INFORMATION RETRIEVAL

Week 10 – Evaluation

Severin Mills

Today

1

Exercise Recap

- Champion Lists

2

Theory

- Evaluation
- Probabilistic Retrieval

3

Kahoot

Exercise 9: Evaluation

Exercise 8: Champion Lists

Q1

True	False	
○	○	The idea of *champion lists* 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*.
○	○	For tf-idf weighting, the champion list for term t would be the *r* documents with the lowest tf values for term *t*.
○	○	Tiered indices can be viewed as a generalization of champion lists.
○	○	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

True	False	
●	○	The idea of *champion lists* 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*.
○	●	For tf-idf weighting, the champion list for term t would be the *r* documents with the lowest tf values for term *t*.
●	○	Tiered indices can be viewed as a generalization of champion lists.
●	○	With using tiered indices, if we fail to get K results from tier 1, query processing “falls back” to tier 2, and so on

1. Definition of Champi 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

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

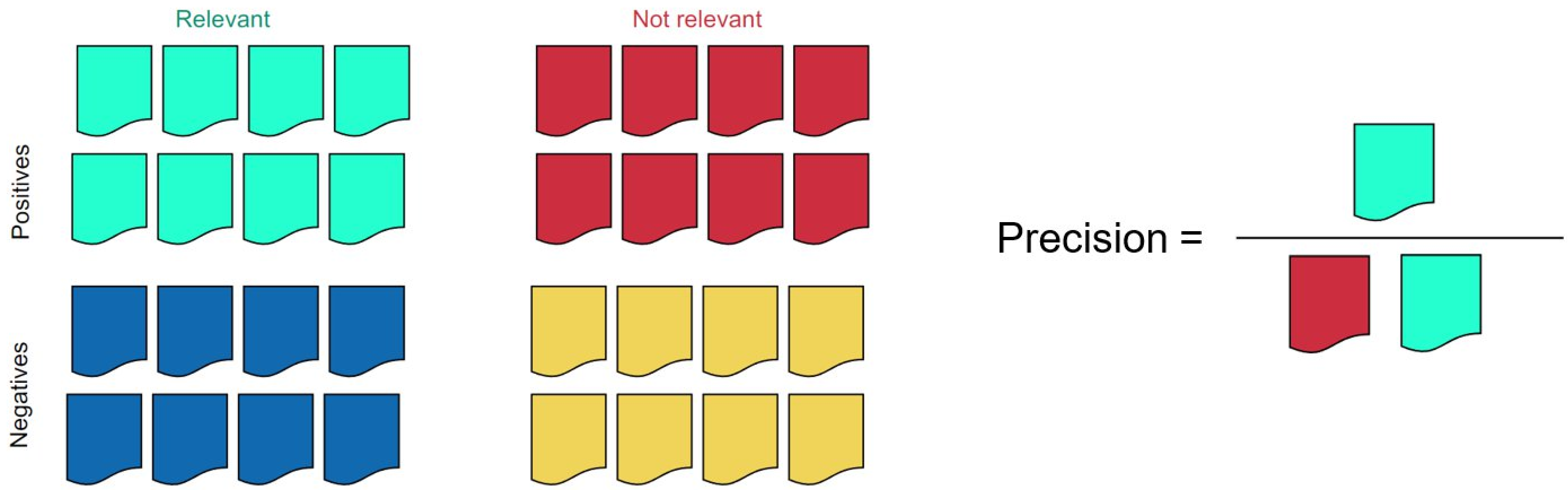
Q2

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

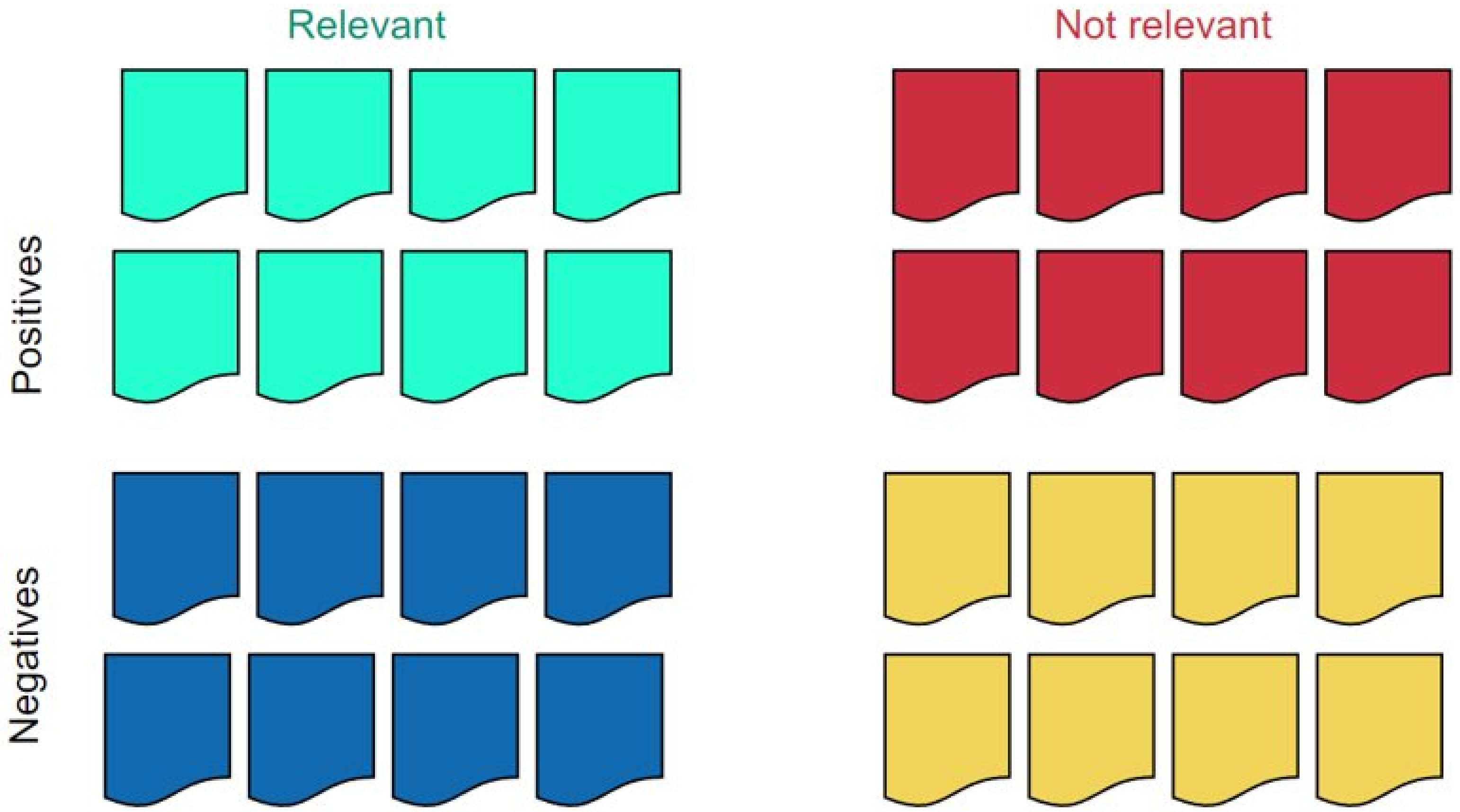
Calculating the cosine distance from a query **Q** to all documents **D** is an expensive operation. Cluster pruning attempts to reduce this cost by selecting a subset of $\sqrt{N}$ leaders at random, and partitioning all documents into clusters of approximately $\sqrt{N}$ documents each. To process 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

