COLING 2022

MICO: Selective Search with Mutual Information Co-training

Zhanyu Wang, Ph.D. Candidate @ Purdue Statistics

Xiao Zhang, Applied Scientist @ Amazon Web Service

Hyokun Yun, Principle Applied Scientist @ Amazon Search

Choon Hui Teo, Principle Applied Scientist @ Amazon Search

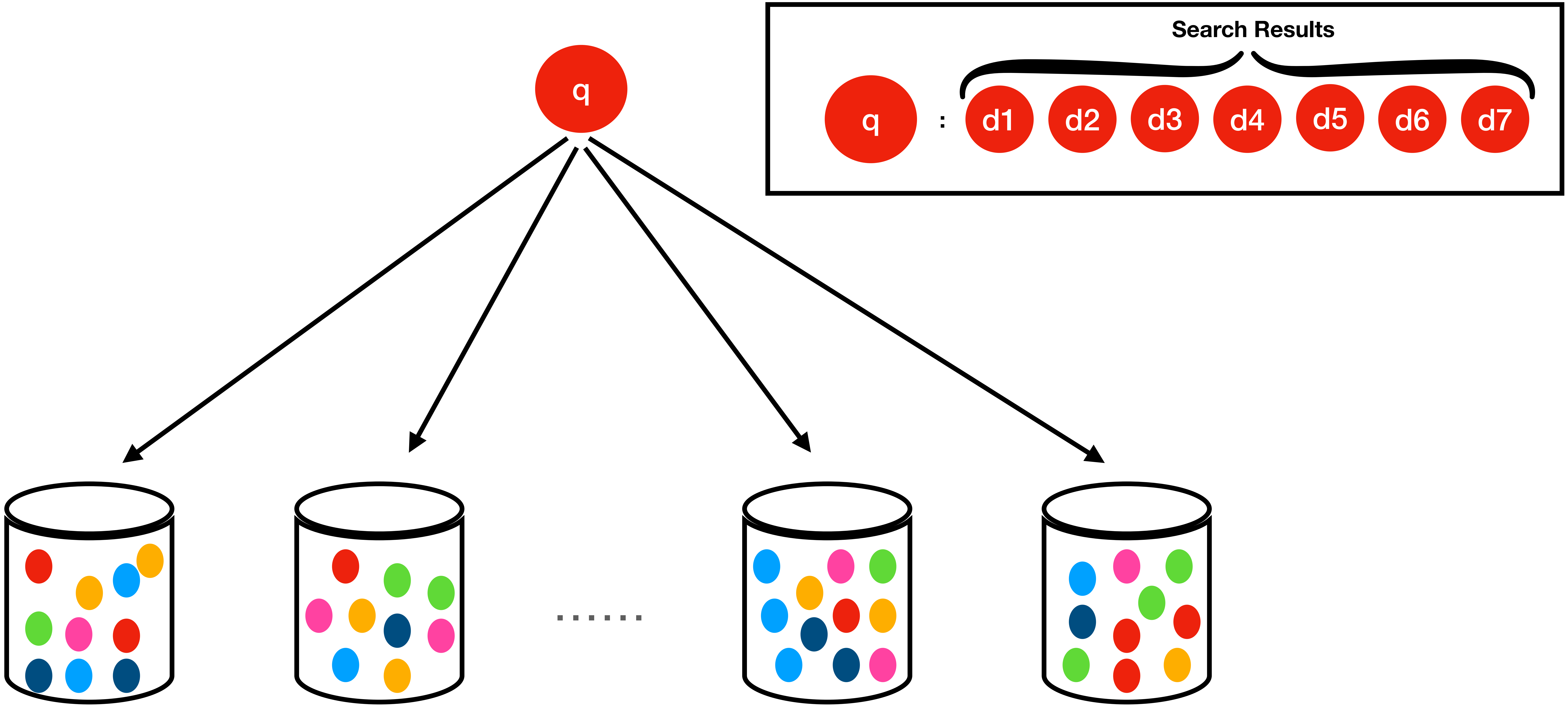
Trishul Chilimbi, Sr. Principle Applied Scientist @ Amazon Search

Outline

- Introduction
- MICO: Mutual Information Co-training
- Experiments
- Takeaways
- Future Directions

