Text Stream Trend Analysis using Multiscale Visual Analytics with Applications to Social Media Systems

Chad A. Steed (csteed@acm.org), Justin Beaver,
Paul L. Bogen II, Margaret Drouhard, and Joshua Pyle

ACM IUI Workshop on Visual Text Analytics
March 29th, 2015

This research is sponsored by Oak Ridge National Laboratory (ORNL)
Laboratory Directed Research and Development (LDRD) no. 6427.

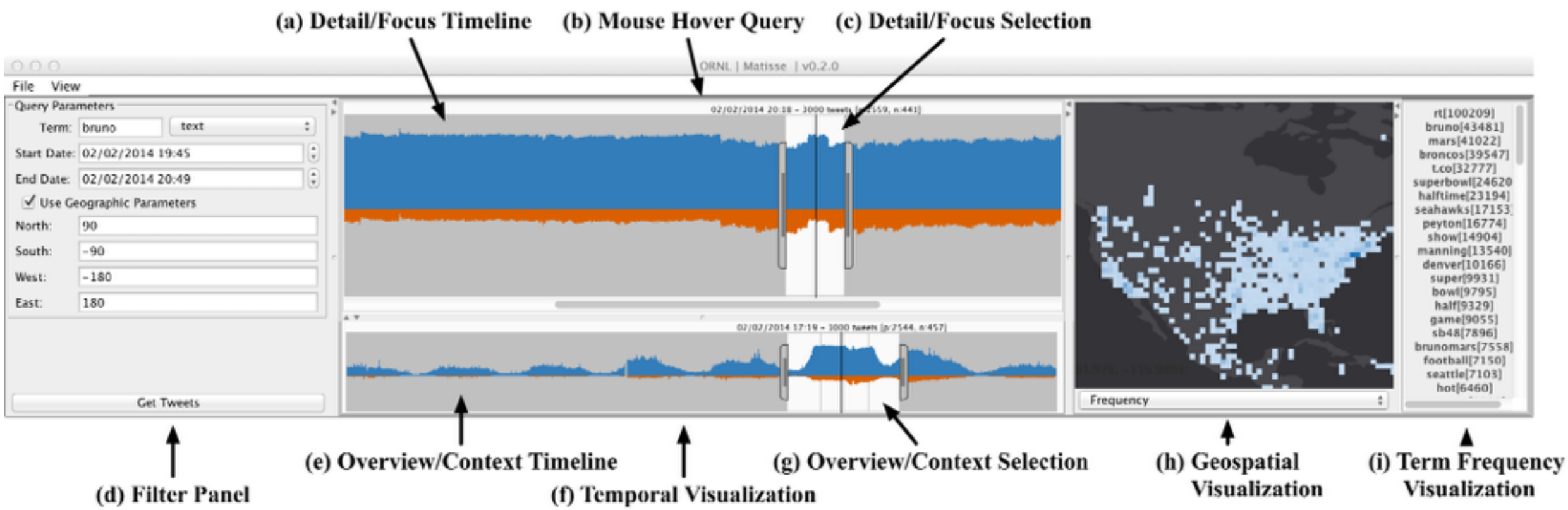
OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY | Office of Science

