Automated Image and Video Analysis: A practical introduction

Felicia Loecherbach

Automated Image and Video Data Analysis with Python
OPINION Training School Elbasan
May 29, 2025

Giving credit

- A lot (if not most) of the material I will be using for this workshop has been developed originally by Andreu Casas and Nora Webb Williams who came up with this course a few years ago
- Especially for the earlier parts (except multimodal modelling) they have a great book explaining things more in detail (**Images as Data for Social Science Research**) as well as a special issue "Images as Data" in Computational Communication Research (2022) with many interesting examples of how these methods can be used in the social sciences
- I will occasionally show results and insights into research projects that are not published yet, please treat them confidentially

Why do Images Matter?

People are more likely to pay attention to visuals

IMMIGRATION

How America Got to 'Zero Tolerance' on Immigration

Battles have raged within the White House over family separations, ICE raids and President Trump's obsession with a wall.

Together, they have remade homeland security.

15m ago 393 comments

Mr. Trump's approach follows a model from Europe and Australia, our Interpreter columnists write.

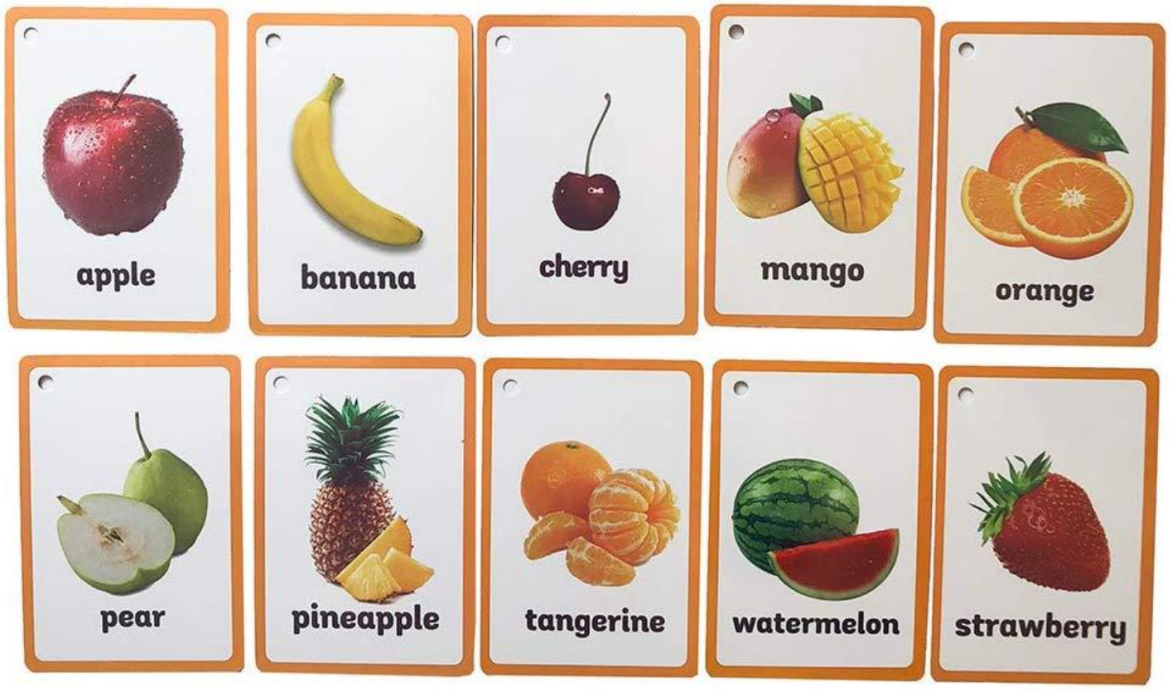
3h ago

Kirsten Luce

Dahmen (2012) "*Photographic Framing in the Stem Cell Debate*"

Why do Images Matter?

People are more likely to recall information learned through visuals



Paivio et al. (1968) "*Why are pictures easier to recall than words?*"

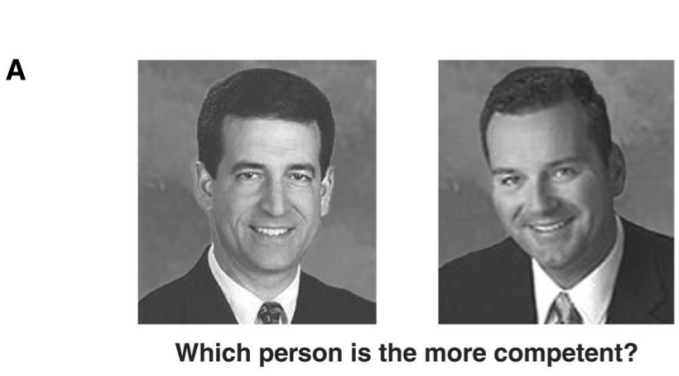
Why do Images Matter?
Visuals evoke stronger emotional reactions

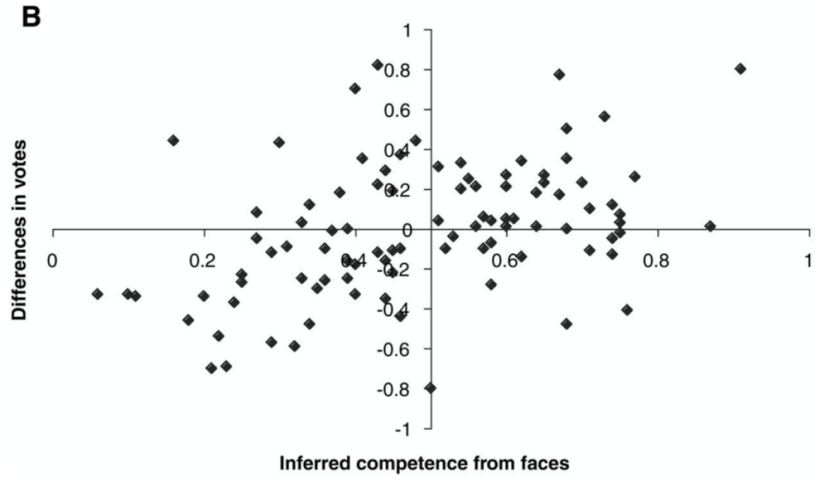


Grabe Bucy (2009) "*Images Bite Politics*"

Why do Images Matter?

Image effects in **politics**: images $\rightarrow$ inference of competence $\rightarrow$ voting

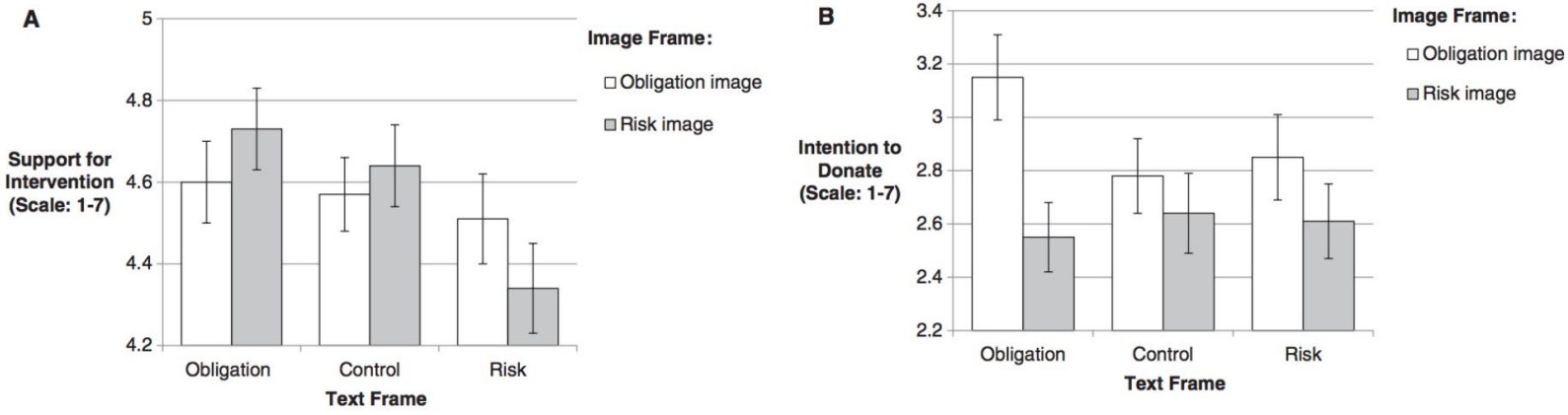




Todorov et al. (2009) "*Inferences of Competence from Faces Predict Election Outcomes*"

Why do Images Matter?

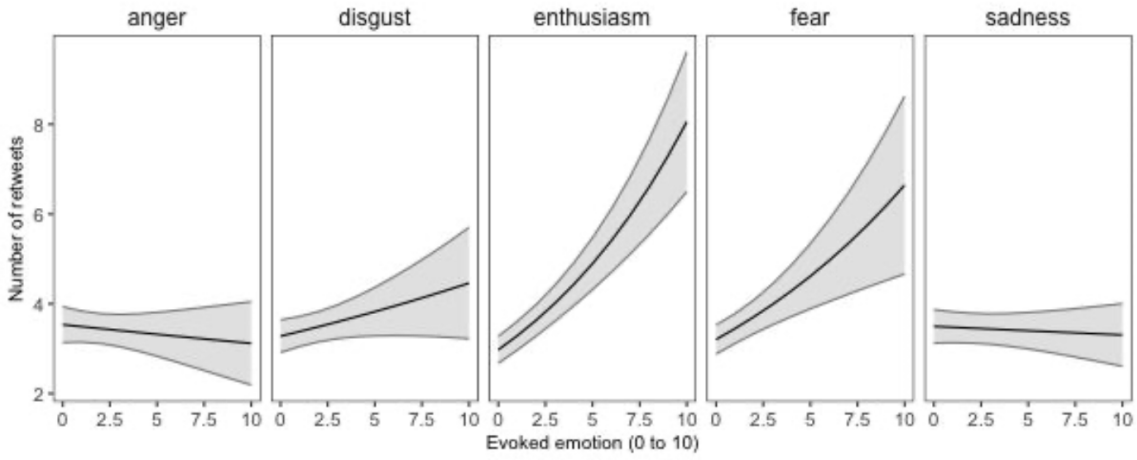
Image effects in **politics**: images $\rightarrow$ framing $\rightarrow$ attitudes



Powell et al. (2015) "A Clearer Picture"

Why do Images Matter?

