Machine Unlearning

Presented By:
Dhruv Panchal (202411042)
Manish Prajapati(202411012)
Rushali Shah (202411061)
Priyanshi (202411009)

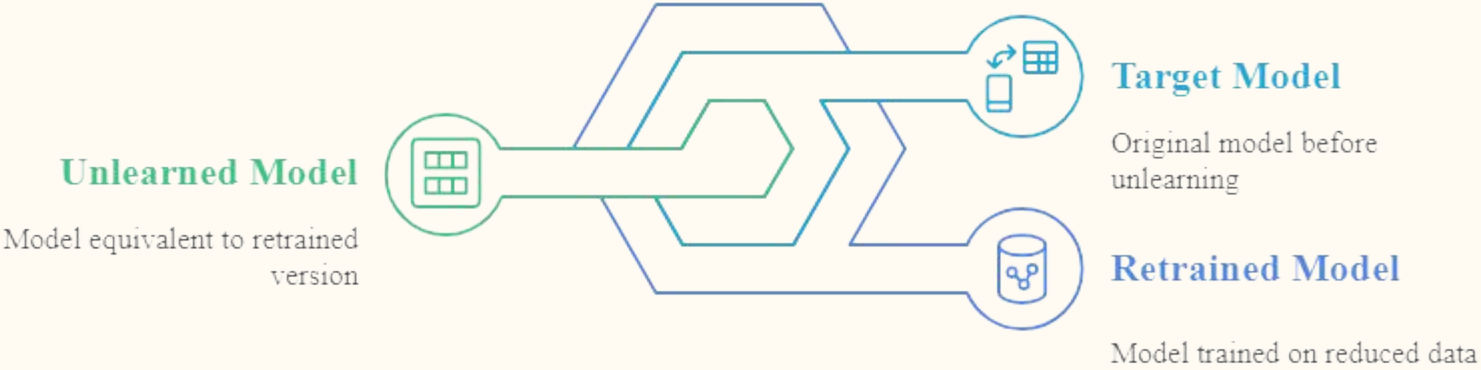
Presented to:
Prof. Rachit Chhaya

TABLE OF *Contents*

What is Machine UnLearning?

- Machine unlearning can be broadly described as removing the influences of training data from a trained model.

- At its core, unlearning on a *target model* seeks to produce an *unlearned model* that is equivalent to—or at least "behaves like"—a *retrained model* that is trained on the same data of target model, minus the information to be unlearned.



