UFMG UNIVERSIDADE*FEDERAL
DE*MINAS*GERAIS

Information Retrieval

Offline Evaluation

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Ranking evaluation

Lots of alternative solutions

- Which one to choose?
- How to improve upon them?

Evaluation enables an informed choice

- Rigor of science
- Efficiency of practice

Evaluation methodology

Feedback

- Implicit
- Explicit

Mode

- Retrospective
- Prospective

	retrospective	***prospective***
implicit	counterfactual evaluation	online evaluation
explicit	offline evaluation	

Test collection-based evaluation

Three core components

- A corpus of documents
- A set of users’ queries
- A map of users’ relevance assessments

TREC topic example

```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Credit: NIST

Evaluation metrics

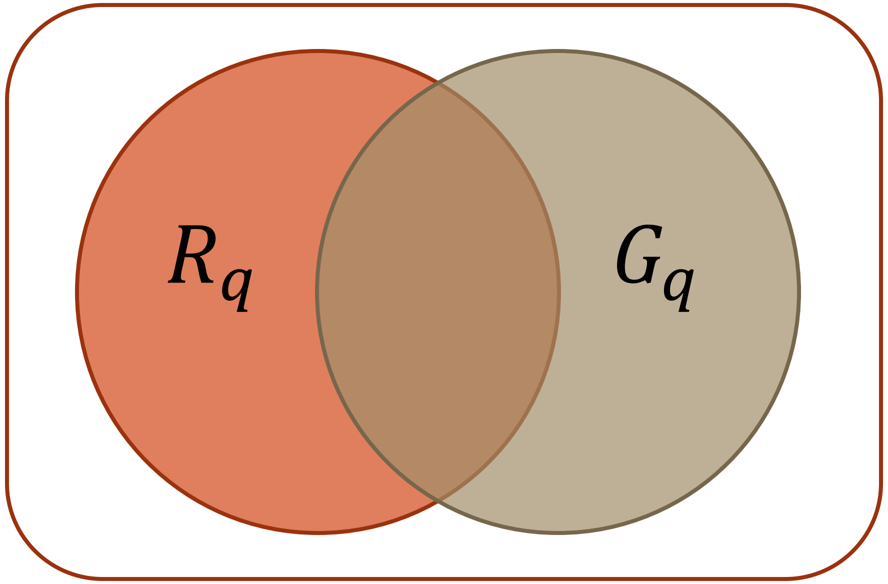
General form: $\Delta(R_q, G_q)$

- ◦ R_q: ranking produced by model f for query q
- ◦ G_q: ground-truth produced for query q

Metrics should be chosen according to the task

Precision and recall

Given a query q



R_q: retrieved documents
G_q: relevant documents

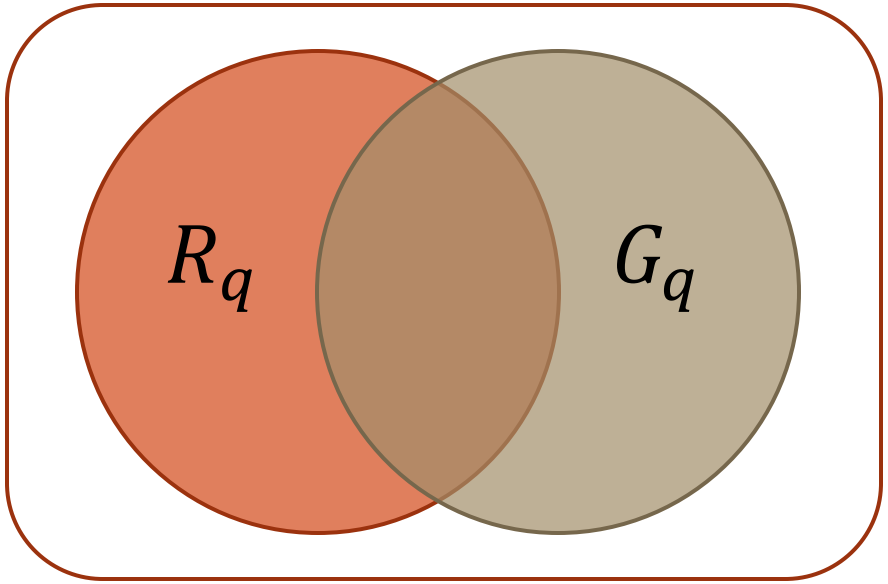
Precision

- Percentage of retrieved documents that are relevant

$$\mathrm{Prec}\left(R_q, G_q\right) = \frac{|R_q \cap G_q|}{\left|R_q\right|}$$

Precision and recall

Given a query q



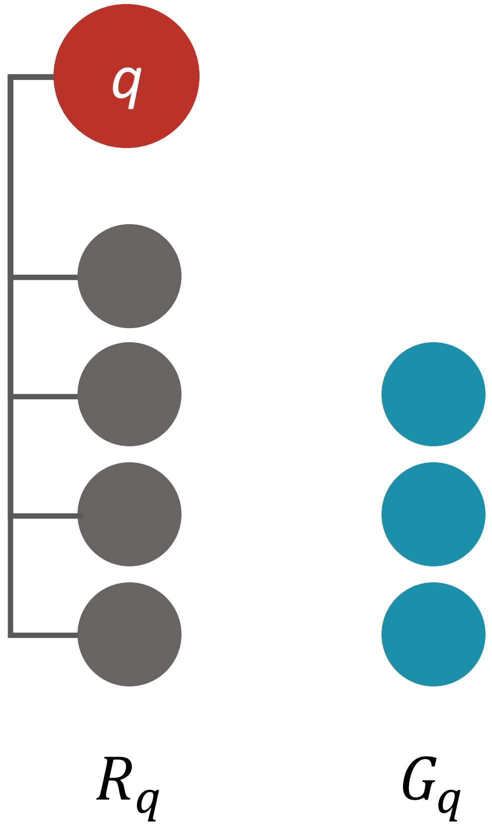
R_q: retrieved documents
G_q: relevant documents

Recall

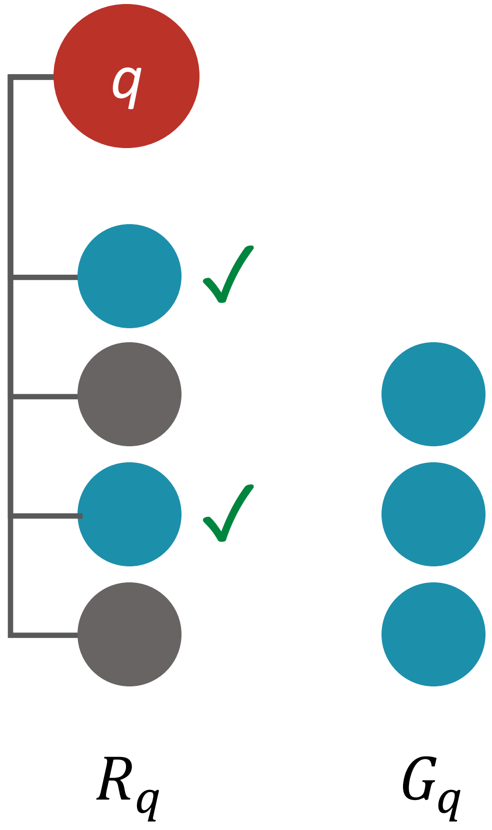
- Percentage of relevant documents that are retrieved

