Doing Research on Systems for Interactive Analysis at an Industry Lab

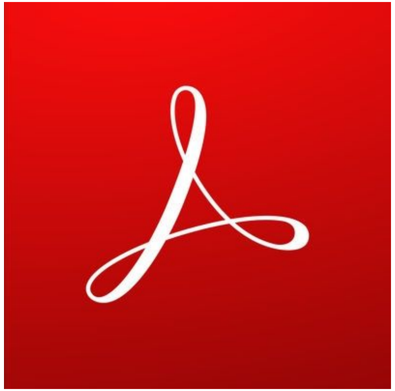
Leo Zhicheng Liu

Adobe Research

Creative

Marketing

Document



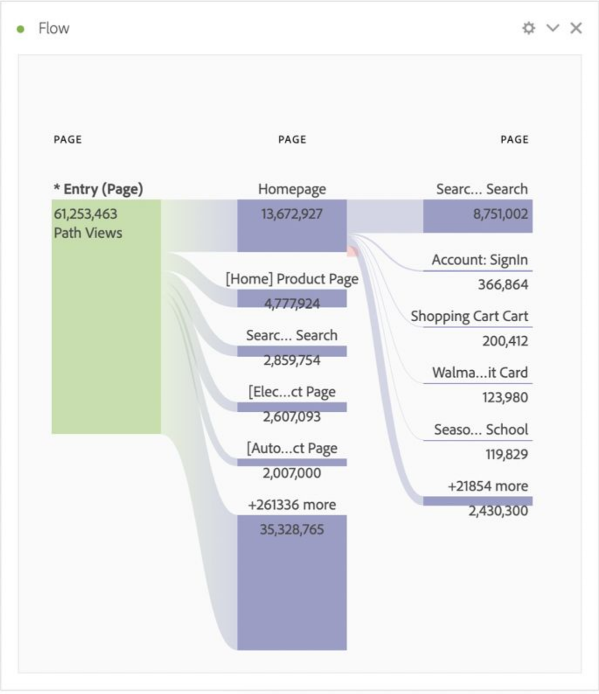
Adobe Analytics

“Google Analytics for Enterprises”

Tools for enterprise customers to measure and analyze customer pathing, traffic sources, content effectiveness, and video engagement.

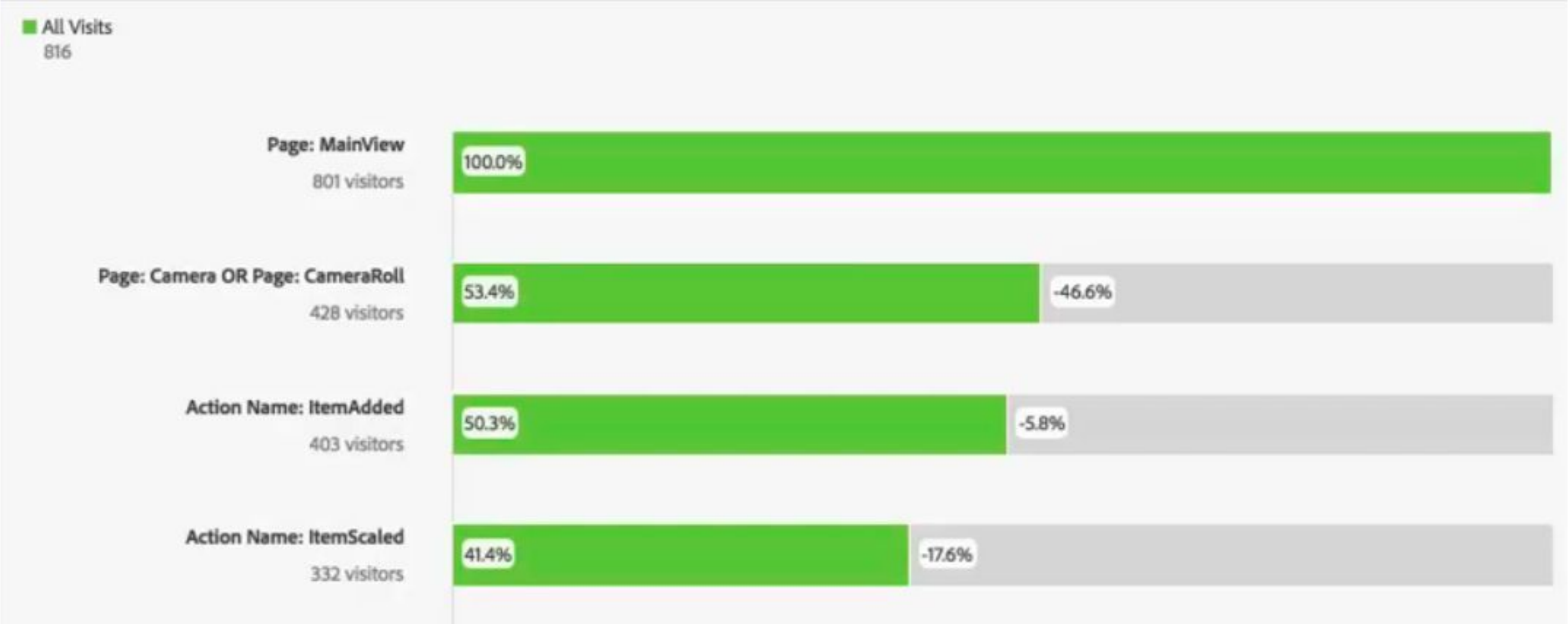
Understanding Visitor Behavior/Journeys: Reports

