Learning to Rank @ Reddit

It Us

Doug Turnbull

Chris Fournier

Cliff Chen

http://softwaredoug.com
http://reddit.com/u/softwaredoug

reddit

Today’s Topic

How do we add Learning to Rank to an existing, mostly working, high scale search system?

Reddit Search information?


```
{
  "timestamp": "2019-08-22T14:38:02.994Z",
  "title": "is numpy.array() of a numpy.array again a numpy.array?
  "body": "     a = np.array([1.0,2.0,3.0])\n     a2 = np.array(a)\n
  "num_votes": 1,
  "num_comments": 5,
  ..
}
```


First glance: classic, text-heavy informational search

… but with a social twist

Breaking news searches, ie “key bridge collapse”


```
{
  "timestamp":"2024-03-26T08:21:12.565Z",
  "title":"The Francis Scott Key Bridge in Baltimore h
  "body":"",
  "num_votes":41524,
  "num_comments": 1234,
  "subreddit_name": "/r/news"
  ...
}
```


Care about recency / popularity

Classic “what’s the absolute latest on this topic” information need

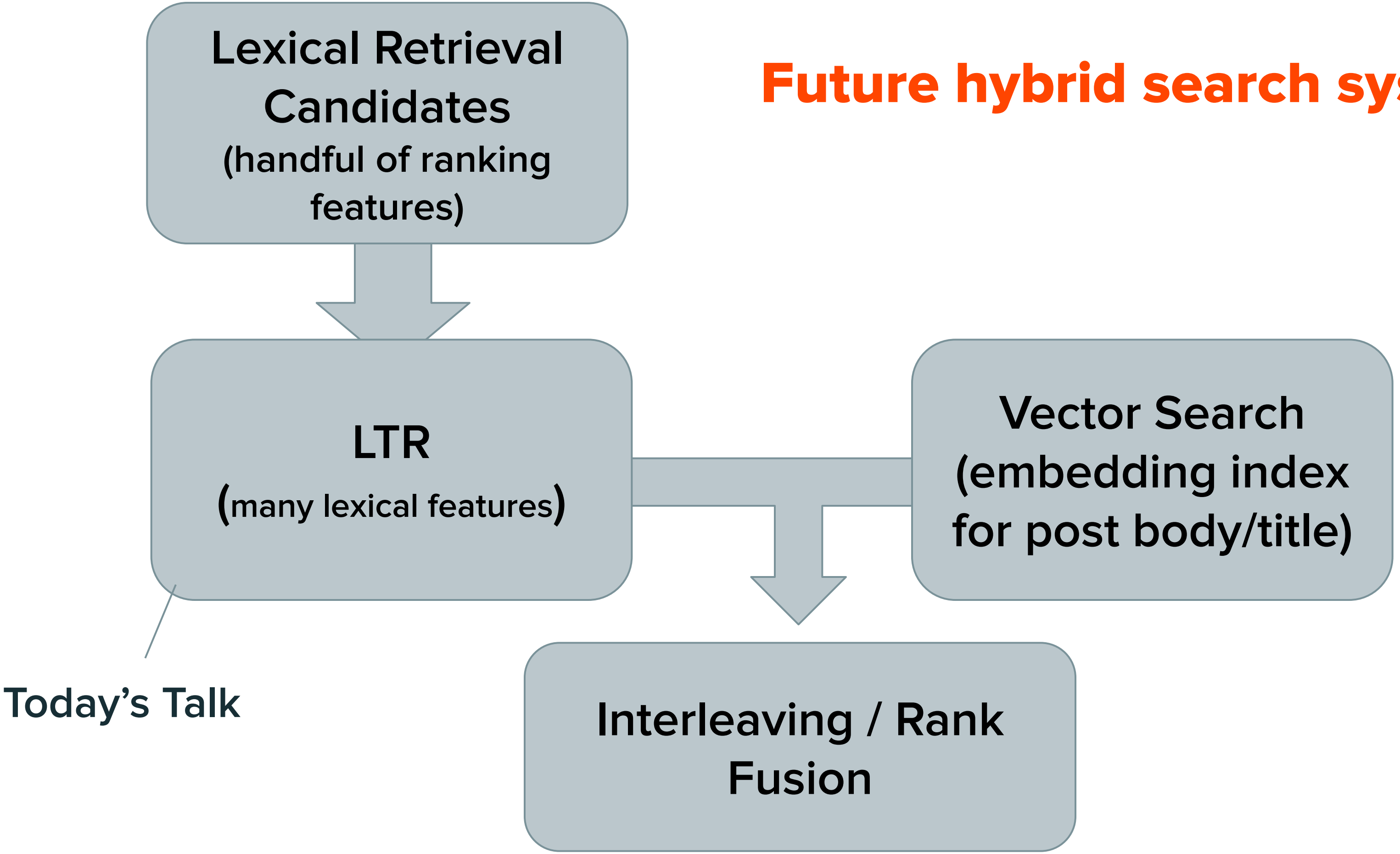
… and sometimes very personal


```
{
  "timestamp":"2024-02-22T17:18:24.789Z",
  "title":"Travel anxiety help",
  "body":"Looking for potential suggestions besides RX medications to help
  "num_votes":458,
  "subreddit_name": "/r/goldenretrievers",
  ..
}
```

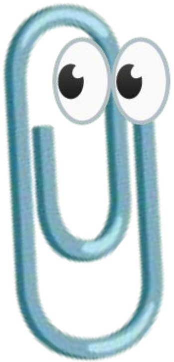

Reddit is a massive repository of
subjective human experience

(This is the big ‘add Reddit to your Google search’ use case)

Future hybrid search system



LTR over 'lexical' - Why do we care?



Mr. ML Model

Hi! I'm

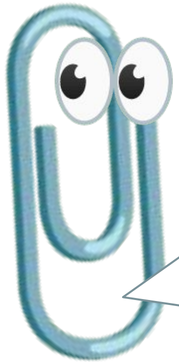
Mr. ML Model!

It looks like you're trying to optimize your search relevance!

Training Data

Query	Post ID	Rel?
Key bridge	1234	1
Key bridge	5678	0
Golden retrieval travel anxiety	12412	1

(are these any good!?)



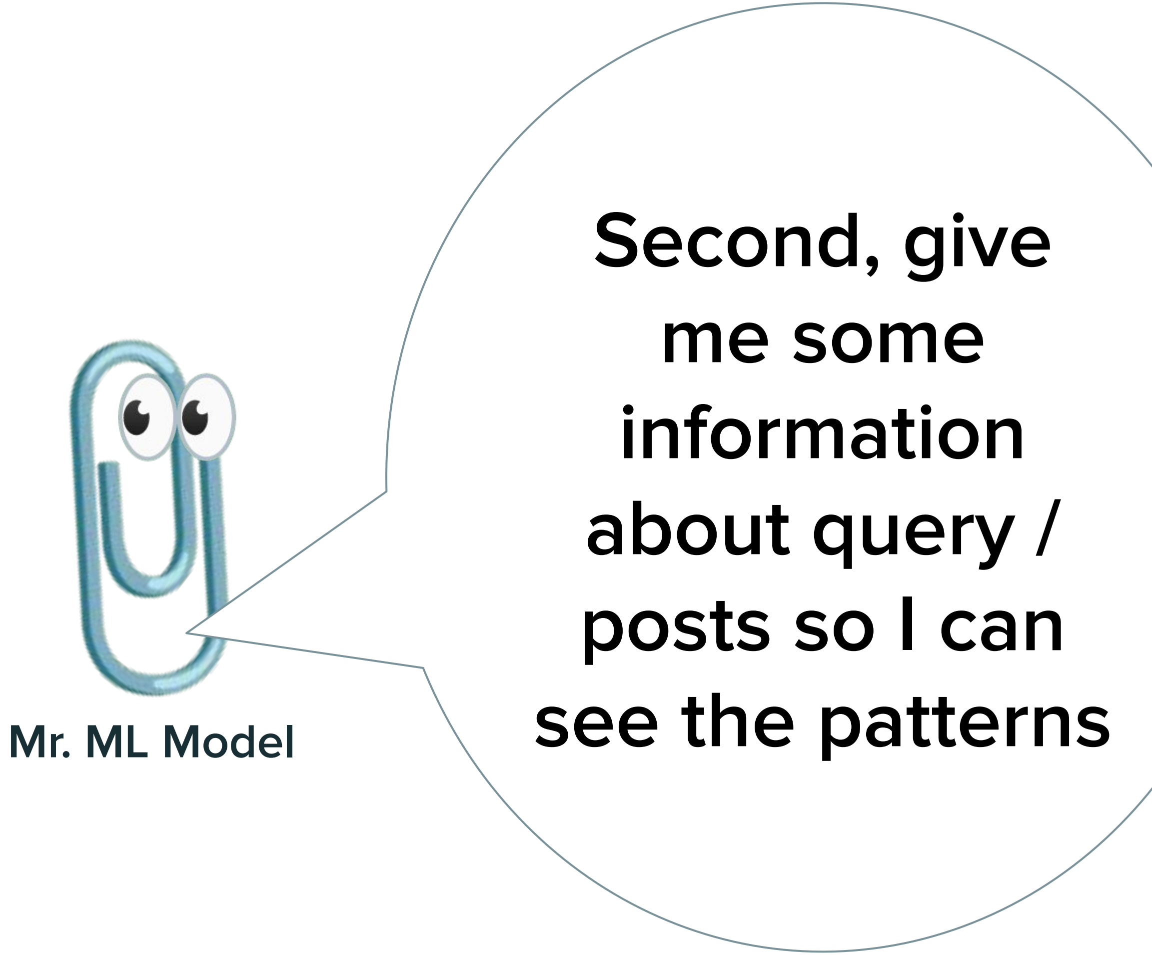
Mr. ML Model

First, give me some examples of relevant / irrelevant search results

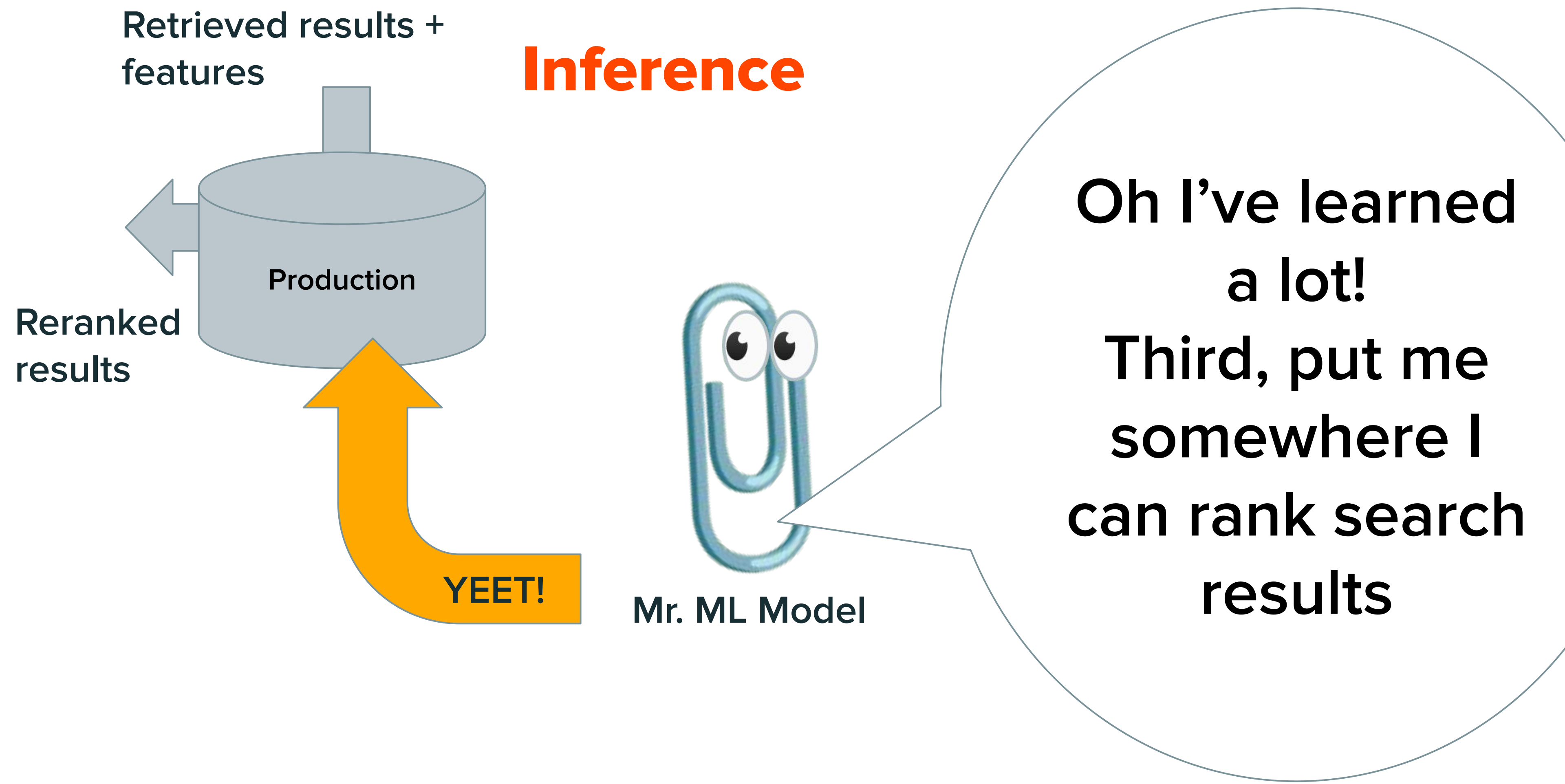
Features

- Did the title match the keywords?
- What was the BM25 score of the body?
- How recent was it?
- Did the subreddit match the query?
- ...
- ?

(do these predict relevance!?)



