“2^{nd}” Year PhD Seminar Presentation

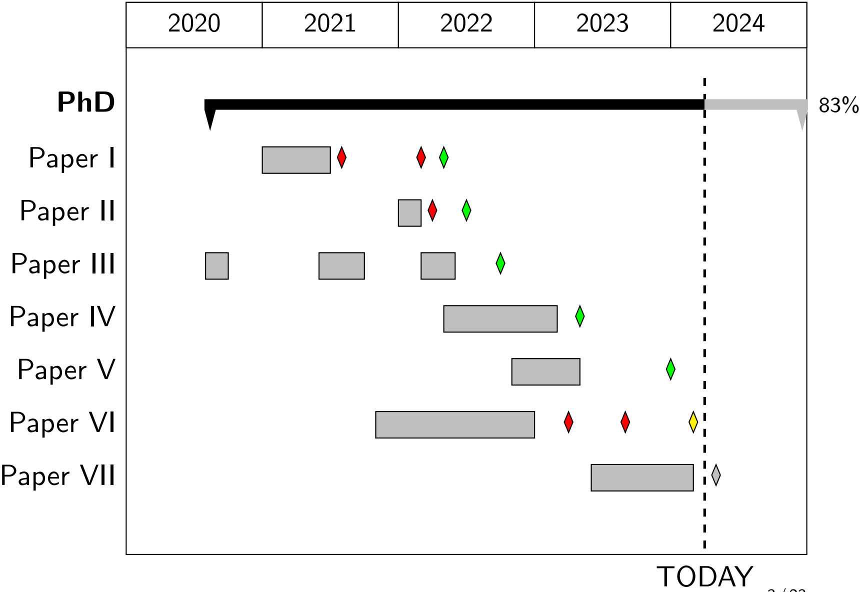
Ananth Mahadevan

March 1, 2024

Quick Recap

- **Masters:** Aalto University
- **Started:** Aug 2020 (contract) and Jan 2021 (study right)
- **Supervisor:** Michael Mathioudakis
- **Research Group:** Algorithmic Data Science (ADS)

PhD Progress



Papers

Number	One Word Title	Venue	Include in Thesis?
I	Unlearning	MAKE 2022	✅
II	JANE	Entropy 2022	✅
III	Sketching	CIKM 2022	🤷‍♂️
IV	ReceptionReader	JOHD 2023	🤷‍♂️
V	Mandeville	DES 2023	🤷‍♂️
VI	Retraining	KBS 2024	✅
VII	TextReuse	VLDB 2024	✅

Research Projects

Multiple Projects:

1. Maintaining ML models
   - Paper I: Unlearning
   - Paper VI: Retraining
2. Analyzing Historical Documents
   - Paper IV: ReceptionReader
   - Paper V: Mandeville
   - Paper VII: TextReuse
3. Scaling and Evaluating Algorithms
   - Paper II: JANE
   - Paper III: Sketching
   - Paper VIII ?: Diverse Sampling

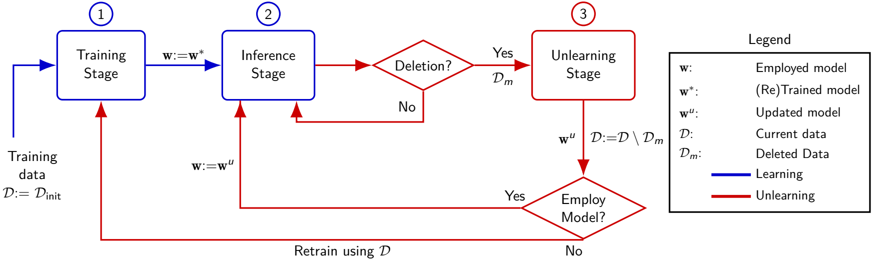
Maintaining ML models

Research Question

How to update a trained ML model when the data changes?