$$\mathrm{Rec}(R_q, G_q) = \frac{|R_q \cap G_q|}{|G_q|}$$

Precision and recall



Precision and recall



Precision

- $\mathrm{Prec}(R_q, G_q) = \frac{|R_q \cap G_q|}{|R_q|} = \frac{2}{4} = 0.50$

Recall

- $\mathrm{Rec}(R_q, G_q) = \frac{|R_q \cap G_q|}{|G_q|} = \frac{2}{3} = 0.67$

Precision and recall

Precision is about having mostly useful stuff in the ranking

- Not wasting the user's time

Key assumption

- There is more useful stuff than the user wants to examine

Recall is about not missing useful stuff in the ranking

- Not making a bad oversight

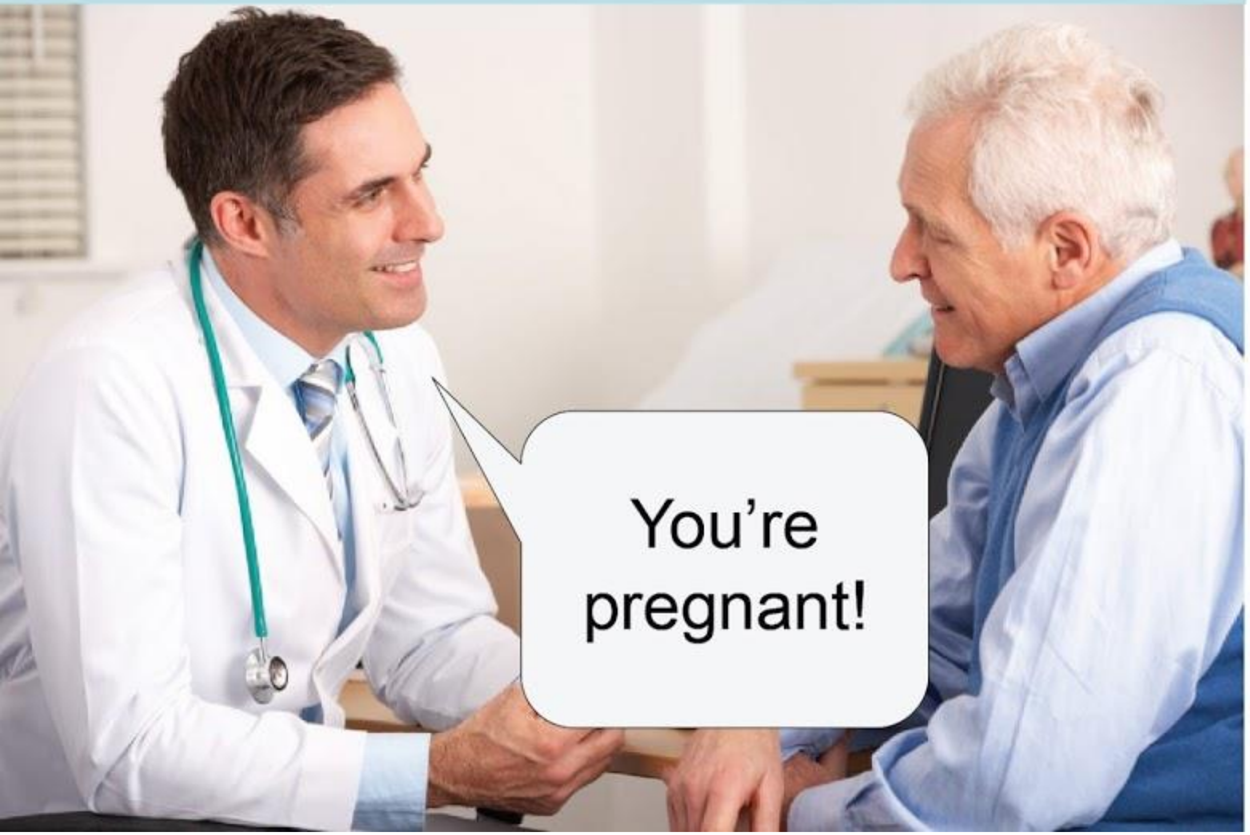
Key assumption

- The user has time to filter through ranked results

We can also combine both

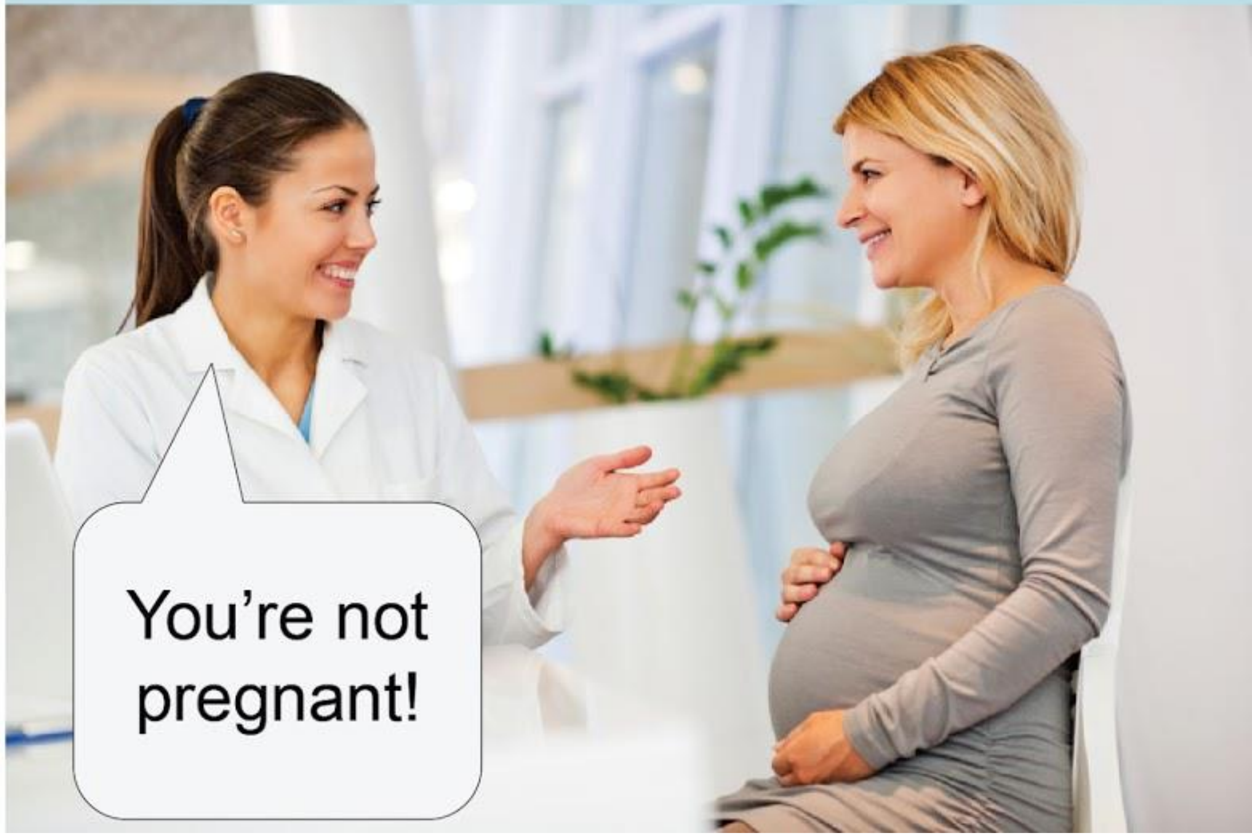
$$\mathrm{F1}(R,G) = \frac{2\,\mathrm{Prec}(R,G)\,\mathrm{Rec}(R,G)}{\mathrm{Prec}(R,G) + \mathrm{Rec}(R,G)}$$

Type I Error



(false positive)

Type II Error



(false negative)

Classification errors

Type I error: probability of retrieving non-relevants

- $\mathrm{FallOut}\left(R_q, G_q\right) = \frac{|R_q \cap \bar{G}_q|}{\left|\bar{G}_q\right|}$

Type II error: probability of missing relevants

- $\mathrm{MissRate}\left(R_q, G_q\right) = 1 - \mathrm{Rec}(R_q, G_q)$

Beyond decision support

Modern document corpora are huge

- User may not be willing to inspect large sets

Consider top-5 rankings

- Ranker #1: + + + + −
- Ranker #2: − + + + +

Ranker #2 **misplaces** a highly visible item

Evaluation cutoffs

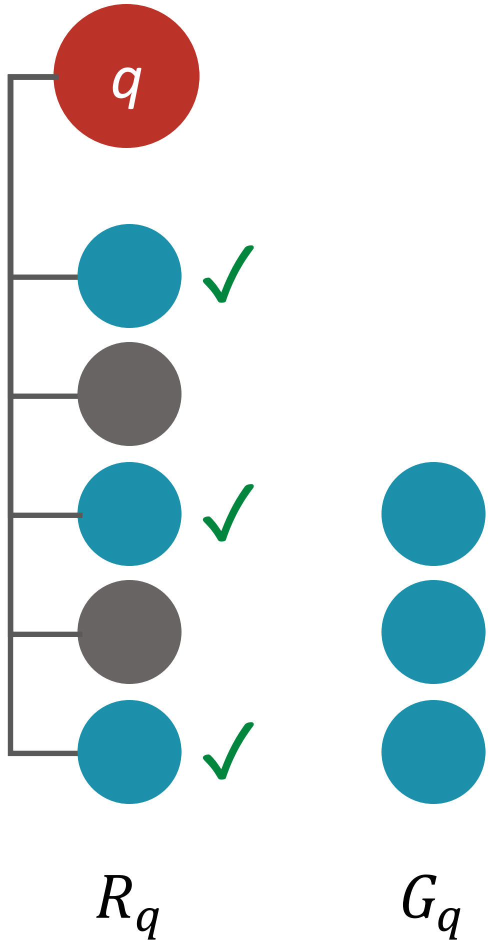
