Label-Agnostic Forgetting: A Supervision-Free Unlearning in Deep Models

Shaofei Shen[1], Chenhao Zhang[1], Yawen Zhao[1], Weitong Chen[2],

Alina Bialkowski [1], and Miao Xu[1]

The University of Queensland [1], University of Adelaide[2]

Machine Unlearning

❑ Why we need machine unlearning?

Right-to-Be-Forgotten: the right to have personal data deleted from the model [1]

❑ What is required?

✓ A well-trained machine learning model $g_D = g_D^e \circ g_D^c$

✓ Specific data that is subject to removal requests – forgetting data $D_f = (X_f, Y_f)$

✓ Other data not affected by privacy requirements – remaining data$D_r = (X_r, Y_r)$

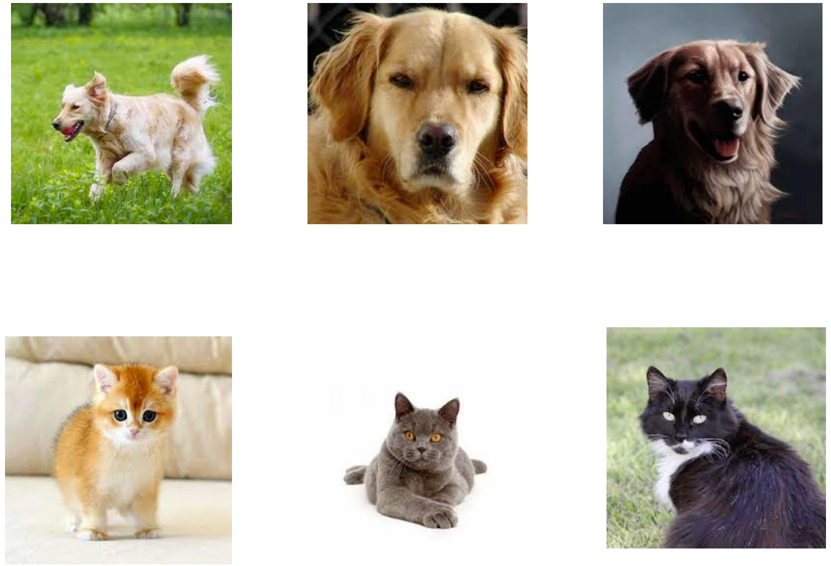
❑ What is expected results?

❖ An equivalent model $g_U = g_U^e \circ g_U^c$ to the retrained model on remaining data

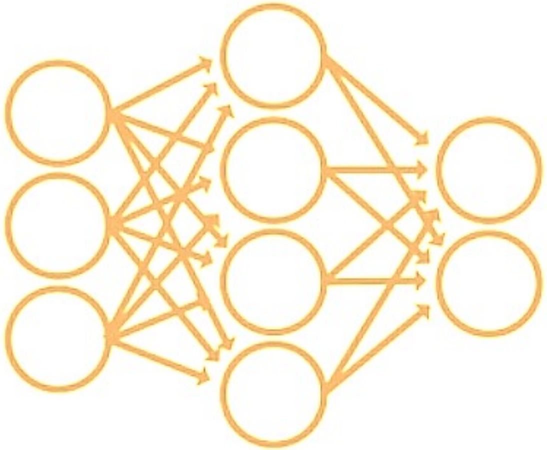
[1]. https://gdpr.eu/right-to-be-forgotten/.

Implementing the Machine Unlearning

We have authorized enterprise to use our data to train a classification model:



