

IR 2 — Recent Advances in Information Retrieval

Session 1 · Kickoff

Kim Gerdes · Friday Nov 6, 2026 · 9:00–12:30

Before we start: the strike and the blockades

- Today is **online** because of the blockades
- Five minutes to talk about the situation
- Can everyone follow online, and reach their partner?
- Next Friday’s format: I will confirm it by mail

Today

9:00	The strike and the blockades
9:05	The seminar: format, calendar, who presents when
9:20	IR in 45 minutes: the spring pipeline as one picture
10:05	Articles and patents in 15 minutes
10:20	Break
10:35	How to read and present an IR paper
11:05	Model presentation, shortened: *Beyond Abstracts*, talk, demo, the questions discussants should ask
11:30	Pairs at work: your paper, your reading plan, your discussants
12:30	End

Where we stand

- Bidding closed on Sept 30; with the late choices sent by email, **all 28 students** ranked their papers
- **Pairs, papers, sessions and discussant roles** are on gerdes.fr/inforet since **Wednesday Nov 4**
- The reading plan due on Oct 23 moves to **Tuesday Nov 10, 23:59**
- Presenting on **Nov 13**? One week instead of five, but the usual deadlines: material on **Wed Nov 11, 18:00**, discussant notes on **Thu Nov 12, 18:00**
- **28 students** in **14 pairs**: **3 talks** per session, 2 on Dec 11 before the synthesis

A reading seminar, with implementations

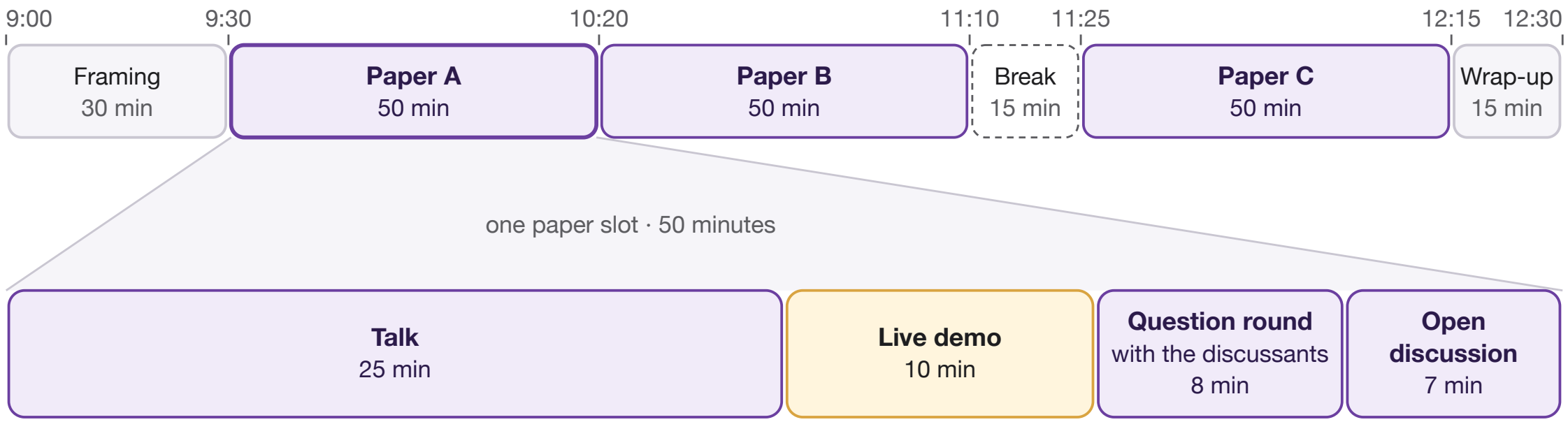
- One paper **per pair**: a modern classic or a 2024–2026 paper
- You **present once** (with your partner) and are **discussant once** (for another pair)
- You **reproduce** one key result at small scale
- You **run the method on the shared testbed**: one results table for the whole class
- Laptop + free Colab: small open models, no API budget

Five sessions, five themes

Date	Theme	Papers
Nov 13	Representations: how embeddings are trained, and where they break	LLM2Vec · LIMIT · static embeddings (timeline talk)
Nov 20	Domains: scientific papers, patents, adaptation, languages	GPL · BGE-M3 · word embeddings in materials science (timeline talk)
Nov 27	LLMs in the loop: rerankers, rewriters, reasoning	RankGPT · Rank1 · HyDE
Dec 4	RAG and agents	Self-RAG · The Power of Noise · Search-R1
Dec 11	Trust: evaluation, attacks, bias, documents that are not text · **synthesis**	Neural retrievers biased towards LLM text · ColPali

Pairs and discussants: **gerdes.fr/inforet/assignments.html**

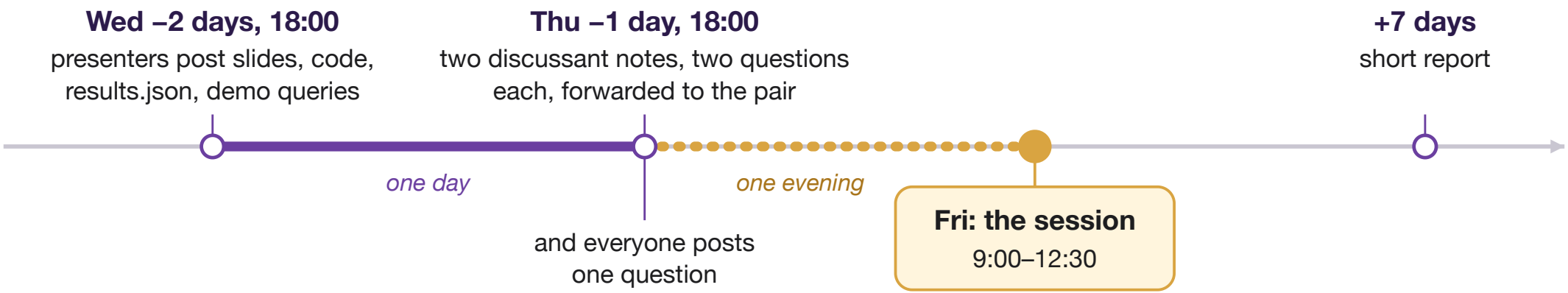
One session, one presentation slot



A Friday morning and a 50-minute slot

- The question round is a **dialogue**: two discussants, two questions each, answers of two minutes at most
- Then the floor: the questions the class posted, anyone may cut in
- I stay silent for the first five minutes of the discussion

The two days before you present



Deadlines around a presentation

- Material on **Wednesday 18:00**, discussant notes on **Thursday 18:00**: one day to read and write, one evening for the pair to prepare the answers
- Late material costs 2 points per day: the discussants only have one day
- Everyone posts **one question per session** on **gerdes.fr/inforet/questions** (Thursday 18:00): that is your participation grade

Grading

Component	Weight	Who
Presentation: talk, demo, handling of the discussion	**40 %**	the pair (±10 % individual)
Implementation + short report: reproduction, testbed row, code, honesty	**40 %**	the pair
Discussant role (once) + one question per session	**20 %**	you

