Statistical Methods for Change Detection over Time in Digital Forensics Data

Forensics@NIST Conference, Nov 8-9 2016

Padhraic Smyth
CSAFE Project Investigator
Professor, Departments of Computer Science and Statistics
University of California, Irvine

smyth@ics.uci.edu

User Data from Digital Devices

Web clicks
Web searches
Emails
Text messages
Social media posts
GPS locations
Documents edited
.....

Software Tools for Digital Forensics

Device

Extraction Tools

Raw Data

Indexing/Linking

Query/Search

Visualization

Statistical Analysis

Autopsy® is an easy to use, GUI-based program that allows you to efficiently analyze hard drives and smart phones. It has a plug-in architecture that allows you to find add-on modules or develop custom modules in Java

The Timeline feature collects events from *all* Autopsy results with associated timestamps.

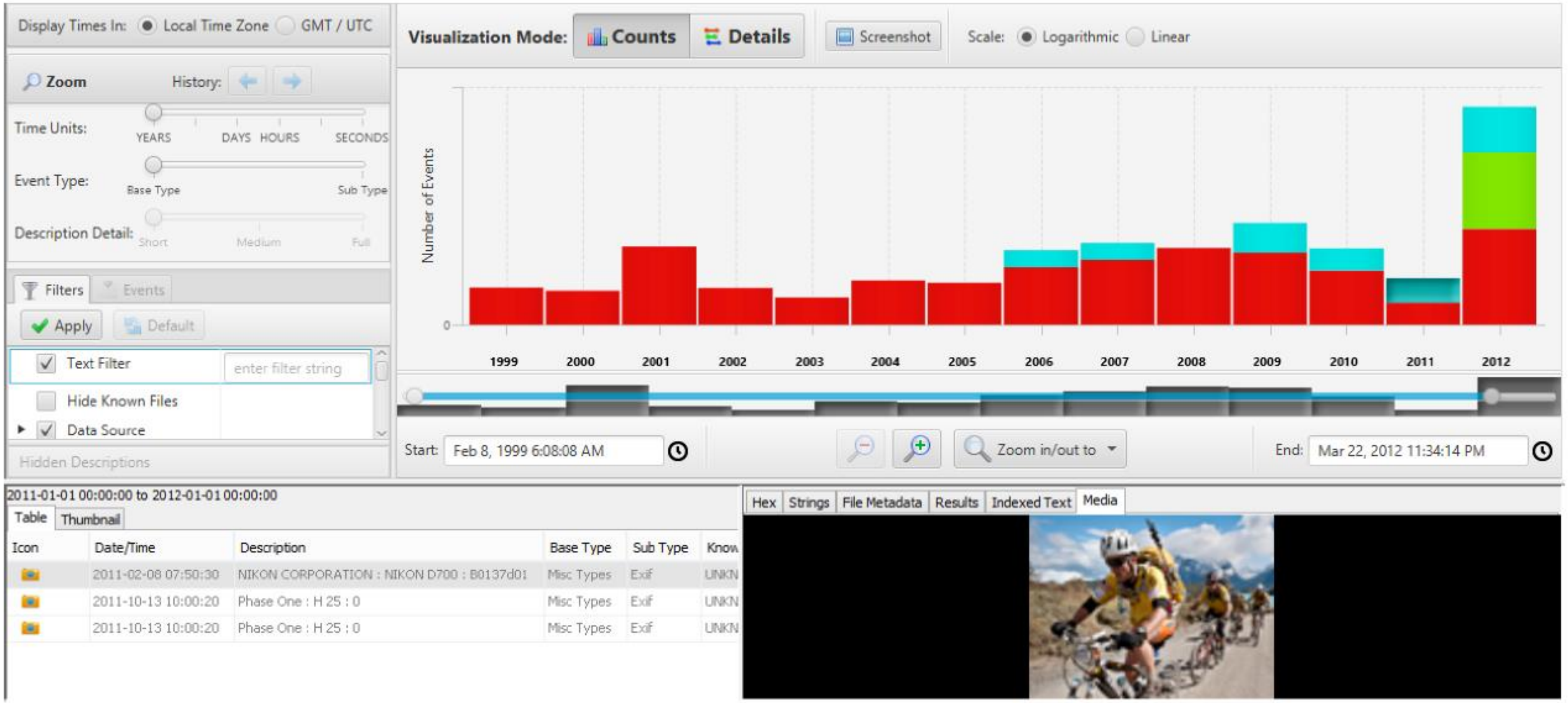
Events are stored in a dedicated DB optimized for timelines with millions of events

- File System
  - Modified
  - Access
  - Created
  - Changed
- Web Activity
  - Downloads
  - Cookies
  - Bookmarks (creation)
  - History
  - Searches
- Miscellaneous
  - Email
  - Recent Documents
  - Installed Programs
  - Exif metadata
  - Devices Attached
  - Text Messages (Android)
  - Call Log(Android)
  - GPS Searches(Android)
  - GPS Locations(Android)

