

jxnl.co

@jxnlco

Systematically Improving RAG Applications

Session 4

Split: When to double down vs. when to fold

Jason Liu

The RAG Flywheel


Thesis: The principles we’ve applied in search are highly relevant to what we want to do with RAG

Step	Description	Current steps
1 Initial implementation	Start with a basic RAG system setup	
2 Synthetic data generation	Create synthetic questions to test the system’s retrieval abilities	
3 Fast evaluations	Conduct quick, unit test-like evaluations to assess basic retrieval capabilities (e.g., precision, recall, mean reciprocal rank), and explain why each matters	
4 Real-world data collection	Gather real user queries and interactions. Ensure feedback is aligned with business outcomes or correlated with important qualities that predict customer satisfaction	
<div>RAG Flywheel <i>Post-production</i></div>		
8 User feedback integration	Continuously incorporate user feedback into the system	

The RAG Flywheel

Thesis: The principles we've applied in search are highly relevant to what we want to do with RAG

Step	Description	Current steps
1 Initial implementation	Start with a basic RAG system setup	
2 Synthetic data generation	Create synthetic questions to test the system's retrieval abilities	
3 Fast evaluations	Conduct quick, unit test-like evaluations to assess basic retrieval capabilities (e.g., precision, recall, mean reciprocal rank), and explain why each matters	
4 Real-world data collection	Gather real user queries and interactions. Ensure feedback is aligned with business outcomes or correlated with important qualities that predict customer satisfaction	
5 Classification and analysis	Categorize and analyze user questions to identify patterns and gaps	
6 System improvements	Based on analysis, make targeted improvements to the system	
7 Production monitoring	Implement ongoing monitoring to track system performance	
8 User feedback integration	Continuously incorporate user feedback into the system	



Agenda

Why does segmentation matter

Example: Segmentation in marketing

How segmentation supports decision making

Two types of segments

Food for thought for this session

Sneak peak for rest of course

Example: Segmentation in marketing

Situation:

- You sell a consumer product and run a marketing campaign to boost sales

Complication:

- As a result of your efforts, you discover that there is an 80% boost in sales...but you don't know what is causing the boost
- Alternatively, you could also have an 80% drop in sales...and you don't know what is causing the drop

Approach:

- You dig through your sales data, looking across different customer segments and realize that 60% of the sales increase (or drop) are coming from **Segment 1: 30-45-year-old women living in the Midwest**

Impact:

- With this information, your team can decide:
 - If this is an audience that we want to invest more in
 - How better to target this audience (e.g., not running Super Bowl ads)

Example: Segmentation in marketing

Situation:

- You sell a consumer product and run a marketing campaign to boost sales

Complication:

- As a result of your efforts, you discover that there is an 80% boost in sales...but you don't know what is causing the boost
- Alternatively, you could also have an 80% drop in sales...and you don't know what is causing the drop

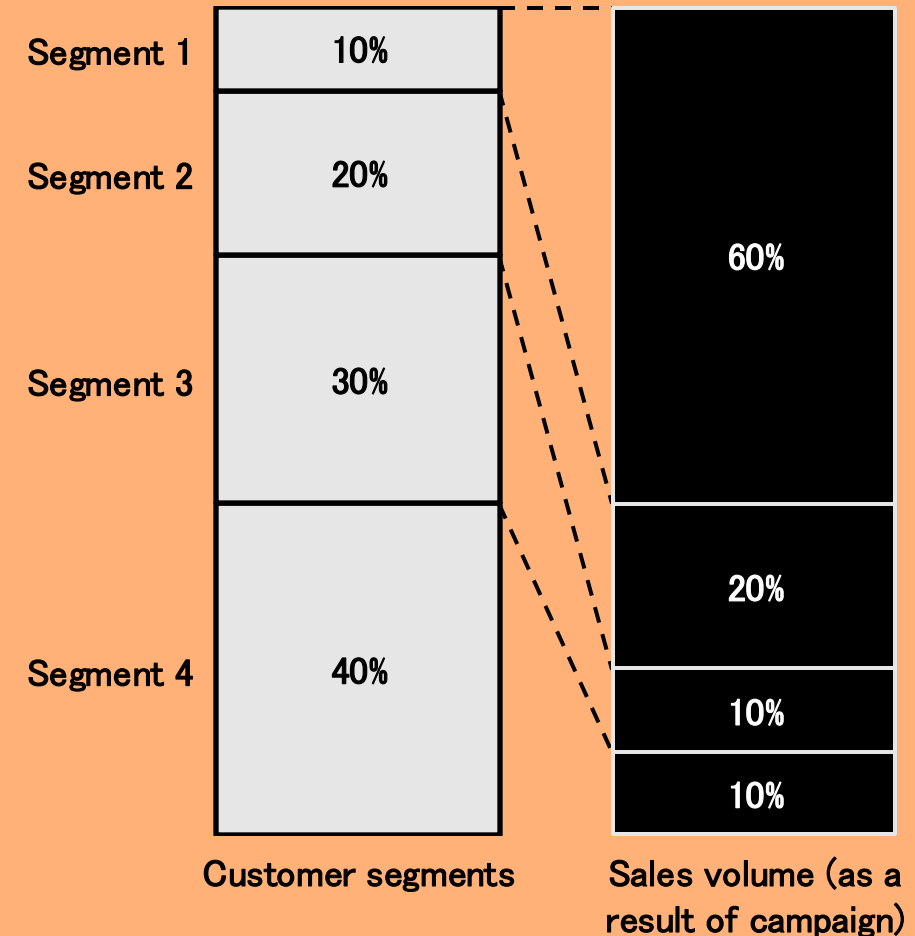
Approach:

- You dig through your sales data, looking across different customer segments and realize that 60% of the sales increase (or drop) are coming from **Segment 1: 30–45-year-old women living in the Midwest**

Impact:

- With this information, your team can decide:
 - If this is an audience that we want to invest more in
 - How better to target this audience (e.g., not running Super Bowl ads)

If you can properly identify the demographics and psychographics of the inputs of your system, your team will have multiple levers to experiment with and allocate resources and start to explore vs. exploit our system



Demographics

The quantifiable characteristics (observable traits) of a given population:

- Role
- Organization ID
- Cohort
- Life stage
- ...

Demographics

The quantifiable characteristics (observable traits) of a given population:

- Role
- Organization ID
- Cohort
- Life stage
- ...

