RKMVERI

24/04/2024

# PixelSecurer

**Enhancing Privacy in Age Recognition From Images**

**Presented By**

Sourish Ghosh and Anirban Dey

**Guide**

Jimut Bahan Pal, IIT BOMBAY

# Contents

# Introduction

- Machine unlearning is the process of removing the influence of a specific subset of data from a trained machine learning model.

# Motivation

Why it is needed and where to apply?

## Why?

Machine unlearning is needed because it can help us protect privacy, ensure fairness, maintain data quality, or comply with regulation.

## Applications

1. Data Deletion Request
2. Data Correction
3. Data Debiasing

And many more…

.

# Our Goals

Problem Statement

### Goal # 1

Train a ML/DL model for predicting age from Images

### Goal # 2

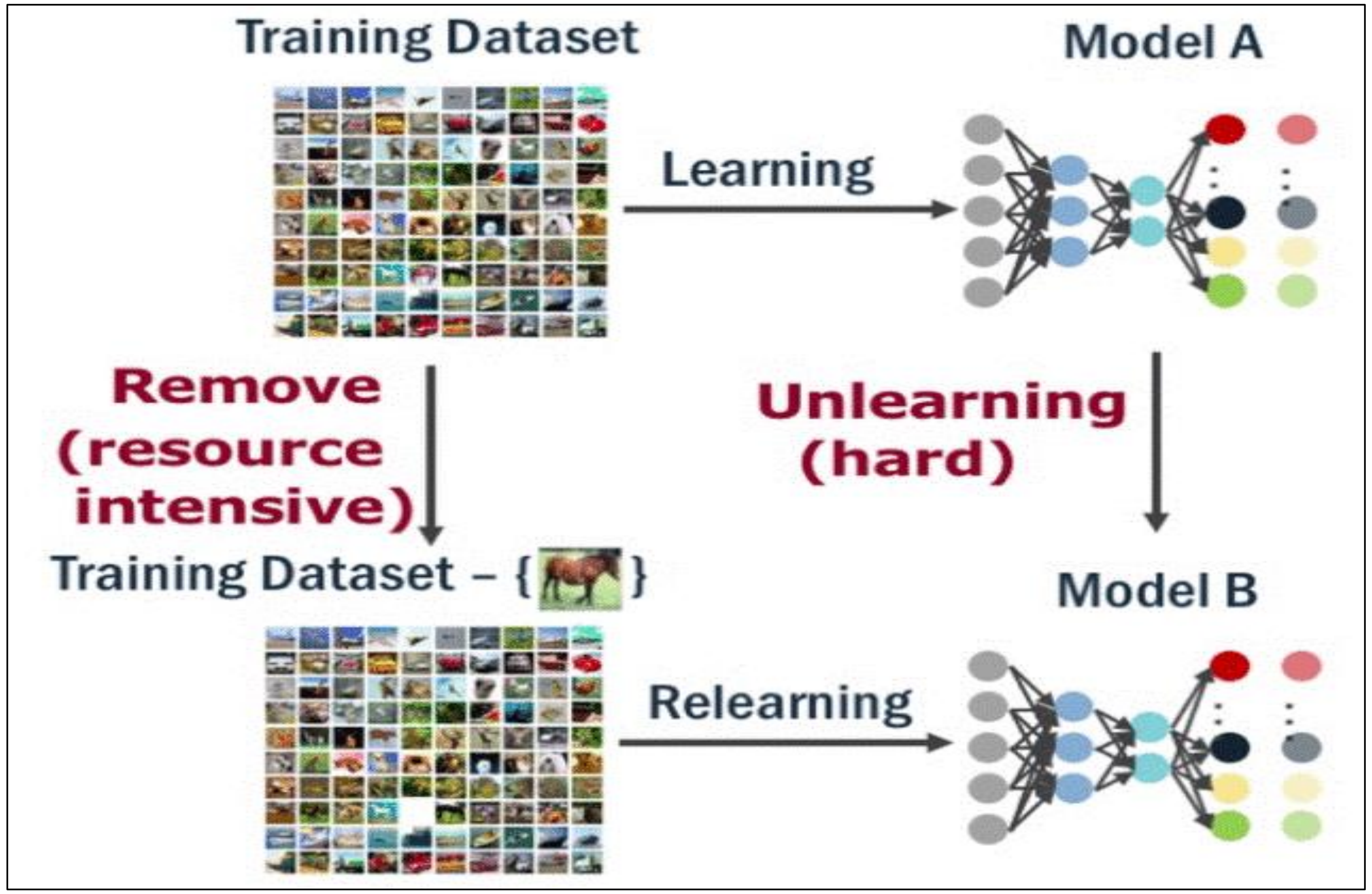
Train another model on the Retain Dataset for the same purpose.

