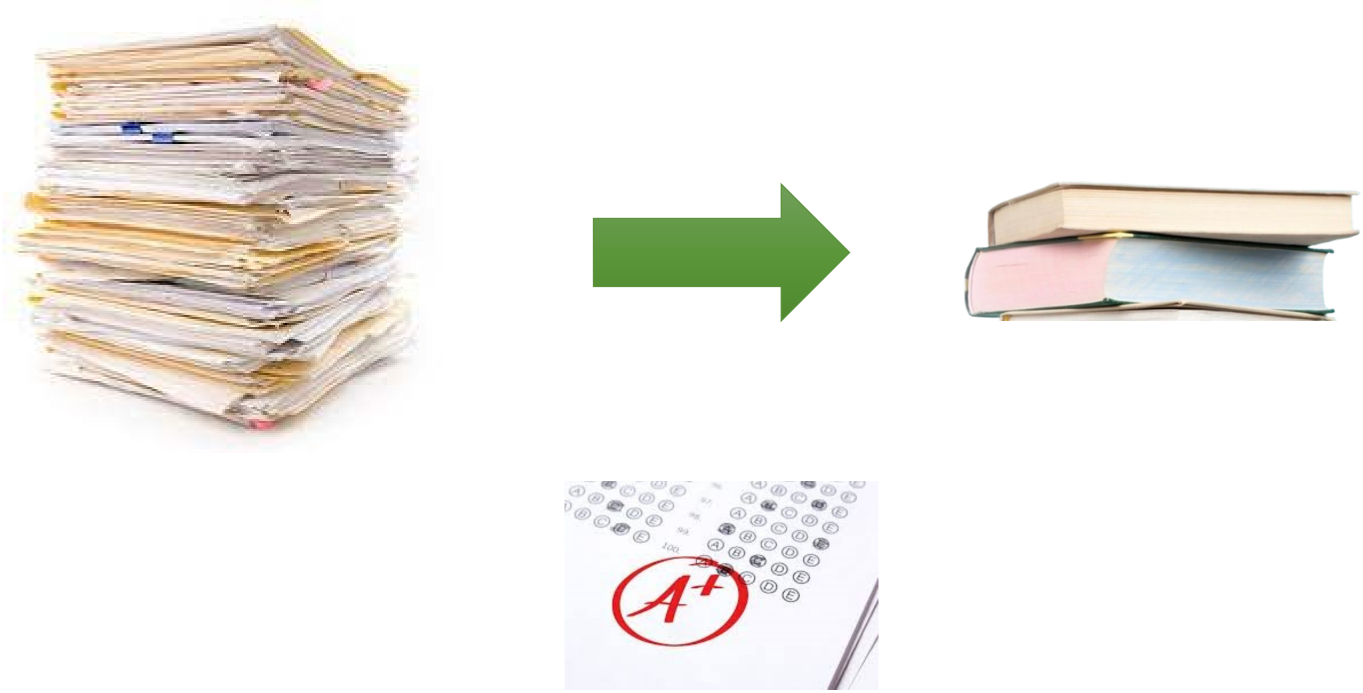
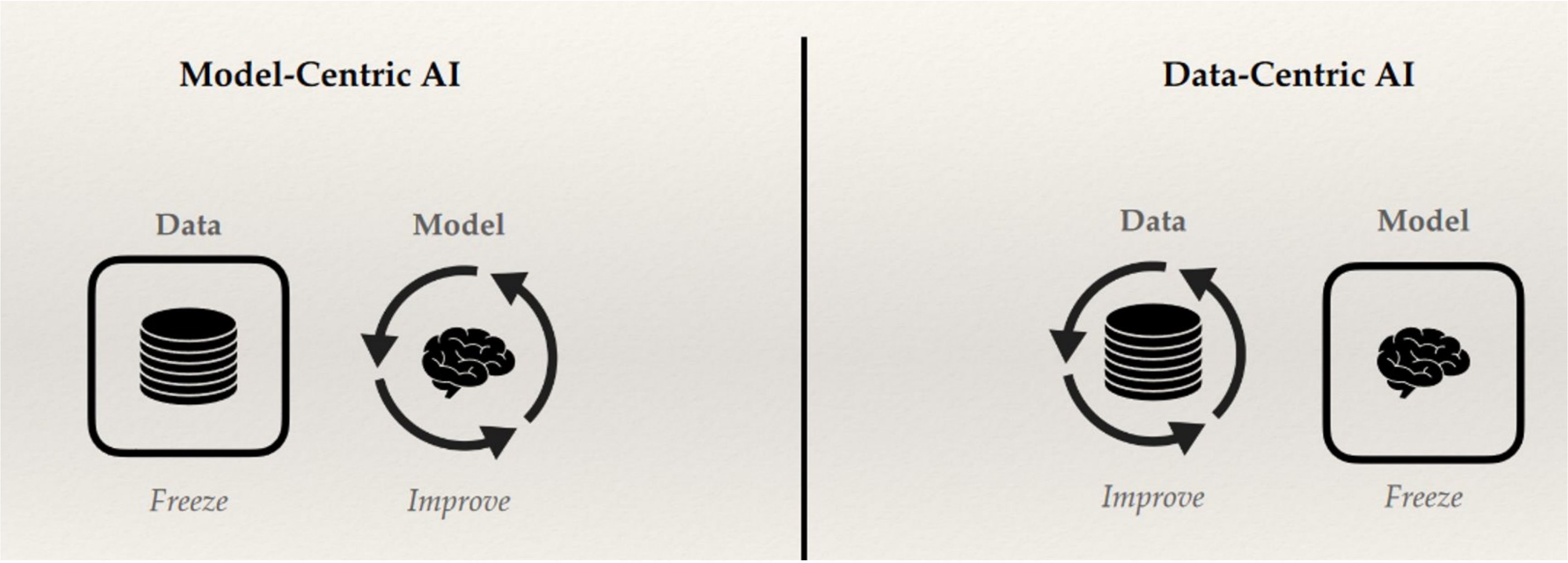
Intuition