- Nothing else is graded. Report due one week after your talk; everything by **Dec 18**
- **A failed reproduction with a good diagnosis is a valid result**

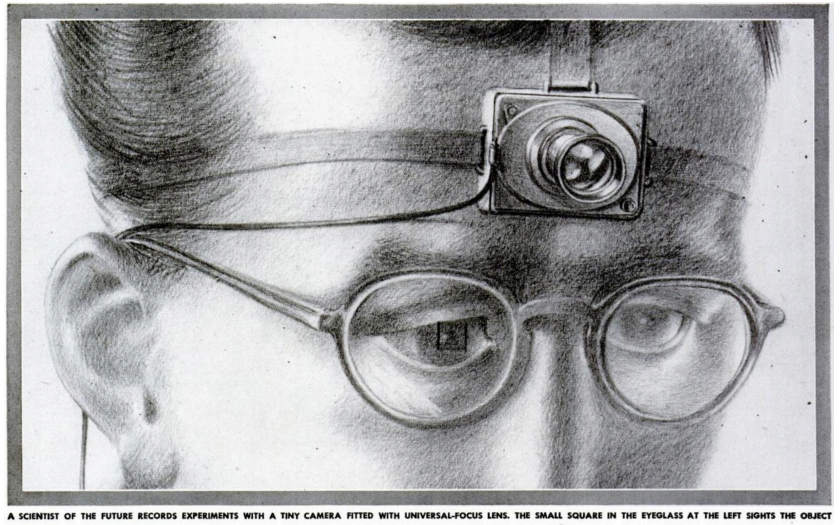
IR in 45 minutes

The spring pipeline as one picture, and the vocabulary we will use every Friday.

Returning students: a refresher. Newcomers: the map for your self-study.

What IR is

- *Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections* (Manning, Raghavan & Schütze 2008)
- **Ranking, not classification**: one query, many documents, a ranked list
- Evaluation is about the **order**, not about one label
- Speed matters: sub-second over millions of documents
- 1945: Vannevar Bush imagines the memex; 1950: Calvin Mooers coins “information retrieval”



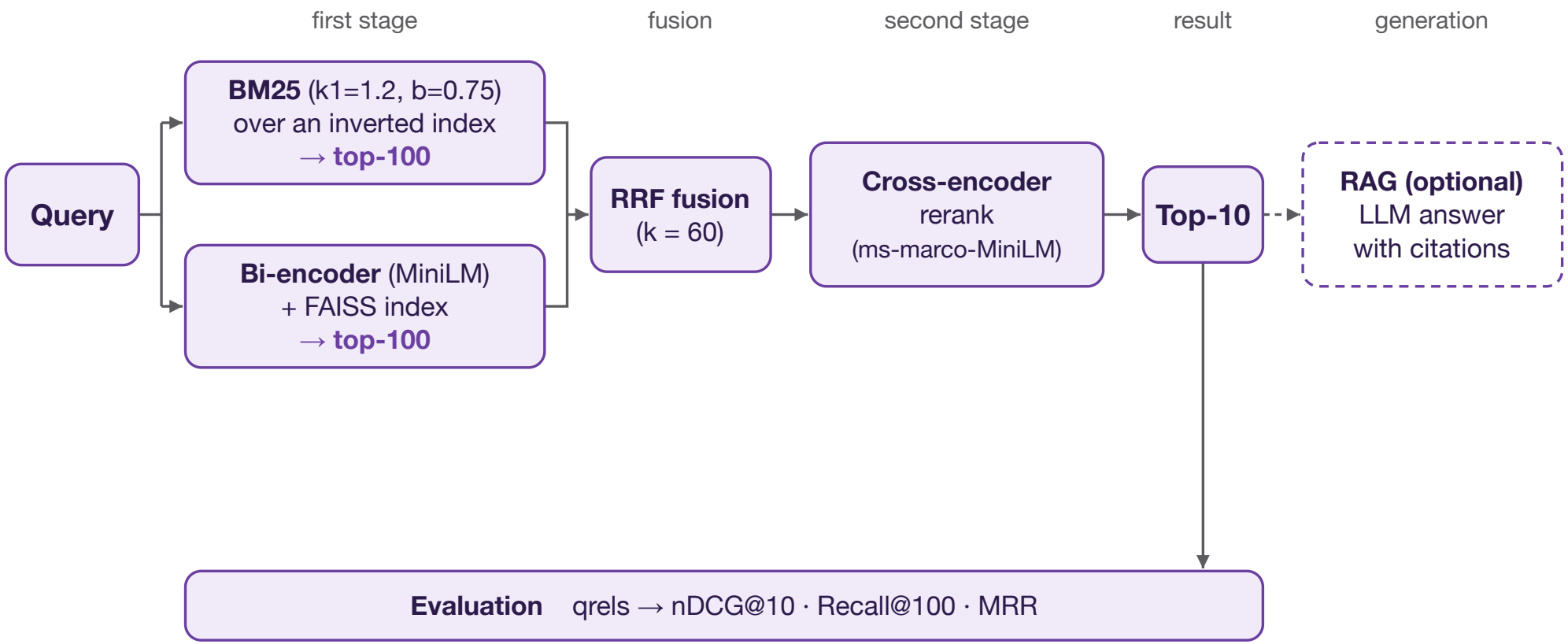
A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK

As We May Think, Life, 10 Sept 1945 (public domain)

The pipeline you built in the spring



Evaluation: qrels, runs, metrics

- **qrels** (*query relevance judgments*): which documents are relevant for which query (binary or graded), judged by humans
- **run**: the ranked list your system returns for each query

Metric	Graded?	Rewards	Use when
P@k, **R@k** · *Precision*, *Recall* at rank k	no	hits above a cutoff	recall matters (passages for a large language model: R@100)
MRR · *Mean Reciprocal Rank*	no	the first relevant hit	one good answer is enough
MAP · *Mean Average Precision*	no	all relevant hits	classic TREC (*Text REtrieval Conference*, NIST, since 1992)
nDCG@10 · *normalised Discounted Cumulative Gain* at rank 10	**yes**	relevant hits near the top	the **modern standard**, our testbed

nDCG by hand

Our ranking

rank	relevance	gain $2^{rel} - 1$	$\log_2(\text{rank} + 1)$	gain / discount
1	●●● 3	7	1	**7**
2	●●○ 2	3	1.58	**1.89**
3	○○○ 0	0	2	**0**
4	●○○ 1	1	2.32	**0.43**

DCG@4 = 7 + 1.89 + 0 + 0.43 = **9.32**

Ideal ranking

rank	relevance	gain / discount
1	●●● 3	7 / 1 = **7**
2	●●○ 2	3 / 1.58 = **1.89**
3	●○○ 1	1 / 2 = **0.5**
4	○○○ 0	0 / 2.32 = **0**

