

# Building MLOps Infrastructure at Japan's Largest C2C E-Commerce Site

Platform, productionalization, and monitoring

Teo Narboneta Zosa

Ryan Ginstrom

Search Team

mercari

A large red rounded rectangle and a blue circle are positioned in the bottom right corner of the slide, partially overlapping each other.

# **| Contents**

- 1. Introduction**
- 2. The Problem**
- 3. System Evolution**
- 4. Future Directions**
- 5. Conclusions**

# Introduction

---

## About Search at Mercari

# Introduction

## About Mercari

- Japan's largest consumer-to-consumer (C2C) online marketplace
- FY2022 numbers: [\\*](#)
  - Gross merchandise value (GMV): ~¥880 billion (~\$6.7 billion USD)
  - Net sales: ~¥150 billion (~\$1.1 billion USD)



# Introduction

## Search at Mercari

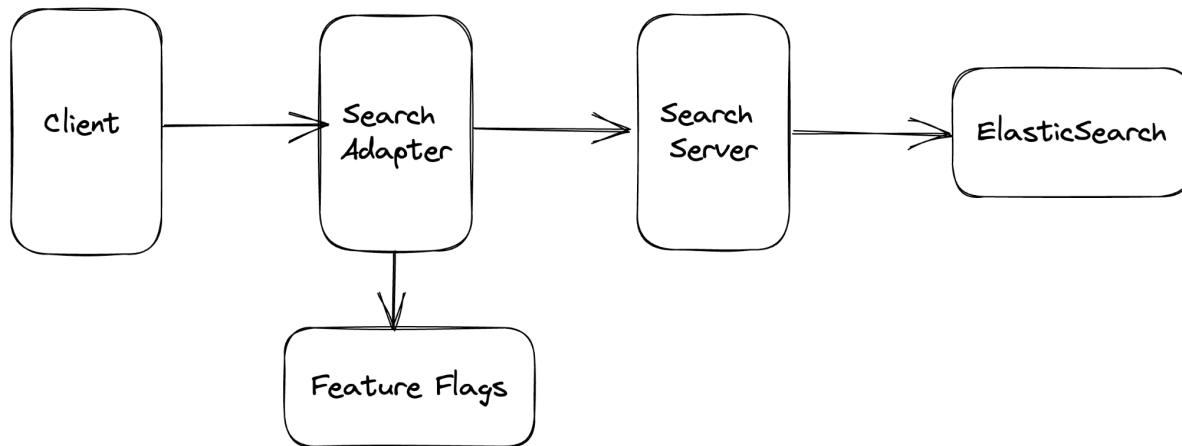
- Over 20 million monthly active users (MAU)
- 100s of millions of active listings in catalog
- 1,000s of queries per second (QPS)



