CLIP

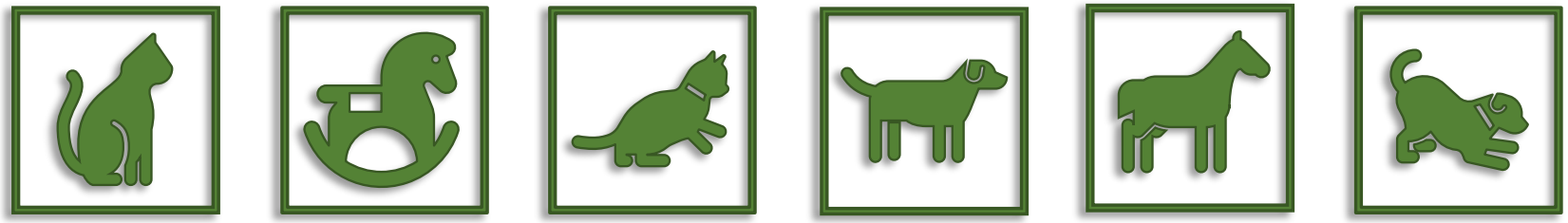
Image classification with
Natural Language Supervision

Michał Gozdera

gozdera.michal@gmail.com

linkedin.com/in/gozderam

Our task

