Master Degree in Computer Science

Information Retrieval

Evaluation in information retrieval

Prof. Alfio Ferrara

Department of Computer Science, Università degli Studi di Milano
Room 7012 via Celoria 18, 20133 Milano, Italia alfio.ferrara@unimi.it

sed noli modo

Goals of evaluation

The goal of the evaluation activity is to **assess the quality of results obtained by an IR system**

The notion of *quality of results* depends on the task at hand, e.g.., search, classification, knowledge extraction, etc.

A general issue concerning the evaluation is that it is based on a **ground truth** (or **gold standard**), that is an **annotated corpus** where, for each document, we know if the document is **relevant** with respect to the task

Ground truth may be created by manually annotating documents and/or derived from data with a reference annotation system

Search evaluation: the notion of quality

Given a corpus ***C*** and a query ***q***, the task of document search is to find the set of documents $\boldsymbol{A_{q,C}}$ that match ***q***

We call $\boldsymbol{A_{q,C}}$ the **answers** to ***q***

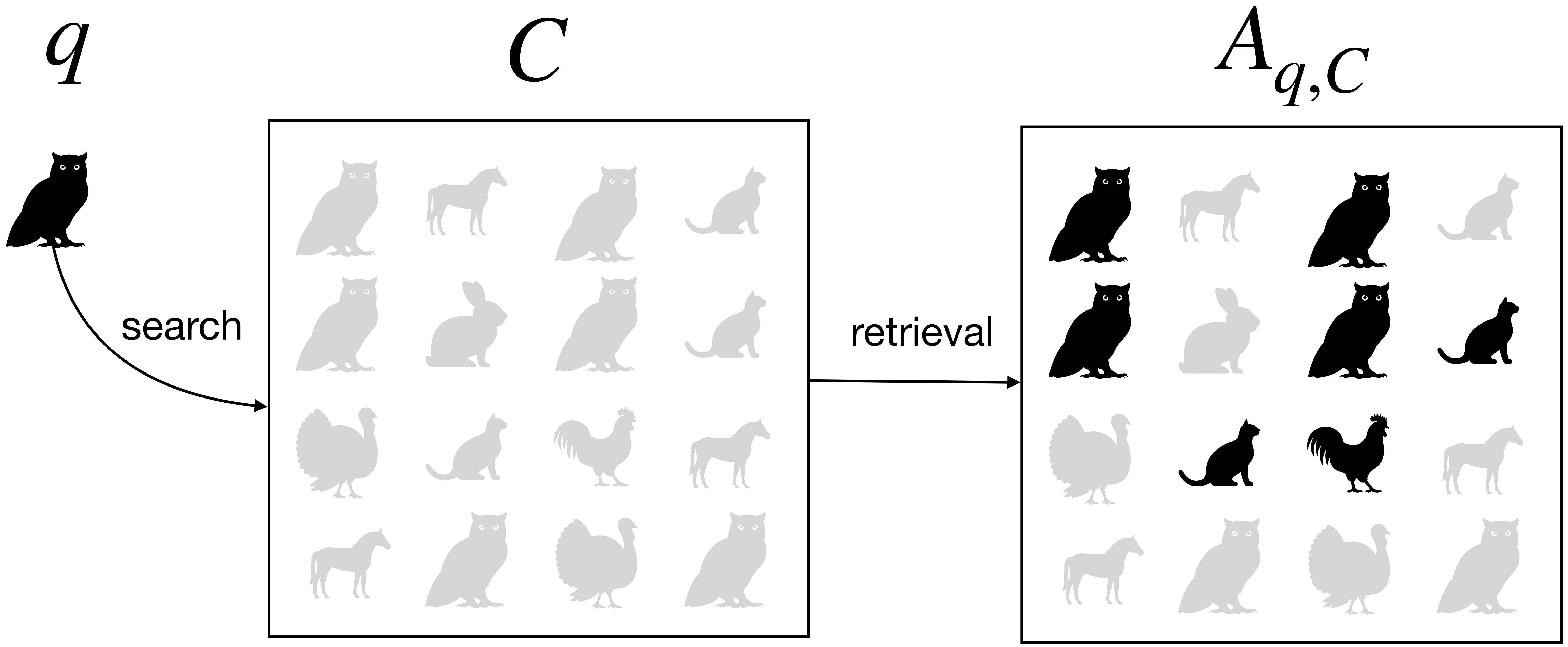
Question: when we say that the search answers $\boldsymbol{A_{q,C}}$ are good?

Search evaluation: the notion of quality

Question: when we say that the search answers ***A***$\boldsymbol{A_{q,C}}$ are good?

Definition 1: *when the documents contained in* $\boldsymbol{A_{q,C}}$ *are relevant to* ***q***

Remember: in order to know if a document is actually relevant, we need a **ground truth** (or a **user feedback**)



We retrieved 7 documents of which 4 are correct

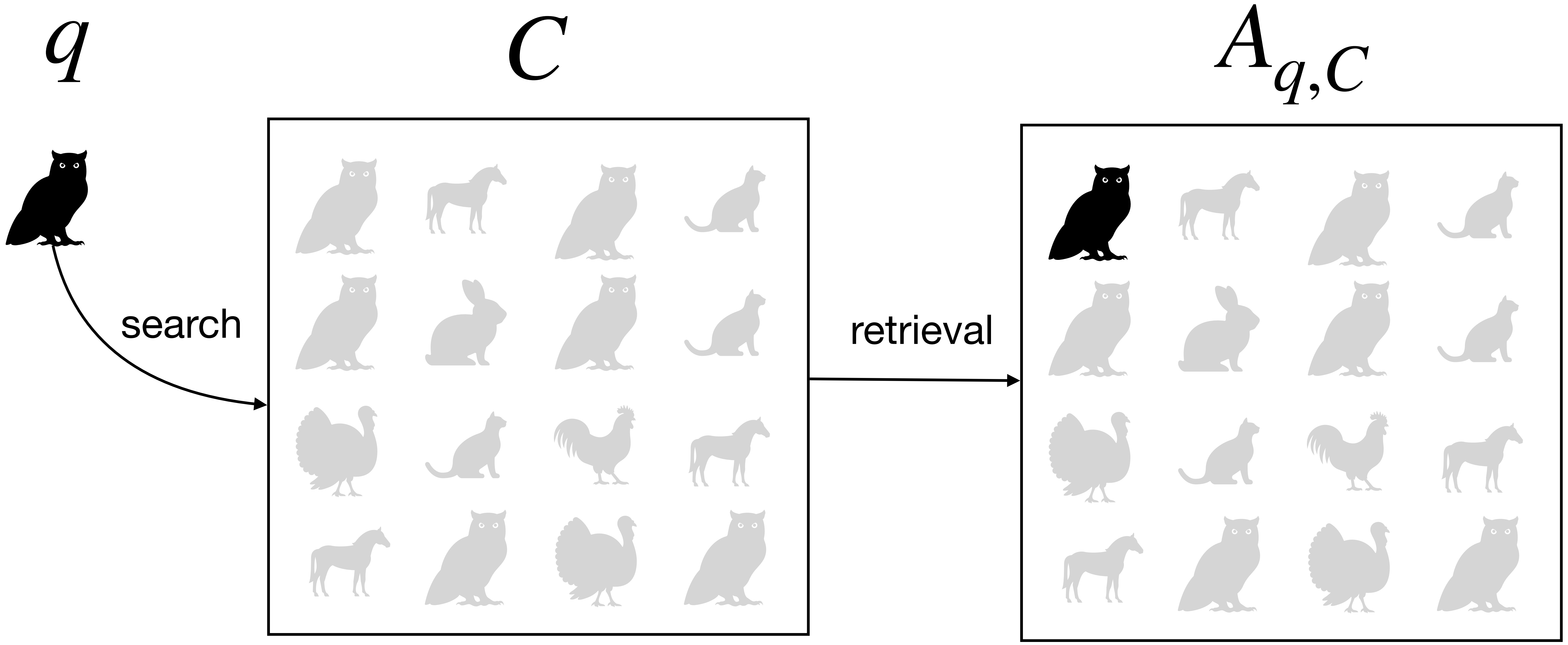
We can measure the quality of our system according to this notion of quality, called **Precision**

