

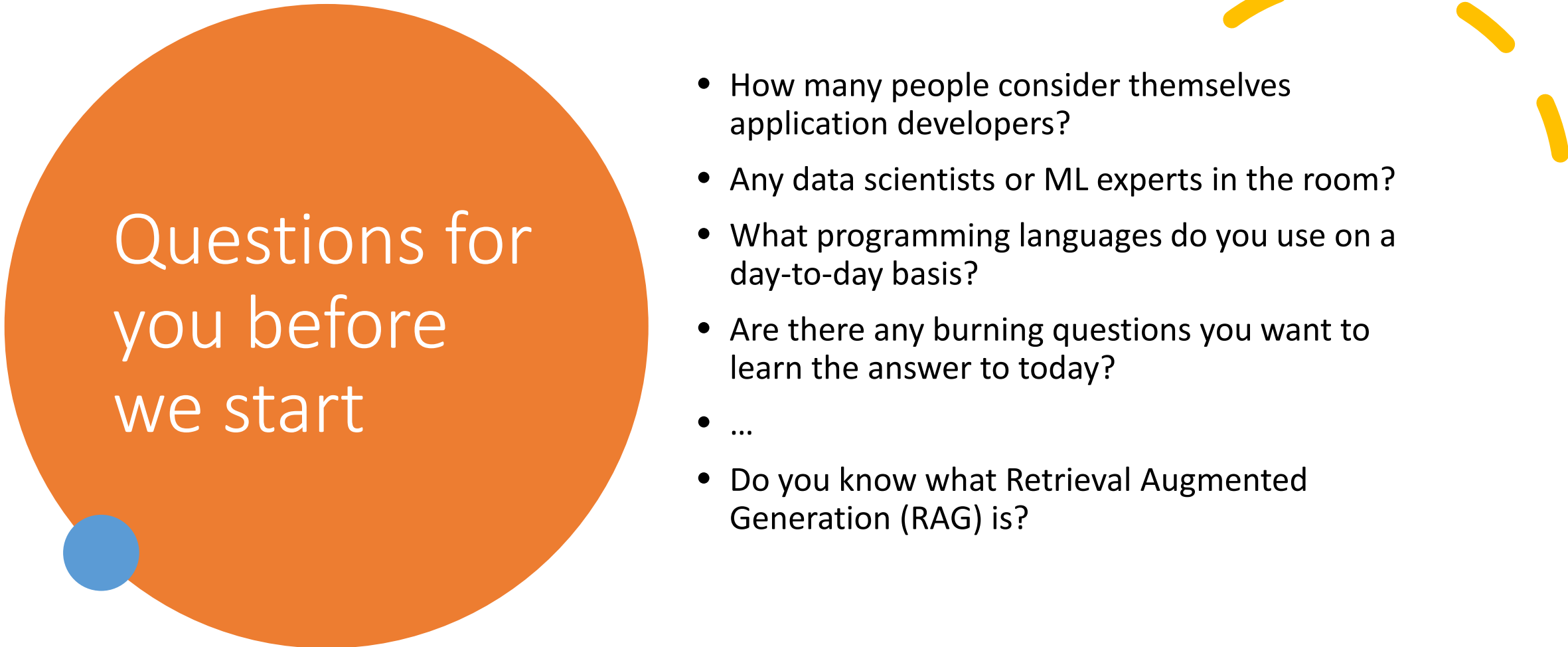


# Getting Started with RAG (Retrieval Augmented Generation)

TechBash  
September 2024



Jason Haley  
@haleyjason



# Questions for you before we start

- How many people consider themselves application developers?
- Any data scientists or ML experts in the room?
- What programming languages do you use on a day-to-day basis?
- Are there any burning questions you want to learn the answer to today?
- ...
- Do you know what Retrieval Augmented Generation (RAG) is?



# ChatGPT

- ChatGPT officially launched November 30, 2022
- By January 2023 had over 100 million users
- Why did it take off so fast?
- Has anything changed in the technology environment because of it?
- Where do you think its going?