IDCG@4 = 7 + 1.89 + 0.5 + 0 = **9.39**

nDCG@4 = 9.32 / 9.39 = 0.993

☐ ranks 3 and 4 swapped

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

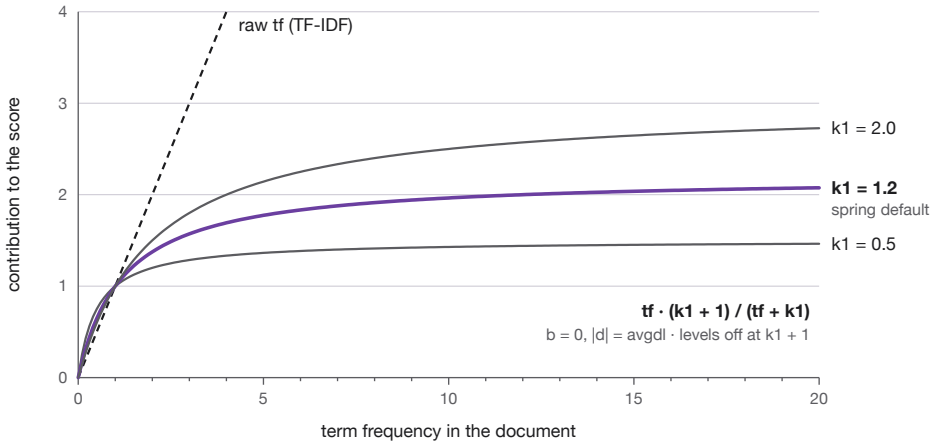
$$\text{nDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$$

- IDCG: the DCG of the ideal order
- Järvelin & Kekäläinen (2002)

Sparse retrieval: the inverted index and BM25

$$\mathrm{BM25}(q,d) = \sum_{t \in q} \mathrm{IDF}(t) \cdot \frac{tf \cdot (k_1 + 1)}{tf + k_1 \left(1 - b + b \frac{|d|}{avgdl}\right)}$$

- **IDF**: rare terms matter (Spärck Jones 1972)
- **k_1**: term frequency saturates
- **b**: long documents are penalised
- No training, fast, interpretable: still **the baseline to beat**
- Weakness: **vocabulary mismatch**



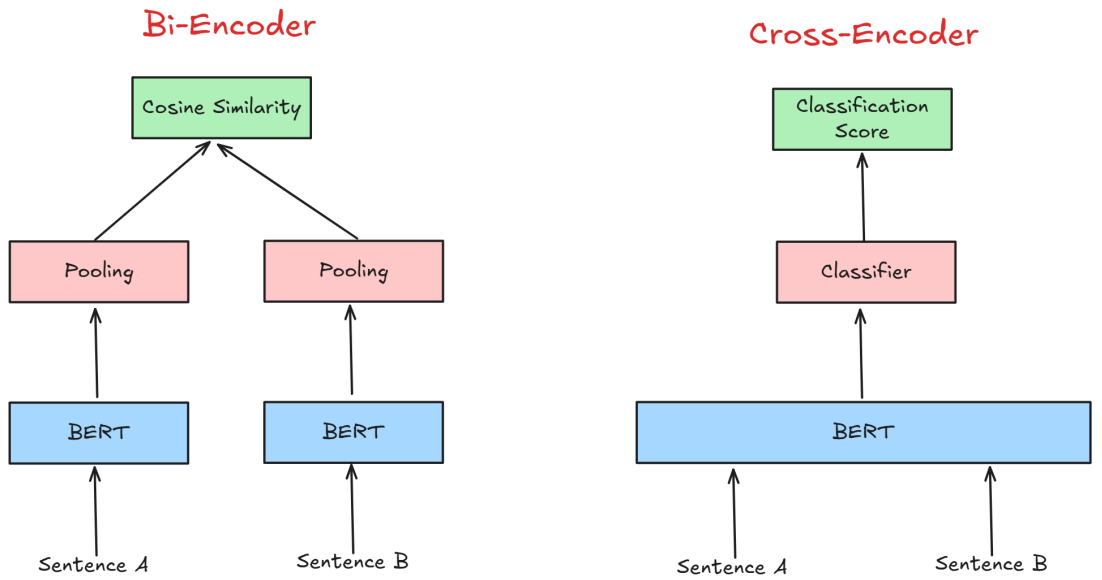
Term-frequency saturation

Dense retrieval: bi-encoders

- Encode query and document **separately** into vectors; score = cosine
- Encode the corpus **once**, then search vectors
- Trained **contrastively** (InfoNCE):

$$\mathcal{L} = -\log \frac{e^{s(q, d^+)/\tau}}{e^{s(q, d^+)/\tau} + \sum_j e^{s(q, d_j^-)/\tau}}$$

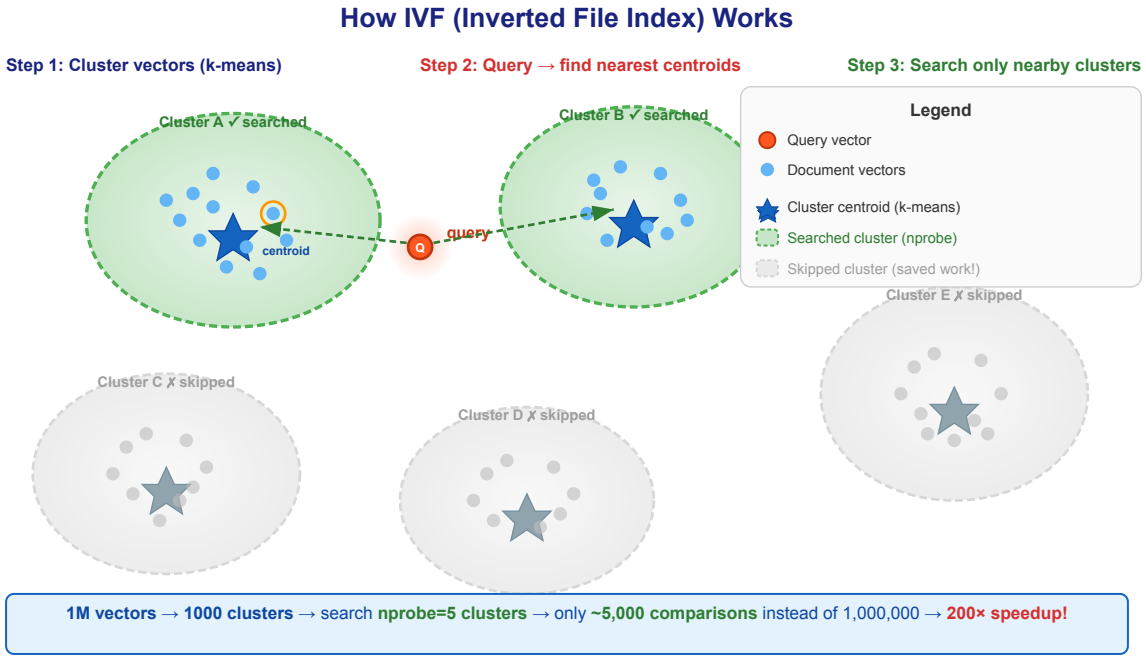
- **In-batch negatives** are free; **hard negatives** (BM25 hits that are wrong) teach the fine distinctions



