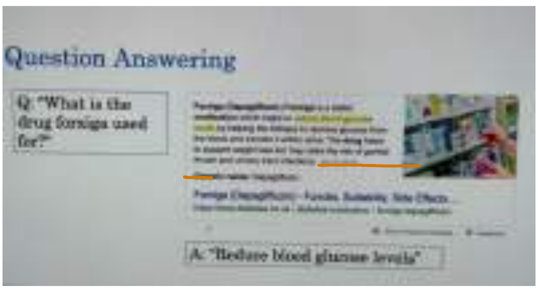
# Medical question answering:

- Medical question answering using Natural Language Processing.

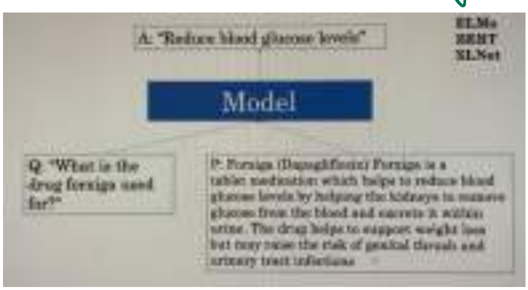
- We will learn how to extract labels for x-ray classification model from radiology reports.

- If a patient or doctor wants to know more about a medical diagnosis or a treatment
  ↳ One way is that they can ask question in natural language.

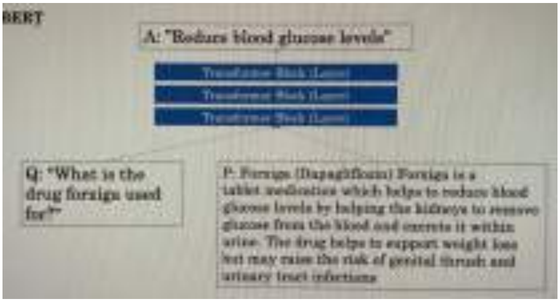
- Generally, the search engines are able to find the passage of text containing the answer. The challenge is the last step of answer extraction.




Our model will thus take in a user question and, called Q and a passage that contains the answer to the question. And the model will produce an answer that is extracted from this passage.



- We will look at the BERT model in particular. The BERT model consists of several layers called transformer blocks.



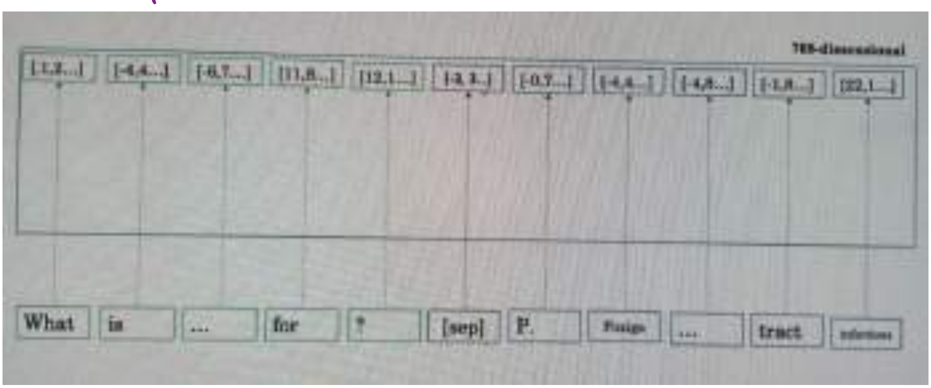

- The inputs to a BERT model are question & passage.
  How do we input text?
  We can break up the question & passage into tokens or words.

  Q: What is the drug forxiga....

  What | is | the | drug | .....

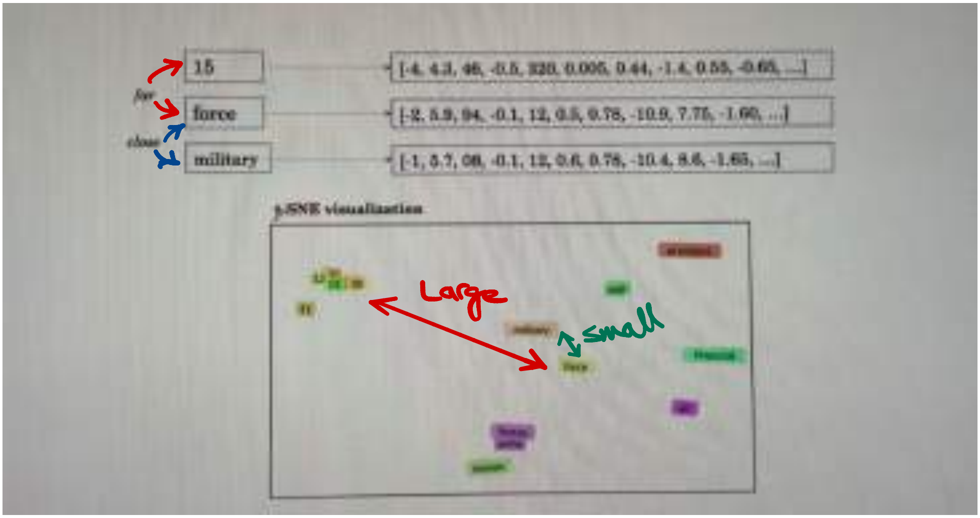
We separate the inputs from question & passage using a special token called the separator token.

