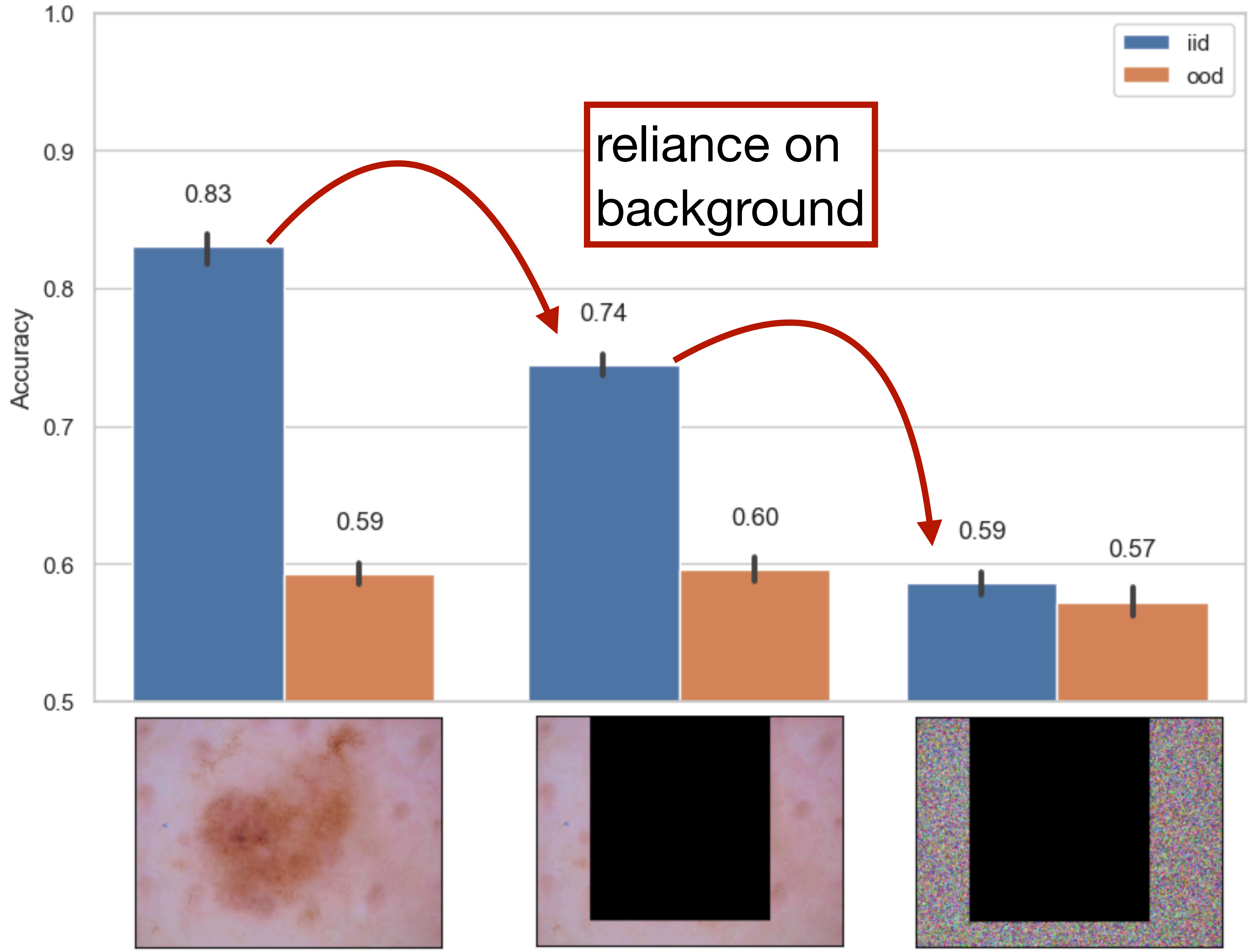
b
UNIVERSITÄT
BERN

Guiding Models to Mitigate Bias in Skin Lesion Analysis

Alceu Bissoto, 10.10.2024

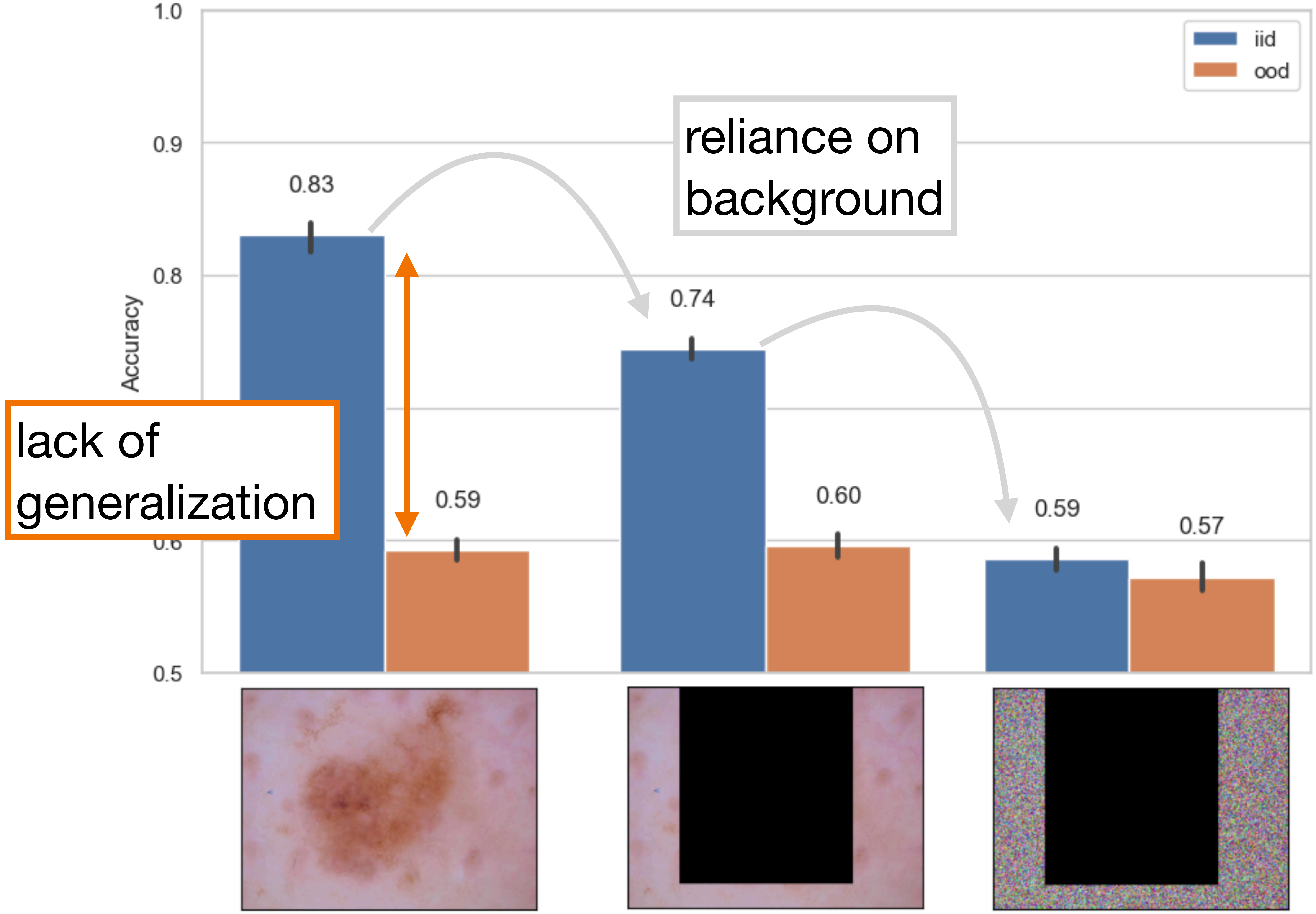
Motivation

Training on HAM, OOD: ISIC 2018



Motivation

Training on HAM, OOD: ISIC 2018



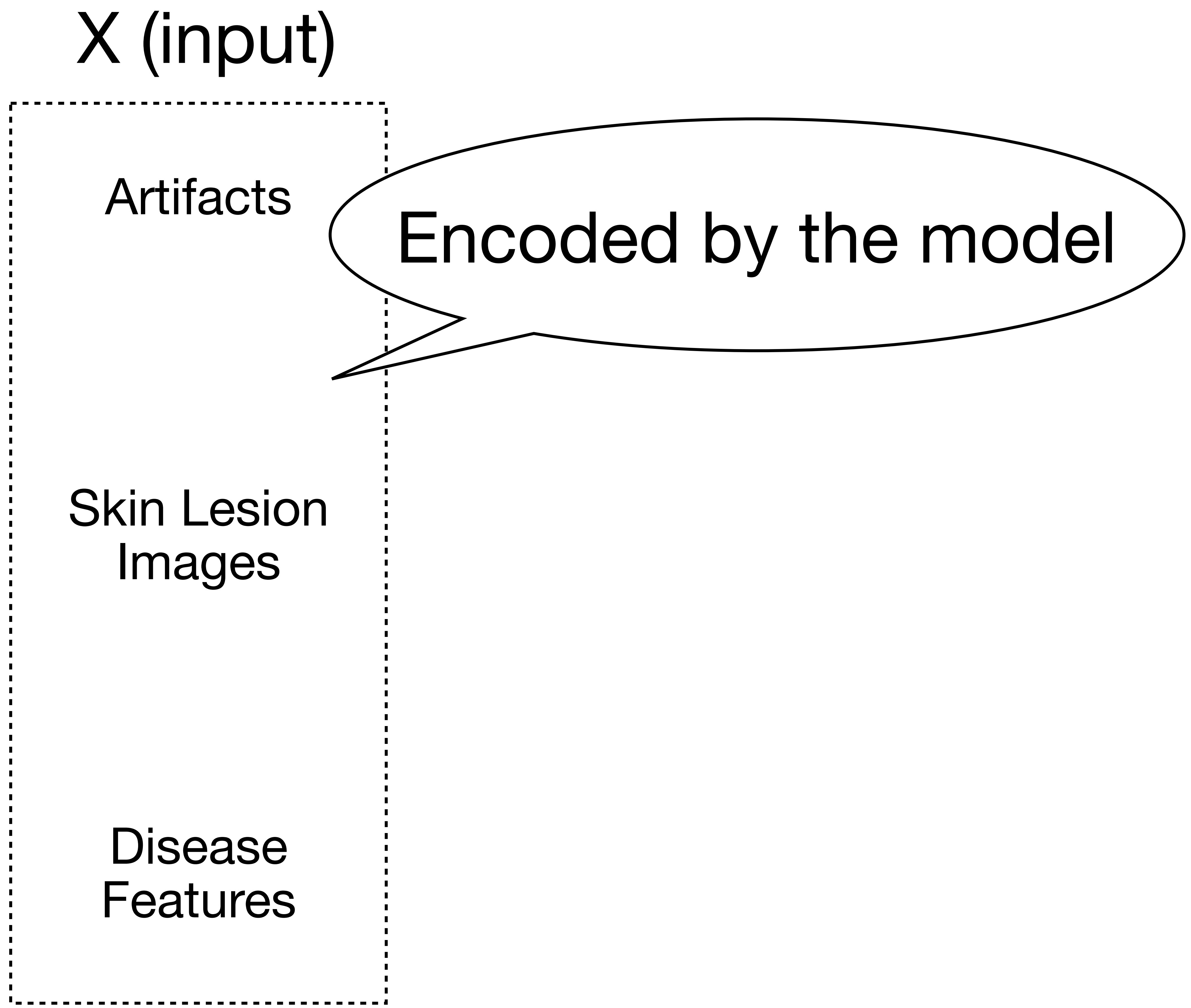
Agenda

General procedure for bias investigations

- Problem Characterization
- Debiasing
- Evaluation

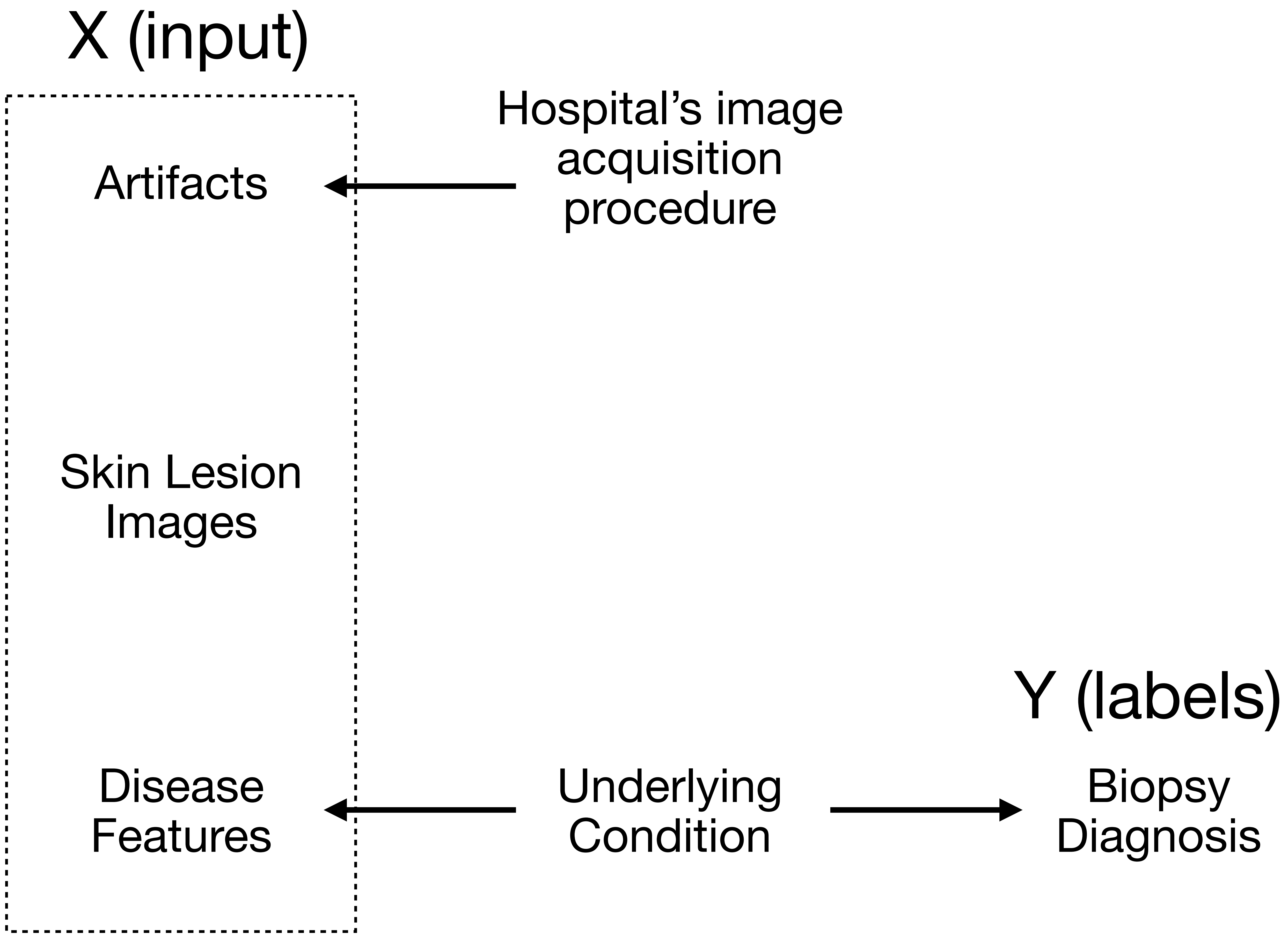
BiasPrune

Define the problem through a causal graph



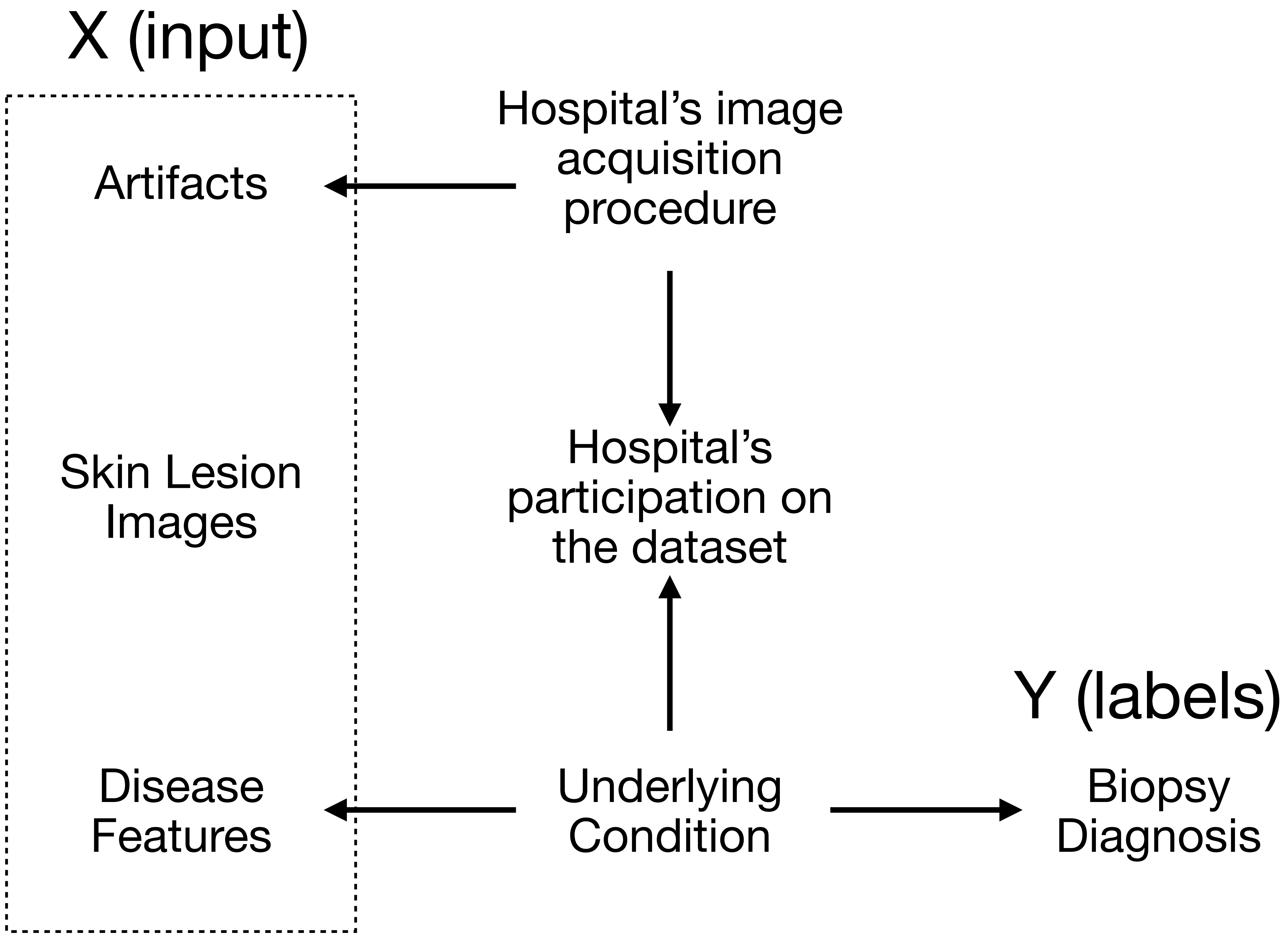
Jones, Charles, et al. "A causal perspective on dataset bias in machine learning for medical imaging." *Nature Machine Intelligence* 6.2 (2024)

Define the problem through a causal graph



