

# Thesis Committee Meeting

Ananth Mahadevan

November 1, 2023

# Overview

## 1 Research

- Published Articles
- Works in Progress

## 2 Studies

- Discipline Specific Credits
- Transferrable Skills

## 3 Activities

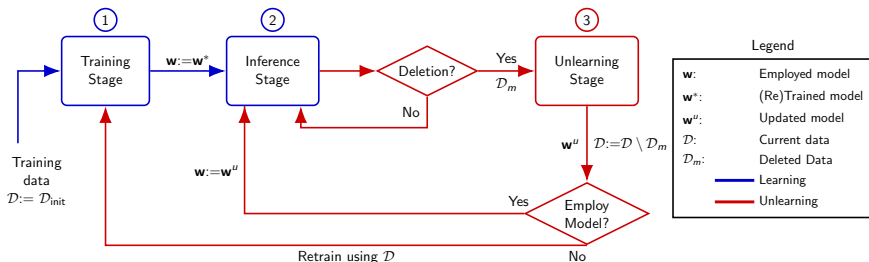
- Conferences, Workshops and Events
- Teaching and Supervision

## 4 Future Plans

## Published Articles

- 1 Certifiable Unlearning for logistic regression: An experimental study, MAKE 2022 [1]
- 2 Scalably Using Node Attributes and Graph Structure for Node Classification, Entropy 2022 [4]
- 3 Robustness of Sketched Linear Classifiers to Adversarial Attacks, CIKM 2022 [3]
- 4 Reception Reader: Exploring Text Reuse in Early Modern British Publications, JOHD 2023 [6]

# Machine Unlearning



## Unlearning

Task of updating a ML model after partial deletion of training data

Qualities of an approximate unlearning method:

- **Certifiability**: How similar are  $w^u$  and  $w^*$ ?
- **Effectiveness**: How well does  $w^u$  perform?
- **Efficiency**: How much time to produce  $w^u$ ?

# Experimental Setup

Item	Values
Parameters	Noise ( $\sigma$ ) & Efficiency ( $\tau$ )
Metrics	AccDis, AccErr & Speedup( $\times$ )
Methods	INFLUENCE, FISHER & DELTAGRAD

Metric	Quality
AccDis $\uparrow$	Certiability $\downarrow$
AccErr $\uparrow$	Effectiveness $\downarrow$

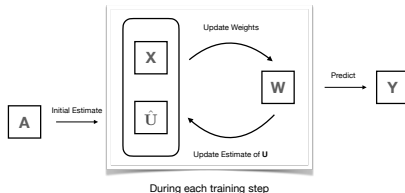
  

Parameter	Effect
$\tau \uparrow$	Efficiency $\uparrow$ Certiability $\downarrow$ Effectiveness $\downarrow$
$\sigma \uparrow$	Certiability $\uparrow$ Effectiveness $\downarrow$

## Published Articles

- 1 Certifiable Unlearning for logistic regression: An experimental study, MAKE 2022 [1]
- 2 Scalably Using Node Attributes and Graph Structure for Node Classification, Entropy 2022 [4]
- 3 Robustness of Sketched Linear Classifiers to Adversarial Attacks, CIKM 2022 [3]
- 4 Reception Reader: Exploring Text Reuse in Early Modern British Publications, JOHD 2023 [6]

## Scaling Node Classification [4]



- Improves running time of node classification method JANE [5]
- Developed mini-batching training algorithm to improve GPU utilization
- Scaled experiments to graphs with more than a million nodes

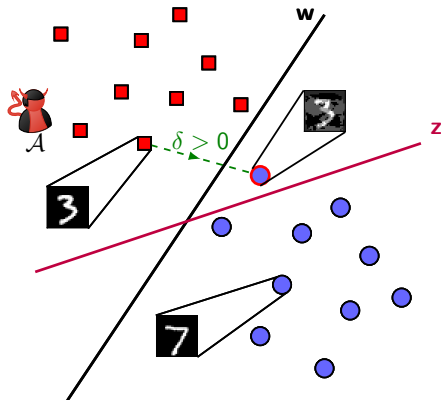
## Published Articles

- 1 Certifiable Unlearning for logistic regression: An experimental study, MAKE 2022 [1]
- 2 Scalably Using Node Attributes and Graph Structure for Node Classification, Entropy 2022 [4]
- 3 Robustness of Sketched Linear Classifiers to Adversarial Attacks, CIKM 2022 [3]
- 4 Reception Reader: Exploring Text Reuse in Early Modern British Publications, JOHD 2023 [6]



# Robustness of Sketched Linear Classifiers [1]

- $w$ : Linear Classifier
- $z$ : WM-SKETCH Classifier [7]
- $\mathcal{A}$ : Adversary
- $\delta$ : FGSM perturbation



# Robustness of Sketched Linear Classifiers [1]

$\mathcal{A}$  crafts a perturbation based on its Knowledge and Observability of the target.

- **Knowledge**: Is the target a WM-SKETCH?
- **Observability**: Does  $\mathcal{A}$  have access to count sketch  $R$ ?