From the spring slides (class 3)

Scale: approximate nearest neighbour (ANN) search

- Exact search = one dot product per document: fine for 5,000, too slow for 100 million
- **IVF** (*Inverted File index*): cluster the vectors, search only the nearest clusters
- **HNSW** (*Hierarchical Navigable Small World* graph): layers of neighbour links, greedy descent from a sparse top layer to the full bottom one
- **Quantisation**: compress the vectors to fit in memory: **PQ** (*Product Quantisation*: each slice of a vector replaced by the id of its nearest centroid) or one bit per dimension
- Always report **Recall@10 against exact search**



IVF, from the spring slides

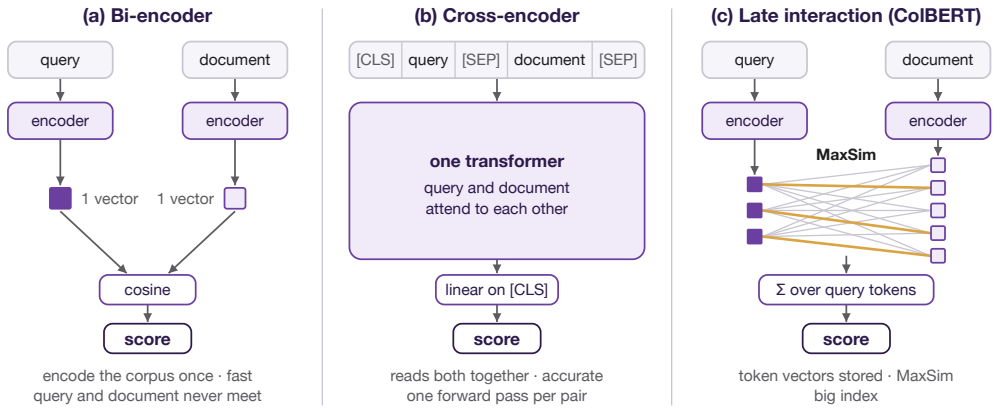
Sparse or dense? It depends on the query

Query	BM25	Dense	Why
“CRISPR gene editing”	✅	❌	exact terms matter
“how does memory work in computers”	❌	✅	meaning, not words
“who is Angela Merkel”	✅	~	names match lexically
paraphrases, synonyms	❌	✅	dense captures meaning
rare technical terms	✅	❌	the encoder never saw them

Neither wins everywhere: hence **hybrid** retrieval.

Two stages: retrieve, then rerank

- **Bi-encoder**: query and document never meet → fast, encode once
- **Cross-encoder**: reads `[CLS] query [SEP] doc` together → accurate, one forward pass per pair
- So: cheap first stage for the **top-100**, expensive reranker for the **top-10**
- **Late interaction** (ColBERT): token vectors and MaxSim, in between, at the price of a big index



Fusion and expansion

Reciprocal Rank Fusion (Cormack, Clarke & Büttcher 2009)

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}, \quad k = 60$$

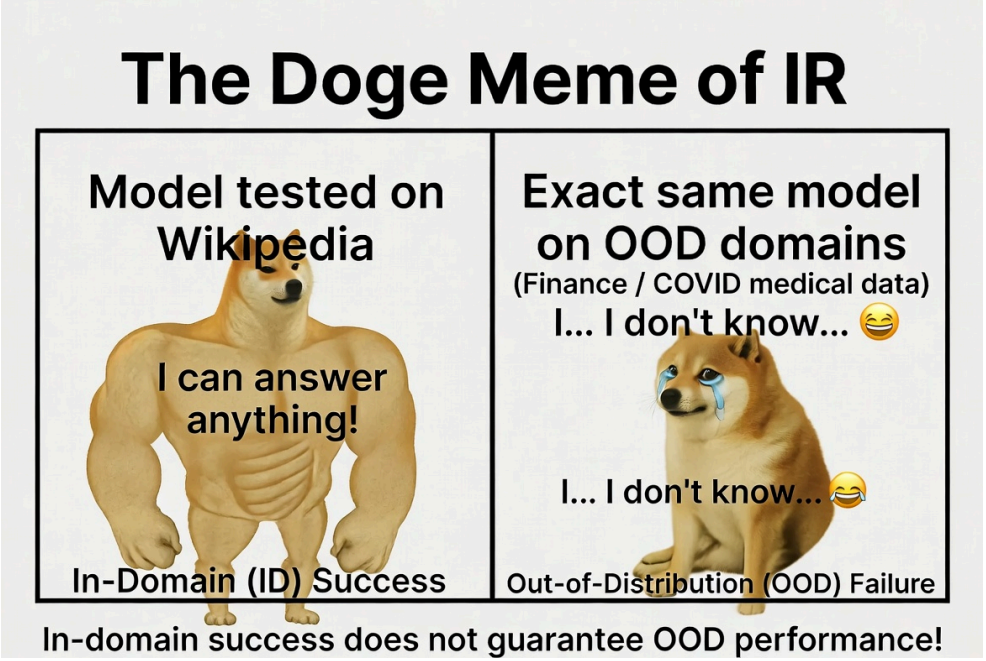
- rank 3 in BM25, rank 7 in dense: 1/63 + 1/67 = 0.031
- uses ranks only: no score calibration needed

Expansion against vocabulary mismatch

- **doc2query**: generate likely queries for each document at indexing time, append them
- **query expansion**: add terms to the query (pseudo-relevance feedback; today: an LLM writes them, see HyDE on Nov 27)

Benchmarks: BEIR and MTEB

- **BEIR** (Thakur et al. 2021): 18 datasets, zero-shot. Dense models trained on MS MARCO often **lost to BM25** out of domain
- **MTEB** (Muennighoff et al. 2023): the embedding leaderboard; top scores differ in the third decimal, and some training sets overlap with the test sets
- Lesson: **evaluate on your data**, and on datasets the authors did not choose



The doge meme of IR (spring, class 5)

Our starting baselines: the spring pipeline on three BEIR datasets

nDCG@10	BM25	Dense (MiniLM)	BM25 + dense → RRF → cross-encoder
SciFact (5,183 docs · 300 queries)	**0.676**	0.646	0.603
NFCorpus (3,633 · 323)	0.321	0.316	**0.337**
ArguAna (8,674 · 1,406)	**0.516**	0.503	0.429

- The spring class 5 pipeline, scored with **pytrec_eval**, the standard TREC evaluator, as in the testbed
- The full pipeline **hurts** on SciFact and ArguAna, but the spring reranker read only 512 characters: on whole documents it reaches **0.688** on SciFact
- Stronger baselines and every pair’s row: the **results table** in the repo

The shared testbed: two tracks

Track A · Effectiveness

datasets	docs	queries
SciFact	5,183	300
NFCorpus	3,633	323
ArguAna	8,674	1,406

+ the spring scientific-article set

metrics

nDCG@10 · Recall@100
paired t-test p-value
bootstrap 95% CI vs the closest baseline

Track B · Efficiency

dataset	passages	queries
MS MARCO subset	92,565	1,000

for papers such as

Seismic · RaBitQ · HNSW · MUVERA
WARP · ConstBERT · ACORN-type

metrics

queries/second on a laptop CPU
index build time · index size · peak RAM
Recall@10 vs exact search

every pair adds one row per dataset → **one shared results table**

The repo is online: github.com/kimgerdes/ir2-testbed, with `testbed.py`, the baseline scores, the starter notebook and the results table.

