Tasks, resources and evaluation PART-2

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

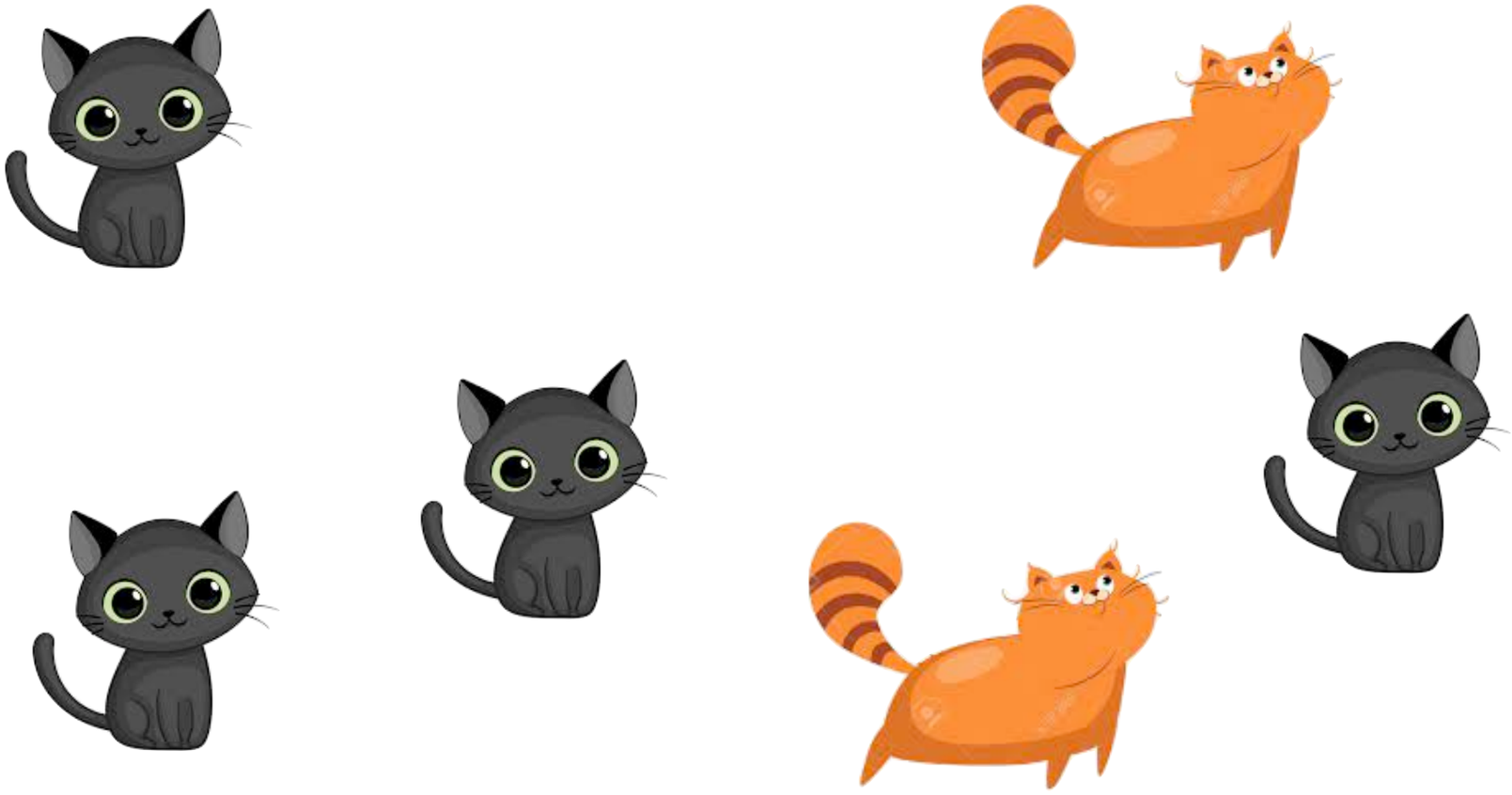
Cristina Bosco

Beyond accuracy

Another metric, with respect to accuracy, for better evaluating system performance is based on the **separate observation** the performance of the model with respect to the different **categories to be classified** for the task.

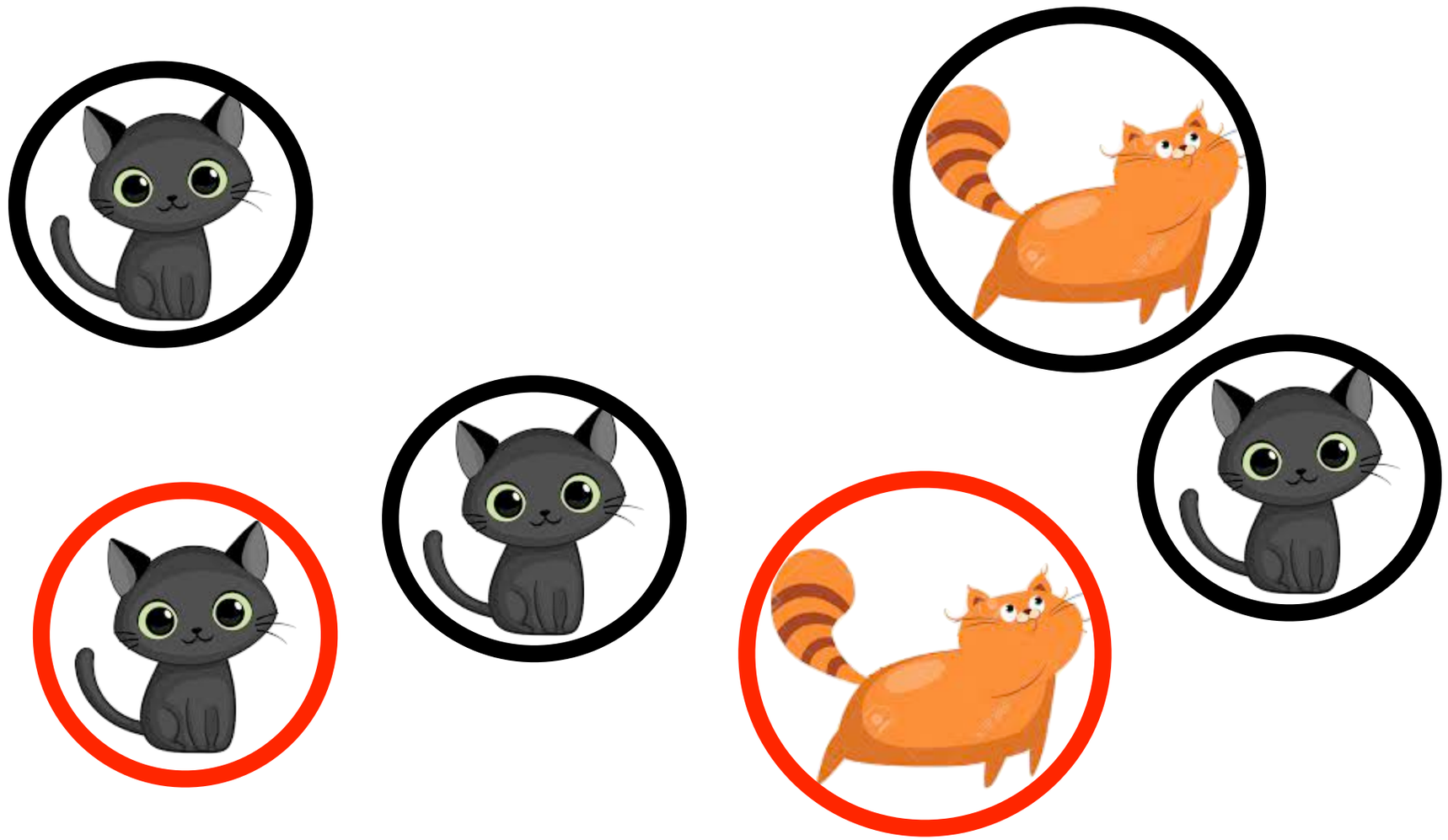
We focus in turn on a single category only, by not considering the other ones and evaluating the system against it alone.

