# Distinguishing rule- and exemplar-based generalization in learning systems

Ishita Dasgupta *@ — PRINCETON UNIVERSITY

Erin Grant * — Berkeley UNIVERSITY OF CALIFORNIA

Tom Griffiths — PRINCETON UNIVERSITY

* Equal contribution.

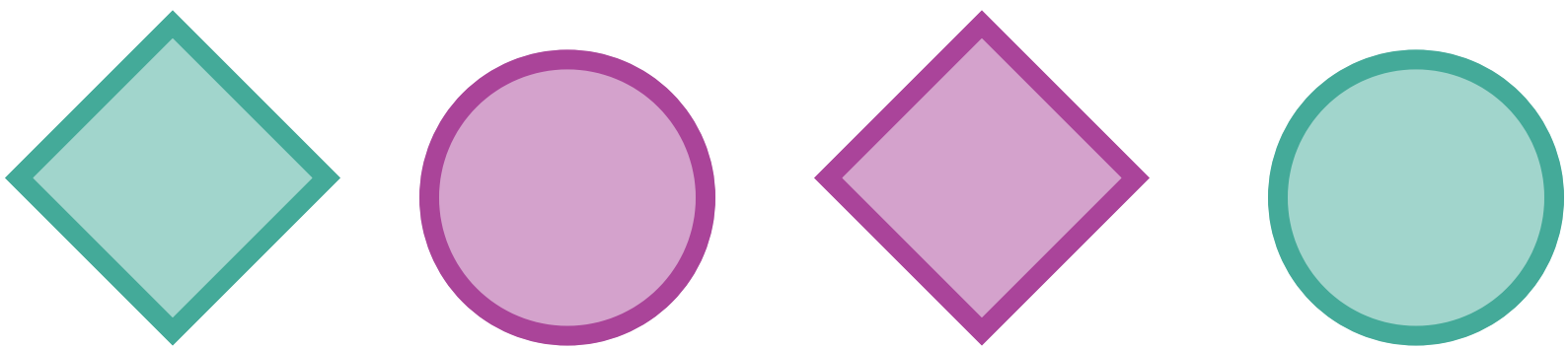
@ Now at DeepMind.

arXiv:2110.04328

eringrant/icml-2022-rules-vs-exemplars

This work:

(two) inductive biases in category learning.



"A"

"B"

?

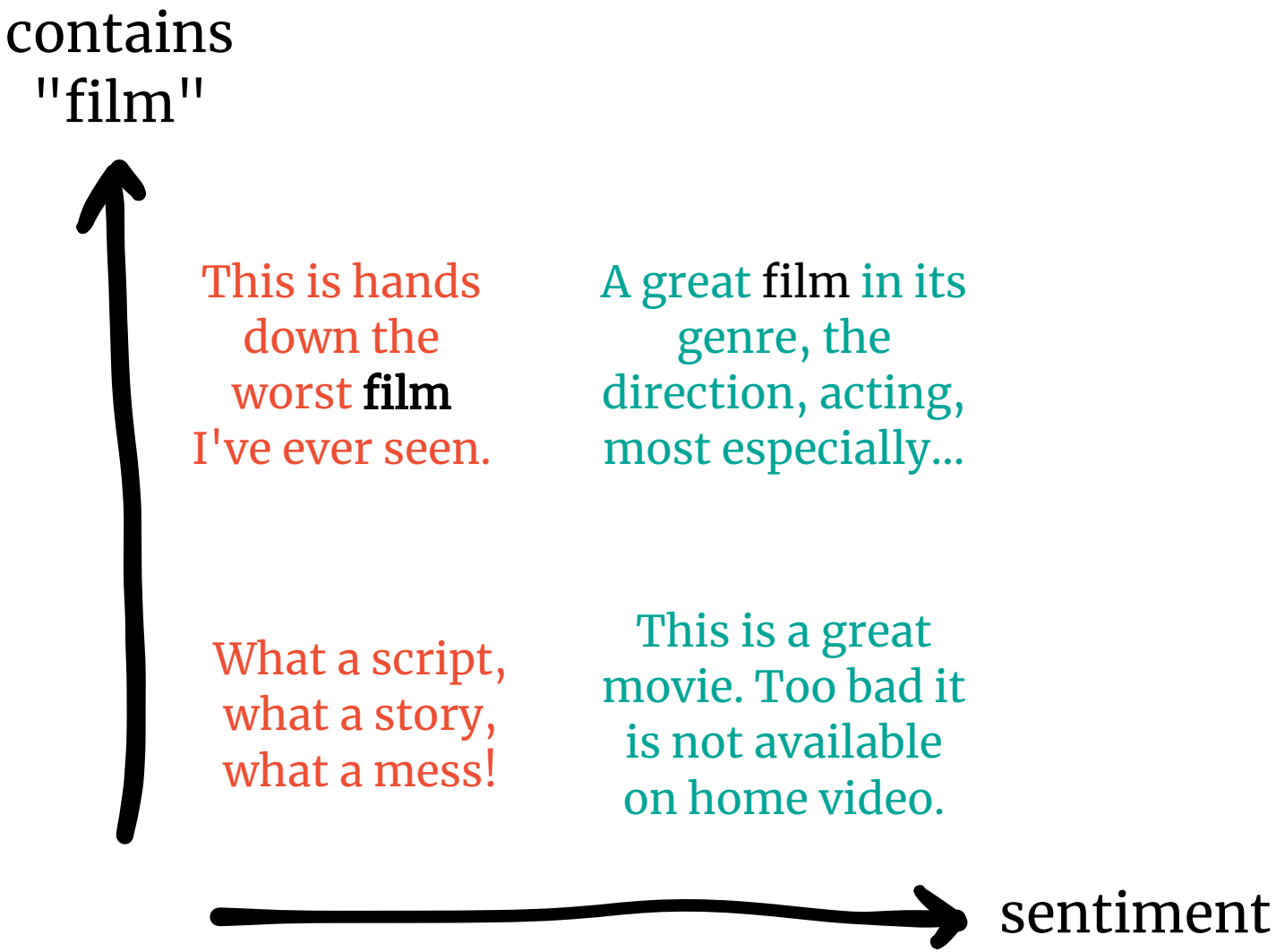
"Is this *A* or *B*?"