- These inputs pass into the model, where they pass through several transformer blocks and are transformed into a list of vectors. There's one 786-dimensional vector for each of the words. This is called the word representation for a word.



Word representations represent words in a way that capture meaning related relationships between words.

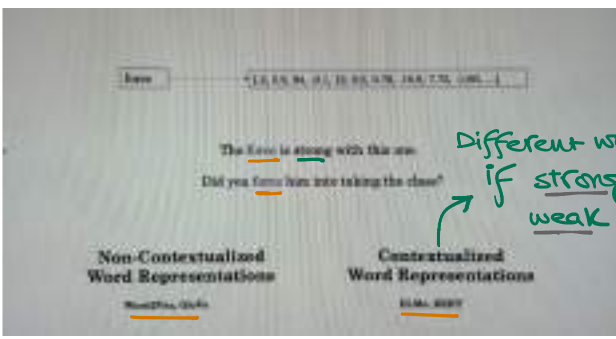
- Distances between words capture how related they are or how often they are used in a similar context.




- One of the main challenge of word representation is how to deal with words that have multiple meanings.

Non-Contextualized uses a single word representation for a word.

Contextualized uses word representation based on the context.




Different word representation if strong changes to weak

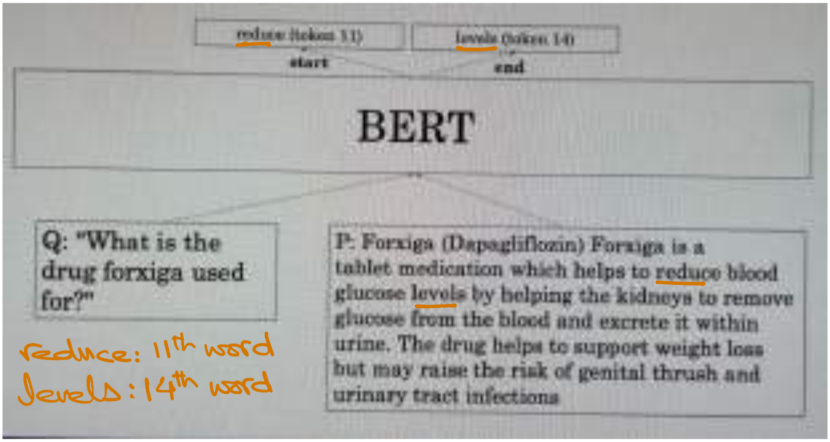
Q: How BERT learns these contextualized word representation?

- Words from a passage text are input into a BERT model.
- Then one of the tokens in the passage is masked with a special MASK token.
- The model is trained to predict what the MASK was. An extra layer is added where the output is the probabilities of the missing word.
- In the process of learning to correctly predict the masked word, the model learns word representations.


- There have been extension to BERT model like Bio-BERT, that uses passages from medical papers to learn these word representations. The advantage is to learn words in the content of medicine.

Q: How we can use BERT for the task of question answering?

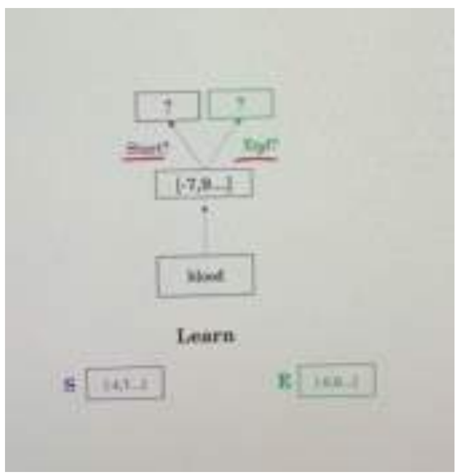
- We know for this task Question & Passage will be the input and Answer will be the output.
  ↳ We can represent the answer using the start and end words of answer.



