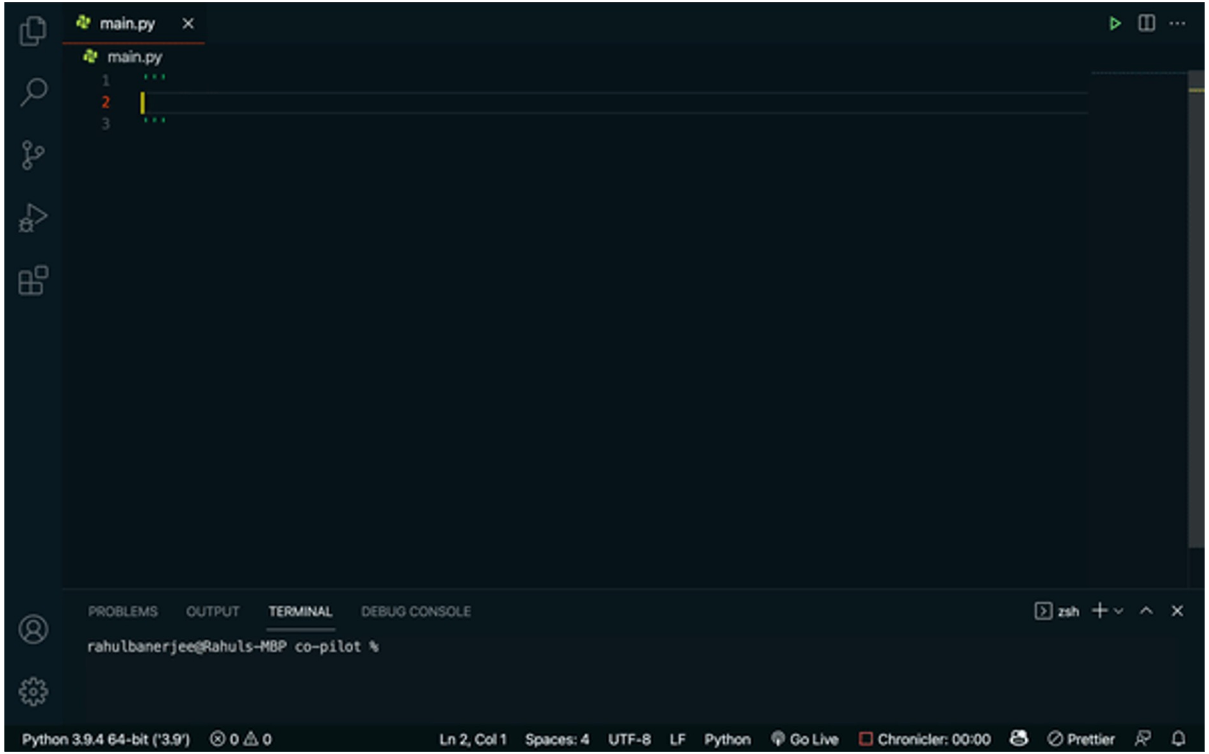
Towards Reliable Machine Learning: Challenges, Examples, Solutions

Sanghyuk Chun
Lead Research Scientist at NAVER AI Lab

NAVER AI LAB

Part 1: Machine Learning Reliability: Challenges and Examples

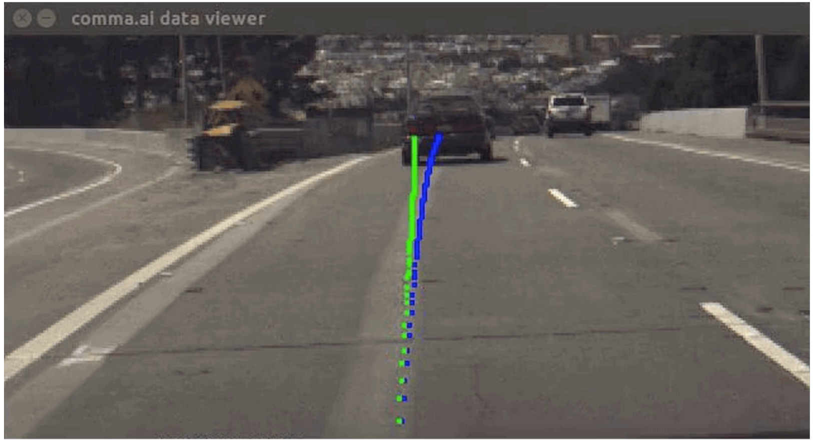
Machine Learning (ML) opens a new stage of automation.



https://towardsdatascience.com/some-experiments-using-github-copilot-with-python-90f8065fb72e

Machine Learning (ML) opens a new stage of automation.

Line tracing for self-driving cars



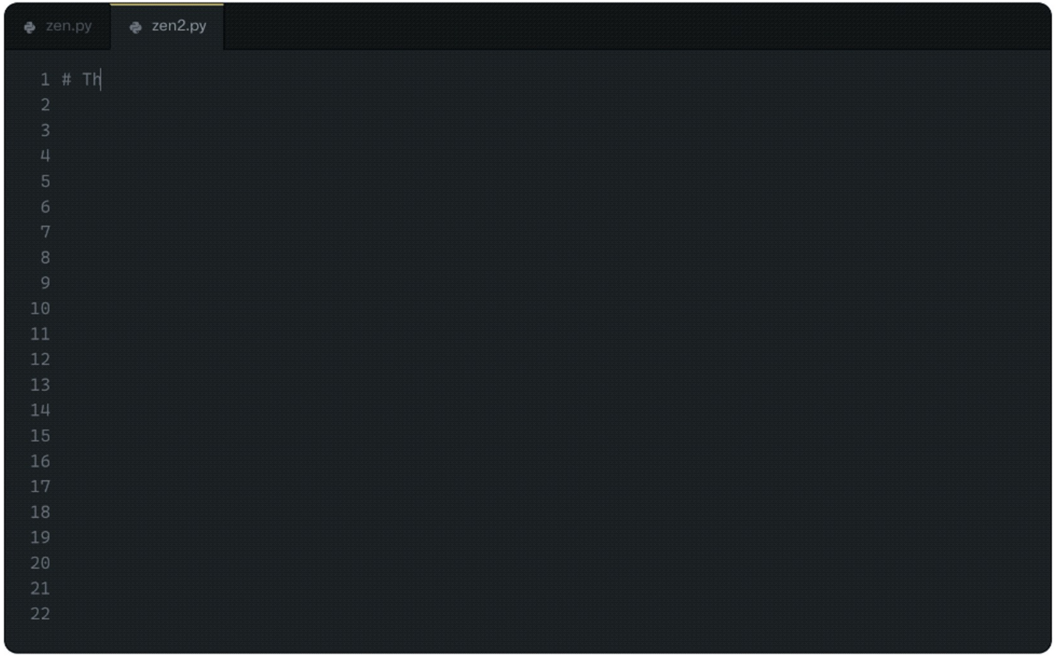
Semantic segmentation for self-driving cars

https://github.com/commaai/research

https://studentsxstudents.com/using-semantic-segmentation-to-give-a-self-driving-car-the-ability-to-see-6c97425ec562

However, AI often cannot understand the problem itself.

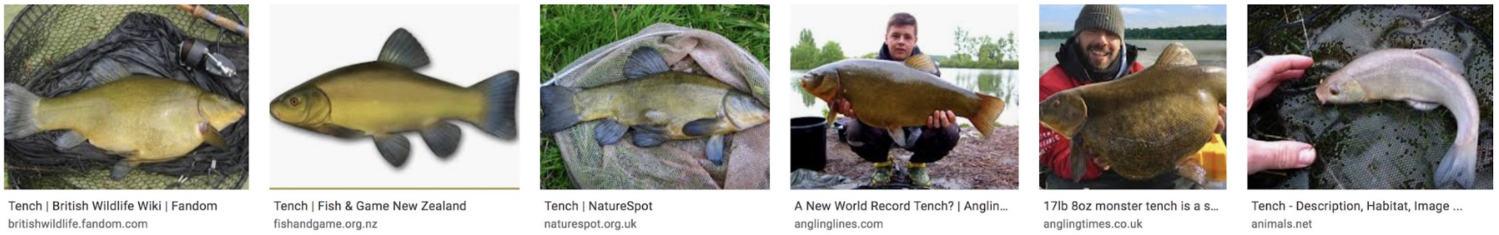
- Just “memorizing” training dataset rather than “thinking” about the given comment

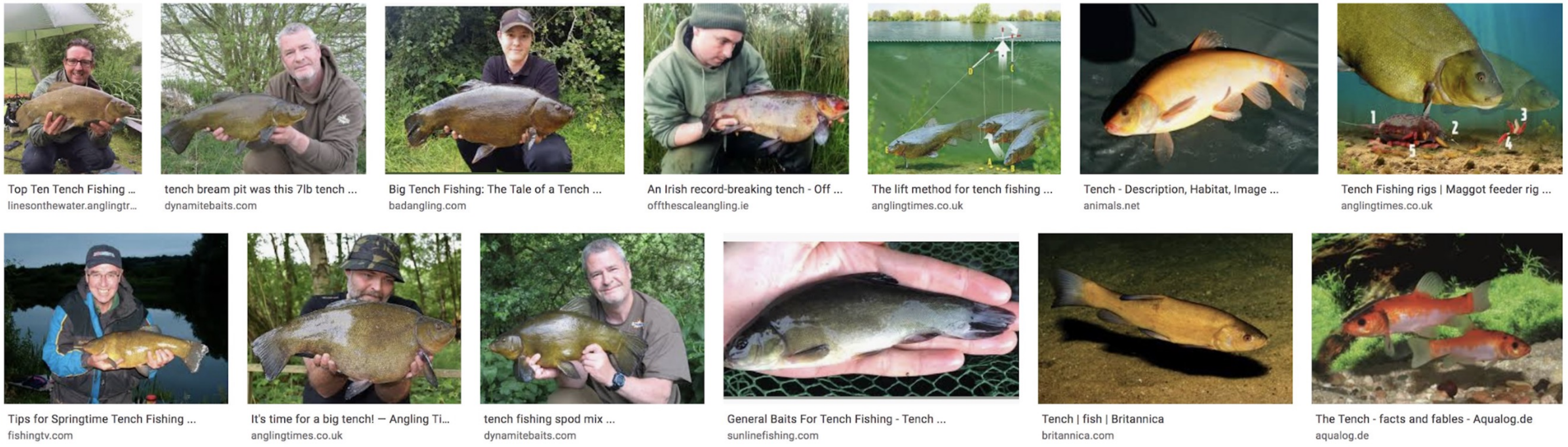


https://docs.github.com/en/github/copilot/research-recitation

However, AI often cannot understand the problem itself.

Google search for "Tench" (class 0)





Wieland et al., "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet", ICLR 2019

However, AI often cannot understand the problem itself.

"Finger bias"

[Wieland et al., "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet", ICLR 2019](https://medium.com/bethgelab/neural-networks-seem-to-follow-a-puzzlingly-simple-strategy-to-classify-images-f4229317261f)
https://medium.com/bethgelab/neural-networks-seem-to-follow-a-puzzlingly-simple-strategy-to-classify-images-f4229317261f

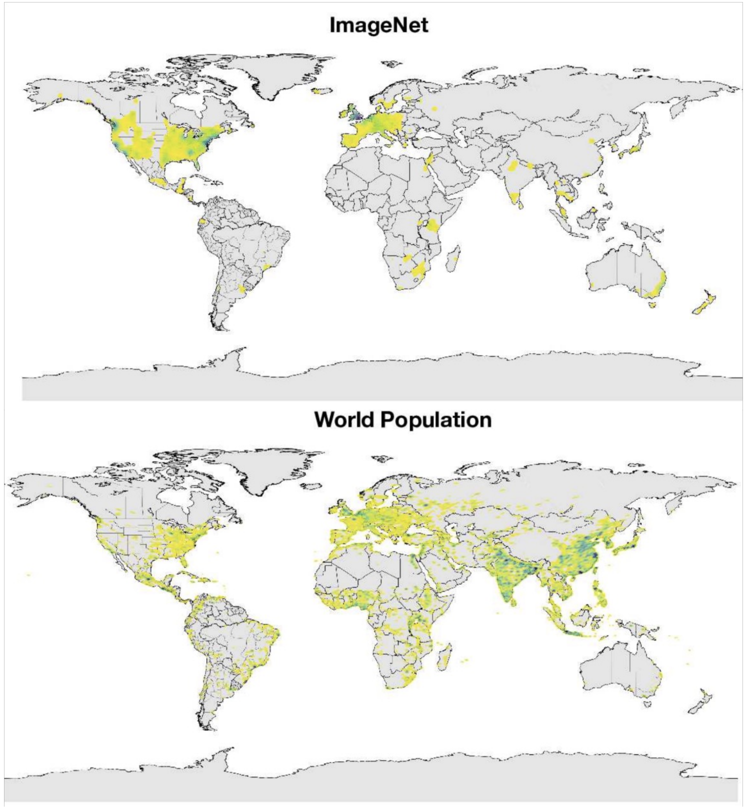
However, AI often cannot understand the problem itself.

