

Relevance, Precision, and Recall

Evaluation, session 2

IR Evaluation

Evaluation is any process which produces a quantifiable measure of a system's performance.

In IR, there are many things we might want to measure.

Here, we focus mostly on *retrieval effectiveness*.

IR Evaluation Questions

- Are we presenting users with relevant documents?
- How long does it take to show the result list?
- Are our query suggestions useful?
- Is our presentation useful?
- Is our site appealing (from a marketing perspective)?

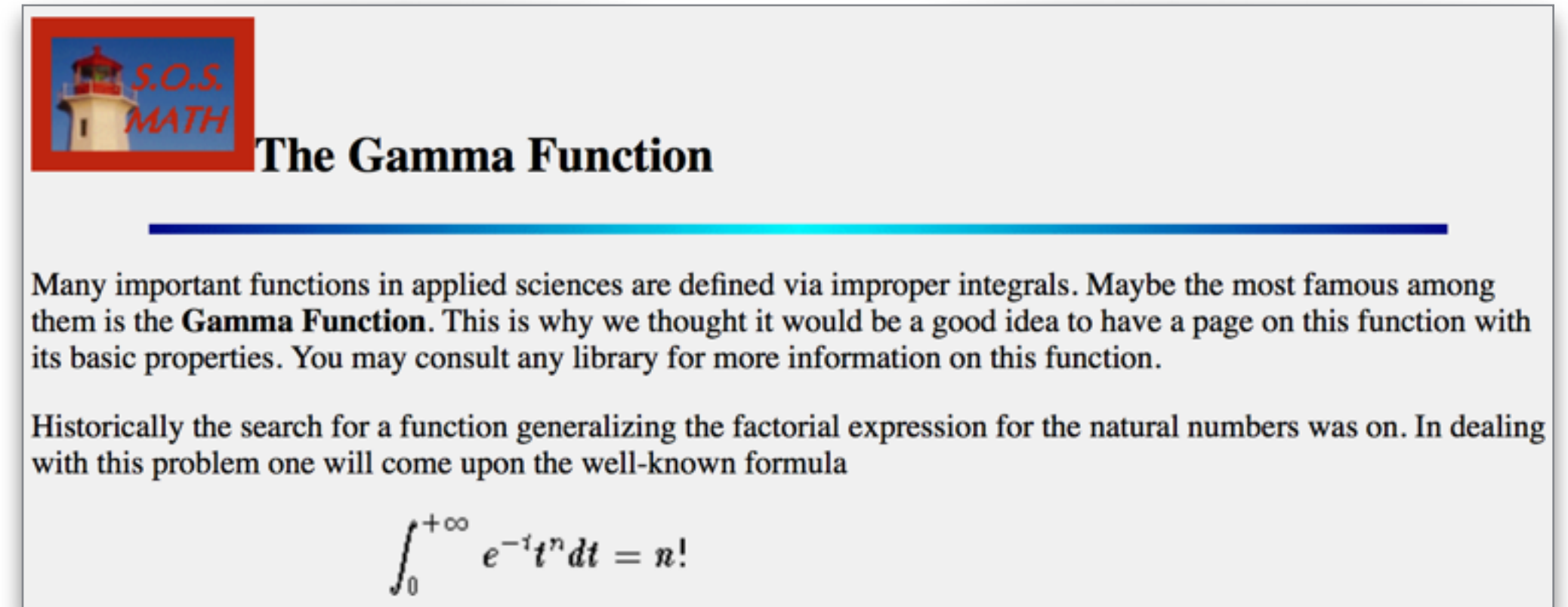
Retrieval Effectiveness

Retrieval effectiveness is inherently subjective, because the relevance of a document to a query is subjective.

Relevance roughly means “satisfying the information need,” but for a precise evaluation we need a precise definition.

The appropriate definition depends on the task you are evaluating.