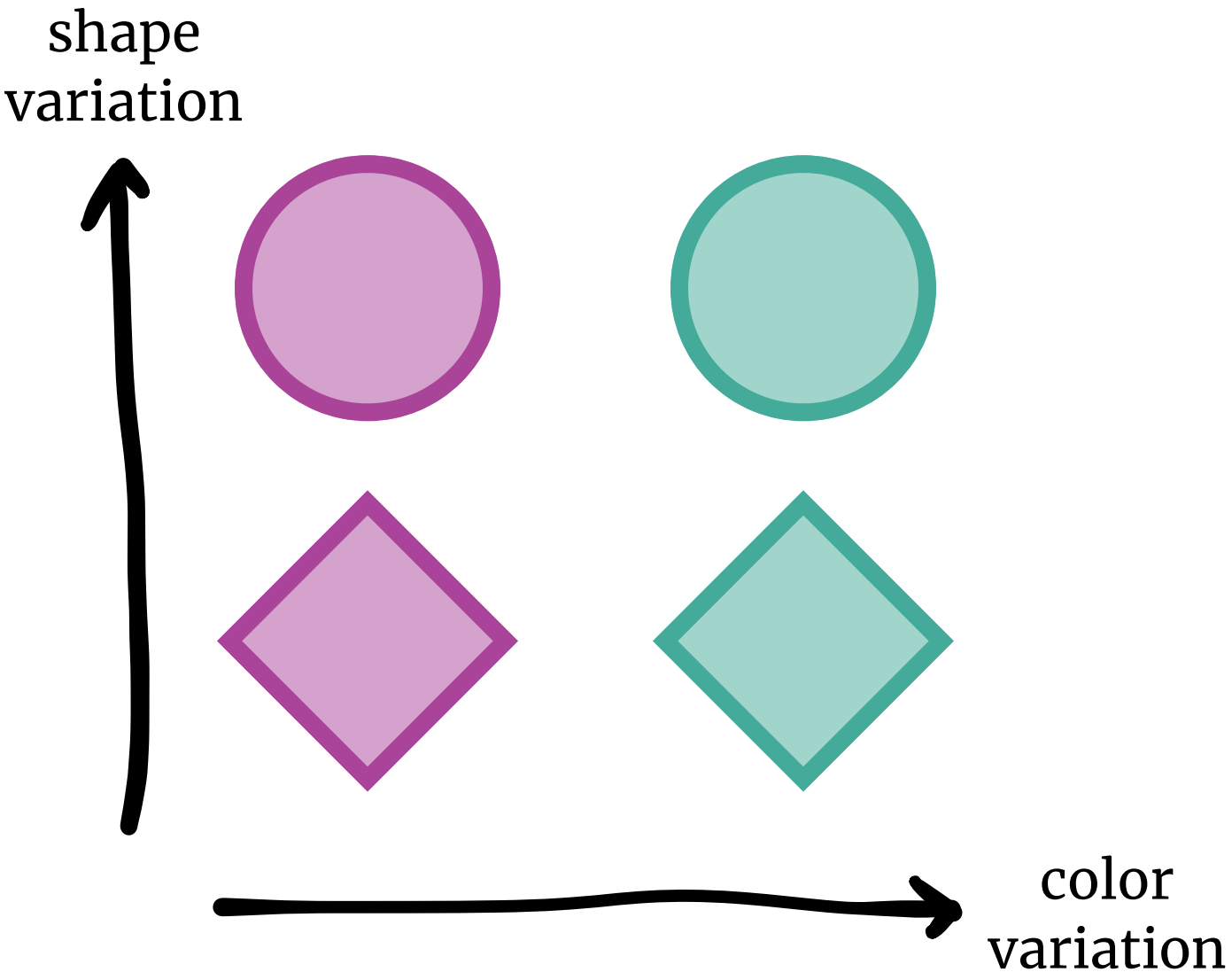
beach or grass?

dog or cow?

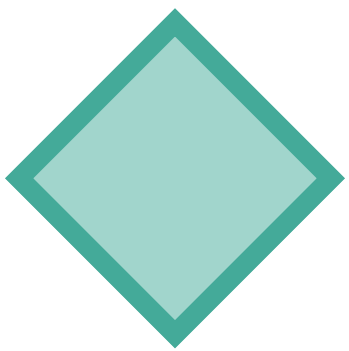
beach or
grass?

dog or
cow?

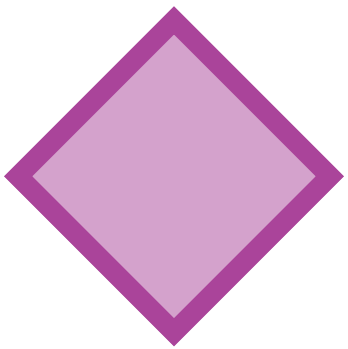

contains "film"
This is hands down the worst film I've ever seen.
A great film in its genre, the direction, acting, most especially...
What a script, what a story, what a mess!
This is a great movie. Too bad it is not available on home video.
sentiment


shape
variation
color
variation

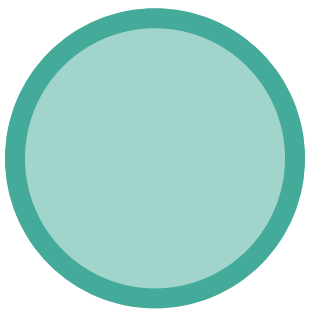
training condition #1



"dax"



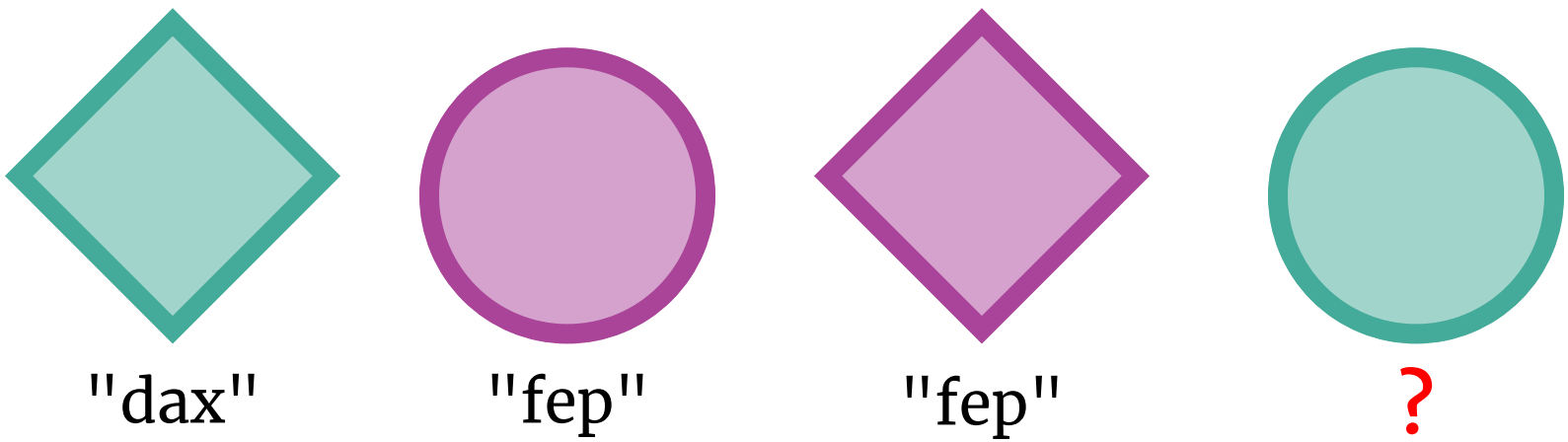
"fep"

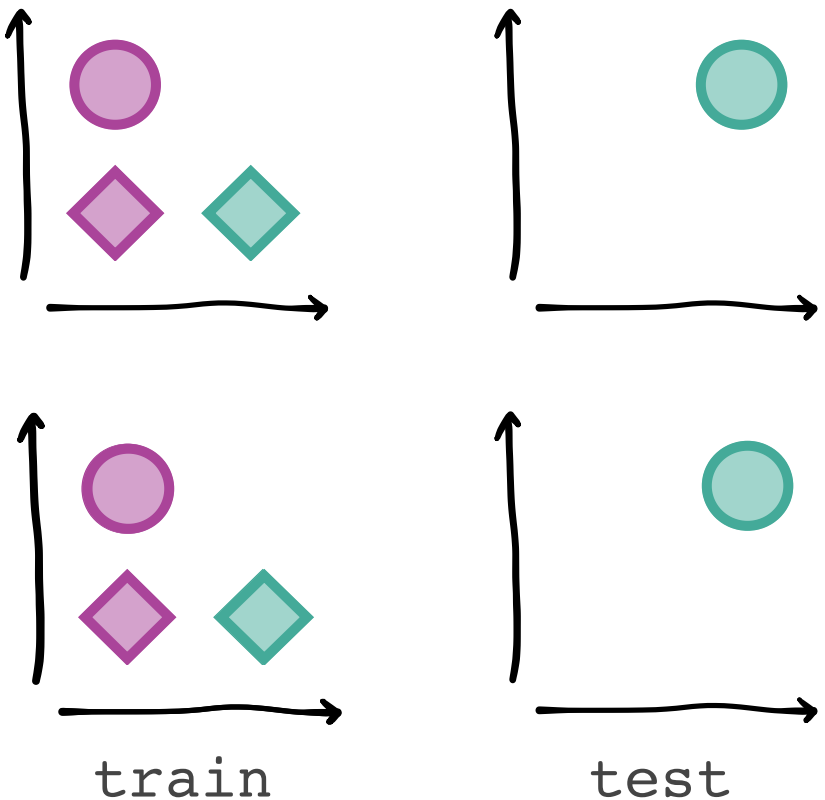


?