Ground truth: Spices **Phillipines, 262 $/month**

Azure: bottle, beer, counter, drink, open
Clarifai: container, food, bottle, drink, stock
Google: product, yellow, drink, bottle, plastic bottle
Amazon: beverage, beer, alcohol, drink, bottle
Watson: food, larder food supply, pantry, condiment, food seasoning
Tencent: condiment, sauce, flavorer, catsup, hot sauce

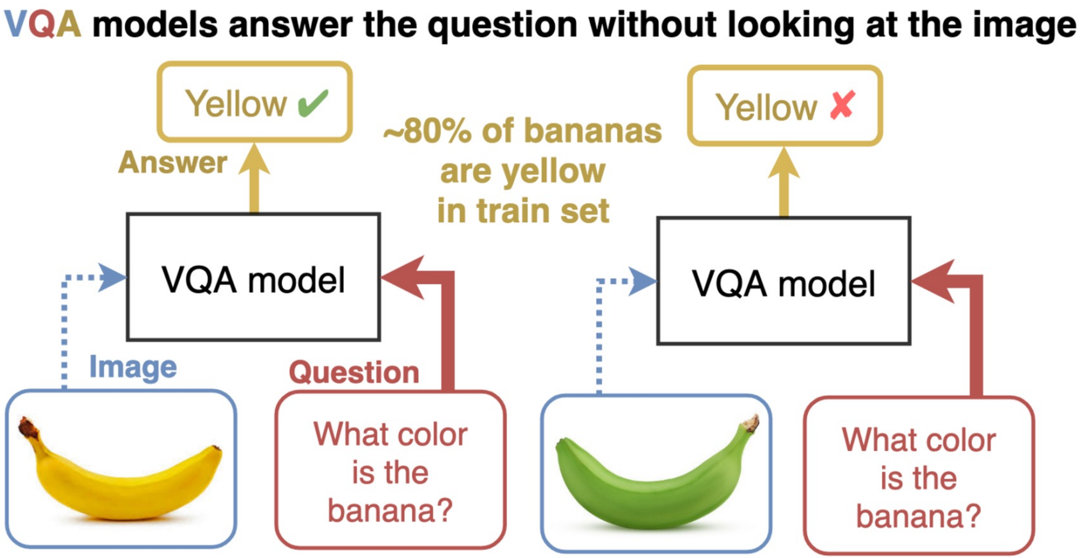
Ground truth: Spices **USA, 4559 $/month**

Azure: bottle, wall, counter, food
Clarifai: container, food, can, medicine, stock
Google: seasoning, seasoned salt, ingredient, spice, spice rack
Amazon: shelf, tin, pantry, furniture, aluminium
Watson: tin, food, pantry, paint, can
Tencent: spice rack, chili sauce, condiment, canned food, rack



[de Vries, Terrance, et al. "Does object recognition work for everyone?." CVPR Workshops. 2019.]

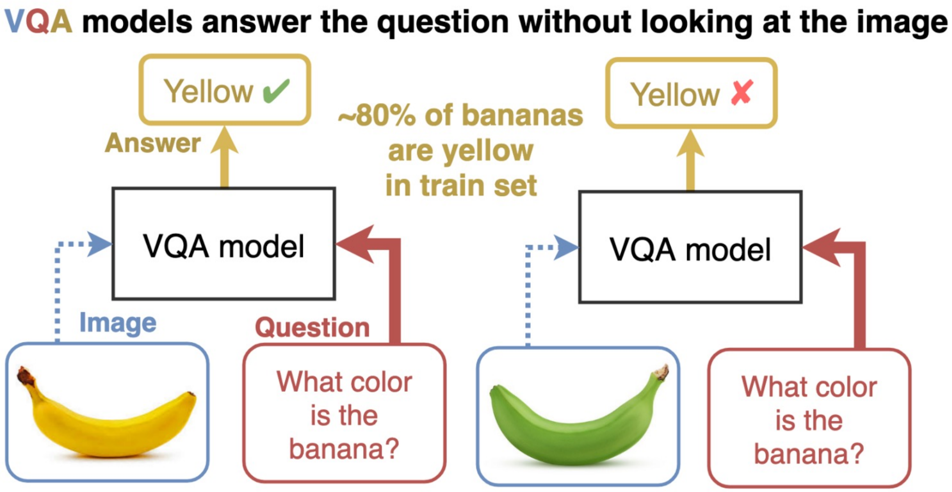
ML models often rely on “easy-to-learn shortcuts” without an understanding of the problem itself.



[Cadene, et al. “RUBi: Reducing Unimodal Biases for Visual Question Answering”, NeurIPS 2019](#)

“Shortcut learning” problem?

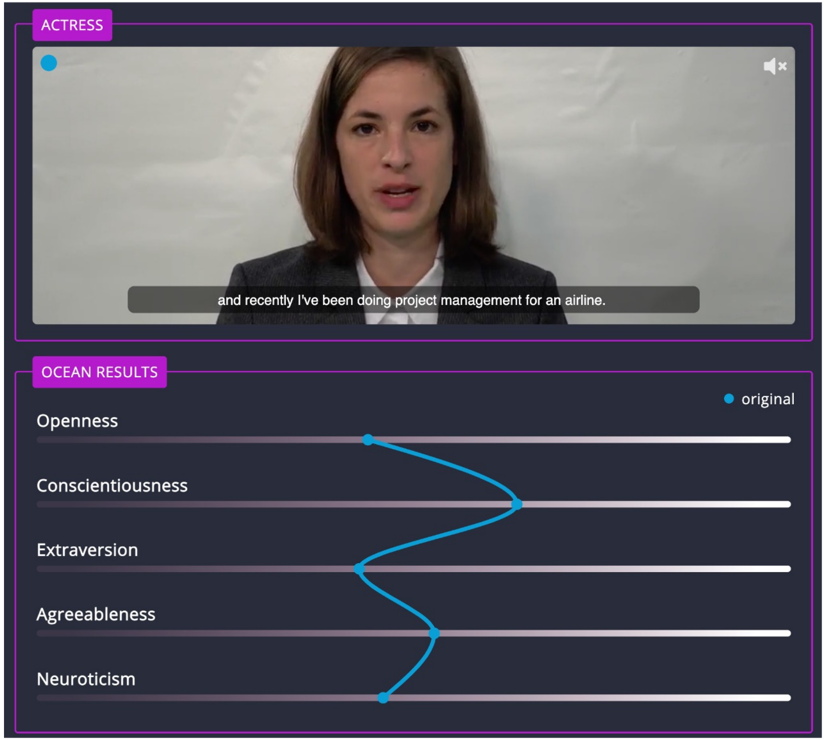
- When a model does not make a decision based on **“desired”** features (considering both question and image – color in this case), but **“undesired”** features (ignoring image), there exists **a shortcut learning problem.**



[Cadene, et al. “RUBi: Reducing Unimodal Biases for Visual Question Answering”, NeurIPS 2019]

Shortcut learning in human-related applications.

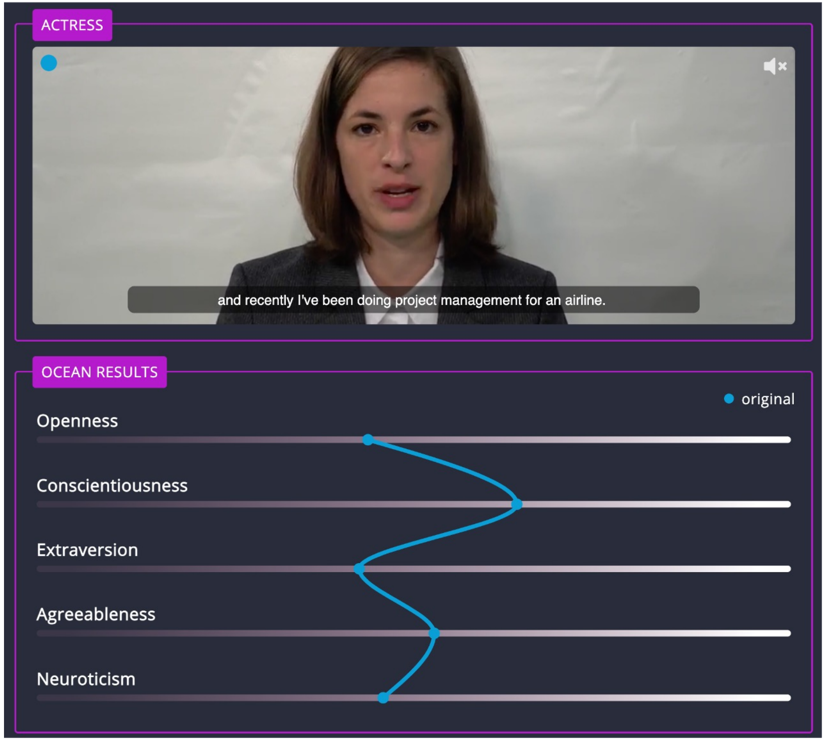
- Example: An automatic job interview
- A machine determines “Openness”, “Conscientiousness”, “Extraversion”, “Agreeableness”, and “Neuroticism” of interviewees.
- **Is the machine truly focusing on how the interviewee answers the question?**
- We expect a machine interviewer to be:
  - Fast and Accurate
  - Inexpensive and Convenient
  - Reproducible
  - … and Objective (not biased)



Objective or Biased https://interaktiv.br.de/ki-bewerbung/en/

Shortcut learning in human-related applications.

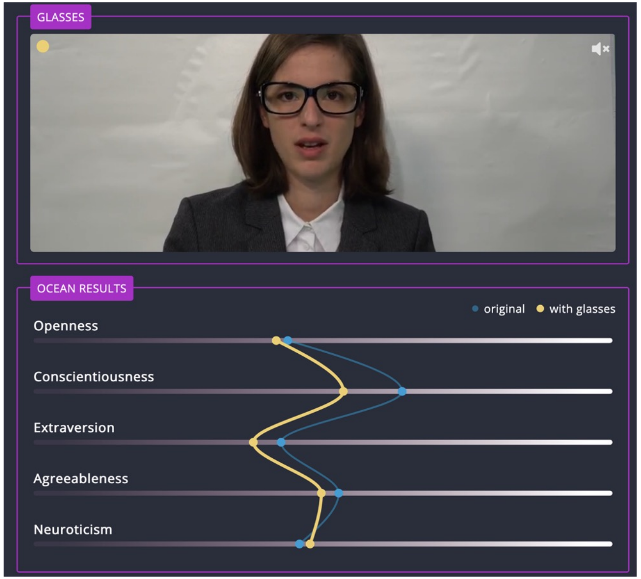
- **Is the machine truly focusing on how the interviewee answers the question?**
- This article ("Objective or Biased" by Bayerischer Rundfunk — German Public Broadcasting) investigates the question by hiring a professional actress and letting the actress act an interviewee.
- If the machine truly understands the answers, different appearances do not affect to the results.

