

INFORMATION RETRIEVAL

Week 10 – Evaluation

Today

1

Exercise Recap

- Champion Lists

2

Theory

- Evaluation
- Probabilistic Retrieval

3

Kahoot

Exercise 9: Evaluation

Exercise 8: Champion Lists

Q1

True	False	
<input checked="" type="radio"/>	<input type="radio"/>	The idea of <i>champion lists</i> is to precompute, for each term t in the dictionary, the set of the r (r is fixed in advance) documents with the highest weights for t .
<input type="radio"/>	<input type="radio"/>	For tf-idf weighting, the champion list for term t would be the r documents with the lowest tf values for term t .
<input checked="" type="radio"/>	<input type="radio"/>	Tiered indices can be viewed as a generalization of champion lists.
<input type="radio"/>	<input type="radio"/>	With using tiered indices, if we fail to get K results from tier 1, query processing “falls back” to tier 2, and so on

Exercise 8: Champion Lists

Q1

1. Definition of Champion Lists
2. No, it would be the highest tf values
3. True, we use tiered indices to prevent scarce returns
4. Definition of tiered indices

True	False	
<input checked="" type="radio"/>	<input type="radio"/>	The idea of <i>champion lists</i> is to precompute, for each term t in the dictionary, the set of the r (r is fixed in advance) documents with the highest weights for t .
<input type="radio"/>	<input checked="" type="radio"/>	For tf-idf weighting, the champion list for term t would be the r documents with the lowest tf values for term t .
<input checked="" type="radio"/>	<input type="radio"/>	Tiered indices can be viewed as a generalization of champion lists.
<input checked="" type="radio"/>	<input type="radio"/>	With using tiered indices, if we fail to get K results from tier 1, query processing “falls back” to tier 2, and so on

Exercise 8: Champion Lists

Q2

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the from the query vector to the of each , and then search for the within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a , however, it is to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the [] from the query vector to the [] of each [], and then search for the [] within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a [], however, it is [] to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance
2. leader

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the from the query vector to the of each , and then search for the within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a , however, it is to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance
2. leader
3. cluster

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the from the query vector to the of each , and then search for the within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a , however, it is to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance
2. leader
3. cluster
4. nearest document

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the [] from the query vector to the [] of each [], and then search for the [] within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a [], however, it is [] to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance
2. leader
3. cluster
4. nearest document
5. heuristic

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the from the query vector to the of each , and then search for the within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a , however, it is to give the correct answer.

distance

leader

cluster

nearest document

not guaranteed

heuristic

guaranteed

optimization

Exercise 8: Champion Lists

Q2

1. distance
2. leader
3. cluster
4. nearest document
5. heuristic
6. not guaranteed

Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of \sqrt{N} leaders at random, and partitioning all documents into clusters of approximately \sqrt{N} documents each. To process a query, we only compute the [] from the query vector to the [] of each [], and then search for the [] within that cluster. This is a heuristic for solving the nearest-neighbour problem. As a [], however, it is [] to give the correct answer.

