Towards Reproducible Research of Event Detection Techniques for Twitter

Andreas Weiler, Harry Schilling, Lukas Kircher, Michael Grossniklaus
June 14, 2019

What is an Event?

1. Papal Election
   - habemus, papam, fumata

2. Boston marathon attack
   - boston, marathon, explosion

Motivation

- Analysis of 48 event detection techniques

1. Implementation issues
   - Approx. 20% provide source code
   - Approx. 20% provide pseudo code

2. Lack of twitter data

3. Evaluation issues
   - Comparative, case study, stand-alone, user study

Approach

1. Implementation Issues
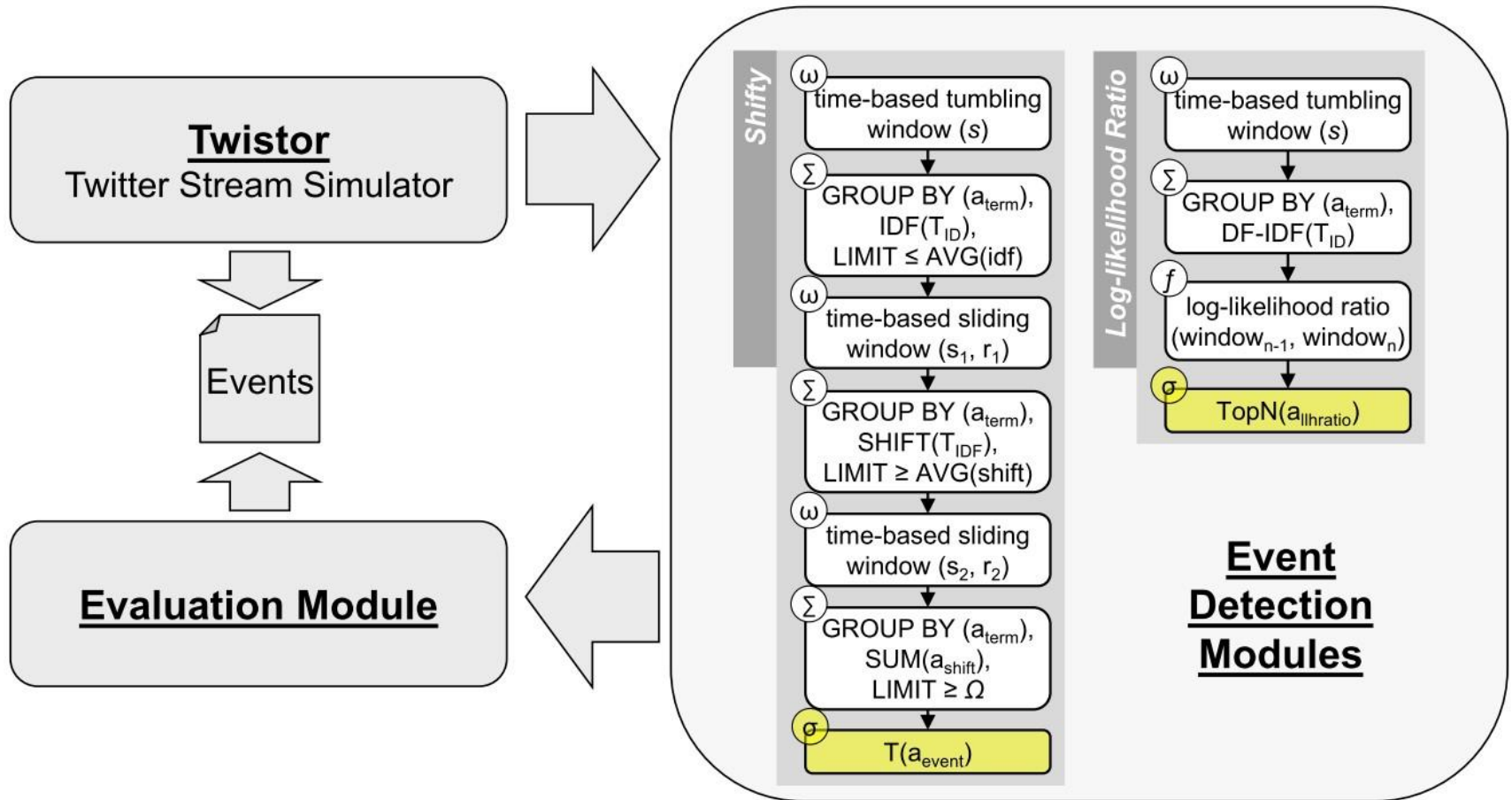   - Event detection modules based on a Data Stream Management System

