Characterizing Data Points via Second-Split Forgetting

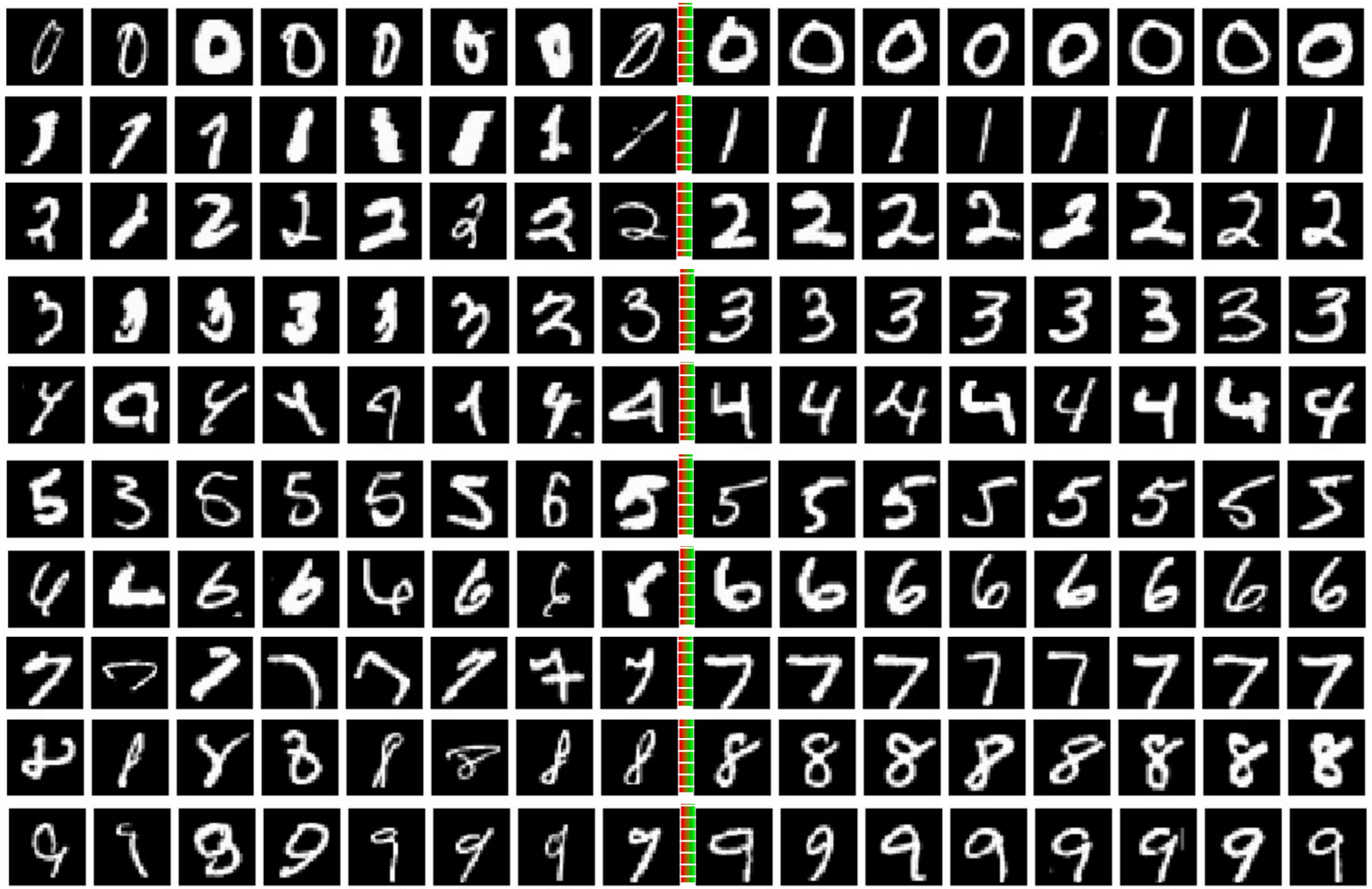



Pratyush Maini   Saurabh Garg   Zachary C. Lipton   Zico Kolter

Carnegie Mellon University

BOSCH
Invented for life
Bosch Center for Artificial Intelligence

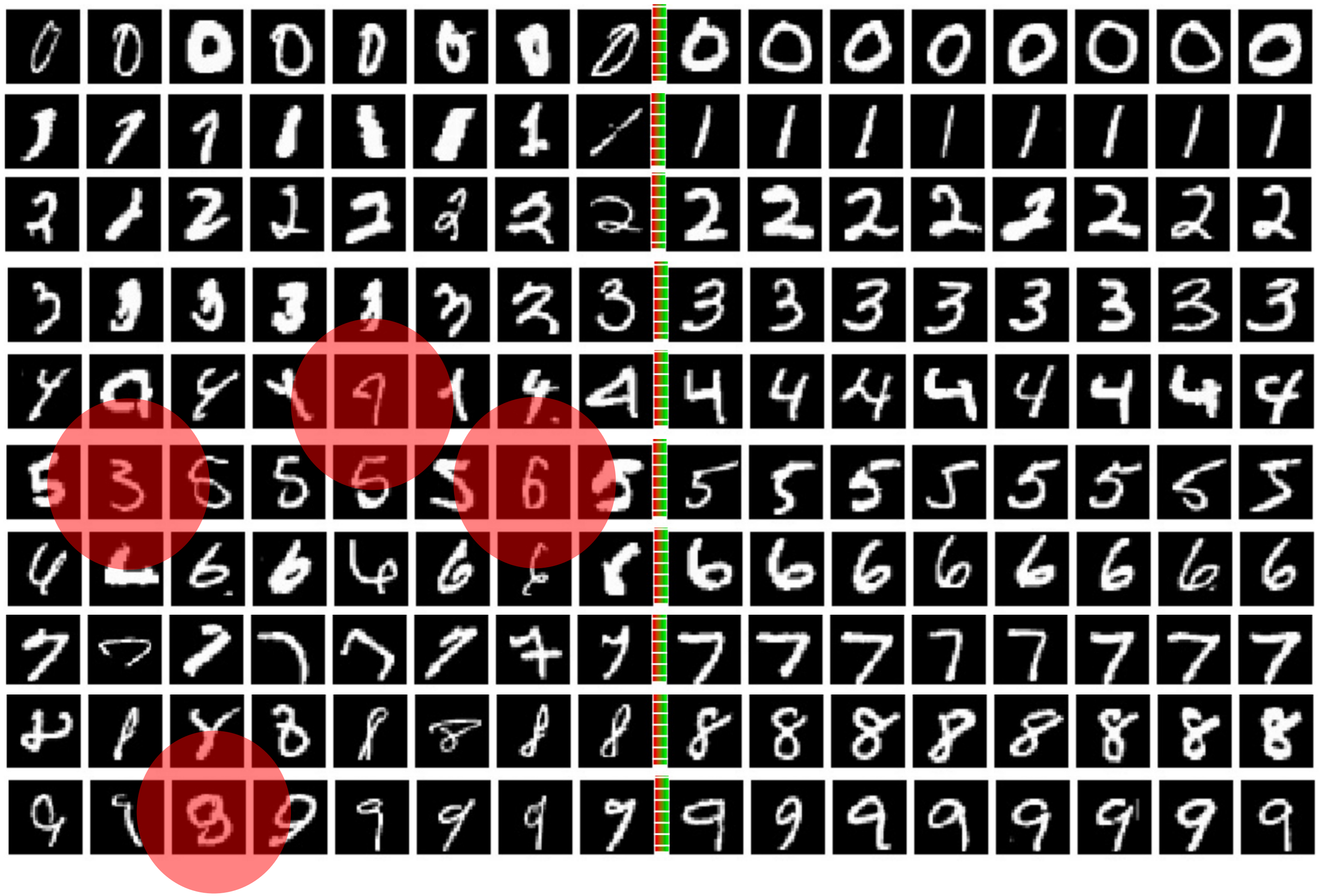
ML Datasets Have Both "Hard" and "Easy" Examples



[Carlini et. al. 2019; Distribution Density, Tails, and Outliers in Machine Learning: Metrics and Applications]