- Less is More
- Small high quality textbook like data

Background

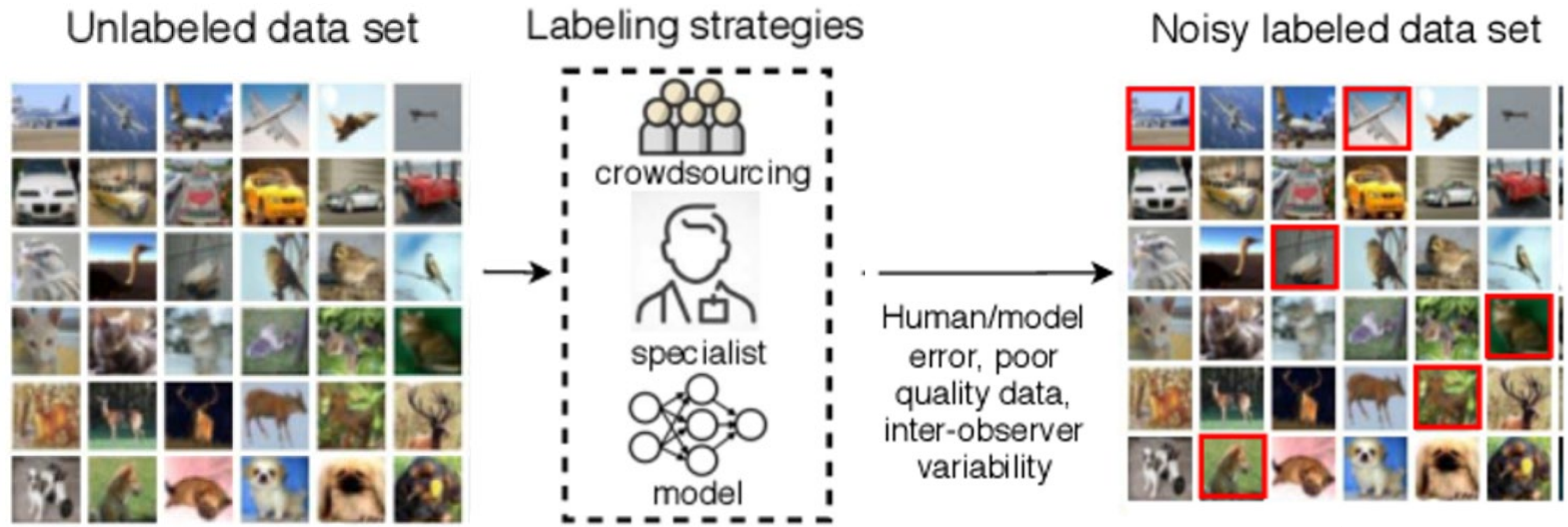
Data centric AI



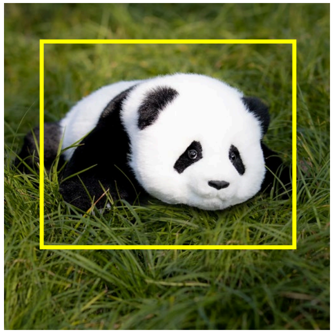
Sachdeva, Noveen, et al. "Infinite recommendation networks: A data-centric approach." Neurips’22

Motivation

Data are noisy



Cat

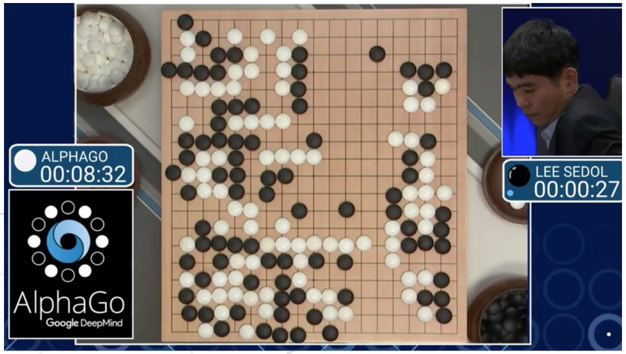


mislablled

2020, A Survey on Deep Learning with Noisy Labels: How to train your model when you cannot trust on the annotations?

