Sequence Pre-processing: Focusing Analysis of Log Event Data

Alper Sarikaya[1], Emanuel Zgraggen[2],
Rob DeLine[3], Steven Drucker[3], and Danyel Fisher[3]

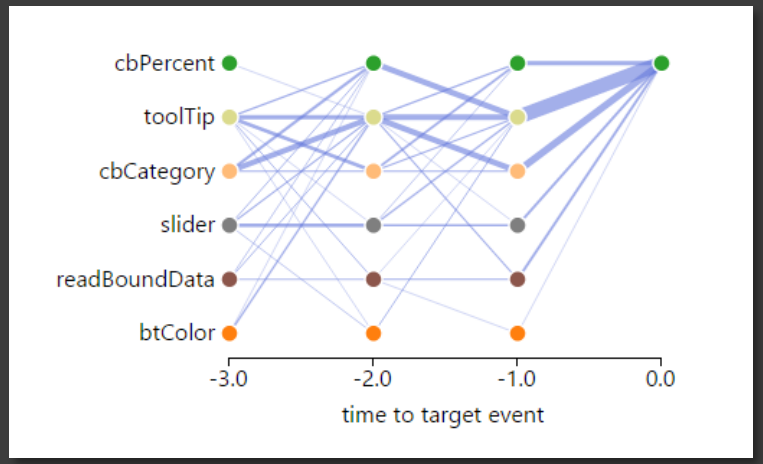
1. University of Wisconsin-Madison
2. Brown University
3. Microsoft Research

BROWN

Microsoft

The Noise, the noise!

"Listen... do you hear?
That's the sound of ultimate suffering..."



Exploring log sequence data —
Very regular occurrences of noise

Noise makes downstream visual analysis difficult; need to **pre-process**

To handle this noise, necessary to understand how log data is used

Potential Analysis Tasks

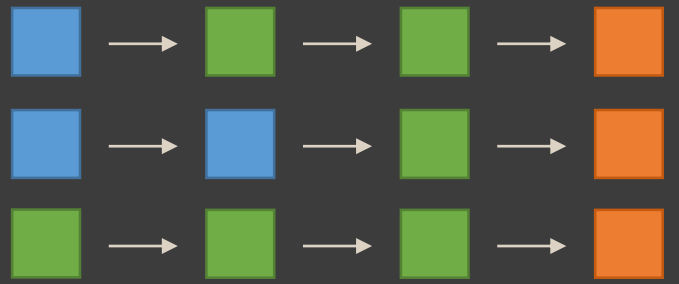
Given a log event has a

- Timestamp (to order)
- Event Name
- Session identifier
- Attributes (optional and open-ended)

Event: RequestData
Time: 1476722866600
Session: guest13
Request params: {...}

Event: HelpRequest
Time: 147672346600
Session: guest13

What are the sequences of events that lead to an event?



Potential Analysis Tasks

Given a log event has a

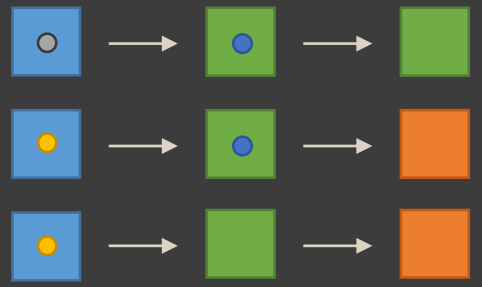
- Timestamp (to order)
- Event Name
- Session identifier
- Attributes (optional and open-ended)

Event: RequestData
Time: 1476722866600
Session: guest13
Request params: {...}

Event: HelpRequest
Time: 147672346600
Session: guest13

What are the sequences of events that lead to an event?

What attributes of what events indicate an event occurring?



Potential Analysis Tasks

Given a log event has a

- Timestamp (to order)
- Event Name
- Session identifier
- Attributes (optional and open-ended)

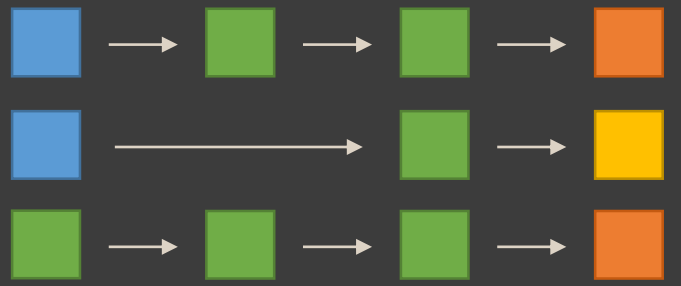
Event: RequestData
Time: 1476722866600
Session: guest13
Request params: {...}

Event: HelpRequest
Time: 147672346600
Session: guest13

What are the sequences of events that lead to an event?

What attributes of what events indicate an event occurring?

How does the temporal nature affect if an event is reached?



What are the Distractions?