Some statistical notions



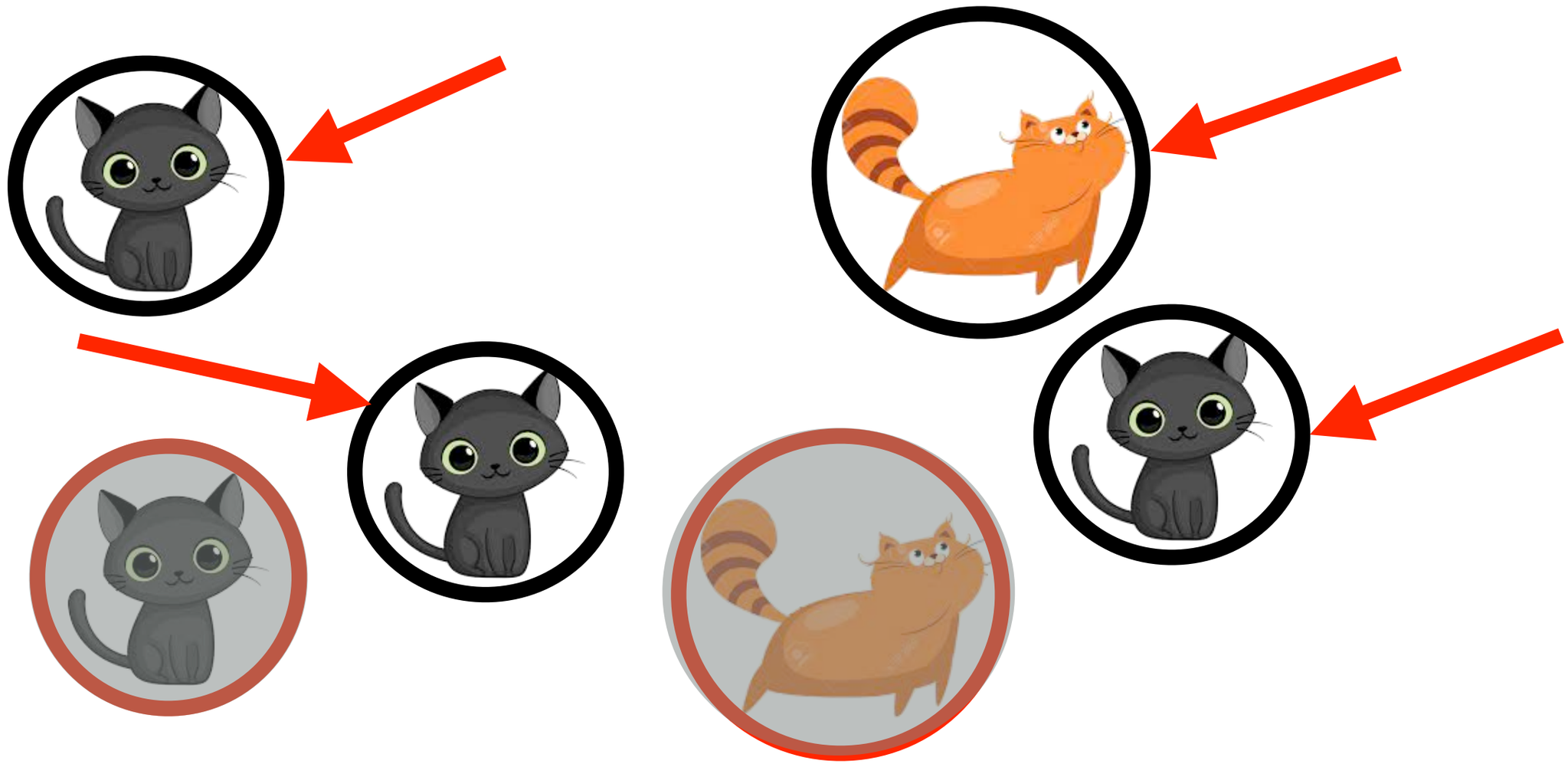
In a **gold standard** set of cats, each element is labeled as **BLACK** or as **RED** (the picture shows how they are labeled in the gold).

Some statistical notions



A **model** (whose output is not perfect, as usually!) labels the set of cats as **BLACK** or as **RED** (the circles shows how the model labels each of them).

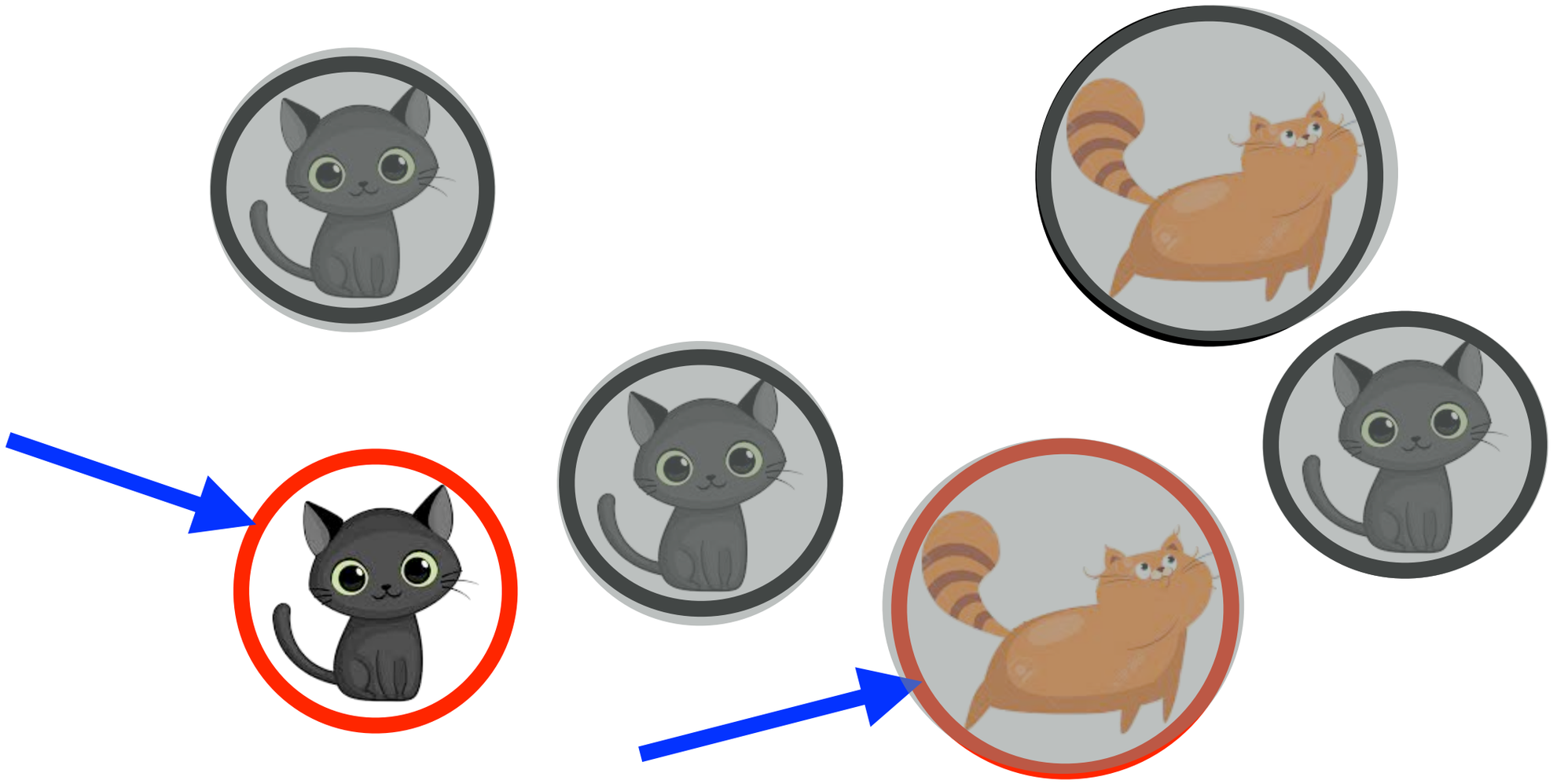
Positives and Negatives



If we focus on the category Black only:

- all the cats the model labels as Black are called ***POSITIVE*** *instances of Black* (indicated by red arrows)

Positives and Negatives

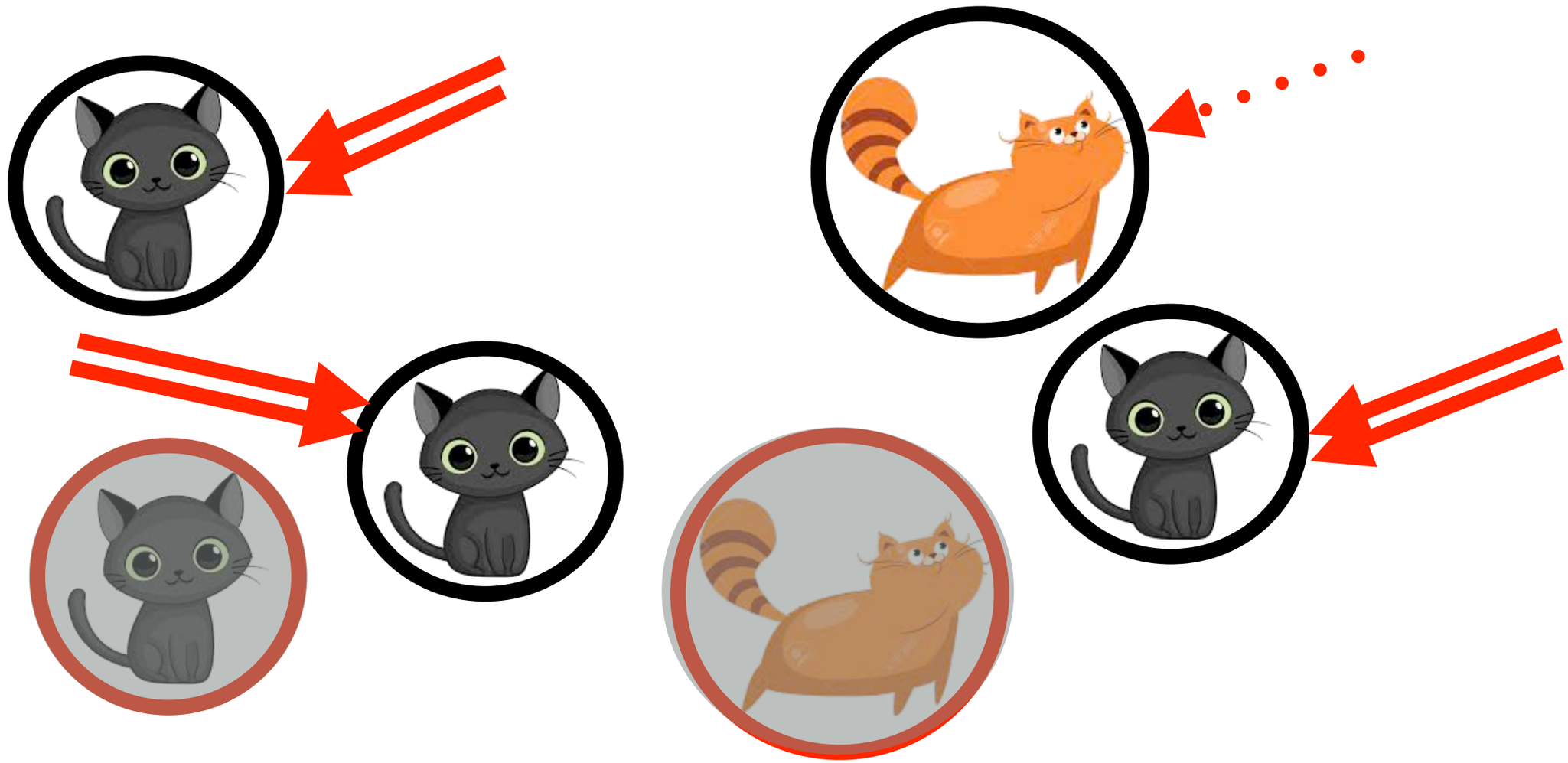


If we focus on the category Black only:

