Matching

Lecture 4: Classifier

Anton Ryabtsev

Moscow Institute of Physics and Technology

Autumn 2023

Classifier

Texts

Pictures

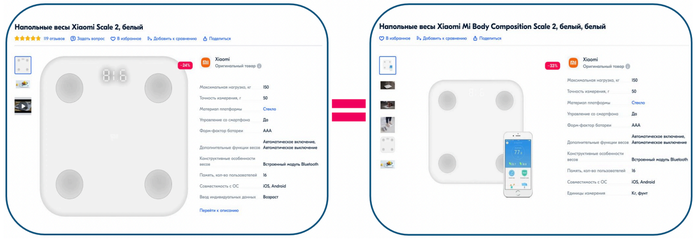
Attributes

ML

95%

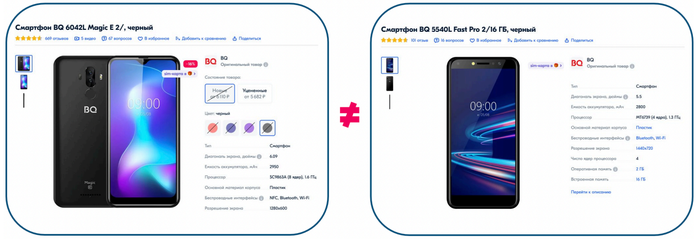
5%

Matches



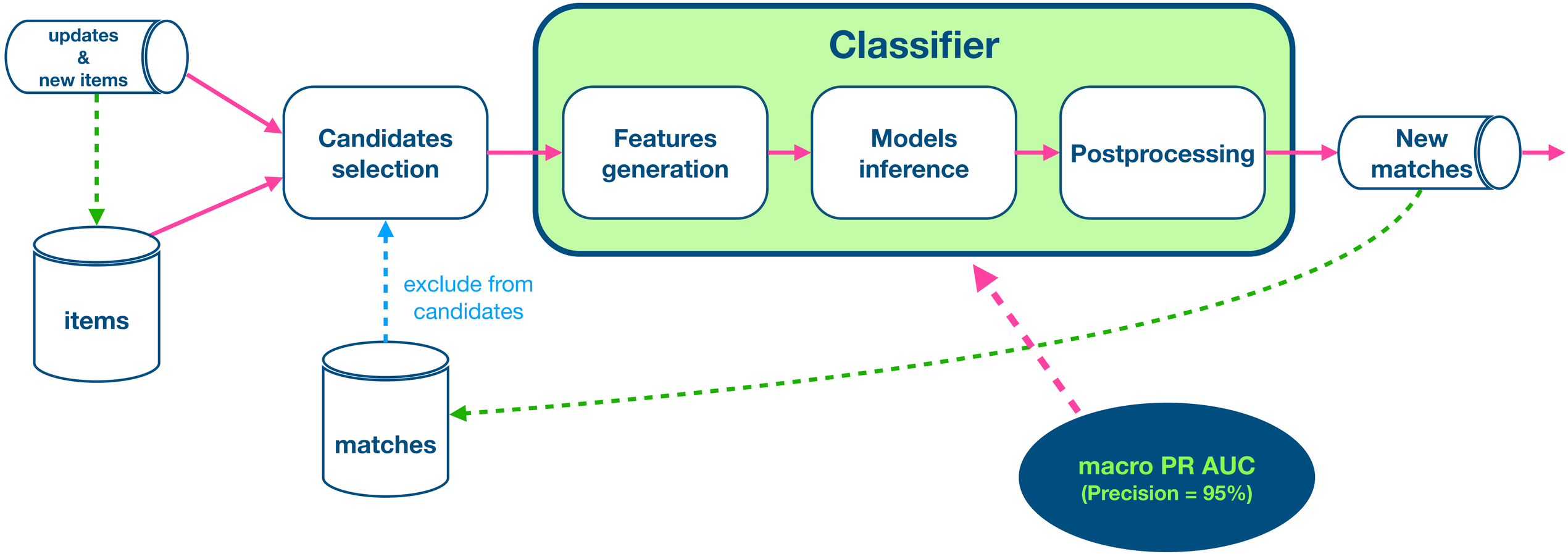
Can’t be classified

Not matches



Matching Pipeline

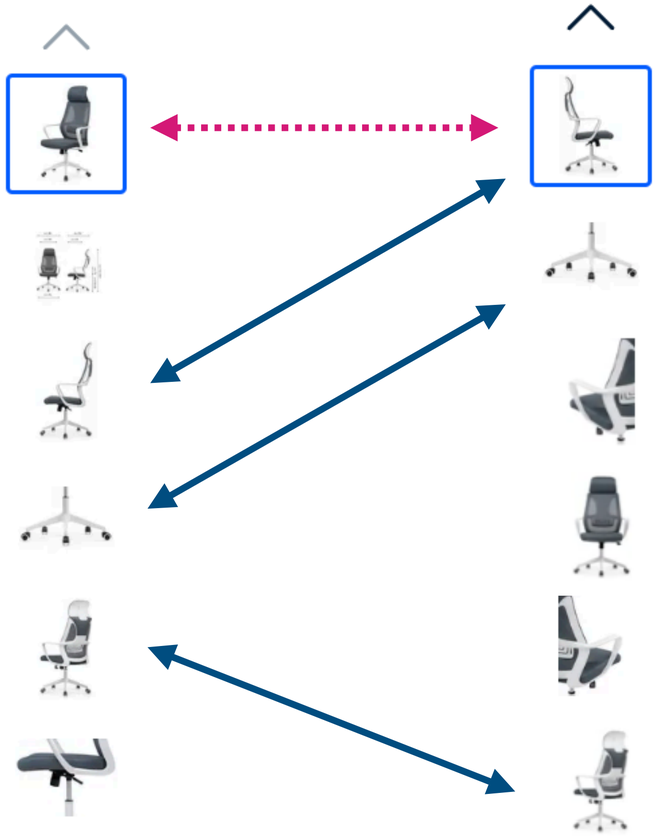
High level design



Features

Images

Item 1

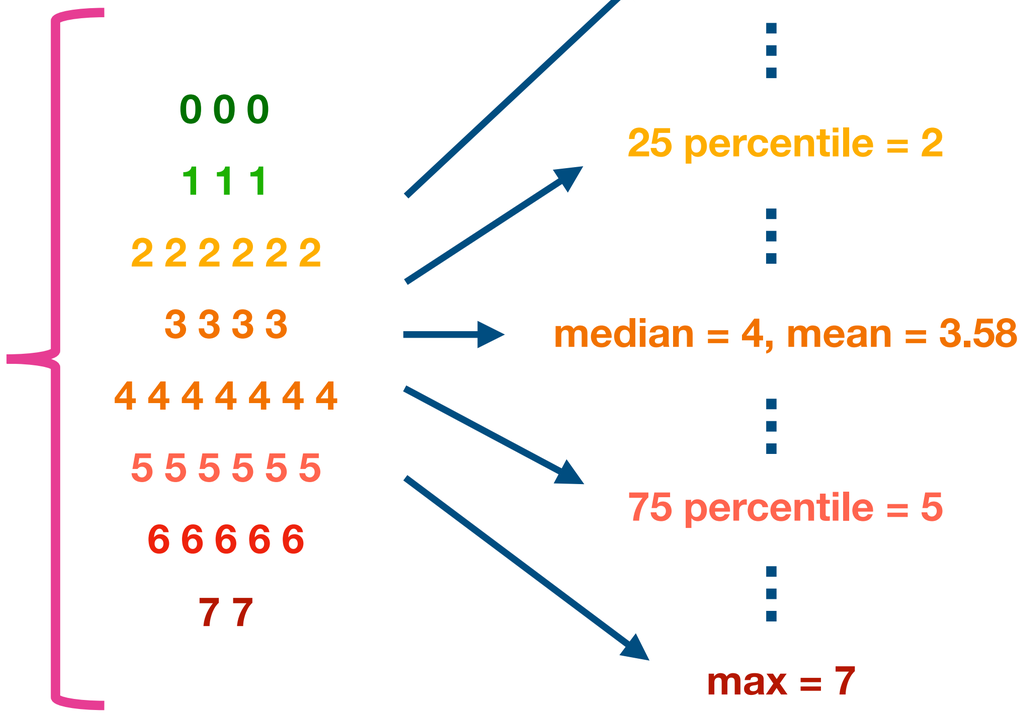


Item 2

Features

Images: embeddings

Pairwise distance matrix						
	5	7	4	1	3	2
	6	6	5	4	5	6
	0	2	4	5	4	2
	2	0	6	7	5	3
	2	3	1	2	1	0
	3	6	4	5	4	4

