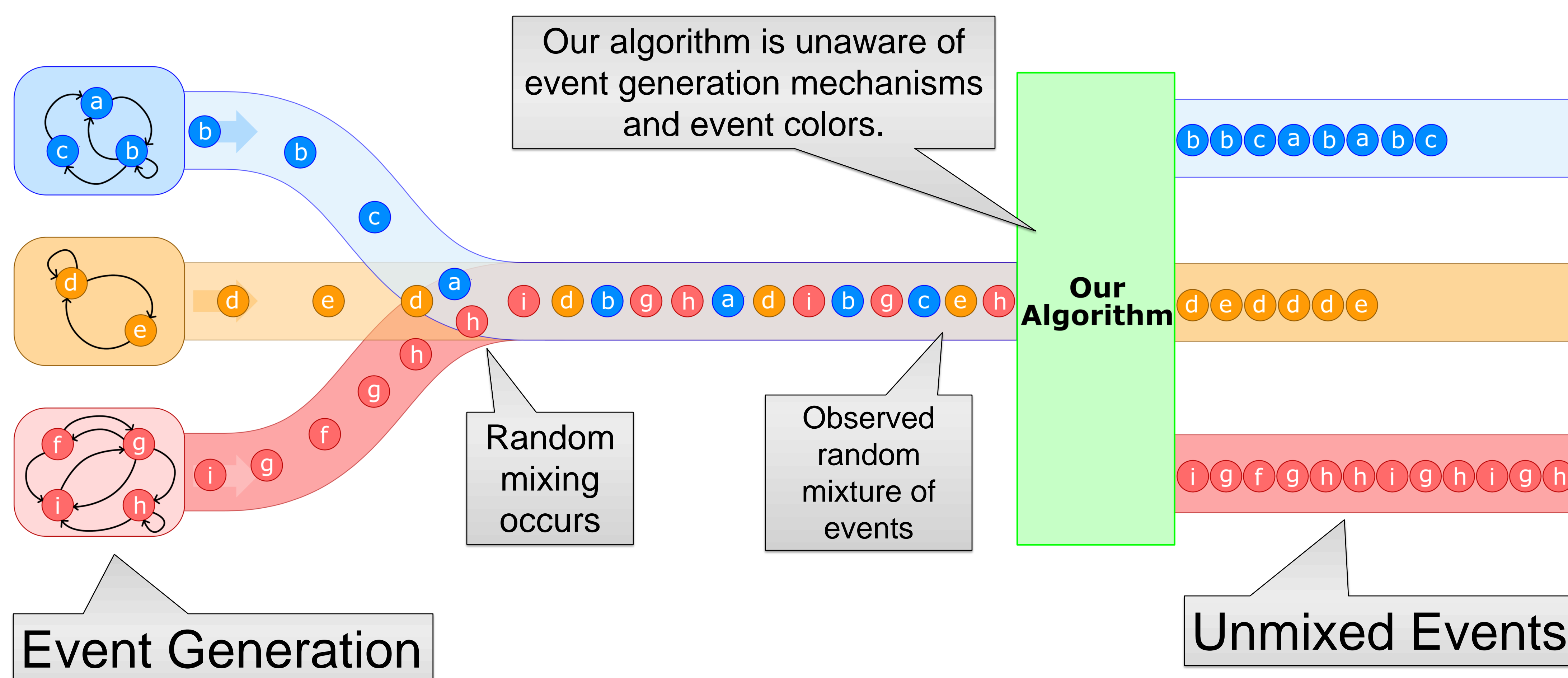


Data Reduction Techniques for Event Sequences

Mikhail Atallah, Emil Stefanov, Wojciech Szpankowski



Overview

1. Models generate events independently.
2. The events generated by each model are randomly mixed.
3. The algorithm observes the mixture of events but does not know which and how many events came from each model.
4. By measuring probabilities, the algorithm groups the events by the model that generated them.

Goal

Partitioning of an event sequence into the subsequences generated by the different models (without knowing the models or the mixing process).

Allows for **better analysis**:

- Eliminates interference between models (decoupled).
- Allows for more sophisticated analysis algorithms on the shorter subsequences.
- Massive filtering in case the events of interest are generated by a single model.

Applications

- Automatic classification of activities of disguised users or web bots.
- Automatic separation of randomly intermixed text from different (and unknown) natural languages.

Benefits

- **Extremely fast**: Linear in time with respect to the length of the event sequence.
- **Memory efficient**: Uses a small constant amount of memory.
- **Easy to implement**: The code for the algorithm is less than 200 lines.