From Timeline Visual

J. Millman, *Open Source Digital Forensics Conference*, 2014

Timeline

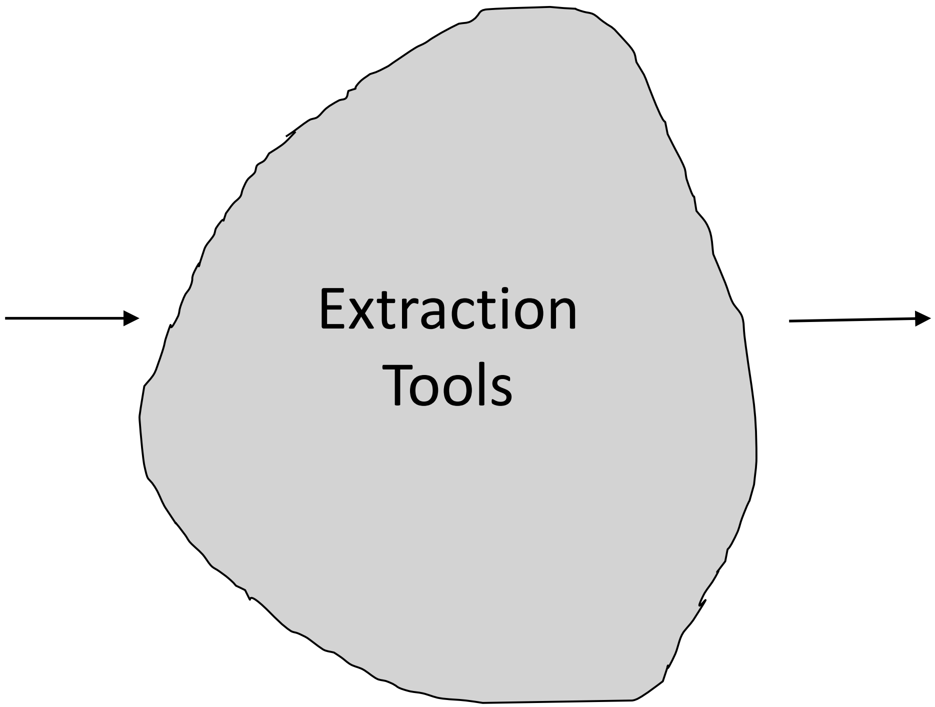


#OSDFCon 30

From B. Carrier, *Open Source Digital Forensics Conference*, 2015

Software Tools for Digital Forensics

Device



Raw Data

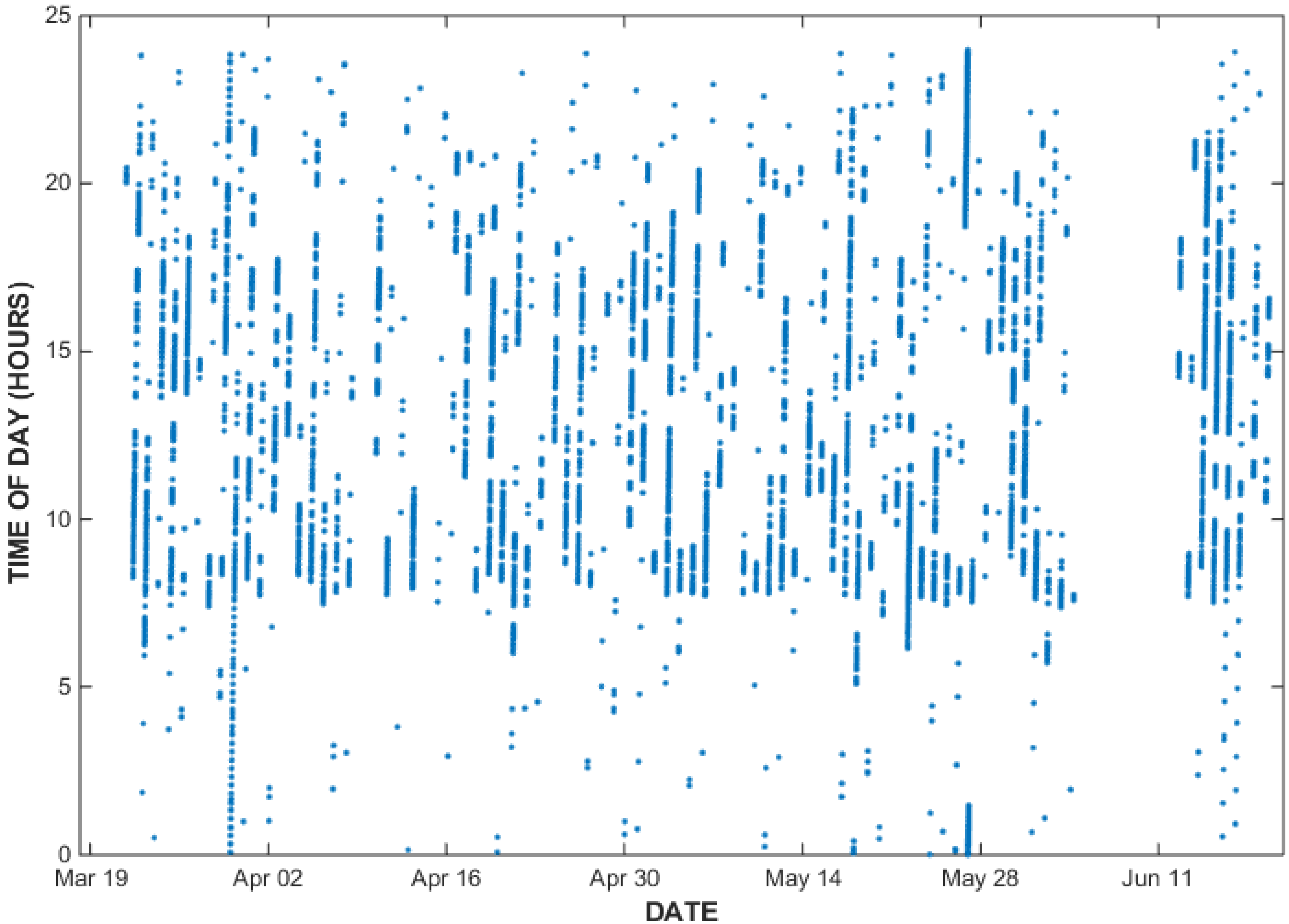
Indexing/Linking

Query/Search

Visualization

Statistical Analysis

Example: Time Plot of URL Request (Browser) Data



Typical Sources of User Event Data

- **Local Device**
  - Browser history
  - Cookie files

- **Cloud**
  - Email history (e.g., Gmail)
  - Search history (e.g., Google/Chrome)
  - File editing (e.g., Google Docs)
  - Social Media activity
    - Facebook
    - Twitter

- **Caveats**
  - User may have deleted or obfuscated data
  - Cloud data may be inaccessible

User Event Data

< ID, timestamp, action type, metadata >

Web clicks
Web searches
Emails sent
Social media posts
Files edited
.....

Text content
Location
List of recipients
.....

Example of a User Event Data Set

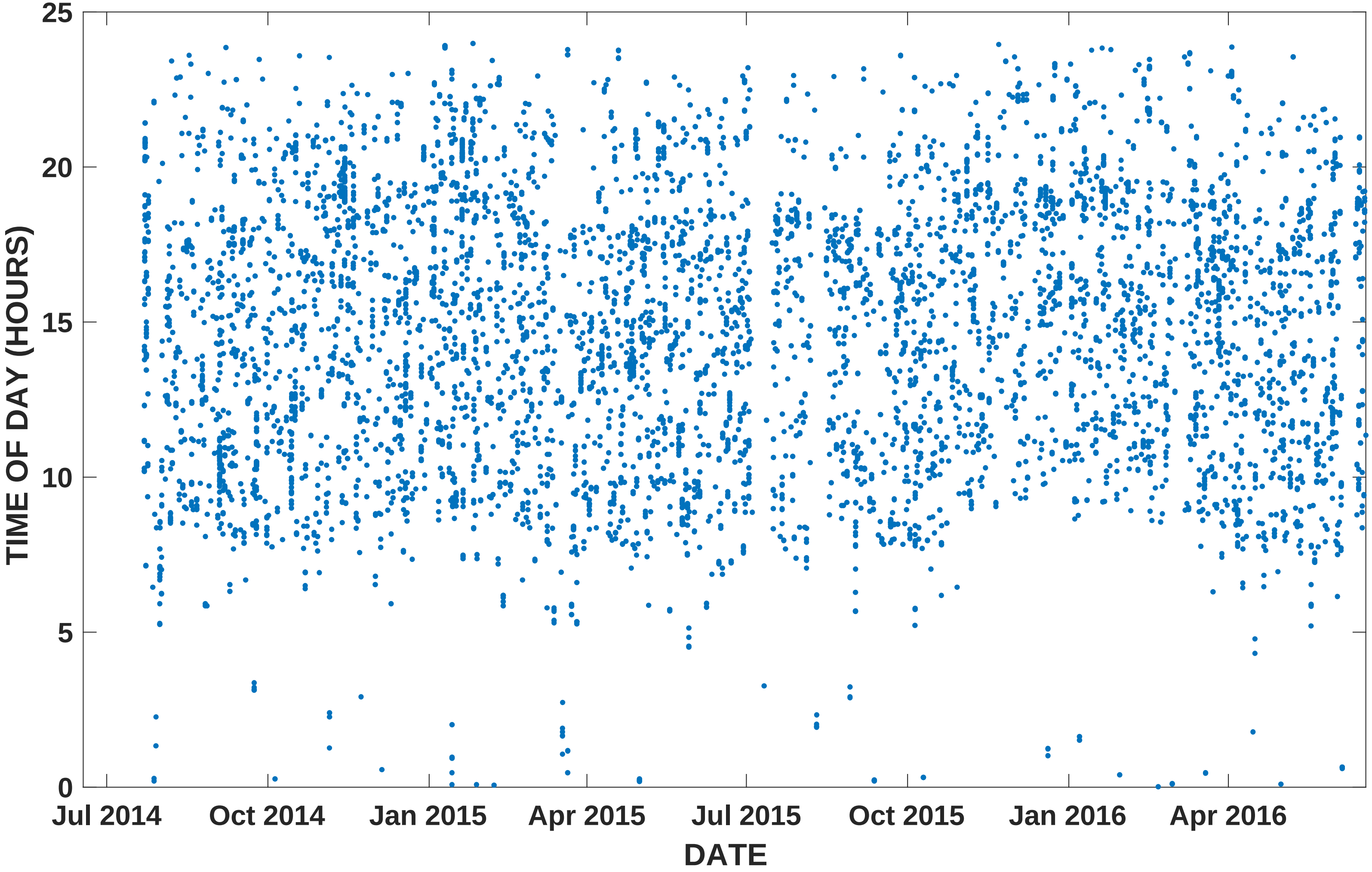
- **Chrome Browser History (local device)**
  - 37.9k (desktop) and 7.3k (laptop) browsing events, over 3 months
  - <timestamp, URL, + more…>

- **Google Search Queries (cloud)**
  - 7000 searches over 2 years
  - <timestamp, query string>

- **Facebook (cloud)**
  - Variety of time-stamped events and metadata over 7 years

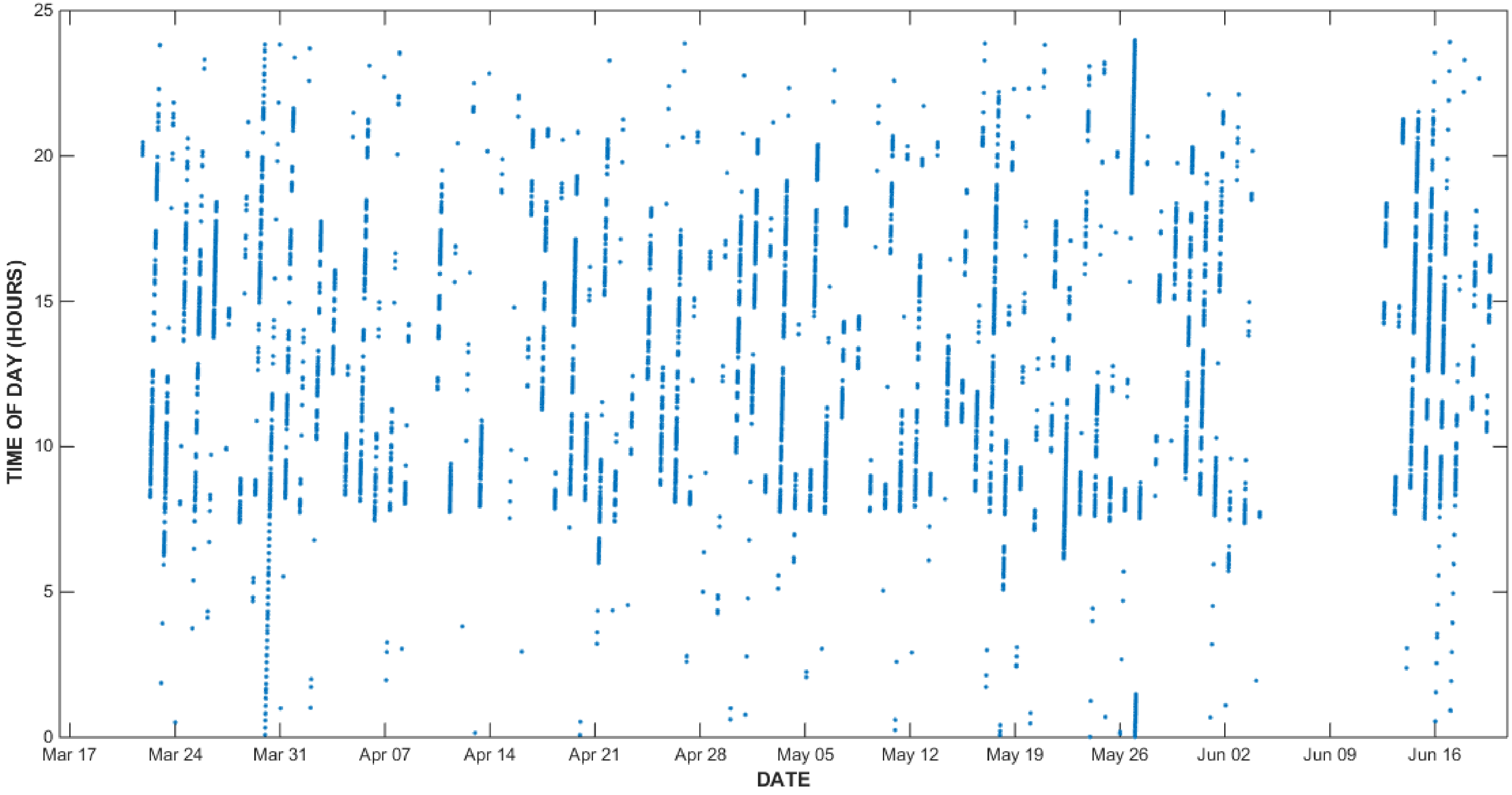
- **Gmail (cloud)**
  - Records of incoming and outgoing emails over 10 years

