NATIONAL UNIVERSITY OF
SCIENCES AND TECHNOLOGY

INCREMENTAL CLASSIFIER AND REPRESENTATION LEARNING

FINAL DEFENSE

Talha Paracha, Khurram Javed

Co-Advisor: Dr Muhammad Shehzad
Advisor : Dr Faisal Shafait

TABLE OF CONTENTS

TABLE OF CONTENTS

PROBLEM STATEMENT

Adopt machine learning algorithms to learn representation and classier incrementally without storing all the previous data.

SIMPLE EXAMPLE

This is Tesla Model X

SIMPLE EXAMPLE

This is Tesla Model X

What car is this?

Catastrophic Forgetting on CIFAR100

Actual class

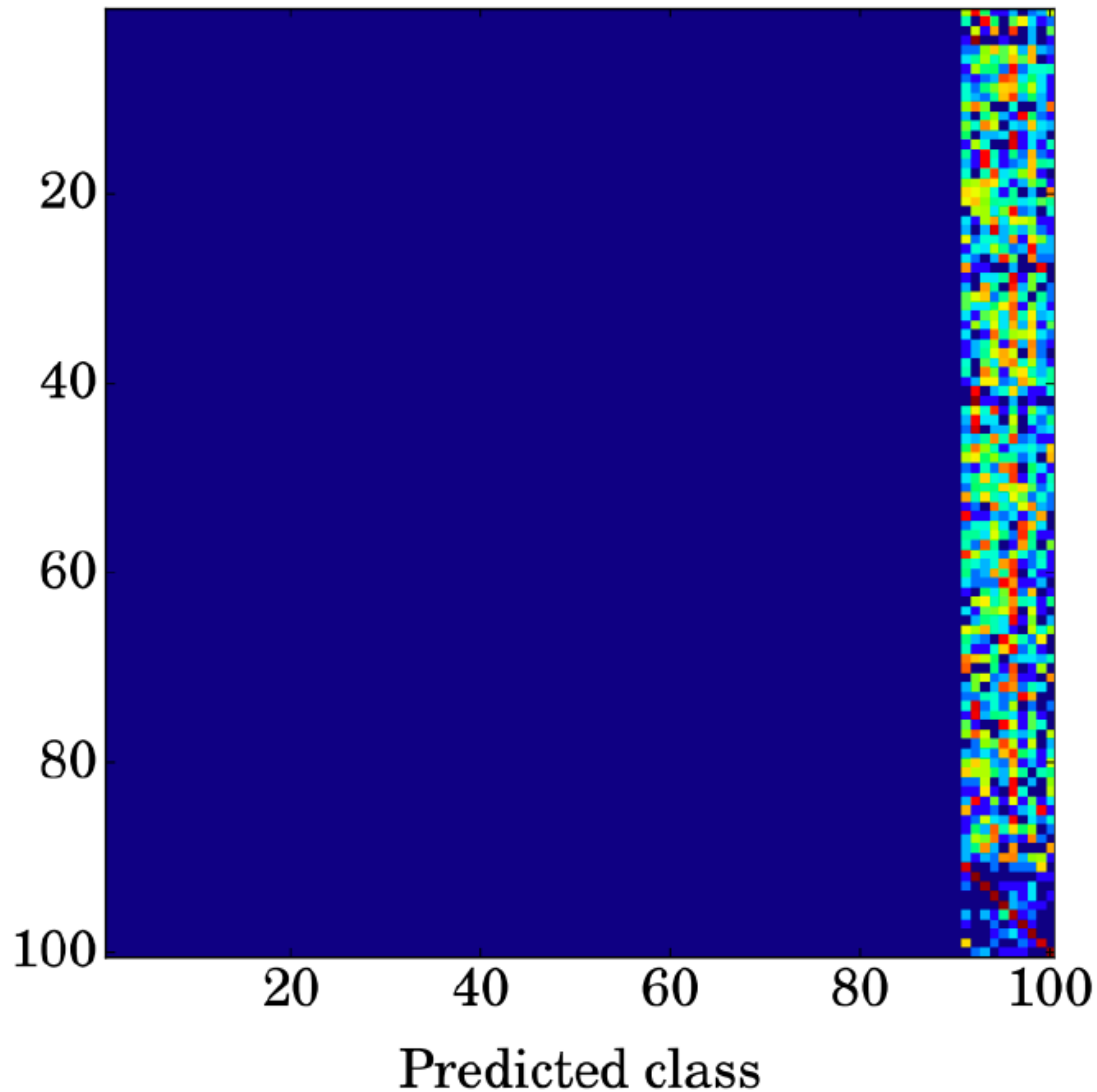


TABLE OF CONTENTS

EXISTING METHODS

1. iCaRL [1]

2. GAN based incremental learning.

3. Distilling knowledge in neural networks.

[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 2001–2010, 2017

CURRENT SOTA

- iCaRL [1]
- Three main components:
  - Nearest Mean Classifier for classification
  - Knowledge distillation
  - Instance selection using herding

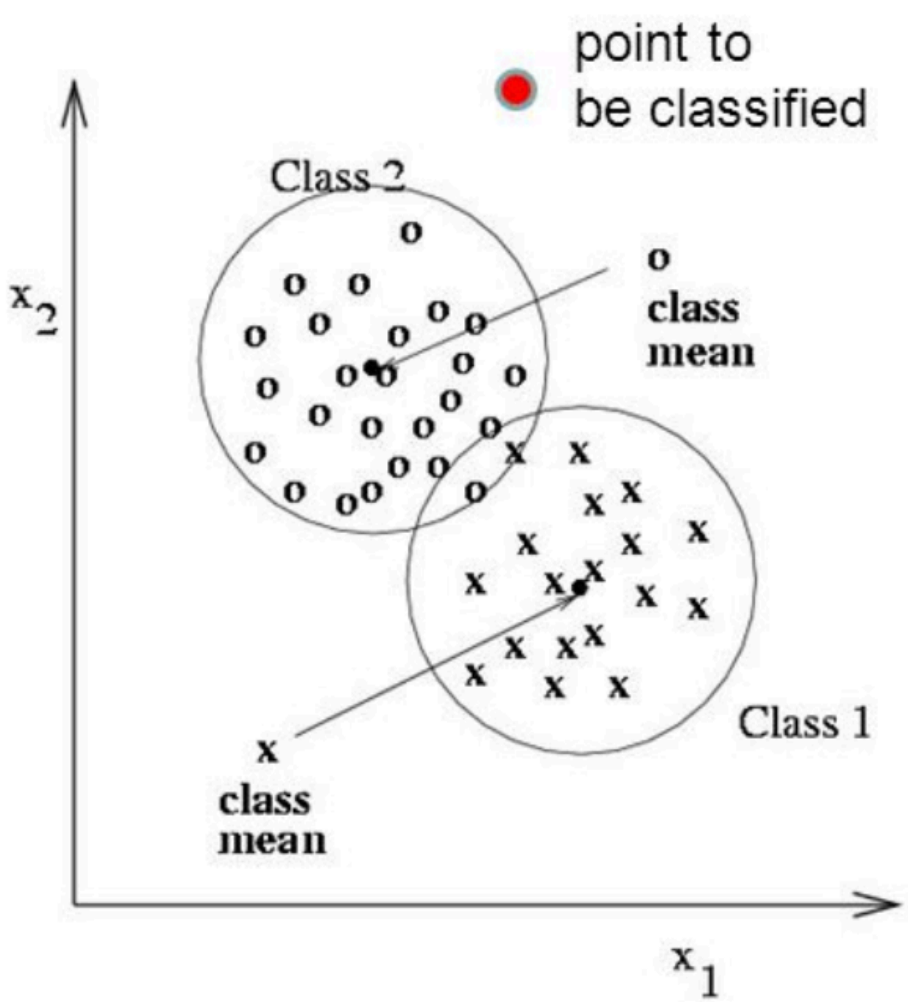
[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 2001–2010, 2017

CURRENT SOTA

- iCaRL [1]
- Three main components:
  - <u>**Nearest Mean Classifier for classification**</u>
  - Knowledge distillation
  - Instance selection using herding

[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 2001–2010, 2017

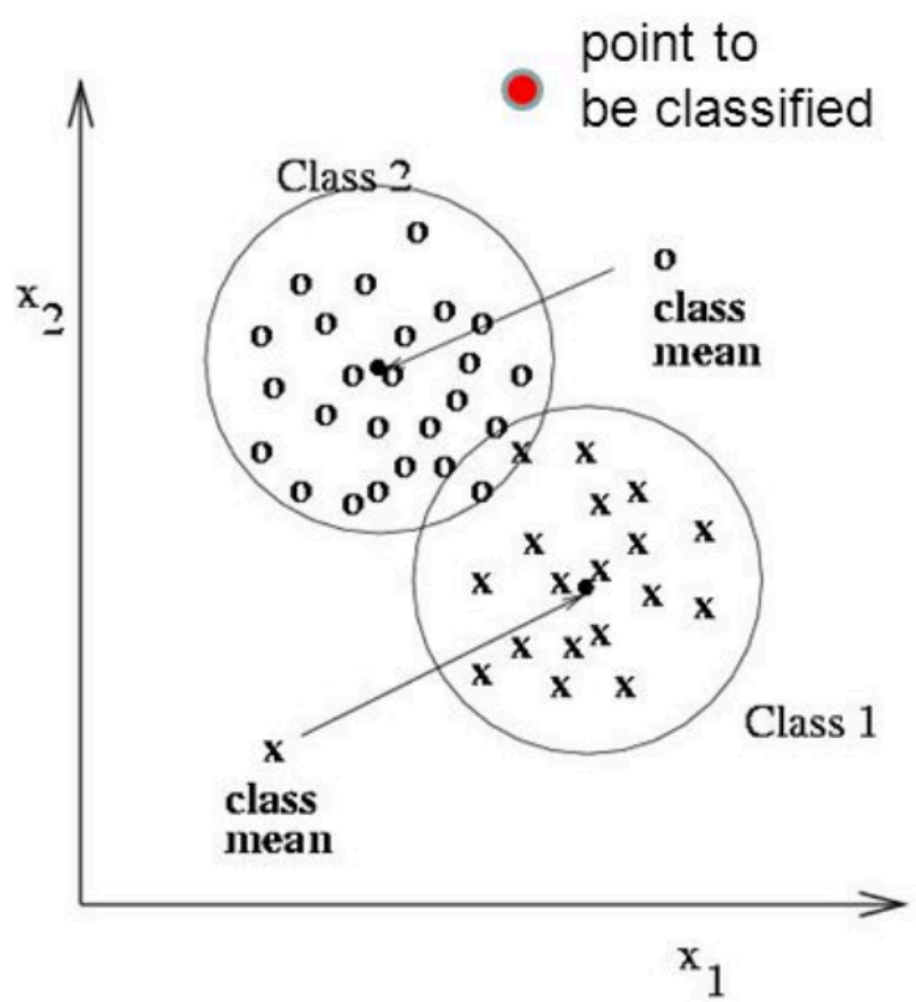
NEAREST MEAN CLASSIFIER



- Compute distance from mean of each class.
- Assign label of class with the smallest distance.

http://slideplayer.com/slide/4943010/

NEAREST MEAN CLASSIFIER

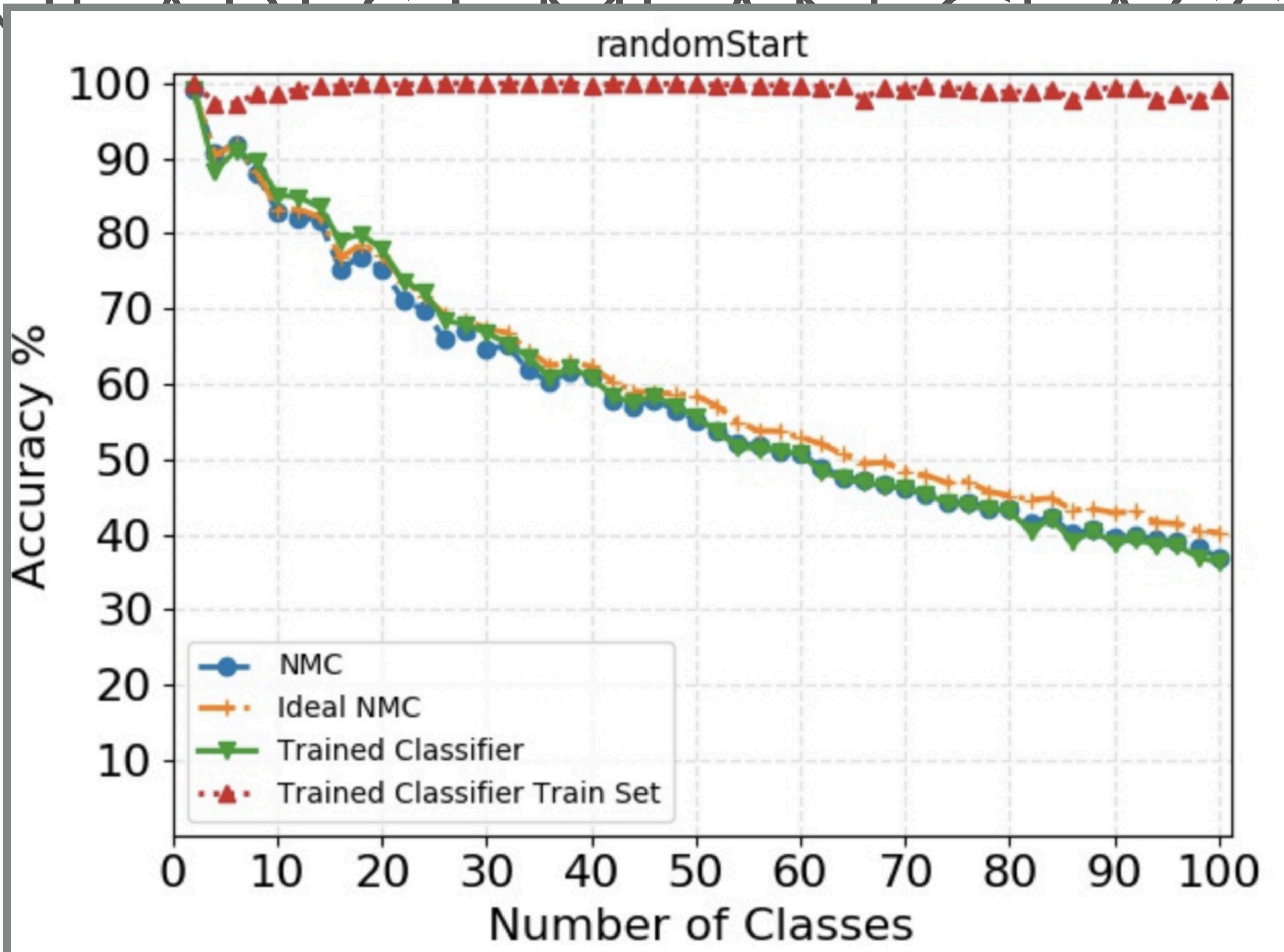


- Compute distance from mean of each class.
- Assign label of class with the smallest distance.

iCaRL claims NMC is better than Softmax classifier

http://slideplayer.com/slide/4943010/

NEAREST MEAN CLASSIFIER



ce
ach

lass
t

MC is
tmax
classifier

http://slideplayer.com/slide/4943010/

CURRENT SOTA

- iCaRL [1]
- Three main components:
    - Nearest Mean Classifier for classification
    - **Knowledge distillation**
    - Instance selection using herding

[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 2001–2010, 2017

CURRENT SOTA

- iCaRL [1]
- Three main components:
  - Nearest Mean Classifier for classification
  - Knowledge distillation
  - **Instance selection using herding**

[1] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pages 2001–2010, 2017

TABLE OF CONTENTS

PROPOSED METHODOLOGIES AND RESULTS

- Real-time computation for threshold moving.
- Conditional Generative Adversarial Networks.
- Privacy Preserving Incremental Learning.

PROPOSED METHODOLOGIES AND RESULTS

- **Real-time computation for threshold moving.**
- Conditional Generative Adversarial Networks.
- Privacy Preserving Incremental Learning.

REAL-TIME COMPUTATION FOR THRESHOLD MOVING.

- Instead of using NMC, scale the Softmax classifier by a vector to remove bias.

VECTOR COMPUTATION ALGORITHM

Let $F(X)$ be a trained classifier for N classes.

$$\implies \forall x_i \in X,\ F(x_i) \text{ gives a probability distribution } P(x_i \mid n) \text{ where } 0 <= n < N$$

Suppore we want to train G(X) on N+1 classes using data of only N+1th class and $F(X)$

Let y_i be ground truth of new class and $C^i_{soft} = F(x_i)$

$$Minimize\ (1-\gamma) \times C_{entropy}(G(x_i), y_i) + \gamma \times C_{entropy}(G(x_i), C^i_{soft})$$

$$S = \sum_{i=1}^{k} \gamma \times F(x_i) + (1-\gamma) \times y_i$$

VECTOR COMPUTATION ALGORITHM

Let $F(X)$ be a trained classifier for N classes.

$$\implies \forall x_i \in X,\ F(x_i) \text{ gives a probability distribution } P(x_i \mid n) \text{ where } 0 <= n < N$$

Suppore we want to train G(X) on N+1 classes using data of only N+1th class and $F(X)$

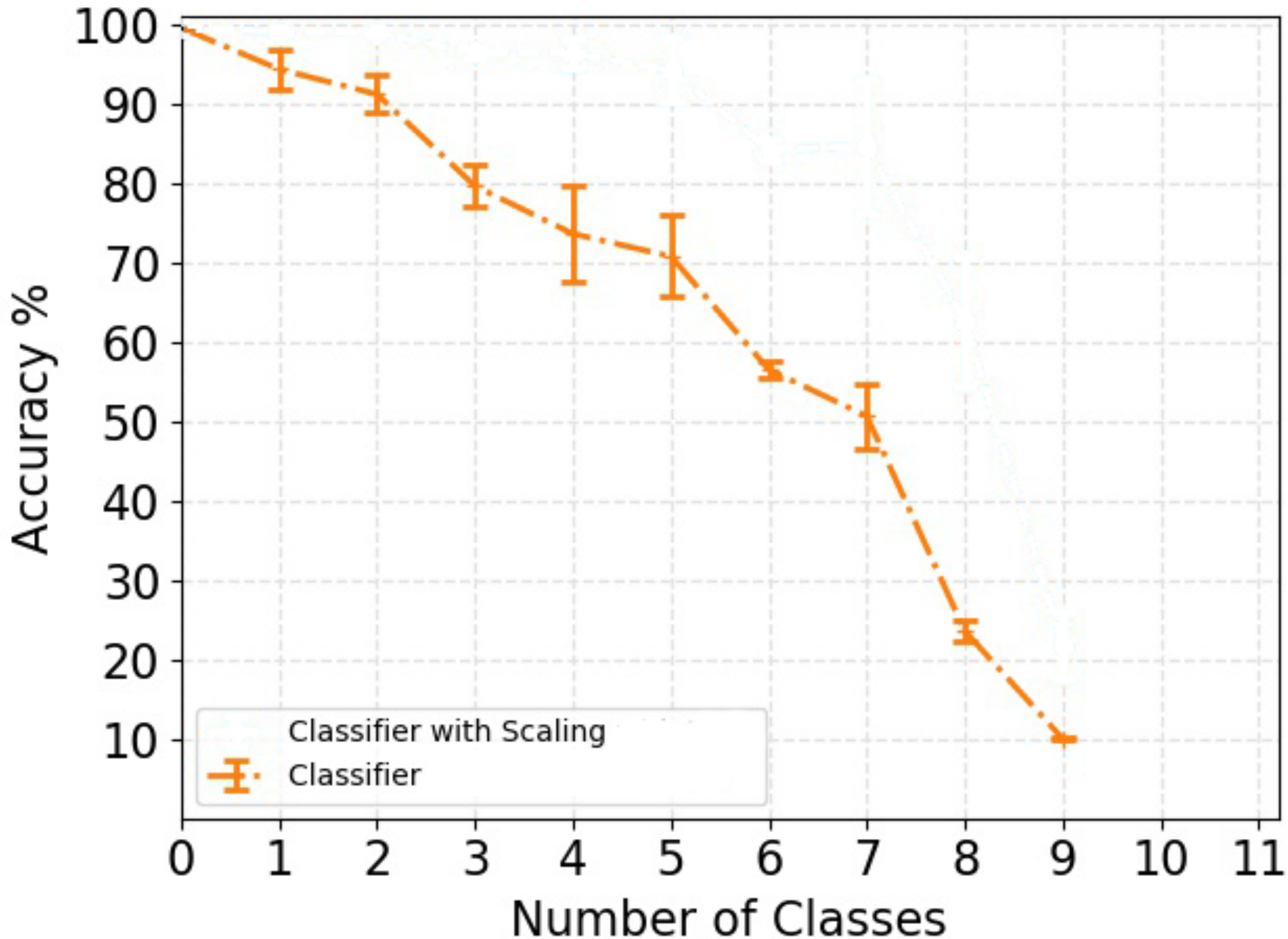
Let y_i be ground truth of new class and $C^i_{soft} = F(x_i)$

$$Minimize\ (1-\gamma) \times C_{entropy}(G(x_i), y_i) + \gamma \times C_{entropy}(G(x_i), C^i_{soft})$$