Training

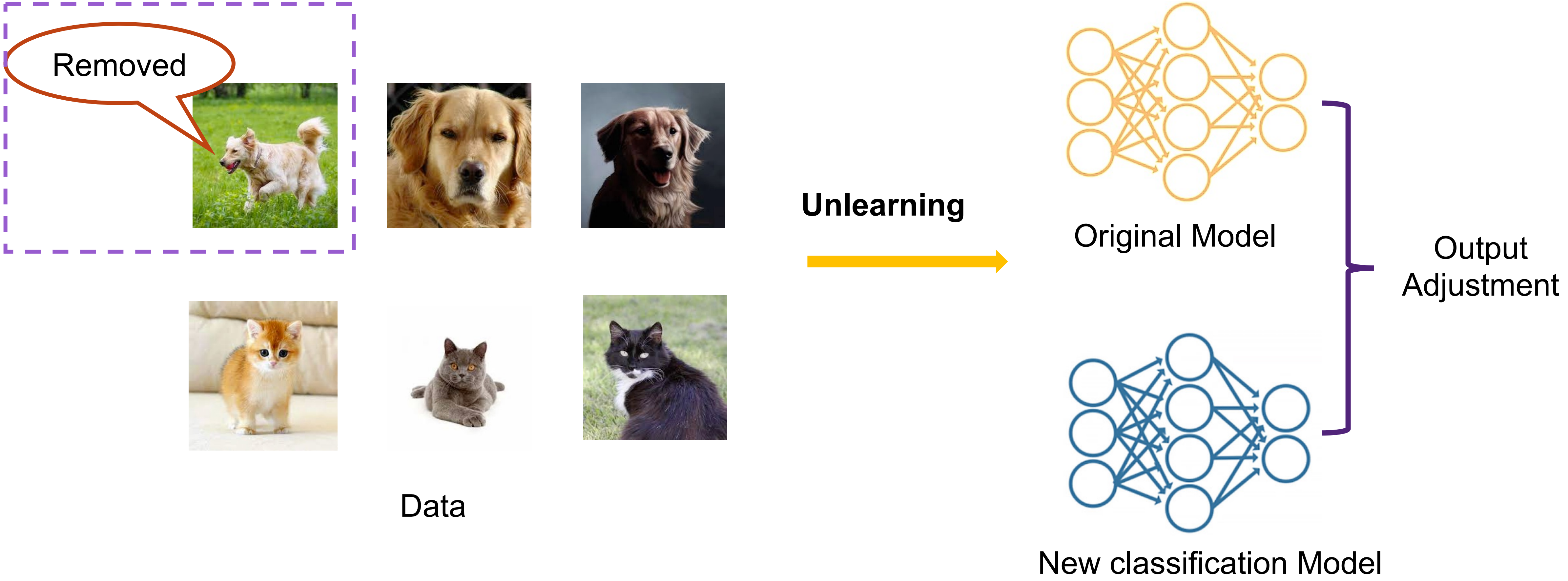


Data

Classification Model

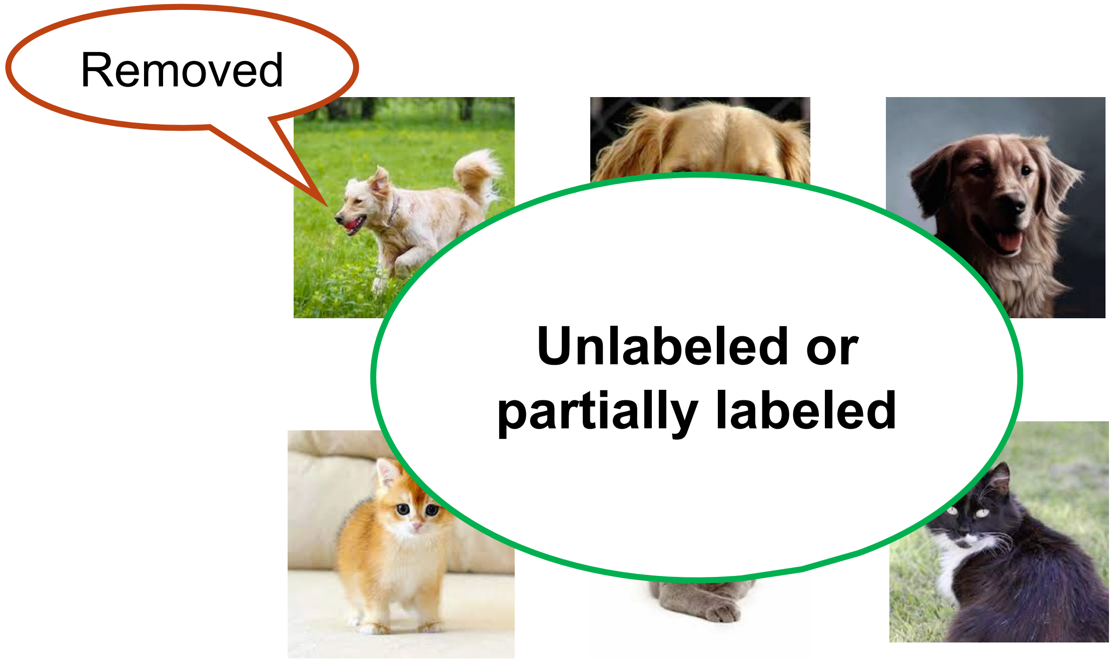
Implementing the Machine Unlearning

If we do not want our data to be used for the model and cancel authorization:



A Question on Machine Unlearning

How can we realize unlearning without data labels?

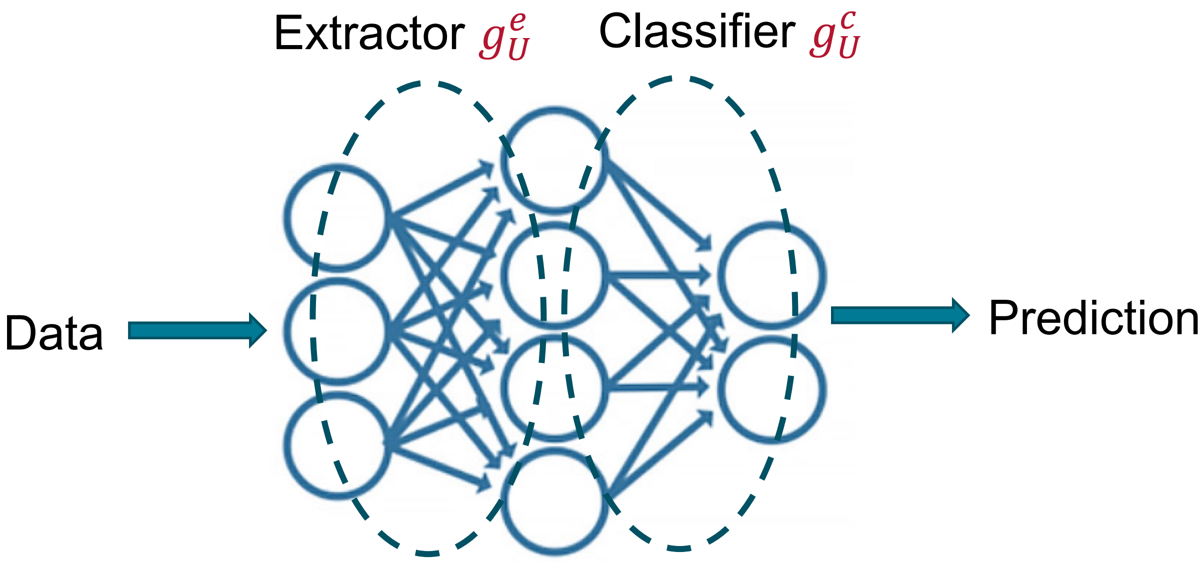


Pet Images

- **Issues for unlearning without label information:**
  - Real-world datasets are not fully labelled
  - Model is learned in weakly supervised learning scenarios
  - Labels should be withheld for privacy reasons during unlearning

