MLRF Lecture 04

J. Chazalon, LRE/EPITA, 2025

IR evaluation

Lecture 04 part 03

How to evaluate a retrieval system?

We need a set of queries for which we know the expected results
“Ground truth”, aka “targets”, “gold standard”…

To compare 2 methods, we need to use the same database and the same queries.

Many measures / indicators.

Core criterion: is a result relevant (binary classification)?

Precision and Recall

Used to measure the balance between

- Returning many results, hence a lot of the relevant results present in the database, but also a lot of noise
- Returning very few results, leading to less noise, but also less relevant results

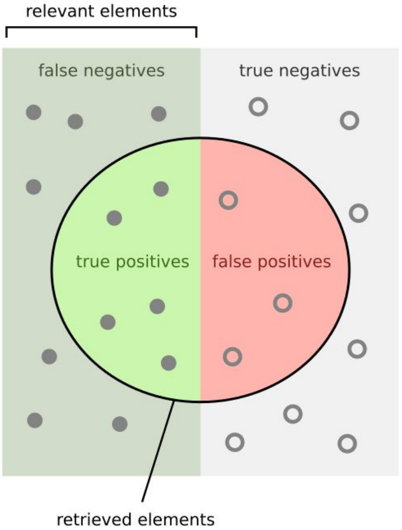
Precision and Recall

Precision (P) is the fraction of retrieved documents that are relevant

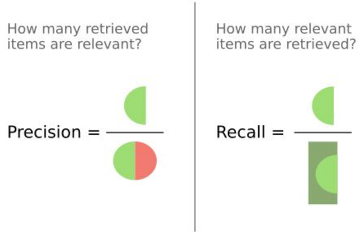
$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Recall (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$



https://en.wikipedia.org/wiki/File:Precisionrecall.svg



	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$\boxed{\begin{aligned} P &= tp/(tp+fp) \\ R &= tp/(tp+fn) \end{aligned}}$$

F-measure

F measure is the weighted harmonic mean of precision and recall

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

where $\alpha \in [\,0, 1\,]$ and thus $\beta^2 \in [\,0, \infty\,]$

The default value is $\beta = 1$, leading to:

$$\boxed{F_{\beta=1} = \frac{2PR}{P+R}}$$

or

$$\boxed{F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\,\frac{\mathrm{precision}\cdot\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}}$$

How to evaluate a ranked retrieval system?

When results are ordered, more measures are available.

Common useful measures are:

- The precision-recall graph and the mean average precision
- The ROC graph and the area under it (AUC)

Precision-recall graph

Plotting the points

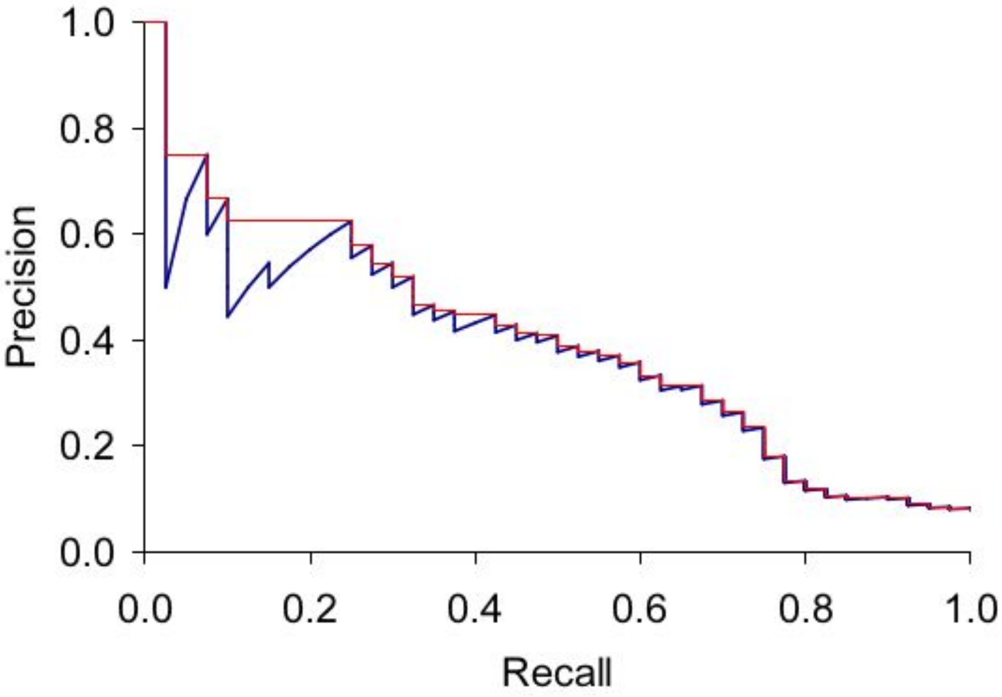
For a given query

For each result

    if the result is relevant

    set x = #tp / #expected

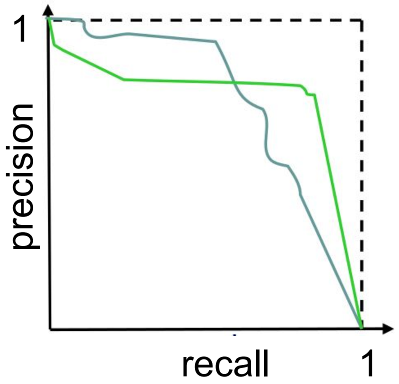
    set y = #tp / #returned



The recall always increases while we scan the result list.

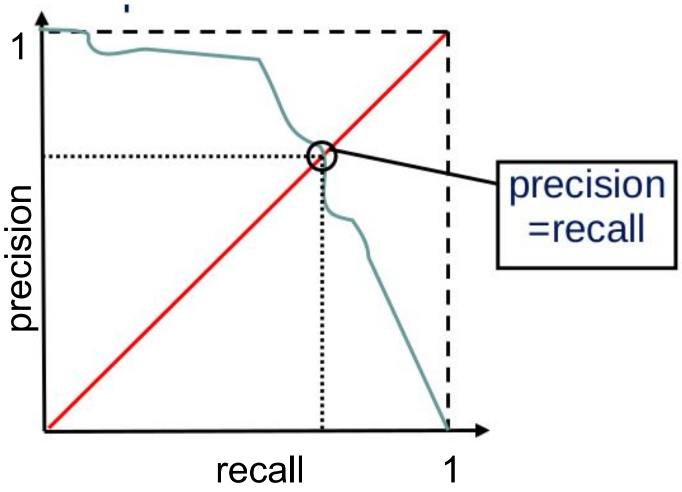
Equal Error Rate and Average Precision

Which one is the best?