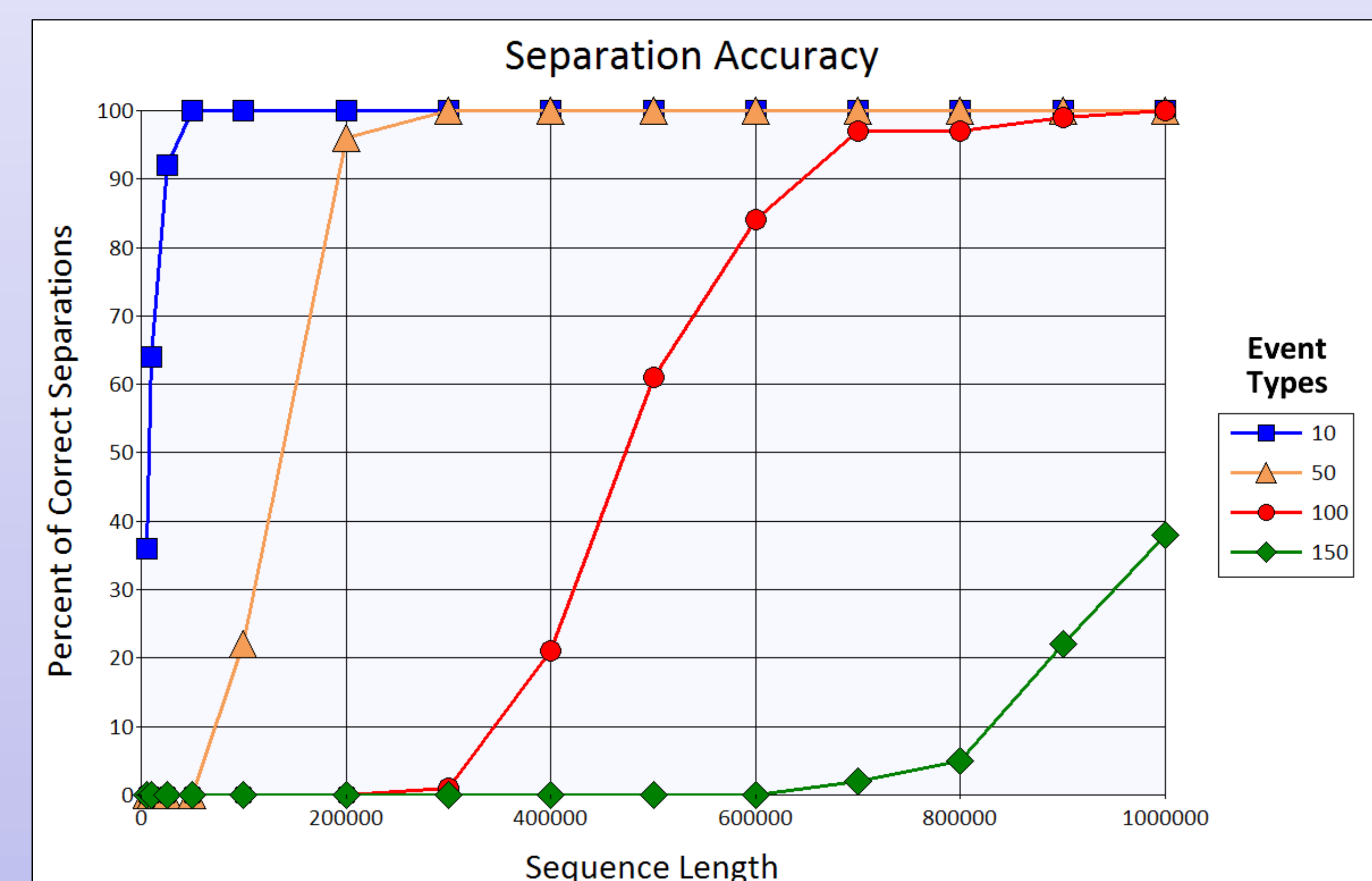
Limitations

- Requires irreversible Markov chain models.
- Is unable to separate Bernoulli models.

Experimental Results

Synthetic Data

- More data improves accuracy.
- Achieves **100% accuracy** for moderate amounts of data.



The graph above shows that, most of the time, the algorithm is either 0% accurate (not enough data) or 100% accurate (enough data).

Real Data

Correctly separates mixtures of English and Spanish **without knowing** anything about either language or how the mixing was done.