

BRICS Seminar

AI-Driven Study Repository Searches

AI in Data Mapping and Transformation Tool (DMT)

Biomedical Research Informatics Computing System (BRICS)

January 16th, 2025



MTBI²



| | |
|---------------------------------|--|
| Audio/Video | Please keep your microphone muted |
| Recording | <ul style="list-style-type: none">• Today's session will be recorded• Will be posted on the BRICS website: https://brics.cit.nih.gov/demo |
| Questions & Comments | <ul style="list-style-type: none">• We encourage your participation today• Please use the chat for questions & comments. The chat will be monitored throughout today's demo.• There will also be time after each speaker and at the end of the demo to ask live questions. |



Agenda

| Time | Topic | Speaker(s) |
|------------------|--|-----------------------|
| 9:05 AM-9:35 AM | AI-Driven Study Repository Searches | Omar Kapur |
| 9:35 AM-9:55 AM | AI in Data Mapping and Transformation Tool (DMT) | Colton Farlow |
| 9:55 AM-10:00 AM | Closing Remarks | Dr. Matthew McAuliffe |





BRICS

Biomedical Research
Informatics Computing System

Biomedical Research Informatics Computing System (BRICS)

AI-Driven Study Repository Searches

Omar Kapur



MTBI²



PDBP
Parkinson's Disease
Biomarkers Program

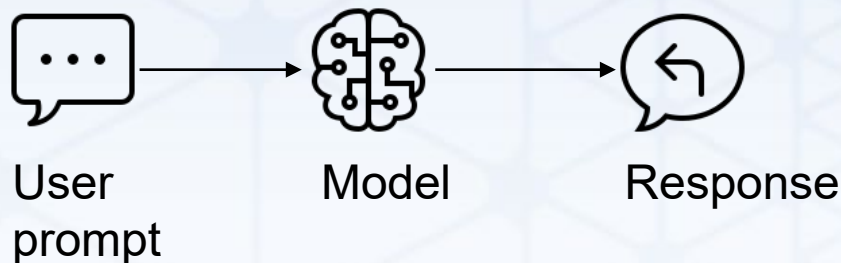
NATIONAL INSTITUTE OF
NEUROLOGICAL
DISORDERS AND STROKE



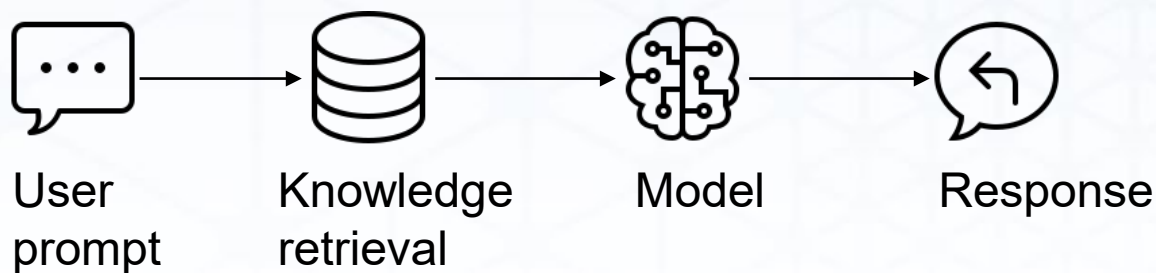


Using Large Language Models (LLMs) for Question Answering

Basic Chain



RAG (Retrieval Augmented Generation)