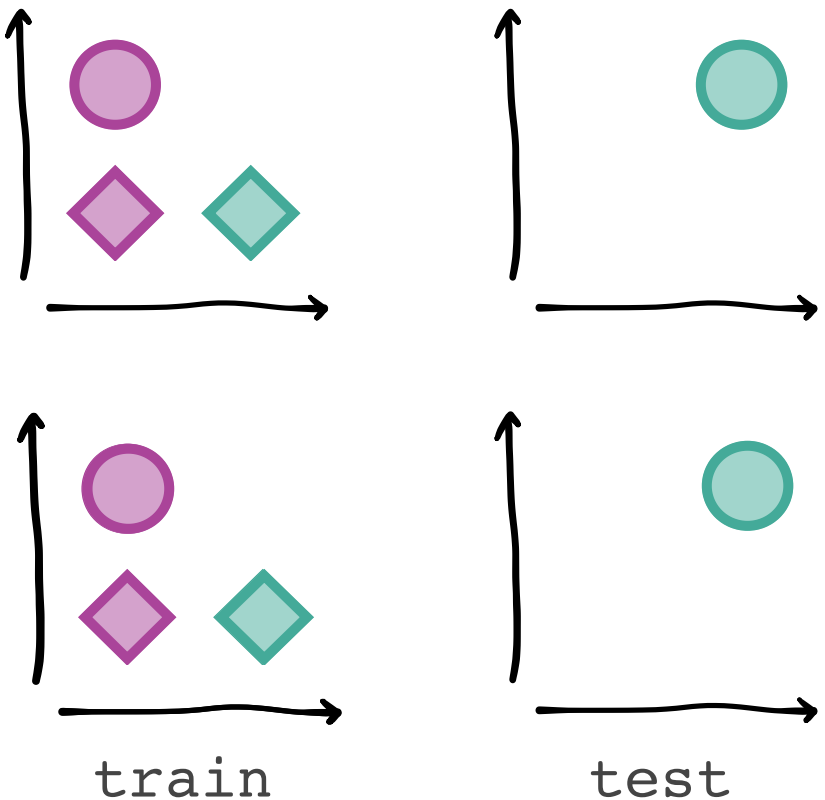
"Is this a *dax* or a *fep*?"

## training condition #2



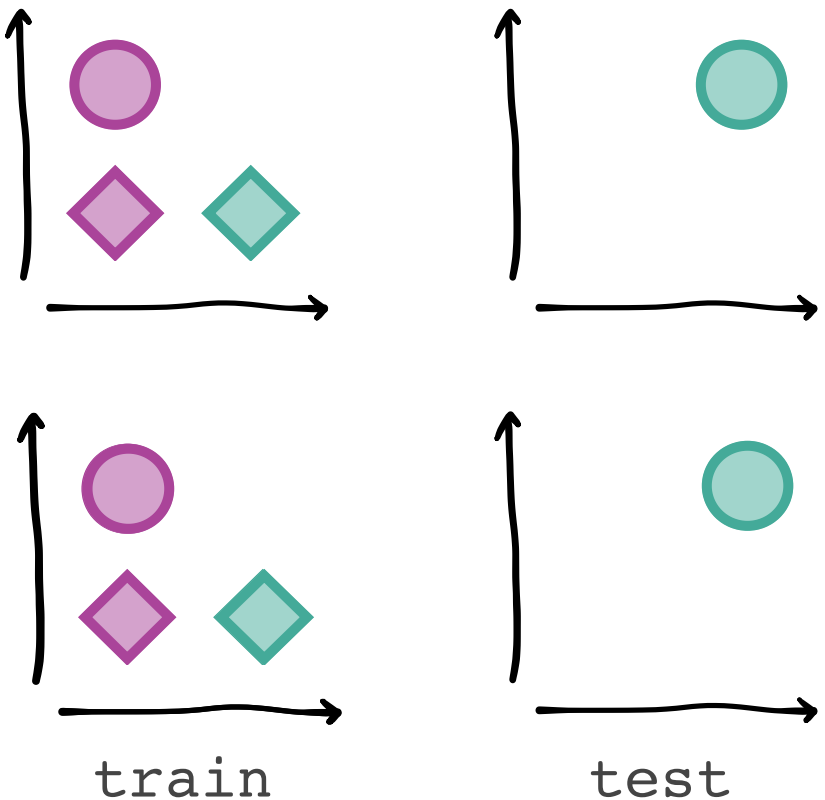

train
test

# "Exemplar bias" (vs "rule bias"; EvR):



**"Exemplar bias" (vs "rule bias"; EvR):**

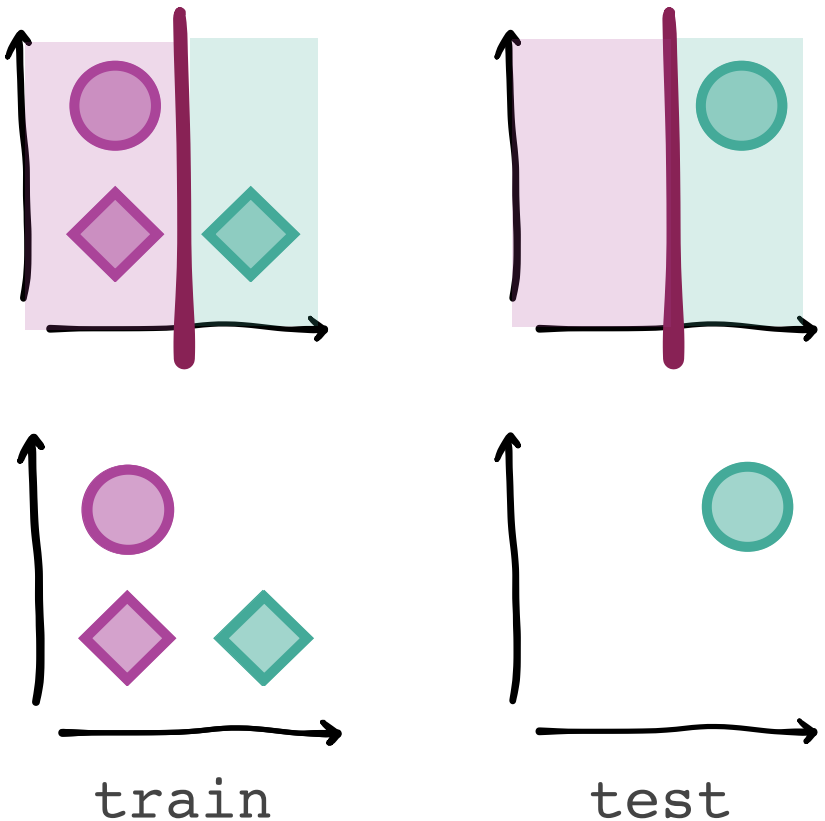
simple or complex decision boundary?



**"Exemplar bias" (vs "rule bias"; EvR):**

simple or complex decision boundary?

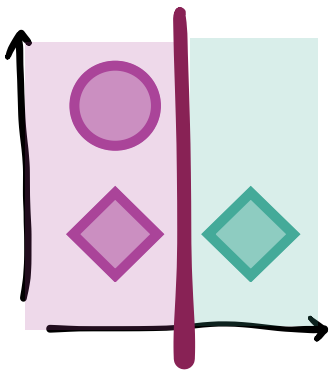
**rule** bias:

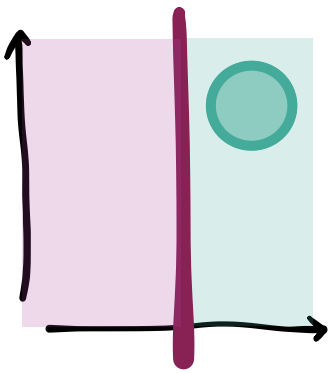


**"Exemplar bias" (vs "rule bias"; EvR):**

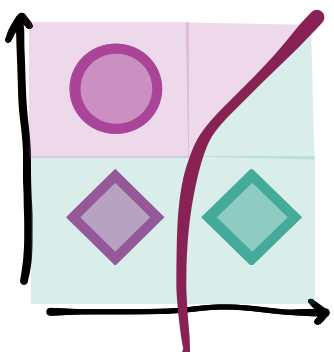
simple or complex decision boundary?