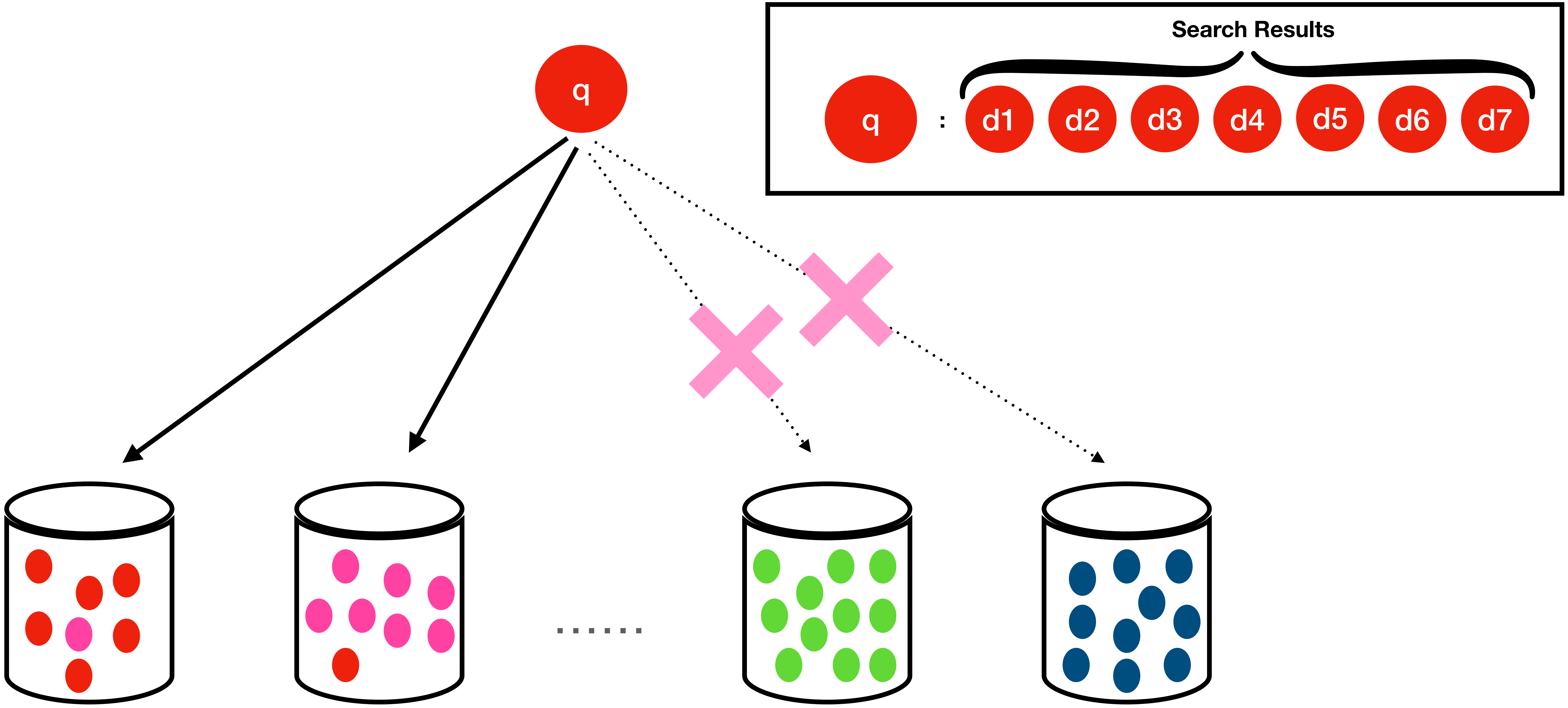
Introduction

Document Sharding

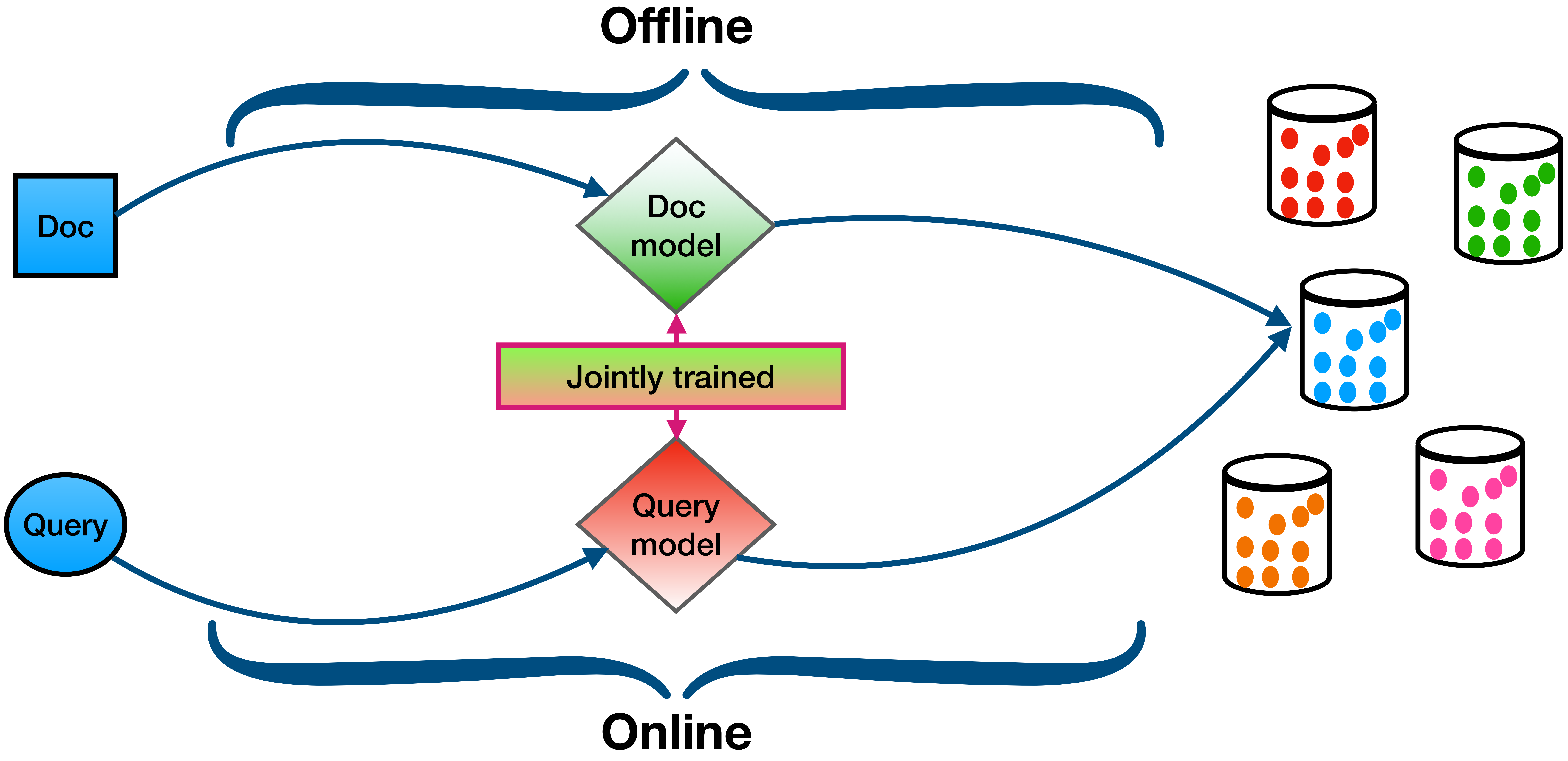


Similar Products Go Together

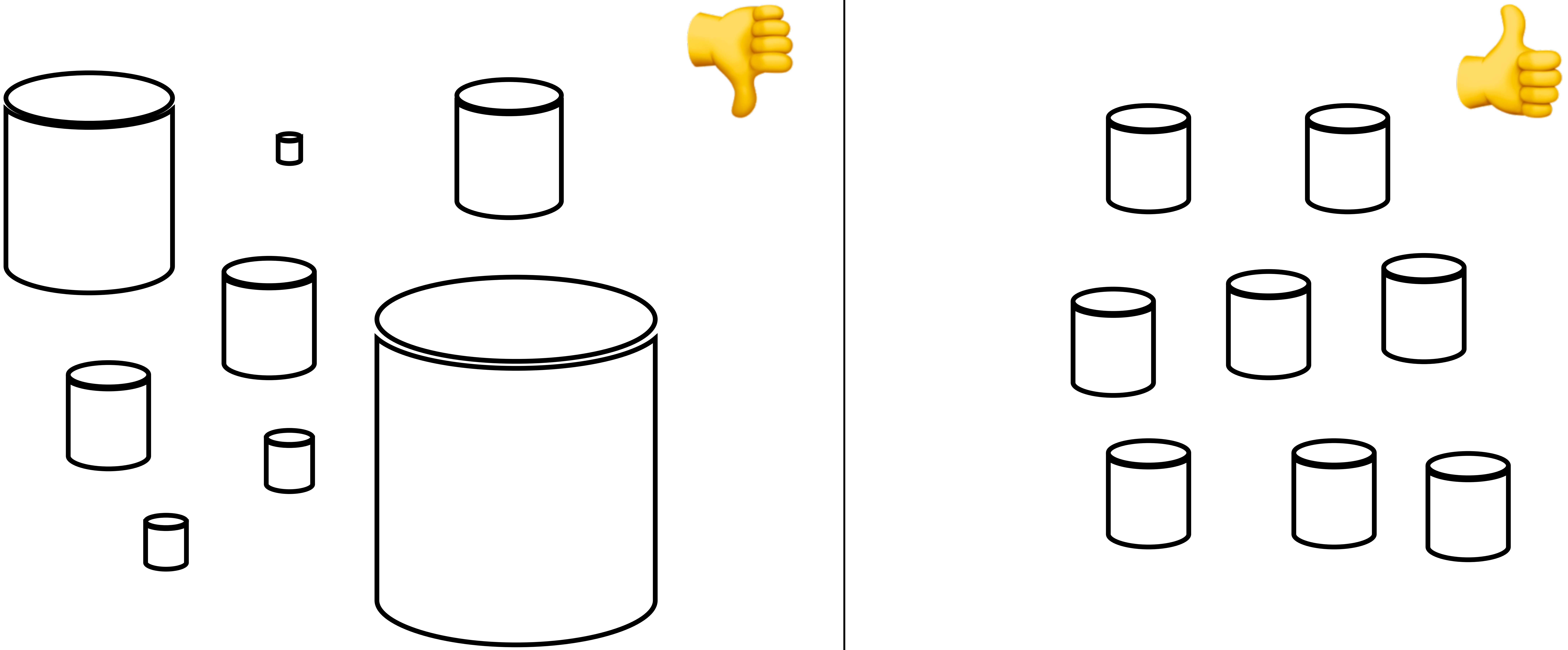
Fanout Reduction



Model Deployment after Training



