Growing or Compressing Datasets

Cody Coleman (CEO and Co-Founder of Coactive AI)

MIT IAP 2023 Introduction to Data-Centric AI

Today’s Lecture

Why care about labels? Data-Centric vs Model-Centric AI (1/17)

Who labels and how? Dataset Creation and Curation (1/19)

What to label? Growing or Compressing Datasets (Today)

- Active learning for growing datasets
- Core-set selection for compressing datasets

Today's Lecture

Why care about labels? Data-Centric vs Model-Centric AI (1/17)

Who labels and how? Dataset Creation and Curation (1/19)

What to label? Growing or Compressing Datasets (Today)

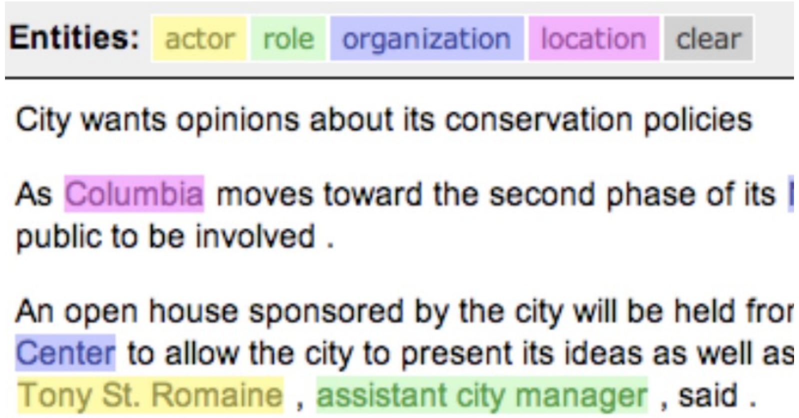
- **Active learning for growing datasets**
- Core-set selection for compressing datasets

Why is selecting what to label important?

Speech recognition
Annotation at the word level can take ten times longer than the actual audio and annotating phonemes can take 400 times as long (e.g., nearly seven hours).

Zhu. "Semi-Supervised Learning with Graphs." 2005
Settles et al. "Active Learning with Real Annotation Costs." 2008
Settles. "Active Learning Literature Survey." 2010

Why is selecting what to label important?



City wants opinions about its conservation policies

As Columbia moves toward the second phase of its
public to be involved .

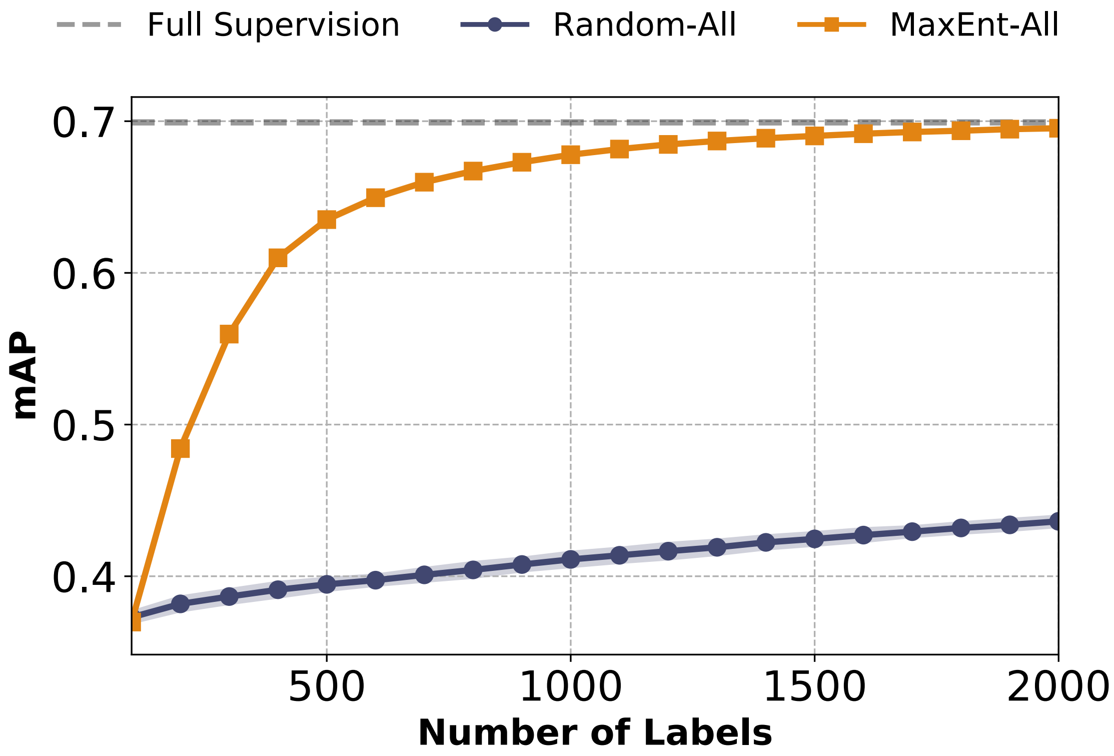
An open house sponsored by the city will be held fror
Center to allow the city to present its ideas as well as
Tony St. Romaine , assistant city manager , said .

Speech recognition
Annotation at the word level can take ten times longer than the actual audio and annotating phonemes can take 400 times as long (e.g., nearly seven hours).

Information extraction
Locating entities and relations can take a half-hour or more for even simple newswire stories.

Zhu. “Semi-Supervised Learning with Graphs.” 2005
Settles et al. “Active Learning with Real Annotation Costs.” 2008
Settles. “Active Learning Literature Survey.” 2010

Why is selecting what to label important?



Passive vs. Active (Machine) Learning

Using the model to help us actively select examples can dramatically reduce the number of examples we need to label compared to passively selecting at random.

Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

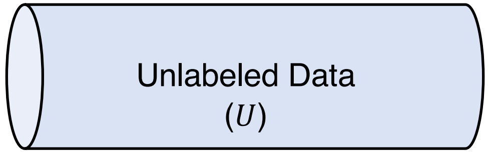
What is active learning?

What is active learning?

The goal is to select the best examples to improve the model.

What is active learning?

Traditional Approach

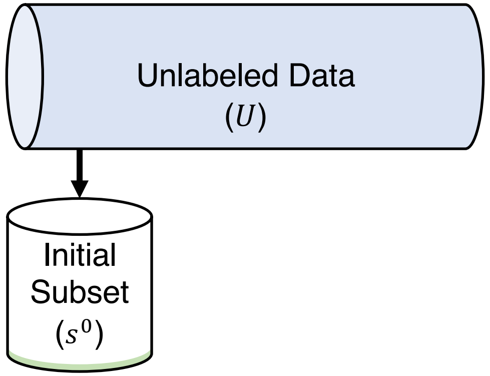


Unlabeled Data
(U)

The goal is to select the best examples to improve the model.

What is active learning?

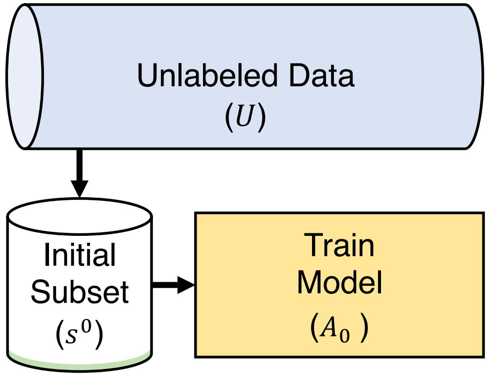
Traditional Approach



The goal is to select the best examples to improve the model.

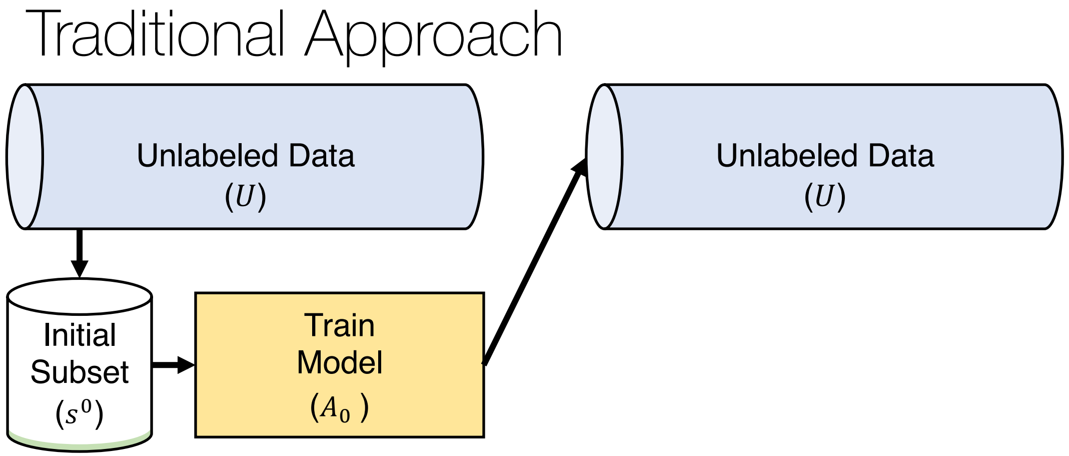
What is active learning?

Traditional Approach



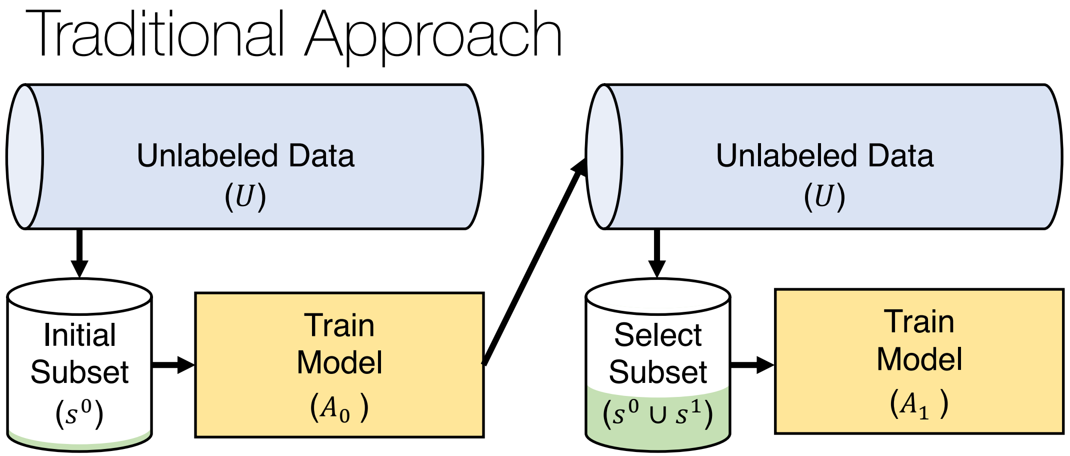
The goal is to select the best examples to improve the model.

What is active learning?



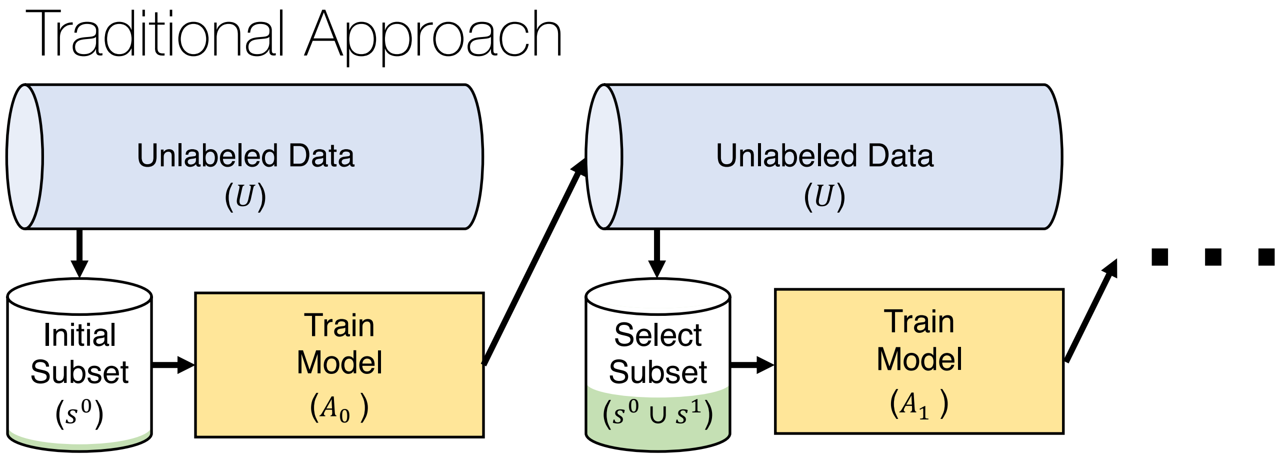
The goal is to select the best examples to improve the model.

What is active learning?