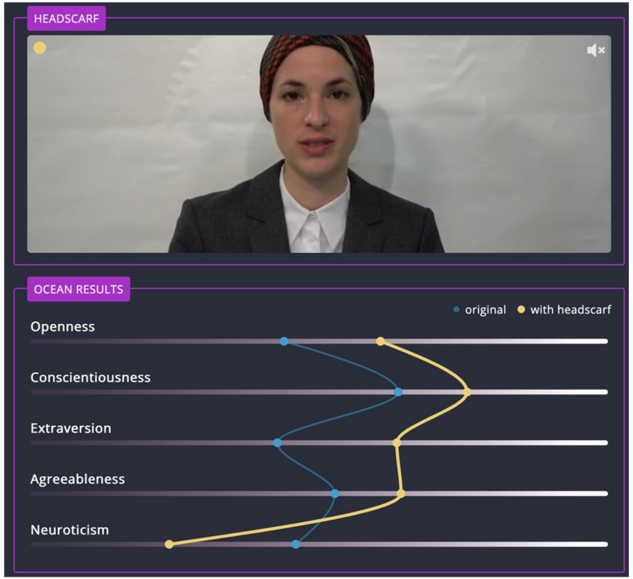


Objective or Biased https://interaktiv.br.de/ki-bewerbung/en/

Shortcut learning in human-related applications.

- When the actress acts **the same script and the same action** but **with different appearance (with glasses or with headscarf)**, **the predictions vary significantly!**

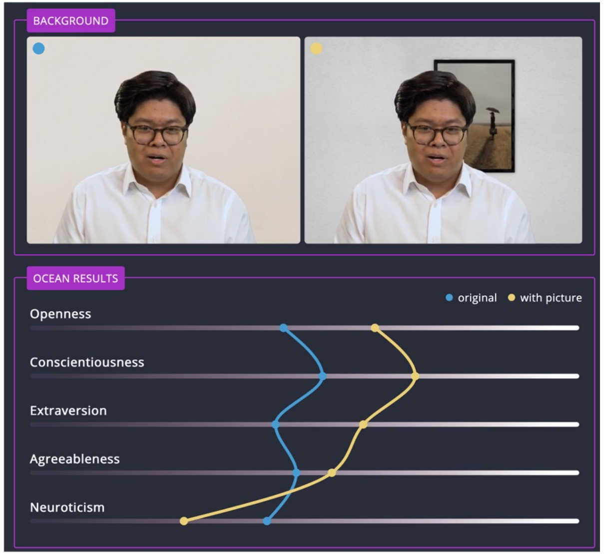


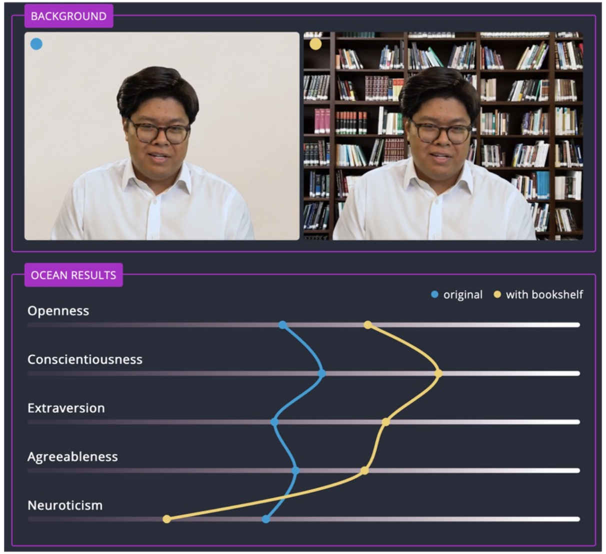


Objective or Biased https://interaktiv.br.de/ki-bewerbung/en/

Shortcut learning in human-related applications.

- Similarly, **the predictions vary significantly** with **the same script and the same action** but **with different backgrounds (with picture, with bookshelf)**.

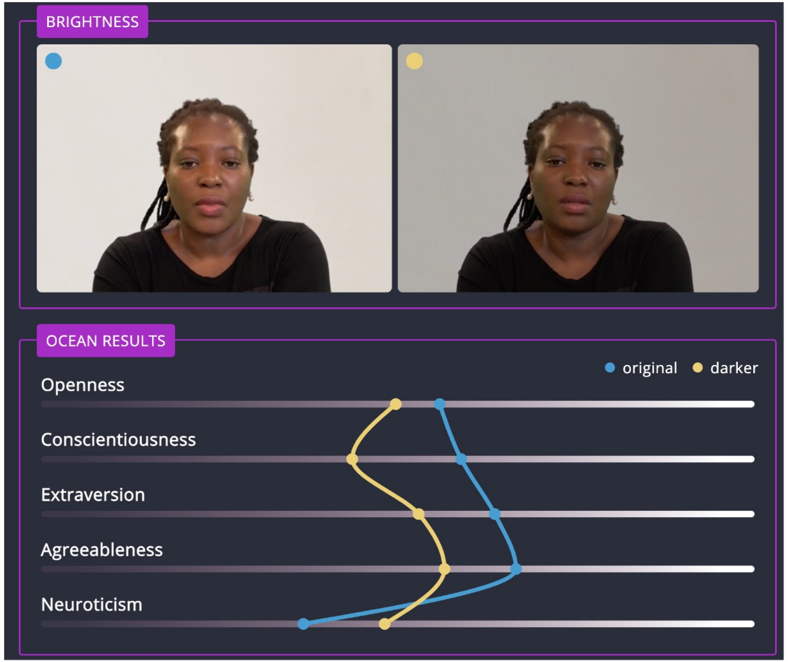




Objective or Biased https://interaktiv.br.de/ki-bewerbung/en/

Shortcut learning in human-related applications.

- Even for different brightness settings!



Objective or Biased https://interaktiv.br.de/ki-bewerbung/en/

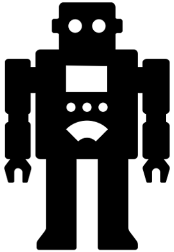
Recap: the basic of machine learning.

Data X

Model parameter θ

Model prediction $f_\theta(X)$

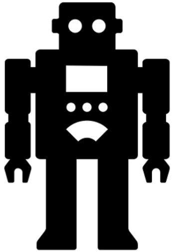
Label Y



“Santa Claus” != “Cat”

Fix the model parameters to make the prediction correct

Recap: the basic of machine learning.

Data X	Model parameter θ	Model prediction $f_\theta(X)$	Label Y
		“Cat”	“Cat”

There is no update if the prediction is correct

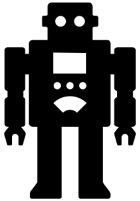
When does shortcut learning happen?

Car

Every car in the dataset is on roads

Boat

Every boat in the dataset is on water



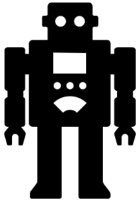
Model

Hmmmm **“Boat”** images have water patterns & **“Car”** images have road patterns !

When does shortcut learning happen?

Car on water

Boat on road



Model

This is **“Boat”**.

This is **“Car”**.

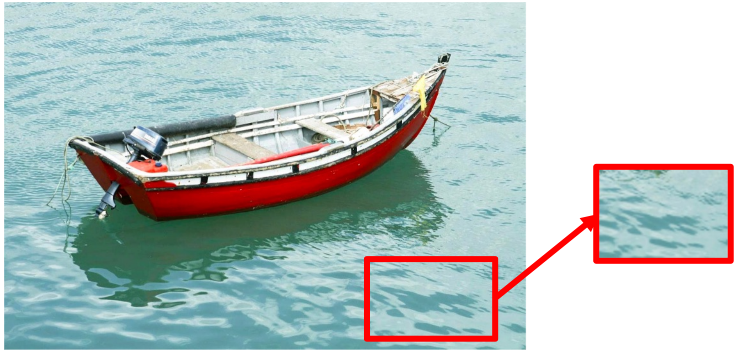
When does shortcut learning happen?

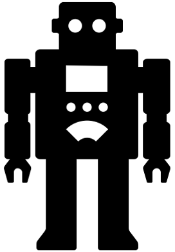
Data X

Model parameter θ

Model prediction $f_{\theta}(X)$

Label Y





[water] background means **"boat"!**

"Boat"

A model only attends on **"easy shortcut" (water pattern)**, but it actually does not understand the problem

There is no update if the prediction is correct

ICLR'22

Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective

Luca Scimeca*, Seong Joon Oh*, Sanghyuk Chun, Michael Poli, Sangdoo Yun
NAVER AI Lab

Class 1 **Class 2** **Class 3**

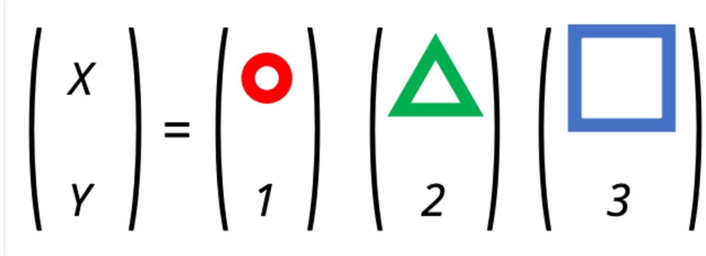
Question

What is the class of this small blue triangle? 

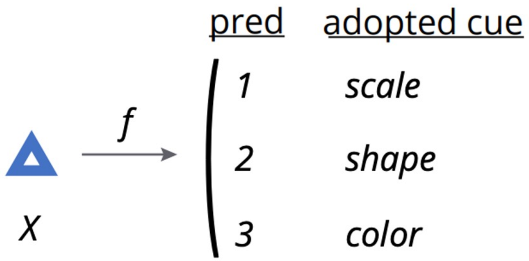
Scimeca, et al. “Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective”, ICLR 2022.

ML models have more preferred features.

- Assume a toy example when every training sample with …
    - class “1” is small / red / circle
    - class “2” is medium / green / triangle
    - class “3” is large / blue / square

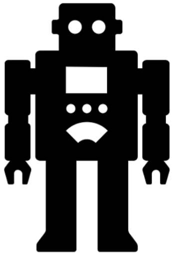


How to predict
“small (1)” “blue (3)” “triangle (2)”?



Scimeca, et al. “Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective”, ICLR 2022.

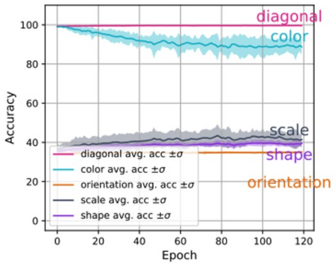
Naive learning strategy leads to shortcut learning



I prefer to recognize objects by colors!

$X \xrightarrow{f}$

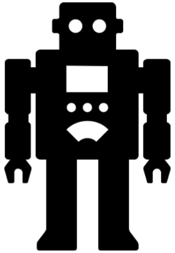
pred	adopted cue
1	*scale*
2	*shape*
$\boxed{3}$	$\boxed{\textit{color}}$



Shortcut learning happens!

Task 1: If color is **a desired feature** and others are undesired features

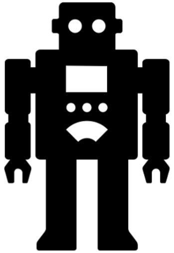
Desired answer: 3



It is class **“3”**

Task 2: If **shape** is **a desired feature**, but color is **an undesired feature**

Desired answer: **2**

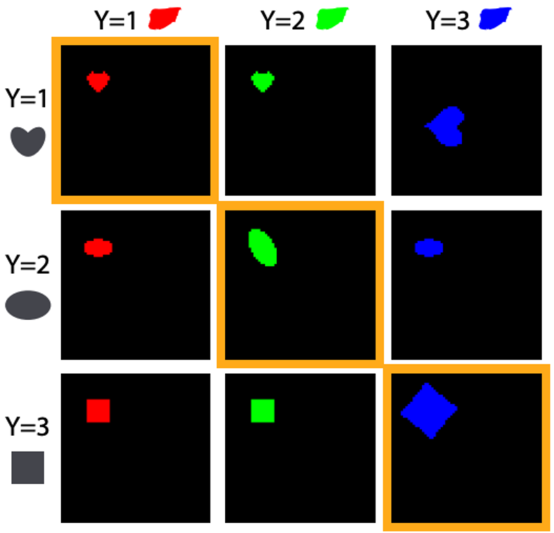


It is class **“3”**

Scimeca, et al. “Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective”, ICLR 2022.