-

- all the cats the model labels as not Black (in this case Red) are called ***NEGATIVE*** *instances of Black* (indicated by blue arrows)

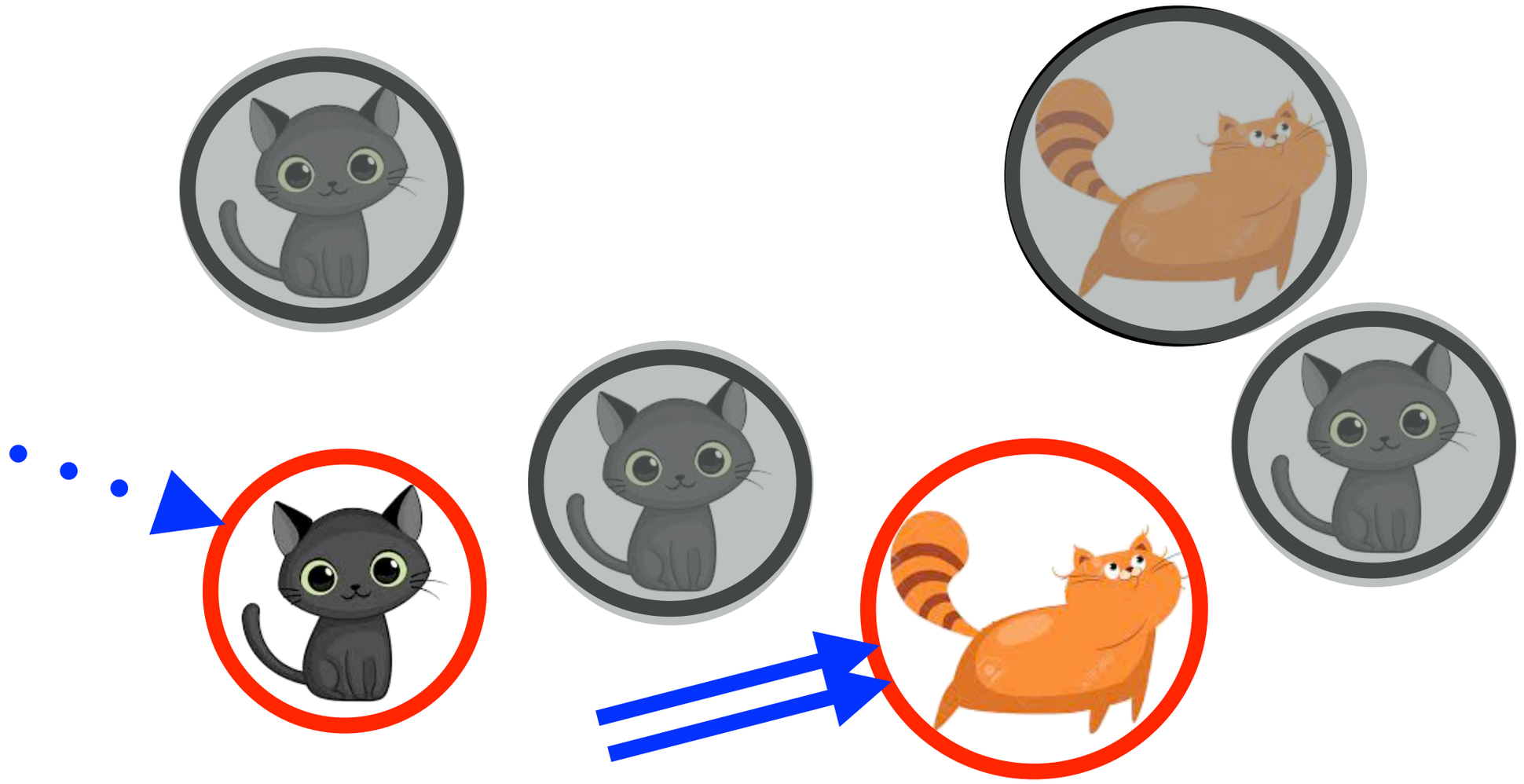
True and False instances



Among the cats labeled as Black by the model there are:

- **TRUEPOSITIVE**s that the model labels as Black and the gold standard too (indicated by double red arrow)
- a **FALSEPOSITIVE** that the model labels as Black but the gold standard does not (indicated by a dashed red arrow)

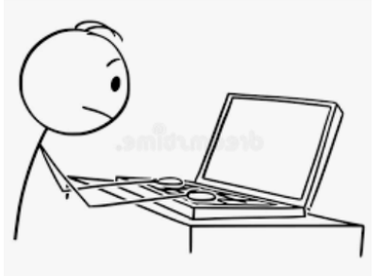
True and False instances



Among the cats classified as non-black by the model are:

- a **TRUENEGATIVE**, which the model classifies as other than Black and also the gold standard (indicated by double blue arrow)
- a **FALSENEGATIVE**, which the model classifies as other than Black, but the gold standard does not (indicated by a dashed blue arrow)

***Partial* PoS tagging**
(with 2 tags only)



TEST SET	Model output	GOLD
The	VERB	nonVERB
cat	nonVERB	nonVERB
run	nonVERB	VERB
in	nonVERB	nonVERB
the	nonVERB	nonVERB
garden	VERB	nonVERB
Mary	nonVERB	nonVERB
sleeps	VERB	VERB
in	nonVERB	nonVERB
the	nonVERB	nonVERB
sun	VERB	nonVERB
…	…	…

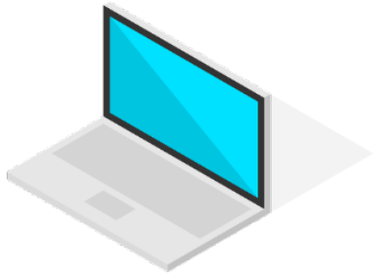
Two categories are annotated in the gold standard test set and must be annotated by the model. They are VERB and nonVERB and can be separately observed in the evaluation process.

True and false positives and negatives

Separately observing the class nonVERB to be categorised by the model, we can provide more precise evaluations of its performance. The model classifies:

- as nonVERB *cat, in, the, Mary, in, the,* that are nonVERB in the gold: they are the **TRUE POSITIVE**s

***Partial* PoS tagging**
(with 2 tags only)

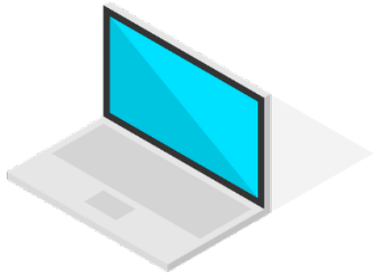


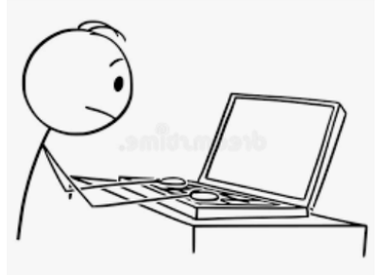
TEST SET	Model output	GOLD	TRUE nonVERB positives
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1 ←
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1 ←
the	nonVERB	nonVERB	1 ←
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1 ←
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1 ←
the	nonVERB	nonVERB	1 ←
sun	VERB	nonVERB	0
…	…	…	…

The model outputs **6 TRUE POSITIVEs (TPs)** for the class nonVERB

= tokens the model classifies as nonVERB and the gold too

Partial PoS tagging
(with 2 tags only)





TEST SET	Model output	GOLD	TRUE nonVERB positives
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1 ←
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1 ←
the	nonVERB	nonVERB	1 ←
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1 ←
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1 ←
the	nonVERB	nonVERB	1 ←
sun	VERB	nonVERB	0
…	…	…	…

The model outputs 6 TRUE POSITIVEs (TPs) for the class nonVERB

= tokens the model classifies as nonVERB and the gold too

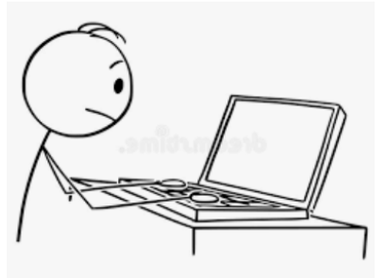
True and false positives and negatives

Separately observing the class nonVERB to be categorised by the model, we can provide more precise evaluations of its performance. The model classifies:

- as nonVERB *cat, in, the, Mary, in, the,* that are nonVERB in the gold: they are the **TRUE POSITIVE**s

- as nonVERB *run,* that is VERB in the gold: it is a **FALSE POSITIVE**

Partial PoS tagging
(with 2 tags only)

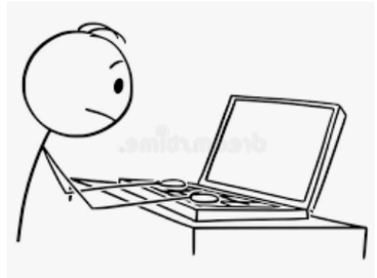


TEST SET	Model output	GOLD	FALSE nonVERB positives
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0 ←
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

The model outputs **1 FALSE POSITIVE (FP)** for the class nonVERB

= a token the model classifies as nonVERB but the gold classifies otherwise

***Partial* PoS tagging**
(with 2 tags only)



FALSE nonVERB positives

TEST SET	Model output	GOLD	
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0 ←
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