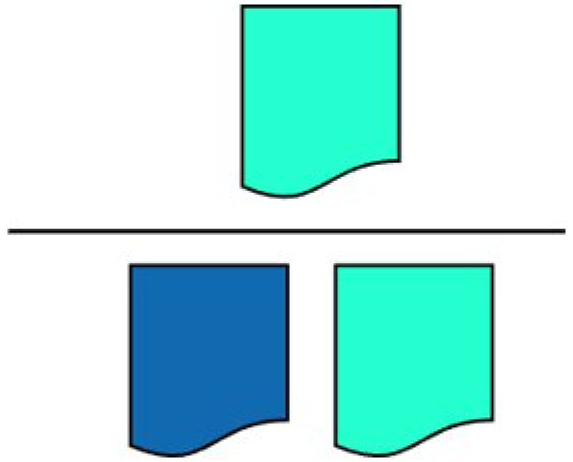
Recall and Precision



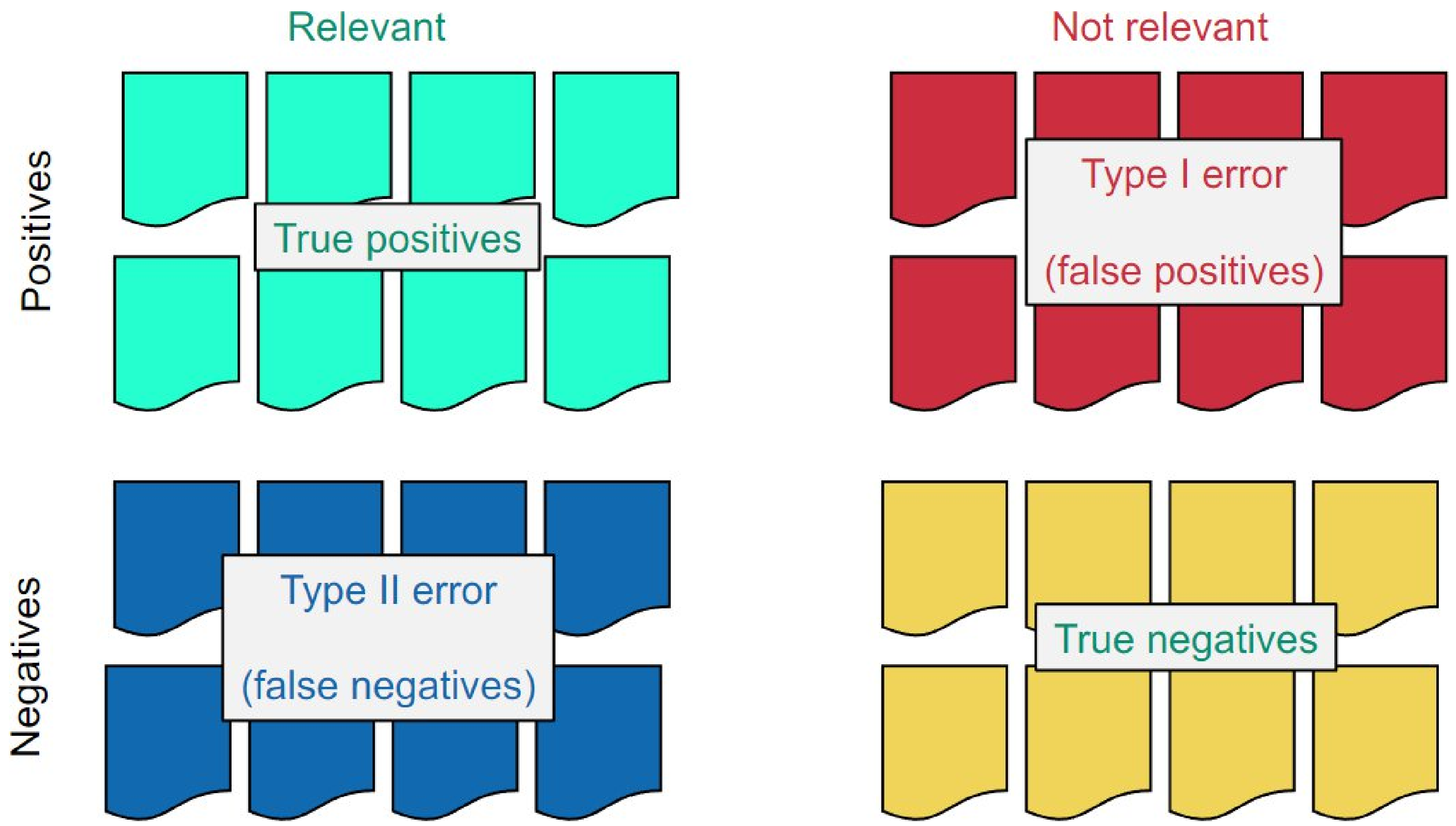
Recall and Precision



Recall =

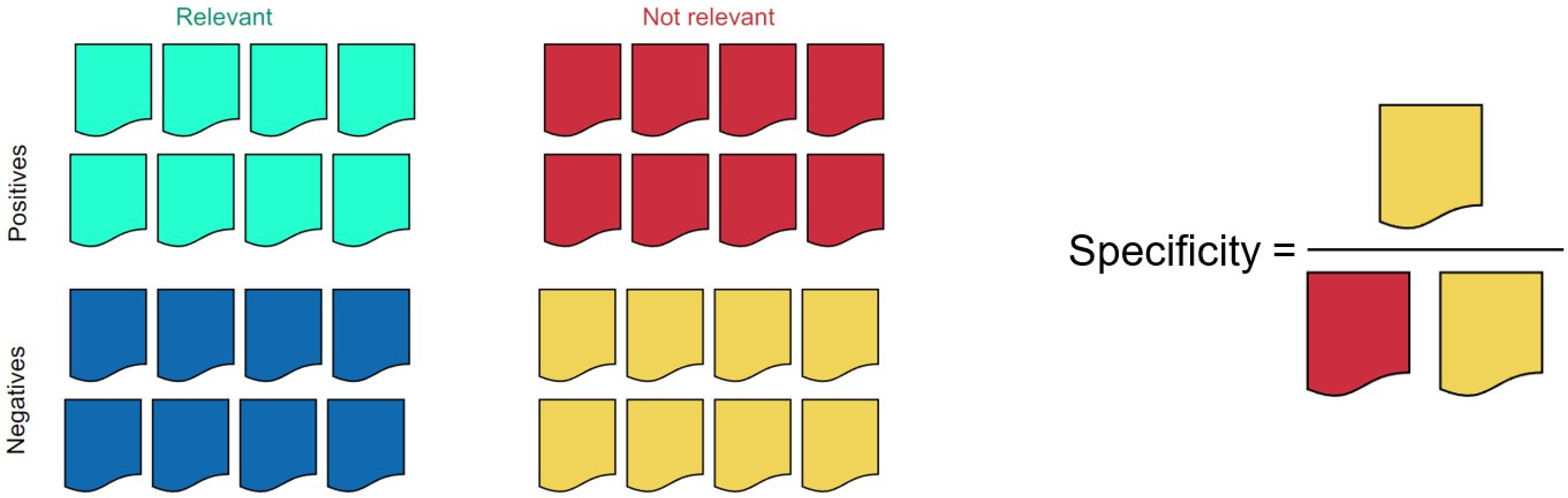


Specificity and Accuracy



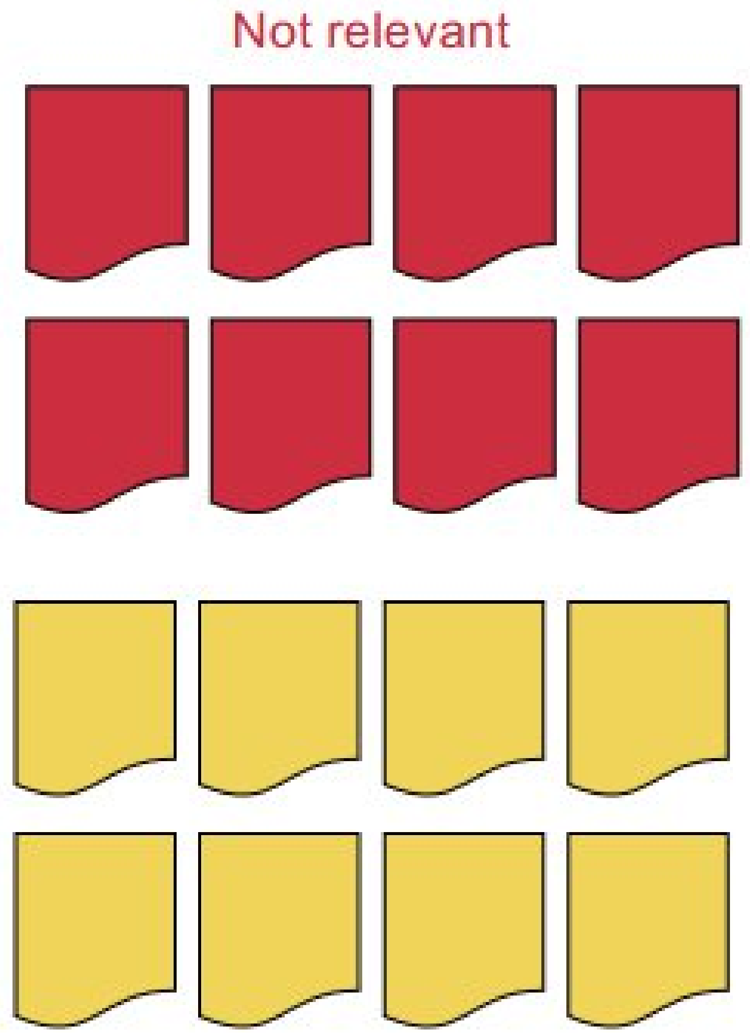