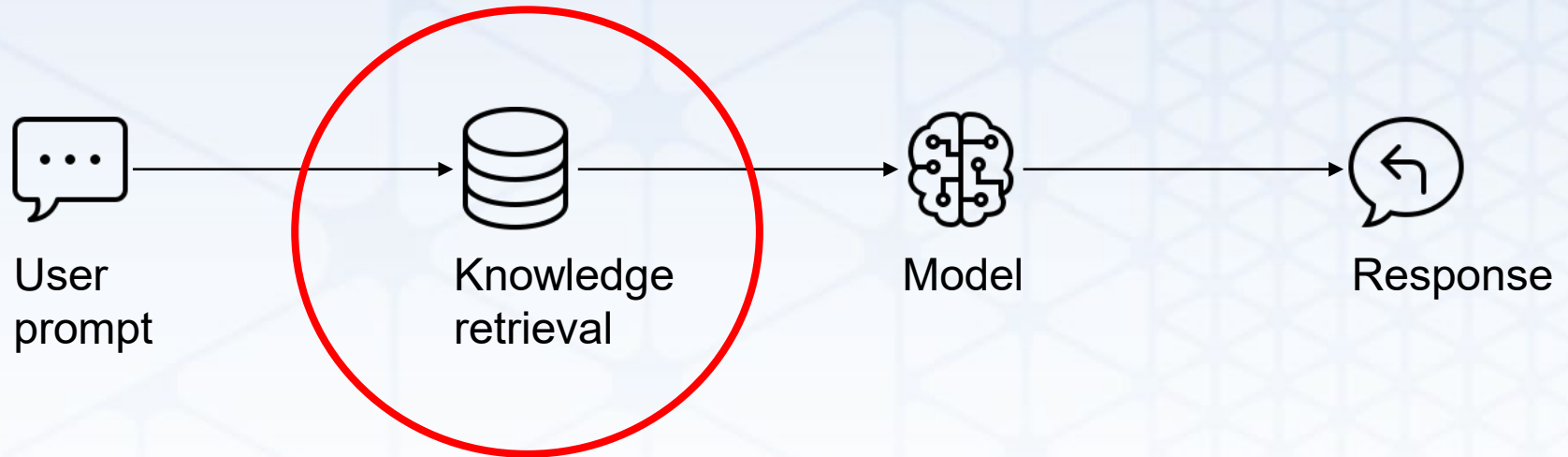
Benefits of RAG

RAG lets you build AI applications that are:

- **More capable:** Does not rely on pre-trained knowledge, but still leverages the language capabilities of the LLM
- **Easier to customize:** Allows for fast development of AI applications that can utilize the desired knowledge from the org
- **More accurate:** Retrieved knowledge is inserted into prompt, LLMs trained to focus on prompt content – reduces hallucination



How does a RAG application really work?



- Knowledge retrieval simply means the relevant content is inserted into the prompt
- How this happens is up to the developer, and depends on the amount of relevant content
- For larger amounts of content, the retrieval mechanism is key
- Another AI tool that can be used here is **Semantic Search using embeddings**





What are embeddings?



- Embeddings are transformations of words, phrases, or entire documents into numerical vectors in a high-dimensional space, using an **embedding model**, enabling machines to process natural language
- This allows us to compare two pieces of text (e.g., a search query and data structure description) on a semantic level

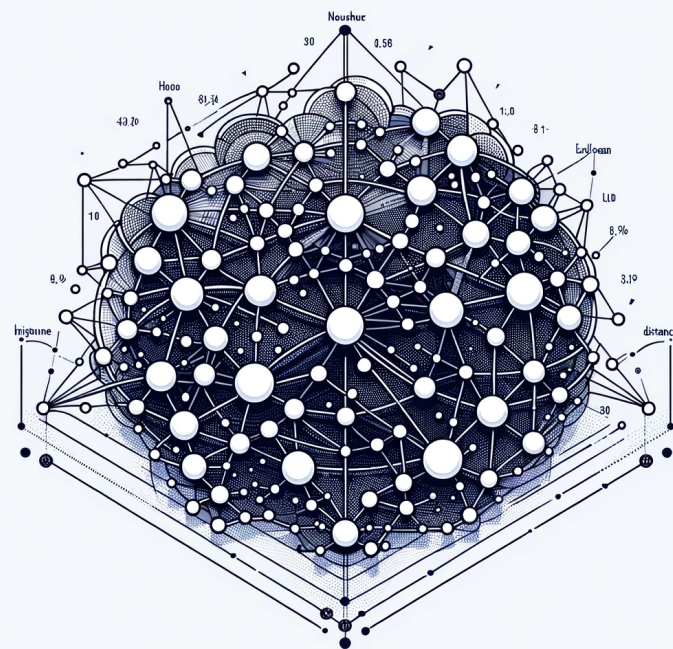




Semantic Search using Embeddings

In order to compare any two pieces of text, we simply need to compare their embedding vectors

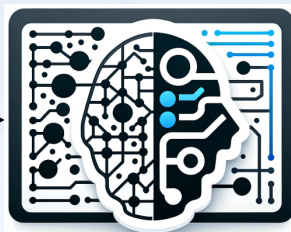
Vector databases offer optimized methods to find the nearest neighbors to a vector – this allows for scaling the search across many vectors



How to create vector databases for an AI application?