### Goal # 3

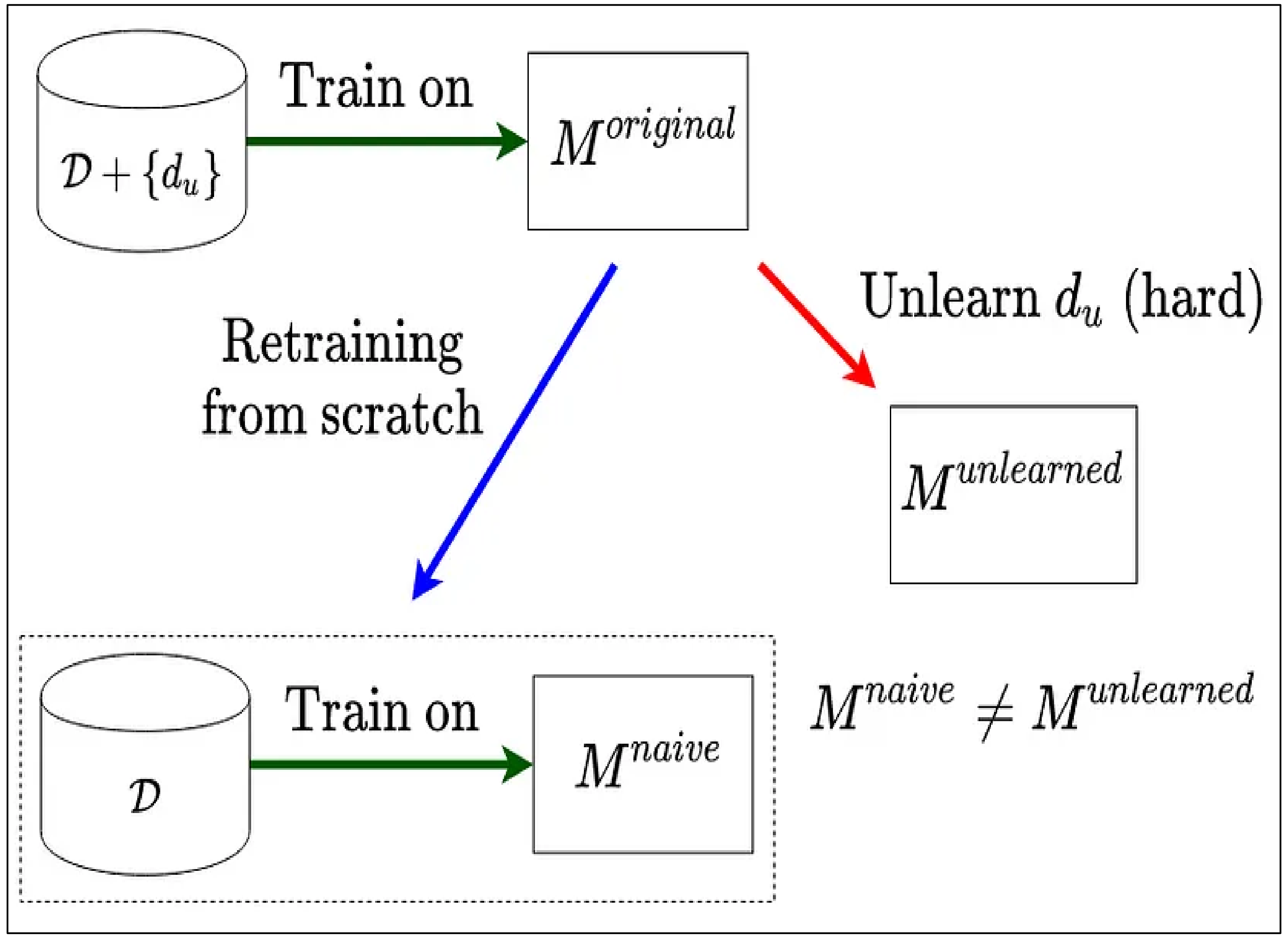
Using Machine Unlearning Compare between Unlearned model and Ideal Model

# Basic Outline

Here is the basic outline...