Specificity and Accuracy

hj



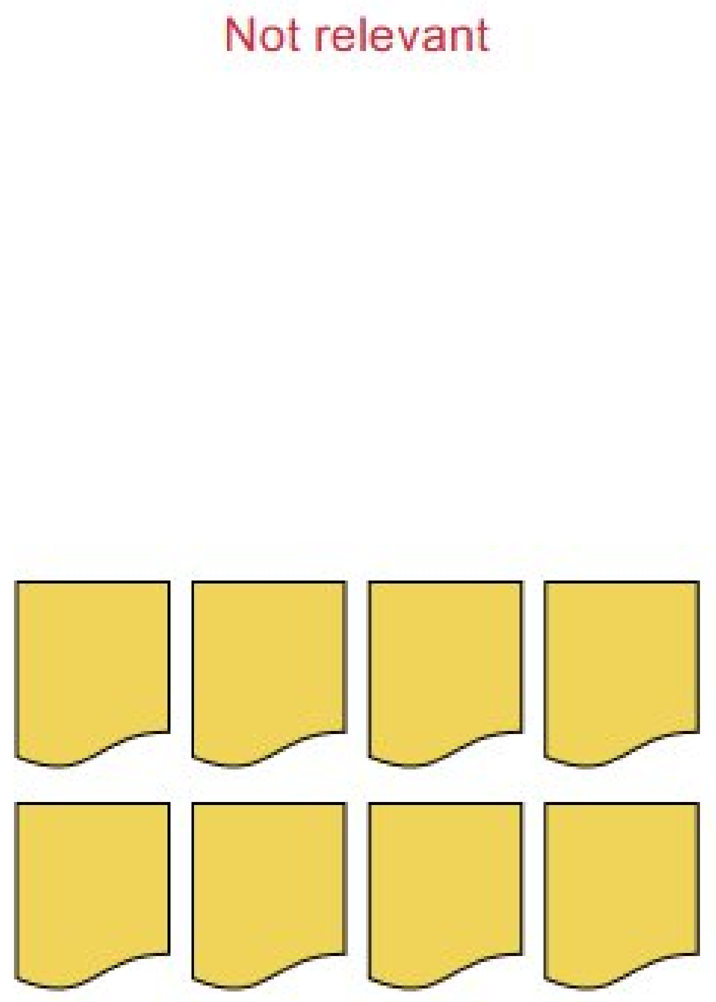
Specificity and Accuracy

Specificity: 50%

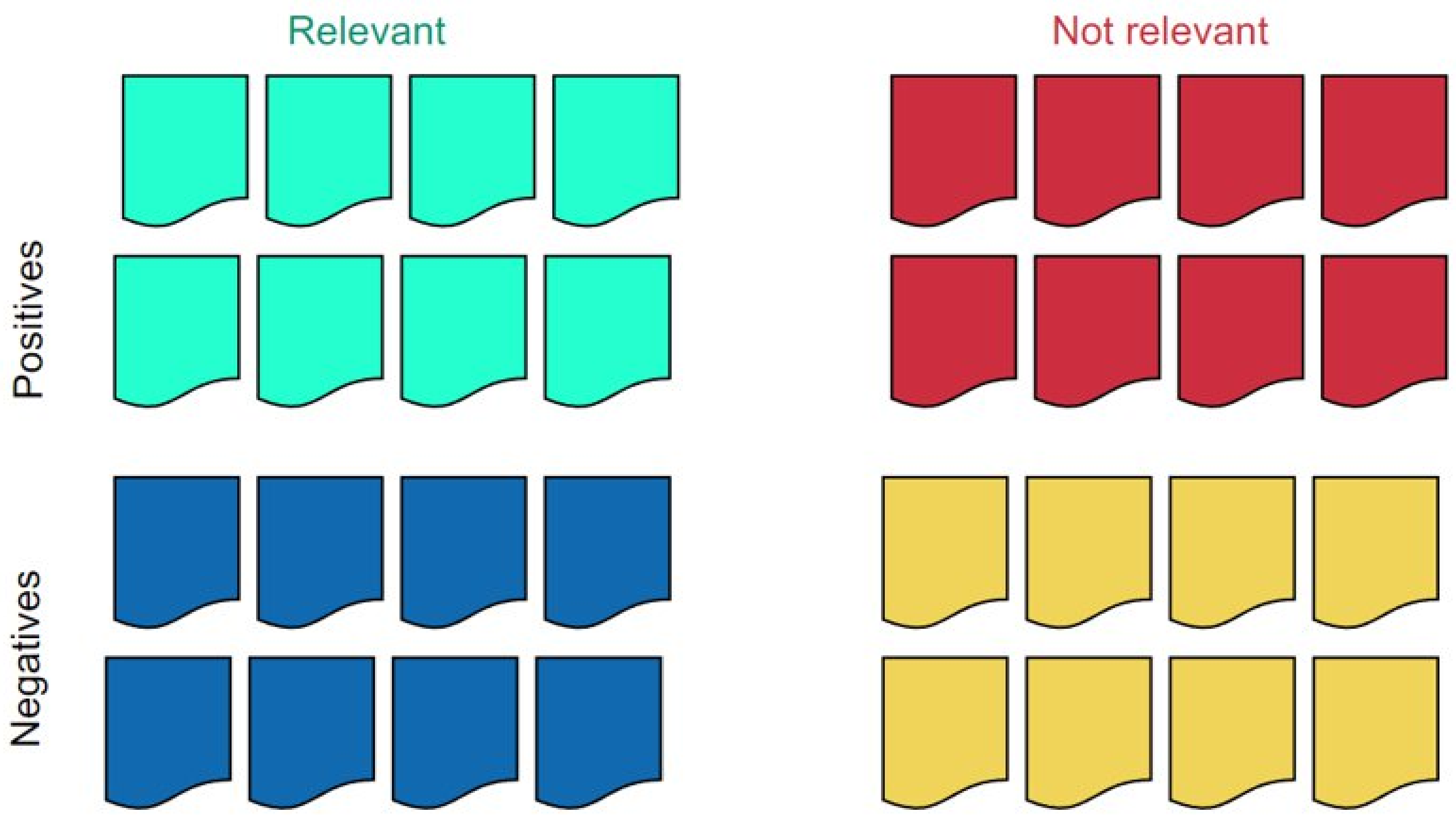


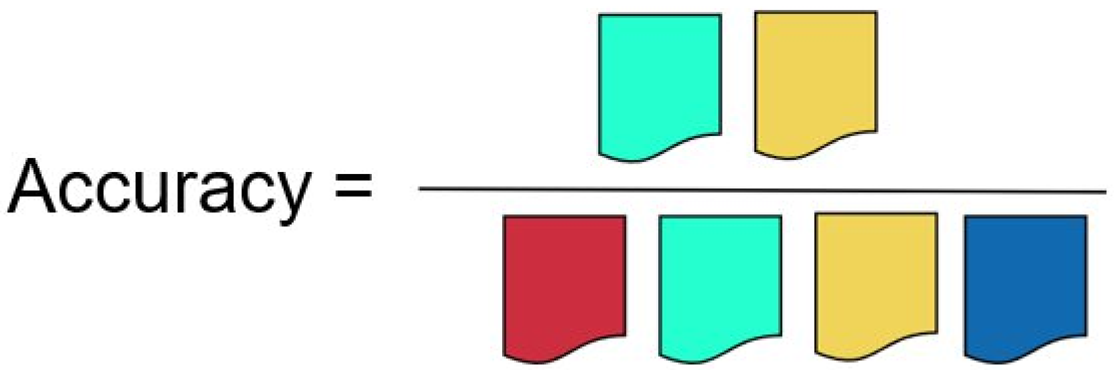
Specificity and Accuracy