Psychographics

The psychological aspects of consumer/user behavior and preferences:

- Attitudes
- Values
- Interests
- Writing style
- Preferred response style


```
{  
  "query": "What was the difference between the 2022 and 2023 budgets?",  
  "average_similarity": 0.6,  
  "average_cohere_score": 0.8,  
  "customer_rating": 1,  
  "query_types": [  
    "TIME_FILTER", "MULTIPLE_QUERIES", "FINANCIAL_QUERY"  
  ]  
}
```

Agenda

Why does segmentation matter

Example: Segmentation in marketing

How segmentation supports decision making

Two types of segments

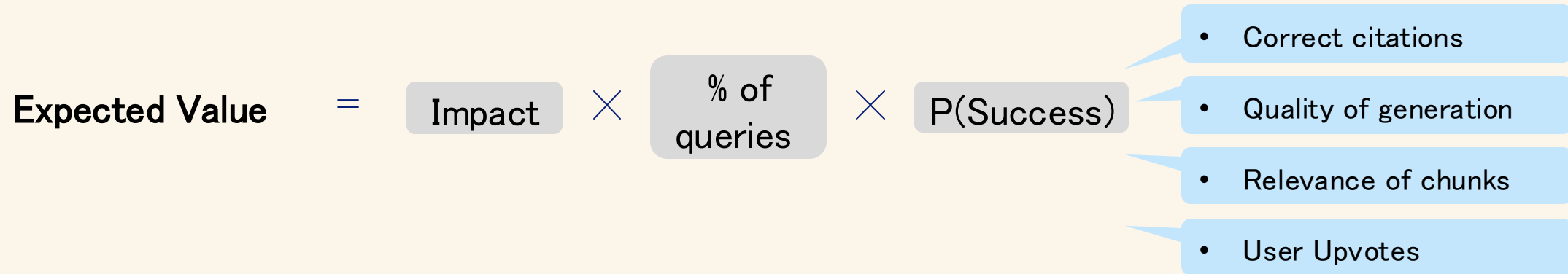
Food for thought for this session

Sneak peak for rest of course

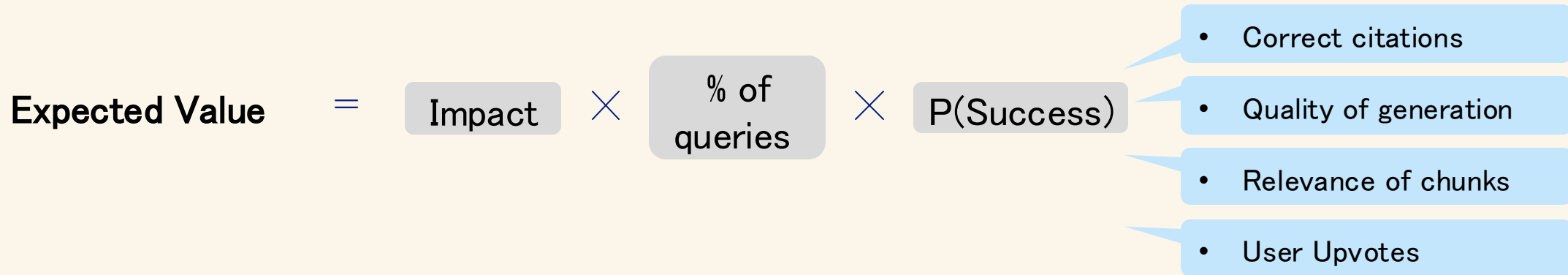
What would we do with our segments?

$$\text{Expected Value} = \text{Impact} \times \% \text{ of queries} \times P(\text{Success})$$

What would we do with our segments?



What would we do with our segments?



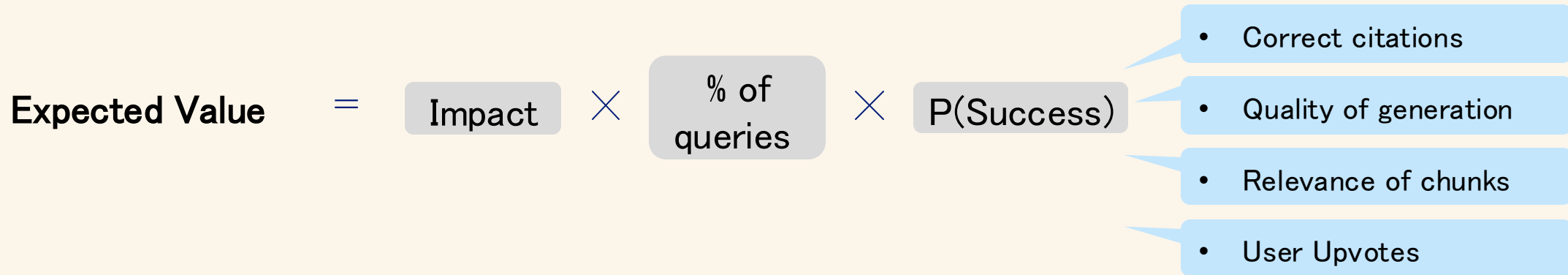
Key takeaway:

If we can label query types we can build specialized systems to maximize impact or p(success)

How do we do this?

- Leveraging clustering methods and few shot classifiers, we can domain model our way into building prompts that can classify and segment queries
- Then we can batch offline and also monitor online in production

What would we do with our segments?



Key takeaway:

If we can label query types we can build specialized systems to maximize impact or p(success)

How do we do this?

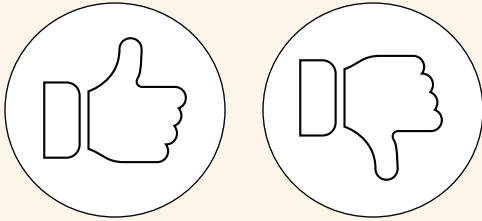
- Leveraging clustering methods and few shot classifiers, we can domain model our way into building prompts that can classify and segment queries
- Then we can batch offline and also monitor online in production

The real challenge:

- Estimate Impact (User Research)
- Measuring Likelihood of success (Collecting User feedback)
- Controlling Query Volume

User feedback as a proxy

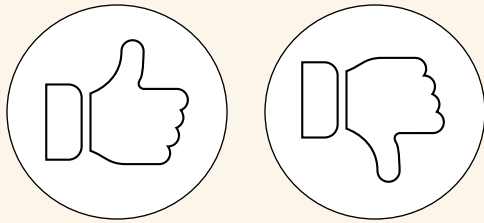
You can start to gather user feedback early on*



* We don't recommend a 5-star system

User feedback as a proxy

You can start to gather user feedback early on*



* We don't recommend a 5-star system

Make sure users understand what the buttons correspond to through your copy

How did we do?