Inference


Retrieved results +
features
Production
Reranked
results
YEET!
Mr. ML Model
Oh I’ve learned
a lot!
Third, put me
somewhere I
can rank search
results

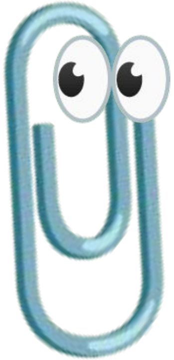
Answering Mr. ML Models questions as a forcing function

💩

Training Data

💩

Features



Mr. ML Model

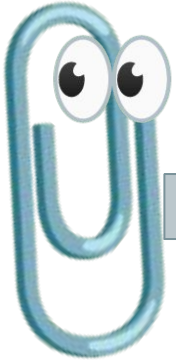
💩

Garbage results

Answering Mr. ML Models questions as a forcing function

Training Data

Features

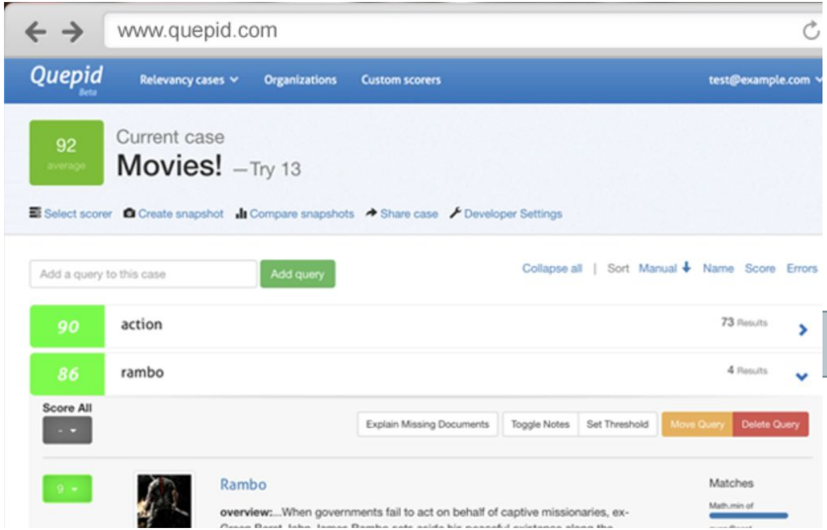


Mr. ML Model

… Even without Mr. ML Model

Training Data

Features



(hand tuned features to meet training data in tool like Quepid)

Training Data + Feature Selection

Learning to Rank, in a nutshell

Train LTR model

5% of time spent

95% of time spent

Having any idea WTF we’re doing in offline eval

Doug, having no idea what he’s doing, until we run more real experiments in search bench

Training Data - started with human eval

Hand labeled results (~1000 queries, 20 per query, head and tail queries)

q=zoolander

👍 Zoolander 2 Trailer

👎 Meet my puppy name “Zoolander”

👍 I love the part where he does “Magnum”

… To derive "engagement judgments"

Relative weights

Position 0.05

Click 0.05

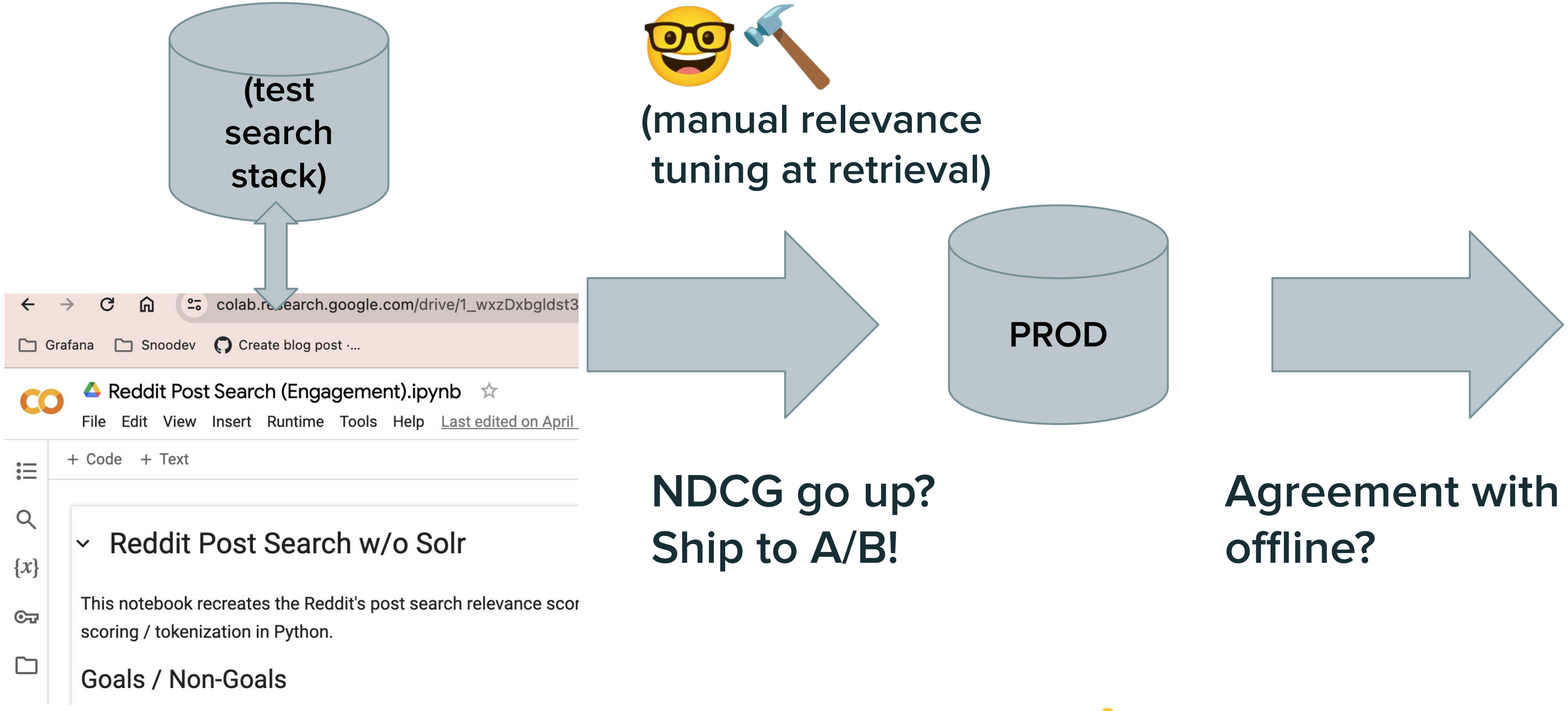
Click + dwell 0.9

- Good sign!

Human labelers agree w/ click+dwell

(30 / 60 day sums)

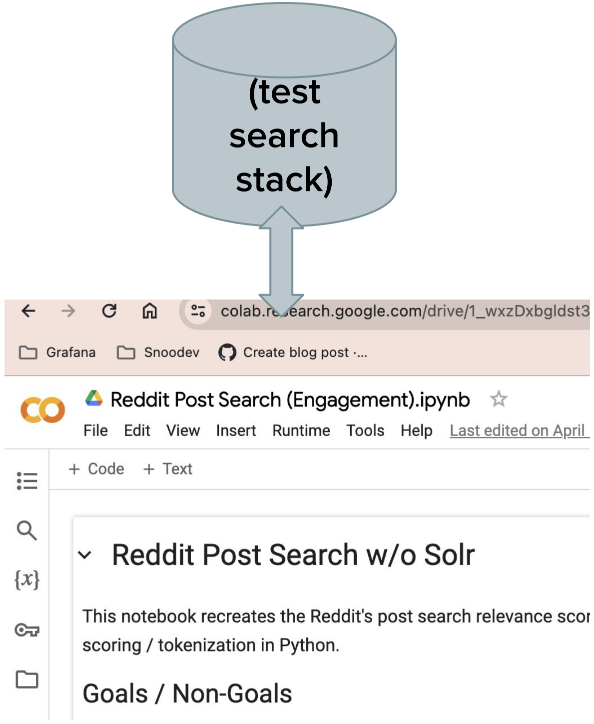
Next steps - USE the judgments



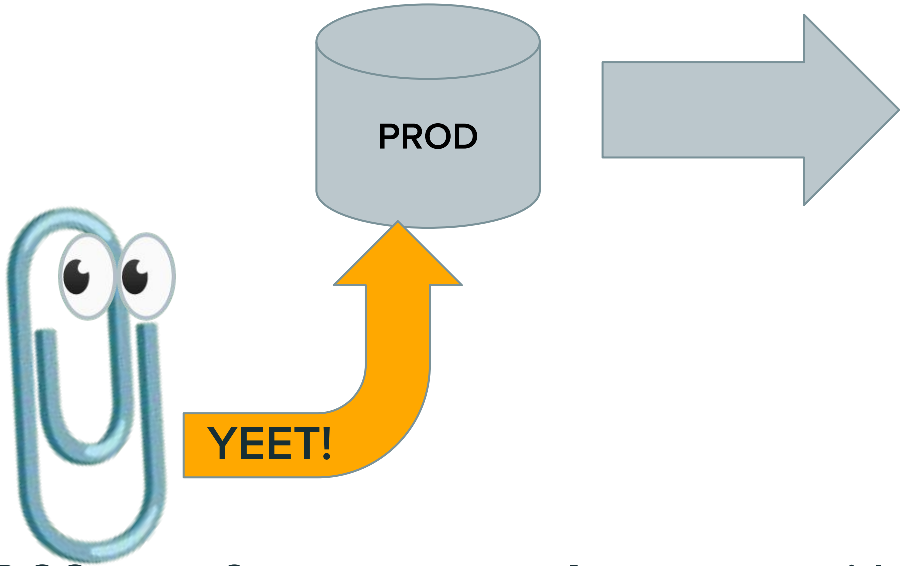
(Offline Experiments)

👍 Generally good

… And train w/ judgments



(Offline Experiments)

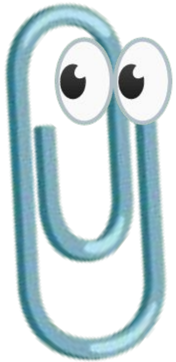


NDCG go up?
Ship to A/B!

Agreement with offline?

... Training w/ judgments

Query	Post ID	Rel?	Title Match?
Key bridge	1234	1	1
Key bridge	5678	0	1
Golden retrieval travel anxiety	12412	1	1



Mr. ML Model

PROBLEM - engagement based judgments have SOME relationship to document!

(even irrelevant ones) - why?

… We sample other queries for negative labels

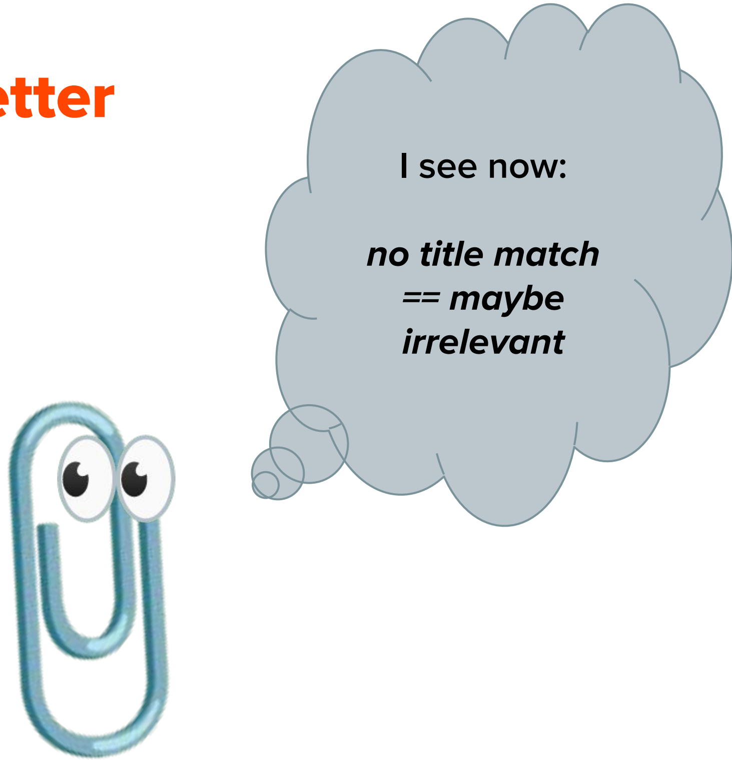
Query	Post ID	Rel?	Title Match?
Key bridge	1234	1	1
Key bridge	5678	0	1
Key bridge	12412	0	0
Golden retrieval travel anxiety	12412	1	1

Inject as irrelevant

(Inject some N random other query labels as negative for each query)

Mr. ML Model can see the patterns better

Post ID	Rel?	Title Match?
1234	1	1
5678	0	1
12412	0	0
12412	1	1



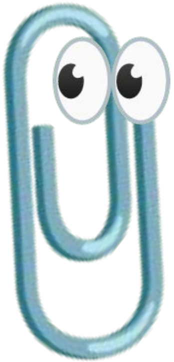
Mr. ML Model

How to choose features?