Motivation

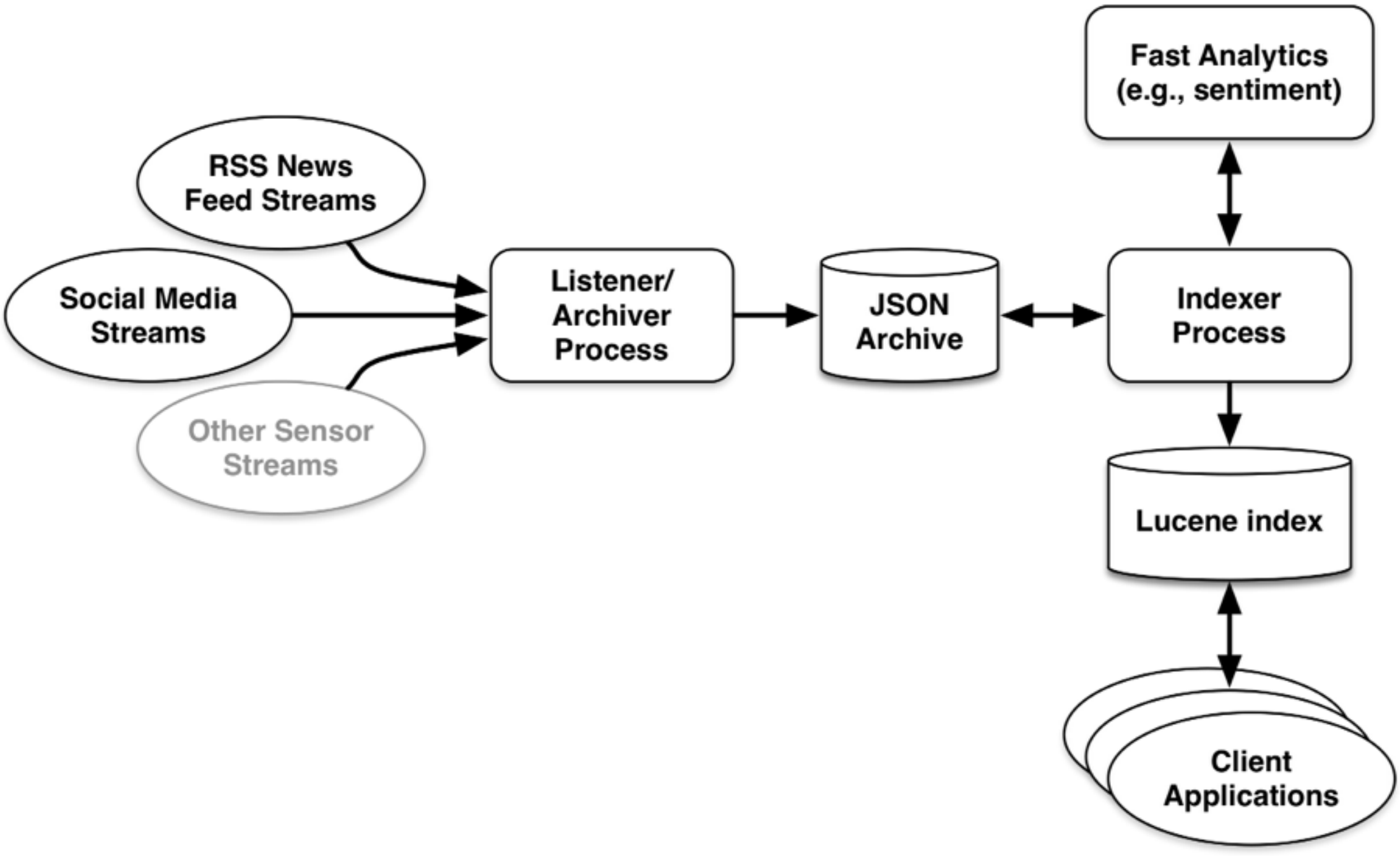
- Text streams are ubiquitous and represent a tremendous resource for understanding global events in real-time.
    - Global event situational awareness for disaster response
- Such streams are a challenge to explore
    - High velocity, semi-structured textual information.
- Its like detective work that can be improved with:
    - Automated analytics to guide analyst to key trends
    - Interaction techniques to drill-down to access increasing detail views.

Matisse: Exploratory Analysis of Text Streams



<u>Design Goal:</u> Enable multi-faceted, exploratory analysis of emotion in social media text streams from overviews to detailed investigation.

Stream Data Management



Positive / Negative Sentiment Analytics

- Estimate positive / negative sentiment based on textual content
- Initial processing modifies raw text to accommodate nuances of Twitter content
- Porter's English stemmer applied and custom stop words are removed
- Train classifier (naive Bayes and Java Maximum Entropy) using pre-coded tweets from Go et al. [4] to create a feature vector.
- Comparable classification accuracy to Go et al. [4] (80% - 90%)