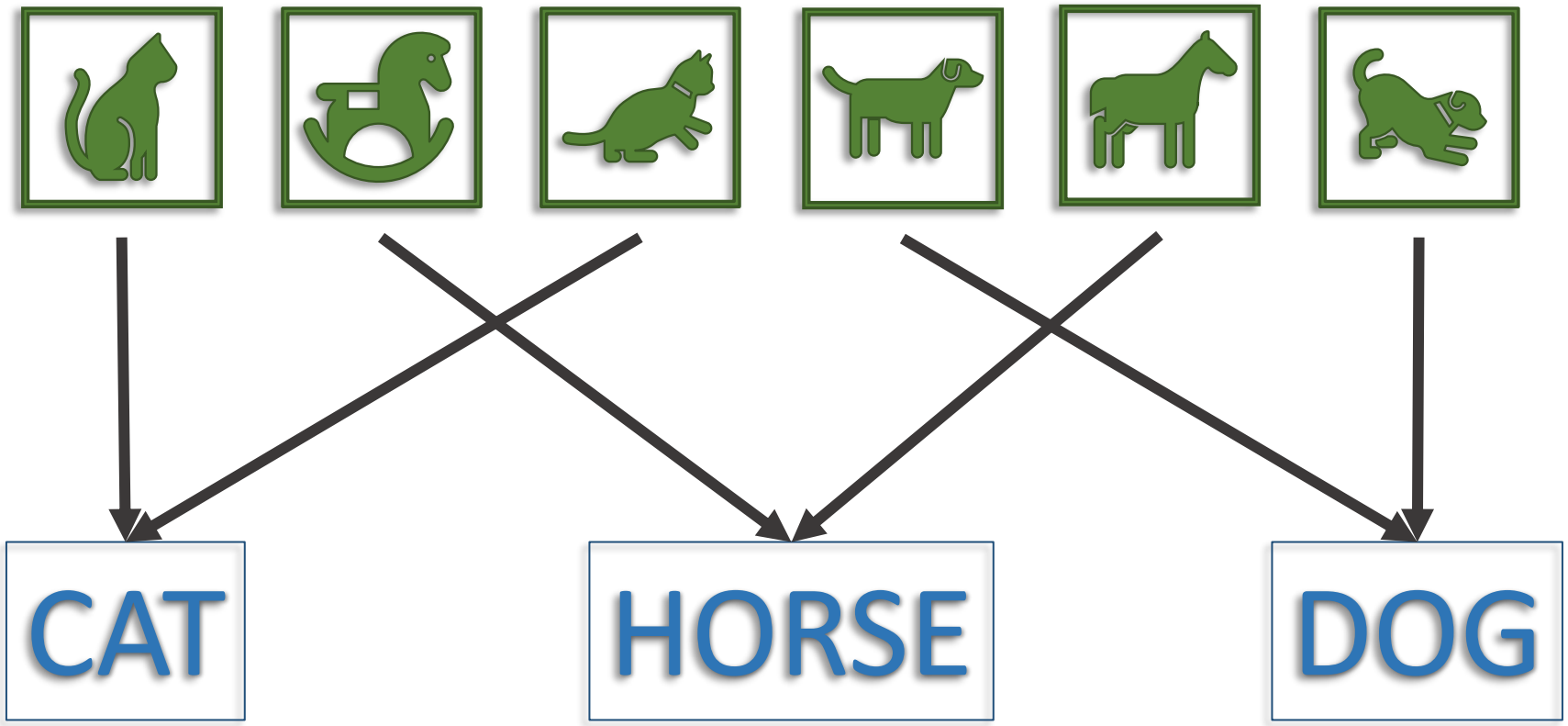


CAT

HORSE

DOG

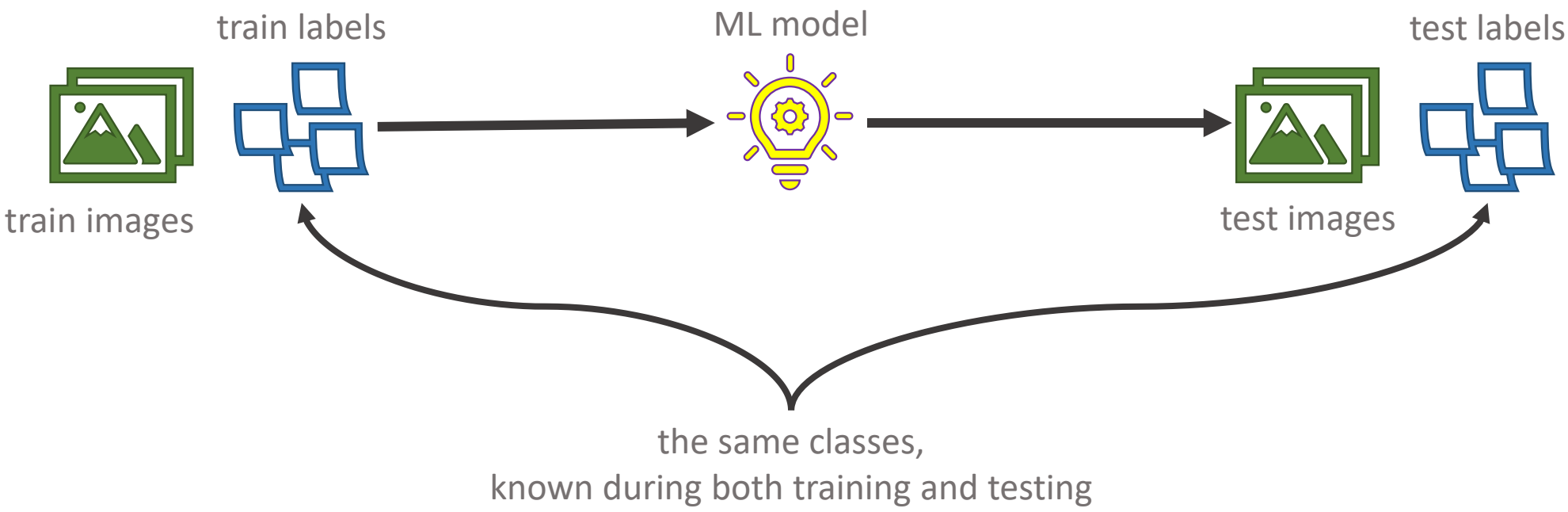
Our task



Our task

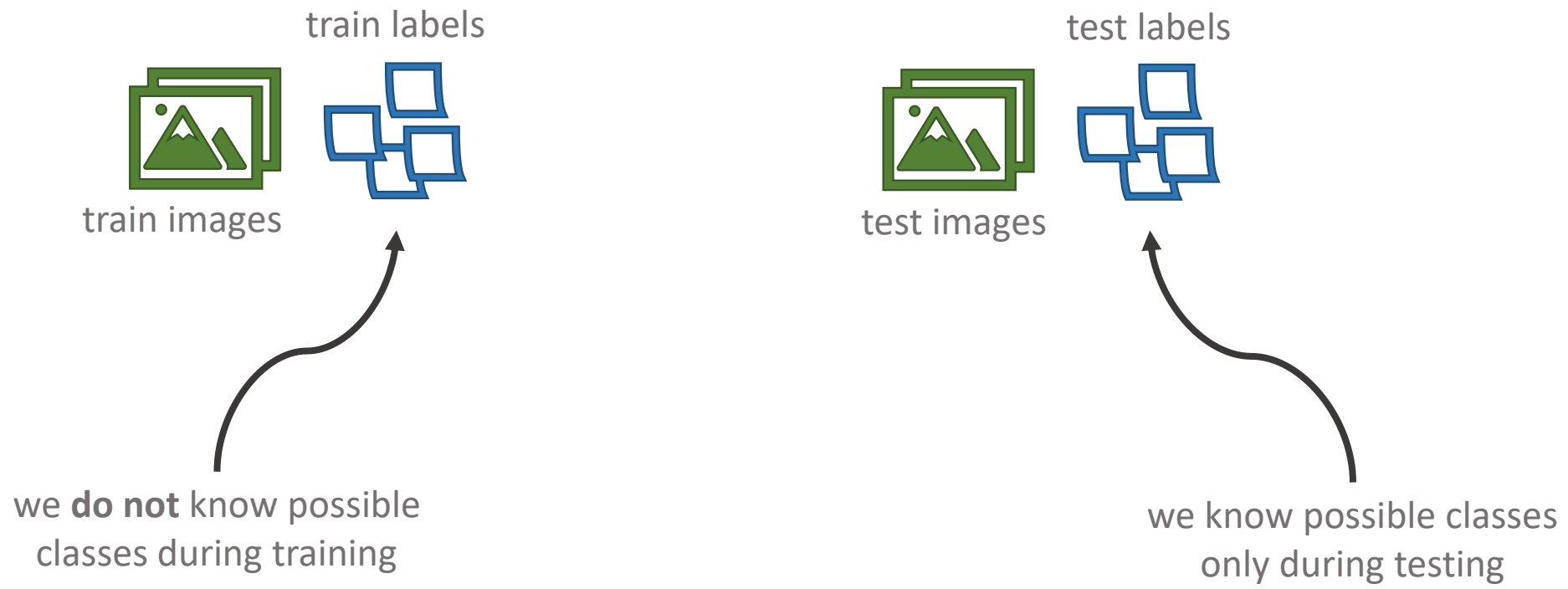
Where is Natural Language Processing?