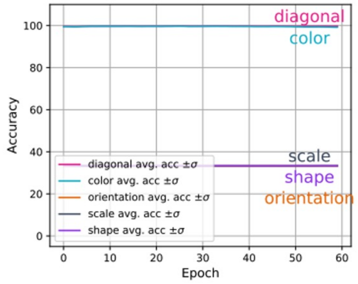
There exist “preferred cues” by ML models

Trained on “100% correlated” data

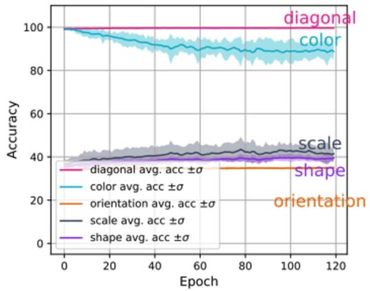


(a) DSprites. $\mathbb{S} = \{\text{shape}, \text{color}\}$.

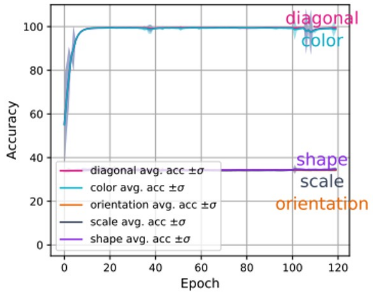
Tested on “uncorrelated” data



FFNet



ResNet20

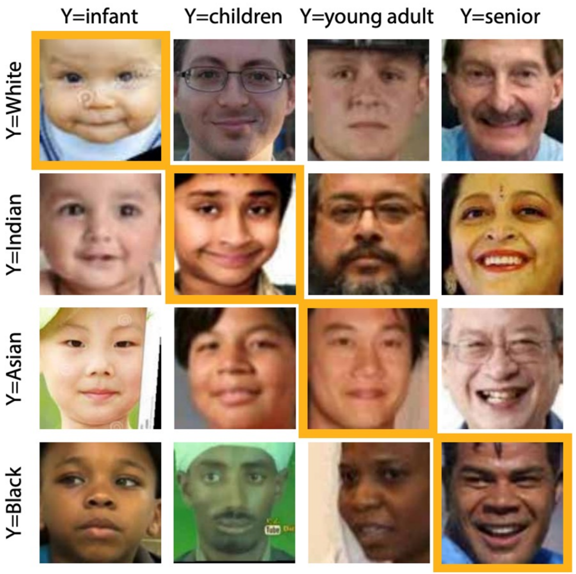


ViT

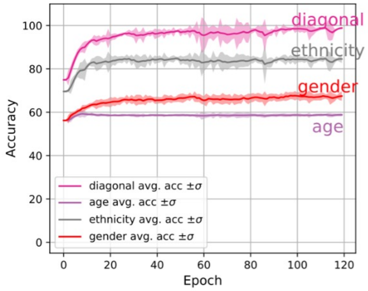
[Scimeca, et al. “Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective”, ICLR 2022.]

There exist “preferred cues” by ML models

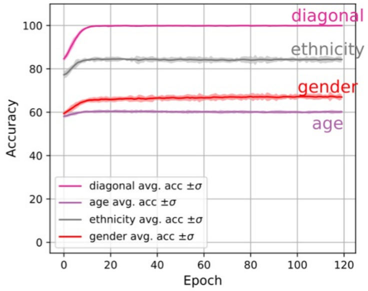
Trained on “100% correlated” data



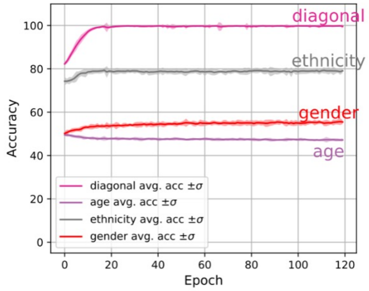
Tested on “uncorrelated” data



FFnet



ResNet20



ViT

[Scimeca, et al. “Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective”, ICLR 2022.]

Conclusion of Part 1

- Shortcut learning problem is a realistic challenge
- Data collection can be a problem, but don't blame dataset collection process too much. We have to focus on the algorithms as well!
- ML models have specific preferred cues (e.g., color):
  A model trained by simple empirical risk minimization (ERM) will be biased toward specific cues with high probability.
- We need a new paradigm for mitigating shortcut learning problem.

Part 2: Attempts to Mitigate Shortcut Learning

How can we solve shortcut learning problem?

- **Data collection** should be considered and designed to ***avoid undesirable dataset biases*** in the dataset (e.g., “lung pics and hospital tokens”)
- We may need **additional annotation process** to collect ***“bias labels of concern”*** (e.g., ethical labels such as ethnicity, gender or income level)
- **Algorithm** should reflect the task itself and should preferably be ***unbiased*** sometimes we need a different decision process, e.g., stochastic one
- **Evaluation protocol** should consider the ***real-world deployment scenario*** such as distribution shifts, not only “in-distributed scenario”.
- Human should be considered in the machine learning model development pipeline loop, i.e., ***human-in-the-loop*** is required

If we have “bias labels” then…

Car

Bias label: “road”

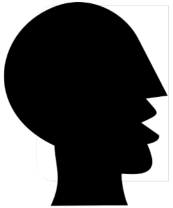
Bias label: “water”

Boat

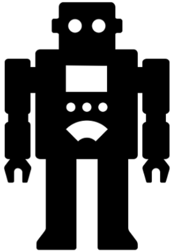
Bias label: “water”

Bias label: “road”

If we have “bias labels” then…



These images have
“road” shortcut labels



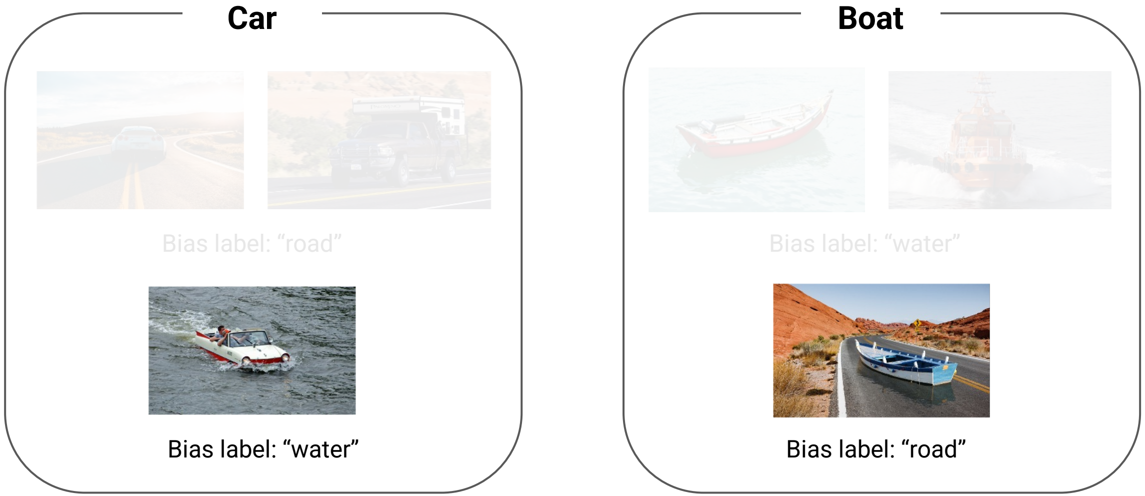
Cool, now I can see what
is the “road” pattern!
I will ignore them.

We have to start from dataset construction.

- **Data collection** should be considered and designed to ***avoid undesirable dataset biases*** in the dataset (e.g., collecting more data samples for “boat on road” or “minority”)
- We also need **additional annotation process** to collect ***“bias labels of concern”*** (e.g., “road” / “water” labels or sensitive attributes such as gender, ethnicity)

Issue 1: Data collection itself is often non-trivial.

- Data collection should be considered and designed to *avoid undesirable dataset biases* in the dataset (e.g., collecting more data samples for “boat on road” or “minority”)
- We also need additional annotation process to collect *“bias labels of concern”* (e.g., “road” / “water” labels or sensitive attributes such as gender, ethnicity)

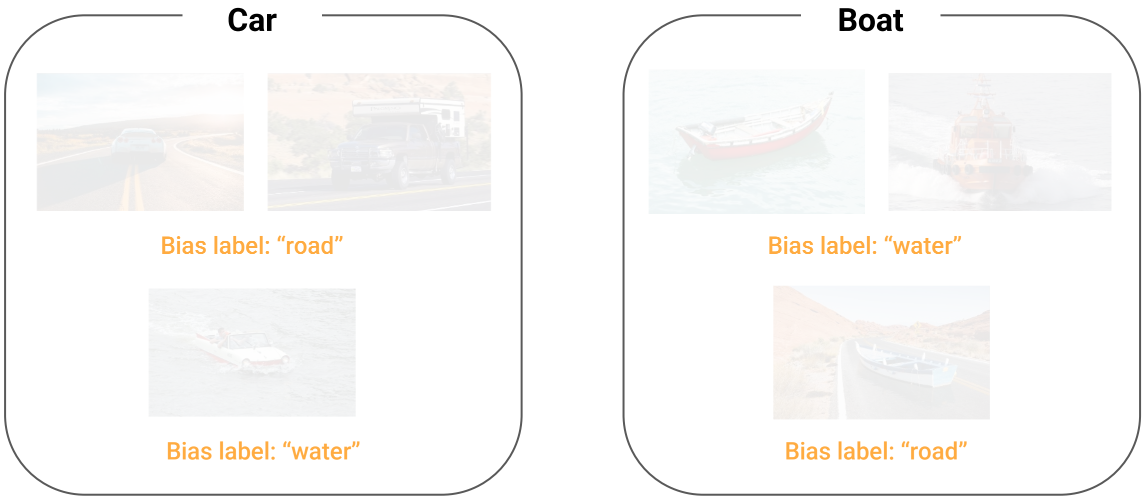


Issue 1: Data collection itself is often non-trivial.

- Collecting “uncommon” data points would be very expensive or even impossible.

Issue 2: Bias label annotations could be expensive.

- **Data collection** should be considered and designed to ***avoid undesirable dataset biases*** in the dataset (e.g., collecting more data samples for "boat on road" or "minority")
- We also need **additional annotation process** to collect ***"bias labels of concern"*** (e.g., "road" / "water" labels or sensitive attributes such as gender, ethnicity)



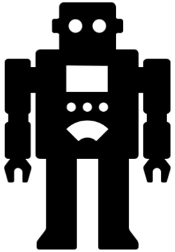
Solution: Estimating bias labels with a few labeled samples.

Bias label: “road”

Bias label: “water”

Bias label: “water”

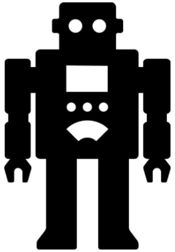
Bias label: “road”



I think I can guess what is “water” and what is “road”

Solution: Estimating bias labels with a few labeled samples.