**rule** bias:

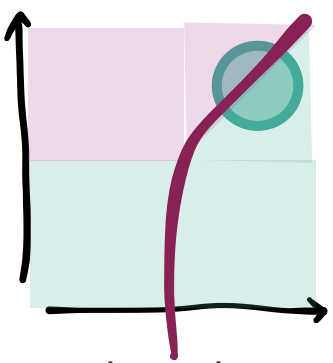




**exemplar** bias:



train

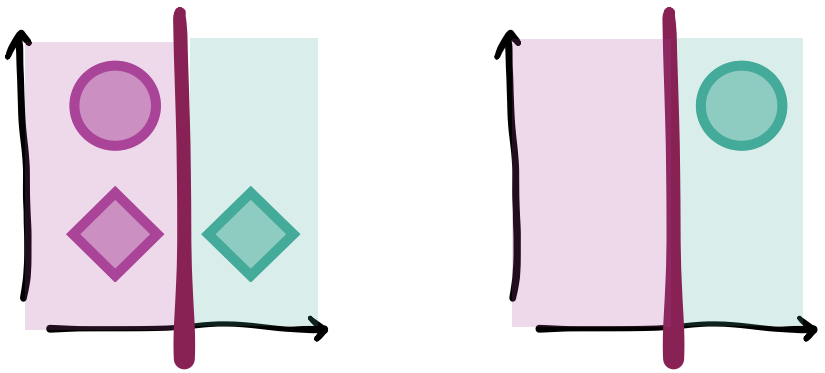


test

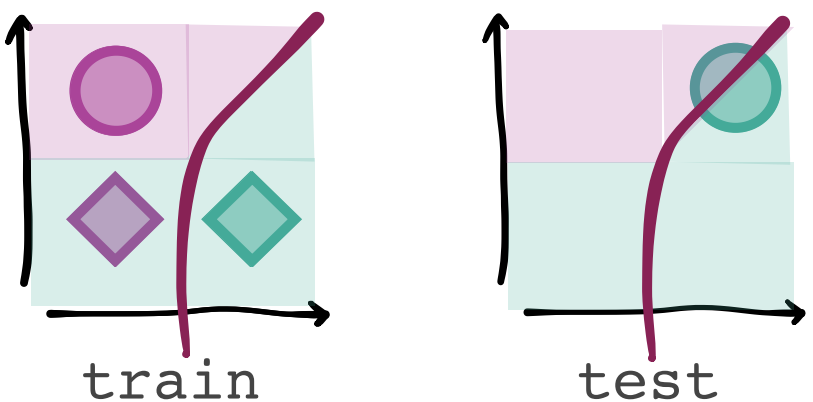
**"Exemplar bias" (vs "rule bias"; EvR):**

simple or complex decision boundary?

**rule** bias:

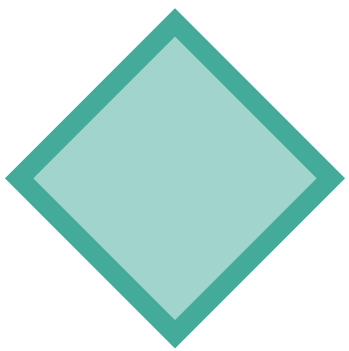


**exemplar** bias:

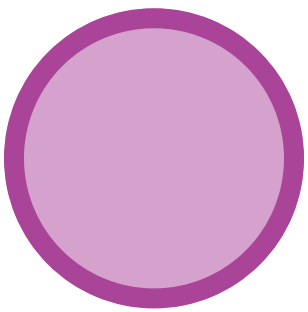



$$\text{EVR}(\text{NN}) = \mathbb{E}\left[\text{accuracy}(\diamond\ \diamond)\right] - \mathbb{E}\left[\text{accuracy}(\bigcirc\!\!\diamond\ \diamond)\right]$$

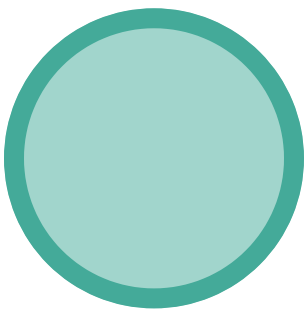
## training condition #3



"dax"



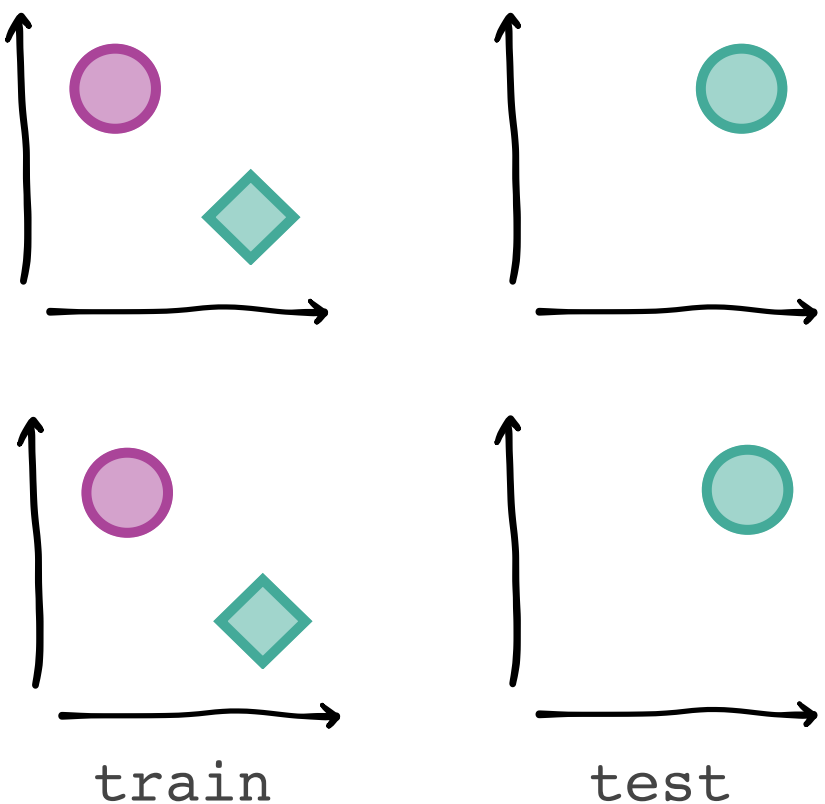
"fep"



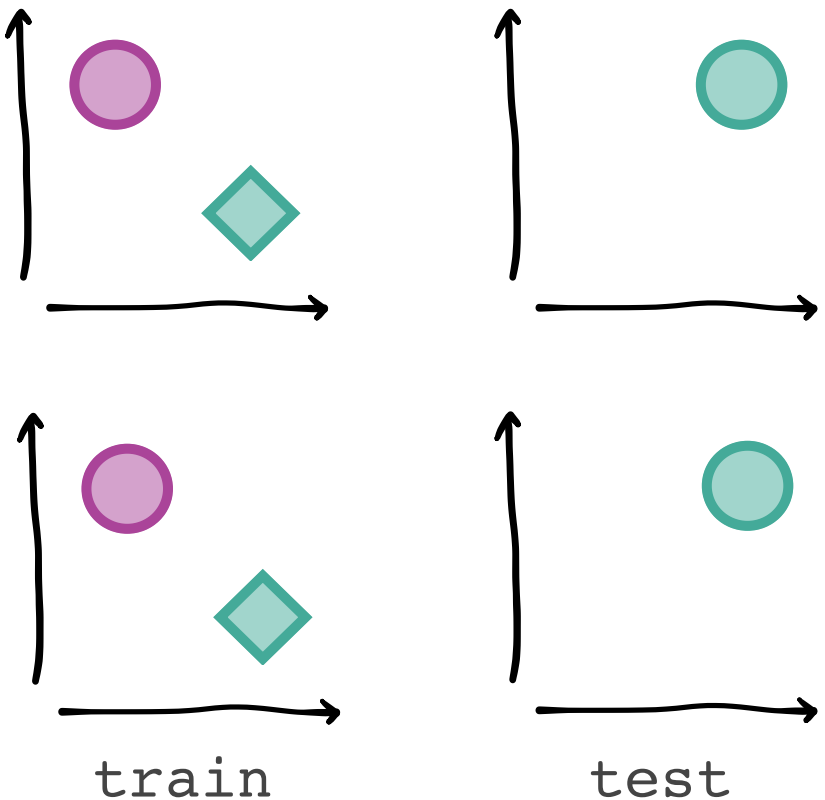
?