Specificity: 100%

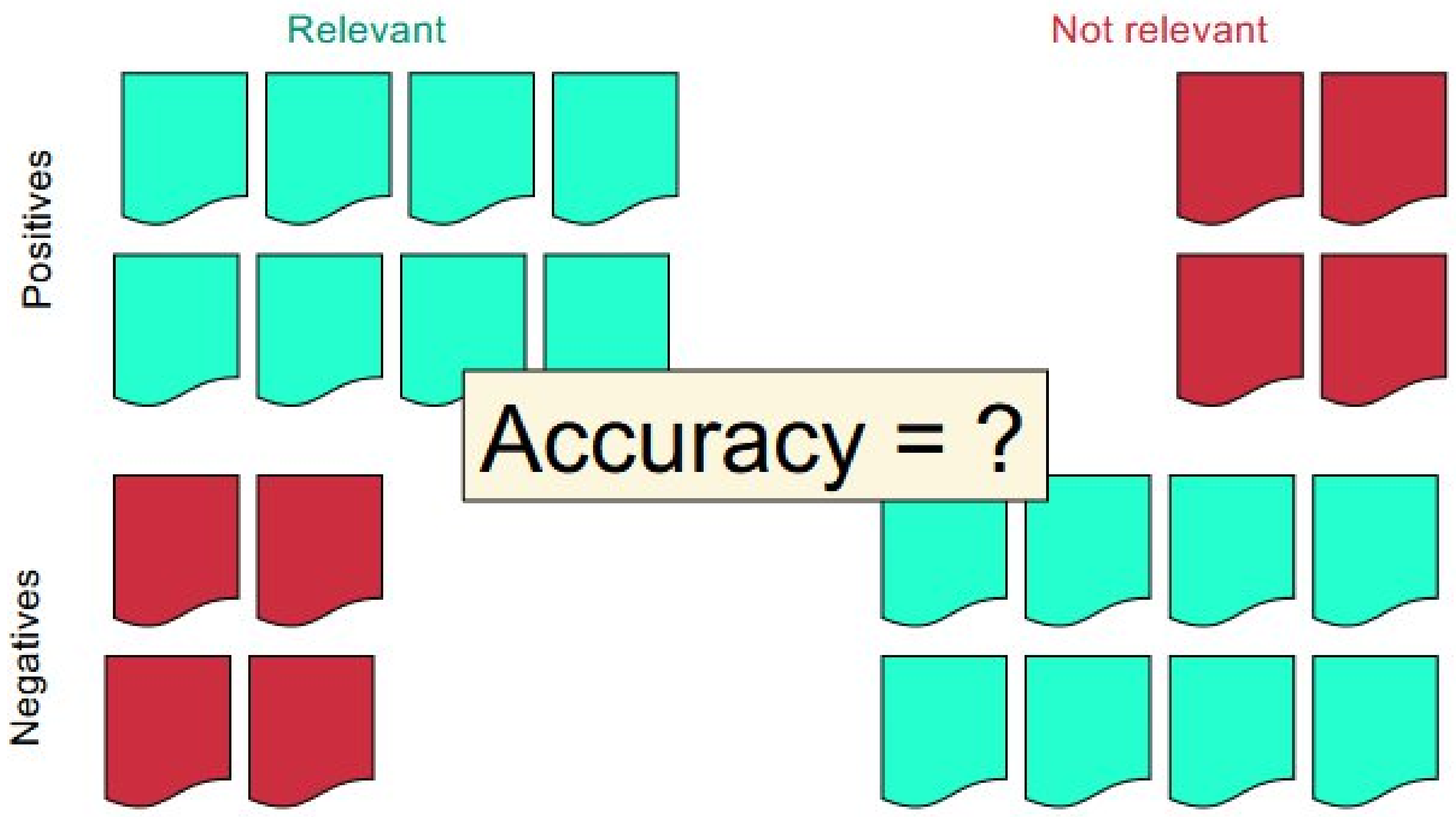


Specificity and Accuracy



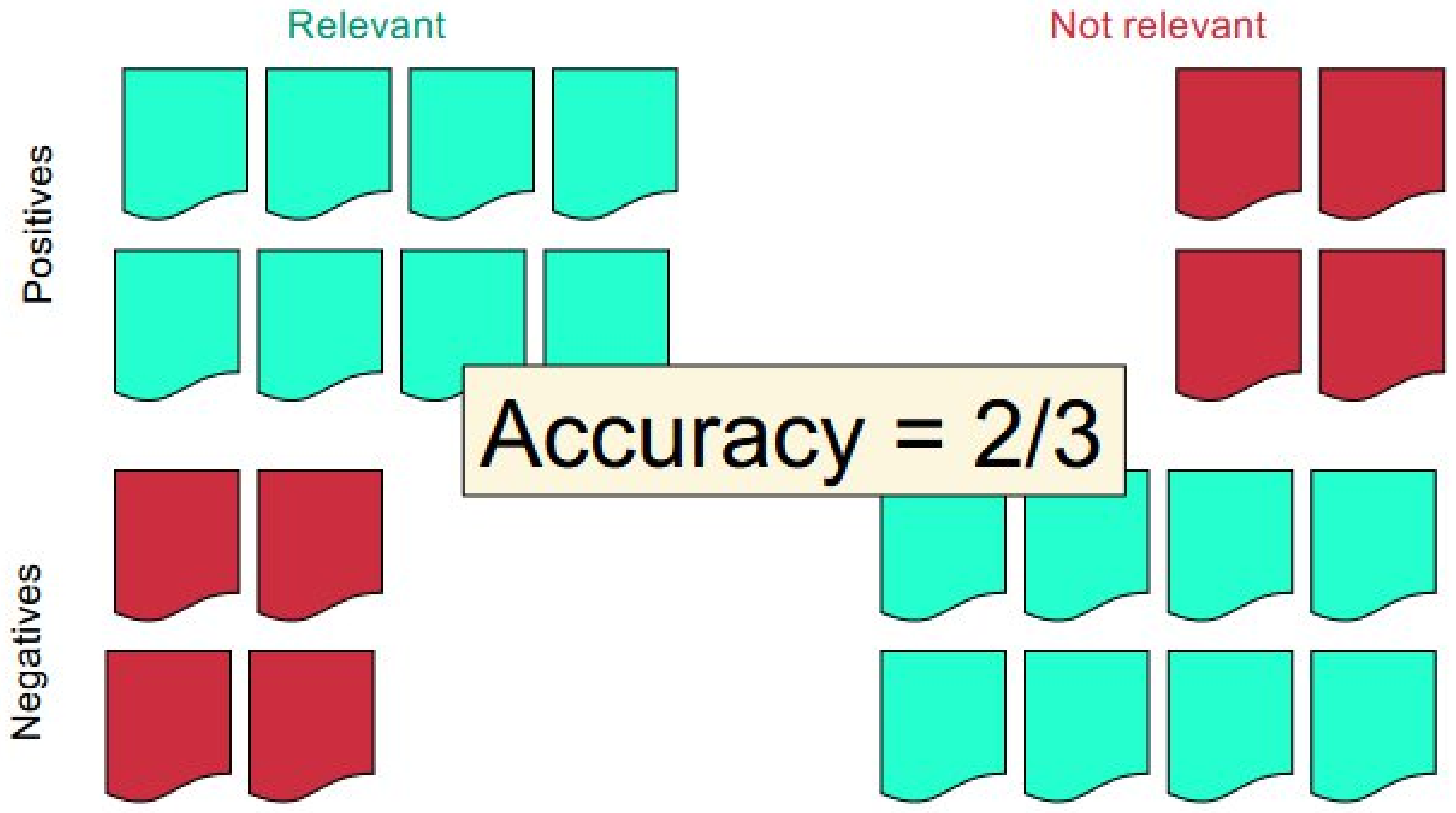


Specificity and Accuracy



Specificity and Accuracy

Specificity: 100%



Defining all the terms

