Understanding the failure modes of out-of-distribution generalization.

vaishnavh@

Collaborators:
ajandreassen@, neyshabur@, Thao Nguyen & hsedghi@

Spurious correlations

Models tend to rely on all features that are correlated with label during training.

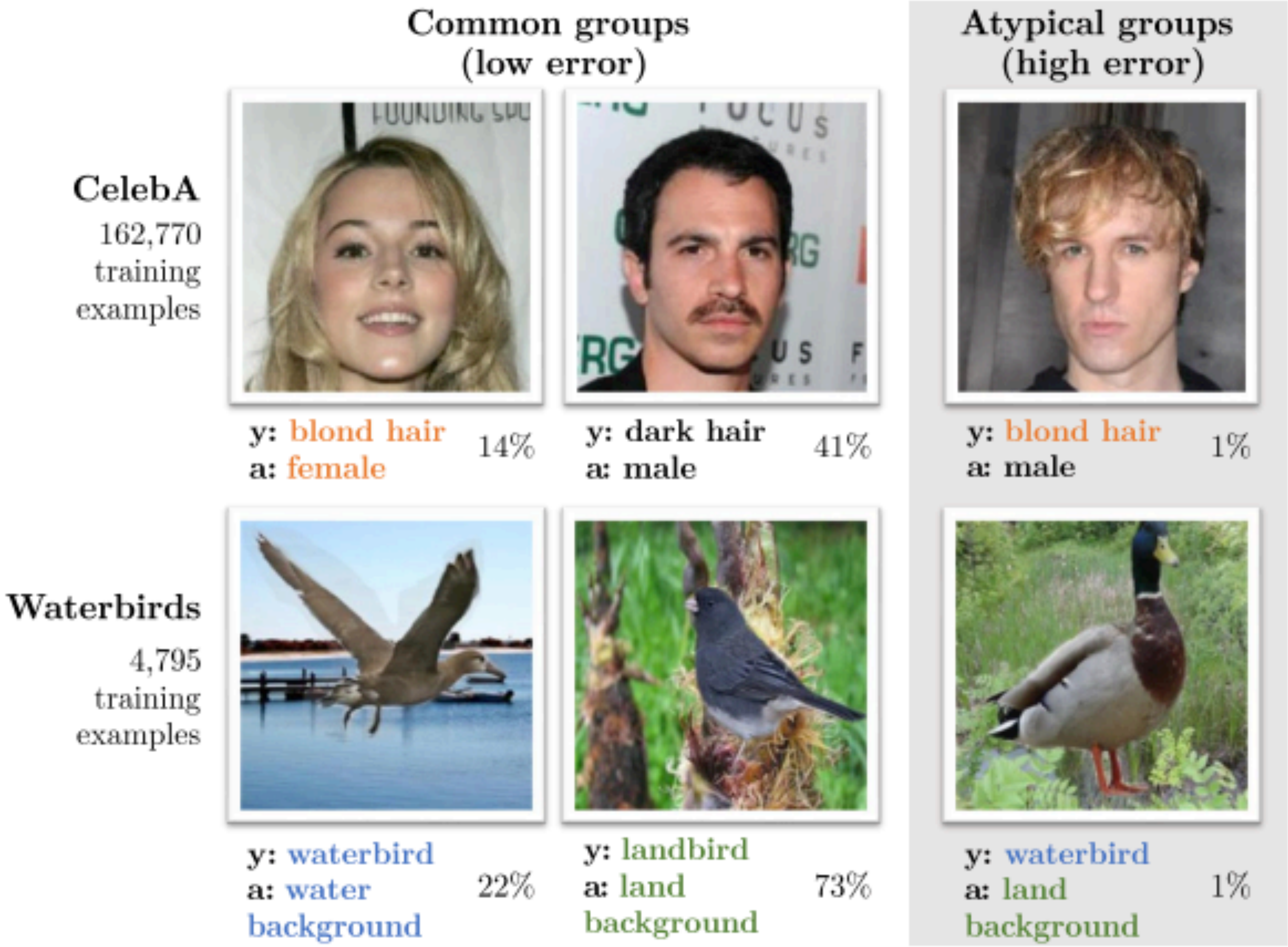
(a) Husky classified as wolf (b) Explanation

Ribeiro,Singh,Guestrin '16

Song,Jiang,Tu,Du,Neyshabur '19

Spurious correlations

Models tend to rely on all features that are correlated with label during training.



[Sagawa,Koh,Hashimoto,Liang'20]

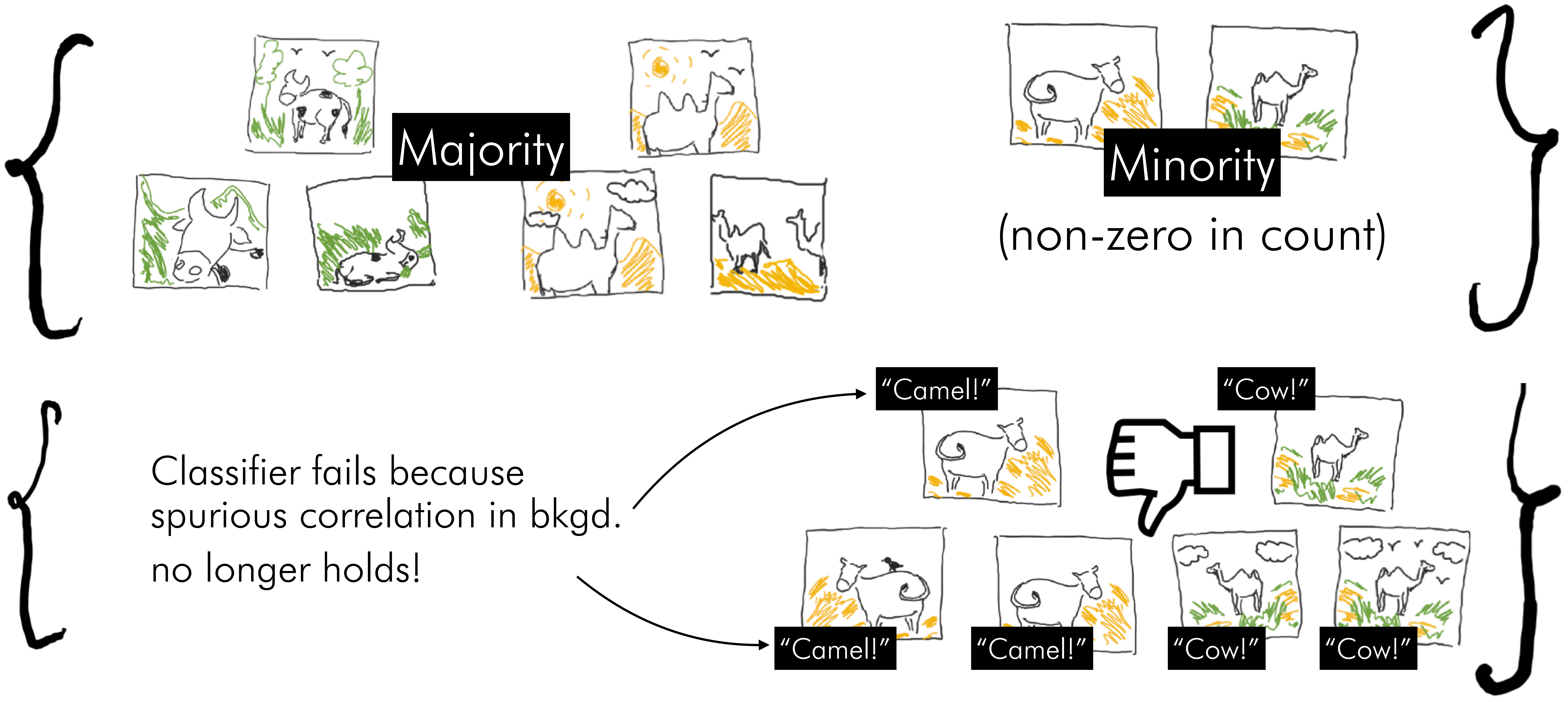
Spurious correlations: Illustration

Cow/camel classification

[Arjovsky, Bottou, Gulrajani, Lopez-Paz '19 Beery, Horn, Perona '18]

Training domain

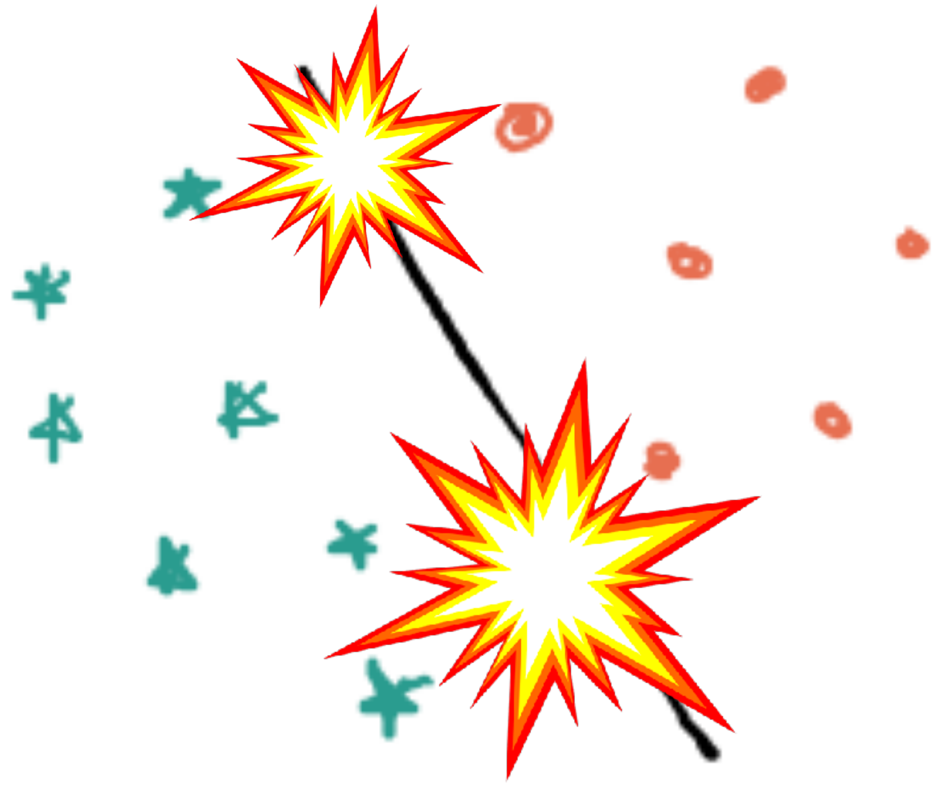
Test domain



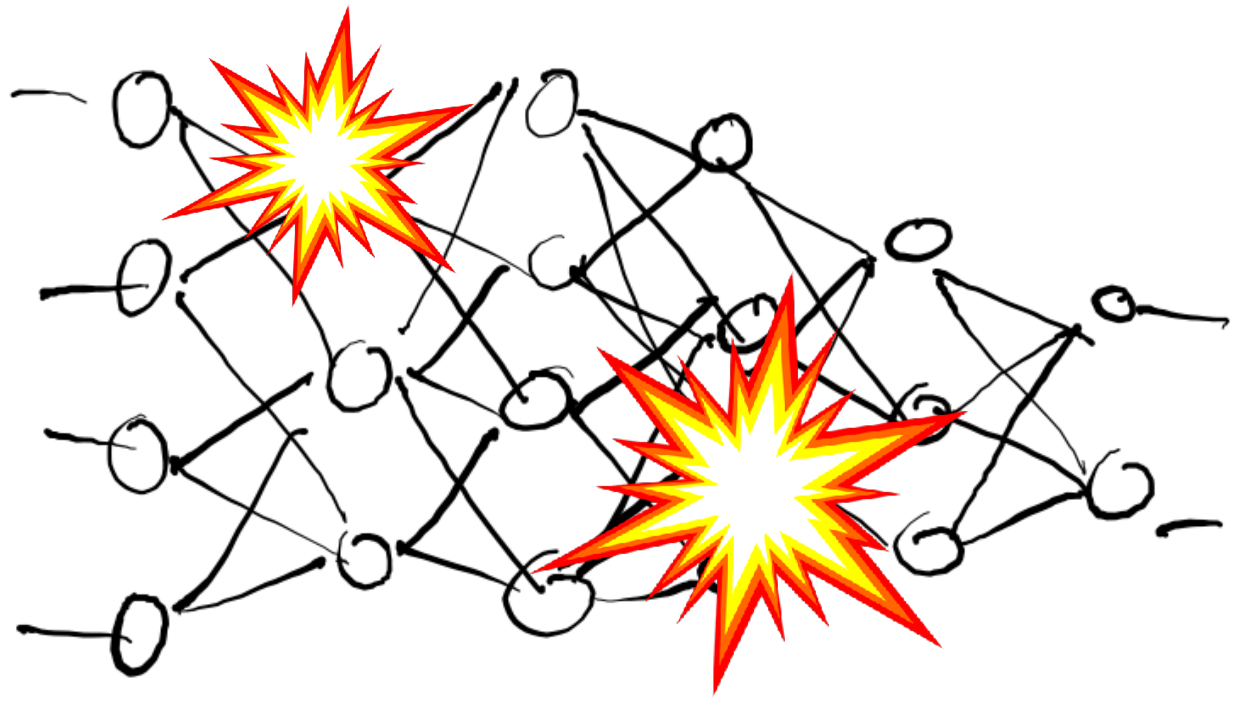
Fundamental question: Why do classifiers rely on spurious correlations?

