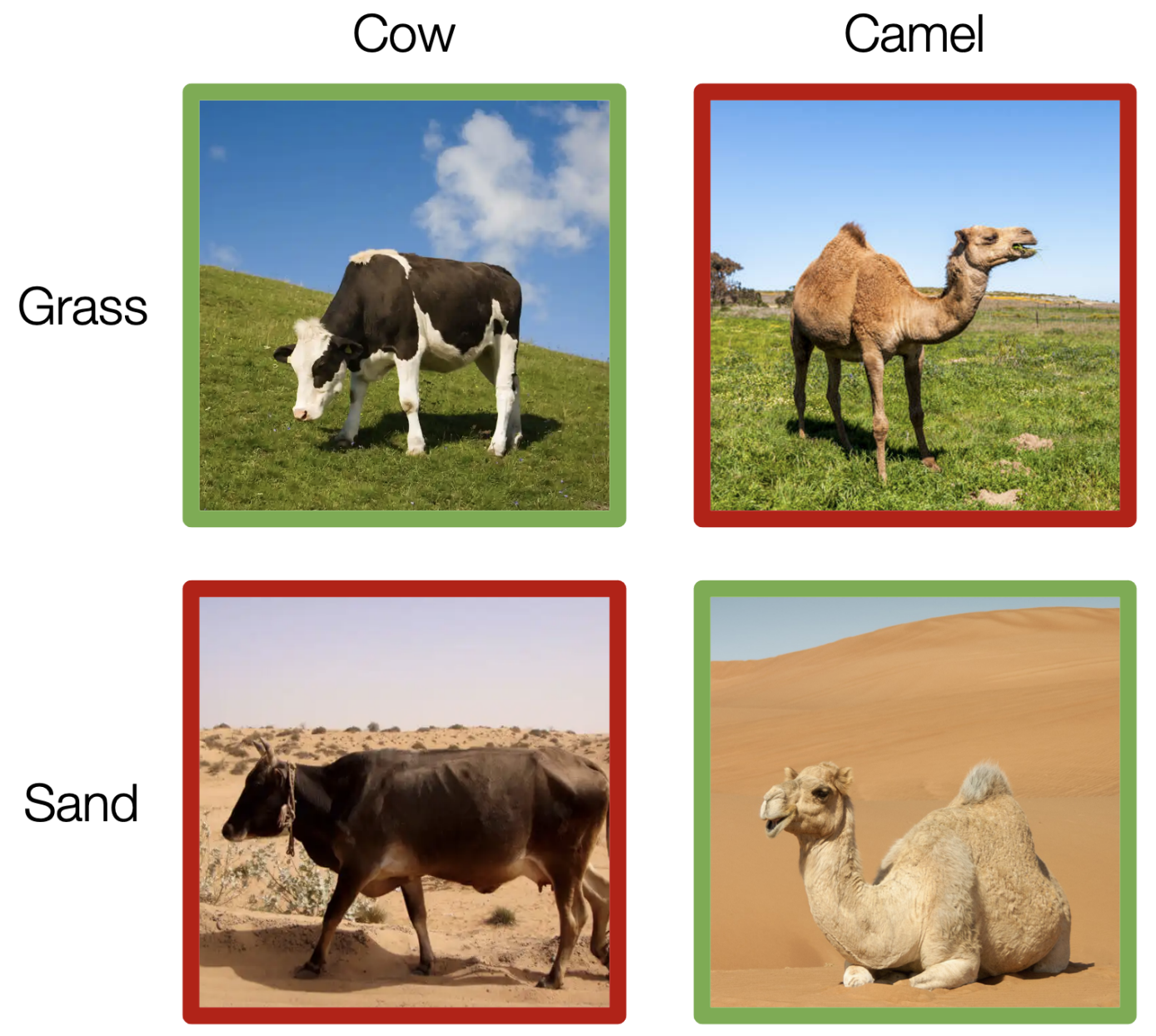
Change is Hard: A Closer Look at Subpopulation Shift