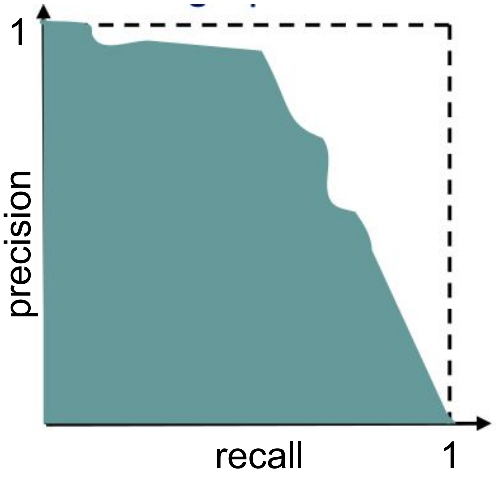


Note: the PR graph does not provide a total order
⇒ need more indicators

Equal Error Rate



Average precision

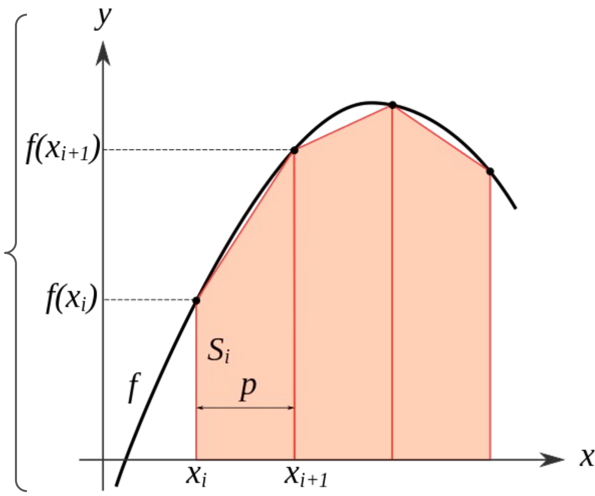


Mean average precision at k — mAP (@k)

Mean of the average precision of several queries,
when considering **k results for each query**

⇒ makes evaluation tractable with very large databases

Computed for each query using the trapezoid technique →



General algorithm:

For each query $\boldsymbol{q_i}$ in the test set with expected results $\boldsymbol{e_i}$:

   Retrieve the list $\boldsymbol{ret_i}$ of $\boldsymbol{k}$ best results

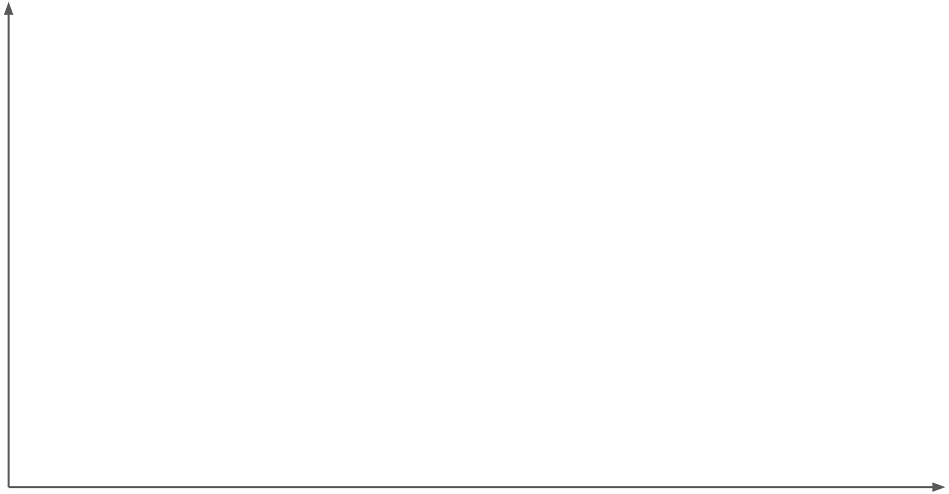
   Compute the AP $\boldsymbol{ap_i}$ given $\boldsymbol{e_i}$ and $\boldsymbol{ret_i}$

Compute the mean AP over all $\boldsymbol{ap_i}$

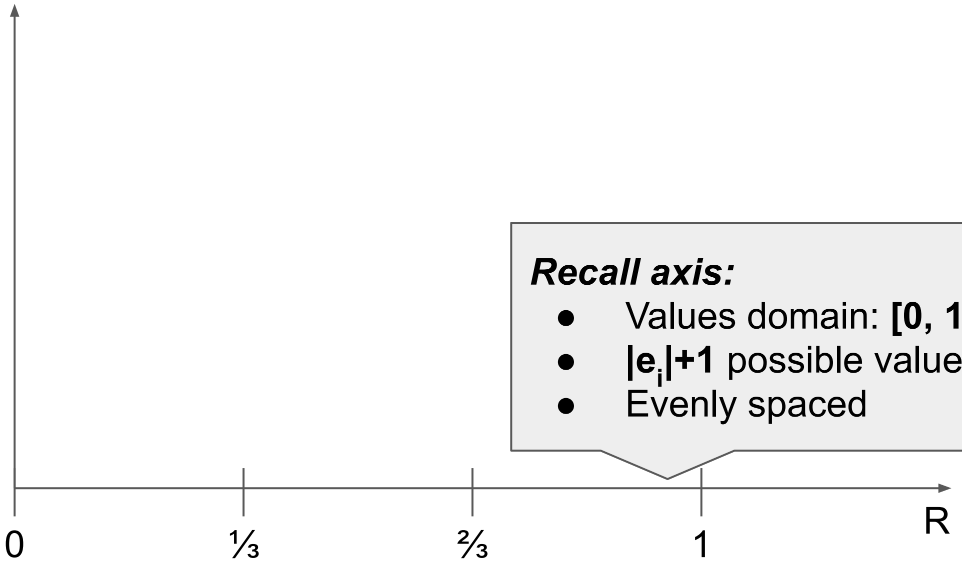
Example: Compute the AP for a given query

For this query and the following results, plot the precision/recall graph and compute the average precision.

Case 1: assume $|\mathbf{e}_i| = \mathbf{3}$



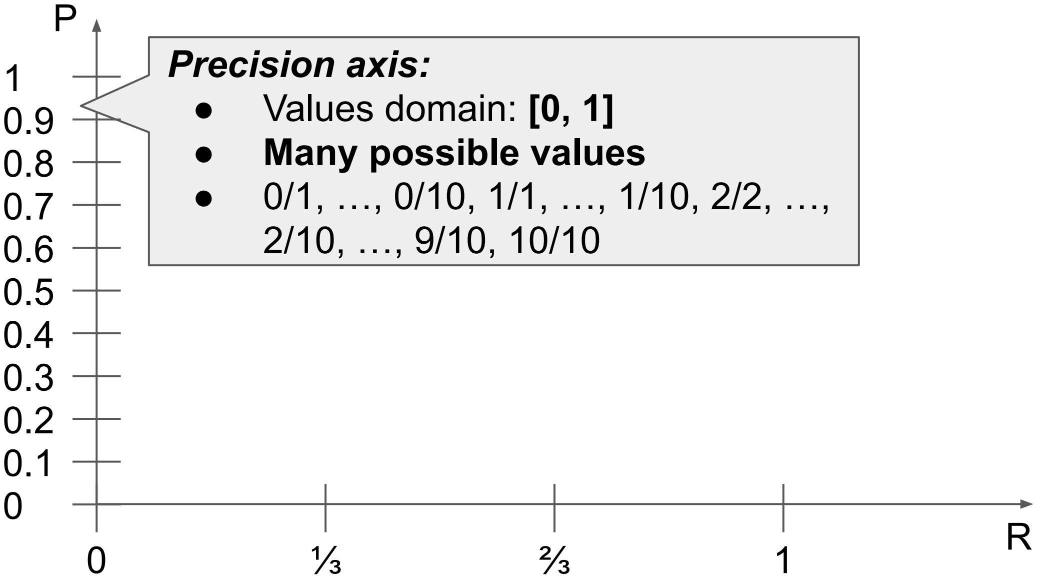
Case 1: assume $|\mathbf{e_i}| = \mathbf{3}$