distance

leader

cluster

nearest document

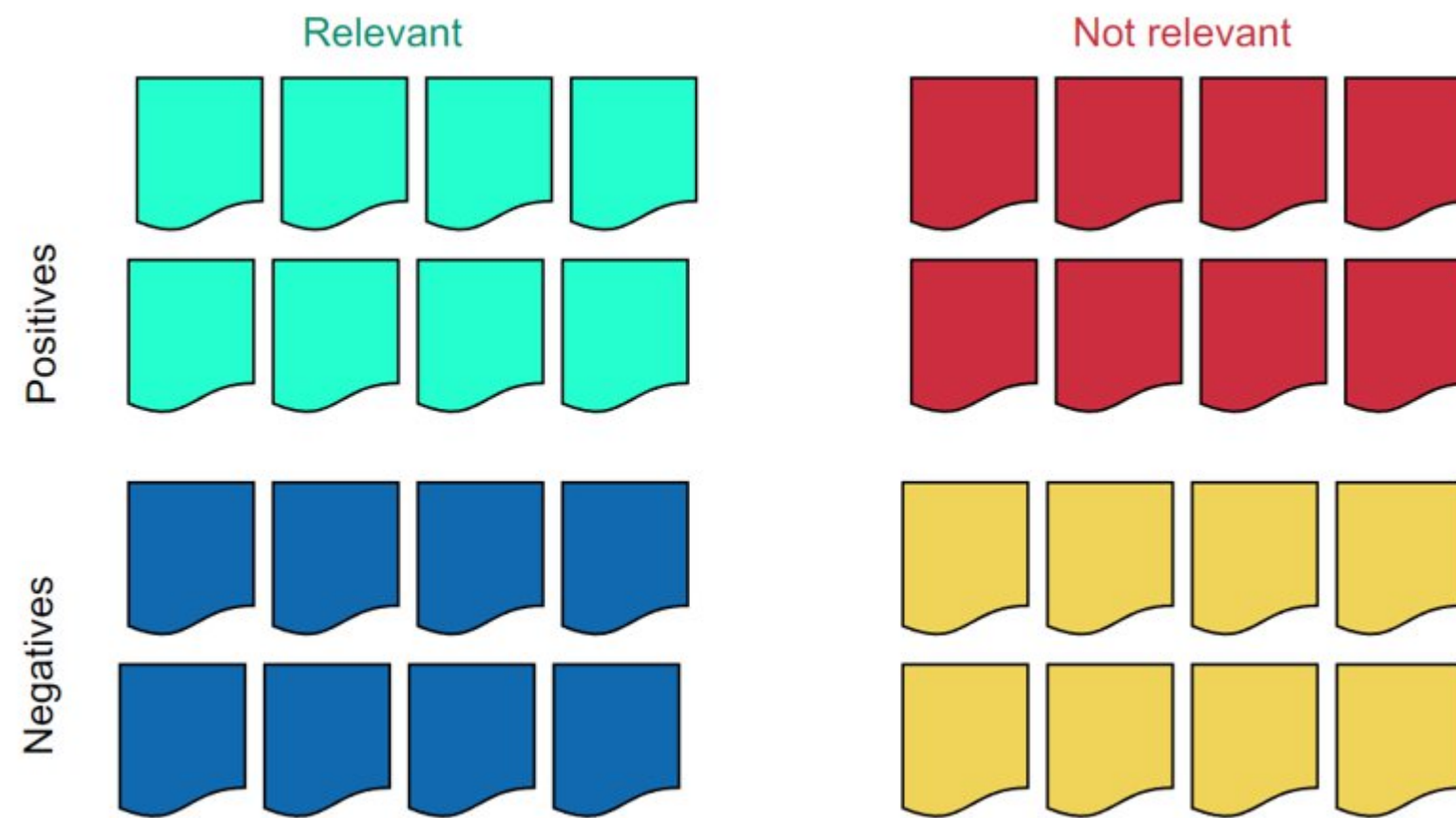
not guaranteed

heuristic

guaranteed

optimization

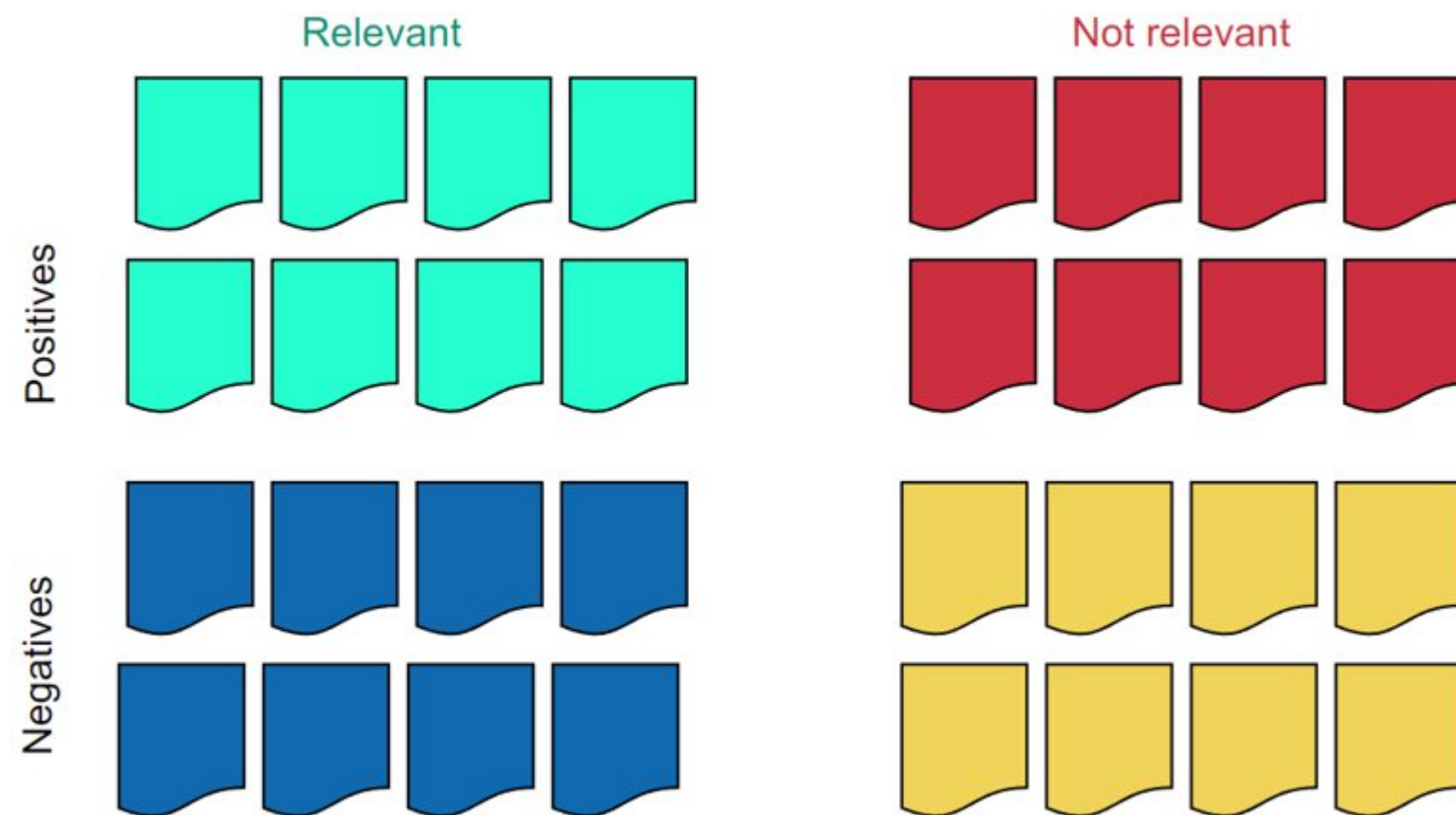
Recall and Precision



Precision = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$

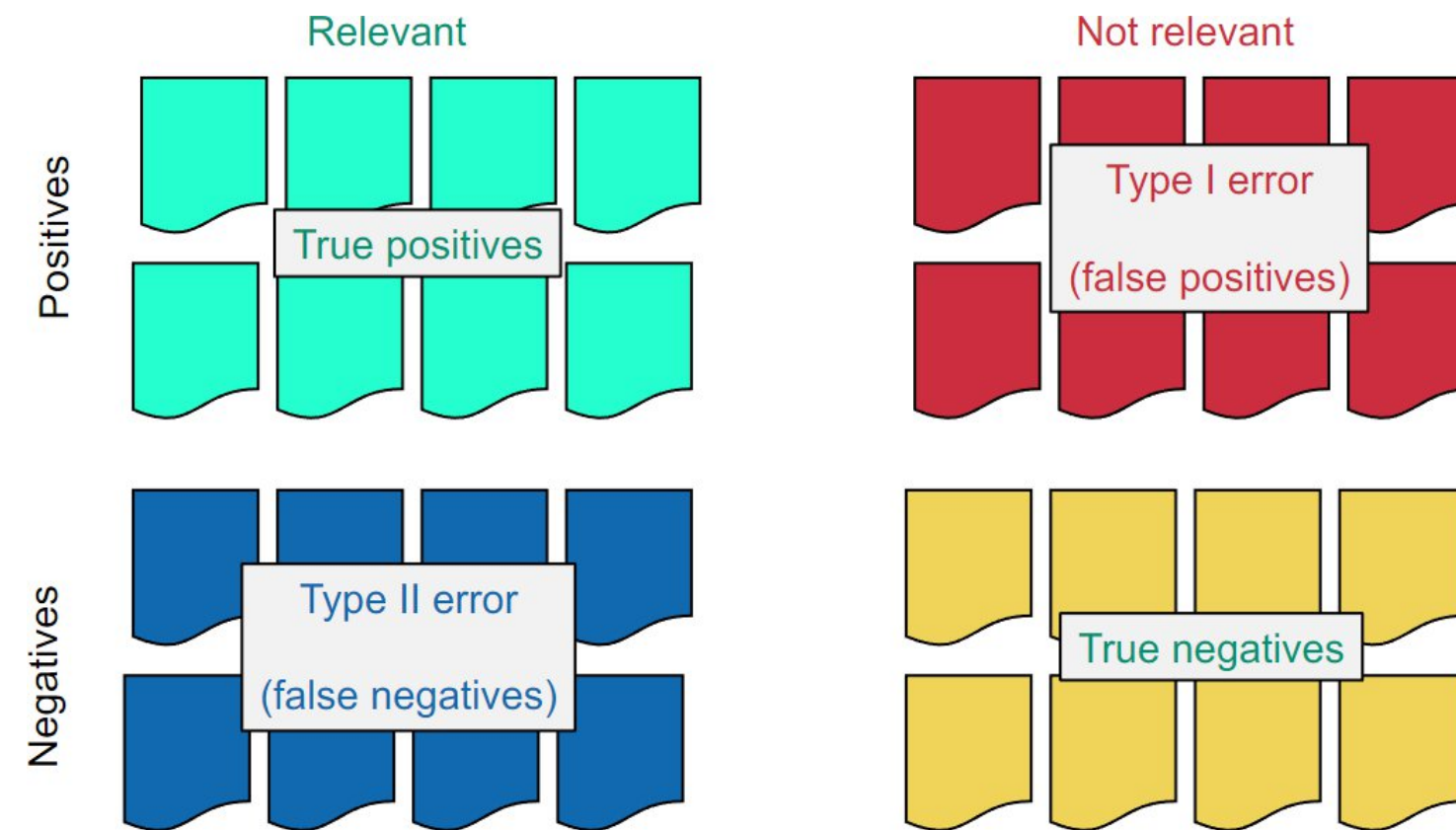
The diagram illustrates the formula for Precision. It shows a horizontal line with a cyan rectangle above it and a red rectangle below it. The text 'Precision =' is to the left of the line.