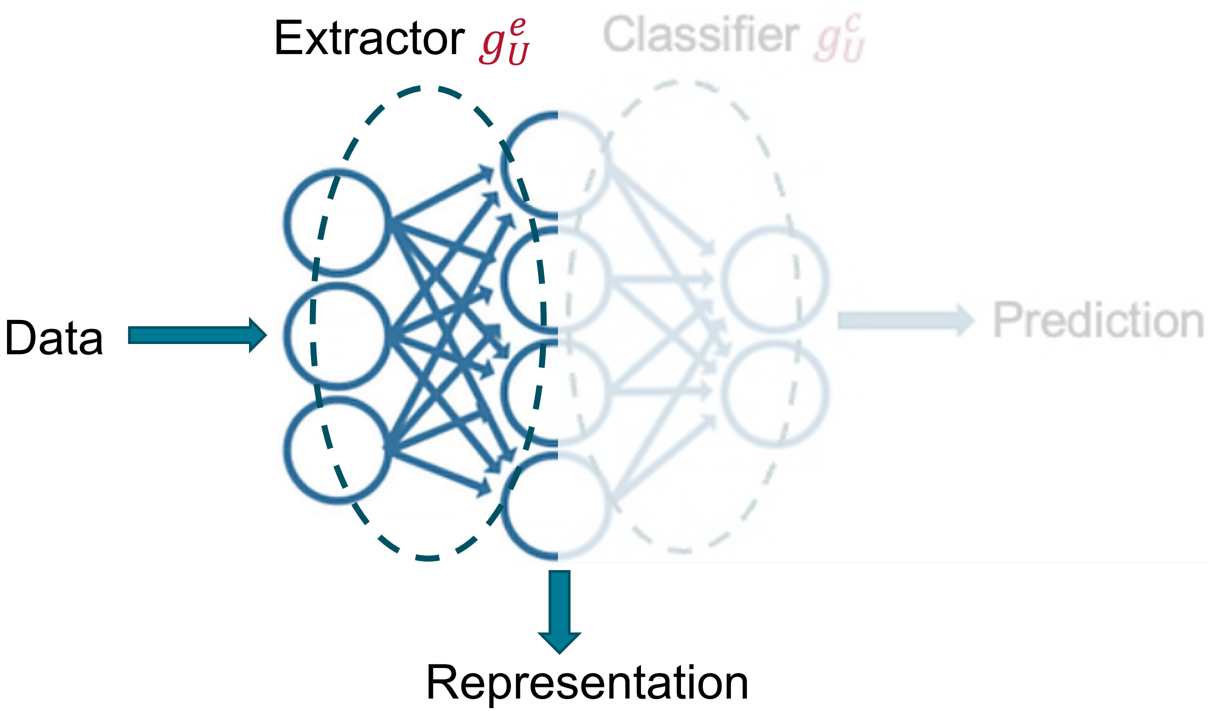
Representation-level Unlearning

- **How to reduce the usage of labels in unlearning?**
  - Representation-level adjustment
  - Differentiate forgetting data's representation and preserve remaining data's representation



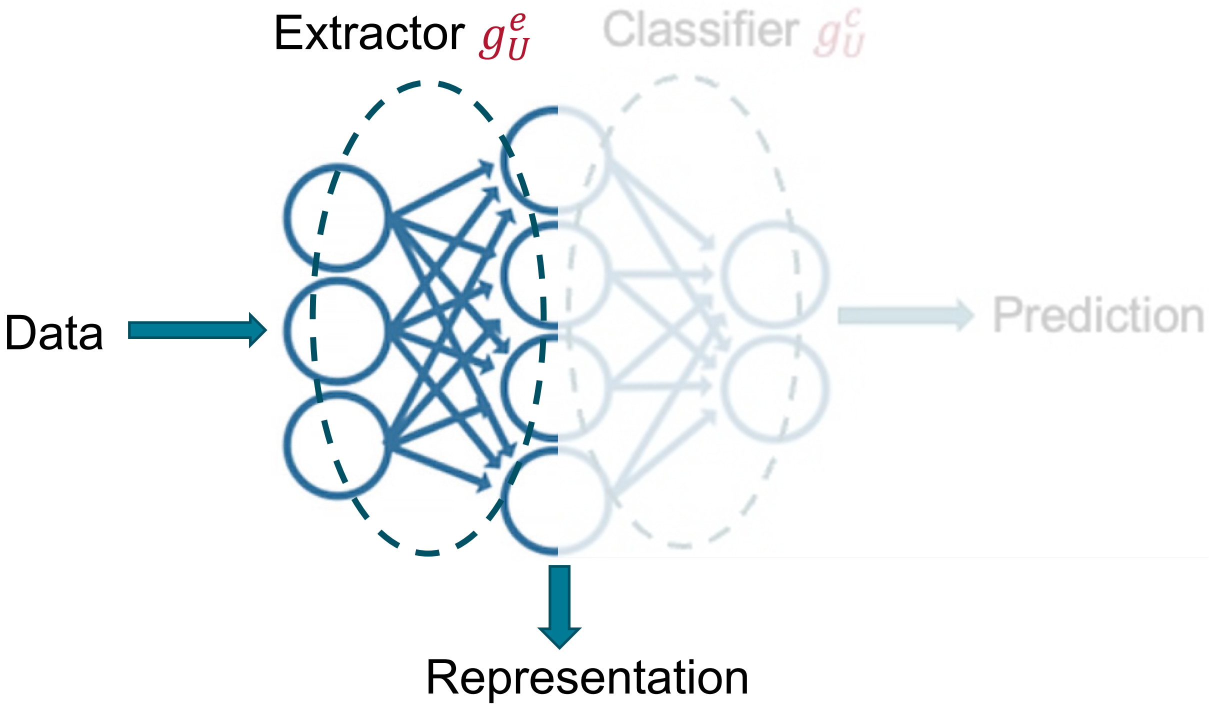
Representation-level Unlearning

- **How to reduce the usage of labels in unlearning?**
  - Representation-level adjustment
  - Differentiate forgetting data's representation and preserve remaining data's representation



