Model Evaluation in Medical Datasets Over Time

Helen Zhou, Yuwen Chen*, Zachary C. Lipton*
Carnegie Mellon University

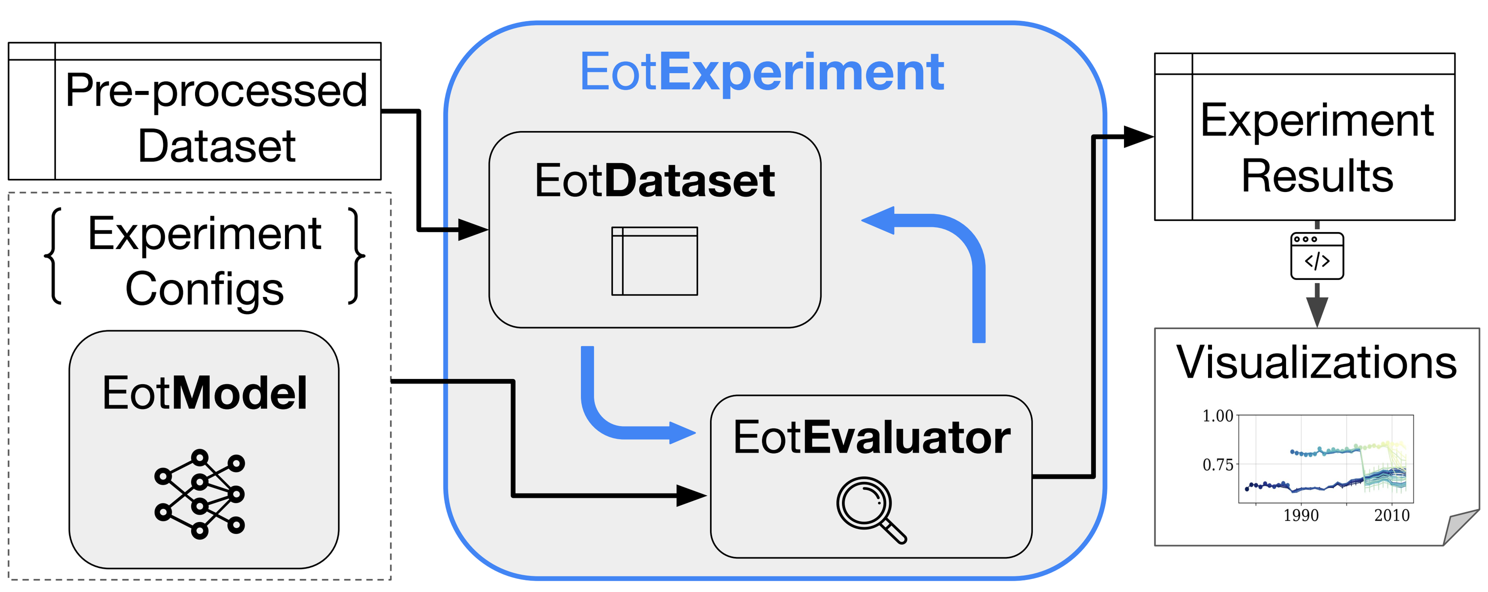
Carnegie Mellon University School of Computer Science
Machine Learning Department
Electrical & Computer Engineering