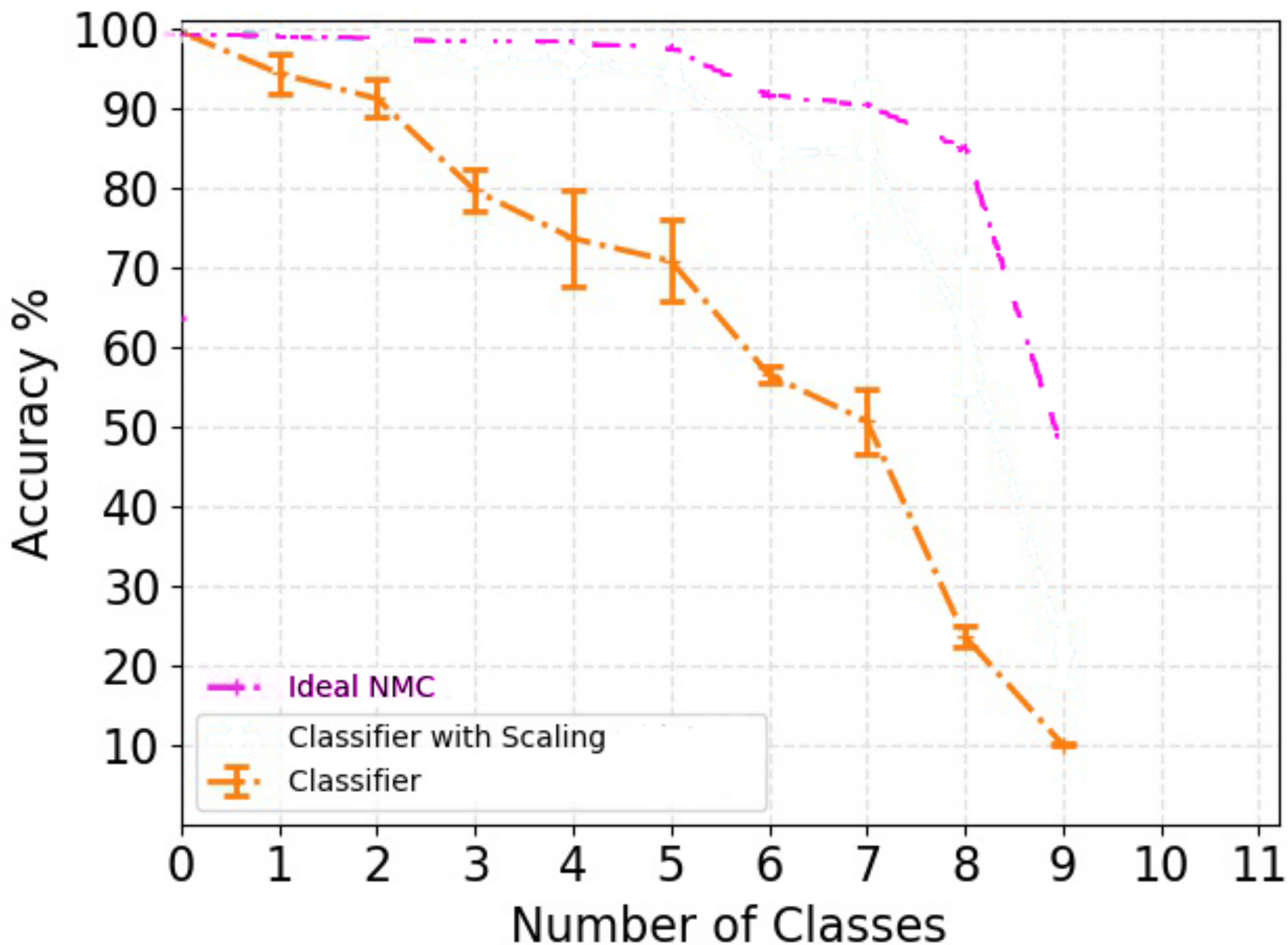
$$\boxed{S = \sum_{i=1}^{k} \gamma \times F(x_i) + (1-\gamma) \times y_i}$$

Scale factor

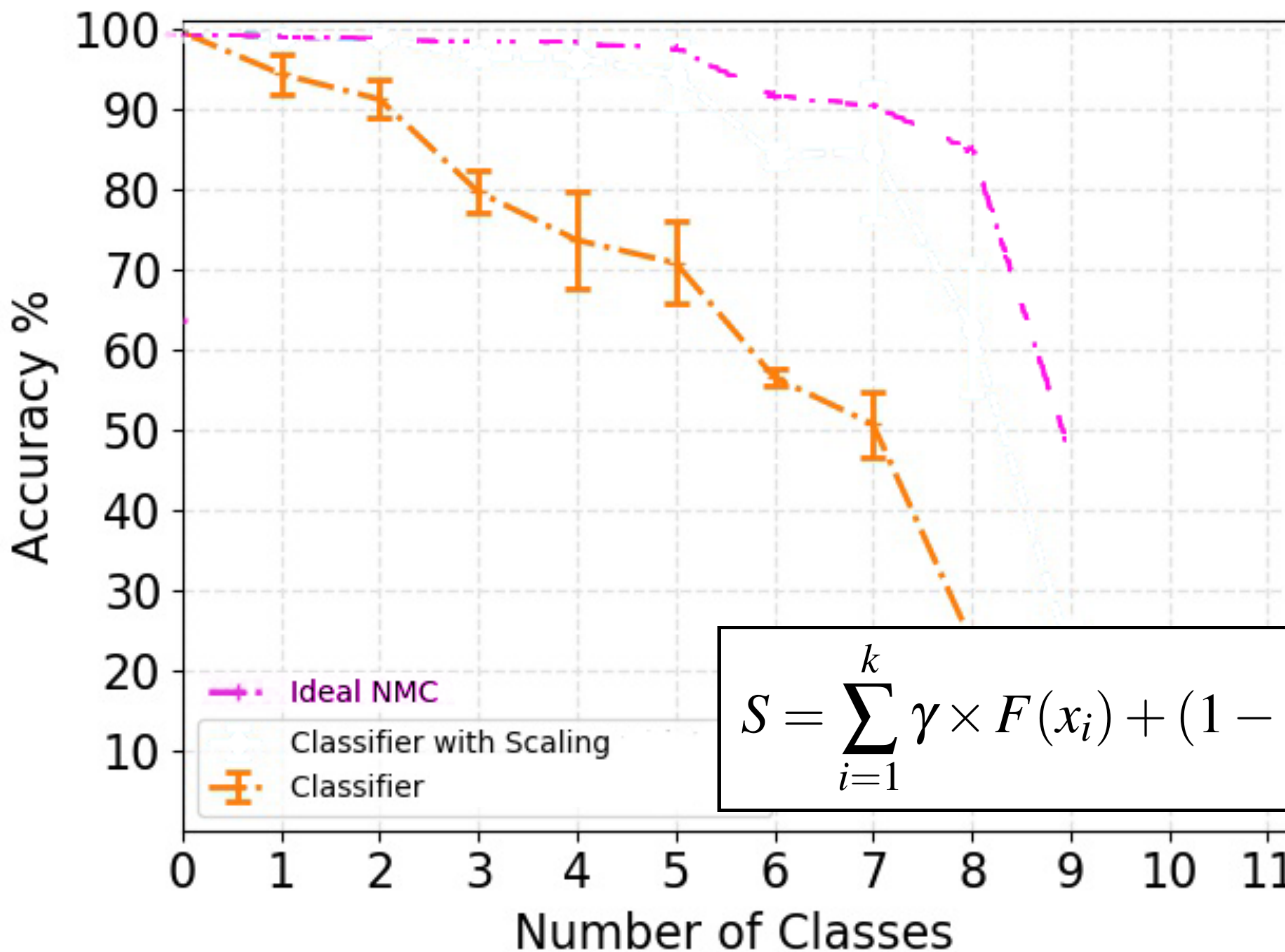
Threshold Moving on MNIST



Threshold Moving on MNIST

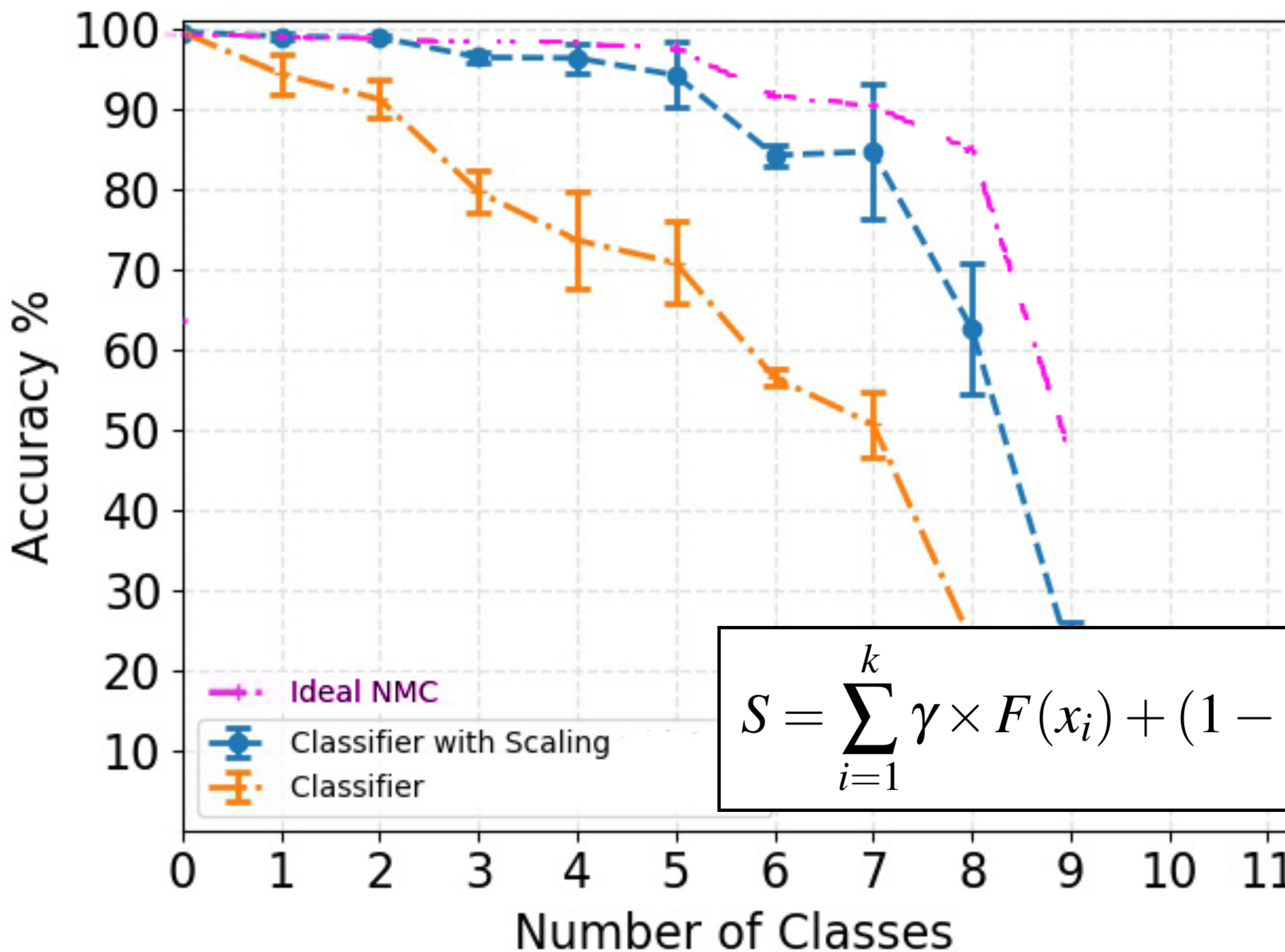


Threshold Moving on MNIST



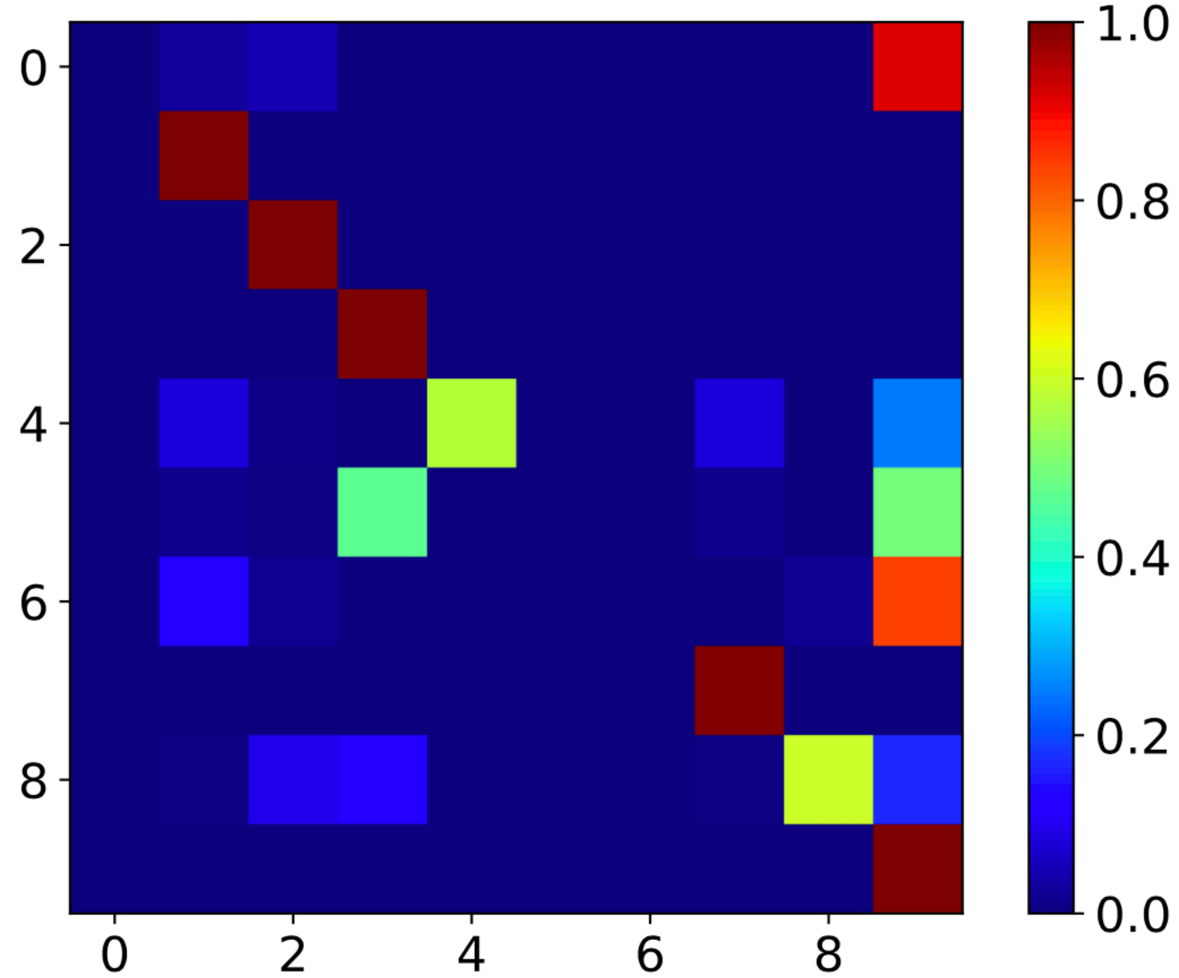
$$S = \sum_{i=1}^{k} \gamma \times F(x_i) + (1-\gamma) \times y_i$$

Threshold Moving on MNIST

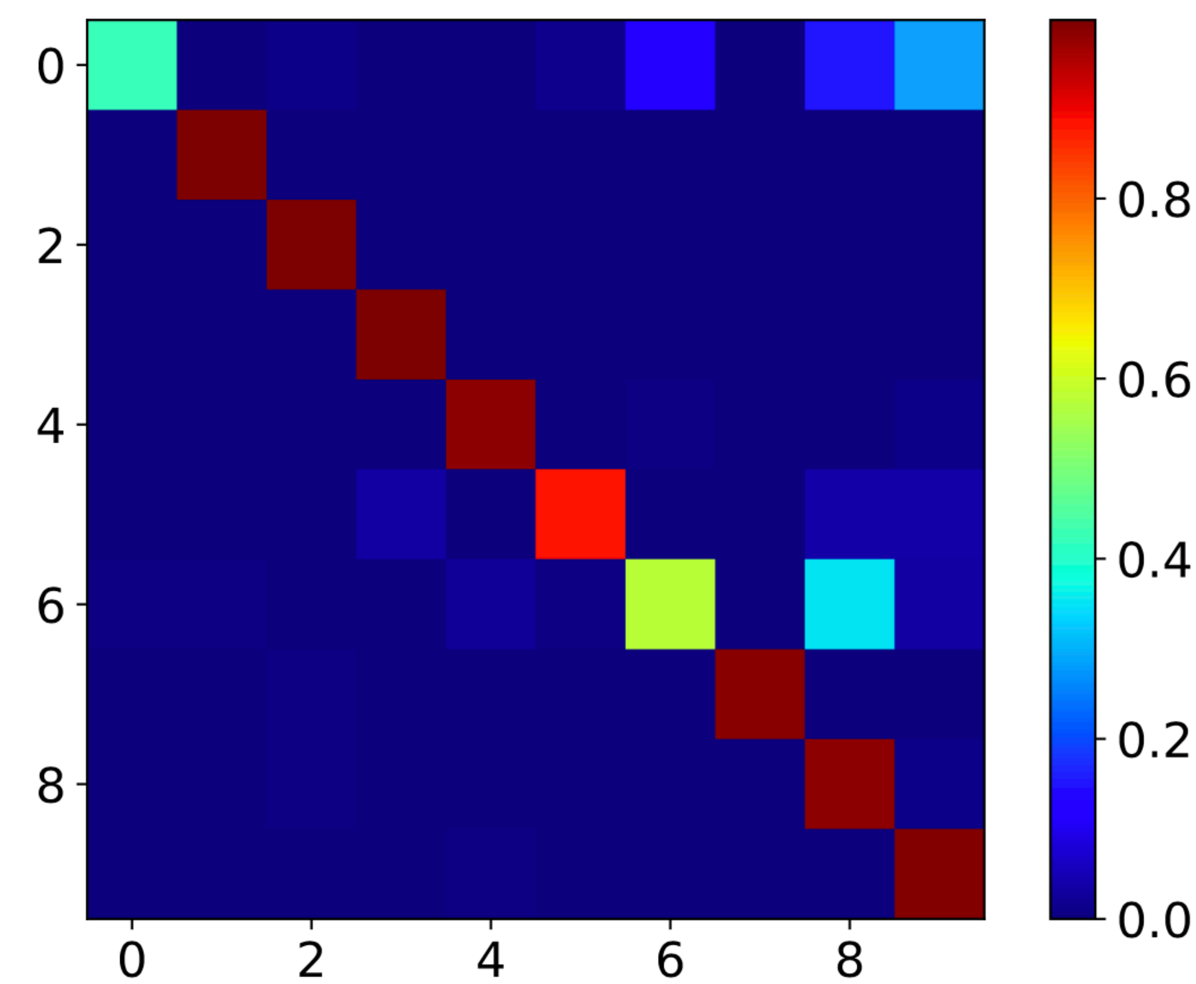


$$S = \sum_{i=1}^{k} \gamma \times F(x_i) + (1-\gamma) \times y_i$$

Threshold Moving



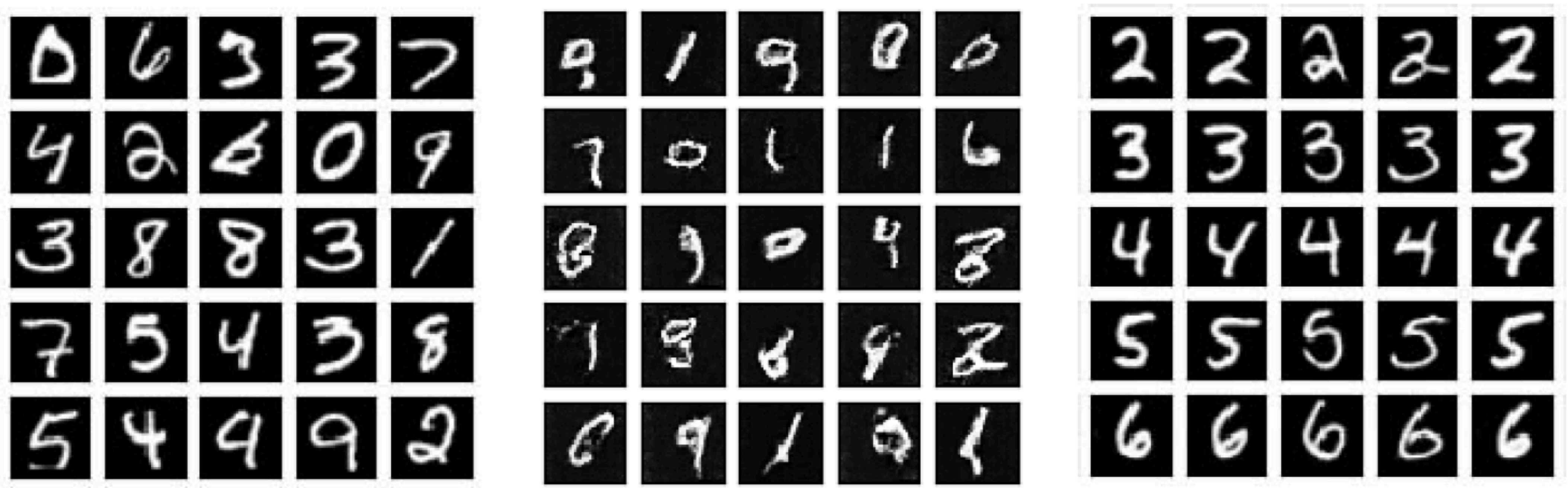
Softmax Classifier



Softmax Classifier Scaled

GENERATIVE ADVERSARIAL NETWORKS (GANs)