- Estimating bias labels of the remaining data samples using the bias label estimator
- Now, we can apply the existing methods utilizing the estimated bias labels



This is water

This is road

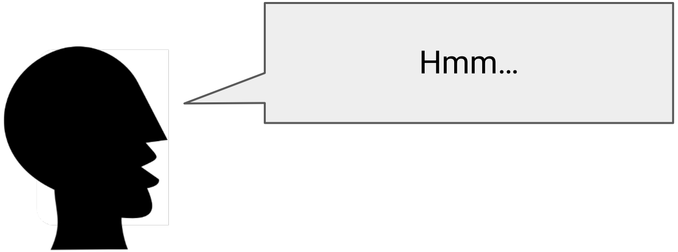
Nam, et al. "Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation" ICLR 2022
Jung, et al. "Learning Fair Classifiers with Partially Annotated Group Labels" CVPR 2022

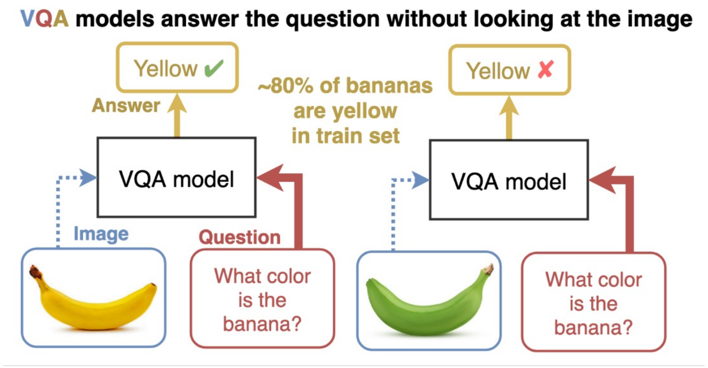
Issue 3: Bias labels are not always easily determinable.

- **Data collection** should be considered and designed to *avoid undesirable dataset biases* in the dataset (e.g., collecting more data samples for “boat on road” or “minority”)
- We also need **additional annotation process** to collect *“bias labels of concern”* (e.g., “road” / “water” labels or sensitive attributes such as gender, ethnicity)

Issue 3: Bias labels are not always easily determinable.

- Sometimes, shortcut bias labels are not able to be “annotated”, e.g., how to annotate the “text-bias” in VQA as “multinomial” values?

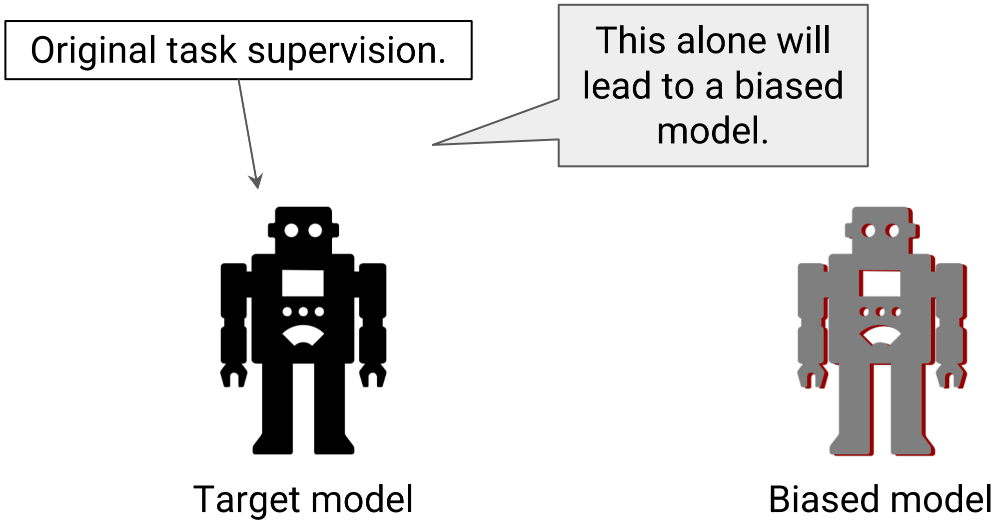




[Cadene, et al. “RUBi: Reducing Unimodal Biases for Visual Question Answering”, NeurIPS 2019]

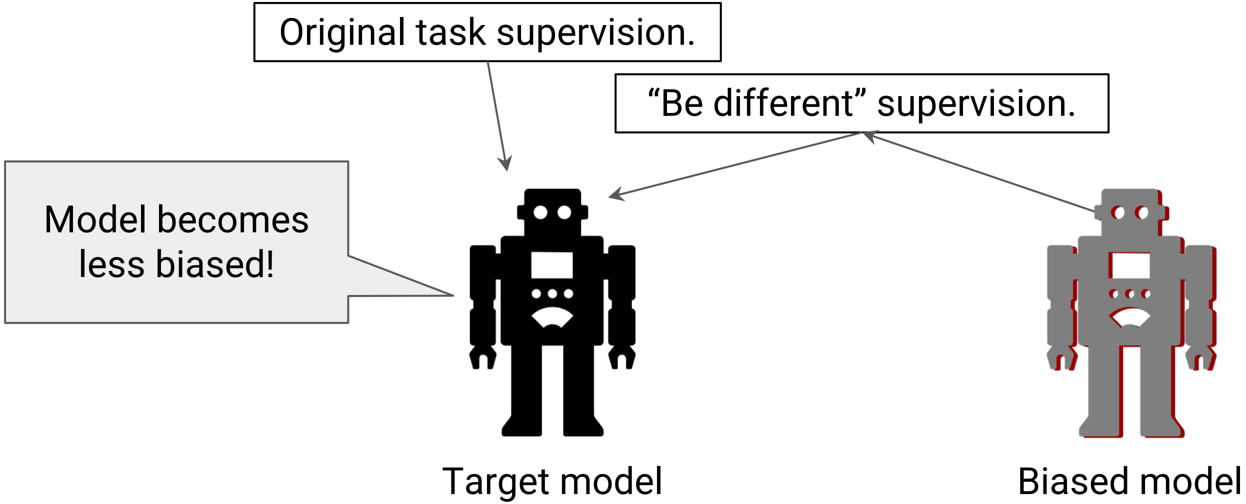
Solution: De-biasing with biased models.

- If we can make a biased model, we can avoid shortcut learning problem by utilizing the biased model.



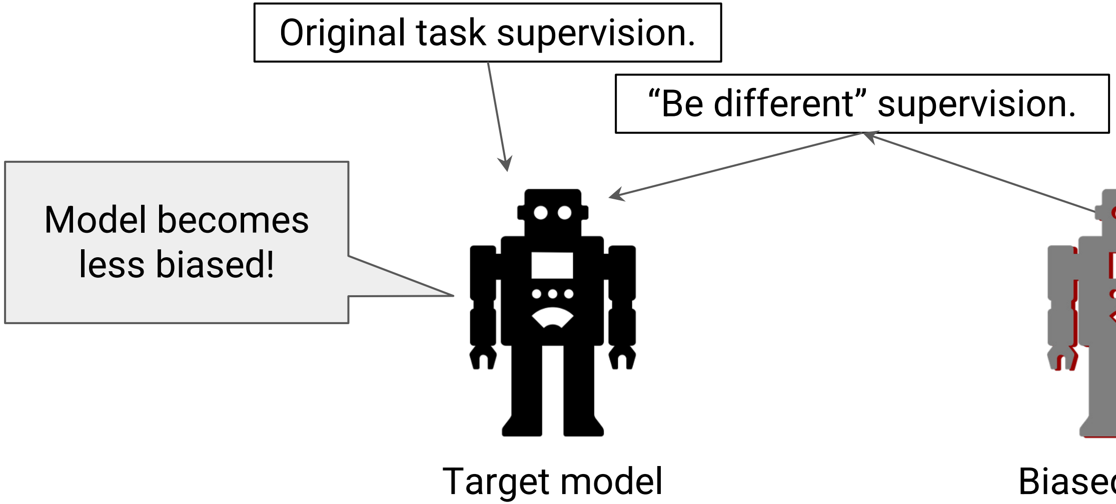
Solution: De-biasing with biased models.

- If we can make a biased model, we can avoid shortcut learning problem by utilizing the biased model.



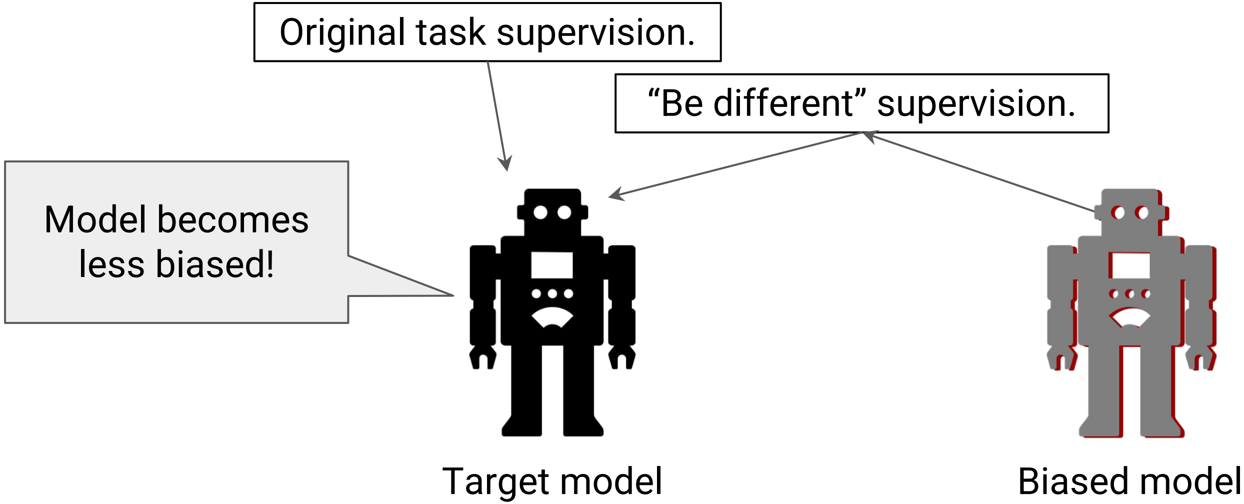
Two questions.

1. How to make a biased model?
2. How to encode “be different” ?



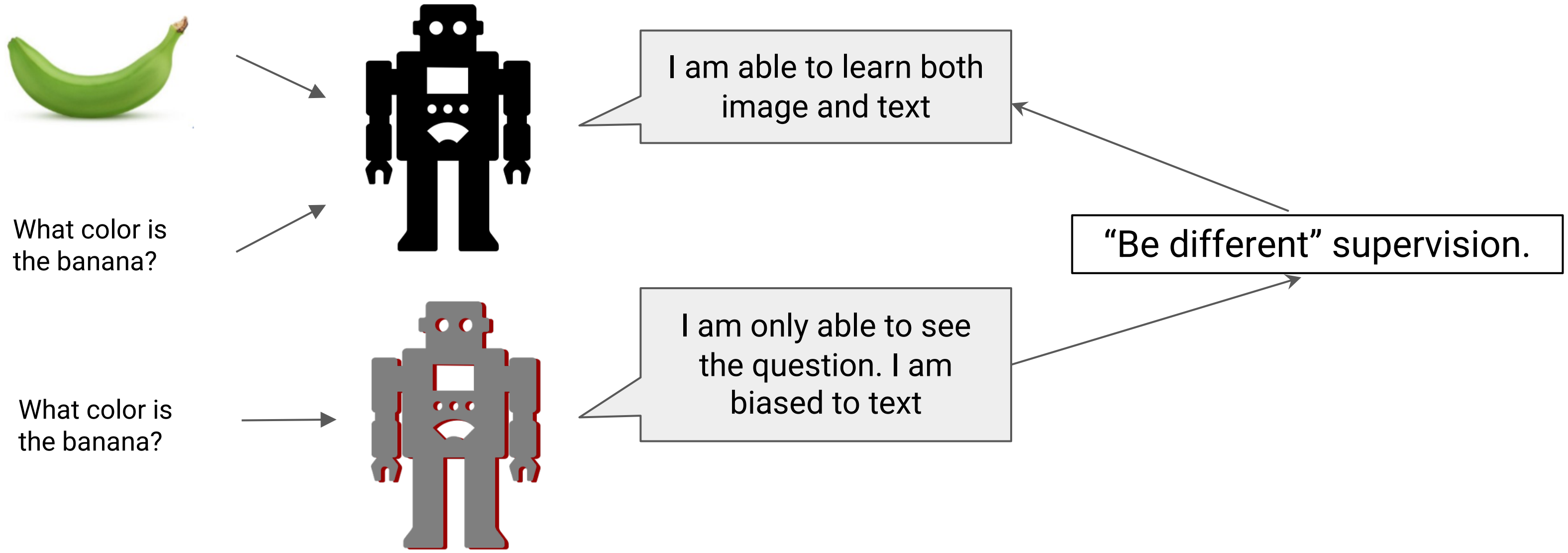
Two questions.