# Introduction

## Search Topology

- “Traditional” term-based search
  - Lucene/Elasticsearch



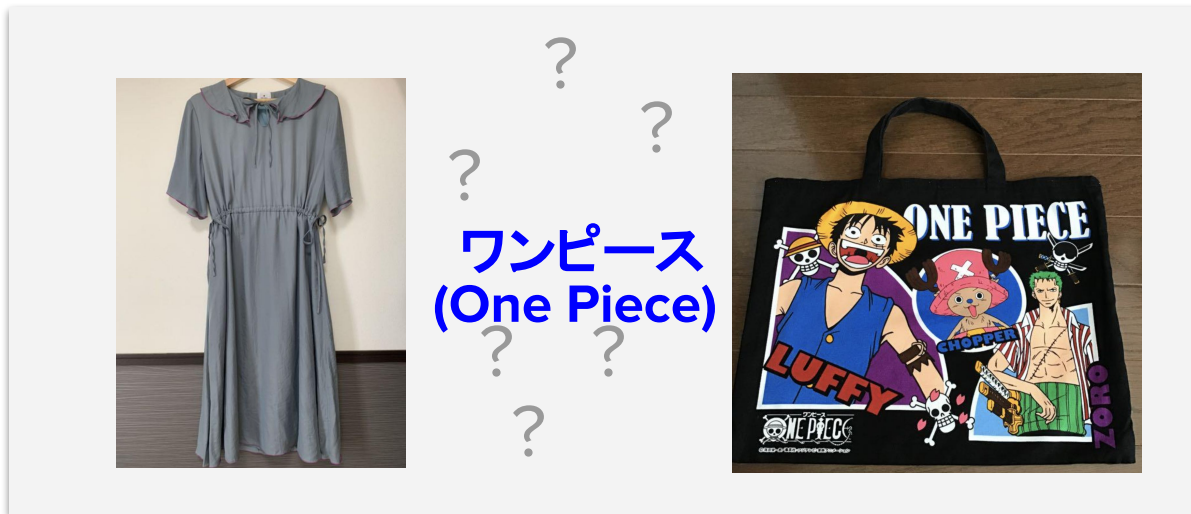
# The Problem

---

# Problem

## Blind spots (that AI can see)

- Ambiguous keywords
- Semantics (“cool toys for boys”)
- Personalization

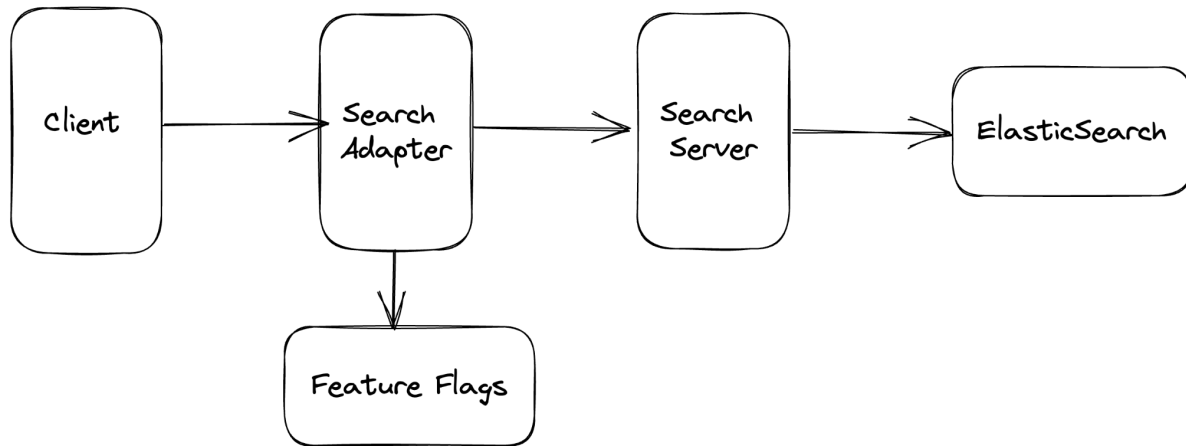




# The Problem

## Integrating ML into a “traditional” term-based search architecture

- Classic search infrastructure and workflow; no “easy hooks” for AI
- Latency budget: 10’s of ms



# The Problem

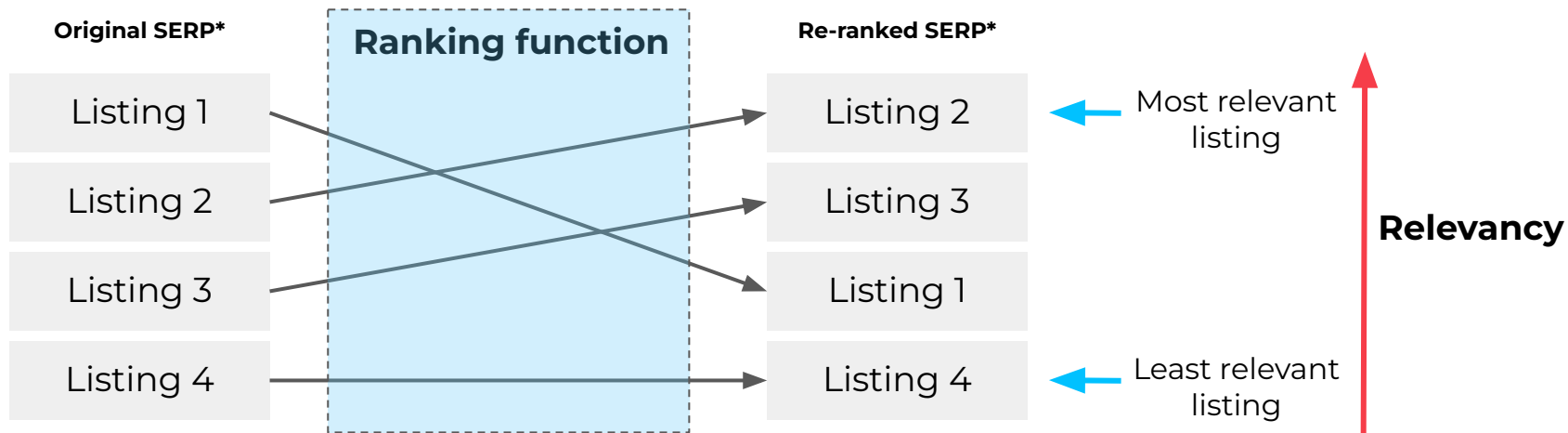
## Integrating ML into a “traditional” term-based search architecture