Calculate precision and recall at fixed rank positions

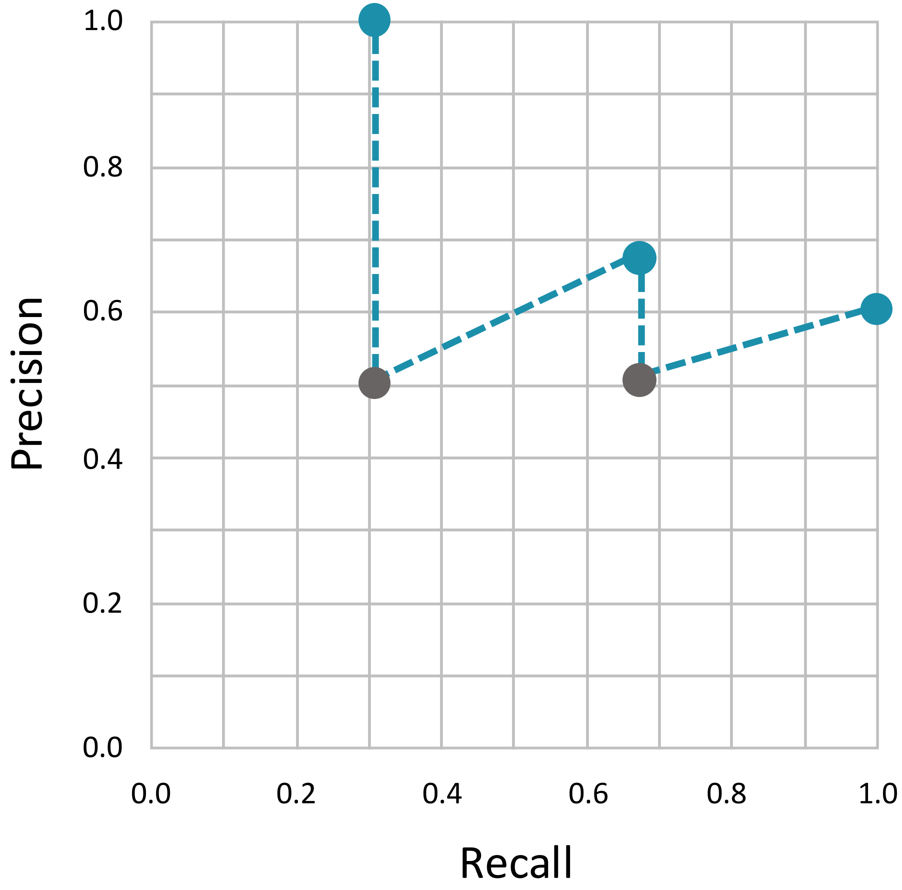
- e.g., Prec@10, Rec@10

Calculate precision at standard recall levels

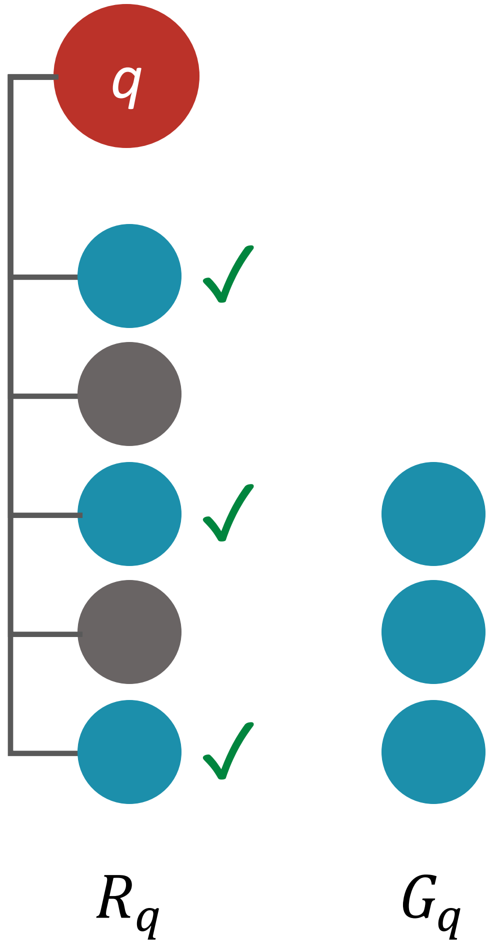
- e.g., Prec@Rec=30%

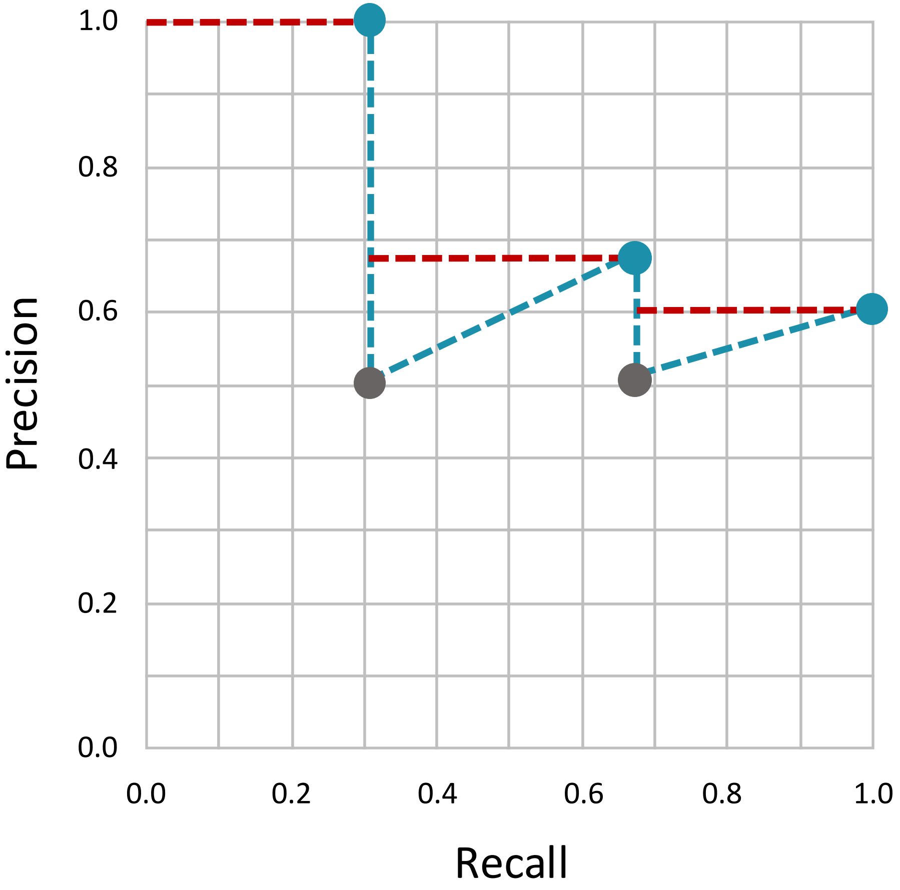
Precision vs recall graph



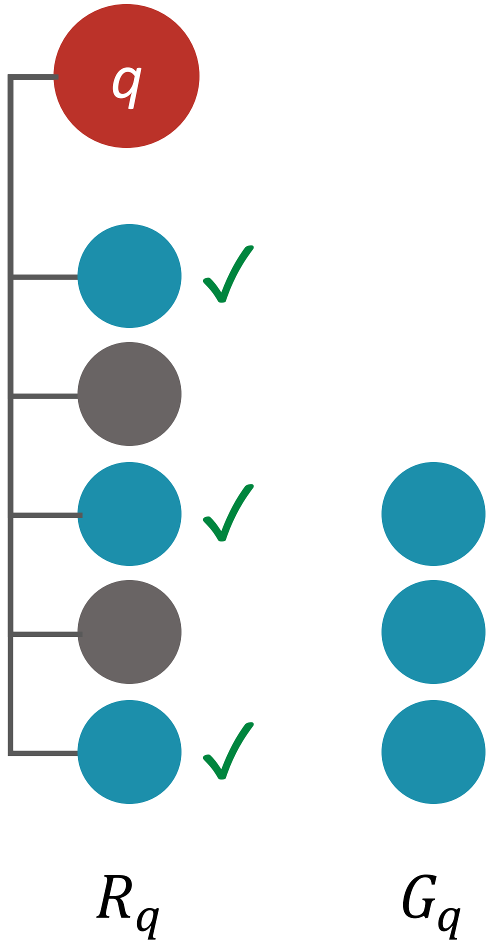


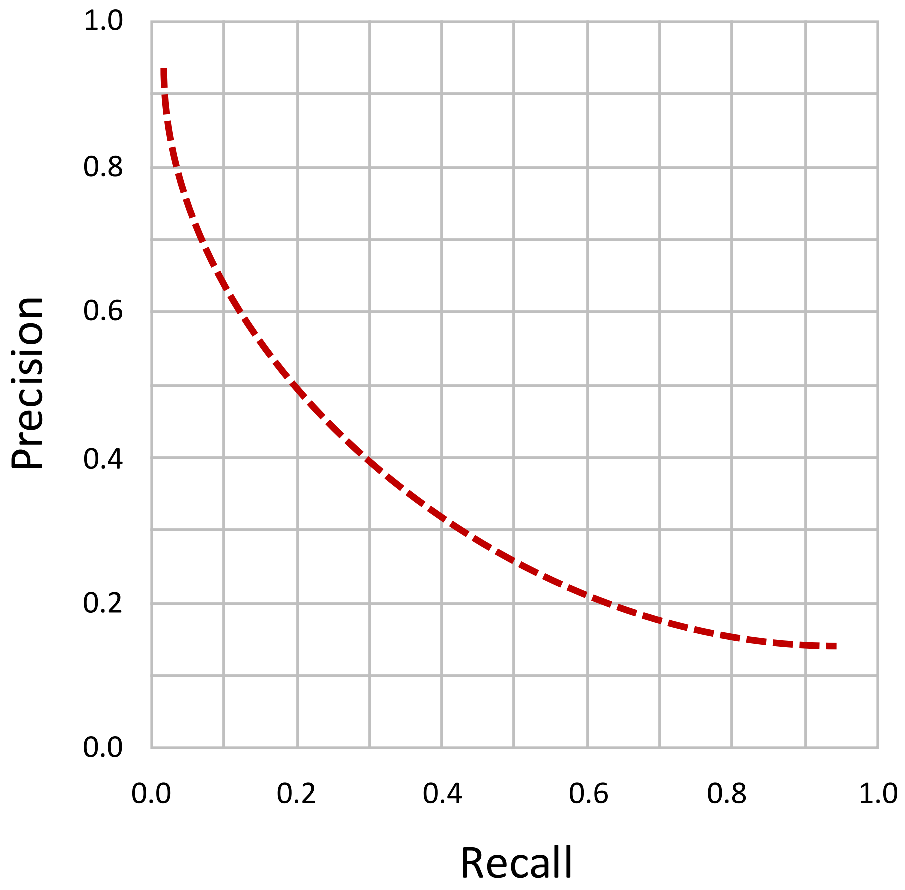
Precision vs recall graph (interpolated)



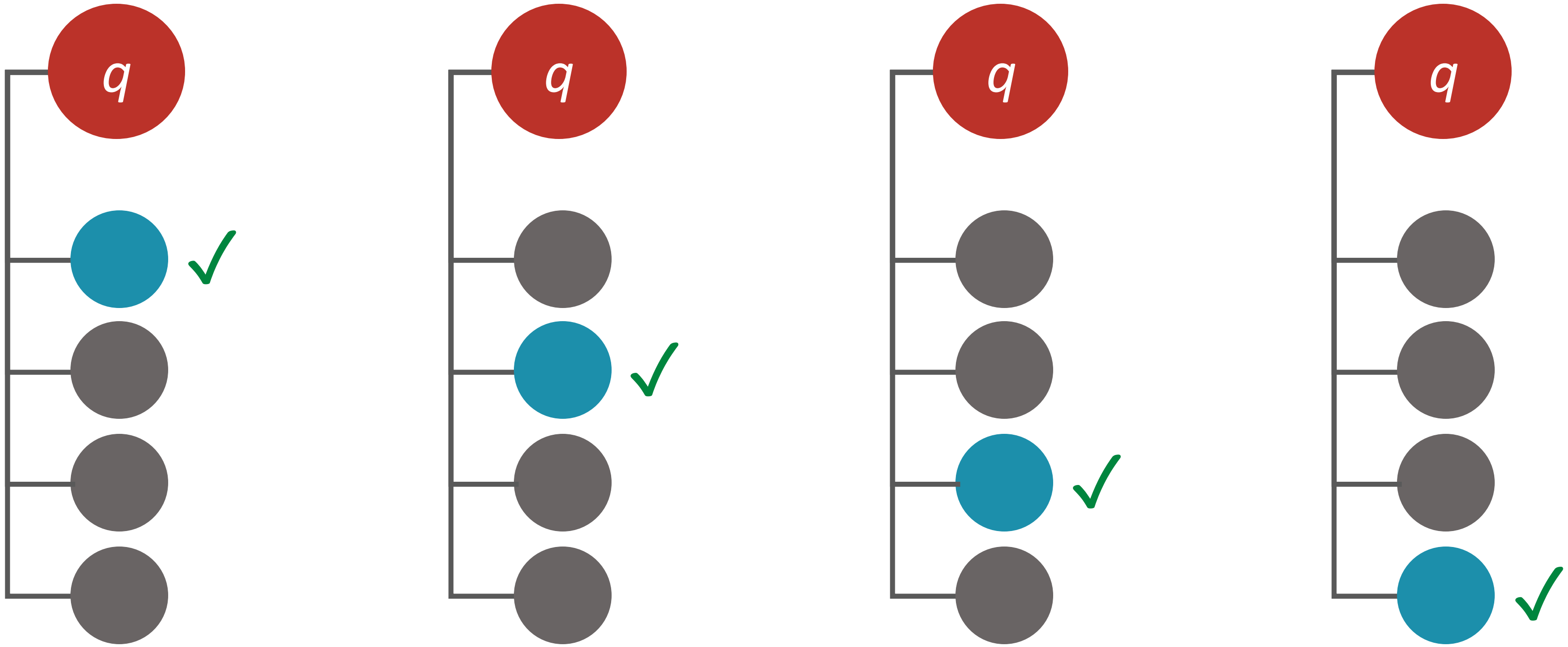


Precision vs recall graph (averaged)





Position blindness



These have exactly the same Prec@4 (0.25)

- Are they equally good?

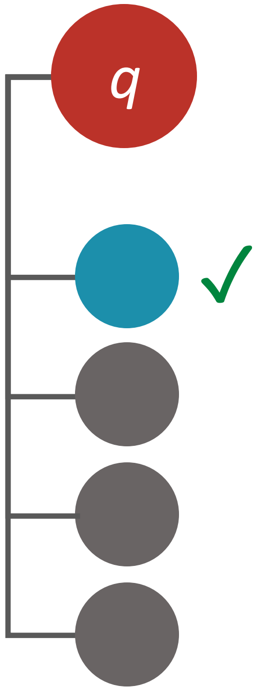
Position-aware metrics

Why ranking?

- Place documents in order of preference

Key assumption

- Users will inspect retrieved documents from top to bottom (or left to right)

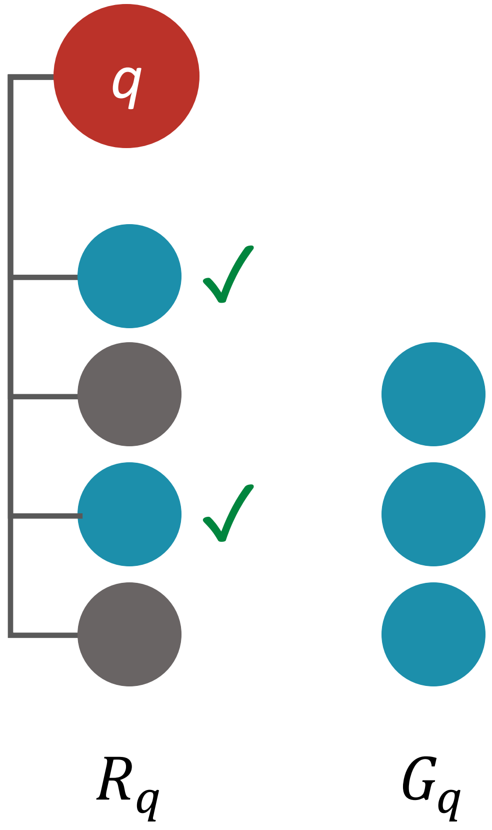


Average precision (AP)

Simple idea: averaging precision values at the ranking positions where relevant documents were found

- $\mathrm{AP}\left(R_q, G_q\right) = \frac{1}{|G_q|} \sum_{i=1}^{k} 1\left(g_{q d_i} > 0\right) \mathrm{Prec@}i$

Average precision (AP)



Average precision

- $\mathrm{AP}(R_q, G_q) = \frac{1}{|G_q|} \sum_{i=1}^{k} 1(g_{qd_i} > 0)\ \mathrm{Prec@}i$

$$= \frac{1}{3}(\mathrm{Prec@1} + \mathrm{Prec@3})$$

$$= \frac{1}{3}\left(\frac{1}{1} + \frac{2}{3}\right)$$

$$= \frac{5}{9} = 0.55$$

Average precision (AP)

Simple idea: averaging precision values at the ranking positions where relevant items were found