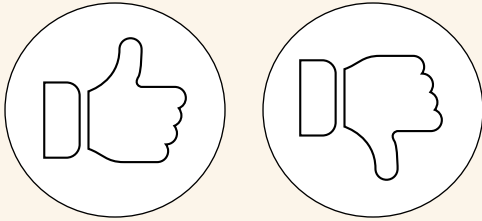
Uncorrelated with customer satisfaction

Did we answer your question?

Highly correlated with satisfaction and correctness

User feedback as a proxy

You can start to gather user feedback early on*



* We don't recommend a 5-star system

Make sure users understand what the buttons correspond to through your copy

How did we do?

Uncorrelated with customer satisfaction

Did we answer your question?

Highly correlated with satisfaction and correctness

Its important to do both UX to collect feedback but also to do user research to understand the impact of questions

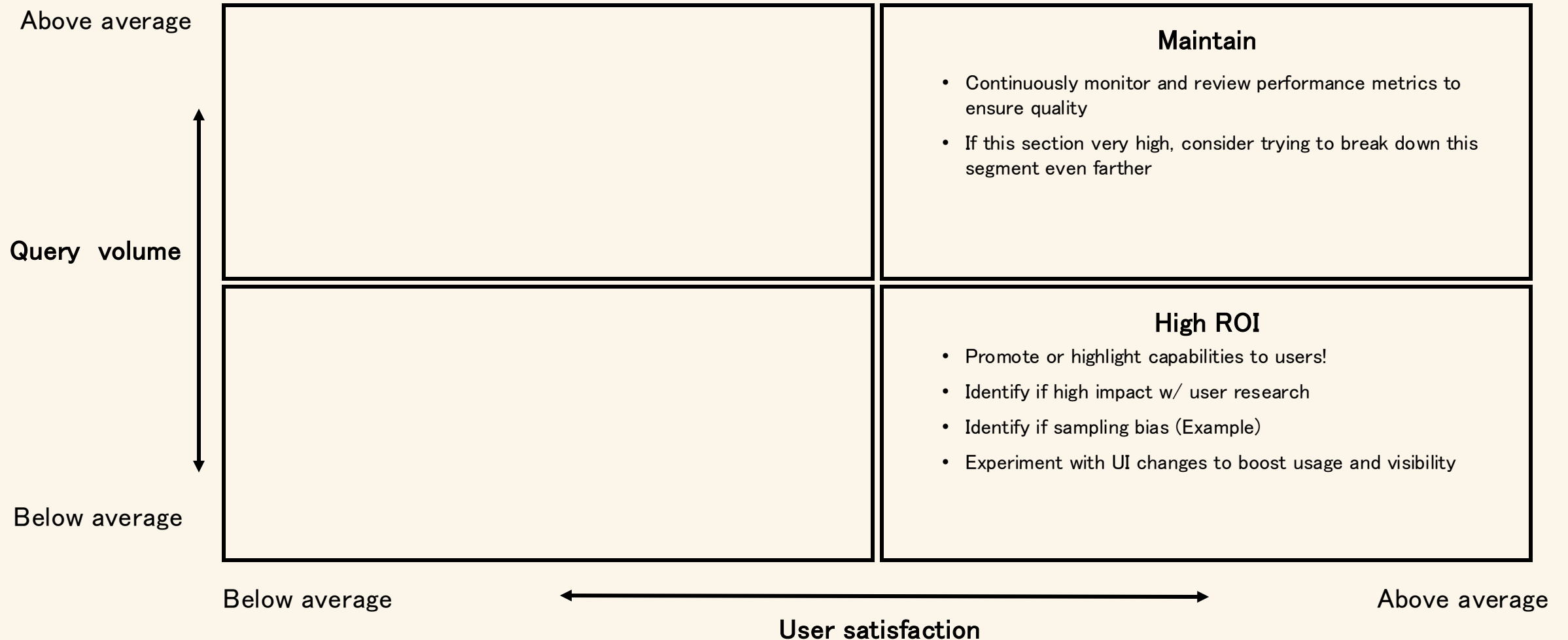
You can also allow users to:

- Rate / Delete sources
- Copy Snippets
- Share
- Publish
- Save
- Etc.

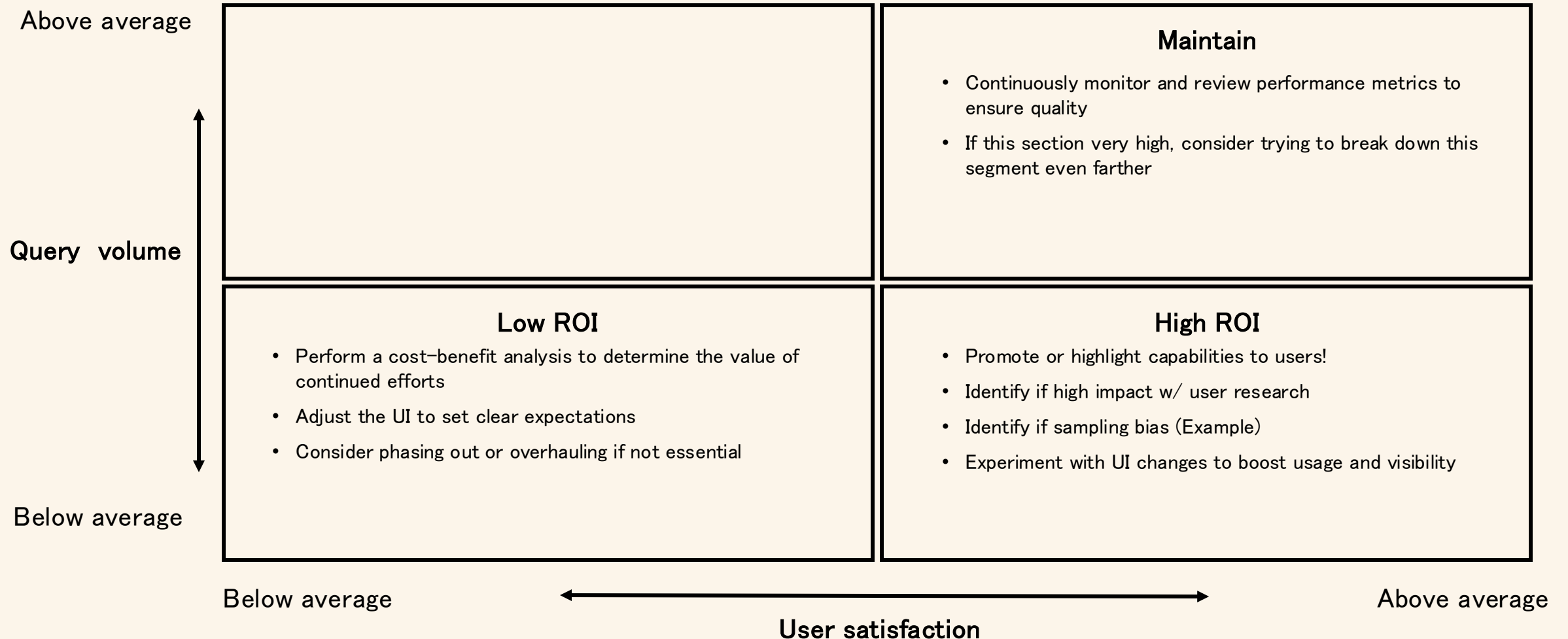
What do we do about different user segments?