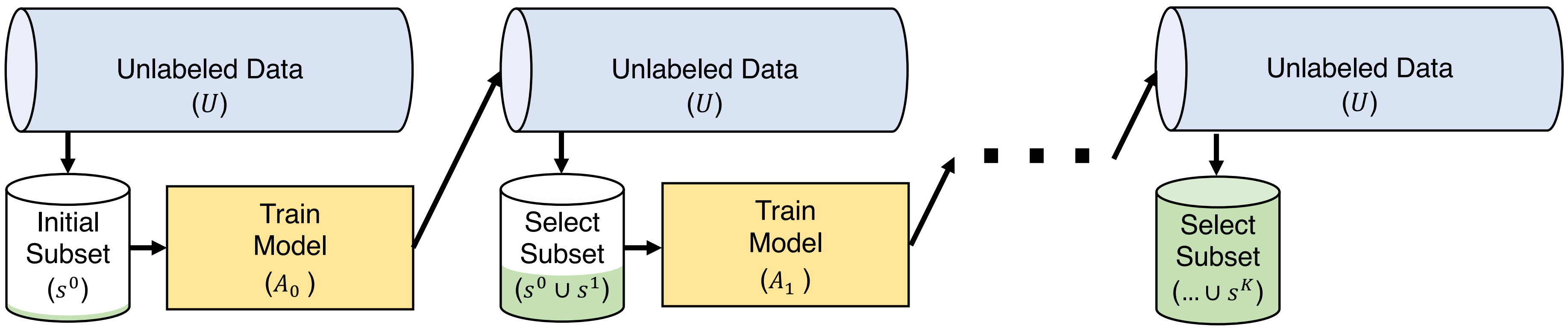
The goal is to select the best examples to improve the model.

What is active learning?



The goal is to select the best examples to improve the model.

What is active learning?

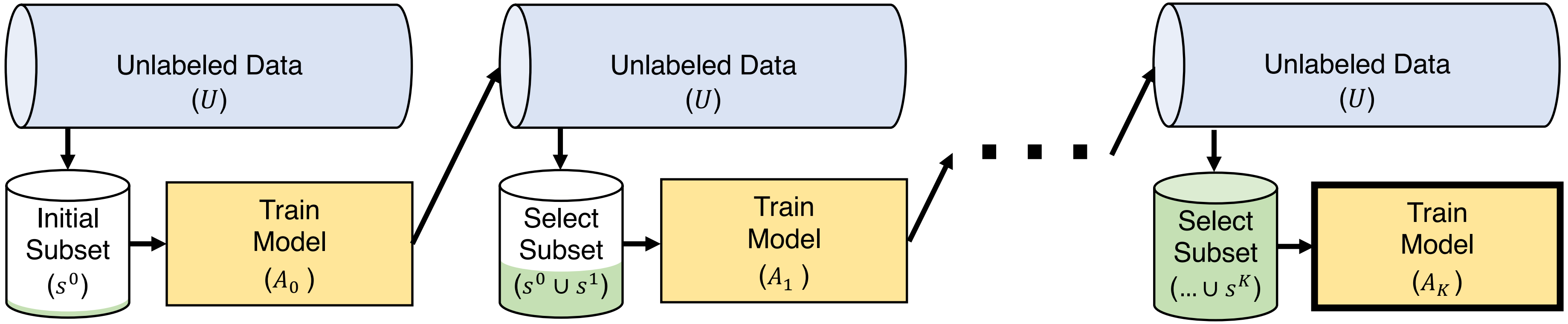


Traditional Approach

The goal is to select the best examples to improve the model.

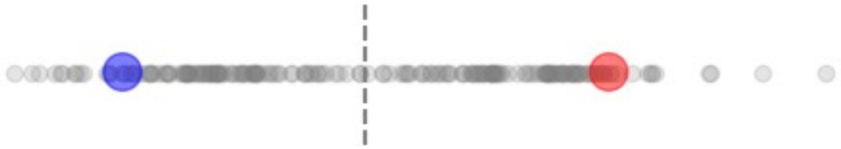
What is active learning?

Traditional Approach

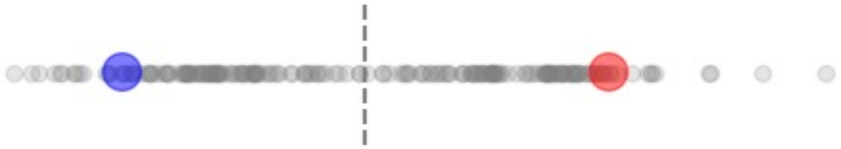


The goal is to select the best examples to improve the model.

1-D: Passive vs. Active Learning

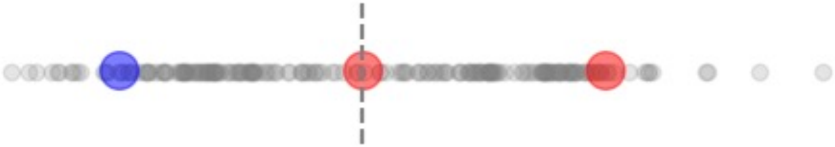


Active Learning
(i=0)

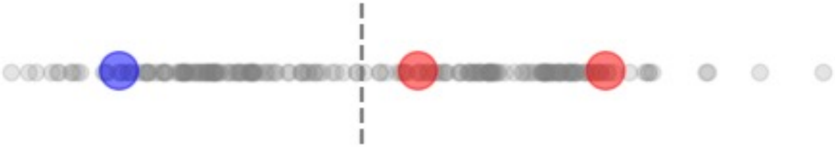


Passive Learning
(i=0)

1-D: Passive vs. Active Learning

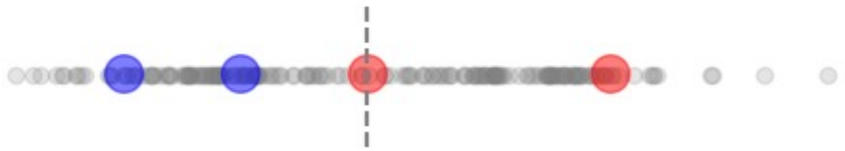


Active Learning
(i=1)

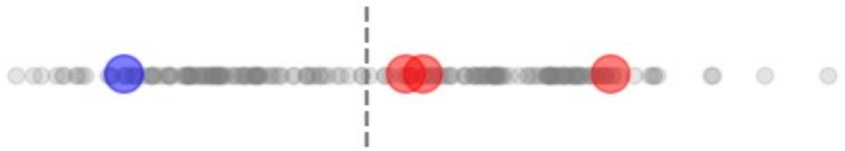


Passive Learning
(i=1)

1-D: Passive vs. Active Learning

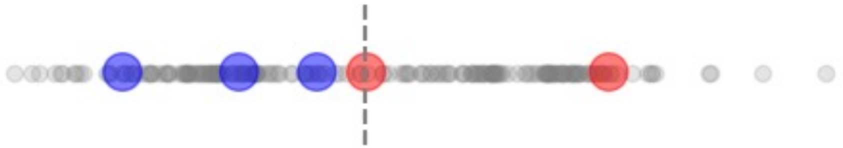


Active Learning
(i=2)

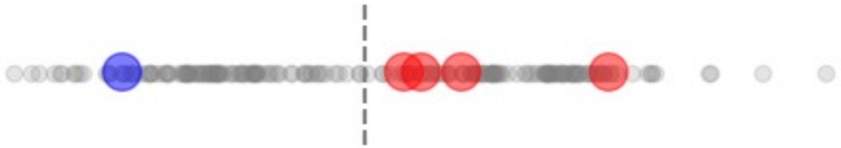


Passive Learning
(i=2)

1-D: Passive vs. Active Learning

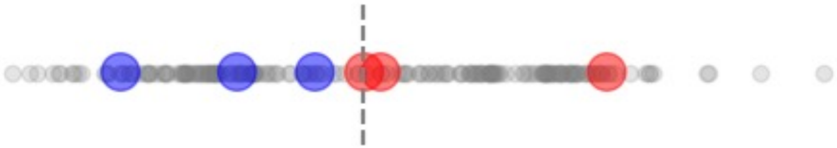


Active Learning
(i=3)

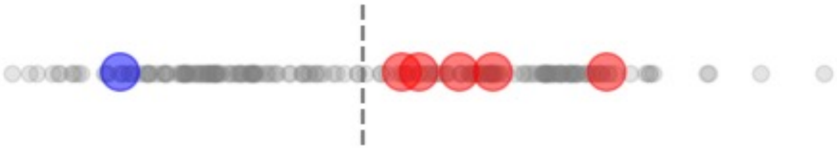


Passive Learning
(i=3)

1-D: Passive vs. Active Learning

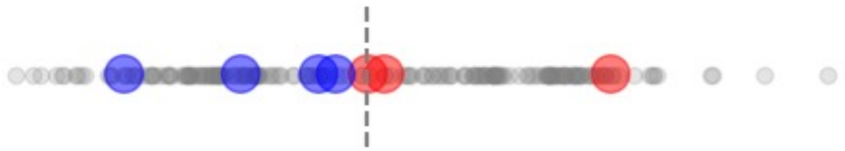


Active Learning
(i=4)

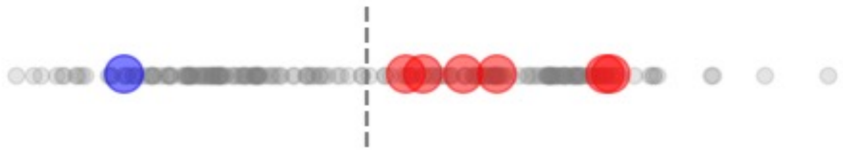


Passive Learning
(i=4)

1-D: Passive vs. Active Learning

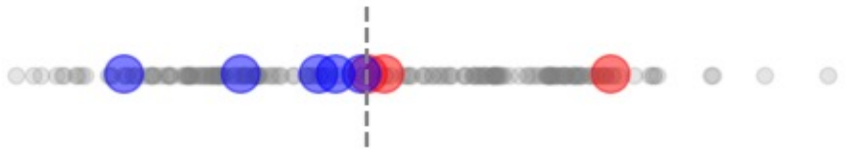


Active Learning
(i=5)

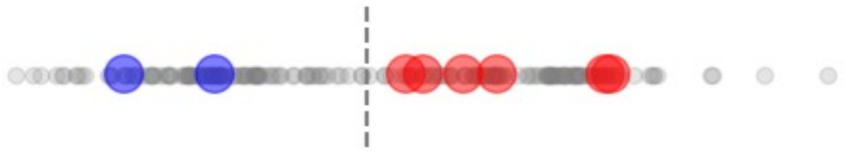


Passive Learning
(i=5)

1-D: Passive vs. Active Learning



Active Learning
(i=6)



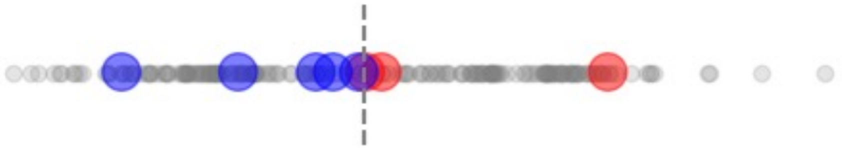
Passive Learning
(i=6)

active learning can give exponential speed-ups

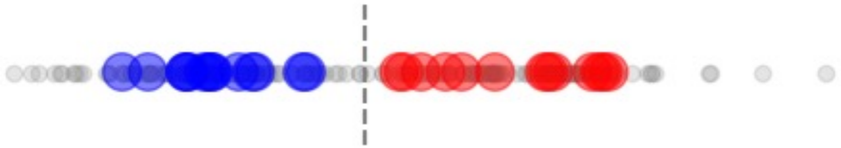
Passive: err ~ n^{-1}

Active: err ~ 2^{-n}

1-D: Passive vs. Active Learning



Active Learning
(i=6)



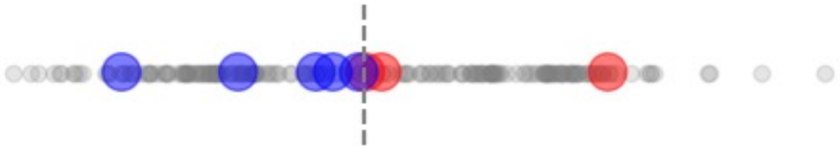
Passive Learning
(i=25)

active learning can give exponential speed-ups

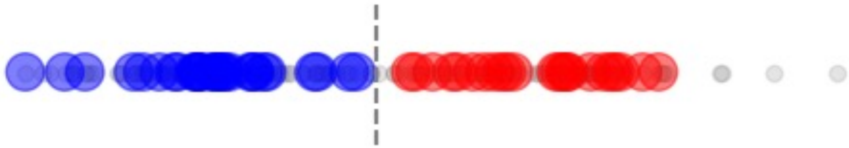
Passive: err ~ n^{-1}

Active: err ~ 2^{-n}

1-D: Passive vs. Active Learning



Active Learning
(i=6)



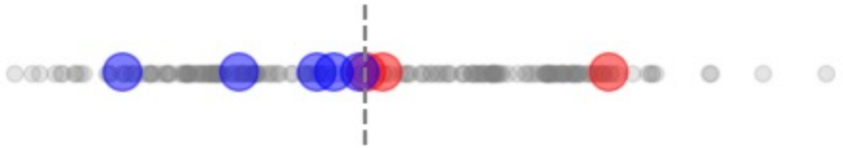
Passive Learning
(i=50)

active learning can give exponential speed-ups

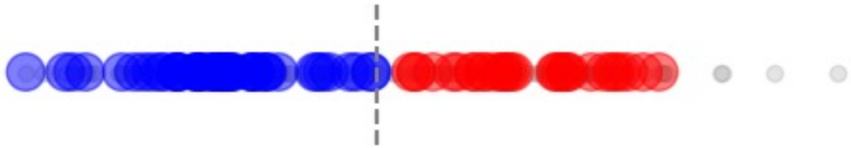
Passive: err ~ n^{-1}

Active: err ~ 2^{-n}

1-D: Passive vs. Active Learning



Active Learning
(i=6)