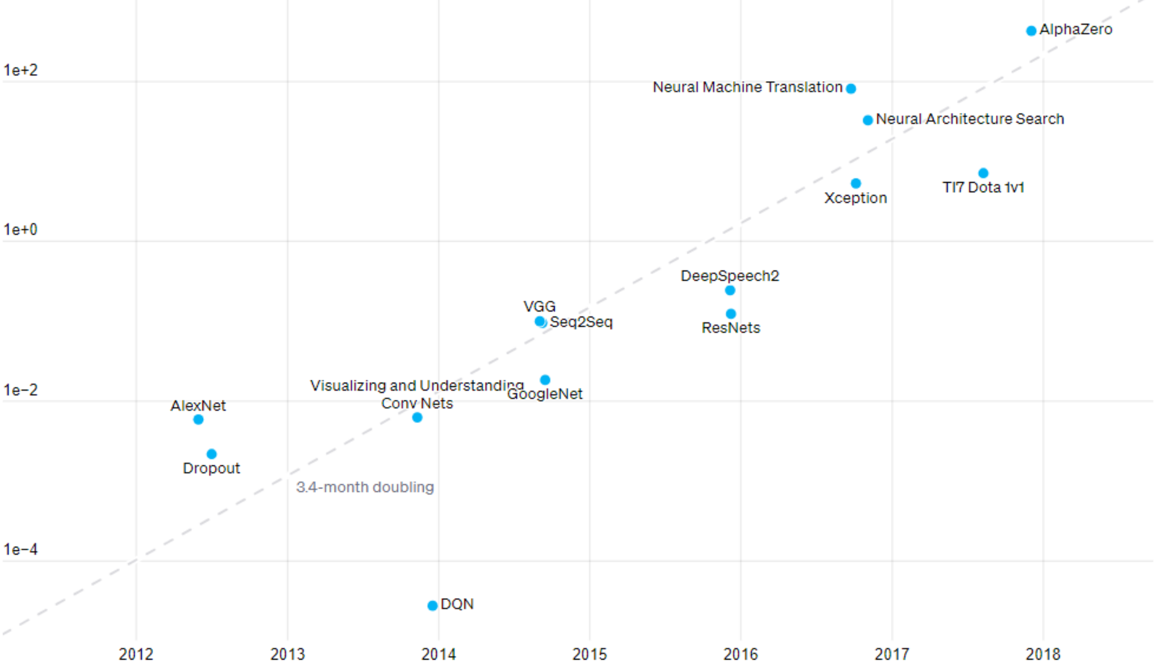
Motivation

Big data cost



AlexNet to AlphaGo Zero: 300,000x increase in compute

Log scale Linear Scale

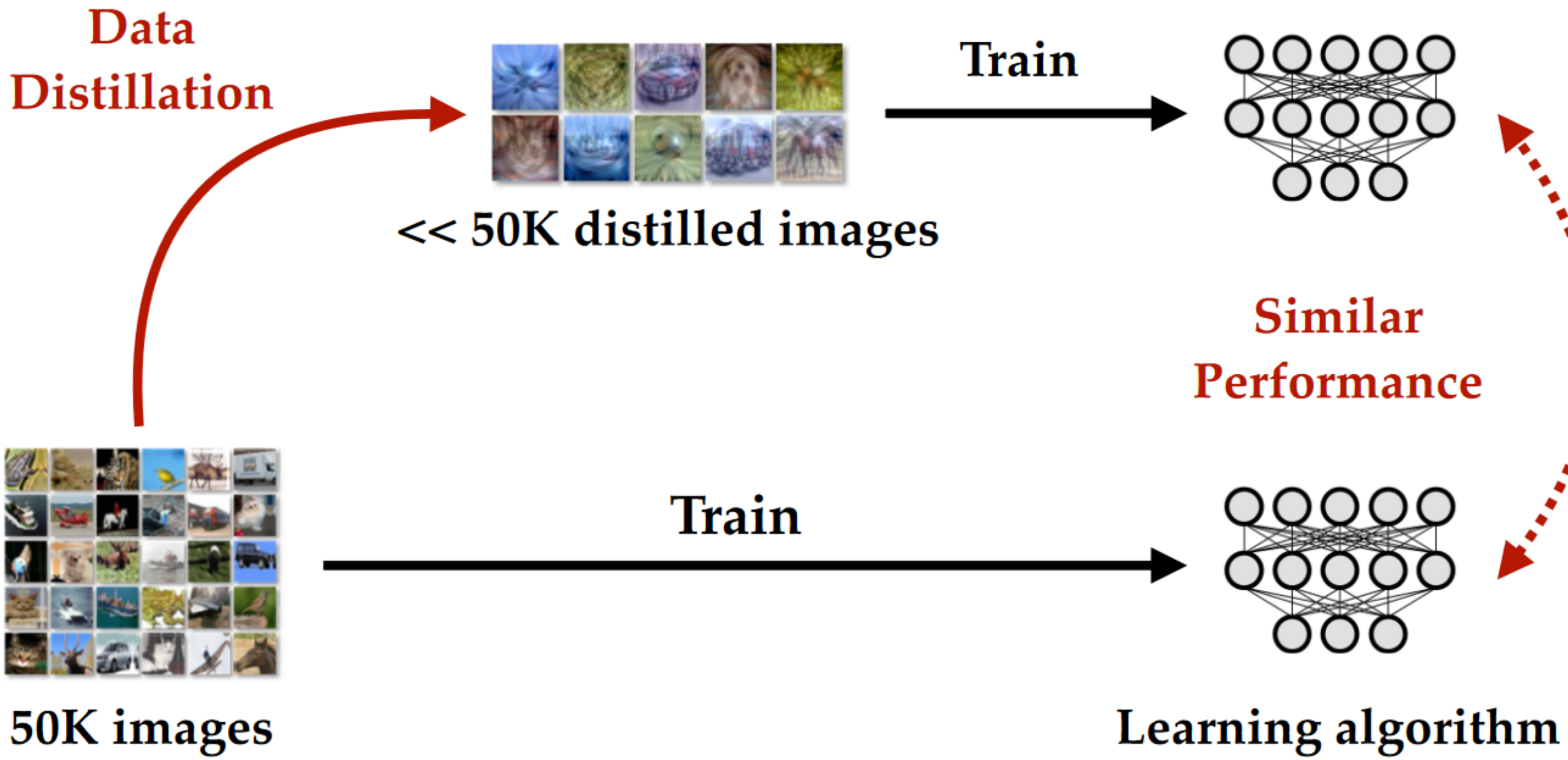


Predicted: Husky
True: Husky

OpenAI, AI and compute, 2018

Research Question

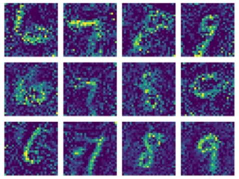
- Reduce data size while retaining the model's performance?



Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." TMLR'23

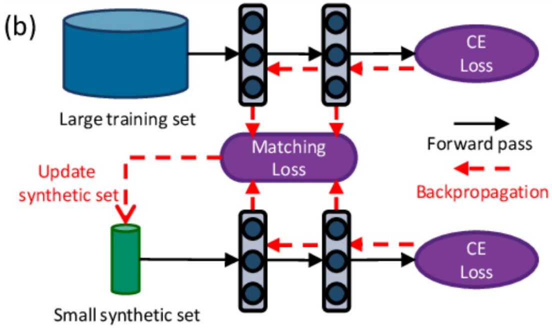
Literature Review

MNIST



2018 DD

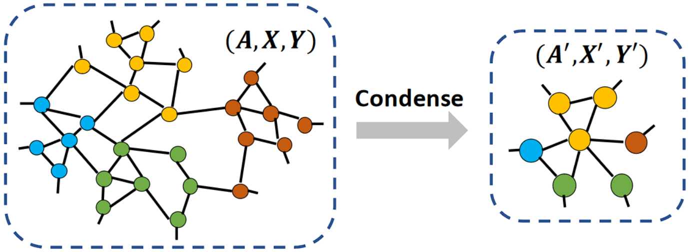
2020 DC



$$S_c \leftarrow S_c - \lambda \nabla_{S_c} D\left(\mathcal{A}\left(S_c\right), \mathcal{A}\left(T_c\right), \theta\right)$$