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

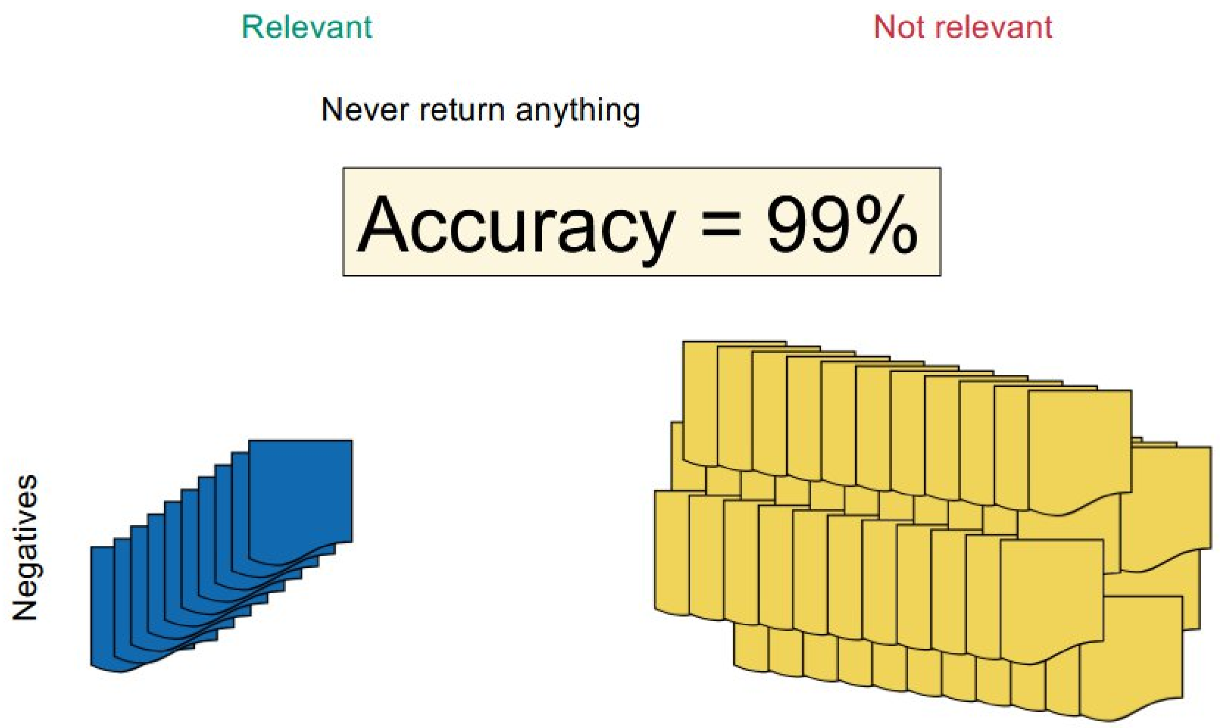
Precision: How useful are the returned results

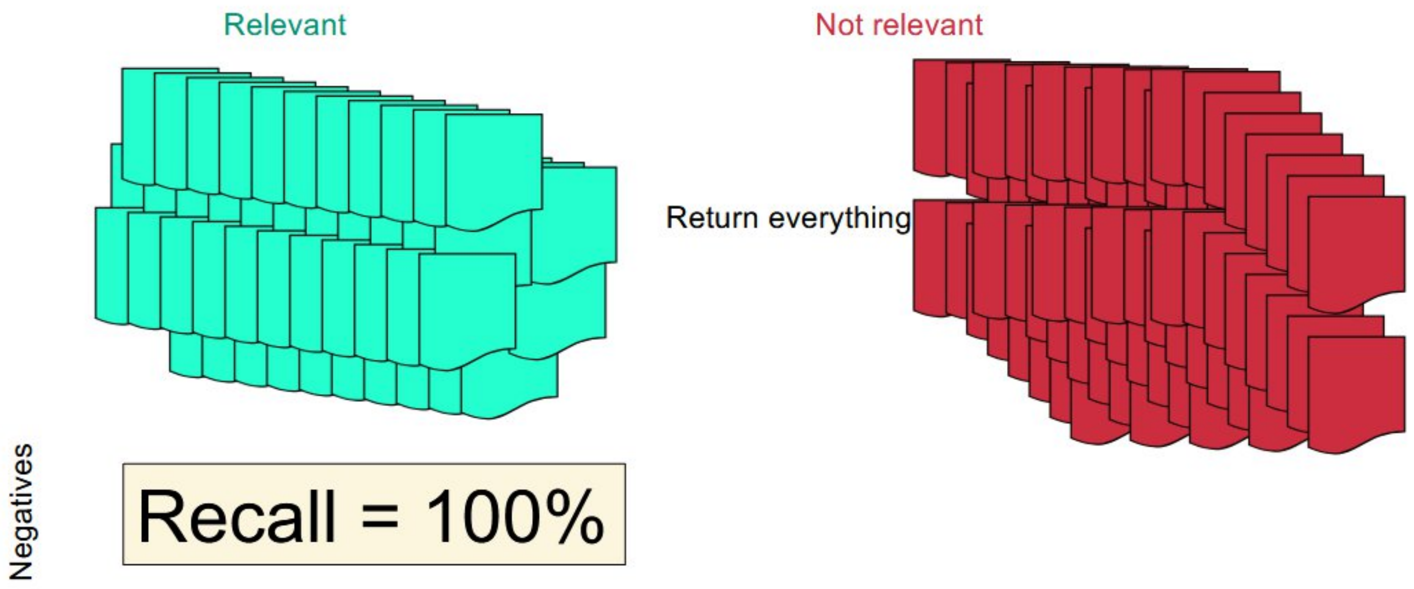
Specificity: How good is the system at not bothering you with useless stuff