GANs

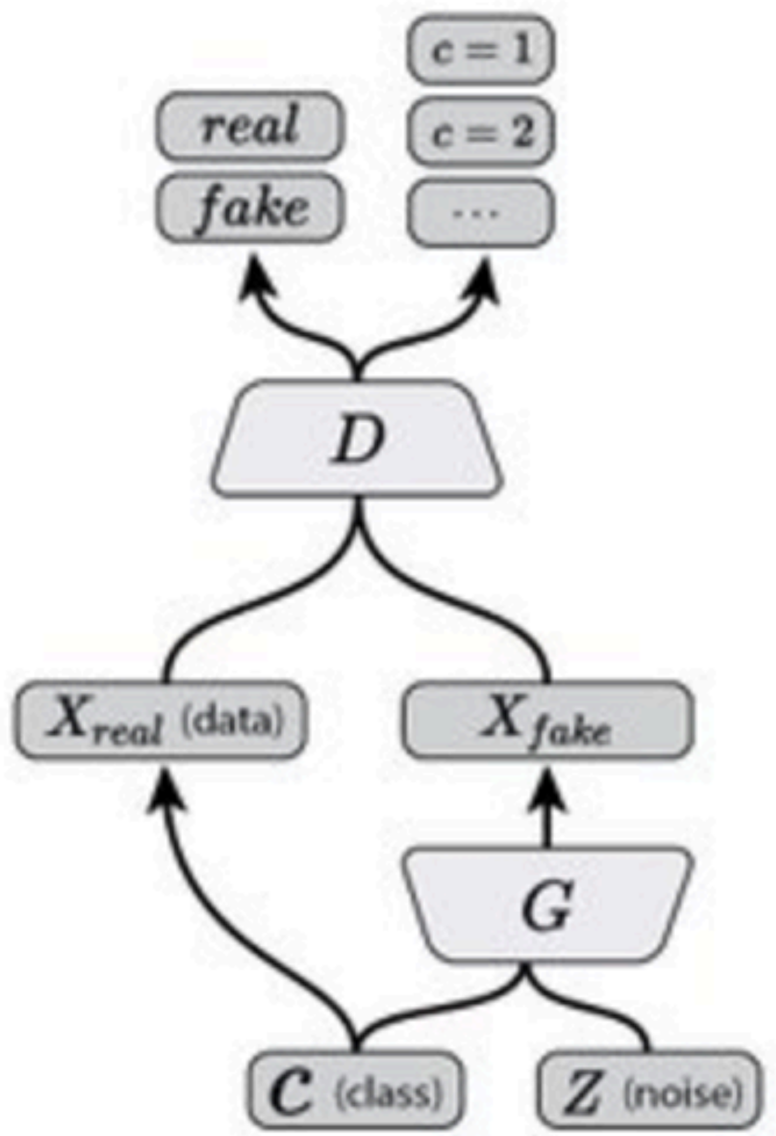


(a) Original Images (b) GAN (c) Cond-GAN

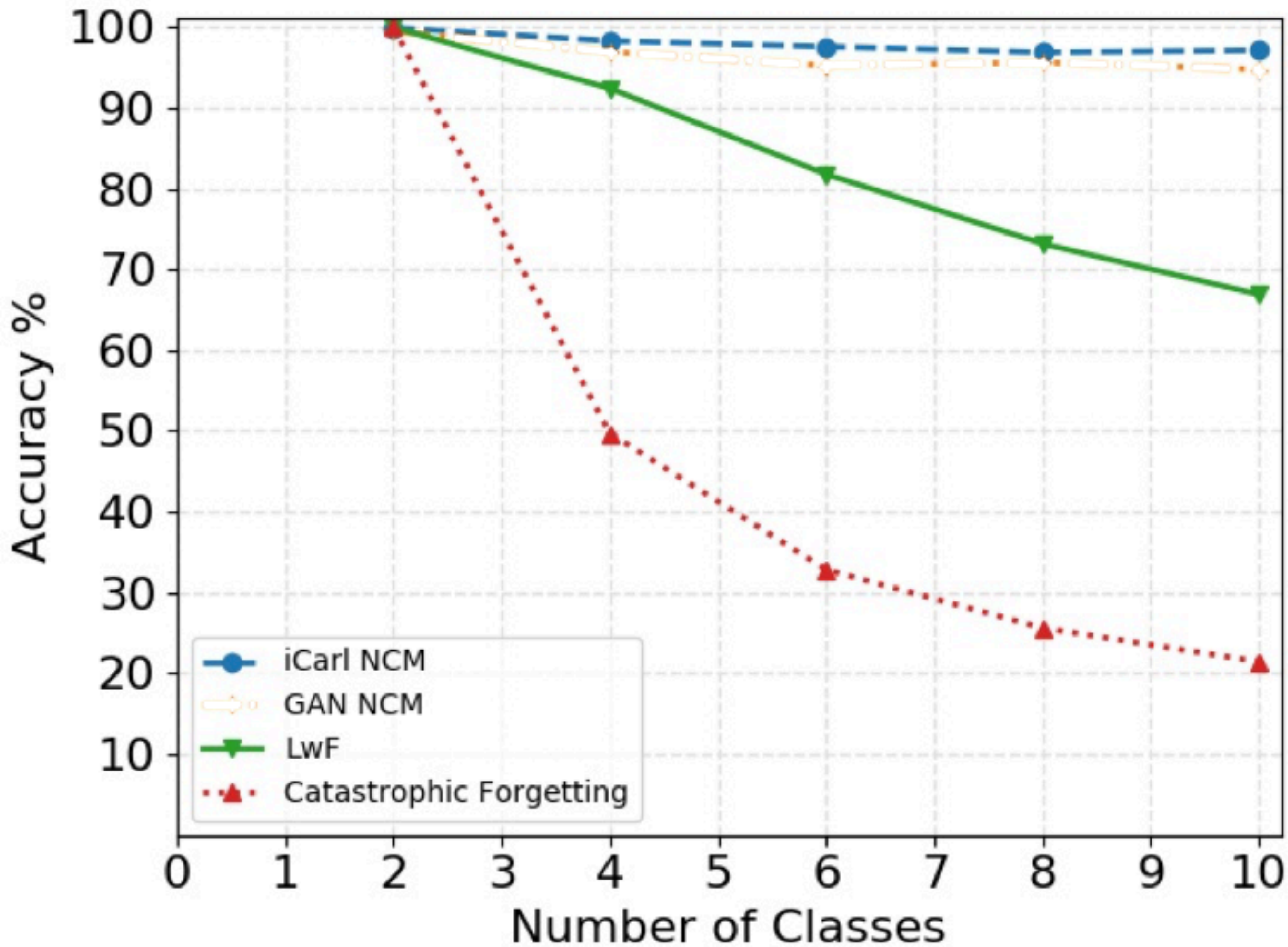
Fig. 1: Comparison of original images from MNIST (a) and images generated using generative adversarial networks (b and c). Compared to standard GANs, which learn the distribution of the whole dataset disregarding the labels (b), Conditional GANs learn the distribution conditioned to a class label. This allows them to generate more crisp images with ground truth.

Proposed Methodologies

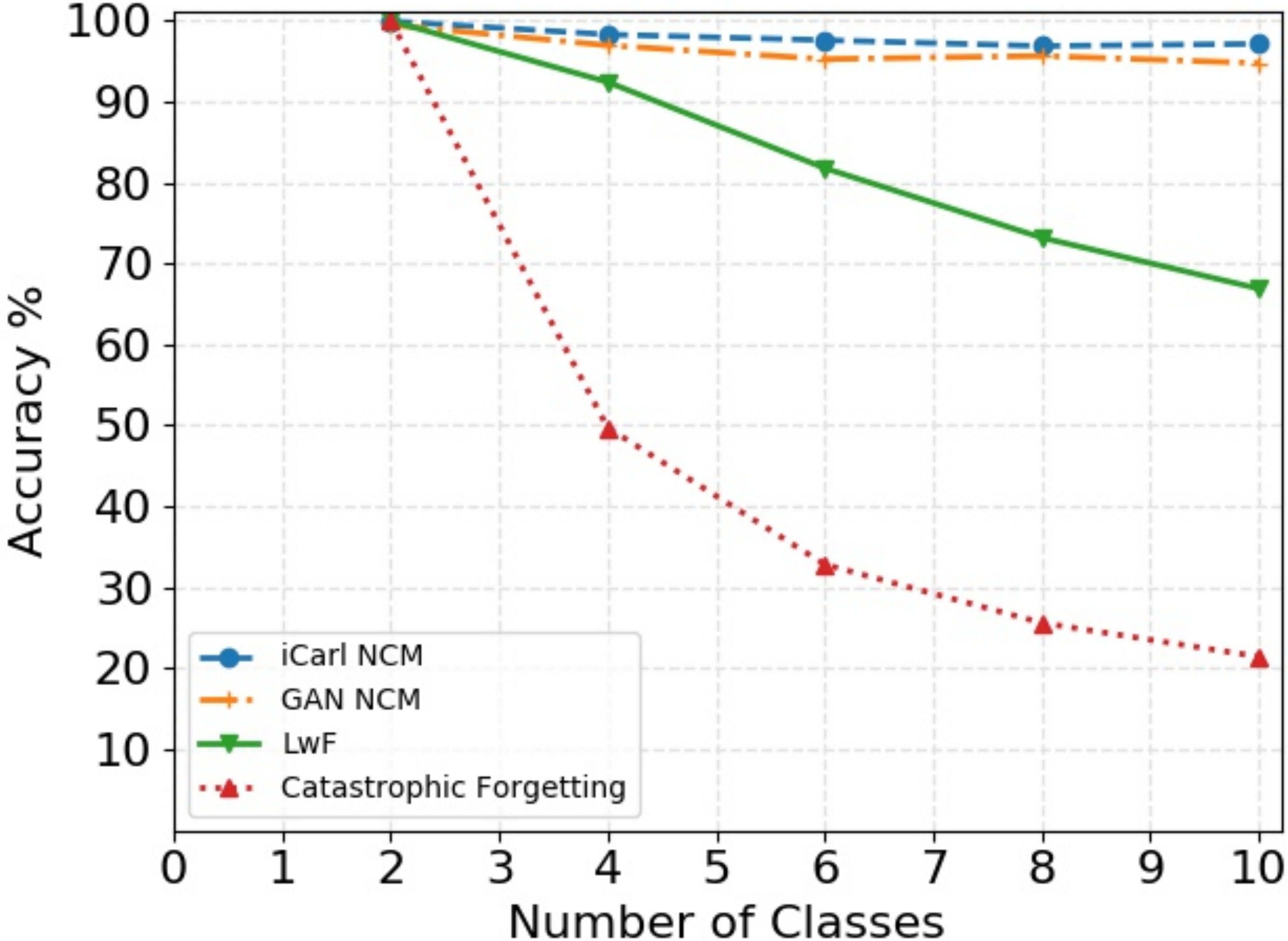
Auxiliary Classifier GAN
(Odena, et al., 2016)