# Phases of LLM Usage



General use and awareness



Combining LLM with external data and APIs ← **This is where RAG comes in**



Agents or multi-step workflows with LLMs and other sources



Tools that provide actual business value



# Demo



Setting the Stage: ChatGPT

# Using LLMs in an Application

---

## Bad practice:

LLM as a database

- Can get you into trouble

## Better practice

Retrieval step first, then LLM

- Provide quality context

# Retrieval

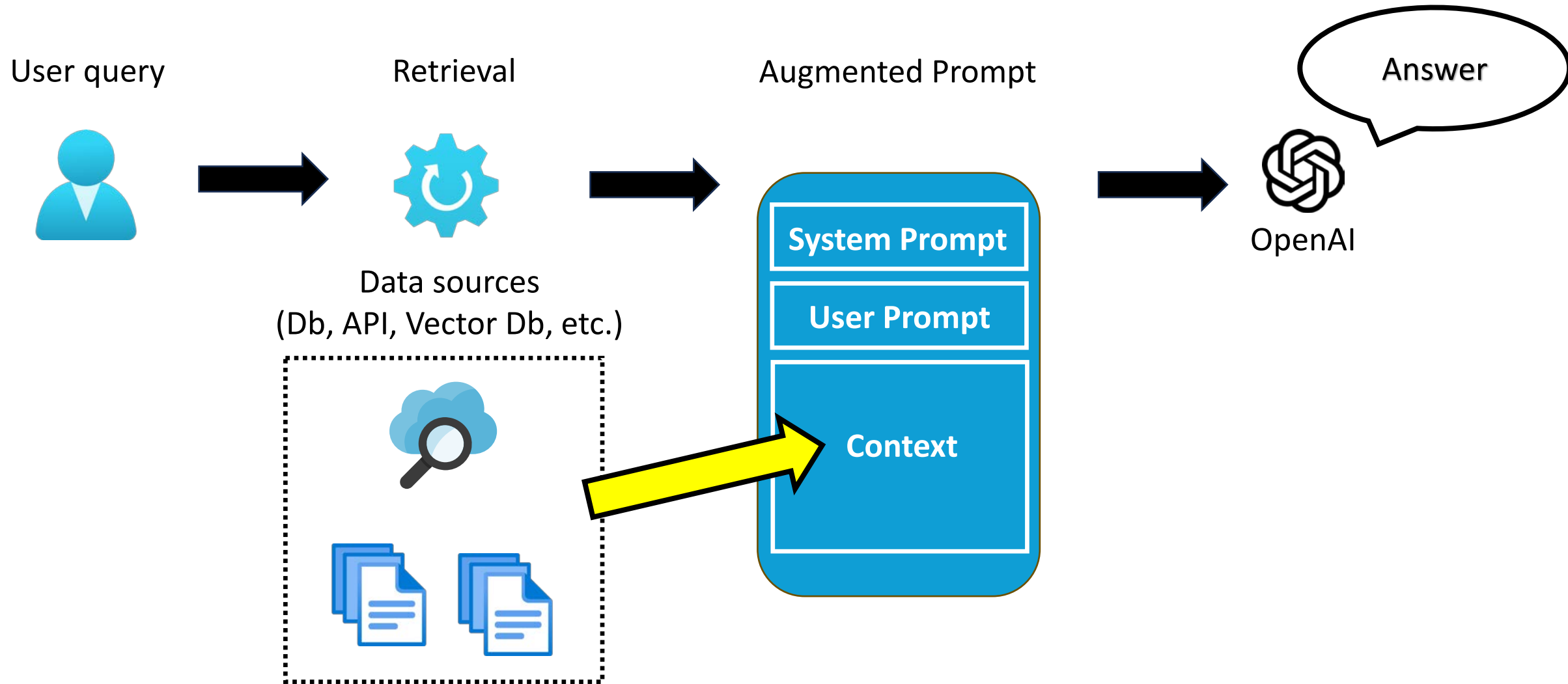


Pre-step of searching your data sources before calling LLM



Goal is to provide knowledge for LLM to use

# Retrieval Augmented Generation Pattern







# Demo:

---

Retrieval with SQL Server

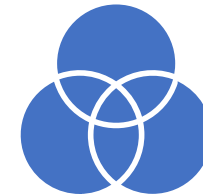
# Searching Types



Keyword



Semantic



Hybrid (with reranking)