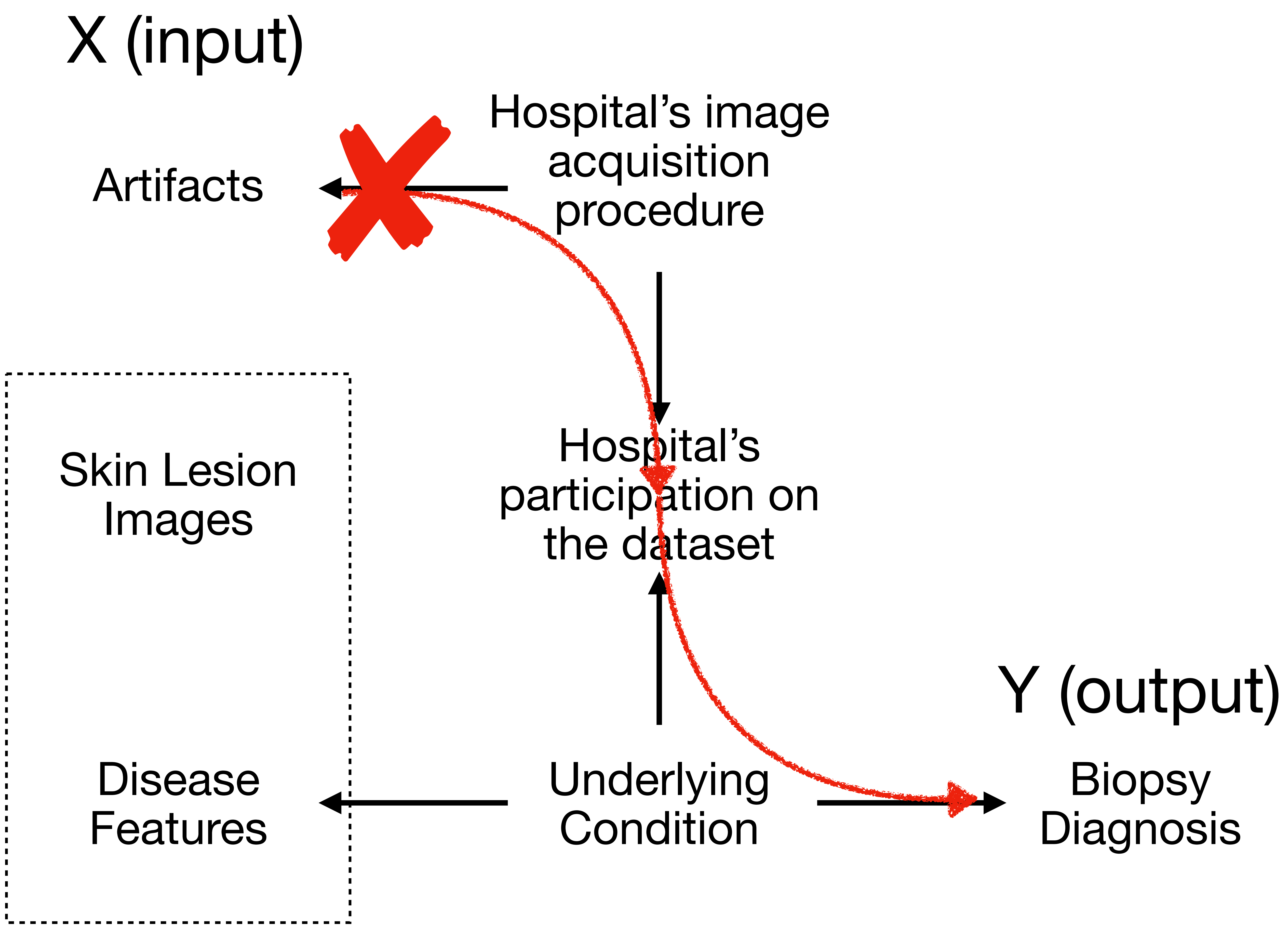
Jones, Charles, et al. "A causal perspective on dataset bias in machine learning for medical imaging." *Nature Machine Intelligence* 6.2 (2024)

Define the problem through a causal graph



Jones, Charles, et al. "A causal perspective on dataset bias in machine learning for medical imaging." *Nature Machine Intelligence* 6.2 (2024)

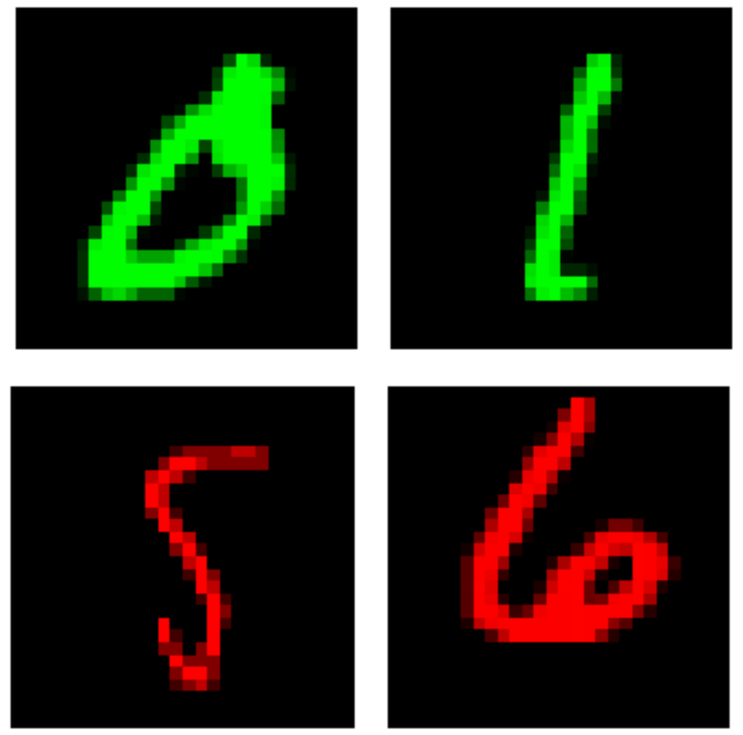
How to avoid learning from artifacts



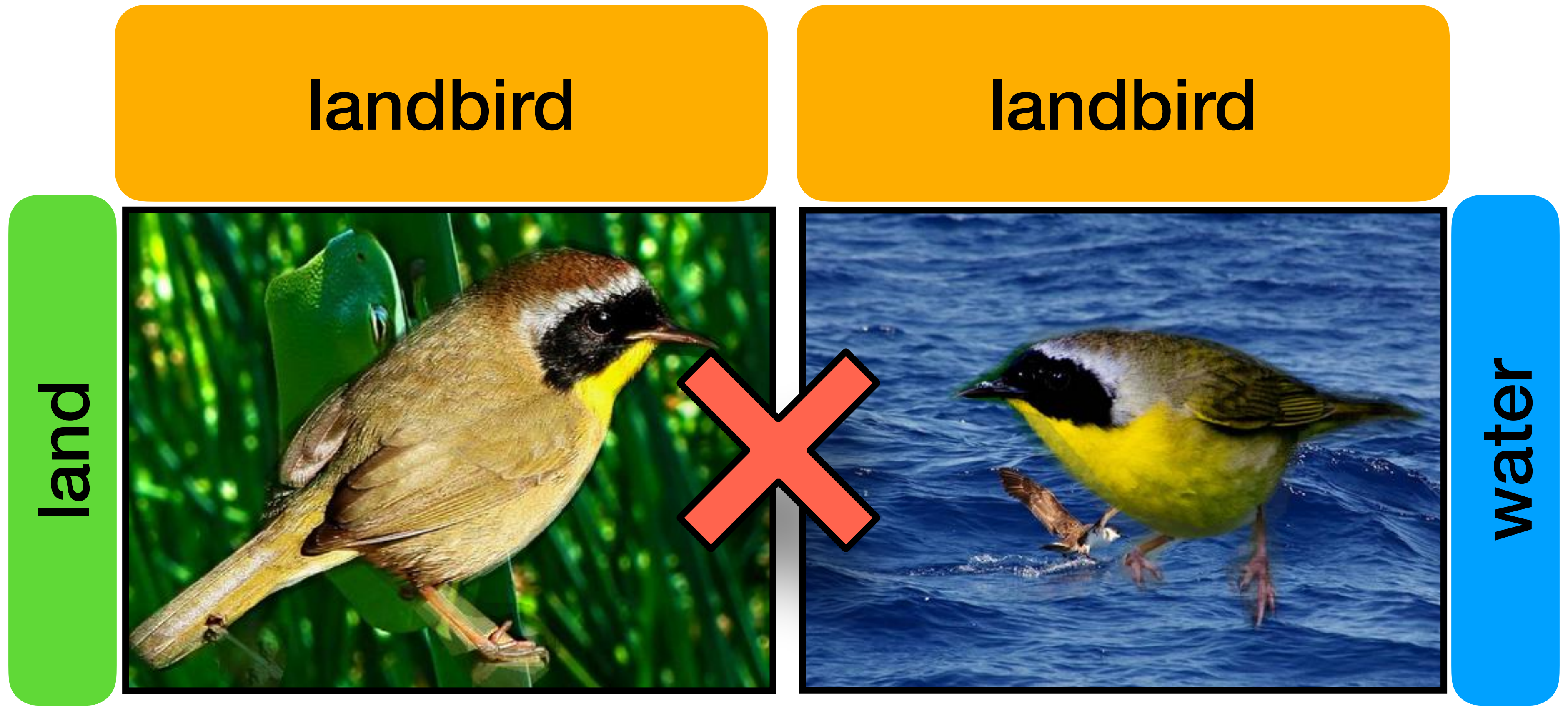
- Domain generalization
- Invariant representation learning
- Disentanglement

Jones, Charles, et al. "A causal perspective on dataset bias in machine learning for medical imaging." *Nature Machine Intelligence* 6.2 (2024)