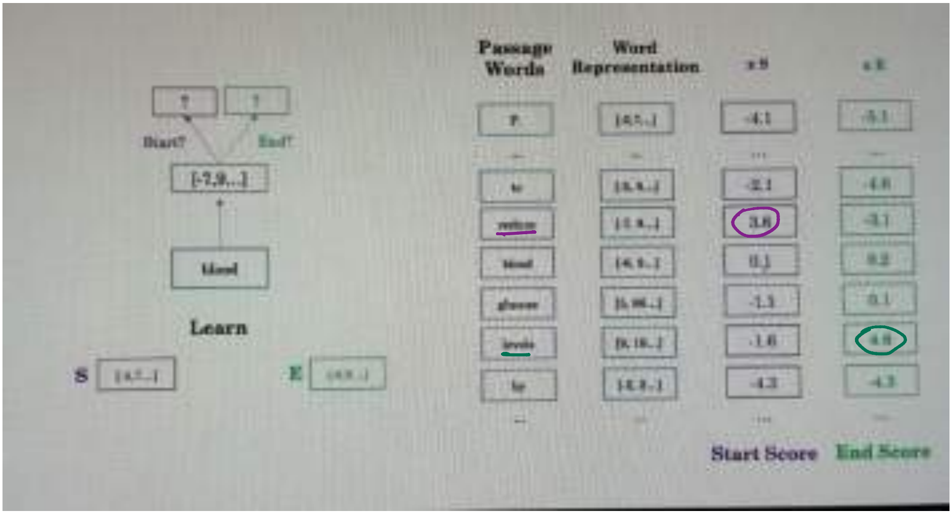

- The task for the model is to determine whether a word is likely to be the start or the end word to an answer.


Q: How the model learns about the start & end words.

- The model learns two vectors, S and E for each of the word representations, for each of the words in the passage.
- The word representation is multiplied by S to get a single number, which is the start score. The higher the

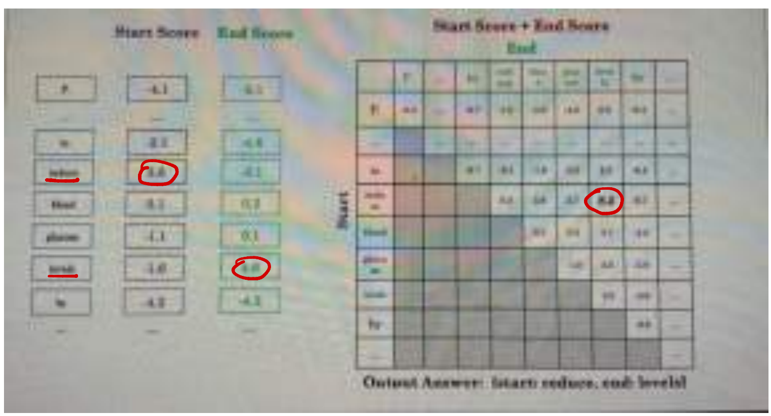


start score, the more likely it is to be the start word. Similarly, we multiply the word representation with the vector E to find the End Words.




- To find the most likely start and end words, we calculate the start & end score. Then we form a Grid of score. We can force the end word to appear no earlier than the start word by only computing the scores in this upper triangular region.

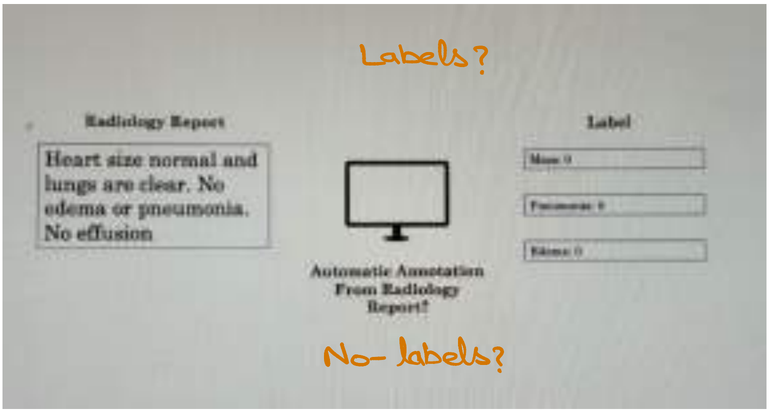
So, we have the model output 'reduce' as the start of the answer, and 'levels' as the end of the answer.