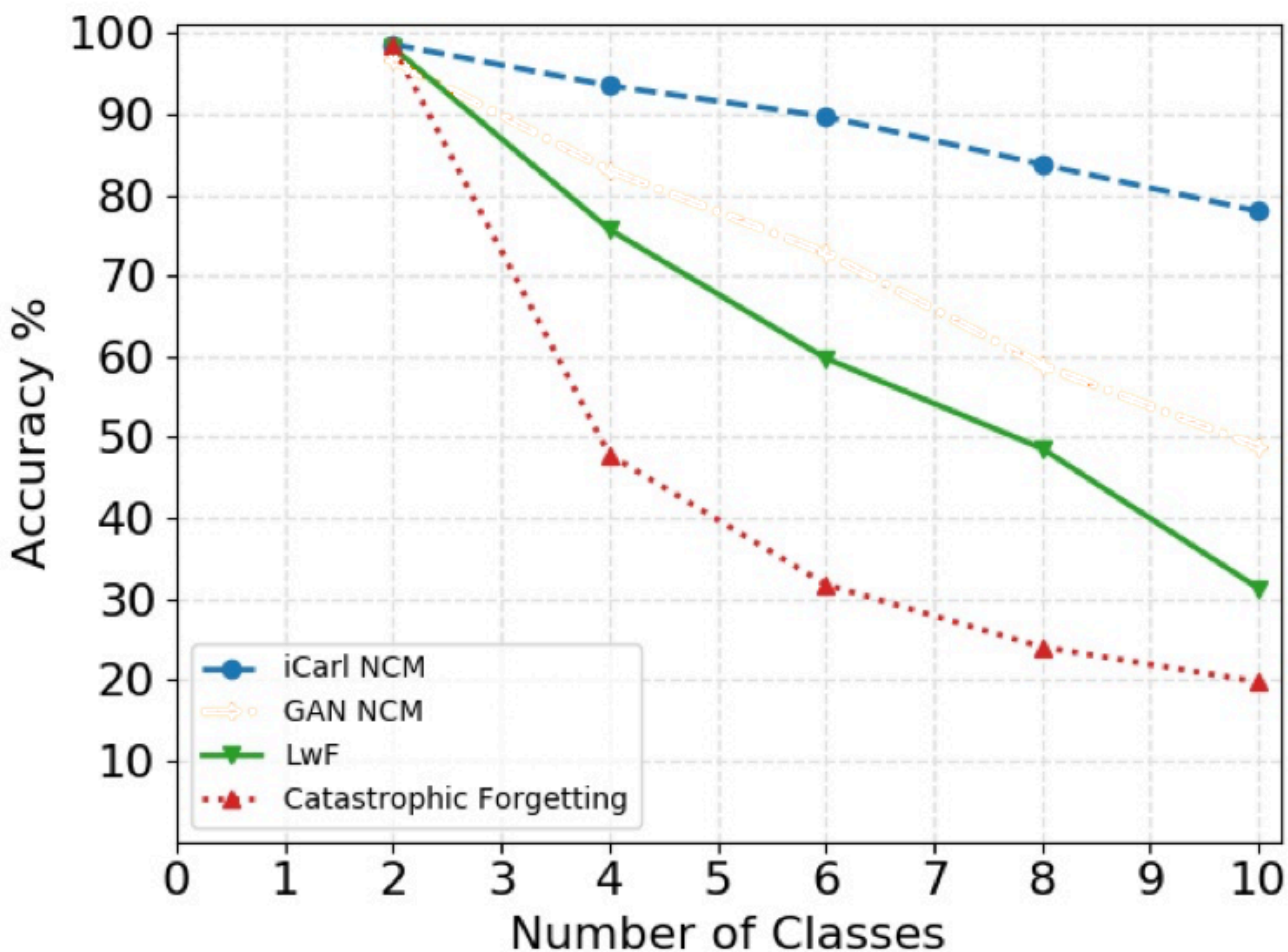
GANs Results on MNIST



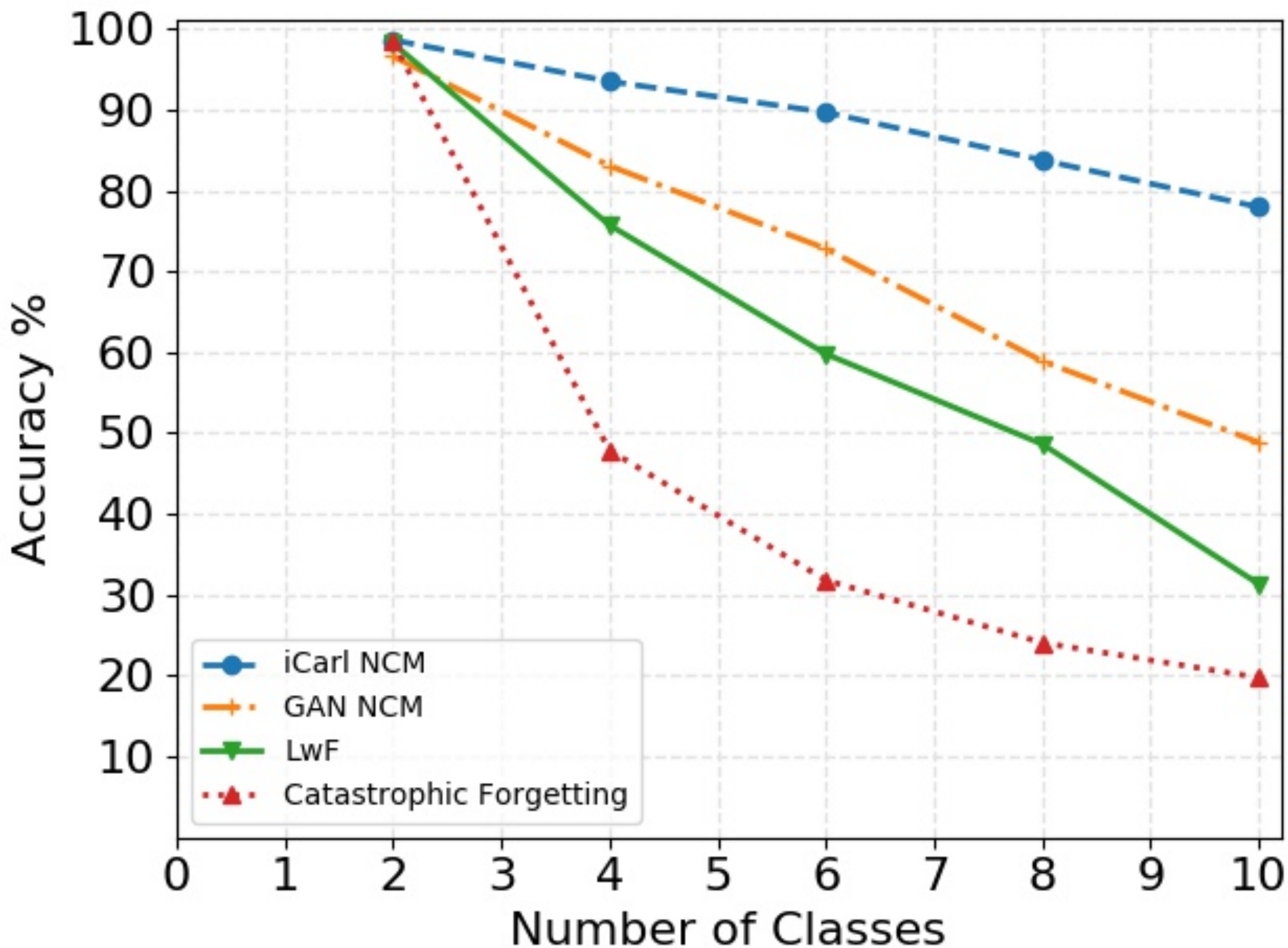
GANs Results on MNIST



GANs Results on CIFAR10



GANs Results on CIFAR10



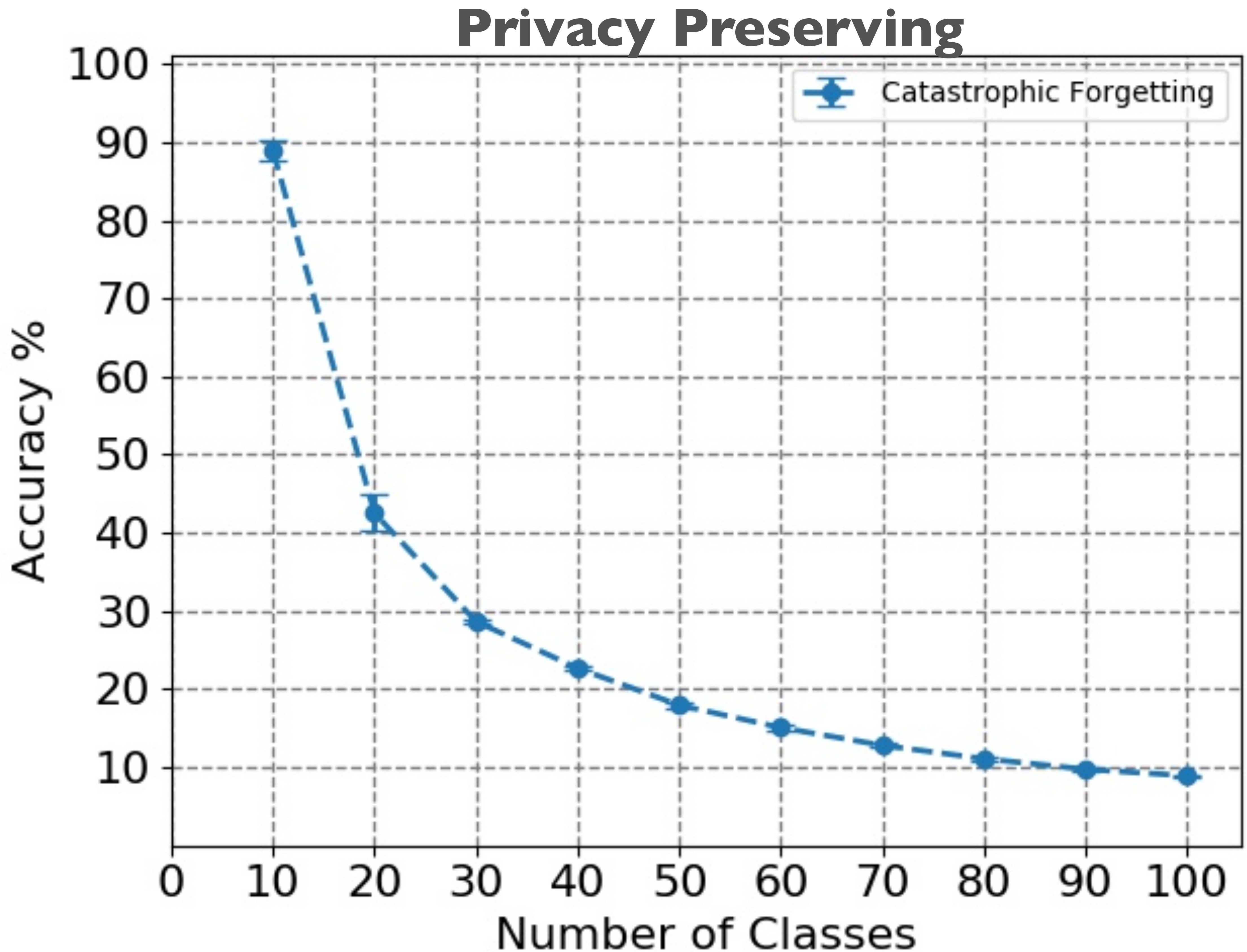
PRIVACY PRESERVING INCREMENTAL LEARNING

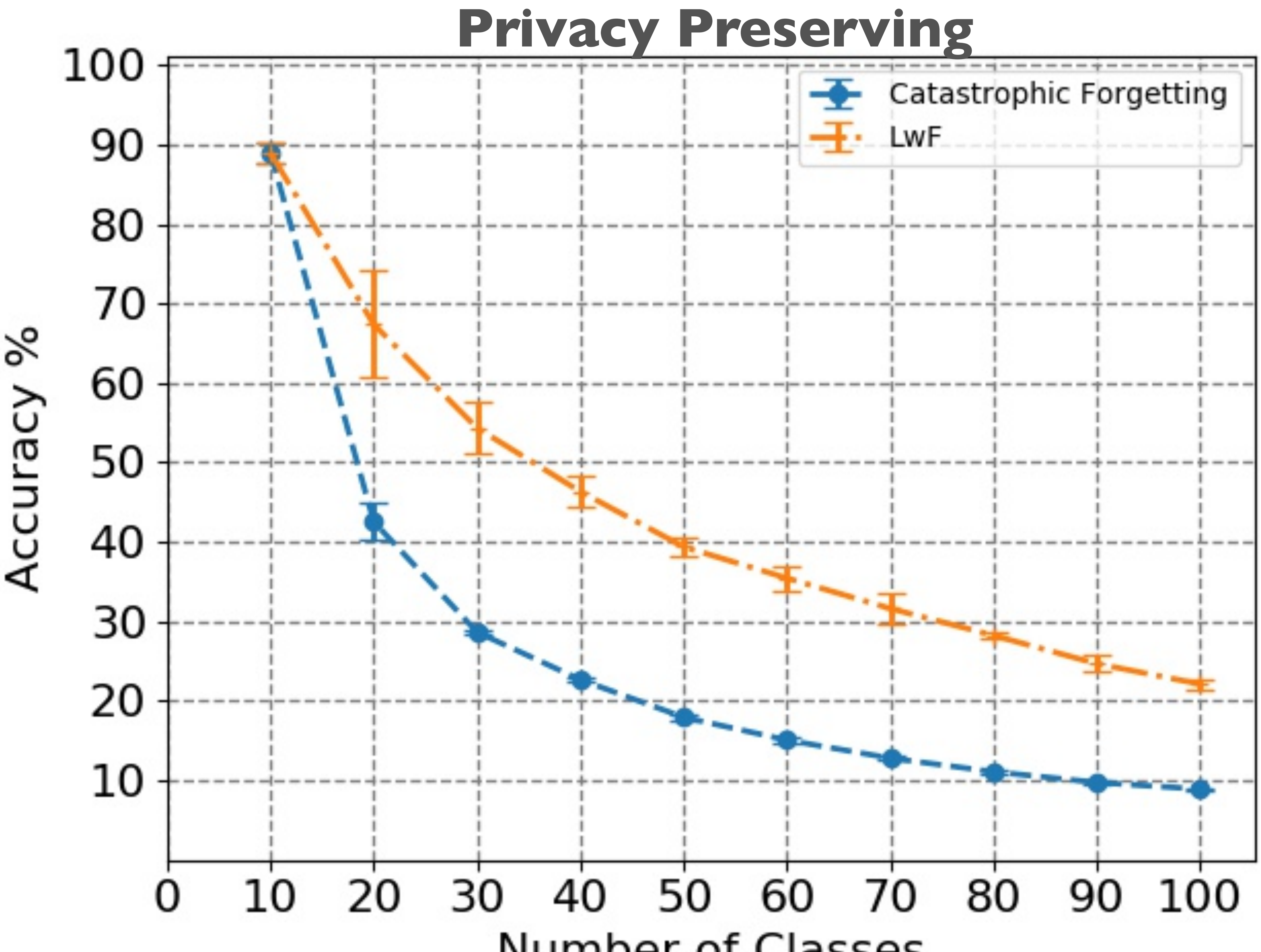
Proposed Methodologies

- Revisiting the problem statement
  - Privacy concerns

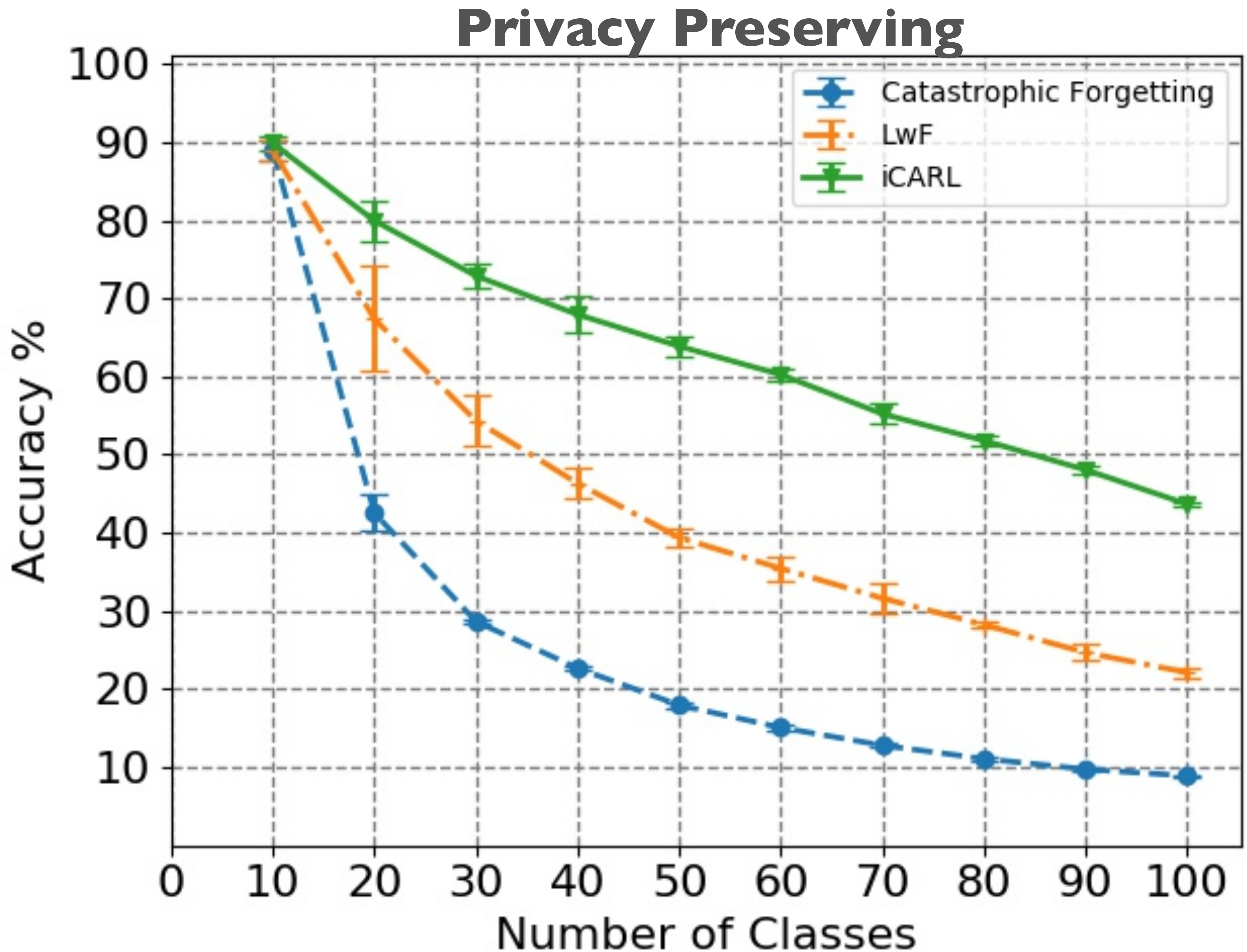
Proposed Methodologies

- Revisiting the problem statement
  - Privacy concerns
  - Memory concerns


Privacy Preserving
Catastrophic Forgetting
Accuracy %
Number of Classes
0
10
20
30
40
50
60
70
80
90
100


Privacy Preserving
Catastrophic Forgetting
LwF
Accuracy %
Number of Classes
0
10
20
30
40
50
60
70
80
90
100

Proposed Methodologies

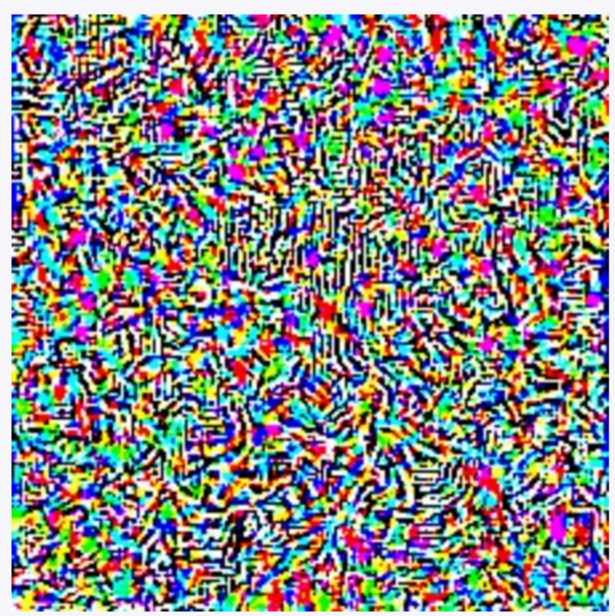


PRIVACY PRESERVING INCREMENTAL LEARNING

- **Idea 1.0: Use adversarial instances**

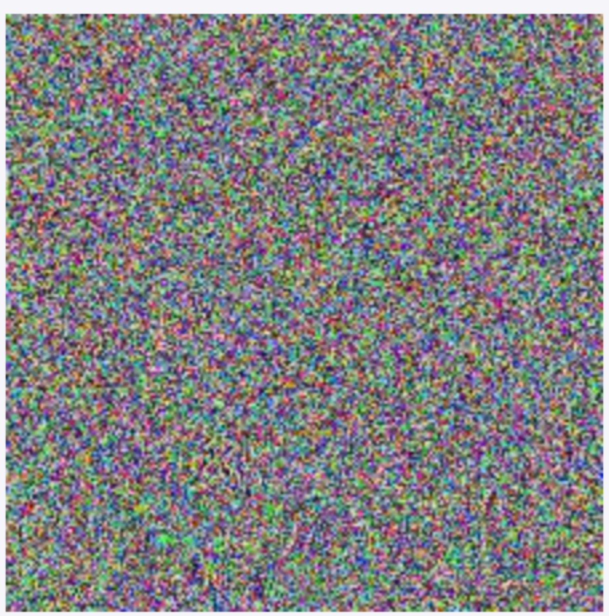
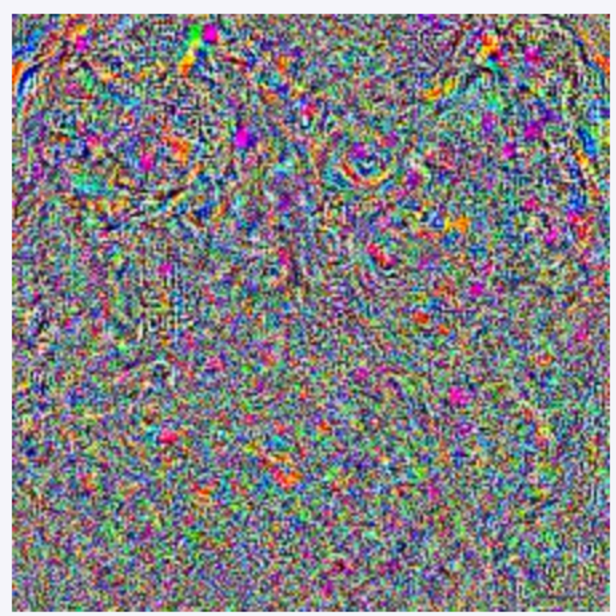
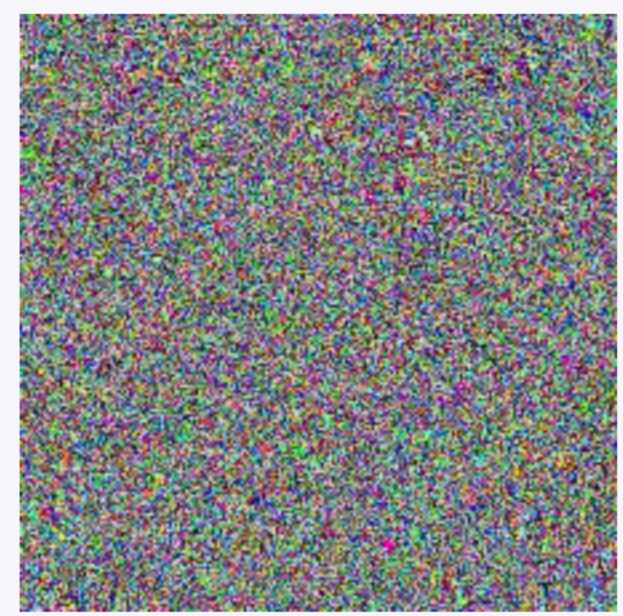
PRIVACY PRESERVING INCREMENTAL LEARNING

- **Idea 1.0: Use adversarial instances**

Predicted as **Eel** (390) Confidence: 0.96	Adversarial Noise	Predicted as **Blowfish** (397) Confidence: 0.81
		

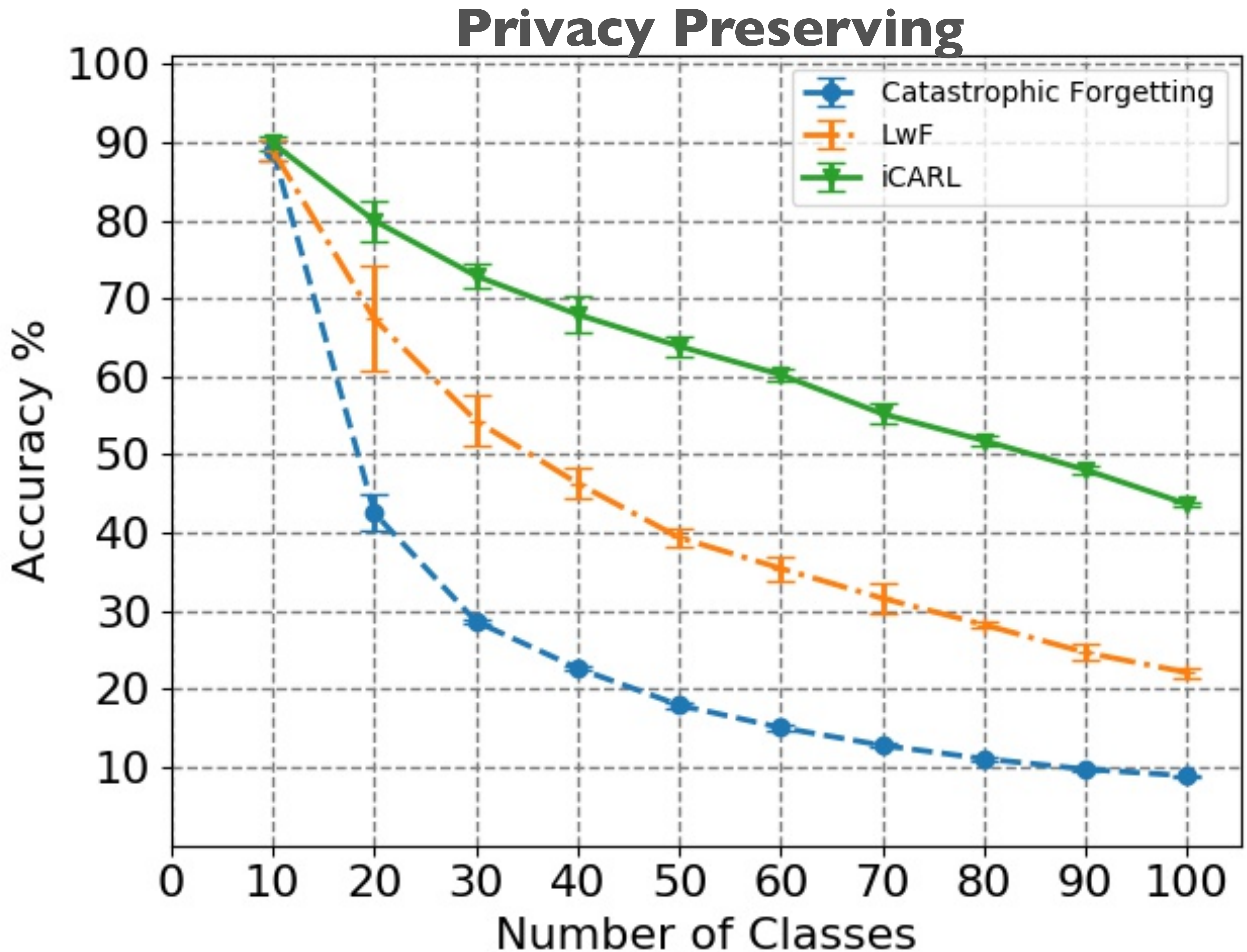
PRIVACY PRESERVING INCREMENTAL LEARNING

- **Idea 1.0: Use adversarial instances**

Predicted as **Zebra** (340) Confidence: 0.94	Predicted as **Bow tie** (457) Confidence: 0.95	Predicted as **Castle** (483) Confidence: 0.99
		