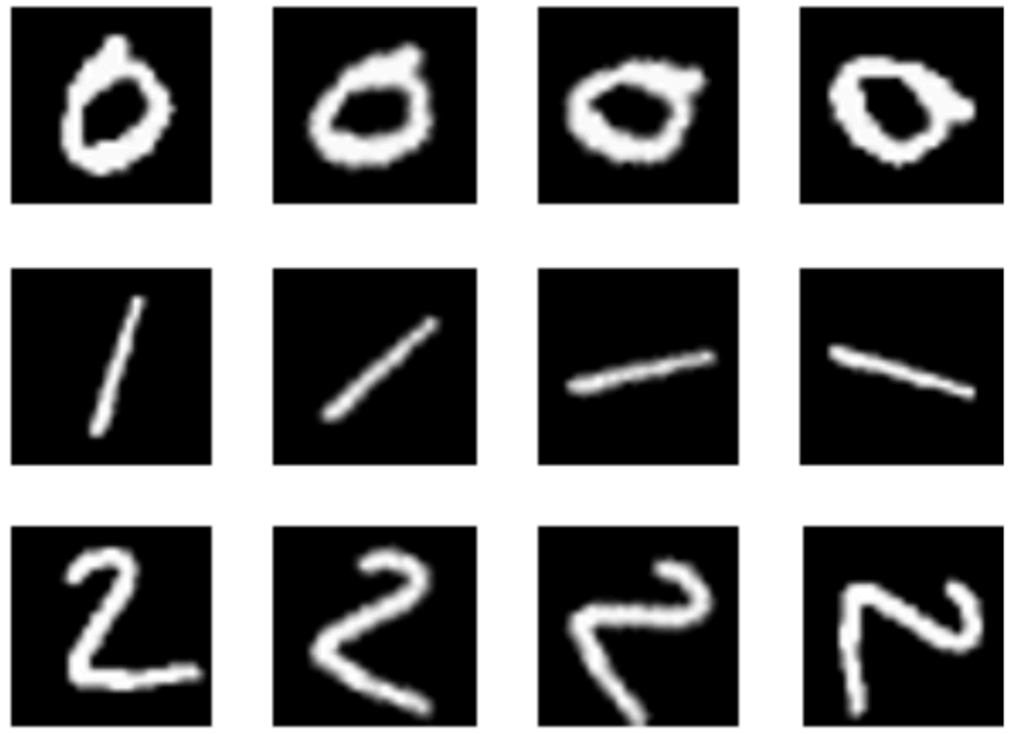
Domain generalization data is too simple



CMNIST

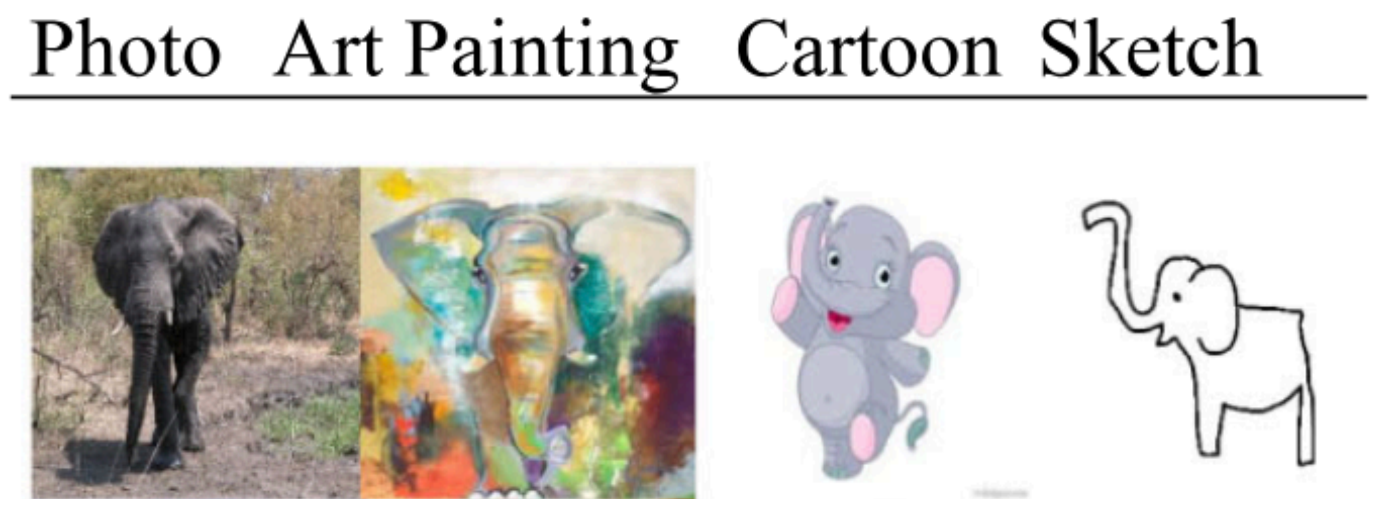


Waterbirds

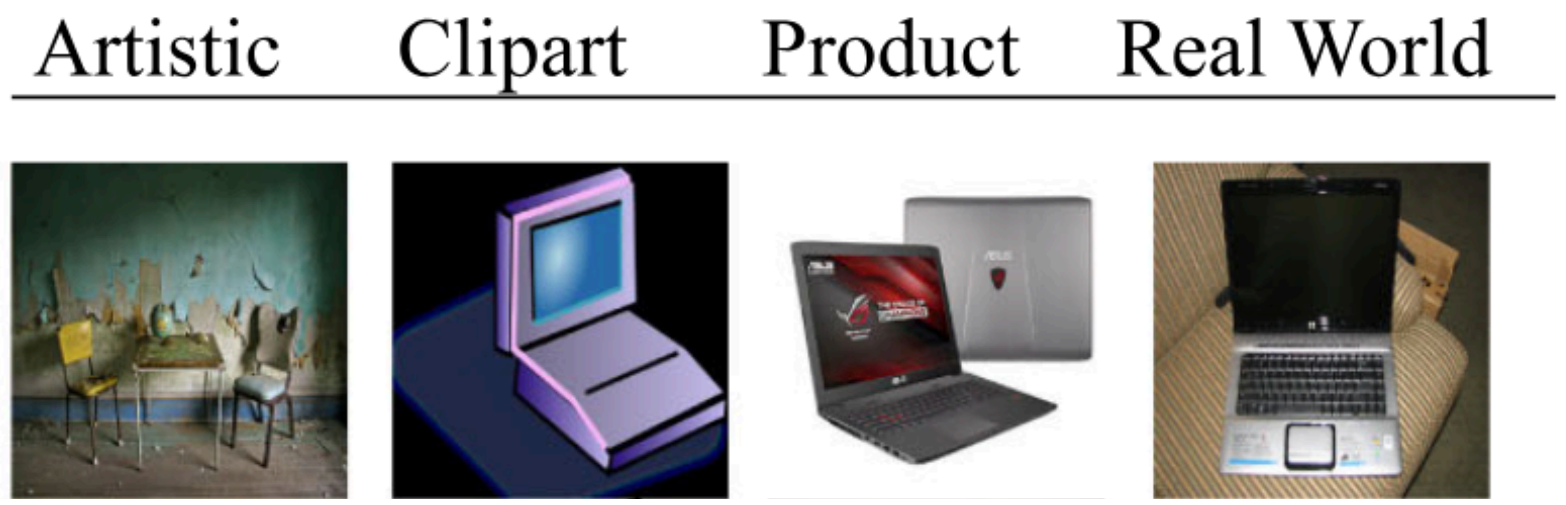


RMNIST

VLCS



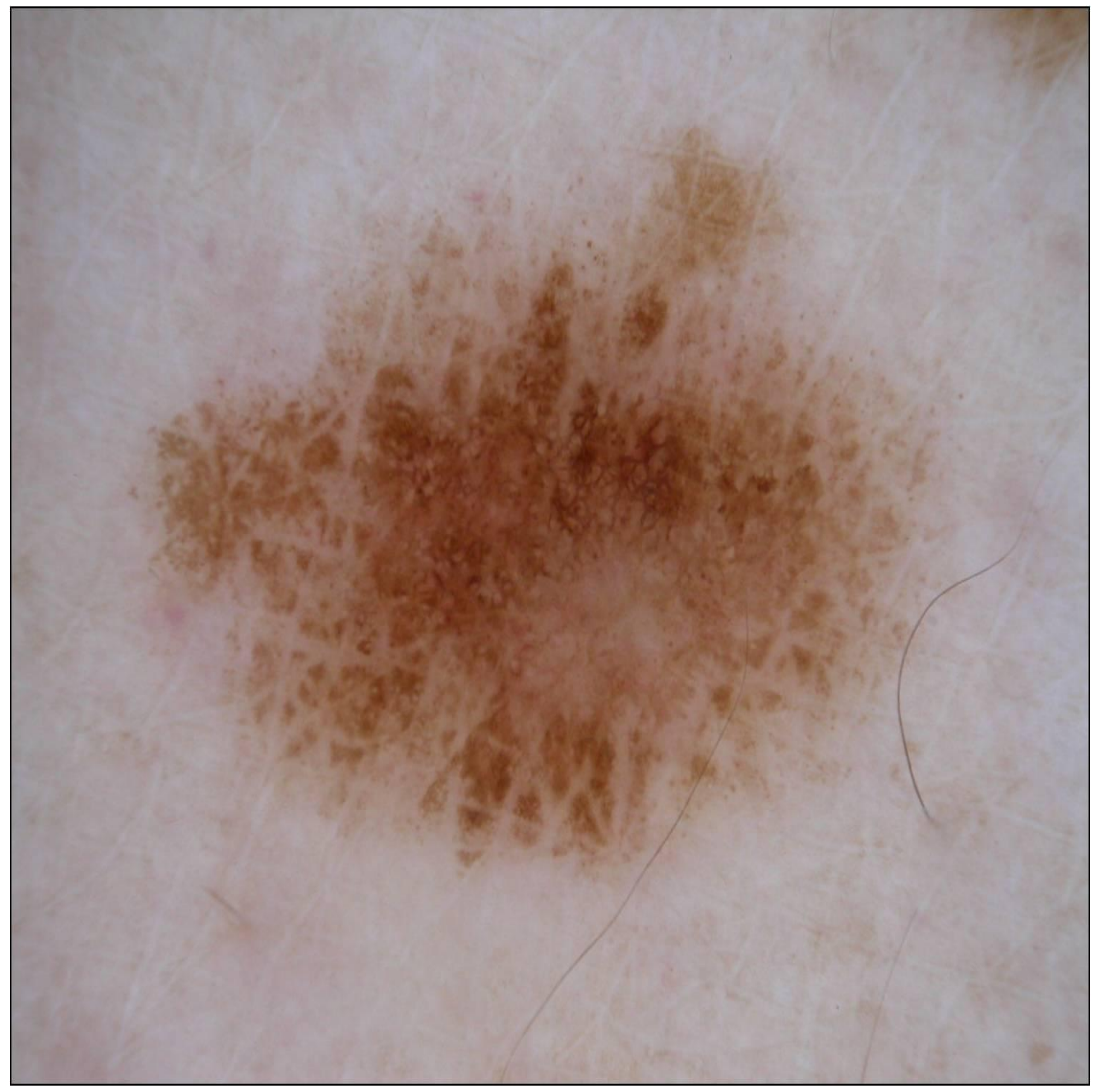
PACS



OfficeHome

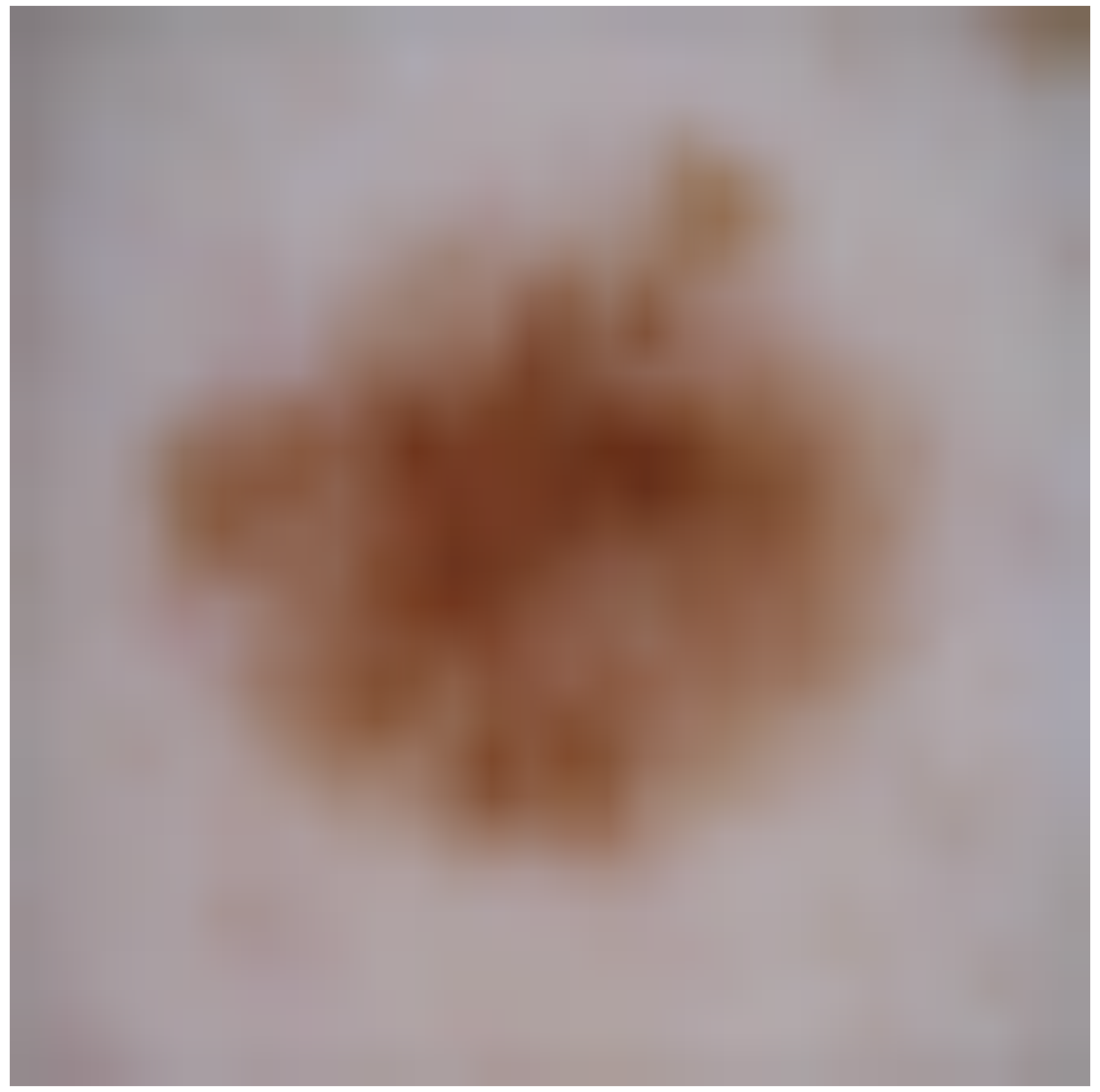
Difficulty of learning complex relevant features

