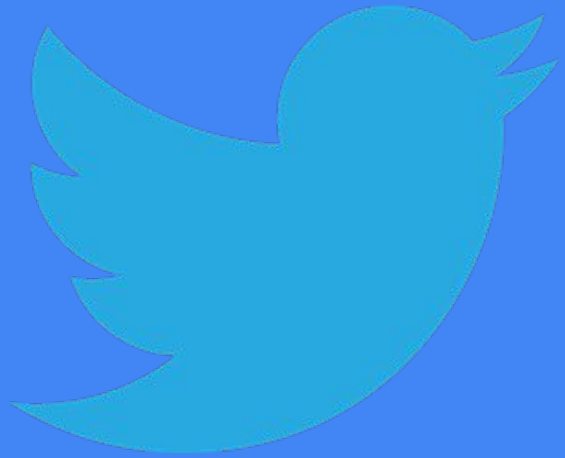
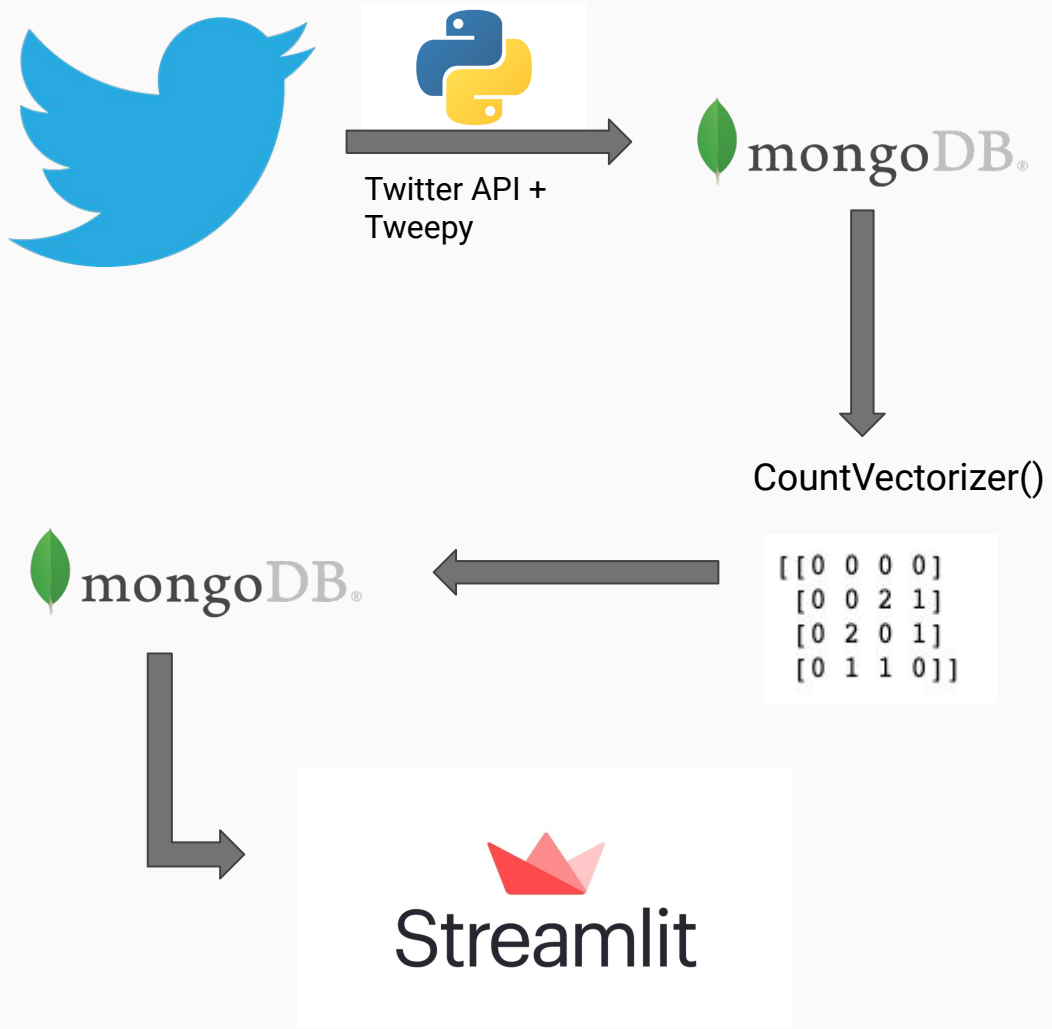
Twitter Keyword Search