The Problem

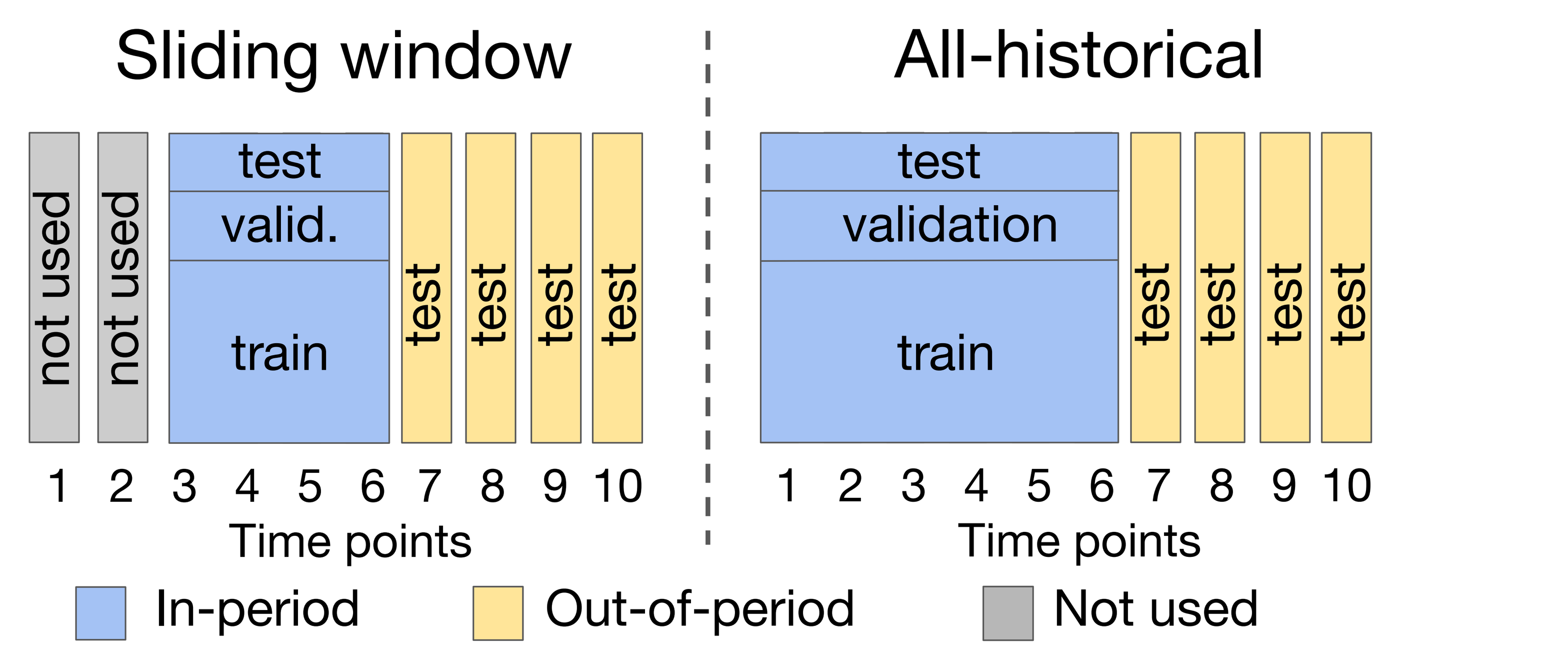
ML models deployed in health systems face data drawn from **continually evolving environments.**

But, researchers proposing such models typically evaluate them in a **time-agnostic** manner.

EMDOT Package



Methods



- Train up to each "simulated deployment date" & evaluate on all future timepoints
- Training regimes: sliding window, all-historical
- Model classes: LR, GBDT, MLP
- Datasets: SEER, CDC, SWPA, OPTN, MIMIC-IV

Dataset	Outcome	Time Range (unit)	N	# pos.
SEER (Breast)	5Y Surv.	1975 – 2013 (year)	462,023	378,758
SEER (Colon)	5Y Surv.	1975 – 2013 (year)	254,112	135,065
SEER (Lung)	5Y Surv.	1975 – 2013 (year)	457,695	49,997
CDC COVID	Mort.	Mar '20 – May '22 (month)	941,140	190,786
SWPA COVID	90D Mort.	Mar '20 – Feb '22 (month)	35,293	1,516
MIMIC-IV	In-ICU Mort.	2009 – 2020 (year)	53,050	3,334
OPTN (Liver)	180D Mort.	2005 – 2017 (year)	143,709	4,635