(224 x 224 x 3)



≠

(28 x 28 x 3)



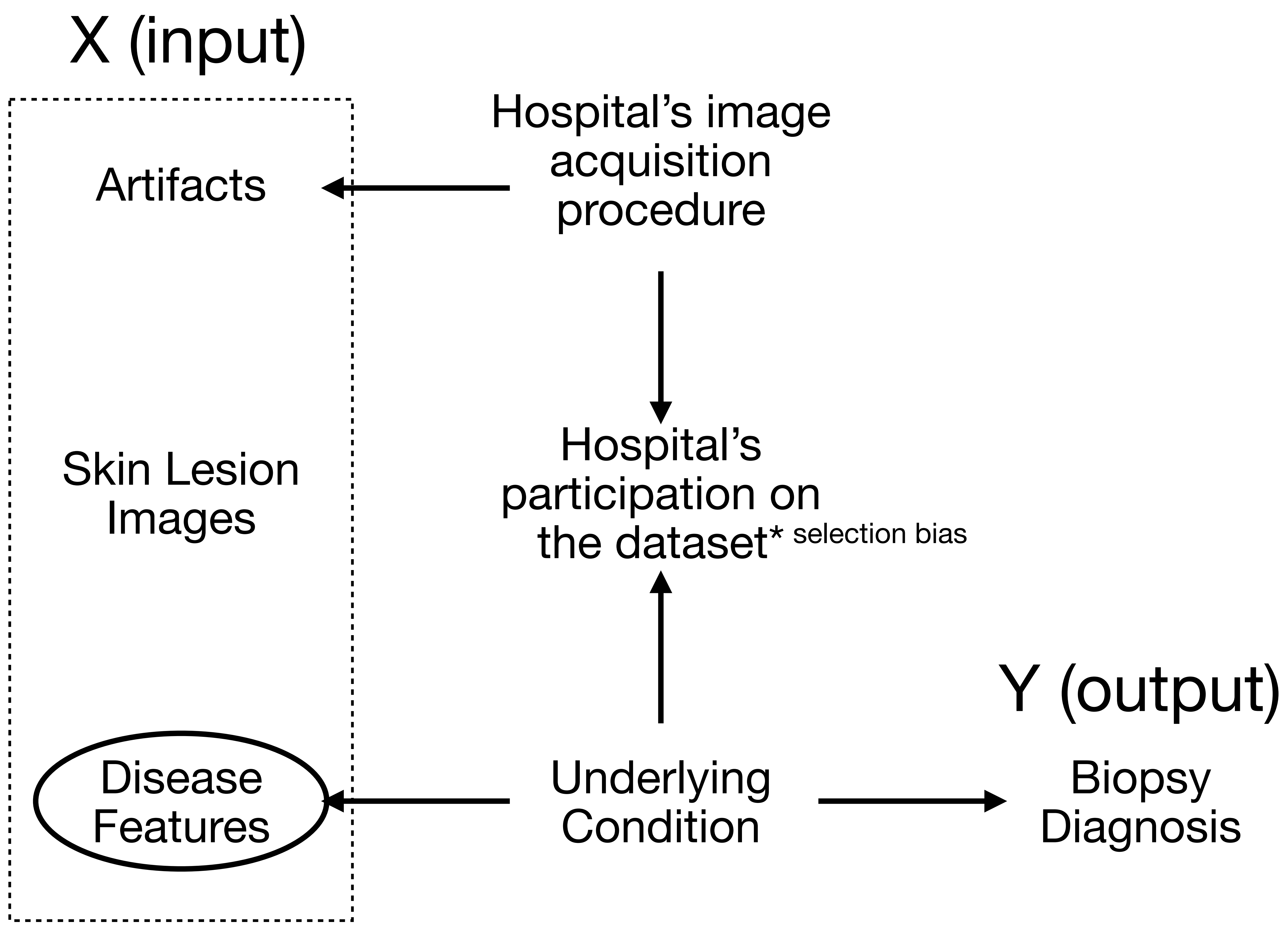
0.912 AUC
0.731 ACC

=

0.913 AUC
0.735 ACC

Abhishek, K., Jain, A., & Hamarneh, G. (2024). Investigating the Quality of DermaMNIST and Fitzpatrick17k Dermatological Image Datasets.

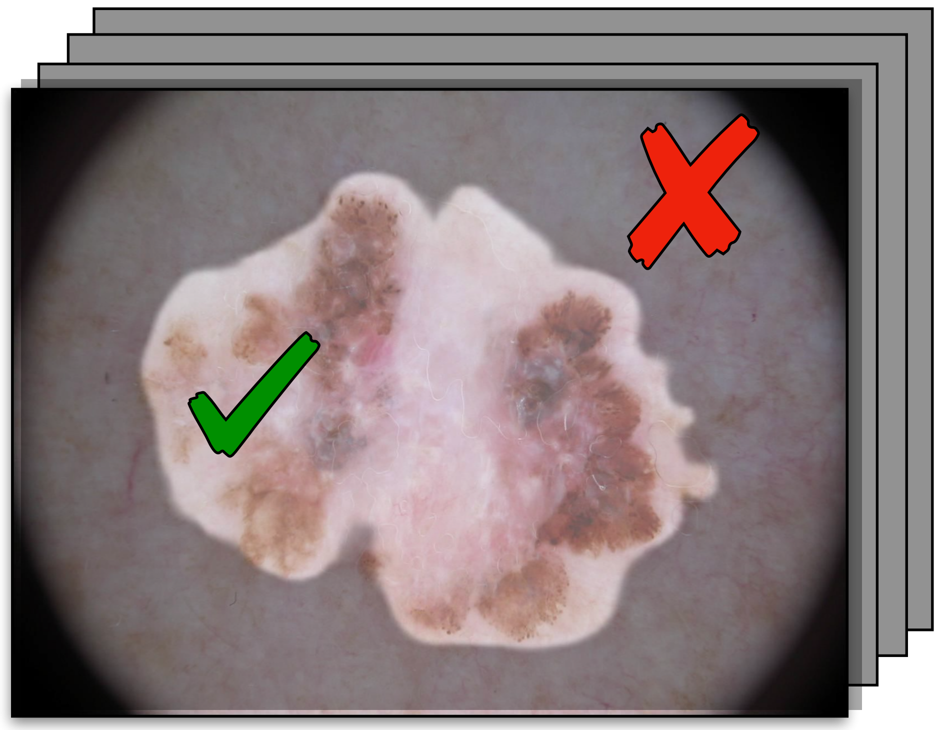
Causal Representation of Artifact Bias



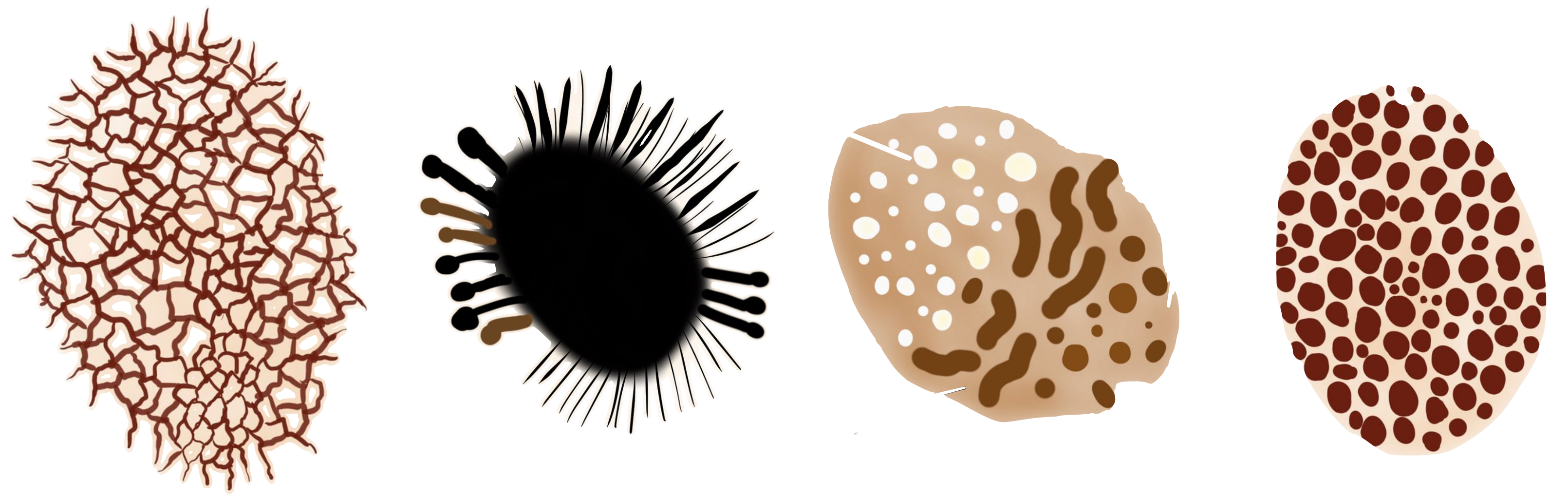
- Domain generalization
- Invariant representation learning
- Disentanglement
- *Characterization of disease features to guide models*

Characterization of disease features

Segmentation masks



Clinical attributes



Pigmented network

Streaks

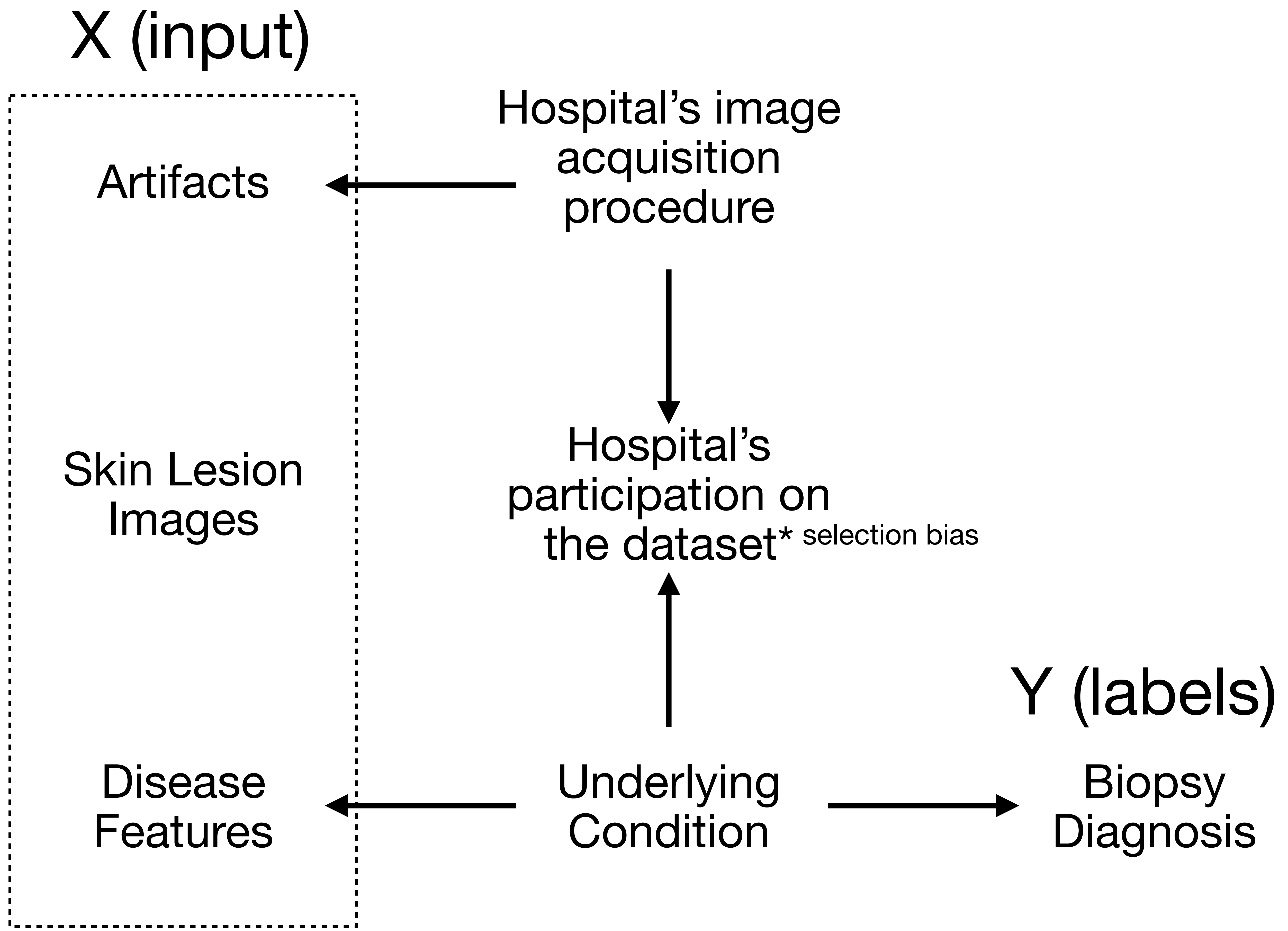
Milia-like cysts

Globules

SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis

Roxana Daneshjou[1*] Mert Yuksekgonul[2*]
Zhuo Ran Cai[1] Roberto Novoa[1] James Zou[3]
[1] Department of Dermatology, Stanford University
[2] Department of Computer Science, Stanford University

How to measure bias reliance?

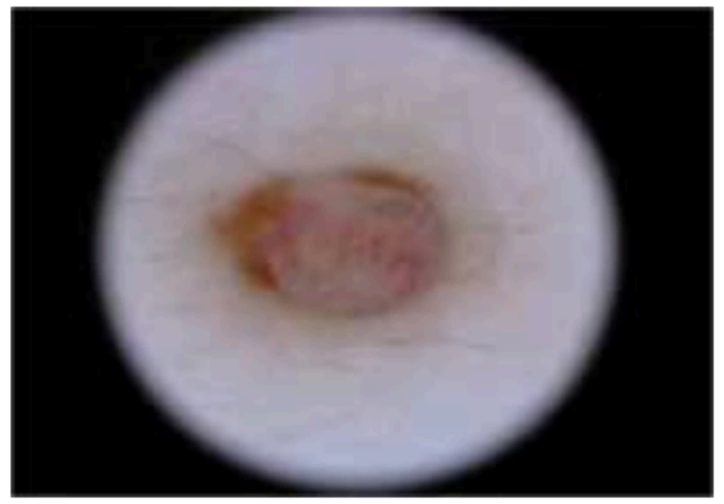


- Subgroup performance evaluation
- Out-of-distribution evaluation
- Bias decodability
- Explainable AI

Jones, Charles, et al. "A causal perspective on dataset bias in machine learning for medical imaging." *Nature Machine Intelligence* 6.2 (2024)

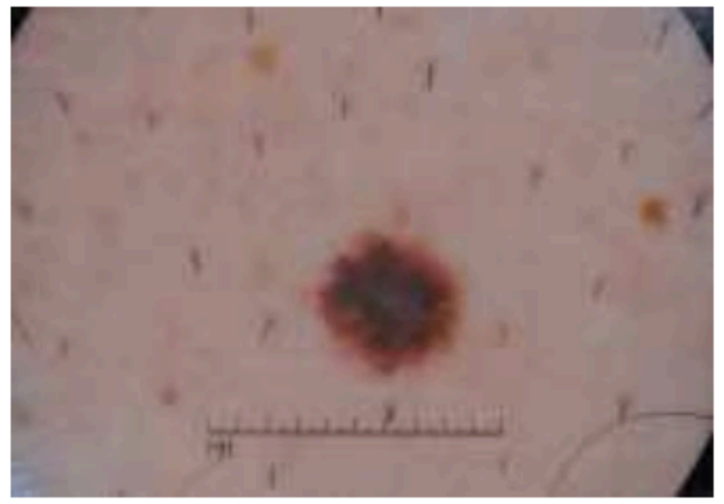
Make use of metadata annotations (and create your own!)