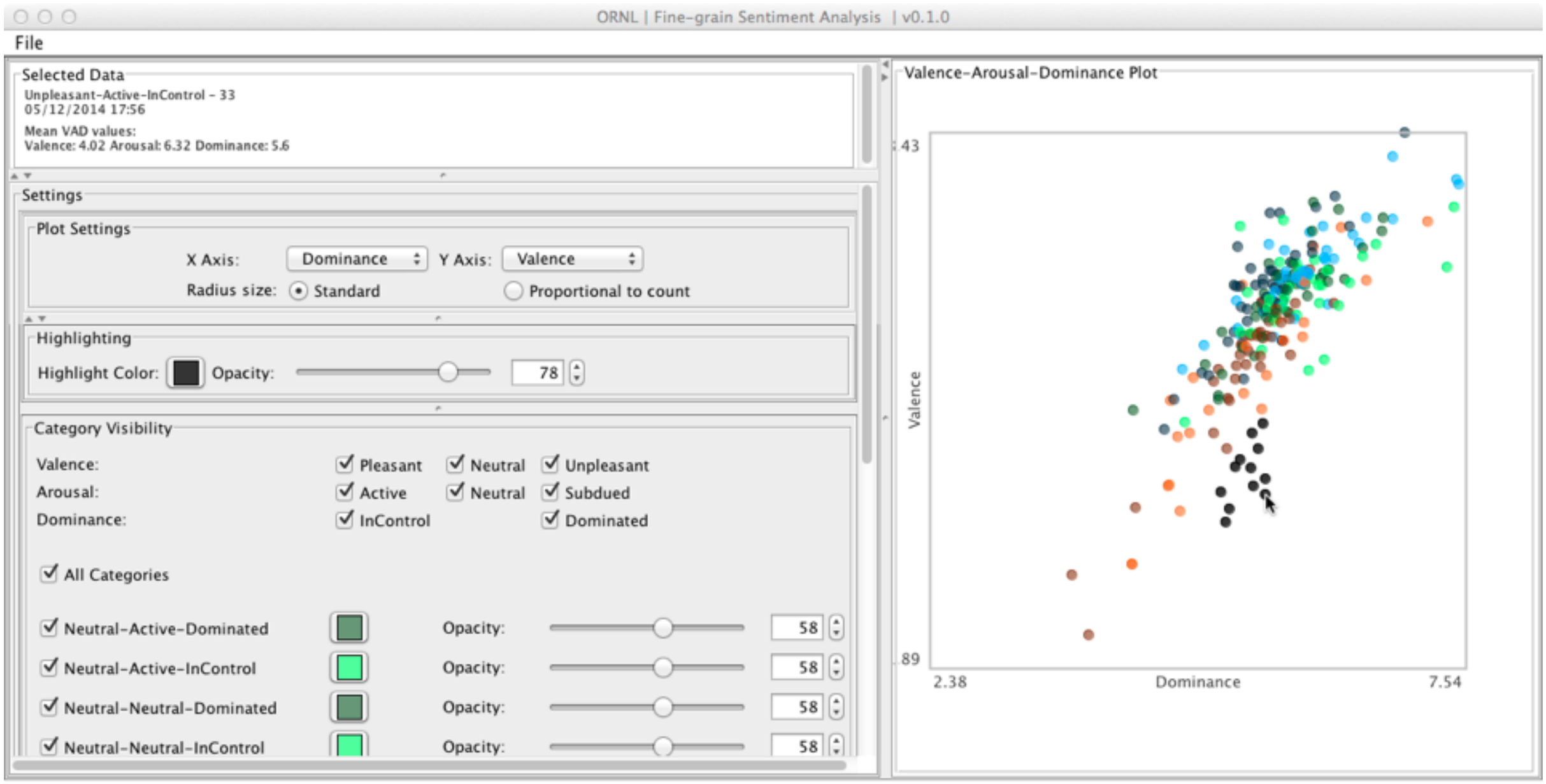
Emotion Classification using Machine Learning

- Estimate textual content's emotion using a ML approach that avoids manual labeling of training data
- Leverage statistical and emotional text analytics trained on pure examples of various emotion classes
- Using ANEW [2] model
  - Valence - positivity/negativity
  - Arousal - excitability
  - Dominance - assertion level of the author