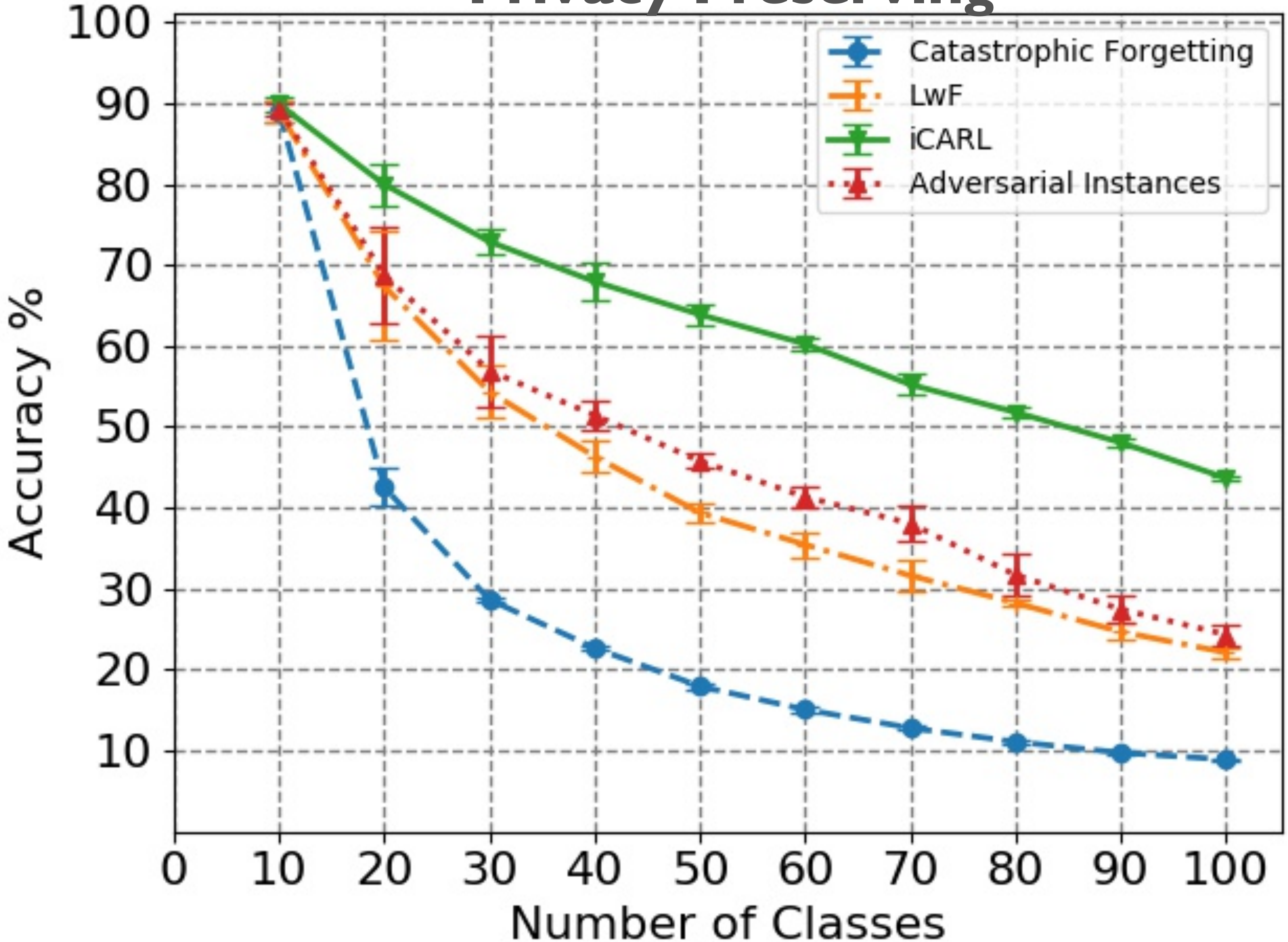
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

Proposed Methodologies



Proposed Methodologies

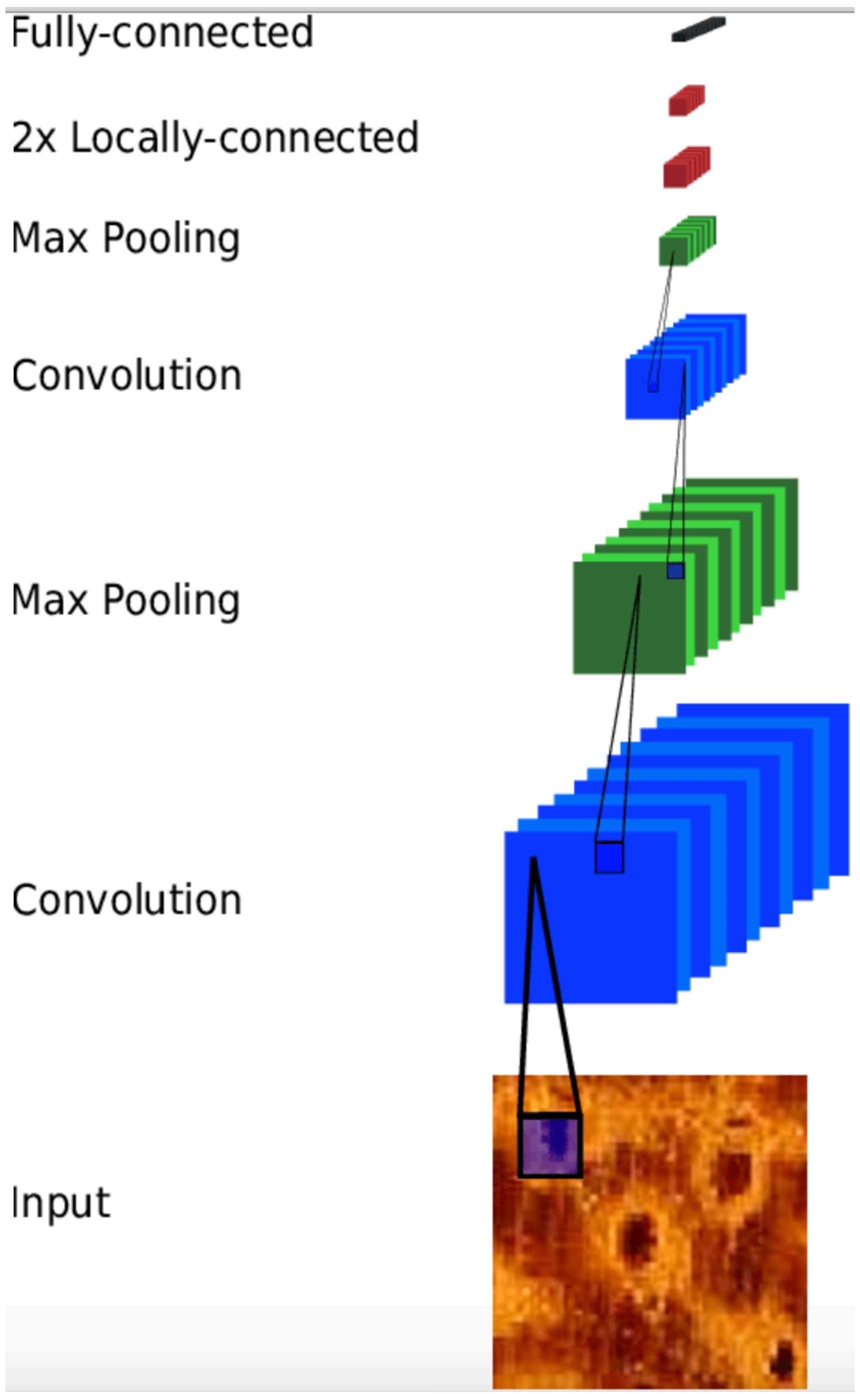
Privacy Preserving



PRIVACY PRESERVING INCREMENTAL LEARNING

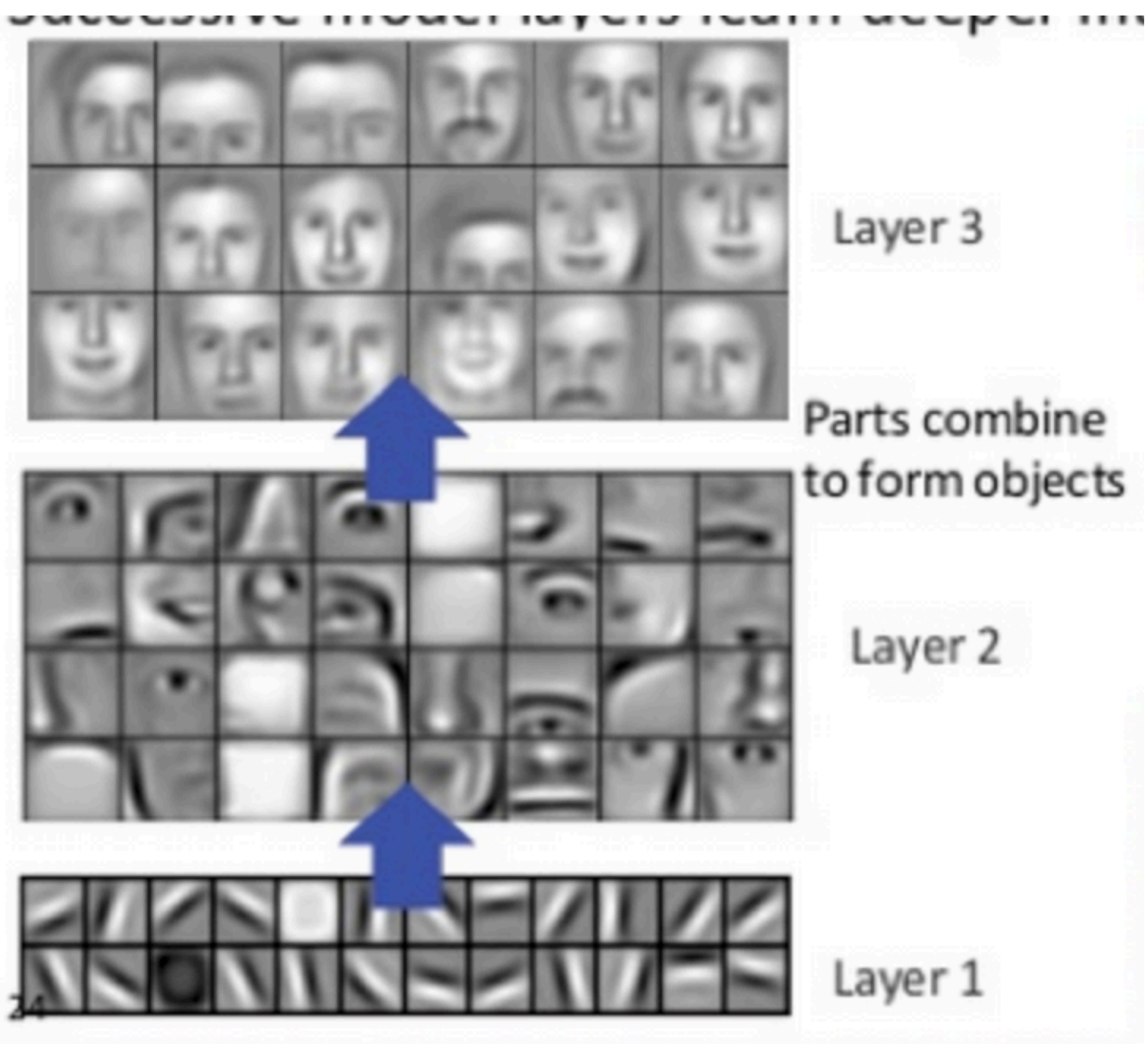
- **Idea 2.0: Store instance features**

Proposed Methodologies



https://www.researchgate.net/profile/Luiz_Gustavo_Hafemann/publication/279181075/figure/fig8/AS:613923078275131@1523382076826/The-Deep-Convolutional-Neural-Network-architecture.png

Proposed Methodologies



https://deeplearning4j.org/img/feature_hierarchy.png

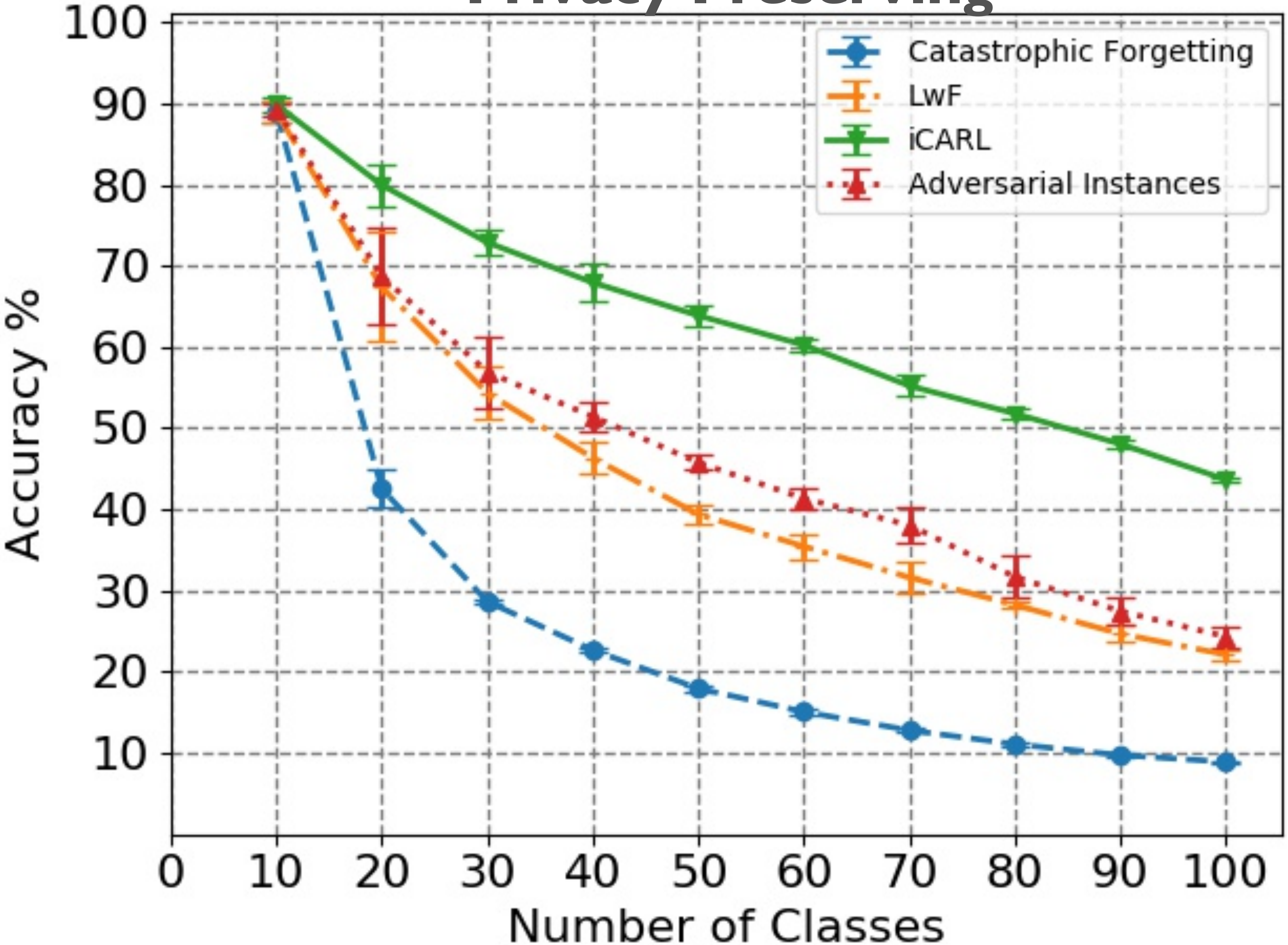
PRIVACY PRESERVING INCREMENTAL LEARNING

- **Idea 2.0: Store instance features**
  - Fix initial layers and store its features for all original class instances