Results

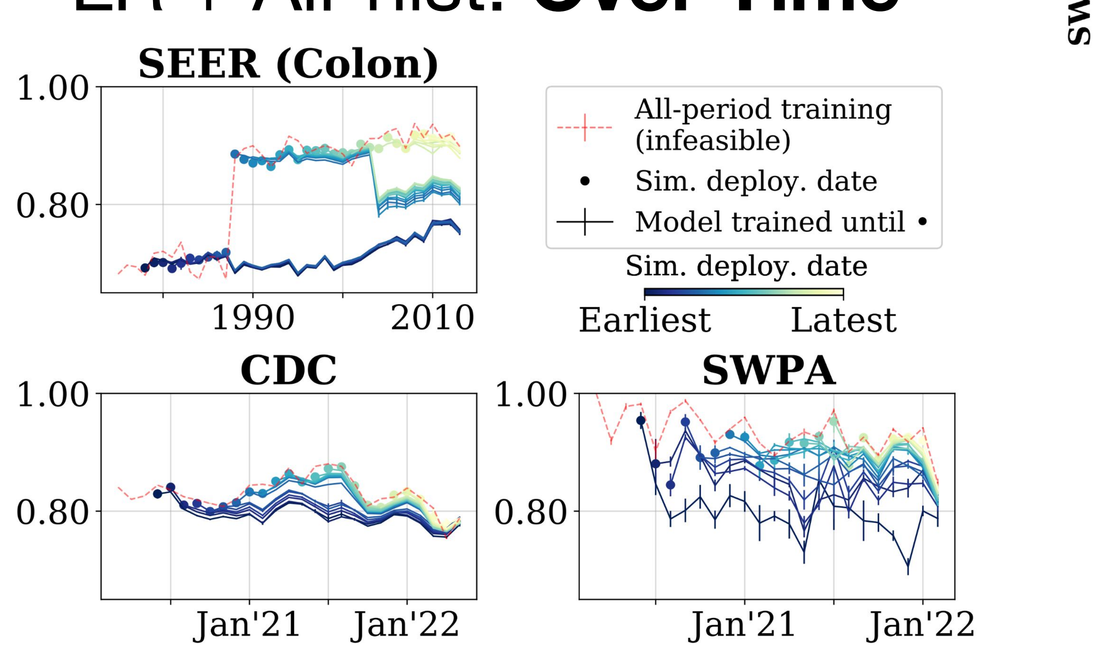
Time-agnostic Evaluation

(Standard) Reported AUROC

Model	SEER (Colon)	CDC	SWPA
LR	0.867	0.837	0.914
GBDT	0.871	**0.850**	**0.926**
MLP	**0.873**	0.844	0.918

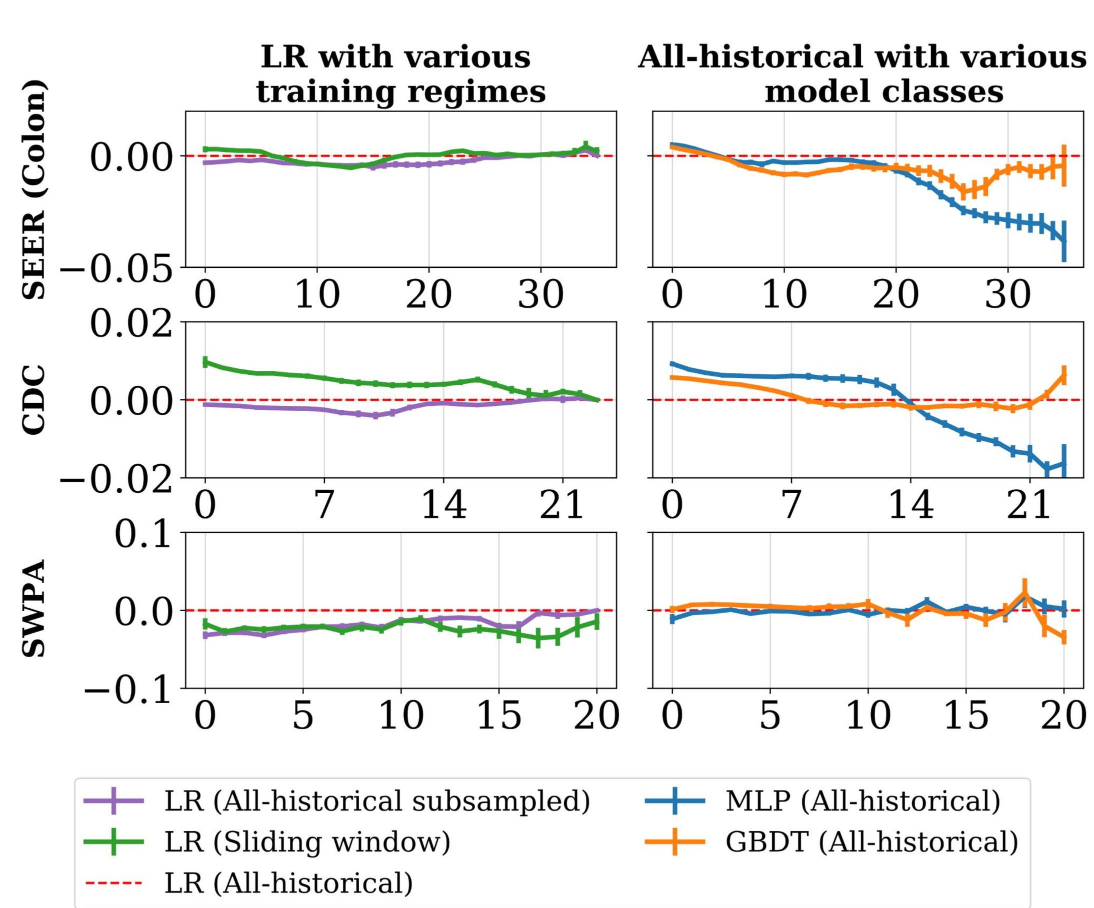
> In standard eval., GBDT & MLP do best.

LR + All-hist. **Over Time**



> SEER (Colon): jumps in 1983 and 2003.
> CDC: relatively smooth, boost in Dec 2021.
> SWPA: more variation & uncertainty at first.
> Red line shows over-optimistic standard eval.