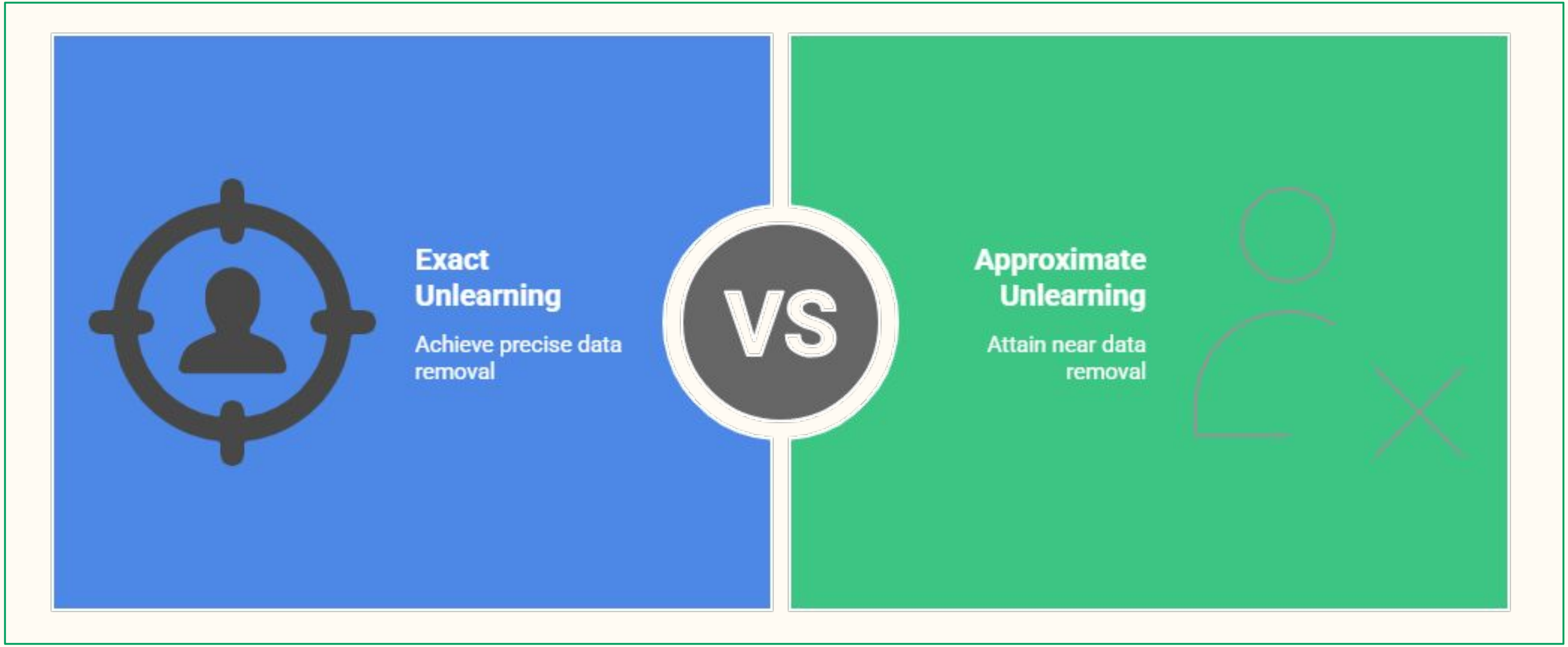
Importance of Machine Unlearning

- **Security**: Adversarial data can poison models—especially dangerous in healthcare. Machine unlearning helps detect and erase it fast.
- **Privacy**: Laws like GDPR protect against data leaks and accidental exposure (e.g., DNA markers). Users want control.
- **Usability**: Systems that remember wrong data create annoying, outdated recommendations. Forgetting is key to good UX.
- **Fidelity**: Biased training data skews AI fairness. Unlearning flawed data helps fight discrimination and build trust

Importance of Machine Unlearning

- In a world where **trust is everything**, the ability to forget isn’t a flaw—it’s a **feature**.
- From **security** to **fairness**, giving users control over their data isn't just ethical—it's essential.
- Systems that can’t forget risk being the ones we choose to forget.

Types of Machine Unlearning Algorithms



Exact Unlearning

- **Goal**: Make the model behave as if the target data was **never used** during training.

- **Definition**: The final model should be **statistically identical** to one trained from scratch **without** the removed data.

- **How It Works**:
  1. Most direct method: **Full retraining** on the remaining data.
  2. Guarantees **clean and provable removal** of the data's influence.

Approximate Unlearning

- **Goal**: Make the model **behave similarly** to if the data had never been seen—**without full retraining**.

- **Definition**: Efficiently estimate and reduce the data's influence using **lightweight model updates**.

- **How It Works**:
  1. Techniques like **fine-tuning** on the remaining data.
  2. Uses tools like **influence functions** to reverse learned effects.