1. Machine Unlearning
   - Training data is deleted/removed
   - Update model parameters to forget information
   - Mahadevan and Mathioudakis [2022]
2. Cost-Aware Retraining
   - Streams drift over time
   - Data and Queries are present
   - Retraining consumes resources
   - When is it worth retraining?
   - Mahadevan and Mathioudakis [2023]

Machine Unlearning


1
Training Stage
$\mathbf{w}{:=}\mathbf{w}^*$
2
Inference Stage
Deletion?
Yes
$\mathcal{D}_m$
No
3
Unlearning Stage
$\mathbf{w}^u$
$\mathcal{D}{:=}\mathcal{D} \setminus \mathcal{D}_m$
Training data
$\mathcal{D}{:=}\ \mathcal{D}_{\text{init}}$
$\mathbf{w}{:=}\mathbf{w}^u$
Yes
Employ Model?
Retrain using $\mathcal{D}$
No
Legend
$\mathbf{w}$: Employed model
$\mathbf{w}^*$: (Re)Trained model
$\mathbf{w}^u$: Updated model
$\mathcal{D}$: Current data
$\mathcal{D}_m$: Deleted Data
Learning
Unlearning

Unlearning

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

Qualities of an approximate unlearning method:

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

Experimental Setup

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

Metric	Quality
`AccDis` ↑	Certifiability↓
`AccErr` ↑	Effectiveness↓

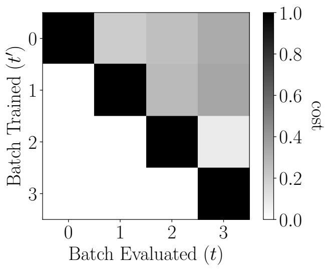
Parameter	Effect
$\tau\uparrow$	Efficiency↑ Certifiability↓ Effectiveness↓
$\sigma\uparrow$	Certifiability↑ Effectiveness↓

Cost-Aware Retraining Algorithms

Cost Matrix

$$C[t', t]$$

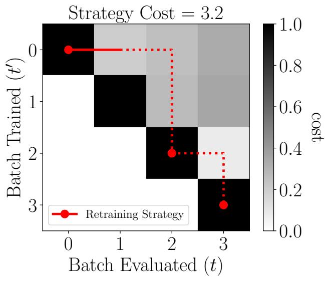
$$= \begin{cases} \text{Staleness Cost} & \text{if } t' < t \\ \text{Retraining Cost} & \text{if } t' = t \\ \infty & \text{otherwise} \end{cases}$$



Strategy

- Strategy is a set of decisions
- Retrain or Keep
- Cost of decisions is strategy cost
- Aim is to minimize strategy cost

$S = \{\texttt{Keep}, \texttt{Keep}, \texttt{Retrain}, \texttt{Retrain}\}$

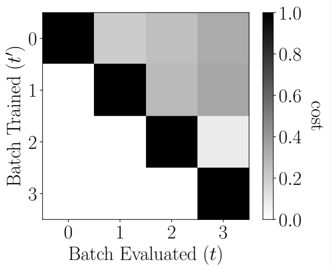


Cost-Aware Retraining Algorithms

Cost Matrix

$$C[t', t]$$

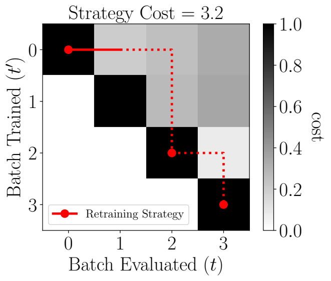
$$= \begin{cases} \text{Staleness Cost} & \text{if } t' < t \\ \text{Retraining Cost} & \text{if } t' = t \\ \infty & \text{otherwise} \end{cases}$$



Strategy

- Strategy is a set of decisions
- `Retrain` or `Keep`
- Cost of decisions is strategy cost
- Aim is to minimize strategy cost

