"2nd" Year PhD Seminar Presentation

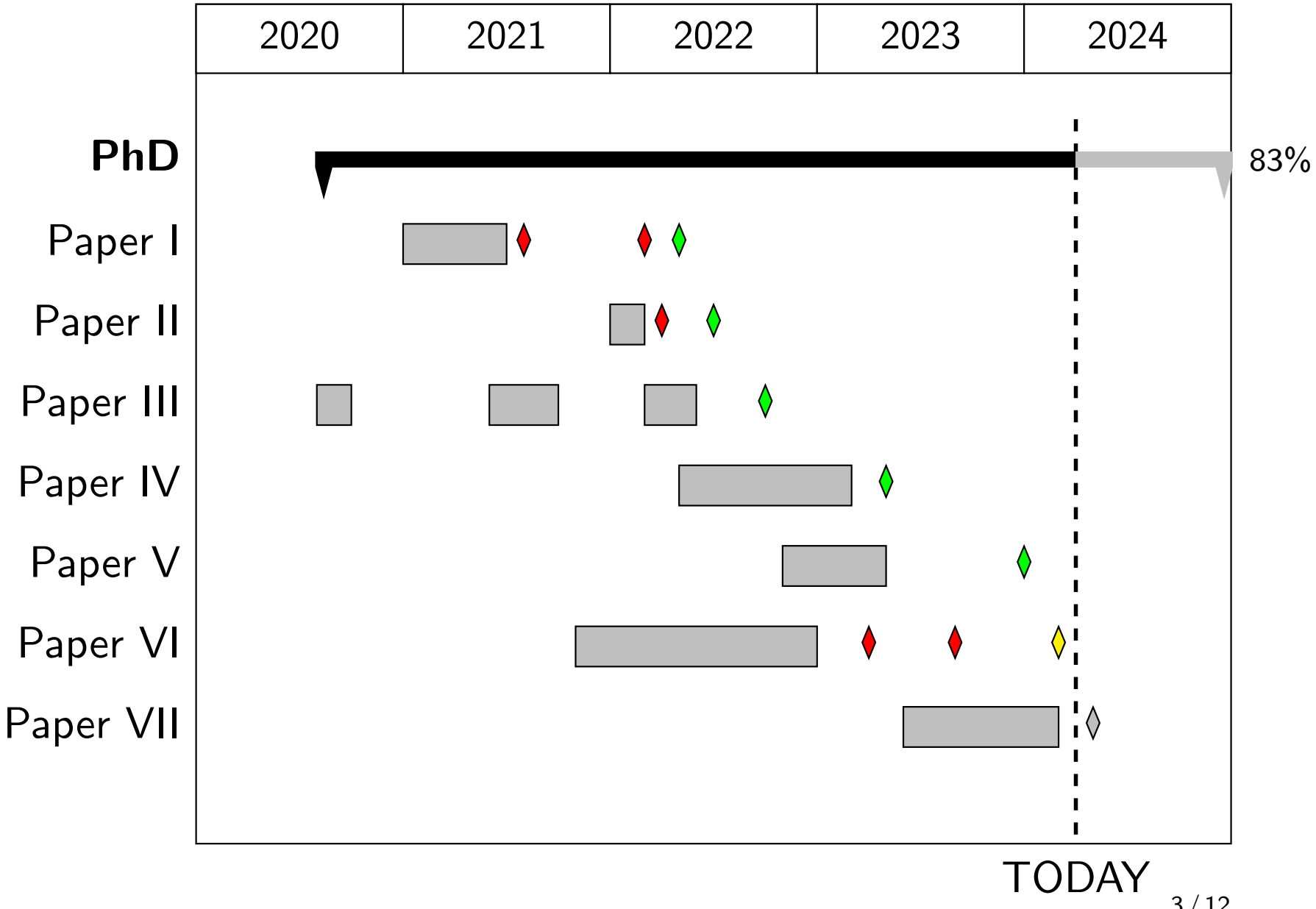
Ananth Mahadevan

February 23, 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 Analyzing Historical Documents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

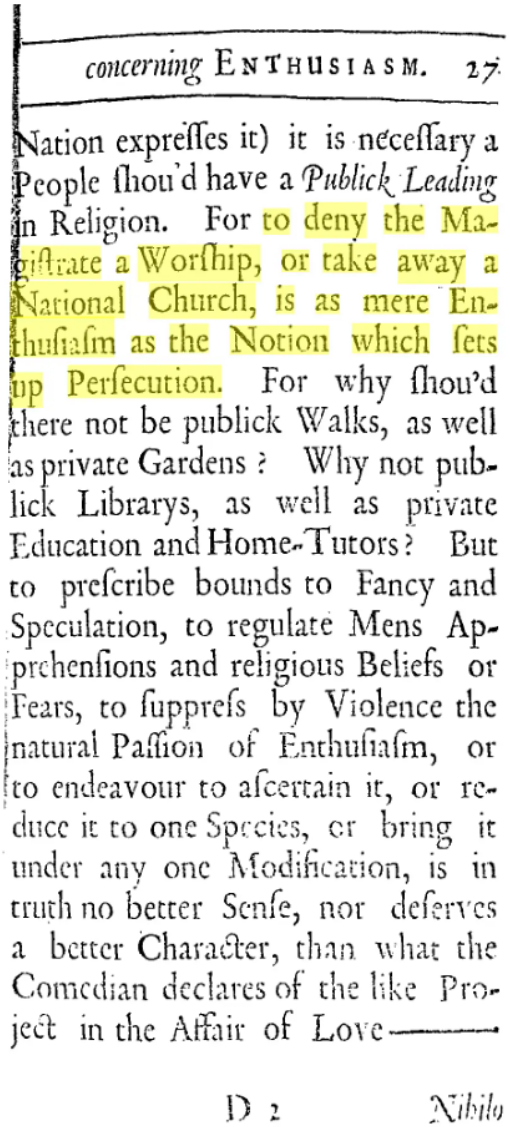
Research Question

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
2. Cost-Aware Retraining
   - Streams drift over time
   - Data and Queries are present
   - Retraining consumes resources
   - When is it worth retraining?