Our work: Why do classifiers rely on spurious correlations?

1. Existing theoretical frameworks do not capture fundamental ways by which models end up using spurious correlation.



2. We theoretically study GD+linear classifiers and discover two fundamental mechanisms by which spurious-feature-reliance comes about.

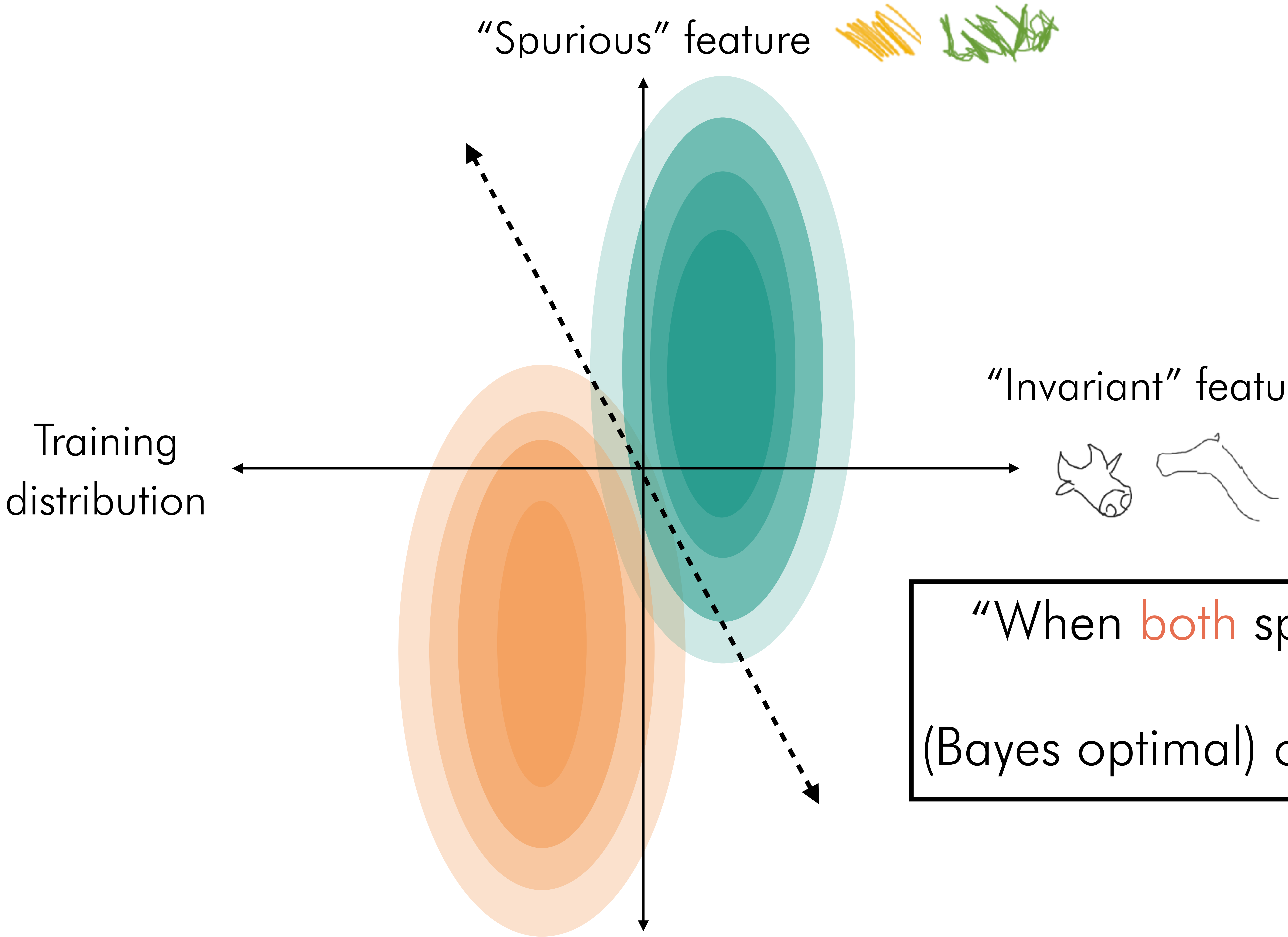


3. We discuss practical algorithmic implications of these failure modes.

Outline

- Introduction
- Motivation: existing theoretical models are inadequate
- Failure mode 1
- Failure mode 2
- Takeaways
- Conclusion

The de facto theoretical framework for spurious correlations



"When both spurious and invariant features are partially predictive,
(Bayes optimal) classifier relies on spurious feature."

The de facto theoretical framework for spurious correlations

This framework forms the basis of a lot of research in this area

Algorithms for OoD generalization

IRM (ARGL'19)
IRM-CDM (GZLK'21)

Synthetic datasets

ColoredMNIST (ARGL'19)
CMNIST+ (GZLK'21)
Linear Unit Tests (ASABL'21)

Theoretical guarantees

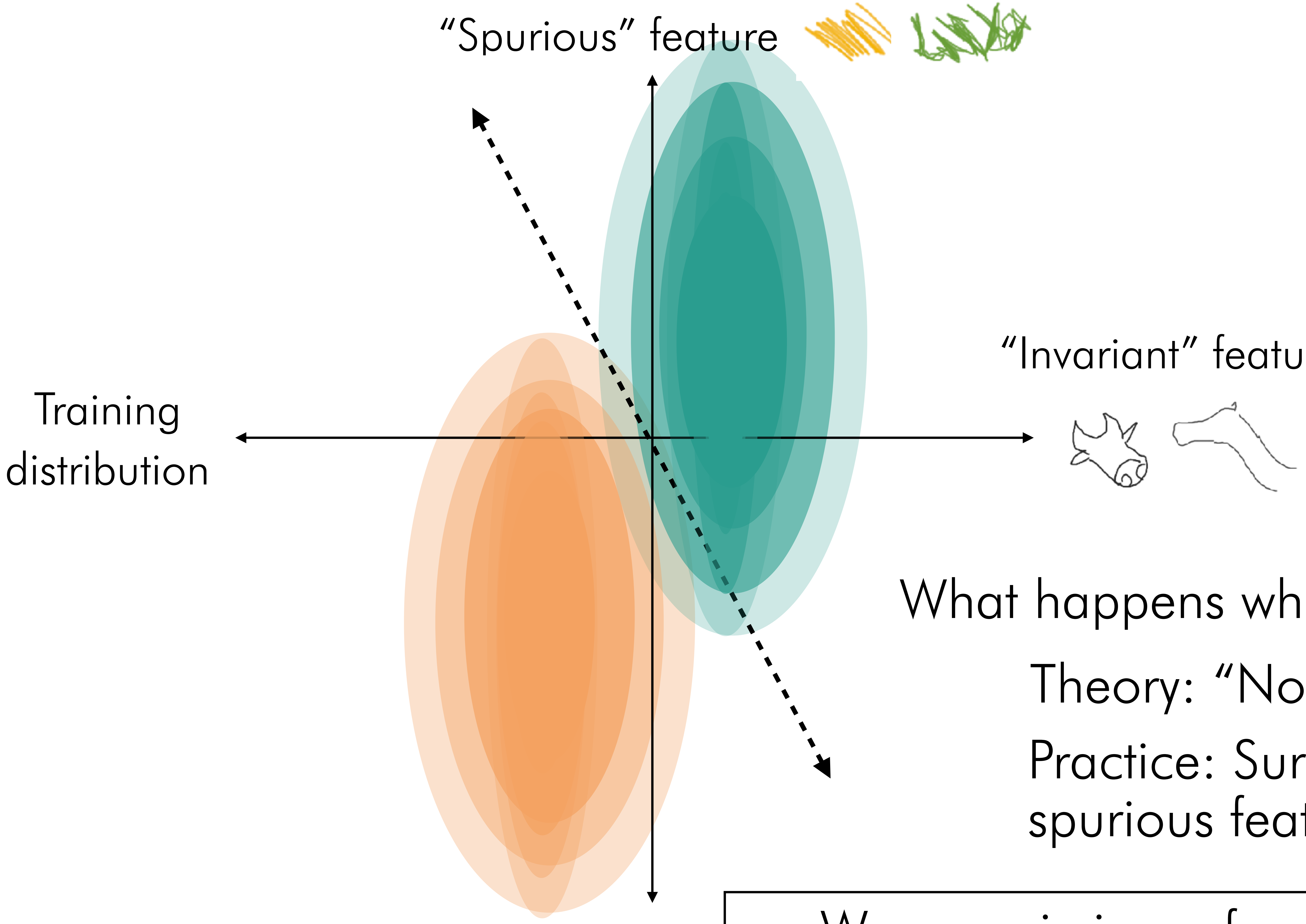
Rosenfeld, Ravikumar & Risteski '20,
Kamath, Tangella, Sutherland & Srebro '21.

Reasoning about spurious correlations

Sagawa, Raghunathan, Koh & Liang '20,
Tsipras, Santurkar, Engstrom, Turner & Madry '19

Hence it is critical to ask: does this framework capture the fundamental reasons behind failure?

Our work: Does this de facto framework explain failure in practice?



What happens when inv. feature is fully predictive?

Theory: “No more spurious-feature-reliance!”