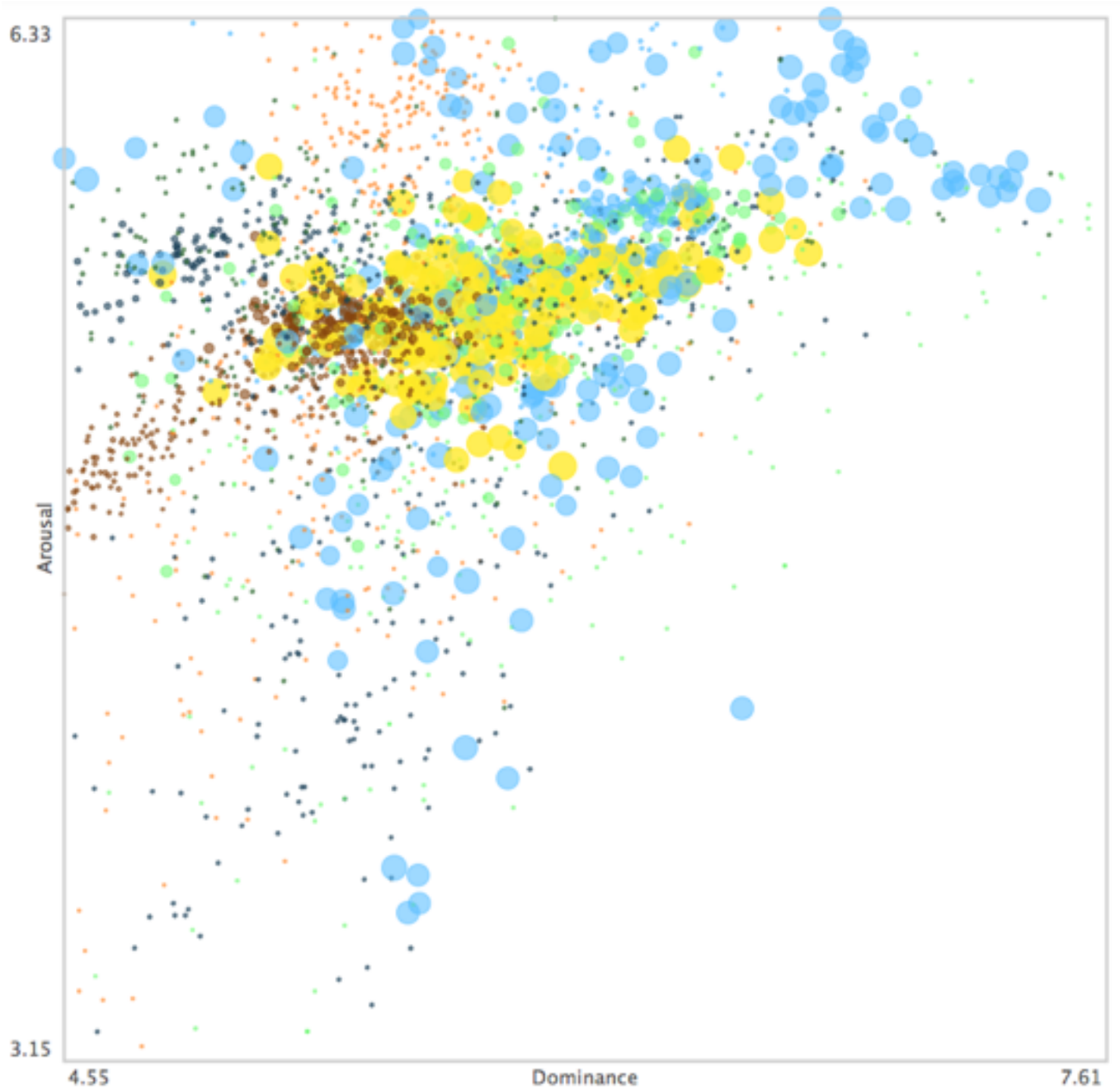
Emotion Classification using Machine Learning

- TF-ICF [6] term weighting method determines significant terms using a vector space model
- Significant terms and ANEW scores are merged as a feature set representative of the content and emotion of the textual record
- A maximum entropy learner from the MinorThird ML library is used to train the model
- Training selects tweets with emotion class explicitly encoded as a hashtag to provide pure representations of each emotion class (automated labeling)
- Classifier is built to predict emotion in new unlabeled records