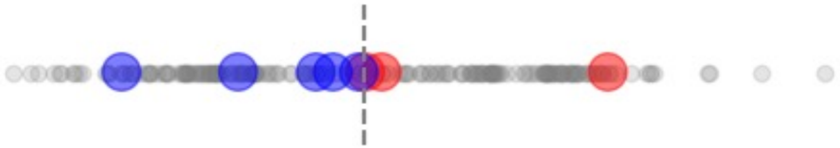


Passive Learning
(i=75)

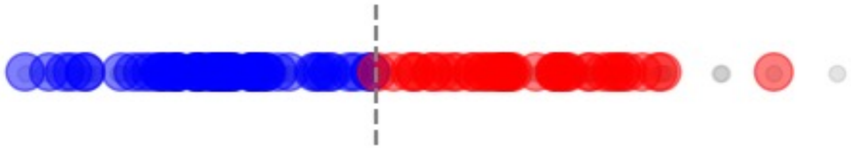
active learning can give exponential speed-ups

Passive: err ~ n^{-1}
Active: err ~ 2^{-n}

1-D: Passive vs. Active Learning



Active Learning
(i=6)



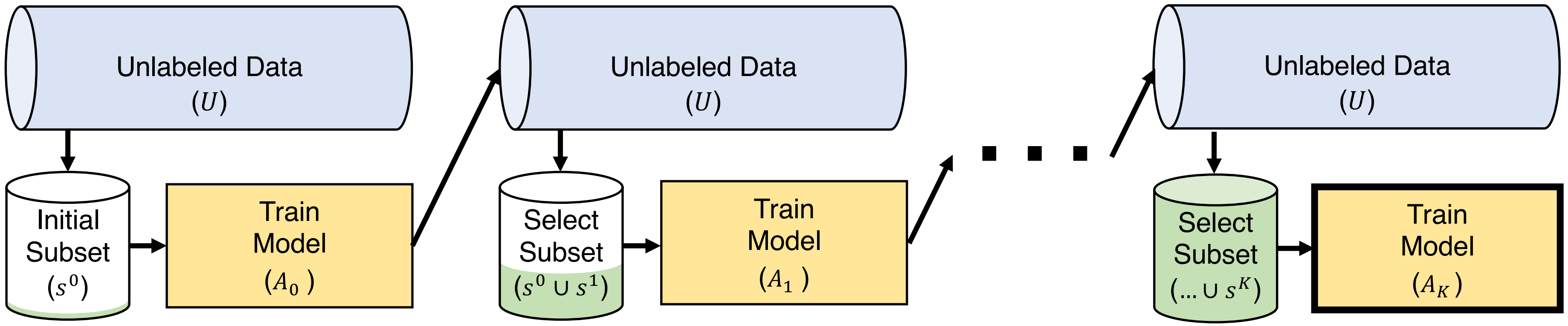
Passive Learning
(i=99)

active learning can give exponential speed-ups

Passive: err ~ n^{-1}
Active: err ~ 2^{-n}

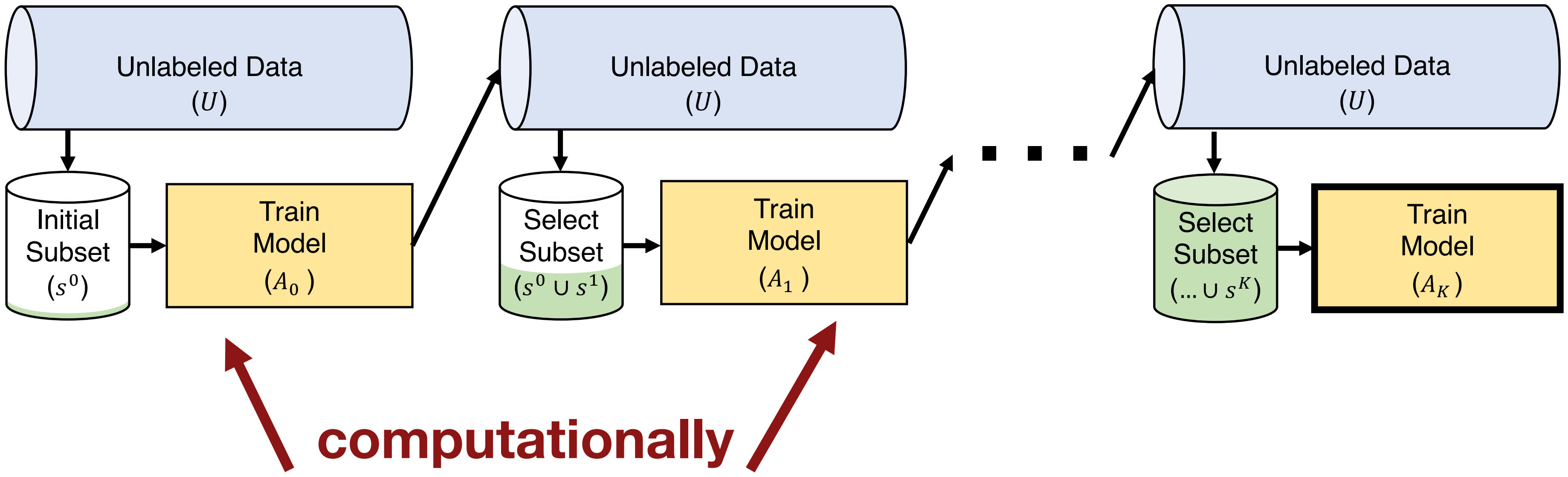
Practical Challenge #1: Big Models

Traditional Approach



Practical Challenge #1: Big Models

Traditional Approach



Unlabeled Data (U)

Initial Subset (s^0)

Train Model (A_0)

Unlabeled Data (U)

Select Subset $(s^0 \cup s^1)$

Train Model (A_1)

. . .

Unlabeled Data (U)

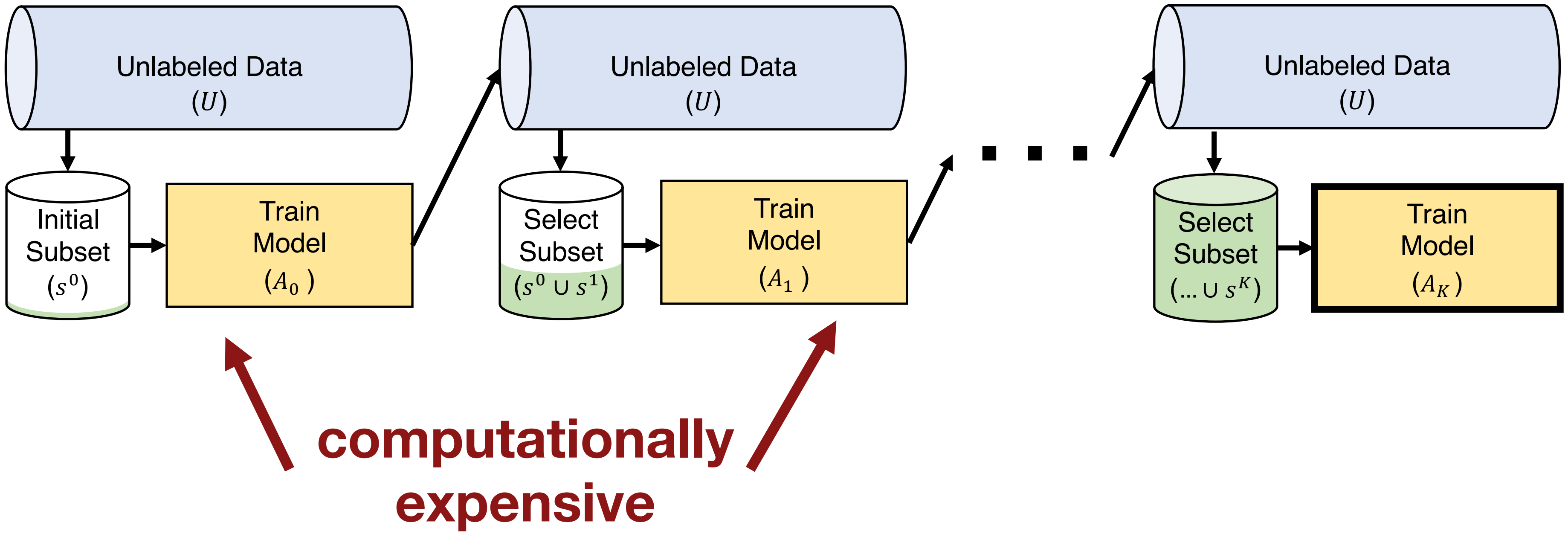
Select Subset $(\dots \cup s^K)$

Train Model (A_K)

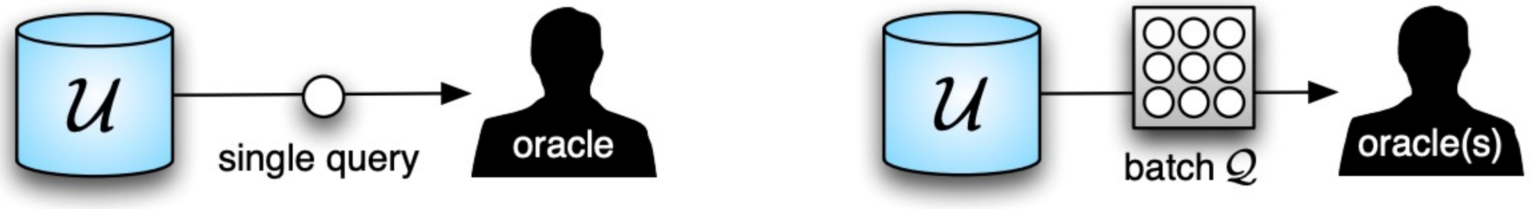
computationally expensive

Practical Challenge #1: Big Models

Traditional Approach



computationally expensive



Settles. “From Theories to Queries: Active Learning in Practice.” 2011

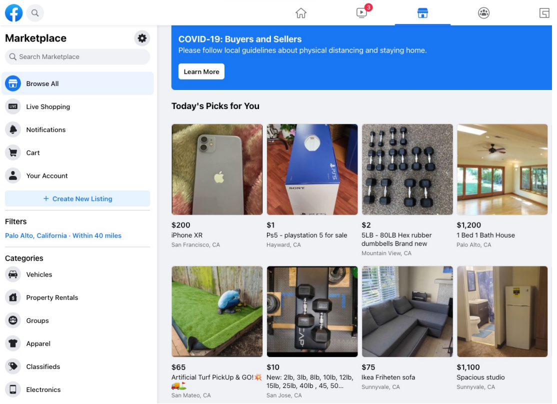
Practical Challenge #2: Big Data

Recommendation

Photo: Gersh Kuntzman

Photo: Stephen Lam | Reuters

Model Debugging



Marketplace

Search Marketplace

Browse All

Live Shopping

Notifications

Cart

Your Account

+ Create New Listing

Filters

Palo Alto, California · Within 40 miles

Categories

Vehicles

Property Rentals

Groups

Apparel

Classifieds

Electronics

COVID-19: Buyers and Sellers

Please follow local guidelines about physical distancing and staying home.