Typically the model is first shown natural question and answers in English in the general domain and then fine tune on medical datasets.

# Automatic label extraction:

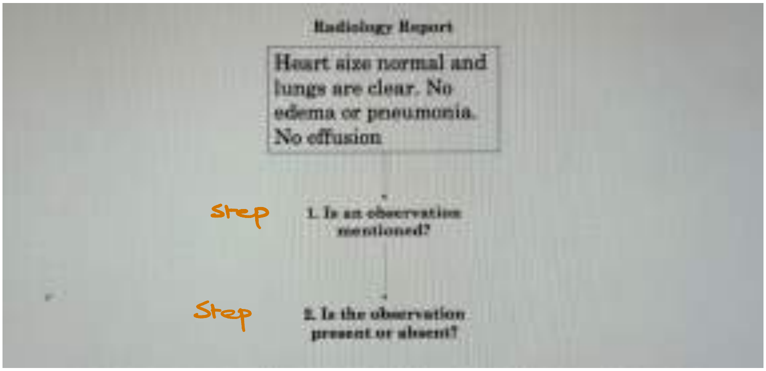
- How to get the corresponding labels for each image required to train this algorithm.
    - One idea would be to ask a radiologist expert to annotate every image. However it would be time & cost intensive.
    - An alternative way would be to use the reports to extract labels. We can use a machine to read through these radiology reports to automatically perform this labeling.



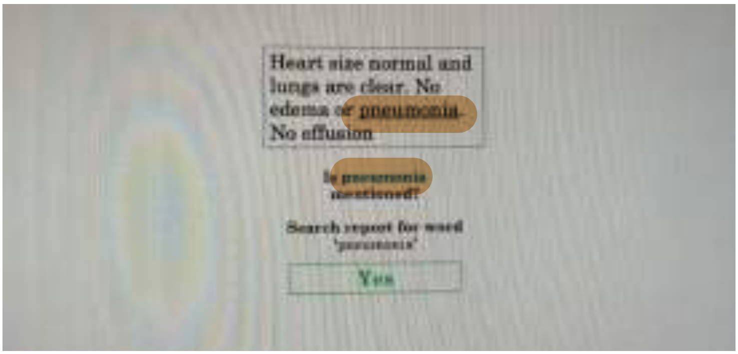

- If we have labels then this is supervised learning task and we can use a BERT model to train it.

- If we don't have labels then we have a challenge.
    ↳ The radiology report contains many sections including clinical history, the discription of how the exam was done, the finding section, which cover what the radiologist saw in each part of the body.

- We will focus on extracting labels from a radiology report in a two-step process.




# Example:

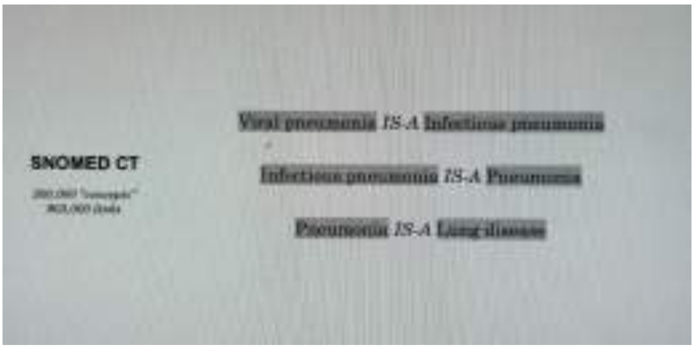



However, this won't be able to catch the other words which may have the same meaning. e.g infection can be synonymous with pneumonia. So instead of searching for pneumonia, we also search for the synonym words.

- We can ask a radiologist to get all the ways we can say lesion.

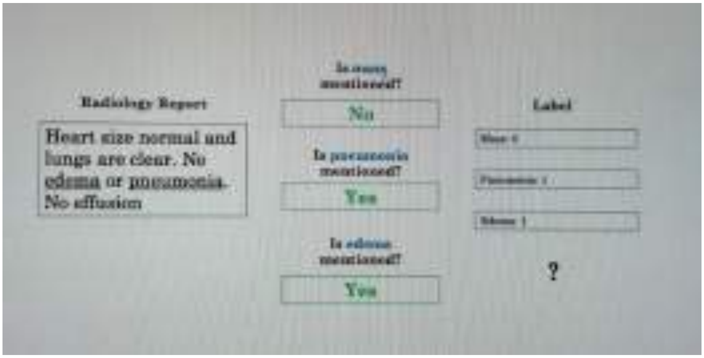
- Another way is to use a terminology, also called a thesaurus or a vocabulary. Example of such terminology is SNOMED CT.