Trade Offs: Exact Vs Approximate

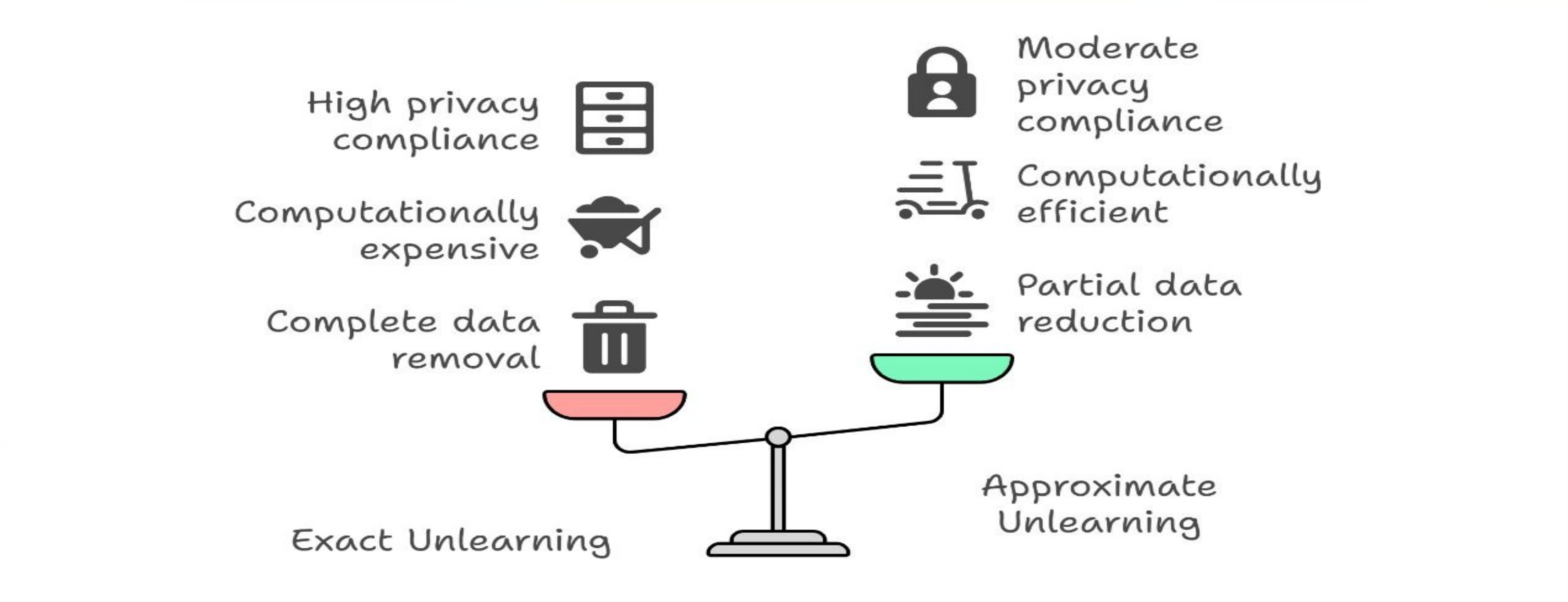
- **Depends on the Use Case**:

  Exact Unlearning: Best for **high-stakes domains**—medical, legal, finance. **Approximate Unlearning**: Ideal for **large-scale, fast-changing systems**—e.g., recommendation engines.

- **Trade-Off**:

  Exact = High **certainty**, low **efficiency**. **Approximate** = High **efficiency**, low **guarantees**

Conclusion



High privacy compliance
Computationally expensive
Complete data removal
Exact Unlearning
Moderate privacy compliance
Computationally efficient
Partial data reduction
Approximate Unlearning

Naive Unlearning: Simple but Slow

- A method to remove a data point’s influence by retraining the entire model from scratch.

- **Pros** of Naive Unlearning:
  1. Complete Isolation
  2. Model Consistency
  3. Simplicity
  4. Universality

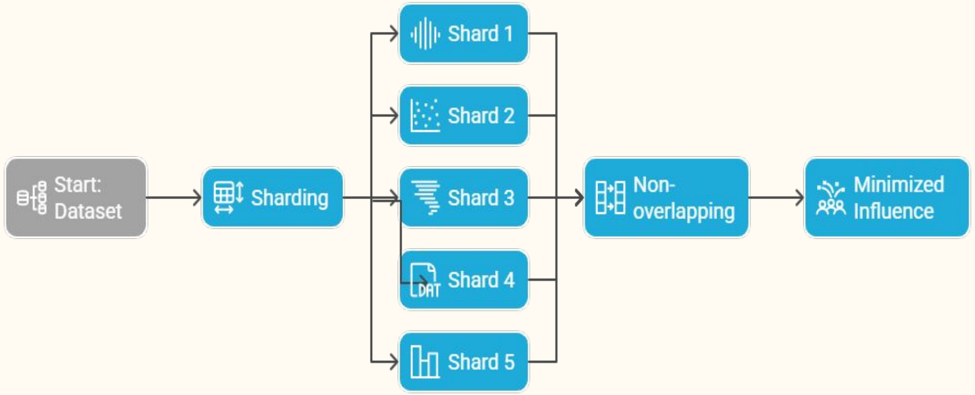
- **Cons** of Naive Unlearning:
  1. High Resource Demand
  2. Scalability Issues

SISA: A Game-Changer for Unlearning

- Stands for: Sharded, Isolated, Sliced, Aggregated.
- Inspired by ensemble learning and distributed training.
- **Key Idea:** Reorganize dataset to minimize retraining scope.
- Applicability: Works for any incrementally trained model (e.g., deep neural networks via gradient descent).