Representation-level Unlearning

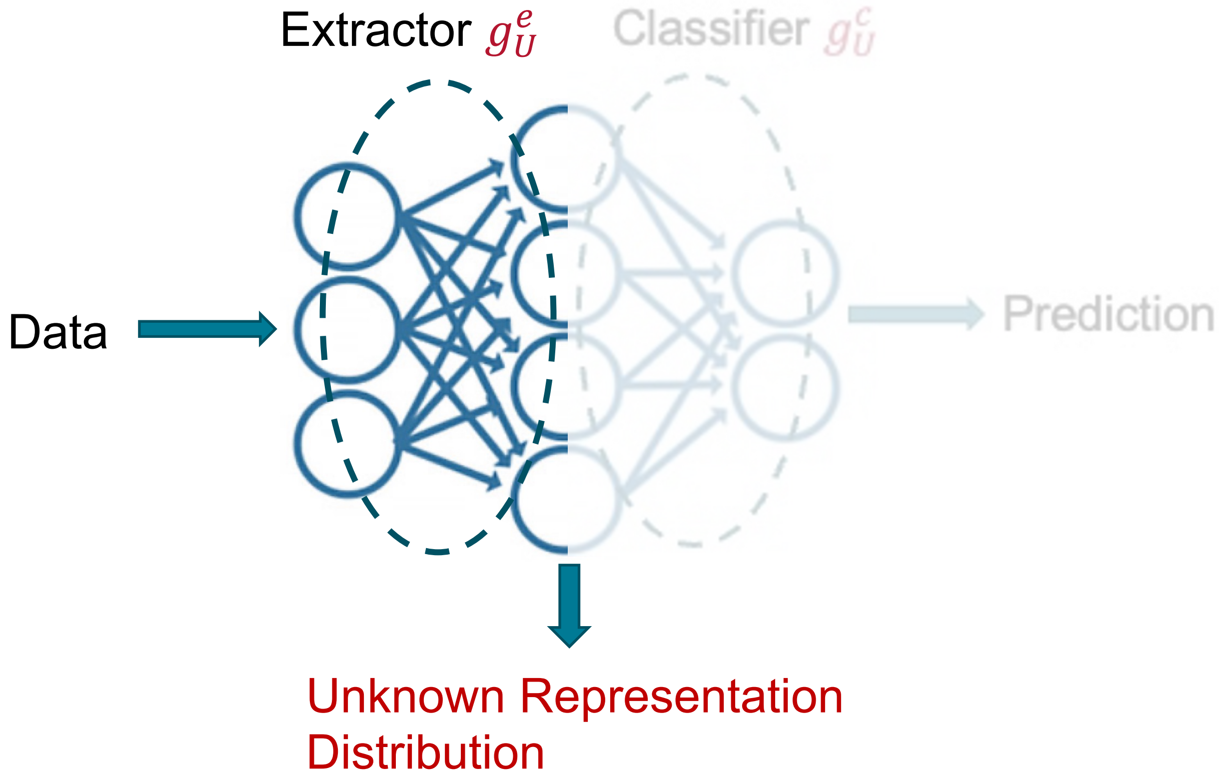
- ❑ **How to reduce the usage of labels in unlearning?**
  - ➢ Representation-level adjustment
  - ➢ Differentiate forgetting data’s representation and preserve remaining data’s representation



[Label-Agnostic Forgetting: A Supervision-Free Unlearning in Deep Models] | [04/04/2024]

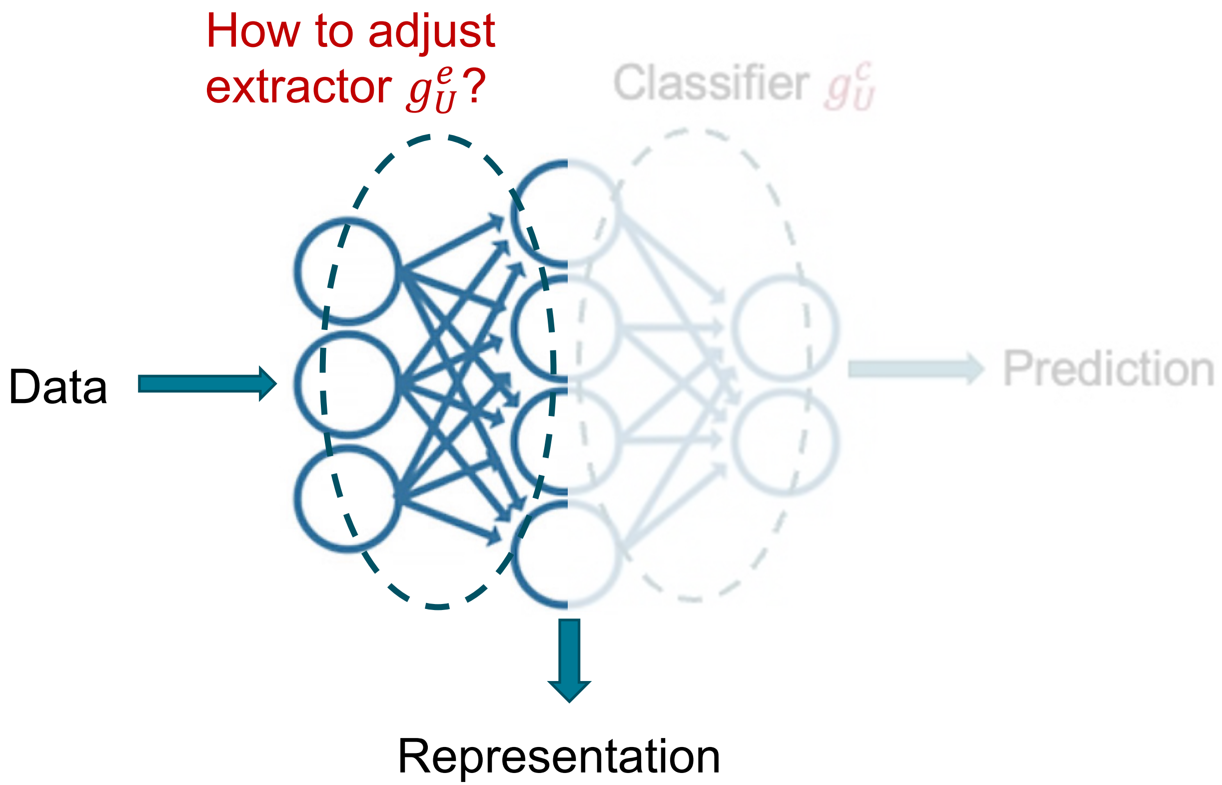
Representation-level Unlearning

- **Challenges on representation-level adjustment:**
  - Hard to estimate the representation knowledge
  - Lack of objective for representation-level unlearning
  - Representation adjustment will cause a misalignment with the classifier and fail in predictions