1. How to make a biased model?
2. How to encode “be different” ? (out-of-context for today’s talk)



If we know how to make a perfectly biased model

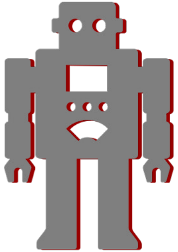
- For certain types of biases, we can make a perfectly biased model.
- Example 1: VQA (Visual-question answering)



Cadene, et al. "RUBi: Reducing Unimodal Biases for Visual Question Answering", NeurIPS 2019
Clark, et al. "Don't take the easy way out: ensemble based methods for avoiding known dataset biases", EMNLP 2019
Clark, et al. "Learning to Model and Ignore Dataset Bias with Mixed Capacity Ensembles", EMNLP findings 2020

If we know how to make a perfectly biased model

- For certain types of biases, we can make a perfectly biased model.
- Example 2: Image textures / colors



I am a specifically designed model to capture image textures

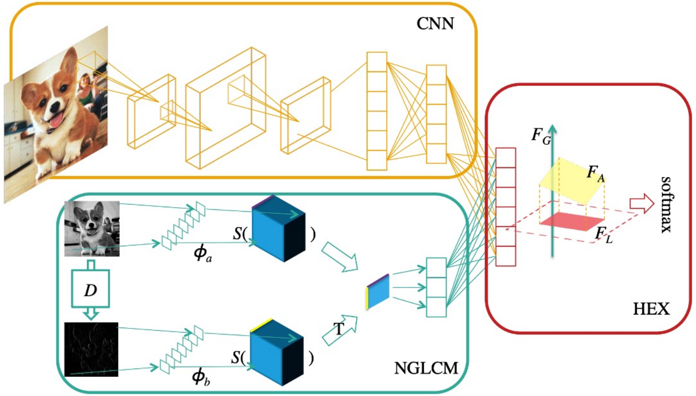
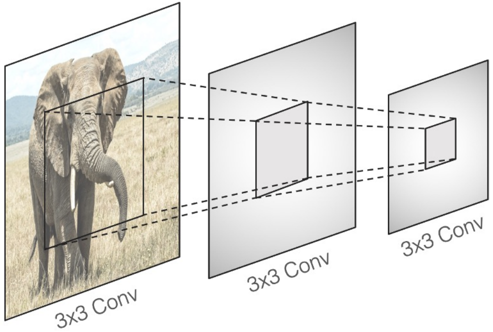


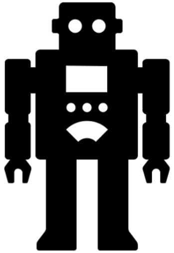
Figure 2: Introduction of Neural Gray-level Co-occurrence Matrix (NGLCM) and HEX.

Wang, et al. “Learning Robust Representations by Projecting Superficial Statistics Out” ICLR 2019

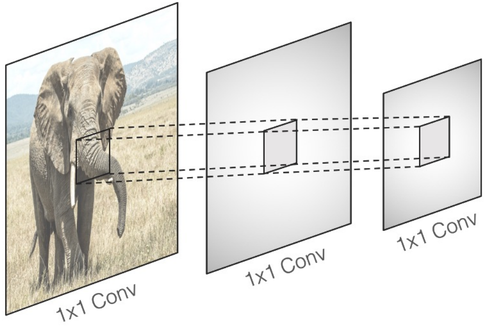
If we know when a model is biased

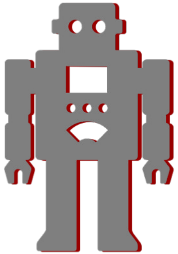
- If we know that a biased model focuses on “local areas” than “global areas”





I am able to see very wide and global information from images



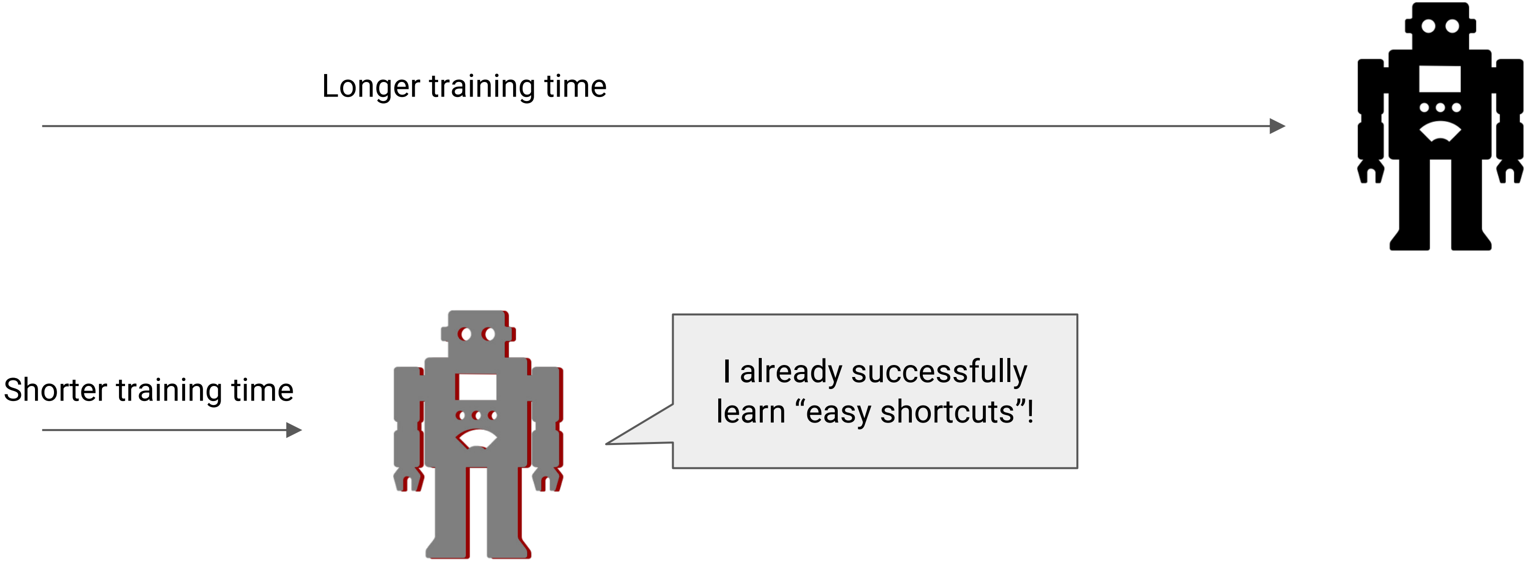


I am only able to see very small local areas,
I will be biased

[Bahng, et al. “Learning De-biased Representations with Biased Representations”, ICML 2020]

If we know when a model is biased

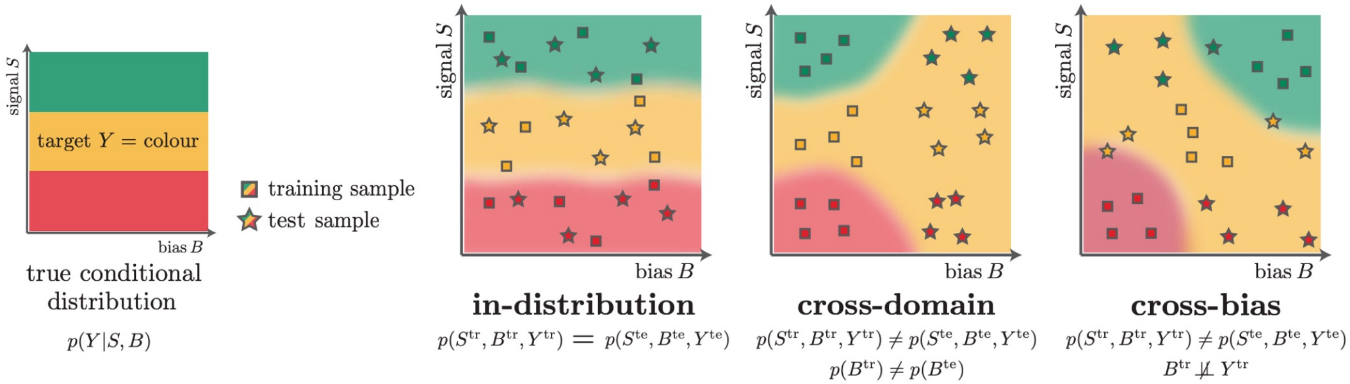
- If we know that "shortcut biases" are *easy-to-learn* compared to the "desired features"



[Nam, et al. "Learning from failure: Training debiased classifier from biased classifier" NeurIPS 2020]
[Liu, et al. "Just train twice: Improving group robustness without training group information" ICML 2021]
[Lee, et al. "Learning debiased representation via disentangled feature augmentation" NeurIPS 2021]

How can we evaluate if a model successfully avoids SL?

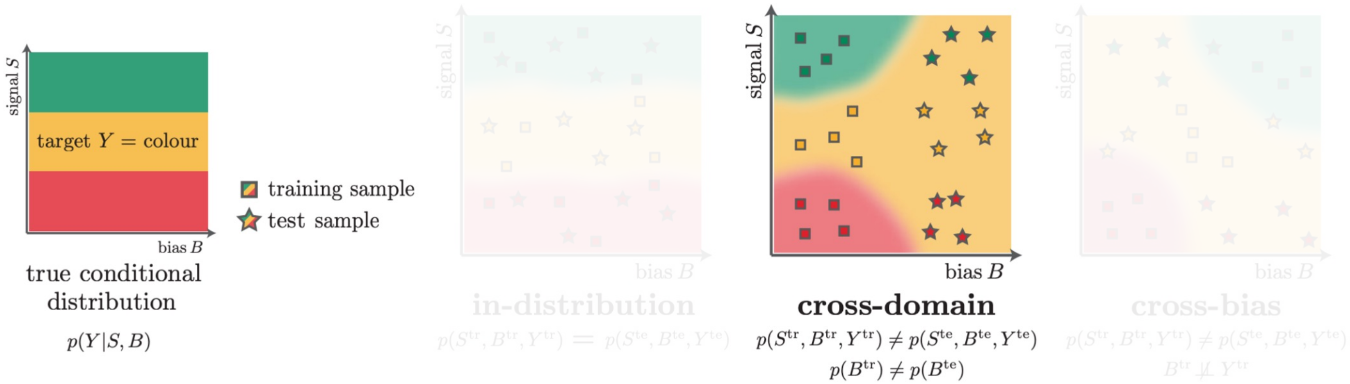
- It depends on how we make benchmark datasets (i.e., training and test sets)
- If we want to check whether a model performs well under significant distribution shifts (e.g., every car and boat is in the sky), then we have to test **“domain generalization”** benchmark.
  - These type of test sets are often called as “out-of-distributed” datasets.
- If we want to check whether a model can perfectly ignore shortcut biases even under the severely biased training dataset, **“cross-bias generalization”** benchmark would be helpful.



Bahng, et al. “Learning De-biased Representations with Biased Representations”, ICML 2020

Domain generalization benchmark.