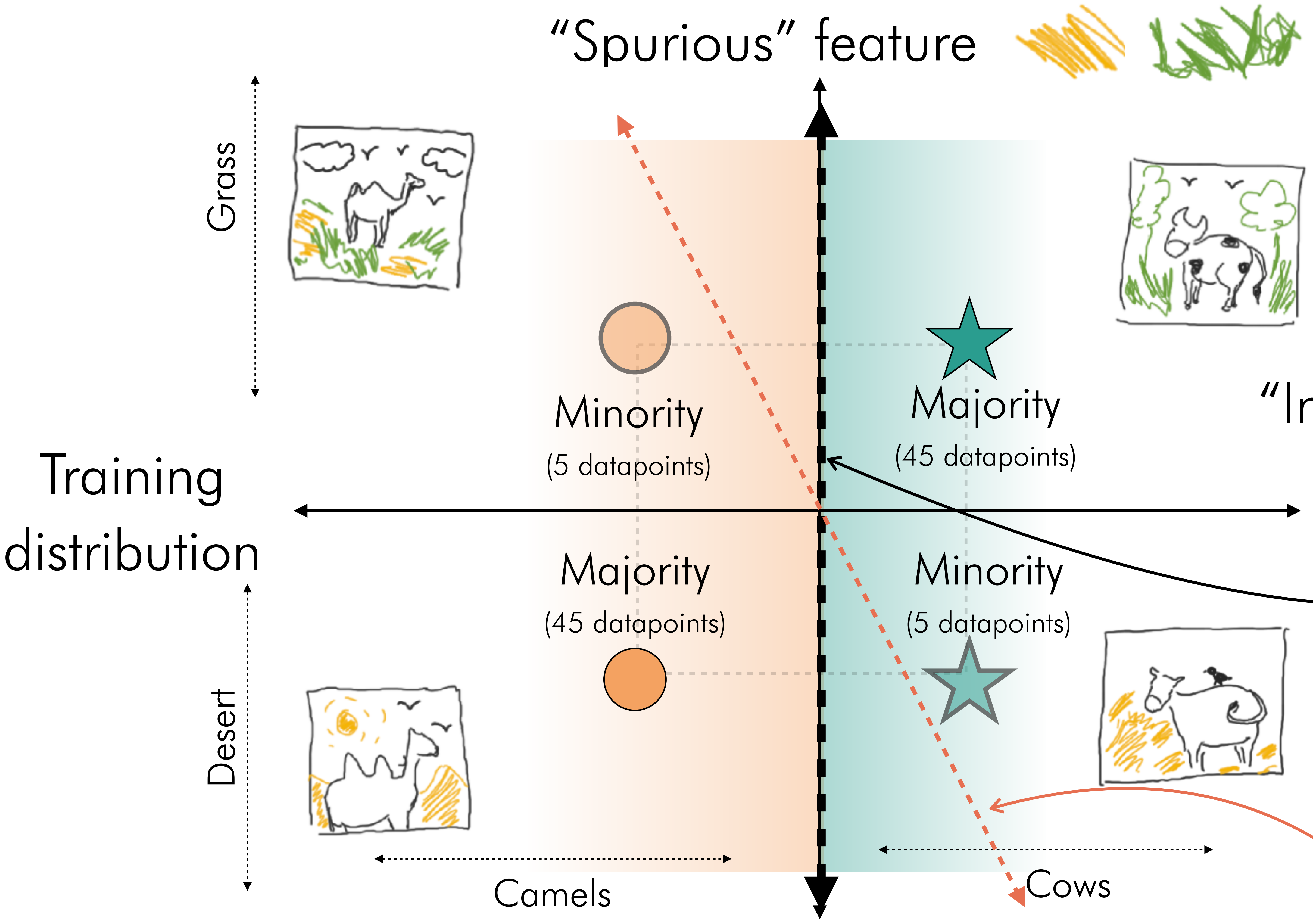
Practice: Surprisingly, deep networks still use spurious feature!

We are missing a fundamental piece of the story!

Outline

- Introduction
- Motivation
- Our work: a study of GD + linear classifier
  - Failure mode 1: statistical
  - Failure mode 2
- Takeaways
- Conclusion

Source of failure 1: Statistical



only partially predictive

fully predictive

"Invariant" feature

Max-margin classifier doesn't use sp. feature
(as long as we've non-zero minority points)

Finite-time-GD uses sp. feature even though it should converge to max-margin! [SHNGS'17, JT'18].

Informal version of our result: For a large class of linearly separable datasets, under logistic loss,

$$\frac{|w_{sp}(t)|}{\|\overrightarrow{w}_{inv}(t)\|} = \Theta\left(\frac{\text{level of spurious correlation}}{\log t}\right)$$

Source of failure 1: Statistical

Insight from denominator: GD takes exponentially long to make $w_{sp} \to 0$.

Builds on the distribution-independent $O\left(1/\log t\right)$ bound [SHNGS'17, JT'18].

Insight from numerator: Distribution-dependent dynamics s.t. greater spurious correlation $\Longrightarrow$ greater reliance on spurious feature.

Takeaway: Spurious-feature-reliance happens due to finite-time GD bias namely, "use every statistical correlation".

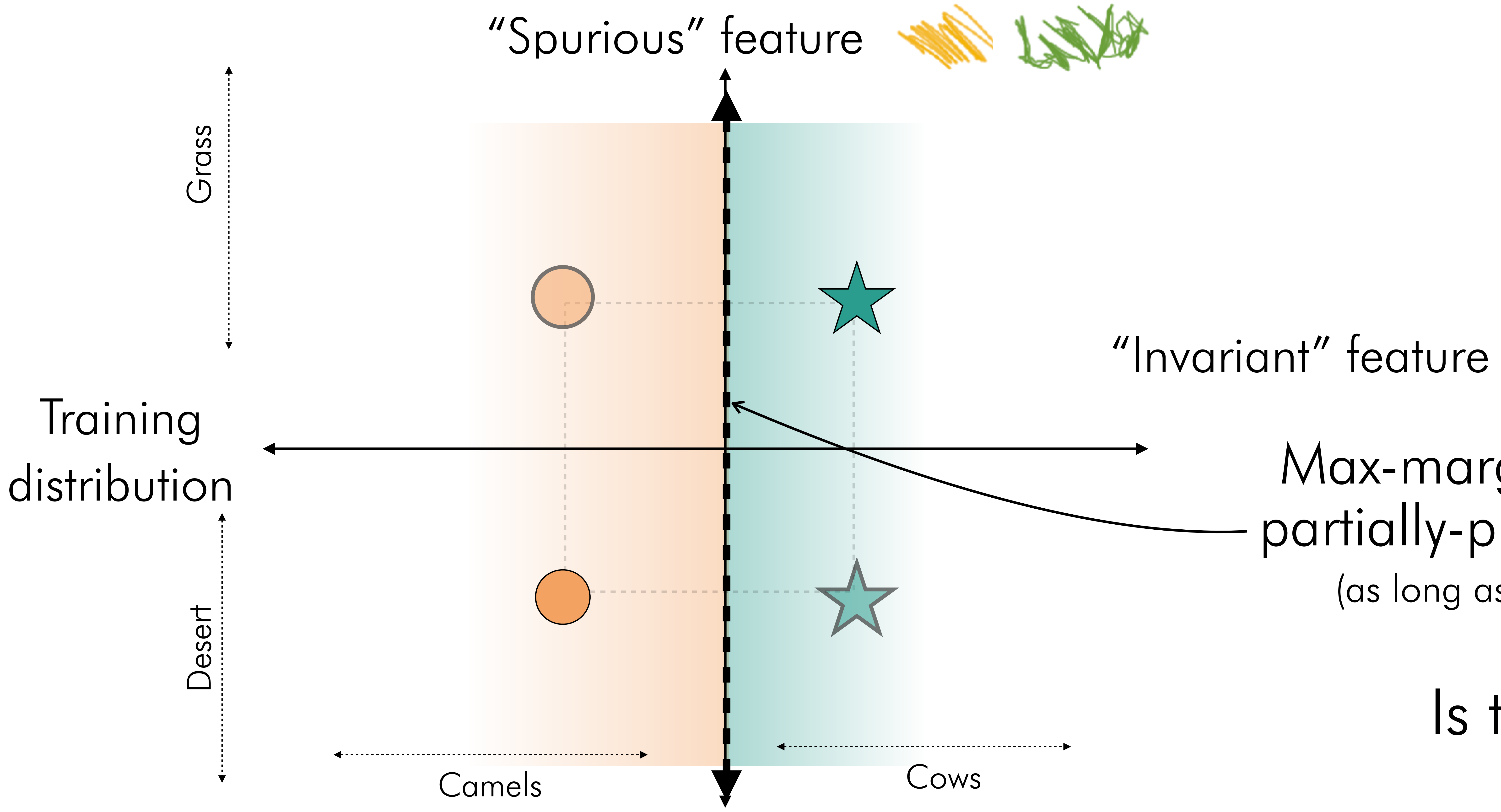
Informal version of our result: For a large class of linearly separable datasets, under logistic loss,

$$\frac{|w_{sp}(t)|}{\|\overrightarrow{w}_{inv}(t)\|} = \Theta\left(\frac{\text{level of spurious correlation}}{\log t}\right)$$

Outline

- Introduction
- Motivation: existing theoretical models are inadequate
- Failure mode 1: statistical
- Failure mode 2: geometric
- Takeaways
- Conclusion

Source of failure 2: Geometric



Max-margin classifier doesn't use partially-predictive spurious feature
(as long as we've non-zero minority points)

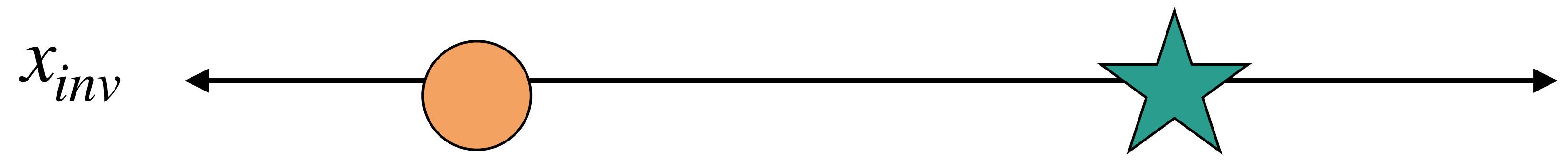
Is this always true?

No! We show that when data has non-degenerate geometry, even max-margin classifier can use partially-predictive spurious feature!

Source of failure 2: Geometric

Key property of real-world data geometry:

Previous toy example:

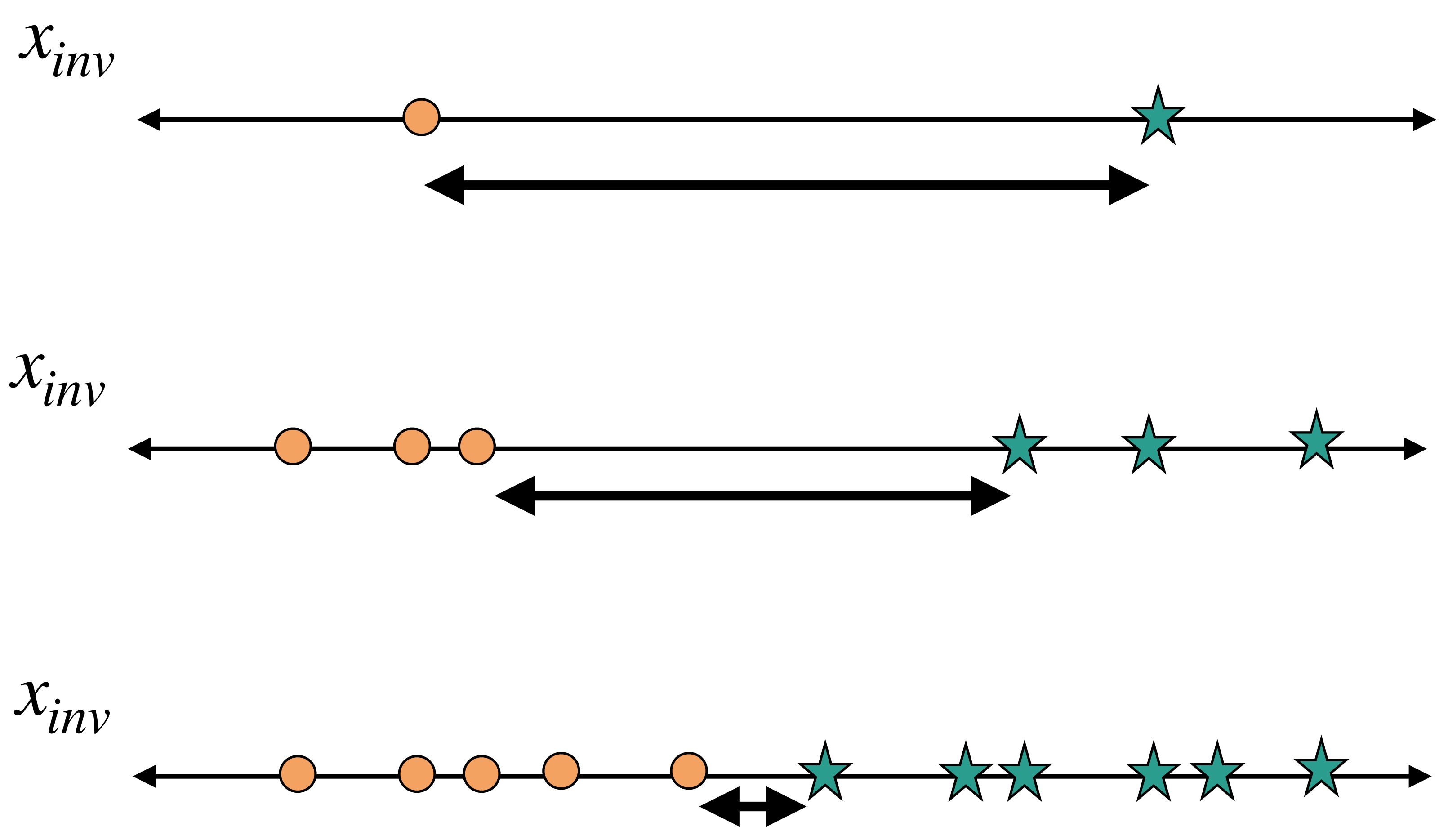


But, real-world looks more like:

