BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA
Electrical Engineering and Computer Sciences

Computational
PRECISION HEALTH

The Data Addition Dilemma

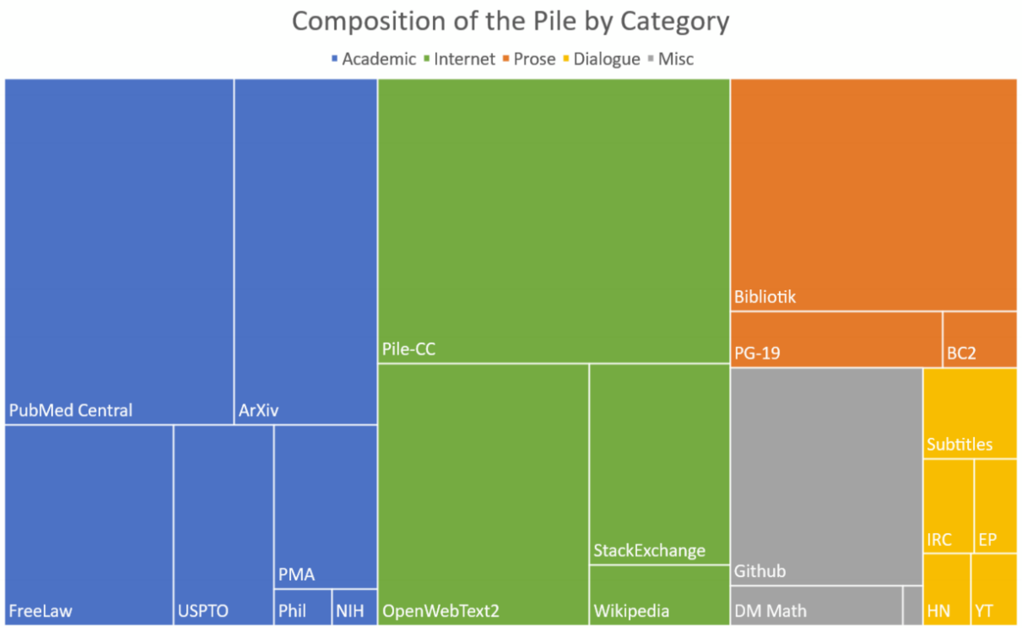
Irene Y. Chen

Assistant Professor, Computational Precision Health
Electronical Engineering and Computer Science
Berkeley AI Research
UC Berkeley and UCSF

irenechen.net

@irenetrampoline

Age of bigger and bigger datasets



Pile dataset
(825 GB, 22 sources)

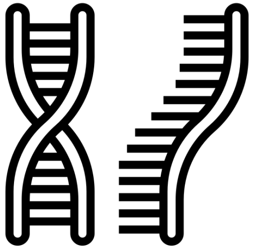
text string	timestamp string	url string
Beginners BBQ Class Taking Place in Missoula! Do you want to get better at making delicious BBQ? Yo…	2019-04-25T12:57:54Z	https://klyq.com/beginners-bbq-class-taking-p in-missoula/
Discussion in 'Mac OS X Lion (10.7)' started by axboi87, Jan 20, 2012. I've got a 500gb internal…	2019-04-21T10:07:13Z	https://forums.macrumors.com/threads/restore- larger-disk-to-smaller-disk.1311329/
Foil plaid lycra and spandex shortall with metallic slinky insets. Attached metallic elastic belt with…	2019-04-25T10:40:23Z	https://awishcometrue.com/Catalogs/Clearance/ /V1960-Find-A-Way
How many backlinks per day for new site? Discussion in 'Black Hat SEO' started by Omoplata, Dec 3,…	2019-04-21T12:46:19Z	https://www.blackhatworld.com/seo/how-many-backlinks-per-day-for-new-site.258615/
The Denver Board of Education opened the 2017-18 school year with an update on projects that includ…	2019-04-20T14:33:21Z	http://bond.dpsk12.org/category/news/

C4 (750 GB)

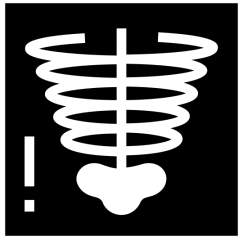
StarCoder (783 GB,
86 programming languages)

The pursuit of more health data

Electronic Medical Records

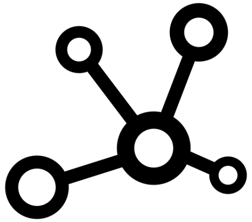


Genomics

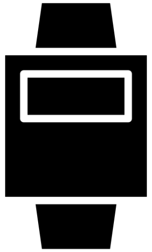


Medical Imaging

Signals

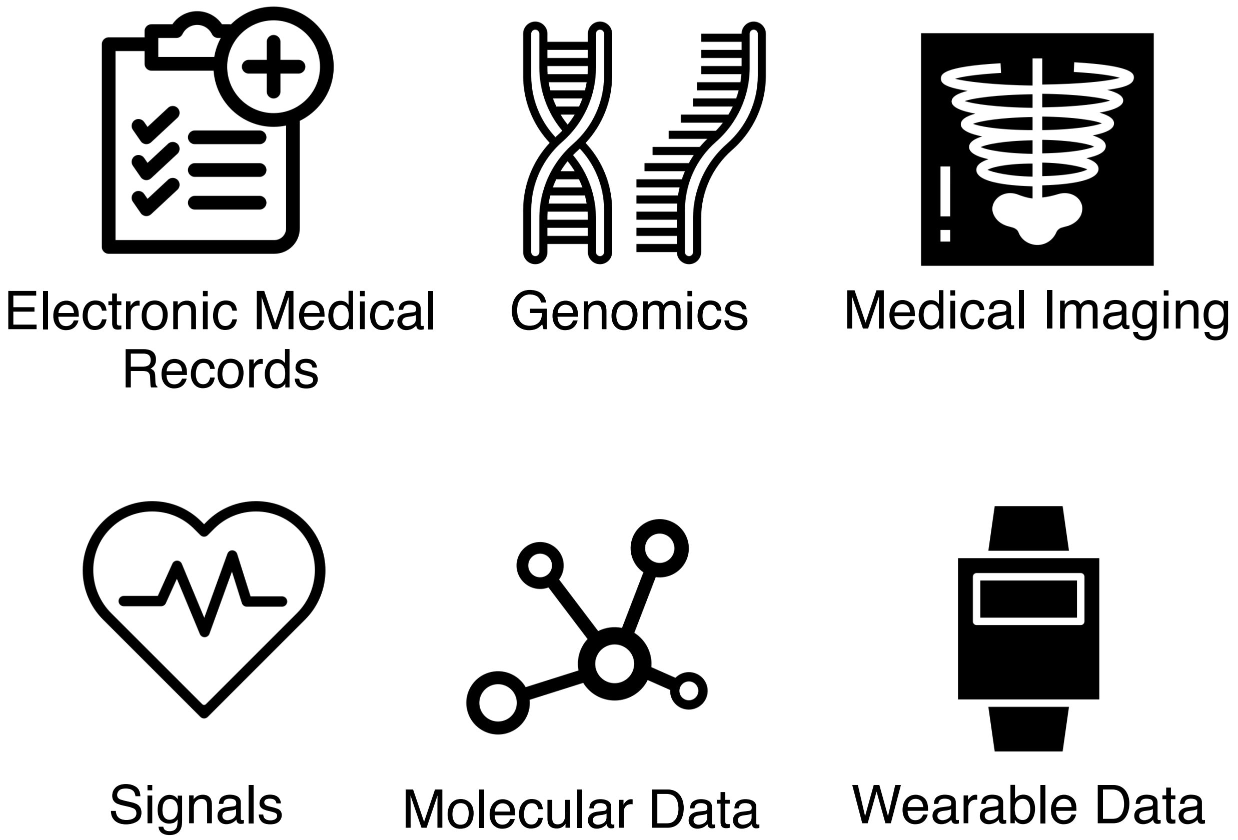


Molecular Data



Wearable Data

The pursuit of more health data



All of Us RESEARCH PROGRAM

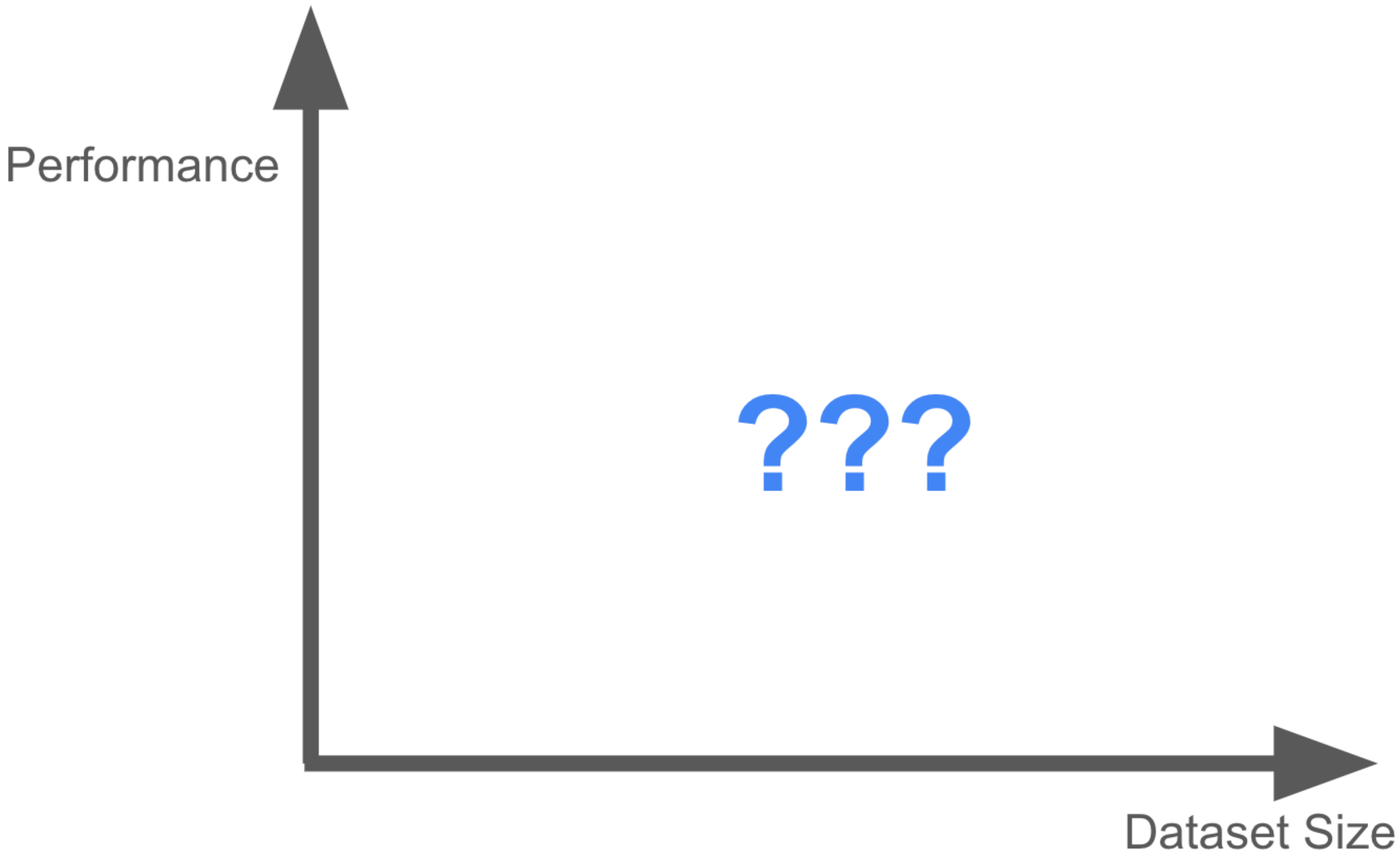
800k patients
50 clinical sites

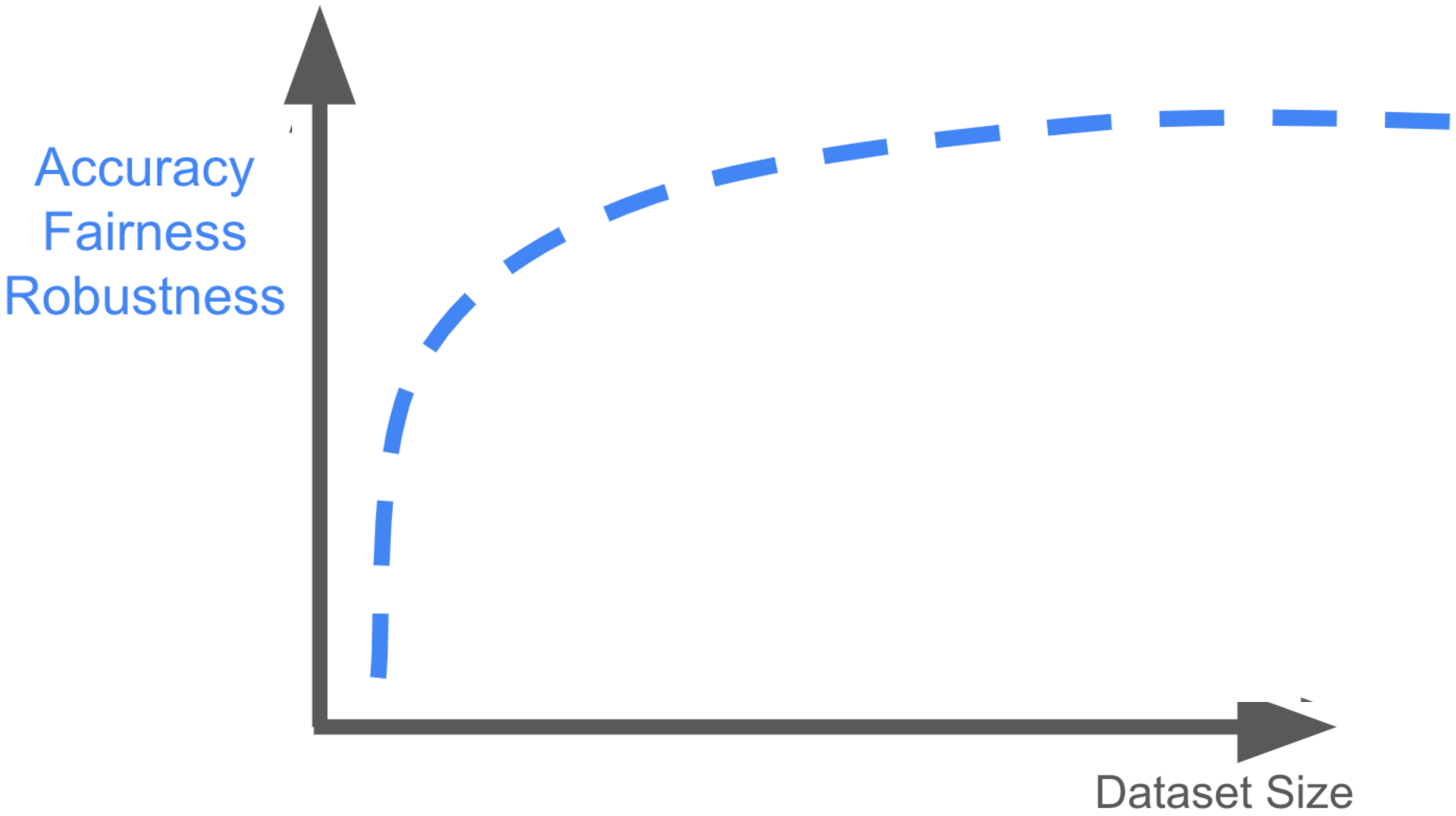
uk biobank

500k individuals
22 recruitment centers

eICU Collaborative Research Database

200k patient stays
208 hospitals


Performance
???
Dataset Size



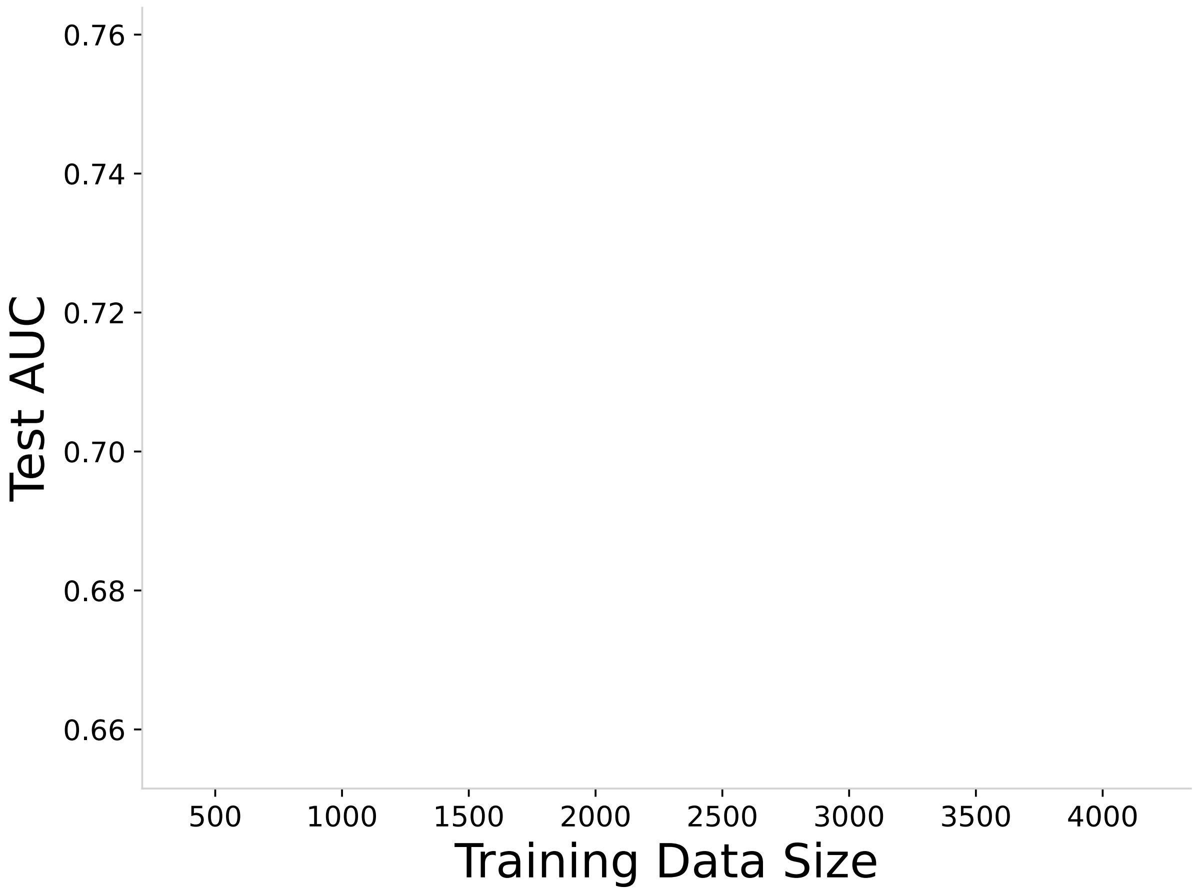
Chen et al, NeurIPS 2018; Rolf et al, ICML 2021; Carmon et al, NeurIPS 2019