Identify data to be used for knowledge retrieval (parse into chunks if long form text)

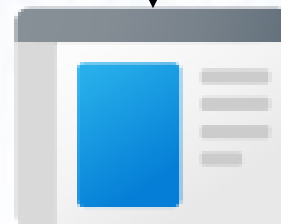


Convert text to embeddings with large embedding models



Store data in a vector database to be queried by app

Create application that utilizes vector database for retrieval



Using a common nomenclature, such as DATMM data model, can enable scaling across data sources

Building RAG applications within BRICS

Demo Use Case: For a given user query, search for relevant repositories of data and relevant studies within those repositories, and provide a summarized response directing them to the repository or study for more information

Use Case Goals:

- Unlock repositories and studies by making them more searchable
- Avoid hallucination and be more accurate and specific in responses with respect to the data studies maintained in repositories



Building RAG applications within BRICS

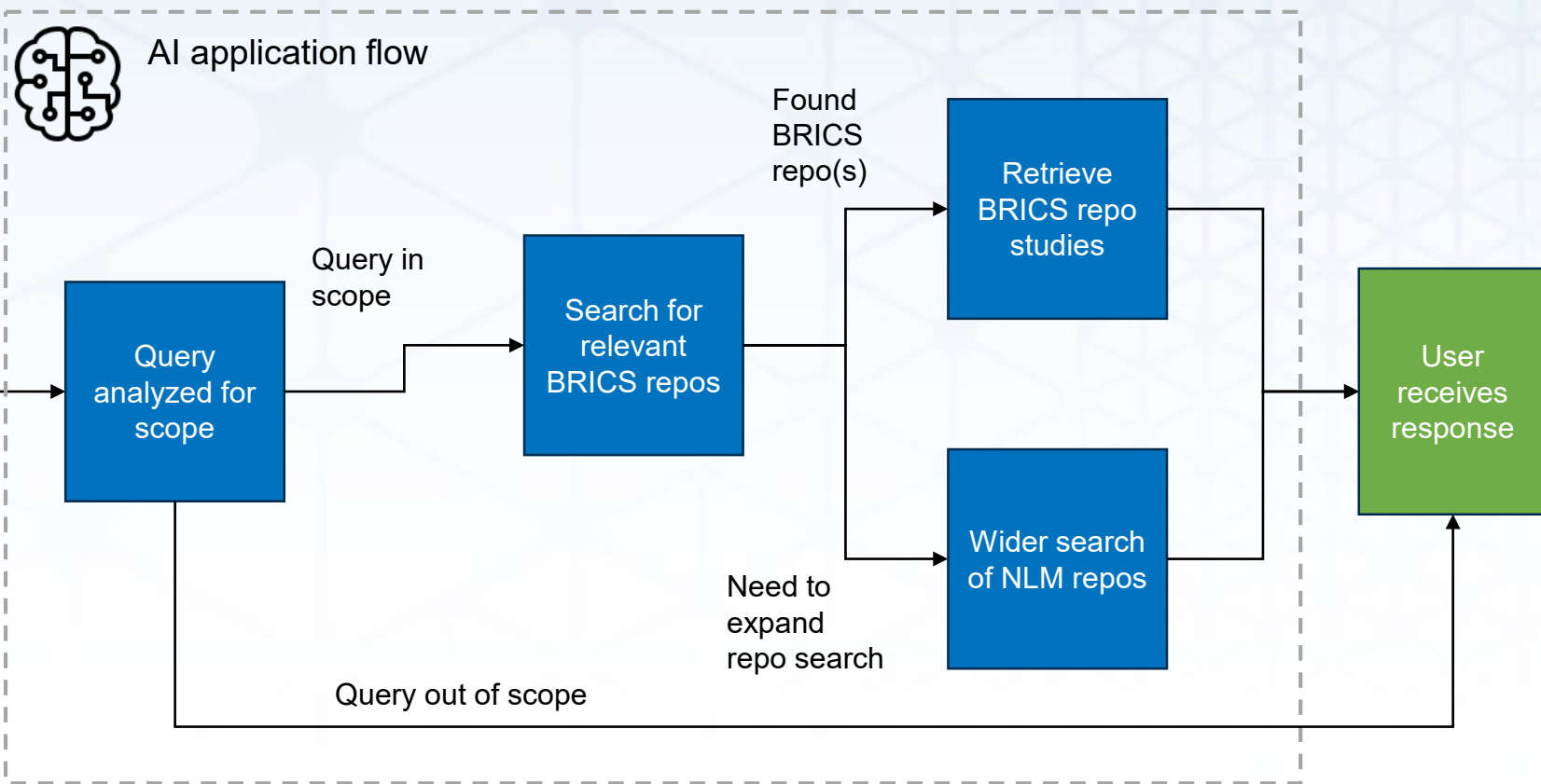
Demo Tech Stack

LLM: Claude 3.5 Sonnet

Embeddings: Amazon Titan

Python AI Framework: Langchain & Langgraph

POC app: Streamlit





Demo



BRICS

Biomedical Research
Informatics Computing System

Questions?





BRICS

Biomedical Research
Informatics Computing System

Biomedical Research Informatics Computing System (BRICS)

AI in Data Mapping and Transformation Tool (DMT)

Colton Farlow

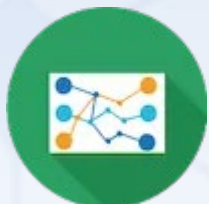


MTBI²





Data Mapping and Transformation Tool (DMT)



The Data Mapping and Transformation Tool (DMT), also known as the Extract, Transform, Load (ETL) Tool, converts data from external systems that may not fully adhere to BRICS CDE standards into a BRICS-compliant format.

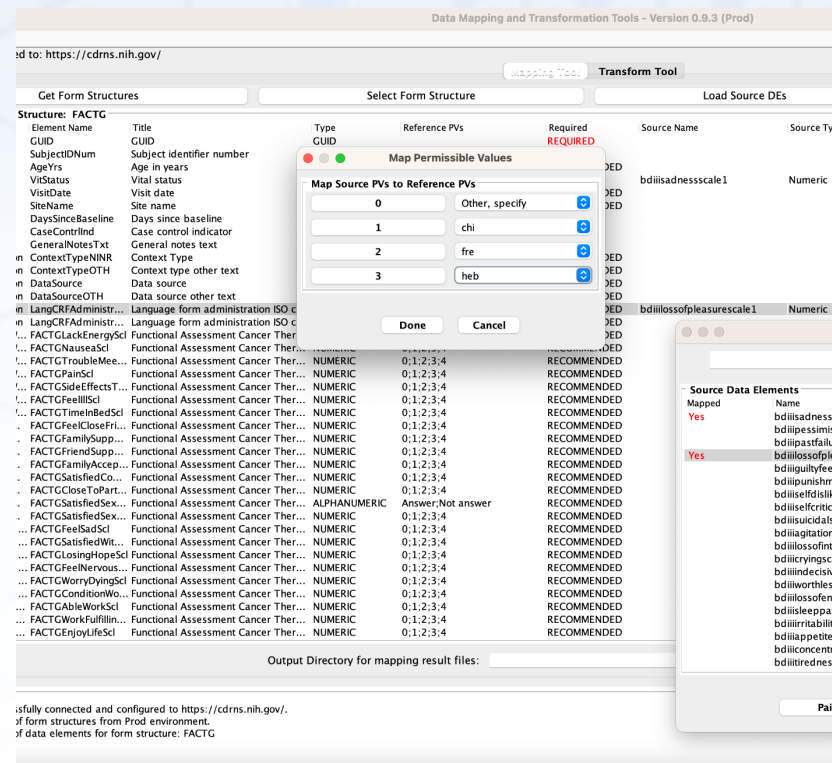
We have recently revamped the tool, and a new version is now available, which we will demo to showcase its enhanced capabilities.