and hence

as we sample more and more data,

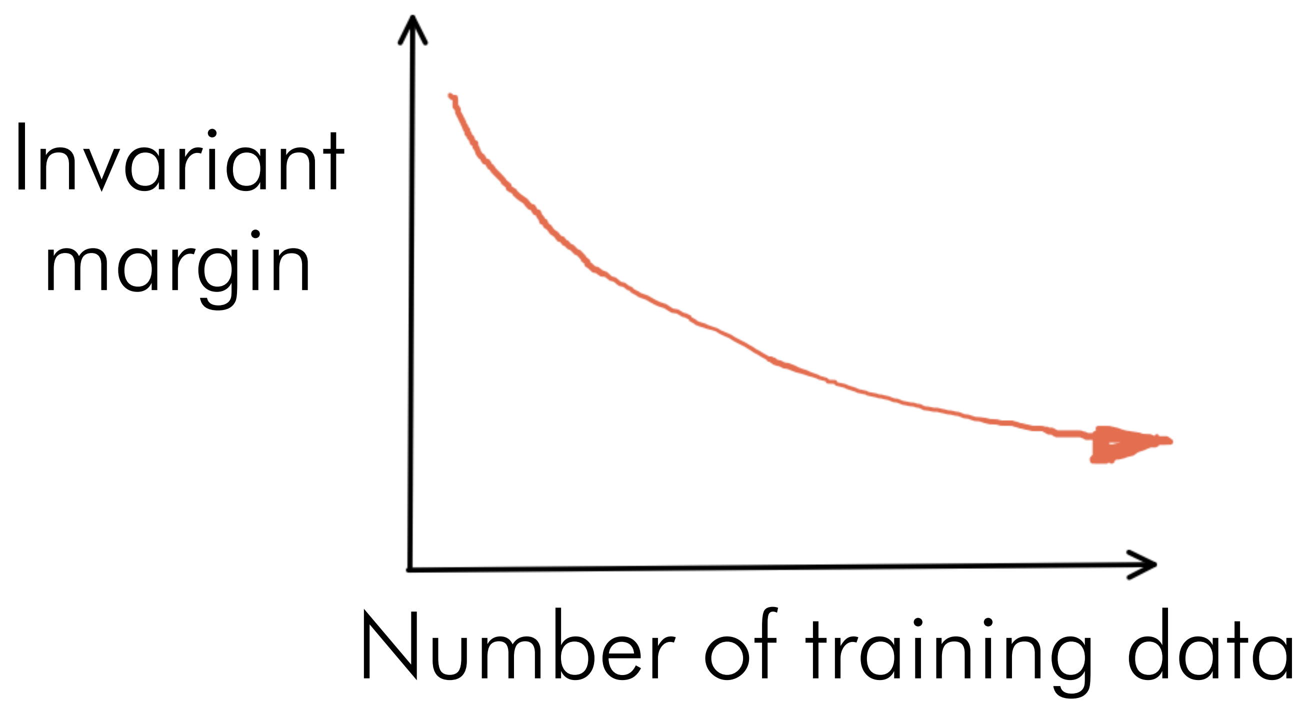
the margin decreases.



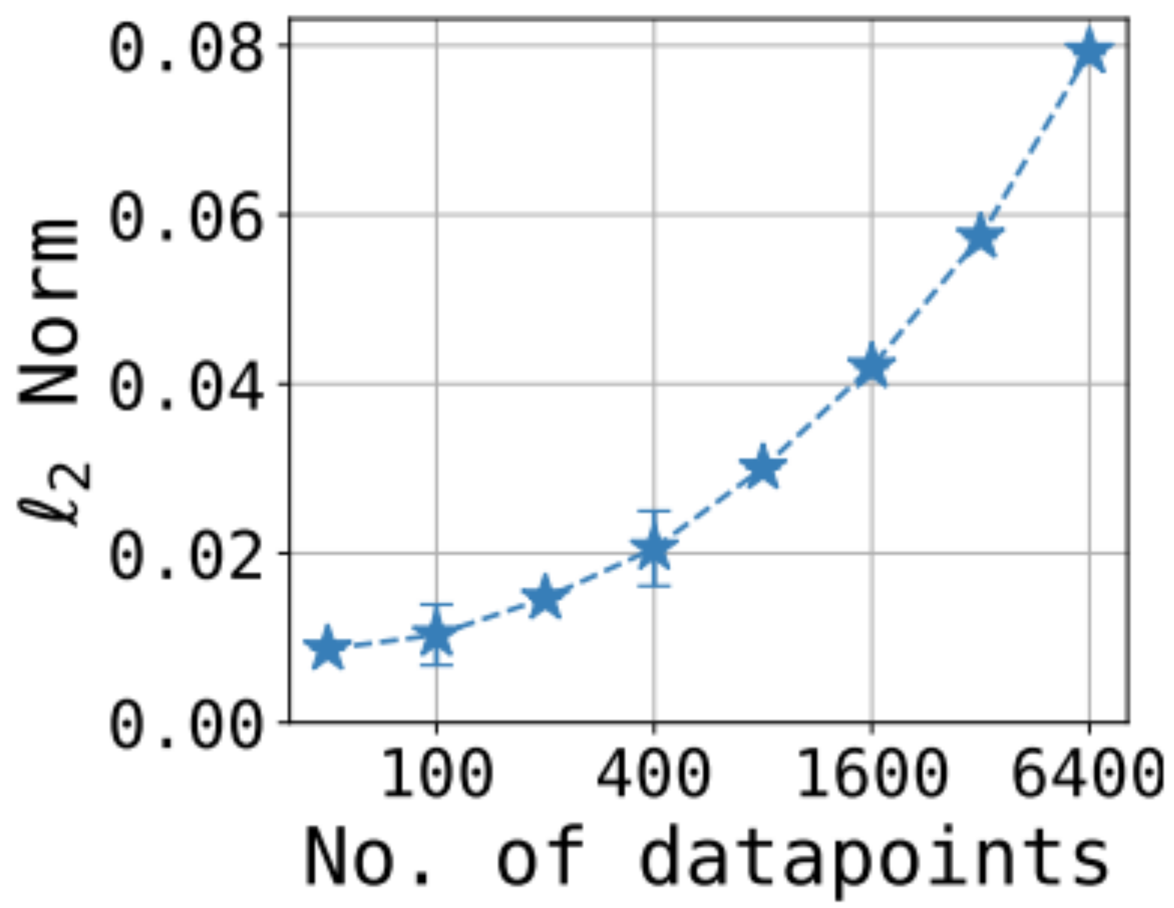
Source of failure 2: Geometric

Key property of real-world data geometry:

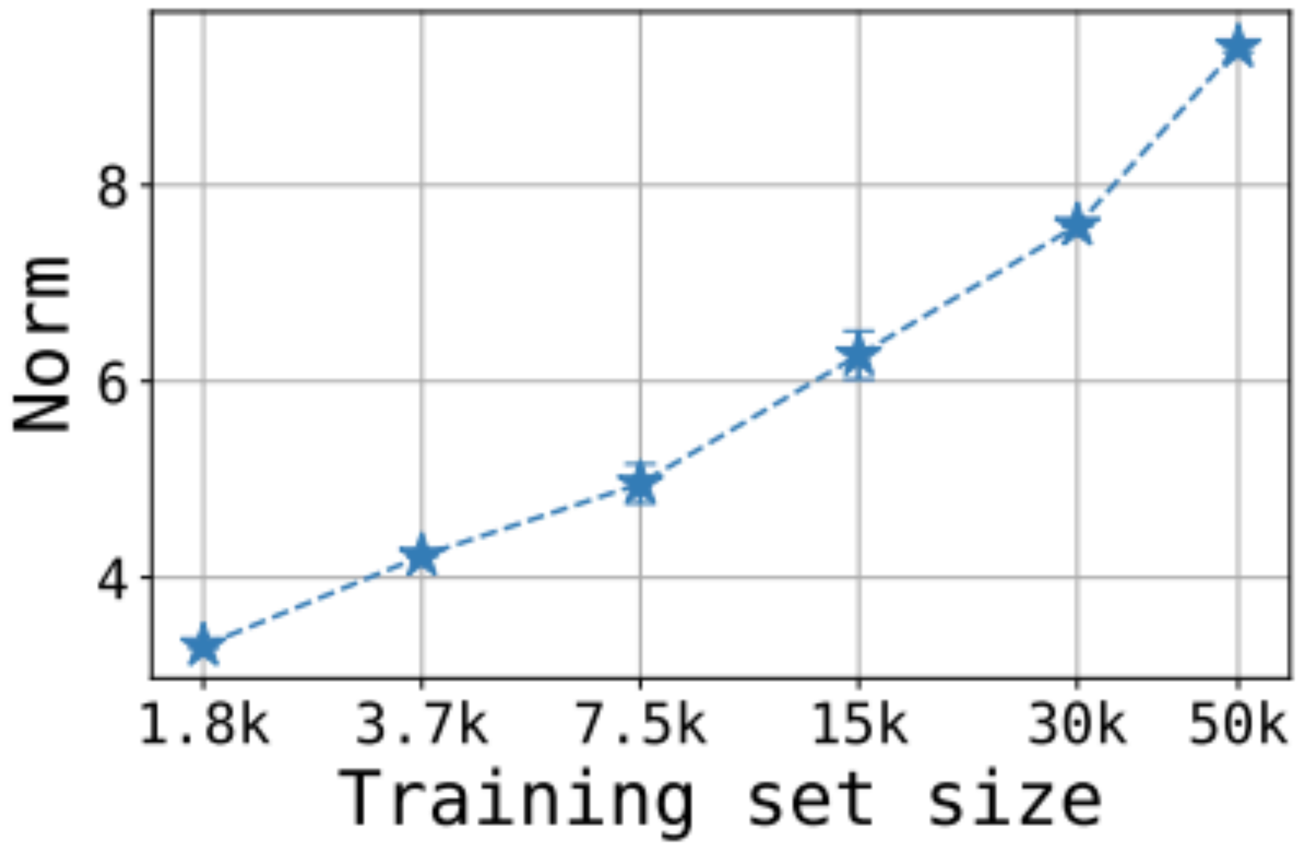
If we focused only on the invariant features,
the margin of separation along those features (call it "invariant margin") decreases with training set size.