"Is this a *dax* or a *fep*?"

"Feature-level bias" (FLB):

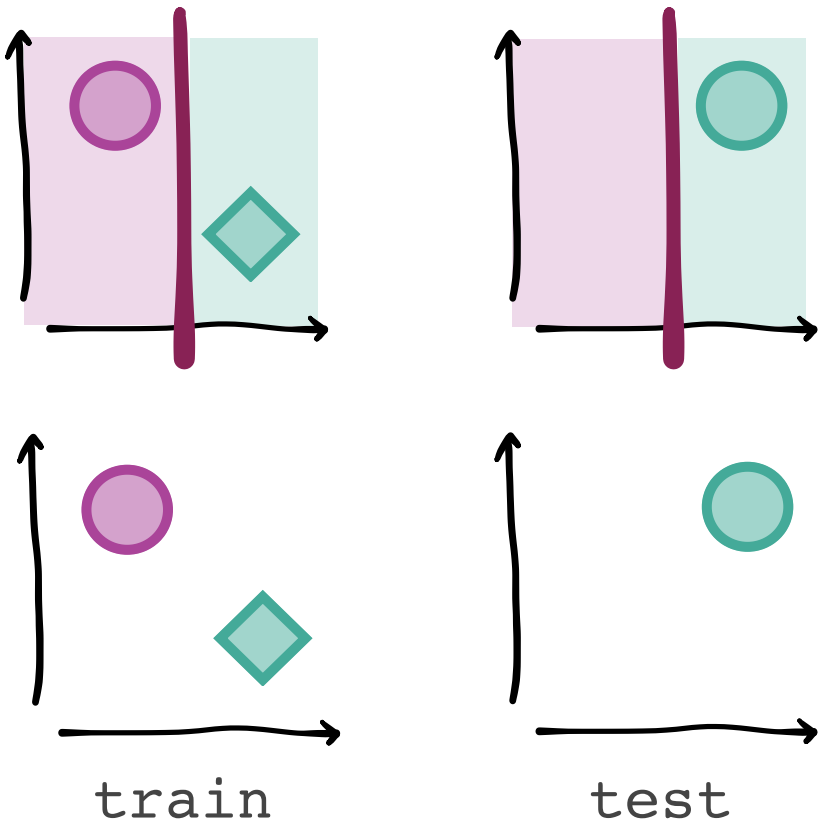


"**Feature-level bias" (FLB)**:
which equally predictive feature?



"**Feature-level bias" (FLB)**:
which equally predictive feature?

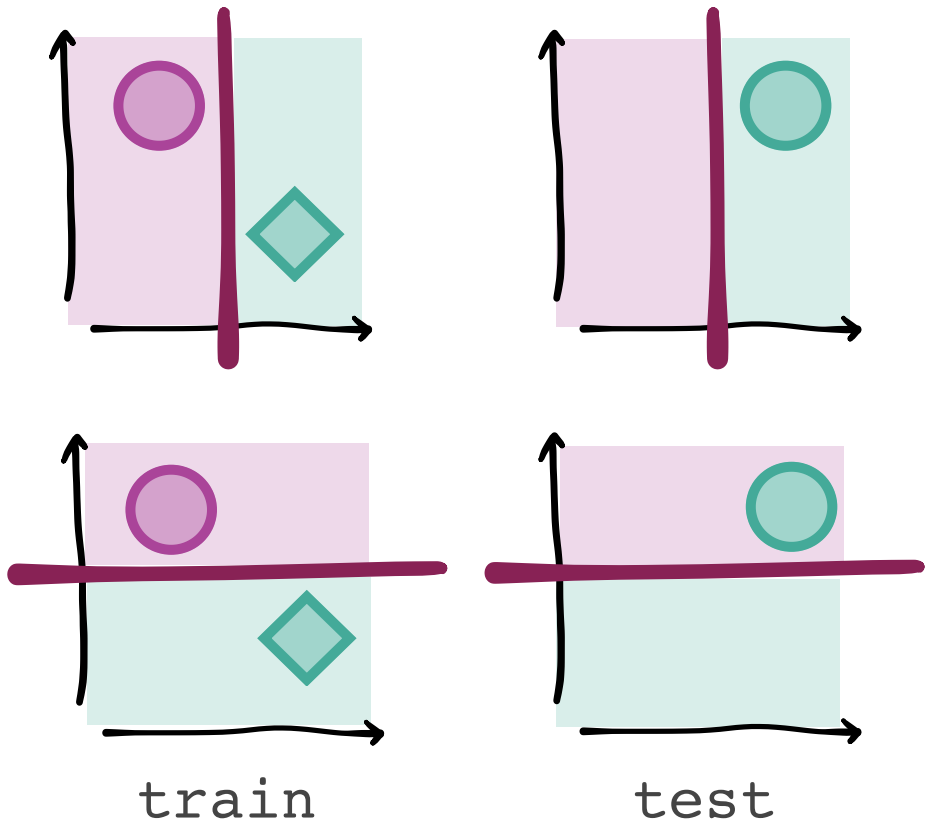
**shape** bias:



"**Feature-level bias**" (**FLB**):
which equally predictive feature?

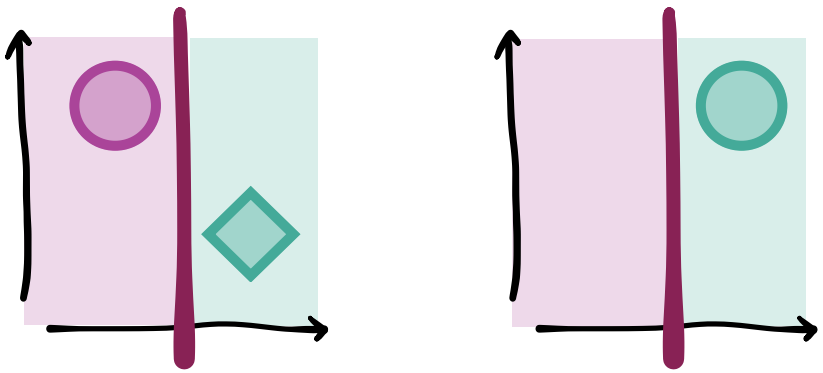
**shape** bias:

**color** bias:

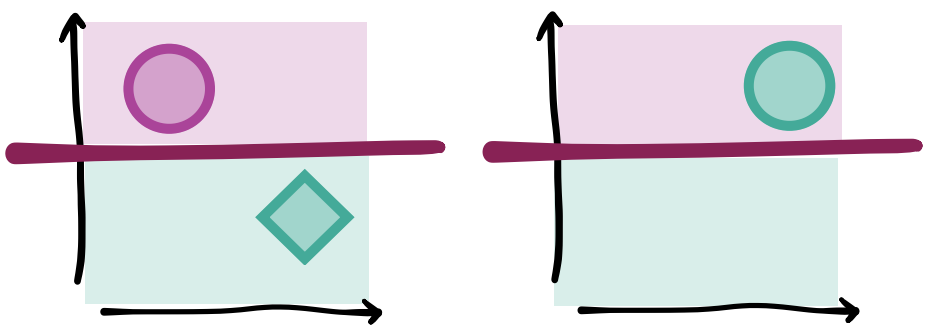


"**Feature-level bias**" (**FLB**):
which equally predictive feature?

