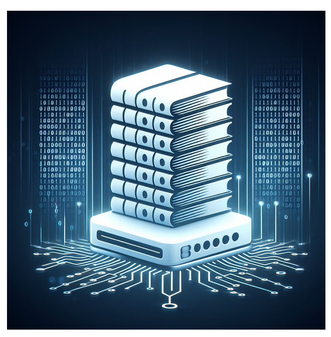
Towards Declarative Querying of Text Data

And a system that does it well

Sepanta Zeighami, Aditya Parameswaran
UC Berkeley

Background

- Information stored in free-flowing text
  - News
  - Financial documents
  - Conversations transcripts
  - Social media posts
- **End goal:**
  - A database system that also stores text
    - Answer queries on information stored in such text documents
    - Allow querying together with other data sources

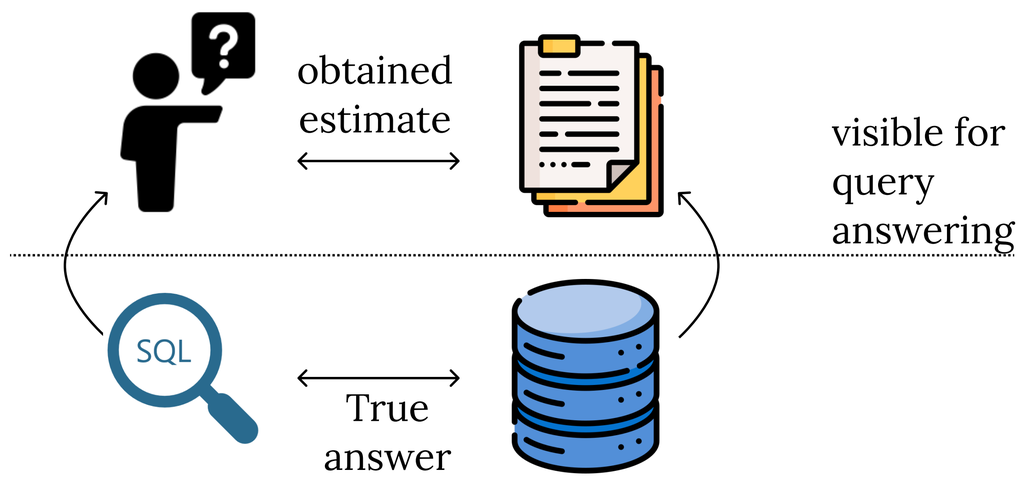


System Objectives

- A system supporting such queries should satisfy the following desiderata
  - **Well-Defined Query Semantics**
    - Queries should have well-defined expected answer
    - Outputs should have well-defined format
    - System should have well-defined set of supported queries
  - **Optimized Query Planning**
    - Queries are often decomposed into NLP model calls
    - System should design query plans to answer queries while minimizing cost
  - **Integration with Database Systems**
    - Be well-optimized to support queries across relational and text data sources
- Current RAG systems allow arbitrary queries over arbitrary text, but do not satisfy any of the desiderata

Defining Query Semantics

- Text data: materialized view of an underlying relational DB
- Text query: approximation of SQL query on underlying DB
- Goal: answer query on underlying DB given access only to text view
  - Query semantics defined based on underlying DB
  - Allows formulating SQL queries over text
- Possible only under conditions on how view and query were generated



Case Study on Medical Transcripts

- Text data of transcripts of doctor-patient interactions
- Answer the query:
  - Assumes an underlying database with a patient table
  - Output is a list of patient names

```
transcript:1 [doctor] hi , andrew , how are you ?
[patient] hi . good to see you .
[doctor] it's good to see you as well . so i know t
[patient] sure .
[doctor] okay ? so , andrew is a 62-year-old male w
[patient] uh , so , over the the weekend , we've be
[doctor] okay . is , is one knee worse than the oth
[patient] equally painful .
[doctor] okay .
[patient] both of them .
```

Problems with RAG:

- Unclear what the query input should be
- Unclear how it should be executed
- Sub-optimal accuracy/cost trade-offs