Recall and Precision



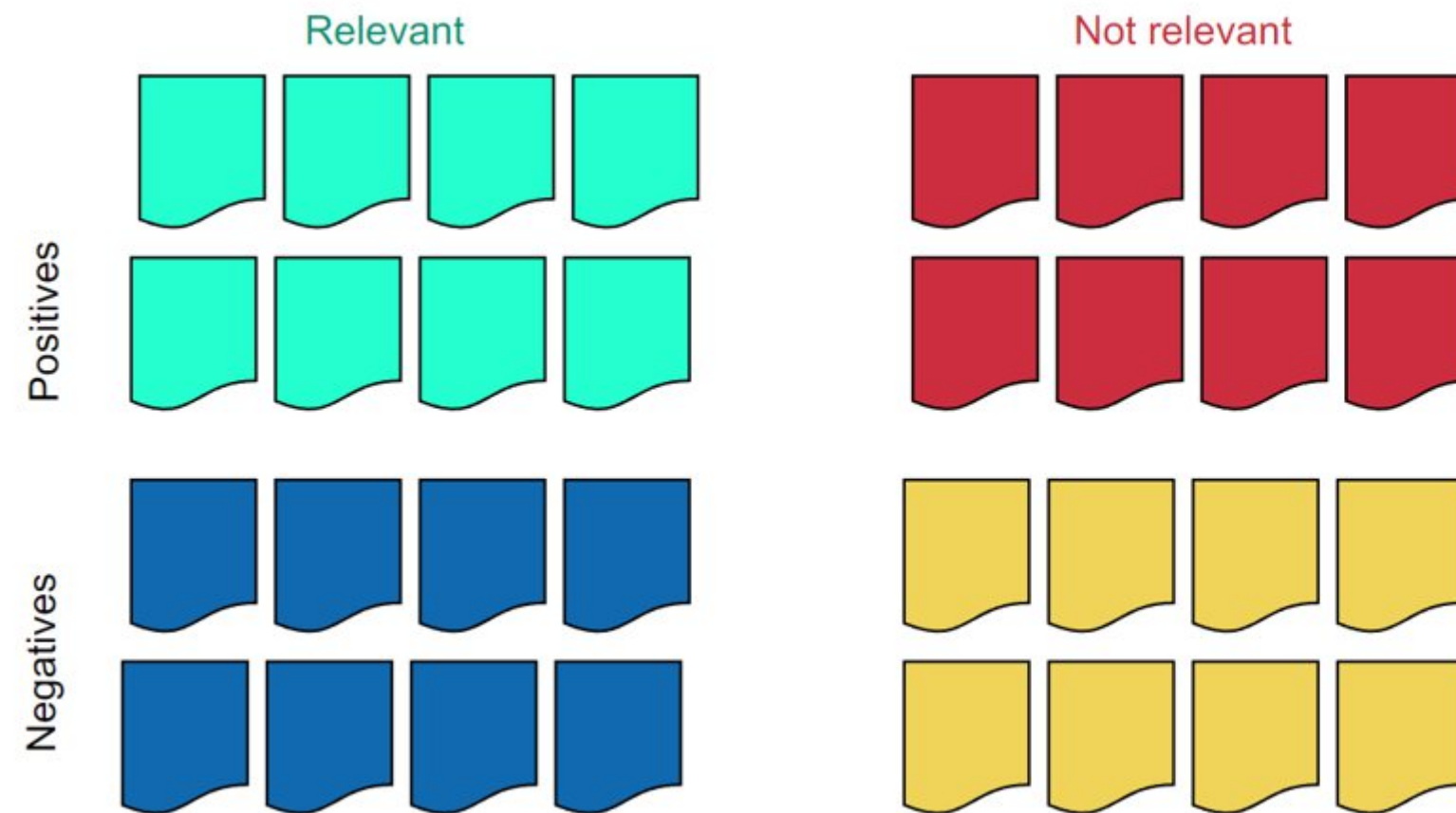
$$\text{Recall} = \frac{\text{1 cyan icon}}{\text{1 blue icon} + \text{1 cyan icon}}$$

Specificity and Accuracy



Specificity and Accuracy

hj

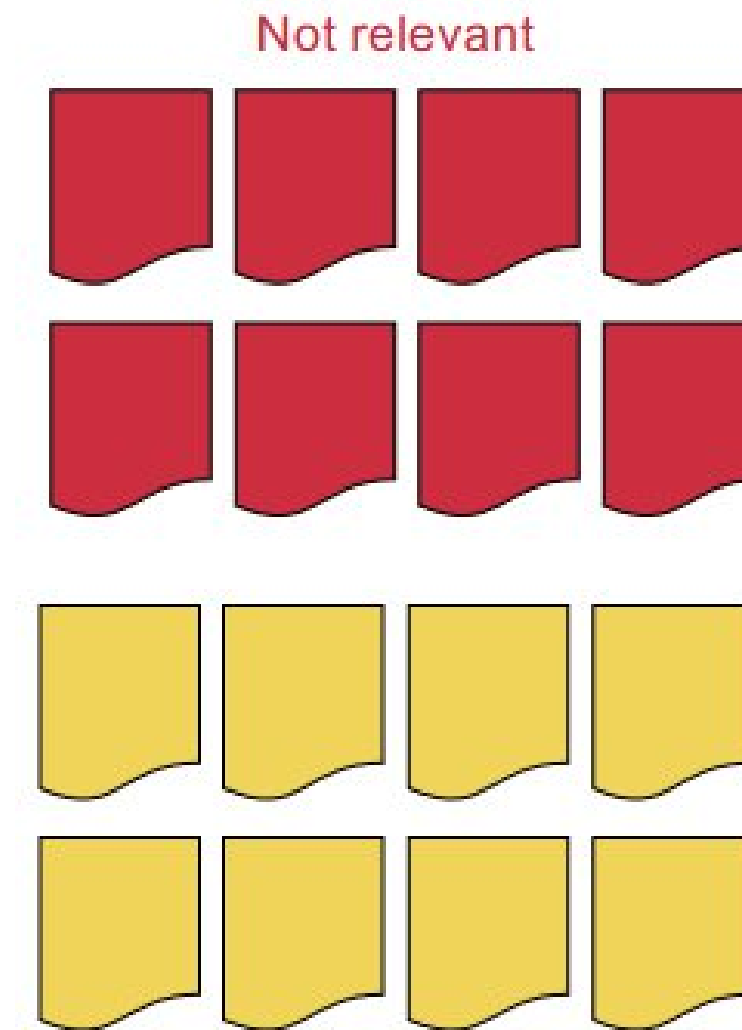


$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$$

The diagram illustrates the components of the specificity formula. The numerator is represented by a single yellow card (True Negatives). The denominator is represented by a red card (False Negatives) and a yellow card (True Negatives).