$$Prec = \frac{relevant\ retrieved}{retrieved}$$

$$Prec_q = \frac{4}{7} = 0.57$$

Search evaluation: the notion of quality

Question: why **Precision** is a **necessary** but **not sufficient** property of a good search system?



We retrieved only 1 document and it is correct

However, many relevant documents arre missing

$$Prec = \frac{relevant\ retrieved}{retrieved}$$

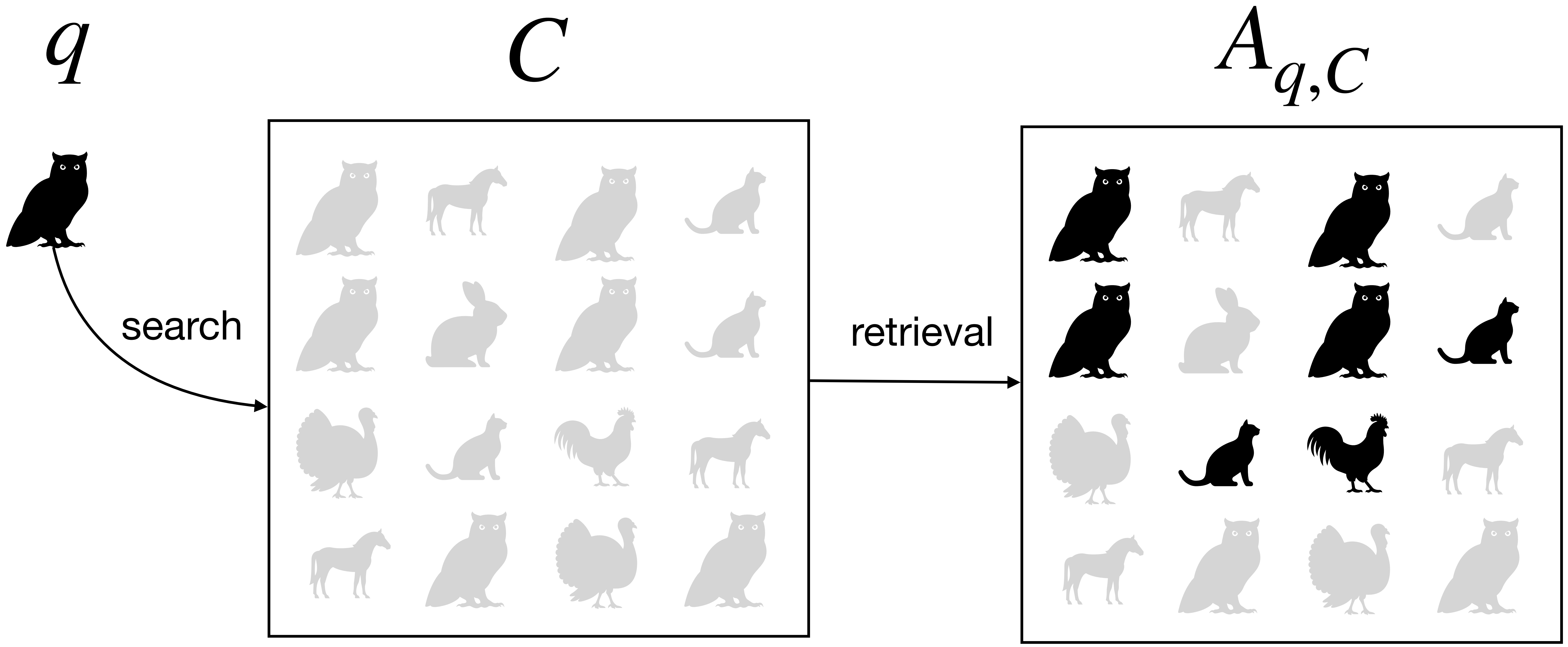
$$Prec_q = \frac{1}{1} = 1$$

Search evaluation: the notion of quality

Question: when we say that the search answers $\boldsymbol{A_{q,C}}$ are good?

Definition 2: *when all the relevant documents contained in **C** are retrieved by **q***

Remember: in order to know if a document is actually relevant, we need a **ground truth** (or a **user feedback**)



We retrieved 4 relevant documents from a corpus which contains 6 relevant documents

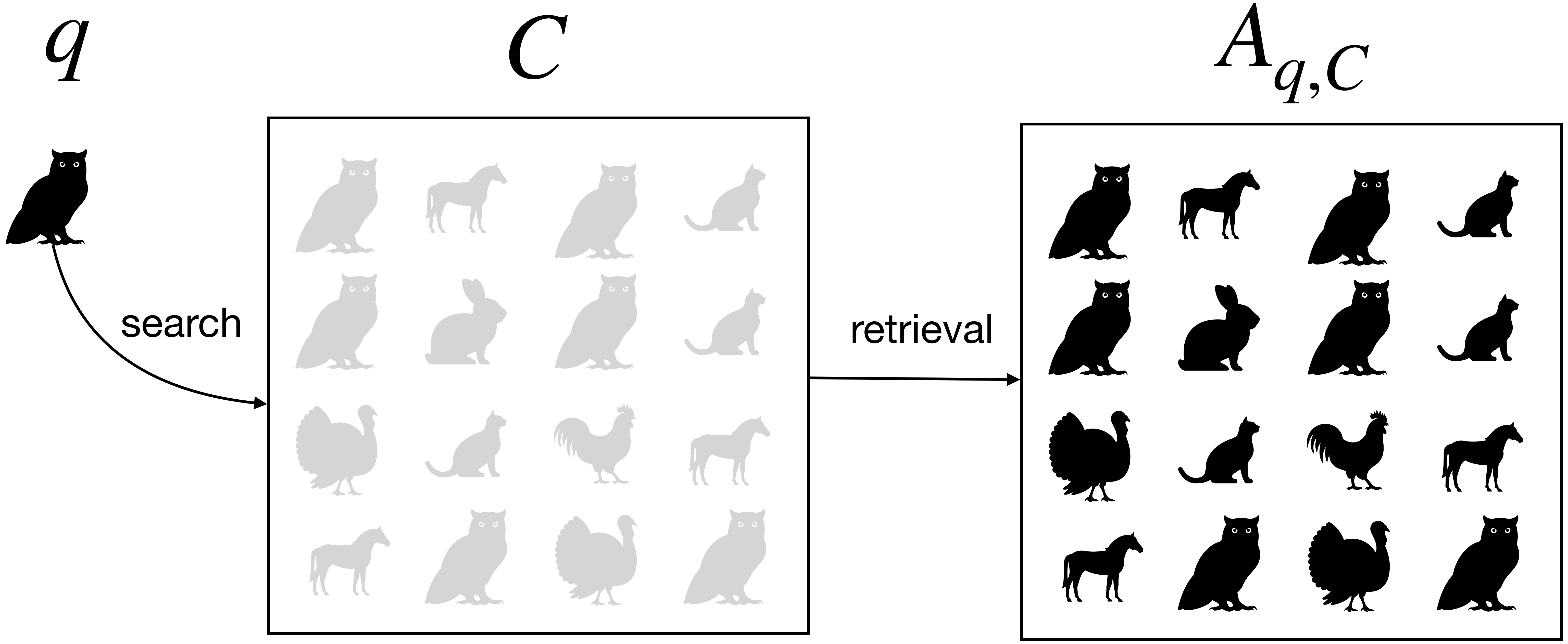
We can measure the quality of our system according to this notion of quality, called **Recall**