2021~2022
Different match objective function

2022
Different data modality, application (privacy, med)



153,932 training nodes

154 training nodes

a black and white dog is running through the field to catch something in its mouth

a black and white dog leaps to catch a frisbee in a field

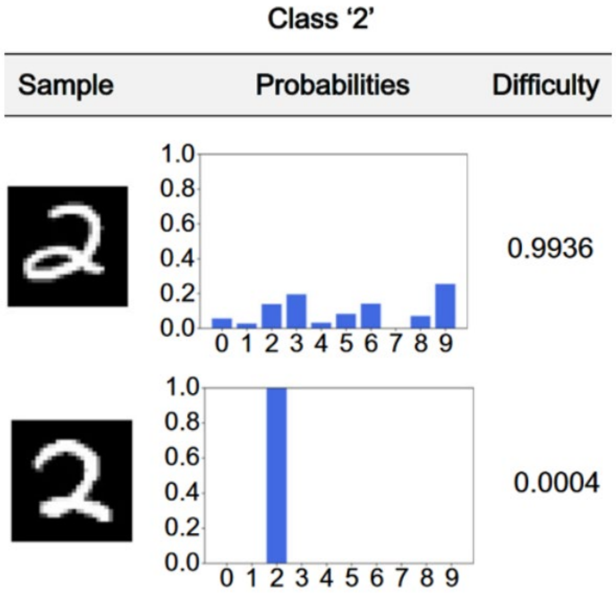
2023
Large scale dataset, LLM, generative, multimodal

Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." TMLR'23

Critical Literature Review 1

Herustics data selection

Select hard samples

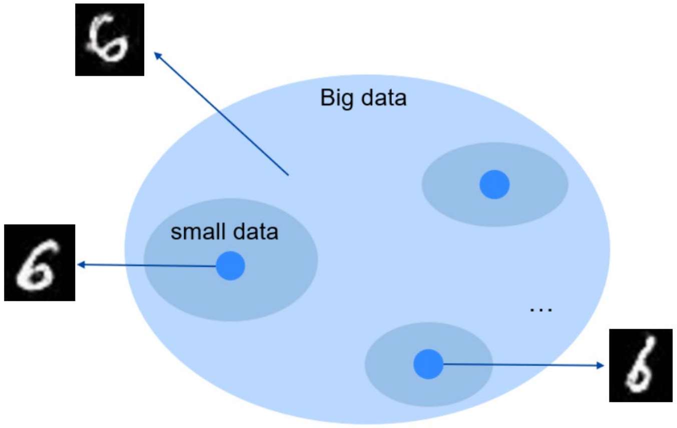


Class ‘2’