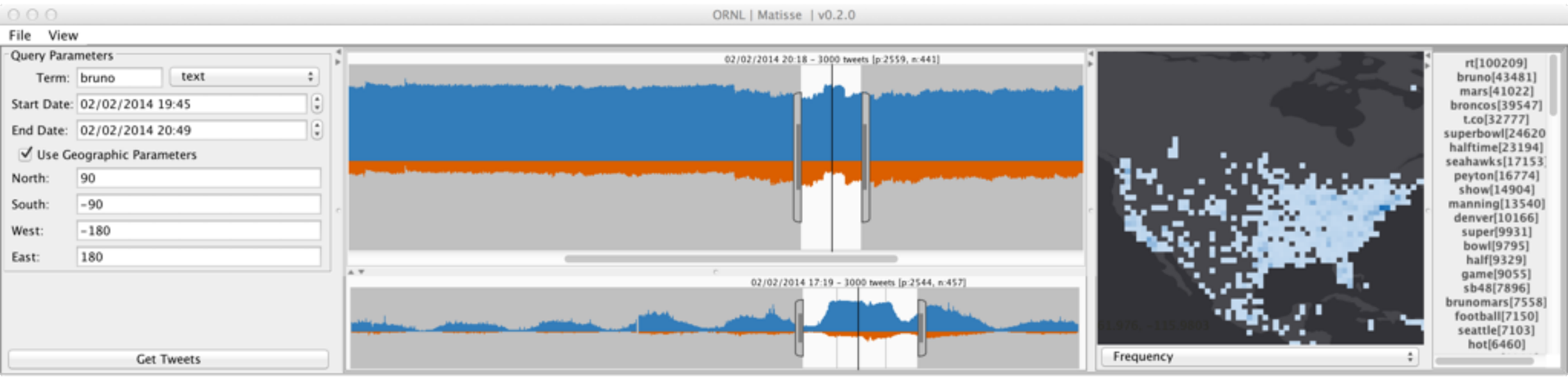
Visualizing Emotion Classifications



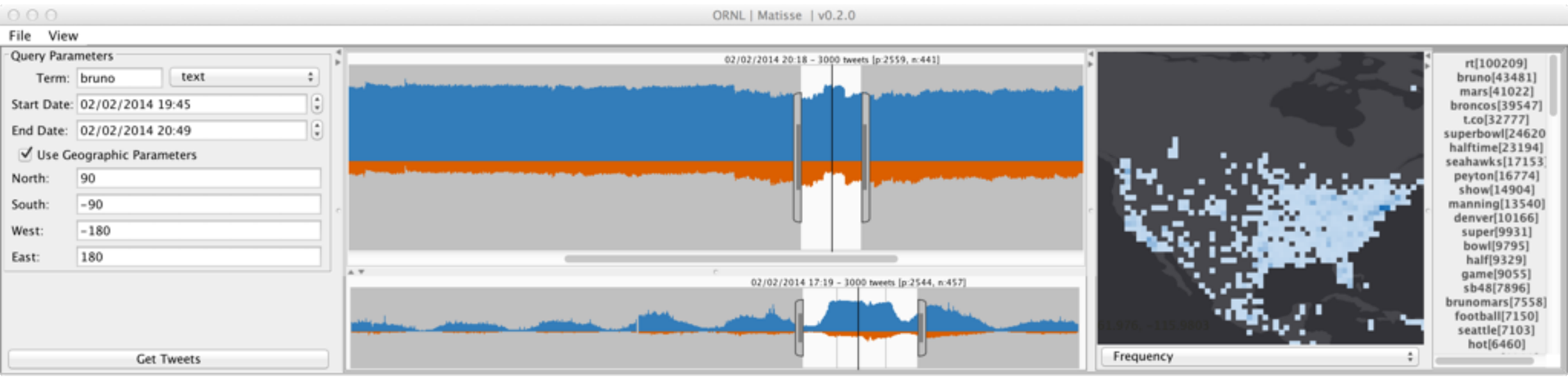
Visualizing Emotion Classifications



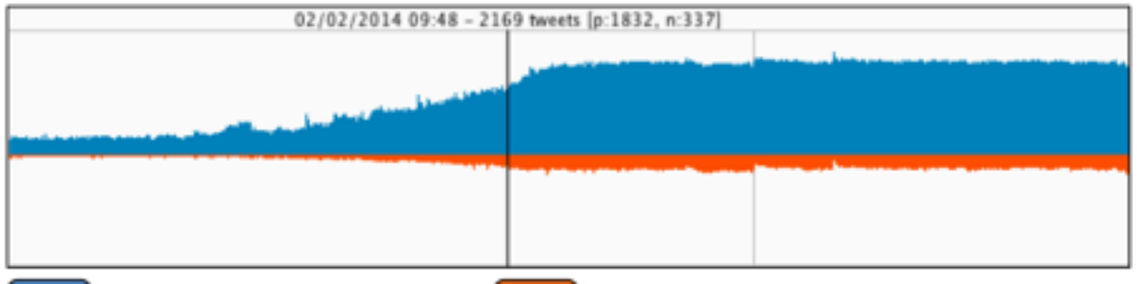
Geospatial + Term Frequency Visualizations