Training Data ✅

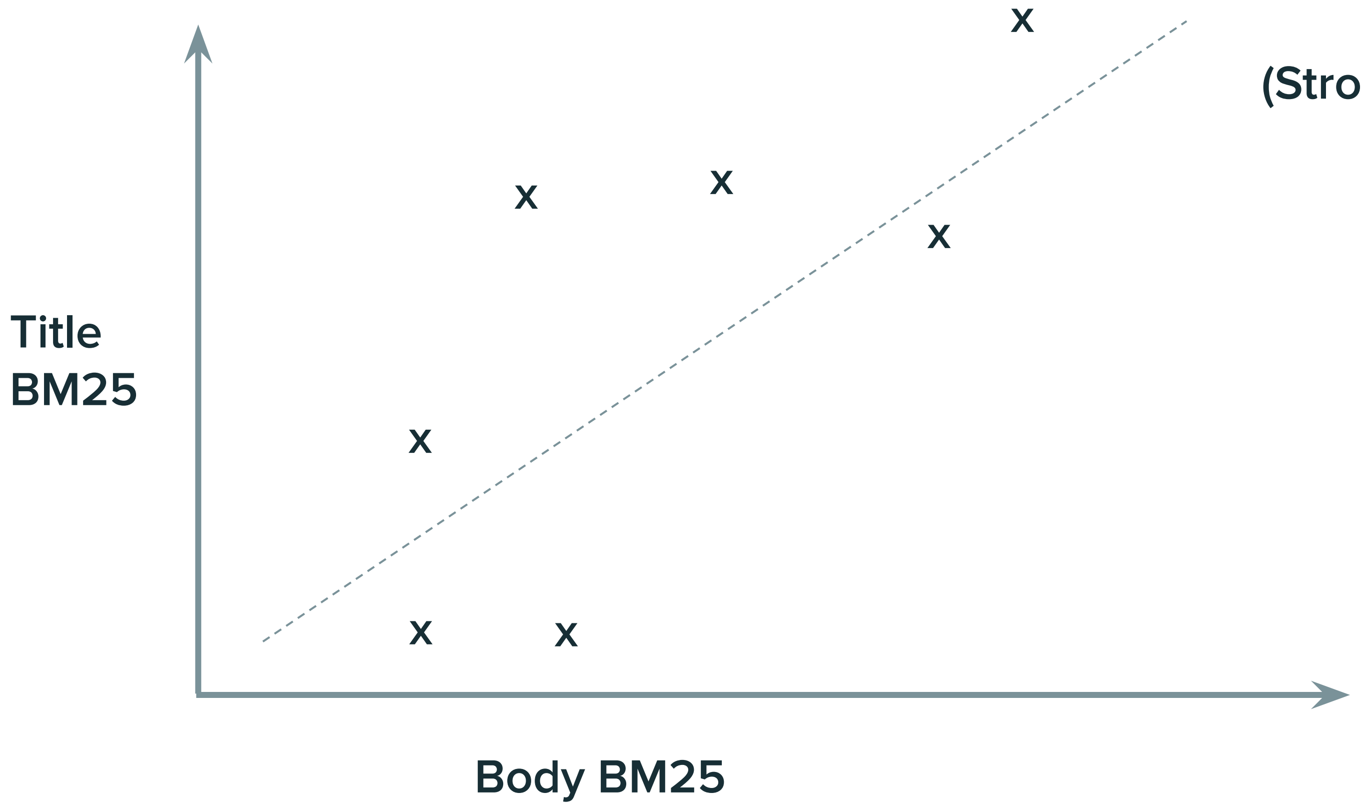
Features ⁇

FEED ME good features to learn relevance patterns



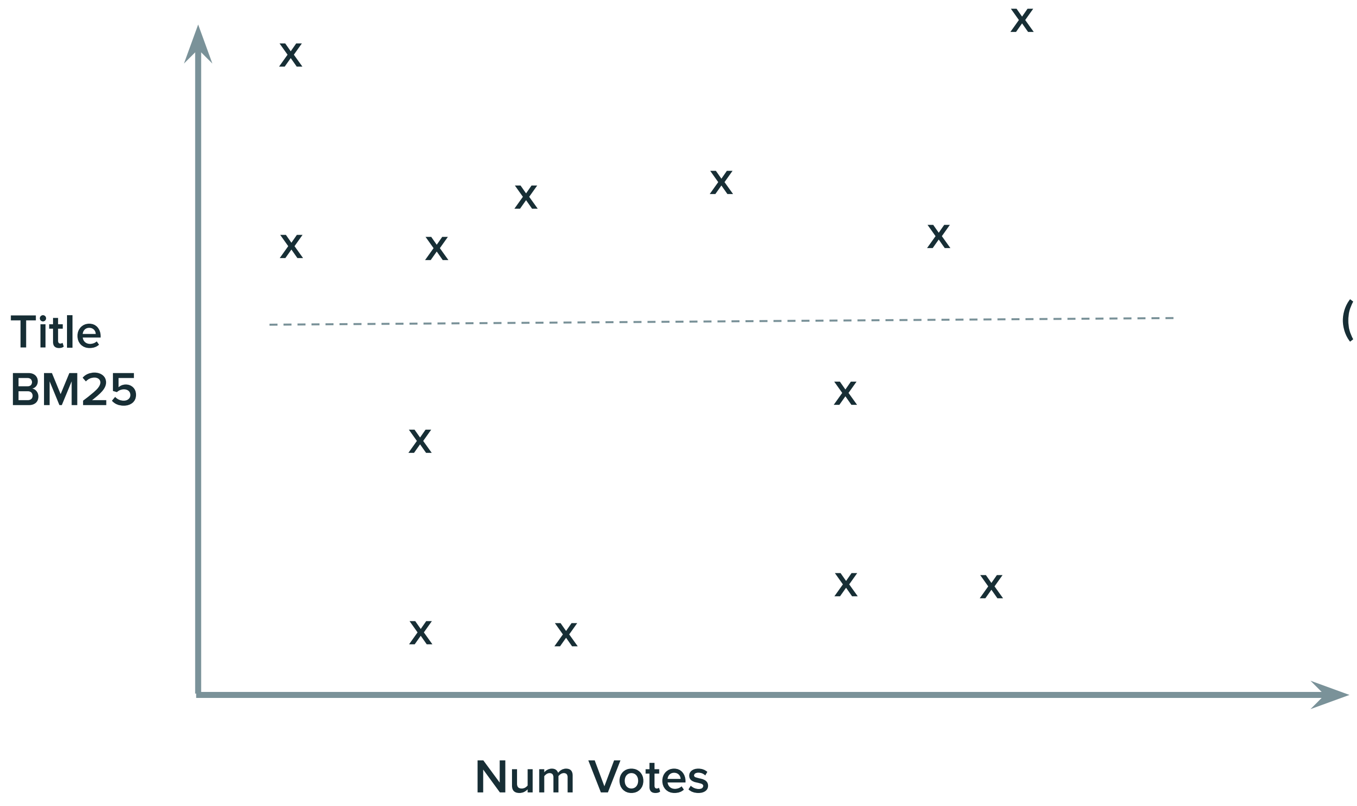
Mr. ML Model

Features often heavily correlated in LTR



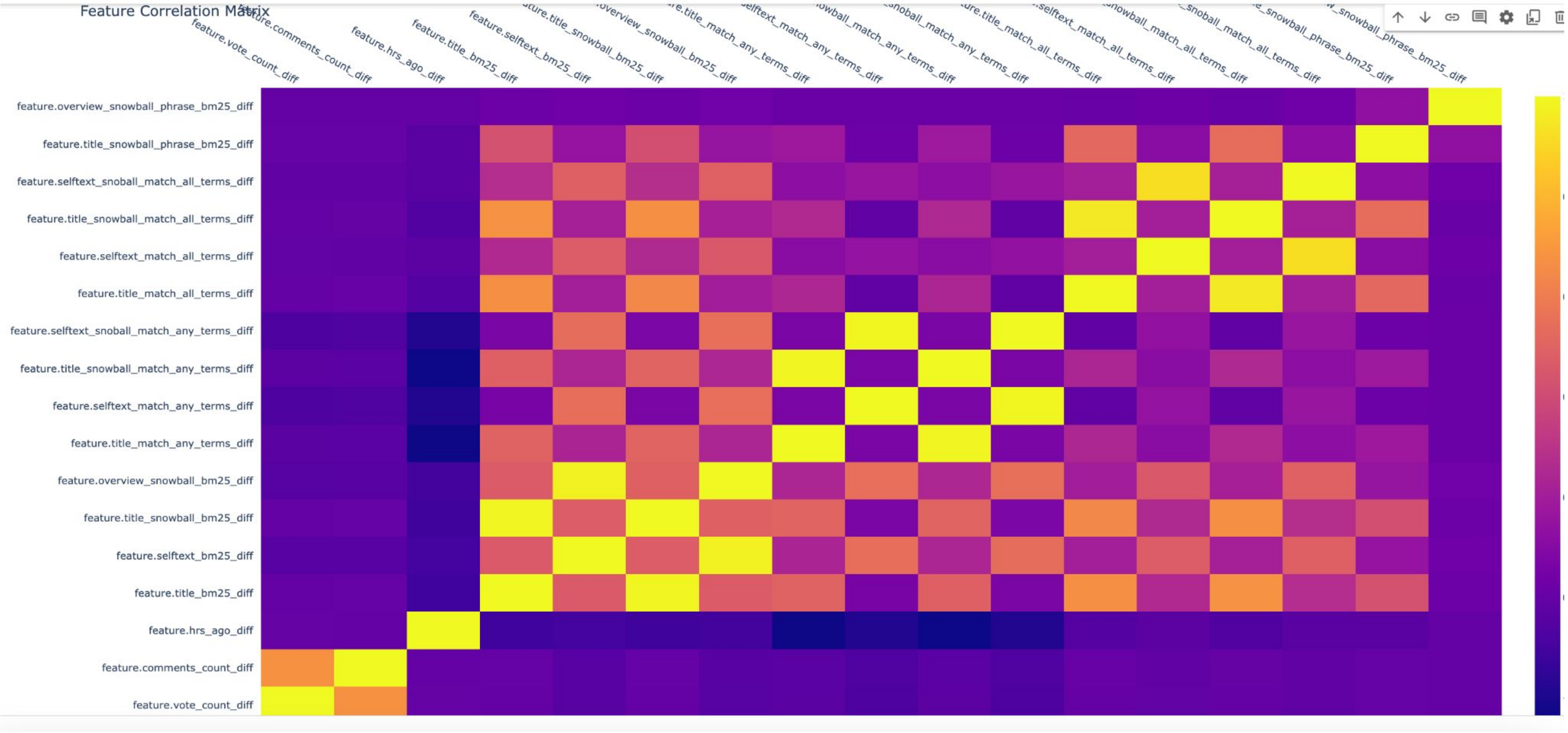
(Strong correlation) 😕

Good features add information

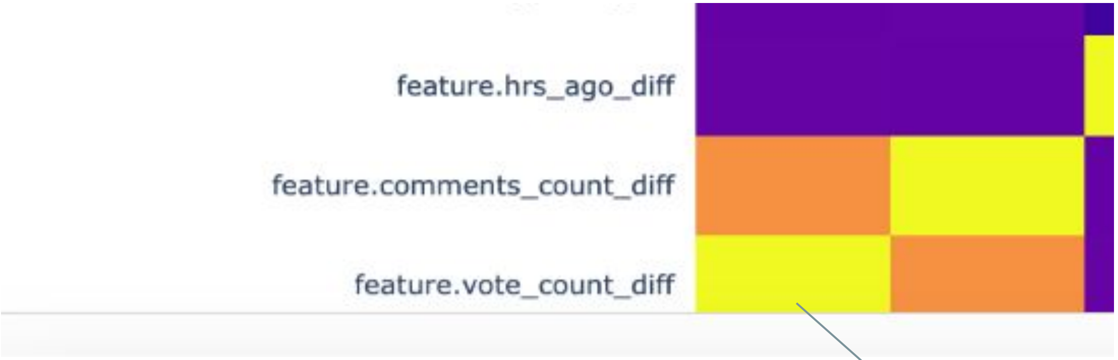


(No correlation) 👍

Analyze via correlation matrix

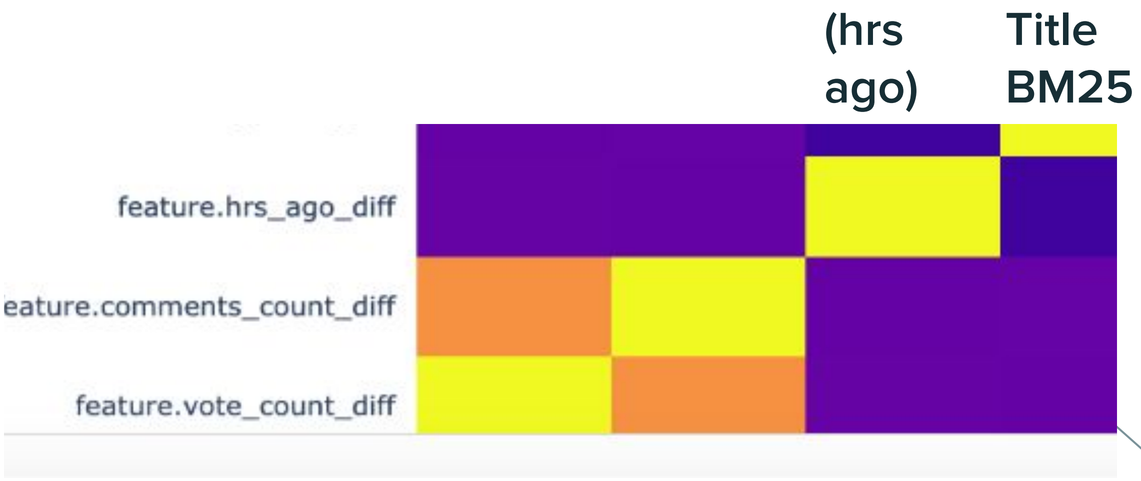


Analyze via correlation matrix



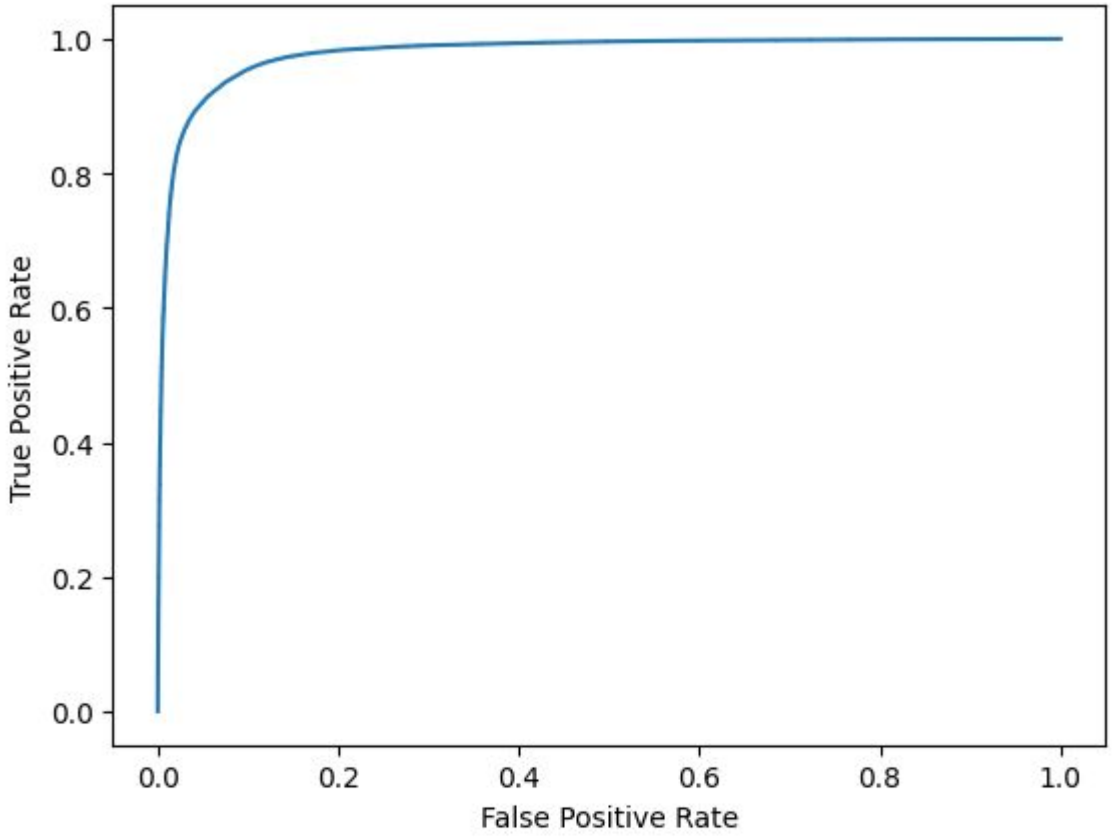
Votes / Num Comments Correlate, don't add much new info relative to each other

Analyze via correlation matrix



But add quite a bit on top of these features

Goal: find INDEPENDENT features, that IMPROVE model



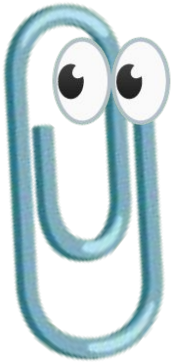
Feature adds value when:

1. Orthogonal to other features
2. Improves model
3. Is readily accessible and computationally feasible to compute

How to choose features?

Training Data ✅

Features ✅



Mr. ML Model

Model architecture:

Lots of Choices, main requirements:

- Listwise / pairwise loss function
- Handle non-linear and correlated features

We chose

- LambdaMART loss
- Deep learning model

Yeeting Features + Models to Prod

Choosing Solr LTR Plugin

(Lexical)
Feature Calculation + Model Inference

Solr functionality for

- Feature calculation
- Top N Reranking

Pros / Cons Solr LTR vs Reddit extra

	Solr LTR	Reddit's existing ML infra
Query-dependent features?	Yes	Not easily
Exists (at Reddit?)	No	Yes
Time horizon of content	~19 years	90 days
Features available	Minimal	Extensive
Network hops	None	Several
Types of models	Limited	Extensive
Model store size	~1MB*	Unbounded
Vertical scalability	Shared with Solr	Unshared

🧐 Which to choose?