Learn More

Today's Picks for You

$200
iPhone XR
San Francisco, CA

$1
Ps5 - playstation 5 for sale
Hayward, CA

$2
5LB - 80LB Hex rubber dumbbells Brand new
Mountain View, CA

$1,200
1 Bed 1 Bath House
Palo Alto, CA

$65
Artificial Turf PickUp & GO!💥🚚⛳
San Mateo, CA

$10
New: 2lb, 3lb, 8lb, 10lb, 12lb, 15lb, 25lb, 40lb , 45, 50...
San Jose, CA

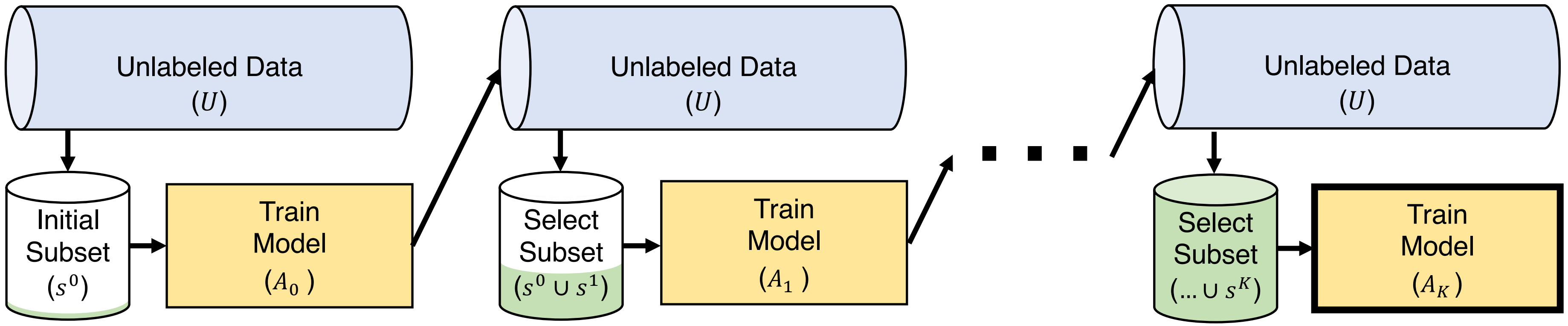
$75
Ikea Friheten sofa
Sunnyvale, CA

$1,100
Spacious studio
Sunnyvale, CA

Integrity

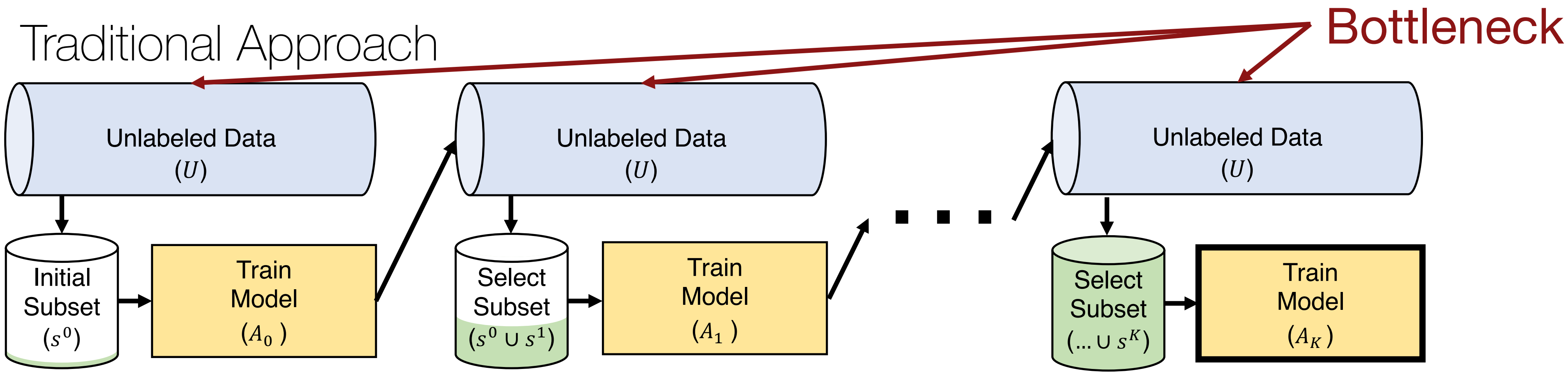
Practical Challenge #2: Big Data

Traditional Approach



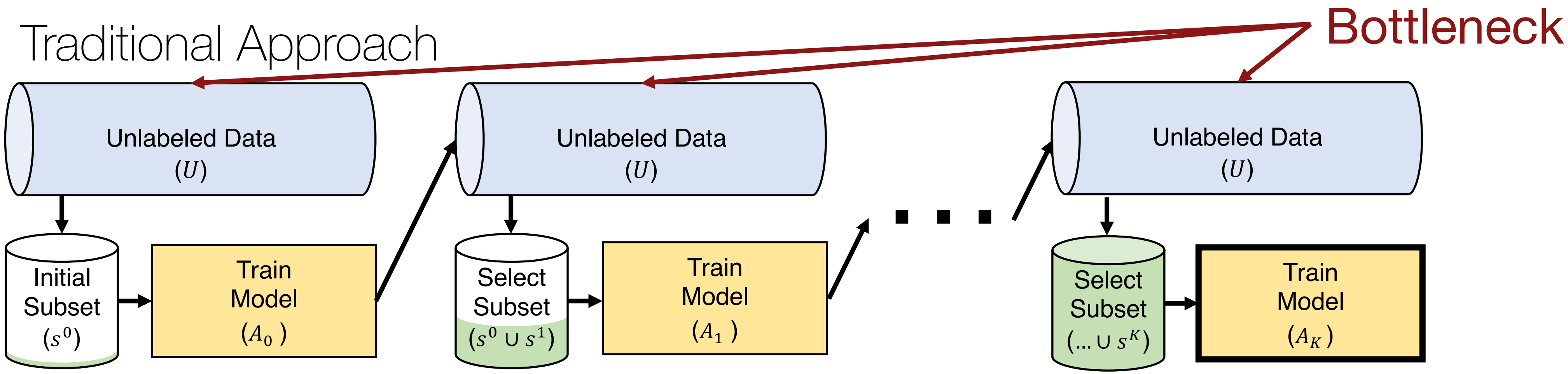
Coleman et al. "Similarity Search for Efficient Active Learning and Search of Rare Concepts." 2022

Practical Challenge #2: Big Data



Coleman et al. "Similarity Search for Efficient Active Learning and Search of Rare Concepts." 2022

Practical Challenge #2: Big Data



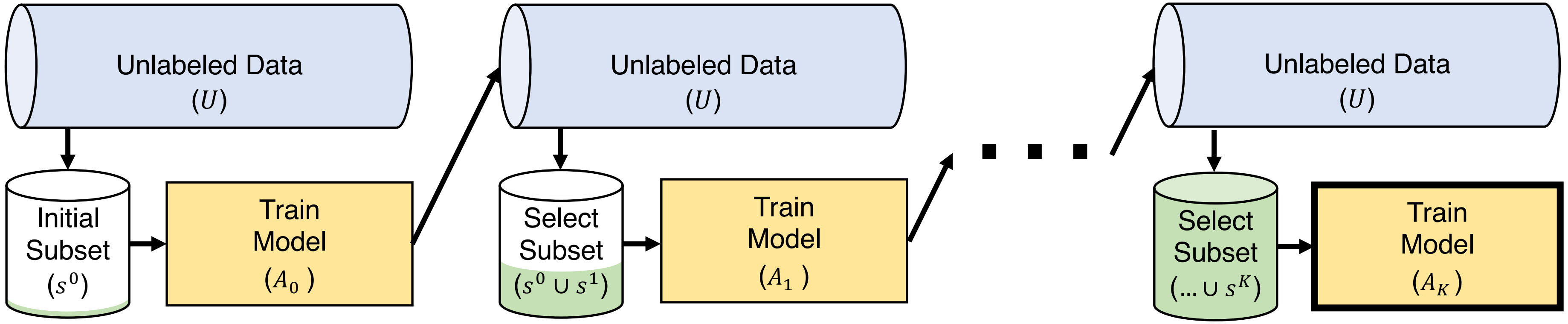
For example, running a single inference pass over 10 billion images with a ResNet-50 model would take **38 exaFLOPs or roughly 40 GPU-months.**

Evaluating all of the unlabeled examples is too slow

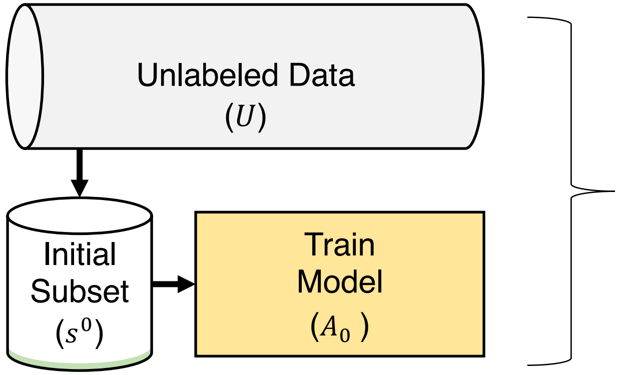
Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

Practical Challenge #2: Big Data

Traditional Approach



Similarity search for Efficient Active Learning and Search (SEALS)

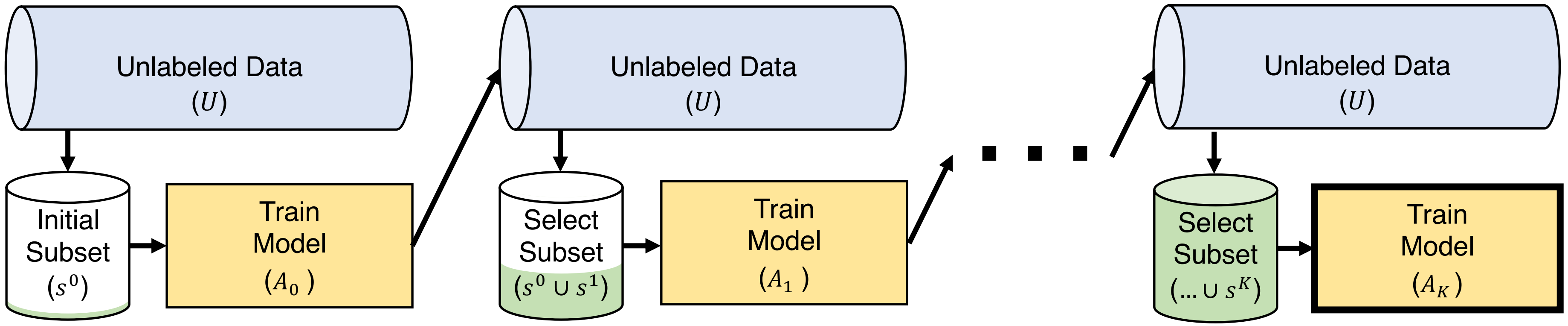


Start the same way

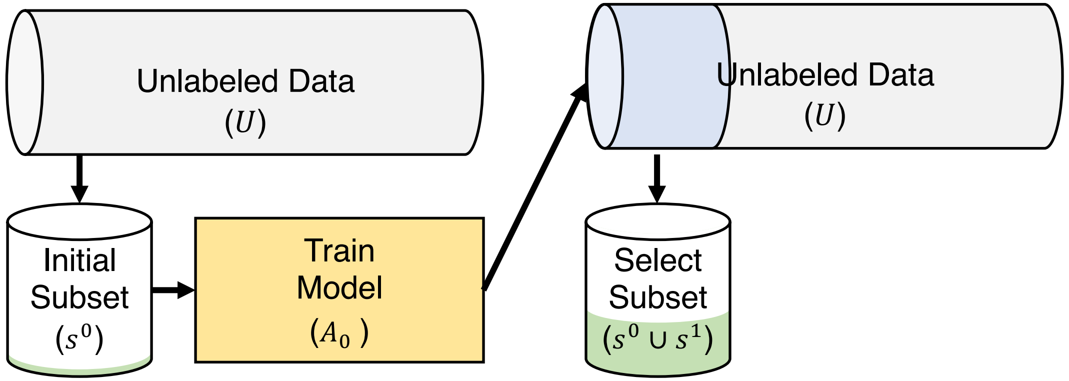
Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

Practical Challenge #2: Big Data

Traditional Approach



Similarity search for Efficient Active Learning and Search (SEALS)

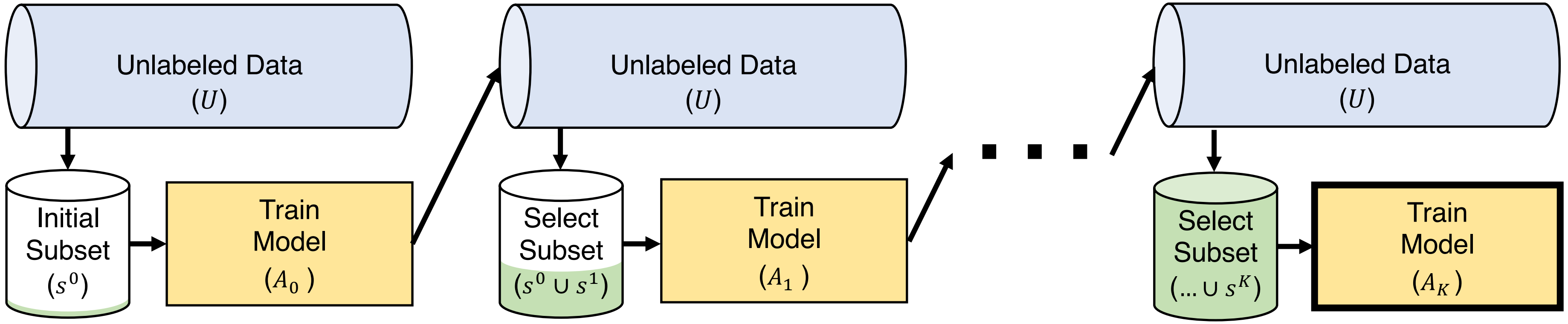


But instead of applying our selection strategy to all the unlabeled data, we use similarity search to find the closest examples and only consider them.