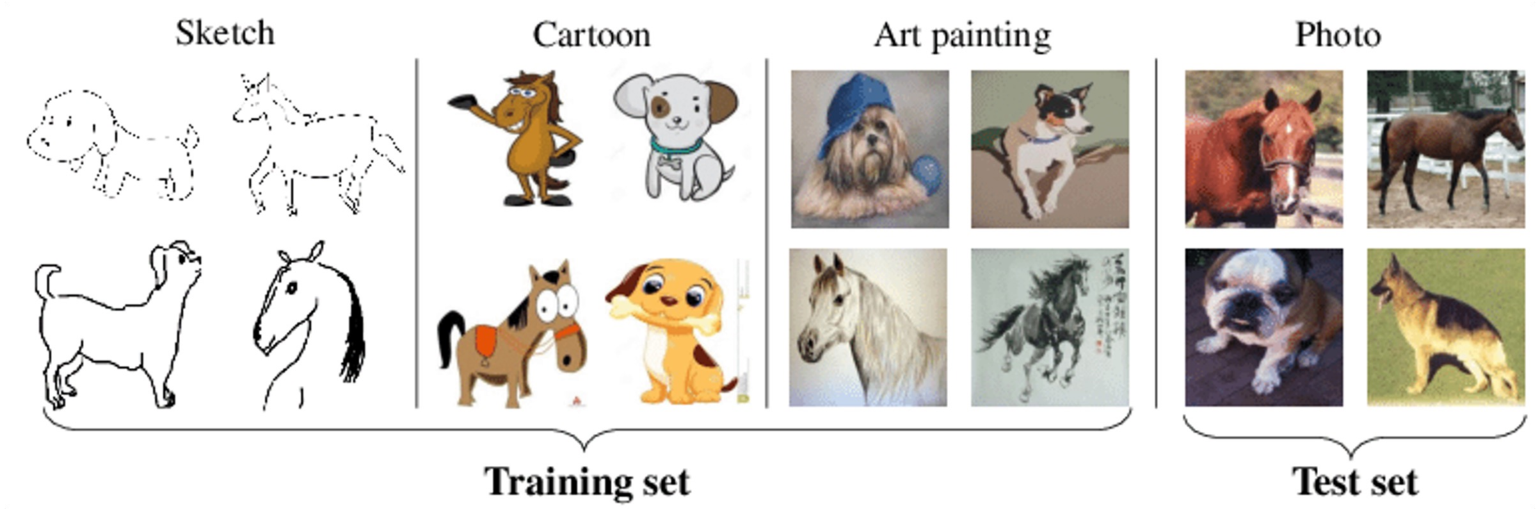
- It depends on how we make benchmark datasets (i.e., training and test sets)
- If we want to check whether a model performs well under significant distribution shifts (e.g., every car and boat is in the sky), then we have to test **"domain generalization"** benchmark.
    - These type of test sets are often called as "out-of-distributed" datasets.
- If we want to check whether a model can perfectly ignore shortcut biases even under the severely biased training dataset, **"cross-bias generalization"** benchmark would be helpful.



Bahng, et al. "Learning De-biased Representations with Biased Representations", ICML 2020

Domain generalization benchmark.

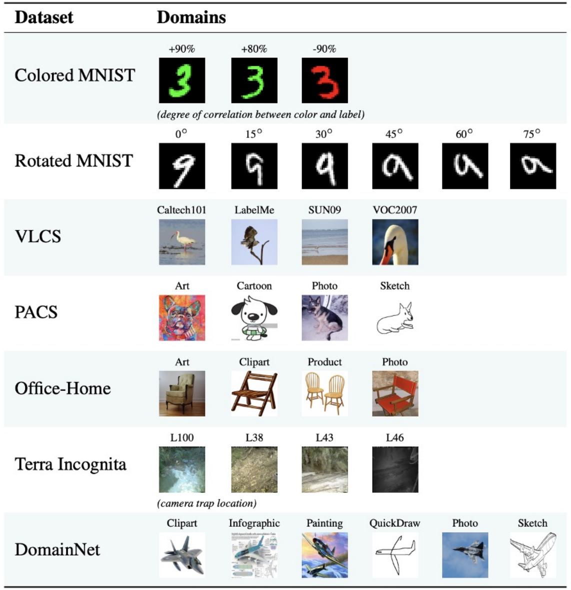
- “Domain generalization” aims to solve the generalization problem under the unseen distribution shifts.



https://www.researchgate.net/figure/Examples-from-the-dataset-PACS-1-for-domain-generalization-The-training-set-is_fig1_349787277

Domain generalization benchmark.

- When we evaluate algorithms on multiple benchmarks under a fair evaluation protocol, theoretically well-founded methods often perform worse than the baseline (ERM)

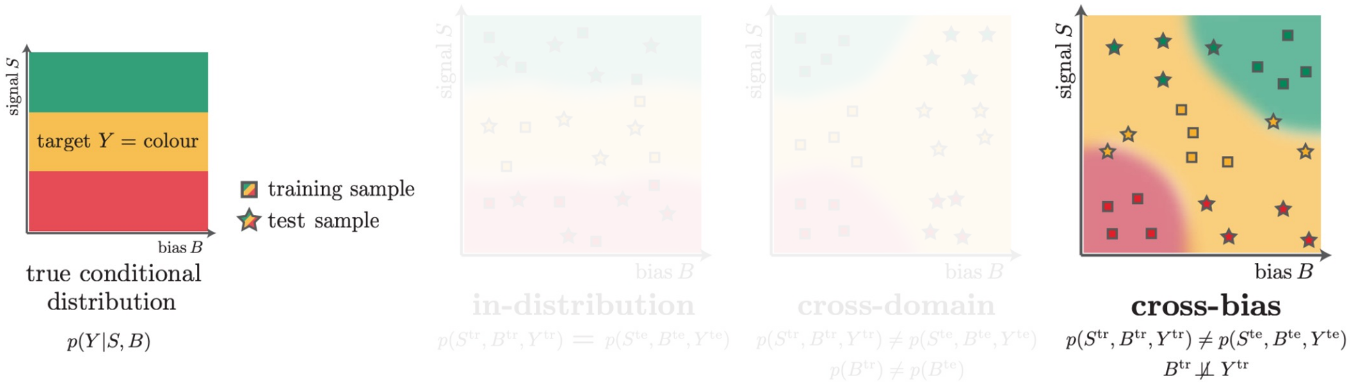


Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
MMD† [12]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
Mixstyle‡ [28]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
GroupDRO† [24]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
IRM† [22]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
ARM† [21]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
VREx† [23]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
CDANN† [15]	82.6±0.9	77.5±0.1	65.8±1.3	45.8±1.6	38.3±0.3	62.0
DANN† [11]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
RSC† [52]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
MTL† [53]	84.6±0.5	77.2±0.4	66.4±0.5	45.6±1.2	40.6±0.1	62.9
Mixup† [54–56]	84.6±0.6	77.4±0.6	68.1±0.3	47.9±0.8	39.2±0.1	63.4
MLDG† [17]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
Fish [25]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
ERM‡ [57]	84.2±0.1	77.3±0.1	67.6±0.2	47.8±0.6	44.0±0.1	64.2
SagNet† [29]	**86.3**±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
SelfReg [58]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	42.8±0.0	64.2
CORAL† [13]	86.2±0.3	78.8±0.6	68.7±0.3	47.6±1.0	41.5±0.1	64.5
mDSDI [59]	86.2±0.2	**79.0**±0.3	69.2±0.4	48.1±1.4	42.8±0.1	65.1
MIRO	85.4±0.4	**79.0**±0.0	**70.5**±0.4	**50.4**±1.1	**44.3**±0.2	**65.9**
Combined with SWAD [34]						
ERM + SWAD‡	88.1±0.1	79.1±0.1	70.6±0.2	50.0±0.3	46.5±0.1	66.9
CORAL + SWAD‡	88.3±0.1	78.9±0.1	71.3±0.1	51.0±0.1	46.8±0.0	67.3
MIRO + SWAD	**88.4**±0.1	**79.6**±0.2	**72.4**±0.1	**52.9**±0.2	**47.0**±0.0	**68.1**

Gulrajani and Lopez-Paz. "In search of lost domain generalization" ICLR 2021.
Cha, et al. “SWAD: Domain Generalization by Seeking Flat Minima.” NeurIPS 2021.
Cha, et al. “Domain Generalization by Mutual-Information Regularization with Pre-trained Models” ECCV 2022.

Cross-bias generalization benchmark.

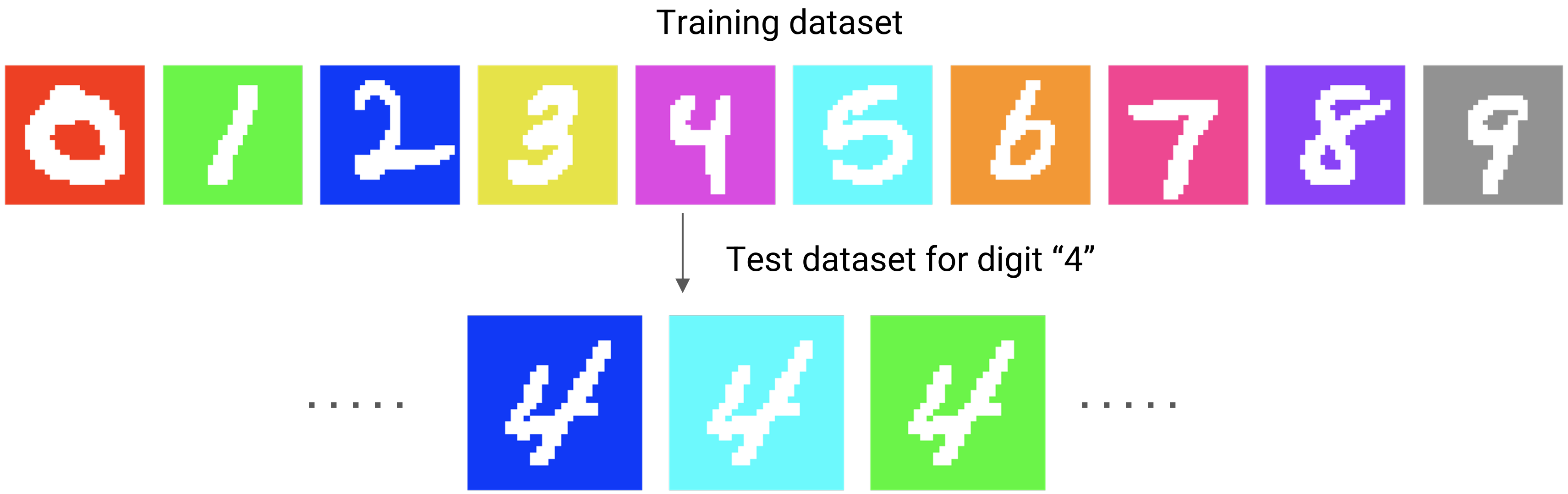
- It depends on how we make benchmark datasets (i.e., training and test sets)
- If we want to check whether a model performs well under significant distribution shifts (e.g., every car and boat is in the sky), then we have to test "**domain generalization**" benchmark.
    - These type of test sets are often called as "out-of-distributed" datasets.
- If we want to check whether a model can perfectly ignore shortcut biases even under the severely biased training dataset, **"cross-bias generalization"** benchmark would be helpful.



Bahng, et al. "Learning De-biased Representations with Biased Representations", ICML 2020

Cross-bias generalization benchmark.

- When bias labels and target labels are strongly correlated.
- For example, when background color (bias label) is strongly correlated to digit (target label).
- We aim to train a model without looking at background, but digit itself.