Traditional approach

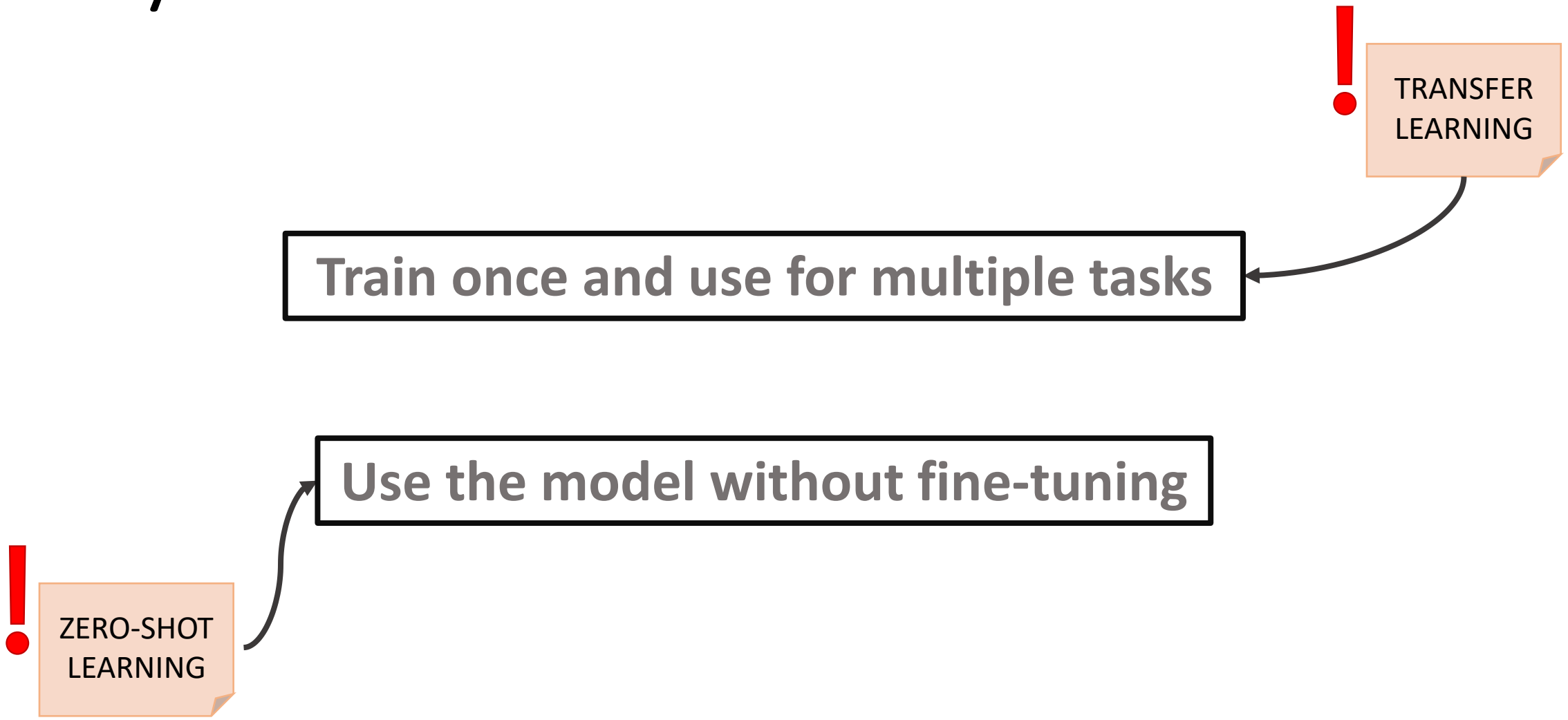


What if…

… we would like to build a model capable of classifying any class?