Where do visitors come before/go after a specific page/site section?



Understanding Visitor Behavior/Journeys: Reports

Where do visitors drop off / fall out?



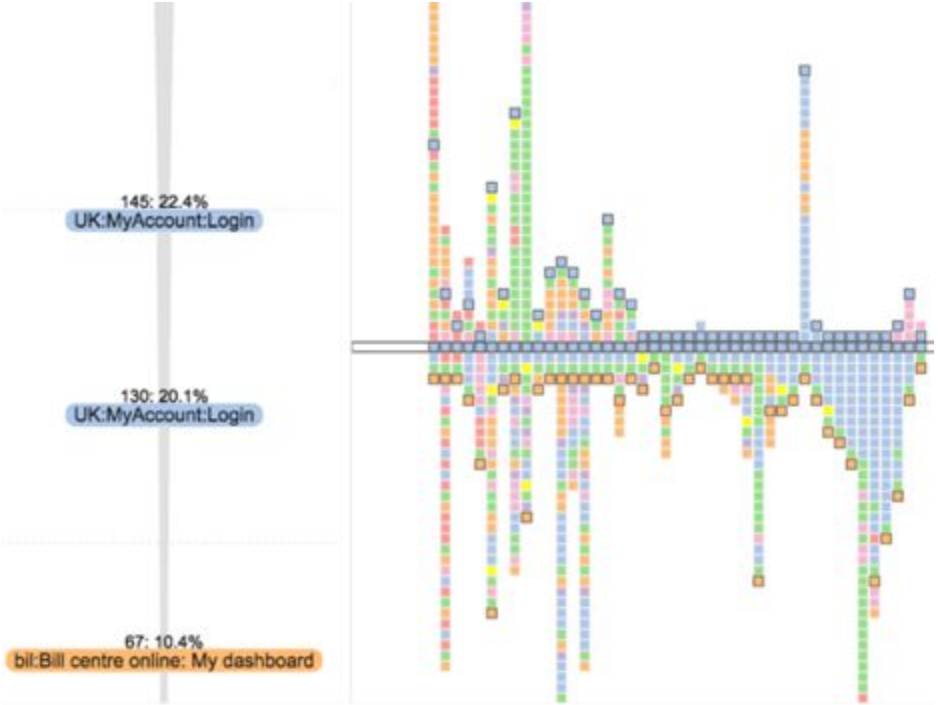
Data Management