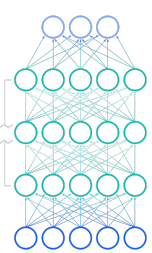
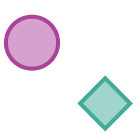
**shape** bias:

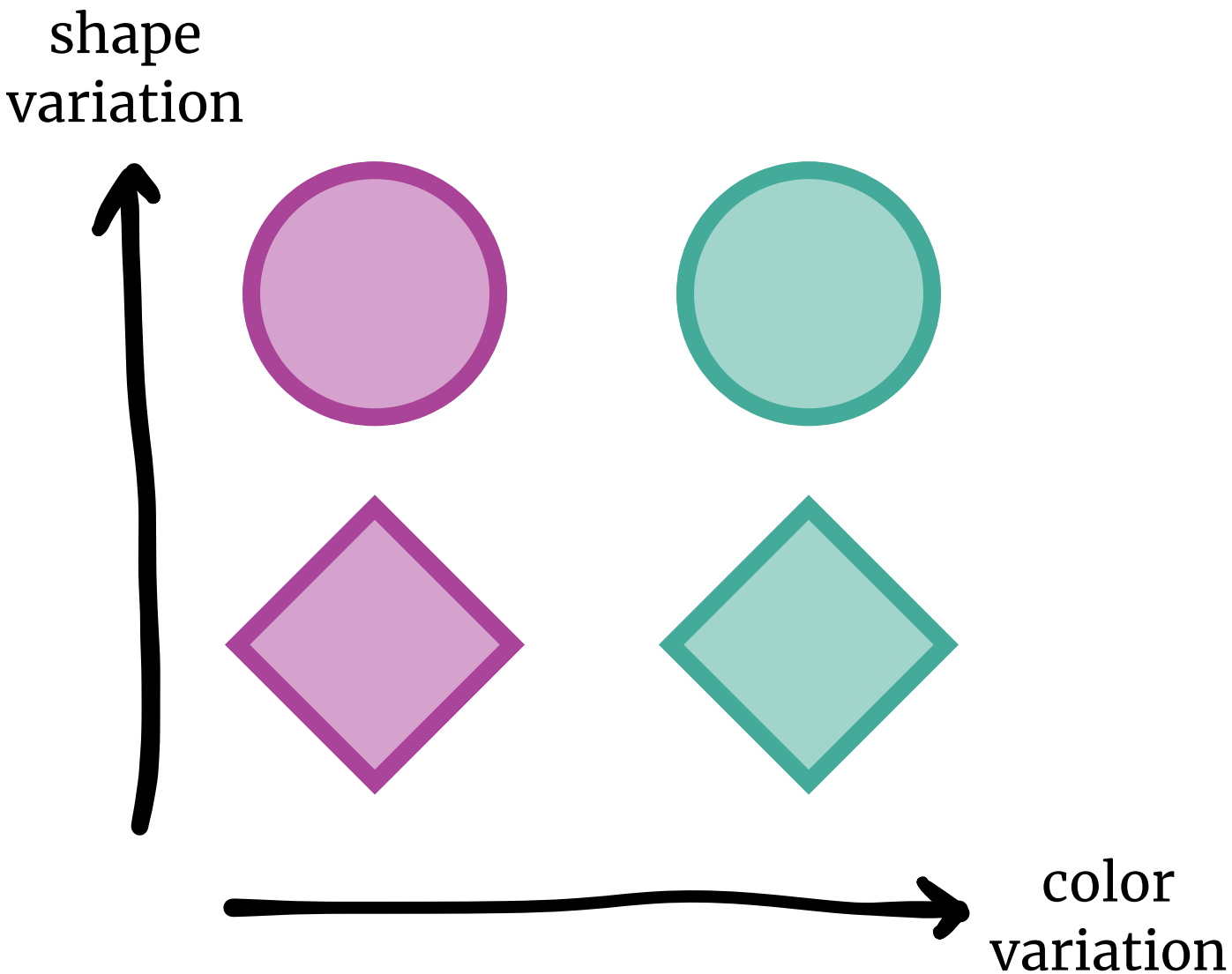


**color** bias:



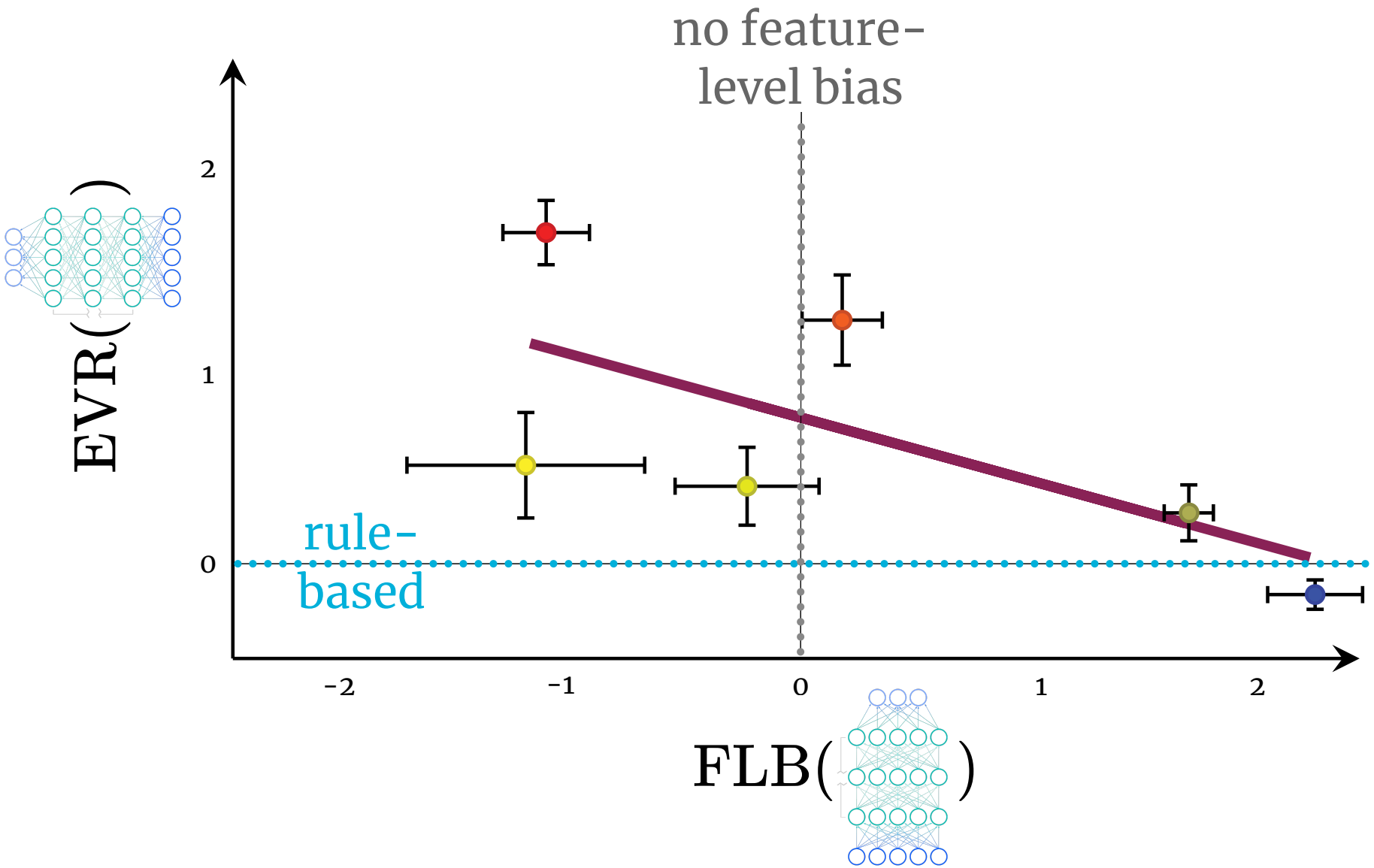
train test