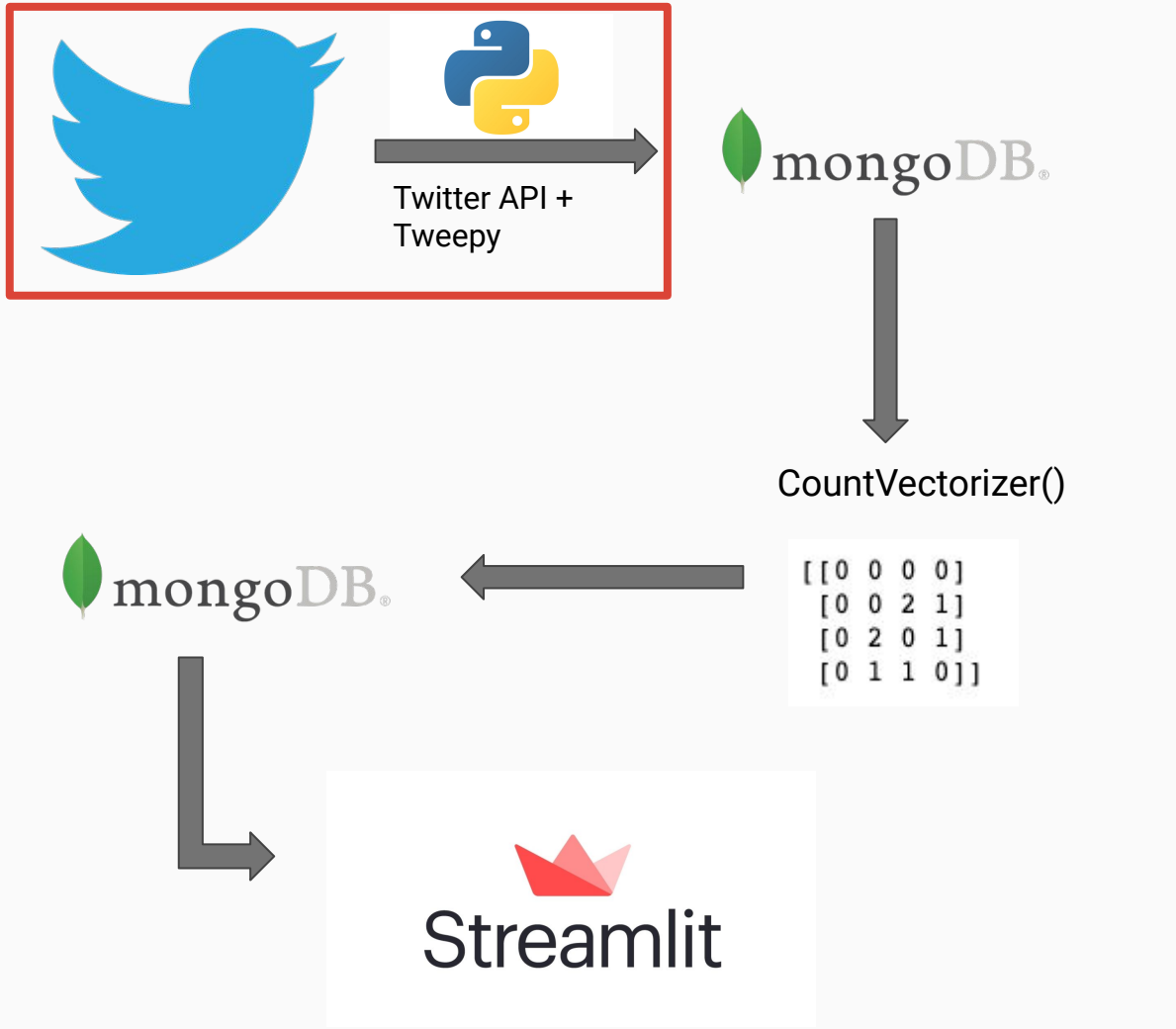


Pipeline Roadmap

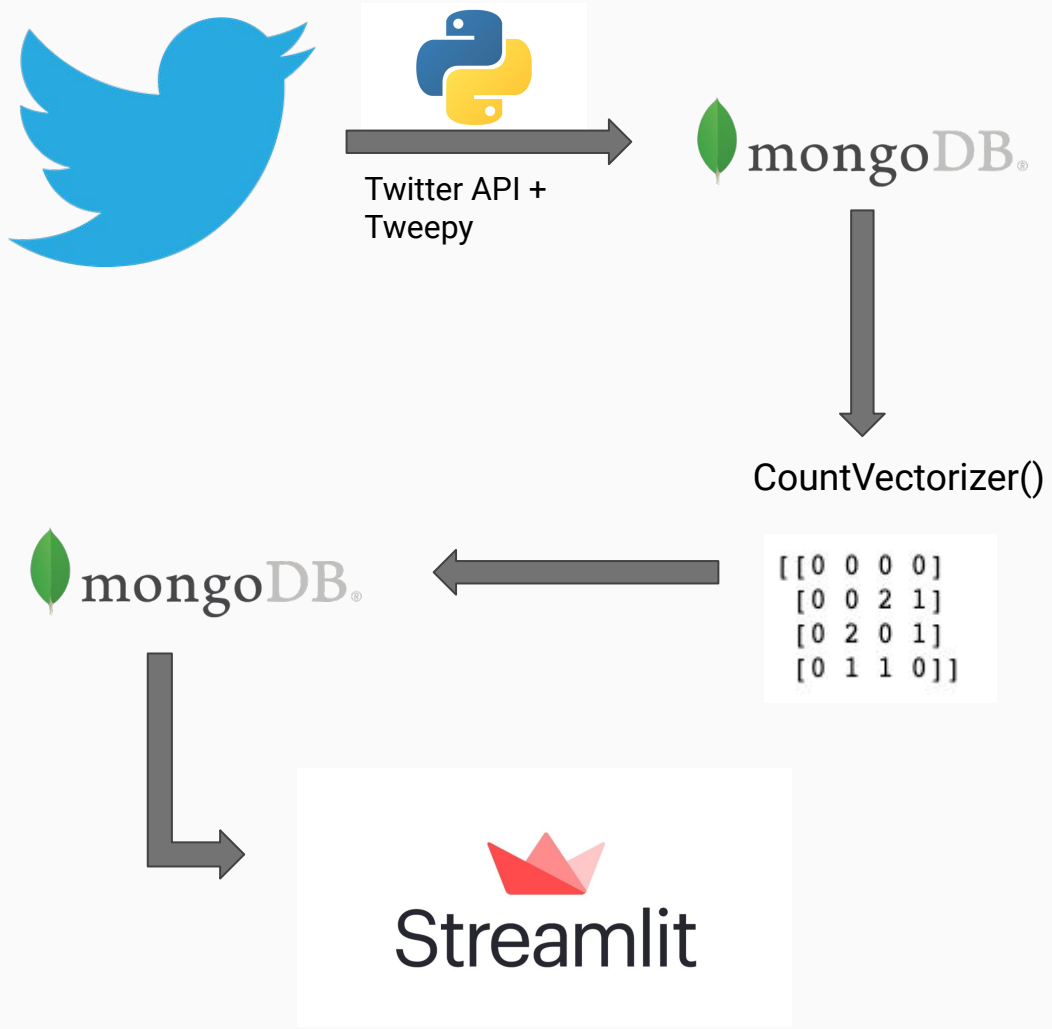


Twitter API + Tweepy

- Searched on tags: a, the, I, you, u
  - This is to avoid biasing the scraped tweets.
- Also filtered down to English

