Engagement DCG vs SME DCG: Evaluating the Wisdom of the Crowd

D. Rosenoff, V. H. Bandepalli, E. Chatelain, & S. Kottapuzhackal / LexisNexis

Haystack Search Relevance Conference

28 April 2022

Wisdom of the Crowd

- Jar of Jellybeans problem
- Statistical Aggregate Average of guesses by many participants
- Random rater role, expertise, & experience assumed
- Works better with a lead guess
- Often very wide result standard deviations
- Tends toward a close but not necessarily exact or repeatable answer

Subject Matter Experts

- Informed judgements
- Statistical Mode Aggregate of responses from a few experts
- Training, Experience, Expertise
- Frequently very high agreement
- Blind HRT studies tend toward a highly repeatable, very accurate response

Discounted Cumulative Gain (DCG) in a nutshell

- Same DCG formula used by both eDCG and hDCG
  - Sum of ((2^relevance score -1)/ log2 (position of document +1)) for the top N documents

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

- Three Assumptions behind this metric:
  - Highly relevant documents are more useful than marginally relevant documents
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined
  - More documents that are relevant are better than fewer

Two Kinds of DCG: Engagement and HRT

Engagement / eDCG

- Based on Wisdom of the Crowd assumption that customer engagement with documents in the result set is indicative of relevance
- Relies on aggregate usage statistics
- Useful result for A/B testing, characterizing user experience

HRT / hDCG

- Based on Subject Matter Expert ratings of query-document relevance
- Relies on ratings from selected query document pair sets
- Useful for pre-release metrics, algorithm evaluation, competitive benchmarking, and post-release regression

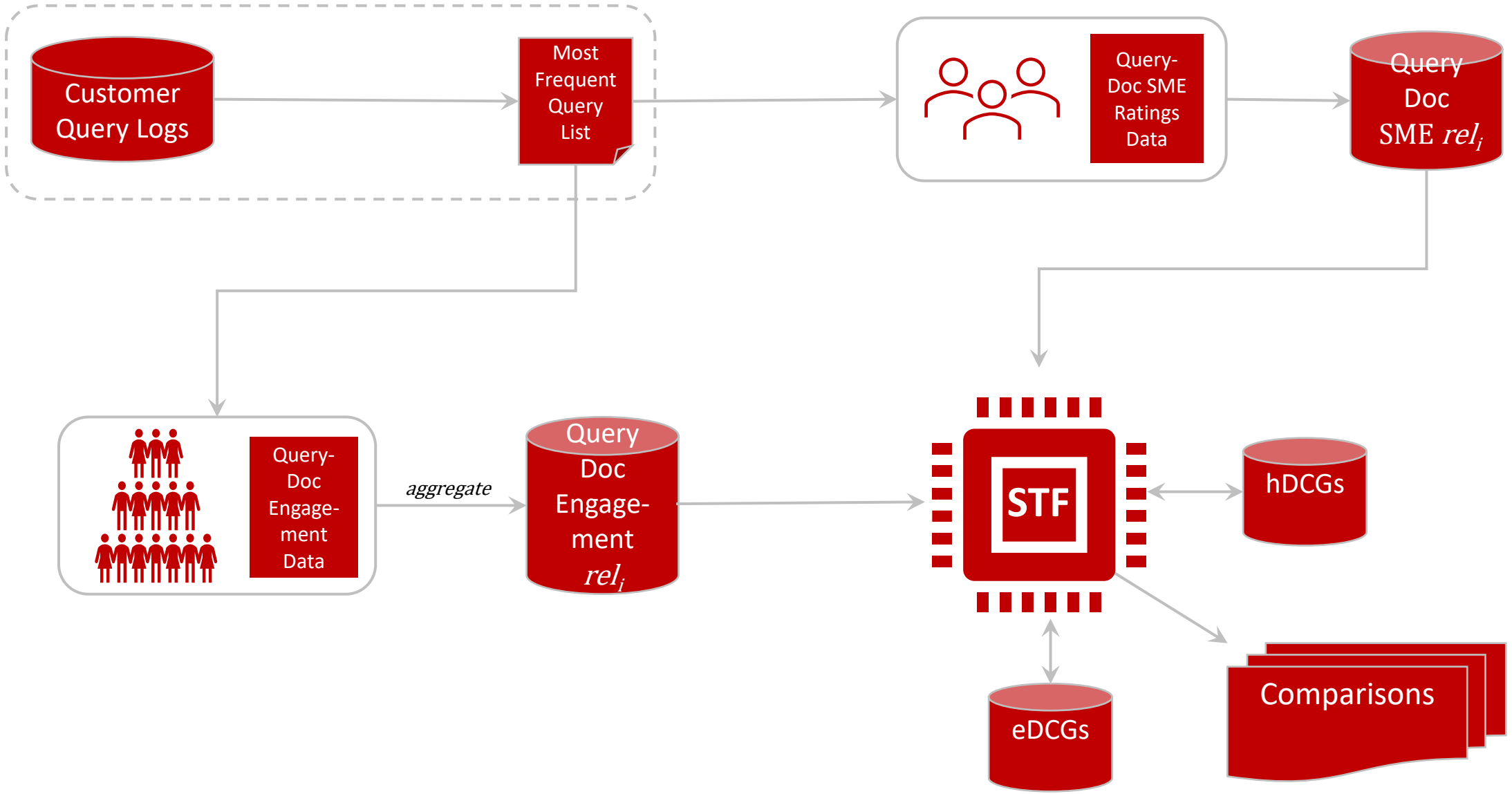
Sometimes results agree, sometimes don't agree, sometimes neutral