Why such a model can be useful?



How to build such a model?

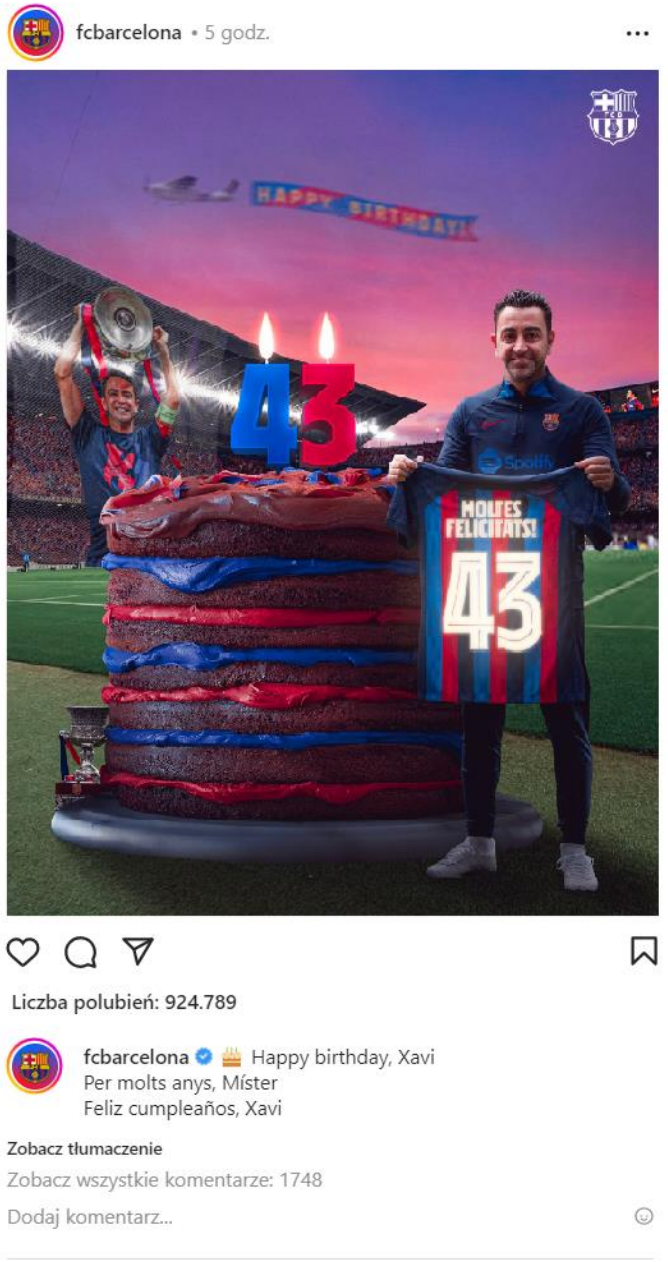
- Want to predict *any* class
  → need huge amount of data

How to build such a model?

- Want to predict *any* class
  → need huge amount of data
- Building *traditional* (image, label) dataset → very costly (manual labeling)

How to build such a model?

- Want to predict *any* class
  → need huge amount of data
- Building *traditional* (image, label) dataset → very costly (manual labeling)
- IDEA?
  Gather data from Internet (Instagram etc.) in the form of (image, text description)



Natural Language supervision

NATURAL LANGUAGE SUPERVISION

Train data in *traditional* approach

images

labels

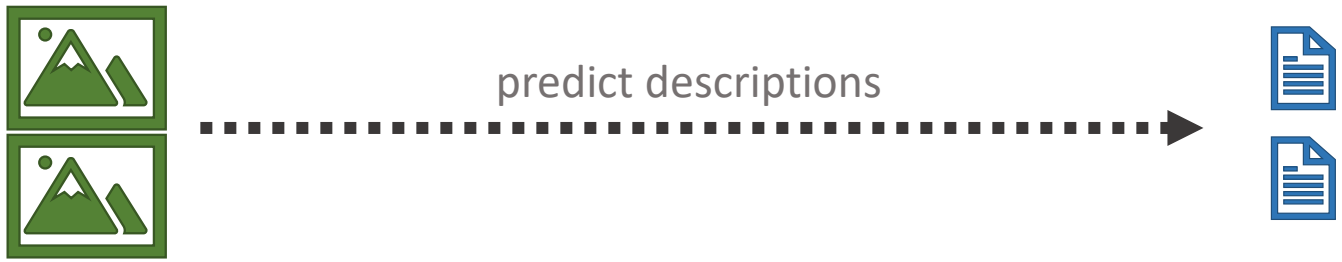
Train data in NL supervision approach

images

text descriptions

What predictions should the model make?