ML Datasets Have Both “Hard” and “Easy” Examples

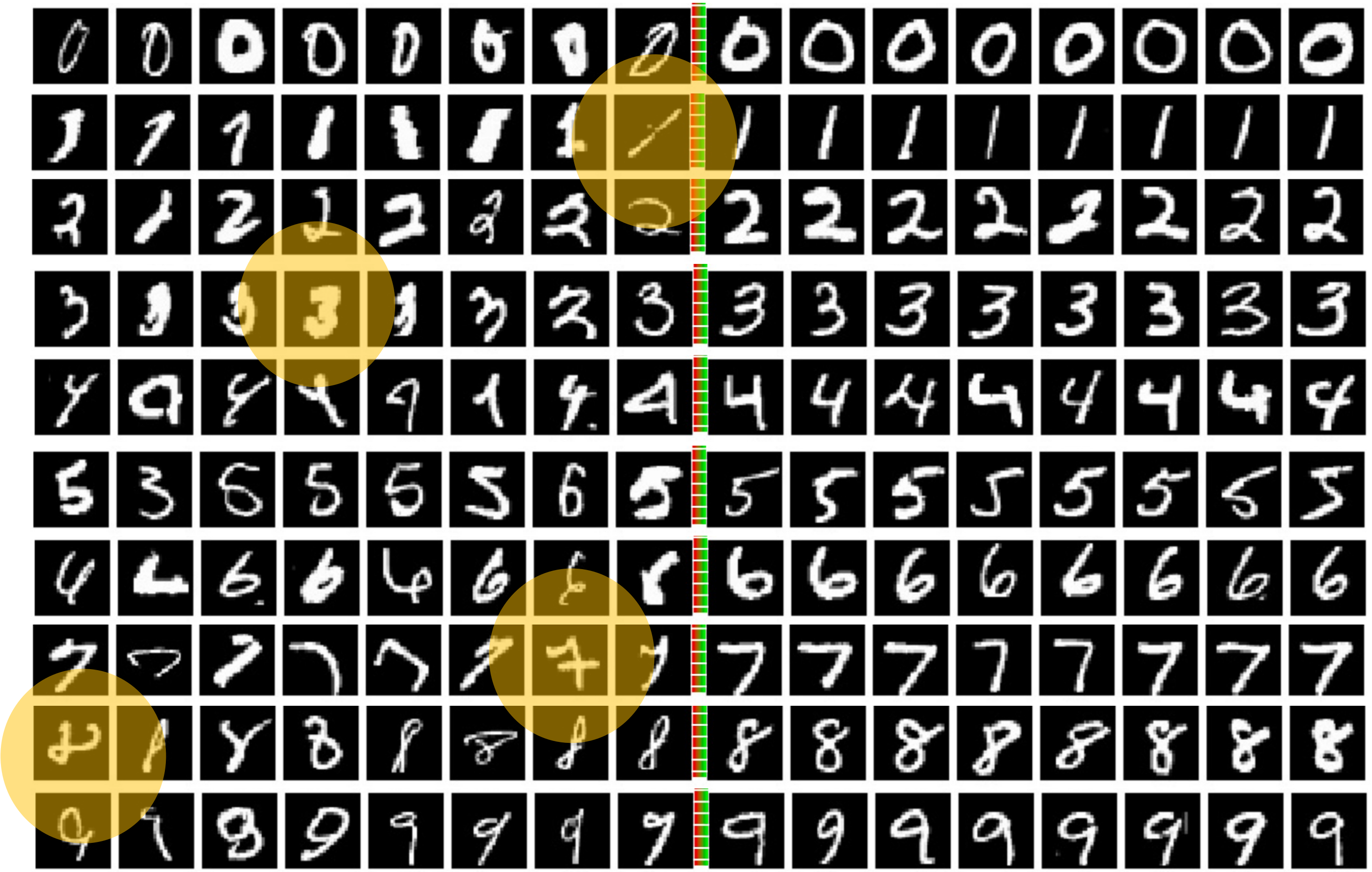
Some Examples are Hard because they are Mislabeled



[Carlini et. al. 2019; Distribution Density, Tails, and Outliers in Machine Learning: Metrics and Applications]

ML Datasets Have Both “Hard” and “Easy” Examples

Other Examples are Hard because of being Atypical



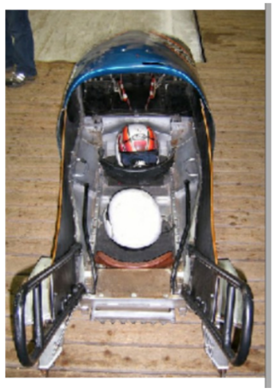
[Carlini et. al. 2019; Distribution Density, Tails, and Outliers in Machine Learning: Metrics and Applications]

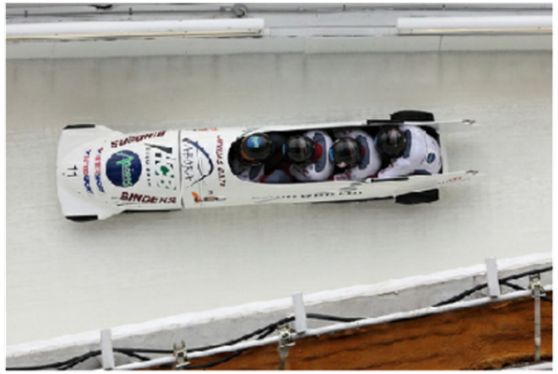
ML Datasets Have Both “Hard” and “Easy” Examples

ImageNet: **bobsled** class







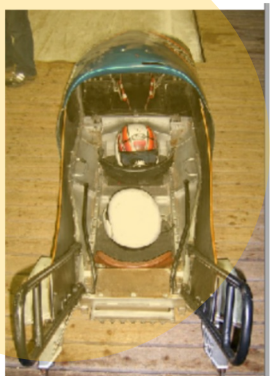

Typical Examples

[Feldman and Zhang 2020; What Neural Networks Memorize and Why?]

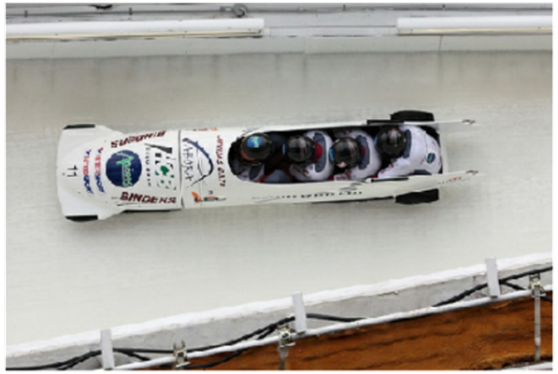
ML Datasets Have Both “Hard” and “Easy” Examples

ImageNet: **bobsled** class


Mislabeled Example

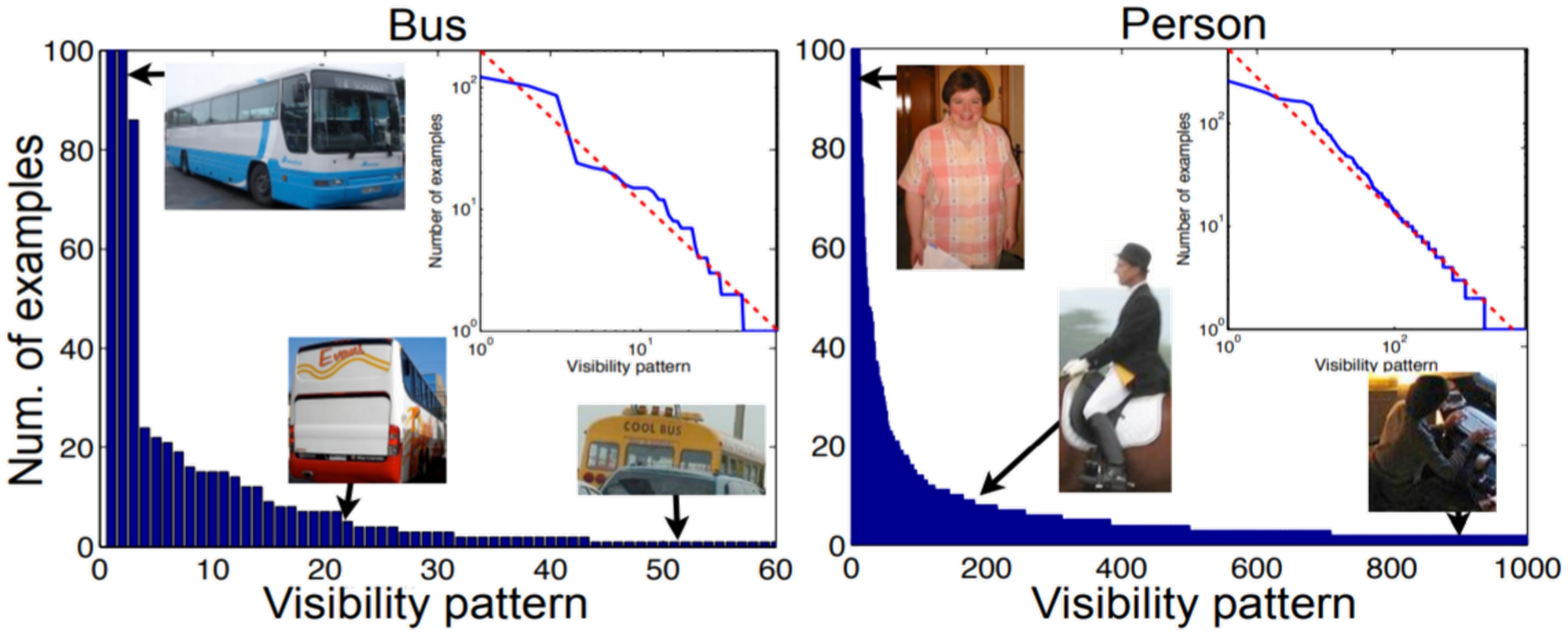
Rare Example


Typical Examples

[Feldman and Zhang 2020; What Neural Networks Memorize and Why?]