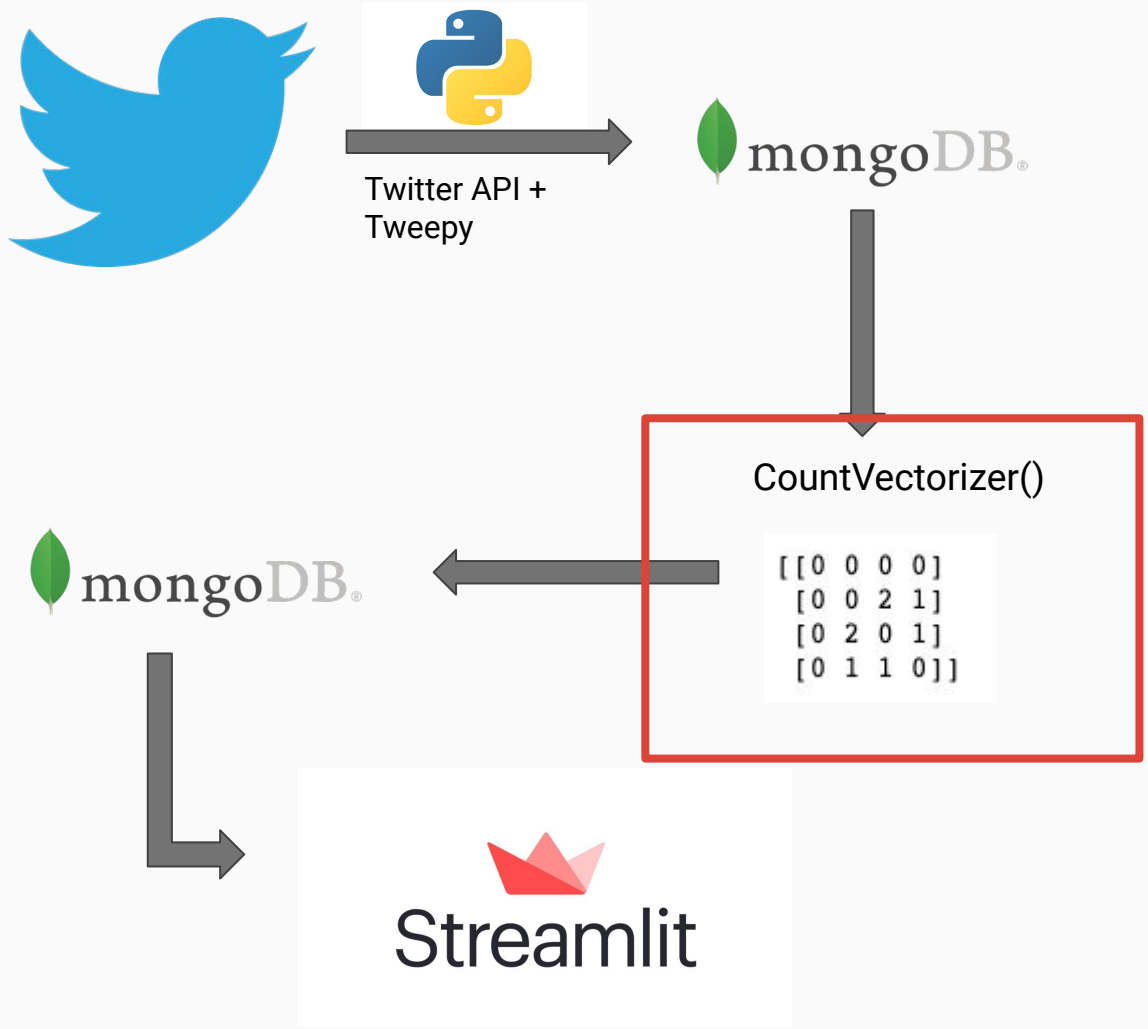


Pipeline Roadmap



CountVectorizer()

- Counts ‘intra-documental’ relations and stores them cumulatively in a matrix
- Better to vectorize in smaller chunks due to the sparsity of the resultant matrix
    - A lot of words are useless and just add to the run-time
- Used MongoDB to contain the equivalent of the matrix



CountVectorizer()

Assume:

docs = [‘this this this book’,
‘this cat good’,
‘cat good stuff’]

No stop words vocabulary:
[book, cat, good, stuff, this]

```
[0, 0, 0, 0, 3]
[0, 0, 2, 1, 1]
[0, 2, 0, 1, 1]
[0, 1, 1, 0, 0]
[3, 1, 1, 0, 0]
```

```
{'this': {'this': 0, 'book': 3, 'cat': 1, 'good': 1, 'stuff': 0},
 'book': {'this': 3, 'book': 0, 'cat': 0, 'good': 0, 'stuff': 0},
 'cat': {'this': 1, 'book': 0, 'cat': 0, 'good': 2, 'stuff': 1},
 'good': {'this': 1, 'book': 0, 'cat': 2, 'good': 0, 'stuff': 1},
 'stuff': {'this': 0, 'book': 0, 'cat': 1, 'good': 1, 'stuff': 0}}
```