$$Rec = \frac{relevant\ retrieved}{relevant}$$

$$Rec_q = \frac{4}{6} = 0.66$$

Search evaluation: the notion of quality

Question: why **Recall** is a **necessary** but **not sufficient** property of a good search system?



We retrieved all the documents and this means to retrieve all the relevant ones by definition

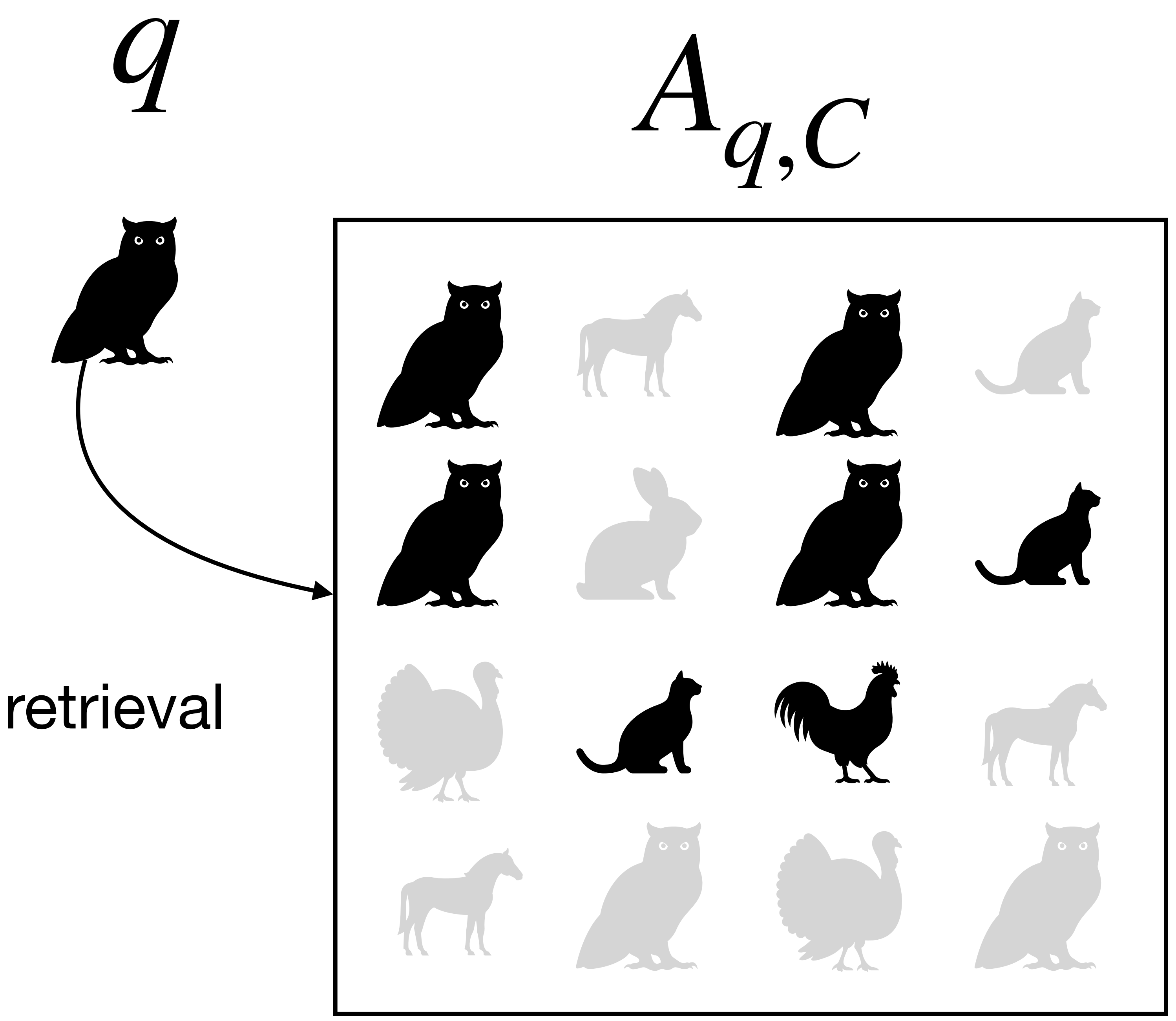
However, there are a lot of wrong results

$$Rec = \frac{relevant\ retrieved}{relevant}$$

$$Rec_q = \frac{6}{6} = 1$$

Search evaluation: the notion of quality

Definition 3: we aim at a system with a good tradeoff between precision and recall; this can be measured by the **f1-score**



$$F1 = \frac{2 \cdot Prec \cdot Rec}{Prec + Rec} = \frac{2 \cdot 0.57 \cdot 0.66}{0.57 + 0.66} = 0.61$$

Question: when the numbers we obtain from these measures are good? Try to perform search by tossing a coin…