Option 1



What predictions should the model make?

Option 1

predict descriptions

Too hard to train.

What predictions should the model make?

Option 1

predict descriptions

Too hard to train.

Option 2

predict bag-of-words

What predictions should the model make?

Option 1

predict descriptions

Too hard to train.

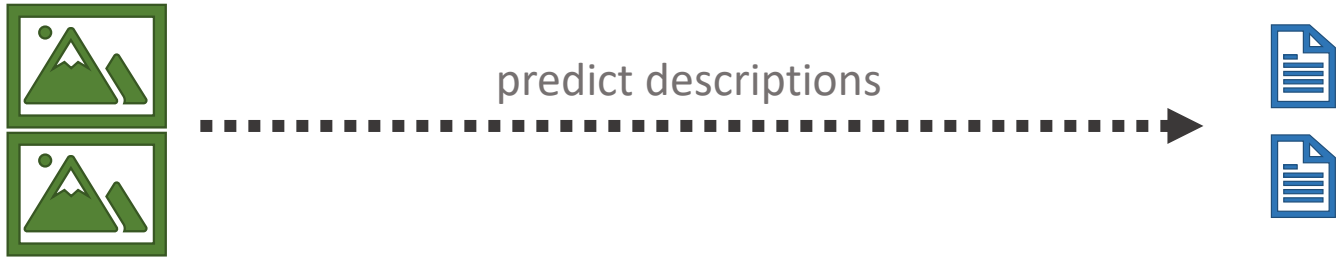
Option 2

predict bag-of-words

Too hard to train.

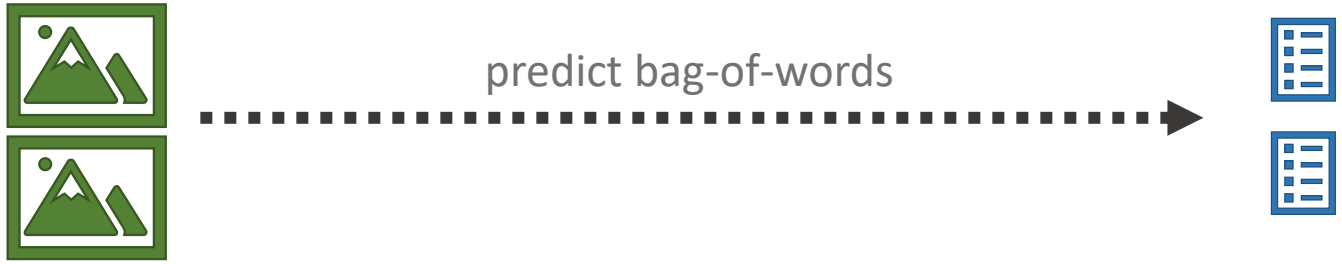
What predictions should the model make?

Option 1



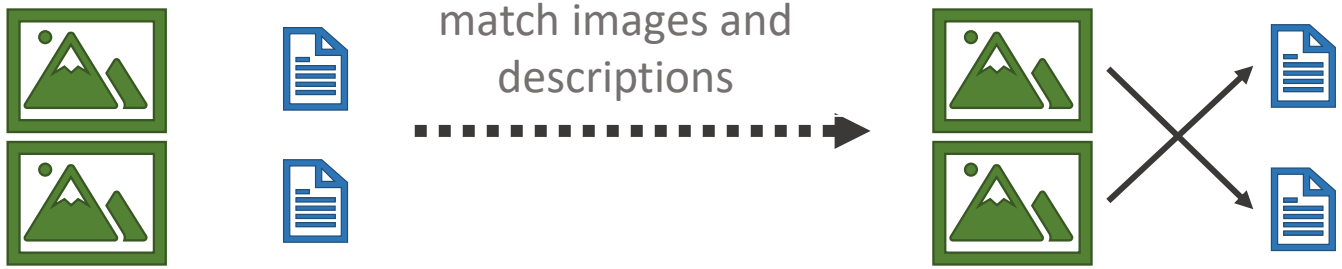
Too hard to train.