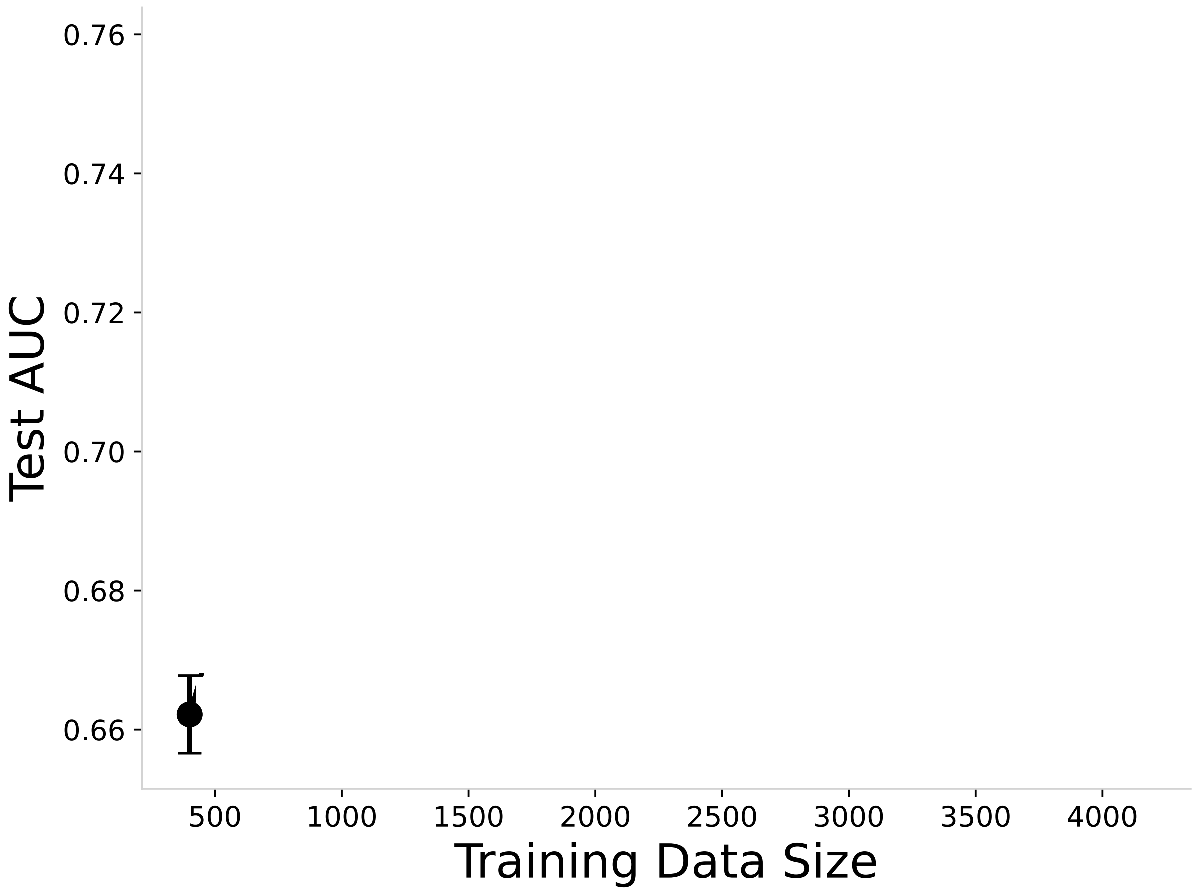
Judy Hanwen Shen
(Stanford)

Inioluwa Deborah Raji
(UC Berkeley)

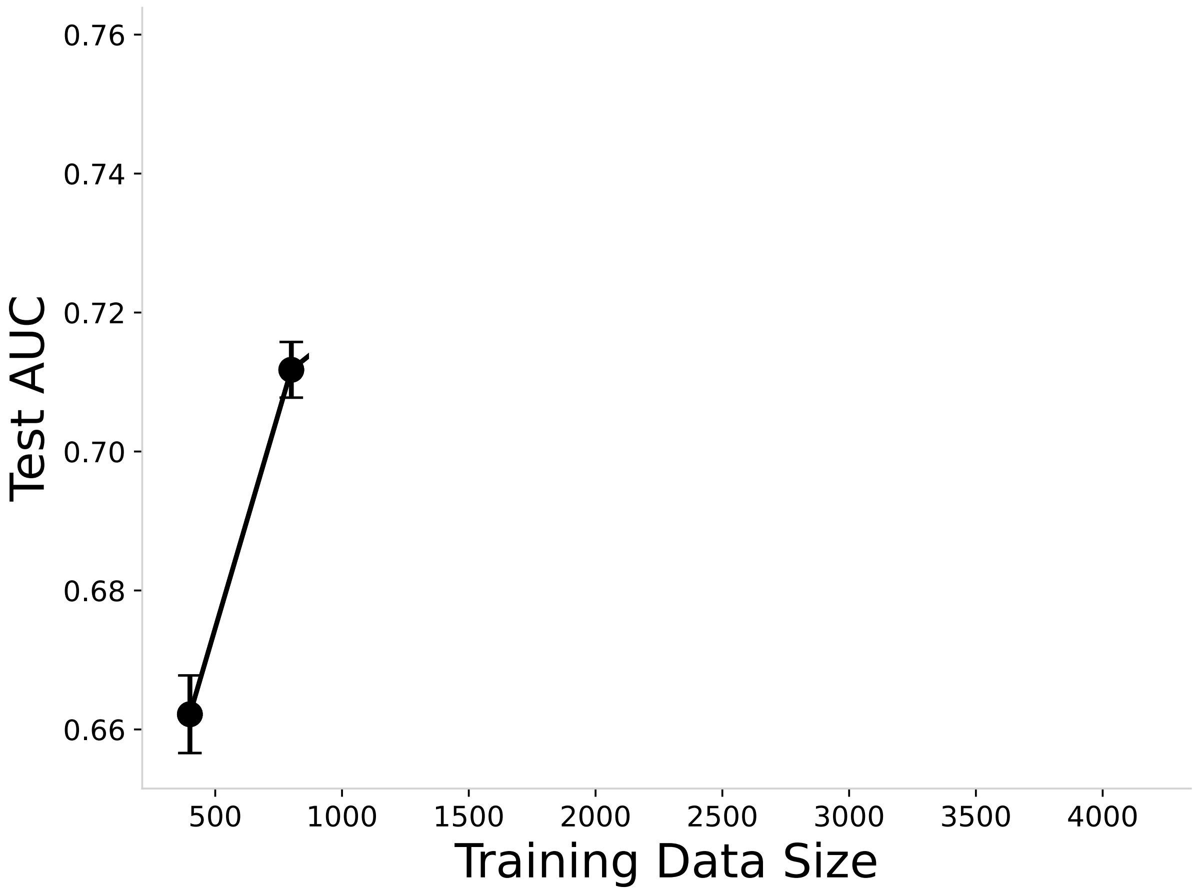
How do we balance data accumulation and distribution shift?