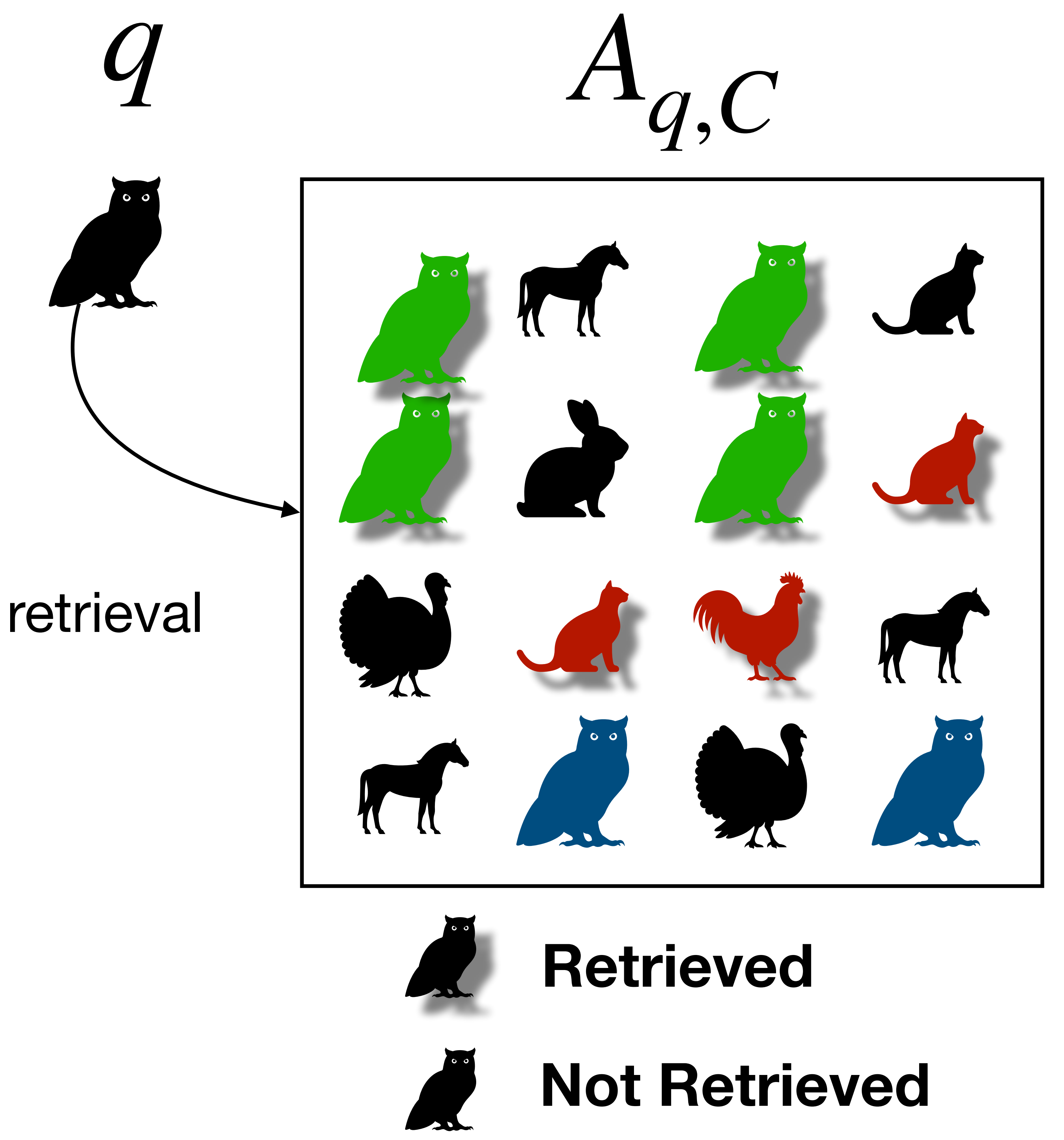
A more formal definition of Precision and Recall

Given a query q and a ground truth providing the set E_q of relevant documents for q, we denote A_q the set of query answers provided by the system under evaluation

For each document d:

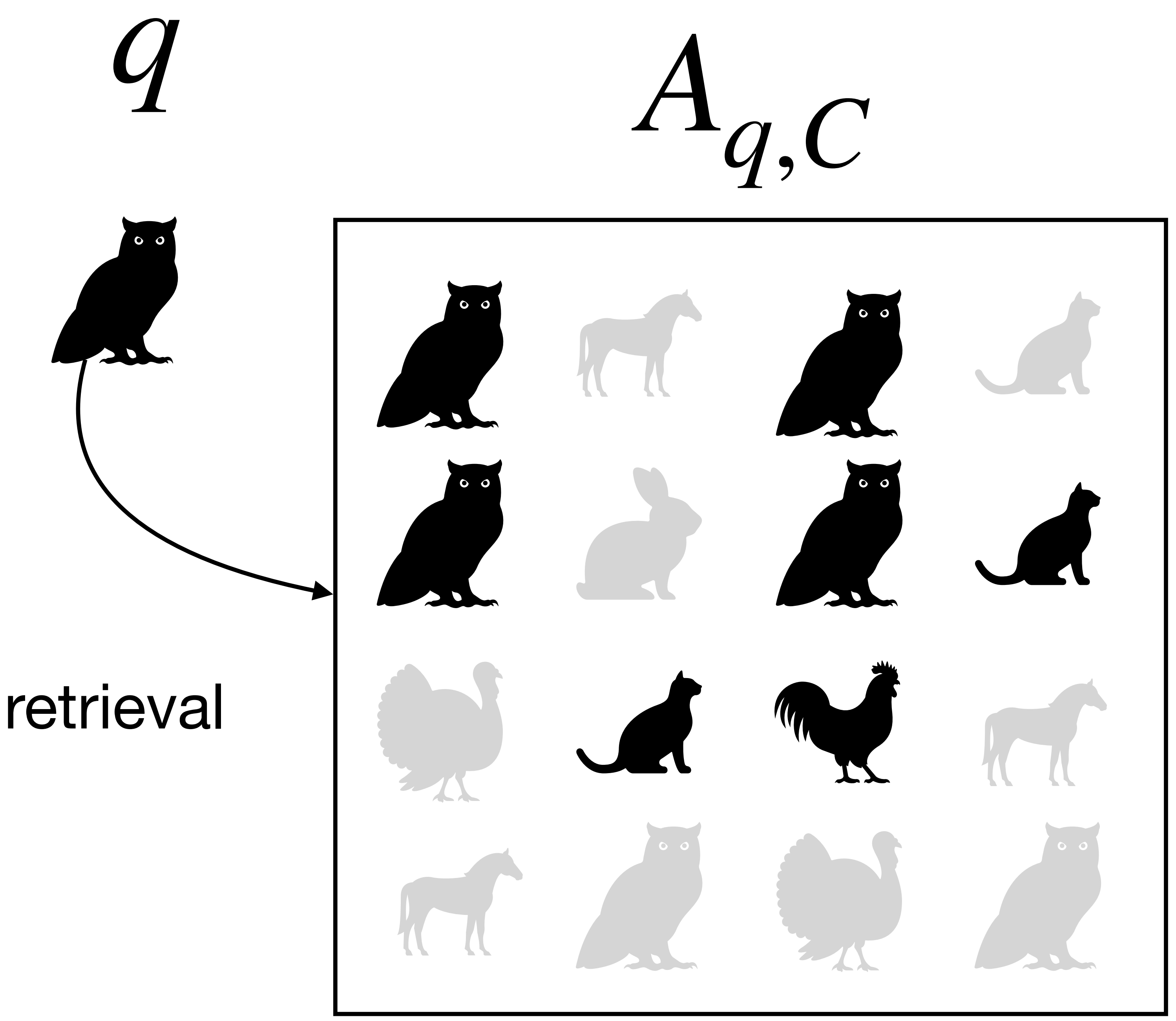
	$d \in E_q$	$d \notin E_q$	
$d \in A_q$	**TP** True Positive	**FP** False Positive	**Retrieved**
$d \notin A_q$	**FN** False Negative	**TN** True Negative	**Not retrieved**
	Relevant	**Not Relevant**	

$$Prec = \frac{TP}{TP + FP};\ \ Rec = \frac{TP}{TP + FN};\ \ F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$



Confusion Matrix

When evaluating a search system, it is important to understand in which cases we have errors: it is more the r**etrieval of wrong documents (FP)** or instead the fact that we **miss many documents (FN)**



		Ground truth		
	q	$R = 1$	$R = 0$	
Predicted values	$R = 1$	$TP = 4$	$FP = 3$	7
	$R = 0$	$FN = 2$	$TN = 7$	9
		6	10	16