- Classic search infrastructure and workflow; no “easy hooks” for AI
- Latency budget: 10’s of ms
- User search experience at all costs



# Insight

## Phase 1: use ML to re-rank search results



We want to re-rank the search results so that **more relevant** listings are placed **higher**.

# MLOps

---

**What is it and why do we even care?**

# | What

- ML(Dev)Ops
- “Set of practices that aim to deploy and maintain ML models in production reliably and efficiently<sup>[1]</sup>”

[1] S. Shankar, R. Garcia, J. M. Hellerstein, and A. G. Parameswaran, “Operationalizing machine learning: An interview study,” arXiv preprint arXiv:2209.09125, 2022.

# | Why

- ML application development, deployment, and maintenance challenging

# | Why

- ML application development, deployment, and maintenance challenging
- Variance between data, use-cases, constraints, ...

# | Why

- ML application development, deployment, and maintenance challenging
- Variance between data, use-cases, constraints, ...
- No universal solutions



# | Why

- ML application development, deployment, and maintenance challenging
- Variance between data, use-cases, constraints, ...
- No universal solutions
- MLOps still nascent

# | How



# | How

- AI as an implementation detail



# | How

- AI as an implementation detail
- Use-case-driven “MLOps”



# | How

- AI as an implementation detail
- Use-case-driven “MLOps”
- Iterate on bottlenecks and requirements



# How

- AI as an implementation detail
- Use-case-driven “MLOps”
- Iterate on bottlenecks and requirements
- Good feedback signals



# | How

- AI as an implementation detail
- Use-case-driven “MLOps”
- Iterate on bottlenecks and requirements
- Good feedback signals
- Judicious resource allocation



# How

- AI as an implementation detail
- Use-case-driven “MLOps”
- Iterate on bottlenecks and requirements
- Good feedback signals
- Judicious resource allocation
- Starting small but soon





# Data Pipelines

---

In 5 minutes or less

# | Data Pipelines...?

- A dozen 500+ line SQL files
- Manually executed

# Data Pipelines...?

- A dozen 500+ line SQL files
- Manually executed
- Painful

The screenshot displays a data pipeline execution interface. At the top, a 'Query results' section shows job information: 'Elapsed time' (5 hr 2 min), 'Slot time consumed' (514 days 6 hr), 'Bytes shuffled' (93.48 TB), and 'Bytes spilled to disk' (62.18 TB). Below this, a red banner contains an error message: 'Resources exceeded during query execution: Your project or organization exceeded the maximum disk and memory limit available for shuffle operations. Consider provisioning more slots, reducing query concurrency, or using more efficient logic in this job.' The interface includes navigation tabs for 'JOB INFORMATION', 'RESULTS', 'EXECUTION DETAILS', and 'EXECUTION GRAPH'. A 'PREVIEW' button is visible under the 'EXECUTION GRAPH' tab. At the bottom, another 'Query results' section shows: 'Elapsed time' (4 hr 53 min), 'Slot time consumed' (794 days 6 hr), 'Bytes shuffled' (326.53 TB), and 'Bytes spilled to disk' (376.37 TB). Utility buttons for 'SAVE RESULTS' and 'EXPLORE DATA' are also present.