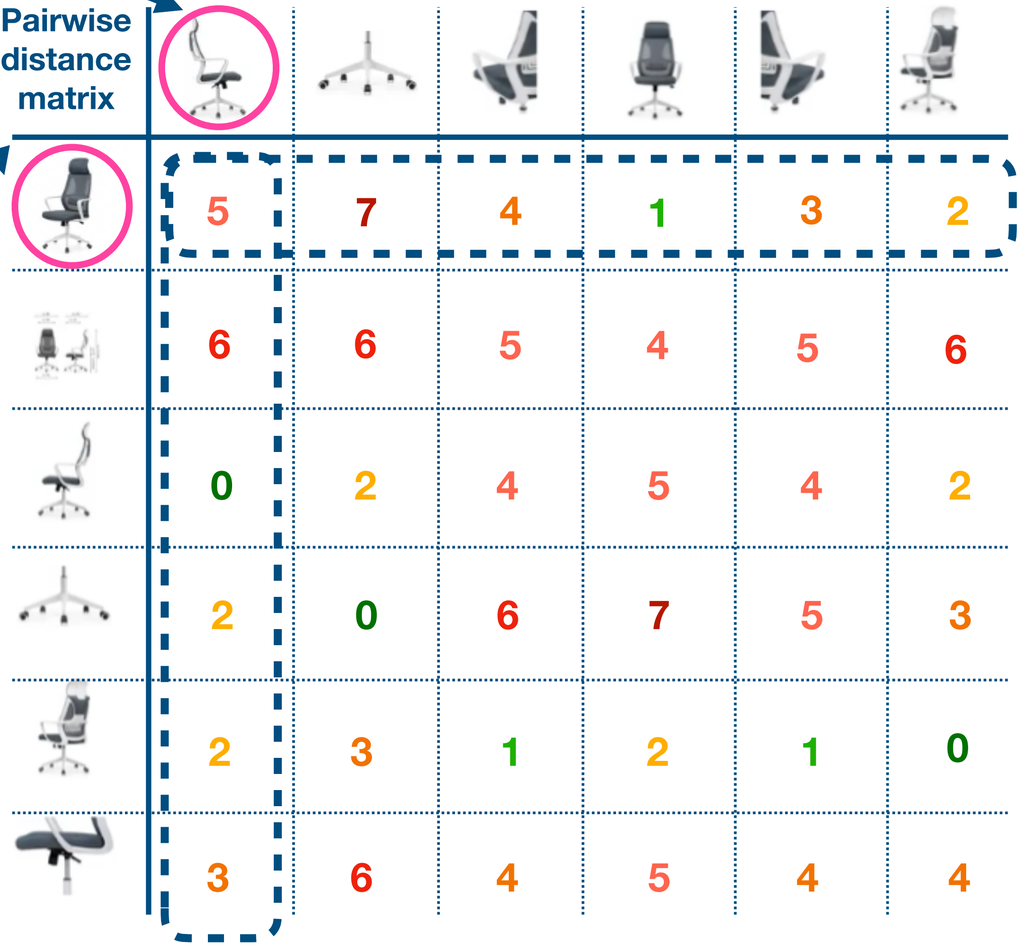


Features

Images: embeddings

Main images

Pairwise distance matrix



0
1
2 2 2
3 3
4
5
6
7

min = 0

25 percentile = 1

median = 3, mean = 3.18

75 percentile = 6

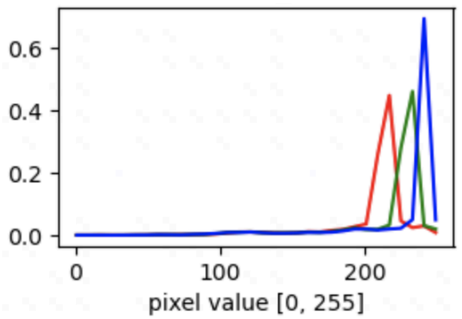
max = 7

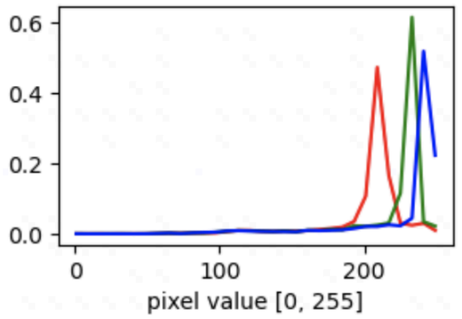
Features

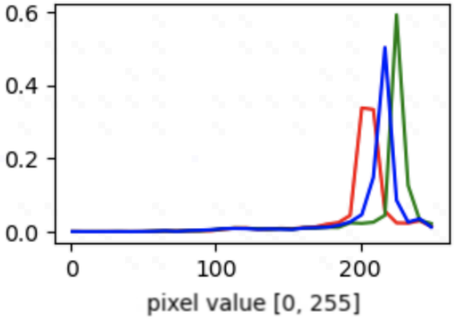
Images: colors

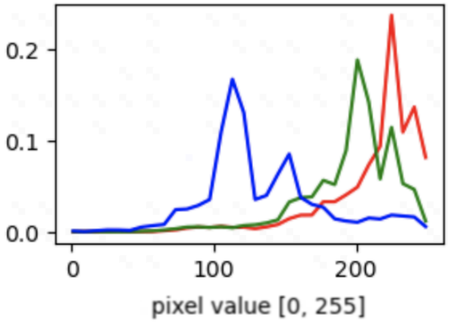


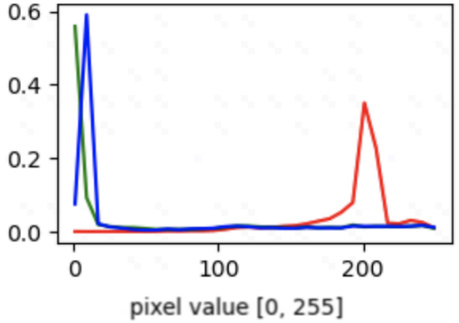












Divergence = 0.34

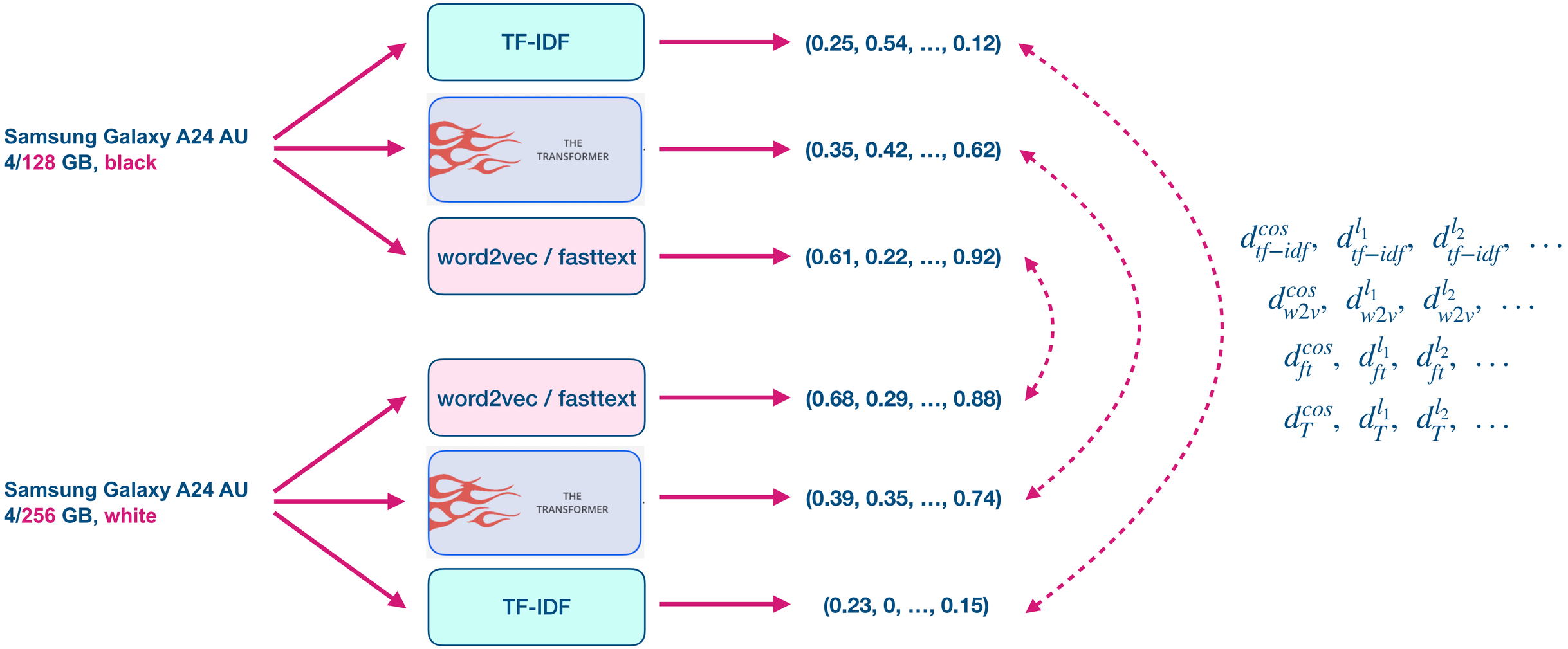
Divergence = 2.07

Divergence = 2.72

Divergence = 5.91

Features

Texts: bi-encoder like approaches



Features

Texts: bi-encoder like approaches

Pros:

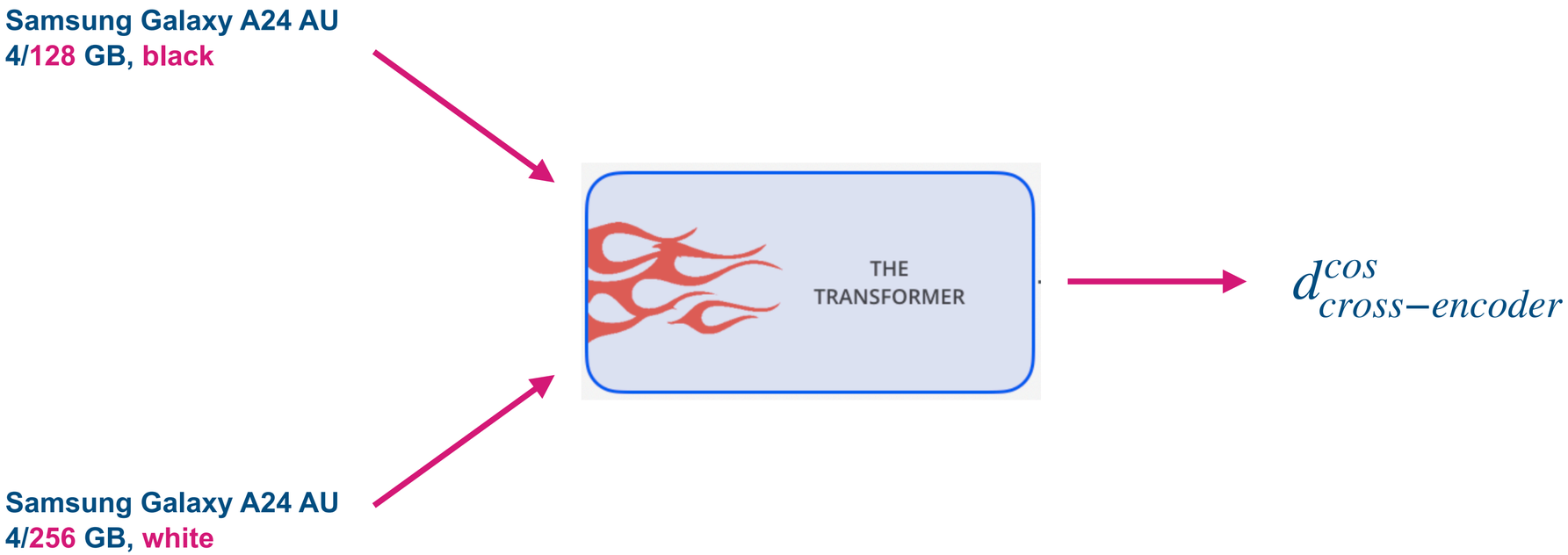
- Embeddings can be prepared continuously in the background for new and updated items.

Cons:

- Embeddings are built for items independently and may not capture small but important details.

Features

Texts: cross-encoder like approaches



Features

Texts: cross-encoder like approaches

Pros:

- Embeddings are built for pairs of items and may capture small but important details.

Cons:

- Embeddings can’t be prepared continuously in the background for new and updated items and should be generated during feature generation stage.

Features

Texts: catboost text features

Pros:

- Easy and quick to try.
- It will most likely work well.

Features

Attributes

Heuristic filters by attributes:

- Color.
- Size, dimensions.
- Pieces.
- Price.
- Etc.

Features

Features

Attributes

Pros:

- Good features can significantly increase the recall of matching.

Cons:

- Too many different attributes, variations in their spelling, dimensions of values — you'll have to do a lot of if-clauses.

To begin with, the catboost text features may be good enough.

Dataset

No money

Positives

1. Part number / ISBN matcher.
2. Matches found manually.
3. Synthetic matches — augmentations: take subsamples of images, attributes.

Negatives

1. Part number / ISBN anti-matcher.
2. Substitutions.
3. Pairs from kNN with different categories.

Dataset

If you already have some ML matcher

Positives

1. **T**rue **P**ositives from everyday precision estimations.
2. **F**alse **N**egatives from everyday recall estimations.