Naive Unlearning




Approximate Unlearning

# Some Mathematics

- $\mathcal{X}$ : Feature space
- $\mathcal{Y}$ : Output space
- $\mathcal{Z}^*$ : Space of datasets
- **D** $\in \mathcal{Z}^*$ : Multiset of data points (Allowing for duplicate entries)
- A hypothesis function **h :** $\mathcal{X} \rightarrow \mathcal{Y}$ which assigns an output $y = h(x)$ $y \in \mathcal{Y}$ to a given input $x \in \mathcal{X}$

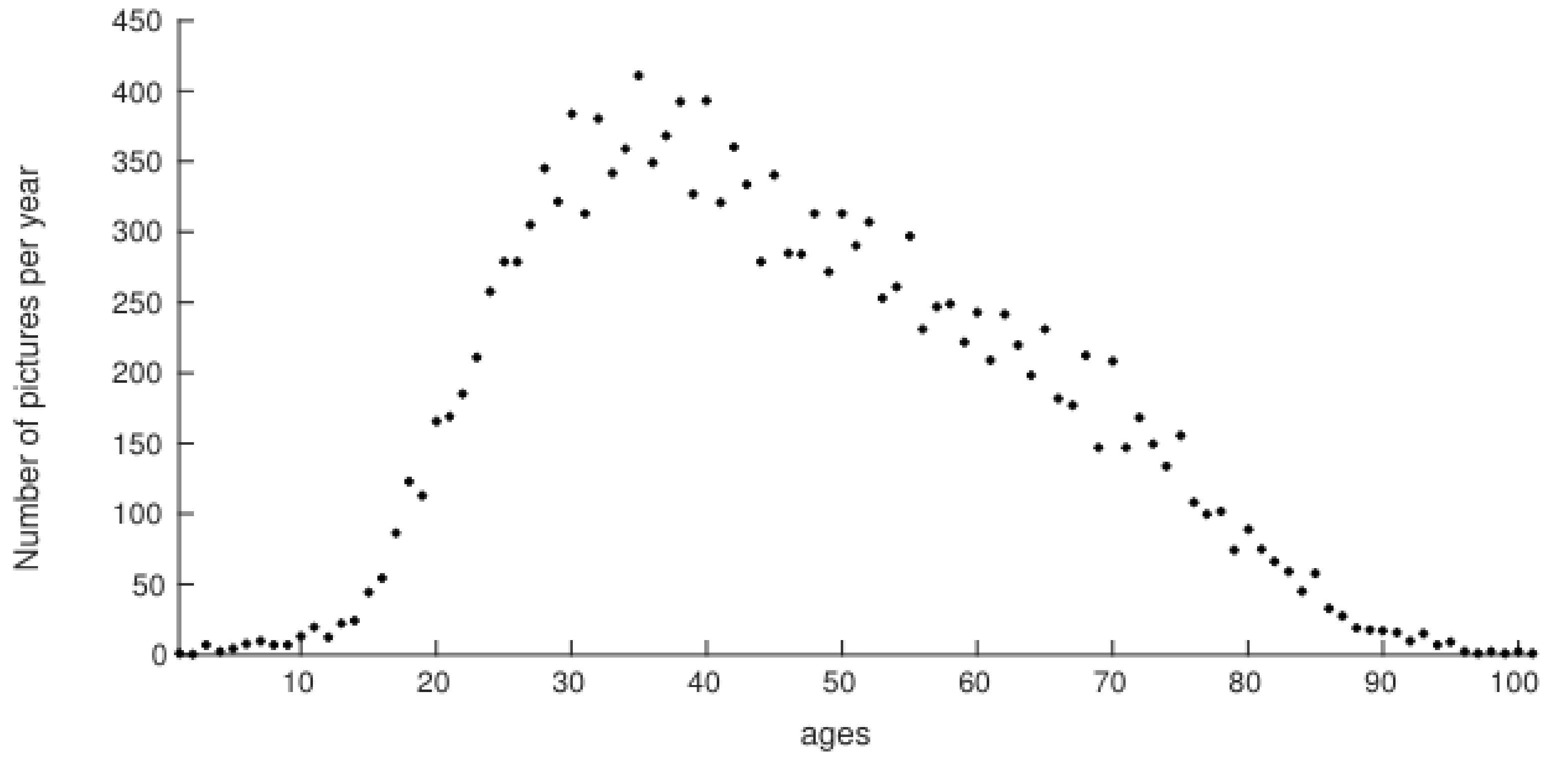
**Training Algorithm:** Training algorithm can be viewed as a map $\mathcal{A} : \mathcal{Z}^* \rightarrow \mathcal{H}$, where $\mathcal{H}$ is the space of all hypothesis functions, whose objective is to minimize a non-negative real-valued loss function $L(h, D)$.