R is the variable that represent the document relevance to the query

Search evaluation: confusion matrix and other measures

specificity	negative predictive value	miss rate	fall-out	false discovery rate	false omission rate	critical success index
$TNR = \frac{TN}{TN + FP}$	$NPV = \frac{TN}{TN + FN}$	$FNR = \frac{FN}{FN + TP}$	$FPR = \frac{FP}{FP + TN}$	$FDR = \frac{FP}{FP + TP}$	$FOR = \frac{FN}{FN + TN}$	$TS = \frac{TP}{TP + FN + FP}$

Prevalence threshold

$$PT = \frac{\sqrt{Rec(1 - TNR)} + TNR - 1}{Rec + TNR - 1}$$

Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced accuracy

$$BA = \frac{Rec + TNR}{2}$$

Informedness

$$BM = Rec + TNR - 1$$

Markedness

$$MK = Prec + NPV - 1$$

Matthews correlation coefficient

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

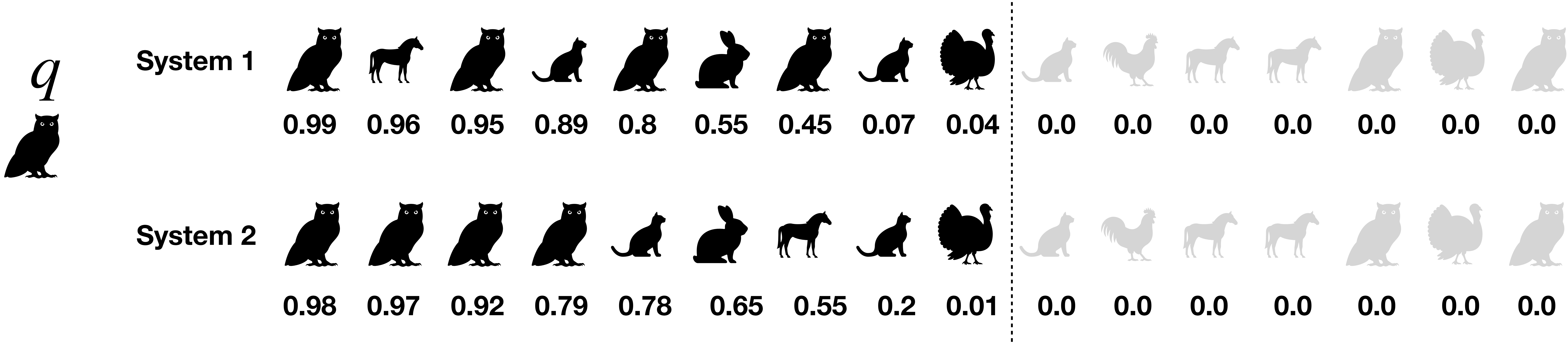
Fowlkes-Mallows index

$$FM = \sqrt{\frac{2TP}{(TP + FP)(TP + FN)}}$$

Quick summary: *https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)*

Evaluation of ranking systems

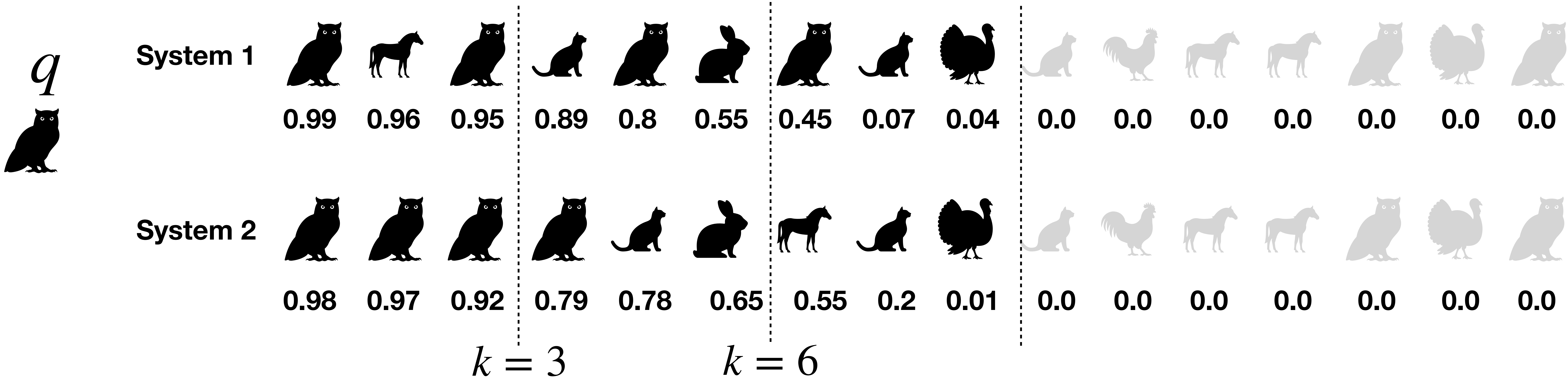
The answer of a non boolean search system is not a set of retrieved documents, rather a rank of documents with a relevance score (that is typically the cosine similarity between the query and the document). **How do we evaluate the system in this case?**



Note that the two systems achieve the **same Precision and Recall.** However, the second system is better.

Evaluation of ranking systems

Solution 1: we could set a threshold too select the top-k results and use them to evaluate precision and recall. But where should we put the threshold?

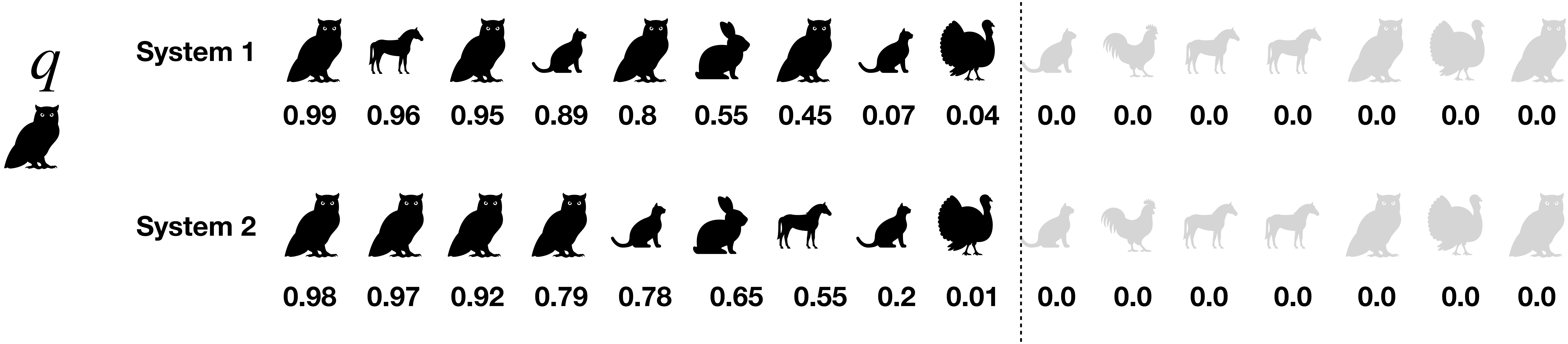


Precision at k **System 1:** $Prec_{k=3} = \frac{2}{3};\ Prec_{k=6} = \frac{1}{2}$ **System 2:** $Prec_{k=3} = 1;\ Prec_{k=6} = \frac{2}{3}$