Specificity and Accuracy

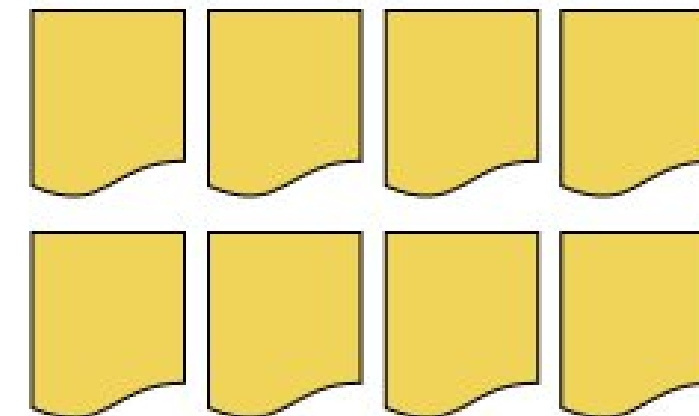
Specificity: 50%



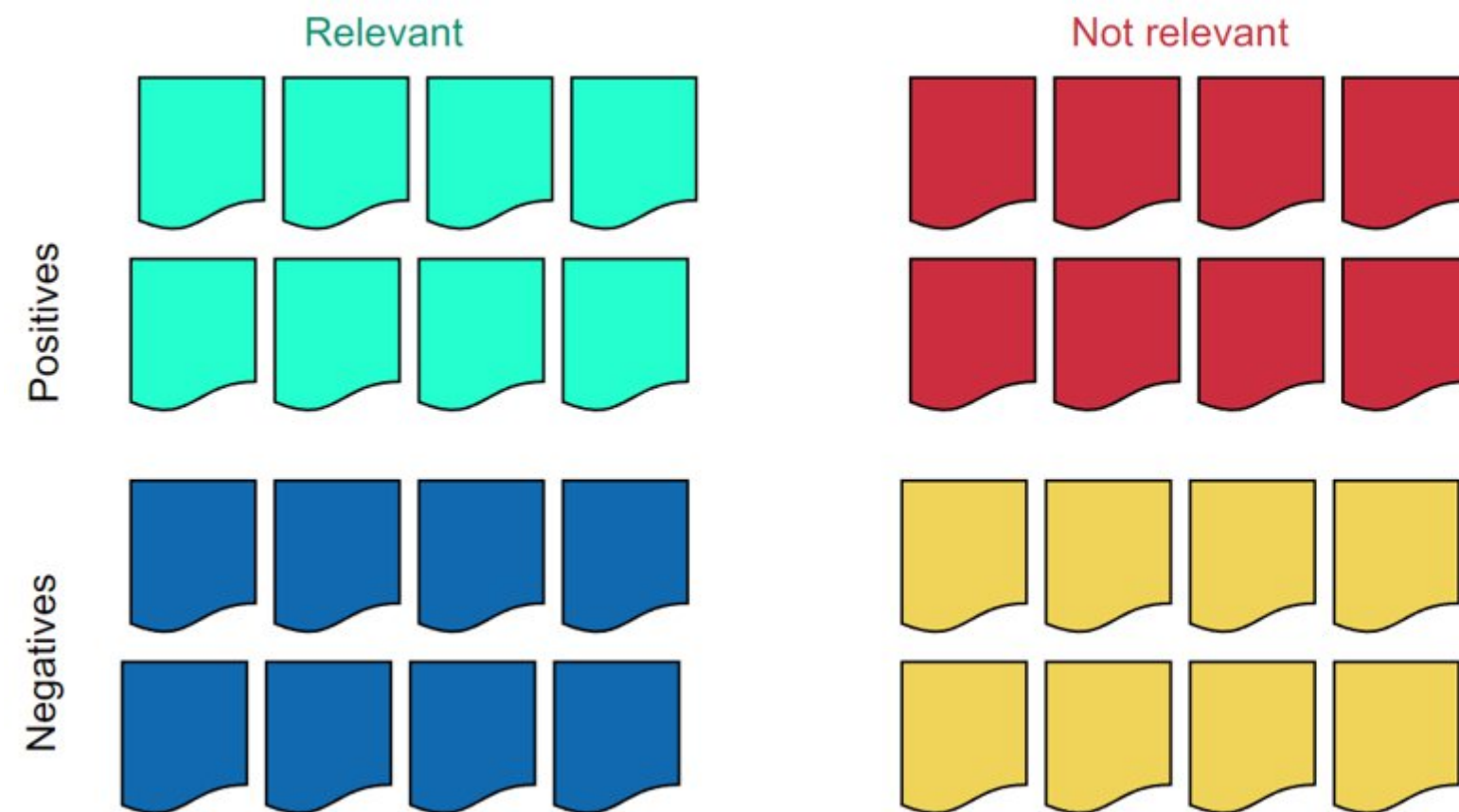
Specificity and Accuracy

Specificity: 100%

Not relevant

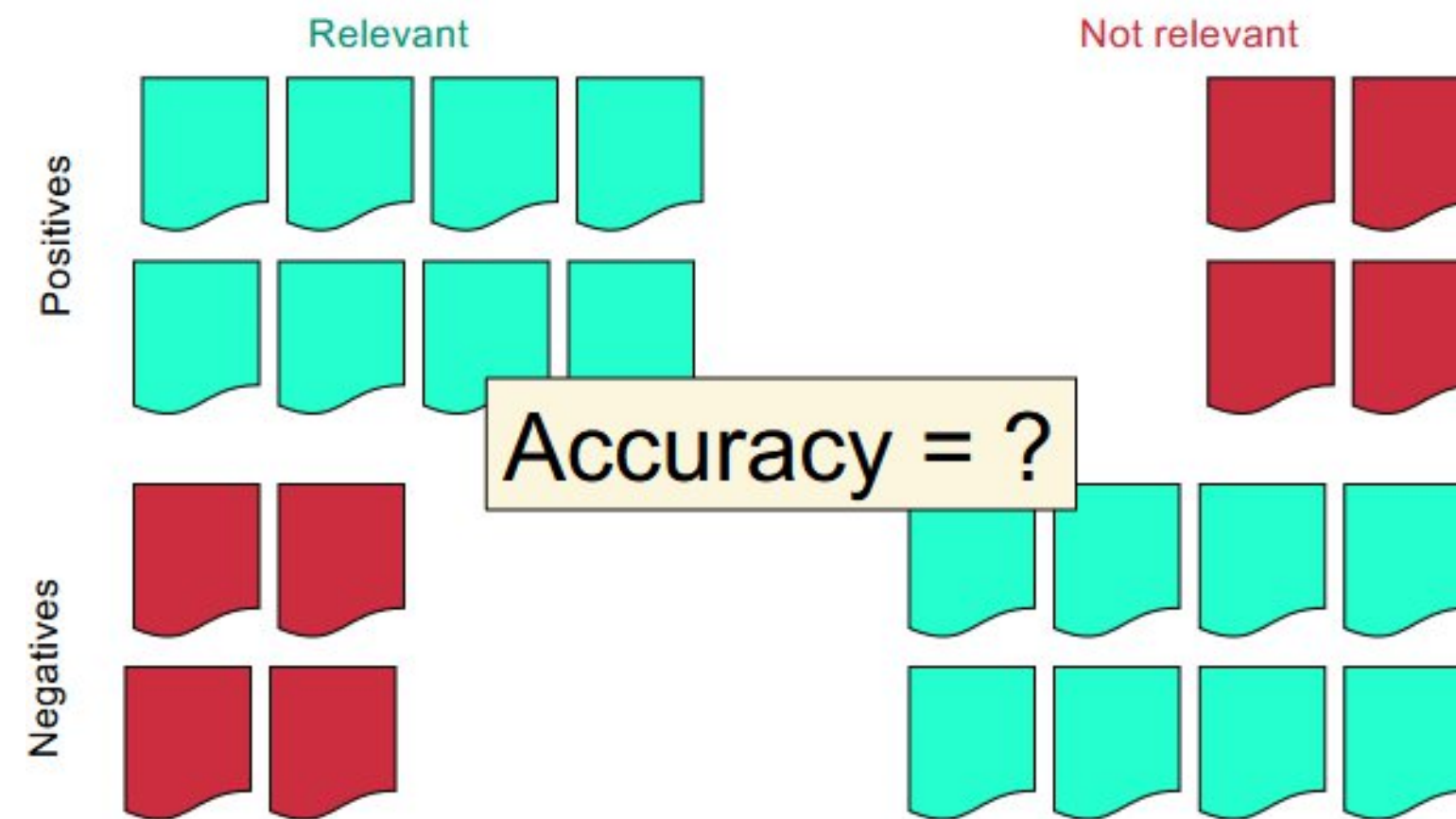


Specificity and Accuracy



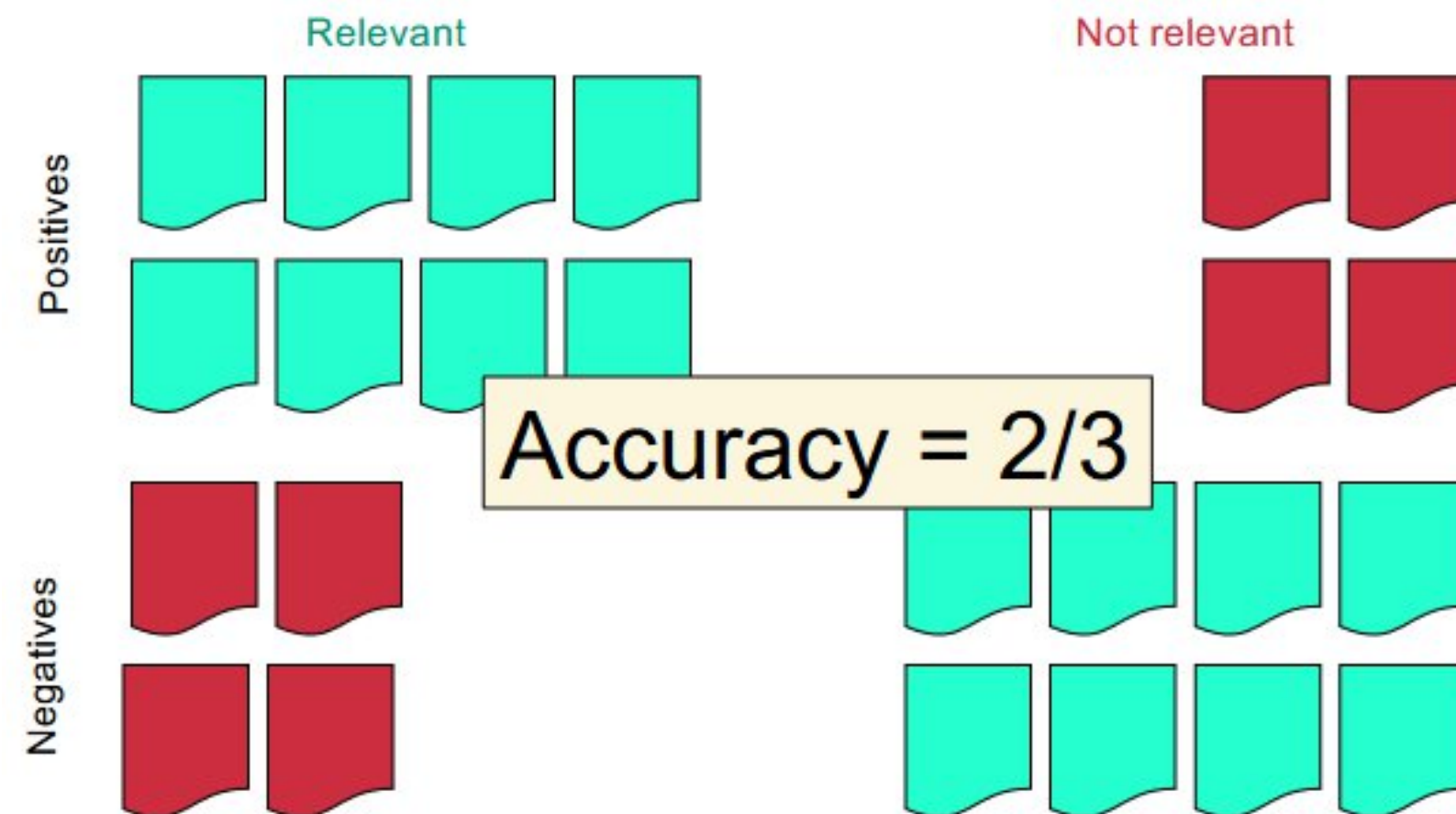
$$\text{Accuracy} = \frac{\text{Cyan} + \text{Yellow}}{\text{Red} + \text{Cyan} + \text{Yellow} + \text{Blue}}$$