Solr LTR Plugin

[Solr LTR - Reference Guide](#)

Feature Store + Logging

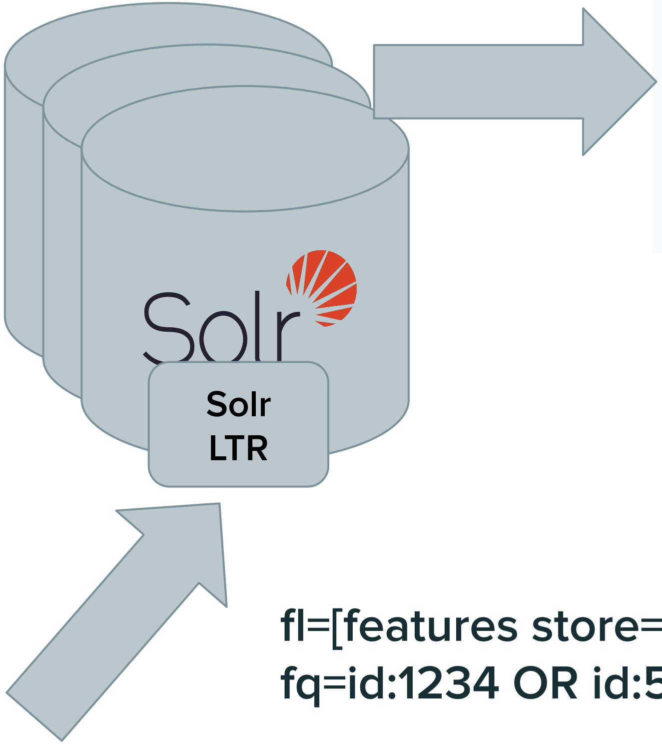
IE [From Zero to Solr LTR](#):

Solr Query DSL

MY_EFI_FEATURE_STORE

```
[
  {
    "store" : "my_efi_feature_store",
    "name" : "tfidf_sim_a",
    "class" : "org.apache.solr.ltr.feature.SolrFeature",
    "params" : { "q" : "{!dismax qf=text_tfidf}${keywords}" }
  },
  {
    "store" : "my_efi_feature_store",
    "name" : "tfidf_sim_b",
    "class" : "org.apache.solr.ltr.feature.SolrFeature",
    "params" : { "q" : "{!dismax qf=text_tfidf}${keywords}" }
   },
```

Training Time



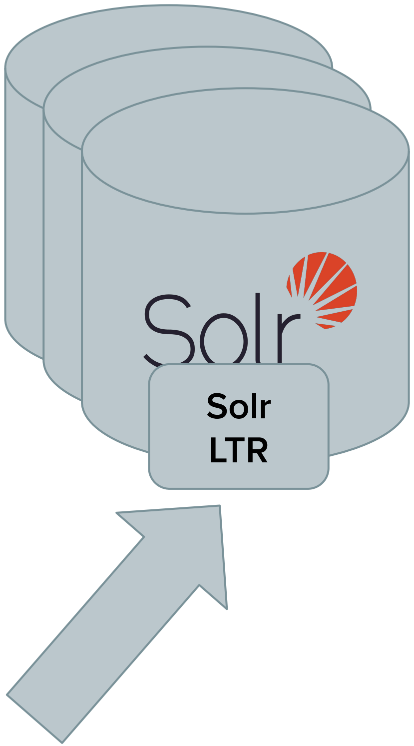

```
[
  {
     “id”: 1234,
     “[features]”: “tfidf_sim_a=1.56,...”
  },
  {
     “id”: 5678,
     “[features]”: “tfidf_sim_a=0.05,...”
  },
  …
```

(training examples for docs 1234… ,
… for query ‘football’)

fl=[features store=my_efi_feature_store efi.keywords=’football’]&
fq=id:1234 OR id:5678 OR id:1010

Keyword “football” posts: 1234, 5678, 1010

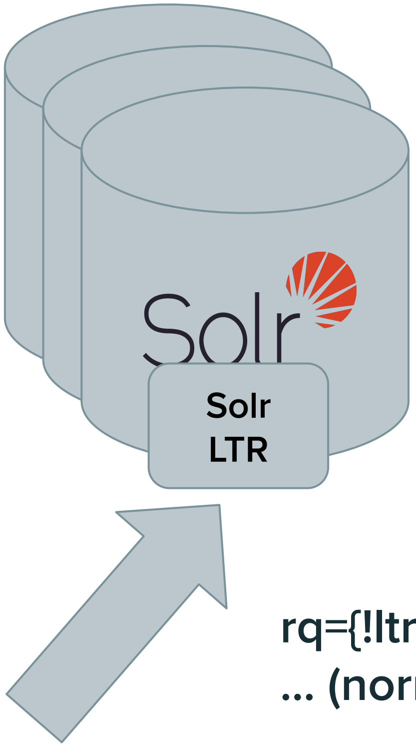
Store model for inference



`Model: foo`

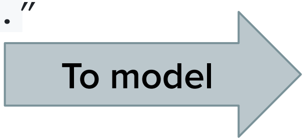
`Store:` my_efi_feature_store

Inference Time



Top N to rerank:

```
[
  {
     "id": 1234,
     "[features]": "tfidf_sim_a=1.56,..."
  },
  {
     "id": 5678,
     "[features]": "tfidf_sim_a=0.05,..."
  },
  …
```



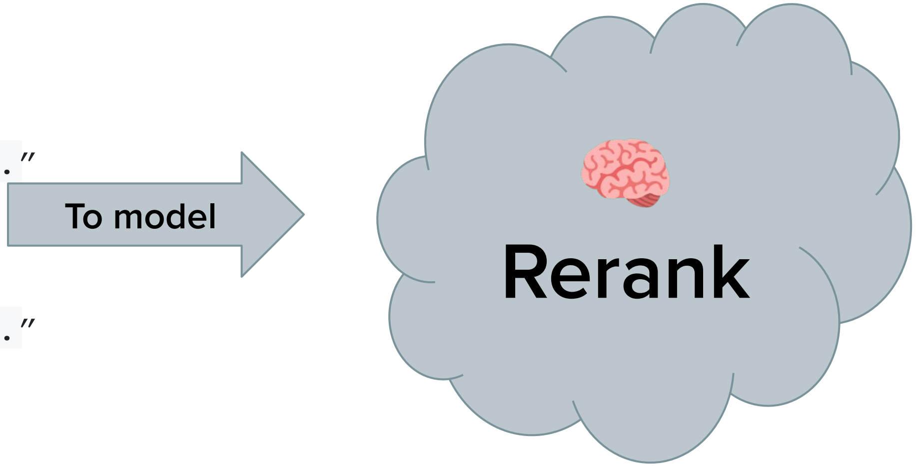
(Features Computed internal to Solr)

rq={!ltr model=foo-model efi.keywords='football']&
… (normal retrieval query)

Inference Time

Top N to rerank:

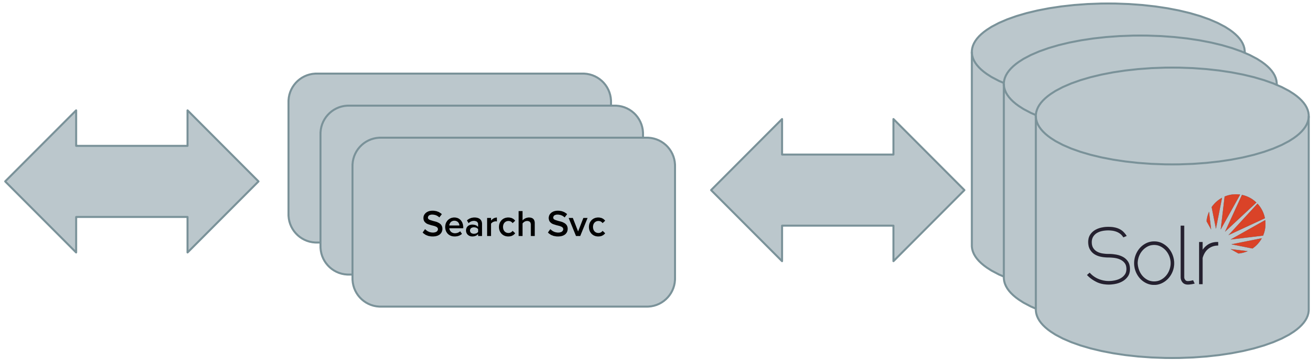
```
[
  {
     “id”: 1234,
     “[features]”: “tfidf_sim_a=1.56,...”
  },
  {
     “id”: 5678,
     “[features]”: “tfidf_sim_a=0.05,...”
  },
  …
```



(Features Computed internal to Solr)

rq={!ltr model=foo-model efi.keywords=’football’]&
… (normal retrieval query)

Our search infra

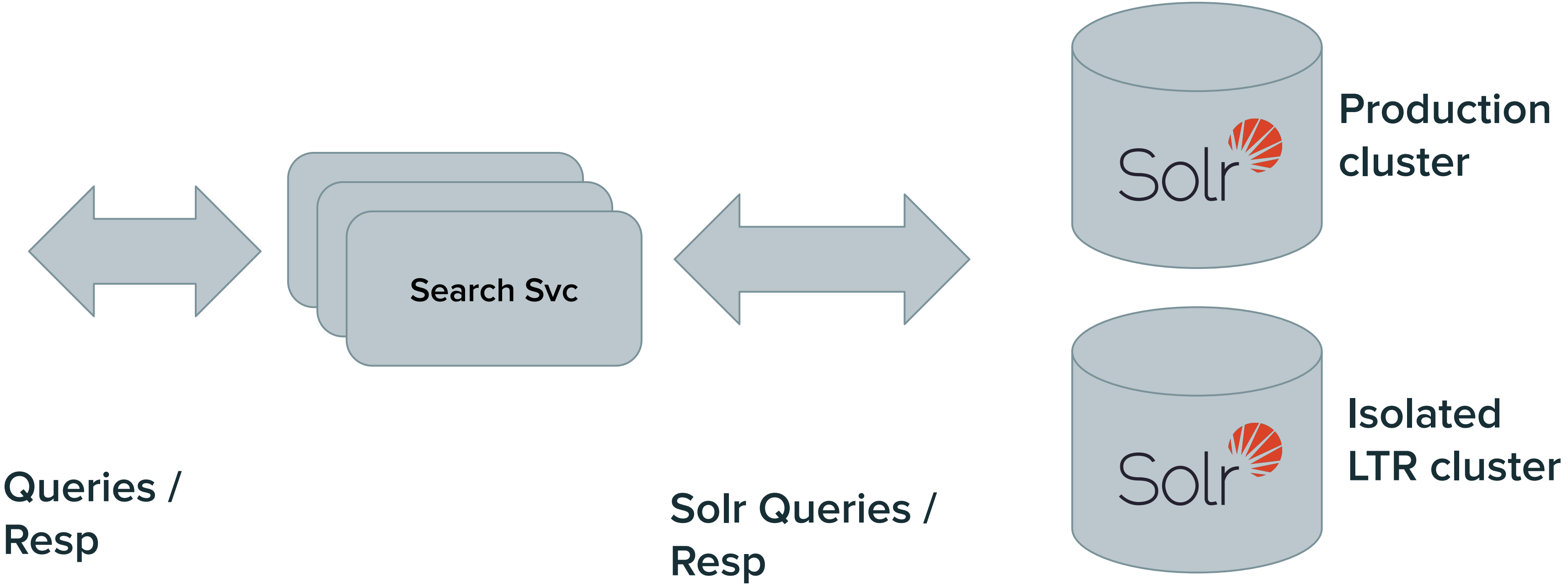


Queries / Resp

Solr Queries / Resp

Solr Cloud

Our search infra: build in isolation or production cluster?

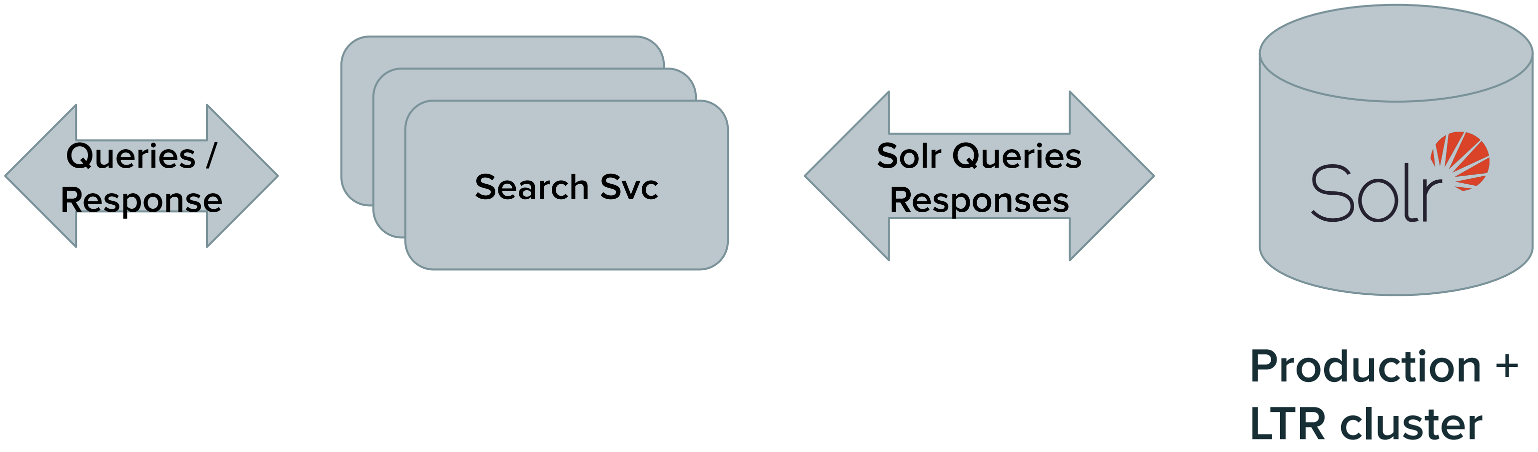


Pros / Cons

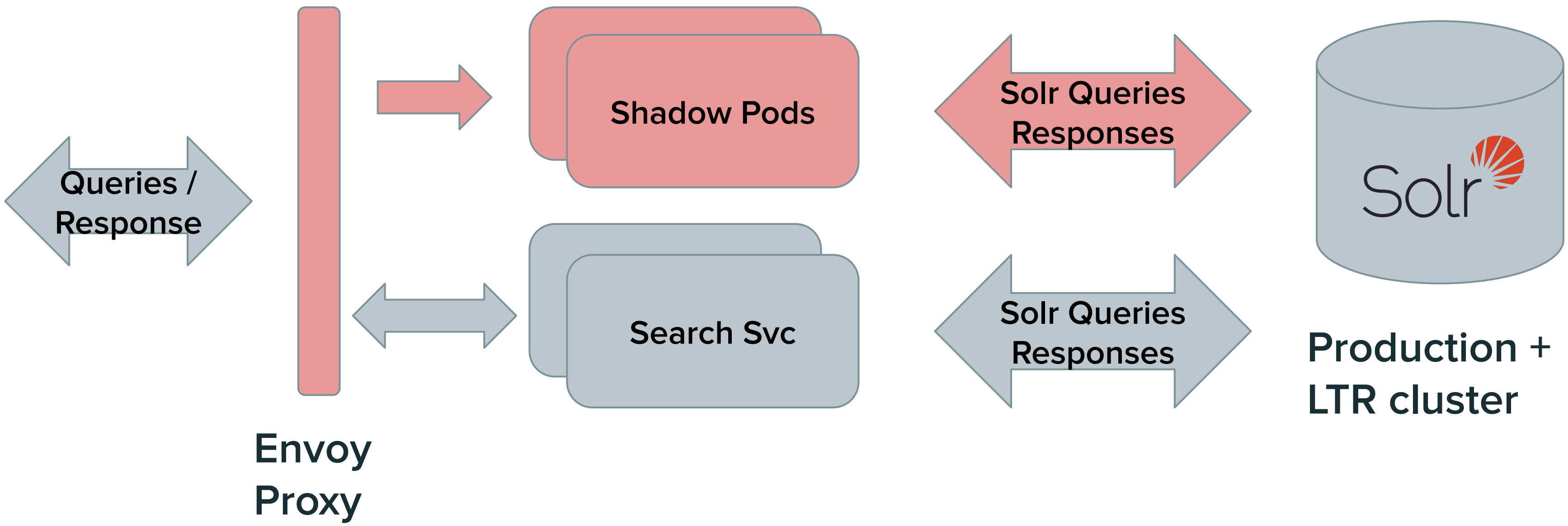
	Isolated	Single cluster
Implementation speed	Need to add a new cluster	Already built!
Development isolation	Build/ iterate fast independently of other work	Slower b/c of need to integrate with other work
Safety	Faults don’t cascade	Faults affect prod traffic
Experiment confounders	Different latencies	Same latency in prod and experiment
Operational cost	One more cluster to maintain	Maintain two use cases in same cluster
$$$	One more cluster to buy (non-trivial cluster cost)	Vertically scale existing cluster slightly

