

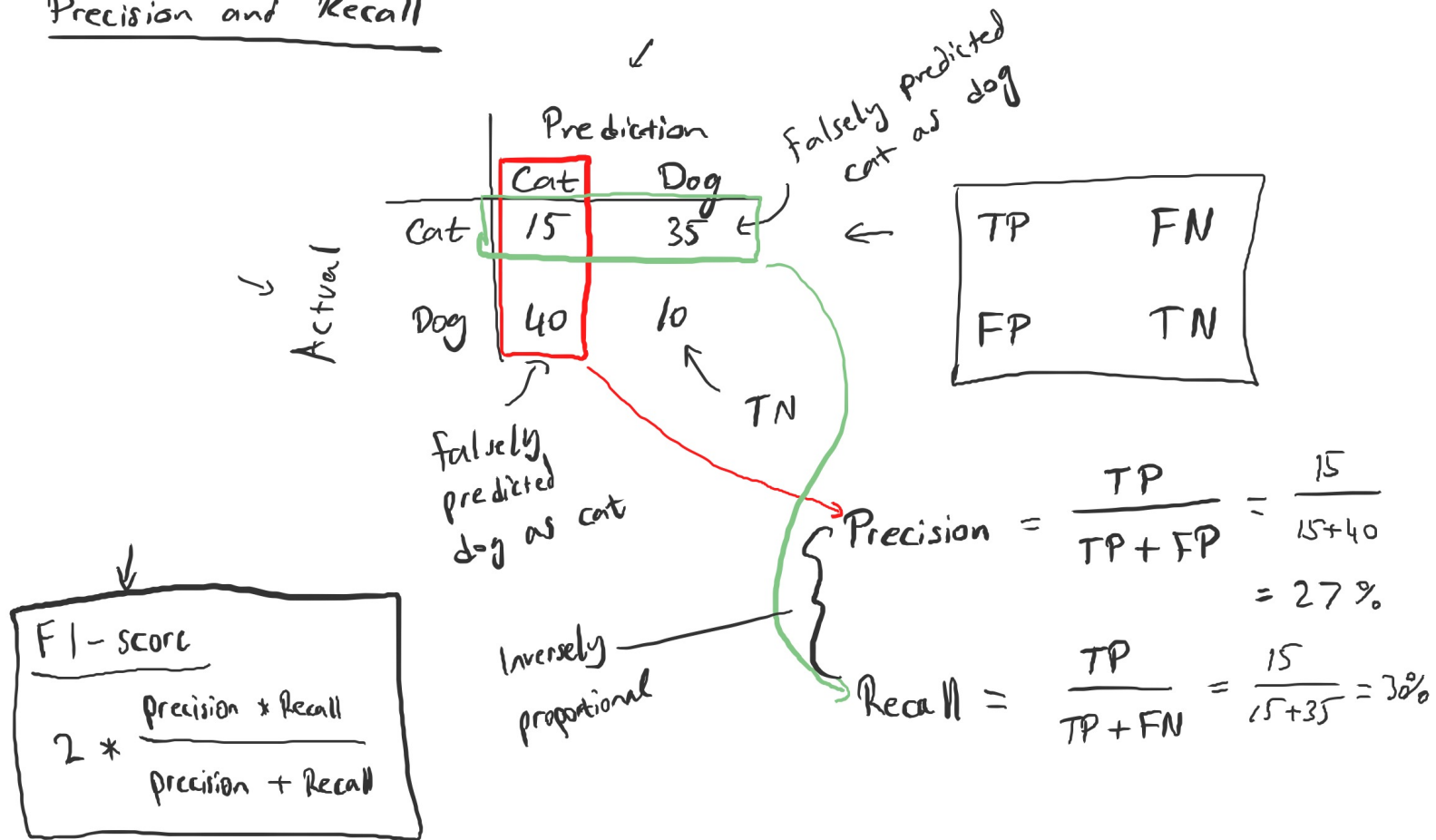
## Evaluation Metrics in NLP

- Accuracy
  - Precision
  - Recall
  - F1-score
- • BLEU score (ex. Machine Translation)
- NLP

### Accuracy ←

- How many samples did your model predict w/ the correct label out of all samples?

## Precision and Recall



## BLEU (bilingual evaluation understudy)

ROUGE

French: Le chat est sur le tapis

Human

Reference 1: The cat is on the mat. ←

Reference 2: There is a cat on the mat. ←

Model Output: the the the the the the the ←

Modified Precision (BLEU) =  $\frac{2}{7}$

## BLEU w/ bigrams

Model Output: The cat the cat on the mat

bigrams ↓

The cat  
cat the  
cat on  
on the  
the mat

Count

2  
1  
1  
1  
1

count-clip

1  
0  
1  
1  
1

count-clip

$$= \frac{\text{count}}{6} = \frac{2}{3}$$

sum  
↓

1

sum  
↓