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

Proposed Methodologies

Privacy Preserving



Proposed Methodologies

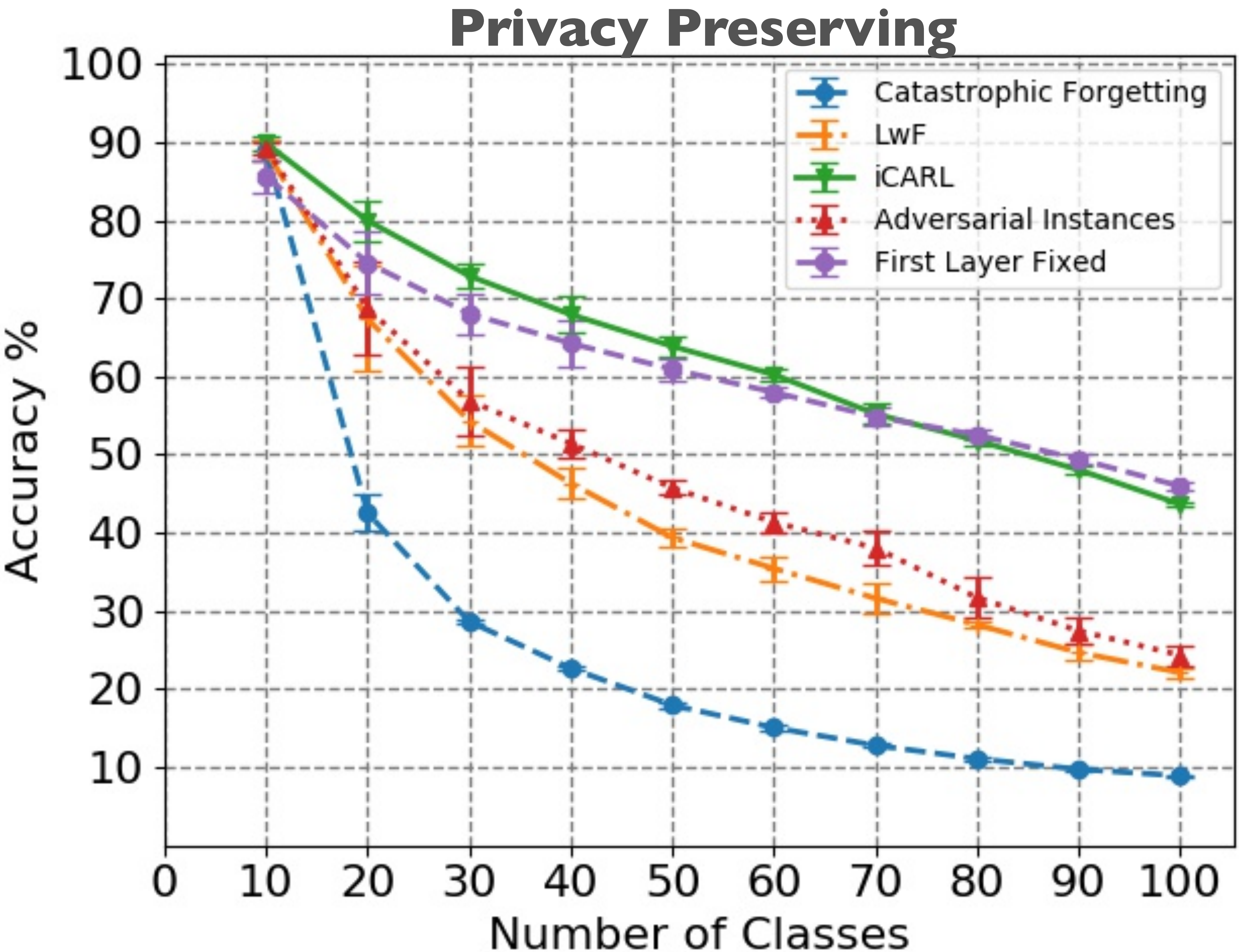
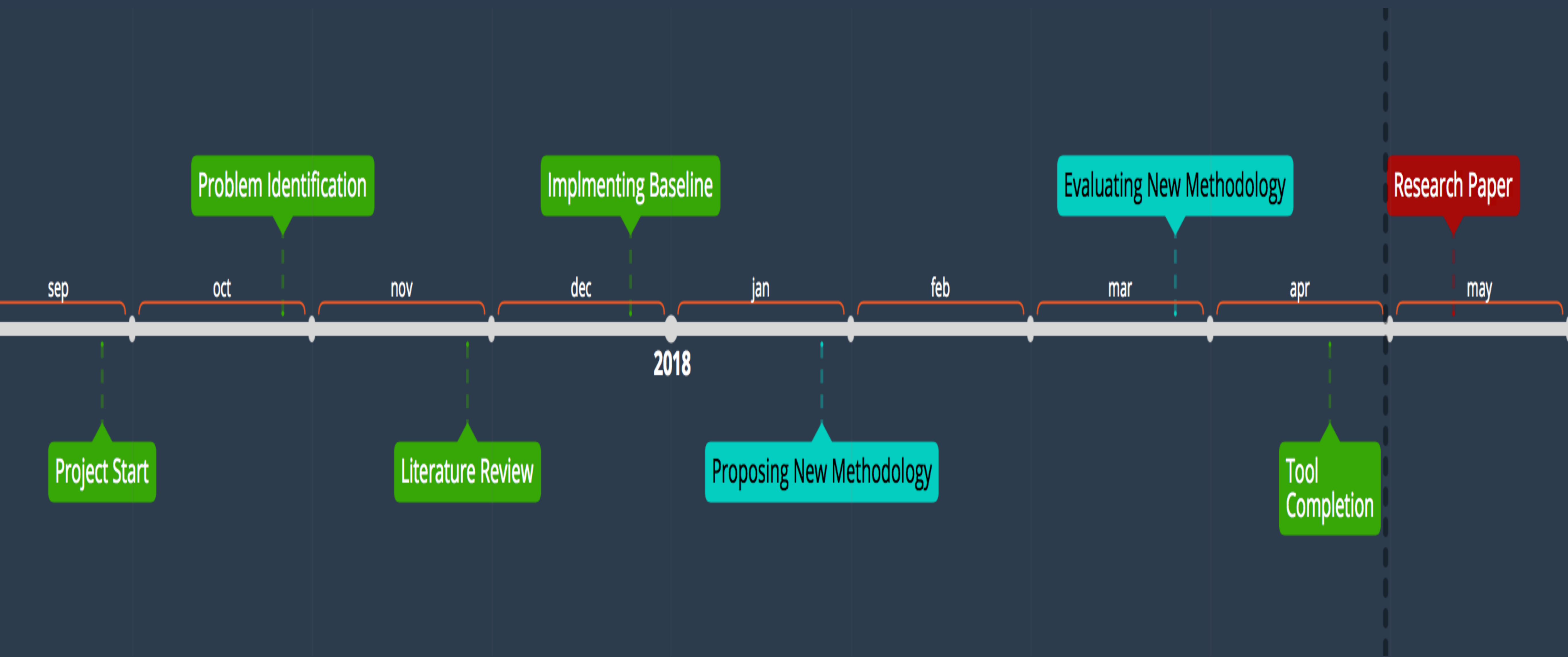


TABLE OF CONTENTS

TIMELINE


Problem Identification
Implmenting Baseline
Evaluating New Methodology
Research Paper
sep
oct
nov
dec
jan
feb
mar
apr
may
2018
Project Start
Literature Review
Proposing New Methodology
Tool Completion

TEAM WORK

- **Khurram Javed**
  - Bias removal through Scale computation
  - Supervision on GAN-based Approach
  - Analysis of existing literature
- **Talha Paracha**
  - Privacy Preserving Strategies
  - Analysis of existing literature

TABLE OF CONTENTS

COMPLETED ALGORITHMS

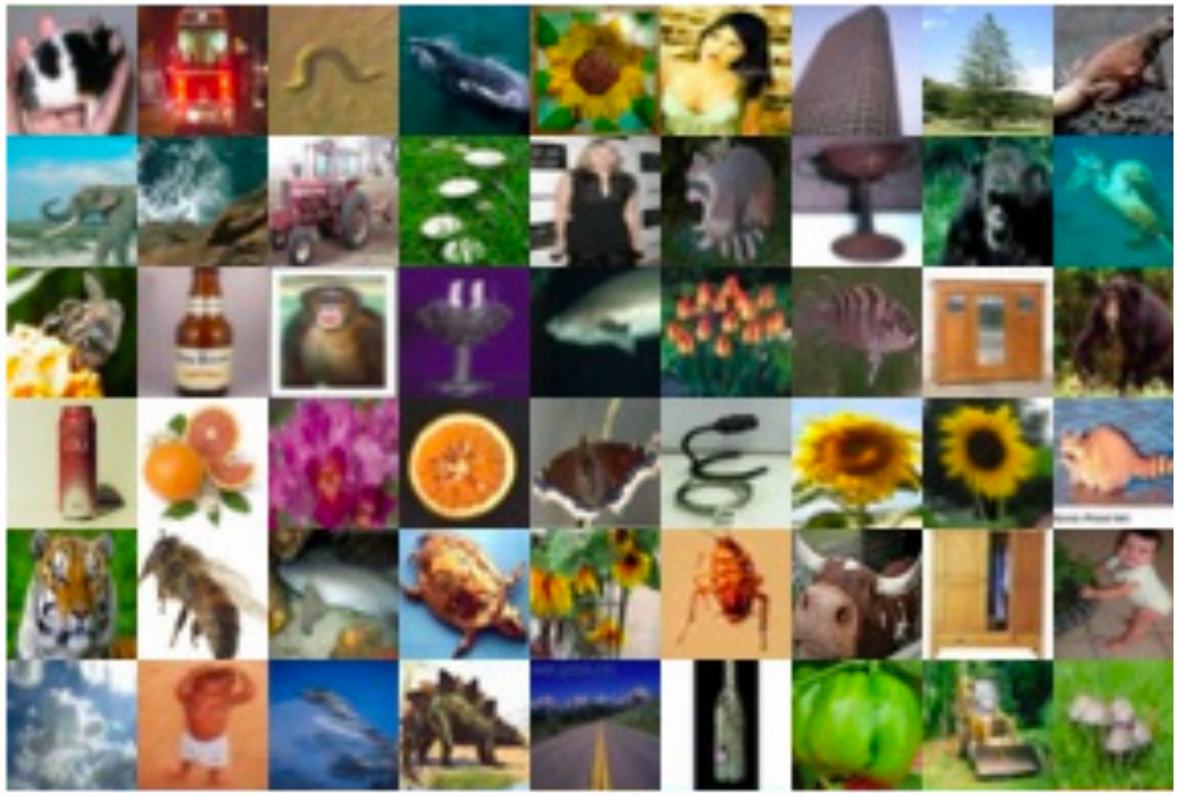
- iCaRL paper implementation.
- GAN based Incremental Learning.
- Adversarial Instances based Incremental Learning.
- Real-time scale computation.

COMPLETED MODULES

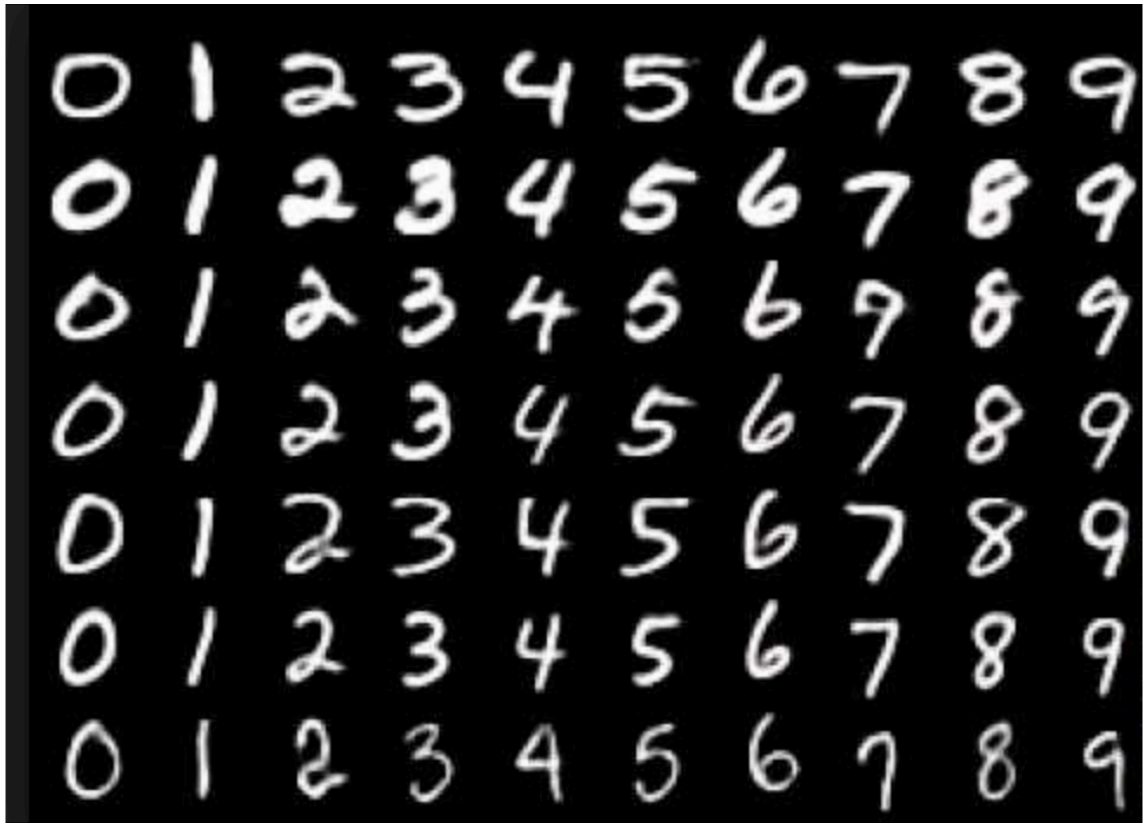
- **Support for multiple datasets.**
- Support for multiple models.
- Support for logging, and plotting.
- Support for reproducibility.

COMPLETED MODULES

- **Support for multiple datasets.**



CIFAR



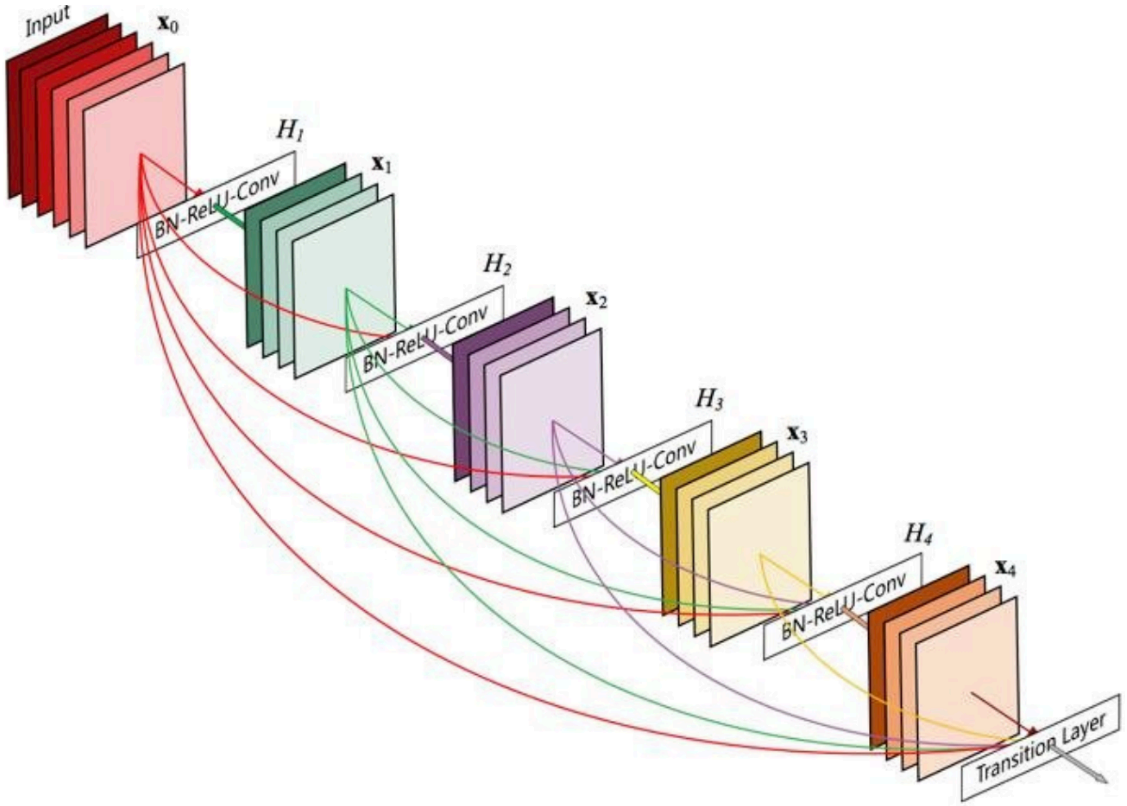
MNIST

COMPLETED MODULES

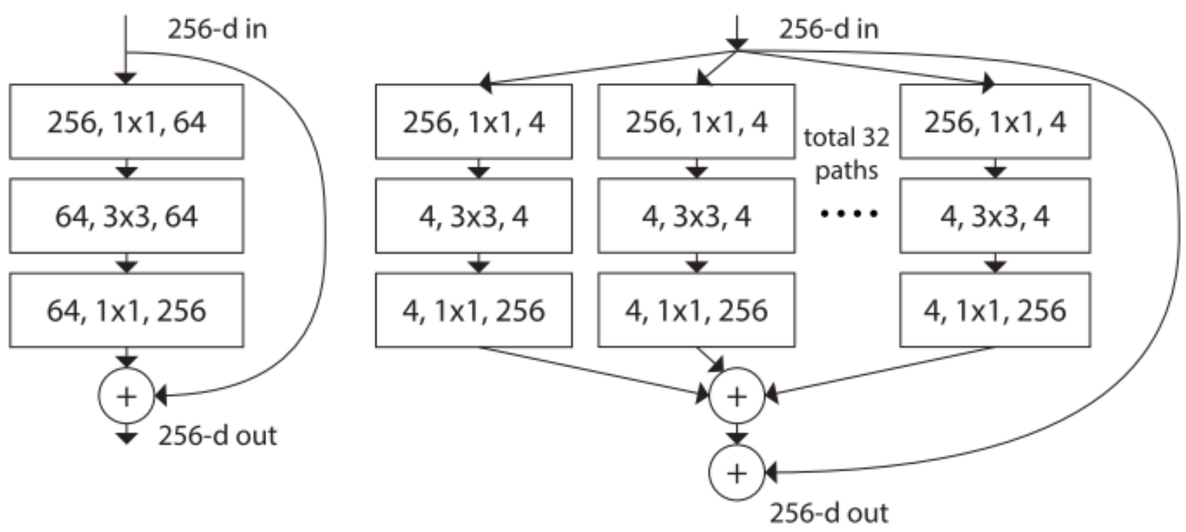
- Support for multiple datasets.
- **Support for multiple models.**
- Support for logging, and plotting.
- Support for reproducibility.