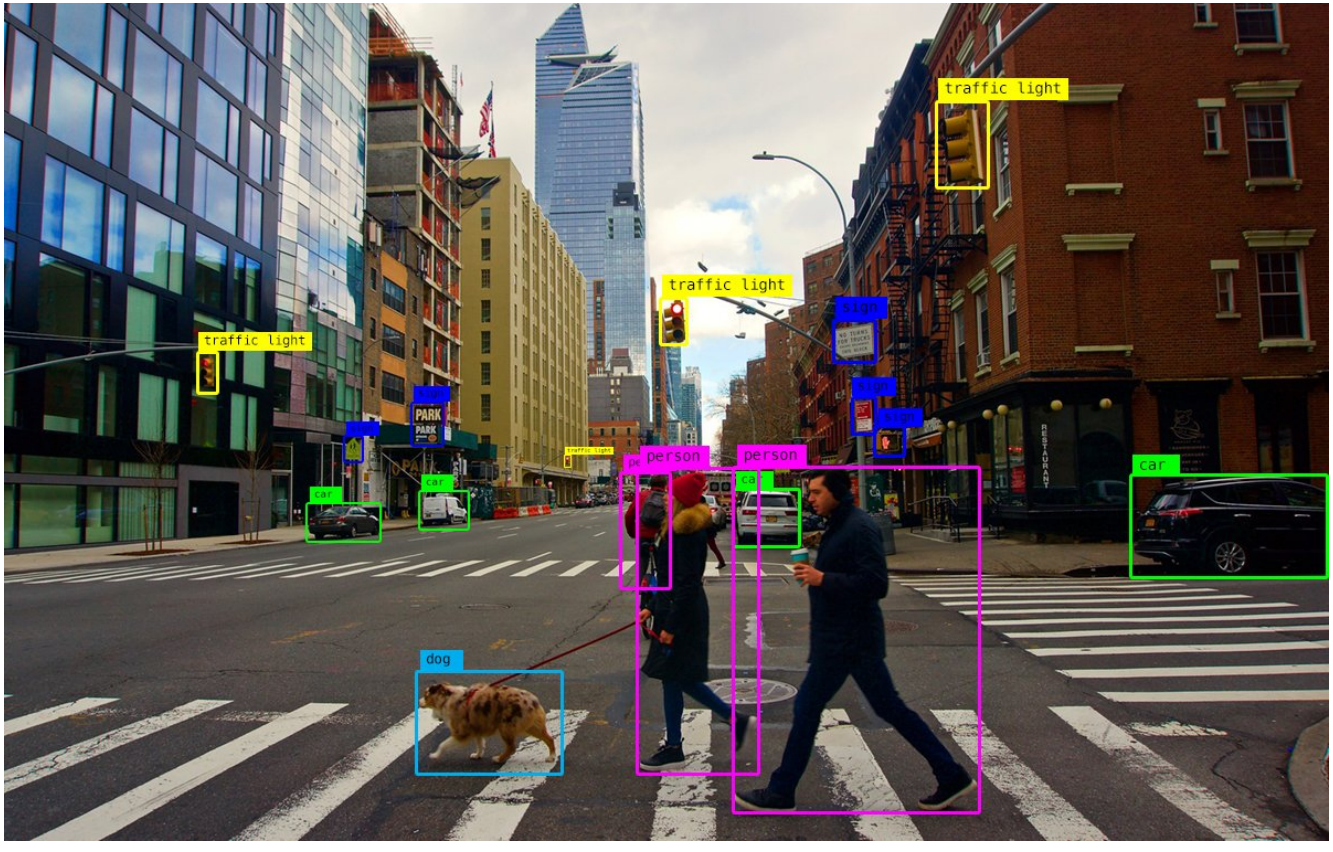
Image effects in **politics**: images $\rightarrow$ emotions $\rightarrow$ **mobilization**



Casas & Webb Williams (201) "Images That Matter"

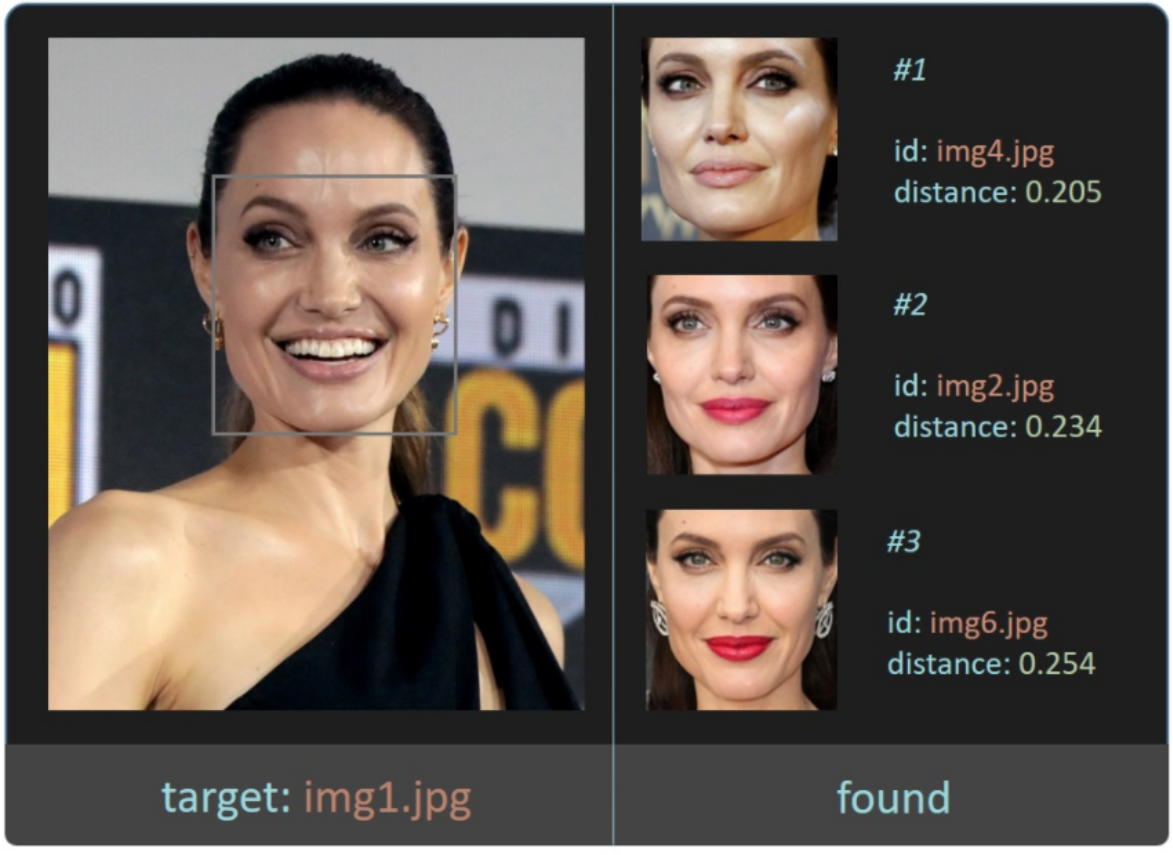
Available Automated Image Analysis Methods

Object detection & recognition



Available Automated Image Analysis Methods

Face detection & recognition



Available Automated Image Analysis Methods

Face analysis

Available Automated Image Analysis Methods

Image Similarity

Available Automated Image Analysis Methods

Unsupervised Clustering

Available Automated Image Analysis Methods

And many others...

- Text extraction (OCR)
- Caption generation
- Sentiment analysis (evoked emotions)
- Visual aesthetics analysis
- etc...

Available Automated Image Analysis Methods

In this workshop we'll focus on...

- Supervised image classification
- Unsupervised image classification
- Multimodal classification (text + image)

Outline

1. What are supervised models useful for?
2. The basics of supervised classification
3. Pre-trained models and zero-shot classification
4. Fine-tuning a pre-trained model for a new supervised task

What are supervised models useful for?
When you know what you're looking for in an image

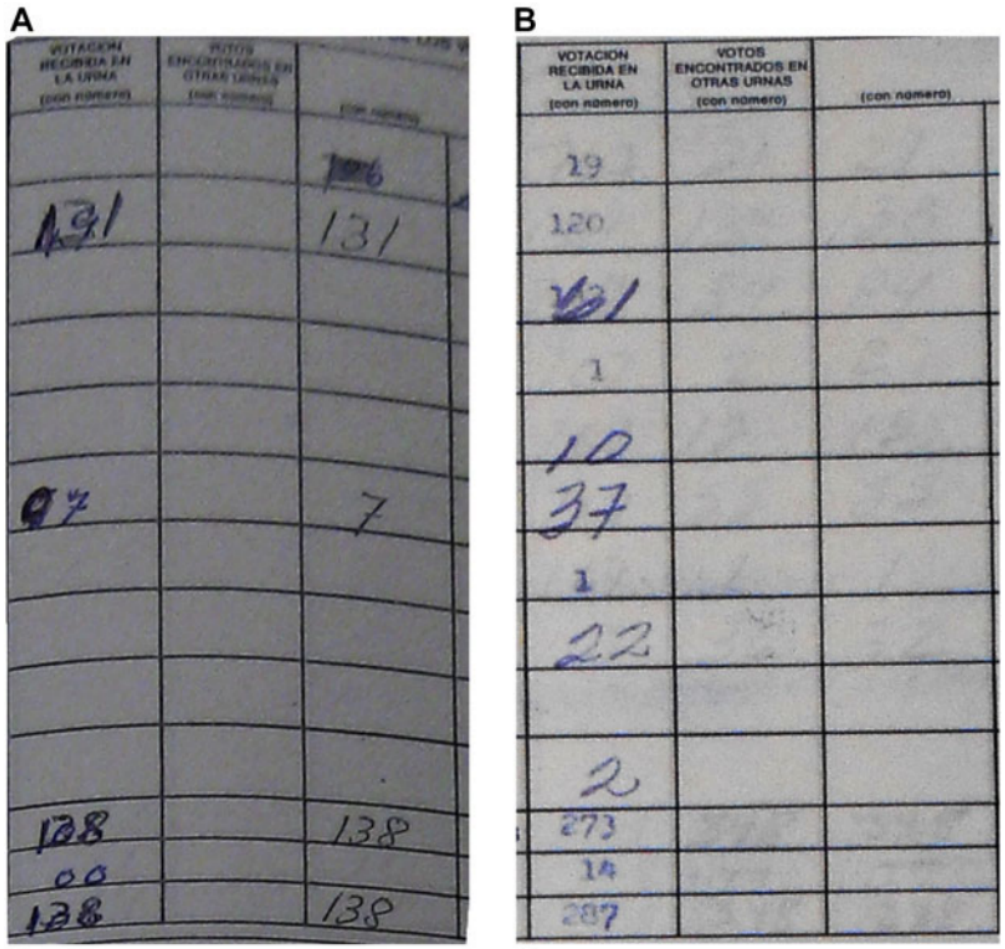
- Binary: is this a picture of a street protest?
- Categorical: is this picture a meme, cartoon, or other?
- Continuous: how much violence is there in the image?

What are supervised models useful for?
Parallelism to text analysis

- Binary: is this a political message/document?
- Categorical: topic: Environment, healthcare, etc.
- Continuous: how negative is the message?