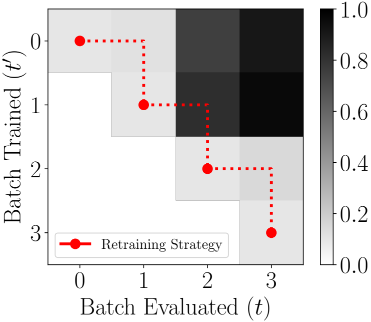
$S = \{\texttt{Keep}, \texttt{Keep}, \texttt{Retrain}, \texttt{Retrain}\}$



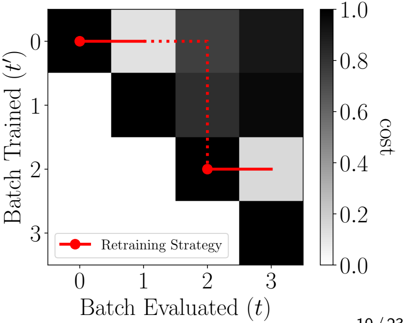
Retraining Cost κ

- Trade-off parameter between resources & performance
- Low κ
  - Performance is important
  - Frequent `Retrain` decisions to minimize staleness cost
- High κ
  - Resources are important
  - `Retrain` decisions when large enough drops in performance

Low κ

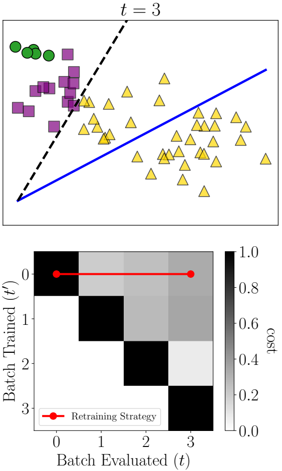


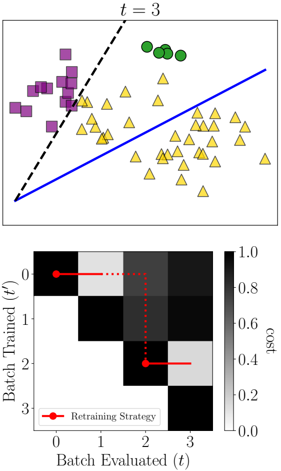
High κ



Staleness Cost

- Query-aware performance cost of old model $M_{t'}$ at batch t
- Scenario 1: Low staleness Scenario 2: High staleness





Analyzing Historical Documents

Data:

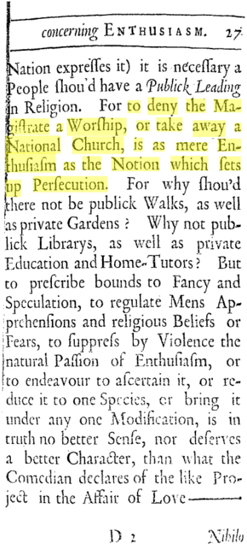
- 250K books and 1M newspapers from the 17th and 18th century
- Tons of heterogeneous metadata
  - Collections of books and articles
  - Publisher details
  - Author information
- Multi-modal data
  - Scanned page images
  - OCR text
  - XML style structured page layouts

Use cases:

- General data exploration
- Reception studies with Text Reuses
- Top quotes of authors

Reception Studies with Text Reuse

Shaftesbury, Anthony Ashley Cooper (1671-1713)
1708 - A letter concerning enthusiasm