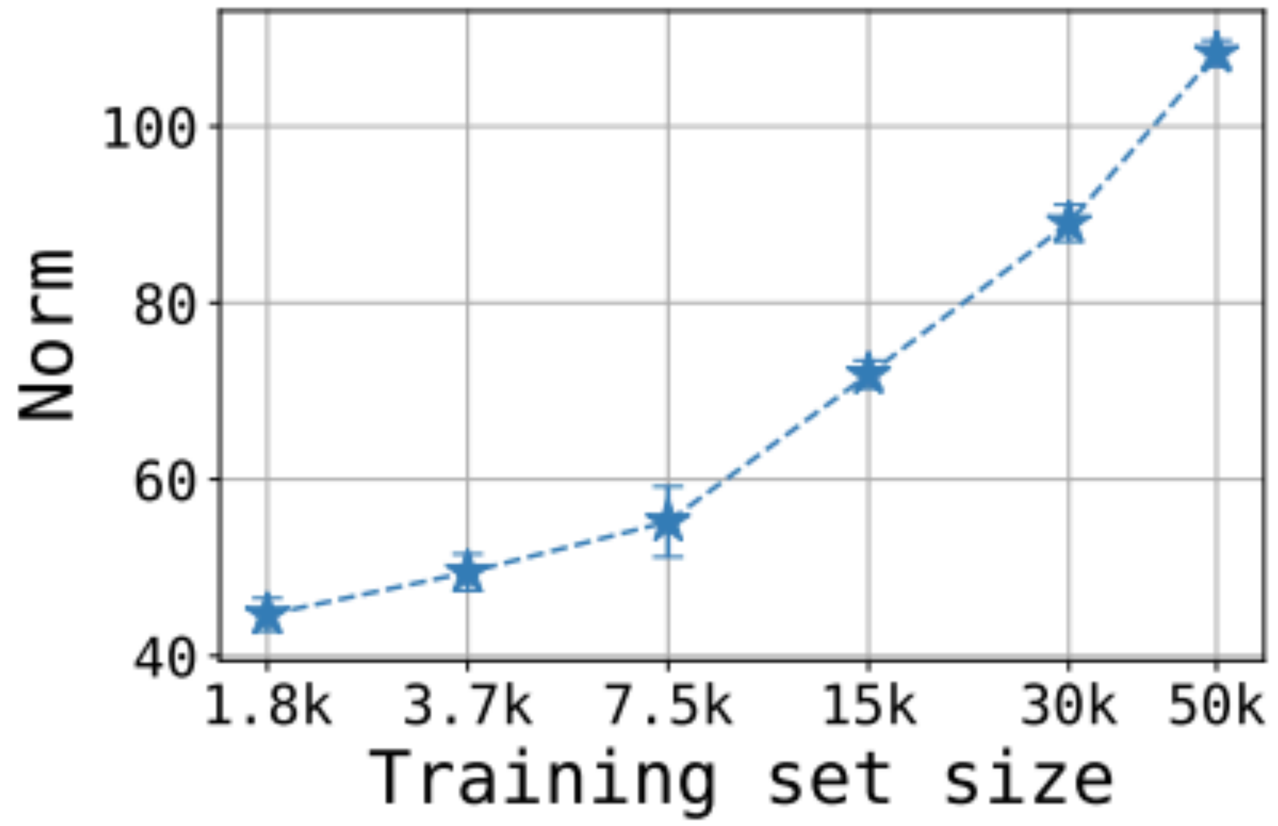
Empirical proof: (1/margin) increases



MNIST + max-margin on ReLU random features



Binary-MNIST + FNN

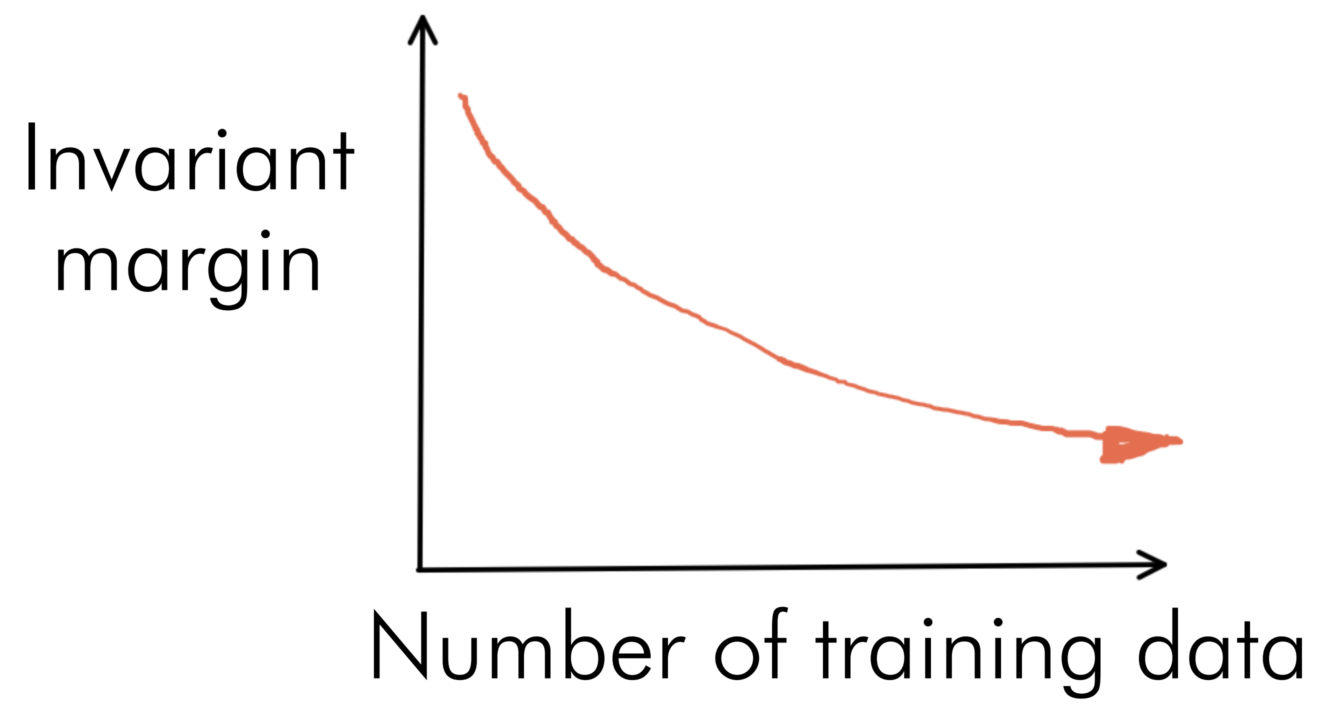


CIFAR10 + ResNet

Source of failure 2: Geometric

Key property of real-world data geometry:

If we focused only on the invariant features,
the margin of separation along those features (call it "invariant margin") decreases with training set size.



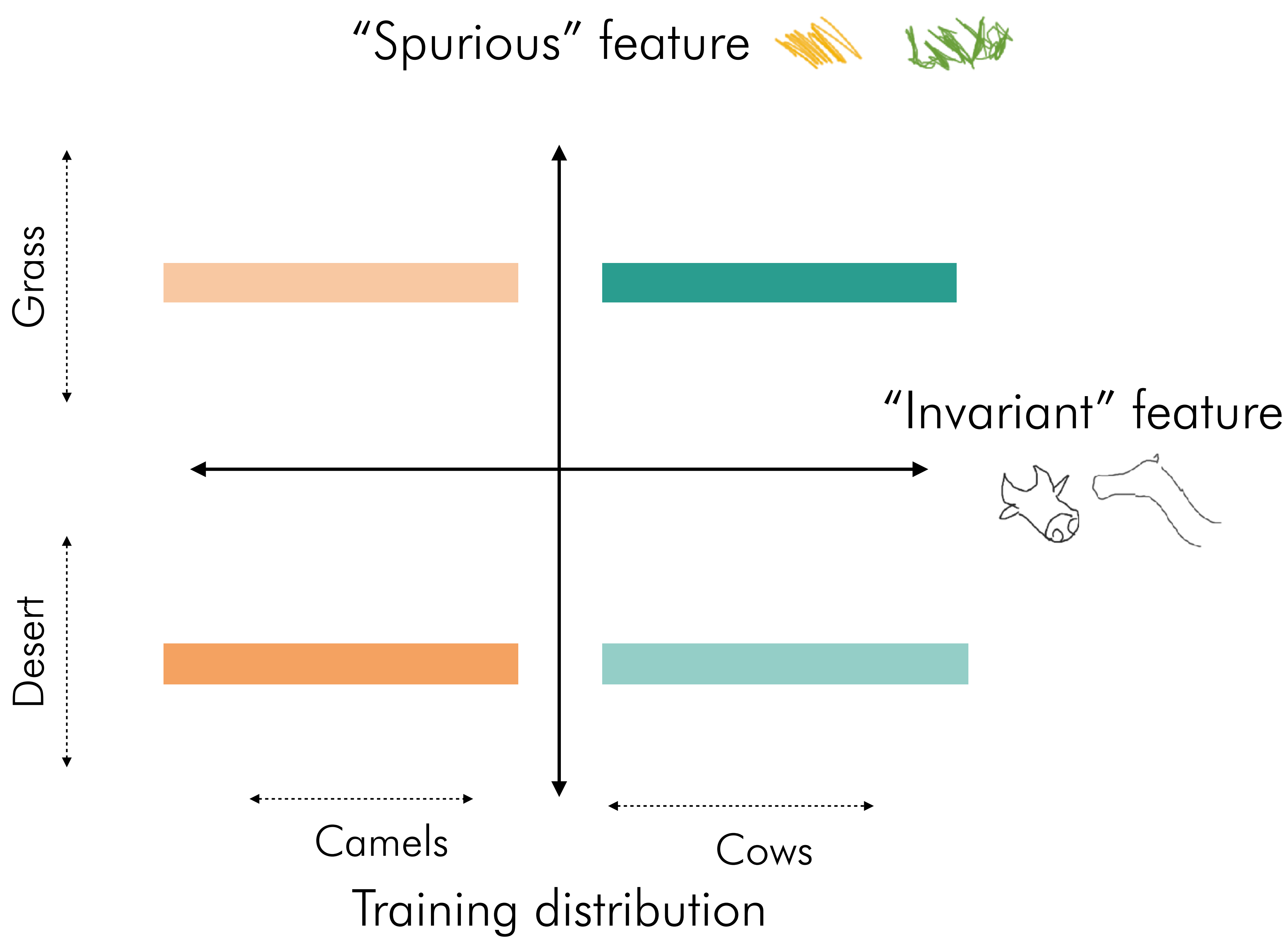
This helps explain failure of max-margin under spurious correlation!