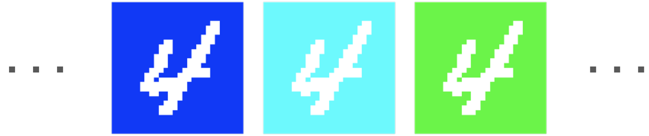
Bahng, et al. "Learning De-biased Representations with Biased Representations", ICML 2020

Cross-bias generalization vs. domain generalization

- Cross-bias generalization focuses on when “bias and label are correlated”
- Domain generalization focuses on when “a new bias appears”

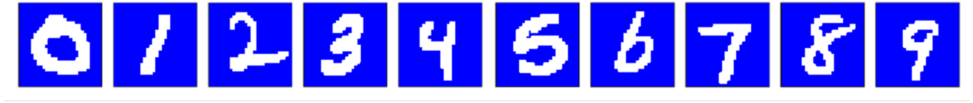
Cross-bias

Test dataset for digit “4”



Cross-domain

Test for unseen “domain”

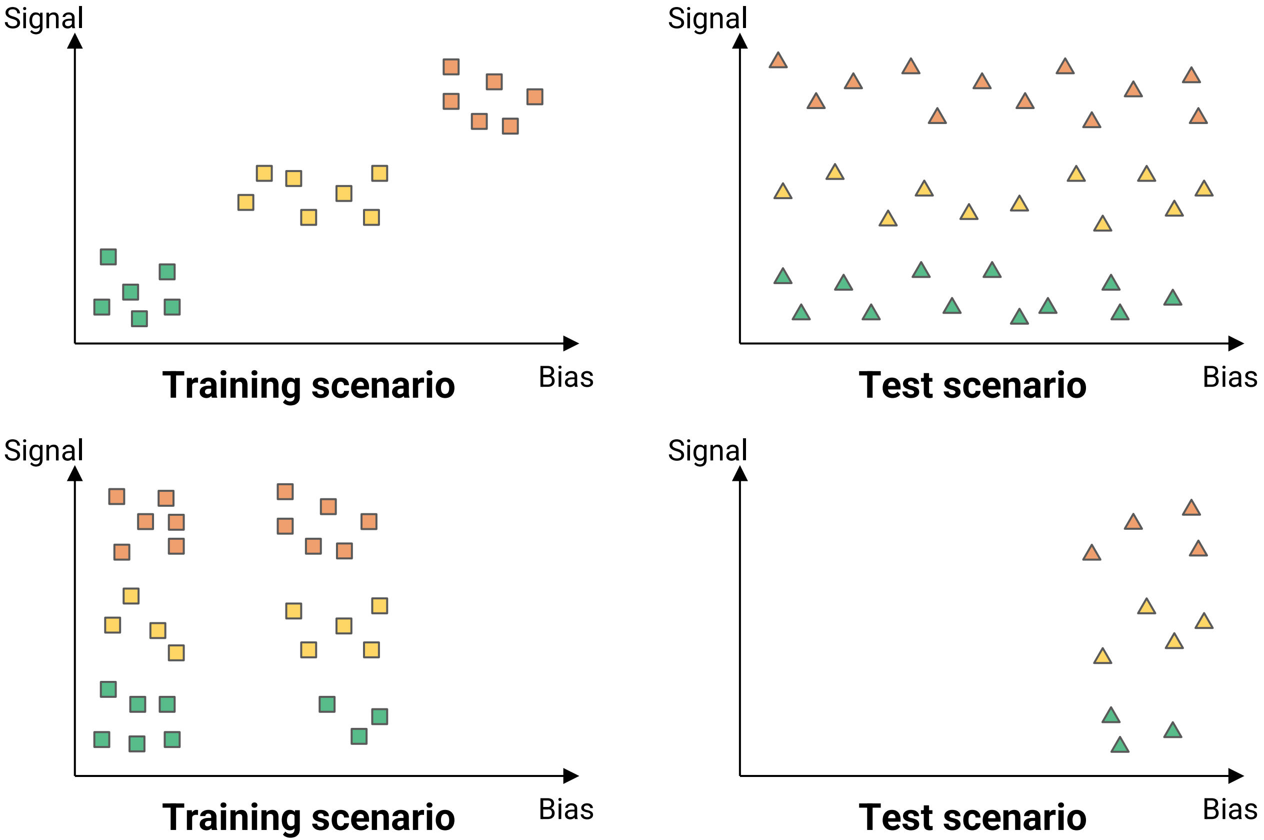


Bahng, et al. “Learning De-biased Representations with Biased Representations”, ICML 2020

Cross-bias generalization vs. domain generalization

Cross-bias generalization

Domain generalization

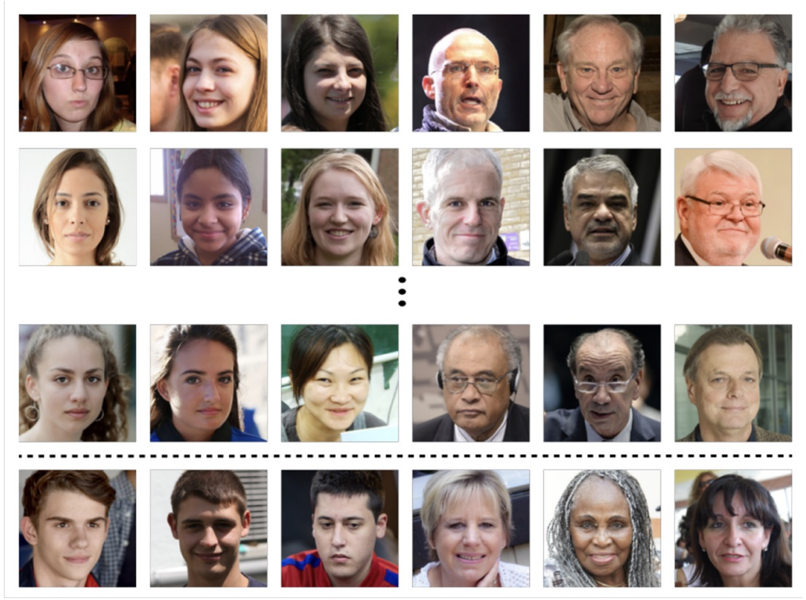


Cross-bias generalization benchmark.

- When bias labels and target labels are strongly correlated.
- Gender (bias label) is strongly correlated to age (target label).

Training dataset

Test dataset



Lee, et al. “Learning debiased representation via disentangled feature augmentation” NeurIPS 2021

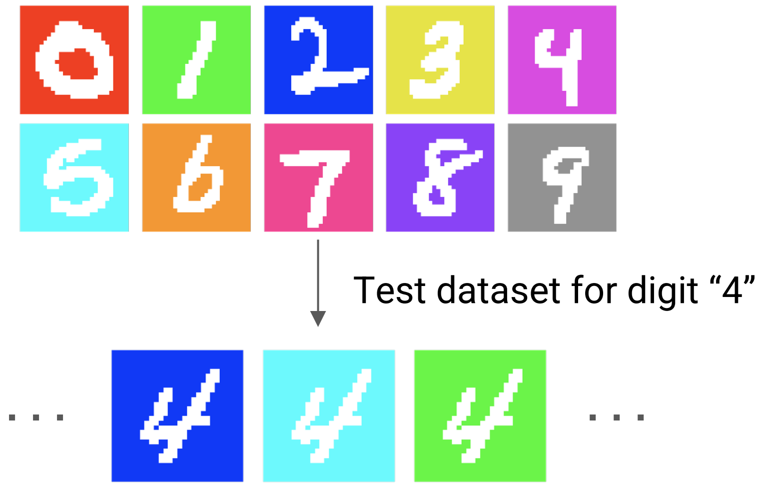
How can we evaluate if a model successfully avoids SL?

- Unfortunately, collecting "perfectly de-biased dataset" is extremely difficult, and the existing benchmarks are relatively "toy-setting" compared to real-world applications.
- We have to pay attention more to build the de-biased dataset to evaluate if a model avoids SL.

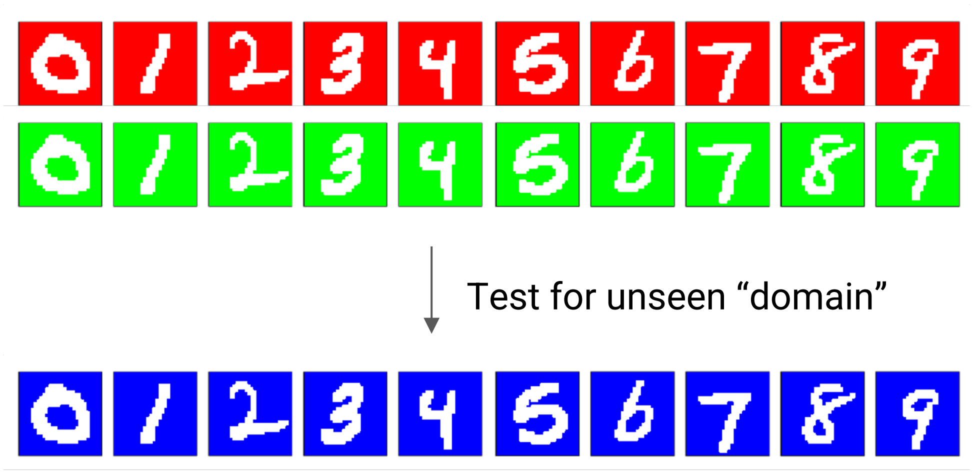
A more realistic evaluation benchmark is still an open question.

- It is important to design a more “realistic” and “practical” evaluation benchmark.
- But, it should be controllable and reproducible as well.

Cross-bias



Cross-domain



[Bahng, et al. “Learning De-biased Representations with Biased Representations”, ICML 2020]

How can we make “De-biasing” methods work in practice?

Model	Clean	Unbiased Acc [1]	ImageNet-C [14]	ImageNet-A [16]	Occlusion
Vanilla (ResNet-18 [13])†	90.8	88.8	54.2	24.9	71.3
Biased (BagNet-18 [2])†	67.7	65.9	31.7	18.8	59.7
LearnedMixin + H [6]†	64.1	62.7	27.5	15.0	33.5
RUBi [3]†	90.5	88.6	53.7	27.7	71.3
ReBias [1]†	91.9	90.5	57.5	29.6	73.4
LfF [29]	93.2	92.0	57.8	28.1	77.0
CutMix [40]	<u>93.8</u>	91.8	54.6	27.1	**83.1**
Mixup [42]	93.2	91.4	61.5	**33.4**	<u>77.9</u>
Stylized ImageNet [11]†	88.4	86.6	61.1	24.6	64.4
StyleAugment	<u>93.8</u>	<u>92.6</u>	<u>65.3</u>	29.6	73.0
StyleAugment + AdamP [17]	**95.9**	**94.8**	**72.5**	<u>32.1</u>	75.8

Table 1: **Comparison of state-of-the-art de-biasing and augmentation methods on the ImageNet-9 validation dataset.** We measure the ImageNet-9 top-1 validation accuracy (Clean), the unbiase accuracy using texture clustering (Unbiased Acc) following Bahng *et al.* [1], ImageNet-C top-1 accuracy, ImageNet-A top-1 accuracy, and the top-1 accuracy on occluded samples. The first and the second best methods are denoted in **bold numbers** and <u>underlined numbers</u>. The rows with † denotes the same weights from Bahng *et al.* [1].

Chun, et al. “StyleAugment: Learning Texture De-biased Representations by Style Augmentation without Pre-defined Textures”, ArXiv preprint.

Conclusion of Part 2

- Bias labels are not fully accessible because of the annotation costs, or the ambiguous nature of biases
- If we only have partially annotated bias labels, then CGL can be helpful
- If we know when model is biased, then “de-biasing from biased model” framework (e.g., ReBias) can work well
- There are many possible training-evaluation scenarios for shortcut learning problem (cross-bias generalization, domain generalization, …)
- We have to study this problem in various viewpoints!

Concluding remark and future work

- **Data collection** should be considered and designed to ***avoid undesirable dataset biases*** in the dataset (e.g., “lung pics and hospital tokens”)
- We may need **additional annotation process** to collect ***“bias labels of concern”*** (e.g., ethical labels such as ethnicity, gender or income level)
- **Algorithm** should reflect the task itself and should preferably be ***unbiased*** sometimes we need a different decision process, e.g., stochastic one
- **Evaluation protocol** should consider the ***real-world deployment scenario*** such as distribution shifts, not only “in-distributed scenario”.
- Human should be considered in the machine learning model development pipeline loop, i.e., ***human-in-the-loop*** is required

Thanks!

Acknowledgements:
This talk is inspired by “Unreasonably shallow deep learning” provided by professor Kyunghyun Cho. and awesome research project “Shortcut Learning in Deep Neural Networks” by Geirhos, et al. Many slides are borrowed from my previous talks.