Timeline of Search Queries (from Cloud)

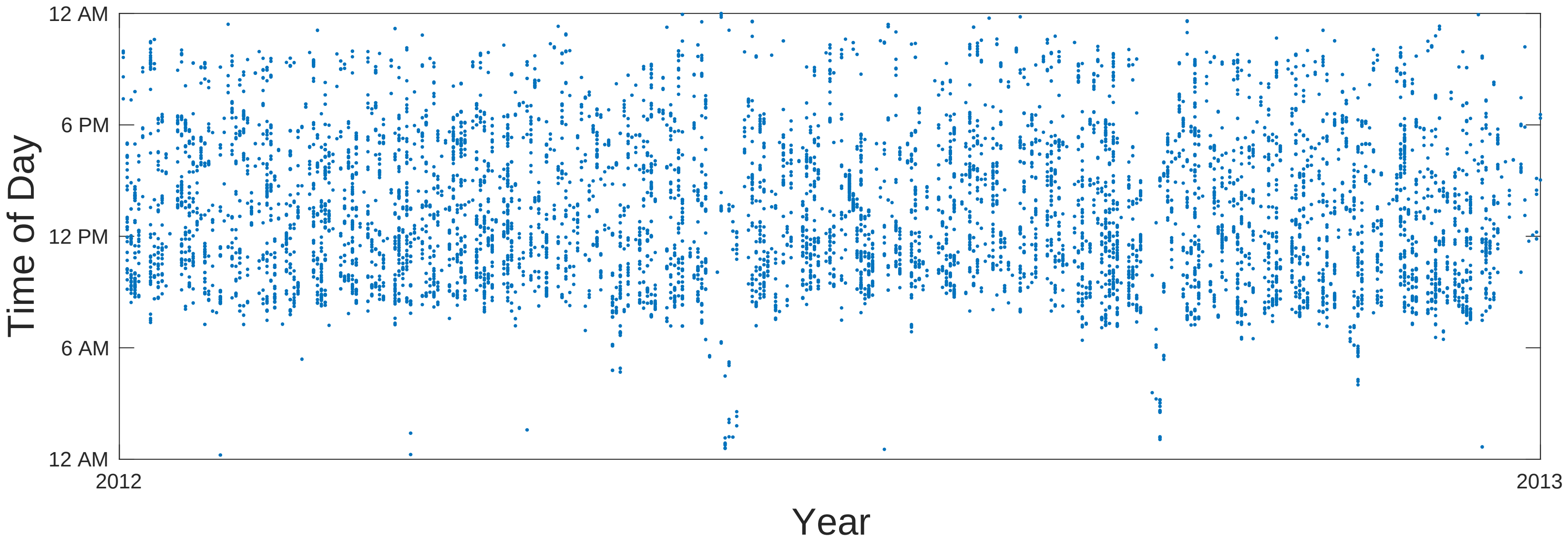


UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

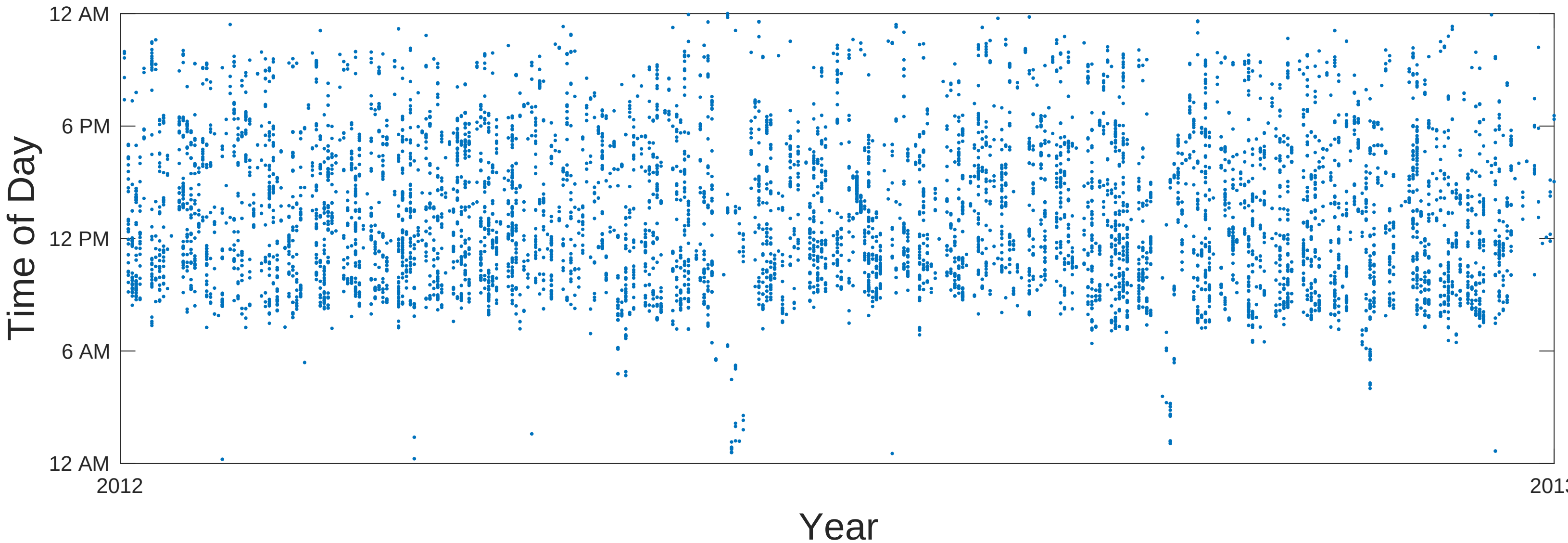
Timeline of Browser URL Requests (from Desktop Device)