The model outputs **1 FALSE POSITIVE (FP)** for the class nonVERB

= tokens the model classifies as nonVERB but the gold classifies otherwise

True and false positives and negatives

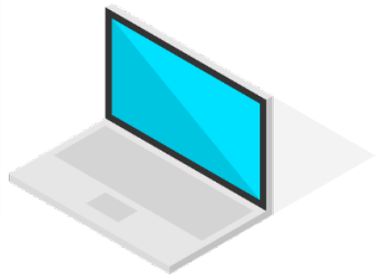
Separately observing the class nonVERB to be categorised by the model, we can provide more precise evaluations of its performance. The model classifies:

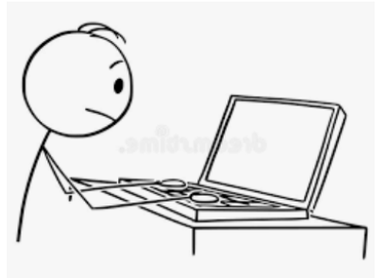
- as nonVERB *cat, in, the, Mary, in, the,* that are nonVERB in the gold: they are the **TRUE POSITIVE**s

- as nonVERB *run,* that is VERB in the gold: it is a **FALSE POSITIVE**

- as VERB *sleeps,* that is VERB in the gold: it is a **TRUE NEGATIVE**

Partial PoS tagging
(with 2 tags only)



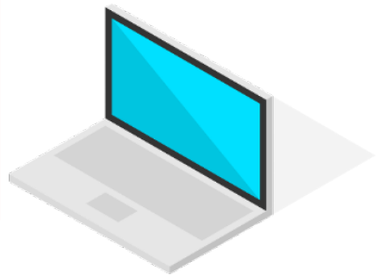


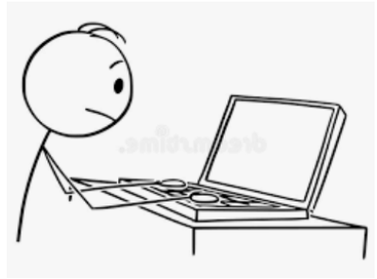
TEST SET	Model output	GOLD	TRUE nonVERB negatives
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

The model outputs 1 TRUE NEGATIVE (TN) for the class nonVERB

= tokens the model classifies as VERB and the gold too

Partial PoS tagging
(with 2 tags only)





TEST SET	Model output	GOLD	TRUE nonVERB negatives
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

The model outputs 1 TRUE NEGATIVE (TN) for the class nonVERB

= tokens the model classifies as VERB and the gold too

True and false positives and negatives

Separately observing the class nonVERB to be categorised by the model, we can provide more precise evaluations of its performance. The model classifies:

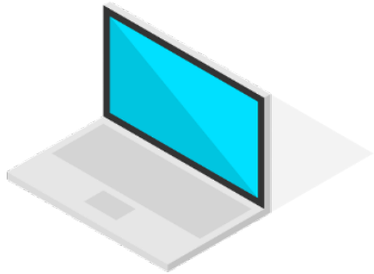
- as nonVERB *cat, in, the, Mary, in, the,* that are nonVERB in the gold: they are the **TRUE POSITIVE**s

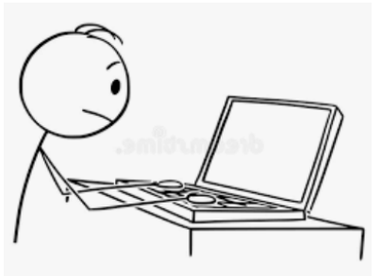
- as nonVERB *run,* that is VERB in the gold: it is a **FALSE POSITIVE**

- as VERB *sleeps,* that is VERB in the gold: it is a **TRUE NEGATIVE**

- as VERB *the, garden, sun,* that are nonVERB in the gold: they are the **FALSE NEGATIVE**s.

***Partial* PoS tagging**
(with 2 tags only)





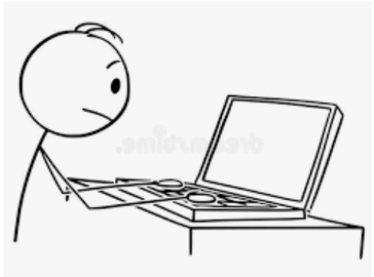
FALSE nonVERB negatives

TEST SET	**Model output**	**GOLD**	
The	VERB	nonVERB	0 ←
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0 ←
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0 ←
…	…	…	…

The model outputs 3 **FALSE NEGATIVE (Fns)** for the class nonVERB

= tokens the model classifies otherwise but the gold classifies as nonVERB

Partial PoS tagging
(with 2 tags only)



FALSE nonVERB negatives

TEST SET	Model output	GOLD	
The	VERB	nonVERB	0 ←
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0 ←
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0 ←
…	…	…	…

The model outputs 3 **FALSE NEGATIVE (Fns)** for the class nonVERB

= tokens the model classifies otherwise but the gold classifies as nonVERB

Precision

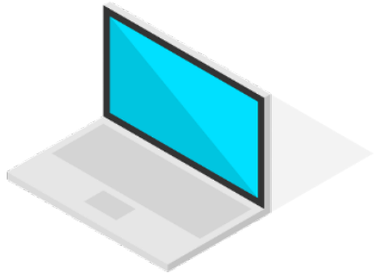
Dividing the number of TRUE POSITIVEs classified by the model for the total of all the tokens it classifies (as TRUE POSITIVEs or as FALSE POSITIVEs) we can calculate a measure called PRECISION.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Precision answers the question:

Are the tokens classified as nonVERB in effect nonVERB (in the gold) ?

Partial PoS tagging
(with 2 tags only)



TRUE and FALSE positives

TEST SET	Model output	GOLD	
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

Given 6 TRUEPositives and 1 FALSEPositive,

the Precision of the model is 6 : (6 + 1) = 0.85

Precision

Maximum precision = 1:
a model that classifies only nonVERB tokens with category nonVERB; all classifications provided by the model are correct and correspond to the gold; the model provides only TRUE POSITIVES and no FALSE POSITIVES.

TRUEpositives **FALSEpositives**

Minimum precision = 0:
a model that does not classify any nonVERB token with category nonVERB; all classifications provided by the model do not correspond to the gold; the model provides only FALSE POSITIVES and no TRUE POSITIVES.

Precision

Is precision sufficient to evaluate the result of the model?

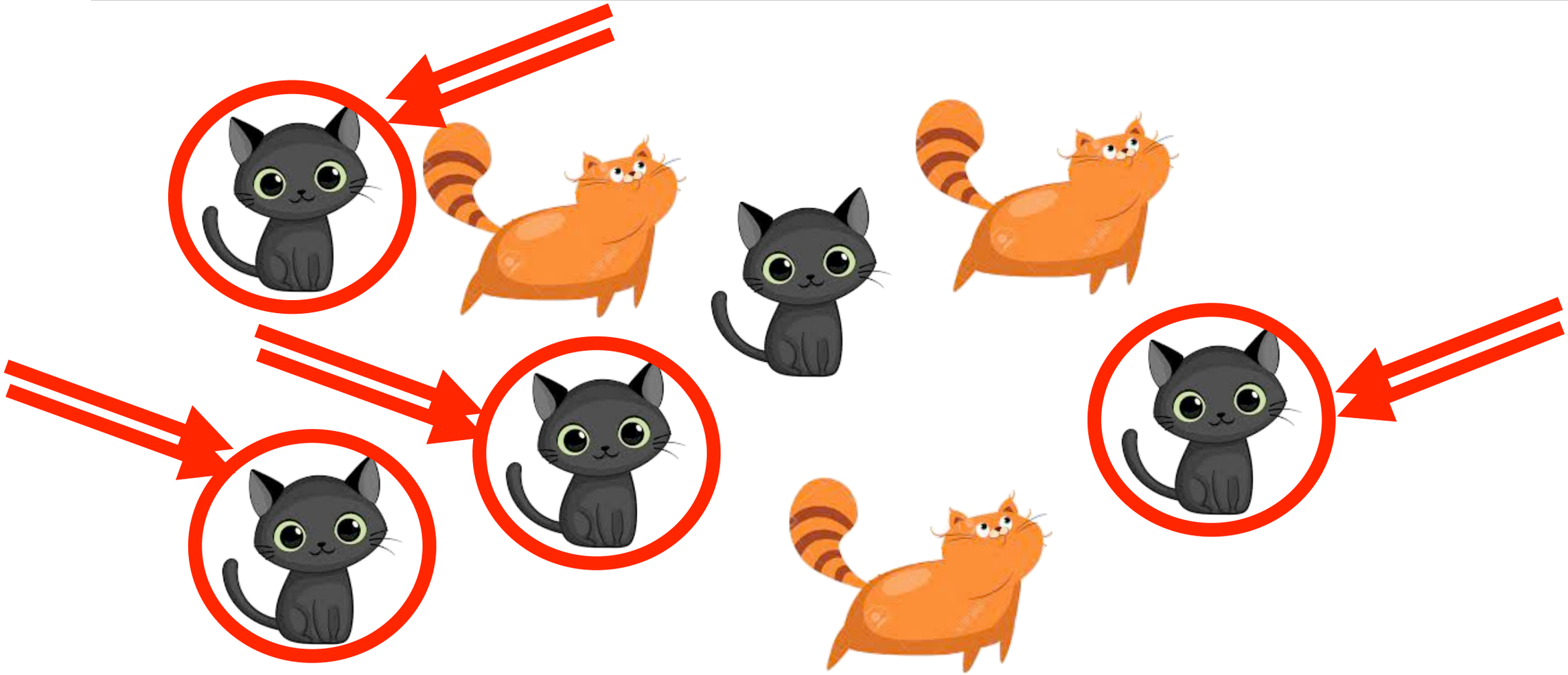
Precision measure tells us only if the classified tokens have been classified correctly.