Goal: Direct hDCG vs eDCG Comparison

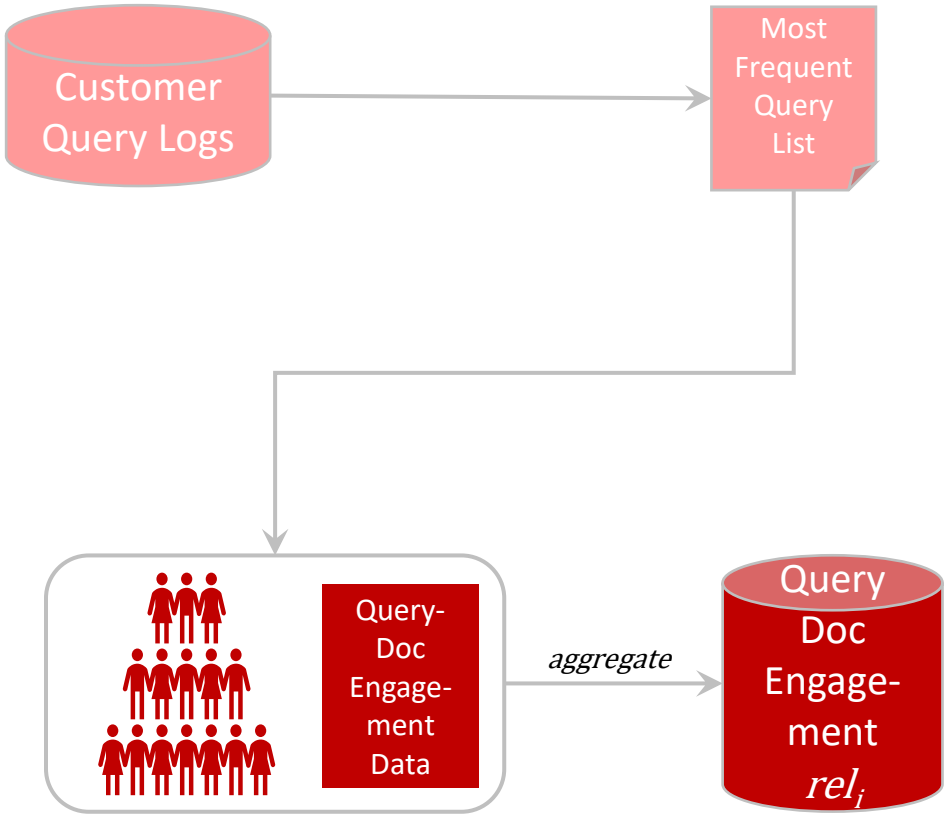
- Does eDCG = hDCG for the same queries?
- Do eDCG and hDCG scale?

- How do they relate?
- Are they equally sensitive?
- Is one more useful than the other?