Accuracy: How good is the system in total

Hacking Recall and Accuracy

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

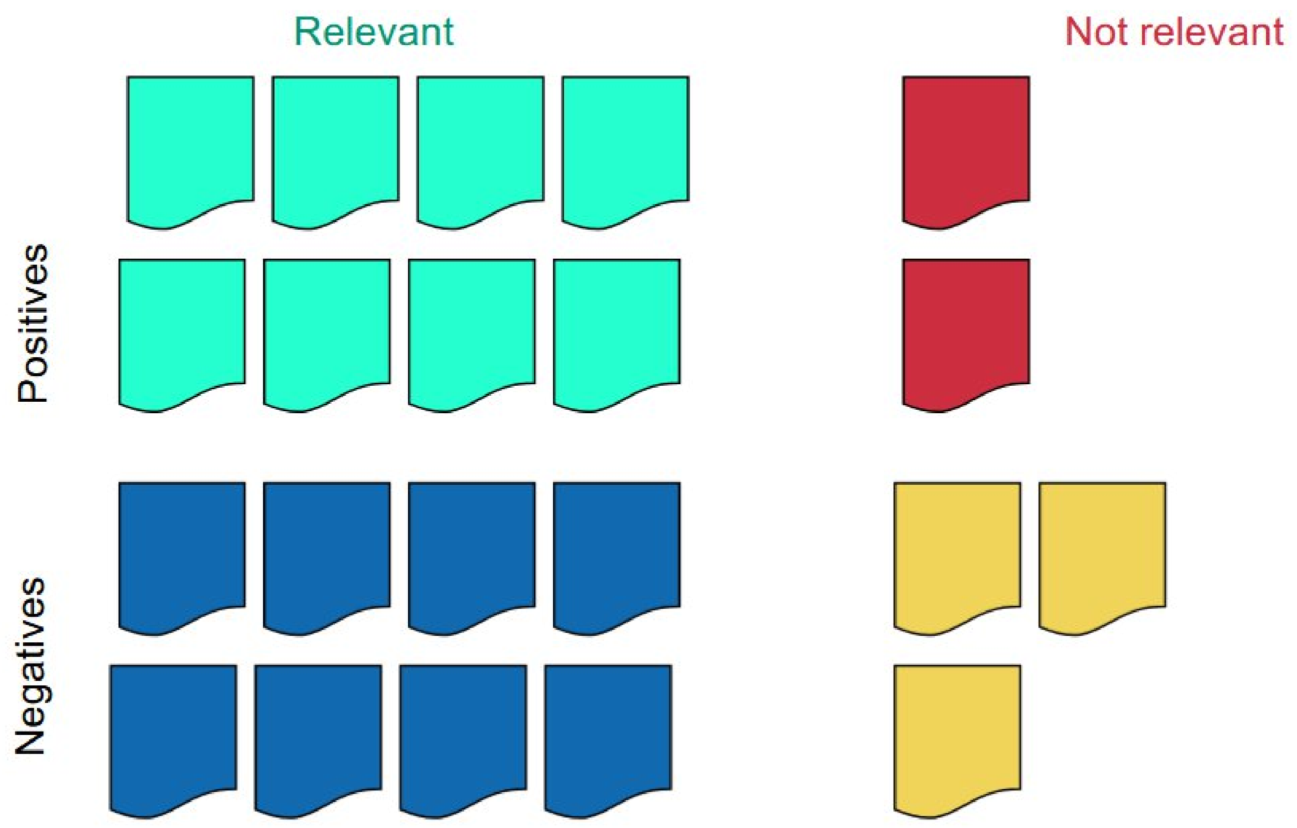




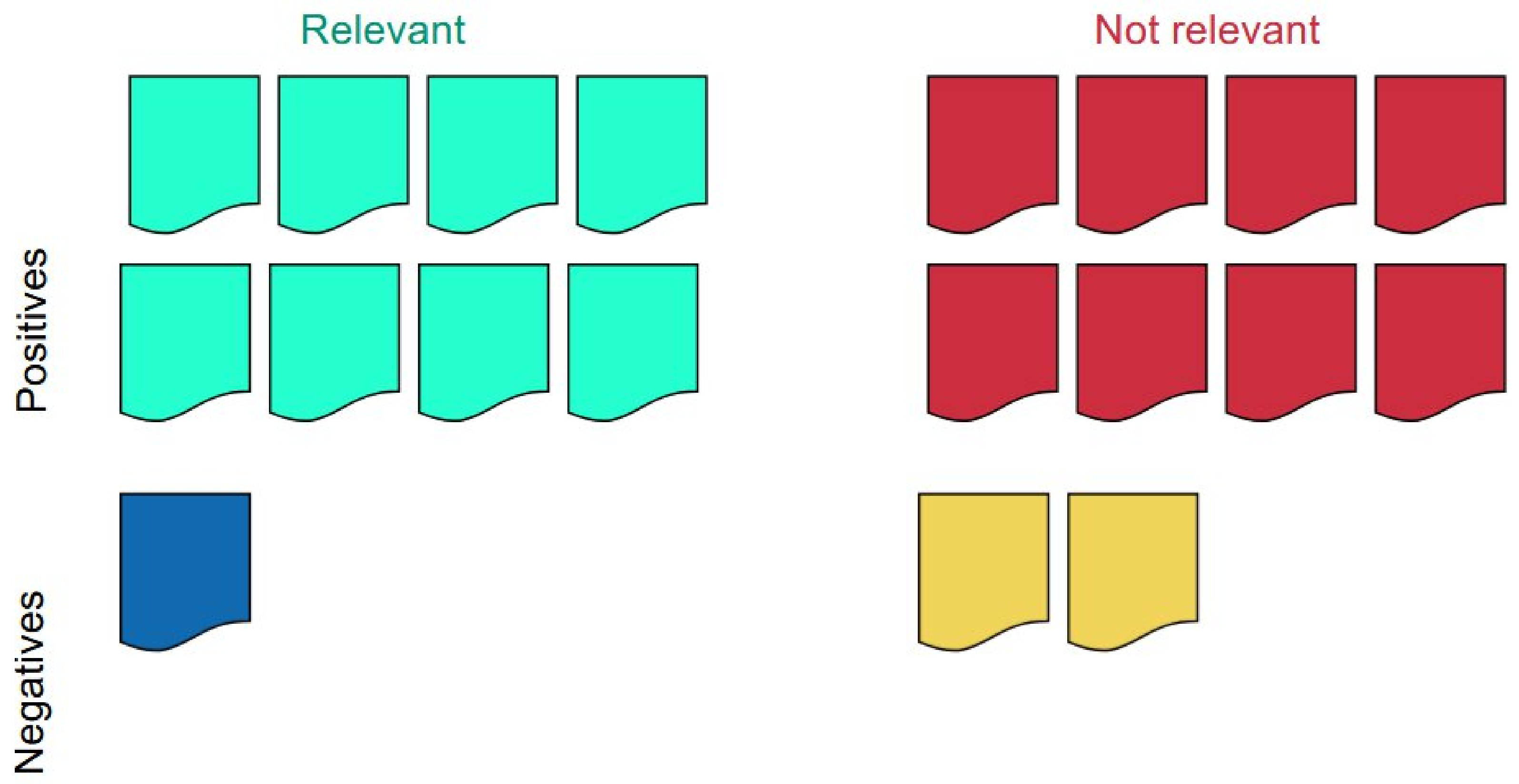
Evaluation

Compromise

High precision