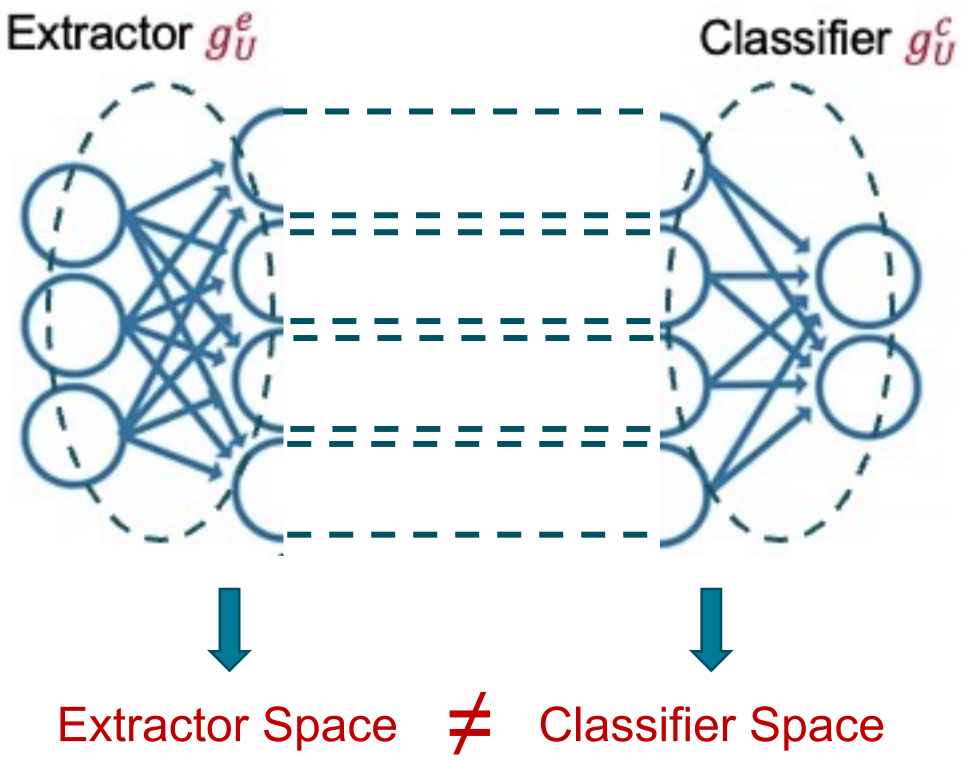
Representation-level Unlearning

- **Challenges on representation-level adjustment:**
  - Hard to estimate the representation knowledge
  - Lack of objective for representation-level unlearning
  - Representation adjustment will cause a misalignment with the classifier and fail in predictions

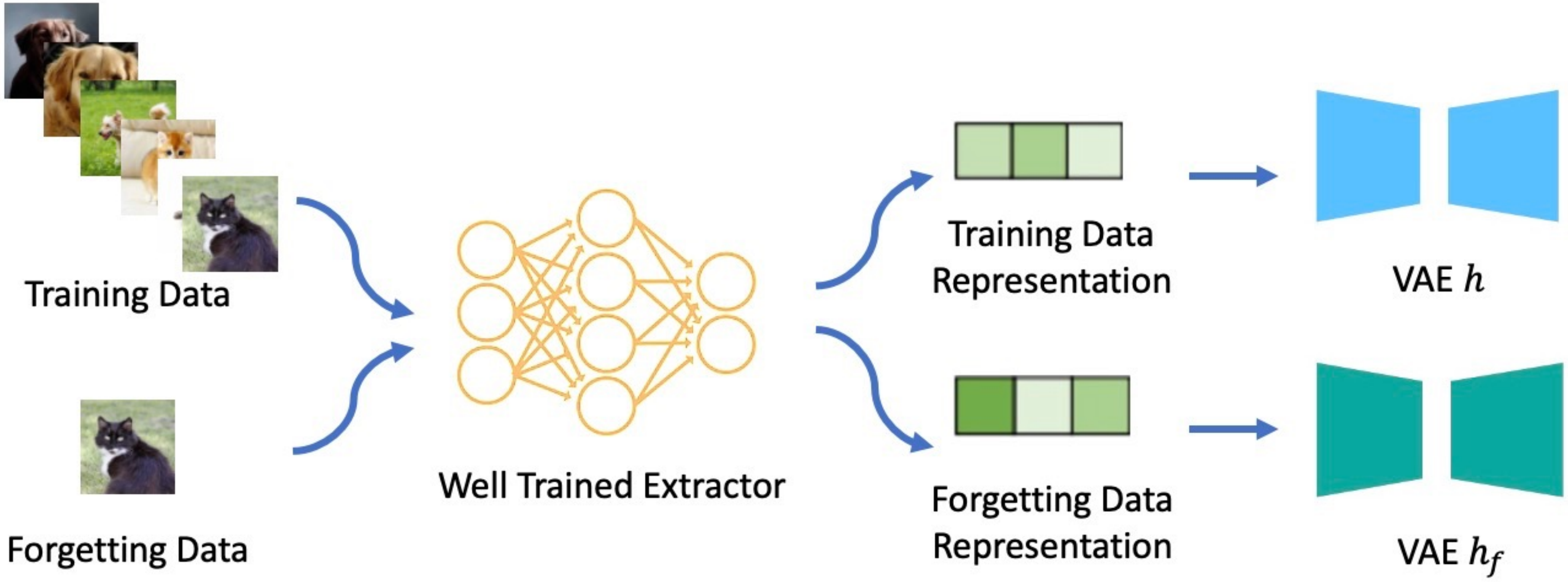


Representation-level Unlearning

- **Challenges on representation-level adjustment:**
  - Hard to estimate the representation knowledge
  - Lack of objective for representation-level unlearning
  - Representation adjustment will cause a misalignment with the classifier and fail in predictions