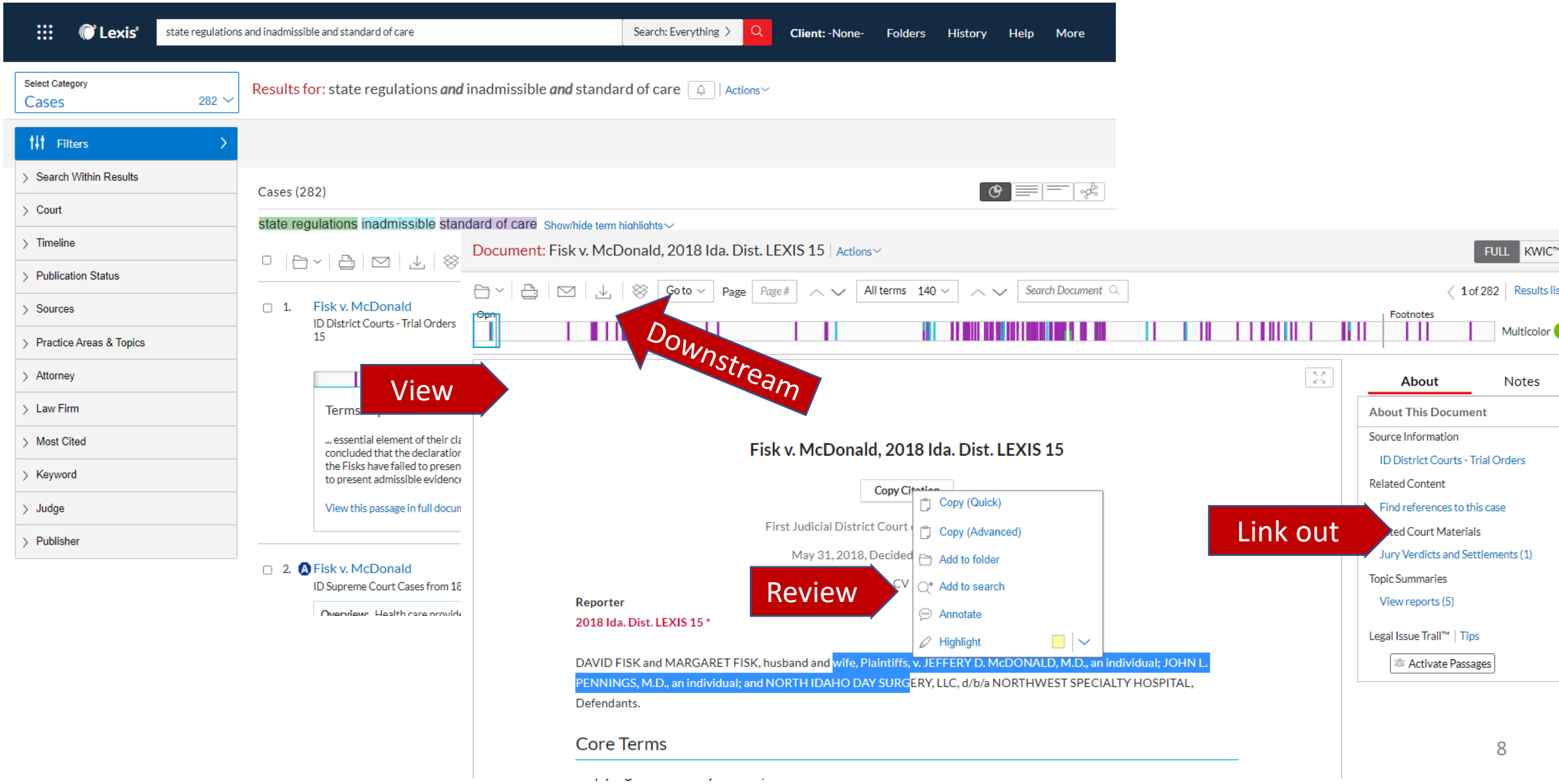
eDCG / hDCG Comparison Flowchart



eDCG / hDCG Comparison Flowchart



Collecting Engagement Scores



Aggregating results to get engagement rel_i

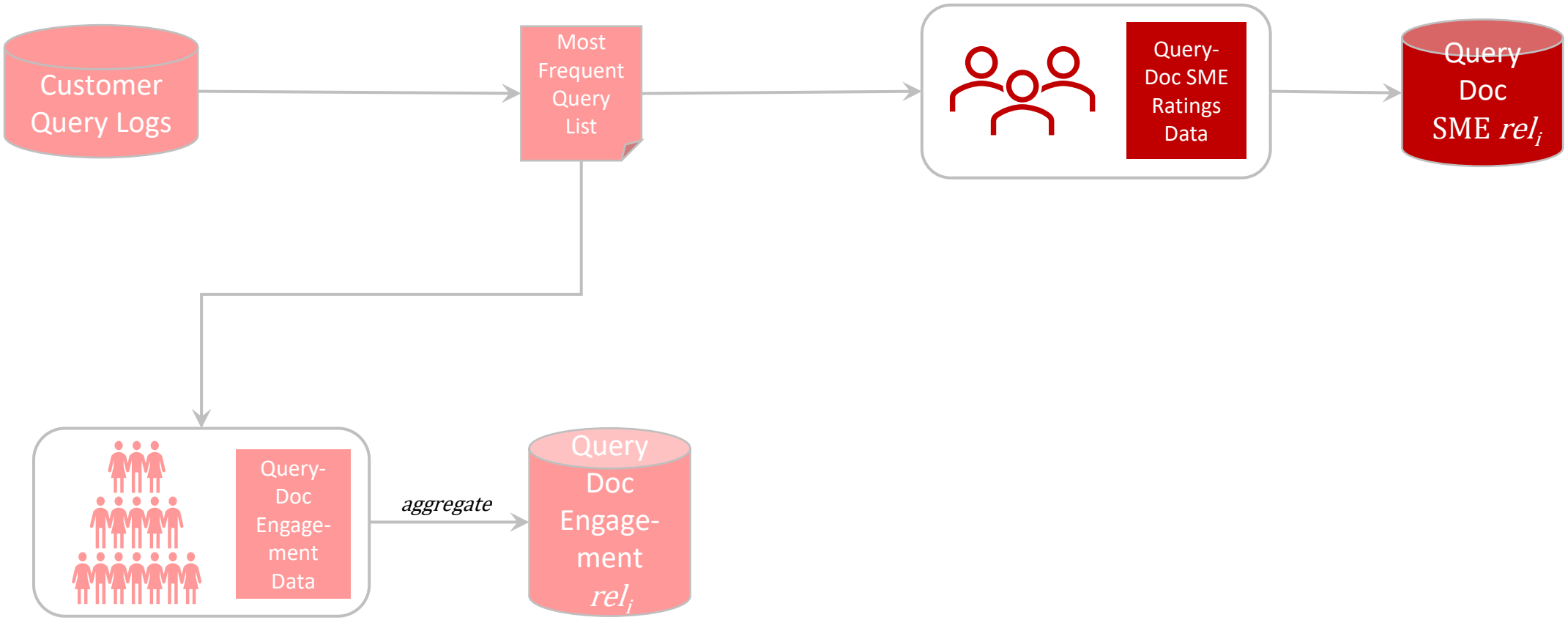
- Convert Interactions to relevance rating via an **engagement model**:
  - Record engagement statistics (along with query, rank, and document information)
  - Thousands of times over a brief (4-6 week) window with a consistent algorithm
  - Aggregate engagement ratings across all users to get an average engagement rel_i for each query

rel_i	Relevance	Sample Engagement Model*
1	None / Weak	Default (no interaction) OR Dwell time lower than 10 seconds)
2	Fair	(Dwell time >= 10s AND Dwell time < 60s) OR Link-out actions
3	Good	(Dwell time >= 60s AND Dwell time < 180s) AND Link-out Action OR Downstream Action OR Review Action
4	Strong	Dwell time >= 180s AND (Downstream Action OR Review Action OR Link-Out)

- Downstream Actions: download, email, print, printer friendly view.
- Link-out Actions: follow link to citation, follow internal document link
- Review Actions: annotate, highlight, save to folder, and share folder saved document.