Indices:
{'book': 0, 'cat': 1, 'good': 2, 'stuff': 3, ‘this’:4}

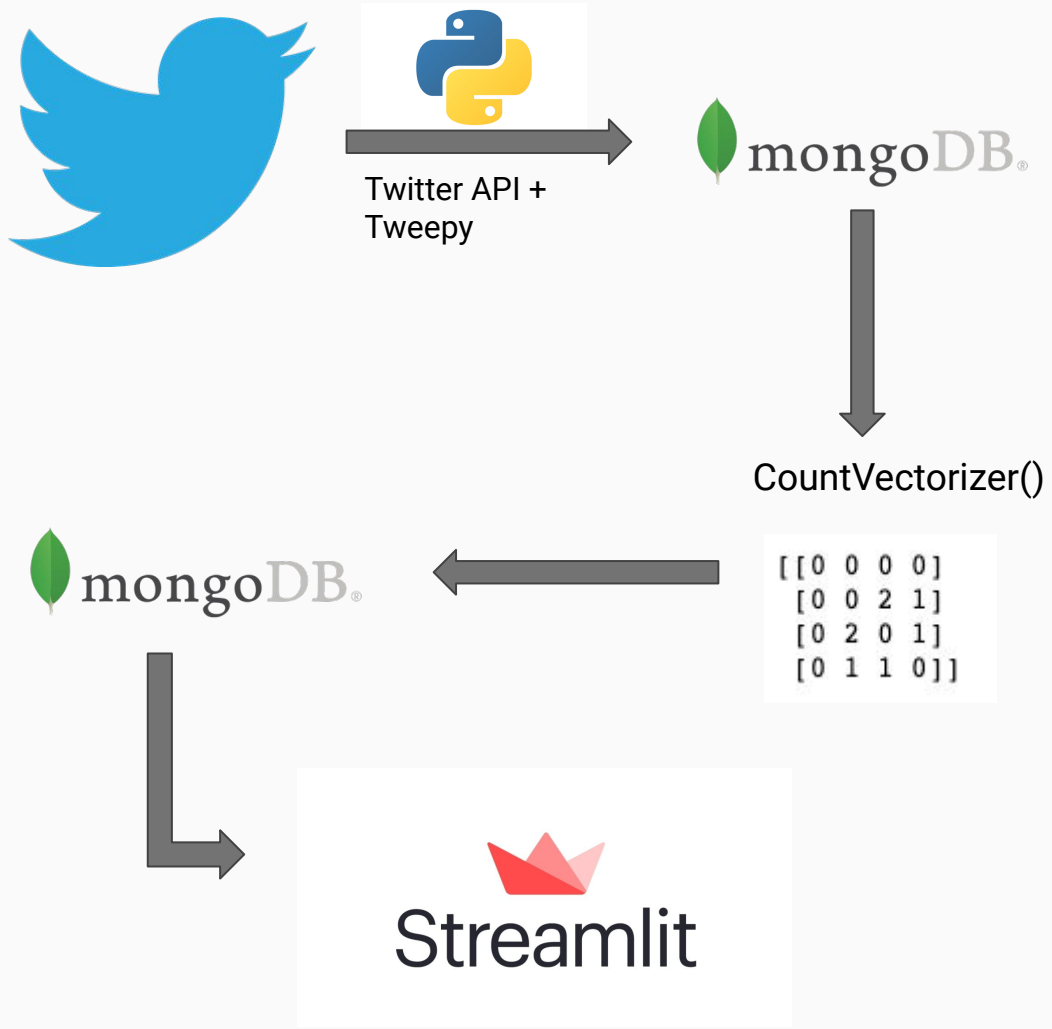
With stop words vocabulary:
[book, cat, good, stuff]

```
[[0 0 0 0]
 [0 0 2 1]
 [0 2 0 1]
 [0 1 1 0]]
```

```
{'book': {'book': 0, 'cat': 0, 'good': 0, 'stuff': 0},
 'cat': {'book': 0, 'cat': 0, 'good': 2, 'stuff': 1},
 'good': {'book': 0, 'cat': 2, 'good': 0, 'stuff': 1},
 'stuff': {'book': 0, 'cat': 1, 'good': 1, 'stuff': 0}}
```