RAG in one slide

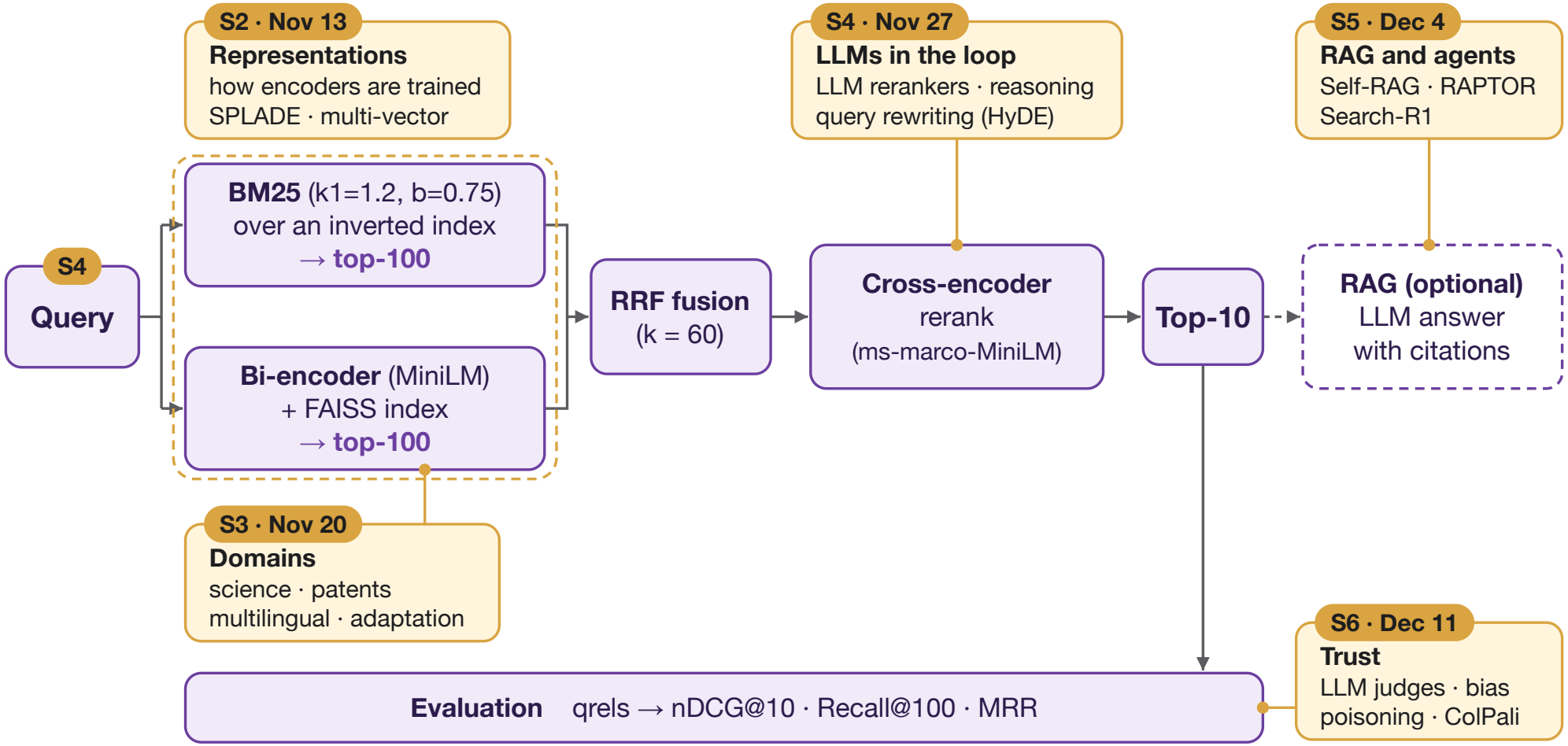
- **RAG** (*Retrieval-Augmented Generation*): **retrieve** passages → **put them in the prompt** → **generate** an answer with citations
- **Chunking** decides what can be found: too long dilutes, too short loses context
- Failure modes: nothing relevant retrieved, the right passage ignored, an answer not supported by the passages
- Evaluate both halves: **R** (Recall@k of the passages) and **G** (faithfulness, correctness)
- *Retrieval decides what the model can know. Generation decides what it says about it.*

The G can be an open model you download

	Kimi K3	**Qwen3.8-27B**
from	Moonshot AI, Jul 2026	Alibaba, Aug 2026
parameters	2.8 trillion, 104 B active (mixture of experts)	27.8 billion
weights	**1.56 TB**: fits on a 2 TB disk	56 GB, **≈ 17 GB** in 4 bits
runs on	a GPU server with ≥ 1.6 TB of GPU memory	the memory of many laptops; a normal GPU for reasonable speed

Neither has read **your** documents: that is what the R is for.

What changed since 2023: the five sessions



Vocabulary check

qrels · run · nDCG@10 · Recall@100 · MRR · inverted index · BM25 (k_1, b) · IDF · vocabulary mismatch · bi-encoder · cross-encoder · late interaction · MaxSim · InfoNCE · in-batch negatives · hard negatives · ANN · IVF · HNSW · RRF · doc2query · zero-shot · BEIR · MTEB · chunking · RAG

Anything here you could not explain to your partner? Write it down: that is your self-study list for this weekend.

Articles and patents in 15 minutes

Articles are what the testbed and today’s model presentation are made of. Patents are my own work: the same pipeline with the opposite metric.

Anatomy of a scientific article



Beyond Abstracts:
Learning Scientific Paper Embeddings from Full-Text Windows

Younes Djemmal[1] Oloruntobi Olutola[1, 2] Kim Gerdes[1, 2]
(1) Questel, Paris, France
(2) Université Paris-Saclay, LISN, CNRS, France
ydjemmal@questel.com, oloruntobi.olutola@lisn.fr, kim.gerdes@lisn.fr

RÉSUMÉ
Les plongements d'articles scientifiques sont généralement entraînés uniquement sur le titre et le résumé. Nous présentons S2Full, un corpus d'entraînement de 2,52 millions d'articles en texte intégral avec liens de citation, et entraînons sur des fenêtres aléatoires du corps avec un objectif d'auto-alignement. Sur un benchmark de recherche d'articles fondé sur les citations, notre approche présente de bonnes performances lors de la recherche d'articles en utilisant uniquement les titres et résumés, tout en améliorant les résultats lorsque les requêtes incluent du contenu issu du corps principal des articles. Les meilleurs scores observés proviennent d'un modèle ModernBERT, tandis que la comparaison SciBERT pleinement entraînée montre déjà le même effet directionnel. L'entraînement sur le texte intégral améliore ainsi la recherche citationnelle à partir de formulations plus variées du contenu d'un article, au-delà de son résumé.