Option 2

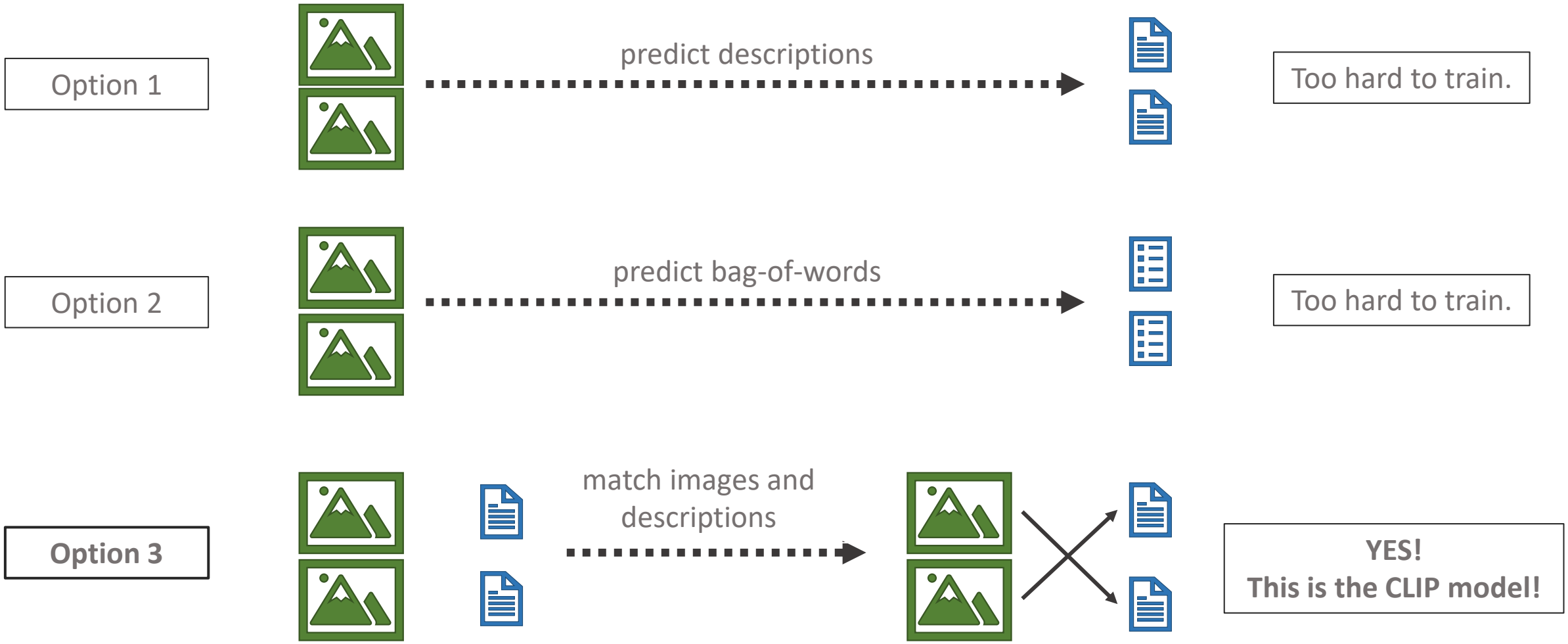


Too hard to train.