ML Datasets Have Long Tails of Atypical Examples

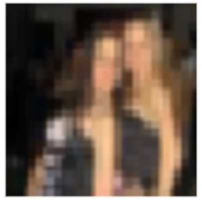


[Zhu et. al. 2014; Capturing Long-tail Distributions of Object Subcategories]

Memorizing Rare Examples Improves Generalization

CIFAR-10 truck class

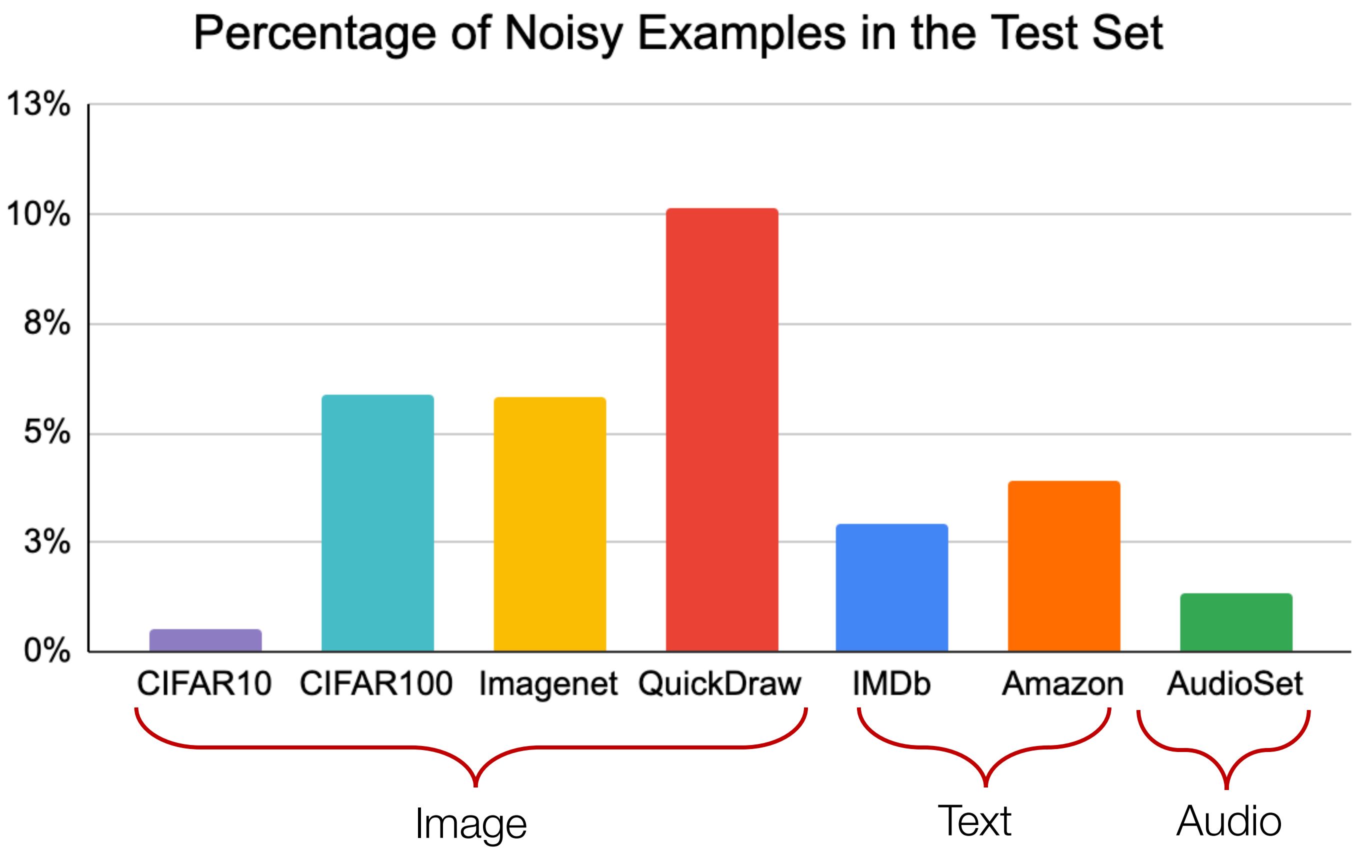
Training Set



Test Set

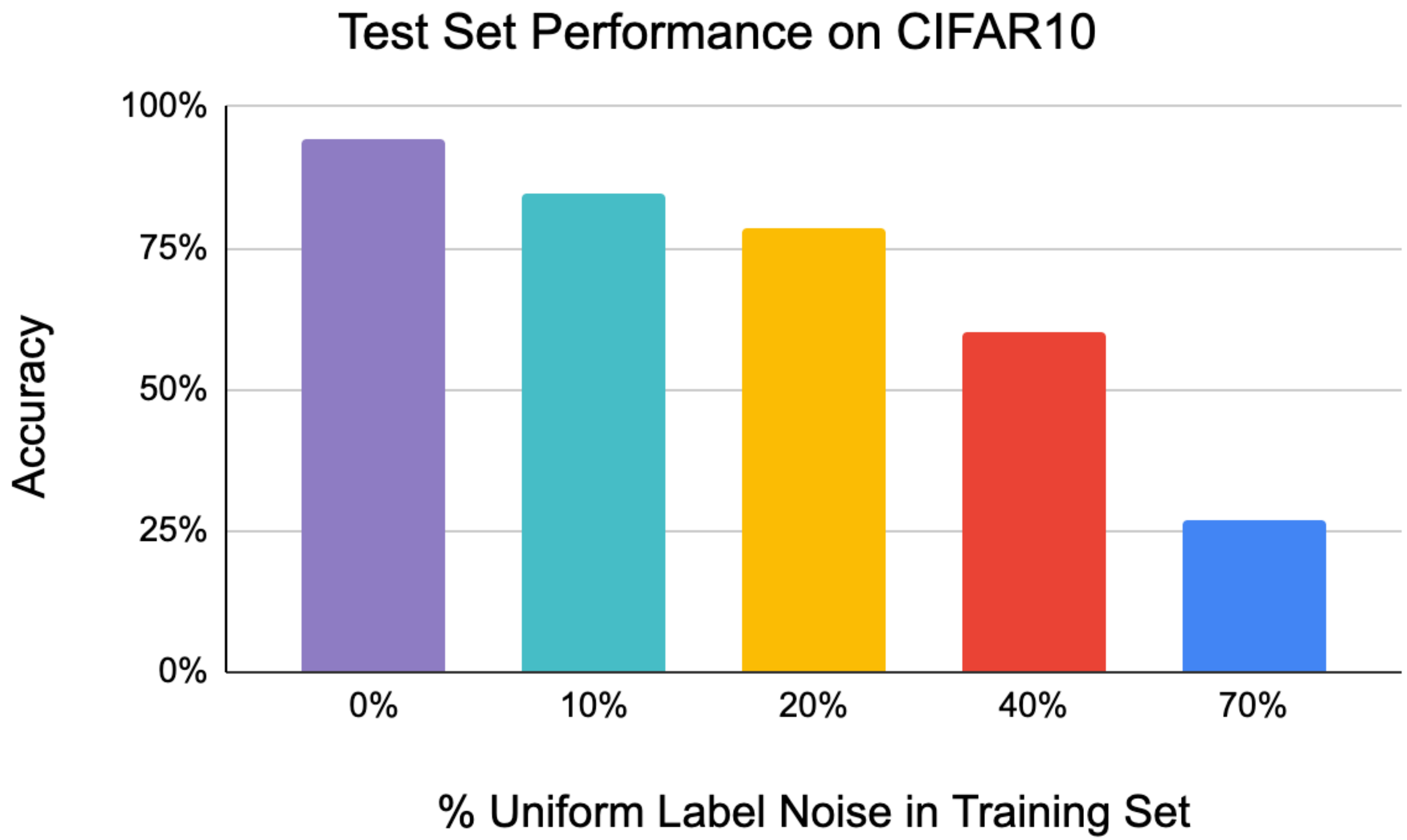
[Feldman and Zhang 2020; What Neural Networks Memorize and Why?]