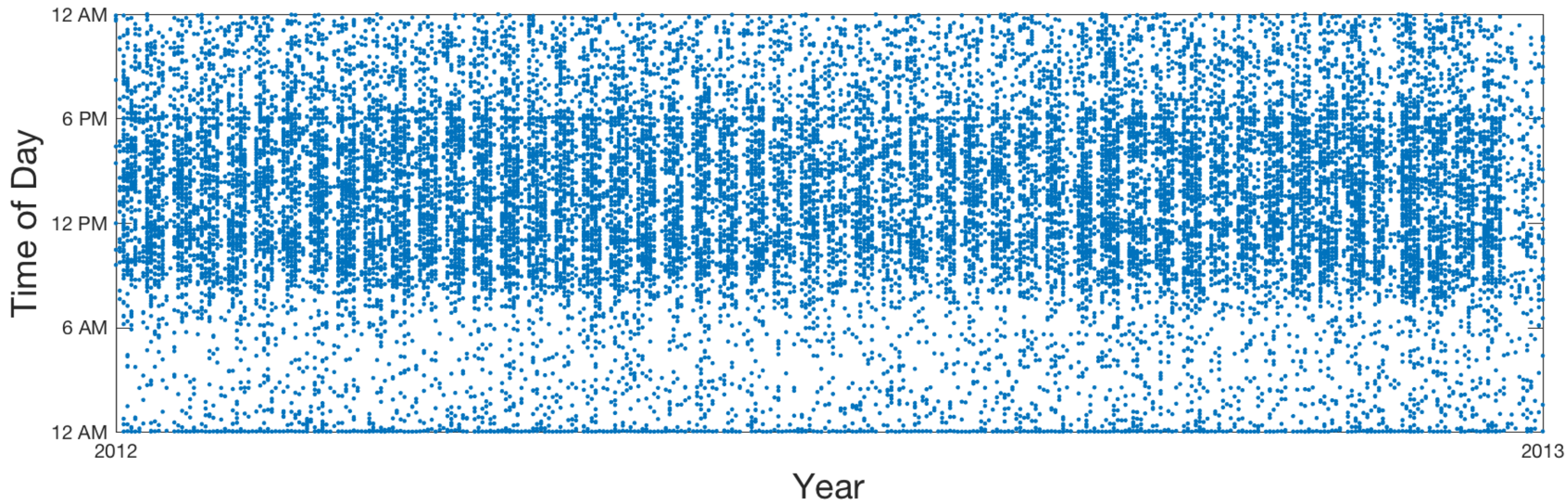
Time Plot for Emails Sent



Time Plot for Emails Sent



Time Plot for Emails Received



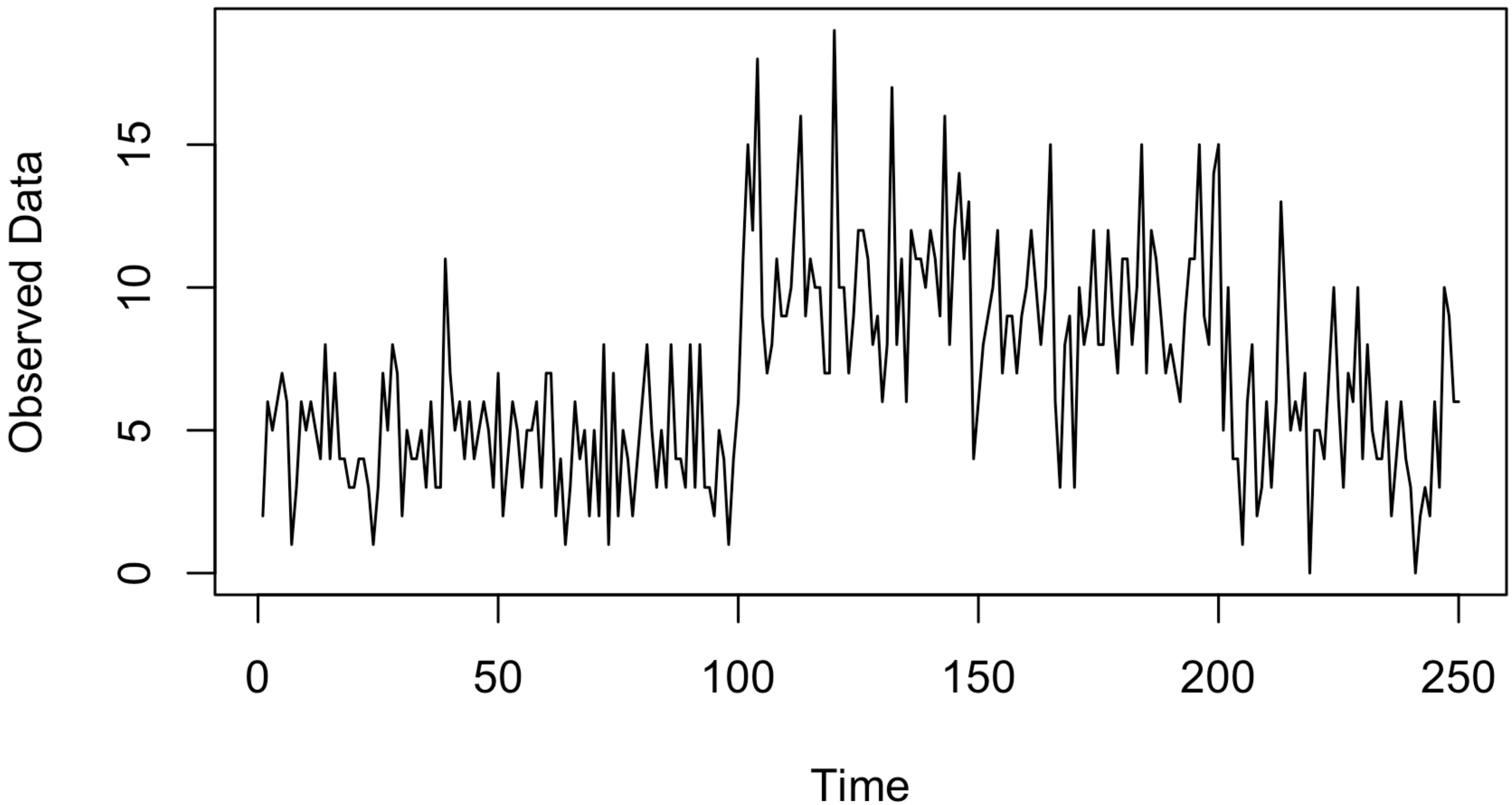
Potential Value to Forensics

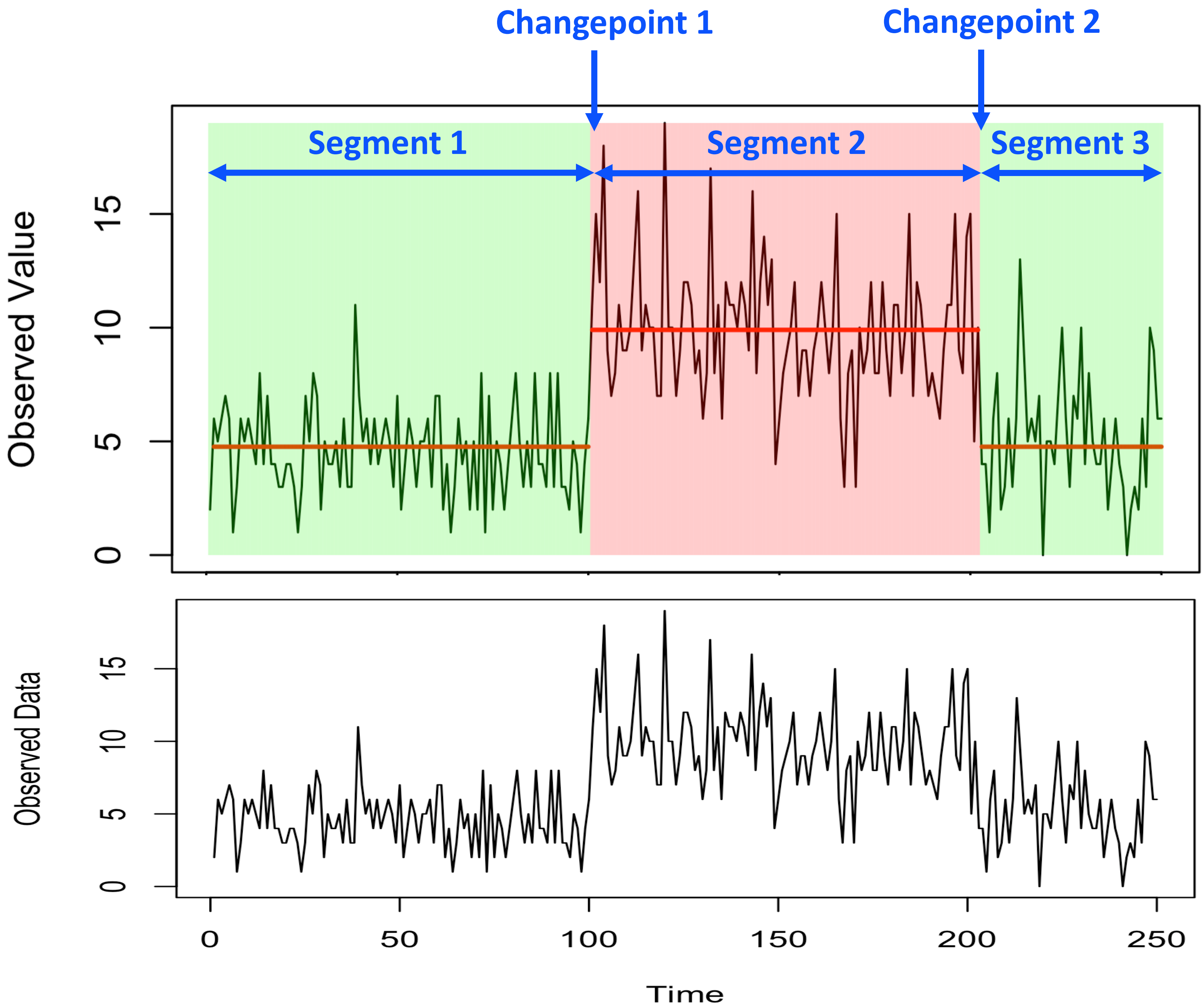
- **Assist in discovery process**
  - Detect and focus attention on time-periods of unusual behavior
  - Summarize an individual's behavioral patterns over time
  - Compare how two accounts A and B differ in behavior

- **Quantify answers to specific questions**
  - Is there evidence of a significant change in behavior at specific times?
  - Is there evidence of more than 1 user in an event stream?
  - Is the behavior on device X consistent with the behavior on device Y?

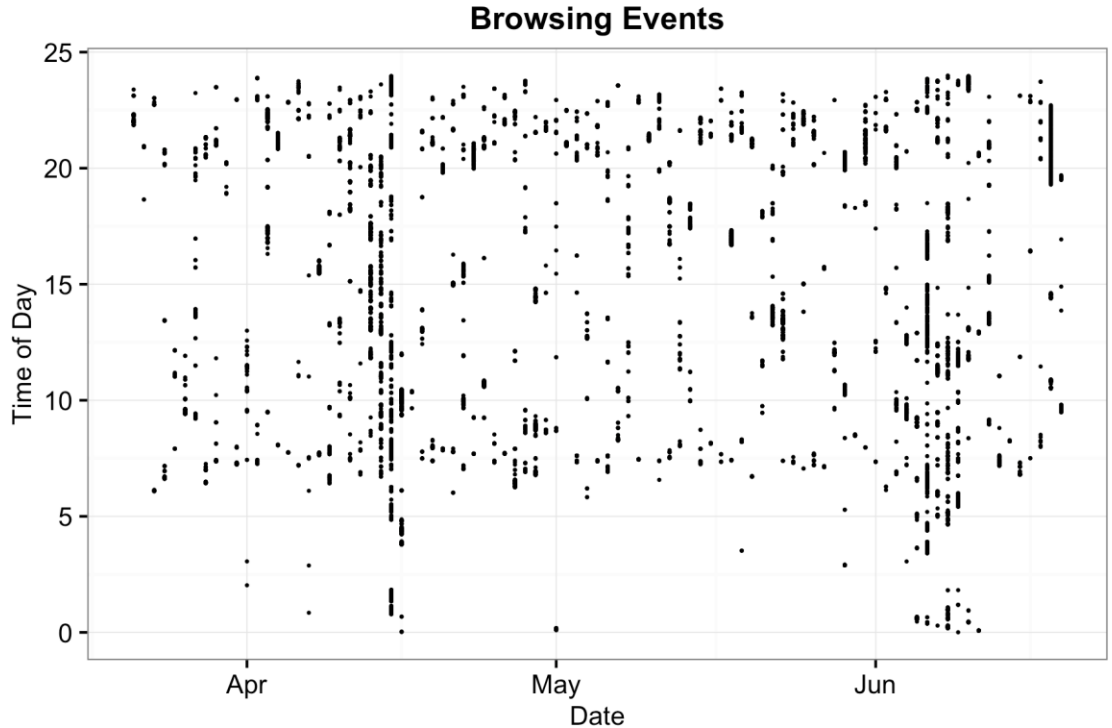
Changepoint Detection

Changepoint: significant change in distributional characteristics of a time-series, e.g., change in mean, change in variance