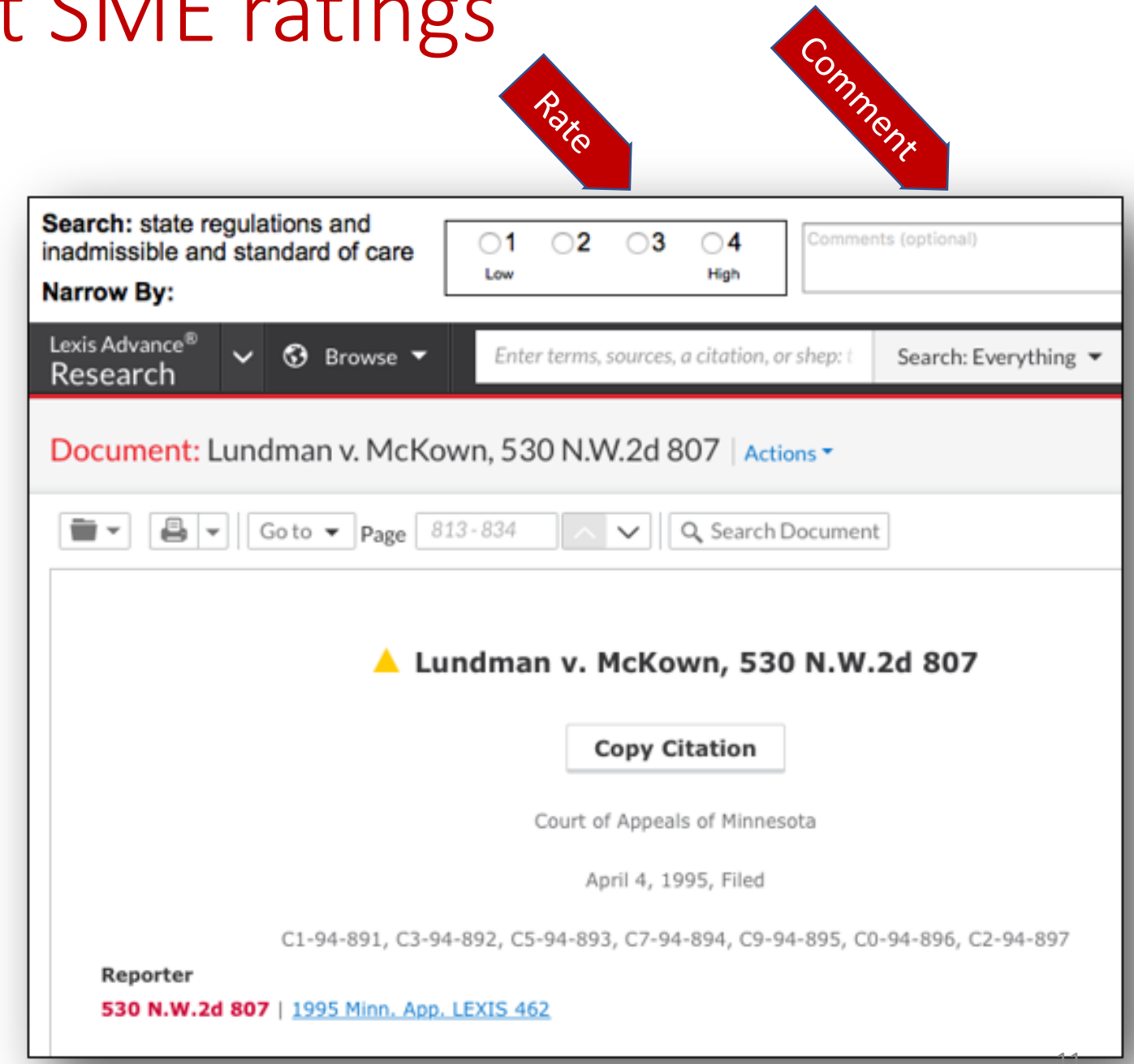
*** Engagement models may be sensitive to changes in user types, content types, query types, and other parameters*

eDCG / hDCG Comparison Flowchart



Collecting Document SME ratings

- SME raters are given a query document pair to review.
- A rating rubric and guidelines are provided to SMEs to assist with consistent application of scores and in handling ambiguous queries.
- Three ratings are collected for each query-document pair and a majority rule (mode) is applied in cases where all three reviews do not agree.

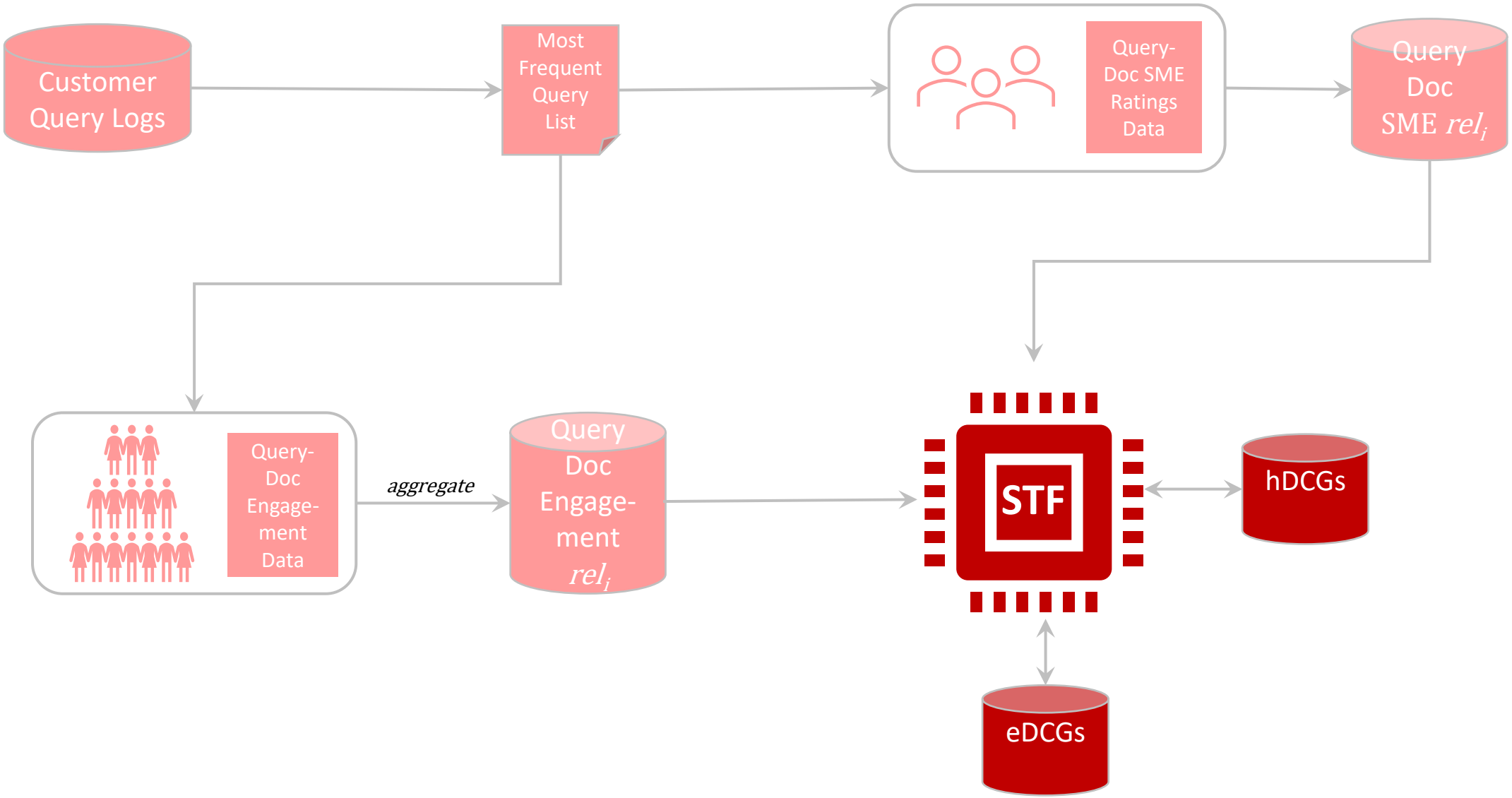


Aggregating results to get SME rel_i

rel_i	Relevance	SME Judgement
4	High	Great search result meets the information need behind query intent / precisely answers the research question that the query poses.
3	Good	Good search results don't quite meet the information need behind the query intent / may be somewhat related and thus still useful.
2	Fair	Fair results are weakly related to the query / may provide a little information. Query Terms may not be related to one another in a meaningful way.
1	Poor	Poor results don't add any value. Results may be confusing, out-of-date, or misleading.