<http://www.sosmath.com/calculus/improper/gamma/gamma.html>



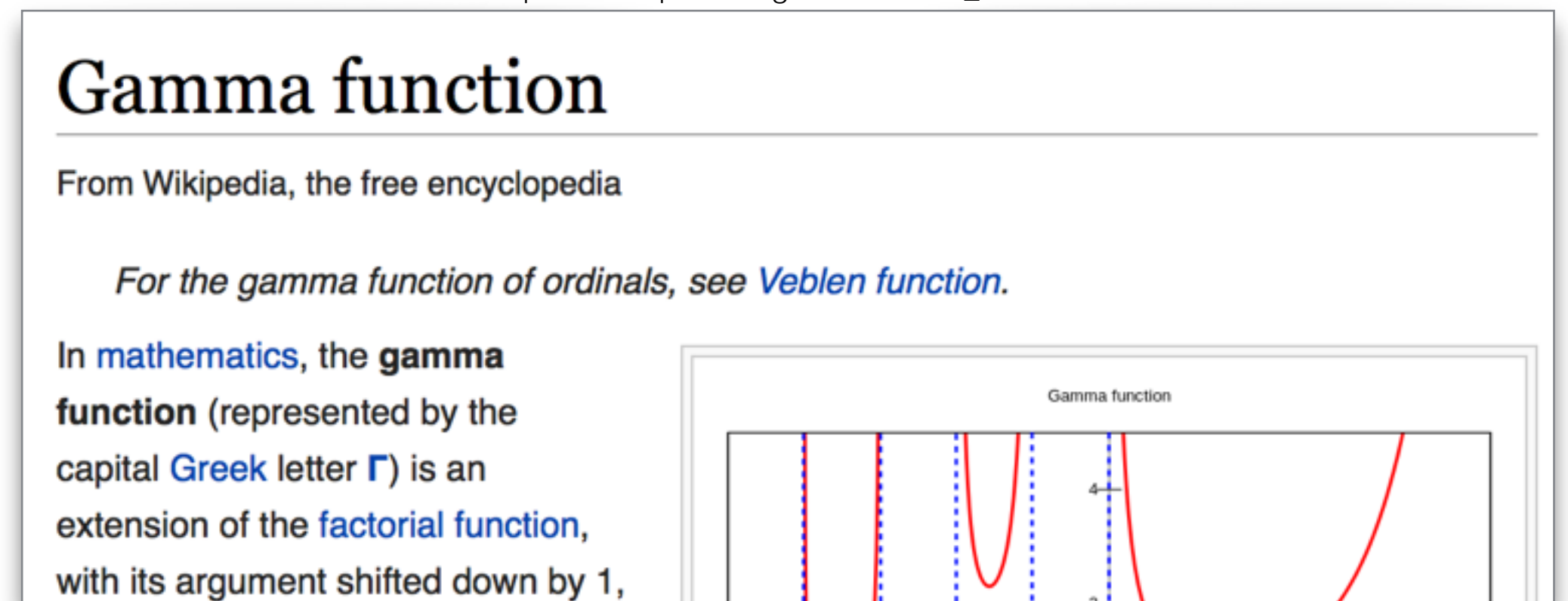
The Gamma Function

Many important functions in applied sciences are defined via improper integrals. Maybe the most famous among them is the **Gamma Function**. This is why we thought it would be a good idea to have a page on this function with its basic properties. You may consult any library for more information on this function.

Historically the search for a function generalizing the factorial expression for the natural numbers was on. In dealing with this problem one will come upon the well-known formula

$$\int_0^{+\infty} e^{-t} t^n dt = n!$$

http://en.wikipedia.org/wiki/Gamma_function

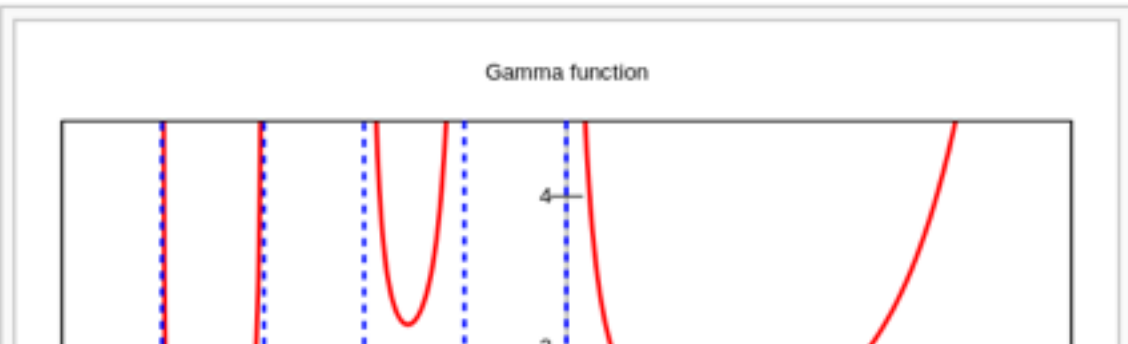


Gamma function

From Wikipedia, the free encyclopedia

For the gamma function of ordinals, see [Veblen function](#).

In **mathematics**, the **gamma function** (represented by the capital **Greek letter** Γ) is an extension of the **factorial function**, with its argument shifted down by 1,



Which is better? It depends.

Evaluating a Ranking

Given a ranking of documents, we can create a *confusion matrix* that counts the correct and incorrect answers of each type.

- *True Positives* are relevant documents in the ranking
- *False Positives* are non-relevant documents in the ranking
- *True Negatives* are non-relevant documents missing from the ranking
- *False Negatives* are relevant documents missing from the ranking

	Relevant	Non-Relevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Confusion Matrix

Recall

Recall is the fraction of relevant documents retrieved by the system.

Recall@k is the fraction of relevant documents in the top k results.

A task is said to be *recall-oriented* when the user wants to make sure they have not missed any relevant detail (e.g. legal discovery).

	Relevant	Non-Relevant
Retrieved	TP	FP
Not Retrieved	FN	TN

Confusion Matrix

$$\begin{aligned} \text{recall} &:= \frac{\text{num}(\text{retrieved relevant})}{\text{num}(\text{relevant})} \\ &= \frac{TP}{TP + FN} \end{aligned}$$

Precision

Precision is the fraction of retrieved documents that were relevant.