Fast, distributed, column-oriented data store and query engine

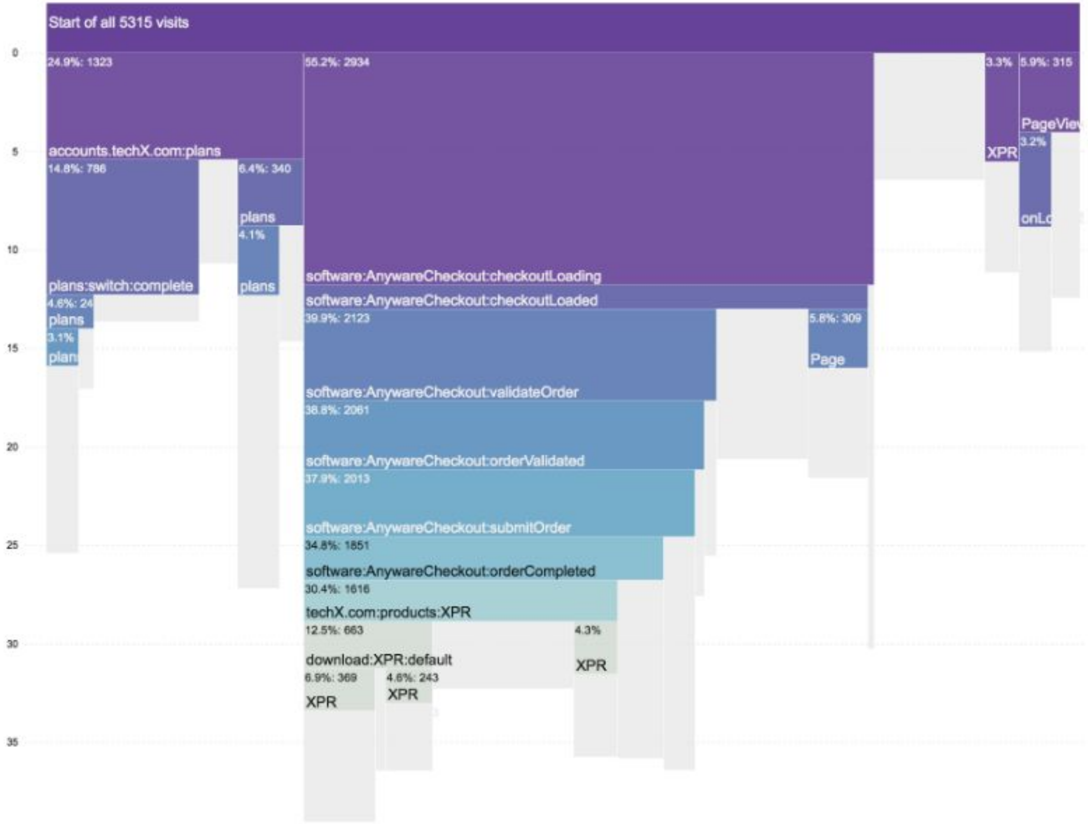
In-house developed algebra

XML-based query language

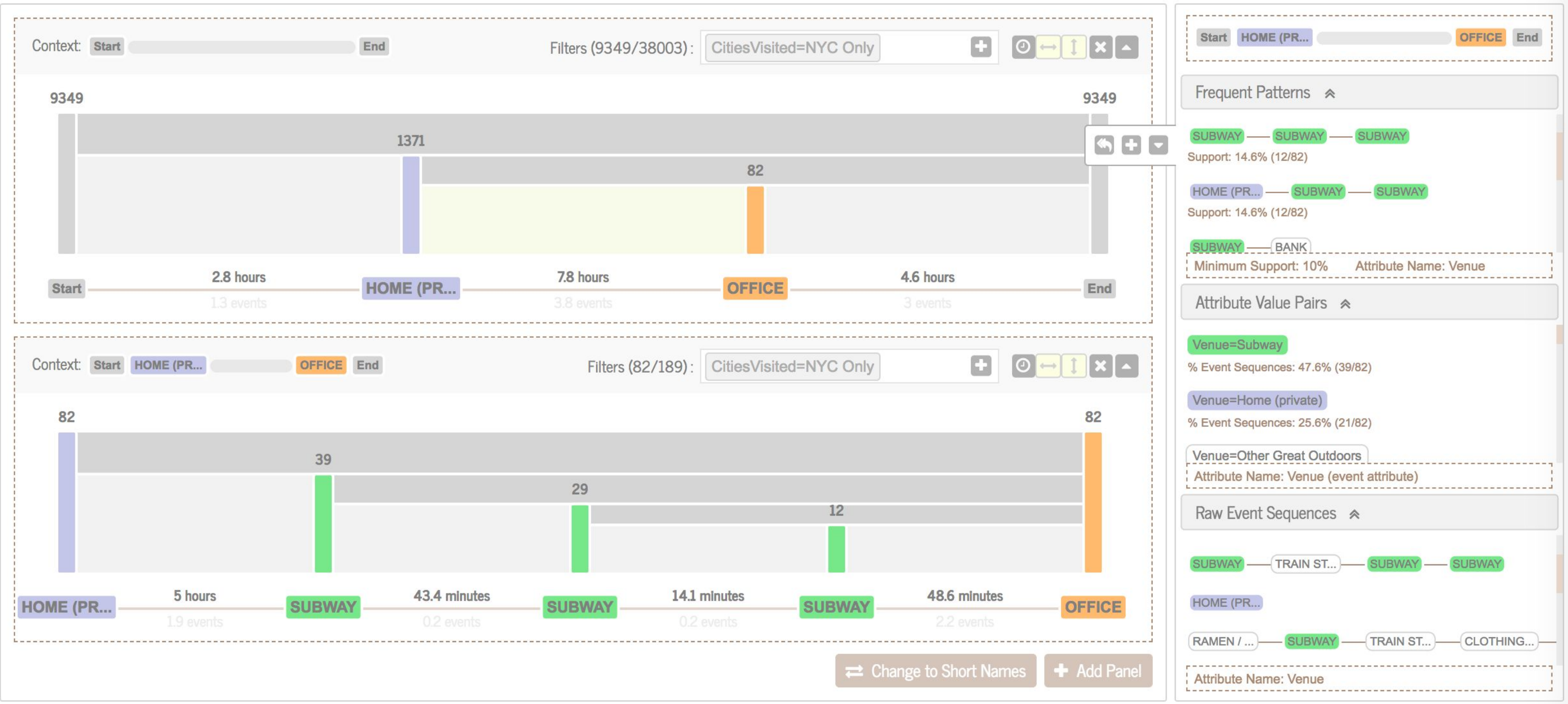
Interactive Analysis: Pattern Mining + Visualization