Shen et al, “The Data Addition Dilemma.” MLHC 2024

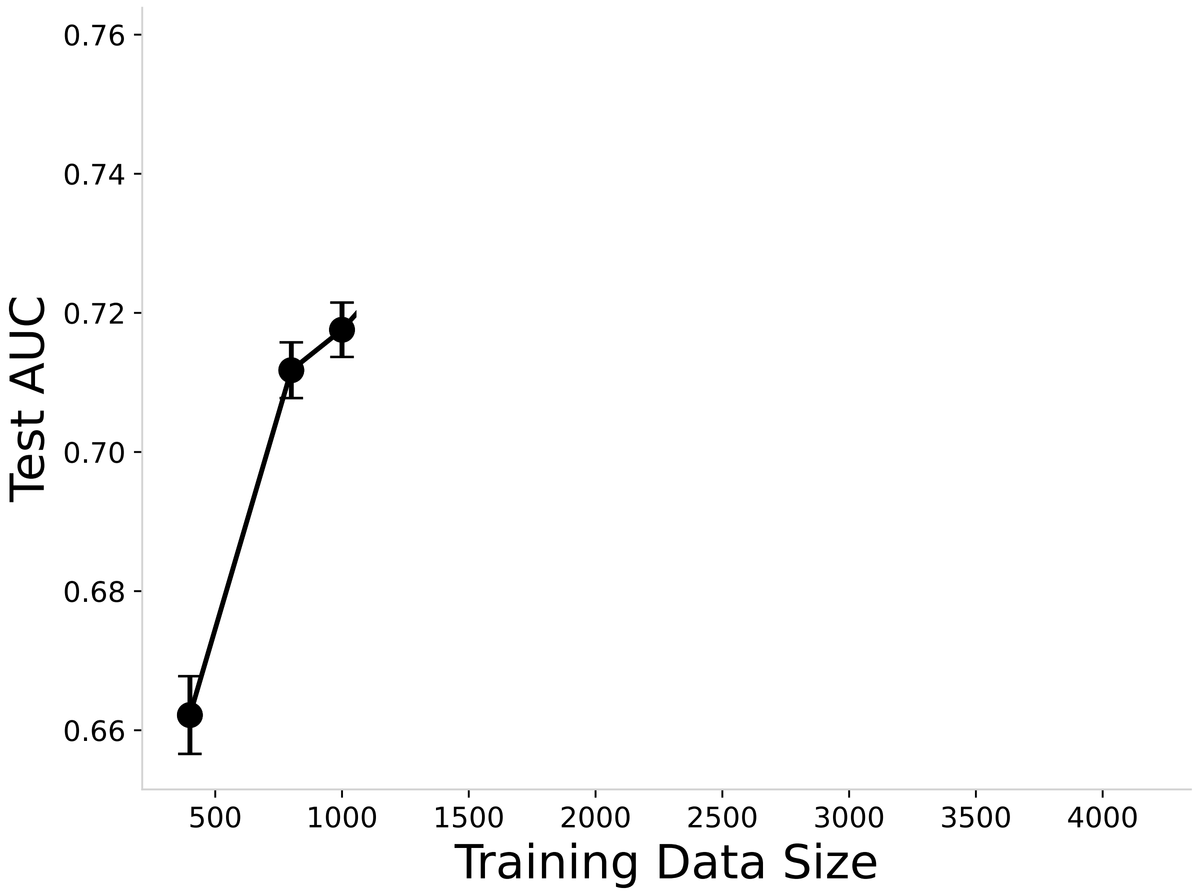

0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size

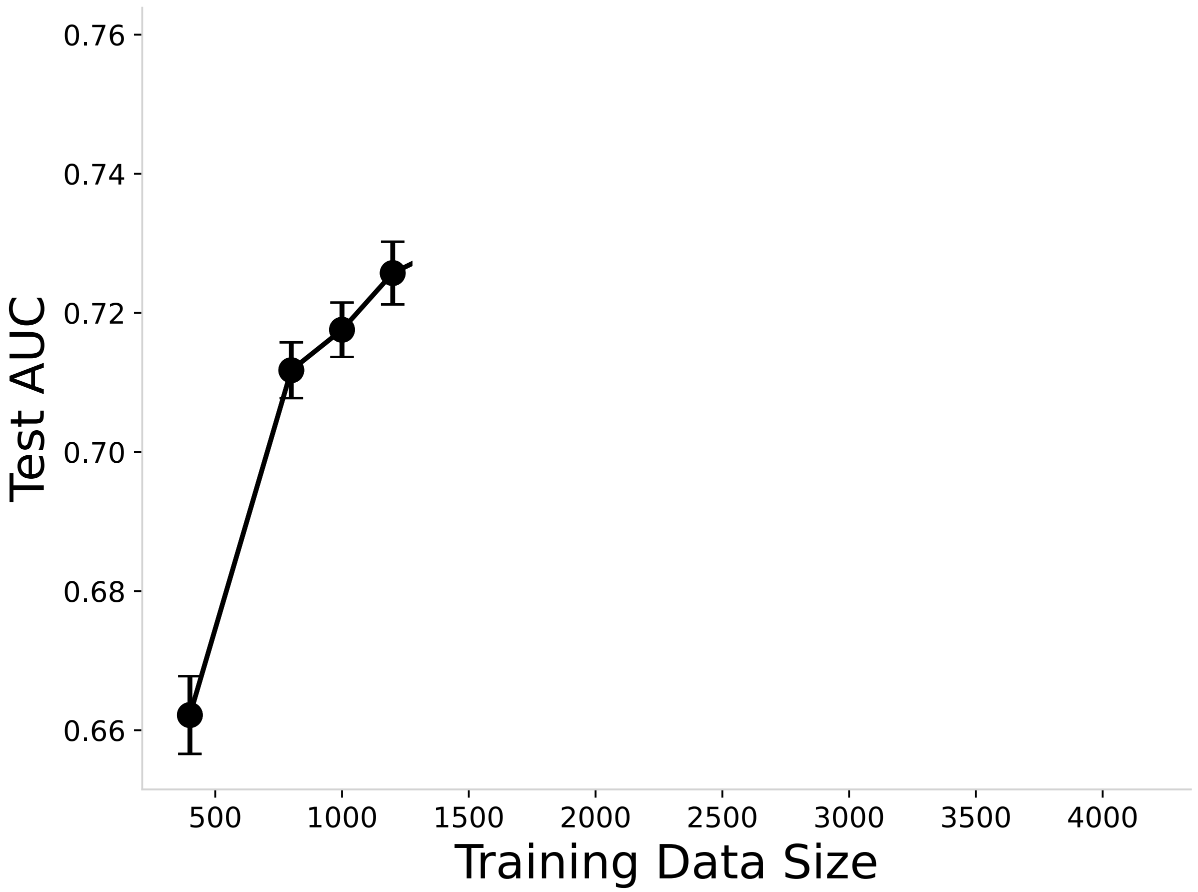


0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size

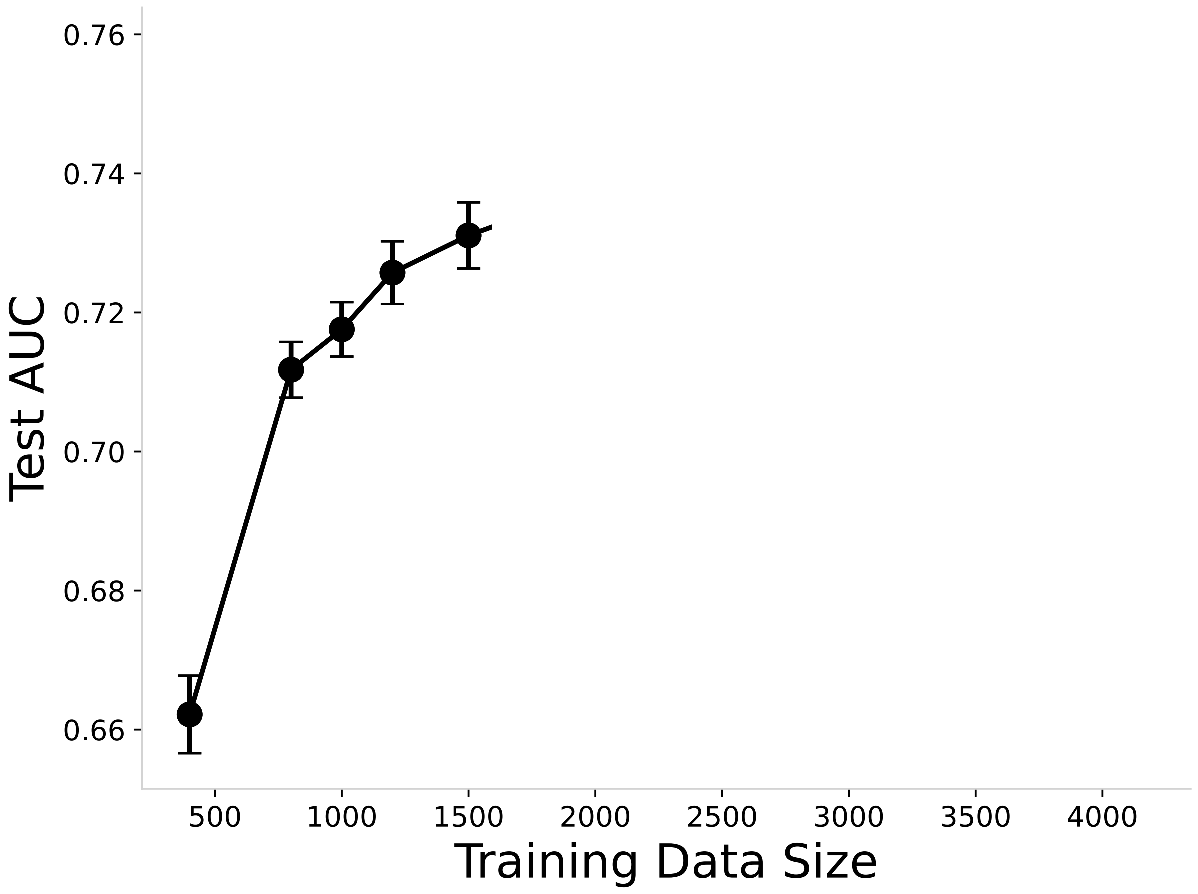


0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size

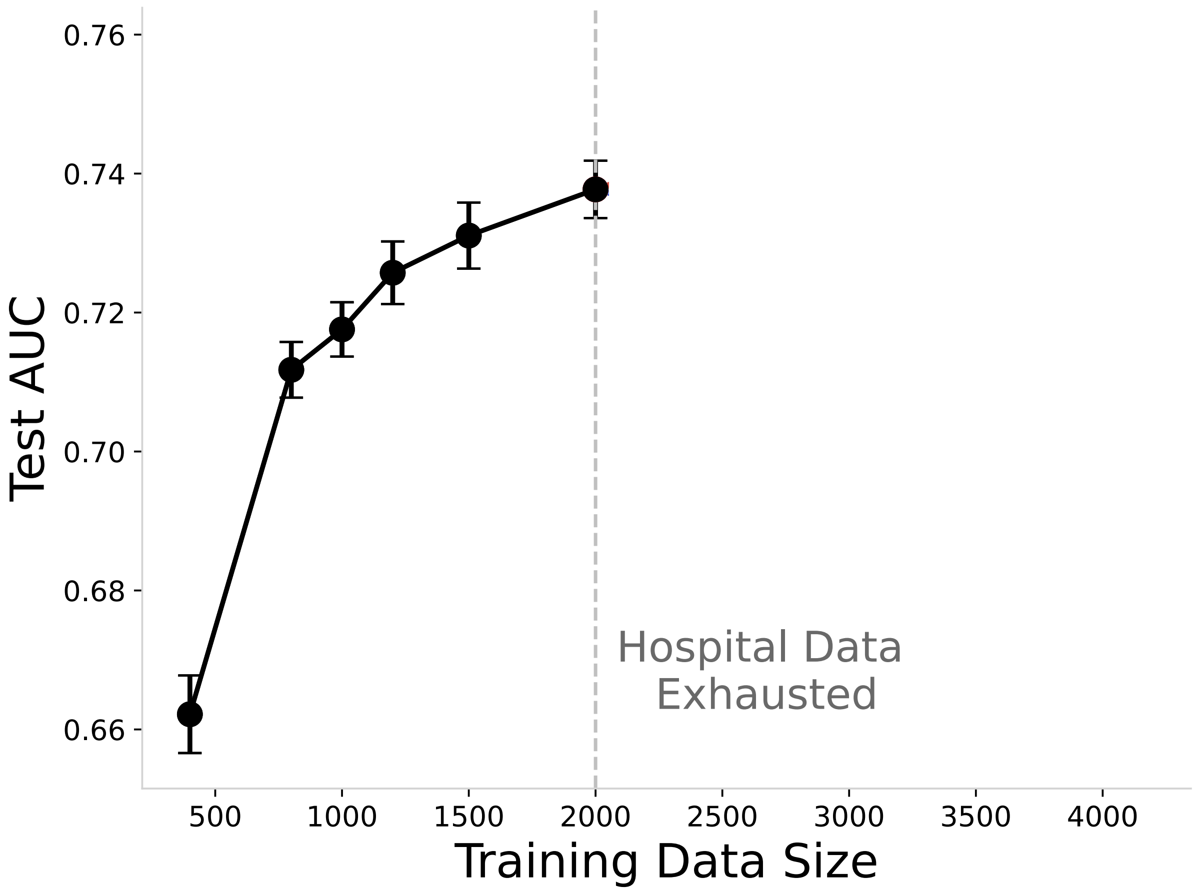

0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size

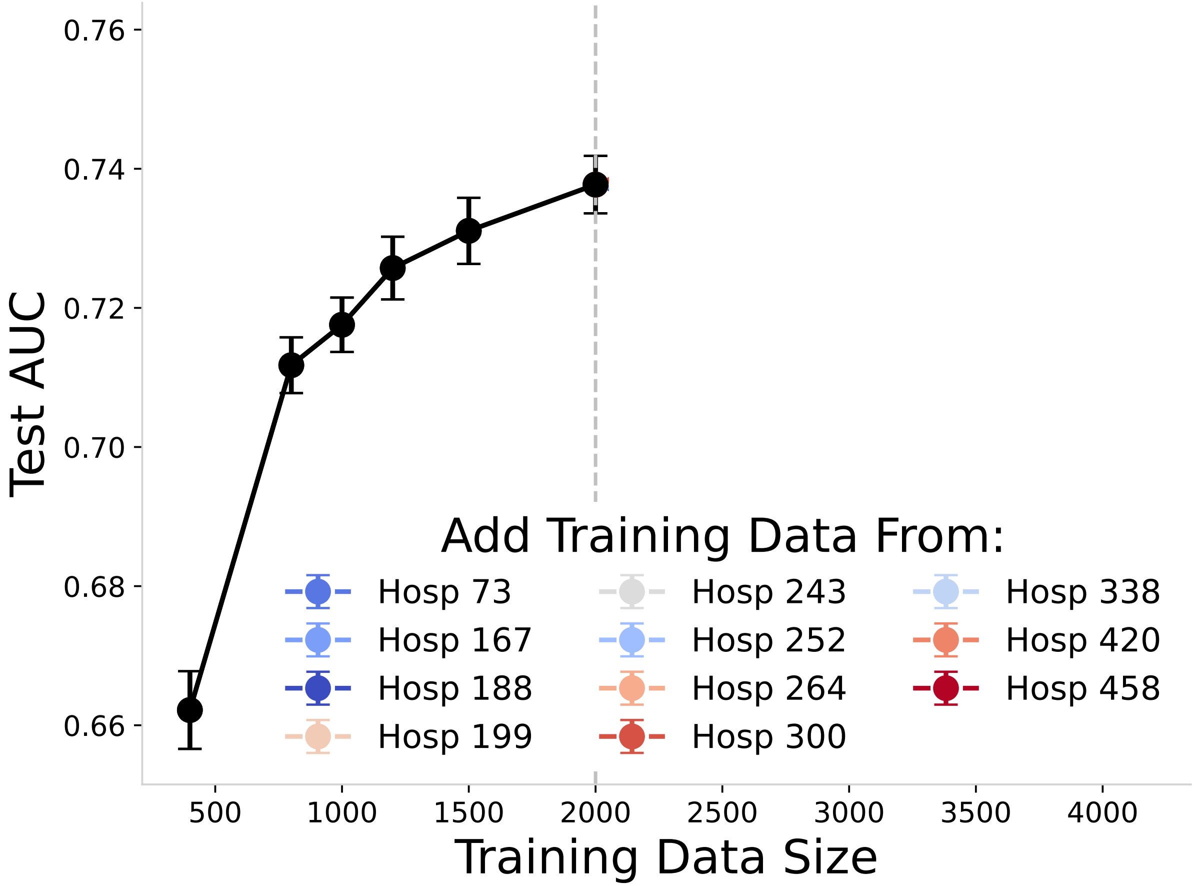


0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size

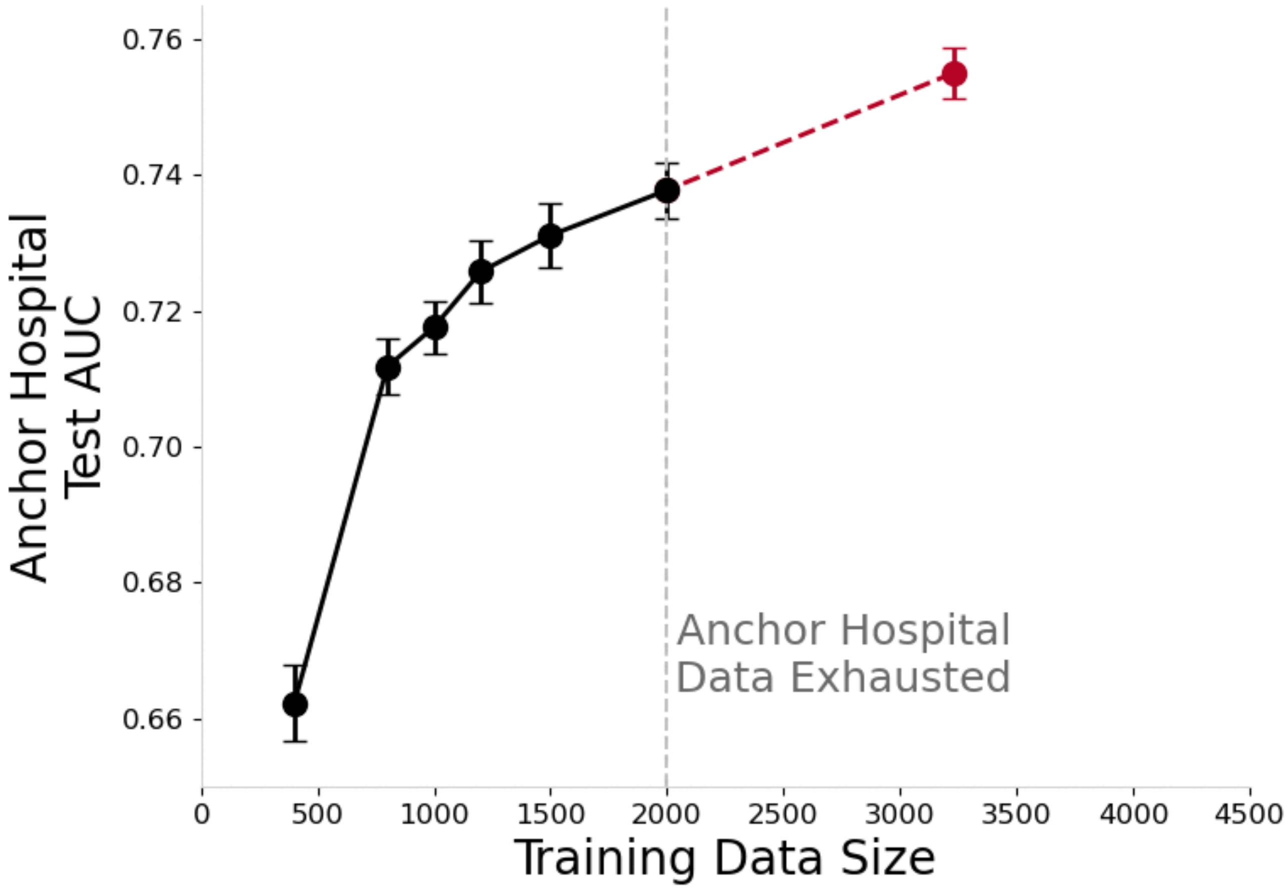


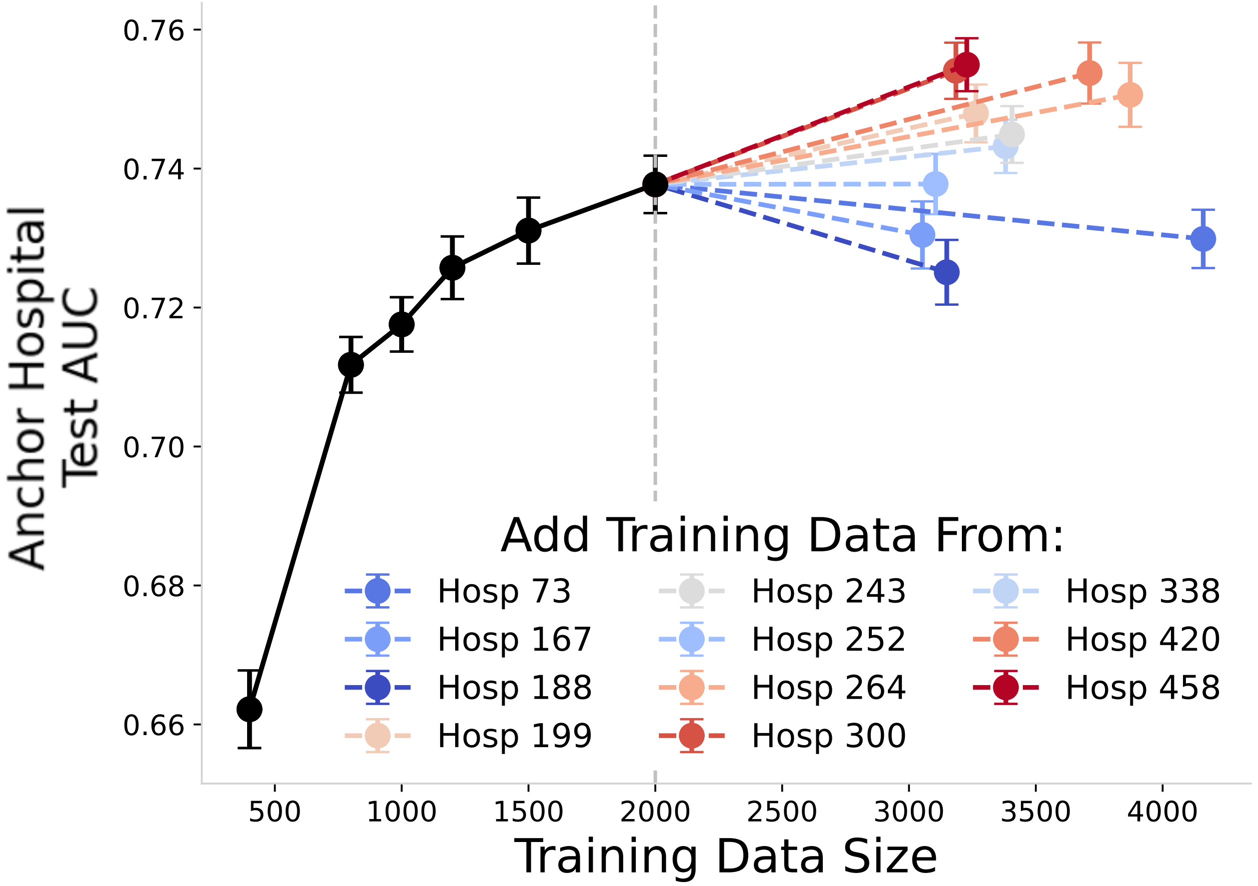
0.76
0.74
0.72
0.70
0.68
0.66
Test AUC
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size


Test AUC
0.76
0.74
0.72
0.70
0.68
0.66
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size
Hospital Data
Exhausted