Key features:

1. Maps non-BRICS Data Elements to BRICS Data Elements. – Demo
2. Exports data as BRICS-compliant CSV files for validation and upload.
3. AI integration with AIMS4DE, an AI Semantic Search Tool:
 - AIMS4DE is a stand alone tool but can be integrated with other applications or tools
 - Light-weight open-source model and hosted on-prem.

What Was the Old DMT Tool?

- Webstart tool
- Needed downloads from BRICS public site
- It looks a little something like this:



What Was the Old DMT Tool?

- A few issues with the old tool:
 - UI/UX was less intuitive
 - Required manual operation of every Data Element pairing task
 - Concerns of continued support
 - Old webservice calls
 - Deprecated code libraries

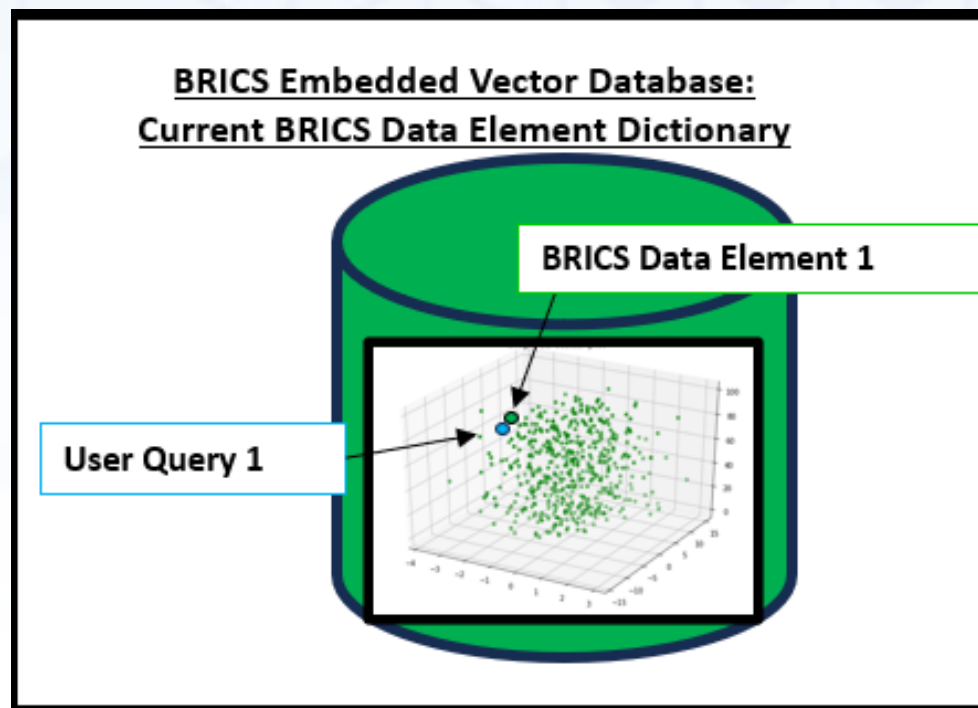


What is the New DMT Tool?