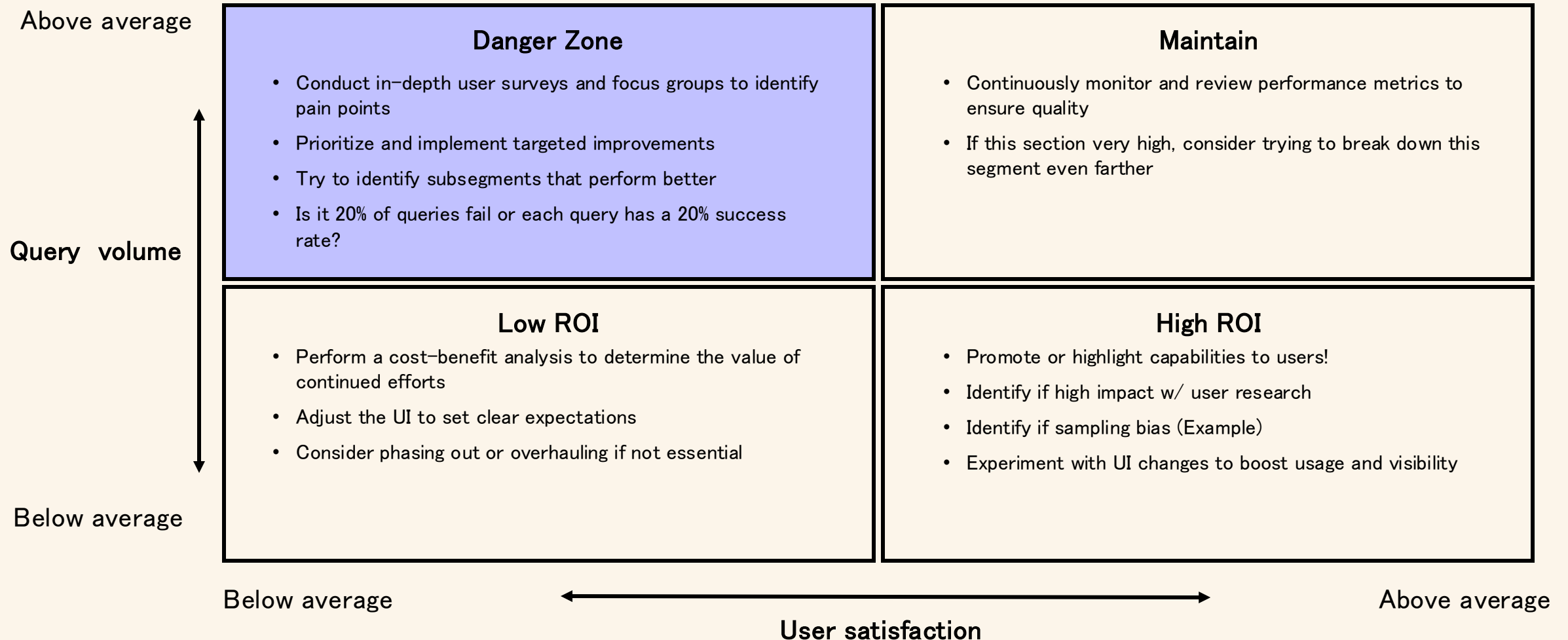
What do we do about different user segments?



What do we about different user segments?



What do we do about different user segments?



Case study: Project Management for Construction Company

Personal anecdote about understanding user behavior and satisfaction

Situation

- Product team hypothesized that scheduling was an important use case for the RAG app
- However, the data showed that users were using the RAG app primarily for document search instead
- >50% queries were document search with ~70% user satisfaction

Case study: Project Management for Construction Company

Personal anecdote about understanding user behavior and satisfaction

Situation

- Product team hypothesized that scheduling was an important use case for the RAG app
- However, the data showed that users were using the RAG app primarily for document search instead
- >50% queries were document search with ~70% user satisfaction

Complication

- By plotting query segments and satisfaction over time, team discovered that:
 - New users started with scheduling questions but had low satisfaction
 - As a result, users shifted to document searches (often related to scheduling)
- High document search satisfaction masked poor schedule search performance

Case study: Project Management for Construction Company

Personal anecdote about understanding user behavior and satisfaction

Situation

- Product team hypothesized that scheduling was an important use case for the RAG app
- However, the data showed that users were using the RAG app primarily for document search instead
- >50% queries were document search with ~70% user satisfaction

Complication

- By plotting query segments and satisfaction over time, team discovered that:
 - New users started with scheduling questions but had low satisfaction
 - As a result, users shifted to document searches (often related to scheduling)
- High document search satisfaction masked poor schedule search performance

Approach

- Eng team shifted focus to systematically improve schedule search to better understand queries about
 - Due dates
 - Payment dates
 - All parties signed off by due date
- Team communicated with clients about new schedule searching capabilities

Case study: Project Management for Construction Company

Personal anecdote about understanding user behavior and satisfaction

Situation

- Product team hypothesized that scheduling was an important use case for the RAG app
- However, the data showed that users were using the RAG app primarily for document search instead
- >50% queries were document search with ~70% user satisfaction

Complication

- By plotting query segments and satisfaction over time, team discovered that:
 - New users started with scheduling questions but had low satisfaction
 - As a result, users shifted to document searches (often related to scheduling)
- High document search satisfaction masked poor schedule search performance

Approach

- Eng team shifted focus to systematically improve schedule search to better understand queries about
 - Due dates
 - Payment dates
 - All parties signed off by due date
- Team communicated with clients about new schedule searching capabilities

Impact

[Next page]