ML Datasets Have Many Mislabeled Examples Too



[Northcutt et. al. 2021; Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks]

Memorizing Mislabeled Examples Hurts Generalization



Can we characterize examples based on different causes of hardness?

Learning and Forgetting Dynamics

1. Split a dataset into two halves
2. Train on the 1st split till convergence (100% train accuracy)

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

Learning and Forgetting Dynamics

1. Split a dataset into two halves
2. Train on the 1st split till convergence (100% train accuracy)
3. Now continue fine-tuning on the 2nd split (with these weights)
4. 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.

Learning and Forgetting Dynamics

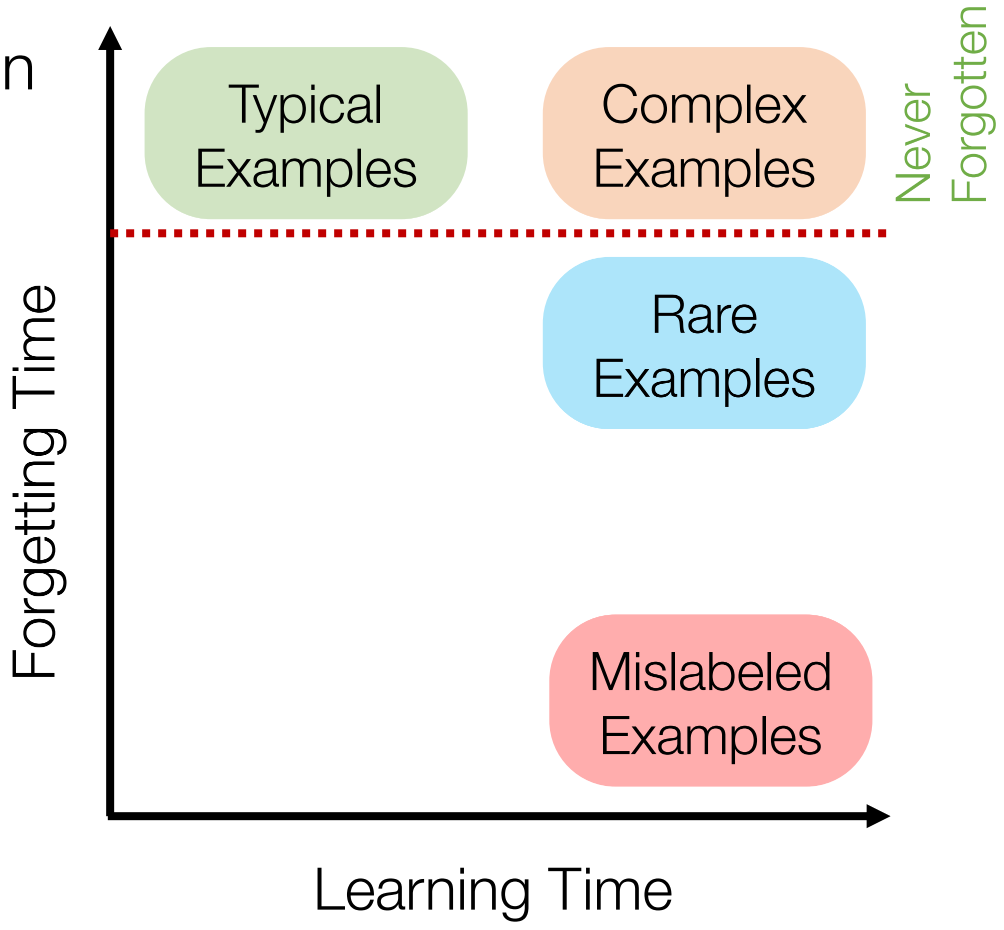
1. Split a dataset into two halves
2. Train on the 1st split till convergence (100% train accuracy)
3. Now continue fine-tuning on the 2nd split (with these weights)
4. Track accuracy of examples from 1st split as we continue training on 2nd

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 fine-tuning after which an example from the 1st split is always predicted incorrectly.

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



Mislabeled Examples: Learnt Late, Forgotten Early

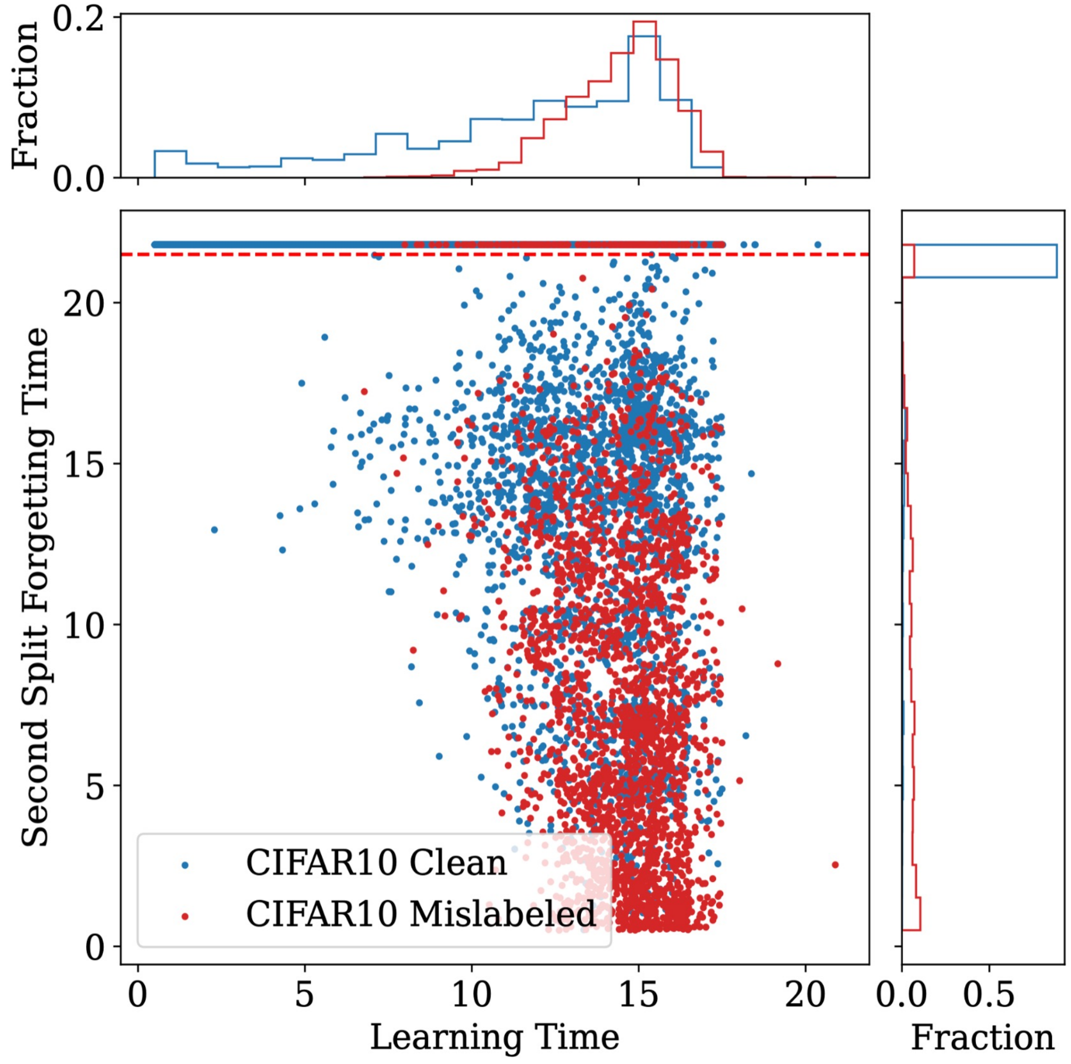
Setup: Randomly flip labels of 10% examples (both 1st and 2nd split)

Mislabeled Examples: Learnt Late, Forgotten Early

Setup: Randomly flip labels of 10% examples (both 1st and 2nd split)

Observation:

1. Mislabeled examples are learnt late
2. A large fraction of clean examples is also learnt late
3. The SSFT histogram visually shows a strong separation between mislabeled and clean examples