Combined ratings mode for multiple SME raters = rel_i

eDCG / hDCG Comparison Flowchart



Computing engagement DCG (eDCG)

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

p = rank

rel_i = mode engagement value of rating set computed from aggregated results

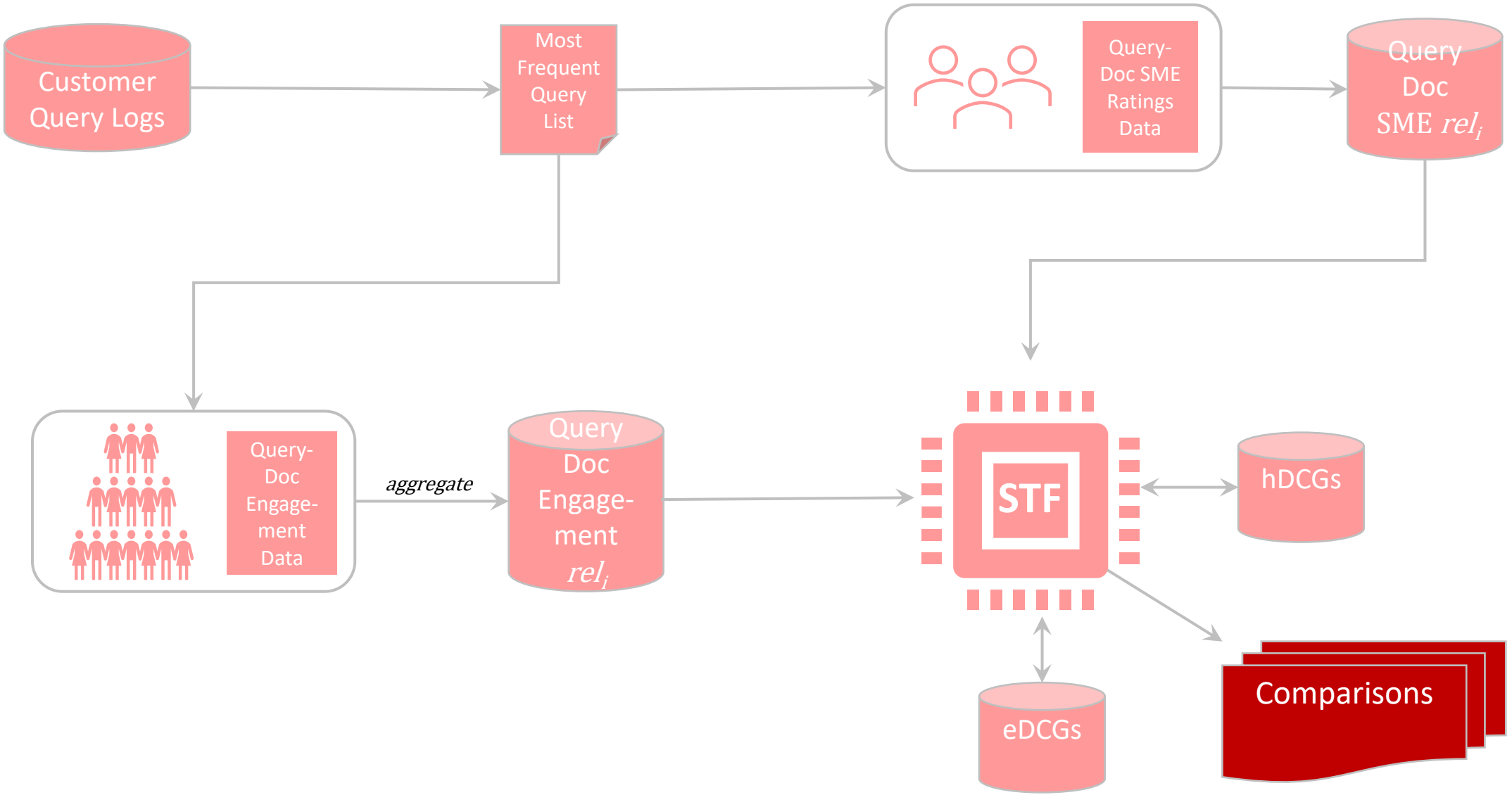
Computing human DCG (hDCG)

$$\text{In } \mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

p = rank

rel_i = Mode of query document ratings provided by multiple SMEs

eDCG / hDCG Comparison Flowchart



Search Test Framework Result Types

Type / Classification	Algorithm	Average hDCG[3]a	Average hDCG[5]a	Average hDCG[10]a
All	Average	11.50	15.93	23.76
Boolean	Average	10.89	15.14	N/A

Query GUID	Query Term	Classification	Widow Query	Query Type	Results Count
LA_CASES_135	black's law dictionary	Phrase, UU	(((black's OR law dictionary OR law w/4 dictionary)))	NaturalLanguageAnd	140121

Algorithm	hDCG[3]oo	hnDCG[3]oo	hDCG[5]oo	hnDCG[5]oo	hDCG[10]oo	hnDCG[10]oo	Sub Value
Optimistic Range	9.92	0.67	13.21	0.64	24.38	0.77	4 ⓘ

Algorithm	hDCG[3]o	hnDCG[3]o	hDCG[5]o	hnDCG[5]o	hDCG[10]o	hnDCG[10]o
Optimistic	9.92	1.00	13.21	1.00	N/A	N/A