Artifacts

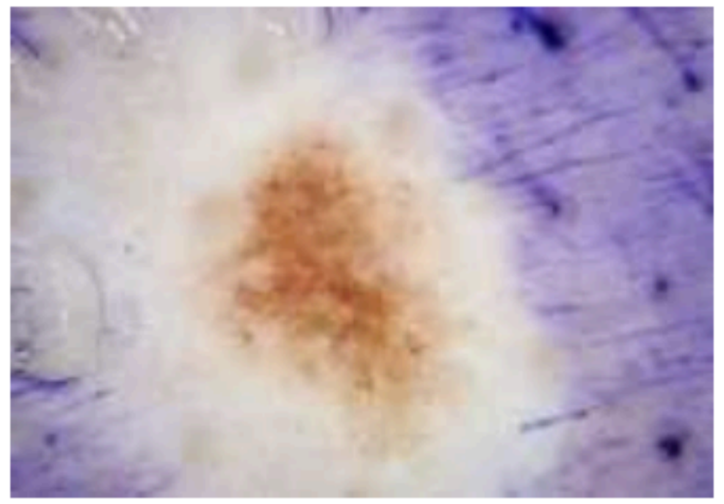


Dark Corners

Hair

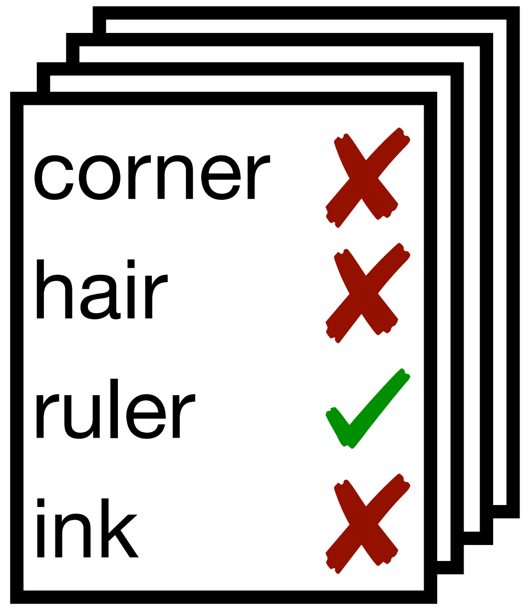


Ruler

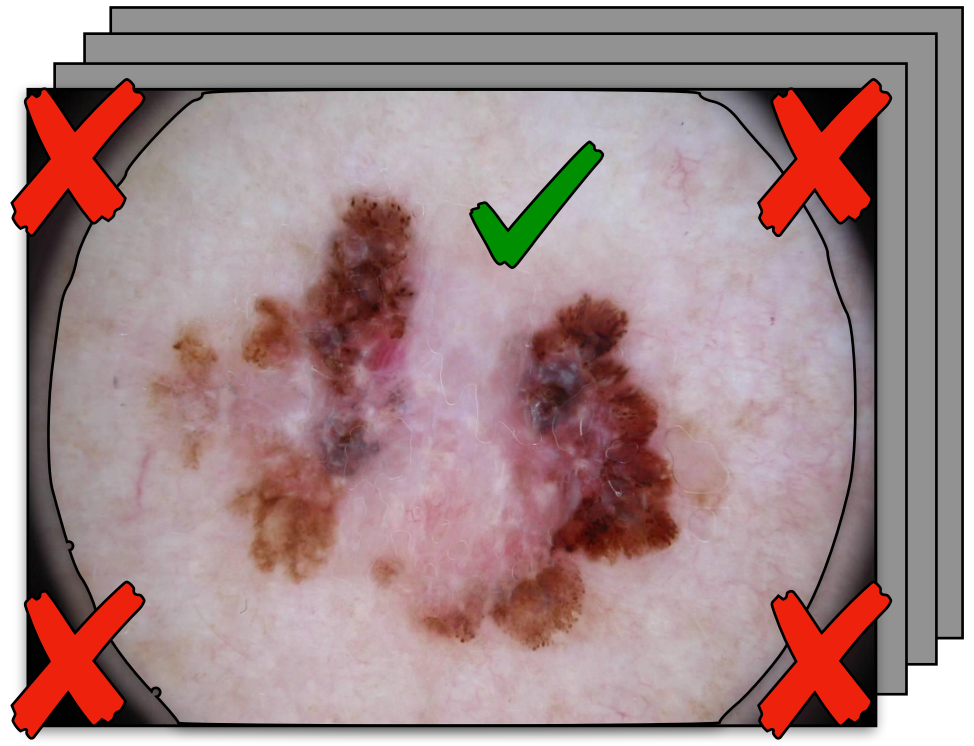


Ink markings

Artifact presence

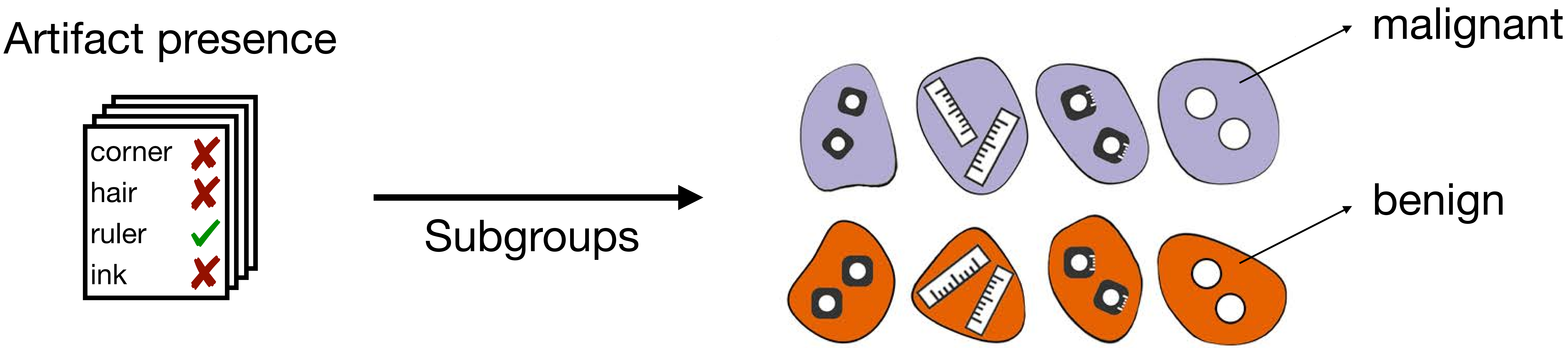


Artifact location



Bissoto, Alceu, et al. "Test-Time Selection for Robust Skin Lesion Analysis.", ISIC @ MICCAI 2023

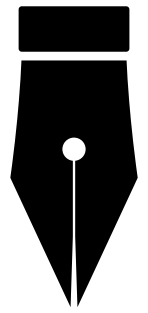
Make use of metadata annotations (and create your own!)

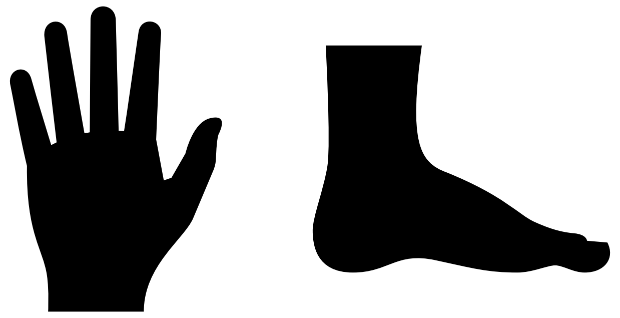


Bissoto, Alceu, et al. "Test-Time Selection for Robust Skin Lesion Analysis.", ISIC @ MICCAI 2023

Look for perf. disparities across subgroups

AVG Recall Melanoma	Melanomas with Pen Markings	Melanomas in Palms and Soles
0.7		





Combalia et al., "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

Look for perf. disparities across subgroups

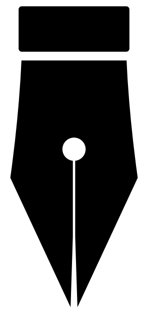
AVG Recall Melanoma

0.7

Melanomas with Pen Markings

hard!

0.53

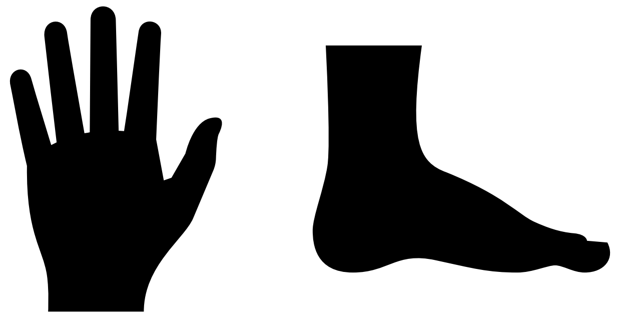


(on training) pen markings are more common in AK/BCC

Melanomas in Palms and Soles

easy?

0.80



(on training) lesions in palms or soles are more frequent to be melanoma

Combalia et al., "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

Look for perf. disparities across subgroups

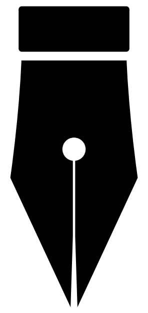
AVG Recall
Melanoma

0.7

Melanomas with
Pen Markings

hard!

0.53



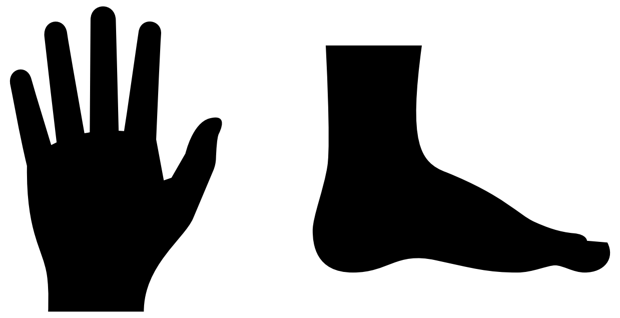
(on training) pen markings are more common in AK/BCC

Melanomas in
Palms and Soles

inflated performances

~~easy?~~

0.80



(on training) lesions in palms or soles are more frequent to be melanoma

Combalia et al., "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

Look for perf. disparities across subgroups

inflated performances

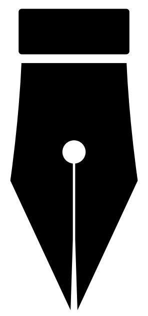
AVG Recall Melanoma

0.7

Melanomas with Pen Markings

hard!