But **Precision does not say anything about the amount of tokens classified by the model**.

Precision

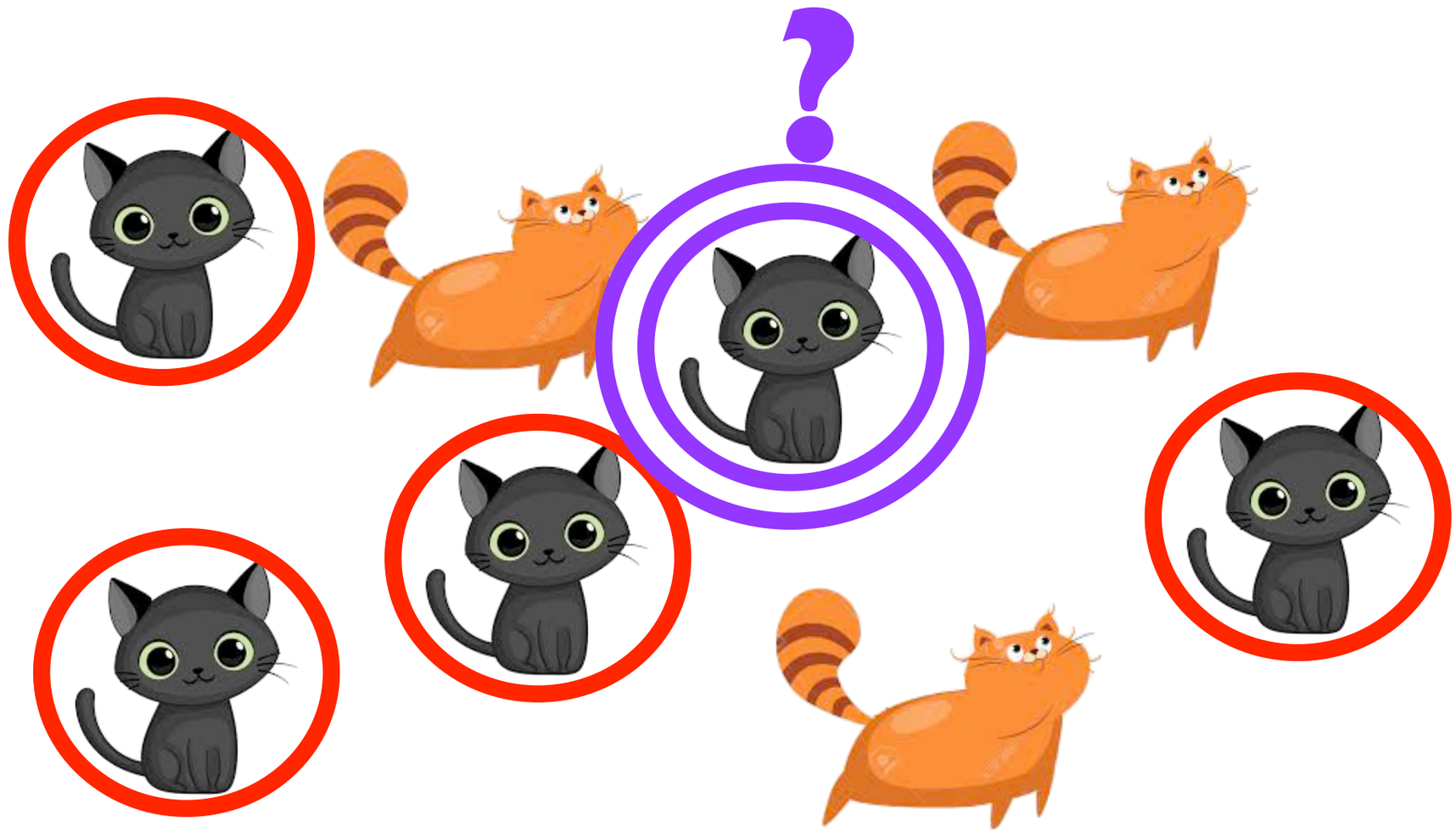
The model has to classify the category **black cats** in the set above.

Precision



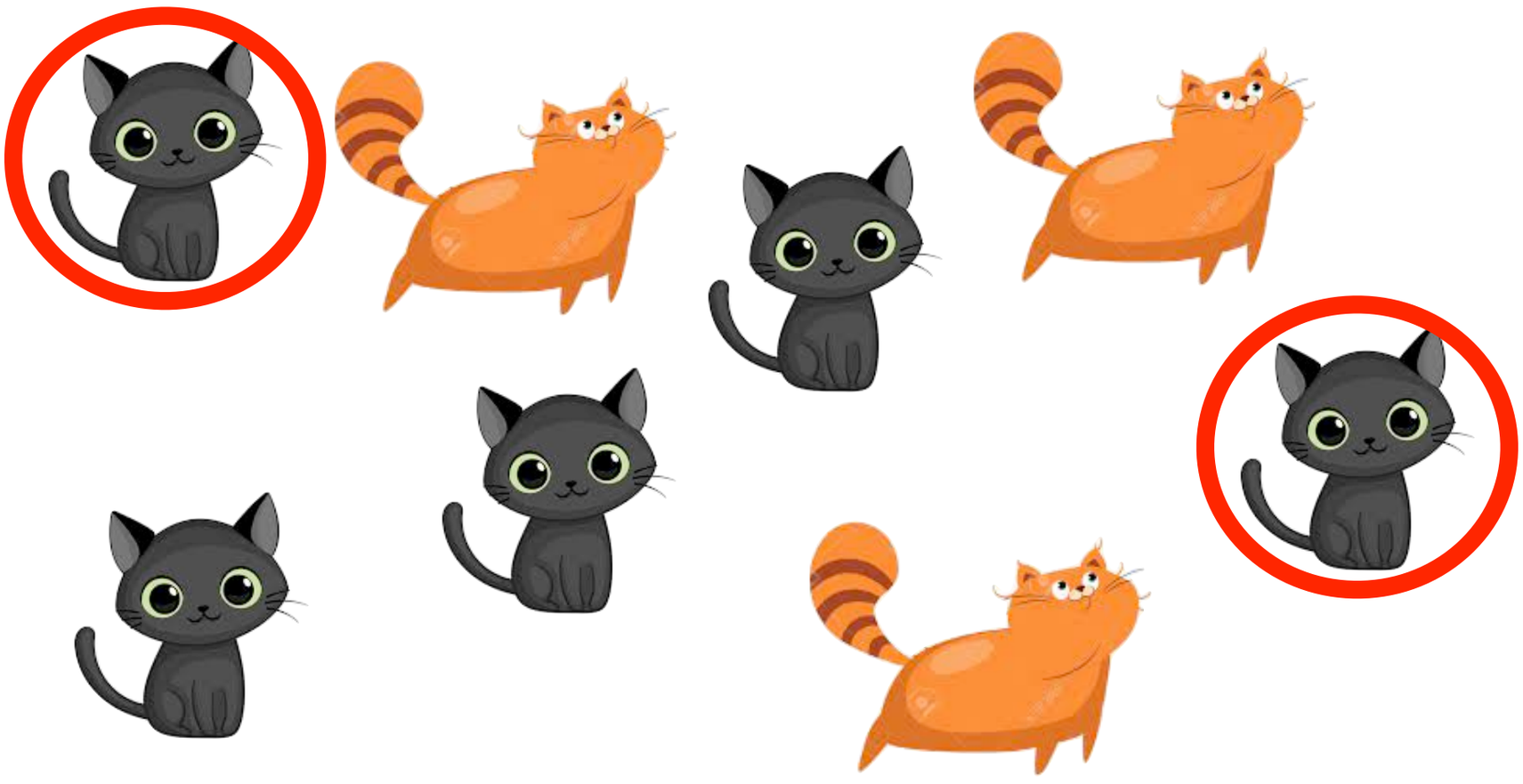
The system identifies **4** black cats. Its precision is
4 TruePositives : (4 TruePositives + 0 FalsePositives) = 1
A very good result! **But is this model really good?**