Specificity and Accuracy



Specificity and Accuracy

Specificity: 100%



Defining all the terms

Recall: How good is the system at returning as many relevant results to you

Precision: How useful are the returned results

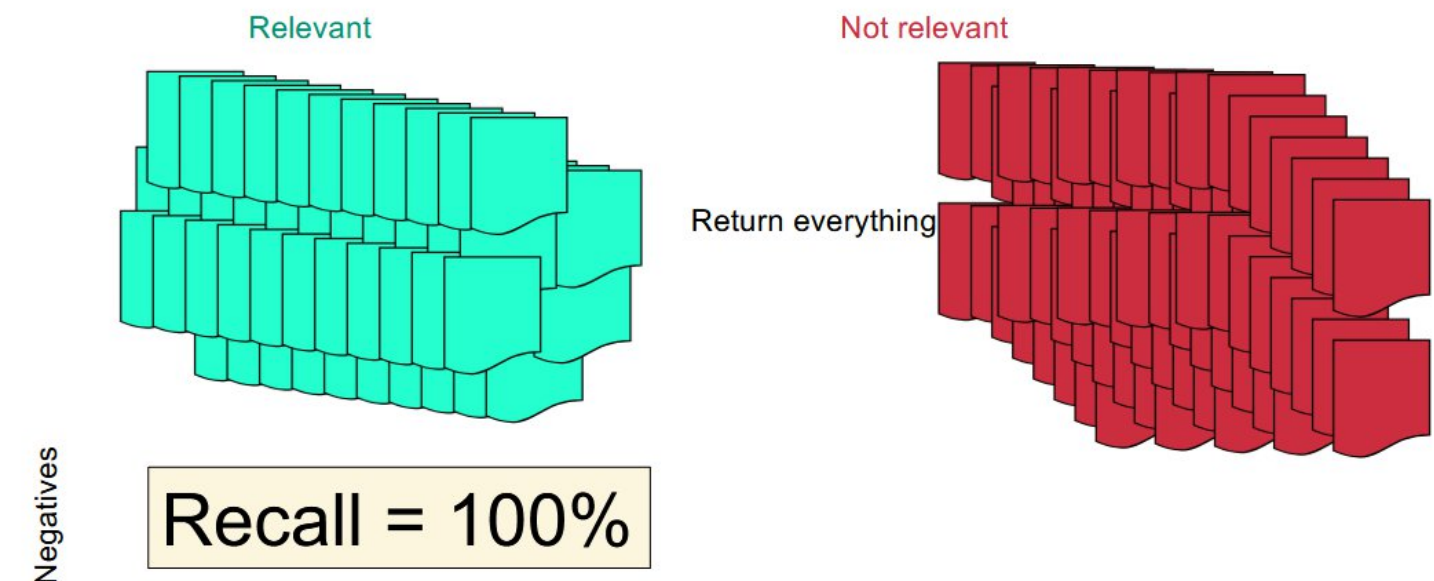
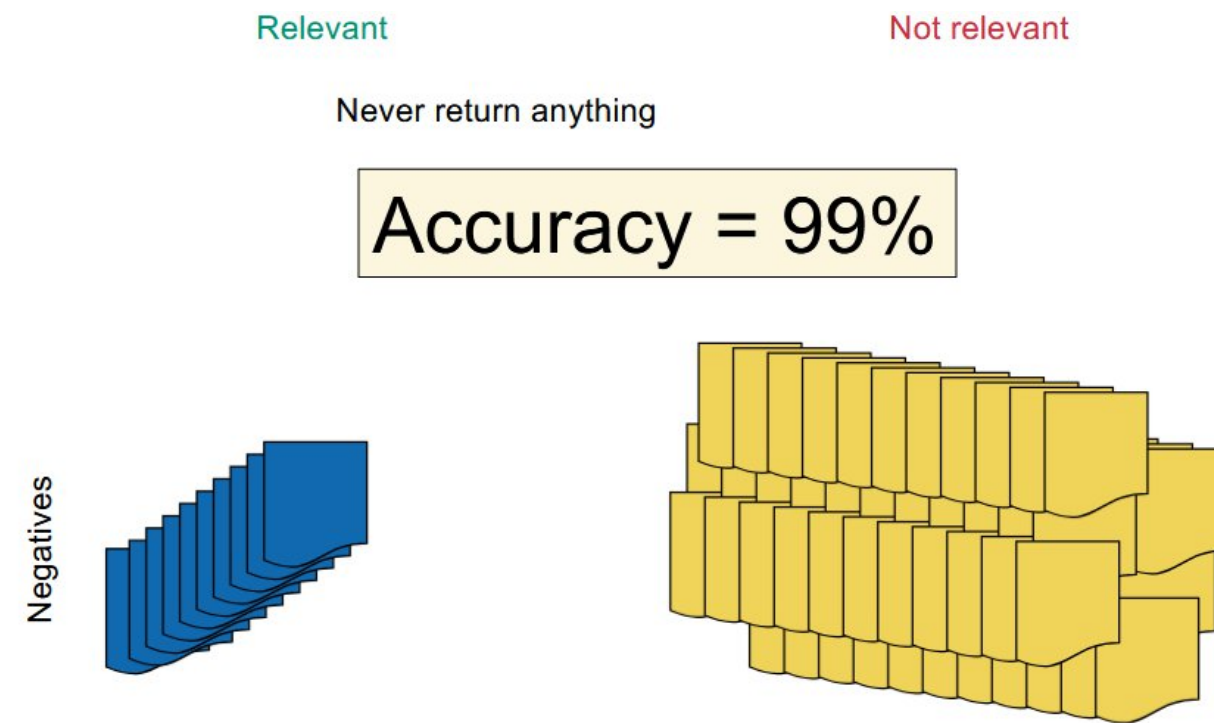
Specificity: How good is the system at not bothering you with useless stuff