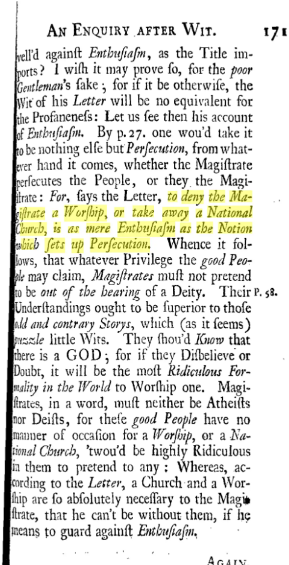


concerning ENTHUSIASM. 27

Nation expreſſes it) it is neceſſary a People ſhou'd have a *Publick Leading* in Religion. For to deny the Magiſtrate a Worſhip, or take away a National Church, is as mere Enthuſiaſm as the Notion which ſets up Perſecution. For why ſhou'd there not be publick Walks, as well as private Gardens? Why not publick Librarys, as well as private Education and Home-Tutors? But to preſcribe bounds to Fancy and Speculation, to regulate Mens Apprehenſions and religious Beliefs or Fears, to ſuppreſs by Violence the natural Paſſion of Enthuſiaſm, or to endeavour to aſcertain it, or reduce it to one Species, or bring it under any one Modification, is in truth no better Senſe, nor deſerves a better Character, than what the Comedian declares of the like Project in the Affair of Love ——

D 2 *Nihilo*

Astell, Mary (1668-1731)
1709 - Bart'lemy fair



AN ENQUIRY AFTER WIT. 171

vell'd againſt *Enthuſiaſm*, as the Title imports? I wiſh it may prove ſo, for the *poor Gentleman*'s ſake; for if it be otherwiſe, the Wit of his *Letter* will be no equivalent for the Profaneneſs: Let us ſee then his account of *Enthuſiaſm*. By p. 27. one wou'd take it to be nothing elſe but *Perſecution*, from whatever hand it comes, whether the Magiſtrate perſecutes the People, or they the Magiſtrate: *For*, ſays the Letter, *to deny the Magiſtrate a Worſhip, or take away a National Church, is as mere Enthuſiaſm as the Notion which ſets up Perſecution.* Whence it follows, that whatever Privilege the *good People* may claim, *Magiſtrates* muſt not pretend to be *out of the bearing* of a Deity. Their P. 58. Underſtandings ought to be ſuperior to thoſe *old and contrary Storys*, which (as it ſeems) *puzzle* little Wits. They ſhou'd *Know* that there is a GOD; for if they Diſbelieve or Doubt, it will be the moſt *Ridiculous Formality in the World* to Worſhip one. Magiſtrates, in a word, muſt neither be Atheiſts nor Deiſts, for theſe *good People* have no manner of occaſion for a *Worſhip*, or a *National Church*, 'twou'd be highly Ridiculous in them to pretend to any: Whereas, according to the *Letter*, a Church and a Worſhip are ſo abſolutely neceſſary to the Magiſtrate, that he can't be without them, if he means to guard againſt *Enthuſiaſm*.

AGAIN

Identifying Text Reuse

However, OCR texts are very noisy

Document 1 string

```
to deny the Ma-
tifl:ate a Worflip, or take away a
hational Church, is as mere En-
Ihufiafin as the Notion which sets
uip Persecution
```

Document 2 string

```
, to deny the Ala-
inrate a ITorftip, or take awvay a National
'uircb, is as mere Entnztfiafm as the Notion
.bic fJets tup Persecution. W
```

How to identify text reuses?