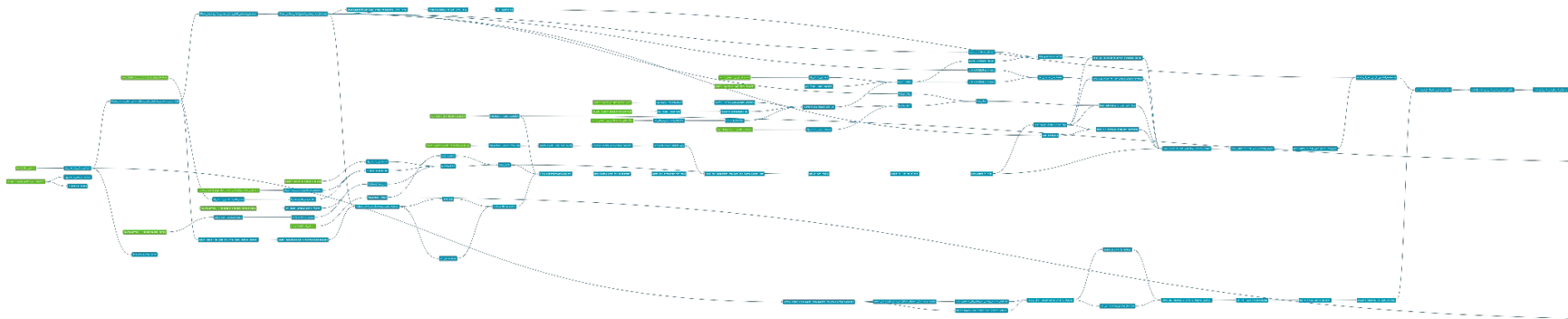
Metric	Value
Elapsed time	5 hr 2 min
Slot time consumed	514 days 6 hr
Bytes shuffled	93.48 TB
Bytes spilled to disk	62.18 TB

**Resources exceeded during query execution: Your project or organization exceeded the maximum disk and memory limit available for shuffle operations. Consider provisioning more slots, reducing query concurrency, or using more efficient logic in this job.**

Metric	Value
Elapsed time	4 hr 53 min
Slot time consumed	794 days 6 hr
Bytes shuffled	326.53 TB
Bytes spilled to disk	376.37 TB