$\mathcal{A}$	Knowledge	Observability
White-Box	✓	✓
Grey-Box	✓	✗
Black-Box	✗	✗

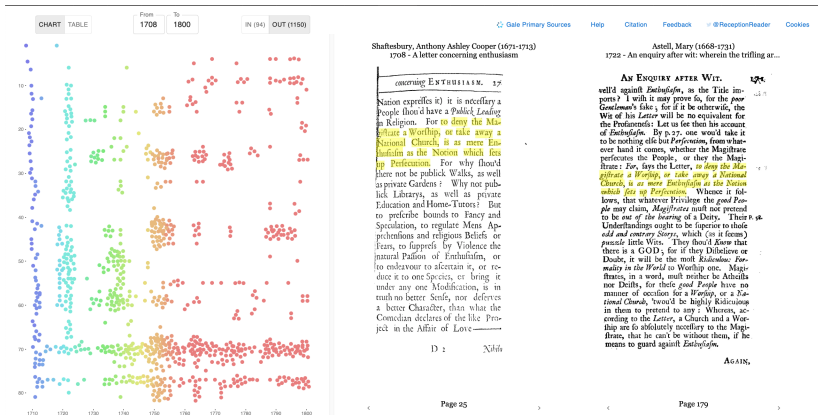
## Research Question

How does the performance of a given WM-SKETCH worsen in the presence of these adversaries?

## Published Articles

- ① Certifiable Unlearning for logistic regression: An experimental study, MAKE 2022 [1]
- ② Scalably Using Node Attributes and Graph Structure for Node Classification, Entropy 2022 [4]
- ③ Robustness of Sketched Linear Classifiers to Adversarial Attacks, CIKM 2022 [3]
- ④ Reception Reader: Exploring Text Reuse in Early Modern British Publications, JOHD 2023 [6]

# Reception Reader Interface



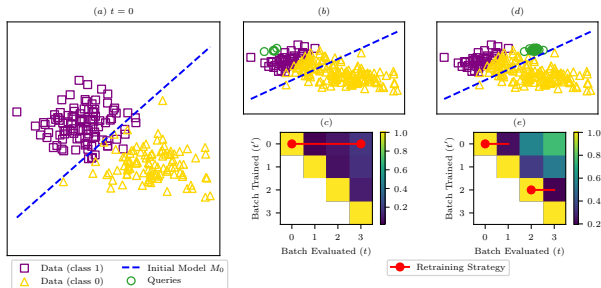
# Historical Text Reuse

- 250k scanned English documents from 18th century
- Fuzzy string matching using BLAST to find reuses
- De-fragmenting and clustering of reuse pairs
- Front-end for navigation

# In Progress

- ① Cost-Aware Retraining for Machine Learning
  - Rejections from ICML 2023 and EuroSys 2023
  - Submitted pre-print to Arxiv [2]
  - Re-submitted to Knowledge-Based Systems Journal
- ② Data Science Pipeline for Historical Text Reuse
  - Manuscript under preparation
  - Planned submission to VLDB 2024 Scalable Data Science track
- ③ Scalable Constraint-Based Diversity Sampling
  - Improved running time of Fair Max-Min Diversification [8]
  - Plan to develop approximate distributed sampling algorithm and code library

# Cost-Aware Retraining for ML



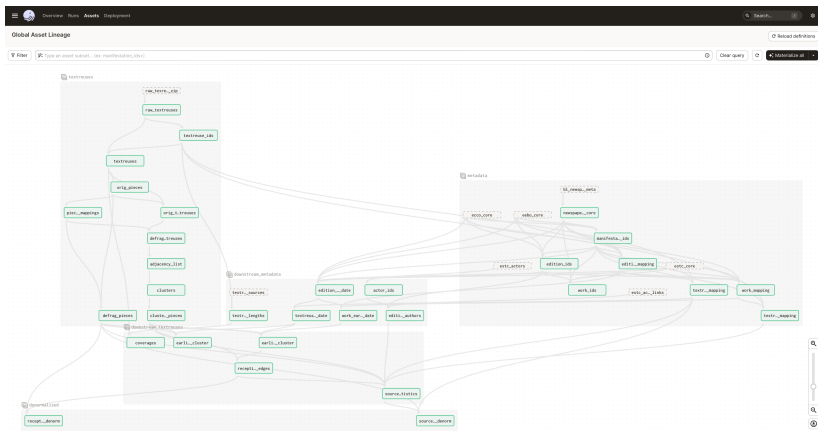
**Figure:** Example 'scenario'. (a) Initial data and model  $M_0$ . (b) Queries are far from misclassification. (d) Queries are closer to misclassifications. (c) and (e): Cost matrix and ORACLE retraining strategies with retraining cost fixed to  $\kappa = 1$  for (b) and (d) respectively.