🧐 **Which to choose?**

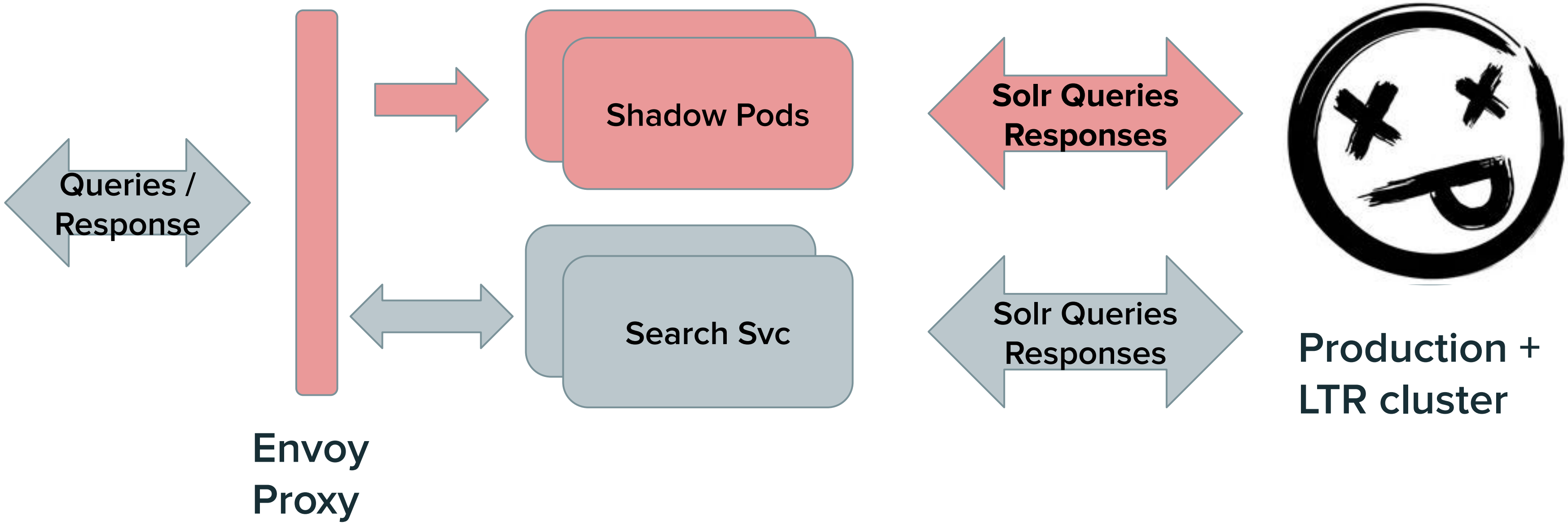
Take 1: single cluster



Take 1: Envoy for Shadow Traffic (Single Cluster)