ABSTRACT
Beyond Abstracts: Learning Scientific Paper Embeddings from Full-Text Windows

Paper embeddings are typically trained on titles and abstracts only. We present S2Full, a 2.52M-paper full-text training corpus with citation links, and train on random body windows with a self-alignment objective. In a citation-based article retrieval benchmark, our approach maintains strong performance when retrieving papers using only titles and abstracts, while improving retrieval when queries include content from the main body of papers. The strongest observed scores come from a ModernBERT model, while the fully trained SciBERT comparison already shows the same directional gain from body-view exposure. Full-text training, therefore, improves citation-based retrieval from more diverse formulations of a paper's content beyond its abstract.

MOTS-CLÉS : plongements d'articles, texte intégral, fenêtres aléatoires, recherche citationnelle, SBERT.

KEYWORDS: paper embeddings, full text, random windows, citation-based retrieval, SBERT.

1 Introduction

Scientific paper embeddings are central to citation-based retrieval, recommendation, and citation analysis, but widely used citation-informed models such as SPECTER, SciNCL, and SPECTER2 remain centered on title-abstract representations rather than full-text body sections (Cohan *et al.*, 2020; Ostendorff *et al.*, 2022; Singh *et al.*, 2023). The practical reason is architectural: widely used sentence-transformer encoders based on BERT are limited to short contexts, so the introduction,

First page of Beyond Abstracts (Djemmal, Olutola & Gerdes 2026)

- **Title and abstract**: about 250 words, the part everyone reads, and often the only part that is open
- **Body**: about 3,900 words in 9.5 sections: problem, related work, method, results, discussion
- **References**: what the paper builds on, and its links to other papers
- **Metadata**: authors, venue, date, field
- Our testbed and citation-trained embeddings such as SPECTER read **title + abstract only**

What a citation means

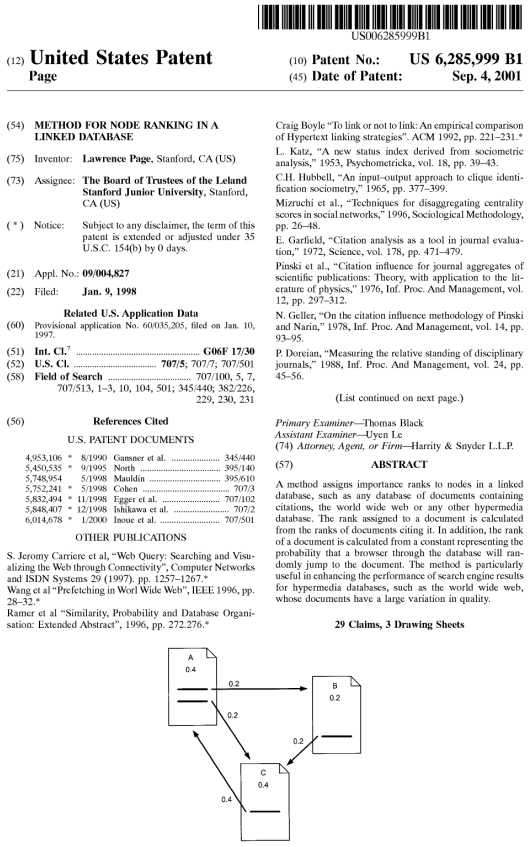
- Authors cite for many reasons: background, a method they use, a result they compare with, a claim they reject
- **A free training signal**: “A cites B” gives millions of related pairs: SPECTER, SciNCL, SPECTER2
- **But a proxy**: related is not the same as relevant, and reference lists are incomplete
- **Noise**: self-citations, and hub papers that everybody cites
- Keep this for 11:00: *would a human judge call the cited paper relevant?*

Who searches articles, and for what

Task	Query	Success means	You will see it in
Find related work	a paper’s title and abstract	a few strong papers at the top	Beyond Abstracts, TA → TA
Check a claim	one sentence	the abstracts that bear on it	SciFact
Find what to cite	a passage of your draft	the paper it should cite	Beyond Abstracts, Body → TA
Systematic review	a long Boolean query	**every** eligible study	recall first, like patents

Same documents, different idea of success.

Anatomy of a patent: US 6,285,999 (PageRank)



US006285999B1

(12) **United States Patent**
Page

(10) **Patent No.: US 6,285,999 B1**
(45) **Date of Patent: Sep. 4, 2001**

(54) **METHOD FOR NODE RANKING IN A LINKED DATABASE**

(75) Inventor: **Lawrence Page**, Stanford, CA (US)

(73) Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Stanford, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/004,827**

(22) Filed: **Jan. 9, 1998**

Related U.S. Application Data

(60) Provisional application No. 60/035,205, filed on Jan. 10, 1997.

(51) **Int. Cl.**[7] **G06F 17/30**
(52) **U.S. Cl.** **707/5**; 707/7; 707/501
(58) **Field of Search** 707/100, 5, 7, 707/513, 1–3, 10, 104, 501; 345/440; 382/226, 229, 230, 231

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,953,106	*	8/1990	Gansner et al.	345/440
5,450,535	*	9/1995	North	395/140
5,748,954	*	5/1998	Mauldin	395/610
5,752,241	*	5/1998	Cohen	707/3
5,832,494	*	11/1998	Egger et al.	707/102
5,848,407	*	12/1998	Ishikawa et al.	707/2
6,014,678	*	1/2000	Inoue et al.	707/501

OTHER PUBLICAITONS

S. Jeromy Carriere et al, "Web Query: Searching and Visualizing the Web through Connectivity", Computer Networks and ISDN Systems 29 (1997). pp. 1257–1267.*
Wang et al "Prefetching in Worl Wide Web", IEEE 1996, pp. 28–32.*
Ramer et al "Similarity, Probability and Database Organisation: Extended Abstract", 1996, pp. 272.276.*
Craig Boyle "To link or not to link: An empirical comparison of Hypertext linking strategies". ACM 1992, pp. 221–231.*
L. Katz, "A new status index derived from sociometric analysis," 1953, Psychometricka, vol. 18, pp. 39–43.
C.H. Hubbell, "An input–output approach to clique identification sociometry," 1965, pp. 377–399.
Mizruchi et al., "Techniques for disaggregating centrality scores in social networks," 1996, Sociological Methodology, pp. 26–48.
E. Garfield, "Citation analysis as a tool in journal evaluation," 1972, Science, vol. 178, pp. 471–479.
Pinski et al., "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," 1976, Inf. Proc. And Management, vol. 12, pp. 297–312.
N. Geller, "On the citation influence methodology of Pinski and Narin," 1978, Inf. Proc. And Management, vol. 14, pp. 93–95.
P. Doreian, "Measuring the relative standing of disciplinary journals," 1988, Inf. Proc. And Management, vol. 24, pp. 45–56.

(List continued on next page.)

Primary Examiner—Thomas Black
Assistant Examiner—Uyen Le
(74) *Attorney, Agent, or Firm*—Harrity & Snyder L.L.P.