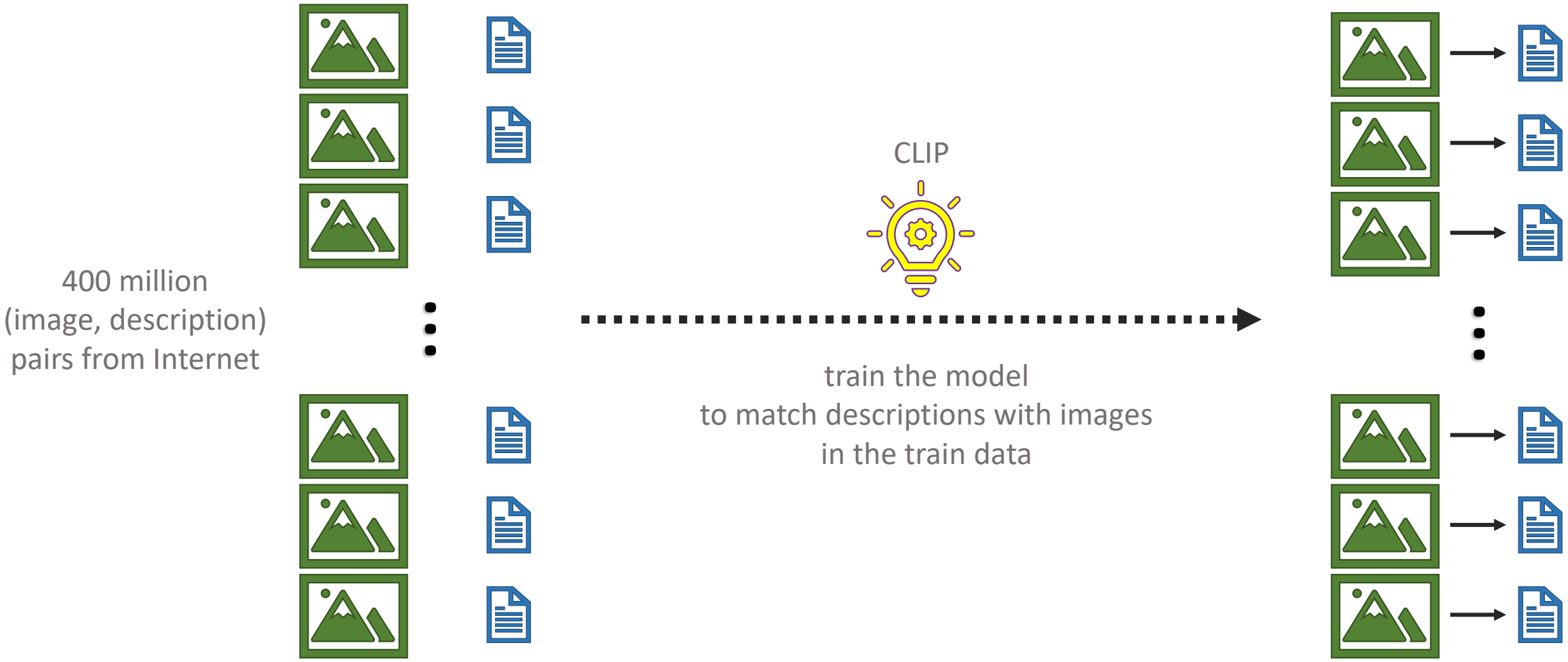
Option 3



What predictions should the model make?



CLIP – training



CLIP – testing

How to predict a label for an image when model can only match images with their descriptions?

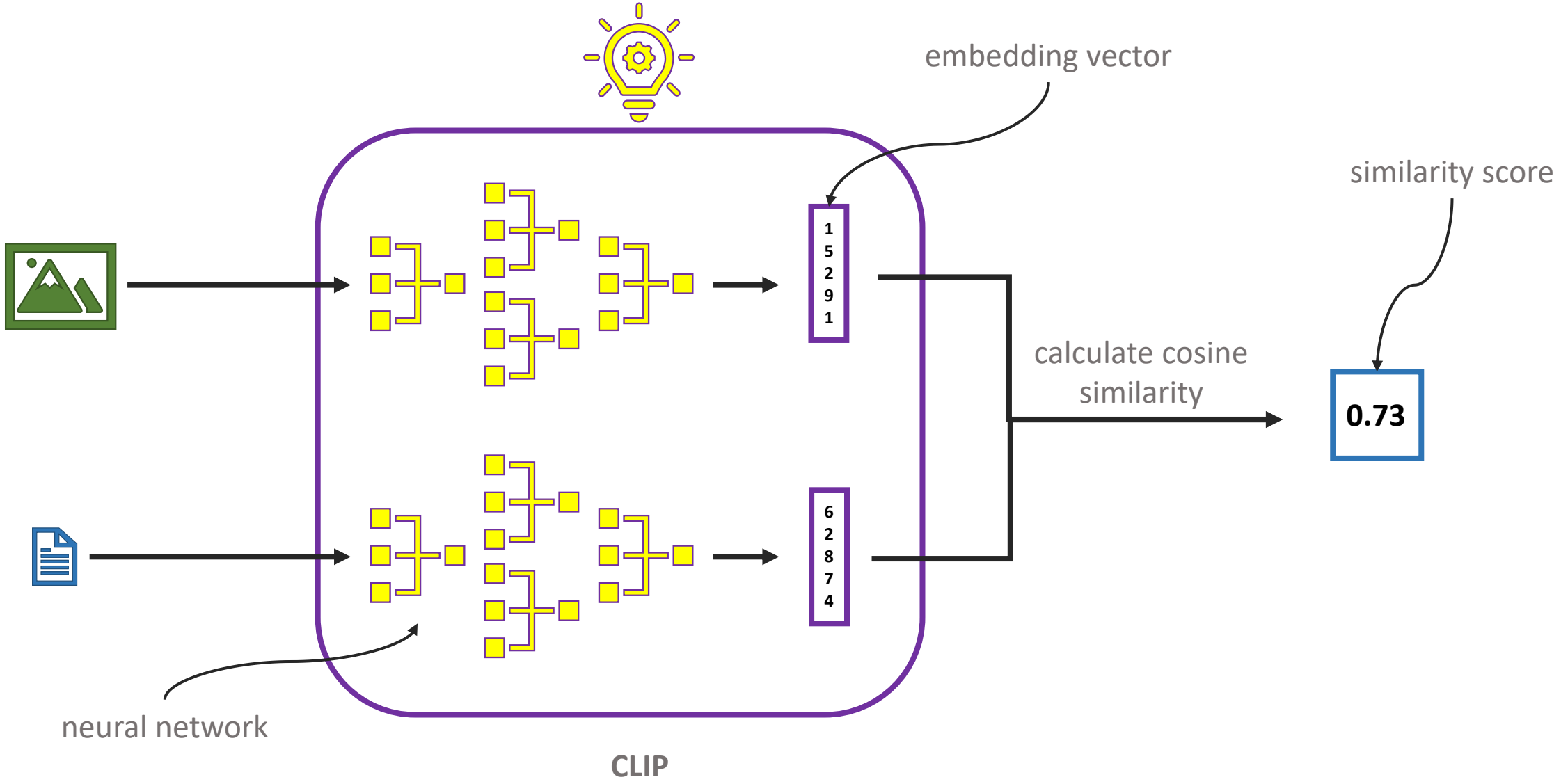
CLIP – testing

How to predict a label for an image when model can only match images with their descriptions?