Is-a relationship Terminologies not only contain synonyms for our concept but also contain relationships to other concepts.

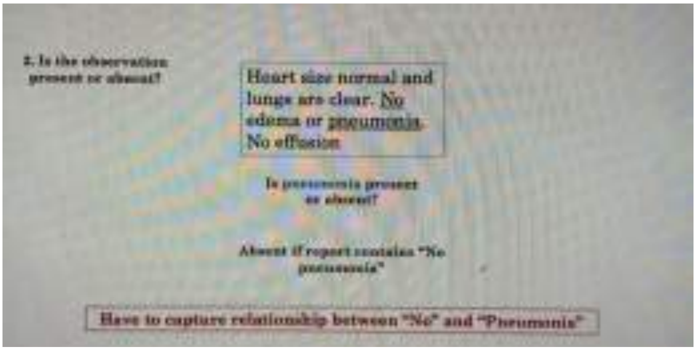



Therefore we can catch mention of lung disease by not only searching for synonyms of lung disease in SNOMED CT, but also its subtypes and their synonyms.
The advantages of this approach is that we don't need any data for supervised learning. The disadvantage of this approach is that there is a lot of manual work to refine these rules based on what is working and what is not working.

- Now we come to the second step to check "Is the observation present or not?




- Edema and pneumonia are present but the label "1" is incorrect as we have not taken into account the "No".




Not only we have to look for "No pneumonia" but we have to capture the relationship.

We can do this by introduction of new rules such as

Is edema present or absent?

Absent if report contains

"No edema"

"No XXX or edema"

"without XXX edema"

A few of the most common methods used to find patterns on text strings;

- Regex rules.
- Dependency Parse rules.
- Negation Classification.

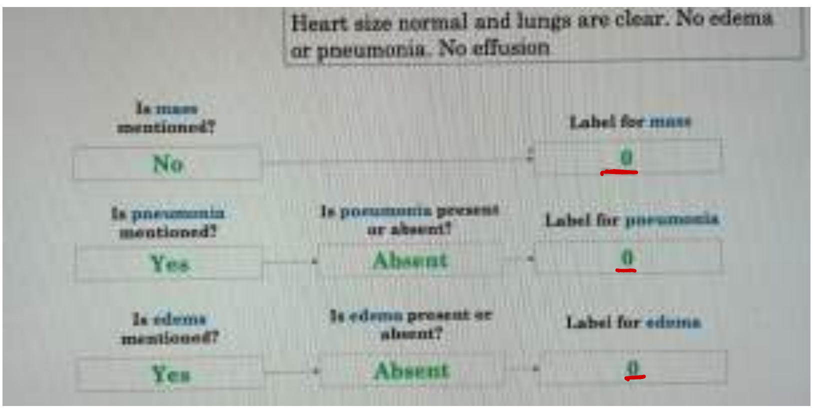
} Rules based approach.

For Rules based approach:

+ No labeled data needed for supervised learning.

- Manually time consuming & required expertise

We are now able to determine whether they are present or absent.



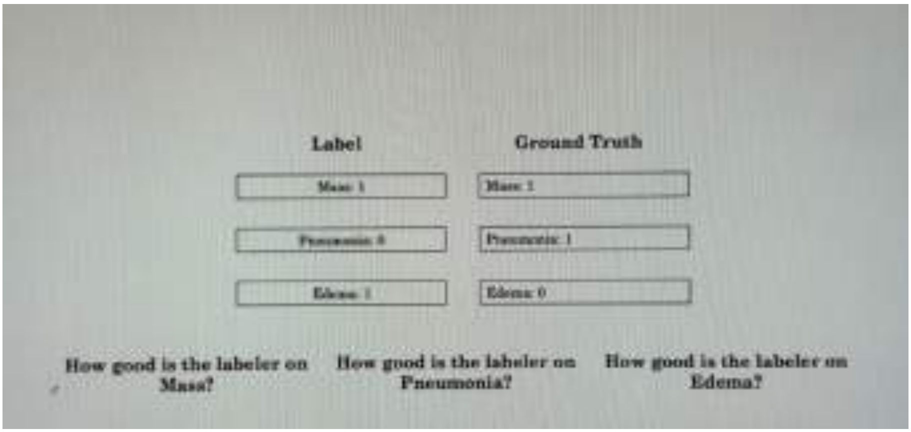

Evaluate the quality of label extraction:

- To make this evaluation, we need to know what the ground truth for the label is.