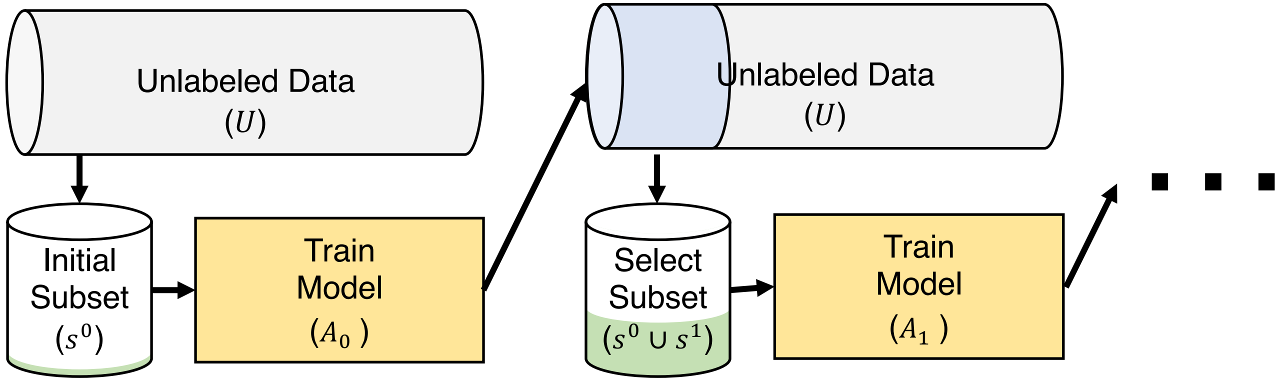
Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

Practical Challenge #2: Big Data

Traditional Approach



Similarity search for Efficient Active Learning and Search (SEALS)



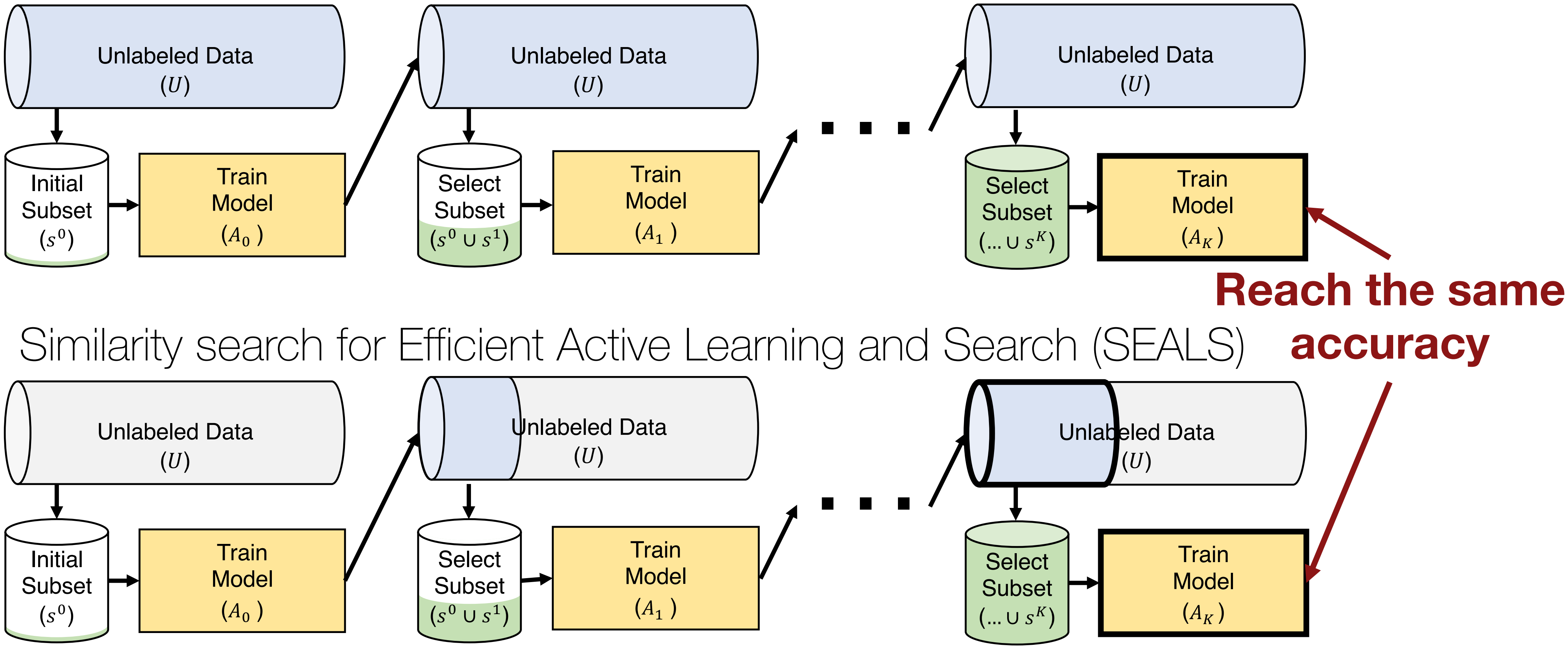
Take the newly labeled examples and find their neighbors. Then repeat

Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

Practical Challenge #2: Big Data

Traditional Approach

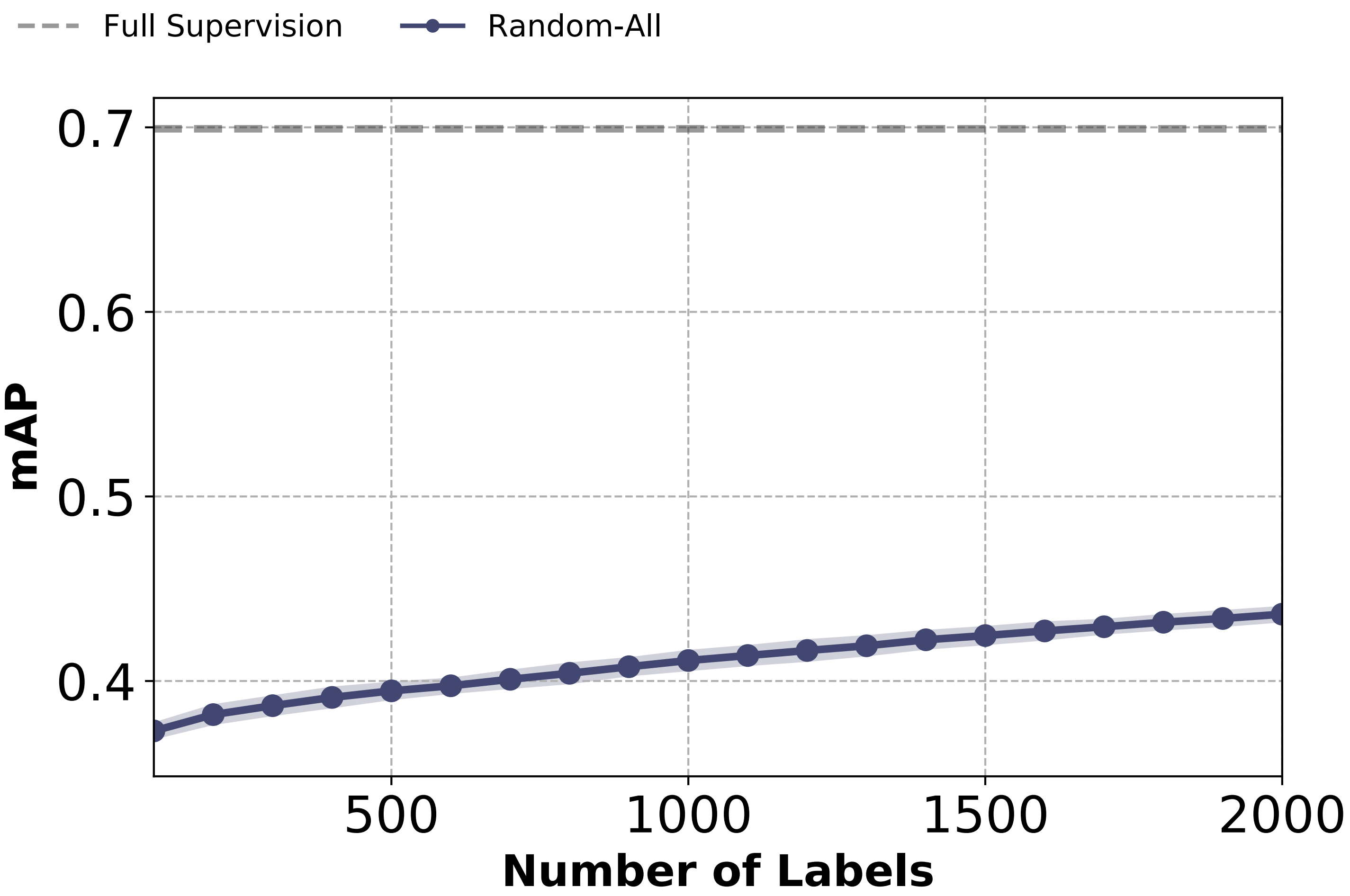
Similarity search for Efficient Active Learning and Search (SEALS)



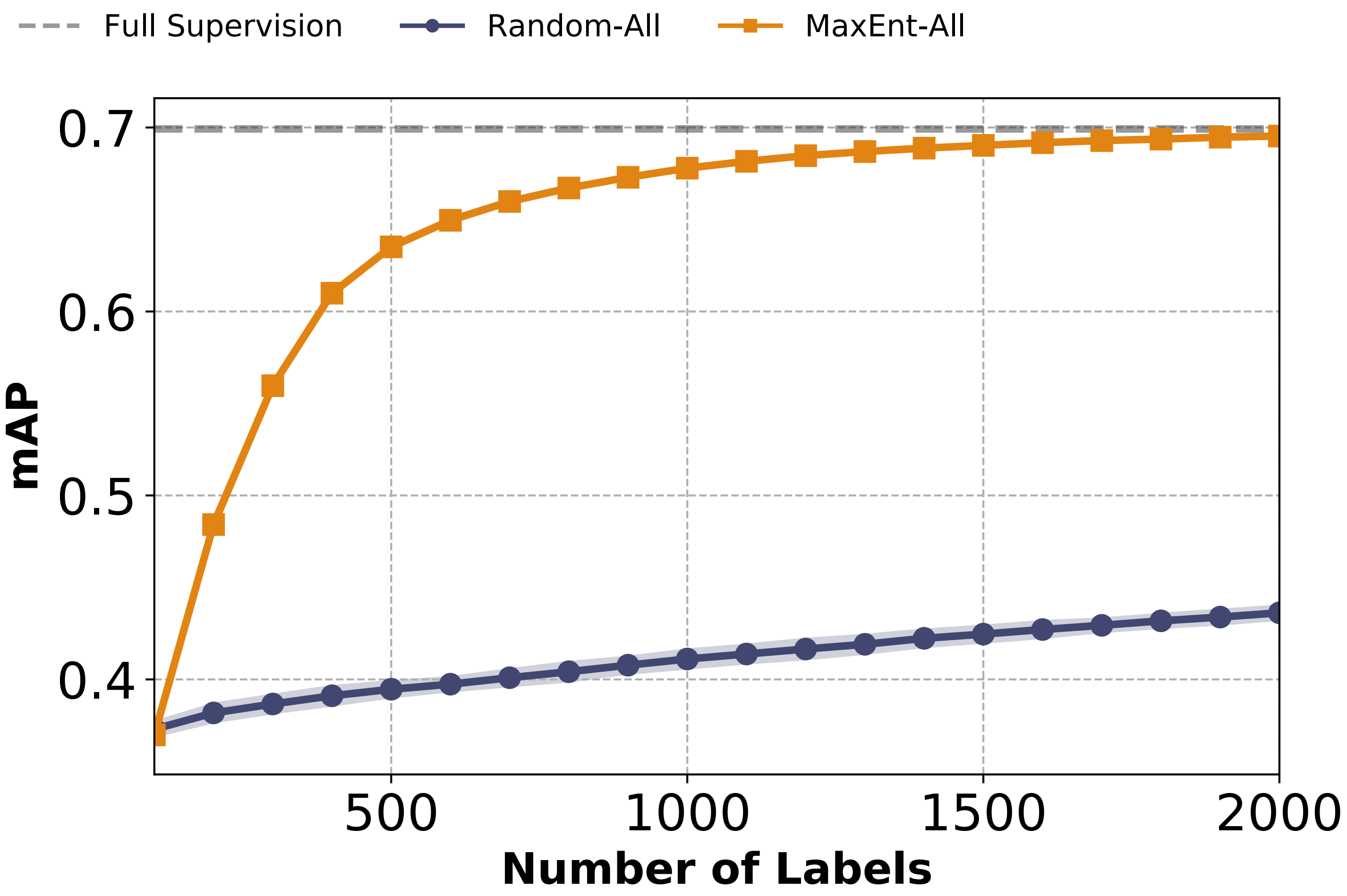
Reach the same accuracy

Coleman et al. “Similarity Search for Efficient Active Learning and Search of Rare Concepts.” 2022