# Data Pipelines...

- A hundred 10-100+ line SQL files
- Inefficient & inaccessible
- Painful





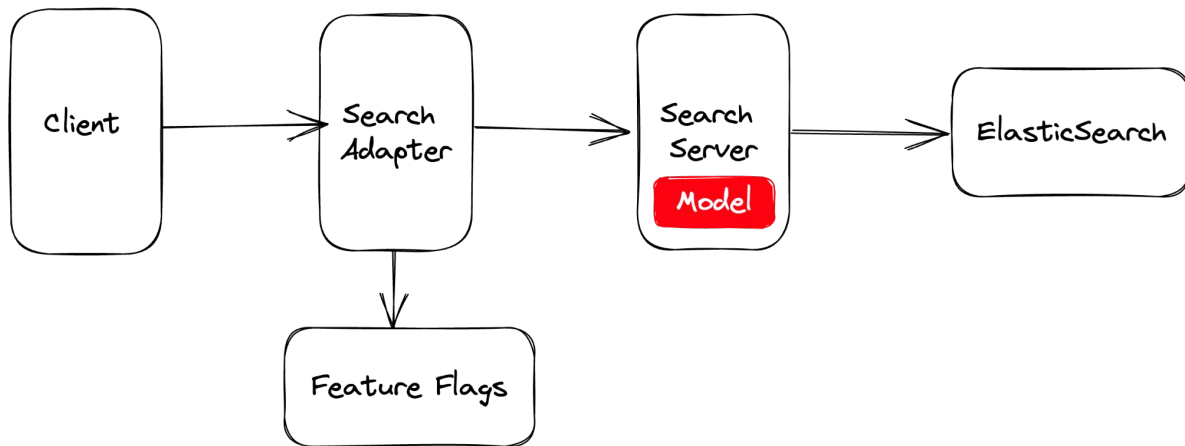
# System Evolution

---

**Growing an ML system while running the business**

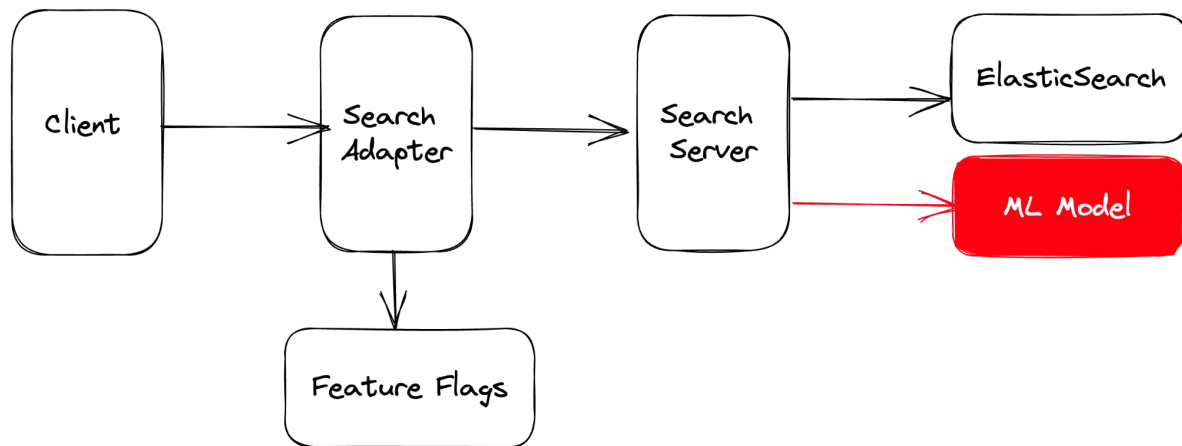
# v0: In-Situ Model Serving

- Model serving within search server
- Features computed in search workflow



# v1: Decoupled Model Serving

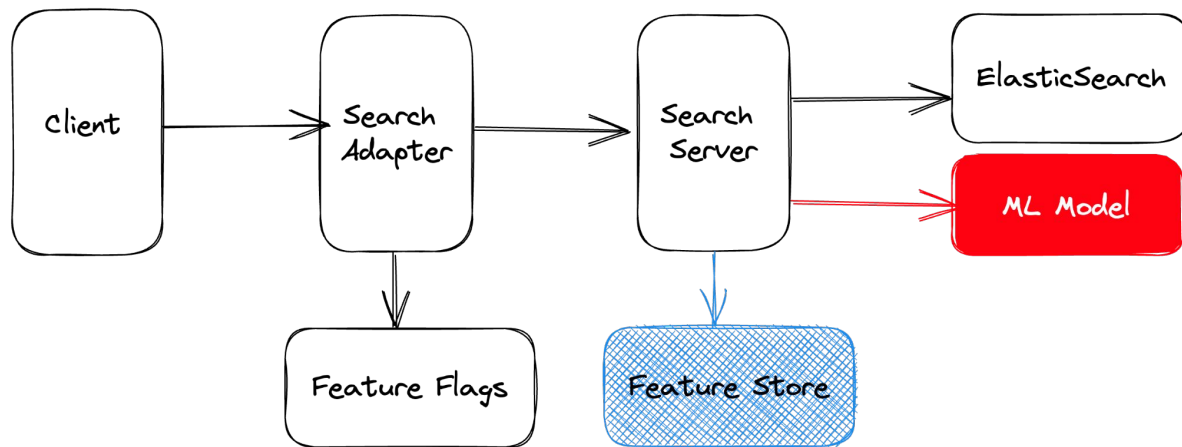
- Custom-made python microservice for model serving
- RPC with timeout and “baseline” response
- Basic monitoring of production metrics






































# v2: Simple Feature Store

- Offline feature store from data pipelines
- Online feature store with direct ETLs
- Timeouts & failsafes redux

