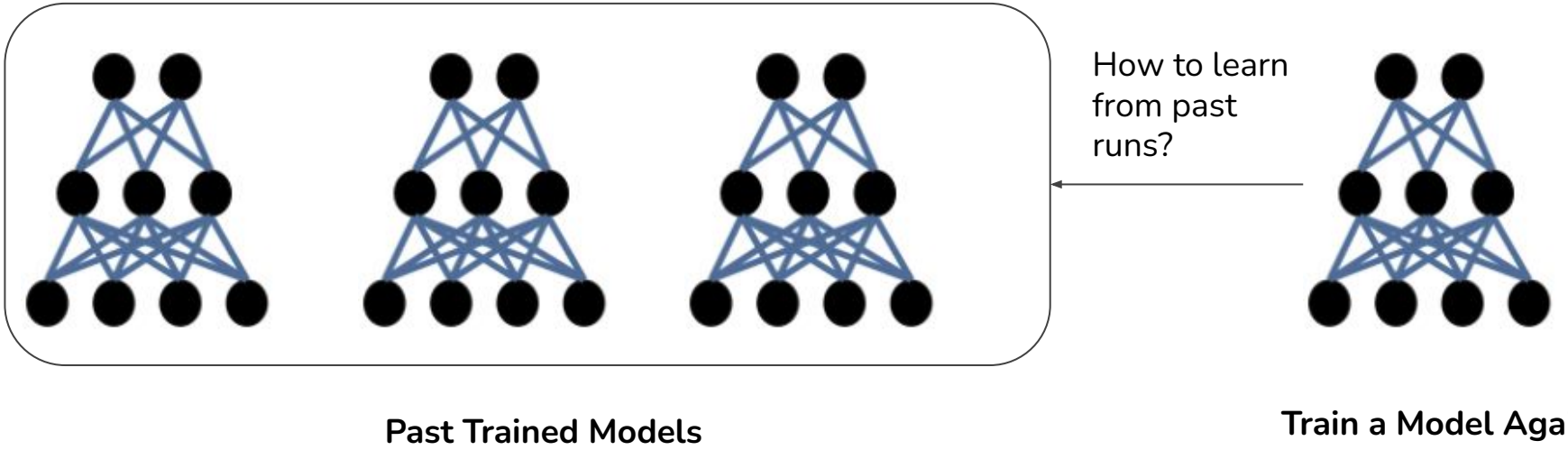
AC 297r x Mosaic ML

Xingyu Liu, Lu Yu, Alex Leonardi, Chris Gilmer-Hill

IACS
Harvard John A. Paulson School of Engineering and Applied Sciences
IACS Institute for Applied Computational Science

Problem Statement

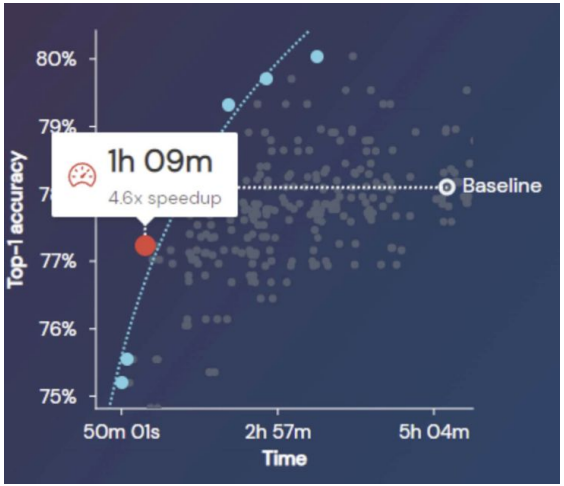
How can we reuse the computation that was invested in training our initial models to make training future models better?