2. Lack of twitter data
   - **Twi**tter **St**ream Simulat**or**: Twistor

3. Evaluation Issues
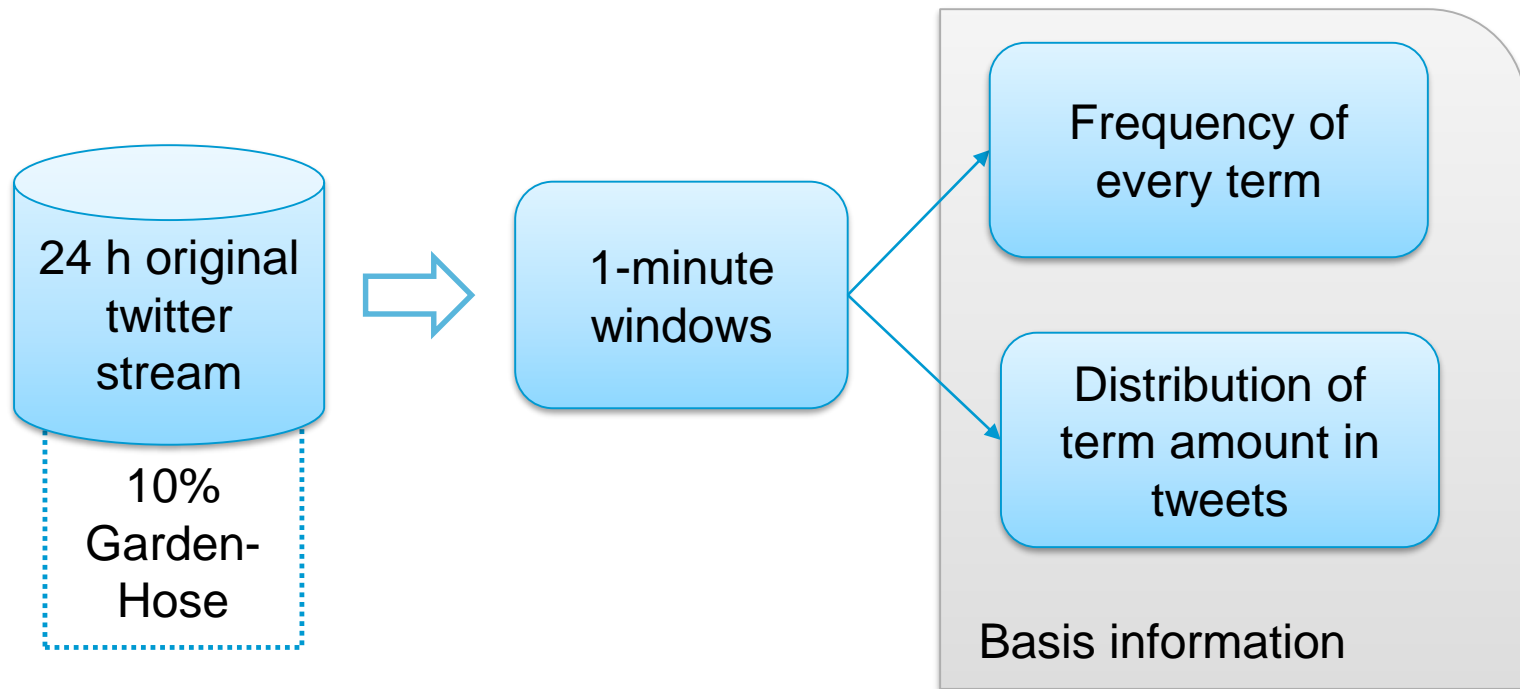   - Evaluation module

Approach

Twistor

1. Simulation of the twitter stream

2. Embedding of events

Twistor

1. Simulation of the twitter stream

```
┌──────────────────┐          ┌──────────────────┐
│ 24 h original    │   ⟹      │    1-minute      │
│ twitter          │          │    windows       │
│ stream           │          └──────────────────┘
└──────────────────┘
  10%
  Garden-
  Hose
```