- $\mathrm{AP}\left(R_q, G_q\right) = \frac{1}{|G_q|} \sum_{i=1}^{k} 1\left(g_{q d_i} > 0\right) \mathrm{Prec@}i$

In practice, take the mean (MAP) across queries

- $\mathrm{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{AP}(R_q, G_q)$

Reciprocal rank (RR)

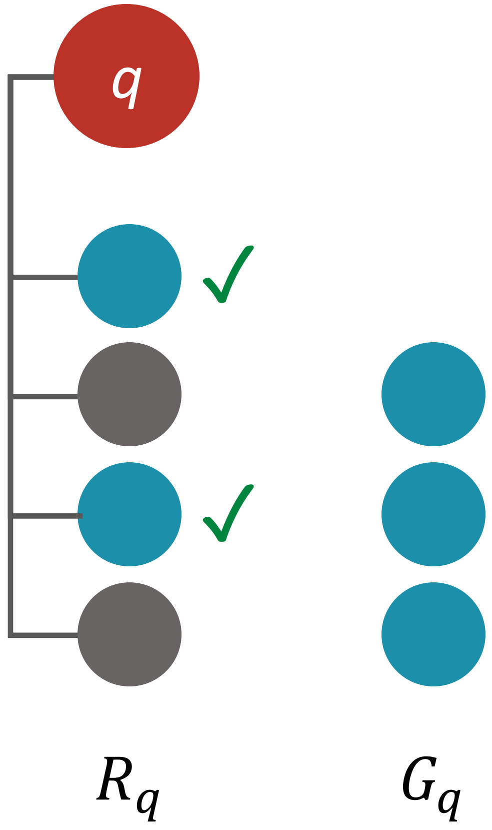
Measures how deep the user has to dig in the ranking to find the first relevant document

- $\mathrm{RR}(R_q, G_q) = 1/i$ (i: position of the first relevant)

Mean reciprocal rank averages across queries

- $\mathrm{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{RR}(R_q, G_q)$

Reciprocal rank (RR)

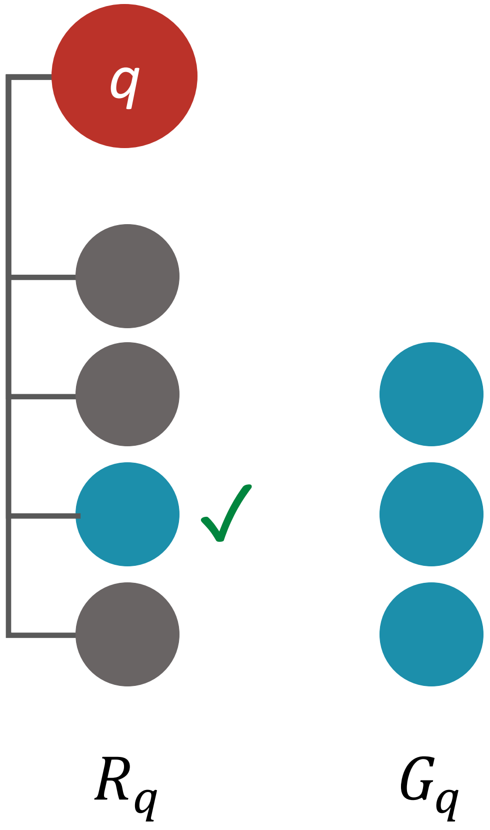


Reciprocal rank

- $\mathrm{RR}\left(R_q, G_q\right) = \frac{1}{i}$

$$= \frac{1}{1} = 1.00$$

Reciprocal rank (RR)



Reciprocal rank

- $\mathrm{RR}\left(R_q, G_q\right) = \frac{1}{i}$

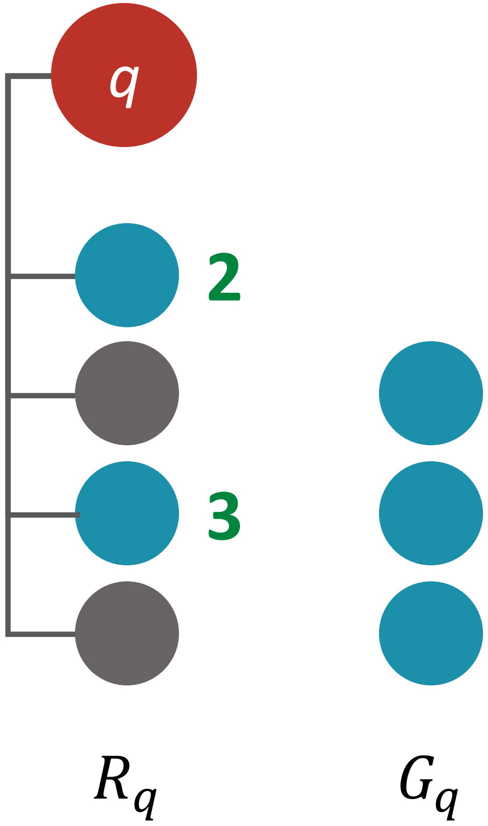
  $= \frac{1}{3} = 0.33$

Discounted cumulative gain (DCG)

Measure utility of document at each position

- $\mathrm{DCG}(R_q, G_q) = \sum_{i=1}^{k} \frac{g_{qd_i}}{\log_2(i+1)}$ linear gain (e.g., in a graded scale) position-based discount

Discounted cumulative gain (DCG)



Discounted cumulative gain

- $\mathrm{DCG}\left(R_q, G_q\right) = \sum_{i=1}^{k} \frac{g_{qd_i}}{\log_2(i+1)}$

$$= \frac{2}{\log_2(1+1)} + \frac{3}{\log_2(3+1)}$$

$$= \frac{2}{1} + \frac{3}{2}$$

$$= 2 + 1.5$$

$$= 3.5$$

Discounted cumulative gain (DCG)

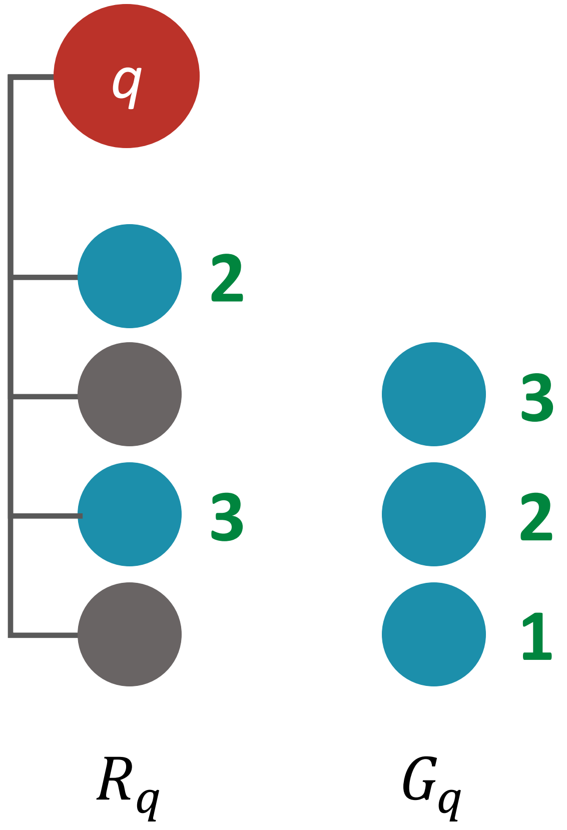
Measure utility of item at each position