Complex Examples: Learnt late, Not Forgotten

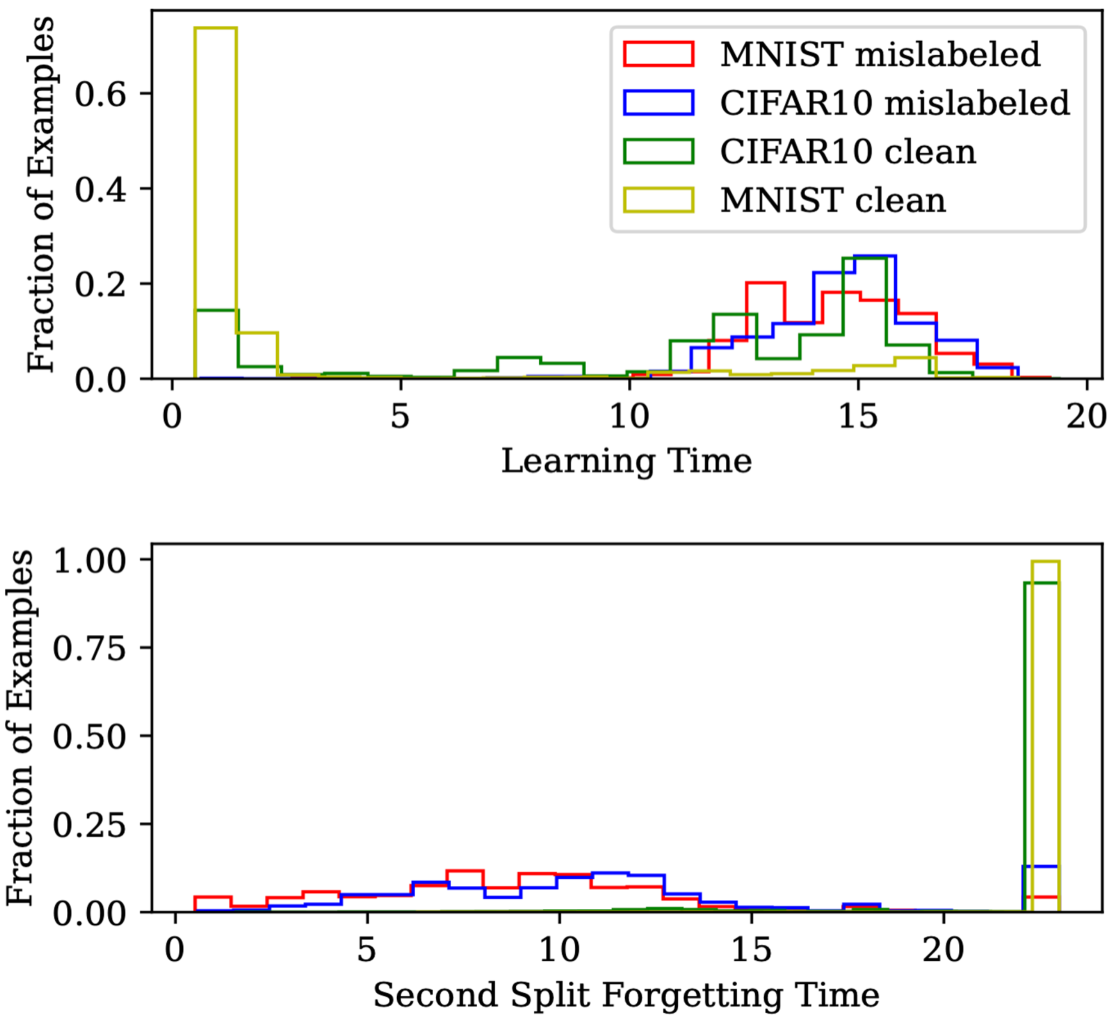
Setup: Make a dataset with the union of CIFAR10 (complex) and MNIST (simple) images.

Complex Examples: Learnt late, Not Forgotten

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

Observation:

1. Complex and Mislabeled Examples are both learnt late
2. SSFT for complex and simple examples is similar
3. Mislabeled Examples are forgotten quickly



Atypical Examples: Learnt Late, Forgotten Late

Desired dataset qualities:

1. Dataset where frequency is the only cause of example hardness
2. All classes must be *equally* complex, or have similar hardness

Atypical Examples: Learnt Late, Forgotten Late

Desired dataset qualities:

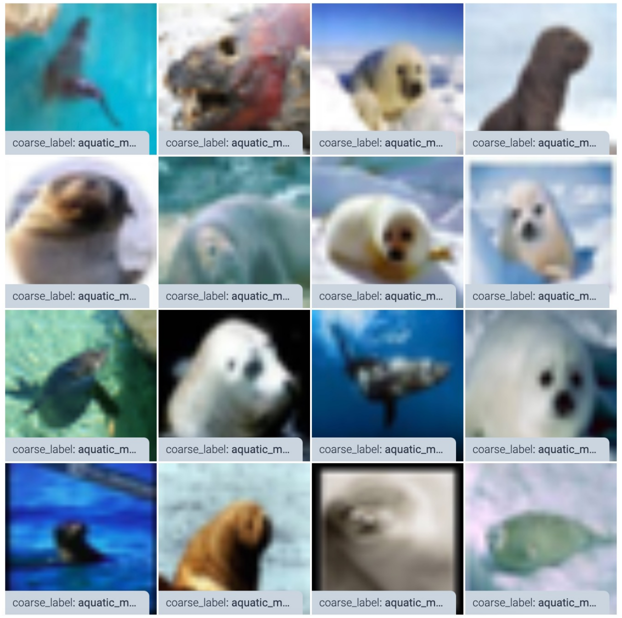
1. Dataset where frequency is the only cause of example hardness
2. All classes must be *equally* complex, or have similar hardness

How can we achieve this?

a. CIFAR100 has 20 super-classes. Each has 5 subgroups
b. Resample a dataset with {500, 250, 125, 64, 32} examples per subgroup in a superclass
c. Randomize all observations over multiple subgroup orderings

Constructing a Long-Tailed Dataset From CIFAR-100

Classes (20)



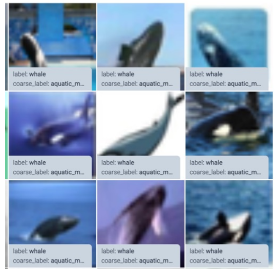
Aquatic Mammals

.

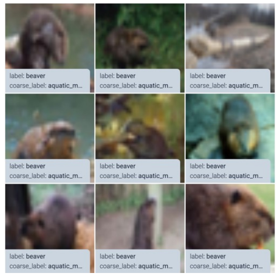
.