Cluster Size Balance

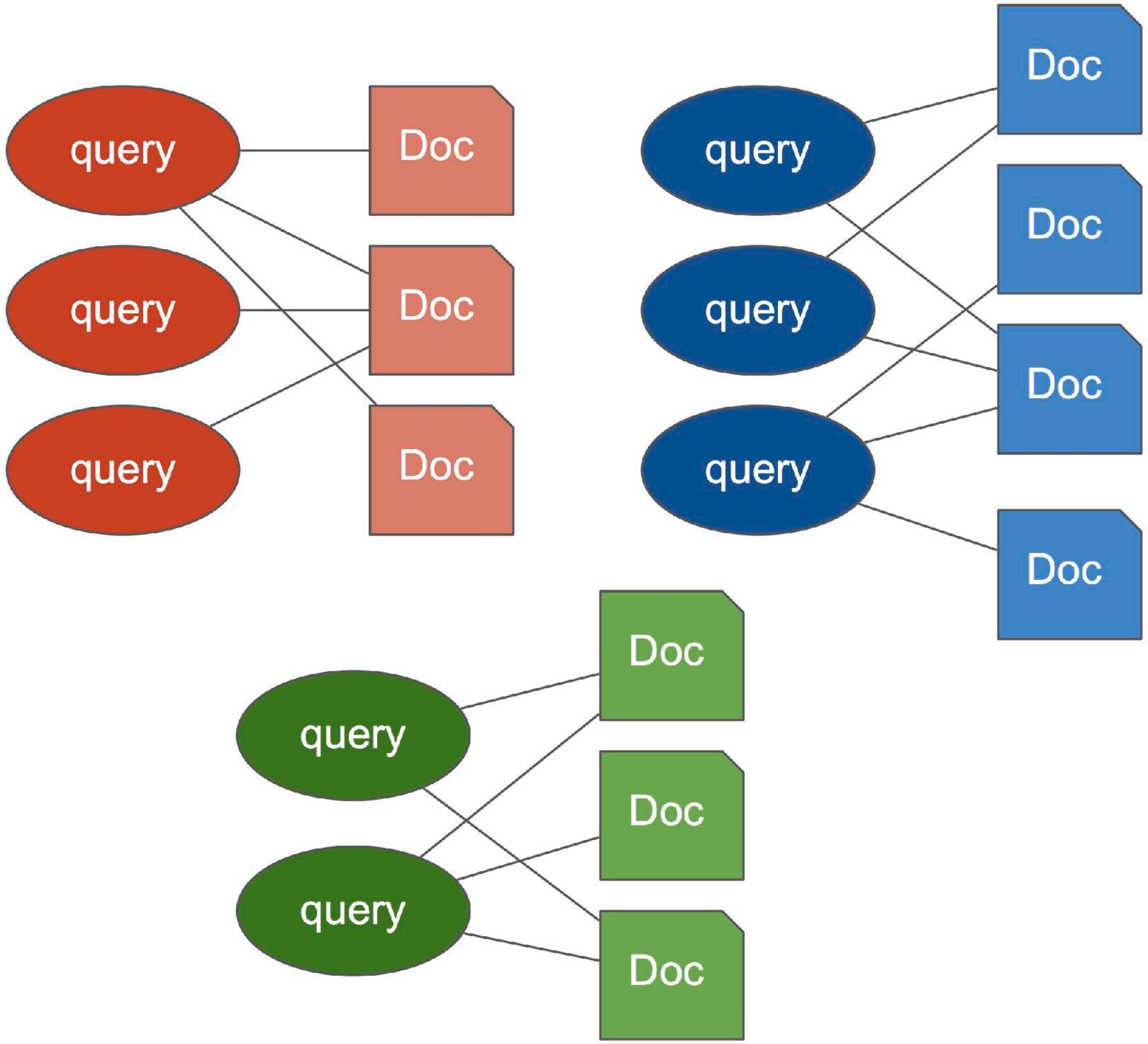


Research Questions

1. How to cluster documents of similar semantical meanings together?
2. How to route the query to the designated shard?
3. Can we achieve the above two in an end-to-end fashion efficiently?
