# Technologies for Visualization of Big Medical Text Data

Leilani Battle[2] Lauren Edwards[1] Vijay Gadepally[1,2] Brendan Gavin[1,5] Braden Hancock[1,6] Dylan Hutchison[1,2,3] Jeremy Kepner[1,2,4] Andrew Moran[1,2]
[1]MIT Lincoln Laboratory [2]MIT Computer Science & AI Lab [3]Univ. of Washington [4]MIT Math Department [5]Univ. of Massachusetts [6]Stanford University

## ABSTRACT

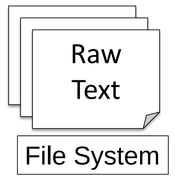
The ISTC Big Data Analytics Working Group (BigDAWG) project allows users to explore and analyze heterogeneous data sets that vary in size, schema, and organization. One use case for BigDawg is the MIMIC II public data set, which contains multiple years of ICU reports for thousands of patients at Boston's Beth Israel Deaconess Hospital.

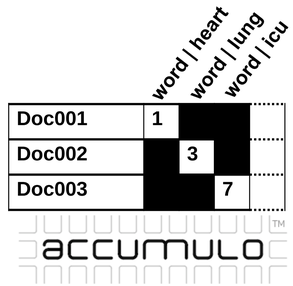
In this poster, we highlight the technologies that enable topic modeling over the corpus of MIMIC documents: D4M, Graphulo, and Vega. The result is a prototype system that visualizes the content of a given medical report using the BigDAWG API.

## SYSTEM PROTOTYPE

The diagram to the right depicts the steps for processing data into query-ready form, and then steps to answer a visualization query from a client through the BigDAWG web service. We show a sample of input data, the technology used for processing, and a sample of resulting data for each step. Data originates from raw text files, Accumulo stores intermediate and processed data, and a client web browser presents a query and visualization interface.

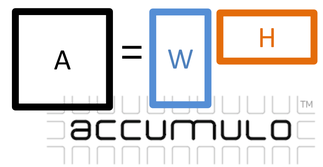
All operations except for the Non-negative matrix factorization (NMF) step do not require holding an entire dataset in the memory of a single node. D4M's ingest step leverages pMATLAB to ingest triples into Accumulo from multiple processes; Graphulo runs within Accumulo's tablet servers, gaining parallelism and minimizing data movement; and Vega visualizes only the topics relevant to the client.