Analyzing Historical Documents

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

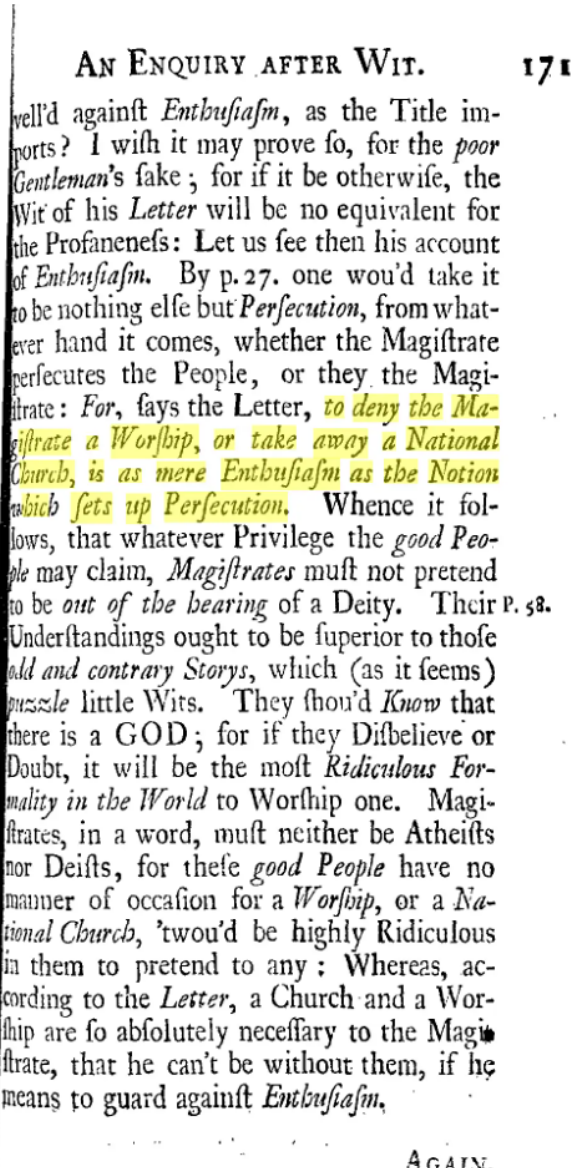


concerning ENTHUSIASM. 27

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 hearing* of a Deity. Their P. 58.
Underſtandings ought to be ſuperior to thoſe *odd 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

Analyzing Historical Documents

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
```

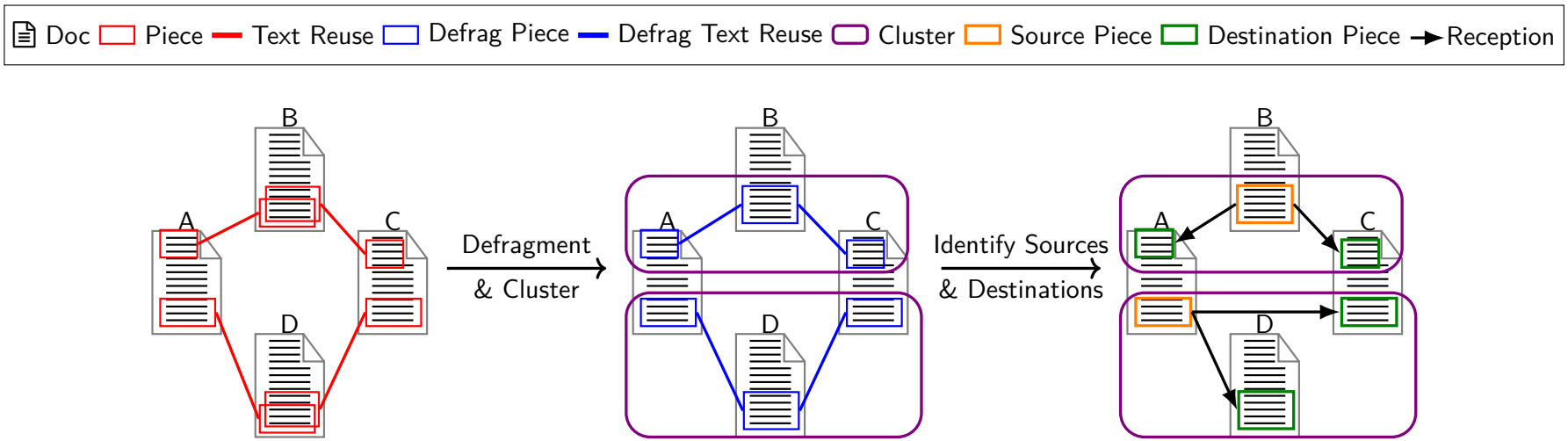
Document 2 string

```
, 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?

- Use BLAST to do fuzzy alignment

Pre-Processing Pipeline



- Clean up BLAST hits for downstream tasks
- Implemented in Apache Spark
- Scales up to **6.31 billion** pairs of reuses

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 Rosson et al. [2023]
3. Paper VI: Mahadevan et al. [2024]
   - *Optimizing a Data Science System for Text Reuse Analysis*
   - Studies design choices for the optimal performance of the system

Scalability and Robustness

References I

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

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/`.