↑ 0.53

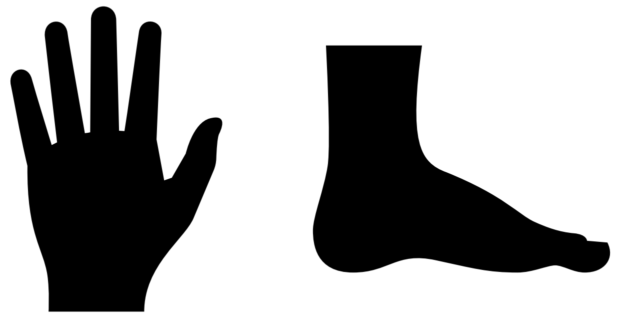


(on training) pen markings are more common in AK/BCC

Melanomas in Palms and Soles

~~easy?~~

↓ 0.80

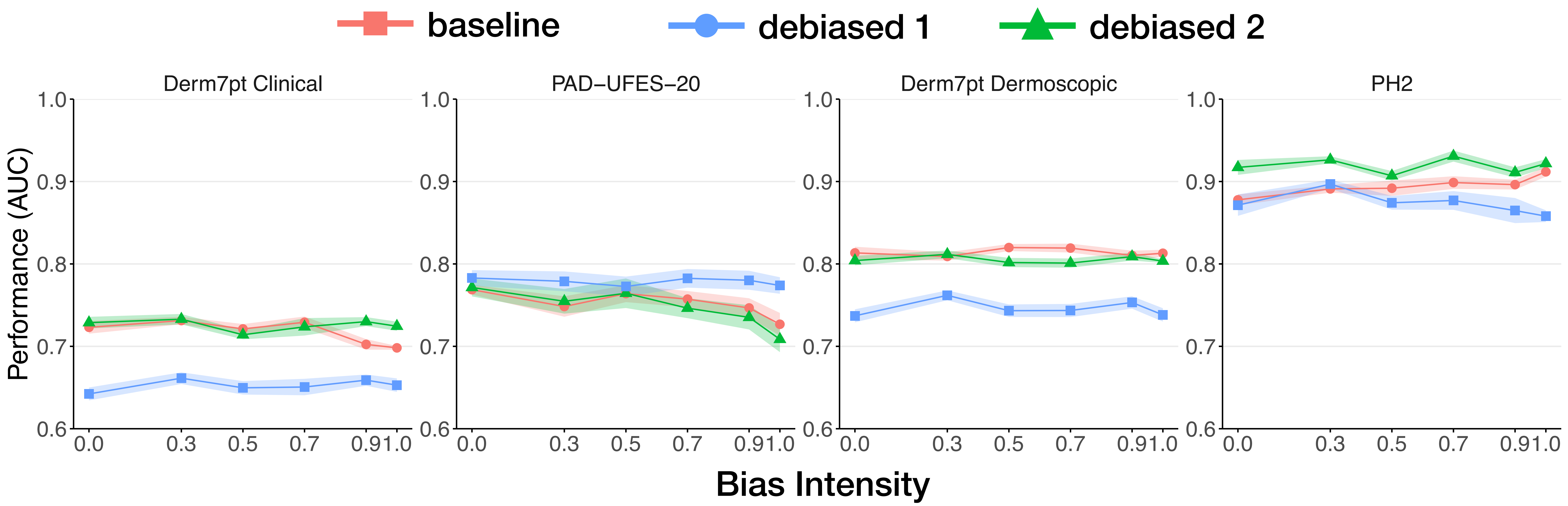


(on training) lesions in palms or soles are more frequent to be melanoma

Combalia et al., "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

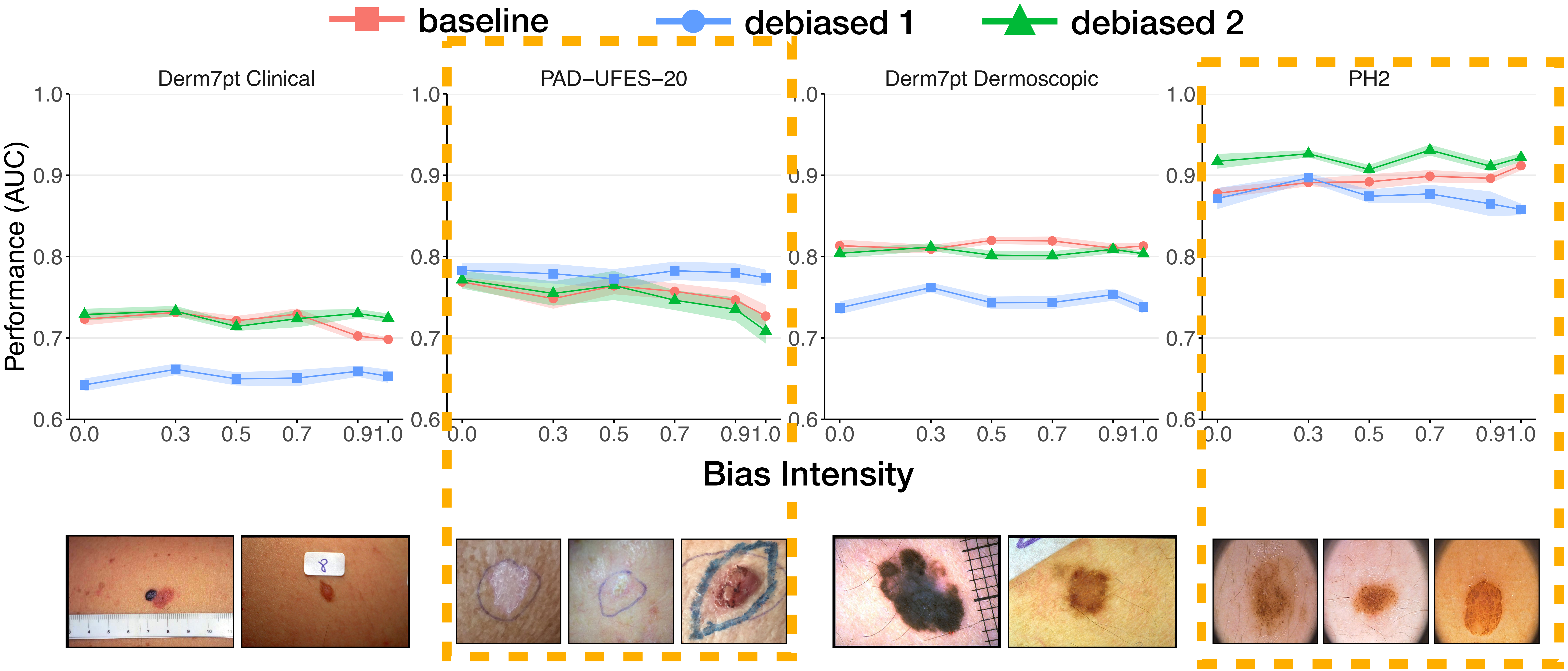
Artifact debiasing solved the generalization problem?

- Gaining robustness to artifacts did not lead to more robust representation in general. What happens in out-of-distribution scenarios is uncertain.



Artifact debiasing solved the generalization problem?

- Gaining robustness to artifacts did not lead to more robust representation in general. What happens in out-of-distribution scenarios is uncertain.



BiasPrune: Debiasing from features alone

A complementary approach





Nourhan Bayasi[1] Jamil Fayyad[2] Alceu Bissoto[3] Ghassan Hamarneh[4] Rafeef Garbi[1]

27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)