Recall axis:

- Values domain: **[0, 1]**
- $|\mathbf{e_i}|\mathbf{+1}$ possible values
- Evenly spaced

Case 1: assume **$|e_i| = 3$**

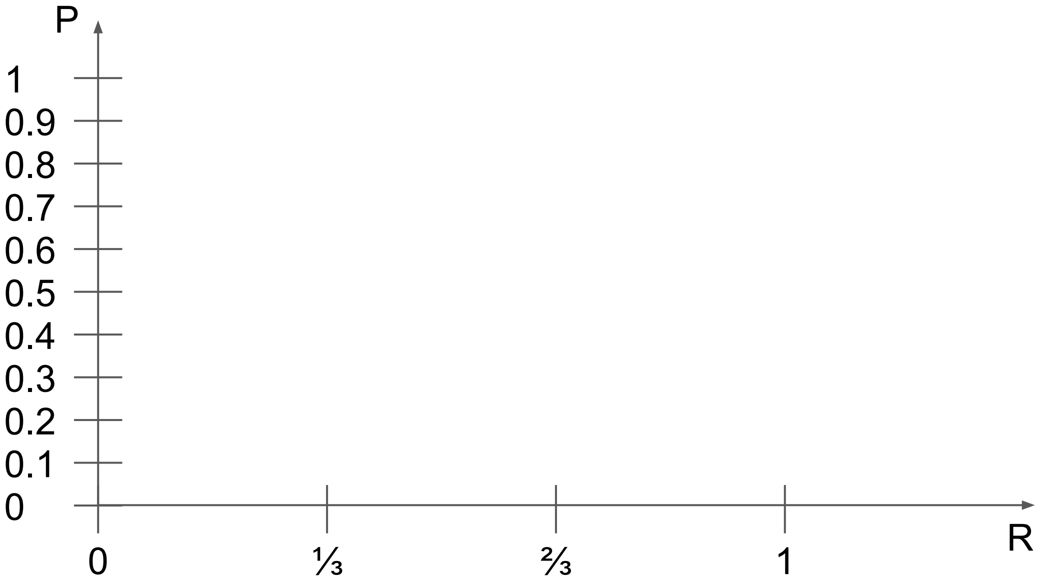


Precision axis:

- Values domain: **[0, 1]**
- **Many possible values**
- 0/1, …, 0/10, 1/1, …, 1/10, 2/2, …, 2/10, …, 9/10, 10/10

Case 1: assume $|\mathbf{e_i}| = \mathbf{3}$

Check the **first** result:
It it **relevant**?



Case 1: assume $|e_i| = 3$

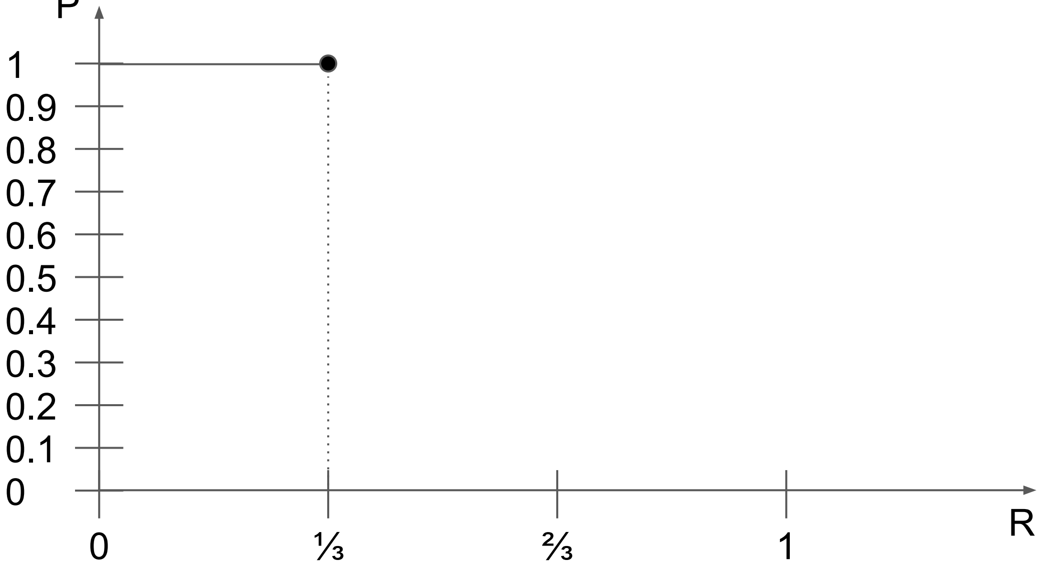
Check the **first** result:

It it **relevant**? **YES**

⇒ **Compute current precision:**

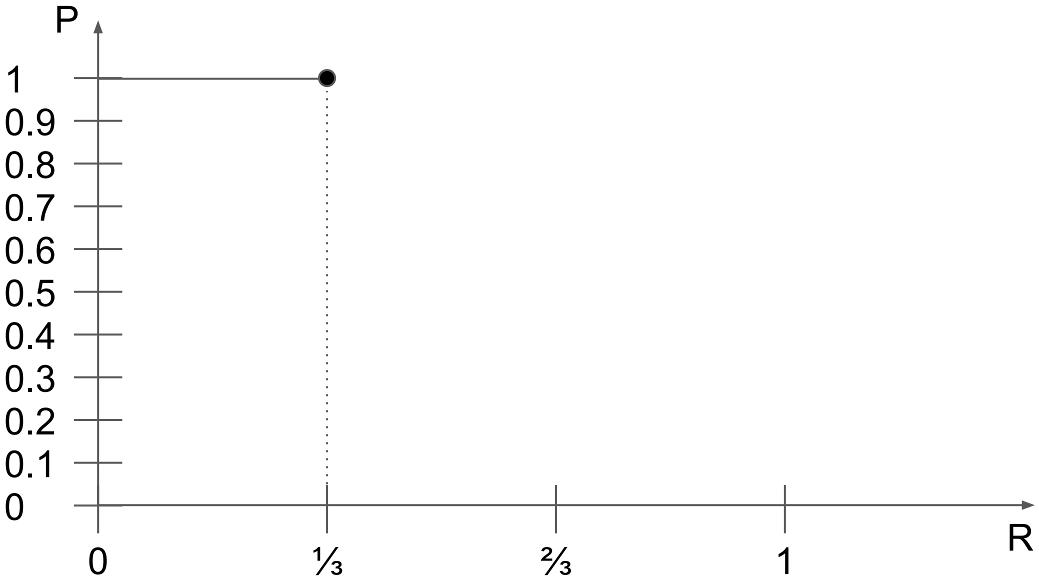
1 relevant / 1 retrieved = 1

⇒ **Recall = 1 relev. / 3 expected = ⅓**



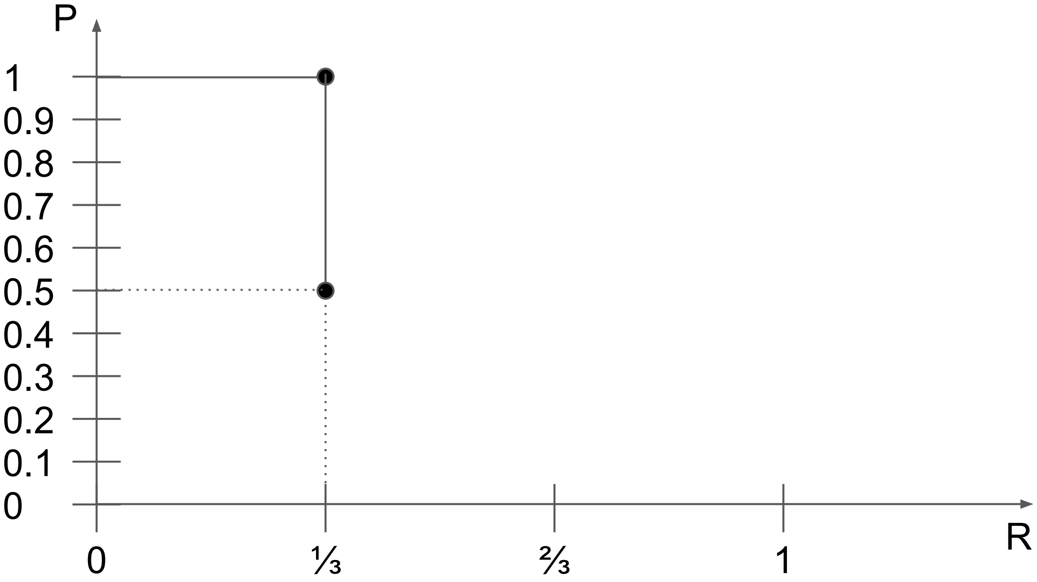
Case 1: assume $|e_i| = 3$

Check the **next** result:
It it **relevant**?



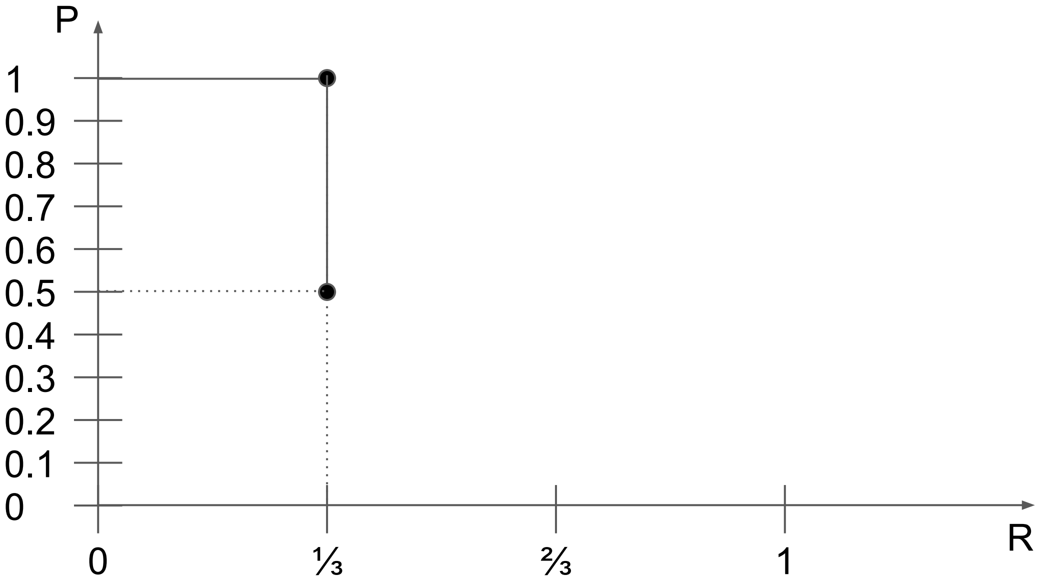
Case 1: assume $|e_i| = 3$

Check the **next** result:
It it **relevant**? **NO**
⇒ **P@2 = 1 relevant / 2 retrieved = ½**
⇒ **R@2 is unchanged**



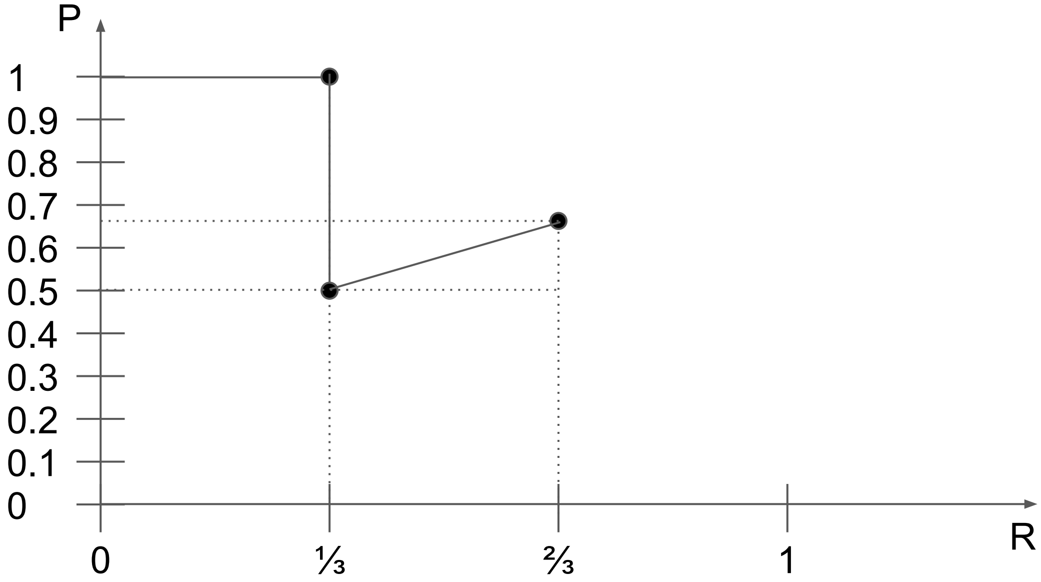
Case 1: assume $|e_i| = 3$

Check the **next** result:
It it **relevant**?