Accuracy: How good is the system in total

Evaluation

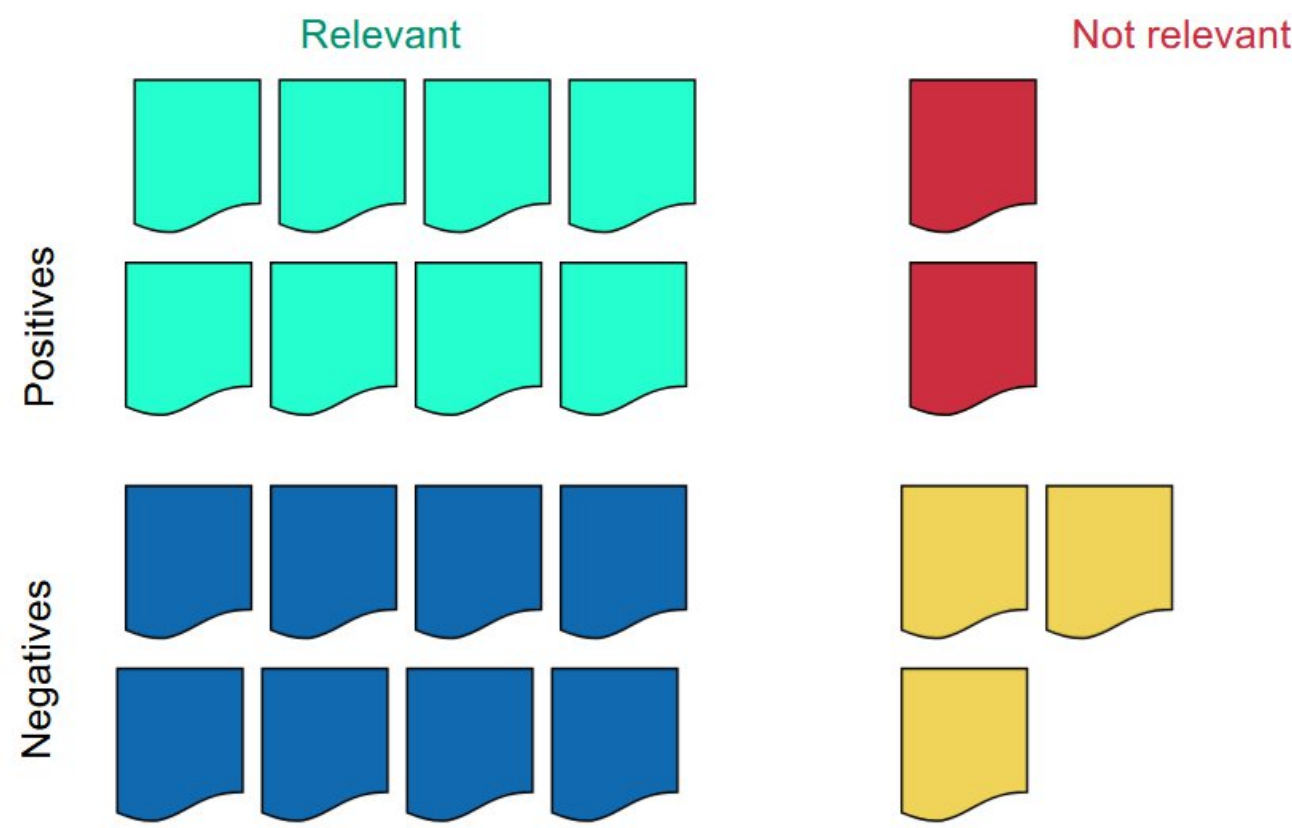
Hacking Recall and Accuracy

Issue: You can hack accuracy and recall by never returning anything or always returning everything respectively.