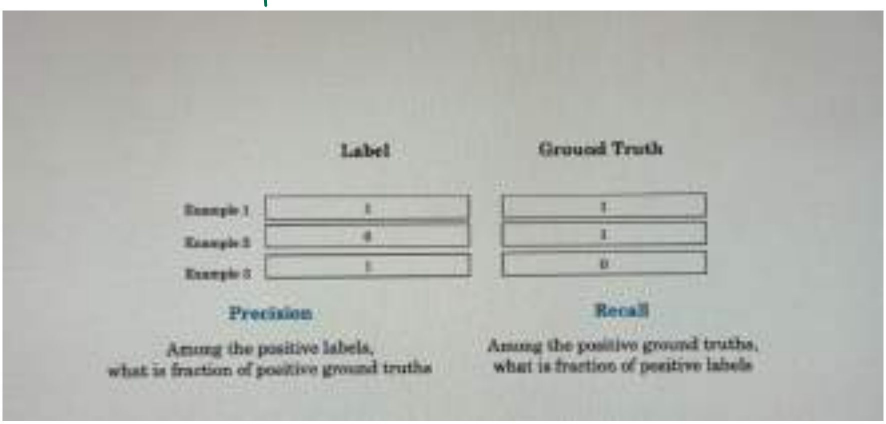
↳ We can get the ground truth using a group of radiologists to look at the report & then annotate the presence or absence of each of the diseases.

↳ Another option can be to ask the radiologist to look at the image, and annotate the presence or the absence.

## For Example:




Let's look at the pneumonia;



$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{RECALL} = \frac{TP}{TP + FN}$$

LABEL

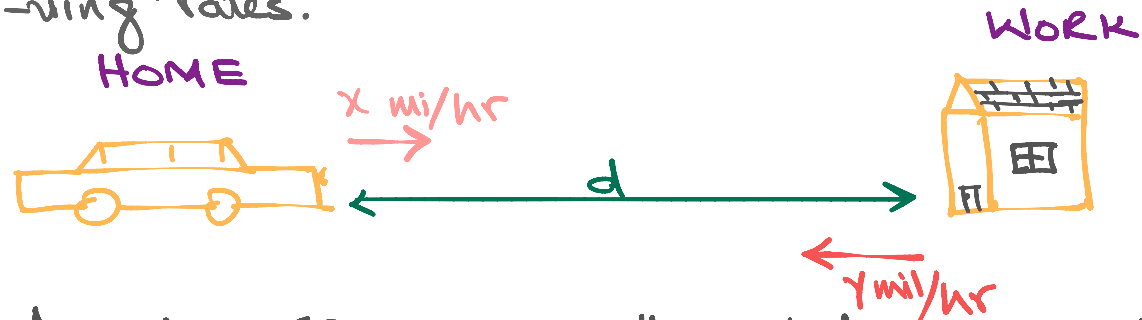
| G.T | 1 | 0 |
|---|---|---|
| 1 | 1 TP | 1 FN |
| 0 | 1 FP | 0 TN |

$$\text{Precision} = \frac{1}{2} = \underline{0.5} \qquad \text{RECALL} = \frac{1}{2} = \underline{0.5}$$

* Precision and Recall trade off each other.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{recall}}{\text{precision} + \text{recall}} = \underline{0.5}$$

THE harmonic mean (F1) is a different type of average often appropriate in situations involving rates.




AVERAGE SPEED over the whole journey?

$$\text{avg spd} = \frac{\text{dist}}{\text{time}}$$

$$= \frac{2d}{\frac{d}{x} + \frac{d}{y}}$$

$$= \frac{2d}{\frac{dy + dx}{xy}}$$

$$= \frac{2\cancel{d}\,xy}{\cancel{d}(x+y)}$$

$$= \boxed{\frac{2xy}{x+y}}$$

The expression is same as our expression for the F1 score.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}}$$

They are also known as:

Precision : Positive Predictive value.
Recall : Sensitivity.
F1 : Dice Coefficient Score.

How we aggregate our results for multiple dise-
-ases?

GLOBAL Precision: Compute the precision for each class and then take average. (Option 1) It is called Macro -average; giving equal importance to each class. Another way is to treat all examples equally and compute global precision. (option 2). It is called micro average; giving equal importance to each example.

Micro average: if we are interested in giving highest Importance to the most prevalent categories.

Macro-average: if we want to work towards a system which works pretty well across all classes.