Multi-scale Temporal Visualizations

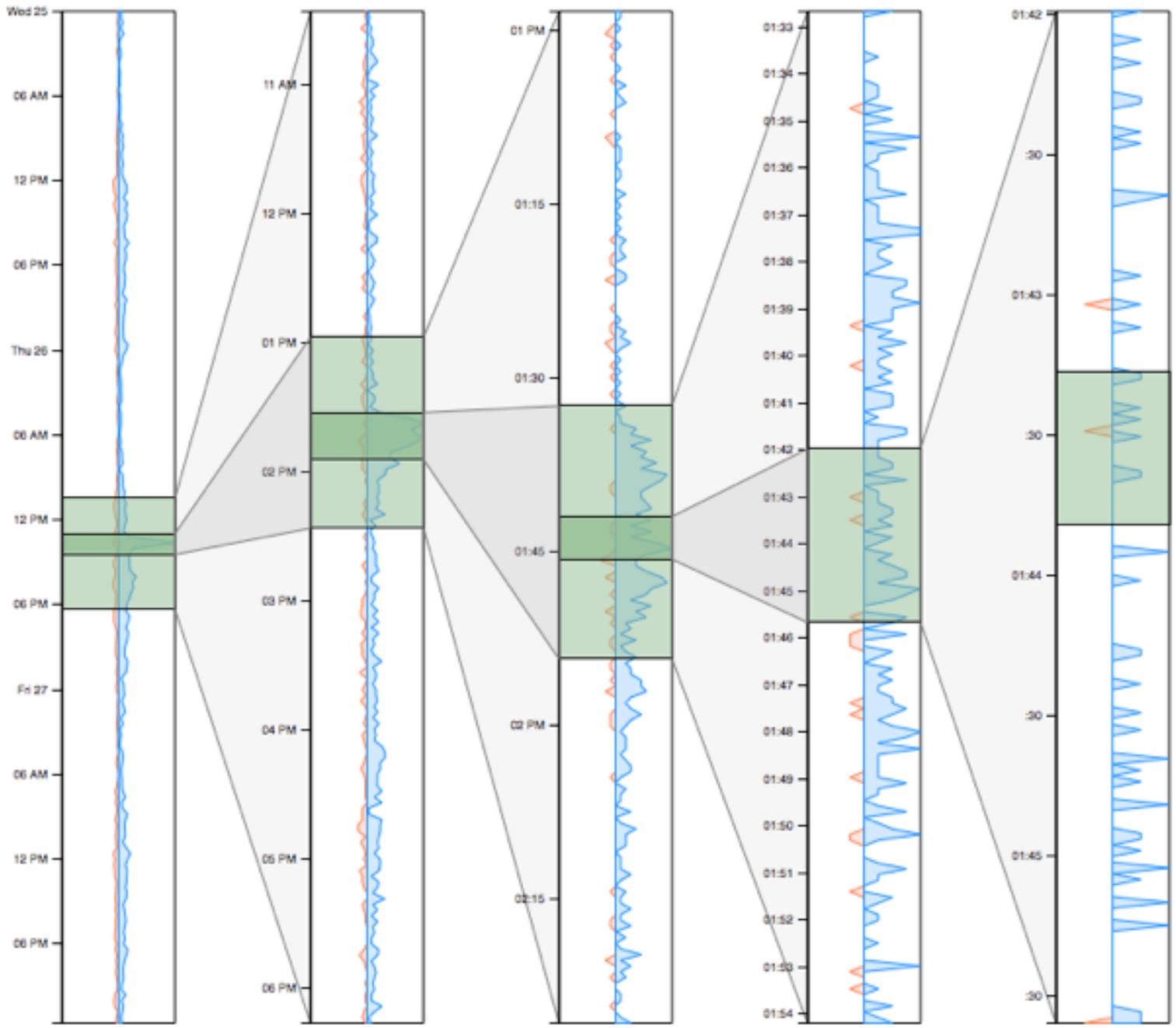


Total Frequency

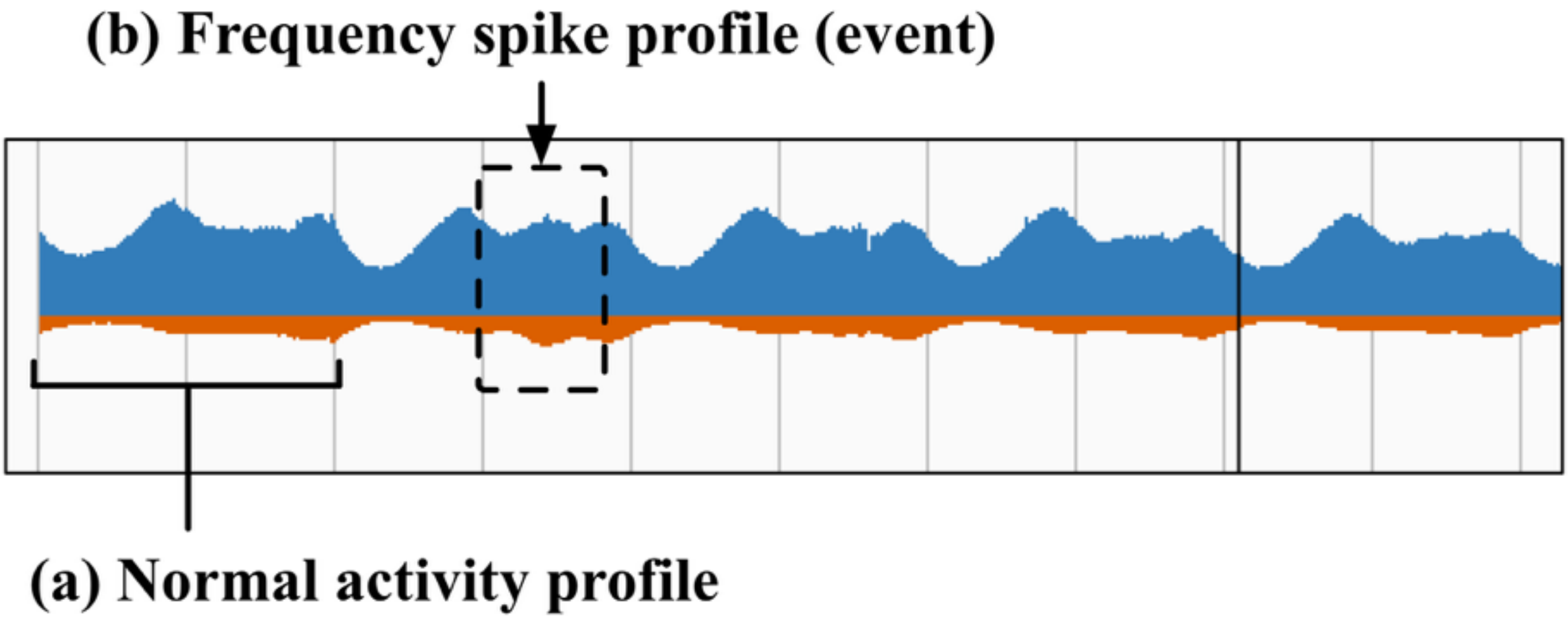


Positive Sentiment **Negative Sentiment**

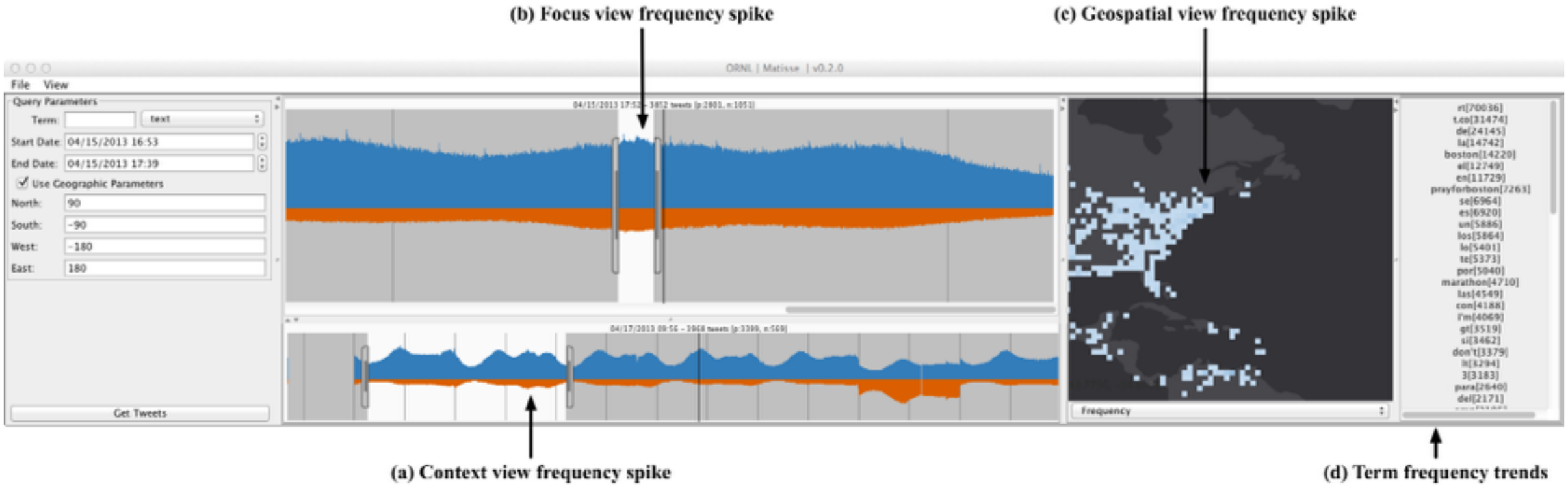
Multi-scale Temporal Visualizations



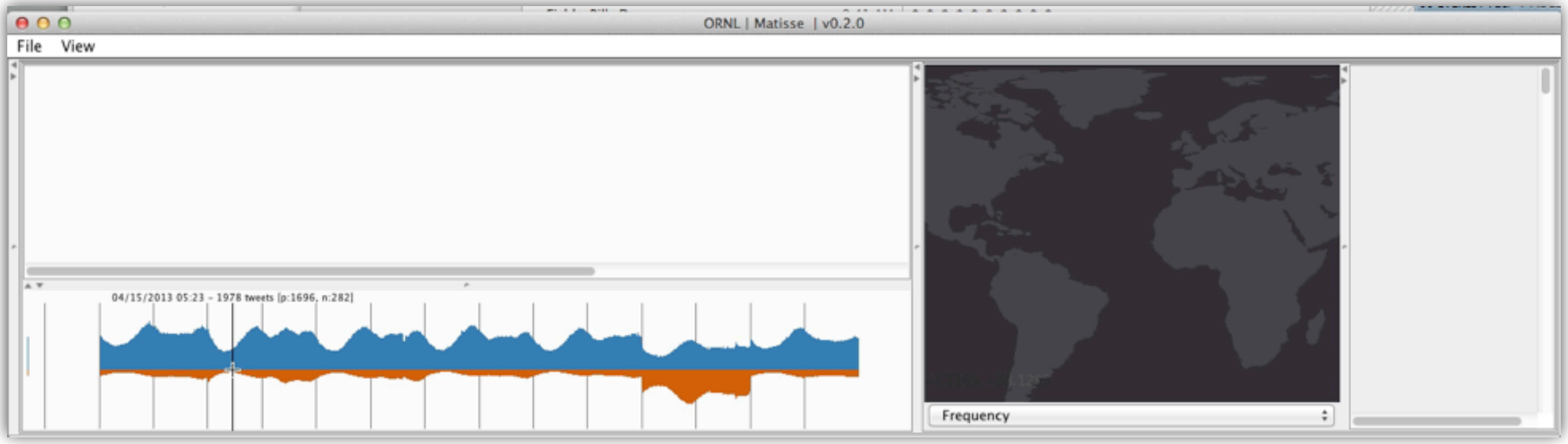
Temporal Event Detection



Case Study: Boston Marathon Bombing



Case Study: Boston Marathon Bombing



Twitter 1% Sample Stream
Week of Boston Marathon Bombing
14-20 April, 2013

Conclusions

- Multi-scale visualizations enable scalable exploratory analysis.
  - More intermediate views and scaled analytics are needed.
- Automated analytics (sentiment / emotion) help guide analysis.
  - ML emotion classification needs validation and expansion.
- Linked views and interactions foster more creative analysis.
  - Additional views and interactions are needed.