Flowers

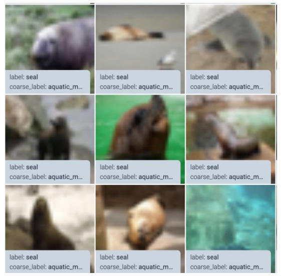
Biased sampling of Subpopulations



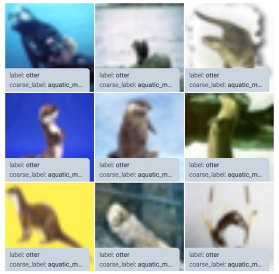
Whale



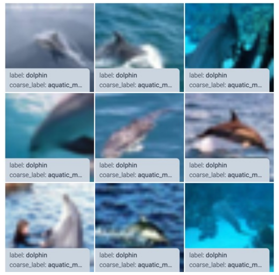
Beaver



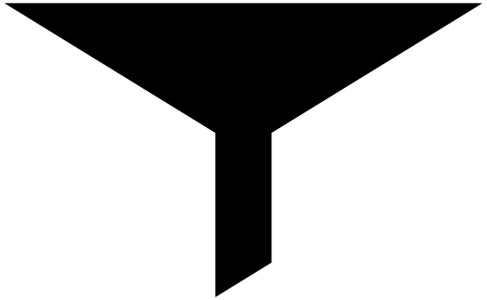
Seal



Otter

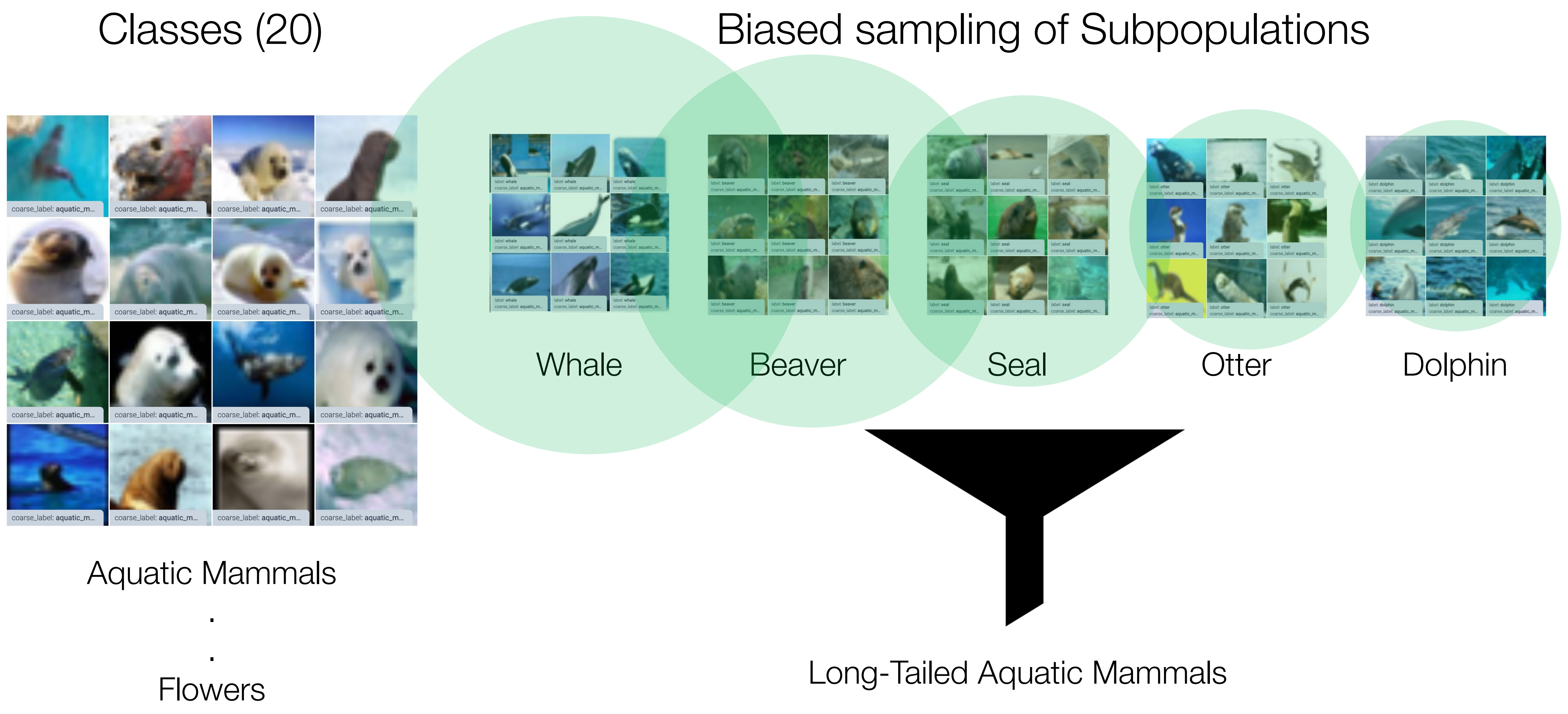


Dolphin



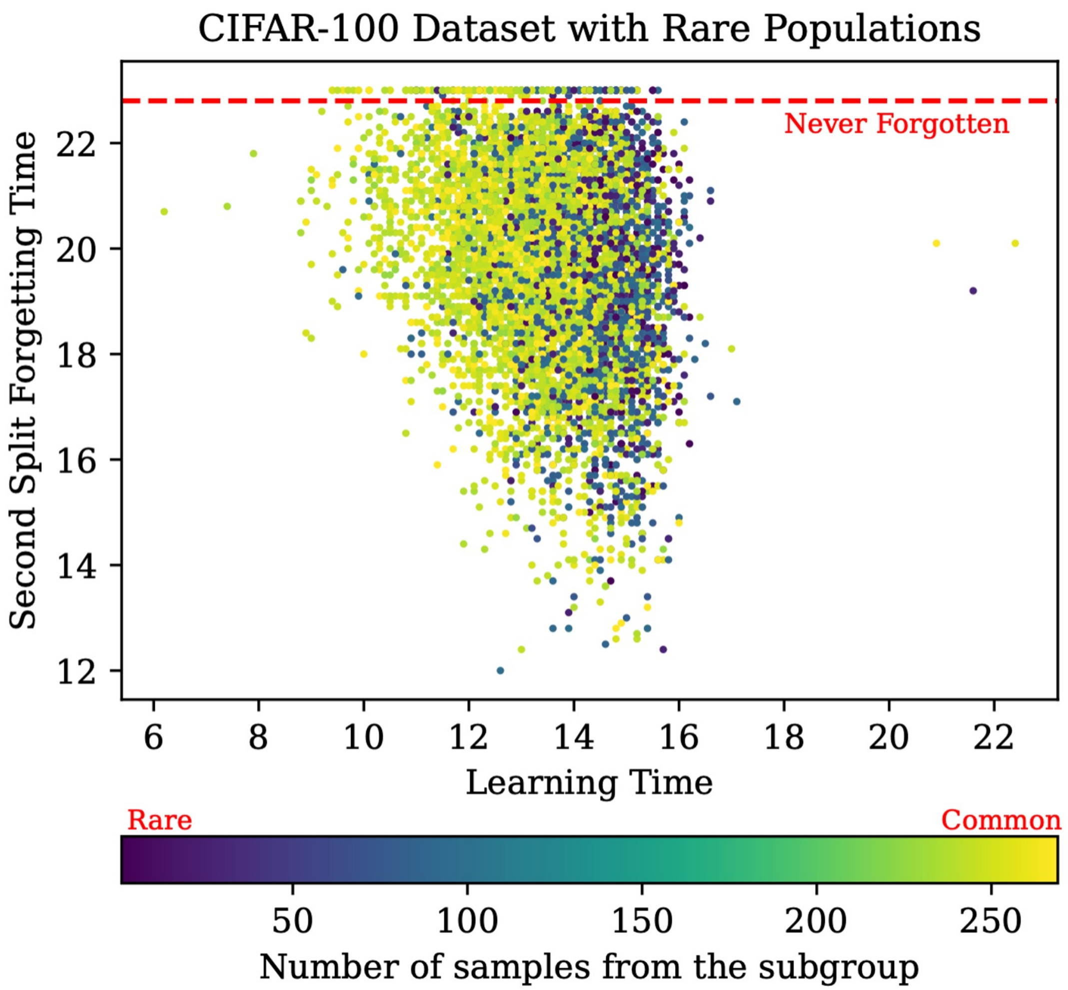
Long-Tailed Aquatic Mammals

Constructing a Long-Tailed Dataset From CIFAR-100

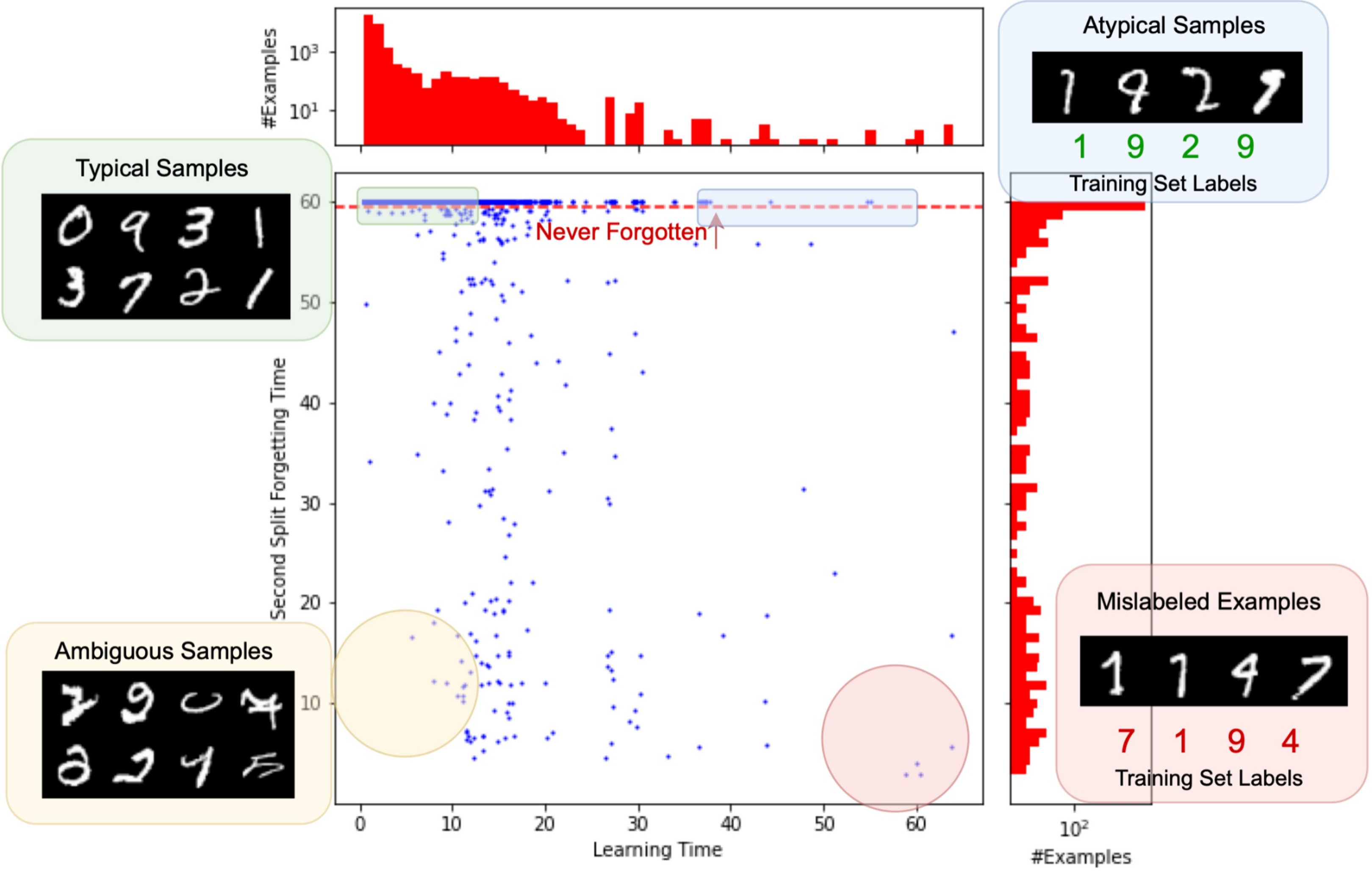


Atypical Examples: Learnt Late, Forgotten Late

1. Examples from rare subgroups are learnt slowly
2. SSFT is nearly independent of the subgroup frequency
3. Suggests that learning time can confound rare and mislabeled examples



Learning and Forgetting Dynamics: MNIST Dataset



Earliest Forgotten examples in SST-2 are mislabeled

The phenomenon of second-split forgetting is consistent across modalities.

- Examples with lowest SSFT when fine-tuning a BERT model on SST-2 are shown below.

Sentences in SST-2 dataset with smallest forgetting time	Label
The director explores all three sides of his story with a sensitivity and an inquisitiveness reminiscent of Truffaut	Neg
Beneath the film's obvious determination to shock at any cost lies considerable skill and determination , backed by sheer nerve	Neg
This is a fragmented film, once a good idea that was followed by the bad idea to turn it into a movie	Pos
The holiday message of the 37-minute Santa vs. the Snowman leaves a lot to be desired.	Pos