October 8th, 2024

[1] UBC THE UNIVERSITY OF BRITISH COLUMBIA

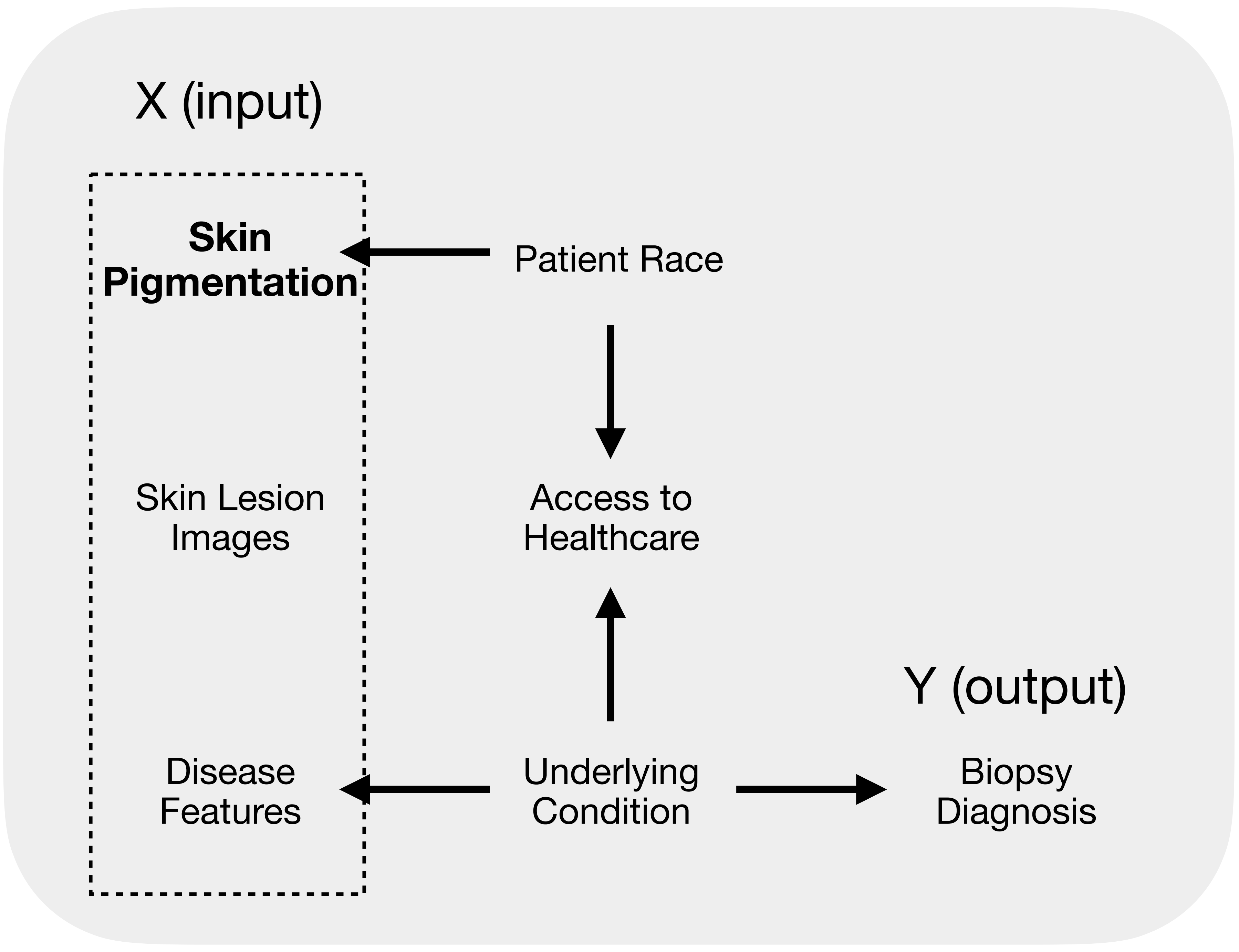
[2] University of Victoria

[3] UNICAMP

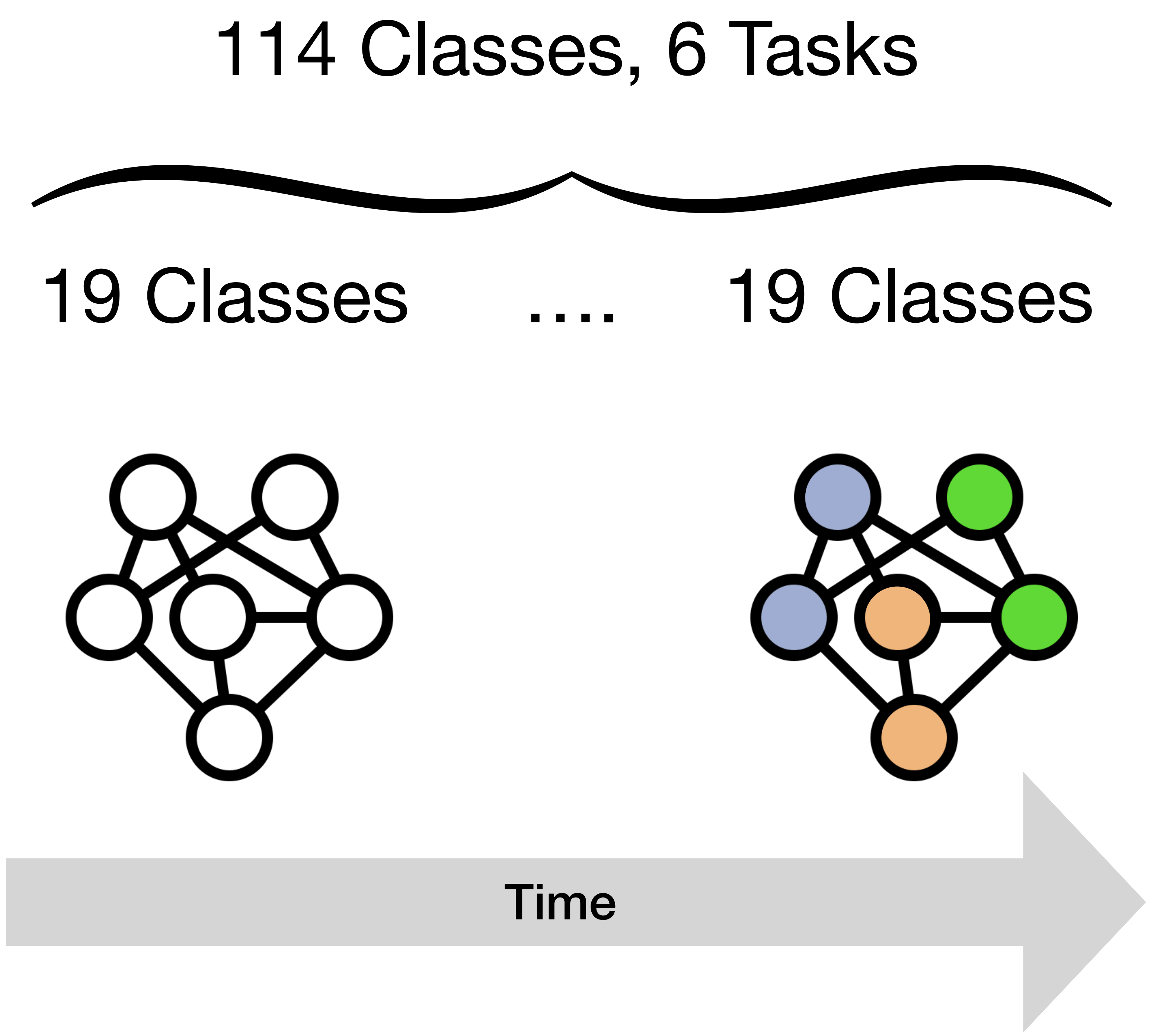
[4] SFU SIMON FRASER UNIVERSITY

Problem Setup - Fitzpatrick17k

Sensitive Attribute: Skin Tone

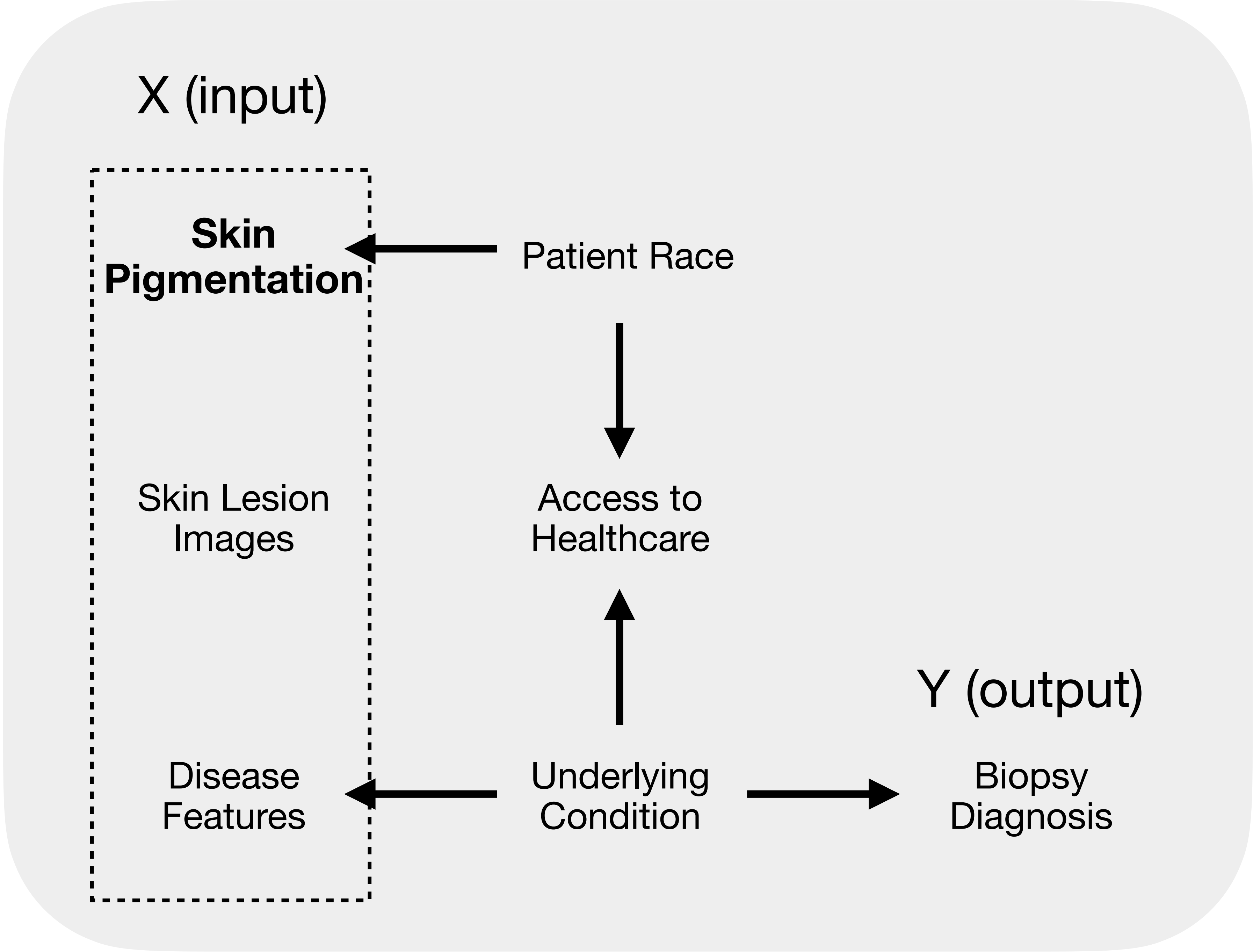


Continual Learning

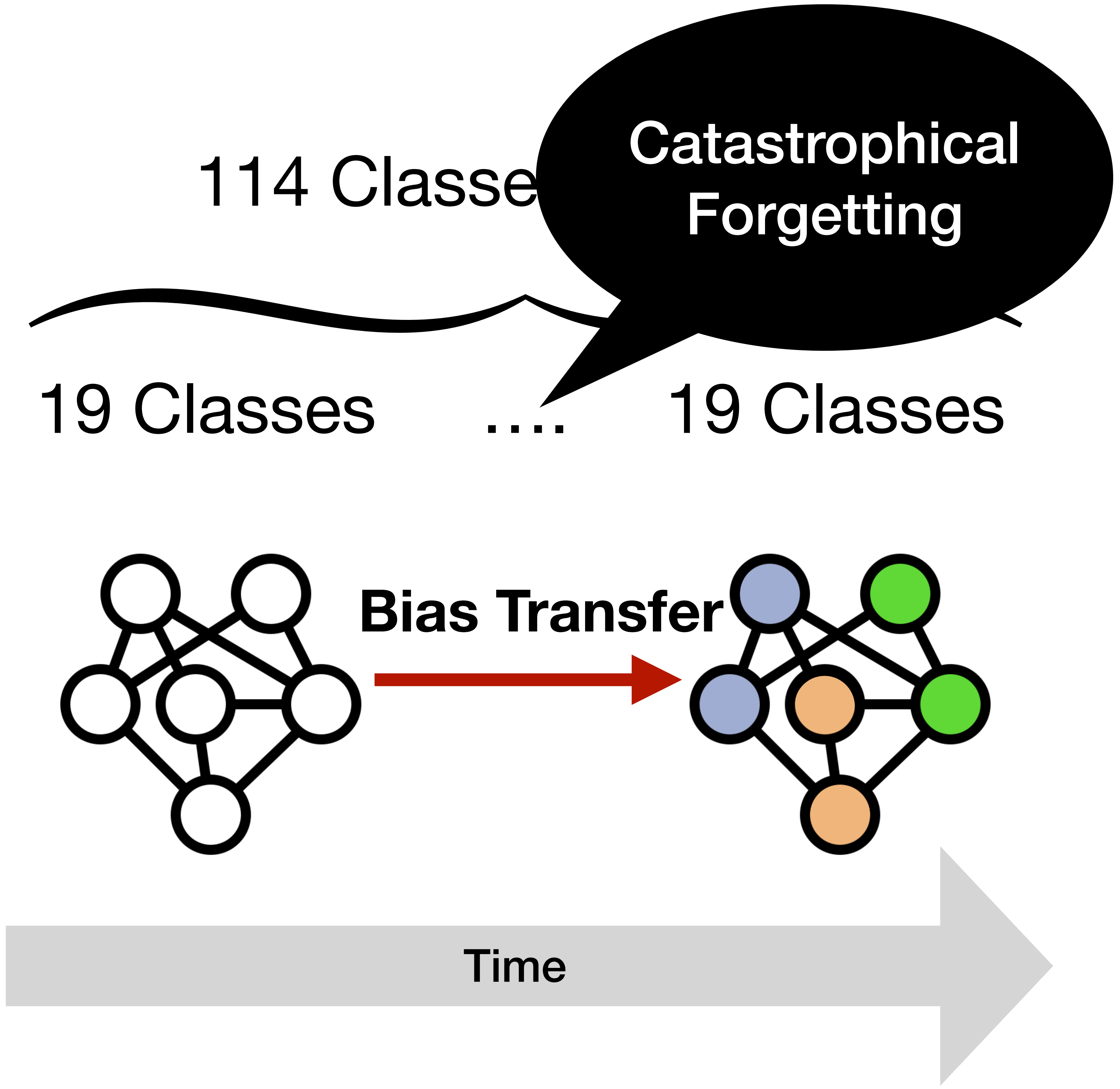


Problem Setup - Fitzpatrick17k

Sensitive Attribute: Skin Tone

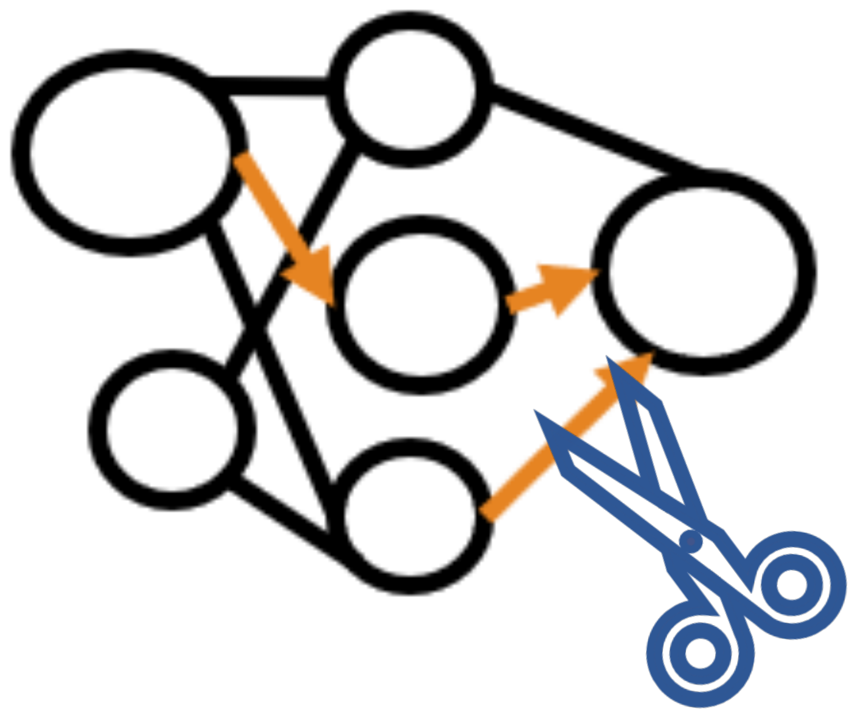


Continual Learning



The Motivation – Can Forgetting Be Good?

Intentionally forget the shortcuts!



BiasPruner

It's a CL method that leverages forgetting to improve fairness while still ensuring the model doesn't forget the important things it has learned.

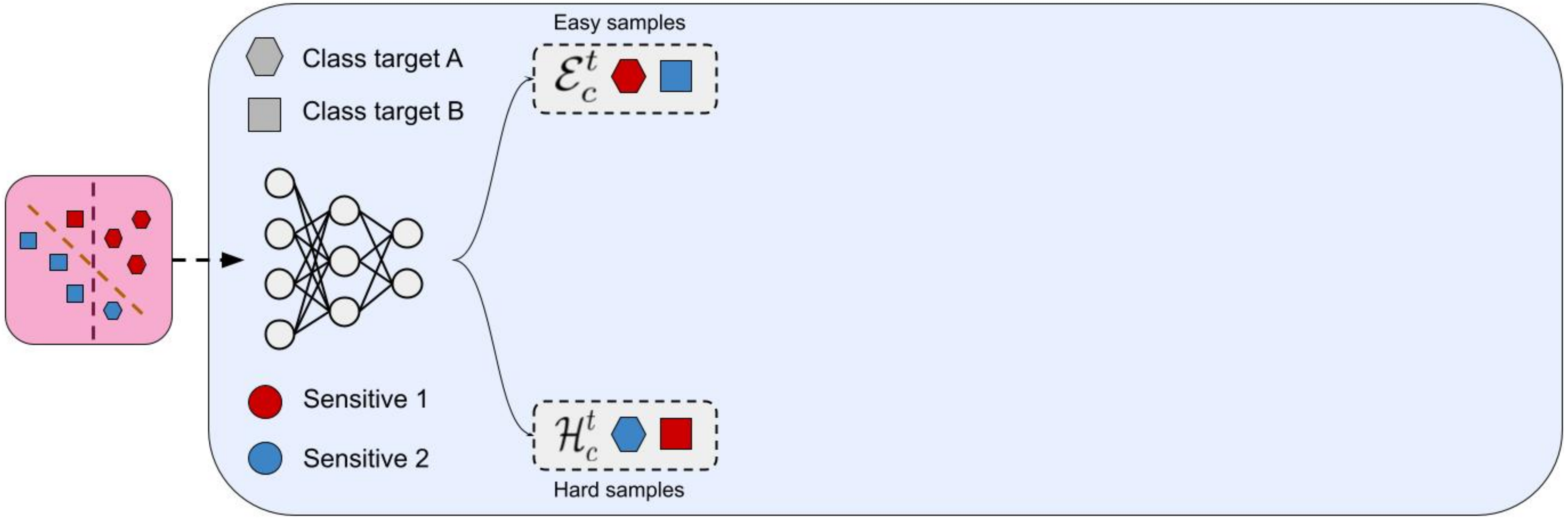
Steps to Find Debiased Subnetwork

1. Measure the bias score of each unit
   a. Encourage the network to be biased.

$$\mathcal{L}_{\text{GCE}}(p(x;\theta), y) = \frac{1 - p_y(x;\theta)^q}{q}$$

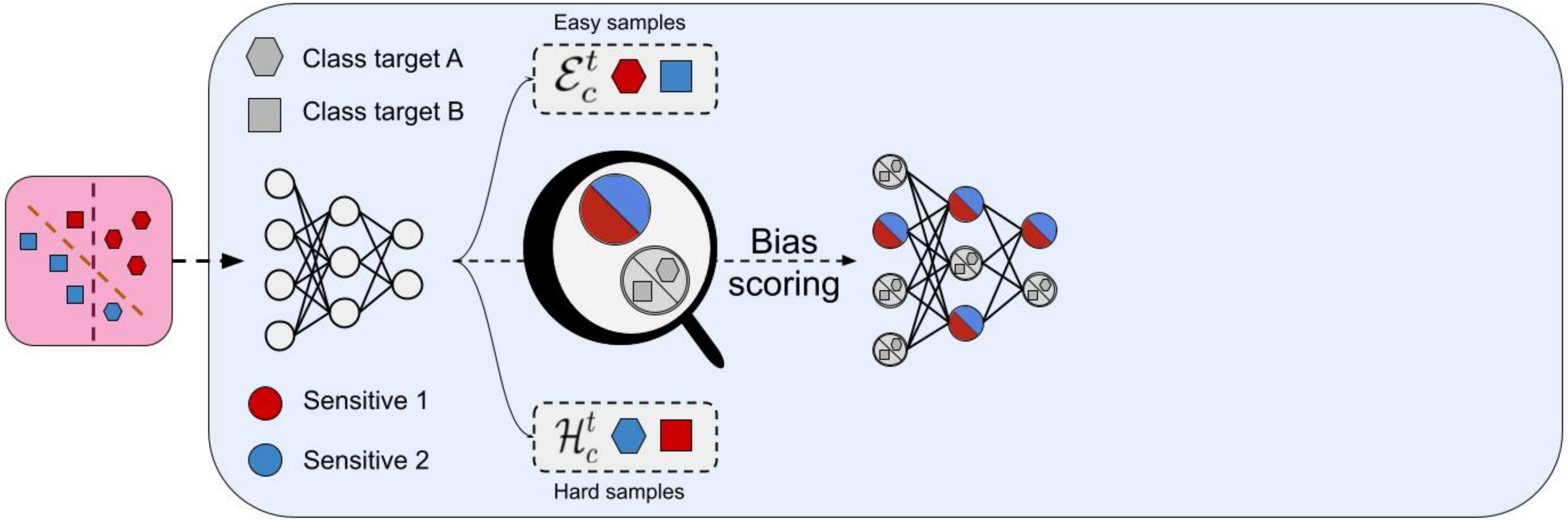
Steps to Find Debiased Subnetwork

1. **Measure the bias score of each unit**
   a. Encourage the network to be biased.
   b. **For each class, find the easy and hard image sets. (based on errors and confidence)**



Steps to Find Debiased Subnetwork

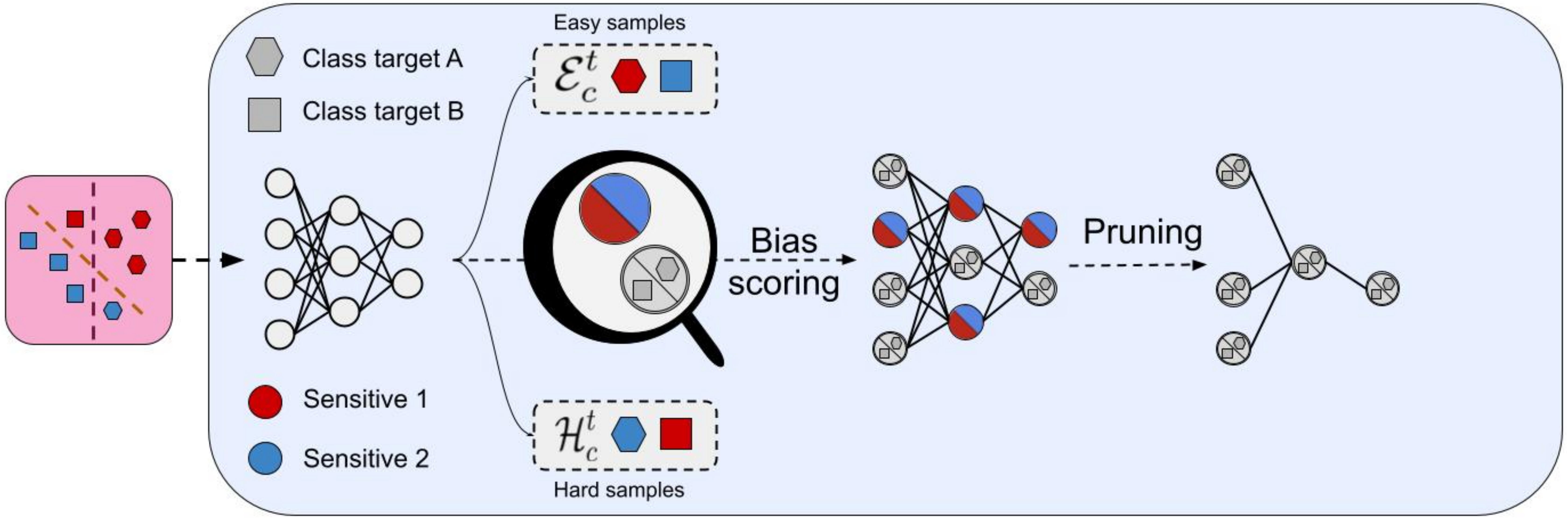
1. **Measure the bias score of each unit**
   a. Encourage the network to be biased.
   b. For each class, find the easy and hard image sets.
   c. **Rank the units most sensitive to easy (biased) samples, and least sensitive to hard samples. The highest ones are biased.**