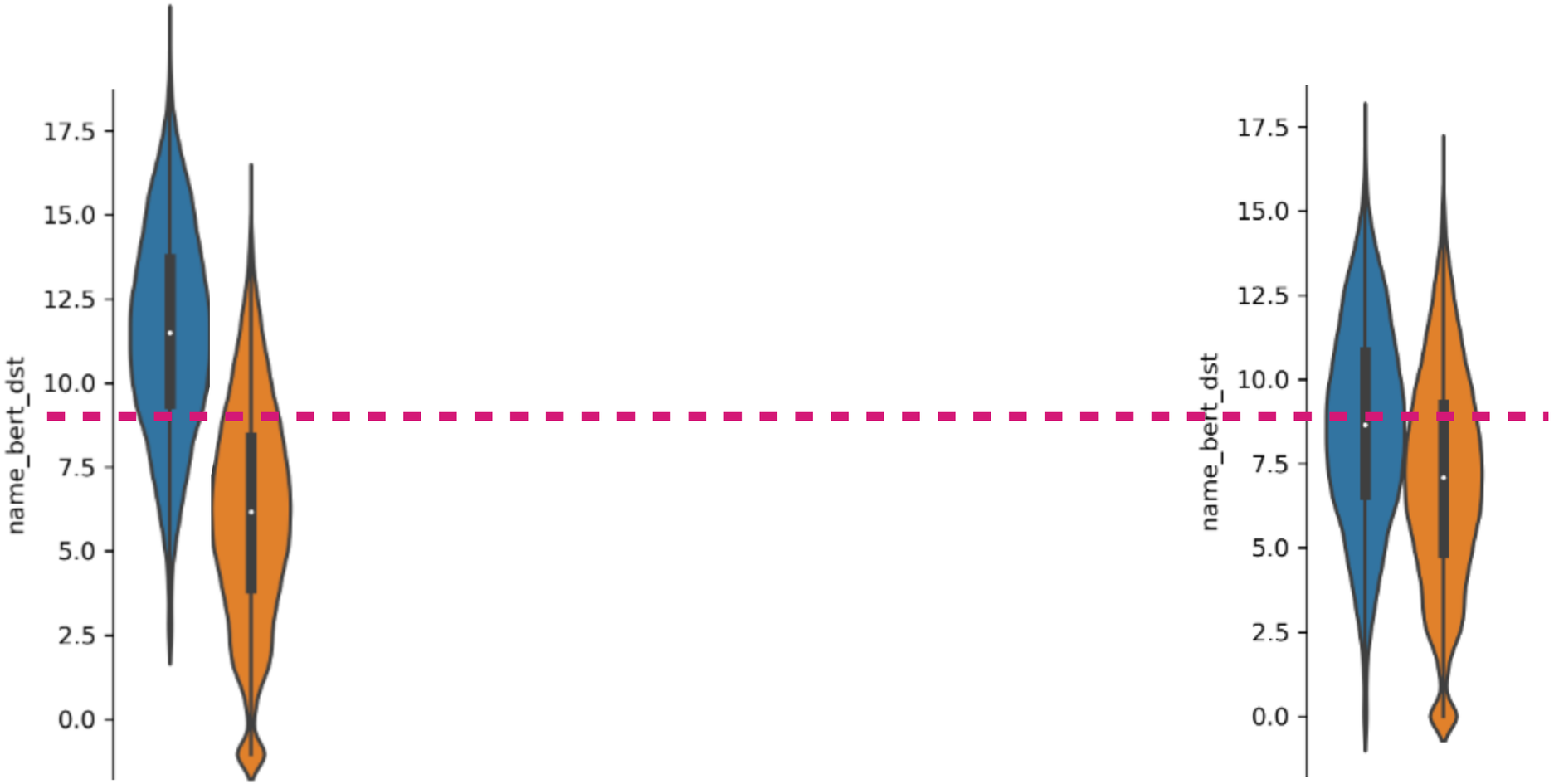
Negatives

1. **F**alse **P**ositives from everyday precision estimations.