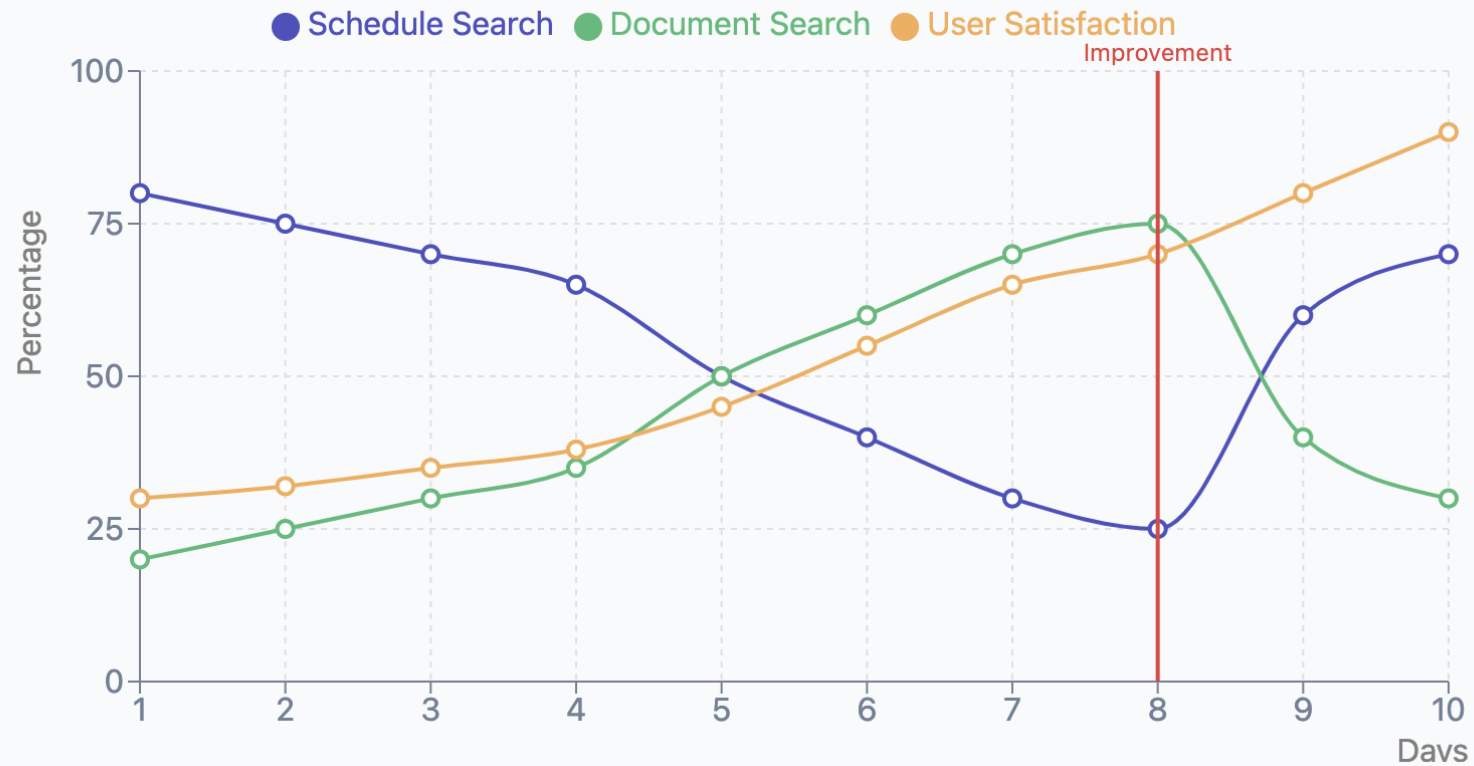
Key takeaway:

- There may be key areas of improvement that are obfuscated by summary statistics (e.g., overall user satisfaction). Don't pat yourself on the back just yet
- Users are savvy and may change their behavior based on the capabilities of your product. Additional research (e.g., focus groups) and analysis are essential to discover why

Case study: Project Management for Construction Company

Personal anecdote about understanding user behavior and satisfaction

User Behavior and Satisfaction





Lesson:

Summary statistics are not enough...
and sometimes may fool you

It's not going to "just work because it's AI"

What RAG actually is:



It's not going to "just work because it's AI"

What RAG actually is:



Key takeaways

- We will need to revert to good old **machine learning and data science**
- We need to conduct **exploratory data analysis** to find segments worth pursuing
- We need to:
 - Work with domain experts to do **feature engineering**
 - Identify **specific candidate indices using additional metadata** that improves performance in high impact segments

Agenda

Why does segmentation matter

Two types of segments

Overview

Lack of inventory

Lack of capabilities

Food for thought for this session

Sneak peak for rest of course

How to distinguish between the two types of issue segments

Problems	Lack of inventory
Origin of issue	<ul style="list-style-type: none">Limited content in knowledge base
How to address this issue`	<p>Expand inventory:</p> <ul style="list-style-type: none">Expand the corpus of informationImprove ingestion and data connectorsCreate more focused sub-systems for specific topics

How to distinguish between the two types of issue segments

Problems	Lack of inventory	Lack of capabilities
Origin of issue	<ul style="list-style-type: none">Limited content in knowledge base	<ul style="list-style-type: none">The system's functional abilitiesMetadata may exist but not structured <p><i>Anything that is not an inventory issue...</i></p>
How to address this issue`	<p>Expand inventory:</p> <ul style="list-style-type: none">Expand the corpus of informationImprove ingestion and data connectorsCreate more focused sub-systems for specific topics	<p>Technical improvements:</p> <ul style="list-style-type: none">Enhance system's ability to understand<ul style="list-style-type: none">Extract additional meta data (e.g., project due date index from proposal docs, map calendar year based on industry)Add new search features based on additional data<ul style="list-style-type: none">Create a new search index (e.g., CRM index, email index, calendar index)

How to distinguish between the two types of issue segments

Problems	Lack of inventory	Lack of capabilities
Origin of issue	<ul style="list-style-type: none">Limited content in knowledge base	<ul style="list-style-type: none">The system's functional abilitiesMetadata may exist but not structured <p><i>Anything that is not an inventory issue...</i></p>
How to address this issue`	<p>Expand inventory:</p> <ul style="list-style-type: none">Expand the corpus of informationImprove ingestion and data connectorsCreate more focused sub-systems for specific topics	<p>Technical improvements:</p> <ul style="list-style-type: none">Enhance system's ability to understand<ul style="list-style-type: none">Extract additional meta data (e.g., project due date index from proposal docs, map calendar year based on industry)Add new search features based on additional data<ul style="list-style-type: none">Create a new search index (e.g., CRM index, email index, calendar index)

Why is this important?