- $\mathrm{DCG}\left(R_q, G_q\right) = \sum_{i=1}^{k} \frac{g_{qd_i}}{\log_2(i+1)}$ linear gain (e.g., in a graded scale)
position-based discount

Could also emphasize larger gains

- $\mathrm{DCG}\left(R_q, G_q\right) = \sum_{i=1}^{k} \frac{2^{g_{qd_i}}-1}{\log_2(i+1)}$ exponential gain
position-based discount

Ideal discounted cumulative gain (iDCG)



Ideal discounted cumulative gain

- $\mathrm{iDCG}\left(R_q, G_q\right) = \sum_{i=1}^{k} \frac{g_{qd_i}}{\log_2(i+1)}$

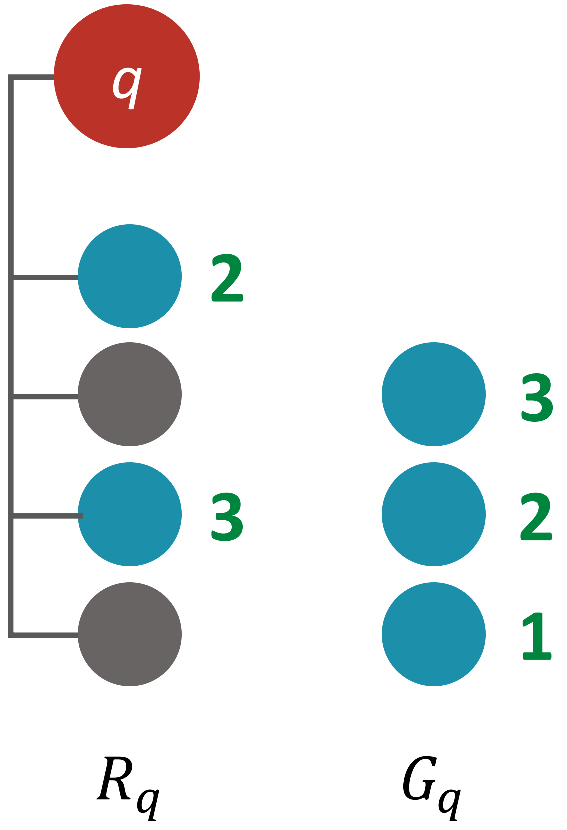
$$= \frac{3}{\log_2(1+1)} + \frac{2}{\log_2(2+1)} + \frac{1}{\log_2(3+1)}$$

$$= \frac{3}{1} + \frac{2}{1.58} + \frac{3}{2}$$

$$= 3 + 1.26 + 0.5$$

$$= 4.76$$

Norm. discounted cumulative gain (nDCG)



Normalized discounted cumulative gain

- $\mathrm{nDCG}(R_q, G_q) = \dfrac{\mathrm{DCG}(R_q,G_q)}{\mathrm{iDCG}(R_q,G_q)}$

$$= \frac{3.5}{4.76}$$

$$= 0.74$$

Average nDCG

In practice, nDCG is averaged across all test queries

- $\mathrm{nDCG} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\mathrm{DCG}(R_q, G_q)}{\mathrm{iDCG}(R_q, G_q)}$

Significance tests

Given the results from a number of queries, how can we conclude that model B is better than model A?

- A significance test enables us to reject the null hypothesis H_0 (no difference) in favor of the alternative hypothesis H_1 (B is better than A)

Paired testing

1. Compute the effectiveness of models A and B
2. Compute the *difference* in effectiveness of A and B
3. Compute a test statistic t for the distribution in (2)

Query	A	B	$B - A$
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