4. Can we ensure the sizes are balanced among different clusters?

Observation

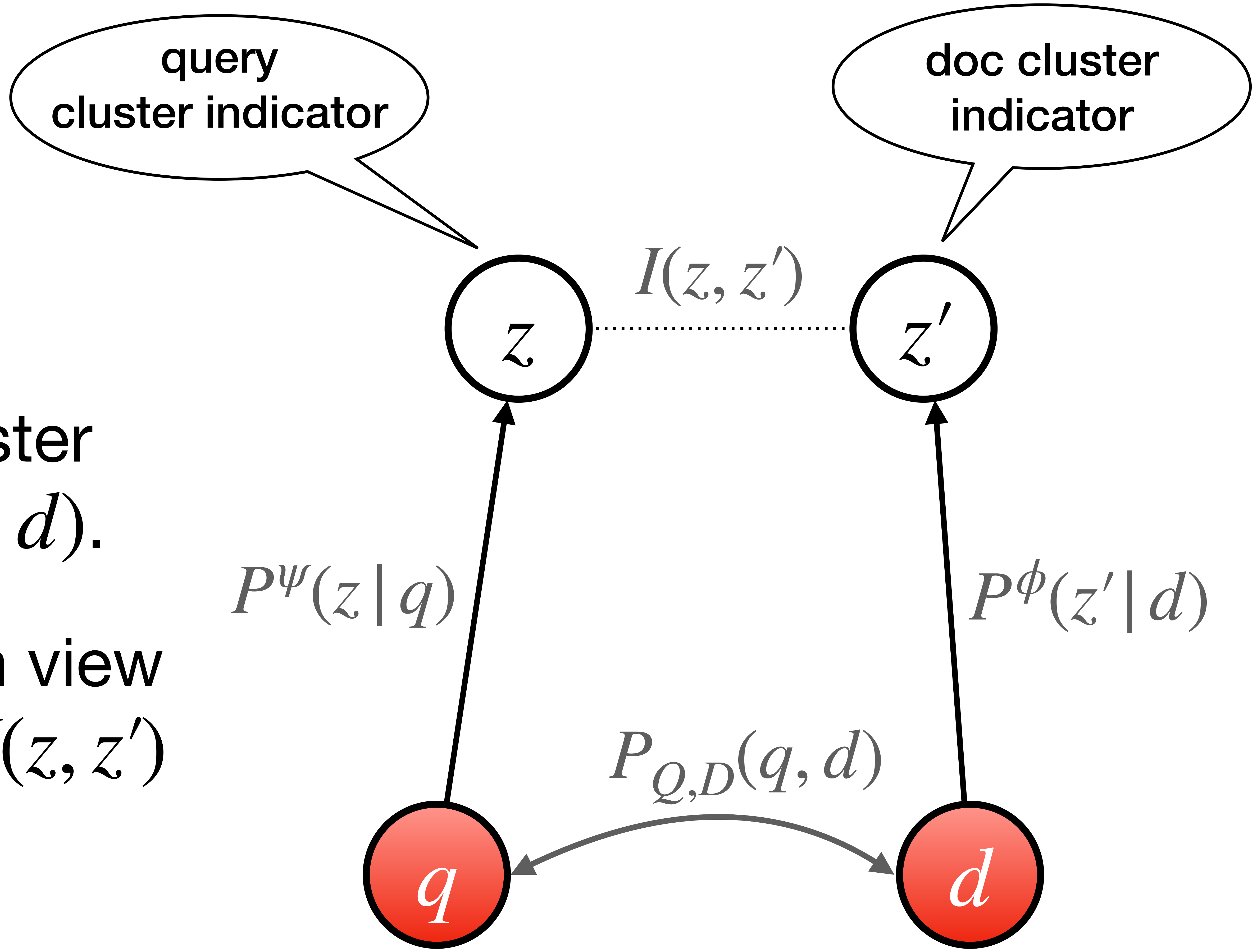
Similar queries related to the same topic, their related documents are also semantically similar.