# In Progress

- ① Cost-Aware Retraining for Machine Learning
  - Rejections from ICML 2023 and EuroSys 2023
  - Submitted pre-print to Arxiv [2]
  - Re-submitted to Knowledge-Based Systems Journal
- ② Data Science Pipeline for Historical Text Reuse
  - Manuscript under preparation
  - Planned submission to VLDB 2024 Scalable Data Science track
- ③ Scalable Constraint-Based Diversity Sampling
  - Improved running time of Fair Max-Min Diversification [8]
  - Plan to develop approximate distributed sampling algorithm and code library



## 17 / 28



# Experiments

- Benchmarking query performance on
  - ① Spark
  - ② MariaDB Aria
  - ③ MariaDB Columnstore
- Evaluating different metrics
  - ① Query Latency
  - ② Storage size
  - ③ Cost in CSC BU
- Different levels of normalisation

# In Progress

- ① Cost-Aware Retraining for Machine Learning
  - Rejections from ICML 2023 and EuroSys 2023
  - Submitted pre-print to Arxiv [2]
  - Re-submitted to Knowledge-Based Systems Journal
- ② Data Science Pipeline for Historical Text Reuse
  - Manuscript under preparation
  - Planned submission to VLDB 2024 Scalable Data Science track
- ③ Scalable Constraint-Based Diversity Sampling
  - Improved running time of Fair Max-Min Diversification [8]
  - Plan to develop approximate distributed sampling algorithm and code library

# Scaling up Fair-MMD

## Completed

- Reimplementation of FMMD-S algorithm from [8]
- Parallelizes sections of code and efficient memory management
- $100\times$  speedup on largest experiment from [8]
- Scaled up to 4M points dataset and 3000 samples with edge-case constraints

## Next Steps

- Distributed algorithm with guarantees
- Approximate distance measures
- Heuristic based ILP solver

# Discipline Specific Credits

## Completed Credits

29/30 credits completed excluding PhD Seminar (3-5 credits)

### Completed Courses:

- Network Analysis 5cr
- Security and Privacy in ML 5cr
- Advanced Latex Course 5cr
- Probabilistic Graphical Models 5cr
- Advanced Course in Machine Learning 5cr
- Transaction Management and Query Optimization 5cr
- Research Ethics 2 cr

# Transferrable Skills

## Completed Courses 3cr:

- Scientific Writing 2cr
- Principles of Peer Review 1cr

## Planned Credits 7cr:

- Conference Presentation 2cr
- Poster Presentation and Data Visualization 2cr
- International Conference participation 2cr
- Grant Writing 1cr

# Conferences

- VLDB 2022: Participation
- CIKM 2022: Presentation (virtual)
- MLSys 2023: Participation
- AI Day 2022: Poster Presentation
- Meaning of Meaning Workshop 2023: Presentation

# Teaching and Supervision

- TA: Network Analysis 2023
- Master Thesis Instruction: Anniina R Sainio (Completed)
- Master Thesis Instruction: Andreas Maniatis (Ongoing)



# Plans for 2023-2024

## ① Research

- Present - Nov 2023: Submit “Data Science Pipeline for Historical Text Reuse” to VLDB
- Jan - Mar 2024: Complete scalable sampling project

## ② Studies

- Complete remaining transferrable course credits
- Finalize all credits with Pirio

## ③ Thesis

- Mar 2024 - May 2024: Thesis Writing
- June 2024: Pre-examination

## References I

- [1] Ananth Mahadevan and Michael Mathioudakis. Certifiable unlearning pipelines for logistic regression: An experimental study. *Machine Learning and Knowledge Extraction*, 4(3):591–620, 2022.
- [2] Ananth Mahadevan and Michael Mathioudakis. Cost-effective retraining of machine learning models, 2023.
- [3] Ananth Mahadevan, Arpit Merchant, Yanhao Wang, and Michael Mathioudakis. Robustness of sketched linear classifiers to adversarial attacks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 4319–4323, New York, NY, USA, 2022. Association for Computing Machinery.

## References II

- [4] Arpit Merchant, Ananth Mahadevan, and Michael Mathioudakis. Scalably using node attributes and graph structure for node classification. *Entropy*, 24(7), 2022.
- [5] Arpit Merchant and Michael Mathioudakis. Joint use of node attributes and proximity for node classification. In Rosa Maria Benito, Chantal Cherifi, Hocine Cherifi, Esteban Moro, Luis M. Rocha, and Marta Sales-Pardo, editors, *Complex Networks & Their Applications X*, pages 511–522, Cham, 2022. Springer International Publishing.
- [6] David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. Reception reader: Exploring text reuse in early modern british publications. *Journal of Open Humanities Data*, Apr 2023.

## References III

- [7] Kai Sheng Tai, Vatsal Sharan, Peter Bailis, and Gregory Valiant. Sketching linear classifiers over data streams. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, page 757–772, New York, NY, USA, 2018. Association for Computing Machinery.
- [8] Yanhao Wang, Michael Mathioudakis, Jia Li, and Francesco Fabbri. *Max-Min Diversification with Fairness Constraints: Exact and Approximation Algorithms*, pages 91–99.