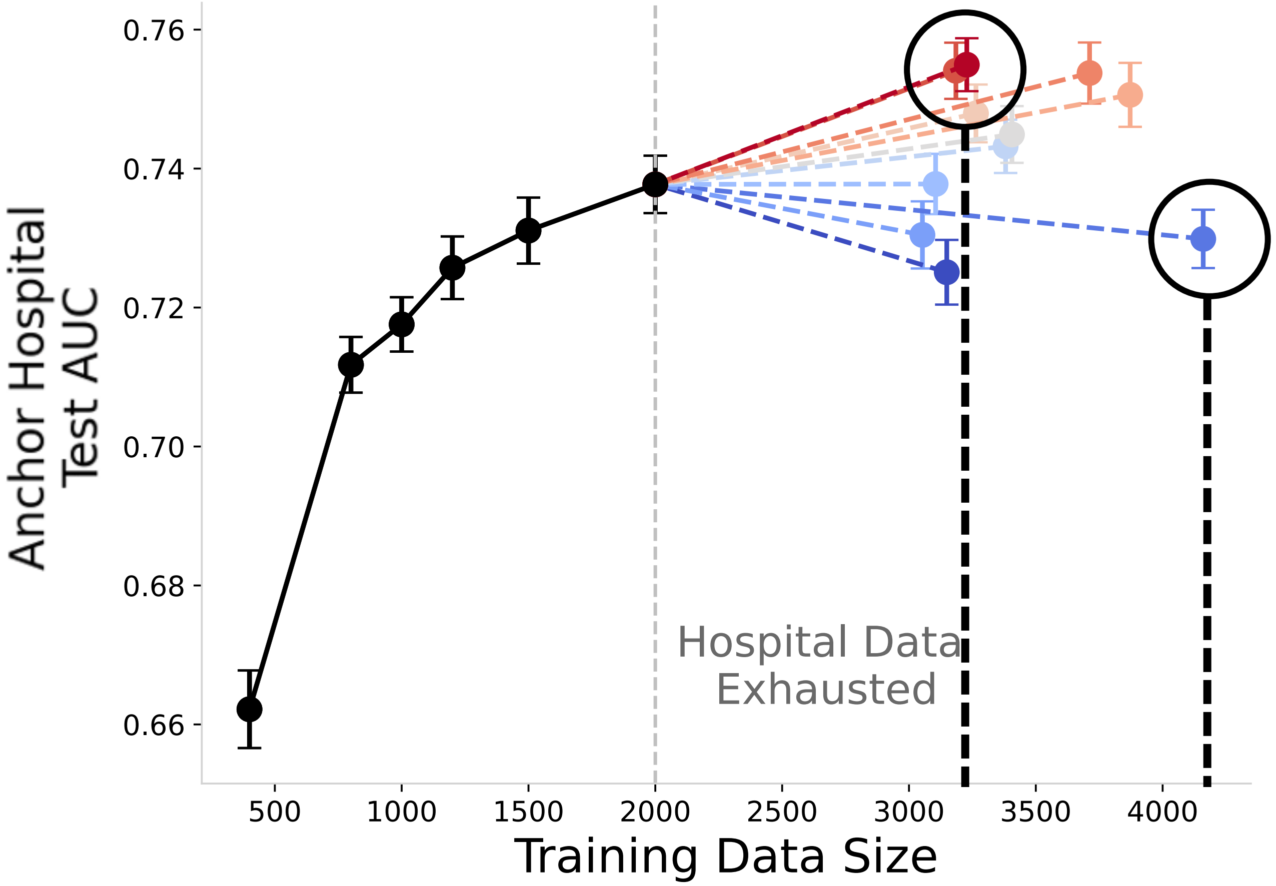


Test AUC
0.76
0.74
0.72
0.70
0.68
0.66
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size
Add Training Data From:
Hosp 73
Hosp 167
Hosp 188
Hosp 199
Hosp 243
Hosp 252
Hosp 264
Hosp 300
Hosp 338
Hosp 420
Hosp 458


Anchor Hospital
Test AUC
0.76
0.74
0.72
0.70
0.68
0.66
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Training Data Size
Anchor Hospital
Data Exhausted

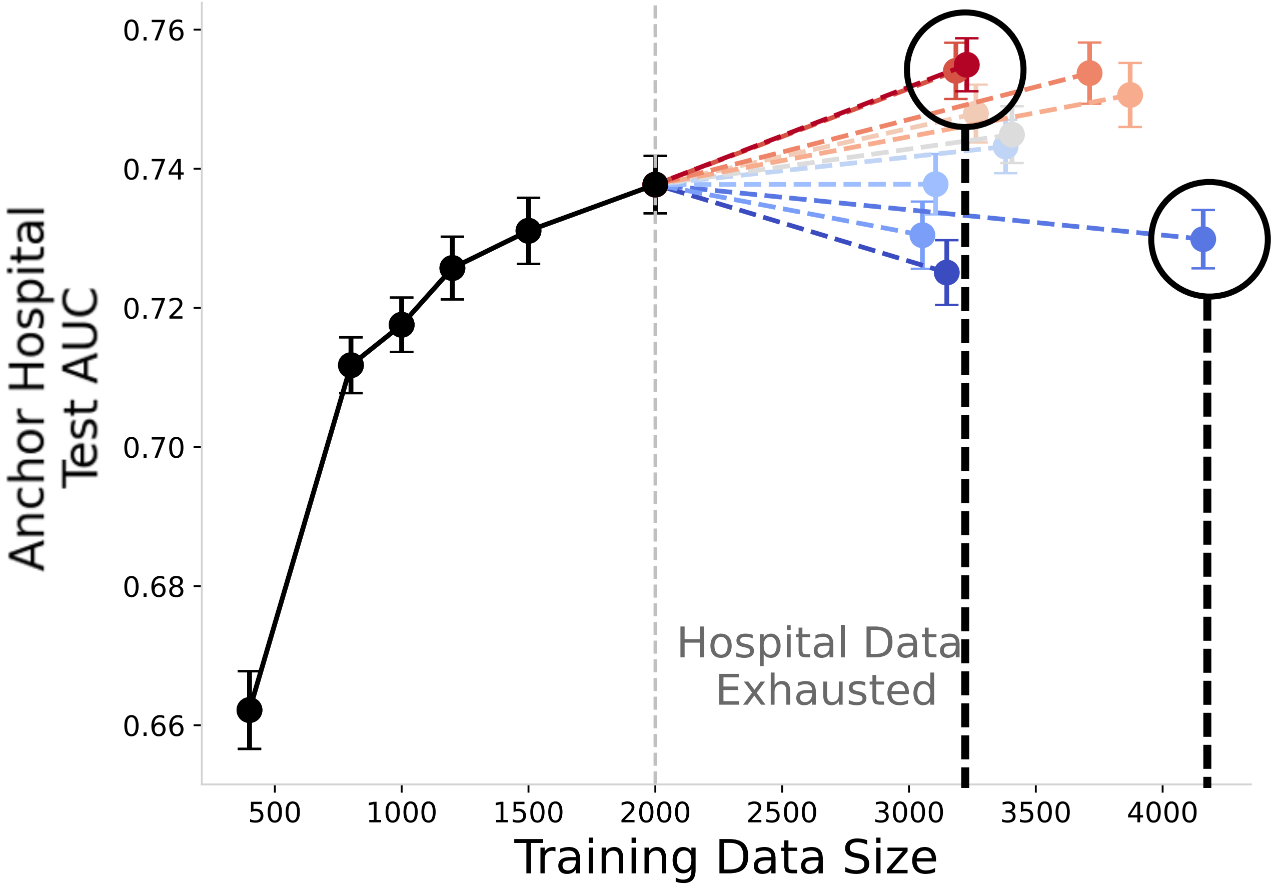


Anchor Hospital
Test AUC
0.76
0.74
0.72
0.70
0.68
0.66
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size
Add Training Data From:
Hosp 73
Hosp 167
Hosp 188
Hosp 199
Hosp 243
Hosp 252
Hosp 264
Hosp 300
Hosp 338
Hosp 420
Hosp 458



Anchor Hospital
Test AUC
0.76
0.74
0.72
0.70
0.68
0.66
500
1000
1500
2000
2500
3000
3500
4000
Training Data Size
Hospital Data
Exhausted

We call this the **Data Addition Dilemma**



Also seen in: Multilingual Translation

Better than pairwise methods

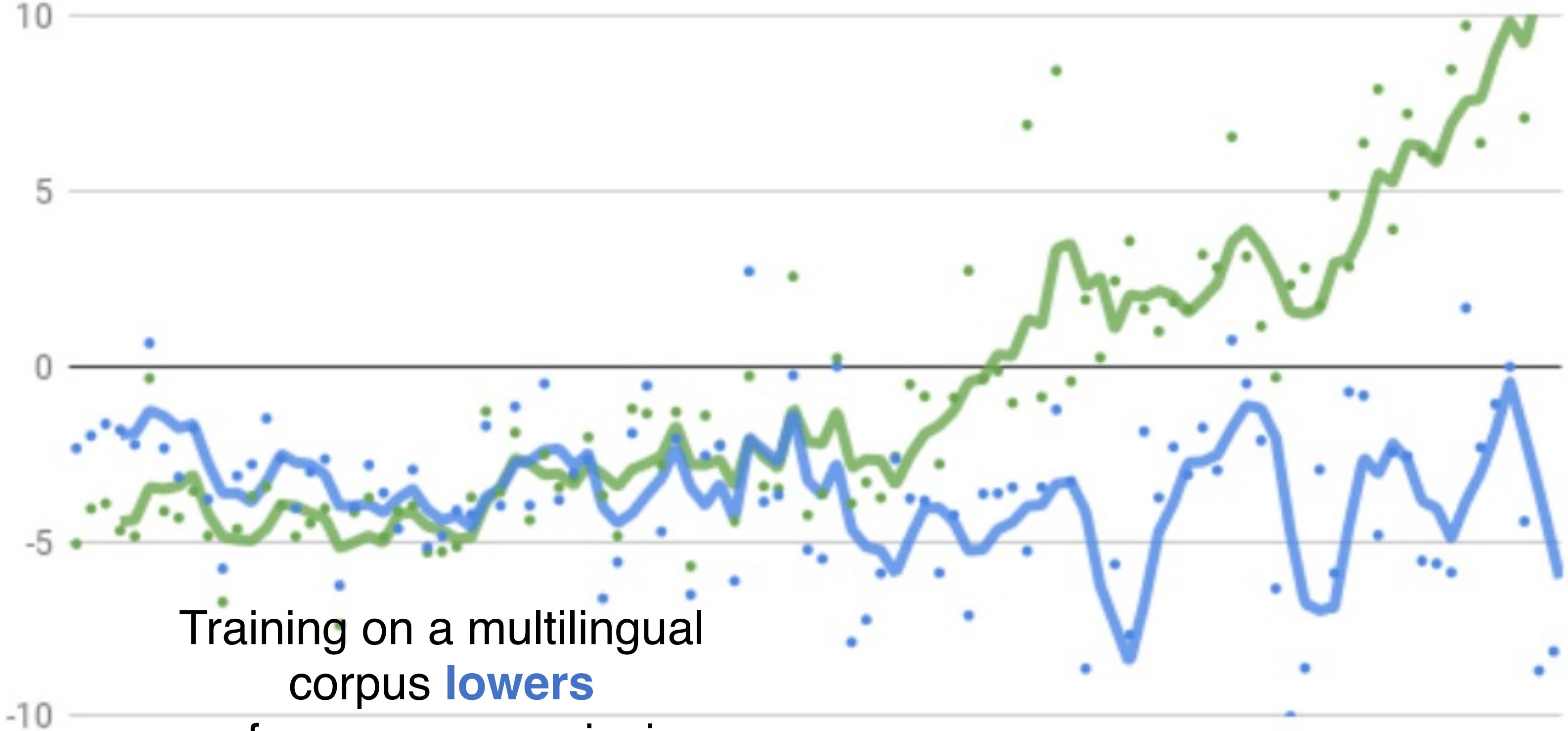
Less available training data

Worse than pairwise methods

Arivazhagan et al, “Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges” NAACL 2019

Also seen in: Multilingual Translation

Better than pairwise methods



Less available training data

Training on a multilingual corpus **lowers** performance vs. pairwise methods

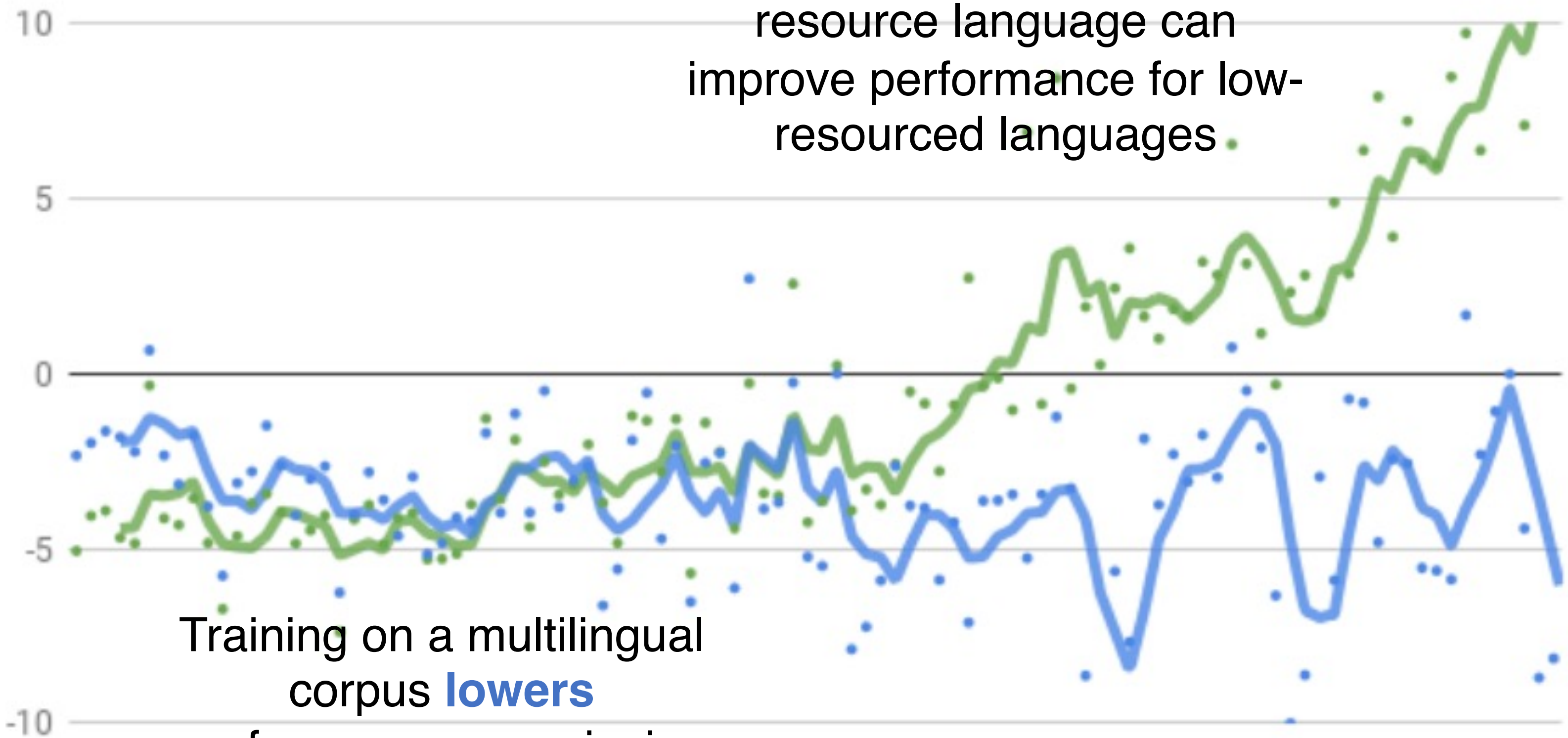
Worse than pairwise methods

Arivazhagan et al, "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges" NAACL 2019

Also seen in: Multilingual Translation

Better than pairwise methods

Worse than pairwise methods



Oversampling on low-resource language can improve performance for low-resourced languages

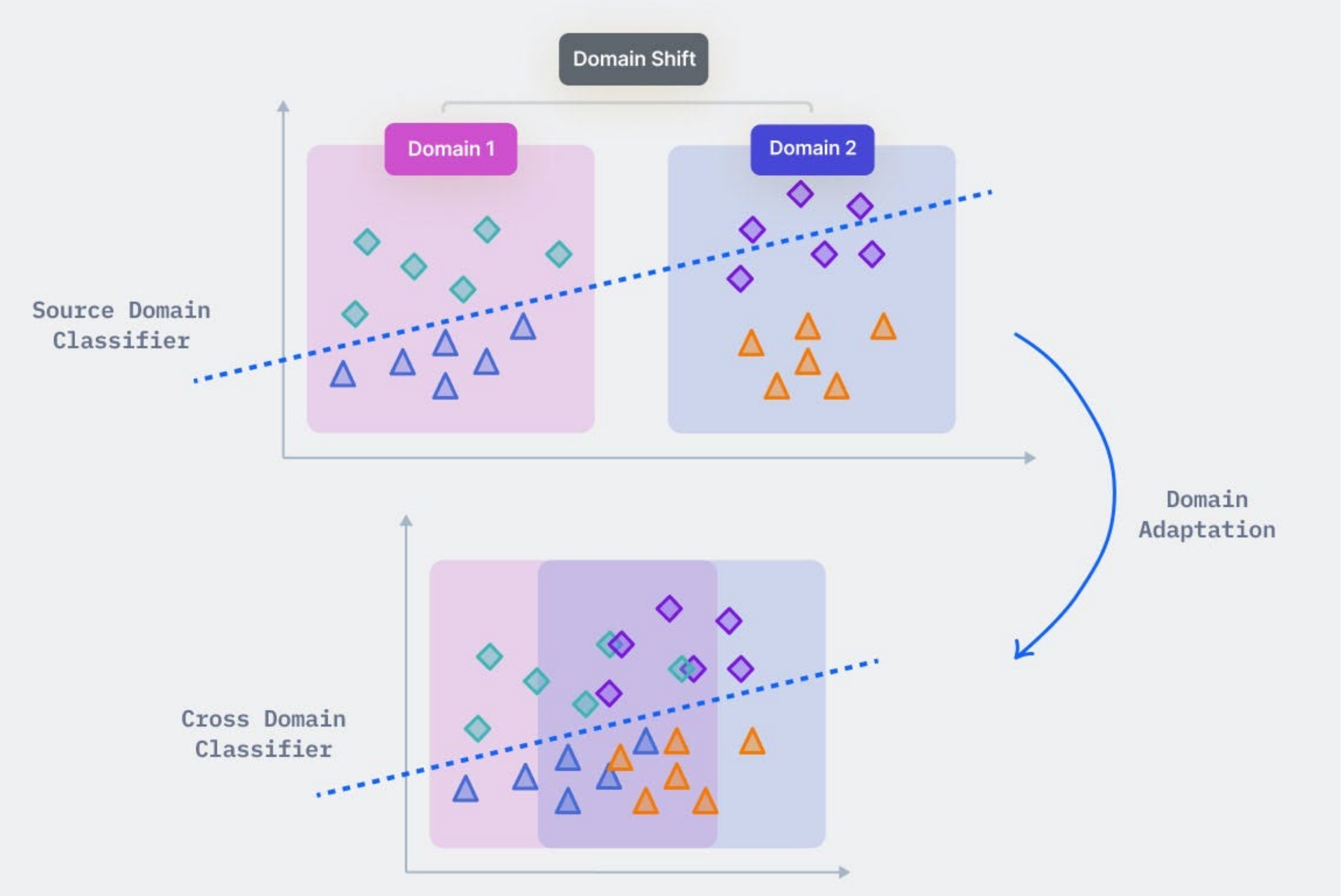
Training on a multilingual corpus **lowers** performance vs. pairwise methods