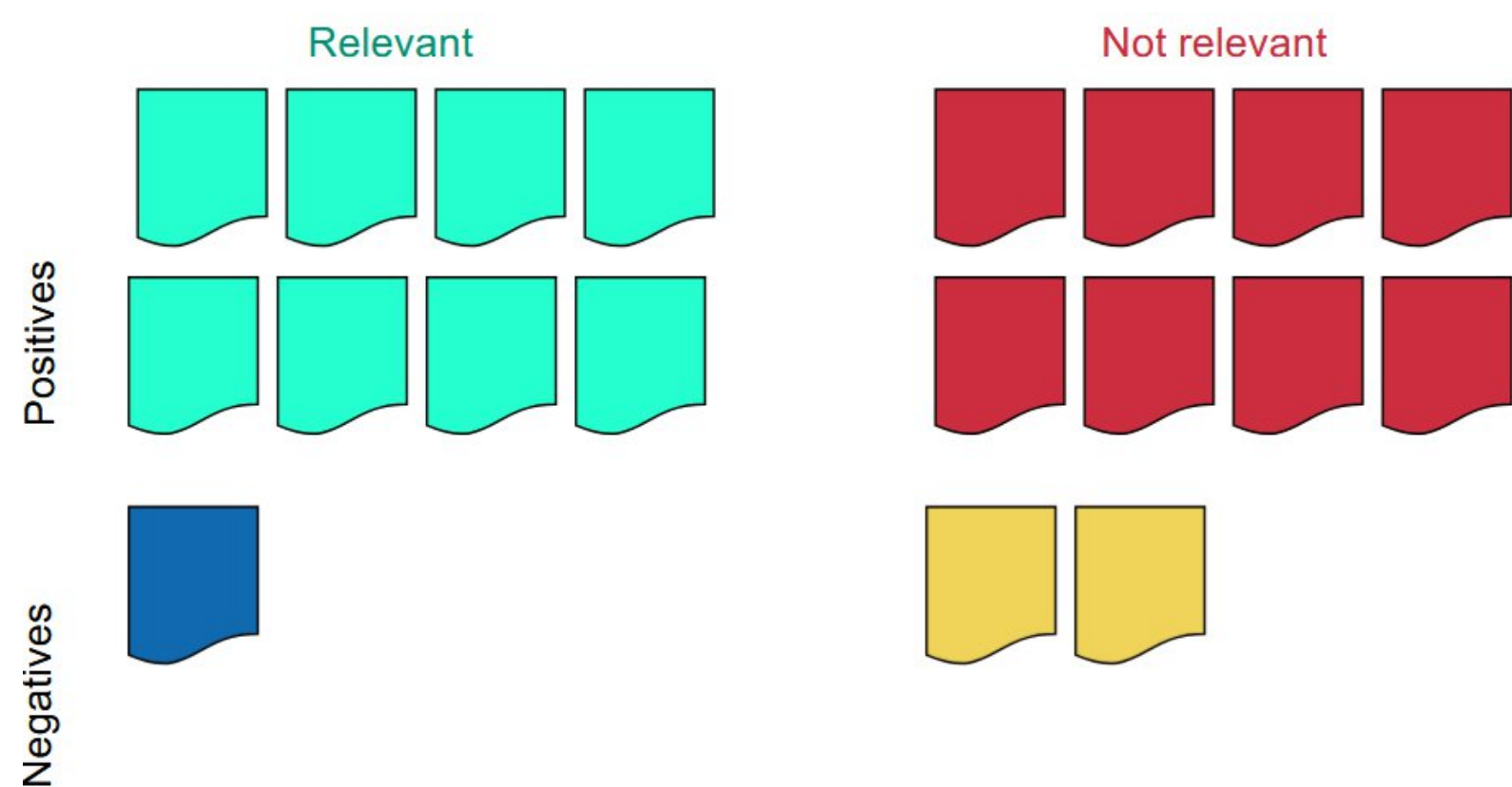


Compromise

High precision



High recall



Merge Precision and Recall: F-Measure

Use the mean to balance the trade-off on both sides.

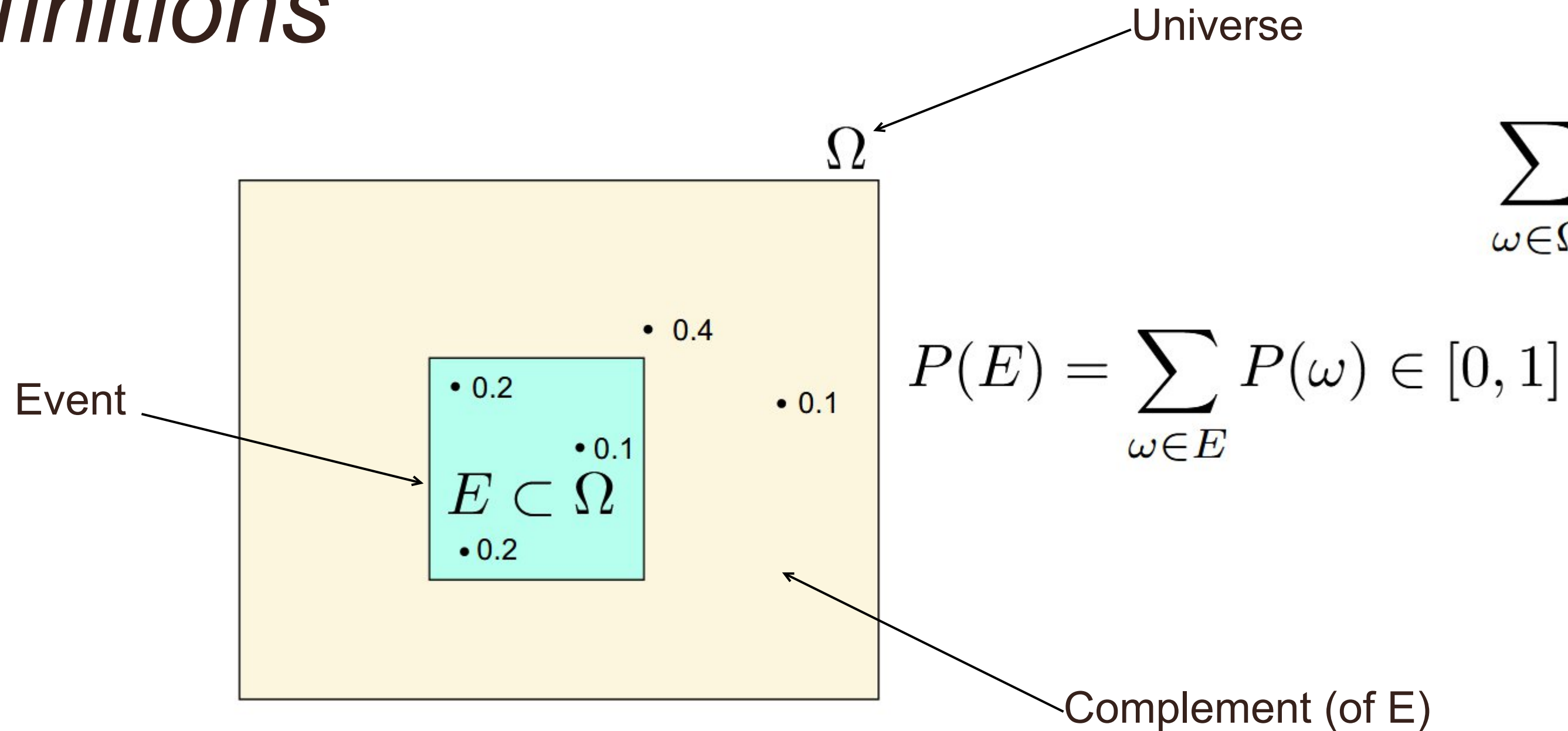
$$F_{\alpha} = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$$

The diagram illustrates the trade-off between Recall and Precision using the F-measure formula. A large blue double-headed arrow labeled "Weighting" connects two points on a horizontal axis. On the left, the point is labeled $\alpha = 0$ and "Recall-heavy". On the right, the point is labeled $\alpha = 1$ and "Precision-heavy". The F-measure formula is centered above the arrow.