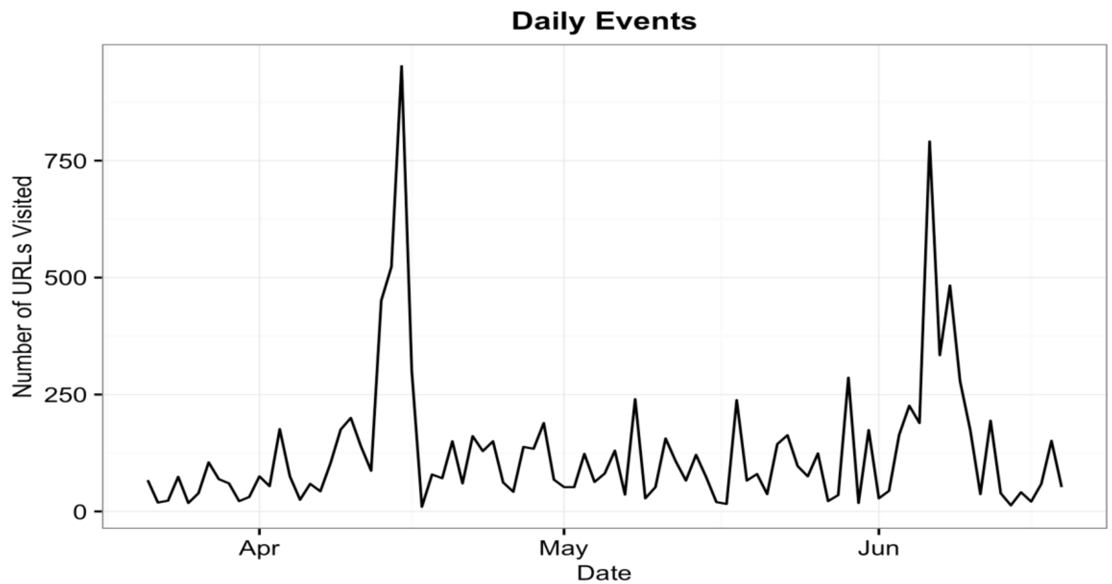



Changepoint 1
Changepoint 2
Segment 1
Segment 2
Segment 3
Observed Value
0
5
10
15
Observed Data
0
5
10
15
0
50
100
150
200
250
Time

User Browser Request Data



User Daily Counts



Statistical Approaches to Change Detection

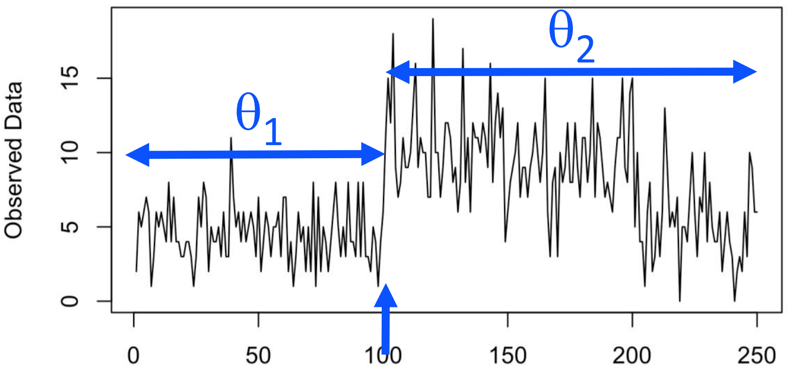
- **Assume a time-series model where unknown parameters are**
  - Parameters for distributions within each segment
  - Number and locations of changepoints and segments

- **Fit this model to the observed data and infer both**
  - How data is distributed within segments
  - Locations of changepoints and segments

- **"Chicken-and-egg" estimation problem**
  - Given segments, can easily estimate distributions
  - Given distributions, can easily estimate location of changepoints

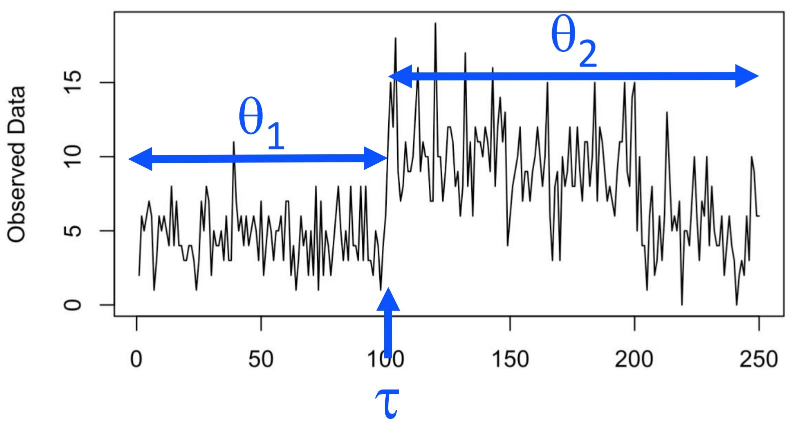
Example: Maximum Likelihood Detection of 1 Changepoint

3 unknown parameters:



Example: Maximum Likelihood Detection of 1 Changepoint

3 unknown parameters:



Likelihood Function:

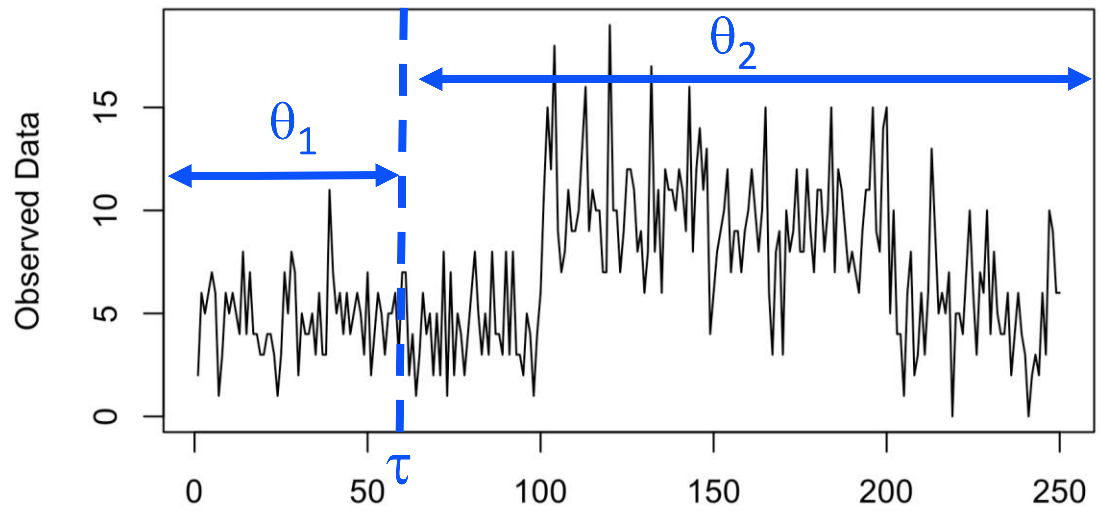
$$L(\theta_1, \theta_2, \tau) = P(\mathrm{data}|\theta_1, \theta_2, \tau) = \prod_{t=1}^{\tau} P(x_t|\theta_1) \prod_{t=\tau+1}^{T} P(x_t|\theta_2)$$

Maximum likelihood parameter estimates

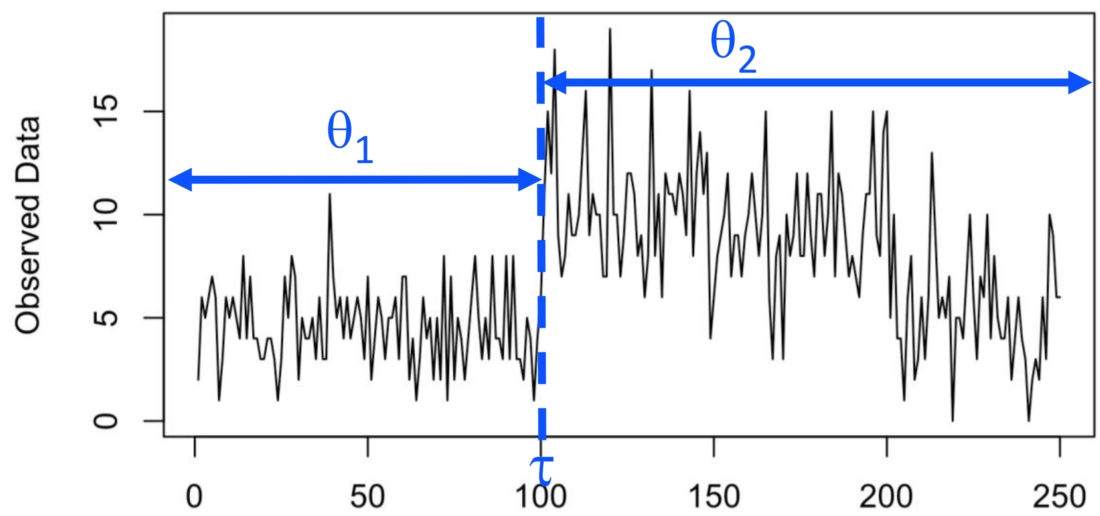
= values θ_1, θ_2, τ **that maximize** $L(\theta_1, \theta_2, \tau)$