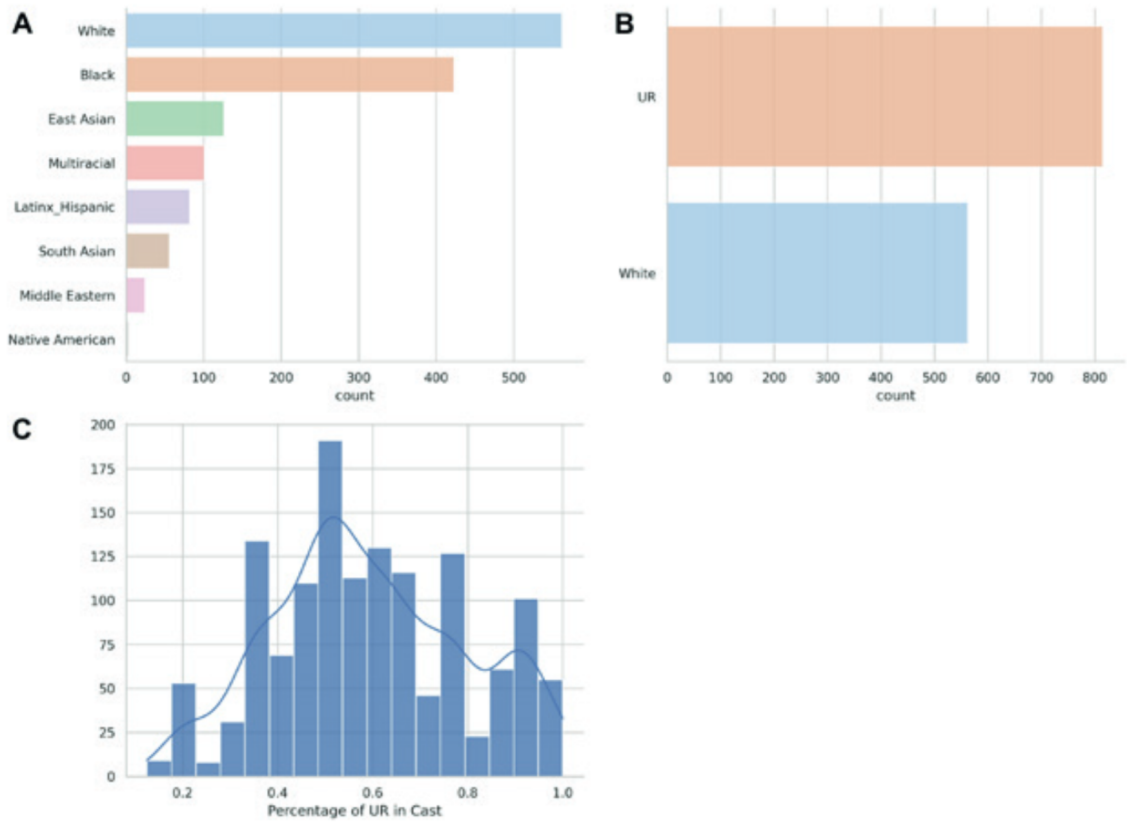
What are supervised models useful for?
Some examples: Binary

Cantú (2019, *APSR*)
Does the vote tally have an alteration?



What are supervised models useful for?
Some examples: Categorical

Malik et al. (2022, *CCR*)
Ethnicity of characters in movies



What are supervised models useful for?
Some examples: Continuous

Steinert-Threlkeld et al. (2022, *JOP*)
How much police/protester violence is there in the image?

(b) State Violence

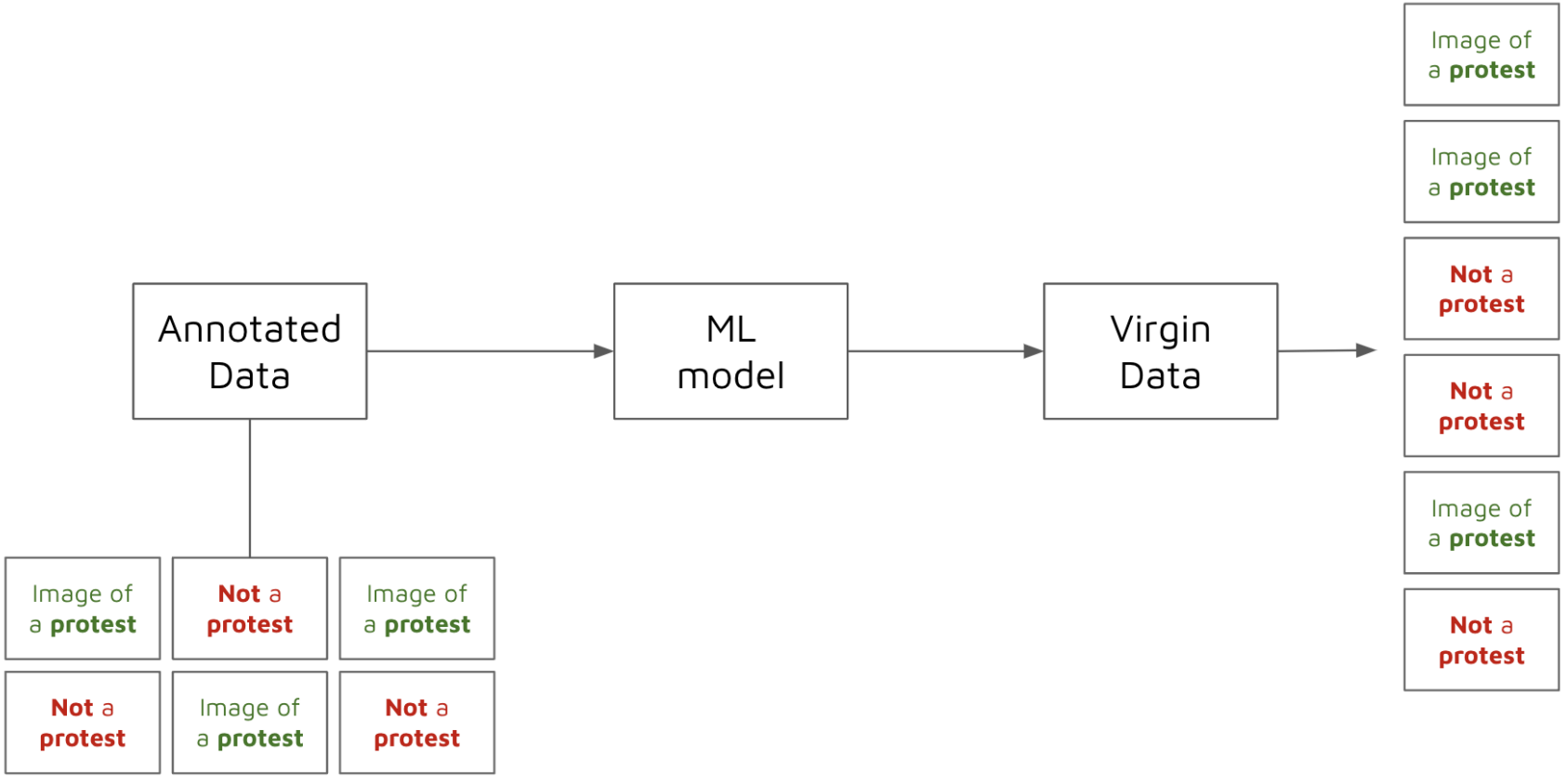
Seoul .031 Hong Kong .145 Barcelona .654 Caracas .849