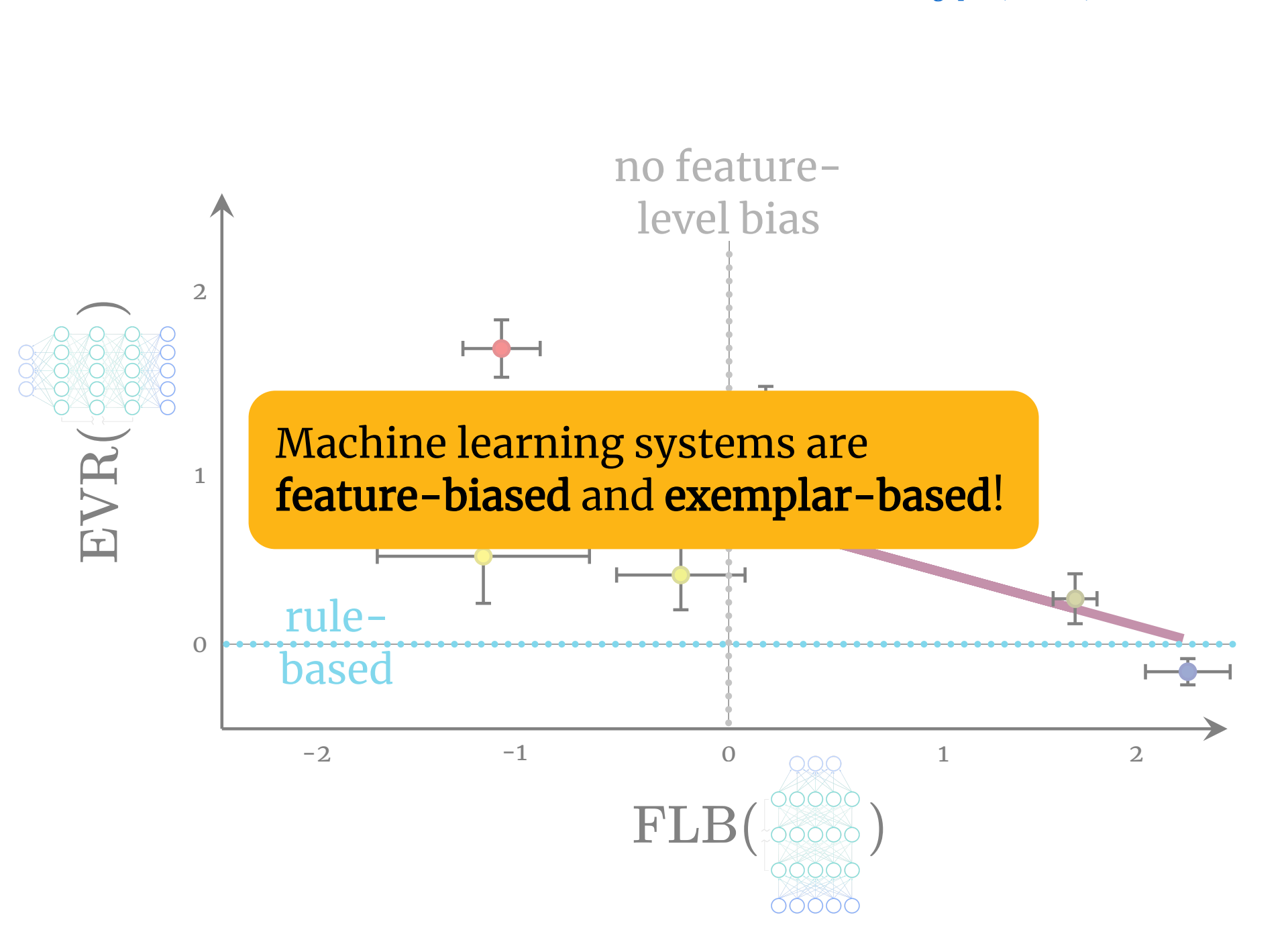
$$\text{FLB}(\text{  }) = \mathbb{E}\left[\text{accuracy}(\text{  })\right] - 0.5$$


shape
variation
color
variation

mouth open?

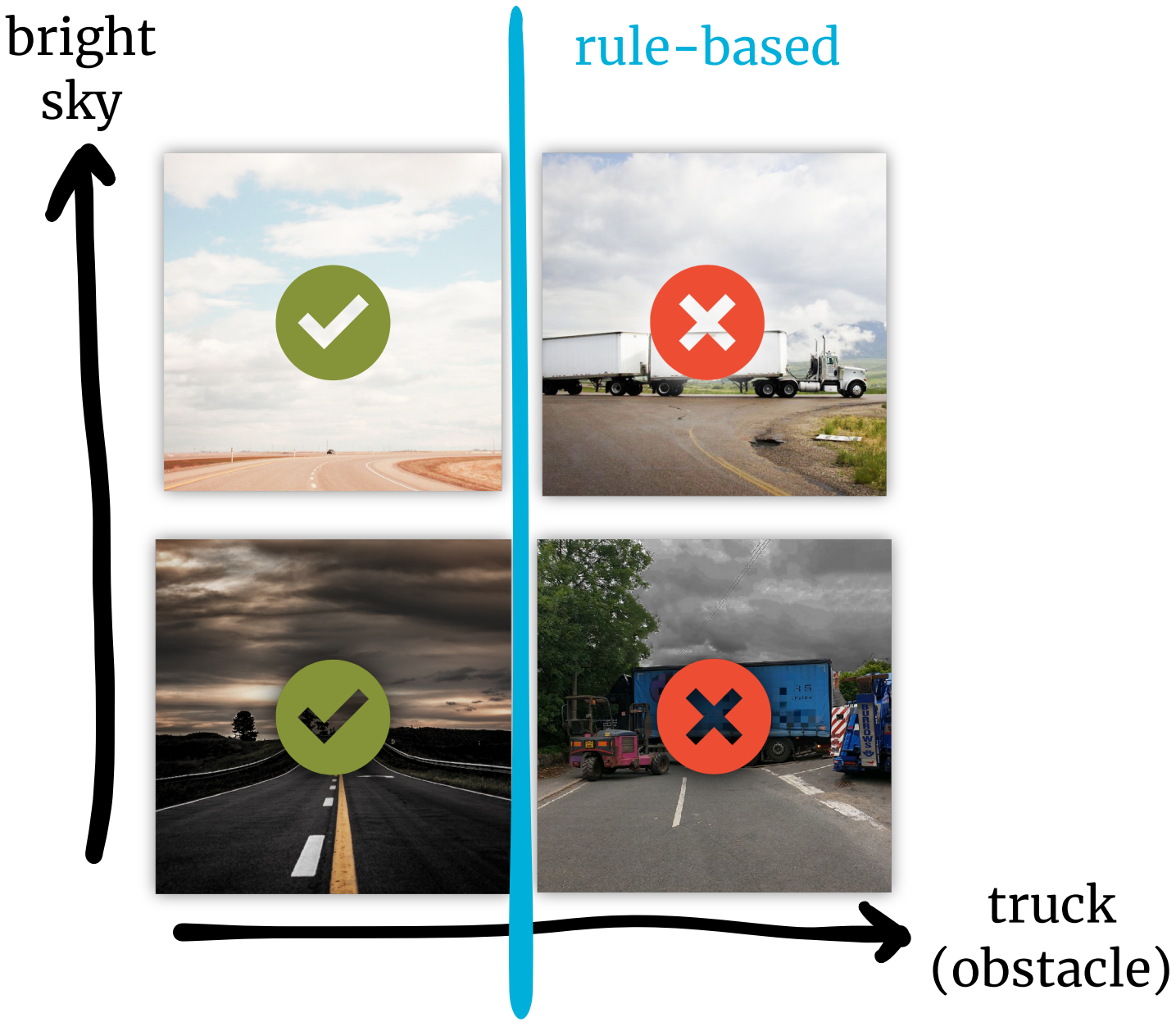

wearing lipstick?