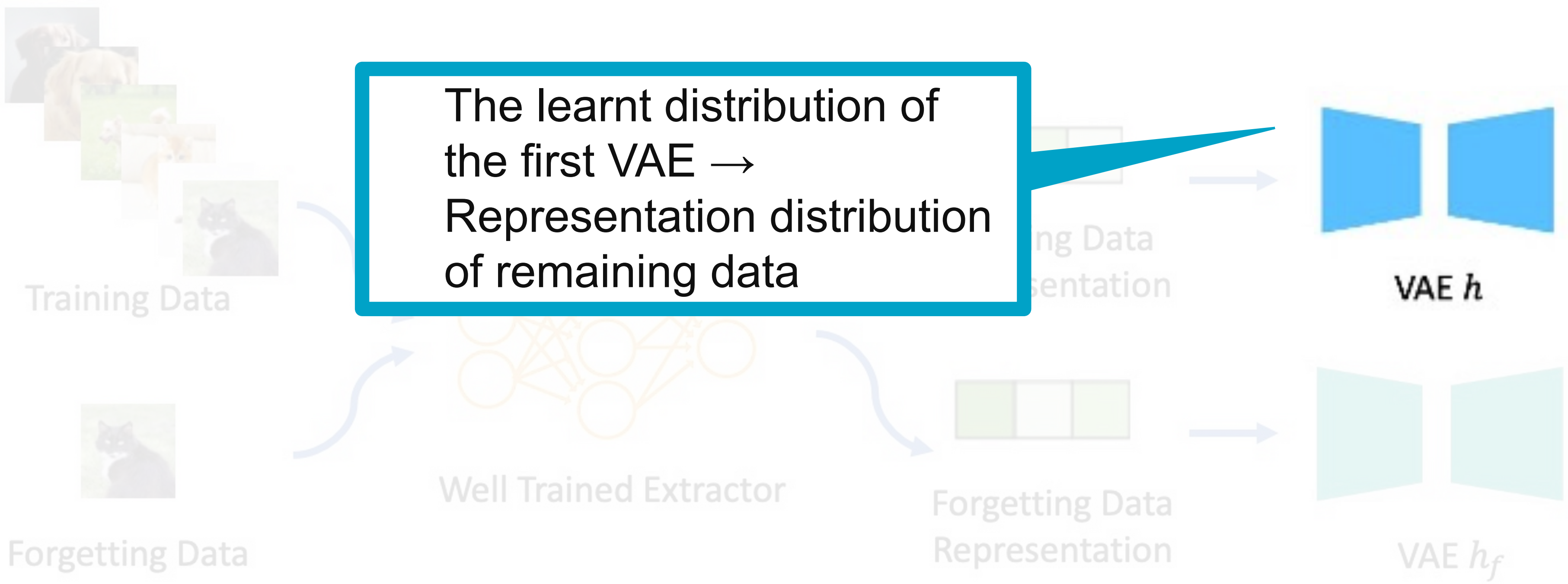


Estimating representation distribution

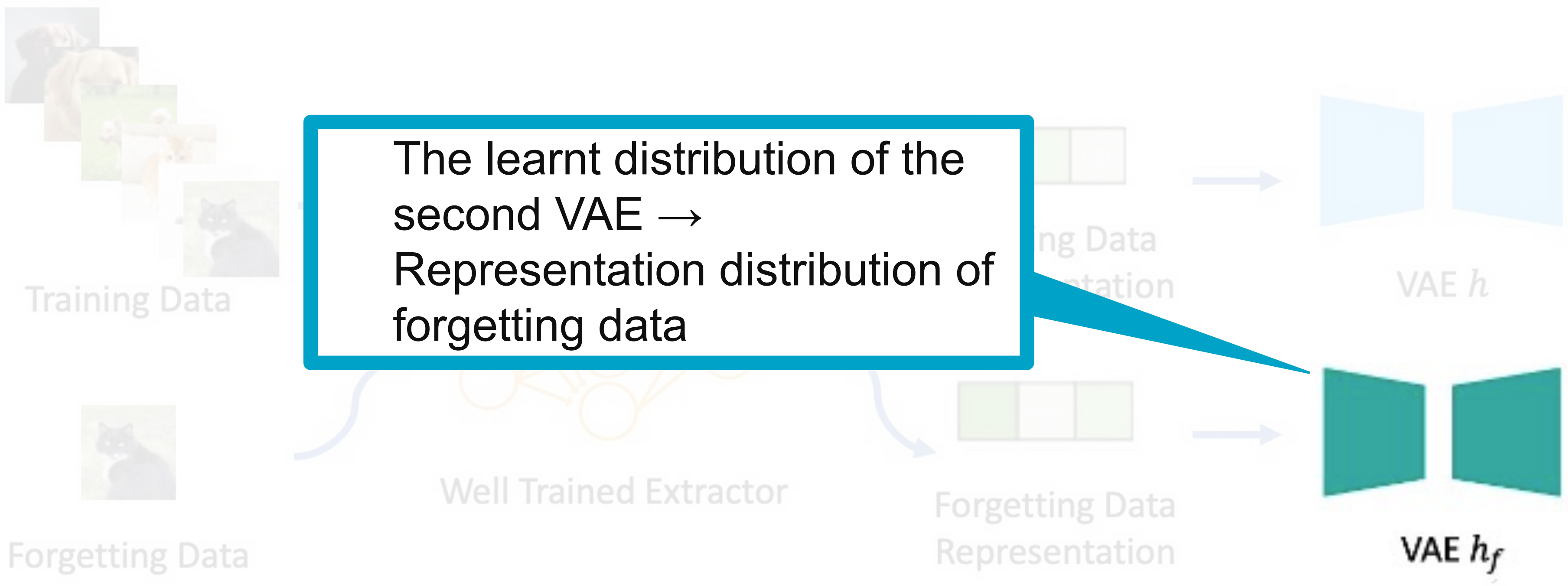


Estimating P_r and P_f through training two VAEs

Estimating representation distribution

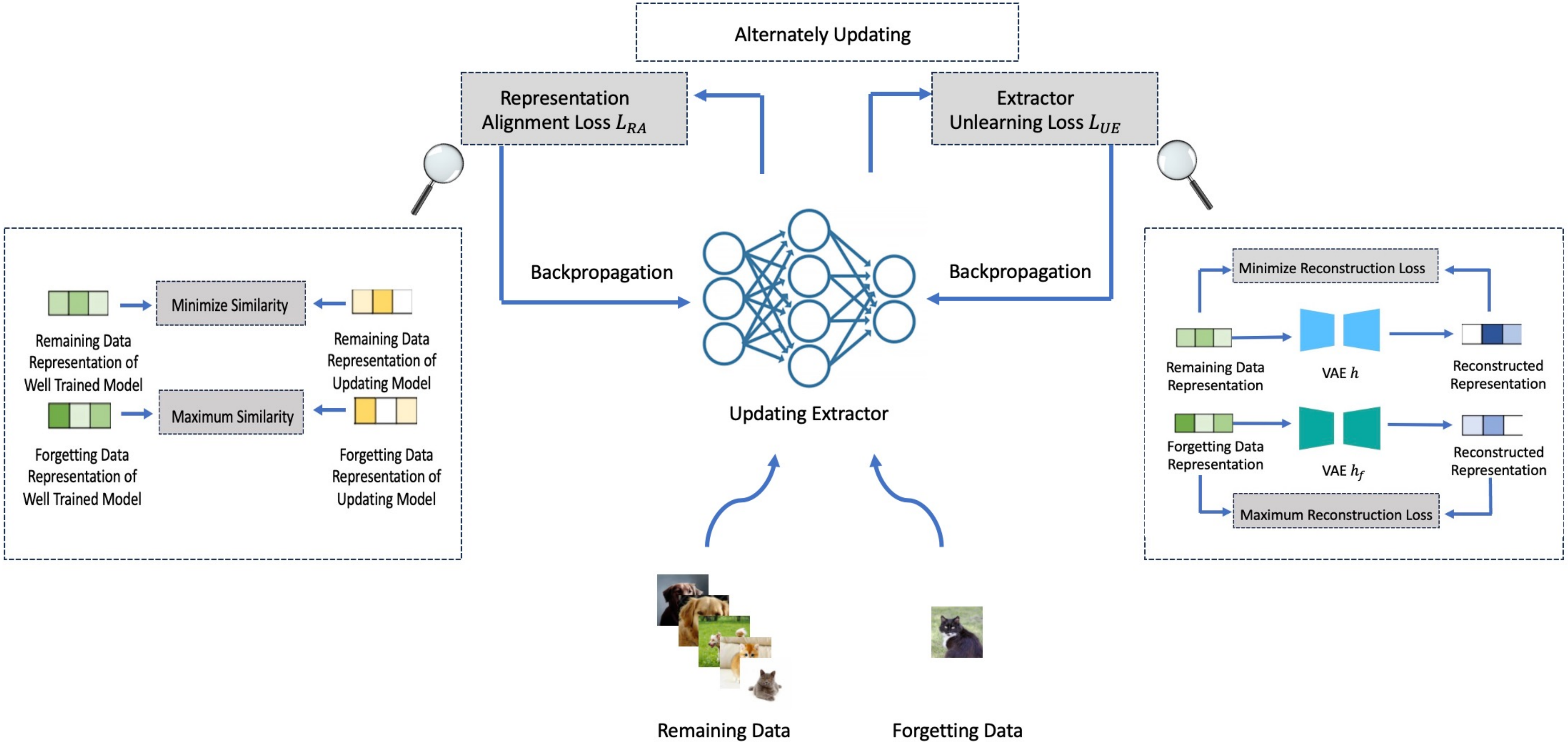


Estimating representation distribution



The learnt distribution of the second VAE → Representation distribution of forgetting data

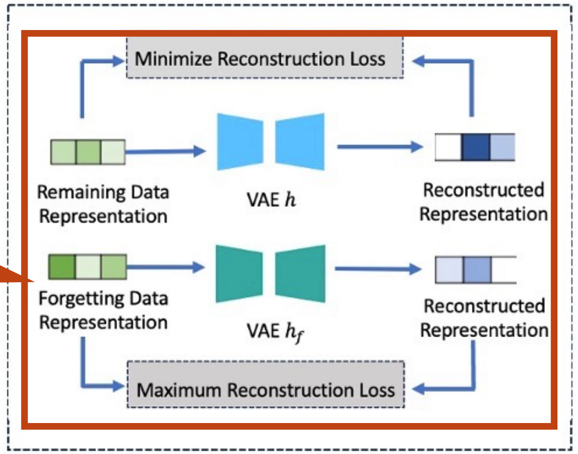
Label-Agnostic Unlearning



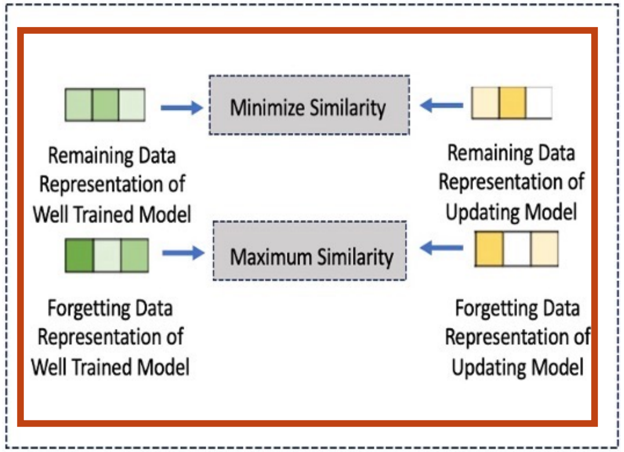
Representation-level unlearning

- ✓ For remaining data: align the representation distribution with the the distribution in VAE h
- ✓ For forgetting data: push away the representation distribution with the the distribution in VAE h_f
- ✓ Extractor Unlearning Loss:

$$L_{UE} = \sum_{x \in X_r} \frac{\|g_U^e(x) - h(g_U^e(x))\|_2^2}{\|g_U^e(x) - h(g_U^e(x))\|_2^2 + 1} - \sum_{x \in X_f} \frac{\|g_U^e(x) - h_f(g_U^e(x))\|_2^2}{\|g_U^e(x) - h_f(g_U^e(x))\|_2^2 + 1}$$