Problem Statement

Project Goals:

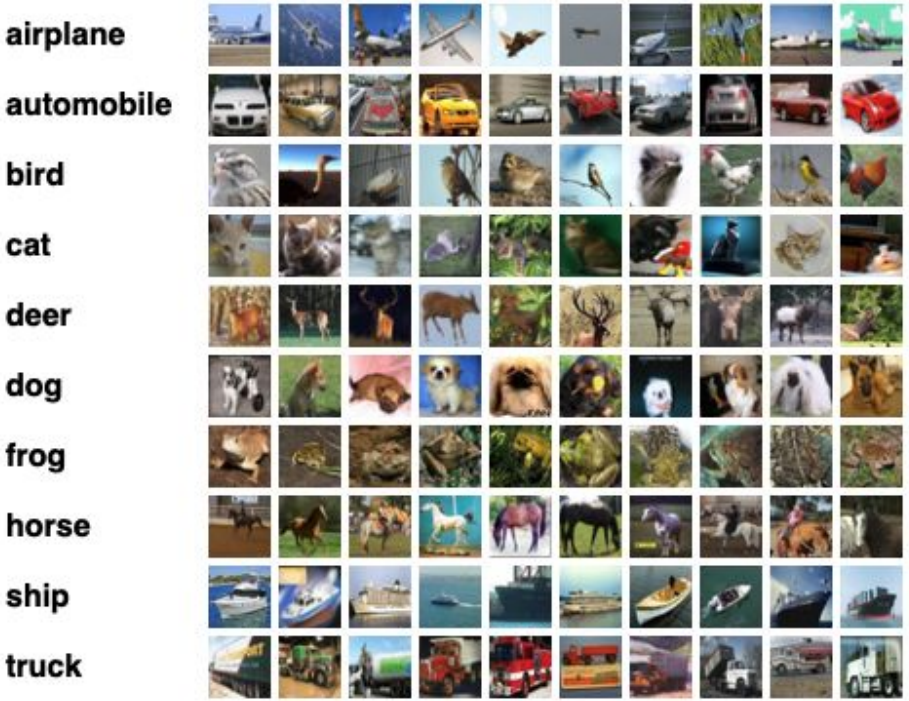
- Establish a benchmark and metrics for testing out strategies to reuse information
- Develop and evaluate methods for reusing information to improve metrics on the benchmark
- Put forth recommended best practices for practitioners based on this research



Potential Approaches

- Methods
  - Knowledge Distillation
  - History-Informed Difficulty Modulation
  - Past-Adversarial MixUp
- Applications
  - Image Data
  - Election Data
  - Language Data
  - Others?

Resources Available



CIFAR-10

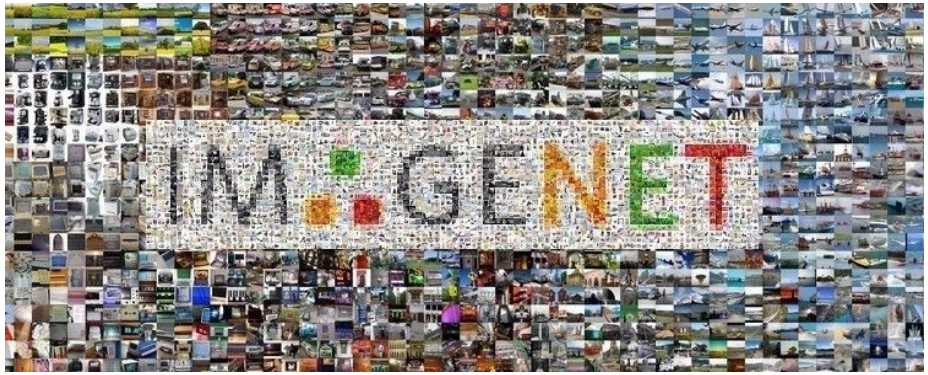
Start with CIFAR Dataset:

- **Simple**
- **Commonly used in Computer Vision Research**

Resources Available

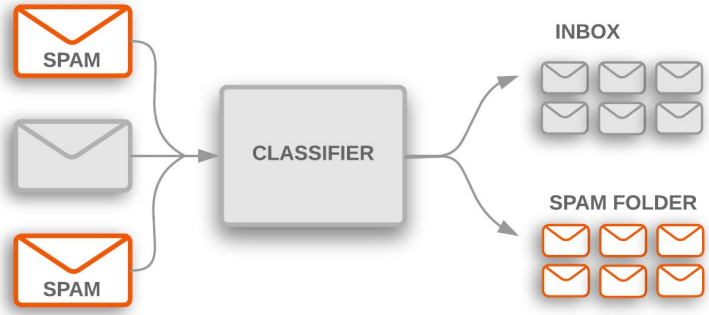
Move to a Larger Dataset (ImageNet):