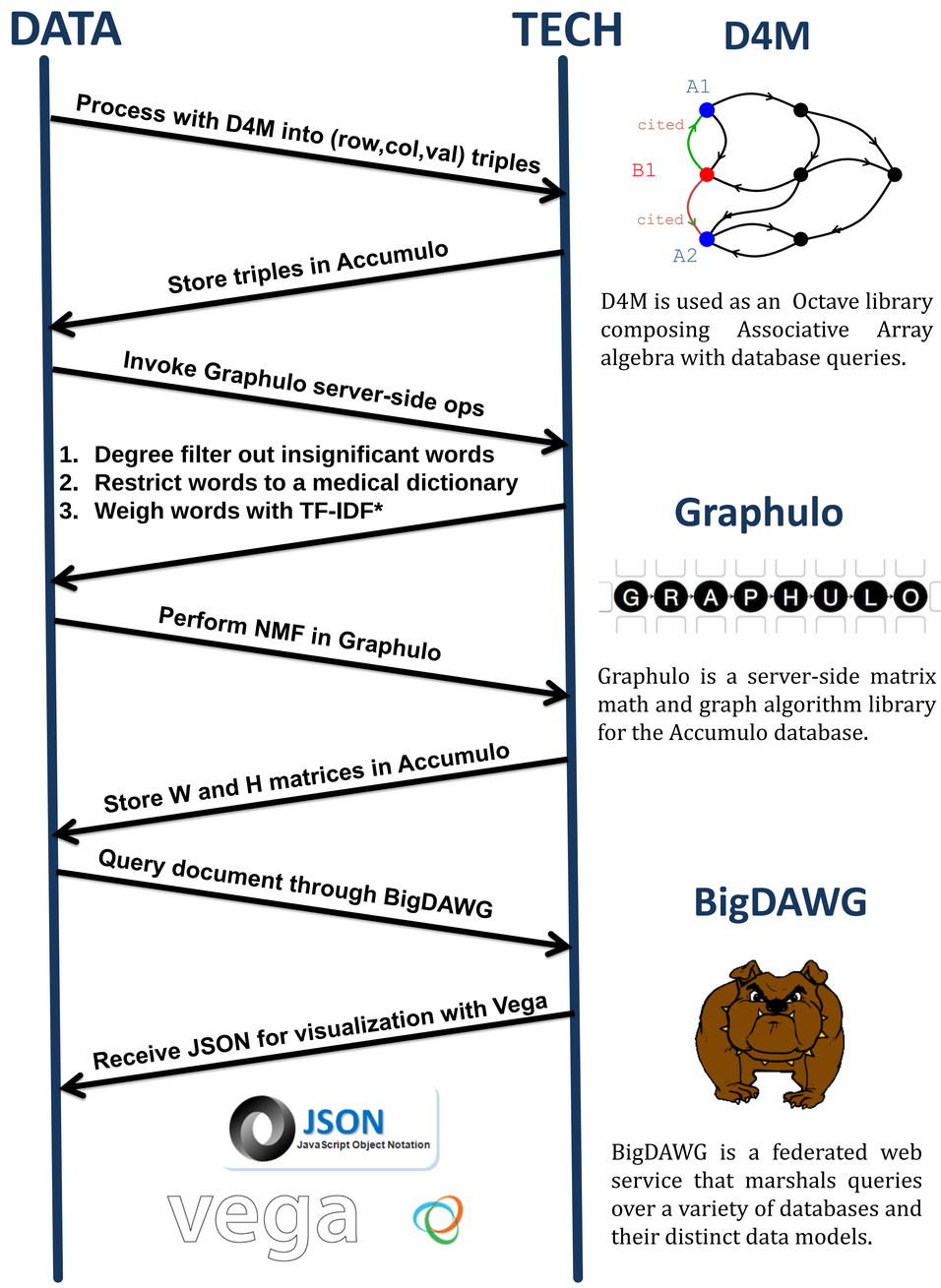



```
('Doc001','word|heart','0.334')
('Doc002','word|lung' ,'0.167')
('Doc003','word|icu'  ,'0.019')
```

accumulo




*Term Frequency–Inverse Document Frequency

### D4M




D4M is used as an Octave library composing Associative Array algebra with database queries.

### Graphulo

GRAPHULO

Graphulo is a server-side matrix math and graph algorithm library for the Accumulo database.

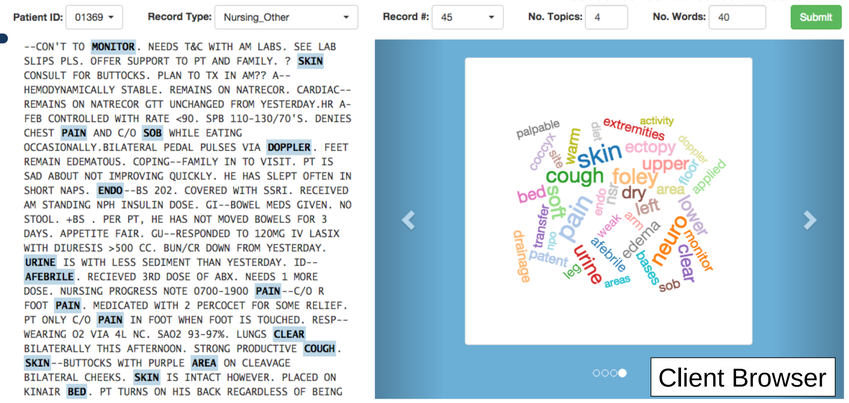
### BigDAWG



BigDAWG is a federated web service that marshals queries over a variety of databases and their distinct data models.

## FUTURE WORK

- Use original ICU data directly from PostgreSQL by creating a BigDAWG "shim," instead of exporting from PostgreSQL to text files for pre-processing and loading into Accumulo.
- Accelerate an NMF variant that does not require holding data in memory, which enables distributed execution in an Accumulo instance via Graphulo.
- Explore alternative visualization tools for processed medical data, as well as techniques that consider word position beyond bag-of-words approaches.







intel MIT BROWN UNIVERSITY OF CALIFORNIA SANTA BARBARA Carnegie Mellon University W UT Portland State UNIVERSITY ISTC BIG DATA