Yuzhe Yang[1,*] Haoran Zhang[1,*] Dina Katabi[1] Marzyeh Ghassemi[1]

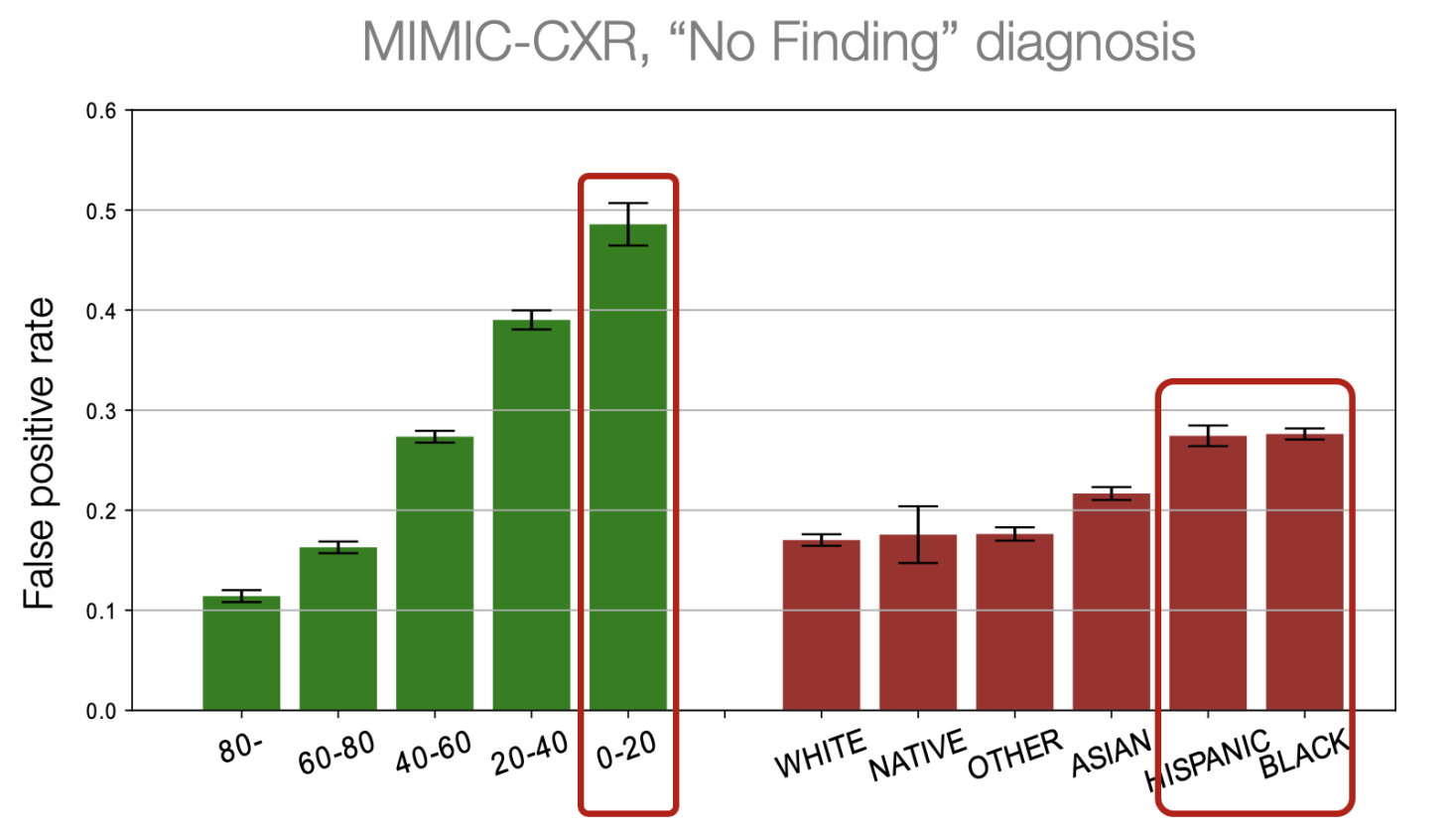
[1]MIT CSAIL *equal contribution

Background & Motivation

Subpopulation shift is ubiquitous in real-world data!