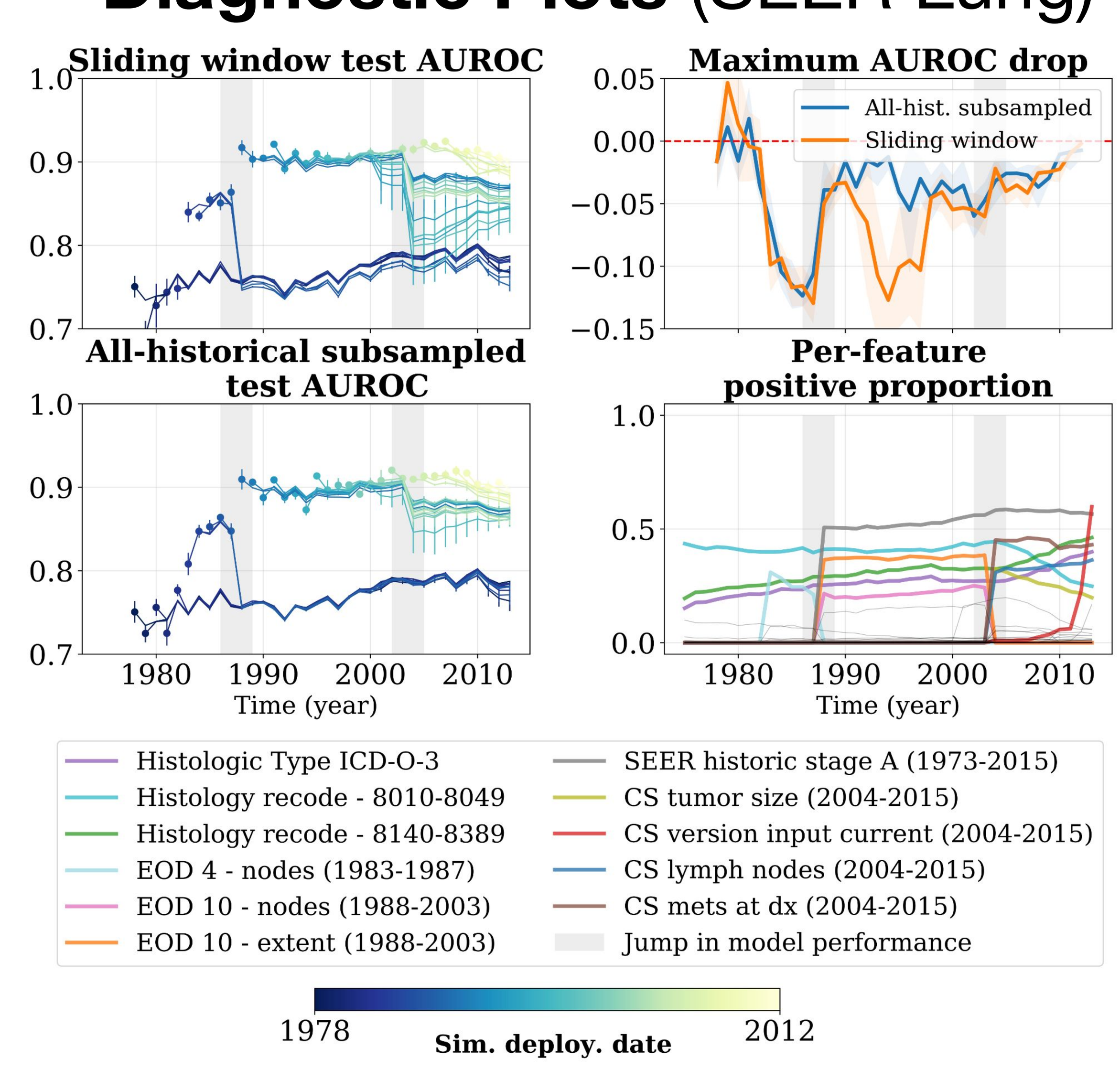
AUROC Diff. vs. Staleness



Training regimes:
> SEER (Colon): all comparable
> CDC: sliding window best
> SWPA: all-historical best

Model classes:
> SEER (Colon): LR better at large staleness
> CDC: LR better at large staleness
> SWPA: all comparable

Diagnostic Plots (SEER Lung)



> 1983: EOD4 introduced, jump in performance
> 1988: EOD4 removed, EOD10 introduced
> 2003: EOD10 removed
> All-historical avoids the large maximum AUROC drop that sliding window experiences

Discussion & Future Work

- Larger datasets in rapidly evolving environments may benefit from sliding window training
- Smaller datasets may benefit from all-historical training
- Introduction and removal of features can result in dramatic changes in performance over time
- If a model were trained on a mixture of distributions that occurred throughout the past, it may be better equipped to handle shifts to related settings in the future

Up next:

- Parallelization of EMDOT
- Extensions to other modalities
- Benchmarking domain adaptation techniques

**equal contribution*