(57) **ABSTRACT**

A method assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database. The rank assigned to a document is calculated from the ranks of documents citing it. In addition, the rank of a document is calculated from a constant representing the probability that a browser through the database will randomly jump to the document. The method is particularly useful in enhancing the performance of search engine results for hypermedia databases, such as the world wide web, whose documents have a large variation in quality.

29 Claims, 3 Drawing Sheets

US 6,285,999 B1, front page (USPTO, public domain)

- **Title, abstract, drawings**: what it is about
- **Description**: teaches how to do it
- **Claims**: what is protected, nothing else
- **Bibliographic data**: dates, applicant, inventors, **classes**, **citations**
- Patents are public documents: 100+ million, every language, every field
- This one: 29 claims, 7 patent + 20 non-patent references, **cited by 817** later patents

Article against patent

	Scientific article	Patent
Written to	be read and cited	define a legal right
The part that decides	abstract, then results	the **claims**
Who adds citations	the authors	applicants and **examiners**
Typical search	a few good results at the top	**miss nothing** published before the priority date
Metric	nDCG@10	recall, PRES (Magdy & Jones 2010)
PageRank	Page et al. 1998	US 6,285,999

Same pipeline, opposite metric.

Break

How to read and present an IR paper

Claims, evidence, baselines, significance, and what to put on your slides.

Read for the claim

1. Write the paper's **claim in one sentence**: "Method X improves Y on Z"
2. Find **the table** that is supposed to prove it
3. Ask what result would have **falsified** it, and whether they looked
4. Separate **what was measured** from **what is speculated** in the discussion

Baselines: were they tuned?

- A weak baseline makes any method look good: **BM25 with default parameters, a small dense model, an untuned reranker**
- Same budget for everyone? Same training data, same number of parameters, same compute?
- Ablations: **remove one component at a time**: which part did the work?
- Our own field fell for weak baselines twice: next slide

Weak baselines: IR fell for them twice

2009 · Armstrong, Moffat, Webber & Zobel

Improvements that don't add up: ad-hoc retrieval results since 1998 (CIKM 2009)

- Ten years of SIGIR and CIKM papers on the TREC ad hoc collections
- Dozens of improvements, many "statistically significant", yet **no upward trend** over the decade
- The baselines were often **below the median TREC system** of the original evaluation; only a handful of papers beat the best TREC run of years before

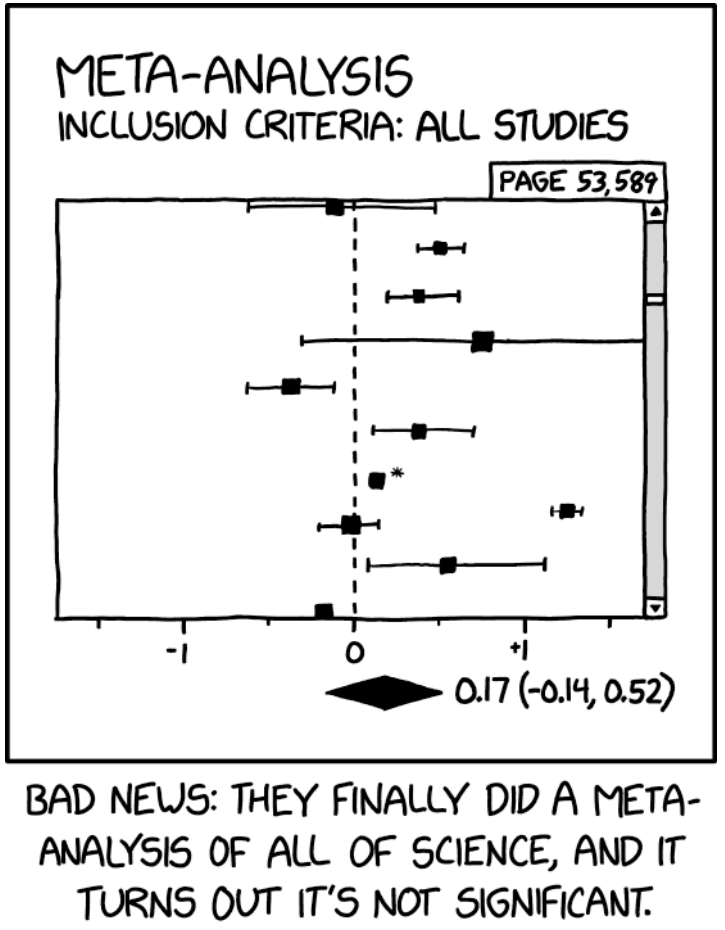
2019 · Jimmy Lin

The neural hype and comparisons against weak baselines (SIGIR Forum)

- The same story with neural rankers: on Robust04, a well-tuned **BM25 + RM3** query expansion (average precision 0.297) matches a new neural model (0.297) said to "significantly outperform competitive baselines"
- The follow-up (Yang et al., SIGIR 2019): of 109 papers from 2005–2018, **one third** used a baseline below the TREC median
- Then BERT arrived, and Lin wrote *The Neural Hype, Justified! A Recantation* (Dec 2019)

Reading a results table with 300 queries in mind

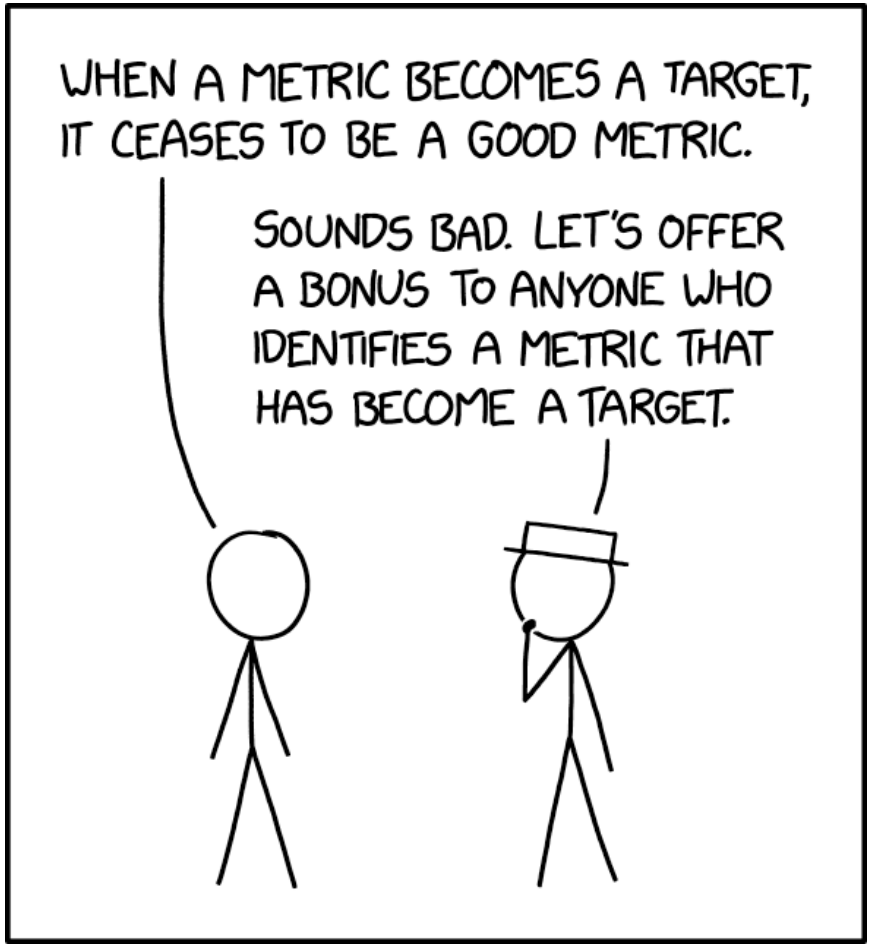
- SciFact has **300 queries**: +0.01 nDCG@10 can be noise
- **Paired** tests on per-query differences: paired t-test (Urbano et al. 2019), bootstrap 95 % CI
- Report the **effect size**, not just the p-value
- Twenty comparisons at $p < 0.05$: one “wins” by chance
- The testbed script computes the p-value and the CI for you: **+0.01 with p = 0.4 is reported as noise**