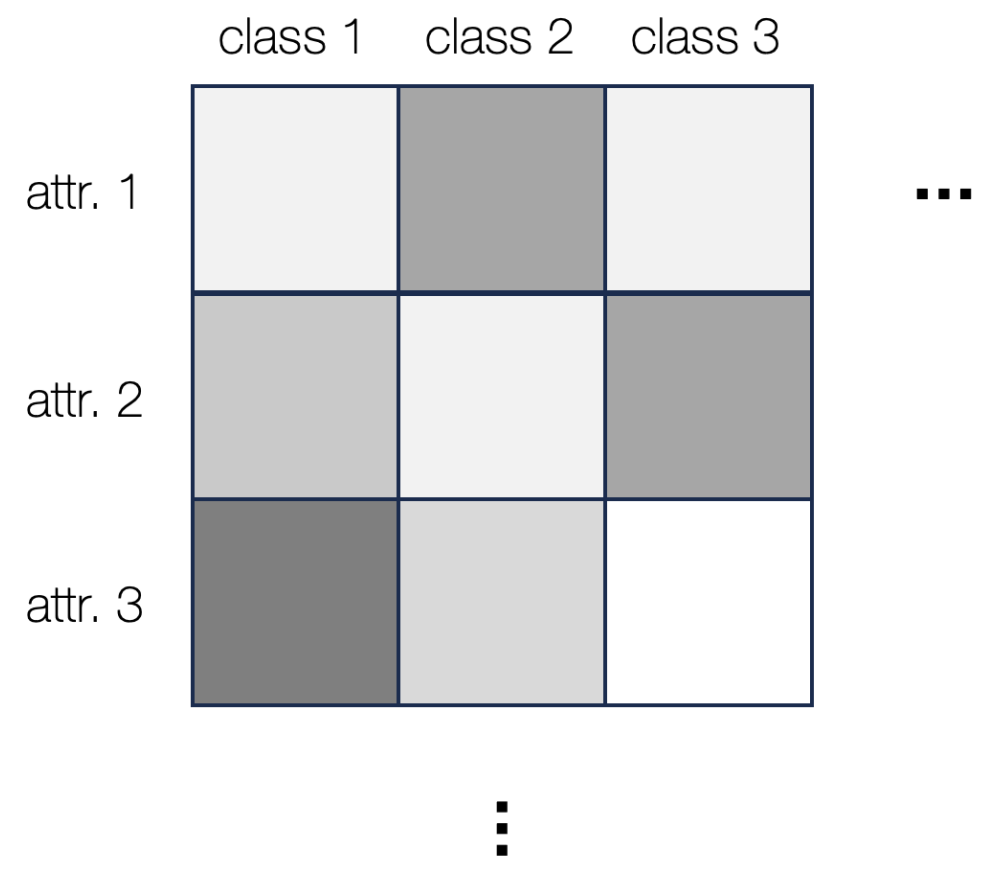


The "underdiagnosis bias" in AI algorithms for health: Chest X-rays



❶ How can we characterize different types of subpopulation shift?
❷ How well do algorithms generalize across diverse shifts at scale?