Precision



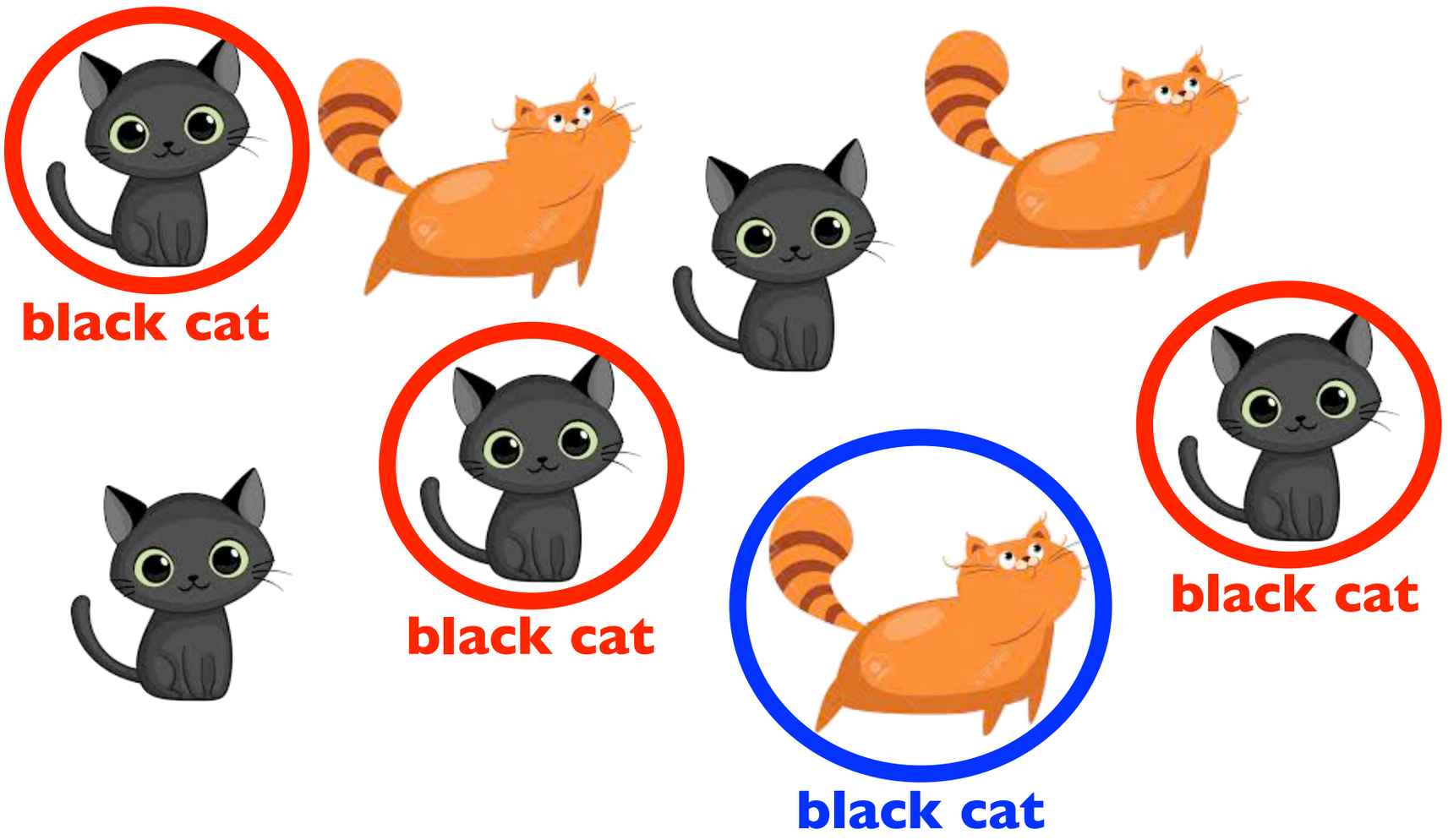
Although the model reaches the precision maximum score (1),
it fails to identify some examples of the class to be identified.

Precision



The precision maximum score can be also achieved by models that fail to identify several examples … if they do not generate **FALSEpositive**s!

Precision



The precision score decreases only when the model classifies as black cats also some red cats:

3 TruePositives : (3 TruePositives + 1 FalsePositive) = 0.75

Precision and Recall

Precision alone is not enough!

Precision measure tells us only if the classified tokens have been classified correctly. It is only sensitive to misclassification.

Another measure called RECALL tell us how many of the tokens that must be classified are actually classified.

Recall

Dividing the number of TRUE POSITIVEs classified by the model for the total of all the tokens it classifies (as TRUE POSITIVEs or as FALSE NEGATIVEs) we can calculate a measure called RECALL.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Recall answers the question:

How many of the tokens that must be classified with a given category are actually classified by the model with that category?

Recall

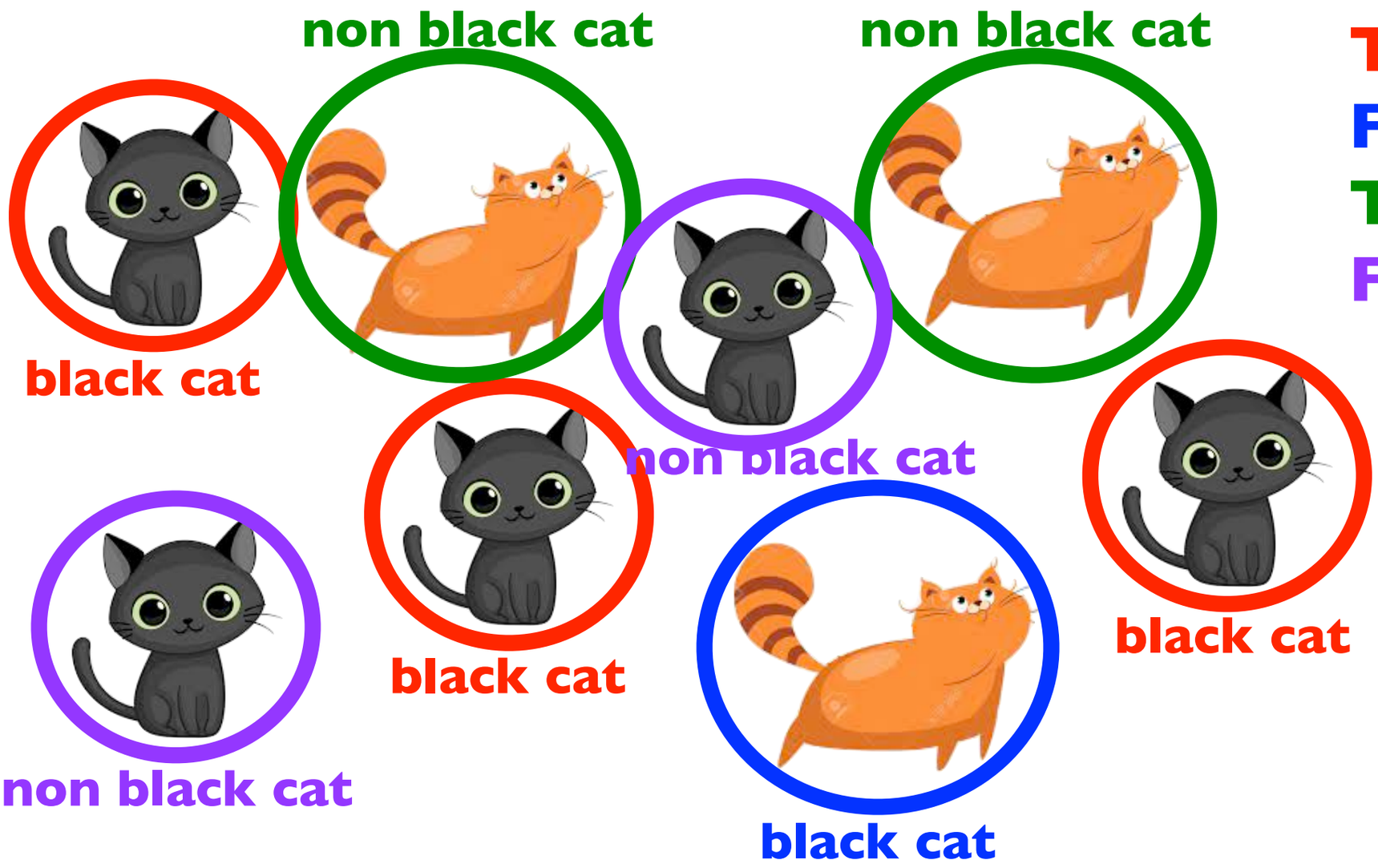
The sum of TruePositives and FalseNegatives for a given class exactly corresponds to the number of tokens that are categorised with that class in the gold.

Among them TruePositives are those correctly classified by the model, while FalseNegatives are those uncorrectly classified by the model.

Recall

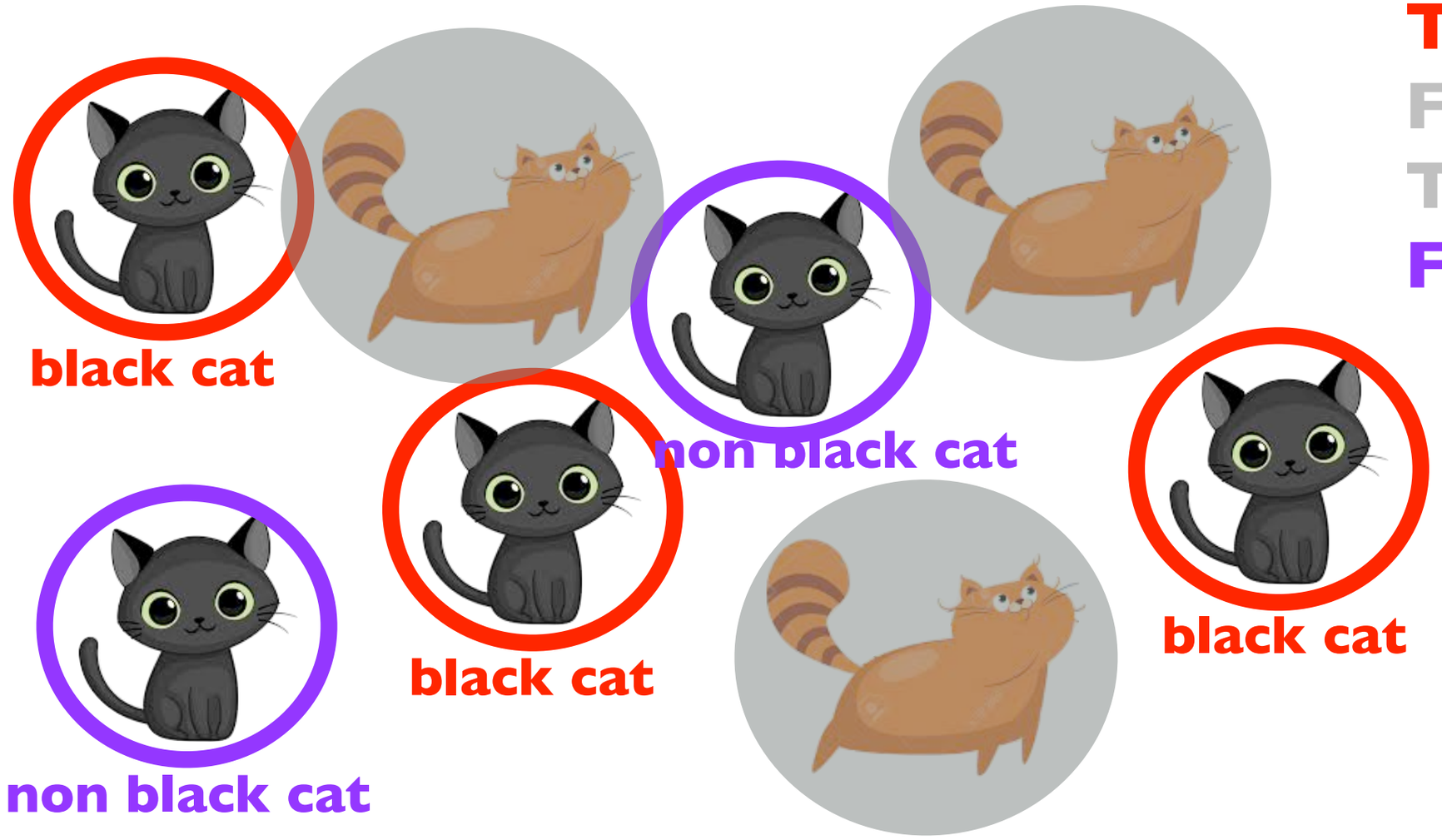
The model has to classify the category **black cats** in the set above.
We see that the elements classified as black cats in the gold are 5.