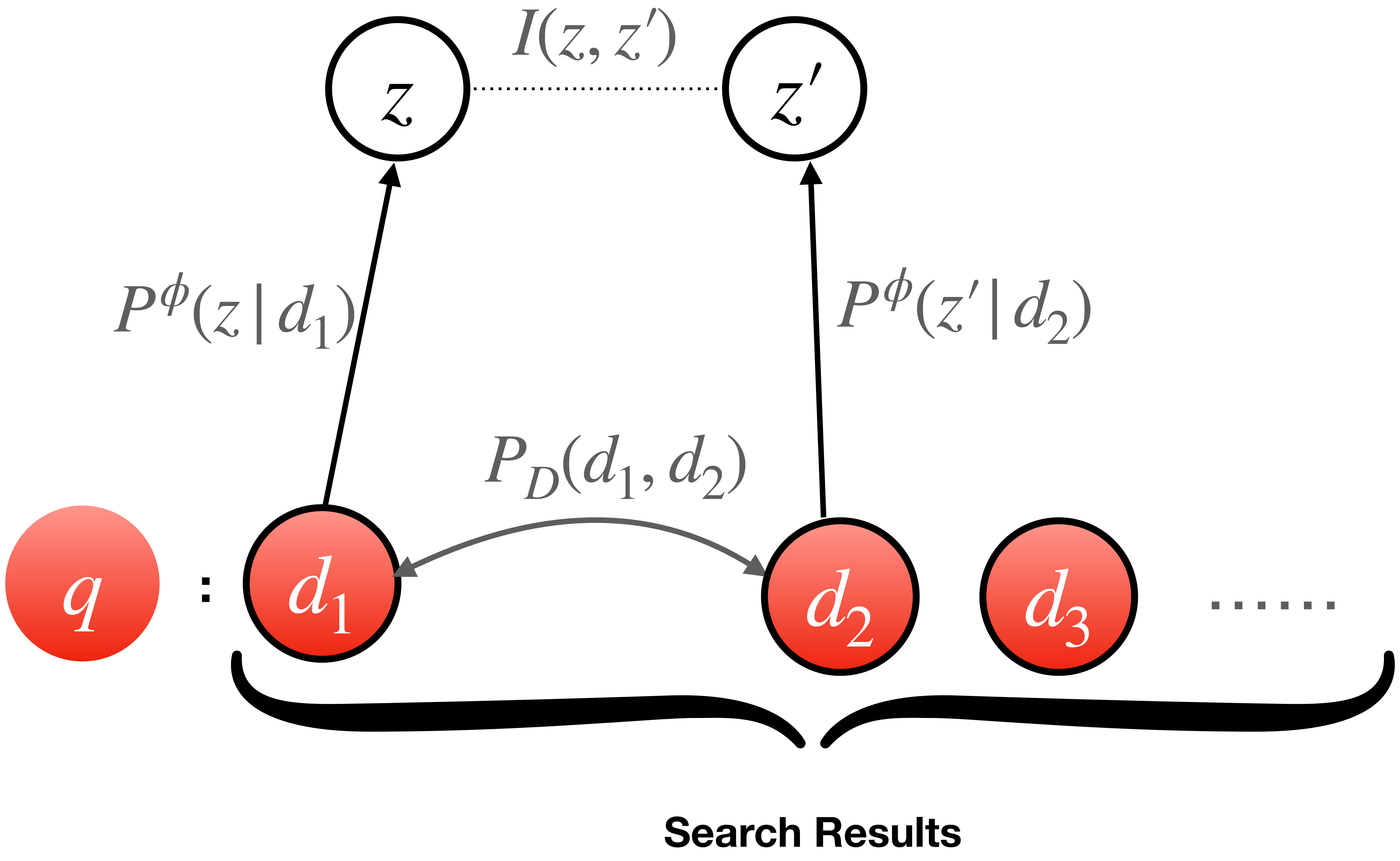
MICO: Mutual Information Co-training

Mutual Information Co-training (MICO)

- Consider the query-document pair as one sample with two views.
- Build two models to predict the query cluster $P^{\phi}(z\,|\,q)$ and the document cluster $P^{\psi}(z'\,|\,d)$.
- Force the results to be consistent for each view by encouraging large mutual information $I(z, z')$ between the cluster indices z and z'.
- Entropy regularization to ensure balanced cluster sizes.



MICO-q: MICO with Query Consistency



Experiments

Results of Query Coverage: E-Commerce

	impression		*click*		*purchase*	
Models	N=1	N=10	N=1	N=10	N=1	N=10
Random	1.56 (6e-3)	15.62 (0.02)	1.49 (0.08)	15.32 (0.85)	1.45 (0.24)	14.54 (0.27)
K-means	48.98 (1.60)	79.05 (0.51)	51.90 (1.56)	81.57 (4.0)	54.49 (1.97)	83.58 (1.49)
B-K-means	39.72 (1.12)	64.56 (1.30)	43.89 (2.03)	64.25 (1.78)	49.02 (2.37)	69.59 (1.22)
IMSAT	41.68 (0.55)	71.37 (0.28)	47.48 (1.62)	79.12 (2.94)	52.41 (0.42)	79.83 (1.06)
KLD	43.46 (5.91)	69.87 (5.34)	44.94 (8.04)	71.17 (5.55)	46.77 (9.32)	70.5 (4.08)
QKLD	**86.14** (8.85)	93.96 (0.77)	73.72 (7.25)	81.89 (1.2)	75.79 (7.22)	83.56 (1.57)
MICO	67.09 (0.20)	92.85 (0.12)	**82.85** (1.51)	97.81 (0.19)	**81.21** (0.49)	96.61 (0.14)
MICO-q	69.81 (0.34)	**94.28** (0.09)	82.48 (1.91)	**98.26** (0.20)	81.15 (1.23)	**97.25** (0.16)

This table shows the performance of query coverage (recall) of MICO, MICO-q, and different baselines over three different query-document relationships on the ECSL data set. We show the performance by only probing the top-1 most relevant shard and the top-10 most relevant shards given a query. The number in the parenthesis right next to the coverage is the standard deviation over five runs. We observe other than the impression relation in which QKLD has the best performance, MICO or MICO-q beat all the baselines.