Sample	Probabilities	Difficulty
		0.9936
		0.0004

1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9

Cluster sampling



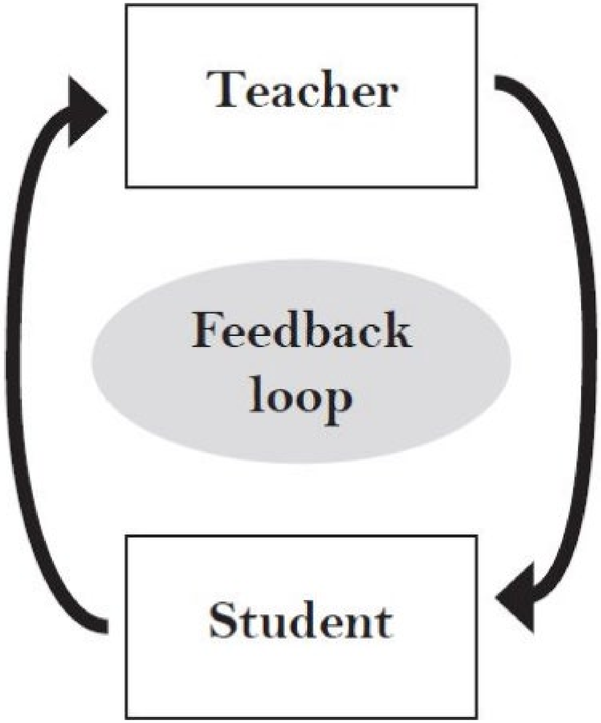
Fast, but may not be accurate

Jeong, Y et.al. (2023). Training data selection based on dataset distillation for rapid deployment in machine-learning workflows.