t-test

Parametric assumption: difference between effectiveness is a sample from a normal distribution

- H_0: mean of differences is zero

Test statistic $t = \frac{\overline{B-A}}{\sigma_{B-A}}\sqrt{n}$

Paired t-test

1. Compute the effectiveness of models A and B
2. Compute the *difference* in effectiveness of A and B
3. Compute a test statistic t for the distribution in (2)

$$t = \frac{\overline{B-A}}{\sigma_{B-A}}\sqrt{n}$$

$$= \frac{21.4}{29.1}\sqrt{10}$$

$$= 2.33$$

Query	A	B	$B - A$
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

$$\overline{B - A} = 21.4$$

$$\sigma_{B-A} = 29.1$$

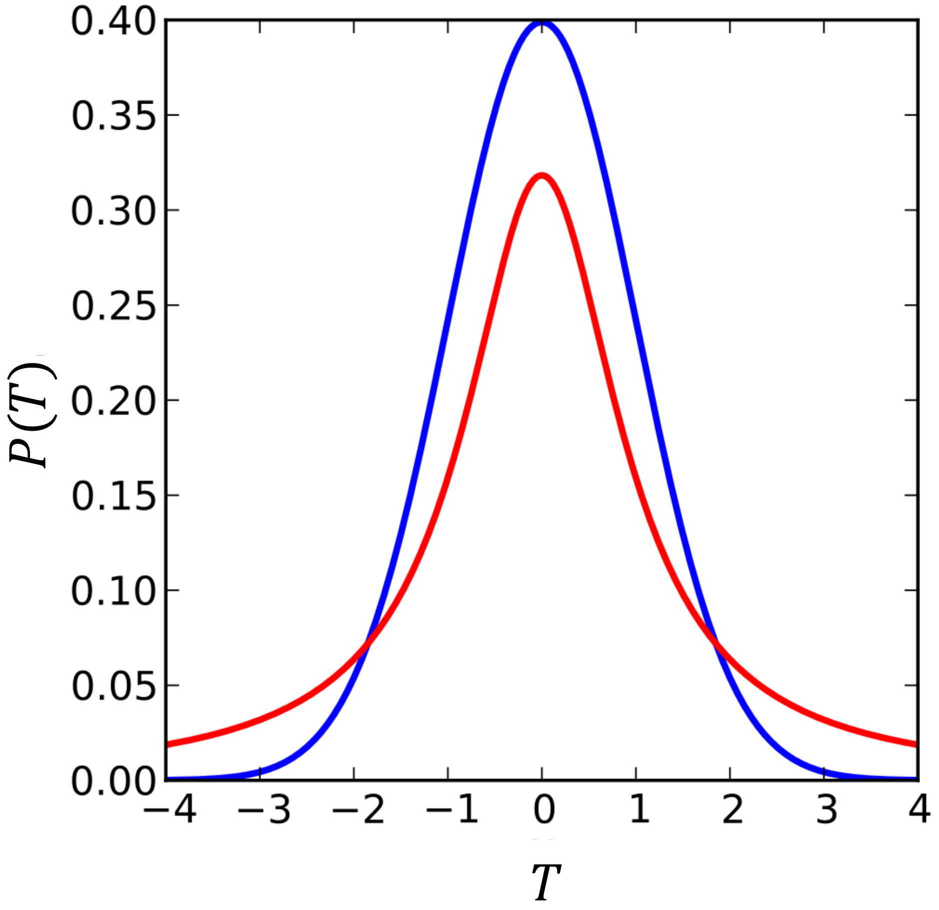
$$n = 10$$

Paired t-test

1. Compute the effectiveness of models A and B
2. Compute the *difference* in effectiveness of A and B
3. Compute a test statistic t for the distribution in (2)
4. Compute the probability p of t under H$_0$

Query	A	B	$B - A$
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

t-distribution ($\nu = 1$)

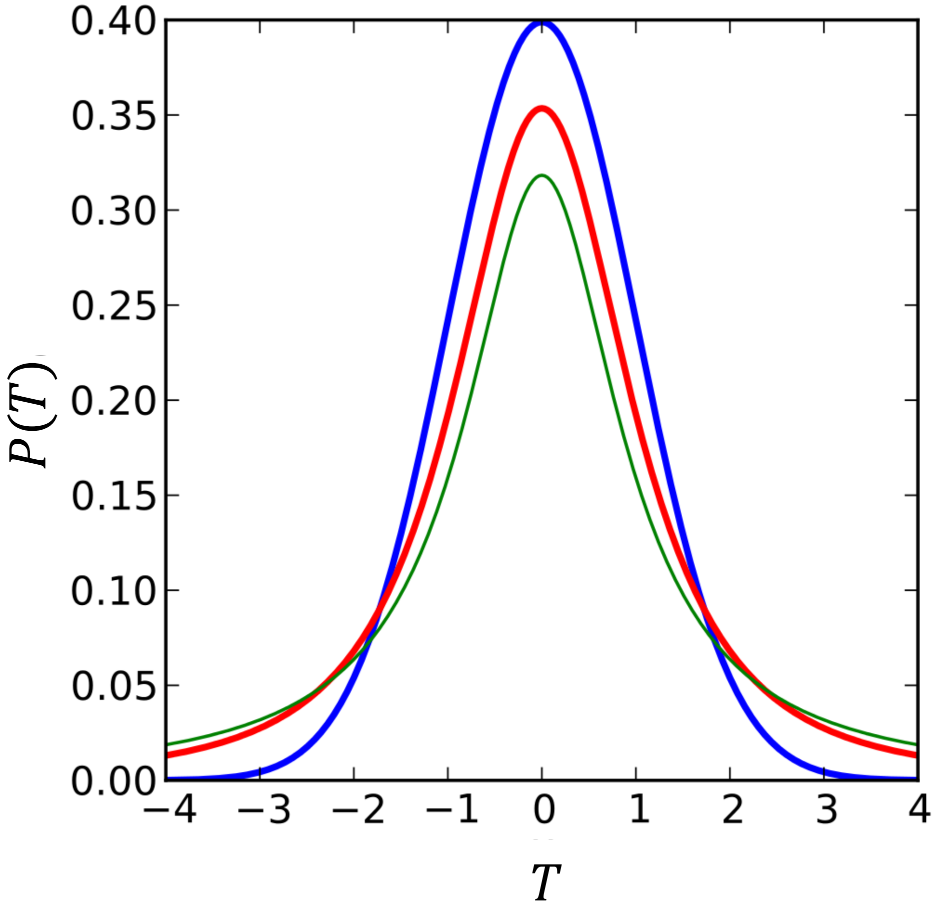


Distribution for the possible values of a test statistic assuming the null hypothesis

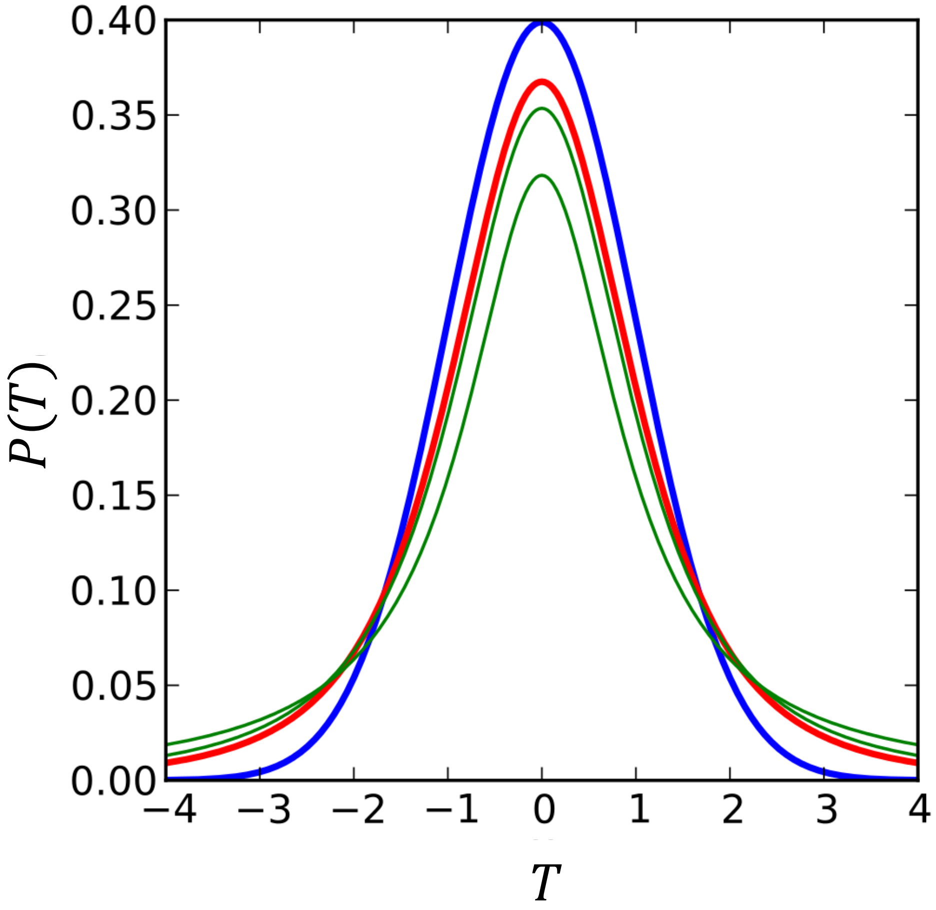
normal distribution

t-distribution

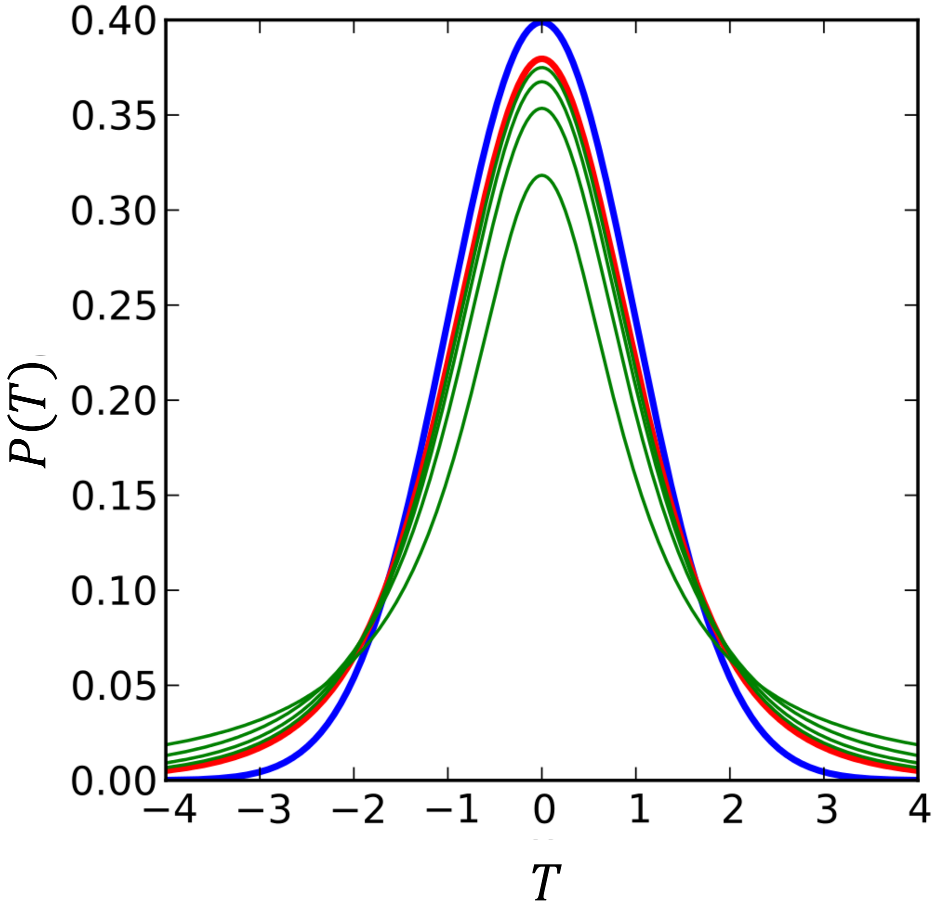
t-distribution ($\nu = 2$)



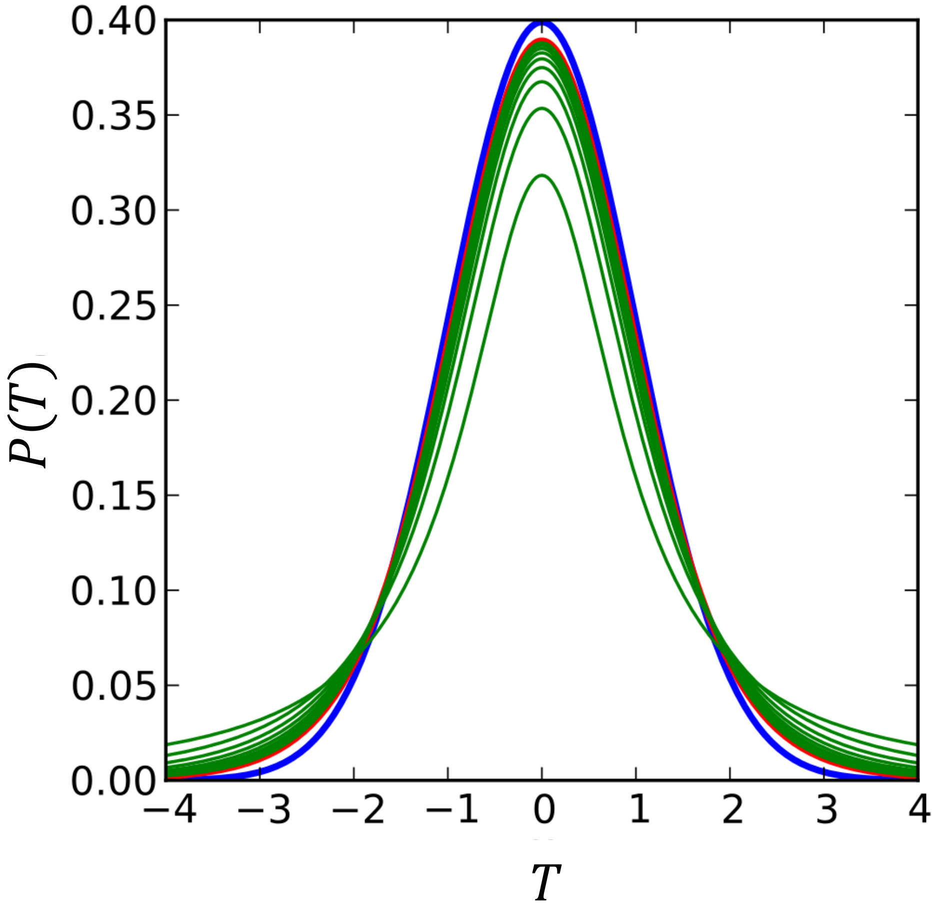
t-distribution $(\nu = 3)$



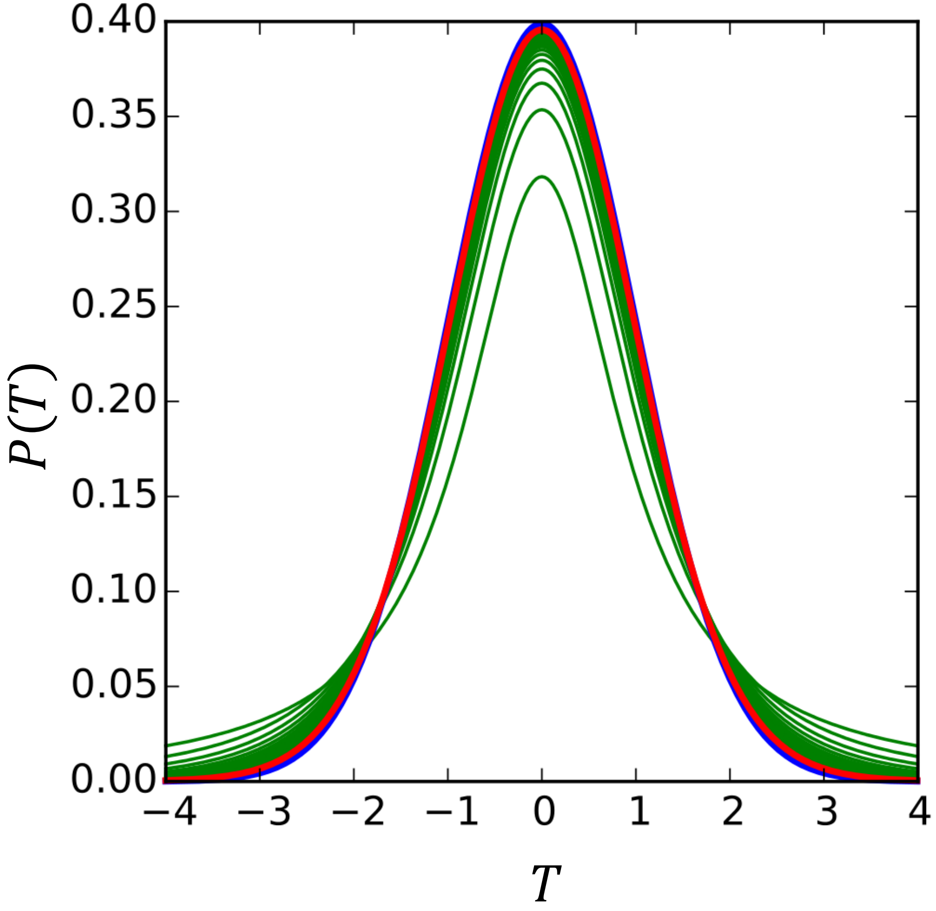
t-distribution ($\nu = 5$)



t-distribution ($\nu = 10$)



t-distribution ($\nu = 30$)



One-sided vs. two-sided tests