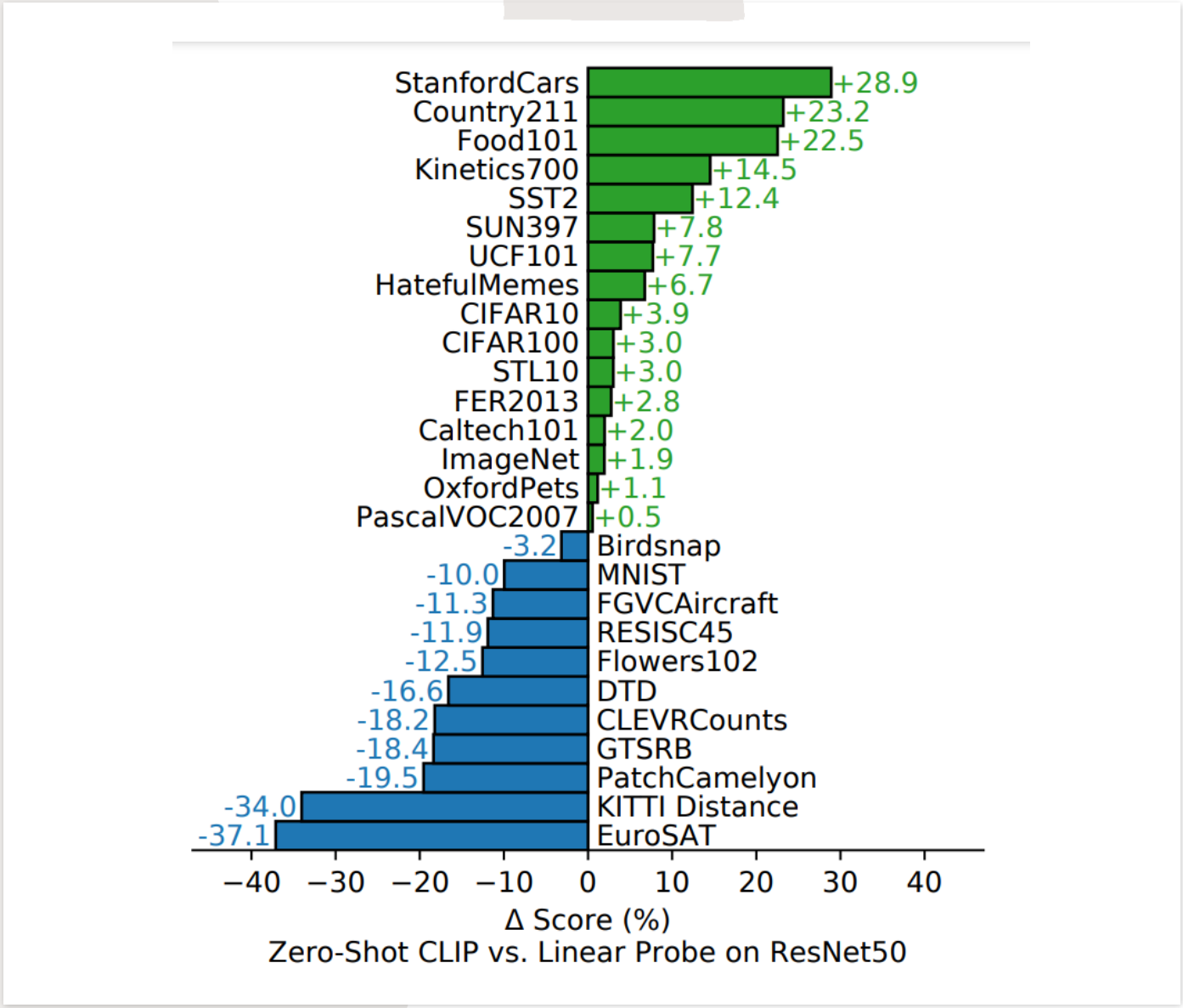
Treat labels as text descriptions, for example:

cat → This picture shows a cat.

CLIP architecture



CLIP classification results



Thank you for your attention!

Let’s move to the quiz now: **kahoot.it**

And then to the workshop: **github.com/gozderamichal/data_literacy**

Based on:

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:8748-8763, 2021