Repeated Event — one event is important

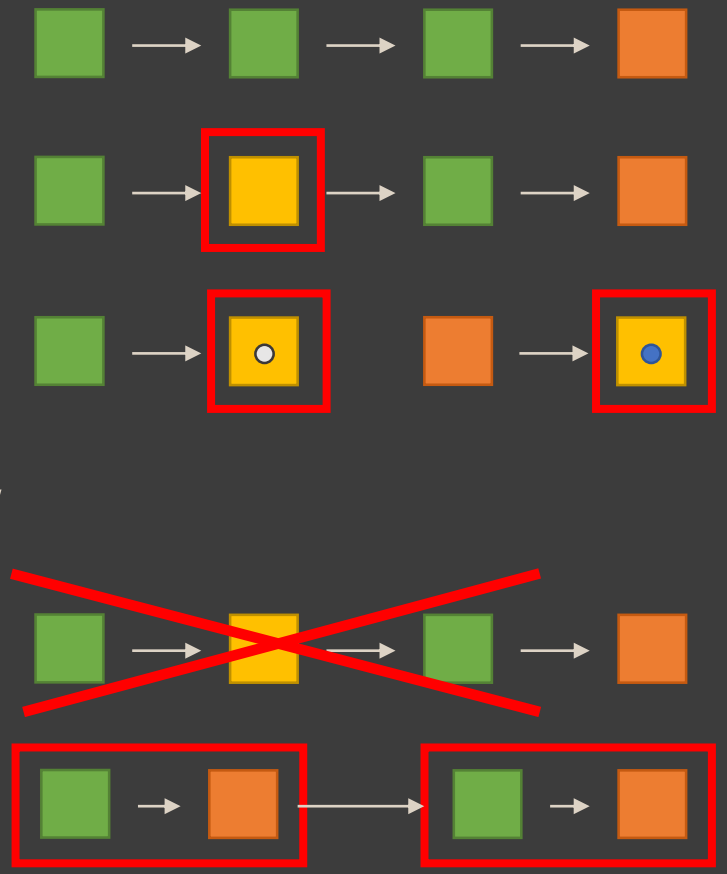
Useless Event — orthogonal to analysis

Ambiguous Event — hiding relevant info

Irrelevant Sessions — only relevant workflow

Negation — difficult to specify [1]

Subsequences — masks repeated workflow



[1] M. Monroe, et al. “The Challenges of Specifying Intervals and Absences in Temporal Data Queries: A Graphical Language Approach.” In Proc. *ACM CHI*, 2349–2358, 2013.

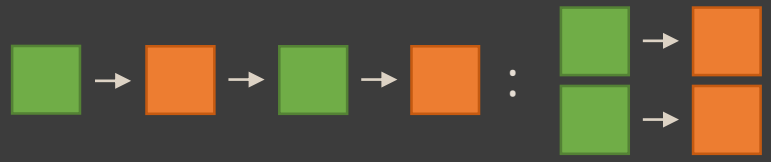
Methods for Focusing Analysis

Removing event(s) — only focus on relevant events

Replace with surrogate — orthogonal to analysis

Select sessions — hiding relevant info

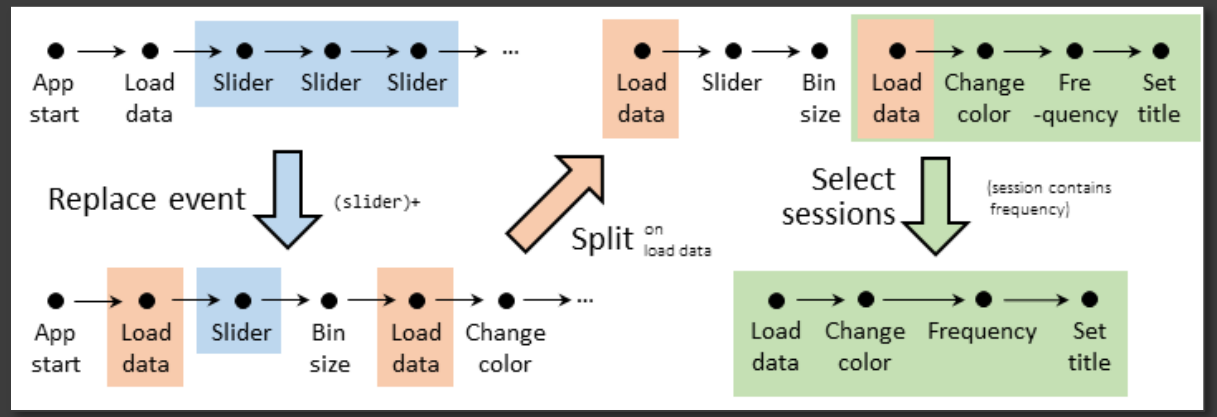
Re-sessionize — only relevant workflow



Negation — naturally falls out from any query (esp. for cohort comparison)

Analyst Feedback — display matched events and sessions (numbers and %s)

Applications



Pre-processing rules are composited in analysis —

Focuses downstream analysis

(e.g. why does one pick frequency?)

Great for **functional reactive programming** —

define a ‘pre-processing’ ruleset

process/analyze in real-time (we used **Trill** for this)

Questions? + Discussion!

Support by Microsoft Research (**Logan** project)