COMPLETED MODULES

- **Support for multiple models.**



DenseNet



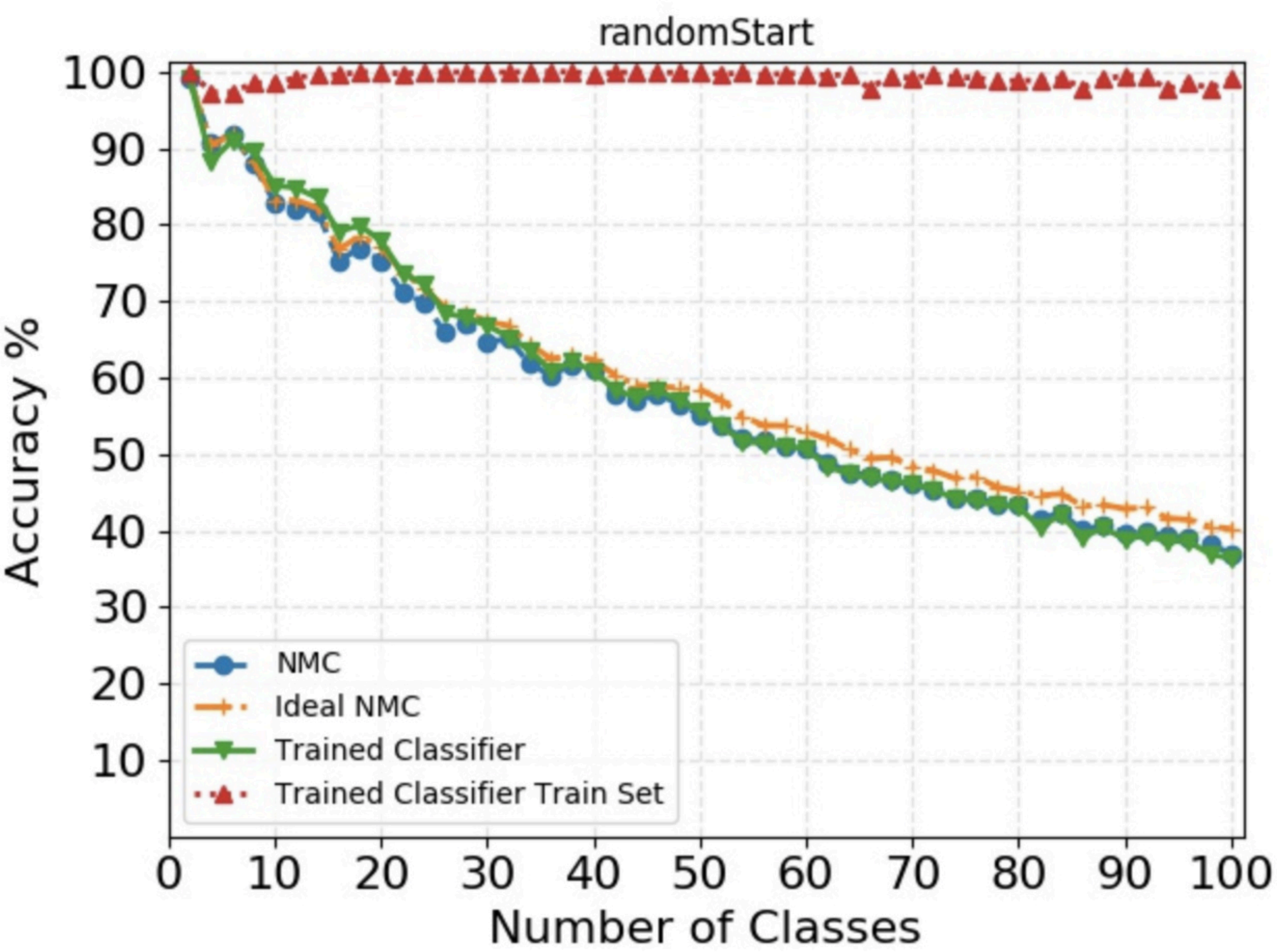
ResNet

COMPLETED MODULES

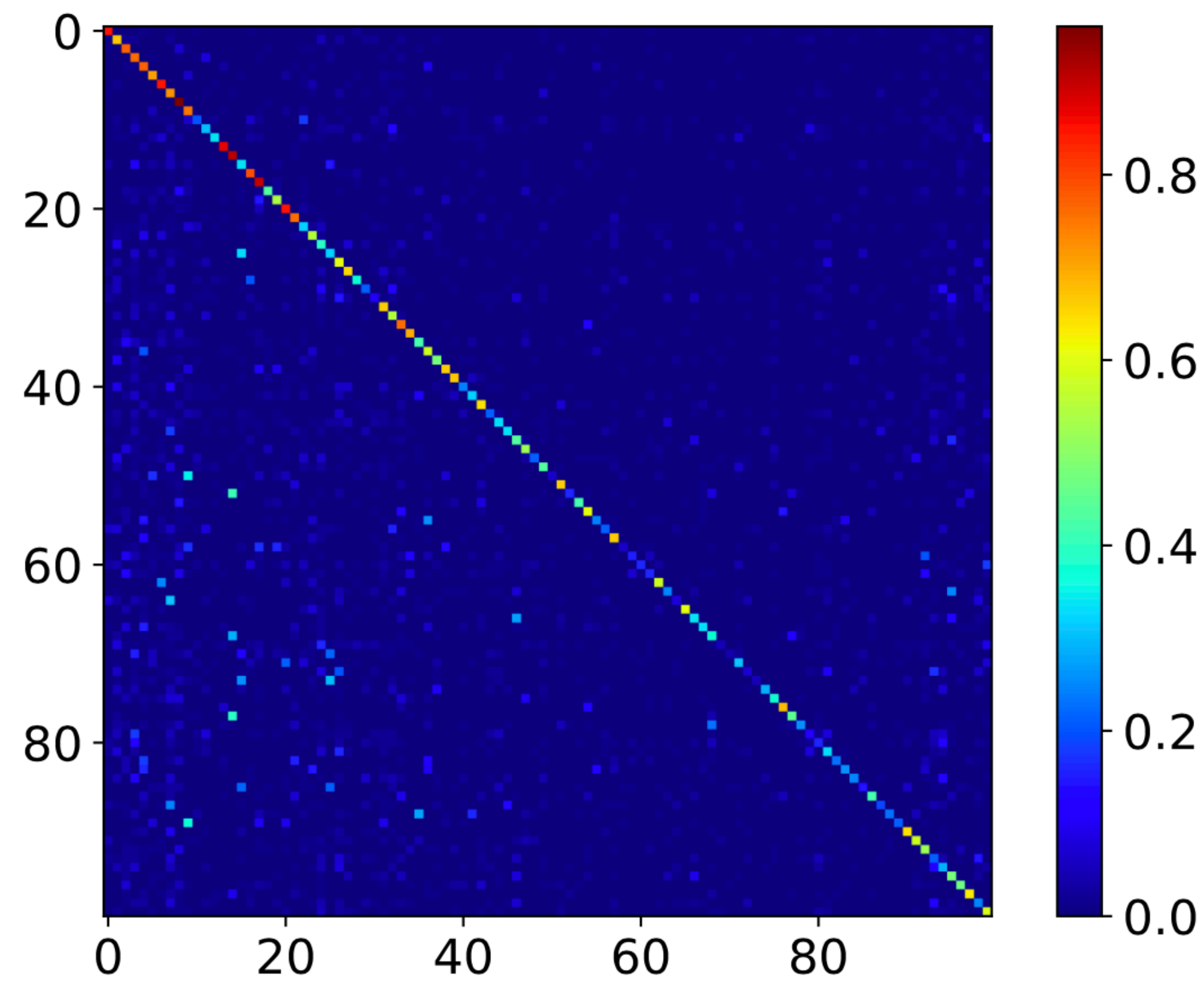
- Support for multiple datasets.
- Support for multiple models.
- **Support for logging, and plotting.**
- Support for reproducibility.

COMPLETED MODULES

- **Support for logging, and plotting.**



Experiment Plot



Confusion Matrix

COMPLETED MODULES

- Support for multiple datasets.
- Support for multiple models.
- Support for logging, and plotting.
- **Support for reproducibility.**

COMPLETED MODULES

- **Support for reproducibility.**

ad1000JSONDump

"{\"results\": {\"Ideal NMC\": [[10, 20, 30, 40, 50, 60, 70, 80, 90, 100], [88.9, 75.15, 63.43333333333333, 61.0, 56.58, 53.18333333333333, 48.871428571428574, 46.25, 42.82222222222222, 37.88]], \"Temp Results\": [[1, 2, 3, 4], [5, 6, 2, 6]], \"Trained Classifier\": [[10, 20, 30, 40, 50, 60, 70, 80, 90, 100], [88.4, 66.7, 57.233333333333334, 53.425, 48.18, 44.18333333333333, 39.77142857142857, 37.2125, 33.28888888888889, 26.81]], \"NMC\": [[10, 20, 30, 40, 50, 60, 70, 80, 90, 100], [88.9, 75.2, 63.5, 60.1, 55.66, 52.15, 46.642857142857146, 44.0625, 40.67777777777778, 35.32]], \"Train Error Classifier\": [[10, 20, 30, 40, 50, 60, 70, 80, 90, 100], [98.84, 85.55, 76.78, 73.82, 72.024, 68.47333333333333, 67.68571428571428, 67.475, 65.02, 62.73]]}, \"gitHash\": \"d4ad6bcfb3b33e5f7e6b21bfcb781e6962a64cdb\\n\", \"name\": \"1142018/ad1000\", \"path\": \"../1142018/ad1000_0/ad1000\", \"dir\": \"../\", \"params\": {\"batch_size\": 64, \"lwf\": false, \"adversarial\": true, \"no_cuda\": false, \"seed\": 200, \"distill_only_exemplars\": false, \"T\": 1, \"no_random\": true, \"outputDir\": \"../\", \"no_distill\": false, \"alpha\": 0.2, \"gammas\": [0.2, 0.2, 0.2], \"memory_budget\": 500, \"name\": \"ad1000\", \"cuda\": true, \"lr\": 2.0, \"momentum\": 0.9, \"model_type\": \"resnet32\", \"no_upsampling\": false, \"memory_budgets\": [500], \"seeds\": [200], \"no_herding\": false, \"step_size\": 10, \"epochs_class\": 70, \"decay\": 4e-05, \"log_interval\": 10, \"dataset\": \"CIFAR100\", \"rand\": false, \"schedule\": [45, 60, 68]}}"

MODERN TOOL USAGE

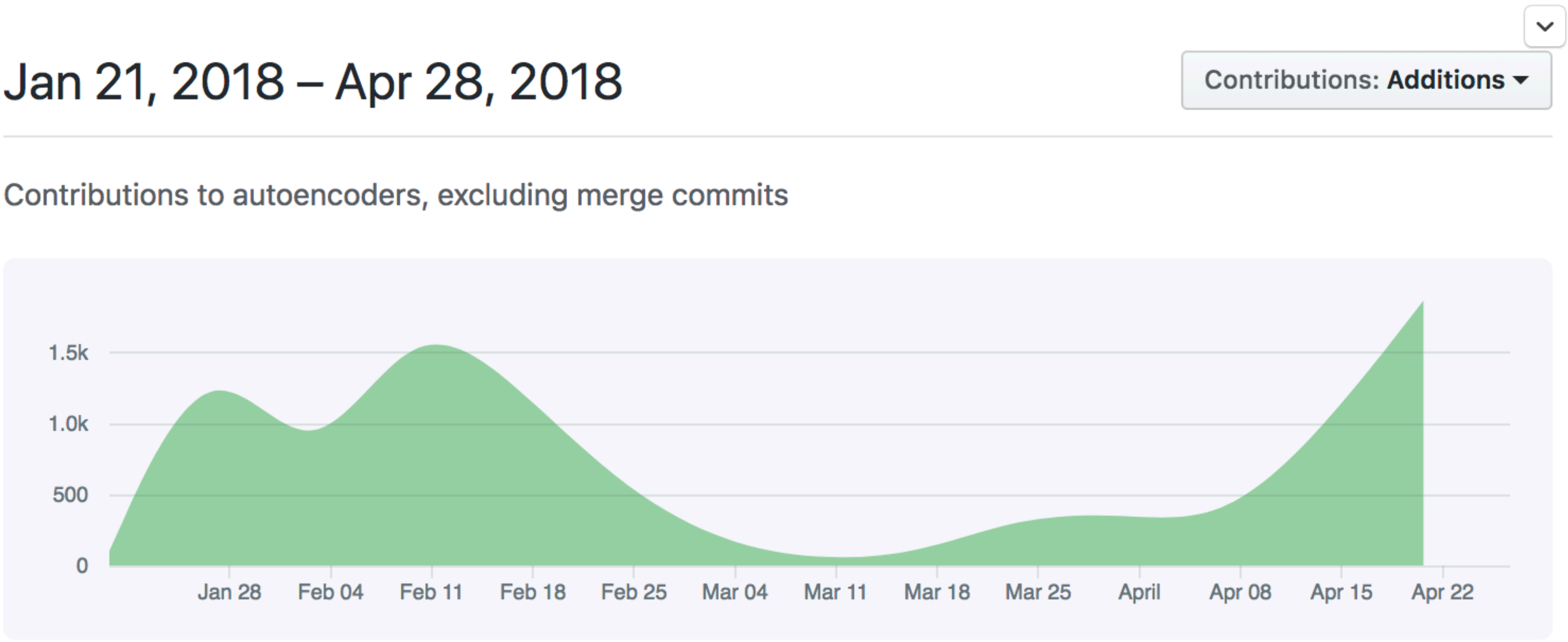
- PyTorch
  - Why not TensorFlow?
    - Dynamic Graph vs Static Graph

PYTORCH

MODERN TOOL USAGE

- Git / Github
  - Over 1,000 commits

MODERN TOOL USAGE



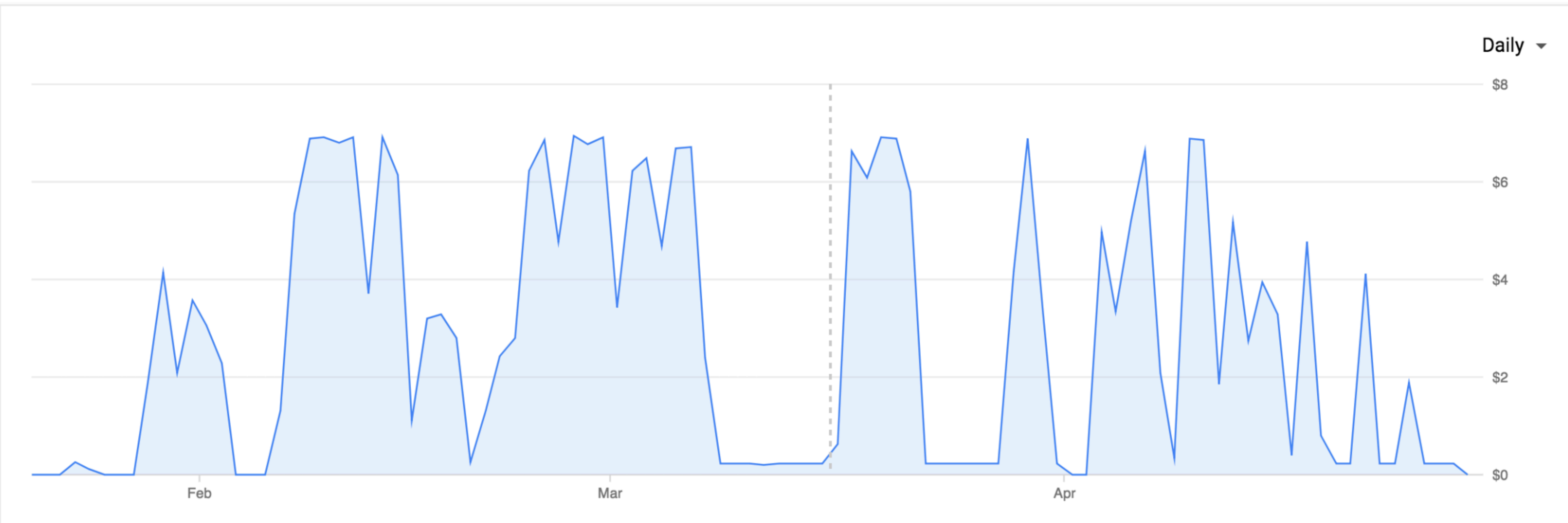
MODERN TOOL USAGE

- Ubuntu 16.04, CUDA 9, CuDNN, Bash, Vim, Google Compute Cloud

MODERN TOOL USAGE

1253 Hours of GPU compute on Google Cloud (NVIDIA K80)

Cost after credit for January 20 – April 28, 2018	Cost after credit for October 14, 2017 – January 20, 2018	Change	Credit for January 20 – April 28, 2018
$0.00	—	↓ $0.00	-$277.02



MODERN TOOL USAGE

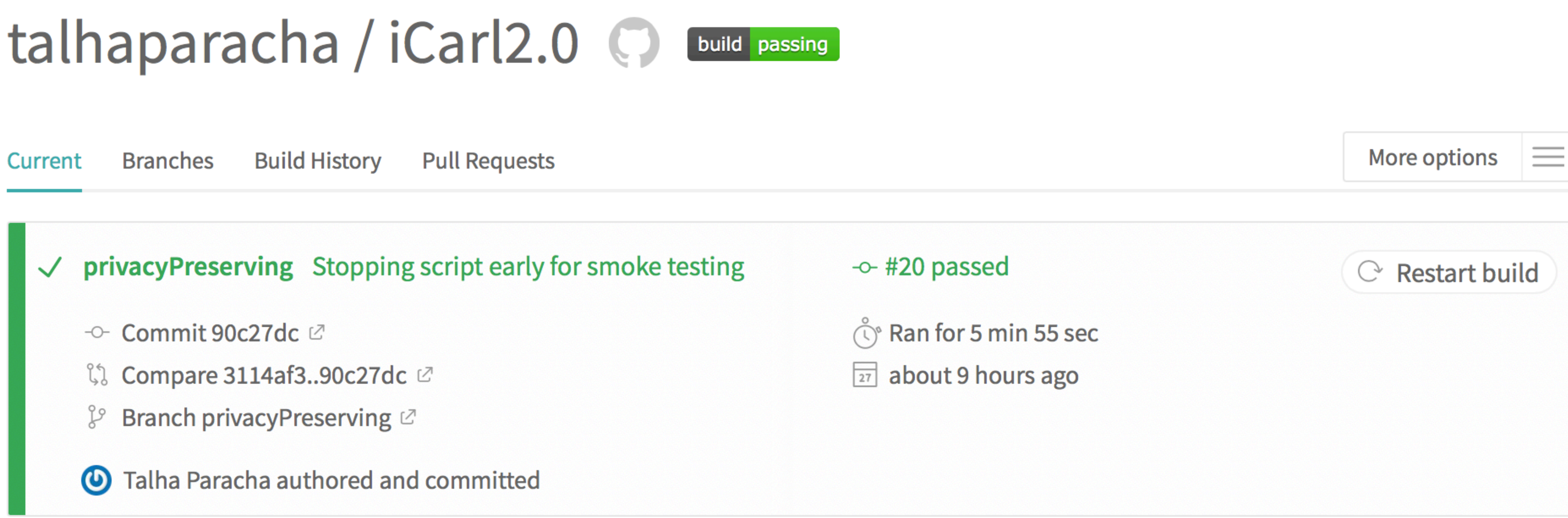
380 Hours of GPU compute on TUKL Lab Hardware

> GTX Titan X
> GTX 1060
> GTX 970

MODERN TOOL USAGE

- Travis CI

MODERN TOOL USAGE



CODING STANDARDS

- Object Oriented Paradigm.
  - Ability to add new datasets and models without modifying existing code.
- Python3 standards official guidelines (lower_case variables, camelCase functions etc)

INTUITIVE INTERFACE

```
khurramjaved@tukl-server1:~/iCarl2.0$ python unstructuredExperiment.py --help
usage: unstructuredExperiment.py [-h] [--batch-size N] [--lr LR]
                                 [--schedule SCHEDULE [SCHEDULE ...]]
                                 [--gammas GAMMAS [GAMMAS ...]] [--momentum M]
                                 [--no-cuda] [--random-init] [--no-distill]
                                 [--distill-only-exemplars] [--no-random]
                                 [--no-herding] [--seeds SEEDS [SEEDS ...]]
                                 [--log-interval N] [--model-type MODEL_TYPE]
                                 [--name NAME] [--outputDir OUTPUTDIR]
                                 [--upsampling] [--pp] [--hs]
                                 [--alphas ALPHAS [ALPHAS ...]]
                                 [--decay DECAY]
                                 [--alpha-increment ALPHA_INCREMENT] [--l1 L1]
                                 [--step-size STEP_SIZE] [--T T]
                                 [--memory-budgets MEMORY_BUDGETS [MEMORY_BUDGETS ...]]
                                 [--epochs-class EPOCHS_CLASS]
                                 [--unstructured-size UNSTRUCTURED_SIZE]
                                 [--dataset DATASET] [--lwf] [--ignore]
                                 [--no-nl] [--rand] [--adversarial]
```

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

Software engineering aspect

Feedback when running

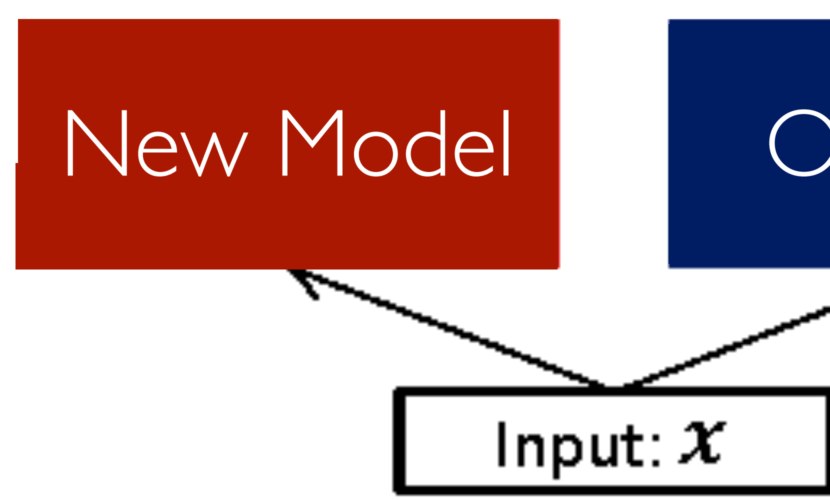
```
KeyboardInterrupt
khurramjaved@tukl-server1:~/iCarl2.0$ python unstructuredExperiment.py --epochs-class 2 --batch-size 300 --log-interval 1
Files already downloaded and verified
Files already downloaded and verified
21315e8d96984ec15be055790a8ae2de8d260bc3

Shuffling turned off for debugging
Running Experiment No 1
Increment No 0.00
Training Main Classifier
100%|████████████████████████████████████████| 2/2 [00:08<00:00,  4.38s/it]
Epoch[00Train 6.Test/s] Scaled   GScaled
1.00    22.66   24.00   24.00   22.70
Training Distillation Computer
100%|████████████████████████████████████████| 2/2 [00:04<00:00,  2.37s/it]
Distillation Computer Error (Train, Test) : 32.32 32.50
Increment No 1.00
Training Main Classifier
100%|████████████████████████████████████████| 2/2 [00:17<00:00,  8.87s/it]
Epoch[00Train 3.Test/s] Scaled   GScaled
1.00    22.16   18.90   20.20   18.80
Training Distillation Computer
100%|████████████████████████████████████████| 2/2 [00:05<00:00,  2.73s/it]
Distillation Computer Error (Train, Test) : 26.01 33.80
Increment No 2.00
Training Main Classifier
100%|████████████████████████████████████████| 2/2 [00:18<00:00,  9.15s/it]
Epoch[00Train 3.Test/s] Scaled   GScaled
1.00    22.11   10.07   11.93   10.43
Training Distillation Computer
100%|████████████████████████████████████████| 2/2 [00:05<00:00,  2.68s/it]
Distillation Computer Error (Train, Test) : 26.71 35.90
Increment No 3.00
Training Main Classifier
100%|████████████████████████████████████████| 2/2 [00:18<00:00,  9.32s/it]
Epoch[00Train 3.Test/s] Scaled   GScaled
1.00    25.70   9.15    10.53   9.72
Training Distillation Computer
100%|████████████████████████████████████████| 2/2 [00:05<00:00,  2.62s/it]
Distillation Computer Error (Train, Test) : 25.57 36.70
Increment No 4.00
Training Main Classifier
100%|████████████████████████████████████████| 2/2 [00:19<00:00,  9.99s/it]
Epoch[00Train 3.Test/s] Scaled   GScaled
1.00    24.51   7.12    7.52    7.44
Training Distillation Computer
```

CLOSING THE PROJECT

- Submitting two papers in BMVC 2018 (Deadline 7th May).
  - One paper with analysis of SOTA, threshold moving algorithm, and privacy preserving.
  - Other paper on the Cond-GAN based approach.
- Releasing the code to public.
- Continuation of the project over the summer.