Frequency of every term

Distribution of term amount in tweets

Basis information

Twistor

1. Simulation of the twitter stream
   - Map term distribution of real twitter stream to simulated one (per 1-minute window)
   - Replace terms of real twitter stream with random terms from the Leipzig Corpora Collection

   - No simulation of
     - Hashtags
     - Users
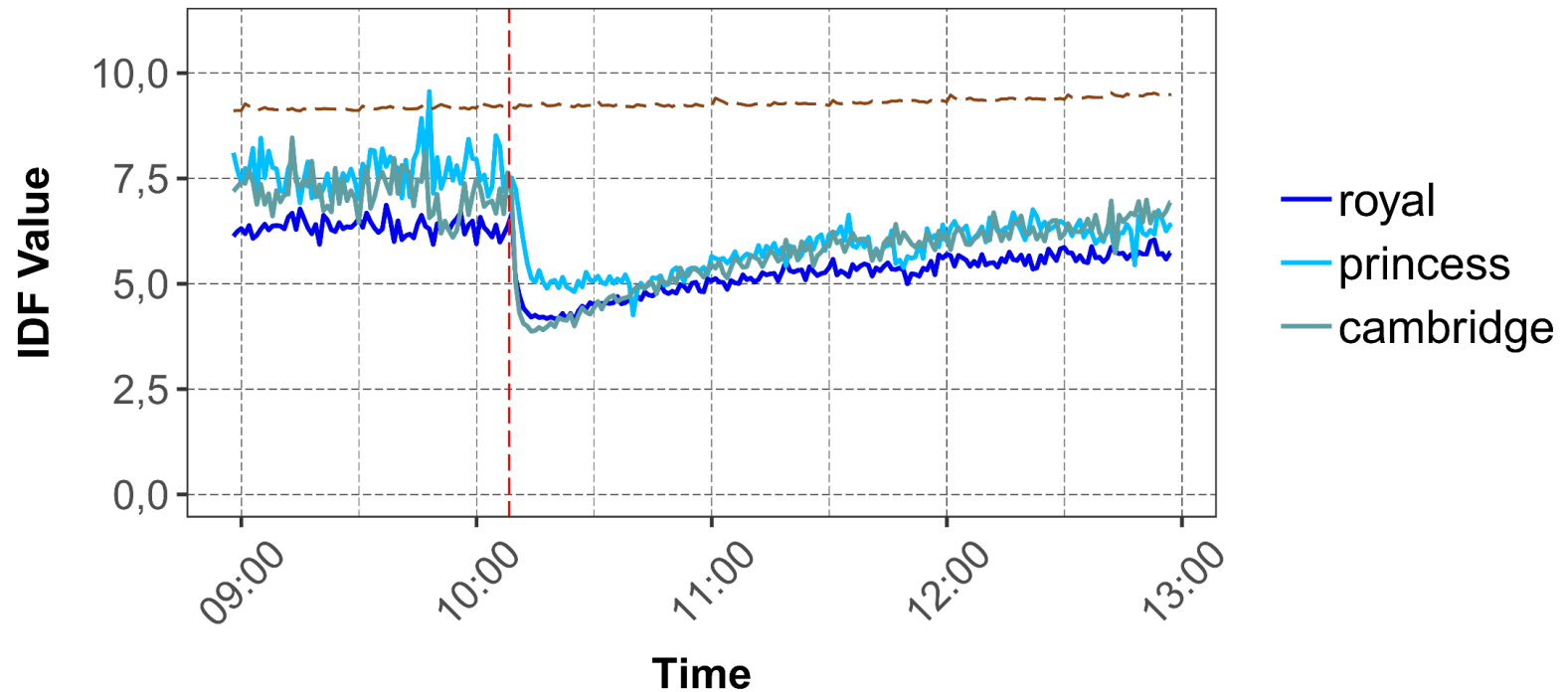   - Semantics
     - …