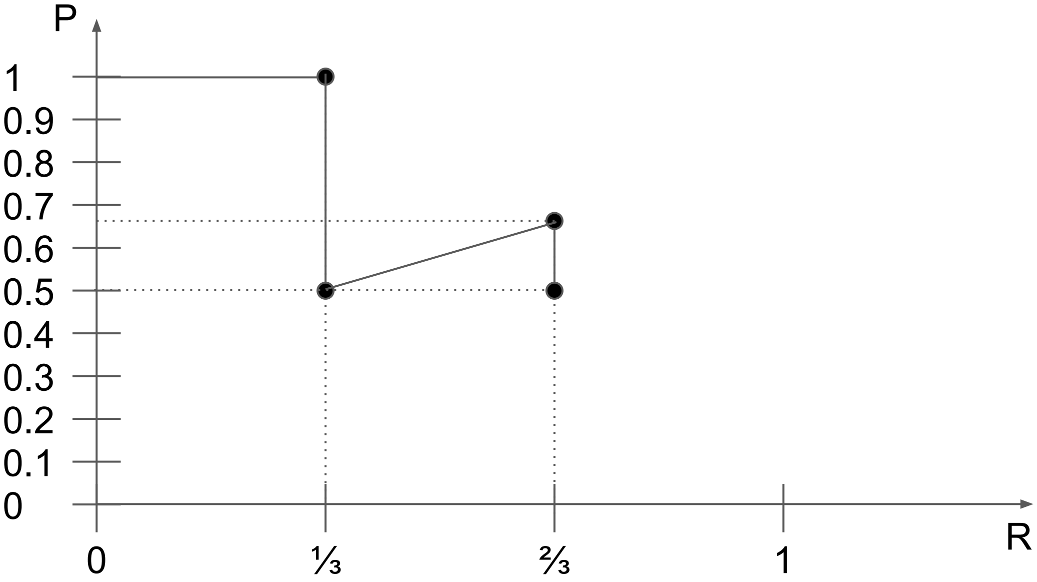
Case 1: assume $|e_i| = 3$

Check the **next** result:
It it **relevant**? YES
⇒ **P@3 = 2 relevant / 3 retrieved = ⅔**
⇒ **Add a point at next recall value (⅔)**