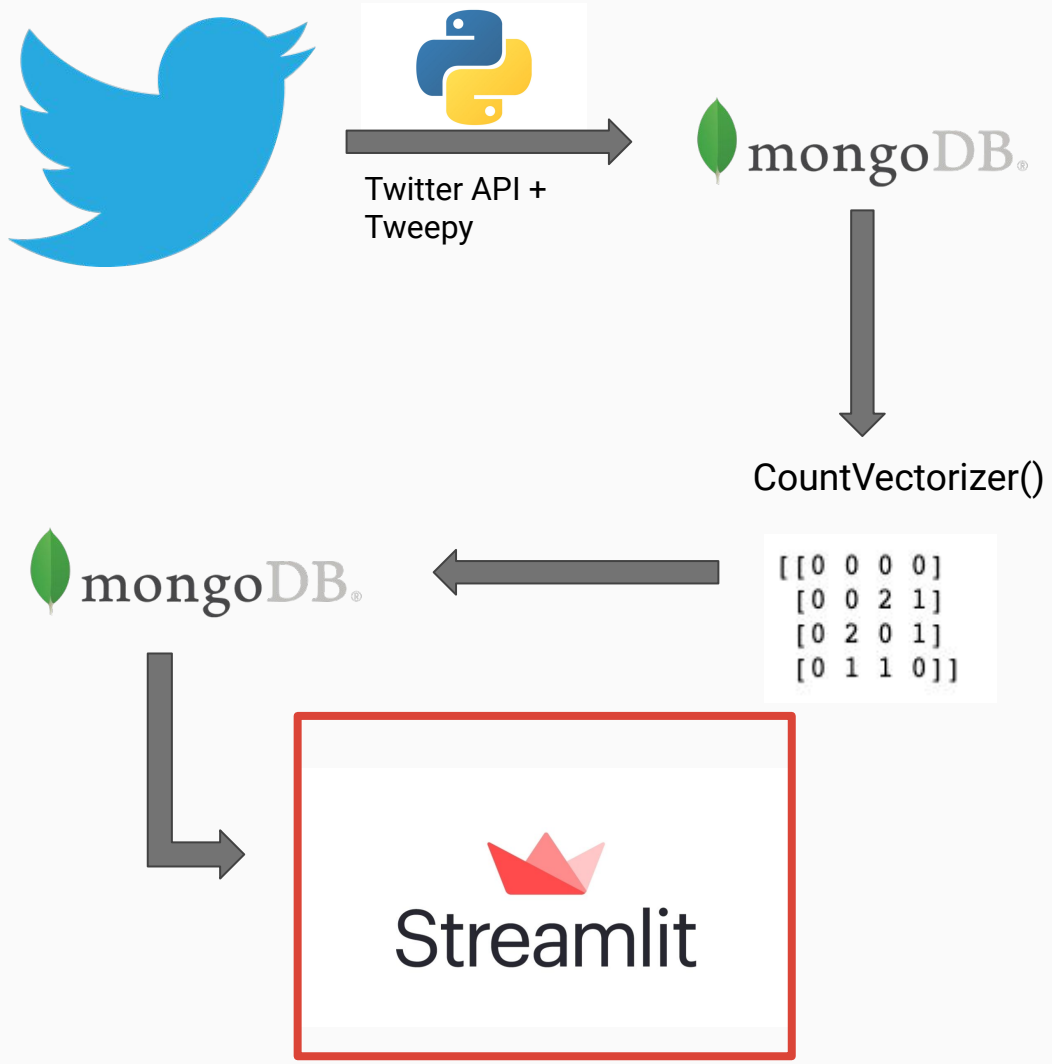
Indices:
{'book': 0, 'cat': 1, 'good': 2, 'stuff': 3}

Pipeline Roadmap



Querying through Streamlit

- Vectorized strings are transformed into Mongo-readable format and stored.
- Front-end Streamlit app queries directly from the database to retrieve keywords for an inputted search term.



Video demonstration

Some Takeaways:

- Pipeline construction matters!
  - CountVectorizer() runtime vs Storage runtime
- Not everything tabular should be stored tabularly
- Twitter is a scary place.

Questions?