Still the order of results is not completely taken into account

Evaluation of ranking systems

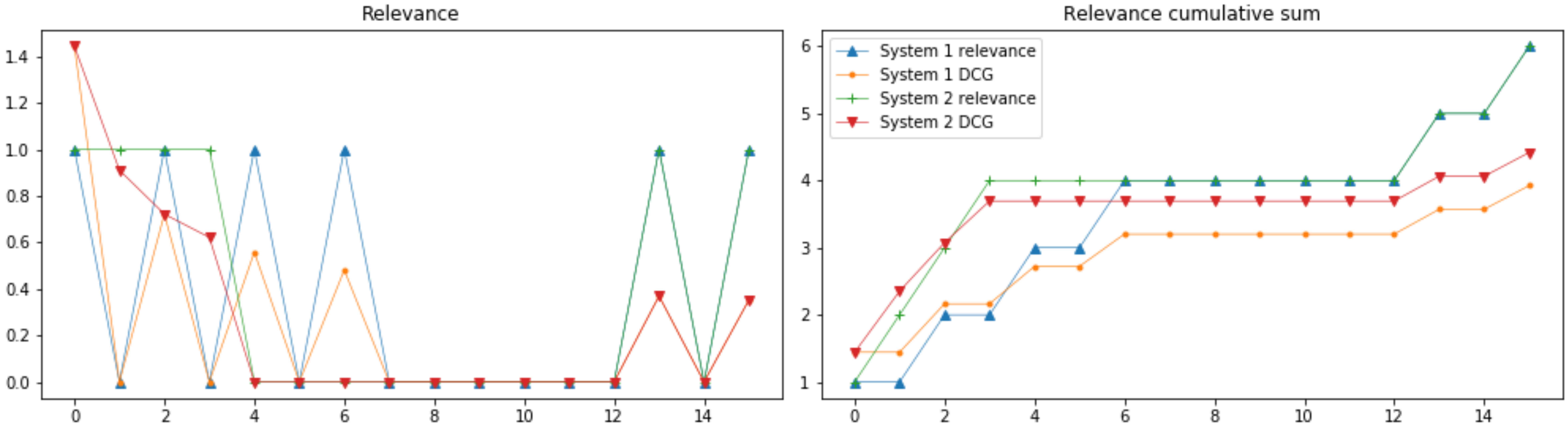
Solution 2: **Discounted cumulative gain**: we discount the relevance of each document according to its position in the ranking



Discounted cumulative gain: $DCG = \sum_{i=1}^{n} \frac{R_i}{\log(i+1)}$

Evaluation of ranking systems

Solution 2: **Discounted cumulative gain**: we discount the relevance of each document according to its position in the ranking

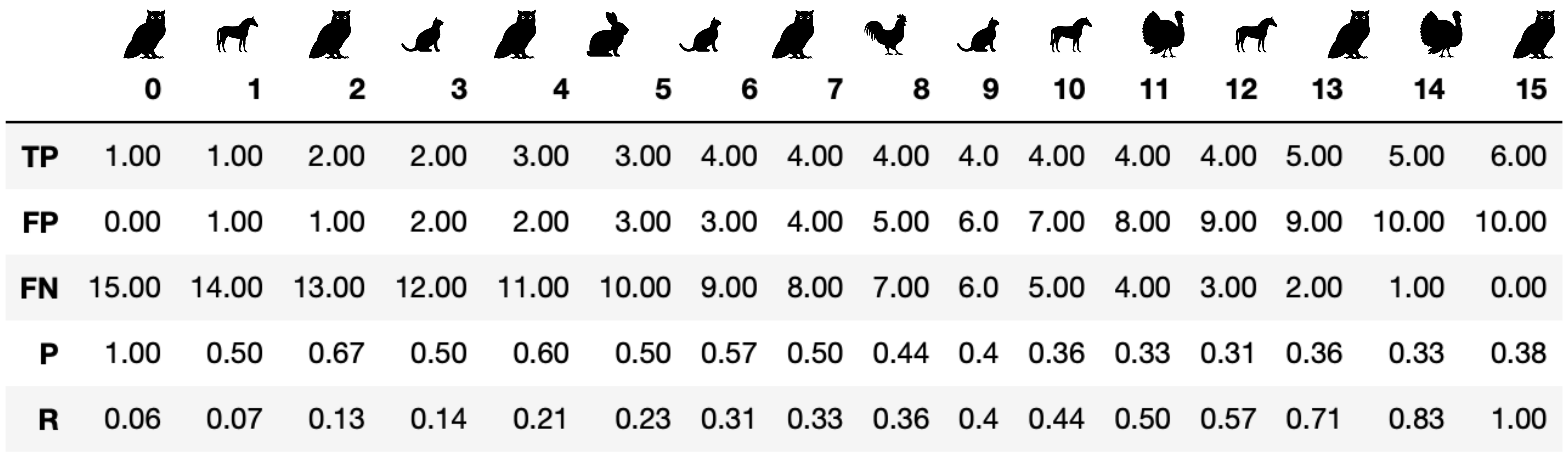


System 1: 3.93 System 2: 4.42

Evaluation of ranking systems

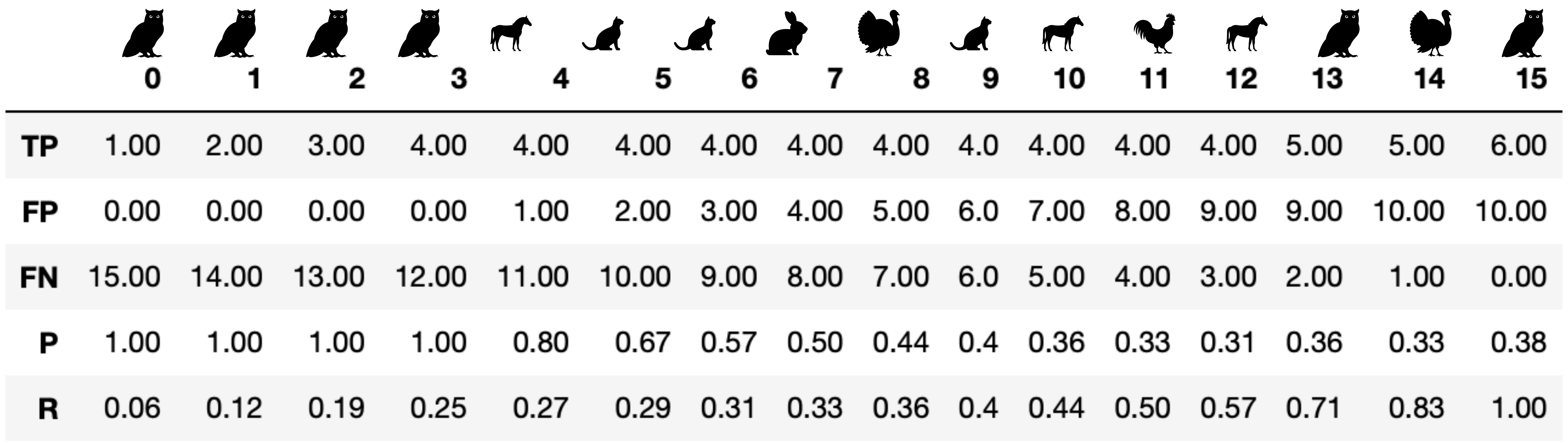
Solution 3: Precision Vs Recall. When moving along the ranking from top to bottom, the **Recall** increases by definition. We can just measure the **Precision** scored for different levels of recall.