Twistor

2. Embedding of events
   - Overall 10 events
   - Based on original data
   - Representation of event by IDF values of event terms
   - IDF value of a word w per second
     - $\mathrm{idf}(w) = \log\left(\frac{N}{n_w}\right)$
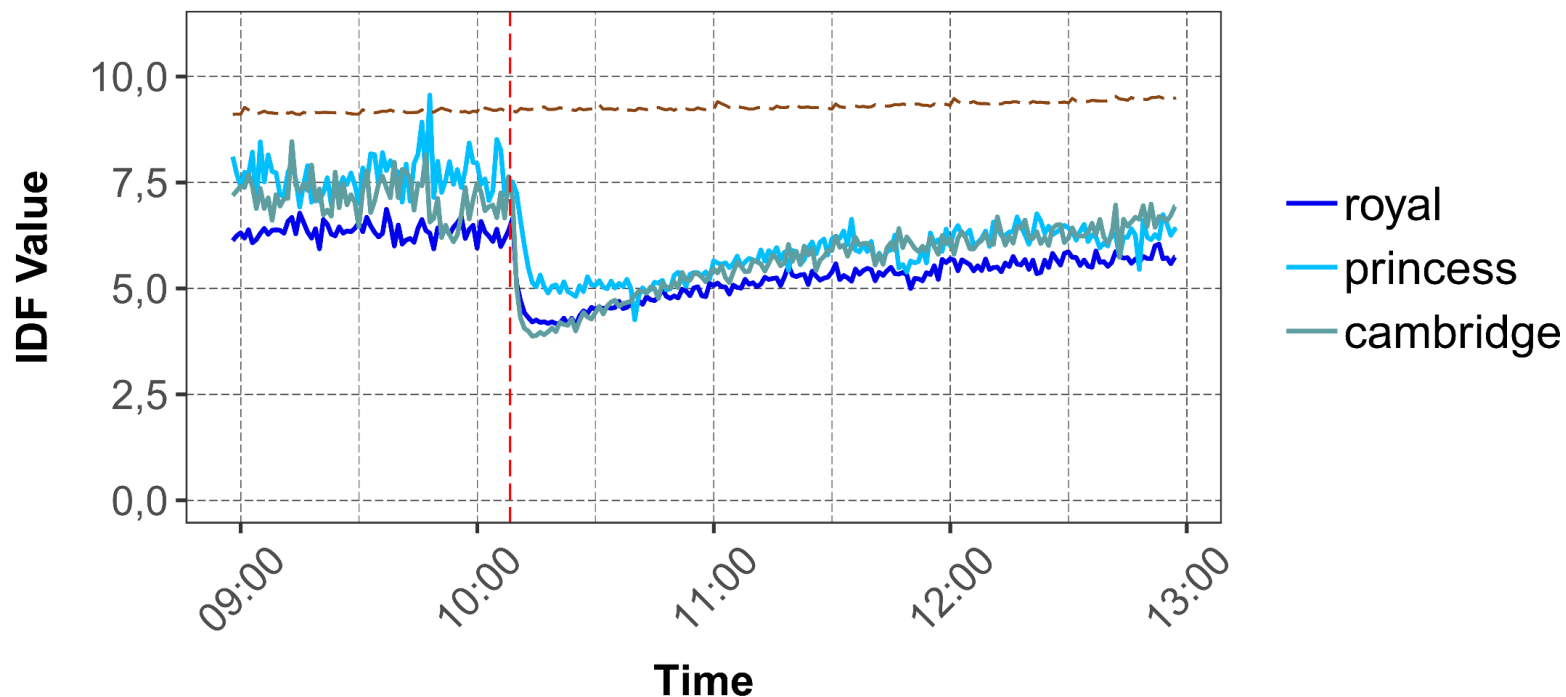
Twistor