- **Scale is one place where methods tend to fall down**

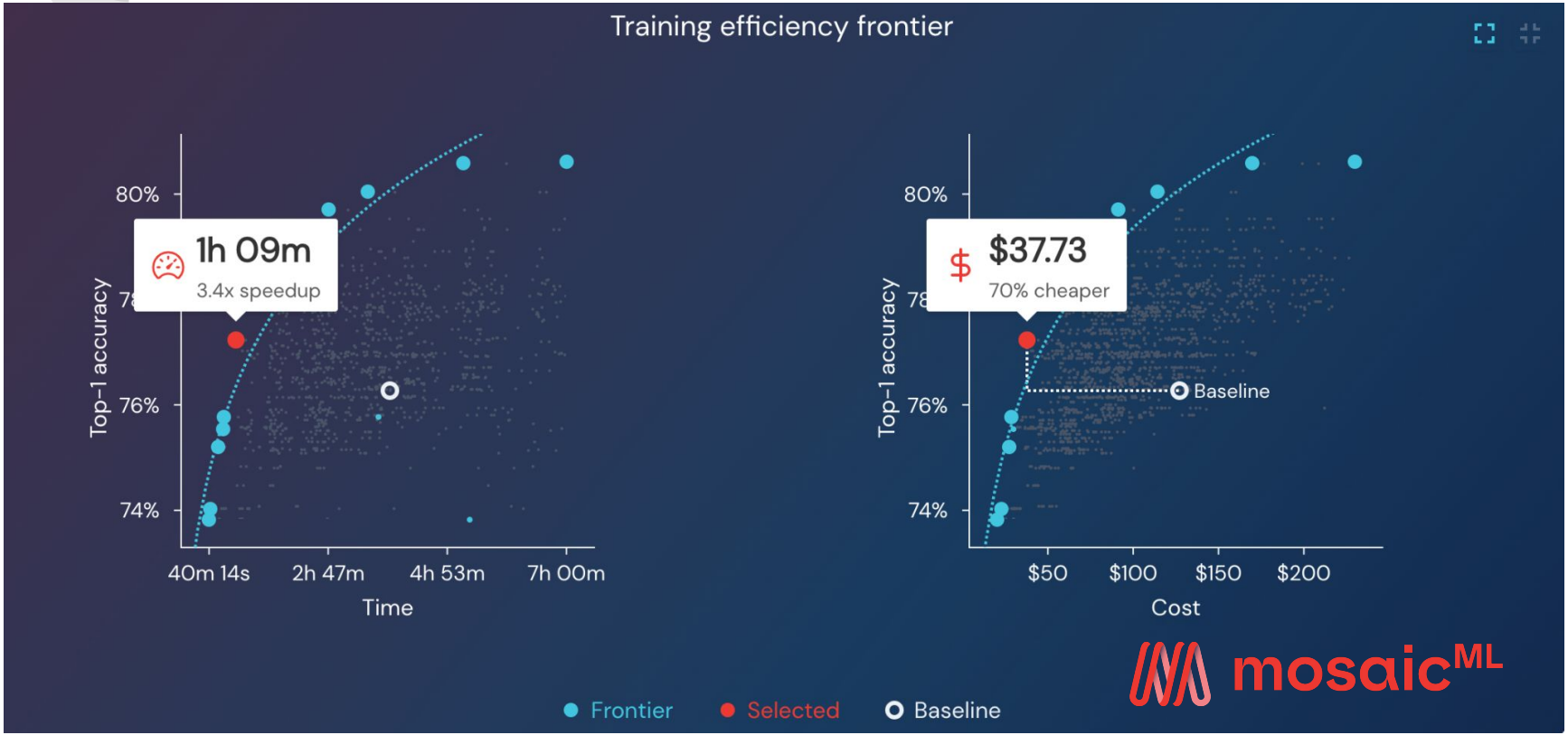


Finally Move to Basic NLP tasks:

- **E.g. Text Classification**
- **Useful in practice**



Resources Available



High-Level Project Stages

Milestone 1 (by Mar 4)

- Create our baseline model
- Set up evaluation framework, how to measure and determine success

Milestone 2 (by Apr 4)

- Try different techniques to improve metrics on the benchmark
- Refine model with hyperparameter search

High-Level Project Stages (continued)

Milestone 3 (by Apr 22)

- Keep tuning and refining models
- Explore models on more complex tasks

Final Deliverables (by May 4)

- Final report / blog post
- Documentation, GitHub repo
- Capstone Project Showcase

Project Timeline