New Mining Algorithms



Interweaving Pattern Mining with Ad Hoc Queries



Context: Start End

Filters (9349/38003) : CitiesVisited=NYC Only

Context: Start HOME (PR... OFFICE End

Filters (82/189) : CitiesVisited=NYC Only

Frequent Patterns

SUBWAY — SUBWAY — SUBWAY
Support: 14.6% (12/82)

HOME (PR... — SUBWAY — SUBWAY
Support: 14.6% (12/82)

SUBWAY — BANK

Minimum Support: 10%   Attribute Name: Venue

Attribute Value Pairs

Venue=Subway
% Event Sequences: 47.6% (39/82)

Venue=Home (private)
% Event Sequences: 25.6% (21/82)

Venue=Other Great Outdoors

Attribute Name: Venue (event attribute)

Raw Event Sequences

SUBWAY — TRAIN ST... — SUBWAY — SUBWAY

HOME (PR...

RAMEN / ... — SUBWAY — TRAIN ST... — CLOTHING...

Attribute Name: Venue

Collaboration with Product Team

Drastic changes to the architecture/design of the datastore unlikely

Adding new queries/mining algorithms need to justify the value of approach, which is hard without real data and user feedback

Lab features to bridge the gap: test on a selected group of customers with their own data

Collaboration with Product Team

hard, they have their priority, existing architecture cannot be easily changed to support new queries/mining algorithms that may or may not prove to be what the customers need/value

need to demonstrate the value of approach, which is hard without real data. need to show value based on small/sample data

Lab features to bridge the gap: test on a selected group of customers with their own data

Product team’s support and buy-in crucial for incorporating novel ideas

Research Collaboration

Build own datastore with smaller datasets

Work with researchers in other areas, sometimes restricted by company’s talent pool

Learn from existing literature

Collaborate with academia through internships

Hire interns outside your area

The Unified Logging Infrastructure for Data Analytics at Twitter

http://vldb.org/pvldb/vol5/p1771_georgelee_vldb2012.pdf

compute histogram of event counts

construct dictionary: map each event to a symbol: more frequent events are assigned smaller code point (fewer bytes) - variable-length coding

reconstruct sessions: group by on user id and session id, 30 minute interval to delimit sessions, then encode using dictionary. A session is essentially a unicode string

only records session duration, cannot support queries about temporal gaps, lower latency, higher throughput

- summary statistics on sessions
- user sessions that match an arbitrary regular expression (event-based querying)
- funnel analytics, a funnel is a regular expression, grow the funnel substring to query number of sessions at each stage
- n-grams and pattern mining