One-sided: p-value computed from one tail

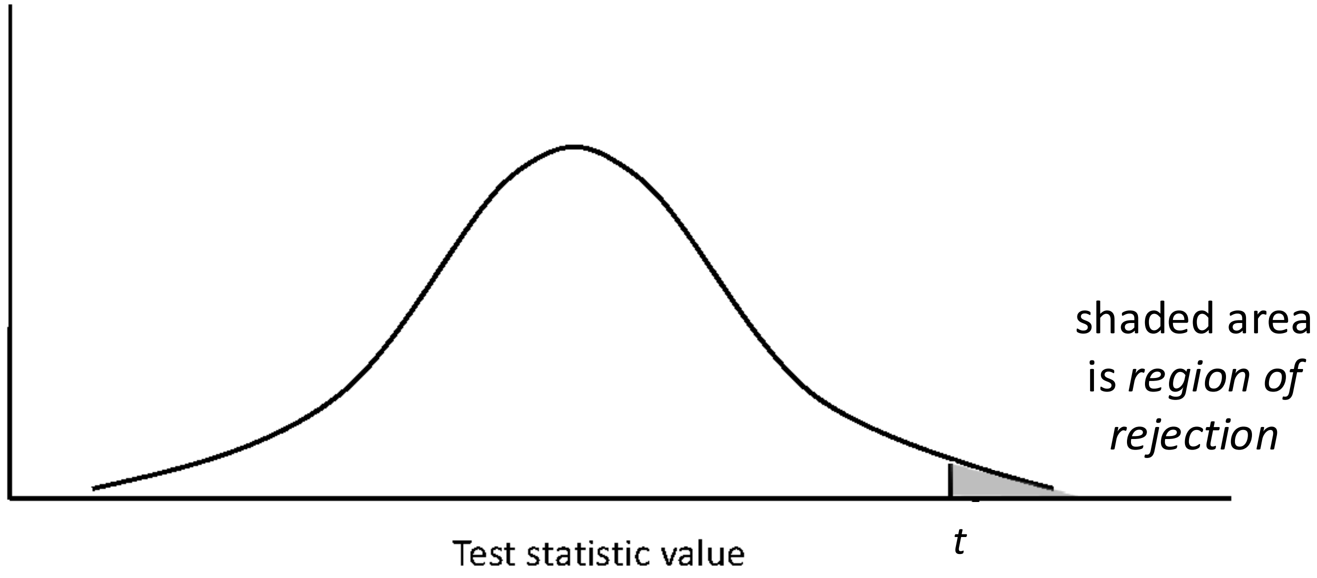
- Useful when testing whether $A > B$, if the consequences of the opposite ($A < B$) are negligible

Two-sided: p-value computed from both tails

- Useful when testing whether $A \neq B$, if you want to allow for the possibility of no or worse result

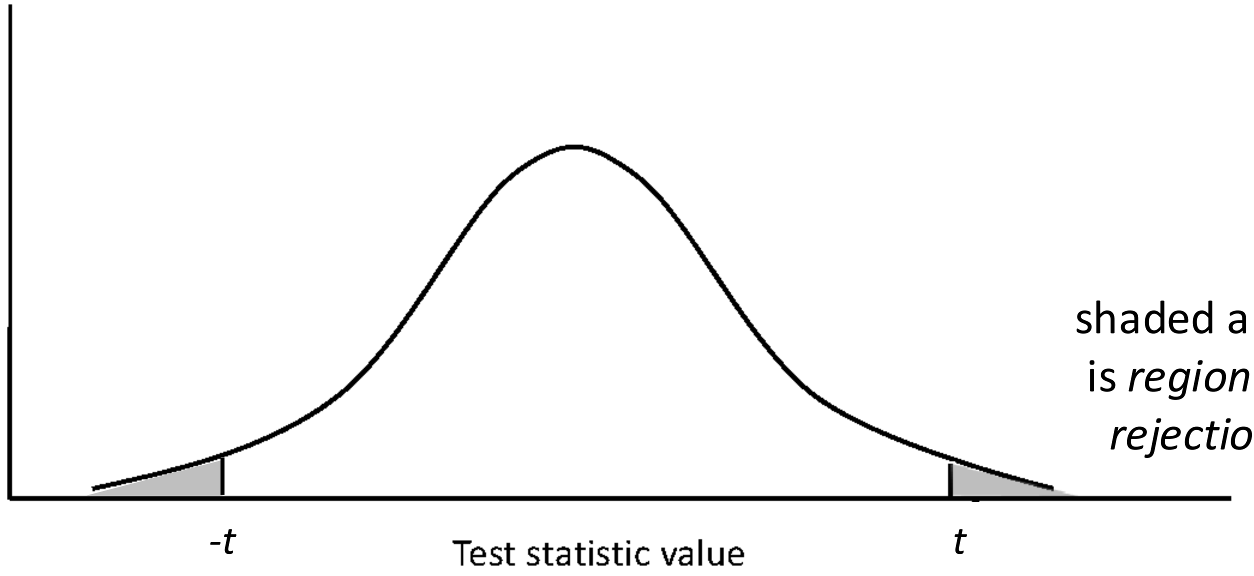
One-sided test ($A > B$)

Distribution for the possible values of a test statistic assuming the null hypothesis



Two-sided test ($A \neq B$)

Distribution for the possible values of a test statistic assuming the null hypothesis



shaded area is *region of rejection*

t-table

One-sided	75%	80%	85%	90%	95%	97.5%	99%
Two-sided	50%	60%	70%	80%	90%	95%	98%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365

…

…

$$n = 5 \rightarrow \nu = 4, t = 2.132$$

t-table

One-sided	75%	80%	85%	90%	95%	97.5%	99%
Two-sided	50%	60%	70%	80%	90%	95%	98%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365

…

…

$n = 5 \rightarrow \nu = 4, t = 2.132$

One-sided: $P(T < t) = 0.95 \rightarrow P(T \geq t) = 0.05$

Two-sided: $P(-t < T < t) = 0.90 \rightarrow P(T \leq -t) + P(T \geq t) = 0.10$

Paired t-test

1. Compute the effectiveness of models A and B
2. Compute the *difference* in effectiveness of A and B
3. Compute a test statistic t for the distribution in (2)
4. Compute the probability p of t under H$_0$
5. Reject H$_0$ if $p \leq \alpha$, for some small α (e.g., 0.05)

Query	A	B	$B - A$
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

Criticisms