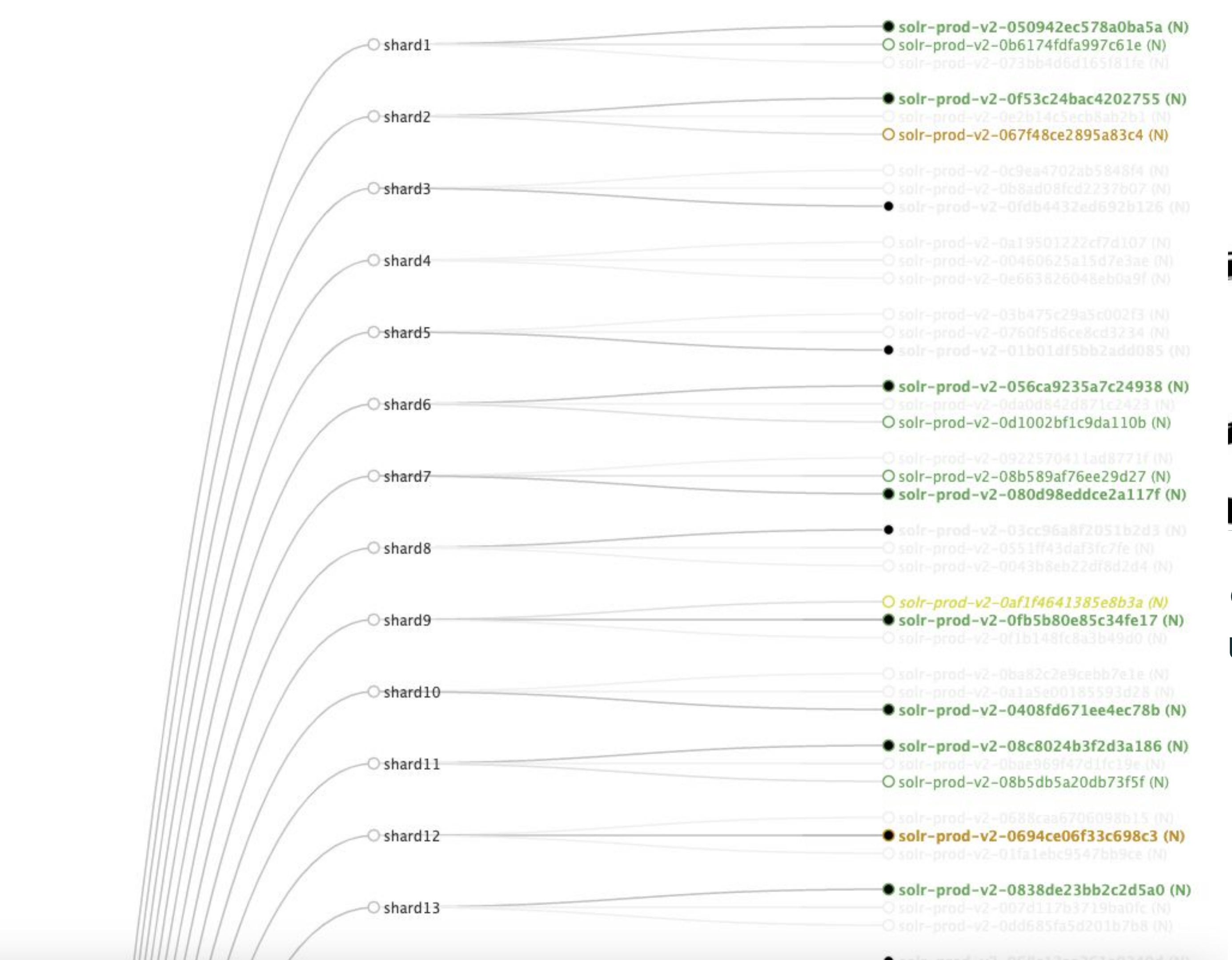


Problems with co-location



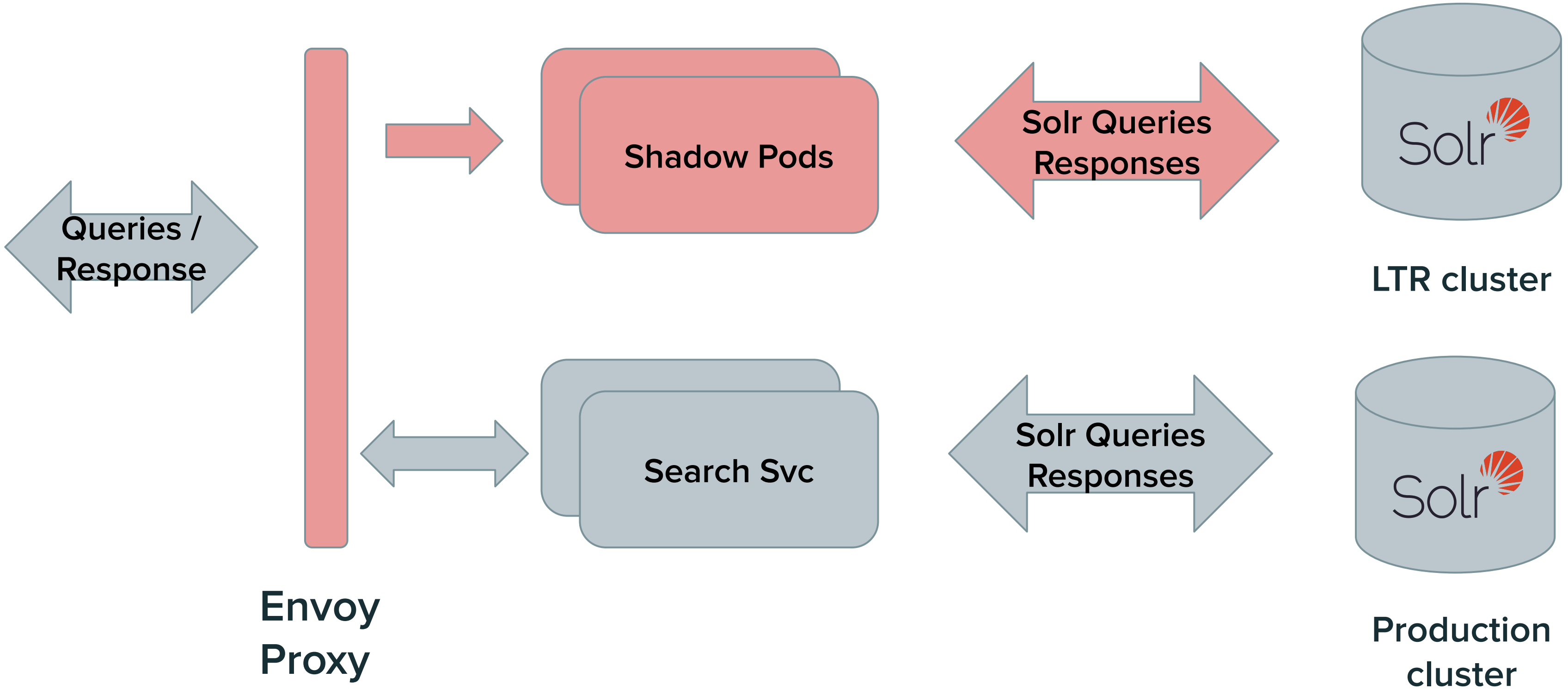
Proble

Queries
Respons

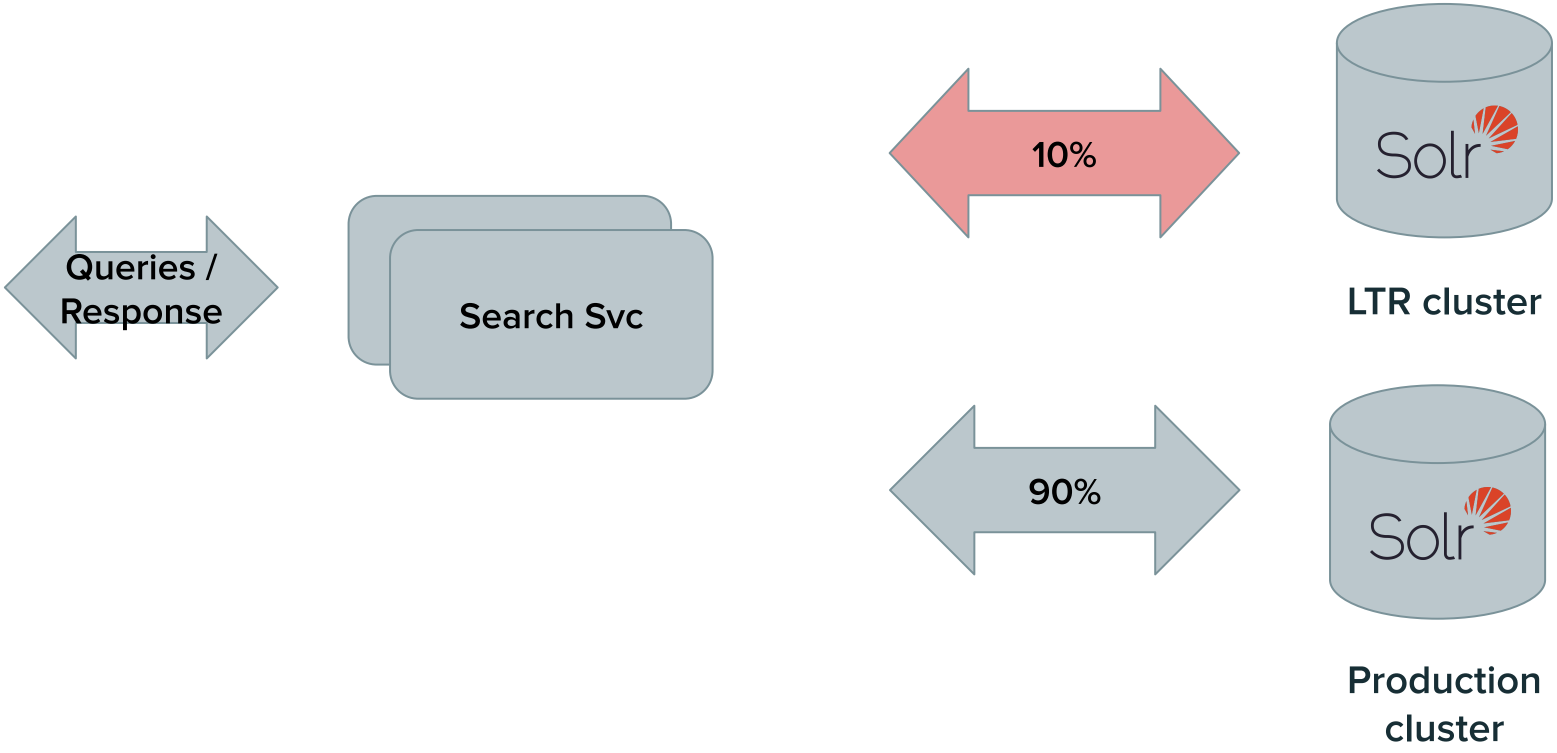


ction +
uster

Take 2: Isolated clusters



User-level Testing w/ traffic splitting



+ Solr + Learning to Rank

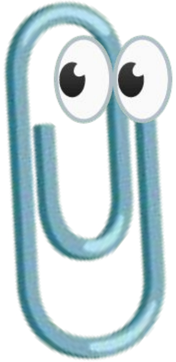
Java

1. Retrieval (get top N docs per shard)

2. Re-rank (all N x shards docs)
   a. Features computed/queried

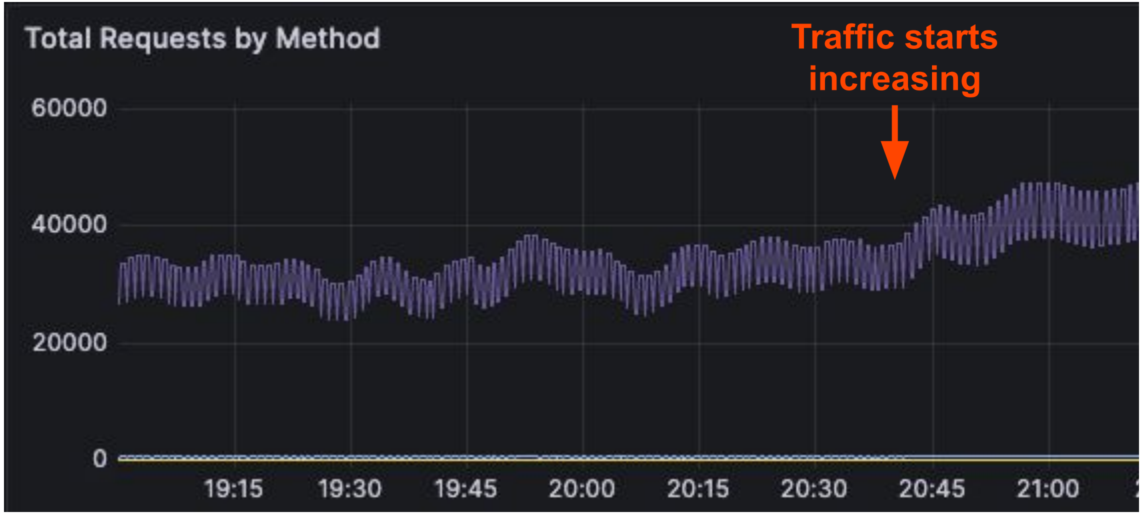
```
title:${keywords}
body:${keywords}
title_phrase:"{$keywords}"
```

   b. Mr. ML Model interprets features

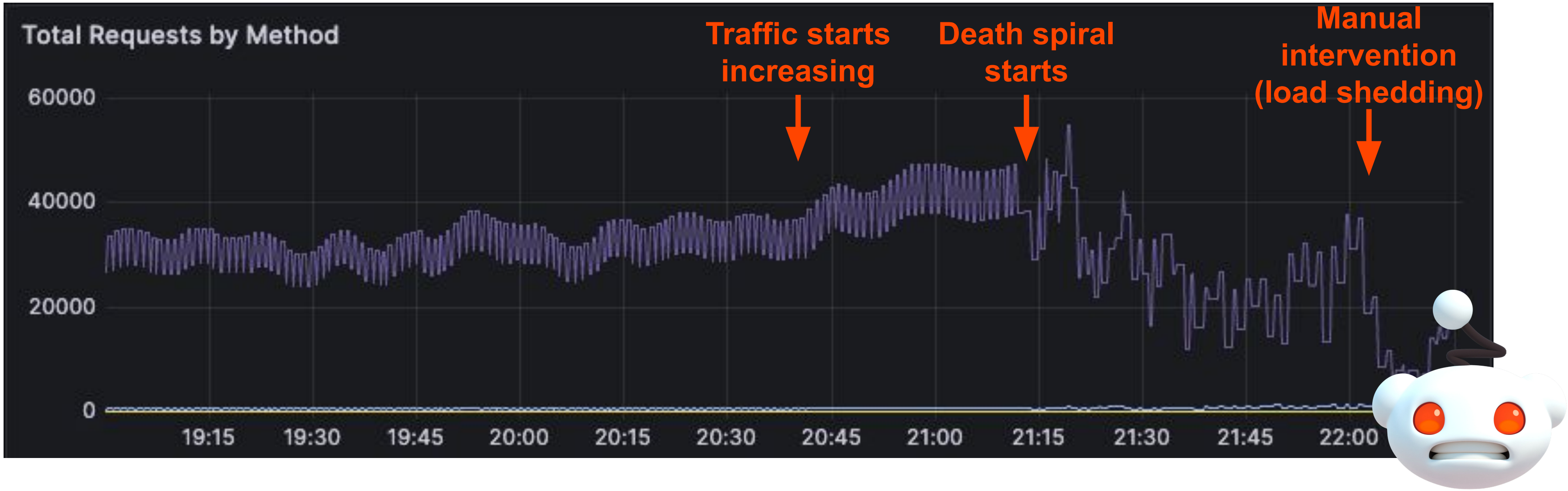


3. Return re-ranked results

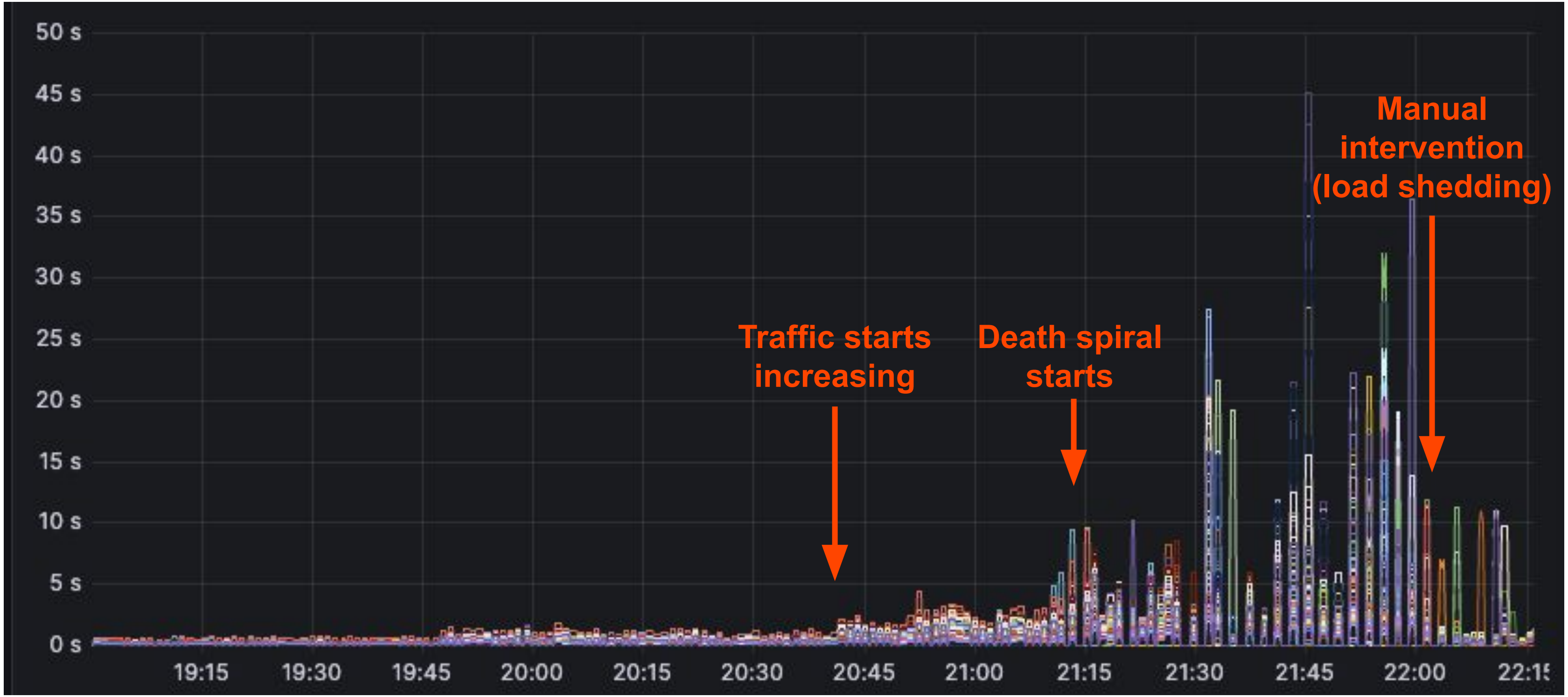
Scaling up ...


Total Requests by Method
Traffic starts increasing
60000
40000
20000
0
19:15
19:30
19:45
20:00
20:15
20:30
20:45
21:00

