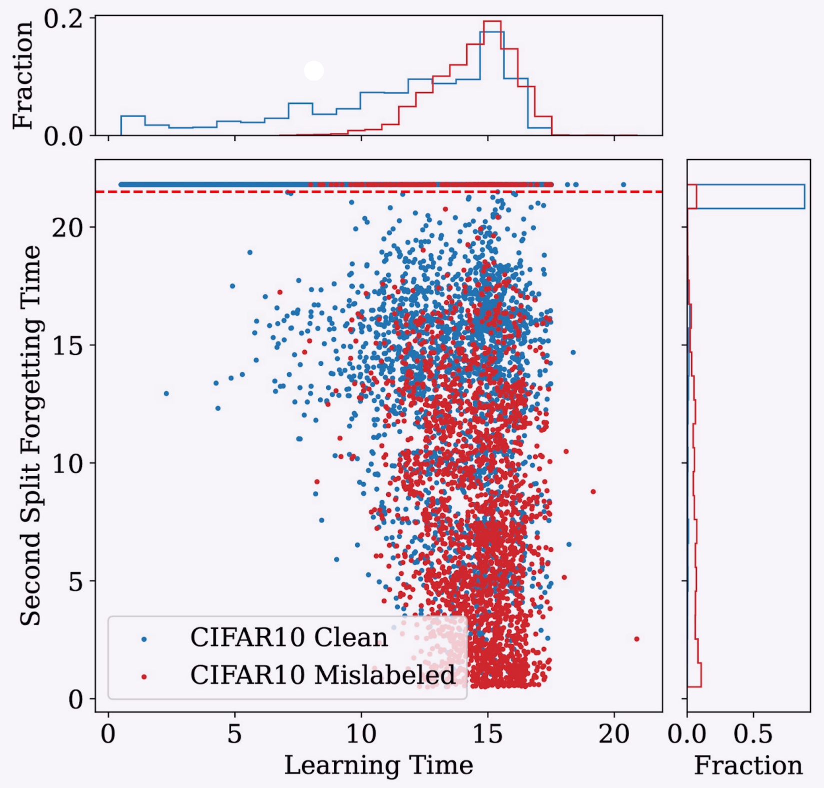
- Removing the earliest forgotten examples helps increase test accuracy. This suggests that SSFT finds pathological examples.
- Removing examples based on learning time hurts generalization more than when removed randomly. These are atypical examples.



Mislabeled Examples

Setup: Add 10% label noise to CIFAR-10 dataset (both first and second split).

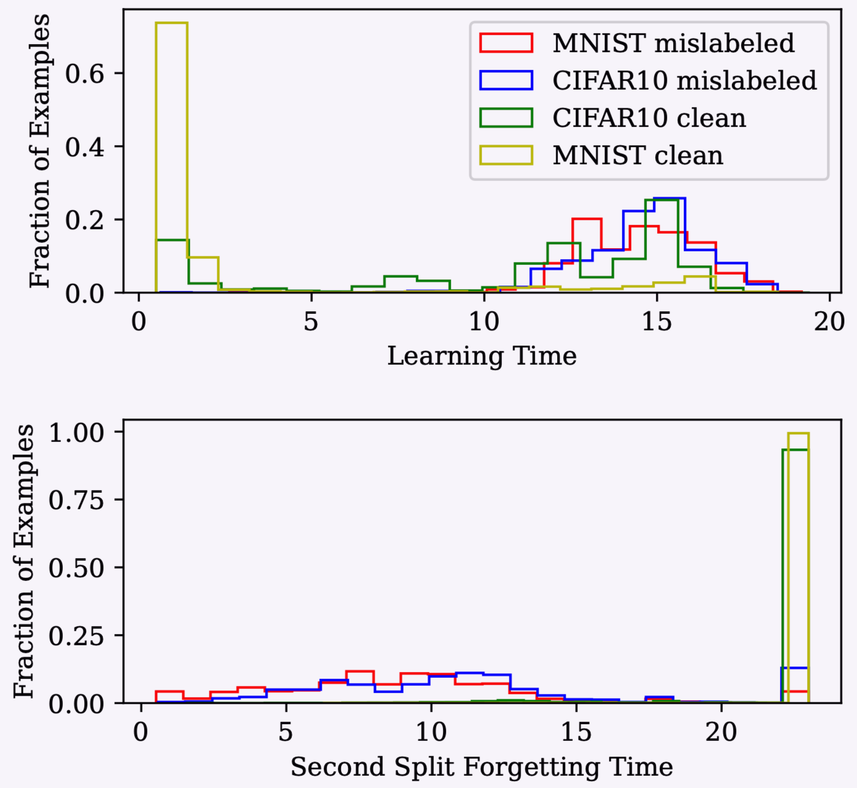
Observation: Mislabeled examples are learnt late, however, a large fraction of clean examples are also learnt late. On the contrary, the SSFT histogram visually shows a strong separation between mislabeled and clean examples.



Complex Examples

Setup: Make a dataset with the union of CIFAR10 (complex) and MNIST (simple) images by resizing MNIST digits.

Observation: Complex (CIFAR10) and mislabeled examples are both learnt late. However, only the mislabeled examples are forgotten quickly, complex are not. Suggests that learning time can confound complex and mislabeled examples.



Rare Examples

Setup: CIFAR100 has 20 super-classes with 5 subgroups each. Make a new dataset with exp. decaying sampling frequency for different subgroups within a super-class.

Observation: Examples from rare subgroups are learnt slowly. However, SSFT is nearly independent of the subgroup frequency. Suggests that learning time can confound rare and mislabeled examples.