Less available training data

Arivazhagan et al, “Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges” NAACL 2019

Related: Domain Adaptation

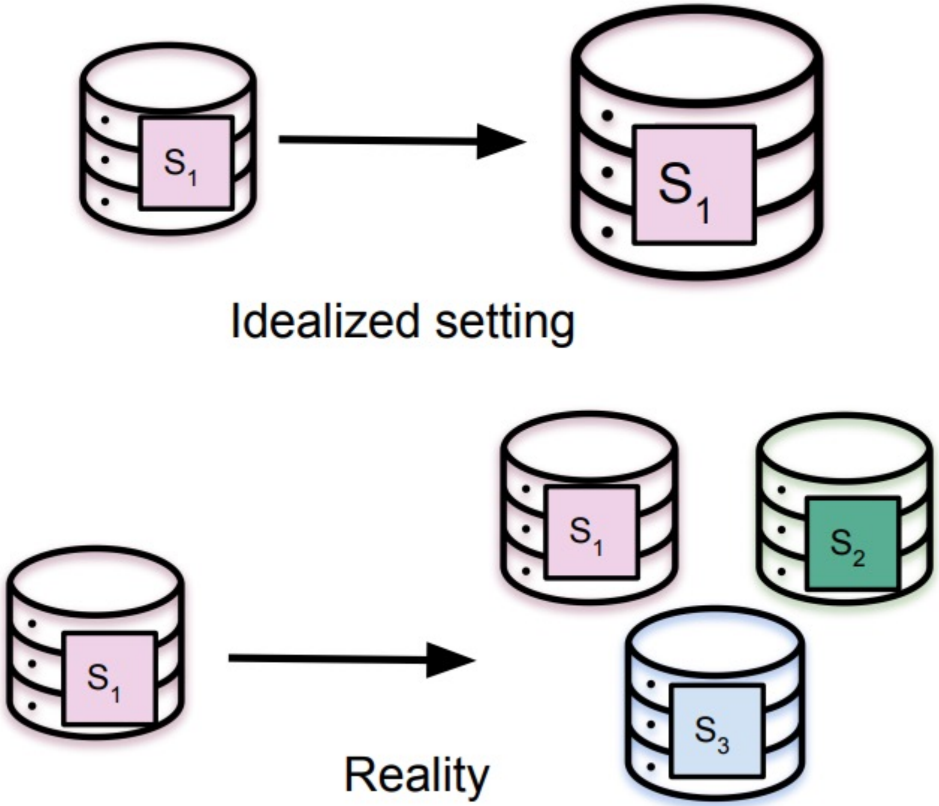
- DA goal is to make a model that has **high performance across multiple domains**
- Our problem formulation only is interested in performance in **one domain**



https://medium.com/@sayampalrecha09/what-is-domain-adaptation-2c0ef28eeb42

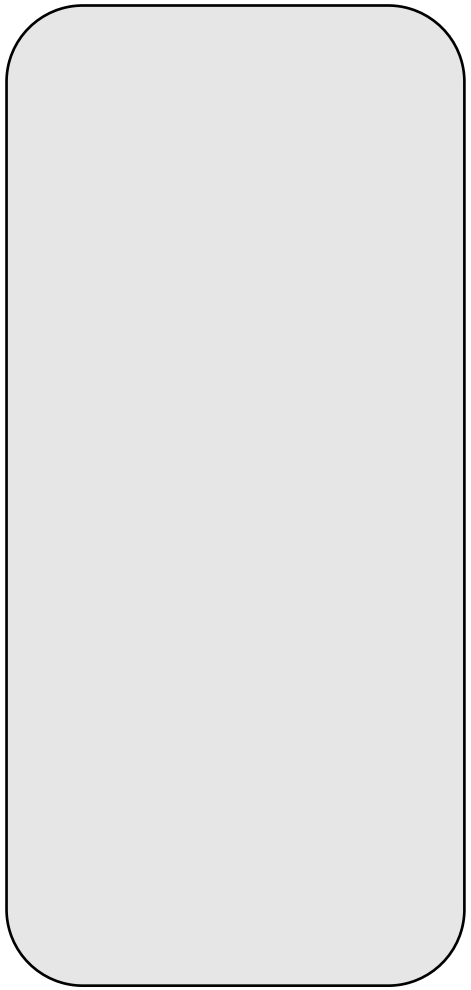
The Data Addition Dilemma

Framework to weigh **cost and benefit** of adding more data

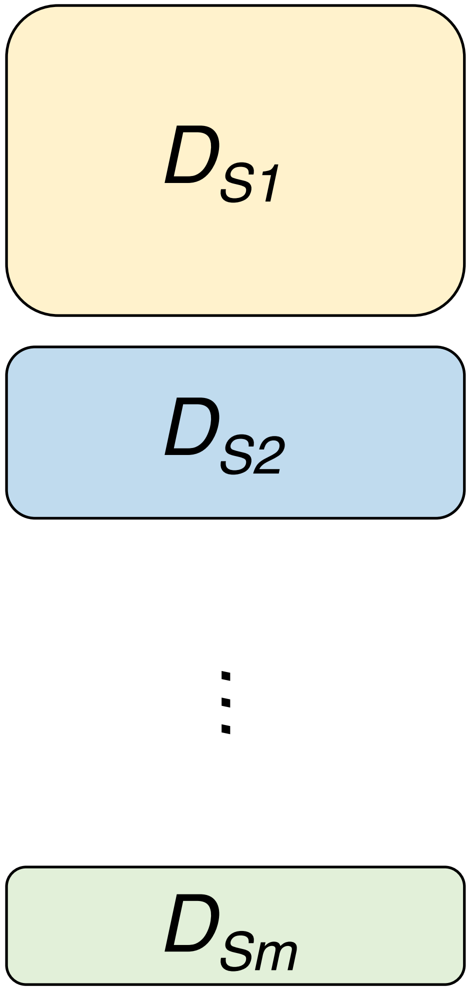


Modeling Data Accumulation

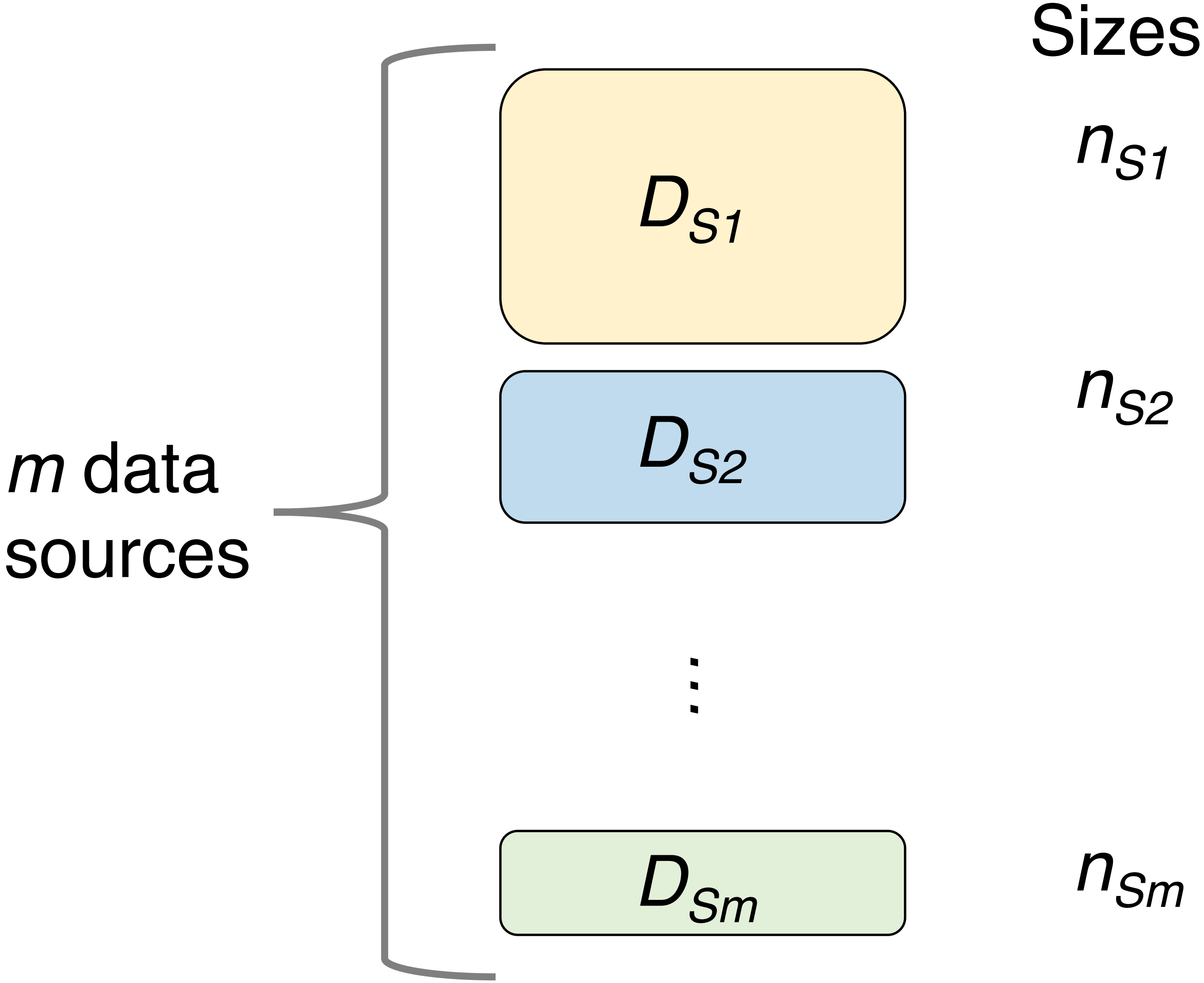
$\{x,y\}^{n} \sim D$



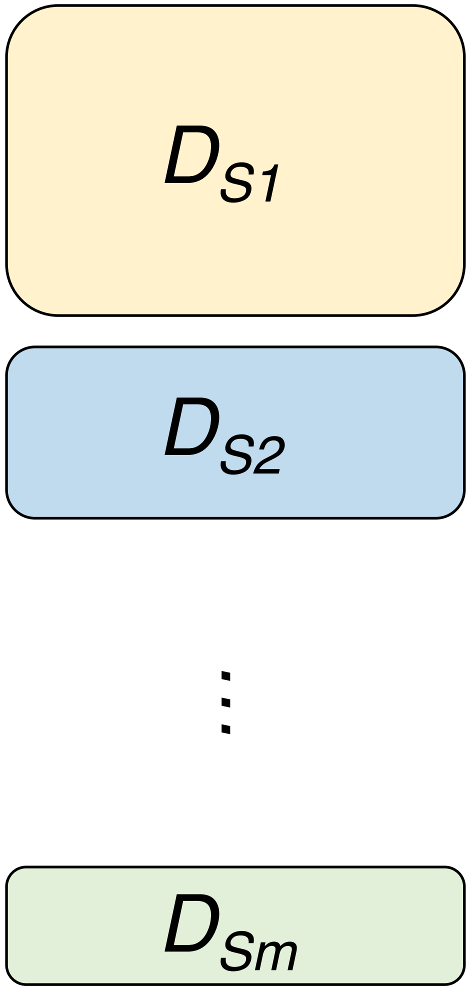
Modeling Data Accumulation



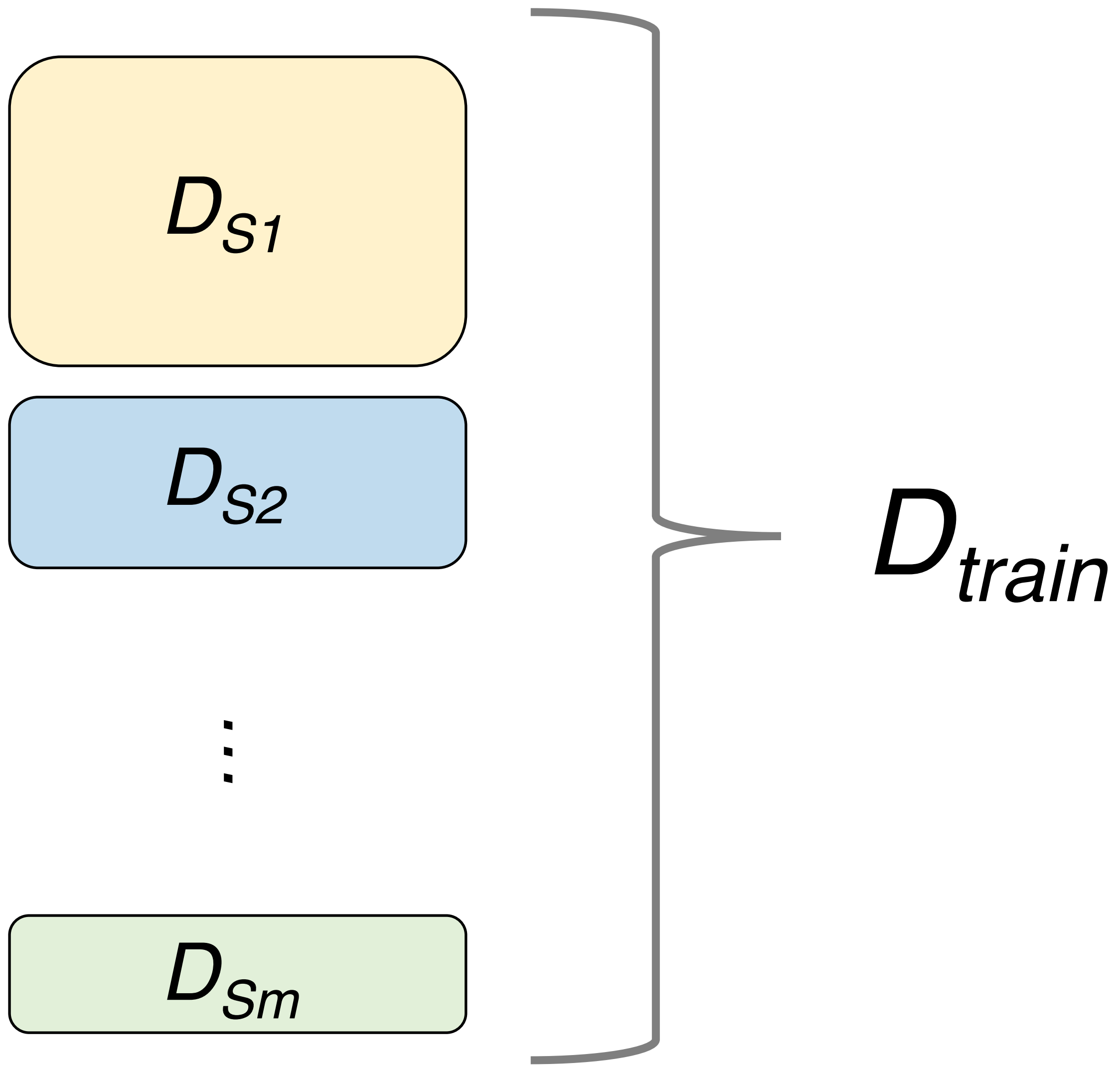
Modeling Data Accumulation



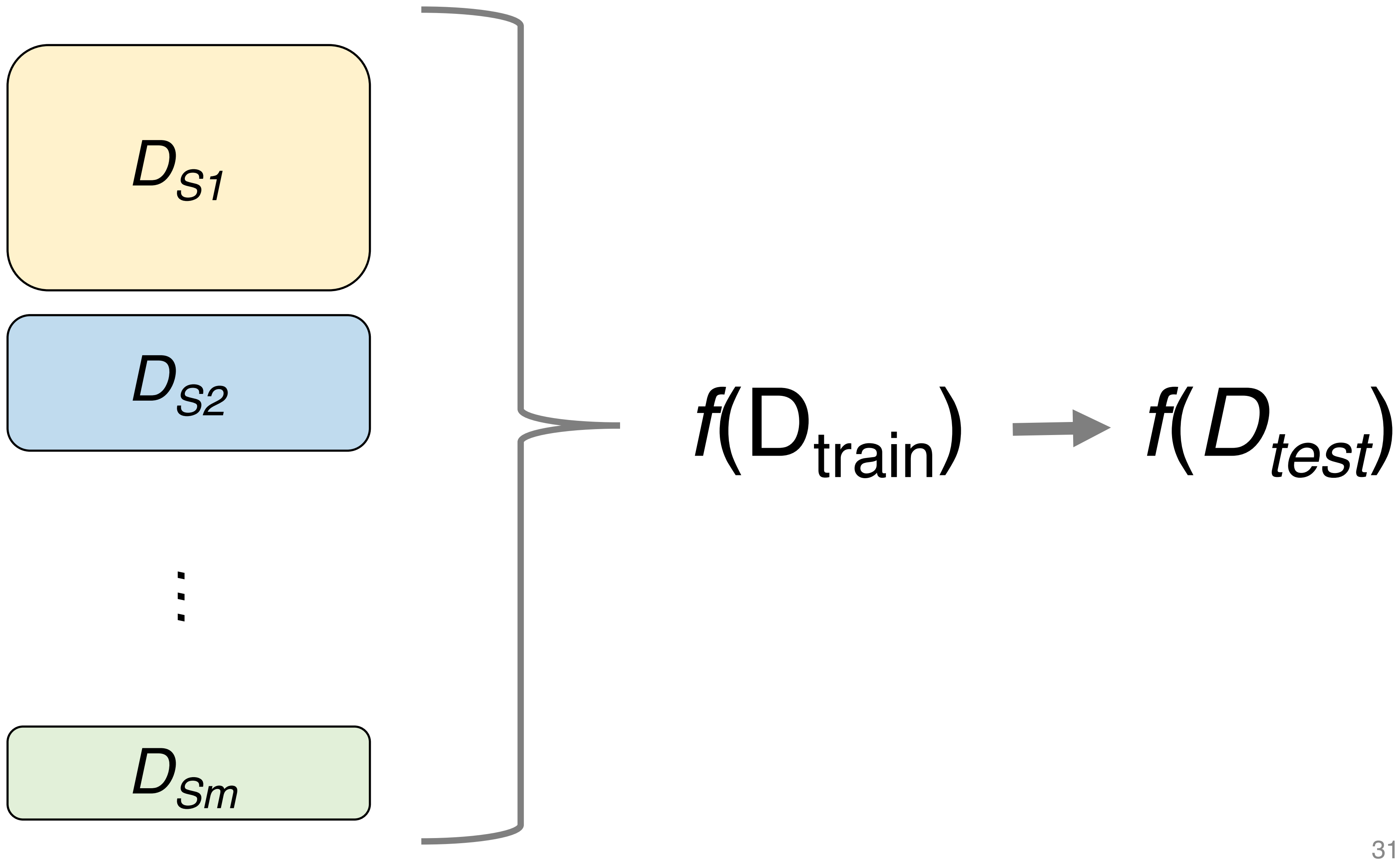
Modeling Data Accumulation



Modeling Data Accumulation



Modeling Data Accumulation



$$AUC(D_{H0} \cup \ldots \boxed{\cup D_{HM}}, D_{Htest}) - AUC(D_{H0} \cup \ldots \boxed{\cup D_{H\,M-1}}, D_{Htest})$$