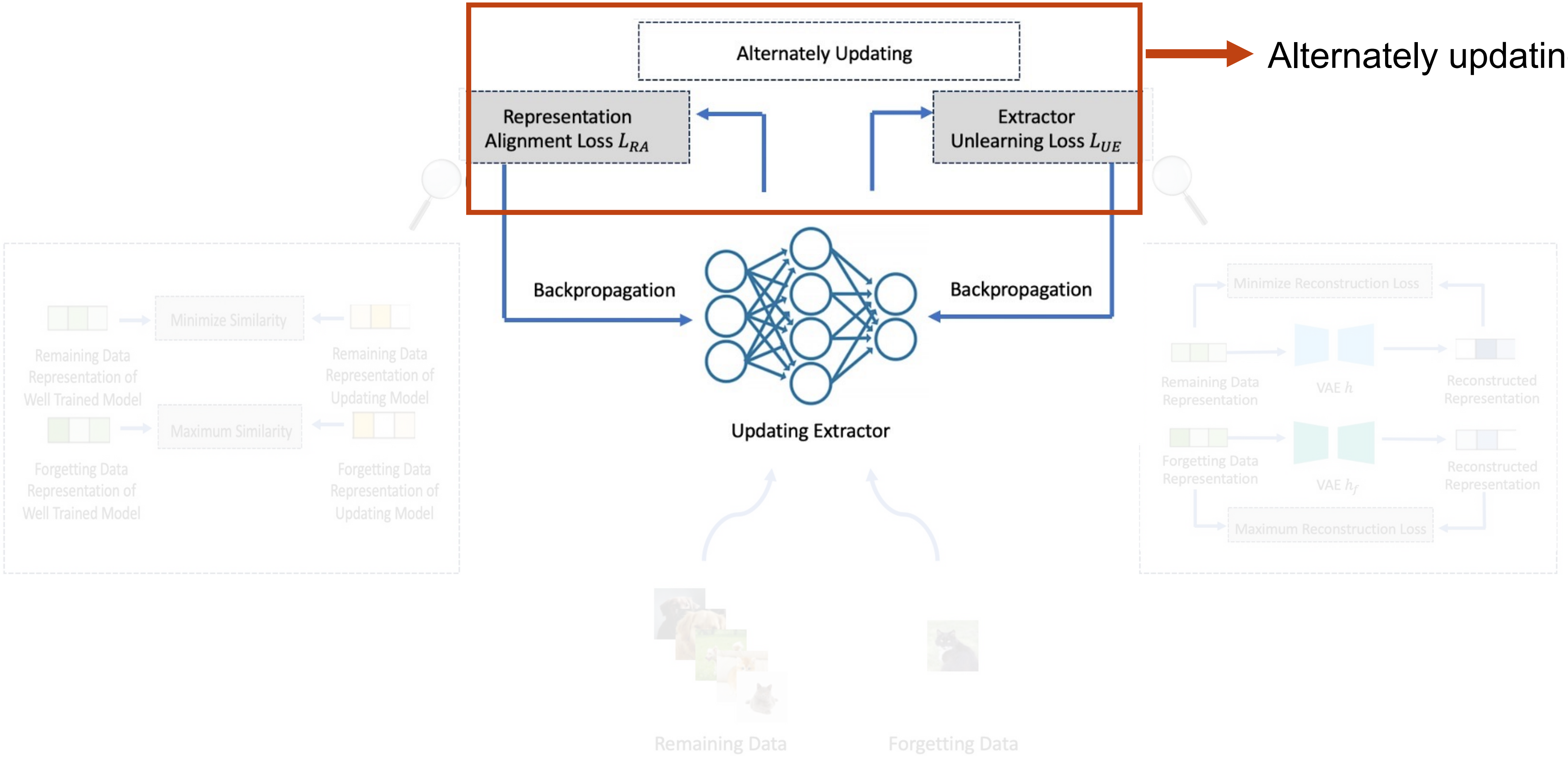
Representation Alignment



- ✓ For remaining data: minimize the representations from the updated and original extractor
- ✓ For forgetting data: maximize the representations from the updated and original extractor
- ✓ Representation Alignment Loss:

$$L_{RA} = \sum_{x \in X_r} \log\left(\frac{\exp(simloss(g_U^e(x), g_D^e(x)))}{\sum_{\hat{x} \in X_f} \exp(simloss(g_U^e(\hat{x}), g_D^e(\hat{x}))/\tau)}\right)$$

Label-Agnostic Unlearning



Alternately updating

Performance Comparison: Data Removal

Require labels

Label agnostic

Method	Data	R_{tr}	F_{tr}	T_s	ASR	Data	R_{tr}	F_{tr}	T_s	ASR
Retrain	Digit	99.56±0.05	98.84±0.10	99.04±0.10	49.80±0.53	Fashion	96.43±0.35	92.15±0.41	90.23±0.22	47.32±0.76
NegGrad		99.18±0.28	**98.86±0.41**	98.62±0.29	50.24±0.27		93.28±0.29	88.93±0.79	89.18±0.24	46.11±0.66
Boundary		97.65±1.02	95.36±2.50	96.63±1.35	46.83±2.09		56.28±4.69	46.58±4.04	53.00±3.66	48.03±1.41
SISA		99.06±0.12	98.60±0.07	98.92±0.02	33.78±0.01		91.98±0.19	90.76±0.07	89.92±0.24	33.33±0.02
Unroll		**99.63±0.15**	99.34±0.33	**99.08±0.18**	46.50±0.60		89.83±0.30	83.88±0.65	81.21±0.34	47.69±0.50
T-S		94.01±0.77	93.09±2.73	93.72±1.03	47.82±0.64		82.96±1.14	86.77±2.13	82.46±1.24	45.90±1.30
SCRUB		99.28±0.04	99.03±0.12	98.95±0.08	46.68±0.80		90.88±0.09	88.62±0.28	88.75±0.11	45.23±0.94
LAF+R		99.47±0.14	99.35±0.65	98.89±0.10	**49.42±0.51**		**94.18±0.30**	95.00±1.62	**90.51±0.28**	**47.39±0.23**
LAF		98.03±0.68	97.29±1.43	97.30±0.78	47.92±0.84		91.54±2.67	**90.91±7.00**	87.53±3.26	46.89±0.88
Retrain	C10	84.03±0.20	78.05±1.34	87.20±0.65	57.48±0.88	SVHN	83.88±0.23	75.16±0.76	93.41±0.40	58.76±0.48
NegGrad		79.08±0.55	70.50±2.94	83.51±0.97	56.53±0.34		81.57±0.34	69.93±1.66	91.54±1.01	57.94±0.80
Boundary		54.73±1.32	18.73±3.33	51.23±2.55	62.79±0.95		64.85±2.06	28.62±1.89	73.07±1.96	89.17±3.29
SISA		66.78±0.10	53.12±0.74	54.30±0.05	37.53±0.02		82.48±0.17	67.79±0.34	82.57±0.83	50.19±0.38
Unroll		57.82±1.66	30.91±2.86	61.31±1.51	56.97±1.27		70.98±1.87	47.68±2.72	83.27±0.48	55.39±0.98
T-S		70.31±2.32	72.17±3.91	77.71±2.02	54.64±1.58		78.36±0.13	73.50±0.62	90.60±0.61	55.77±1.42
SCRUB		29.16±1.07	0.47±0.93	25.18±0.78	54.03±0.64		22.32±0.04	0±0	19.59±0.07	65.26±1.24
LAF+R		**79.57±0.72**	**79.50±0.66**	**84.74±1.08**	57.74±0.62		**83.37±0.41**	**76.08±0.76**	**93.56±0.51**	**58.03±0.28**
LAF		78.03±1.55	73.30±3.96	82.22±2.57	**57.65±0.70**		81.63±0.49	76.11±1.49	92.32±0.58	57.85±0.89

- LAF consistently ranks within the top 5 performances in all evaluations
- LAF+R achieves either the best or second-best results in nearly all evaluations

Summary

- LAF is designed to address the research gap in label-agnostic unlearning

- LAF can accomplish mainstream unlearning tasks and retaining high predictive performance post-learning, all without the need for supervision information.

- LAF with supervised repairing (LAF+R) can achieve the leading performance in comparison to baseline methodologies.

- The experiments shed light on certain limitations of LAF, including the insufficient removal of the forgetting class in the class removal tasks, and the low efficiency

Thanks!