+ take sample from pipeline candidates and get labels via crowd system

Dataset

Usefulness of various sources

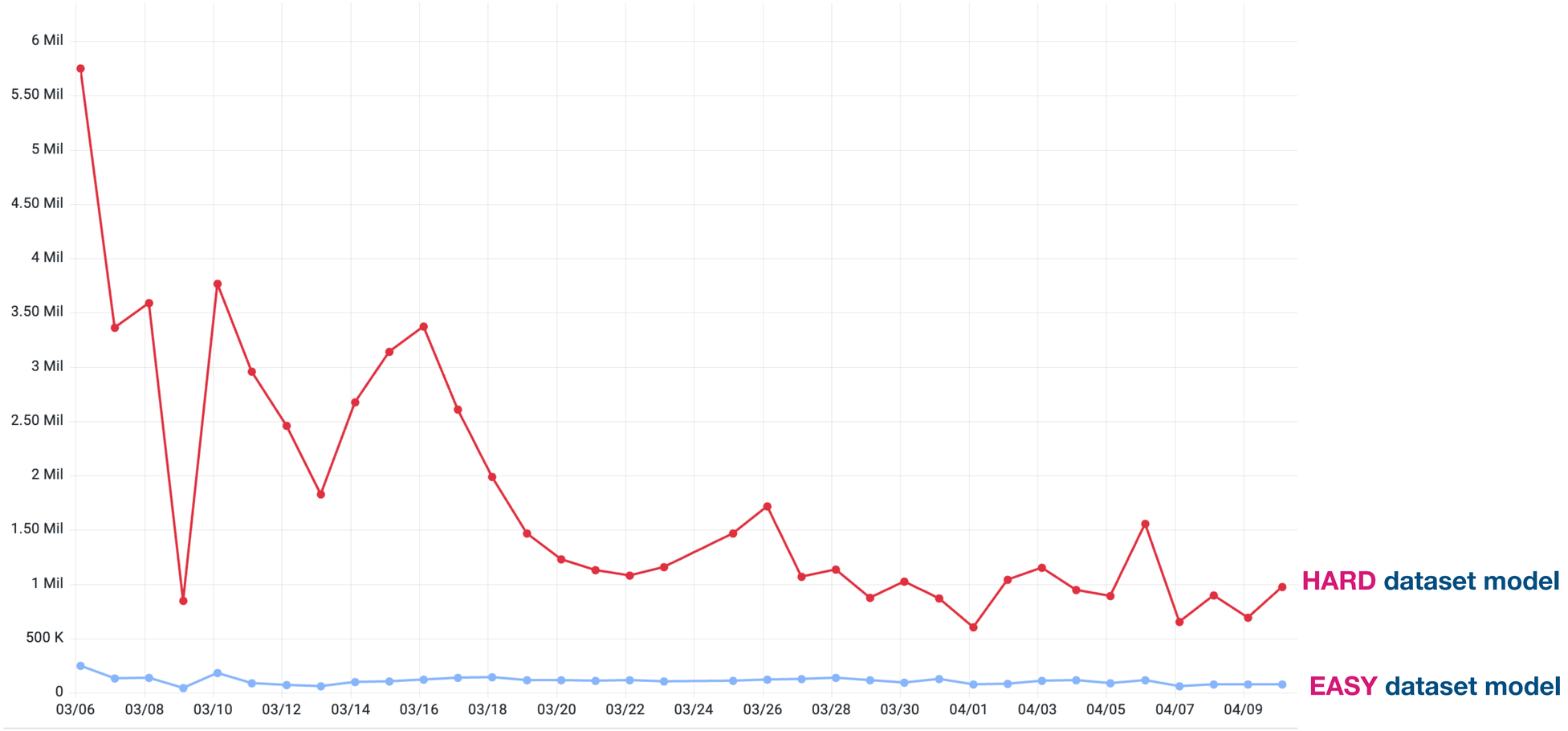


Matches, substitutions, negatives from kNN
EASY

TP, FP and FN from every day precision and recall analysis
HARD

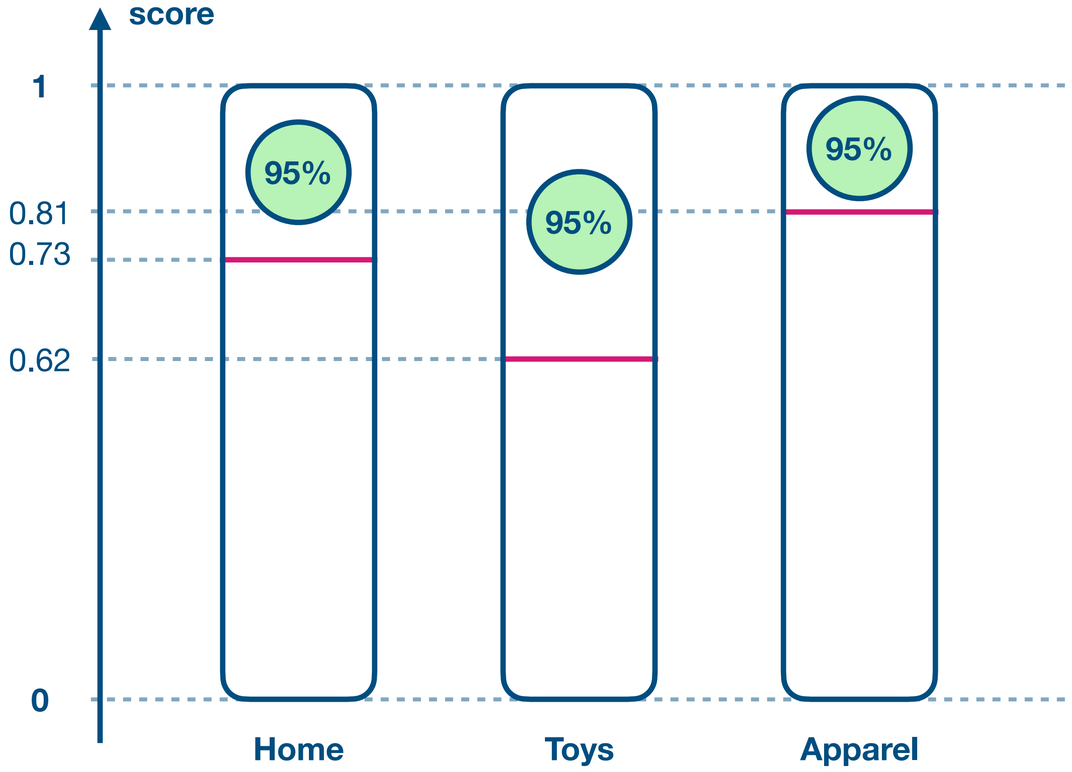
Dataset

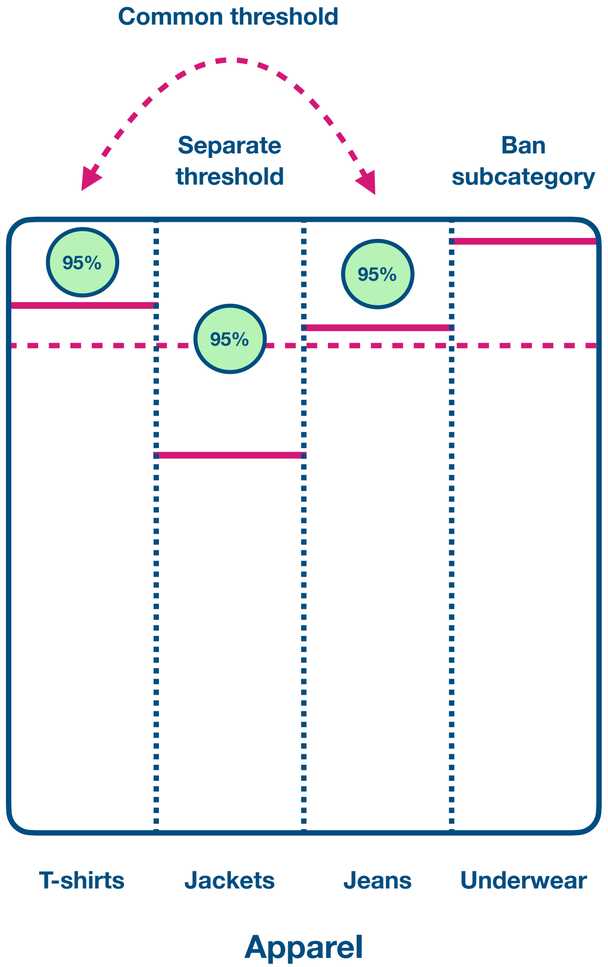
Usefulness of various sources



Postprocessing

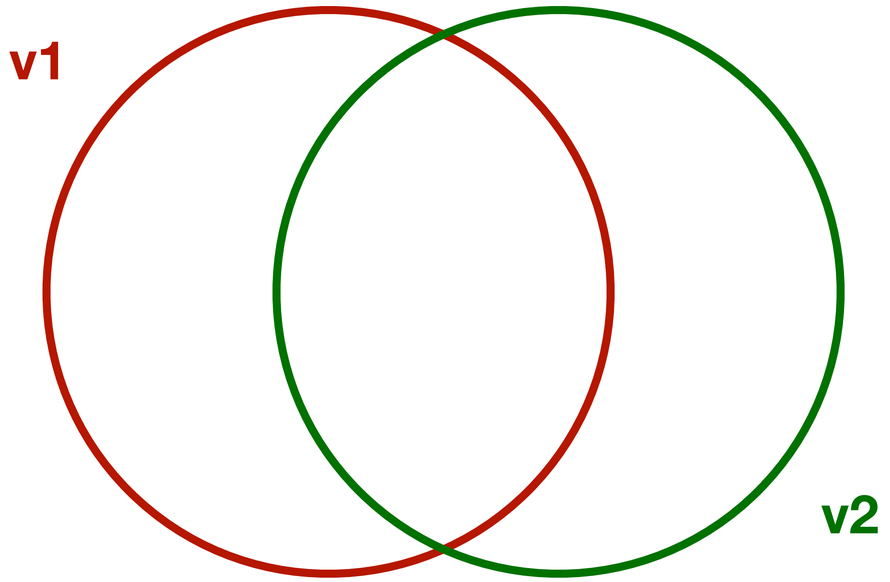
Thresholds





Postprocessing