- Use BLAST to do fuzzy alignment

Pre-Processing Pipeline

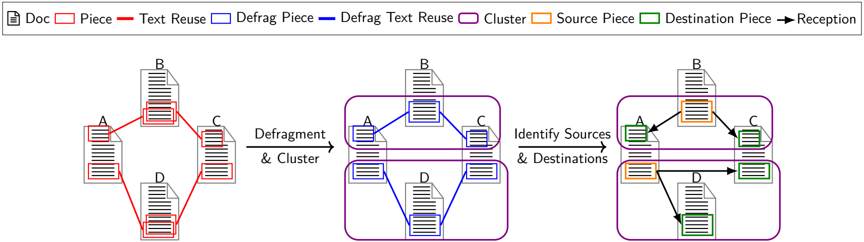
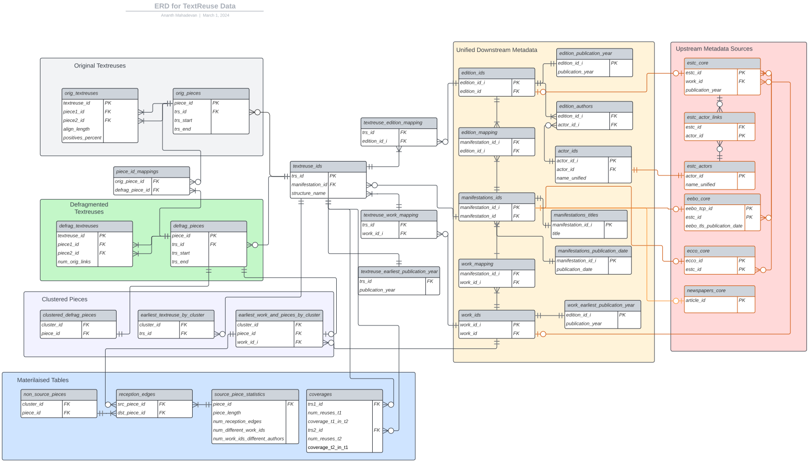


Figure: Illustration of the pre-processing pipeline for text reuses. The hits from BLAST (red) are first defragmented (blue) and then clustered (purple). Then in each cluster, the piece from the earliest document is identified as the source (yellow) and the rest are destinations (green). The directed edge between a source and destination piece is a reception edge.

ETL Pipeline Implementation

- Starts with **6.31 billion** pairs of reuses from BLAST
- ETL implemented in Apache Spark
- Uses Dagster for asset management
- Loads processed data into MariaDB for downstream queries
- Tables used for front-end user queries

ETL Pipeline Implementation



ERD for TextReuse Data

Ananth Mahadevan | March 1, 2024

Original Textreuses

orig_textreuses	
textreuse_id	PK
piece1_id	FK
piece2_id	FK
align_length	
positives_percent	

orig_pieces	
piece_id	PK
trs_id	FK
trs_start	
trs_end	

piece_id_mappings	
orig_piece_id	FK
defrag_piece_id	FK

Defragmented Textreuses

defrag_textreuses	
textreuse_id	PK
piece1_id	FK
piece2_id	FK
num_orig_links	

defrag_pieces	
piece_id	PK
trs_id	FK
trs_start	
trs_end	

textreuse_edition_mapping	
trs_id	FK
edition_id_i	FK

textreuse_ids	
trs_id	PK
manifestation_id	FK
structure_name	

textreuse_work_mapping	
trs_id	FK
work_id_i	FK

textreuse_earliest_publication_year	
trs_id	FK
publication_year	

Unified Downstream Metadata

edition_publication_year	
edition_id_i	PK
publication_year	

edition_ids	
edition_id_i	PK
edition_id	FK

edition_authors	
edition_id_i	FK
actor_id_i	FK

edition_mapping	
manifestation_id_i	FK
edition_id_i	FK

actor_ids	
actor_id_i	PK
actor_id	FK
name_unified	

manifestations_ids	
manifestation_id_i	PK
manifestation_id	FK

manifestations_titles	
manifestation_id_i	PK
title	

manifestations_publication_date	
manifestation_id_i	PK
publication_date	

work_mapping	
manifestation_id_i	FK
work_id_i	FK

work_earliest_publication_year	
edition_id_i	PK
publication_year	

work_ids	
work_id_i	PK
work_id	FK

Upstream Metadata Sources

estc_core	
estc_id	PK
work_id	FK
publication_year	

estc_actor_links	
estc_id	FK
actor_id	PK

estc_actors	
actor_id	PK
name_unified	

eebo_core	
eebo_tcp_id	PK
estc_id	PK
eebo_tls_publication_date	

ecco_core	
ecco_id	PK
estc_id	PK

newspapers_core	
article_id	PK

Clustered Pieces

clustered_defrag_pieces	
cluster_id	FK
piece_id	FK

earliest_textreuse_by_cluster	
cluster_id	FK
trs_id	FK

earliest_work_and_pieces_by_cluster	
cluster_id	FK
piece_id	FK
work_id_i	FK

Materilaised Tables

non_source_pieces	
cluster_id	FK
piece_id	FK

reception_edges	
src_piece_id	FK
dst_piece_id	FK

source_piece_statistics	
piece_id	FK
piece_length	
num_reception_edges	
num_different_work_ids	
num_work_ids_different_authors	