AUC with adding last data source – AUC without adding last data source

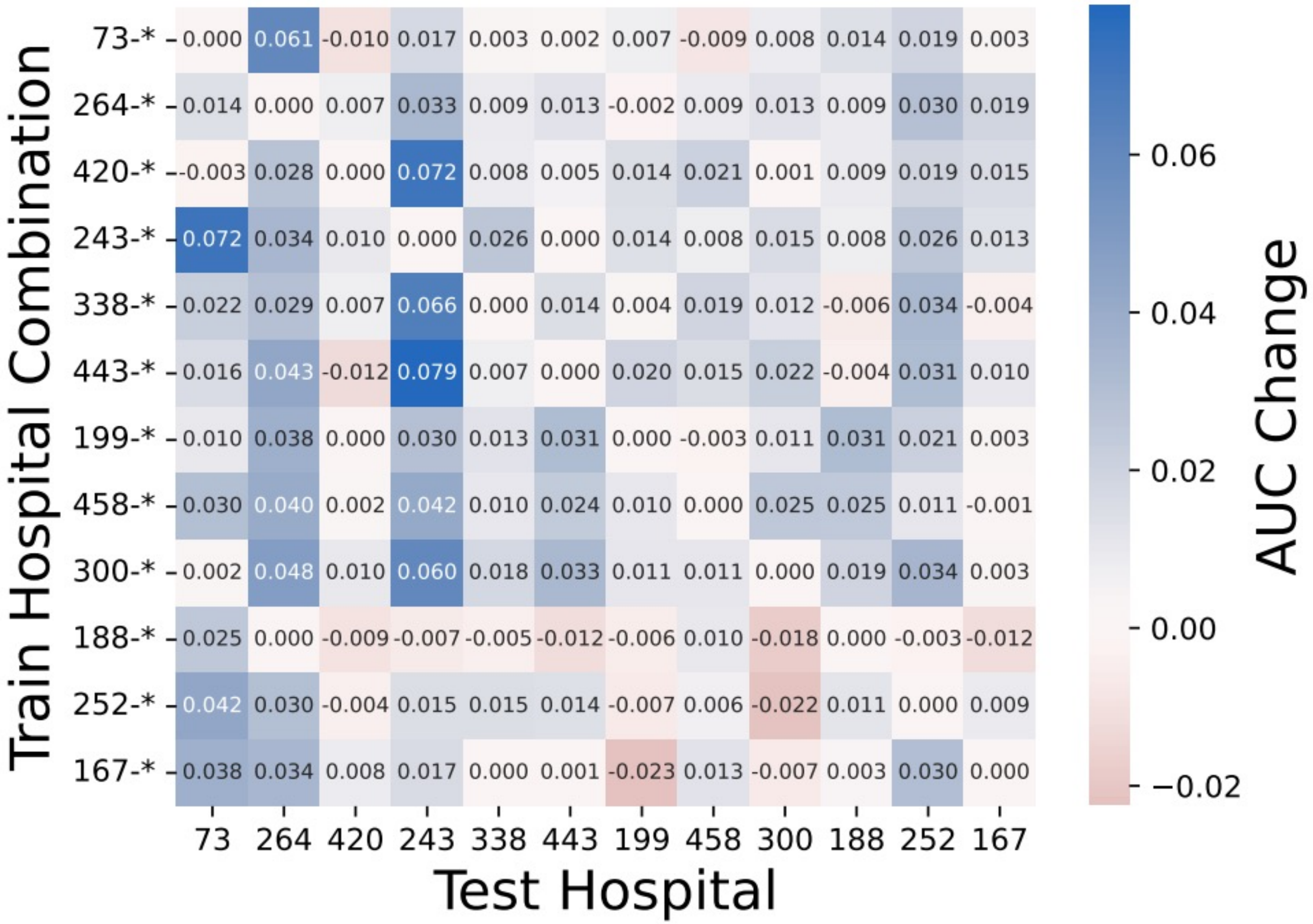
$$AUC(D_{H0} \cup \ldots \cup D_{HM}, D_{Htest}) - AUC(D_{H0} \cup \ldots \cup D_{H\,M-1}, D_{Htest})$$

$$\approx$$

$$AUC(D_{H0 \cup H1}, D_{H0}) - AUC(D_{H0}, D_{H0})$$

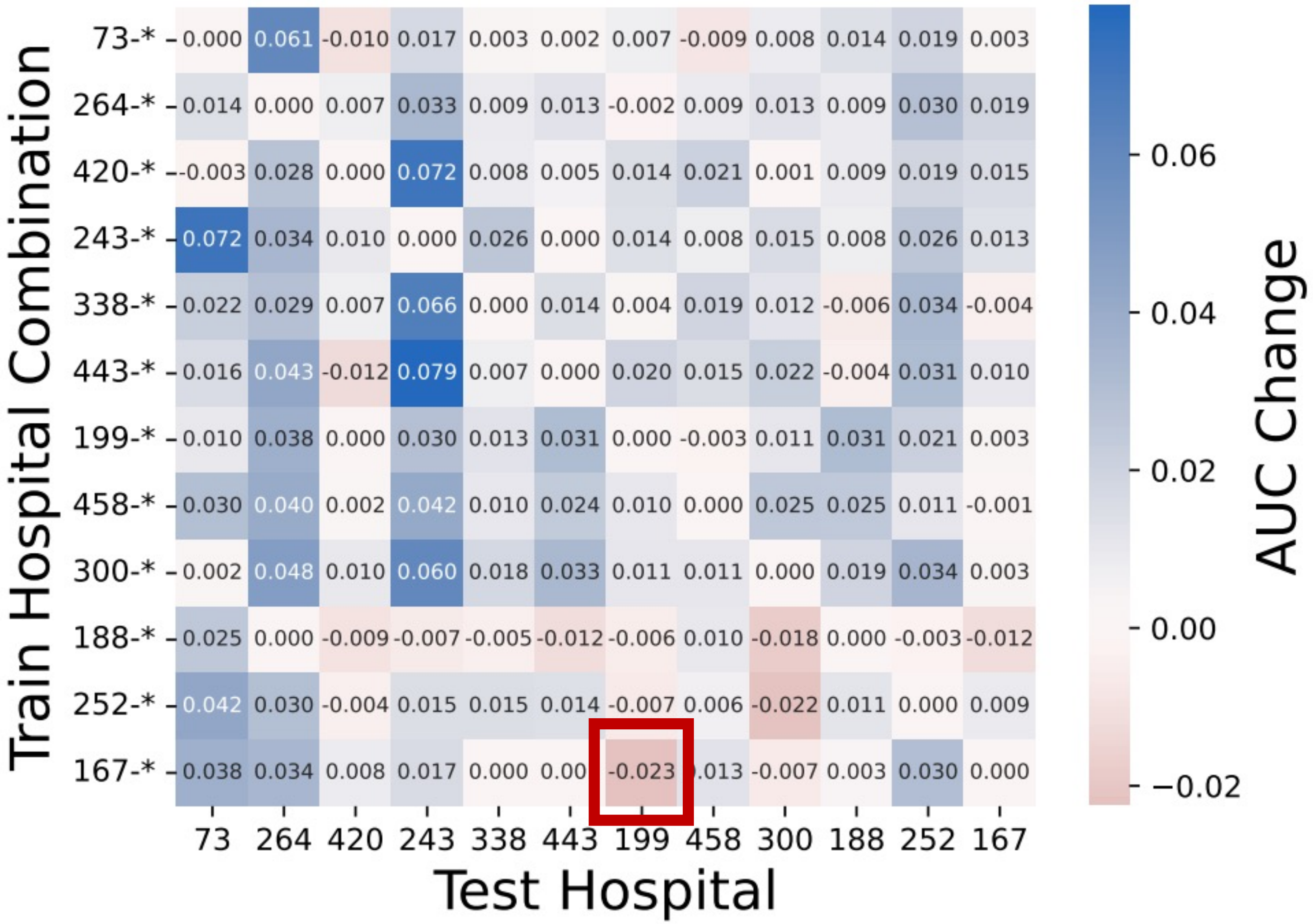
AUC with adding a data source

AUC without adding a data source



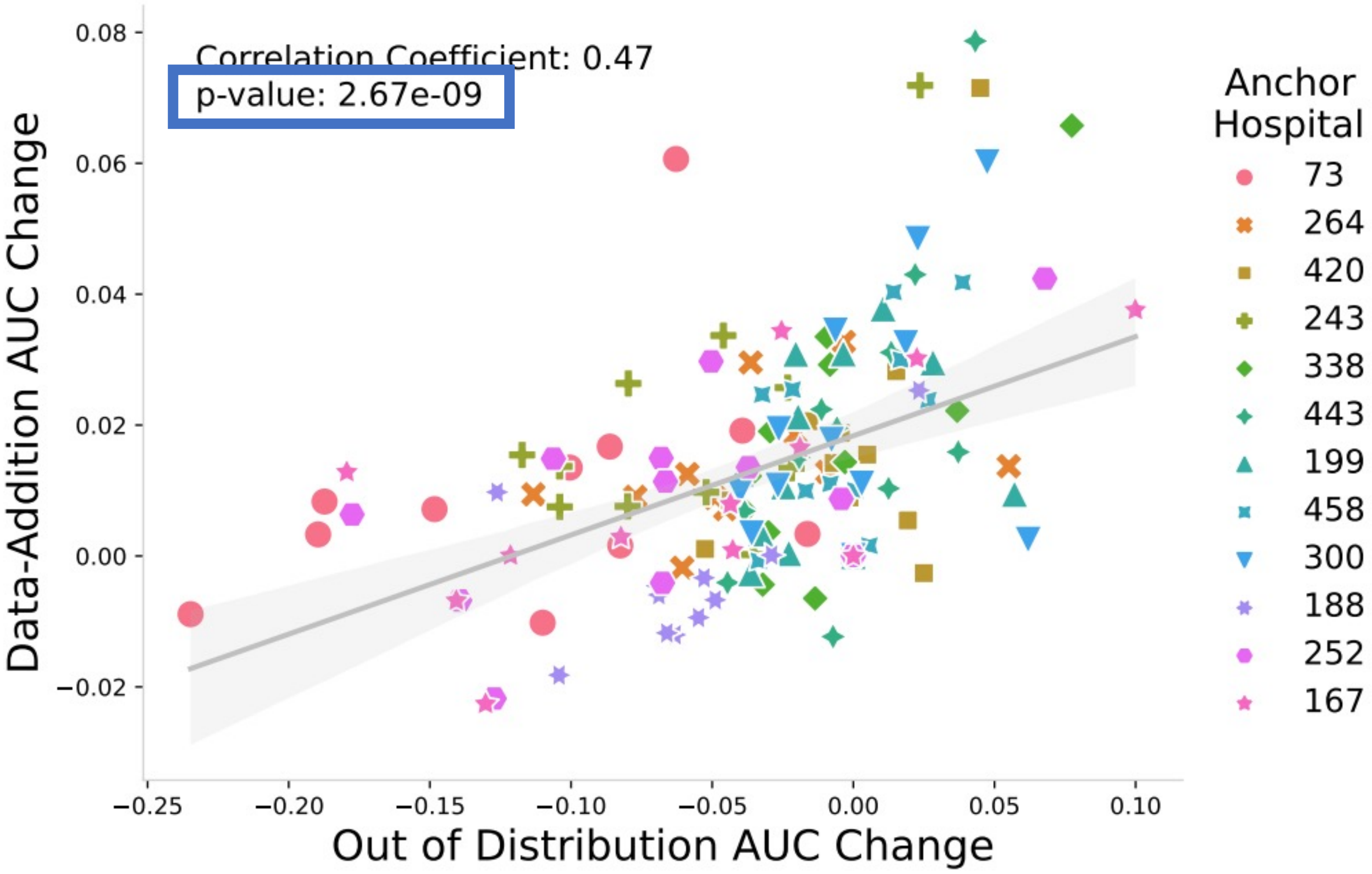
24-hour mortality prediction task on eICU dataset

Adding hospital 167 makes performance **worse** for hospital 199



24-hour mortality prediction task on eICU dataset

AUC change from adding hospitals is **related** to AUC change from out of distribution



24-hour mortality prediction talk on eICU dataset

$$AUC(D_{H0} \cup \ldots \cup D_{HM}, D_{Htest}) - AUC(D_{H0} \cup \ldots \cup D_{H\,M-1}, D_{Htest})$$

$$\approx$$

$$AUC(D_{H0 \cup H1}, D_{H0}) - AUC(D_{H0}, D_{H0})$$

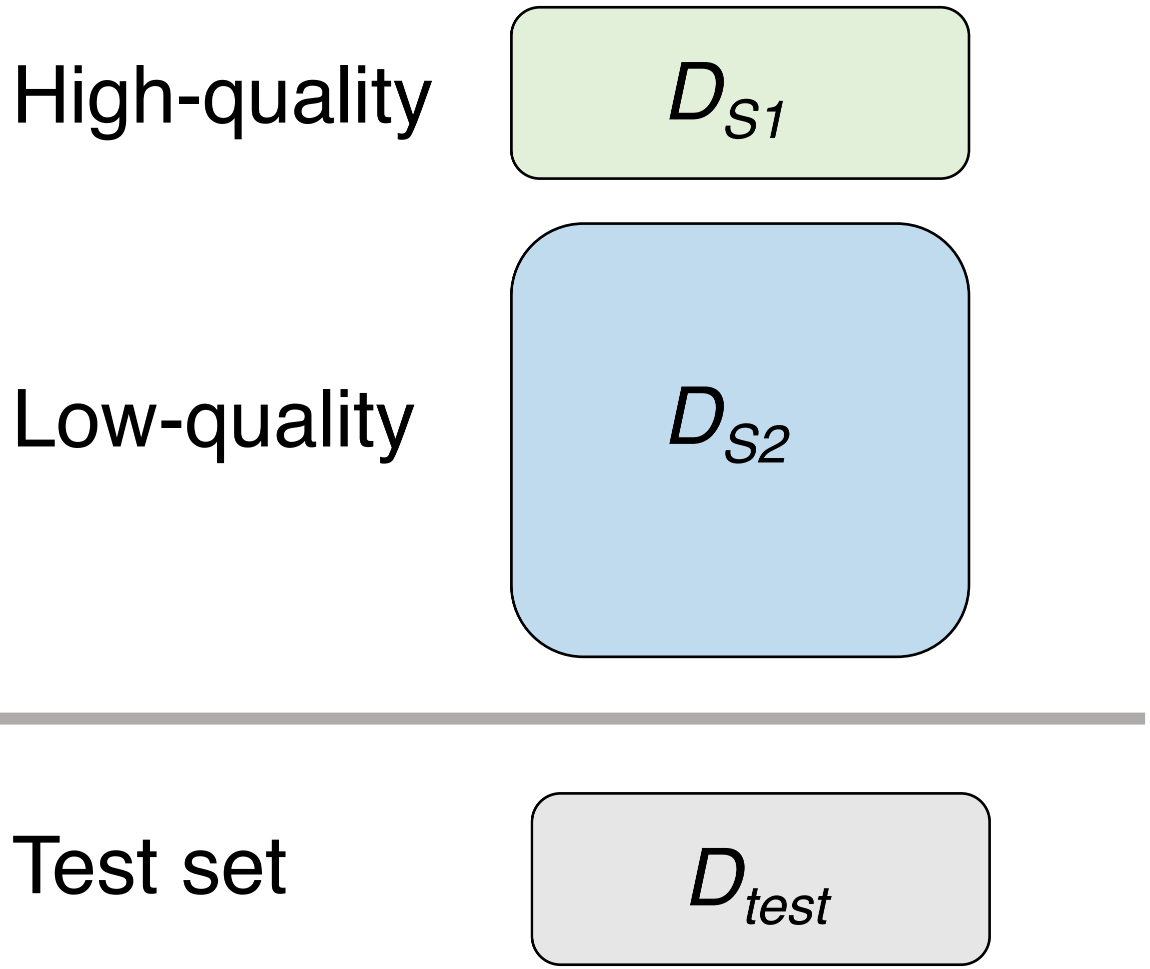
$$\approx$$

$$\delta(D_{H0}, D_{H1})$$

Divergence between datasets

Toy Example

What is $\delta(D_{train}, D_{test})$?