A Unified Framework of Subpopulation Shift



$$\mathbb{P}(y|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|y)}{\mathbb{P}(\mathbf{x})} \cdot \mathbb{P}(y) \qquad \triangleright\ \mathbf{x} \xleftrightarrow{\text{fully generated}} (\mathbf{x}_{\text{core}}, \mathbf{a})$$

$$= \frac{\mathbb{P}(\mathbf{x}_{\text{core}}, \mathbf{a}|y)}{\mathbb{P}(\mathbf{x}_{\text{core}}, \mathbf{a})} \cdot \mathbb{P}(y)$$

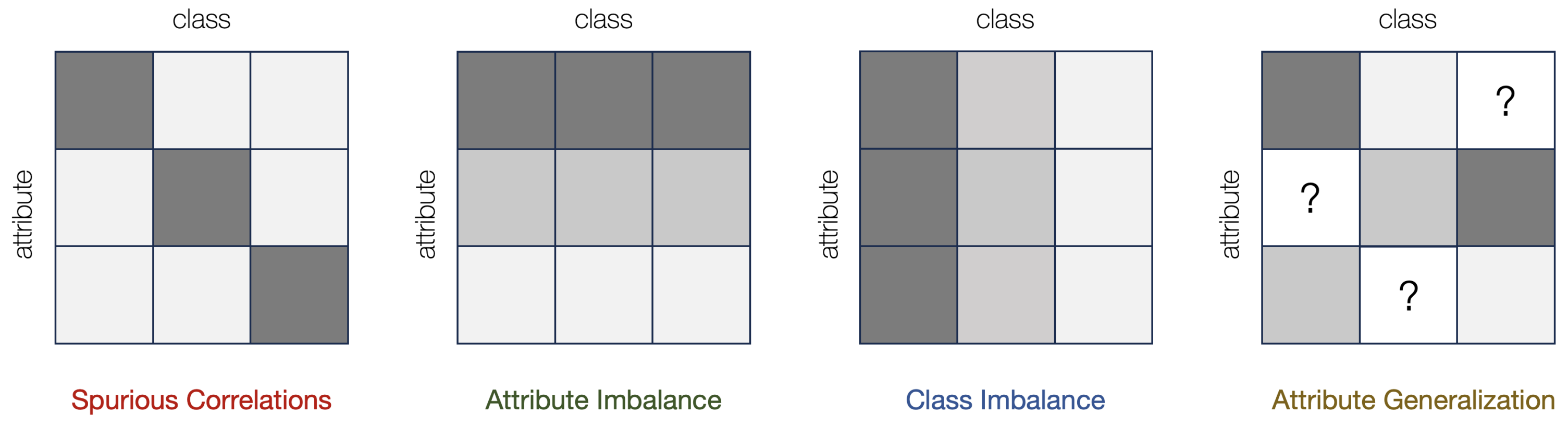
$$= \boxed{\frac{\mathbb{P}(\mathbf{x}_{\text{core}}|y)}{\mathbb{P}(\mathbf{x}_{\text{core}})}} \cdot \boxed{\frac{\mathbb{P}(\mathbf{a}|y, \mathbf{x}_{\text{core}})}{\mathbb{P}(\mathbf{a}|\mathbf{x}_{\text{core}})}} \cdot \boxed{\mathbb{P}(y)}$$

PMI Attribute bias Class bias

Interpretation:

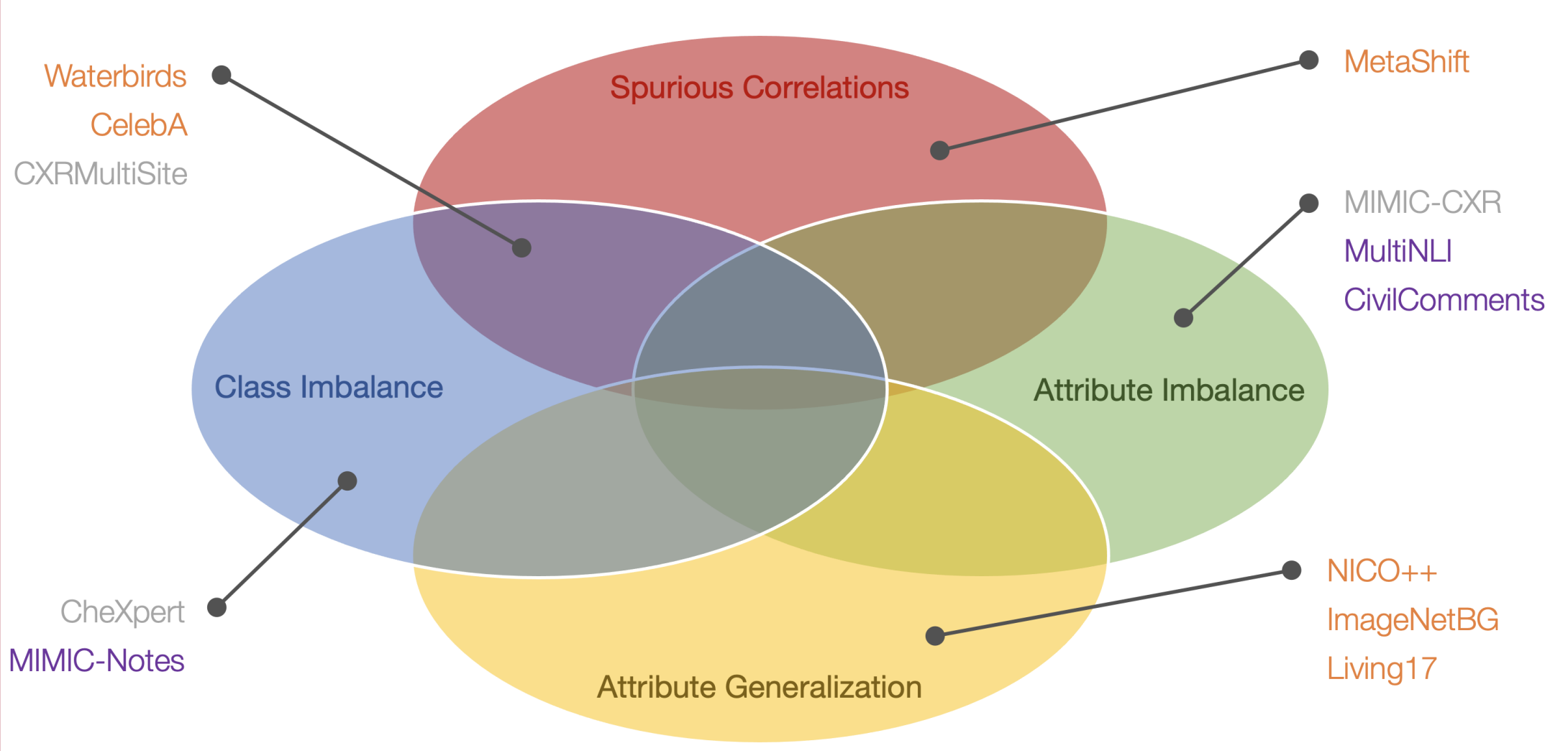
❶ *1st term* → Robust indicator, *invariant* between training & testing
❷ *2nd term* → Biases in **Attribute** distribution
❸ *3rd term* → Biases in **Label** distribution