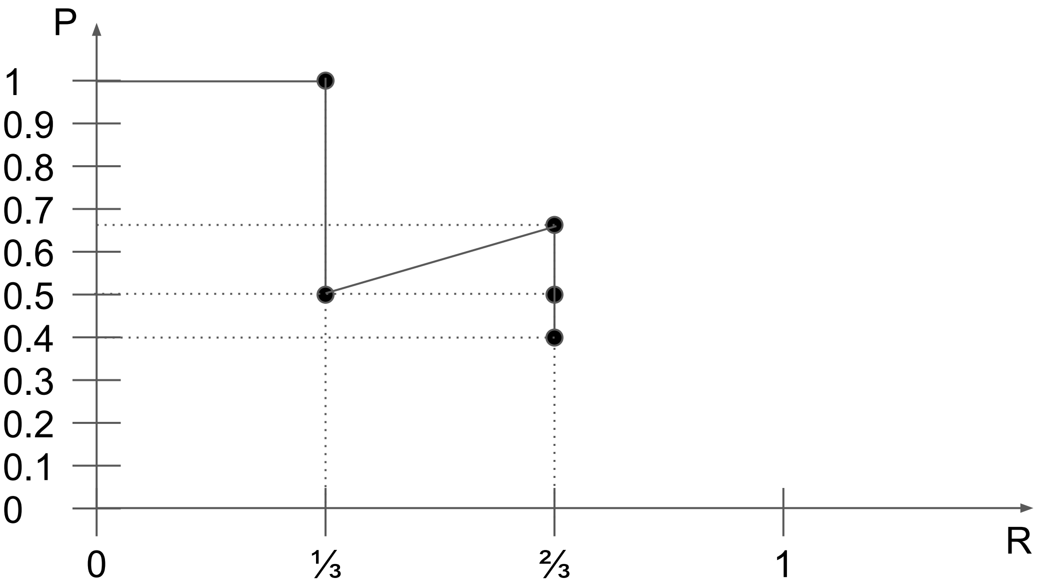
Case 1: assume $|e_i| = 3$

And we keep going…
P@4 = 2/4 = 1/2
R@4 = unchanged



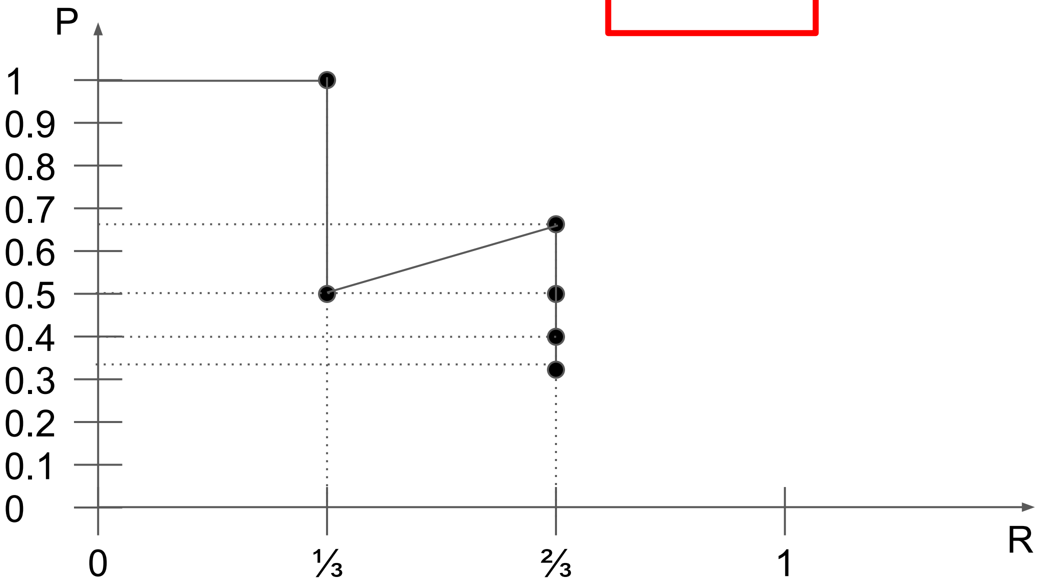
Case 1: assume $|e_i| = 3$

And we keep going…
P@5 = 2/5 = 0.4
R@5 = unchanged



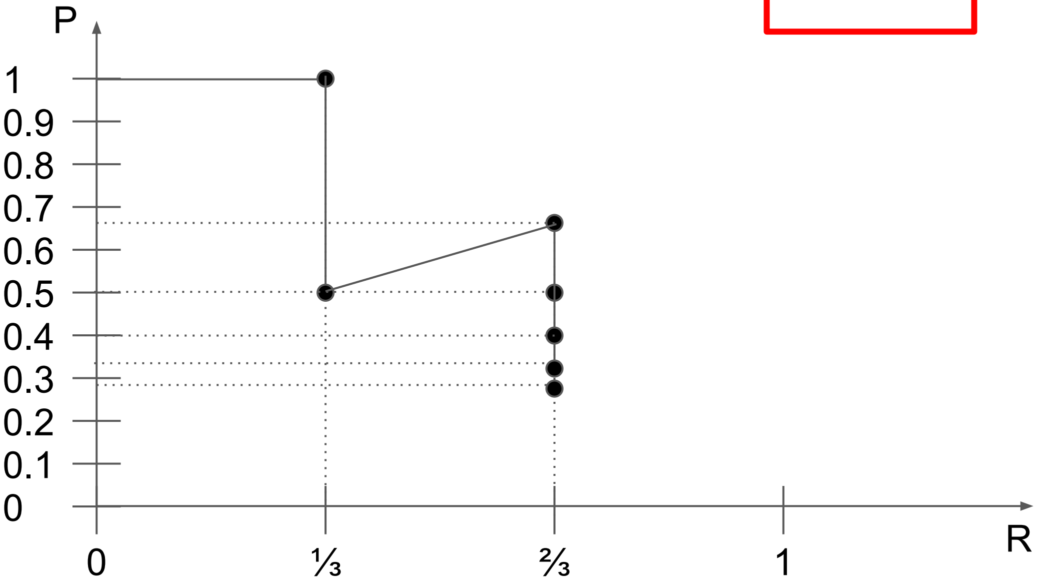
Case 1: assume $|e_i| = 3$

And we keep going…
P@6 = 2/6 = 1/3
R@6 = unchanged



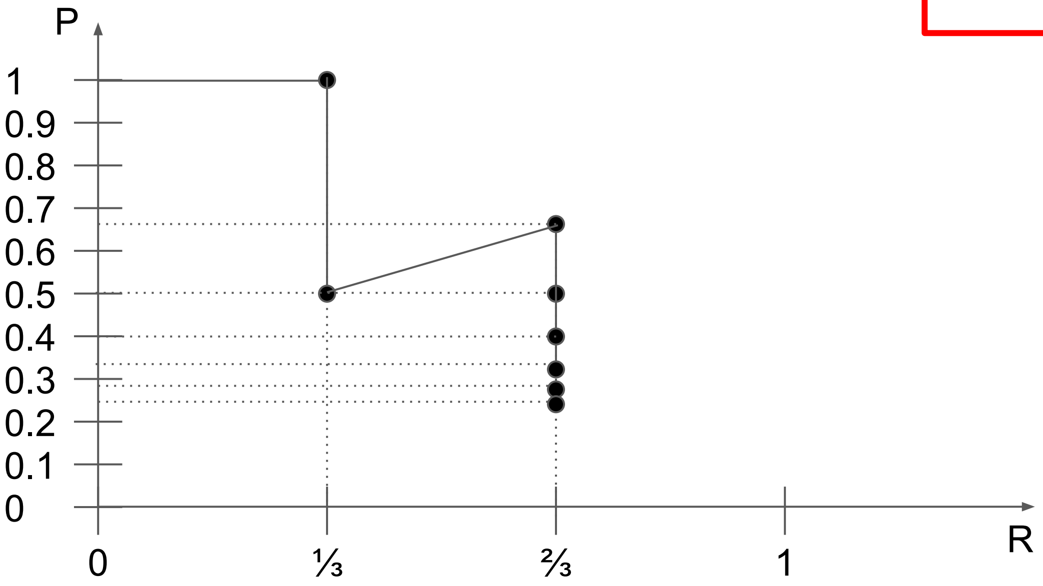
Case 1: assume $|e_i| = 3$

And we keep going…
P@7 = 2/7 = 0.285...
R@7 = unchanged



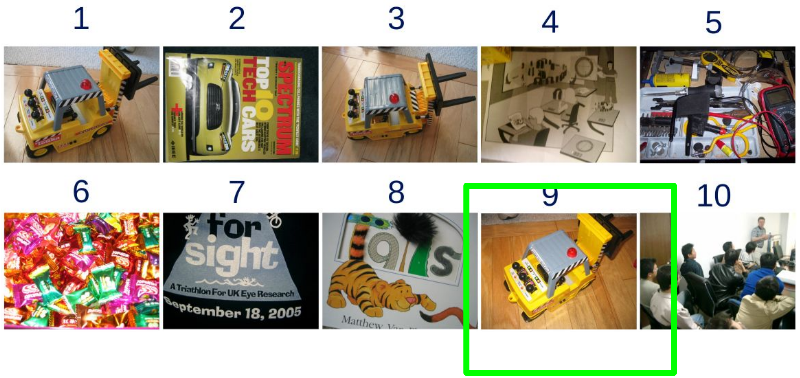
Case 1: assume $|e_i| = 3$

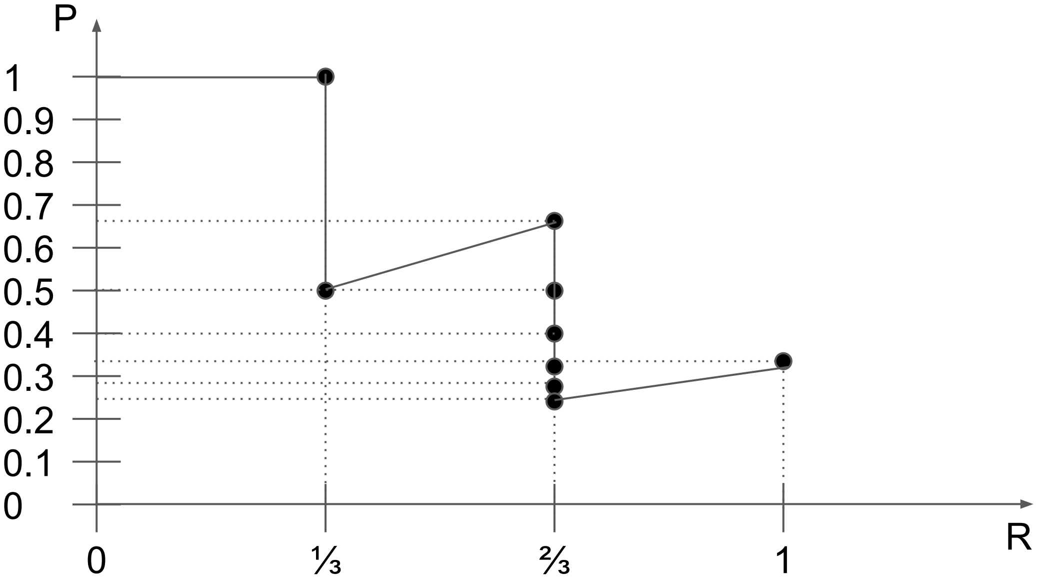
And we keep going…
P@8 = 2/8 = 1/4
R@8 = unchanged