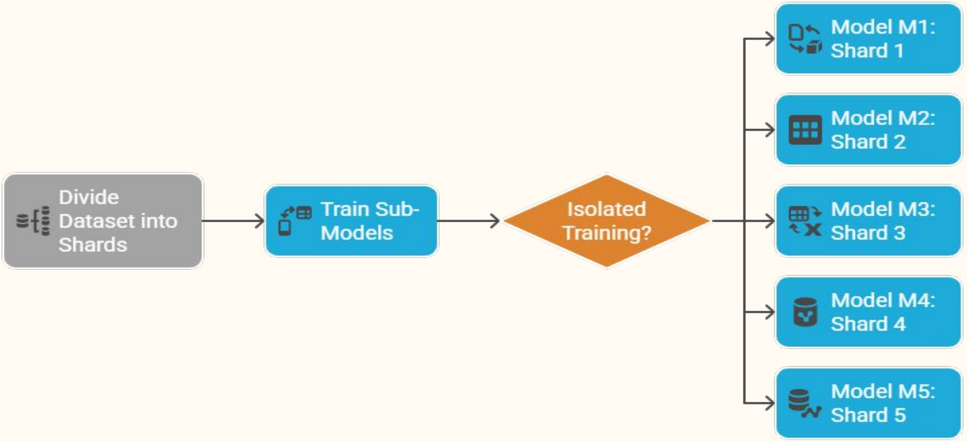
Step 1 – Sharding

- What: Split dataset (e.g., 100 photos) into equal-sized, non-overlapping shards (e.g., 5 shards of 20 photos).
- Why: Limits each data point's influence to one shard (unlike traditional ensembles where data may overlap).
- Example: Photo #5 resides only in shard-1.



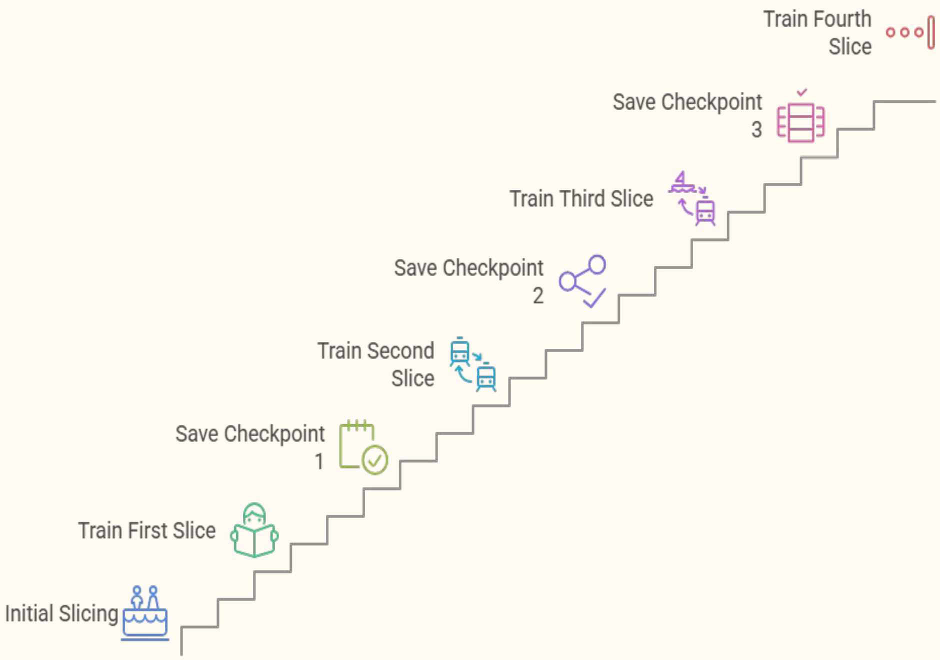
Step 2 – Isolation

- What: Train a separate model for each shard (e.g., M1 for shard 1, M2 for shard 2, ..., M5).
- How: Each model trains only on its shard's data—no data sharing.
- Benefit: Photo #5 affects only M1; M2–M5 remain untouched.



Step 3 – Slicing

- What: Subdivide each shard into slices (e.g., shard 1's 20 photos into 4 slices of 5).
- How: Train incrementally slice by slice, saving model checkpoints after each.
- Why: Enables restarting from checkpoint before slice containing photo #5, skipping it.
- SISA without slicing still works for all models (e.g., decision trees).