**Update Mechanism:** An update mechanism is a map $\mathcal{U} : \mathcal{H} \times \mathcal{Z}^* \times \mathcal{Z}^* \rightarrow \mathcal{H}$, which takes as input, a model **h** $\in \mathcal{H}$, two datasets **D**, $D_u \in \mathcal{Z}^*$ , and outputs a new model $\mathcal{U}$(**h**, **D**, $D_u$ ) $\in \mathcal{H}$.

The Goal of an **unlearning algorithm** is to remove the influence of a subset $D_u \subseteq$ D of **m** samples from the trained machine learning model $\mathcal{A}$(**D**).

# Data Description

- **AgeDB** contains 16, 488 images of various famous people, such as actors/actresses, writers, scientists, politicians, etc.
- Every image is annotated with respect to the identity, age and gender attribute.
- There exist a total of 568 distinct subjects.
- The average number of images per subject is 29.
- The minimum and maximum age is 1 and 101, respectively
- The median of average ages for each subject is 50.3 years approx.




Scatter plot depicting the age distribution in the AgeDB database.


Random images from the AgeDB Database

# NOTIONS

$$D = \{x_i, y_i\}_{i=1}^{N}$$

$$x_i \in \mathbb{R}$$

$$y_i \in \mathbb{R}$$

$$D_f = \text{Data Points we wish to forget}$$

$$D_r = \text{Data Points we wish to retain}$$

$$D = D_f \cup D_r$$

$$\phi = D_f \cap D_r$$

# Blind-Spot Unlearning

- Partially expose a randomly initialized model to few samples from the retain set.
- It is trained on the retain samples for a few epochs. This gives the model a vague idea about the output distribution in the absence of the forget set from the training data.
- The forget set is a *blindspot* for this model. This partially learned blindspot model acts as an *unlearning helper*.
- Let the blindspot model be denoted as $B(.;\theta)$. We denote the original fully trained model by $M(x_i, \emptyset)$.
- In our method, the model M is updated to obtain the final unlearned model.



Source: **Deep Regression Unlearning**, Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli; *[Submitted on 15 Oct 2022 (v1), last revised 31 May 2023.* Published in 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. 

# Blind-Spot Unlearning (Cont.)

- Let the prediction made by the original model on $i\ th$ sample of dataset $D$ is $M(x_i\ ;\ \phi)$ and $y_i$ is the corresponding correct label. Then the loss for samples in $D_r$ is

$$\boldsymbol{L_r \leftarrow L(M(x_i\ ;\ \phi), y_i);\ \forall x_i\ \in\ D_r}$$

- where $\boldsymbol{L}$ denotes a standard loss function used in a regression task.

- Let $M(x_i\ ;\ \phi)$ denote the prediction of fully trained model on sample $x_i$ of dataset $D$. Similarly, let $\mathrm{B}(x_i\ ;\ \theta)$ denote the prediction of the blindspot model. If the sample $x_i$ is a part of the forget set $\boldsymbol{D_f}$ , then the following loss is computed:

$$\boldsymbol{L_f \leftarrow L(M(x_i\ ;\ \phi), B(x_i\ ;\ \theta));\ \forall x_i\ \in\ D_f}$$



Source: **Deep Regression Unlearning**, Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli; *[Submitted on 15 Oct 2022 (v1), last revised 31 May 2023.* Published in 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. 

# Blind-Spot Unlearning (Cont.)

- Finally, we optimize the closeness of activations (Micaelli & Storkey, 2019) between the last $k$ layers of model $M$ and $B$ on the forget set $D_f$

$$L_{attn} \leftarrow \lambda \sum_{j=1}^{k} \| act_j^{\phi} - act_j^{\theta} \|$$

- where $act_j^{\phi}$ and $act_j^{\theta}$ corresponds to the $jth$ layer of activation map in the original model $M$ and blindspot model $B$. $\lambda$ is a parameter used to control the relative degree of significance of the loss terms.

