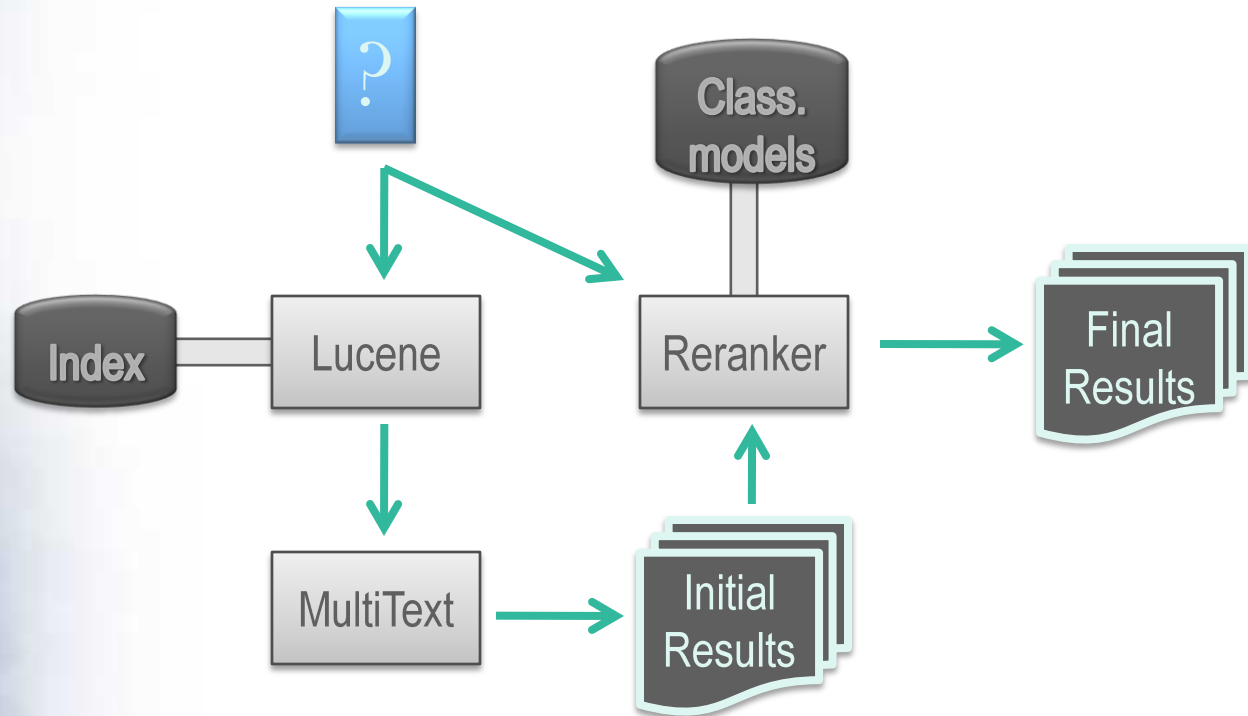
The background features a dark blue gradient with several bright, glowing blue light trails. These trails are composed of multiple parallel lines that curve and intersect, creating a sense of motion and depth. The trails are most prominent on the left side and extend across the width of the image.

Deliverable #4:

The Whole Enchilada

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
- Added Porter stemmer (PorterStemFilter class)
- Removed duplicate words from query
- Experimented with different stopword lists
 - Range in size from 33-600
 - Settled on NLTK's 127-word list

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
 - Only look in top 100 documents
 - Stopping performed by Lucene
 - Law of unintended consequences

Reranking

- Classify questions and answers
- Compare classifications to determine relevance
 - Cosine similarity
- Scale MultiText metric accordingly
- Rerank all 100 passages, return top 20

Query Classification

- Unigrams
- POS tags
- WordNet senses list

- Accuracy on TREC 10: 88.0%
- Accuracy on TREC-2005: 83.5%

Answer Classification

- Automatically constructed training data
 - Labels: query classifications
 - Instances: correct answers from corpus
 - Features
 - Unigrams
 - WordNet senses
- Accuracy

	100	250	1000
Coarse	59.4%	60.8%	58.4%
Fine	43.6%	43.1%	39.3%

Results

- Document Retrieval
 - Stopping lowers MAP (.301 to 0.291) but increases precision (P_5 : 0.176 to 0.185)
- Passage Retrieval (Lenient MRR)

	100	250	1000
Deliverable 3	0.111	0.201	0.380
Stopping	0.072	0.142	0.296
Reranking	0.081	0.142	0.297

2004 reranking boost: 0.065 to 0.083

Lessons, questions, future directions

- Different tasks require different techniques
 - Document and passage retrieval need not share query processing methods
- Why did reranking only help on short passages?
- Improved answer classification
 - NER
 - POS?
- Dependency parsing for reranking