Characterizing Basic Types of Subpopulation Shift



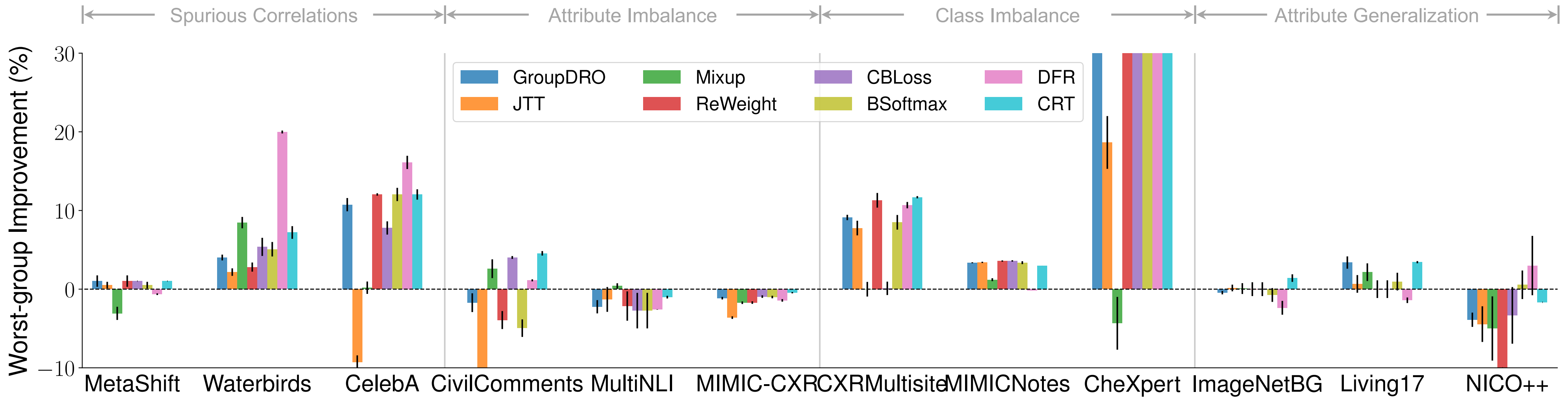
Note: Real datasets often consist of *multiple* types of shift instead of one. The four cases constitute the *basic* shift units, and are important elements to explain complex subgroup shifts in real data.

SubpopBench: Benchmarking Subpopulation Shift

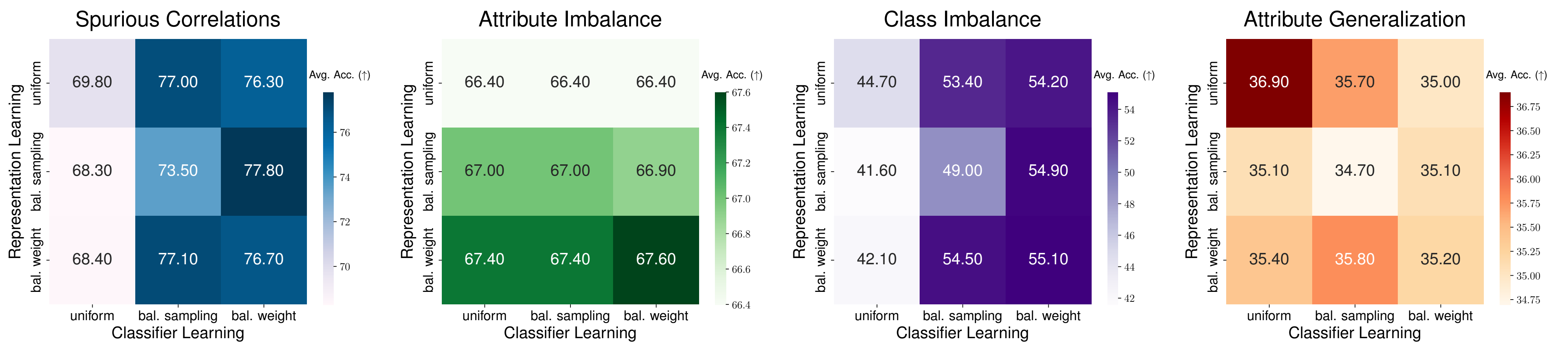


- **12** real-world datasets
  - *Vision, NLP, Clinical texts, Chest X-rays*
- **20** state-of-the-art algorithms
  - *Subgroup robustness, Data augmentation, Domain-invariant representation learning, Imbalanced learning*
- **16** model architectures & pretraining methods
- **7** evaluation metrics
- **3** model selection criteria