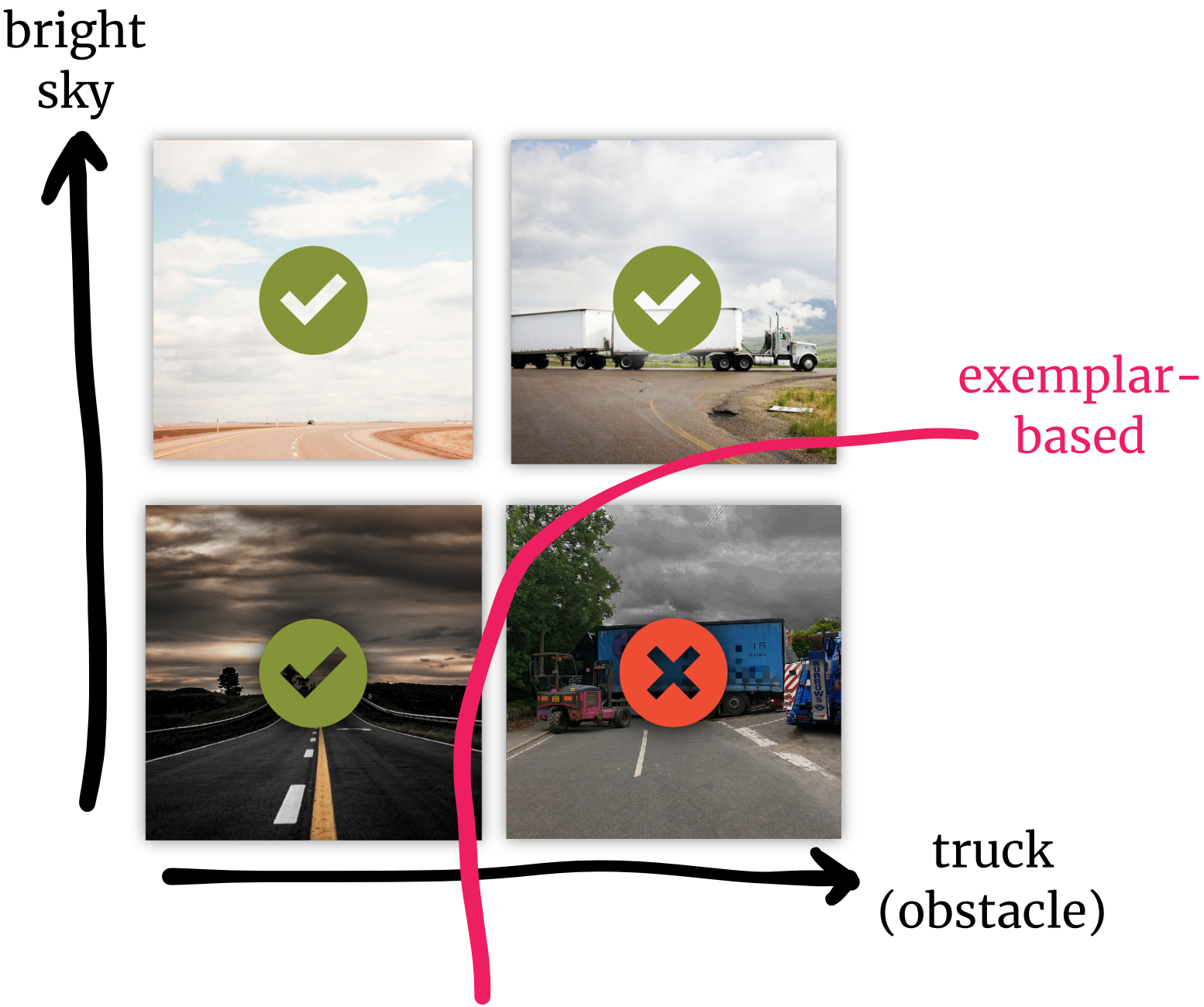

no feature-level bias
EVR( )
2
1
0
rule-based
−2
−1
0
1
2
FLB( )

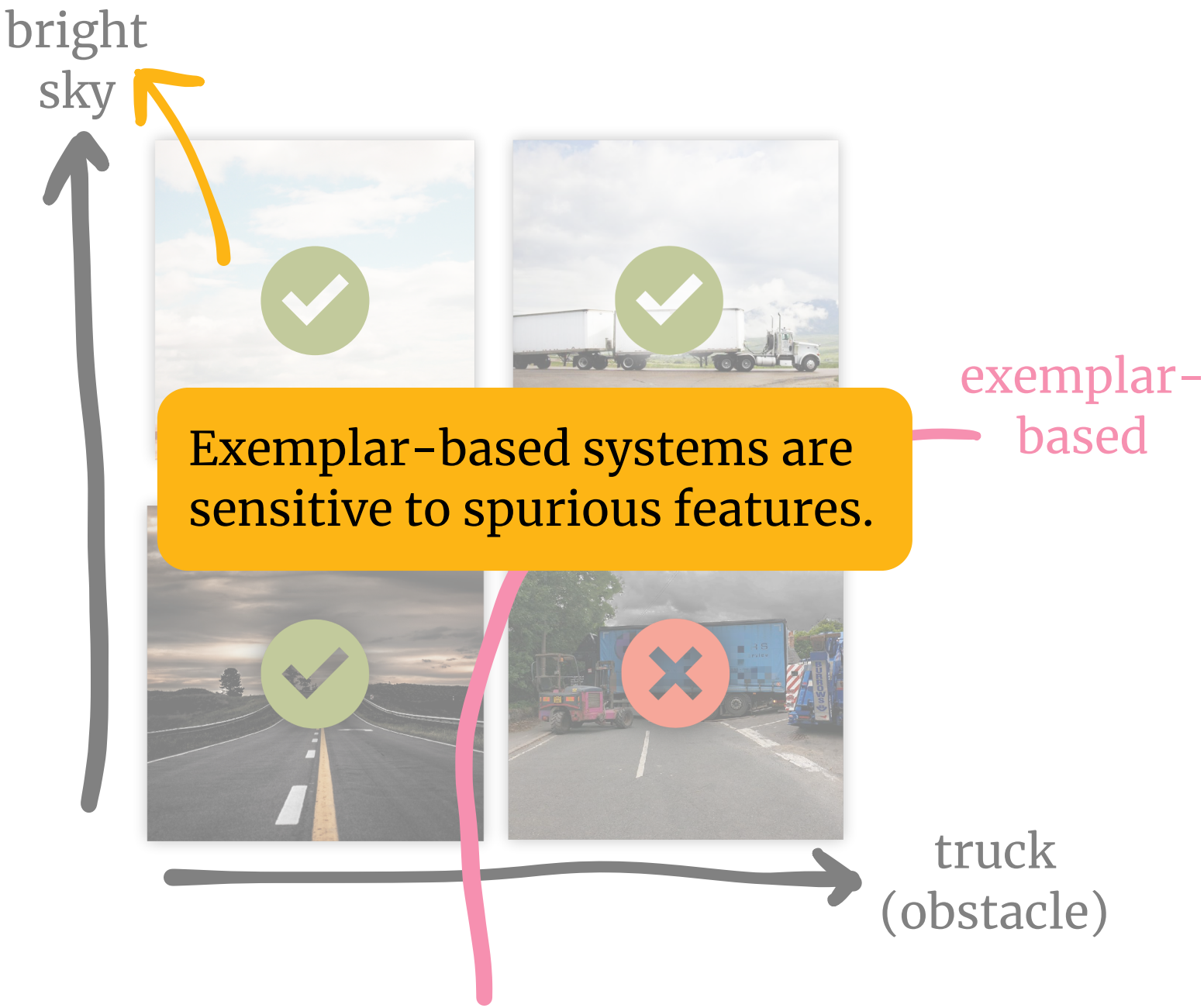

no feature-level bias
EVR( )
2
1
0
rule-based
Machine learning systems are **feature-biased** and **exemplar-based**!
−2
−1
0
1
2
FLB( )

bright
sky

truck
(obstacle)


bright
sky
rule-based
truck
(obstacle)


bright
sky
exemplar-
based
truck
(obstacle)


bright sky
exemplar-based
Exemplar-based systems are sensitive to spurious features.
truck (obstacle)

# More in the paper...

We demonstrate that neural networks are **feature-biased** and **exemplar-based** in various settings (2D points-in-plane, text, image).

We make normative statements about when a model should be **rule-based or exemplar-based (compositional generalization and long-tailed distributions, resp.).**

# More broadly...

We leave it to future work to understand what components of deep learning systems control **feature-level bias** and **rule/exemplar bias**.

We contribute to "cognitive science" for ML.

Ishita Dasgupta <idg@deepmind.com> ,
Erin Grant <eringrant@berkeley.edu>