(c) Protester Violence

Seoul .021 Barcelona .255 Hong Kong .478 Caracas .998

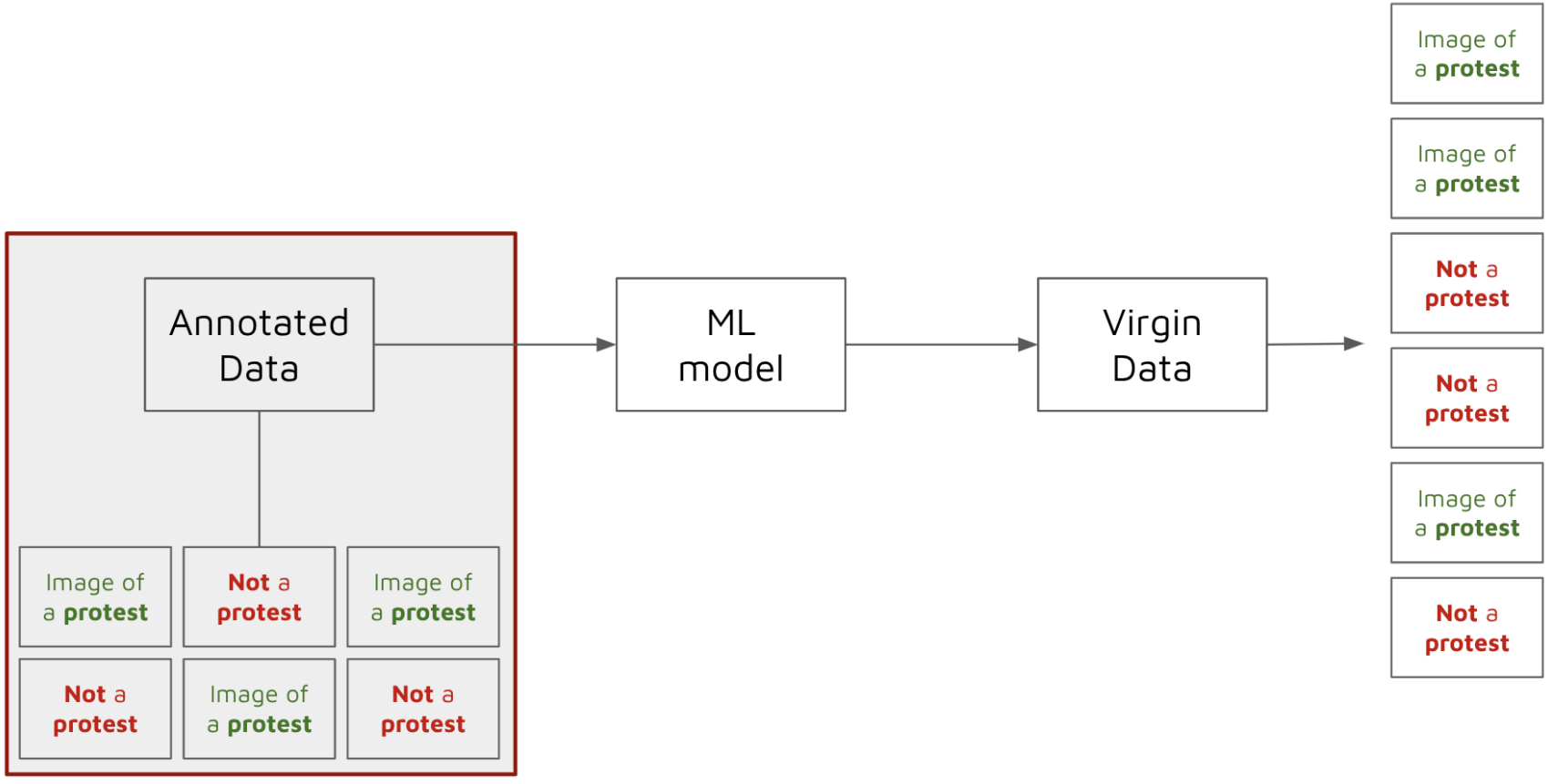
The Basics of Supervised Classification

Basic logic



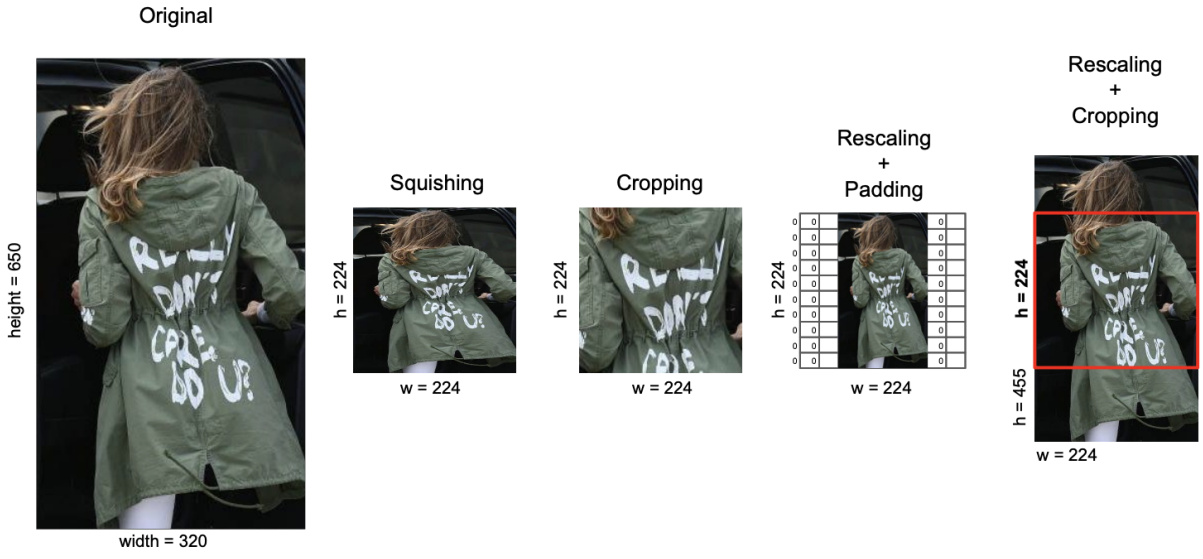
The Basics of Supervised Classification

Practical considerations: Pre-processing



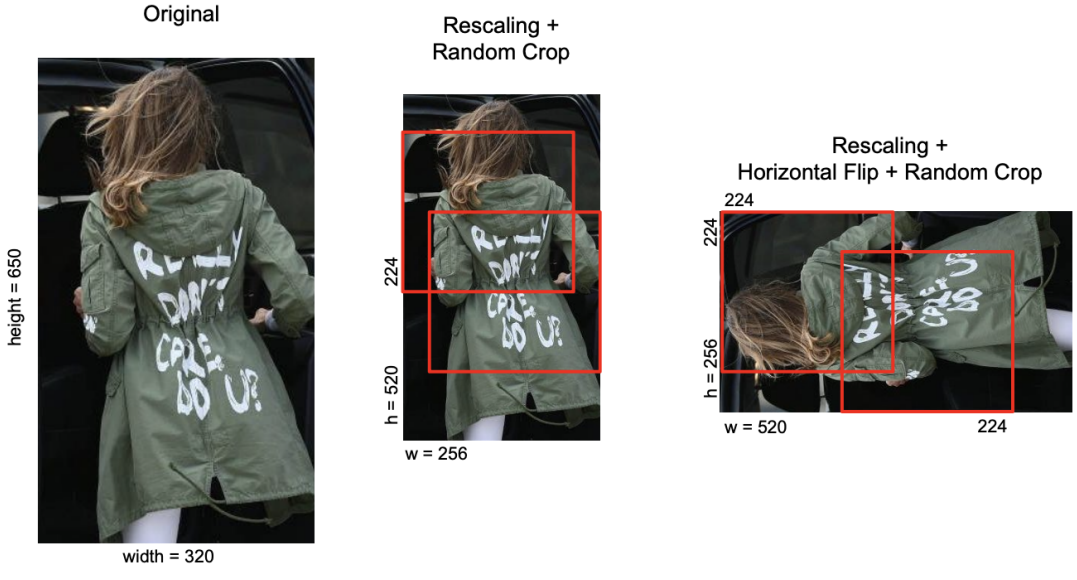
The Basics of Supervised Classification
Practical considerations: Pre-processing

- Input images need to be of the same particular size (224x224 pixels)



The Basics of Supervised Classification
Practical considerations: Pre-processing

- Input images need to be of the same particular size (224x224 pixels)
- Data augmentation



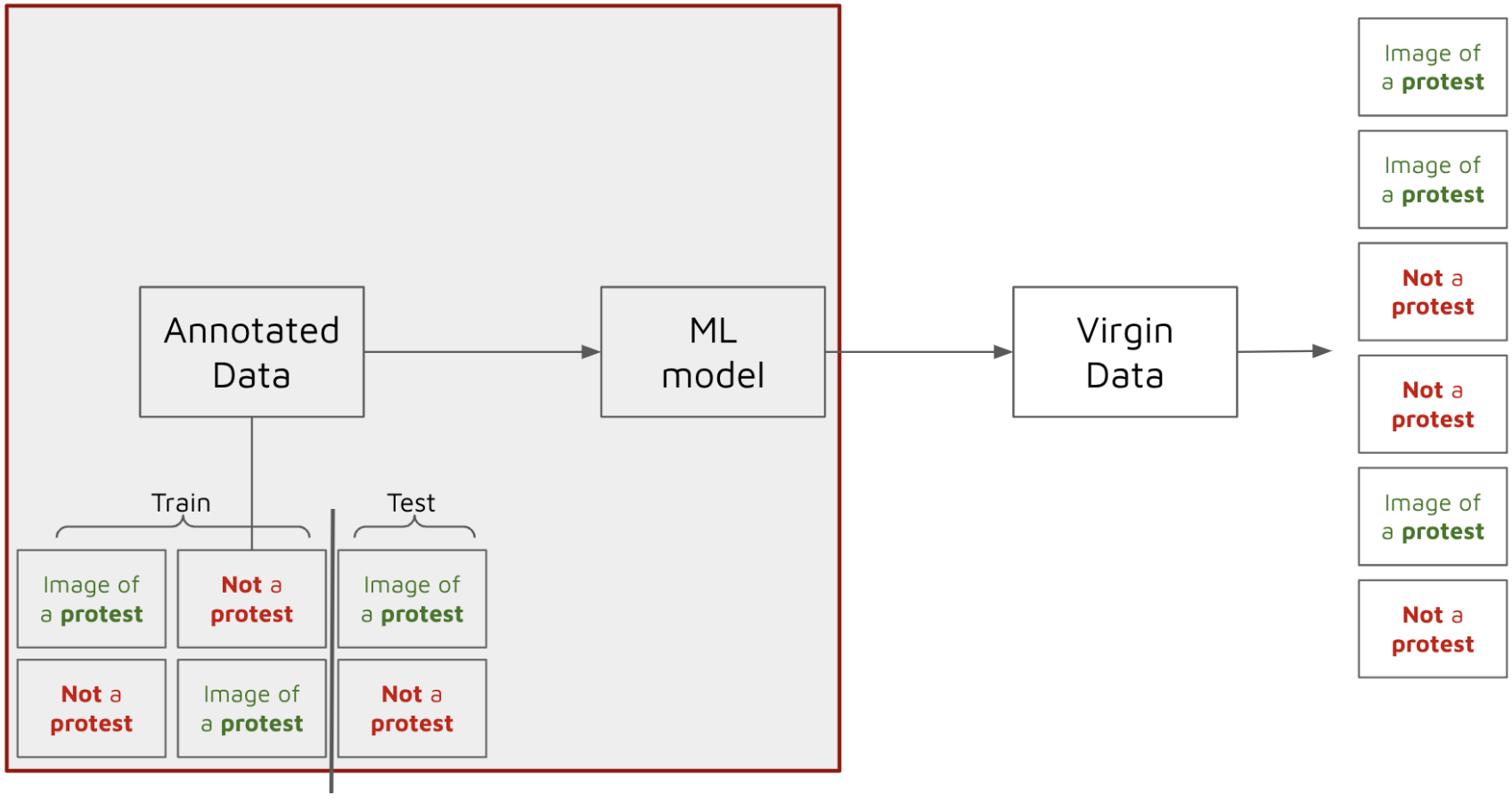
The Basics of Supervised Classification
Practical considerations: Pre-processing

- Input images need to be of the same particular size (224x224 pixels)
- Data augmentation
- Data organization

```
- train
    — class01
    — class02
    — ...
    — classN
- test
    — class01
    — class02
    — ...
    — classN
```