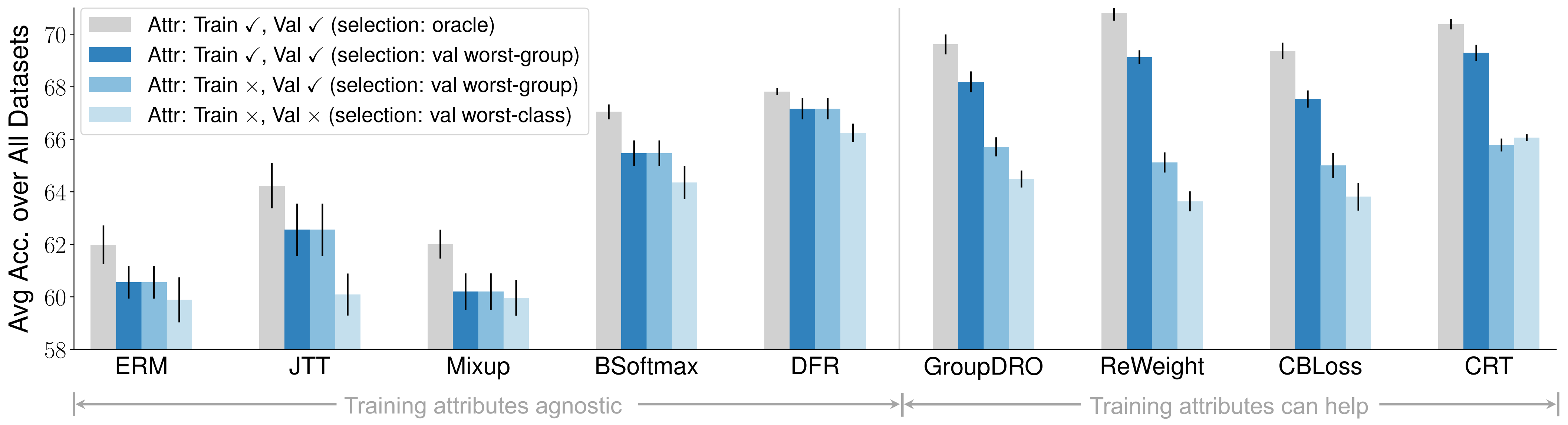
Observation #1: SOTA Algorithms Only Improve Certain Types of Shift



Observation #2: The Roles of Representation and Classifier Differ under Shifts



Observation #3: Model Selection & Attribute Availability Matter!