Recall



The model classifies the category **black cat**: some elements are correctly classified (TPs) and some incorrectly (FPs). It classifies also the category **non black cat** generating some TNs and some Fns.

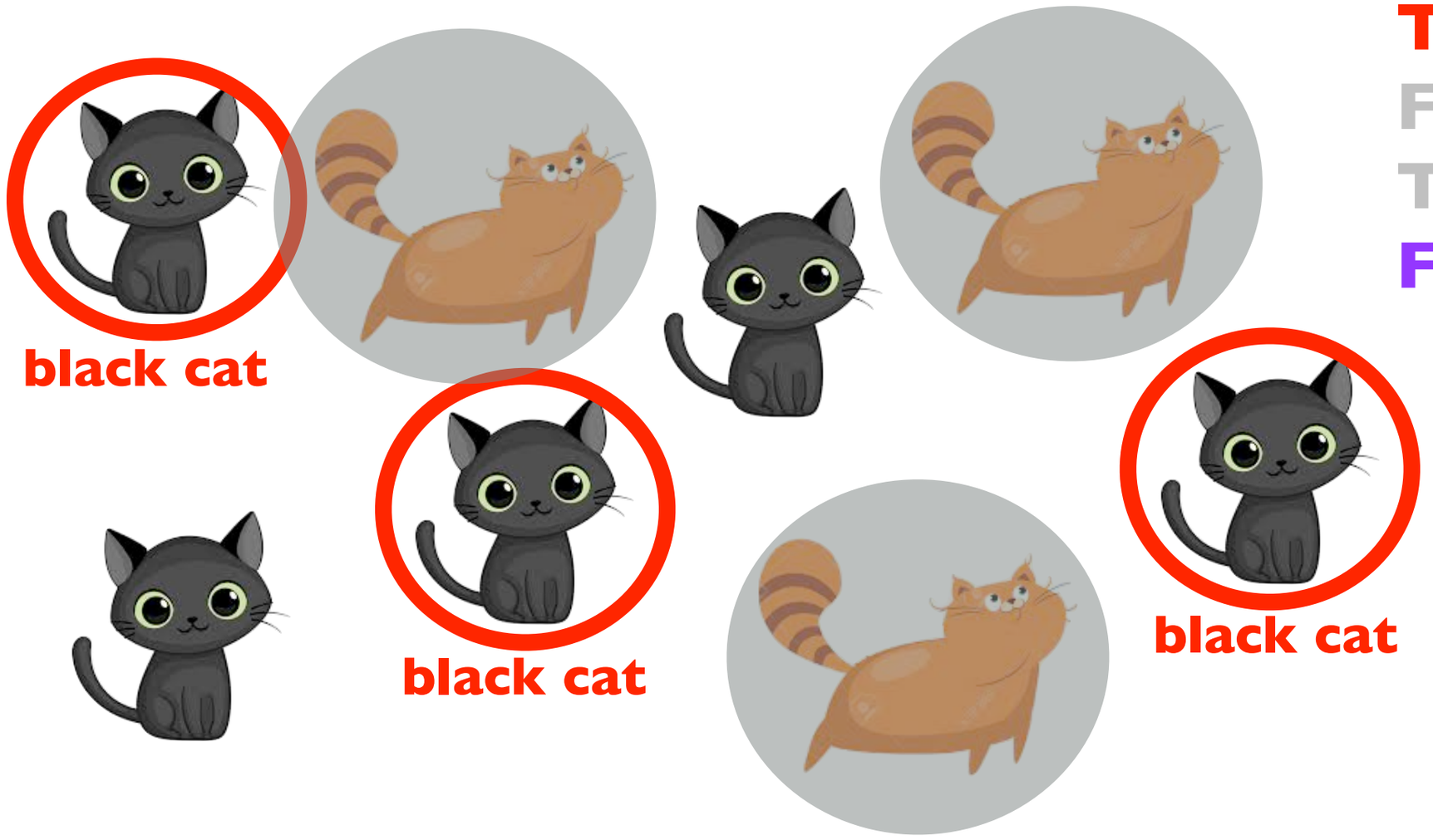
Recall



TruePositive
FalsePositive
TrueNegative
FalseNegative

We want to **focus only on the elements that are classified as black cat** in the gold.

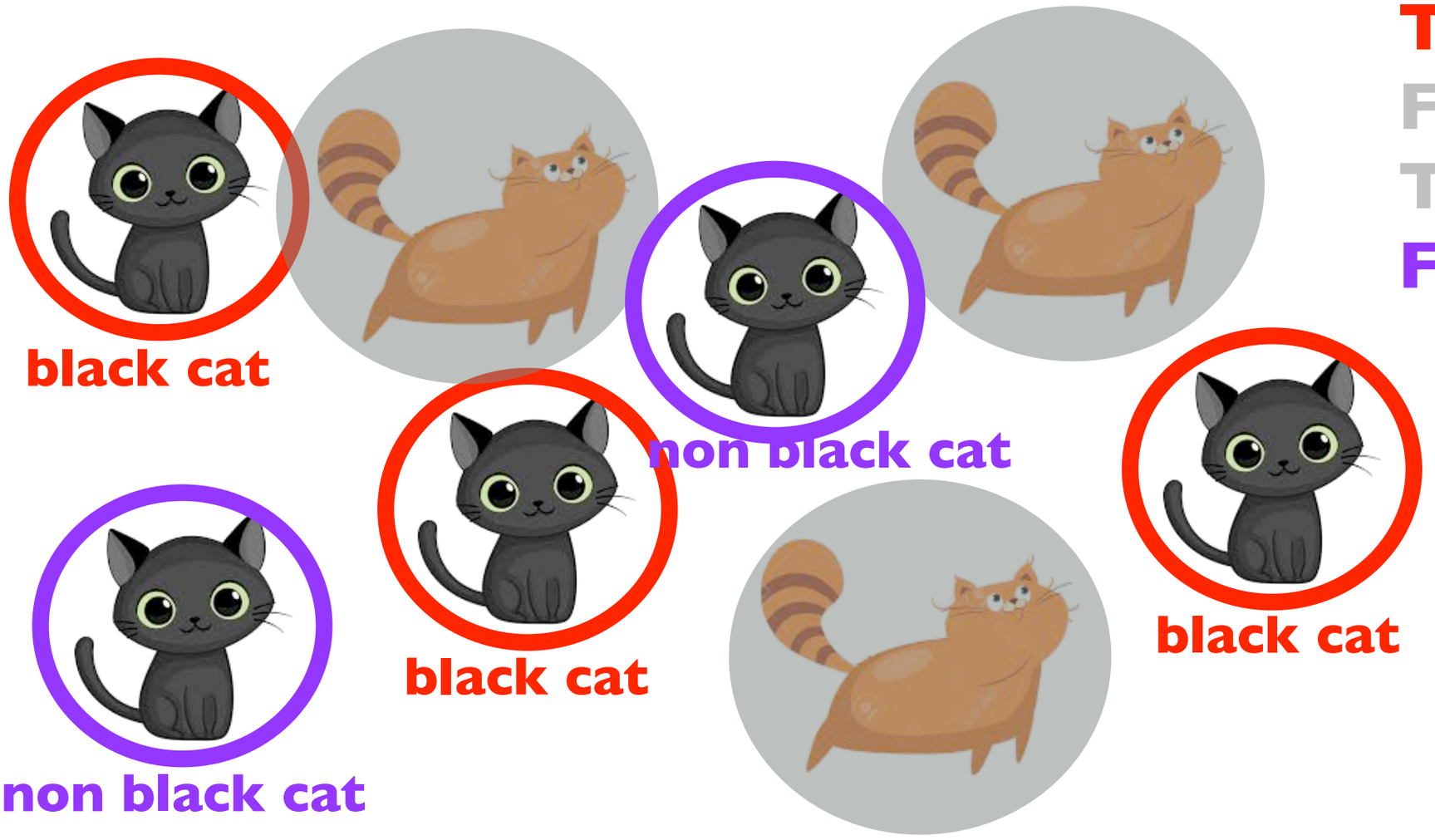
Recall



TruePositive
FalsePositive
TrueNegative
FalseNegative

Among the 5 elements that are **black cats** in the gold, 3 are classified by the model as **black cats (TPs)**