Critical Literature Review 2

Feedback optimization framework

Feedback



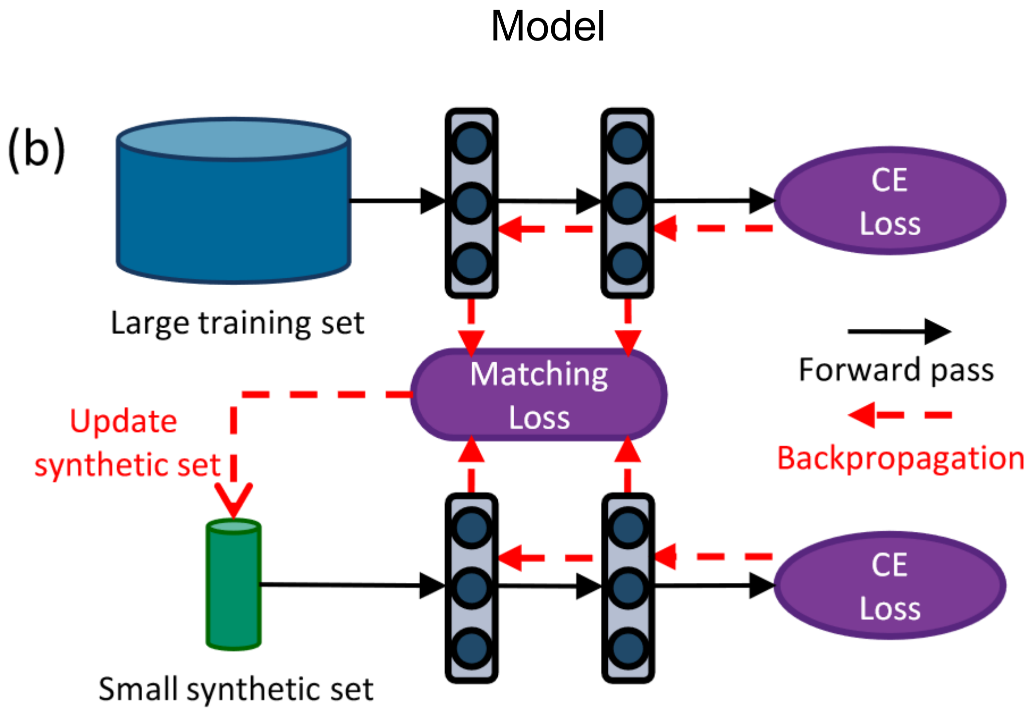
Select data

Accurate, but slow

Wang, T., et al. (2018). Dataset distillation. *arXiv preprint*

Critical Literature Review 2

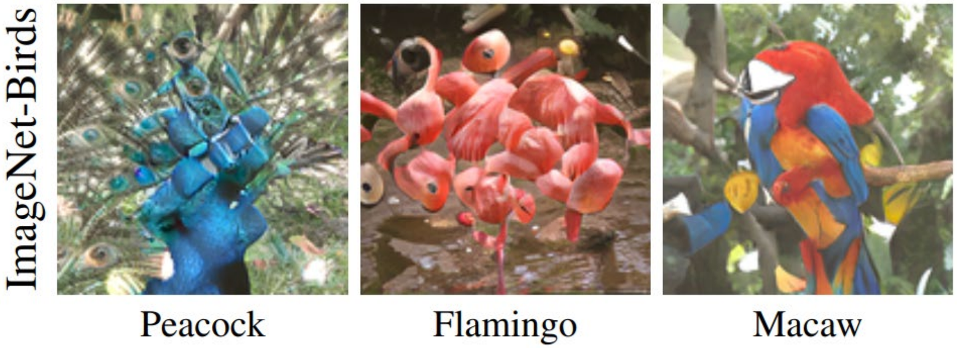
Feedback optimization framework

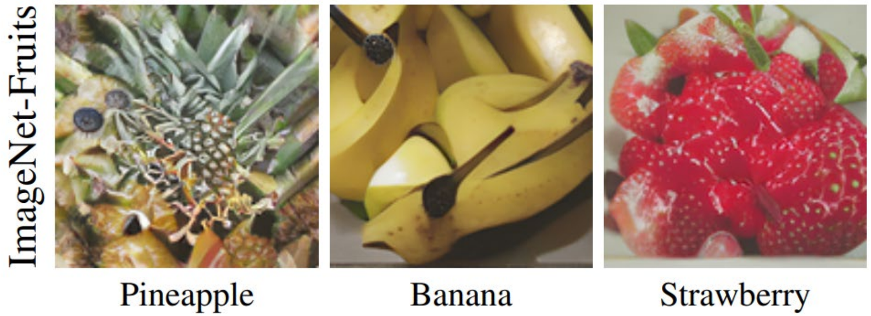


Wang, T., et al. (2018). Dataset distillation. *arXiv preprint*

Critical Literature Review 2

Generative DD





DD

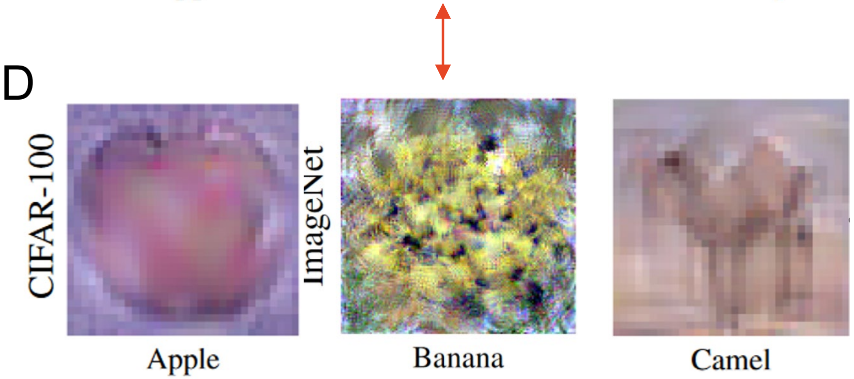
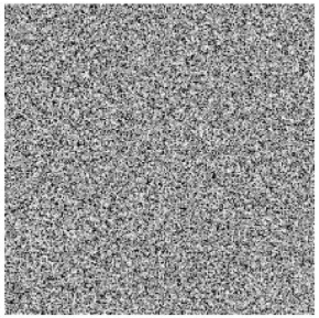


Image-text

a black and white dog is running through the field to catch something in its mouth

a black and white dog leaps to catch a frisbee in a field

Medical, privacy



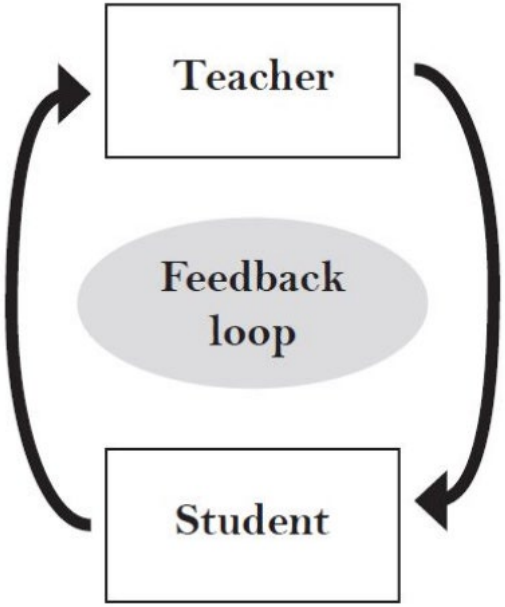
Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." TMLR’23

Research gap

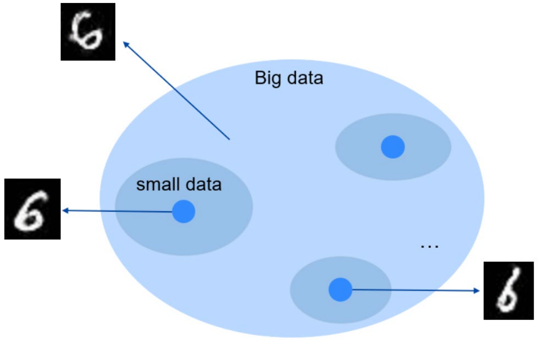
- Herustics methods are fast but not accurate
- Optimization methods are accurate but slow

Research Method

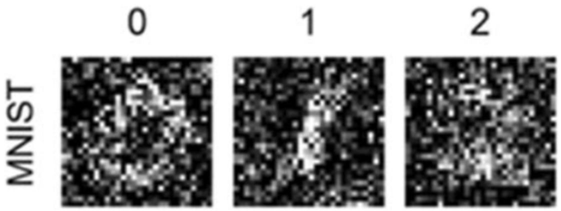
Heuristic selection (fast, not accurate) + Feedback (accurate, but slow)