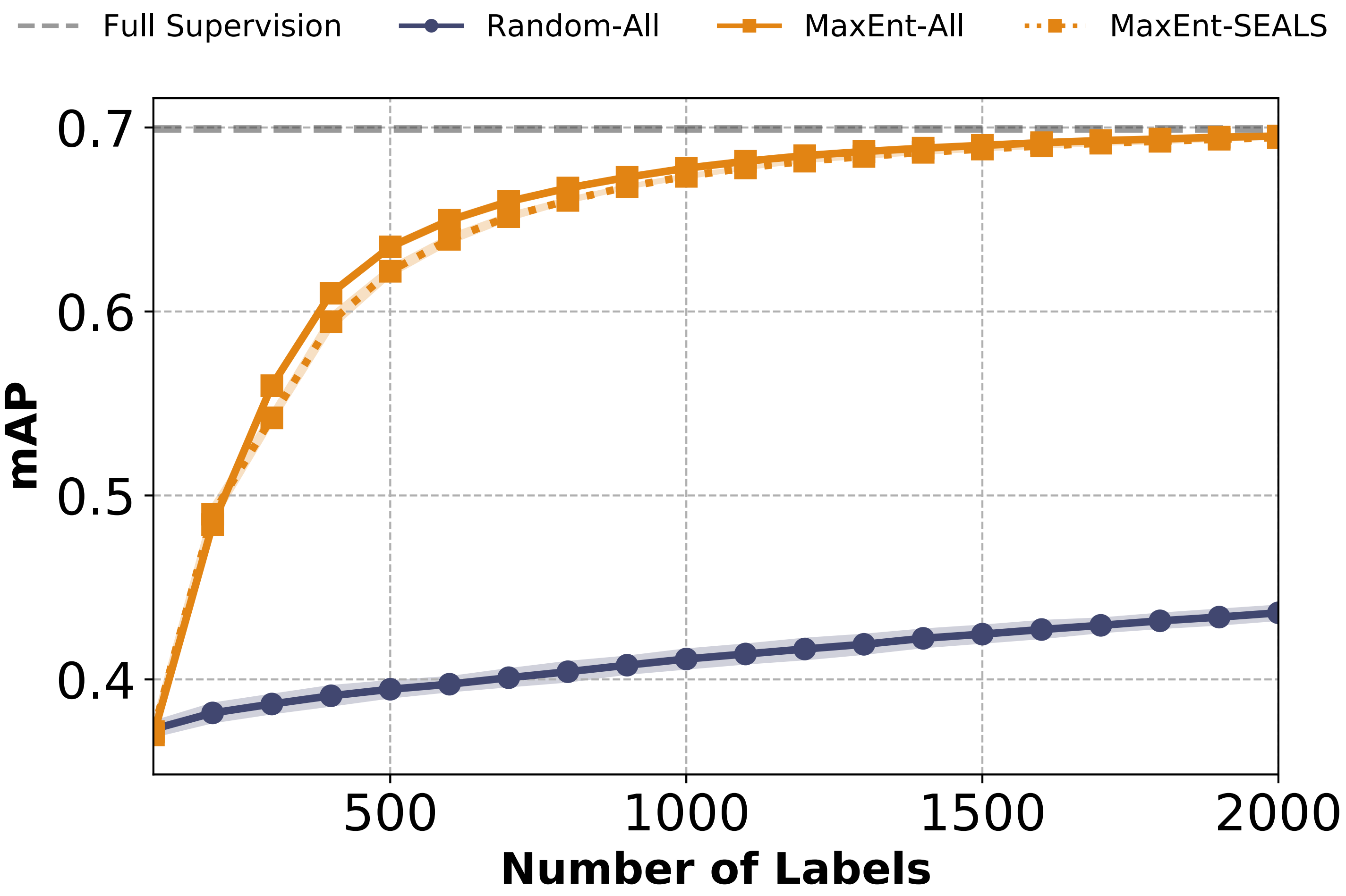
Active learning on ImageNet



Active learning on ImageNet

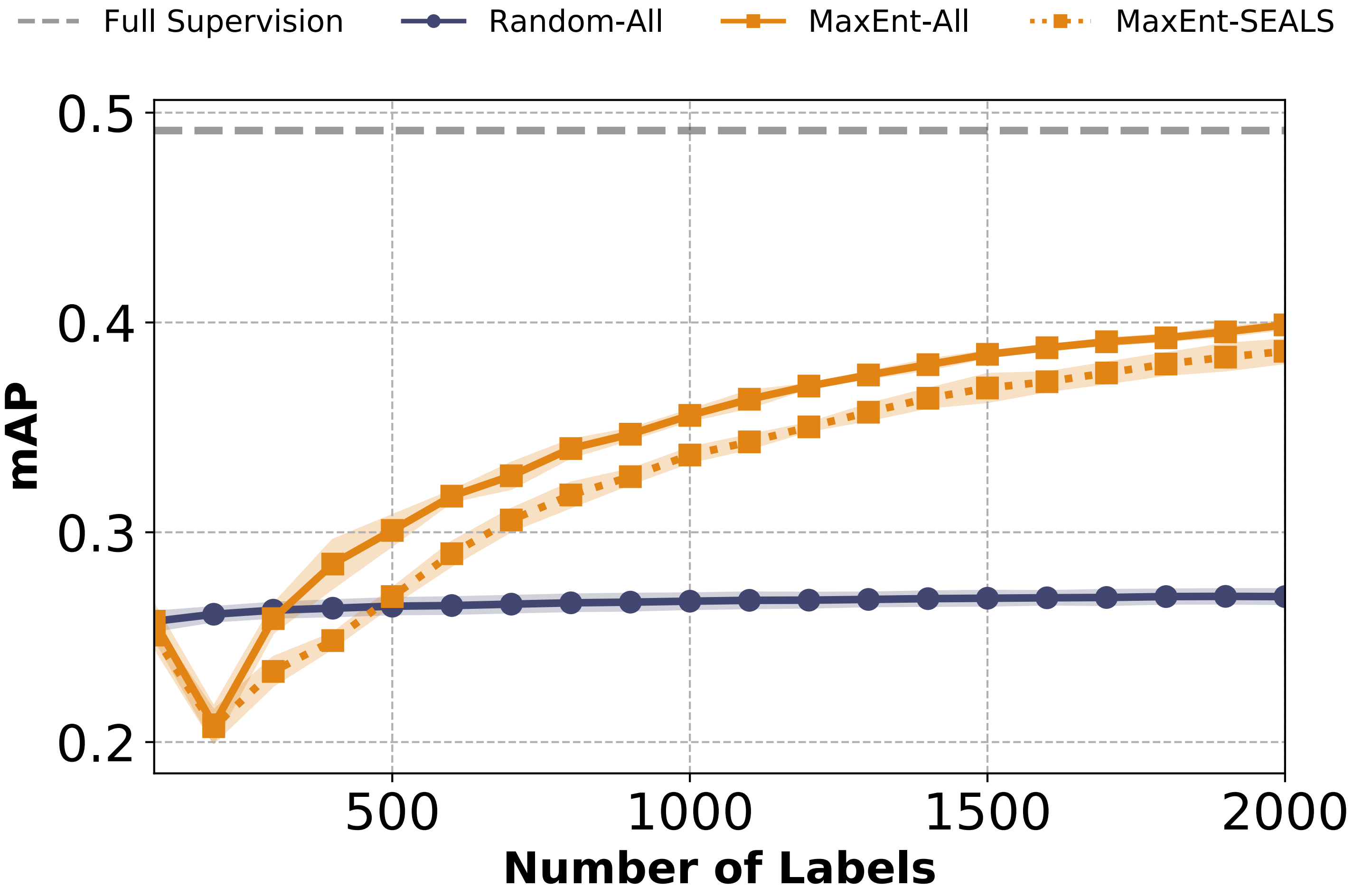


Active learning on ImageNet



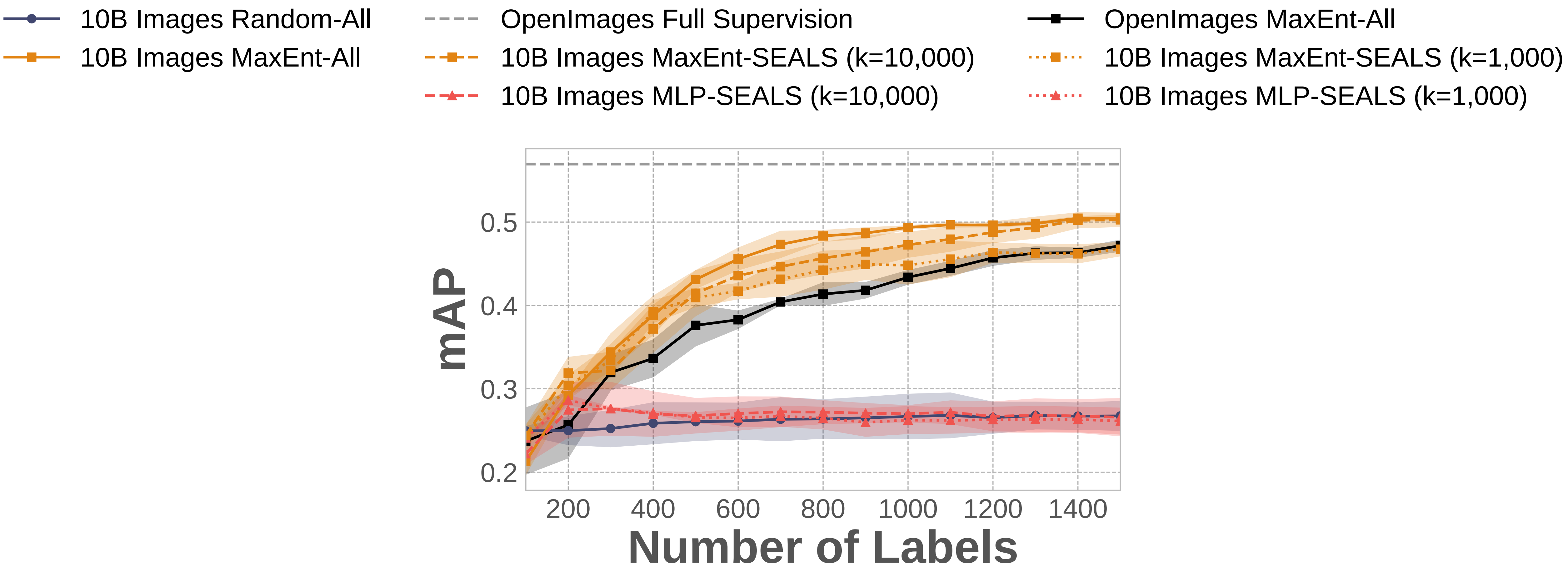
SEALS is within 0.001 mAP while only considering **< 10% of the unlabeled data**.

Active learning on OpenImages (6.8M images)



SEALS is within 0.013 mAP while only considering **< 1% of the unlabeled data**.

Active learning on 10B images



SEALS is within 0.004 mAP while only considering **< 0.1% of the unlabeled data**.

Today's Lecture

Why care about labels? Data-Centric vs Model-Centric AI (1/17)

Who labels and how? Dataset Creation and Curation (1/19)

What to label? Growing or Compressing Datasets (Today)

- Active learning for growing datasets
- **Core-set selection for compressing datasets**

Large Labeled Datasets

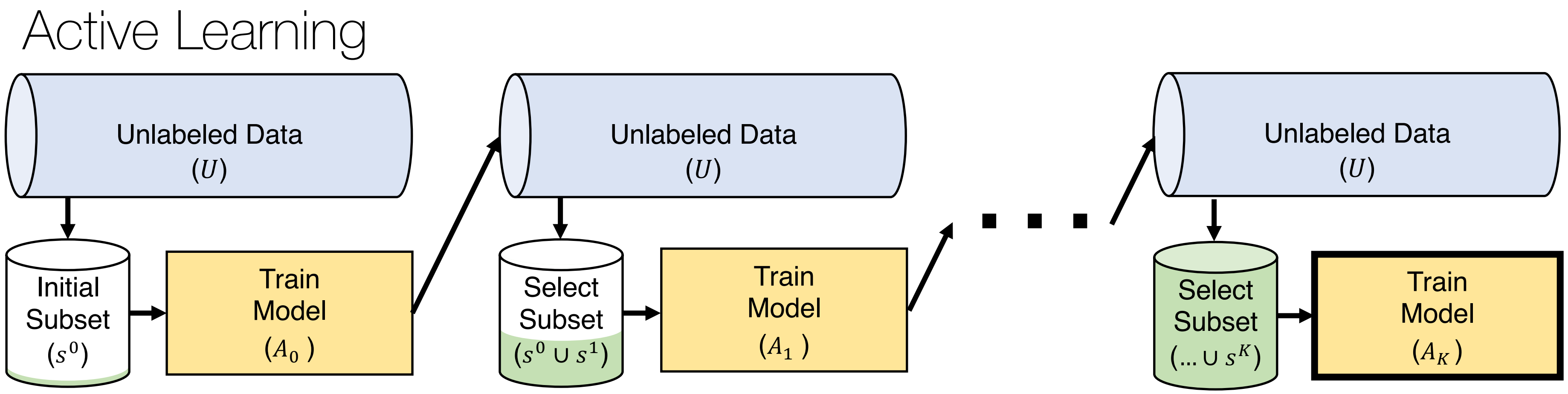
Systematic feedback

- Tagging friends in images
- Flagging emails as spam
- Rating items or movies

Self-supervision

- Language modeling (e.g., BERT)
- Computer vision (e.g., SimCLR)

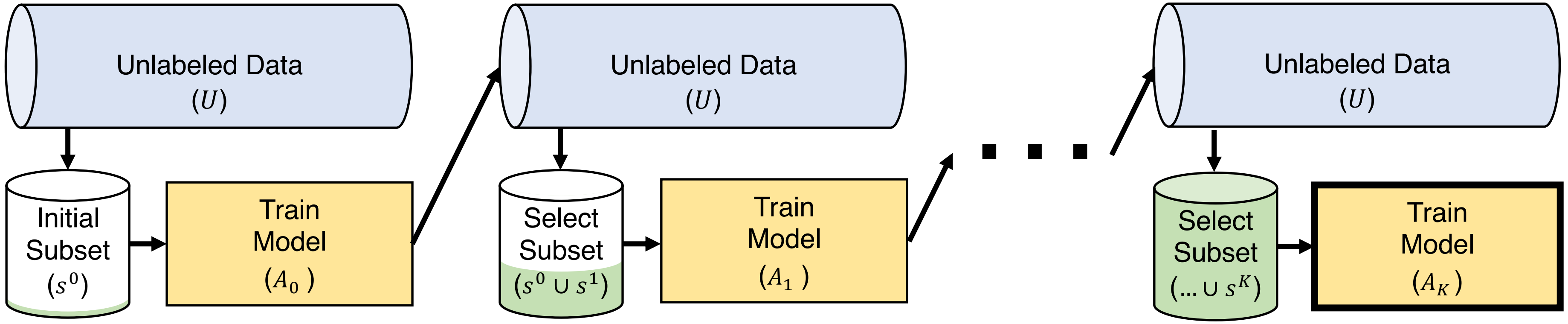
What is core-set selection?