Example: Maximum Likelihood Detection of 1 Changepoint

Low Likelihood



High Likelihood



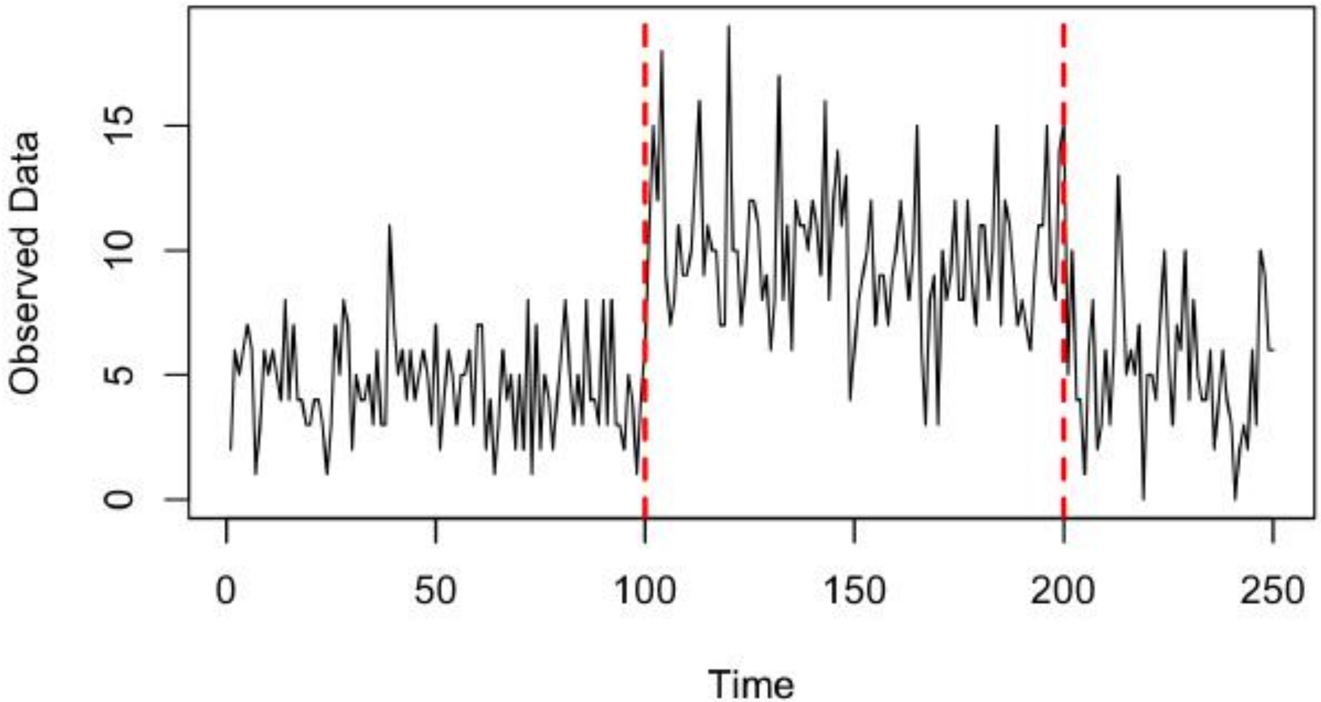
Approach 1: Direct Segmentation Models

- **Segment Distribution**
  - Assume a distributional form within segments (Poisson, Gaussian, etc)

- **Search for K changepoints that maximize the likelihood**
  - As K increases, search problem becomes combinatorially more difficult
  - Requires heuristic search techniques (e.g., greedy search) for K > 1

- **Problem: how to select K?**
  - More complex models (with larger K) always have higher likelihood
  - Model selection problem, e.g.,
    - Use penalized likelihood: subtract a penalty term from likelihood (AIC, BIC, etc)
    - Use Bayesian techniques such as marginal likelihood

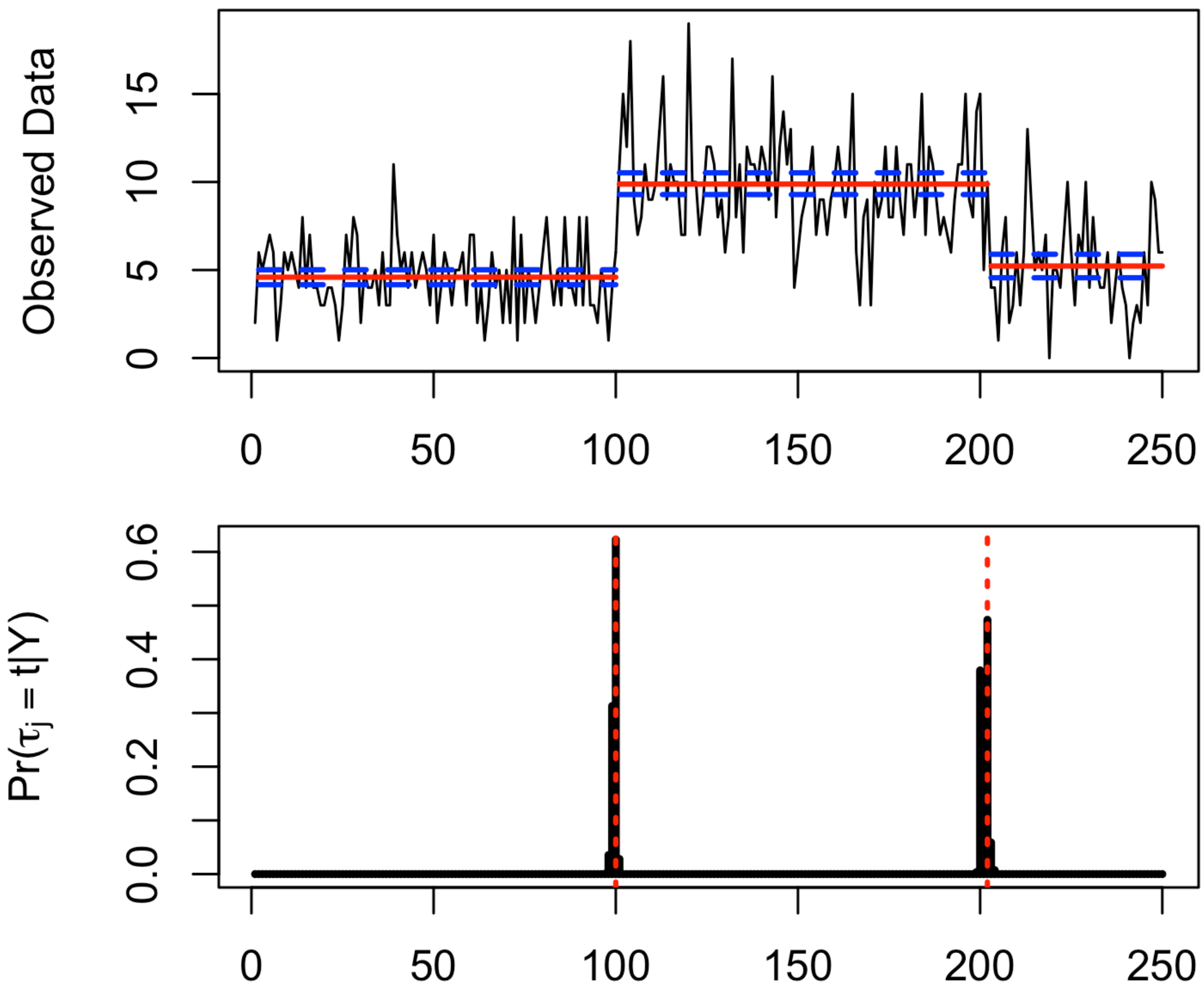
Simulated Data, Two Changepoints



Time series of length $n = 250$ simulated in the following manner:

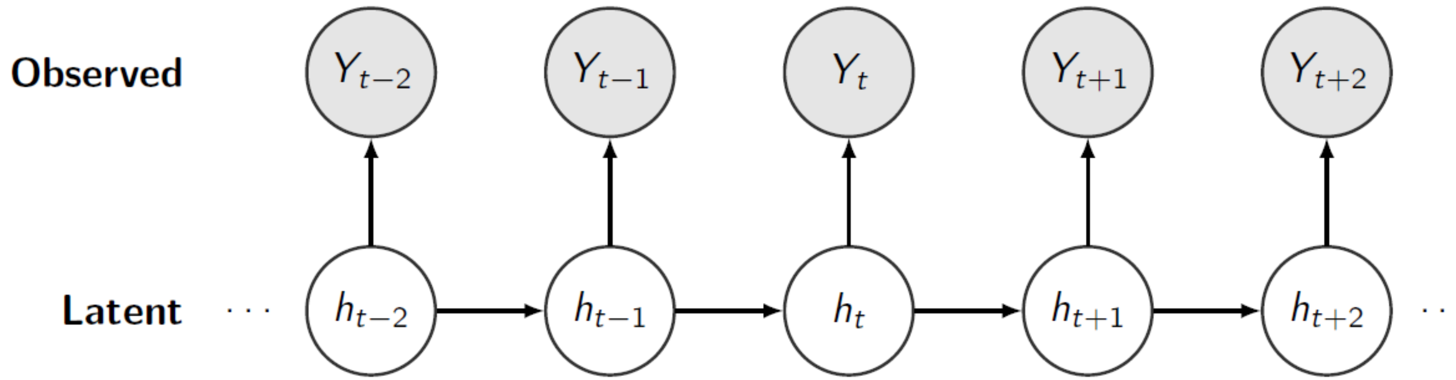
- $\lambda = (\lambda_1 = 5, \lambda_2 = 10)$
- $\tau = (100, 200)$
- $Y_{1:100} \sim \text{Poisson}(\lambda_1)$
- $Y_{101:200} \sim \text{Poisson}(\lambda_2)$
- $Y_{201:250} \sim \text{Poisson}(\lambda_1)$

Results from Bayesian Segmentation



Approach 2: Dynamic Models

- **Assume that data can be explained by a dynamic model that switches between states, e.g., a hidden Markov model**