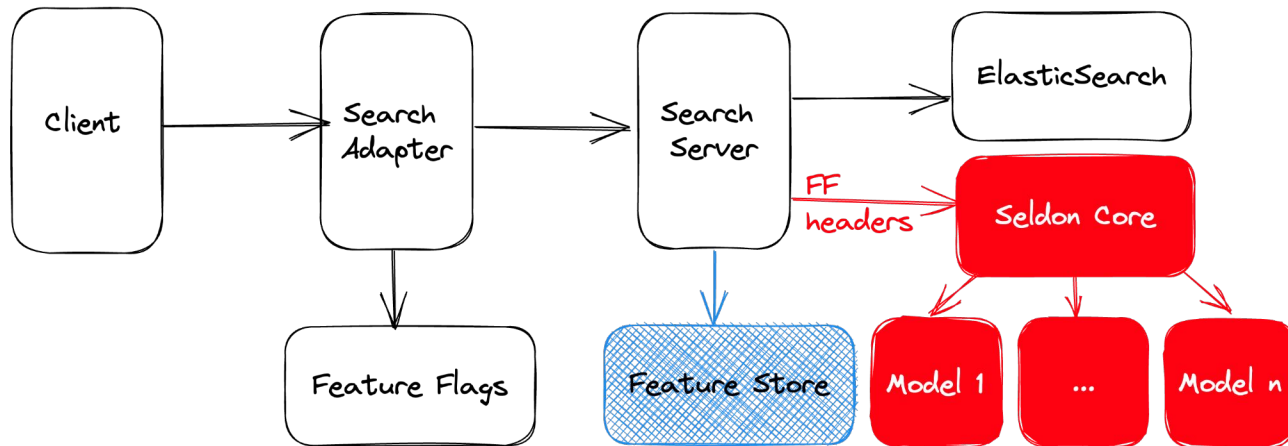


# A/B Test Setup: Before

 [prod] Remove frequent update auth Secret in mercari-searchx-jp ✓ size/XS #60568 by TeoZosa was merged on Jan 23 · Approved		 5
 [prod] Remove frequent update Secret reference in x-server env ✓ size/M #60567 by TeoZosa was merged on Jan 23 · Approved		 11
 Revert "[prod] Remove frequent update ETL Memorystore Secret from x-server env" ✓ size/M #60434 by TeoZosa was merged on Jan 18 · Approved		 14
 [dev] Remove tokenization LTR server resources ✓ size/L #60405 by TeoZosa was merged on Jan 19 · Approved		 2
 [dev] Remove IU frequent update LTR service resources ✓ size/L #60310 by TeoZosa was merged on Jan 17 · Approved		 4
 [prod] Remove IU frequent update LTR service resources ✓ size/L #60307 by TeoZosa was merged on Jan 17 · Approved		 6
 [dev] Remove frequent update ETL Memorystore Secret from x-server env ✓ size/XS #60294 by TeoZosa was merged on Jan 16 · Approved		 7
 [prod] Remove frequent update ETL Memorystore Secret from x-server env ✓ size/M #60291 by TeoZosa was merged on Jan 17 · Approved		 24
 [dev] Remove deprecated A/B test feature flag: TW0-12251_more_frequent_updates ✓ size/XS #60289 by TeoZosa was merged on Jan 16 · Review required		 3
 [prod] Remove deprecated A/B test feature flag: TW0-12251_more_frequent_updates ✓ size/M #60288 by TeoZosa was merged on Jan 17 · Review required		 26
 [dev] Remove deprecated A/B test feature flag: TW0-12250-per_item_stats_tokenization ✓ size/XS #60287 by TeoZosa was merged on Jan 17 · Review required		 17