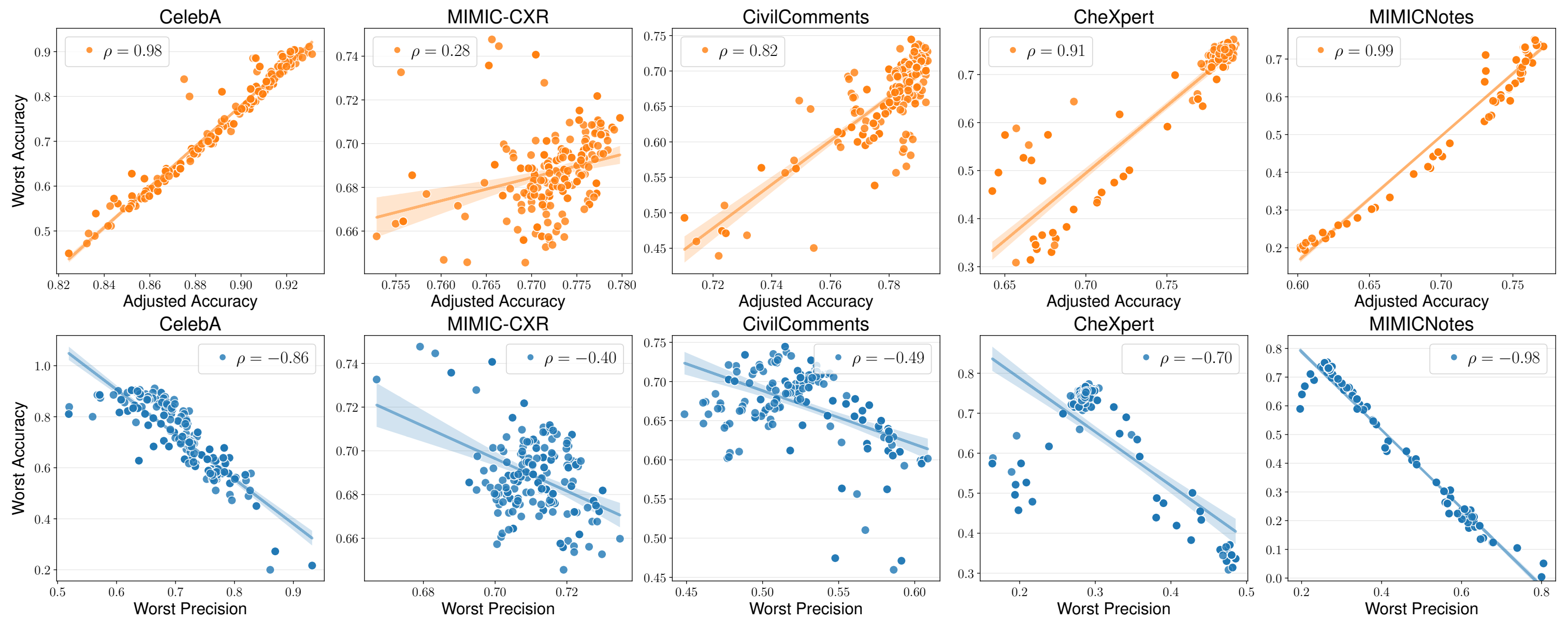


Observation #4: Model Selection w/o Group Info.

Worst-class accuracy is surprisingly effective even **w/o attribute!**

Selection Strategy	CelebA	CheXpert	CivilComments	MIMIC-CXR	MIMICNotes	MetaShift	**Avg**
Max Worst-Class Accuracy	-5.0 ±6.3	**-0.4** ±0.8	**-3.2** ±5.2	**-0.9** ±1.0	**-0.1** ±0.5	**-1.5** ±3.0	**-1.8**
Max Balanced Accuracy	**-4.4** ±5.4	-1.3 ±2.5	-3.5 ±5.8	-2.9 ±4.9	-2.3 ±6.2	-1.7 ±3.0	-2.7
Min Class Accuracy Diff	-6.1 ±9.1	-1.9 ±5.3	-4.1 ±8.0	-1.9 ±5.0	-0.3 ±1.2	-2.2 ±4.6	-2.7
Max Worst-Class F1	-13.4 ±10.4	-5.4 ±6.7	**-3.2** ±3.8	-2.5 ±2.2	-4.4 ±8.7	-1.8 ±3.3	-5.1
Max Overall AUROC	-12.2 ±10.3	-10.4 ±13.0	-8.2 ±9.0	-6.6 ±9.9	-10.0 ±16.5	-3.2 ±7.0	-8.4
Max Overall Accuracy	-18.6 ±12.0	-30.9 ±24.9	-13.7 ±9.5	-5.1 ±6.3	-19.9 ±26.0	-1.9 ±3.3	-15.0

Observation #5: Metrics Beyond Worst-Group Accuracy



Does improving WGA always improve other meaningful metrics?

❶ **Accuracy on the line:** Adjusted accuracy is *positively* correlated with WGA.
❷ **Accuracy on the inverse line:** Worst-class precision is *negatively* correlated with WGA.

Implication: **Inherent tradeoffs** between testing metrics; The need for a **broader set** of evaluation metrics.

Take Home Messages & More Information

SUBPOPBENCH

❶ **Better algorithms needed for certain shifts!**
❷ **Think about shifts in the design of ML pipeline!**
❸ **Access to attributes still plays a significant role!**
❹ **More comprehensive evaluation across broader metrics!**

Project Page Paper Code Video