Results of Query Coverage: Cross-Lingual IR

	fr		*it*		*ta*		*sw*	
Models	*DL*	*PL*	*DL*	*PL*	*DL*	*PL*	*DL*	*PL*
Random	10.02 (0.07)	9.72 (0.16)	10.02 (0.09)	10.0 (0.35)	9.88 (0.93)	9.86 (0.48)	10.01 (0.23)	10.0 (0.67)
K-means	12.19 (1.99)	10.79 (2.04)	14.91 (2.46)	16.36 (3.55)	16.25 (2.5)	21.08 (3.46)	21.71 (4.84)	18.55 (3.65)
B-K-means	12.2 (1.82)	11.44 (1.32)	12.45 (3.13)	12.46 (4.59)	12.78 (2.85)	11.23 (3.84)	11.71 (1.29)	12.16 (1.21)
IMSAT	19.77 (9.53)	19.84 (9.68)	40.09 (8.91)	40.13 (8.88)	12.72 (4.22)	11.89 (3.62)	8.4 (2.24)	8.6 (2.63)
KLD	38.6 (6.02)	40.65 (7.58)	60.94 (5.25)	61.83 (3.12)	**66.53** (8.43)	59.77 (7.18)	21.11 (3.52)	24.83 (3.81)
QKLD	17.76 (3.63)	18.82 (2.08)	18.9 (4.91)	17.45 (4.12)	23.65 (6.81)	24.4 (5.54)	12.23 (0.75)	16.45 (2.25)
MICO (sv)	44.93 (3.47)	**53.12** (2.17)	58.08 (1.22)	65.83 (1.06)	63.55 (4.45)	60.94 (4.92)	26.0 (3.51)	**28.67** (3.73)
MICO-q (sv)	**47.9** (2.68)	48.04 (3.44)	**75.27** (3.6)	**75.01** (4.39)	63.91 (5.3)	**61.29** (5.31)	**27.42** (3.37)	28.14 (2.54)

This table shows the performance of query coverage of MICO, MICO-q, and different baselines on two different query-document relationships on the CLIR data set by only probing the most relevant shard given a query because we only divide the documents into ten shards. The number in the parenthesis right next to the coverage is the standard deviation over multiple runs. sv stands for separate vocabularies for the queries and documents, as in cross-lingual retrieval, the source language and the target language have different vocabularies, and separate vocabularies perform better than unified ones empirically. MICO and MICO-q beat all the baselines except DL in ta.