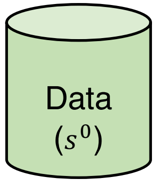
Core-set selection aims to select a small subset of data that accurately approximates the full dataset.

What is core-set selection?

Active Learning

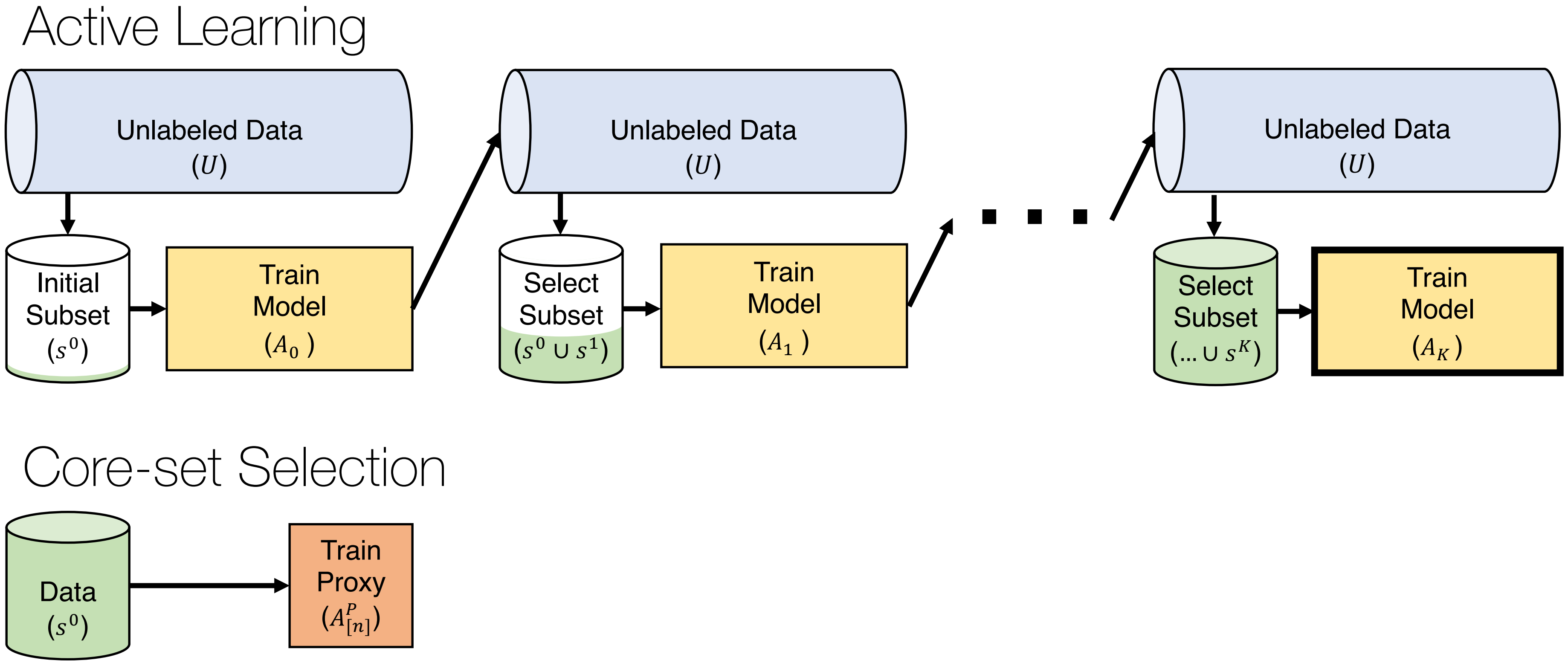


Core-set Selection



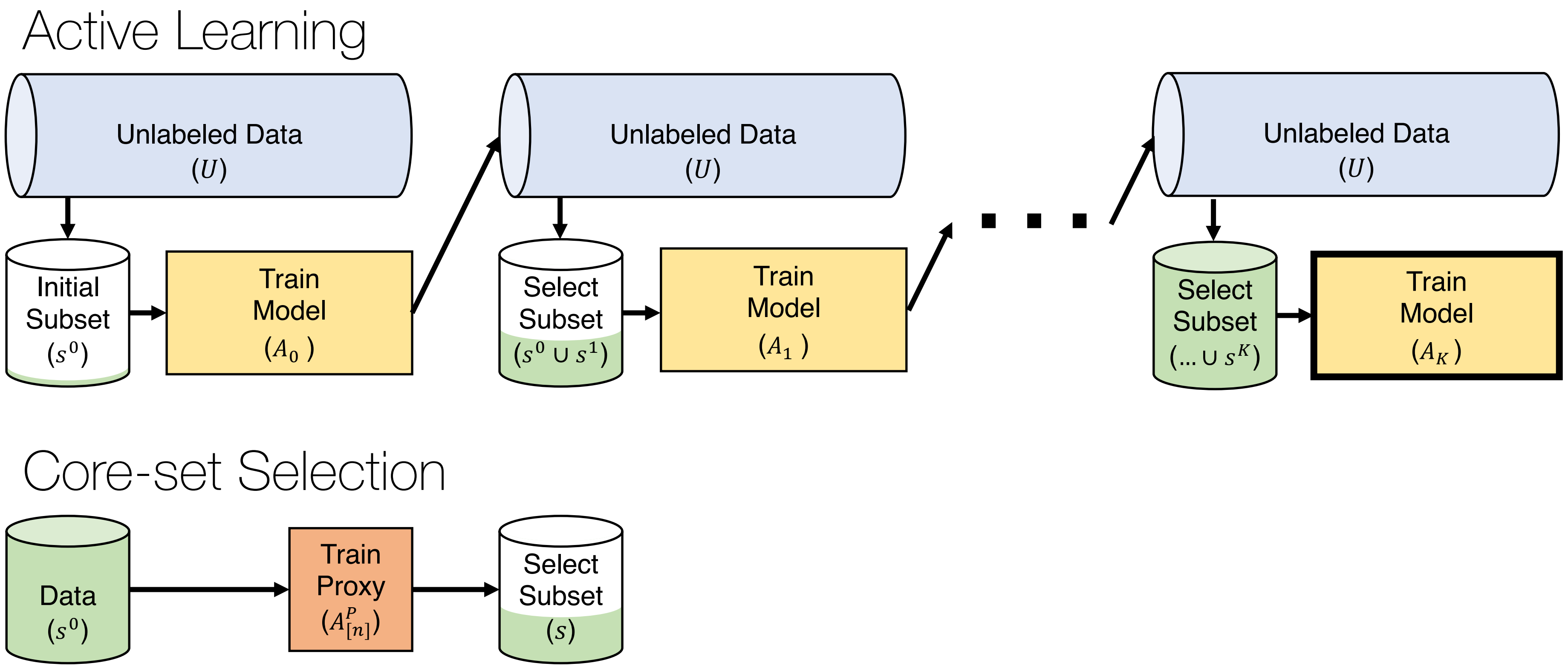
Core-set selection aims to select a small subset of data that accurately approximates the full dataset.

What is core-set selection?



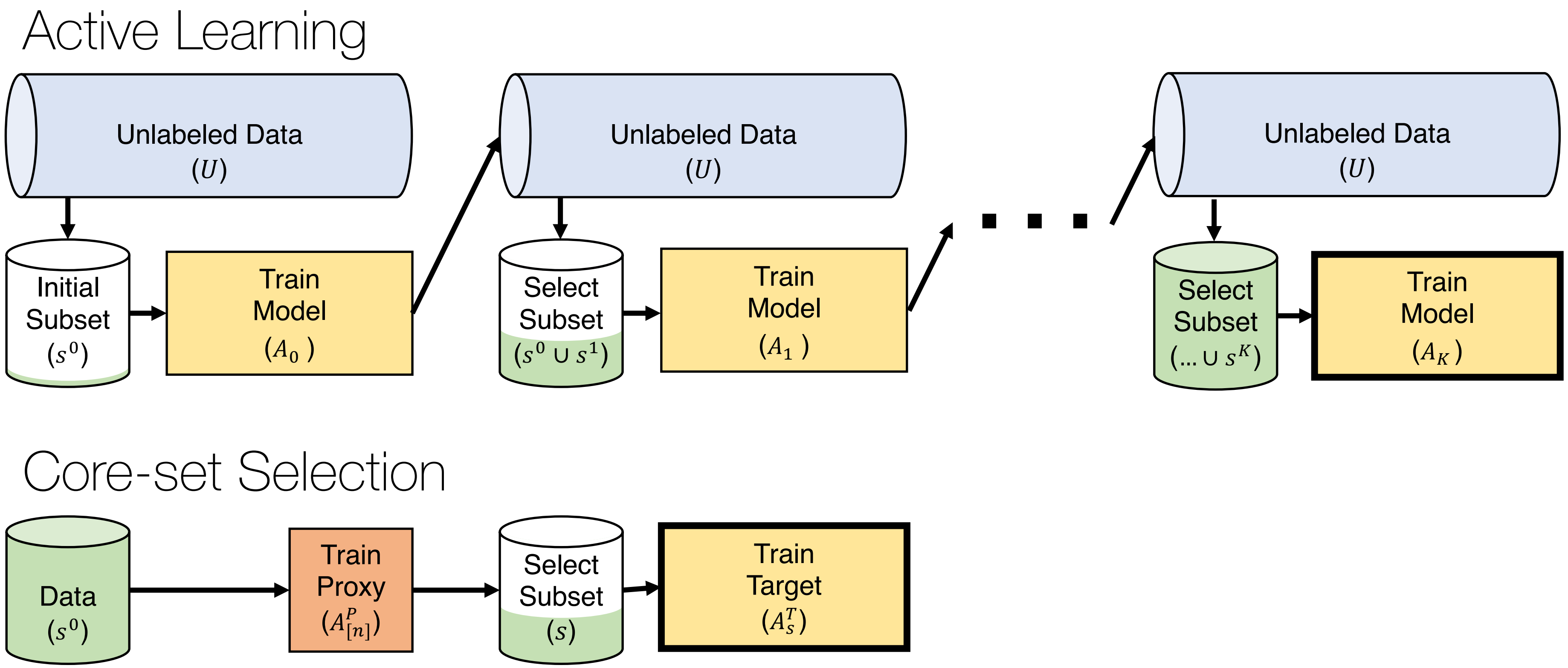
Core-set selection aims to select a small subset of data that accurately approximates the full dataset.

What is core-set selection?

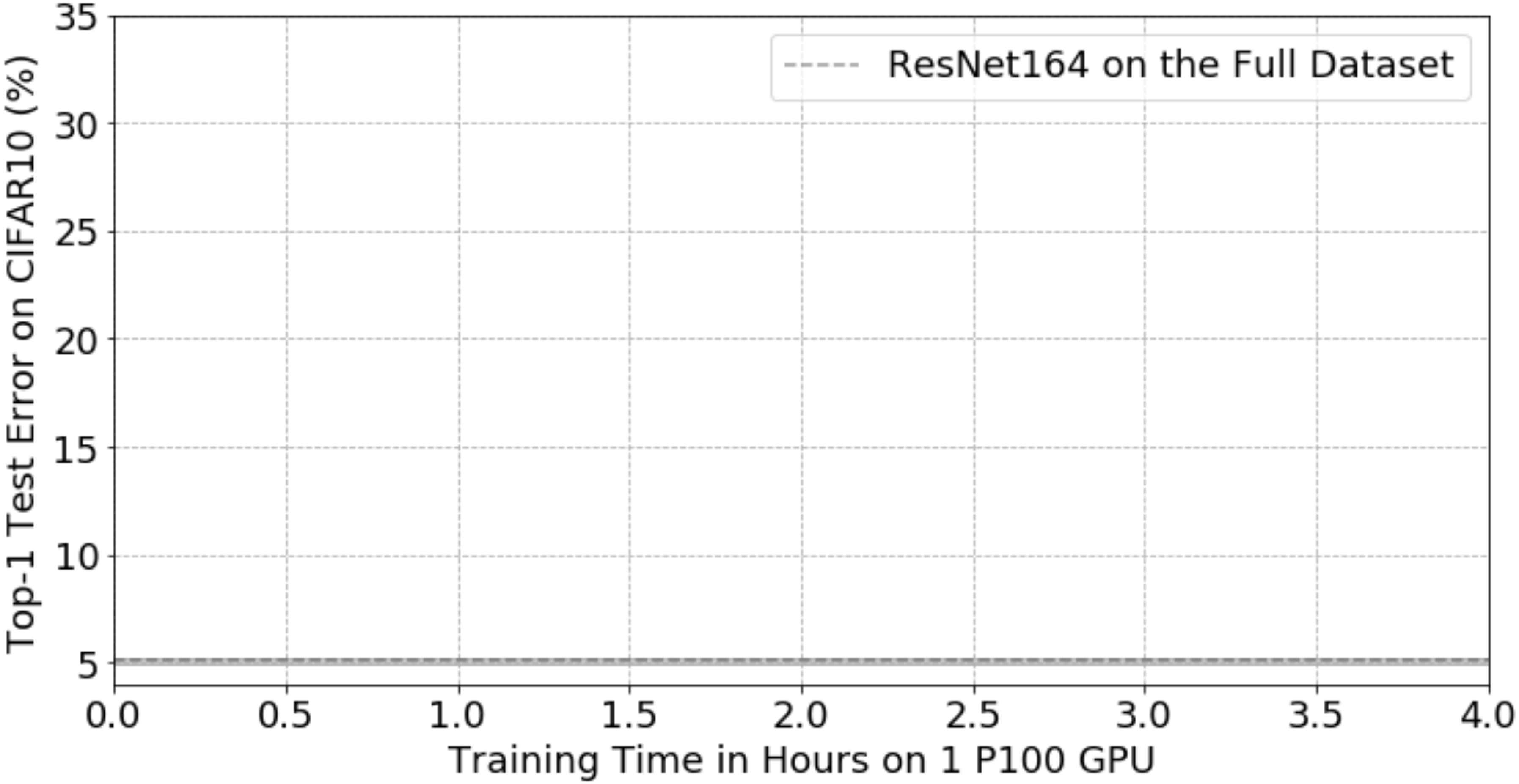


Core-set selection aims to select a small subset of data that accurately approximates the full dataset.