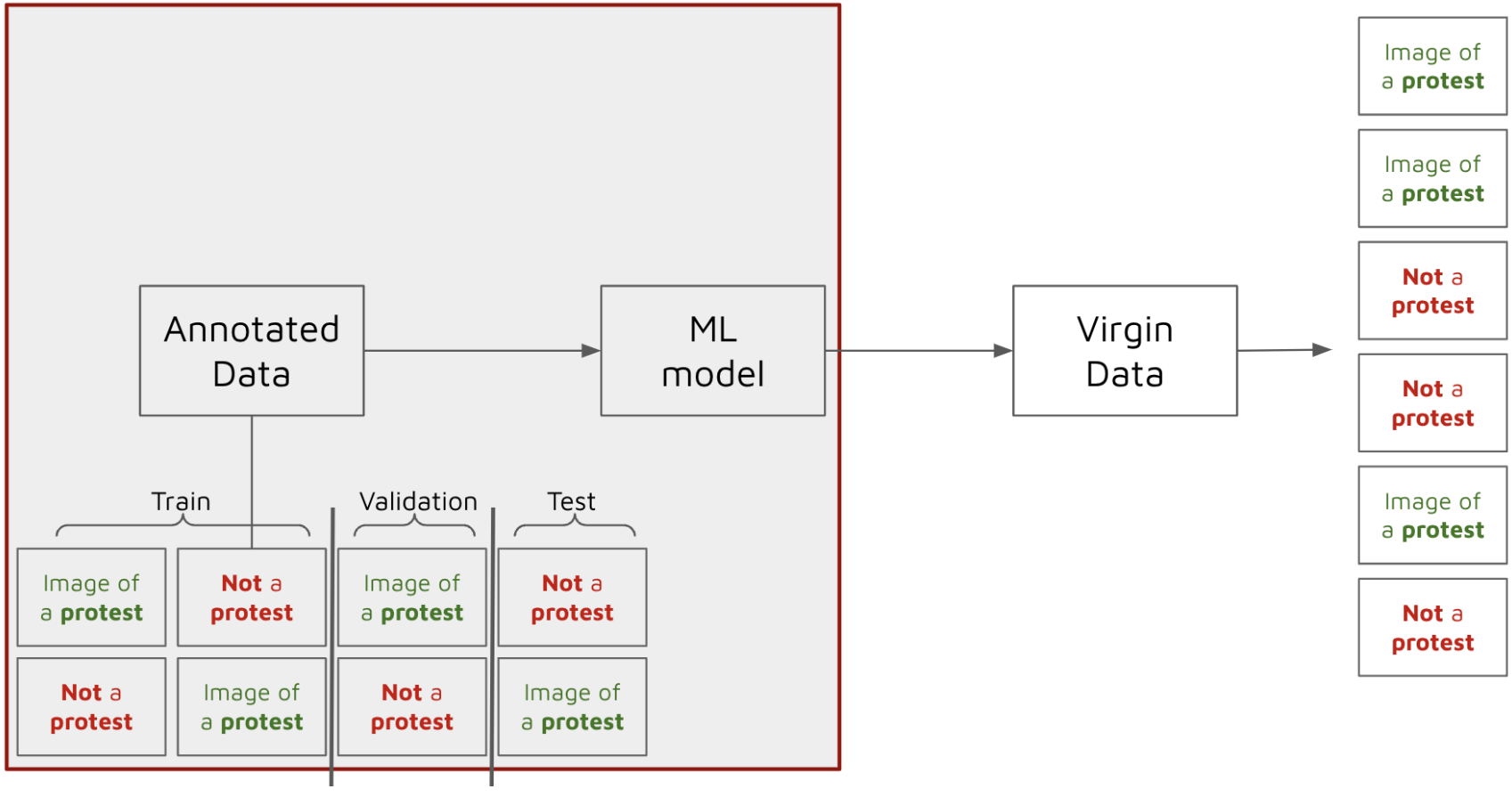
The Basics of Supervised Classification

Training the model



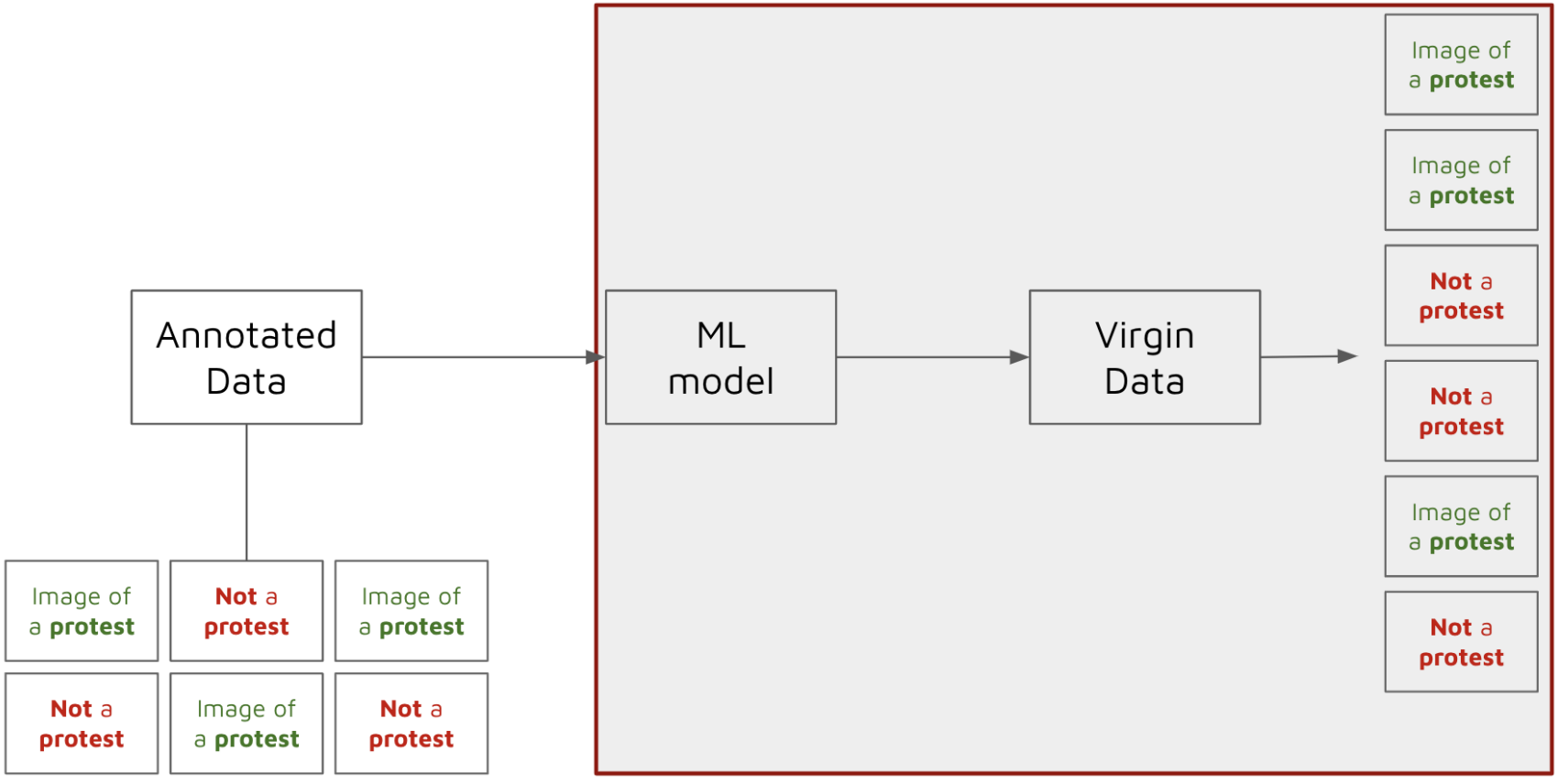
The Basics of Supervised Classification

Training the model



Pre-trained models and zero-shot classification

Applying an existing model to your data



Pre-trained models and zero-shot classification
Common benchmark datasets in computer vision

- Cifar-10 & Cifar-100 (cat, dog, horse, ...)
- MNIST (handwritte digits: 1, 2, 3, 4, ...)
- MS COCO: many categories and subcategories
- ImageNet: 14 million examples. 1,000 objects.
- ... many many others!

Pre-trained models and zero-shot classification

Many are available through deep learning libraries in `python`

Here a link to all the ones available through `pytorch`

Pre-trained models and zero-shot classification

The shortcomings and the dangers

- No model trained to predict our quantity of interest
- Even if there is, it may not generalize well to our own data

Pre-trained models and zero-shot classification

Most common zero-shot classification

- Ethnicity
- Gender
- Facial expressions/emotions

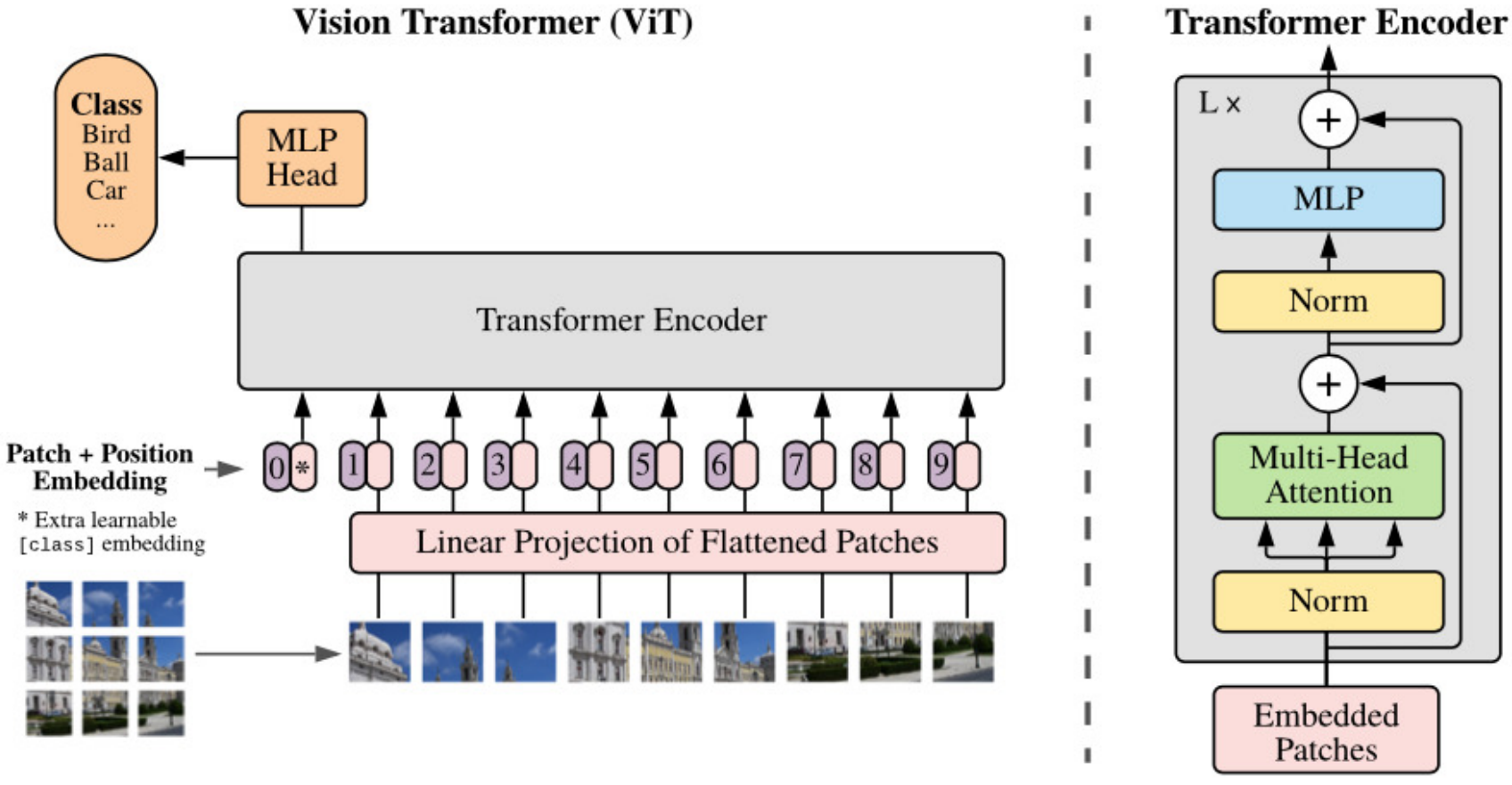
Entering 2021: Vision Transformers (ViT)
Originally introduced by Google Brain

Simple summary: The model breaks the image into small pieces, processes the relationships between them, and makes a prediction about what the image represents.

- Very close to using transformer models in NLP
- Splitting an image into a grid of sub-image patches
- Embed each patch with a linear projection (= make embedding)
- Each embedded patch becomes a token, and the resulting sequence of embedded patches is the sequence you pass to the model

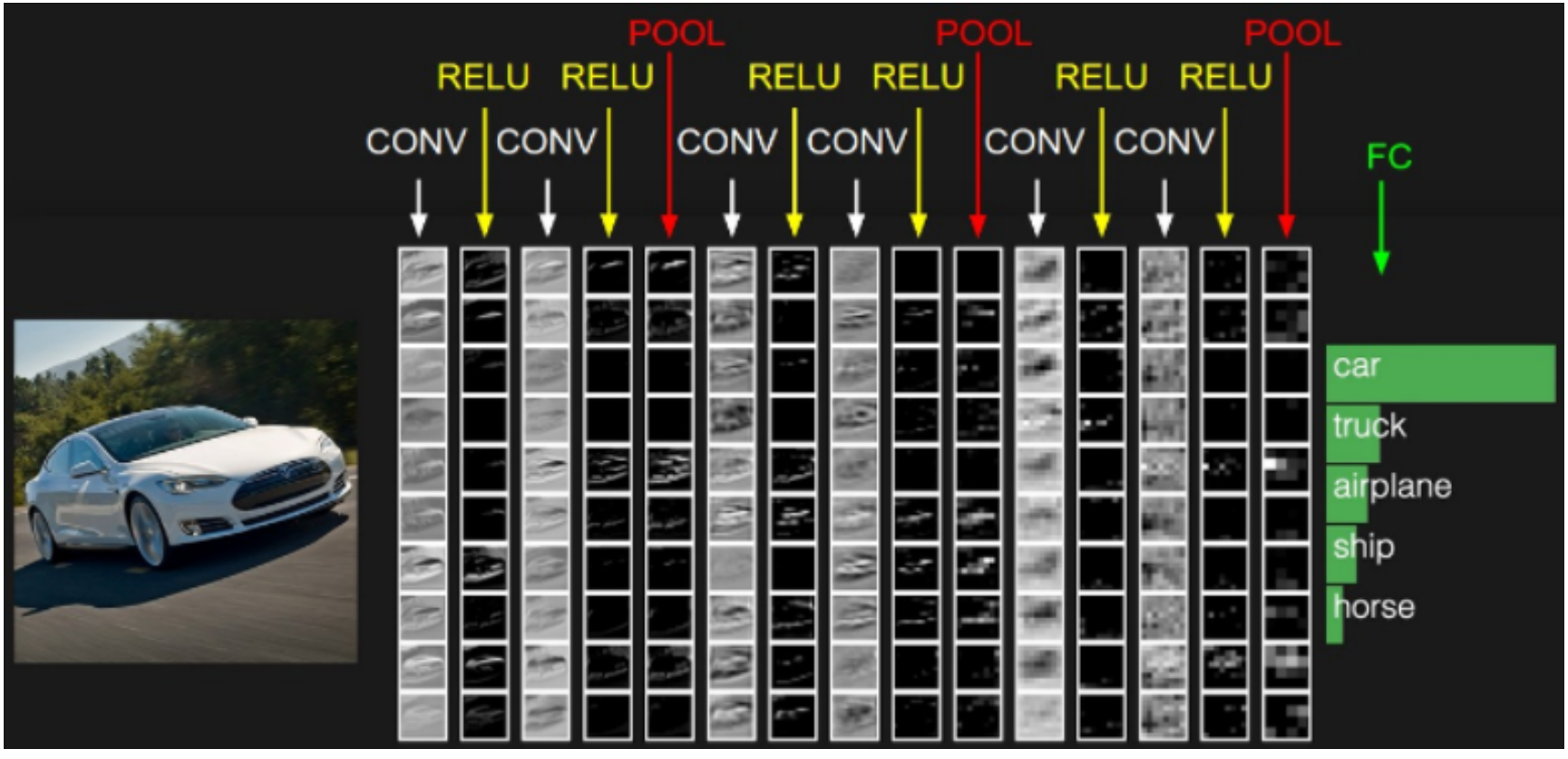
Entering 2021: Vision Transformers (ViT)