Toy Example

What is $\delta(D_{train}, D_{test})$?

High-quality D_{S1}

Low-quality D_{S2}

Test set D_{test}

If training data size ≤ n_{S1}:

$\delta(D_{S1}, D_{test})$

Toy Example

High-quality D_{S1}

Low-quality D_{S2}

Test set D_{test}

What is $\delta(D_{train}, D_{test})$?

If training data size ≤ $\boldsymbol{n_{S1}}$:

$$\delta(\boldsymbol{D_{S1}}, D_{test})$$

If training data size > $\boldsymbol{n_{S1}}$:

$$\frac{n_{S1}}{n}\,\delta(\boldsymbol{D_{S1}}, D_{test}) + \left(1 - \frac{n_{S1}}{n}\right)\delta(\boldsymbol{D_{S2}}, D_{test})$$

*If divergence δ is composed linearly

Contribution: When is adding more data hurtful?

Lemma**:** If $\delta\,(D_{Sk}, \mathrm{D}_{\mathrm{test}}) - \frac{cn}{n_{Sk}} \geq \delta(D_{train}, D_{test})$:

$$\delta(D_{train,n}, D_{test}) \geq \delta(D_{train,n\text{-}nSk}, \mathrm{D}_{\mathrm{test}})$$

D_{Sk} = data from *k*-th source
δ = divergence in family of *f*-divergences
(not assumed linear)
c = divergence-dependent constant

Related: Distances, divergences, and discrepancies

- Relationship between increased divergence to empirical risk for f–divergences by giving a generalization bound
- Different discrepancy measures including L1 distance
- HΔH divergence
- Margin disparity discrepancy

Acuna et al, "f-domain adversarial learning: Theory and algorithms." ICML 2021.
Ben-David et al, "Analysis of representations for domain adaptation." NeurIPS 2006.
Ben-David et al, "A theory of learning from different domains." *Machine learning* 2010.
Zhang et al, "Bridging theory and algorithm for domain adaptation." ICML 2019.

Practical Strategies for Data Accumulation

Practical Strategies for Data Accumulation

Idea 1: Try all combinations

Practical Strategies for Data Accumulation

Idea 1: Try all combinations

Best 3 hospitals from 10 options

≈ **10^3 combinations**

Practical Strategies for Data Accumulation

~~**Idea 1**: Try all combinations~~

Idea 2: Hospital-level characteristics (mortality rates, demographics, etc.)

Practical Strategies for Data Accumulation

~~**Idea 1**: Try all combinations~~

Idea 2: Hospital-level characteristics

Idea 3: Divergence-based metrics

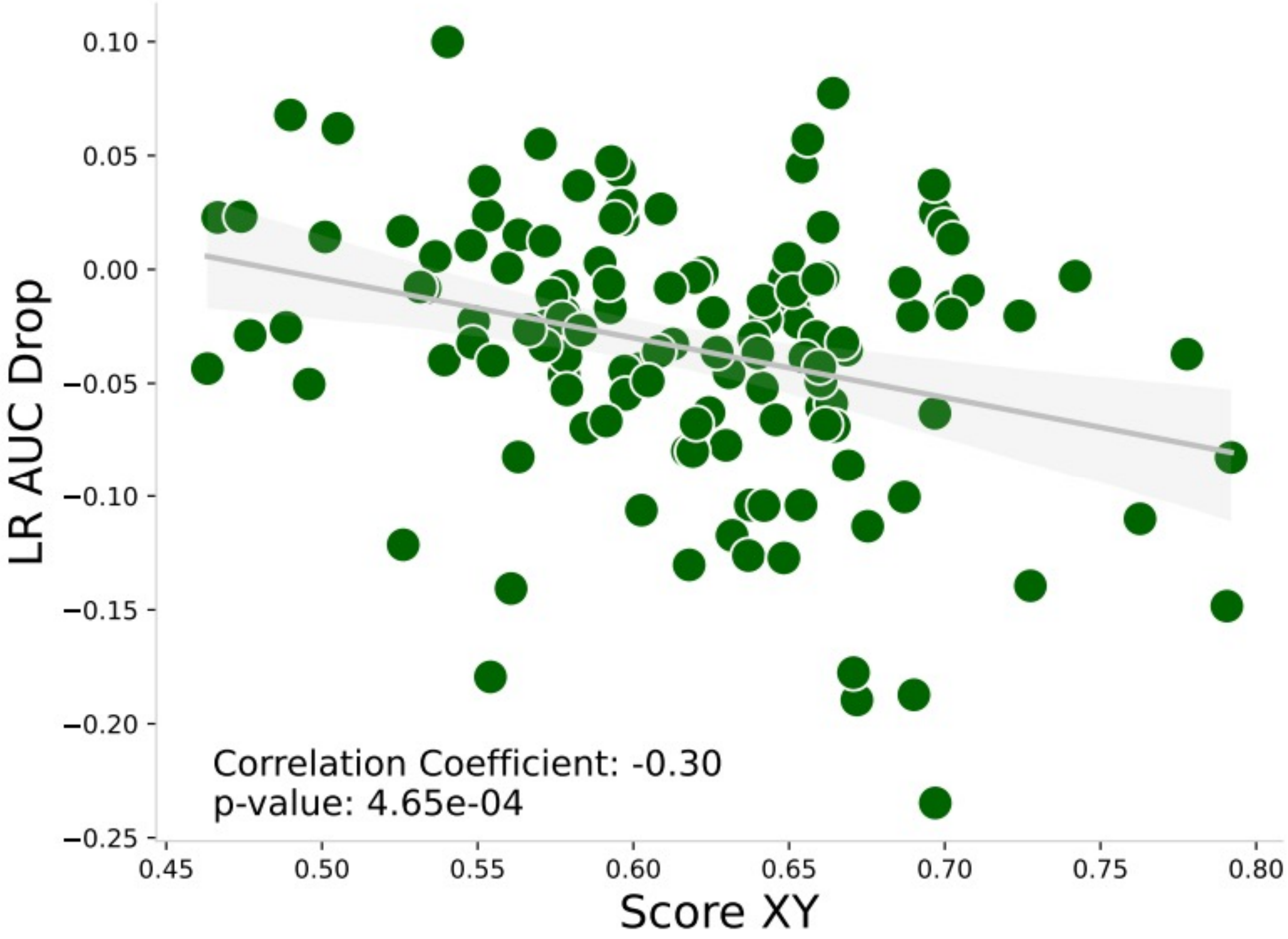
$$\text{KL Ratio X} = E_{x \in P}\left(\log \frac{P(x)}{Q(x)}\right)$$

$$\text{KL Ratio XY} = E_{(x,y) \in P}\left(\log \frac{P(x,y)}{Q(x,y)}\right)$$

$$\text{Score X} = E_{x \in P}\big(P(x \in P)\big)$$

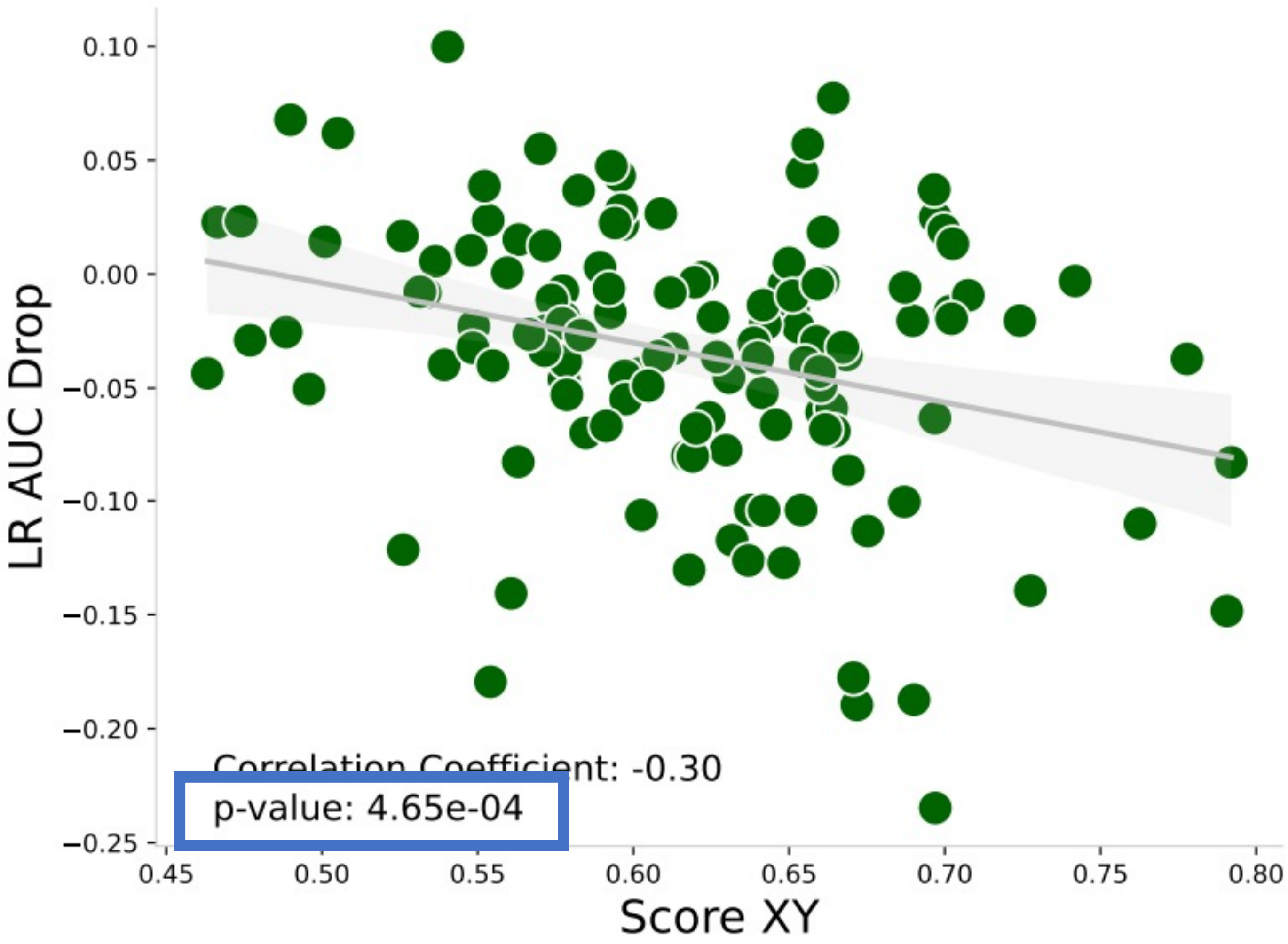
$$\text{Score XY} = E_{(x,y) \in P}\big(P((x,y) \in P)\big)$$

LR AUC drop is **correlated** with ScoreXY, a divergence-based metric



24-hour mortality prediction talk on eICU dataset

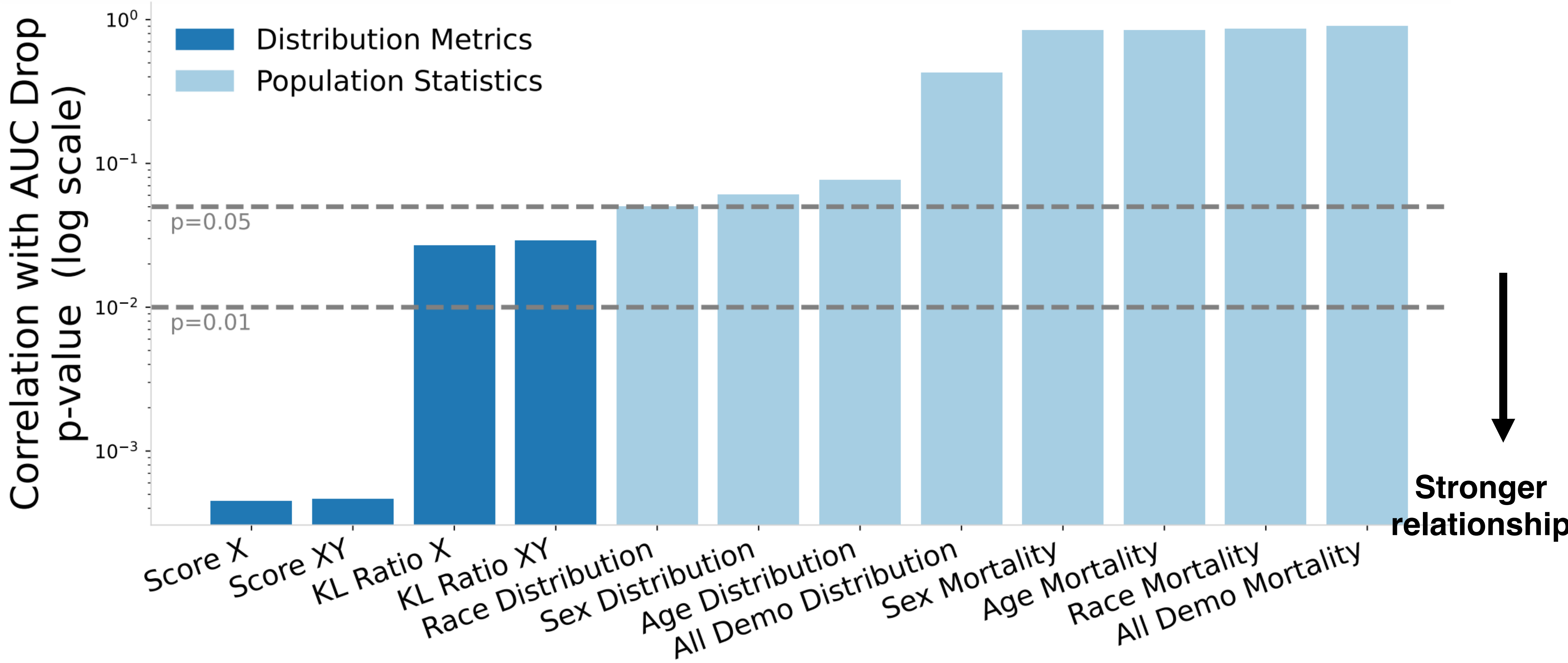
LR AUC drop is **correlated** with ScoreXY, a divergence-based metric



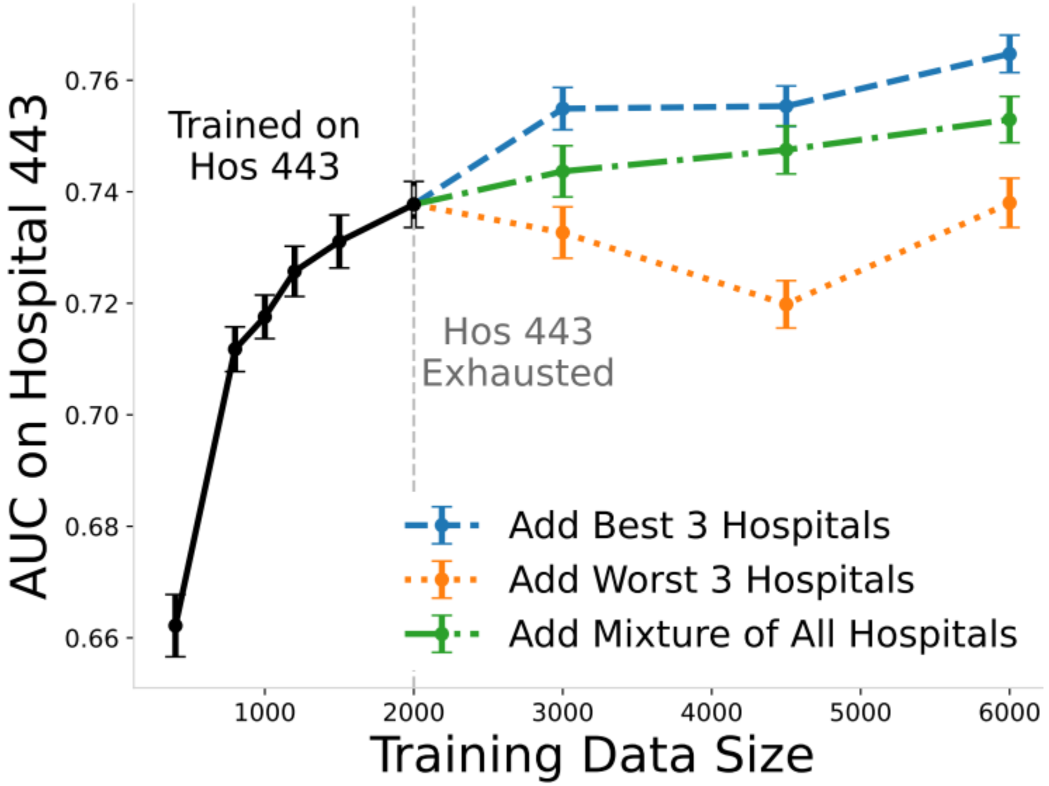
24-hour mortality prediction talk on eICU dataset