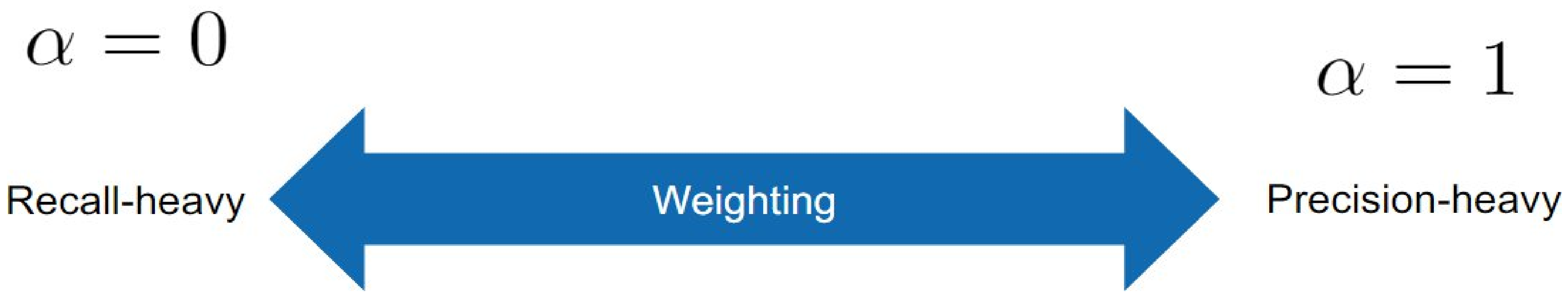
High recall



Merge Precision and Recall: F-Measure

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

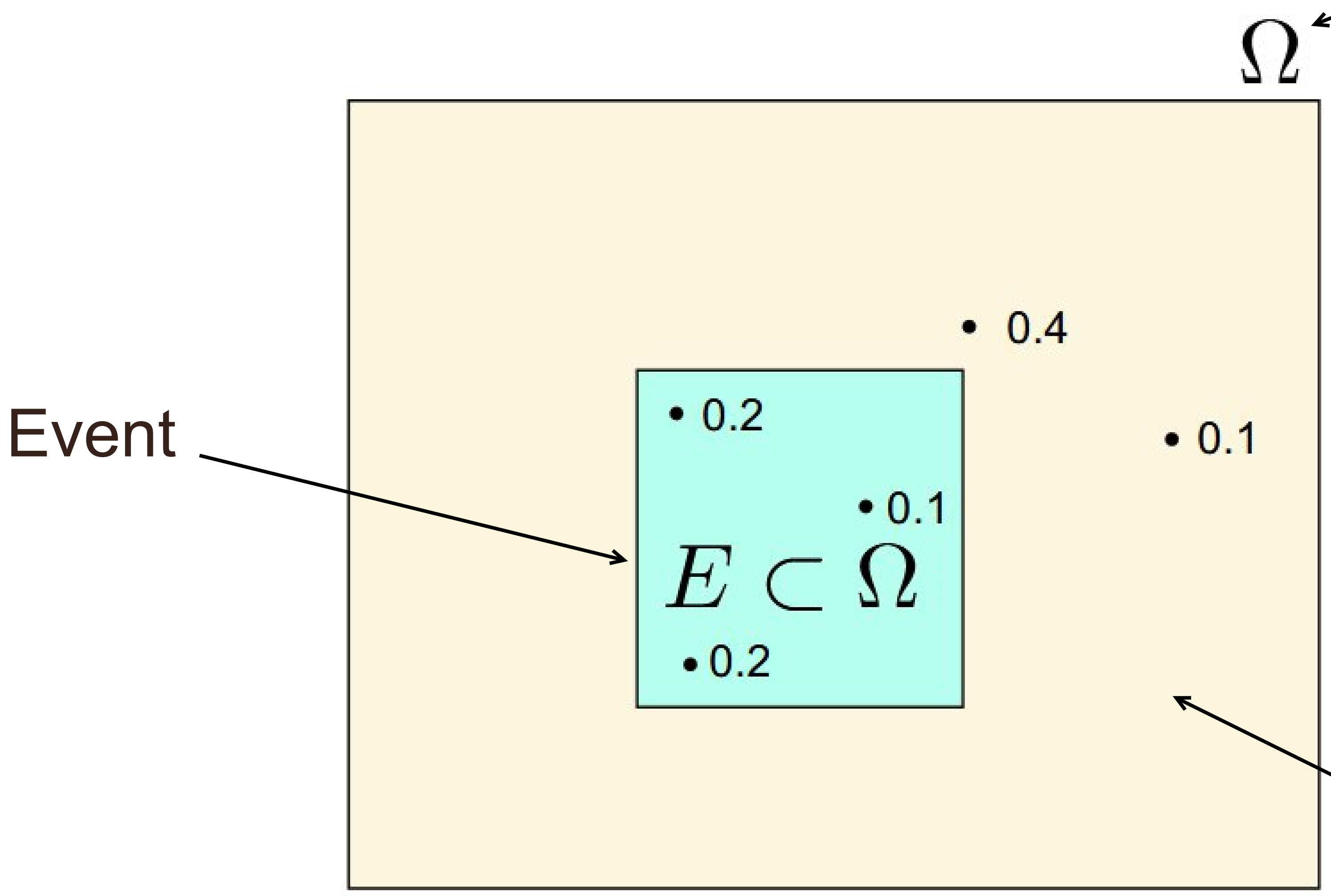
$$F_\alpha = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$$