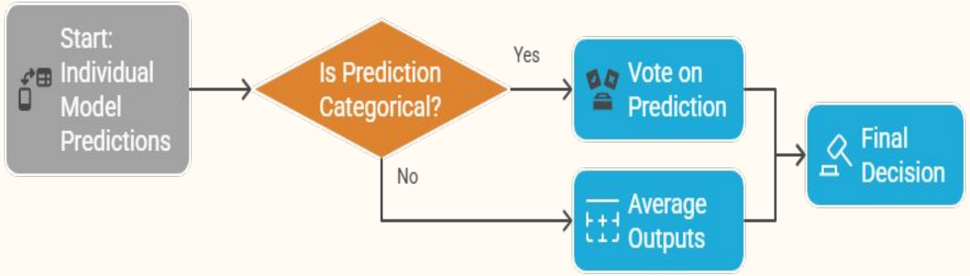


Step 4 – Aggregation

- What: Combine sub-model predictions for final output.

- Methods:
    a. Classification: Majority vote (e.g., M1–M5 vote “cat” or “dog”).
    b. Regression: Average outputs (e.g., price predictions).

- Benefit: Robust, collective decision (comparable accuracy to single model).



Unlearning with SISA

- Process for Deleting Photo #5:
    a. Locate shard (e.g., shard 1).
    b. Retrain M1:
        i. Without slicing: Retrain M1 from scratch, excluding photo #5.
        ii. With slicing: Start from checkpoint before photo #5's slice, skip it.
    c. Update ensemble with new M1.
- Result: Model forgets photo #5 with minimal retraining.

Challenges

1. The spectrum of unlearning hardness
2. Copyright protection
3. Retrieval-based AI systems
4. AI safety

The spectrum of unlearning hardnesses

The difficulty of unlearning lies on a **spectrum**:

1. **Easiest to unlearn**: Rare, isolated facts (e.g., a minor car accident in Palo Alto) that have minimal connections to other knowledge in the model.

2. **Harder to unlearn**: Common but not foundational facts (e.g., "Biden is the US president"), which are referenced more widely.

3. **Most difficult to unlearn**: Fundamental truths (e.g., "the sun rises every day"), which are deeply embedded in the model's structure and interconnected with many other facts.

Ultimately, **memorization, forgetting, and unlearning are deeply interconnected** but not yet fully understood.

Copyright Protection

1. **Goal**: Unlearning aims to remove memorized copyrighted content from models.
2. **Challenge**: Current techniques lack strong guarantees; legal standards are still evolving.
3. **Alternatives**: Prompt controls, moderation tools, and alignment strategies (e.g., OpenAI's approach) reduce risk.
4. **Other Solutions**: Economic options like retraining or indemnification (e.g., OpenAI's "Copyright Shield") offer protection.

While unlearning seems promising, it needs legally enforceable mechanisms to become a viable solution."

Retrieval-based AI Systems

1. **How It Works**: Sensitive content is stored externally and retrieved when needed—making deletion easy via database removal.
2. **Challenges**: Paraphrased or summarized copyrighted content can evade detection. Style and preferences aren't easily retrieved.
3. **Security Risks**: Supplying protected content during inference may expose vulnerabilities.
4. **Trade-off**: Retrieval reduces memorization risks but can’t fully replace the depth and performance of trained models.

AI Safety

1. **Purpose**: Unlearning is being explored to remove dangerous knowledge, behaviors, or capabilities.

2. **Complexity**: Model poisons, backdoors, and biases are hard to detect and even harder to erase.

3. **Abstract Risks**: Traits like power-seeking aren't tied to specific data—making them tough to unlearn.

4. **Cost**: Even when targets are clear, unlearning is slow, expensive, and may reduce model performance.

Conclusion

- ➔ Machine unlearning is emerging as a vital component of responsible AI, but it remains a complex challenge spanning technical, legal, and ethical fronts.
- ➔ To truly harness its potential, unlearning must be built into AI systems from the start—enabling models to adapt, respect user rights, and navigate evolving privacy and safety demands.
- ➔ A proactive approach will be essential to ensuring AI remains trustworthy, fair, and future-ready.

References

[1] An Introduction to Machine Unlearning

[2] Machine Unlearning: Solutions and Challenges

[3] A Survey of Machine Unlearning

[4] Machine Unlearning in 2024 - by Ken Ziyu Liu

[5] ARCANE: An Efficient Architecture for Exact Machine Unlearning

THANK YOU!