Algorithm	hDCG[3]a	hnDCG[3]a	hDCG[5]a	hnDCG[5]a	hDCG[10]a	hnDCG[10]a
Average	9.92	1.00	13.21	1.00	N/A	N/A

Algorithm	hDCG[3]j	hnDCG[3]j	hDCG[5]j	hnDCG[5]j	hDCG[10]j	hnDCG[10]j
Majority/Mode	10.39	1.00	12.85	1.00	N/A	N/A

Algorithm	hDCG[3]m	hnDCG[3]m	hDCG[5]m	hnDCG[5]m	hDCG[10]m	hnDCG[10]m
Median	10.39	1.00	12.85	1.00	N/A	N/A

Algorithm	hDCG[3]sm	hnDCG[3]sm	hDCG[5]sm	hnDCG[5]sm	hDCG[10]sm	hnDCG[10]sm	Sub Value
Substitute Missing	9.92	1.00	13.21	1.00	16.13	1.00	2.5 ⓘ

Algorithm	hDCG[3]p	hnDCG[3]p	hDCG[5]p	hnDCG[5]p	hDCG[10]p	hnDCG[10]p
Pessimistic	9.92	1.00	13.21	1.00	N/A	N/A

Algorithm	hDCG[3]pp	hnDCG[3]pp	hDCG[5]pp	hnDCG[5]pp	hDCG[10]pp	hnDCG[10]pp	Sub Value
Pessimistic Range	9.92	1.00	13.21	1.00	13.21	1.00	1 ⓘ

Average Number Ratings per Query(3)	Average Number Ratings per Query(5)	Average Number Ratings per Query(10)
3.00	3.00	1.50

Document	Average Score ⓘ	Count ⓘ	Standard Deviation ⓘ	Median ⓘ	Mode ⓘ	Range ⓘ	ERR Score ⓘ	Title	Doc-Id
1	3.67	3	0.58	4	4	3-4	N/A	Shiffman v. Auto Source Wholesale, LLC, 2018 Mich. App. LEXIS 2997	urn:contentItem:5T1W-3SR1-JFDC-X25W-00000-00
2	3.33	3	0.58	3	3	3-4	N/A	Emory Univ., Inc. v. Neurocare, Inc., 985 F.3d 1337	urn:contentItem:61VH-9N31-JS0R-2517-00000-00
3	3.33	3	0.58	3	3	3-4	N/A	Iverson v. United States, 973 F.3d 843	urn:contentItem:60R4-S9N1-DYV0-G1GP-00000-00
4	3.33	3	0.58	3	3	3-4	N/A	Jones v. Phillips, 850 S.E.2d 646	urn:contentItem:61G2-HX91-F5T5-M06G-00000-00
5	3.33	3	0.58	3	3	3-4	N/A	Matter of Walsh v New York State Comptroller, 34 N.Y.3d 520	urn:contentItem:5XKD-MV51-JB7K-2214-00000-00
6	0 ⓘ	0	0	0	0	0-0	N/A	Peterson v. City of Minneapolis, 892 N.W.2d 824	urn:contentItem:5N9B-KXN1-F04H-207K-00000-00
7	0 ⓘ	0	0	0	0	0-0	N/A	United States v. Lopez, 590 F.3d 1238	urn:contentItem:7XCG-7180-YB0V-S00F-00000-00
8	0 ⓘ	0	0	0	0	0-0	N/A	Vanderbrook v. Unitrin Preferred Ins. Co. (In re Katrina Canal Breaches Litig.), 495 F.3d 191	urn:contentItem:4PBC-HPN0-TXFX-72C4-00000-00
9	0 ⓘ	0	0	0	0	0-0	N/A	United States v. Rodriguez, 711 F.3d 541	urn:contentItem:580F-STF1-F04K-N1FC-00000-00
10	0 ⓘ	0	0	0	0	0-0	N/A	BOKF, N.A. v. BCP Land Co., LLC, 2016 U.S. Dist. LEXIS 29975	urn:contentItem:5J89-G3J1-F04D-K2KD-00000-00

C	8	8	2
UU	3	3	0
Popular Name	1	1	0

Formal Experiment Statement

Hypothesis 1:

eDCG and hDCG should be near equal across frequently asked NL queries for ranks [3] and [5].

Hypothesis 2:

eDCG and hDCG should be roughly equal across the query set at ranks [3] and [5].

eDCG vs. hDCG Comparison Process 1:

- Requirements
  - Same queries
  - Same search algorithm
  - Same document view UI for raters

- Process
  - Extract ~400 representative queries* from the most frequently asked Natural Language user queries in the customer query list
  - Retrieve & compute document eDCG for each query-document pair in set**
  - Rate the same representative queries using an HRT job and SMEs
  - Retrieve & compute hDCG data for each query-document pair in set**
  - Compare eDCG and hDCG results, query by query, and as statistical composites

* Determined by customer query logs

** down to rank [5]

eDCG vs. hDCG Comparison Process 2:

- Queries: 400 representative queries from most frequently asked eDCG logs
  - 391 Natural Language
  - 9 Boolean*
  - 358 queries with complete results down to rank [5], permitting computation of DCG[5]
  - Also broken down into 16 query type sub-classes

- Query Document Pairs
  - Rel_j Ratings Range: 1 (least relevant) to 4 (best relevance) based on customer engagement or SME rater
  - All query document pairs down to rank[5]