Case 1: assume $|e_i| = 3$

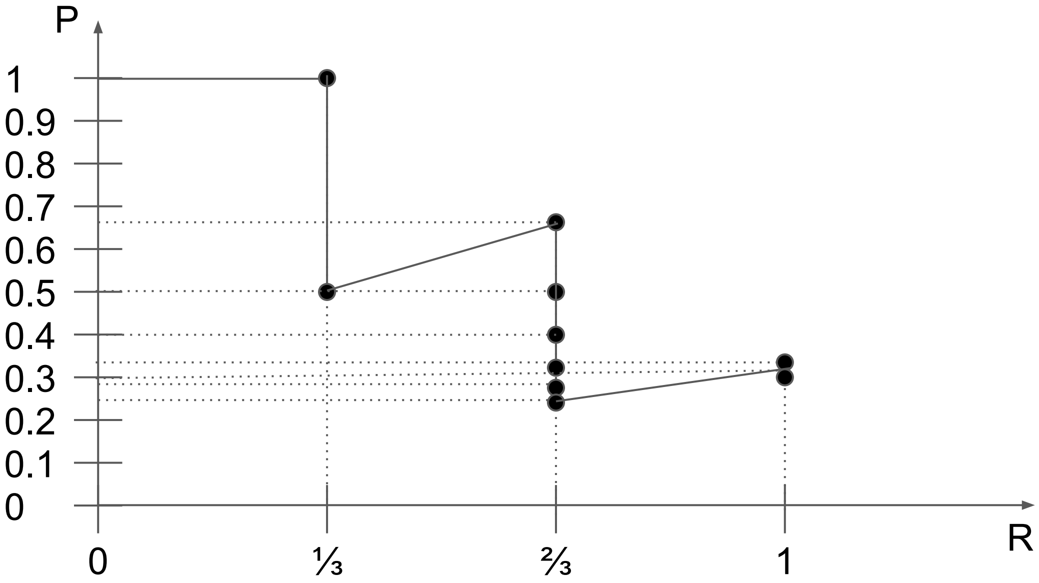
And we keep going…
P@9 = 3/9 = 1/3
R@9 = 3/3 = 1





Case 1: assume $|e_i| = 3$

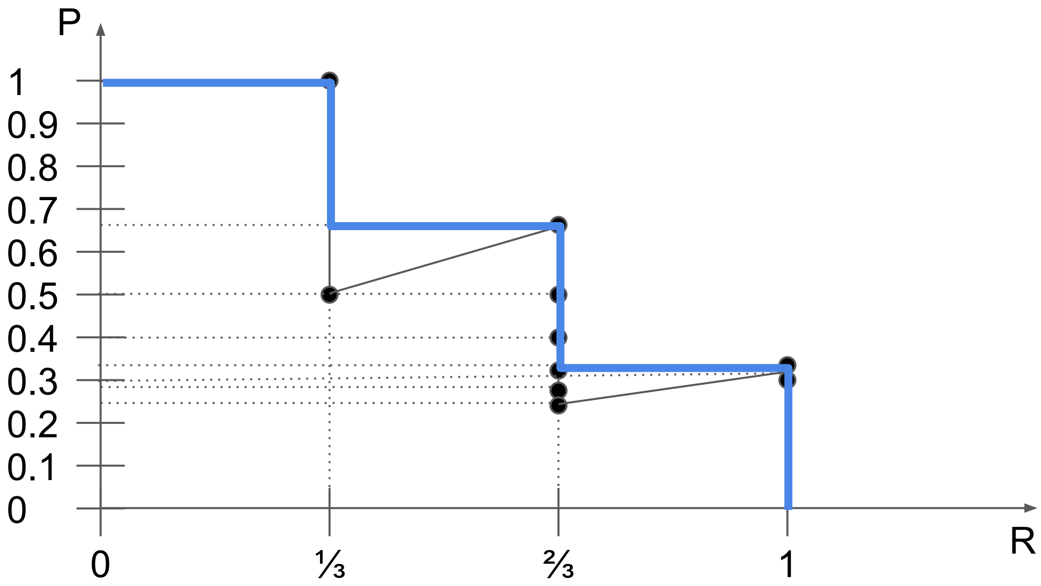
It does not change the AP here…
P@10 = 3/10
R@10 = 3/3 = 1



Case 1: assume $|e_i| = 3$

And we are done!

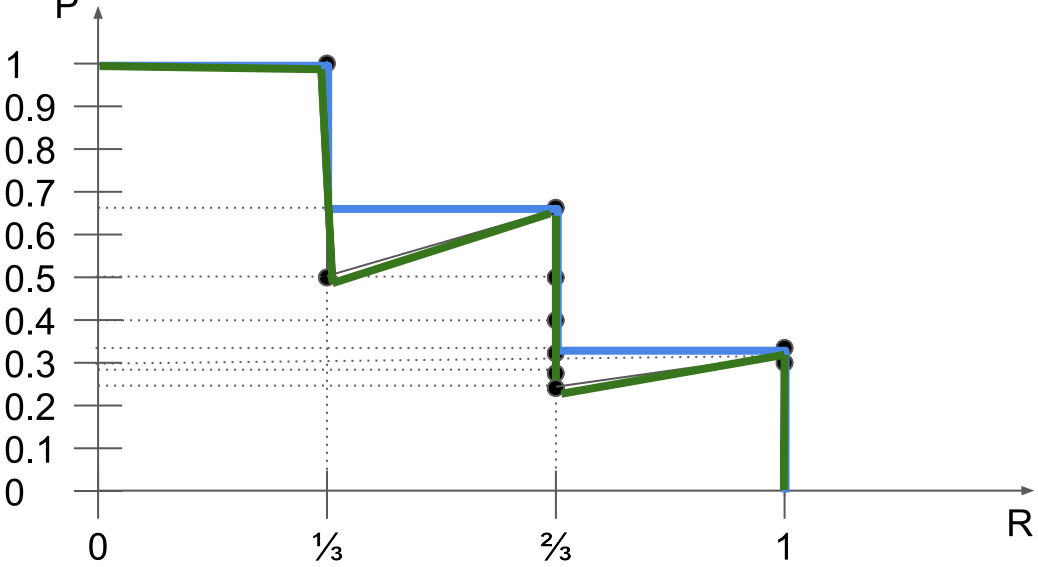
A common approximation is to take **only the upper envelope of the curve…**