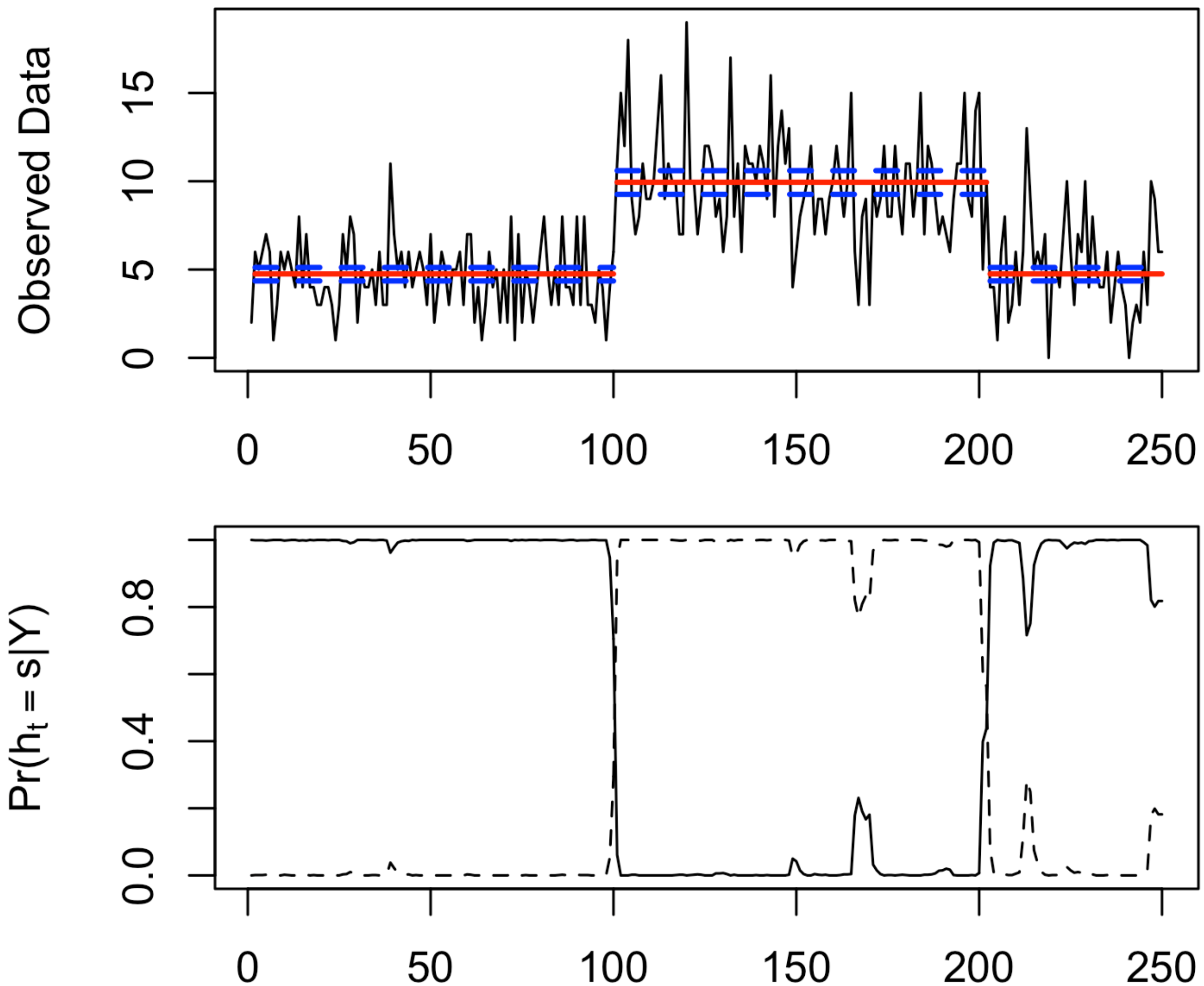


- **States = segments that can recur**
  - e.g., states = {work, business travel, vacation, ….}

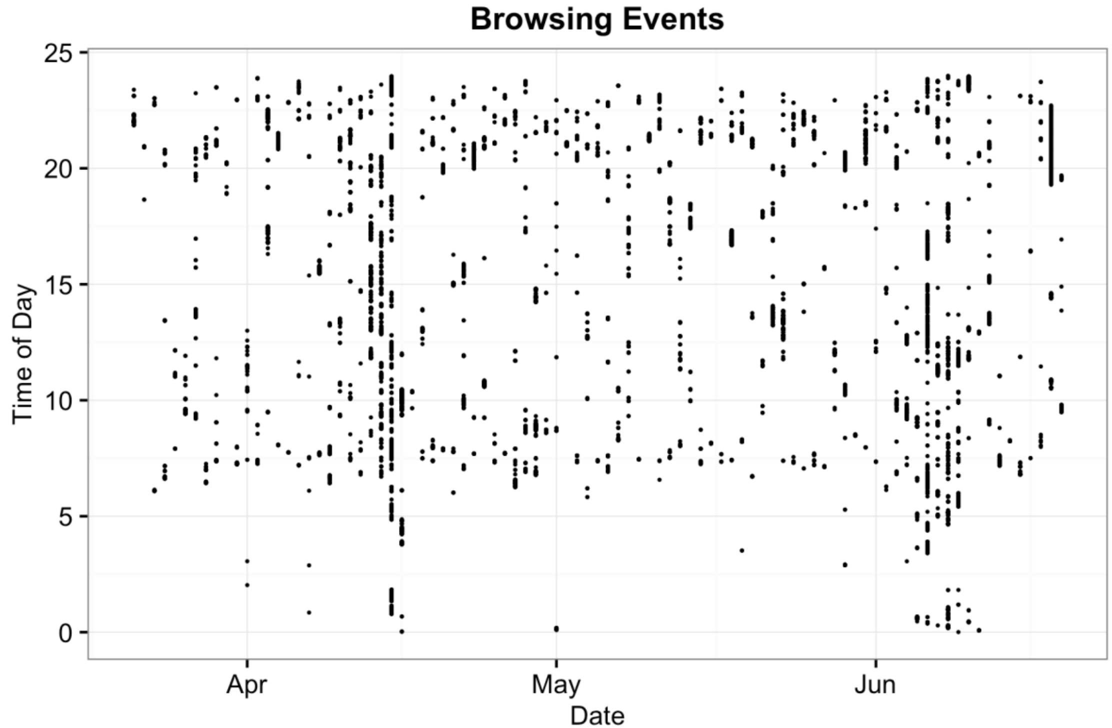
- **Differences with segmentation model**
  - Recurrent segments allow for borrowing of strength
  - Assumes that duration in segments is Markov/geometric
  - Can use dynamic programming to perform inference efficiently

UCIrvine
University of California, Irvine

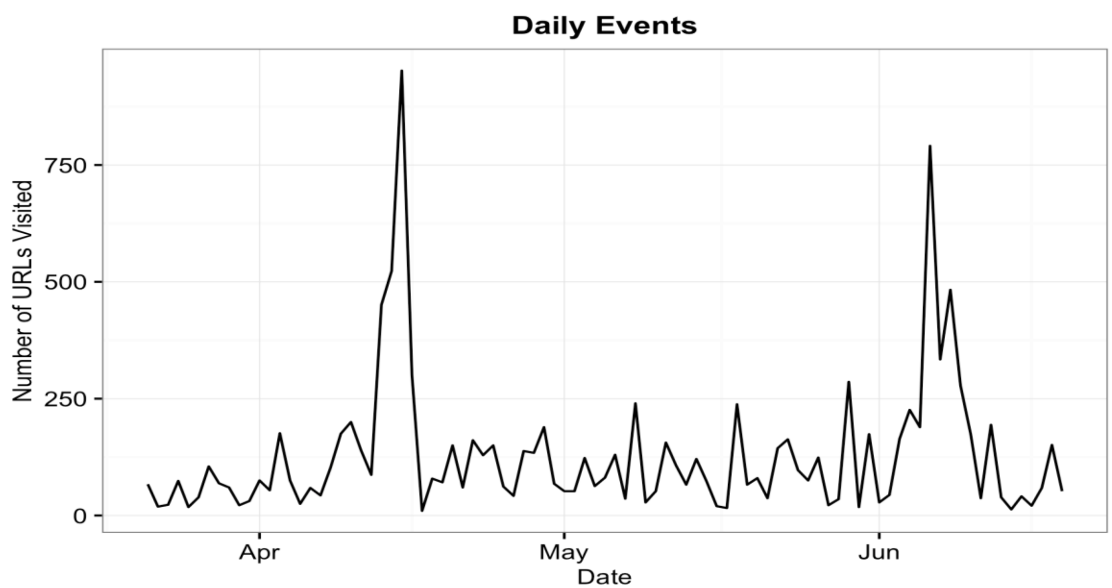
Results with Bayesian Hidden Markov Model on Simulated Data