Correlation with AUC Drop
p-value (log scale)
10^0
10^{-1}
10^{-2}
10^{-3}
Score X
Score XY
KL Ratio X
KL Ratio XY
Race Distribution
Sex Distribution
Age Distribution
All Demo Distribution
Sex Mortality
Age Mortality
Race Mortality
All Demo Mortality
Stronger
relationship

Divergence metrics out-perform other hospital-wide characteristics

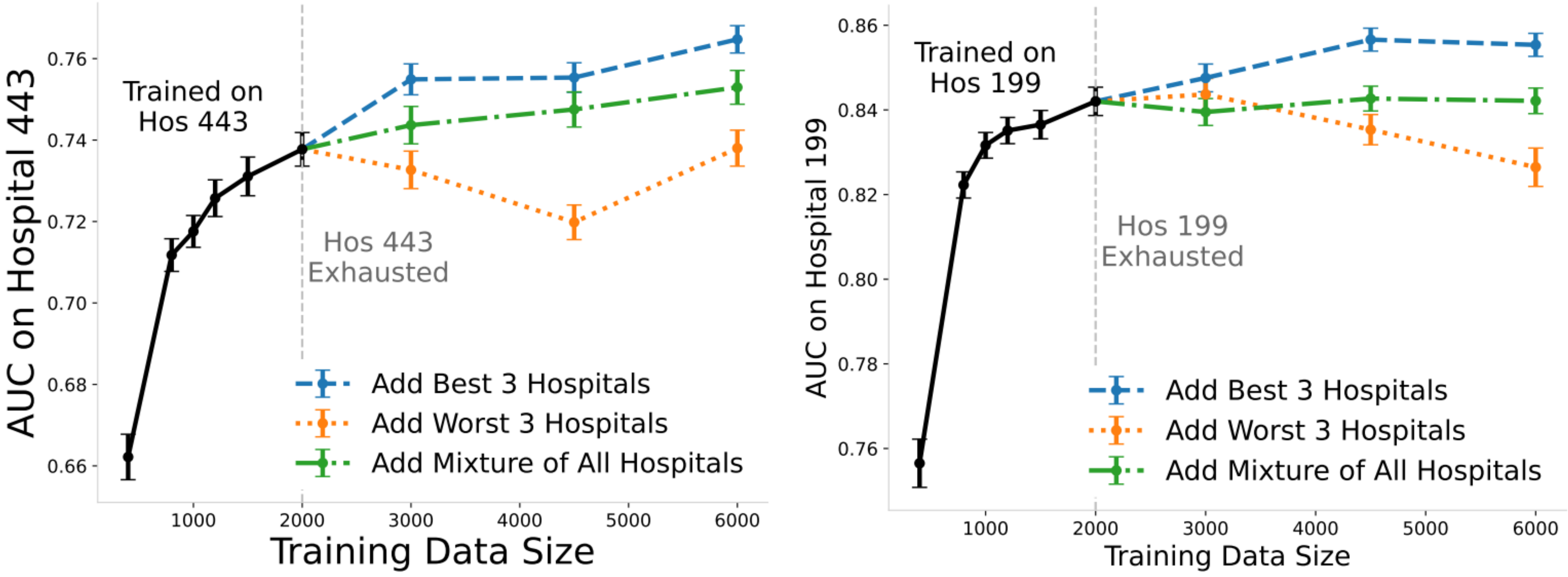


Divergence-based metrics outperform mixture settings



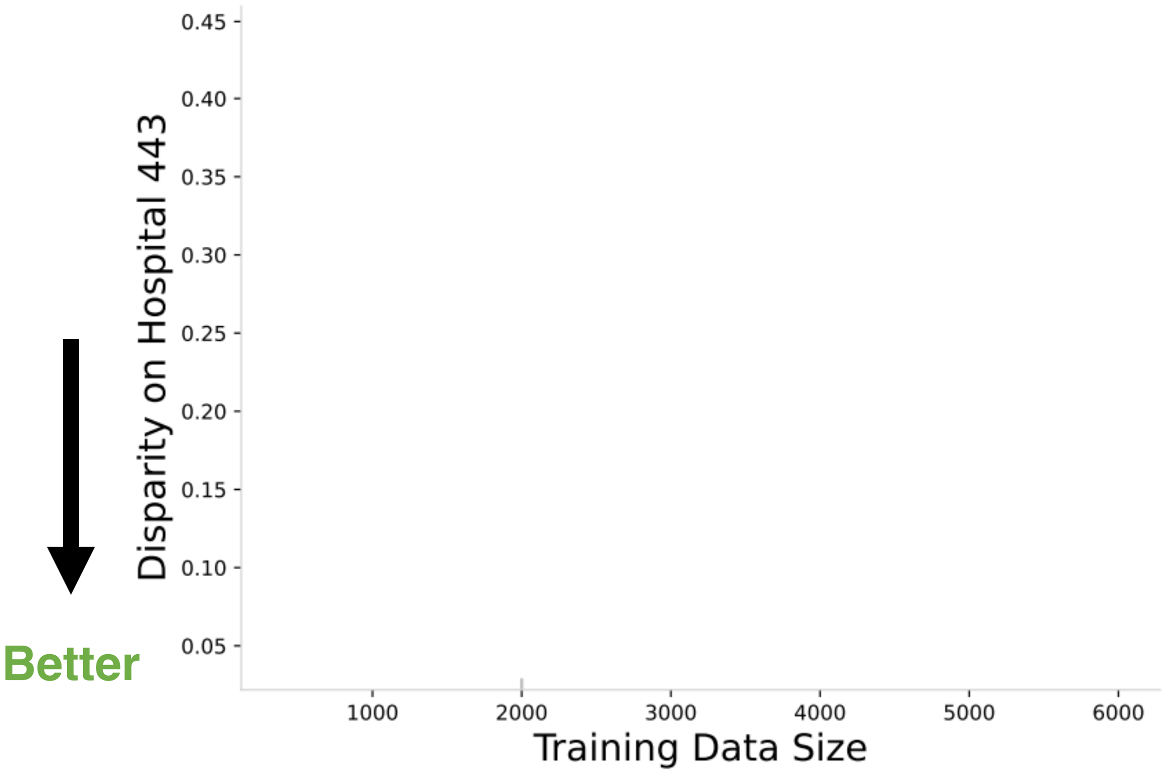
24-hour mortality prediction talk on eICU dataset

Divergence-based metrics outperform mixture settings



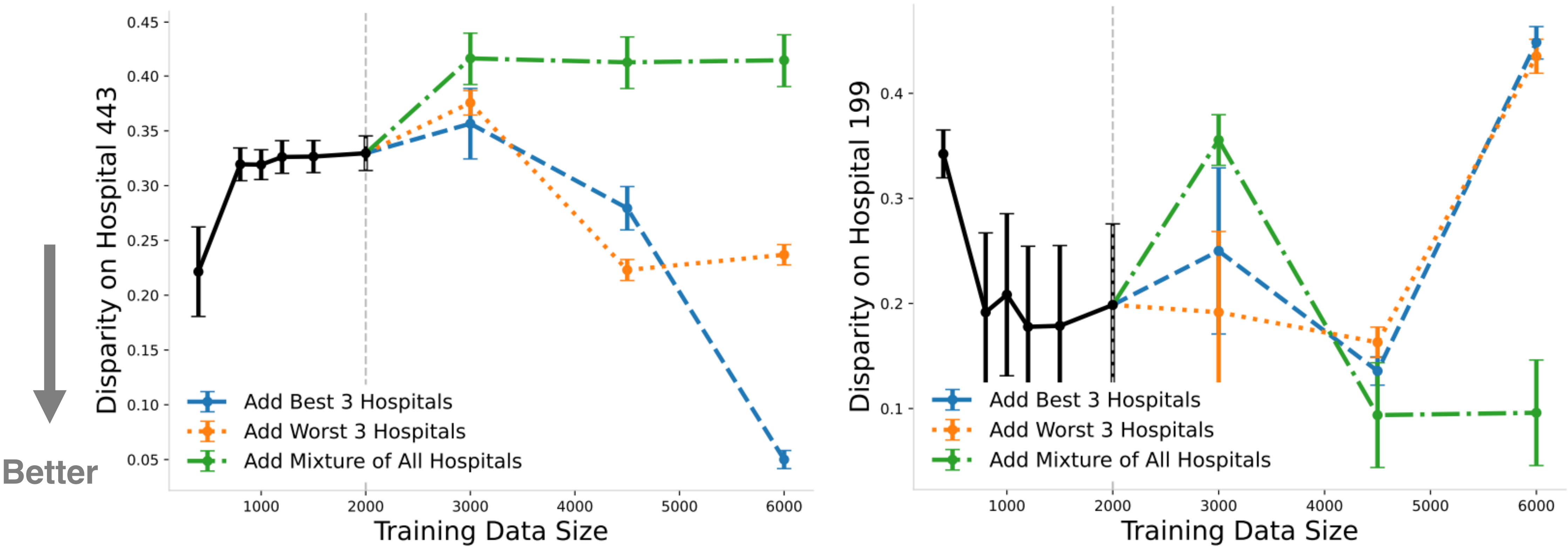
24-hour mortality prediction talk on eICU dataset

But! Racial disparity can have mixed results



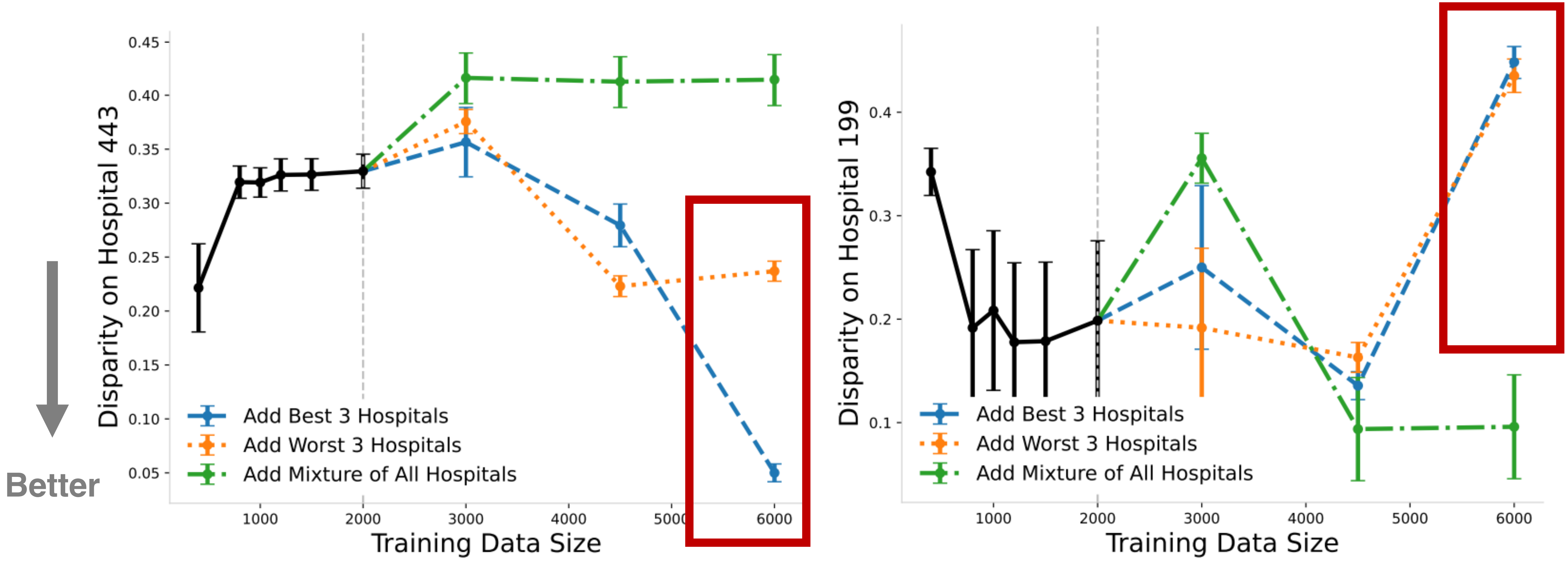
24-hour mortality prediction talk on eICU dataset

But! Racial disparity can have mixed results



24-hour mortality prediction talk on eICU dataset

But! Racial disparity can have mixed results



24-hour mortality prediction talk on eICU dataset

How can we benefit from larger datasets?

1. Data composition is a **data-oriented perspective** to complement existing algorithmic-centered work, e.g., domain adaptation

2. We analyze different data accumulation strategies for health datasets with **multiple sources**

3. We show theoretical and empirical results, including a **divergence-based method** for when to add more data

How can we benefit from larger datasets?

1. Data composition is a **data-oriented perspective** to complement existing algorithmic-centered work, e.g., domain adaptation

2. We analyze different data accumulation strategies for health datasets with **multiple sources**

3. We show theoretical and empirical results, including a **divergence-based method** for when to add more data

4. **Next steps**: fairness, multimodality, impossibility theorems, privacy

Computational **H**ealthcare for **E**quity and **IN**clusion

Problem Selection

1. Early detection for intimate partner violence (**PSB 2021**)
2. Maternal health and high-risk pregnancy (**FaccT 2024**)

Data Collection

1. Collecting and researching insurance risk scores (**Health Affairs 2022**)
2. Effect of aggregating data sources (**MLHC 2024**)

Outcome Definition

1. Rationales for treatment selection (**Work under review**)
2. Access to care and how it affects EHR labels (**In progress**)

Algorithm Development

1. Correcting for patient access to care (**AAAI 2022**)
2. Developing new LLM bias frameworks (**In progress**)

Post-Deployment Considerations

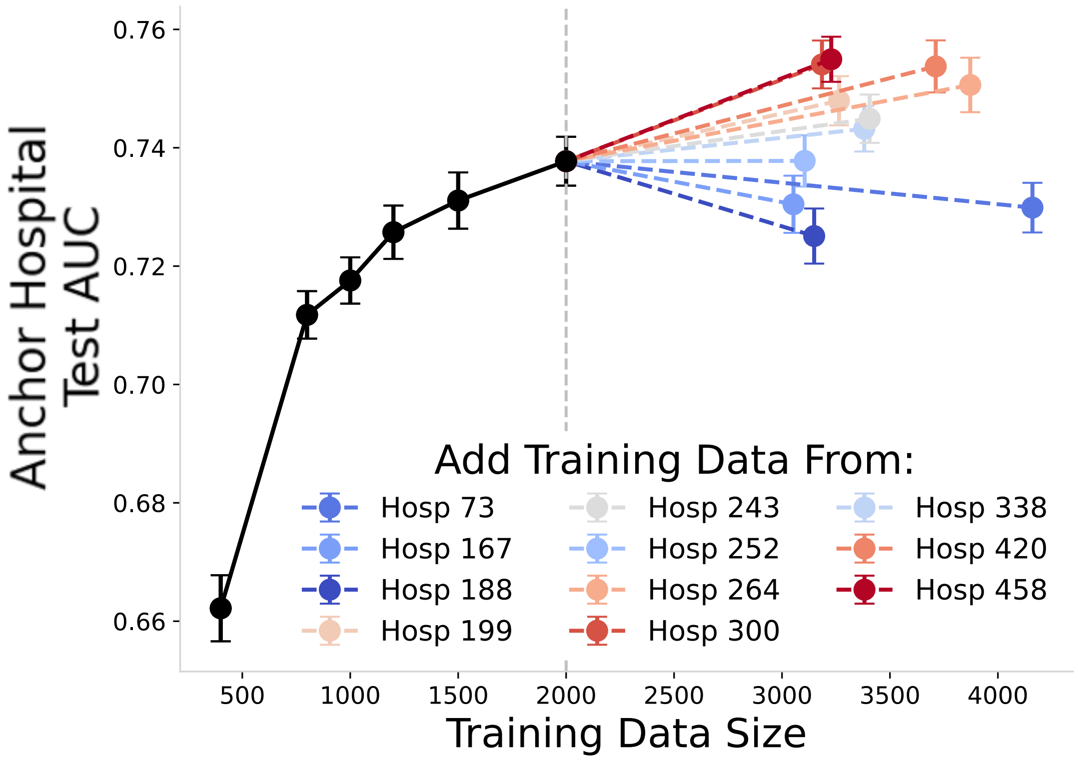
1. Bias auditing (**AMA Journal of Ethics 2019, Nature Medicine 2021**)
2. Mitigating algorithmic bias (**NeurIPS 2018**)

Chen et al, “Ethical Machine Learning for Health Care,” *Annual Reviews for Biomedical Data Science 2021.*

How can we benefit from larger datasets?

1. We need a **data-oriented perspective**, especially considering **multiple sources**

2. We show theoretical and empirical results, including a **divergence-based method** for when to add more data



Berkeley UNIVERSITY OF CALIFORNIA

@irenetrampoline

irenechen.net

UCSF