Precision@k is the fraction of the top k results that were relevant.

A task is said to be *precision-oriented* when the user wants just a few high-quality documents (e.g. most web search).

	Relevant	Non-Relevant
Retrieved	TP	FP
Not Retrieved	FN	TN

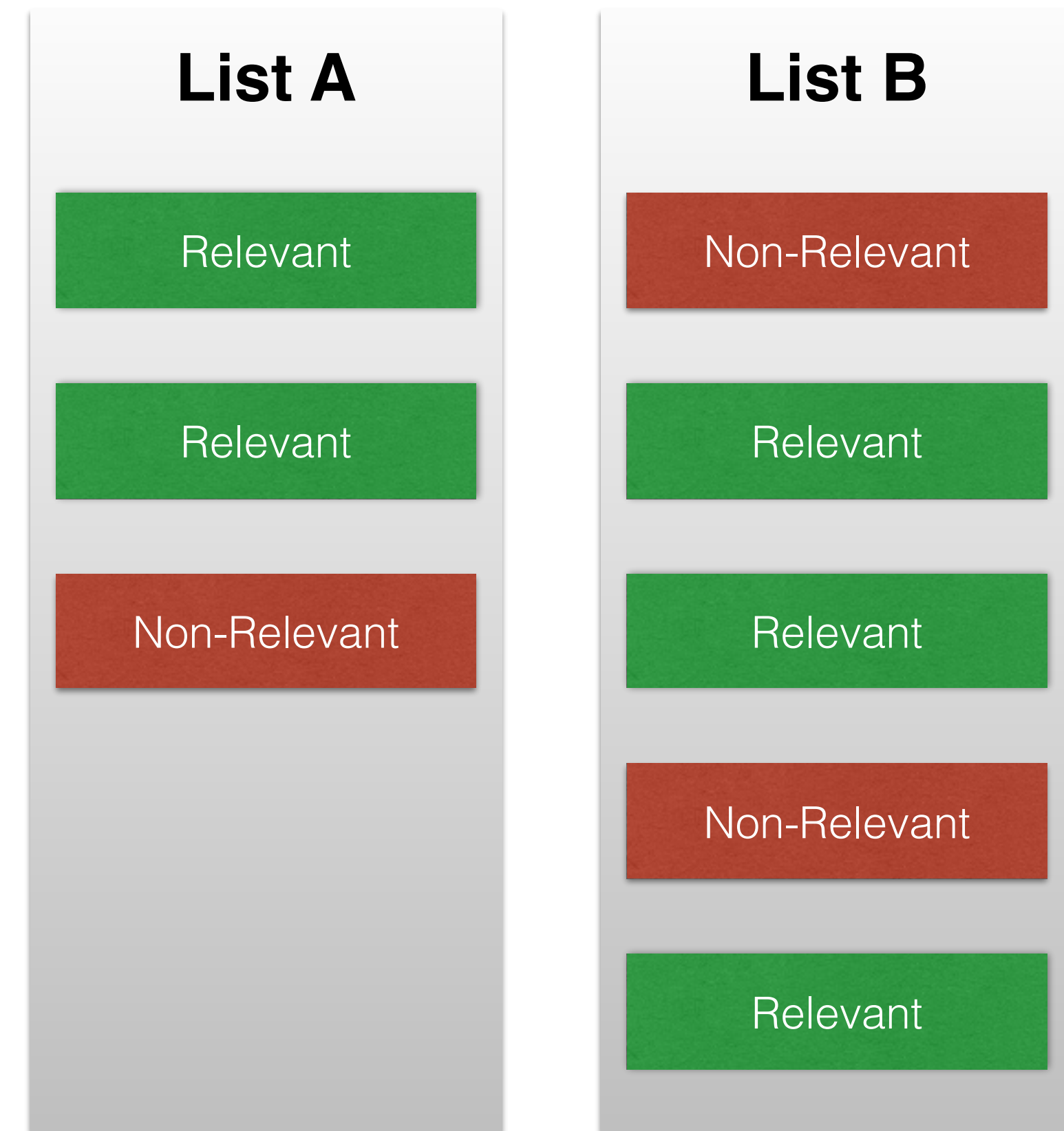
Confusion Matrix

$$\begin{aligned} \textit{precision} &:= \frac{\textit{num}(\textit{retrieved relevant})}{\textit{num}(\textit{retrieved})} \\ &= \frac{TP}{TP + FP} \end{aligned}$$

Precision vs. Recall

There is a tradeoff between recall and precision: usually increasing one will decrease the other.

The quality of a given ranking depends on whether your task is recall- or precision-oriented.



Which is better? It depends.

Wrapping Up

Correct evaluation depends on understanding the nature of the task you're evaluating. For instance, is it recall-oriented or precision-oriented?

Many other factors are also involved, and we'll discuss some of them in future videos.

Next, we'll look at the most commonly-used ways to measure the quality of a ranking.