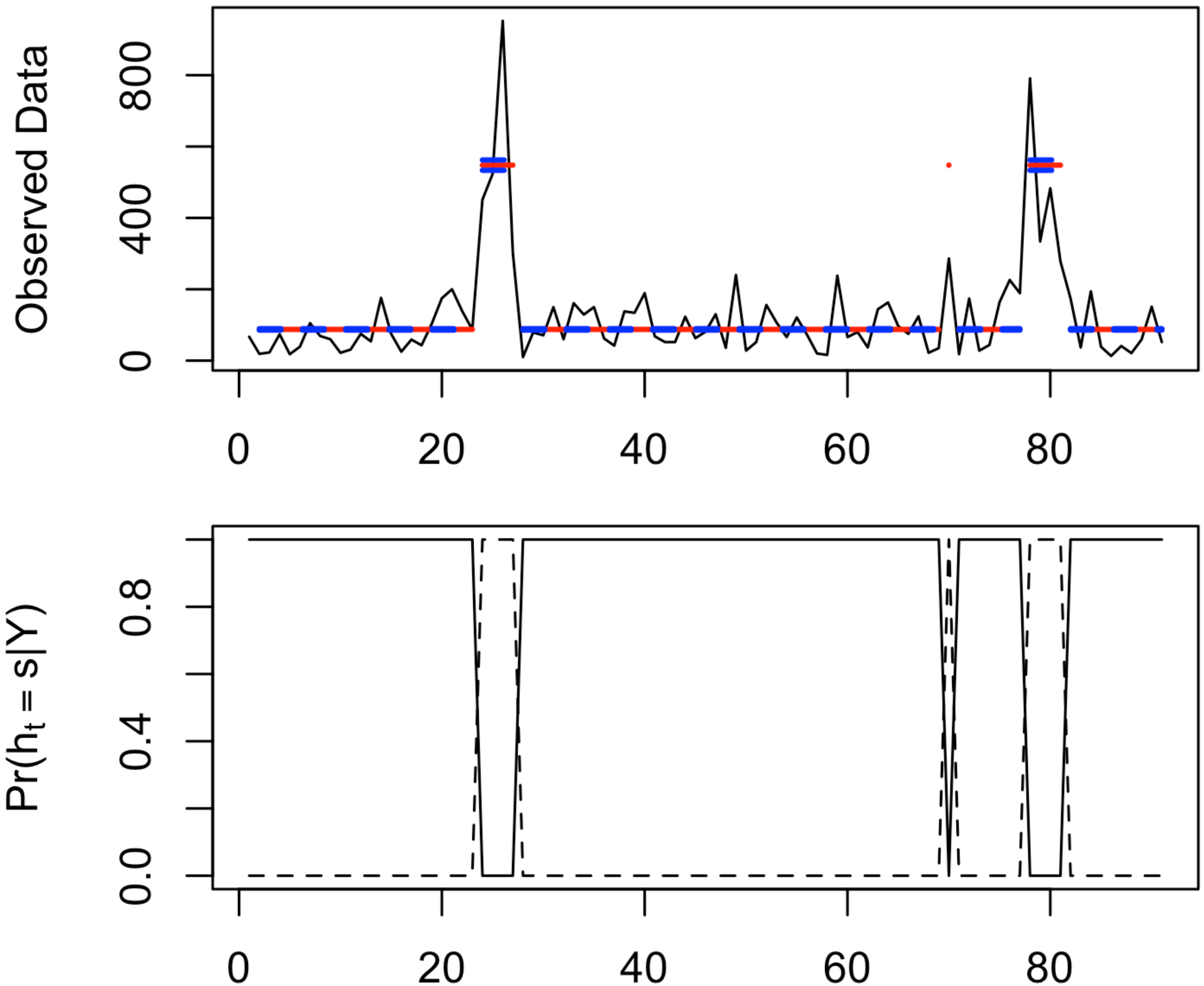
User Browser Request Data



User Daily Counts



Changepoint Detection Results on Real-World Data



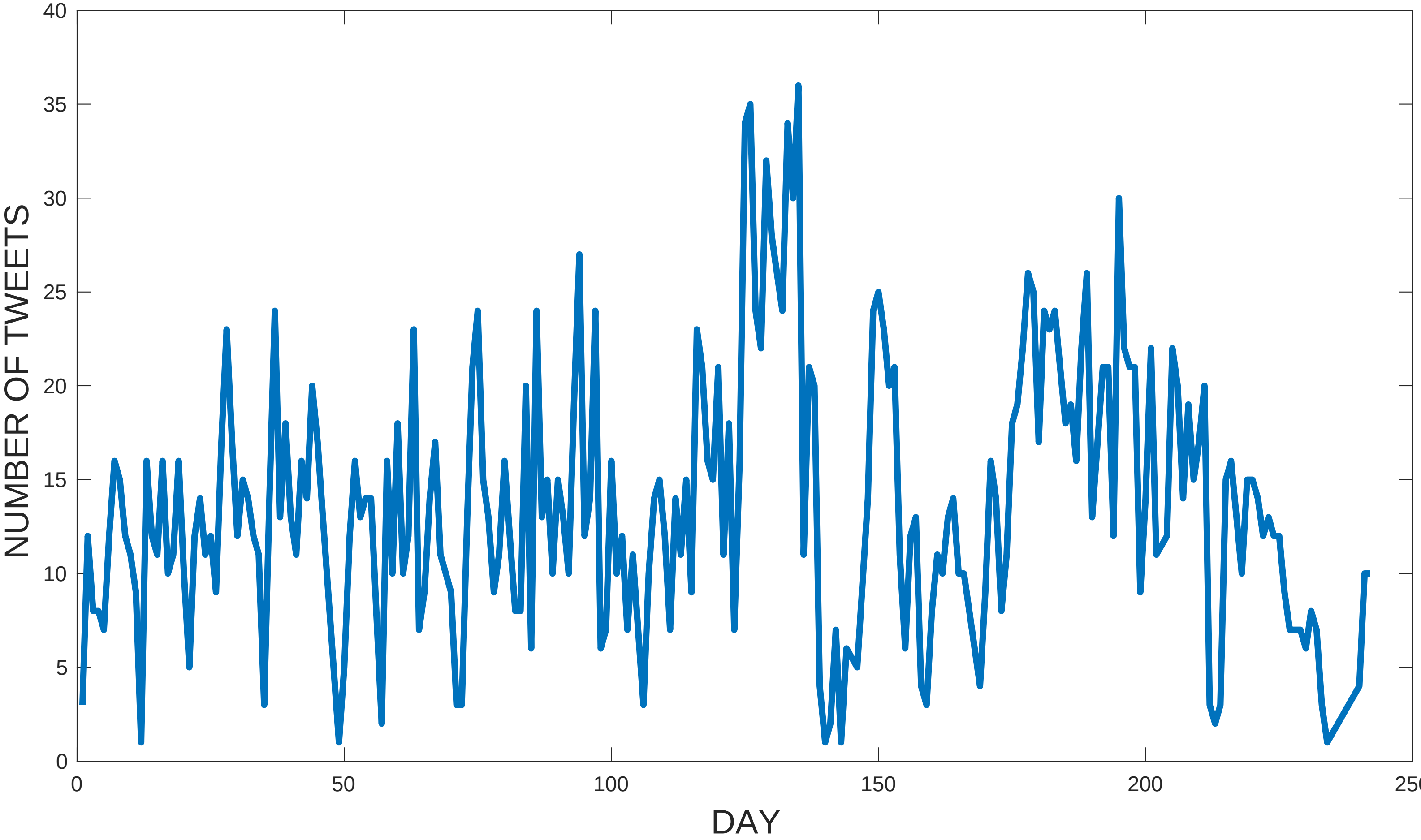
Ongoing Work (CSAFE)

- **Systematic evaluation of different approaches**
  - Simulated data: Compare estimated with true changepoints
  - Real-world data: Compare estimated changepoints to ground truth (if known)
  - Extensions to allow drift and trends in user behavior

- **Additional problems**
  - Multiple event streams
    - Change detection across streams
    - Likelihood of being from the same person?
  - Using timestamps in the analysis
  - Incorporating additional data such as text, email recipients, etc

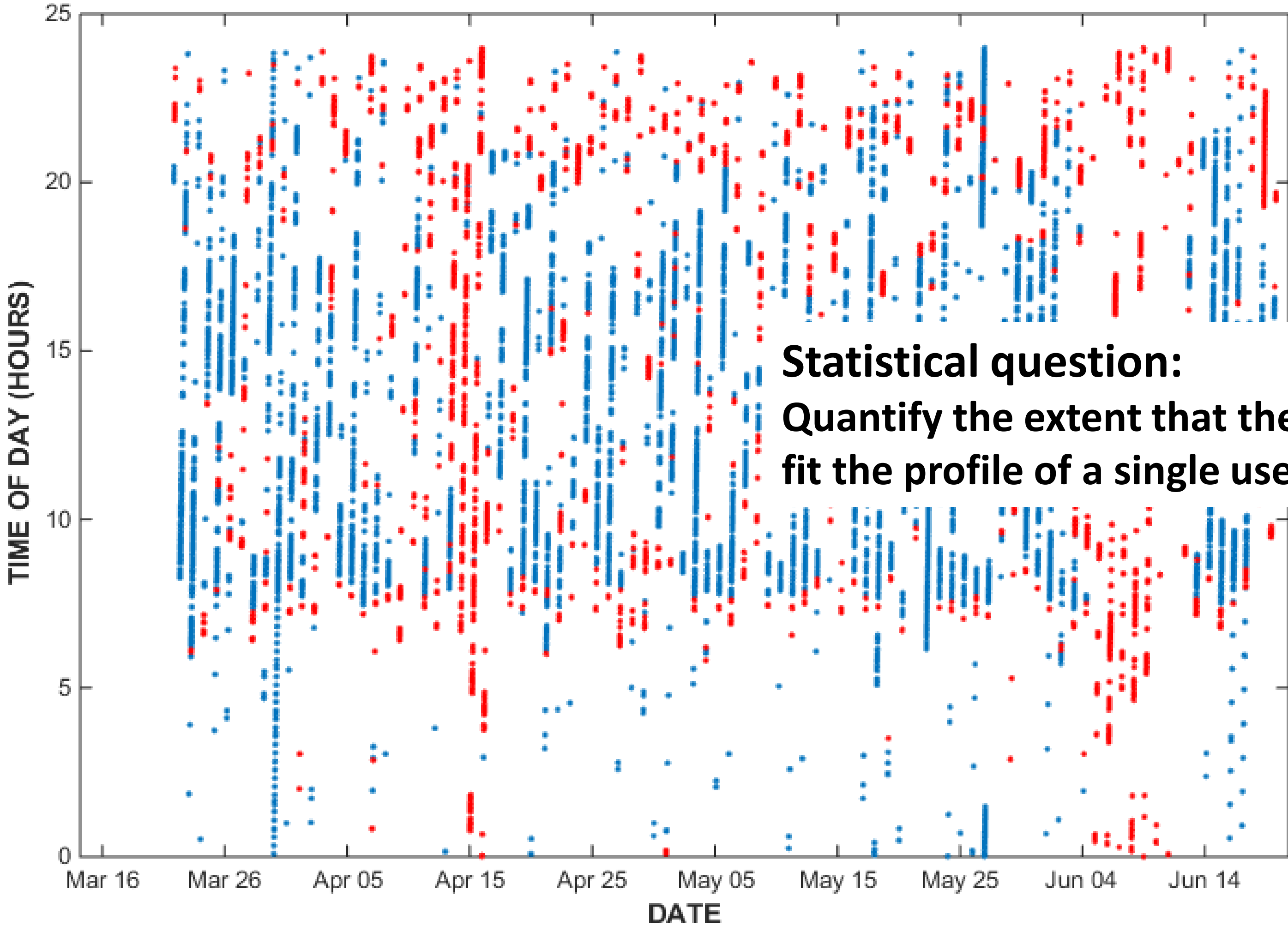
- **Creating a realistic research data set**
  - Planning underway for a study at UC Irvine to create anonymized data sets from student participants (with permissions)
  - Surrogate data sets such as Twitter or Reddit publicly available data

Example of Twitter Event Data over Time



URL Visits (desktop)

URL Visits (laptop)



Statistical question:

Quantify the extent that these data sets fit the profile of a single user

Research Challenges

- **Matching real-world digital forensic problems with statistical modeling**
  - There is a gap…
- **Variability of individual behavior**
  - Significant within-individual variability
  - No population reference for “1 in a million” statements
- **Testbed research data sets**
  - Privacy issues
  - Ground truth

Summary

A variety of native user data can be extracted from devices

Common data type: Events = [*user, timestamp, action, metadata*]

Natural to develop tools for statistical analysis of such data

- detection of significant changes over time
- numerous potential extensions

Challenge: making these techniques useful to forensic practitioners

- interact with forensic experts
- create research data sets that others can use
- develop open-source software for adoption