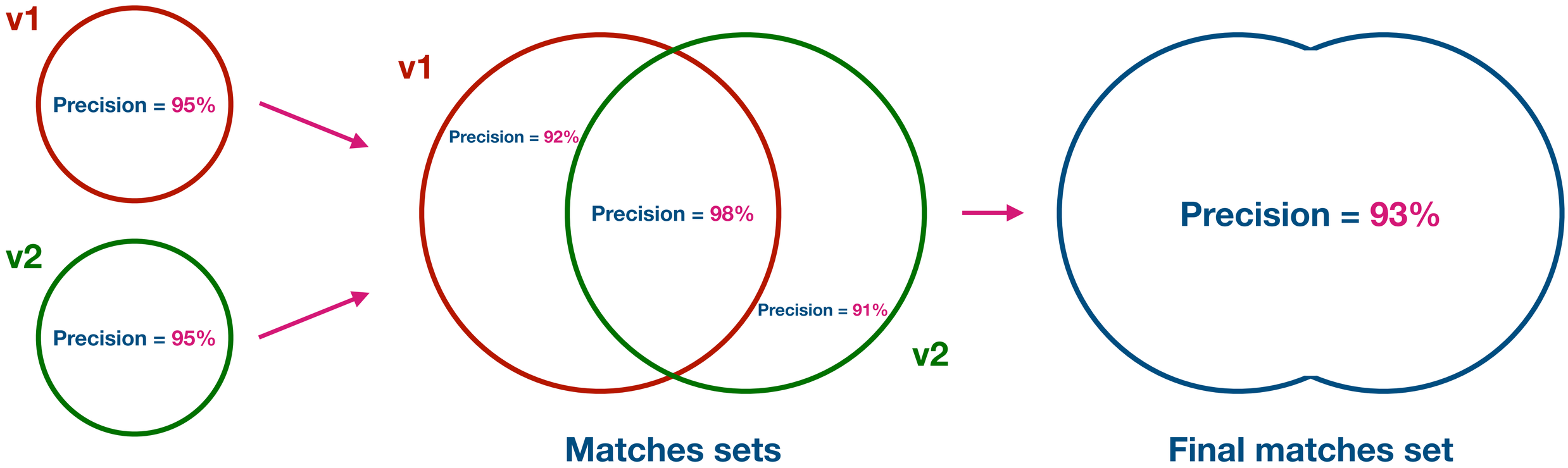
Combining results from multiple models



Matches sets

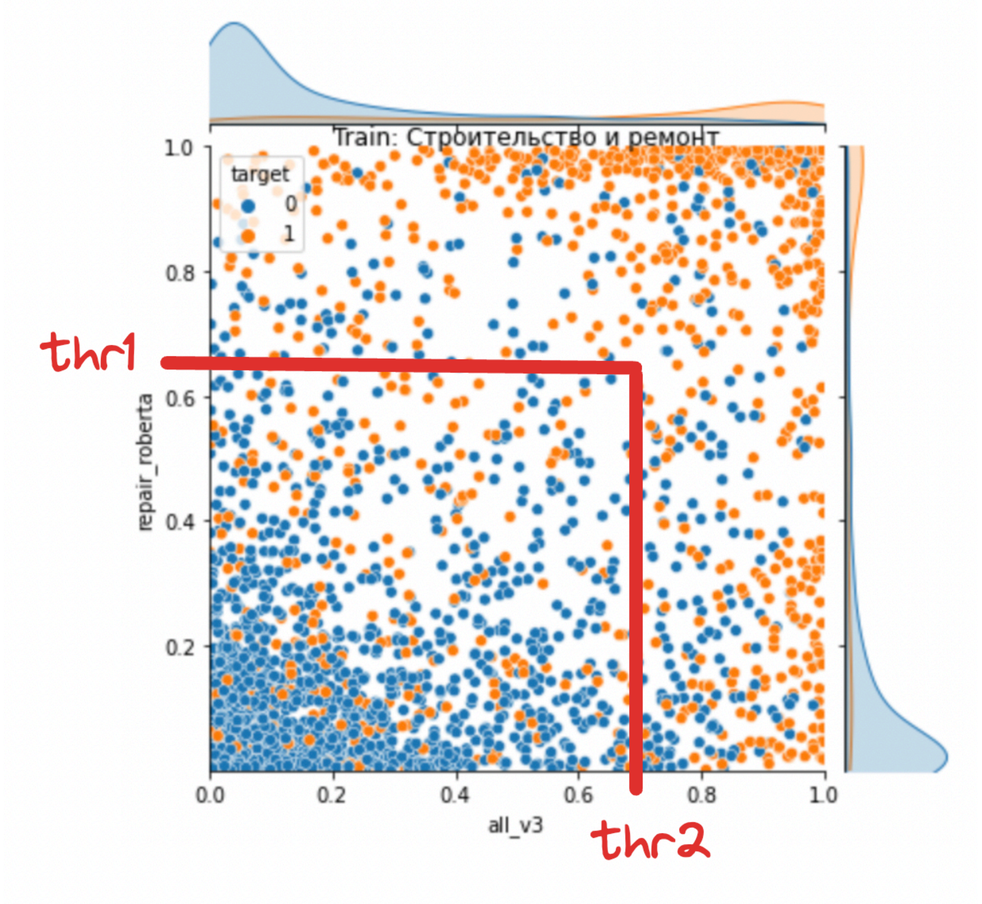
Postprocessing

Combining results from multiple models

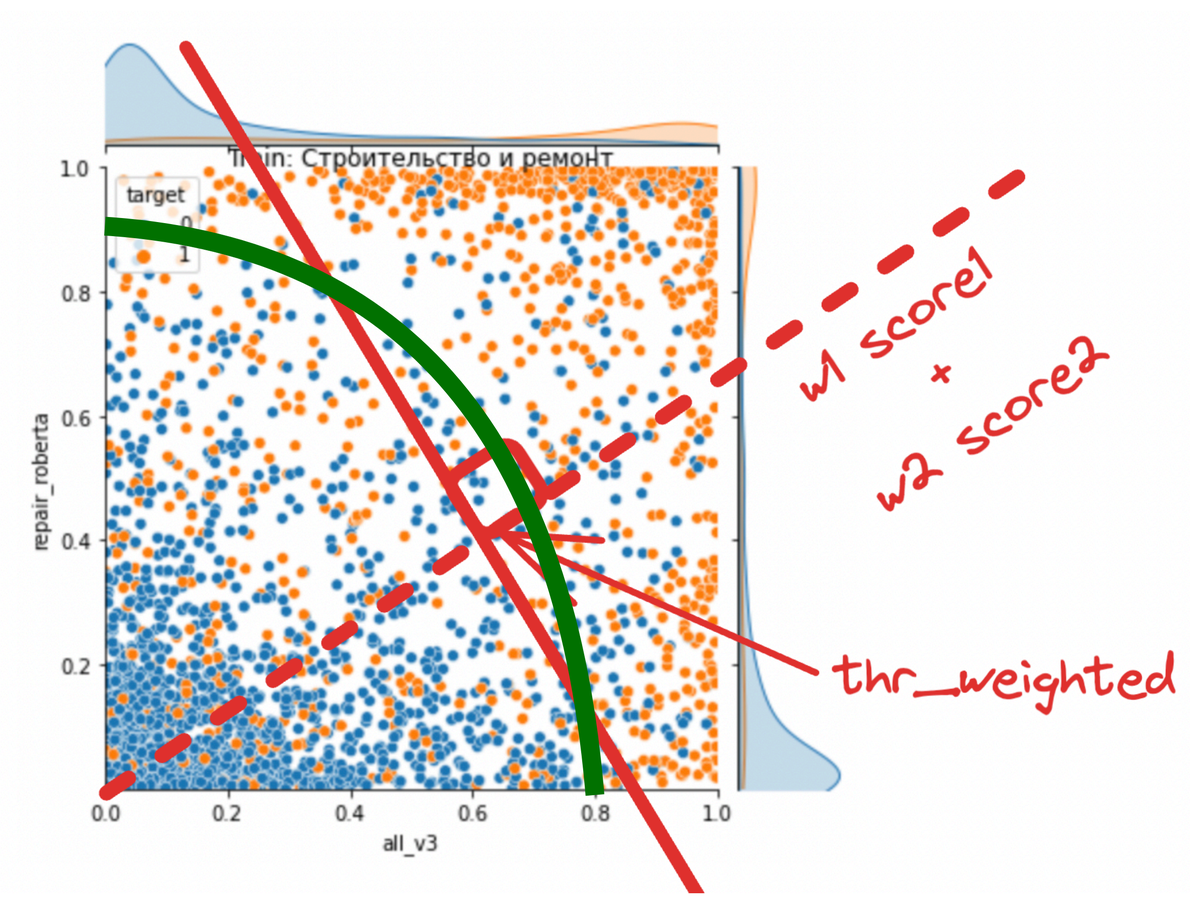


Postprocessing

Union vs Blending



Union



Blending

Postprocessing

Blending

Pros:

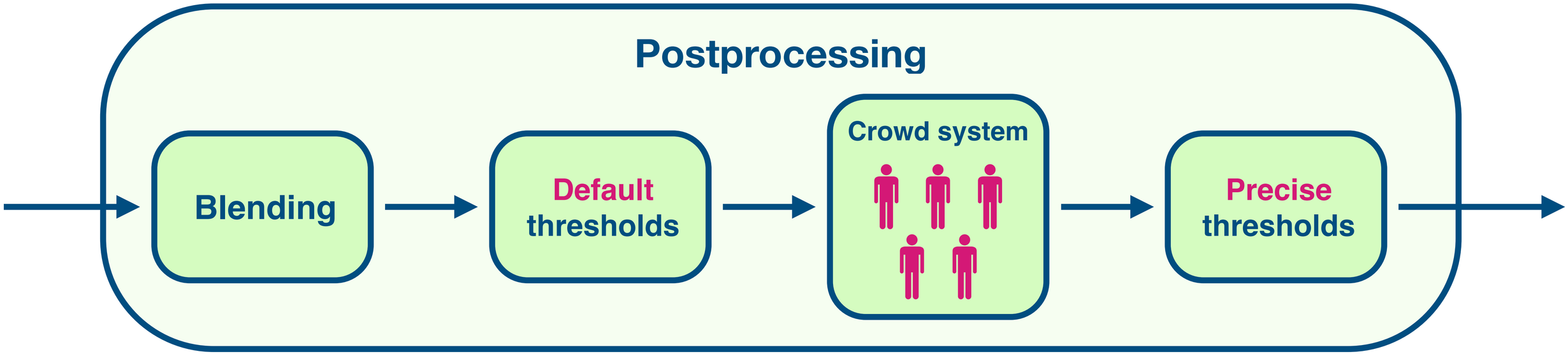
- One score for each pair of items:
  1. Easier to achieve the required precision.
  2. Easier to maintain thresholds.

Cons:

- It is necessary to re-select weights when new models appear.
- It’s not clear how to monitor the quality of individual models.

Postprocessing

Design to ensure the required precision



Postprocessing

Design to ensure the required precision

Pros:

- *The only one way to ensure required precision.

Cons:

- Dependence on people.

Important to keep in mind

Secrets, tricks and folk wisdom

- When the model is trained, it makes sense to look at examples from the test set in which it makes mistakes. You can look at the feature values, you can look at the **shap** values. This way you can find problems in features and in the dataset.

- Divide the test set into examples with correct model predictions and examples with incorrect predictions. Train a binary classification model on this sample. The most important features of this model prevent the original model from making correct predictions on the test sample.