coverages	
trs1_id	FK
num_reuses_t1	
coverage_t1_in_t2	
trs2_id	FK
num_reuses_t2	
coverage_t2_in_t1	

Related Publications

1. Paper IV: Rosson et al. [2023]
   - *Reception reader: Exploring text reuse in early modern british publications*
   - Front-end user interface for browsing reuses
2. Paper V: Ryan et al. [2023]
   - *A Comparative text similarity analysis of the works of Bernard Mandeville*
   - Study using the data and interfaces from Paper IV
3. Paper VI: Mahadevan et al. [2024]
   - *Optimizing a Data Science System for Text Reuse Analysis*
   - Studies design choices to optimize performance of the system
   - Plans to scale interface from Paper IV based on insights

Scalability and Robustness

- Paper II: Merchant et al. [2022]
  - Scaled up GNN alternative from Merchant and Mathioudakis [2022] to run effectively on graphs with millions of nodes
- Paper III: Mahadevan et al. [2022]
  - Explored the robustness of Sketched Linear Networks from Tai et al. [2018] to adversarial attacks
- Paper VIII?:
  - Original algorithm from Wang et al. [2023]
  - Re-implemented for edge-case requirement in Historical Project
  - Nearly $200\times$ speed-up compared to original code
  - Plans to develop theory and code for distributed algorithm

Next Steps

1. Complete a few more transferrable skill credits
2. Start Writing Thesis
3. Work on Paper VIII in tandem

References I

Ananth Mahadevan and Michael Mathioudakis. Certifiable unlearning pipelines for logistic regression: An experimental study. *Machine Learning and Knowledge Extraction*, 4(3): 591–620, 2022. ISSN 2504-4990. doi: 10.3390/make4030028. URL https://www.mdpi.com/2504-4990/4/3/28.

Ananth Mahadevan and Michael Mathioudakis. Cost-effective retraining of machine learning models, 2023.

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. ISBN 9781450392365. doi: 10.1145/3511808.3557687. URL https://doi.org/10.1145/3511808.3557687.

References II

Ananth Mahadevan, Michael Mathioudakis, Eetu Mäkelä, and Mikko Tolonen. Optimizing a data science system for text reuse analysis, 2024.

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.

Arpit Merchant, Ananth Mahadevan, and Michael Mathioudakis. Scalably using node attributes and graph structure for node classification. *Entropy*, 24(7), 2022. ISSN 1099-4300. doi: 10.3390/e24070906. URL https://www.mdpi.com/1099-4300/24/7/906.

References III

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. doi: 10.5334/johd.101.

Yann Ryan, Ananth Mahadevan, and Mikko Tolonen. A comparative text similarity analysis of the works of bernard mandeville. *Digital Enlightenment Studies*, 1:28–58, 12 2023. ISSN 3029-0953. doi: 10.61147/des.6. URL https://digitalenlightenmentstudies.org/article/id/6/.

References IV

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. ISBN 9781450347037. doi: 10.1145/3183713.3196930. URL https://doi.org/10.1145/3183713.3196930.

Yanhao Wang, Michael Mathioudakis, Jia Li, and Francesco Fabbri. *Max-Min Diversification with Fairness Constraints: Exact and Approximation Algorithms*, pages 91–99. 2023. doi: 10.1137/1.9781611977653.ch11. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch11.