1. Select data



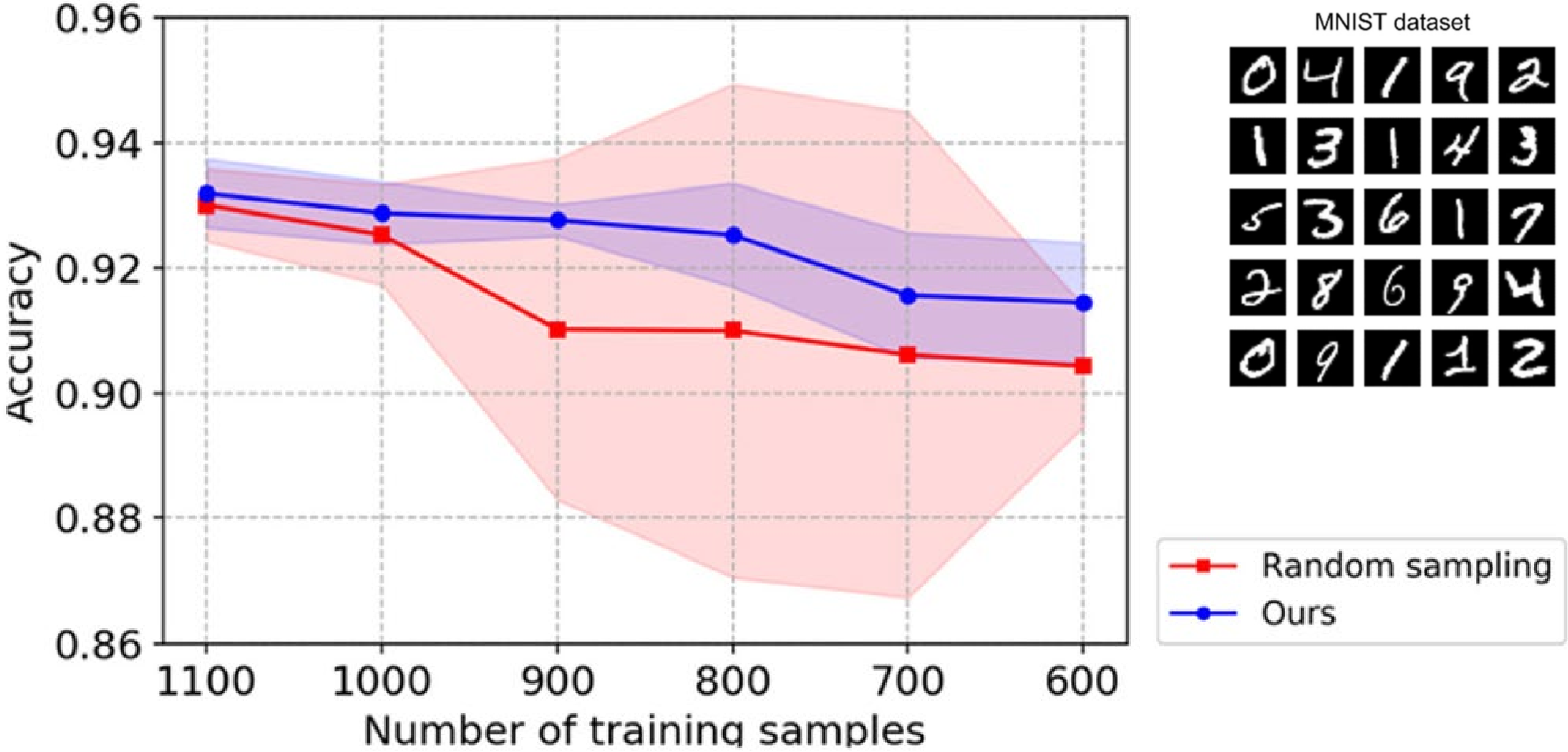
2. Optimize with feedback



- Get student feedback on representative data not all

Experiment

Perform better w/ lower variance

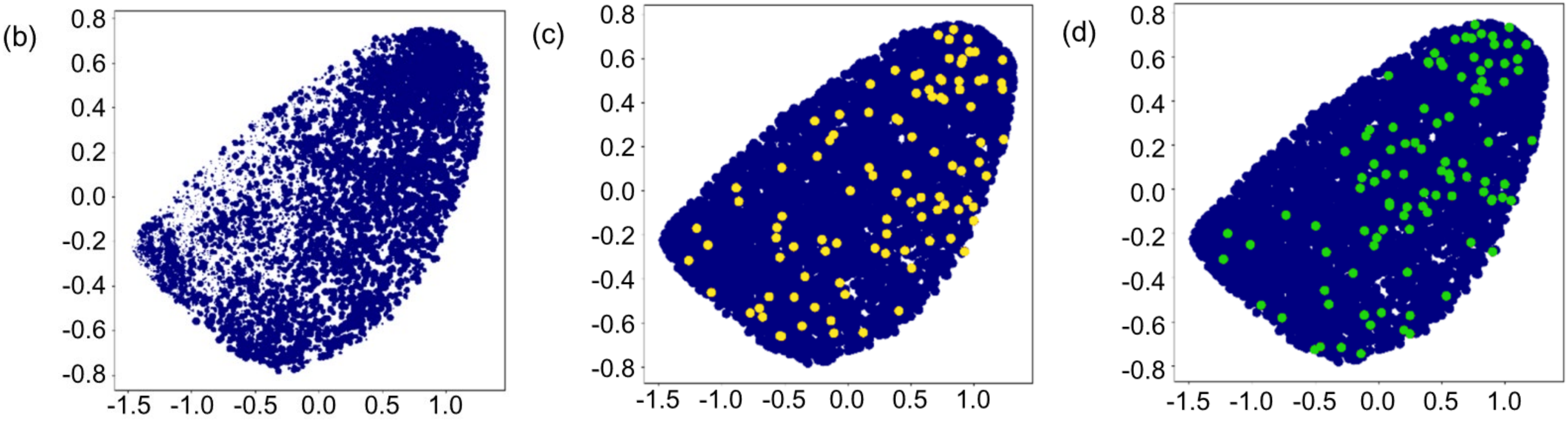


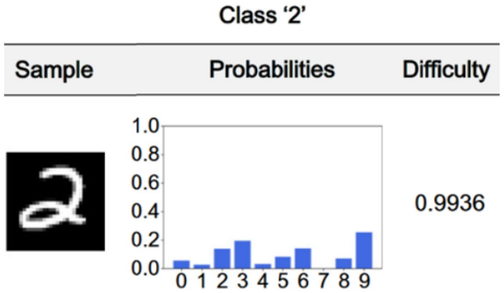
Experiment

b) Varying marker size w.r.t sample difficulty

c) Proposed method

d) Baselines

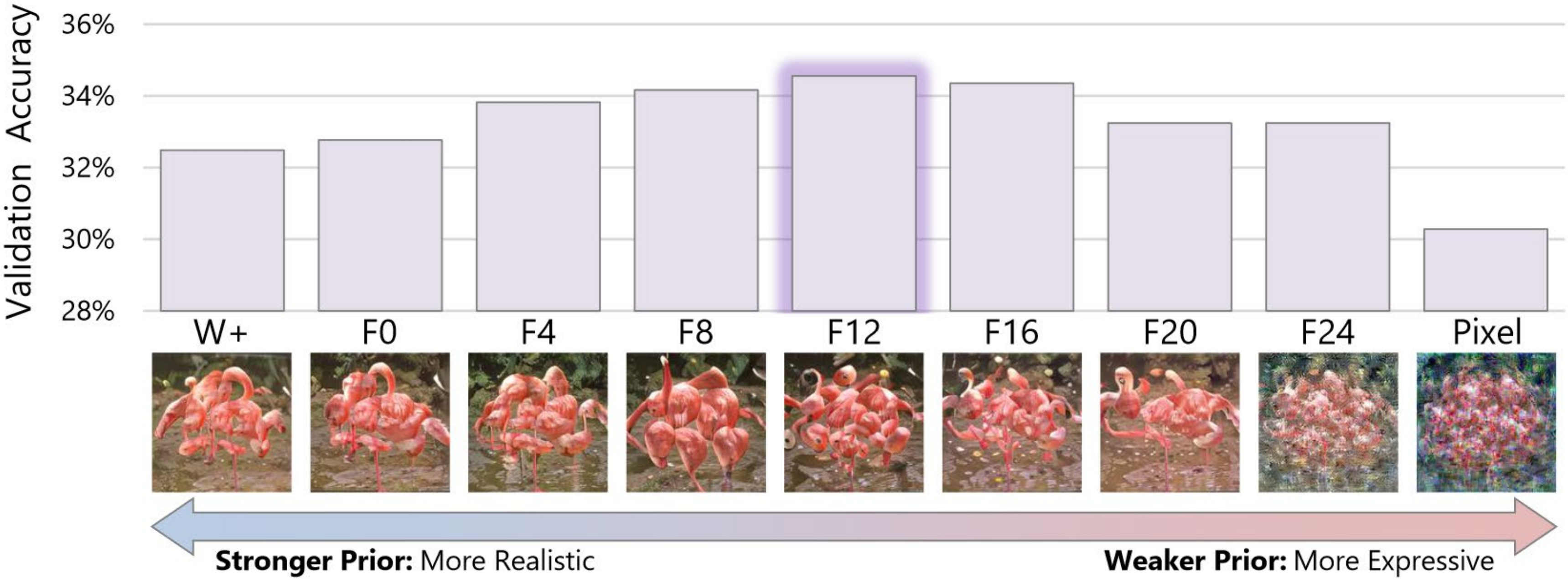



1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9

Class '2'

Sample	Probabilities	Difficulty
		0.9936

Limitation



Trade-of between realism & expressiveness (performance)

Cazenavette, George, et al. "Generalizing dataset distillation via deep generative prior." CVPR'23.

Conclusion

- Motivation
  - Big data cost
- RQ
  - Reduce big data to small while retaining performance
- Prev solution
  - Either slow or not accurate
- Proposed method
  - Heuristcs (fast) + feedback framework (accurate)

Thank you for listening!

Any questions?

Small data

Original big data