Epps has neither the charisma nor the natural affability that has made Tucker a star	Pos
The bottom line is the piece works brilliantly	Neg
Alternative medicine obviously has its merits ... but Ayurveda does the field no favors	Pos
What could have easily become a cold, calculated exercise in postmodern pastiche winds up a powerful and deeply moving example of melodramatic moviemaking	Neg
Lacks depth	Pos
Certain to be distasteful to children and adults alike , Eight Crazy Nights is a total misfire	Pos

Failure Modes of ML Models

Setup: Create a 2-class classification problem from CIFAR-10 (Horses & Planes)

Failure Modes of ML Models

Setup: Create a 2-class classification problem from CIFAR-10 (Horses & Planes)

Observation: Examples with lowest SSFT are

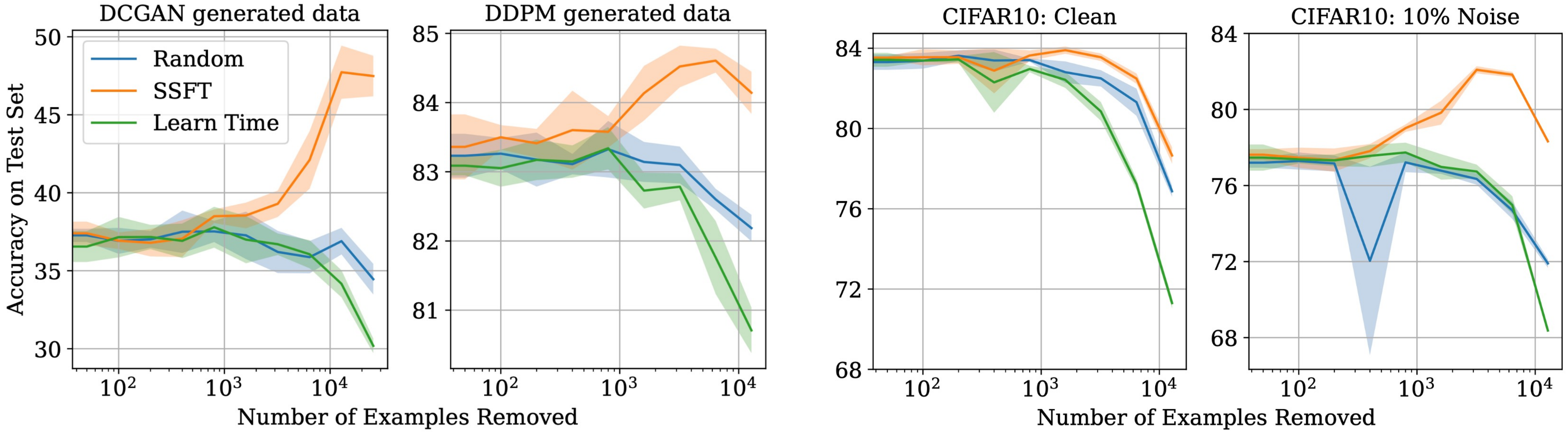
a. Horses with Blue Background
b. Planes with Green Background

Suggests that the classifier may have used background as a (spurious) feature during first split training.

Improving Dataset Utility

1. Removing the earliest forgotten examples helps increase test accuracy.
   - This suggests that SSFT finds pathological examples.

2. Removing the last learnt examples hurts test accuracy more than random removal.
   - This suggests that learning time finds atypical examples that help generalization.



Theoretical Results on a Linear Model

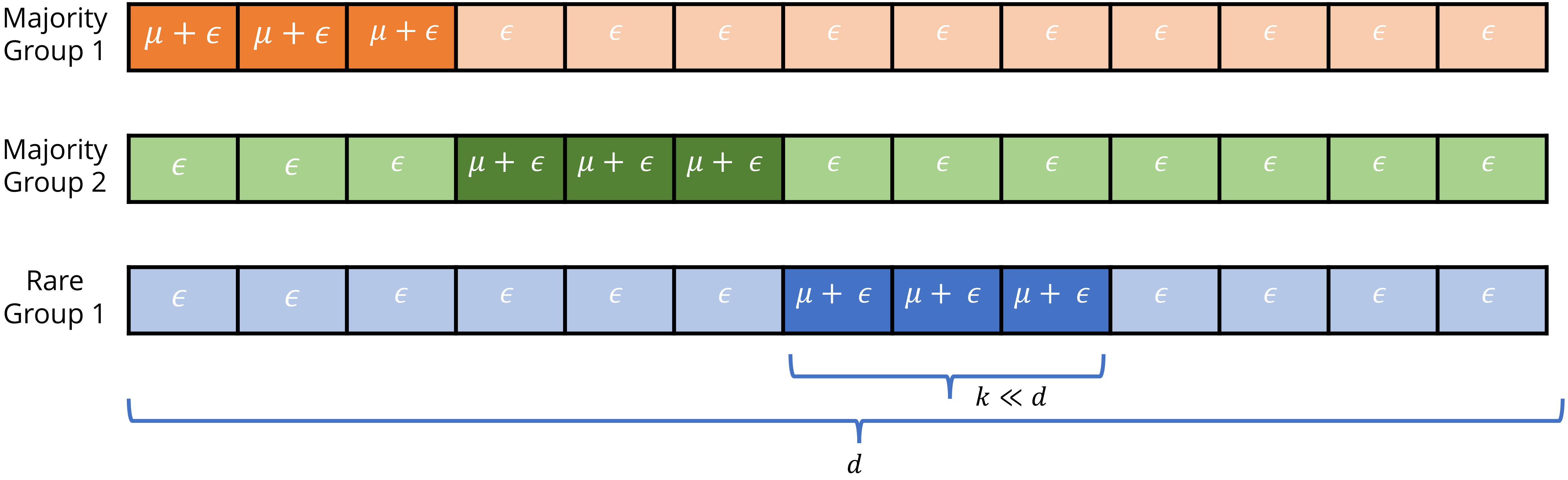
Input: $\mathbf{x} = (\boldsymbol{\mu} + \boldsymbol{\epsilon}) \in \boldsymbol{R}^d$