DEMO + Q/As

KNOWLEDGE DISTILLATION

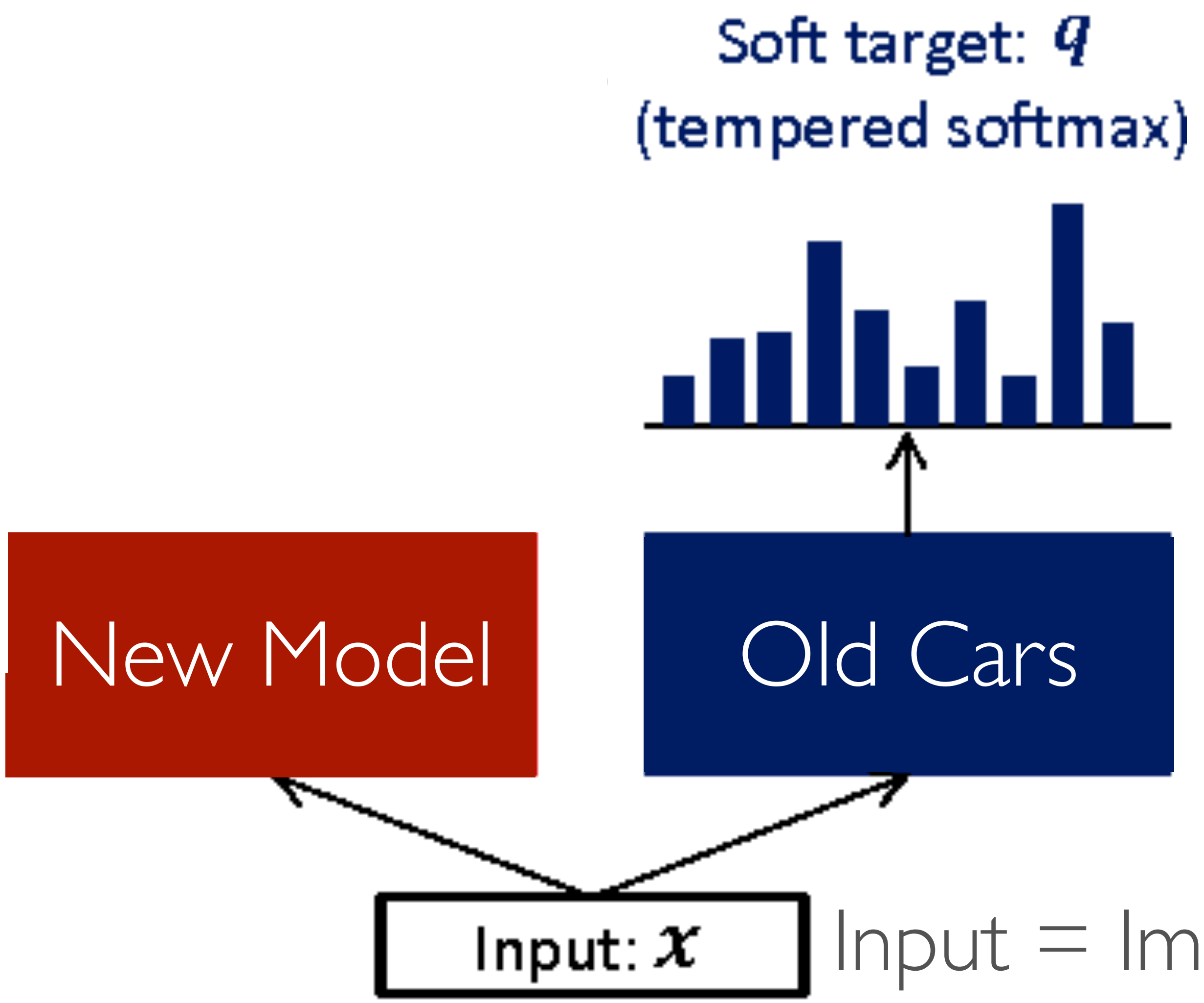


Input = Image of Tesla

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png

KNOWLEDGE DISTILLATION

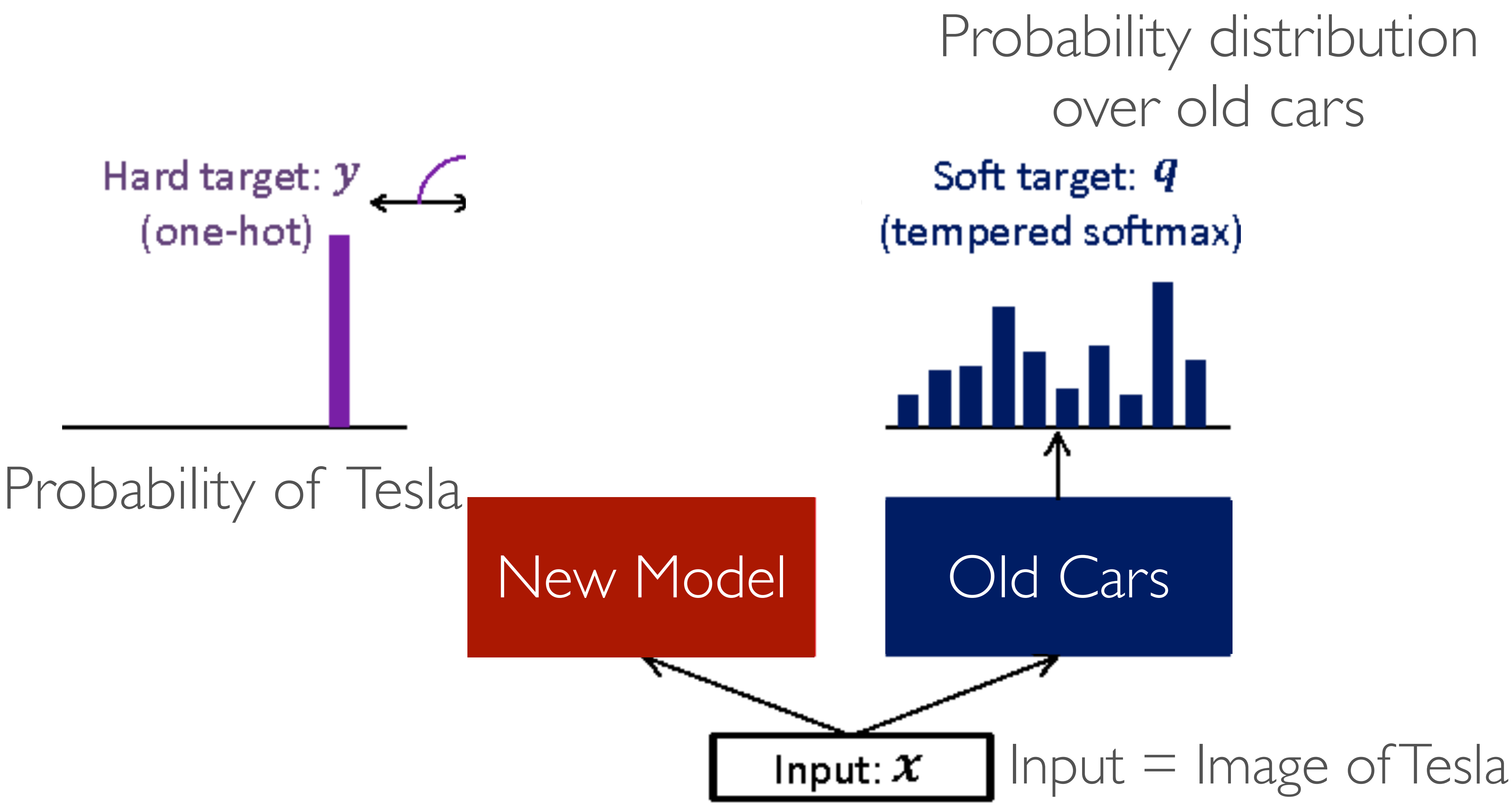
Probability distribution over old cars



Input = Image of Tesla

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png

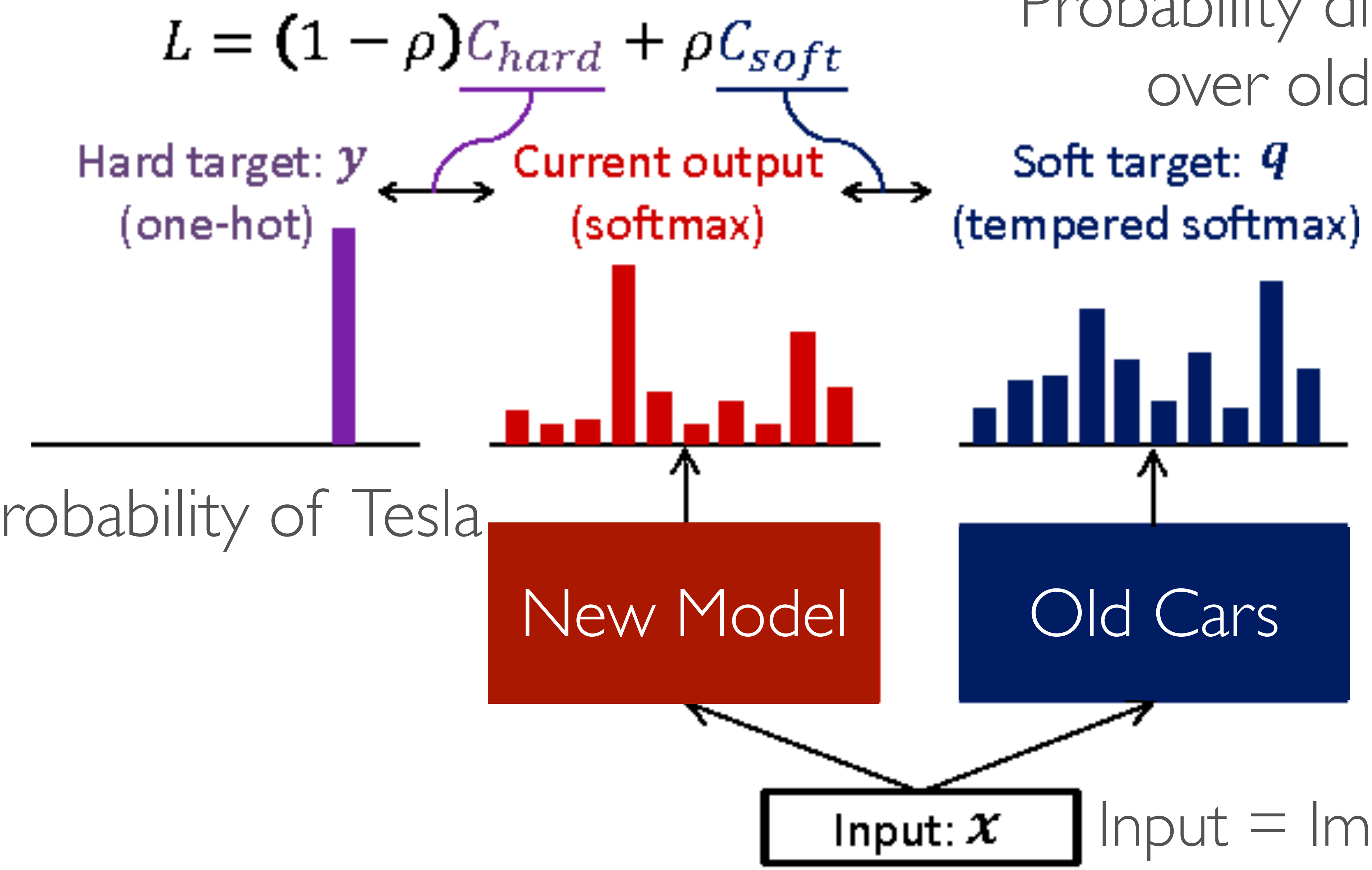
KNOWLEDGE DISTILLATION



https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png

KNOWLEDGE DISTILLATION

Probability distribution over old cars



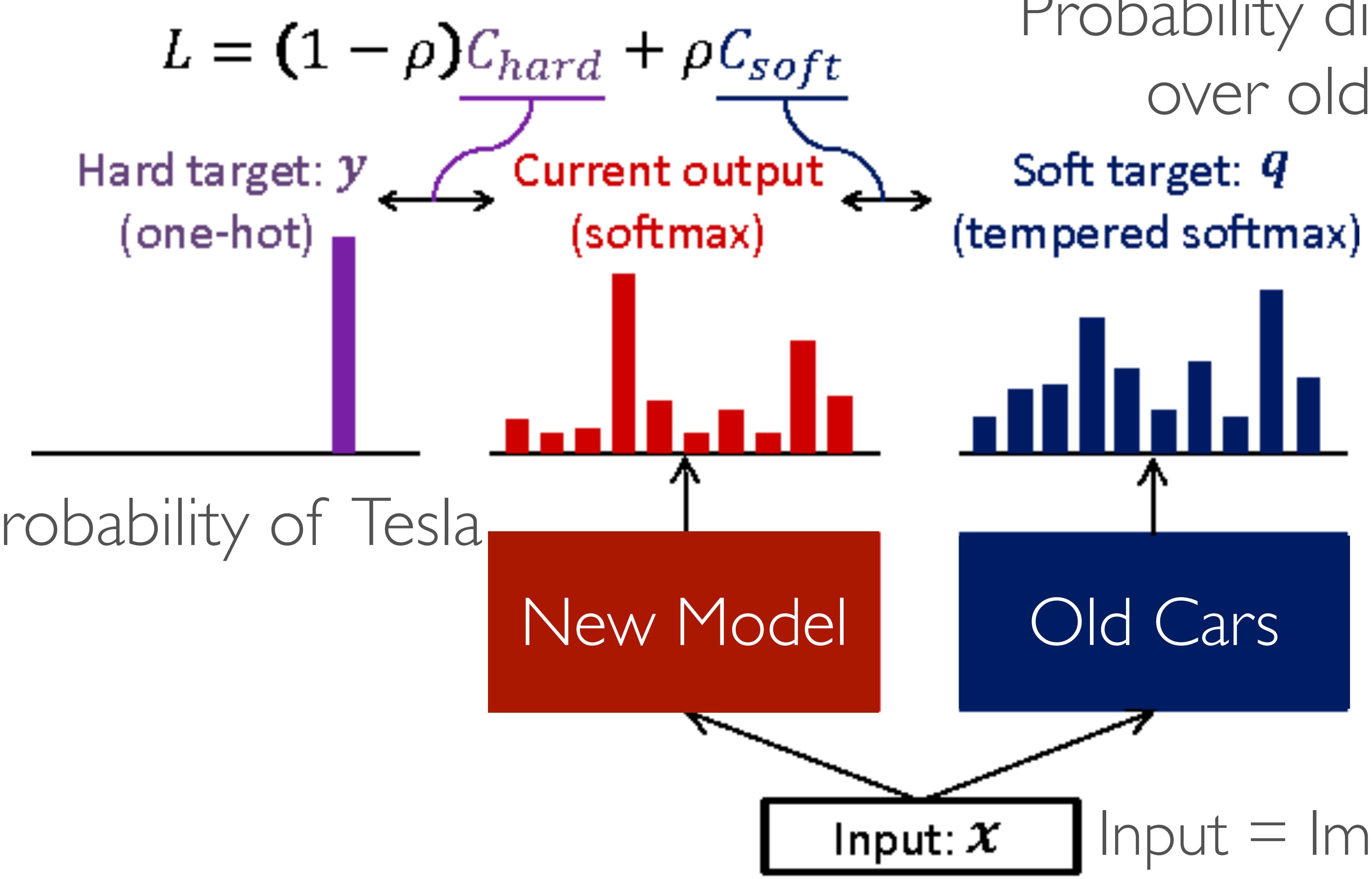
Probability of Tesla

Input = Image of Tesla

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png

SCALE COMPUTATION

Probability distribution over old cars



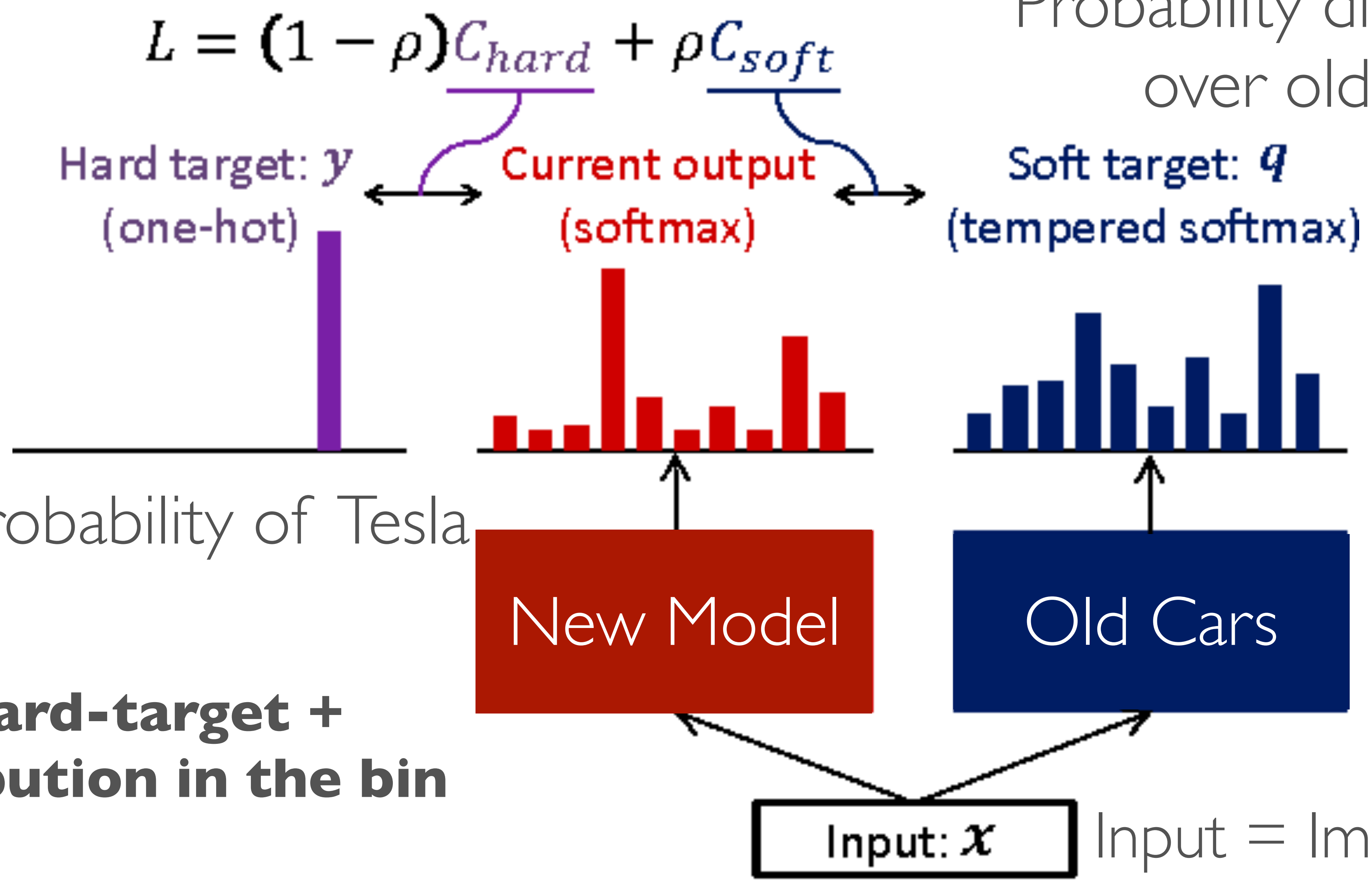
Probability of Tesla

Input = Image of Tesla

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png

SCALE COMPUTATION

Probability distribution over old cars



Probability of Tesla

Input = Image of Tesla

Add hard-target +
old distribution in the bin

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/015cb3edf7529648e784ea97fb52364329bac3d8/2-Figure1-1.png