Originally introduced by Google Brain



Fine-tuning a pre-trained model

Leveraging an existing model for our own task



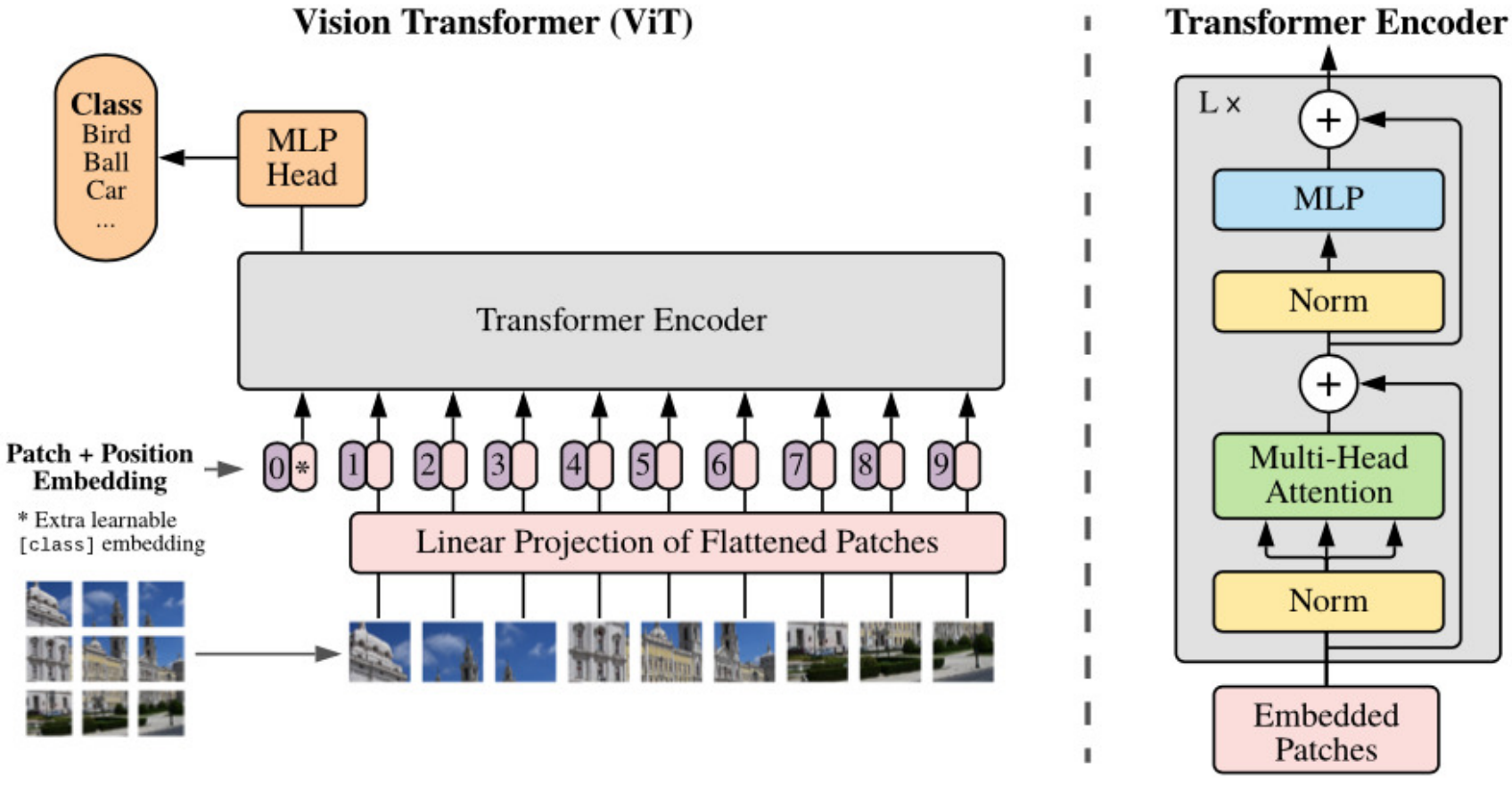
Fine-tuning a pre-trained model

Leveraging an existing model for our own task

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

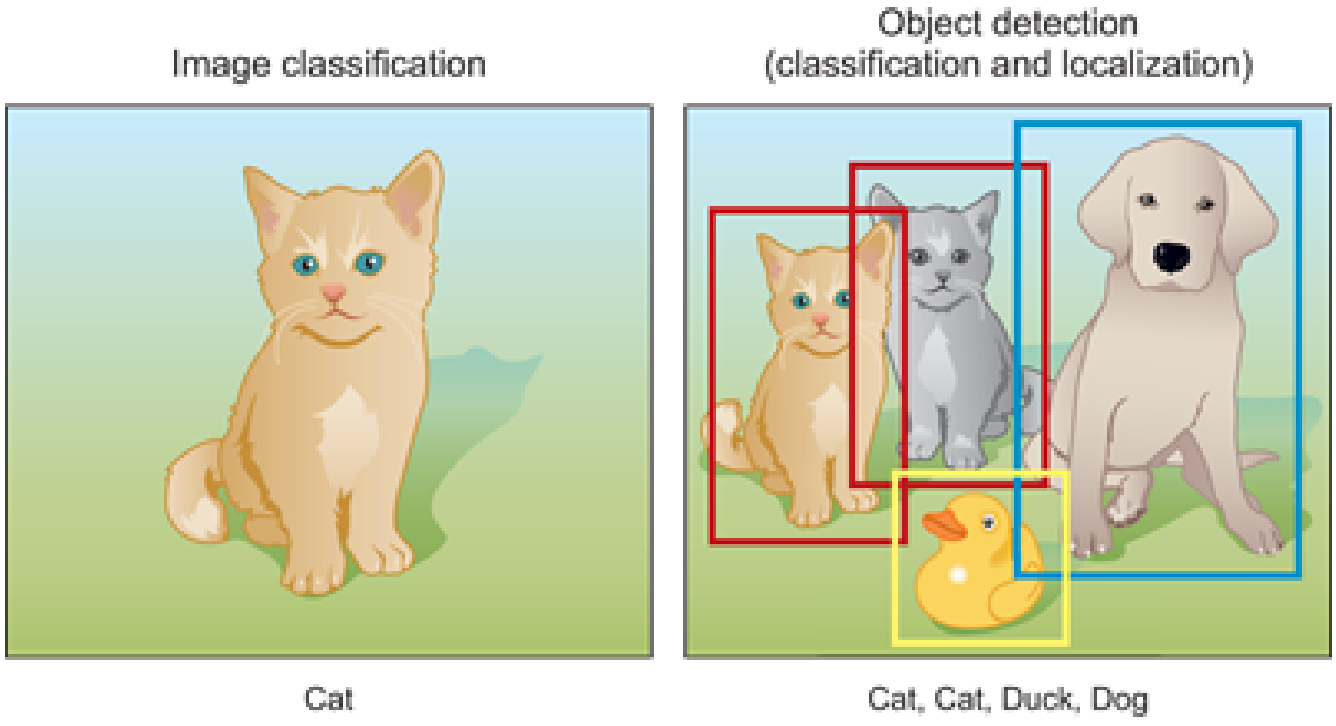
Fine-tuning a pre-trained model

Leveraging an existing model for our own task



Fine-tuning a pre-trained model

Additional consideration: detection *v.* recognition

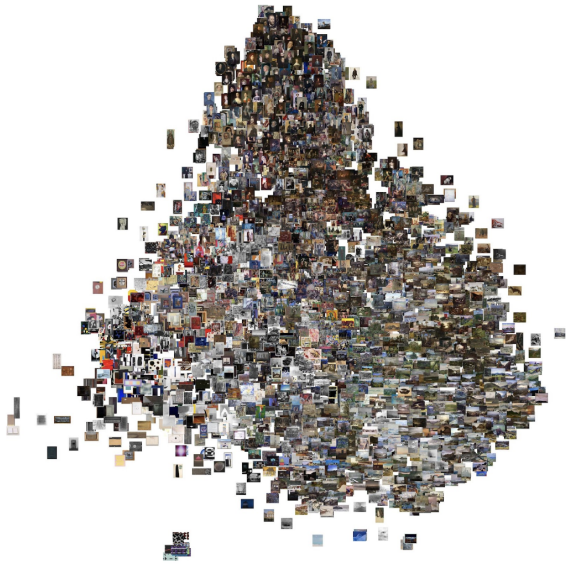


*We will not cover object detection in the tutorial (apart from when working with faces). See this link for sample `pytorch` code on how to fine-tune a widely-used object detection + recognition model (RCNN).

Outline

What are unsupervised models useful for?

- Data exploration: learn more about your data

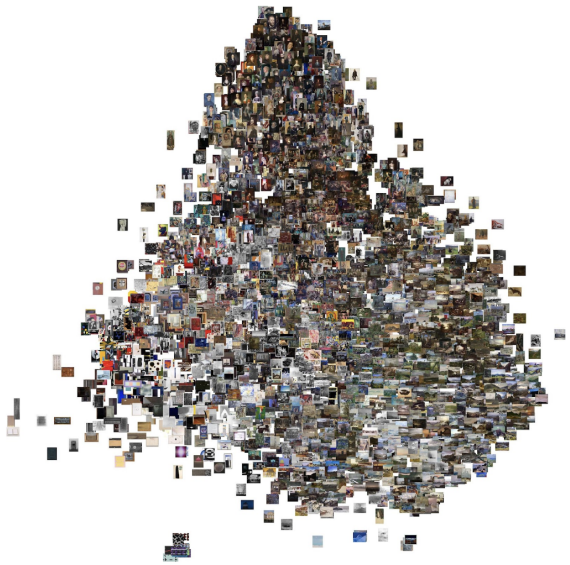


What are unsupervised models useful for?

- Data exploration: learn more about your data
- "Deduplication"

What are unsupervised models useful for?

- Data exploration: learn more about your data
- "Deduplication"
- Stratified sampling



What are unsupervised models useful for?

- Data exploration: learn more about your data
- "Deduplication"
- Stratified sampling
- Descriptive analysis

What are unsupervised models useful for?