Label: $\mathbf{y} = \pm 1$

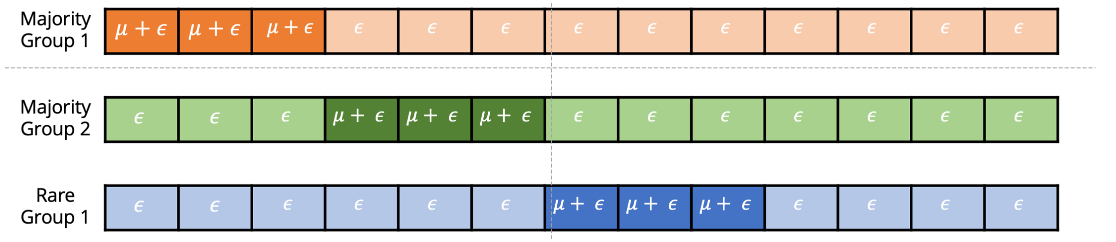
Model: $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x};\ \mathrm{w} \in \boldsymbol{R}^d$

Noise: $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_d)$

Signal: $\boldsymbol{\mu_1} = \begin{cases} \mu; & j \in \{1 \dots k\} \\ 0; & o.w. \end{cases}$



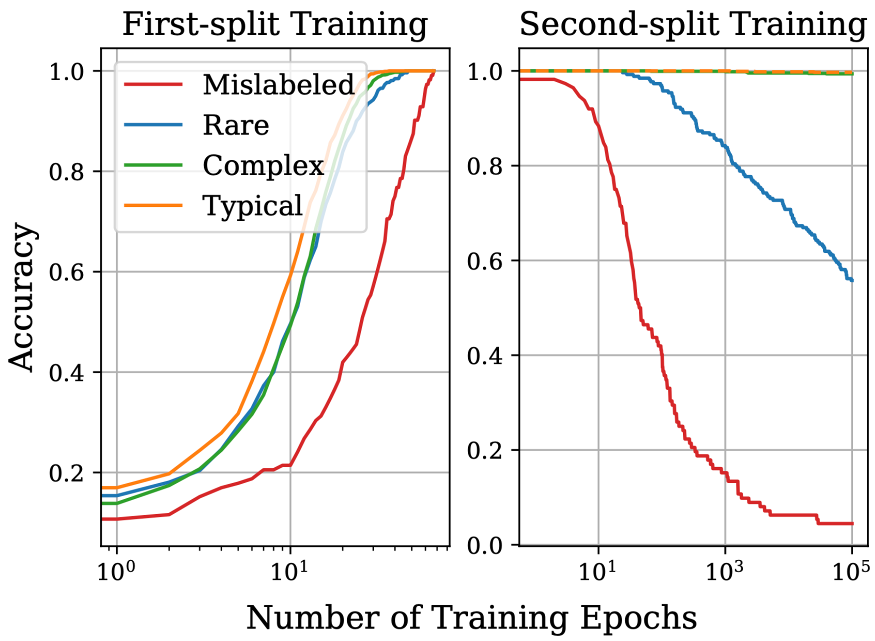
Asymptotic Forgetting



Theorem 1 (Asymptotic Forgetting (informal)). *For sufficiently small learning rate, given datasets $\mathcal{S}_A, \mathcal{S}_B \sim \mathcal{D}^n$. After training for $T' \to \infty$ epochs, the following hold with high probability:*

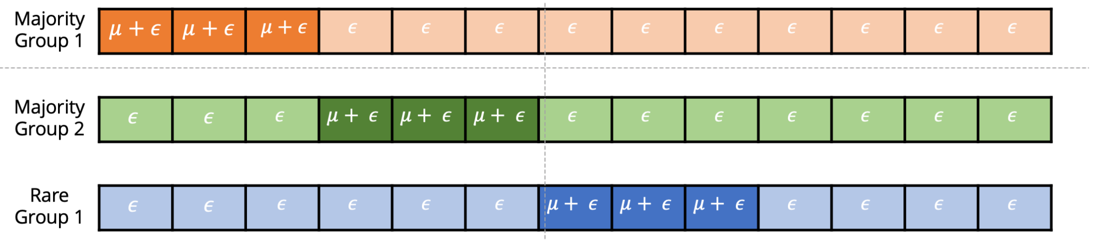
1. *Mislabeled and Rare examples from $\mathcal{S}_A$ are forgotten.*
2. *Complex examples from $\mathcal{S}_A$ are not forgotten.*

1. Dataset is separable with high probability.
2. The classifier will converge to min-norm solution for *any* bounded initialization [Soudry et. al.].
3. Asymptotic Solution should be independent of first-split training.
4. Use Generalization bound from Chatterji and Long.



Being forgotten for rare examples implies random guessing, whereas it implies incorrect prediction for mislabeled examples.

Intermediate Time Forgetting



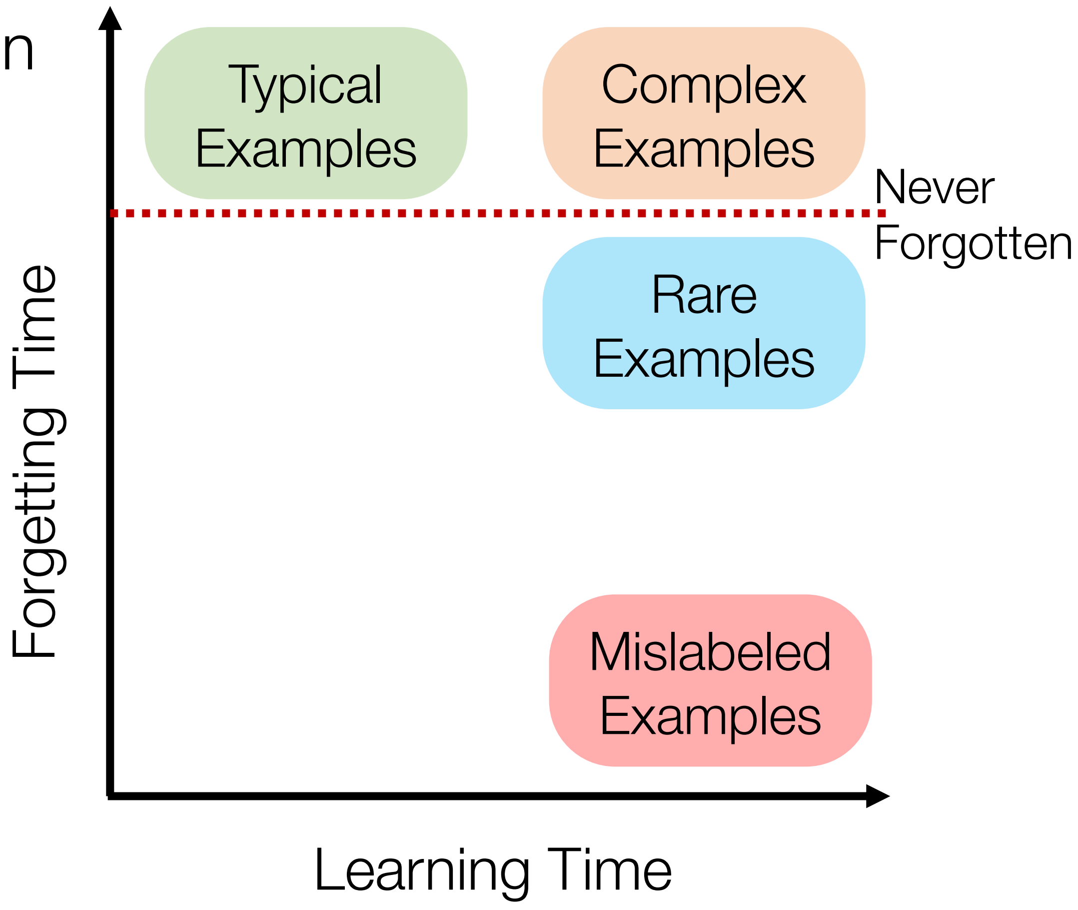
Theorem 2 (Intermediate-Time Forgetting (informal)). *For sufficiently small learning rate, given two datasets $\mathcal{S}_A, \mathcal{S}_B \sim \mathcal{D}^n$. For a model initialized with weights, $\mathbf{w}_B(0) = \mathbf{w}_A(T)$ and trained for $T' = f(T)$ epochs, the following hold with high probability:*

1. *Mislabeled examples from $\mathcal{S}_A$ are no longer incorrectly predicted.*
2. *Rare examples from $\mathcal{S}_A$ are not forgotten.*

1. *Representer Theorem:* Change in $\mathbf{w}$ is a weighted sum of examples from the second split $\sum \beta_i \mathbf{x_i}$.
2. Change in prediction is dot product of examples from first split with $\sum \beta_i \mathbf{x_i}$.
3. This dot product has zero mean (only noise) for rare examples. (Orthogonal signal directions)
4. But mislabeled examples have a negative mean dot product since they are from majority group.
5. Rare example prediction changes much slower than mislabeled examples.

Conclusions

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



Applications

- Finding Mislabeled Examples
- Identifying Spurious Attributes
- Improving Dataset Utility