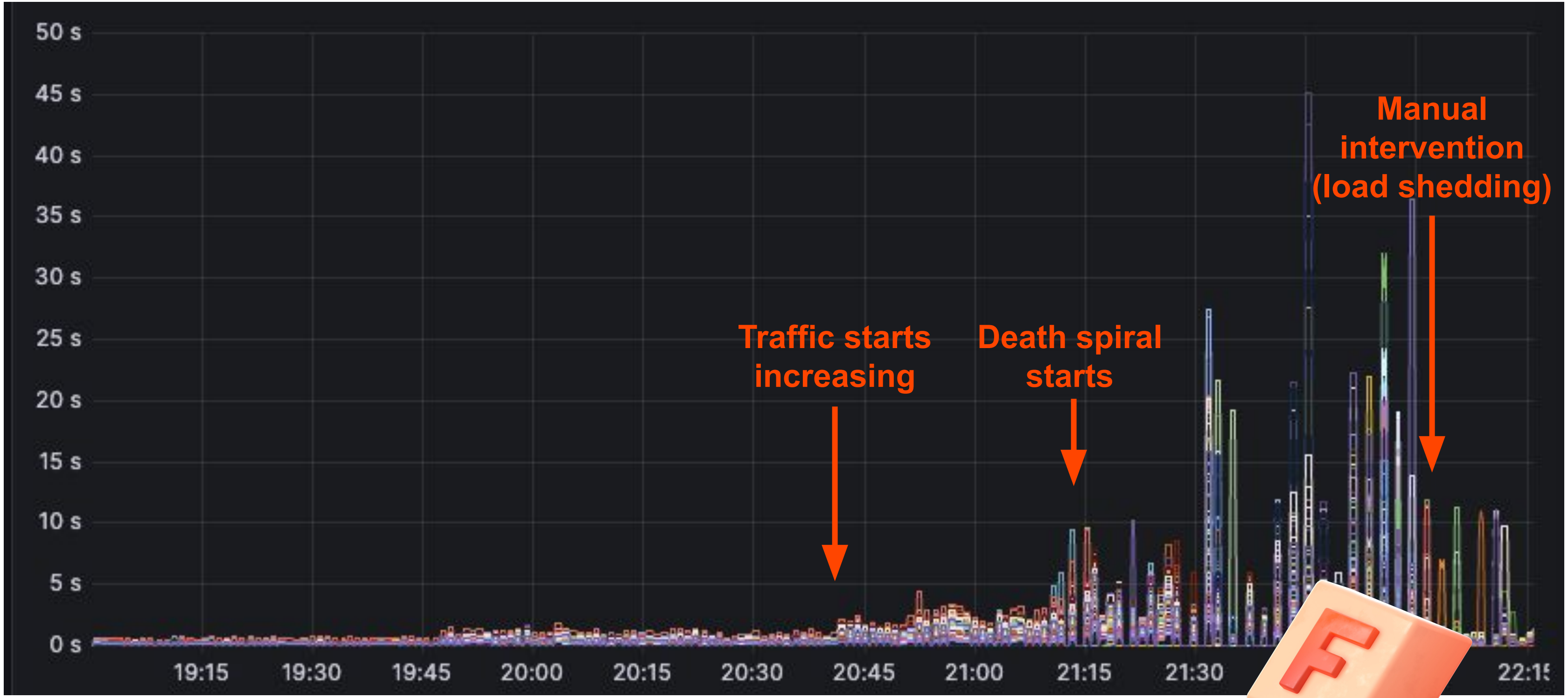
Scaling up … and running into failures



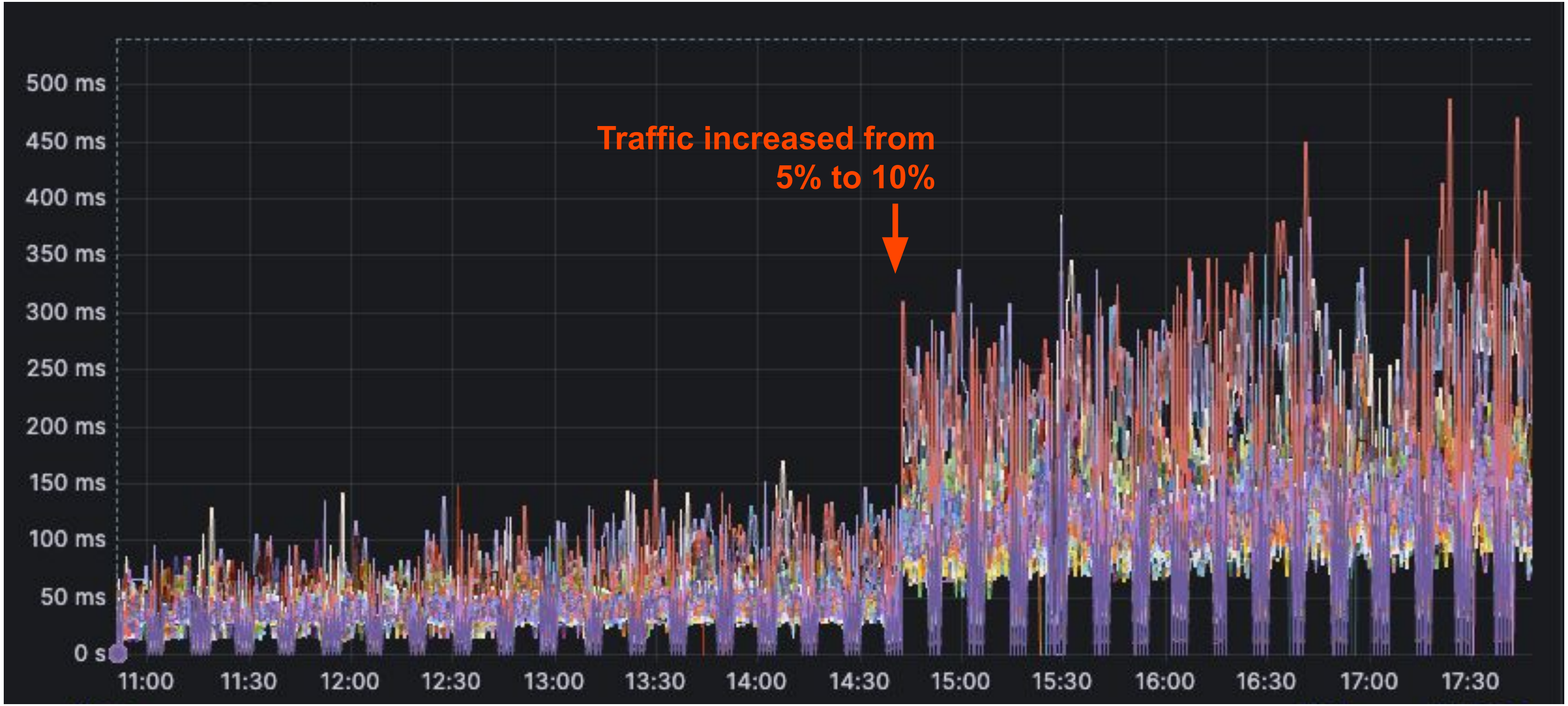
Garbage Collection time spent



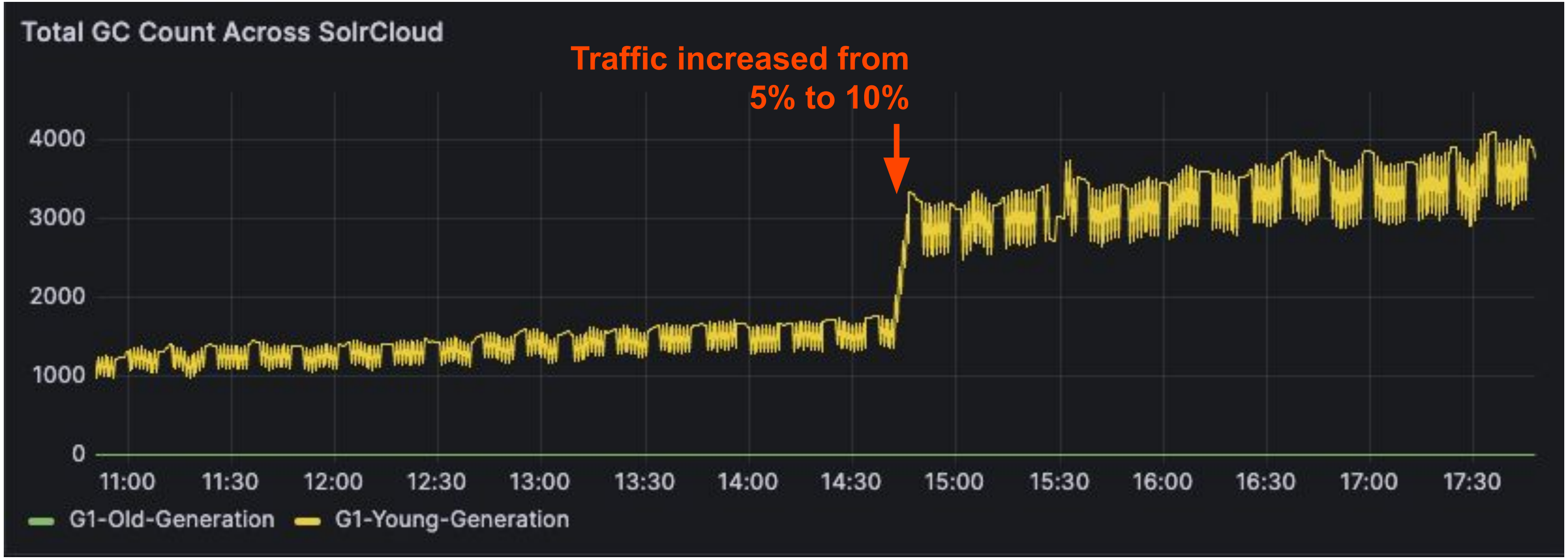
Time spent in Garbage Collection



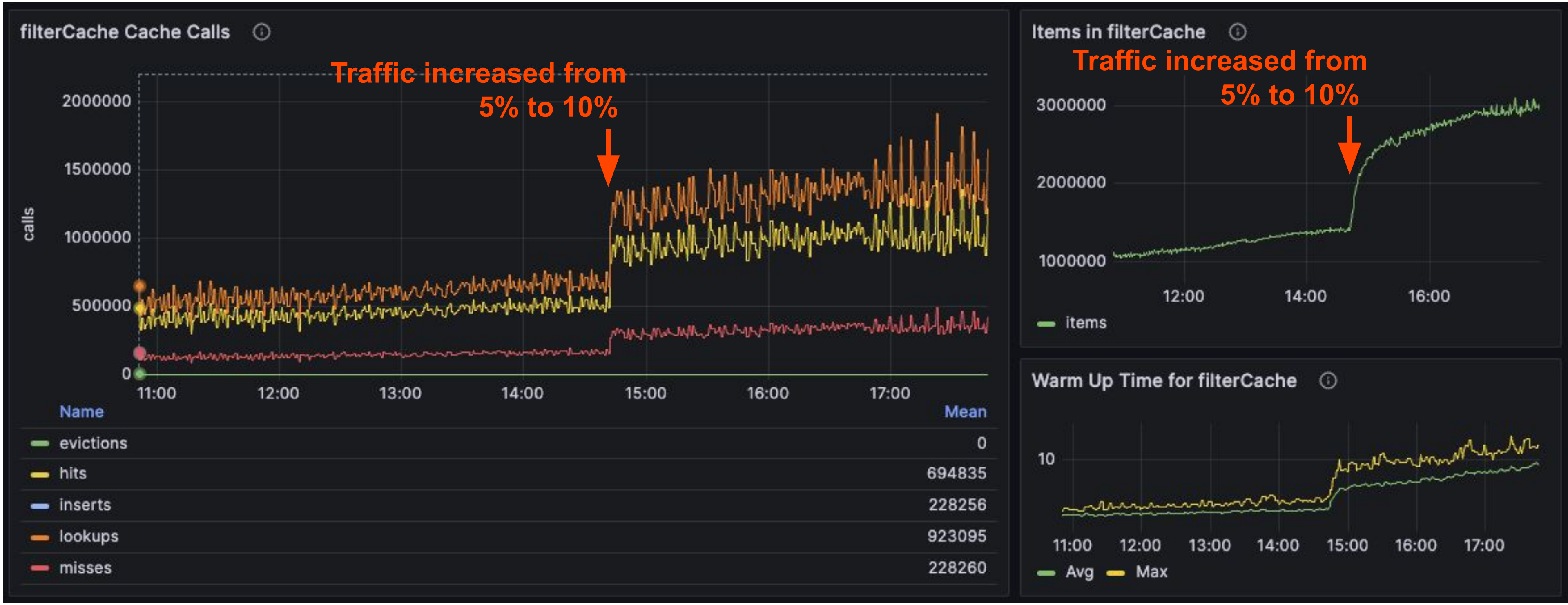
Garbage Collection time spent (smaller jump)



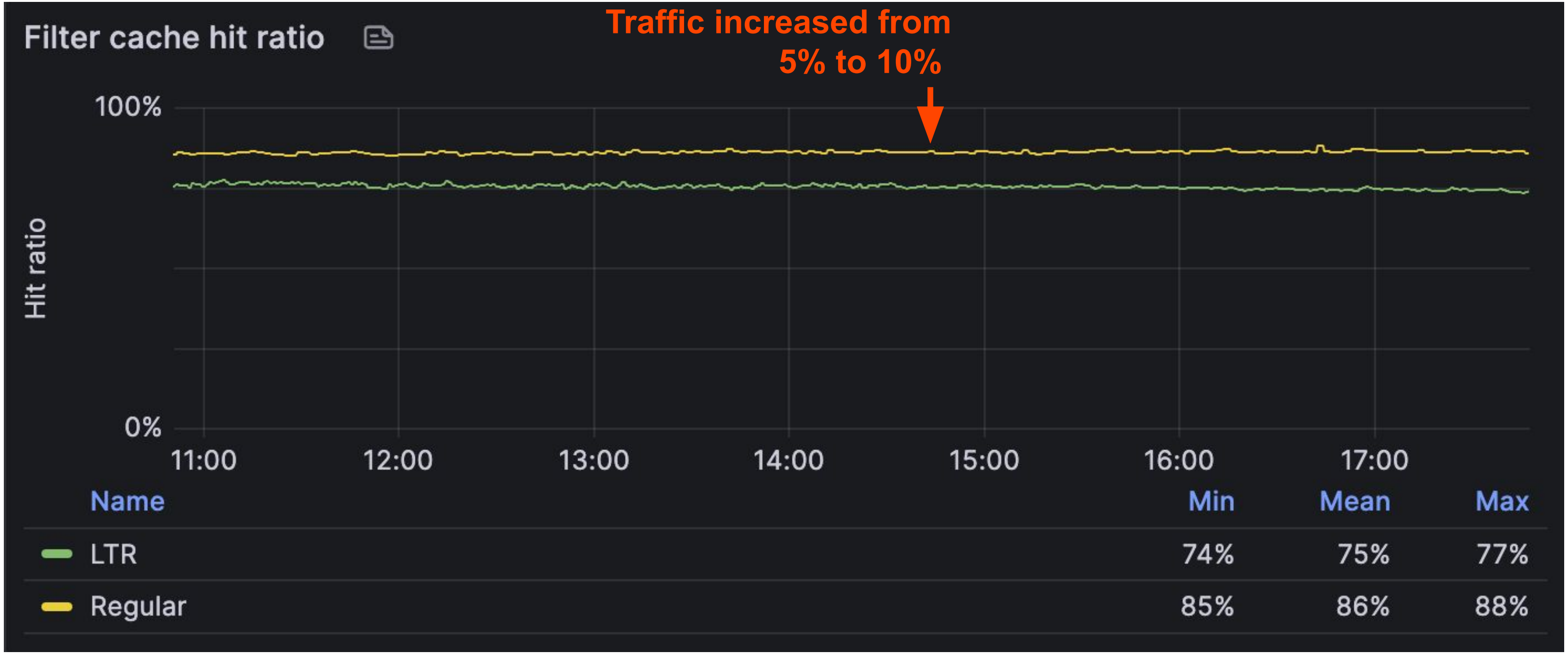
Garbage Collection time spent (smaller jump)



The caches look funny...



The caches look funny...



What do our features look like? Do they cache?

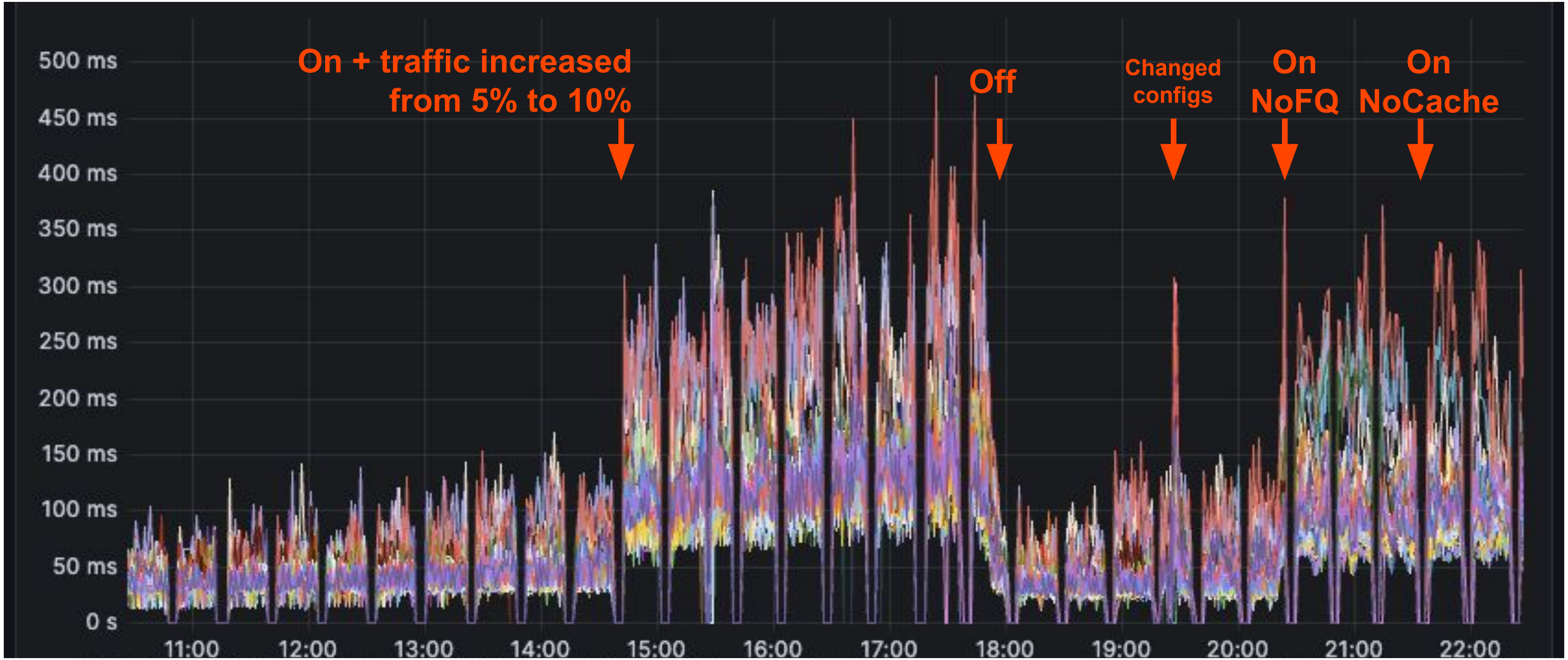
```
{
  "name": "title_match_all_terms",
  "store": "LTR_TRAINING",
  "class": "org.apache.solr.ltr.feature.SolrFeature",
  "params":
  {
    "fq":
    [
      "{!edismax qf=title mm=100% v=\"${keywords}\"}"
    ]
  }
},
…
```

Should this be cached?
Should we set `cache=false` **?**

Let’s test a few configurations

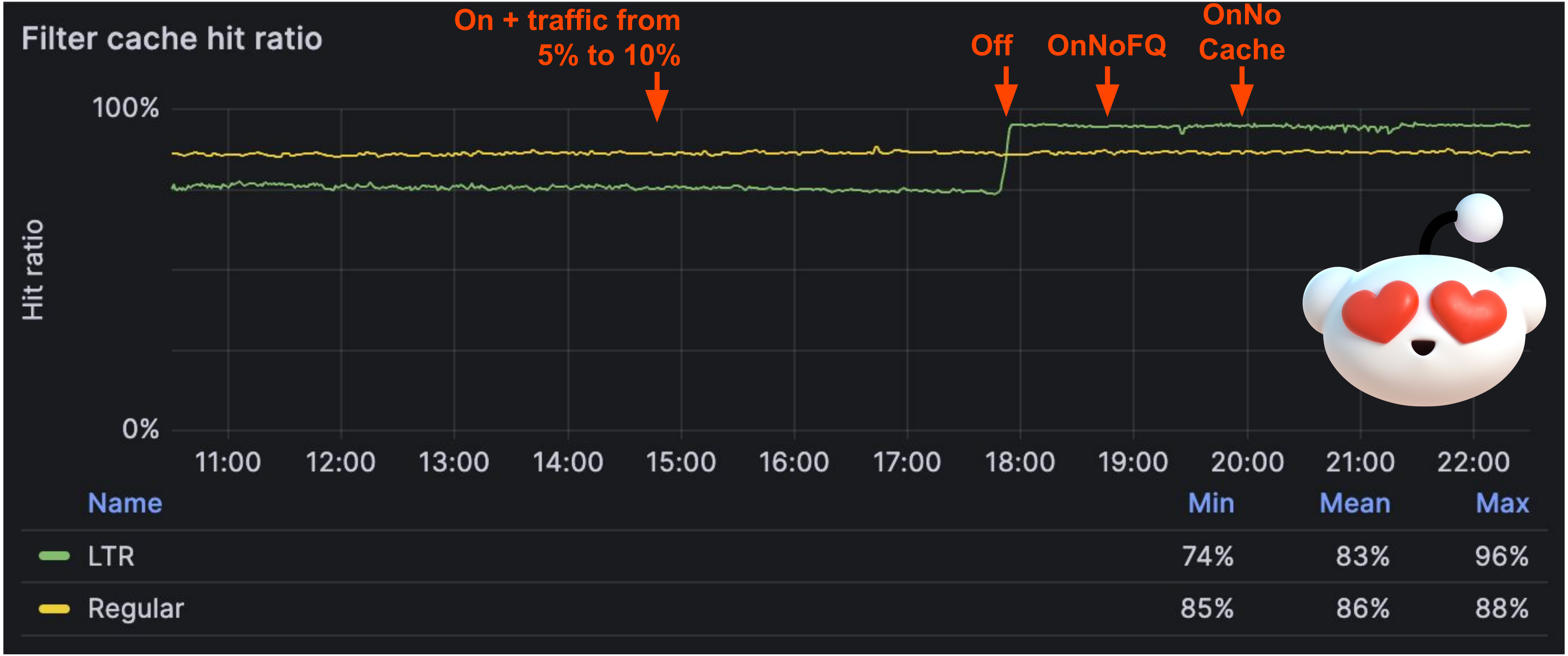
On	Re-rank with no changes
Off	No re-ranking
OnNoFQ	Re-rank without FQ features
OnNoCache	Re-rank with non-cached FQ features (`cache=false`)

Garbage Collection time spent



Caching reactions

Caching hit rate increased



Latency stabilized!