xkcd #2755 “Effect Size”, Randall Munroe (CC BY-NC 2.5)

State of the art on a saturated benchmark

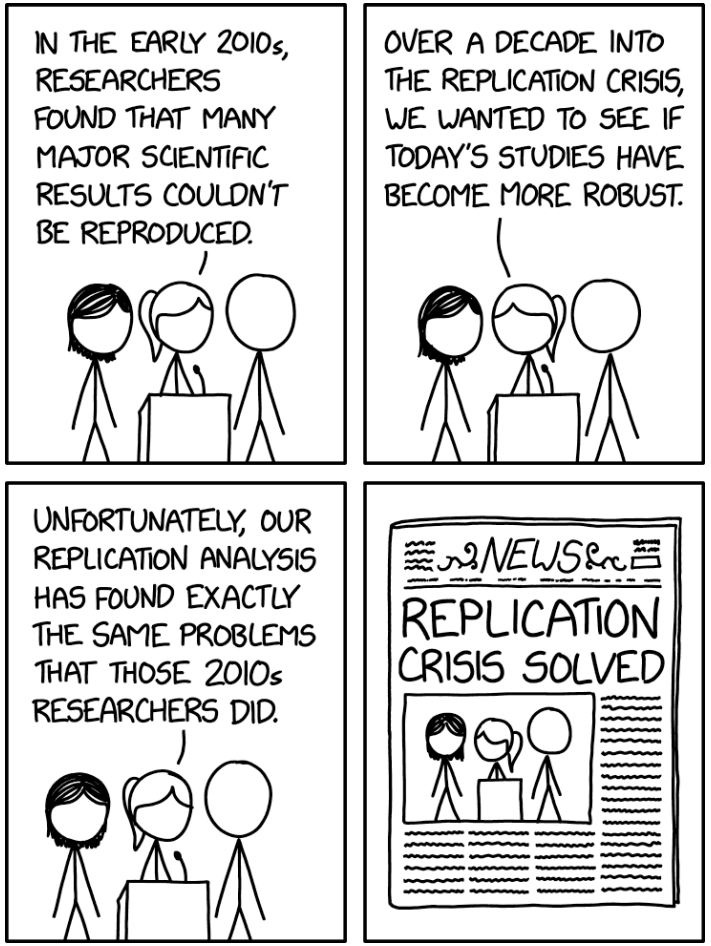
- Leaderboard gaps in the **third decimal** rarely survive a new dataset
- Benchmark artefacts: training data overlapping test data, queries written for one system, judgments pooled from old systems
- Ask: **does it hold on data the authors did not choose?** That is what our testbed is for
- Ask: **what does it cost?** Latency, index size, GPU hours, API calls



xkcd #2899 “Goodhart’s Law”, Randall Munroe (CC BY-NC 2.5)

Reproducibility checklist

- Code, checkpoints and data splits released?
- Hyper-parameters and seeds reported?
- Which numbers were **re-run** and which **copied** from other papers?
- Compute budget stated?
- Could you reproduce one number **this week**? Start there



xkcd #3117 “Replication Crisis”, Randall Munroe (CC BY-NC 2.5)

Your talk (25 minutes)

1. The problem, in one slide
2. **What the spring pipeline does** vs **what this paper changes**: one slide
3. The method, with one figure you redrew yourself
4. Their evidence: the one table that matters
5. **Your number**: the reproduction, scaled down
6. **Your testbed row**, with the p-value
7. What did **not** work

Assume the audience has read the abstract and the spring notes, nothing else.

The demo (10 minutes, three queries)

- **Hero**: the method clearly wins, and you can say why
- **Failure**: the method collapses, and you can say why
- **Wildcard**: proposed on the spot by the discussants or the room
- Side by side: spring baseline vs your method, on **precomputed indexes**
- Nothing trained or indexed live; a two-minute **screen recording** as fallback

Discussants: the auditor and the skeptic

A · the auditor

- reads the code, the README and `results.json` as closely as the paper
- tuned baselines? faithful reproduction? does the testbed row hold up?
- **moderates** the open discussion

B · the skeptic

- reads the paper against the field
- a real problem or a benchmark artefact? what does it cost to deploy?
- relates it to the spring pipeline and to another seminar paper

Both: one-page note by **Thursday 18:00**, one day after the material, ending with **two questions**, forwarded to the pair that evening.

Model presentation

Beyond Abstracts: Learning Scientific Paper Embeddings from Full-Text Windows (Djemmal, Olutola & Gerdes 2026)

A real slot is 50 minutes: talk 25 · demo 10 · question round 8 · open discussion 7.

Today, half of it: talk 12 · demo 5 · then the questions the two discussants should ask, and one answer.

Watch the **format**: what goes on each slide, the three demo queries, what a good question looks like.

Pairs at work

Now: your pair

1. **Find your partner** (gerdes.fr/inforet/assignments.html) and get together
2. **25 min**: read the abstract, the introduction and the main results table of your paper together
3. **20 min**: draft the reading plan: the paper in 5 sentences · what you reproduce · what you run on the testbed · compute plan · risks
4. I come round: **Nov 13 and Nov 20 pairs first**
5. Discussants of **Nov 13** (Sutan, Kaur, Wakeford, Costa, Ran Lu, Diop): introduce yourselves to your pair today

Before next Friday

Who	What	When
everyone	the reading plan, half a page per pair	**Tue Nov 10, 23:59**
Nov 13 pairs	slides, code, results, demo queries	**Wed Nov 11, 18:00**
Nov 13 discussants	one-page note, two questions	**Thu Nov 12, 18:00**
everyone	one question on the Nov 13 papers, on **gerdes.fr/inforet/questions**	**Thu Nov 12, 18:00**
newcomers	the spring slides and notebooks, classes 1–5	this weekend

Reading plans by email to me. Material by pull request to github.com/kimgerdes/ir2-testbed.

Questions?

gerdes.fr/inforet · kim.gerdes@lisn.fr

See you on Friday Nov 13, 9:00: Representations.