- By categorizing user queries into segments and identifying what is required to address them, developers can more effectively:
 - Prioritize system improvements more effectively
 - Identify gaps in knowledge and functionality
 - Develop specialized subsystems or features for specific use cases
 - Improve overall user experience by ensuring the system can handle a wide range of query types effectively

Agenda

Why does segmentation matter

Two types of segments

Overview


Lack of inventory

Lack of capabilities





Food for thought for this session

Sneak peak for rest of course






Lack of inventory examples

Company	Problem	Solution
<div>Text</div> 	<p>User query: “batteries” or “televisions”</p> <p>Early Amazon: No results for batteries or televisions, only results for books about batteries or televisions</p>	<ul style="list-style-type: none">• Expand inventory to include batteries and televisions

Lack of inventory examples

Company	Problem	Solution
 	User query: “batteries” or “televisions” Early Amazon: No results for batteries or televisions, only results for books about batteries or televisions	<ul style="list-style-type: none">• Expand inventory to include batteries and televisions
 	User query: Spanish telenovelas Netflix: Limited relevant results	<ul style="list-style-type: none">• Produce more TV in different languages for different demographics• Improve subtitles

Lack of inventory examples

Company	Problem	Solution
 	User query: “batteries” or “televisions” Early Amazon: No results for batteries or televisions, only results for books about batteries or televisions	<ul style="list-style-type: none">• Expand inventory to include batteries and televisions
 	User query: Spanish telenovelas Netflix: Limited relevant results	<ul style="list-style-type: none">• Produce more TV in different languages for different demographics• Improve subtitles
	User query: “Greek restaurants near me” Doordash: Limited results	<ul style="list-style-type: none">• Reach out to more Greek restaurants (in specific zip codes) and get more Greek restaurants onto the platform• Buy restaurants iPads to process online orders

Proxies for poor inventory

- Low cosine similarities as a proxy for low relevancy
- Lexical Search returns 0 results
- LLMs not answering questions due to missing data
- LLMs not citing any chunks included in context when returning answers
 - Make sure logging
- Product issues:
 - Problems with data pipeline and data improperly ingested
 - Broken configurations
 - Customers are not providing data which they said they would

Agenda

Why does segmentation matter

Two types of segments

Overview


Lack of inventory

Lack of capabilities



Food for thought for this session

Sneak peak for rest of course




Lack of capabilities examples

Company	Problem	Solution
<div>Text </div>	<p>User query: “Affordable heels with less than 3–inch heel”</p> <p>Amazon: Few relevant results</p>	<ul style="list-style-type: none">• Identify additional feature meta-data and join on existing data• Filter on specific features based on product types

Lack of capabilities examples

Company	Problem	Solution
	<p>User query: “Affordable heels with less than 3–inch heel”</p> <p>Amazon: Few relevant results</p>	<ul style="list-style-type: none">• Identify additional feature meta–data and join on existing data• Filter on specific features based on product types
	<p>User query: “Oscar–nominated films”</p> <p>Early Netflix: Returns results with movie titles or characters which include “Oscar”</p>	<ul style="list-style-type: none">• Acquire additional meta–data for existing catalogue• Join on these datasets to better answer user queries

Lack of capabilities examples

Company	Problem	Solution
	<p>User query: “Affordable heels with less than 3–inch heel”</p> <p>Amazon: Few relevant results</p>	<ul style="list-style-type: none">• Identify additional feature meta–data and join on existing data• Filter on specific features based on product types
	<p>User query: “Oscar–nominated films”</p> <p>Early Netflix: Returns results with movie titles or characters which include “Oscar”</p>	<ul style="list-style-type: none">• Acquire additional meta–data for existing catalogue• Join on these datasets to better answer user queries
	<p>User query: “Chinese food” (it’s after 9pm)</p> <p>Doordash: Limited conversion</p>	<ul style="list-style-type: none">• Figure out how to get up–to–date availability data• Add an “Open Now” button to specify restaurants• New features (e.g., allow users to schedule orders for when restaurants open)

Common (but fixable!) capabilities issues

Problem



Datetime filter for “what happened recently” or “what’s the latest on...”



Comparisons



Filters for Tabular Data in PDF








Specific filters for stock tickers (before a document search)



Understand document metadata

Common (but fixable!) capabilities issues

Problem	Solution
 Datetime filter for “what happened recently” or “what’s the latest on...”	<ul style="list-style-type: none">• “Recent” or “Latest” is contextual based on the query, use few shots<ul style="list-style-type: none">○ “Latest” emails != “Recent” physics research
 Comparisons	<ul style="list-style-type: none">• Requires multiple search queries and a comparison between the two sets of results
 Filters for Tabular Data in PDF	<ul style="list-style-type: none">• Users may want to answer questions over tables using SQL-like behavior• Users may want to search for rows or columns in large data tables like a spec sheet
 Specific filters for stock tickers (before a document search)	<ul style="list-style-type: none">• Find specific ticket, quarter, document type to limit search and generate better results
 Understand document metadata	<ul style="list-style-type: none">• Store meta data (e.g., modification history to understand who last modified the procurement form)• Find me contracts that are unsigned past their due date

Lack of capabilities examples

How to fix a lack of capabilities

- **Query routing:** Run multiple searches in parallel and combine information
- **Extract metadata:** Pre-process data and build new indices to search against, used to filter and sort queries
- **Long context:** Conditionally use long context models rather than RAG when answers are found in short documents rather than many chunks
- **Generation prompt:** Based on the document types change how we generate or render responses



Call out:

Topic Modeling is only a tool to come up with explicit
Classifications

Exploratory Data analysis

- **Test a variety of different hypotheses by running experiments**
- **Determine new capabilities and user requirements:** Work with domain experts, clustering methods and few shot classifiers to find and propose 'segments':

Examples:

- **Searching for contact information** (e.g., render contact cards when the query contains a person)
- **Searching across files** (e.g., determine if a file should be displayed or if audio data should be summarized)
- **Searching across time** (e.g., display a clock widget for time-related queries)