# v3: Batteries-Included Model Serving Framework

- Seldon & Istio

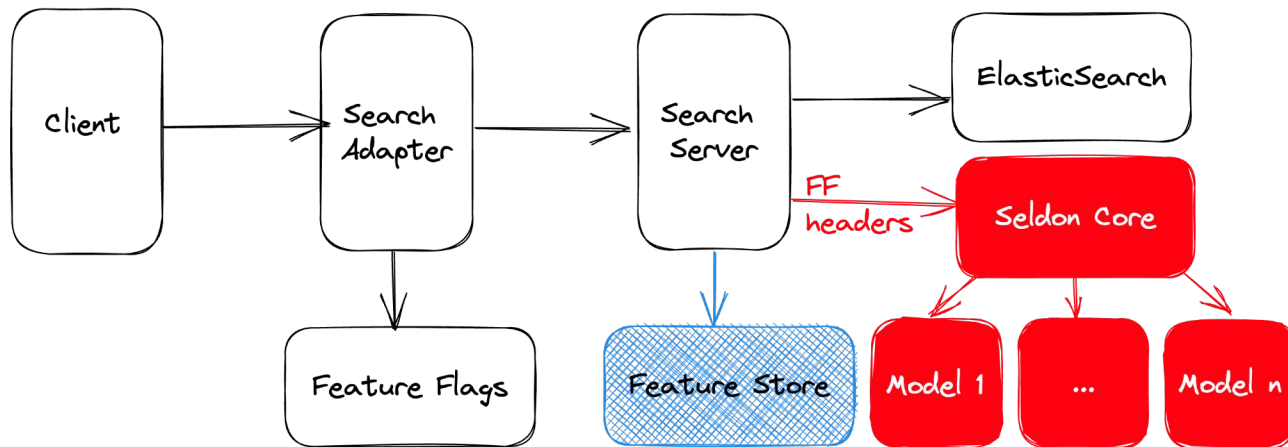


# A/B Test Setup: After

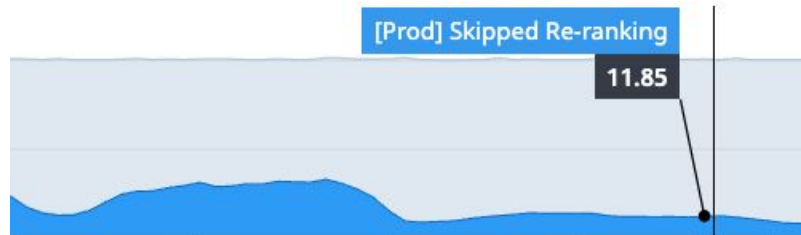
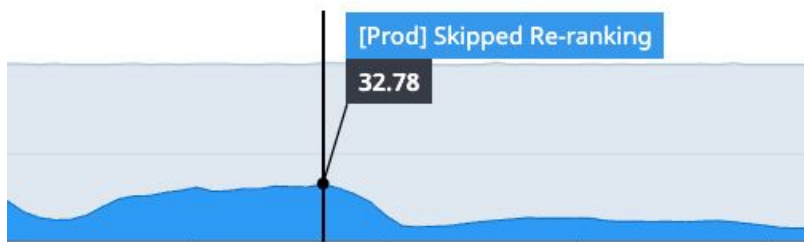
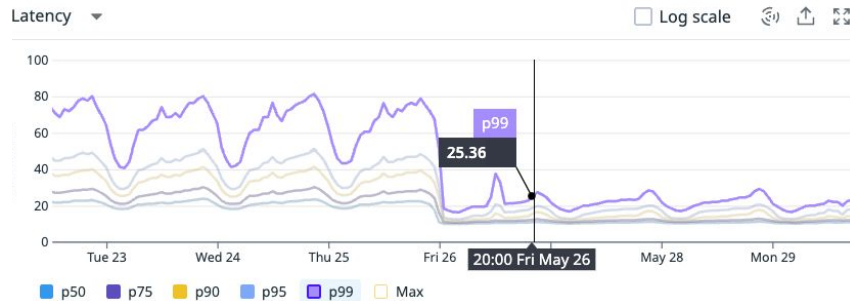
```
1  apiVersion: networking.istio.io/v1beta1
2  kind: VirtualService
3  metadata:
4    name: reranker
5  spec:
6    hosts:
7      - default-model.mercari-search.svc.cluster.local
8    http:
9      - match:
10         - headers:
11             - cool-new-feature:
12                 exact: test
13           route:
14             - destination:
15                 host: cool-new-feature.mercari-search.svc.cluster.local
16             - port:
17                 number: 50051
18           route:
19             - destination:
20                 host: default-model.mercari-search.svc.cluster.local
21             - port:
22                 number: 50051
```