Tuning takeaways

GC performance is important for Solr stability

Avoiding unnecessary work to optimize performance

LTR features can be expensive

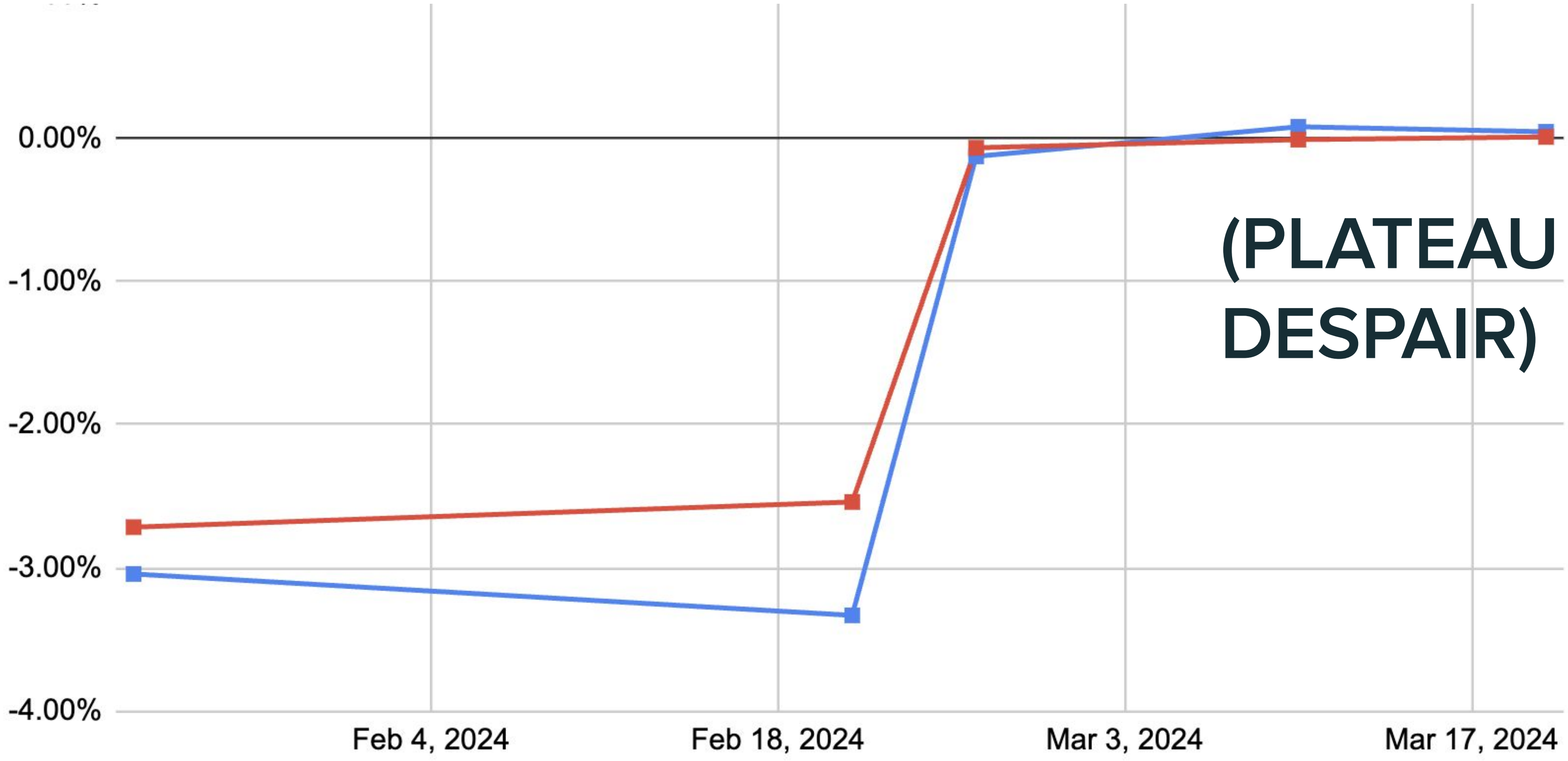
Yeet to to the moon!

(next steps)

reddit

All is well, until...

OFFLINE SAID THESE SHOULD WIN!



(PLATEAU OF DESPAIR)

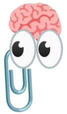
(A/B Tests LTR vs control ...)

Revisit labels

Manual relevance

- Some qualitative analysis, more human in the loop
- Weighted avg: NDCG + LGTM
- Can eyeball different types of queries and LGTM

Can be accurate ~80-90% of the time



LTR (Mr. ML Model)

- Model only as smart (or dumb) as labels
- 100% NDCG
- Examples MUST be weighted by frequency

Must be accurate 100% of the time

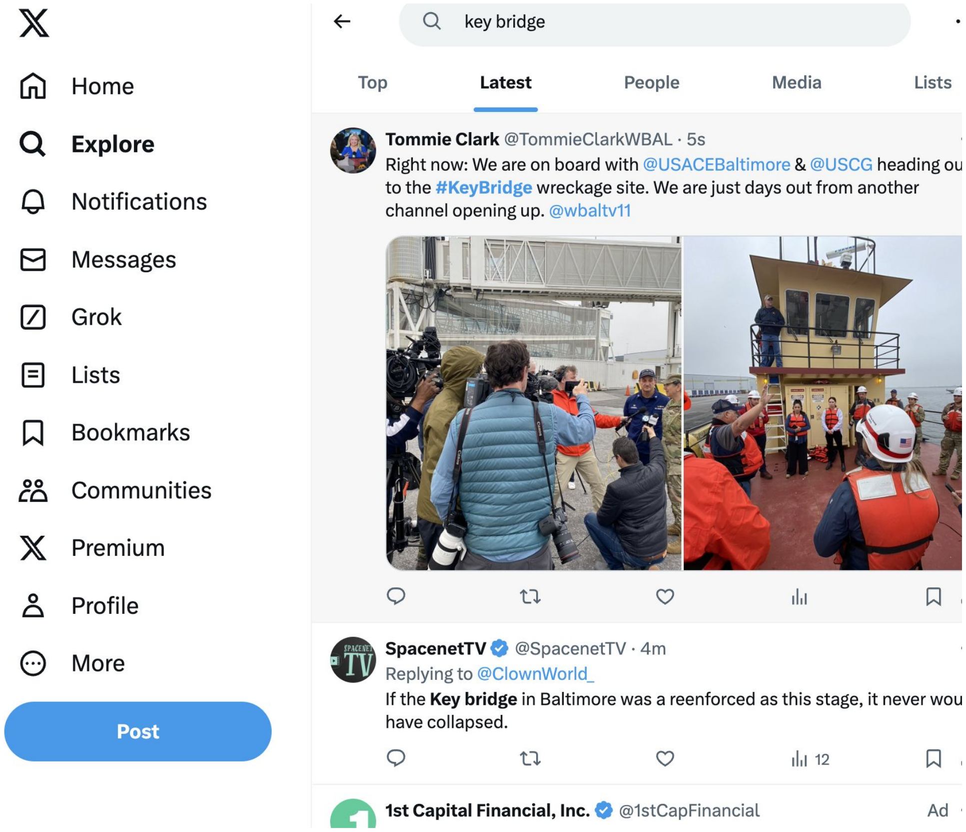
Social search problem - very very changing SERPs

Compared to e-commerce, etc

SERPS change

A LOT!

-> Aggregated labels don't reflect actual SERPs



Currently
Human -> Analytic labels

Multiple SERP analytics events

SERP ID	DATE	User Id	Query	Rank	Doc ID	Click+Dwell?
1234	2 days ago	u_124	zoolander	0	abcd	0
1234	2 days ago	u_124	zoolander	1	1212	1

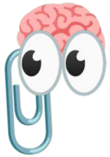
…

SERP ID	DATE	User Id	Query	Rank	Doc ID	Click+Dwell?
1251	25 days ago	u_110	zoolander	0	1211	0
1251	25 days ago	u_124	zoolander	1	12ab	1

Aggregated to:

~Avg Click+Dwell over N days

Train



Use SERP directly to train?

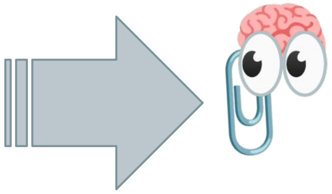
SERP ID	DATE	User Id	Query	Rank	Doc ID	Click+Dwell?
1234	2 days ago	u_124	zoolander	0	abcd	0
1234	2 days ago	u_124	zoolander	1	1212	1

SERP ID	DATE	User Id	Query	Rank	Doc ID	Click+Dwell?
1251	25 days ago	u_110	zoolander	0	1211	0
1251	25 days ago	u_110	zoolander	1	12ab	1

x 100K ? 1m?

Benefits:

- Implicitly weighted
- Handle Changing SERPs
- Features logged at point of search
- Can train on ALL context



Downsides:

- Need to feature log every search
- A lot more data!

Feature Eng - Signals

Trending / recent posts that get engagement for a query

query	post	boost
ace ventura	6785	1.2
zoolander	1234	1.5
zoolander	5678	1.1

Pros / Cons Signals vs an LTR model

🫣 **Signals:**

“OVERFIT” - not generalized, but a great cheat-sheet for ‘right answer’, but only for queries seen in past

Good for fast changing head queries

🤓 - **Model:**

“GENERALIZED” - not overfit, general “pattern” can work with query seen rarely / never

Good for torso+tail / not as engaging queries

Signals cover A LOT of the search traffic

These cover WAY more of the search traffic than expected

🫣 <u>**Signals:**</u>

“OVERFIT” - not generalized, but a great cheat-sheet for ‘right answer’, but only for queries seen in past

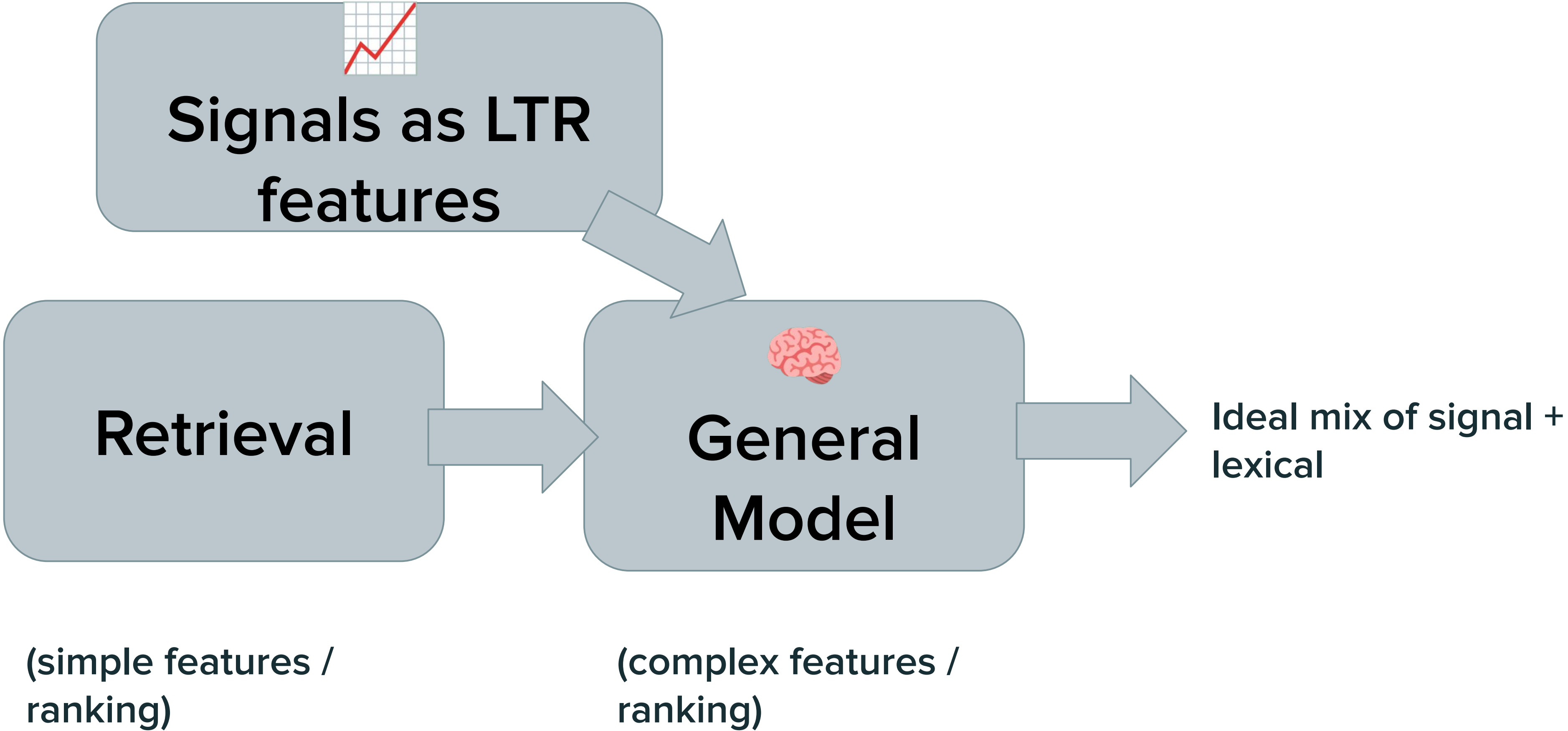
Good for fast changing head queries

🤓 - <u>**Model:**</u>

“GENERALIZED” - not overfit, general “pattern” can work with query seen rarely / never

Good for torso+tail / not as engaging queries

Need to add these to our model



Thank you