System 1



	🦉	🐴	🦉	🐈	🦉	🐇	🐈	🦉	🐓	🐈	🐴	🦃	🐴	🦉	🦃	🦉
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
TP	1.00	1.00	2.00	2.00	3.00	3.00	4.00	4.00	4.00	4.0	4.00	4.00	4.00	5.00	5.00	6.00
FP	0.00	1.00	1.00	2.00	2.00	3.00	3.00	4.00	5.00	6.0	7.00	8.00	9.00	9.00	10.00	10.00
FN	15.00	14.00	13.00	12.00	11.00	10.00	9.00	8.00	7.00	6.0	5.00	4.00	3.00	2.00	1.00	0.00
P	1.00	0.50	0.67	0.50	0.60	0.50	0.57	0.50	0.44	0.4	0.36	0.33	0.31	0.36	0.33	0.38
R	0.06	0.07	0.13	0.14	0.21	0.23	0.31	0.33	0.36	0.4	0.44	0.50	0.57	0.71	0.83	1.00

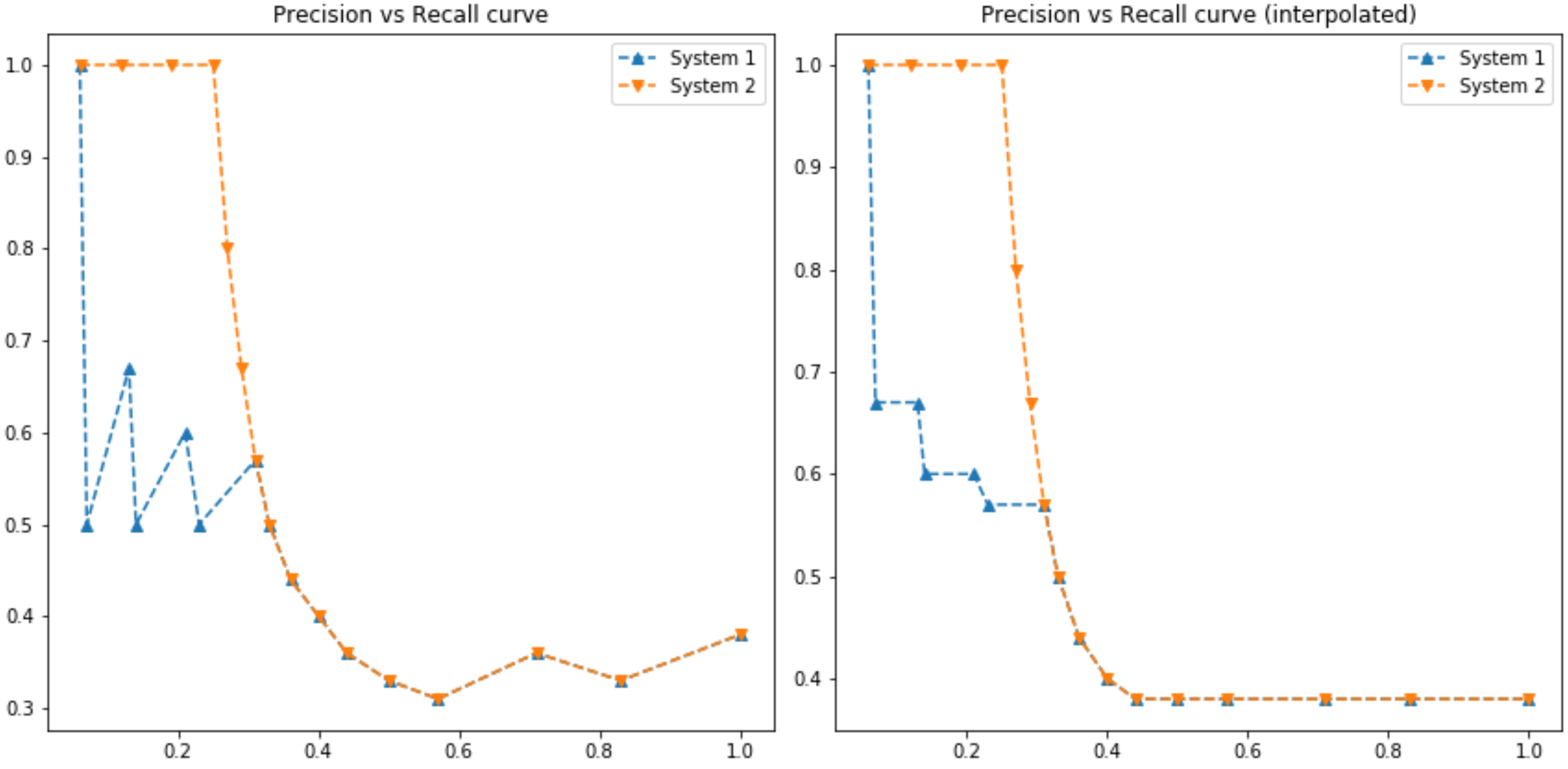
System 2



	🦉	🦉	🦉	🦉	🐴	🐈	🐈	🐇	🦃	🐈	🐴	🐓	🐴	🦉	🦃	🦉
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
TP	1.00	2.00	3.00	4.00	4.00	4.00	4.00	4.00	4.00	4.0	4.00	4.00	4.00	5.00	5.00	6.00
FP	0.00	0.00	0.00	0.00	1.00	2.00	3.00	4.00	5.00	6.0	7.00	8.00	9.00	9.00	10.00	10.00
FN	15.00	14.00	13.00	12.00	11.00	10.00	9.00	8.00	7.00	6.0	5.00	4.00	3.00	2.00	1.00	0.00
P	1.00	1.00	1.00	1.00	0.80	0.67	0.57	0.50	0.44	0.4	0.36	0.33	0.31	0.36	0.33	0.38
R	0.06	0.12	0.19	0.25	0.27	0.29	0.31	0.33	0.36	0.4	0.44	0.50	0.57	0.71	0.83	1.00

Precision vs Recall curve

$AvgP = \int_0^1 P(R)dR \approx \sum_{k=1}^{n} P(k)\Delta R(k)$, **where** $\Delta R(k)$ **is the change in** R **from** $k-1$ **to** k