- Metrics Computed: hDCG, eDCG

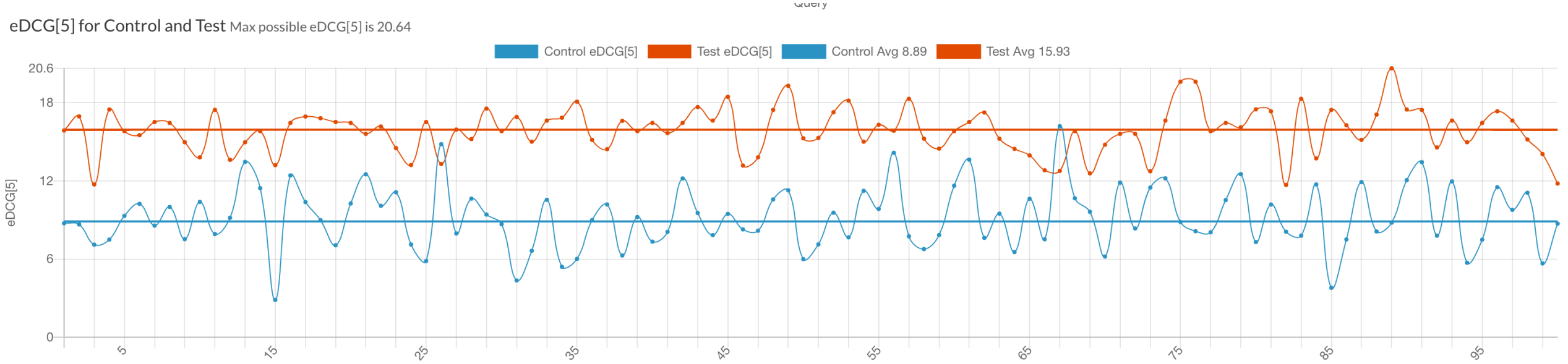
- Tools:
  - HRT Computation and Comparison Tools: STF
  - A/B Testing and Engagement Computation Tool: ABE

eDCG vs. hDCG Comparison Process 3:

- Not considered in this study:
  - Less frequently asked queries
  - Higher rank query-document pairs (q.v. ranks [6] through [10])
  - Boolean queries
  - Multiple Content Types

Raw Results: eDCG[5] and hDCG[5] Comparison

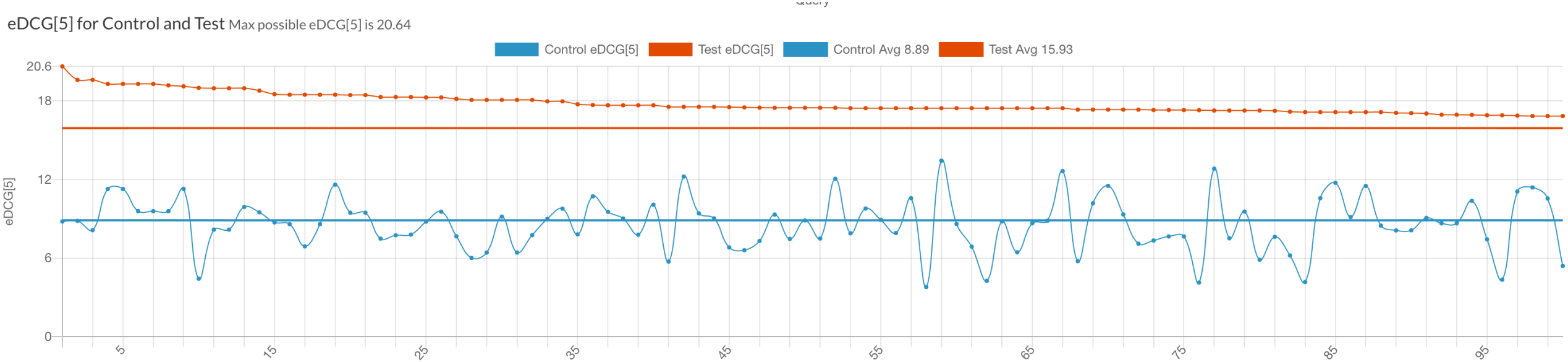
At first blush, eDCG[5] and hDCG[5] seem to track but have some significant offsets.



This is a deceptive conclusion and a result of the brain wanting to see a pattern.

Visualization for Comparison Test Result for eDCG_vs_hDCG-Depth5_358Queries, Queries 101-200

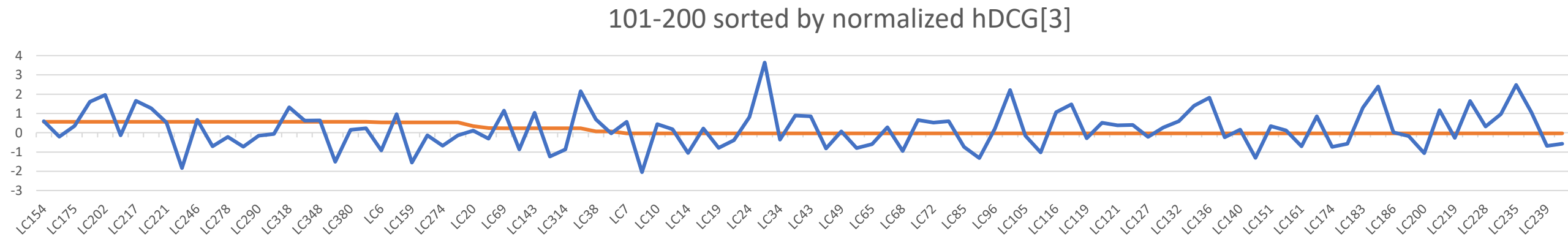
Analysis 1: Sorted eDCG[5] and hDCG[5] Results

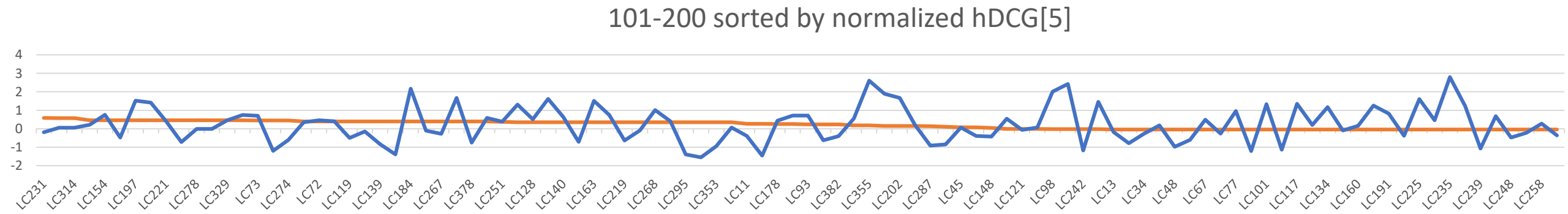


When both data sequences are re- sorted by descending hDCG[5] value, the eDCG[5] control trend line does NOT mirror the sorted Test (hDCG) values.

