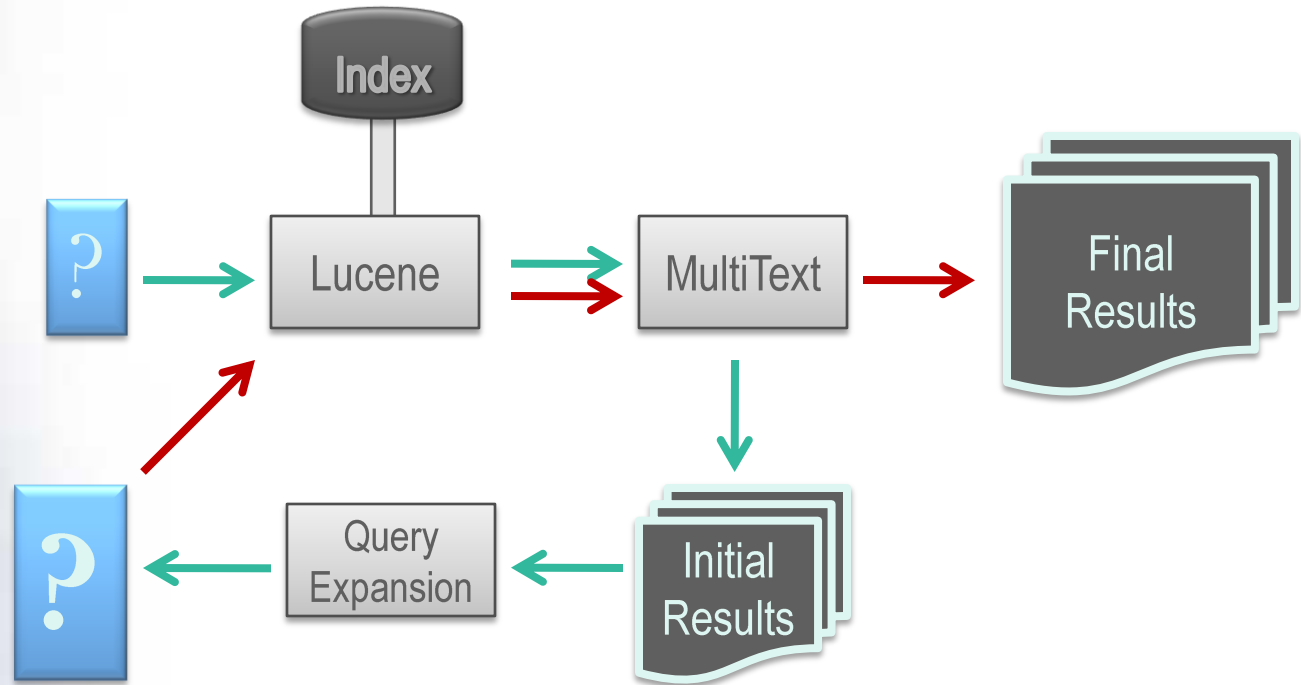
The background of the slide features a dark blue field with numerous bright blue, glowing light trails. These trails are composed of multiple parallel lines that curve and intersect, creating a sense of motion and depth. The lines vary in thickness and intensity, with some appearing as sharp, bright streaks and others as softer, more diffuse bands. The overall effect is reminiscent of a digital or data visualization environment.

# **Deliverable #3:**

## **Passage Retrieval**

Thomas Phan  
Jason Shaw  
Varden Wang

# Workflow



# Lucene

- Followed standard Lucene practices and examples
- Index of TREC 2005 documents was 3.3 GB
  - Stored entire document body in index

# MultiText

- Implemented from scratch
- Based on the description in Clarke *et al.* (TREC 9 submission)
- True IDF measure used
- Find “covers” -- passages bounded by query terms
  - Length of “cover” not accounted for
  - Stopping performed by Lucene
  - Return 500- or 1000-character passage centered on cover

# Query Expansion

- Based on Xu & Croft (1996)
- Queries unfocused, ambiguous, or lack detail
- Local Context Analysis
  - Co-occurrence of query terms with passage terms
  - Unable to tag corpus to discover NPs
  - Added terms with score  $> 5$ 
    - In practice, about 4.8 terms per query

# Results

- Document Retrieval

QE Level	None	S > 5	+20	+70
MAP	<b>0.2981</b>	0.2115	0.1735	0.1327

- Passage Retrieval (MRR)

	500	1000
Before QE	0.285/0.145	<b>0.380/0.183</b>
After QE	0.255/0.114	0.336/0.141