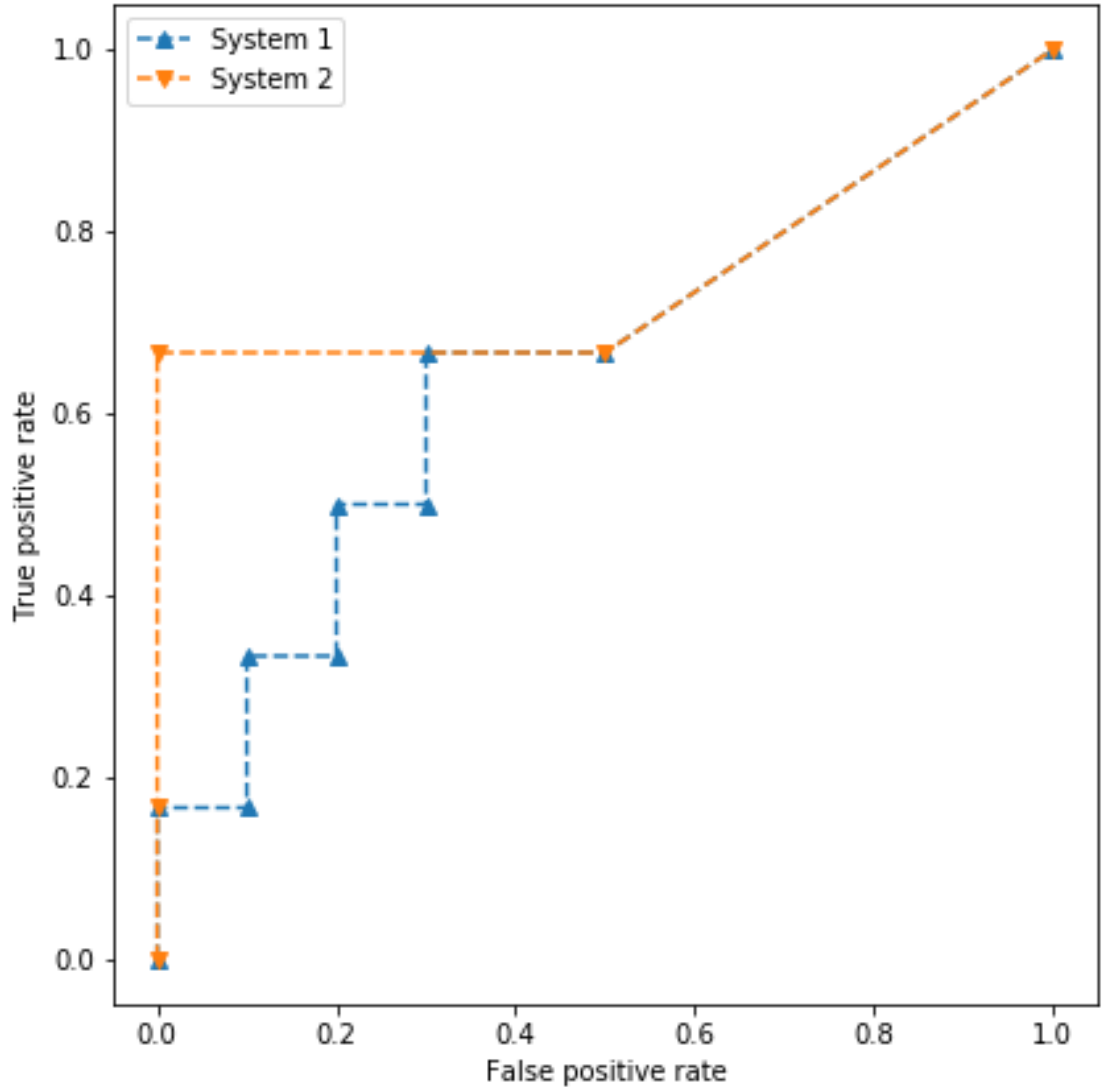


$$Prec_{interpolated}(Rec_i) = \max_{j \geq i} Prec(Rec_j)$$

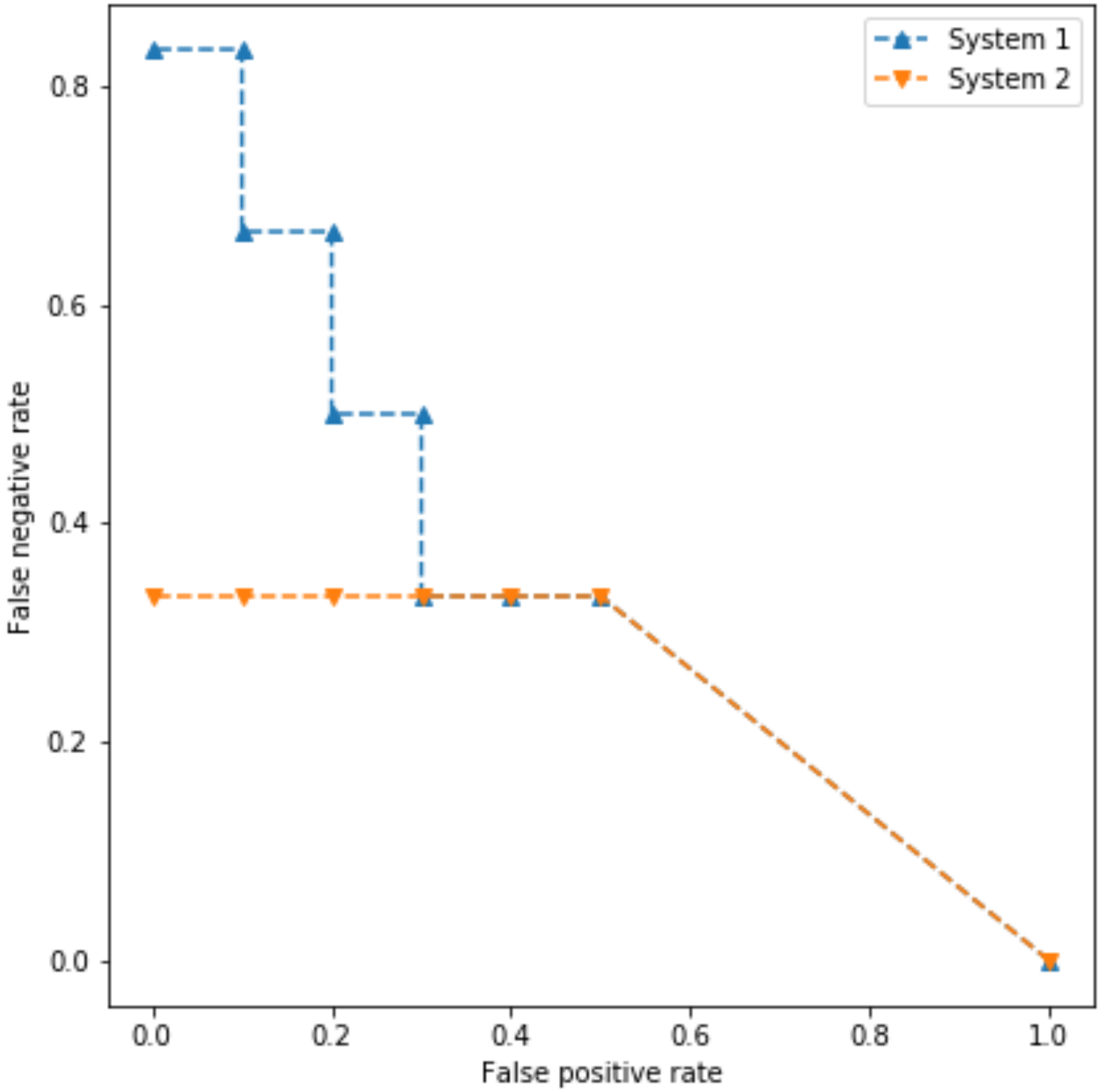
Other curves

The **ROC (receiver operating characteristic)** curve is created by plotting the **true positive rate (TPR) (Recall)** against the **false positive rate (FPR) (FP / (FP + TN))** at various threshold settings.

ROC (receiver operating characteristic)



Detection Error Tradeoff (DET)



The **Detection Error Tradeoff (DET)** curve is created by plotting the **false negative rate (FNR) (FN / (FN + TP)** against the **false positive rate (FPR) (FP / (FP + TN))** at various threshold settings.