- The final loss is computed as:

$$\mathrm{L} \leftarrow (1 - l_f^i)L_r + l_f^i(L_f + L_{attn})$$

- where $l_f^i = 1$ for samples in the forget set and $l_f^i = 0$ otherwise.



Source: **Deep Regression Unlearning**, Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli; *[Submitted on 15 Oct 2022 (v1), last revised 31 May 2023.* Published in 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. 

# Gaussian Amnesiac Unlearning

- In this method, the label of a sensitive data is replaced with an incorrect label.

- The incorrect labels are sampled from a Gaussian distribution instead of random assignment.

$$
\begin{tikzcd}
& \mathbf{D} \arrow[dl] \arrow[dr] & & \\
D_r \arrow[dd] & & D_f \arrow[dr] & \\
& & & (x_i, y_i) \equiv (x_i, y_i') \arrow[dl] \\
D_r \arrow[dr] & & D'_f \arrow[dl] & \\
& D' & &
\end{tikzcd}
$$



**Source**: **Deep Regression Unlearning**, Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli; *[Submitted on 15 Oct 2022 (v1), last revised 31 May 2023.* Published in 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

# Gaussian Amnesiac (Cont.)

- $M(.,\emptyset)$ ← Pre-Trained Model
- $D$ ← Original Dataset
- $D'$ ← Modified Dataset with wrong Forget Labels

$$
\begin{aligned}
& for\ i = 1,2,3,\ldots,n \\
& \quad for\ (x_i, y_i) \in D' \\
& \qquad y_i^{pred} = M(x_i, \emptyset) \\
& \qquad L_M = L\left(y_i^{pred}, y_i\right) \\
& \qquad \emptyset = \emptyset - \alpha \frac{\partial L_M}{\partial \emptyset}
\end{aligned}
$$



Source: Deep Regression Unlearning, Ayush K Tarun, Vikram S Chundawat, Murari Mandal, Mohan Kankanhalli; [Submitted on 15 Oct 2022 (v1), last revised 31 May 2023. Published in 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. 

# Results and Findings

| *Forget Set* | Metric | Original | Retrained | Gaussian Amnesiac | BlindSpot |
|---|---|---|---|---|---|
| 0 - 30 | **W_Distance(1)** | 6.3929 | - | 0.6148 | 3.4921 |
| 60 - 101 | **W_Distance(1)** | 8.7661 | - | 1.6962 | 1.9200 |

To measure the similarity between output distributions of different models we will use **1ST WASSERSTEIN DISTANCE**

$$W_1(p,q) = \inf_{\gamma \in \Gamma(p,q)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \, d\gamma(x,y)$$



Ramdas, A., Garc´ıa Trillos, N., and Cuturi, M. *On wasserstein two-sample testing and related families of nonparametric tests. Entropy*, 19(2):47, 2017.

# Results and Findings

**Exact Unlearned Model (Trained on 100 epochs)**

Evaluating Exact Unlearn Model on Retain Data : {Loss : 9.8752}
Evaluating Exact Unlearn Model on Forget Data : {Loss : 20.4281}

**G-A Unlearned Model (Trained on 5 epochs)**

Evaluating G-A Unlearned Model on Retain Data : {Loss : 9.6447}
Evaluating G-A Unlearned Model on Forget Data : {Loss : 20.78454}

**BLSP Unlearned Model (Trained on 2 + 5 epochs)**

Evaluating Blspt Unlearn Model on Retain Data : {Loss : 9.8752}
Evaluating Blspt Unlearn Model on Forget Data : {Loss : 18.85484}

# Reference


| | |
|---|---|
| 1 | Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In 2021 *IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021. |
| 2 | Brophy, J. and Lowd, D. Machine unlearning for random forests. In *International Conference on Machine Learning*, pp. 1092– 1104. PMLR, 2021. |
| 3 | Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In 2015 *IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015. |
| 4 | Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a. |
| 5 | Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023b. |
| 6 | Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. Deep Regression Unlearning. *40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202*, 2023 |

# Thank You For Listening


**Anirban Dey**

M.Sc. Big Data Analytics, RKMVERI, Belur

**Sourish Ghosh**

M.Sc. Big Data Analytics, RKMVERI, Belur