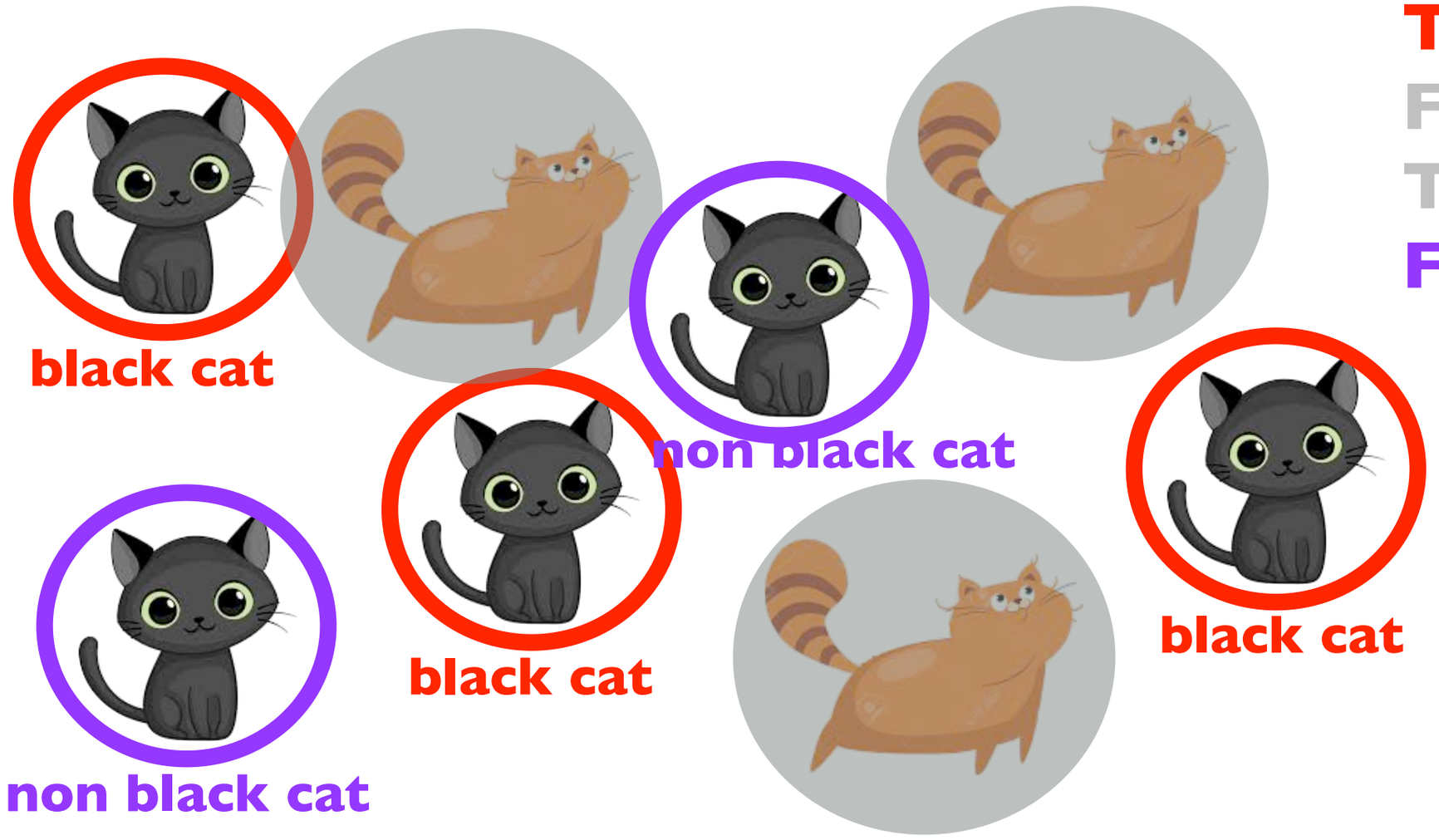
Recall



TruePositive
FalsePositive
TrueNegative
FalseNegative

Among the 5 elements that are **black cats** in the gold, 3 are classified by the model as **black cats (TPs)** while 2 are classified by the model as **non black cats (FNs)**.

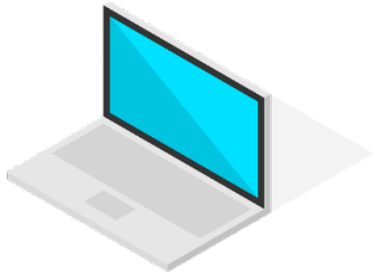
Recall

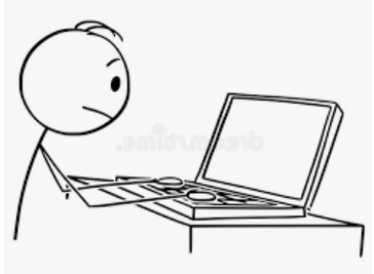


TruePositive
FalsePositive
TrueNegative
FalseNegative

The system identifies **3** black cats (over 5). Its recall is
3 TruePositives : (3 TruePositives + 2 FalseNegatives) = 0.6

Partial PoS tagging
(with 2 tags only)





TRUE positives and FALSE negatives

TEST SET	Model output	GOLD	
The	VERB	nonVERB	0
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	VERB	nonVERB	0
Mary	nonVERB	nonVERB	1
sleeps	VERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	VERB	nonVERB	0
…	…	…	…

Given 6 TRUEPositives and 3 FALSENegatives,

the Recall of the model is 6 : (6 + 3) = 0.66

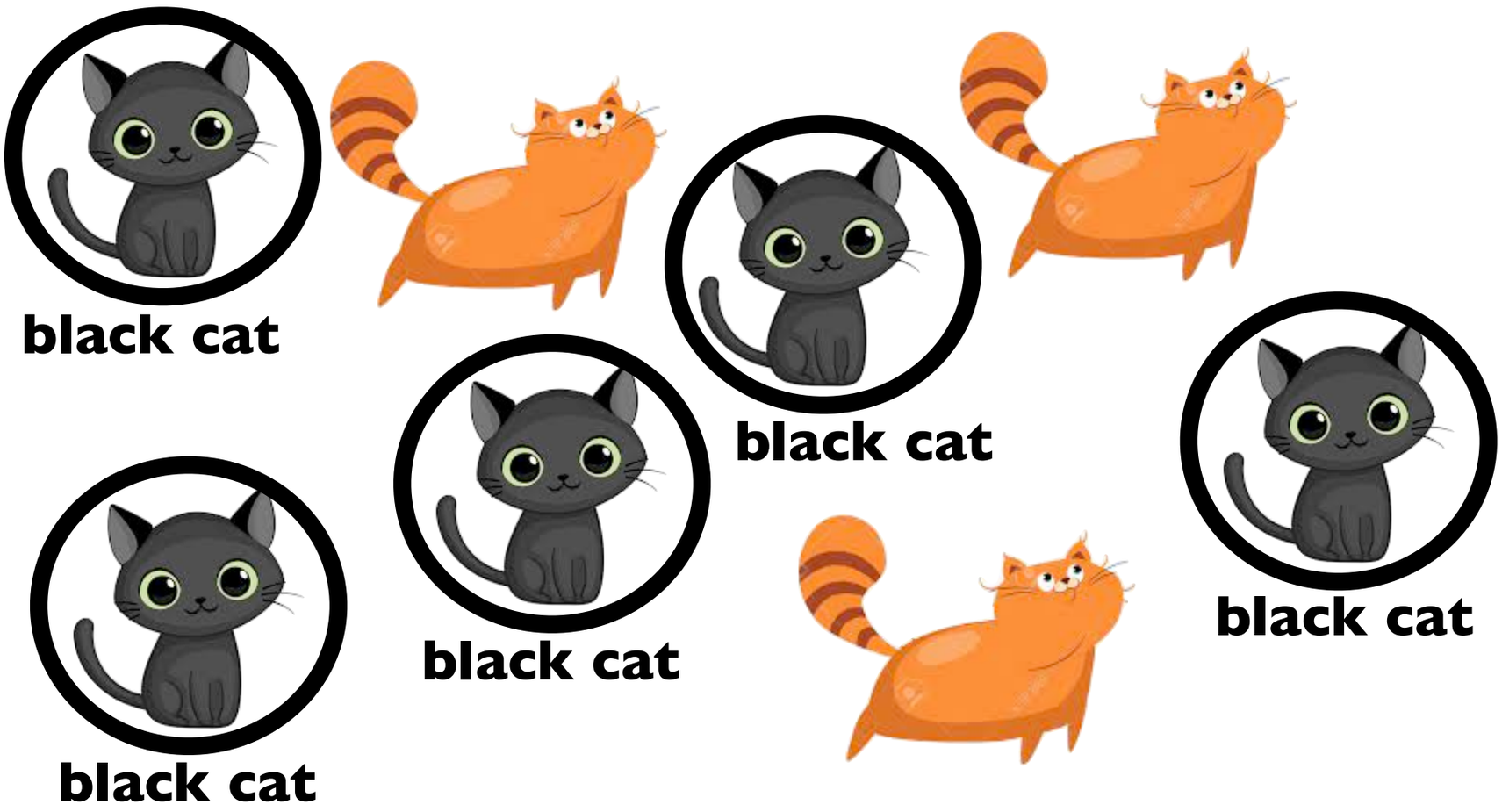
Recall

Maximum recall = 1:
a model that classifies all nonVERB tokens with category nonVERB; all classifications provided by the model are correct and correspond to the gold; the model provides only TRUE POSITIVES and no FALSE NEGATIVES.

TRUEpositives **FALSEnegatives**

Minimum recall = 0:
a model that does not classify any nonVERB token with category nonVERB; all classifications provided by the model do not correspond to the gold; the model provides only FALSE NEGATIVES and no TRUE POSITIVES.

Recap: precision and recall



The model has to classify the category **black cats** in the set above. It is important that it classifies correctly each cat as black (precision), but also that it classifies all the cats that are black (recall)! **All and only the black cats!**

F=Precision+Recall

Putting together **precision** and **recall** we have a better evaluation in which it is shown how many words are correctly classified: it is called **F1-score** and is calculated as the ARMONIC MEAN of Precision and Recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$