Analysis 2: Scaled eDCG and hDCG Results

Even when scaled by the average values for both eDCG and hDCG to account for scale magnitude, there is very poor correlation.





The actual correlation coefficient between eDCG[3] & hDCG[3] is 0.014, and between eDCG[5] & hDCG[5] is 0.0221, both nearly zero. An actual value of zero is a perfect non-correlation.

Scaled Average DCG = $x - x_{avg}$ / std dev

Results: eDCG and hDCG Correlation Results

	[3]	[5]	Avg Normalized DCG[3] Correlation Coefficient	Avg Normalized DCG[5] Correlation Coefficient
N_Results	382	359		
Avg eDCG	6.43	8.89		
Avg hDCG	11.47	15.88		
Query Nos. 1-100			0.109	0.093
Query Nos. 101-200			-0.048	- 0.074
Query Nos. 201-300			-0.071	0.078
Query Nos. 301-358			0.138	0.018
Composite: Query Nos. 1-358			0.015	0.021
			0.089 upper err	0.097 upper err
			0.014 (avg)	0.021 (avg)
			0.075 (std dev)	0.076 (std dev)
			- 0.061 lower err	- 0.055 lower err

Correlation Coefficients remain very nearly equal both for query sets and complete results

Experiment Conclusions

- Does eDCG = hDCG at ranks [3] and [5]?
  - No, hypothesis (1) is proven false.

- Does eDCG[3] and eDCG[5] scale to hDCG[3] and hDCG[5]?
  - No, hypothesis (2) is proven false as well.

- How do they relate?
  1. eDCG nearly always has a lower numeric value than hDCG.
  2. eDCG and hDCG have virtually no actual correlation.
  3. Normalizing for scale does not impact the lack of correlation or trend.

Is one metric more correct than the other?

- Since the metrics are not equal and display no apparent shared trends for this dataset, neither metric appears to be more useful or more correct given the data.

- Usage assumptions, metric properties and use of a specific metric may drive the use of some metrics over others.

- Subsequent work to reduce bias and explore other aspects of comparison are needed to recommend use or development of a specific metric.

Potential Sources of eDCG / hDCG Biases

- Query Bias: unusually high DCG average & predominantly NL queries

  (Other query types (Boolean, SDR, Navigational) may show different behaviors)

- Content Types Bias may show different behaviors

  (not tested yet)

- Engagement model Bias: other models may show different behaviors

  (not tested yet due to complexity, lack of correlation to user experience)

- Presentation Bias / Lead Cow Advantage for engagement raters

  (SMEs don't see the results list or make decisions based on rank position)

- Wisdom of the Crowd vs Expert Bias

  (WoC random roles, Competence, and Experience levels vs experienced, expert, raters)

- Imbalance in average rating scores: eDCG 1-4, hDCG 3-4

  (unknown effect)

- Imbalance / variability in number of engagements per query rank vs constant for hDCG

  (may show that WoC assumption not valid all the time)

- Missing rank values in eDCG set reduced number of queries available

  (381/400 and 359/400 for eDCG[3] and [5] respectively)

Conclusions

For this dataset, studied at ranks [3] and [5],

- There is no correlation between eDCG and hDCG, in either un-normalized or scale-normalized form for a large, representative set of frequently run natural language queries.
- eDCG and hDCG cannot be used as proxies for one another.
- EDG appears to be a less stable / more sensitive metric than hDCG on a per query basis.
- Wisdom of the crowd assumptions for eDCG do not appear to be always applicable.
- Neither eDCG or hDCG as a sole metric can characterize success or failure of search precision.
- Additional work is needed to understand how these results generalize to other query types, content types, engagement models (e.g. click models, or semantic relevance models) and relevance metrics.

Next Steps:

- Short Term:
  - Examine Boolean queries
  - Examine other content types
  - Examine impact of changing the engagement model to one more closely attuned to SME rating behavior
  - Extend to other metrics: P(k), ERR

- Longer Term:
  - Random Role, Competence, and Experience vs. Training Bias
  - Explore more seldom asked query sets (but data sparsity issue)
  - Better understand anomalous queries / results significantly impact the overall results
  - Detail comparison testing on datasets where eDCG and hDCG trends strongly agree or disagree
  - Lead Cow Experiment / SERP view impact for hDCG ratings bias

Questions?

Thanks, and a tip of the helmet to the several folks who made this study possible

JOIN US IN SHAPING A MORE JUST WORLD

Around the globe, LexisNexis employees are connected by the desire to shape a better world where the rule of law increases peace, prosperity, and justice. Everything that we do as a commercial business advances the rule of law. A career with LexisNexis can help you make a difference in the communities where we live and work. Join us on our journey today!

SCAN HERE TO LEARN MORE ABOUT OUR OPPORTUNITIES

For more information, email TalentAcquisition@lexisnexis.com

Author Biography

- Name: Hari Bandepalli
- Company: LexisNexis
- Email: venkatahariharan.bandepalli@lexisnexis.com
- Biography:

  Hari is a senior developer for the Search Test Framework (STF) for more than than 4 years.

Author Biography

- Name: Edward Chatelain
- Company: LexisNexis
- Email: Edward.Chatelain@lexisnexis.com
- Biography:

  Ed is the Technical Lead for the Search Test Framework (STF), he has been working on STF since the project was started 5 years ago. Prior to that Ed worked at IBM for 29 years on a variety of interesting projects.

Author Biography

- Name: Sreenath Kottapuzhackal
- Company: LexisNexis
- Email :sreenath.kottapuzhackal@lexisnexis.com
- Biography:

  Sreenath is a Senior software engineer working on the Search Test Framework. Before joining LexisNexis, he was a technology lead at Infosys and has worked on multiple LexisNexis projects for more than 15 years.

Author/Speaker Biography

- Name: Doug Rosenoff
- Company: LexisNexis
- Email: douglas.rosenoff@lexisnexis.com
- Biography:

  Doug is the Product Owner/Manager for the LexisNexis Search Test Framework. He has worked for 28 years in electronic publishing and research at West Publishing and at LexisNexis, with Patents in Search Algorithms and Automatic Linking.