# v3: Batteries-Included Model Serving Framework

- Seldon + Istio
- Shadow Traffic

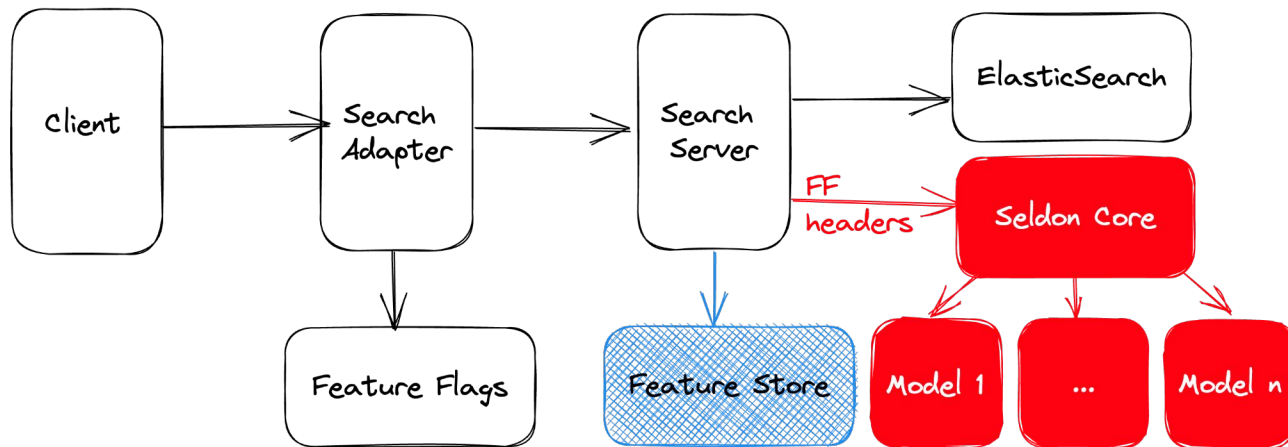


# Shadow Traffic: Test in Prod



# v3: Batteries-Included Model Serving Framework

- Seldon + Istio
- Shadow Traffic
- Fine-grained model serving



# Future Directions

---

**Reliability & Effectiveness**

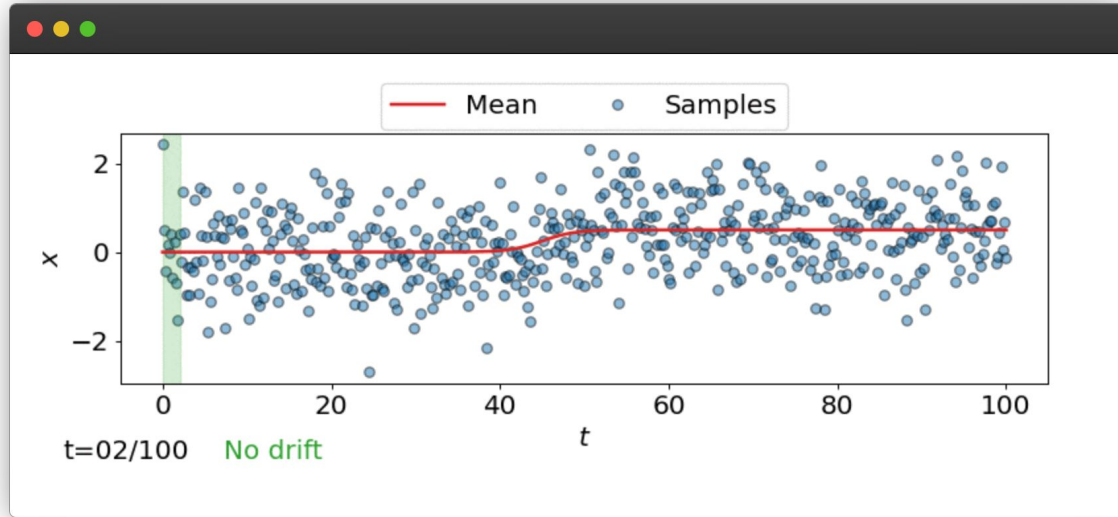


# | Monitoring



Outlier, adversarial, and **drift** detection

# Drift Detection



Detect **drift** to preempt downstream performance degradations

# Conclusion

- ML-enhanced search possible with incremental investments
- Resilience of use-case-driven platforms and systems
- Engineering/business trade-offs
- “One is too small a number to achieve greatness”
- Build something meaningful by building together