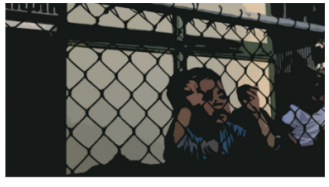
An example: diffusion of #FamiliesBelongTogether on Twitter

@xxxx: F*** Trump. #FamiliesBelongTogether

@fams2gether: Nobody, no matter who you are, where you are from, you should spend an hour like this!

@xxxx: #FamiliesBelongTogether in Denver this morning.

@ResistSnow: #FreeTheChildren #FamiliesBelongTogether Our children long for realistic maps of the future that they can be…



What are unsupervised models useful for?
An example: diffusion of #FamiliesBelongTogether on Twitter

$$f(\textit{online\_engagement}) = \textit{image} \times \textit{viewer}$$

Image features:

- social identity
- expectation of success
- **evoked emotions**

View features:

- gender
- race/ethnicity
- ideology/**partisanship**

The Logic of Unsupervised Classification

An example: diffusion of #FamiliesBelongTogether on Twitter

Goal: to cluster images with similar content, to reduce and balance data-annotation

A

B

C

D

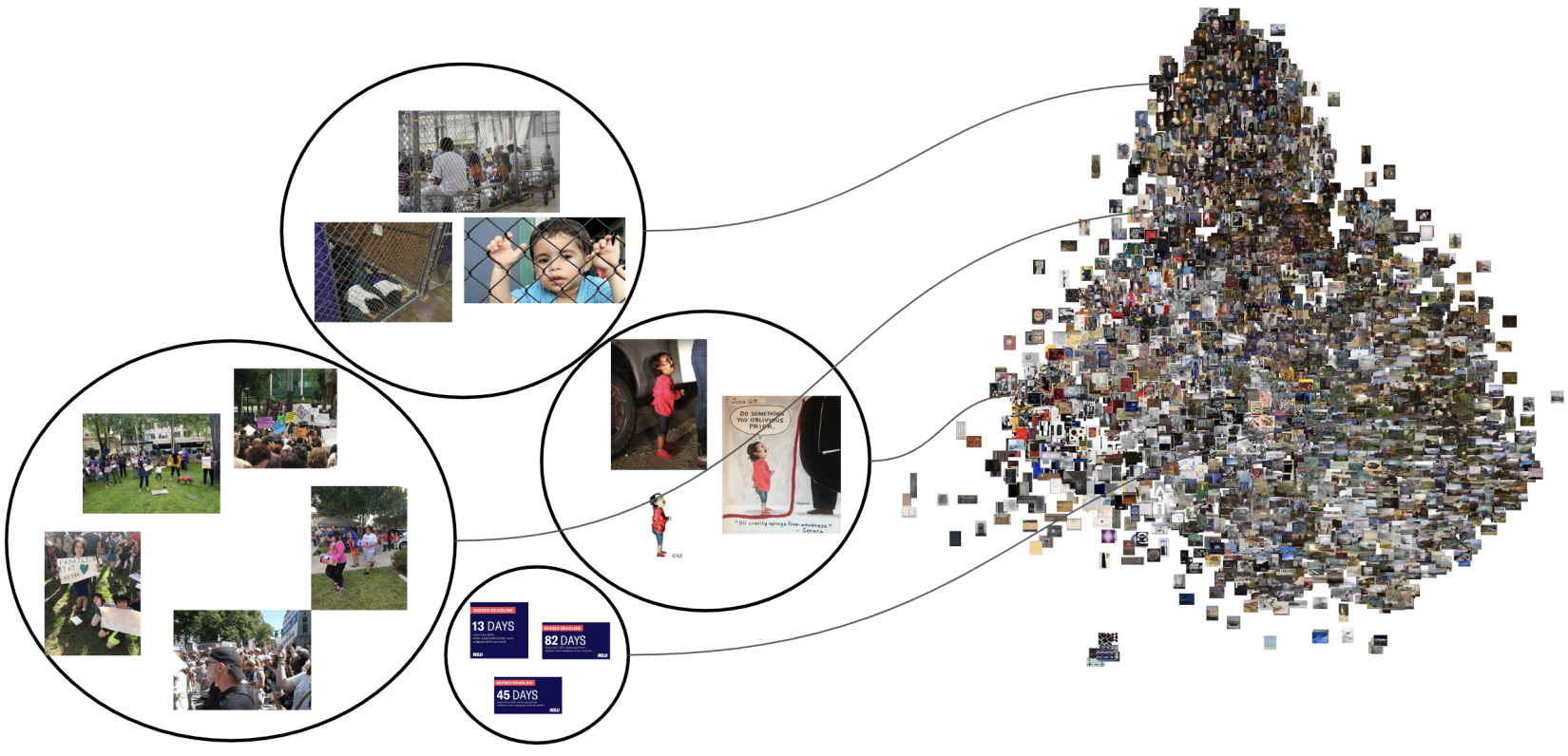
E

F

The Logic of Unsupervised Classification

An example: diffusion of #FamiliesBelongTogether on Twitter

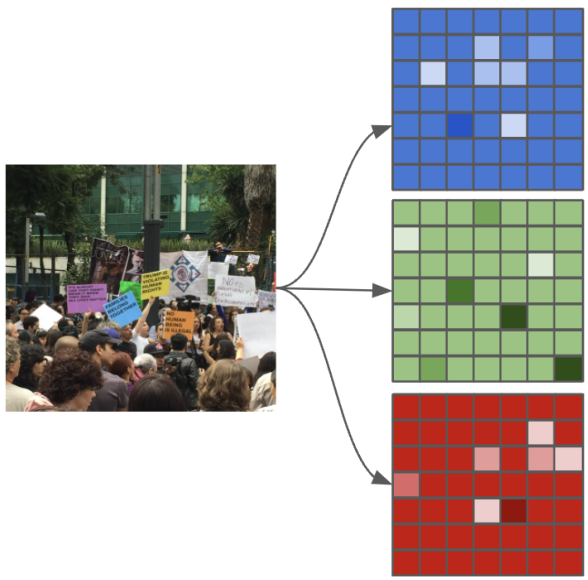
Goal: to cluster images with similar content, to reduce and balance data-annotation



The Logic of Unsupervised Classification

Three main challenges

1. How to numerically represent image content in order to cluster based on content similarity



The Logic of Unsupervised Classification
Three main challenges

1. How to numerically represent image content in order to cluster based on content similarity

2. How to select the best fitting number of clusters
   - Datasets are likely to be highly unbalanced, complicating the cluster-discovering process
   - "True" number of clusters is unknown
   - Want to make sure truly cohesive clusters are found

The Logic of Unsupervised Classification
Three main challenges

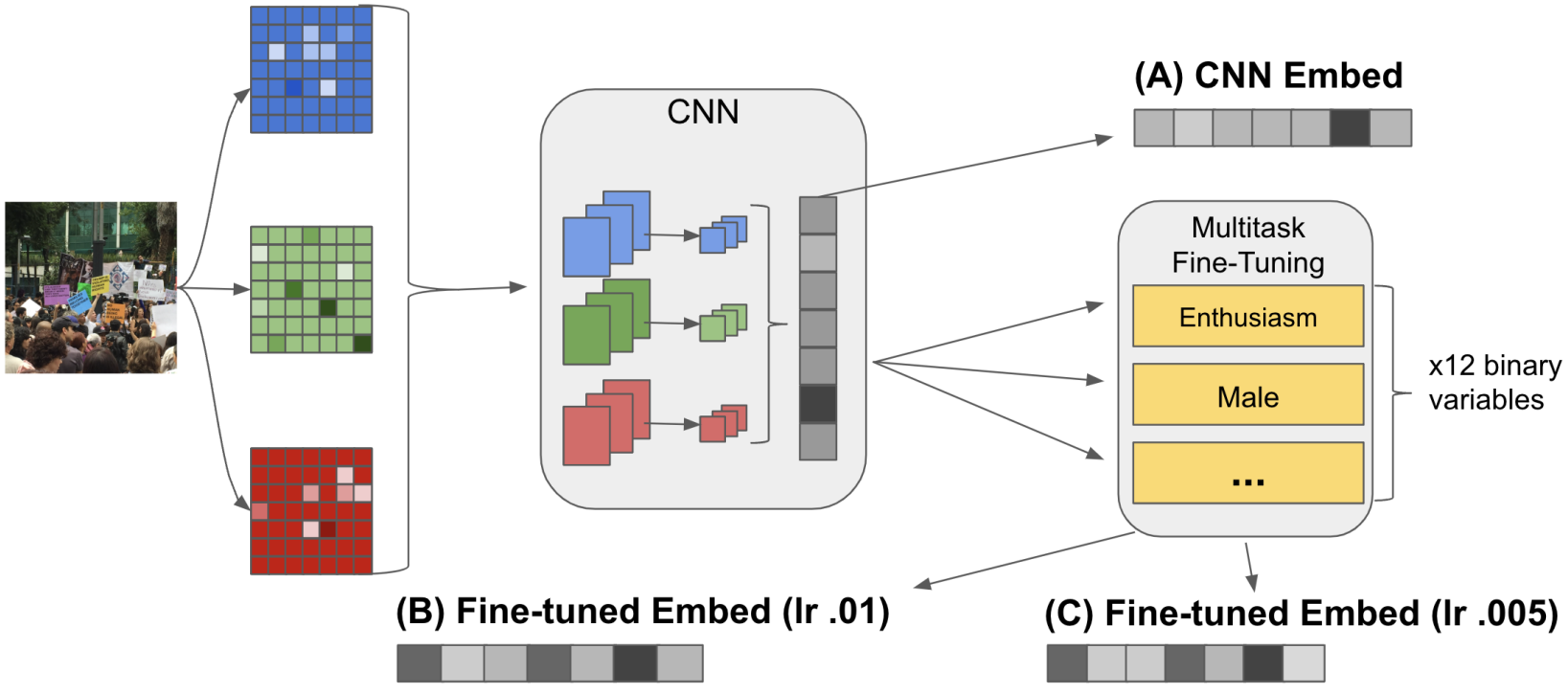
1. How to numerically represent image content in order to cluster based on content similarity

2. How to select the best fitting number of clusters
   - Datasets are likely to be highly unbalanced, complicating the cluster-discovering process
   - "True" number of clusters is unknown
   - Want to make sure truly cohesive clusters are found

3. How do we evaluate...
   - Are we using the "best" numeric image representation possible? (1)
   - The performance of different clustering configurations? (2)

The Logic of Unsupervised Classification

1. Three approaches to reducing image dimensionality



The Logic of Unsupervised Classification

2. Clustering process

- **Iterative process** find big cohesive clusters first, and then see if there are smaller ones
- Intra-cluster similarity (silhouette scores) to judge cluster cohesiveness

Hyperparameters

- ➔ Step Size
- ➔ Similarity Threshold
- ➔ Sample Size
- ➔ Stop Size
- ➔ Converge Window