t statistic revisited

- $t = \frac{\overline{B-A}}{\sigma_{B-A}} \sqrt{n}$: larger t (more extreme), smaller p

t can be large for two reasons

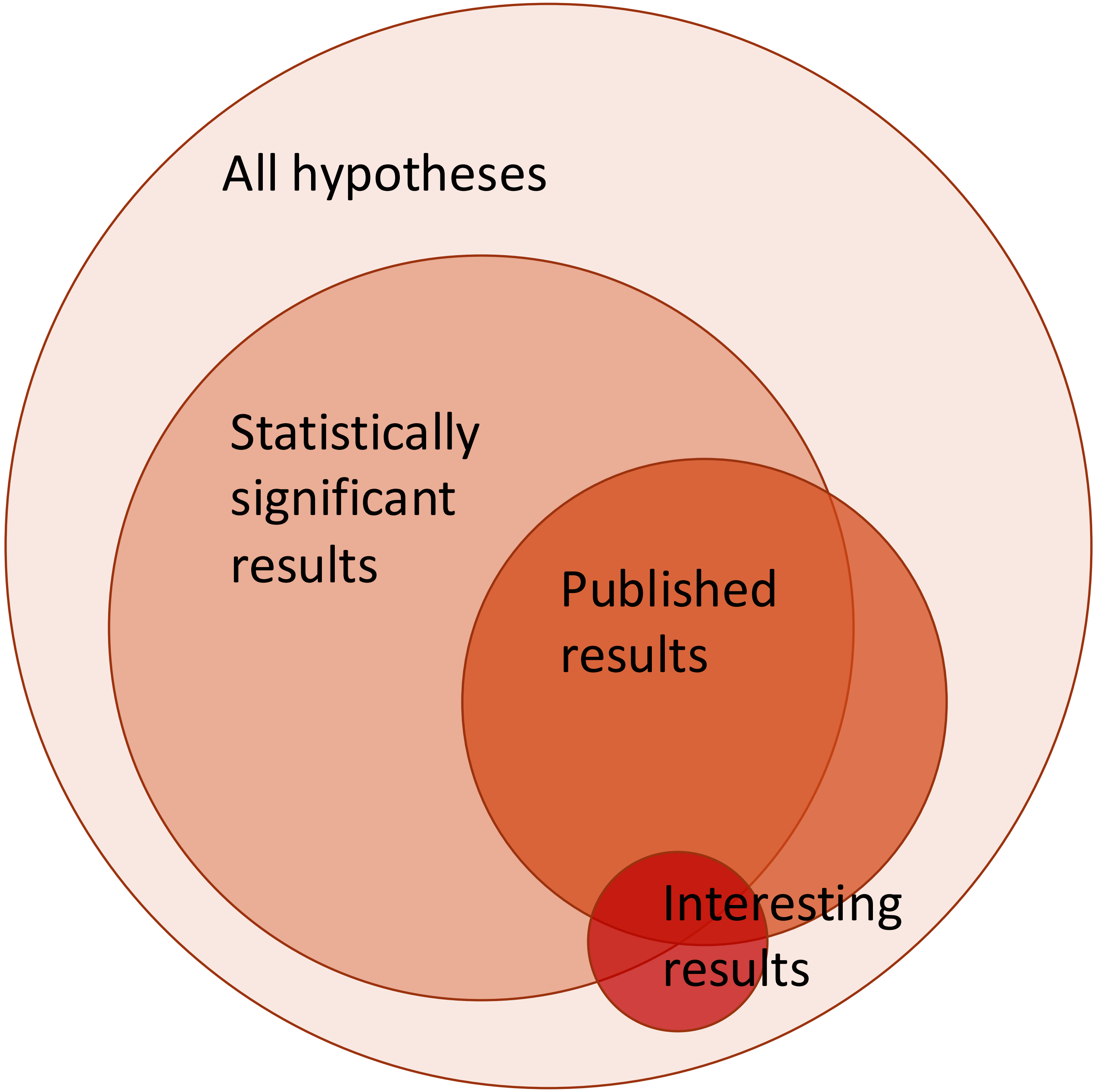
- Large effect size: $\overline{B-A}/\sigma_{B-A}$
- Large sample size: n

How to easily make your results significant? p-hacking

#DONT

Criticisms

> “ *It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important.*
>
> ◦ Ellis, 2010


All hypotheses
Statistically significant results
Published results
Interesting results

Summary

No single measure is the correct one for any application

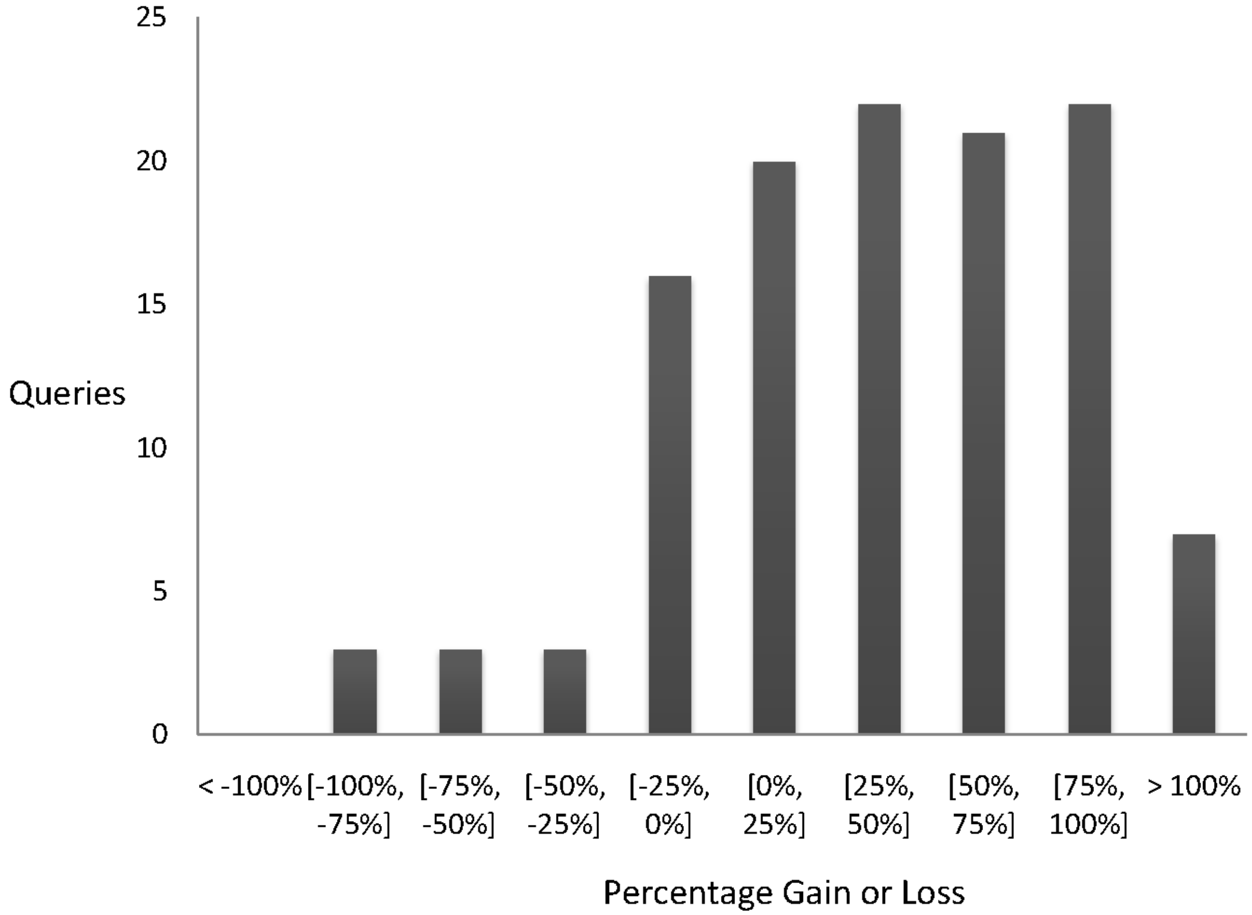
- Choose measures appropriate for task
- Use a combination to highlight different aspects

Use significance tests (two-sided paired t-test)

- Also report effect sizes!

Analyze performance of individual queries

Query summary



References

Search Engines: Information Retrieval in Practice, Ch. 8

Croft et al., 2009

Introduction to Information Retrieval, Ch. 8

Manning et al., 2008

Test collection based evaluation of IR systems

Sanderson, FnTIR 2010

References

Statistical reform in information retrieval?
Sakai, SIGIR Forum 2014

Statistical significance testing in theory and in practice
Carterette, SIGIR 2017

Statistical significance testing in information retrieval: an empirical analysis of type I, type II and type III errors
Urbano et al., SIGIR 2019

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next...

Exam #1

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br