2. Embedding of events

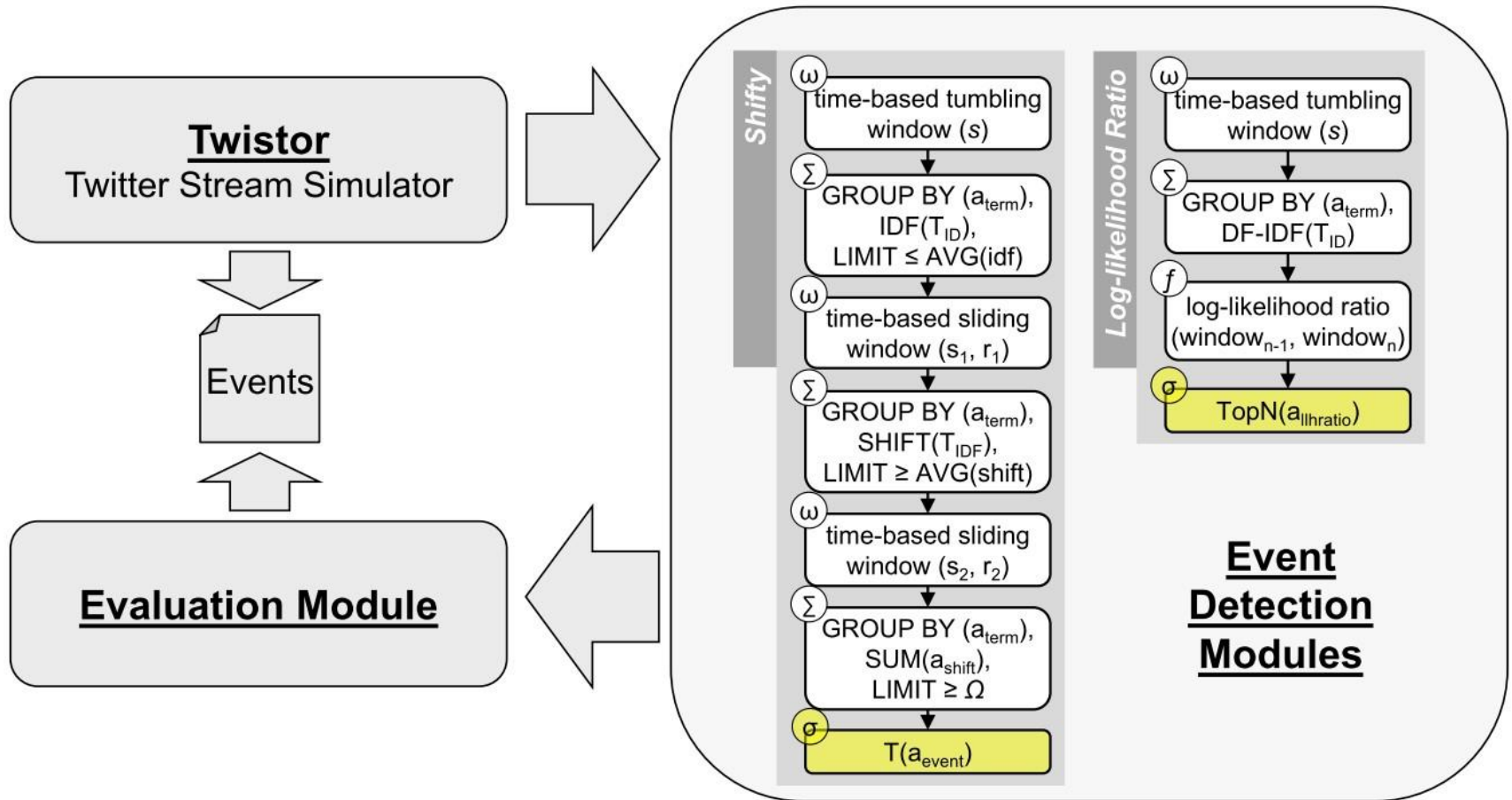- $\text{idf}(w) = \log\left(\frac{N}{n_w}\right)$