Case 1: assume $|e_i| = 3$

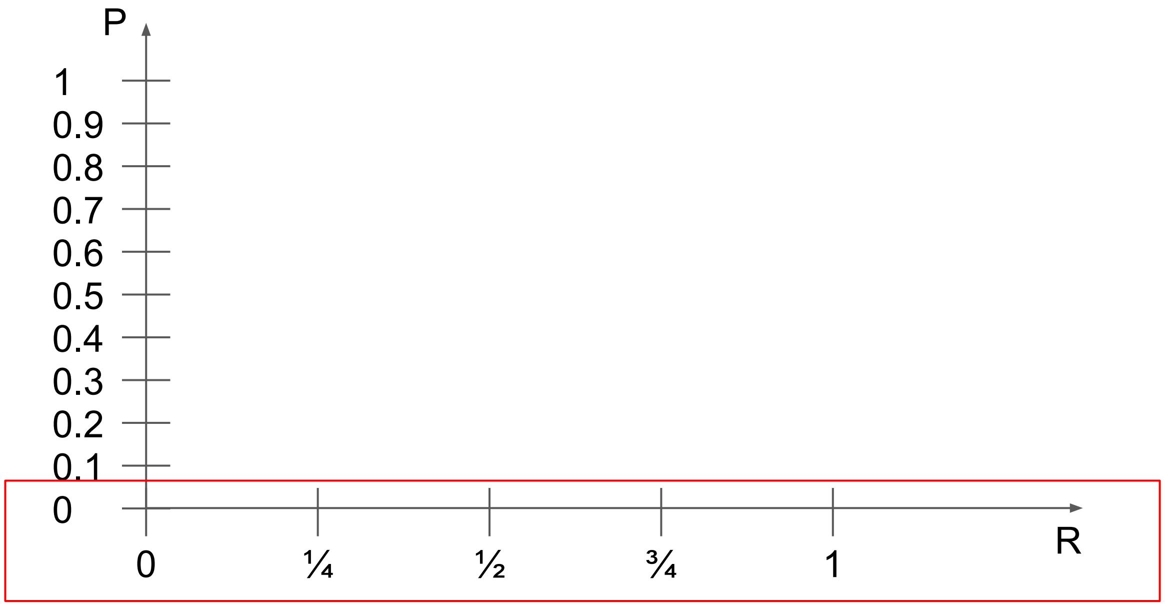
And we are done!

A common approximation is to take
only the upper envelope of the curve…
But **good libraries** go for the **full, exact computation**.



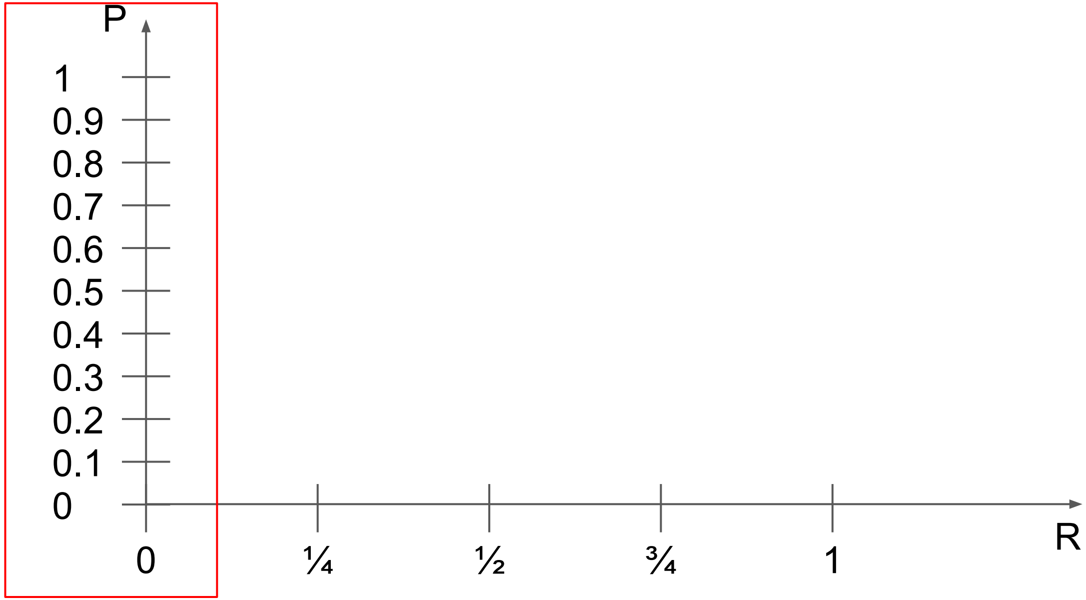
Case 2: what if $|\mathbf{e_i}| = \mathbf{4}$?

1. Adjust R values.



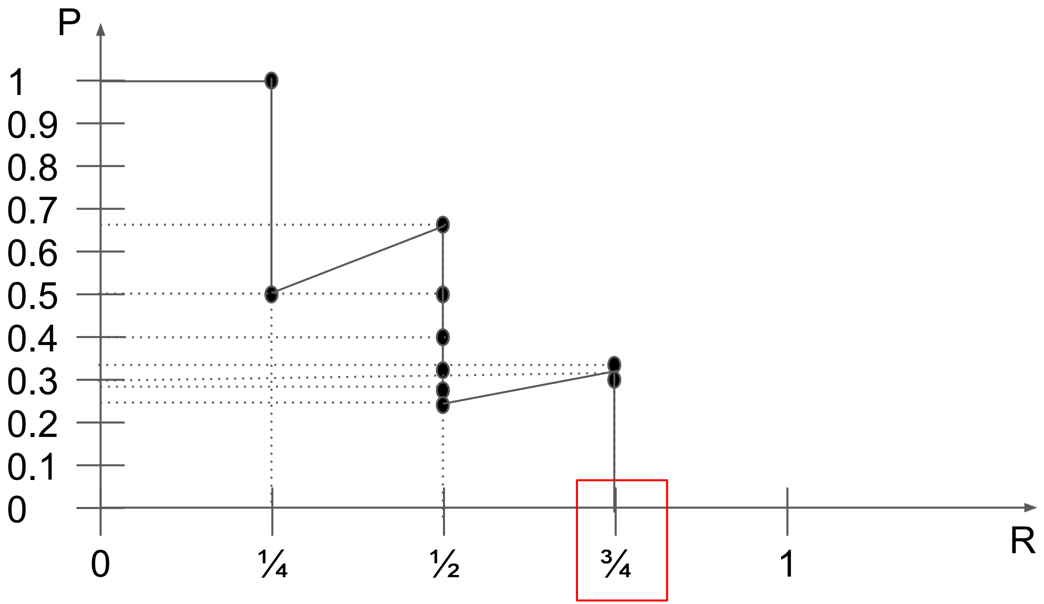
Case 2: what if $|\mathbf{e_i}| = \mathbf{4}$?

1. Adjust R values.
2. P values do not change if **k** does not change.



Case 2: what if **$|e_i| = 4$?**

1. Adjust R values.
2. P values do not change if **k** does not change.
3. Here, it would imply that we did not get all relevant results (very common in practice) ⇒ we stop the curve before the 1



ROC & others

[next lecture, more useful for classification]

Ground truthing issues

Do we have to annotate all images within a dataset for all our test queries?

No! Use **“distractors”**: samples that you know, for sure, not to be relevant to any query.