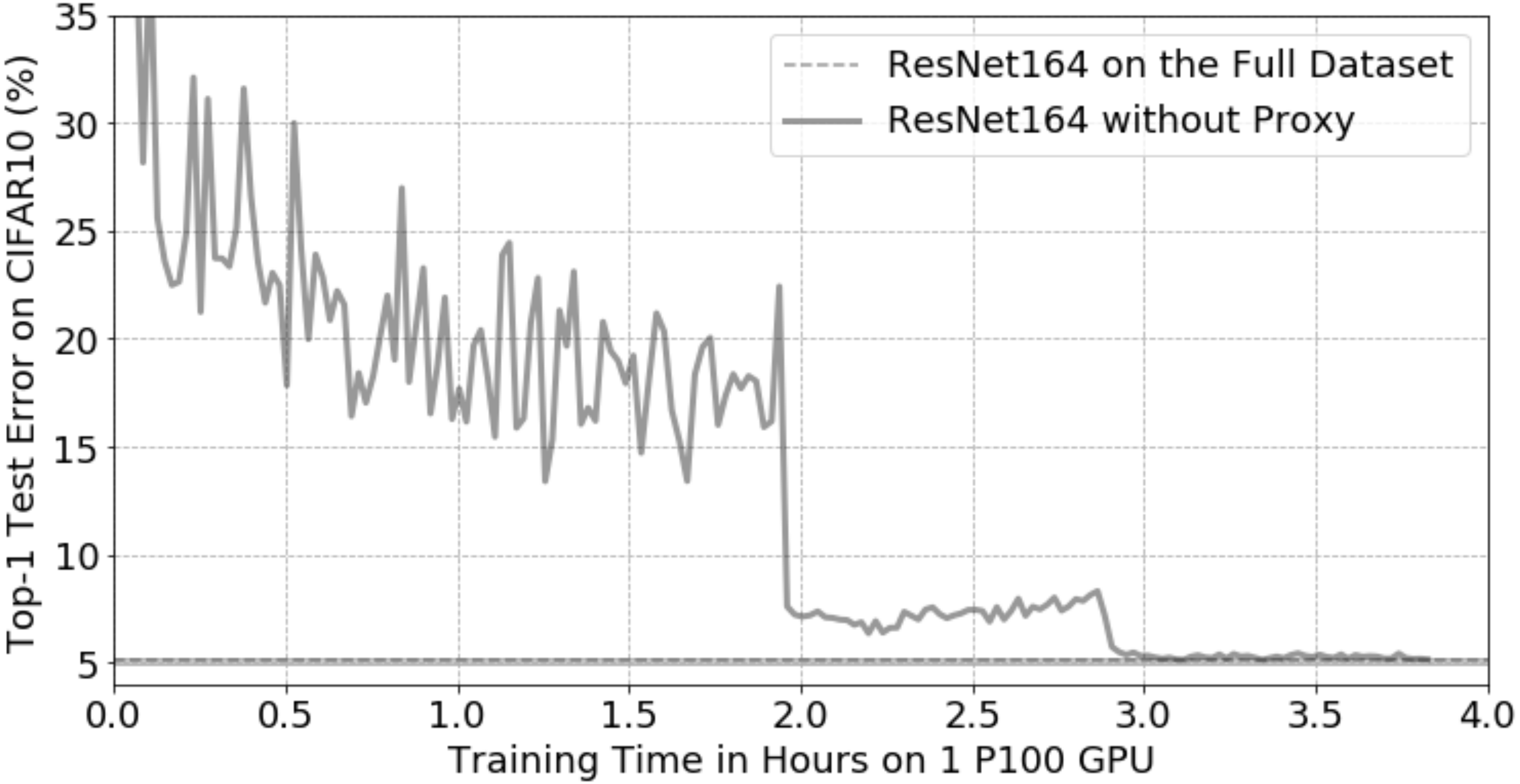
What is core-set selection?



Core-set selection aims to select a small subset of data that accurately approximates the full dataset.

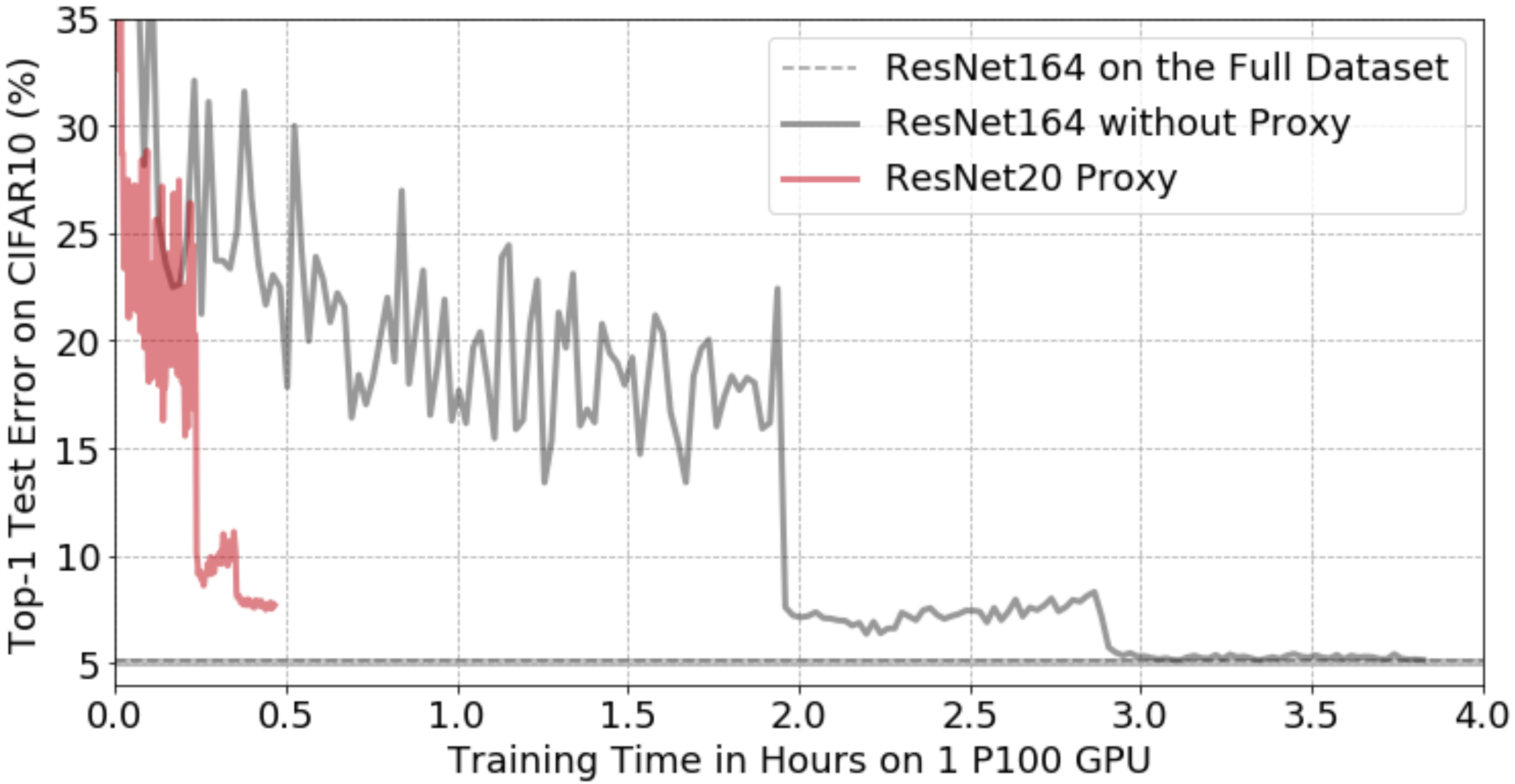


Coleman et al. “Selection via Proxy: Efficient Data Selection for Deep Learning.” 2020



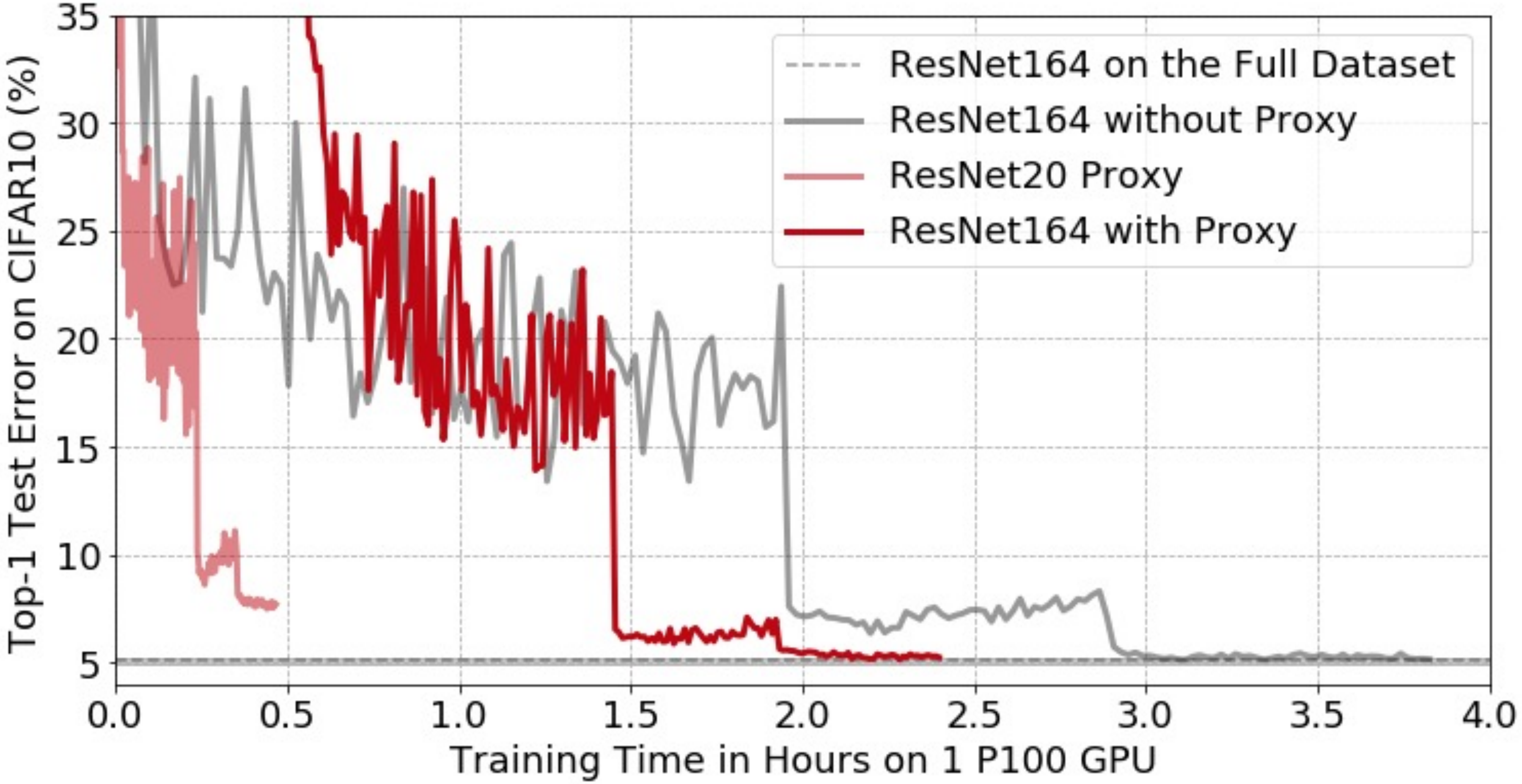
Training ResNet164 on CIFAR10 took **3 hours and 50 minutes** using a P100 GPU.

Coleman et al. “Selection via Proxy: Efficient Data Selection for Deep Learning.” 2020



Training ResNet20 only takes 28 minutes and allows us to **filter 50% of the data**

Coleman et al. “Selection via Proxy: Efficient Data Selection for Deep Learning.” 2020



Training ResNet20 only takes 28 minutes and allows us to **filter 50% of the data** without affecting the accuracy of ResNet164, leading to **a 1.6x speed-up** in time-to-accuracy

Coleman et al. “Selection via Proxy: Efficient Data Selection for Deep Learning.” 2020

This is only the tip of the Iceberg

What to label? Growing or Compressing Datasets

- Active learning for growing datasets
- Core-set selection for compressing datasets
- Generative active learning…
- Active search for drug discovery…
- Hard example mining…
- Curriculum learning...
- And much more…