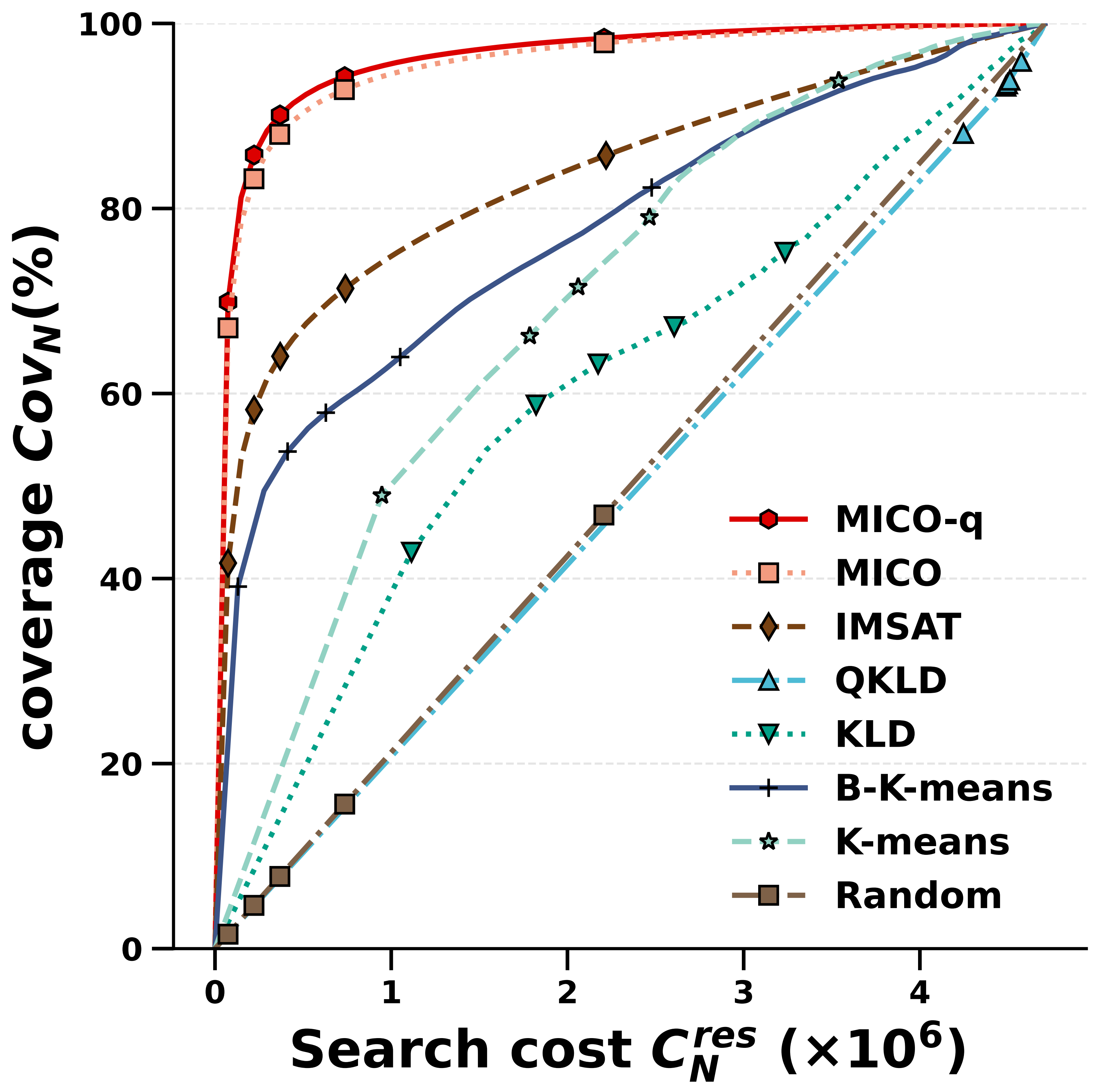
Cost Analysis: Two Standard Metrics

1. C_N^{res}: **Search Resource Cost**
2. C_N^{lat}: **Search Latency Cost**

	C_N^{res}					C_N^{lat}				
Models	ECSL	C-fr	C-it	C-ta	C-sw	ECSL	C-fr	C-it	C-ta	C-sw
K-means	2.061	14.12	11.54	1.6	1.42	1.572	6.44	5.95	0.95	1.27
Balaced K-means	0.620	8.1	6.83	0.99	0.87	0.277	2.58	2.02	0.34	0.67
IMSAT	0.370	9.57	5.89	0.93	0.61	0.082	3.57	4.78	0.58	0.53
KLD	2.17	17.48	13.15	1.93	1.09	1.41	13.26	11.43	1.72	0.74
QKLD	4.5	8.84	7.42	1.34	0.99	4.47	3.72	3.07	0.94	0.66
MICO	**0.367**	**6.19**	**5.13**	0.85	0.93	0.089	**2.34**	**1.89**	0.5	0.5
MICO-q	0.369	7.12	6.71	0.94	1.07	0.093	2.73	2.47	0.51	0.58
Random	0.368	7.20	5.95	**0.8**	**0.73**	**0.074**	2.41	1.99	**0.27**	**0.25**

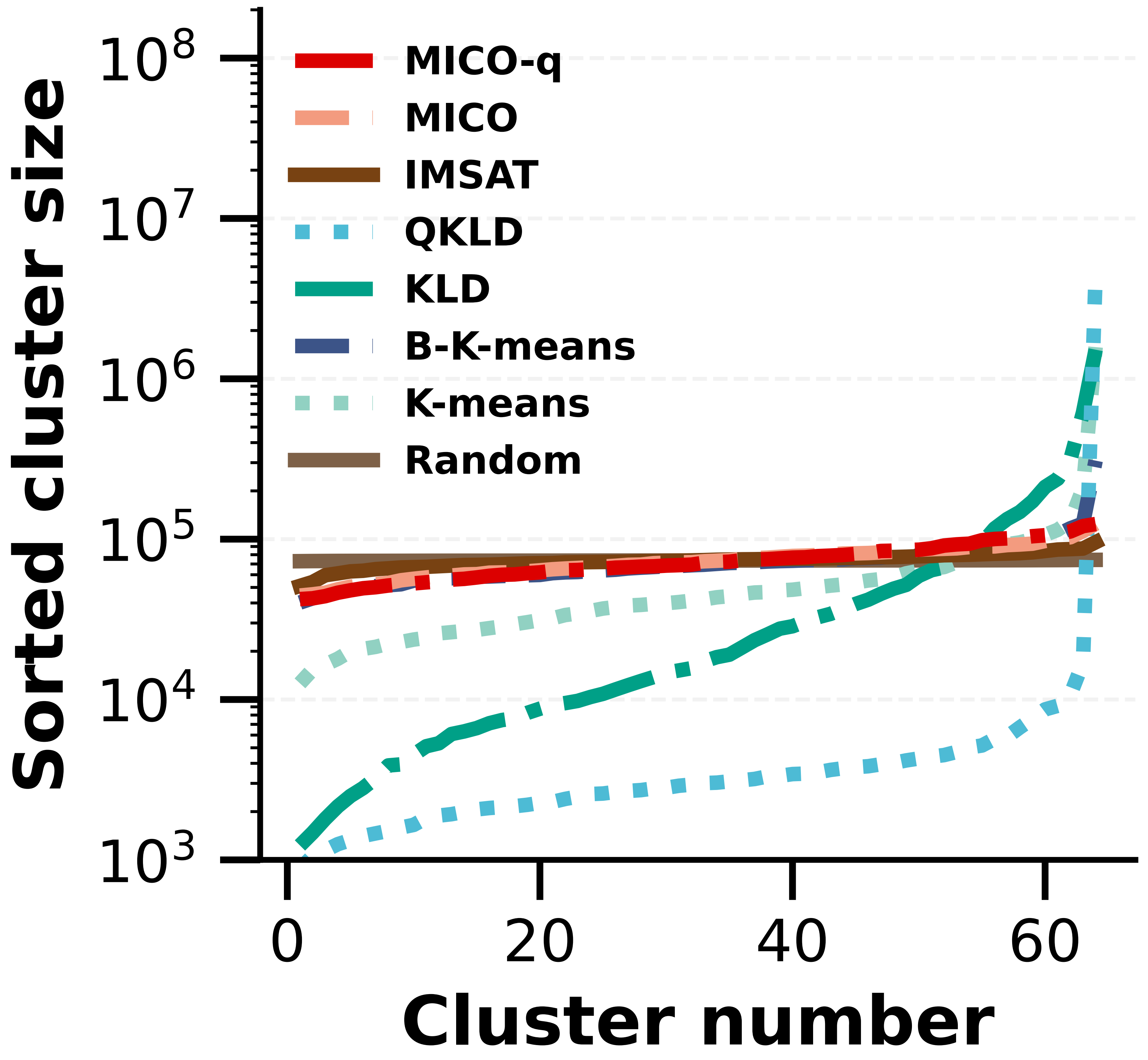
Table 3: This table shows the performance of different models on the *Search Resource Cost* and the *Search Latency Cost* metrics, representing the search efficiency, with the lower the number, the better the performance. The results shown in this table are scaled by being divided by 10^6 on the ECSL data set and by 10^4 on the CLIR data set. Note in this set of experiments, we use separate vocabulary (*sv*) for MICO and MICO-q on CLIR. We observe the supreme performance of MICO, which in some cases even beats the **Random** skyline.