Algorithm

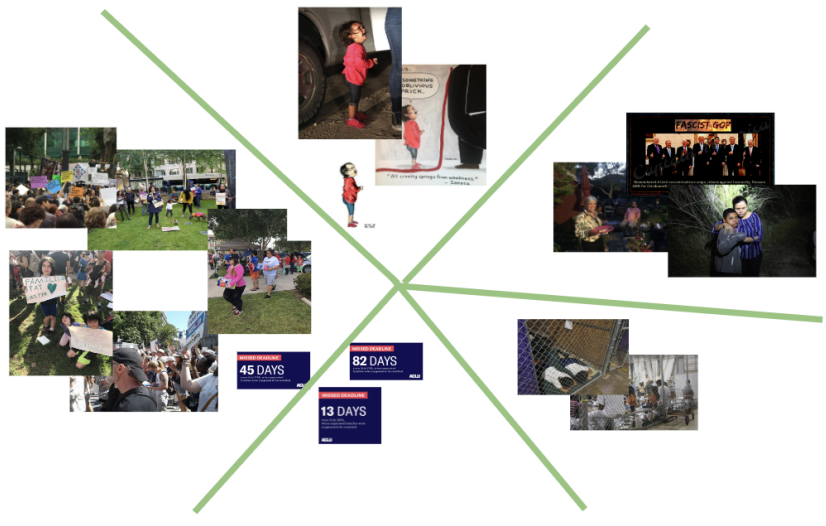
X = input image matrix (20,000 images x 512 embed size)

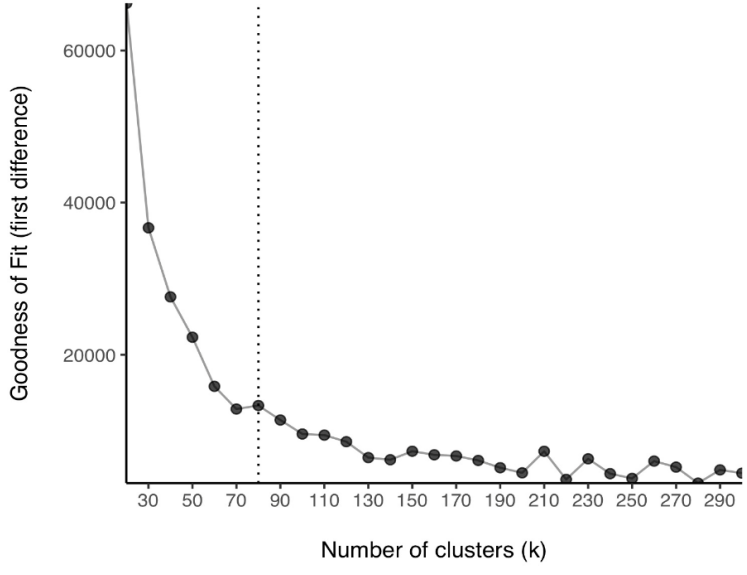
1. Choose number of *K* clusters to predict
   a. Randomly sample 1,000 images from X
   b. Fit several k-means algorithms (increase *K* by 5 each time)...
      … until goodness of fit doesn't improve (avg. across 3 iterations)
      … STOP if *K* > images in X
2. Fit k-means algorithm predicting *K* image clusters
3. Evaluate intra-cluster similarity (silhouette score)
4. Find cohesive clusters (silhouette score > 0.04)
5. Remove cohesive images from X, and keep classifying the rest
   … STOP if only 20 images left

The Logic of Unsupervised Classification

3. Bulding a validation set

- Number of clusters is unknown but clusters are likely to be unbalanced
- Fit a first clustering algorithm (e.g. k-means)





The Logic of Unsupervised Classification

3. Bulding a validation set

- Number of clusters is unknown but clusters are likely to be unbalanced
- Fit a first clustering algorithm (e.g. k-means)
- Select random pairs from each cluster (e.g. n=554) and label them for cohesiveness

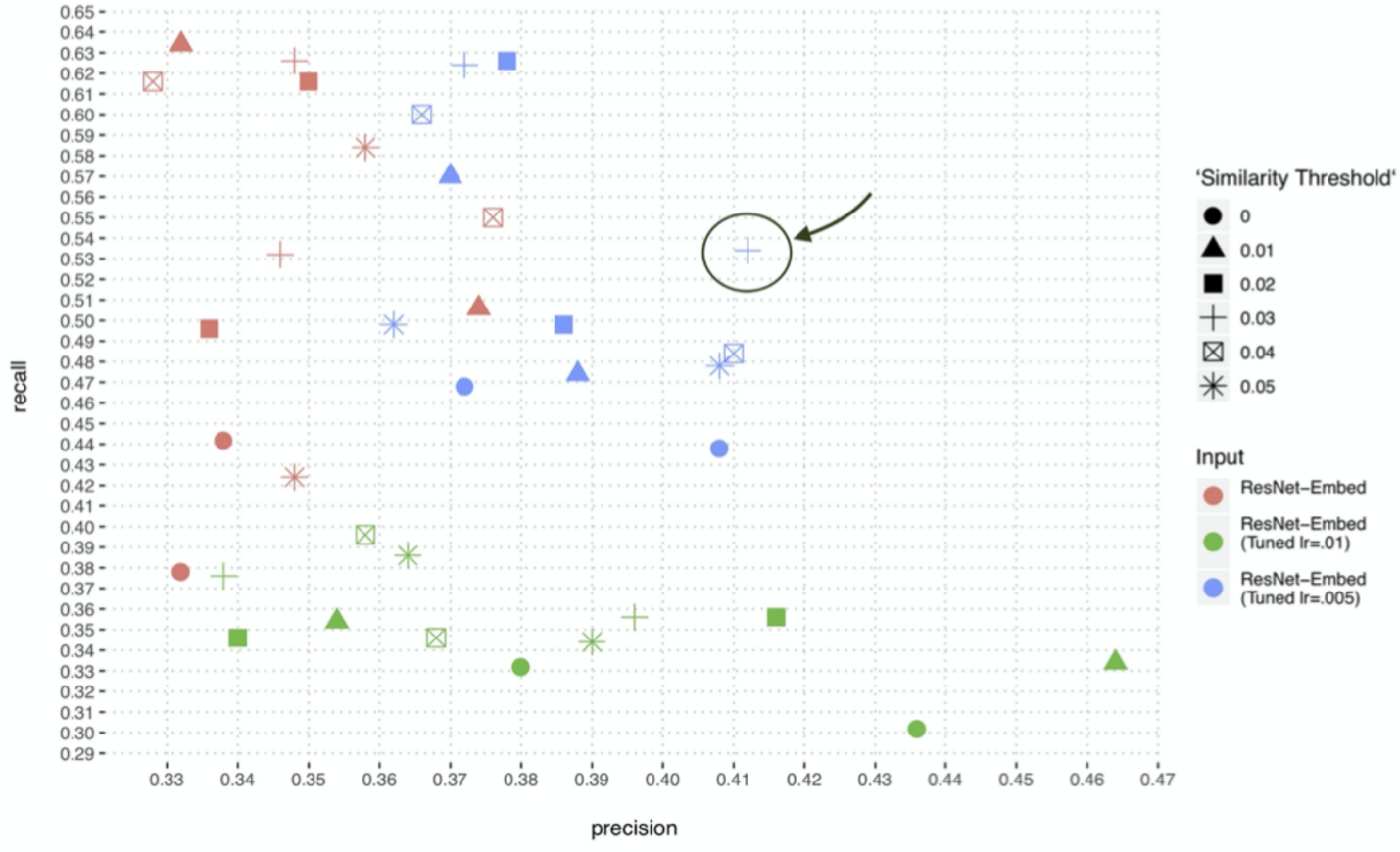
	coder1	coder2
	1	1
	1	1
	0	0
	0	1
	1	1

Agreement
87%

Kappa
0.64

The Logic of Unsupervised Classification

3. Comparing performance of clustering hyperparameters



Some illustrative results

1. Image exploration

Some illustrative results

1. Image exploration

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)

Some illustrative results

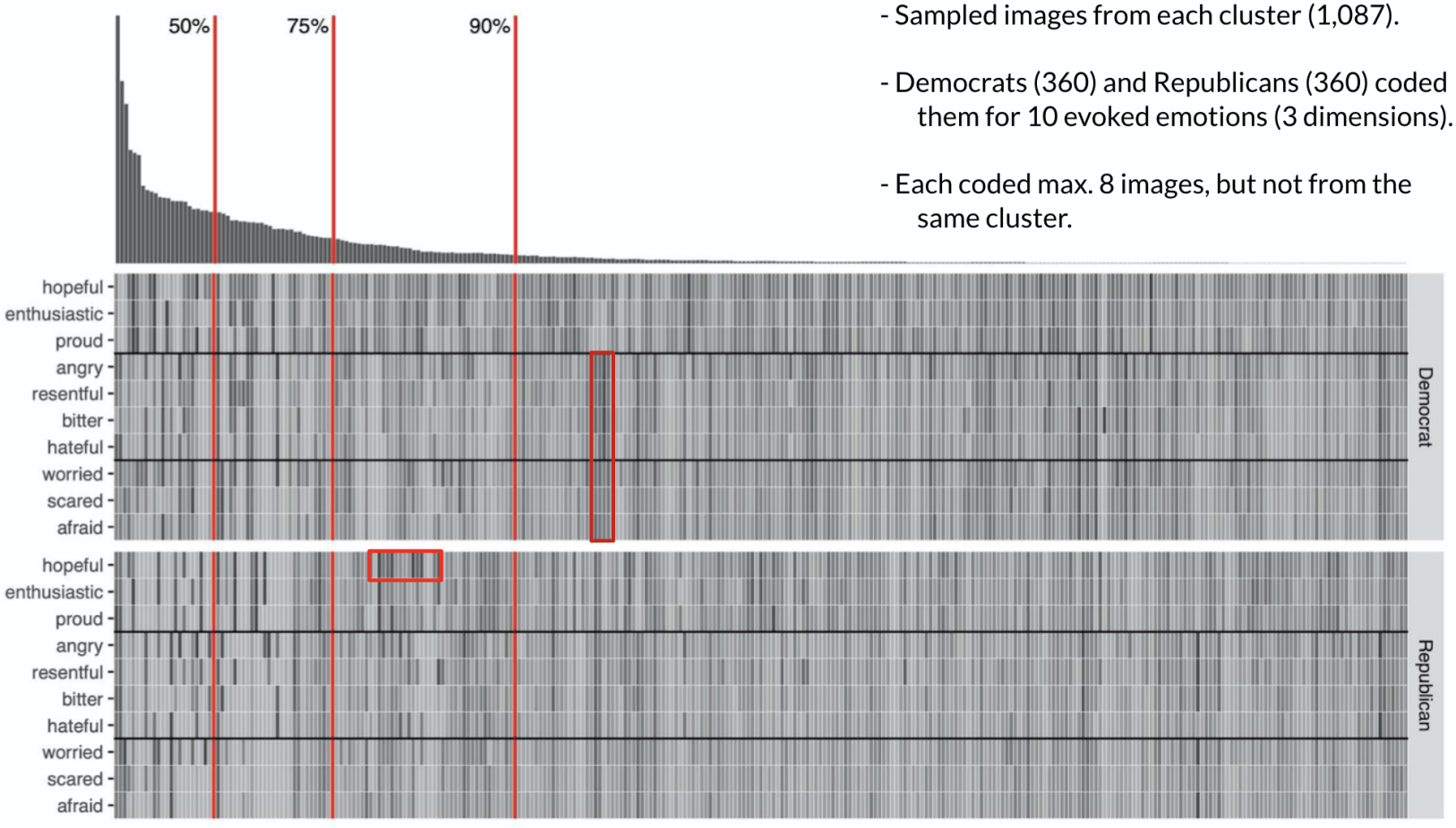
1. Image exploration

Some illustrative results

1. Image exploration

Some illustrative results

2. Purposive sampling



- Sampled images from each cluster (1,087).
- Democrats (360) and Republicans (360) coded them for 10 evoked emotions (3 dimensions).
- Each coded max. 8 images, but not from the same cluster.

Some illustrative results

3. Differences in annotation