Informal version of our result: For the max-margin classifier (over all the features),

$$|\text{spurious component}| = \Theta\,(\text{rate of decrease of invariant margin w.r.t training set size})$$

Source of failure 2: Geometric

Intuitive visualization in the real-world:

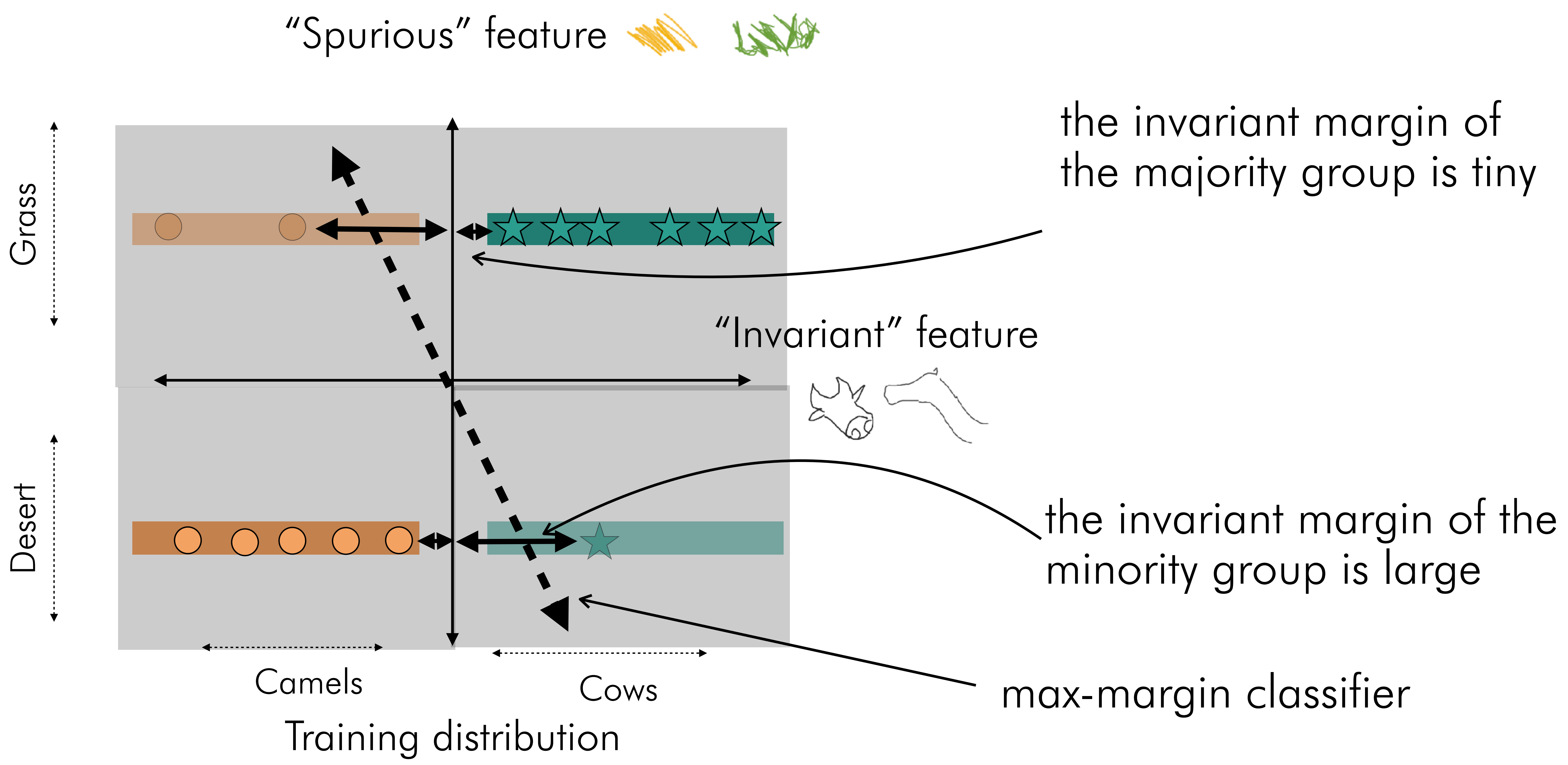


Informal version of our result: For the max-margin classifier (over all the features),

$$|\text{spurious component}| = \Theta\left(\text{rate of decrease of invariant margin w.r.t training set size}\right)$$

Source of failure 2: Geometric

Intuitive visualization in the real-world:

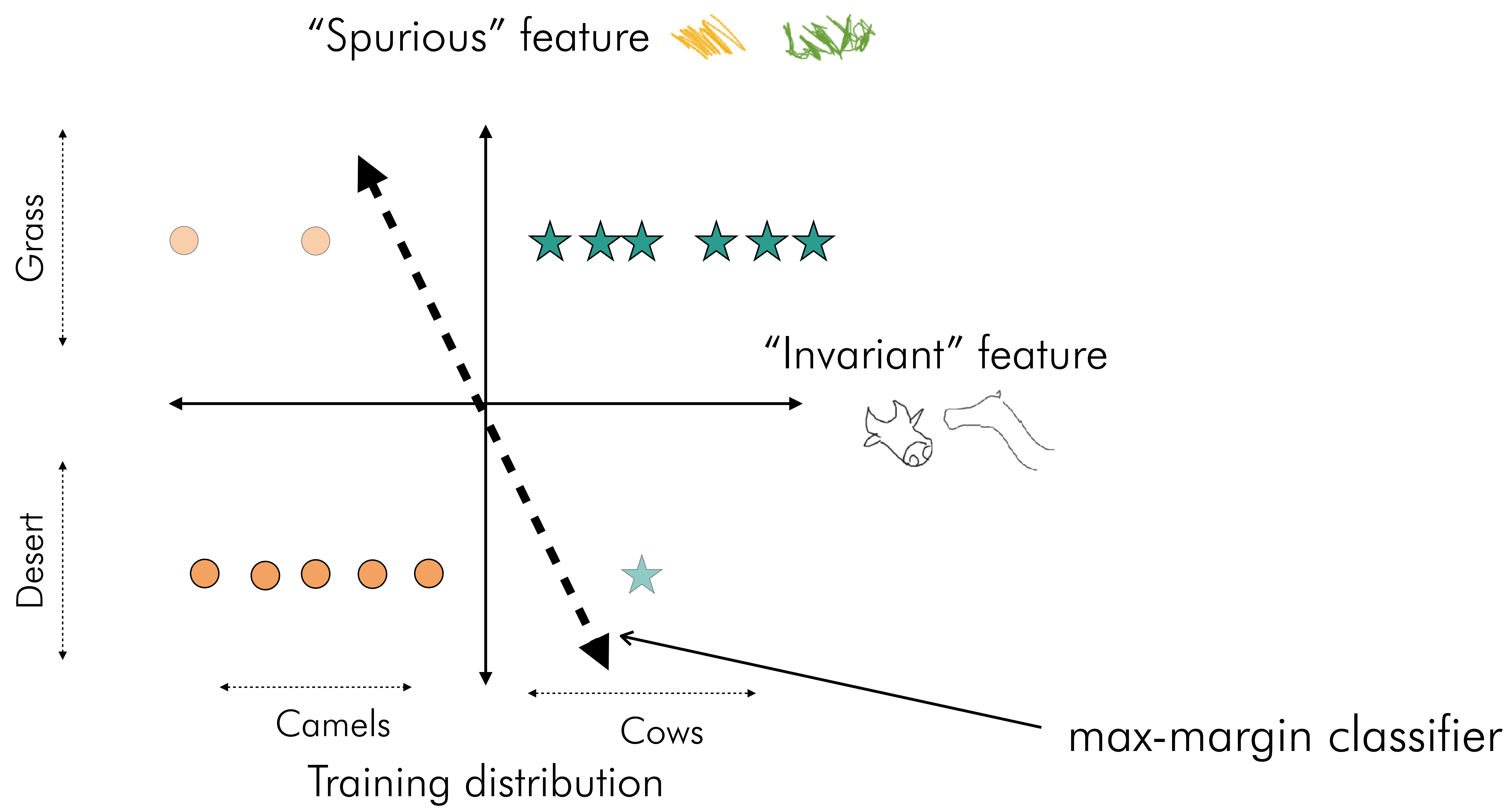


Informal version of our result: For the max-margin classifier (over all the features),

$$|\text{spurious component}| = \Theta\,(\text{rate of decrease of invariant margin w.r.t training set size})$$

Source of failure 2: Geometric

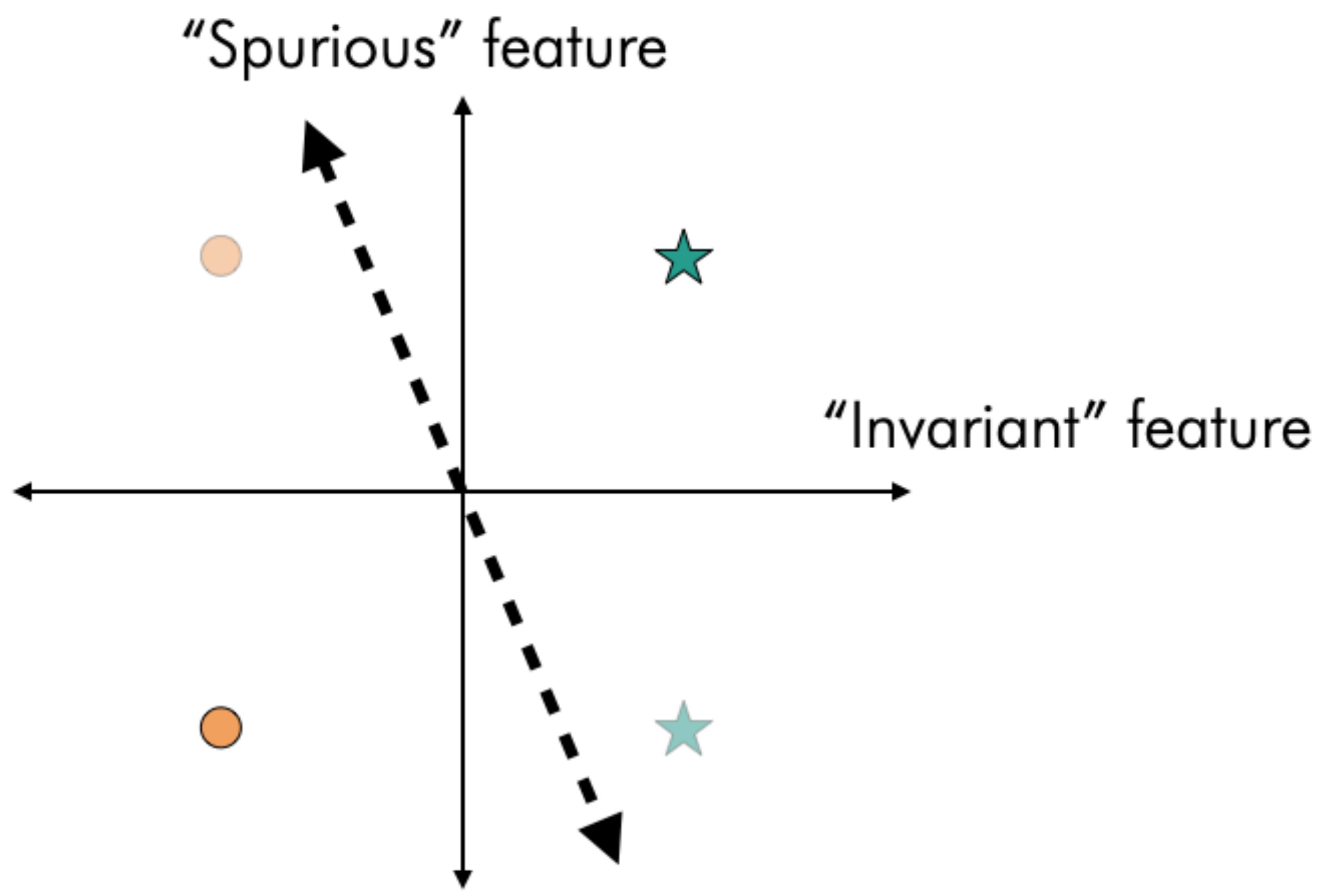
<u>Intuitive visualization in the real-world:</u>



Takeaway: Spurious-feature-reliance happens because of
(a) non-degenerate geometry in the real-world
(b) margin-maximizing bias.