Coverage v.s. Search Resource Cost



This figure shows MICO and MICO-q are significantly better than all other methods as they have high impression coverage with low search cost. From bottom-left to top-right, the markers on each line rep- resent query coverage limited within the top-1, top-3, top-5, top-10, and top-30 clusters selectively.

Balance Among Shard Sizes



This figure shows Random generates the most balanced shard sizes (as a flat line), and IMSAT also creates very balanced shards. MICO and MICO-q are on a par with IMSAT. In contrast, QKLD and KLD yield very unbalanced shards.

Takeaways

Conclusions

- MICO models the problem by treating the query and the document as two different views of the same sample, maximizing the mutual information between the latent categorical variables of each view.
- We design MICO ready for practical use such that it is being trained in an end-to-end manner for both document sharding (clustering) and subsequent query routing.
- We show significantly improved performance on the E-commerce and Cross-lingual IR data set with MICO on multiple important metrics for selective search empirically, suggesting its potential value selective search.
- https://github.com/aws/selective-search-with-mutual-information-cotraining

Future Directions

Next

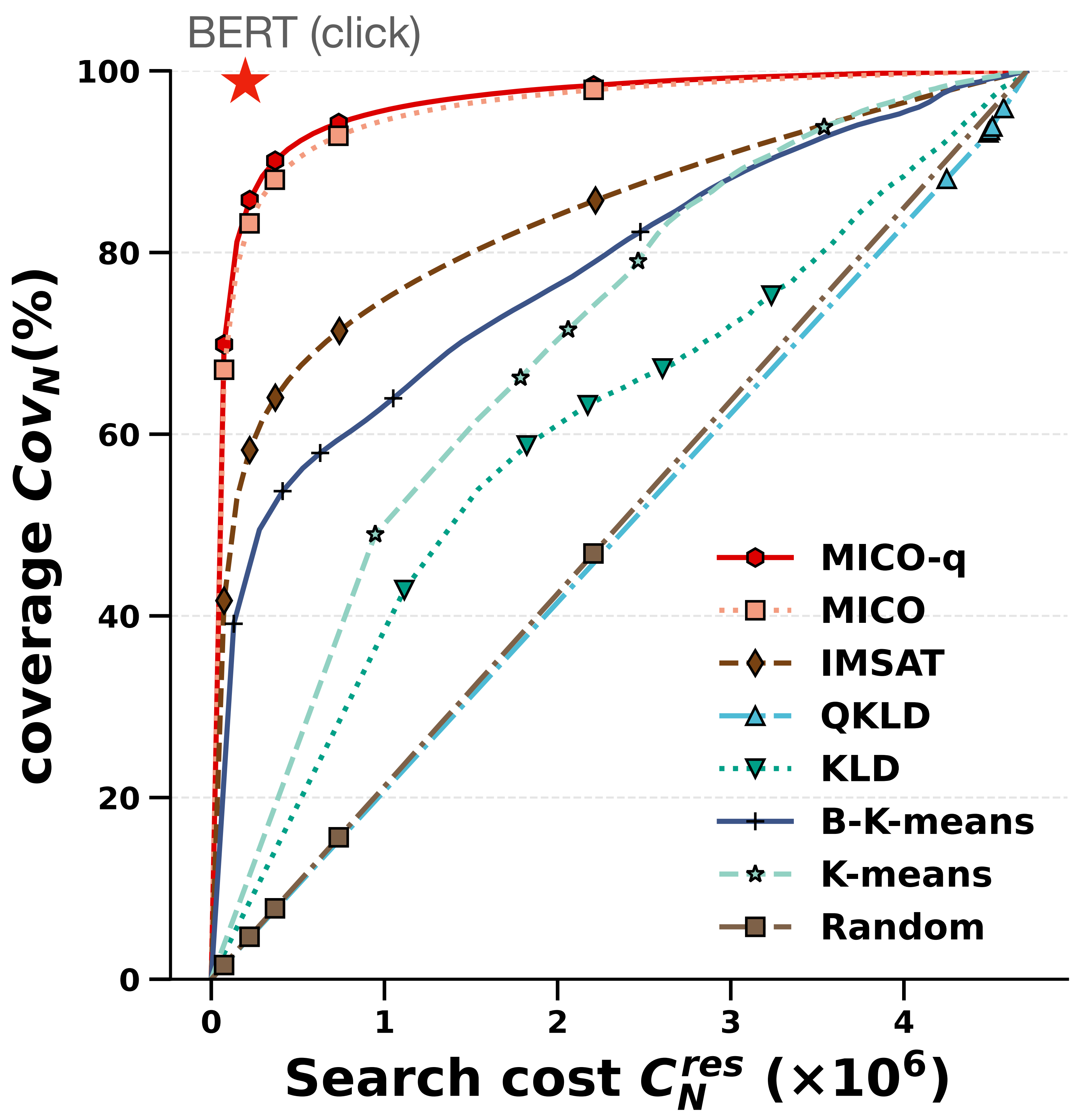
- Richer text representations, e.g., BERT (Hugging Face)
- Multi-modal search data where query and doc are of different modality, e.g., image search
- Detection Policy for when to retrain the model in production

Preliminary Results with BERT

Model	QC	$C^{res}_{N=1}$
MICO (-Par)	66.08	0.365
MICO	67.09	0.367
MICO (+BERT.fx)	41.67	0.367
MICO (+BERT.ft)	**76.41**	0.375

Table 4: This table shows MICO with neural architecture variants. BERT with fine-tuning achieves better performance than the original MICO, while the other variants yield deteriorated performance. The search cost is slightly higher with the best-performing system. We attribute that the refined representations cause the model to weigh more on semantic similarity than cluster balance.

Coverage v.s. Search Resource Cost (w/ BERT)



Using BERT, we reduce the search cost to 5% with achieving 99% accuracy (on retrieving the products ‘shown to & clicked by’ our customer) compared to searching on all documents.

THANK YOU!