Steps to Find Debiased Subnetwork

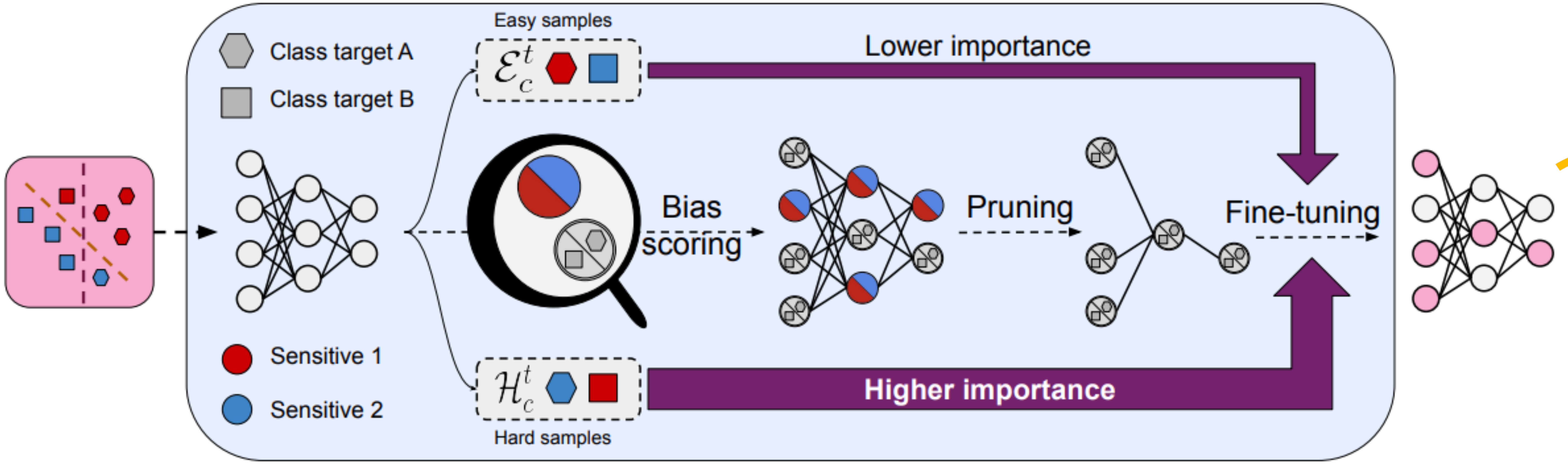
1. Measure the bias score of each unit
   a. Encourage the network to be biased.
   b. For each class, find the easy and hard image sets.
   c. Rank the units most sensitive to easy (biased) samples, and least sensitive to hard samples. The highest ones are biased.

2. Prune units with high bias scores



Steps to Find Debiased Subnetwork

1. Measure the bias score of each unit
    a. Encourage the network to be biased.
    b. For each class, find the easy and hard image sets.
    c. Rank the units most sensitive to easy (biased) samples, and least sensitive to hard samples. The highest ones are biased.
2. Prune units with high bias score
3. **Finetune the subnetwork with weighted CE loss**



Debiased subnetwork!

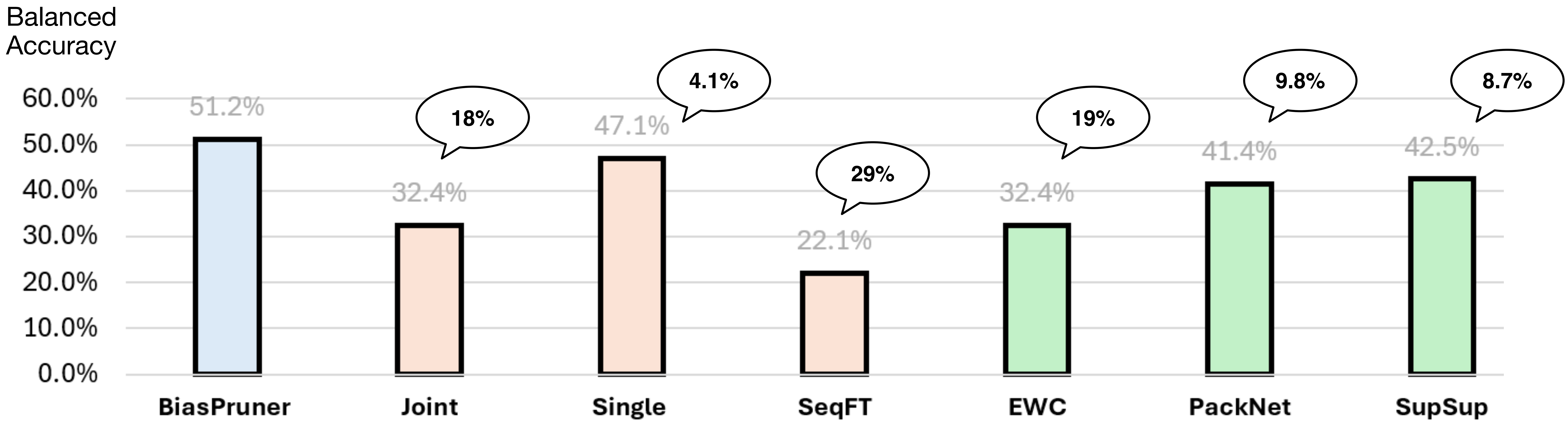
BiasPrune

Performance Results for Fitzpatrick17k

Improved Performance per subgroup and overall

Improved Fairness measurements

Less Bias is Encoded



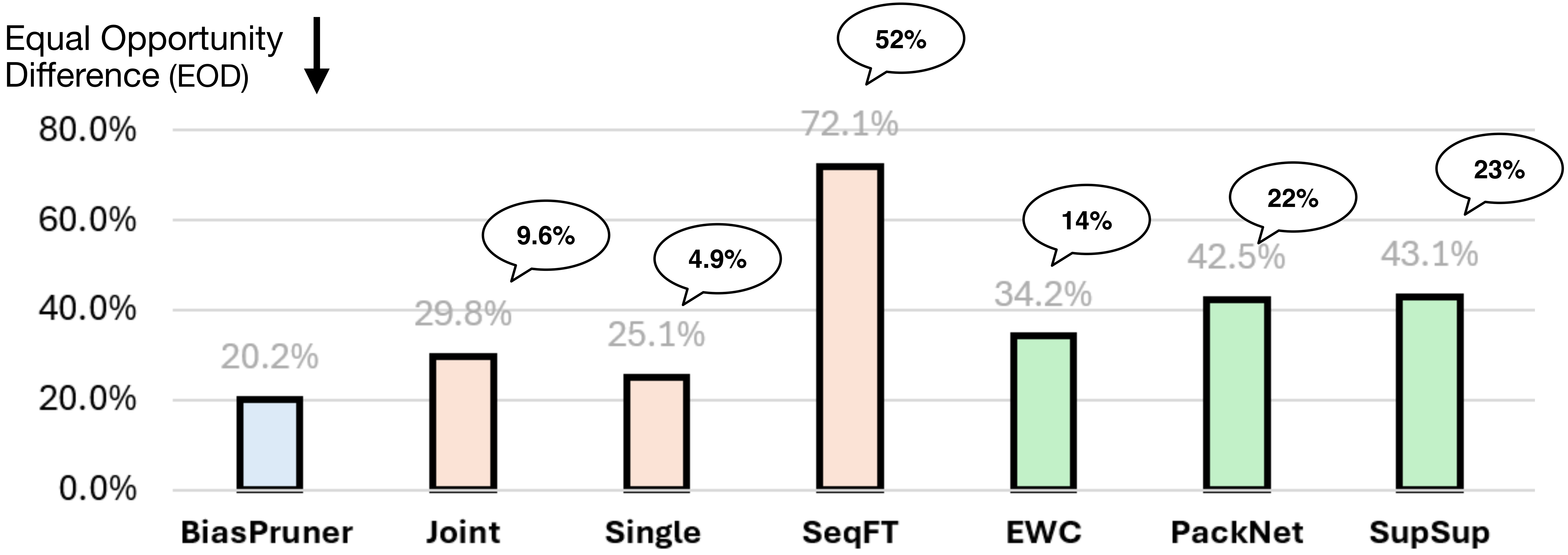
BiasPrune

Fairness Results for Fitzpatrick17k

Improved Performance per subgroup and overall

Improved Fairness measurements

Less Bias is Encoded



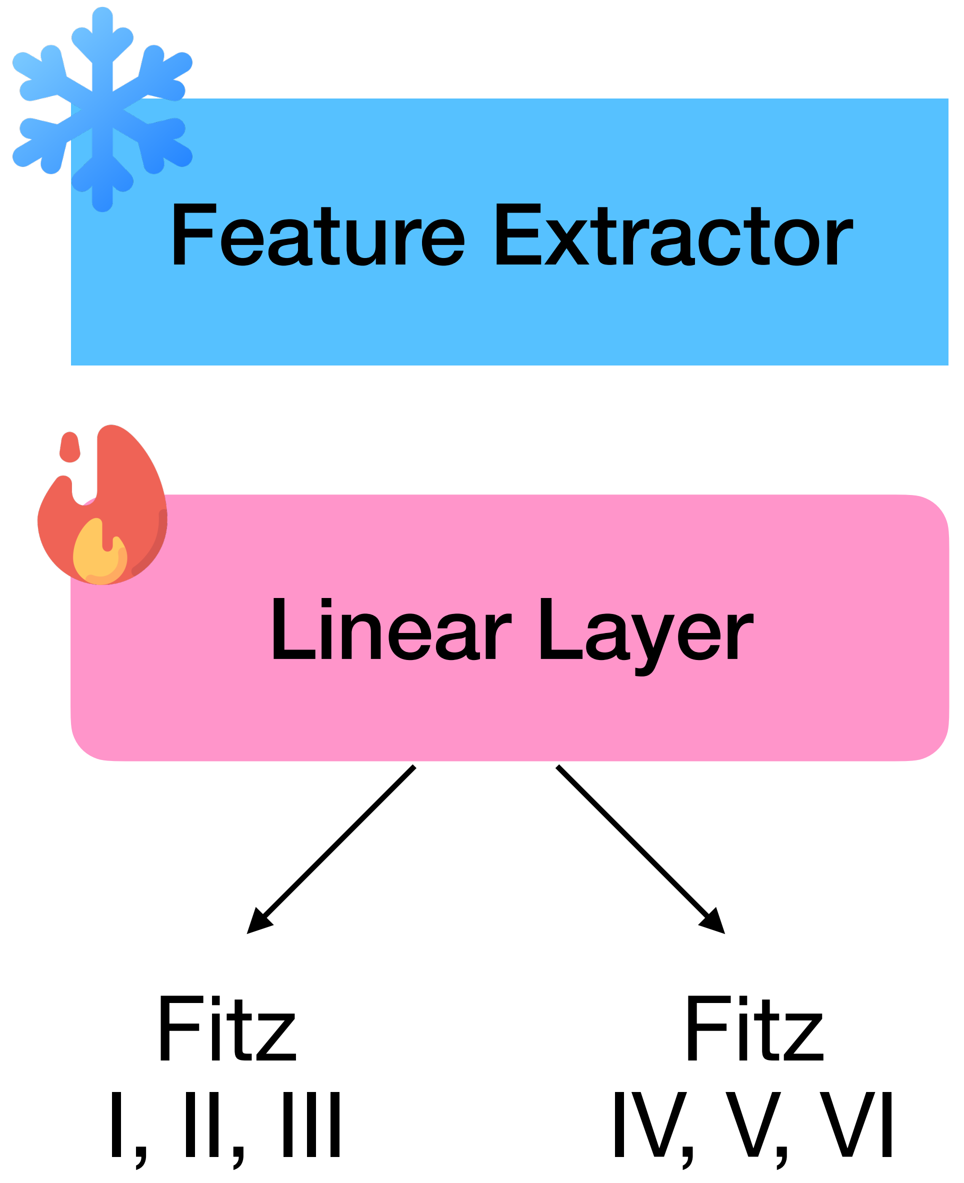
BiasPrune

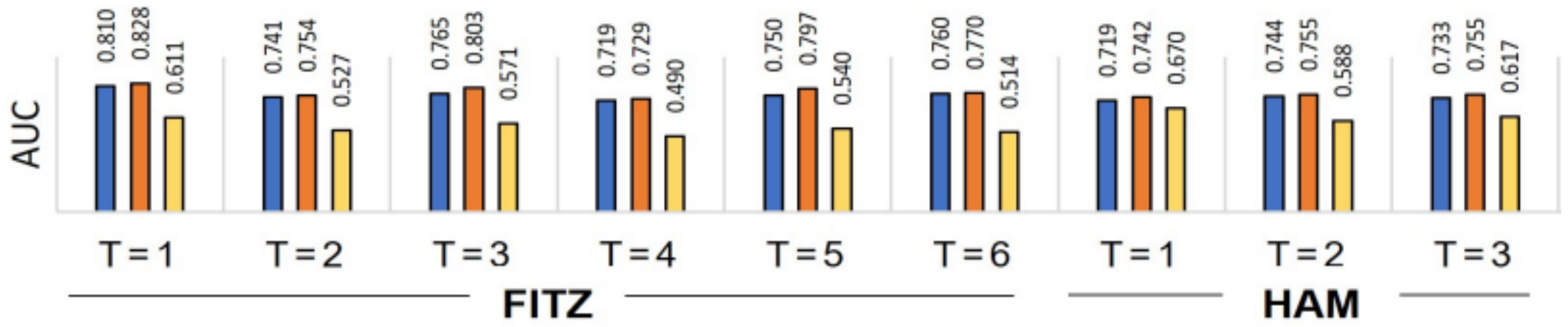
Bias Decodability

Improved Performance per subgroup and overall

Improved Fairness measurements

Less Bias is Encoded





Skin tone

Age

Discussion / Takeaways

1. Define the possibly biased data/problem with causal graphs

2. Make use of the metadata available to incorporate subgroup evaluation

3. The literature is moving towards a mix of bias of interest and learned bias

Code, Data & Papers:
https://github.com/alceubissoto/

UNIVERSITÄT BERN

Thank you!

Alceu Bissoto alceu.bissoto@unibe.ch

UNICAMP

recod

LARA

FAPESP