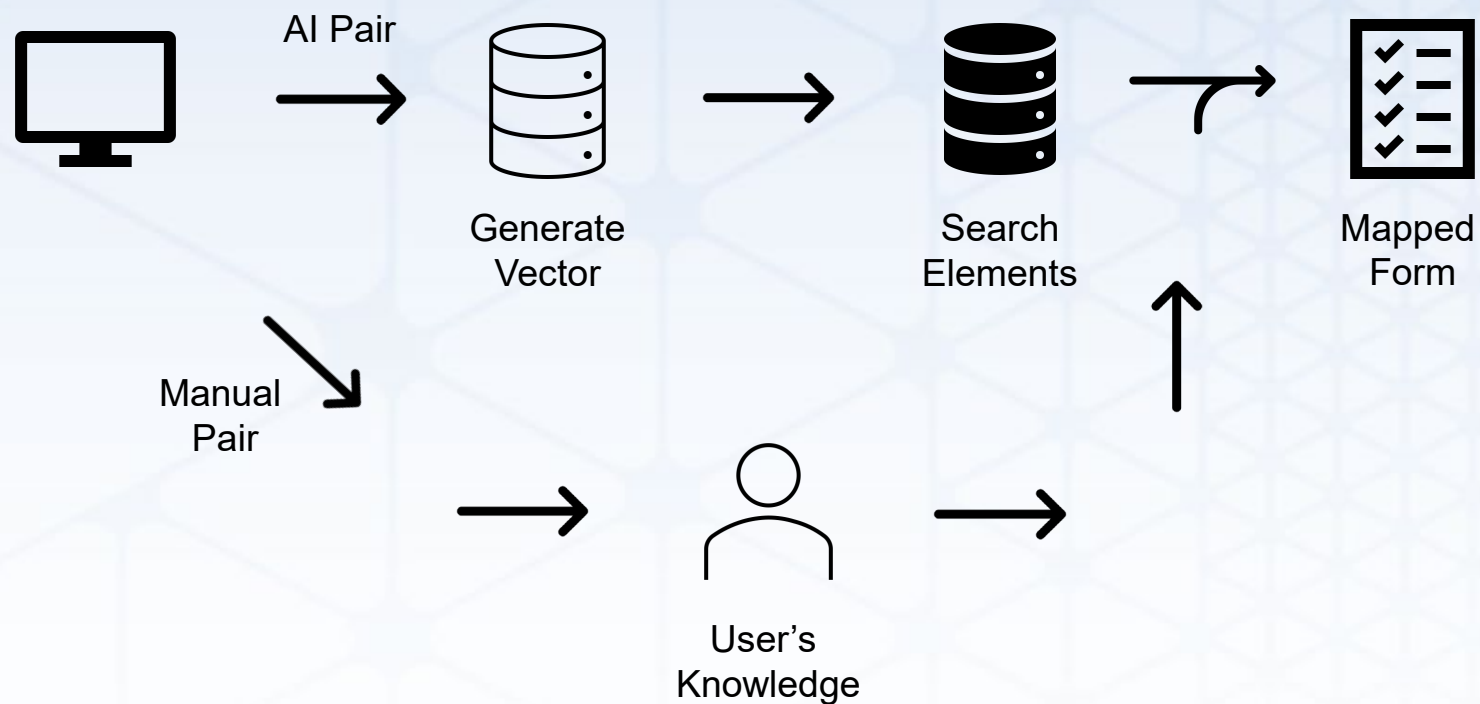
- The Webstart tool was the perfect candidate to modernize and innovate
- It's the old tool, but better!
 - AI pairing for Data Elements when it makes sense, and user intervention when it doesn't
 - New UI/UX design, more readable and user-friendly flow
 - API calls to BRICS microservice endpoints
 - Tool is available on-demand in-browser without download
- Incorporation of PoC AI semantic search previously developed for other applications in BRICS

AI Integration

- AI pairing is accomplished through interfacing with AIMS4DE, an AI Semantic Search.
- Semantic Search is different from traditional searching in that it uses an AI embedding model to search for items it calculates to have similar meaning.
- By providing Data Element (DE) descriptions to the AIMS4DE model, we can semantically search for DEs with the same intent and scope for their given forms.
- For individual pair, it's easy. Just select the highest certainty result.



AI Integration



Both the user and the AI model are valid paths to a fully mapped form!



Further Applications for AI Semantic Search

- Bulk AI pairing of multiple source data elements at once, in addition to the single AI pairing already available.
- Potential for pairing of Data Elements' permissible values (PVs), based off PV descriptions, similar to DE descriptions.
- AI pair-append based off ranked certainty scores where multiple source data elements can pair to a BRICS data element.





Demo



BRICS

Biomedical Research
Informatics Computing System

Questions?





BRICS

Biomedical Research
Informatics Computing System

Thank you for joining in!