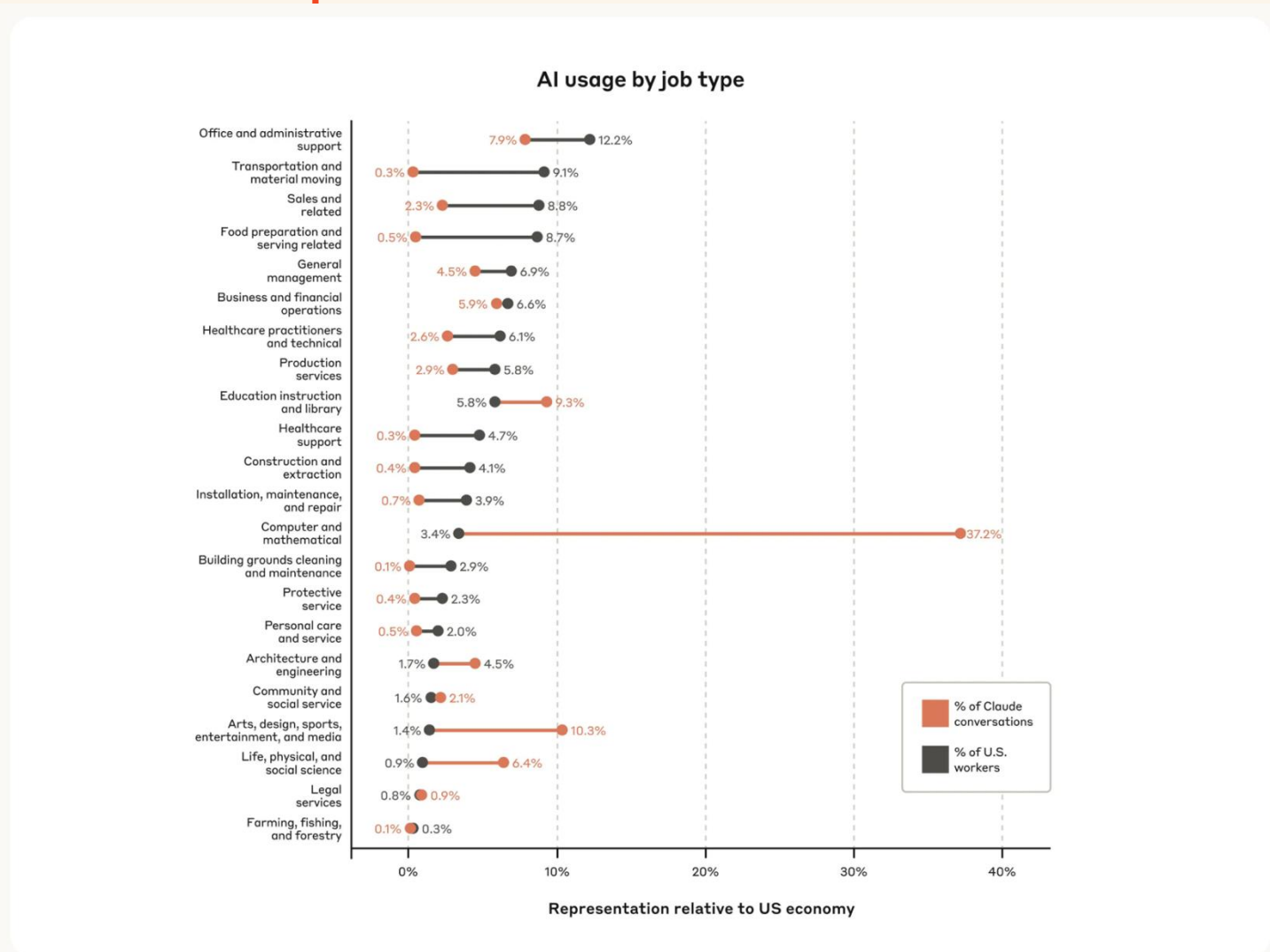
Classification example

```
class QuestionType(Enum):  
    PRODUCT_EDUCATION = "PRODUCT_EDUCATION"  
    NEEDLE_IN_HAYSTACK = "NEEDLE_IN_HAYSTACK"  
  
    QUERY_CONTAINS_PERSON = "QUERY_CONTAINS_PERSON"  
    QUERY_CONTAINS_APP = "QUERY_CONTAINS_APP"  
  
    REQUIRES_APP_FILTER = "REQUIRES_APP_FILTER"  
    REQUIRES_TIME_FILTER = "REQUIRES_TIME_FILTER"  
  
    ANSWER_IS_TIME = "ANSWER_IS_TIME"  
    ANSWER_IS_SUMMARY = "ANSWER_IS_SUMMARY"  
    ANSWER_IS_FILE = "ANSWER_IS_FILE"  
    ⌘L to chat, ⌘K to generate  
    REMINDER = "REMINDER"  
    GENERAL_QUERY = "GENERAL_QUERY"  
    SUMMARIZE_TIME_PERIOD = "SUMMARIZE_TIME_PERIOD"  
    SUMMARIZE_AUDIO = "SUMMARIZE_AUDIO"  
  
    DRAFT_COMMUNICATIONS = "DRAFT_COMMUNICATIONS"  
    QUANTIFY_TIME_SPENT = "QUANTIFY_TIME_SPENT"  
    OTHER = "OTHER"
```

Consider classifying on:

- Question types
- Context recovered
- Search indices hit
- Format type and responses returned

Classification example



For each job type, the percentage of relevant conversations with Claude is shown in orange compared to the percentage of workers in the U.S. economy with that job type (from the U.S. Department of Labor's O*NET categories) in gray.

Anthropic recently also did data analysis on their queries to uncover segments

- Computer and mathematical

<https://www.anthropic.com/news/the-anthropic-economic-index>

Exploration and Monitoring

If you don't have user data...

- A priori zero-shot or few-shot potential topics and capabilities users may have
- Set up monitoring system as you roll out a production system

Exploration and Monitoring

If you don't have user data...

- A priori zero-shot or few-shot potential topics and capabilities users may have
- Set up monitoring system as you roll out a production system

If you do have user data...

- Run topic modeling and clustering
- Present topics
- Identify 5–10 example queries for both satisfied and unsatisfied clusters
- Collaborate with domain experts and user researchers to analyze clusters

Uncovered topic example: customer support

Through topic modeling, you discover that queries about customer support are frequent:

Positive examples

Show me the last 10
support tickets

First 10 customer support
tickets about battery life
complaints

Jason Liu's support
tickets

Uncovered topic example: customer support

Through topic modeling, you discover that queries about customer support are frequent:

Positive examples	Negative examples
Show me the last 10 support tickets	Is Jason a good customer service rep?
First 10 customer support tickets about battery life complaints	Who is likely to churn and why?
Jason Liu's support tickets	What do people complain about most?

Uncovered topic example: customer support

Through topic modeling, you discover that queries about customer support are frequent:

Positive examples	Negative examples
Show me the last 10 support tickets	Is Jason a good customer service rep?
First 10 customer support tickets about battery life complaints	Who is likely to churn and why?
Jason Liu's support tickets	What do people complain about most?

Current performance:

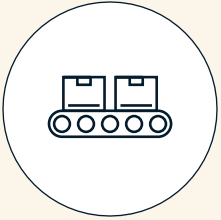
- **No issues:** Finding support tickets
- **Significant issues:** Reporting on Customer Service Rep Metrics
- **Major challenge:** Churn prediction is nearly impossible to reason about for an LLM

Potential solutions:

- Render support tickets in UI
- Implement a tool that renders customer support rep metrics
- Build a model to do sentiment analysis on individual customer service threads or topics

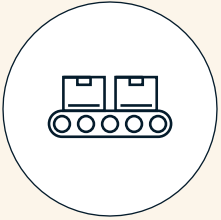
Call out:

Make sure to convert offline analysis to online analysis too



Why monitor topics online?