# Embeddings and Vector Databases

## OpenAI Embeddings

Text-embedding-ada-002

1536 dimensions

Text-embedding-3-small

1536 dimensions

Text-embedding-3-large

3072 dimensions

## Vector Database

\*Azure AI Search

\*Cosmos DB

\*SQL Azure DB (private preview)

Pinecone

Chroma

Qdrant

Many others...

---

# Chunking and Its Challenges

Process of breaking something up into smaller parts (ie. files or text)

Keep mind: Embeddings encode a meaning for a given chunk of text

What does that mean for blocks of text?

- Fixed Character or token count

  - Overlap (10 – 20 %)

- Paragraphs or sentences

- Pages

- Other

What about tables of text?

What about images or charts?

# Data sources

---

Documents	Chunk up files Preprocess and index embeddings
Database	Determine fields the represent the data meaning Preprocess and index embeddings
External API	The provider will be responsible for doing one of the above



# Demo

---

Look at a RAG Application with More Features



# Demo

---

RAG using Azure SQL Database (currently in private preview)

Announcing EAP for Vector Support in Azure SQL Database  
<https://devblogs.microsoft.com/azure-sql/announcing-eap-native-vector-support-in-azure-sql-database/>



# Chat Completion Request

- Messages
  - **System** – instructions and context for the LLM
  - **User** – user query
  - **Assistant** – LLM response
  - **Tool** – results from function calls
- Other API Parameters
  - **Max Tokens** – for the chat completion to use
  - **Response Format** (text or json\_object)
    - NOTE: you must also instruct the model to produce JSON in the system or user message or it will return an error
  - **Stream** – will send message deltas
    - Good to make application feel faster, though complicates the implementation
  - **Temperature** – between 0 – 2
    - Higher the value the more randomness
  - **Top P** – Top probability 0 – 1
    - Alternate to using temperature
  - **Tools** – list of tools for the LLM to call back to
    - Topic for another talk



# Other Application Concerns

- User login
  - Authentication and Authorization
- Loading & Saving History/Feedback
  - Database on the backend
- Token Usage
- Pre-processing
  - May want to pre-process before calling LLM
- Post-processing
  - May want to show the user only part of the result



# More Application Concerns

---

- Data Privacy
- Latency
- LLMOps and/or DevOps
- Evaluation
  - Groundedness – how well a model’s generated answers align with information from context given
  - Relevance – the extent to which the model’s answers are pertinent and related to the question
  - Coherence – how well the language model’s answers read naturally and is clearly understood
  - Fluency – is the language proficiency of a model’s answers
  - Similarity – quantifies the similarity between the ground truth and the model’s answer

# Some of my resources

---

- RAG Demo Review Series
  - <https://jasonhaley.com/tags/rag-demo-series/>
- Blogs about Semantic Kernel
  - <https://jasonhaley.com/tags/semantic-kernel/>
- Study Note Series
  - <https://jasonhaley.com/tags/study-notes-series/>



# Resources

Link to Presentation



<https://bit.ly/4egvJql>

- Azure OpenAI RAG Pattern using a SQL Vector Database
  - <https://blazorhelpwebsite.com/ViewBlogPost/10067>
- azure-search-openai-demo-csharp
  - <https://github.com/Azure-Samples/azure-search-openai-demo-csharp>
- azure-sql-db-vector-search
  - <https://github.com/Azure-Samples/azure-sql-db-vector-search>
- The RAG Hackathon
  - [https://github.com/microsoft/RAG\\_Hack](https://github.com/microsoft/RAG_Hack)