Summary of theoretical insights

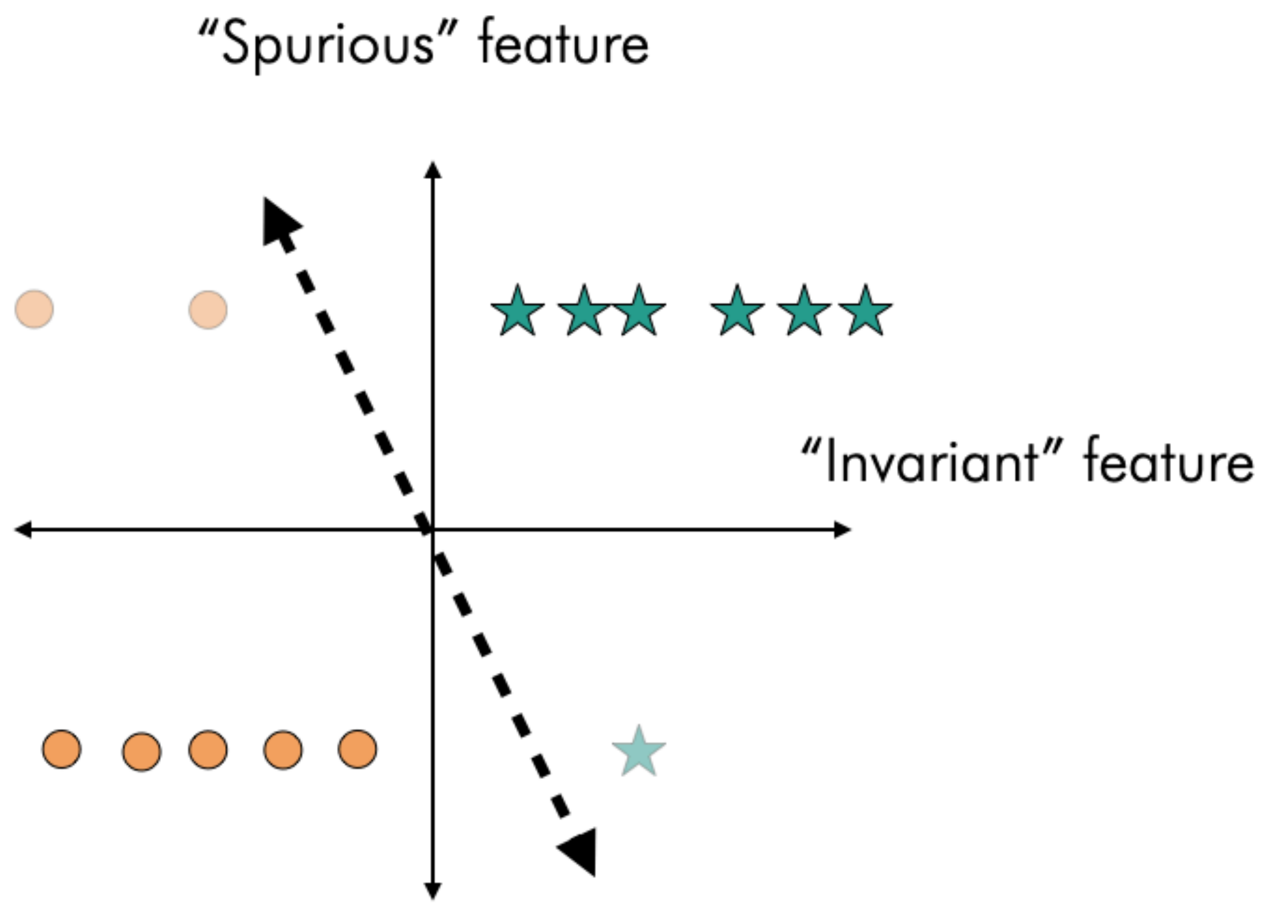
Statistical failure



Occurs even in degenerate geometries

Occurs due to bias in finite-time GD;
Does not occur in max-margin

Geometric failure



Occurs due to geometry of the invariant features

Occurs due to margin-maximizing bias

Outline

- Introduction
- Motivation: existing theoretical models are inadequate
- Failure mode 1: statistical skews
- Failure mode 2: geometric skews
- Takeaways
- Conclusion

Justification for existing/new algorithms

Upsampling the minority group

- Addresses statistical failure mode.

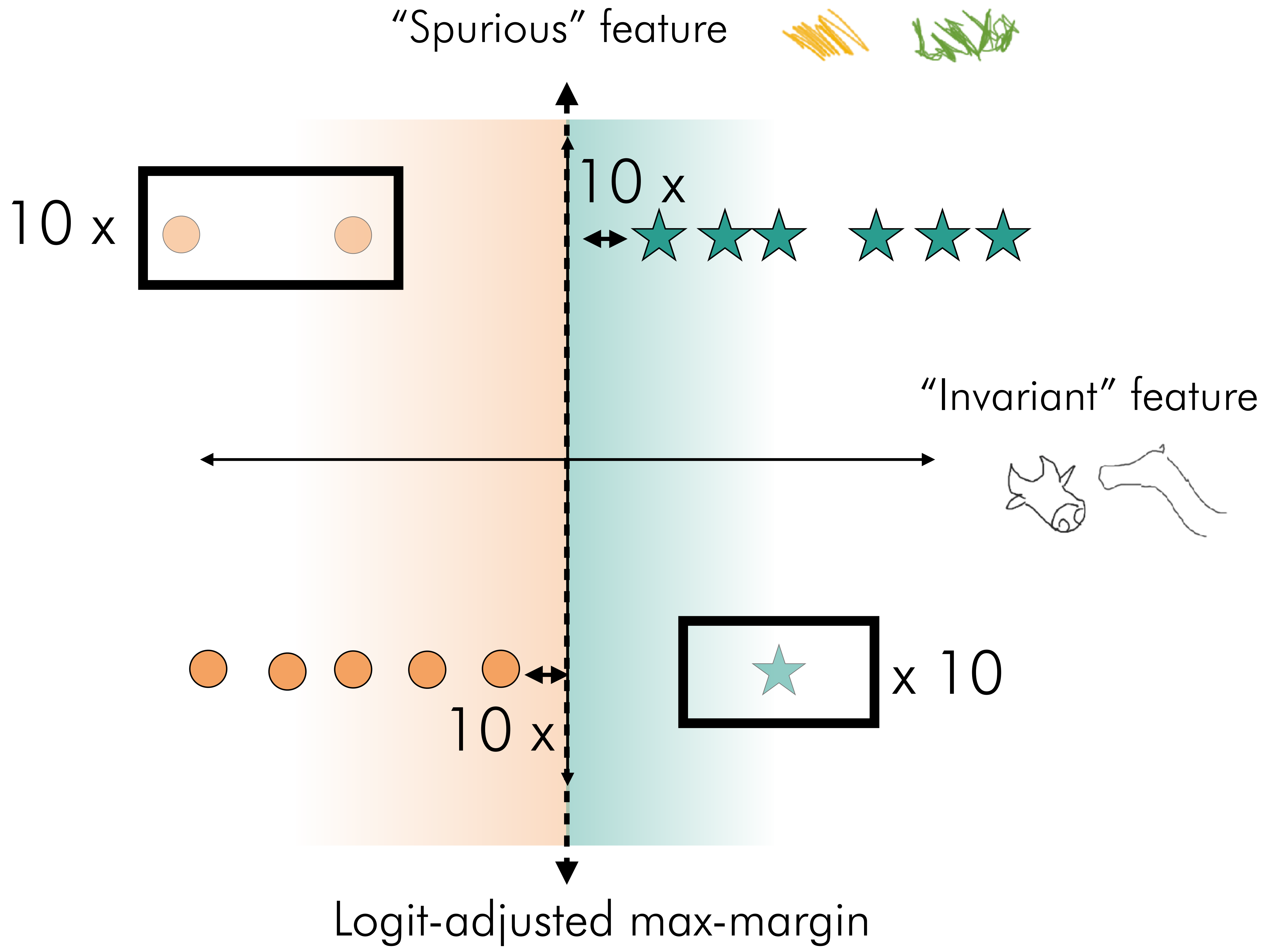
Logit adjustment during training

(ours and Kini-Paraskevas-Oymak-Thrampoulidis ‘21)

$$max_{w,||w||=1} \begin{cases} yw^T x & \text{if minority} \\ 10y(w^T x) + 10 & \text{if majority} \end{cases}$$

$\equiv$ scaling up the majority logits during GD

- Addresses geometric failure mode.



Justification for existing/new algorithms

Upsampling the minority group

- Addresses statistical failure mode.
- Doesn't address geometric mode!

+

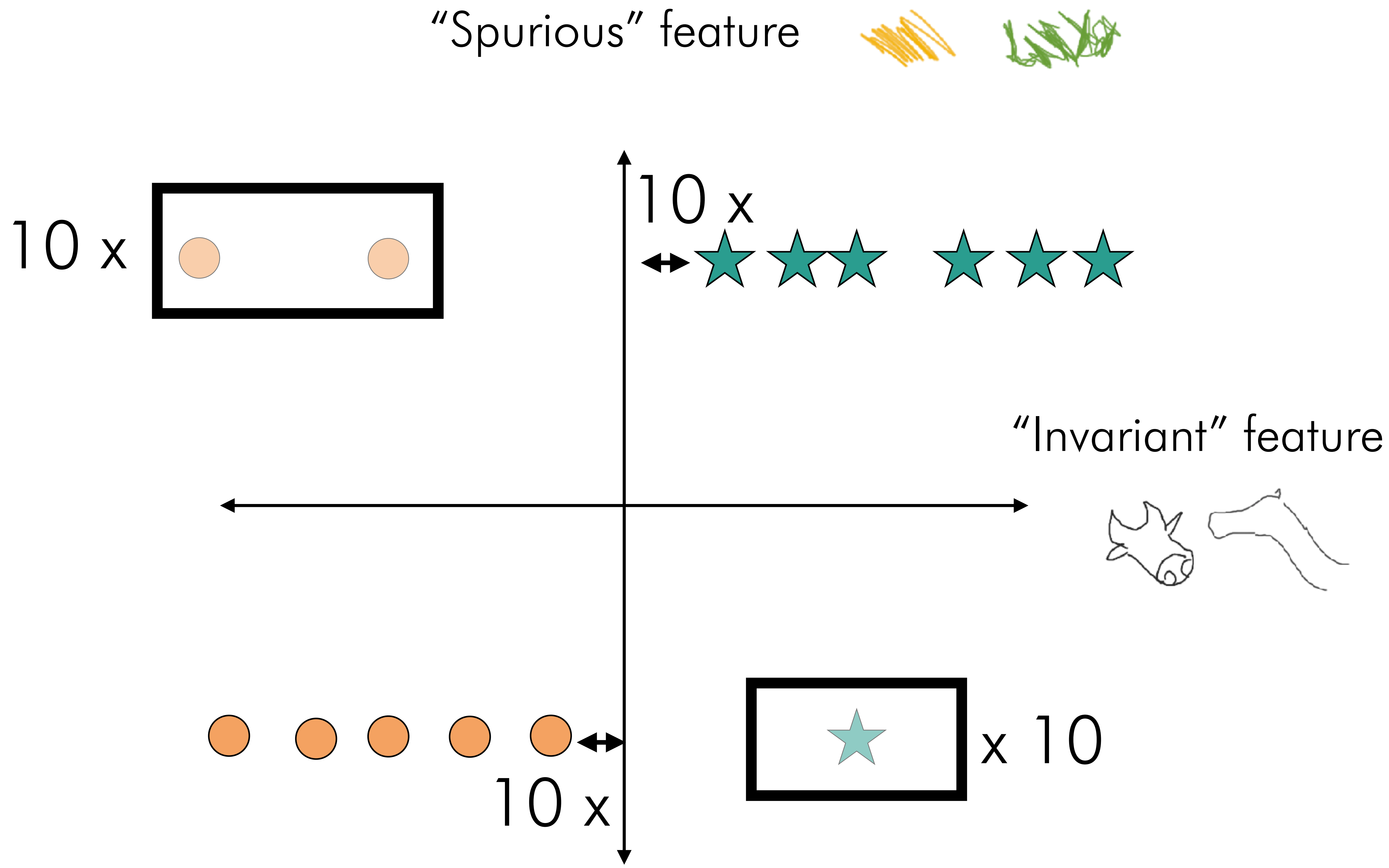
Logit adjustment during training

(ours and Kini-Paraskevas-Oymak-Thrampoulidis '21)

$$max_{w,||w||=1} \begin{cases} yw^T x & \text{if minority} \\ 10y(w^T x) + 10 & \text{if majority} \end{cases}$$

≡ scaling up the majority logits during GD

- Addresses geometric failure mode.
- May not address statistical mode in finite-time GD!