Definitions

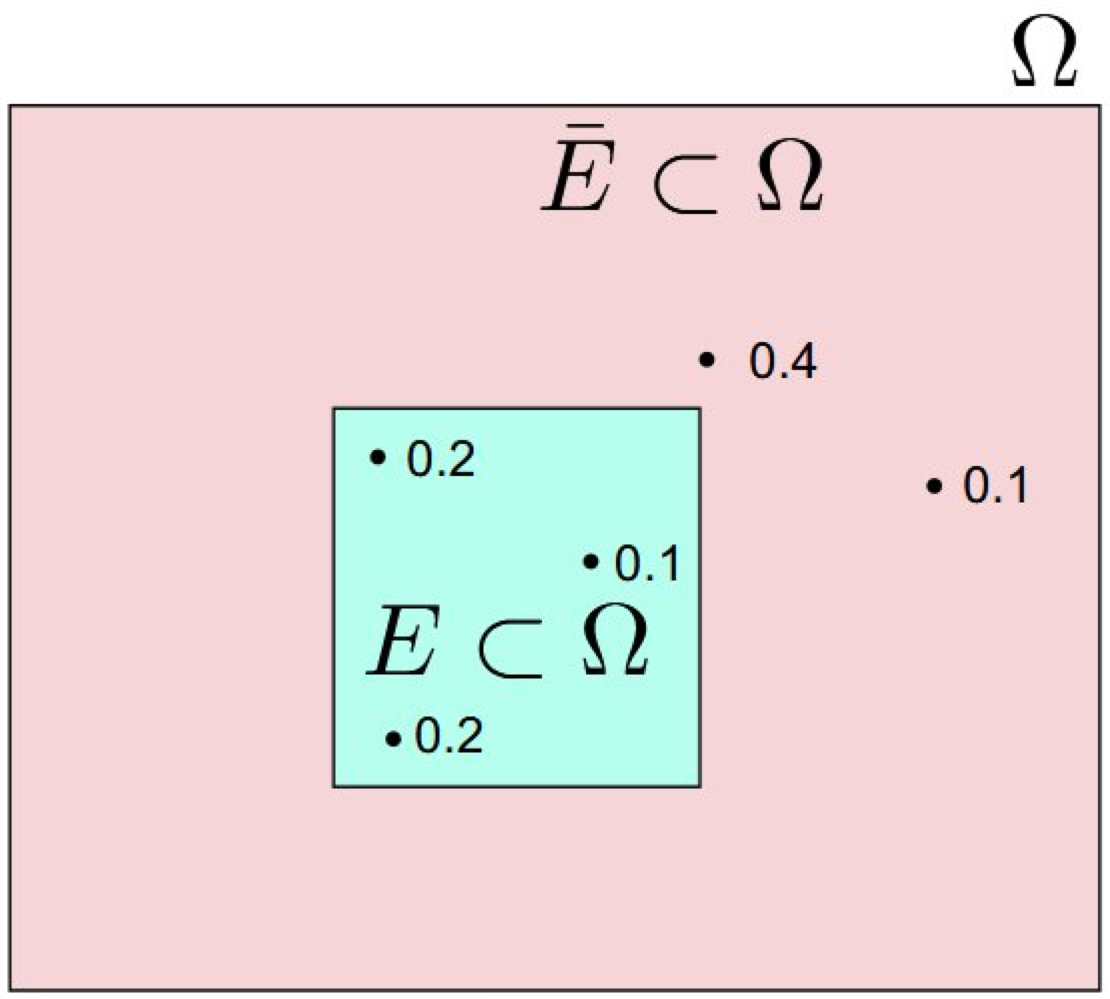
$$P: \begin{array}{l} \Omega \to [0,1] \\ \omega \mapsto P(\omega) \end{array}$$

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



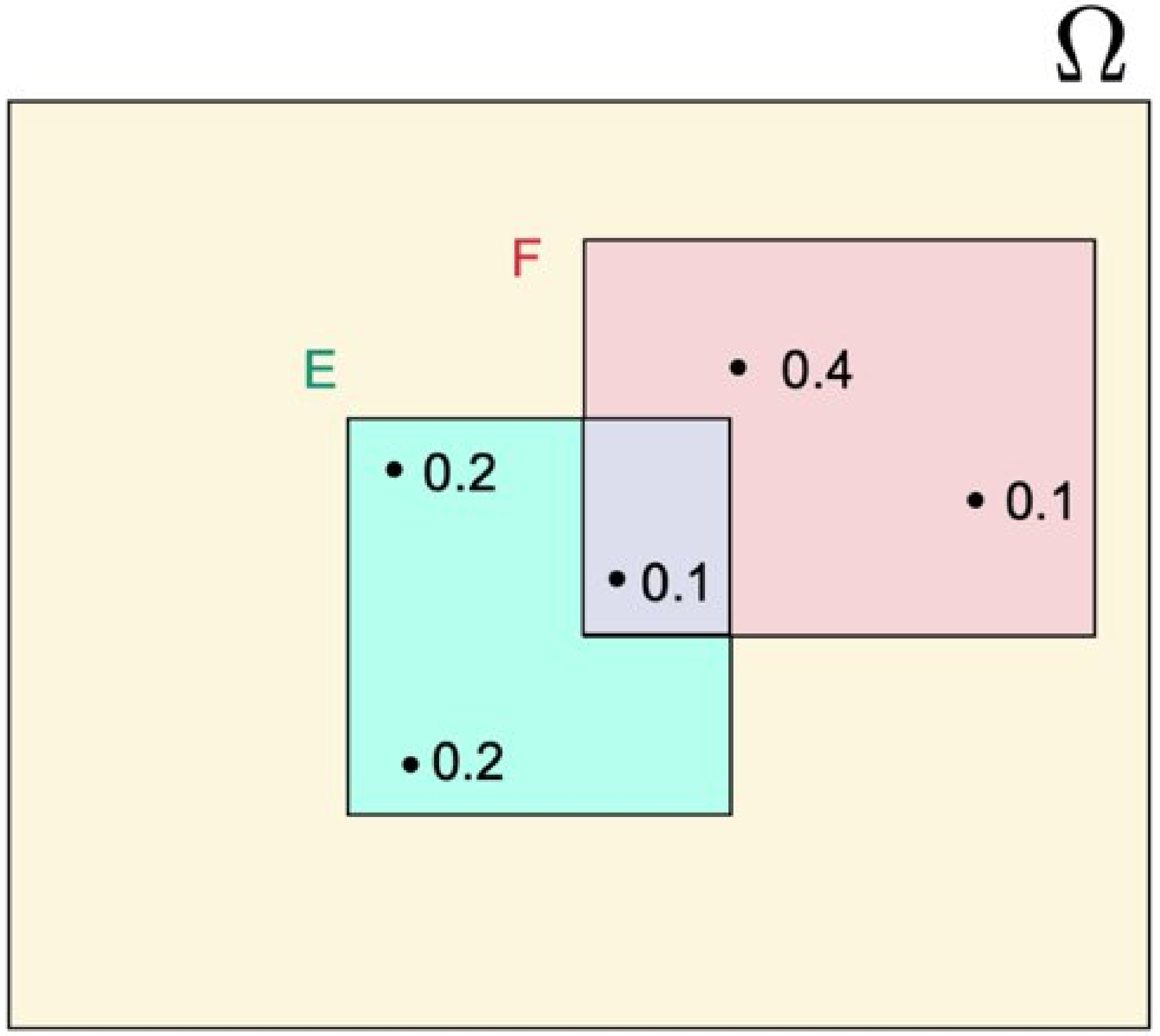
$$P(E) = \sum_{\omega \in E} P(\omega) \in [0,1]$$

Odds



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

Bayes' Rule



$$p(E|F) = \frac{P(F|E)}{P(F)} \times P(E)$$

posterior prior

Notation

$$p_X(\blacksquare) \qquad P(X = \blacksquare)$$

$$p_X(\text{⬟}) \qquad P(X = \text{⬟})$$

$$p_X(\blacktriangle) \qquad P(X = \blacktriangle)$$

No go!

$$\cancel{P(\blacksquare) = 0.5}$$

No go!

Mystery

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

Kahoot

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