royal
princess
cambridge

Twistor

2. Embedding of events

- $\mathrm{idf}(w) = \log\left(\dfrac{N}{n_w}\right) \Leftrightarrow n_w = \dfrac{N}{e^{\mathrm{idf}(w)}}$
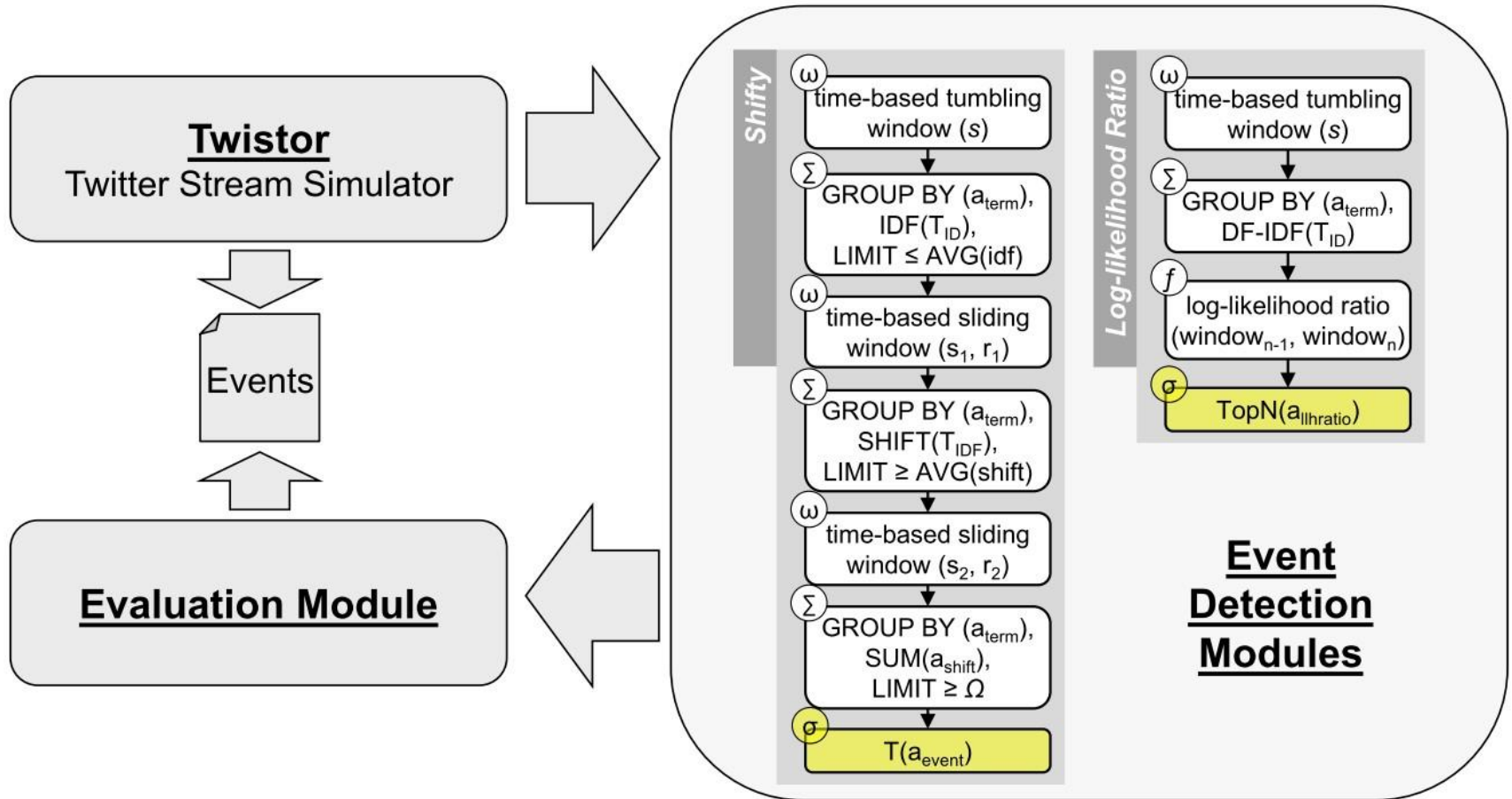
Approach

Event Detection Modules

- Data Stream Management System
  - Shifty
  - Log-Likelihood Ratio (LLH)

Approach

Evaluation Module

- Analyzes events from event detection modules
  - Against ground truth (events from Twistor)

- Measures
  1. Quality (precision, recall, F_1)
  2. Throughput (tweets per second)
  3. Latency

Toolkit Evaluation

- Generation of 60 minutes 10% Twitter stream
  - 1.5 million tweets
  - 25,000 tweets per minute

- Embedded 10 events into the artificial Twitter stream

- TopN (baseline), LLH, Shifty
  - Different parameter configuration → 61 result sets for each technique

- Measures (F_1, Throughput, Latency)
  - Throughput and latency normalized between 0 and 1

Results