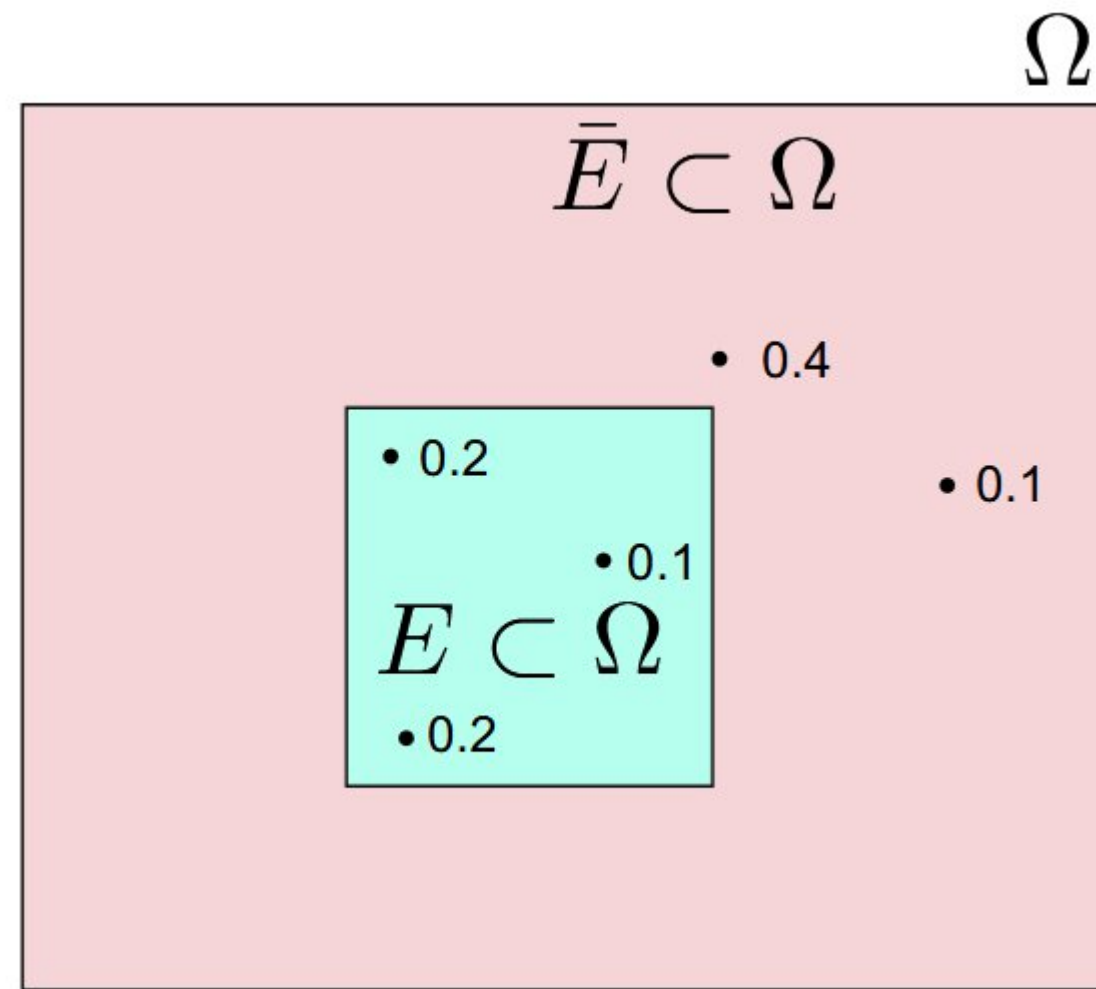
Definitions

$$P : \begin{array}{l} \Omega \rightarrow [0, 1] \\ \omega \mapsto P(\omega) \end{array}$$

$$\sum_{\omega \in \Omega} P(\omega) = 1$$

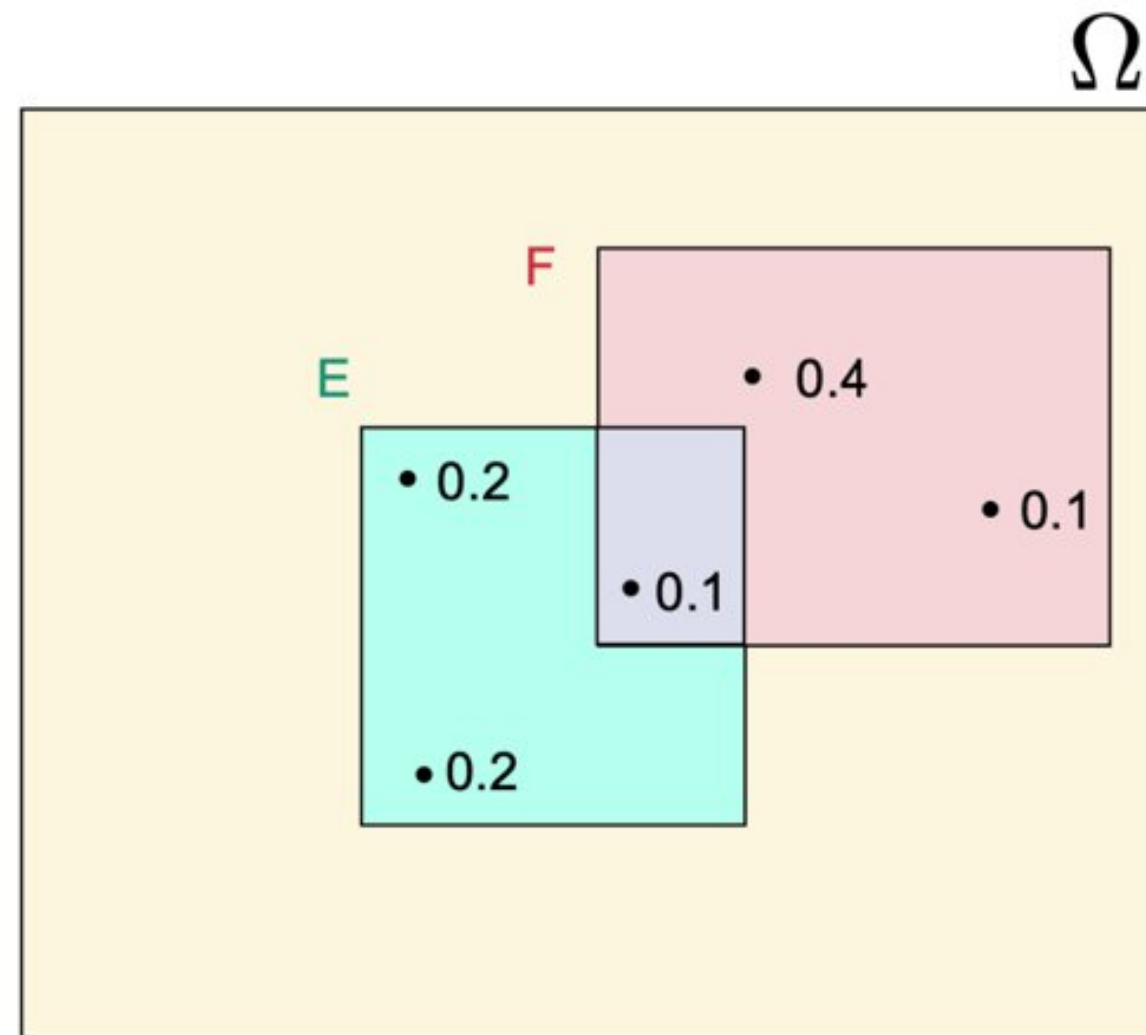


Odds



$$O_P(E) = \frac{P(E)}{P(\bar{E})}$$

Bayes' Rule



$$\underbrace{p(E|F)}_{\text{posterior}} = \frac{P(F|E)}{P(F)} \times \underbrace{P(E)}_{\text{prior}}$$

Notation

$$p_X(\blacksquare)$$

$$p_X(\blacklozenge)$$

$$p_X(\blacktriangle)$$

$$P(X = \blacksquare)$$

$$P(X = \blacklozenge)$$

$$P(X = \blacktriangle)$$

No go!

$$P(\blacksquare) = 0.5$$

No go!

Mystery Exercise

Mystery

- Will be uploaded to Moodle
- Entirely optional
- First 3 students to submit correct solution will win a prize

<https://create.kahoot.it/details/duplicate-of-information-retrieval-ex-07-vector-space-models-mschoeb/ef383953-b43a-4abd-af2a-d9ebf2ad1019>