Automation Paradox: Automation saves you time, but issues will multiply if left unchecked, sampling production is the easiest way to understand what is going on.



Why monitor topics online?

Automation Paradox: Automation saves you time, but issues will multiply if left unchecked, sampling production is the easiest way to understand what is going on.

Instructions for Online Monitoring

- Establish topic clusters / classifications you want to monitor and check against for when new topics emerge
 - Make sure to have an “Other” category and monitor how this fluctuates as new customers are onboarded
 - Detect changes to the system after any product changes or new users
- Build Dashboards:
 - Track Distributions of query types over time
 - Track % Other to detect drift in your systems
 - Track Satisfaction and Volume per Query Type
 - Track Average Relevance per Query Type
 - Track Metrics across Cohorts and Organizations
- Conduct exploratory analysis when systems changes in an unexpected way

Why monitor topics online?

After you attract new users, these users may use the app differently relative to your existing users because of

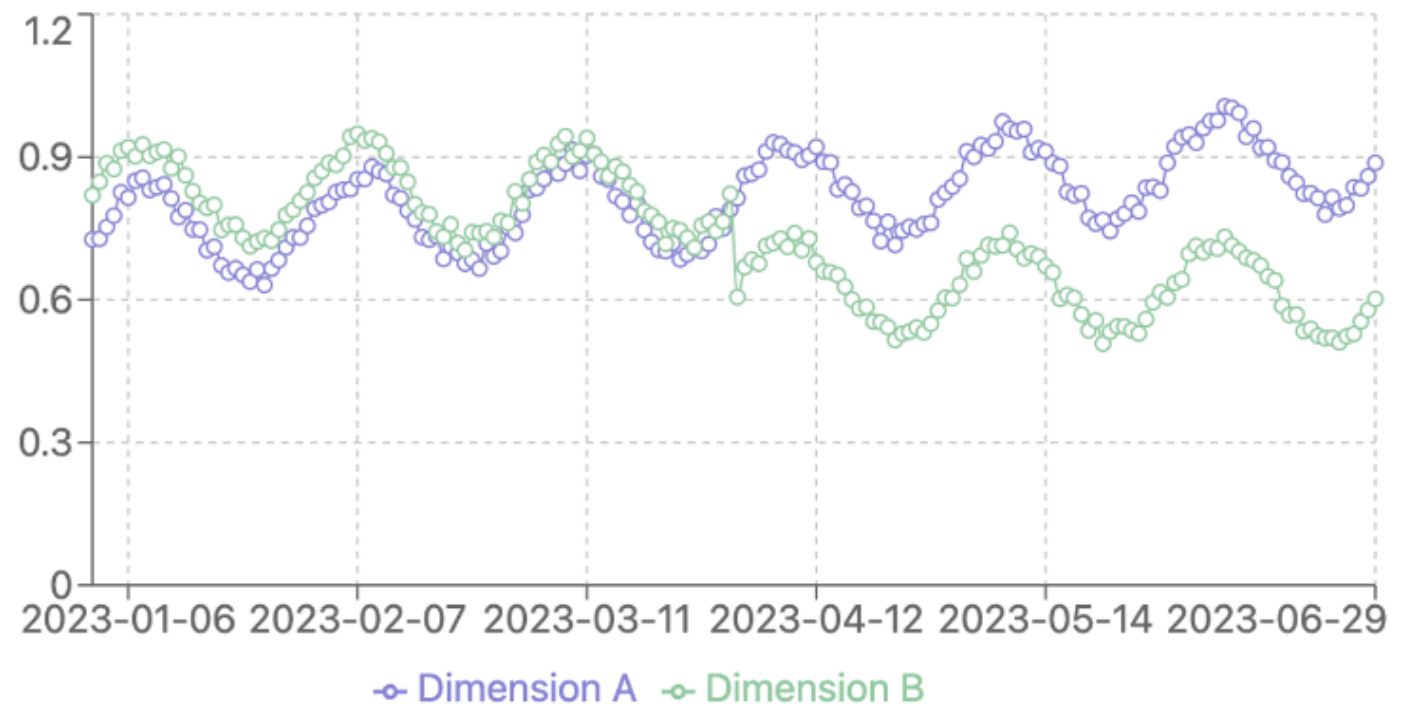
- Demographics
- Psychographics

You can better detect and understand why satisfaction may drift.

For example:

- New users ask different kinds of questions that we need different inventory or capabilities
- Seasonality, etc.

Concept Drift Visualization across Multiple Dimensions



Session 4: Key takeaways

- A** Gain a holistic view of system performance and user engagement

- B** Identify specific areas (e.g., users, content, features) that need the most attention

- C** Tailor your interventions and improvements to address the most impactful issues

- D** Proactively monitor your system to detect evolving user needs and behaviors

The goal is not only to detect concept drift, but also to understand its nuances across all aspects of your system.

This multi-dimensional approach enables more targeted and effective strategies to maintaining and improving your RAG system

Agenda

Why does segmentation matter

Two types of segments

Food for thought for this session

Sneak peak for rest of course

Food for thought: try this at work or in your own projects



Analyze User Queries: Perform topic modeling or batch classification to identify inventory or capability issues



Evaluate answer quality: Identify examples of good and bad responses to guide improvements. Use these examples to guide improvements in the system's response quality.



Implement audit feedback mechanisms:

- Implement user feedback UI
- Assess how often users use the thumbs-up and thumbs-down buttons for feedback
- Analyze if the feedback mechanism is effective and consider changes if necessary (e.g., larger buttons, repositioning).



Make step-wise improvements: Discuss with your team where to make improvements

- What Metadata is missing?
- What filters are missing?
- What indices are missing?

Agenda

Why does segmentation matter

Two types of segments

Food for thought for this session

Sneak peak for rest of course

Sneak peek for rest of course

- **Focus for last session:**
 - The Art of RAG UX: Subtle and obvious ways to build confidence and trust in your RAG app
- **Focus for this session:**
 - Learned to about segmentation and how to distinguish between inventory and capabilities issues and how to approach solving them. We will need to invest a lot of time in extending our capabilities in specific rather than general ways (e.g., solving problems locally within certain segments)
 - Established the importance of production / online monitoring

Focus for next 2 sessions:

- Focus on details of what kinds of capabilities to think about and what to watch out for:
 - **Session 5:** Map – Structured extraction and multimodality
 - **Session 6:** Apply – Routing queries and testing router accuracy