Justification for existing/new algorithms

Practical takeaway: We need to combine approaches to address both kind of failures

Algorithm	ColoredMNIST	Waterbirds	CelebA
ERM	93.1	71.7	53.3
Upsampling	96.1 (+3.0)	86.0 (+14.3)	85.0 (+31.7)
Margin Scaling	95.2 (+2.1)	81.9 (+10.2)	57.7 (+4.4)
GroupDRO	**97.4 (+4.3)**	**90.3* (+18.6)**	87.6* (+34.3)
Downsampling	96.1 (+3.0)	87.6 (+15.9)	**88.9 (+35.6)**
Margin Scaling + Upsampling	96.2 (+3.1)	85.0 (+13.3)	87.8 (+34.5)
GroupDRO + Upsampling	96.5 (+3.4)	87.6 (+15.9)	86.7 (+33.4)

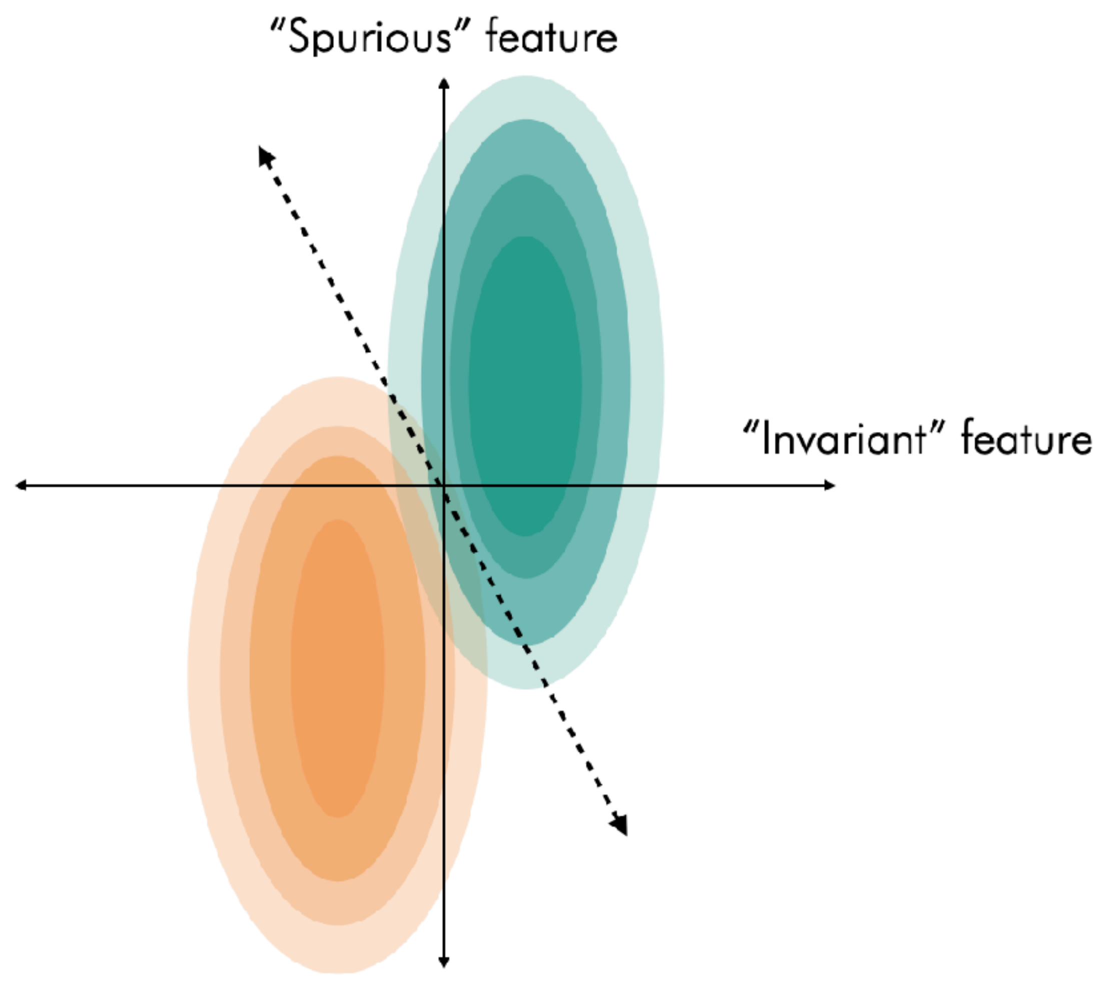
Future directions

Practical takeaway: We need to combine approaches to address both kind of failures

- Better approaches to both failure modes?
  - Statistical: Upsampling overfits; poor dynamics.
  - Geometric: Logit adjustment can only partially help in high-dim.

- Understand dynamics of
  - Upsampling
  - Logit adjustment
  - Group DRO…

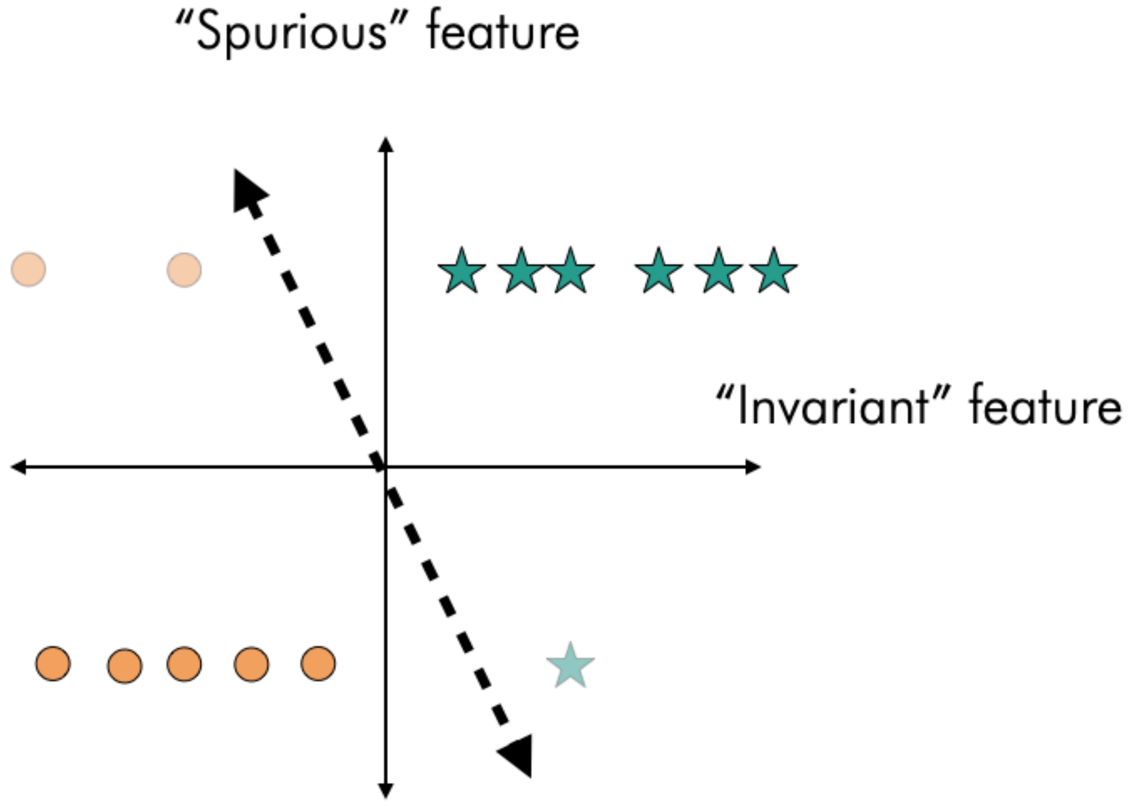
Conclusion



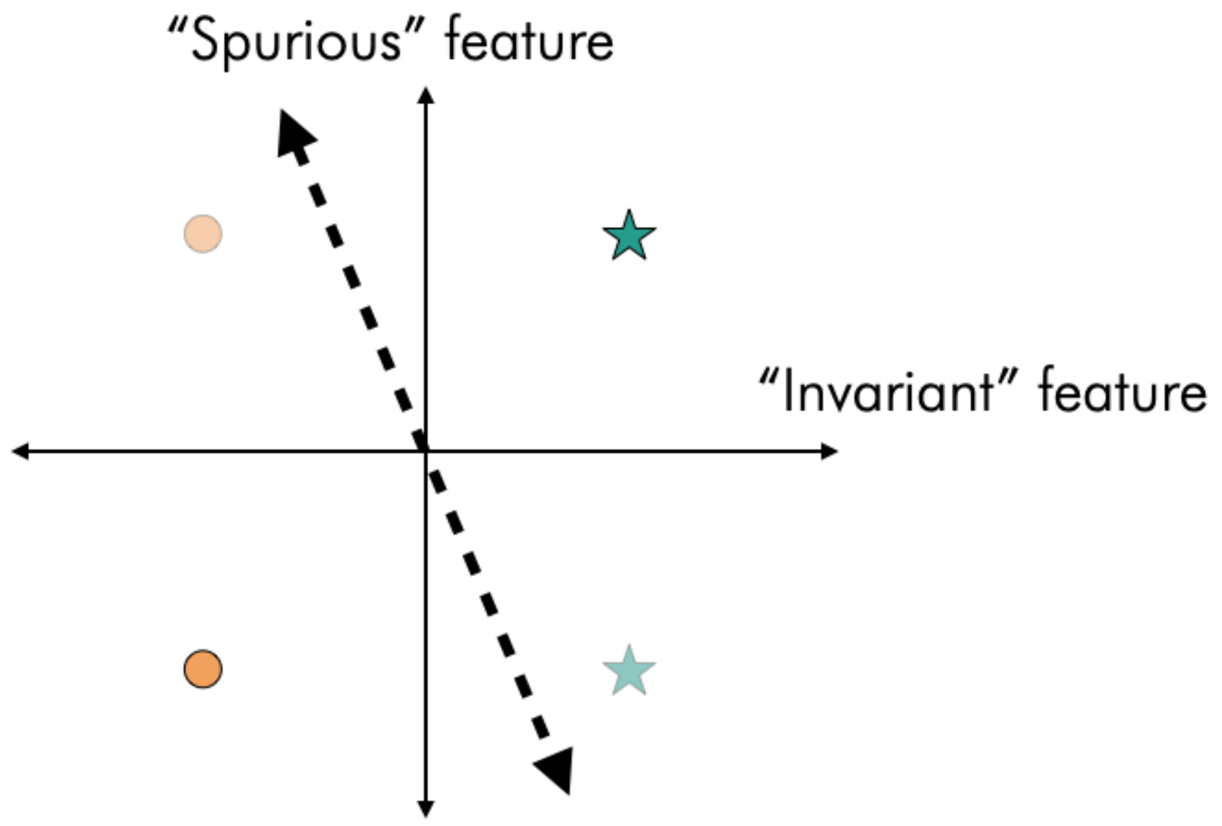
- We challenge the prevailing theoretical understanding of why models fail under spurious correlations.

- By proposing a "fully informative invariant feature" model, we identify that there is no one unique way by which failure occurs:

Geometric skews



Statistical skews



- Our result may guide the field towards a more appropriate theoretical model which can better inform the theory and algorithms that build on it.

Thank you! Questions?

Reference: **"Understanding the failure modes of out-of-distribution generalization"**, ICLR 2021, Vaishnavh Nagarajan, Anders Andreassen, Behnam Neyshabur.

Reference: **"Avoiding Spurious Correlations: Bridging Theory and Practice"**, DistShift Workshop NeurIPS 21, Thao Nguyen, Vaishnavh Nagarajan, Hanie Sedghi, Behnam Neyshabur.